Section 0. References

Impact of Rain on NYC Subway Ridership – Udacity Course Project:
https://saxenarajat99.wordpress.com/2014/09/21/impact-of-rain-on-nyc-subway-ridership-udacity-course-project/

Analyzing the NYC Subway Dataset:
http://sebasibarguen.github.io/udacity-nanodegree-nyc-subway/

Analyzing MTA Subway Data:
http://www.jasondamiani.com/portfolio/analyzing-mta-subway-data/

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 – Wittney U Test. I used two-tail P value.
Null hypothesis is: The subway ridership when it is raining is identical when it's not raining.

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.

Because the distributions of two samples are not normally distributed. Therefore, non-parametric test will be the best fit. Also just drawing out the histogram can clearly tell that the two samples are not normally distributed.

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.

After Mann-Whitney U test, U = `1924409167.0`, P = `0.024999912793489721`.
Since it's a two tail, so the P value should be `0.049999825586979442`
The mean of 'ENTRIESn_hourly' when it's rain is `1105.4463767458733`.

The mean of 'ENTRIESn_hourly' when it's not raining is `1090.278780151855`.

1.4 What is the significance and interpretation of these results?

Given the significance level at 0.05, the p value of `0.049999825586979442` falls in the below the p critical value. Therefore, we can reject null hypothesis, and infer that subway ridership is not the same when it's raining or not raining.

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?

Rain, mean temperature, pressure, fog. Yes, unit is dummy variable

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."***

I decided to use rain because if it's raining people will prefer to use subway since it's underground. Meanwhile, as soon as I included rain, R squared jumped up a lot.

precipi: precipitation decides people's anticipation and expectation of the weather. If precipi is high in one area, people in general may more likely to use public transportation.

meantempi: if the temperature is very low or very high, people are more likely to use transportation, but if temperature is comfortable, maybe people are more willing to bike or walk.

fog: if the weather is foggy, driving may be dangerous, people will be more likely to use public transportation like subway.

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

rain	226.993381
precipi	-5986.819862
meantempi	-12.817060
fog	39.728617

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

0.3789

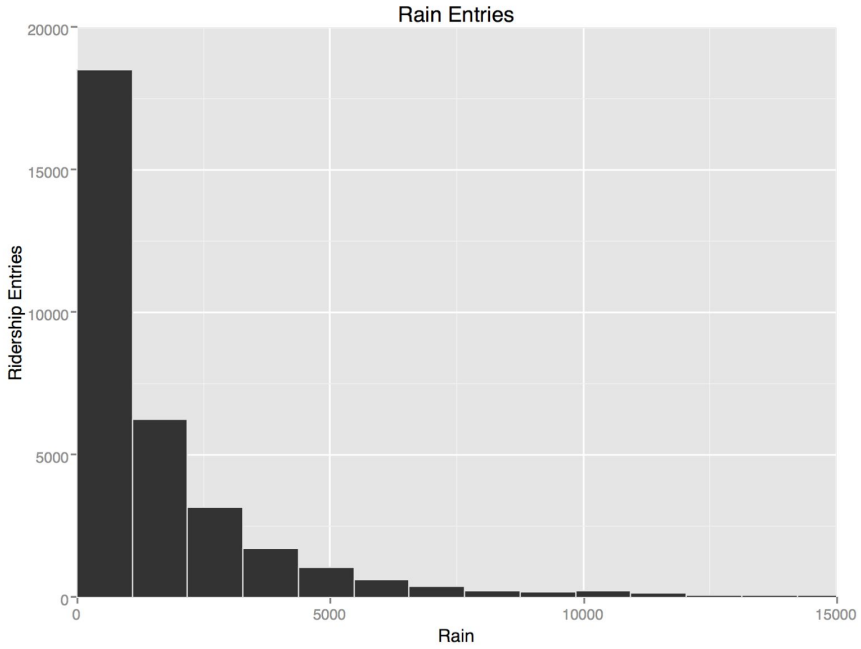
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

R square is 1 minus the ratio of residual variability. Here the residual variability is 1-0.3789 = 0.6211. When the residual variability relative to the overall variability is small, the predictions from the regression equation are good. In other words, the more close R square is to 1, the better the model fits.
In this case, rain, precipi, meantempi, fog and unit explains around 38% of the NYC subway ridership, and are left with 62% residual variability. I think this linear model to predict ridership is appropriate for this dataset, but it could still be improved.

