|  |
| --- |
| МИНОБРНАУКИ РОССИИ |
| Федеральное государственное бюджетное образовательное учреждение высшего образования **«МИРЭА − Российский технологический университет»**  **РТУ МИРЭА** |

**Институт информационных технологий (ИИТ)**

**Кафедра прикладной математики (ПМ)**

**ОТЧЕТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ**

по дисциплине «Технологии и инструментарий анализа больших данных»

**Практическая работа № 4**

|  |  |  |  |
| --- | --- | --- | --- |
| Выполнил студент группы ИНБО-01-17 | ИВБО-06-20, Чурилов А.К. | (подпись) | |
| Принял | Парамонов А.А. | (подпись) | |
| Отчет представлен | «9» ноября 2023 г. | |  | |

Москва 2022 г.

**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.

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

1. Найти и интерпретировать корреляцию между переменными «Улица» и «Гараж» (подсчитать корреляцию по Пирсону).
2. Построить диаграмму рассеяния.
3. Найти и выгрузить данные. Вывести, провести предобработку и описать признаки.
4. Построить корреляционную матрицу по одной целевой переменной. Определить наиболее коррелирующую переменную, продолжить с ней работу в следующем пункте.
5. Реализовать регрессию вручную, отобразить наклон, сдвиг и MSE.
6. Визуализировать регрессию на графике.
7. Загрузить данные: 'insurance.csv'. Вывести и провести предобработку. Вывести список уникальных регионов.
   1. Выполнить однофакторный ANOVA тест, чтобы проверить влияние региона на индекс массы тела (BMI), используя первый способ, через библиотеку Scipy.
   2. Выполнить однофакторный ANOVA тест, чтобы проверить влияние региона на индекс массы тела (BMI), используя второй способ, с помощью функции anova\_lm() из библиотеки statsmodels.
   3. С помощью t критерия Стьюдента перебрать все пары. Определить поправку Бонферрони. Сделать выводы.
   4. Выполнить пост-хок тесты Тьюки и построить график.
   5. Выполнить двухфакторный ANOVA тест, чтобы проверить влияние региона и пола на индекс массы тела (BMI), используя функцию anova\_lm() из библиотеки statsmodels.
   6. Выполнить пост-хок тесты Тьюки и построить график.
   7. Оформить отчет о проделанной работе, написать выводы.

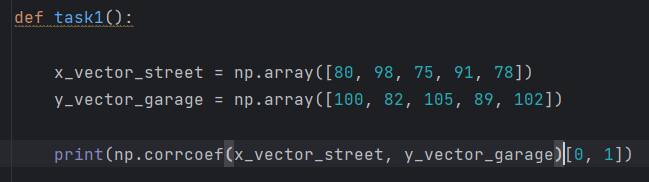
**Ход работы**

**Задание 1**

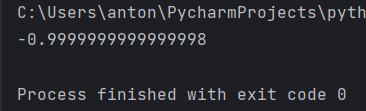
Определить два вектора, представляющие собой число автомобилей, припаркованных в течении 5 рабочих дней у бизнес-центра на уличной стоянке и в подземном гараже.

