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, and 4 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://pandas-docs.github.io/pandas-docs-travis/groupby.html#dataframe-column-selection-in-groupby
http://ggplot.yhathq.com/docs/index.html
http://influentialpoints.com/Training/Wilcoxon-Mann-Whitney_U_test_use_and_misuse.html
https://uk.answers.yahoo.com/question/index?qid=20100109093525AAEufIF
http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

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.
Mann-Whitney U Test by default is one tailed, but we can make it two tailed by doubling up the value obtained in the result, so here I am using two tailed test by doubling the p-value got in result.
Null hypothesis was that the ridership is no different on a rainy day then a non rainy day.
P critical value is 0.05
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.
T-test is used to test the mean of two groups of data, while Mann whitney u test is used to test the median between two groups of data. As t-test is dealing in means so when the data is more skewed the mean of the data would tend towards the skewed part of the data so it is more sensitive towards extreme outliers or extreme values so that could make the result more misleading, this condition is more relaxed in the Mann whitney u test and it is more robust against the violations of the assumptions.
t-test is done on these assumptions.

a. The population follows normal distribution.
b. The variance of both populations are identical

Mann-whitney u test is done on these assumptions

a. The populations do not follow any specific parameterized distributions
b. The populations of interest have the same shape

c. The populations are independent of each other

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.
With Rain Mean – 1105.446
Without Rain Mean –1090.278
p-value – 0.05
1.4 What is the significance and interpretation of these results?
From the test we can deduce that both the distribution of entries is statistically different, and the p-value is 0.025 which makes the null hypothesis wrong and states that both distribution are different statistically. The larger values of p builds the confidence in the null hypothesis and as the value here is very low it defies 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:

1. OLS using Statsmodels or Scikit Learn
2. Gradient descent using Scikit Learn
3. Or something different?

I used OLS using Statsmodels

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?
I used Hour, Precipi (rain in mm at that hour) and meanwindspdi (wind speed), also I used the dummy variable UNIT as well
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."

Time at which wind is high with precipitation is happening at that time as well should result in people deciding to use subway for that time for it being safer option.

2.4 What are the parameters (also known as "coefficients" or "weights") of the non-dummy features in your linear regression model?
Hour 65.386178
precipi 40.466678
meanwindspdi 35.792983

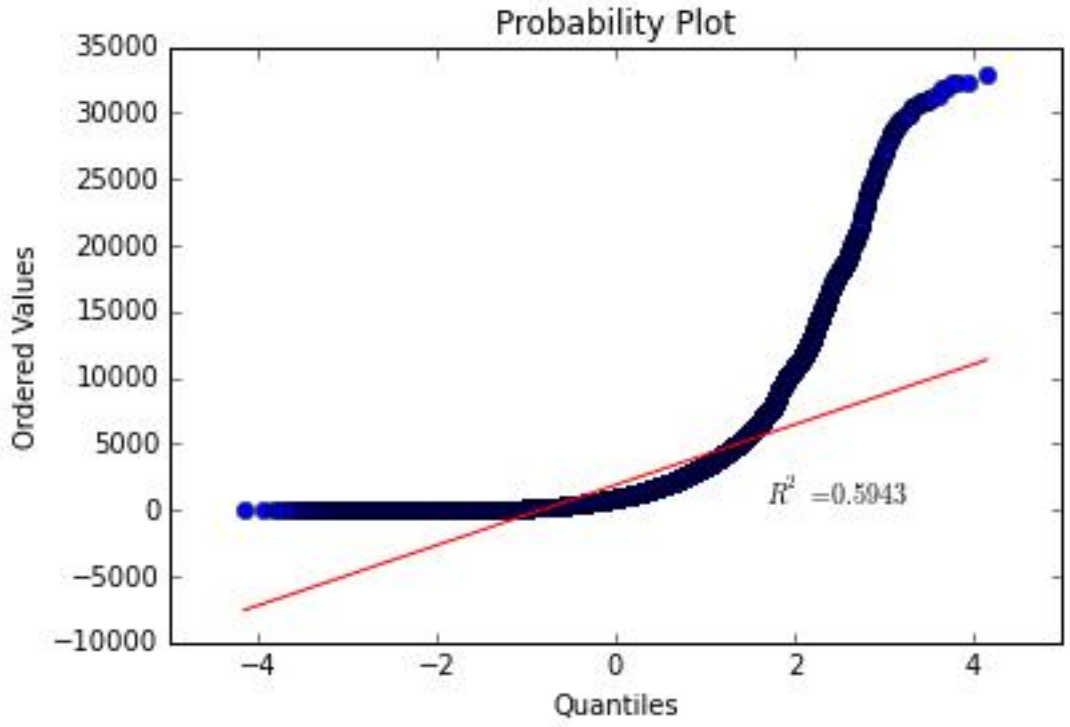
2.5 What is your model's R^2 (coefficients of determination) value?
0.479217847323

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?

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination. It is the percentage of the response variable variation that is explained by a linear model in our case the linear regression model. From our experiment we are plotting our predicted values and the actual values, so the scatter dots in this plane have a 50% variance from the regression line made by our model.

We have chosen linear regression model in our case due to presenting the logic that during rain and high winds would tend to make people to use subway more, so even though our variance is around 50% from the predictions but still it shows that the inclination is there that people do tend to use subway more while it is raining or if the winds are high.

Also if we create the probability plot from the actual it can be seen that the plot describes the non-linearity of the model.

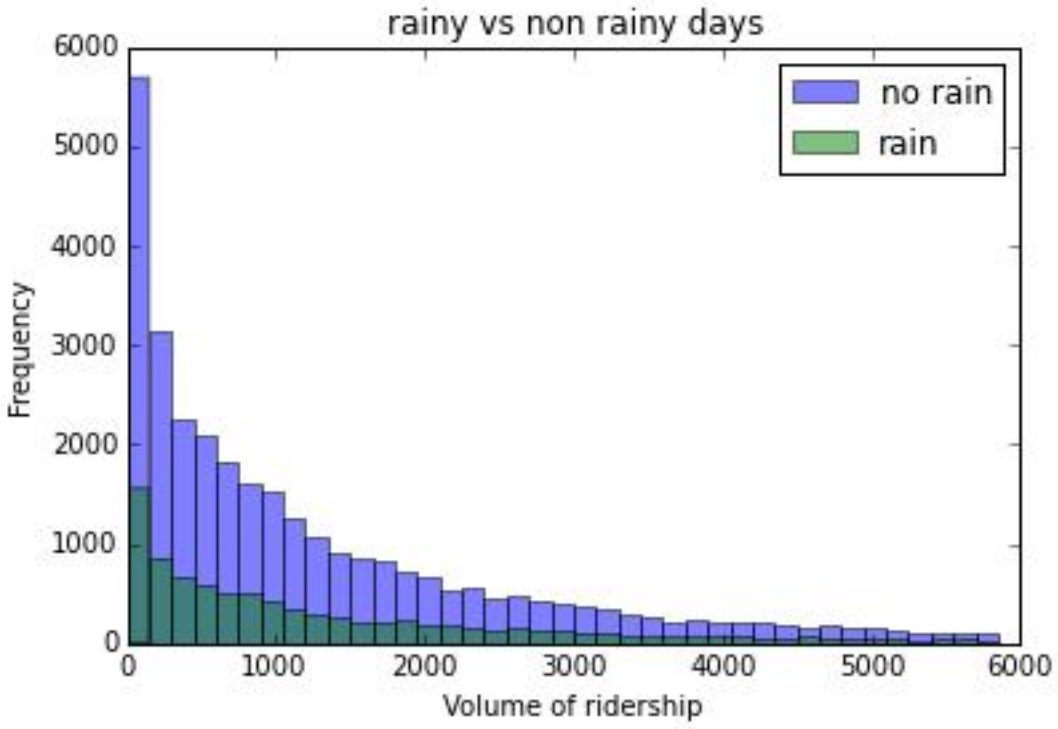


Section 3. Visualization

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.

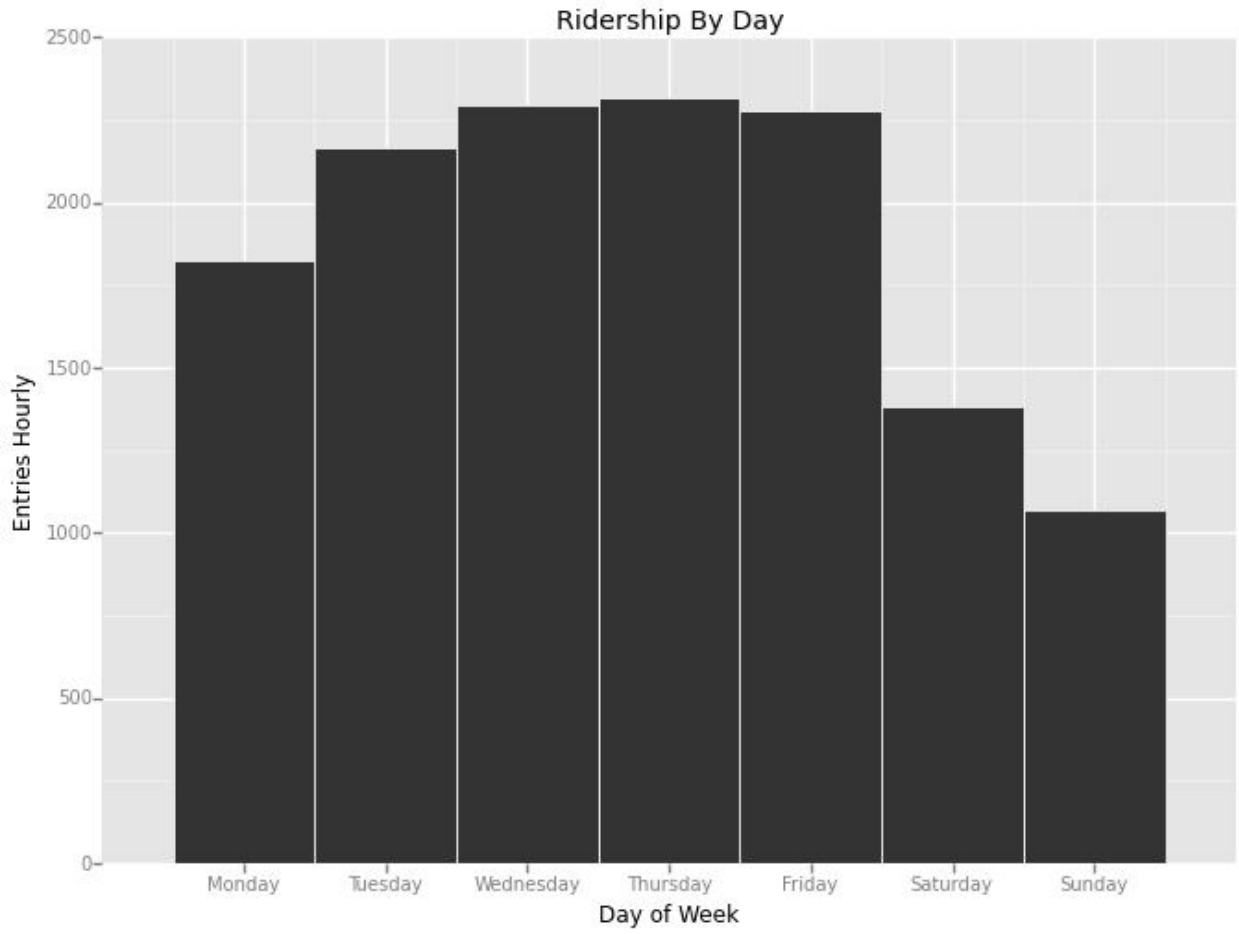
- Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.



