

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

**ОТЧЁТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ 3**

**по дисциплине «Технологии и инструментарий анализа больших данных»**

Выполнил студент группы ИКБО-02-20 Злобин А.О.

Принял

*Ассистент*

Горячев А.А.

Практические работы выполнены « » 2023 г.

(подпись студента)

«Зачтено» « » 2023 г.

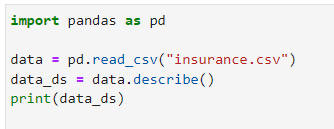
(подпись руководителя)

Москва 2023

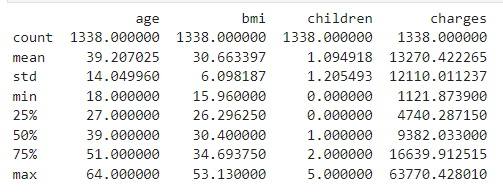
# Загрузить данные из файла “insurance.csv”.

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

**Код**

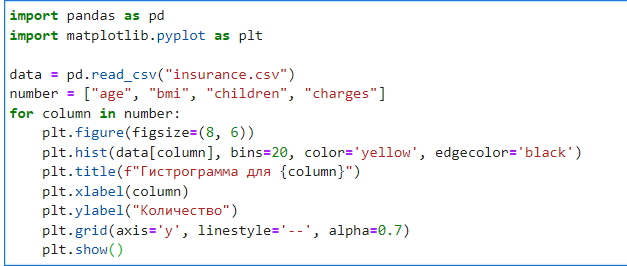
****

**Результат**

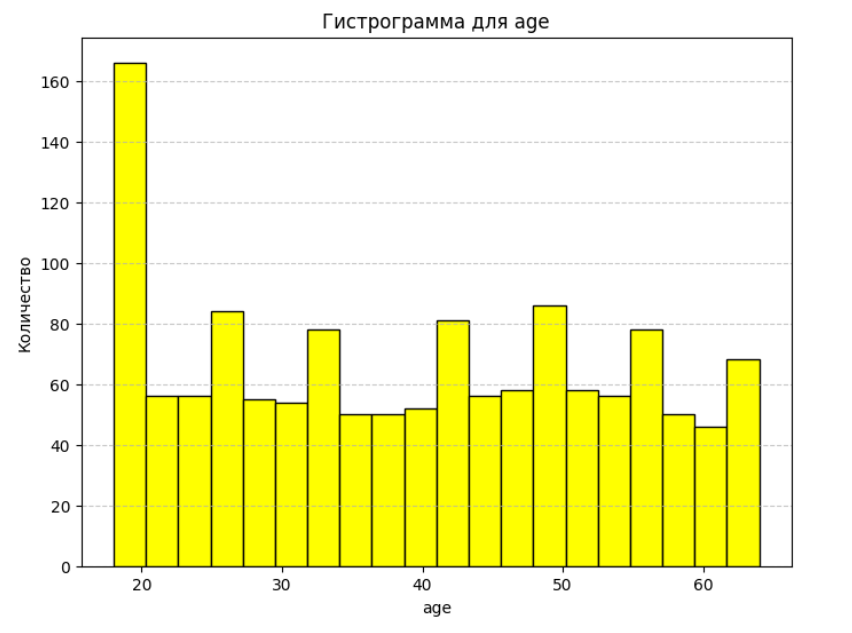
****

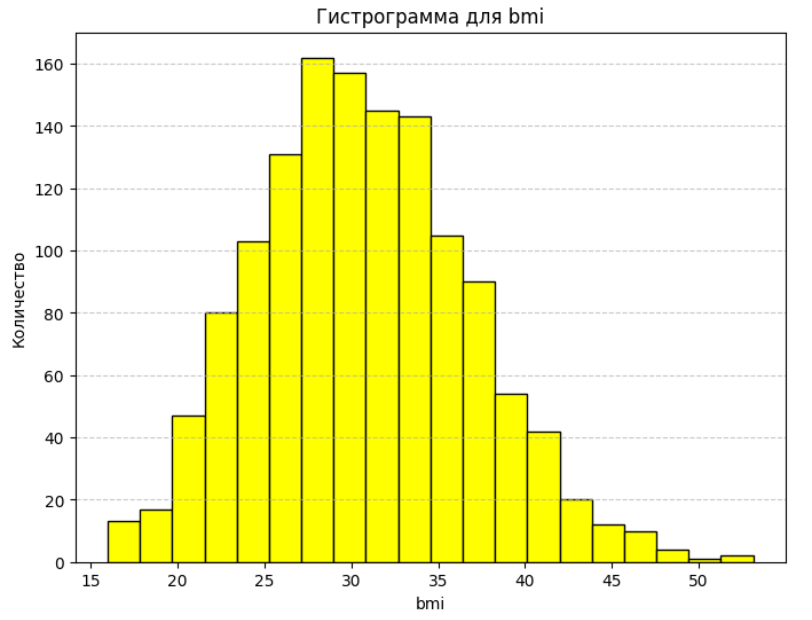
# Построить гистограммы для числовых показателей. Сделать выводы.

