МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА

Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»

Кафедра математичного моделювання та статистики

Освітньо-професійна програма «Економічна кібернетика і Дата Сайнс»

Галузь знань 05 «Соціальні та поведінкові науки»
Спеціальність 051 «Економіка»

Форма навчання: очна (денна)

КВАЛІФІКАЦІЙНА МАГІСТЕРСЬКА РОБОТА

на тему **«Оцінювання детермінант оплати праці методами Дата Сайнс»**

здобувача Коваленко Анастасії Арсеніївни

Науковий керівник: д.ф.-м.н., професор Ольга ПРИТОМАНОВА
(науковий ступінь, учене звання, ПІБ)

_____
(підпис)

Робота допущена до захисту перед екзаменаційною комісією з атестації здобувачів вищої освіти (ЕК)

Завідувач кафедри: к. ф.-м. н., професор
Галина ВЕЛИКОІВАНЕНКО

_____
(підпис)

Київ 2024

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА

Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»

Кафедра математичного моделювання та статистики

Освітньо-професійна програма «Економічна кібернетика і Дата Сайнс»

Галузь знань 05 «Соціальні та поведінкові науки»
Спеціальність 051 «Економіка»

ПОГОДЖЕНО
Керівник проектної групи (гарант)
освітньо-професійної програми
Олена ПІСКУНОВА

_____
(підпис)

_____ 2024 р.

ЗАТВЕРДЖУЮ
Завідувач кафедри математичного
моделювання та статистики
Галина ВЕЛИКОІВАНЕНКО

_____
(підпис)

_____ 2024 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

здобувачу вищої освіти Коваленко Анастасії Арсеніївні
(прізвище, ім'я, по батькові)
_____очної (денної)_____ **форми навчання**

на підготовку кваліфікаційної магістерської роботи
на тему
«Оцінювання детермінант оплати праці методами Дата Сайнс»

Тему затверджено наказом ректора Університету від «19» вересня 2024 р .№ 1610 / ст

Кваліфікаційна магістерська робота виконується на матеріалах даних офіційних сайтів Національного банку України https://bank.gov.ua/, Державної служби статистики https://www.ukrstat.gov.ua/ та інших відкритих джерел**.**

План кваліфікаційної магістерської роботи

Розділ 1	Теоретичні та методичні основи аналізу та оцінювання детермінант оплати праці
Розділ 2	Статистичний аналіз головних факторів, що впливають на рівень оплати праці
Розділ 3	Оцінювання детермінант оплати праці

Об'єкт дослідження:	оплата праці та фактори, що на неї впливають
Предмет дослідження:	методи дата сайнс для оцінювання головних факторів оплати праці
Мета виконання кваліфікаційної магістерської роботи:	застосування методів машинного навчання та дата сайнс для аналізу та оцінювання детермінант оплати праці

Конкретні завдання, які здобувач повинен виконати для досягнення поставленої мети:
У розділі 1:

1) огляд джерел та економічний аналіз оплати праці та її детермінант;
2) огляд методів статистичного аналізу детермінант оплати праці;
3) огляд методів машинного навчання та дата сайнс для оцінювання детермінант оплати праці.

У розділі 2:

1) інформаційна база для моделювання; графічний аналіз динаміки обраних у першому розділі факторів, що впливають на рівень оплати праці;
2) методи машинного навчання для аналізу факторів, що впливають на рівень оплати праці (розглянути конкретні методи).

У розділі 3:

1) описати обрані методи дата сайнс для оцінювання детермінант оплати праці;
2) порівняння застосованих методів та отриманих результатів моделювання.

Завдання підготував
науковий керівник _____ Ольга ПРИТОМАНОВА
(підпис) *(ім'я, прізвище)*
«__» _____ 2024 р.

Завдання одержав
здобувач _____ Анастасія КОВАЛЕНКО
(підпис) *(ім'я, прізвище)*
«__» _____ 2024 р.

Реферат

Кваліфікаційна магістерська робота містить 73 сторінки, 3 таблиці, 45 рисунків, список використаних джерел з 50 найменувань.

«Оцінювання детермінант оплати праці методами Дата Сайнс»

Об'єктом дослідження кваліфікаційної магістерської роботи є оплата праці та фактори, що на неї впливають.

Предметом дослідження є методи дата сайнс для оцінювання основних факторів оплати праці.

Мета і завдання дослідження. Основною метою кваліфікаційної магістерської роботи є застосування методів машинного навчання та дата сайнс для аналізу та оцінювання детермінант оплати праці.

Відповідно до поставленої мети визначені такі *завдання*:

- визначення основних факторів, що впливають на оплату праці;
- проведення статистичного та візуального аналізу даних;
- оцінити точність побудованих моделей.

Теоретична, методична та практична значущість отриманих результатів. Теоретична значущість полягає в поглибленні розуміння чинників, що визначають рівень оплати праці на основі даних опитаних фахівців з ІТ-сфери. Методична значущість полягає в розробці підходів до використання методів Дата Сайнс для аналізу соціально-економічних явищ.

Практичне значення одержаних результатів дослідження полягає у розробці математичних моделей для оцінювання детермінант оплати праці на вибірках великого обсягу із застосуванням методів дата сайнс.

Рік виконання кваліфікаційної бакалаврської роботи – 2024.

Рік захисту роботи – 2024.

Ключові слова: оплата праці, детермінанти оплати праці, дата сайнс, машинне навчання, економетрична модель, регресія, градієнтний бустинг

В і д г у к
про кваліфікаційну магістерську роботу
здобувача освітньо-професійної програми «Економічна кібернетика і Дата Сайнс»
навчально-наукового інституту
«Інститут інформаційних технологій в економіці»
Коваленко Анастасії Арсеніївни
на тему **«Оцінювання детермінант оплати праці методами Дата Сайнс»**

1. Актуальність теми обумовлена тим, що обсяги даних, необхідні для оцінювання детермінант оплати праці, постійно зростають і сягають десятків тисяч реалізацій. Тому застосування методів Дата Сайнс надає можливість побудувати адекватні математичні моделі у цій сфері.

2. Позитивні риси кваліфікаційної магістерської роботи: у роботі розроблено математичні моделі із застосуванням методів машинного навчання та дата сайнс, розрахунки проведено у програмному середовищі R Studio, що показало достатній рівень автора працювати з бібліографічними джерелами, аналізувати теоретичний та практичний матеріал, обґрунтовувати висновки, застосовувати сучасні інформаційні технології.

3. Наявність самостійних розробок автора: вибір теми та виконання кваліфікаційної роботи проведено автором повністю самостійно. Кваліфікаційна робота відповідає затвердженому індивідуальному завданню та оформлена в основному відповідно до вимог до кваліфікаційних робіт.

4. Цінність теоретичних висновків та практичних рекомендацій, отриманих у роботі, пов’язана з розробленими адекватними математичними моделями оцінювання детермінант оплати праці, що сприятиме розумінню впливу основних факторів на рівень оплати праці.

5. Наявність недоліків: у висновках бажано було б представити результати оцінювання детермінант оплати праці кількісно, не тільки описово, та обґрунтувати, чому досвід роботи та рівень освіти є найважливішими детермінантами оплати праці.

Робота Коваленко А.А. пройшла апробацію на XII Всеукраїнській науково-практичній конференції форумі молодих економістів-кібернетиків «Моделювання економіки: проблеми, тенденції, досвід» (22-23 листопада 2024 року, Львів). За результатами роботи опубліковано тези у збірнику конференції.

6. Загальна оцінка кваліфікаційної магістерської роботи та її допущення до захисту перед ЕК: кваліфікаційна магістерська робота Коваленко А.А. виконана на *належному рівні* та може бути допущена до захисту перед ЕК. Робота свідчить про відповідність набутих здобувачем компетентностей вимогам освітньої програми «Економічна кібернетика і Дата Сайнс», а її автор Коваленко А.А., за умови успішного захисту, заслуговує на присвоєння освітньої кваліфікації «Магістр з економіки».

Науковий керівник: професор кафедри
математичного моделювання та статистики,
доктор фізико-математичних наук, професор _____ Ольга ПРИТОМАНОВА
«12» грудня 2024 р

ЗМІСТ

ВСТУП **3**

РОЗДІЛ 1. ТЕОРЕТИЧНІ ТА МЕТОДИЧНІ ОСНОВИ АНАЛІЗУ ТА ОЦІНЮВАННЯ ДЕТЕРМІНАНТ ОПЛАТИ ПРАЦІ 5

1.1 Економічний аналіз оплати праці та її детермінант 5

1.2 Методи статистичного аналізу детермінант оплати праці 12

1.3 Методи машинного навчання та дата сайнс для оцінювання детермінант ... 18

РОЗДІЛ 2. СТАТИСТИЧНИЙ АНАЛІЗ ОСНОВНИХ ФАКТОРІВ, ЩО ВПЛИВАЮТЬ НА РІВЕНЬ ОПЛАТИ ПРАЦІ 25

2.1 Графічний аналіз динаміки факторів, що впливають на рівень оплати праці 25

2.2 Застосування алгоритмів машинного навчання для аналізу детермінант оплати праці 58

РОЗДІЛ 3. ОЦІНЮВАННЯ ДЕТЕРМІНАНТ ОПЛАТИ ПРАЦІ 60

3.1 Метод регресійних моделей 60

3.2 Метод градієнтного бустингу 67

3.3 Порівняння застосованих методів та отриманих прогнозів 69

ВИСНОВКИ 72

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ 74

ВСТУП

У сучасному світі стрімкого цифрового розвитку постійно з’являються нові засоби аналізу та оцінки ринкових явищ, зокрема у сфері оплати праці. Ось де методи Дата Сайнс можуть забезпечити краще розуміння факторів, які відіграють роль у визначенні зарплати, особливо в IT у цей період зростаючого попиту на кваліфікованих фахівців. Це головним чином сприяє підвищенню ефективності управління людськими ресурсами на основі даних.

Науковці активно досліджують детермінанти заробітної плати за допомогою методів машинного навчання та аналізу великих даних. Закордонні автори, такі як Ернст Берндт, та українські, Балан О. Д. та Савченко Ю. К., аналізували зв'язок між соціально-економічними факторами та доходами. В Україні інтерес до теми також зростає, і дослідження в цій галузі здійснюються на основі даних про ринок праці, зокрема аналітики OLX, Work.ua та міжнародних платформ. Існуючі праці підкреслюють значення таких факторів, як досвід, рівень освіти та спеціалізація, але комплексного аналізу за допомогою Дата Сайнс бракує.

Метою дослідження є застосування методів машинного навчання та дата сайнс для аналізу та оцінювання детермінант оплати праці.

Основні завдання дослідження включають: дослідження теоретичних та методичних підходів до аналізу та оцінювання детермінант оплати праці, зокрема провести економічний аналіз оплати праці та її детермінант, розглянути методи статистичного аналізу детермінант оплати праці, охарактеризувати методи машинного навчання та Data Science для оцінювання детермінант; виконати статистичний аналіз основних факторів, що впливають на рівень оплати праці, а саме: здійснити графічний аналіз динаміки факторів, що впливають на рівень оплати праці, застосувати алгоритми машинного навчання для аналізу детермінант оплати праці; провести оцінювання детермінант оплати праці за

допомогою сучасних методів аналізу даних, зокрема використати метод регресійних моделей для оцінювання детермінант, застосувати метод градієнтного бустингу для аналізу впливу факторів, порівняти результати застосованих методів і зробити висновки щодо отриманих результатів.

Об'єктом дослідження є оплата праці та фактори, що на неї впливають.

Предметом дослідження є методи дата сайнс для оцінювання основних факторів оплати праці.

У дослідженні застосовано методи Дата Сайнс, зокрема регресійний аналіз, методи машинного навчання, візуалізація даних, кластеризація та кореляційний аналіз. Також використовувались статистичні методи для попередньої обробки даних та оцінки результатів.

Теоретична значущість полягає в поглибленні розуміння чинників, що визначають рівень оплати праці на основі даних опитаних фахівців з IT-сфери. Методична значущість полягає в розробці підходів до використання методів Дата Сайнс для аналізу соціально-економічних явищ. Практична значимість роботи полягає у розробці математичних моделей для оцінювання детермінант оплати праці на вибірках великого обсягу із застосуванням методів дата сайнс.

Інформаційною базою дослідження стали статистичні дані з міжнародних і національних джерел, зокрема з платформ Work.ua, OLX, DOU.ua, а також відкритих баз даних, таких як Kaggle, що містять інформацію про фактори, які можуть впливати на рівень заробітної плати фахівців IT-сфери.

РОЗДІЛ 1
ТЕОРЕТИЧНІ ТА МЕТОДИЧНІ ОСНОВИ АНАЛІЗУ ТА ОЦІНЮВАННЯ ДЕТЕРМІНАНТ ОПЛАТИ ПРАЦІ

1.1 Економічний аналіз оплати праці та її детермінант

Детермінанти оплати праці, або ж головні чинники, що визначають рівень заробітної плати, значно еволюціонували протягом століть [1]. Їх розвиток тісно пов'язаний з етапами економічного розвитку суспільства, змінами в соціальних структурах, а також з глобалізацією і прогресом у трудовому законодавстві. Зараз, коли на ринку праці з'являються нові індустрії та технології, традиційні фактори визначення заробітної плати, як-от професійні об’єднання чи кваліфікація працівника, набувають нових форм, а на передній план виходять нові детермінанти, наприклад, гендерна рівність та особисті навички [2].

У період доіндустріального розвитку, особливо в середньовічній Європі, заробітна плата формувалася в рамках ремісничих гільдій. Гільдії контролювали кількість майстрів, які могли працювати в певній сфері, а також встановлювали стандарти якості та ціни на продукцію, що забезпечувало певний рівень стабільності доходів [3]. На той час оплату праці визначали такі чинники, як статус у гільдії, наявність зв’язків та рівень кваліфікації [4]. Пізніше, у 18-19 століттях, зростання масштабів виробництва і зменшення ролі гільдій призвело до створення профспілок, які стали новим механізмом колективного визначення рівня оплати праці.

З початком індустріальної революції, коли виробництво зосередилося у великих фабриках, зросла потреба в нових типах робочої сили. Професії стали більш спеціалізованими, а тому кваліфікація та освіта почали відігравати значну

роль у визначенні оплати праці [5]. На рівень зарплат також впливали робочі умови і статус роботи (наприклад, фізична праця була менш оплачуваною порівняно з інженерними та управлінськими посадами) [6]. В цей період виникають перші закони щодо мінімальної заробітної плати та обмеження тривалості робочого дня, що стало важливим фактором формування державної політики у сфері оплати праці [7].

Після Другої світової війни роль держави у встановленні та регулюванні оплати праці зросла. Зокрема, у багатьох західних країнах було запроваджено політику соціального захисту, яка включала мінімальну заробітну плату, соціальні виплати, а також заходи з охорони здоров'я [8]. Державне втручання дозволило знизити нерівність у доходах і забезпечити базові стандарти життя для всіх працівників. У цей період на рівень зарплат почали впливати такі фактори, як інфляція, економічний ріст і зайнятість, а також підтримка робочих місць у державному секторі [9].

Глобалізація та розвиток інформаційних технологій призвели до нових змін на ринку праці. Високий рівень конкуренції та доступність робочої сили з різних регіонів світу послабили вплив традиційних профспілок, але створили нові детермінанти, такі як рівень володіння мовами, культурна гнучкість і здатність працювати у глобальних командах [10]. Одночасно важливими факторами стали навички у сфері технологій і досвід роботи з новітніми системами автоматизації.

Крім того, дедалі більшого значення набуває питання гендерної рівності та рівності можливостей у різних групах населення [11]. У сучасних умовах, коли дедалі більше жінок активно залучені до ринку праці, різниця в оплаті праці чоловіків і жінок, а також проблеми з кар'єрним просуванням стали значними аспектами, що впливають на політику оплати праці в багатьох компаніях і державах. Наприклад, дослідження показують, що, попри значний прогрес, розрив у заробітній платі між чоловіками та жінками в деяких країнах зберігається [12].

З розвитком дистанційної роботи і нових моделей зайнятості (наприклад, фріланс, гіг-економіка) детермінанти оплати праці продовжують розвиватися.

Наприклад, компанії все більше звертають увагу на гнучкість графіку, можливості віддаленої роботи та індивідуальні мотиваційні програми, що враховують особисті якості та досягнення працівника [13]. Крім того, зростає роль персоналізації компенсаційних пакетів, що включають не тільки базову заробітну плату, а й різні додаткові пільги, як-от оплачуване навчання чи страхування.

Таблиця 1.1 - Еволюція детермінант оплати праці

Період	Фактори, які впливають на оплату праці	Причина появи
Доіндустріальний період	Статус у гільдії, зв'язки, рівень кваліфікації	Контроль гільдіями кількості майстрів, стандарти якості та цін на продукцію [14].
Індустріальна революція	Освіта, кваліфікація, статус роботи	Зростання масштабів виробництва, поява спеціалізованих професій, закони про мінімальну зарплату [15].
Післявоєнний період	Інфляція, економічний ріст, державне втручання	Запровадження мінімальної зарплати, соціальних виплат, базових стандартів життя [16].
Глобалізація	Конкуренція, технологічні навички, гнучкість	Високий рівень конкуренції, автоматизація, потреба у глобальній взаємодії [17].
Сучасність	Рівність можливостей, персоналізація винагород	Розвиток гіг-економіки, боротьба з гендерною нерівністю, нові моделі зайнятості [18].

Джерело: розроблено автором на основі [18]

Рівень оплати праці є важливим показником економічного розвитку, що відображає попит і пропозицію на ринку праці, а також здатність економіки забезпечити достатній рівень добробуту населення [19]. Оплата праці залежить від багатьох чинників, які можна умовно розділити на соціально-економічні, демографічні, інституційні та макроекономічні.

Соціально-економічні фактори визначаються характером і структурою суспільства. Вони впливають на рівень заробітної плати через соціальні норми, рівень життя та інші аспекти [20]. До основних соціальних детермінант відносяться освіта і кваліфікація так як високий рівень освіти та наявність професійних навичок збільшують конкурентоспроможність працівника на ринку праці і, як наслідок, можуть підвищувати рівень оплати праці [21] . У сучасному

світі, де англійська є глобальною мовою, працівники з високим рівнем знання англійської мають переваги та можуть розраховувати на вищу оплату праці.

Демографічні фактори впливають на ринок праці через вікову структуру населення, гендер, рівень міграції та інші аспекти. Наприклад, у багатьох країнах жінки часто отримують меншу заробітну плату порівняно з чоловіками на аналогічних посадах, що обумовлено низкою соціальних і економічних причин, таких як дискримінація та різниця в професійних напрямах [22]. Вік також є суттєвим фактором: молоді працівники, як правило, отримують менше через брак досвіду, тоді як з віком їх дохід зростає до певного рівня, а потім може почати знижуватись. Місце проживання також грає важливу роль: у великих містах, де вартість життя вища, заробітні плати зазвичай вищі, ніж у менших містах або сільській місцевості. Працівники-мігранти зазвичай отримують нижчу заробітну плату, особливо якщо вони працюють у неформальному секторі економіки або в галузях з меншою захищеністю праці [23].

До інституційних факторів належать законодавчі обмеження, вплив профспілок та колективні договори. Наприклад, мінімальна заробітна плата, встановлена державою, є одним з найважливіших інструментів для забезпечення захисту найманих працівників. Вона створює нижню межу заробітної плати, нижче якої роботодавець не може платити, що позитивно впливає на добробут працівників з низькими доходами [24]. Профспілки можуть підвищувати рівень заробітної плати для своїх членів шляхом переговорів і захисту їх прав, що часто є визначальним фактором у великих компаніях і державних установах. Уряди можуть надавати субсидії або підтримку працівникам у низькооплачуваних галузях, що частково компенсує низький рівень оплати праці [25] Податкова політика має прямий вплив на рівень оплати праці, оскільки підвищення податків на доходи працівників може знижувати їх чистий заробіток і створювати додатковий тиск на роботодавців для перегляду зарплат [26].

Макроекономічні фактори включають рівень інфляції, економічне зростання, рівень безробіття та стан економіки в цілому [27]. Наприклад, у

періоди економічного зростання спостерігається збільшення попиту на робочу силу, що сприяє зростанню зарплат. З іншого боку, висока інфляція може знижувати реальну вартість зарплат, навіть якщо номінальні показники зростають [28]. Важливим фактором є також рівень безробіття: при високому рівні безробіття працівники мають менше можливостей вимагати вищу заробітну плату, оскільки пропозиція робочої сили перевищує попит на неї. У країнах з високим рівнем економічного розвитку зазвичай спостерігаються вищі рівні заробітної плати, що обумовлено більш високою продуктивністю праці, наявністю технологій і розвинутим ринком праці [29].

Також виділяють галузеві та технологічні фактори.

Галузеві фактори також значно впливають на рівень оплати праці. Наприклад, у галузях з високою капіталомісткістю і складними технологічними процесами, таких як IT, фінансові послуги та інженерія, рівень оплати праці є вищим через необхідність залучення висококваліфікованих спеціалістів. Технологічний прогрес також має значний вплив на рівень заробітних плат, оскільки він може збільшувати попит на спеціалістів, здатних працювати з новітніми технологіями, і знижувати потребу в менш кваліфікованій робочій силі [30].

Технологічні чинники, такі як автоматизація, інновації та технологічний розвиток, суттєво впливають на оплату праці, трансформуючи структуру ринку праці та створюючи нові вимоги до працівників. Автоматизація призводить до скорочення попиту на низькокваліфіковану працю, адже машини та алгоритми виконують рутинні завдання швидше та ефективніше [31]. Це спричиняє зростання попиту на спеціалістів, здатних обслуговувати ці технології, що підвищує їхню конкурентоспроможність та рівень заробітної плати. Інновації та технологічний розвиток стимулюють появу нових професій та галузей, зумовлюючи високі зарплати у сферах, пов’язаних із програмуванням, аналізом даних і робототехнікою. Однак цей процес може також призводити до розширення нерівності в оплаті праці між працівниками, які володіють

необхідними цифровими навичками, та тими, хто їх не має [32]. Загалом, технологічні зміни сприяють підвищенню продуктивності праці та ефективності економіки, проте вони висувають жорсткі вимоги до адаптації працівників через навчання і перенавчання.

Таблиця 1.2 – Класифікація детермінант оплати праці

Категорія детермінант	Детермінант	Опис
Соціально-економічні	Освіта	Вища освіта зазвичай асоціюється з вищими заробітками, оскільки дає змогу здобути конкурентні навички та знання, підвищуючи вартість працівника для роботодавця.
	Рівень англійської мови	Знання англійської мови є важливим фактором для багатьох міжнародних компаній, де англійська є робочою мовою. Працівники з високим рівнем знань мови мають перевагу на ринку праці та отримують вищу оплату.
Демографічні	Вік	Молоді працівники отримують менші зарплати, але з віком, досвідом та кваліфікацією заробіток зростає. Максимальні зарплати часто припадають на середній вік.
	Стать	Жінки, як правило, заробляють менше за чоловіків на аналогічних посадах, що пов'язано з гендерною нерівністю та розподілом обов'язків у сім'ї. Деякі країни приймають закони для усунення гендерного розриву.
	Місце проживання	У великих містах зарплати зазвичай вищі через високі витрати на життя, у сільських районах – нижчі.
Інституційні	Мінімальна заробітна плата	Визначає нижню межу доходу на ринку праці
	Субсидії	Державні субсидії можуть компенсувати низькі зарплати, покращуючи фінансове становище працівників
	Податки	Вищі податки зменшують чисту зарплату, а знижені податкові ставки можуть підвищити дохід працівників
Макроекономічні	Інфляція	Високий рівень інфляції знижує реальну купівельну спроможність заробітної плати
	Рівень безробіття	Високий рівень безробіття знижує тиск на роботодавців підвищувати зарплату
Галузеві	Галузь	У галузях з високою капіталомісткістю і складними технологічними процесами, таких як IT, фінансові послуги та інженерія, рівень оплати праці є вищим через необхідність залучення висококваліфікованих спеціалістів.
Технологічні	Автоматизація	Автоматизація призводить до скорочення попиту на низькокваліфіковану працю. Це спричиняє зростання попиту на спеціалістів, здатних обслуговувати ці технології, що підвищує їхню конкурентоспроможність

		та рівень заробітної плати.

Джерело: розроблено автором на основі [32]

Державне регулювання оплати праці спрямоване на забезпечення соціальної справедливості, економічного зростання та зменшення нерівності через політики, що включають мінімальні стандарти, захист прав працівників та забезпечення гендерної рівності [33]. Історично воно виникло з потреби захисту вразливих груп, таких як жінки та діти, й надалі розвивалося, охоплюючи мінімальну заробітну плату, оплату понаднормової роботи та галузеві стандарти. Сучасні виклики, зокрема зростання тіньової економіки, гіг-економіки й автоматизації, вимагають адаптації цих механізмів, що включає нові форми захисту та обговорення універсального базового доходу [34]. Ефективність регулювання залежить від балансу між захистом прав працівників і конкурентоспроможністю бізнесу, водночас його роль у розвинутих і країнах, що розвиваються, суттєво різниться.

Культурні відмінності суттєво впливають на механізми формування, сприйняття та регулювання оплати праці в різних країнах [35]. У розвинених економіках, як США чи Велика Британія, акцент робиться на індивідуальні досягнення, тоді як у країнах з колективістською культурою, як Японія чи Південна Корея, важливіша лояльність до компанії та стабільність [36]. Гендерні аспекти також відіграють важливу роль: у прогресивних країнах, як скандинавські, гендерний розрив у зарплаті мінімальний, тоді як у традиційних суспільствах, як країни Близького Сходу, жінки можуть стикатися з дискримінацією [37]. Організаційна культура, мобільність робочої сили та соціальні гарантії також різняться залежно від культурного контексту. У глобалізованому світі міжнародні компанії часто поєднують стандарти оплати праці з місцевими культурними особливостями, адаптуючи свої практики для підтримання конкурентоспроможності та забезпечення справедливості в різних країнах.