This is chart is comparing the mean of volume of ridership from a station on rainy and non rainy days.

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



Ridership is more on weekdays than on weekends and maximum on Monday while minimum on Sunday

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 ride subway when it is raining, this can be concluded from the Mann Whitney U test and the R^2 value test.

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.

From the Statistical Mann Whitney U we got these results –

With Rain Mean – 1105.446

Without Rain Mean –1090.278

p-value – 0.025

So from the p value we can say that null hypothesis which is that the median of the two groups is significantly different as our p value is 0.025 which makes the median of without rain group out of the 95% confidence interval or less than 0.05 p-critical value.

Linear regression analysis

Coefficients for the features are –

Hour 65.386178

precipi 40.466678
meanwindspdi 35.792983

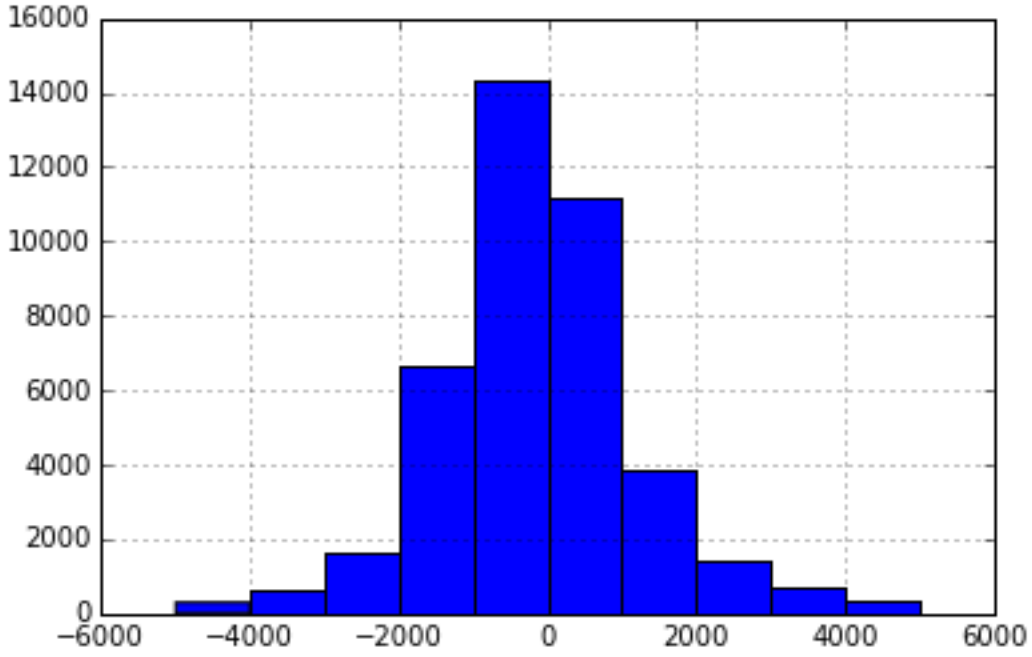
Positive coefficient states that if the precipitation or wind increases then there is positive increase in the ridership.

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.

We are only covering few parameters here like rain, fog or wind but many other factors can also contribute to the analysis, for e.g. if there is some road blockage near a certain station people living around that area would tend to take subway that day to get away from the traffic jam, also the ridership can also be divided between categories, some of the riders have only subway as an option so if we can somehow get some statistics on how many of these people do have some other options and they choose subway then we can deduce better analysis over our statistics. Also out model could make better predictions if we can include this categorical ridership as a feature as well. From the current data here is residual plot which can be built which deduces that though the predictions are more concentrated towards the actual values but still the plot is not that much narrow.



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