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

**Институт информационных технологий (ИИТ)**

**Кафедра практической и прикладной информатики (ППИ)**

**ОТЧЕТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ**

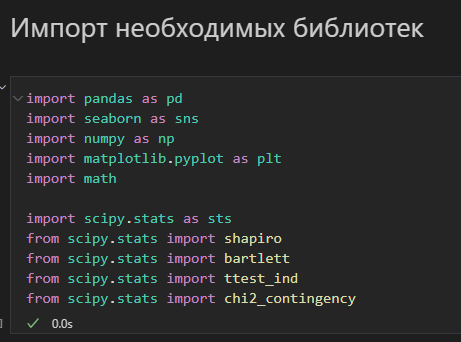
по дисциплине «Технологии и инструментарий анализа больших данных»

**Практическое занятие № 3**

|  |  |  |  |
| --- | --- | --- | --- |
| Студент группы ИНБО-01-17 | *ИНБО-03-20, Першутов Н.С.* | (подпись) | |
| Преподаватель | *Парамонов В.В* | (подпись) | |
| Отчет представлен | «\_\_\_»\_\_\_\_\_\_\_\_2023г. | |  | |

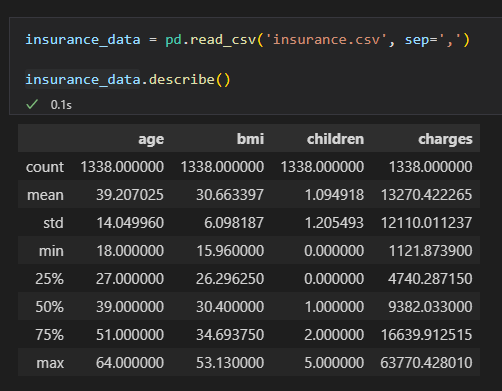
Москва 2023г.

**Результат работы:**

****

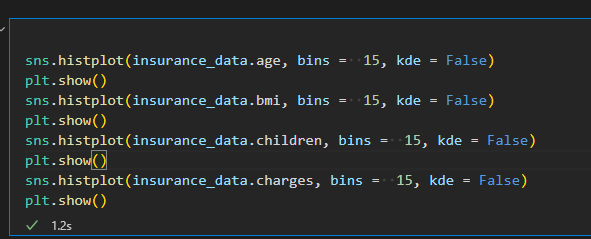
**Рисунок – Импорт необходимых библиотек**

**Задание 1. Загрузить данные из файла “insurance.csv”. С помощью метода describe() посмотреть статистику по данным.**

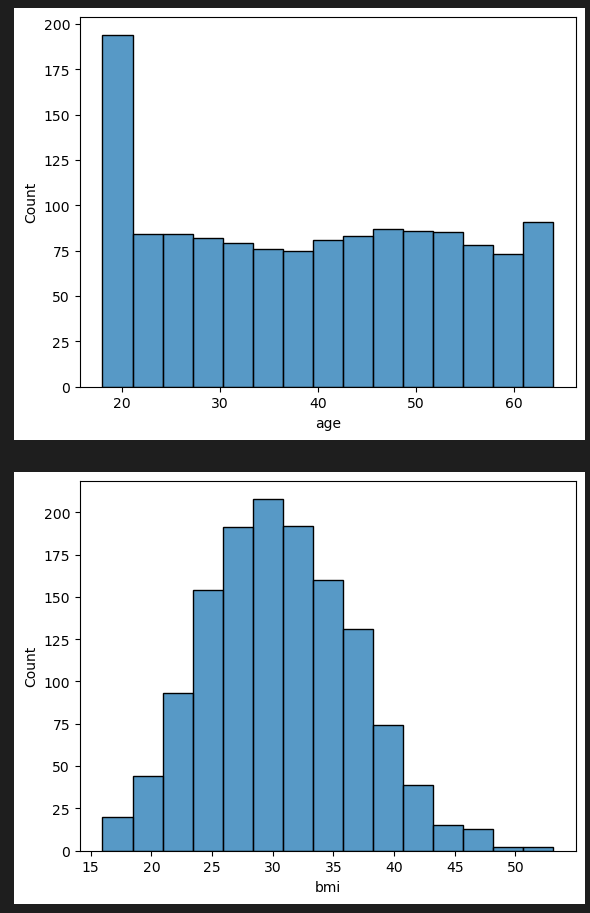
****

**Рисунок 1 – Импортирование данных**

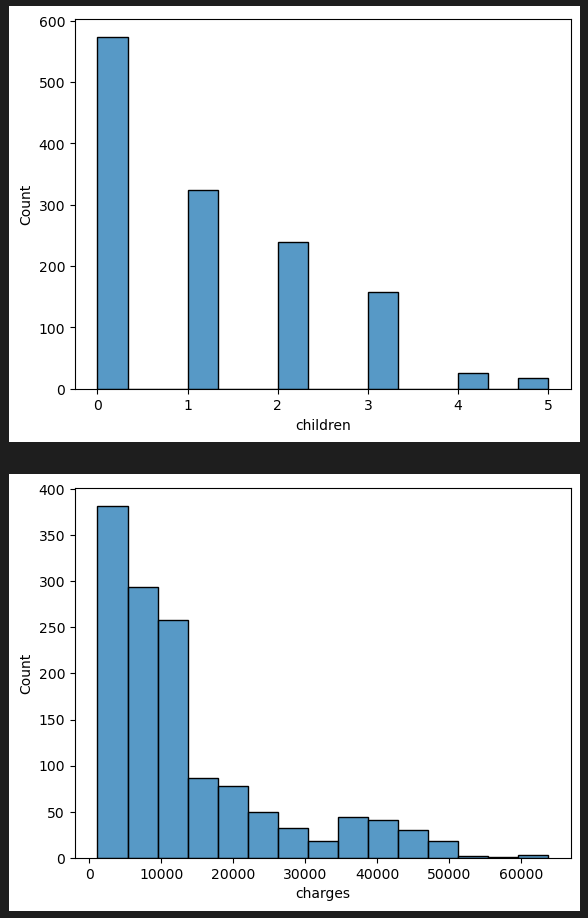
**Задание 3. Построить гистограммы для числовых показателей.**

****

**Рисунок 3.1 – Построение гистограмм**

****

**Рисунок 3.2 – Отрисованные гистограммы**

****

**Рисунок 3.2 – Отрисованные гистограммы**

**Задание 4. Найти меры центральной тенденции и меры разброса для индекса массы тела (bmi) и расходов (charges). Отобразить результаты в виде текста и на гистограммах (3 вертикальные линии). Добавить легенду на графики.**

Меры центральной тенденции:

• Мода – это значение, которое наиболее часто встречается в выборке.

• Медиана. Для нечетного количества элементов медиана равна центральному элементу в отсортированном массиве. Для четного количества элементов медиана равна среднему двух центральных элементов в отсортированном массиве.

