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 = 0.025.

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) = 2028.2
mean(rides_no_rain) = 1845.5.

1.4 What is the significance and interpretation of these results?

As the p-value (5.5e-06) < p-critical (0.025), the null hypothesis is rejected. This implies that the two samples (rides_rain and rides_no_rain) people do not ride subway equally [2]. From the

mean samples, it is observed that the mean of ridership with it rain (2028.2) is higher than without rain (1845.5). This infer that the people ride subway more when it does not rain.

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?

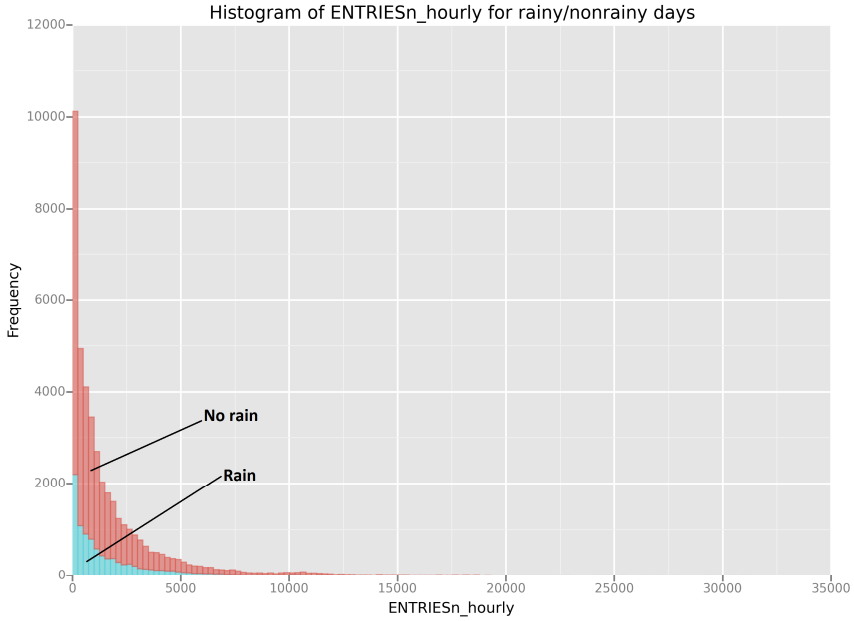
The closer the R^2 value to 1 indicate the regression model predicts the linear relationship between independent and dependent variables. On the other hand, the closer the R^2 value to 0 indicate the regression model does not predict the linear relationship between independent and dependent variables. As the R^2 value (0.481) is nearly midway, it indicates that the relationship between the features and values is fairly linear [4]. In other words, the ridership can be predicted with 48.1% confidence for a combination of rain, hour, and weekday.

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.

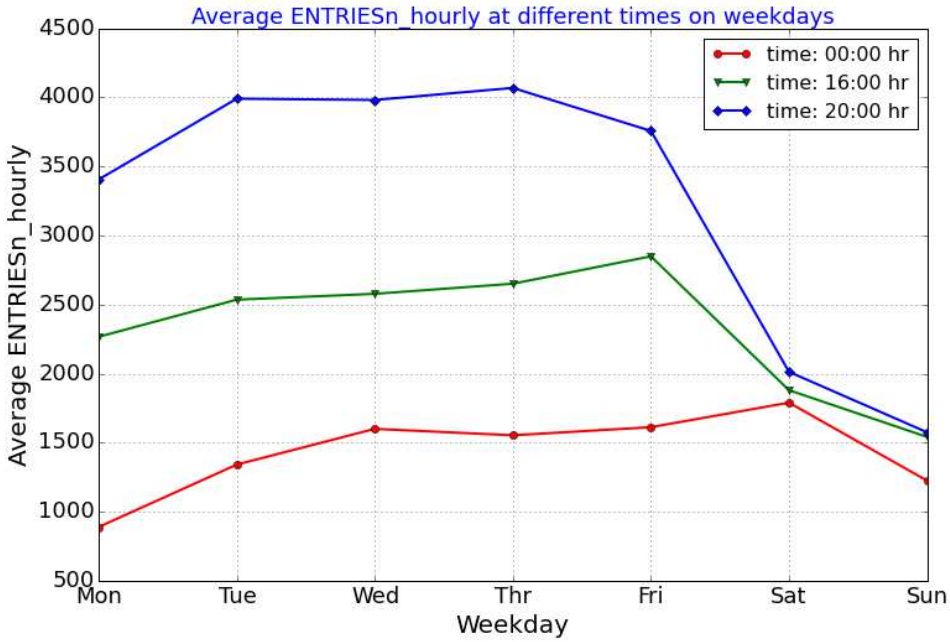
The binwidth = 250 is used for the following plot. (Note: There is a bug in the ggplot in labeling the plots).
From the graph it can be observed that the number of riders using subway when it rains is less than when it does not rain. This is contrary to our premise that riders may use the subway when it rains.
The frequency of ridership is high for ENTRIESn_hourly below ~ 5000 and decreases asymptocially 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?

We started with a premise that rain may cause more ENTRIESn_hourly. The hypothesis testing rejects the null hypothesis inferring that the riderships are different for when it rains and when it doesn't. The mean values of ENTRIESn_hourly when it rain is higher than without rain. This indicates that the riders prefer to use subway when it does not rain.
In addition, observing the histogram plot reveal that the ENTRIESn_hourly when it rains has distribution with lower frequency than when it does not rain.
From the statistical analysis and data visualization, it is concluded that the people ride the NYC subway more when it does not 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.

The Mann-Whitney U test lead to the conclusion that the null hypothesis is rejected. This implies that the two samples (rides_rain and rides_no_rain) people do not ride subway equally.
In addition, the linear regression indicates the coefficient/weight of 'rain' is θ_1 (rain) = 15.4661 which is less than θ_2 (hour) = 856.2425 and θ_3 (weekday) = 441.2410. Thus the ENTRIESn_hourly has weak relationship with the 'rain' variable and stronger relationship with hour and weekday.

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 relationship between the dependent and independent variables is not strongly linear. The linear regression is a modest choice in such case. Advanced machine learning algorithms can be useful to obtain further insights from the data.
The other limitations of the linear regression are:

- It looks at the mean of the dependent variable without looking at the extremes in the values
- The outlier cause the increase/decrease in weight of the coefficients.

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