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

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



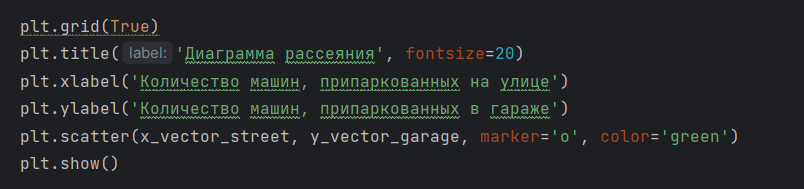
**Рисунок 1**

****

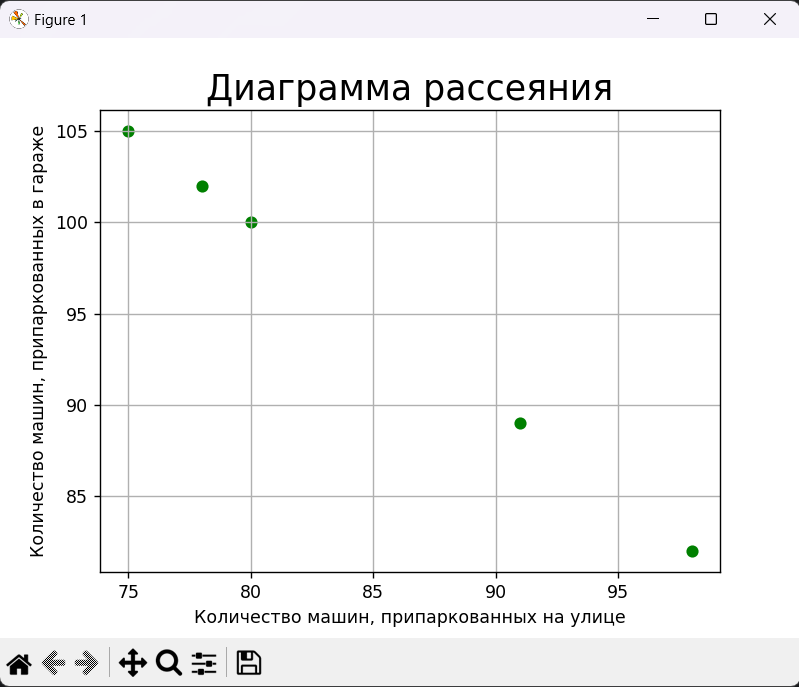
**Рисунок 2**

**Задание 1.2**

Построить диаграмму рассеяния.



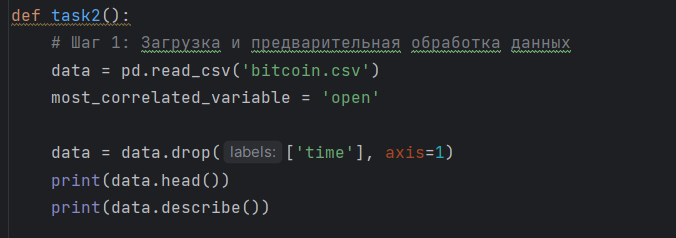
**Рисунок 3**

****

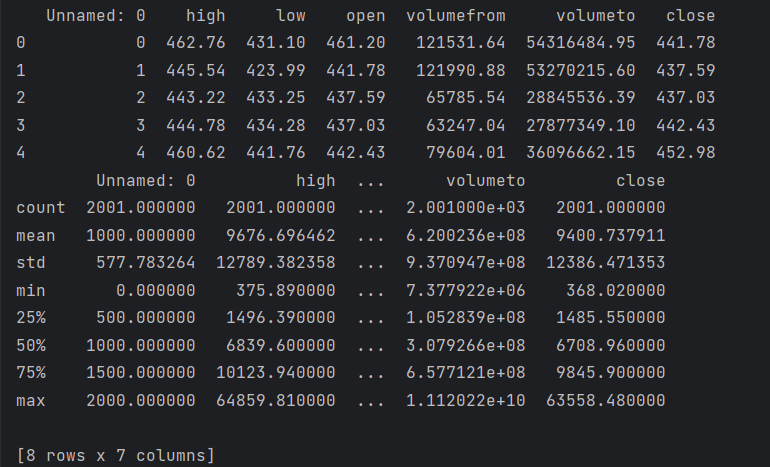
**Рисунок 4**

**Задание 2**

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

****

**Рисунок 5**

****

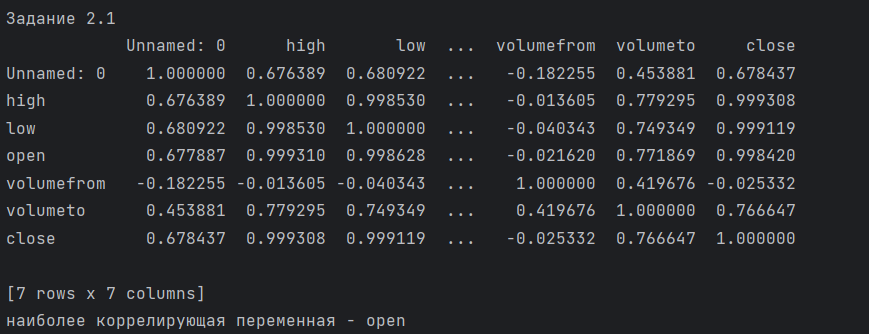
**Рисунок 6**

**Задание 2.1**

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

****

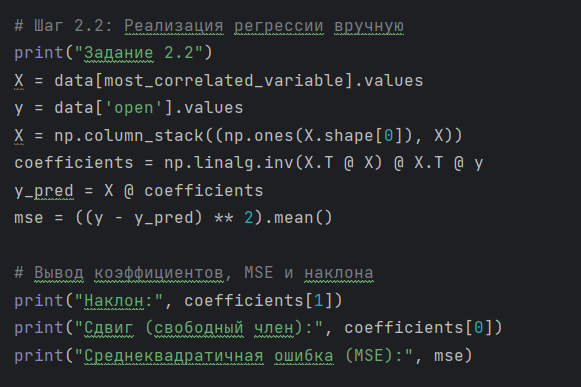
**Рисунок 7**

****

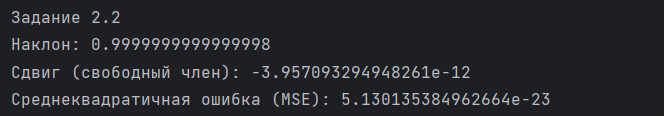
**Рисунок 8**

**Задание 2.2**

Реализовать регрессию вручную, отобразить наклон, сдвиг и MSE.

