

РОЗВІДУВАЛЬНИЙ АНАЛІЗ ДАНИХ ДЛЯ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ АНАЛІЗУ ТА ПЕРЕДБАЧЕННЯ ЗАРОБІТНОЇ ПЛАТИ DATA DEVELOPERS МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Вінницький національний технічний університет

Анотація

Дослідження присвячене підготовці та розвідувальному аналізу даних для подальшого використання в інформаційній технології аналізу та передбачення заробітної плати Data Developers методами машинного навчання. Проведено аналіз структури датасету та його основних характеристик, а також досліджено можливі взаємозв'язки між рівнем оплати праці Data Developers та таких ключових чинників як професійний досвід, формат зайнятості, можливість віддаленої роботи та масштаб компанії.

Ключові слова: інформаційна технологія, машинне навчання, аналіз даних, передбачення, заробітна плата, розвідувальний аналіз даних, Data Developers.

Abstract

The study is devoted to the preparation and exploratory data analysis for further use in the information technology of analysis and prediction of Data Developers' salaries using machine learning methods. The dataset structure and its key features were analyzed, as well as investigates possible relationships between the level of remuneration of Data Developers and such key factors as professional experience, employment format, the possibility of remote work and the scale of the company.

Keywords: information technology, machine learning, data analysis, prediction, salary, exploratory data analysis, Data Developers.

Вступ

У сучасному світі, що швидко розвивається, роль фахівців з обробки даних (Data Developers) стає дедалі важливішою для успіху організацій у всіх галузях. Зростаючий попит на аналіз великих обсягів даних, розробку інтелектуальних систем та машинне навчання призводить до динамічних змін на ринку праці, що безпосередньо впливає на рівень компенсації та умови зайнятості цих висококваліфікованих спеціалістів.

У зв'язку з цим, виникає потреба у побудові інтелектуальних моделей, здатних виявляти закономірності та прогнозувати рівень заробітної плати на основі об'єктивних характеристик розробника даних. Методи машинного навчання дозволяють ефективно обробляти великі масиви даних, виконувати розвідувальний аналіз, виявляти найвпливовіші фактори та будувати точні прогностичні моделі.

Розробка інформаційної технології для аналізу та передбачення заробітної плати сприятиме підвищенню прозорості ринку праці, допоможе фахівцям краще оцінювати власний потенціал на ринку праці, а компаніям – формувати обґрунтовану політику оплати праці [1-3].

Отже, головною метою роботи є проведення розвідувального аналізу даних (EDA) для подальшого використання в інформаційній технології аналізу та передбачення заробітної плати Data Developers з метою виявлення ключових закономірностей у компенсації, зайнятості та географічному розподілі, з урахуванням впливу досвіду, посади, умов праці та інших факторів впливу на зарплату Data Developers.

Результати дослідження

Для проведення аналізу було обрано набір даних, що має назву «Data Developer Salary in 2024», опублікований автором «Shahzad Aslam» на платформі Kaggle у відкритому доступі [4]. Даний датасет містить 16534 записів і 11 атрибутів, охоплює такі ключові змінні, як рівень зарплати, назва посади, рівень досвіду, тип зайнятості, місце проживання працівника, коефіцієнт віддаленої роботи, розташування компанії та її розмір.

Цільовою змінною у цьому дослідженні є заробітна плата у доларах США (`salary_in_usd`), яка відображає загальний рівень компенсації фахівців з обробки даних. Ця змінна може залежати від багатьох факторів, таких як рівень досвіду (`experience_level`), тип зайнятості (`employment_type`), місце проживання (`employee_residence`), відсоток віддаленої роботи (`remote_ratio`), а також місце розташування компанії (`company_location`) і її розмір (`company_size`). Серед інших атрибутів, що можуть впливати на зарплату, важливе значення мають посада (`job_title`) та загальний контекст ринку, представлений у даному датасеті. На рисунку 1 наведено фрагмент датасету "Dataset salary in 2024", що включає ключові атрибути.

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	2024	SE	FT	AI Engineer	202730	USD	202730	US	0	US	M
1	2024	SE	FT	AI Engineer	92118	USD	92118	US	0	US	M
2	2024	SE	FT	Data Engineer	130500	USD	130500	US	0	US	M
3	2024	SE	FT	Data Engineer	96000	USD	96000	US	0	US	M
4	2024	SE	FT	Machine Learning Engineer	190000	USD	190000	US	0	US	M

Рис.1. Приклад атрибутів, що містить набір даних

Попередньо було проведено очищення датасету: усунуто пропущені значення, приведено назви стовпців до зручного формату, виконано базову перевірку наявності аномалій. Для виявлення викидів було застосовано IQR-метод [5]. Побудовано гістограму, яка ілюструє загальний розподіл цільової змінної – заробітної плати (`salary_in_usd`) та окремо виділяє викиди червоним кольором (рис. 2).

