# Задание 2

Проведите предварительный анализ данных из задания 1. На основе анализа подготовьте презентацию.

Переменные (столбцы в файле) имеют следующие значения:

К – количество комнат;
Метро – ближайшее метро;
От М – расстояние от метро в минутах ходьбы пешком или на транспорте; например, 10п это 10 минут пешком, а 5т – 5 минут на транспорте;
Улица, номер дома, дом, - адрес;
Общ – общая площадь;
Жил – жилая площадь;
Кух – площадь кухни;
Б – балкон;
Л – лифт;
Т – телефон;
С – санузел;
П – пол;
И – ипотека;
Цена – зависимая переменная
Статус – прямая продажа или возможна альтернатива – размен, например;
Примечание – текстовое описание квартиры.

## Проведите анализ распределения значений отдельных переменных

- Для категориальных переменных: постройте частотные таблицы по категориям. Подумайте, есть ли смысл объединять категории с малым числом примеров. Подумайте, можно ли графически проверить, «безопасно» ли такое объединение.
- Для количественных переменных представьте описательные статистики, гистограммы распределения, ящичковые диаграммы и/или другие графики похожего типа (например, violinplot из библиотеки seaborn).
  Посмотрите, есть ли в данных явные (или не очень явные) выбросы, возможно стоит построить две серии графиков – по всем данным и по «основной части», с отброшенными выбросами.
  Подумайте, есть ли смысл применить к данным какие-нибудь нелинейные преобразования (часто используют логарифмирование), чтобы сделать их распределение более похожим на «стандартные» вероятностные распределения – равномерное, нормальное или экспоненциальное.
- Для всех переменных определите число пропусков.

## Оцените зависимости между парами переменных

Исследуйте такие вопросы, как:

- Как связаны друг с другом и с ценой разные площади (жилая, общая, кухни)? Что из них оставить, а что лучше исключить при построении модели?
- Влияют ли на цену число балконов или наличие телефона?
- …

Для предварительной оценки удобно строить scatterplot’ы, считать корреляцию или другие статистические критерии.

## Создание новых переменных

Подумайте, можно ли создать на основе имеющихся данных какие-то новые переменные, как было сделано в предыдущем задании с переменными, связанными с числом этажей в доме.

Например, подумайте, как получить единое значение для расстояния от метро. Есть предложение, что можно время пешком оставить как есть, а время на транспорте – умножать на 4. Можно ли графически проверить такое предположение? Может быть лучше использовать другую константу?

Можно ли построить какие-то признаки, связанные с географическим расположением квартиры?

Можно ли получить из поля «примечания» какие-то полезные признаки, например, описывающие состояние ремонта? Какие-то ещё признаки?

## Что интересного можно узнать про Москву

Придумайте, что интересного можно узнать из этих данных про Москву?

В каких районах какая высота домов? Где больше всего новостроек (их можно выявить по наличию большого числа продающихся квартир в одном доме)? …