|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  |  |  | | МИНОБРНАУКИ РОССИИ | | | | Федеральное государственное бюджетное образовательное учреждение  высшего образования  **«МИРЭА** – **Российский технологический университет»**  **РТУ МИРЭА** | | | | |
| Институт информационных технологий (ИТ) | |
| Кафедра прикладной математики | |

|  |  |
| --- | --- |
| **ОТЧЁТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ № 3** | |
| **по дисциплине «Технологии и инструментарий анализа больших данных»** | |
|  | |
| Выполнил студент группы ИКБО-20-21 | Фомичев Р.А. |
| Проверил ассистент кафедры ПМ ИИТ | Тетерин Н.Н. |

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

Код программы представлен на рисунке 1.



Рисунок 1 – Код программы

Результат работы программы представлен на рисунке 2.

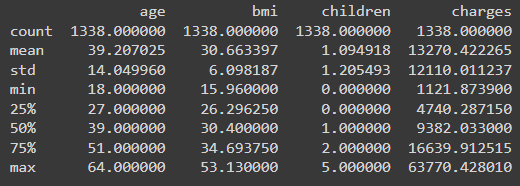


Рисунок 2 – Результат работы программы

Основываясь на полученных данных, следует следующий вывод, что средний возраст опрошенных 39 лет, средние выплаты составляют 13270 у.е., в среднем опрошенные имеют не более 2 детей.

2. Построить гистограммы для числовых показателей. Сделать выводы.

Код представлен на рисунке 3.

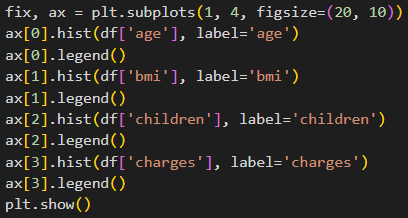


Рисунок 3 – Код программы

Результат работы программы представлен на рисунке 4.