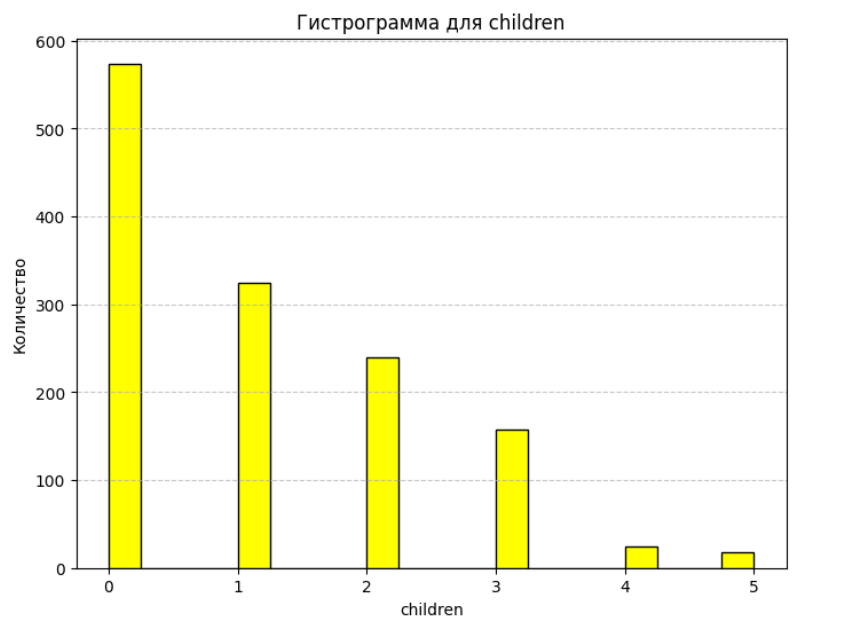
**Код**

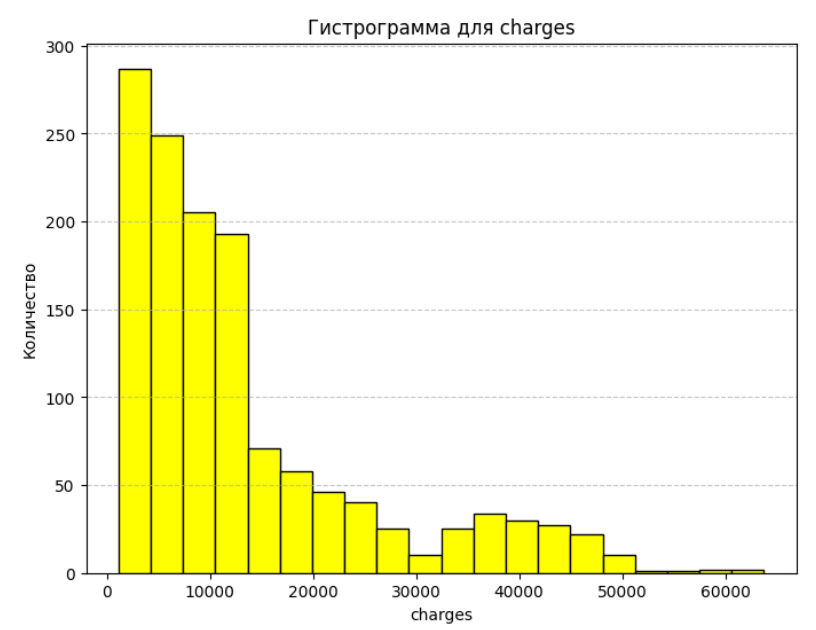