Соціальні очікування та нематеріальна мотивація значно впливають на визначення рівня оплати праці, формуючи як внутрішні, так і зовнішні фактори,

що визначають сприйняття працівниками своїх трудових прав і можливостей [38]. Соціальні очікування базуються на суспільних нормах і переконаннях, які різняться залежно від країни, культури й економічної ситуації: наприклад, у розвинених країнах більше уваги приділяється справедливості в оплаті та соціальних пільгах, тоді як у країнах з перехідною економікою очікування зосереджені на стабільності [39]. Нематеріальна мотивація, включно з можливостями кар'єрного росту, гнучким графіком, сприятливим робочим середовищем та корпоративною культурою, може компенсувати обмеження фінансових ресурсів, підвищуючи лояльність і ефективність працівників [40]. Такі підходи є особливо важливими для компаній, які прагнуть утримувати таланти та адаптуватися до соціальних і економічних реалій, забезпечуючи конкурентоспроможність і продуктивність навіть у складних умовах [41].

1.2 Методи статистичного аналізу детермінант оплати праці

Дата Сайнс (наука про дані) – це галузь, яка охоплює широкий спектр методів і технік для збору, аналізу, інтерпретації та візуалізації даних (рис. 1.1)

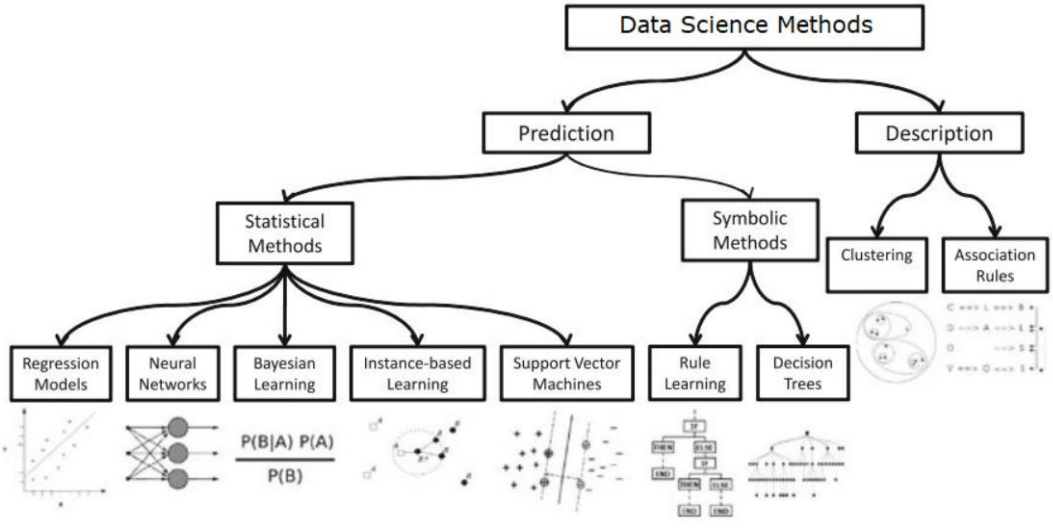

Рисунок 1.1 – Методи Дата Сайнс

Джерело: розроблено автором на основі [42]

Завдяки швидкому розвитку інформаційних технологій, методи Дата Сайнс все частіше використовуються для вирішення складних економічних завдань, таких як прогнозування оплати праці, оцінка факторів, що впливають на заробітну плату, та аналіз тенденцій на ринку праці. У цьому розділі розглянемо основні методи Дата Сайнс, які є актуальними для проведення дослідження, а також детально опишемо ті з них, які застосовуються у роботі.

Статистичний аналіз детермінант оплати праці є основою для вивчення економічних та соціальних детермінант, що впливають на рівень винагороди працівників. Використання різноманітних методів статистичного аналізу дає змогу оцінити не лише рівень та розподіл заробітної плати в окремих групах, але й виявити важливі фактори, що можуть впливати на цей показник. У цьому підрозділі розглядаються основні методи, які використовуються для аналізу детермінант оплати праці, зокрема, описова статистика, кореляційний та регресійний аналізи, а також статистичні тести.

Описова статистика є першим етапом аналізу даних і включає методи для обчислення основних статистичних показників, таких як середнє значення, медіана, дисперсія, стандартне відхилення та інші. Ці показники дозволяють отримати загальне уявлення про розподіл і основні характеристики змінних у наборі даних [43].

Показники центральної тенденції

Середнє арифметичне - найбільш поширений показник центральної тенденції, який визначає середнє значення всіх елементів у вибірці. Використовується для оцінки "типового" рівня заробітної плати в популяції чи вибірці. Наприклад, це може бути середній рівень зарплати в певному секторі чи для конкретної категорії працівників:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad (1.1)$$

де X_i – окремі значення заробітної плати,

n – кількість спостережень

Медіана — це значення, яке розділяє набір даних на дві рівні частини. Це значення, яке стоїть посередині набору даних після їх сортування. Медіана менш чутлива до викидів і може давати більш точну характеристику "типової" зарплати в разі значної варіативності даних. Якщо кількість елементів непарна, то медіаною є середнє значення. Якщо парна — це середнє арифметичне двох центральних значень.

Мода — це значення, яке зустрічається найчастіше в наборі даних. Мода може бути корисною для визначення найбільш популярного рівня заробітної плати в групі працівників.

Показники розсіювання

Дисперсія вимірює варіативність або розсіювання даних навколо середнього:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad (1.2)$$

де σ^2- дисперсія,
X_i – значення заробітної плати,
$\bar{X}$ – середнє значення

Дисперсія є важливим показником для розуміння, наскільки варіативними є зарплати серед працівників.

Стандартне відхилення — це квадратний корінь із дисперсії, що дає змогу оцінити середню відстань значень заробітної плати від середнього:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \qquad (1.3)$$

Це зручний показник для порівняння варіативності різних груп, що дозволяє зробити висновки про рівень нерівності в оплаті праці.

Коефіцієнт варіації - це відношення стандартного відхилення до середнього арифметичного, яке дозволяє порівнювати варіативність між різними наборами даних:

$$CV = \frac{\sigma}{\bar{X}} \times 100\% \quad (1.4)$$

Коефіцієнт варіації часто використовується для порівняння розкиду між різними групами чи періодами, де одна група може мати значно більшу середню зарплату, але також велику варіативність.

У цій роботі описова статистика застосовується для попереднього аналізу даних про оплату праці та фактори, що впливають на неї. За допомогою візуалізації даних можна отримати уявлення про розподіл зарплат, вплив досвіду, рівня освіти та інших факторів.

Кореляційний аналіз використовується для виявлення сили та напрямку зв’язку між різними змінними. Наприклад, можна визначити, наскільки досвід роботи або рівень освіти корелюють із рівнем заробітної плати. У Дата Сайнс кореляційний аналіз часто виконується на початковому етапі дослідження, щоб зрозуміти, які змінні мають найбільший вплив на цільову змінну (у нашому випадку – зарплату).

Кореляційний аналіз дозволяє оцінити ступінь зв'язку між двома або більше змінними, наприклад, між рівнем освіти, досвідом та заробітною платою. Одним з основних індикаторів у цьому аналізі є коефіцієнт кореляції Пірсона.

Коефіцієнт кореляції Пірсона - це міра лінійної залежності між двома змінними. Коефіцієнт кореляції Пірсона може варіюватися від -1 (від'ємна кореляція) до +1 (позитивна кореляція), де 0 означає відсутність лінійної залежності:

$$r = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2 \Sigma(Y_i - \bar{Y})^2}} \quad (1.5)$$

де X_i та Y_i – значення двох змінних,
$\bar{X}$ та $\bar{Y}$ – їх середні значення

Коефіцієнт кореляції Пірсона визначає, наскільки сильно змінюється заробітна плата залежно від зміни іншої змінної (наприклад, освіти чи досвіду).

У даній роботі кореляційний аналіз використовується для виявлення взаємозв’язків між зарплатою та такими змінними, як досвід роботи, освіта та тип компанії. Це допомагає визначити найбільш значущі фактори, які враховуються у подальших етапах моделювання.

Регресійний аналіз – це статистичний метод, що дозволяє моделювати та виявляти залежність між залежною змінною та одним або декількома незалежними змінними. Найбільш поширеним підходом є лінійна регресія, що дозволяє визначити лінійну залежність між змінними. Проте в реальних дослідженнях часто застосовують поліноміальну, логістичну та багатофакторну регресію для складніших взаємозв'язків.

Регресійний аналіз використовується для оцінки залежності заробітної плати від кількох факторів. Це дозволяє не лише виявити зв'язки між змінними, але й передбачити рівень заробітної плати для нових даних.

Лінійна регресія дозволяє моделювати залежність між однією незалежною змінною та залежною змінною (заробітною платою):

$$y = \beta_0 + \beta_1 X + \epsilon \tag{1.6}$$

де Y – залежна змінна (заробітна плата),
X – незалежна змінна,
β_0 – вільний член,
β_1 – коефіцієнт регресії,
ϵ – випадкова помилка

Лінійна регресія дозволяє визначити, чи існує лінійний зв'язок між залежною змінною (яку ми намагаємося передбачити) та однією або кількома незалежними змінними. Лінійна регресія також допомагає оцінити, наскільки кожен з незалежних факторів впливає на залежну змінну, і який напрямок має цей вплив (позитивний чи негативний).

Множинна регресія використовується для оцінки впливу кількох незалежних змінних (наприклад, досвіду, освіти та статі) на залежну змінну (заробітну плату):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \quad (1.7)$$

Множинна регресія дозволяє одночасно враховувати декілька факторів, що впливають на заробітну плату.

У цій роботі регресійний аналіз використовується для побудови моделі, яка оцінює вплив різних факторів на рівень заробітної плати. Зокрема, застосовуються лінійні та поліноміальні регресійні моделі для моделювання залежності між зарплатою та досвідом роботи, освітою та іншими змінними.

Статистичні тести

Т-тест використовують для порівняння середніх значень заробітної плати між двома групами, наприклад, між чоловіками та жінками:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (1.8)$$

де $\bar{X}$ – середнє значення,
σ^2 – дисперсія,
n – розмір вибірки

Це дозволяє визначити, чи є статистично значущі відмінності між групами.

ANOVA (аналіз дисперсії) використовується для порівняння середніх значень між кількома групами:

$$F = \frac{\text{Міжгрупова варіація}}{\text{Внутрішньогрупова варіація}} \quad (1.9)$$

Це дає змогу оцінити, чи існують відмінності між групами на основі аналізу варіацій у даних.

1.3 Методи машинного навчання та дата сайнс для оцінювання детермінант оплати праці

Машинне навчання є невід'ємною частиною Дата Сайнс, що включає широкий набір алгоритмів для виявлення закономірностей у великих масивах даних. Серед найбільш популярних методів – дерево рішень, випадковий ліс, методи кластеризації (наприклад, k-середніх), метод опорних векторів (SVM) та нейронні мережі. Машинне навчання дозволяє побудувати більш точні предиктивні моделі та оптимізувати процес прийняття рішень.

У рамках цієї роботи застосовуються кілька методів машинного навчання для покращення точності прогнозів щодо рівня оплати праці з запропонованих: метод дерево рішень, кластерний аналіз, метод k-середніх, аналіз головних компонентів, логістична регресія, дискримінантний аналіз та градієнтний бустинг.

Дерево рішень — це структура даних, що використовується для прийняття рішень і для побудови прогностичних моделей у машинному навчанні. Воно виглядає як діаграма, що розгалужується у вигляді дерева, де кожен вузол представляє умову або запитання, а кожна гілка — можливі відповіді на це питання. Листи дерева рішень показують кінцевий результат або прогноз.

Дерево рішень — це модель, яка розбиває простір даних на підгрупи, щоб знайти оптимальні шляхи для прийняття рішень або класифікації на основі вхідних змінних. Ця модель використовується для побудови алгоритмів класифікації та регресії.

Формули для дерева рішень зазвичай не існує у вигляді єдиної алгебраїчної виразу, оскільки воно працює через ітеративний процес розділення даних. Проте основні поняття можна виразити так:

$$I(T) = H(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} H(S_i) \tag{1.10}$$

де $H(S)$ – ентропія набору даних S
S_i – підмножини, на які розділяється S

Цей метод дозволяє визначити категорії, до якої належить новий зразок даних, а також розуміння важливих характеристик та залежностей між змінними.

Кластерний аналіз використовується для поділу об’єктів на групи на основі схожості їх характеристик. Для дослідження оплати праці кластерний аналіз може бути застосований для сегментації ринку праці та виявлення груп працівників з подібними характеристиками, наприклад, за рівнем освіти, досвідом або спеціалізацією.

Кластерний аналіз — це метод статистичного аналізу, який використовується для групування об'єктів у групи (кластери) на основі їх схожості. Мета методу — зробити так, щоб об’єкти в одному кластері були максимально схожими між собою, а об’єкти в різних кластерах — максимально відмінними.

Кластеризація не передбачає заздалегідь визначених груп і належить до методів некерованого навчання.

Метод k-середніх є алгоритмом кластеризації, який групує об’єкти у кластери за схожістю. У контексті оплати праці цей метод може бути використаний для кластеризації працівників за профілями (наприклад, за рівнем освіти, віком, стажем роботи) і виявлення схожих груп.

Алгоритм роботи**:**

1. Вибір кількості кластерів *k*.
2. Ініціалізація центроїдів кластерів.
3. Призначення кожного об’єкта до найближчого центроїда.

4. Оновлення положень центроїдів.
5. Повторення кроків 3-4 до збіжності.

Формула для оновлення центроїдів:

$$C_j = \frac{1}{n_j} \sum_{i \in C_j} x_i \tag{1.11}$$

де C_j – центр кластеру j,
n_j – кількість об'єктів у кластерів j,
x_j – об'єкти, що належать до кластери j

У цьому дослідженні кластерний аналіз використовується для сегментації працівників на групи, що мають схожі профілі, що допомагає краще зрозуміти різні категорії працівників і їх вплив на рівень заробітної плати.

Аналіз головних компонент (Principal Component Analysis, PCA) – це метод зниження розмірності даних, що дозволяє зменшити кількість змінних без втрати значущої інформації. Це особливо корисно для великих наборів даних, де присутні численні корельовані змінні. PCA допомагає виділити найважливіші компоненти, що пояснюють найбільшу частину варіацій у даних.

Центрування даних: Для початку від кожної змінної віднімається її середнє значення:

$$x'_{ij} = x_{ij} - \bar{x}_j \tag{1.12}$$

де x_{ij} – значення j-ї змінної для і-го спостереження,
$\bar{x}_j$ – середнє значення j-ї змінної

Обчислення коваріаційної матриці:

$$C = \frac{1}{n-1} X'X \tag{1.13}$$

де C – коваріаційна матриця,
X – матриця центрованих даних

Знаходження власних векторів і власних значень: Розв'язується задача:

$$C\omega = \lambda\omega \tag{1.14}$$

де λ – власні значення, які відповідають дисперсіям головних компонент,
ω – власні вектори, які визначають напрямки головних компонент

Формування головних компонент: Головні компоненти обчислюються як лінійна комбінація змінних:

$$PC_k = X\omega_k \tag{1.15}$$

де PC_k – k-та головна компонента,
ω_k – власний вектор, відповідний k-му власному значенню

У даному дослідженні PCA застосовується для оптимізації моделі та підвищення її точності шляхом виділення основних компонентів, які найбільше впливають на рівень зарплати. Це допомагає зменшити обсяг обчислень і спростити аналіз, зосереджуючись на ключових факторах.

Логістична регресія використовується для моделювання залежності бінарної залежної змінної від одного або кількох незалежних змінних. У контексті аналізу заробітної плати логістична регресія може бути застосована для прогнозування ймовірності належності працівника до певної категорії (наприклад, рівень заробітної плати: низький, середній чи високий):

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}} \tag{1.16}$$

де $P(Y = 1|X)$ – ймовірність позитивного результату,
β_0 – константа,
β_1, β_p – коефіцієнти регресії,
X_1, X_p – незалежні змінні

Логістична регресія є ефективним інструментом для ідентифікації детермінант, які найбільш суттєво впливають на заробітну плату, наприклад, стать, освіта, стаж роботи [44].

Лінійний дискримінантний аналіз (LDA) використовується для класифікації та зниження розмірності даних. Він виявляє лінійні комбінації змінних, які максимально розділяють класи [45]. У контексті заробітної плати LDA можна застосовувати для класифікації працівників за рівнем доходу:

$$Z = a_0 + a_1X_1 + a_2X_2 + \cdots + a_pX_p \quad (1.17)$$

де Z – дискримінантний бал,
$a_0, a_1, \ldots, a_P$ *– коефіцієнти дискримінантної функції,*
X_1, X_p *– незалежні змінні*

Модель Lasso (Least Absolute Shrinkage and Selection Operator) — це метод регресії, який виконує одночасну регуляризацію та вибір ознак. Lasso додає до функції втрат штраф, що пропорційний сумі абсолютних значень коефіцієнтів моделі. Це допомагає зменшити кількість ознак, зробивши частину коефіцієнтів рівними нулю, що сприяє інтерпретованості моделі:

$$L(\beta) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{P}|\beta_j| \quad (1.18)$$

де n – кількість спостережень,
p – кількість ознак,
y_i *– спостережуване значення цільової змінної,*
x_{ij} *– значення j-ї ознаки для і-го спостереження,*
β_j *– коефіцієнт моделі,*
λ *– параметр регуляризації, який контролює ступінь штрафу*

Коли λ збільшується, Lasso примушує більше коефіцієнтів β_j ставати нульовими.

Її застосовують для запобігання перенавчанню моделей шляхом накладання штрафу на абсолютні значення коефіцієнтів, що зменшує складність моделі. Це дозволяє ефективно виділяти найзначущіші змінні, роблячи модель інтерпретованою та зручною для аналізу [46]. У прогнозних задачах Lasso забезпечує високу точність за рахунок відкидання неінформативних ознак, що робить її надзвичайно корисною у галузях, де кількість змінних значно перевищує кількість спостережень.

Градієнтний бустинг — це алгоритм машинного навчання, що поетапно створює сильну модель шляхом комбінації слабких моделей, найчастіше дерев рішень, для підвищення точності передбачень.

Градієнтний бустинг базується на побудові послідовності дерев рішень, де кожне наступне дерево фокусується на помилках попередніх:

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{1.19}$$

де y_i – фактичне значення,
$\hat{y}_i$ – прогнозоване значення

Його основне призначення полягає в усуненні помилок попередніх моделей, поліпшенні точності прогнозів, зменшенні біасу та дисперсії, а також врахуванні нелінійних взаємозв'язків між змінними [47]. Завдяки гнучкості алгоритму, він широко використовується для завдань класифікації, регресії, оцінювання важливості факторів, і може ефективно працювати з різними типами даних, роблячи його універсальним інструментом аналізу.

Також було розглянуто деякі ключові методи аналізу даних. Це статистичні показники точності моделі: крос-валідацію, метрики оцінювання моделей (RMSE, MAE, R^2, AUC) та SHAP (Shapley Additive Explanations).

Крос-валідація (Cross-Validation) — це метод перевірки ефективності моделей машинного навчання шляхом розбиття даних на навчальну та тестову вибірки [48]. Основна ідея — забезпечити, щоб кожна частина даних використовувалася і для навчання, і для тестування.

Формально, крос-валідація поділяє дані на k частин (фолдів). Для кожного фолду модель тренується на $k-1$ частинах, а тестується на одній залишеній. Середній показник ефективності обчислюється за всі фолди:

$$CV_{mean} = \frac{1}{k} \sum_{i=1}^{k} Error_i \tag{1.20}$$

Метрики оцінювання моделей (RMSE, MAE, R^2, AUC)

Оцінювання точності моделей є ключовим аспектом аналізу даних. У R реалізовано велику кількість метрик для цього [49].

RMSE (Root Mean Square Error) - вимірює середню квадратичну похибку:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1.21}$$

Менше значення вказує на кращу модель.

MAE (Mean Absolute Error) - вимірює середню абсолютну похибку:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1.22}$$

R^2 (коефіцієнт детермінації) - показує, яку частку варіації залежної змінної пояснює модель:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1.23}$$

AUC (Area Under Curve) - використовується для оцінки класифікаційних моделей, зокрема їх здатності розрізняти класи. Значення від 0 до 1, де більше означає кращу класифікацію.

SHAP (Shapley Additive Explanations) — це метод пояснення прогнозів моделей машинного навчання. Заснований на теорії кооперативних ігор, SHAP оцінює внесок кожної змінної у результат моделі:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!}\left[f(S \cup \{i\}) - f(S)\right] \tag{1.24}$$

де N – набір усіх змінних,
S – підмножина змінних,
$f(S)$ – прогноз моделі для набору S

РОЗДІЛ 2
СТАТИСТИЧНИЙ АНАЛІЗ ОСНОВНИХ ФАКТОРІВ, ЩО ВПЛИВАЮТЬ НА РІВЕНЬ ОПЛАТИ ПРАЦІ

2.1 Графічний аналіз динаміки факторів, що впливають на рівень оплати праці

У цьому розділі було використано програмну мову R разом із інтерактивним середовищем розробки RStudio. Ці інструменти забезпечують широкі можливості для аналізу даних, створення візуалізацій та побудови моделей машинного навчання, що дозволяють виконувати прогнозування обраних показників із високою точністю та ефективністю.

Для подальшого дослідження чинників, що впливають на оплату праці в Україні, було використано дані, отримані внаслідок анонімного опитування, організованого платформою DOU.ua протягом останніх п’яти років. У рамках цього опитування було зібрано понад 70 тисяч анкет від спеціалістів IT-сфери різних рівнів кваліфікації та напрямків діяльності [50].

Інформація була отримана з відкритого репозиторію devua/csv/salaries, розміщеного на платформі GitHub [50]. Цей ресурс, створений українською спільнотою розробників DOU.ua, містить детальні відповіді респондентів, кожен запис у датасеті відповідає окремому учаснику опитування.

Після попередньої обробки даних за останні п’ять років, обраний датасет став містити 70 132 рядків та 14 стовпчиків (рис. 2.1а та рис. 2.1б):

Рік	Вік	Стать	Місто	Освіта	Посада	Загальний стаж роботи в IT (в роках)	Ваш тайтл на цій посаді	Основна спеціалізація
2021	29	Чоловік	Вінниця	Вища	Software Engineer	9	Team Lead	Front-end
2021	31	Чоловік	Харків	Вища	Software Engineer	10	Team Lead	Full Stack
2021	22	Жінка	Київ	Вища	Project Manager	6	Немає тайтлу	Не вказано
2021	23	Чоловік	Київ	Вища	Software Engineer	3	Middle	Front-end
2021	23	Жінка	Дніпро	Вища	Software Engineer	2	Middle	Front-end
2021	22	Жінка	Київ	Вища	Designer	1.5	Немає тайтлу	Не вказано

Рисунок 2.1а – Початок датасету

Джерело: розроблено автором на основі [50]

Основна мова програмування	Рівень англійської	Зарплата у $ за місяць, лише ставка після сплати податків	Розмір компанії	Тип компанії
TypeScript	Upper-Intermediate	646792	до 50	Стартап
C# / .NET	Upper-Intermediate	470191	понад 1000	Аутсорсингова
Не вказано	Upper-Intermediate	308125	до 50	Продуктова
JavaScript	Upper-Intermediate	117658	до 200	Аутсорсингова
TypeScript	Upper-Intermediate	109314	понад 1000	Продуктова
Не вказано	Intermediate	77251	до 50	Продуктова

Рисунок 2.1б – Кінець датасету

Джерело: розроблено автором на основі [50]

Довгі назви стовпців для більшої зручності було перейменовано (рис. 2.2):

```
data <- data %>%
  rename(Стаж = `Загальний стаж роботи в IT (в роках)`,
         `Ваш тайтл` = `Ваш тайтл на цій посаді`,
         Зарплата = `Зарплата у $ за місяць, лише ставка після сплати податків`)
```

Рисунок 2.2 – Перейменування довгих назв стовпців

Джерело: розроблено автором на основі [50]

Набір даних після оновлення назв колонок для зручності аналізу (рис. 2.3а та рис. 2.3б):

Рік	Вік	Стать	Місто	Освіта	Посада	Стаж	Ваш тайтл	Основна спеціалізація
2021	29	Чоловік	Вінниця	Вища	Software Engineer	9	Team Lead	Front-end
2021	31	Чоловік	Харків	Вища	Software Engineer	10	Team Lead	Full Stack
2021	22	Жінка	Київ	Вища	Project Manager	6	Немає тайтлу	Не вказано
2021	23	Чоловік	Київ	Вища	Software Engineer	3	Middle	Front-end
2021	23	Жінка	Дніпро	Вища	Software Engineer	2	Middle	Front-end
2021	22	Жінка	Київ	Вища	Designer	1.5	Немає тайтлу	Не вказано

Рисунок 2.3а – Початок набору даних із оновленими назвами стовпців

Джерело: розроблено автором на основі [50]

Основна мова програмування	Рівень англійської	Зарплата	Розмір компанії	Тип компанії
TypeScript	Upper-Intermediate	646792	до 50	Стартап
C# / .NET	Upper-Intermediate	470191	понад 1000	Аутсорсингова
Не вказано	Upper-Intermediate	308125	до 50	Продуктова
JavaScript	Upper-Intermediate	117658	до 200	Аутсорсингова
TypeScript	Upper-Intermediate	109314	понад 1000	Продуктова
Не вказано	Intermediate	77251	до 50	Продуктова

Рисунок 2.3б – Кінець набору даних із оновленими назвами стовпців

Джерело: розроблено автором на основі [50]

Для визначення типів даних у кожному стовпчику датасету можна скористатися функцією str(). Ця функція відображає структуру об'єкта,

включаючи типи даних змінних, що дозволяє швидко оцінити, які саме типи даних містить кожен стовпчик (рис. 2.4а):

```
tibble [70,132 × 14] (S3: tbl_df/tbl/data.frame)
 $ Рік                       : chr [1:70132] "2021" "2021" "2021" "2021" ...
 $ Вік                       : chr [1:70132] "29" "31" "22" "23" ...
 $ Стать                     : chr [1:70132] "Чоловік" "Чоловік" "Жінка" "Чоловік" ...
 $ Місто                     : chr [1:70132] "Вінниця" "Харків" "Київ" "Київ" ...
 $ Освіта                    : chr [1:70132] "Вища" "Вища" "Вища" "Вища" ...
 $ Посада                    : chr [1:70132] "Software Engineer" "Software Engineer" "Proje
 $ Стаж                      : chr [1:70132] "9" "10" "6" "3" ...
 $ Ваш тайтл                 : chr [1:70132] "Team Lead" "Team Lead" "Немає тайтлу" "Middle'
 $ Основна спеціалізація     : chr [1:70132] "Front-end" "Full Stack" "Не вказано" "Front-e
 $ Основна мова програмування: chr [1:70132] "TypeScript" "C# / .NET" "Не вказано" "JavaScr
 $ Рівень англійської        : chr [1:70132] "Upper-Intermediate" "Upper-Intermediate" "Upp
 $ Зарплата                  : chr [1:70132] "646792" "470191" "308125" "117658" ...
 $ Розмір компанії           : chr [1:70132] "до 50" "понад 1000" "до 50" "до 200" ...
 $ Тип компанії              : chr [1:70132] "Стартап" "Аутсорсингова" "Продуктова" "Аутсор
```

Рисунок 2.4а – Перший результат функції str()

Джерело: розроблено автором на основі [50]

Результати функції str() показують, що датасет складається з 70,132 рядків та 14 стовпців, усі з яких мають тип даних chr (character), що вказує на текстові значення. Стовпці "Стать", "Місто", "Освіта", "Посада", "Ваш тайтл", "Основна спеціалізація", "Основна мова програмування", "Рівень англійської", "Тип компанії" дійсно містять текстові дані. Однак навіть стовпці, що мають містити числову інформацію, як-от "Рік", "Вік", "Стаж" та "Зарплата", також мають тип chr, що вказує на необхідність їх перетворення у числовий формат для подальшого аналізу. В результаті ми бачимо, що деякі стовпці потребують попередньої обробки даних, щоб їх можна було правильно використовувати для моделювання або аналізу. Наприклад, стовпці "Рік", "Вік", "Стаж" та "Зарплата" повинні бути переведені у формат numeric, тоді як текстові дані, такі як "Ваш тайтл" або "Рівень англійської", краще перетворити на фактори, що дозволить їх ефективно використовувати для категоризації.

