|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  |  |  | | 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 | Мухаметшин А. Р. |
| Проверил ассистент кафедры ПМ ИИТ | Тетерин Н.Н. |

**Задание 1**

Загрузить данные из файла insurance.csv

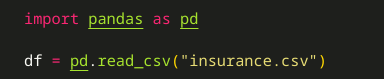


Рисунок 1 – выполнение задания 1

**Задание 2**

С помощью метода describe() посмотреть статистику по данным.

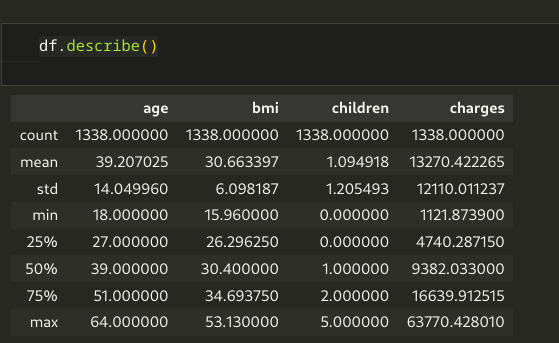


Рисунок 2 – Решение задания 2

Выводы:

1. count – количество значений.

2. mean– среднее значение.

3. std – стандартное отклонение.

4. min – минимальное значение.

5. 25% – первый квартиль.

6. 50% – медиана (второй квартиль).

7. 75% – третий квартиль.

8. max – максимальное значение.

**Задание 3**

Построить гистограммы для числовых показателей. Сделать выводы



Рисунок 3 – Код к заданию 3

Пример работы представлен на рисунке 5.

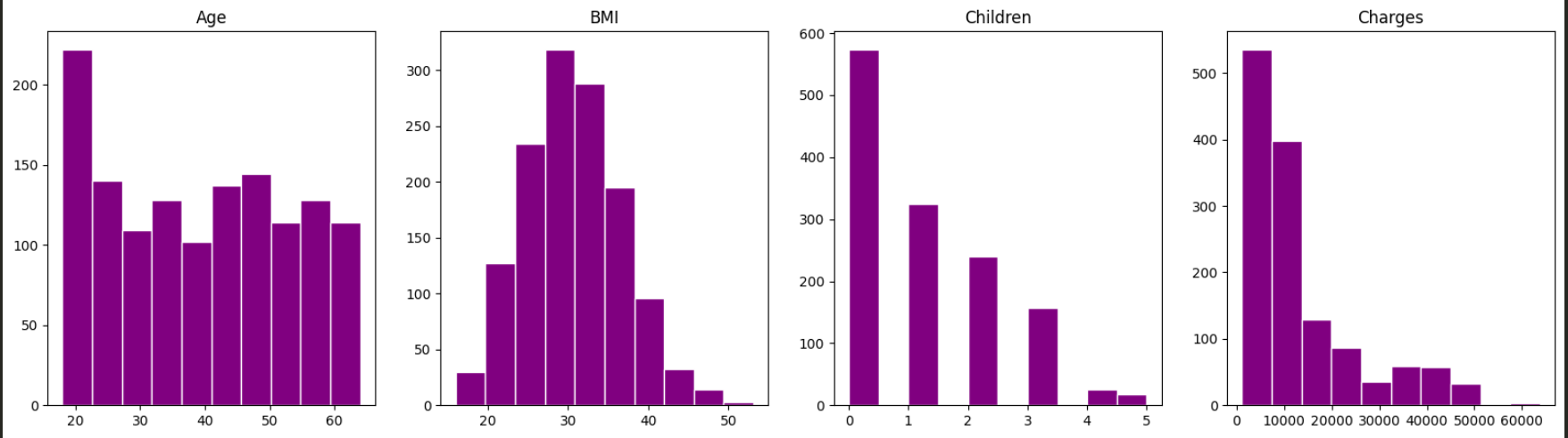


Рисунок 5 – Пример работы

Вывод:

Средний возраст участников составляет примерно 39 лет, с диапазоном от 18 до 64 лет. Это говорит о разнообразии в возрастной категории, с большинством людей зрелого возраста

Средний BMI равен 30.66, что соответствует категории избыточного веса или ожирения. Минимальное значение BMI - 15.96, что указывает на наличие людей с низкой массой тела. Максимальное значение - 53.13

В среднем у респондентов 1 ребенок. Большинство - либо бездетные, либо имеют одного или двух детей (максимум - 5)

Средний уровень страховых взносов составляет около 13270 долларов с большим разбросом (от 1121 до 63770 долларов), что указывает на значительное разнообразие в стоимости страхования, скорее всего связанное с возрастом, BMI и количеством детей

**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 6 – Код к заданию 4

Пример работы представлен на рисунке 7.

