Udacity NanoDegree: Data Analysis
Project 1: Intro to Data Science

Submitted by: Rakesh Dhote
Email: rakesh.dhote@gmail.com

Problem statement: Figure out if more people ride the subway when it is raining versus when it is not raining.

Section 1. Statistical Test

1.1 Which statistical test did you use to analyse the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis? What is your p-critical value?

The two populations for the analysis are defined as:
rides_rain: people ride the subway when it rain
rides_no_rain: people ride the subway when it does not rain

The Mann-Whitney U-test is used for the analysis.
A two-tailed test is used for the analysis.
The null hypothesis for the analysis is
Ho: Two populations (rides_rain and rides_no_rain) have same population mean (people ride subway equally when it is raining or not raining)
The significance level of 95% is used for the analysis thus the p-critical = 0.05.

1.2 Why is this statistical test applied to the data set? In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.

The Shapiro-Wilk test for normality of the data (rides_rain and rides_no_rain) reveal that the data is not normally distributed (refer to the figure in Section 3.1) for the distribution of the data). The Mann-Whitney U-test is an appropriate statistical test for such non-parametric distribution of the data.

1.3 What results did you get from this statistical test? These should include the following numerical values: p-values, as well as the means for each of the two samples under test.

The statistical test results are as follows:
    Mann-Whitney U-test statistic =153635120.5
    P-value = 5.5e-06 (Using Scipy function the p-value returned is NaN. The p-value is calculated following the procedure in [1])
The means of the two samples are
    mean(rides_rain) = 2028
    mean(rides_no_rain) = 1846.

1.4 What is the significance and interpretation of these results?

As the p-value (5.5e-06) < p-critical (0.05), the null hypothesis is rejected. This implies that the two samples (rides_rain and rides_no_rain) people do not ride subway equally [2]. In addition, it is observed from the mean samples that the mean ridership on the rainy days (2028) is higher than non- rainy days (1846).

Section 2. Linear regression

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:

1. Gradient descent (as implemented in exercise 3.5)
2. OLS using Statsmodels
3. Or something different?

The regression was carried out using the Ordinary Least Square (OLS).

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

The OLS regression analysis was conducted by fitting different features by trial-and-error method. The 'rain', 'hour', 'weekday' are selected as features for the model.
The 'UNIT' is selected as the dummy variable [3].

2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that the selected features will contribute to the predictive power of your model.

- Your reasons might be based on intuition. For example, response for fog might be: "I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often."
- Your reasons might also be based on data exploration and experimentation, for example: "I used feature X because as soon as I included it in my model, it drastically improved my R^2 value."

The reasons for the choice of the features are as follows:

rain: If it rains, more people may decide to use subway to avoid traffic congestion and accidents to minimize strain and time of travel.

hour: This feature is selected as it improved the R^2 value.

weekday: If it is weekday (Mon-Fri), more people may decide to use the subway to avoid traffic congestion, accidents to minimize strain and time of travel.

UNIT: The dummy variable drastically improved the R^2 value.

2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?

The coefficients for the non-dummy features obtained using the OLS method for regression are

θ_0 (constant) = 1886.7228
θ_1 (rain) = 15.4661
θ_2 (hour) = 856.2425
θ_3 (weekday) = 441.2410

2.5 What is your model's R^2 (coefficients of determination) value?

The coefficients of determination R^2 value is 0.481.

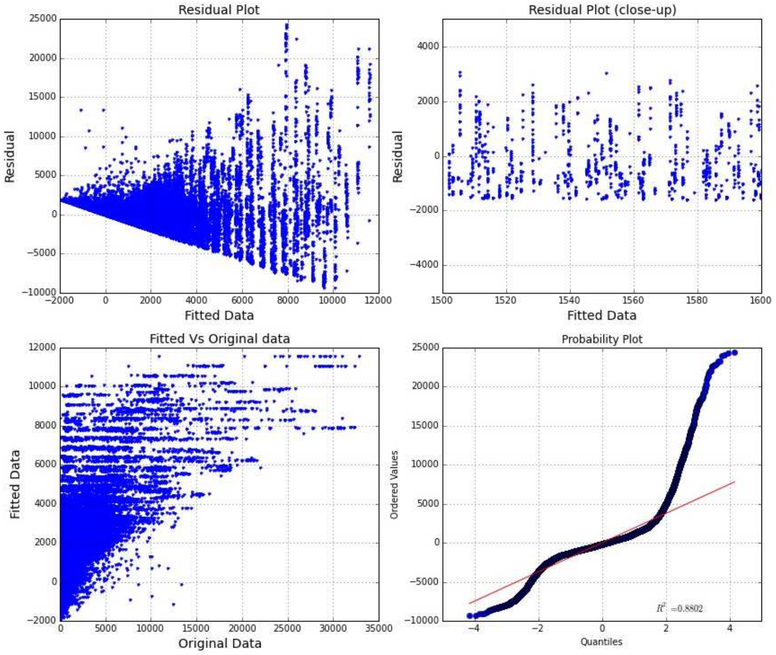
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

