МГТУ им. Н. Э. Баумана кафедра ИУ5

курс «Технологии машинного обучения»

Лабораторная работа №1

«Разведочный анализ данных. Исследование и визуализация данных»

ВЫПОЛНИЛ:

Гаранин А. В. Группа ИУ5Ц-81Б

ПРОВЕРИЛ:

Гапанюк Ю. Е.

Москва, 2020 г.

**Цель лабораторной работы:** изучение различных методов визуализации данных.

**Краткое описание:** Построение основных графиков, входящих в этап разведочного анализа данных. Корреляционный анализ данных.

Формирование выводов о возможности построения моделей машинного обучения и о возможном вкладе признаков в модель.

# Задание:

Выбрать набор данных (датасет). Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного набора данных
2. Основные характеристики датасета
3. Визуальное исследование датасета. Необходимо использовать не менее 2 различных библиотек и не менее 5 графиков.
4. Информация о корреляции признаков.

Сформировать отчет и разместить его в репозитории на github.

# Выбор набора данных:

Источник:

<https://www.kaggle.com/hugoncosta/price-of-flats-in-moscow>

Следующий набор данных дает небольшую выборку цен на квартиры в Москве.

Показатели:

*price*

цена квартиры в $1000

*totsp*

общая площадь квартиры, кв.м.

*livesp*

жилая площадь квартиры, кв.м.

*kitsp*

площадь кухни, кв.м.

*dist*

расстояние от центра в км.

*metrdist*

расстояние до метро в минутах

*walk*

1 – пешком от метро, 0 – на транспорте

*brick*

1 – кирпичный, монолит ж/б, 0 – другой

*floor*

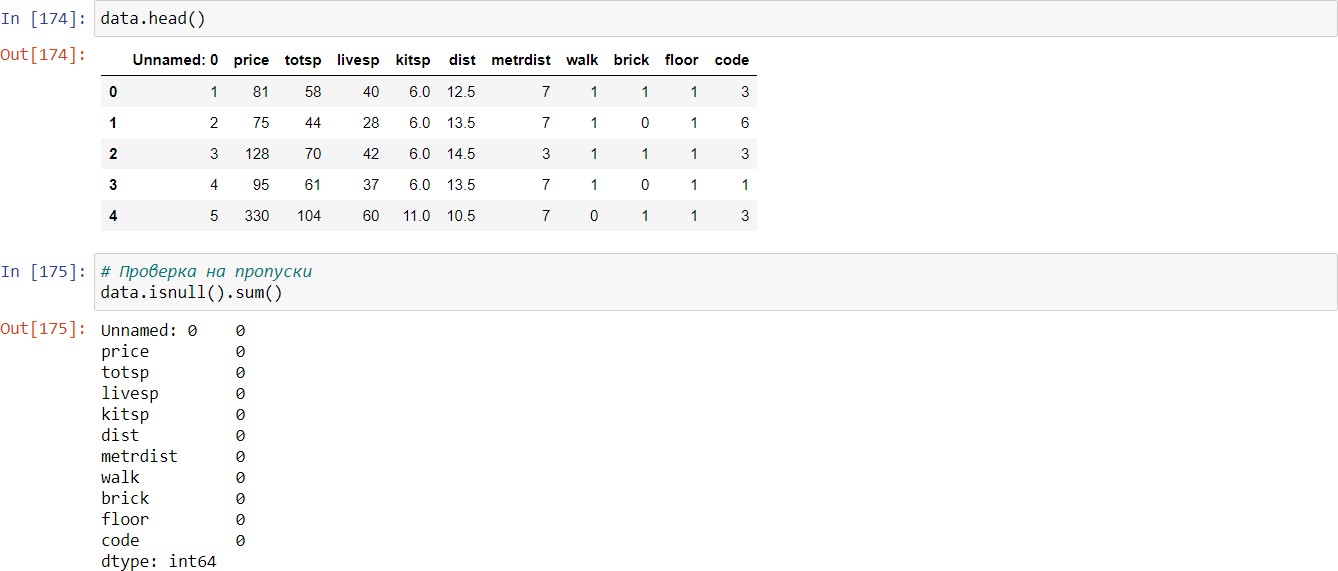
1 – этаж кроме первого и последнего, 0 – иначе.

