МИНИСТЕРСТВО ЦИФРОВОГО РАЗВИТИЯ, СВЯЗИ И МАССОВЫХ КОММУНИКАЦИЙ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Сибирский государственный университет телекоммуникаций и информатики»

Кафедра телекоммуникационных систем и вычислительных средств (ТС и ВС)

Отчет
по дисциплине
«Системы искусственного интеллекта»

по теме:
ЛАБОРАТОРНАЯ РАБОТА 2. ВИЗУАЛИЗАЦИЯ ДАННЫХ

Студент:
Группа ИА331 *Р.К. Рубцов*

Предподаватель:
. *К.И. Брагин*

Новосибирск 2026 г.

СОДЕРЖАНИЕ

1 ВИЗУАЛИЗАЦИЯ ДАННЫХ

1.1 Цель работы

Изучение программных средств для визуализации наборов данных, освоение основных типов графиков библиотек Matplotlib и Seaborn, а также получение навыков анализа данных на основе их графического представления.

1.2 Описание набора данных

В работе использован датасет `name_gender_dataset.csv`, содержащий 147 269 записей о популярных именах. Для каждого имени указаны пол, количество носителей и вероятность встречаемости.

1.2.1 Назначение набора данных

Набор данных предназначен для обучения методам анализа и визуализации информации. На его основе можно решать следующие задачи:

- анализ распределения имён по полу;
- исследование статистики популярности имён;
- выявление зависимостей между количественными признаками;
- анализ распределения частот;
- построение моделей классификации пола по имени.

1.2.2 Описание признаков

Таблица 1 — Характеристики признаков датасета

Признак	Описание
Name	Имя (категориальный, строковый)
Gender	Пол: M (мужской), F (женский)
Count	Количество носителей имени (целочисленный)
Probability	Вероятность встречаемости (вещественный)

1.3 Общая статистика набора данных

- Количество объектов — 147 269
- Количество признаков — 4
- Мужские имена — 57 520 (39%)
- Женские имена — 89 749 (61%)
- Пропущенные значения отсутствуют

1.3.1 Основные статистические показатели

Таблица 2 — Статистики количественных признаков

Признак	Среднее	Минимум	Максимум
Count	$\approx 2\,481$	1	5 304 407
Probability	$\approx 6.8 \times 10^{-6}$	2.7×10^{-9}	1.45×10^{-2}

1.3.2 Наиболее популярные имена

Таблица 3 — Топ популярных имён

Имя	Пол	Количество
James	M	5 304 407
John	M	5 260 831
Robert	M	4 970 386
Michael	M	4 579 950
William	M	4 226 608
Mary	F	4 169 663

1.4 Программная реализация

Для анализа данных использовались библиотеки Python: NumPy, Pandas, Matplotlib и Seaborn.

Листинг 1.1 — Листинг кода

```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns

plt.style.use('seaborn-v0_8-darkgrid')
sns.set_palette("husl")

data_path = "name_gender_dataset.csv"
data = pd.read_csv(data_path)

print(data.head(10))
print(data.info())
print(data.describe())
print(data.isnull().sum())

data.plot.scatter(x='Count', y='Probability')
plt.xscale('log')
plt.title('      Probability   Count')
plt.grid(alpha=0.3)
plt.tight_layout()
```

```
plt.show()

gender_counts = data['Gender'].value_counts()
gender_counts.plot(kind='bar', color=['pink', 'skyblue'])

plt.title('           ')
plt.xlabel('  ')
plt.ylabel('          ')
plt.tight_layout()
plt.show()

stats_by_gender = data.groupby('Gender')[['Count']].agg(['mean',
   'median', 'std'])

stats_plot = data.groupby('Gender')[['Count']].mean()
stats_plot.plot(kind='bar', color=['skyblue'], alpha=0.8)
plt.title('          Count      ')
plt.xlabel('  ')
plt.ylabel('         ')
plt.legend(title='    ')
plt.grid(axis='y', alpha=0.3)
plt.tight_layout()
plt.show()

stats_by_gender = data.groupby('Gender')[['Probability']].agg(['
   mean', 'median', 'std'])

stats_plot = data.groupby('Gender')[['Probability']].mean()
stats_plot.plot(kind='bar', color=['pink'], alpha=0.8)
plt.title('          Probability      ')
plt.xlabel('  ')
plt.ylabel('         ')
plt.legend(title='    ')
plt.grid(alpha=0.3)
plt.show()

top5_m_names = data[data['Gender'] == 'M'].nlargest(5, 'Count')

top5_m_names.plot.barh(x='Name', y='Count', color='skyblue')
plt.xlabel('Count')
plt.title('  -5                ')
plt.gca().invert_yaxis()
```

```
plt.grid()
plt.show()

top5_f_names = data[data['Gender'] == 'F'].nlargest(5, 'Count')

top5_f_names.plot.barh(x='Name', y='Count', color='pink')
plt.xlabel('Count')
plt.title('  -5                ')
plt.gca().invert_yaxis()
plt.grid()
plt.tight_layout()
plt.show()
```

