**Міністерство освіти і науки України**

**Національний технічний університет України «КПІ» імені Ігоря Сікорського Кафедра обчислювальної техніки ФІОТ**

**ЗВІТ**

**з лабораторної роботи №7**

**з навчальної дисципліни «Технології Data Science»**

**Тема:**

**РОЗРОБКА ПРОГРАМНОГО МОДУЛЯ ПРОГНОЗУВАННЯ ДИНАМІКИ ЗМІНИ**

**ПОКАЗНИКІВ ЕФЕКТИВНОСТІ ТОРГІВЕЛЬНИХ КОМПАНІЙ**

**Виконав:**

Студент 4 курсу кафедри ФІОТ,

Навчальної групи ІП-11

Олександр Головня

**Перевірив:**

Професор кафедри ОТ ФІОТ Олексій Писарчук

**Київ 2024**

**І. Мета:**

дослідити виявити та узагальнити особливості реалізації проектного практикуму в галузі аналізу часових (стохастичних рядів), як характеристика показників ефективності діяльності торгівельних компаній.

**ІІ. Завдання:**

Розробити програмний скрипт мовою Python, що реалізує функціонал за обраним

рівням складності:

**ІІ рівень складності 9 балів.**

Відповідно до технічних умов, табл.2 додатку.





**ІІІ. Результати виконання лабораторної роботи.**

Блок схема алгоритму:



Рис.1 – Блок-схема алгоритму програми

* 1. **Опис структури проекту програми.**

Для реалізації розробленого алгоритму мовою програмування Python з використанням можливостей інтегрованого середовища сформовано проект.

Проект базується на лінійній бізнес-логіці функціонального програмування та має таку структуру.

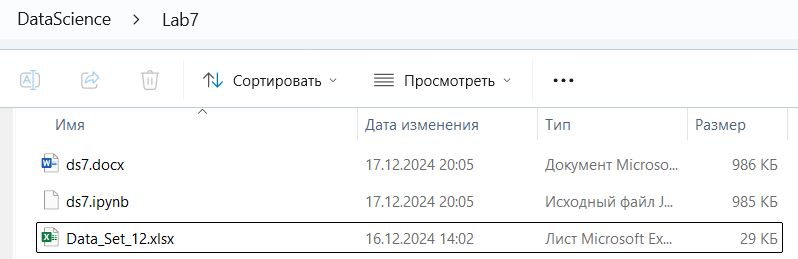


Рис.2 – Структура проєкту

ds7.ipynb – файл програмного коду лабораторної роботи;

ds7.docx – файл звіту лабораторної роботи

Data\_Set\_12.xlsx – dataset

* 1. **Результати роботи програми відповідно до завдання.**

Спочатку я зчитав дані з відповідного xlsx файлу в датайрейм за допомогою бібліотеки pandas.

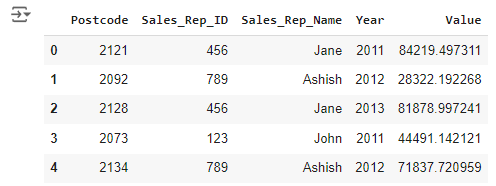


Рис 3.1 – Зчитування даних у датафрейм

**Перейменування колонок**:

Колонки перейменовано для зручності: Sales\_Rep\_Name → Sales\_Representative, Sales\_Rep\_ID → Sales\_Representative\_ID

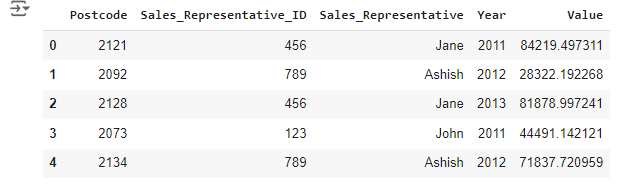


Рис. 3.2 – Перейменування стовпців

**Перевірка на дублі**:

* Підрахунок кількості дублікатів за допомогою duplicated().

Потім я переглянув розмірність датасету, перевірив його на наявність дублікатів та вивів інформацію. по кількості значень (не пропущених) та типу даних кожного стовпця.

**Перевірка типів даних**:

* info() дозволяє переглянути типи даних та пропуски.

