|  |
| --- |
| 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 № 4**

|  |  |  |  |
| --- | --- | --- | --- |
| Выполнил студент группы ИНБО-01-17 | ИВБО-06-20, Чурилов Антон Константинович | (подпись) | |
| Принял |  | (подпись) | |
| Отчет представлен | «29» сентября 2022 г. | |  | |

Москва 2022 г.

**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.

1. 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 Шопиро-Уилка) 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().

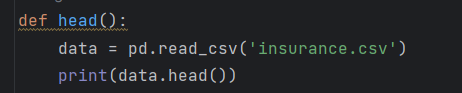
15. С помощью критерия Хи-квадрат проверить, являются ли переменные зависимыми. Создать датафрейм, используя следующий код: data = pd.DataFrame({'Женат': [89,17,11,43,22,1], 'Гражданский брак': [80,22,20,35,6,4], 'Не состоит в отношениях': [35,44,35,6,8,22]}) data.index = ['Полный рабочий день','Частичная 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.

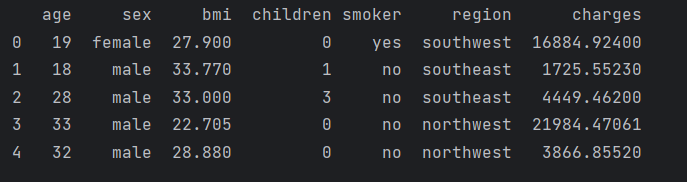
**Ход работы**

**Задание 1**

Загрузить данные из файла “insurance.csv”.

****

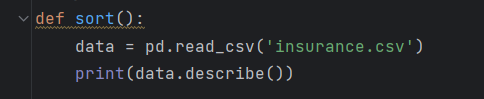
**Рисунок 1 – Выгрузка датафрейма из csv**

****

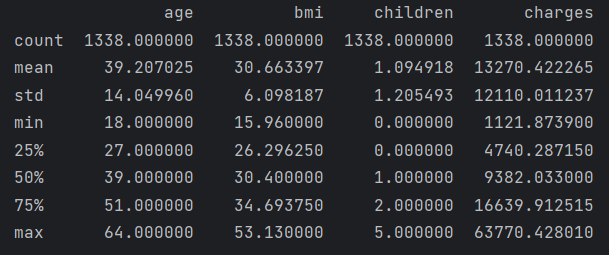
**Рисунок 2 – Задание 1**

**Задание 2**

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

****

**Рисунок 3 – Задание 2**

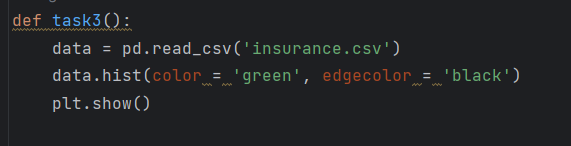
****

**Рисунок 4 – Вывод**

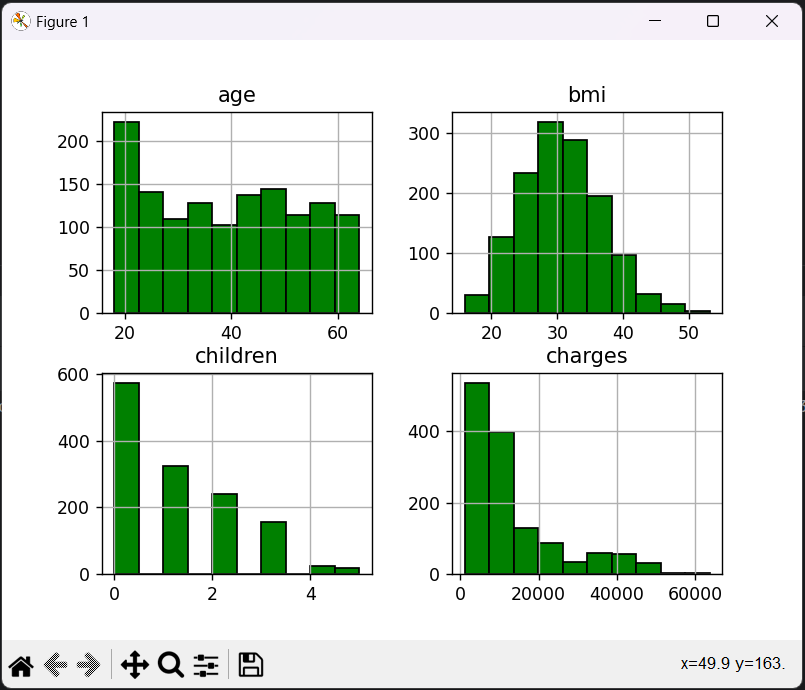
В приведённых данных представлена статистика по столбцам age, bmi, children и charges, а именно – count (количество записей), mean (среднее значение), std (стандартное отклонение), min (минимальное значение), max (максимальное значение), 25% (1 квартиль – 25% генеральной совокупности меньше указанных значений), 50% (2 квартиль – 50% генеральной совокупности меньше указанных значений), 75% (3 квартиль – 75% генеральной совокупности меньше указанных значений).

**Задание 3**

Построить гистограммы для числовых показателей. Сделать выводы.