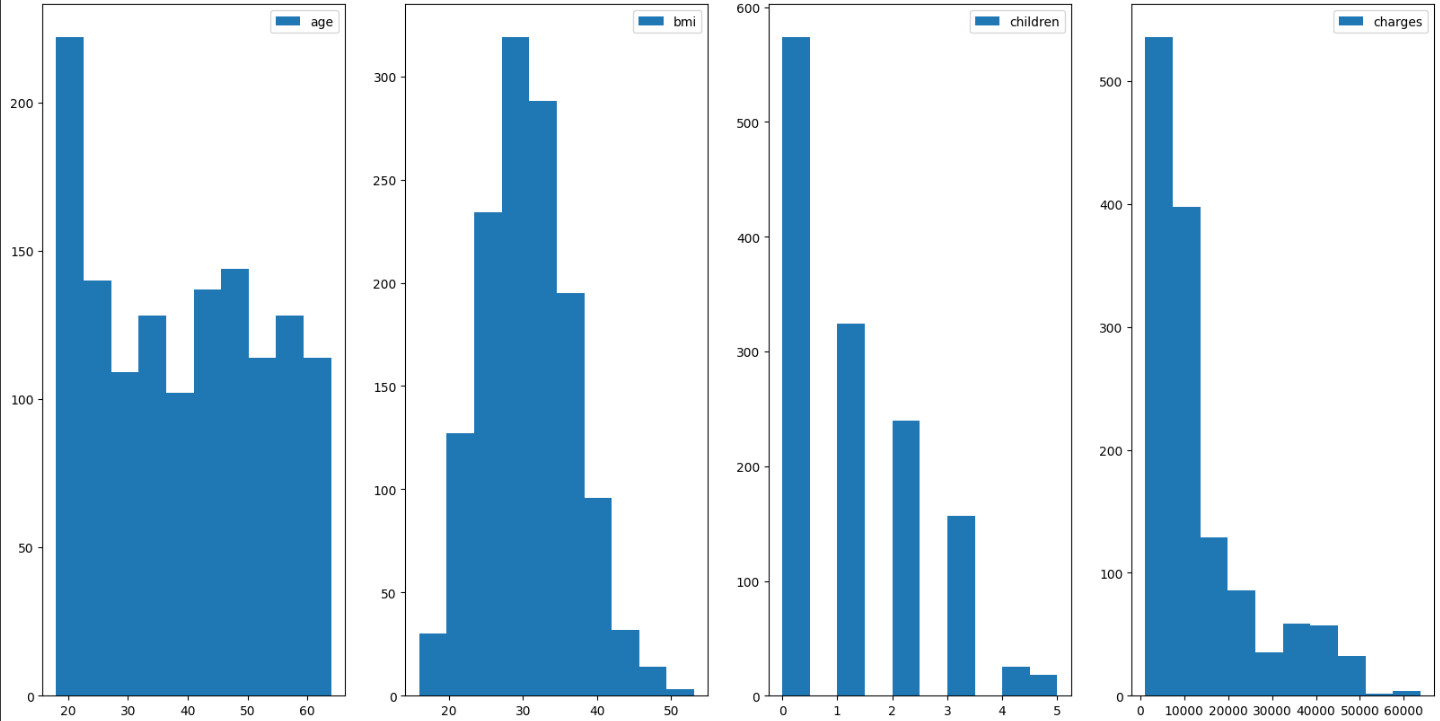


Рисунок 4 – Результат работы программы

Основываясь на полученных данных, следует следующий вывод, что больше всего опрошенных было возраста до 20 лет, имт больше всего от 25 до 30, больше всего людей из выборки без детей, а самая частая выплатой является до 10000 долларов.

3. 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.

Код программы представлен на рисунке 5.

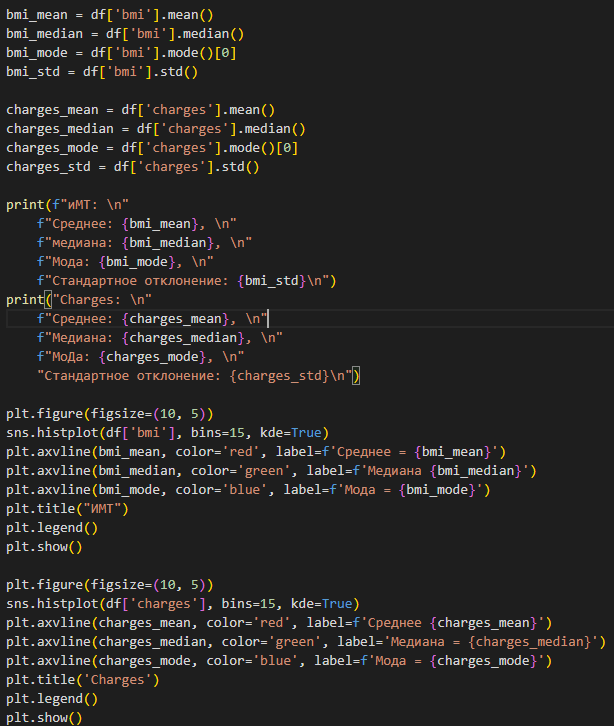


Рисунок 5 – Код программы

Результат представлен на рисунке 6.

