Analyzing the NYC Subway Dataset

Project 2 Report - by Casey Faist

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.

References:

http://stanford.edu/~mwaskom/software/seaborn/
http://stanford.edu/~mwaskom/software/seaborn/tutorial.html
http://jupyter.org/qtconsole/stable/index.html
http://statsmodels.sourceforge.net/devel/generated/statsmodels.regression.linear_model.RegressionResults.html
http://chrisalbon.com/python/seaborn_color_palettes.html
http://stanford.edu/~mwaskom/software/seaborn/tutorial/distributions.html
http://stanford.edu/~mwaskom/software/seaborn-dev/generated/seaborn.distplot.html
http://matplotlib.org/users/text_intro.html
http://stanford.edu/~mwaskom/software/seaborn/generated/seaborn.barplot.html
http://stanford.edu/~mwaskom/software/seaborn/tutorial/categorical.html
http://stanford.edu/~mwaskom/software/seaborn/tutorial/color_palettes.html#qualitative-color-palettes
http://stanford.edu/~mwaskom/software/seaborn/tutorial/distributions.html#plotting-bivariate-distributions

Also find references pasted as comments next to relevant lines in code file.

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?

After plotting an exploratory histogram comparing the number of entries when rain was observed and when no rain was observed (fig 1), I decided that a Welch's t-test or other t-test assuming a normal distribution would be inappropriate for this dataset. Instead, I opted for the Mann-Whitney statistical analysis.

My hypothesis is that ridership increases when it rains. Therefor, I will be using a one-tailed P value at a P-critical value of 5%, a standard alpha level for this test, to interpret the Mann-Whitney results.

The null hypothesis, that there is no significant difference in ridership between rainy conditions and non-rainy conditions:

$H_0 : x_{bar1} - x_{bar2} = 0$

My alternative to the null is that subway ridership is greater when it is raining than when it is not:

$H_A : x_{bar1} > x_{bar2}$

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.

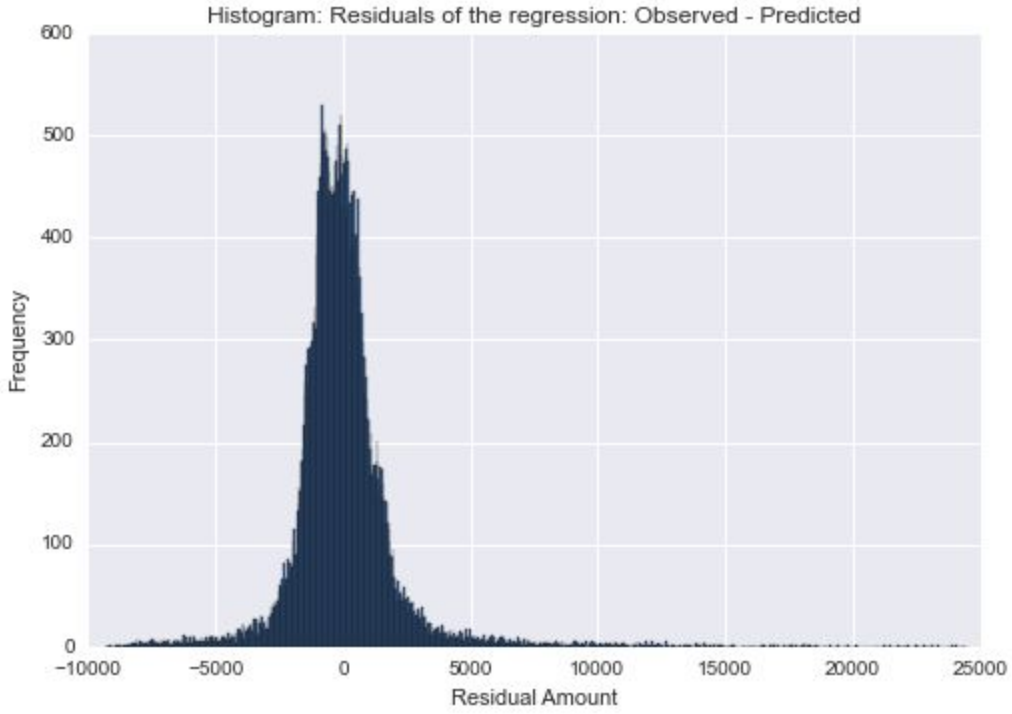
The histogram (fig 1) revealed that the data is skewed, so testing whether both samples came from the same normally distributed population would lead to inaccurate representation of trends in the data. The Mann-Whitney test is able to test if two samples came from the same population for nonparametric data, which makes it more appropriate for this dataset.

