Министерство науки и образования РФ

Федеральное государственное бюджетное учреждение

высшего образования

**«Тверской государственный технический университет»**

(ТвГТУ)

Кафедра программного обеспечения

**Отчет по лабораторной работе №1**

По дисциплине: «Анализ больших данных»

Тема: «**Реляционные данные. Исследовательский анализ данных.**

**Построение визуализаций данных OLAP**»

|  |
| --- |
| Выполнил:  студент группы  Б.ПИН.РИС - 21.06  Иванов А.В. |
| Проверила:  старший преподаватель  кафедры ПО  Корнеева Е.И. |

Тверь 2025

**Формулировка задач, описание условий**

**Задание на лабораторную работу**

1. Создать базу из файлов или скриптов согласно варианту. Сложность программы выбрать из представленных в тексте лабораторной работы.

2.  Подключиться к базе данных из python.

3.  Сделать описание данных. Из каких таблиц и полей состоят данные таблиц? Какие из данных являются признаками? К какому типу данных и к какой шкале относятся признаки?

4. Одномерный анализ. Построить гистограммы распределения количественных признаков, которые важны для задачи. Сделать вывод после построения. Какое распределение для каждого из признаков? Почему, по вашему мнению, признаки важны для задачи?

5. Многомерный анализ. Построить графики из 3-4 признаков. Выбрать категориальные (номинальные, порядковые или бинарные) признаки и количественные. Что получилось на каждом графике? Почему, по вашему мнению, признаки важны для задачи?

**Вариант и условия задач. Указать какой сложности выполняется задание.**

Вариант 5.

**Сложность: Rare**

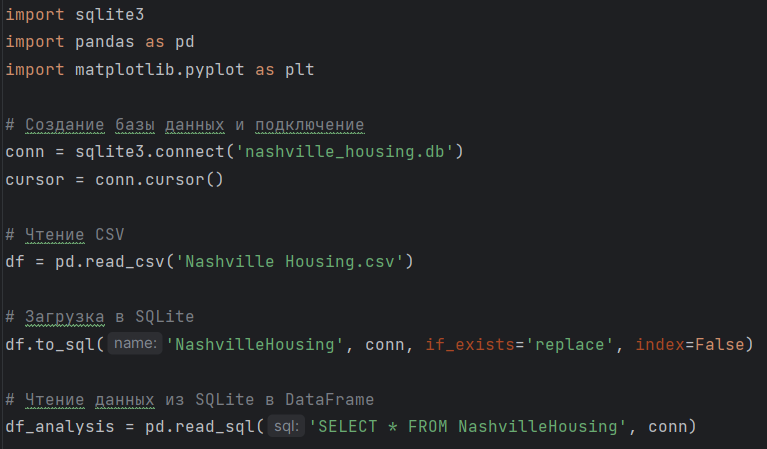
* Реализовать Jupiter Notebook или консольное приложение для выполнения задания.
* В качестве базы данных выбрать sqlite3. Создать базу из файлов или скриптов согласно варианту.
* Подключиться к базе данных из python.
* Сделать описание данных и выводы по заданию.
* Соединить признаки в 1 таблицу pandas для анализа
* Одномерный анализ. Построить 2 гистограммы распределения количественных признаков, которые важны для задачи и сделать их описание по заданию.
* Многомерный анализ. Построить хотя бы 1 график из 3-4 признаков и сделать его описание по заданию.

**Ссылка на репозиторий с программной реализацией**

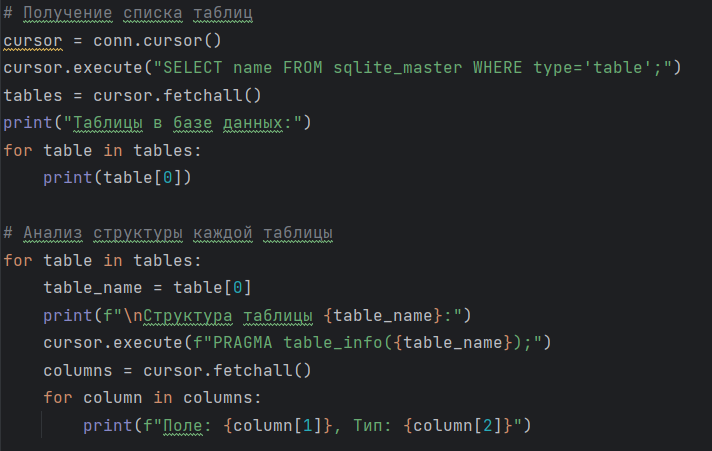
https://github.com/ArtyomI69/data\_science