Згідно результатів вище за допомогою функцій as.numeric() і as.factor() тип змінних певних стовпців було перетворено у числовий або факторний відповідно (рис. 2.5):

```
data$Рік <- as.numeric(data$Рік)
data$Вік <- as.numeric(data$Вік)
data$Стаж <- as.numeric(data$Стаж)
data$Зарплата <- as.numeric(data$Зарплата)
data$Стать <- as.factor(data$Стать)
data$Місто <- as.factor(data$Місто)
data$Освіта <- as.factor(data$Освіта)
data$`Ваш тайтл` <- as.factor(data$`Ваш тайтл`)
data$`Рівень англійської` <- as.factor(data$`Рівень англійської`)
data$`Розмір компанії` <- as.factor(data$`Розмір компанії`)
data$`Тип компанії` <- as.factor(data$`Тип компанії`)
```

Рисунок 2.5 – Перетворення певних стовпців у потрібний формат

Джерело: розроблено автором на основі [50]

При повторному використанні функції str() було перевірено чи змінився формат змінних у той, який ми вказали на рисунку 2.5 (рис. 2.4б):

```
tibble [70,132 × 14] (S3: tbl_df/tbl/data.frame)
 $ Рік                      : num [1:70132] 2021 2021 2021 2021 2021 ...
 $ Вік                      : num [1:70132] 29 31 22 23 23 22 26 29 38 45 ...
 $ Стать                    : Factor w/ 2 levels "Жінка","Чоловік": 2 2 1 2 1 1 2 2
 $ Місто                    : Factor w/ 25 levels "Вінниця","Дніпро",..: 1 20 9 9 2
 $ Освіта                   : Factor w/ 7 levels "Вища","Дві вищі",..: 1 1 1 1 1 1 1
 $ Посада                   : chr [1:70132] "Software Engineer" "Software Engineer"
 $ Стаж                     : num [1:70132] 9 10 6 3 2 1.5 8 10 13 10 ...
 $ Ваш тайтл                : Factor w/ 10 levels "Architect","Head",..: 8 8 10 6 6
 $ Основна спеціалізація    : chr [1:70132] "Front-end" "Full Stack" "Не вказано" "F
 $ Основна мова програмування: chr [1:70132] "TypeScript" "C# / .NET" "Не вказано" "J
 $ Рівень англійської       : Factor w/ 6 levels "Advanced","Elementary",..: 5 5 5 5
 $ Зарплата                 : num [1:70132] 646792 470191 308125 117658 109314 ...
 $ Розмір компанії          : Factor w/ 6 levels "до 10 спеціалістів",..: 4 6 4 3 6
 $ Тип компанії             : Factor w/ 7 levels "Аутсорсингова",..: 6 1 5 1 5 5 1 1
```

Рисунок 2.4б – Другий результат функції str()

Джерело: розроблено автором на основі [50]

Після виконання змін за допомогою функцій для перетворення типів даних за допомогою функцій as.numeric() і as.factor(), ми бачимо, що деякі стовпці, які були раніше представлені як текстові дані (character), тепер стали числовими або факторними змінними.

Для кращого розуміння структури даних та виявлення зв’язків між окремими змінними було проведено комплексний аналіз із використанням методів візуалізації.

Було створено гістограму для аналізу розподілу респондентів за віком (рис. 2.6):

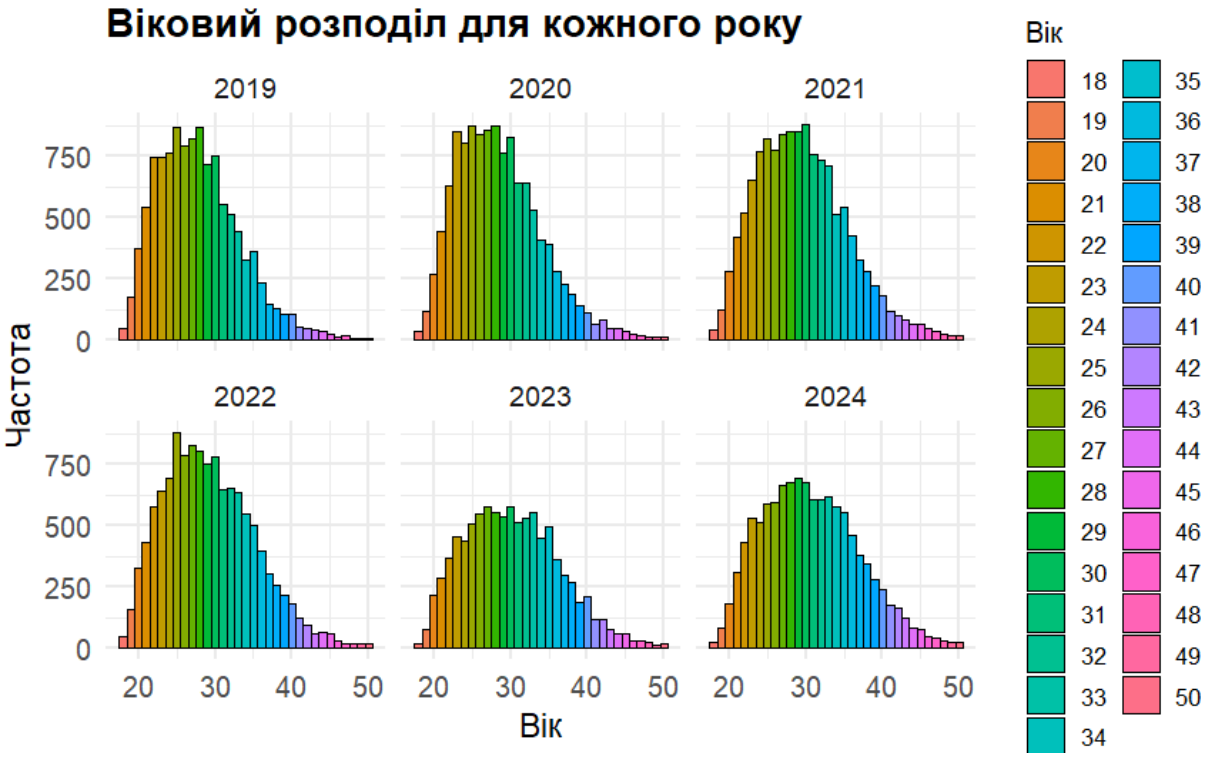

Рисунок 2.6 – Віковий розподіл респондентів

Джерело: розроблено автором на основі [50]

Діаграма демонструє динаміку вікового розподілу респондентів протягом 2019-2024 років. Спостерігається загальна тенденція до зсуву розподілу у бік старших вікових груп. Якщо у 2019 році пік припадав на вікову категорію 20-25 років, то у 2024 році максимум спостерігається у віковій групі 25-30 років. Це може свідчити про старіння вибірки або про зміну цільової аудиторії дослідження. Також варто зазначити, що форма розподілу залишається відносно стабільною протягом усього періоду, що вказує на збереження загальної структури вибірки.

На гістограмах видно, що основна частина респондентів належить до вікової категорії 25-35 років, причому піковий вік у 2024 році становить близько

30 років. Ця вікова група, ймовірно, є найактивнішою в IT-сфері, яка традиційно приваблює молодих фахівців із кількарічним досвідом роботи. Рівень активності респондентів зменшується як у молодшому віці (до 20 років), так і в старшому віці (після 40 років), що можна пояснити як меншим досвідом у молодих спеціалістів, так і ймовірним переходом старших фахівців на керівні посади чи в інші галузі.

Гістограма також ілюструє достатньо рівномірний розподіл між віковими групами 25-40 років, що свідчить про стабільність у заповненні вакансій IT-галузі серед фахівців у цьому діапазоні. Дані підтверджують, що респонденти переважно є людьми працездатного віку, а значна частина молоді, ймовірно, ще навчається або не має достатнього досвіду для входу в IT-сферу.

Потім було досліджено залежність рівня оплати праці від віку. Результати представлено на рисунку 2.7:

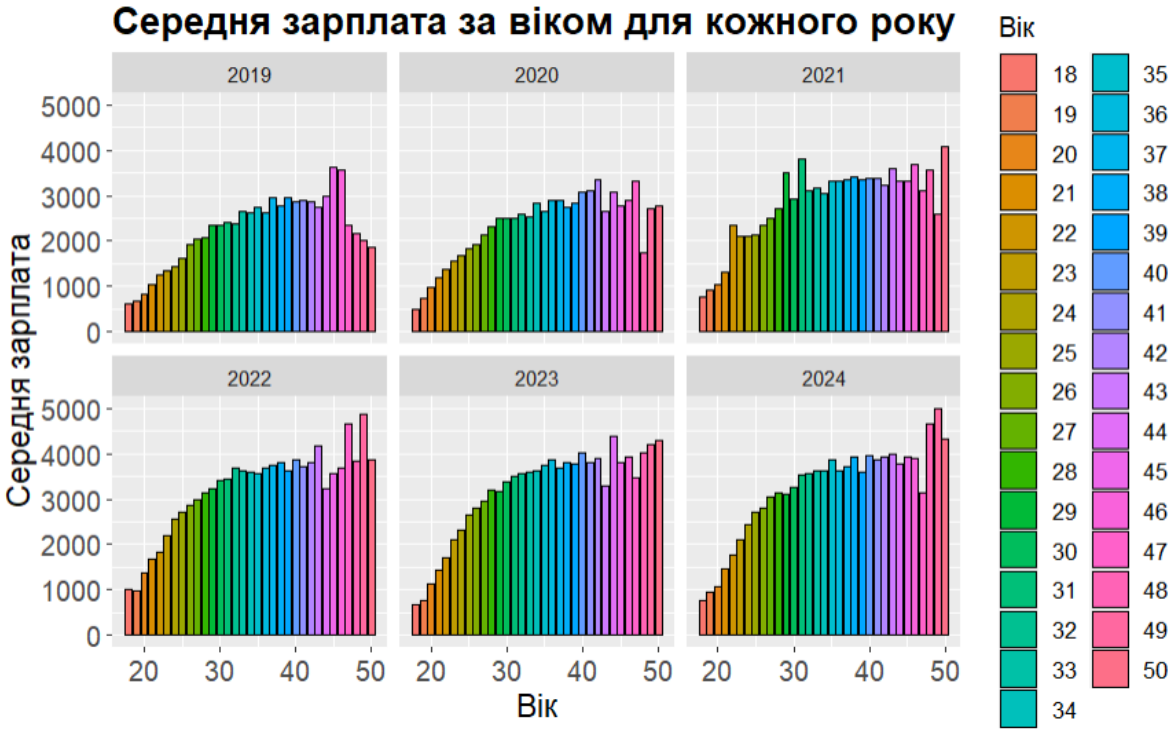

Рисунок 2.7 – Зміни середньої заробітної плати в залежності від віку працівників протягом 2019-2024 років

Джерело: розроблено автором на основі [50]

Загалом, можна спостерігати, що з віком рівень заробітної плати зростає. Це цілком логічно, оскільки з досвідом та набуттям нових навичок працівники стають більш цінними для компаній і, відповідно, отримують вищу оплату. Однак, якщо порівнювати різні роки, то можна помітити деякі нюанси. Наприклад, у 2021 та 2022 роках спостерігається більш виражене зростання зарплат для молодих спеціалістів віком до 30 років. Це може свідчити про підвищення попиту на молодих фахівців з певними навичками та готовність компаній пропонувати їм більш конкурентоспроможну зарплату. Також, можна відзначити, що у більш старших вікових категорій ріст зарплат відбувається більш плавно.

Цікавою особливістю є те, що в останні роки спостерігається тенденція до згладжування різниці в заробітній платі між різними віковими групами. Тобто, молоді спеціалісти отримують вищі зарплати, а досвідчені працівники не так сильно відриваються від них за рівнем оплати праці. Це може свідчити про зміну підходів до формування заробітної плати в компаніях та про більшу увагу до молодих фахівців.

Було проаналізовано розподіл респондентів за гендерною ознакою за період 2019-2024 років (рис. 2.8):

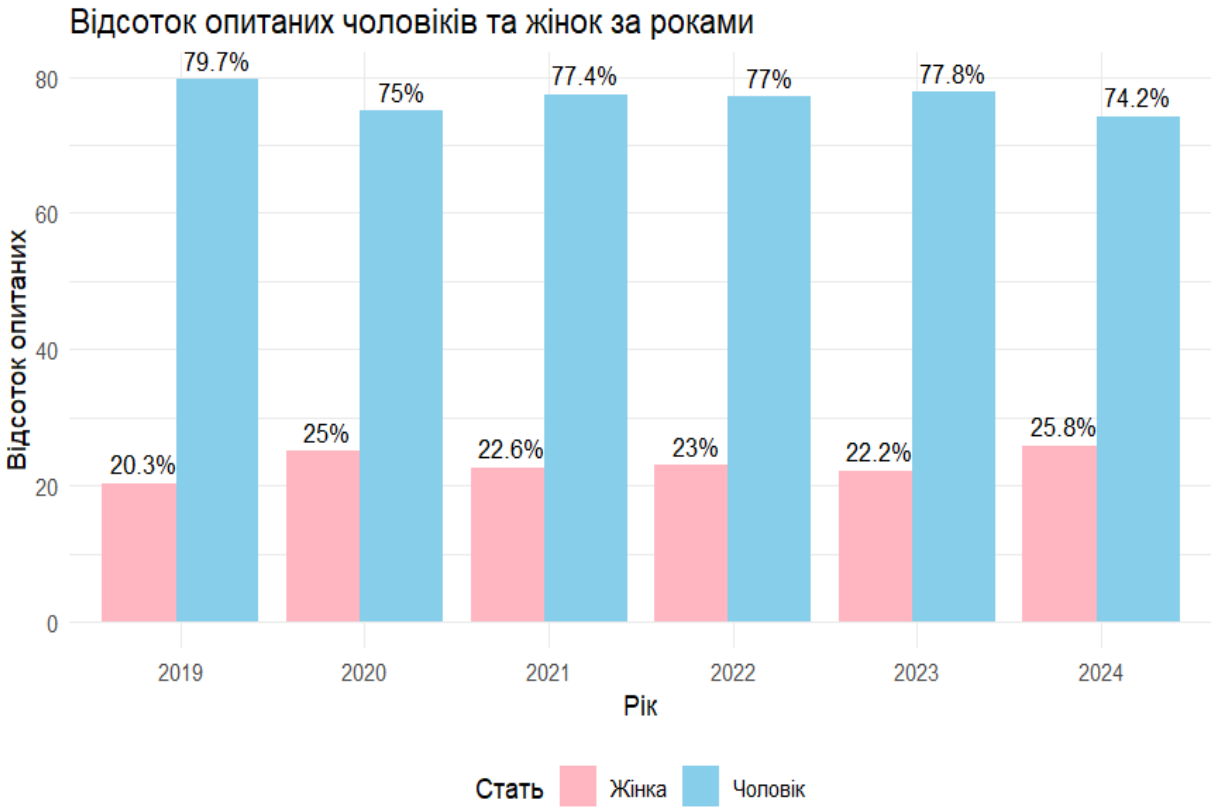

Рисунок 2.8 - Розподіл респондентів за гендерною ознакою за період з 2019 по 2024 роки

Джерело: розроблено автором на основі [50]

Аналіз розподілу респондентів за гендерною ознакою у дослідженні детермінант оплати праці за період 2019-2024 років виявляє значну диспропорцію у представленості чоловіків та жінок у вибірці. Спостерігається стабільно високий відсоток участі чоловіків (в середньому 77.2%) порівняно з суттєво нижчим показником залучення жінок (в середньому 22.8%), при цьому різниця у представленості становить близько 48 відсоткових пунктів протягом усього періоду дослідження, що може вказувати на потенційне зміщення у даних та необхідність застосування відповідних методів корекції при побудові предиктивних моделей для забезпечення репрезентативності результатів та мінімізації впливу гендерного дисбалансу на точність оцінювання факторів оплати праці.

За результатами аналізу гендерного фактору як однієї з ключових детермінант оплати праці виявлено стійку диференціацію в рівнях заробітних плат між чоловіками та жінками (рис. 2.9):

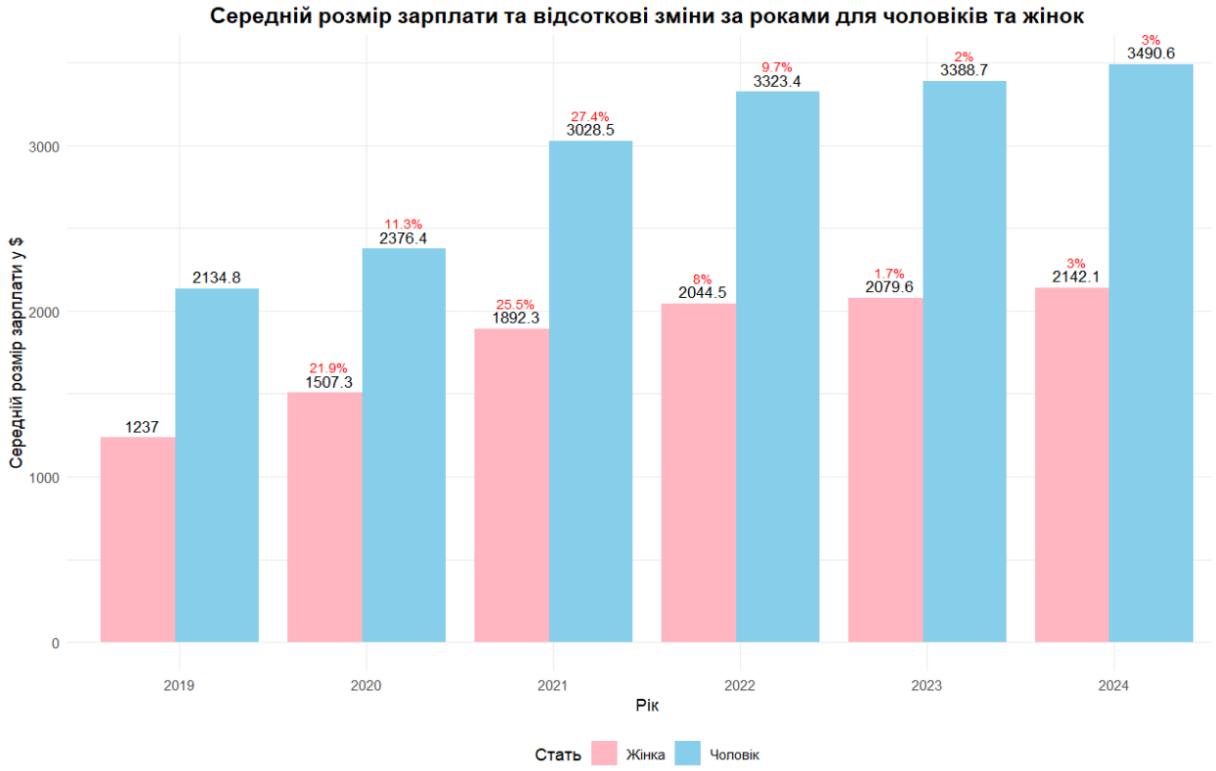

Рисунок 2.9 – Динаміка розміру середньої заробітної плати та відсоткові зміни для чоловіків та жінок за період з 2019 по 2024 роки

Джерело: розроблено автором на основі [50]

Дослідження часового ряду за 2019-2024 роки демонструє, що найменший відносний розрив спостерігався у 2020 році, коли заробітна плата жінок складала 63.4% від заробітної плати чоловіків. Станом на 2024 рік різниця складає 1348.5 одиниць, що є найбільшим абсолютним показником за весь період спостереження. Незважаючи на загальне зростання рівня оплати праці для обох статей (на 73.2% для жінок та 63.5% для чоловіків), гендерний розрив залишається суттєвим - заробітна плата жінок складає лише 61.4% від заробітної плати чоловіків станом на 2024 рік, що свідчить про збереження гендерної нерівності як вагомого чинника у формуванні рівня оплати праці та вказує на необхідність врахування цього фактору при розробці моделей прогнозування заробітних плат методами Data Science.

Було побудовано розподіл респондентів за містами (рис. 2.10):

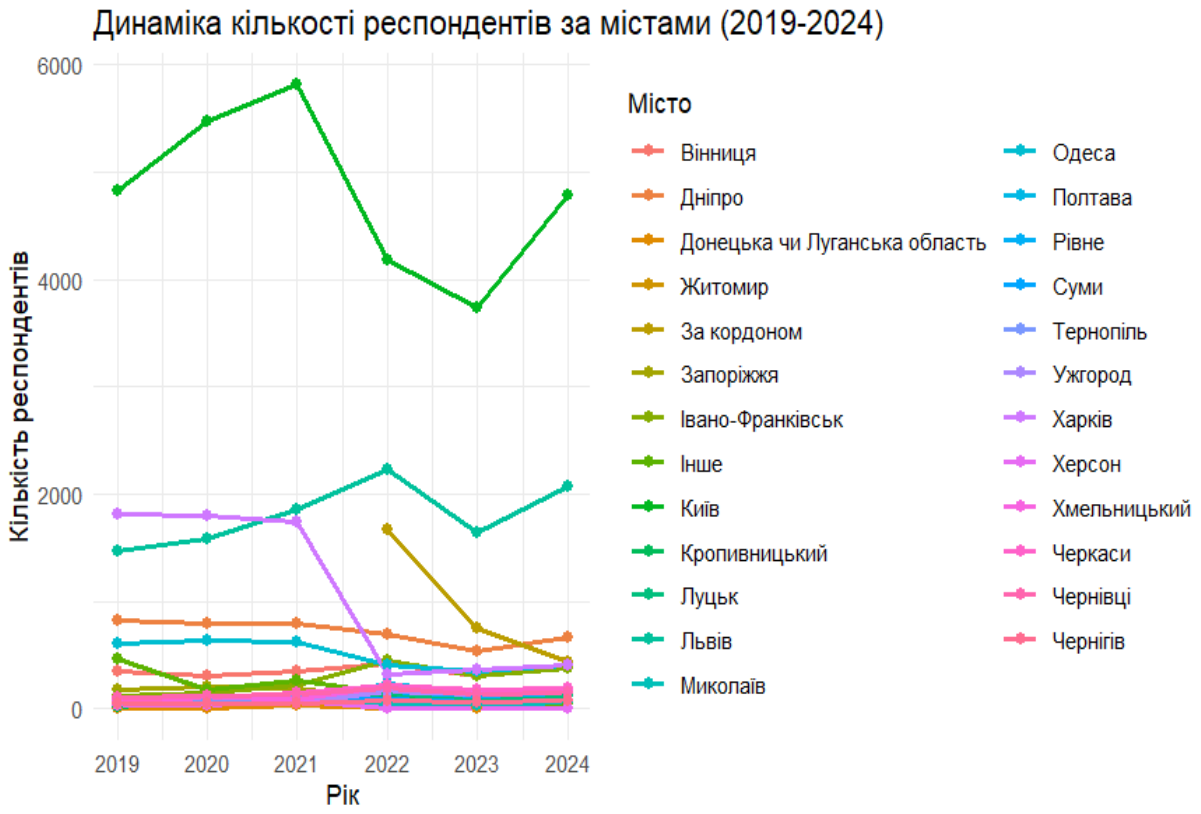

Рисунок 2.10 – Динаміка кількості респондентів за містами (2019-2024)

Джерело: розроблено автором на основі [50]

Бачимо, що стабільно найбільша кількість респондентів спостерігається у Києві, але з початком повномасштабного вторгнення в Україну 24 лютого 2022 року ця кількість різко впала, хоча вже ще залишається на найвищому рівні у порівнянні з іншими містами. Так само різко велика кількість респондентів переїхала зі східних регіонів, такі як Харків та Дніпро, у більш західніші регіони, такі як Львів, Івано-Франківськ, Ужгород та Луцьк, що видно з динаміки. Також з початком повномасштабного вторгнення з’явилась динаміка респондентів за кордоном, яка з часом йде на спад.