****

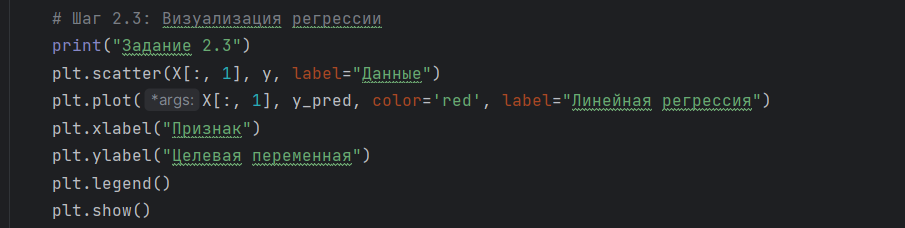
**Рисунок 9**

****

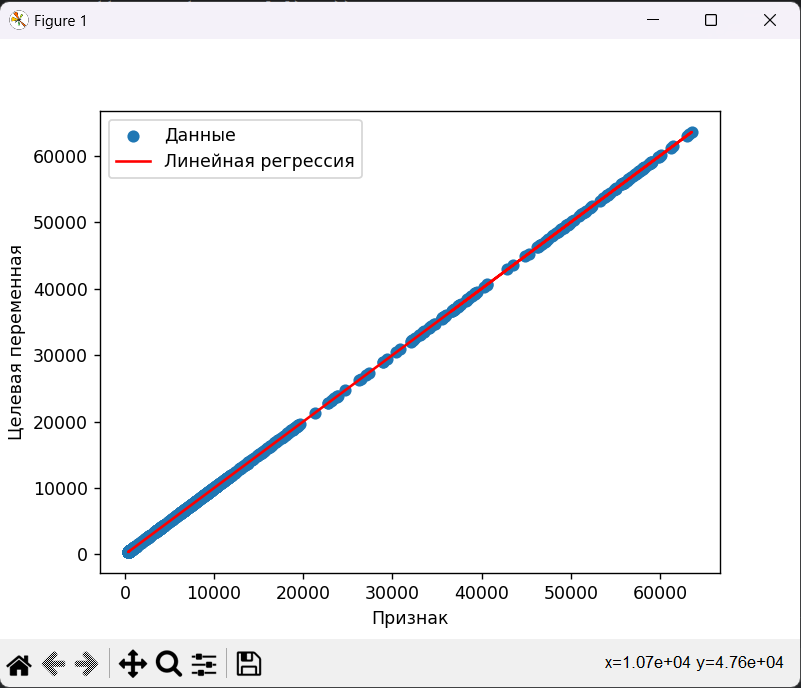
**Рисунок 10**

**Задание 2.3**

Визуализировать регрессию на графике.