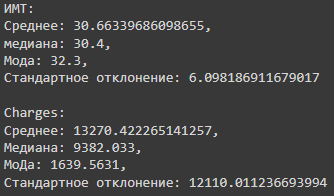


Рисунок 6 – Результат работы программы

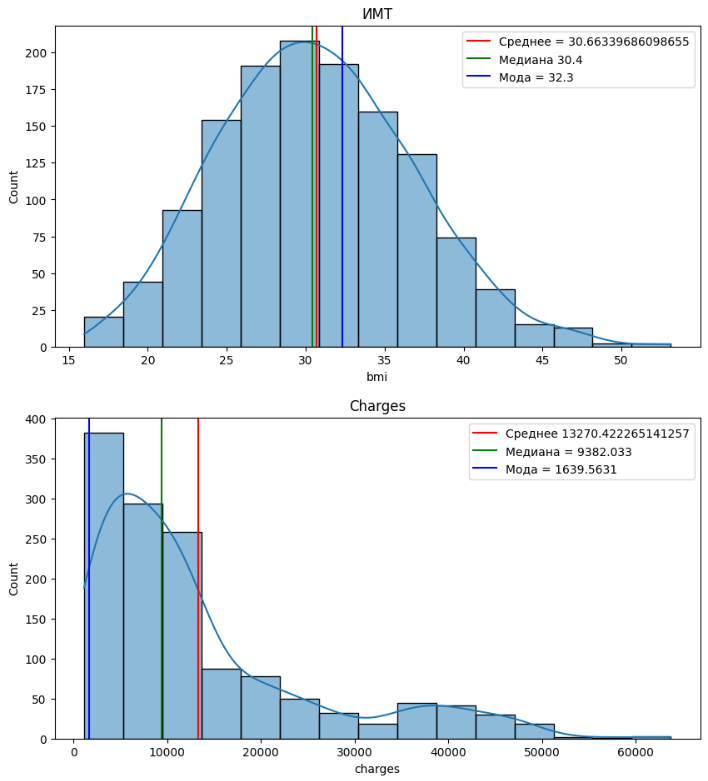


Рисунок 7 – Результат работы программы

Основываясь на полученных данных, следует следующий вывод, что средняя и медианная ИМТ приблизительно равны 30, стандартное отклонение приблизительно равно 6. Средняя и медианная значений выплат отличаются и равны приблизительно 13270 и 9382 соответственно, стандартное отклонение равно приблизительно 12110.

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

Код программы представлен на рисунке 8.

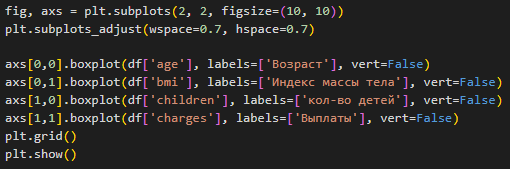


Рисунок 8 – Код программы

Результат работы программы представлен на рисунке 9.

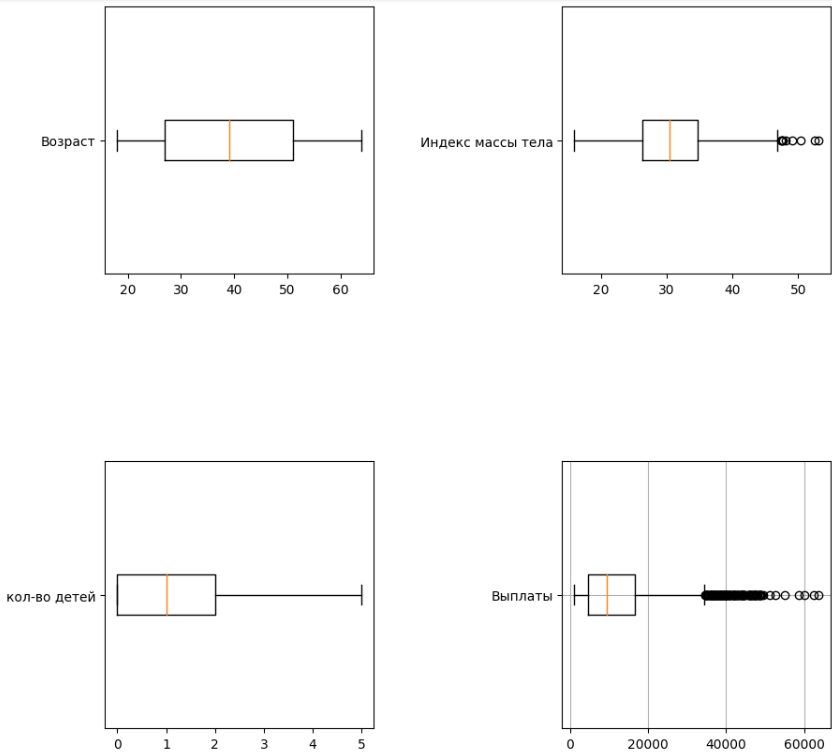


Рисунок 9 – Результат работы программы

Основываясь на полученных данных, следует следующий вывод, что медианный возраст приблизительно равен 40, в квартиль попадает возраст от 30 до 50 приблизительно. ИМТ приблизительно равен 32, в квартиль попадает ИМТ от 26 до 38, ИМТ более 47 является вбросами. Количество детей приблизительно равно 1, в квартиль попадает количество детей от 0 до 2. Средние выплаты приблизительно равны 10000 у.е, выплаты более 36000 являются вбросами.