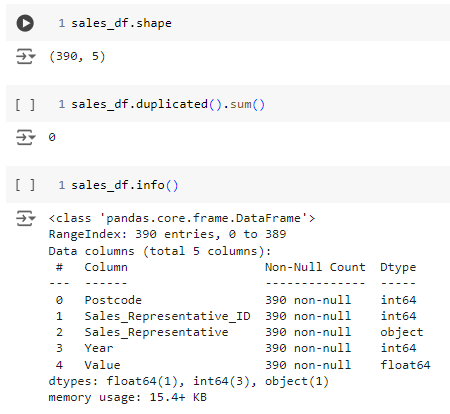


Рис. 3.3 – Інформація про датасет

Можна побачити, що дублікати відсутні, як і пропущені значення, адже усі стовпці містять 390 непропущених значень, а всього рядків у датасеті також 390. Крім того, я переглянув значення описових статистик для стовпця, що містить продажі у доларах

**Аналіз статистики значень продажів**:

* describe() відображає базову статистику для стовпця Value.

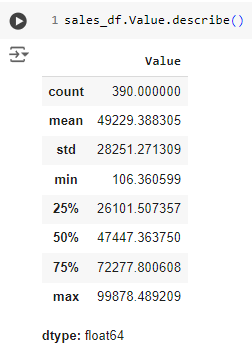


Рис. 3.4 – Значення описових статистик для стовпця Value

Як бачимо, середнє значення не сильно відрізняється від медіани. Також варто зазначити, що мінімальне та максимальне значення відрізняються від середнього менше, ніж на 2 середньоквадратичних відхилення, що може свідчити, що як таких викидів немає., проте також варто поглянути на гістрограму розподілу та діаграму розмаху.

***Візуалізація розподілу продажів***

**Гістограма значень**:

* Побудовано гістограму для стовпця Value з оптимальною кількістю бінів за допомогою формули Стерджеса.

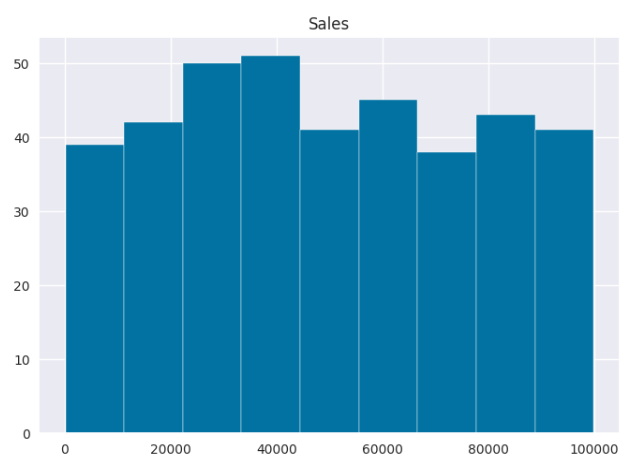


Рис. 3.5 – Гістограма розподілу продажів

Розподіл продажів не дуже схожий на нормальний візуально та є дещо асиметричним в ліву сторону

**Боксплот (Boxplot)**:

* Графік для виявлення викидів у стовпці Value

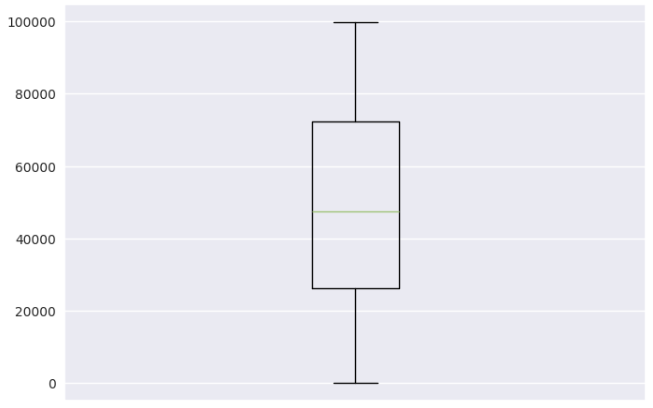


Рис. 3.6 – Діаграма розмаху продажів

На діаграмі розмаху можна чітко побачити, що викидів немає. Тепер поглянемо на лінійний графік продажів та гістограму розподілу одночасно.

**Функція для аналізу продажів за роками**

* **plot\_sales\_by\_year** будує два графіки:

Лінійний графік продажів.

Гістограму значень продажів.

