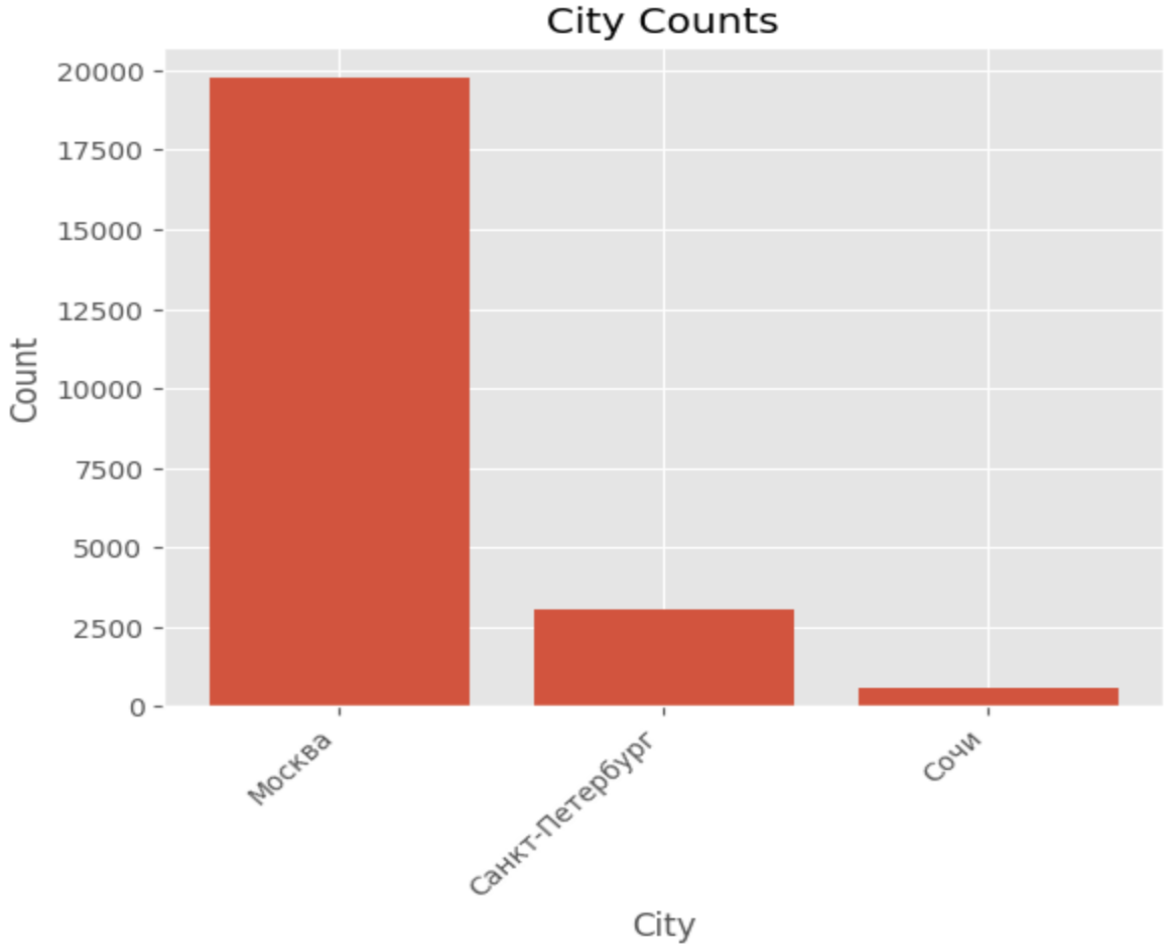
Exploratory Data Analysis

Датасет _data.csv представлен в виде таблицы, которая содержит в себе данные о жилых помещениях, используемых для аренды.
Таблица также содержит в себе столбцы с несколькими типами данных (int, str, float итд), поэтому некоторые столбцы были вынесены в отдельную таблицу.

Более подробно о датасете описано в следующих шагах:

1. Так как для исследования нужен только один город, проверили на наличие других городов в датафрейме, всего три города – ***Москва, Сочи, Санкт-Петербург.***
   На графике можно увидеть сколько объявлений представлено в каждом городе.



2. Колонка ***Цена*** содержит в себе много типов данных (str, int) и разные данные, для удобства вынесли её в отдельную таблицу с колонками: *“'Стоимость', 'Залог', 'Предоплата в мес', 'Коммунальные услуги включены'”*

	Стоимость	Залог	Предоплата в мес	Коммунальные услуги включены
0	500000	True	1	True
1	500000	True	1	True
2	500000	True	1	False
3	400000	True	1	False
4	225000	True	1	True
...	...	...	...	...
23363	42000	True	1	True
23364	45000	True	1	True
23365	50000	True	1	True
23366	55000	True	2	True
23367	57000	True	1	True

23368 rows × 4 columns

3. Распределение количества объявлений по станциям метро Москвы
4. Средняя цена за месяц аренды по станциям метро Москвы

5. Количество отсутствующих данных по колонкам (NaN)

Заключение

1. После изучения датасета пришли к выводу – в данных много NaN значений, которые в будущем стоит заменить или исключить
2. Информативные колонки (например, “Цена”) необходимо вынести в отдельную таблицу и добавить новые фичи(признаки) с информацией в исходную таблицу. Датасет станет более точным.
3. Много повторений в датасете – необходимо исключить наличие дубликатов и оставить только уникальные значения