# Описание проделанной работы

## 1. Создание базы данных SQLite3 и подключение из Python



## 2. Описание данных

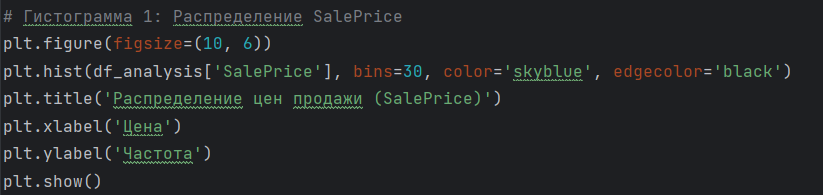


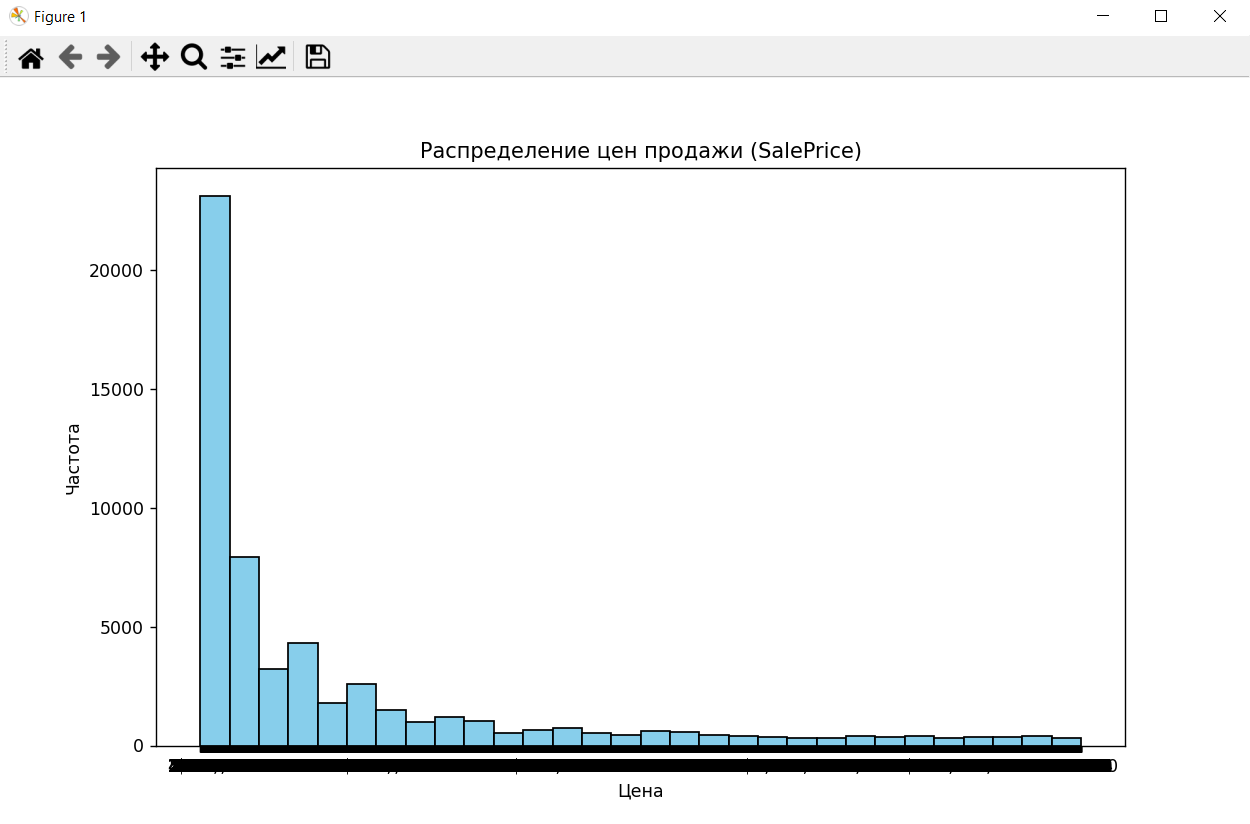
**Признаки, важные для задачи прогнозирования стоимости:**

* SalePrice (целевая переменная).
* YearBuilt, Acreage, Bedrooms, LandValue, BuildingValue.

