|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  |  |  | | МИНОБРНАУКИ РОССИИ | | | | Федеральное государственное бюджетное образовательное учреждение  высшего образования  **«МИРЭА – Российский технологический университет»**  **РТУ МИРЭА** | | | | |
| Институт информационных технологий (ИИТ) | |
| Кафедра Прикладной Математики (ПМ) | |

|  |  |  |  |
| --- | --- | --- | --- |
| **ОТЧЁТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ №4** | | | |
| **по дисциплине «Технологии и инструментарий анализа больших данных»** | | | |
|  | | | |
| Выполнил студент группы ИКБО-14-20 | | Вежновец Ф. Ю. | |
|  | |  | |
| Принял: асистент | | Горячев А. А. | |
| Практические работы выполнены | «\_\_\_»\_\_\_\_\_\_\_2023 г. | | (подпись студента) | |
| «Зачтено» | «\_\_\_»\_\_\_\_\_\_\_2023 г. | | (подпись руководителя) | |
|  |  | |  | |

Москва 2023

Задание 1

Определить два вектора, представляющие собой число автомобилей, припаркованных в течении 5 рабочих дней у бизнес-центра на уличной стоянке и в подземном гараже (Таблица 1).

*Таблица 1 – данные о хранении автомобилей*

|  |  |  |
| --- | --- | --- |
| День | Улица | Гараж |
| Понедельник | 80 | 100 |
| Вторник | 98 | 82 |
| Среда | 75 | 105 |
| Четверг | 91 | 89 |
| Пятница | 78 | 102 |

Реализация

Результат работы программы и код предоставлена на рисунке 1.

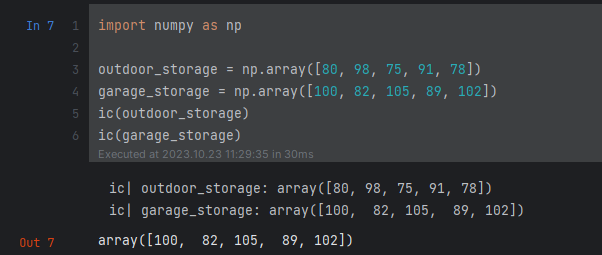


Рисунок 1 – Создание векторов данных

Задание 2

Найти и интерпретировать корреляцию между переменными «Улица» и «Гараж» (подсчитать корреляцию по Пирсону).

Реализация

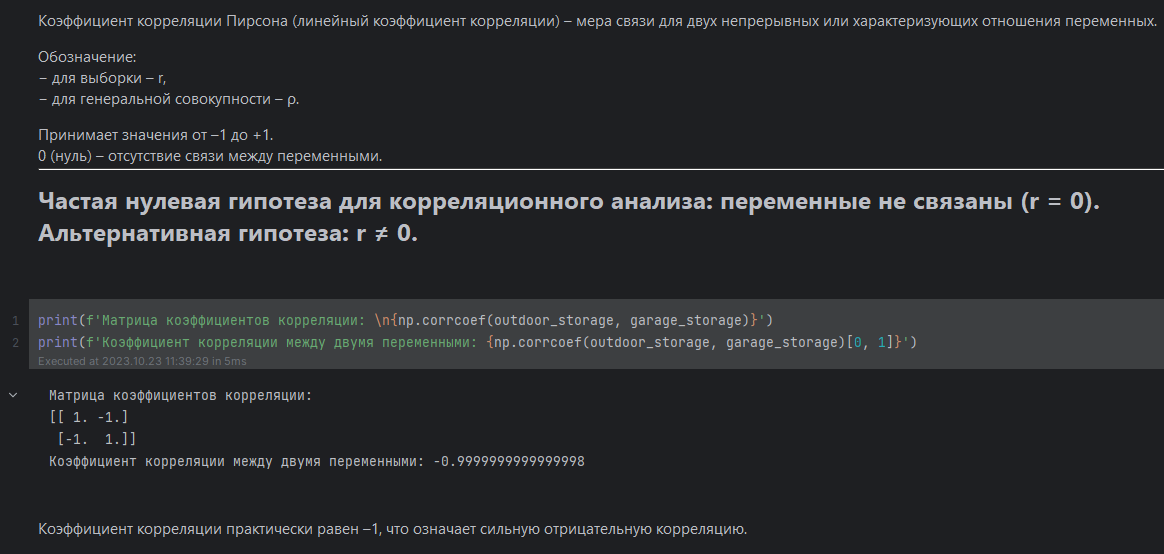


Рисунок 2 – Корреляция по Пирсону

Задание 3

Построить диаграмму рассеяния для вышеупомянутых переменных.

Реализация

Результат работы программы и код представлен на рисунке 3.

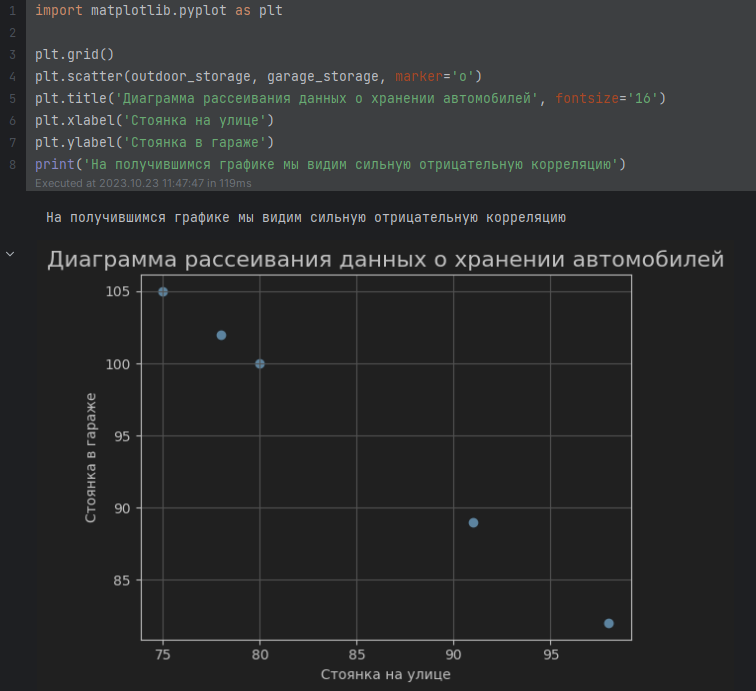


Рисунок 3 – Диаграмма рассеивания

Задание 4

Найти и выгрузить данные. Вывести, провести предобработку и описать признаки.