**Рисунок 5 – Код задания 3**

****

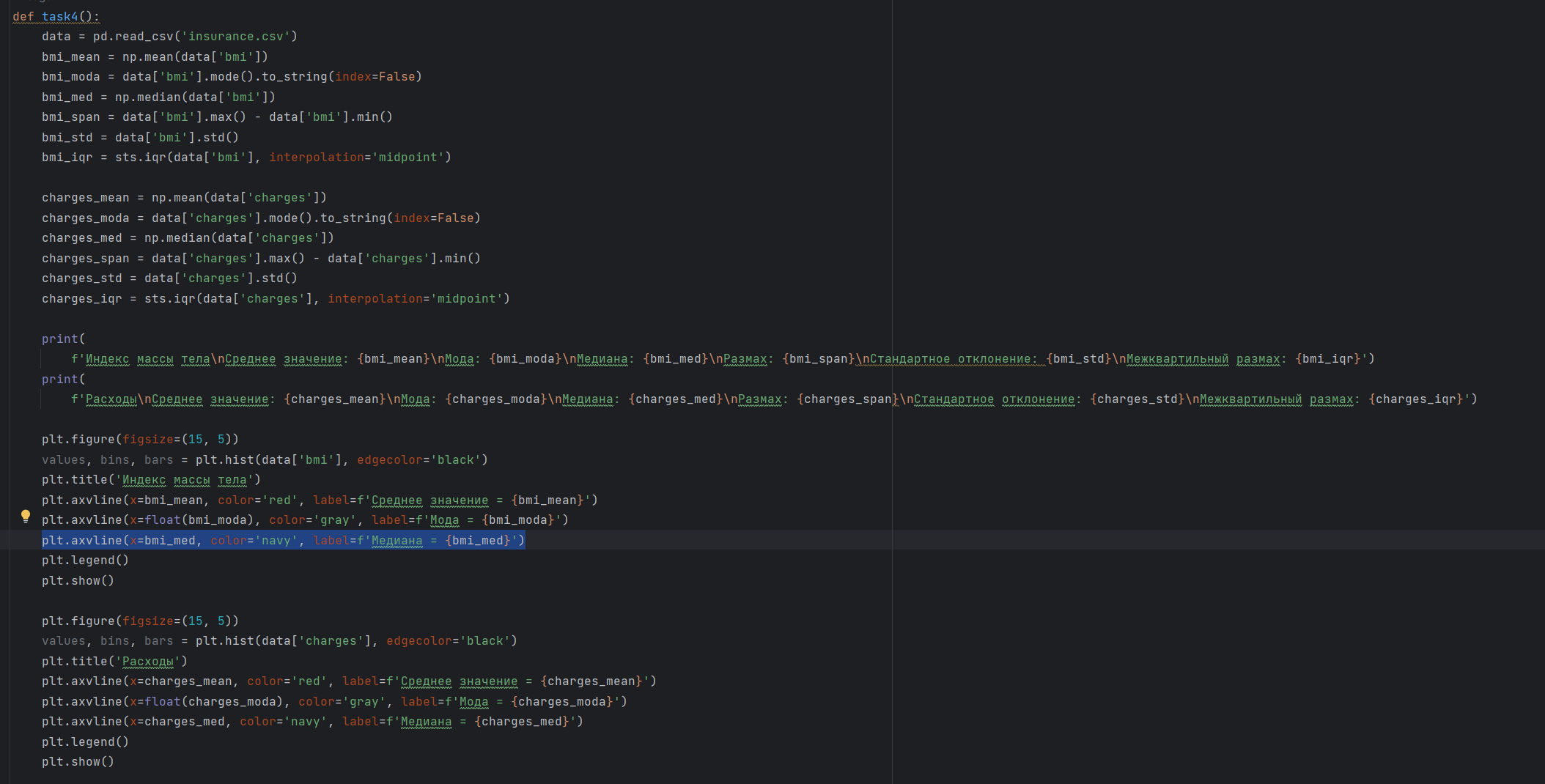
**Рисунок 6 – Графики**

Исходя из графиков, можно сделать вывод, что в приведённых данных преобладают записи:

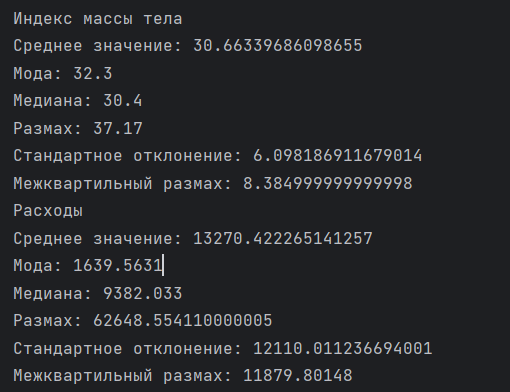
* со значением возраста 20 лет
* со значением индекса массы тела близким к 30
* с нулевым значением количества детей
* со значением расходов меньше 20 тысяч

**Задание 4**

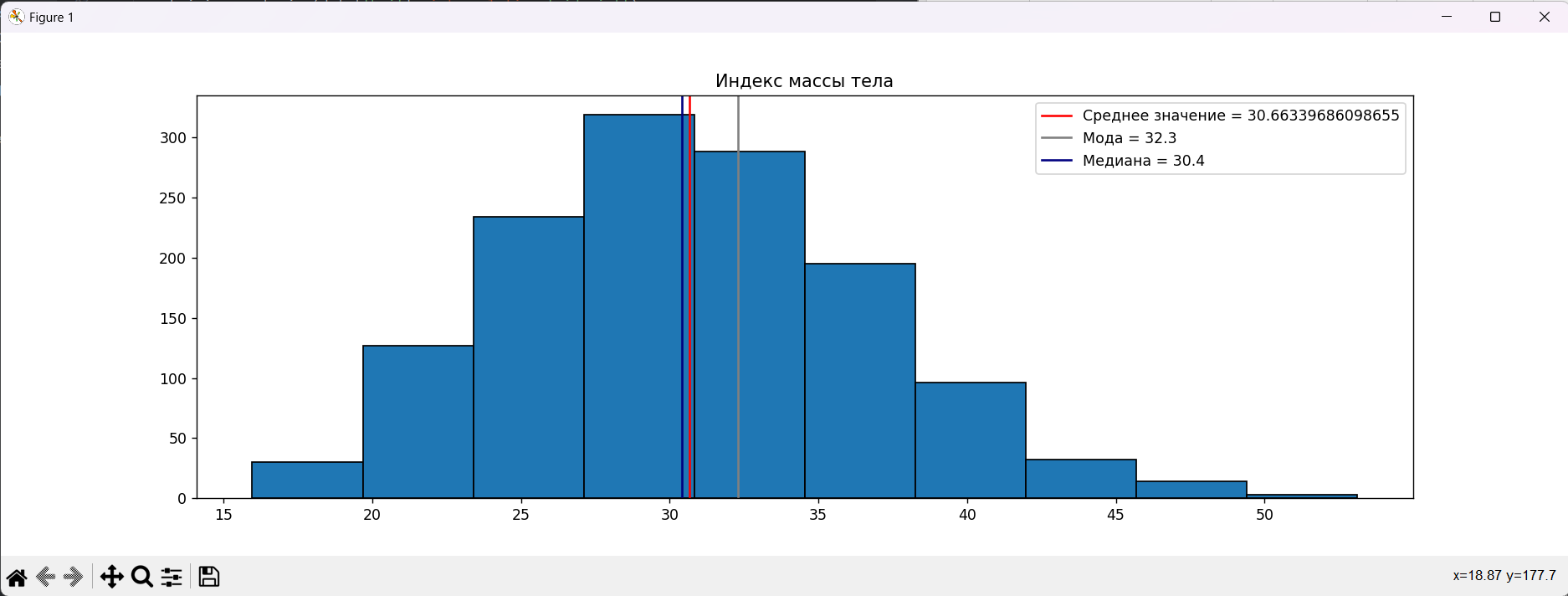
Найти меры центральной тенденции и меры разброса для индекса массы тела (bmi) и расходов (charges). Отобразить результаты в виде текста и на гистограммах (3 вертикальные линии). Добавить легенду на графики. Сделать выводы.