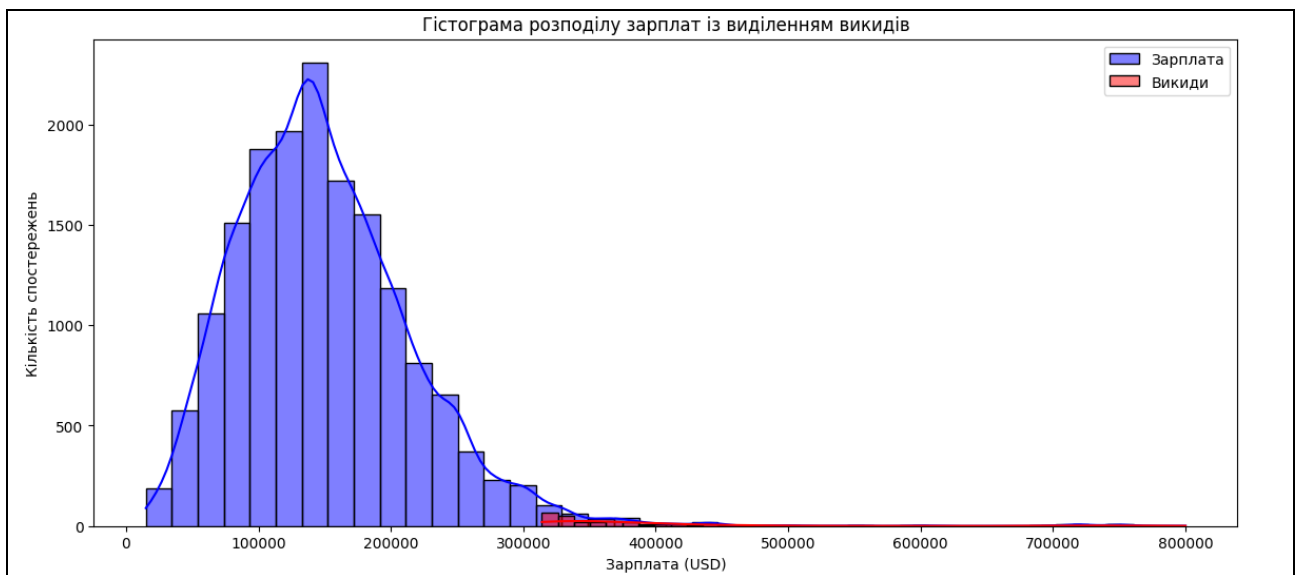


Рис.2. Гістограма розподілу зарплат із виділенням викидів (IQR-метод)

Отриманий графік показує, що більшість зарплат фахівців з обробки даних у 2024 році зосереджена в межах 60 000 – 130 000 доларів США, що свідчить про типовий рівень оплати праці в цій сфері. Розподіл є асиметричним із подовженим правим «хвостом», що вказує на наявність обмеженої кількості надзвичайно високих зарплат. Виявлені викиди, зокрема значення понад 300 000 доларів, можуть бути пов'язані з керівними посадами або роботою у великих міжнародних компаніях. Цей графік дозволяє не лише побачити загальну структуру зарплат, а й виокремити крайні значення, які впливають на середні показники й потребують додаткового аналізу.

Ще одним важливим аспектом попереднього аналізу є оцінка розподілу спостережень за роками працевлаштування зображено на рисунку 3 [6-8].