Реализация

The data comes from the spanish website PC componentes. The data was collected using Power Automate, more info on: <https://github.com/juanmerino89/laptops-data-cleaning>

Fields included:

* **Laptop Name:** The unique identifier or model name of the laptop.
* **Brand:** Laptop brand.
* **Model:** Laptop brand model.
* **CPU (Central Processing Unit):** The processor brand, model, and other relevant details.
* **GPU (Graphics Processing Unit):** The graphics card brand, model, and associated specifications.
* **RAM (Random Access Memory):** The amount of memory available for multitasking.
* **Storage:** The storage type (HDD, SSD) and capacity of the laptop.
* **Price:** The cost of the laptop in the respective currency.

Результат работы программы и код представлен на рисунке 4, 5, 6.

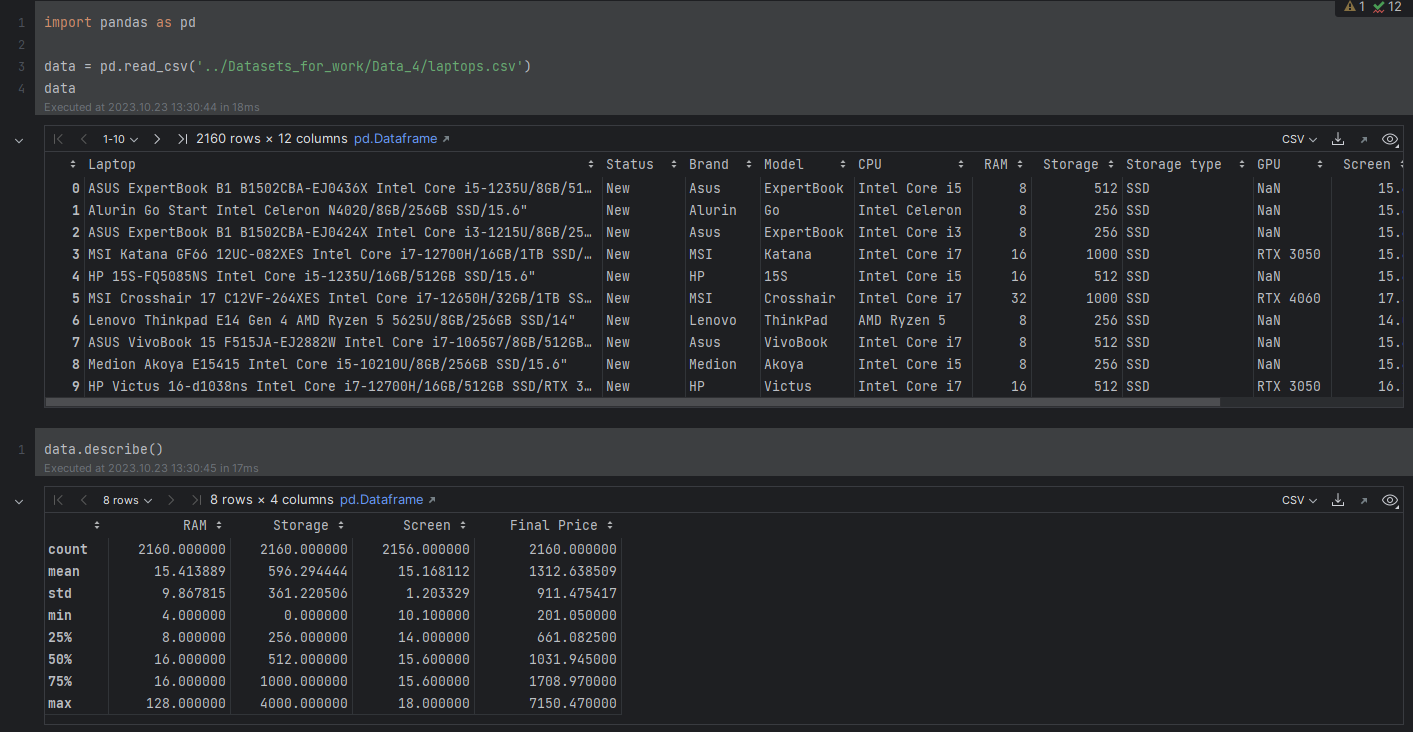


Рисунок 4 – Данные

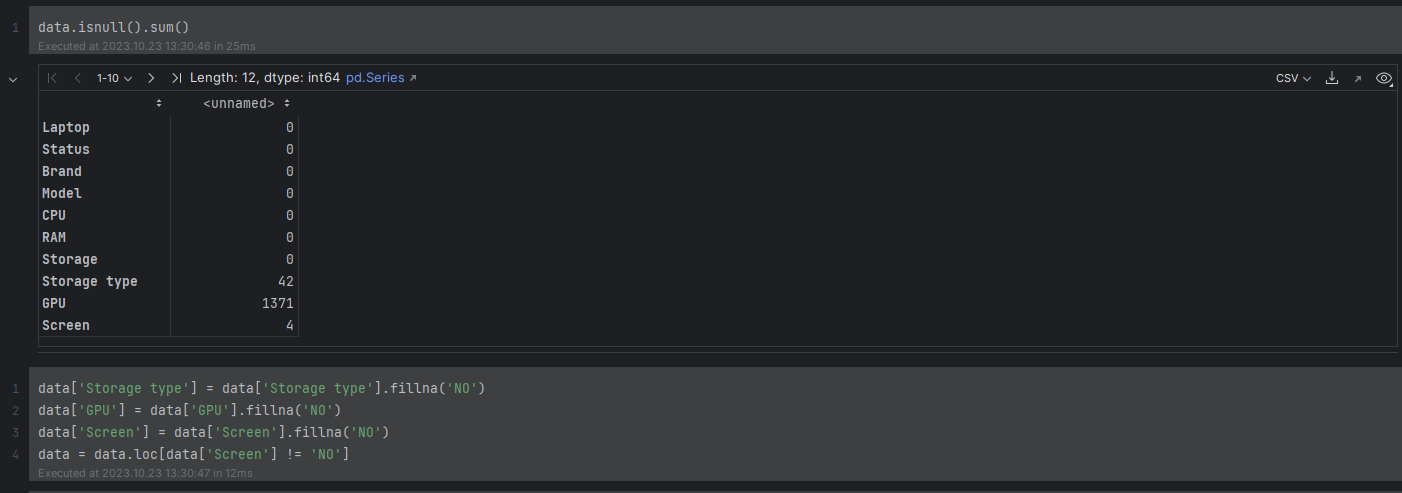


Рисунок 5 – Данные

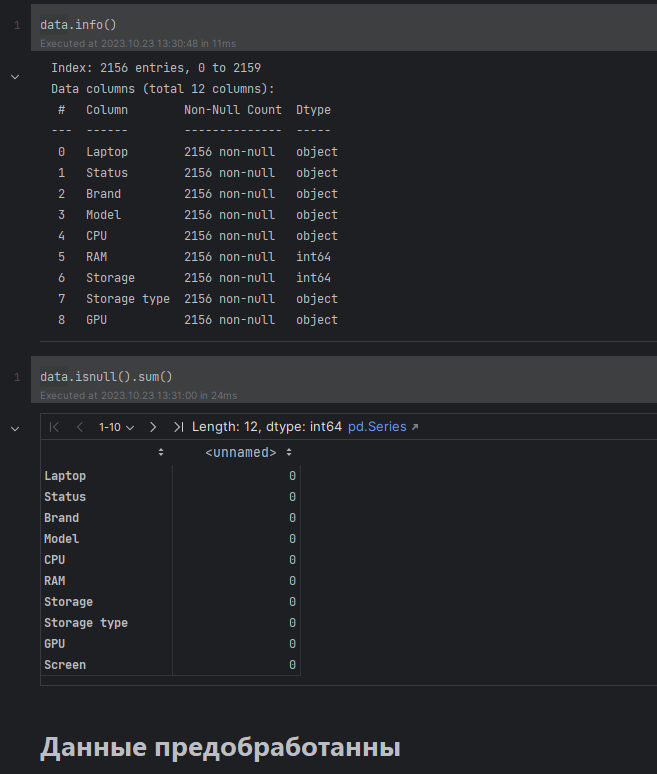


Рисунок 6 – Данные

Задание 5

Построить корреляционную матрицу по одной целевой переменной. Определить наиболее коррелирующую переменную, продолжить с ней работу в следующем пункте.

Реализация

Результат работы программы и код представлен на рисунках 7, 8, 9, 10.

