|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  |  |  | | МИНОБРНАУКИ РОССИИ | | | | Федеральное государственное бюджетное образовательное учреждение  высшего образования  **«МИРЭА – Российский технологический университет»**  **РТУ МИРЭА** | | | | |
| Институт информационных технологий (ИИТ) | |
| Кафедра Прикладной Математики (ПМ) | |

|  |  |  |  |
| --- | --- | --- | --- |
| **ОТЧЁТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ №3** | | | |
| **по дисциплине «Технологии и инструментарий анализа больших данных»** | | | |
|  | | | |
| Выполнил студент группы ИКБО-14-20 | | Вежновец Ф. Ю. | |
|  | |  | |
| Принял: асистент | | Горячев А. А. | |
| Практические работы выполнены | «\_\_\_»\_\_\_\_\_\_\_2023 г. | | (подпись студента) | |
| «Зачтено» | «\_\_\_»\_\_\_\_\_\_\_2023 г. | | (подпись руководителя) | |
|  |  | |  | |

Москва 2023

Задание 1

Загрузить данные из файла “insurance.csv”.

Реализация

Результат работы программы и код предоставлена на рисунке 1.

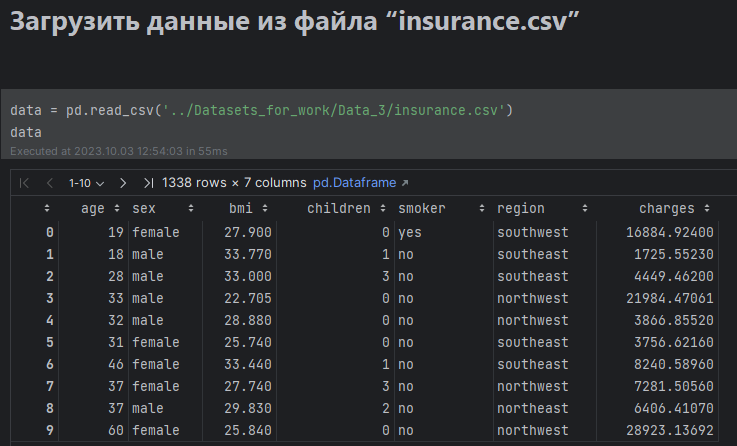


Рисунок 1 – Загрузка данных

Задание 2

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

Реализация

Результат работы программы и код представлен на рисунке 2.

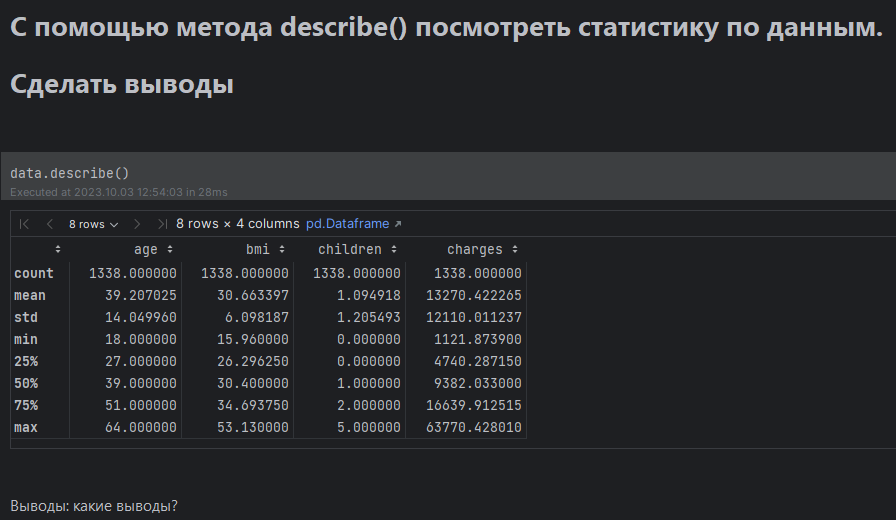


Рисунок 2 – Результат работы программы

Задание 3

Построить гистограммы для числовых показателей. Сделать выводы.

Реализация

Результат работы программы и код представлен на рисунке 3.

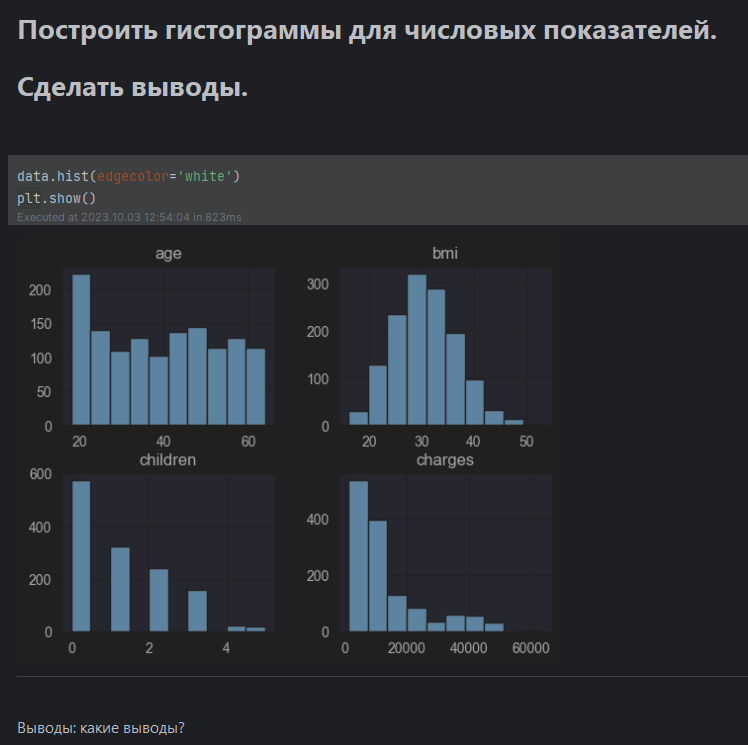


Рисунок 3 – Результат работы программы

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. Сделать выводы.