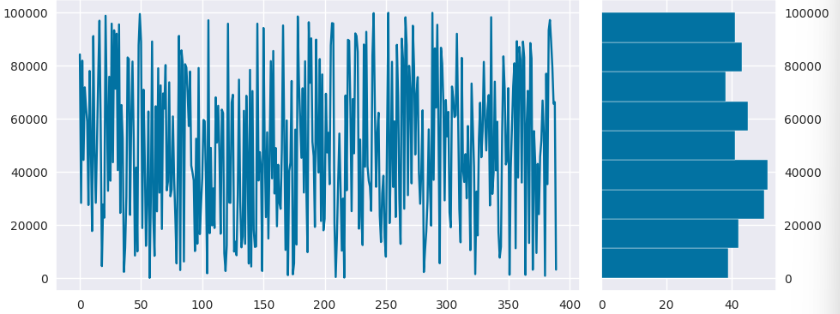


Рис. 3.7 – Лінійний графік та гістограма розподілу продажів

Аналогічно якихось аномальних вимірів не виявлено

**Аналіз продажів за кожен рік**

1. Виклик функції plot\_sales\_by\_year() для кожного року: **2011**, **2012**, **2013**.
2. Підсумок статистики продажів за кожен рік за допомогою describe().

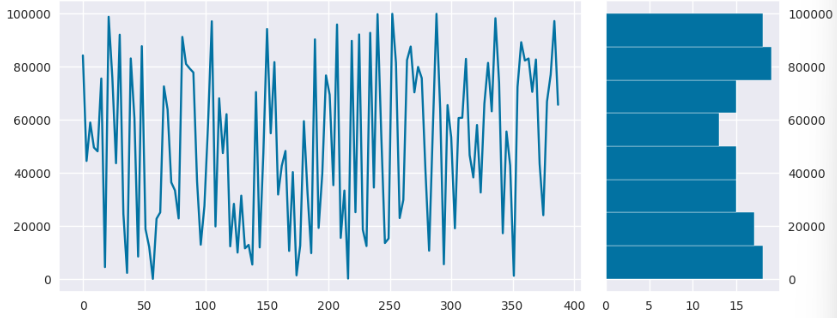
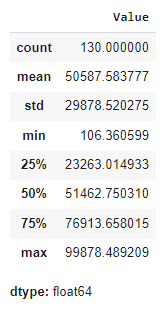
 

Рис. 3.8 – Лінійний графік та гістограма розподілу продажів за 2011 рік та його статистичні характеристики продажів

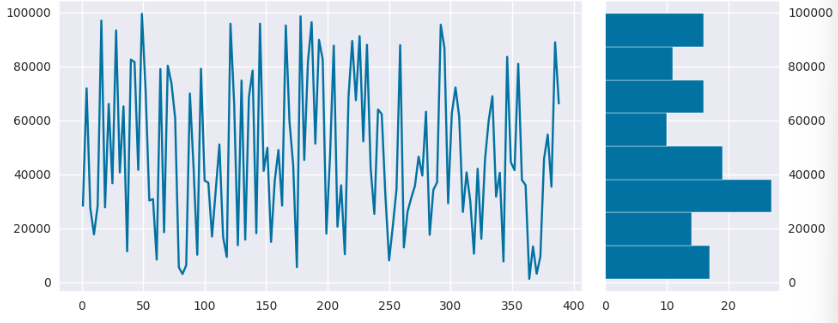
 

Рис. 3.9 – Лінійний графік та гістограма розподілу продажів за 2012 рік та його статистичні характеристики продажів

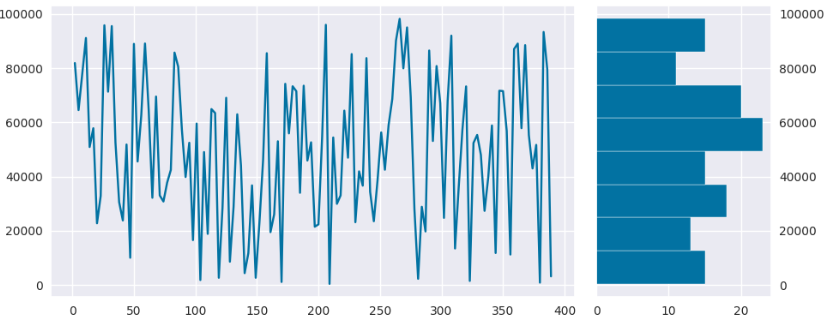
 

Рис. 3.10 – Лінійний графік та гістограма розподілу продажів за 2013 рік та його статистичні характеристики продажів

