Explanatory Data Analysis | Разведочный анализ данных

1. Обзор набора данных
2. Географическое разбитие данных
3. Обработка отсутствующих данных
4. Характеристики отдельных столбцов
5. План дальнейших действий:

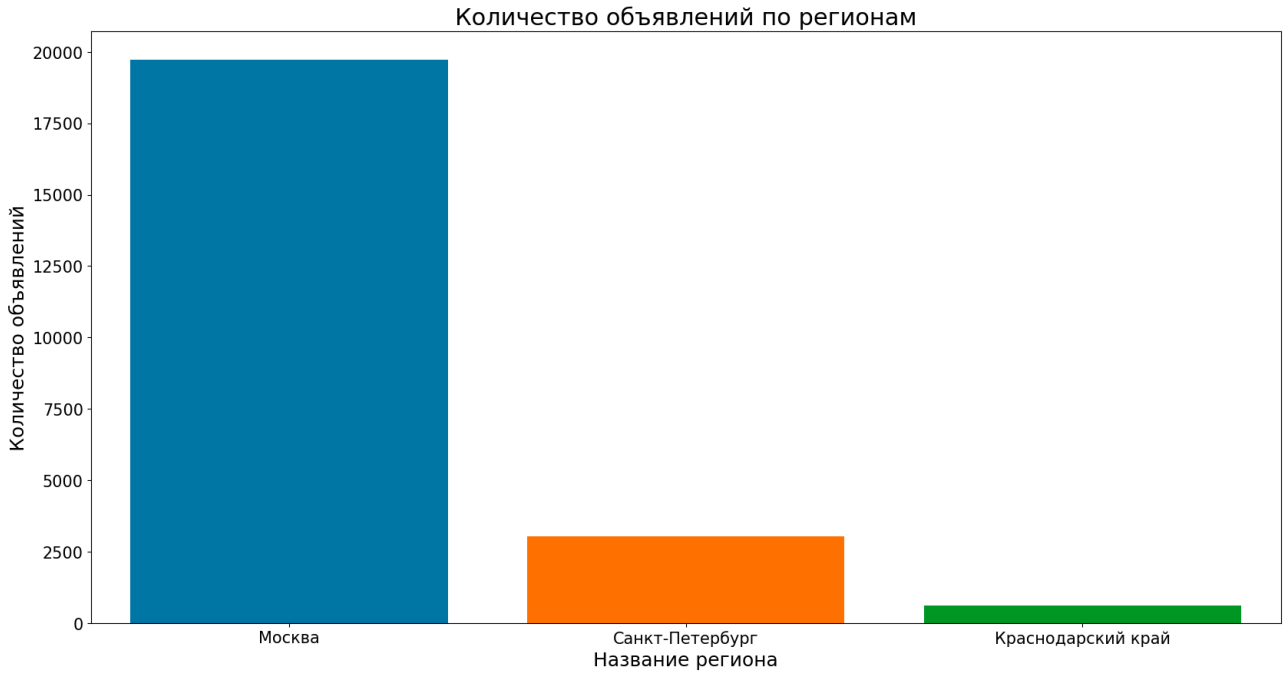
Описание датасета

Дата сет представлен в виде таблице, включающий в себя 23368 квартир и 24 колонок содержащих характеристики на каждую из квартир. Большинство столбцов необходимо будет в дальнейшем разделять на несколько для удобства обработки данных, так как на данный момент в них разные типы данных (int и str). Есть столбцы со значением Nan. Кроме того, есть индивидуальные данные для каждой строки - ID объявления и Ссылка на объявление.
Доступны следующие колонки:

```
Index(['ID  объявления', 'Количество комнат', 'Тип', 'Метро', 'Адрес',
       'Площадь, м2', 'Дом', 'Парковка', 'Цена', 'Телефоны', 'Описание',
       'Ремонт', 'Площадь комнат, м2', 'Балкон', 'Окна', 'Санузел',
       'Можно с детьми/животными', 'Дополнительно', 'Название ЖК',
       'Серия дома', 'Высота потолков, м', 'Лифт', 'Мусоропровод',
       'Ссылка на объявление'],
      dtype='object')
```

Географическое разбитие данных

Так как в качестве пилотного региона выбрана Москва, проверим адрес на наличие других регионов в датафрейме:



Убираем объявления с Санкт-Петербургом и Краснодарским краем

Для того, чтобы просмотреть информацию о размерности данных воспользуемся функцией info()

```
<class 'pandas.core.frame.DataFrame'>
Index: 19737 entries, 0 to 23367
Data columns (total 24 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   ID  объявления          19737 non-null  int64
 1   Количество комнат       19202 non-null  object
 2   Тип                     19737 non-null  object
 3   Метро                   19391 non-null  object
 4   Адрес                   19737 non-null  object
 5   Площадь, м2             19737 non-null  object
 6   Дом                     19737 non-null  object
 7   Парковка                8563 non-null   object
 8   Цена                    19737 non-null  object
 9   Телефоны                19737 non-null  object
 10  Описание                19737 non-null  object
 11  Ремонт                  17274 non-null  object
 12  Площадь комнат, м2      12509 non-null  object
 13  Балкон                  13107 non-null  object
 14  Окна                    14587 non-null  object
 15  Санузел                 17696 non-null  object
 16  Можно с детьми/животными  14822 non-null  object
 17  Дополнительно           19465 non-null  object
 18  Название ЖК             4456 non-null   object
 19  Серия дома              2091 non-null   object
...
 22  Мусоропровод            11730 non-null  object
 23  Ссылка на объявление    19737 non-null  object
dtypes: float64(1), int64(1), object(22)
```

У нас имеется только два столбца с числовыми данными, но можно предположить, что должно быть больше таких столбцов. Возможно, потребуется изменить тип данных или извлечь дополнительную информацию из этих столбцов в дальнейшем. Кроме того, следует обратить внимание на то, что некоторые столбцы содержат большое количество значений NaN (например, Серия дома и Название ЖК), что указывает на отсутствие информации в этих ячейках.

Обработка отсутствующих данных

Количество полностью заполненных объектов из всей выборки: 6
Процент полностью заполненных объектов из всей выборки: 0.03

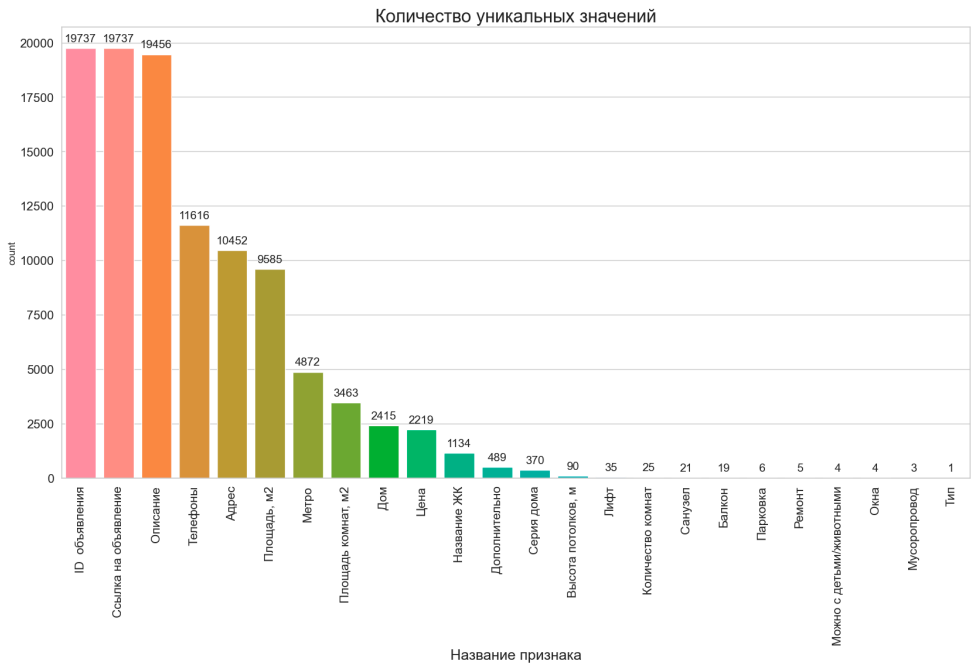
Процент пропущенных значений:

Серия дома	89.41%
Название ЖК	77.42%
Парковка	56.61%
Высота потолков, м	46.62%

Мусоропровод	40.57%
Площадь комнат, м2	36.62%
Балкон	33.59%
Окна	26.09%
Можно с детьми/животными	24.90%
Лифт	21.24%
Ремонт	12.48%
Санузел	10.34%
Количество комнат	2.71%
Метро	1.75%
Дополнительно	1.38%
ID объявления	0.00%
Описание	0.00%
Телефоны	0.00%
Цена	0.00%
Дом	0.00%
Площадь, м2	0.00%
Адрес	0.00%
Тип	0.00%
Ссылка на объявление	0.00%

Из всего набора объявлений у нас заполнены все колонки только в 3% от общего числа. Также видно, что более 50% данных отсутствуют в трех конкретных колонках. Исходя из этого, эти колонки могут быть кандидатами на удаление в дальнейшем исследовании данных, которое поможет принять более обоснованное решение.

Посмотрим количество уникальных значений по колонкам



Выводы

1. Колонка "Тип" имеет только одно значение, поэтому ее можно удалить, так как она не несет информационной ценности.
2. Колонки "Мусоропровод", "Окна", "Ремонт", "Парковка", "Можно с детьми/животными" содержат небольшое количество значений. Это может указывать на то, что эти колонки могут быть категориальными признаками, которые могут быть использованы для классификации или анализа данных.
3. Колонки "Ссылка на объявление", "Описание" и "ID объявления" имеют большое количество уникальных значений. Вероятно, эти колонки несут слишком много разнородной информации и могут быть удалены в дальнейшем анализе.