Реализация

Результат работы программы и код представлен на рисунках 4, 5, 6, 7.

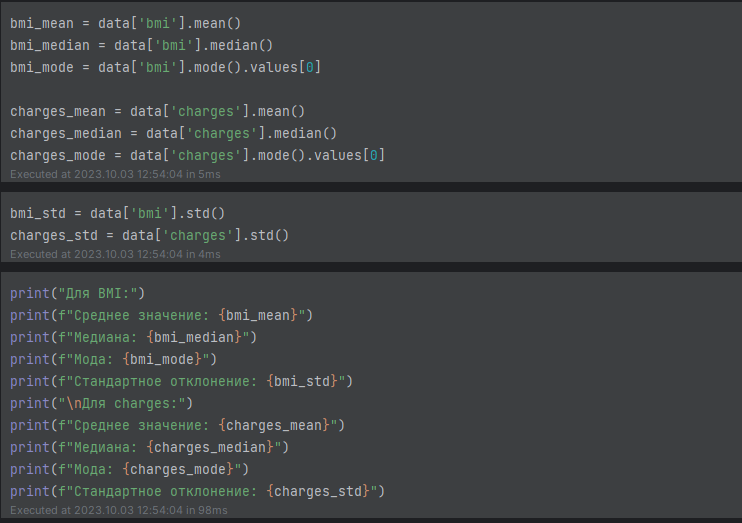


Рисунок 4 – Результат работы программы

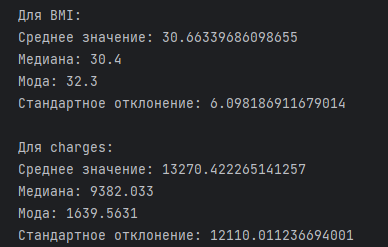


Рисунок 5 – Результат работы программы



Рисунок 6 – Результат работы программы

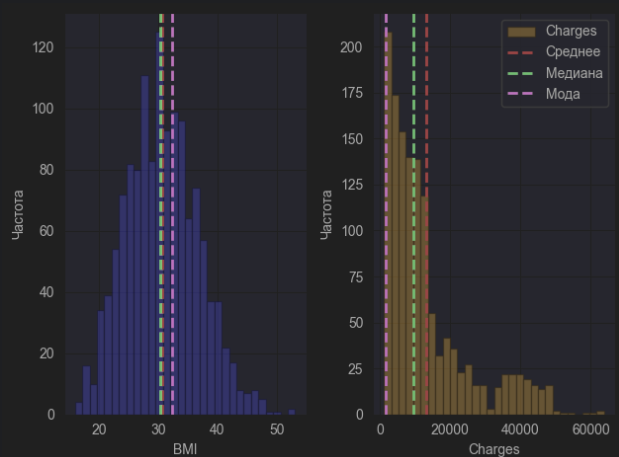


Рисунок 7 – Результат работы программы

Задание 5

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

Реализация

Результат работы программы и код представлен на рисунках 8 и 9.

|  |
| --- |
|  |

Рисунок 8 – Результат работы программы

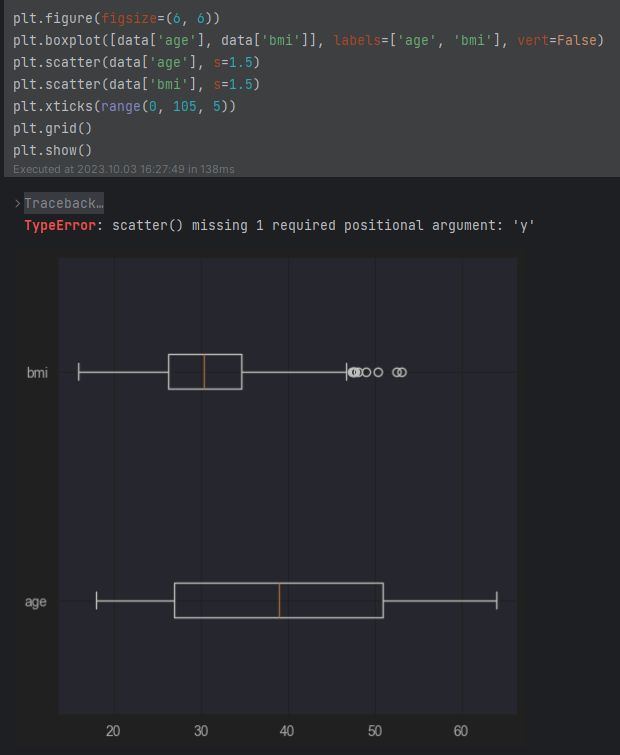


Рисунок 9 – Результат работы программы

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. Сделать выводы.

