**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: |  | Проверил: |
| студентка группы ИУ5-25М |  | преподаватель каф. ИУ5 |
| Зозуля О.А. |  | Гапанюк Юрий Евгеньевич |
| Подпись и дата: |  | Подпись и дата: |

Москва, 2023 г.

# ЗАДАНИЕ

* Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов [здесь.](https://github.com/ugapanyuk/courses_current/wiki/DSLIST)

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)

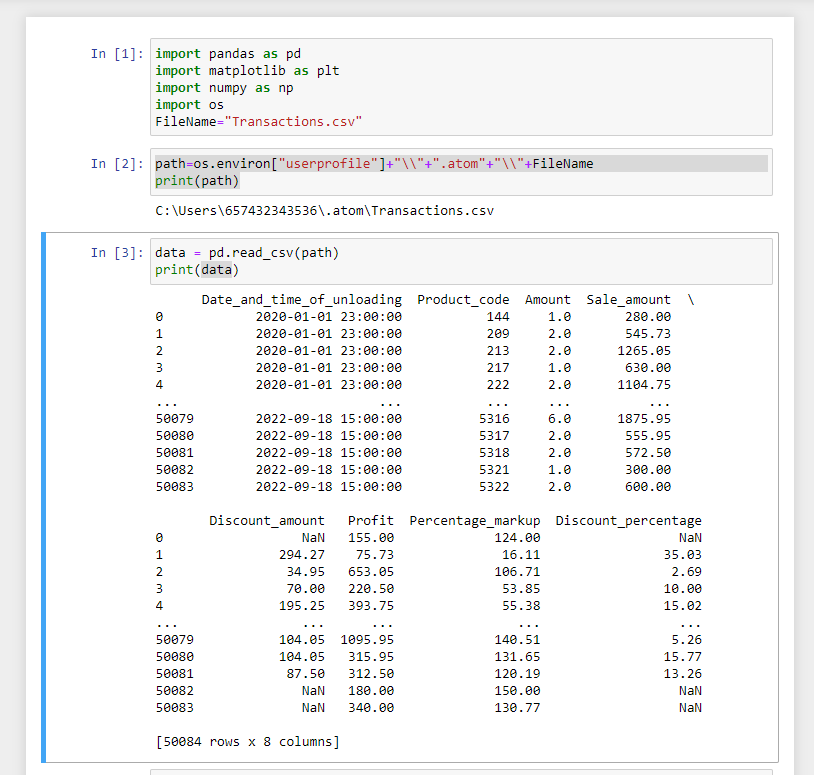
data = pd.read\_csv(path)

print(data)

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

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

data = pd.read\_csv(path)  
print(data)



