Московский государственный технический
университет им. Н.Э. Баумана.

Факультет «Информатика и управление»

Кафедра ИУ5. Курс «Методы машинного обучения»

Отчет по лабораторной работе №1

«Создание "истории о данных"»

Выполнила:
студентка группы ИУ5-25М
Зозуля О.А.
Подпись и дата:

Проверил:
преподаватель каф. ИУ5
Гапанюк Юрий Евгеньевич
Подпись и дата:

Москва, 2023 г.

ЗАДАНИЕ

- Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь.

Для лабораторных работ не рекомендуется выбирать датасеты очень большого размера.

- Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований:

1. История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
2. На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
3. Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
4. Выбор графиков должен быть обоснован использованием методологии data-to-viz. Рекомендуется учитывать типичные ошибки построения выбранного вида графика по методологии data-to-viz. Если методология Вами отвергается, то просьба обосновать Ваше решение по выбору графика.
5. История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

- Сформировать отчет и разместить его в своем репозитории на github.

1. Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований:

```
2. import matplotlib
   from mpl_toolkits import mplot3d
   import pandas as pd
   from matplotlib import pyplot as plt
   import numpy as np
   import seaborn as sns
   import os
```

```
import pandas as pd
from matplotlib import pyplot as plt
import numpy as np
import os
path=os.environ["userprofile"]+"\\"+".atom"+"\\"+FileName
print(path)
```

data = pd.read_csv(path)

print(data)

```
path=os. environ["userprofile"]+"\\"+".atom"+"\\"+"Transactions.csv"
print(path)
```

C:\Users\657432343536\.atom\Transactions.csv

```
data = pd.read_csv(path)
print(data)
```

In [1]:
```
import pandas as pd
import matplotlib as plt
import numpy as np
import os
FileName="Transactions.csv"
```

In [2]:
```
path=os.environ["userprofile"]+"\\"+".atom"+"\\"+FileName
print(path)
```

```
C:\Users\657432343536\.atom\Transactions.csv
```

In [3]:
```
data = pd.read_csv(path)
print(data)
```

```
      Date_and_time_of_unloading  Product_code  Amount  Sale_amount  \
0            2020-01-01 23:00:00           144     1.0       280.00
1            2020-01-01 23:00:00           209     2.0       545.73
2            2020-01-01 23:00:00           213     2.0      1265.05
3            2020-01-01 23:00:00           217     1.0       630.00
4            2020-01-01 23:00:00           222     2.0      1104.75
...                          ...           ...     ...          ...
50079        2022-09-18 15:00:00          5316     6.0      1875.95
50080        2022-09-18 15:00:00          5317     2.0       555.95
50081        2022-09-18 15:00:00          5318     2.0       572.50
50082        2022-09-18 15:00:00          5321     1.0       300.00
50083        2022-09-18 15:00:00          5322     2.0       600.00

       Discount_amount   Profit  Percentage_markup  Discount_percentage
0                  NaN   155.00             124.00                  NaN
1               294.27    75.73              16.11                35.03
2                34.95   653.05             106.71                 2.69
3                70.00   220.50              53.85                10.00
4               195.25   393.75              55.38                15.02
...                ...      ...                ...                  ...
50079           104.05  1095.95             140.51                 5.26
50080           104.05   315.95             131.65                15.77
50081            87.50   312.50             120.19                13.26
50082              NaN   180.00             150.00                  NaN
50083              NaN   340.00             130.77                  NaN

[50084 rows x 8 columns]
```

Linear regression

In [13]:
```
plt.plot(data["Product_code"][:1000],
         data["Sale_amount"][:1000],
        )

plt.title("Line graph")
plt.ylabel('Sale_amount')
plt.xlabel('Product_code')
plt.show()
```

