Analyzing the NYC Subway Dataset

Questions

Overview

This project consists of two parts. In Part 1 of the project, you should have completed the questions in Problem Sets 2, 3, 4, and 5 in the Introduction to Data Science course.

This document addresses part 2 of the project. Please use this document as a template and answer the following questions to explain your reasoning and conclusion behind your work in the problem sets. You will attach a document with your answers to these questions as part of your final project submission.

Section 0. References

Please include a list of references you have used for this project. Please be specific - for example, instead of including a general website such as stackoverflow.com, try to include a specific topic from Stackoverflow that you have found useful.

http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
http://stats.stackexchange.com/questions/53053/mann-whitney-or-two-tailed-t-test
http://blog.minitab.com/blog/adventures-in-statistics/choosing-between-a-nonparametric-test-and-a-parametric-test
http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
http://www.clockbackward.com/2009/06/18/ordinary-least-squares-linear-regression-flaws-problems-and-pitfalls/

Section 1. Statistical Test

1.1 Which statistical test did you use to analyze 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?

- *Mann–Whitney U test*
- *p-critical = 0.05 (two tailed)*
- *H0 : P(x > y) = 0.5*
- *H1 : P(x > y) != 0.5*

  where x is the subway riders during hours with rain and y is subway riders during hours with no rain

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

- *The **distribution** of the 2 samples are not normal, both samples have a pareto distribution. The Mann-Whitney U test is more efficient on non-normal distributions.*
- *In a pareto distribution the median is more meaningful than the mean. This kind of distribution increases the bias of outliers to the mean. The **median is a better measure**. In these samples, the mean for hours with rain and hours without rain are 1105.45 and 1090.28 respectively whereas the medians are 282 and 278 respectively. This illustrates the strong effect of the outliers in the sample on the mean which is not representative of the centre of the distribution.*
- *Both groups have **sufficient size** to perform a nonparametric test*
- *The **sample sizes are not equal** where size of the rain sample is half (33% of the total data) of the size as the no rain sample (66% of the total data)*
- *A **two tailed test** is appropriate because we do not know if rain will increase or decrease ridership.*

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 number of subway riders was greater during hours where there was rain (mean = 1105, median = 282) than for hours without rain (mean = 1090, median = 278), U =1924409167, p = .05

1.4 What is the significance and interpretation of these results?

The significance level was p = .05. This means that, assuming that the null hypothesis is true, if the samples were randomly sampled from the population again, there is a probability of 5% that the same results will be achieved again. In other words, there is a probability of 5% that our results were purely from random chance and that rain does not effect how many people use the subway in New York.

With p = .05 it is unlikely that these results were achieved through random chance. Therefore, we can reject the null hypothesis.

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:
OLS

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?
Features: rain, precipi, Hour
Dummy - UNIT

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.”

Hour*: The time of the day greatly influence when people use the subway. There are certain times in peak times in the day for subway usage eg. commute times and lunch time.*
rain*: I would assume that people would be more likely to use the subway when it is raining instead of walking or cycling. Including his feature greatly improved my R^2 value*
precipi*: A large amount of rain should influence people to use the subway over other forms of transport compared to a small amount of rain*
Unit*: This was used as a dummy variable because different stations will have different number of entries due to other factors such as physical location and number of subway lines it has access to. Therefore, data related to each station is independent from each other and cannot be analysed together.*

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

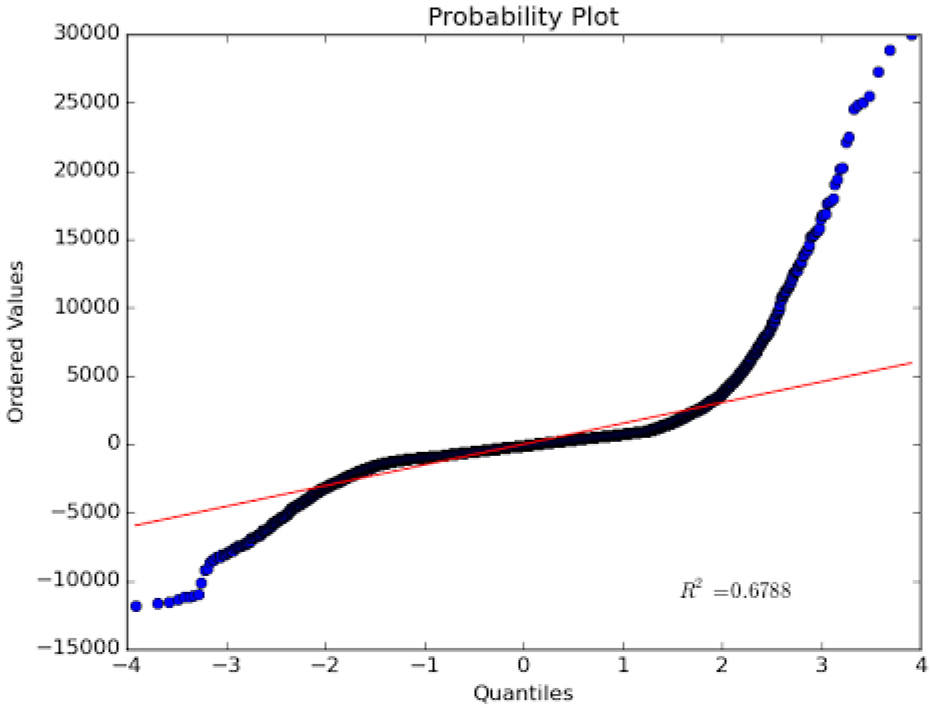
rain	*34.331402*
Hour	*62.309276*
precipi	*59.811168*