Досліджено залежність між рівнем оплати праці та містом. Результати представлені на рисунку 2.11:

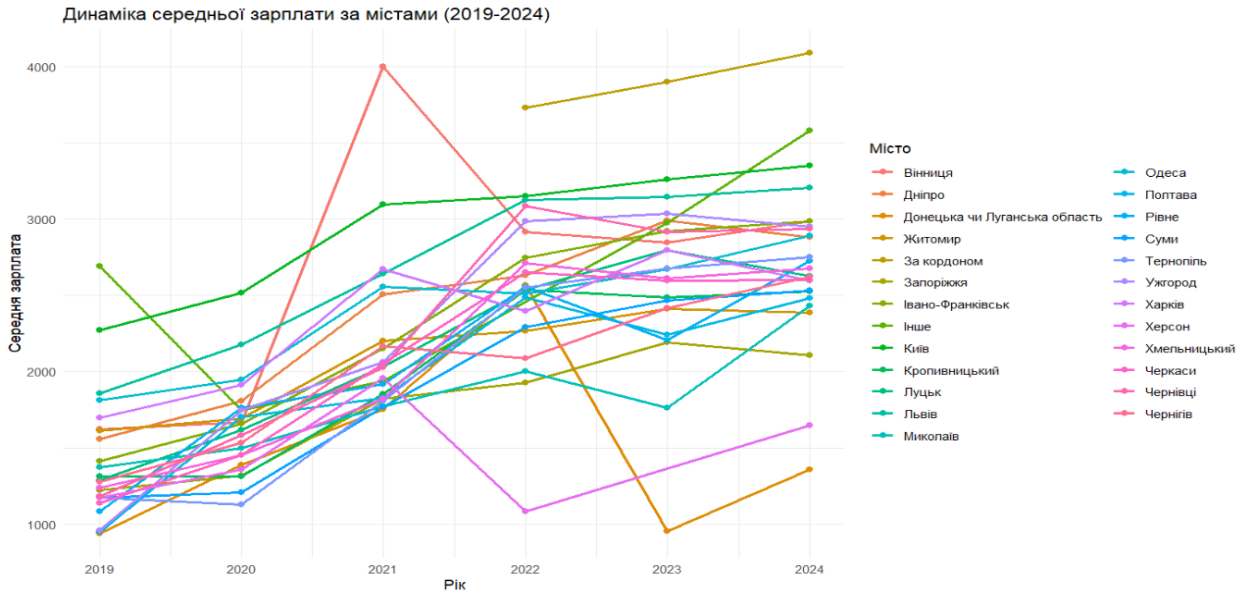

Рисунок 2.11 – Динаміка зміни середньої заробітної плати в різних містах України протягом 2019-2024 років

Джерело: розроблено автором на основі [50]

Можна помітити, що рівень заробітної плати суттєво відрізняється між різними містами, причому цей розрив з часом може як збільшуватися, так і зменшуватися.

Загалом, спостерігається тенденція до зростання середньої зарплати в більшості міст, що свідчить про загальне покращення економічної ситуації в країні. Однак, темпи зростання зарплат відрізняються в різних містах. Найбільш динамічне зростання заробітних плат спостерігається в великих містах, таких як Київ, Львів, Дніпро, що може бути пов'язано з концентрацією в них великих підприємств та IT-компаній. Водночас, в менших містах зростання зарплат відбувається більш повільно, що може бути пов'язано з меншим розвитком промисловості та послуг.

Побудовано розподіл респондентів за освітою (рис. 2.12):

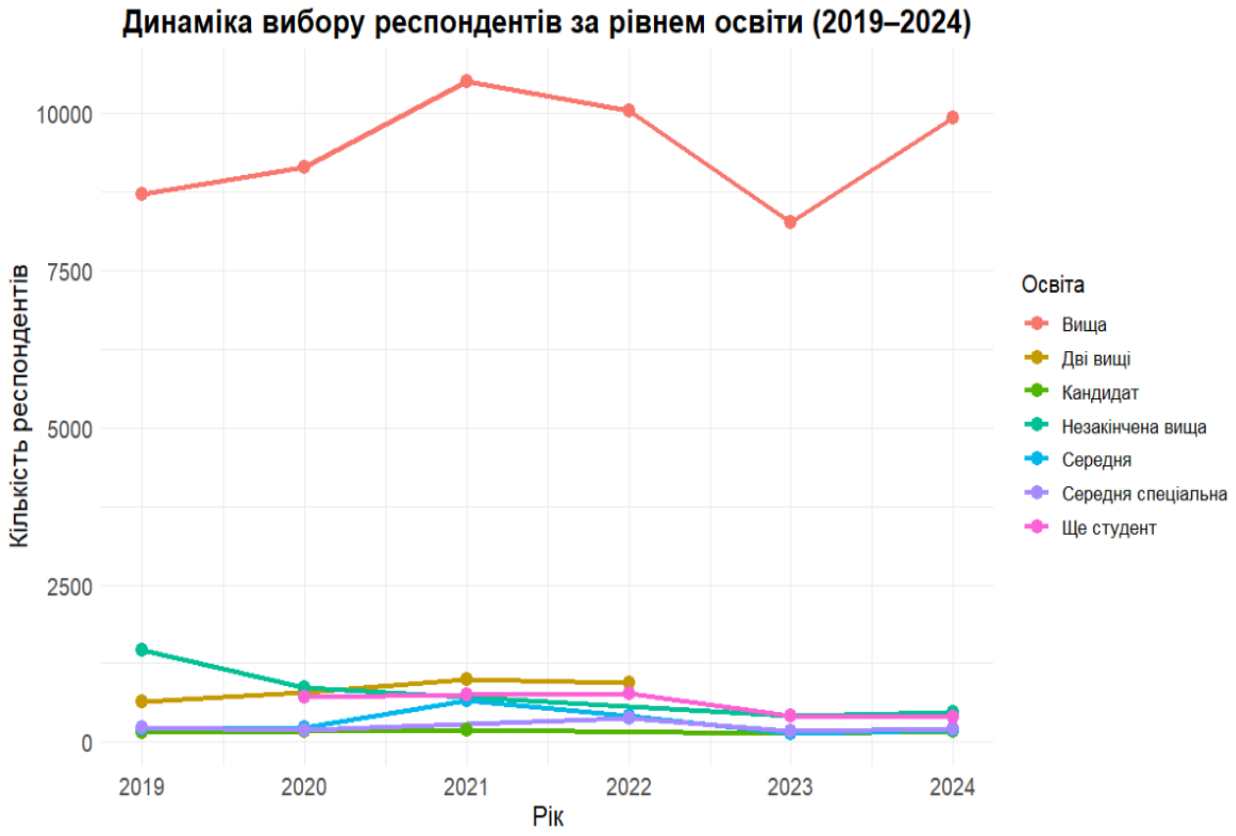

Рисунок 2.12 – Розподіл респондентів за рівнем освіти

Джерело: розроблено автором на основі [50]

Найбільшу кількість респондентів традиційно складають особи з вищою освітою, хоча спостерігається певне коливання їхньої кількості протягом років. Особливо помітне зростання кількості респондентів з вищою освітою у 2019 та 2021 роках, після чого спостерігається деяке зниження. Це може свідчити про зміни в загальній освітній структурі суспільства або про зміни в аудиторії, яка бере участь в дослідженнях.

Щодо інших рівнів освіти, то їхні частки в загальній кількості респондентів є значно меншими. Спостерігається загальна тенденція до зменшення кількості респондентів з незакінченою вищою, середньою та середньою спеціальною освітою, що може бути пов'язано з процесами урбанізації та зростанням доступності вищої освіти. Водночас, частка студентів залишається відносно стабільною протягом усього періоду.

Діаграма на рисунку 2.13 наочно демонструє, як змінювалася середня заробітна плата в залежності від рівня освіти протягом 2019-2024 років:

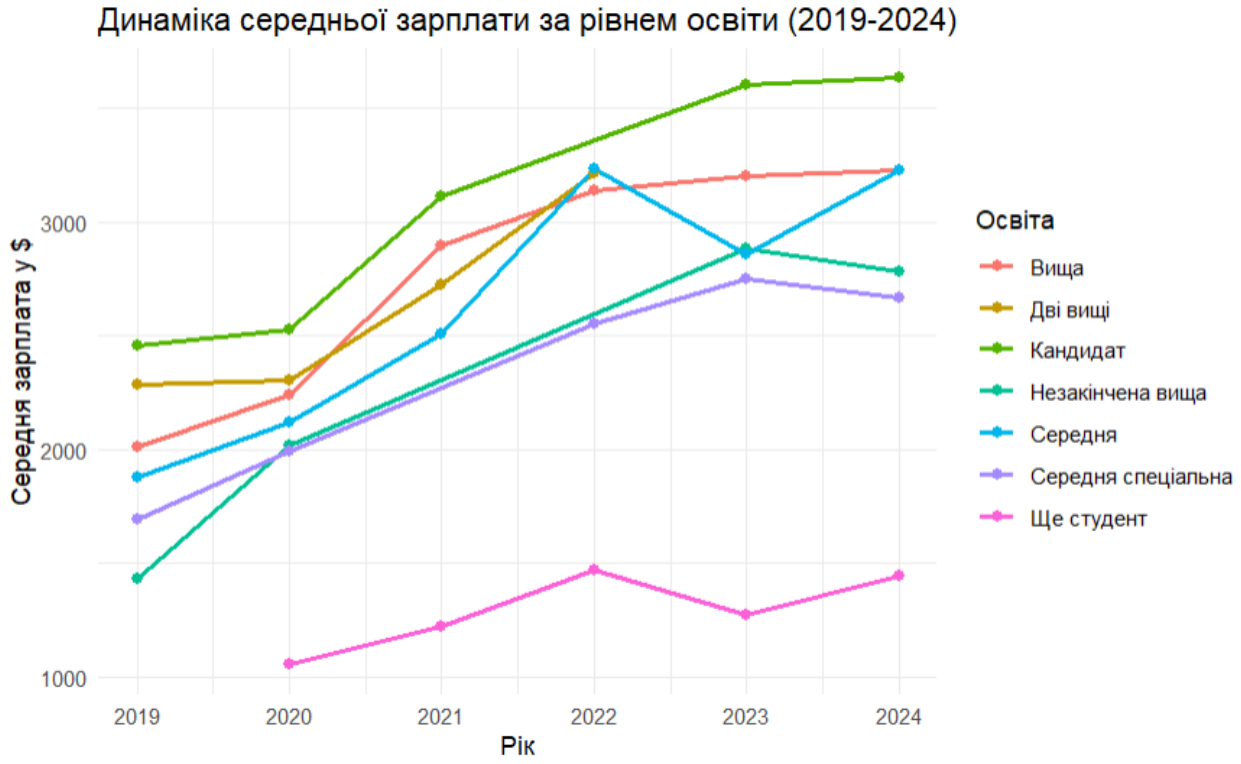

Рисунок 2.13 – Динаміка середньої зарплати за рівнем освіти (2019-2024)

Джерело: розроблено автором на основі [50]

Перше, що впадає в око, це загальна тенденція до зростання середньої зарплати для всіх рівнів освіти. Це свідчить про загальне покращення економічної ситуації та підвищення оплати праці в країні. Однак, якщо порівнювати різні рівні освіти, то можна помітити, що вища освіта традиційно пов'язана з більш високою заробітною платою. При цьому, найбільш помітне зростання зарплат спостерігається у осіб з вищою освітою та кандидатським ступенем. Це підтверджує тезу про те, що інвестиції в освіту окупаються у вигляді більш високої заробітної плати в майбутньому.

Було створено діаграму розподілення респондентів за вказаною посадою (рис. 2.14):

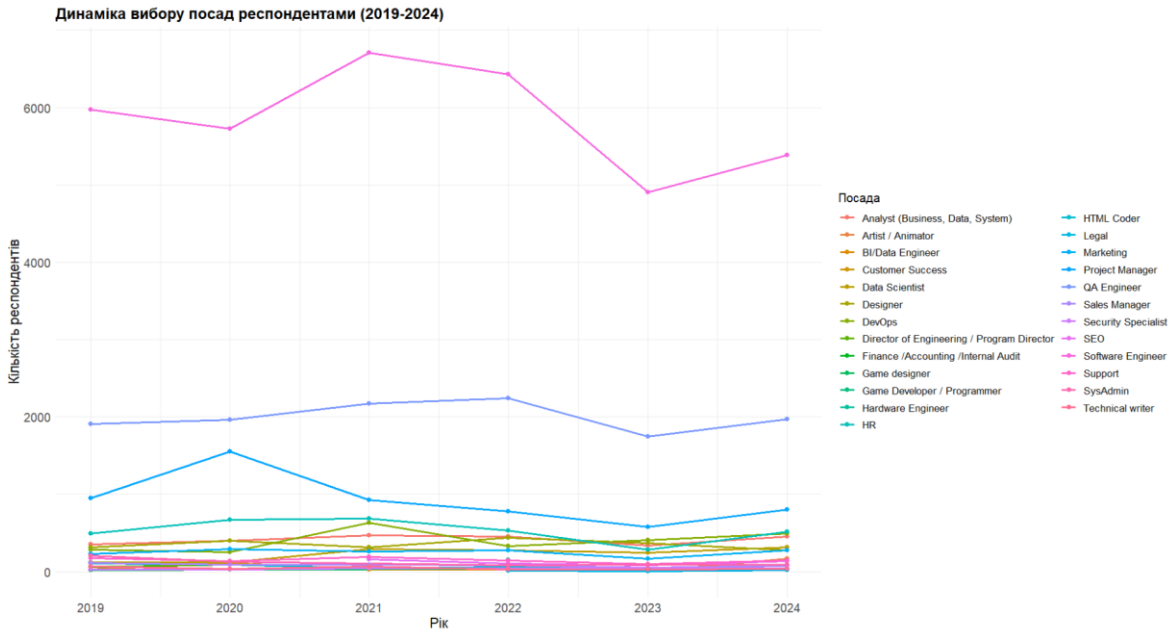

Рисунок 2.14 – Розподіл респондентів за посадою

Джерело: розроблено автором на основі [50]

На основі діаграми динаміки вибору посад респондентами за період з 2019 по 2024 роки можна зробити кілька ключових висновків. Найбільш популярною посадою протягом зазначеного періоду є Software Engineer, що значно випереджає всі інші ролі за кількістю опитуваних. Її популярність зростала з 2019 року, досягаючи піку у 2021 році, після чого почала знижуватися, але все ще залишається на першому місці. Це свідчить про стабільний попит на цю посаду в IT-галузі.

Інші посади, такі як QA Engineer, HR, і Project Manager, демонструють меншу, але все ж значну динаміку зростання та коливань у популярності. Посади, представлені нижче, як-от Technical Writer, BI/Data Engineer і Game Developer, мають значно менше опитуваних, що свідчить про їхню специфічність або менший попит на ринку праці. Загалом, діаграма ілюструє зростаючий інтерес до технічних і управлінських посад, при цьому помітно, що посади з вузькою спеціалізацією мають обмежену популярність.

Побудовано діаграму розподілення респондентів за посадою, за винятком Software Engineer, QA Engineer та Project Manager, аби було зручніше аналізувати інші посади (рис. 2.15):

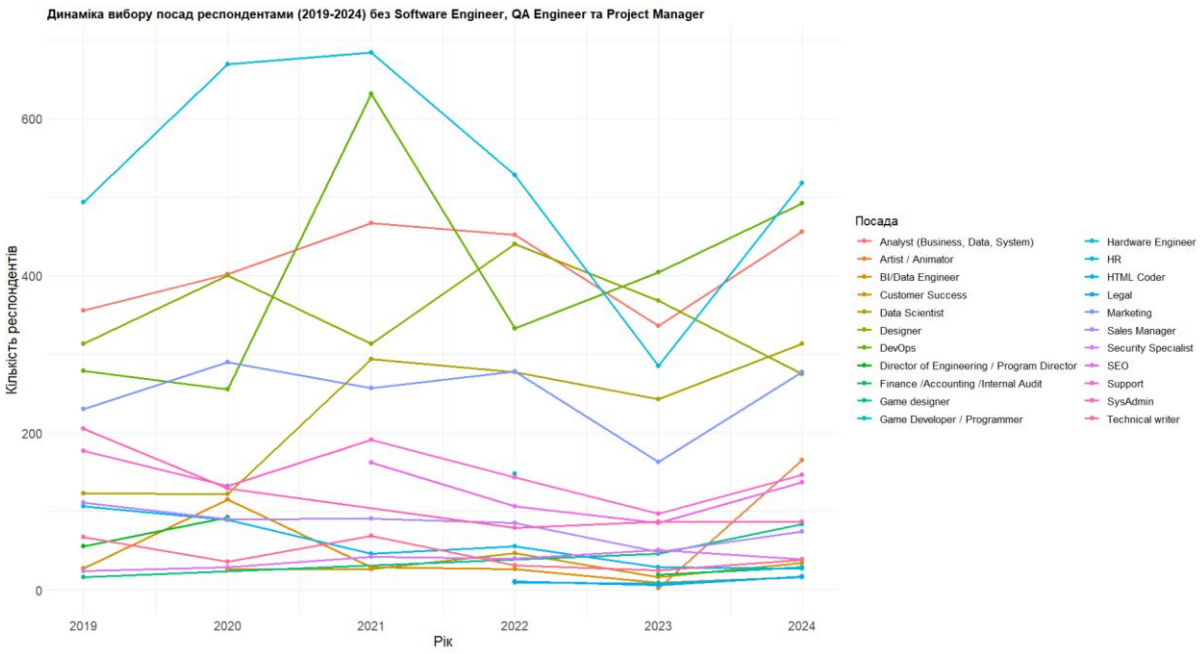

Рисунок 2.15 – Розподіл респондентів за посадою, окрім Software Engineer, QA Engineer та Project Manager

Джерело: розроблено автором на основі [50]

В період з 2019 по 2024 роки різні категорії посад демонструють неоднакові тенденції. Деякі посади, такі як HR та Data Scientist, показують стабільний або навіть зростаючий інтерес, що свідчить про попит на фахівців цих напрямків. З іншого боку, позиції на кшталт Legal і Technical Writer залишаються відносно малопопулярними, без значних коливань.

Також можна помітити, що до 2022 року інтерес до деяких посад, наприклад, DevOps і Customer Success, суттєво зростав, але з 2023 року спостерігається певне зниження. Важливо зазначити, що в 2024 році для окремих категорій, таких як Marketing, помітний значний стрибок, можливо, через зміни ринкових потреб. В цілому, діаграма ілюструє важливі зміни в ринкових трендах і перевагах респондентів у виборі професійних ролей.

Діаграма на рисунку 2.16 наочно демонструє зміну середньої заробітної плати за різними посадами протягом 2019-2024 років:

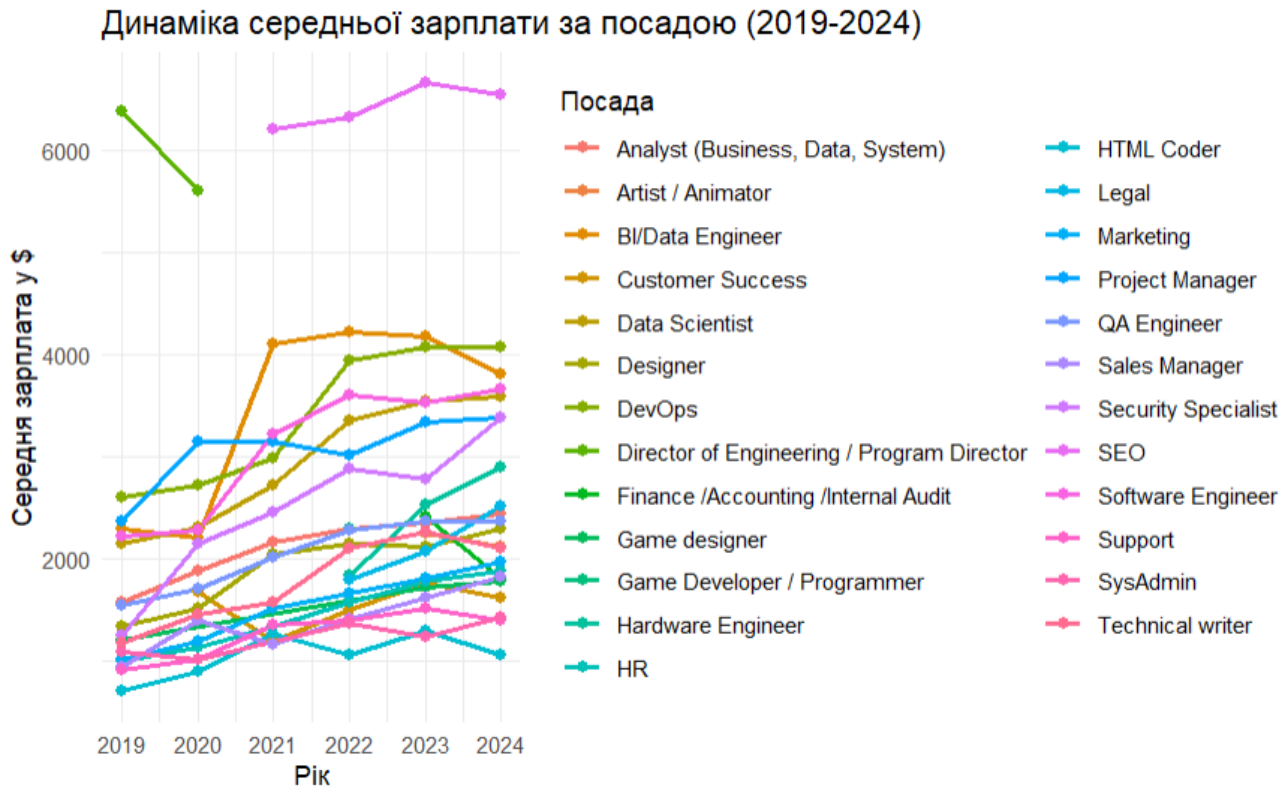

Рисунок 2.16 – Динаміка середньої зарплати за посадами (2019-2024)

Джерело: розроблено автором на основі [50]

Можна помітити, що загальна тенденція для більшості професій – це зростання середньої зарплати. Це свідчить про загальне покращення економічної ситуації та підвищення оплати праці на ринку праці. Посади, пов'язані з програмуванням, розробкою, аналізом даних демонструють стійкий тренд на підвищення оплати праці. Це пояснюється високим попитом на кваліфікованих IT-спеціалістів та постійним розвитком технологій. По-друге, помітно, що зарплати у сфері управління та бізнесу (наприклад, Project Manager, Director of Engineering) також демонструють стійкий ріст. Це свідчить про зростання ролі менеджменту в сучасному бізнесі та підвищення вимог до кваліфікації керівників.

Побудовано динаміку кількості респондентів за різним стажем роботи за період із 2019 до 2024 року (рис. 2.17):

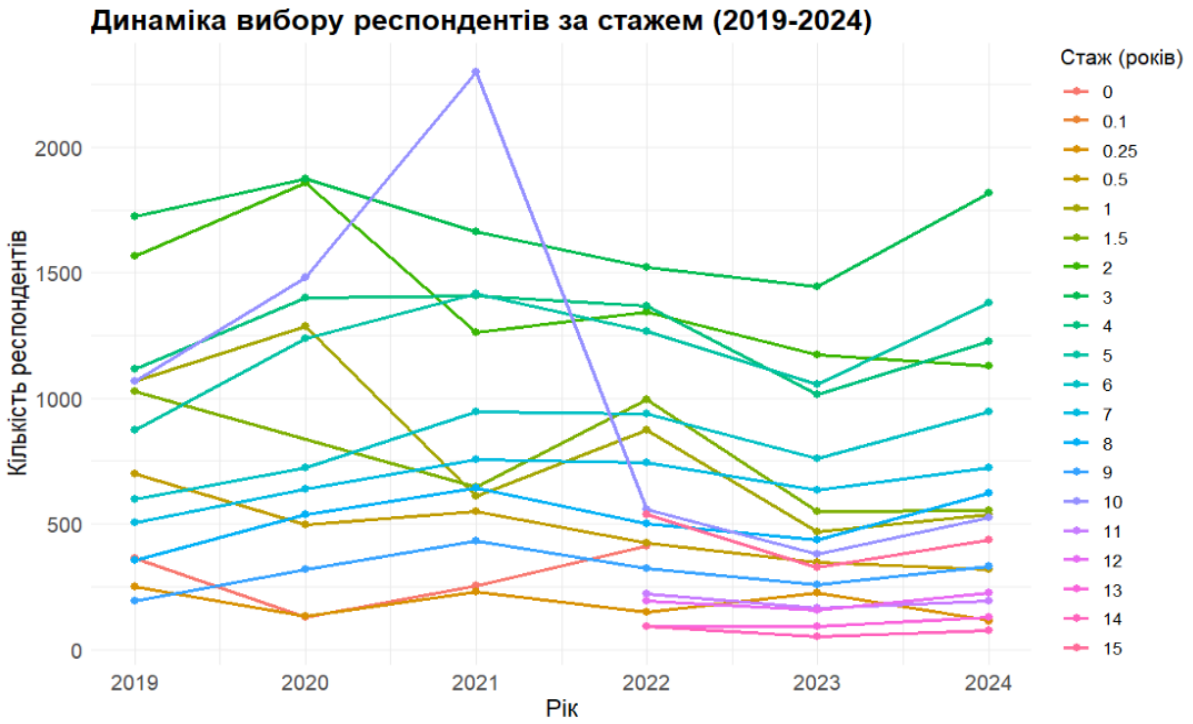

Рисунок 2.17 – Розподіл респондентів за стажем

Джерело: розроблено автором на основі [50]

Видно, що найбільшу кількість респондентів складають особи зі стажем 1–3 роки, особливо зі стажем 2 роки, який демонструє значну стабільність, а в 2024 році навіть зростає. Для коротших проміжків стажу, таких як 0 років (менше 1 місяця), 0.1 року (1 місяць), 0.25 року (3 місяці) та 0.5 року (півроку), спостерігається стабільно низька кількість респондентів із деякими коливаннями, але вони залишаються менш популярними у виборі.