Linear regression

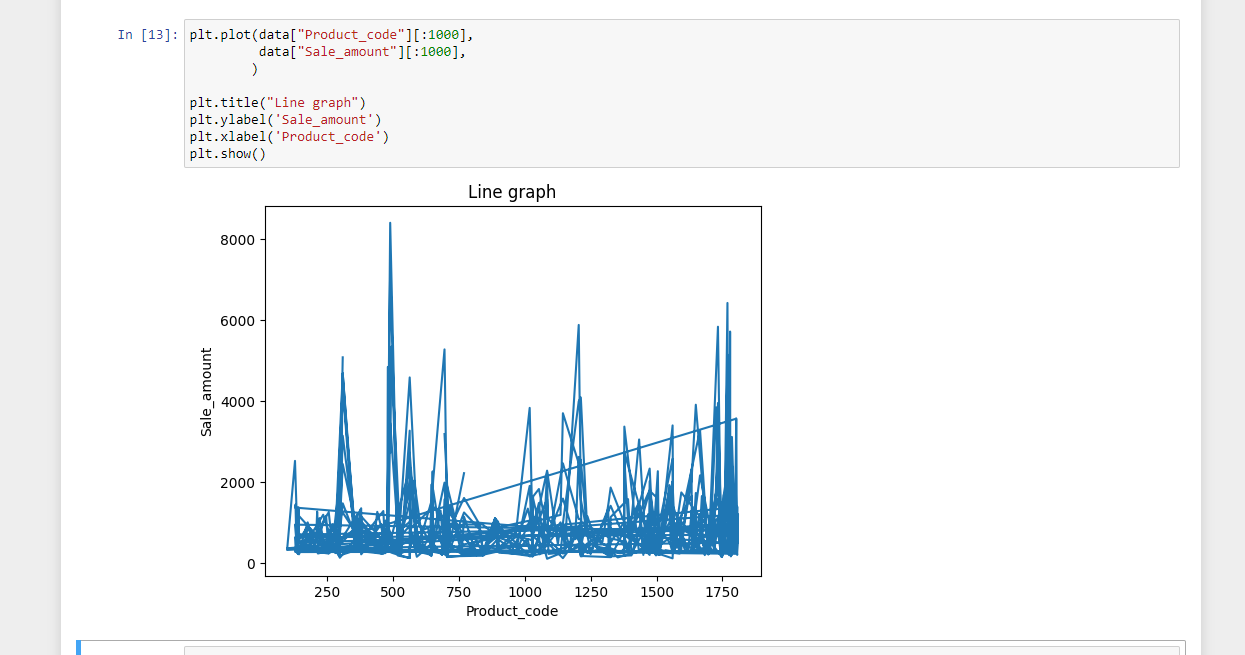


Рисунок 1. Зависимость кода продукта от суммы продажи

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**

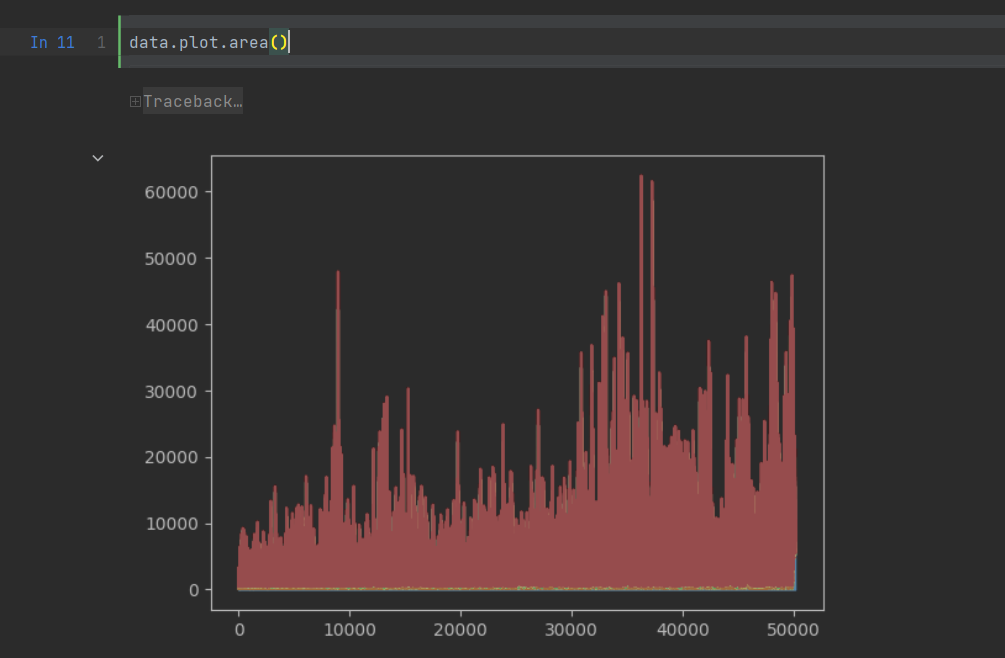


Рисунок 2. Правильный график Area plot

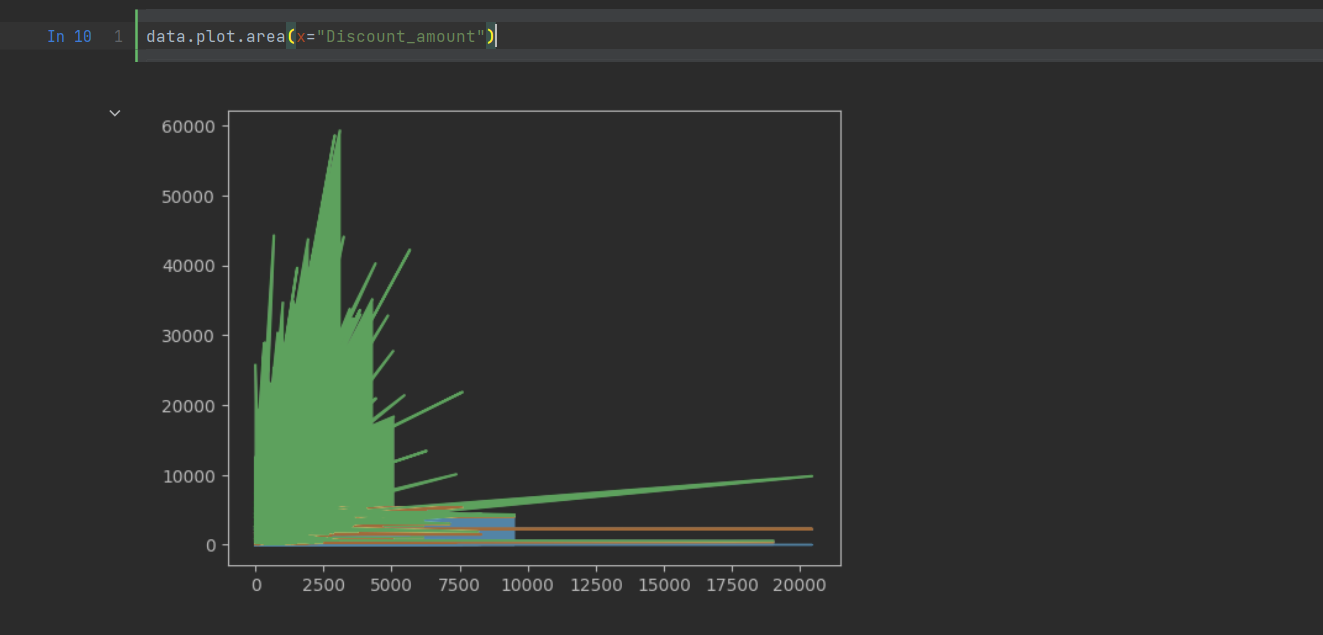


Рисунок 3. Неправильный график Area plot от суммы скидки

**FILL BETWEEN**

plt.fill\_between(data["Product\_code"][:50], data["Profit"][:50])  
  
plt.show()

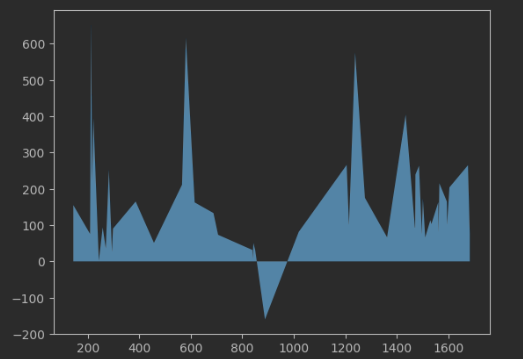


Рисунок 4. Зависимость кода продукта от прибыли

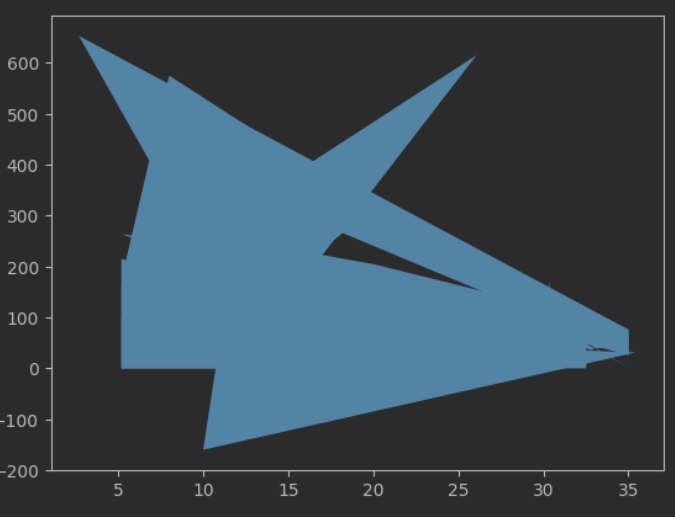


Рисунок 5. Зависимость Discount\_percentage от прибыли

**SCATTER PLOT**

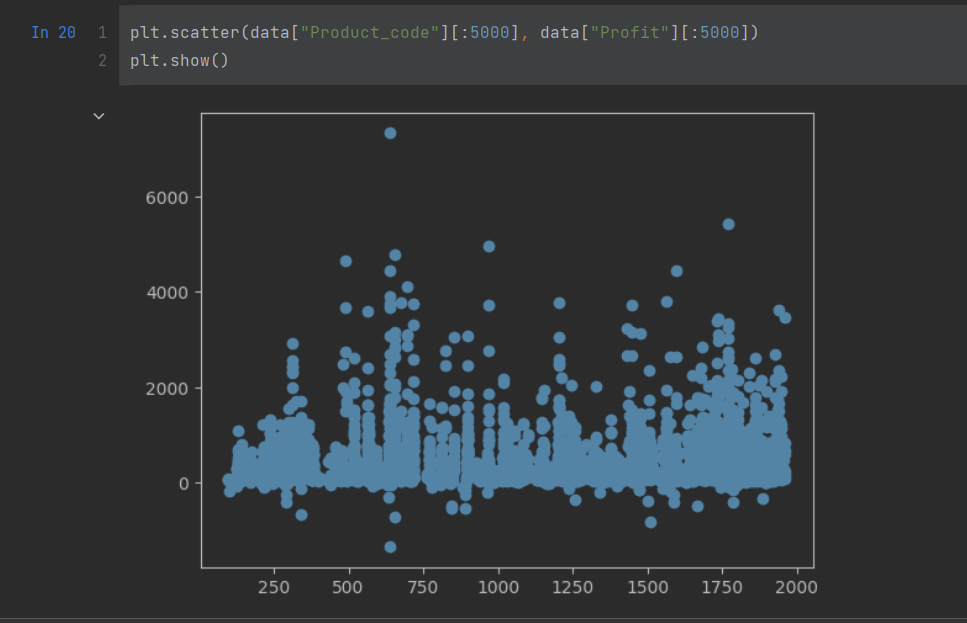


Рисунок 5. Зависимость кода продукта от прибыли

Далее я буду пользоваться библиотекой seaborn

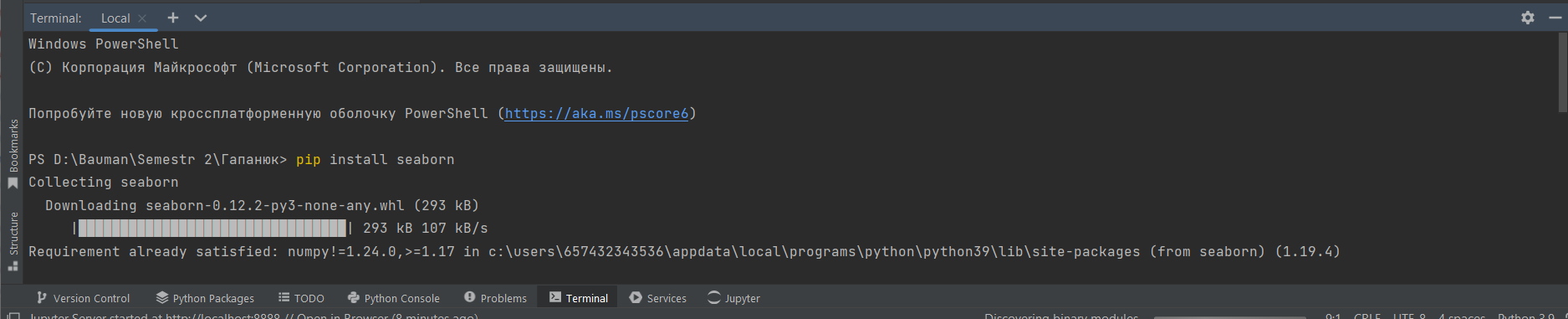


Рисунок 6. Загрузка новой библиотеки

**HEATMAP**

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

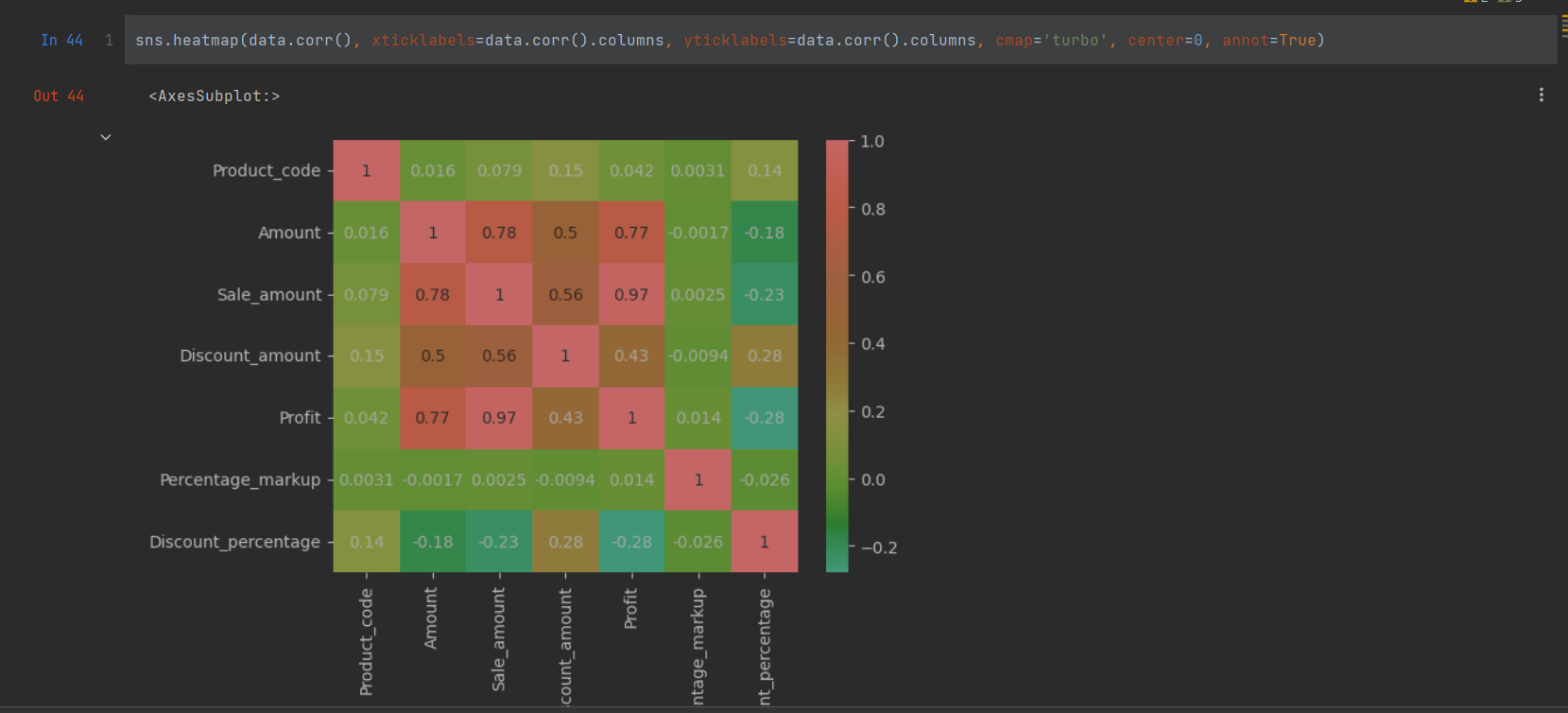