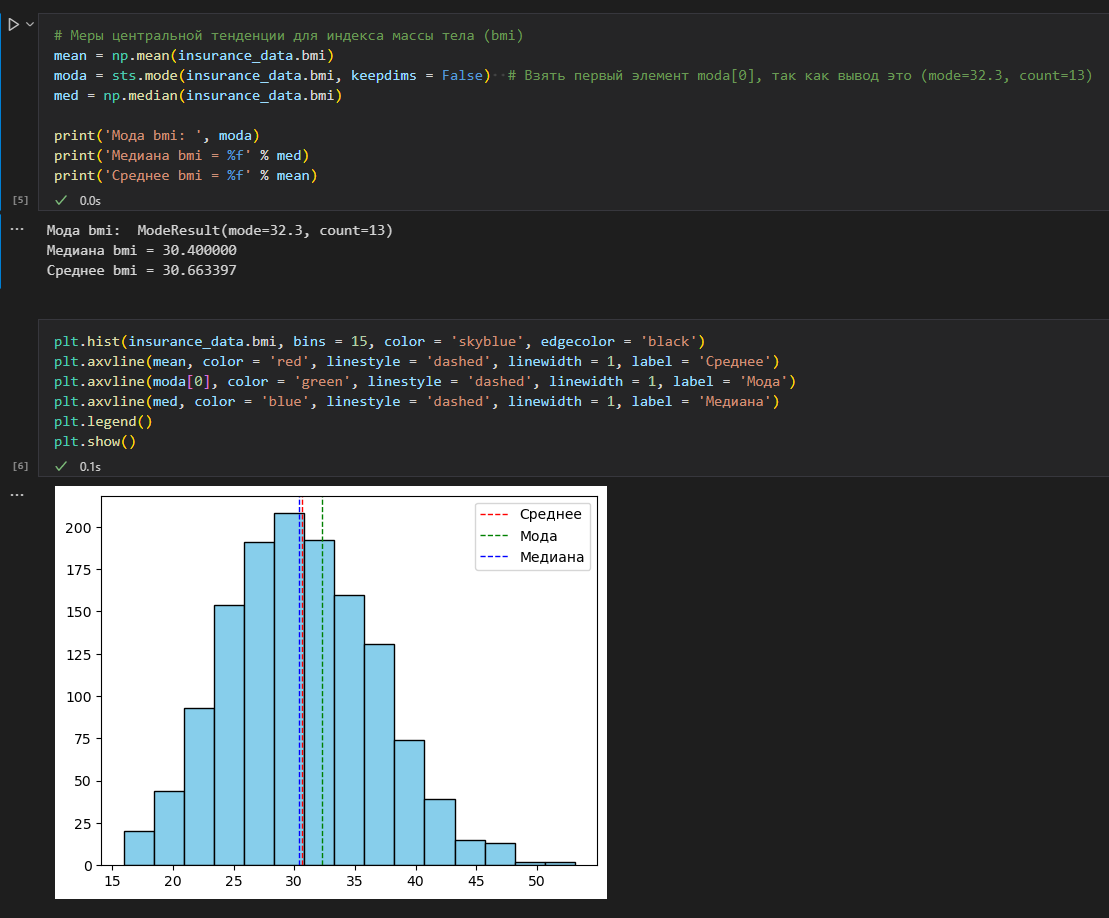
• Среднее – сумма значений всех элементов выборки, деленное на их количество.

Меры разброса:

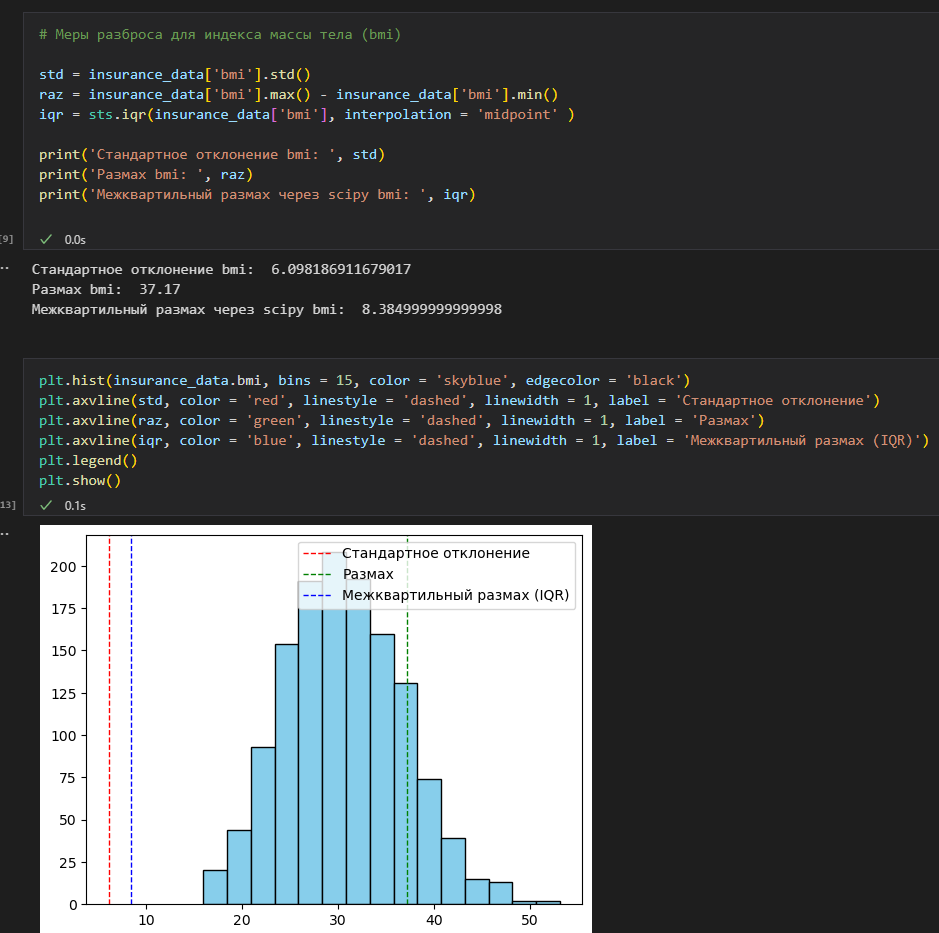
• Размах – разница между максимальным и минимальным значением выборки.

• Стандартное отклонение – Это корень из дисперсии, которая вычисляется по формуле (∑(𝑥𝑖−𝑥̅)^2)/(𝑛−1), где 𝑥̅– это среднее выборки. Для генеральной совокупности такой показатель называется среднеквадратическим отклонением.