The R^2 value is a statistical measure of how close the data is to the fitted line. The R^2 value of 1 indicates the fitted model explains all variability of the response data around its mean. The R^2 value of 0 indicates none of the variability in the response data around its mean. As the R^2 value (0.481) is nearly midway, the model captures some variability of the response data around the mean. In other words, around 51.9% of the variation is ridership is accounted for by variables that weren't included in the regression model.

In addition to the R^2 value, additional diagonostic plots are presented here. The residual plot of the multilinear regression (top left) is not random as expected. The residual is neither centered to zero and nor dispersed symmetrically throughout the range of fitted values. Because of the long-tailed histogram, the residuals at large fitted value are very high! The close-up of the residual plot (top right) between 1500-1600 fitted value range indicates systematic, low and high. Thus non-random patterns in the residual indicate biased results [7].

The plot of fitted data-original data (bottom left) also does not show a symmetric dispersion, thus indicating that the fit is not good.

The check for normality of residuals is plotted in the Q-Q plot (bottom right). The plot does not follow a straight line as expected for the linear regression fit. The"S" shaped curve indicates that the fitted distribution is more skewed than the data.



Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like. Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

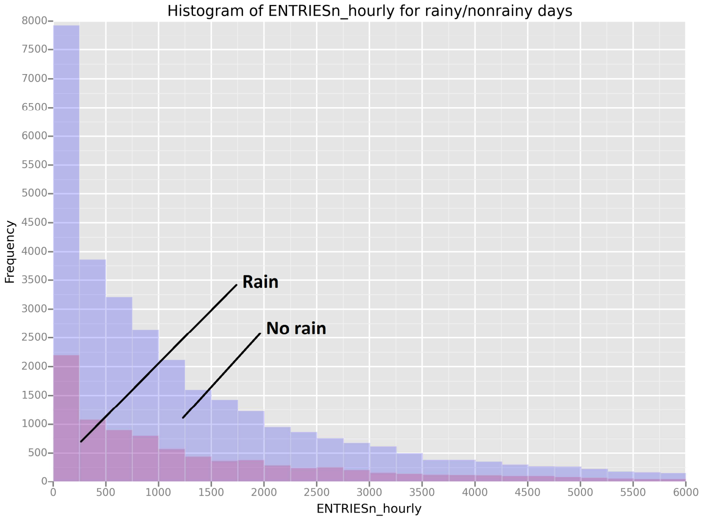
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that falls in this interval.
- Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.

The binwidth = 250 is used for the following plot. (Note: There is a bug in the ggplot in labeling the plots).

From the graph it can be observed that the frequency count of number of riders using subway when it does not rain is higher than when it does rain. This is due to the fact that the number of non-rainy days are higher (33,064) than the rainy days (9,585).

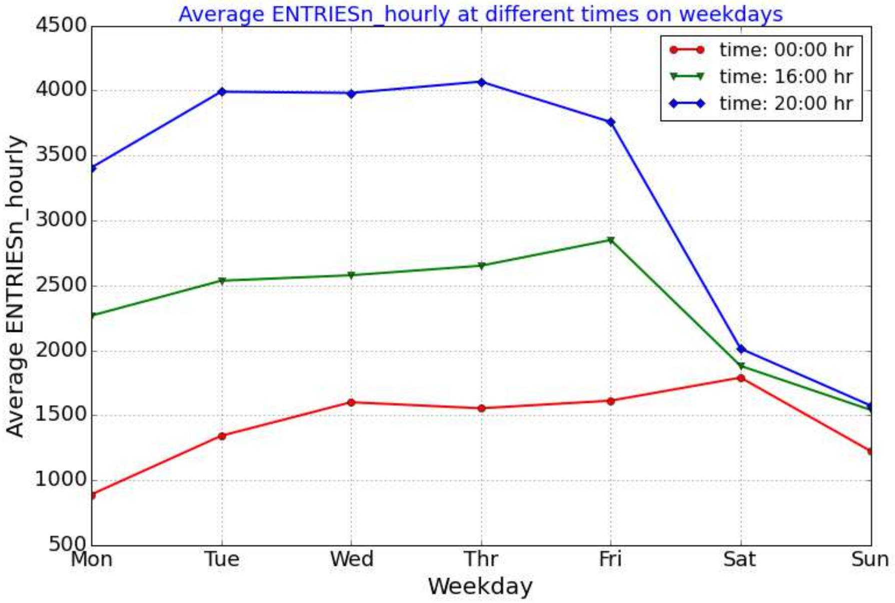
The frequency of ridership asymptocially decreases for higher ENTRIESn_hourly.