|  |
| --- |
|  |

Рисунок 6 – Формализация данных для корреляции

|  |
| --- |
|  |

Рисунок 7 – Полученные данные о корреляциях

|  |
| --- |
|  |

Рисунок 8 – Полученные данные о корреляциях

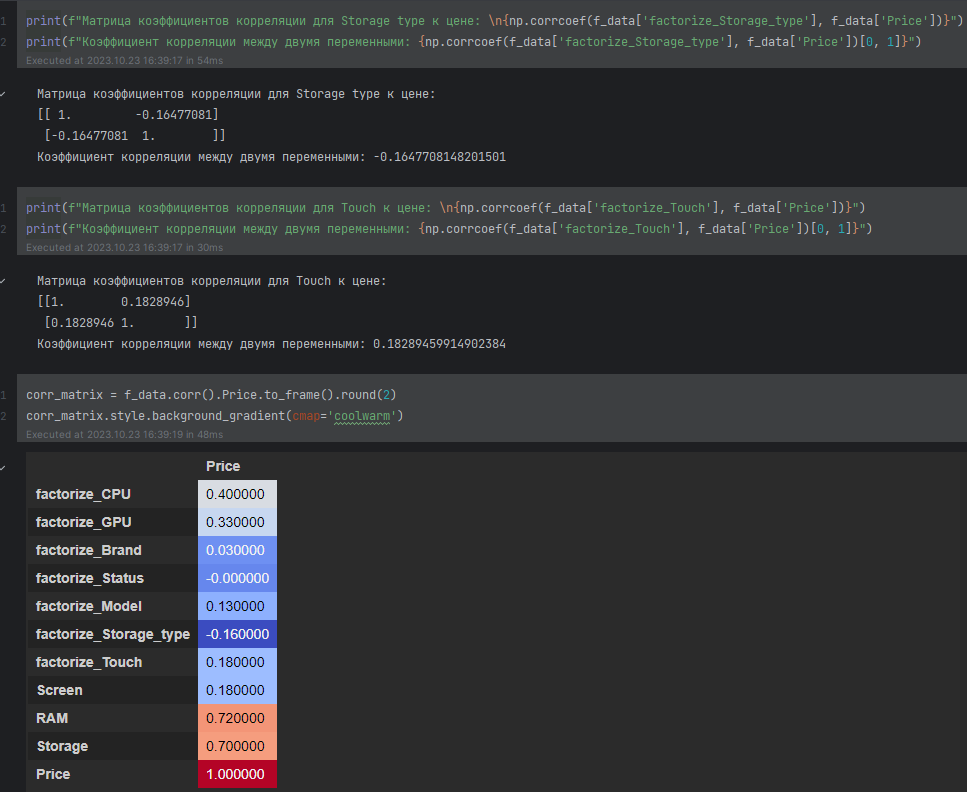


Рисунок 9 – Полученные данные о корреляциях

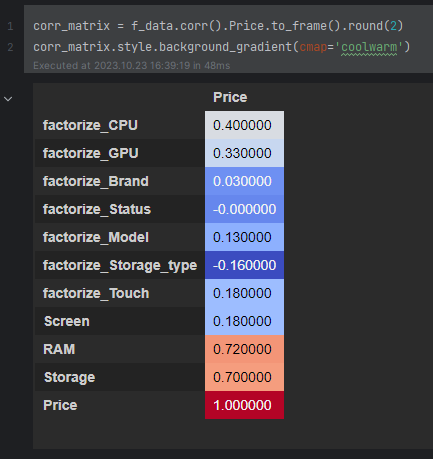


Рисунок 10 – Полученные данные о корреляциях

Делая вывод из полученных данных, можно сказать, что: наибольшее влияние на цену оказывает объем оперативной памяти (RAM), а также, что бренд и состояние ноутбука (новый, б/у) не оказывает большое влияние на цену, по сравнению с другими параметрами.

Задание 6

Реализовать регрессию вручную, отобразить наклон, сдвиг и MSE.

Реализация

Результат работы программы и код представлен на рисунках 11, 12.