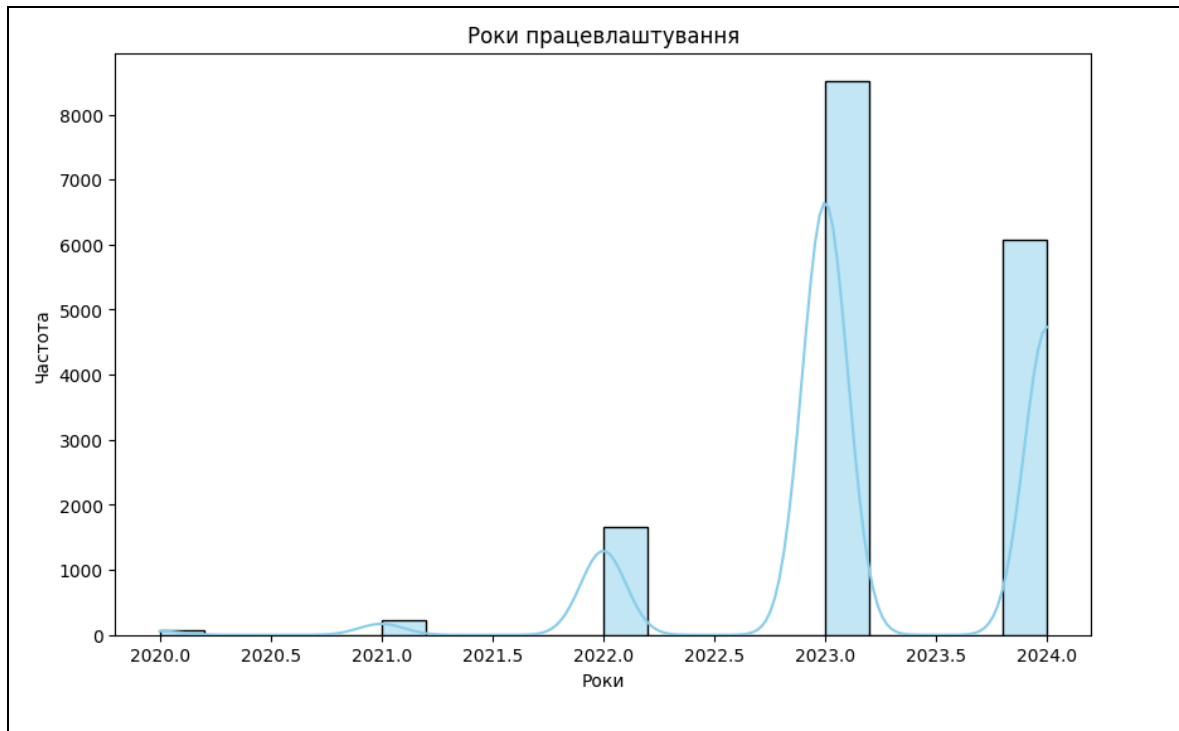


Рис.3. Гістограма частоти за роками працевлаштування

На графіку чітко видно, що найбільша кількість записів припадає на 2023 рік, за яким іде 2024. Значно менше даних спостерігається за 2020–2022 роки. Це може свідчити про активізацію збору даних або збільшення попиту на фахівців із Data Development саме у 2023–2024 роках. Такий перекис розподілу може впливати на аналіз змін у зарплатах з часом і має враховуватись при побудові моделей прогнозування або порівнянні тенденцій.

Побудовано графіки впливу рівня досвіду (experience_level) та типу зайнятості (employment_type) для більш детального аналізу на заробітню плату (рис. 4, 5).

