DS 203 : Programming for Data Science
Tutorial and Assignment Sheet – 6
Basic Statistics

Submission guidelines:

- Refer to the video lecture, https://online.stat.psu.edu/stat415/ and https://www.itl.nist.gov/div898/handbook/prc/prc.htm for theory
- Quick reference of tests: http://www.biostathandbook.com/testchoice.html and https://www.scribbr.com/statistics/statistical-tests/
- Package reference: https://scipy-lectures.org/packages/statistics/index.html
- Prepare an ipython notebook and name it <roll no.>.ipynb and submit it on Moodle before 11:59pm on September 30, 2020.

1. For the data source at https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016, visualize the following data with the appropriate type of graph, and use the right options to make the graph look good (such as using legends and axes titles with the legible font size, and exploring color palettes):
   a. Pick six populous countries (population > 10 million), and for their suicide per 100,000 people over the same set of years (i.e. do not select countries with inadequate number of years), compute the point-estimates (data-driven) of their means and variances using code.
   b. Plot Q-Q plots for these countries to check if the yearly rates of suicide are Gaussian distributed, as six different single variables.
   c. Compare the log-likelihoods of the data for each country picked in the previous part with respect to a Gaussian distribution with the computed means and variances. Is there a relationship between the visual obtained from Q-Q plot and the log-likelihood?
   d. Determine 95% C.I. of the mean yearly suicide rates for each of these countries.
   e. Pick two countries with the closest mean suicide rates. Using Welch's t-test and Wilcoxon signed-rank test, confirm if the mean suicide rates really different from each other.
   f. Repeat the previous part with an appropriate paired test (paired by year).
   g. For all countries, compute the yearly suicide rate (one number per country per year), and explore its correlation with human development index (HDI) and GDP per capita. Use linear regression or correlation and determine its significance.

2. Repeat the exercise for data at URL https://www.kaggle.com/jmmvutu/summer-products-and-sales-in-ecommerce-wish. Here, select variables and relations of your choice. Add one more part (h):
   h. For two discrete variables, test an appropriate hypothesis with chi-squared test.