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.

Copied from iPython console:

```
In [144]: %run proj2_main.py
```

P-result = 2.74106957124e-06 At a one-tailed alpha level of 5%, the statistical analysis is significant.



The mean hourly entries with rain was: 2028.19603547 The mean hourly entries without rain was: 1845.53943866 R^2 value = 0.484944517921 Coefficients: weekday 9.471582e+02
hour 1.236949e+02
latitude -4.368886e+13
longitude -2.424438e+13
precipi 5.314337e+02
fog -3.020458e+02
rain 7.788432e+01
meanpressurei -4.323586e+02
unit_R003 -3.688474e+10
unit_R004 6.843996e+10
unit_R005 1.747252e+11
unit_R006 3.132797e+11
unit_R007 4.301532e+11
unit_R008 -2.560743e+11
unit_R009 7.399700e+11
unit_R011 2.196827e+10
unit_R012 -5.925817e+10
unit_R013 -5.925817e+10
unit_R016 1.838806e+11
unit_R017 1.838806e+11

```
unit_R018                 9.847204e+11
unit_R019                 1.542645e+12
unit_R020                 1.537504e+11
unit_R021                 1.991215e+10
unit_R022                -8.275433e+10
unit_R023                -8.275434e+10
unit_R024                 9.835044e+11
unit_R025                 1.106235e+12
unit_R027                -8.601385e+11
unit_R029                -1.193990e+12
                             ...
Station_PRESIDENT ST      -1.455151e+12
Station_PRINCE ST-B'WAY   -7.919306e+11
Station_PROSPECT AVE      -2.695585e+12
Station_PROSPECT PARK     -1.733072e+12
Station_QUEENSBORO PLZ     4.765353e+11
Station_RECTOR ST         -1.803918e+12
Station_ROCKAWAY AVE      -1.064308e+12
Station_ROOSEVELT AVE      9.830135e+11
Station_ROOSEVELT IS       5.058337e+11
Station_SARATOGA AVE      -1.179917e+12
Station_SMITH-9 ST        -1.879768e+12
Station_SPRING ST         -1.102115e+12
Station_ST. GEORGE        -3.474177e+12
Station_STEINWAY ST        8.500235e+11
Station_STERLING ST       -1.566979e+12
Station_SUTPHIN BLVD       9.899057e+11
Station_SUTTER AVE        -1.188774e+12
Station_UNION ST          -1.643708e+12
Station_VAN SICLEN AVE    -7.678660e+11
Station_VAN WYCK BLVD      1.023086e+12
Station_VERNON/JACKSON     1.421783e+11
Station_W 8 ST-AQUARIUM   -3.765880e+12
Station_WALL ST           -1.763993e+12
Station_WASHINGTON-36 A    7.487664e+11
Station_WESTCHESTER SQ     3.607441e+12
Station_WHITLOCK AVE       2.786524e+12
Station_WILSON AVE        -4.374478e+11
Station_WOODHAVEN BLVD     3.129193e+11
Station_WOODLAWN ROAD      4.177401e+12
Station_WORLD TRADE CTR   -1.199629e+12
dtype: float64
```

In [145]:

1.4 What is the significance and interpretation of these results?

At the standard one-tailed alpha level of 5%, these results are significant and we would reject the null hypothesis in favor of the alternative. Thus, according to this report, ridership increases when it rains.

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 the OLS statsmodels method, taken directly from my answers to Problem Set 3 in the Intro to Data Science course.

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

I found that the following variables translated to the most accuracy for all of the combinations I tested:
"weekday", "hour", "latitude", "longitude", "precipi", "fog", 'rain', 'meanpressurei'

The ‘UNIT’ data category, “Conds” category, and “Station” category were included as dummy variables for this regression.

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 R2 value."*

I selected these features through a combination of common sense and trial and error. It makes intuitive sense that on days with higher precipitation, more people might decide to use the subway; however, my guess that day of the week would affect ridership was not supported by the reported coefficient.