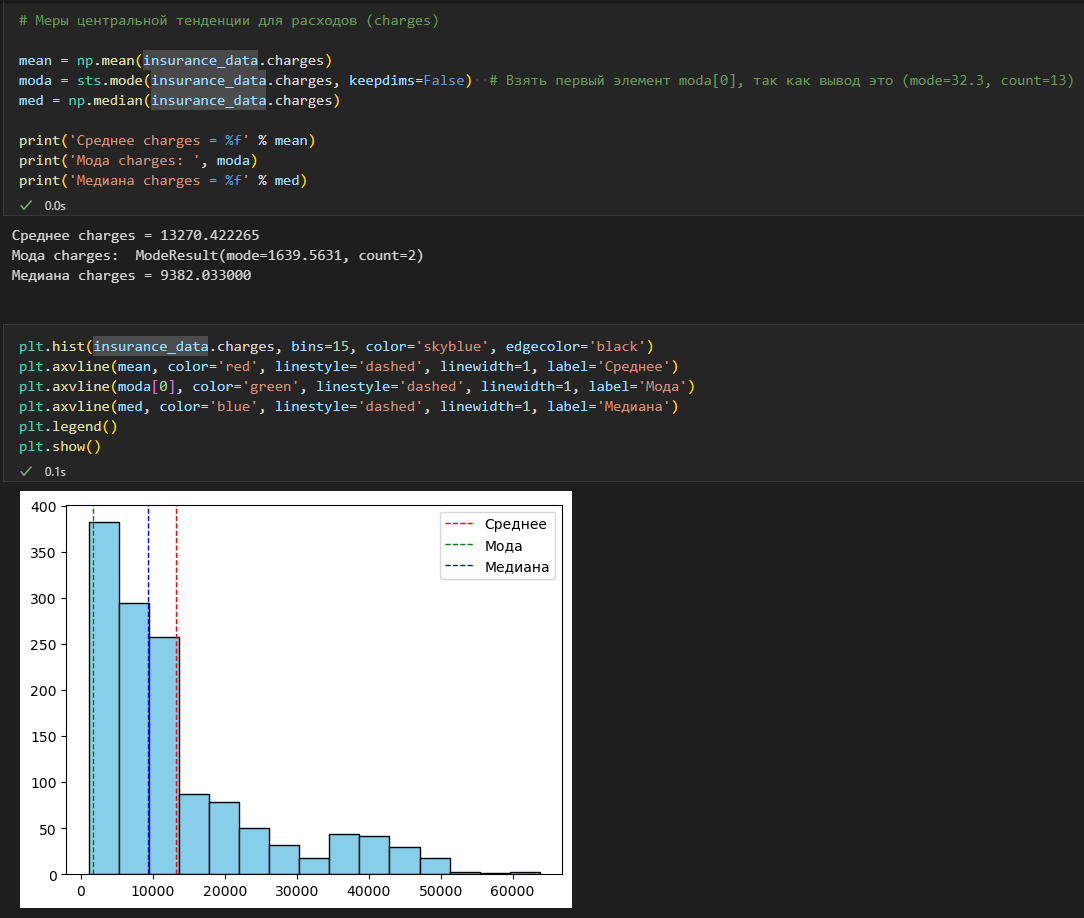
• Межквартильный размах (IQR). Для всех выборок существуют такие отсечки, которые называются «квартили», их всего три: Q1, Q2 и Q3. Межквартильный размах – разность между Q3 (75%) и Q1 (25%), это ширина интервала, который содержит 50% данных.

****

**Рисунок 4.1 – Меры центральной тенденции для индекса массы тела (bmi)**

****

**Рисунок 4.2 –** **Меры разброса для индекса массы тела (bmi)**

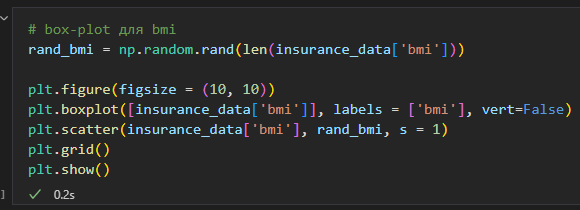
****

**Рисунок 4.3 – Меры центральной тенденции для расходов (charges)**

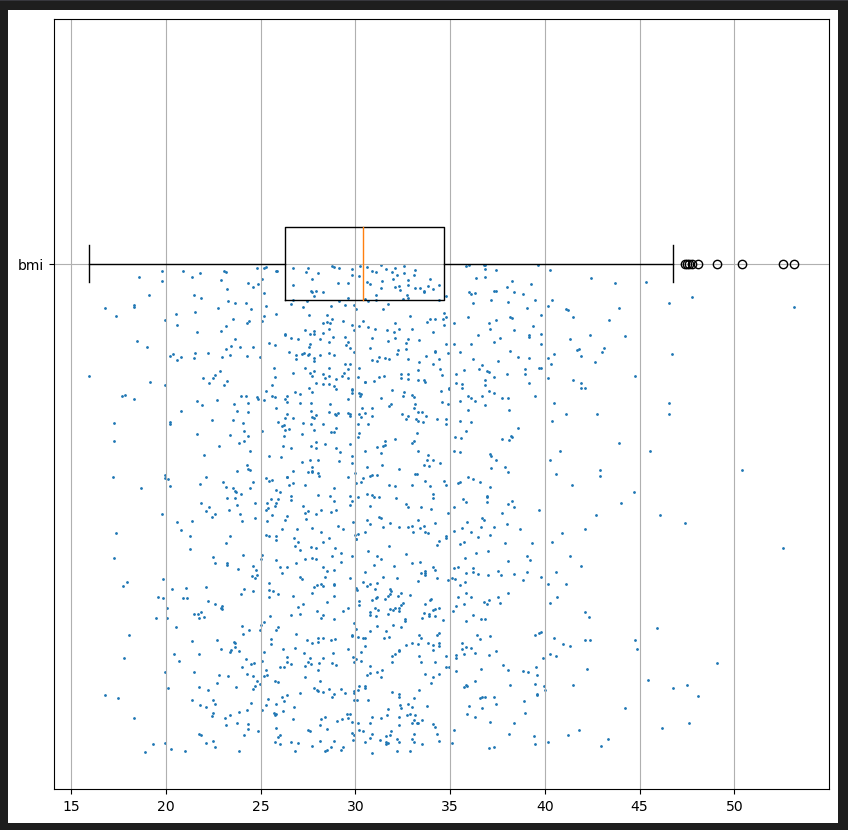
****

**Рисунок 4.4 – Меры разброса для расходов (charges)**

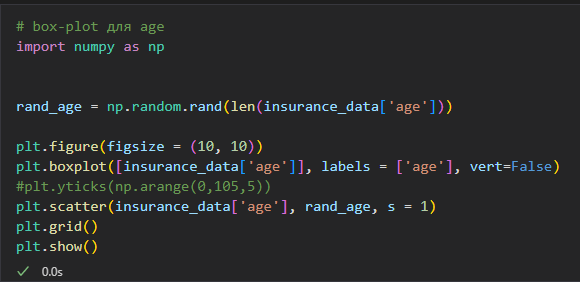
**Задание 5. Построить box-plot для числовых показателей. Названия графиков должны соответствовать названиям признаков.**

