|  |
| --- |
|  |
| МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ |
| 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 №6** | |
| **по дисциплине** |  |
| **«Технологии и инструментарий анализа больших данных»** | |
| Выполнил студент группы ИКБО-20-19 | Московка А.А. |
| Принял семинарист кафедры ИиППО | Десятников А.А. |

|  |  |  |
| --- | --- | --- |
| Практическая работы выполнены | « » 2022г. |  |
| «Зачтено» | « » 2022 г. |  |

Москва 2022

**Постановка задачи**

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

2. 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.

3. 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.

4. Найти дублирование данных. Удалить дубликаты.

5. 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).

6. 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().

7. 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? Оформить отчет о проделанной работе. Отчет должен содержать результаты визуализации для разных значений параметров и выводы.

**Программный код с комментариями и результаты его выполнения**

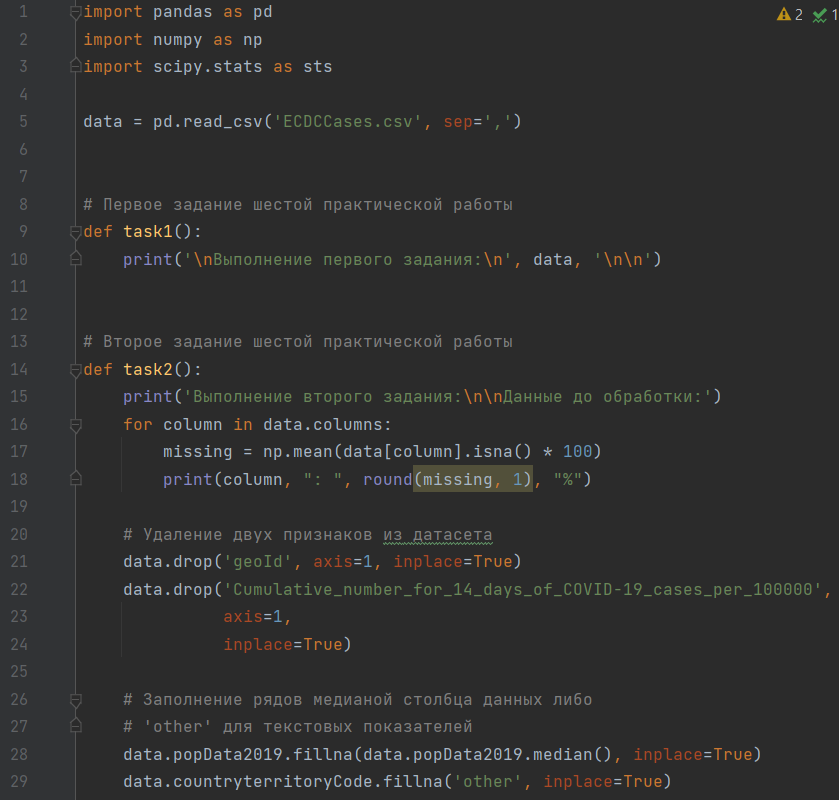


Рисунок 1 – Скриншот программного кода (Часть 1)



Рисунок 2 – Скриншот программного кода (Часть 2)

