

**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 «Вступ до технології Data Science»**

Тема:

**РОЗРОБКА ПРОГРАМНОГО МОДУЛЯ ПРОГНОЗУВАННЯ ДИНАМІКИ ЗМІНИ
ПОКАЗНИКІВ ЕФЕКТИВНОСТІ ТОРГІВЕЛЬНИХ КОМПАНІЙ
(міні проекти в галузі аналізу даних для завдань електронної комерції)**

Виконав:

Студент 3 курсу кафедри ІІІ ФІОТ,
Навчальної групи ІІІ-11
Сідак К.І.

Перевірив:

Професор кафедри ОТ ФІОТ
Писарчук О.О.

Київ 2023

I. Мета роботи:

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

II. Завдання:

Розробити програмний скрипт мовою Python, що реалізує функціонал за обраним рівнем складності:

II рівень складності 9 балів.

Відповідно до технічних умов, табл.2 додатку.

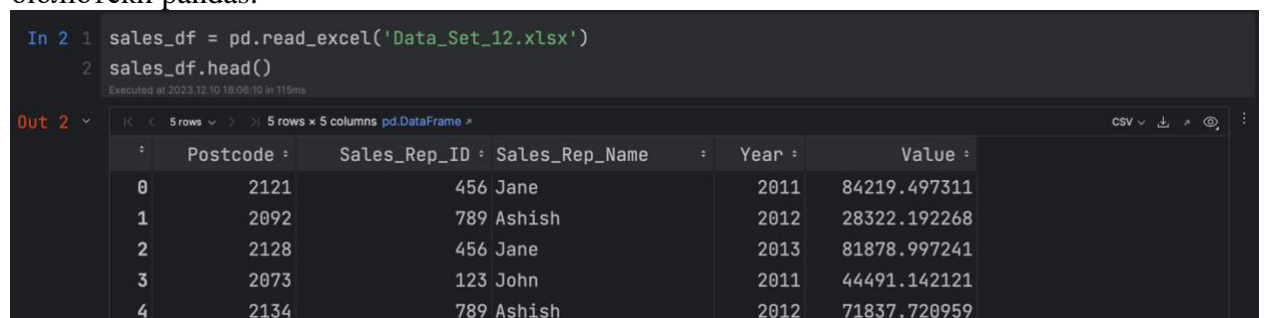
12. Розробити програмний скрипт, що реалізує аналіз даних, поданих у файлі Data_Set_12.xlsx.

Розробити програмний скрипт, що реалізує аналіз даних, самостійно обраних процесів. Обов'язковою вимогою є аналіз множини процесів, поданих часовими рядами із різними властивостями.

III. Виконання лабораторної роботи.

3.1. Зчитування даних та їх нормалізація

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



```
In 2 1 sales_df = pd.read_excel('Data_Set_12.xlsx')
2 sales_df.head()
```

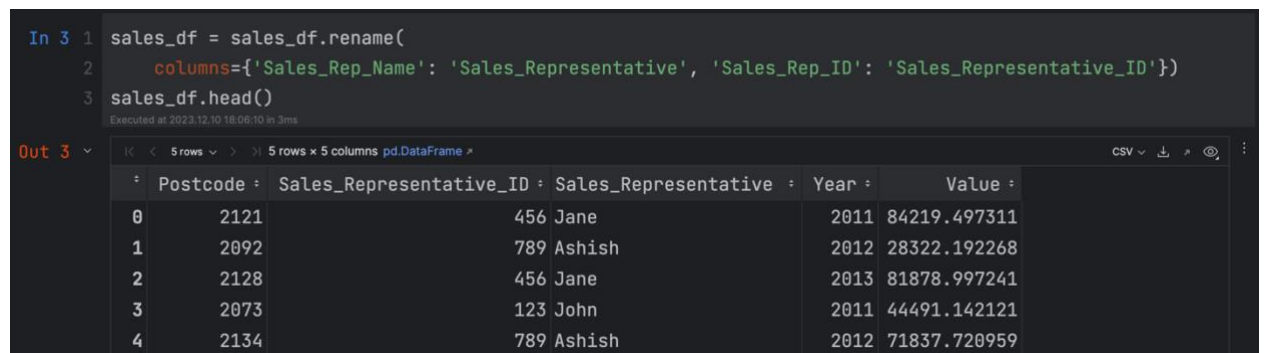
Executed at 2023.12.10 18:06:10 in 115ms

Out 2 5 rows x 5 columns pd.DataFrame

	Postcode	Sales_Rep_ID	Sales_Rep_Name	Year	Value
0	2121	456	Jane	2011	84219.497311
1	2092	789	Ashish	2012	28322.192268
2	2128	456	Jane	2013	81878.997241
3	2073	123	John	2011	44491.142121
4	2134	789	Ashish	2012	71837.720959

Рис 3.1 – Зчитування даних у датафрейм

Наступним кроком я перейменував деякі стовпці на більш змістовні назви, а саме стовпець, що містить ім'я торгового представника, та стовпець, що містить ID цих представників.



```
In 3 1 sales_df = sales_df.rename(
2     columns={'Sales_Rep_Name': 'Sales_Representative', 'Sales_Rep_ID': 'Sales_Representative_ID'})
3 sales_df.head()
```

Executed at 2023.12.10 18:06:10 in 3ms

Out 3 5 rows x 5 columns pd.DataFrame

	Postcode	Sales_Representative_ID	Sales_Representative	Year	Value
0	2121	456	Jane	2011	84219.497311
1	2092	789	Ashish	2012	28322.192268
2	2128	456	Jane	2013	81878.997241
3	2073	123	John	2011	44491.142121
4	2134	789	Ashish	2012	71837.720959

Рис. 3.2 – Перейменування стовпців

Потім я переглянув розмірність датасету, перевірів його на наявність дублікатів та вивів інформацію про кількість значень (не пропущених) та типу даних кожного стовпця.

```
In 4 1 sales_df.shape
      Executed at 2023.12.10 18:06:10 in 2ms

Out 4 (390, 5)

In 5 1 sales_df.duplicated().sum()
      Executed at 2023.12.10 18:06:10 in 4ms

Out 5 0

In 6 1 sales_df.info()
      Executed at 2023.12.10 18:06:11 in 2ms

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 390 entries, 0 to 389
Data columns (total 5 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Postcode                             390 non-null    int64
1   Sales_Representative_ID              390 non-null    int64
2   Sales_Representative                 390 non-null    object
3   Year                                 390 non-null    int64
4   Value                                390 non-null    float64
dtypes: float64(1), int64(3), object(1)
memory usage: 15.4+ KB
```

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.