Водночас стаж від 4 до 6 років показує помірну стабільність, поступово зростаючи в останні роки. Респонденти зі стажем понад 10 років (особливо 10-15 років) є найменш чисельними, хоча і демонструють деяке зростання в 2024 році. Важливо відзначити, що такі категорії, як 1.5 року (півтора роки), також зберігають стабільні, але низькі показники протягом усього періоду. Це свідчить про перевагу серед респондентів із середнім рівнем досвіду, у той час як новачки і досвідчені спеціалісти є менш репрезентативними.

Діаграма на рисунку 2.18 демонструє цікаву тенденцію щодо зміни середньої заробітної плати в залежності від стажу роботи:

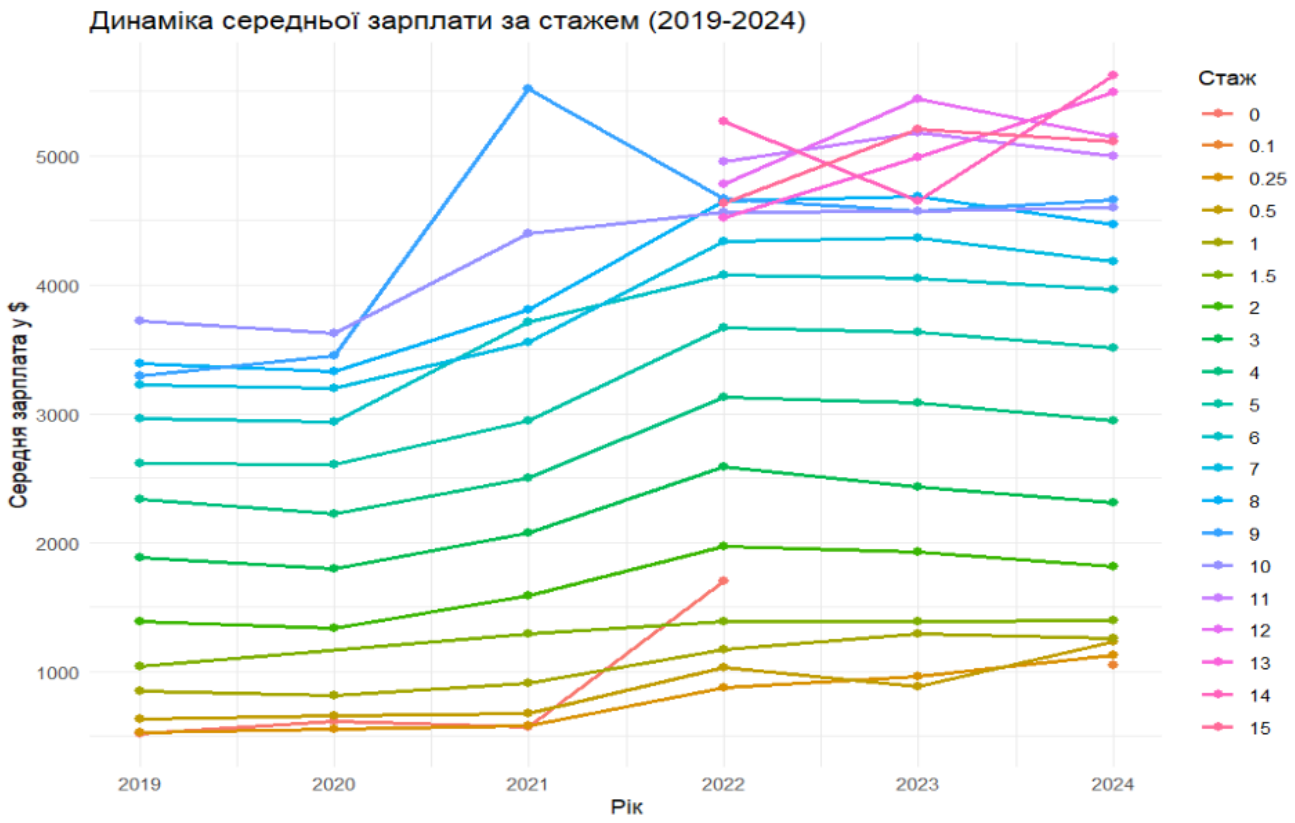

Рисунок 2.18 – Динаміка середньої зарплати за стажем роботи (2019-2024)

Джерело: розроблено автором на основі [50]

Загалом, можна помітити, що з набуттям досвіду рівень заробітної плати зростає. Це цілком логічно, оскільки з досвідом та набуттям нових навичок працівники стають більш цінними для компаній і, відповідно, отримують вищу оплату. Спостерігається, що найбільш різке зростання зарплати відбувається протягом перших років роботи. Далі темпи зростання сповільнюються, а для деяких категорій працівників з великим стажем може спостерігатися навіть незначне зниження зарплати. Це може бути пов'язано з різними факторами, такими як досягнення певного рівня в кар'єрі, зміна пріоритетів працівників або ж циклічні зміни на ринку праці.

Оскільки опитувальник був взятий зі сфери IT, у датасеті є стовпчик під назвою «Ваш тайтл», який також необхідно дослідити. "Тайтл" (від англ. "title") — це посада або звання, яке вказує на рівень кваліфікації та роль працівника в компанії, наприклад, "Junior Developer", "Middle Developer", "Senior Developer", "Lead Engineer", "Software Architect" тощо. У IT-індустрії тайтл часто є важливим

чинником, що впливає на зарплату, оскільки він вказує на досвід, знання та відповідальність спеціаліста.

Тож на основі цього стовпчика було побудовано динаміку кількості респондентів за рівнями їхнього тайтлу (рис. 2.19):

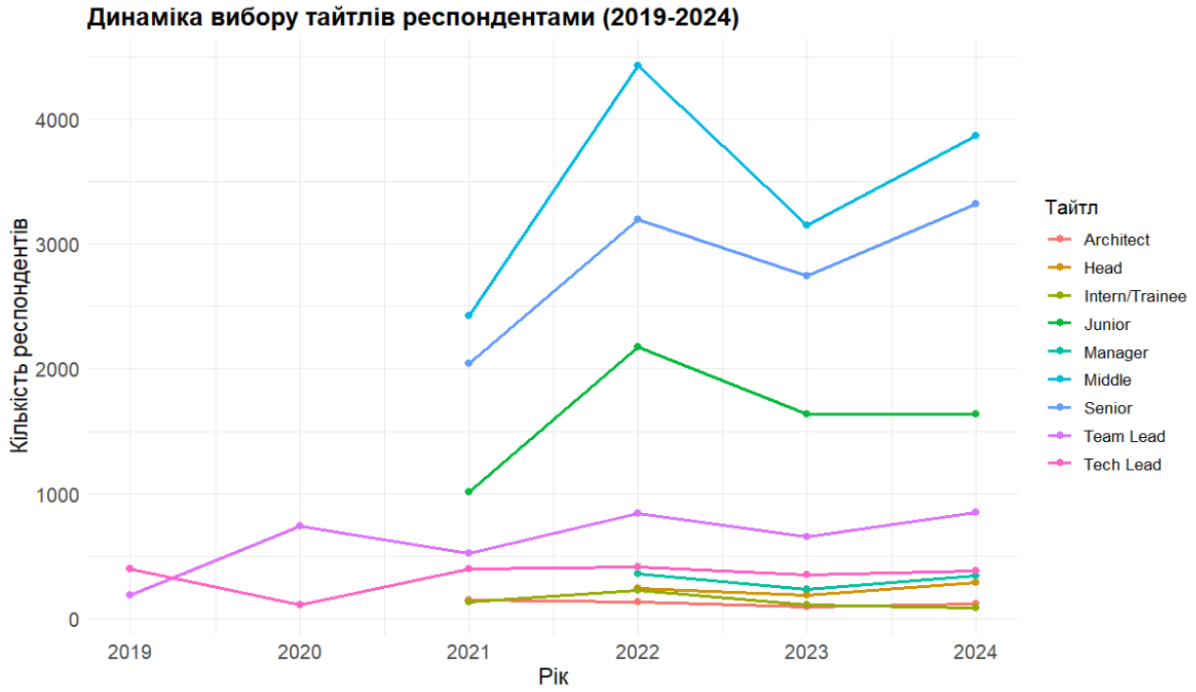

Рисунок 2.19 – Розподіл респондентів за тайтлом

Джерело: розроблено автором на основі [50]

Найбільш помітним є стрімке зростання популярності ролі "Middle" (середній рівень). Це свідчить про збільшення кількості фахівців з середнім рівнем досвіду та відповідальності, що може бути пов'язано з загальним зростанням ринку праці та розширенням компаній.

Водночас, спостерігається коливання популярності інших ролей. Так, ролі "Junior" (молодший рівень) та "Intern/Trainee" (стажер) демонструють певний спад після 2021 року, що може свідчити про зменшення кількості нових випускників або зміни в політиці найму компаній. Ролі "Senior" (старший рівень), "Team Lead" (керівник команди) та "Tech Lead" (технічний лідер) також демонструють певні коливання, що може бути пов'язано зі змінами в структурі компаній та вимогами до керівників.

Діаграма представлена на рисунку 2.20 наочно демонструє зміну середньої заробітної плати за різними тайтлами протягом 2019-2024 років:

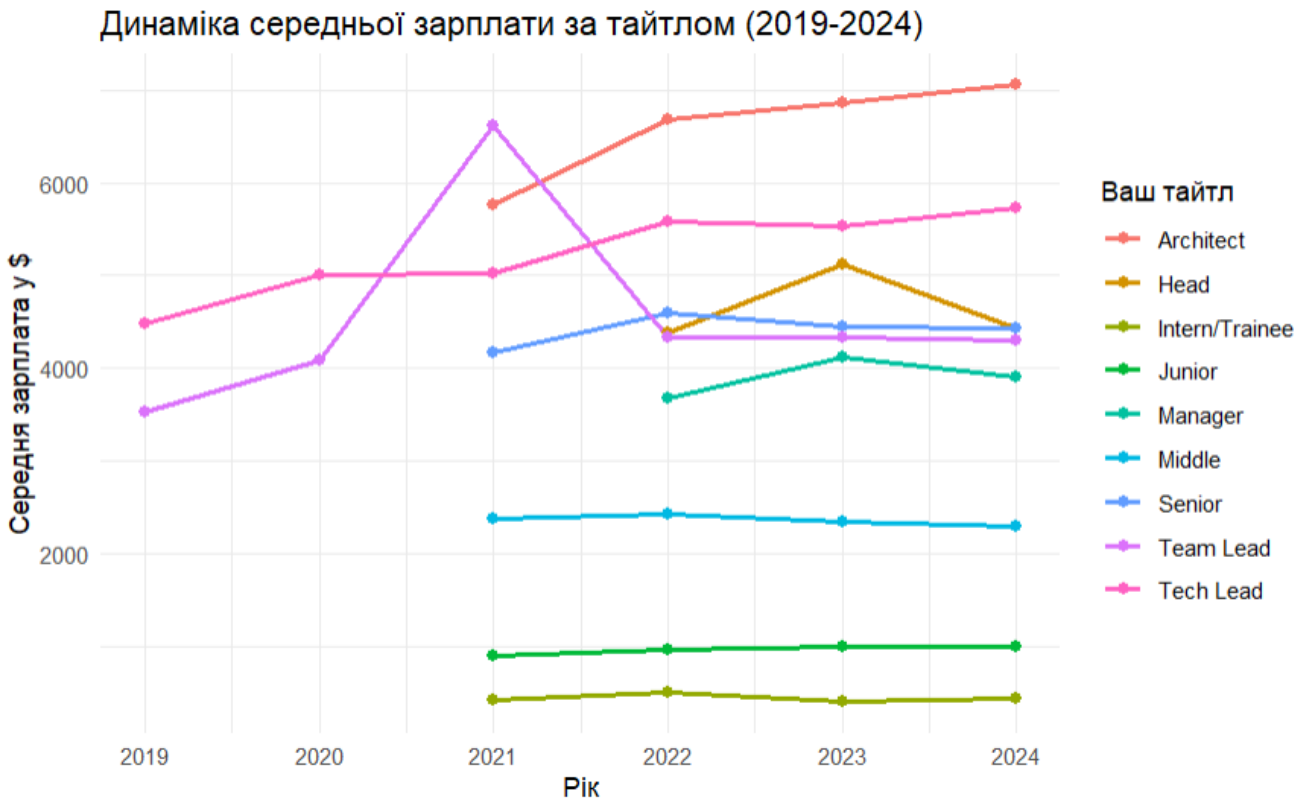

Рисунок 2.20 – Динаміка середньої зарплати за тайтлом (2019-2024)

Джерело: розроблено автором на основі [50]

Проаналізувавши дані, можна виділити кілька цікавих моментів. По-перше, найбільш динамічне зростання зарплат спостерігається у позицій з більш високою відповідальністю та вищими вимогами до кваліфікації, таких як "Architect", "Head", "Tech Lead". Це пояснюється високим попитом на кваліфікованих фахівців та постійним розвитком технологій. По-друге, помітно, що зарплати у середньому та старшому рівнях (Middle, Senior) також демонструють стійкий ріст. Це свідчить про зростання ролі досвіду та експертизи на ринку праці.

Стовпчики «Основна спеціалізація» та «Основна мова програмування» візуально оцінювалась іншим чином, оскільки по результатам функції n_distinct(), за допомогою якої можна дізнатися кількість унікальних елементів у певному

стовпці датасету, у стовпчику «Основна спеціалізація» маємо 173 варіантів спеціалізацій, а у стовпчику «Основна мова програмування» - 81.

Тому було візуалізовано лише найпопулярніші спеціалізації за кожен з років (рис. 2.21):

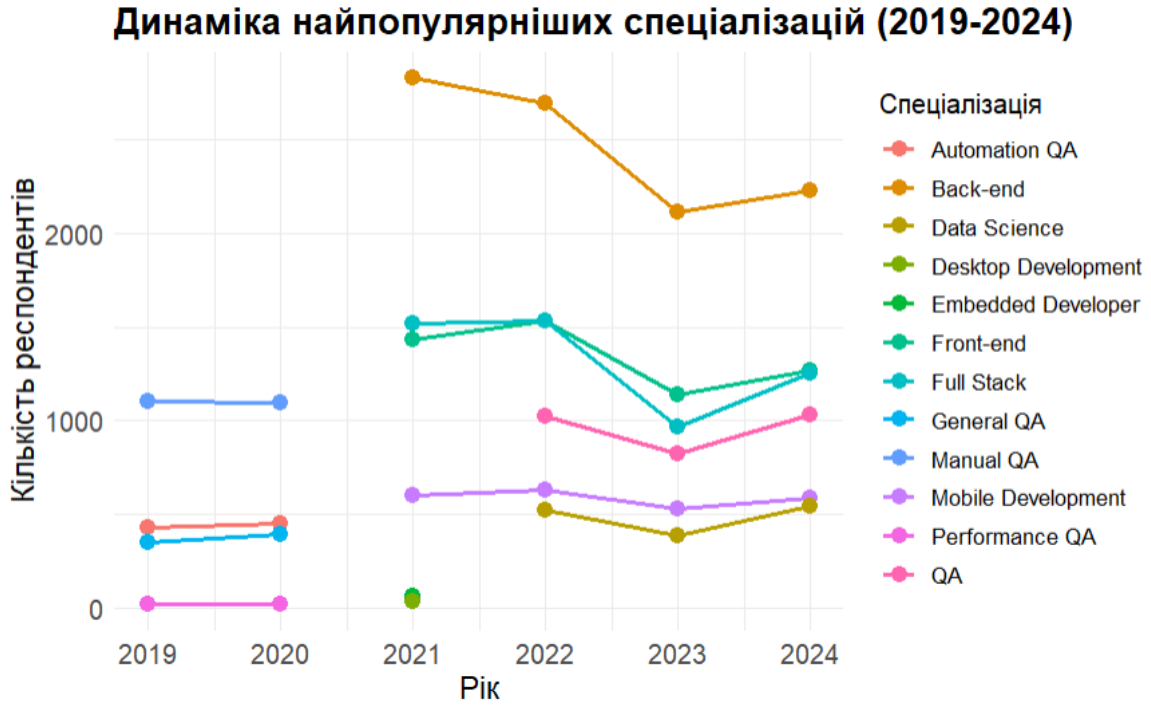

Рисунок 2.21 – Топ найпопулярніших основних спеціалізацій вказаних респондентами

Джерело: розроблено автором на основі [50]

Протягом періоду з 2019 по 2020 роки у спеціалізаціях, пов’язаних з тестуванням програмного забезпечення (Automation QA, General QA, Manual QA, Performance QA) спостерігається відносна стабільність у топ-5 спеціалізацій. Однак, починаючи з 2021 року, відбувається значна зміна в структурі попиту. Спеціалізація Back-end демонструє стрімке зростання і виходить на лідируючі позиції, тоді як деякі інші спеціалізації, які були популярними раніше, починають втрачати свої позиції. З 2022 року по 2023 усі спеціалізації демонструють певний спад, що також певним чином може бути пов’язано із початком повномасштабного вторгнення на території України.

Діаграма на рисунку 2.22 демонструє тенденцію щодо зміни середньої заробітної плати за різними IT-спеціалізаціями протягом 2019-2024 років:

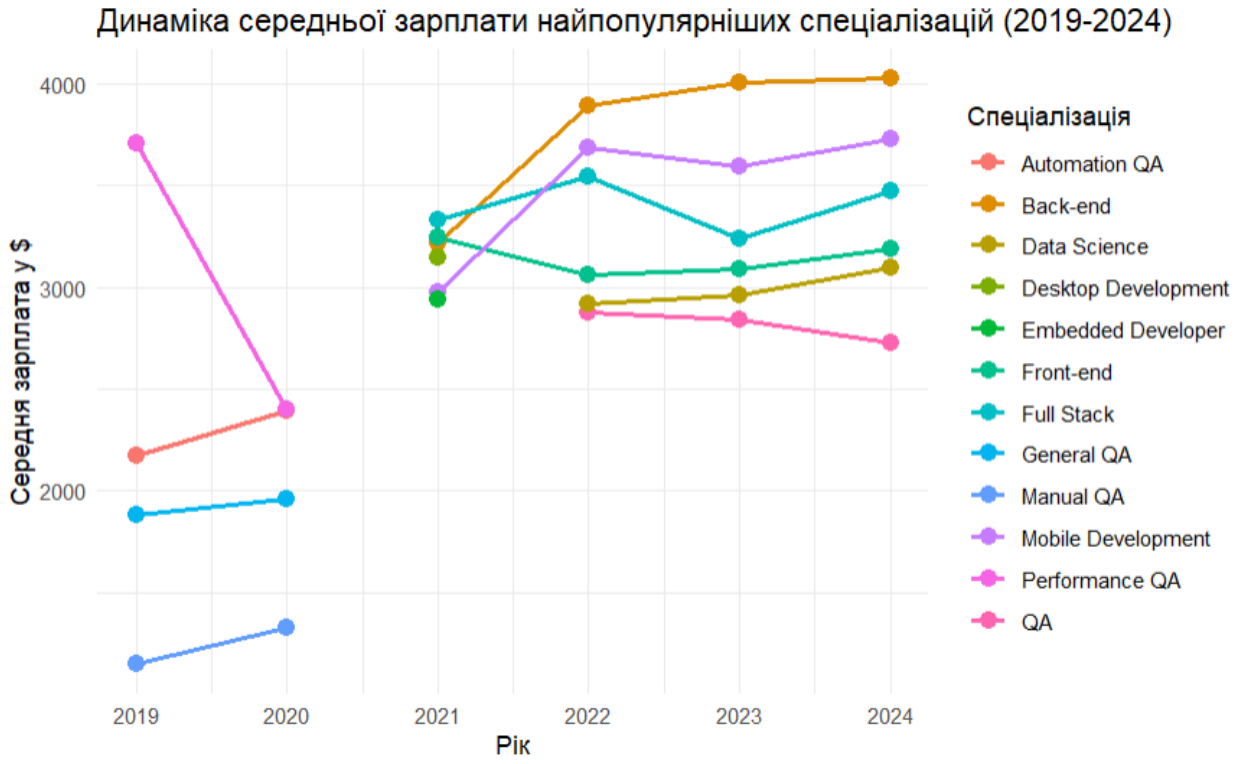

Рисунок 2.22 – Динаміка середньої зарплати найпопулярніших спеціалізацій (2019-2024)

Джерело: розроблено автором на основі [50]

Загалом, можна помітити, що більшість спеціальностей демонструють стійке зростання заробітної плати. Це свідчить про загальне підвищення попиту на IT-фахівців та їхню дедалі більшу цінність на ринку праці. Найбільш динамічне зростання зарплат спостерігається у сферах, пов'язаних з аналізом даних (Data Science), розробкою програмного забезпечення (Back-end, Front-end, Full Stack) та автоматизацією тестування (Automation QA). Це пов'язано з активним розвитком технологій, цифровою трансформацією бізнесу та зростаючою потребою в спеціалістах, які можуть працювати з великими обсягами даних, розробляти складні програмні продукти та забезпечувати їхню якість.

Побудовано і розподіл за найпопулярнішими основними мовами програмування, які обирали опитані респонденти (рис. 2.23):

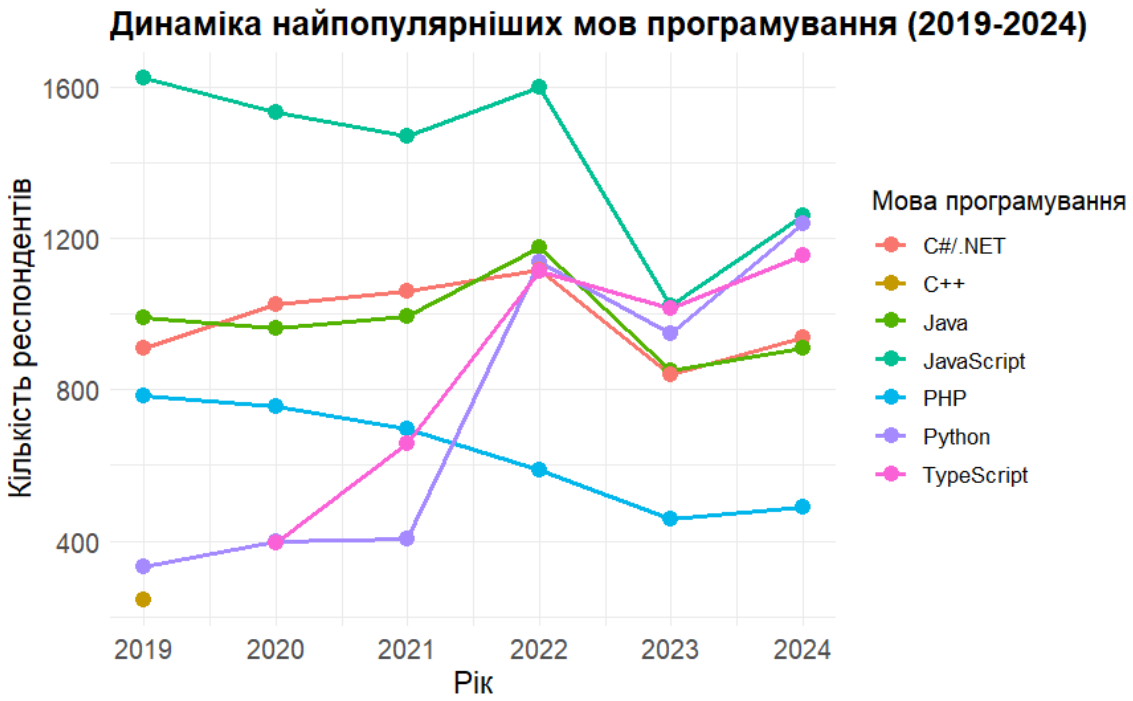

Рисунок 2.23 – Топ найпопулярніших основних мов програмувань вказаних респондентами

Джерело: розроблено автором на основі [50]

Мова програмування JavaScript протягом усього досліджуваного періоду займає лідируючу позицію у найпопулярнішій мові програмування серед усіх опитаних, незважаючи на різкий та великий спад у 2023 році. Найбільш помітним є стрімке зростання популярності мови Python у 2022 році, що свідчить про зростаючу роль аналізу даних, машинного навчання та інших сфер, де Python є домінуючою мовою. Варто також зазначити, що TypeScript стає дедалі популярнішою мовою, зростаючи з 2020 до 2024 року, стаючи однією з найпопулярніших мов програмування серед представлених увійшовши у топ 3.

