Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования

ИРКУТСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

|  |
| --- |
| Институт информационных технологий  и анализа данных |
| наименование института |

**ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ №4**

по дисциплине «Технологии обработки, анализа и визуализации данных»

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Выполнил студент |  | ИИТм-23-1 |  |  |  | Солопов Д.Д. |
|  |  | шифр группы |  | подпись |  | И.О. Фамилия |
| Проверил |  |  |  |  |  | Григорьев С.В. |
|  |  |  |  | подпись |  | И.О. Фамилия |

Отчет защищен с оценкой \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Иркутск 2024 г

**Содержание**

[Введение 3](#_Toc184071610)

[1 Анализ данных с помощью метода многомерной регрессии 4](#_Toc184071611)

Введение

**Цель работы** заключается в изучении способов применения методов многомерного анализа данных из заданного набора, а также известных алгоритмов на их основе, для решения предсказательных задач, в том числе за счёт построения и использования моделей.

**Задачи**:

1. Сформулировать задачу для анализа данных из данного набора и решить ее методом многомерной линейной регрессии;
2. Сформулировать задачу для анализа данных из данного набора и решить ее методом выявления аномалий;
3. Сформулировать задачу для анализа данных из данного набора и решить ее методом иерархической кластеризации;
4. Сформулировать задачу для анализа данных из данного набора и решить ее с помощью наивного байесовского классификатора.

1 Анализ данных с помощью метода многомерной регрессии

**Сформулированная задача для анализа данных**: предсказать цену недвижимости за единицу площади в зависимости от возраста недвижимости, расстояния до ближайшей станции метро и количества круглосуточных магазинов, расположенных поблизости.

Для начала импортируем необходимые библиотеки для последующего использования в среде Google Colab:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import hvplot.pandas

from sklearn.model\_selection import train\_test\_split

from sklearn import metrics

from sklearn.metrics import r2\_score

from sklearn.linear\_model import LinearRegression

%matplotlib inline

Теперь загрузим и предварительно ознакомимся с датасетом:

# Загрузка набора данных

df = pd.read\_csv("/content/Real Estate Dataset.csv")

# Просмотр основной информации о фрейме

df.info()

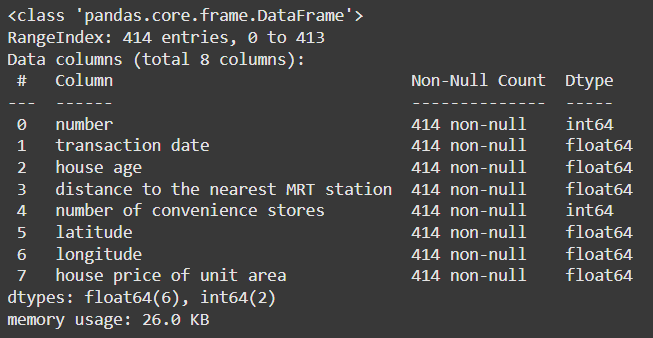


Рисунок 1 – Основная информация о наборе данных

Набор данных содержит информацию о цене недвижимости за единицу площади, которая обладает определёнными характеристиками, влияющими на эту цену, в том числе:

1. **house age** – число лет недвижимости;
2. **distance to the nearest MRT station** – расстояние до ближайшей станции метро;
3. **number of convenience stores** – количество расположенных по близости круглосуточных магазинов

В наборе данных находятся только числовые и вещественные значения, поэтому никаких дополнительных преобразований не требуется. Количественные значения определены с помощью типа данных int64, а вещественные – с помощью float64.

Определим выбросы в исследуемом наборе данных с помощью диаграммы “ящик с усами” (boxplot), предварительно построив график QQ-Plot (квантиль-квантиль график) с помощью следующего программного кода:

import statsmodels.api as sm

import pylab as py

import scipy.stats as stats

Y\_label = 'house price of unit area'

#QQ-Plot

stats.probplot(df[Y\_label], dist='norm', plot=py)

py.show()

**QQ-plot (квантиль-квантиль график**) — это инструмент, который используется для оценки сходства распределения одной числовой переменной с нормальным распределением или между распределениями двух числовых переменных.