**Результат**

****





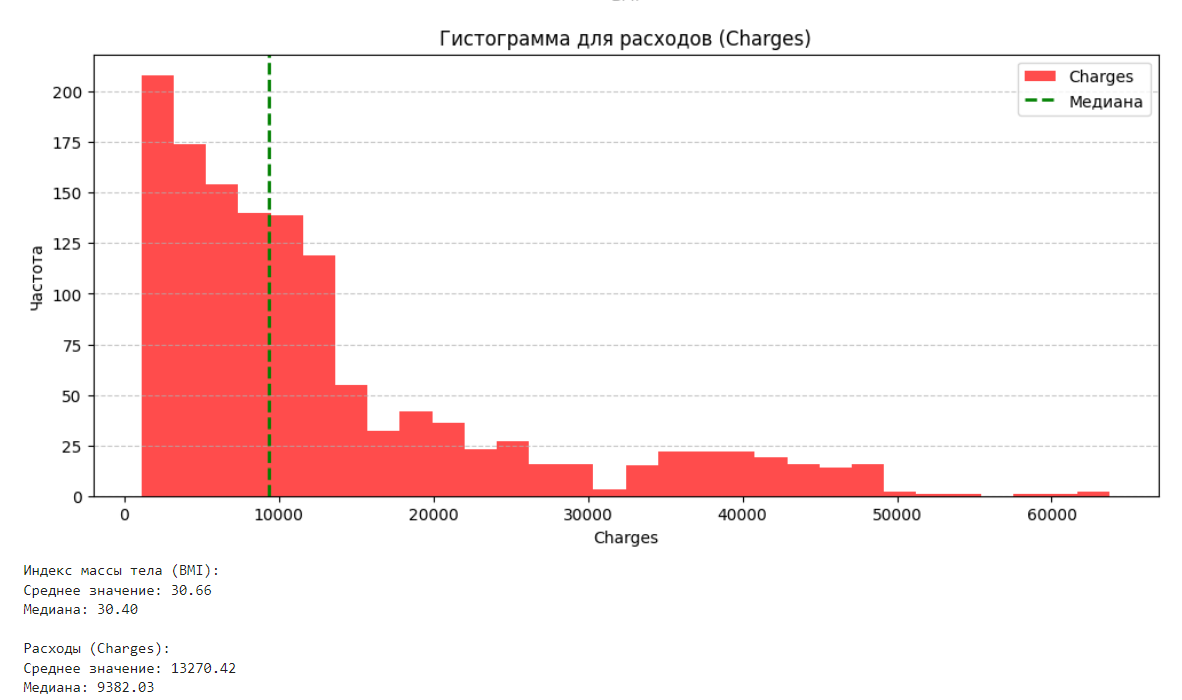
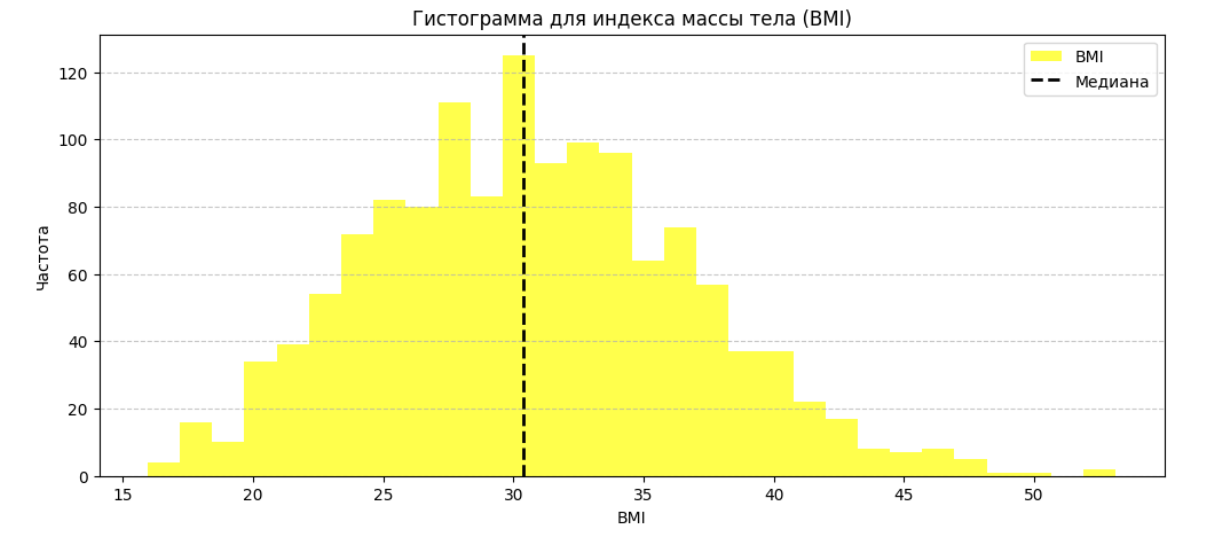
****