Діаграма на рисунку 2.24 наочно демонструє зміну середньої заробітної плати фахівців, які працюють з різними мовами програмування, протягом 2019-2024 років:

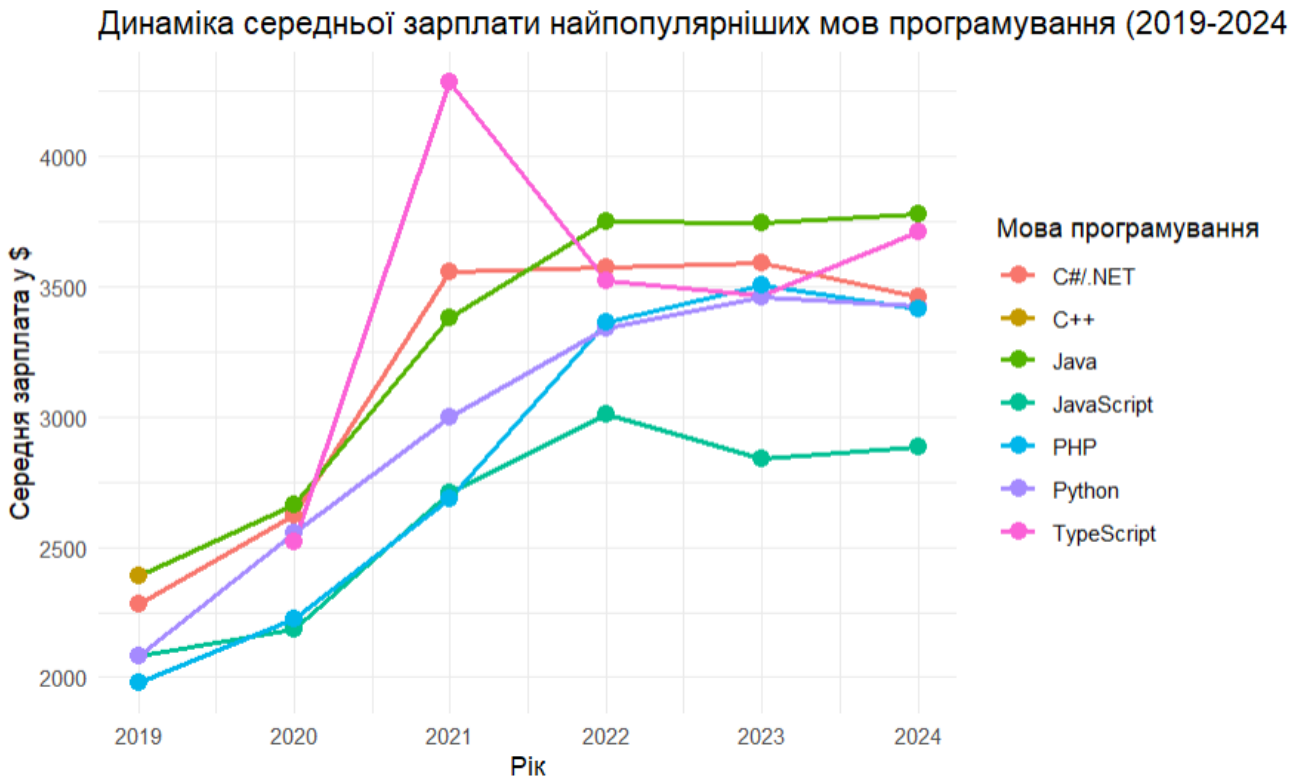

Рисунок 2.24 – Динаміка середньої зарплати найпопулярніших мов програмування (2019-2024)

Джерело: розроблено автором на основі [50]

Можна помітити, що загальна тенденція для більшості мов – це зростання середньої зарплати. Це свідчить про загальне підвищення попиту на IT-фахівців та їхню дедалі більшу цінність на ринку праці.

Найбільш динамічне зростання зарплат спостерігається у фахівців, які працюють з мовами TypeScript. Це пов'язано з активним розвитком веб-розробки, машинним навчанням та даними наук, де ці мови знаходять широке застосування. Водночас, деякі мови, такі як Java, демонструють більш стабільний ріст, що свідчить про їхню зрілість та широке використання в корпоративному секторі.

Було також побудовано розподіл опитаних респондентів за рівнем їхньої англійської мови (рис. 2.25):

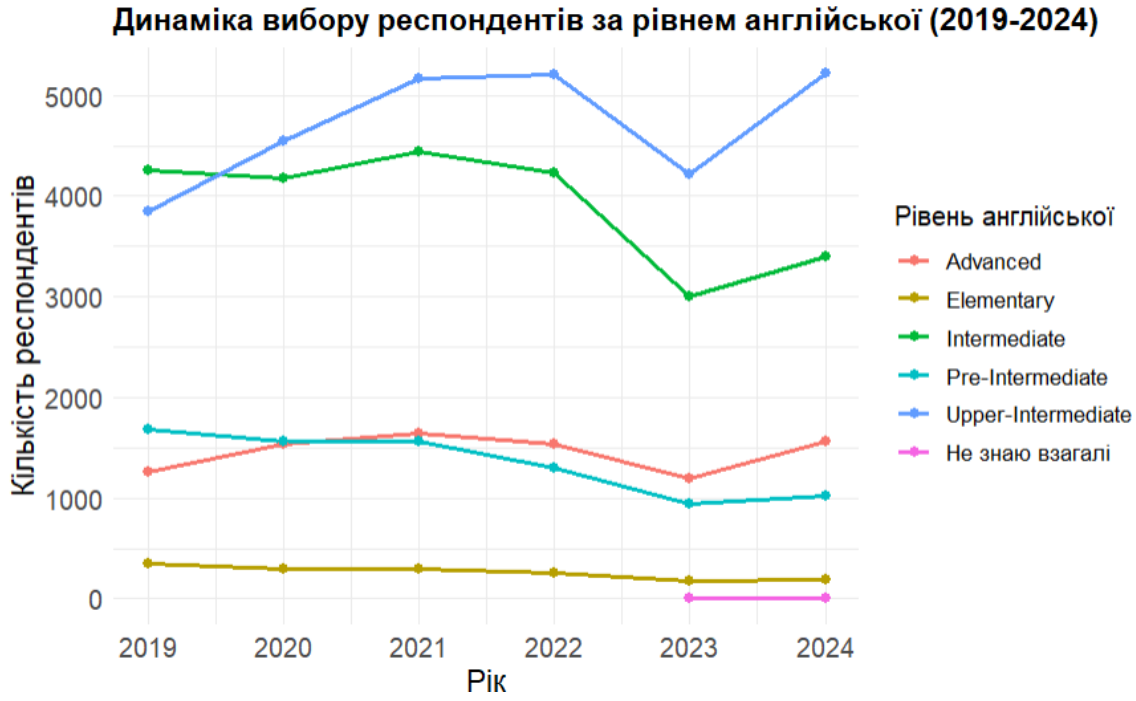

Рисунок 2.25 – Розподіл респондентів за рівнем англійської

Джерело: розроблено автором на основі [50]

Найбільш помітним є зростання кількості респондентів з рівнем Upper-Intermediate (вищий середній) та Advanced (просунутий). Це свідчить про підвищення загального рівня володіння англійською мовою серед респондентів, що може бути пов'язано з зростаючою роллю англійської мови як міжнародної мови спілкування та необхідністю для професійного зростання. Кількість респондентів з рівнем Pre-Intermediate (нищий середній) та Intermediate (середній) демонструє певний спад після 2021 року, що може бути пов'язано з тим, що все більше людей прагне досягти більш високого рівня.

Діаграма представлена на рисунку 2.26 демонструє тенденцію щодо зміни середньої заробітної плати в залежності від рівня володіння англійською мовою протягом 2019-2024 років:

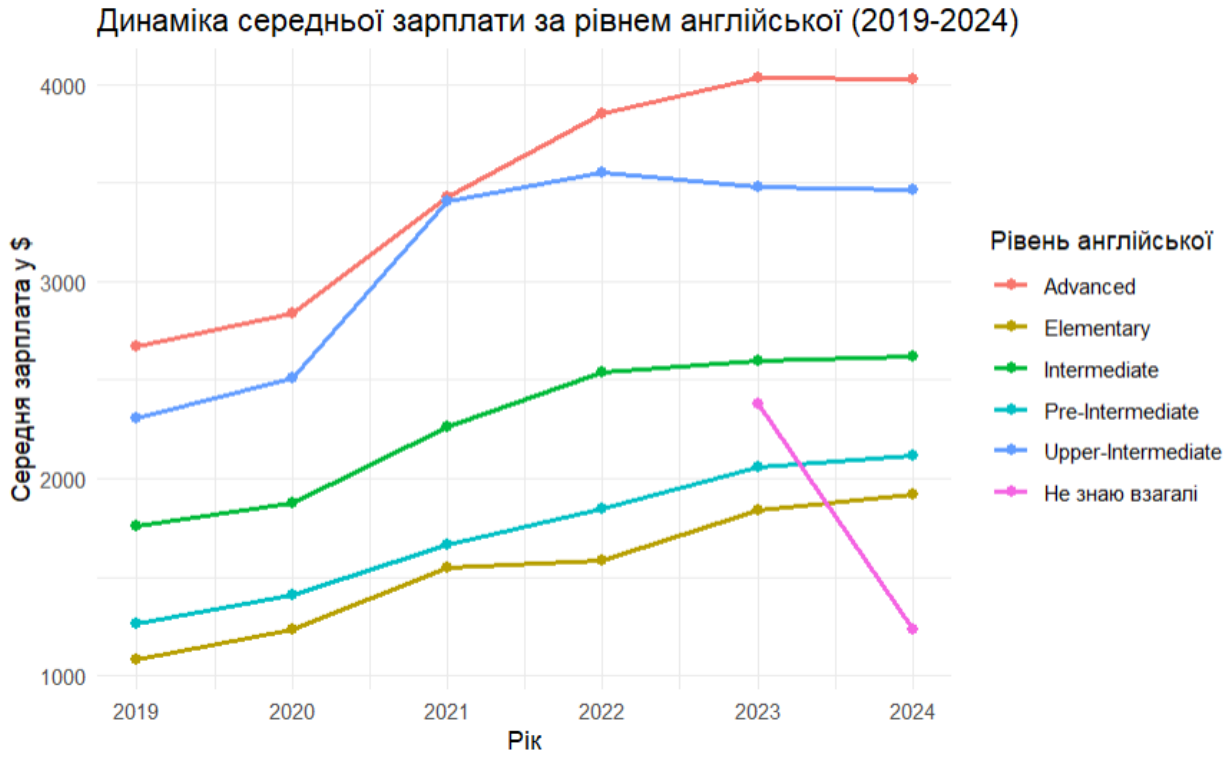

Рисунок 2.26 – Динаміка середньої зарплати за рівнем англійської (2019-2024)

Джерело: розроблено автором на основі [50]

Загалом, можна помітити, що з покращенням рівня англійської мови, як правило, зростає і рівень заробітної плати. Це цілком логічно, оскільки володіння англійською мовою є важливою вимогою для багатьох професій, особливо в IT-сфері, де більшість технічної документації та комунікації відбувається саме англійською мовою.

Бачимо, що найбільш різке зростання зарплати спостерігається у 2021 році для рівня англійської Upper-Intermediate. Далі темпи зростання сповільнюються, а для рівня Upper-Intermediate та Advanced навіть спостерігається незначне зниження у 2024 році. Це може бути пов'язано з тим, що для багатьох позицій достатньо середнього рівня володіння мовою, а більш високий рівень не є критичним фактором для підвищення заробітної плати.

Побудовано діаграму, яка демонструє динаміку змін у виборі компаній різного розміру респондентами протягом 2019-2024 років (рис. 2.27):

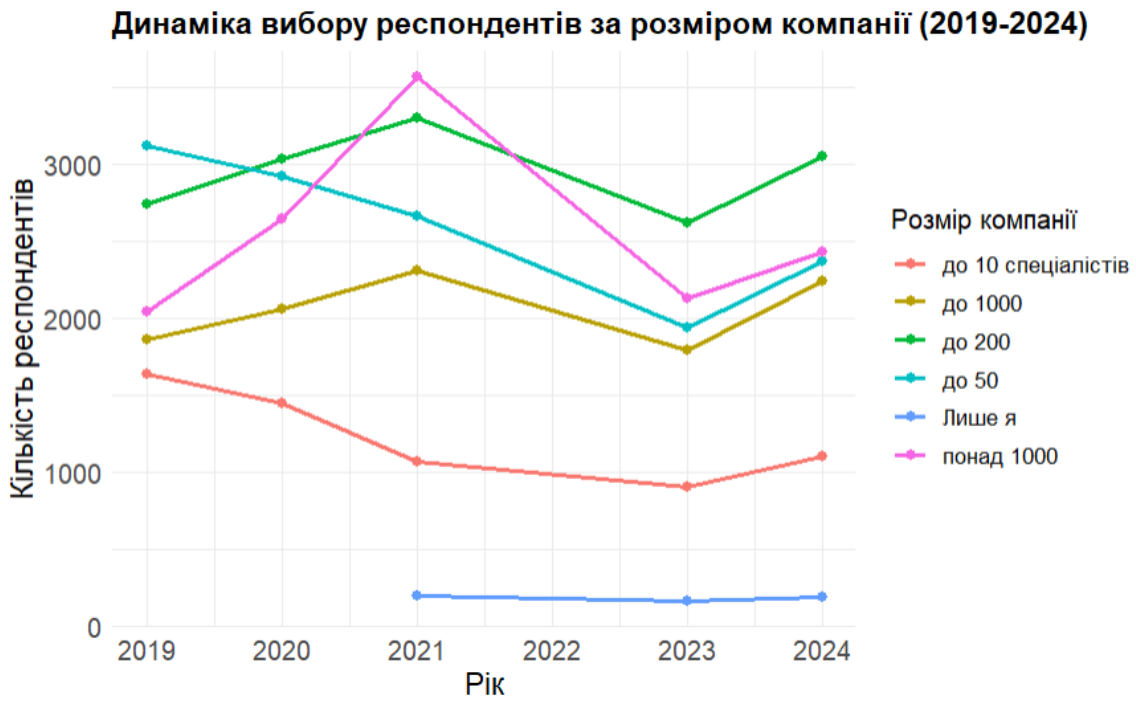

Рисунок 2.27 – Розподіл респондентів за розміром компанії

Джерело: розроблено автором на основі [50]

Найбільш помітним є зростання у 2021 році кількості респондентів, які працюють у компаніях з кількістю співробітників понад 1000. Це свідчить про зростання популярності великих компаній серед працівників, що може бути пов'язано з більшими можливостями для кар'єрного росту, більш стабільними умовами праці та більш високими заробітками. Кількість респондентів, які працюють самостійно, залишається відносно стабільною, що може свідчити про зростання популярності фрілансу та інших форм самозайнятості.

Діаграма на рисунку 2.28 демонструє тенденцію щодо зміни середньої заробітної плати в залежності від розміру компанії протягом 2019-2024 років:

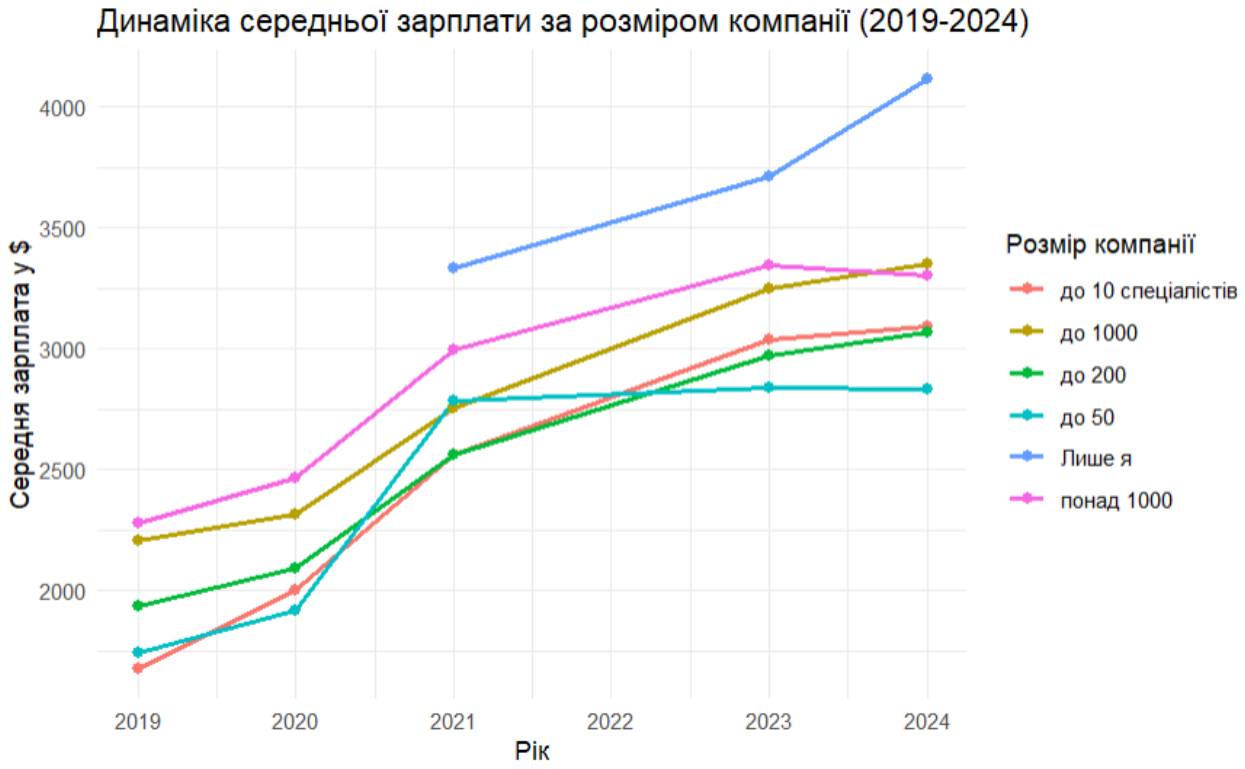

Рисунок 2.28 – Динаміка середньої зарплати за розміром компанії (2019-2024)

Джерело: розроблено автором на основі [50]

Загалом, можна помітити, що з ростом компанії, як правило, зростає і рівень середньої заробітної плати. Це цілком логічно, оскільки великі компанії, як правило, мають більші фінансові можливості і можуть пропонувати більш високу оплату праці.

Найвищий рівень зарплат спостерігається у самозайнятих особах в категорії «Лише я», де середня зарплата значно перевищує інші категорії і зростає швидше, досягнувши понад $4000 у 2024 році. Також на лідируючих позиціях стабільно знаходяться великі компанії у категорії «до 1000» та «понад 1000 працівників», які досягли у 2024 році позначки понад 3000$.

Для середніх і малих компаній (до 10, 50 та 200 працівників) спостерігається схожа тенденція поступового зростання, але їх рівень зарплат помітно нижчий порівняно з великими компаніями. Це вказує на суттєву залежність рівня оплати праці від розміру компанії: у великих компаніях

заробітна плата стабільно вища, що, ймовірно, пояснюється кращими фінансовими можливостями та вищими вимогами до працівників.

Також було побудовано діаграму, яка демонструє динаміку змін у виборі компаній різних типів респондентами протягом 2019-2024 років (рис. 2.29):

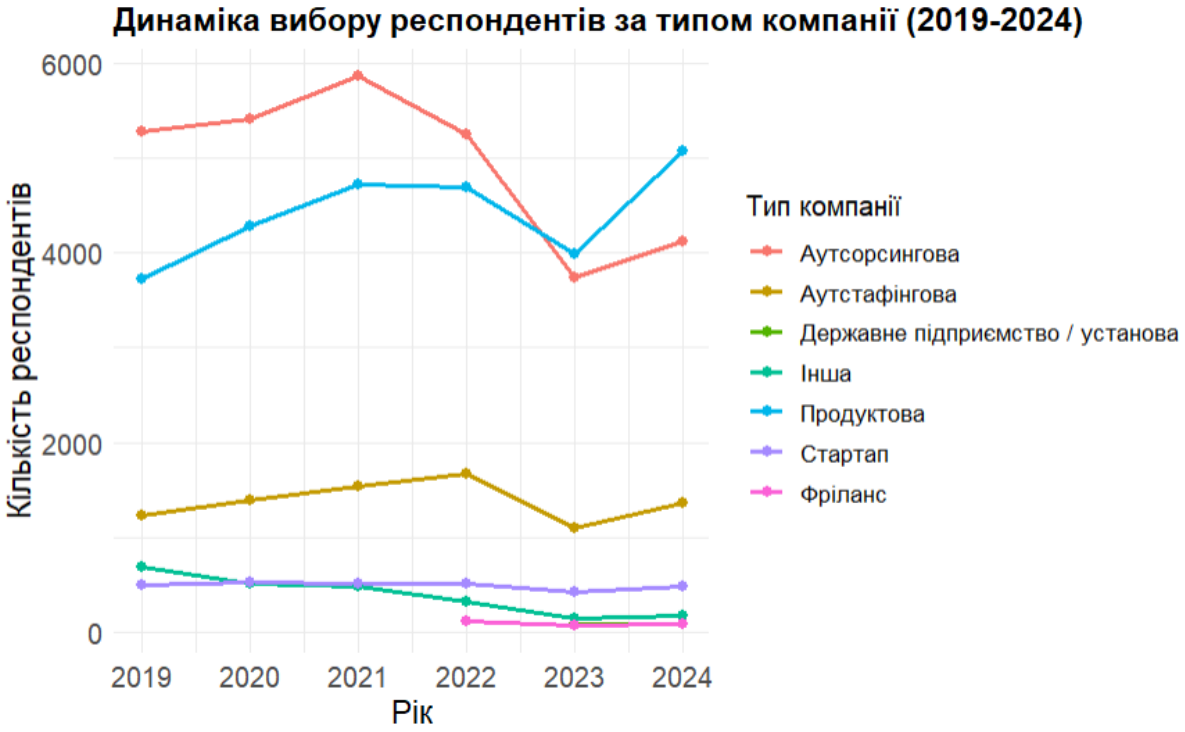

Рисунок 2.29 - Розподіл респондентів за типом компанії

Джерело: розроблено автором на основі [50]

Найбільш помітним є зростання кількості респондентів, які працюють в продуктових компаніях, адже з графіка бачимо, що попри спад у 2023 році, цей спад все ж був меншим ніж у аутсорсингових компаніях, що і вивело продуктові компанії у лідери у 2024 році. Щодо аутсорсингових компаній – після помітно найбільшого спаду серед усіх представлених на діаграмі типів компаній у 2023 році, він так і не набрав своєї колишньої популярності серед опитуваних респондентів, що і призвело до втрати лідируючої позиції. Продуктові компанії, як правило, розробляють власні продукти, що забезпечує їм більшу стабільність та перспективи розвитку, особливо в довгостроковій перспективі. Це може бути особливо привабливим для працівників у нестабільні часи, такі як війна. Вона стимулювала розвиток українських продуктів та послуг, що могло сприяти

зростанню популярності продуктових компаній. Водночас, спостерігається коливання кількості респондентів, які працюють в інших типах компаній. Кількість респондентів, які працюють у стартапах, також демонструє нестабільну динаміку, що може свідчити про високу залежність цього сегмента від інвестицій та загальної економічної ситуації.

Далі було досліджено чи є залежність між типом компанії та рівнем оплати праці (рис. 2.30):

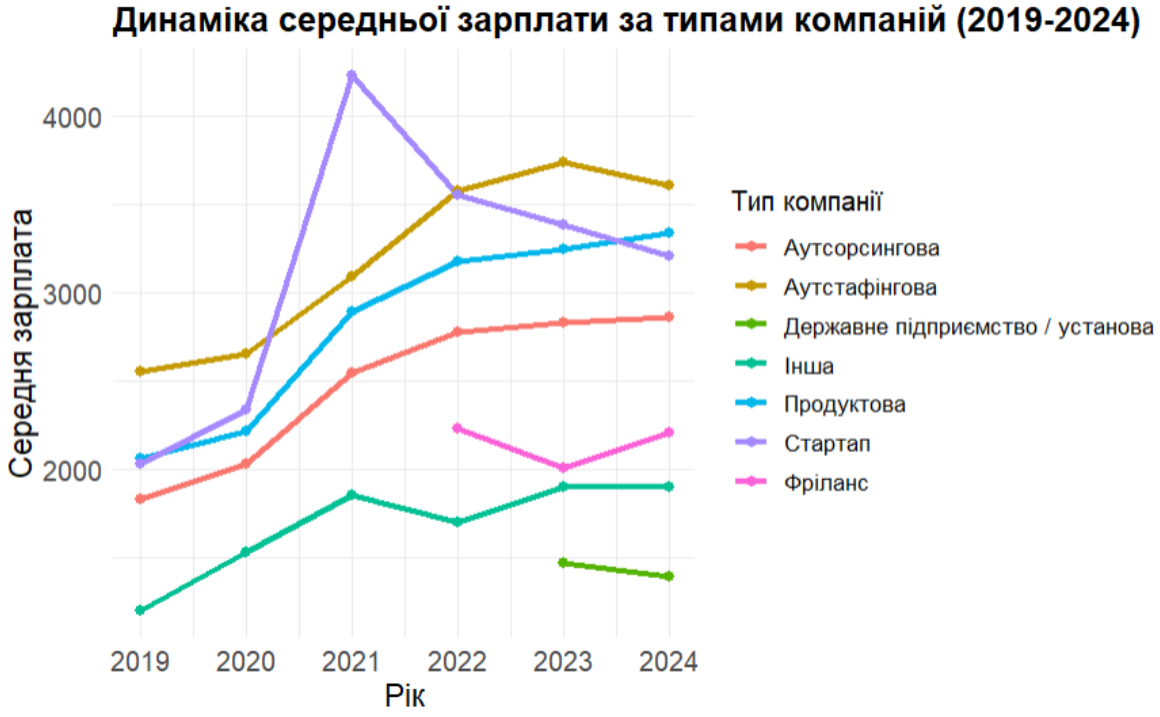

Рисунок 2.30 – Розподіл середньої заробітної плати за типом компанії

Джерело: розроблено автором на основі [50]

Найбільш помітним є стрімке зростання середньої зарплати у стартапах у 2021 році, що може свідчити про високу конкуренцію за талановитих фахівців в цьому сегменті ринку та про готовність стартапів пропонувати привабливі умови праці для залучення інвестицій. Водночас, спостерігається коливання середньої зарплати в інших типах компаній. Так, середня зарплата в аутсорсингових компаніях демонструє більш стабільний ріст, що може бути пов'язано з постійним попитом на аутсорсингові послуги. Середня зарплата в державних підприємствах та установах демонструє найменшу динаміку, що може свідчити про більш жорстку систему оплати праці в державному секторі.