****

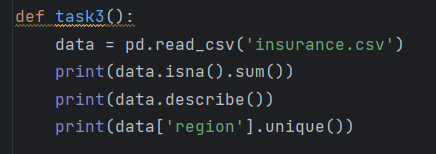
**Рисунок 11**

****

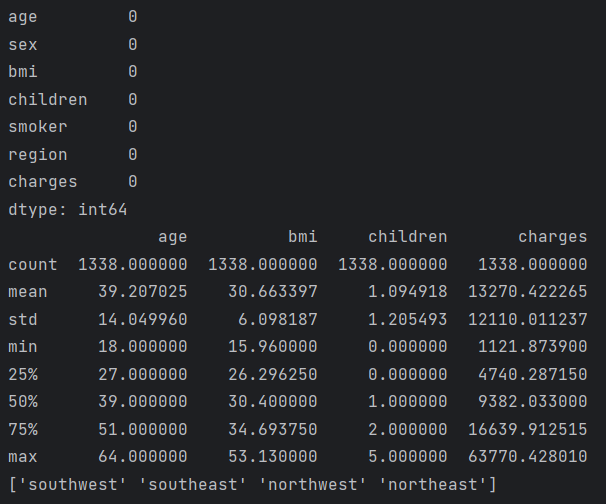
**Рисунок 12**

**Задание 3**

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

****

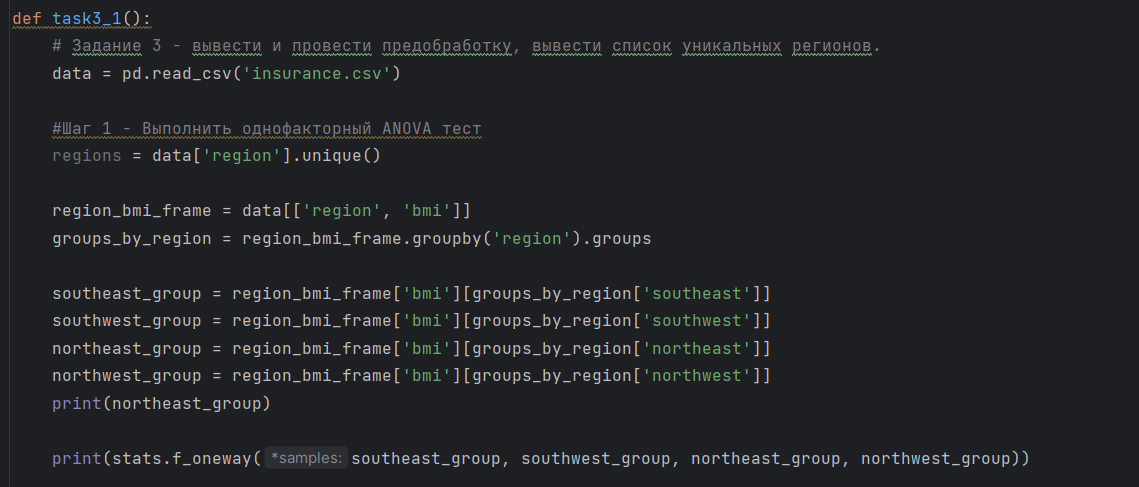
**Рисунок 13**

****

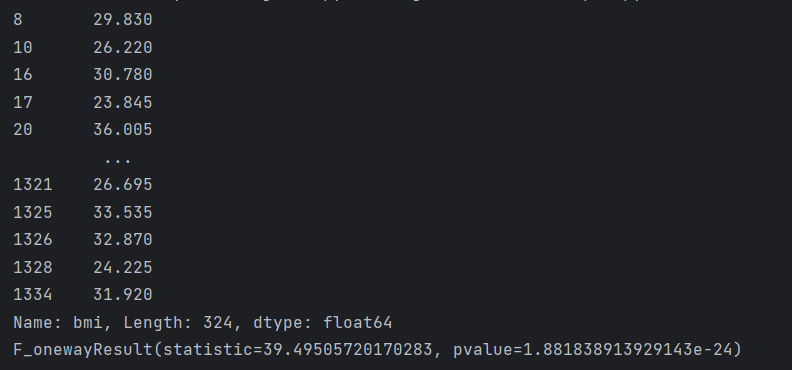
**Рисунок 14**

**Задание 3.1**

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

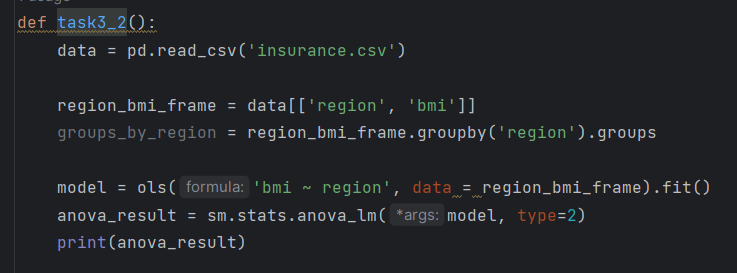
****

**Рисунок 15**

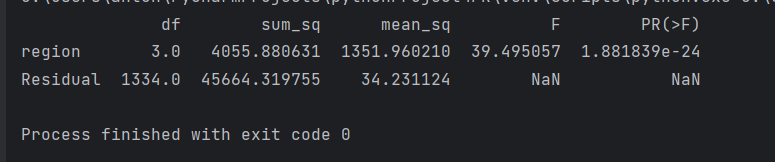
** Рисунок 16**

**Задание 3.2**

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

****

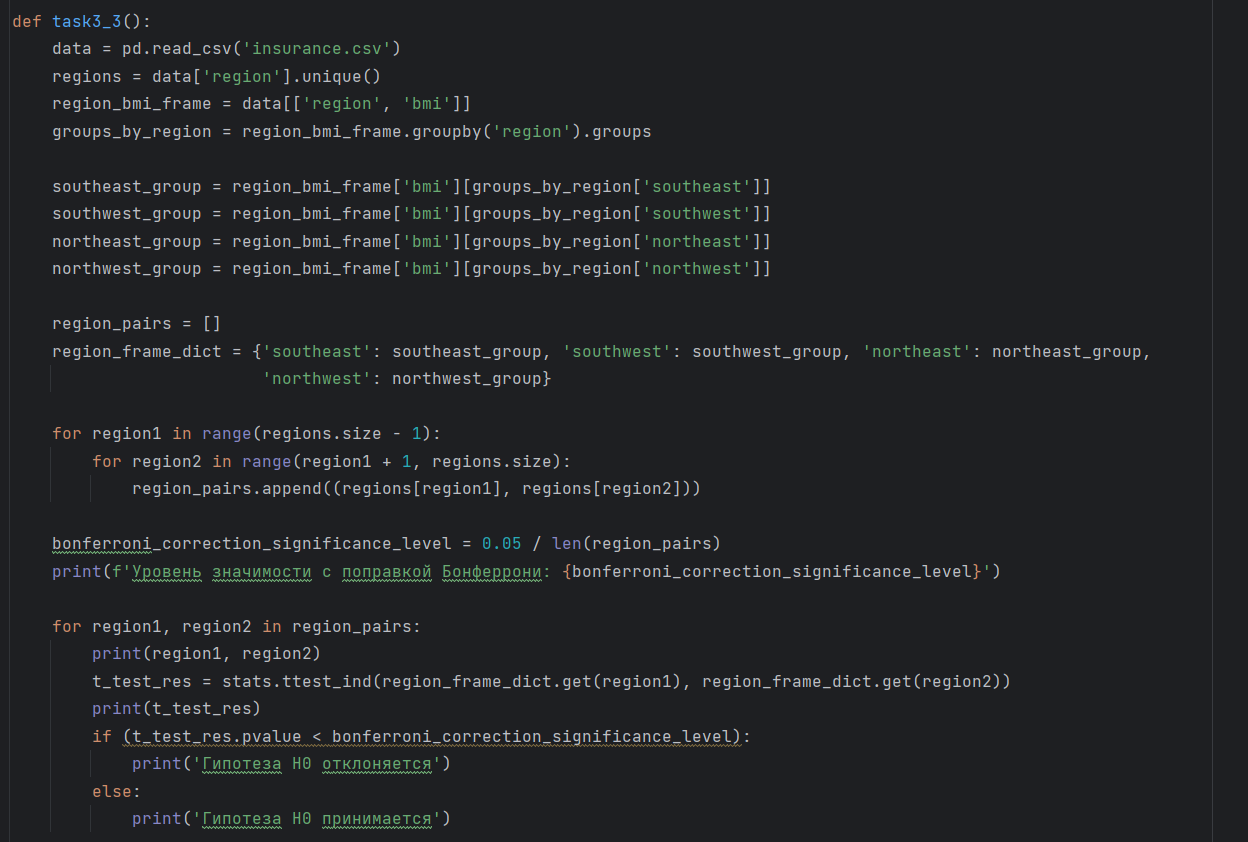