# 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**

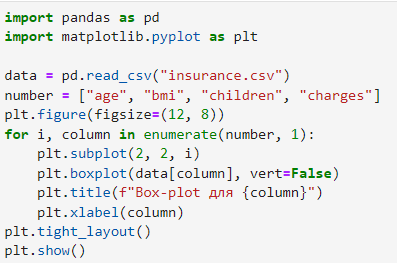


**Результат**

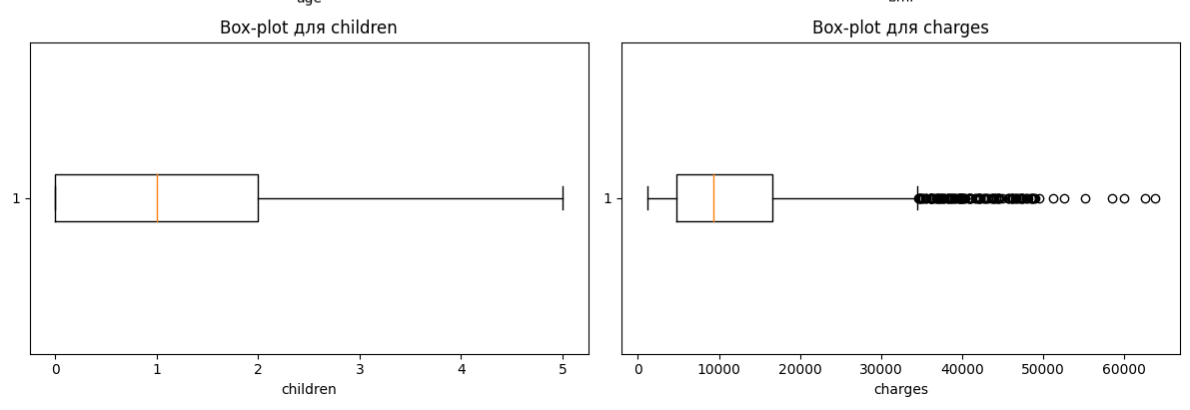
****

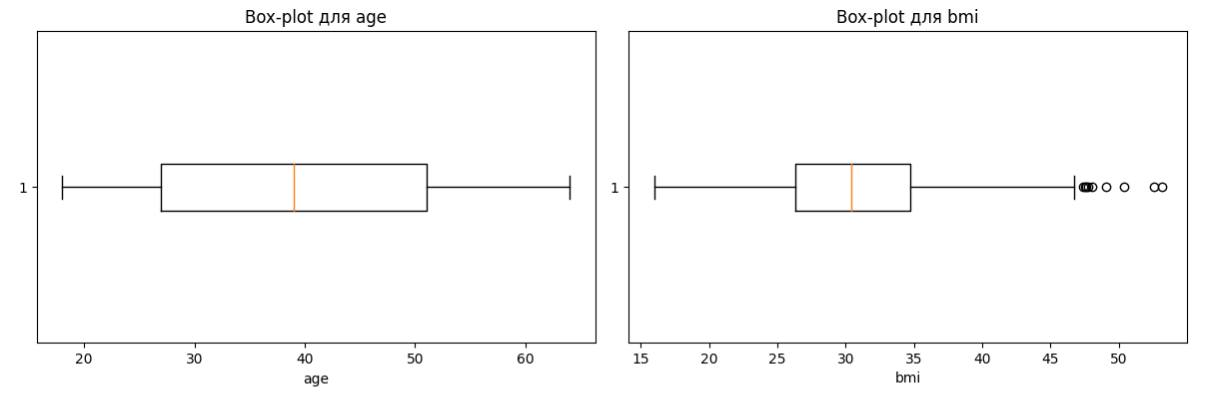
# Построить box-plot для числовых показателей. Названия графиков должны соответствовать названиям признаков. Сделать выводы.