Посмотрим метрики для каждого столбца

	ID объявления	Высота потолков, м
count	19737.000000	10535.000000
mean	267114888.218270	2.992925
std	19801055.081395	7.852740
min	107298592.000000	1.200000
25%	271221229.000000	2.640000
50%	273928403.000000	2.640000
75%	274697333.000000	2.800000
max	275006443.000000	320.000000

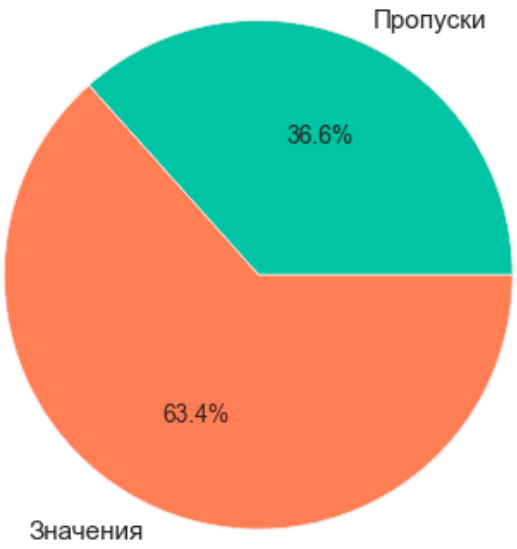
Так как для всех объектов ID объявление уникально, то мы им принебрегаем.
Высота потолков от 1,2 до 320 метров - тут имеем дело с выбросами

Характеристики отдельных столбцов

Можем удалить стобец 'Телефоны', поскольку номер телефона не влияет на стоимость

Столбец 'Тип' содержит только одно значение "Квартира" для всех строк в DataFrame apart_rent он не несет полезной информации для анализа или моделирования.Удаление столбца не повлияет на результаты и выводы.

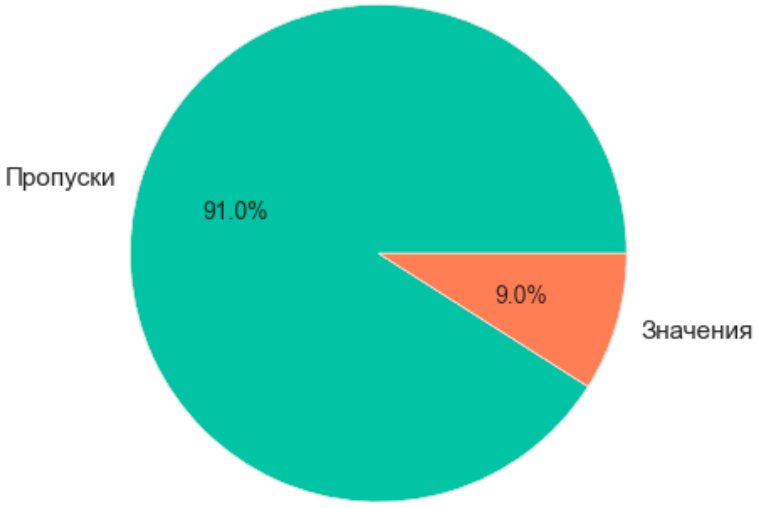
Соотношение информативных и пустых строк в столбце 'Площадь комнат, м2'



Из круговой диаграммы видно, что количество пропусков практически 50%, при этом имеется альтернативный стобец с общей площадью, где количество пропусков 0, следовательно данный можно удалить

Можно удалить стоблец ['Серия дома'] поскольку он не влияет на стоимость жилья, является идентификационным номером. Так же можно увидеть, что 90% информации это пропуски.

Соотношение информативных и пустых строк в столбце 'Площадь комнат, м2'



Удаляем данный столбец из-за большого количесва пропусков и отсустсвием влияния на стоимость

План дальнейшей работы может быть следующим:

1. Заменить названия колонок на английском языке и обработать пропущенные значения (NaN) и значения None. Это поможет создать более удобную и однородную структуру данных.
2. Добавить новые признаки (фичи) на основе имеющихся данных. Например, можно разбить существующие колонки на несколько более детальных, преобразовать текстовые данные в числовые или создать дополнительные признаки, основываясь на имеющихся данных.
3. Удалить дубликаты объявлений, чтобы избежать повторений и сохранить только уникальные данные. Это поможет очистить данные и предотвратить искажения при анализе или обучении модели.
4. Передать файл "data.csv" команде машинного обучения для обучения модели. Подготовленные данные могут быть использованы для тренировки и оценки различных моделей машинного обучения.

План дальнейшей работы охватывает шаги по обработке данных, созданию новых признаков, очистке и подготовке данных для дальнейшего анализа или обучения моделей машинного обучения.