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.

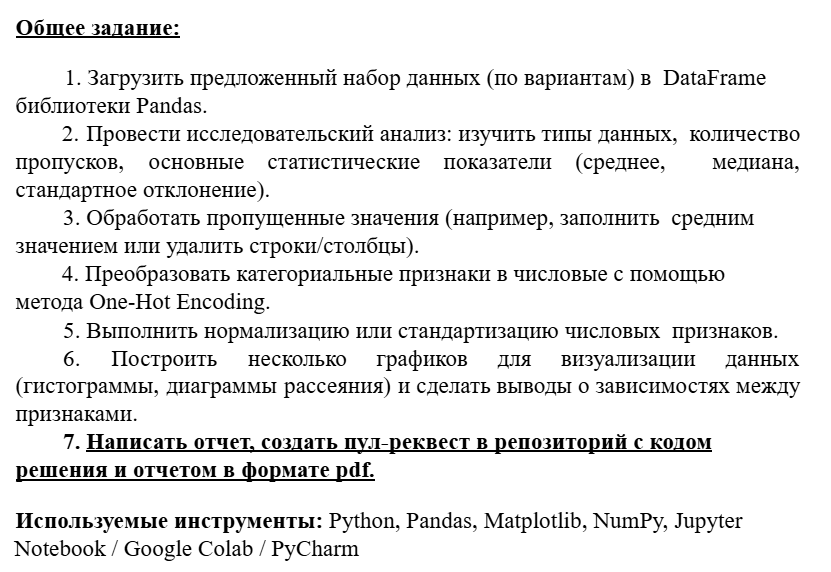
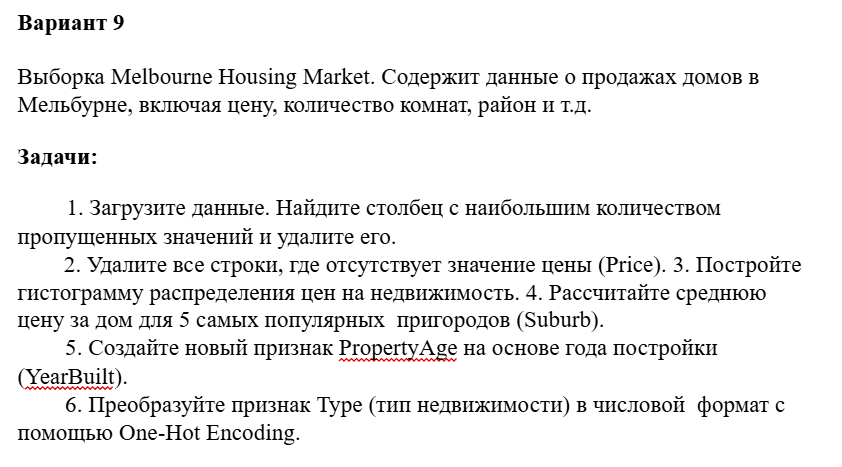
Проверил:

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

Код программы:

import pandas as pd

import matplotlib.pyplot as plt

df = pd.read\_csv("C:\\Users\\wlksm\\OneDrive\\Рабочий стол\\Уник\\OMO\\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())

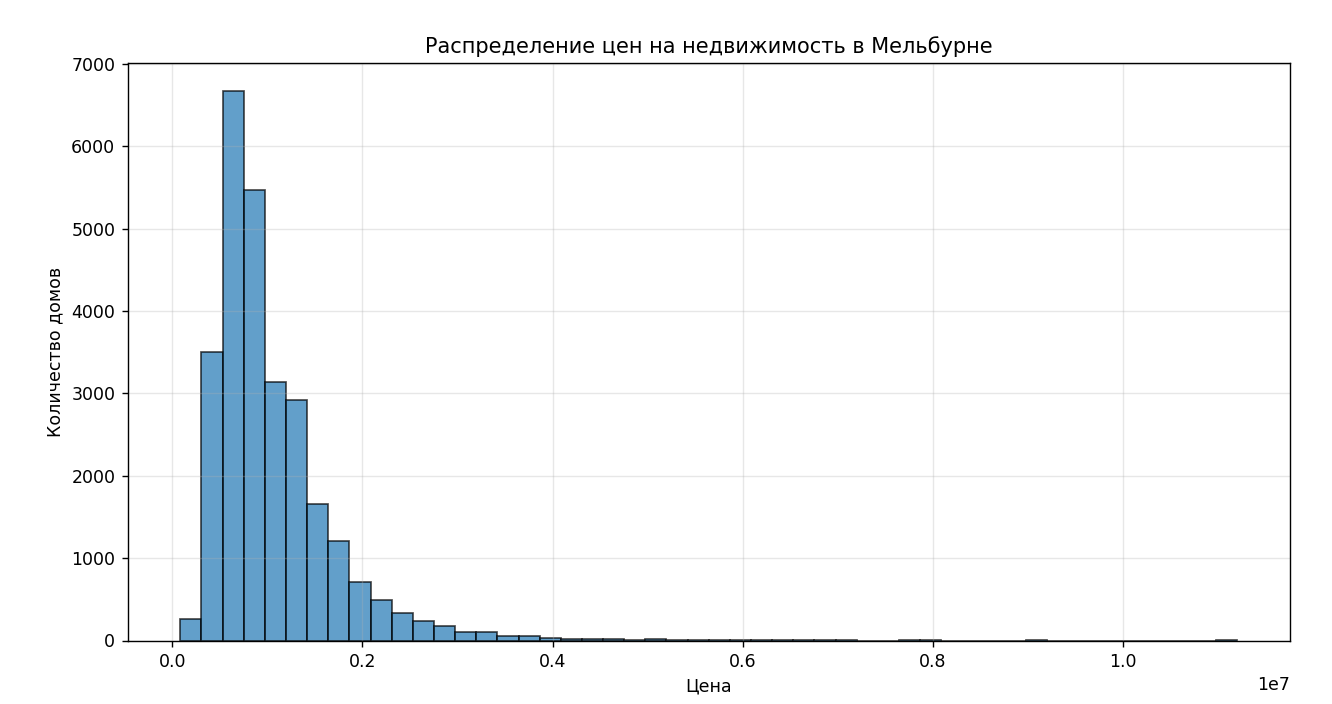
type\_encoded = pd.get\_dummies(df['Type'], prefix='Type')

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']