1.5 Результаты визуализации

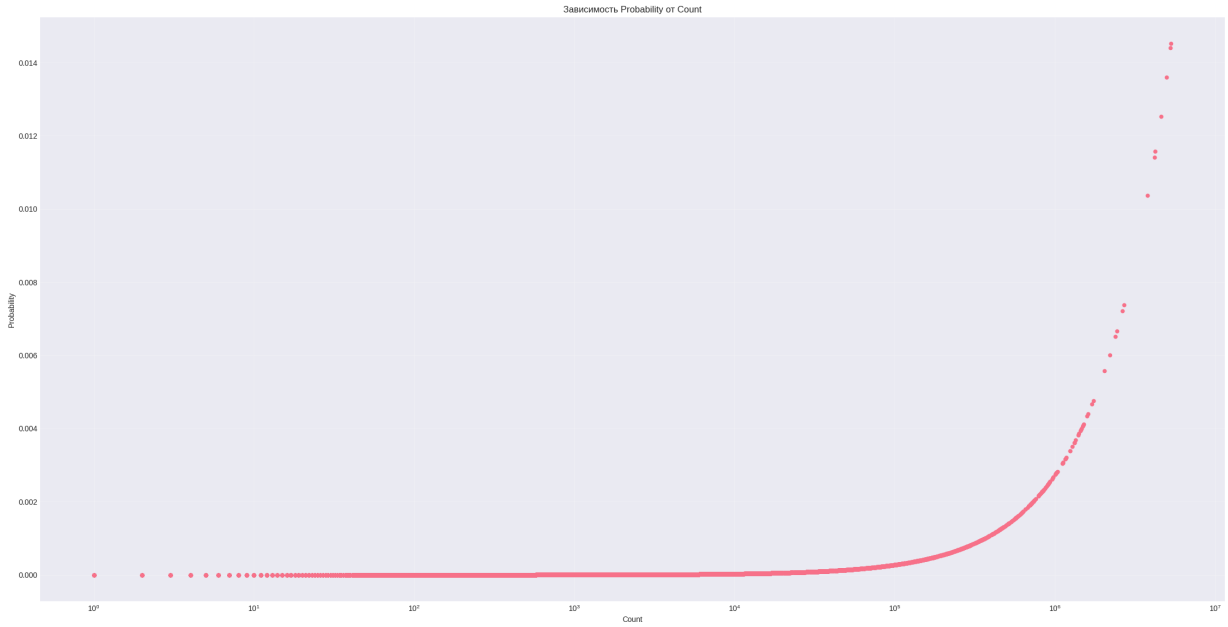

Рисунок 1 — Зависимость Probability от Count

На графике наблюдается выраженная положительная корреляция между количеством носителей имени и вероятностью его встречаемости. Распределение имеет степенной характер.