Як бачимо, датасет містить однакову кількість продажів за кожен з років. Крім того, найбільше середнє значення продаж були у 2011 році, у 2012 найменше, а у 2013 дещо менше, ніж у 2011, але все ж більше за 2012. Спираючись на розподіли по рокам, можна побачити, що 2011 рік мав найбільш близький до рівномірного розподіл та мав найбільшу кількість дорогих продаж, що впливає на тренд по рокам, зображений нижче. Наступним кроком я дослідив динаміку зміни продажів за роки

**Загальна динаміка продажів за роками**

1. Групування даних за роками (groupby) для обчислення **суми продажів**.
2. Побудова лінійного графіку суми продажів.

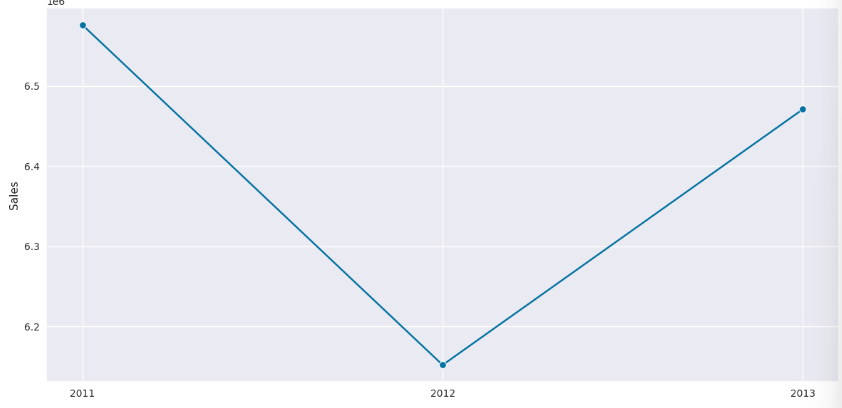


Рис. 3.15 – Динаміка сумарного значення продаж за роки Можна побачити, що в 2012-му році було суттєве зниження сумарного значення продаж у доларах, а у 2013 вже знову відбувся ріст.

**Середні продажі по роках і представниках**

1. **Групування за роками та продавцями**:

Обчислено середні значення продажів.

1. **Візуалізація**:

Лінійний графік середніх продажів.

Стовпчастий графік середніх продажів за роками.



Рис. 3.16 – Середні значення продаж за кожний рік для кожного торгового представника

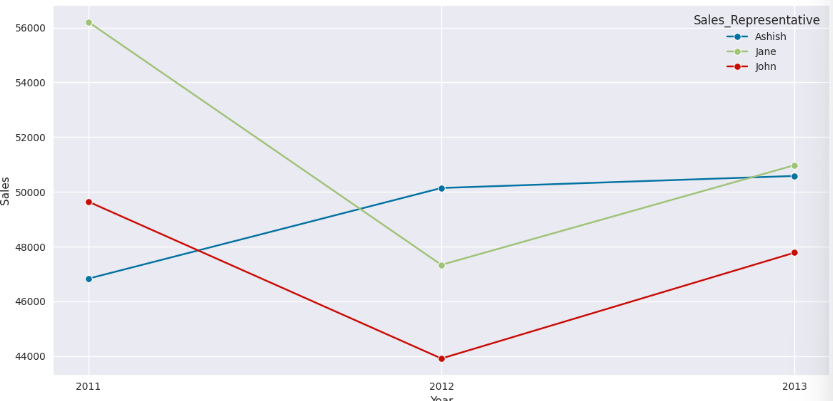


Рис. 3.17 – Динаміка зміни середніх значень продаж для кожного торгового представника