5. 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.

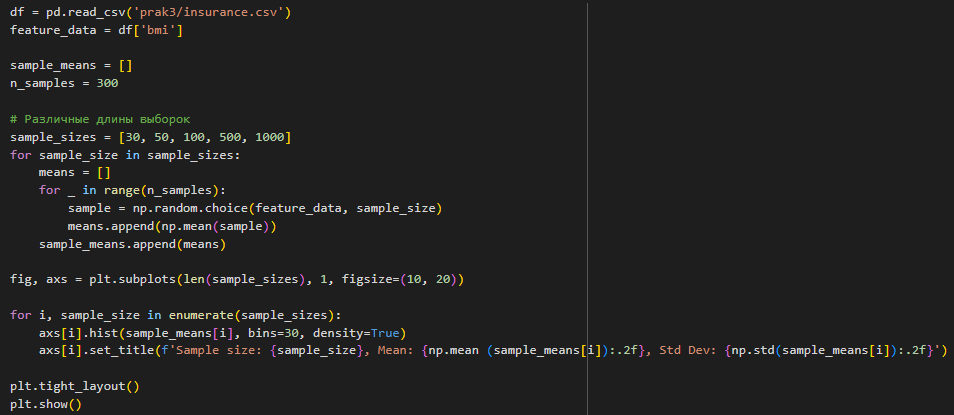


Рисунок 10 – Код программы

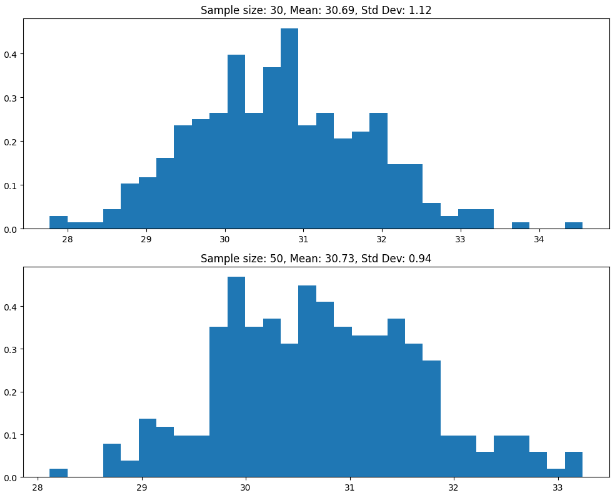


Рисунок 11 – Результат работы программы

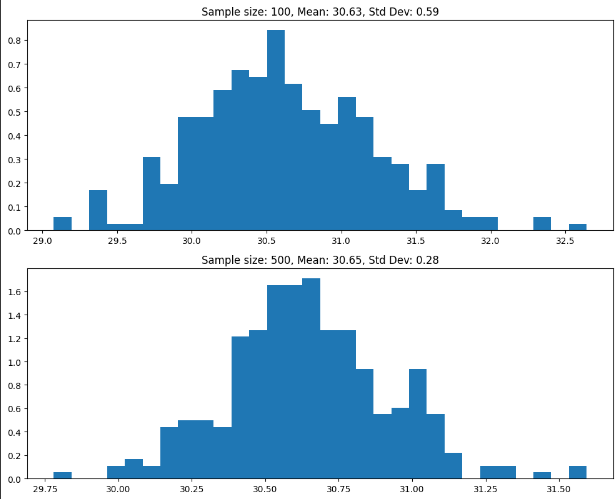


Рисунок 12 – Результат работы программы

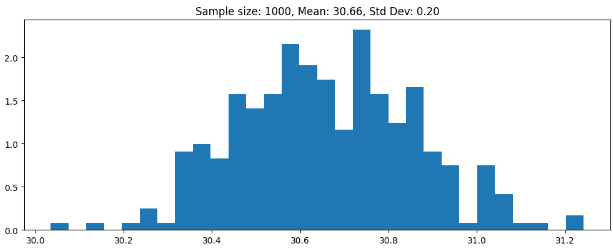


Рисунок 13 – Результат работы программы

Основываясь на полученных данных, следует следующий вывод, что при увеличении размера выборок стандартное отклонение уменьшается, а среднеквадратичное остается неизменным. Центральная предельная теорема выполняется, так как построенные графики близки к распределению по Гауссу.

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

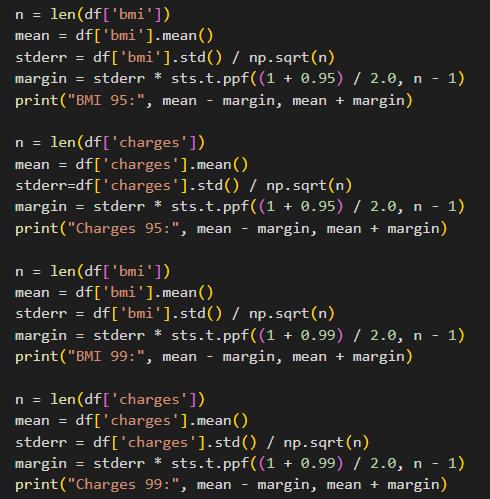


Рисунок 16 – Код программы

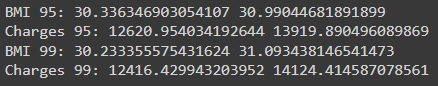


Рисунок 17 – Результат работы программы

7. 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.