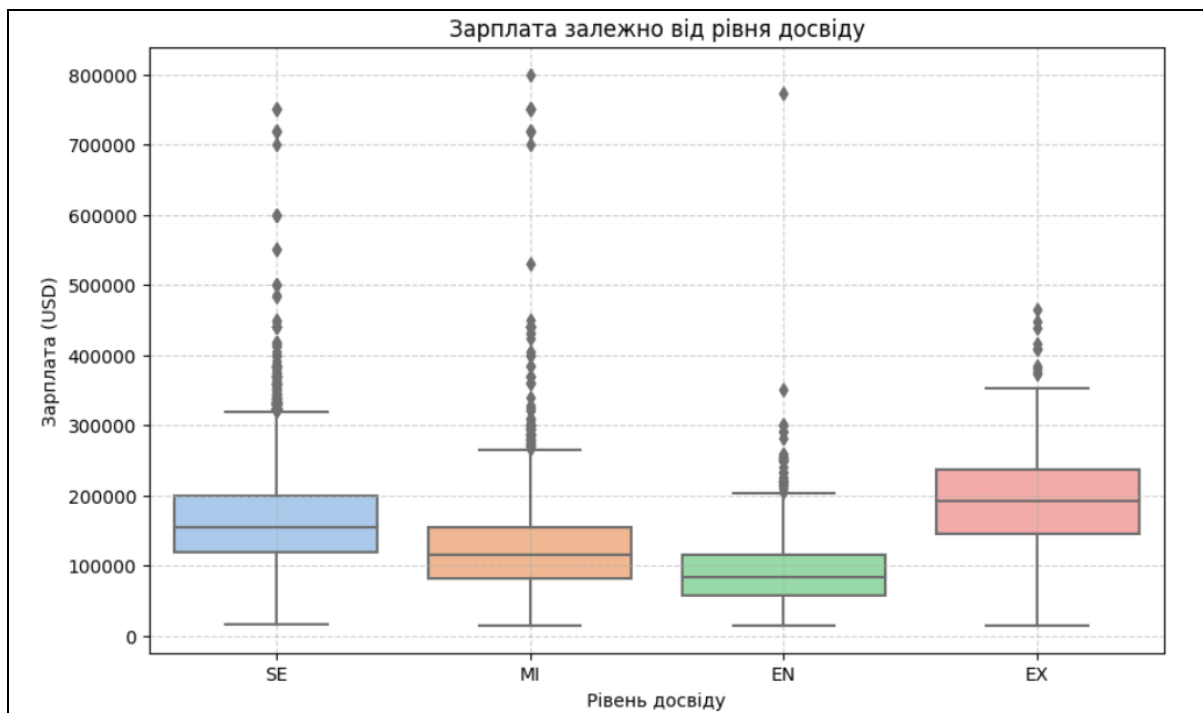


Рис.4. Зарплата залежно від рівня досвіду

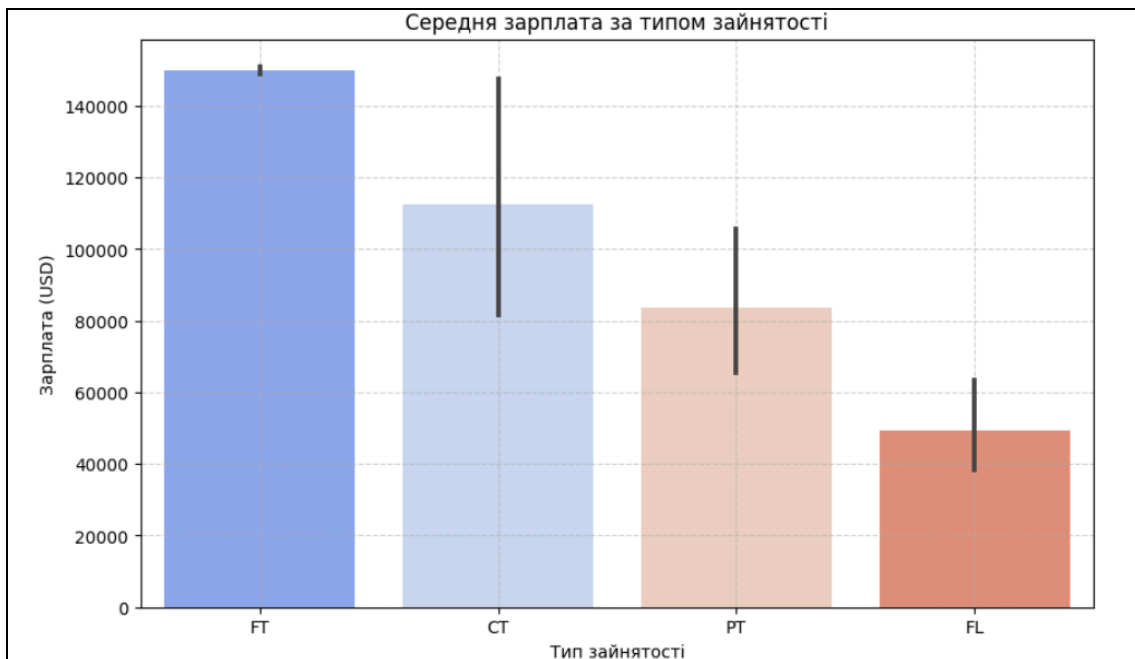


Рис.5. Середня зарплата для різних типів зайнятості

Ці графіки ілюструють ключові аспекти заробітної плати фахівців у сфері даних. Графік, представлений на рисунку 4 візуалізує розподіл зарплат для кожного рівня досвіду – Entry-level (EN), Mid-level (MI), Senior (SE) та Executive (EX) — показуючи, що зі зростанням рівня досвіду медіанна зарплата також зростає. Графік представлений на рисунку 5, у свою чергу, висвітлює вплив форми зайнятості: повна зайнятість (FT) забезпечує найвищу середню заробітну плату, тоді як контрактики (CT) отримують другий за величиною дохід, але з більшою варіативністю. Часткова зайнятість (PT) та фриланс (FL) показують значно нижчі середні зарплати. Тепер доречно перевірити вплив віддаленої роботи на зарплату (рис. 6).