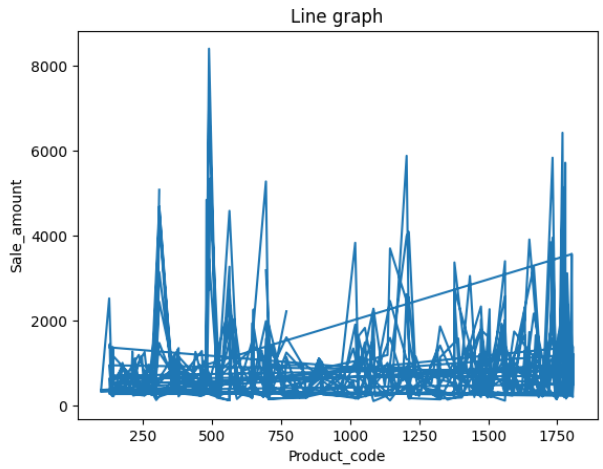

Рисунок 1. Зависимость кода продукта от суммы продажи

```
plt.plot(data["Product_code"][:1000],
         data["Sale_amount"][:1000],
        )
 plt.title("Line graph")
plt.ylabel('Sale_amount')
plt.xlabel('Product_code')
plt.show()
```

```
data.plot.area()
```

Area plot

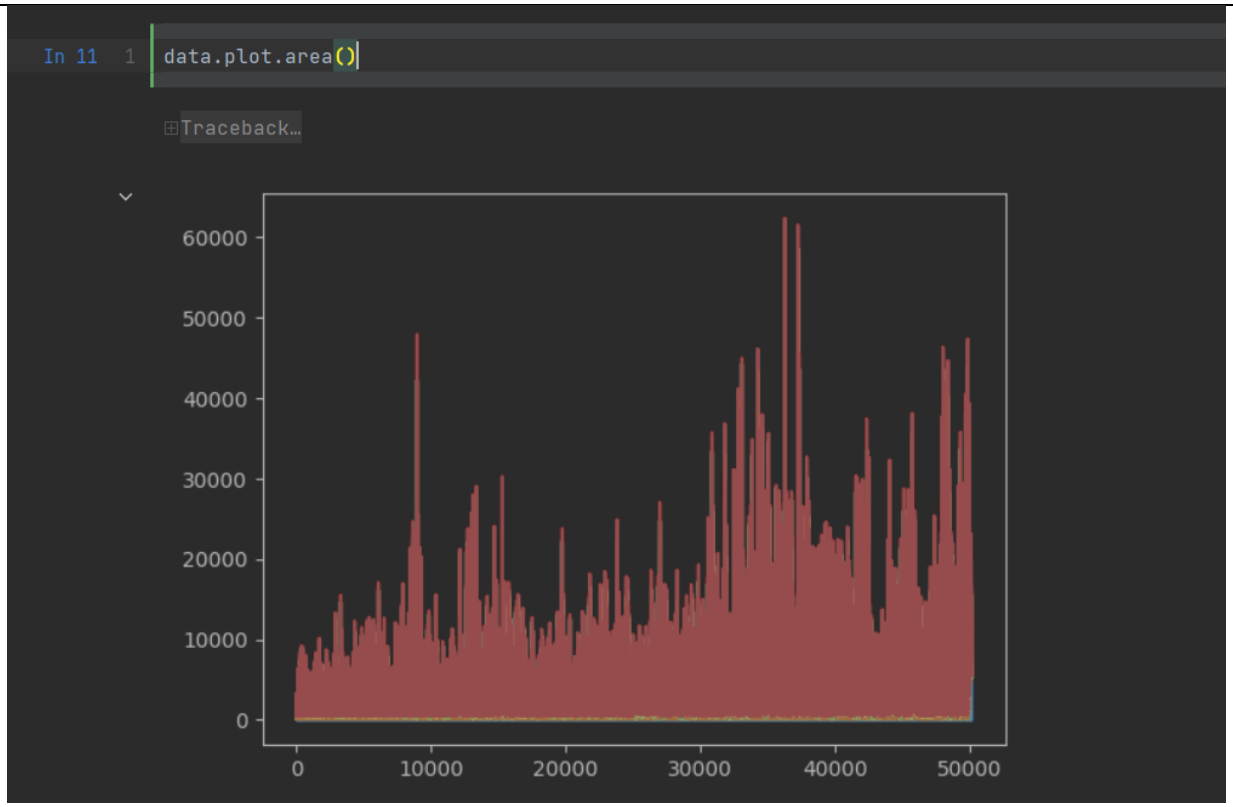

Рисунок 2. Правильный график Area plot

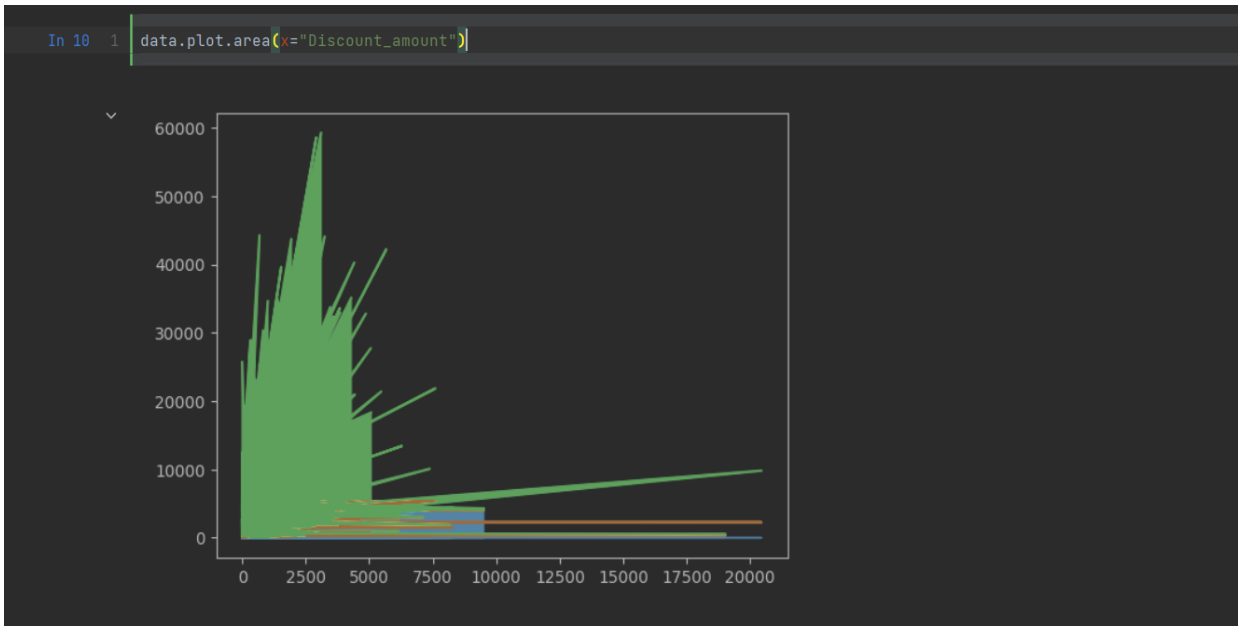

Рисунок 3. Неправильный график Area plot от суммы скидки

FILL BETWEEN

```
plt.fill_between(data["Product_code"][:50], data["Profit"][:50])

plt.show()
```