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

The parameters for the features in my OLS linear regression model were as follows:

(copied from above)

weekday	9.471582e+02
hour	1.236949e+02
latitude	-4.368886e+13
longitude	-2.424438e+13
precipi	5.314337e+02
fog	-3.020458e+02
rain	7.788432e+01
meanpressurei	-4.323586e+02

Given the dummy values I selected, weekday and rain had the greatest effect on the accuracy of predictions.

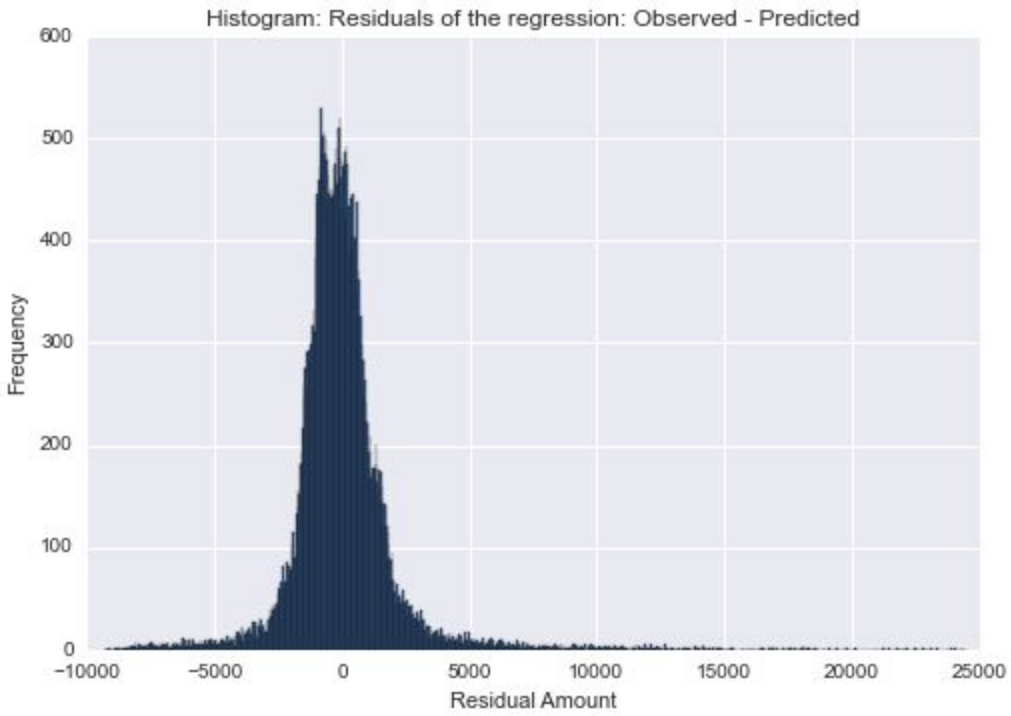
2.5 What is your model's R2 (coefficients of determination) value?

My regression and feature selection produced an R2 value of 0.48494451792.

2.6 What does this R2 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 R2 value?

The closer the R2 value comes to 1, the better your regression model fits the dataset. However, in real-world analysis, achieving a perfect 1 becomes more and more difficult. A truly reliable predictor would lie somewhere over 0.50; however a fit of ~0.48 is not insignificant. Therefore I think this linear model is appropriate for this dataset.

A plot of the residuals supports this conclusion:



My residuals are clustered around 0, with a slight tendency to over-predict shown by the higher frequencies of residuals on the negative side of 0.