In 7 1	sales_df.Value.describe()																		
	Executed at 2023.12.10 18:06:11 in 4ms																		
Out 7	<div> <div>8 rows</div> <div>Length: 8, dtype: float64 pd.Series</div> </div> <table> <tr> <th></th><th>Value</th></tr> <tr> <td>count</td><td>390.000000</td></tr> <tr> <td>mean</td><td>49229.388305</td></tr> <tr> <td>std</td><td>28251.271309</td></tr> <tr> <td>min</td><td>106.360599</td></tr> <tr> <td>25%</td><td>26101.507357</td></tr> <tr> <td>50%</td><td>47447.363750</td></tr> <tr> <td>75%</td><td>72277.800608</td></tr> <tr> <td>max</td><td>99878.489209</td></tr> </table>		Value	count	390.000000	mean	49229.388305	std	28251.271309	min	106.360599	25%	26101.507357	50%	47447.363750	75%	72277.800608	max	99878.489209
	Value																		
count	390.000000																		
mean	49229.388305																		
std	28251.271309																		
min	106.360599																		
25%	26101.507357																		
50%	47447.363750																		
75%	72277.800608																		
max	99878.489209																		

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 гістограму розподілу та діаграму розмаху.

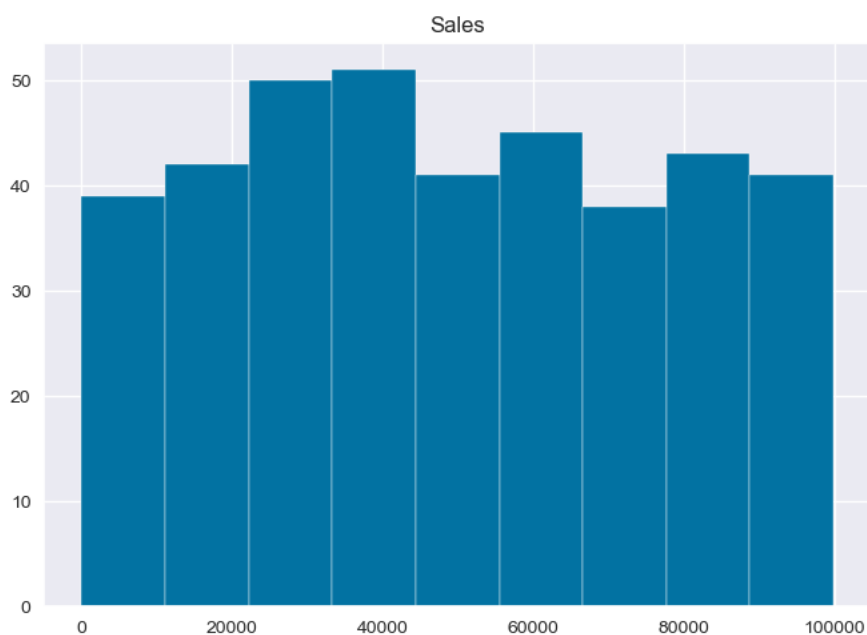


Рис. 3.5 – Гістограма розподілу продажів

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

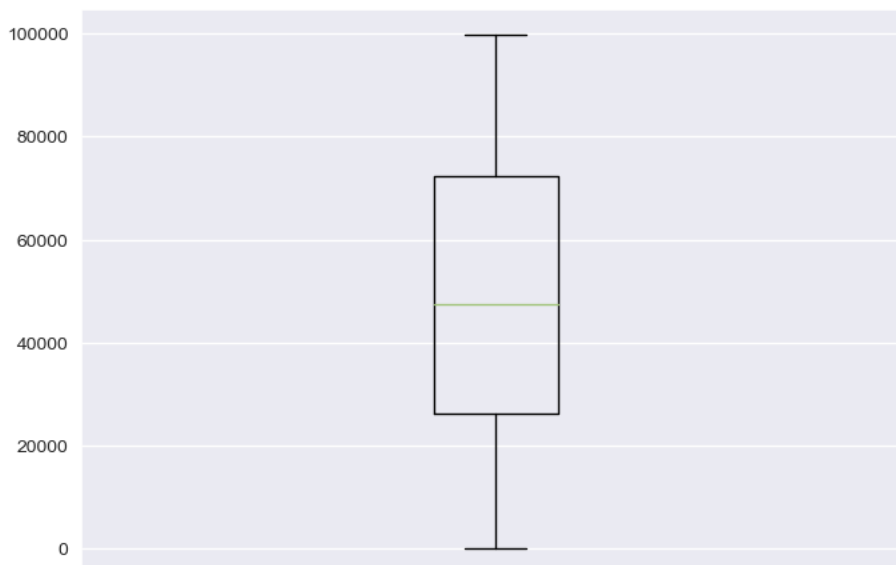


Рис. 3.6 – Діаграма розмаху продажів

На діаграмі розмаху можна чітко побачити, що викидів немає.

Тепер поглянемо на лінійний графік продажів та гістограму розподілу одночасно.