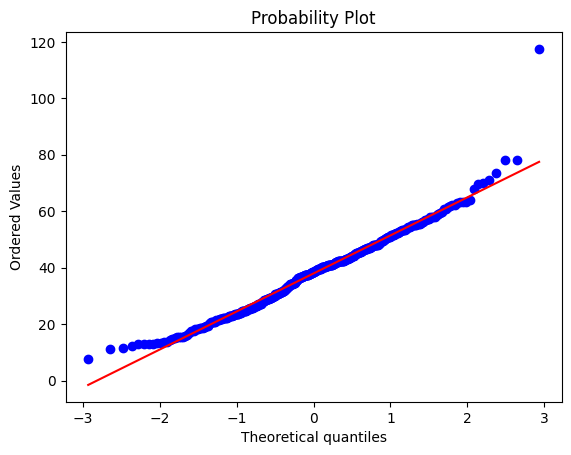


Рисунок 2 – Результат построения квантиль-квантиль графика по цене на недвижимость за единицу площади

Как видно из квантиль-квантиль графика имеются значения со слишком большим отклонением от линии лучшего соответствия (обозначена красным цветом).

Теперь построим диаграмму boxplot, с помощью следующего программного кода:

df[[Y\_label]].boxplot(figsize=(13, 8))

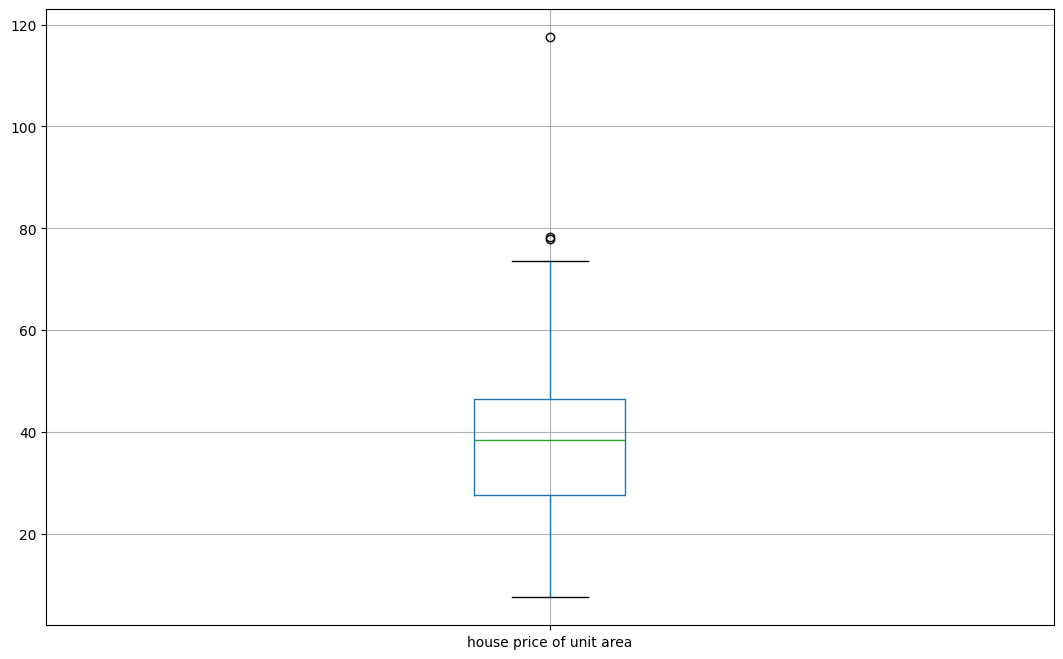


Рисунок 3 – Диаграмма boxplot для цены на недвижимость за единицу площади

Как видно из диаграммы boxplot по цене за единицу площади имеются выбросы, которые могут в дальнейшем повлиять на предсказание цены. Необходимо удалить записи с такими выбросами.

Для удаления выбросов будет использован метод межквартильного размаха (IQR). Вычислим границы IQR и отсеим выбросы на основе этих границ.

определяется как среднее число между наименьшим числом и медианой набора данных

Для начала определим среднее число между наименьшим числом и медианной набора данных (Q1) и среднее значение между медианой и самым высоким значением в наборе данных (Q2):

Q1 =df[Y\_label].quantile(0.25)

Q3 =df[Y\_label].quantile(0.75)

При вычислении получаем, что Q1 = 27.7, а Q2 = 46.6

Теперь вычислим значение IQR, который представляет собой разницу между третьим и первым квартилем:

IQR = Q3 - Q1

IQR

Значение IQR будет равно 18.9.