****

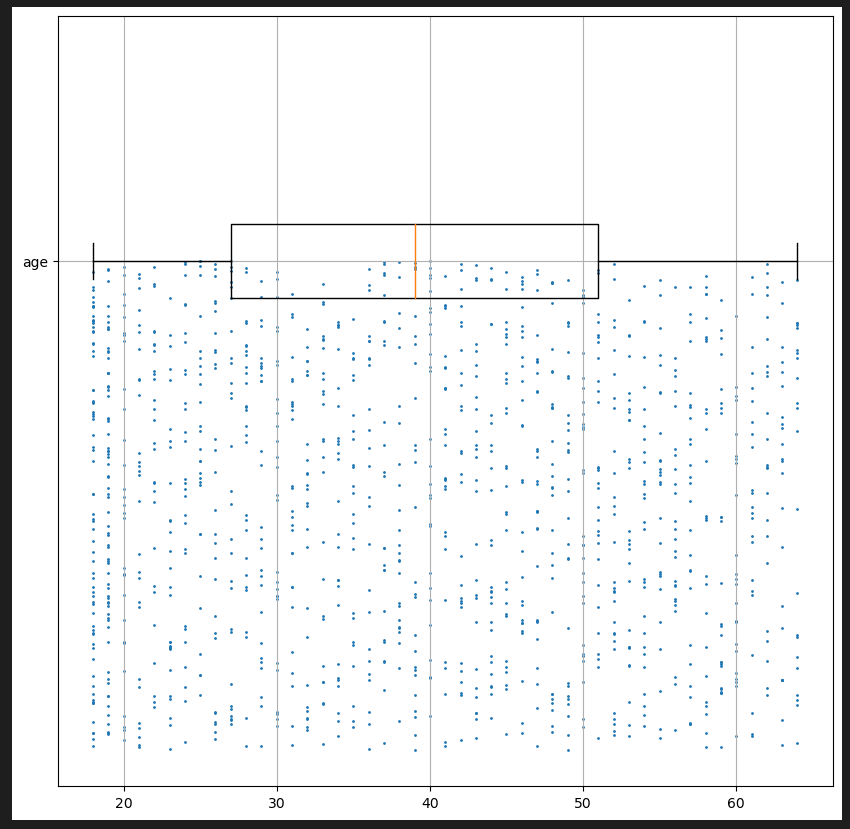
**Рисунок 5.1 – box-plot для bmi**

****

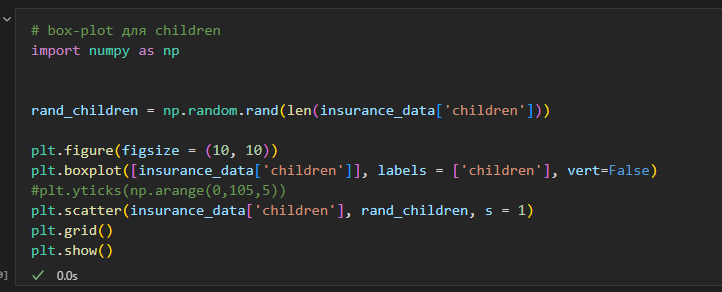
**Рисунок 5.2 – box-plot для bmi**

****

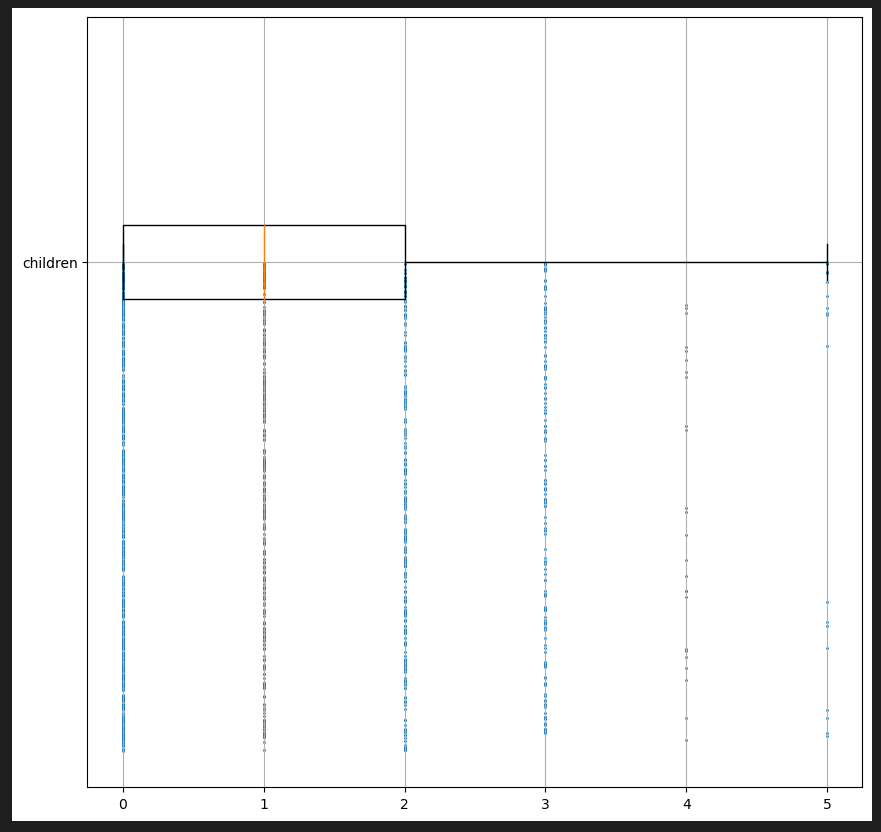
**Рисунок 5.3 – box-plot для age**

****

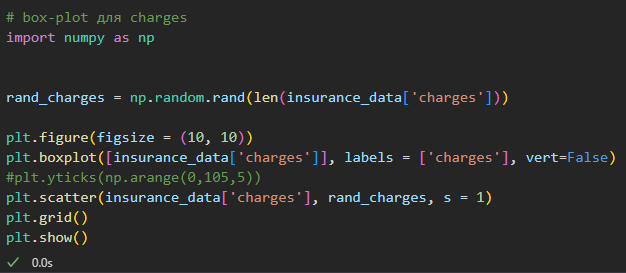
**Рисунок 5.4 – box-plot для age**

****

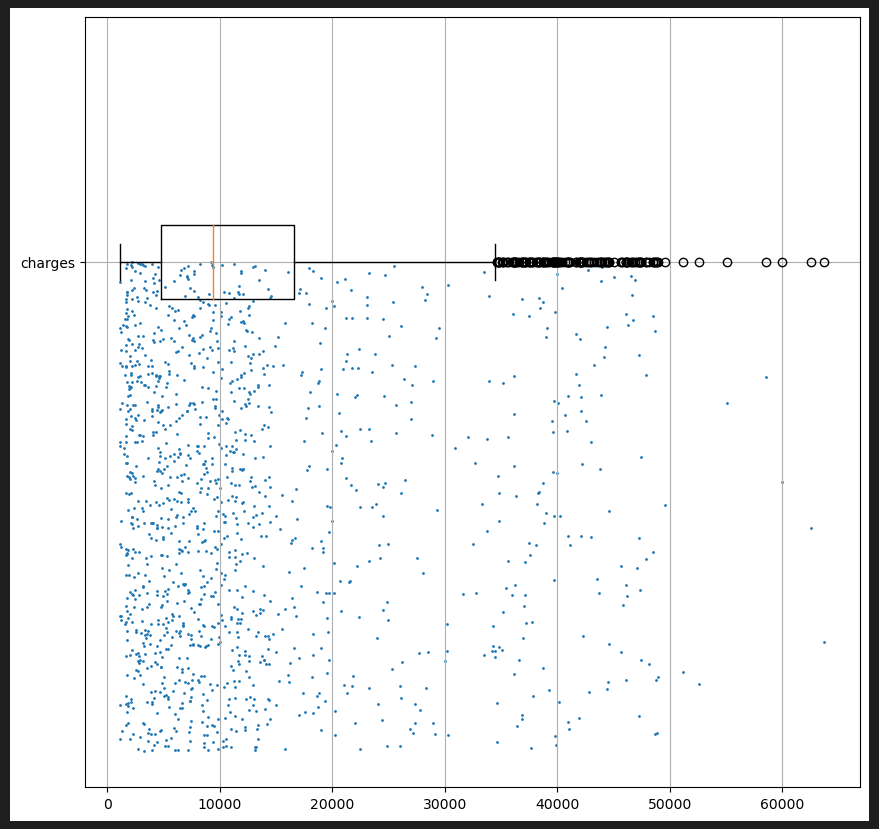
**Рисунок 5.5 – box-plot для children**

****

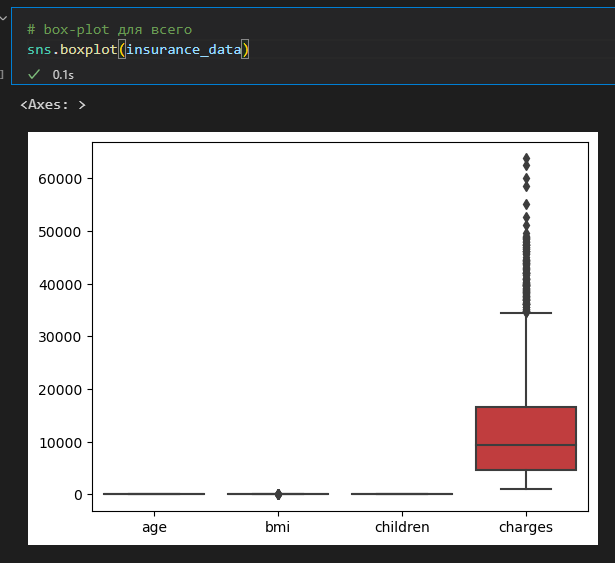
**Рисунок 5.6 – box-plot для children**

****

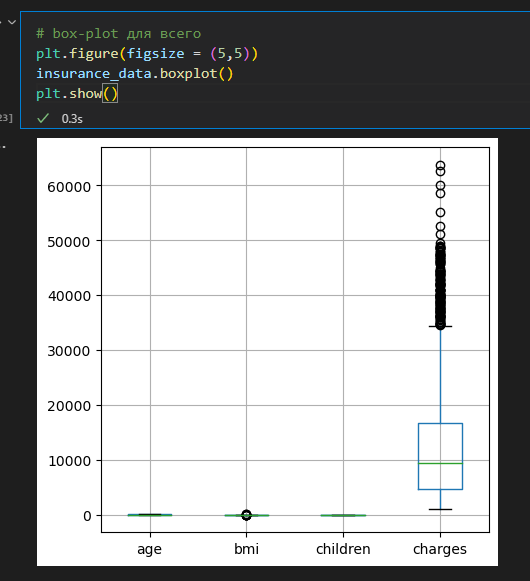
**Рисунок 5.7 – box-plot для charges**

****

**Рисунок 5.8 – box-plot для charges**

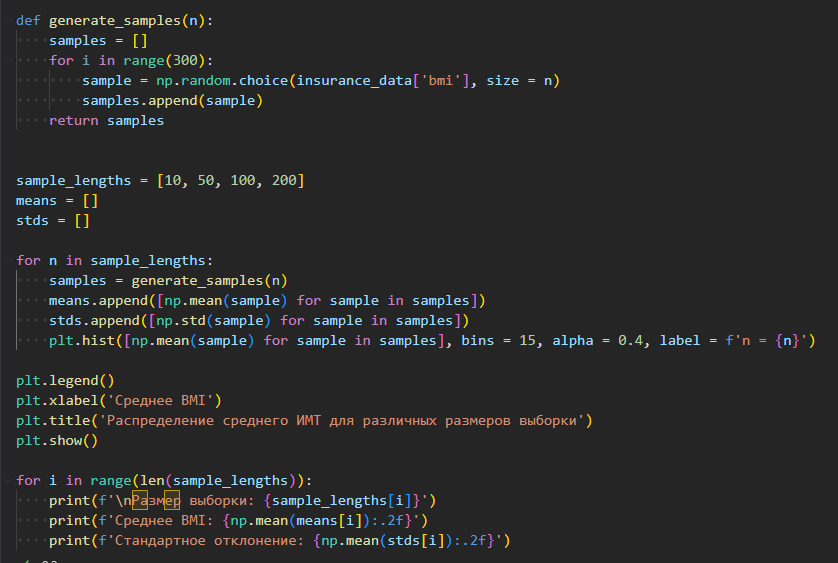
****

**Рисунок 5.9 – общий box-plot используя библиотеку seaborn**

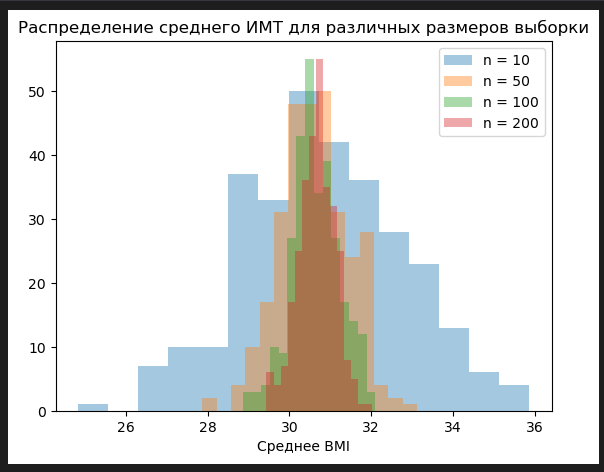
****

**Рисунок 5.10 – общий box-plot используя библиотеку matplotlib**

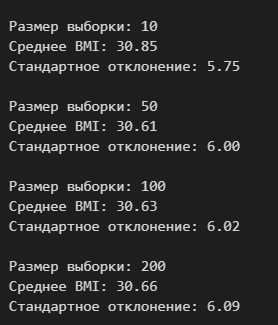
**Задание 6. Используя признак charges или imb, проверить, выполняется ли центральная предельная теорема. Использовать различные длины выборок n. Количество выборок = 300. Вывести результат в виде гистограмм. Найти стандартное отклонение и среднее для полученных распределений.**

****

**Рисунок 6.1 – Проверка выполнения центральной предельной теоремы.**

****

**Рисунок 6.2 – Отрисованный график**

****

**Рисунок 6.3 – Результаты**

**Задание 7. Построить 95% и 99% доверительный интервал для среднего значения расходов и среднего значения индекса массы тела.**

Для расходов "charges":