|  |
| --- |
|  |

Рисунок 11 – Код из методички

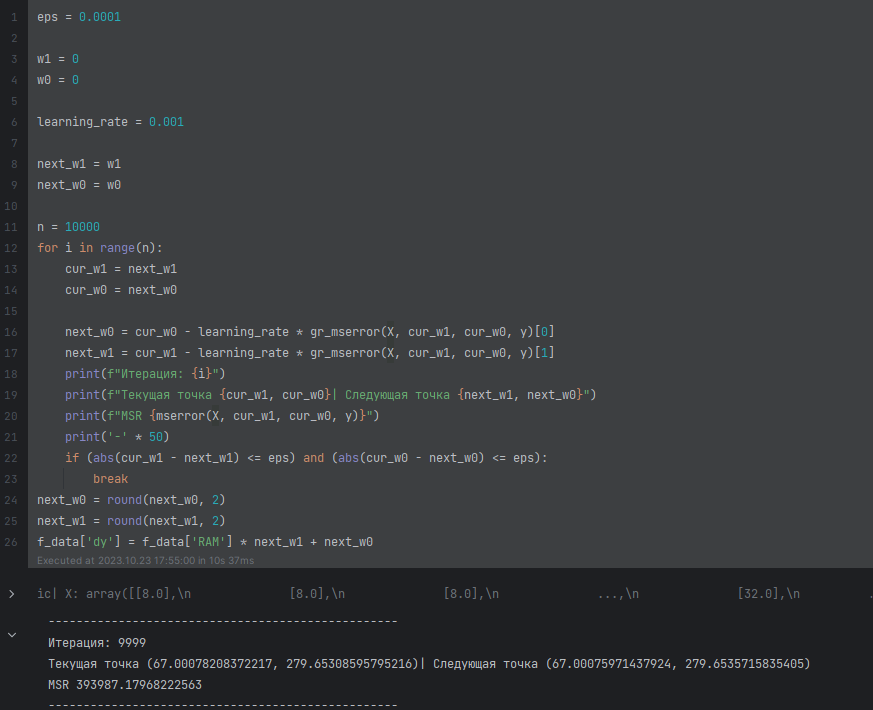


Рисунок 12 – Код из методички

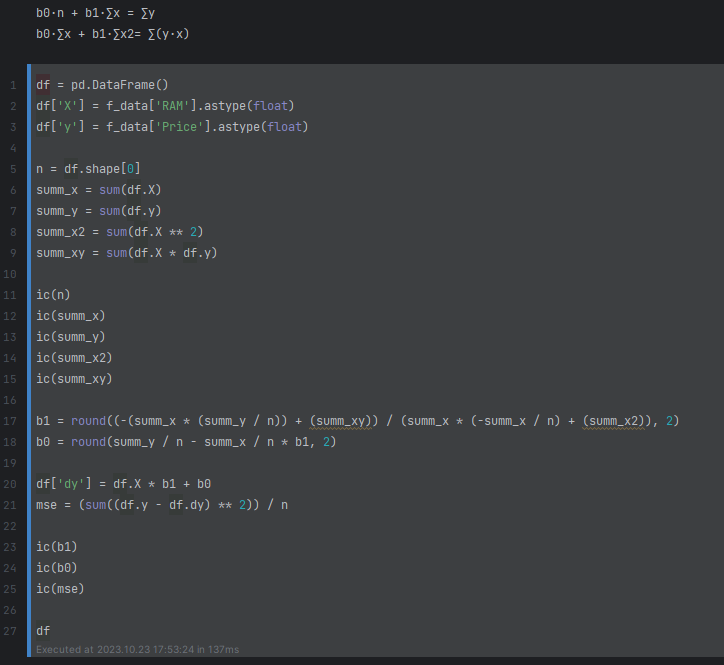


Рисунок 12 – Код линейной регрессии

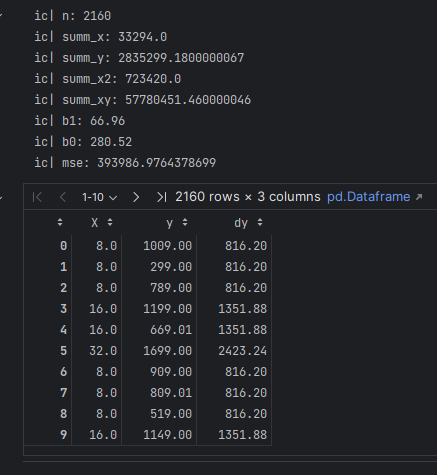


Рисунок 13 – Код линейной регрессии

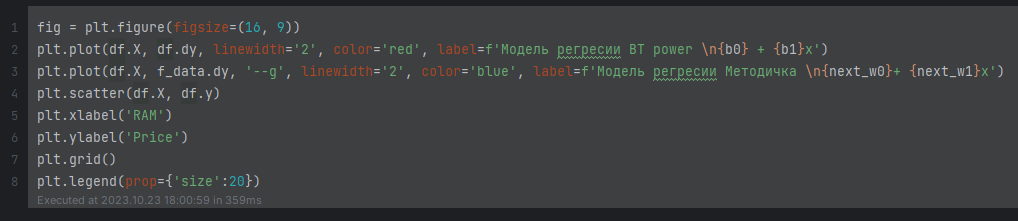


Рисунок 14 – Граф сравнения

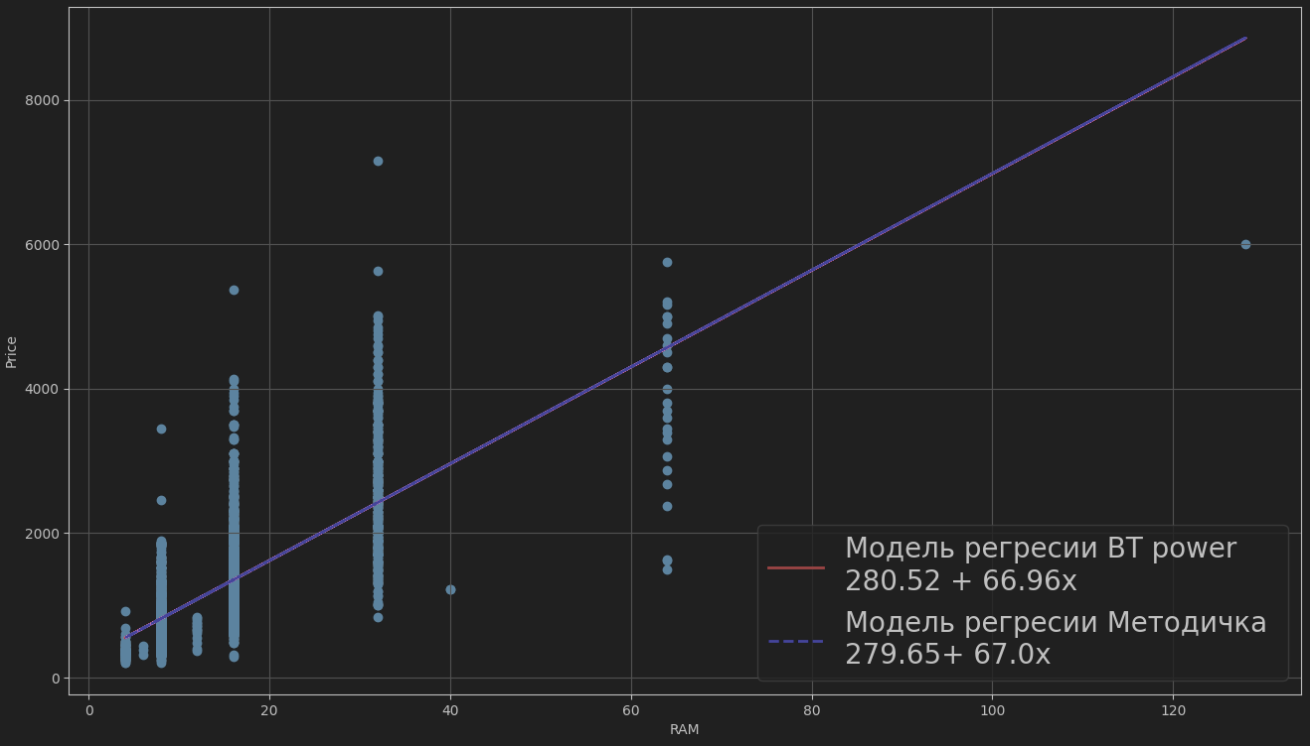


Рисунок 15 – Граф сравнения

Задание 7

