Udacity Nanodegree Project 1

NYC Subway Ridership and Weather Analysis

By: Taylor Somma

This is a project analyzing data obtained from the NYC subway system and provided by Udacity. The objective of this analysis is to determine what facets of the weather, time of day, and day of week most influence the number of people that ride the subway in NYC. After determining the variables with the strongest correlation we will produce a predictive model using linear regression with gradient descent to best predict the number of subway riders.

Sources

Udacity
Pandas Documentation
Statistical Test Flowchart
Parametric vs. non-parametric data
Mann-Whitney U
Interpreting R-Squared Value

Back to magicfilebox.com

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 a two tail p-value because we are testing if there is a difference in means between the datasets without a presumption of which mean is larger.

Assuming a null hypothesis that there is no meaningful difference between the mean number of riders on rainy and non rainy days. I chose to use a p-critical value of 0.01.

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.

A Mann-Whitney test was chosen because the sample sizes of rainy and non-rainy days are different and the frequency distribution shown in the histogram in the answer for 3.1 the data for ENTRIESn_hourly is 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.

As seen here the p value of the Mann-Whitney U test is 2.74e-06, which is much less than 0.05. Since this output gives us a one-tailed result we simply need to double it to get the two-tailed p-value of 5.482e-06. The mean number of hourly riders for rainy and non rainy days is 2028 and 1845 respectively.

1.4 What is the significance and interpretation of these results?

Since the p-value is far less than the p-critical value we can safely reject the null hypothesis of any difference in means of ENTRIESn_hourly during rainy and non rainy days being due to chance. Given the p-value and the mean number of ENTRIESn_hourly during rainy days being more than not rainy days we can make the claim that more people ride the subway during rainy days than days with no rain. We cannot make any claims of why this is yet but one guess is that people who would have otherwise walked to where they needed to go instead chose to take the subway.

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: OLS using Statsmodels or Scikit Learn Gradient descent using Scikit Learn Or something different?

I used gradient descent using Scikit Learn

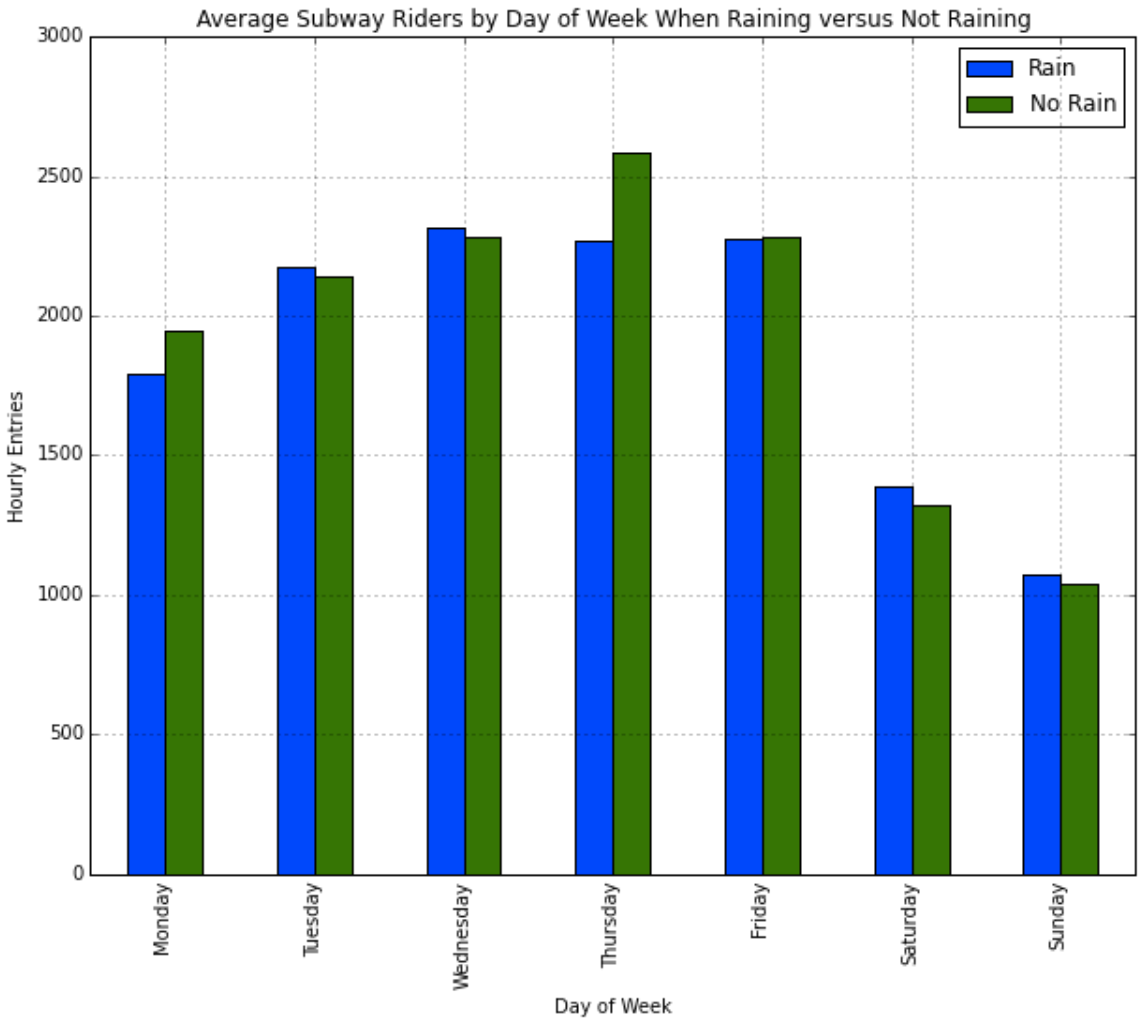
2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

