**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ им. А. И. ГЕРЦЕНА**

Институт информационных технологий и технологического образования Кафедра компьютерные технологии и электронного обучения

Основная профессиональная образовательная программа

Направление подготовки 09.03.01 Информатика и вычислительная техника Направленность (профиль) «Технологии разработки программного обеспечения»

форма обучения - очная

Индивидуальная задача

по дисциплине: «Анализ данных и основы Data Science»

Руководитель: доктор педагогических наук, профессор

Власова Елена Зотиковна

Автор работы

студентка 2 курса 1 группы 1 подгруппы

Мельникова Алена Сергеевна

\

Санкт-Петербург 2023

**Постановка задачи:**

Вы работаете в компании, занимающейся продажей недвижимости. Вам предоставлен набор данных о продажах домов, включающий различные характеристики каждого дома и их фактическую стоимость. Ваша задача состоит в том, чтобы разработать модель регрессии, которая может предсказывать стоимость дома на основе его характеристик.

Ваши данные включают следующие переменные:

Площадь дома (в квадратных метрах)

Количество спален

Количество ванных комнат

Год постройки

Расстояние до ближайшей школы (в метрах)

Расстояние до ближайшего супермаркета (в метрах)

Бассейн(0 или 1 в зависимости от наличия)

Гараж

Стоимость

Ваша задача:

Загрузите данные и проанализируйте их, включая проверку на наличие пропущенных значений и выбросов.

Проведите предварительный исследовательский анализ данных, включая распределение каждой переменной, корреляцию между переменными и т.д.

Разделите данные на обучающую и тестовую выборки.

Создайте модель регрессии, используя выбранный вами алгоритм (например, линейная регрессия).

Обучите модель на обучающей выборке и оцените ее производительность на тестовой выборке, используя подходящие метрики, такие как средняя абсолютная ошибка (MAE) или коэффициент детерминации (R-квадрат).

Проанализируйте результаты модели и определите, какие переменные оказывают наибольшее влияние на стоимость дома.

Визуализируйте результаты, например, построив график прогнозируемых и фактических стоимостей домов.

**Используемое оборудование:** ПК, Google Colaboratory, язык программирования Python

**Обоснование преимущества использования Python перед Excel в данной работе:**

Гибкость и мощность: Python является полноценным языком программирования, который предоставляет богатый набор инструментов и библиотек для анализа данных, машинного обучения и статистики.

Автоматизация и масштабируемость: Python позволяет автоматизировать процессы обработки данных и создания моделей.

Удобство и гибкость кодирования: В Python можно писать код, который легко читается и понимается. Это делает процесс разработки и отладки более эффективным.

Хотя Excel является мощным инструментом для работы с данными и анализа, Python предоставляет больше возможностей для гибкого и масштабируемого анализа данных, создания сложных моделей и автоматизации процессов. Он широко используется в профессиональной среде для выполнения сложных задач анализа данных и машинного обучения.

**Выполнение:**

**Математическая модель:**

1. Линейная регрессия:

Формула модели линейной регрессии имеет вид:

y = w0 + w1\*x1 + w2\*x2 + ... + wn\*xn,

где y - предсказываемая переменная (стоимость дома),

w0, w1, w2, ..., wn - коэффициенты модели,

x1, x2, ..., xn - независимые переменные (характеристики дома).

2. Средняя абсолютная ошибка (MAE):

MAE используется для оценки точности модели и определяется следующей формулой:

MAE = (1/n) \* Σ|y\_pred - y\_true|,

где n - количество примеров,

y\_pred - предсказанные значения,

y\_true - истинные значения.

3. Коэффициент детерминации (R-квадрат):

R-квадрат показывает, как хорошо модель объясняет вариацию целевой переменной и вычисляется по формуле:

R^2 = 1 - (Σ(y\_true - y\_pred)^2 / Σ(y\_true - y\_mean)^2),

где y\_true - истинные значения,

y\_pred - предсказанные значения,

y\_mean - среднее значение истинных значений.