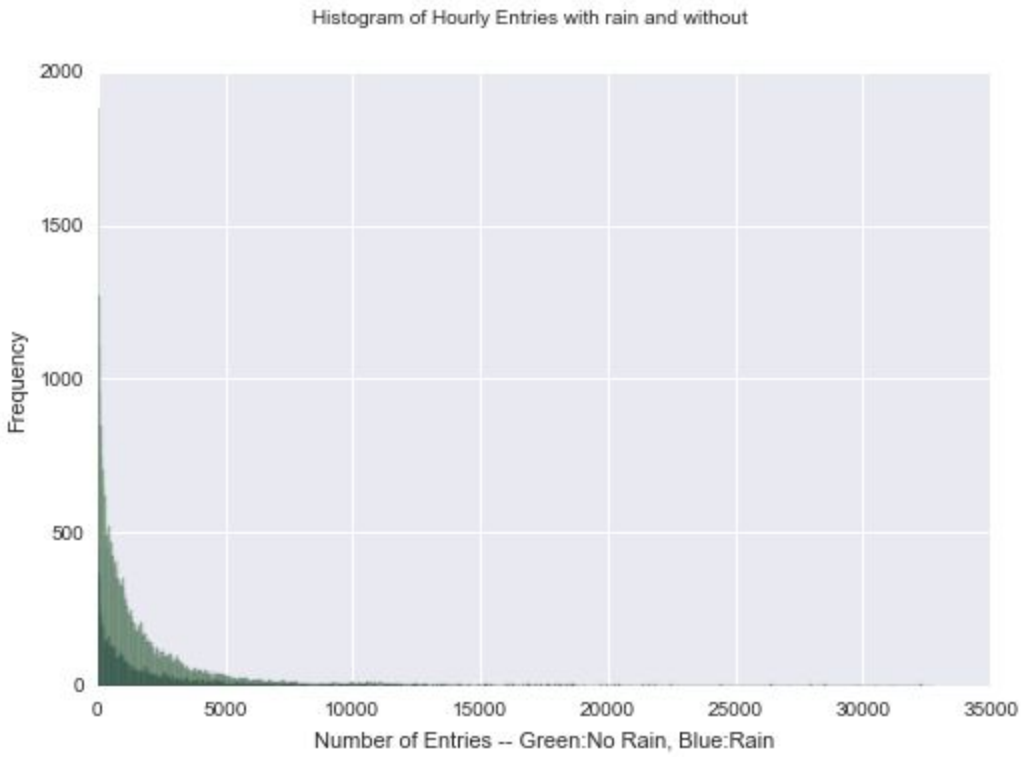
Section 3. Visualization

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.

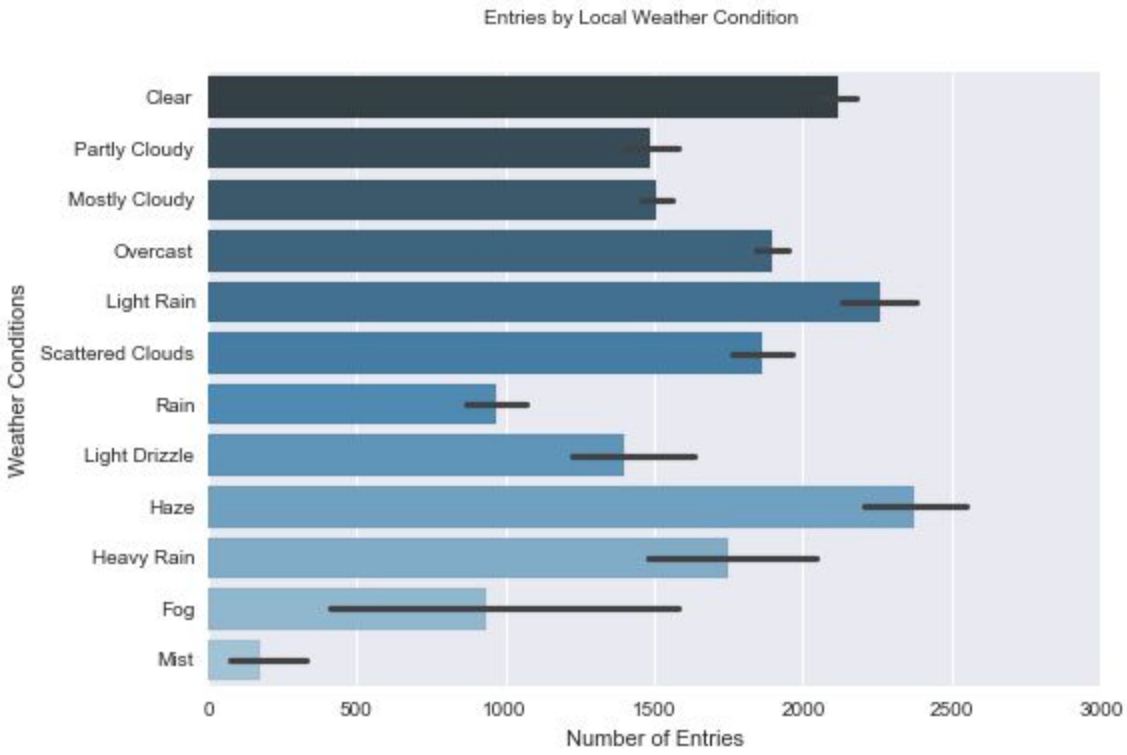
- *You can combine the two histograms in a single plot or you can use two separate plots.*
- *If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.*
- *For the histograms, 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.*



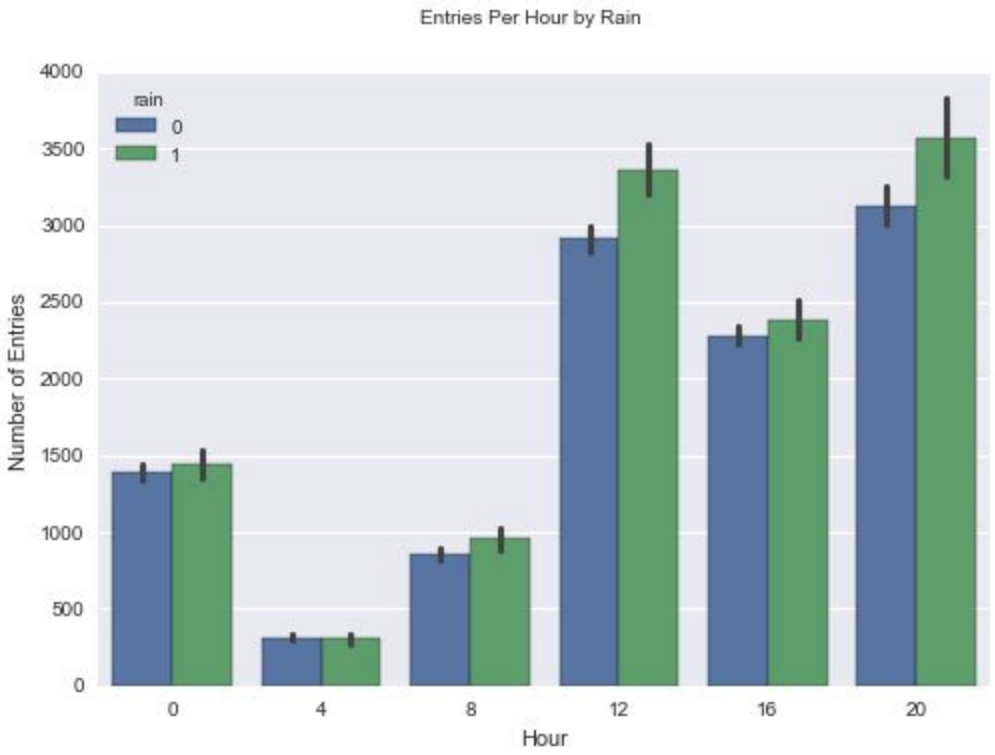
This histogram visually demonstrates fewer data points for rain overall, but a less steep trend - more of the with-rain data points seem to report higher entry counts than without-rain. I based my hypothesis, that ridership increases when it rains out, on that observation.

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*



This bar graph illustrates the number of hourly entries by the condition at the time and location of the units. The most people entered the subway during clear, light rain, and hazy conditions, with the potential error on heavy rain reaching above 2000 as well. Rain itself in this graph seems much less impactful, and mist saw far fewer entries than any other - potentially because mist most often would occur very early morning or late night, when the fewest number of people would be entering the stations regardless.



Across all days, the average hourly entries did increase when it was raining vs. when it was not raining. This graph accounts for spikes in ridership throughout the average day; the biggest jump in ridership when it rains occurs around noon and in the last few hours of the day.

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?

From my statistical analysis and visual interpretation of the data, it seems to be a clear trend that more people ride the NYC subway when it is raining. From the hourly plot, it seems that more people will choose to take the subway in the middle of the day and late at night.

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 one-tailed significant p result of ~0.00000274 from the Mann-Whitney U test, and the roughly normally distributed goodness of fit score and residuals plot conducted after the linear regression, led to my conclusion that ridership does indeed increase when it rains. The OLS regression itself suggests that rain and whether or not it is a weekday are the two biggest predictors of ridership on the NYC 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,*

This dataset leaves out some important pieces of information about ridership. For one, it would be interesting to compare the number of total riders to the number of total travellers on rainy days and non-rainy days. The implication of increased ridership on the subway when it rains is that they would otherwise be travelling some other way - either walking, driving their own cars, or by bus, for example. Then, predictions get a little easier: Would they otherwise be walking, but found the conditions to be unpleasant? If they would instead drive, would they otherwise be facing rush hour or bridges or construction work in more dangerous (wet and less visible) conditions? Then you could predict ridership from windchill or driving conditions directly.

Another interesting aspect to consider is why the riders are traveling. If rainy conditions on the roads would make corporate workers late, it makes sense that ridership would increase - but only at certain times of day or certain days. If the shift is due instead to, for instance, tourists who get stuck out in the rain, then any time of day there was rain there would also be a predictable spike in tourists.

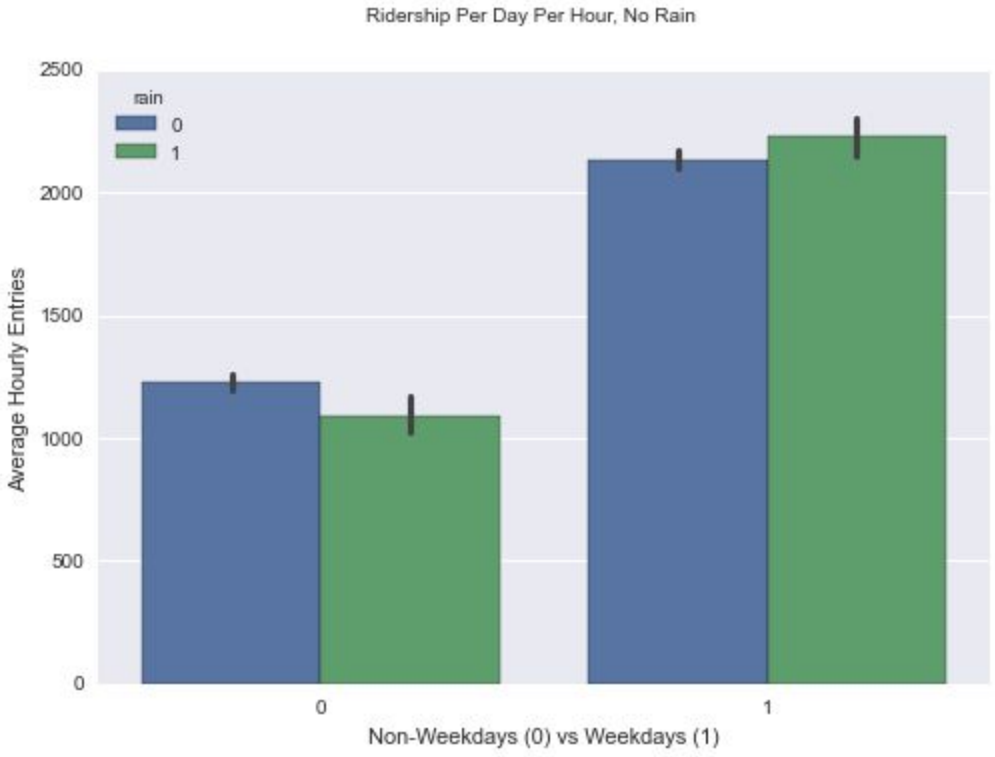
2. *Analysis, such as the linear regression model or statistical test.*

Obviously any analysis can be improved. Perhaps a different statistical test would have been more appropriate. With more testing, I might have found features that better predict ridership and improve the R2 value. I also used the OLS model, which typically gives more accurate results but takes longer to compute.

While the effect of rain on overall ridership is a good first step, an investigation into whether rain increases ridership at specific stations more than others would be an interesting one. It would allow for better crowd-control procedures under the observed conditions, influence the maintenance or train schedules, and could be used to inform the public as to what stations would be most crowded.

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

I think it is possible that the ridership profile for those who would take the subway in rain and would not in clear weather involve people on lunch breaks from work or tourists heading to restaurants, and those who would otherwise call a cab or walk after a night out to bars or cultural attractions.



When graphed, interestingly, ridership overall decreases during rain on weekends, but increases during rain on weekdays.