Analysing the NYC Subway Dataset

P1, Intro to Data Science

Anthony Munnelly

May, 2015 Data Science Nanodegree Tranche

Section 0. References

Python for Data Analysis, by Wes McKinney	http://shop.oreilly.com/product/0636920023784.do
Ordinary Least Squares in Python	http://www.datarobot.com/blog/ordinary-least-squares-in-python/
Linear Regression Models with Python	http://mpastell.com/2013/04/19/python_regression/
Linear Regression with Python	http://connor-johnson.com/2014/02/18/linear-regression-with-python/
Machine Learning with Python - Linear Regression	http://aimotion.blogspot.ie/2011/10/machine-learning-with-python-linear.html
Basic Linear Regressions in Python	http://jmduke.com/posts/basic-linear-regressions-in-python/
Intro to Pandas Data Structures	http://www.gregreda.com/2013/10/26/intro-to-pandas-data-structures/
Problem with ggplot geom_histogram()	https://github.com/yhat/ggplot/issues/417
Regression Analysis with Python	http://work.thaslwanter.at/Stats/html/statsModels.html#linear-regression-analysis-with-python
Finding a Meaningful Model	http://www.esri.com/news/arcuser/0111/findmodel.html
What are t-values and p-values in Statistics?	http://blog.minitab.com/blog/statistics-and-quality-data-analysis/what-are-t-values-and-p-values-in-statistics

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 Welch's T-Test to analyse the NYC subway data. I used a two-tailed P value, with a null hypothesis that weather has no effect on subway ridership. The test was two-tailed because it is not necessarily the case that people always dislike getting caught in the rain, as Rupert Holmes' dreadful *Piña Colada Song* attests. My p-critical value was 0.2695 (if I understand the question correctly).

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.

Welch's T-Test is applicable to the dataset because it is particularly appropriate to use in in samples of unequal variance. If the variance of both samples is the same, Welch still works. If the variances are different, Welch's test is more accurate than the Student's T-Test.

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.

p-value	0.2965
Mean hourly entries for rainy days	1105.4
Mean hourly entries for dry days	1090.3

1.4 What is the significance and interpretation of these results?

The high p-value suggests that there is no evidence to reject the null hypothesis. As such, it is my contention that rain had little if any effect on subway ridership patterns on the New York subway in May, 2011.

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 OLS, ordinary least squares, from statmodels.api.

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 seven features:
Rain;
Hourly Exits;
Mean temperature;
Mean Pressure;
Fog;
Precipitation,
And whether or not the day was a working day or a weekend (a dummy variable).

2.3 Why did you select these features in your model?

Trial and error was my chief method of feature selection. The means for pressure and temperature were obvious, as all three variables for each measurement would lead to collinearity. In one iteration I used days of the week as a dummy variable, but there were too many of them – categorizing them as workdays or

weekends proved much more effective. I would have loved to have used the Units, but there were too many of them. If they had not been anonymous – that is to say, if it were possible to identify which parts of the city they were in – they may have been more useful. More on that below.

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

EXITSn_hourly	1,730.1703
workday	1,155.1175
weekend	953.6157
fog	3.1004
precipi	-8.3820
meanpressurei	-12.5696
meantempi	-14.3733
Rain	-21.9242

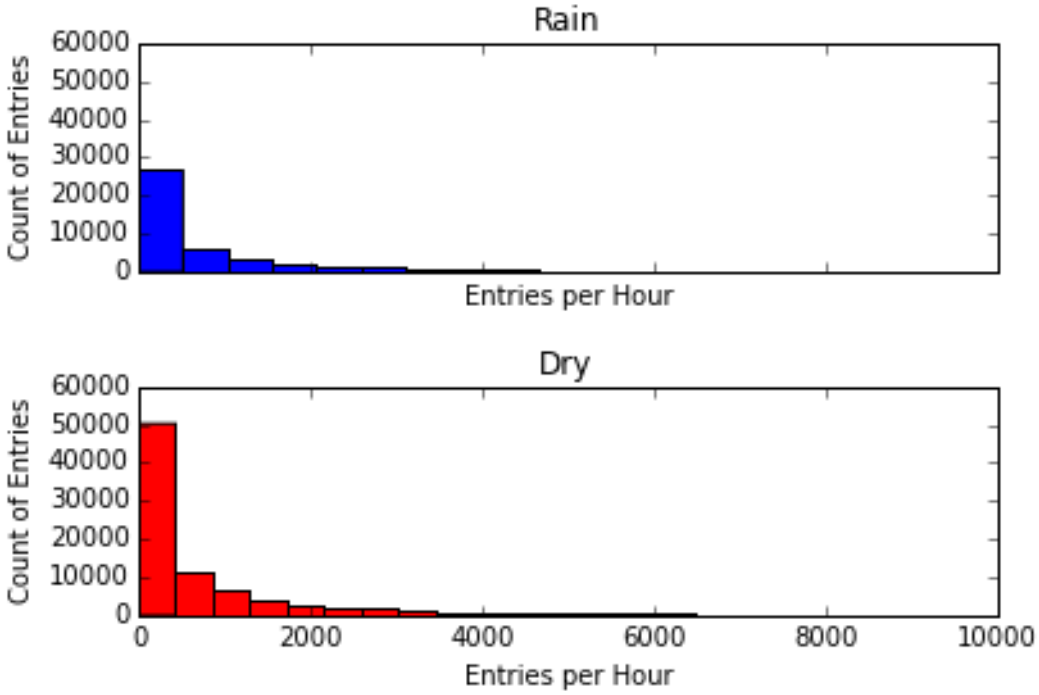
2.5 What is your model's R2 (coefficients of determination) value?

0.636

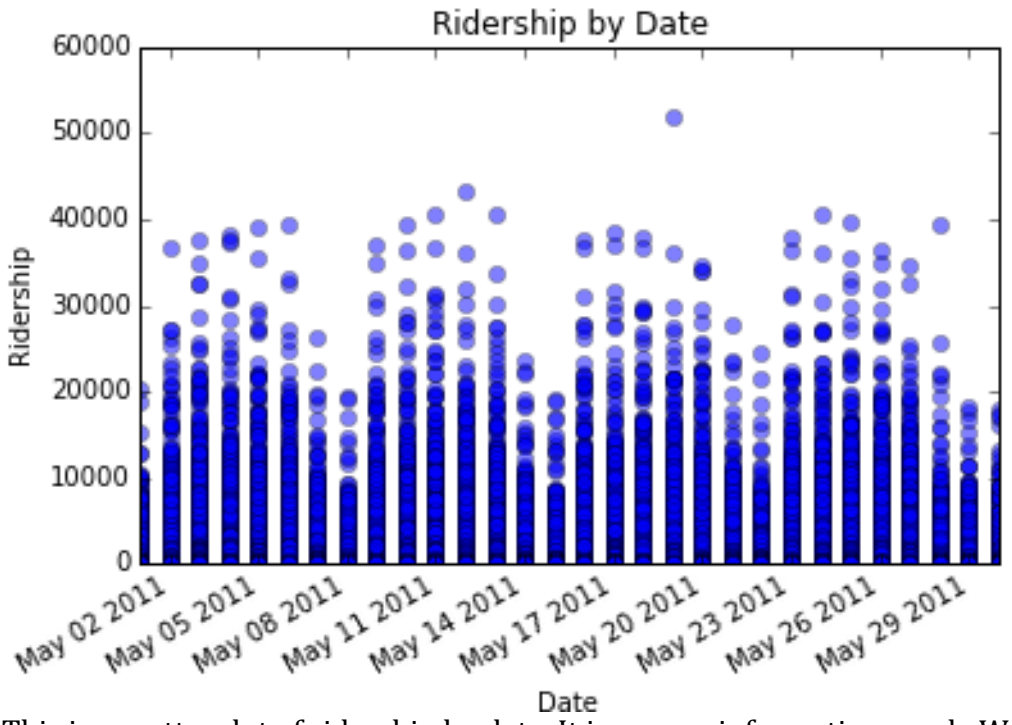
2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

An R2 value of 0.636 is not good enough to say that this model can successfully predict ridership on the subway.

Section 3. Visualization



These histograms show the entries per hour into the New York Subway system in May of 2011 on days which were rainy and days which were dry. The histograms show similar patterns. There is a higher entry count for dry days, as there were more dry days recorded. The rainy days histogram skews slightly more right than the dry. The histograms are both plotted on the same x-axis for clarity.



This is a scatterplot of ridership by date. It is a more informative graph. We can see a repeating pattern that represents five workdays and two weekend days. We also notice a standout outlier on May 19th, which is discussed below.

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?

There is no link between wet weather and subway ridership. The distributions are not that different from each other, and they have similar median values.

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 data show no strong causal link between weather and subway ridership. The distributions are more or less the same, and there are many more factors effecting subway ridership than simply the weather, as addressed below.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis.

1. The current dataset is not ideal for investigating a correlation between weather and subway ridership, for the following reasons.
2. Rain is not a binary event. There are different gradations of rain. In Ireland, we use the lovely term “soft day” to describe a particular type of

rain. Rain should be categorized under more headings than a simple true or false.

3. One month at the beginning of summer is not an ideal date range for a weather study. A data set from a full year would be better.
4. Time and location are, logically, the biggest factors in dictating subway ridership. People will travel more on workdays than weekends. People travel more at the rush hours than at off-peak times. An investigation based on the Units rather than on the weather could have taught us more about subway ridership patterns.
5. There's an outlier on the second graph in section 4. A little subsetting and googling revealed that the busiest data point occurred at nine in the evening at the subway station at Union Square and 14th Street on May 19th, 2011, a Thursday. It's hard to say why this should be, but it's unlikely to have been caused by the weather. A next-stage project would be to look at this outlier, and try to figure out what happened.

Section 6. Code

https://github.com/amunnelly/Udacity_Projects/blob/master/nyc_subway_project.py