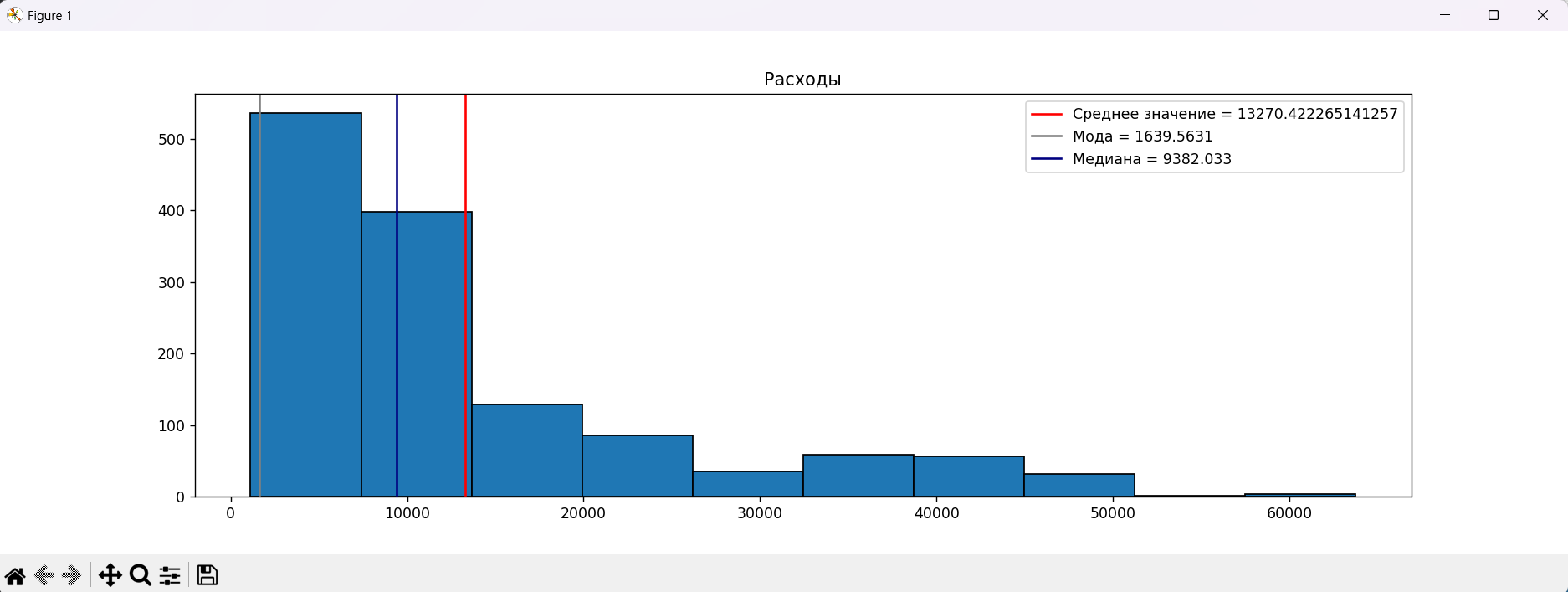


**Рисунок 7 – Код задания 4**



**Рисунок 8 – Результаты в виде текста**



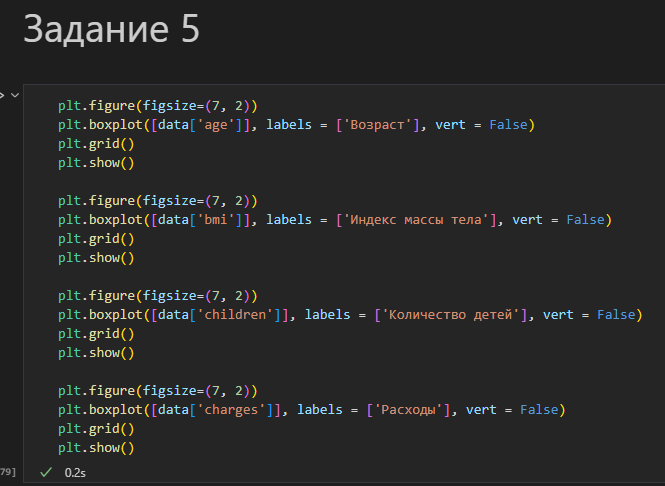


**Рисунок 9– Результаты в виде гистограмм**

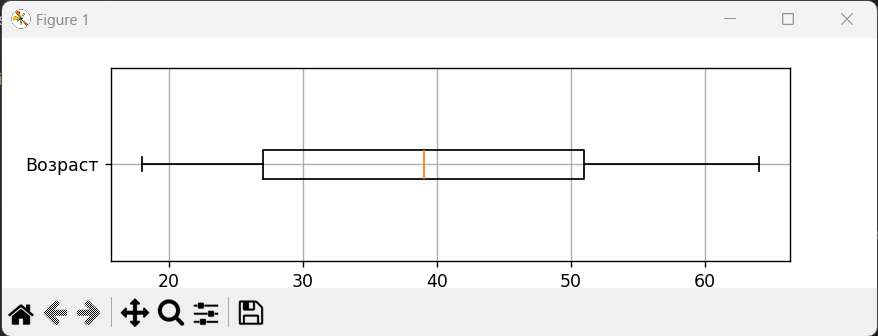
Исходя из полученных значений можно сделать выводы, что для индекса массы тела – среднее значение: ~30.66, мода: 32.3, медиана: 30.4, размах: 37.17, стандартное отклонение: ~6.1, межквартильный размах: ~8.38, а для расходов – среднее значение: ~13270.42, мода: ~1639.56, медиана: ~9382.03, размах: ~62648.55, стандартное отклонение: ~12110.011, межквартильный размах: ~11879.80.

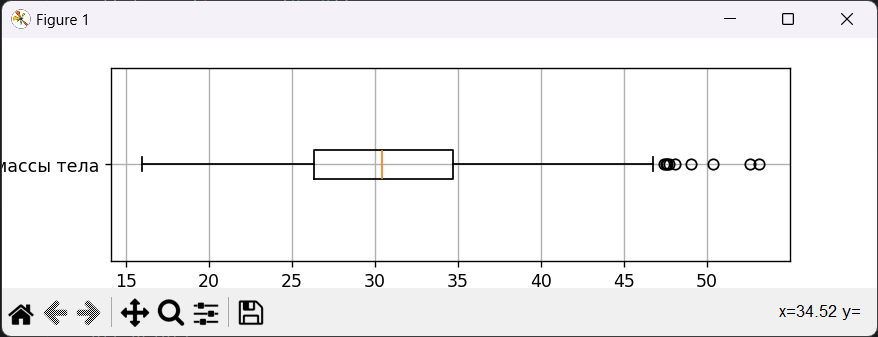
**Задание 5**

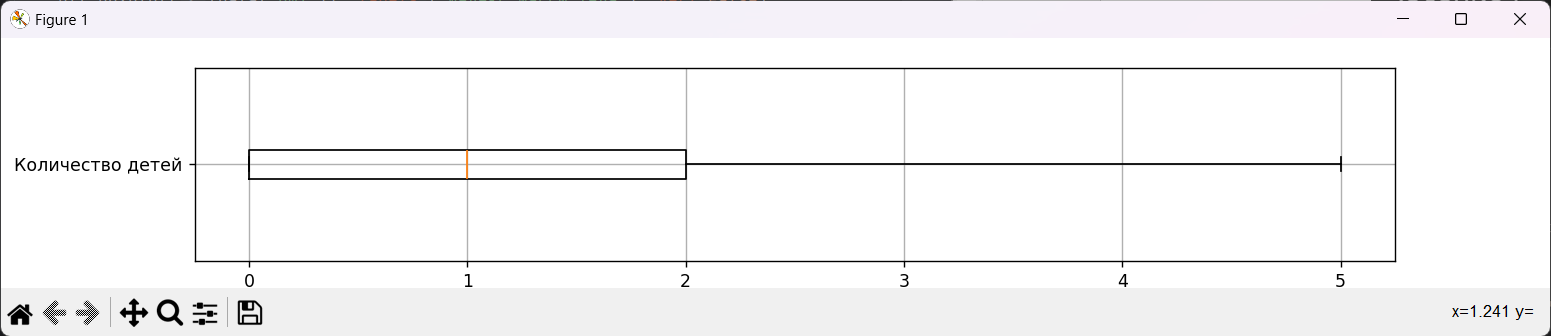
Построить box-plot для числовых показателей. Названия графиков должны соответствовать названиям признаков. Сделать выводы.

