|  |
| --- |
|  |
| МИНОБРНАУКИ РОССИИ |
| Федеральное государственное бюджетное образовательное учреждение  высшего образования  **«МИРЭА - Московский технологический университет»**  **РТУ МИРЭА** |

Институт Информационных Технологий

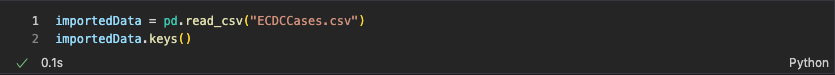
Кафедра Прикладной Математики (ПМ)

**ПРАКТИЧЕСКАЯ РАБОТА № 6**

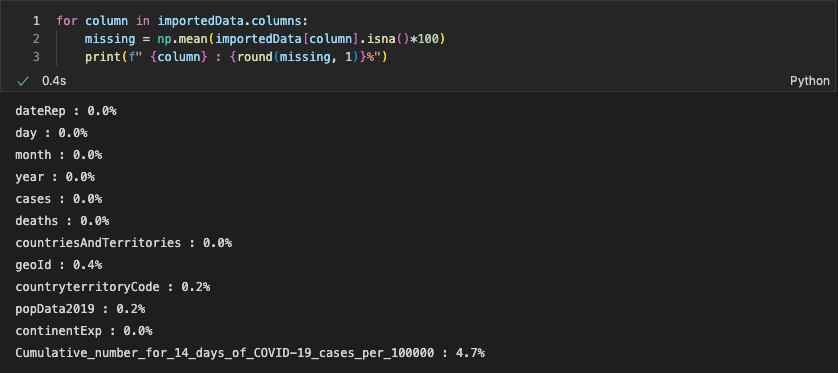
|  |  |
| --- | --- |
| Выполнил студент группы ИКБО-08-19  Борисов А.В. | (\_\_\_\_\_\_\_\_\_\_)  *подпись* |
| Принял Ассистент кафедры ПМ  Высоцкая А.А. | (\_\_\_\_\_\_\_\_\_\_)  *подпись* |
| Практическая работа выполнена | « » 2022 г. |
| «Зачтено» | « » 2022 г. |

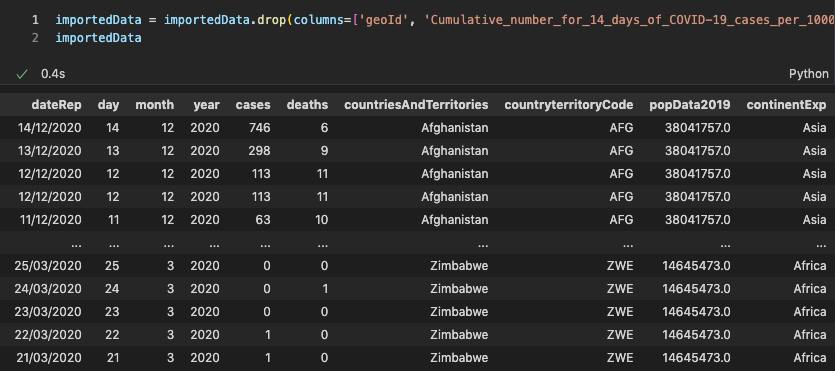
Москва 2022

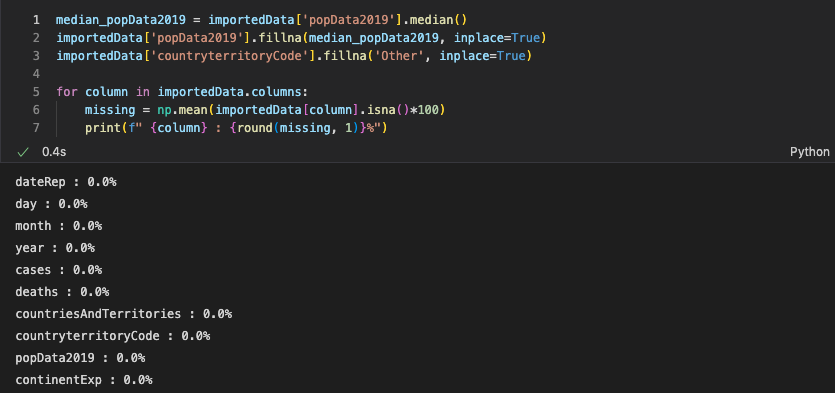
1. *Загрузить данные из файла “ECDCCases.csv”.*



