|  |
| --- |
| Изображение выглядит как эмблема, символ, герб, нашивка  Автоматически созданное описание  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 № 3**

|  |  |  |  |
| --- | --- | --- | --- |
| Студент группы ИНБО-01-17 | *ИКБО-01-22 Прокопчук Роман Олегович* | (подпись) | |
| Ассистент | *Тетерин Николай Николаевич* | (подпись) | |
| Отчёт представлен | «\_\_» октября 2025 г. | |  | |

Москва, 2025 г.

**ЦЕЛЬ**

Ознакомиться с инструментами работы со статистикой в Python.

**ХОД РАБОТЫ**

**Задание**

1. Загрузить данные из файла “insurance.csv”.

2. С помощью метода describe() посмотреть статистику по данным. Сделать выводы.

3. Построить гистограммы для числовых показателей. Сделать выводы.

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. Сделать выводы.

5. Построить box-plot для числовых показателей. Названия графиков должны соответствовать названиям признаков. Сделать выводы.

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. Сделать выводы.

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

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. Сделать выводы на основе полученных p-значений.

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

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.

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.

12. Найти дублирование данных. Удалить дубликаты.

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).

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

15. С помощью критерия Хи-квадрат проверить, являются ли переменные зависимыми. Создать датафрейм, используя следующий код: data = pd.DataFrame({'Женат': [89,17,11,43,22,1], 'Гражданский брак': [80,22,20,35,6,4], 'Не состоит в отношениях': [35,44,35,6,8,22]}) data.index = ['Полный рабочий день','Частичная занятость','Временно не работает','На домохозяйстве','На пенсии','Учёба'] Использовать функцию scipy.stats.chi2\_contingency(). Влияет ли семейное положение на занятость?

16. Оформить отчет о проделанной работе, написать выводы.

**Insurance (задания 1-8)**

На рисунке 1 представлен скрипт для выгрузки и вывода описания данных.

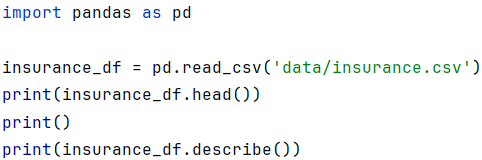


Рисунок 1 – Код загрузки данных

Результат выполнения программы представлен на рисунке 2.

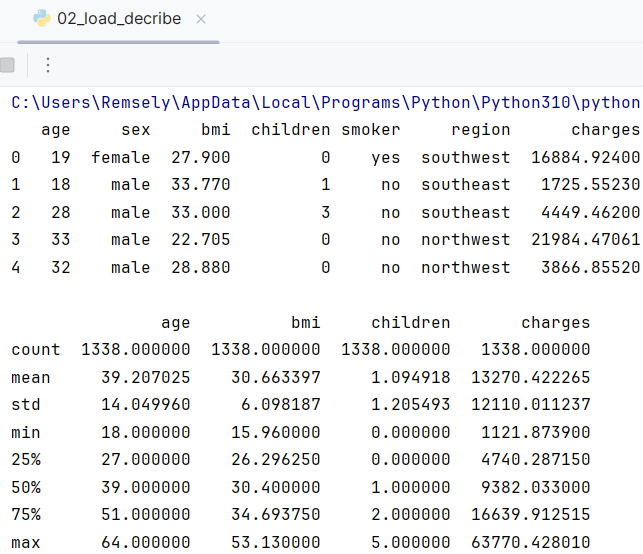


Рисунок 2 –Вывод описания данных

Вывод:

* В датасете 1338 записей о застрахованных лицах;
* Возраст варьируется от 18 до 64 лет, средний возраст ~39 лет;
* Индекс массы теласреднее значение 30.66, стандартное отклонение 6.1 указывает на умеренный разброс значений;
* Большинство застрахованных не имеют детей (медиана = 1), максимум -5 детей;
* Сильная вариативность расходов — от $1,121 до $63,770, среднее ($13,270) значительно выше медианы ($9,382), что указывает на наличие выбросов и правостороннюю асимметрию распределения.

На рисунке 3 представлен код для построения гистограмм для числовых значений. Итоговые диаграммы представлены на рисунке 4.

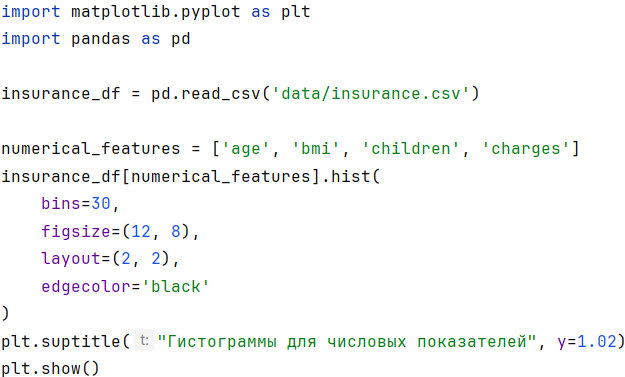


Рисунок 3 – Код для построения гистограмм

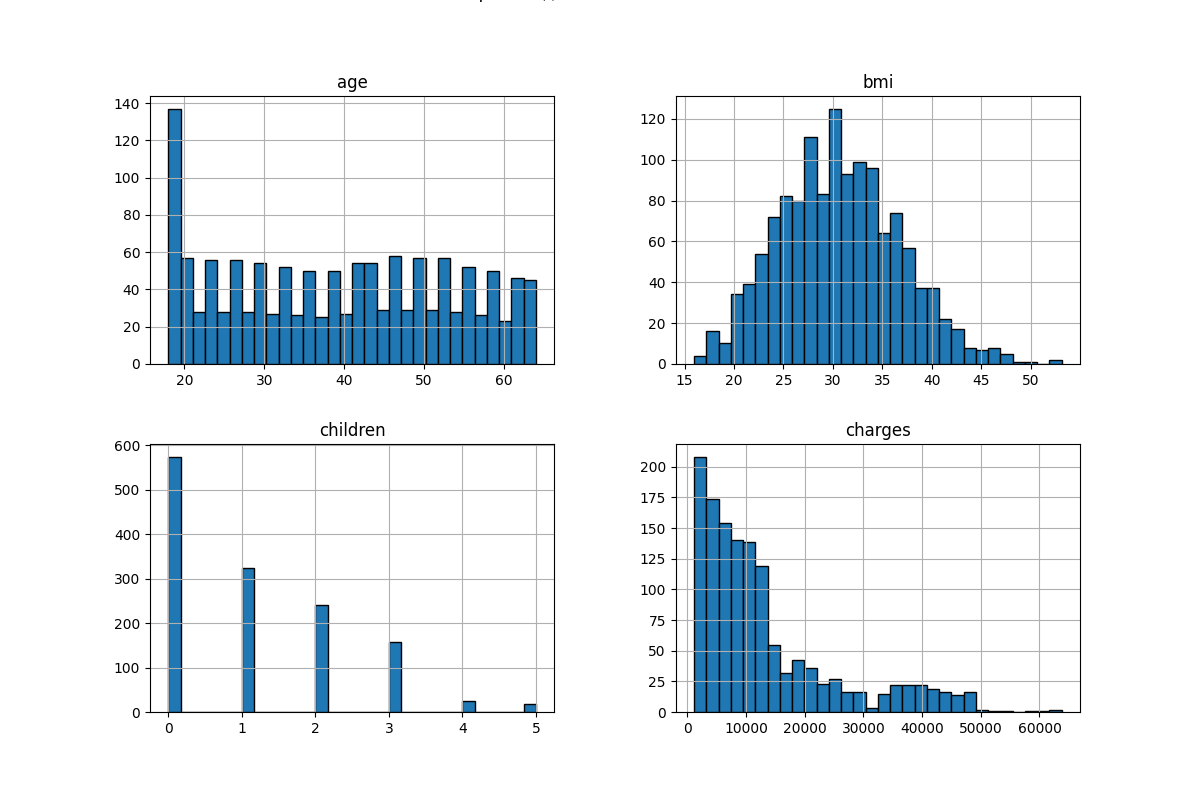


Рисунок 4 – Построенные гистограммы

Выводы:

* age: Распределение относительно равномерное с пиком в молодом возрасте;
* bmi: Распределение близко к нормальному с центром около 30;
* children: Преобладают застрахованные без детей, семьи с 4-5 детьми встречаются редко.
* charges: Большинство расходов концентрируются в низком диапазоне (0-20000$), но есть значительное количество случаев с высокими расходами (35000+ $)/

Код для выполнения четвертого задания представлен на рисунке 5. Вывод программы представлен на рисунке 6.



