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

https://www.moresteam.com/whitepapers/download/dummy-variables.pdf

https://www.khanacademy.org/math/probability/regression/regression-correlation/v/r-squared-or-coefficient-of-determination

https://www.youtube.com/watch?v=dQNpSa-bq4M

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html

http://stattrek.com/hypothesis-test/hypothesis-testing.aspx

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 Mann Whitney U-test to analyze the hourly entries into the NYC subway during rainy versus non-rainy conditions.

I used a two tailed P value because I am only interested to know if the two samples are different and I am not concerned about which sample has the greater number of entries.

The null hypothesis is: The distribution of the rainy sample is the same as the distribution of the non-rainy sample.

> That is, there is a 0.5 probability that an observation (i.e. number of entries) randomly selected from the rainy distribution exceeds an observation randomly selected from the non-rainy distribution.

H_o : $P(x > y) = 0.5$, where x is a random draw from the sample of rainy records and y is a random draw from the sample of not rainy records.

The 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.

The Mann Whitney U-test is applicable to the NYC dataset because it is a non-parametric test that does not assume the data is drawn from any particular underlying probability distribution.

The NYC dataset is non-normal as seen in the histograms in section 3. The non-normal character of the distribution lead me to use the Mann Whitney U-test (i.e. it assumes no

underlying probability distribution). However, the sample sizes are large enough that a Welch's T-test could also be used even though the probability distribution is non-normal.

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 mean of the sample with rain is 1105 entries.

The mean of the sample without rain is 1090 entries.

The one sided p-value returned from the "scipy.statsmodels.mannwhitneyu" is 0.025.

The two sided p-value is 0.05.

1.4 What is the significance and interpretation of these results?

These results indicate that there is a statistically significant difference between the hourly entries when it is raining versus when it is not raining. This interpretation is based on a typical alpha value of 0.05. The two sided p-value result from the Mann Whitney U-test is 0.05 which is equal to the alpha (p-critical) value. This is evidence that the rainy and non-rainy samples are not from the same population.

Hypothesis:

Null hypothesis: The distribution of the rainy sample is the same as the distribution of the non-rainy sample. (I.e., there is a 0.5 probability that an observation randomly selected from the rainy distribution exceeds an observation randomly selected from the non-rainy distribution.)

Alternate hypothesis: The distribution of the rainy sample is not the same as the distribution of the non-rainy sample.

Decision:

Based on the results of the Mann Whitney U-Test (p = 0.05)

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:

a. Gradient descent (as implemented in exercise 3.5)
b. OLS using Statsmodels
c. Or something different?

I used OLS to compute the coefficients theta in my regression model.

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 in my model: 'rain'.

I used dummy variables in my features for the 'UNIT' and the day of the week, and the hour of the day.

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.

I suspect that weather conditions will influence people's decisions about driving/walking versus taking the subway. Also, poor weather may discourage people from going out at all (if they don't have to). I experimented with 'rain', 'fog', 'meantempi', and 'precipi' and found that slightly better R^2 values could be achieved by using a combination weather related features. However, the difference in R^2 values was very small and also I suspect that there is multi-collinearity between these weather related variables. Therefore, I chose 'rain' only.

I chose 'Hour' of the day because it makes sense that people's daily schedule (e.g. travel to and from work) will have a significant effect of subway usage at specific times of the day. However since I don't think that there is a linear relationship between the numerical value of the hour and the ridership, I treated the 'Hour' as categorical data rather than numeric. For this reason I used dummy variables for the hour.

I chose 'UNIT' to capture the geographic variability and population density variations along the subway routes.

I chose the day of the week because there are significant differences in volume between week days and weekends. More generally, people's weekly schedules may affect the number of people who need to travel on each day of the week.

- 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."

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

The coefficients of the non-dummy features are

'rain' 2.49

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

The coefficient of determination (R^2) for my model is 0.537.

```
Your R^2 value is: 0.536800876737
```

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 coefficient of determination (R^2) is the percentage of total variation in ridership that is described by the model. 53.68% of the variations in ridership is described by the model. This seems like it could be useful to predict weather related variations in ridership but a good part of the variability in ridership is due to other factors. I think that a more detailed model would be needed if the model is to be useful for making decisions.

Section 3. Visualization

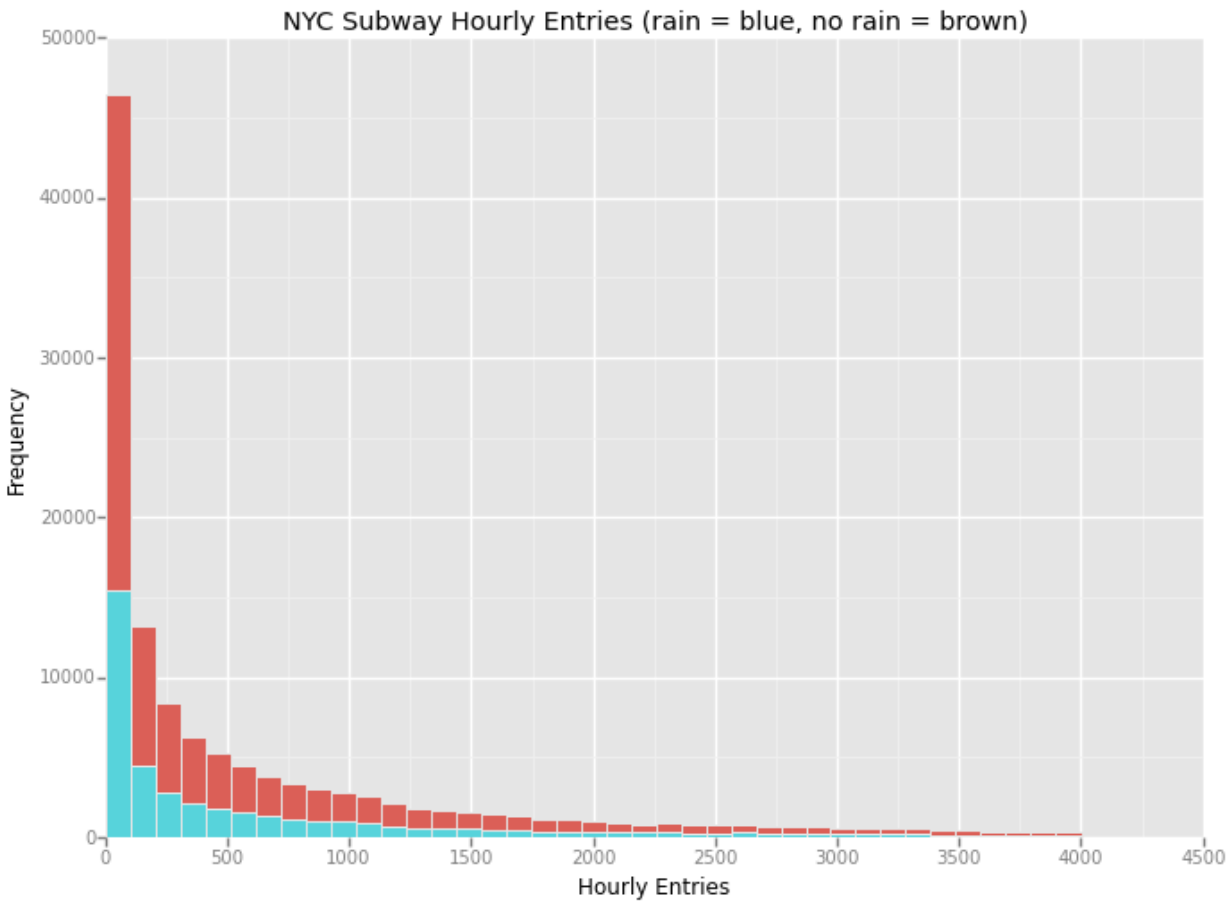
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

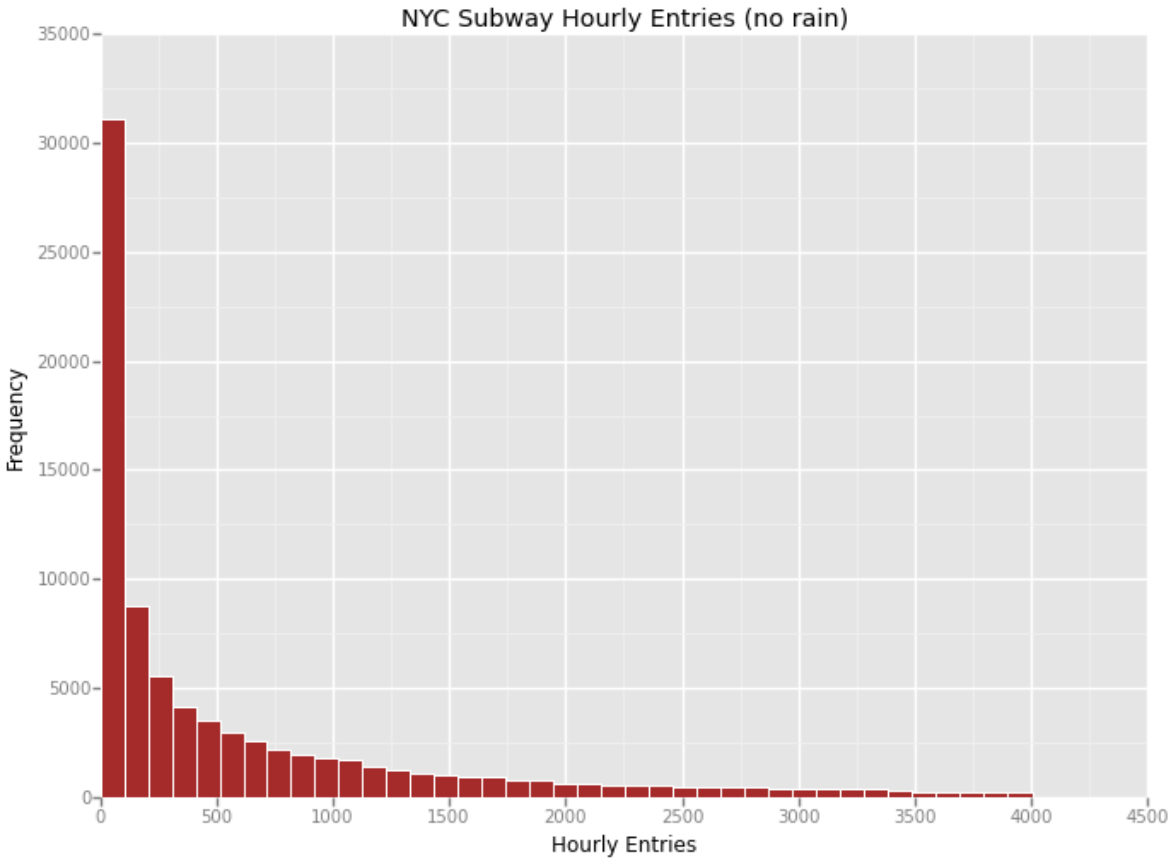
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

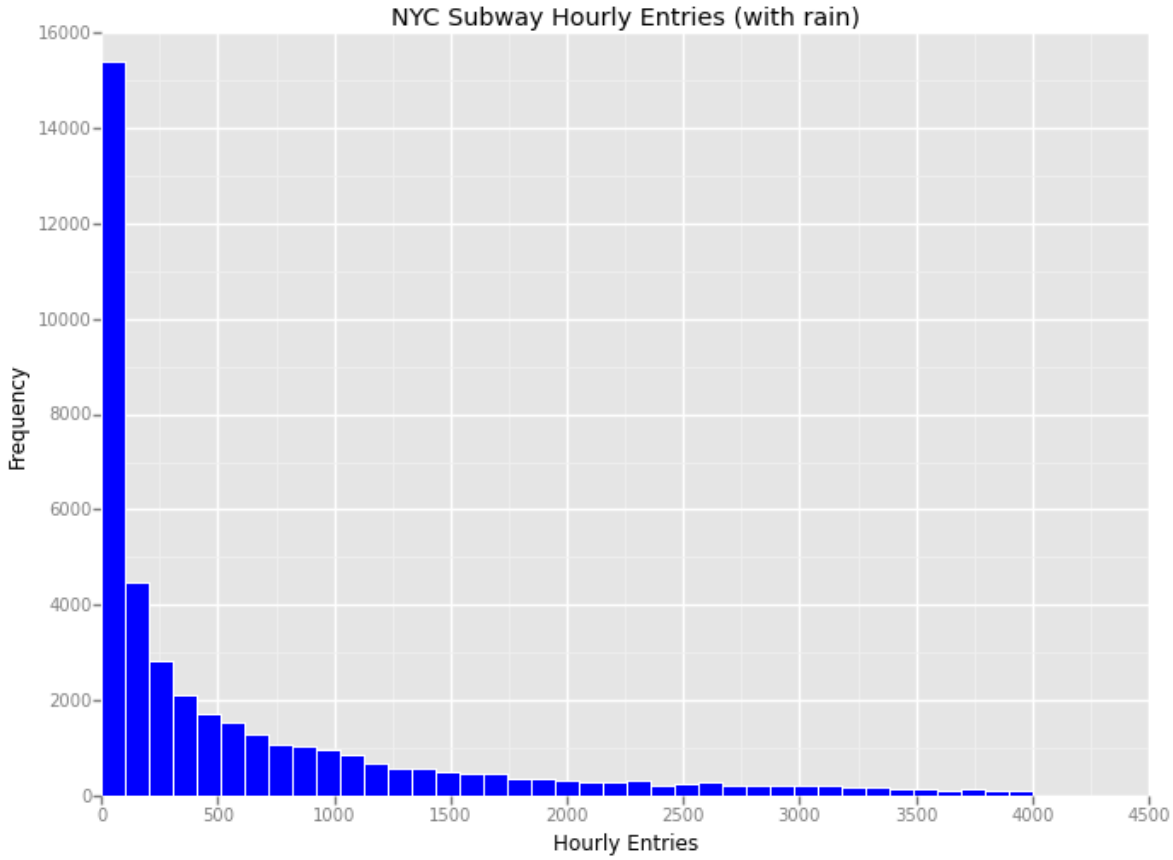
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

- 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.

The following visualizations show histograms of the hourly entries into the New York City subway when it is raining versus when it is not raining. The first visualization shows both 'rain' and 'no rain' data in one histogram. The second and third visualizations show the 'rain' and 'no rain' data in separately.


NYC Subway Hourly Entries (rain = blue, no rain = brown)
Frequency
50000
40000
30000
20000
10000
0
0
500
1000
1500
2000
2500
3000
3500
4000
4500
Hourly Entries


NYC Subway Hourly Entries (no rain)
Frequency
Hourly Entries
0
5000
10000
15000
20000
25000
30000
35000
0
500
1000
1500
2000
2500
3000
3500
4000
4500


NYC Subway Hourly Entries (with rain)
Frequency
Hourly Entries
0
2000
4000
6000
8000
10000
12000
14000
16000
0
500
1000
1500
2000
2500
3000
3500
4000
4500

The key insights depicted in these histogram are:

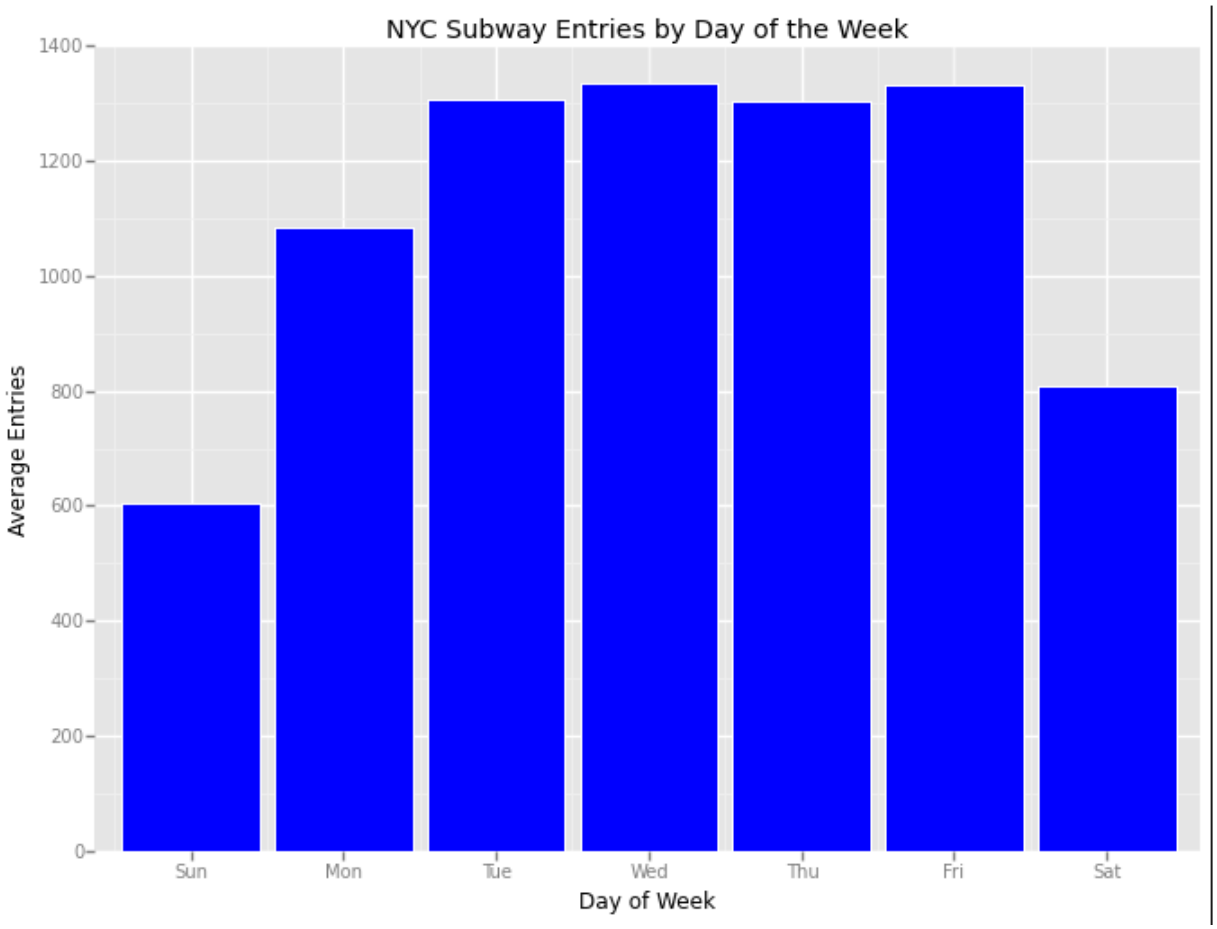
1. More people ride the subway when it is not raining than when it is raining. This is evidenced by the brown bar being consistently larger than the blue bar for all intervals.
2. The hourly entries data are right-skewed. Smaller hourly entries values are seen with higher frequency.

Note that there are more records in the dataset for not raining than for raining conditions.

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 following visualizations show the average hourly entries into the New York City subway with respect to time. The first visualization shows the average hourly entries by day of the week.

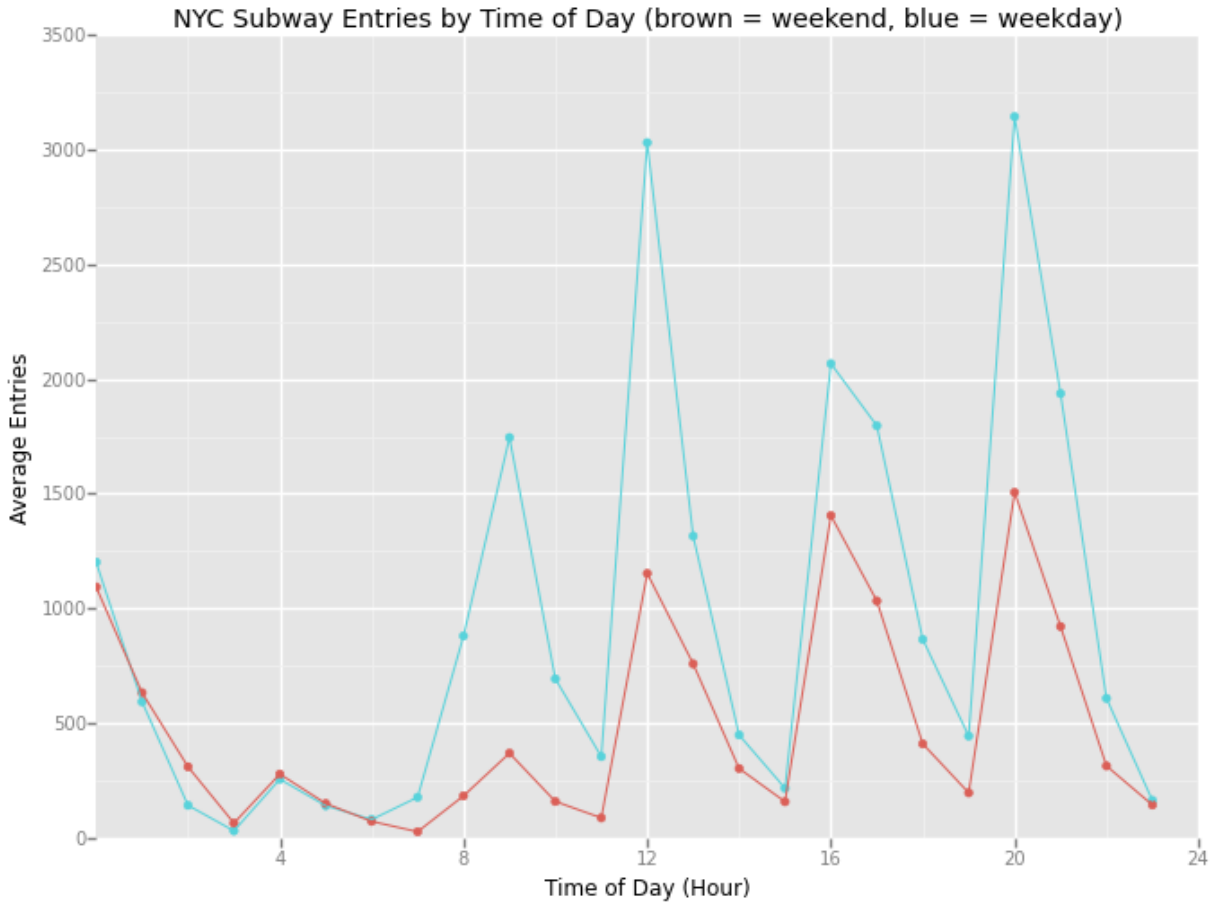


Key insights depicted in this bar chart are:

1. More people ride the subway during the traditional workweek (Monday to Friday) than on the weekend.

2. Average hourly usage of the subway is relatively constant during the work week. However, there appears to be a slightly lower usage on Mondays. Further analysis is needed to identify likely reasons for this. Perhaps there was a statutory holiday on a Monday during the month that affected the average.
3. Sundays see the least number of subway riders.

The second visualization shows the average hourly entries by time of day. Additionally, the data is divided into two groups (weekend ridership and weekday ridership).



Key insights depicted in this line graph are:

1. There are peak times during the day when subway usage is high (hour 9, 12, 16, 20). In between these times the usage is relatively low.
2. The same peak times occur on weekends and on weekdays. However, the magnitudes of the peaks are significantly smaller on the weekends.
3. There appears to be a large number of riders entering the subway at midnight. There is no significant difference between weekend and weekday riders at midnight.

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 the subway when it is raining than when it is not raining as indicated by the positive coefficient (2.49) of the 'rain' feature in the regression model.

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 indicated that there is a statistically significant difference between ridership when it is raining versus when it is not raining ($p = 0.05$). I reject the null hypothesis that the two distributions are the same with a 95% confidence level.

The linear regression resulted in a small positive coefficient for the 'rain' parameter. This indicates that rain positively effects the models prediction of the number of people entering the subway. That is, if it is raining (rain = 1) then the positive coefficient results in predicting more entries into the 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 data set records whether it is raining or not for each time interval. However, the number of intervals when it is raining is not the same as the number of intervals when it is not raining. So comparing the absolute number of entries is not indicative of the effect of the rain.

The linear regression model could be improved by adding more features. However, to avoid the problem of multi-collinearity, the weather related variables in the dataset (e.g. 'fog', 'meantempi', and 'percipi') should be evaluated to ensure there is little or no correlation between them before attempting to add them as features in the regression model.

Some of the weather related variables may actually be closely related and therefore redundant. For example, the 'meantempi' and 'rain' variables may have a very close correlation. Redundant (or collinear) independent variables can confuse the linear regression since it is not clear which redundant variable is responsible for the changes in the dependent variable. The redundant variables will likely skew the coefficients determined by the linear regression.

The use of dummy variables for UNIT, Hour, and day of week is important to the linear regression because these features have a visible effect on the number of entries to the subway as seem by the visualizations.

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

Since the Mann Whitney U-test results were significant at the 95% level (on the line between significant and not significant) and since the linear regression model has a very small coefficient for the 'rain' feature, I believe that rain does not play a strong role in NYC Subway ridership.