Analyzing the NYC Subway Dataset

Section 0. References

http://statsmodels.sourceforge.net/0.5.0/generated/statsmodels.regression.linear_model.OLS.html

http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm

http://www.cookbook-r.com/Graphs/Scatterplots_(ggplot2)/

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?

I used MannWhitney U-test and p-value for analyzing the NYC subway data.
I used two-tail P value to determine if NULL hypothesis holds good or not.
My NULL hypothesis is that there is no difference in the ridership MEANs between rainy and non-rainy days.
I considered a p-critical value of 0.025 (1-sided)

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.

MannWhitney U-test is used to determine whether the samples are different in non-normal distributions.
As p-value is less than 0.025, the samples are different and can be inferred that the distributions are statistically different between rainy & non-rainy days.
The assumption (NULL Hypothesis) is that both the samples are same. It is disproved based on the calculated P-value.
Other assumptions: 1) Distributions are non-normal 2) Chosen sample is random 3) No dependency between the two samples/distributions.

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.

Mean on rainy days = 1105.4463767458733
Mean on non-rainy days = 1090.278780151855
U = 1924409167.0
p(1-sided) = 0.024999912793489721
p(2-sided) = 0.0499998255869794

1.4 What is the significance and interpretation of these results?

MannWhitney U-test assumes there is no variance between the Means of two distributions. Based on the results, we can see that there is variance between the Means of two samples. Also P-value is less than P-critical value of 0.025 and hence we can interpret that the two samples are statistically different.

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?

Gradient descent (a type of machine learning algorithm) is used to run linear regression on NYC subway data.

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 the following features: rain, precipi, Hour, meantempi, precipi
features = dataframe[['rain', 'maxtempi', 'Hour', 'mintempi', 'precipi']]
Yes, I used dummy_units as a dummy variable and joined it to the features (shown above).
dummy_units = pandas.get_dummies(dataframe['UNIT'], prefix='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."

rain, precipi, mintempi, maxtempi, fog – All these features define weather and I assumed weather will have an impact on the ridership predictions.

Hour – Assumed that the ridership during peak hours could be more and hence chose this feature.

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

[-4.44124847 0.91818241]

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

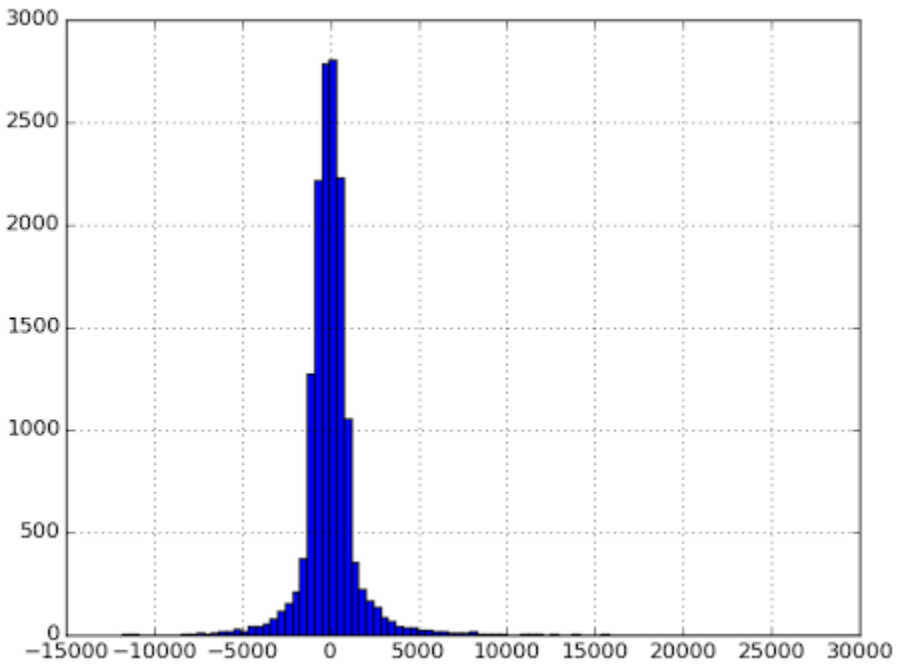
My model's R^2 value is 0.4803

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^2 represents the percentage of variance explained. This model explains only 48.03% of variation. Linear regression model is used to determine how a response (dependent) variable is affected by changes in explanatory (independent) variable. Since we have both predictions and actuals available, After plotting the residuals (which give the difference between actuals and predicted values), I can see that the histogram of residuals has long tails. So, this regression model is not a very good fit.

