Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Лабораторная работа №1
по дисциплине
«Методы машинного обучения»
на тему

«Создание "истории о данных" (Data Storytelling)»

Выполнил:
студент группы ИУ5И-21М
Дун Чжэнянь

Москва — 2024 г.

1. Цель лабораторной работы

Изучение различных методов визуализация данных и создание истории на основе данных.

2. 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.

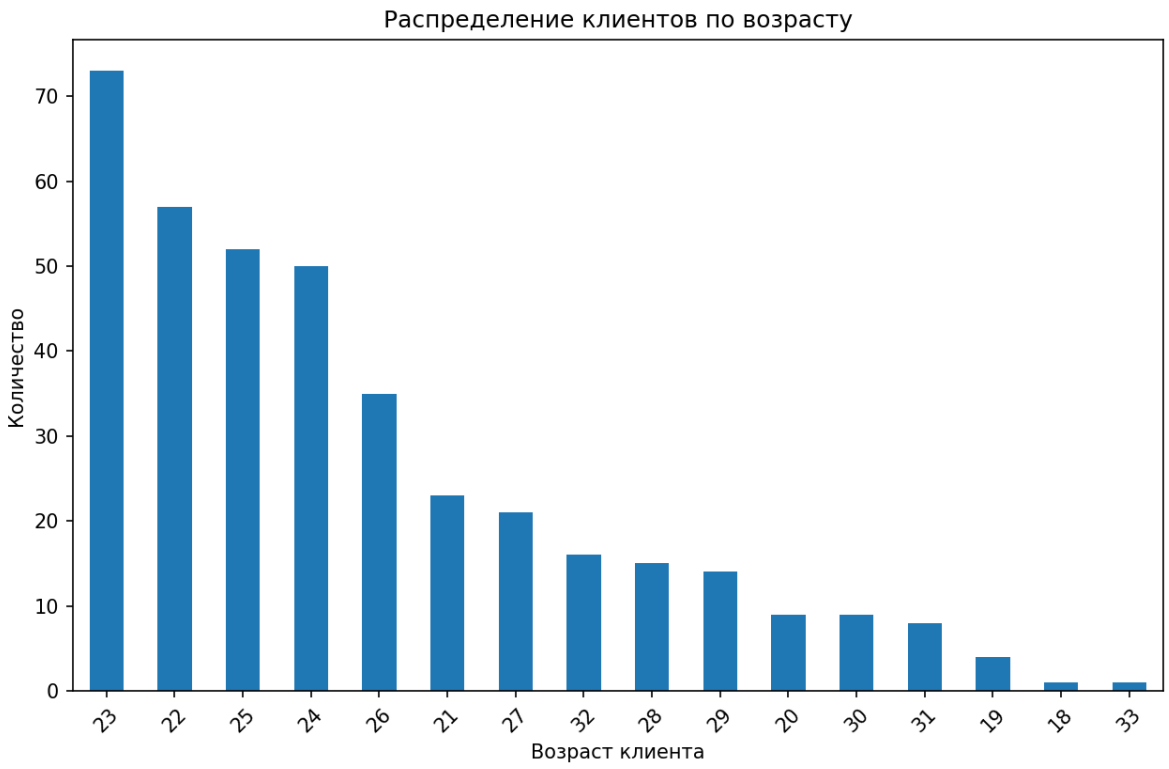
- Сформировать отчет и разместить его в своем репозитории на github.

3. Текст программы

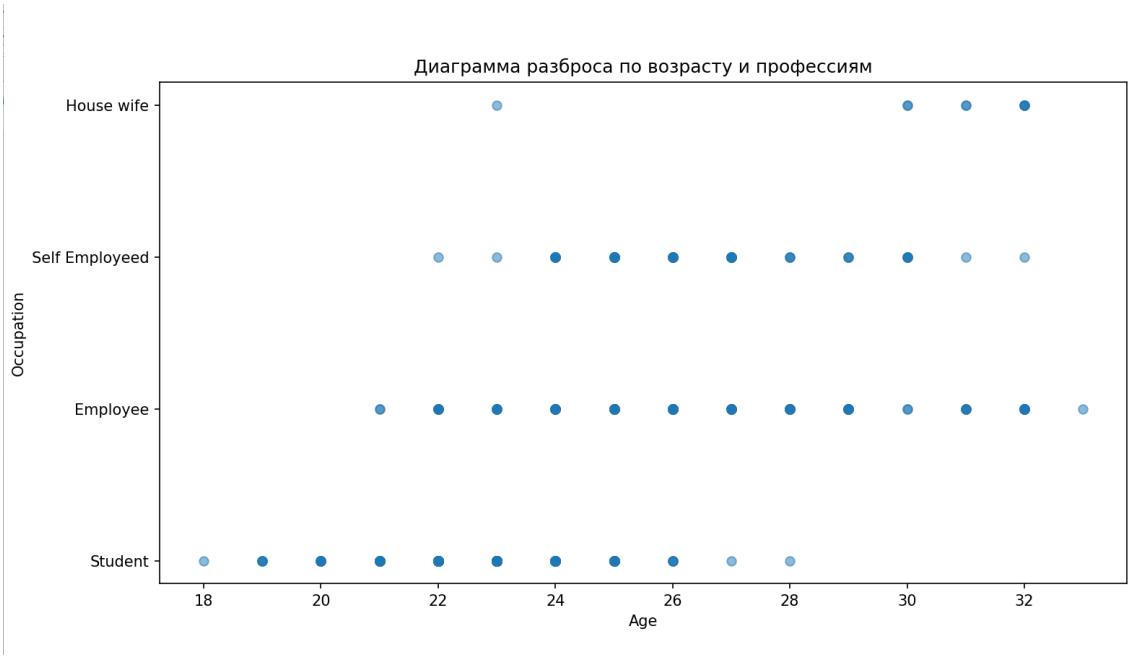
```
import pandas as pd
import matplotlib.pyplot as plt

# Загрузка данных
data = pd.read_csv('onlinefoods.csv')

# Просмотр первых нескольких строк
print(data.head())

# Основные статистические характеристики
print(data.describe())

# Проверка наличия пропущенных значений
print(data.isnull().sum())

plt.figure(figsize=(10, 6))
data['Age'].value_counts().plot(kind='bar')
plt.title('Распределение клиентов по возрасту')
plt.xlabel('Возраст клиента')
plt.ylabel('Количество')
plt.xticks(rotation=45)
plt.show()

plt.figure(figsize=(8, 6))
plt.scatter(data['Age'], data['Occupation'], alpha=0.5)
plt.title('Диаграмма разброса по возрасту и профессиям')
plt.xlabel('Age')
plt.ylabel('Occupation')
plt.show()

plt.figure(figsize=(10, 6))
data.boxplot(column='Age', by='Monthly Income', rot=45)
plt.title('Ящик с усами для возраста в зависимости от ежемесячного
дохода')
```

```
plt.xlabel('Monthly Income')
plt.ylabel('Age')
plt.show()

category_counts = data['Age'].value_counts()

# Построение круговой диаграммы
plt.figure(figsize=(8, 8))
plt.pie(category_counts, labels=category_counts.index,
autopct='%1.1f%%', startangle=140)
plt.title('Distribution of Products by age')
plt.show()
```

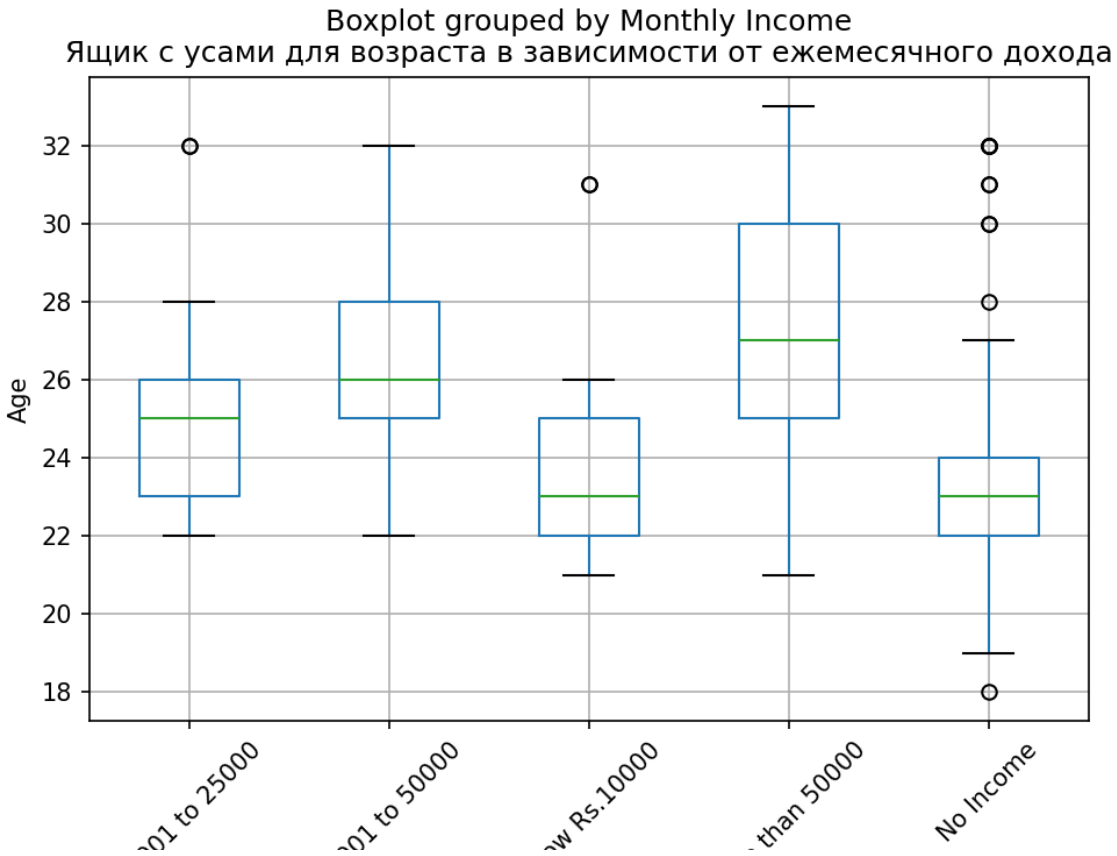
4. Экранные формы с примерами выполнения программы



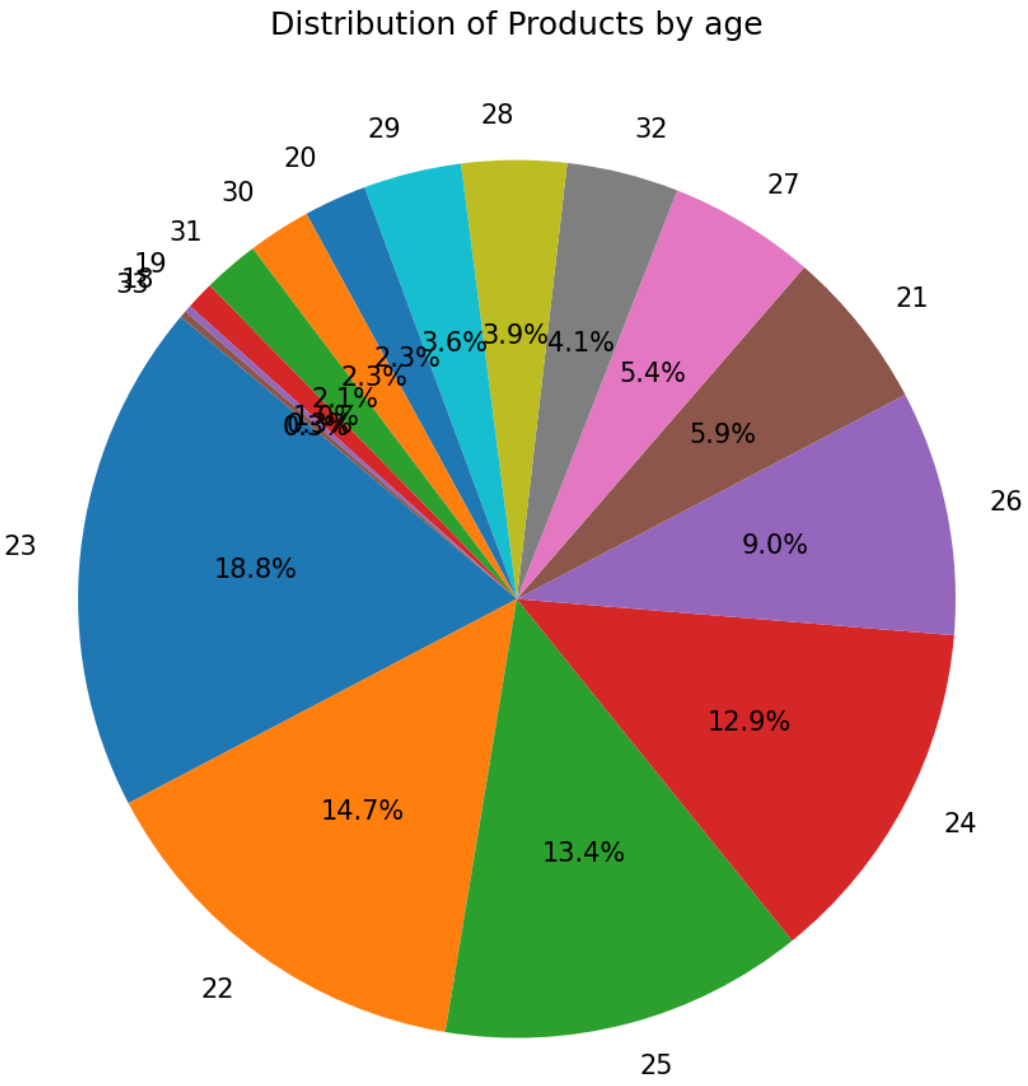
Интерпретация: гистограмма показывает, что 23-летние едят больше всех.



Интерпретация: диаграммы показывают, что нет значительной корреляции между возрастом клиентов и их родом занятий.



Интерпретация: как видно из графика, доля людей, не имеющих дохода, относительно невелика.



Интерпретация: как видно из графика, блюда в основном предназначены для молодых людей.

Список литературы

[1] Гапанюк Ю. Е. LAB_MMO__DATA_STORYЛабораторная работа №1Создание "истории о данных" (Data Storytelling)// GitHub. — 2024. — Режим доступа:https://github.com/ugapanyuk/courses_current/wiki/LAB_MMO__DATA_STORY#%D0%BB%D0%B0%D0%B1%D0%BE%D1%80%D0%B0%D1%82%D0%BE%D1%80%D0%BD%D0%B0%D1%8F-%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D0%B0-1

[2] https://www.kaggle.com/datasets