**Рисунок 17**

****

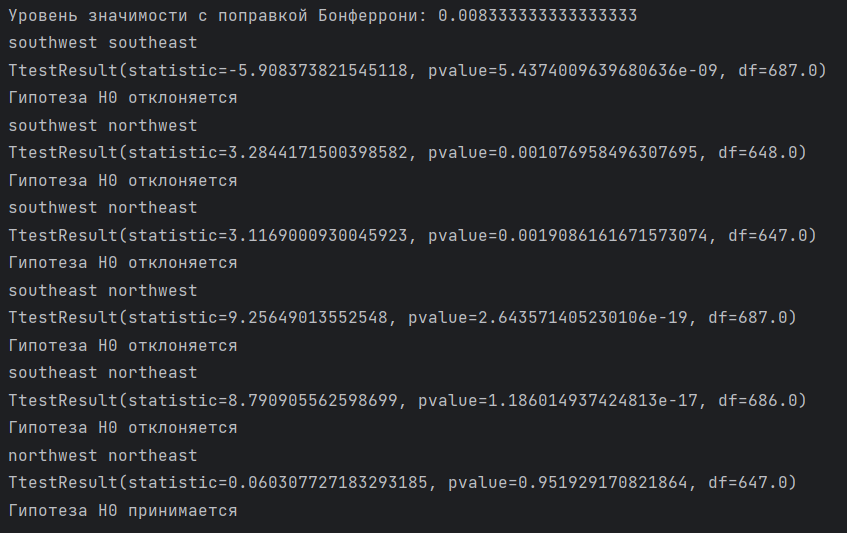
**Рисунок 18**

**Задание 3.3**

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

****

**Рисунок 19**

****

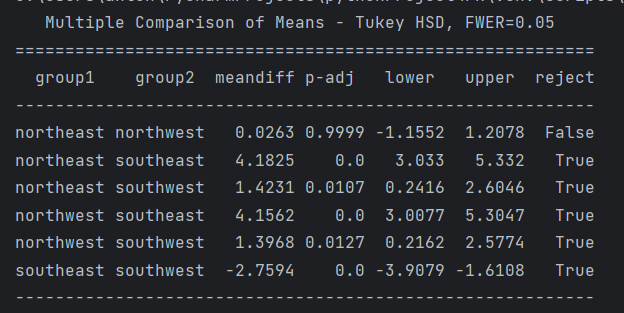
**Рисунок 20**

**Задание 3.4**

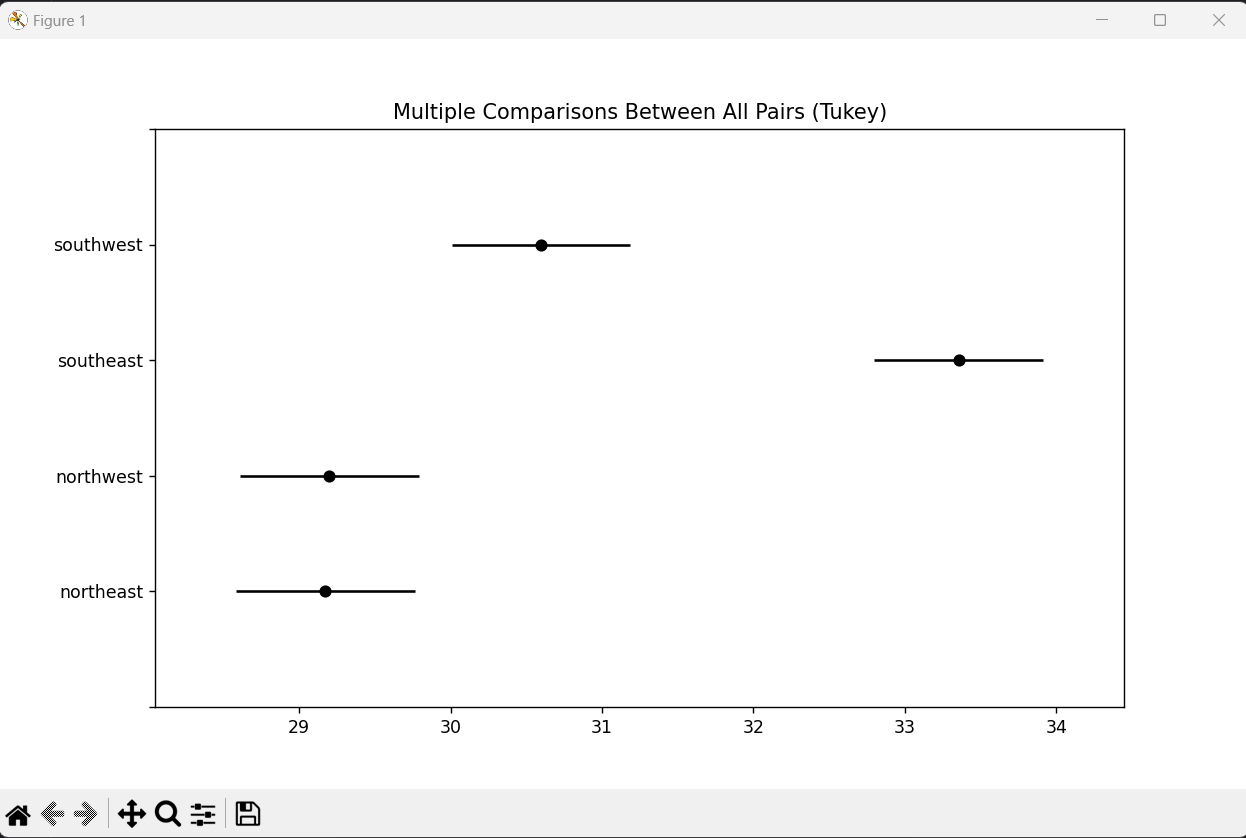
Выполнить пост-хок тесты Тьюки и построить график.

