|  |
| --- |
| https://lh6.googleusercontent.com/QcftzNtI05T0Y6fjdSh1Rr2rt8oqZ1IvnLvbn1jLJ7CCyteVir3k-xBLv4SL1wAgWJsRhmmJSR0UW-RP63_GQenE4vVWv05BRoZTsmIcBccVTnfxwmsnNMvjg599x9SqZd8E3dkd |
| МИНОБРНАУКИ РОССИИ |
| Федеральное государственное бюджетное образовательное учреждение высшего образования  **«МИРЭА - Российский технологический университет»**  **РТУ МИРЭА** |

Институт информационных технологий (ИТ)

Кафедра прикладной математики (ПМ)

|  |  |
| --- | --- |
| **ОТЧЕТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ №9** | |
| **по дисциплине** | |
| «Технологии и инструментарий анализа больших данных» | |
| Выполнил студент группы ИВБО-04-19 | Кузьников М.С. |
|  |  |
| Принял | Трифанова О.А. |

Практическая работа выполнена «\_» \_\_\_\_\_ 2022 г.

(подпись студента)

Зачтено «\_\_» 2022 г.

**СОДЕРЖАНИЕ**

1. [ПОСТАНОВКА ЗАДАЧИ 3](#_Toc116463437)

[2. ЗАДАНИЕ №1 4](#_Toc116463438)

[3. ЗАДАНИЕ №2 6](#_Toc116463439)

[4. ЗАДАНИЕ №3 7](#_Toc116463440)

[5. ЗАДАНИЕ №4 8](#_Toc116463441)

[6. ЗАДАНИЕ №5 11](#_Toc116463442)

[7. ЗАДАНИЕ №6 12](#_Toc116463443)

[8. ЗАДАНИЕ №7 13](#_Toc116463444)

[ЗАКЛЮЧЕНИЕ 16](#_Toc116463445)

[СПИСОК ЛИТЕРАТУРЫ 17](#_Toc116463446)

# ПОСТАНОВКА ЗАДАЧИ

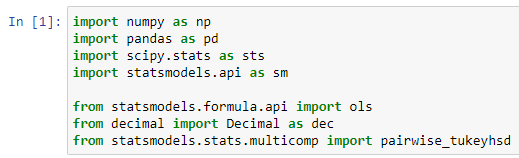
**Цель работы:** изучить основы дисперсионного анализа

**Задание:**

1. Загрузить данные: ***«insurance.csv».*** Вывести и провести предобработку. Вывести список уникальных регионов.
2. Выполнить однофакторный ***«ANOVA»*** тест, чтобы проверить влияние региона на индекс массы тела ***«BMI»***, используя первый способ, через библиотеку ***«Scipy».***
3. Выполнить однофакторный ***«ANOVA»*** тест, чтобы проверить влияние региона на индекс массы тела ***«BMI»***, используя второй способ, с помощью функции ***«anova\_lm()»*** из библиотеки ***«statsmodels».***
4. С помощью ***«t»*** критерия Стьюдента перебрать все пары. Определить поправку Бонферрони.
5. Выполнить пост-хок тесты Тьюки и построить график.
6. Выполнить двухфакторный ***«ANOVA»*** тест, чтобы проверить влияние региона и пола на индекс массы тела ***«BMI»***, используя функцию ***«anova\_lm()»*** из библиотеки ***«statsmodels».***
7. Выполнить пост-хок тесты Тьюки и построить график.

## ЗАДАНИЕ №1

Подключаю необходимые библиотеки для работы ***(Рисунок 1)***.

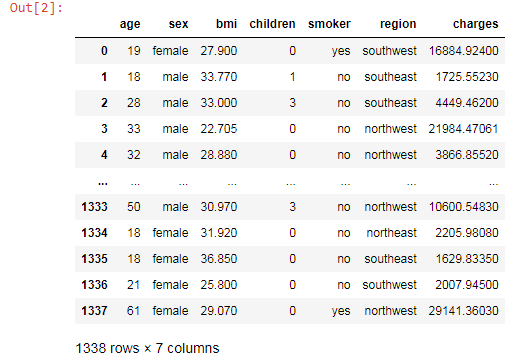


*Рисунок 1 – Листинг программы*

Загружаю данные из файла ***«insurance.csv» (Рисунок 2.1 – 2.2).***

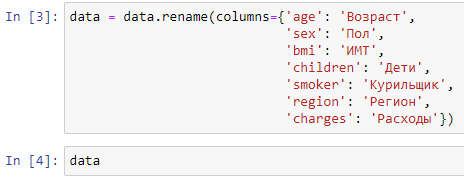
*D:\My_Documents\Учёба\2019\7_Семестр\_+_Технологии и Инструментарий Анализа Больших Данных\ПР\ПР_№8\Screens\Screenshot_2.png*

*Рисунок 2.1 – Листинг программы*

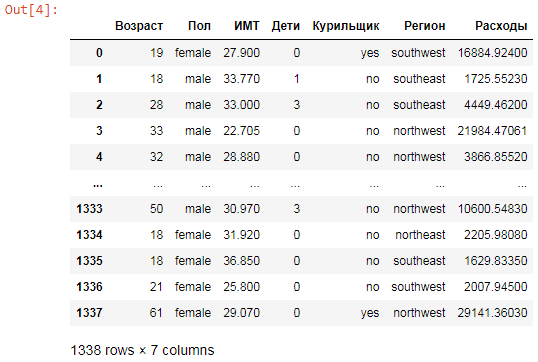
**

*Рисунок 2.2 – Вывод программы*

Переименовываю названия столбцов для читабельности ***(Рисунок 3.1 – 3.2).***

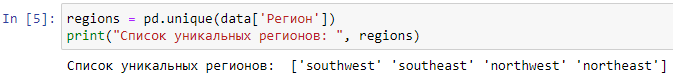


*Рисунок 3.1 – Листинг программы*



*Рисунок 3.2 – Вывод программы*

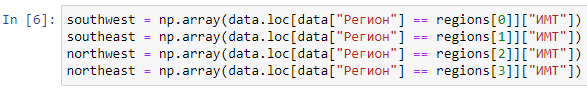
Вывожу список уникальных регионов ***(Рисунок 4).***

**

*Рисунок 4 – Листинг программы с выводом*

## ЗАДАНИЕ №2

Выполняю подготовку данных для ***«ANOVA»*** теста ***(Рисунок 5)***.

**

*Рисунок 5 – Листинг программы*

Провожу ***«ANOVA»*** тест ***(Рисунок 6)***.

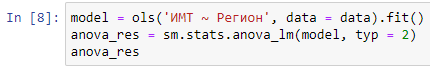
*D:\My_Documents\Учёба\2019\7_Семестр\_+_Технологии и Инструментарий Анализа Больших Данных\ПР\ПР_№8\Screens\Screenshot_8.png*

*Рисунок 6 – Листинг программы с выводом*

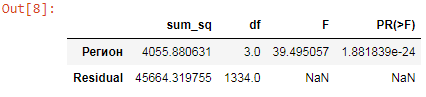
На основе полученных результатов делаю следующий вывод: ***«p-значение»*** ниже ***«0.05»***, значит ***«Регион»*** влияет на ***«ИМТ»***.

## ЗАДАНИЕ №3

Провожу ***«ANOVA»*** тест ***(Рисунок 7.1 – 7.2)***.



*Рисунок 7.1 – Листинг программы*

**

*Рисунок 7.2 – Вывод программы*

На основе полученных результатов делаю следующий вывод: ***«p-значение»*** ниже ***«0.05»***, значит ***«Регион»*** влияет на ***«ИМТ»***.

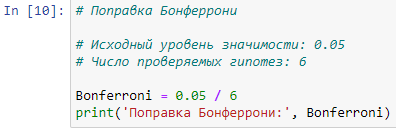
## ЗАДАНИЕ №4

Рассмотрю сравнение среднего значения ***«ИМТ»*** групп жителей по критерию ***«t»-Стьюдента*** для независимых выборок, где группы принадлежат к регионам ***«southwest»***, ***«southeast»***, ***«northwest»*** и ***«northeast»***.

Нулевая гипотеза – регион не влияет на ***«ИМТ»***, исследуемые выборки принадлежат к одной генеральной совокупности.

Альтернативная гипотеза – регион влияет на ***«ИМТ»***, выборки принадлежат к разным генеральным совокупностям.

Вычисляю поправку Бонферрони ***(Рисунок 8.1 – 8.2).***

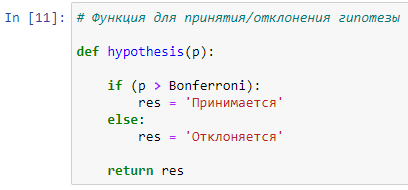
**

*Рисунок 8.1 – Листинг программы*

*D:\My_Documents\Учёба\2019\7_Семестр\_+_Технологии и Инструментарий Анализа Больших Данных\ПР\ПР_№8\Screens\Screenshot_12.png*