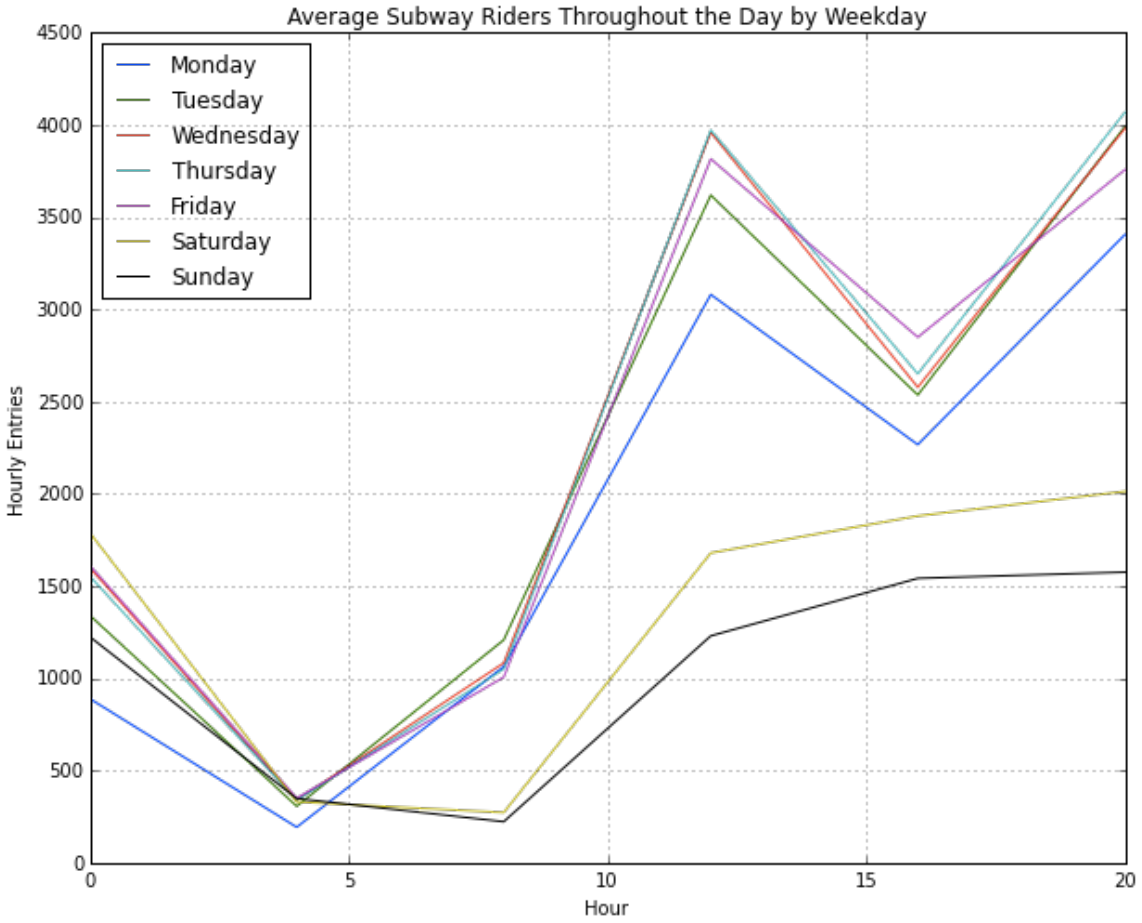
I used rain, weekday, and hour as input variables in my model. My model used unit as a dummy variable.