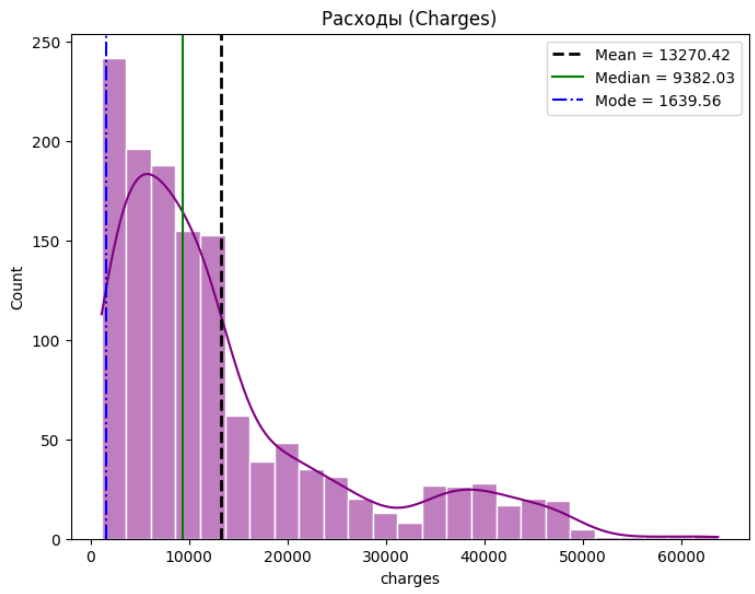
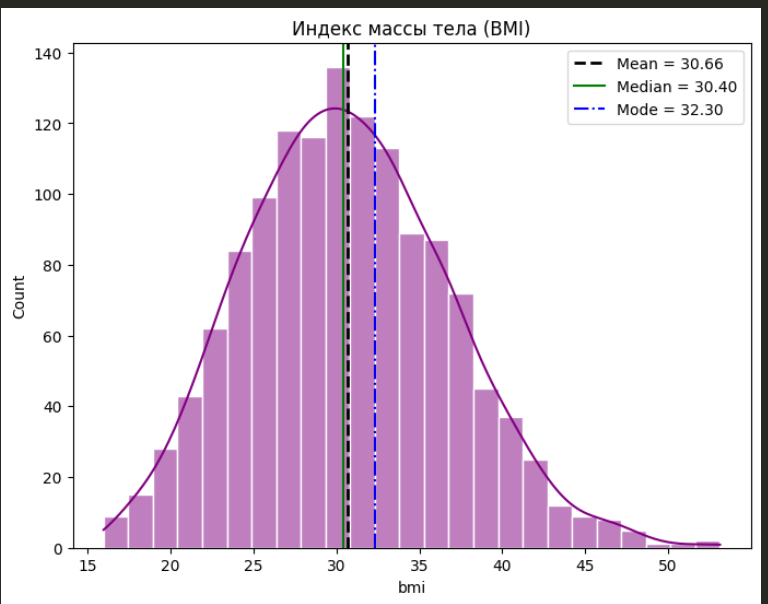
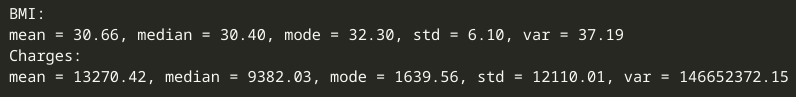


Рисунок 7 – Пример работы

Вывод

Индекс массы тела (BMI):

Среднее и медиана: Средний BMI составляет 30.66, в то время как медиана - 30.40. Это указывает на то, что распределение значений имеет незначительные отклонения и сосредоточено вокруг значения избыточного веса

Наиболее часто встречающееся значение (мода) BMI равно 32.30, что также подтверждает, что большинство участников имеют избыточный вес или находятся на грани ожирения

Стандартное отклонение 6.10 и дисперсия 37.19 указывают на умеренное разнообразие значений BMI среди участников

Страховые взносы:

Среднее и медиана: Средний размер страховых взносов составляет 13270.42 долларов, в то время как медиана - 9382.03 долларов. Это говорит о том, что среднее значение значительно выше медианы, что может свидетельствовать о наличии некоторых участников с высокими страховками, влияющих на среднее

Наиболее распространенное значение страховых взносов (мода) - 1639.56 долларов, что сильно ниже среднего значения. Это подтверждает присутствие группы людей с низкими расходами на страхование

Стандартное отклонение 12110.01 и дисперсия 146652372.15 указывают на широкий разброс данных о страховых взносах, что может означать значительное разнообразие в состоянии здоровья и других факторах участников

**Задание 5**

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

Код к заданию 5 представлен на рисунке 8.