Рисунок 7. Тепловая карта датасета

**Displot**

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

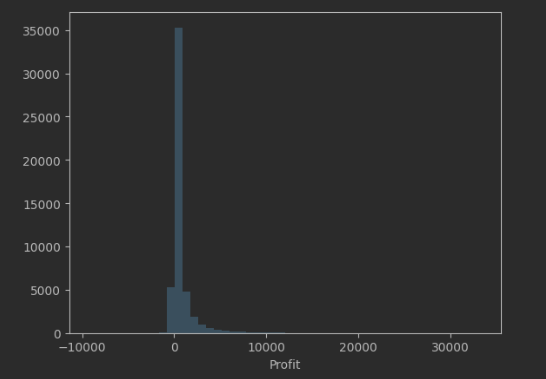


Рисунок 8. Тепловая карта датасета

**Violin plot**

zuh=[data["Profit"][:50],data["Sale\_amount"][:50],data["Discount\_amount"][:50]]  
sns.violinplot(zuh)

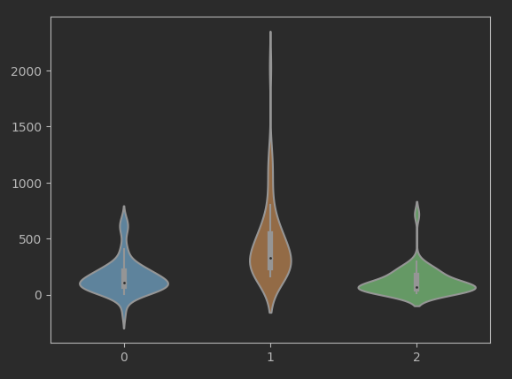


Рисунок 9. Скрипичный график зависимости прибыли

**BUBBLE PLOT**

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

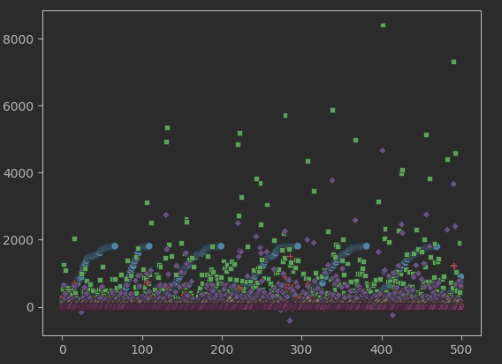


Рисунок 10. Пузырьковая диаграмма