Для побудови розподілу заробітної плати була використана функція n_distinct(), для визначення унікальних значень у стовпчику «Зарплата», за результатами якої отримали значення 2091. Тому було поділено усі унікальні елементи цього стовпчика у діапазони. Для цього за допомогою функції unique() було виділено всі унікальні значення зі стовпця «Зарплата», після чого за допомогою функцій min() та max() було визначено найменше (50) та найбільше (646792) значення відповідно серед усіх унікальних значень, що і визначило наші початкову та останню межу для діапазонів.

Тож було встановлено та візуалізовано динаміку зміни розподілу зарплат за роками (рис. 2.31):

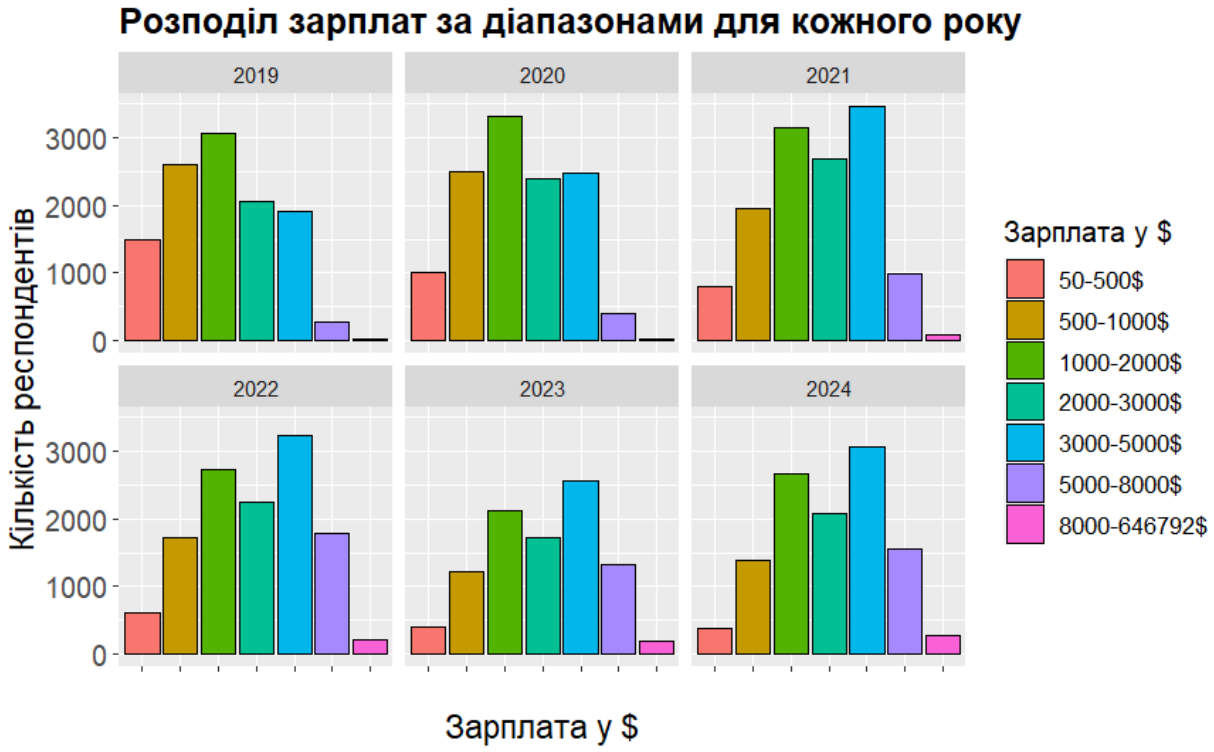

Рисунок 2.31 – Розподіл зарплати опитаних респондентів

Джерело: розроблено автором на основі [50]

Найбільш помітним є зростання кількості респондентів з високими зарплатами (понад 3000 доларів США) у 2021 та 2022 роках. Це може свідчити про загальне підвищення рівня життя та зростання заробітних плат у цей період. Водночас, спостерігається стабільність або незначне зниження кількості

респондентів з низькими зарплатами (до 1000 доларів США), що може вказувати на зменшення розриву між багатими та бідними.

Іншим цікавим моментом є зміна структури заробітних плат у різних діапазонах. Так, у 2021 році спостерігається значне зростання кількості респондентів з зарплатами від 3000 до 5000 доларів США, що може свідчити про появу нових робочих місць з високою оплатою. У наступні роки ця тенденція дещо послаблюється, але загалом рівень зарплат залишається досить високим.

Для аналізу взаємозв’язку між залежною змінною та незалежними змінними було створено кореляційну матрицю (рис. 2.32):

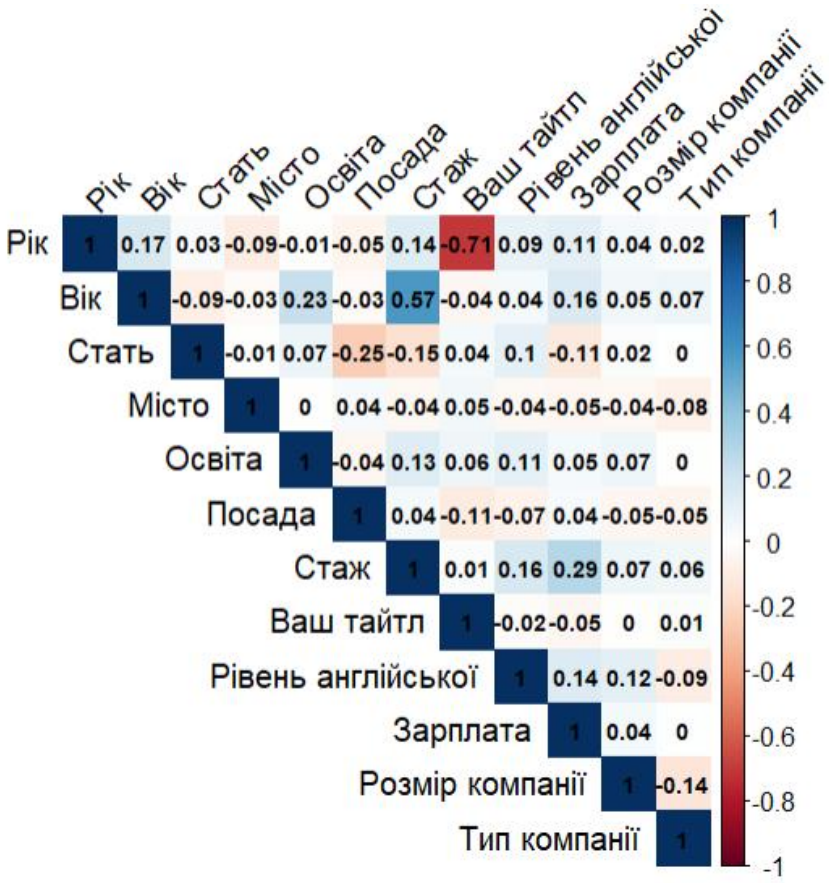

Рисунок 2.32 – Кореляційна матриця

Джерело: розроблено автором на основі [50]

Між стажем роботи та зарплатою існує сильний позитивний зв'язок (0.29). Це означає, що чим більший досвід роботи має людина, тим, як правило, вища її зарплата. Це логічно, оскільки з набуттям досвіду працівники набувають нових

навичок, стають більш цінними для компанії і, відповідно, отримують вищу оплату.

Вік також сильно корелює із зарплатою (0.16). З досвідом та набуттям нових навичок працівники стають більш цінними для компаній і, відповідно, отримують вищу оплату.

Володіння англійською мовою також позитивно впливає на рівень зарплати (0.14). Це особливо актуально для спеціальностей, які передбачають міжнародну співпрацю або роботу з іноземною документацією.

Стать також має негативний зв’язок із зарплатою (-0.11). Це означає, що зі зміною статі (умовно переходом від однієї категорії до іншої, наприклад, від чоловіків до жінок), середня зарплата зменшується.

2.2 Методи машинного навчання для аналізу факторів, що впливають на рівень оплати праці

У цьому підрозділі застосовуються методи машинного навчання для прогнозування заробітної плати. Тут ми використовуємо алгоритми, які можуть навчатися на складних взаємозв'язках між детермінантами.

Метод рішення дерев (Decision Trees) використовують для побудови моделей, що дозволяють візуалізувати, як різні фактори впливають на заробітну плату. Для цього можна використати функцію rpart() (рис. 2.33):

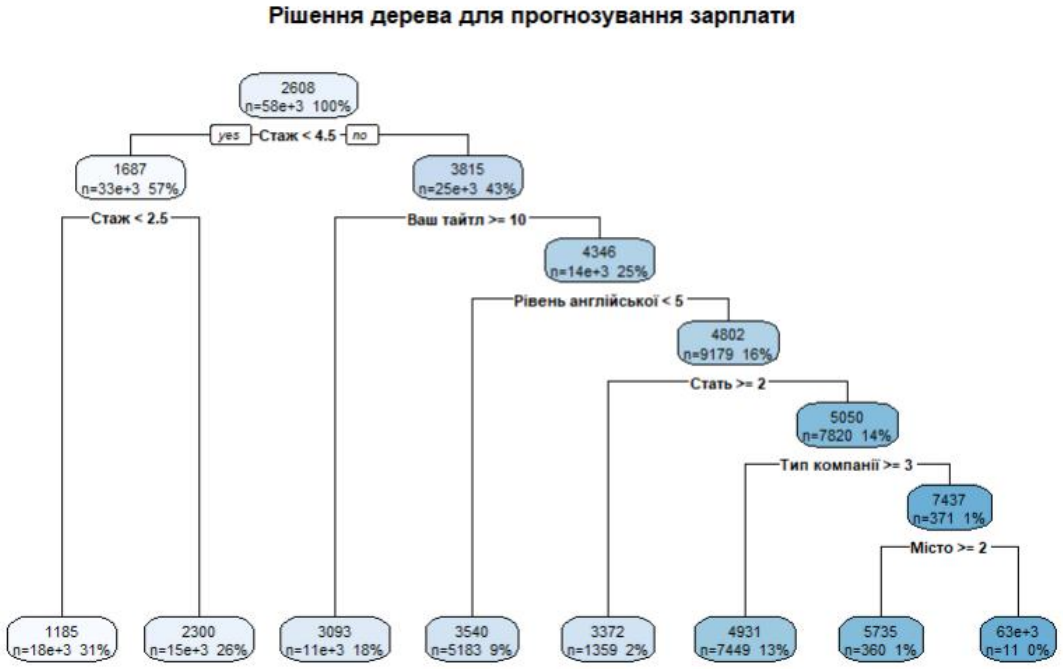

Рисунок 2.33 – Результат методу дерева рішень

Джерело: розроблено автором на основі [50]

Співробітники зі стажем менше 2.5 років мають, як правило, нижчу зарплату порівняно з тими, хто працює довше. Це свідчить про те, що досвід є важливим фактором при визначенні рівня оплати праці.

Співробітники з високим рівнем посади (Ваш тайтл >= 10) мають, як правило, вищу зарплату. Це логічно, оскільки більш високі посади зазвичай пов'язані з більшою відповідальністю та вищими вимогами до кваліфікації.

Співробітники з високим рівнем знання англійської мови також мають тенденцію до отримання вищої зарплати. Це може бути пов'язано з тим, що знання іноземних мов є цінною компетенцією на сучасному ринку праці, особливо в міжнародних компаніях.

Тип компанії також впливає на рівень зарплати. Співробітники компаній з більшим числом працівників (Розмір компанії >= 3) та компаній, що розташовані у великих містах (Місто >= 2), як правило, отримують вищу зарплату. Це може бути пов'язано з більшими можливостями для кар'єрного зростання та більш високими виплатами у великих компаніях та міських центрах.

РОЗДІЛ 3
ОЦІНЮВАННЯ ДЕТЕРМІНАНТ ОПЛАТИ ПРАЦІ

3.1 Метод регресійних моделей

У цьому підрозділі використовувались лінійні регресійні моделі для оцінки впливу різних детермінант на рівень заробітної плати. Моделі регресії дозволяють оцінити значення коефіцієнтів для кожного фактора (гендер, освіта, досвід, місце проживання) та їх вплив на зарплату.

Для подальшого моделювання дані були поділені на дві частини: навчальну та тестову вибірки, використовуючи випадковий розподіл. Тестова вибірка складає 20% від загальної кількості даних, а навчальна – 80%.

Кожен елемент має рівні шанси потрапити в одну з цих груп. Навчальна вибірка використовується для тренування моделі, тоді як тестова вибірка залишається поза процесом навчання і буде використовуватися для валідації моделі. Завдяки такому підходу ми можемо оцінити точність та ефективність моделі на нових, невідомих даних.

Було побудовано модель лінійної регресії, параметри якої було проаналізовано та оцінено (рис. 3.1):

```
Call:
lm(formula = Зарплата ~ Рік + Вік + Стать + Місто + Освіта +
    Посада + Стаж + `Ваш тайтл` + `Рівень англійської` + `Розмір компанії` +
    `Тип компанії`, data = data_numeric)

Residuals:
   Min     1Q Median     3Q    Max
 -6348   -793   -166    549 642245

Coefficients:
                      Estimate Std. Error t value
(Intercept)         -3.550e+05  2.665e+04 -13.321
Рік                  1.755e+02  1.316e+01  13.336
Вік                 -7.351e+00  3.429e+00  -2.144
Стать               -8.123e+02  4.027e+01 -20.171
Місто               -1.998e+01  2.959e+00  -6.752
Освіта               6.532e+01  1.867e+01   3.498
Посада               1.420e+01  2.721e+00   5.218
Стаж                 3.081e+02  6.179e+00  49.863
`Ваш тайтл`          2.061e+01  1.077e+01   1.915
`Рівень англійської` 4.516e+02  1.795e+01  25.164
`Розмір компанії`    3.460e+01  1.231e+01   2.811
`Тип компанії`      -5.573e+00  1.435e+01  -0.388
                    Pr(>|t|)
(Intercept)          < 2e-16 ***
Рік                  < 2e-16 ***
Вік                 0.032053 *
Стать                < 2e-16 ***
Місто               1.47e-11 ***
Освіта              0.000469 ***
Посада              1.81e-07 ***
Стаж                 < 2e-16 ***
`Ваш тайтл`         0.055545 .
`Рівень англійської` < 2e-16 ***
`Розмір компанії`   0.004943 **
`Тип компанії`      0.697735
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3866 on 57581 degrees of freedom
Multiple R-squared:  0.1056,    Adjusted R-squared:  0.1054
F-statistic: 617.8 on 11 and 57581 DF,  p-value: < 2.2e-16
```

Рисунок 3.1 - Основні статистичні показники для моделі лінійної регресії

Джерело: розроблено автором на основі [50]

Рівняння отриманої лінійної регресії має вигляд:

$$\bar{Y} = -3.550e+05 + 1.755e+02 \cdot X_1 - 7.351e+00 \cdot X_2 - 8.123e+02 \cdot X_3 - 1.998e+01 \cdot X_4 + 6.532e+01 \cdot X_5 + 1.420e+01 \cdot X_6 + 3.081e+02 \cdot X_7 + 2.061e+01 \cdot X_8 + 4.516e+02 \cdot X_9 + 3.460e+01 \cdot X_{10} - 5.573e+00 \cdot X_{11}$$

Отримані результати свідчать про те, що модель є статистично значущою, оскільки значення F-статистики є високим, а p-value менше 0.05. Це означає, що

принаймні один з предикторів (незалежних змінних) має статистично значущий вплив на залежну змінну (зарплату).

Чим більший стаж роботи, тим вища зарплата. Коефіцієнт при змінній "Стаж" є значним і додатним, що свідчить про пряму пропорційність між цими двома змінними.

Рівень освіти також позитивно впливає на зарплату. Коефіцієнт при змінній "Освіта" є значним і додатним, що підтверджує, що люди з вищою освітою, як правило, отримують більшу зарплату.

Посада також має позитивний вплив на зарплату. Коефіцієнт при змінній "Посада" є значним і додатним, що свідчить про те, що працівники, які займають більш високі посади, отримують більшу оплату.

R-квадрат показує, яка частина дисперсії залежної змінної (зарплати) пояснюється незалежними змінними. У нашому випадку R-квадрат дорівнює 0.1056, що означає, що модель пояснює близько 10.56% дисперсії зарплати. Це відносно невелике значення, що свідчить про те, що існують інші фактори, які впливають на зарплату, але не були включені в модель.

Стандартна помилка моделі характеризує середнє відхилення фактичних значень зарплати від передбачених моделлю. Чим менше значення стандартної помилки, тим точніше модель описує дані.

Узагальнюючи результати, тестові значення було порівняно з прогнозованими, що дозволило оцінити точність моделі та її здатність робити прогнози на основі незалежних змінних.

На рисунку 3.2 представлено графіки фактичних та спрогнозованих значень місячної заробітної плати. Хоча модель загалом здатна описувати певні закономірності, спостерігається чимало випадків, де розбіжності між реальними та прогнозованими значеннями значні. Це свідчить про обмеження моделі, що можуть бути викликані неврахованими нелінійними залежностями або впливом зовнішніх факторів, які не були включені до аналізу.

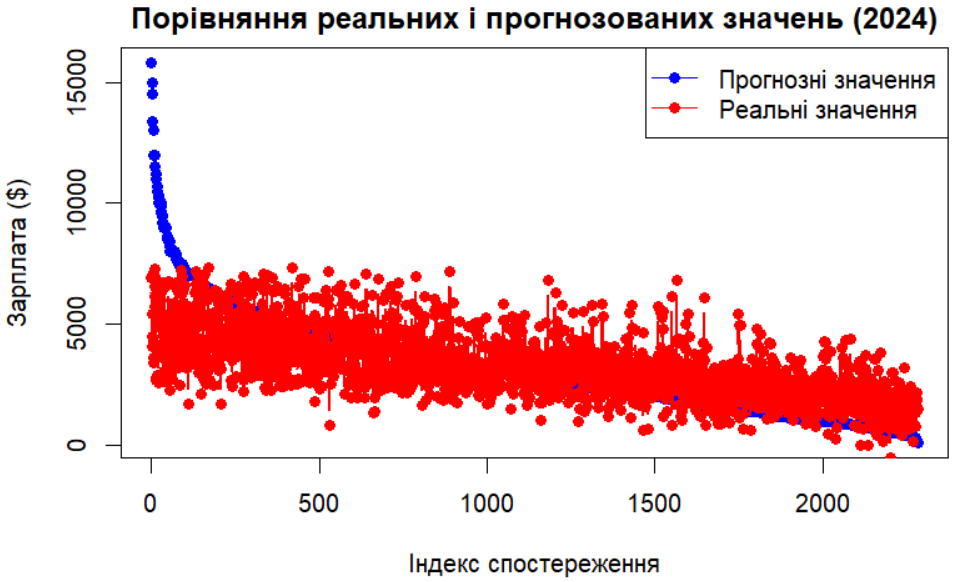

Рисунок 3.2 - Порівняння реальних і прогнозних значень для моделі лінійної регресії

Джерело: розроблено автором на основі [50]

Реальні значення демонструють значну варіативність, тобто зарплати співробітників відрізняються в широкому діапазоні. Прогнозовані значення загалом відтворюють загальну тенденцію реальних даних, проте спостерігається певна відмінність між ними.

Модель, яка використовувалась для прогнозування, здається, досить добре вловлює загальну картину, але не завжди точно передбачає індивідуальні значення. Можливі причини таких відхилень можуть бути пов'язані з наявністю в даних шумів, впливом факторів, які не були враховані в моделі, або з нелінійністю залежностей між змінними. Для більш детального аналізу варто розглянути додаткові метрики оцінки якості моделі, такі як середньоквадратична помилка або коефіцієнт детермінації, а також побудувати графіки залишків.

Застосовано метод LASSO регуляризації, що дозволяє зменшувати складність моделі та запобігати перенавчанню. Це метод регуляризації, який допомагає боротися з мультиколінеарністю та великою кількістю змінних, зменшуючи вагу менш значущих факторів.

Для корекції моделі з багатьма змінними було використано функцію glmnet з пакету glmnet (рис. 3.3):

```r
# Побудова моделі з регуляризацією
lasso_cv <- cv.glmnet(
  X_train, Y_train,
  alpha = 1, # Лассо-регуляризація
  standardize = TRUE, # Стандартизація змінних
  nfolds = 10 # Крос-валідація
)

# Візуалізація результатів крос-валідації
plot(lasso_cv)
```

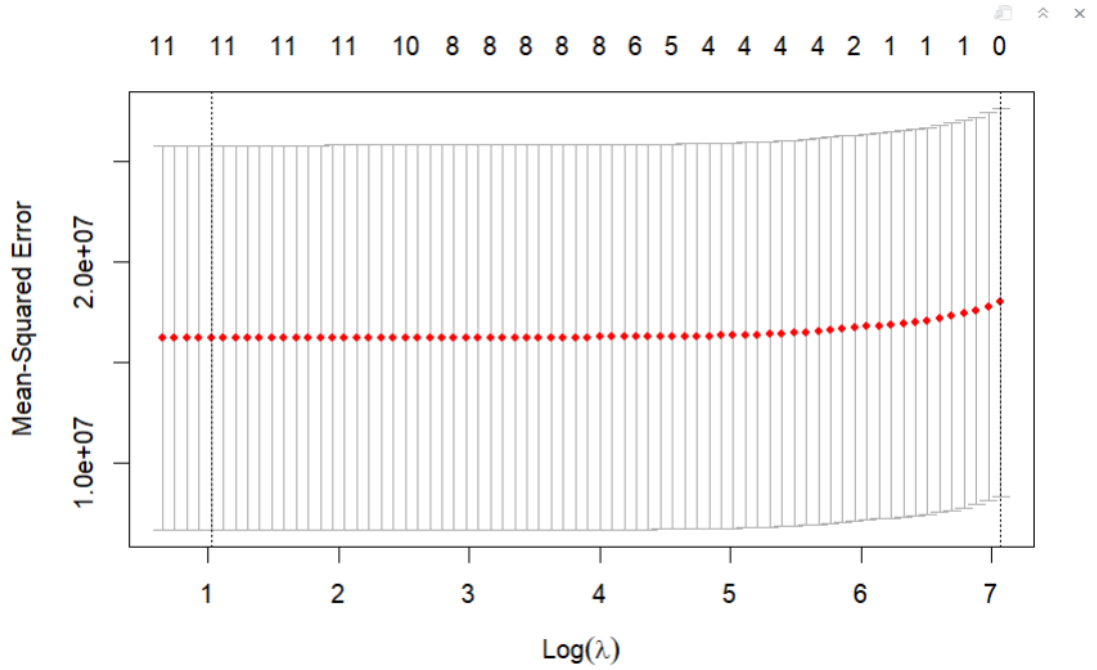

Рисунок 3.3 – Результат крос-валідації

Джерело: розроблено автором на основі [50]

Графік крос-валідації моделі Lasso демонструє відносно стабільну середньоквадратичну помилку (MSE) для широкого діапазону значень параметра регуляризації (lambda). Це вказує на те, що модель не надзвичайно чутлива до вибору цього параметра. Однак, відсутність чітко вираженого мінімуму MSE ускладнює визначення оптимального значення lambda. Така ситуація може бути пов'язана з кількома факторами, такими як обмежений обсяг даних, слабкі кореляції між змінними або наявність сильного шуму в даних.

Для вибору оптимального значення lambda можна розглянути кілька підходів. Наприклад, можна вибрати найменше значення lambda, для якого MSE не відрізняється значно від мінімального значення. Також варто розглянути інші критерії оцінки моделі, такі як середнє абсолютне відхилення (MAE) або коефіцієнт детермінації (R-квадрат).

Після вибору оптимального значення lambda необхідно оцінити якість моделі на незалежному тестовому наборі даних. Крім того, важливо проаналізувати коефіцієнти моделі, щоб зрозуміти, які змінні були відібрані моделлю Lasso і як вони впливають на цільову змінну.

Загалом, результати крос-валідації свідчать про те, що модель Lasso може бути корисна для аналізу ваших даних, але вибір оптимального значення lambda вимагає додаткового аналізу.