**Код**



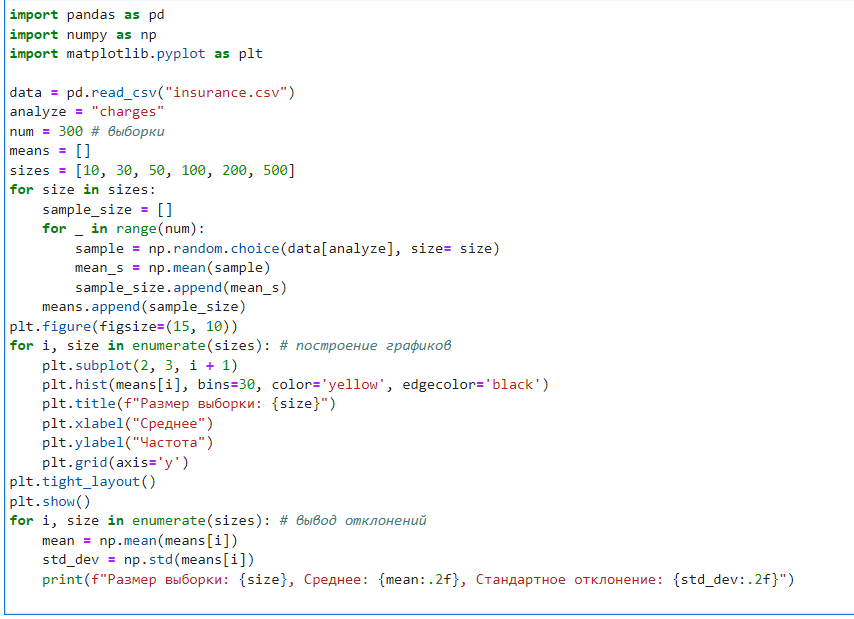
**Результат**



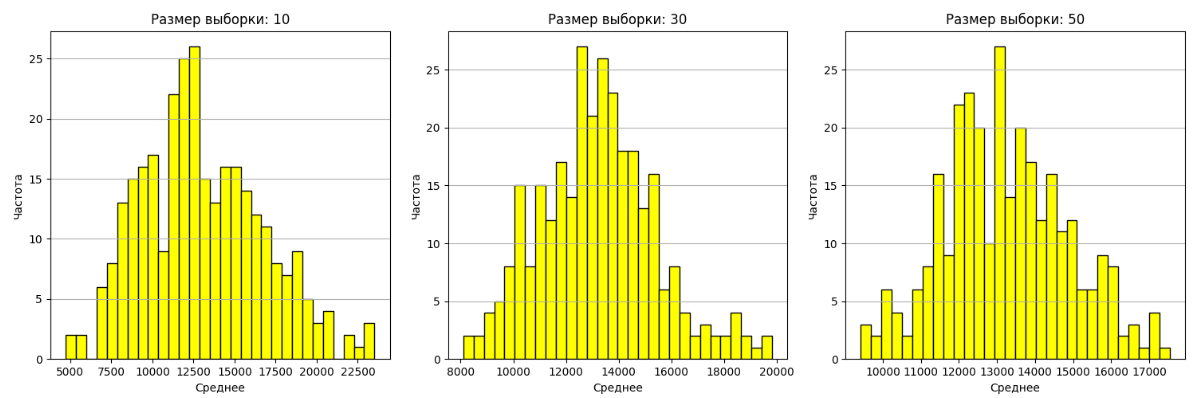


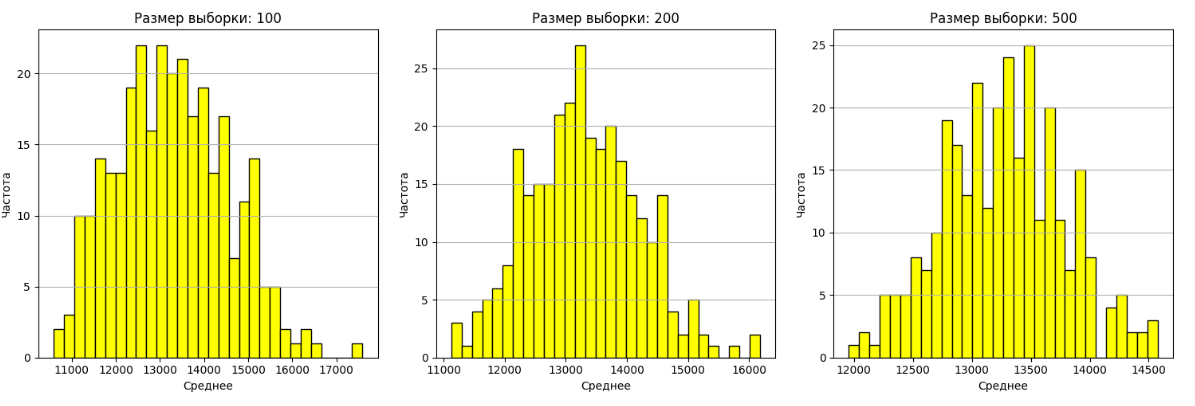
# Используя признак charges или imb, проверить, выполняется ли центральная предельная теорема. Использовать различные длины выборок n. Количество выборок = 300. Вывести результат в виде гистограмм. Найти стандартное отклонение и среднее для полученных распределений. Сделать выводы.