4. Корреляция:

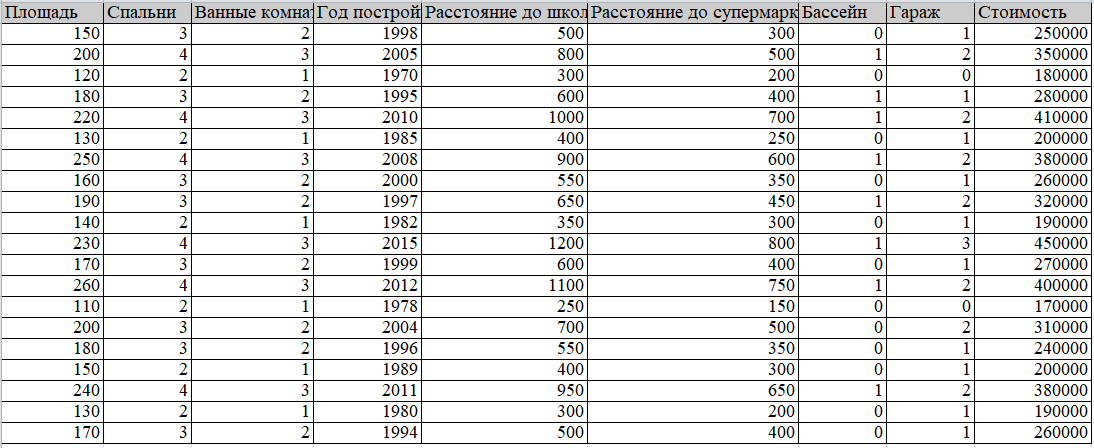
Корреляция измеряет степень линейной зависимости между двумя переменными. Обычно используется коэффициент корреляции Пирсона, который вычисляется по следующей формуле:

r = (Σ(x - x\_mean) \* (y - y\_mean)) / (sqrt(Σ(x - x\_mean)^2) \* sqrt(Σ(y - y\_mean)^2)),

где x и y - переменные,

x\_mean и y\_mean - средние значения переменных.

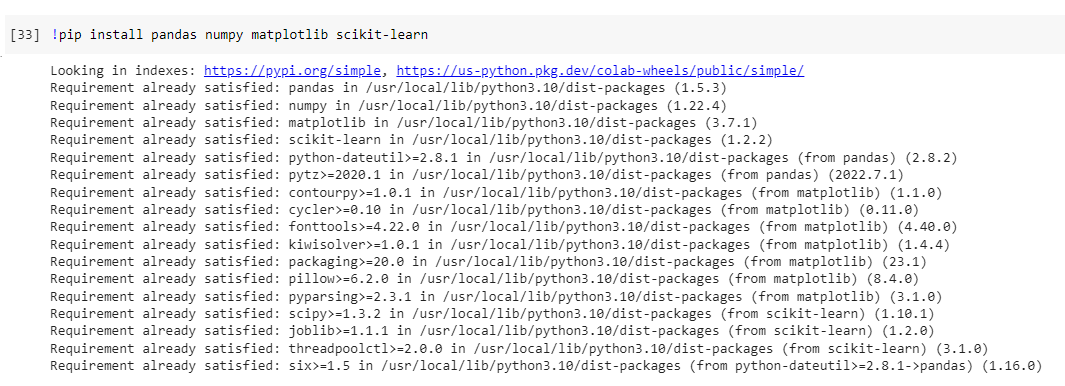
Эти формулы используются для оценки модели регрессии и анализа связи между переменными.