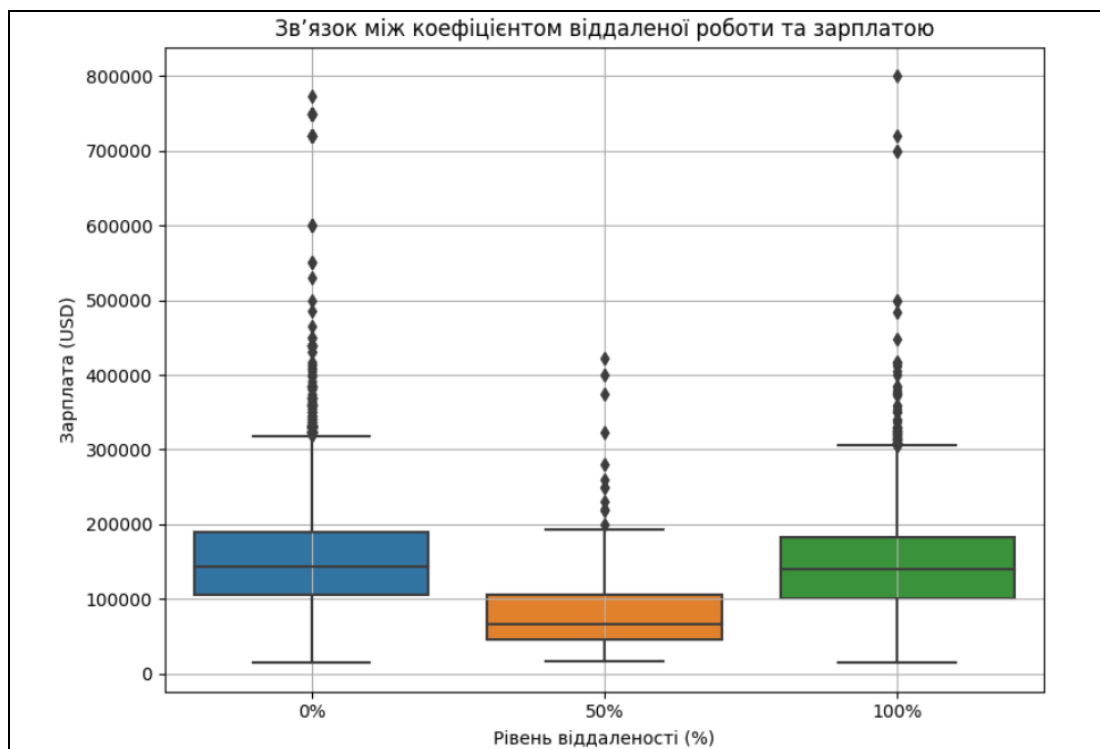


Рис.6. Зв'язок між коефіцієнтом віддаленої роботи та зарплатою

Цей графік показує, що найвищі зарплати спостерігаються у фахівців з 0% (повністю в офісі) та 100% (повністю віддалено) рівнем віддаленості, тоді як для 50% (гібридного) режиму роботи медіанна зарплата є помітно нижчою. При цьому, найбільший розкид зарплат та наявність значних викидів (високооплачуваних фахівців) присутні як для 0%, так і для 100% віддаленості.

Тепер дізнаємось чи впливає розмір компаній на дохід, зображено на рисунку 7.