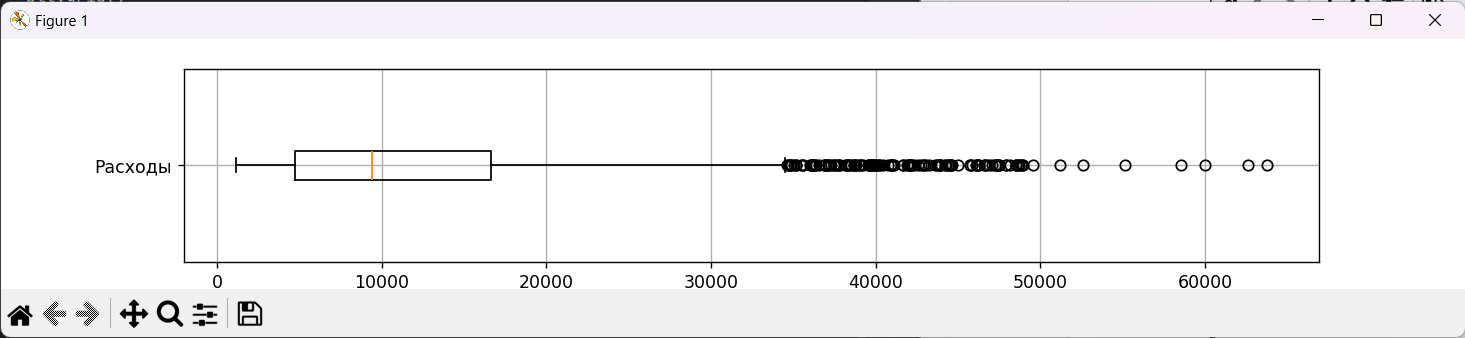


**Рисунок 10 – Код задания 5**









**Рисунок 11 – Полученные графики задания 5**

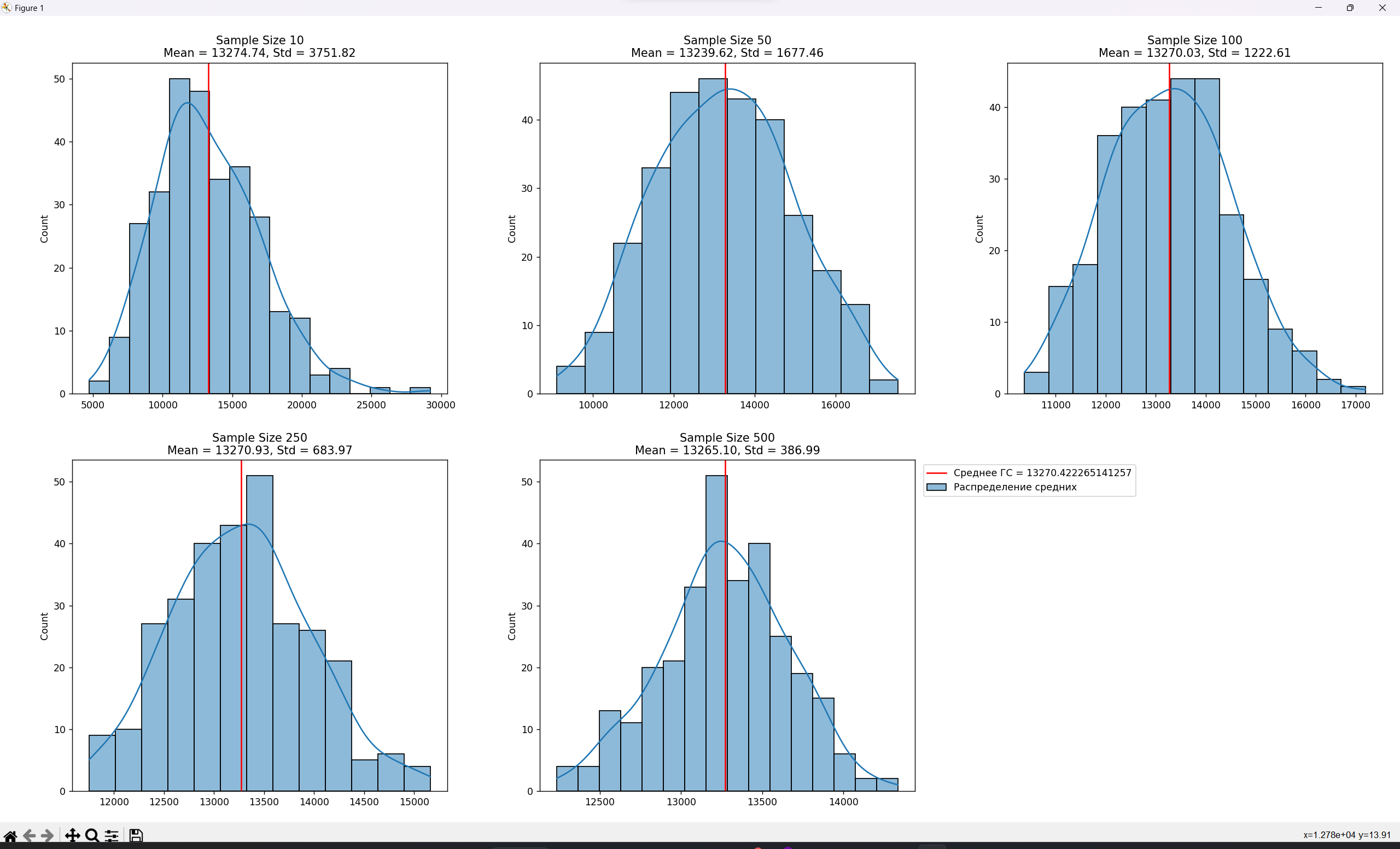
Исходя из графиков, можно сделать вывод, что наборы данных индекса массы тела и расходов имеют выбросы, а также, что выборка индекса массы тела близка к нормальному распределению.