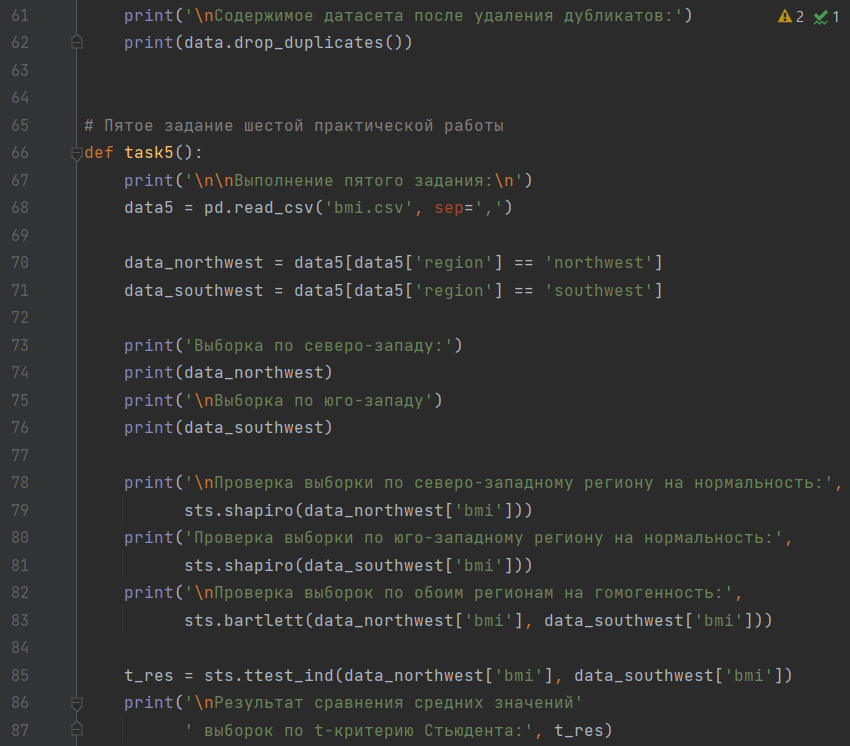


Рисунок 3 – Скриншот программного кода (Часть 3)

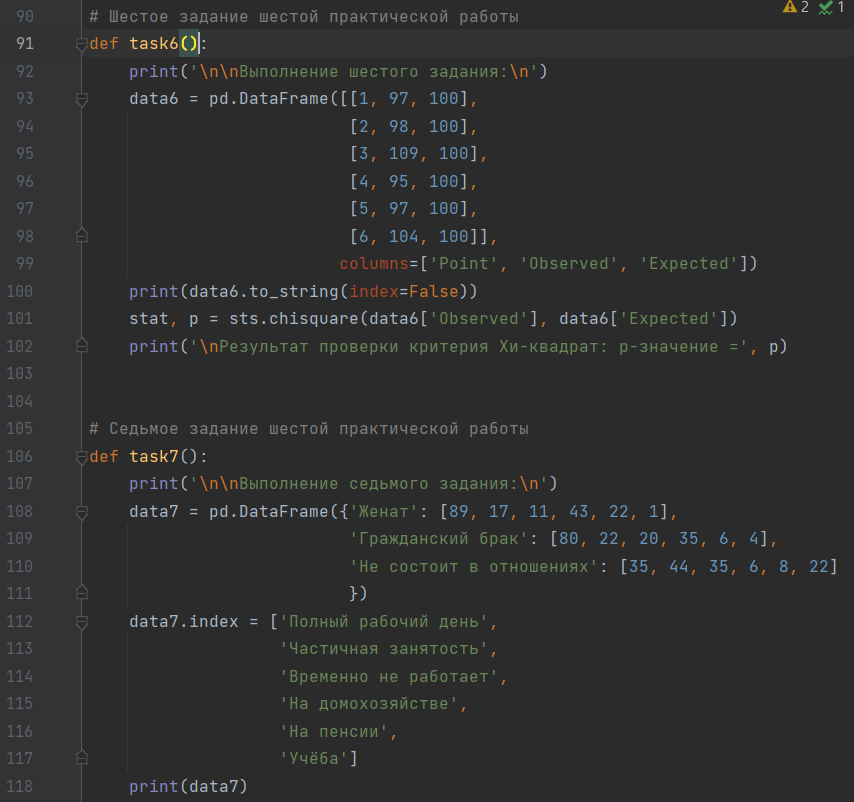


Рисунок 5 – Скриншот программного кода (Часть 4)

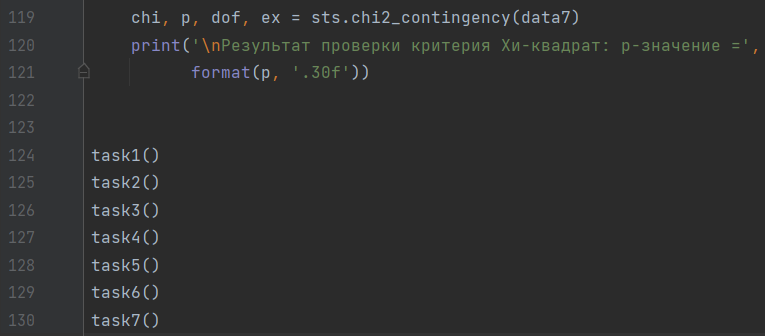


Рисунок 6 – Скриншот программного кода (Часть 5)

На следующих рисунках представлены скриншоты результатов выполнения программы со всеми заданиями с пояснениями (Рисунок 7-14).