Згідно вказаних результатів вище, було здійснено пошук оптимального значення параметра регуляризації (lambda) (рис. 3.4):

```
# Оптимальне значення λ
best_lambda <- lasso_cv$lambda.min
cat("Оптимальне значення λ:", best_lambda, "\n")

# Побудова фінальної моделі з оптимальним λ
lasso_model <- glmnet(
  X_train, Y_train,
  alpha = 1,
  lambda = best_lambda,
  standardize = TRUE
)

# Перегляд коефіцієнтів моделі
coef(lasso_model)

# Прогнозування для тестової вибірки
Y_pred <- predict(lasso_model, s = best_lambda, newx = X_test)

# Оцінка точності моделі
mse <- mean((Y_test - Y_pred)^2)
cat("Середня квадратична помилка (MSE):", mse, "\n")
```

```
Оптимальне значення λ: 2.77781
12 x 1 sparse Matrix of class "dgCMatrix"
                                s0
(Intercept)        -3.375661e+05
Рік                 1.668877e+02
Вік                -5.872627e+00
Стать              -8.278326e+02
Місто              -1.944522e+01
Освіта              6.382345e+01
Посада              1.403442e+01
Стаж                3.066391e+02
Ваш тайтл           1.291580e+01
Рівень англійської  4.493097e+02
Розмір компанії     3.714824e+01
Тип компанії       -8.096591e+00
Середня квадратична помилка (MSE): 9805990
```

Рисунок 3.4 – Пошук оптимального параметра регуляризації (lambda)

Джерело: розроблено автором на основі [50]

Оптимальне значення лямбда було визначено за допомогою крос-валідації і становить 2.77781. Це значення мінімізує середньоквадратичну помилку (MSE)

на валідаційних множинах, що вказує на найкращу здатність моделі узагальнювати на нові дані серед усіх розглянутих значень лямбда.

Побудова фінальної моделі здійснюється з використанням обраного оптимального значення лямбда. Це дозволяє отримати модель, яка має найкращу комбінацію складності та точності передбачення.

Рівняння отриманої лінійної регресії Lasso має вигляд:

$$\bar{Y} = -3.363898e+05 + 1.663070e+02 \cdot X_1 - 5.748191e+00 \cdot X_2 - 8.271234e+02 \cdot X_3 - 1.940024e+01 \cdot X_4 + 6.344803e+01 \cdot X_5 + 1.397734e+01 \cdot X_6 + 3.065166e+02 \cdot X_7 + 1.244758e+01 \cdot X_8 + 4.491843e+02 \cdot X_9 + 3.701267e+01 \cdot X_{10} - 7.877971e+00 \cdot X_{11}$$

Аналіз коефіцієнтів моделі показує, що деякі змінні мають статистично значущий вплив на цільову змінну (зарплата). Зокрема, позитивний вплив мають такі змінні як "Рік", "Освіта", "Стаж", "Рівень англійської", "Розмір компанії", а негативний - "Стать", "Місто", "Тип компанії". Це означає, що збільшення значень цих змінних, як правило, пов'язане зі зростанням зарплати, тоді як зменшення - зі зниженням.

Оцінка точності моделі на тестовій вибірці показала, що середньоквадратична помилка (MSE) становить 9805990. Це значення характеризує середнє квадратичне відхилення прогнозів моделі від фактичних значень. Чим менше значення MSE, тим точніші прогнози робить модель. Отримане значення MSE дозволяє оцінити якість моделі, але для більш детального аналізу варто порівняти його з результатами інших моделей або з деяким еталонним значенням.

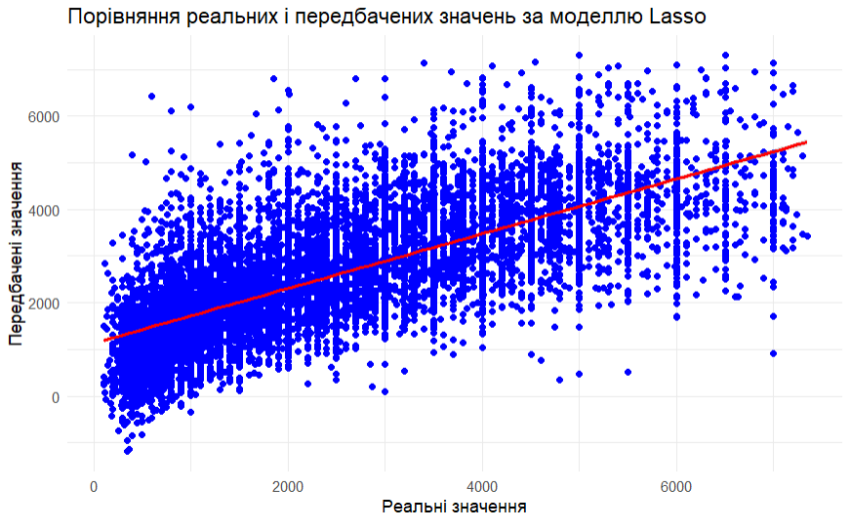

Рисунок 3.5 – Порівняння реальних і передбачених значень за моделлю Lasso

Джерело: розроблено автором на основі [50]

Загалом, результати проведеного аналізу свідчать про те, що модель Lasso, побудована з оптимальним значенням лямбда, є адекватною для прогнозування зарплати.

3.2 Метод градієнтного бустингу

Метод градієнтного бустингу використовують для підвищення точності моделей. Можна застосувати алгоритми на кшталт xgboost або gbm. (рис. 3.6):

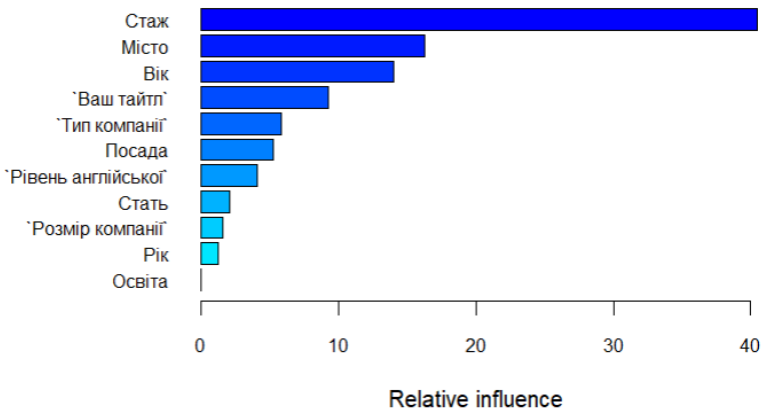

Рисунок 3.6 – Результат градієнтного бустингу

Джерело: розроблено автором на основі [50]

Кожна смужка на графіку відповідає певному фактору (незалежній змінній), а її довжина відображає відносний вплив цього фактора на цільову змінну – зарплату. Чим довша смужка, тим більший вплив відповідного фактора на зарплату.

З графіка видно, що найбільший вплив на рівень зарплати має стаж роботи. Це означає, що чим довший стаж роботи, тим вища, як правило, зарплата. Це логічно, оскільки з досвідом зростає кваліфікація співробітника, що дозволяє йому виконувати більш складні завдання і претендувати на більш високу оплату.

На другому місці за впливом на зарплату стоїть місто. Це свідчить про те, що рівень зарплати значною мірою залежить від географічного розташування компанії. Можливо, в деяких містах вищі ціни на життя, більша конкуренція за робочу силу або просто більш високий рівень оплати праці в певних галузях.

Модель градієнтного бустингу дозволила визначити найважливіші фактори, які впливають на рівень зарплати. Результати моделі підтверджують інтуїтивні уявлення про те, що такі фактори, як стаж роботи, посада та місцезнаходження компанії, мають найбільший вплив на заробітну плату.

Також було побудовано модель багатошарового персептрона (MLP) (рис. 3.7):

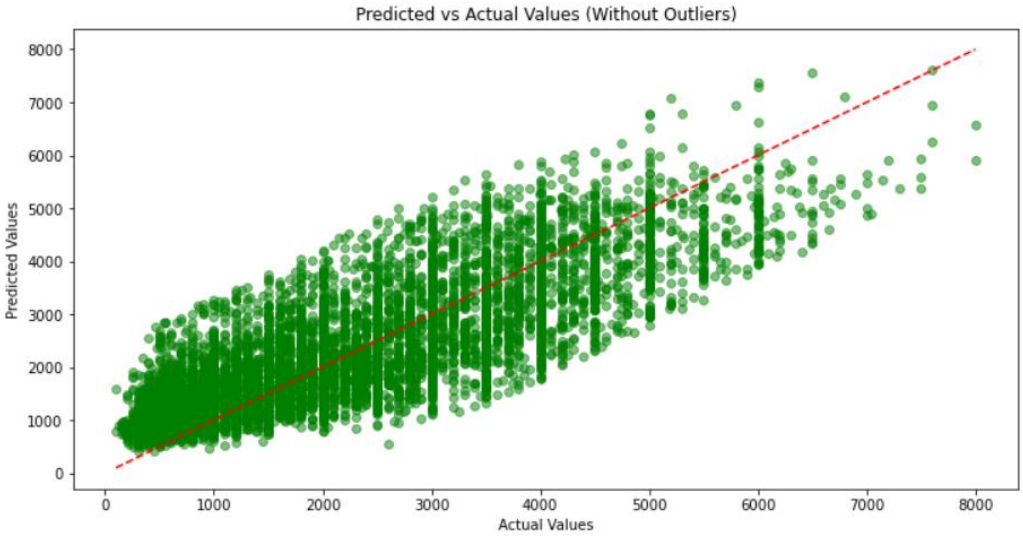

Рисунок 3.7 – Порівняння реальних та прогнозованих значень за моделлю багатошарового персептрона

Джерело: розроблено автором на основі [50]

На діаграмі зображено розсіювання прогнозованих значень моделі багатошарового персептрона відносно реальних значень. Кожна точка на графіку представляє одну пару: фактичне значення та прогноз моделі. Лінія, що проходить по діагоналі графіка, відображає ідеальний випадок, коли прогнозовані значення повністю збігаються з реальними.

Здебільшого, точки на діаграмі розташовані вздовж діагоналі, що свідчить про досить високу кореляцію між прогнозованими та реальними значеннями. Це означає, що модель загалом досить добре уловлює залежності в даних і може робити адекватні прогнози. Точки розподілені досить щільно вздовж діагоналі, що вказує на хорошу точність прогнозування для більшості даних. Невелика кількість точок відхиляється від діагоналі, що може свідчити про наявність викидів у даних або про те, що модель має труднощі з прогнозуванням в деяких областях. Розподіл точок відносно діагоналі здається досить симетричним, що вказує на відсутність систематичної похибки в прогнозах.

На основі представленої діаграми можна зробити висновок, що модель багатошарового персептрона демонструє досить високу точність прогнозування.

3.3 Порівняння застосованих методів та отриманих прогнозів

У цьому підрозділі проводимо перевірку точності та валідності побудованих моделей. Це дозволяє оцінити ефективність наших моделей для прогнозування заробітної плати.

Було порівняно метрики усіх створених моделей (рис. 3.8):

Model <chr>	RMSE <dbl>	MAE <dbl>	MAPE <dbl>	R2 <dbl>
Linear	3865.789	967.6121	59.91487	0.1055662
Lasso	3865.798	967.4333	59.93122	0.1055627
Rpart	3802.402	1019.5102	64.13521	0.1346576
GBM	3799.904	832.6730	47.57522	0.1364278
XGBoost	2623.330	701.4579	36.54055	0.6377283
Sequential	1745.910	814.1600	53.70000	0.3400000

Рисунок 3.8 – Порівняння результатів моделей

Джерело: розроблено автором на основі [50]

Наведена таблиця представляє результати роботи різних моделей машинного навчання, які були використані для прогнозування зарплати. Кожна модель має свої особливості та підходи до прогнозування, а метрики оцінки допомагають нам зрозуміти, наскільки добре кожна модель впоралась зі своїм завданням.

Лінійна модель та Lasso: Ці моделі показали найгірші результати. Це може вказувати на те, що залежність між зарплатою та незалежними змінними є нелінійною і не може бути адекватно описана простою лінійною моделлю.

Дерево рішень (Rpart): Ця модель показала дещо кращі результати, ніж лінійні моделі, але все ще значно поступається XGBoost. Це може бути пов'язано з тим, що дерево рішень може захоплювати нелінійні залежності, але може бути схильне до перенавчання.

Градієнтний бустинг (GBM): Ця модель показала значно кращі результати, ніж лінійні моделі та дерево рішень. Це свідчить про ефективність ансамблевих методів, які дозволяють побудувати більш складні моделі.

Багатошаровий персептрон (Sequential): Ця модель також показала хороші результати, але трохи гірші за XGBoost. Це може бути пов'язано з тим, що налаштування нейронних мереж є більш складним процесом, ніж налаштування дерев рішень або градієнтного бустингу.

На основі представлених результатів можна зробити висновок, що модель XGBoost є найкращою для прогнозування зарплати в нашому наборі даних. Вона дозволяє досягти найвищої точності прогнозування і найкраще пояснює залежності між змінними.

Аналіз визначив фактори, які кожна модель вважає найбільш значущими, і результати цього аналізу представлені в таблиці 3.1:

Таблиця 3.1 - Порівняння важливих факторів для використаних алгоритмів

Priority	**Linear_Regression**	**Lasso_Regression**	**Decision_Tree**	**Gradient_Boosting**	**MLP**
1	Стать	Стать	Стаж	Стаж	Стаж
2	Рівень англійської	Рівень англійської	Місто	Вік	Стать
3	Стаж	Стаж	Вік	Місто	Рівень англійської
4	Рік	Рік	Ваш тайтл	Ваш тайтл	Вік
5	Освіта	Освіта	Рівень англійської	Посада	Посада
6	Розмір компанії	Розмір компанії	Рік	Тип компанії	Рік
7	Ваш тайтл	Місто	Стать	Рівень англійської	Освіта
8	Місто	Посада	Тип компанії	Розмір компанії	Місто
9	Посада	Ваш тайтл	Посада	Стать	Тип компанії
10	Вік	Тип компанії		Рік	Розмір компанії
11	Тип компанії	Вік		Освіта	Ваш тайтл

Джерело: розроблено автором на основі [50]

Фактор "Стаж" є дуже важливим для всіх алгоритмів, займаючи найвищий пріоритет у більшості з них. Це свідчить про те, що досвід роботи є критичним фактором для успішного використання цих алгоритмів.

"Рівень англійської" та "Стать" також є одним з ключових факторів, який входить до провідної трійки для більшості алгоритмів. Це підкреслює важливість мовних навичок для ефективного застосування цих методів.

Такі фактори, як "Вік" та "Посада", мають дещо нижчий пріоритет, але все ще вважаються значущими для певних алгоритмів, таких як "Gradient Boosting" та "MLP".

ВИСНОВКИ

Проведене дослідження спрямоване на аналіз існуючих факторів, що впливають на рівень заробітної плати, розробку моделі прогнозування та проведення детального статистичного та візуального аналізу даних. Отримані результати дозволяють зробити ряд важливих висновків.

Аналіз існуючих факторів показав, що на рівень заробітної плати впливає широкий спектр факторів, таких як: досвід роботи, освіта, галузь, розмір компанії, географічне розташування тощо. Деякі з цих факторів мають більший вплив на рівень зарплати, ніж інші. Зокрема, було встановлено, що досвід роботи та рівень освіти є найважливішими детермінантами заробітної плати.

Розроблена модель прогнозування заробітної плати на основі градієнтного бустингу показала високу точність прогнозування, у якої R-квадрат становить 0,64, що є найближчим з усіх отриманих значень до одиниці та середньоквадратична похибка, яка становить 701. Це свідчить про те, що модель адекватно описує залежність між заробітною платою та незалежними змінними.

Статистичний аналіз даних дозволив виявити ряд цікавих закономірностей. Зокрема, було встановлено, що існує значна різниця в рівні заробітної плати між чоловіками та жінками, а також що співробітники з вищою освітою в середньому заробляють більше, ніж співробітники зі середньою спеціальної освітою. Ці результати можуть бути використані для подальших досліджень та розробки рекомендацій щодо поліпшення системи оплати праці.

Візуальний аналіз даних допоміг наочно представити отримані результати. Були побудовані різноманітні графіки, такі як діаграми розсіювання, гістограми, коробкові діаграми, які дозволили краще зрозуміти взаємозв'язки між змінними та виявити аномалії в даних.

Крім того, результати аналізу також вказують на певні проблеми, які потребують подальшого вивчення та вирішення. Наприклад, виявлена суттєва

розбіжність у рівні заробітної плати між чоловіками та жінками на аналогічних посадах може свідчити про наявність гендерної нерівності на ринку праці. Це питання потребує більш глибокого вивчення причин таких відмінностей та розробки заходів для забезпечення справедливої системи оплати праці незалежно від статі працівників. Аналогічно, різниця в оплаті праці між співробітниками з різним рівнем освіти може відображати необхідність перегляду системи компенсацій з метою кращого стимулювання безперервного професійного розвитку. Ці та інші аспекти, виявлені в ході дослідження, становлять основу для подальшого вдосконалення політики управління персоналом та оплати праці в організаціях.

Проведене дослідження є важливим кроком до розуміння факторів, що впливають на рівень заробітної плати.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Berndt E. R. The Practice of Econometrics: Classic and Contemporary. 1991. Addison-Wesley. 846 p.
2. Commons J. R., Andrews J. B. Principles of Labor Legislation. Harper and Brothers. 1936. 215 p.
3. Hanagan M. P. Guilds, Unions, and the Evolution of the Labor Market. JSTOR, 1980. P. 161 – 166.
4. Кравченко О. О. Інноваційні підходи до мотивації праці. *Фінансовий простір*. 2018: С. 170-178.
5. Близнюк В. В. Людський капітал як фактор економічного розвитку. *Економіка і прогнозування*. 2000: С. 64-74.
6. Левандовська Н. І. Фактори, які впливають на систему оцінки праці робітників, і їх класифікація. *Економічні науки. Випуск 12. Частина 1.* 2007. С. 124-129.
7. Балан О. Д., Савченко Ю. К. Оплата праці та її соціально-економічні аспекти. *Агросвіт*. 2018: С. 22-26.
8. Про оплату праці. Офіційний вебпортал Верховної Ради України. URL: https://zakon.rada.gov.ua/laws/show/108/95-вр#Text (дата звернення: 25.11.2024).
9. Економіка праці й соціально-трудові відносини: підручник / О.В. Шкільова. Київ: Четверта хвиля, 2008. 472 с.
10. Соціально-економічне становище України. Державна служба статистики України. URL: http:/ /www.ukrstat.gov.ua/ (дата звернення: 25.11.2024).
11. Колот А. М. Оплата праці на підприємстві: організація та удосконалення. *Кафедра соціоекономіки та управління персоналом*. 1997. 192 с.

12. Васильчак С. В., Жидяк О. Р., Полянчич Т. М. Теоретичні основи формування оплати праці на підприємстві. *Науковий вісник НЛТУ України*. 2011. С. 152-157.
13. Потриваєва Н. В., Савченко І. В. Стан та перспектива обліку розрахунків з оплати праці: теоретичний аспект. *Економічний форум*. 2014. С. 243-245.
14. Машевська А. А. Теоретичне підґрунтя організації оплати праці суб'єктів господарювання. *Ефективна економіка.* 2019. С. 9-11.
15. Миронова Ю. Ю., Панасенко В. А. Проблеми організації обліку розрахунків з оплати праці на підприємстві. *Економіка і регіон*. 2016. С. 121-126.
16. Скорнякова Ю., Лапшункова, О. Організація обліку розрахунків з персоналом щодо оплати праці. *Економіка та суспільство*. 2023. С.1-7.
17. Ладунка І. С., Зажерило А. І. Напрямки вдосконалення організації оплати праці на підприємствах. *Мукачівський державний університет:«Економіка та управління підприємствами»*. 2018. С. 394-397.
18. Перепадя Ф. Л., Тонких Л. С. Управління фондом оплати праці персоналу промислових підприємств. 2015. С.550-554.
19. Вегера В. М. Оплата праці: поняття, особливості. *Актуальні проблеми держави і права*. 2014. С. 419-425 (дата звернення: 15.11.2024).
20. Autor D. H. Work of the Future: Shaping Technology and Institutions. *MIT Work of the Future*. 2019. P.32-40.
21. Clark G. Farewell to Alms: A Brief Economic History of the World. *Princeton University Press*. 2010. P.133-144.
22. Clegg D. Labor Market Policy. *OECD*. 2015 P.183-189.
23. Stiglitz J. E. Globalization and its Discontents. *Norton & Company*. 2002. P. 293–296.

24. World Economic Forum. Global Gender Gap Report. URL: https://www.weforum.org (accessed: 15.11.2024).
25. Мінімальна заробітна плата в Україні. URL: https://index.minfin.com.ua/ua/labour/salary/min/ (дата звернення: 15.11.2024).
26. European Commission. Equality Strategy 2020-2025. URL: https://ec.europa.eu/newsroom/just/items/682425/en#:~:text=The%20Gender%20Equality%20Strategy%202020%2D2025%20sets%20out%20key%20actions,in%20all%20EU%20policy%20areas.&text=Striving%20for%20a%20Union%20of,all%20their%20diversity%20%2D%20are%20equal. (accessed: 15.11.2024).
27. OECD. Labor Market Regulation and Minimlicy. URL: https://www.oecd.org/en/topics/employment-protection-and-minimum-wages.html (accessed: 15.11.2024).
28. World Economic Forum. The Future of Jobs Report URL: https://www.weforum.org (accessed: 15.11.2024).
29. Choi S. Cultural Influences on Work Motivation and Employee Outcomes. *Human Resource Management Review*. 2019. P. 178-196.
30. Bell D., Montague J. Gender Pay Gap in a Globalized World: Cultural Differences and Economic Factors. *Global Economics Review*. 2020. P. 46-76.
31. Chen S., Zhao Y. The Role of Corporate Culture in Shaping Salary Determinants in Global Firms. *International Journal of Business and Social Science*. 2018. P. 315-366.
32. Kabeer N. Gender and the Role of Social Norms in Determining the Wage Gap. *Development Policy Review*. 2017. P. 152-163.
33. Hofstede G., Hofstede G. J., Minkov M. Cultures and Organizations: Software of the Mind. McGraw-Hill. 2010. P. 32-45.
34. Kohn M., Schooler C. Work and Personality: An Inquiry into the Impact of Social Stratification. Journal of Social Issues. 2018. P. 294-308.

35. Ryan R. M., Deci E. L. Self-determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-being. *American Psychologist*. 2017. P. 88-101.
36. Robbins S. P., Judge T. A. Organizational Behavior. *Pearson Education*. 2019. P. 248-255.
37. Gagné M., Deci E. L. Self-determination Theory in Work Organizations: Promoting Human Potential in Positive Psychology. *APA Handbook of Positive Psychology*. 2017. P. 73-87.
38. Eisenberger R., Cameron J. Detrimental Effects of Reward on Intrinsic Motivation: Reality or Myth?. *American Psychologist*. 2020. P. 113-125.
39. RPubs - R & tidyverse. *RPubs*. URL: https://rpubs.com/pbenavides/r_tidyverse_101 (accessed: 25.11.2024).
40. Hyndman R. J., Athanasopoulos G. Forecasting: Principles and Practice. 3rd ed. *OTexts*. URL: https://otexts.com/fpp3/ (accessed: 25.11.2024).
41. Hadley W. Tidy Data. *The Journal of Statistical Software*. 2014. P. 1-22.
42. Hadley W., Grolemund G. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. *O'Reilly Media, Inc*. 2016. P. 345-419.
43. F Distribution. R Tutorial. *An R Introduction to Statistics*. URL: https://www.r-tutor.com/elementary-statistics/probability-distributions/f-distribution (accessed: 25.11.2024).
44. Logistic Regression. R Tutorial. *An R Introduction to Statistics*. URL: https://www.r-tutor.com/elementary-statistics/logistic-regression (accessed: 25.11.2024).
45. Analysis of Variance. R Tutorial. *An R Introduction to Statistics*. URL: https://www.r-tutor.com/elementary-statistics/analysis-variance (accessed: 25.11.2024).
46. Numerical Measures. R Tutorial. *An R Introduction to Statistics*. URL: https://www.r-tutor.com/elementary-statistics/numerical-measures (accessed: 25.11.2024).

47. Tutorial: Hello, Quarto – Quarto. *Quarto*. URL: https://quarto.org/docs/get-started/hello/rstudio.html (accessed: 25.11.2024).
48. Salary statistics on Djinni. *Djinni | Hire talent or find a job: remotely & on your own*. URL: https://djinni.co/salaries/ (accessed: 25.11.2024).
49. Зарплати українських розробників — літо 2024. *Редакція DOU*. URL: https://dou.ua/lenta/articles/salary-report-devs-summer-2024/ (дата звернення: 25.11.2024).
50. Волошин Сергій. Датасет 2024_june_raw.csv. *GitHub*. URL: https://github.com/devua/csv/blob/master/salaries/2024_june_raw.csv (дата звернення: 25.11.2024).

Короткий звіт за результатами перевірки кваліфікаційної роботи антиплагіатною інтернет-системою Strikeplagiarism:

StrikePlagiarism.com

Дата звіту 11/30/2024
Дата редагування ---

Звіт не був оцінений.

Звіт подібності

метадані

Заголовок
Коваленко_Притоманова_плагіат

Автор
Коваленко

Науковий керівник / Експерт
Притоманова

підрозділ
кафедра математичного моделювання та статистики

Тривога

У цьому розділі ви знайдете інформацію щодо текстових спотворень. Ці спотворення в тексті можуть говорити про МОЖЛИВІ маніпуляції в тексті. Спотворення в тексті можуть мати навмисний характер, але частіше характер технічних помилок при конвертації документа та його збереженні, тому ми рекомендуємо вам підходити до аналізу цього модуля відповідально. У разі виникнення запитань, просимо звертатися до нашої служби підтримки.

Заміна букв		0
Інтервали		0
Мікропробіли		3
Білі знаки		0
Парафрази (SmartMarks)	a	2

Обсяг знайдених подібностей

Коефіцієнт подібності визначає, який відсоток тексту по відношенню до загального обсягу тексту було знайдено в різних джерелах. Зверніть увагу, що високі значення коефіцієнта не автоматично означають плагіат. Звіт має аналізувати компетентна / уповноважена особа.

0.90%
0.90%
КП 1

0.27%
0.27%
КП 2

0.92%
0.92%
КЦ

25
Довжина фрази для коефіцієнта подібності 2

11063
Кількість слів

83675
Кількість символів