**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.



**Рисунок 12 – Код задания 6**

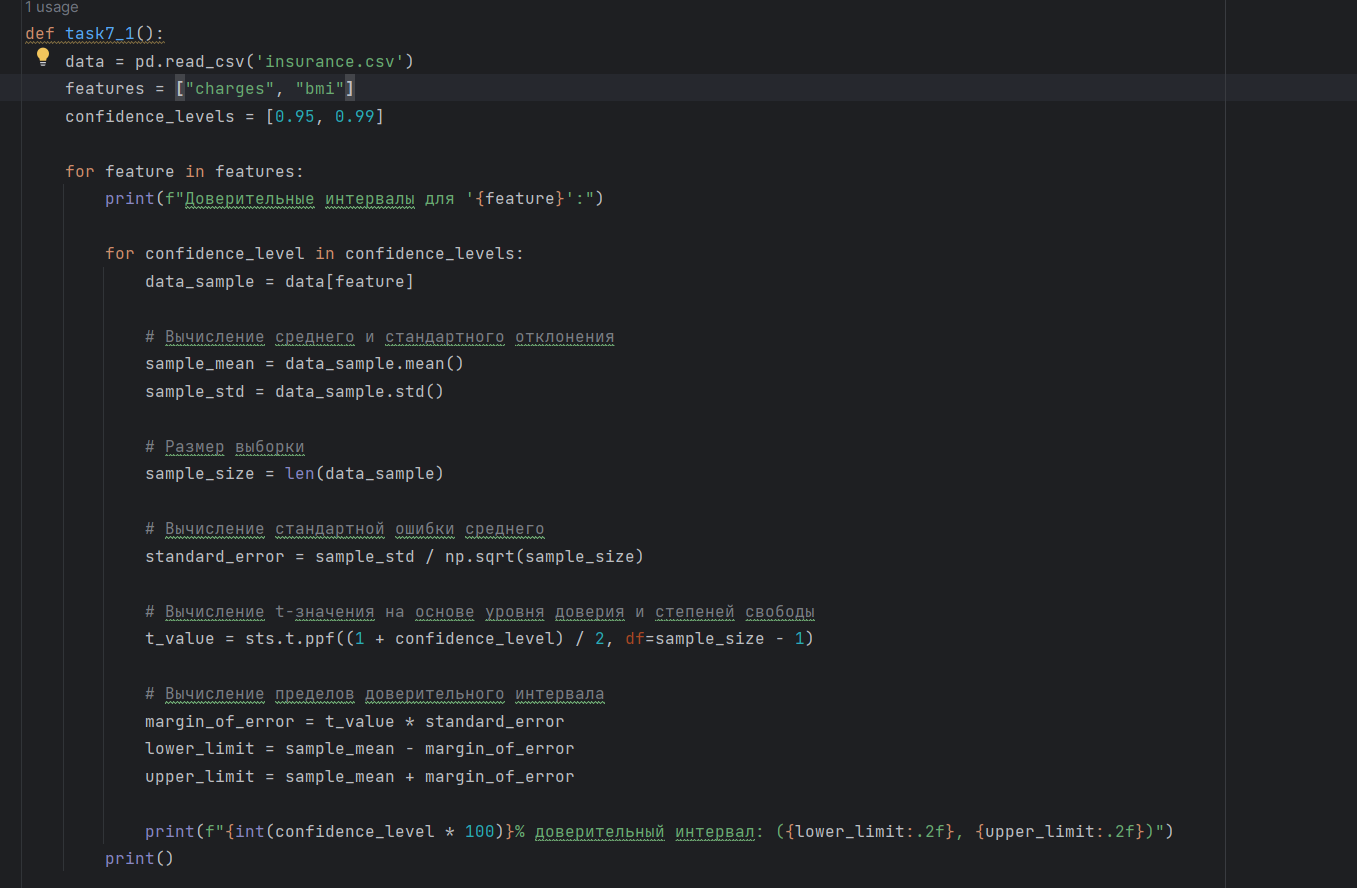
****

**Рисунок 13 – Графики для выборок**

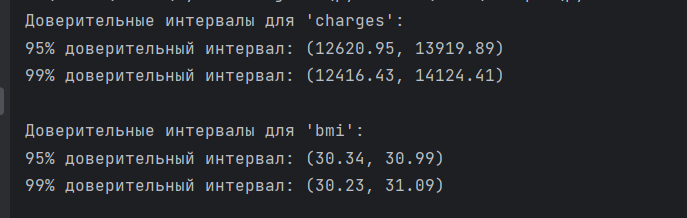
Исходя из графиков, можно сделать вывод, что полученные выборочные средние распределены нормально.