2.5 What is your model’s R^2 (coefficients of determination) value?
0.482478577098

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?

A R^2 value of 0.48 for a model to predict ridership indicates good fit, the model explains 48% of the original variability but there are still 52% unexplained residuals. Human behaviour is difficult to predict so a lower R^2 value is acceptable. If we only had the R^2 to determine the fit of the model then 0.48 is good enough to demonstrate good fit due to the unpredictability of human behaviour. For predictions that require a high degree of accuracy, the R^2 indicates that this model may not have high enough precision to be used to predict ridership.

In addition to the R^2 value the residual plots should also be examined to evaluate fit. A probability plot of the residual values against a theoretical normal distribution was used to determine if the residuals were well behaved. The centre quantiles agree, but the lower and upper quantiles do not agree - both exhibiting large residuals. The residuals are not normally distributed with large residual values at both tails. The residuals are too large to accept the model. Although we had earlier concluded that the model was a marginally good fit based on the R^2 value, the large residuals mean that we can no longer accept the model as a good fit used to predict ridership.



Probability Plot of residual values of Linear Regression Model

Section 3. Visualization

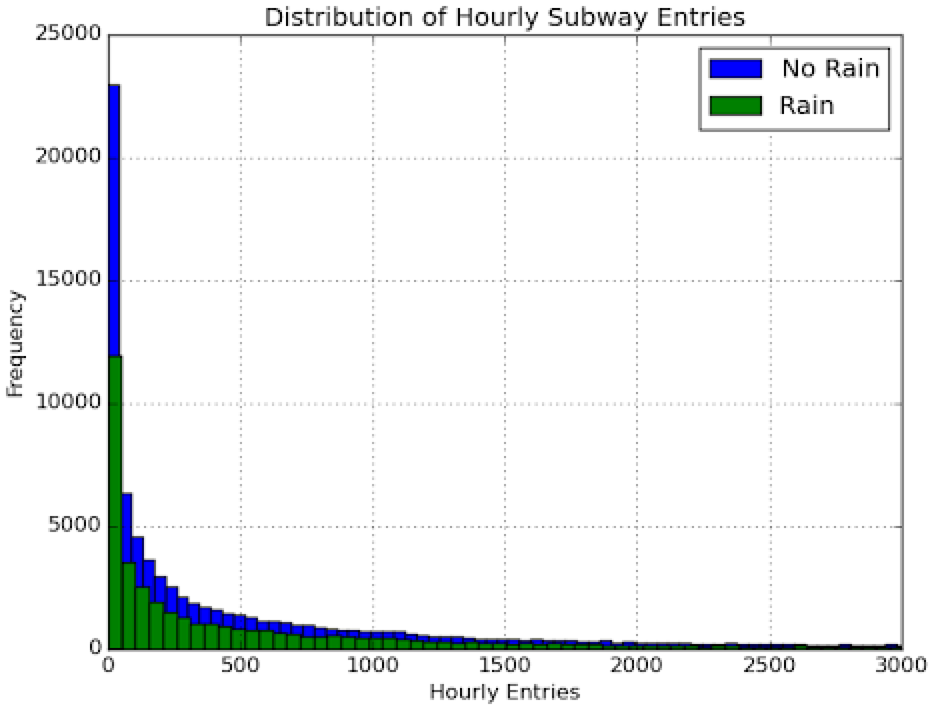
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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.