For checking residual normality, we could plot histogram, probability plot and dot plots. I chose to plot histogram for the residuals.

Below is the histogram plotted for the residuals.



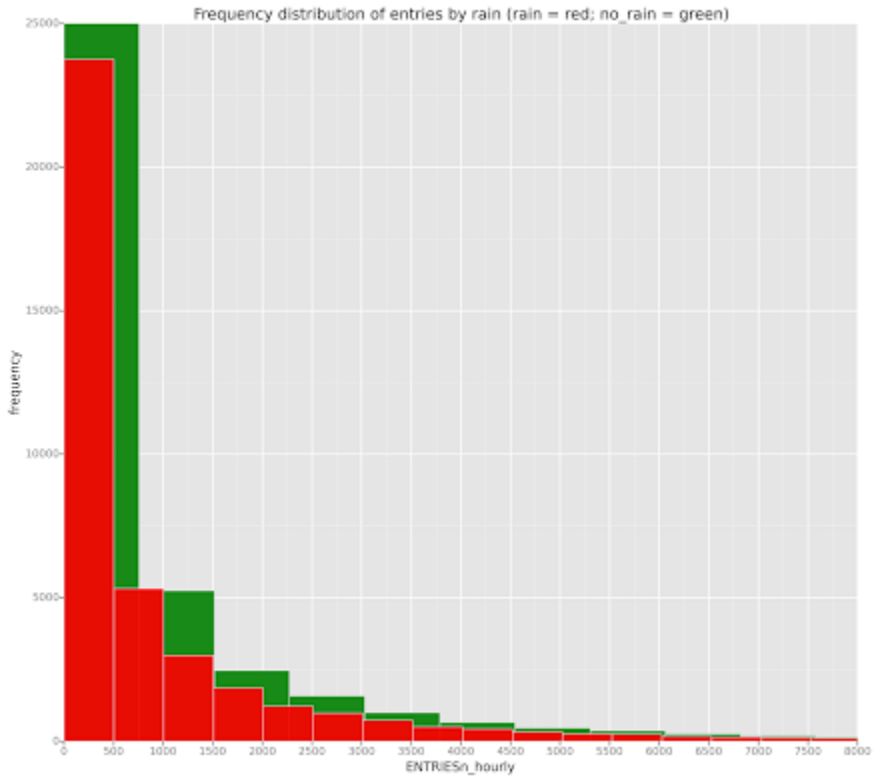
Section 3. Visualization

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.



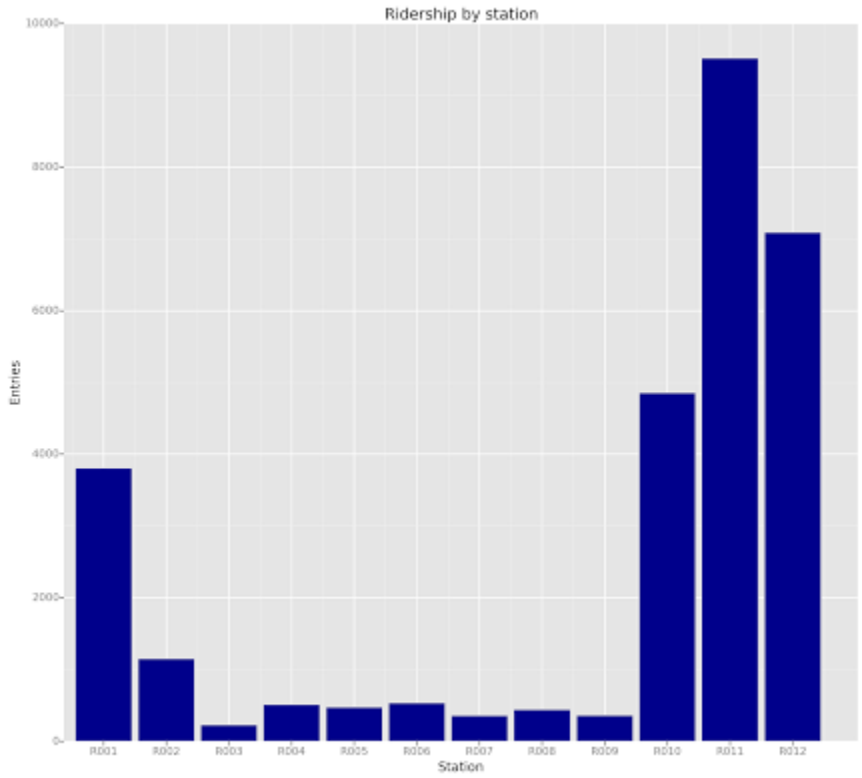
Frequency distribution of entries for rainy and non-rainy days shows that both the distributions are not normal-distributions. The number of rainy days is lesser than the number of non-rainy days and hence we would not be able to conclude from this chart that the ridership during rainy days is less than the ridership during 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:

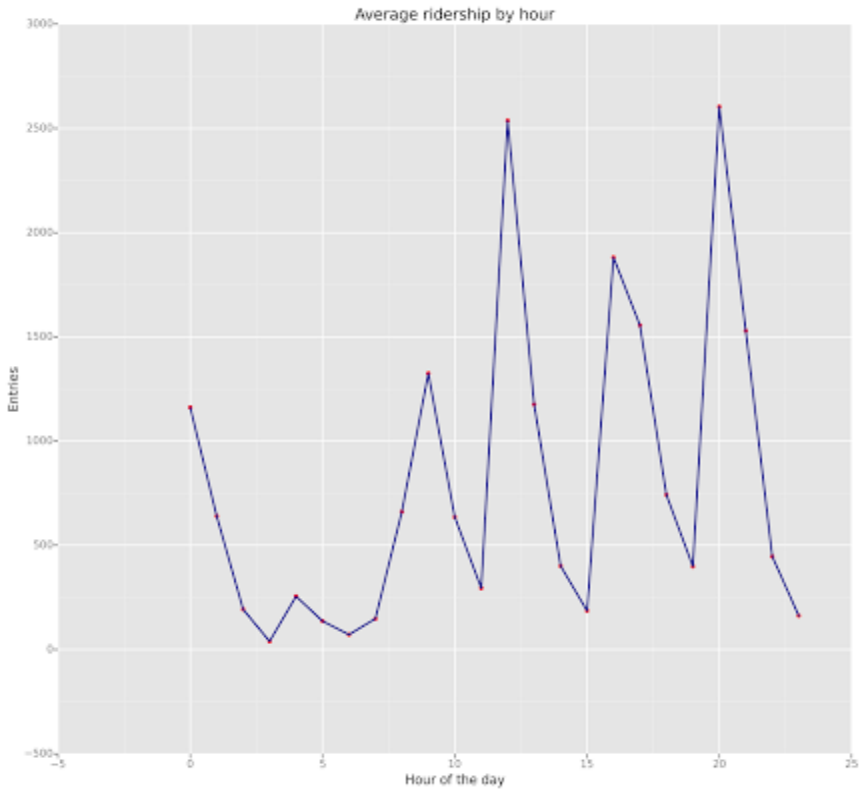
- Ridership by time-of-day
- Ridership by day-of-week

The below chart shows the ridership by subway station

(UNIT). Data considered only for 12 UNITs. R011 has maximum average ridership while R003 has minimum average ridership.



The below chart shows the average ridership by hour of the day. We can see that there are two peaks in the average ridership at 12 PM and 8 PM. This could be because of bad weather or some other demographic which is not captured and hence cannot infer the reason with the current data only.



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 NYC subway when it is raining than when it is not raining.
From the results of MannWhitney U-test (p-value (two-sided) = 0.049), it is evident that both the distributions are statistically different and more people ride NYC subway when it is raining. Though the means of both the data sets do not have a large difference, the distributions are statistically different. Hence, it is important to statistically prove/disprove the hypothesis than inferring intuitively.

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.

MannWhitney U-test assumes there is no variance between the Means of two distributions.