**Задание 7**

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



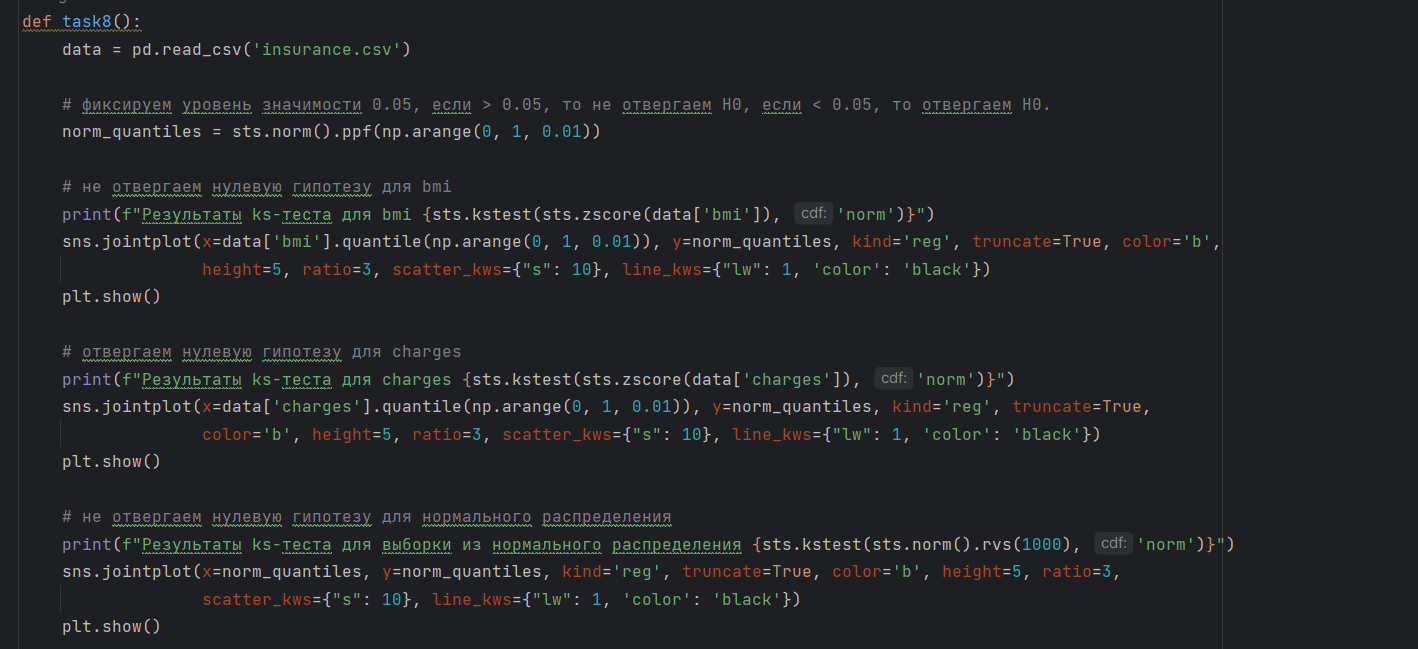
**Рисунок 14 – Задание 7**

****

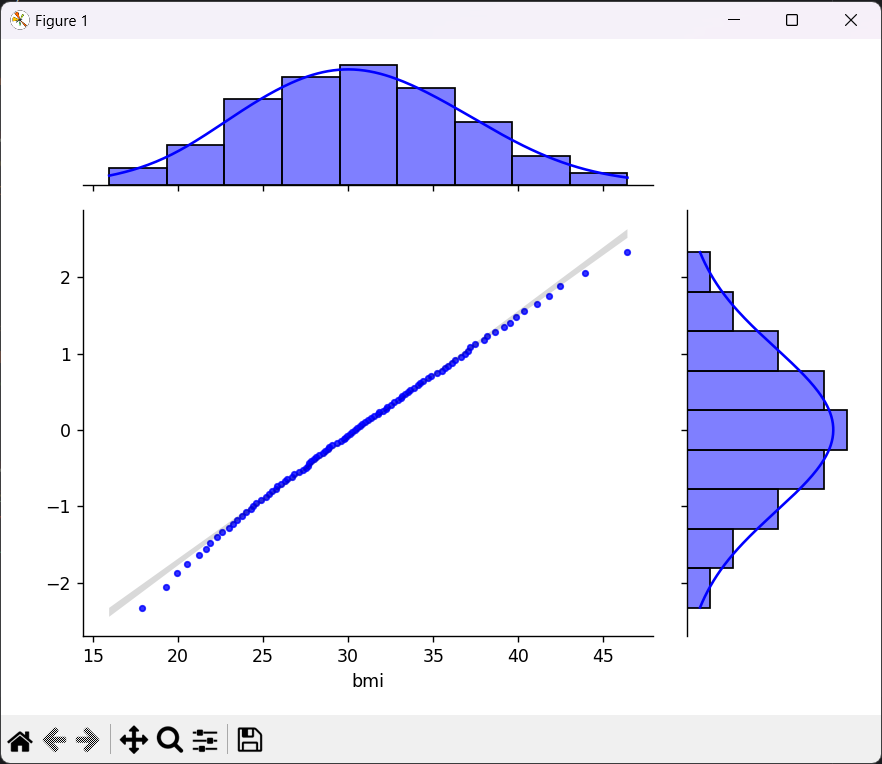
**Рисунок 15 – Результат**

**Задание 8**

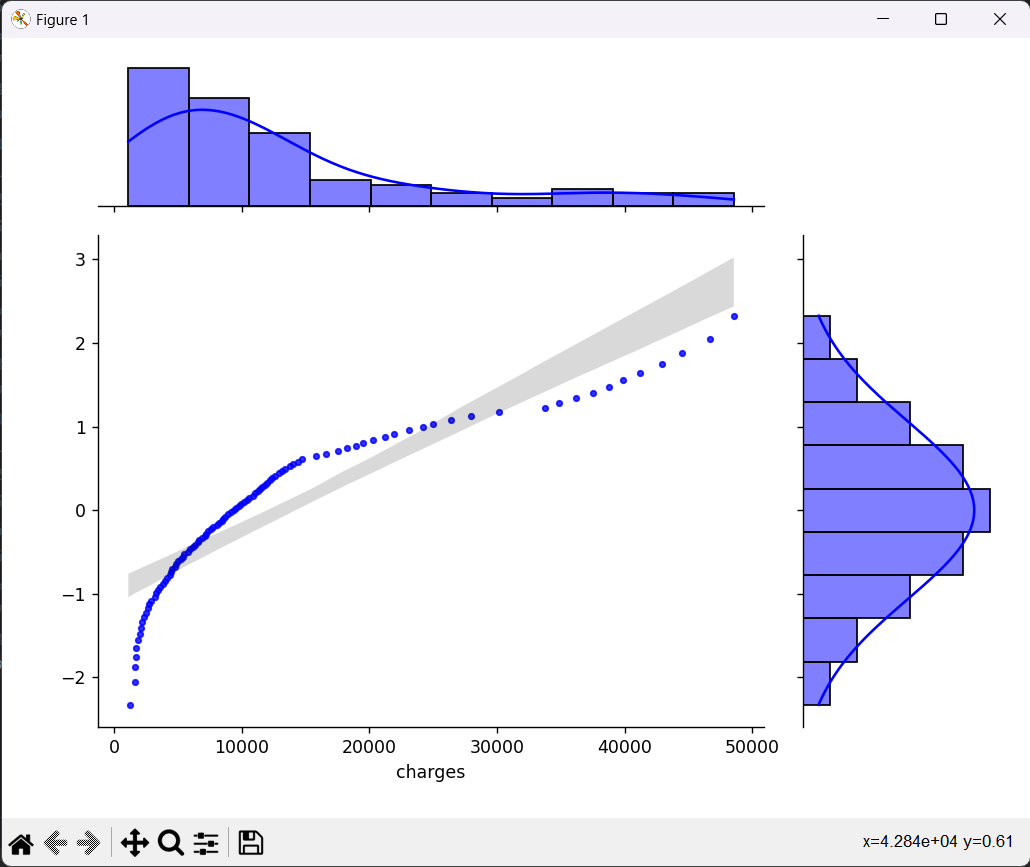
Проверить распределения следующих признаков на нормальность: индекс массы тела, расходы. Сформулировать нулевую и альтернативную гипотезы. Для каждого признака использовать KS-тест и q-q plot. Сделать выводы на основе полученных p-значений.



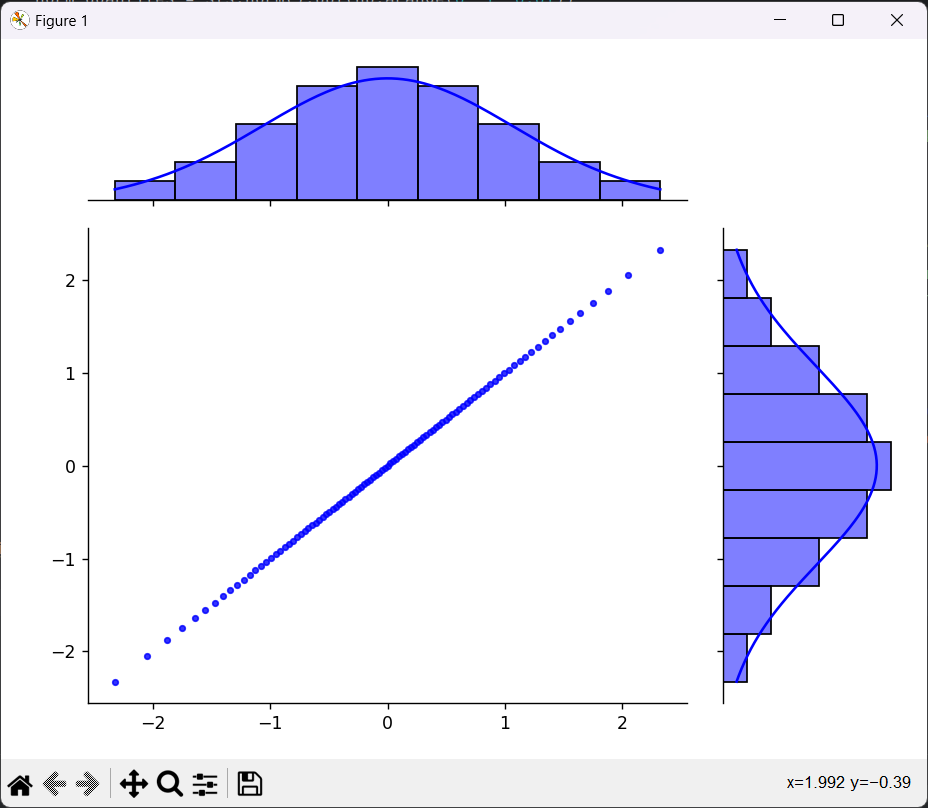
**Рисунок 16 – Код задания 8**



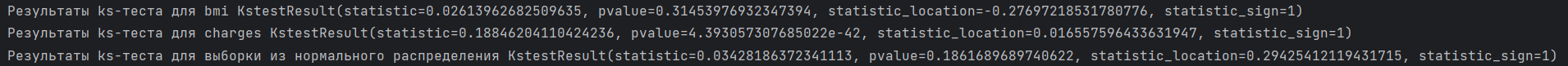
**Рисунок 17 – KS-тест и q-q plot для bmi**



**Рисунок 18 – KS-тест и q-q plot для charges**



**Рисунок 19 – KS-тест и q-q plot для нормального распределения**

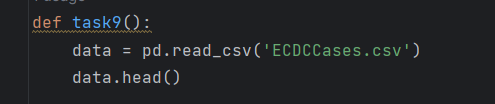
****

**Рисунок 20 – Текстовый результат**

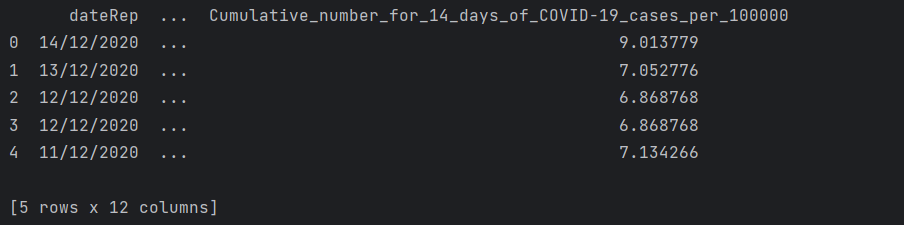
Исходя из полученных результатов, можно сделать вывод, что гипотеза H0 (данные соответствуют нормальному распределению) отвергается для charges, так как значение pvalue меньше зафиксированного уровня значимости, равного 0,05.