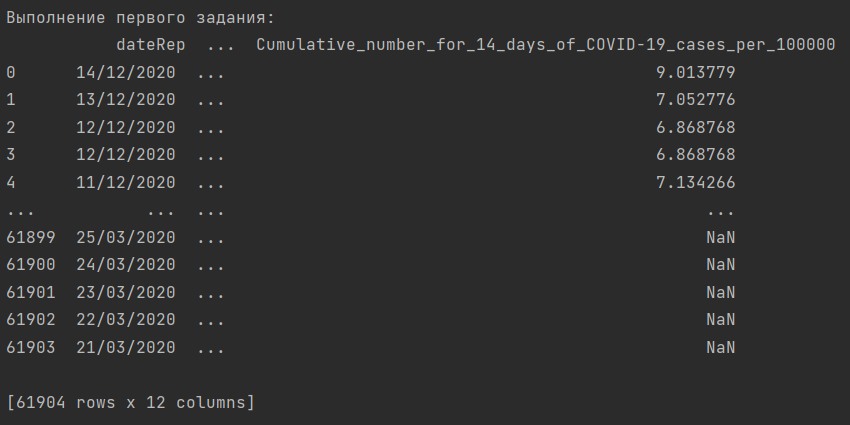


Рисунок 7 – Скриншот результата выполнения первого задания

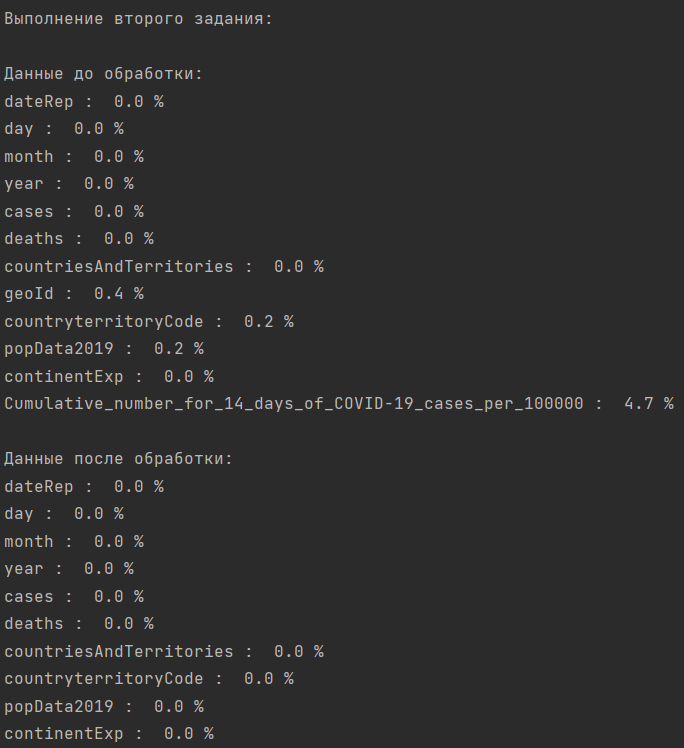


Рисунок 8 – Скриншот результата выполнения второго задания

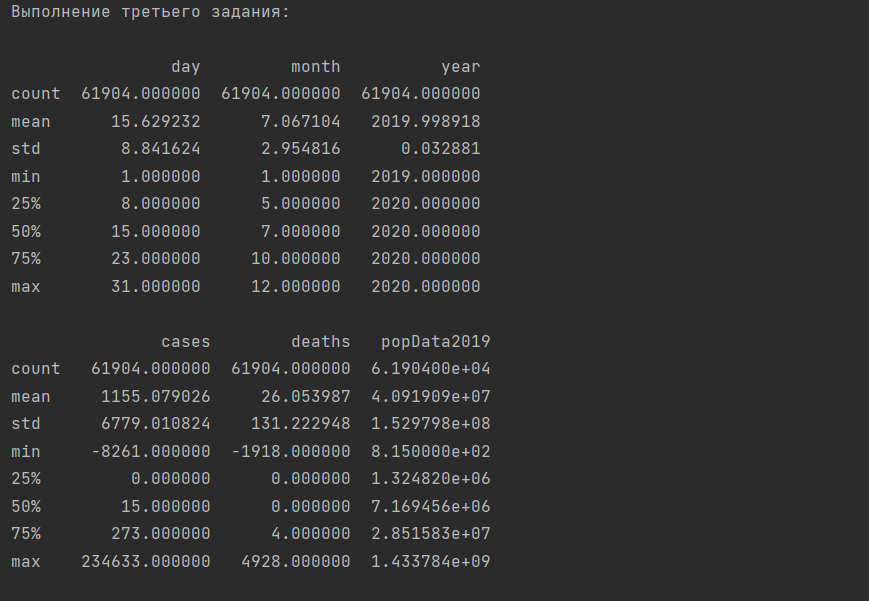


Рисунок 9 – Скриншот результата выполнения третьего задания (Часть 1)