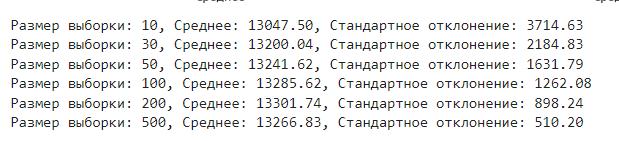
**Код**



**Результат**

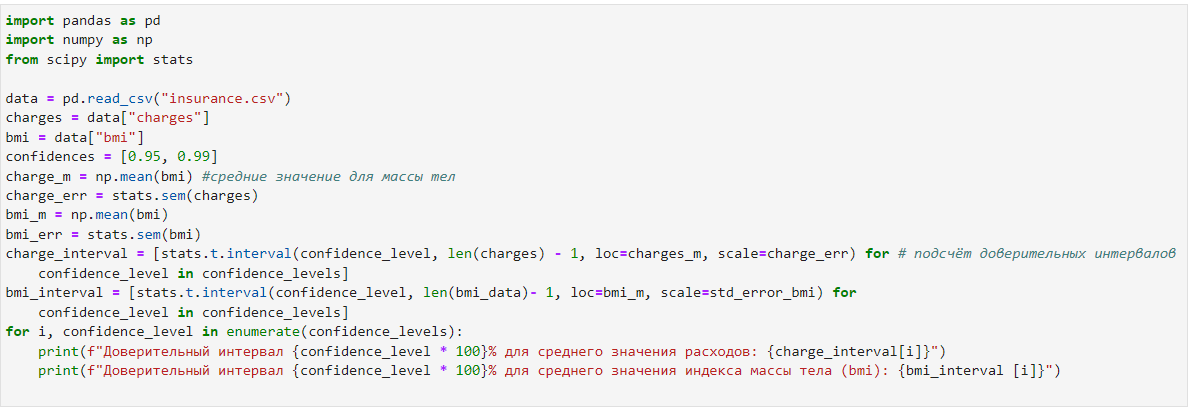
****

****

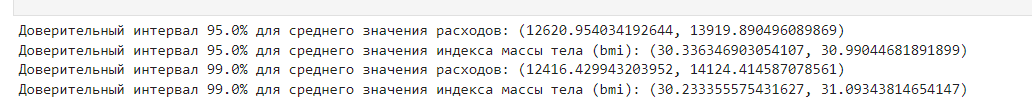
****

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

**Код**

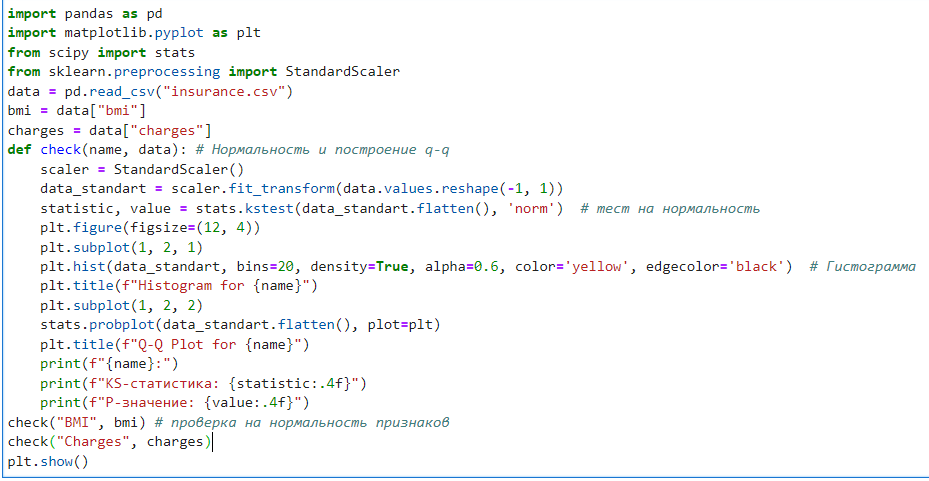


**Результат**

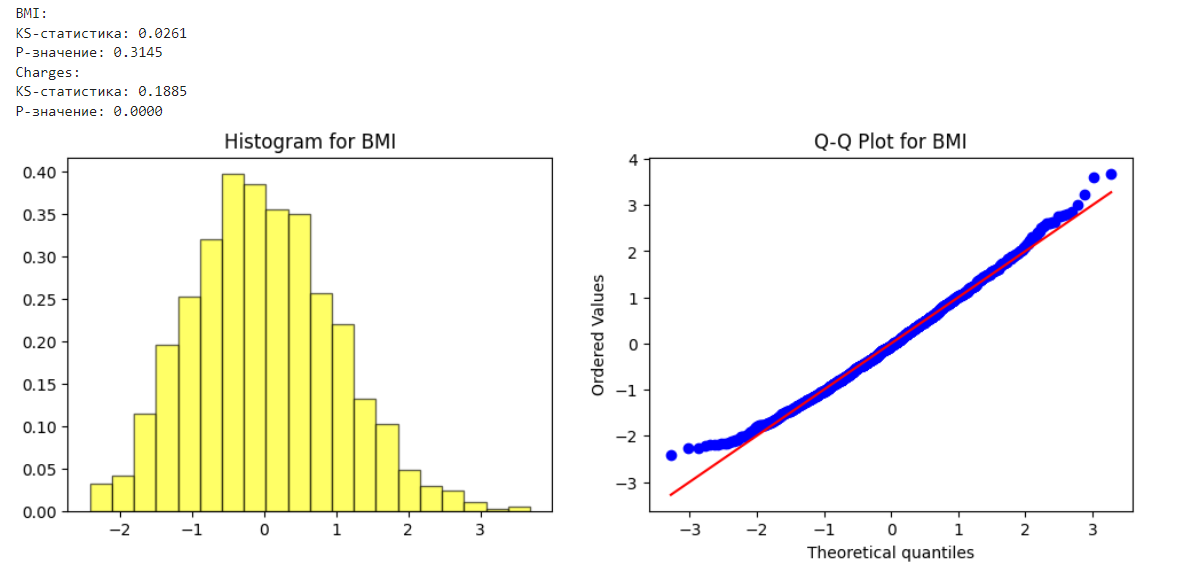
****

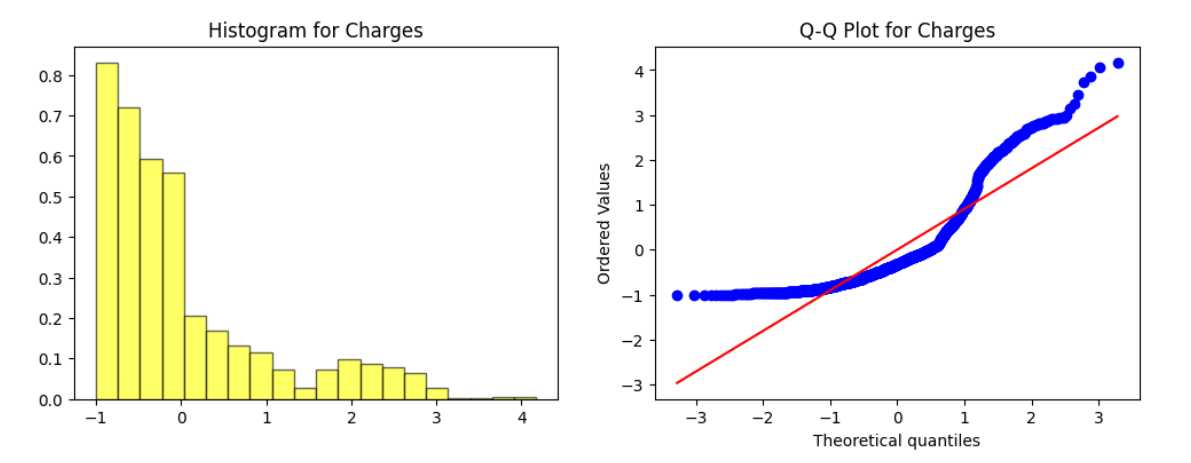
# Проверить распределения следующих признаков на нормальность: индекс массы тела, расходы. Сформулировать нулевую и альтернативную гипотезы. Для каждого признака использовать KS-тест и q-q plot. Сделать выводы на основе полученных p-значений.