*code*

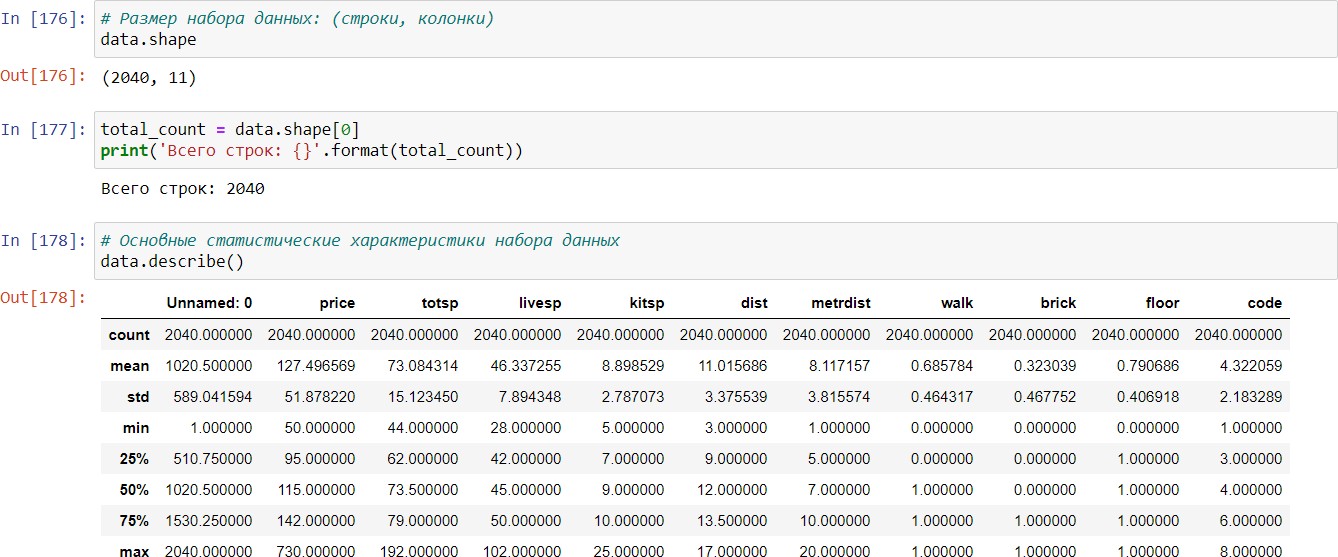
число от 1 до 8, при помощи которого мы группируем наблюдения по подвыборкам: 1. Наблюдения сгруппированы на севере, вокруг Калужско- Рижской линии метрополитена 2. Север, вокруг Серпуховско- Тимирязевской линии метрополитена 3. Северо-запад, вокруг Замоскворецкой линии метрополитена 4. Северо-запад, вокруг Таганско- Краснопресненской линии метрополитена 5. Юго-восток, вокруг Люблинской линии метрополитена 6. Юго-восток, вокруг Таганско- Краснопресненской линии метрополитена 7. Восток, вокруг Калиниской линии метрополитена 8. Восток, вокруг Арбатско-Покровской линии метрополитена

# Выполнение работы

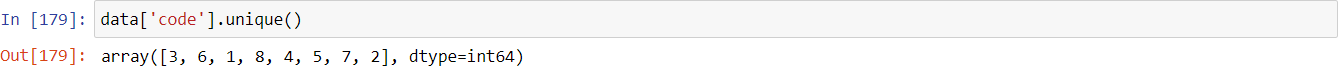
1. Загрузка и первичная обработка данных



=> Пустых значений в наборе данных нет.



Выделим индексы основных регионов:

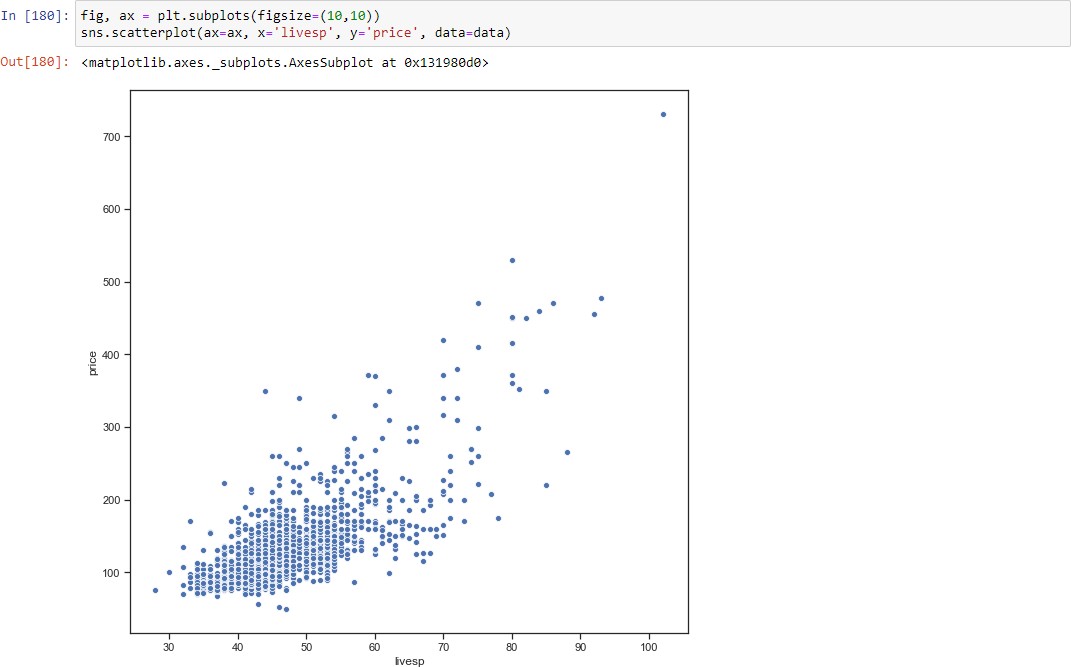


1. Визуальное исследование датасета

Исследование набора данных будет проводиться при помощи диаграмм рассеивания и гистограмм.

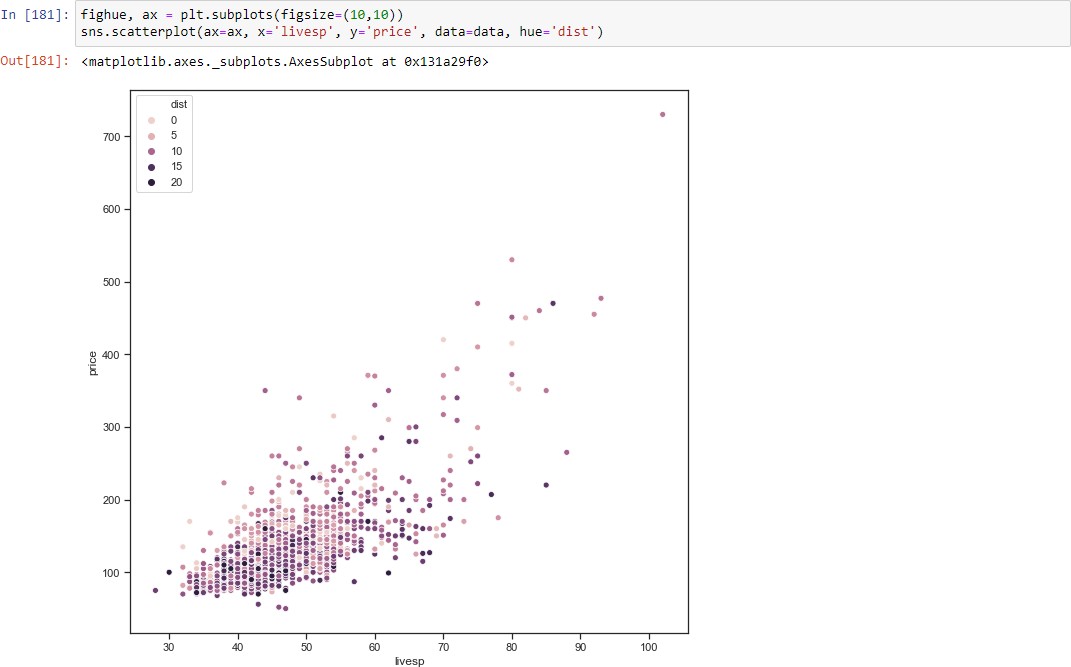
С помощью диаграммы рассеивания можно оценить, коррелируют ли между собой две выбранные переменные.

Исследуем зависимость цены (price) от жилой площади квартиры (livesp)



Как видно, зависимость имеет характер, близкий к экспоненциальному.

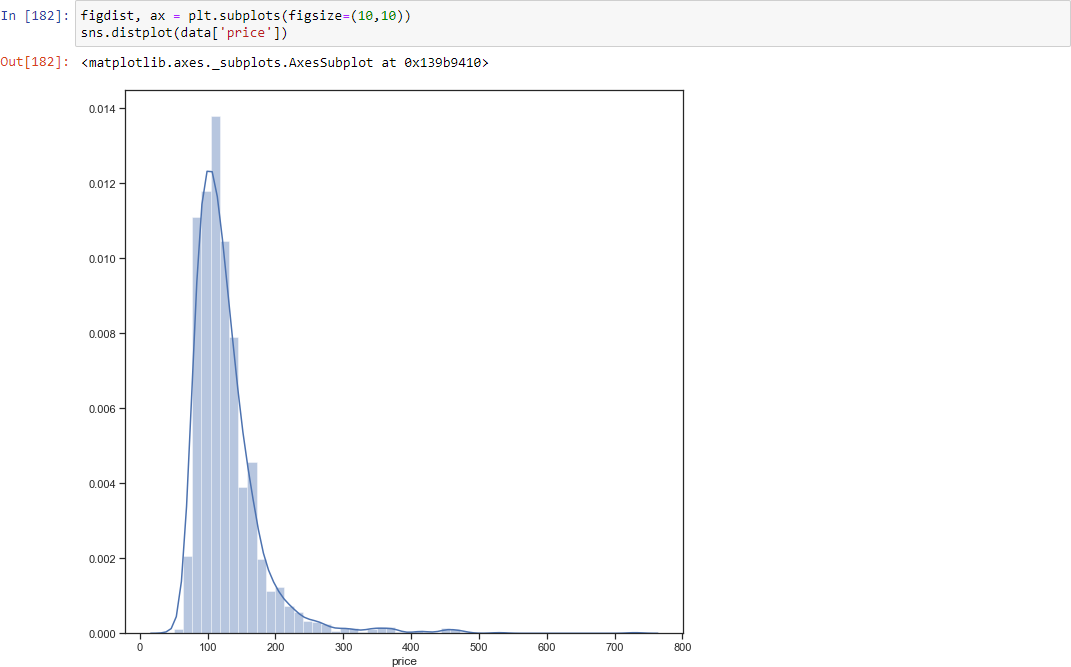
Рассмотрим еще и влияние третьего параметра – расстояния от центра до квартиры (dist).



Из графика видно, что в целом наблюдается слабая, неявная зависимость цены от расстояния от центра.

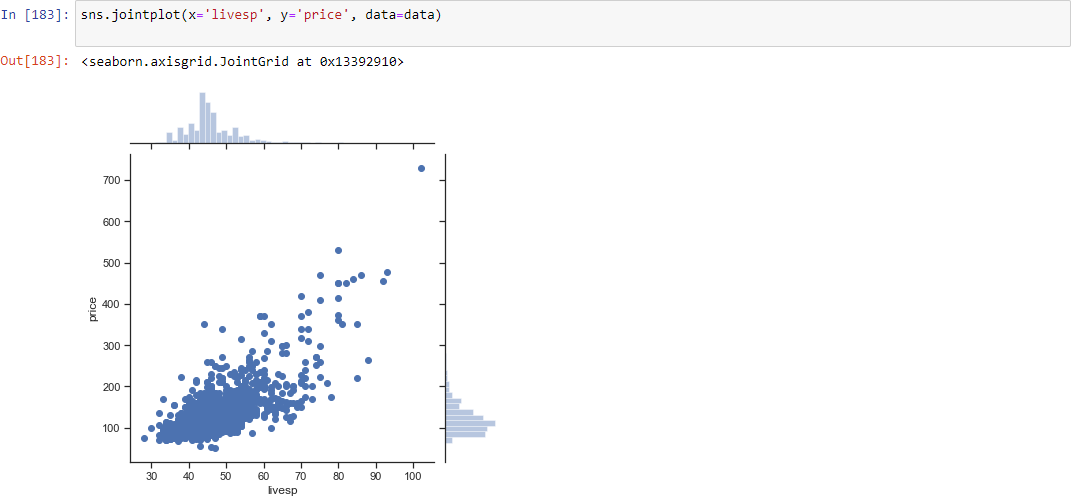
Еще одним способом визуального исследования данных является представление показателей в виде гистограммы. С помощью гистограммы можно оценить плотность вероятности распределения данных.

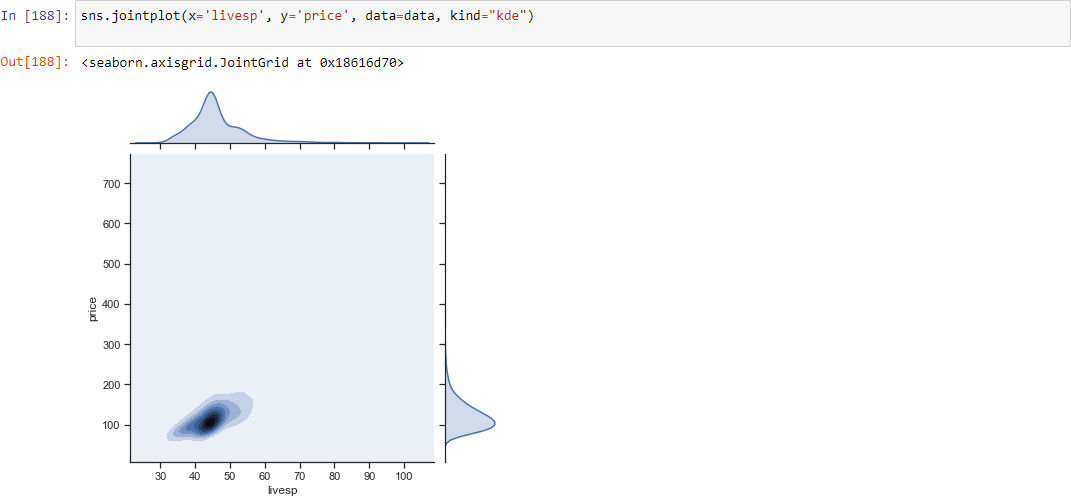
Исследуем вероятность для поля цена (price):



Полученные результаты практически совпадают с нормальным распределением (распределением Гаусса).

Для наглядности можно построить комбинированный график с гистограммой и диаграммой рассеяния:





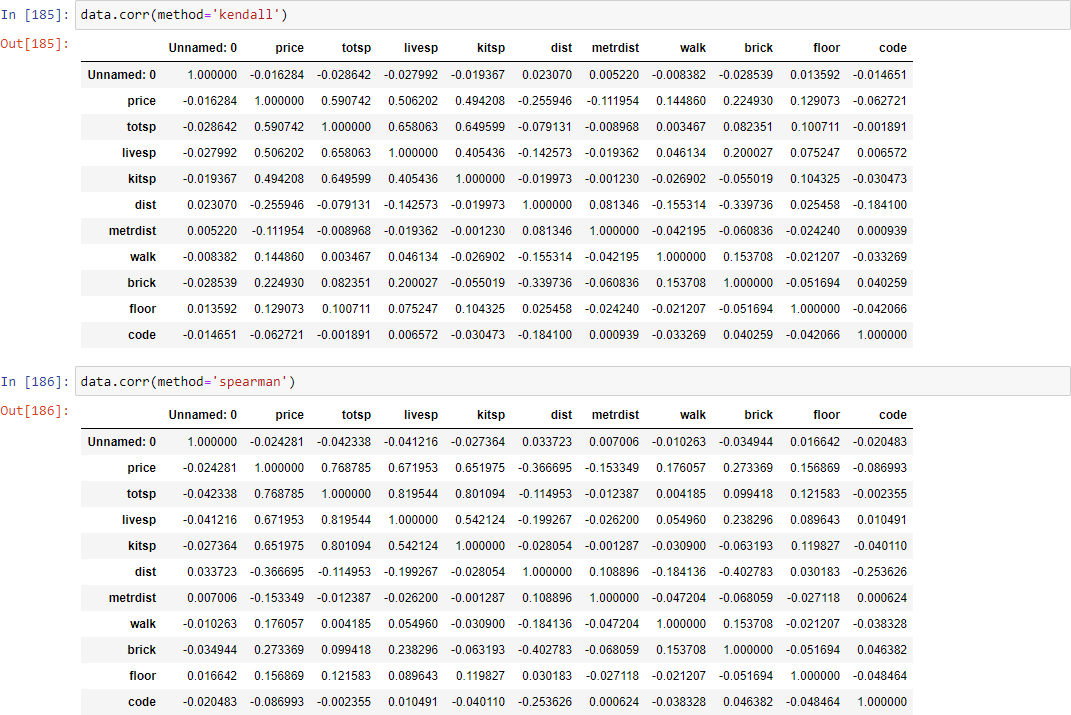
1. Исследование корреляции признаков

Проверка корреляции признаков позволяет решить две задачи:

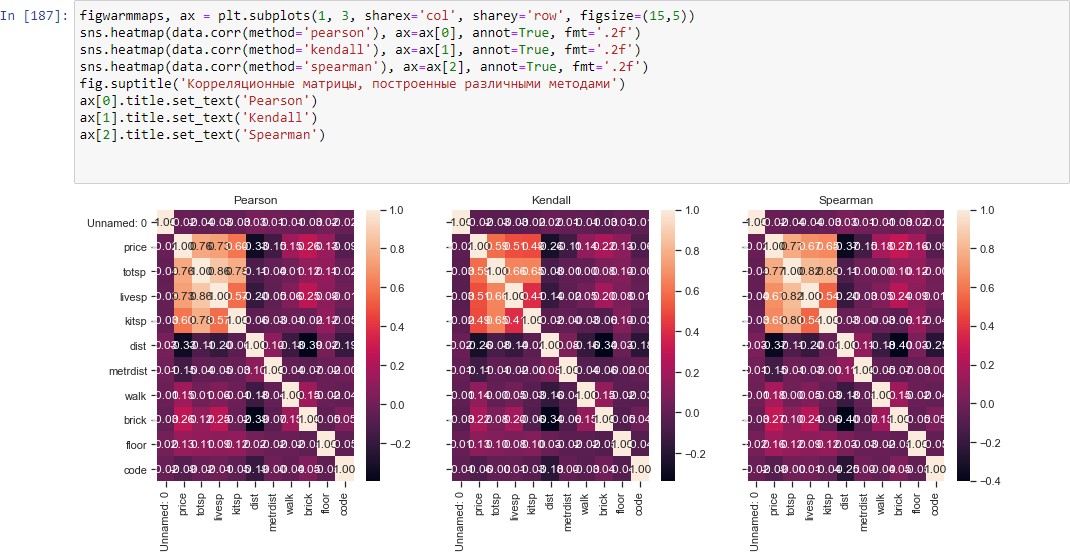
* 1. Понять, какие признаки наиболее сильно коррелируют с целевым признаком. Именно эти признаки будут наиболее информативными для моделей машинного обучения. Признаки, которые слабо коррелируют с целевым признаком, можно попробовать исключить из построения модели, иногда это повышает качество модели. Некоторые алгоритмы машинного обучения автоматически определяют ценность того или иного признака для построения модели.
  2. Понять, какие нецелевые признаки линейно зависимы между собой. Линейно зависимые признаки, как правило, очень плохо влияют на качество моделей. Поэтому если несколько признаков линейно зависимы, то для построения модели из них выбирают какой-то один признак.

Существуют разные алгоритмы построения корреляционной матрицы:





Для наглядности можно составить тепловую карту:



Таким образом, можно выделить признаки, слабо влияющие на целевой признак – цену (price): регион (code), этаж (floor), brick, до метро пешком/на транспорте (walk), расстояние до метро (metrdist) и от центра (dist). В основном цена на квартиру в Москве зависит от площади комнат.