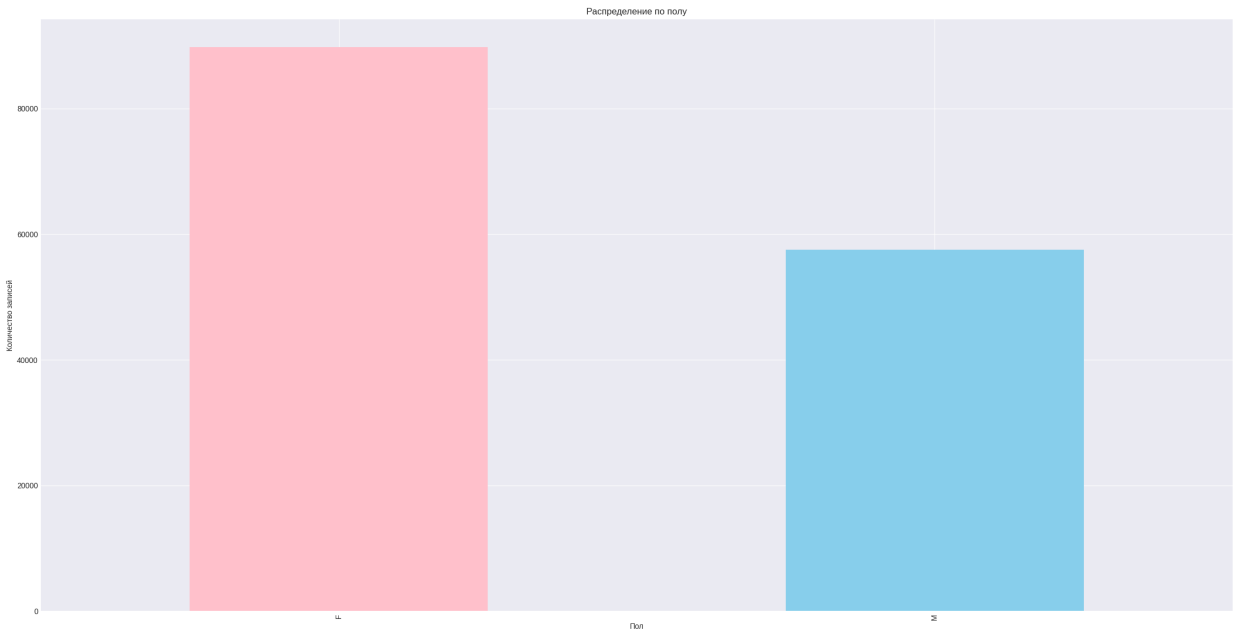

Рисунок 2 — Распределение имён по полу

Женских имён в наборе данных больше, однако мужские имена в среднем имеют более высокие значения показателя Count.

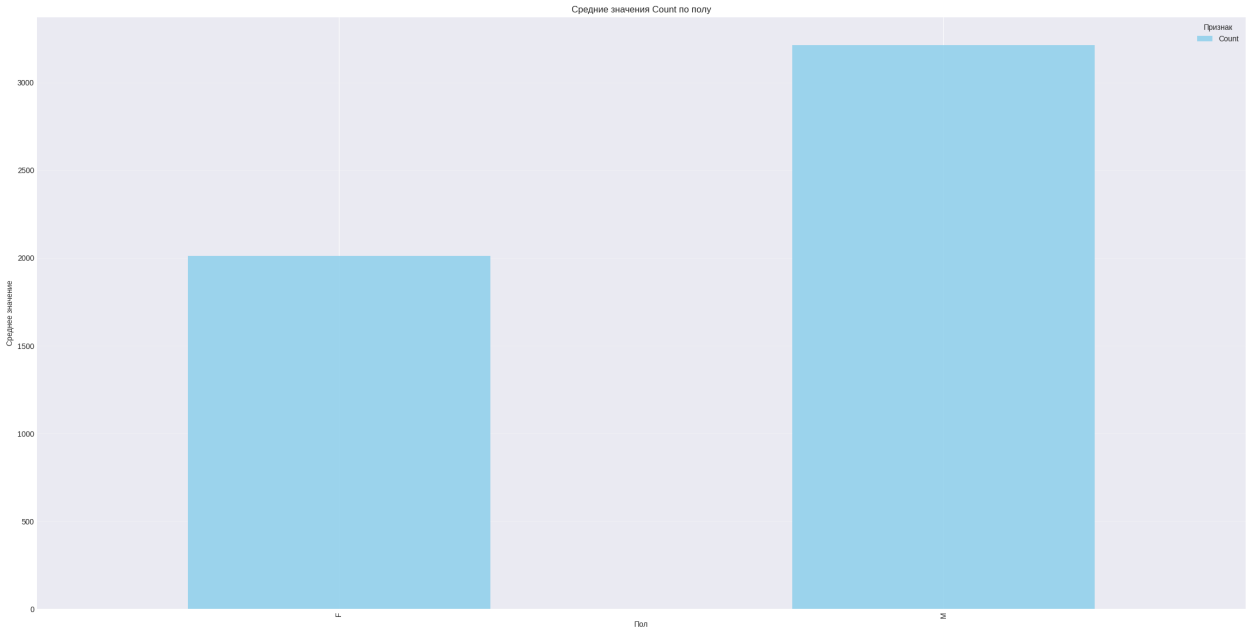

Рисунок 3 — Средние значения Count по полу

Среднее значение количества носителей мужских имён выше по сравнению с женскими, что указывает на большую концентрацию популярности среди мужских имён.