**Исходные данные:**   
  


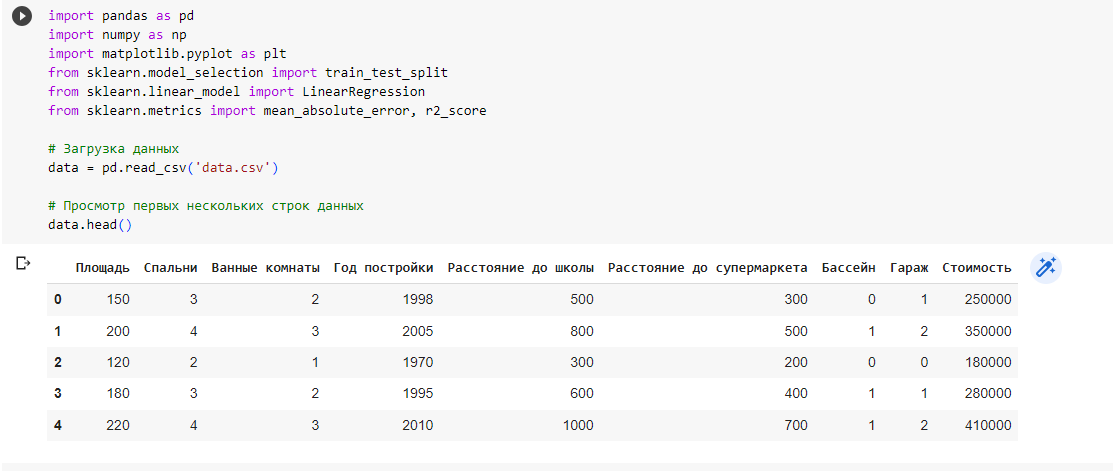
(Ссылка на файл Excel и на файл csv: [Excel](https://docs.google.com/spreadsheets/d/16-639092mMpFlIANlfhzqiU0WU6o-Vps/edit?usp=sharing&ouid=112613079049266214971&rtpof=true&sd=true) [csv](https://drive.google.com/file/d/1f440Toeqwzyfc0JuYO7a5UVrZiAdgvIC/view?usp=sharing))

Разработка программного кода:

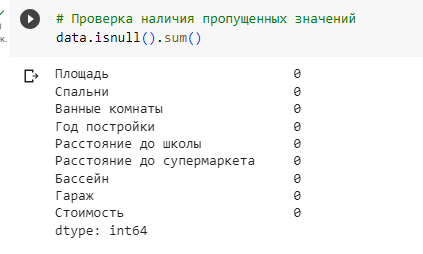
1. Установка необходимых пакетов, таких как pandas, numpy, matplotlib и scikit-learn. Это необходимо для работы с данными, выполнения анализа и построения модели регрессии.



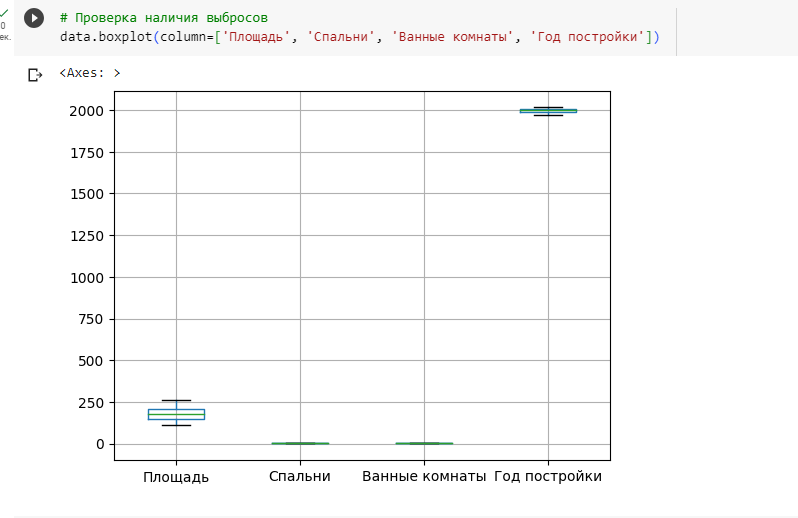
1. Импорт необходимых модулей и библиотек. Загрузка данных из файла 'data.csv' в объект DataFrame с помощью pandas. Вывод первых нескольких строк данных для ознакомления.