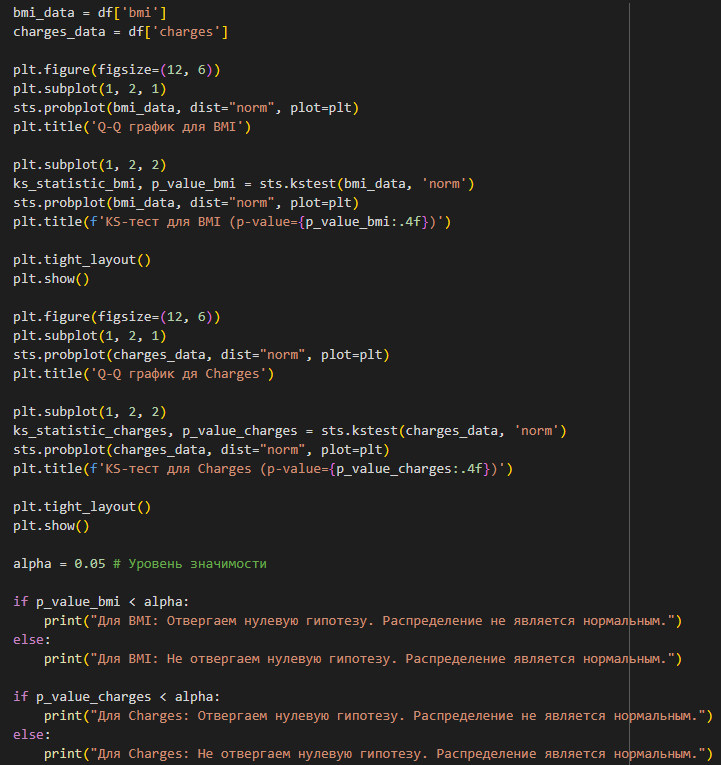


Рисунок 18 – Код программы

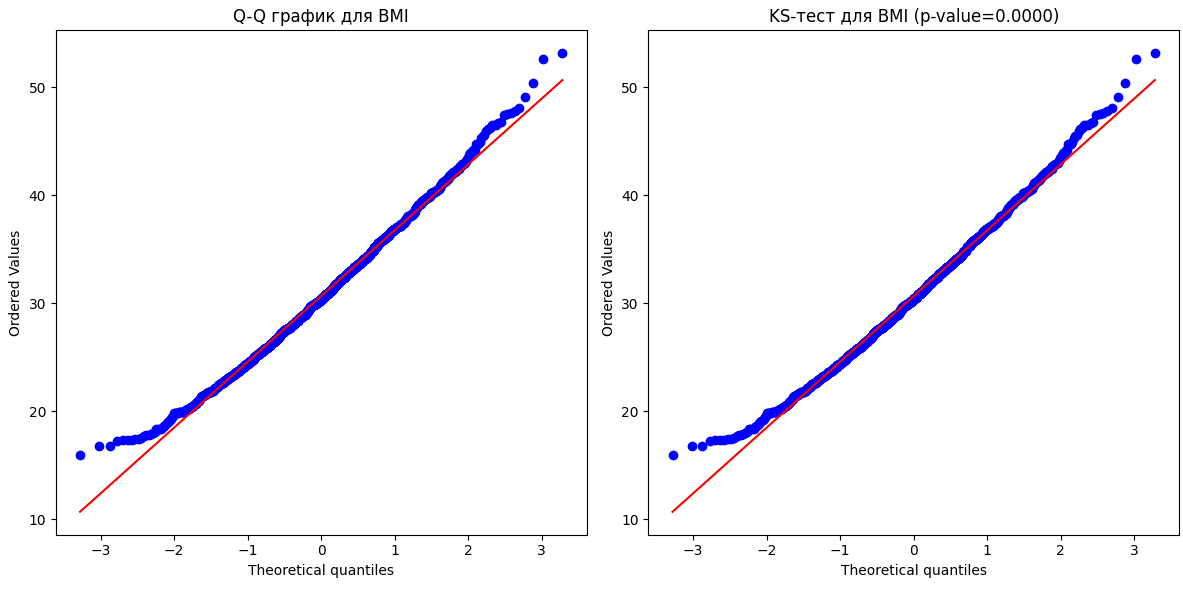


Рисунок 20 – Результат работы программы

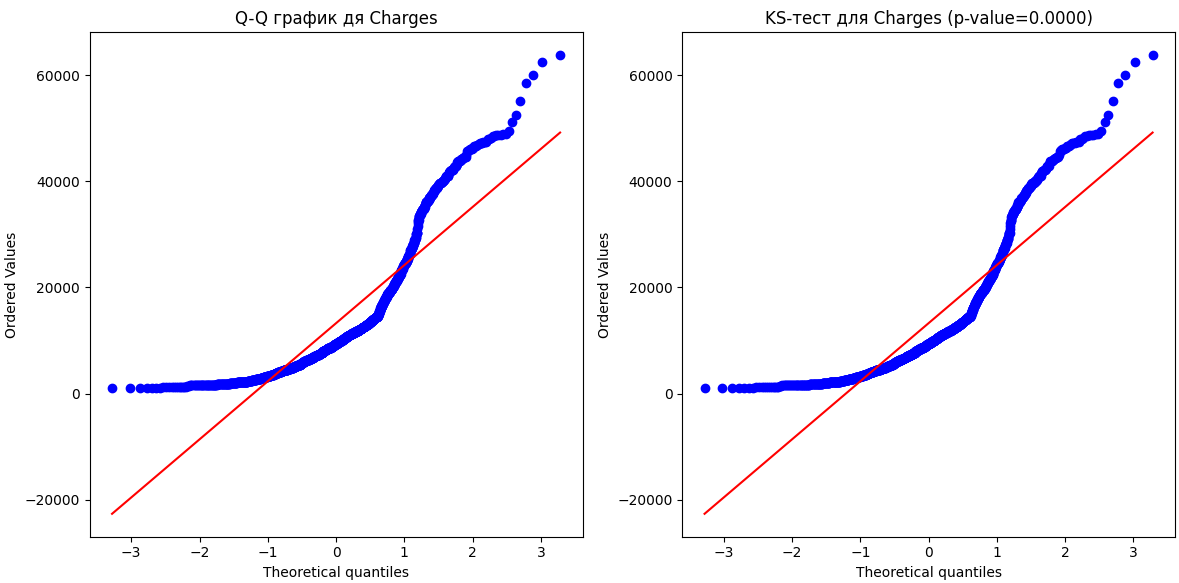


Рисунок 21 – Результат работы программы

Основываясь на полученных данных, следует следующий вывод, что распределение ИМТ соответствует нормальному распределению, а время выплаты не соответствуют.