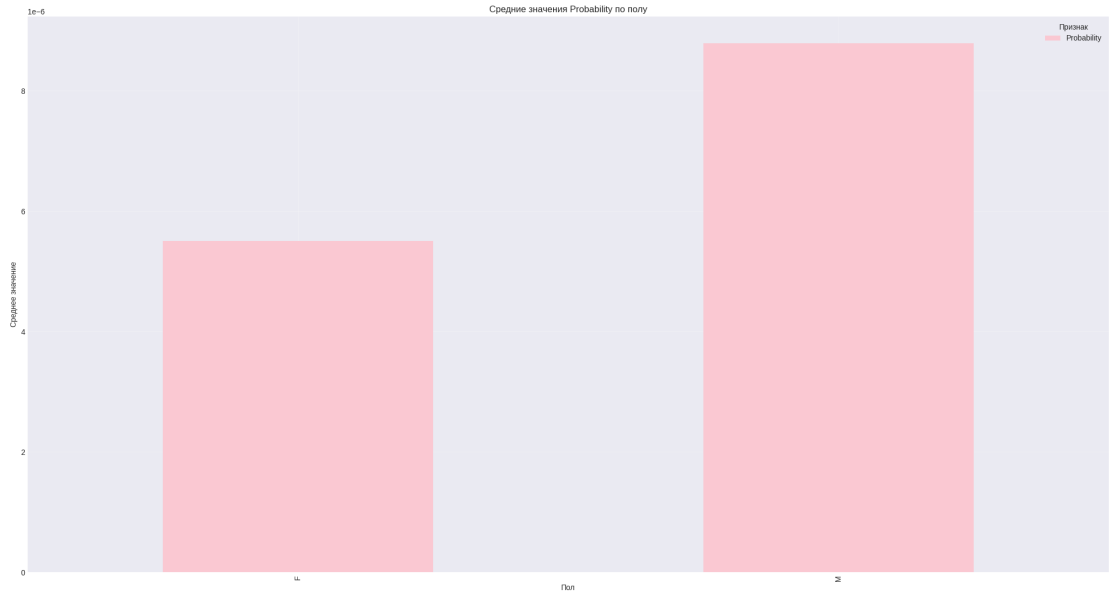

Рисунок 4 — Средние значения Probability по полу

Вероятность встречаемости также в среднем выше у мужских имён, что подтверждает выявленную ранее тенденцию.