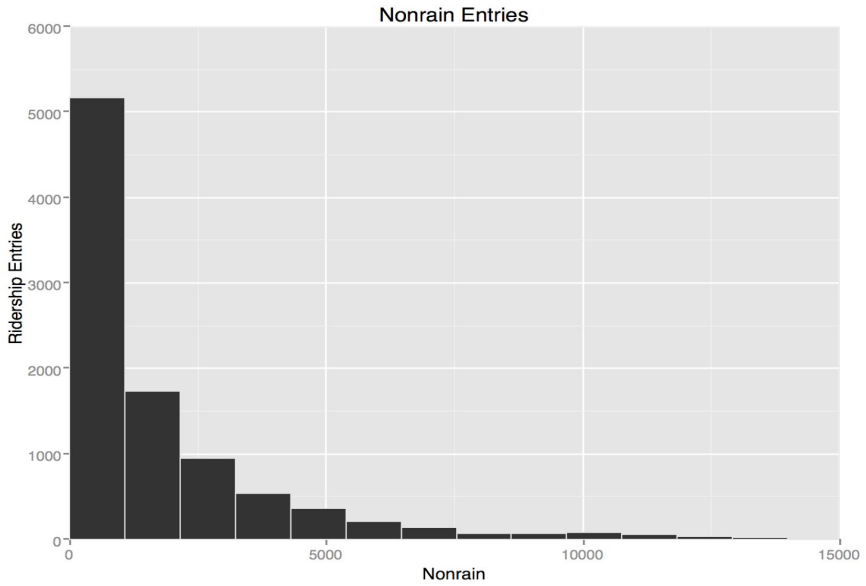
Section 3. Visualization

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.***



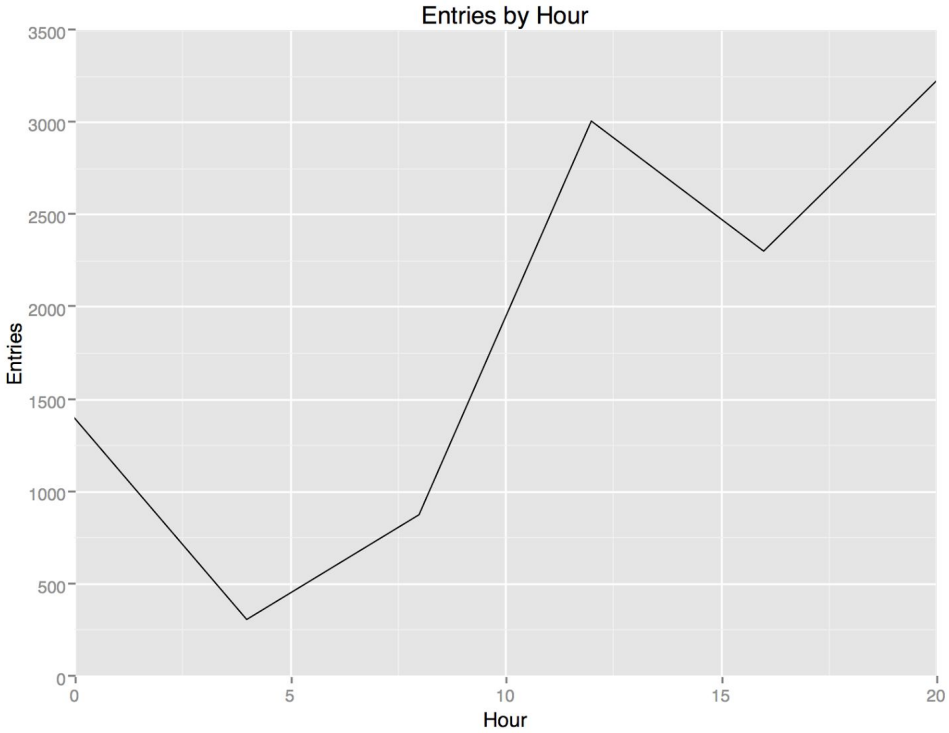
Histogram of ridership when rain: this is a histogram of ridership entries when it's raining. The outliers have been cut out from the graph.



Histogram of ridership entries when not raining: this is a histogram of ridership entries when it's not raining. The outliers have been cut out from the graph.

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 by time-of-day: the x-axis is the hour of the day, y-axis is the ridership within that hour. The x-axis and y-axis are corresponding to 'hour' and 'ENTRIESn_hourly' column from the file.

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?

Yes. More people ride the NYC subway when it is raining.

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 statistical analysis, our null hypothesis is NYC subway ridership is the same when it's rain or not rain. Given the significance level at 0.05, the P value is 0.049999825586979442, smaller than the significance level, thus reject the null hypothesis. Therefore, we can conclude that NYC subway ridership is different when it's raining and not raining.

From linear regression, if we take out dummy variable unit, the R square is only 0.00258996707234, from here, if we take out rain, the R square is 0.00172932911481. Meanwhile, the coefficient of rain is 230.39, whereas the coefficient of precipi is -4708, meantempi -9.465018, and fog -50.525845. Rain has the highest coefficient among these 4 features, and it also means it has the 'biggest weights'.

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 is not large enough. It only has the dataset from a limited period. And the chosen period may not be representative enough to analyze. Also the variables included in the dataset are not that many, and may not be representative enough. It could include more variables such as special events in NYC and some activities from Subway itself.

Regarding Analysis, it is possible that features we chose might be inner related, and we haven't taken that into consideration, so multicollinearity could exist there.

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