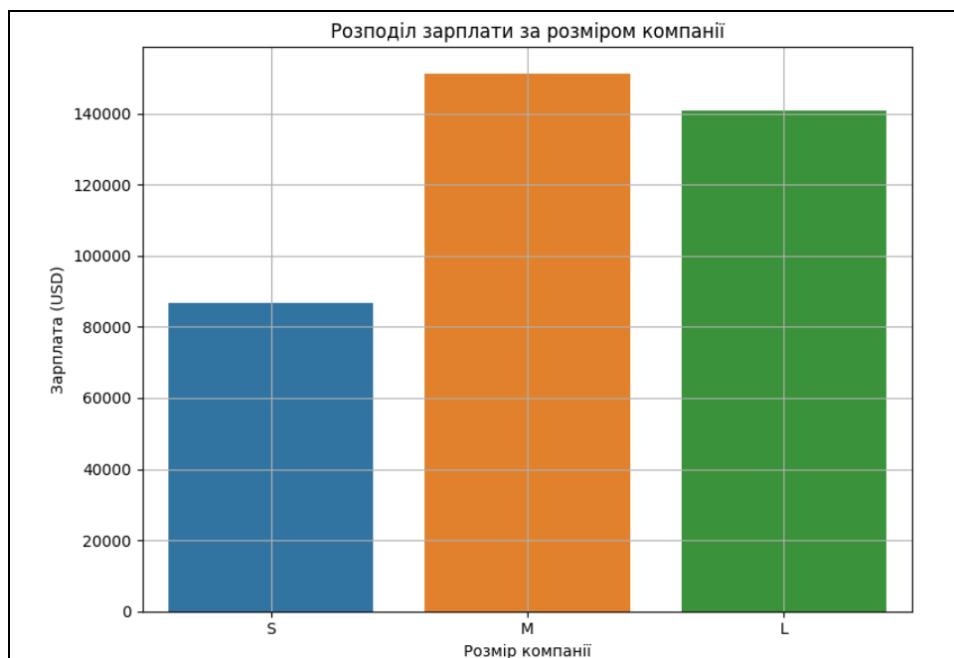


Рис.7. Розподіл зарплати за розміром компанії

Графік демонструє, що компанії середнього (M) та великого (L) розміру пропонують значно вищу середню заробітну плату порівняно з невеликими компаніями (S). Це свідчить про те, що розмір компанії є важливим фактором, що впливає на рівень доходу фахівців у сфері даних.

Дане дослідження також зосереджено на аналізі середньої заробітної плати за різними посадами в галузі даних, з метою виявлення найбільш прибуткових спеціалізацій та тенденцій на ринку. Представлений графік (рис.8) візуалізує топ-10 посад за середньою зарплатою, надаючи цінний огляд поточних ринкових умов станом на кінець 2024.

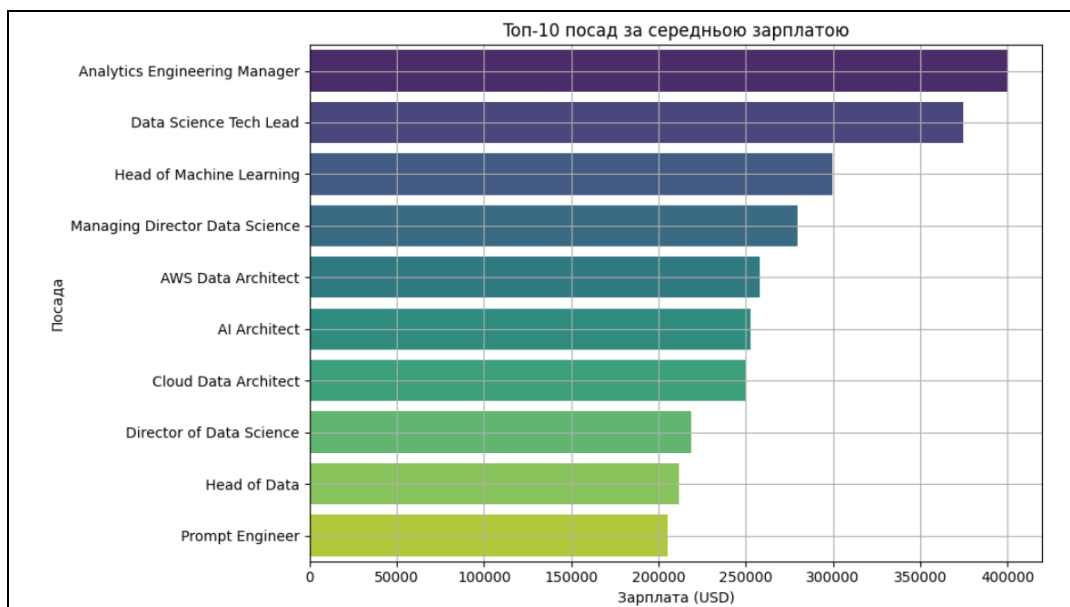


Рис.8. Топ-10 посад за середньою зарплатою

Аналіз графіка чітко демонструє, що лідируючі позиції за рівнем доходу займають ролі, які вимагають високого рівня досвіду, спеціалізації та управлінських навичок. Зокрема, такі посади, як "Analytics Engineering Manager", "Data Science Tech Lead" та "Head of Machine Learning", знаходяться на вершині списку, що підкреслює високу цінність керівних та архітектурних ролей у сфері даних. Також помітно, що інженерні та архітектурні позиції, що пов'язані з хмарними технологіями ("AWS Data Architect", "Cloud Data Architect") та штучним інтелектом ("AI Architect"), мають високий рівень зарплати. Наявність "Prompt Engineer" серед топ-посад свідчить про зростаючу важливість спеціалізацій, пов'язаних із новітніми технологіями, такими як генеративний ШІ. Це дослідження підтверджує, що для досягнення найвищих зарплат у сфері даних необхідна глибока експертиза та фокус на стратегічних та управлінських аспектах.