## 3. Одномерный анализ

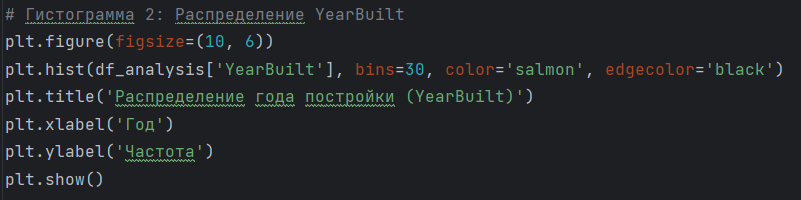
Распределение SalePrice

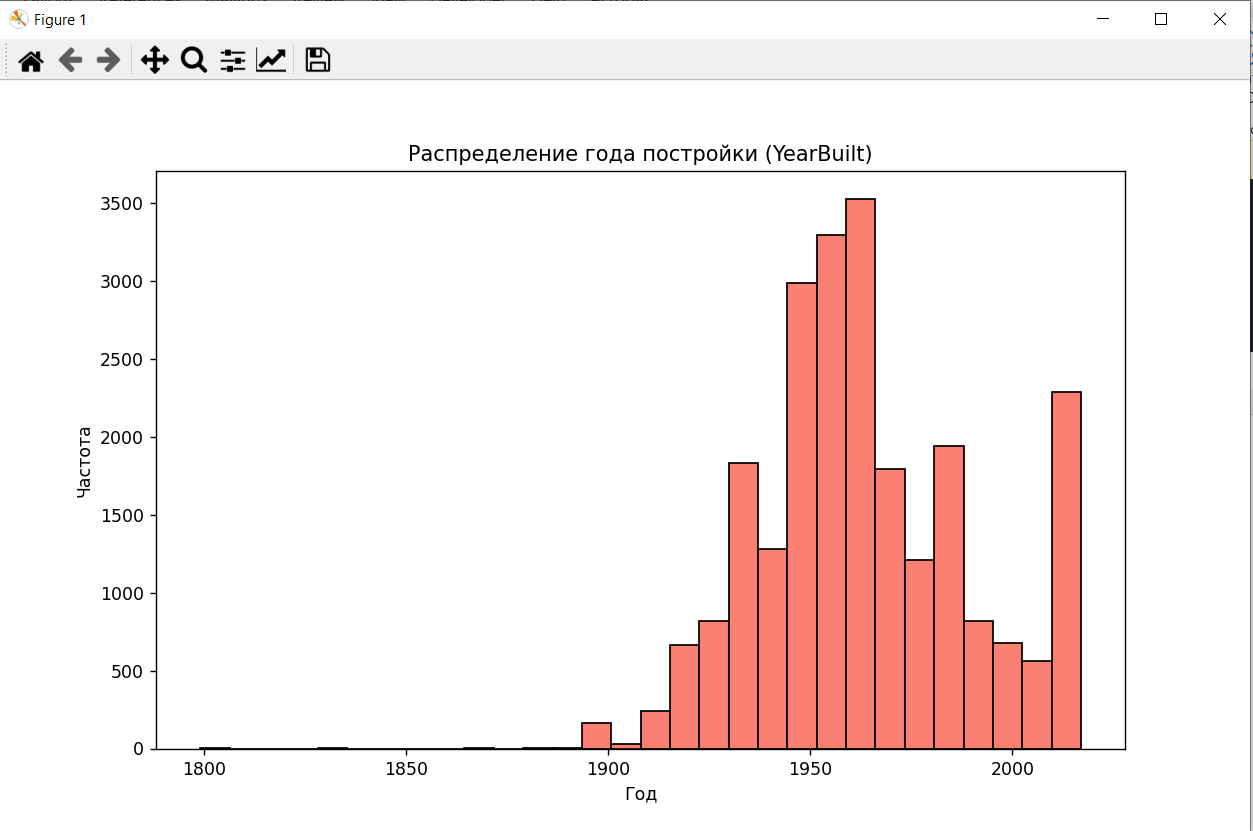




Вывод: Распределение правостороннее с выбросами в высоком ценовом диапазоне. Большинство объектов продаются в диапазоне до $500,000.

Распределение YearBuilt

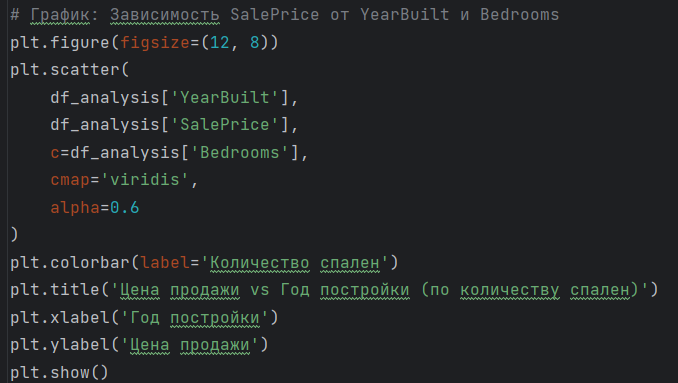


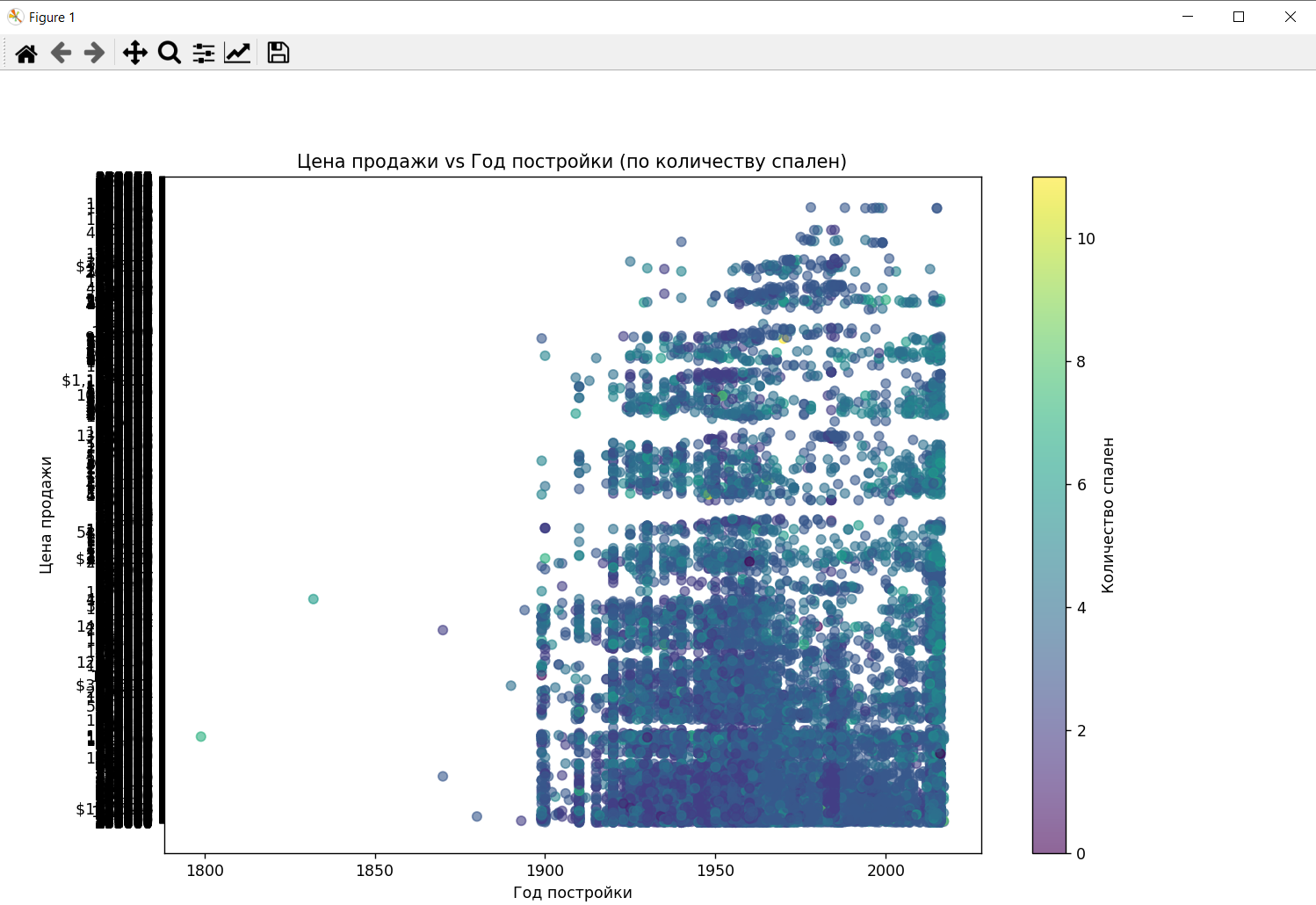


Вывод: Большинство домов построены после 1970 года. Пик строительства приходится на 1980-2000 годы.

## 4. Многомерный анализ

Зависимость SalePrice от YearBuilt и Bedrooms





**Вывод:**

* Более новые дома (после 2000 года) имеют более высокую стоимость.
* Дома с большим количеством спален (3-5) концентрируются в верхнем ценовом диапазоне.
* Видна положительная корреляция между годом постройки и ценой.

**Краткий вывод по работе. Описание реализованной программы и её функций.**

В данной лабораторной работе мы перенесли таблицу из csv в sqlite3.

Нашли признаки.

И сделали одномерный и многомерный анализ.

* **SalePrice** и **YearBuilt** — ключевые признаки для анализа стоимости недвижимости.
* Многомерный анализ показывает, что новые дома с большим количеством спален имеют более высокую стоимость.

**Ссылки на используемые материалы. Документация**

Документация Python sqlite3: [**https://docs.python.org/3/library/sqlite3.html**](https://docs.python.org/3/library/sqlite3.html)

Pandas: [**https://pandas.pydata.org/docs**](https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html)

Matplotlib : [**https://matplotlib.org**](https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.hist.html)

**Листинг кода**

import sqlite3

import pandas as pd

import matplotlib.pyplot as plt

# Создание базы данных и подключение

conn = sqlite3.connect('nashville\_housing.db')

cursor = conn.cursor()

# Чтение CSV

df = pd.read\_csv('Nashville Housing.csv')