**Задание 9**

Загрузить данные из файла “ECDCCases.csv”.

****

**Рисунок 21 – Задание 9**

****

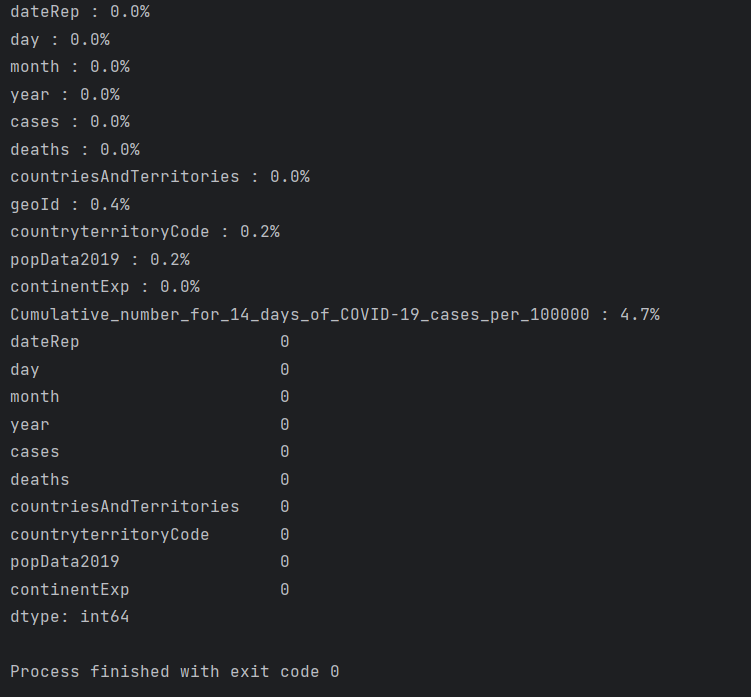
**Рисунок 22 – Результат**

**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.

****

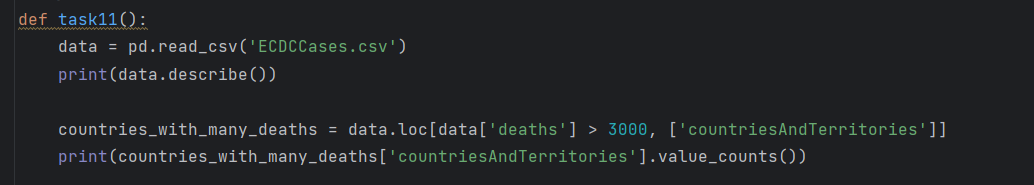
**Рисунок 23 – Задание 10**

****

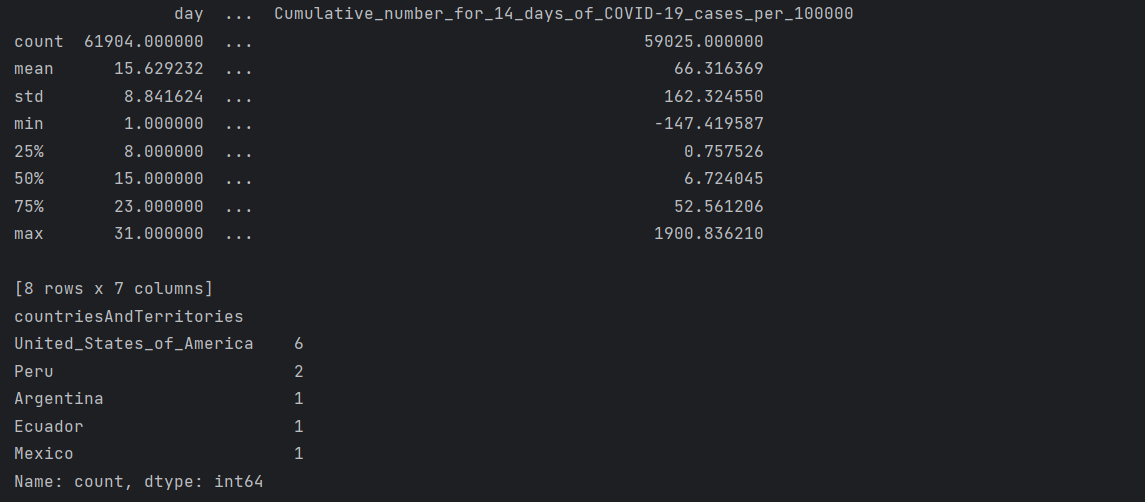
**Рисунок 24 – Результат**

**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.

****

**Рисунок 25 – Задание 11**

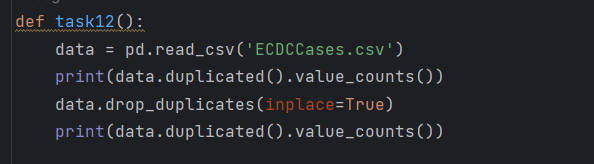
****

**Рисунок 26 – Результат**

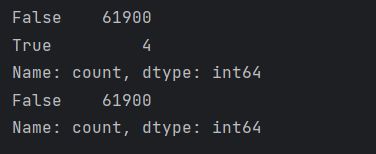
На основе полученных данных, можно сделать вывод о том, что признаки cases, popData2019 и deaths содержат выбросы.

**Задание 12**

Найти дублирование данных. Удалить дубликаты.