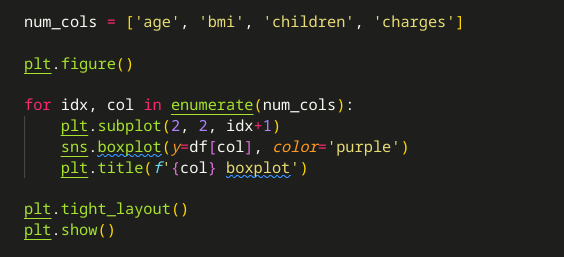


Рисунок 8 – Код к заданию 5

Пример работы представлен на рисунке 9.

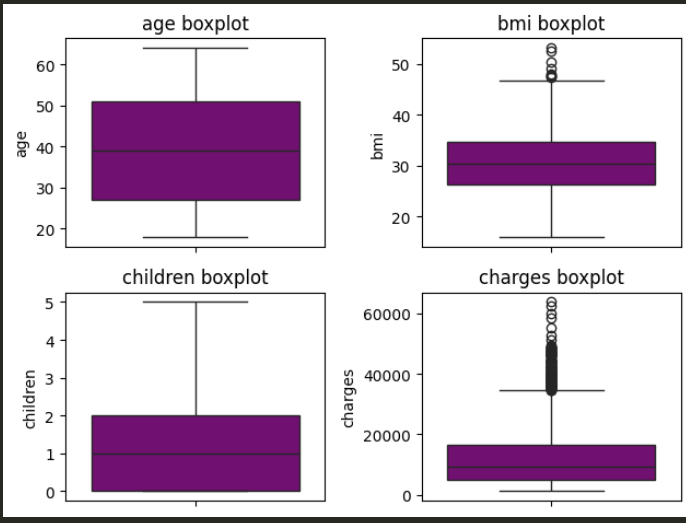


Рисунок 9 – Пример работы

**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 гистрограмм.

Найти стандартное отклонение и среднее для полученных распределений. Сделать выводы

Код к заданию 6 представлен на рисунке 10.

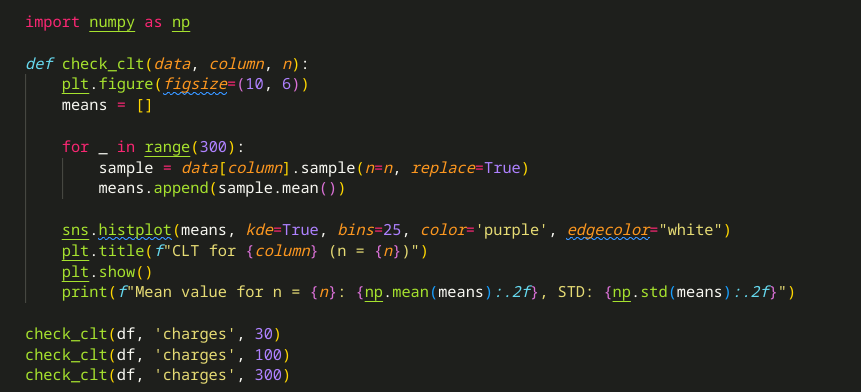


Рисунок 10 – Код к заданию 6

Пример работы представлен на рисунке 11.