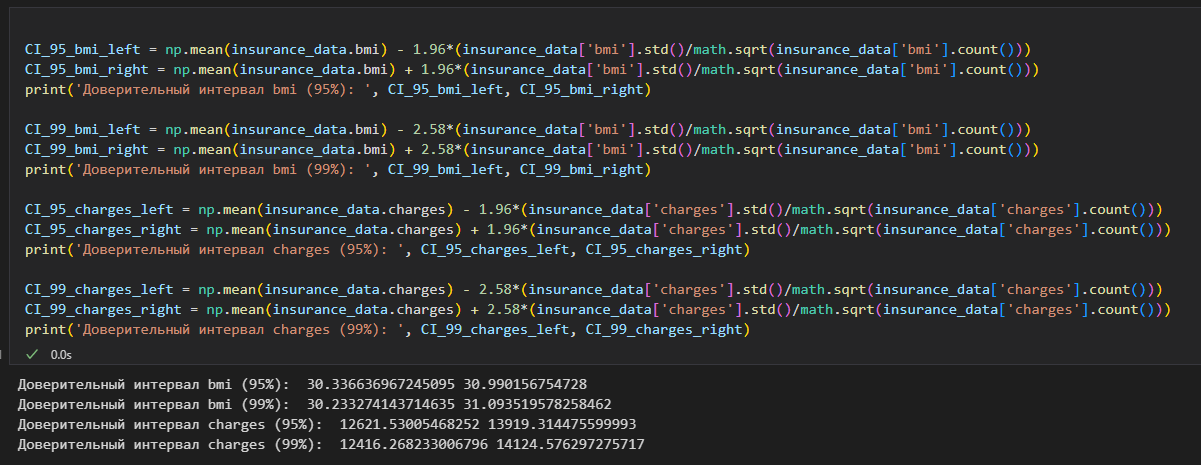
Доверительный интервал (95%): среднее значение charges +- 1.96\*(стандартное отклонение charges / корень из количества наблюдений)

Доверительный интервал (99%): среднее значение charges +- 2.58\*(стандартное отклонение charges / корень из количества наблюдений)

Для индекса массы тела "bmi":

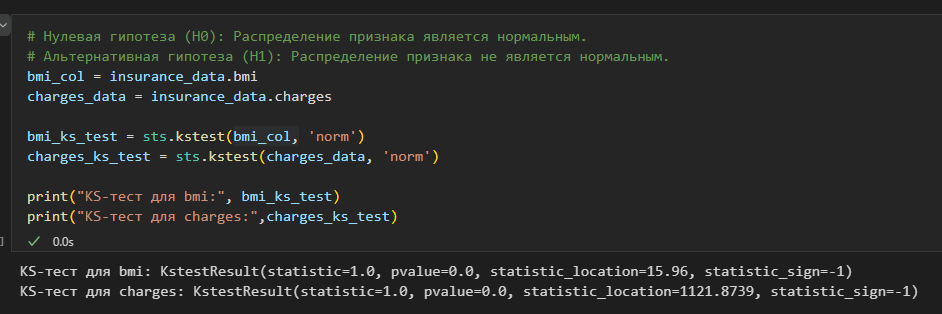
Доверительный интервал (95%): среднее значение bmi +- 1.96\*(стандартное отклонение bmi / корень из количества наблюдений)

Доверительный интервал (99%): среднее значение bmi +- 2.58\*(стандартное отклонение bmi / корень из количества наблюдений)

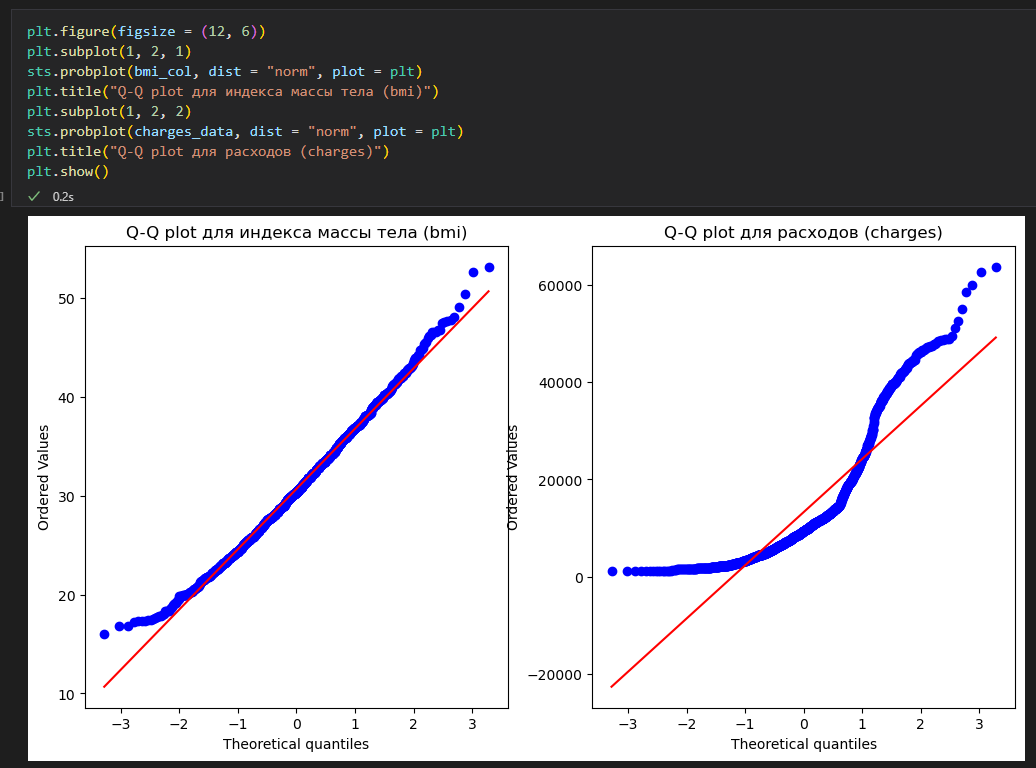
****

**Рисунок 7.1 – Построение доверительного интервала**

**Задание 8. Проверить распределения следующих признаков на нормальность: индекс массы тела, расходы. Сформулировать нулевую и альтернативную гипотезы. Для каждого признака использовать KS-тест и q-q plot.**

****

**Рисунок 8.1 – Проверка распределения признаков на нормальность**

****

**Рисунок 8.2 – Отрисовка графиков**

Правый график показывает, что распределение исследуемой выборки сильно отличается от нормального.

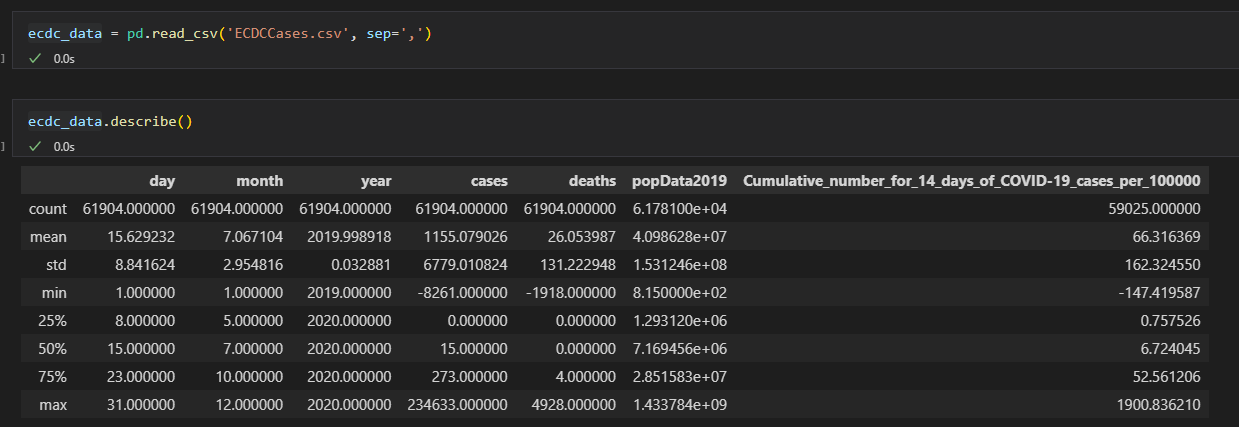
На левом графике середина распределения следует нормальному закону, но его концы отклоняются от него.