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



Рисунок 22 – Код программы

9. 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.

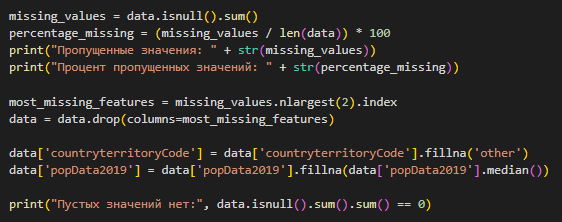


Рисунок 23 – Код программы

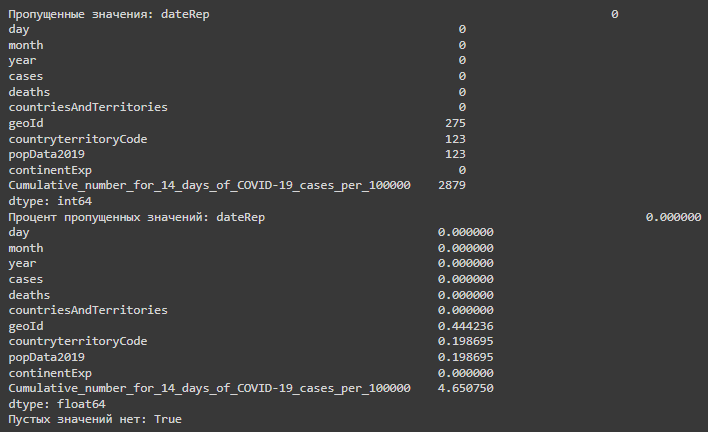


Рисунок 24 – Результат работы программы

10. 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.

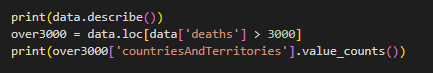


Рисунок 25 – Код программы

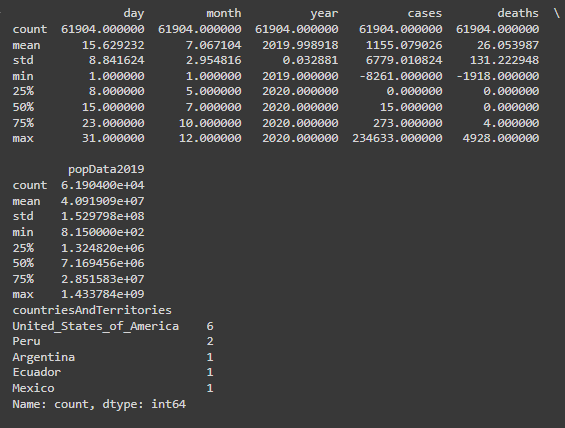


Рисунок 26 – Результат работы программы

Основываясь на полученных данных, следует следующий вывод, что среди данных есть вбросы, потому что существуют отрицательные значения у полей смертей и случаи.

11. Найти дублирование данных. Удалить дубликаты.