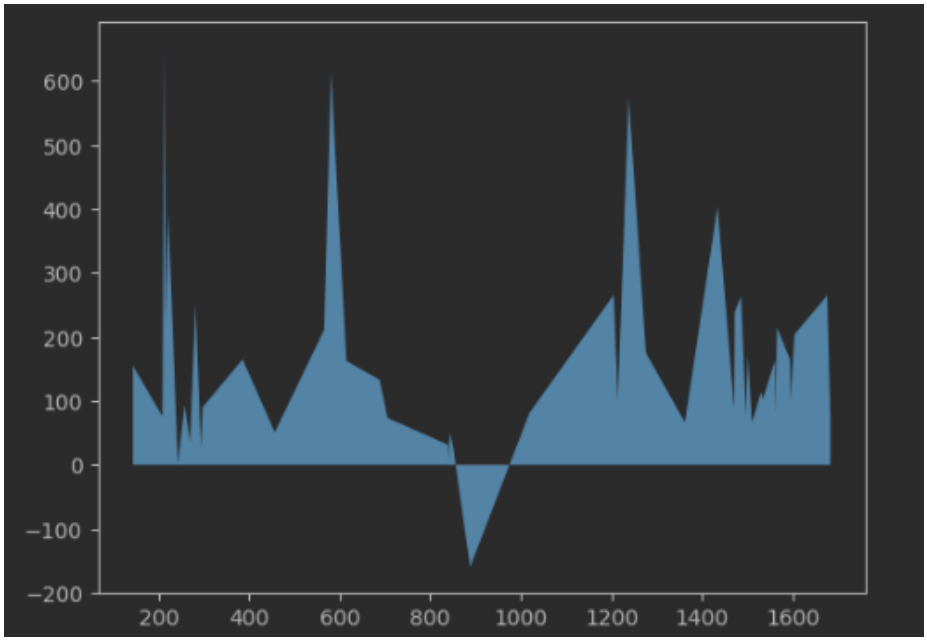

Рисунок 4. Зависимость кода продукта от прибыли

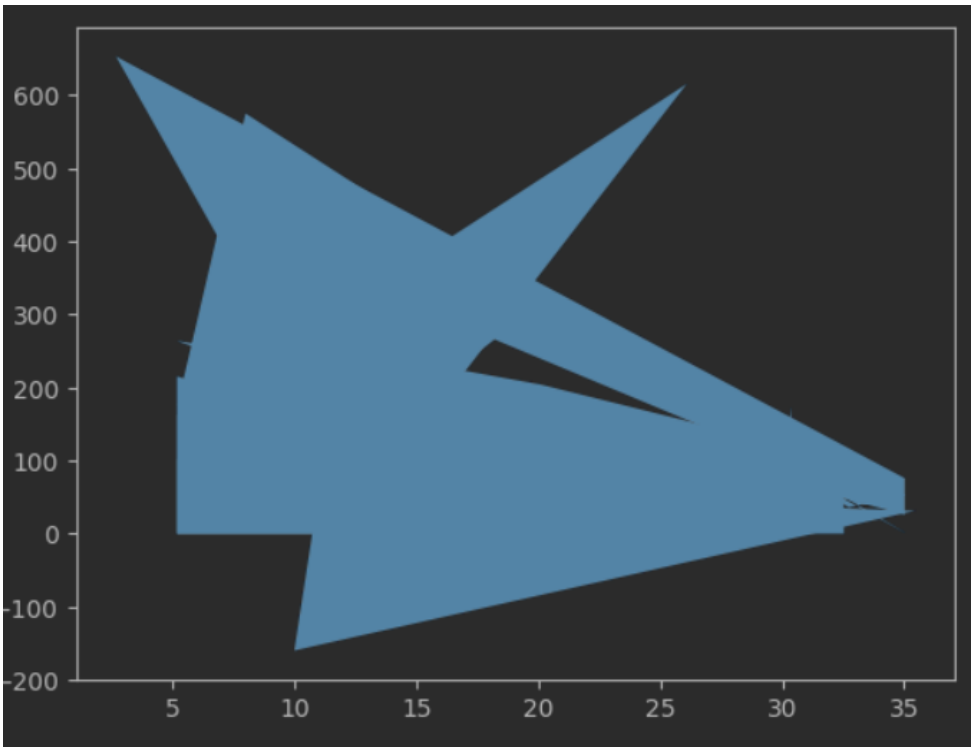

Рисунок 5. Зависимость Discount_percentage от прибыли

SCATTER PLOT

