|  |  |
| --- | --- |
|  | |
| МИНистерство Науки и высшего образования российской федерации | |
| Федеральное государственное бюджетное образовательное учреждениевысшего образования"МИРЭА - Российский технологический университет"РТУ МИРЭА | |
| Институт информационных технологий (ИТ) | |
| Кафедра прикладной математики (ПМ) | |

|  |  |
| --- | --- |
| **ОТЧЕТ**  **ПО ПРАКТИЧЕСКИМ РАБОТАМ** | |
| **по дисциплине** | |
| «Технологии и инструментарий анализа больших данных» | |
|  | |
| Выполнили студенты группы ИВБО-07-19 | Гридасов Е.А. |
| Принял ассистент кафедры ПМ | Горячев А.А. |

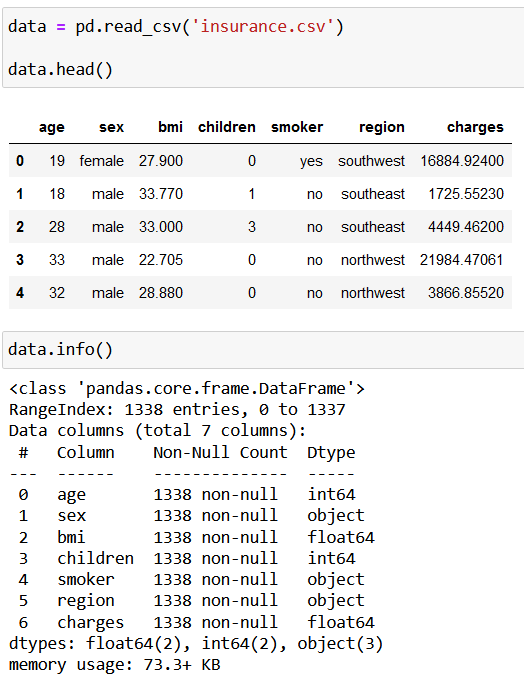
|  |  |  |
| --- | --- | --- |
| Практическая работа выполнена | «\_\_»\_\_\_\_\_\_\_\_\_2022 г. |  |
|  |  |  |
| «Зачтено» | «\_\_»\_\_\_\_\_\_\_\_\_2022 г. |  |

Москва 2022

# ВЫПОЛНЕНИЕ ЗАДАЧ

Целью работы было изучение дисперсионного анализа.

Для выполнения задания был предоставлен файл insurance.csv, который сначала надо был обработать. После чего надо было создать массив уникальных значений по регионам.



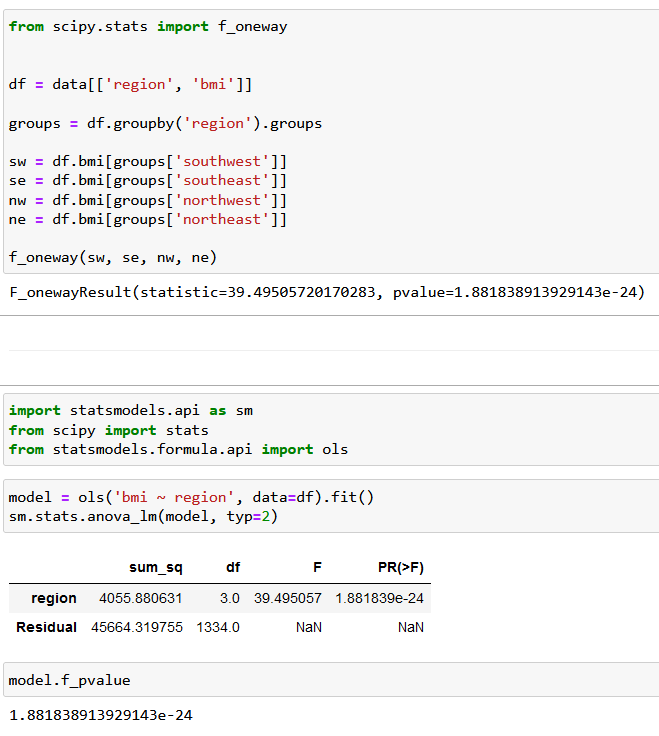
**Рисунок 1 – Загрузка датасета**



**Рисунок 2 — Проверка значений и создание массива уникальных значений по регионам**

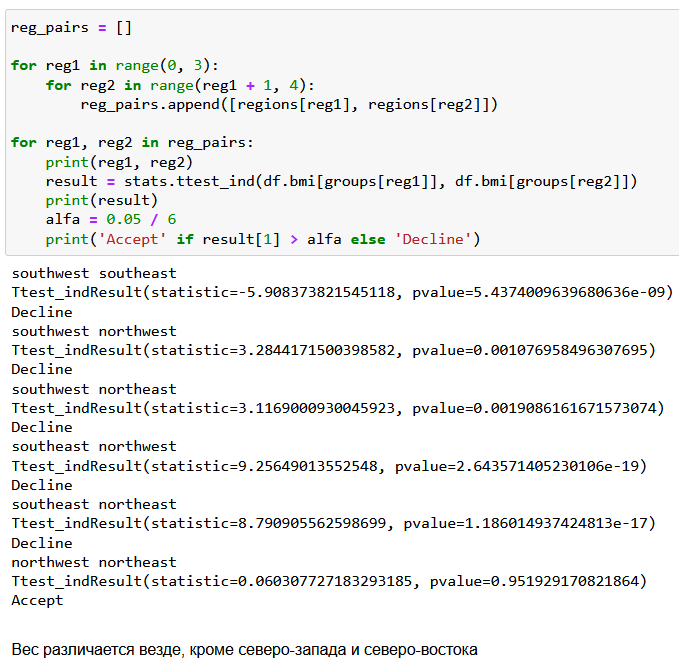
В данных не было никаких проблем, которые нужно было бы решить исправлением или замещением пустых значений.