Based on the calculated p-value it is clear that the two distributions (rain & no-rain) are statistically different and mean on rainy days is greater than the mean on non-rainy days.
Also, a positive co-efficient for rain parameter indicates that rain impacts the subway ridership.

In Linear regression, R-squared value is greater when rain feature is considered against not having rain feature in the regression analysis. Hence with an increase in R-squared value when Rain is considered we can say that on rainy day NYC subway ridership would be greater than non-rainy day.

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.

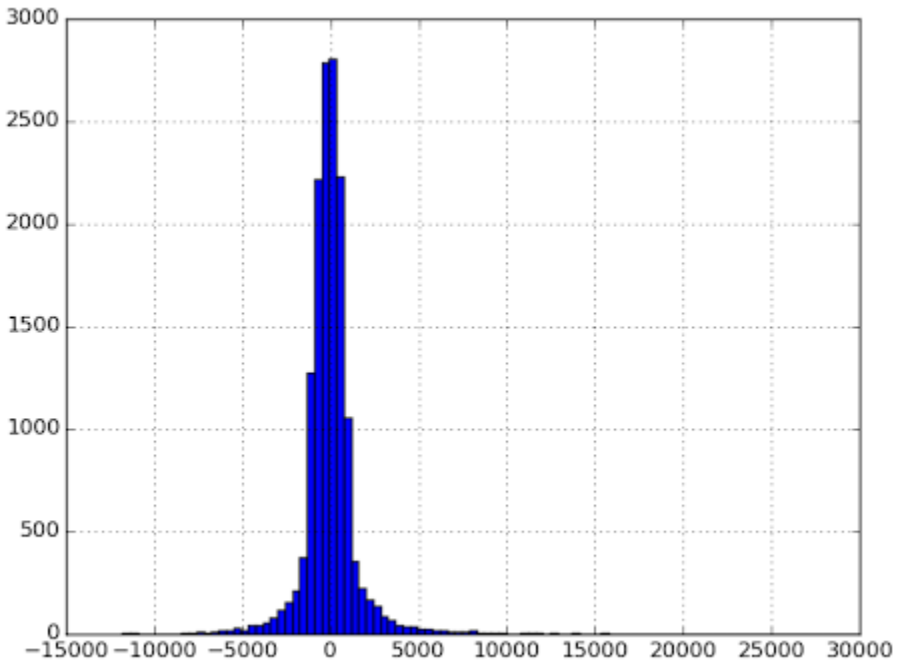
Dataset: Sample drawn corresponds to only one month of data and so it cannot be considered truly random sample. Sample should have been drawn across multiple months and years.
Data doesn't consider other demographics like peak hour rush, population density around each subway station, public work hours, school timings etc, which could also impact the ridership.
To particularly infer if rain increases ridership it would have been better to consider one subway station but draw samples corresponding to this subway station for multiple months and multiple days (both rain & non-rain) and then run statistical tests on this data which could increase our confidence in the statistical test results.
Also as for most of the models only 1/3rd of total turnstile data is used, this might have had some impact on the statistical test results as the sample has become smaller.

Statistical tests: MannWhitney U-test and linear regression tests applied are used for this analysis.
MannWhitney U-test:
This non-parametrical statistical test requires both the datasets to be non-normal distributions and are independent of each other. Based on the obtained p-value and mean values, this statistical test is good enough to infer that the means of rainy days and non-rainy days are different. Hence MannWhitney U-test is a good statistical test for turnstile weather data set.

Linear Regression:
Linear regression is used to determine the effect of dependent variable when there is a change in independent variable. Unlike MannWhitney test, Linear regression doesn't assume anything about the distributions. It only makes assumptions about the residuals (errors) to be normally distributed. Comparing the difference between actuals and predictions and plotting histogram for residuals, I can see that the tails are long and can infer that Linear regression model is not the best fit for this data set.

For checking residual normality, we could plot histogram, probability plot and dot plots. I chose to plot histogram for the residuals.

Below is the histogram plotted for the residuals:



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

Based on the tests run, it looks like Hour of the day has a greater impact on ridership than rain. Intuition could be used while choosing the features that could impact the outcomes but statistical models are important to make inferences.