ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ НИЖЕГОРОДСКИЙ
ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. Р. Е. АЛЕКСЕЕВА

Кафедра «Прикладная математика»

Лабораторная работа №4

по дисциплине «Базы данных»

Тема: «Регрессия, регрессионная модель»

(разведочный анализ, распределение значений, отношения переменных,корреляция)

Студент

_____ Валькова.Н.П.
(Подпись) (Фамилия, И., О.)

ПМ 18-
(Группа) ……. (Дата сдачи)

Проверил

_____ Моисеев А.Е
(Подпись) (Фамилия, И., О.)

Отчет защищен «_____» _____2021__г.
с оценкой _____

Нижний Новгород, 2021

Оглавление

Введение:

Корреля́ция (от лат. *correlatio*), или **к**орреляцио́нная зави́симость — статистическая взаимосвязь двух или более случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин.

При расчёте корреляций пытаются определить, существует ли статистически достоверная связь между двумя или несколькими переменными в одной или нескольких выборках.

Важно понимать, что корреляционная зависимость отражает только взаимосвязь между переменными и не говорит о причинно-следственных связях.

Показатель корреляции. Коэффициент корреляции(r) характеризует величину отражающую степень взаимосвязи двух переменных между собой. Если коэффициент корреляции ближе к 1 (или -1) то говориться о сильной корреляции, а если ближе к 0,то о слабой(равенство 0 - отсутствие корреляции). При положительной корреляции увеличение (или уменьшение) значений одной переменной ведёт к закономерному увеличению (или уменьшению) другой переменной т.е. взаимосвязи типа увеличение-увеличение (уменьшение-уменьшение).

При отрицательной корреляции увеличение (или уменьшение) значений одной переменной ведёт к закономерному уменьшению (или увеличению) другой переменной т.е. взаимосвязи типа увеличение-уменьшение (уменьшение-увеличение).

Разведочный анализ данных (англ. exploratory data analysis, EDA) — анализ основных свойств данных, нахождение в них общих закономерностей, распределений и аномалий, построение начальных моделей, зачастую с использованием инструментов визуализации.

Постановка задачи:

1. Скачать данные с kaggle: Sberbank Russian Housing Market.
2.Провести разведочный анализ датасета: построить графики распределений и отношений переменных,построить тепловую карту корррреляций.

Решение
Скачиваем базу Sberbank Russian Housing Market с сайта kaggle.com.
Для построения графиков было выбрано 5 колонок
timestamp - дата
price_doc - цена продажи
full_sq -общая площадь
life_sq — жилые помещения
max_floor — количество этажей
build_year — дата постройки
green_zone_part — доля зелёной зоны
shopping_centers_raion — сколько в районе магазинов

- Импортируем:

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

- Считываем. Из колонки с ценой убираем выброс. У колонки с датами оставляем только год.

```
data = pd.read_csv("train.csv", index_col = 'id')
# cols = ['timestamp','price_doc', 'full_sq', 'life_sq', 'max_floor', 'build_year','green_zone_part']
columns = ['price_doc', 'full_sq', 'max_floor', 'build_year','shopping_centers_raion']
data = data[columns].dropna()
# data['timestamp'] = data['timestamp'].str.split('-').str[0]
# data = data[data.price_doc < data.price_doc.mean()]
# data = data[data.full_sq < data.full_sq.mean()]
```

- Строим матрицу корреляции и тепловую картц корреляции

```
sns.set(style = 'whitegrid')
sns.pairplot(data[columns],hue="timestamp",diag_kind="hist")
plt.savefig('pair3.png')
plt.show()
```

Строим график распределений и отношений переменных:

```
cor = np.corrcoef(data[columns].to_numpy().T)
print(cor)
plt.subplots(figsize=(10,10))
sns.heatmap(cor, cmap='Spectral', cbar=True, annot=True,
square=True, yticklabels=columns, xticklabels=columns)
plt.xticks(rotation=45)
plt.savefig('pair.png')
plt.show()
```

График распределений и отношений переменных:

Добавлена разбивка по годам.