На основі попереднього аналізу було побудовано кореляційну матрицю для аналізу взаємозв'язків між зарплатою та іншими змінними (рис. 9) [9-10].

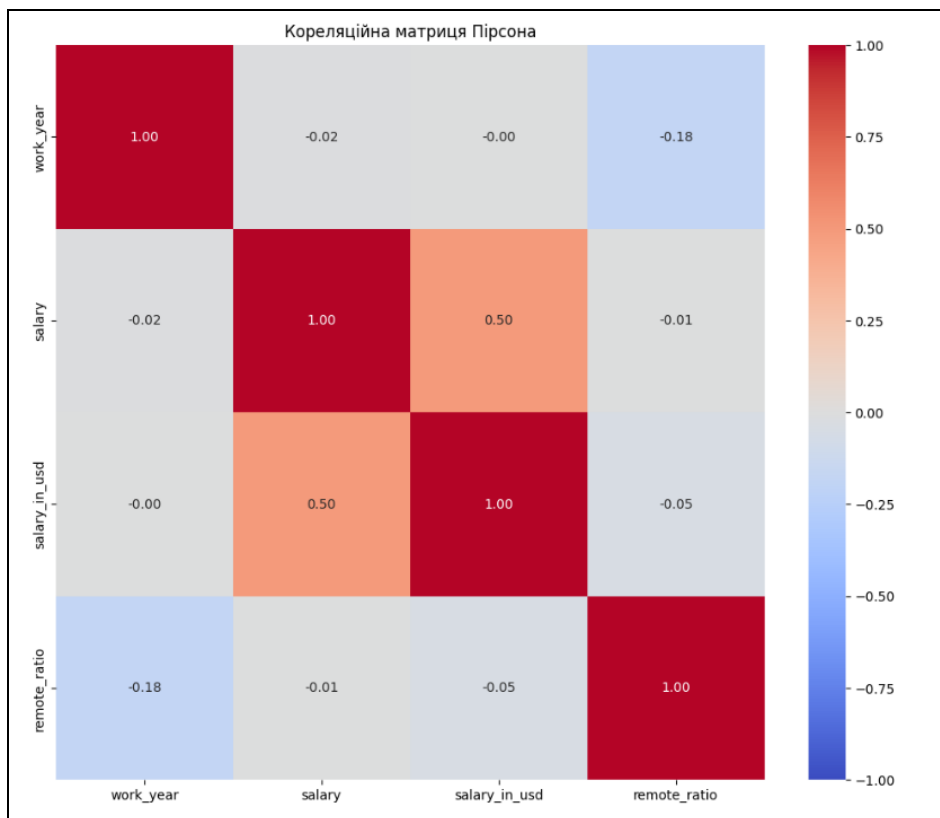


Рис.9. Матриця кореляції

Кореляційна матриця показує, що "salary" та "salary_in_usd" ідеально позитивно корелюють (1.00), що є очікуваним, оскільки вони, ймовірно, представляють одне й те ж значення або пряму конвертацію. Обидві змінні мають помірну позитивну кореляцію з "work_year" (0.50), що вказує на тенденцію зростання зарплати з досвідом.

Натомість, "remote_ratio" демонструє дуже слабку або майже відсутню негативну лінійну кореляцію з "work_year" (-0.18), "salary" (-0.01) та "salary_in_usd" (-0.05), що свідчить про незначний зв'язок між відсотком віддаленої роботи та іншими показниками.

Отримані результати є важливими для подальшого прогнозування та моделювання на ринку праці, оскільки вони підтверджують вплив рівня досвіду, типу зайнятості, розміру компанії та коефіцієнта віддаленої роботи на рівень заробітної плати та можуть бути використані для створення прогнозних моделей компенсації та оптимізації стратегій найму.