Реализация

Результат работы программы и код представлен на рисунках 10, 11, 12, 13.

|  |
| --- |
|  |

Рисунок 10 – Результат работы программы

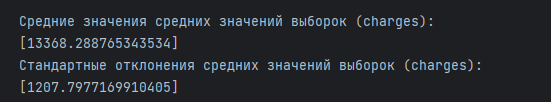


Рисунок 11 – Результат работы программы

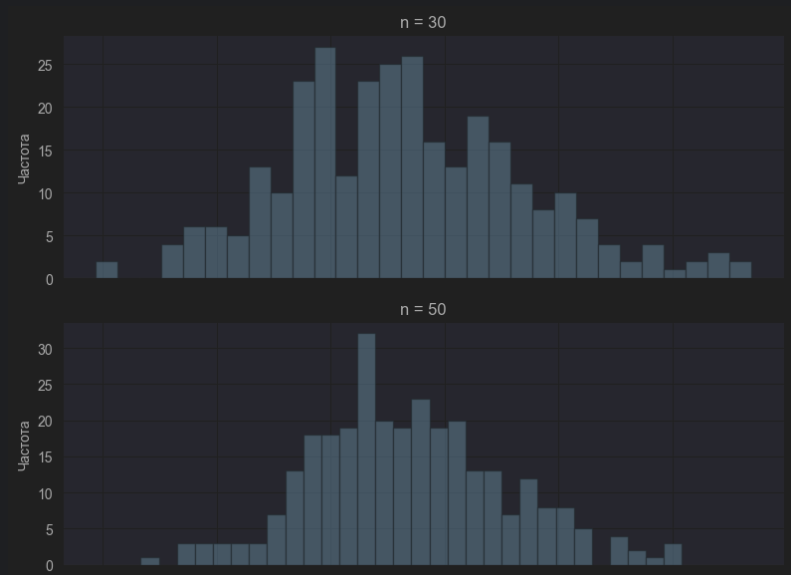


Рисунок 12 – Результат работы программы



Рисунок 13 – Результат работы программы

Задание 7

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

Реализация

Результат работы программы и код представлен на рисунках 14, 15.

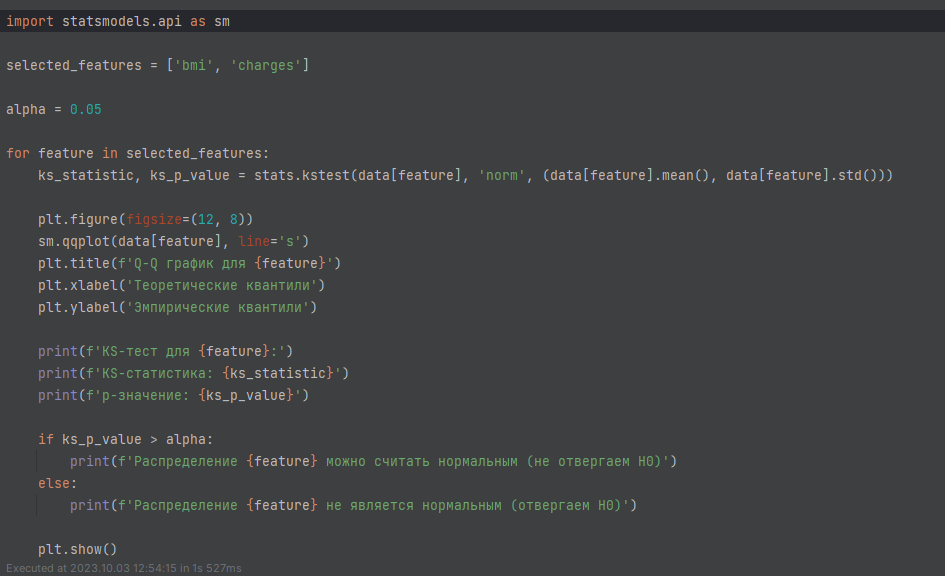
|  |
| --- |
|  |
| Рисунок 14 – Результат работы программы    Рисунок 15 – Результат работы программы |

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. Сделать выводы на основе полученных p-значений.

Реализация

Результат работы программы и код представлен на рисунках 16, 17, 18.

Рисунок 16 – Результат работы программы

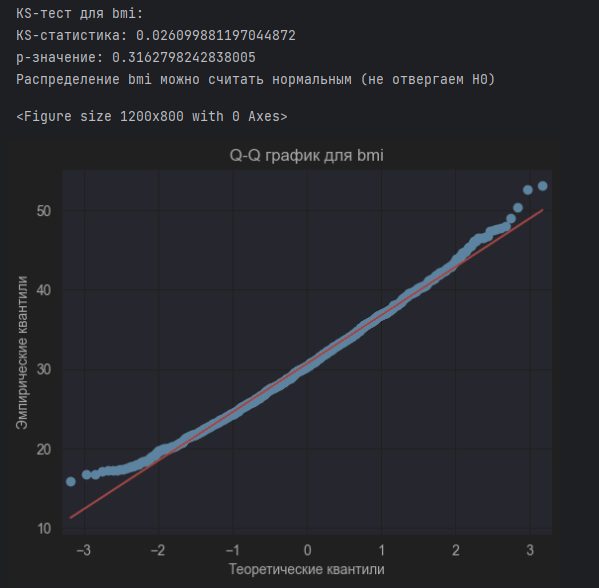


Рисунок 17 – Результат работы программы

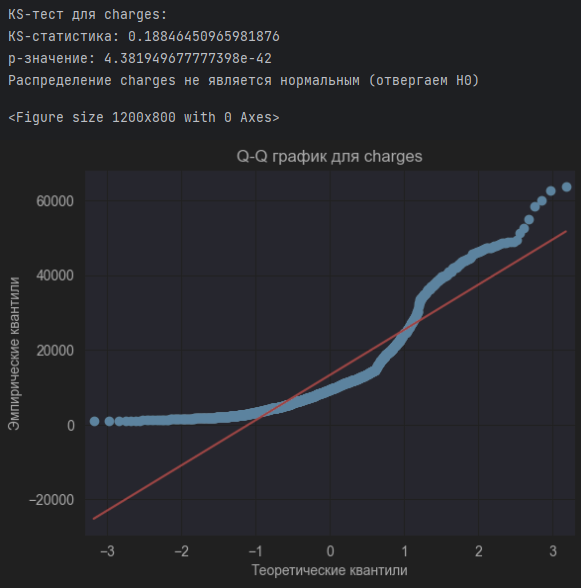


Рисунок 18 – Результат работы программы

Задание 9

Загрузить данные из файла “ECDCCases.csv”.

Реализация

Результат работы программы и код представлен на рисунке 19.



Рисунок 19 – Результат работы программы

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.

Реализация

Результат работы программы и код представлен на рисунках 20, 21, 22, 23.

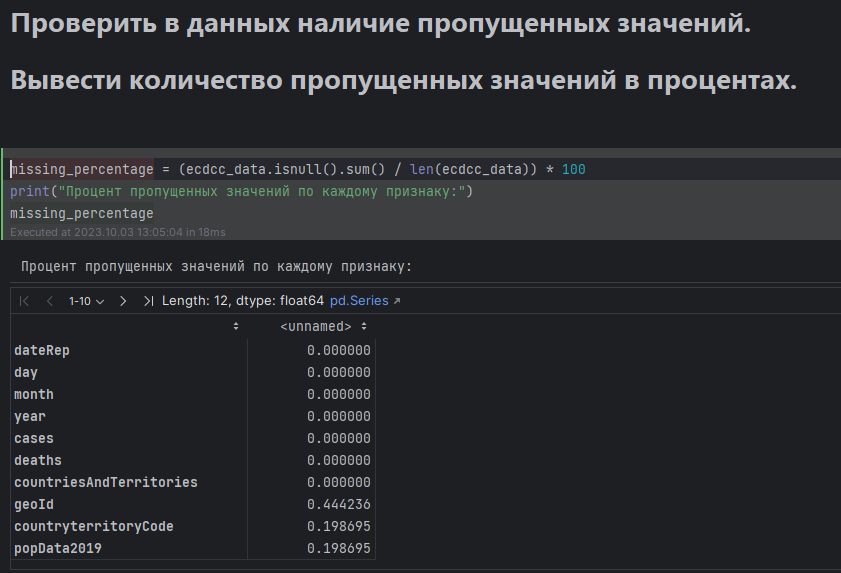


Рисунок 20 – Результат работы программы

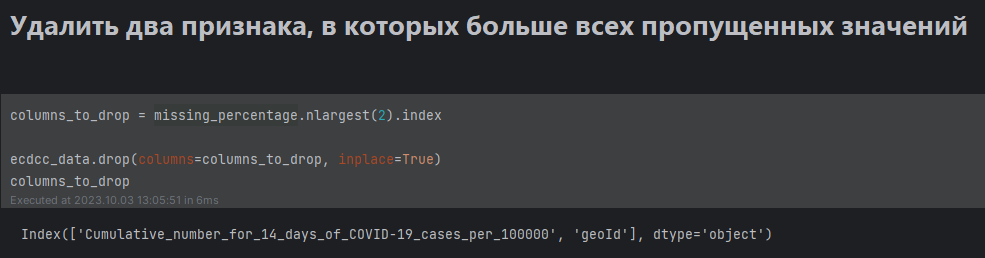


Рисунок 21 – Результат работы программы

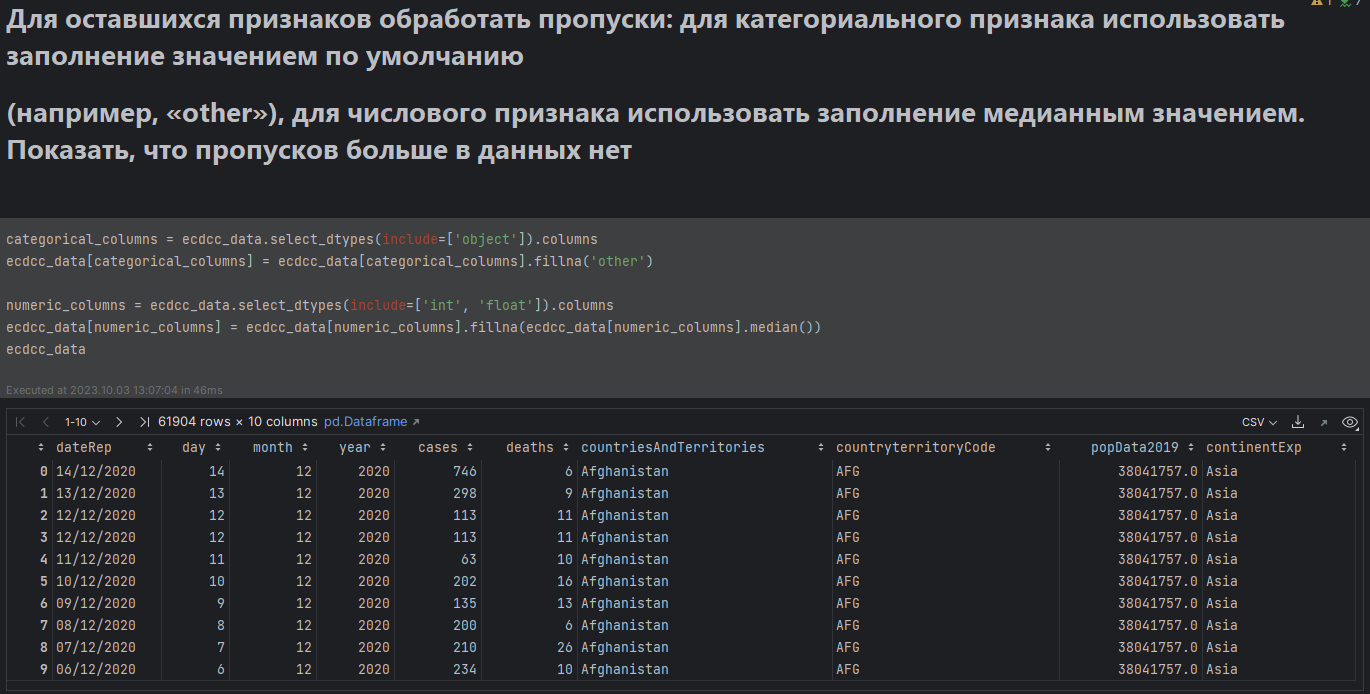


Рисунок 22 – Результат работы программы

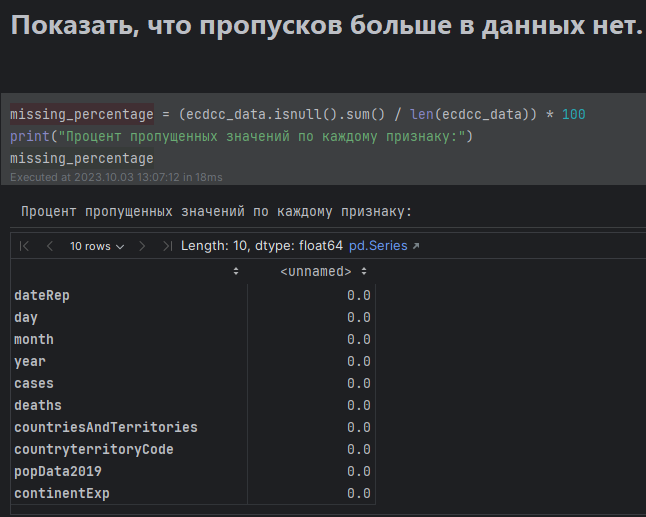


Рисунок 23 – Результат работы программы

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.

Реализация

Результат работы программы и код представлен на рисунках 24, 25, 26.

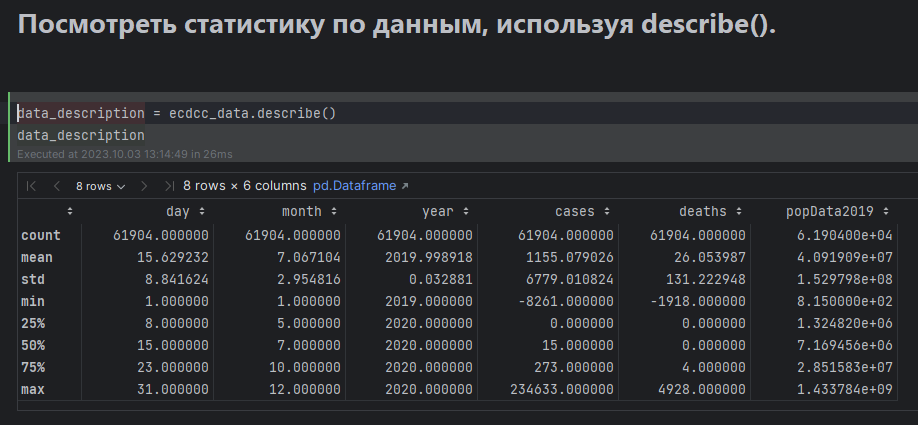


Рисунок 24 – Результат работы программы

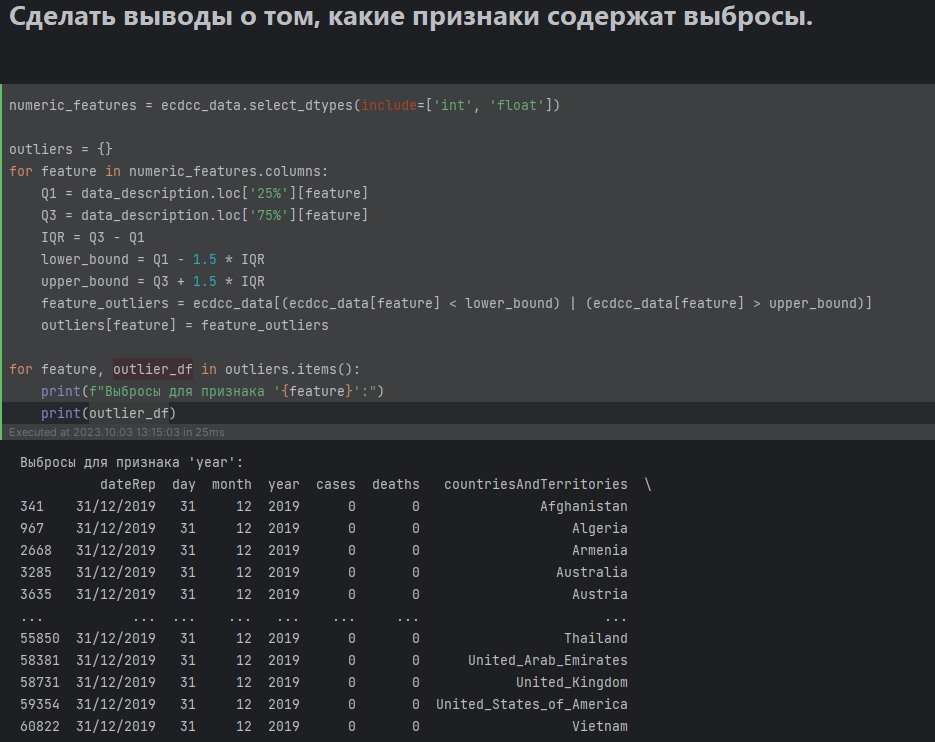


Рисунок 25 – Результат работы программы

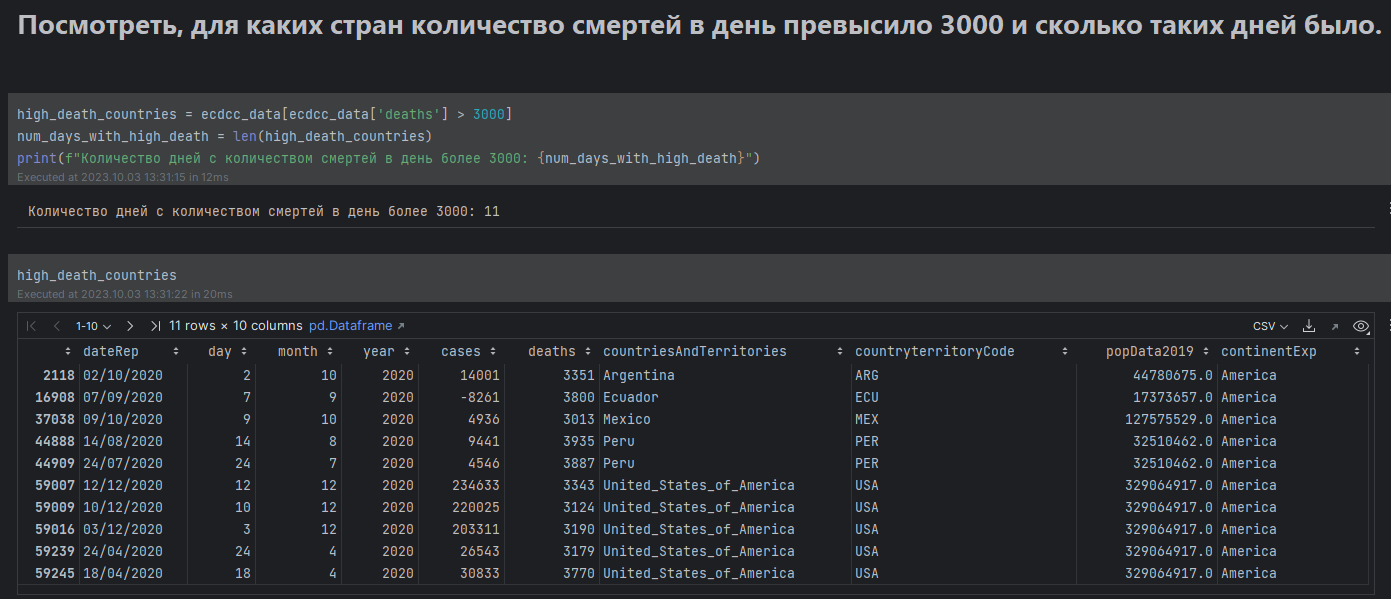


Рисунок 26 – Результат работы программы

Задание 12

Найти дублирование данных. Удалить дубликаты.

Реализация

Результат работы программы и код представлен на рисунке 27.