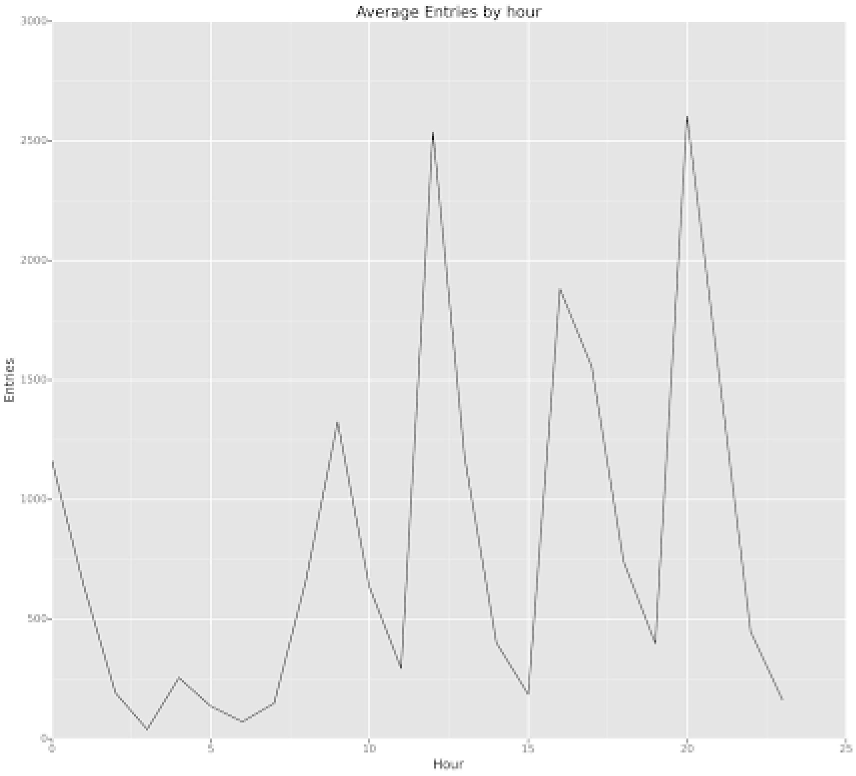
- Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.



- *Both populations exhibit a very similar distribution, both exhibit a pareto distribution.*
- *The "No Rain" population is larger than the "Rain" population*

3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:

- Ridership by time-of-day
- Ridership by day-of-week



On average the busiest hour for entries into subway turnstiles was 8pm followed by 12pm, 4pm, 9am and 12am.

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?
More people use the NYC subway during hours that it is raining compared to hours without 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.

A Mann-Whitney test indicated that the number of entries into the NYC subway system was greater during hours where there was rain (Mdn = 282) than for hours without rain (Mdn = 278), U = 1924409167, p = .05.

A linear regression model calculated to predict the number of entries at hourly intervals into the NYC subway system based on if there had been rain during the hour, the amount of rain and the time of the day. The presence of rain was coded as a boolean value, the amount of rain was recorded as precipitation in inches and the time of the day was coded as the hour number. The linear model was found with R^2 = 0.48, there were also large non-normal residuals illustrated by a Q-Q plot. The residual values was too high and R^2 too low to demonstrate a good fit. So, the linear model cannot be used to contribute to conclusions.

There is a significant (p =.05) probability that more people will use the NYC subway during hours that rained compared to those that did not. Therefore, we can reject H_0 and we can conclude that on hours that have rained, more people use the NYC subway.

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. Analysis, such as the linear regression model or statistical test.

The dataset only contained data from the May 2011 this could have introduced sampling bias where seasonality variables could be a factor. For example, there could have been a number of major events on in May which lead to higher subway use than average, or, more people come to New York City for holidays in May than average . To minimise sampling bias data should be collected for an entire year.

The weather data is aggregated at the daily level whereas the subway data is used at hourly intervals. Statistical calculations were performed on hourly subway data against daily weather data. The weather data is not relevant for the hour which the subway was used. For example, for a given day, there was rain in the afternoon, subway entries earlier in the day cannot be analysed together with this weather data. To increase the validity of the conclusions in this document the weather data will also need to be hourly so that the correct weather conditions can be mapped to the subway data points.

The use of the Mann-Whitney U test has lower power than a parametric test such as Welch's T-Test. If it was appropriate to use the T-test, it may have given us a more accurate result in determining if the two groups were different.

The variables available in the dataset for building a linear regression model were almost all related to the weather. These weather variables are not independent, there is some level of dependence among them. The dependence amongst the independent variables can cause the model to choose a model that is not the most optimal. In the linear regression presented here both rain and precipi were included as independent variables, these are related, since if precepi = 0 then rain is always 0 and if precepi > 0 then rain is always 1.

Outliers in the data will also have a negative effect on the accuracy of the linear regression model, particular using OLS. In the data there is 1 day with precepi = 2.18, the next highest value is 0.89. The high precepi value is an outlier and may have introduced error into the model.

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