Теперь находим нижнюю и верхнюю границу для отсеивания выбросов в исходном наборе данных:

lower\_bound = Q1 - 1.5 \* IQR

upper\_bound = Q3 + 1.5 \* IQR

После вычисления lower\_bound будет равен -0.65, а upper\_bound равен 74.95.

Теперь можно рассмотреть конкретные значения выбросов из датафрейма Pandas с помощью выполнения операции фильтра:

df[(df[Y\_label]<lower\_bound) | (df[Y\_label]>upper\_bound)]

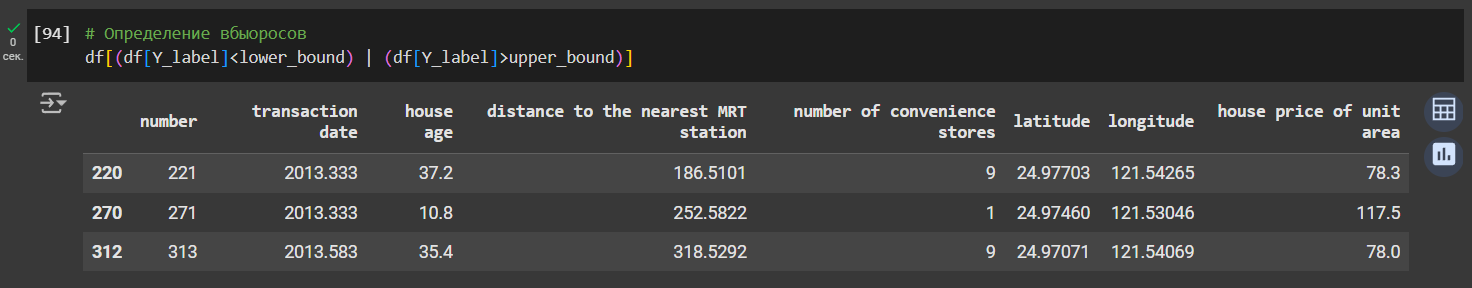


Рисунок 4 – Вывод выбросов в виде значений из датафрейма

Как видно из рисунка 4 выбросов в наборе данных всего 3. Данные выбросы необходимо удалить с помощью исключения их из всех значений набора данных следующим образом:

df = df[~((df[Y\_label] < (lower\_bound)) |(df[Y\_label] > (upper\_bound)))]

После этого на диаграмме “ящик с усами” и квантиль-квантиль графике никаких больших отклонений не будет (см. рис. 5), а из набора данных будет на 3 записи меньше – вместо 414 записей будет 411.

Теперь можно продолжать работу с обработанным набором данных.

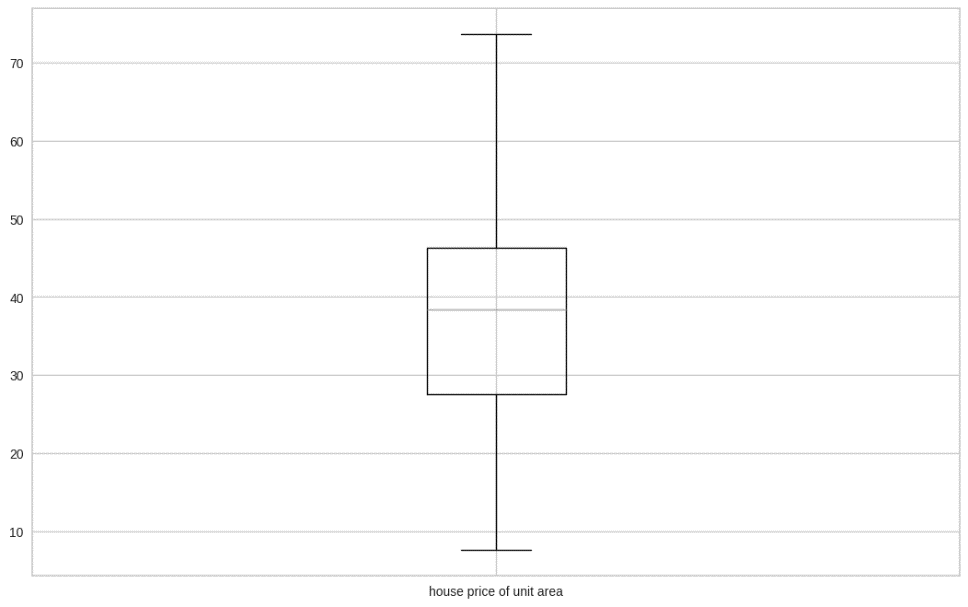


Рисунок 5 - Диаграмма boxplot для цены на недвижимость за единицу площади после удаления выбросов