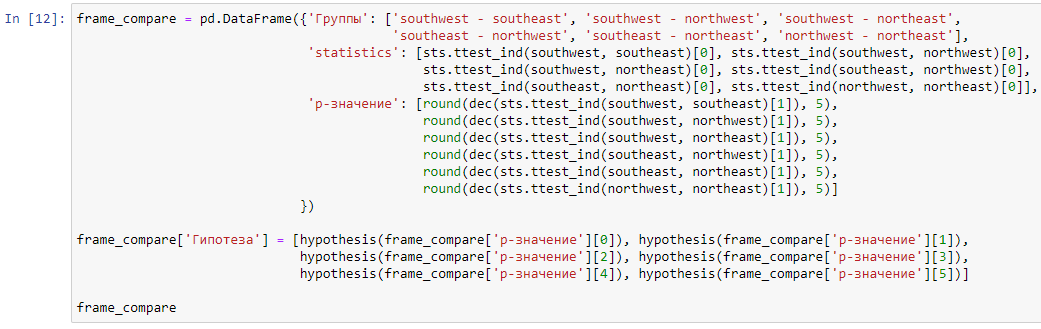
*Рисунок 8.2 – Вывод программы*

Создаю функцию для принятия или отклонения гипотезы на основе ***«p-значения» (Рисунок 9).***

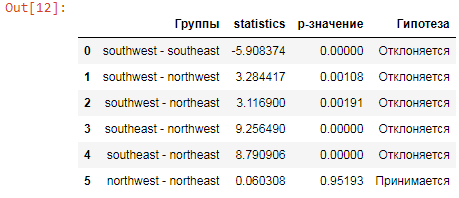
**

*Рисунок 9 – Листинг программы*

С помощью ***«t»*** критерия Стьюдента перебираю все пары и представляю полученные результаты в виде таблицы ***(Рисунок 10.1 – 10.2).***



*Рисунок 10.1 – Листинг программы*



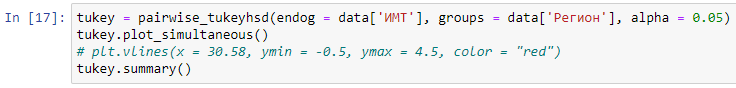
*Рисунок 10.2 – Вывод программы*

На основе полученных результатов делаю следующие выводы:

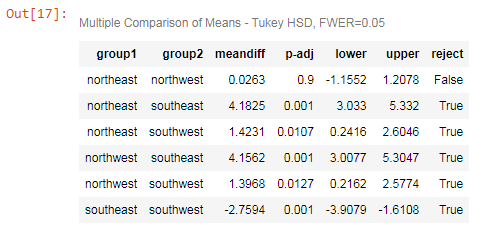
1. ***«p-значение»*** в группе ***«northwest – northeast»*** выше ***«0.00833»***, следовательно, нулевая гипотеза принимается. Значит, регион не влияет на ***«ИМТ».***
2. ***«p-значение»*** в остальных группах ниже ***«0.00833»***, следовательно, нулевая гипотеза отвергается. Значит, регионы в этих группах влияют на ***«ИМТ».***

## ЗАДАНИЕ №5

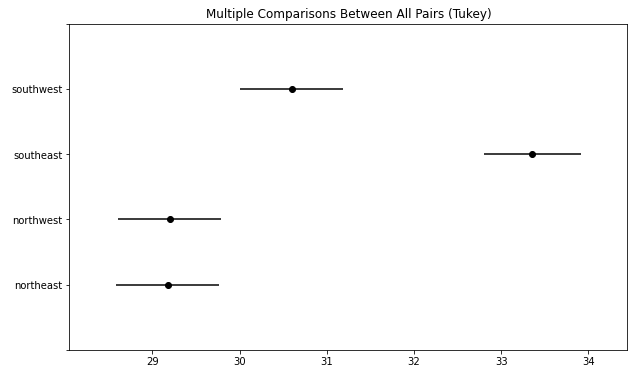
Выполняю пост-хок тест Тьюки и строю график ***(Рисунок 11.1 – 11.3).***

**

*Рисунок 11.1 – Листинг программы*



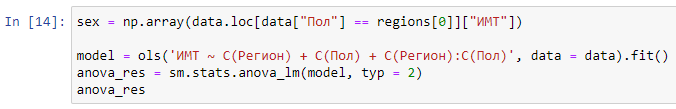
*Рисунок 11.2 – Вывод программы (сводная таблица)*



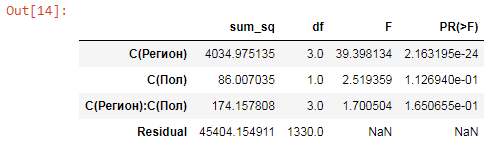
*Рисунок 11.3 – Вывод программы (график)*

## ЗАДАНИЕ №6

Провожу двухфакторный ***«ANOVA»*** тест ***(Рисунок 12.1 – 12.2) .***

**

*Рисунок 12.1 – Листинг программы*



*Рисунок 12.2 – Вывод программы*

На основе полученных результатов делаю следующие выводы:

1. Поскольку ***«p-значение»*** для ***«Региона»*** меньше ***«0.05»***, это означает, что этот фактор оказывает статистически значимое влияние на***«ИМТ».***
2. Поскольку ***«p-значение»*** для ***«Пола»*** выше ***«0.05»***, это говорит об отсутствии значительного влияния ***«Пола»*** на ***«ИМТ».***
3. Поскольку ***«p-значение»*** для ***«Региона/Пола»*** выше ***«0.05»***, это говорит об отсутствии значительного эффекта взаимодействия между воздействием ***«Пола»*** и ***«Региона»*** на ***«ИМТ».***

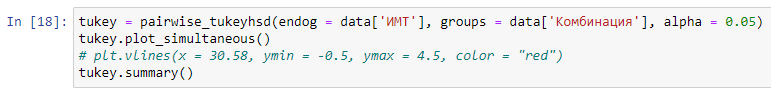
## ЗАДАНИЕ №7

Добавляю столбец в таблицу для исследования эффекта взаимодействия между воздействием ***«Пола»*** и ***«Региона»*** на ***«ИМТ» (Рисунок 13).***

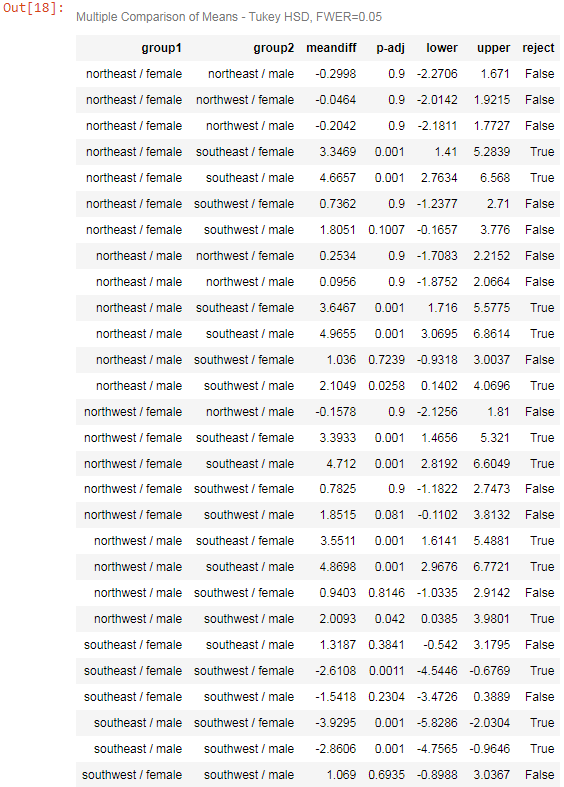
*D:\My_Documents\Учёба\2019\7_Семестр\_+_Технологии и Инструментарий Анализа Больших Данных\ПР\ПР_№8\Screens\Screenshot_21.png*

*Рисунок 13 – Листинг программы*

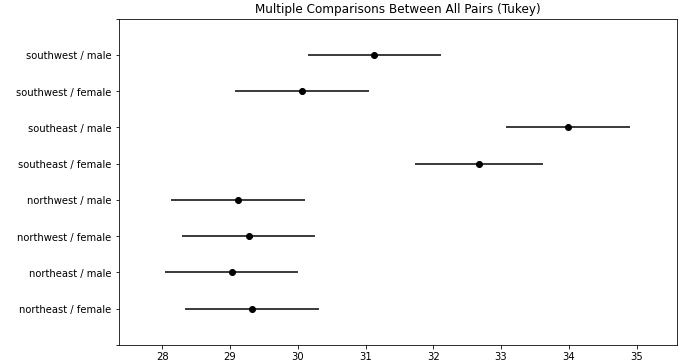
Провожу пост-хок тест Тьюки и строю график. ***(Рисунок 14.1 – 14.3).***

**

*Рисунок 14.1 – Листинг программы*



*Рисунок 14.2 – Вывод программы (сводная таблица)*

**

*Рисунок 14.3 – Вывод программы (график)*

# ЗАКЛЮЧЕНИЕ

В результате выполнения практической работы были проведены однофакторные и двухфакторные ***«ANOVA»*** тесты в рамках изучения дисперсионного анализа.

# СПИСОК ЛИТЕРАТУРЫ

1. [Документация по языку программирования ***«Python***](https://pydocs.ru/)***»*** [Электронный ресурс]: <https://pydocs.ru/>
2. Документация по ***«Jupyter Notebook»*** [Электронный ресурс]: <https://docs.jupyter.org/en/latest/>