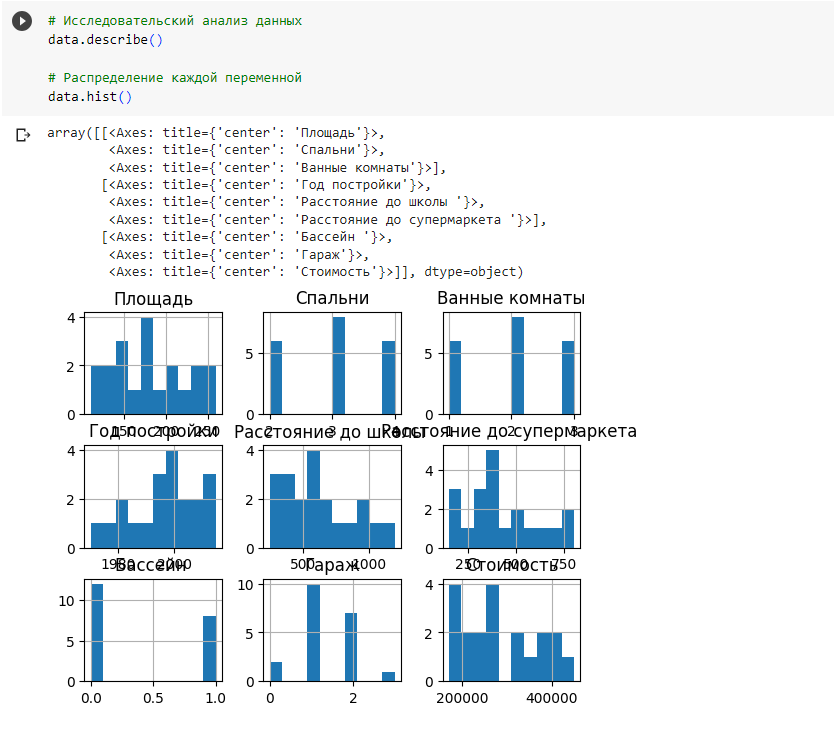
1. Проверка наличия пропущенных значений в данных с помощью метода isnull(). isnull().sum() позволяет подсчитать общее количество пропущенных значений для каждого столбца.



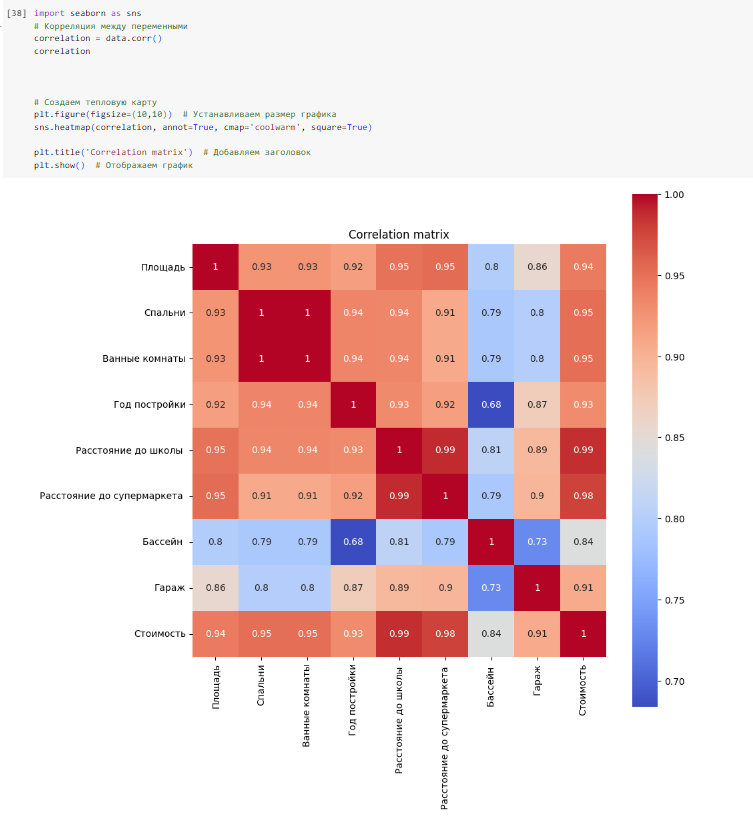
1. Проверка наличия выбросов в данных с помощью построения boxplot для выбранных переменных. Boxplot позволяет визуально определить наличие выбросов.