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.*







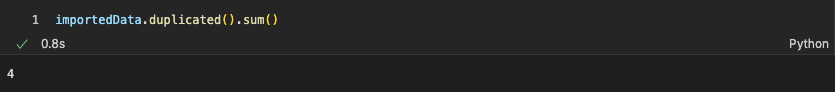
1. *Посмотреть статистику по данным, используя describe(). Сделать выводы о том, какие признаки содержат выбросы. Посмотреть, для каких стран количество смертей в день превысило 3000 и сколько таких дней было.*

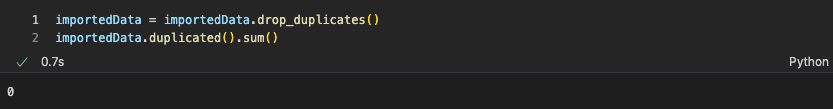


Из .describe() можно сделать вывод, что исходные данные содержат ошибки ввода (отрицательные значения, в колонках «cases» и «deaths», так же что они содержат выбросы. Это можно определить из того, что разница между средним и максимальным значением сильно больше стандартного отклонения.

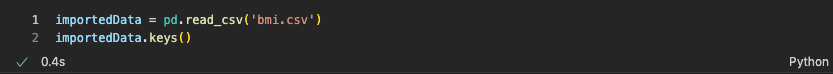


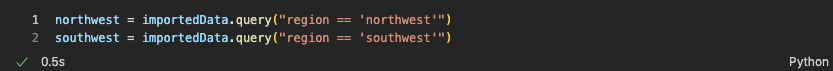
1. *Найти дублирование данных. Удалить дубликаты.*



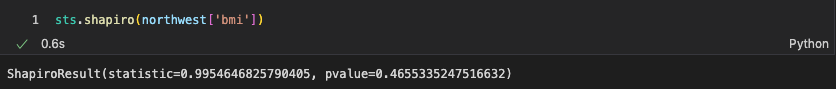


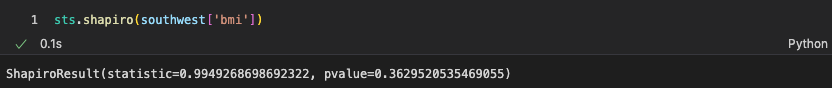
1. *Загрузить данные из файла “ bmi.csv”. Взять оттуда две выборки. Одна выборка – это индекс массы тела людей c региона northwest, вторая выборка – это индекс массы тела людей с региона southwest. Сравнить средние значения этих выборок, используя t-критерий Стьюдента. Предварительно проверить выборки на нормальность (критерий Шопиро-Уилка) и на гомогенность дисперсии (критерий Бартлетта).*





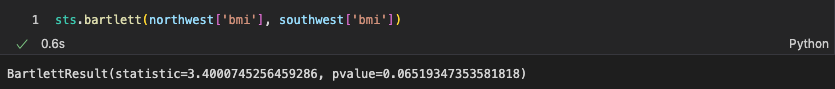
**Тест Шапиро-Уилка.** Нулевая гипотеза состоит в том, что выборка соответствует нормальному распределению, альтернативная же гипотеза исключает принадлежность выборки к нормальному распределению.





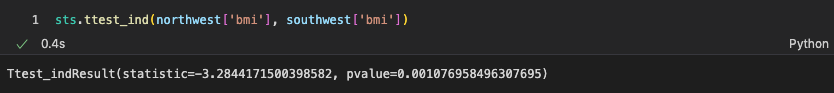
Поскольку у обоих выборок p-критерий больше 0.05 то теорию о том, что выборки соответствуют нормальному распределению можно рассматривать далее.

**Критерий Бартлетта.** Нулевая гипотеза предполагает, что рассматриваемые выборки получены из генеральных совокупностей, обладающих одинаковыми дисперсиями. Альтернативная гипотеза отвергает это.