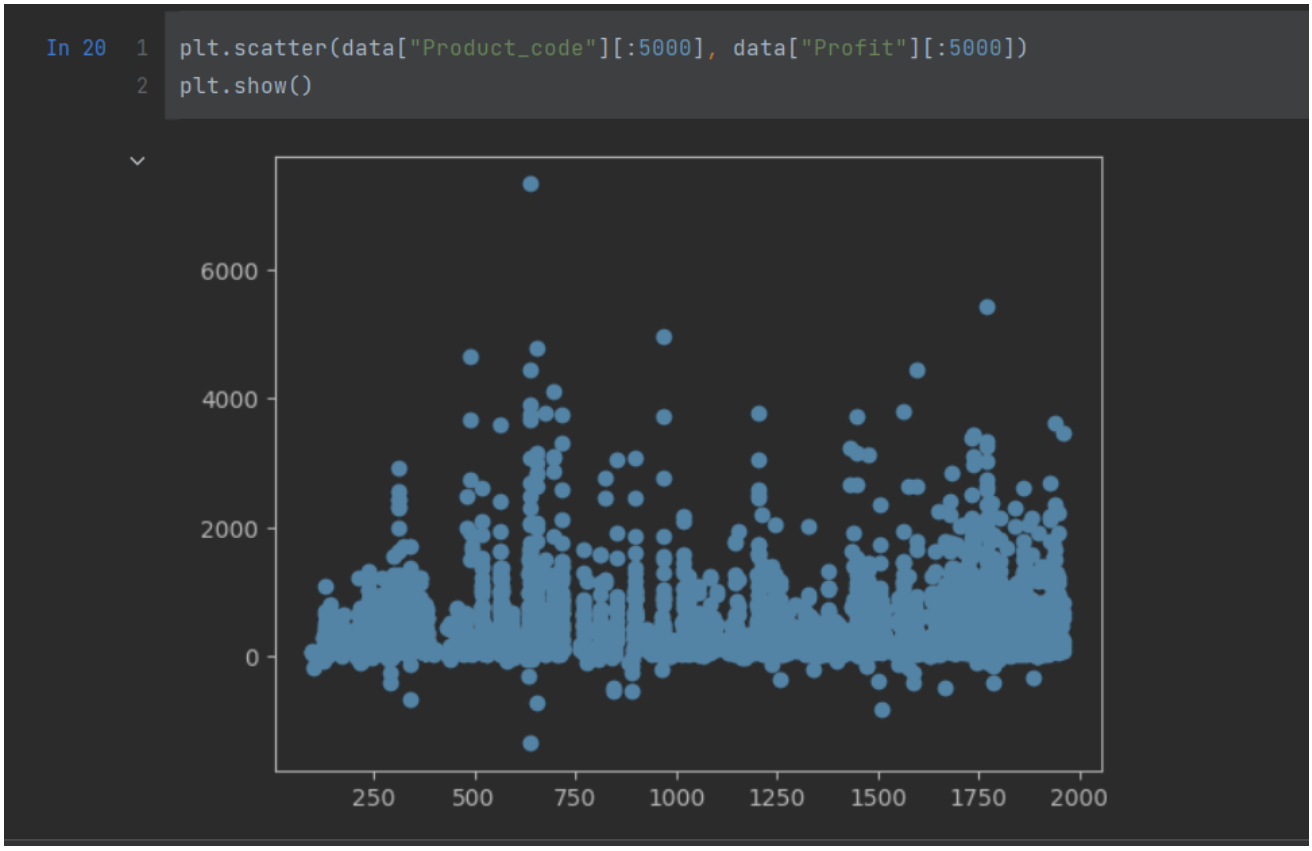

Рисунок 5. Зависимость кода продукта от прибыли

Далее я буду пользоваться библиотекой seaborn

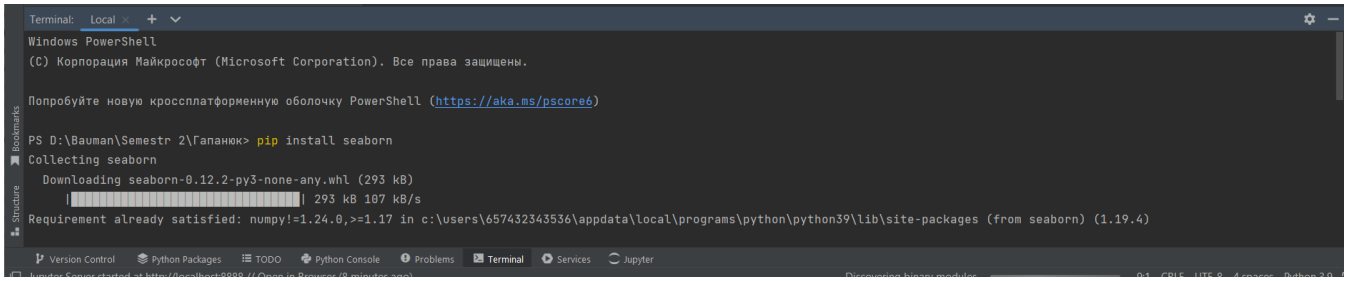

Рисунок 6. Загрузка новой библиотеки

HEATMAP

```
sns.heatmap(data.corr(), xticklabels=data.corr().columns,
            yticklabels=data.corr().columns,
            cmap='turbo',
            center=0,
            annot=True)
```