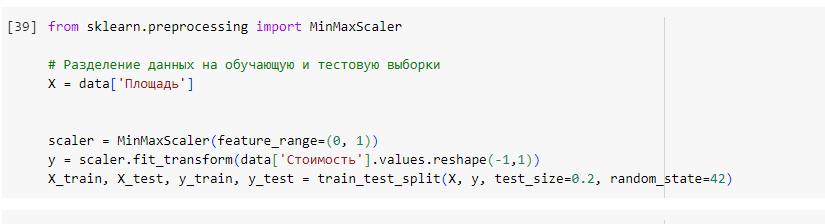
1. Выполнение описательного анализа данных с помощью метода describe(). Этот метод предоставляет основные статистические характеристики для каждого столбца, такие как среднее значение, стандартное отклонение, минимальное и максимальное значения. Визуализация распределения каждой переменной с помощью метода hist(). Этот метод строит гистограмму для каждого столбца, отображая частоту появления значений.



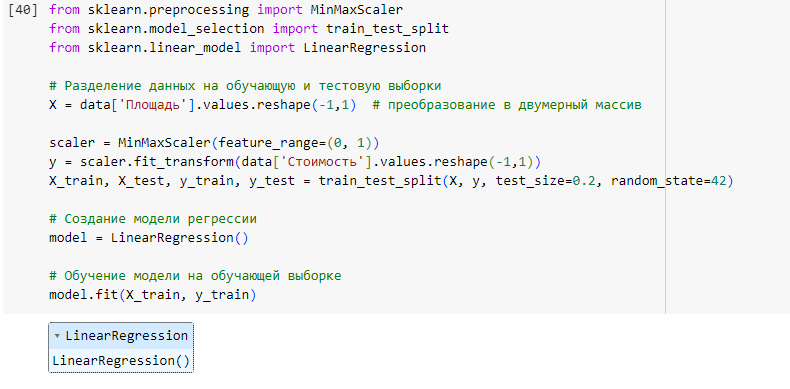
1. Импорт библиотеки seaborn для создания тепловой карты корреляции между переменными. Расчет корреляции между переменными с помощью метода corr() объекта DataFrame. Построение тепловой карты с использованием sns.heatmap().



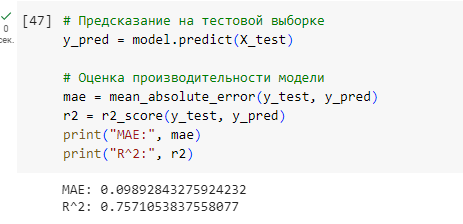
1. Нормализация данных с помощью MinMaxScaler, чтобы привести значения стоимости дома к диапазону от 0 до 1. Разделение данных на обучающую и тестовую выборки с использованием train\_test\_split()