**Код**



**Результат**





**Вывод**

В качестве нулевой гипотезы мы берем, что распределение нормальное, в качестве альтернативной, что нет. Следовательно, при p значении = 0.05, мы модем сказать, что первое распределение – нормальное (>0.05), второе – нет (<0.05). Как вывод гипотеза для BMI не отвергается, а гипотеза для Charges отвергается.

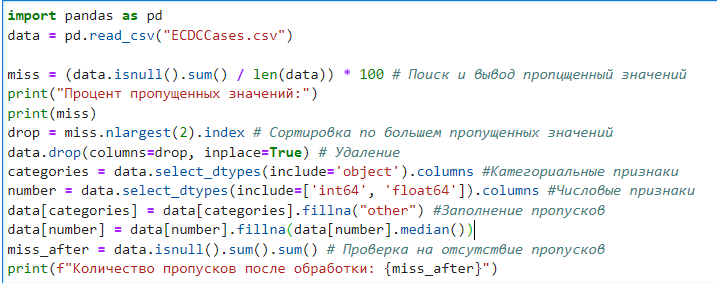
# Загрузить данные из файла “ECDCCases.csv”.

**Результат**

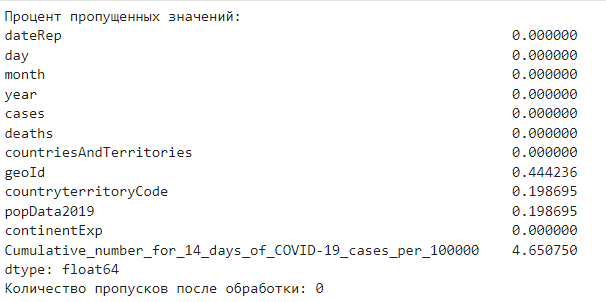
# 

1. **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.**

**Код**

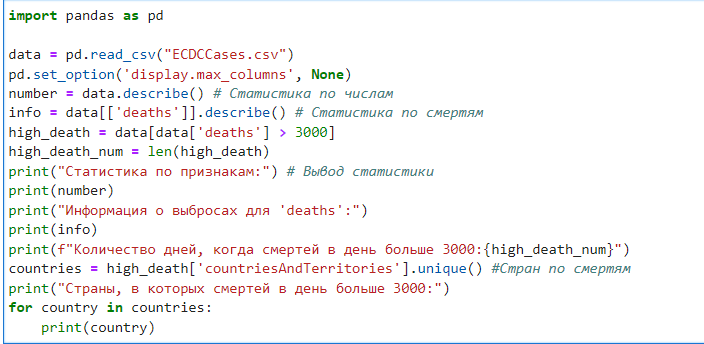


**Результат**

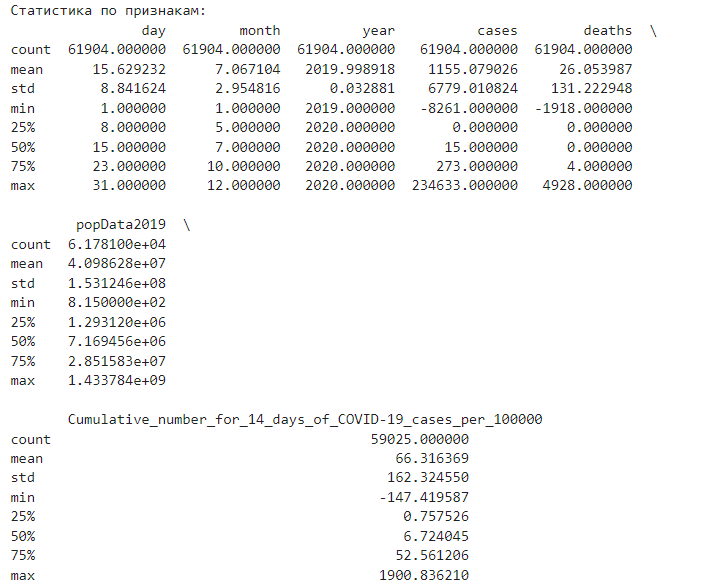
****

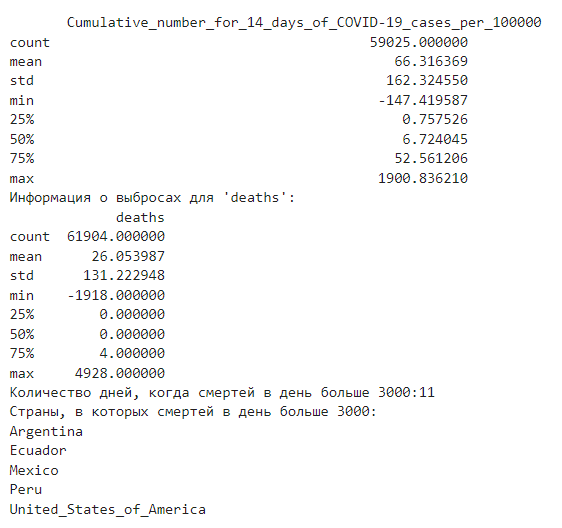
# Посмотреть статистику по данным, используя describe(). Сделать выводы о том, какие признаки содержат выбросы. Посмотреть, для каких стран количество смертей в день превысило 3000 и сколько таких дней было.