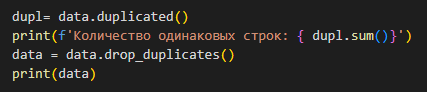


Рисунок 27 – Код программы

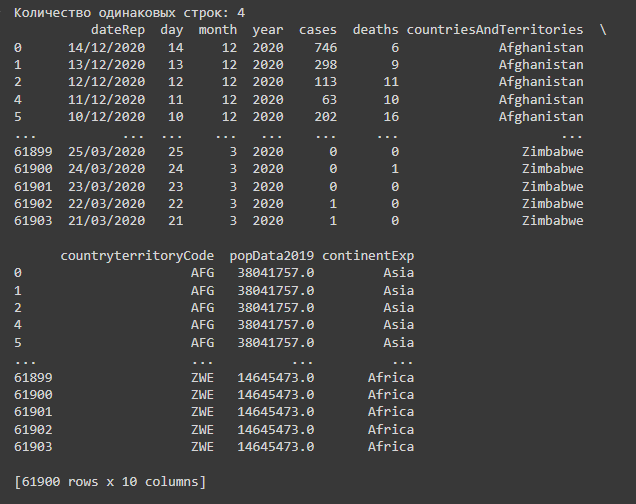


Рисунок 28 – Результат работы программы

12. 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).

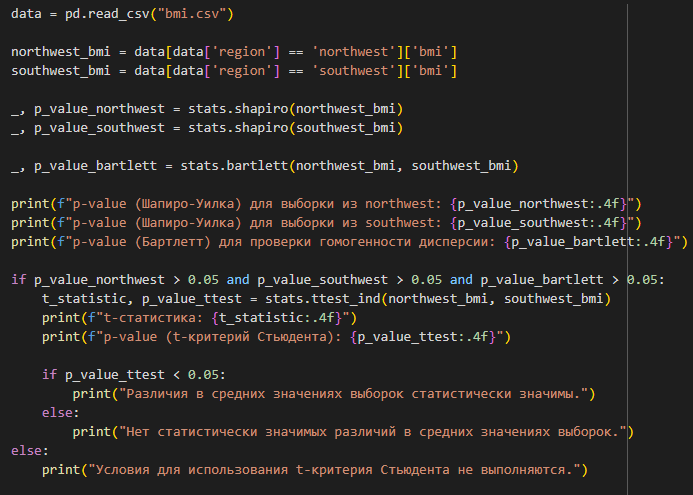


Рисунок 29 – Код программы

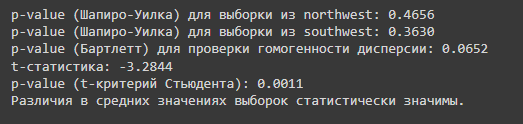


Рисунок 30 – Результат работы программы

13. 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().

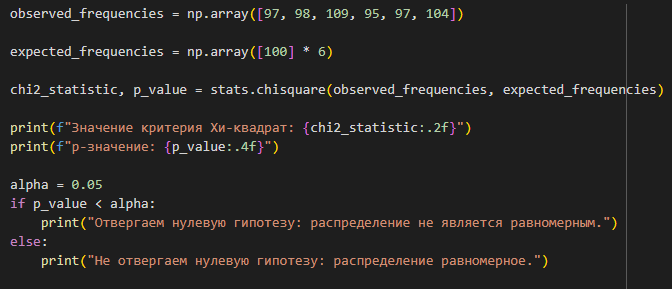


Рисунок 31 – Код программы



Рисунок 32 – Результат работы программы

14. 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?

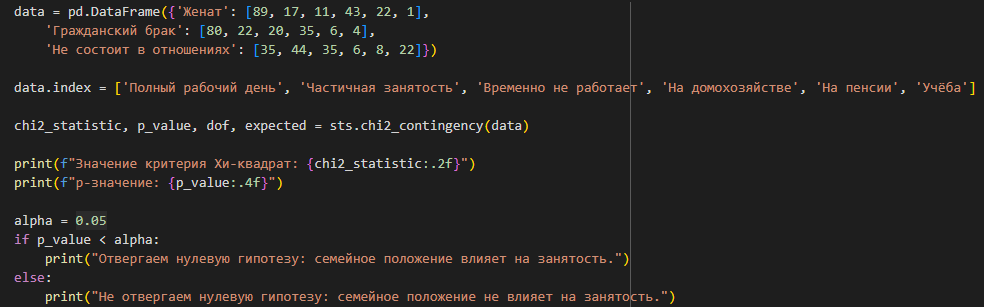


Рисунок 33 – Код программы



Рисунок 34 – Результат работы программы