Московский государственный технический университет им. Н. Э. Баумана

Курс «Технологии машинного обучения»

Отчёт по лабораторной работе №1

Выполнил:
Дувакин А.В.
группа ИУ5-63Б

Проверил:
Гапанюк Ю.Е.

Дата: 14.02.25

Подпись:

Дата:

Подпись:

Москва, 2025 г.

Цель лабораторной работы: изучение различных методов визуализация данных.

Краткое описание. Построение основных графиков, входящих в этап разведочного анализа данных.

Рекомендуемые инструментальные средства можно посмотреть здесь.

Задание:

- Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь.
- Для первой лабораторной работы рекомендуется использовать датасет без пропусков в данных, например из Scikit-learn.
- Пример преобразования датасетов Scikit-learn в Pandas Dataframe можно посмотреть здесь.

Для лабораторных работ не рекомендуется выбирать датасеты большого размера.

- Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного Вами набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета.
4. Информация о корреляции признаков.

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

Ход выполнения:

Загрузка и первичный анализ

```
[ ] import pandas as pd
    import matplotlib.pyplot as plt
    from collections import Counter
    import ast

    df = pd.read_csv("/dataset.csv")
    df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48564 entries, 0 to 48563
Data columns (total 15 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Ids                 48564 non-null  int64
 1   Employer            48564 non-null  object
 2   Name                48564 non-null  object
 3   Salary              48564 non-null  bool
 4   From                15399 non-null  float64
 5   To                  10276 non-null  float64
 6   Experience          48564 non-null  object
 7   Schedule            48564 non-null  object
 8   Keys                48564 non-null  object
 9   Description         48564 non-null  object
 10  Area                48564 non-null  object
 11  Professional roles  48564 non-null  object
 12  Specializations     48564 non-null  object
 13  Profarea names      48564 non-null  object
 14  Published at        48564 non-null  object
dtypes: bool(1), float64(2), int64(1), object(11)
memory usage: 5.2+ MB
```

```
[ ] df.head()
```

	Ids	Employer	Name	Salary	From	To	Experience	Schedule	Keys	Description	Area	Professional roles	Specializat
0	49313809	Space307	Golang Developer (Kimp)	True	251322.0	NaN	От 3 до 6 лет	Полный день	['Docker', 'Golang', 'Redis', 'Английский язык...	Мы в Space307 разрабатываем международную	Санкт-Петербург	['Программист, разработчик']	['Программиров Разраб

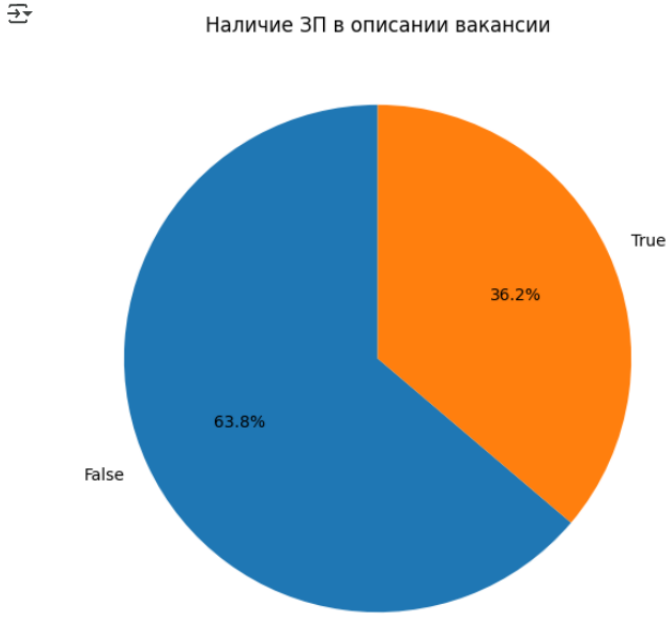
Анализ наличия ЗП в вакансии

```
[ ] value_counts = df['Salary'].value_counts()

    plt.figure(figsize=(7, 7))
    value_counts.plot(kind='pie', autopct='%1.1f%%', startangle=90)

    plt.title('Наличие ЗП в описании вакансии')

    plt.ylabel('')
    plt.show()
```



Определение корреляции между максимальной заработной платой и требуемым опытом.

```
[ ]  experience_order = ['Нет опыта', 'От 1 года до 3 лет', 'От 3 до 6 лет', 'Более 6 лет']
     df['Experience'] = pd.Categorical(df['Experience'], categories=experience_order, ordered=True)

     plt.figure(figsize=(8, 6))
     df.boxplot(column='To', by='Experience', grid=False)

     plt.title('ЗП для различного уровня опыта')
     plt.suptitle('')
     plt.xlabel('Уровень опыта')
     plt.ylabel('Максимальная ЗП (руб)')

     plt.tight_layout()
     plt.show()
```

```
<Figure size 800x600 with 0 Axes>
```