Рисунок 10 – Скриншот результата выполнения третьего задания (Часть 2)

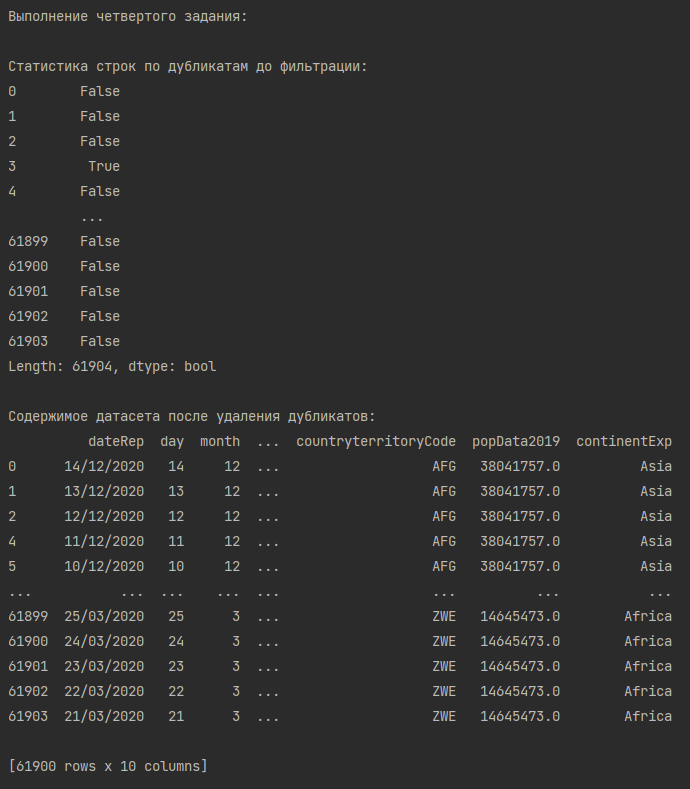


Рисунок 11 – Скриншот результата выполнения четвертого задания

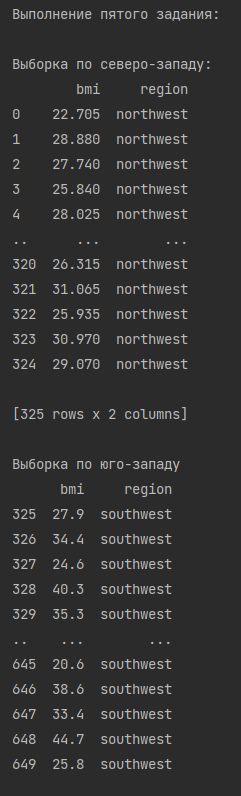


Рисунок 12 – Скриншот результата выполнения пятого задания (Часть 1)