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 R2 value.”

I started by testing several input variables independently. The selection of these variables was based partly on intuition and partly on other visualizations I made. These variables were rain, weekday, fog, hour, and latitude and longitude. Looking at the heat map shown in question 3.2 using mean ENTRIESn_hourly as weights it appears that location is a good indicator of ridership. Latitude and Longitude turned out to not be the best predictors of ridership with r-squared values of 0.375 when using OLS linear regression. As seen in the graph below showing mean ENTRIESn_hourly split by day of week and rain and no rain groups it is apparent that there are less subway riders on the weekend and some minor

variations throughout the week. I also chose to use the time of day as a feature in my model because as shown in the graph displaying ENTRIESn_hourly by hour there is easily apparent variation in entries throughout the day. I dropped fog from the final prediction model because it lowered the r-squared value when paired with the other features. I only tested fog because I thought people might not want to go out on a foggy day. I think fog probably didn't have a significant impact because there were not many foggy days.





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

For some reason I am getting 243 coefficients, they can be seen below.

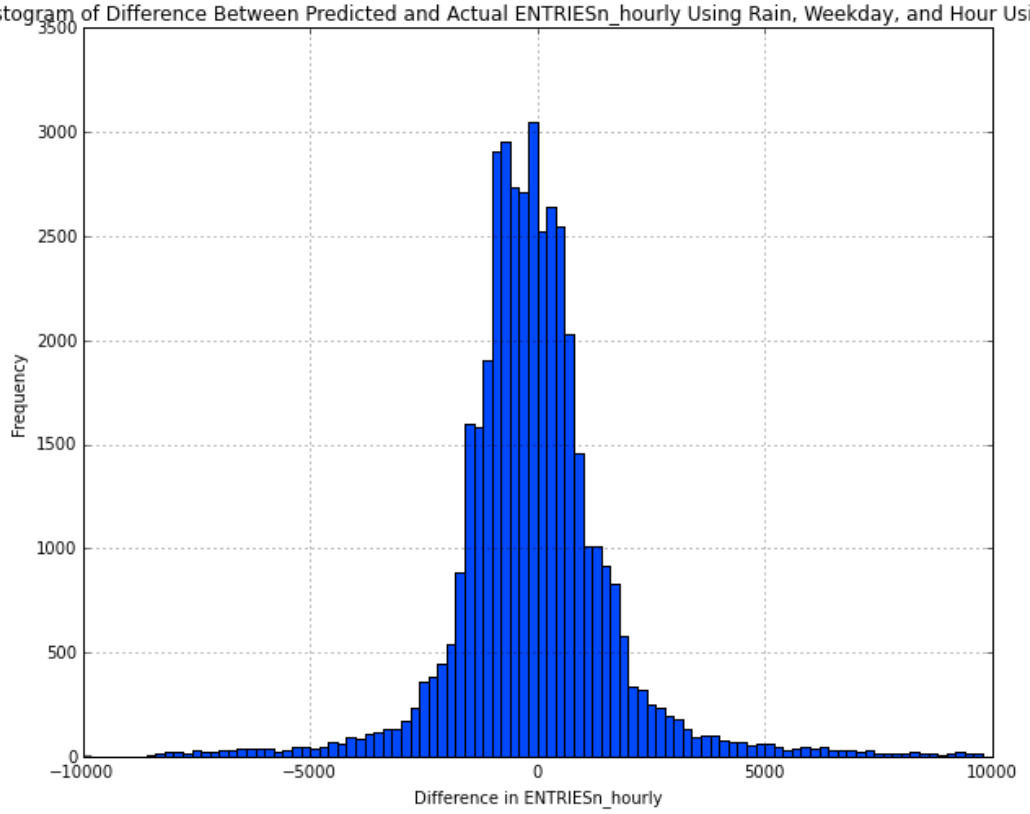
```
[ -29.56435321  199.25262378  240.50039783  -87.40544791  -67.819688
33   -66.79618424  -61.01323903  -77.50707251  -75.89859894  -79.849
97656   350.72037281  444.45482258   47.0542978   -65.31362109  137.
25870273   382.64592381   75.89659556  260.69099991  160.20736384  4
96.55101402   263.86618926   92.23473263  195.80117557   56.16575013
  301.3166379    29.57551781  148.4084839   133.25475088  395.432769
41  -57.98455828    36.60361524  -77.75032228  -74.90825511 -109.235
39939  -58.07837874   -24.66743801   75.22386591  -77.94226276   56.
05039899  165.1414673   395.28052366   36.09014962  114.50661266  17
3.91759816  -53.86471435    68.59336605  -34.96899617  408.55269955
 -38.41406716  202.0065343   -87.34057071  -52.55637957  -76.2533995
8  -80.75862736   45.5225721   -50.53484037  -69.23801519  -64.44303
949  -98.41503017  -56.78461664   -88.4219609   -51.74383596   -6.11
2638    104.28605523   94.39253818   -24.27572044   69.52333274  468
.25672967   23.00880927   59.91049578   -42.07133702  -91.29219326