Рисунок 5 – Код для выполнения четвертого задания

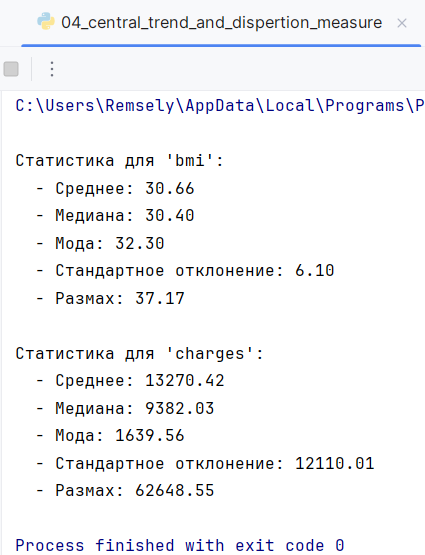
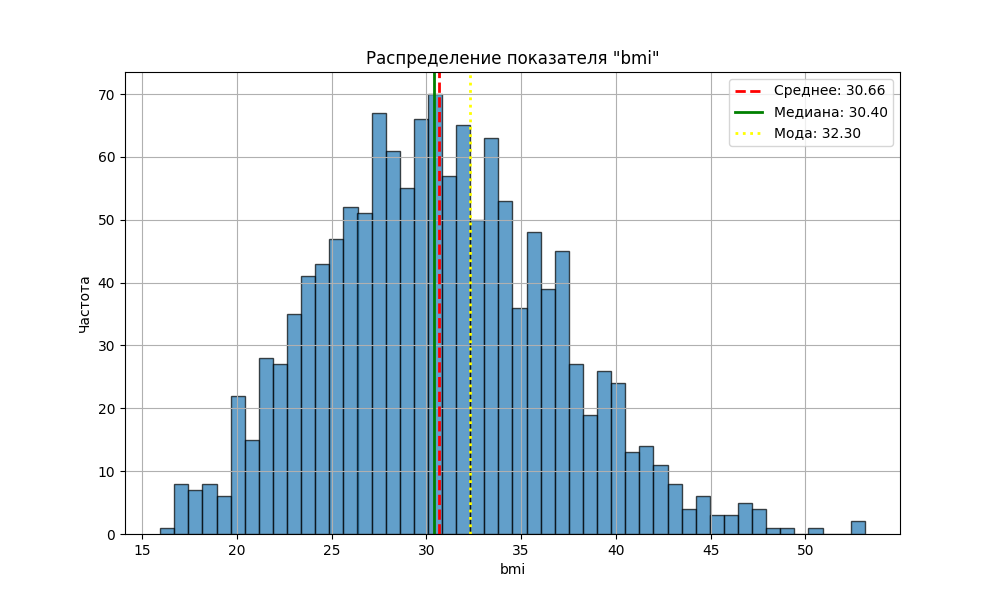
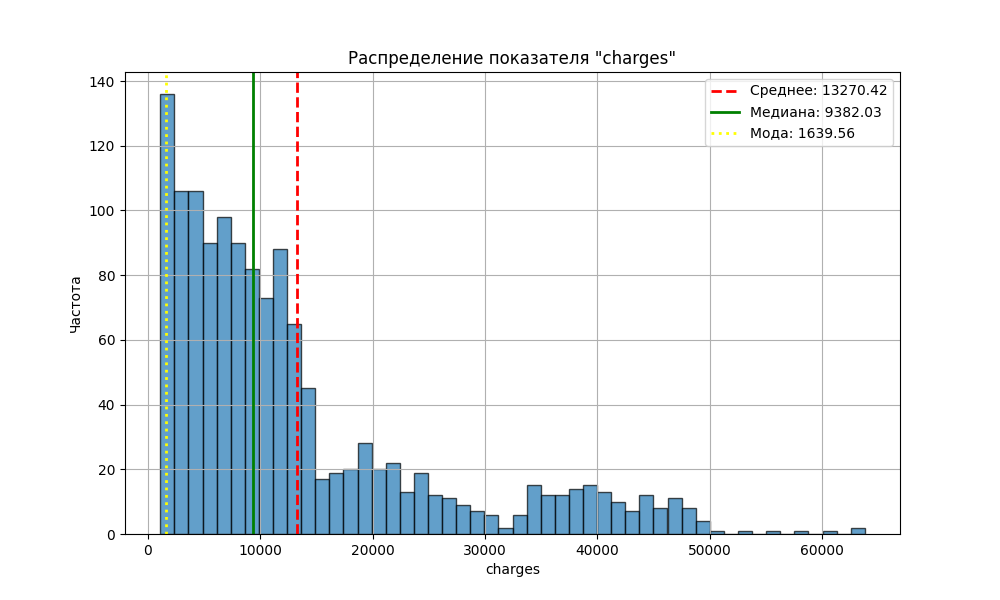


Рисунок 6 – Вывод программы

Созданные диаграммы представлены на рисунках 7-8

  
Рисунок 7 – Распределение показателя bmi

  
Рисунок 8 – Распределение показателя charges

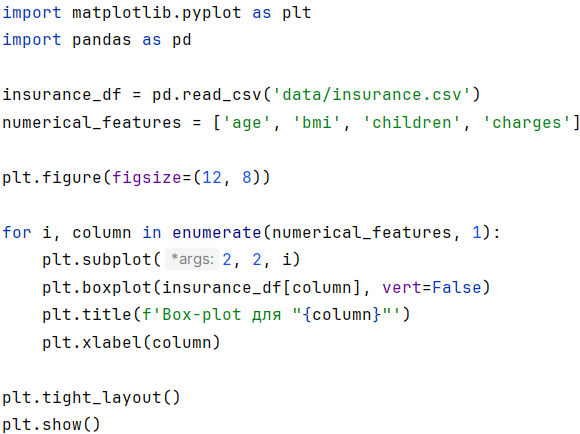
Выводы для bmi:

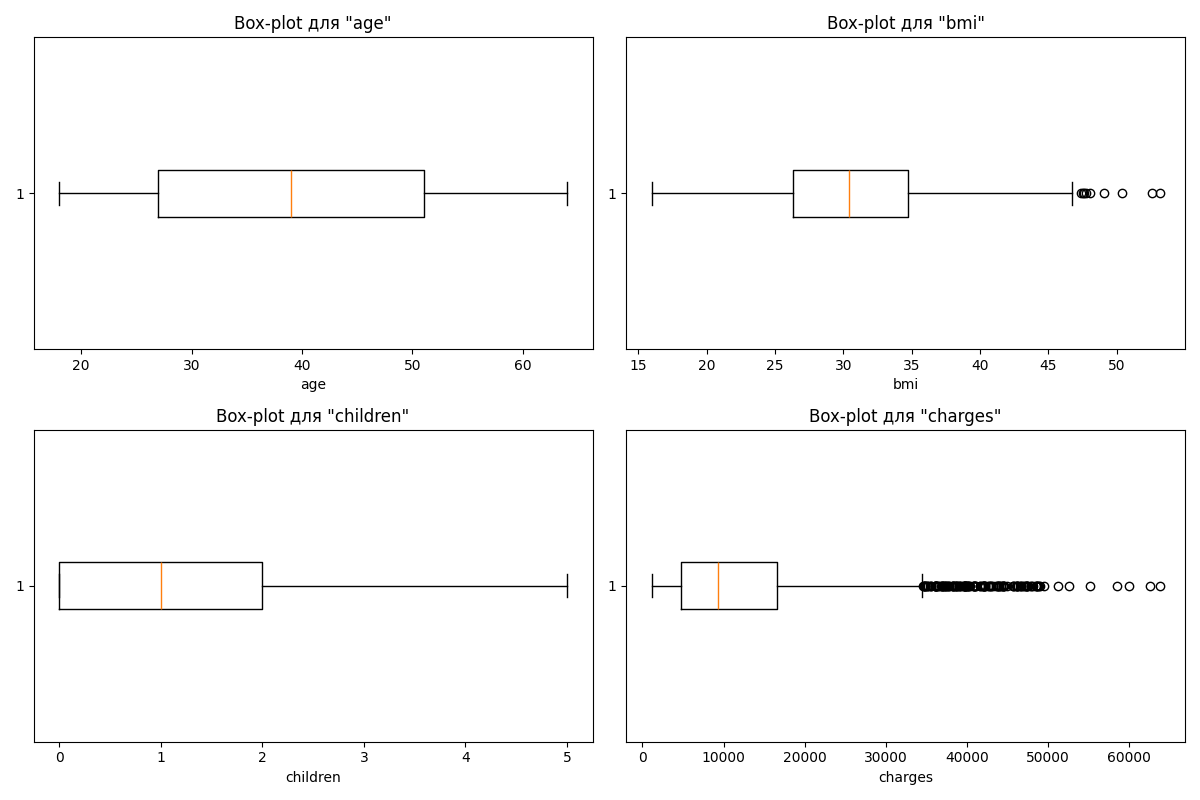
* Среднее (30.66), медиана (30.40) и мода (32.30) находятся близко друг к другу, что подтверждает относительную симметричность распределения;
* Стандартное отклонение 6.10 показывает умеренную вариабельность;
* Размах 37.17 (от 15.96 до 53.13).

**Выводы для charges:**

* Сильное расхождение между средним (13,270) и медианой (9,382) указывает на правостороннюю асимметрию;
* Мода (1,639.56) значительно ниже медианы, что подтверждает скопление данных в нижней части распределения;
* Огромный размах (62,648.55) и очень большое стандартное отклонение (12,110) свидетельствуют о высокой неоднородности расходов;
* Распределение типично для страховых расходов: большинство случаев имеют низкие затраты, но редкие серьезные заболевания создают очень высокие расходы.

На русинке 9 представлен скрипт для создания box-plot. Сами диаграммы представлены на рисунке 10.

  
Рисунок 9 – Код для создания box-plot

  
Рисунок 10 – Созданные box-plot

Выводы:

* Age – распределение симметрично;
* bmi – распределение почти симметрично с небольшим количеством выбросов в верхней части (люди с очень высоким ИМТ >45-50);
* children – Сильно асимметричное распределение — медиана смещена к нижней границе, большинство значений концентрируется на 0-2 детей, значения 4-5 являются выбросами;
* charges – наиболее асимметричное распределение с множеством выбросов в верхней части. Медиана смещена к нижней границе коробки, верхний ус значительно длиннее нижнего, что подтверждает правостороннюю асимметрию. Многочисленные выбросы (>$30,000) соответствуют случаям с серьезными заболеваниями или осложнениями.

Код для проверки выполнения центральной предельной теоремы представлен на рисунке 11. Вывод программы представлен на рисунке 12, получившиеся диаграммы – на рисунке 13.

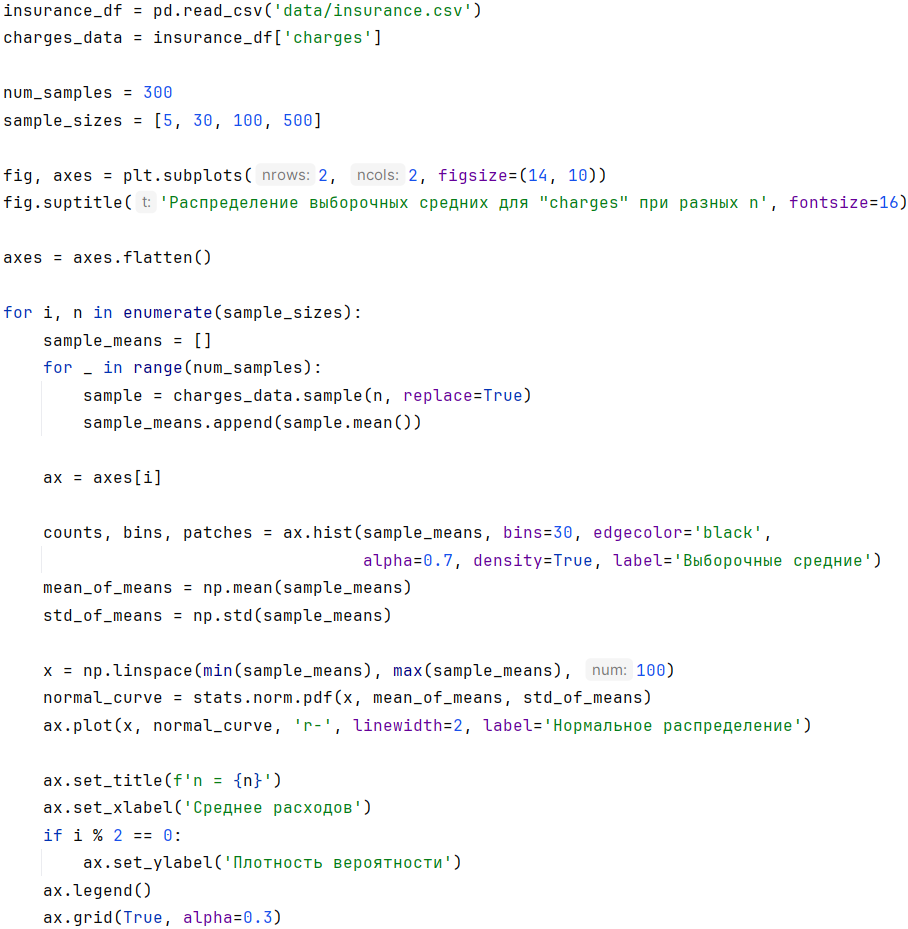


Рисунок 11 – Код для проверки центральной предельной теоремы

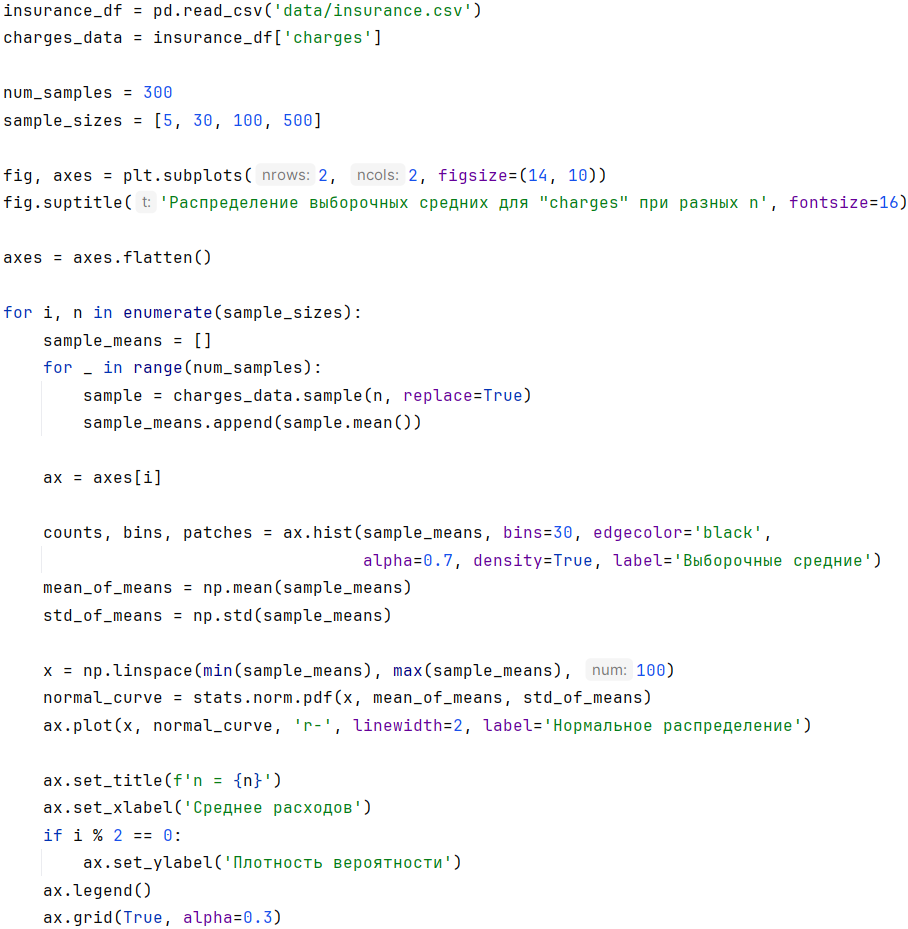


Рисунок 11 – Код для проверки центральной предельной теоремы

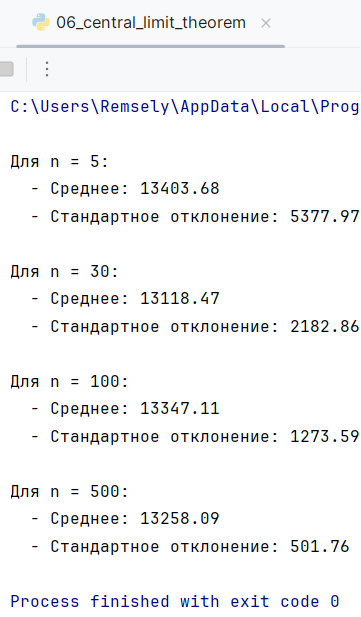


Рисунок 12 – Вывод программы

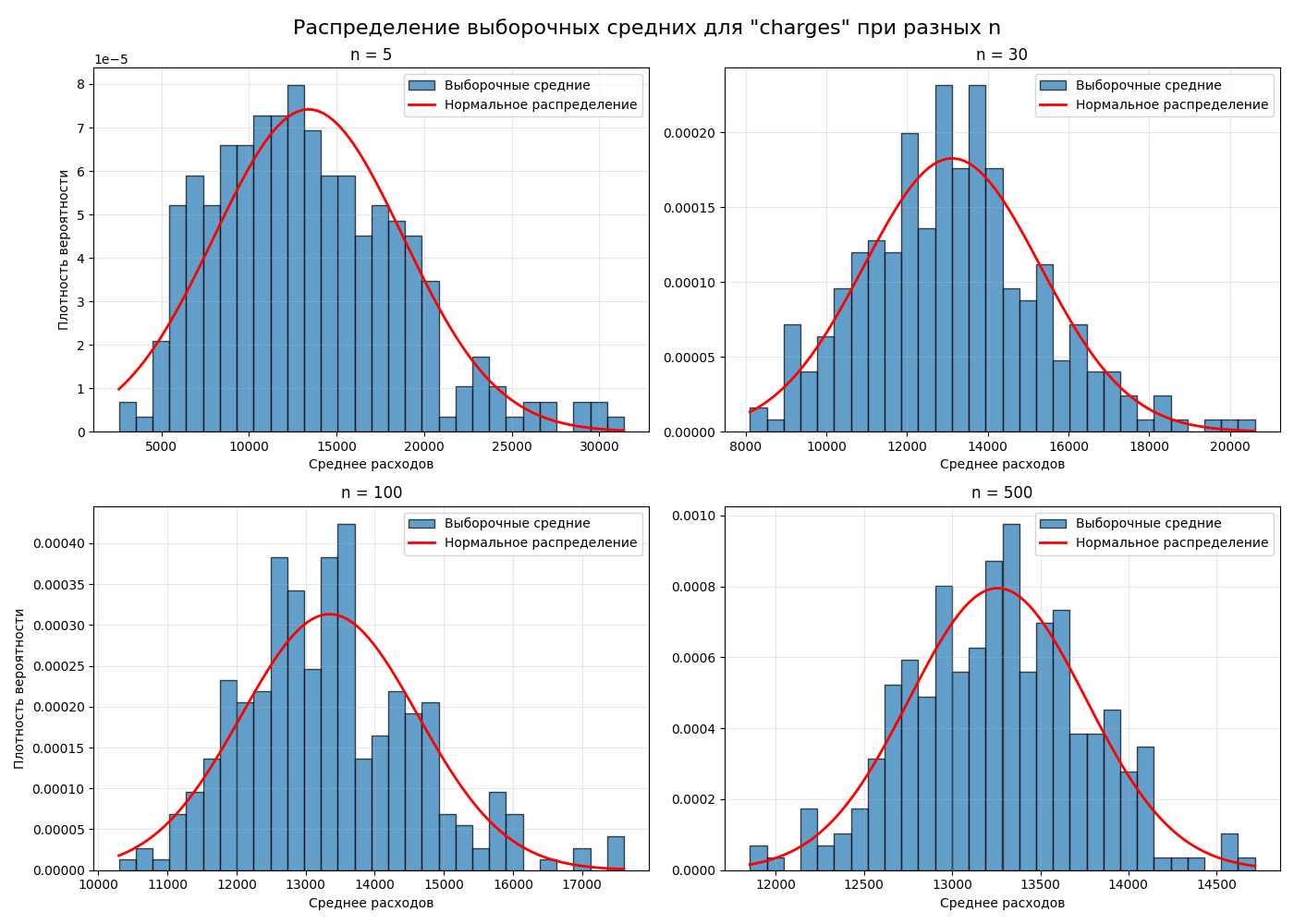


Рисунок 13 – Результаты в виде гистограмм

При увеличении размера выборки n распределение выборочных средних приближается к нормальному, что полностью подтверждает центральную предельную теорему.

Код для нахождения доверительных интервалов представлен на рисунке 14, результат его работы – на рисунке 15.

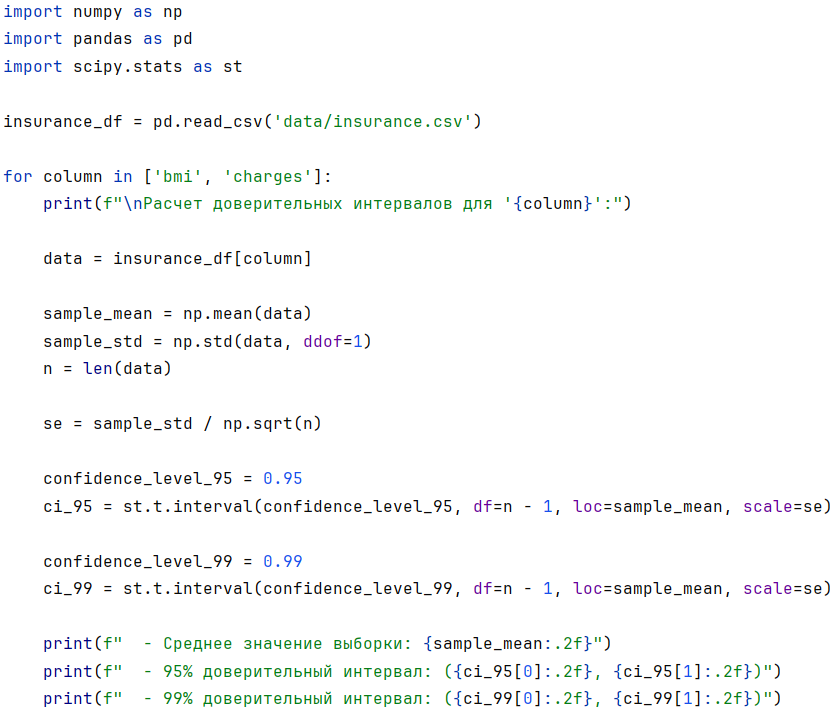
****

Рисунок 14 – Код для нахождения доверительных интервалов

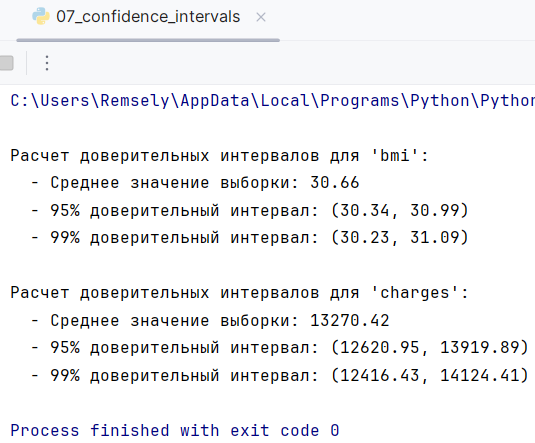


Рисунок 15 – Вывод программы

Код для проверки распределений признаков на нормальность представлен на рисунке 16, результат его работы – на рисунке 17, построенные диаграммы – на рисунках 18-19.

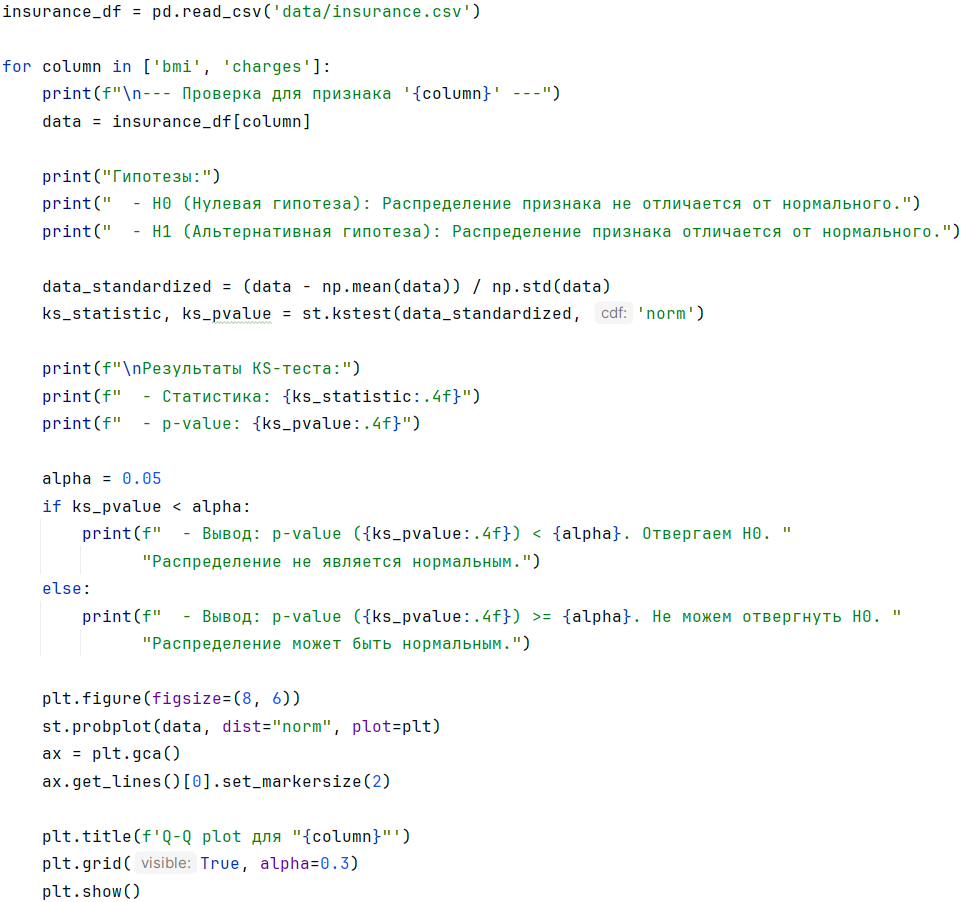


Рисунок 16 – Код для проверки распределений на нормальность

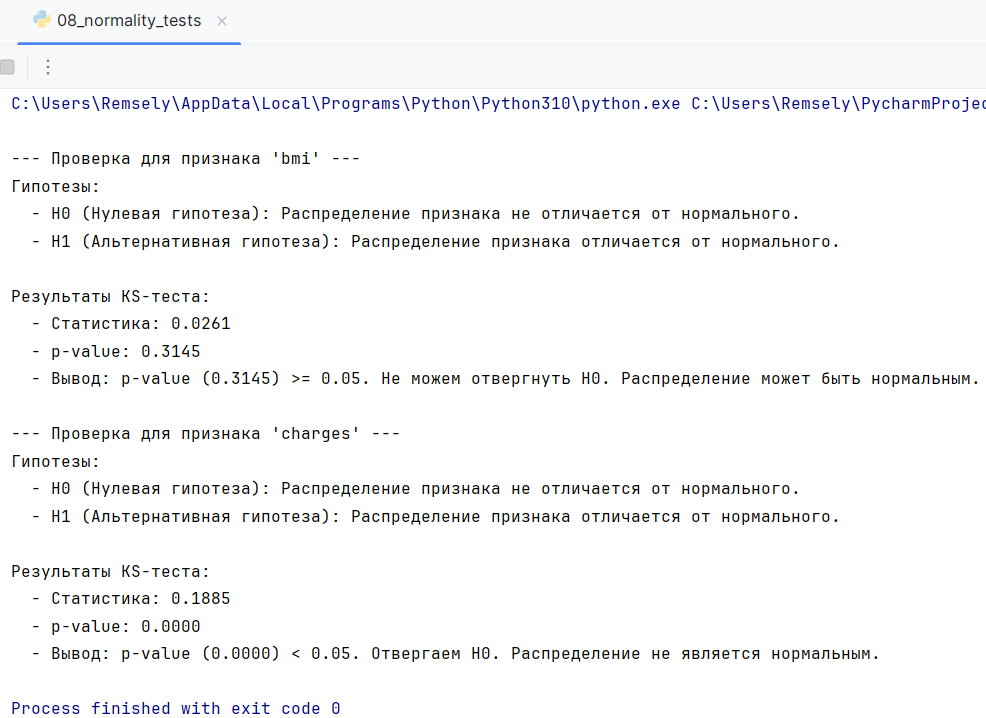


Рисунок 17 – Вывод программы

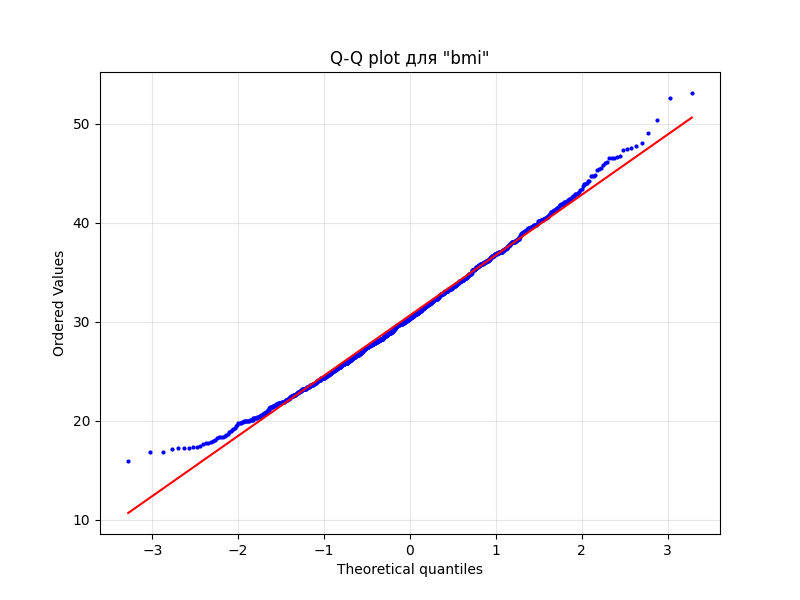


Рисунок 18 – Q-Q plot для bmi

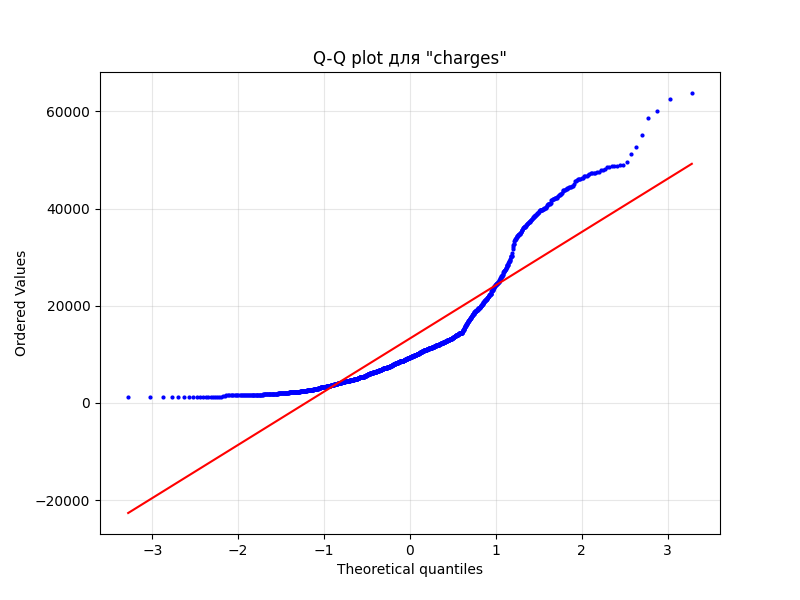


Рисунок 19 – Q-Q plot для charges

bmi: Распределение ИМТ соответствует нормальному закону. Небольшие отклонения на краях Q-Q plot (левый и правый концы) указывают на немного более тяжелые хвосты по сравнению с идеальным нормальным распределением, но эти отклонения не являются критичными

charges: Распределение расходов значительно отличается от нормального. Нарушение нормальности объясняется природой данных: страховые расходы не могут быть отрицательными и имеют редкие, но очень высокие значения.

**ECDC Cases (задания 9-12)**

Код для удаления пропущенных значений представлен на рисунке 20, результат его работы – на рисунке 21.



Рисунок 20 – Код для удаления пропущенных значений

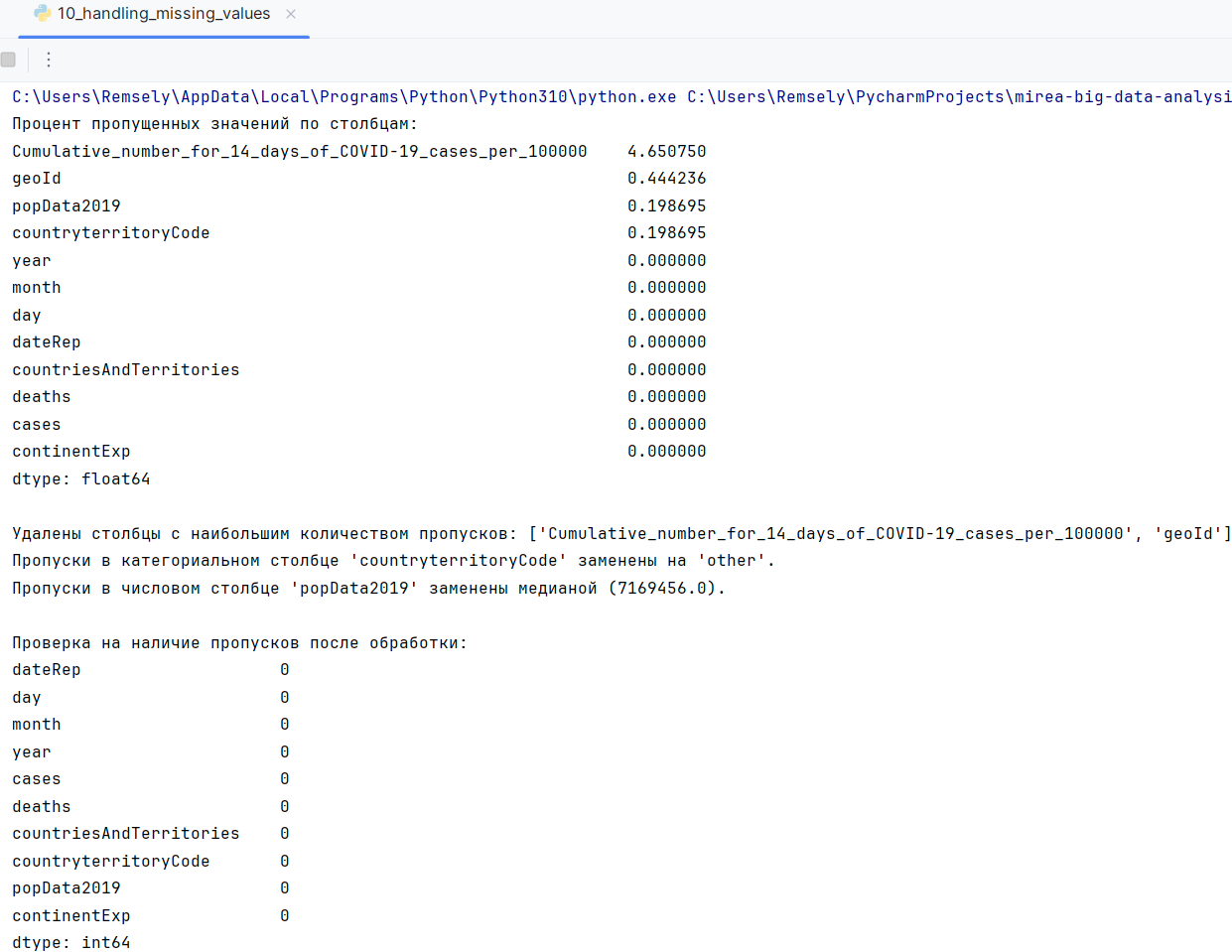


Рисунок 21 – Результат удаления пропущенных значений

Код для нахождения выбросов (задание 11) представлен на рисунке 22, результат его работы – на рисунке 23.

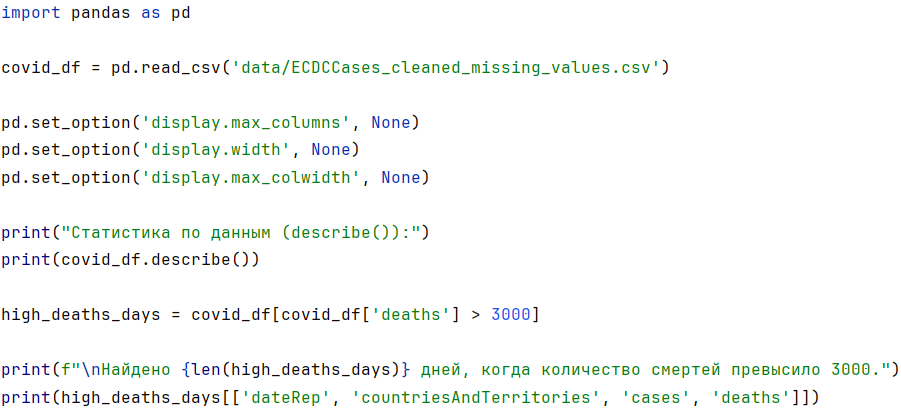


Рисунок 22 – Код для нахождения выбросов

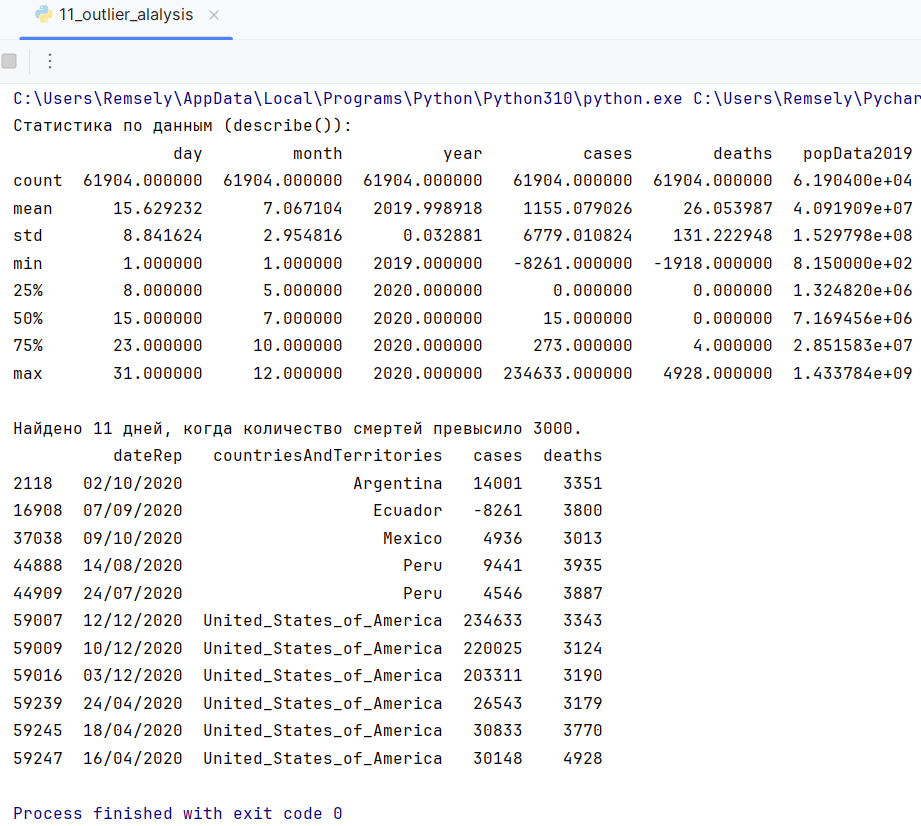


Рисунок 23 – Сравнение времени работы t-SNE и UMAP

Выводы:

* deaths – Минимальное значение -1918 является явной ошибкой данных (отрицательное количество смертей невозможно), требует исправления. Максимум 4928 смертей в день и стандартное отклонение 131.22 при среднем 26.05 указывают на экстремальные выбросы;
* cases – Аналогично содержит отрицательные значения (вероятно, корректировки данных);
* popData2019 – Огромный разброс — от 815 человек (микрогосударства) до 1.43 млрд (Китай, Индия), что естественно для страновых данных.

Код для удаления дублирующихся данных представлен на рисунке 24, результат его работы – на рисунке 25.



Рисунок 24 – Код для удаления дублей

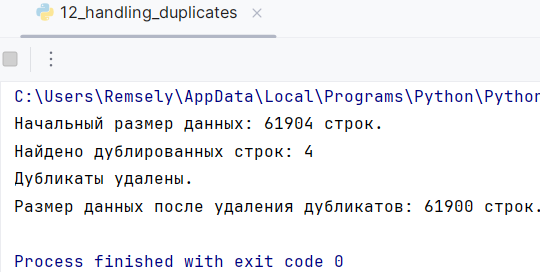


Рисунок 25 – Результат удаления дублей

**Тесты (задания 13-15)**

Код для выполнения задания 13 представлен на рисунке 26, его вывод – на рисунке 27.

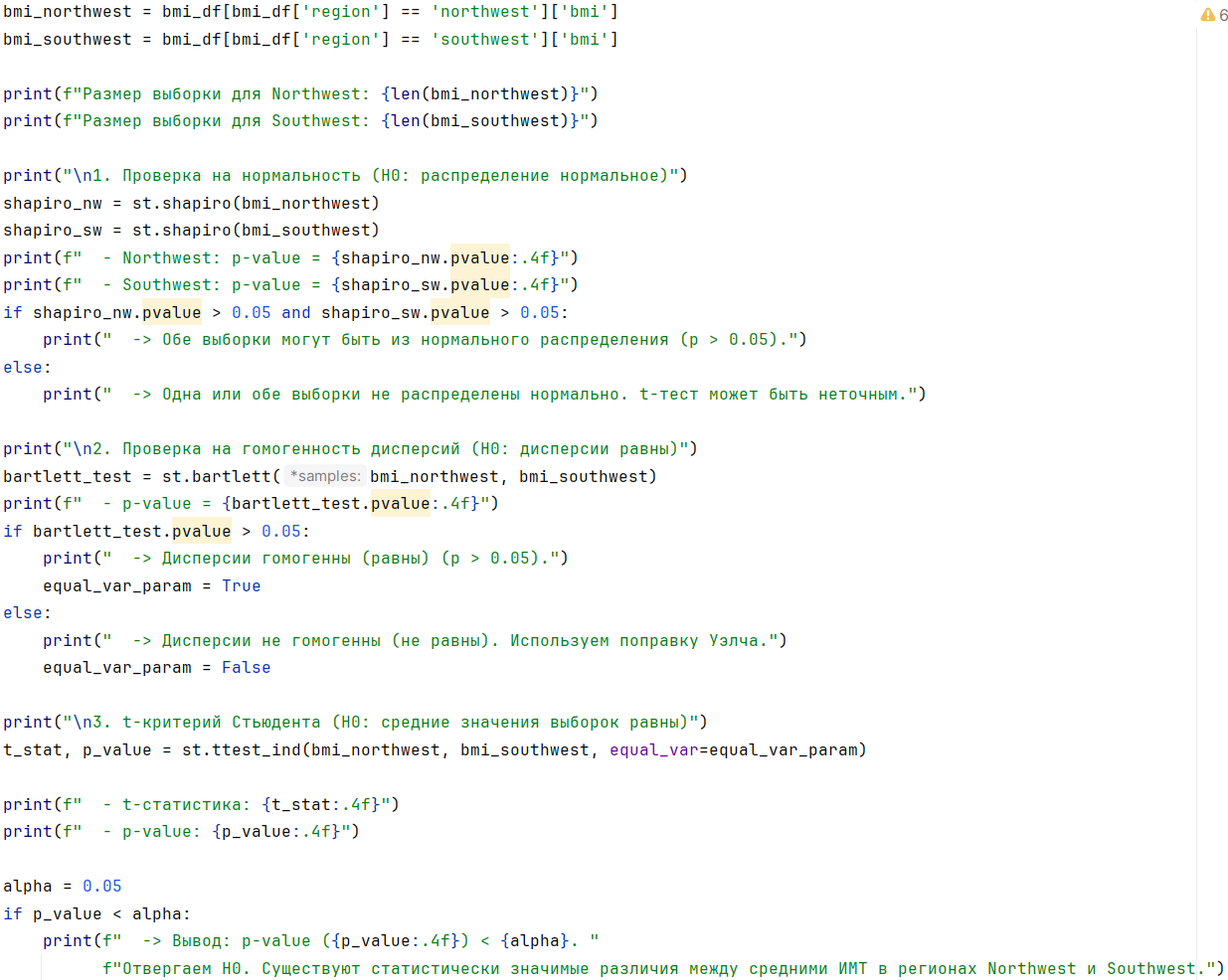


Рисунок 26 – Код для выполнения задания 13

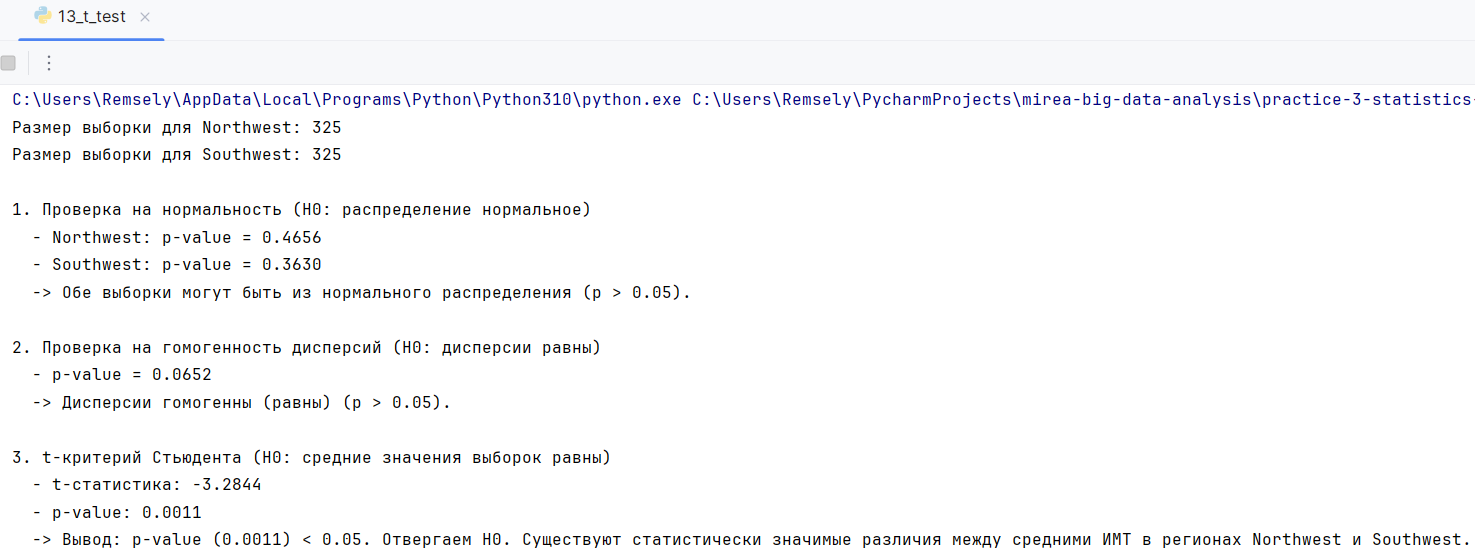


Рисунок 27 – Результаты проверок

Код для выполнения задания 14 представлен на рисунке 28, его вывод – на рисунке 29.

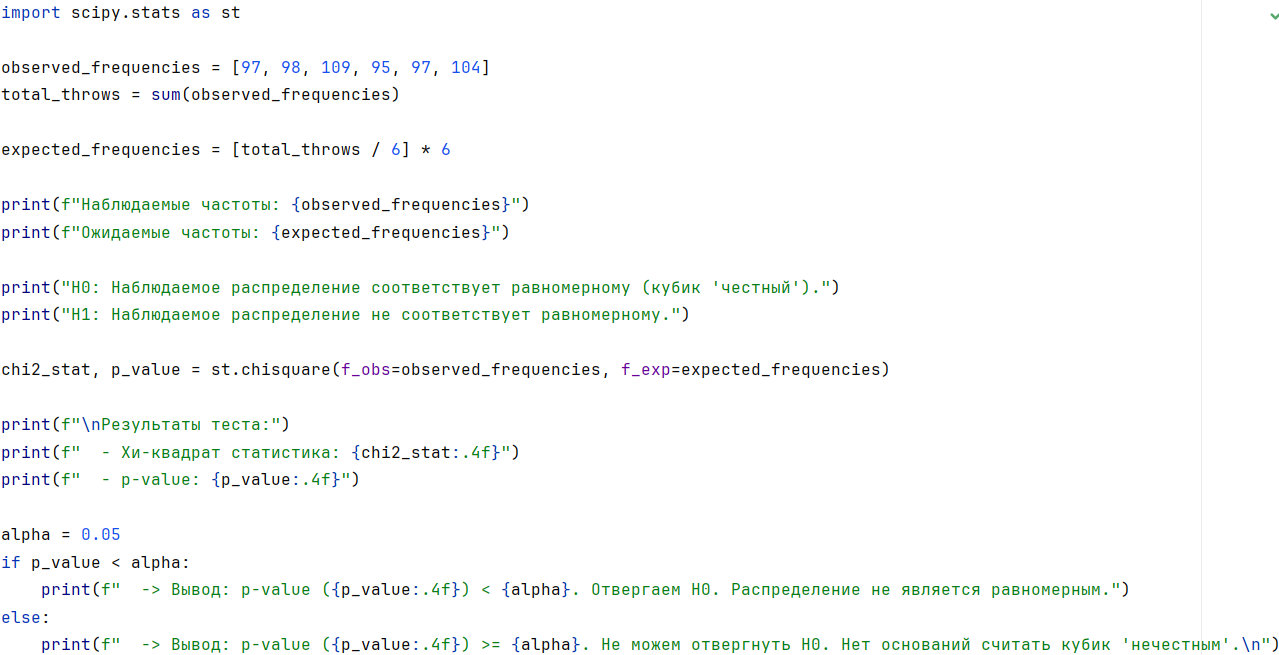


Рисунок 28 – Код для выполнения задания 14

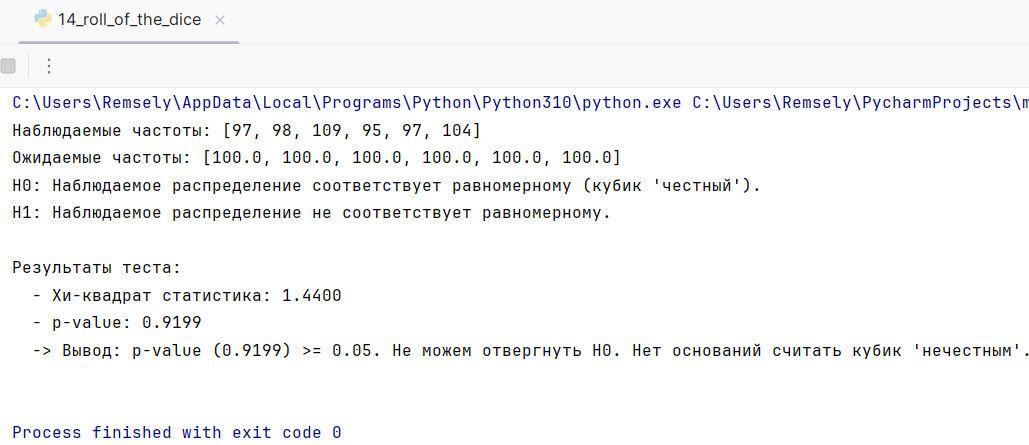


Рисунок 29 – Результаты проверок

Код для выполнения задания 15 представлен на рисунке 30, его вывод – на рисунке 31.

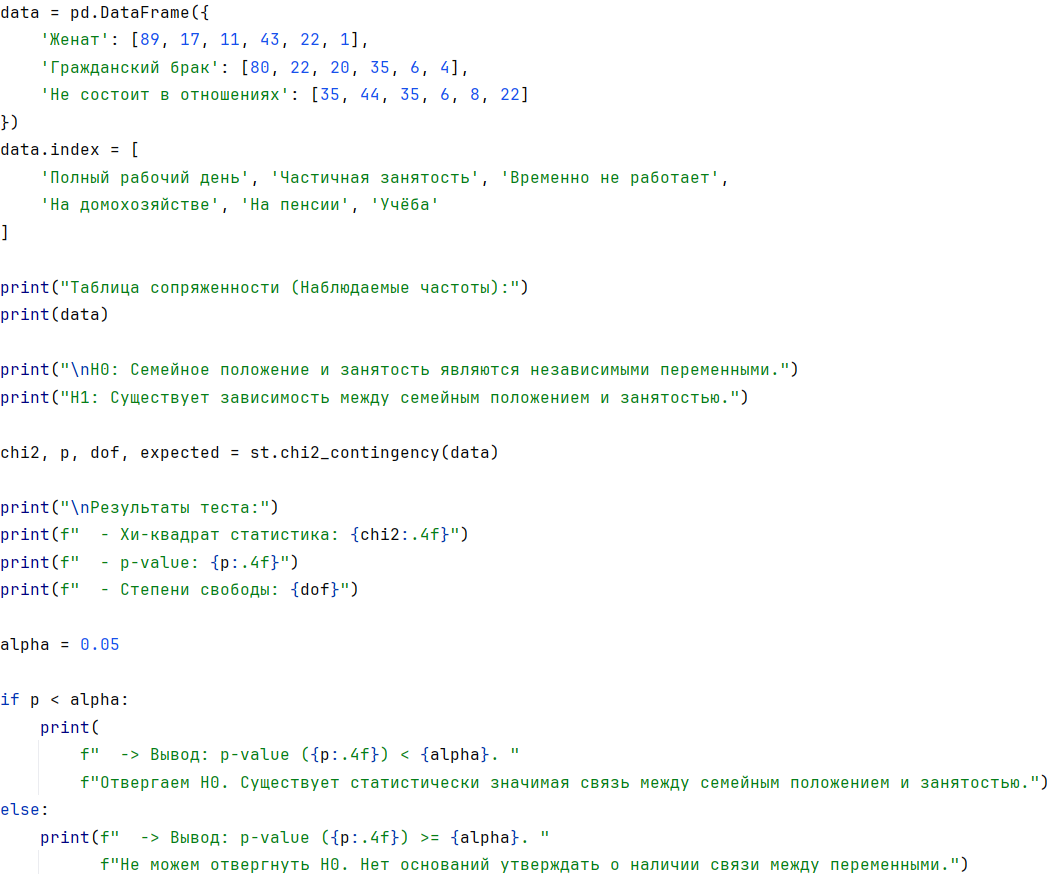


Рисунок 30 – Код для выполнения задания 15

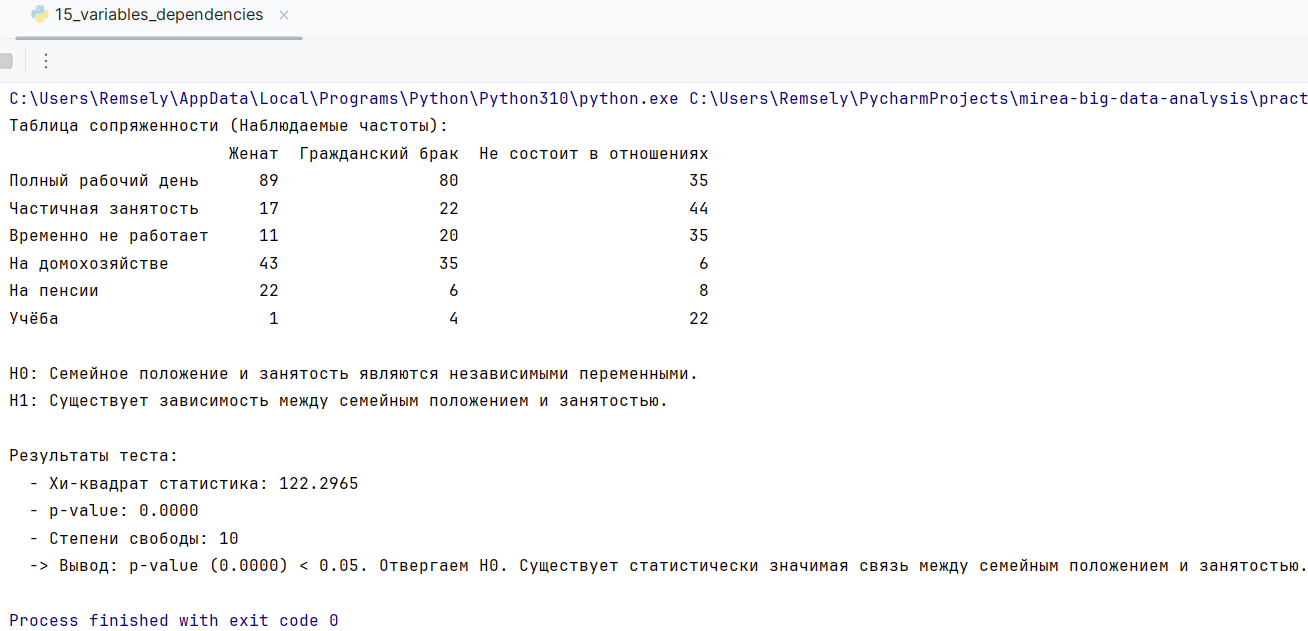


Рисунок 31 – Результаты проверок

**ВЫВОД**

В ходе выполненной практической работы проведено ознакомление с инструментами работы со статисткой в Python.