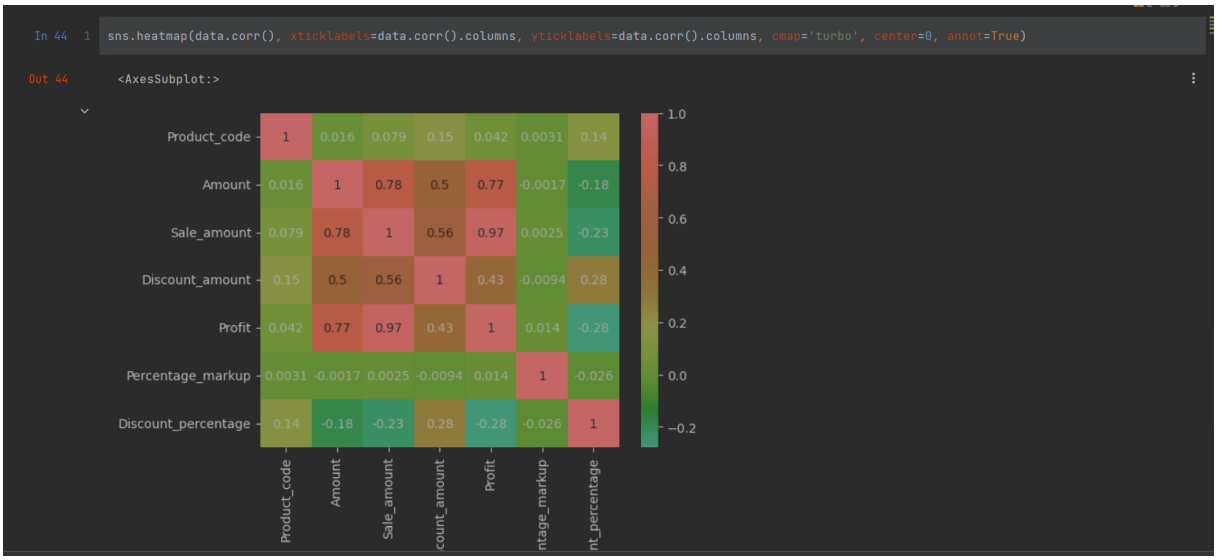

Рисунок 7. Тепловая карта датасета

Displot

```
sns.distplot(data["Profit"], hist=True, kde=False, rug=False )
```

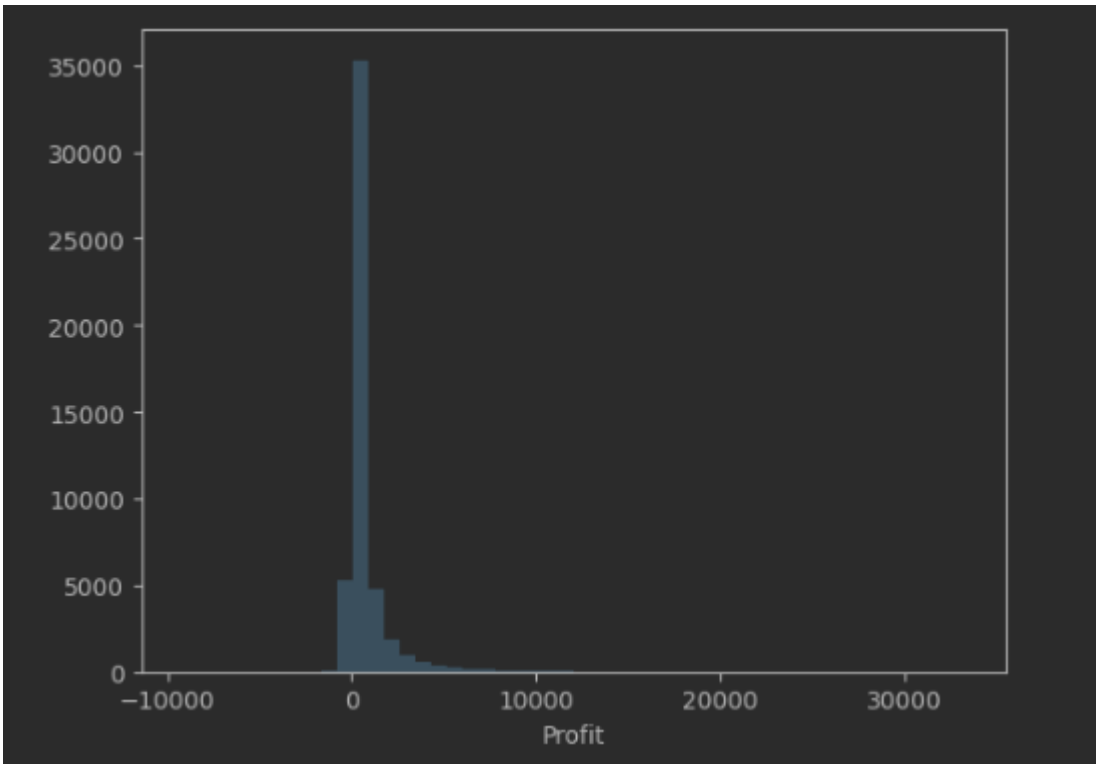

Рисунок 8. Тепловая карта датасета

Violin plot

```
zuh=[data["Profit"][:50],data["Sale_amount"][:50],data["Discount_amount"][:50]]
sns.violinplot(zuh)
```