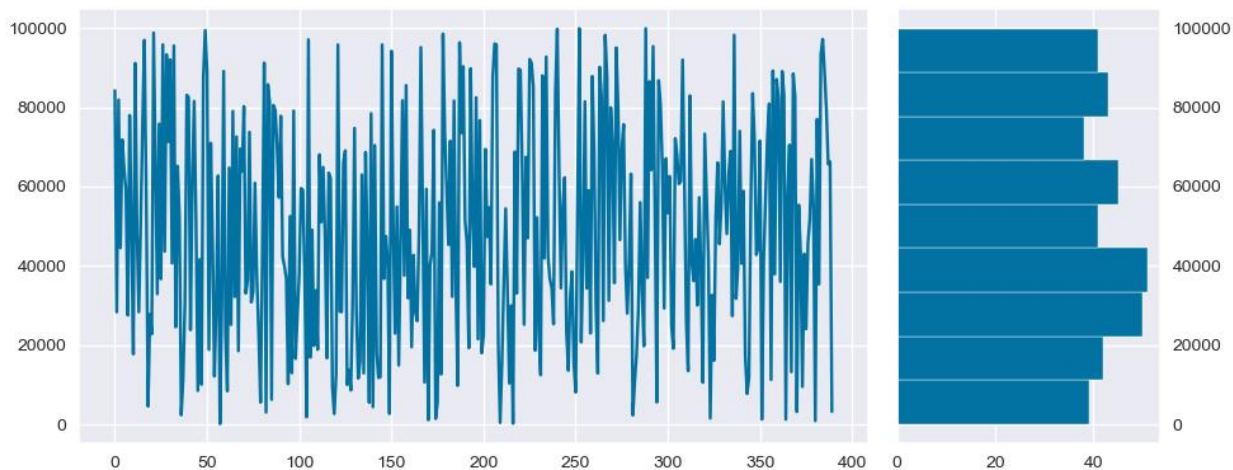


Рис. 3.7 – Лінійний графік та гістограма розподілу продажів

Аналогічно якихось аномальних вимірів не виявлено.

Сам датасет містить продажі за 2011, 2012 та 2013 рік, тому я вирішив переглянути аналогічні графіки та статистичні характеристики стовпця продажів за кожен з цих років.

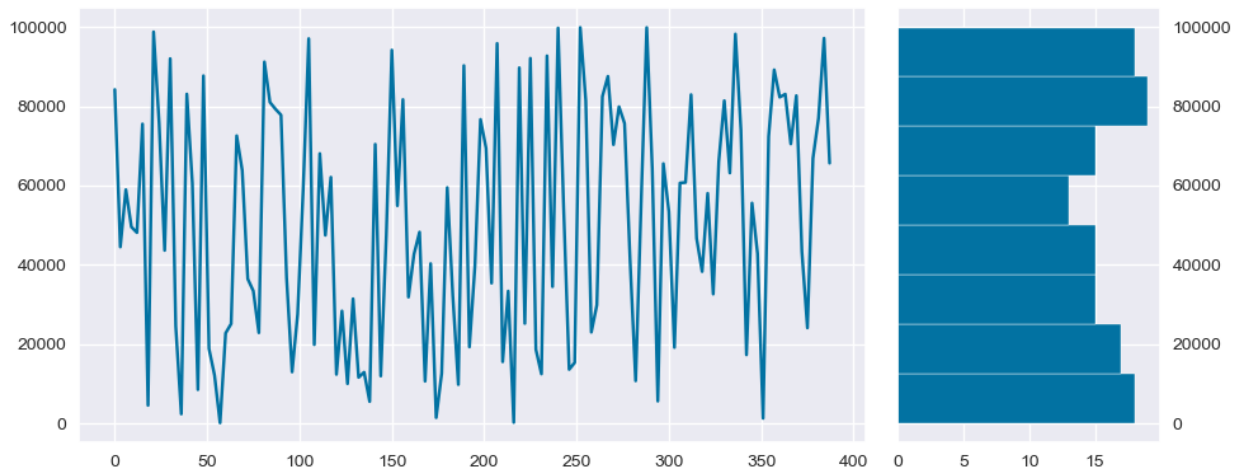


Рис. 3.8 – Лінійний графік та гістограма розподілу продажів за 2011 рік

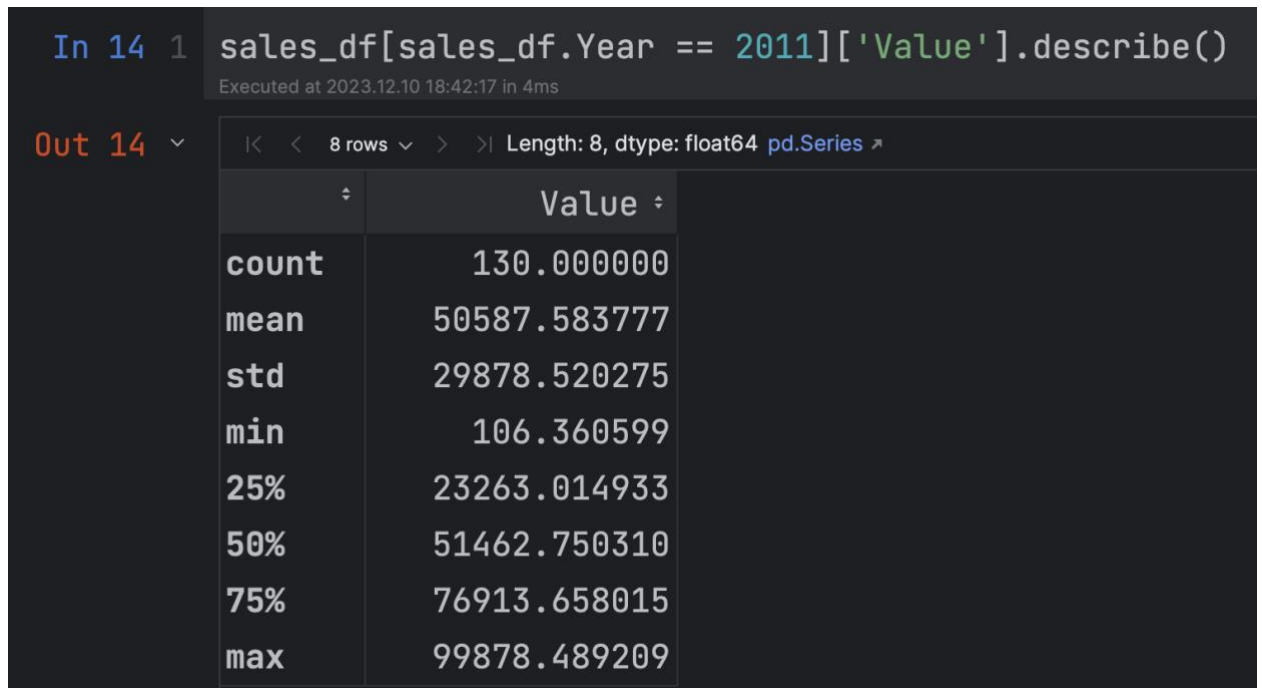


Рис. 3.9 – Статистичні характеристики продажів за 2011 рік

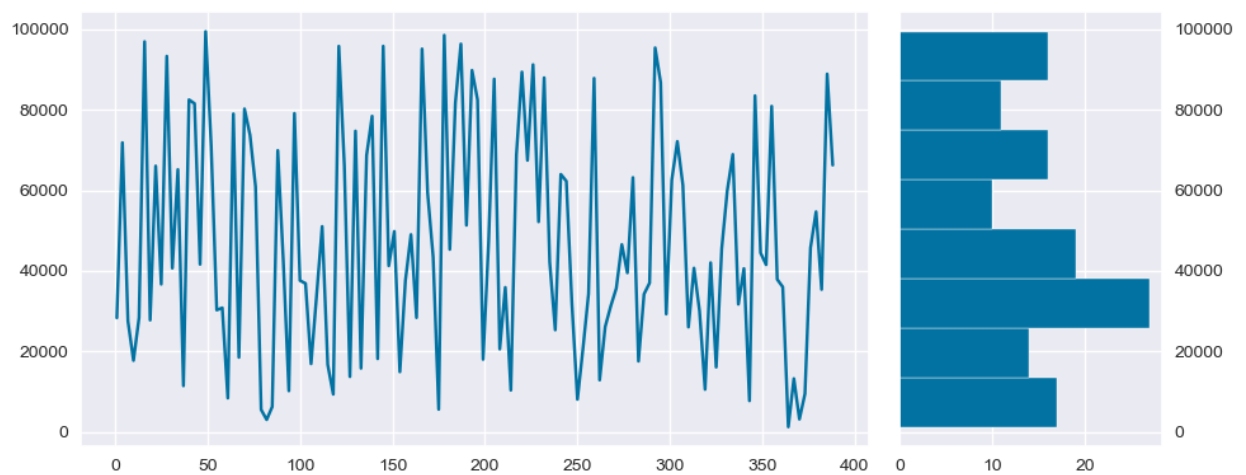


Рис. 3.10 – Лінійний графік та гістограма розподілу продажів за 2012 рік

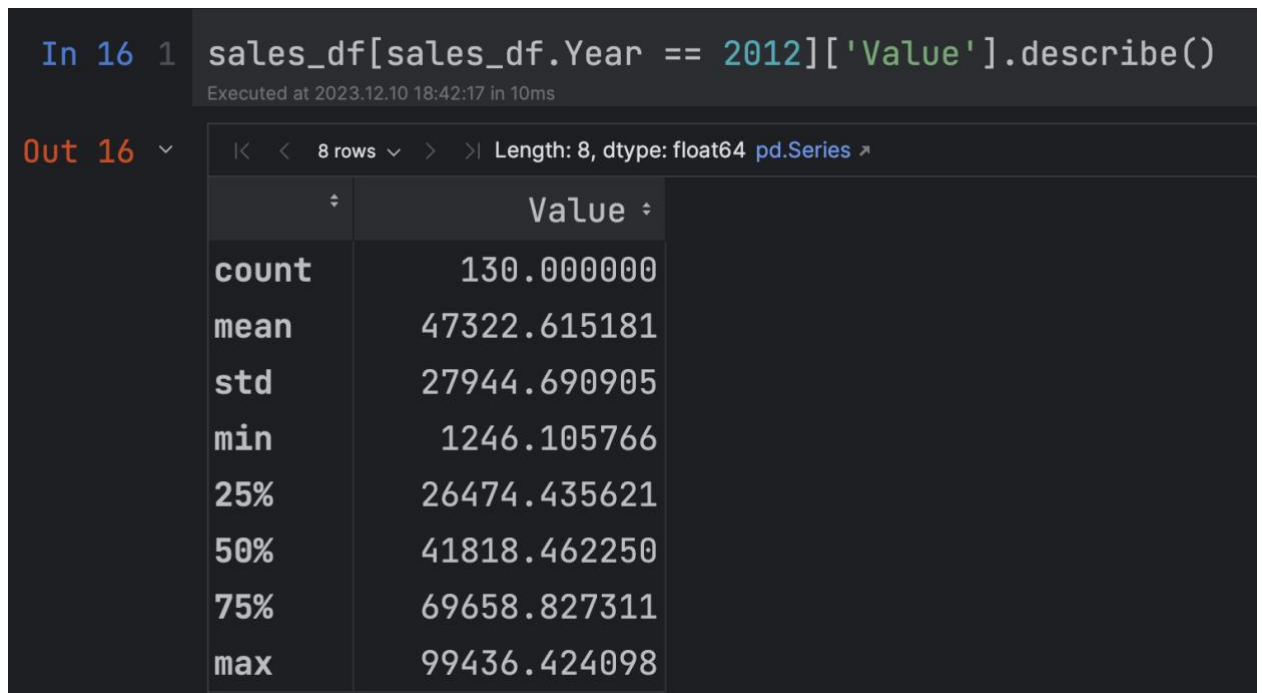


Рис. 3.11 – Статистичні характеристики продажів за 2012 рік

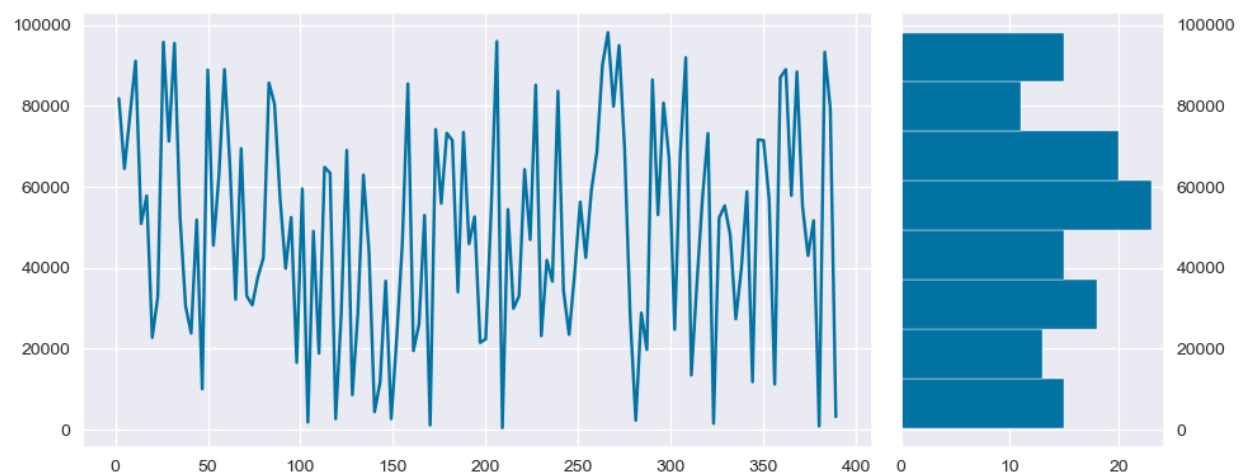


Рис. 3.12 – Лінійний графік та гістограма розподілу продажів за 2013 рік

In 18

1

sales_df[sales_df.Year == 2013]['Value'].describe()

Executed at 2023.12.10 18:42:17 in 8ms

Out 18

▼

<

>

8 rows ▼

>

>

Length: 8, dtype: float64

pd.Series ↗

↕	Value ↕
count	130.000000
mean	49777.965956
std	26968.856306
min	429.356425
25%	28824.251006
50%	52441.187887
75%	70873.817709
max	98199.933992

Рис. 3.13 – Статистичні характеристики продажів за 2013 рік

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.

Настройка параметров и добавление данных о продажах за период

In 19

1

sales_by_year = sales_df.groupby('Year')['Value'].sum().reset_index()

2

sales_by_year

Executed at 2023.12.10 18:42:17 in 3ms

Out 19

<

>

3 rows

>

>

3 rows × 2 columns

pd.DataFrame

	Year	Value
0	2011	6.576386e+06
1	2012	6.151940e+06
2	2013	6.471136e+06

Рис. 3.14 – Сумарне значення продажів у доларах за кожний з років

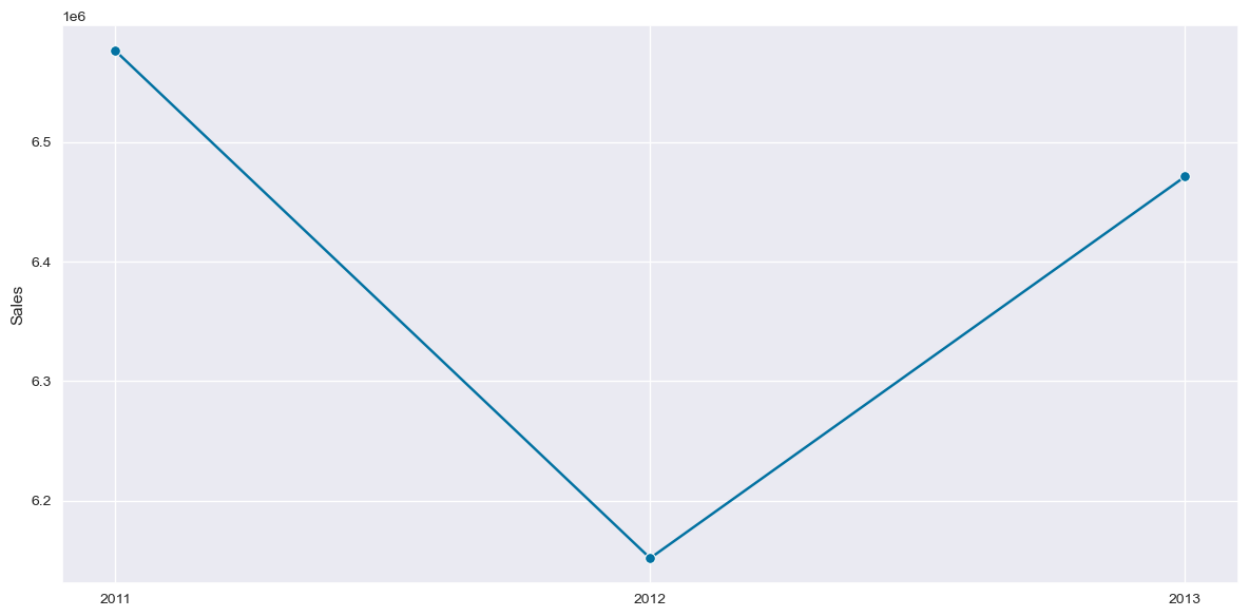


Рис. 3.15 – Динаміка сумарного значення продаж за роки

Можна побачити, що в 2012-му році було суттєве зниження сумарного значення продаж у доларах, а у 2013 вже знову відбувся ріст. Такий результат був очікуваним, адже кількість продаж за кожний рік є однаковою, тобто середні значення відображають динаміку сумарних продаж.

Далі я проаналізував середні значення продаж за кожний рік для кожного торгового представника.

```
In 21 1 avg_sales_year_representative = sales_df.groupby(['Year', 'Sales_Representative'])['Value'].mean()
      2 .reset_index()
      3 avg_sales_year_representative
```

Executed at 2023.12.10 18:42:17 in 14ms

Out 21

	Year	Sales_Representative	Value
0	2011	Ashish	46829.122073
1	2011	Jane	56205.908485
2	2011	John	49643.635644
3	2012	Ashish	50141.586408
4	2012	Jane	47333.772158
5	2012	John	43912.806087
6	2013	Ashish	50580.899288
7	2013	Jane	50974.143450
8	2013	John	47779.602629

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

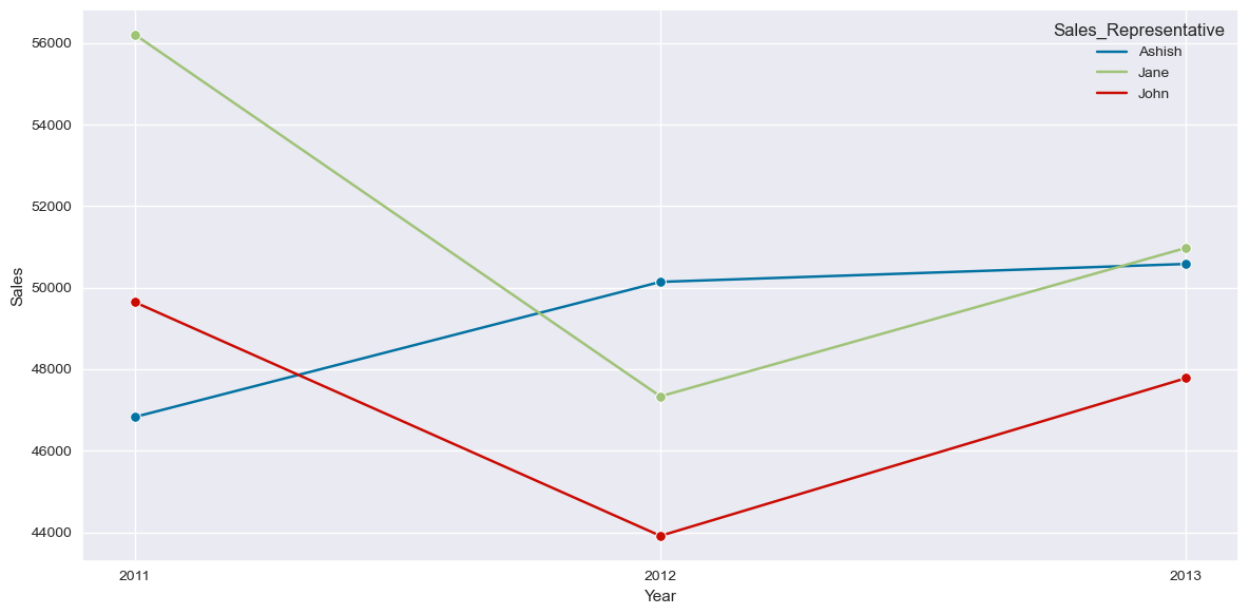


Рис. 3.17 – Динаміка зміни середніх значень продаж для кожного торгового представника

На графіку можна помітити, що, на відміну від загального тренду, для Ashish у 2012 відбувся зріст по значенню продаж у середньому. Для цього представника за всі три роки динаміка позитивна, а інші два представники мають тренд, аналогічний загальному.

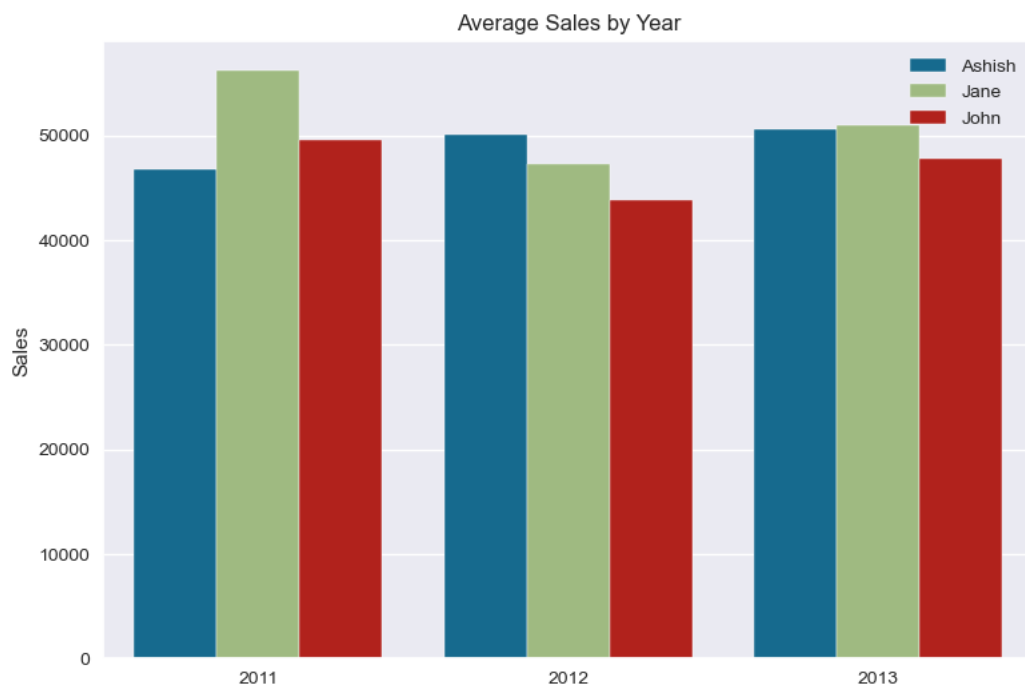


Рис. 3.18 – Стовпчаста діаграма середніх значень продаж для кожного торгового представника за кожний рік

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 виявився 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.

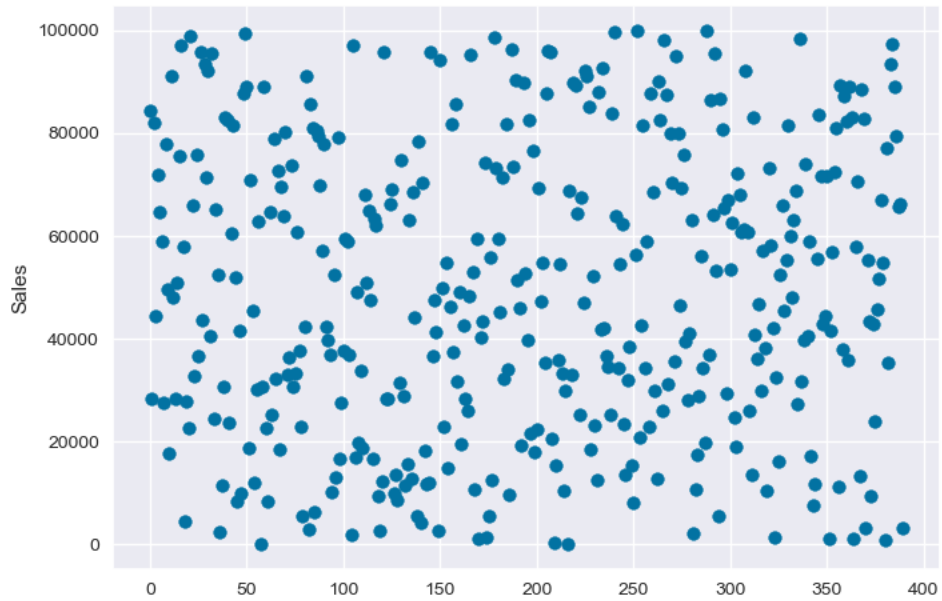


Рис. 3.19 – Діаграма розсіювання

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

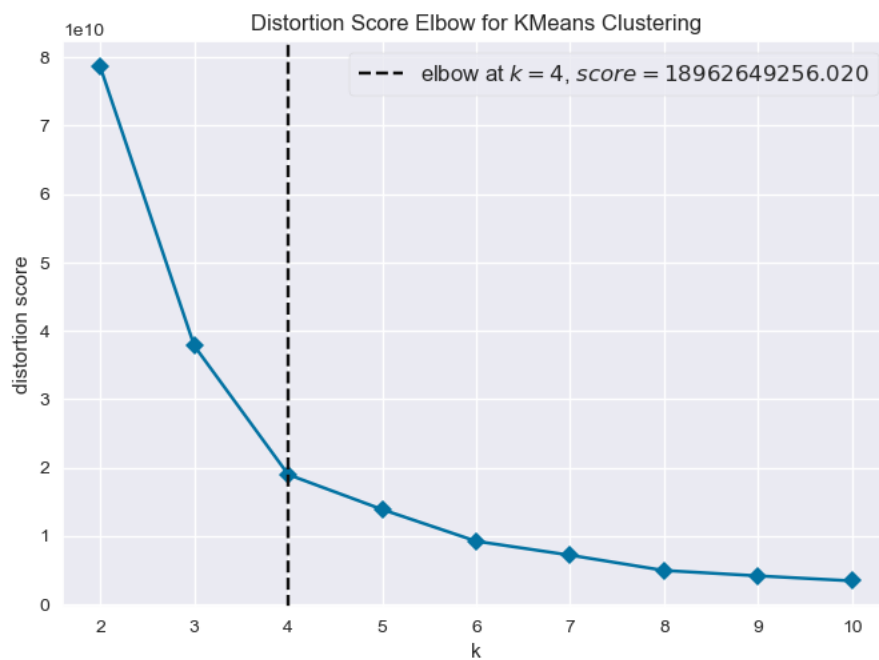


Рис. 3.20 – Підбір оптимальної кількості кластерів

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.