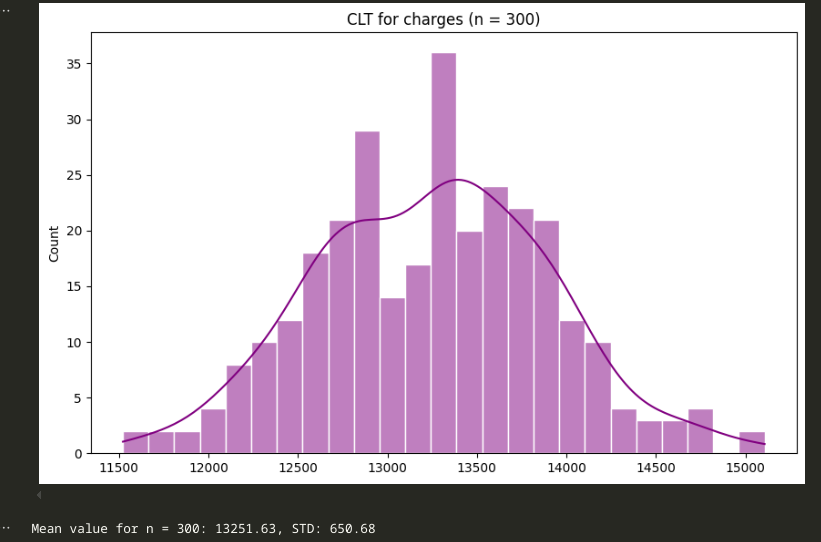
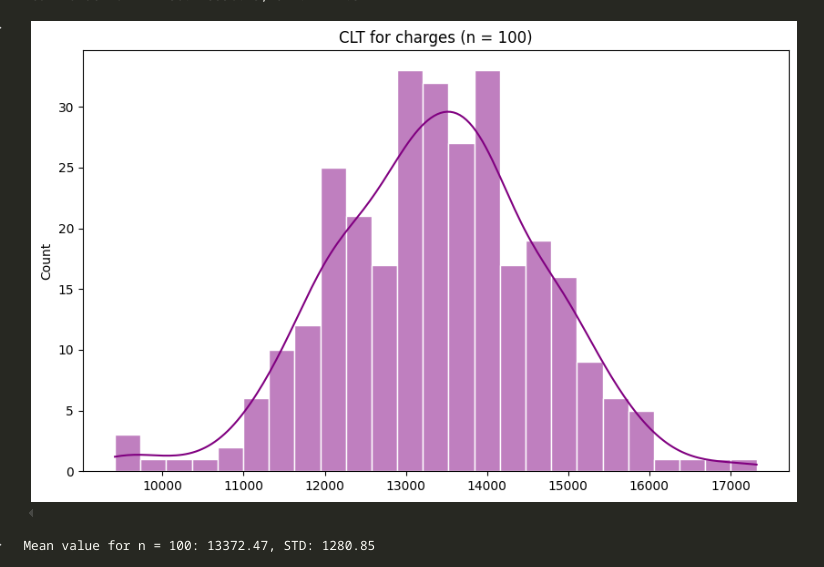
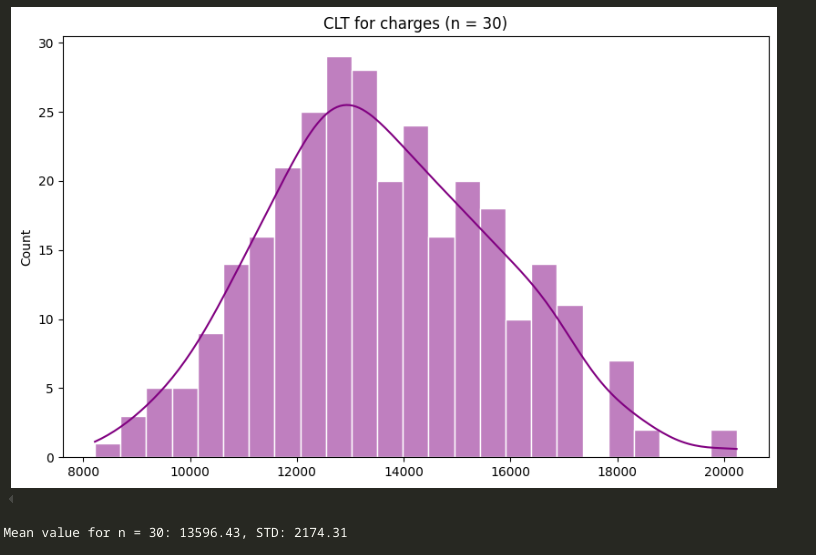


Рисунок 11 – Пример работы

Выводы:

Согласно ЦПТ, распределение средних значений выборок будет стремиться к нормальному распределению по мере увеличения размера выборки, независимо от формы исходного распределения. На гистограммах наблюдается, что при увеличении размера выборки гистограммы становятся более "нормальными" по форме. Это является подтверждением ЦПТ