```

```
-95.61301889  -53.24844503  -12.50080118   -30.9324703     8.2773758
6   17.89033918   32.64101701   80.54556725     1.76855644   28.1509
2047  -79.76221933   55.59743797  113.91774377   -19.80059181  -30.6
486407    79.76140055  -41.5452349   -77.41479213   191.38354882   8
0.8791877   -11.41220644  -58.91651015  -33.9010425    73.02720276
-35.06614455    3.5311573   -22.66723869  -18.45525126    52.3002939
5  -41.02282739  -65.07274057    8.88180974  154.6663004    30.01011
713   31.80769838   78.60689558   44.65464815  303.99316463    -0.78
912476  -65.56039717  -46.06664608  -54.18309092   26.57199846   -30
.93259775   -2.906383    -23.54260338   19.99681919  -62.1987691   -
44.19442701   30.01031792  -41.92581173  -30.22477479  -28.79734909
   11.41217361   47.23449507  -55.58333192  -76.48415982   53.627271
97    -2.45013824  -33.24383882  -63.2458346    -7.05889409  -60.951
81065   -49.34267011   26.6364816   -31.76619075  -16.06800311  -20.
98918132    30.02053006  -61.34507173  -72.19160055  -56.99122683  -
34.72891336   -38.35936407  -66.45286161  -69.05034315  -48.57012793
  -34.6911645   -23.92036976  -74.42440069   47.94034266  -26.220449
17  -59.52638598    40.25247999  -52.12988181   75.25176163  -65.992
65804  -19.82858503   -66.58459477  -59.79814098  -83.90235322   74.
61754695  -33.79999527   -67.46386903  -29.28088643  -44.02366765  -
65.2830277    62.44424668   -56.35052384  -48.47598767   18.91686885
   -8.22050869  -42.50047895   -22.18141704   -4.94111904  -69.48920
198  -95.38101479  -69.87959798   -57.4606695   -58.73321149  -47.74
631968  -67.80095706  -78.17936829   -17.68051381  -76.90859817  -45
.63758829  -22.69616977  -67.1155389   -76.06962398  -67.30458777  -
56.2793822   -21.77313268   -9.41226761   -57.45352009  -62.74911827
  -63.55641366   -5.19755543  -55.68522877   -70.06515674   28.80807
893  -27.47441386  -51.73670807  -76.89787292   -46.64955441  -54.33
321752  -16.46220056  -74.22040745  -56.11682781   -92.13874994  -70
.66860994  -30.23695724  -42.22264044   -2.18024681   -23.0480638
-79.39112473  -48.09893855  -69.85496975  -92.33838537   -86.1000559
6  -98.21498984  -70.16749568  -69.76816362  -71.35117601   -30.1741
3625  -90.9746132   -87.52963674  -36.74100654  -77.52986326   -64.0
194032   -51.18561329  -50.02455891  -58.76302677  -44.43092366   -7
2.57323665  -25.09497915   -5.3130838   -89.99032595  -91.21824883
 -86.99784948  -65.86052717 -100.54763556]
```

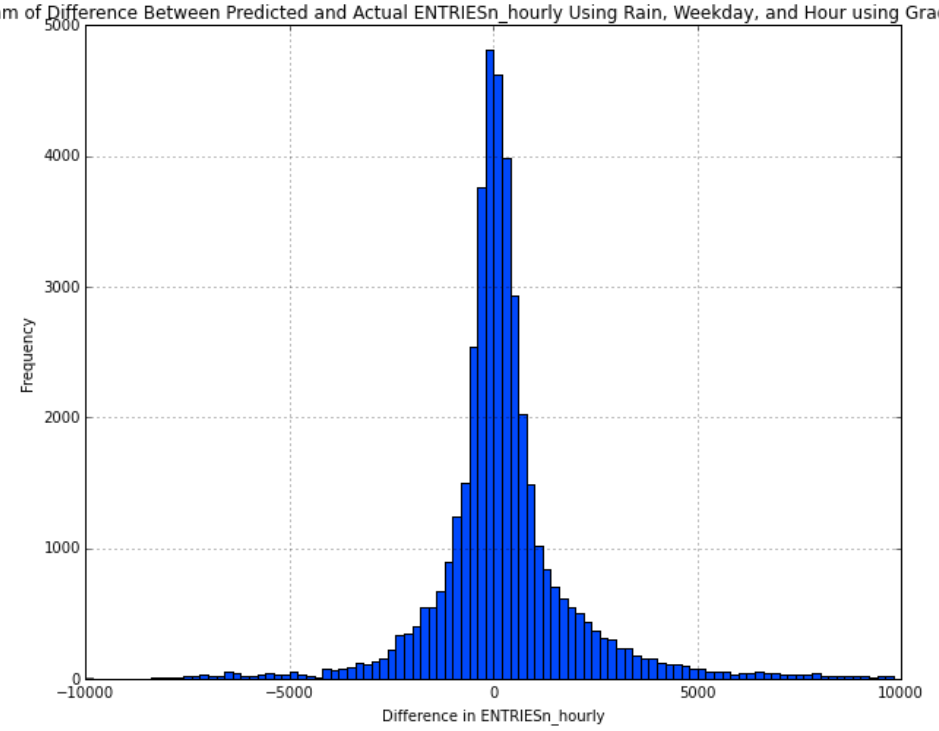
2.5 What is your model's R2 (coefficients of determination) value?

My models R2 is 0.408394170194 when using gradient descent. It is curious that the R2 value with OLS is greater at ~0.48 but appears to have a wider dispersion in the residual histogram.

Histogram of Difference Between Predicted and Actual ENTRIESn_hourly Using Rain, Weekday, and Hour Using OLS



Histogram of Difference Between Predicted and Actual ENTRIESn_hourly Using Rain, Weekday, and Hour using Gradient Descent



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?

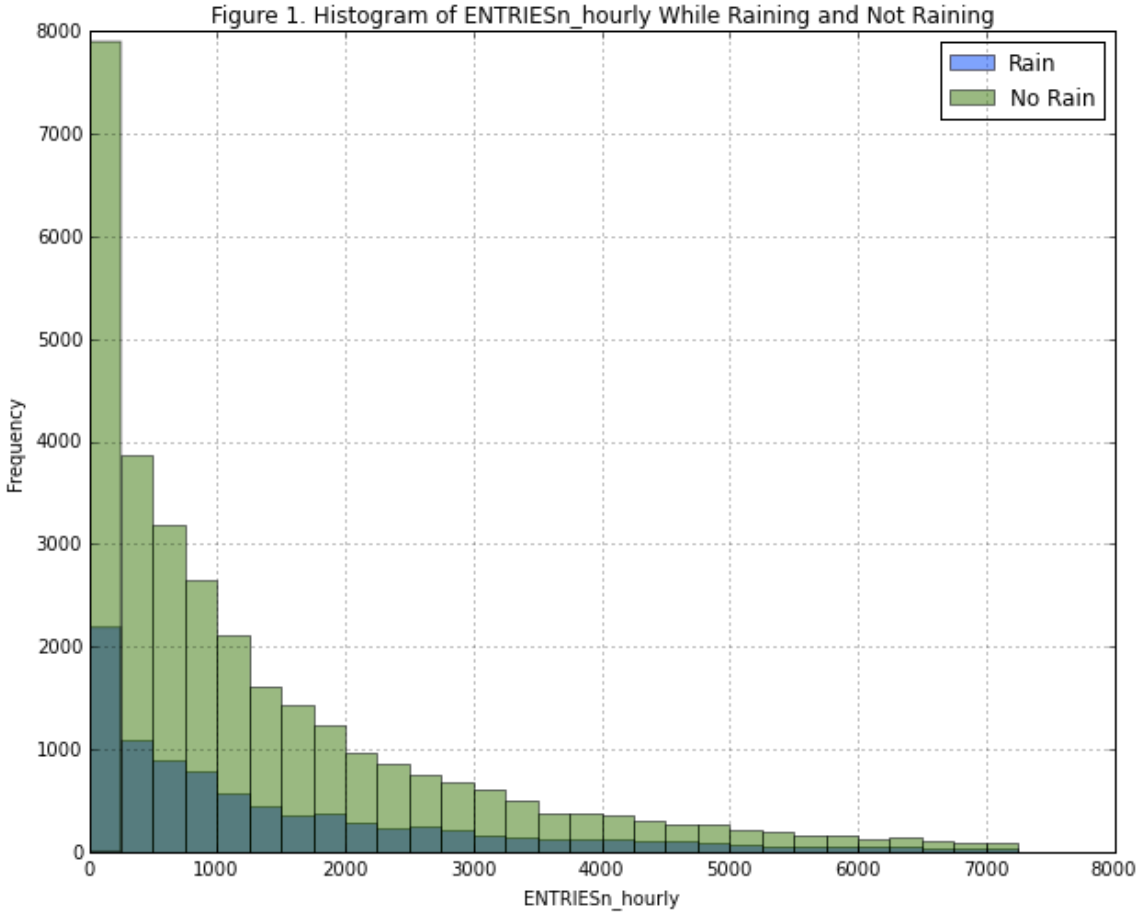
An R-squared value of ~0.408 is not the best and I believe there is room for improvement with my model. However, we are trying to predict human behavior based on the weather, which is difficult because humans can be unpredictable. Also when looking at the plot of residuals it appears that most of the time the model predicts ENTRIESn_hourly within a few hundred of the actual value.

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.

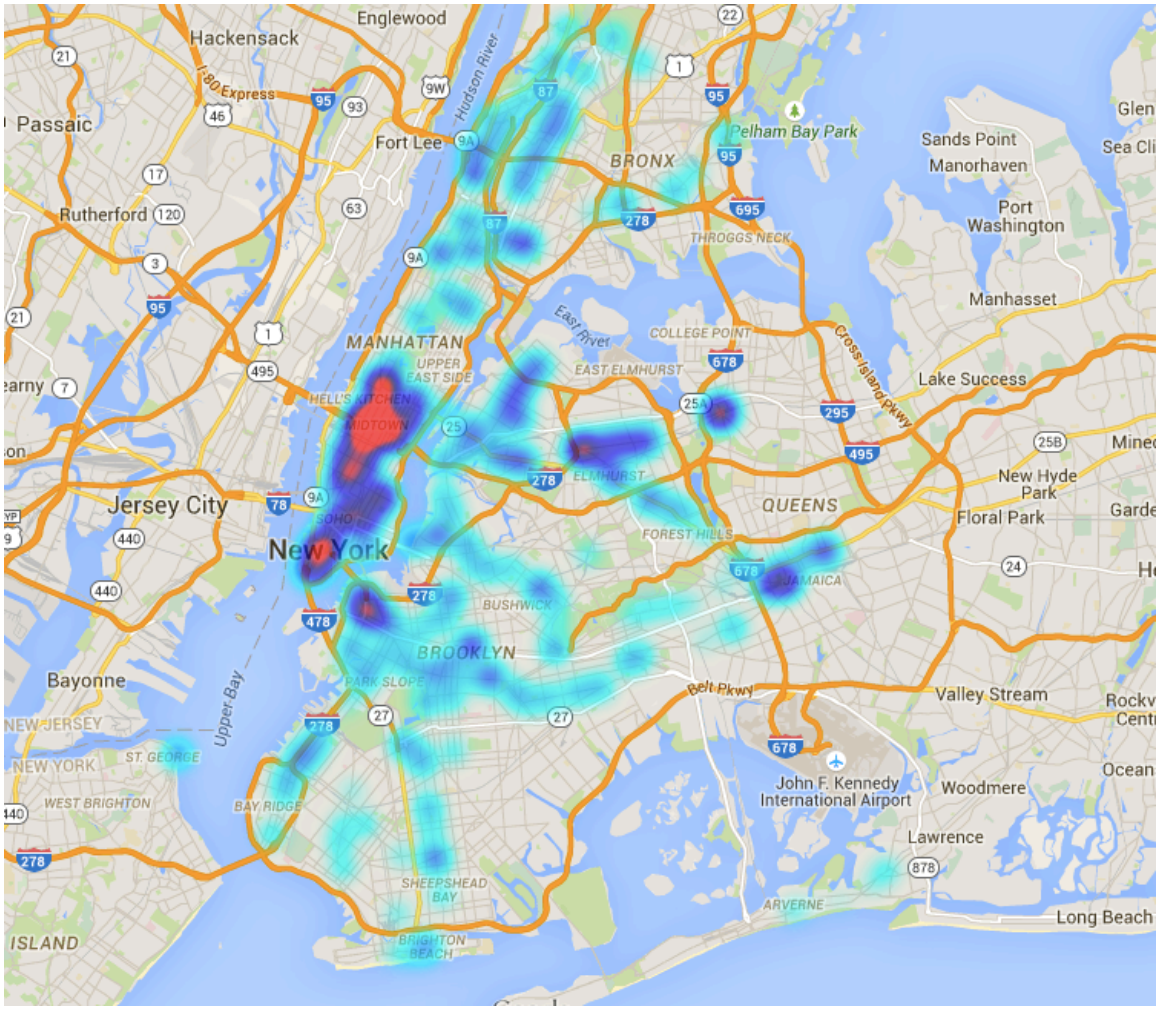
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days. You can combine the two histograms in a single plot or you can use two separate plots. If you decide to use to two separate plots for the two histograms, please ensure 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.



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

Shown in the answer for question 2.3 is a plot showing the mean ENTRIESn_hourly with time of day on the x-axis broken up by day of week. Also shown below is a heat map I made using the Google maps API that shows the locations of all of the stations with the weights of the heat map being the mean ENTRIESn_hourly for that station.



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?
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.

Yes, more people ride the subway on rainy days than when it is not raining. I am using ENTRIESn_hourly to represent ridership. The mean number of riders on rainy days is 2028 versus an average of 1985 on non rainy days. This alone is not statistically significant. This is why I ran a Mann-Whitney U test to assess for a statistically significant difference in the distribution of the datasets representing rainy days and not rainy days. As shown in the histogram plotting the frequency of ENTRIESn_hourly for rainy and non rainy days it is very apparent that the data is not normally distributed. Also apparent is the large difference in the size of the data sets, there are many more non rainy days than rainy days. This is why

the Mann-Whitney U test was chosen to test for a difference between the data sets. A Mann-Whitney U test is better suited to data fitting the previously described criteria than a standard t-test, which requires normally distributed data of relatively equal sample sizes. After finding that there are in fact more subway riders on rainy days the next step was to create a predictive model using this fact as a starting point. I then used scikit learn to create a linear regression model using rain as the only feature to predict ENTRIESn_hourly. This gave a R-squared value of 0.375. I think we can do better than that. Especially since there is probably variation in ridership on different days of the week and different times of the day. Since our dataset included these variables I added them to the linear regression model, which achieved a R-squared value of 0.481 using OLS and 0.408 using gradient descent. When plotting the residuals in a histogram we can see that over 16000 times the model correctly predicted ENTRIESn_hourly within 400 riders.

There are several shortcomings in my regression analysis and in the dataset that could be remedied to create a better model. First of all the time matters greatly to the number of riders on the subway which is an assumption that can be made with intuition alone, there will be more riders during rush hour. The shortfall in the dataset is that the ENTRIESn_hourly relies on readings that are taken every 4 hours. This will skew the results, as it will spread out the number of riders over a much larger time area than rush hour probably lasts. Also I think that the station could be used as a good predictor. Another thing that may have helped my model is to normalize the ENTRIESn_hourly field. It could be the case that the model was inaccurate in general. With a mean ENTRIESn_hourly of 1800-2000 it is clear when looking at the plot of residuals that there are many predicted values that are inaccurate by at least that mean. It could be the case that the few residuals that are off by over 5000 are the few data points that had a number of ENTRIESn_hourly that were that high. This could also mean that the ENTRIESn_hourly predictions that were accurate to less than 1000 ENTRIESn_hourly could be off by as much as 50% when comparing to the mean number of ENTRIESn_hourly. In conclusion I think the model could be improved by having more frequent data points available to better show change in ridership throughout the day. Also being able to include which station is being used at different times might be valuable because different stations might have more or less riders at different times of the day. Another data point that could be valuable is if particular stations connect to other transit systems such as metro north or Amtrak.

My next step to improve my prediction model is to group the dataset by station and take the mean ENTRIESn_hourly for each station. Then I will sort the stations by ENTRIESn_hourly and use the rank as a feature. I think that this will be a good way to further improve the model.