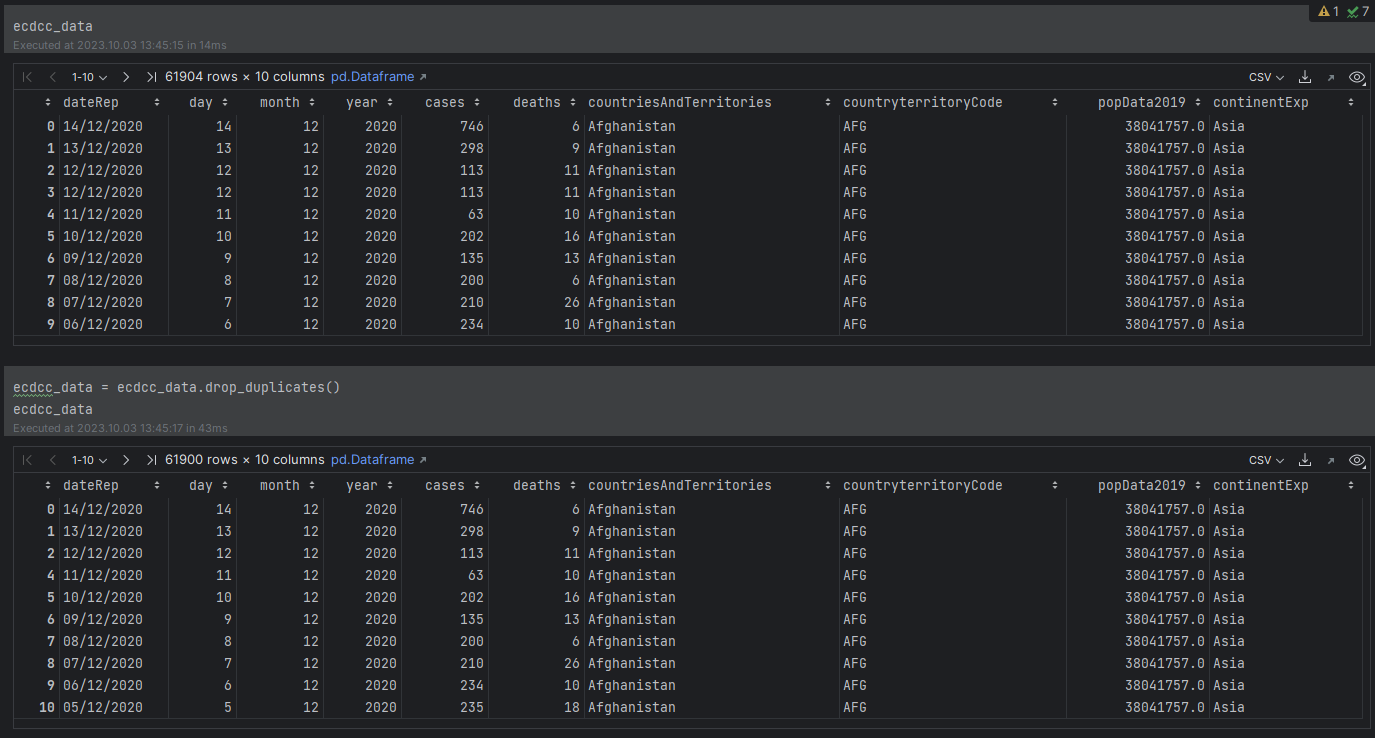


Рисунок 27 – Результат работы программы

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 Шопиро-Уилка) и на гомогенность дисперсии (критерий Бартлетта).

Реализация

Результат работы программы и код представлен на рисунках 28, 29, 30, 31, 32.