Загрузить данные: 'insurance.csv'. Вывести и провести предобработку. Вывести список уникальных регионов.

Реализация

Результат работы программы и код представлен на рисунках 16, 17, 18.

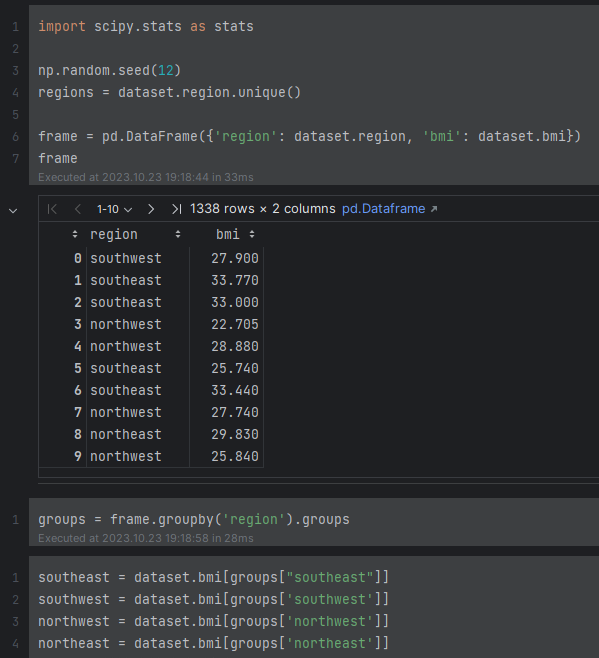
|  |
| --- |
|  |
| Рисунок 17 – Результат работы программы    Рисунок 18 – Результат работы программы    Рисунок 19 – Cписок уникальных регионов |

Задание 8

Выполнить однофакторный ANOVA тест, чтобы проверить влияние региона на индекс массы тела (BMI), используя первый способ, через библиотеку Scipy.