****

**Рисунок 21**

****

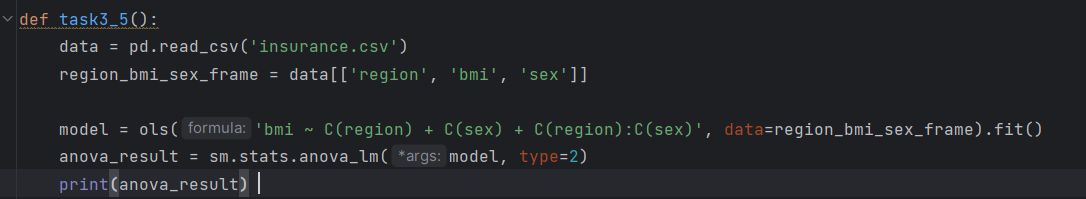
**Рисунок 22**

****

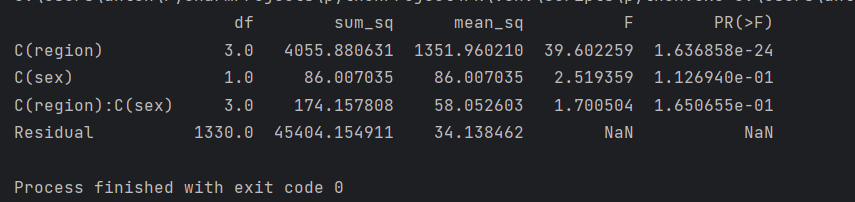
**Рисунок 23**

**Задание 3.5**

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

****

**Рисунок 24**

****

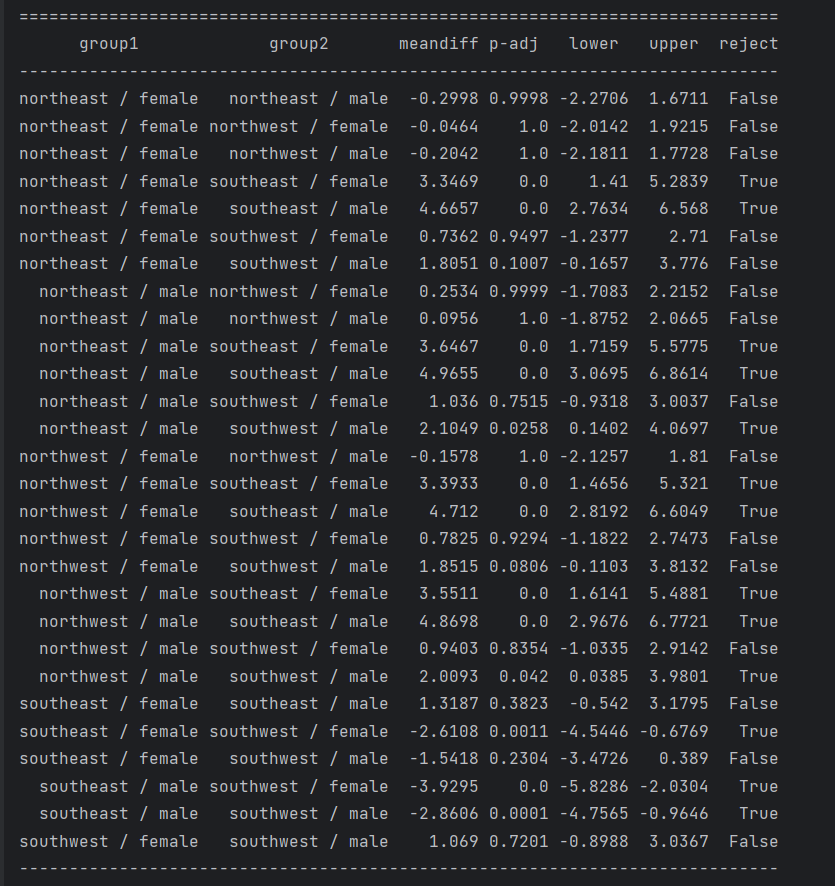
**Рисунок 25**

**Задание 3.6**

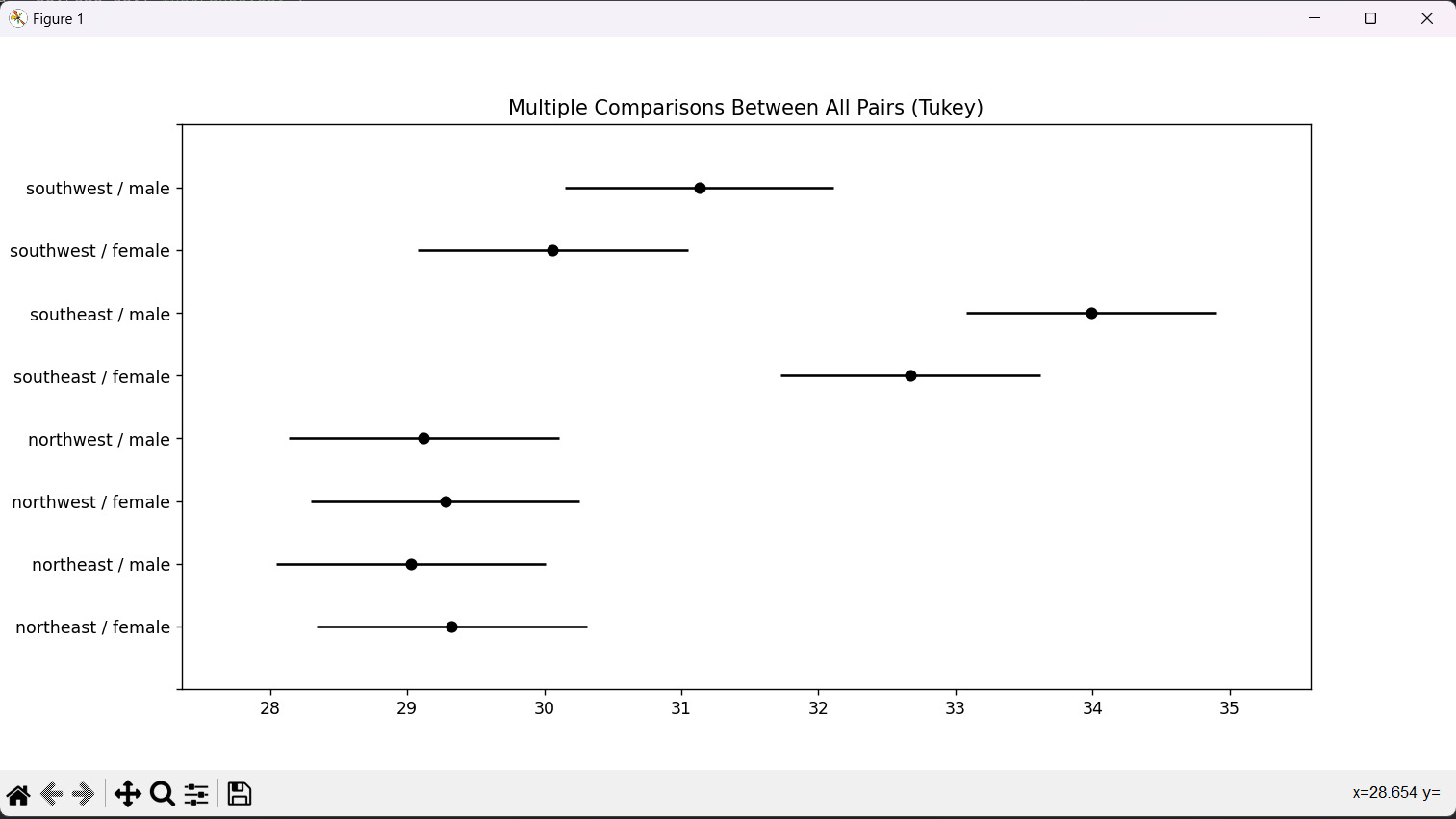
Выполнить пост-хок тесты Тьюки и построить график.

****

**Рисунок 26**

****

**Рисунок 27**

****

**Рисунок 28**

**Вывод**

В ходе выполнения данной практической работы мы научились находить и интерпретировать корреляцию, строить диаграммы рассеяния, реализовывать линейную регрессию как вручную, так и с помощью библиотеки sklearn.linear\_model, а также узнали, как найти угол наклона и y-перехват линейной регрессии.

В ходе выполнения данной практической работы мы выяснили, что такое дисперсионный анализ, какие виды его бывают и как реализовать это в Python, узнали, что такое пост-хок тесты, попрактиковались использовать пост-хок тест Тьюки.

**СПИСОК ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ**

1. Using Python on Windows [Электронный ресурс] – Режим доступа: <https://docs.python.org/3/using/windows.html>, свободный. Языка – английский. Дата обращения: 29.10.2023.
2. Jupyter Notebooks in VS Code [Электронный ресурс] – Режим доступа: <https://code.visualstudio.com/docs/datascience/jupyter-notebooks>, свободный. Языка – английский. Дата обращения: 29.10.2023.
3. Pandas documentation [Электронный ресурс] – Режим доступа: <https://pandas.pydata.org/docs/index.html>, свободный. Язык – английский. Дата обращения: 29.10.2023.
4. NumPy Documentation [Электронный ресурс] – Режим доступа: <https://numpy.org/doc/>, свободный. Язык – английский. Дата обращения: 30.10.2023.
5. Matplotlib 3.6.0 Documentation [Электронный ресурс] – Режим доступа: <https://matplotlib.org/stable/index.html>, свободный. Язык – английский. Дата обращения: 30.10.2023.
6. Statistical models Documentation [Электронный ресурс] – Режим доступа: <https://www.statsmodels.org/stable/index.html>, свободный. Язык – английский. Дата обращения: 30.10.2023.