Рисунок 29 – Результат работы программы

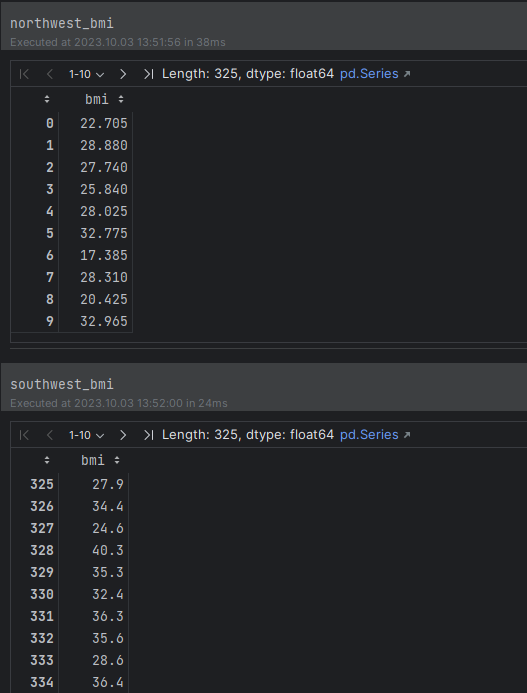


Рисунок 30 – Результат работы программы

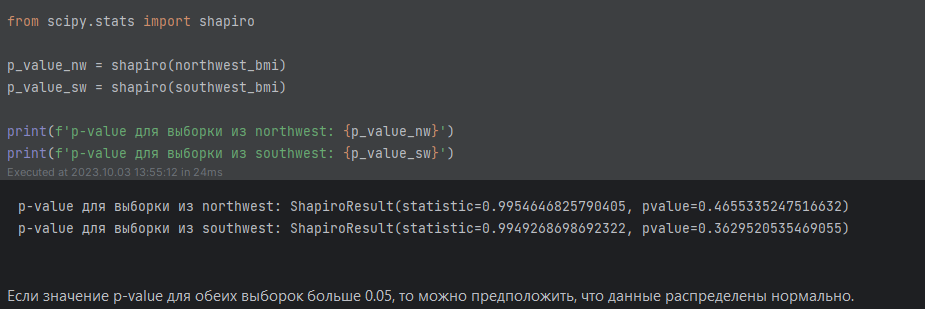


Рисунок 31 – Результат работы программы

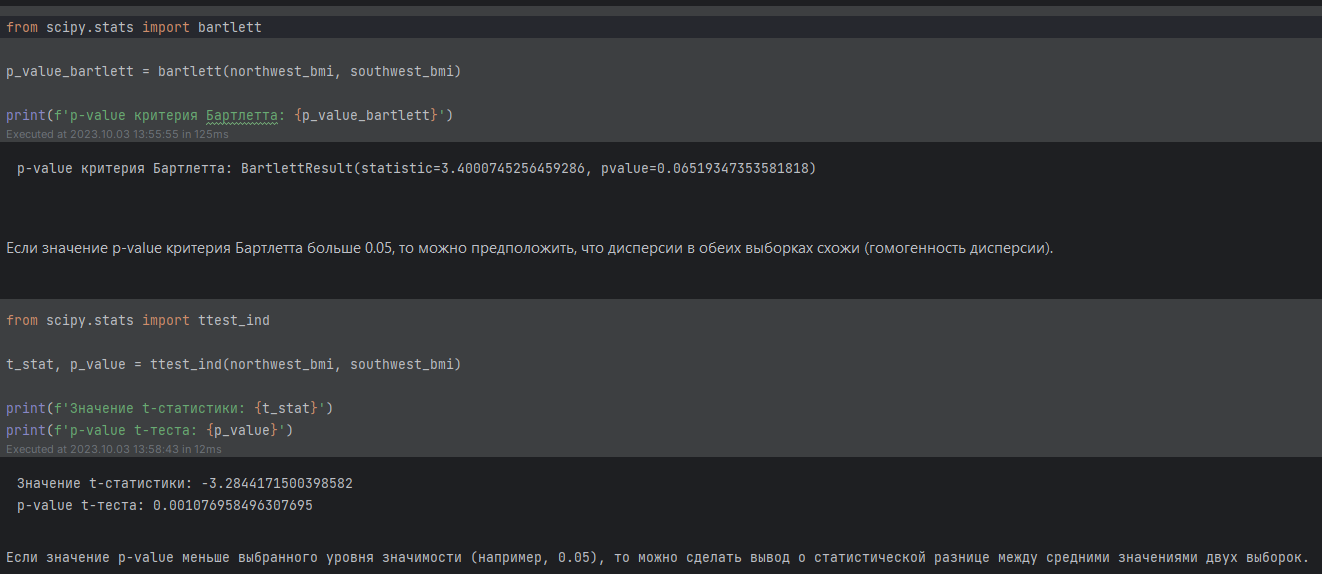


Рисунок 32 – Результат работы программы

Задание 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().

Реализация

Результат работы программы и код представлен на рисунке 33.

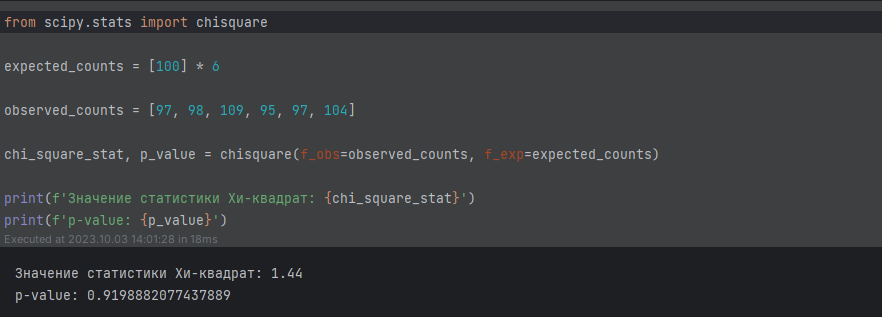


Рисунок 33 – Результат работы программы

Задание 15

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?

Реализация

Результат работы программы и код представлен на рисунке 34.

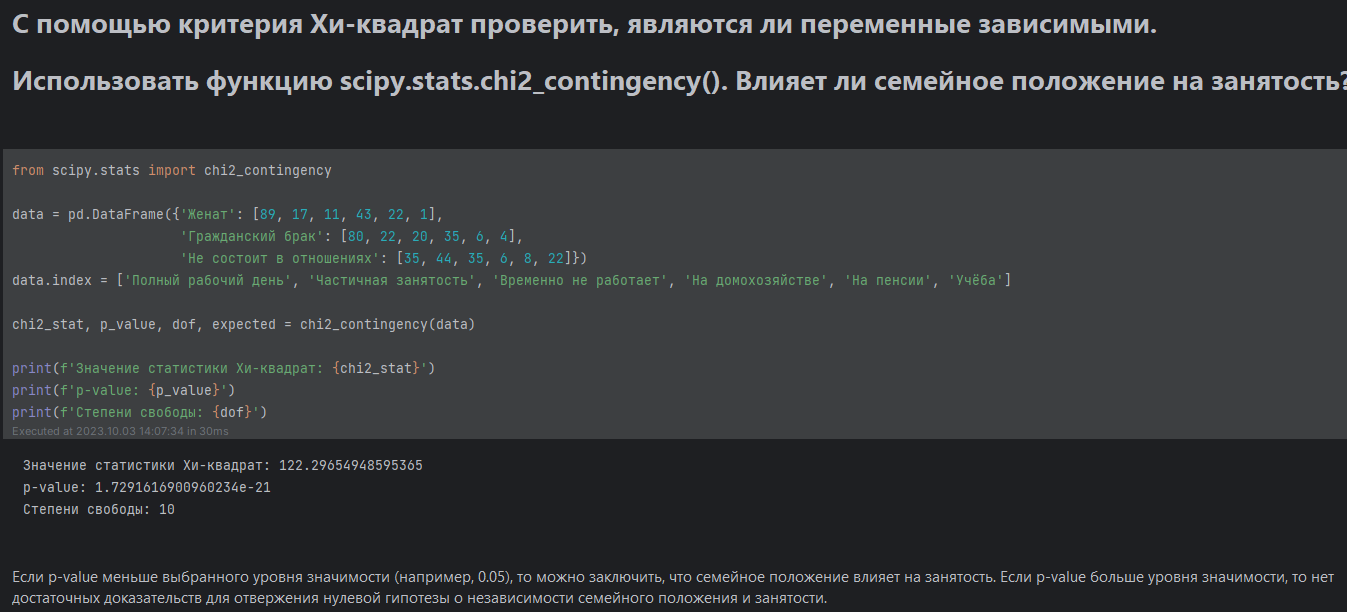


Рисунок 34 – Результат работы программы