Реализация

Результат работы программы и код представлен на рисунках 16, 17.

Рисунок 16 – Результат работы программы

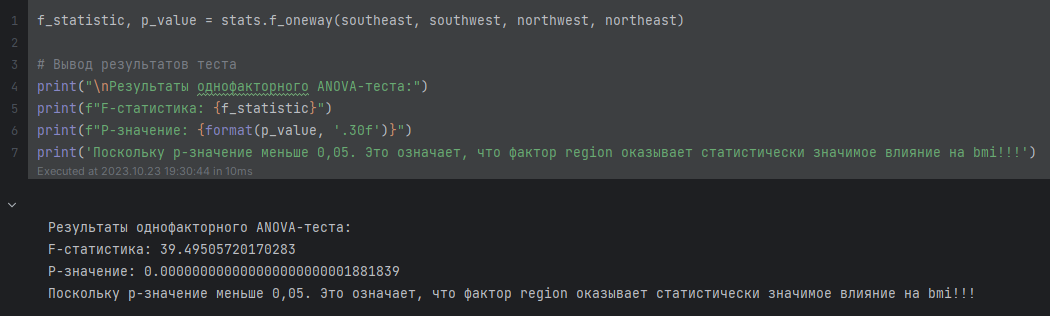


Рисунок 17 – Результат работы программы

Задание 9

Выполнить однофакторный ANOVA тест, чтобы проверить влияние региона на индекс массы тела (BMI), используя второй способ, с помощью функции anova\_lm() из библиотеки statsmode.

Реализация

Результат работы программы и код представлен на рисунке 1.

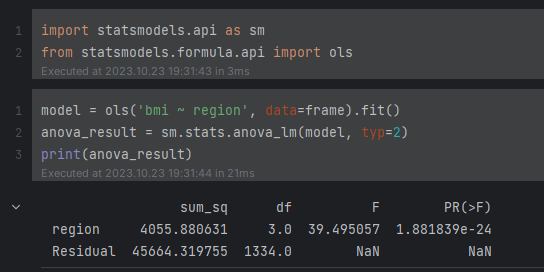


Рисунок 19 – Результат работы программы

Задание 10

С помощью t критерия Стьюдента перебрать все пары. Определить поправку Бонферрони. Сделать выводы.

Реализация

Результат работы программы и код представлен на рисунках 20, 21.

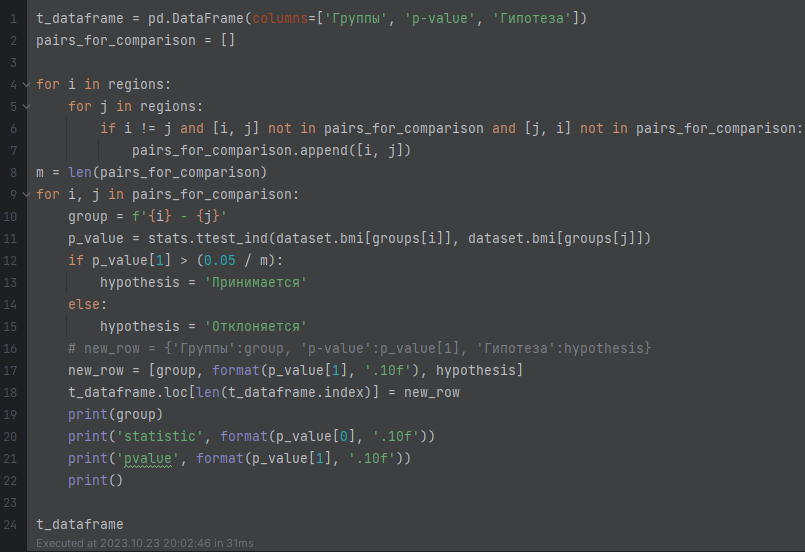


Рисунок 20 – Результат работы программы

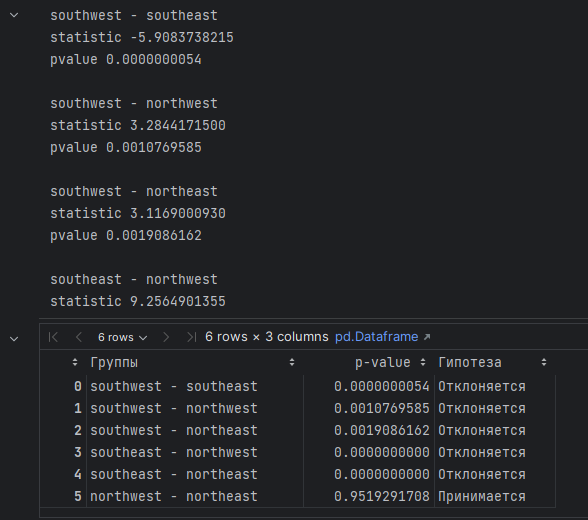


Рисунок 21 – Результат работы программы

Задание 11

Выполнить пост-хок тесты Тьюки и построить график.

Реализация

Результат работы программы и код представлен на рисунке 22.