Как видно из стандартного отклонения, оно уменьшается при увеличении размера выборки: от 2174.31 (n = 30) до 650.68 (n = 300). Это означает, что средние значения выборок становятся более сосредоточенными вокруг истинного среднего

На гистограммах для n = 100 и n = 300 видно, что с увеличением размера выборки распределение становится более концентрированным, что также соответствует ЦПТ: с большим размером выборки мы имеем больше уверенности в оценке среднего значения

**Задание 7**

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

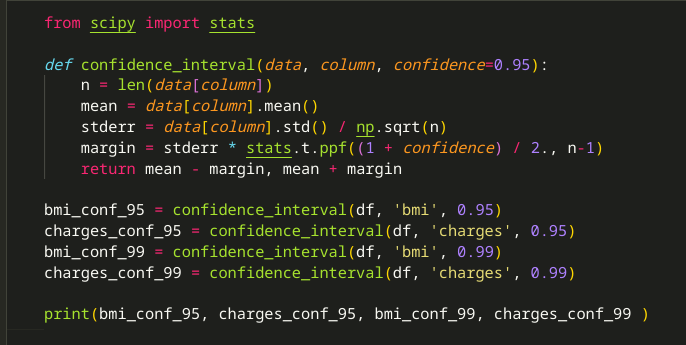


Рисунок 12 – выполнение задания

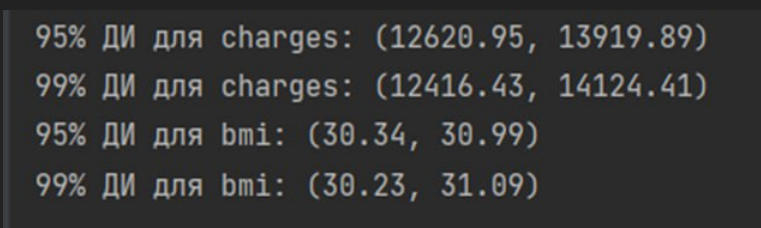


Рисунок 13 – результат работы

**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.

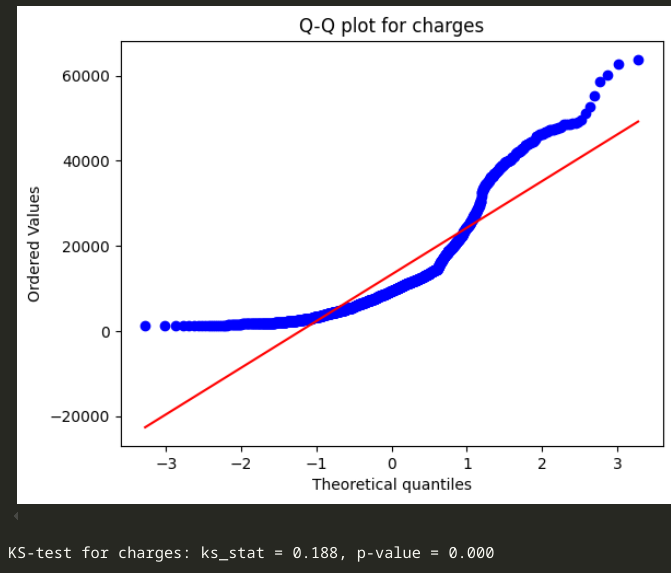
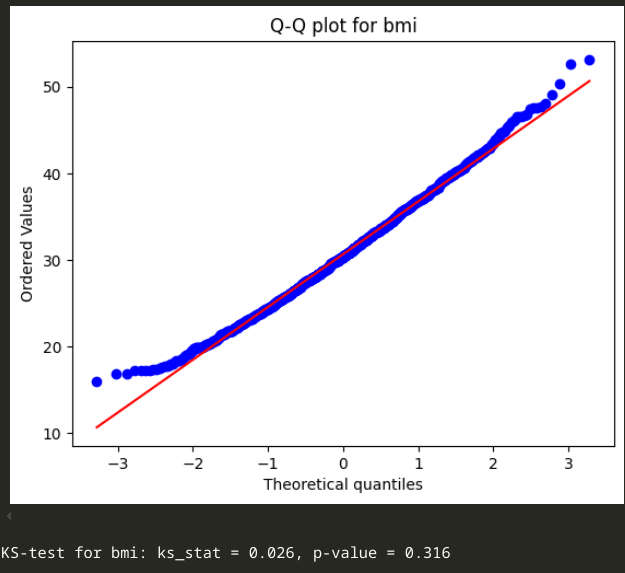


Рисунок – выполнение задания

Основная гипотеза (H0)

> Нет значительных различий/влияний в зависимости от рассматриваемых факторов на изучаемую величину (или переменную)

Альтернативная гипотеза (H1)

> Существуют значительные различия/влияния в зависимости от рассматриваемых факторов на изучаемую величину (или переменную)

Индекс массы тела

Поскольку p-value (0.316) больше уровня значимости (0.05), мы не можем отвергнуть нулевую гипотезу. Распределение индекса массы тела можно считать нормальным

Расходы

Поскольку p-value (0.000) меньше уровня значимости (0.05), мы отвергаем нулевую гипотезу. Распределение расходов не является нормальным

**Задание 9**

Загрузить данные из файла ECDCCases.csv



Рисунок – выполнение задания

**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.

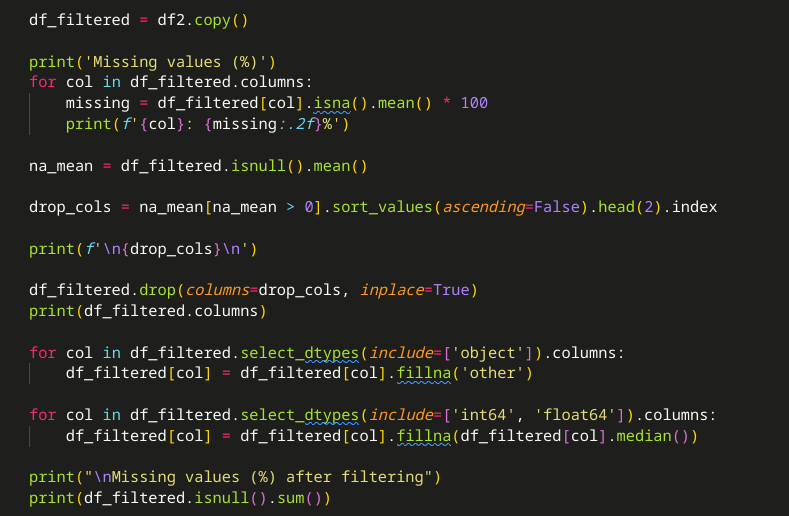


Рисунок – код задания



Рисунок – выполнение задания

**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.

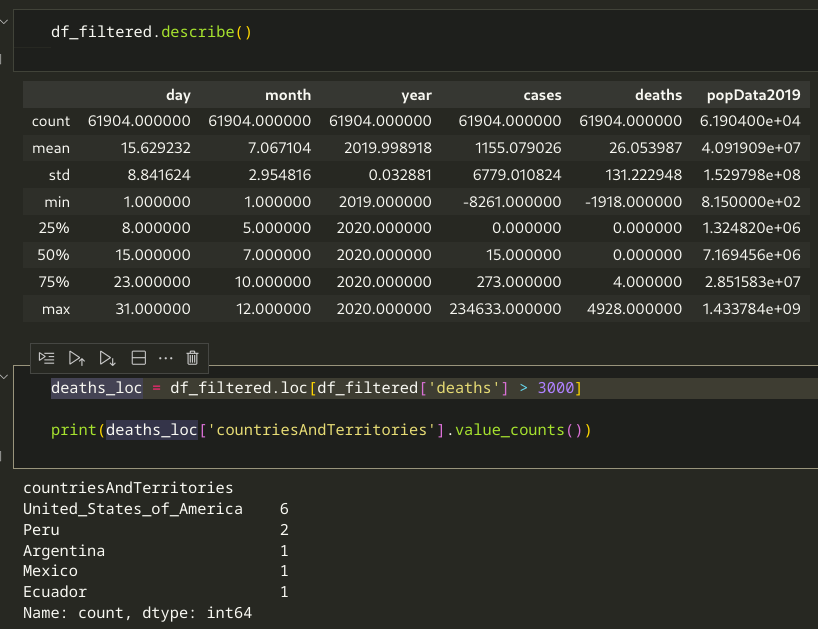


Рисунок – выполнение задания

**Задание 12**

Найти дублирование данных. Удалить дубликаты

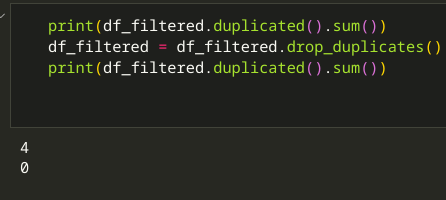


Рисунок – выполнение задания

**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 ШопироУилка) и на гомогенность дисперсии (критерий Бартлетта).

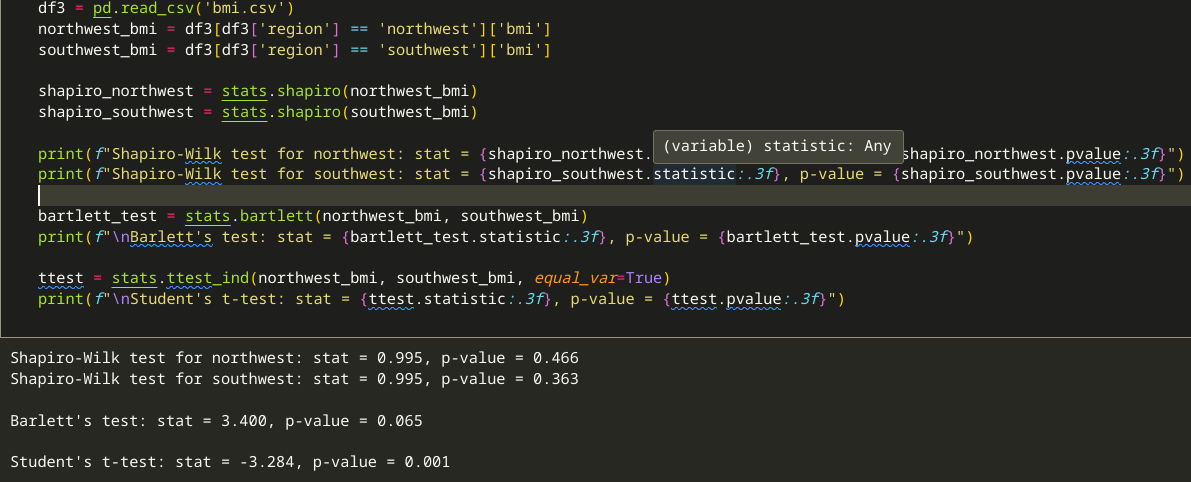


Рисунок – выполнение задания

Вывод:

Результаты теста Шапиро-Уилка показывают, что p-value для обеих выборок (регион northwest: 0.466 и регион southwest: 0.363) выше уровня значимости 0.05. Это указывает на то, что обе выборки можно считать нормально распределенными

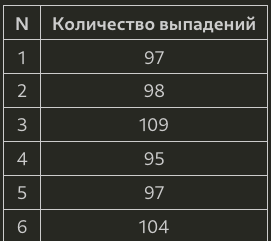
Результат теста Бартлетта (p-value = 0.065) также выше 0.05, что говорит о том, что дисперсии в двух выборках однородны

Результат t-критерия Стьюдента (p-value = 0.001) меньше 0.05, что позволяет отвергнуть нулевую гипотезу. Это свидетельствует о наличии статистически значимой разницы в средних значениях индекса массы тела между людьми из регионов northwest и southwest

Существует значительная разница в среднем индексе массы тела между людьми из регионов northwest и southwest

**Задание 14**

Кубик бросили 600 раз, получили следующие результаты:



С помощью критерия Хи-квадрат проверить, является ли полученное распределение равномерным.

Использовать функцию scipy.stats.chisquare()

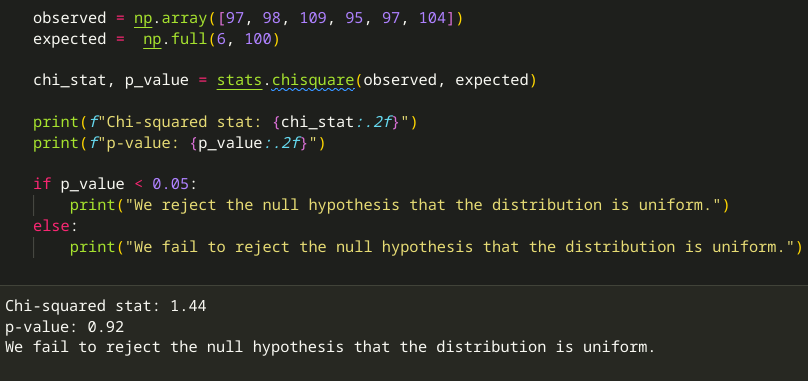


Рисунок – выполнение задания

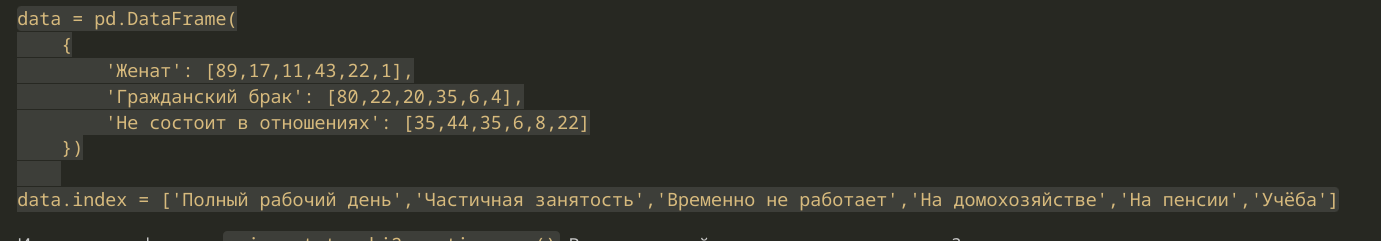
Вывод:

Полученное распределение выпадений значений можно считать равномерным.

**Задание 15**

С помощью критерия Хи-квадрат проверить, являются ли переменные зависимыми.

Создать датафрейм, используя следующий код:



Использовать функцию scipy.stats.chi2\_contingency().

Влияет ли семейное положение на занятость?

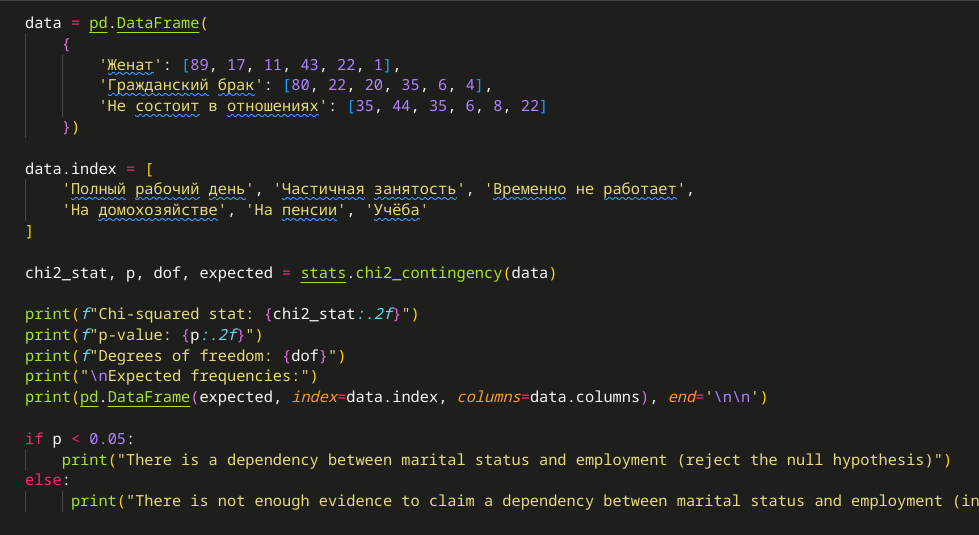


Рисунок – код задания

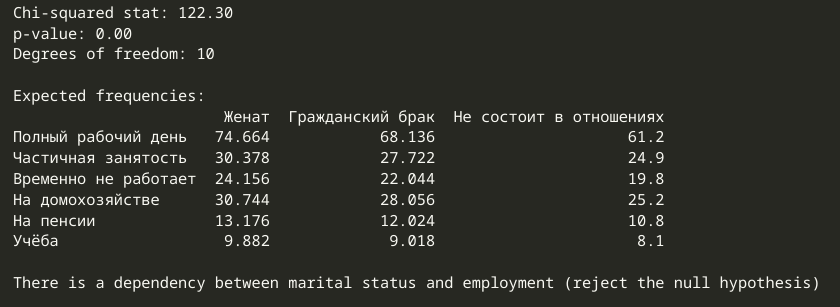


Рисунок – выполнение задания

Вывод

Так как p-value (0.00) меньше 0.05, мы отвергаем нулевую гипотезу и утверждаем, что между семейным положением и типом занятости существует статистически значимая зависимость