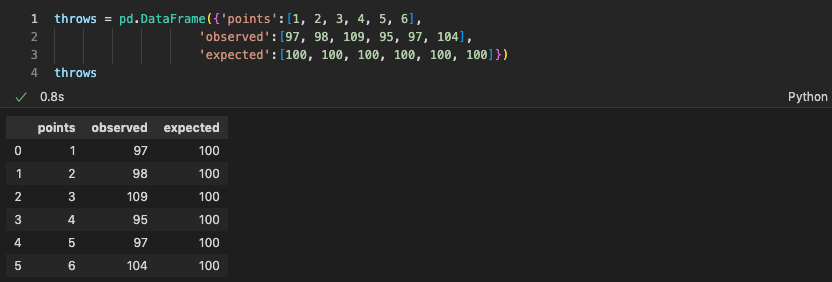
Поскольку p-значение превышает 0.05, можно сделать вывод что дисперсии выборок примерно одинаковы и нулевую гипотезу можно рассматривать далее.

**t-критерий Стьюдента.** Нулевая гипотеза – выборки принадлежат разным генеральным совокупностям, альтернативная теория – выборки принадлежат одной генеральной совокупности.



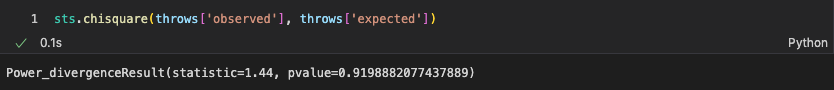
Поскольку p-значение значительно меньше 0.05, можно сделать вывод, что различия сравниваемых величин незначительны и их средние значения схожи, а соответственно нулевая гипотеза отклоняется.

1. *Кубик бросили 600 раз, получили следующие результаты:*



1. *С помощью критерия Хи-квадрат проверить, является ли полученное распределение равномерным. Использовать функцию scipy.stats.chisquare().*

**Хи-квадрат.** Нулевая гипотеза состоит в том, что распределение равномерно, альтернативная же наоборот, это отрицает.



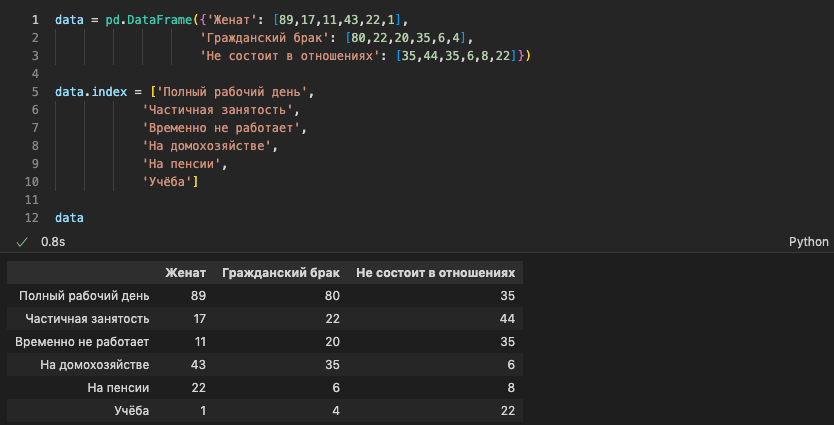
Основываясь на результатах теста (p-значение > 0.05) принимается нулевая гипотеза, соответственно можно сделать вывод, что распределение равномерное.

1. *С помощью критерия Хи-квадрат проверить, являются ли переменные зависимыми. Создать датафрейм, используя следующий код:*

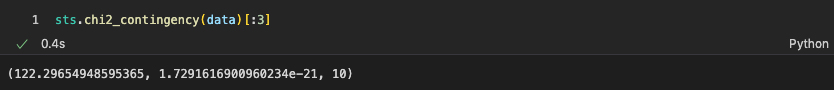
*data = pd.DataFrame({'Женат': [89,17,11,43,22,1], 'Гражданский брак': [80,22,20,35,6,4],*

*'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?*



**Хи-квадрат.** Нулевая гипотеза состоит в том, что распределение равномерно и занятость не зависит от семейного положения, альтернативная же наоборот, это говорит что она зависит.



Поскольку p-значение гораздо меньше 0.05, нулевую гипотезу можно отбросить, и соответственно, можно сделать вывод, что семейное положение влияет на занятость