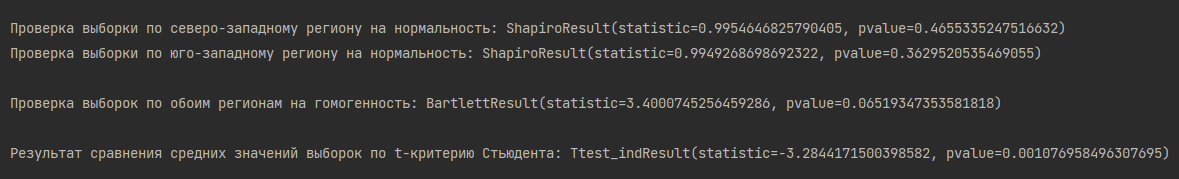


Рисунок 13 – Скриншот результата выполнения пятого задания (Часть 2)

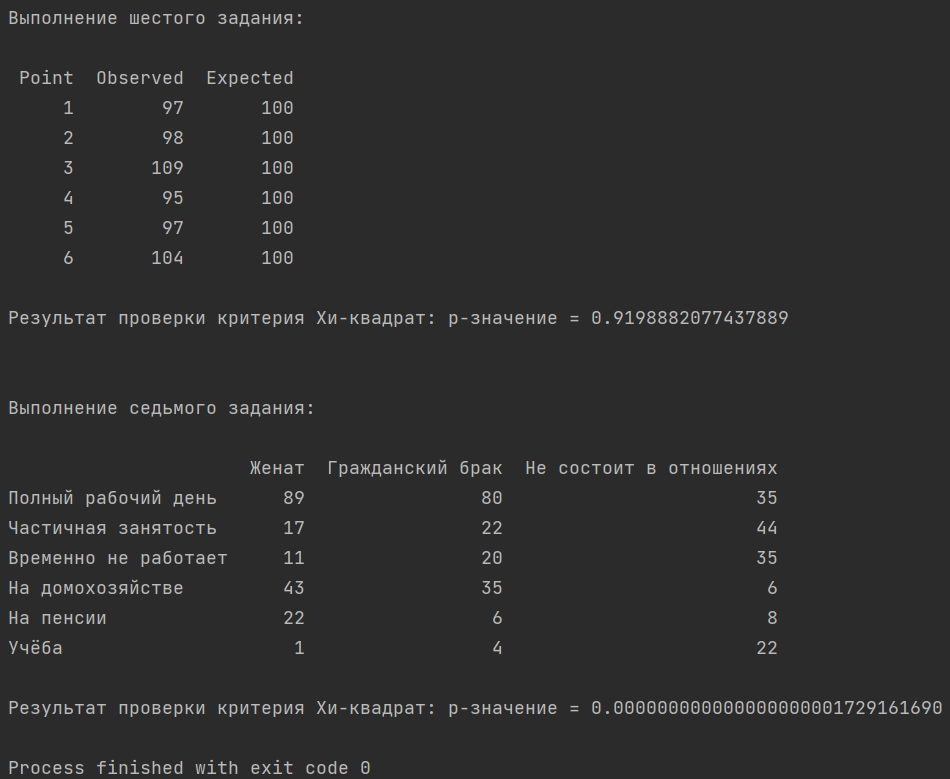


Рисунок 14 – Скриншот результата выполнения шестого и седьмого заданий

**Вывод**

В результате выполнения данной практической работы были проверены данные на наличие пропущенных значений, выведено кол-во пропущенных значений в процентах, удалено два признака, в которых больше всего пропущенных значений, просмотрена статистика данных с использованием метода describe(), найдены и удалены дублированные данные, взяты две выборки данных, проверены выборки на нормальность и гомогенность дисперсии, с помощью критерия Хи-квадрат проверено полученное распределение на равномерность, использована функция scipy.stats.chisquare(), проверены переменные на зависимость, а также создан датафрейм и использована функция scipy.stats.chi2\_contingency(), по результатам работы сделаны выводы и оформлен отчет.