Московский Авиационный Институт

(Национальный Исследовательский Университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

**Лабораторная работа №0 по курсу**

**«Искусственный интеллект (Машинное обучение)»**

Студент: Алексеев В.Е.
Группа: М8О–306Б–19
Преподаватель: Ахмед Самир Халид
Оценка: _____
Дата: _____
Подпись: _____

Москва, 2022

# Лабораторная работа №0

**Задача:** Вам предстоит руками проанализировать данные, визуализировать зависимости, построить новые признаки и сказать хватит ли вам этих данных, и если не хватит найти еще.

**Датасет:** New York City Airbnb Open Data

## Описание входных данных:

- id – id апартаментов.
- name – название апартаментов.
- host ID – id хозяина.
- host_name – имя хозяина.
- neighbourhood_group – район Нью-Йорка.
- neighbourhood – окрестность.
- latitude – широта.
- longitude – долгота.
- room_type – тип апартаментов.
- price – цена апартаментов.
- minimum_nights – минимальное количество ночей.
- number_of_reviews – количество отзывов.
- last_review – дата последнего отзыва.
- reviews_per_month – количество отзывов за месяц.
- calculated_host_listings_count – количество записей на апартаменты.
- availability_365 – количество дней, когда апартаменты доступны для бронирования.

## Типы признаков:

• Категориальные
– name
– host_name
– neighbourhood_group
– neighbourhood
– room_type
– last_review

• Количественные
– id
– host_id
– latitude
– longitude
– price
– number_of_reviews
– reviews_per_month

– calculated_host_listings_count
– availability_365

**Размер:**

- Строк: 48895
- Столбцов: 16

## Решаемая задача:

Задача предсказания - предсказать цену апартаментов (price).

Для решения поставленной задачи нам нужно оставить только те признаки, которые нам понадобятся:

– neighbourhood_group
– neighbourhood
– latitude
– longitude
– room_type
– price
– minimum_nights
– number_of_reviews
– reviews_per_month
– calculated_host_listings_count
– availability_365

Теперь, проанализировав данные, можно понять, что нам необходимо заполнить пропуски.

Будем заполнять reviews_per_month средним значением:

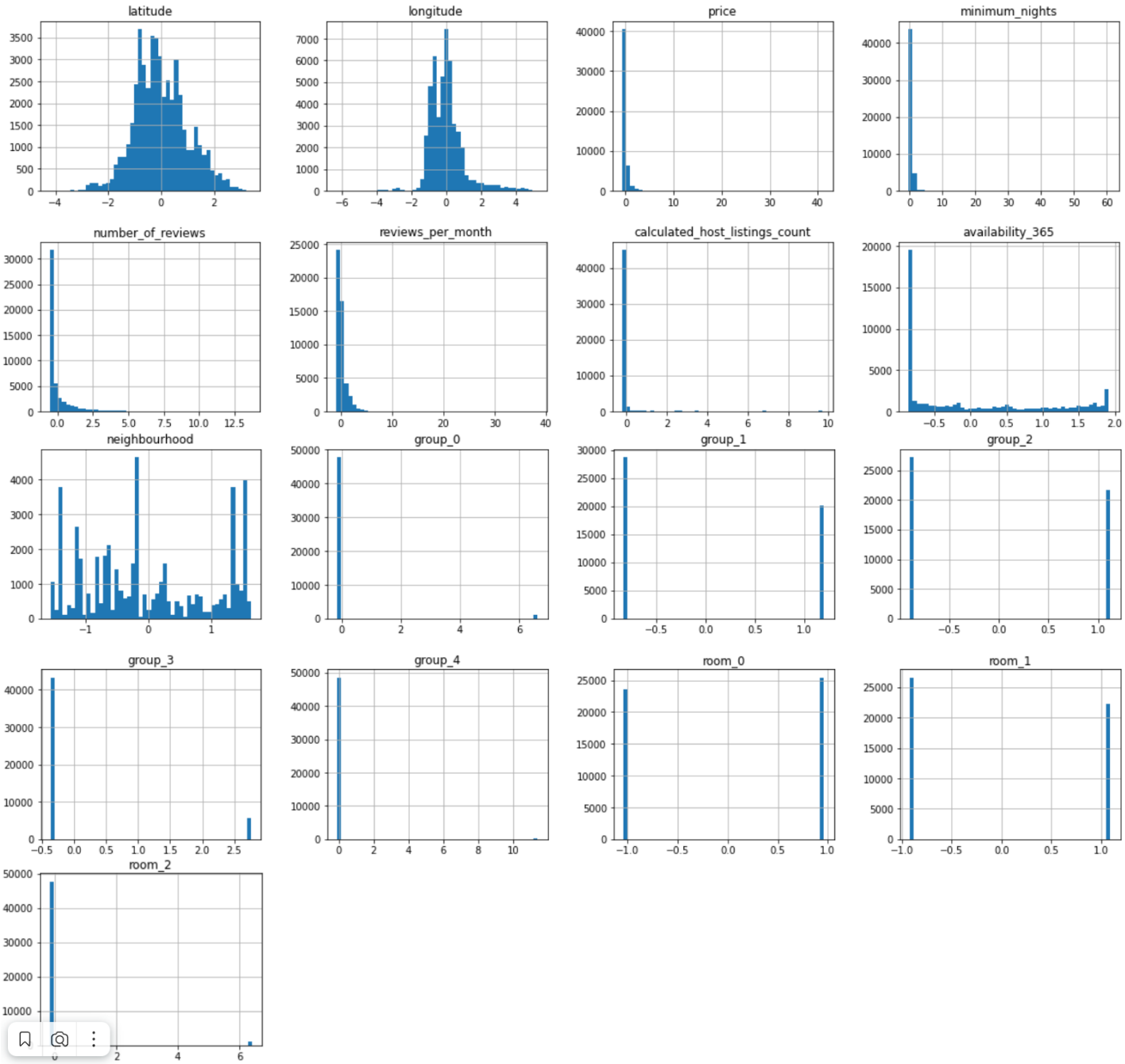
| | Total | Percent |
|---|---|---|
| reviews_per_month | 10052 | 0.205583 |
| latitude | 0 | 0.000000 |
| group_0 | 0 | 0.000000 |
| room_1 | 0 | 0.000000 |
| room_0 | 0 | 0.000000 |
| group_4 | 0 | 0.000000 |
| group_3 | 0 | 0.000000 |
| group_2 | 0 | 0.000000 |
| group_1 | 0 | 0.000000 |
| neighbourhood | 0 | 0.000000 |
| longitude | 0 | 0.000000 |
| availability_365 | 0 | 0.000000 |
| calculated_host_listings_count | 0 | 0.000000 |
| number_of_reviews | 0 | 0.000000 |
| minimum_nights | 0 | 0.000000 |
| price | 0 | 0.000000 |
| room_2 | 0 | 0.000000 |

## Работа с категориальными признаками:

| | neighbourhood_group | neighbourhood | room_type |
|---|---|---|---|
| **count** | 48895 | 48895 | 48895 |
| **unique** | 5 | 221 | 3 |
| **top** | Manhattan | Williamsburg | Entire home/apt |
| **freq** | 21661 | 3920 | 25409 |

Нам следует переделать категориальные данные в числовые и пронормировать. room_type и neighbourhood_group будем переделывать с помощью one_hot_encoder'a, a neighbourhood с помощью label_encoder'a.