Если полученные p-значения меньше выбранного уровня значимости (обычно 0.05), то нулевая гипотеза отвергается,

и можно сделать вывод, что распределение признака не является нормальным. В противном случае, нулевая гипотеза

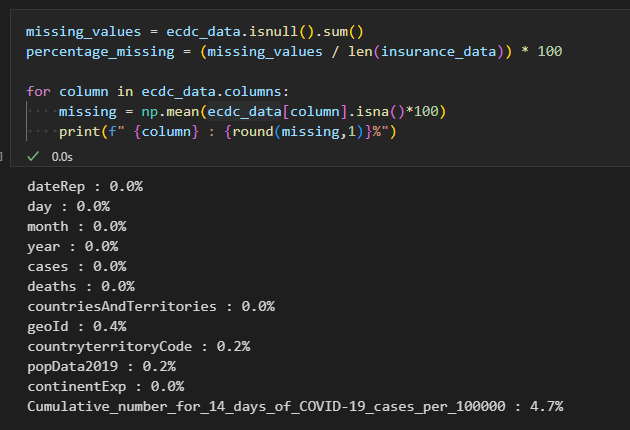
не может быть отвергнута, и можно считать распределение признака нормальным.

**Задание 9. Загрузить данные из файла “ECDCCases.csv”**

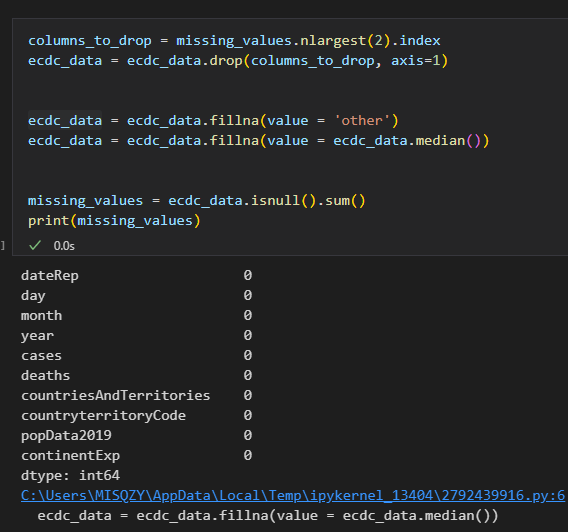
****

**Рисунок 9.1 – Загрузка данных**

**Задание 10. Проверить в данных наличие пропущенных значений. Вывести количество пропущенных значений в процентах. Удалить два признака, в которых больше всех пропущенных значений. Для оставшихся признаков обработать пропуски: для категориального признака использовать заполнение значением по умолчанию (например, «other»), для числового признака использовать заполнение медианным значением. Показать, что пропусков больше в данных нет.**

****

**Рисунок 10.1 – Проверка пропущенных значений в процентах**

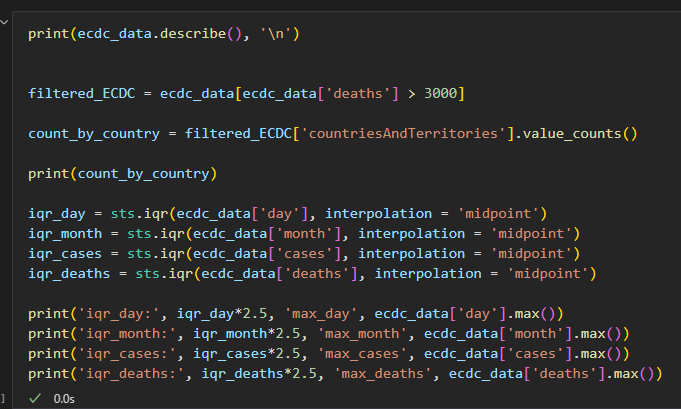
****

**Рисунок 10.2 – Удаление двух признаков, в которых больше всего пропущенных значений**

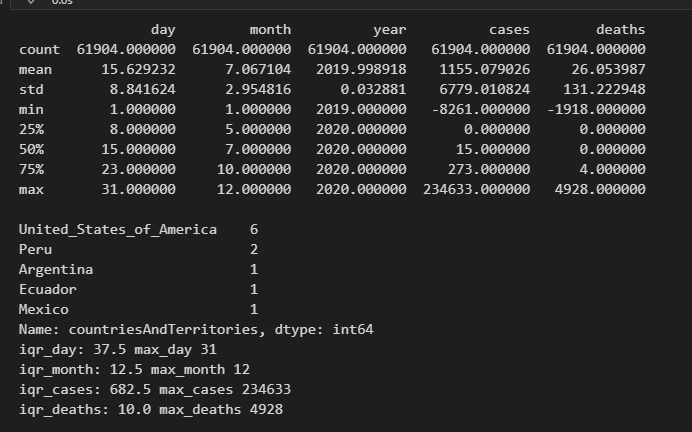
**Задание 11. Посмотреть статистику по данным, используя describe(). Сделать выводы о том, какие признаки содержат выбросы. Посмотреть, для каких стран количество смертей в день превысило 3000 и сколько таких дней было.**

Обычно выбросы могут быть обнаружены по следующим признакам:

* Большое значение максимального значения (которое значительно отличается от среднего и медианы).
* Большое значение стандартного отклонения (которое указывает на большую вариацию данных).

****

**Рисунок 11.1 – Сбор выбросов**

****

**Рисунок 11.2 – Результат выполнения**

Стандартное отклонение (std) для признаков "cases" и "deaths" составляет 6779.010824 и 131.222948 соответственно.

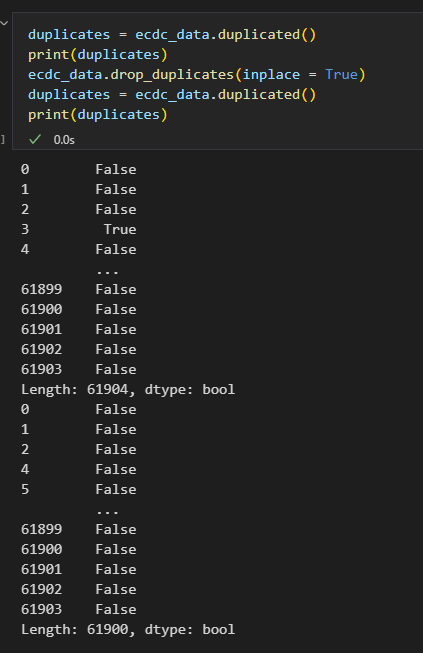
Максимальное значение для "cases" равно 234633, а для "deaths" равно 4928.