После надо было провести однофакторный ANOVA тест двумя способами и по критерию Стьюдента определить взаимоотношение веса и регионов.

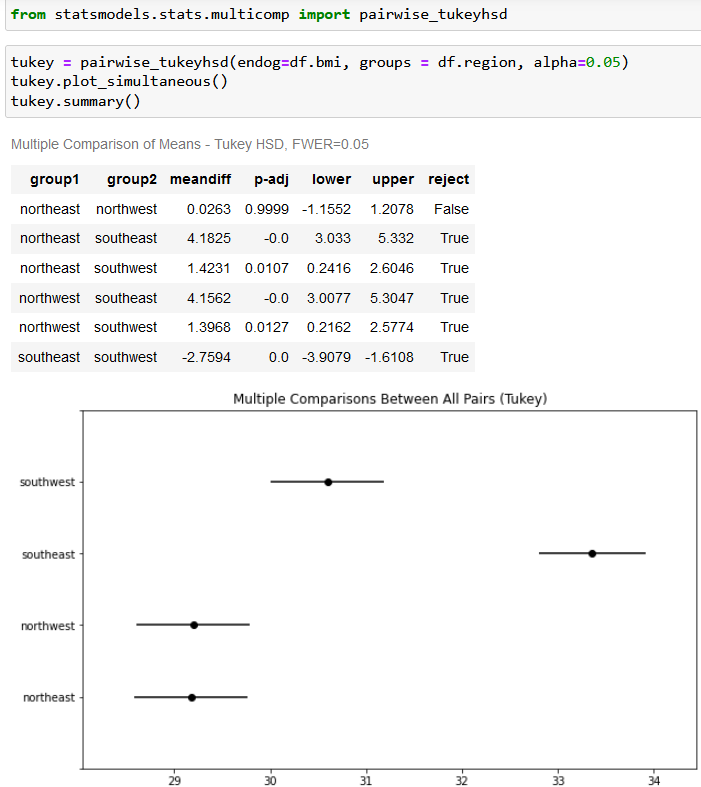


**Рисунок 3 — Проверка обоими способами ANOVA теста**

По значению p-value определяем, что где-то регион влияет на вес. Для понимания где обратимся к критерию Стьюдента и методу Бонферонни.



**Рисунок 4 — t-критерий Стьюдента по методу Бонферонни и выводы**



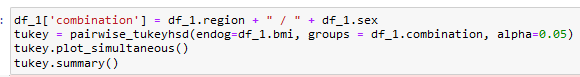
**Рисунок 5 — Проведение пост-хок тестов**

Окончательно приходим к выводу, что от регионов северо-запада и северо-востока есть влияние на вес жителей.

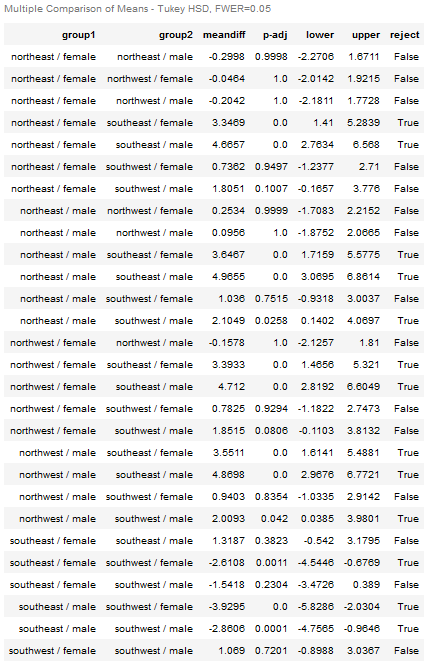
После надо было провести двухфакторный ANOVA тест и пост-хок тесты Тьюки с влиянием пола и региона на массу тела жителей.



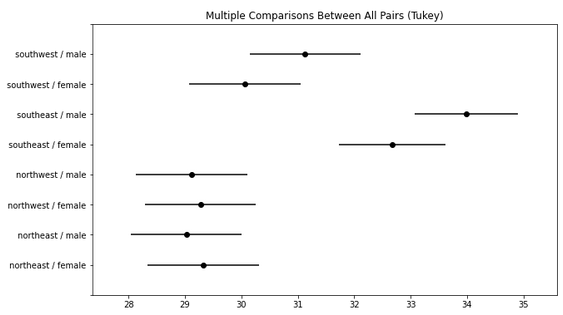
**Рисунок 6 — Генерация данных и проведение ANOVA теста**



**Рисунок 7 — Код для проведения пост-хок теста Кьюки**



**Рисунок 8 — Результаты пост-хок теста Кьюки**



**Рисунок 9 — График пост-хок теста Кьюки**

Получилось, что регион и пол играет роли для женщин в северо-западном регионе по сравнению с остальными.

# ВЫВОДЫ

В ходе выполнения работы были изучены библиотеки matplotlib, seaborn, sklearn, stats. Освоена тема дисперсионного анализа, изучены новые методы и модели.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Python Documentation [Электронный ресурс] - <https://www.python.org/>
2. Jupyter Documentation [Электронный ресурс] - <https://docs.jupyter.org/en/latest/index.html>
3. Pandas Documentation [Электронный ресурс] - <https://pandas.pydata.org/docs/getting_started/index.html>
4. Matplotlib Documentation [Электронный ресурс] - <https://matplotlib.org/>
5. Plotly Documentation [Электронный ресурс] - <https://plotly.com/python/>
6. Numpy Documentation [Электронный ресурс] - <https://numpy.org/doc/>
7. Stats Documentation [Электронный ресурс] - <https://docs.scipy.org/doc/scipy/reference/stats.html>