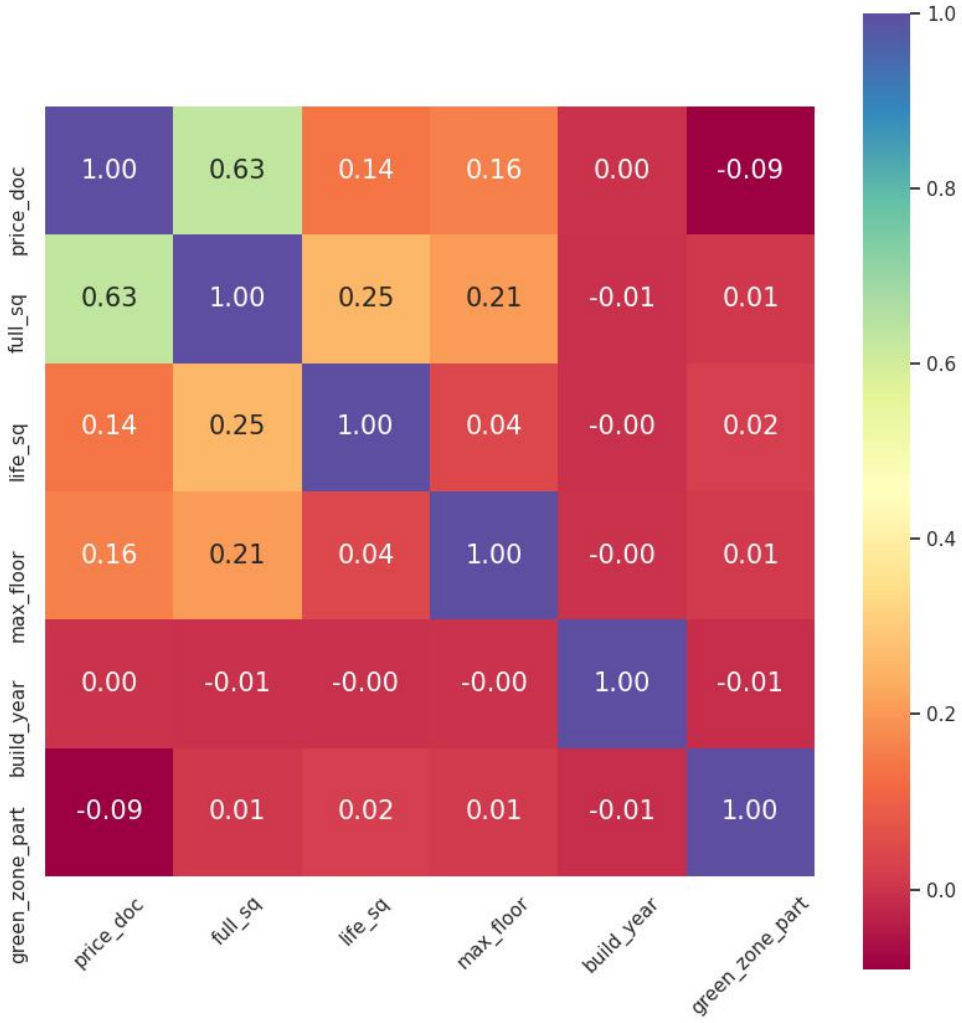
- С зелёной зоной

Можно заметить зависимости между следующими величинами: вся площадь и цена, вся площадь и жилая площадь, а так же площадь и максимальное число этажей.

- С учётом магазинов

Не очень очевидно, но можно заметить связь между числом магазинов в районе и ценой.

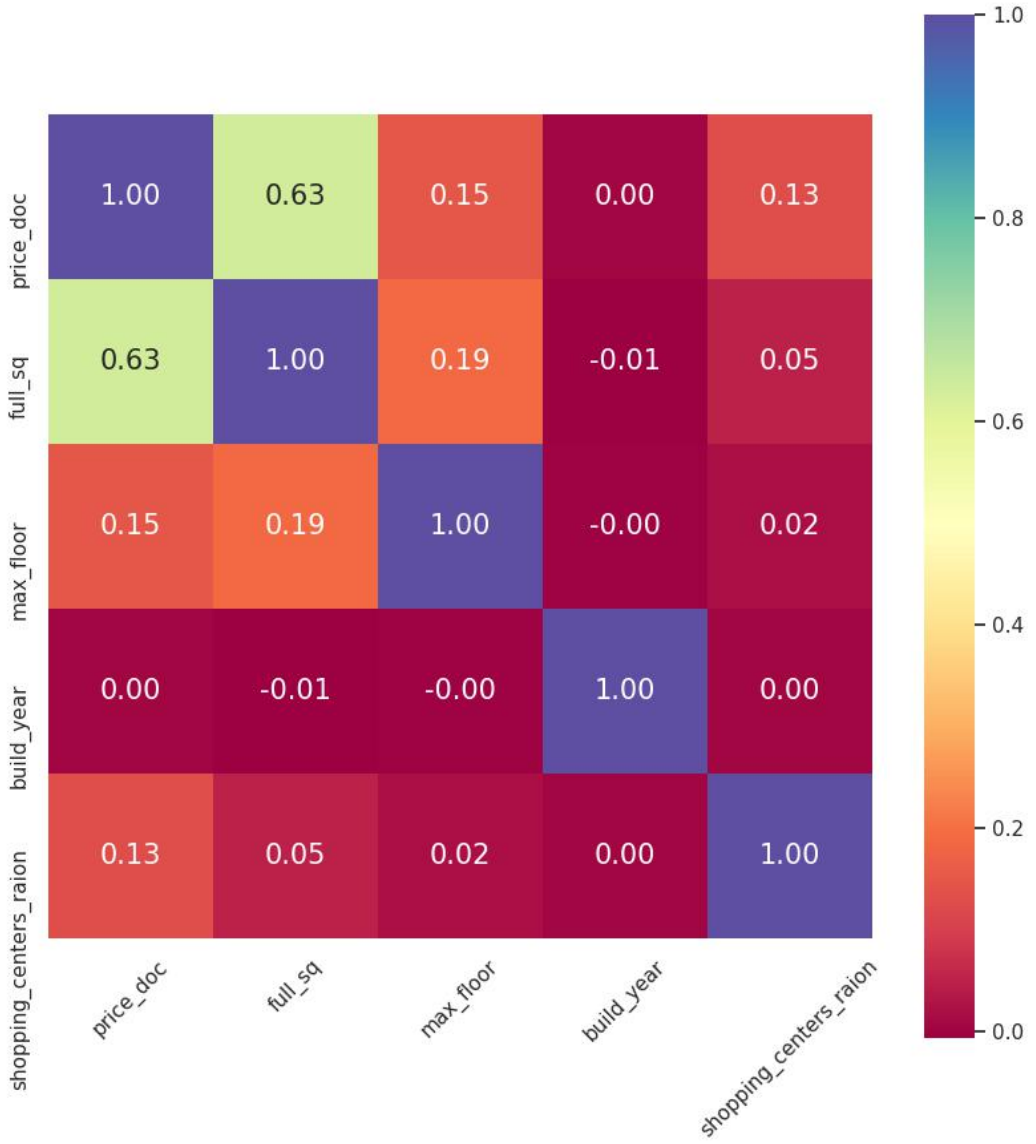
Тепловая карта корреляции:



По карте корреляции можем наблюдать связь между ценой и площадью, а так же слабую связь цены и числа этажей. Так же довольно логично, что мы наблюдаем связь, хоть и слабую всей площади с жилой площадью и числом этажей.

```
[[ 1.00000000e+00  6.31759865e-01  1.40582039e-01  1.60530937e-01
   2.13399425e-03 -9.13570676e-02]
 [ 6.31759865e-01  1.00000000e+00  2.54971752e-01  2.08166353e-01
  -6.06503733e-03  8.34140185e-03]
 [ 1.40582039e-01  2.54971752e-01  1.00000000e+00  4.32370739e-02
  -2.40106709e-03  2.23467579e-02]
 [ 1.60530937e-01  2.08166353e-01  4.32370739e-02  1.00000000e+00
  -2.83353770e-04  1.12069860e-02]
 [ 2.13399425e-03 -6.06503733e-03 -2.40106709e-03 -2.83353770e-04
   1.00000000e+00 -7.21726850e-03]
 [-9.13570676e-02  8.34140185e-03  2.23467579e-02  1.12069860e-02
  -7.21726850e-03  1.00000000e+00]]
```

Как и было замечено ранее присутствует слабая связь цены и числа ТЦ в районе.




```
[[ 1.00000000e+00  6.31430002e-01  1.48177797e-01  2.16093517e-03
   1.30389332e-01]
 [ 6.31430002e-01  1.00000000e+00  1.86678624e-01 -6.04089967e-03
   5.15794036e-02]
 [ 1.48177797e-01  1.86678624e-01  1.00000000e+00 -2.61118784e-04
   1.88537144e-02]
 [ 2.16093517e-03 -6.04089967e-03 -2.61118784e-04  1.00000000e+00
   2.45640056e-03]
 [ 1.30389332e-01  5.15794036e-02  1.88537144e-02  2.45640056e-03
   1.00000000e+00]]
```

Листинг:

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

data = pd.read_csv("train.csv", index_col = 'id')
# columns = ['timestamp', 'price_doc', 'full_sq', 'life_sq', 'max_floor',
 'build_year', 'green_zone_part']
columns = ['price_doc', 'full_sq', 'max_floor', 'build_year', 'shopping_centers_raion']
data = data[columns].dropna()
data['timestamp'] = data['timestamp'].str.split('-').str[0]
data = data[data.price_doc < data.price_doc.mean()]
data = data[data.full_sq < data.full_sq.mean()]

sns.set(style = 'whitegrid')
sns.pairplot(data[columns], hue="timestamp", diag_kind="hist")
plt.savefig('pair3.png')
plt.show()

cor = np.corrcoef(data[columns].to_numpy().T)
print(cor)
plt.subplots(figsize=(10,10))
sns.heatmap(cm, cmap='Spectral', cbar=True, annot=True, square=True,
yticklabels=columns, xticklabels=columns)
plt.xticks(rotation=45)
plt.savefig('pair.png')
plt.show()
```