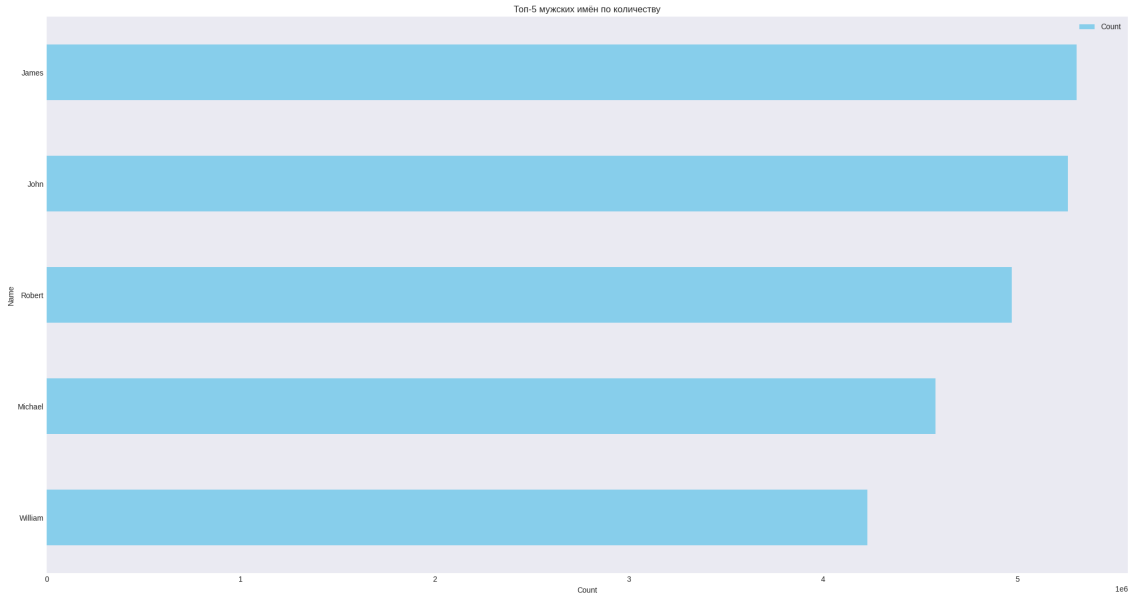

Рисунок 5 — Топ-5 мужских имён по количеству носителей

Диаграмма демонстрирует наиболее популярные мужские имена с максимальным значением Count.

