Analyzing the NYC Subway Dataset

Questions

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.

Documentation:

http://statsmodels.sourceforge.net/devel/examples/notebooks/generated/ols.html
http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
http://pandas.pydata.org/pandas-docs/stable/generated/pandas.get_dummies.html
http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html
http://docs.ggplot2.org/current/
https://docs.python.org/2/library/datetime.html
http://pandas.pydata.org/pandas-docs/stable/timeseries.html
https://books.google.com/books?id=4wMn27KSnTsC&lpg=PA59&ots=ZfxxFlOvd_&dq=matplotlib%20bar%20chart%20multiple%20variables&pg=PA65#v=onepage&q=matplotlib%20bar%20chart%20multiple%20variables&f=false
http://pandas.pydata.org/pandas-docs/stable/generated/pandas.Series.to_dict.html
http://matplotlib.org/1.3.0/examples/pylab_examples/barchart_demo.html
http://statsmodels.sourceforge.net/devel/generated/statsmodels.graphics.gofplots.qqplot.html

Discussions / Blogs:

https://discussions.udacity.com/t/problem-set-3-5-linear-regression/24118/6
https://discussions.udacity.com/t/linear-regression-prediction-code-only-outputs-errors/26686/2
https://discussions.udacity.com/t/problem-set-3-8/27076
https://discussions.udacity.com/t/problem-set-4-1-visualization-1/23401
https://discussions.udacity.com/t/problem-set-4-ggplot-not-accepting-date-format/19959

https://discussions.udacity.com/t/problem-set-4-1-errors-using-pandasql/19854
http://stackoverflow.com/questions/11346283/renaming-columns-in-pandas
https://discussions.udacity.com/t/problem-set-4-how-to-get-day-of-week/15617
https://discussions.udacity.com/t/cant-display-legend-in-ggplot/13269
https://discussions.udacity.com/t/problem-set-4-1-adjusting-position/19926
http://johnbeieler.org/blog/2013/06/06/using-sql/
http://www.w3schools.com/sql/sql_where.asp
http://stackoverflow.com/questions/19410042/how-to-make-ipython-notebook-matplotlib-plot-inline
http://stackoverflow.com/questions/13740672/in-pandas-how-can-i-groupby-weekday-for-a-datetime-column
http://stackoverflow.com/questions/17691447/get-count-of-values-across-columns-pandas-dataframe

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?

-The statistical test used to analyze the data was the Mann-Whitney U-Test.
-The p-value used was two-tailed
-The null hypothesis is, on average and given two unknown distributions, randomly selected values from one distribution are not likely to be higher than randomly selected values from the other distribution. In other words, the two populations, from which the distributions were created, are statistically the same.

$H_0: P(x > y) = 0.5$

-The p-critical or alpha 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 these two samples as the distribution of the samples were not 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 average number of entries per hour while it was raining was 1105.45 and the average number of entries per hour when it was not raining was 1090.28. The p-value returned from the Mann-Whitney U-Test was 0.024999912793489721. Because this p-value is for a one-tailed test, we will double it and find a true p-value of 0.049999825586979442.

1.4 What is the significance and interpretation of these results?

We reject the null as the p-value was below the alpha level of 0.05. When combined with the means of the two samples, we can state that subway ridership increases by a significant margin when it is raining in New Your City.

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. OLS using Statsmodels or Scikit Learn
b. Gradient descent using Scikit Learn
c. Or something different?

While I did use a model that incorporated Gradient Descent from Scikit, I found a model using OLS from Statsmodels to result in a higher R^2 value. All of the following answers will refer to the OLS model.

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

The features that I used to help improve the prediction of the model were as follows:
Rain, Amount of Precipitation, Hour of the Day, Fog, Maximum Dew Point, Maximum Pressure, Average Wind Speed, Minimum Temperature and Minimum Dew Point.

('rain', 'precipi', 'Hour', 'fog', 'maxdewpti', 'maxpressurei', 'meanwindspdi', 'mintempi', 'mindewpti')

I also used a dummy variable of the subway stations (UNIT).

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.

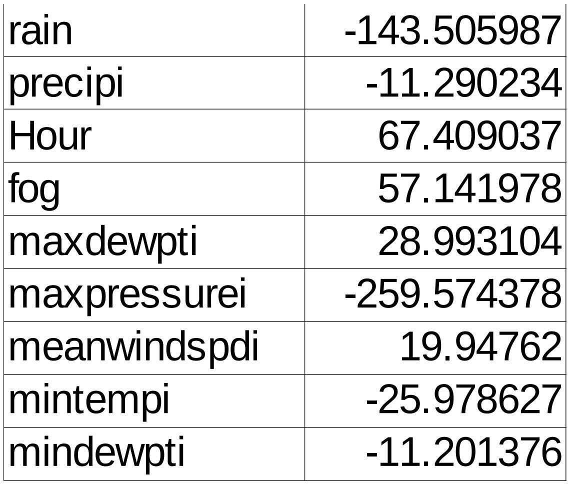
- Your reasons might be based on intuition. For example, response for fog might be: “I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often.”
- Your reasons might also be based on data exploration and experimentation, for example: “I used feature X because as soon as I included it in my model, it drastically improved my R^2 value.”

The following variables were chosen based on intuition:
Maximum Dew Point-Higher dew points would theoretically make walking in the city more uncomfortable and thus lead to more passengers on the subway.

Amount of Precipitation, Rain, Fog – Poor weather conditions should lead to more activity on the subway.
Hour of the Day – trends of activity during different parts of the day should help a prediction model a great deal.
Minimum Temperature for the Day - Different temperatures would theoretically influence one's decision to either use the subway or walk/take other forms of transportation.

The following variables were chosen based on trial and error, and finding that they improved the prediction:
Minimum Dew Point, Maximum Pressure and Average Wind Speed – I was actually surprised to see that these improved the predictability of the model. I assumed these factors would have no effect on one's decision to ride the subway and would thus not improve the predictability of our model.

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



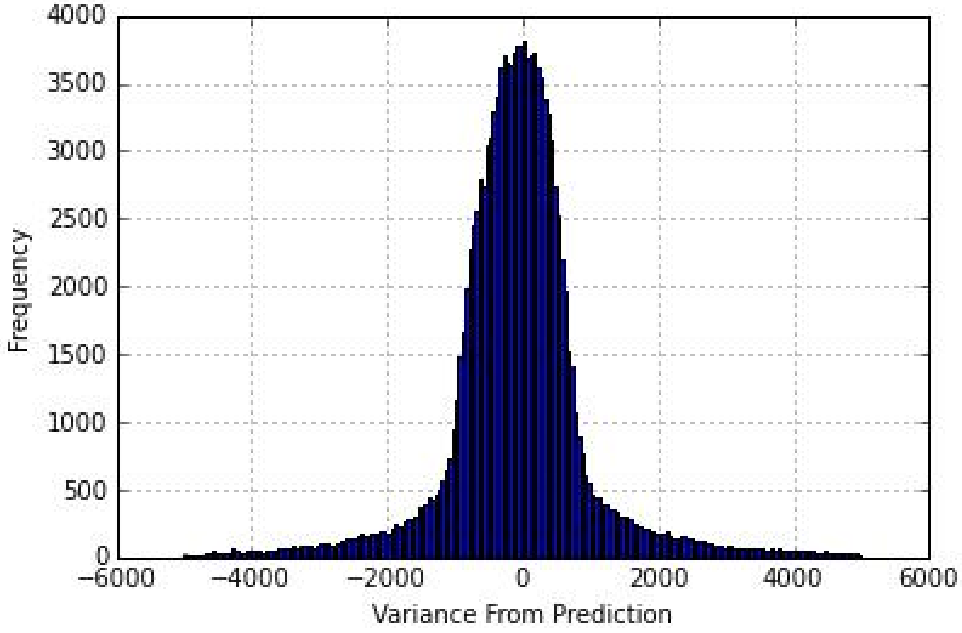
rain	-143.505987
precipi	-11.290234
Hour	67.409037
fog	57.141978
maxdewpti	28.993104
maxpressurei	-259.574378
meanwindspdi	19.94762
mintempi	-25.978627
mindewpti	-11.201376

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

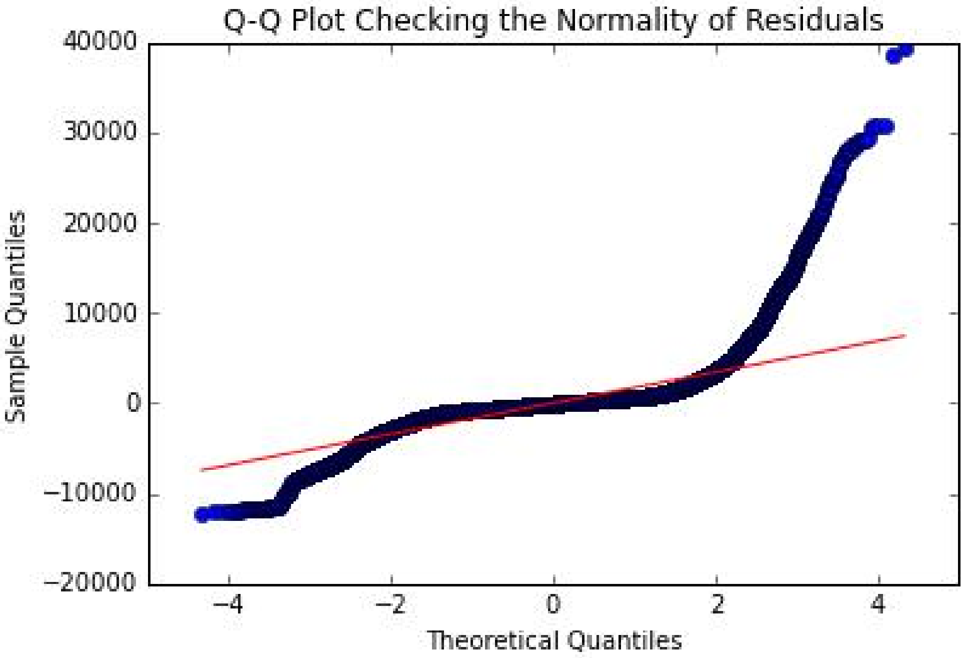
My maximum R^2 value achieved locally was 0.46019524171

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?

While R^2 values closer to 1.0 generally indicate a better fit of a regression model as the model would be accounting for more of the variability of the data, this is not always the case. Further information is required to determine goodness of fit and one such way is analysis of residuals.



When quickly analyzing of a histogram of the residuals of the model, we see long tails, which suggests that there are some large residuals. In order to determine the normality of the plot of residuals and fit of my model, I created a Q-Q Plot comparing actual number of entries given by the dataset with the model's prediction of entries given the same set of circumstances.



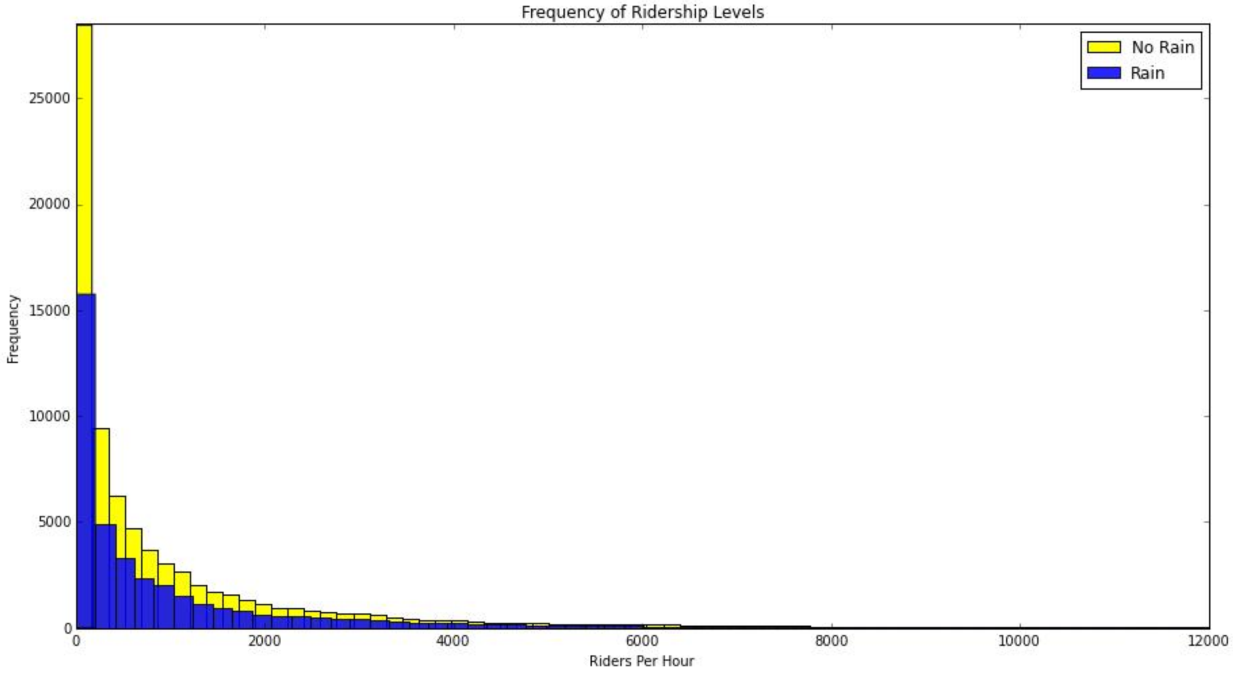
As we can see in the Q-Q Plot, the relationship between the model's predictions and actual data is distinctly not normal. Because of this and a relatively low R^2 value, it would appear that this linear regression model is not appropriate for the dataset.

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.

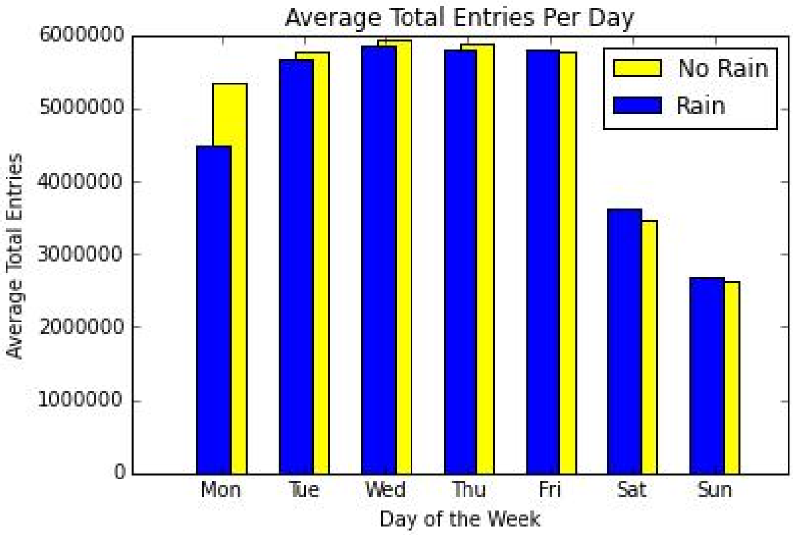
- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms,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



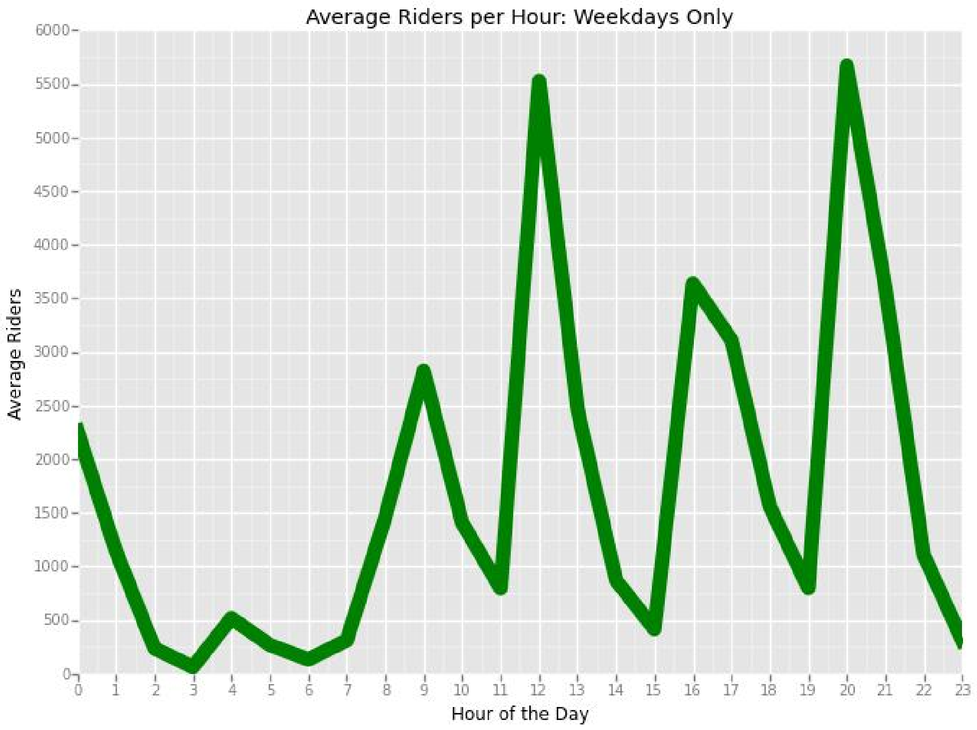
The histogram above shows the frequency of each level of ridership with respect to rain, over the period of time recorded in the dataset. As there were more Non-Rain days than Rain days, the frequency of ridership is higher among Non-Rain days and this visualization shows that fact. Upon closer inspection of the higher Riders Per Hour levels, we see that the trend visible in the visualization above does indeed continue throughout the lower frequency ridership levels.

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



In order to to look deeper at the question of whether NYC subway ridership is affected by rain, I created the visualization above, which shows the averaged number of total entries for each weekday with respect to rain. Interestingly, we see a slight difference in rider behavior during the beginning of the week, when compared with the weekend. On average, ridership is lower on rainy days than non-rainy days from Monday through Thursday, but the opposite is true for the remainder of the week.



I also thought it would be interesting to look at average ridership by time of the day to see if my assumptions about traditional behavior would be confirmed by the data and visualization. As we saw in the previous bar chart, there is a clear difference in behavior from weekdays to weekend days, so I narrowed the data to weekdays only to look at 'typical' rider numbers.

Spikes can be seen at "9-5" rush hour times and while I was initially surprised to see the afternoon peak around 4:00pm, this may be explained by certain stock exchanges closing around this time. I was also surprised to see that the highest average was at 8:00pm, so I excluded Friday, in case it was heavily biasing the data. However, I actually found that from Monday through Thursday, average ridership was higher at 8:00pm than when Friday was included.

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 a wholesale perspective over the span of this set of data, yes, on average more people rode the subway when it is raining than when it is not. This can be seen by looking at average subway entries on days with and without rain. As stated earlier, The average number of entries per hour while it was raining was 1105.45 and the average number of entries per hour when it was not raining was 1090.28.

Given that the coefficient for the rain variable in my regression model was negative, one might assume that rain would actually cause lower subway rider numbers. However, since this contradicts our statistical findings, it is much more likely that this coefficient indicates multicollinearity among the independent variables. This could be one factor leading to a lower than desired R^2 value. When combined with the existence of large residual values, it is clear that the linear regression model was not a good fit for the data.

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.

See Above

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.

1. The biggest shortcoming I can see with the data set used was that it covered a relatively short period of time (May 2011). By looking at only a single month, the data is prone to bias due to anomalies or one time events such as holidays, major sporting events, or civil disturbance. This is shown in the visualization "Average Total Entries Per Day", as one of the rainy days of this month happened to be Memorial Day, when one would assume that there would be less subway riders regardless of the weather. Additionally, by only looking at one month's worth of data, we are unable to make claims about ridership across seasons. One would assume that rider *behavior may be vastly different in the winter than in the summer and we would expect to see spikes during times of the year that are more conducive to tourism, as an example.*

One more issue I had with the dataset, with regard to our specific analytical goals was that any entry in the dataset that occurred on a day when it rained at least once in New York City was marked as an entry from a "rainy day", regardless of whether it was raining at the time that the data was recorded. For example, if it rained on Wednesday, May 4th from 9-11am, but was sunny for the rest of the day, data collected at 4pm would be marked as "rainy day". If the goal is to analyze subway rider behavior based on weather, than this shortcoming is surely detrimental.

2. The first shortcoming I found with the analysis of this data was that simply looking at average ridership on rainy and non-rainy days not only ignores factors that have more to do with ridership (such as time of day or subway station), it overvalues this factor and can lead to inappropriate conclusions. Additionally, using the R^2 value as the only or primary indicator of a model's predictability is foolhardy, as models with very low values can be quite predictable and models with high values can be extremely unpredictable.

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

I felt that looking at sums of data on a per-day basis or averages over the entirety of the data would not properly show the behavioral differences among riders with regard to weather. I wanted to factor in differences in ridership based on the day of the week, given that rider behavior changes during different times of the week and this can be seen when looking at the weekday results in comparison with the weekend.

I was a bit surprised to see that on average from Monday through Thursday, less riders used the subway when it rained than when it did not rain and the opposite was true for weekend days. One easy explanation for this is due to the nature of the ridership on each of these days. One would assume that more riders do so for business during the beginning of the week and pleasure during the end of the week. A rainy early week day in New York City may result in more people working from home or taking taxis instead of riding the subway. This visualization begs the creator (and the viewer) to do further research into the cause of this result.