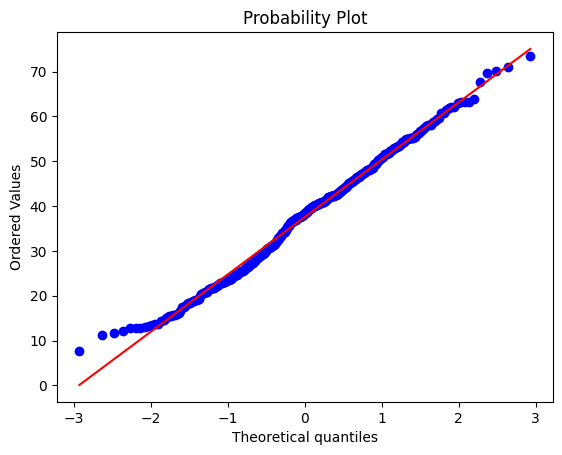


Рисунок 6 - Результат построения квантиль-квантиль графика по цене на недвижимость за единицу площади

Поскольку у нас используется метод многомерной линейной регрессии, то и неизвестных переменных X должно быть более одной, для предсказания цены.

Опишем атрибуты набора данных для большего удобства (см. таблицу 1).

Таблица 1 – Атрибуты набора данных

|  |  |  |
| --- | --- | --- |
| **Поле** | **Описание** | **Тип данных** |
| X1 | Возраст недвижимости | float64 |
| X2 | Расстояние до ближайшей станции метро | float64 |
| X3 | Количество круглосуточных магазинов | int64 |
| Y | Цена недвижимости за единицу площади | float64 |

Удалим из датафрейма те столбцы, которые точно не будут использоваться для решения задачи методом многомерной линейной регрессии: дата осуществления транзакции (transaction date) и номер (number). Данные столбцы не представляют интерес в контексте решаемой задачи, а поэтому их следует удалить:

df = df.drop('number', axis=1)

df = df.drop('transaction date', axis=1)

Теперь выведем тепловую карту корреляции между столбцами датафрейма:

sns.heatmap(df.corr(), annot=True, cmap='crest')

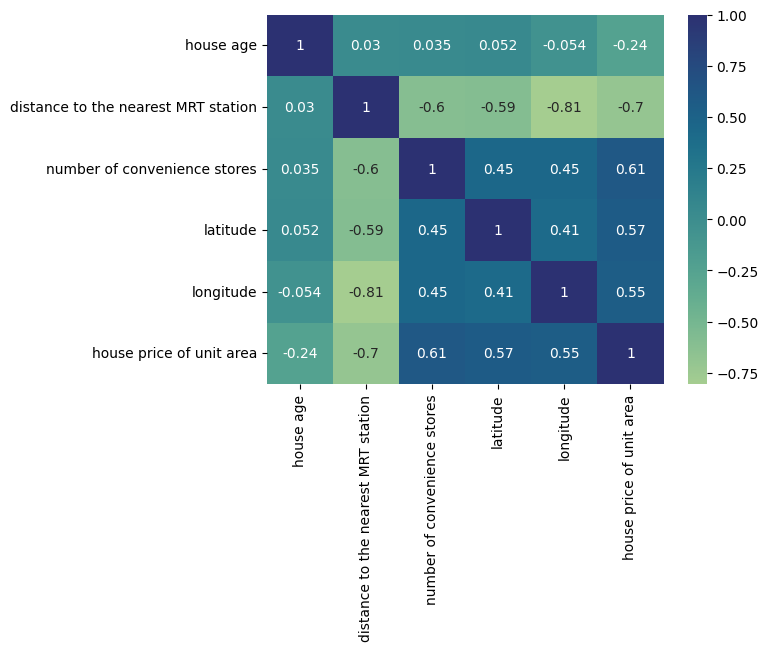


Рисунок 7 – Корреляция между столбцами датафрейма

Из атрибутов X1, X2 и X3 наиболее выраженную корреляцию с атрибутом Y имеет атрибут X2, менее выраженную – атрибут X1.