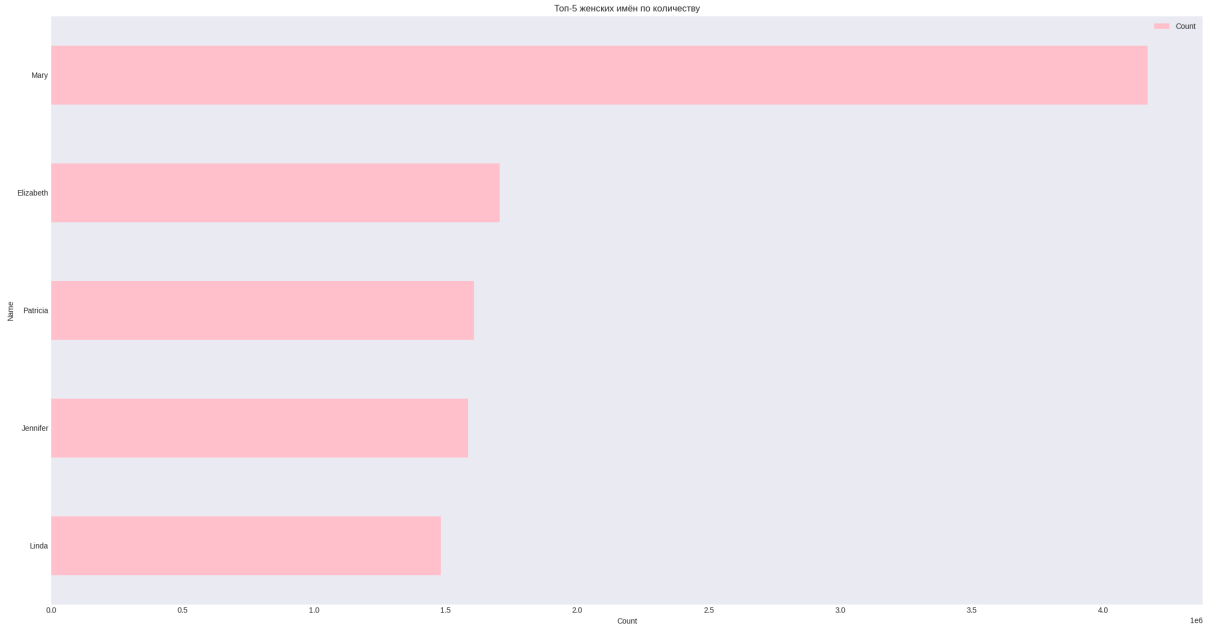

Рисунок 6 — Топ-5 женских имён по количеству носителей

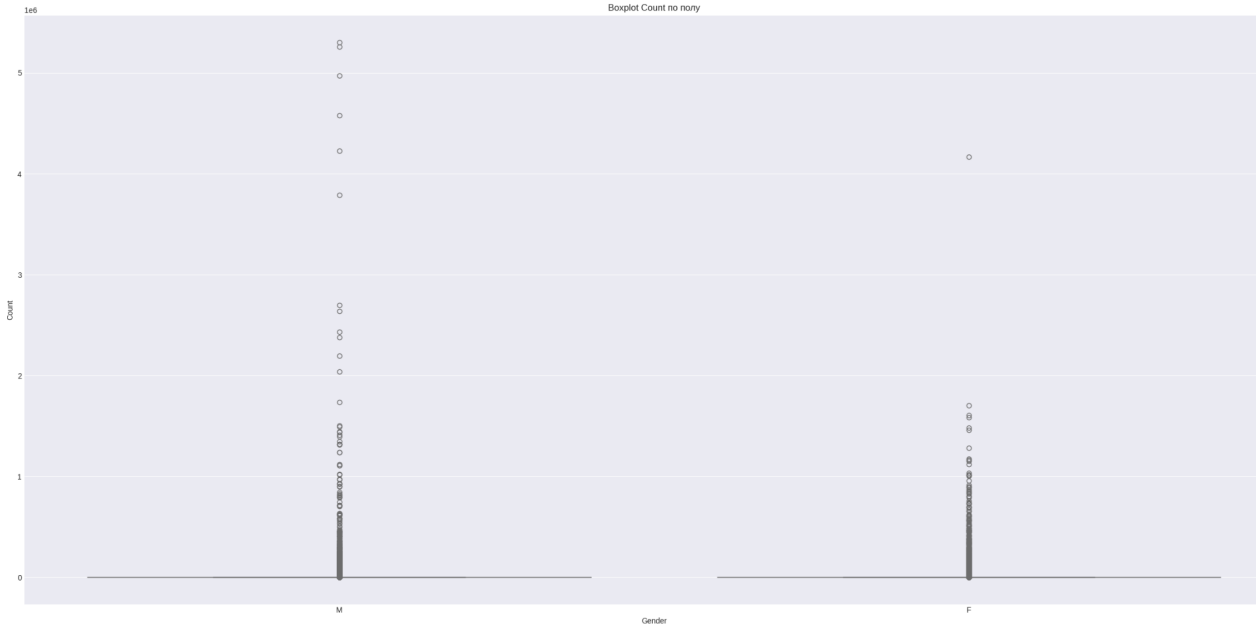

Рисунок 7 — Boxplot распределения Count по полу

На диаграмме размаха видно, что распределение количества носителей имён (Count) имеет выраженную асимметрию и большое количество выбросов.

У мужских имён наблюдаются более высокие максимальные значения по сравнению с женскими, что указывает на наличие более популярных мужских имён. При этом основная масса значений сосредоточена в

нижней части диапазона, что говорит о том, что большинство имён имеют относительно небольшое количество носителей.

Представлены самые популярные женские имена по количеству носителей.

1.6 Ответы на контрольные вопросы

1. Инструментальные средства Data Science

Jupyter Notebook, Google Colab, VS Code, PyCharm, Anaconda, Git, Docker.

2. Библиотеки машинного обучения

- NumPy — работа с массивами и линейной алгеброй.
- Pandas — обработка и анализ табличных данных.
- Matplotlib — базовая визуализация.
- Seaborn — статистическая визуализация.
- Scikit-learn — алгоритмы машинного обучения.
- SciPy — научные вычисления.

3. Причины популярности Python

Простота синтаксиса, развитая экосистема библиотек, активное сообщество, удобство прототипирования.

4. Функции визуализации

plt.scatter(), plt.bar(), plt.barh(), sns.countplot(), sns.heatmap(), plt.xscale(), plt.title() и др.

5. Библиотека для работы с наборами данных

Pandas.

6. Нежелательная стратегия обработки пропусков

Полное удаление строк с пропусками, так как это может привести к потере значительной части информации.

7. Нужно ли применять OneHotEncoder к целевой переменной?

Нет. Для целевой переменной используется LabelEncoder.

8. Разбиение выборки

Оптимальные соотношения: 20:80 или 25:75.

9. Загрузка CSV-файла

```
dataset = pd.read_csv("data.csv")
```

1.7 Выводы

В ходе выполнения лабораторной работы были изучены методы визуализации данных и проведён анализ распределения популярности имён.

Выявлена положительная корреляция между количеством носителей имени и вероятностью его встречаемости. Распределение носит степенной характер. Мужские имена в среднем обладают более высокой популярностью, несмотря на меньшее количество уникальных записей.

Полученные навыки визуализации позволяют эффективно проводить предварительный анализ данных перед построением моделей машинного обучения.