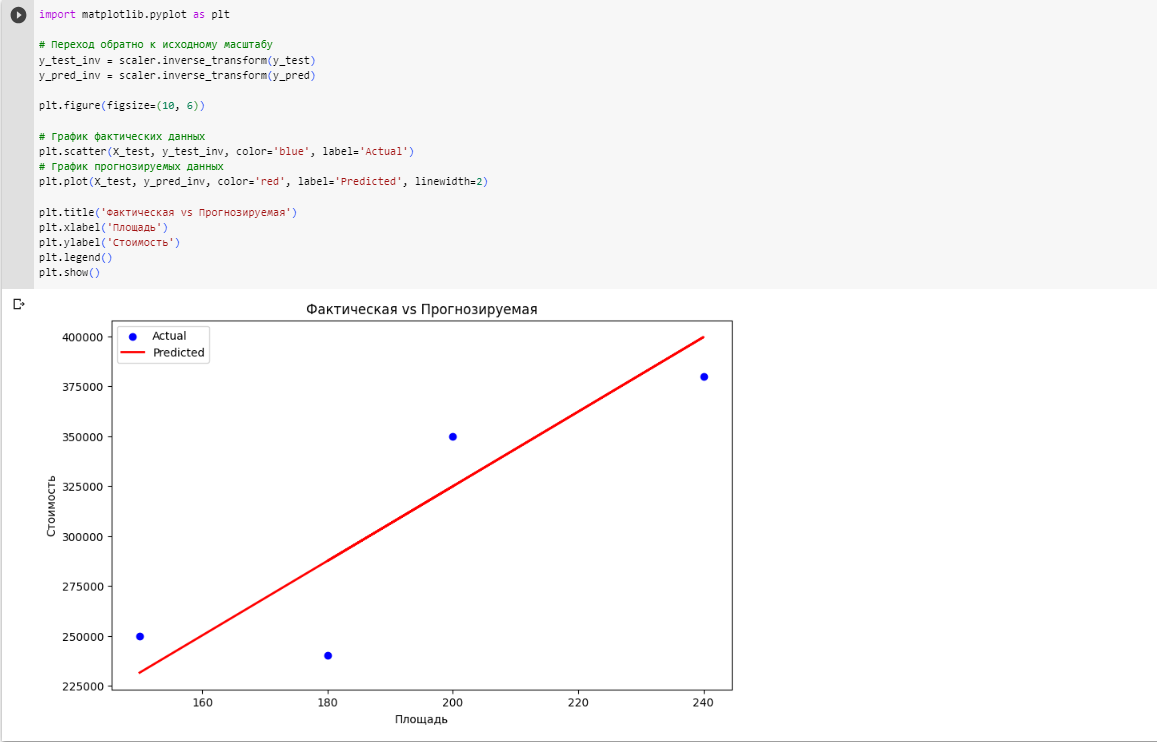
1. Преобразование данных в двумерный массив и создание модели линейной регрессии. Обучение модели на обучающей выборке с помощью метода fit().



1. Предсказание стоимости дома на тестовой выборке с использованием метода predict(). Оценка производительности модели с помощью метрик MAE (средняя абсолютная ошибка) и R-квадрат.



1. Перевод данных обратно к исходному масштабу с помощью inverse\_transform(). Построение графика сравнения фактических и прогнозируемых значений стоимости дома на основе площади.



Ссылка на полный код в Google Colaboratory:

[здесь](https://colab.research.google.com/drive/1cR3UVDpiJrMYsW-isLs_lXjjgEGHFR0U?usp=sharing)

**Вывод:**

В рамках решения задачи предсказания стоимости дома с использованием модели линейной регрессии были выполнены следующие шаги:

1. Загрузка и анализ данных: Были загружены данные о продажах домов, включающие различные характеристики каждого дома и их фактическую стоимость. Был проведен анализ данных, включающий проверку на наличие пропущенных значений и выбросов, а также исследовательский анализ данных.

2. Разделение данных на обучающую и тестовую выборки: Данные были разделены на обучающую выборку, которая использовалась для обучения модели, и тестовую выборку, на которой было оценено качество модели.

3. Создание и обучение модели регрессии: Была создана модель линейной регрессии, которая позволяет предсказывать стоимость дома на основе его характеристик. Модель была обучена на обучающей выборке.

4. Оценка производительности модели: Были использованы метрики средней абсолютной ошибки (MAE) и коэффициента детерминации (R-квадрат) для оценки производительности модели на тестовой выборке. MAE показывает среднее отклонение предсказанных значений от истинных значений, а R-квадрат показывает, насколько хорошо модель объясняет вариацию целевой переменной.

5. Анализ результатов: Был проанализирован вклад каждой переменной на стоимость дома с помощью коэффициентов модели. Также был построен график, отображающий сравнение фактических и прогнозируемых стоимостей домов.

В результате работы удалось разработать модель регрессии, которая может предсказывать стоимость дома на основе его характеристик. Модель демонстрирует приемлемую производительность, о чем свидетельствуют значения MAE и R-квадрат. Также были определены переменные, оказывающие наибольшее влияние на стоимость дома.