# Загрузка в SQLite

df.to\_sql('NashvilleHousing', conn, if\_exists='replace', index=False)

# Чтение данных из SQLite в DataFrame

df\_analysis = pd.read\_sql('SELECT \* FROM NashvilleHousing', conn)

# Получение списка таблиц

cursor = conn.cursor()

cursor.execute("SELECT name FROM sqlite\_master WHERE type='table';")

tables = cursor.fetchall()

print("Таблицы в базе данных:")

for table in tables:

print(table[0])

# Анализ структуры каждой таблицы

for table in tables:

table\_name = table[0]

print(f"\nСтруктура таблицы {table\_name}:")

cursor.execute(f"PRAGMA table\_info({table\_name});")

columns = cursor.fetchall()

for column in columns:

print(f"Поле: {column[1]}, Тип: {column[2]}")

# Одномерный анализ

# Гистограмма 1: Распределение SalePrice

plt.figure(figsize=(10, 6))

plt.hist(df\_analysis['SalePrice'], bins=30, color='skyblue', edgecolor='black')

plt.title('Распределение цен продажи (SalePrice)')

plt.xlabel('Цена')

plt.ylabel('Частота')

plt.show()

# Гистограмма 2: Распределение YearBuilt

plt.figure(figsize=(10, 6))

plt.hist(df\_analysis['YearBuilt'], bins=30, color='salmon', edgecolor='black')

plt.title('Распределение года постройки (YearBuilt)')

plt.xlabel('Год')

plt.ylabel('Частота')

plt.show()

# Многомерный анализ

# График: Зависимость SalePrice от YearBuilt и Bedrooms

plt.figure(figsize=(12, 8))

plt.scatter(

df\_analysis['YearBuilt'],

df\_analysis['SalePrice'],

c=df\_analysis['Bedrooms'],

cmap='viridis',

alpha=0.6

)

plt.colorbar(label='Количество спален')

plt.title('Цена продажи vs Год постройки (по количеству спален)')

plt.xlabel('Год постройки')

plt.ylabel('Цена продажи')

plt.show()