Министерство образования Республики Беларусь
Учреждение образования
«Брестский Государственный технический университет»
Кафедра ИИТ

Лабораторная работа №1
По дисциплине: «Основы машинного обучения»
Тема: «Знакомство с анализом данных: предварительная обработка и визуализация»

Выполнил:
Студент 2 курса
Группы АС-66
Колбашко А. В.
Проверил:
Крощенко А. А.

Брест 2025

Цель работы: получить практические навыки работы с данными с использованием библиотек Pandas для манипуляции и Matplotlib для визуализации. Научиться выполнять основные шаги предварительной обработки данных, такие как очистка, нормализация и работа с различными типами признаков.

Ход работы

Общее задание:

1. Загрузить предложенный набор данных (по вариантам) в DataFrame библиотеки Pandas.
2. Провести исследовательский анализ: изучить типы данных, количество пропусков, основные статистические показатели (среднее, медиана, стандартное отклонение).
3. Обработать пропущенные значения (например, заполнить средним значением или удалить строки/столбцы).
4. Преобразовать категориальные признаки в числовые с помощью метода One-Hot Encoding.
5. Выполнить нормализацию или стандартизацию числовых признаков.
6. Построить несколько графиков для визуализации данных (гистограммы, диаграммы рассеяния) и сделать выводы о зависимостях между признаками.
7. **Написать отчет, создать пул-реквест в репозиторий с кодом решения и отчетом в формате pdf.**

Используемые инструменты: Python, Pandas, Matplotlib, NumPy, Jupyter Notebook / Google Colab / PyCharm

Вариант 9

Выборка Melbourne Housing Market. Содержит данные о продажах домов в Мельбурне, включая цену, количество комнат, район и т.д.

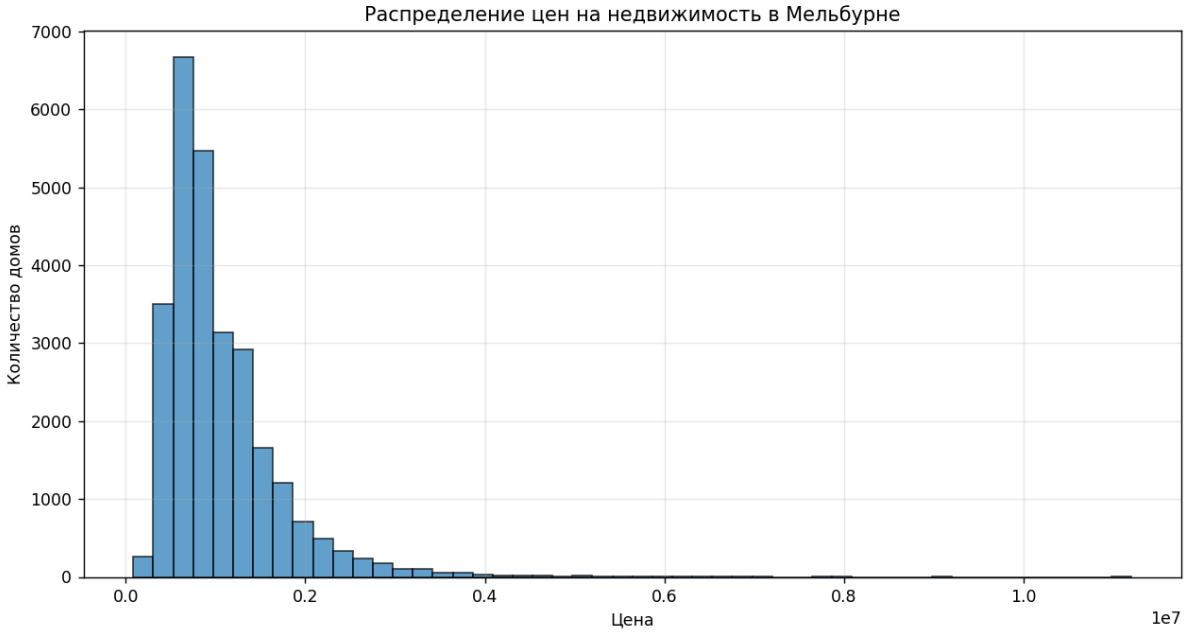
Задачи:

1. Загрузите данные. Найдите столбец с наибольшим количеством пропущенных значений и удалите его.
2. Удалите все строки, где отсутствует значение цены (Price). 3. Постройте гистограмму распределения цен на недвижимость. 4. Рассчитайте среднюю цену за дом для 5 самых популярных пригородов (Suburb).
5. Создайте новый признак PropertyAge на основе года постройки (YearBuilt).
6. Преобразуйте признак Type (тип недвижимости) в числовой формат с помощью One-Hot Encoding.

Код программы:

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

df = pd.read_csv("C:\\Users\\wlksm\\OneDrive\\Рабочий
стол\\Уник\\ОМО\\Kolbashko\\src\\Melbourne.csv")

missing_values = df.isnull().sum()
column_with_most_missing = missing_values.idxmax()
df = df.drop(columns=[column_with_most_missing])
print(f"\nСтолбец с наибольшим количеством пропусков:
{column_with_most_missing} ({missing_values.max()} пропусков)")

df = df.dropna(subset=['Price'])

plt.figure(figsize=(12, 6))
plt.hist(df['Price'], bins=50, edgecolor='black', alpha=0.7)
plt.title('Распределение цен на недвижимость в Мельбурне')
plt.xlabel('Цена')
plt.ylabel('Количество домов')
plt.grid(alpha=0.3)
plt.show()

top_5_suburbs = df['Suburb'].value_counts().head(5).index
print("\n5 самых популярных пригородов:")
print(top_5_suburbs.tolist())

for suburb in top_5_suburbs:
    avg_price = df[df['Suburb'] == suburb]['Price'].mean()
    print(f"{suburb}: средняя цена = {avg_price:,.2f}")

current_year = pd.Timestamp.now().year
df['PropertyAge'] = current_year - df['YearBuilt']
print(f"\nСтатистика по возрасту недвижимости:")
print(df['PropertyAge'].describe())

print(f"\nУникальные значения Type до кодирования:")
print(df['Type'].value_counts())
type_encoded = pd.get_dummies(df['Type'], prefix='Type')
df = pd.concat([df, type_encoded], axis=1)
print(f"\nДобавлены новые столбцы:")
print([col for col in df.columns if col.startswith('Type_')])
```

Диаграммы после выполения программы:



Консольный вывод:

Столбец с наибольшим количеством

пропусков: BuildingArea (21097 пропусков)

5 самых популярных пригородов:

['Reservoir', 'Bentleigh East', 'Richmond',

'Preston', 'Brunswick']

Reservoir: средняя цена = 691,104.48

Bentleigh East: средняя цена = 1,131,418.21

Richmond: средняя цена = 1,067,584.51

Preston: средняя цена = 877,869.85

Brunswick: средняя цена = 977,988.76

Статистика по возрасту недвижимости:

count 12084.000000

mean 58.390847

std 36.762373

min 6.000000

25% 25.000000

50% 55.000000

75% 75.000000

max 829.000000

Name: PropertyAge, dtype: float64

Уникальные значения Type до кодирования:

Type

h 18472

u 5909

t 2866

Name: count, dtype: int64

Добавлены новые столбцы:

['Type_h', 'Type_t', 'Type_u']