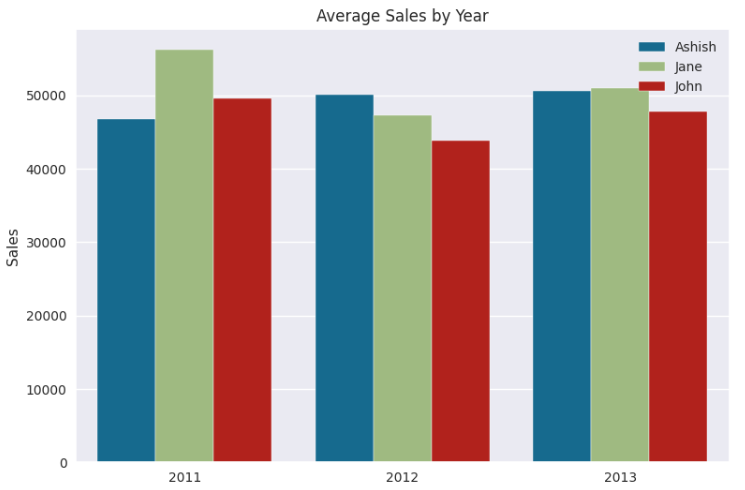


Рис. 3.18 – Стовпчаста діаграма середніх значень продаж для кожного торгового представника за кожний рік

Як на графіку, так і на стовпчастій діаграмі видно, що у 2011 Jane була найбільш ефективною серед усіх інших торгових представників. У 2013 John показав дещо гірші результати за інших торгових представників, в той час як інші два представники показали приблизно схожі результати. 2012 рік виявися роком, коли Ashish мав найкраще середнє значення продаж. Наступним кроком я вирішив кластеризувати значення продаж та дослідити кількість продаж кожного торгового представника по кожному з кластерів.

**Аналіз розподілу продажів**

1. **Розкид продажів**:

Побудовано діаграму розсіювання.

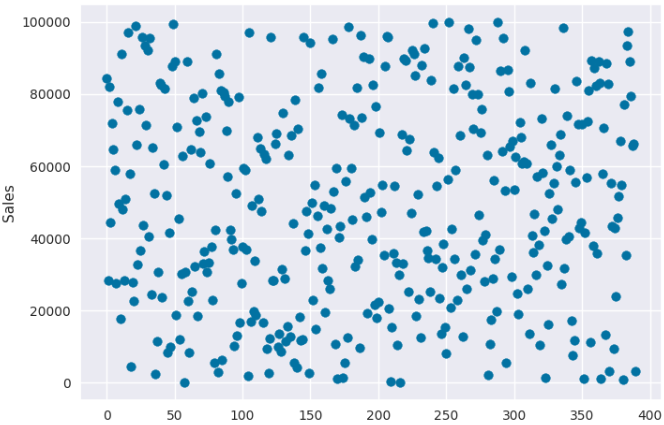


Рис. 3.19 – Діаграма розсіювання

Для кластеризації я використав алгоритм K-Means, для якого треба явно задавати кількість кластерів, тому я підібрав оптимальну кількість кластерів з використанням методу ліктя.

**Кластеризація продажів**

1. **"Elbow method"**:

Використано KElbowVisualizer для визначення оптимальної кількості кластерів (2–11).

1. **KMeans**:

Кластеризація значень продажів на **4 кластери**.

Додавання нового стовпця Cluster до sales\_df.

1. **Візуалізація кластерів**:

Діаграма розсіювання зі значеннями продажів та кластерами.

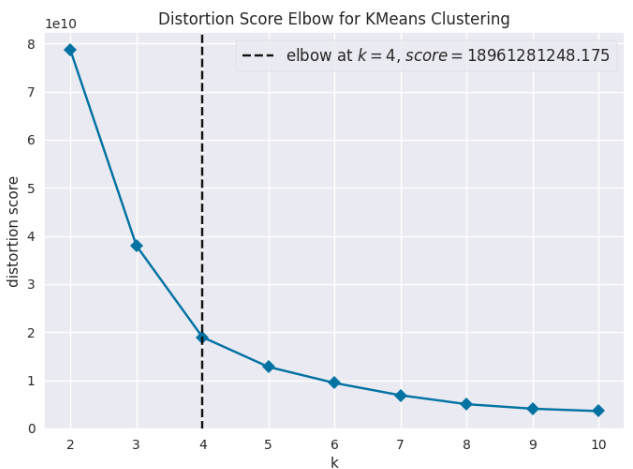


Рис. 3.20 – Підбір оптимальної кількості кластерів

Використовуючи цей метод, я отримав, що 4 кластери – це оптимальне значення, тому я кластеризував продажі алгоритмом K-Means, задавши K рівне 4. Отримані кластери я відсортував по середньому значенню продаж та візуалізував ці кластери.