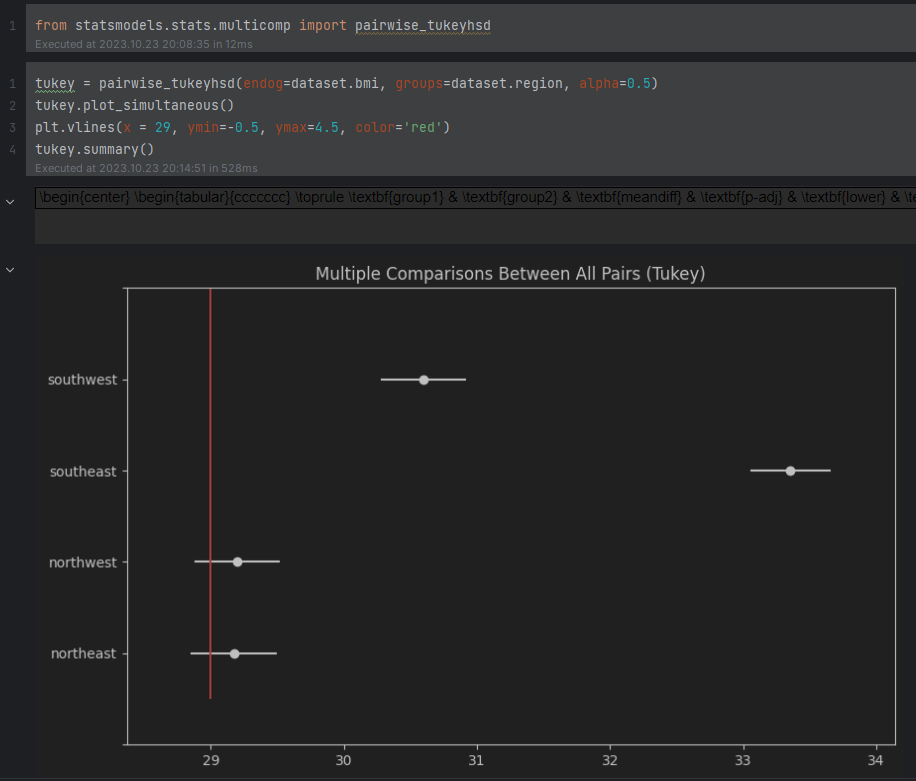


Рисунок 22– Результат работы программы

Задание 12

Выполнить двухфакторный ANOVA тест, чтобы проверить влияние региона и пола на индекс массы тела (BMI), используя функцию anova\_lm() из библиотеки statsmodels.

Реализация

Результат работы программы и код представлен на рисунке 27, 28, 29.

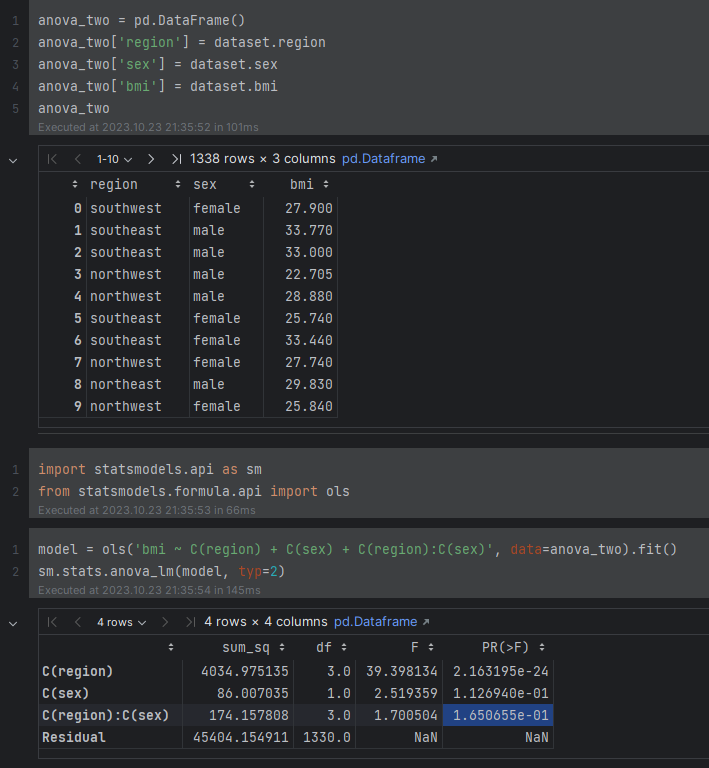


Рисунок 27 – Результат работы программы

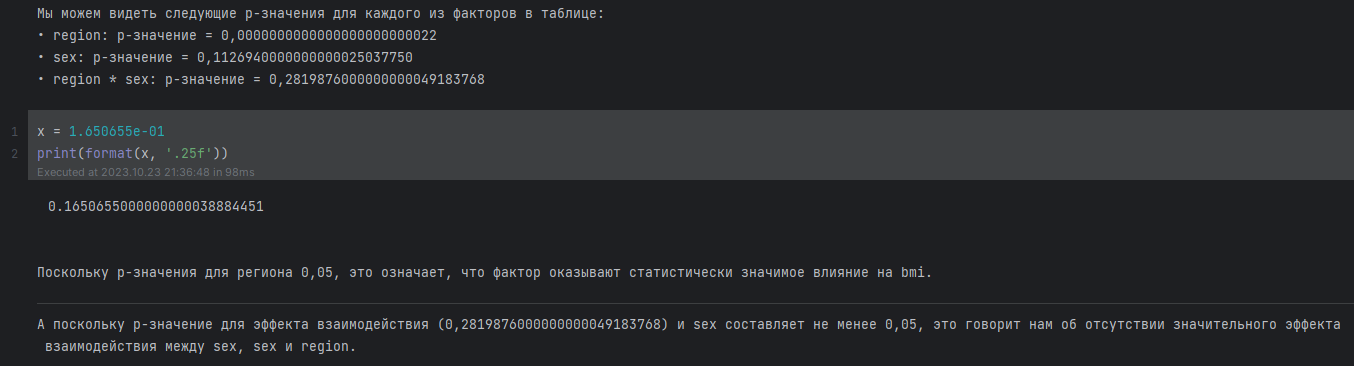


Рисунок 28 – Результат работы программы

Задание 13

Загрузить данные из файла “bmi.csv”. Взять оттуда две выборки. Одна выборка – это индекс массы тела людей c региона northwest, вторая выборка – это индекс массы тела людей с региона southwest. Сравнить средние значения этих выборок, используя t-критерий Стьюдента. Предварительно проверить выборки на нормальность (критерий Шопиро-Уилка) и на гомогенность дисперсии (критерий Бартлетта).