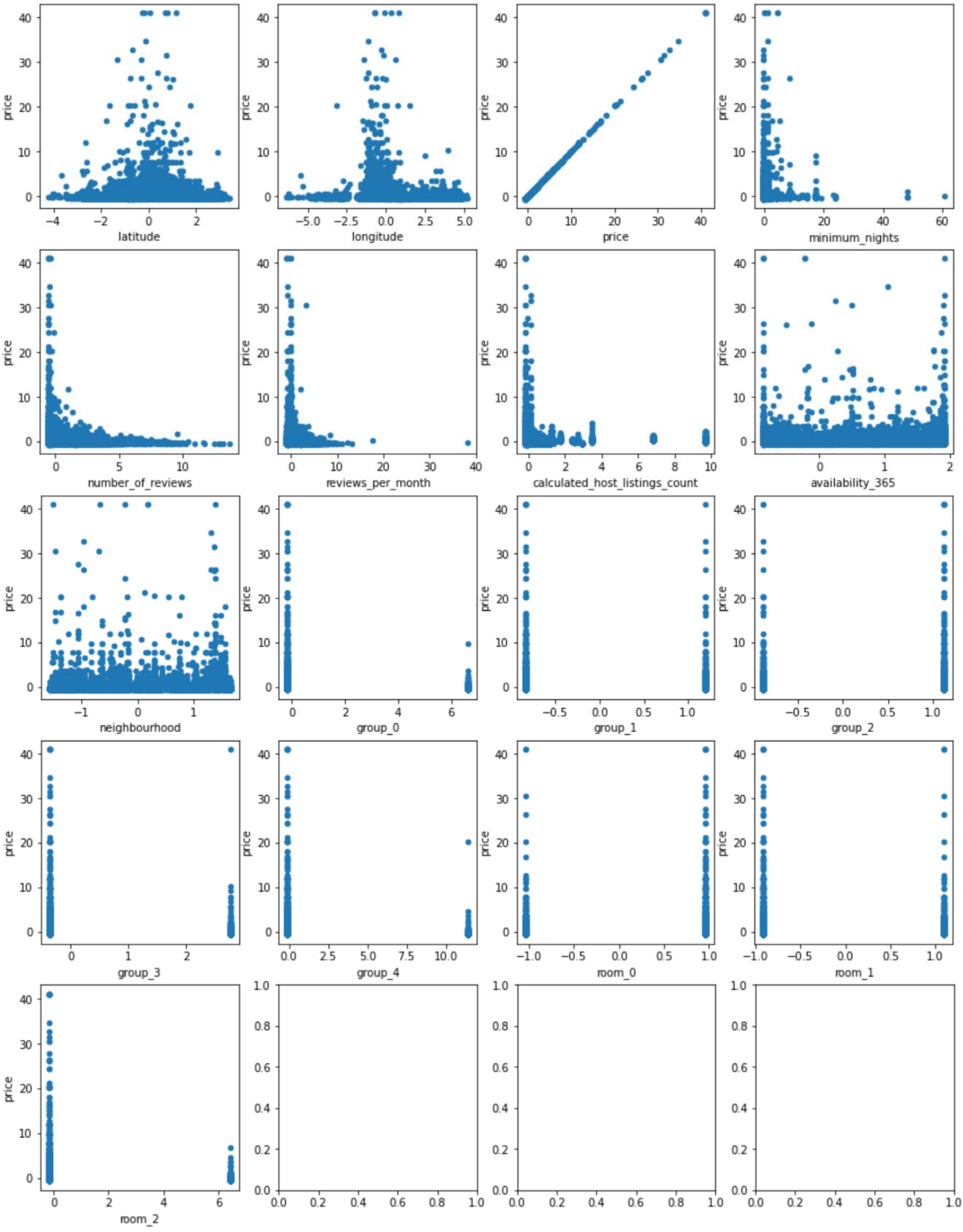
## Распределение признаков после нормировки:



**Корреляция:**

```
price                             1.000000
room_0                            0.255857
group_2                           0.163976
availability_365                  0.081829
neighbourhood                     0.062057
calculated_host_listings_count    0.057472
minimum_nights                    0.042799
latitude                          0.033939
group_4                          -0.013840
reviews_per_month                -0.022373
group_0                          -0.041030
number_of_reviews                -0.047954
room_2                           -0.053613
group_3                          -0.080205
group_1                          -0.098603
longitude                        -0.150019
room_1                           -0.240246
Name: price, dtype: float64
```

## Зависимость главного значения от остальных:



## Выводы:

Данная лабораторная работа была довольно интересной. В ходе ее выполнения я попрактиковался в анализе и подготовке данных, научился выявлять статистические закономерности и наиболее важные признаки. Еще я понял, что для обработки признаков нужны знания статистики, например, чтобы строить различные сложные графики и уметь их понимать. Важно так же уметь создавать новые признаки из имеющихся и отбирать только самые нужные из изначальных.