**Аналіз кластерів**

1. **Підрахунок кількості продажів за кластерами і продавцями**:

Використано groupby() та побудовано стовпчастий графік.

1. **Підрахунок продажів по кластерах і роках**:

Додаткове групування для детальнішого аналізу.

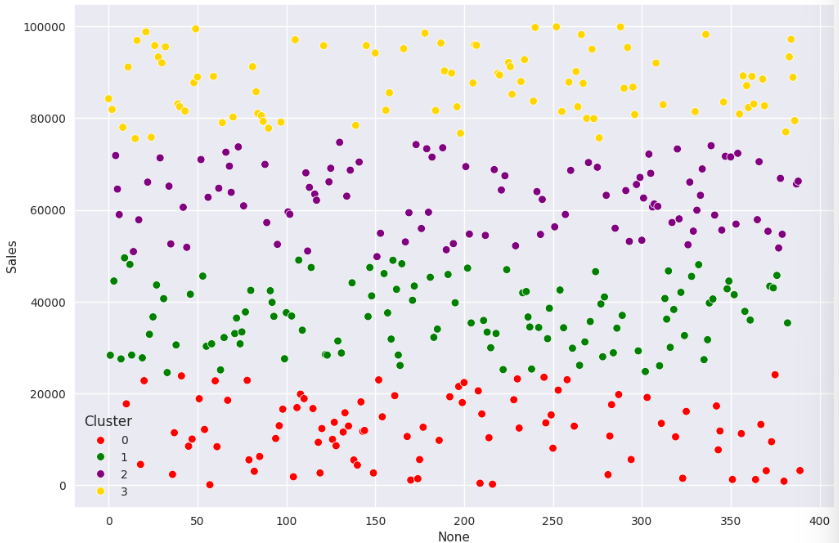


Рис. 3.21 – Візуалізація кластерів

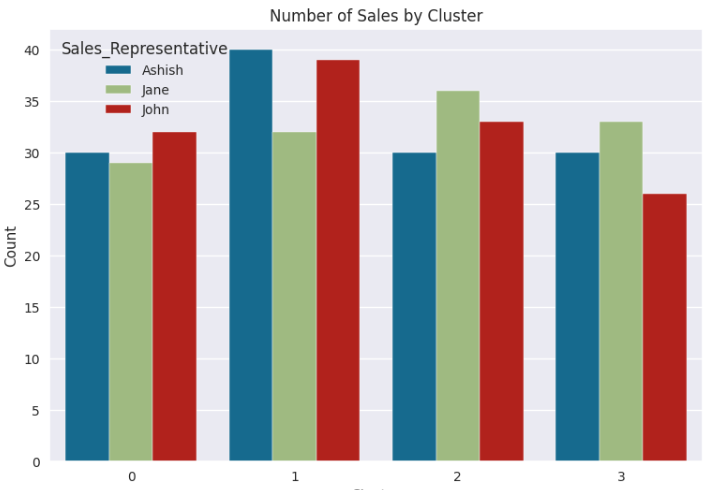


Рис. 3.23 – Стовпчаста діаграма кількості продажів у кожного торгового представника по різним кластерам

**Підсумок**

1. Дані були очищені та підготовлені.
2. Проведено дослідження динаміки продажів у часі.
3. Візуалізовано середні продажі по роках і представниках.
4. Виконано кластеризацію для визначення груп за рівнем продажів.
5. Зроблено висновки щодо розподілу продажів за представниками, роками та кластерами.
   1. **Аналіз результатів відлагодження та верифікації результатів роботи програми.**

Результати відлагодження та тестування довели працездатність розробленого коду. Це підтверджується результатами розрахунків, які не суперечать теоретичним положенням.

Верифікація функціоналу програмного коду, порівняння отриманих результатів з технічними умовами завдання на лабораторну роботу доводять, що усі завдання виконані у повному обсязі.

**IV. Висновки.**

Отже, в ході даної лабораторної я у результаті виконання програми узагальнив особливості реалізації проектного практикуму в галузі аналізу часових рядів та аналізу даних загалом для дослідження показників ефективності діяльності торгівельних компаній. Зокрема, я дослідив динаміку продаж за роки загалом та по кожному торговому представнику. Крім того, за допомогою кластеризації я виявив найбільш часті продажі (інтервал значень, якому вони належать) та дослідив кількості продаж у кожному кластері, здійснені кожним з торгових представників.