****

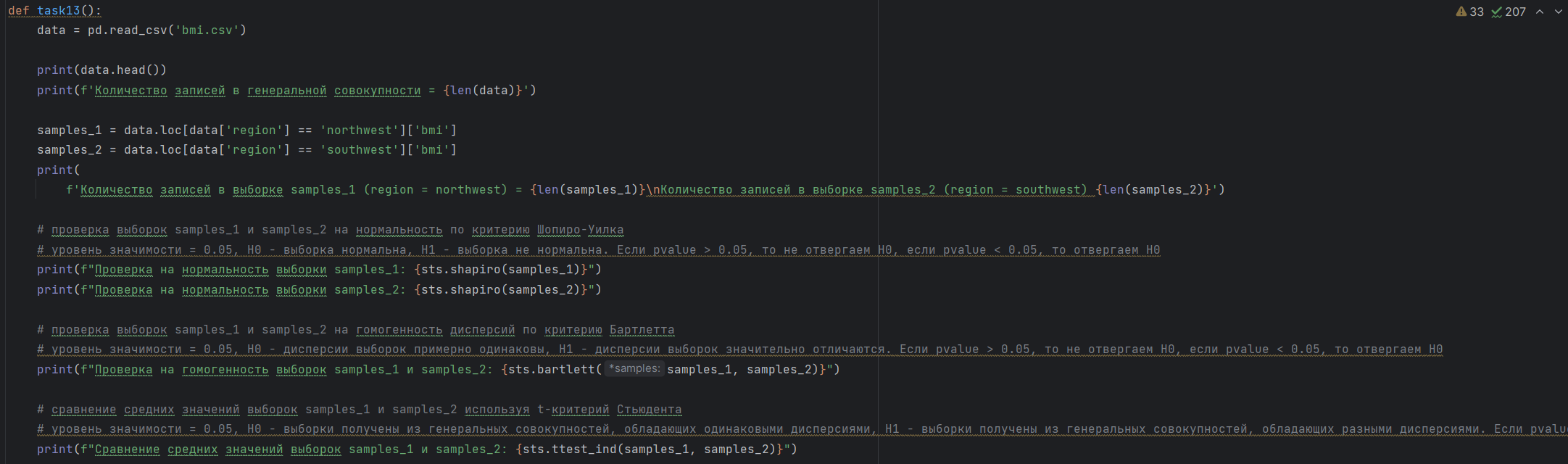
**Рисунок 27 – Задание 12**

****

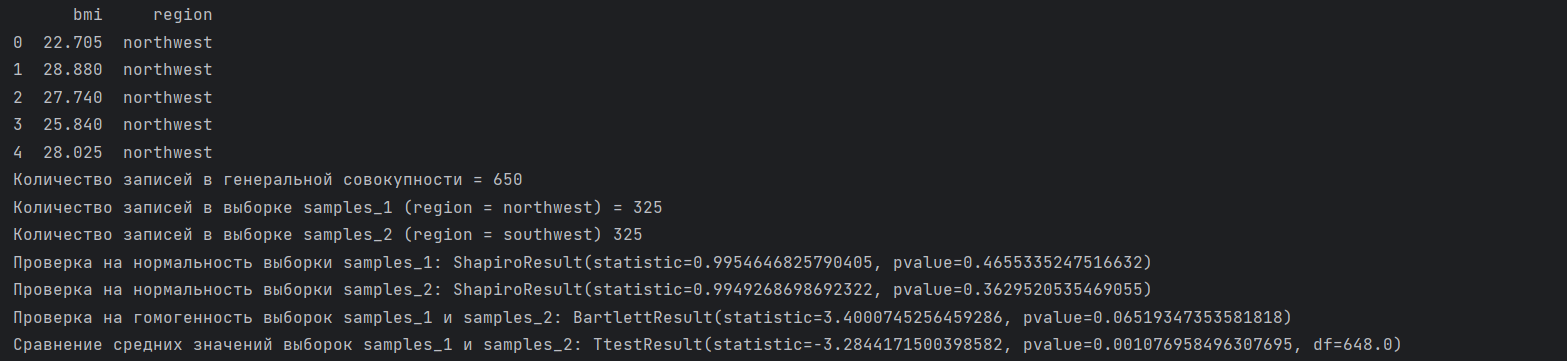
**Рисунок 28 – Результат**

**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).

****

**Рисунок 29 – Задание 13**

****

**Рисунок 30 – Результат**

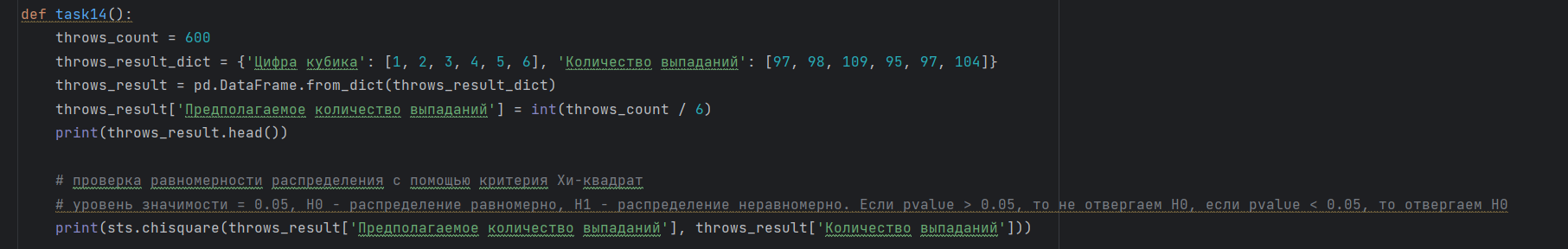
Исходя из полученных значений, можно сделать вывод, что выборки распределены нормально, дисперсии выборок примерно одинаковы, гипотезу H0 о том, что выборки получены из генеральных совокупностей, обладающих одинаковыми дисперсиями, отвергаем.

**Задание 14**

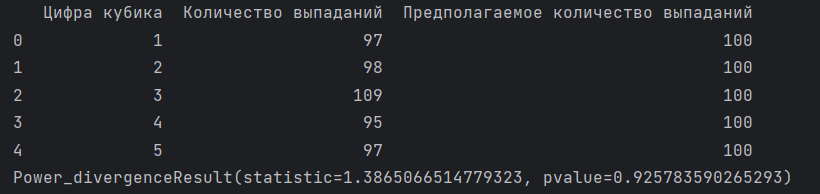
Кубик бросили 600 раз, получили следующие результаты:

|  |  |
| --- | --- |
| 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().

****

**Рисунок 31 – Задание 14**

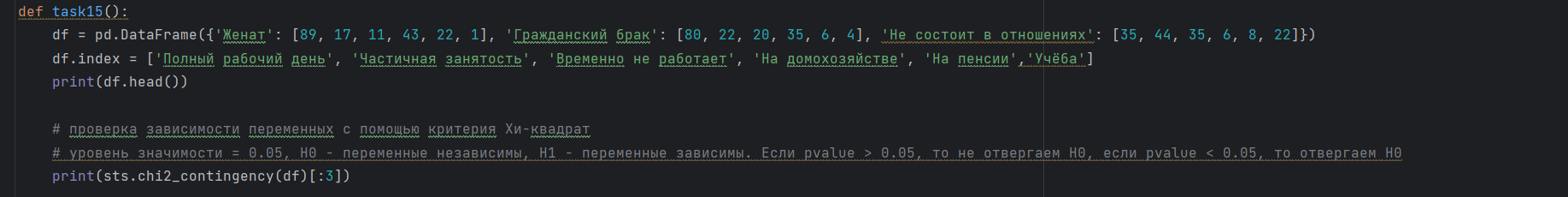
****

**Рисунок 32 – Результат**

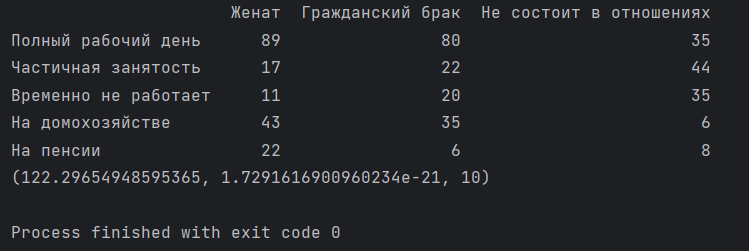
Так как pvalue больше 0,05, мы принимаем гипотезу H0, а следовательно полученное распределение равномерно.

**Задание 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?

****

**Рисунок 33 – Задание 15**

****

**Рисунок 34 – Результат**

Так как pvalue гораздо меньше 0,05, мы отвергаем гипотезу H0 о независимости переменных, следовательно семейное положение влияет на занятость.

**Вывод**

В ходе выполнения данной практической работы мы познакомились c основами статистики, научились строить гистограммы, меры центральной тенденции и меры разброса, строить box-plot, проверять выполняется ли центральная предельная теорема, определять доверительные интервалы с определённым коэффициентом доверия, а также проверять признаки на нормальность ks-тестом и строить q-q plot.

Предобработке данных, а также познакомились с выполнением статистических тестов, таких как: сравнение средних значений выборок по t-критерию Стьюдента, проверка на нормальность по критерию Шапиро-Уилка, проверка на гомогенность дисперсий по критерию Бартлетта, проверка на равномерность распределений с помощью критерия Хи-квадрат, проверка на зависимость с помощью критерия Хи-квадрат.