**Код**

****

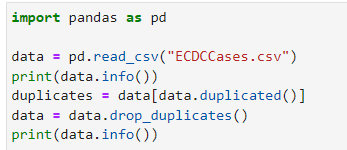
**Результат**

****

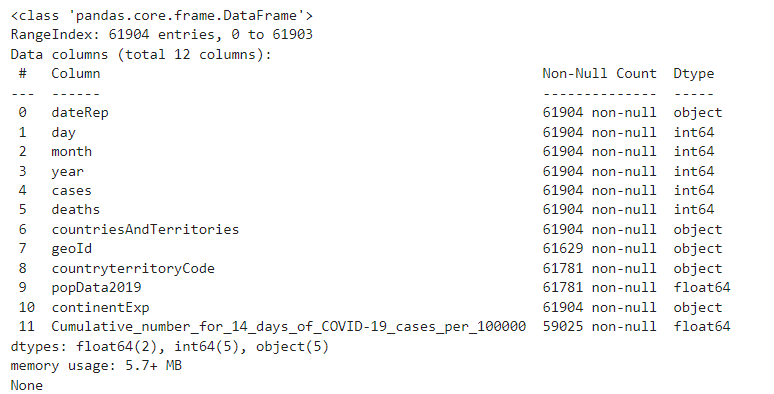
****

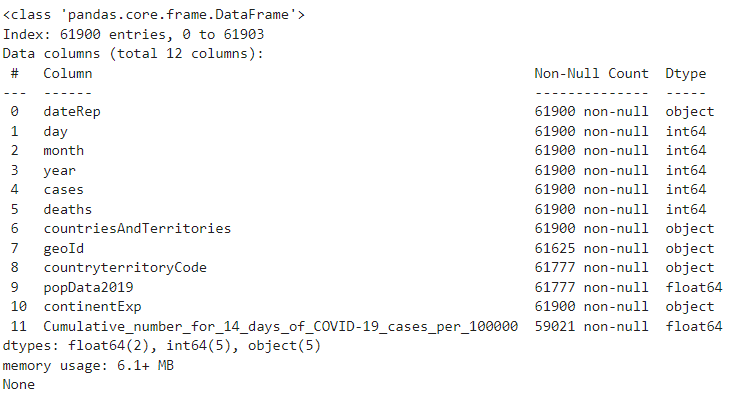
# Найти дублирование данных. Удалить дубликаты.

**Код**

****

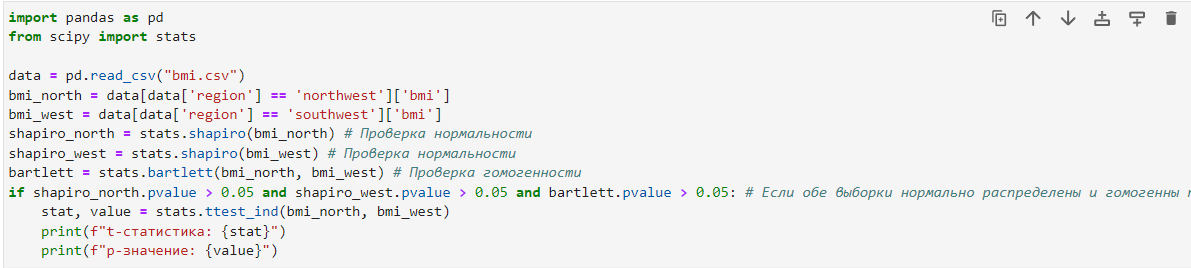
**Результат**

****

****

# 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).

**Код**

****

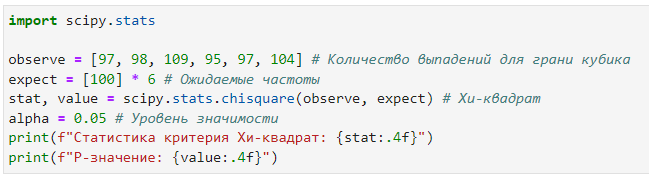
**Результат**

****

**Вывод**Так как среднее значение выборок различается, то мы отвергаем нулевую гипотезу.

# 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().

**Код**

****

**Результат**

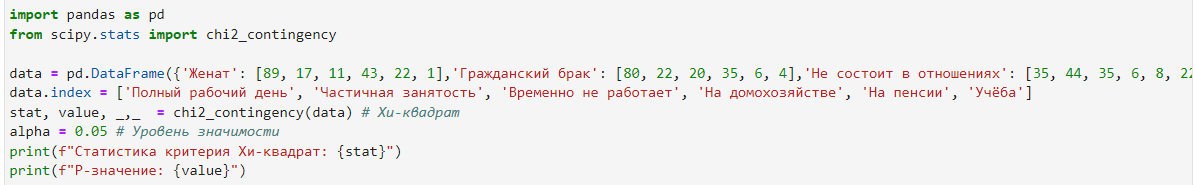
****

**Вывод**

Так как значение P больше, чем уровень значимости, то распределение равномерное.

# 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?

**Код**

****

**Результат**

****

**Вывод**

Так как значение P значительно меньше, чем альфа, то отклоняем нулевую гипотезу и значит семейное положение и занятость зависимы друг от друга.