Problem Set 2: Wrangling Subway Data

Name	File
Problem 2.1	Number_of_Rainy_Days.py
Problem 2.2	Temp_on_Foggy_and_Nonfoggy_Days.py
Problem 2.3	Mean_Temp_on_Weekends.py
Problem 2.4	Mean_Temp_on_Rainy_Days.py
Problem 2.5	Fixing_Turnstile_Data.py
Problem 2.6	Combining_Turnstile_Data.py
Problem 2.7	Filtering_Irregular_Data.py
Problem 2.8	Get_Hourly_Entries.py
Problem 2.9	Get_Hourly_Exits.py
Problem 2.10	Time_to_Hour.py
Problem 2.11	Reformat_Subway_Dates.py

Problem Set 3: Analysing Subway Data

Name	File
Problem 3.1	Exploratory_Data_Analysis.py
Problem 3.2	Welch's t-Test (only questions) **Does entries data from the previous exercise seem normally distributed?** No **Can we run Welch's T test on entries data? Why or why not?** The distribution of the samples according to the histogram, is not normal. The Shapiro Test proofs the same. The Welch's T test is not appropiate for this dataset because the samples have not a normal distribution.
Problem 3.3	Mann_Whitney_U_Test.py
Problem 3.4	Ridership on Rainy vs. Nonrainy Days (only questions) **Is the distribution of the number of entries statistically different between rainy and non rainy days?** Yes. **Describe your results and the methods used.** Using the Mann-Whitney U Test I received a Uequal to 1924409167.0. The p values is 0.024999912793489721 (one tail), also the two tail p value is 0.049999826 which is less than 0.05. The results indicates that there is a significant difference between the two samples.
Problem 3.5	Linear_Regression.py
Problem 3.6	Plotting_Residuals.py
Problem 3.7	Compute_R2.py

Problem Set 4: Visualising Subway Data

Name	File
Problem 4.1	Visualization_1.py
Problem 4.2	Make_Another_Visualization.py

Problem Set 5: MapReduce on Subway Data

Name	File
Problem 5.1	riders_per_station_mapper.py riders_per_station_reducer.py

Problem 5.2	ridership_by_weather_mapper.py ridership_by_weather_reducer.py
Problem 5.3	busiest_hour_mapper.py busiest_hour_reducer.py

Links:

https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test

http://www.statisticsviews.com/details/feature/5722691/Getting-to-the-Bottom-of-Regression-with-Gradient-Descent.html

https://en.wikipedia.org/wiki/Nonparametric_regression

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html

https://www.statisticssolutions.com/mann-whitney-u-test-2/

https://ggplot.yhathq.com/docs/geom_line.html