Висновки

У ході розвідувального аналізу даних було досліджено фактори, що впливають на рівень заробітних плат Data Developers, на основі відкритого датасету «Data Developer Salary in 2024». Аналіз показав, що ключовими чинниками, які впливають на рівень оплати праці, є професійний досвід, формат зайнятості, можливість віддаленої роботи та масштаб компанії.

Попередня обробка даних включала очищення датасету, усунення пропущених значень та перевірку наявності аномалій. Для виявлення викидів застосовано IQR-метод. Аналіз розподілу цільової змінної "salary_in_usd" показав, що більшість зарплат зосереджена в межах 60 000 – 130 000 доларів США. Розподіл є асиметричним з подовженим правим «хвостом», що вказує на наявність обмеженої кількості надзвичайно високих зарплат, які можуть бути пов'язані з керівними посадами або роботою у великих міжнародних компаніях. Також виявлено, що найбільша кількість записів у датасеті припадає на 2023 та 2024 роки, що може свідчити про активізацію збору даних або збільшення попиту на фахівців із Data Development у ці періоди.

Виявлено, що із зростанням досвіду та переходом на керівні посади медіанна зарплата значно зростає, а також, що повна зайнятість забезпечує вищі доходи порівняно з частковою зайнятістю чи фрилансом. Графіки також продемонстрували, що компанії середнього та великого розміру пропонують більш конкурентоспроможні зарплати. Дослідження підкреслило зростаючу роль спеціалізацій, пов'язаних з генеративним ШІ та хмарними технологіями, які забезпечують високий рівень доходів.

Отримані результати можуть бути корисними для розробки інформаційних систем підтримки прийняття рішень у сфері працевлаштування, а також для побудови прогнозних моделей заробітної плати.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Ghorai, P., Barik, R. Salary Prediction Using Machine Learning Techniques. In: Tiwari, S., Trivedi, M.C., Kolhe, M.L., Singh, B.K. (eds) *Advances in Data and Information Sciences. ICDIS 2024. Lecture Notes in Networks and Systems*, 2025, vol 1193. Springer, Singapore. https://doi.org/10.1007/978-981-97-9619-9_17
2. Rukiye Kaya, Mehtap Saatçi, Mehmet Gökhan Bakal. "Improving Salary Offer Processes With Classification Based Machine Learning Models", *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)*, 2024, pp.1-7, doi: 10.1109/IDAP64064.2024.10710706
3. A. Asaduzzaman, M. R. Uddin, Y. Woldeyes and F. N. Sibai, "A Novel Salary Prediction System Using Machine Learning Techniques," *2024 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*, Chiang-mai, Thailand, 2024, pp. 38-43, doi: 10.1109/ECTIDAMTNCN60518.2024.10480058
4. Data Developer Salary in 2024 [Електронний ресурс]. – Режим доступу: <https://www.kaggle.com/datasets/zeesolver/data-eng-salary-2024>
5. Міжквартильний або інтерквартильний розмах (IQR) [Електронний ресурс]. – Режим доступу: <https://uk.economy-pedia.com/11035383-interquartile-range>
6. Pandas Getting started. 2024 [Електронний ресурс] – Режим доступу: https://pandas.pydata.org/docs/getting_started/index.html
7. Seaborn Tutorial. 2023 [Електронний ресурс]. – Режим доступу: <https://seaborn.pydata.org/tutorial.html>
8. Matplotlib Pyplot Documentation. 2024 [Електронний ресурс]. – Режим доступу: https://matplotlib.org/3.5.3/api/as_gen/matplotlib.pyplot.html
9. Мокін В.Б., Дратовий М.В. Наука про дані: машинне навчання та інтелектуальний аналіз даних. Навчальний посібник. м. Вінниця, 2024 р. 263 с. URL: <https://iq.vntu.edu.ua/repository/getfile.php/8163.pdf>
10. A Complete Guide to Data Visualization in Python With Libraries & More. 2023 [Електронний ресурс]. – Режим доступу: <https://www.simplilearn.com/tutorials/python-tutorial/data-visualization-in-python>

Старинець Тетяна Віталіївна – студентка групи 2ICT-226, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м. Вінниця, e-mail: tetanastarynets@gmail.com.

Войцеховська Ольга Олександрівна – PhD, старший викладач кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, м. Вінниця, e-mail: olgav1085@gmail.com.

Starynets Tetiana V. – student of group 2IST-22b, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: tetanastarynets@gmail.com.

Voitsekhovska Olha O. – PhD, Senior Lecturer of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: olgav1085@gmail.com.