Минимальные значения для обоих признаков отрицательны, что может быть необычным для количественных данных о случаях и смертях.

Признаки "cases" и "deaths" содержат выбросы, так как максимальные значения превышают верхнюю границу выбросов,

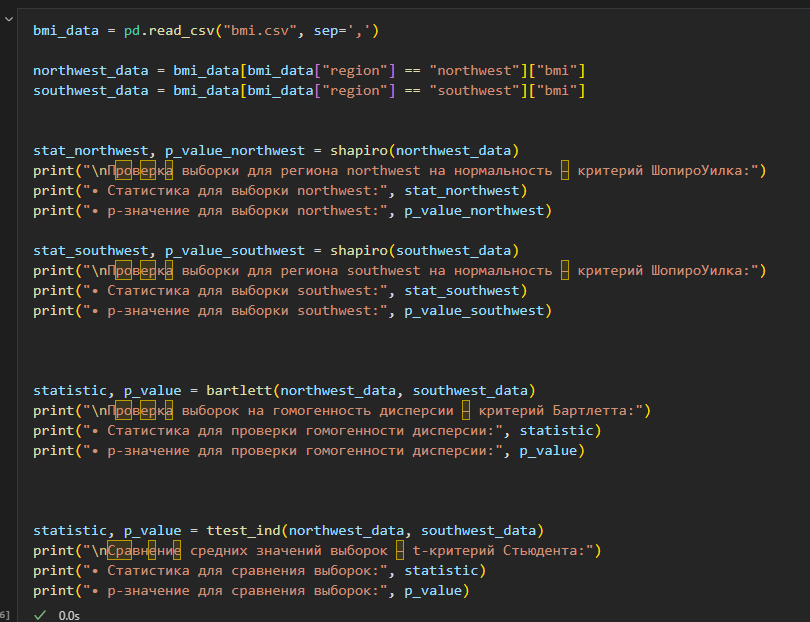
определенную по интерквартильному размаху.

**Задание 12. Найти дублирование данных. Удалить дубликаты.**

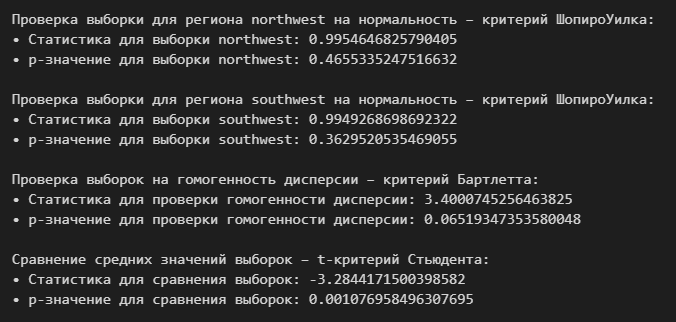
****

**Рисунок 12.1 – Нахождение и удаление дубликатов**

**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 ШопироУилка) и на гомогенность дисперсии (критерий Бартлетта)**

****

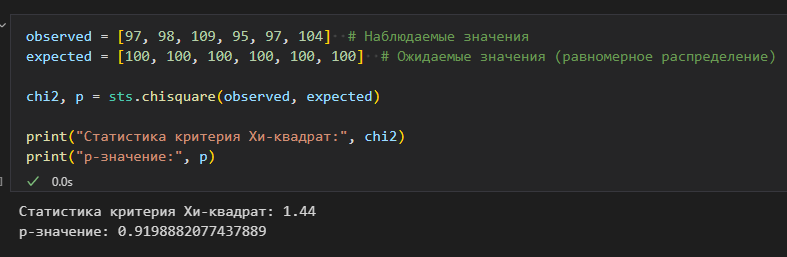
**Рисунок 13.1 – Обработка двух выборок**

****

**Рисунок 13.2 – Результат обработки**

**Задание 14. Кубик бросили 600 раз, получили следующие результаты: N 1 2 3 4 5 6 Количество выпадений 97 98 109 95 97 104. С помощью критерия Хи-квадрат проверить, является ли полученное распределение равномерным. Использовать функцию scipy.stats.chisquare().**

|  |  |
| --- | --- |
| N | Количество выпадений |
| 1 | 97 |
| 2 | 98 |
| 3 | 109 |
| 4 | 95 |
| 5 | 97 |
| 6 | 104 |

****

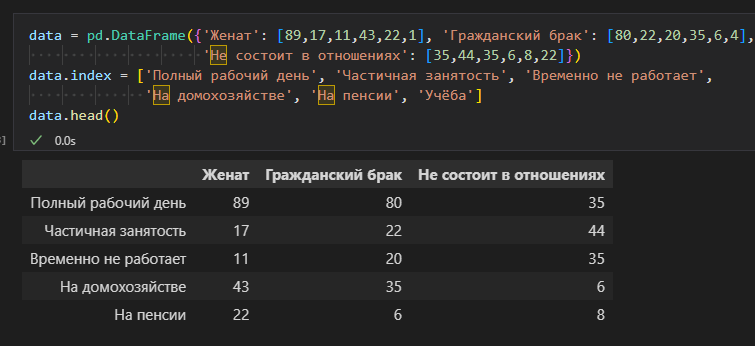
**Рисунок 14.1 – Проверка с помощью критерия Хи-квадрат**

Ожидаемые значения равномерно распределены, поэтому для каждого значения мы используем 100 (600 / 6).

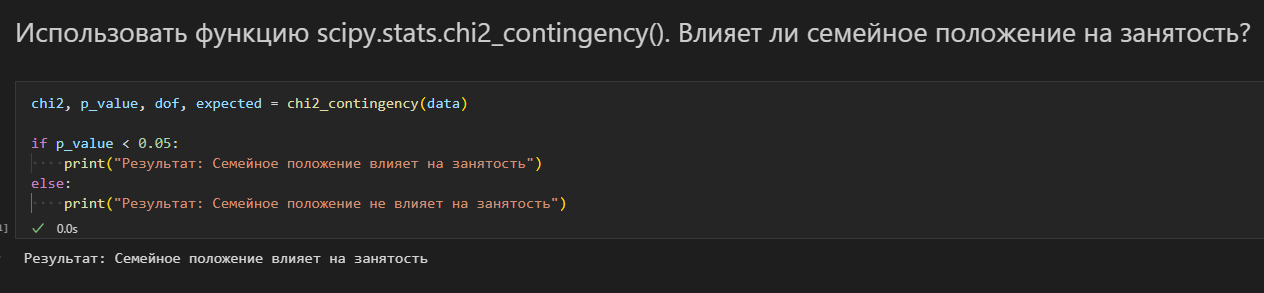
Функция scipy.stats.chisquare() возвращает статистику критерия Хи-квадрат и p-значение.

p-значение больше заданного уровня значимости (обычно 0.05); мы можем отклонить H1 в пользу H0.

**Задание 15. С помощью критерия Хи-квадрат проверить, являются ли переменные зависимыми. Создать датафрейм, используя следующий код:**

****

**Рисунок 15.1 – Создание датафрейма**

****

**Рисунок 15.2 – Проверка с помощью критерия Хи-квадрат**