3.2 One visualization can be more freeform. Some suggestions are:

- Ridership by time-of-day or day-of-week
- Which stations have more exits or entries at different times of day

The following scatter plot shows the Average ENTRIESn_hourly at different times on weekdays in the month of May 2011. Only three time stamps are chosen for plotting.



The following observations are made from the above plot

- Average ENTRIESn_hourly at 16:00 and 20:00 hrs are higher from Monday to Friday than the weekend (Saturday and Sunday)
- Average ENTRIESn_hourly at 00:00 hours is highest on Saturday.

Section 4. Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

4.1 From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining or when it is not raining?

From the statistical analysis and linear regression, it is concluded that the people ride the NYC subway more when it rain.

4.2 What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.

We started with a premise that rain may cause more ENTRIESn_hourly. The hypothesis testing rejects the null hypothesis inferring that the riderships on rainy and non-rainy days are different. From the linear regression, it can be seen that the coefficient for feature ‘rain’ is positive indicating that ridership on the rainy days is higher than non-rainy days (by ~15). In addition, the mean ridership on rainy days (2028) is higher than non- rainy days (1846) indicating higher ridership on rainy days.

Section 5. Reflection

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1. Dataset,
2. Linear regression model,
3. Statistical test.

Dataset: On summarizing the improved-dataset in the Pandas dataframe, the following observations are made:

- The data range from 5/1/2011 to 5/31/2011
- The time entries are at the particular time slots (0, 4, 8, 12, 16 and 20 hrs.) on each day.
- Maximum precipitation ('precipi') = 0.3 inch rain
- Maximum mean precipitation ('meanprecipi') = 0.1575 inch.

The summary indicates that the data set is taken for short periods of one month. It seems that the May 2011 may be relatively dry in NY for the people to ride the subway [5]. It may also be possible that the rain may have taken place between the time slots of measurement. Some of the entries at time slots (ex. At 8:00 hrs) are missing for several days. The data may miss the rain during those missing times.

Linear Regression model:

As the R^2 value using the OLS method is 0.481, it indicates that the model capture 48.1% variance in data with the selected features. Thus, 51.9% of the variation is ridership is accounted for by variables that weren't included in the regression model. The additional diagonostic residual and Q-Q plot indicate that the linear regression is a modest choice in such case. A polynomial regression or finding additional features in the dataset may help better prediction.

Statistical test:

The Mann Whitney U-test for the non-parametric distribution of the data in the above case need to be supplemented with additional descriptive statistic like median, mean and interquarantile range for the conclusions.

In general, the statistical significance test signifies that the statistic is reliable. It does not indicate that the findings are important or it has any decision making ability.

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?

In addition to insights shared in the answer to the Dataset in section 5.1, the following suggestions may help in taking

- frequent readings (ex. Every 15/30/45 mins or one/ two hour interval), and
- readings over few months when it rains and perhaps if feasible over a few years.

References:

[1] Calculation of the p-value for Mann-Whitney U-test: http://stats.stackexchange.com/questions/116315/problem-with-mann-whitney-u-test-in-scipy

[2] Mann-Whitney U-test: http://www.stat.ucla.edu/~rgould/x401f01/mannwhitney.html

[3] Dummy variable - https://www.moresteam.com/whitepapers/download/dummy-variables.pdf

[4] Goodness of fit: http://www.graphpad.com/guides/prism/6/curve-fitting/index.htm?r2_ameasureofgoodness_of_fitoflinearregression.htm

[5] NY weather: http://www.nycbynatives.com/visitors_center/nyc_weather.php

[6] Mann-Whitney U-test: http://www.stat.ucla.edu/~rgould/x401f01/mannwhitney.html

[7] Regression analysis: How to interpret R^2 and assess the goodness of fit: http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit