# Section 0. References

http://stackoverflow.com/questions/14386117/how-to-look-back-at-previous-rows-from-within-pandas-dataframe-function-call

Initially I had some trouble figuring out how to “look back” on a previous row, this topic helped clear that up.

http://statsmodels.sourceforge.net/devel/examples/notebooks/generated/ols.html

Didn’t really understand how to use OLS library, this helped.

http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm

Helped me understand how to interpret the p value for the Mann-Whitney U Test, and p values in general given a null hypothesis.

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

Used this to understand dummy variables.

http://blog.yhathq.com/posts/aggregating-and-plotting-time-series-in-python.html

How to aggregate and plot data

# Section 1. Statistical Test

1.1.1 Which statistical test did you use to analyze the NYC subway data?
**I used the Mann-Whitney U test to analyze the NYC subway data**

1.1.2 Did you use a one-tail or a two-tail P value?
**I used a two-tail P value since I didn’t specify which sample would have a larger mean**

1.1.3 What is the null hypothesis?
**The null hypothesis is that the two samples come from the same population**

1.1.4 What is your p-critical value?
**Since we assume a 95% confidence interval, so the alpha is 0.05, p-critical is 0.025 for each tail.**

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 two samples are not normally distributed, which can be confirmed by the Shapiro-Wilk test, so we are unable to use Welch’s T test, and therefore must use the Mann-Whitney U test to confirm that the two samples came from different populations.**

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.
**U-value == 1924409167.0, p-value ~= 0.025, alpha == 0.05, sample with rain mean ~= 1105.4464, sample without rain mean ~= 1090.2788**

1.4 What is the significance and interpretation of these results?

**There is about a 2.5% chance of randomly obtaining means at least as far apart as observed from the same population. Since we assume a 95% confidence interval and a two-tailed test, we will reject the null hypothesis and assume that the two samples are drawn from two different populations, or that there is actually a difference in number of entries when it rains versus when it does not.**

# 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:
**I used the SGDRegressor function from the scikit learn package using alpha == 0.000001 and 100 iterations.**

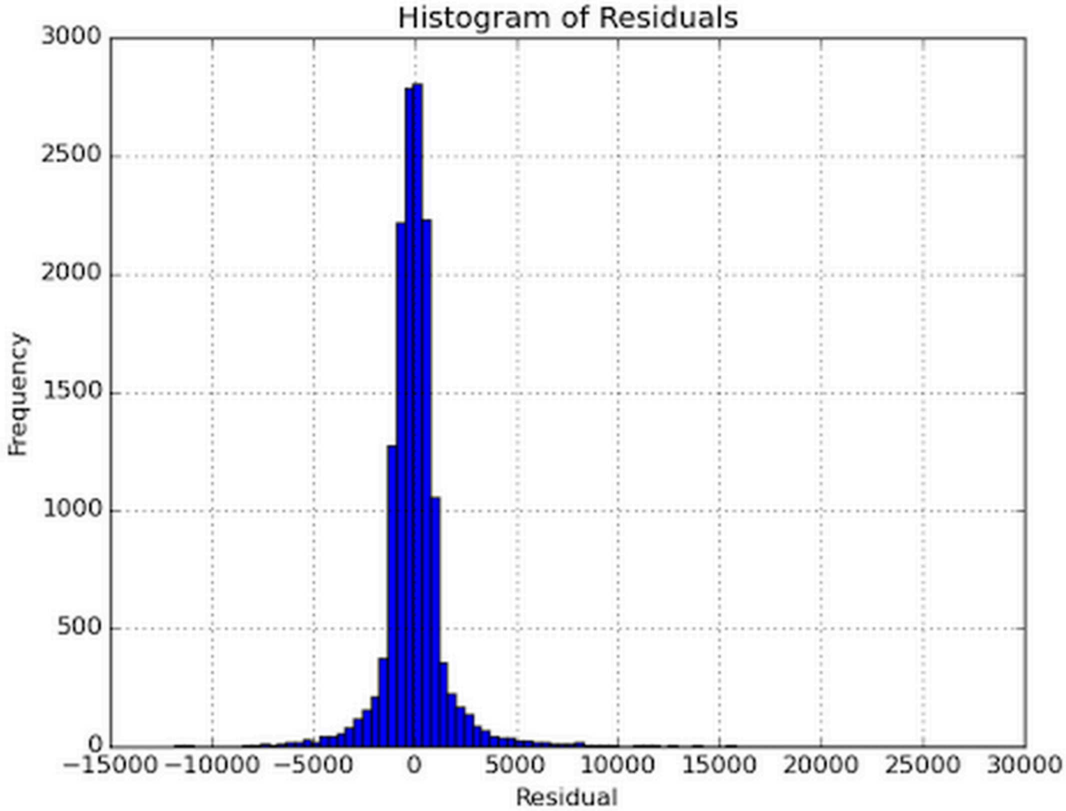
2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?
**I used the features rain, precipitation, hour, mean temperature, mean wind speed, and fog. A dummy variable was used for the UNIT feature, which was important to keep track of since there is great variation between subway stops**

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 maintained the original feature set provided because any change resulted in a lower R^2 value. I then added 'fog' to the list of initially given features intuitively, thinking that any weather pattern could contribute to the number of subway patrons. The assumption that 'fog' also contributes to how many people ride the subway was confirmed when I noticed an increase in R^2 between the feature set without and with 'fog.' This increase in R^2 shows that the feature set used in the linear regression better fits the data provided to construct it; the model describes more of the variance in the data.**

2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?
**[ -7.00970292e+01, -1.20171235e+01, 6.49919841e+01, -5.52324638e+00, 2.84858914e+01, 1.01820613e+02 ]**

2.5 What is your model's $R^2$ (coefficients of determination) value?
**R^2 == 0.431718954406**

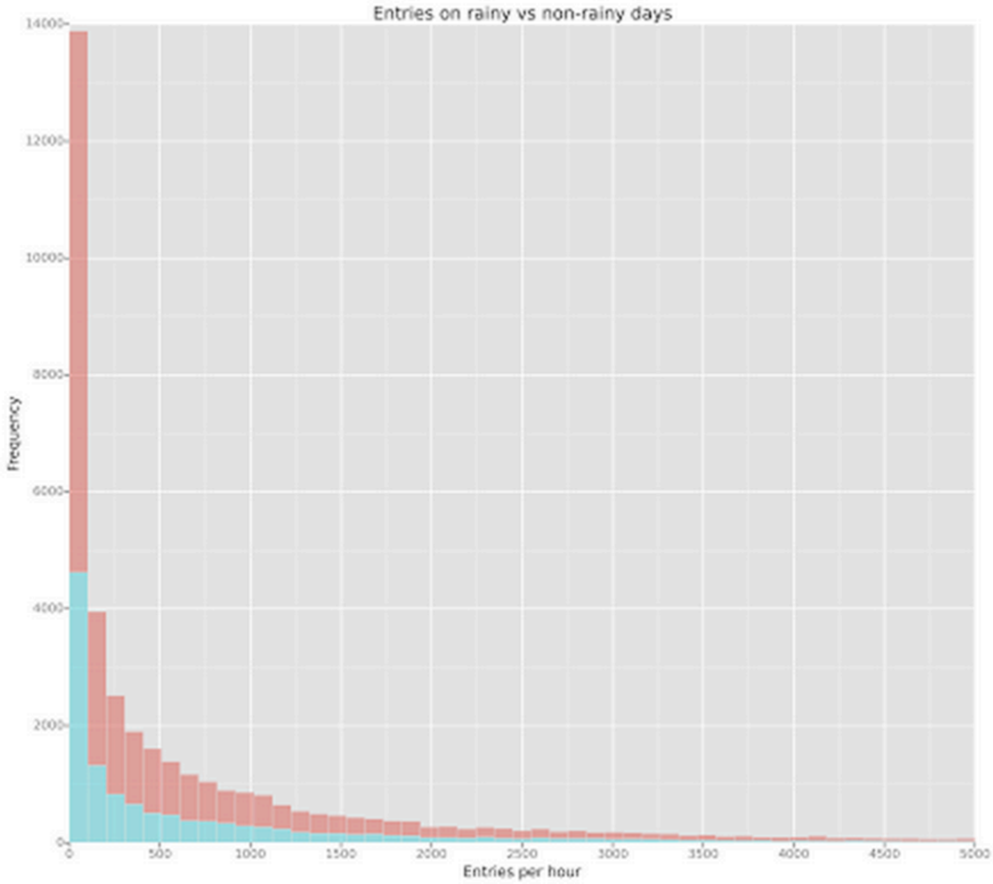
2.6 What does this $R^2$ value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this $R^2$ value?
**The R^2 value helps determine how much of the variance is described by the model by comparing the residuals to the actual variance. To get a better picture of how well our model fits, we can plot a histogram of the residuals.**



**The above figure is a histogram of the residuals with a bin size of 100. This figure shows that although the distribution is normal, it has a wide range, approximately (-5000,5000). One way to ameliorate this skewness is to experiment with non-linear multiple regression. We can determine a feature's linearity by plotting residuals of only that one feature and examining it's behavior. From these data I think it's safe to say that we can give an educated guess about ridership, but not predictions that would be dependable in sensitive systems.**

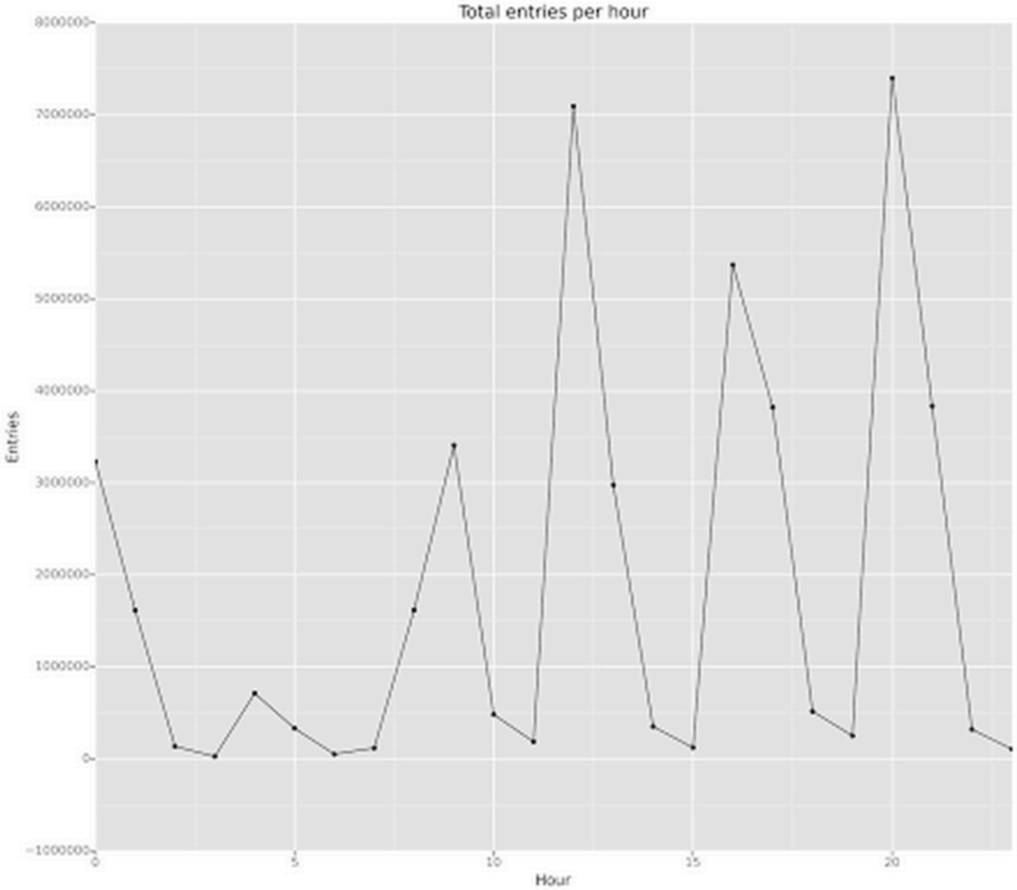
# Section 3. Visualization

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



**The figure above shows a stacked histogram of entries per hour (ENTRIESn_hourly) with green depicting rainy days and red depicting non-rainy days; apologies for the missing legend, I'm not sure if there is support for legends in the python implementation of ggplot. I went through many tutorials for R, but none of the solutions worked in python.  From this figure, we can assume that there are many more entries on non-rainy days than on rainy days. We can also see that the most common bucket is 100 entries per hour, falling very quickly and running into the long tail around 2000 entries per hour. The data has been truncated to 5000 entries per hour. It is important to note that although there is more ridership on non-rainy days according to this visualization, there are fewer rainy days included in this data, so we cannot conclude from this graph that more people ride the subway on 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:



**The above figure shows a line graph of the total number of entries per hour, grouped and summed over all days in the dataset. This shows the raw number of entries in a given hour. Peaks around the morning and afternoon rush hours (9am and 4-5pm) as well as lunch and dinner (12pm and 8pm) are clearly depicted.**

# Section 4. Conclusion

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?

**Using the results from the Mann-Whitney U Test described above, we can reformulate our conclusions from the non-directional alternative hypothesis that the samples are drawn from different populations, to the directional alternative hypothesis that one population**

**has a higher mean than the other. More specifically, we can reject the null hypothesis that the population from which the sample of non-rainy days is drawn from has a mean equal to or higher than the population from which the sample of rainy days is drawn from. With an alpha of 0.05, and the returned one-tail value from above (p == 0.025), we can reject the null hypothesis.**

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.
**As stated above, the Mann-Whitney U test results help us say with more confidence that more people ride the subway when it's rainy. Furthermore, considering the relatively high R^2 value of 0.43 from our linear regression model, we notice that rain, among other factors, does contribute to subway ridership.**

# Section 5. Reflection

**Normalizing the data to control for variation in traffic due to the different subway stations would help account for stations that are just naturally more congested so we can boil out the more subtle effects rain may have on ridership. The dataset I drew my conclusions from were also taken from the online set for the S3 server, which is admittedly only a third of the data. More data over a longer period of time, along with partitioning the data for training and verifying stages would help provide a more reliable model as well. I also used the default regression model, but I am sure there are more complex learning algorithms that could analyze the dataset better. The model could have also accounted for intensity of precipitation, generating groups for a light rain and a heavy rain; a hurricane would probably lead to no ridership whatsoever.**

**The data also only captures ridership that could be recorded, obviously, but this might also influence our model, discounting any riders that hopped the entry way. There might have also been data loss from purely technological reasons, such as an unreliable entry way or outages.**