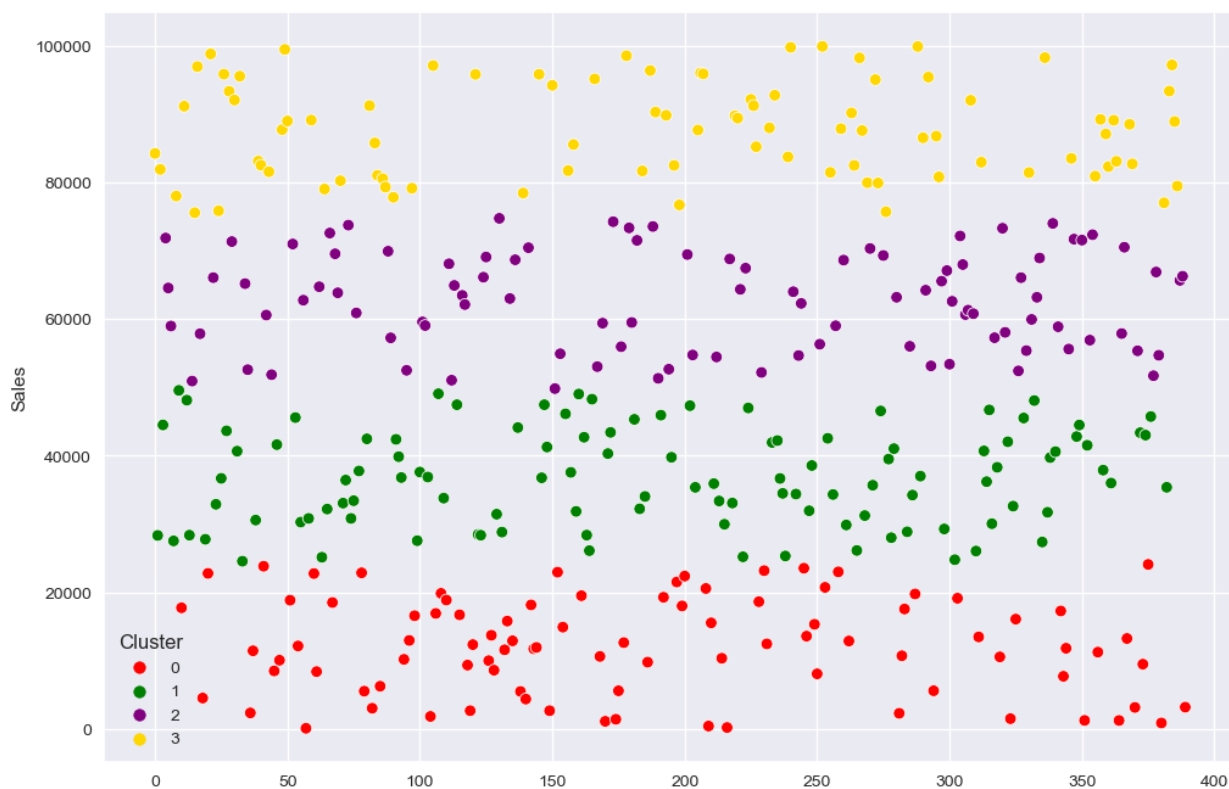


Рис. 3.21 – Візуалізація кластерів

Можна побачити, що було отримано 4 чітко виділені кластери, які розбиті по інтервалам (можна умовно виділити 4 смуги). Кластер під номером 0, наприклад, містить продажі до 20000 з чимось доларів, а останній кластер містить найдорожчі продажі від майже 80000 доларів і більше.

Потім я дослідив кількість продаж у кожного торгового представника в рамках різних кластерів.

```
In 27 1 sales_by_clusters = sales_df.groupby(['Cluster', 'Sales_Representative'])['Value'].count().reset_index()
      2 sales_by_clusters
      Executed at 2023.12.10 18:42:18 in 12ms
```

Out 27

#	Cluster	Sales_Representative	Value
0	0	Ashish	30
1	0	Jane	29
2	0	John	32
3	1	Ashish	40
4	1	Jane	32
5	1	John	39
6	2	Ashish	30
7	2	Jane	36
8	2	John	33
9	3	Ashish	30

Рис. 3.22 – Кількість продаж кожного торгового представника в рамках кластерів

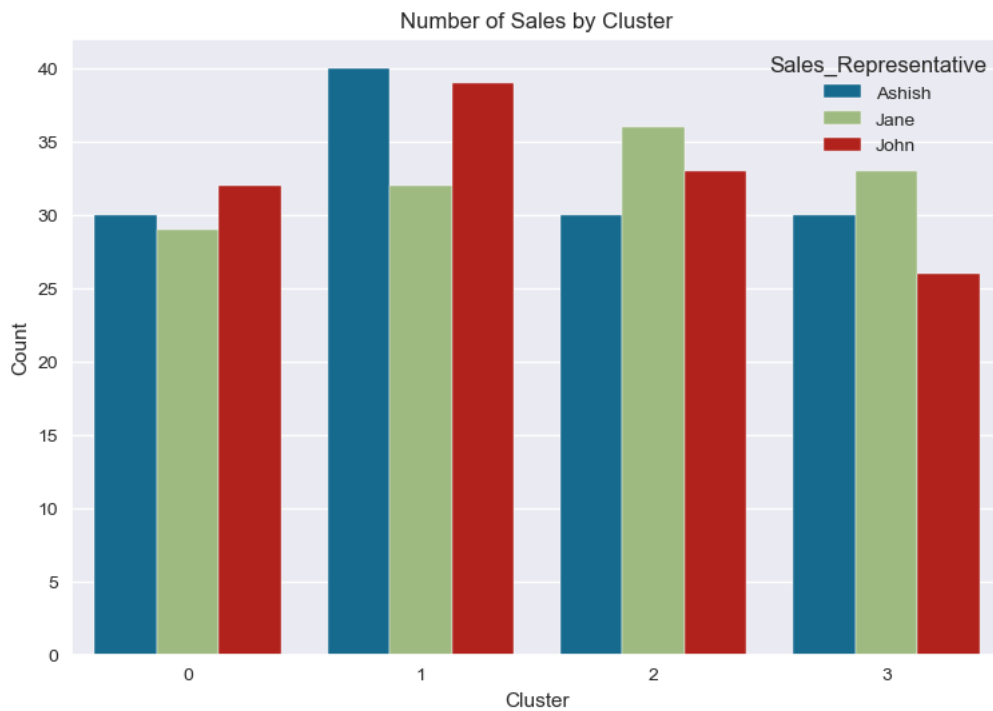


Рис. 3.23 – Стовпчаста діаграма кількості продажів у кожного торгового представника по різним кластерам

Отже, бачимо, що Jane здійснила найбільшу кількість найдорожчих продаж (33), що також є причиною того, що вона за 2 із трьох років є лідером по значенню продаж, в той час як John здійснив найбільшу кількість найдешевших продаж. Загалом найбільше було здійснено продаж у кластері під номером 1 (приблизно від 30000 до 50000 доларів). Варто зазначити, що якраз середнє значення та медіана продаж якраз також належать цьому кластеру, що й пояснює отримані результати. Також можна помітити, що Ashish здійснив найбільшу кількість цих продаж (40), а інших продаж порівну (30 у кожному кластері).

IV. Висновок

Отже, в ході даної лабораторної я 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. Таким чином, на мою думку, досить ефективним з точки зору компанії для отримання стабільної позитивної динаміки буде здійснення майже всіх продаж у межах приблизно від 30000 до 50000 доларів.