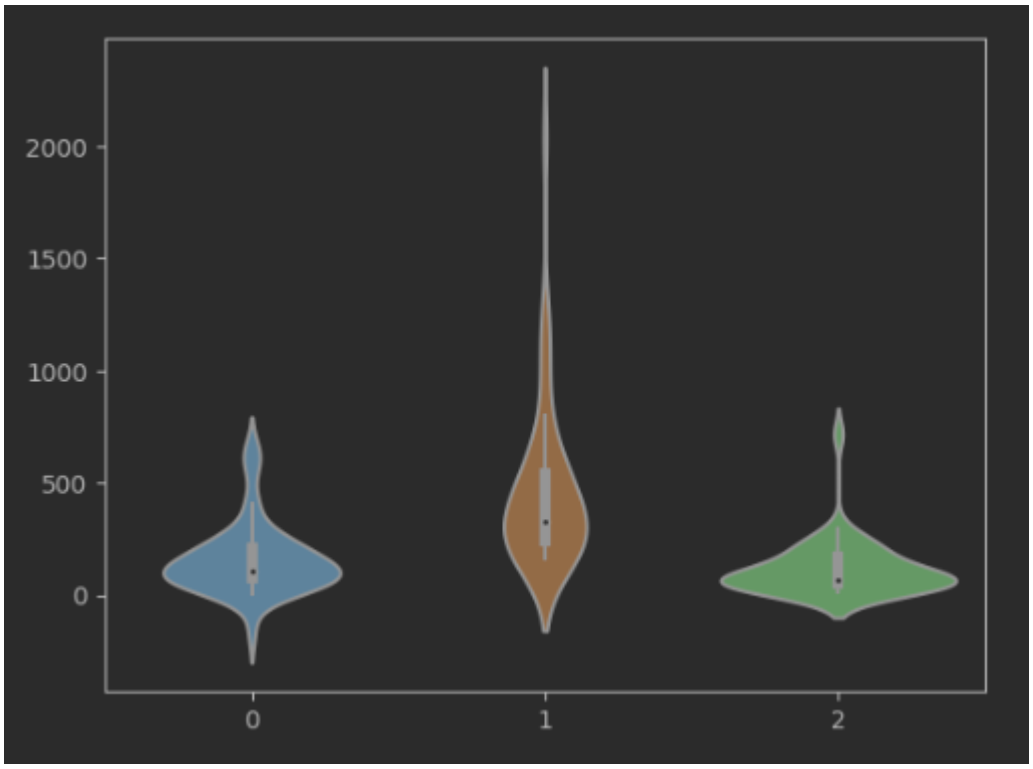

Рисунок 9. Скрипичный график зависимости прибыли

BUBBLE PLOT

```
sns.scatterplot(data=data[:500], legend=False, sizes=(20, 2000))
```

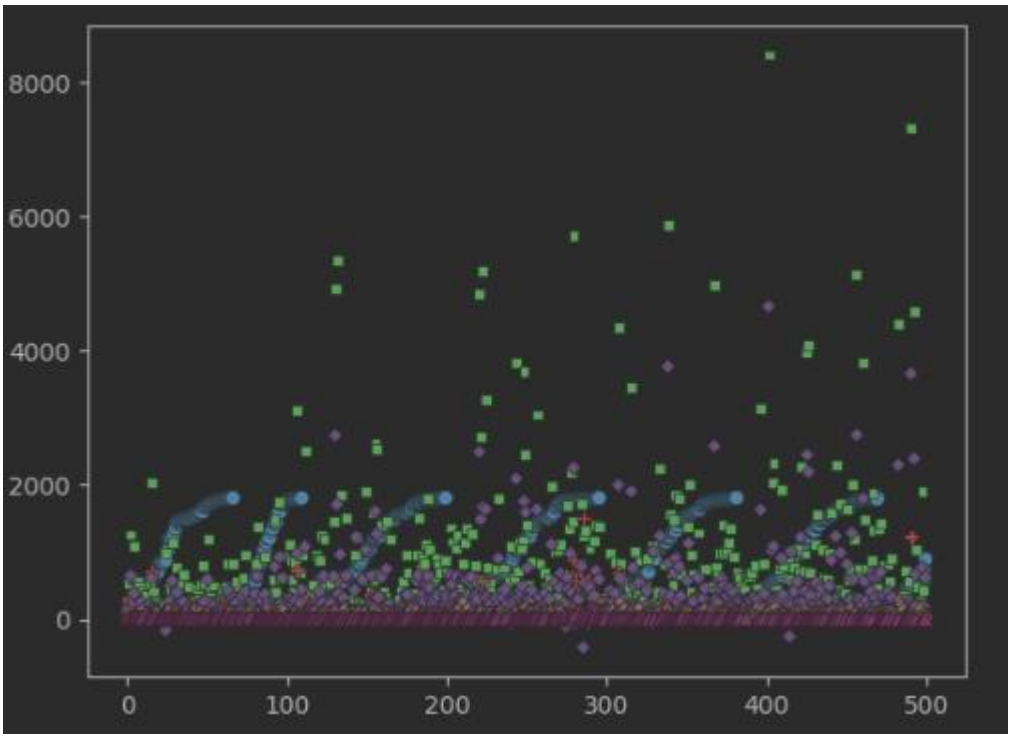

Рисунок 10. Пузырьковая диаграмма