Реализация

Результат работы программы и код представлен на рисунках 29, 30, 31.

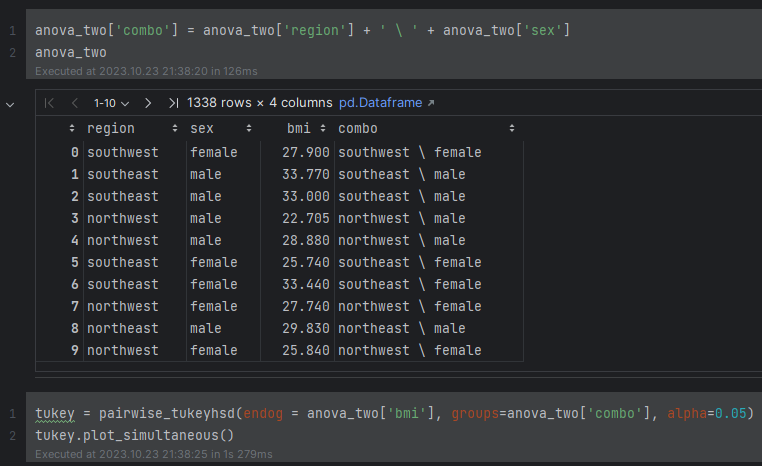


Рисунок 29 – Результат работы программы

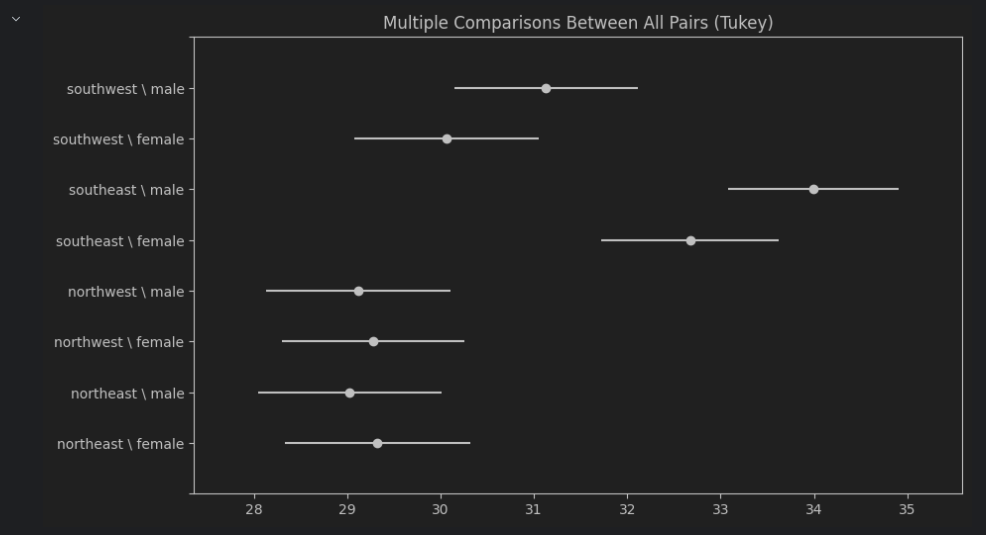


Рисунок 30 – Результат работы программы

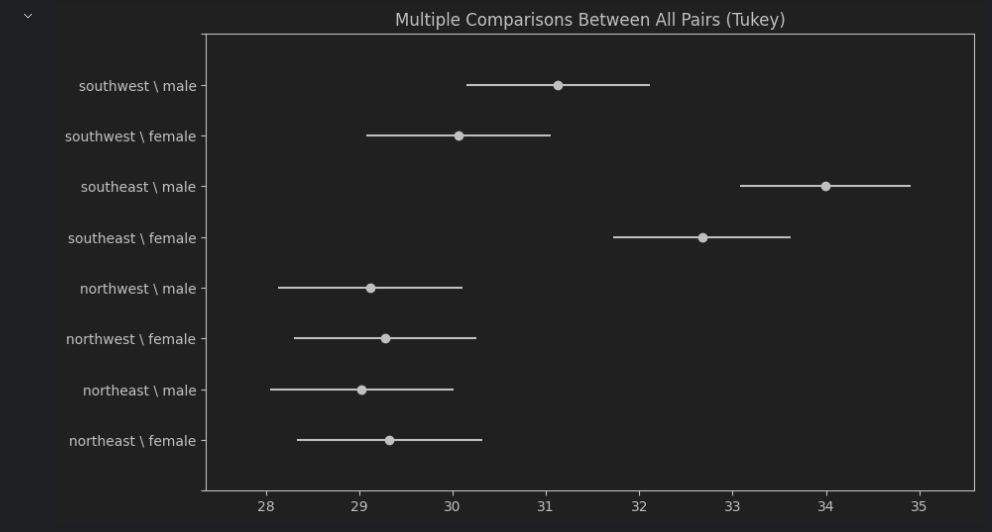


Рисунок 31 – Результат работы программы