ANDREAS STORZ – JUNE 8, 2015

Project 1 – Analyzing the NYC Subway Dataset

Introduction

This is the first project in a series of five as part of Udacity's Data Analyst Nanodegree. The goal is to derive insights into the volume of ridership of the New York City subway based on statistical testing. Specifically, this project is geared toward investigating whether rain has any influence on how many people use New York's subway system on any given day. The data given are for the month of May 2011; the original dataset can be downloaded here. It includes information on the hourly entries and exits of the NYC subway system along with records on the weather and individual unit identifiers.

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?

Comparing numerical averages for the two samples, i.e. hourly entries on rainy days vs. days without rain suggests that the weather does have an effect:

	Mean (hourly entries)	Median (hourly entries)
Rain	2028	939
No rain	1846	893
Difference	182	46

However, only a robust test can tell us whether these differences are statistically significant. I ultimately decided to apply a **Mann-Whitney U test** to check whether there is a difference in the means of ridership on days with and without rain. Additionally, I also performed **Welch's t-test**, a type of t-test comparing means across two different samples. As the Mann-Whitney U test only checks whether there is a difference in means, it cannot give any information about directionality. For Welch's test, on the other hand, I decided to use a two-tail P value. The reason is that we do not know a strict direction that we could test. It may be that more people use the subway when it is raining as most of its facilities are covered (trains, stations, and walkways). But it may also be the case that more people opt to stay at home, use their own cars or organize some type of ride-share when it is raining, therefore ridership might go down on rainy days. Or it could be that people use the subway consistently at levels irrespective of the specific weather situation, therefore rain might not affect ridership at all. As a consequence, a two-tailed test is deemed appropriate.

The underlying null hypothesis is that there is no difference in the true means across the two samples (rain vs. no rain). Assuming a 95%-confidence level, the relevant alpha (and thus the p-critical value) is 0.05. Therefore, we would only reject the null hypothesis if a difference as extreme as the one observed (or larger) would be occurring fewer than 5 times out of 100 as the result of random variation if the true difference was in fact zero.

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 does neither assume equal sample size nor equal variance across the two samples. This is appropriate for the current situation as there are significantly more days without rain as there are rainy days in the data set. (Rain was only observed on 10 out of 31 days.)

But t-tests make the assumption of random samples that follow a normal distribution. (In reality only the sample means need to be randomly distributed, but normality is seen as an indicator of normally distributed means.) A look at the histograms (see Section 3) shows that the two samples do not follow a normal distribution. A **Shapiro-Wilk test** confirms this:

Sample	W	p
Rain	0.594	0.0
No rain	0.596	0.0

The Shapiro-Wilk test returns a highly significant result, which provides evidence that the underlying assumption of normality is violated. (Additional evidence is given by that fact that the observed means are much larger than the corresponding medians, indicating that the data is heavily skewed to the right.)

However, with large samples it is very easy to get significant results from the Shapiro-Wilk test due to even relatively small deviations from normality. Our samples are fairly large with 9585 (rainy) and 33064 (no rain) observations, respectively. Furthermore, t-tests become increasingly robust to deviations from normality with increasing sample sizes: following the Central Limit Theorem, "sample means of moderately large samples are often well-approximated by a normal distribution even if the data are not normally distributed" (http://en.wikipedia.org/wiki/Student%27s_t-test). In our case, Welch's t-test will likely yield valid results despite the non-normal distributions of the two samples.

Nonetheless, I opted to also perform a Mann-Whitney U test, as it is a non-parametric test that "does not assume that the data are drawn from any particular underlying probability distribution" (Lesson 3, Intro to Data Science).

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.

These two tests yielded the following results:

Test	Test statistic (t / u)	p
Welch's t-test	5.043	0.000
Mann-Whitney U test	153635120.5	0.000

As reported above, the means for the two samples were 2028 (rain) and 1846 (no rain) hourly entries into the NYC subway, respectively.

1.4 What is the significance and interpretation of these results?

Both tests yield results indicating that the difference of 182 entries per hour across the two samples is statistically significant, with both p-values approaching zero. Thus, we can be fairly confident that the observed difference is not due to random variation, but instead represents a systematic difference. In real terms then, the main takeaway is that on average more people use the subway when it is raining compared to times without rain. This finding can be further investigated using linear regression, which simultaneously takes into account the effects of other additional variables.

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 performed OLS using Statsmodels.

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

I tested a variety of features in my linear regression model. Most importantly, the inclusion of the dummy variables for the individual units registering entries and exits into the subway system dramatically improves the overall R^2 (cf. below). This makes sense because the units collect turnstile information across stations. Some of the units record much more traffic than others. Ignoring this source of variation results in a much weaker model. Therefore, I retained the dummy variables for the 240 units in my final model. I also tested whether the individual subway stations ('station') might be more suited as dummy variables, but this implementation resulted in an overall lower R^2.

In addition to that, I included the 'hour' variable as dummies because there should not be a linear relationship between the time of day and the volume of the ridership. This, again, resulted in a significantly higher R^2.
The other features in my final model were:

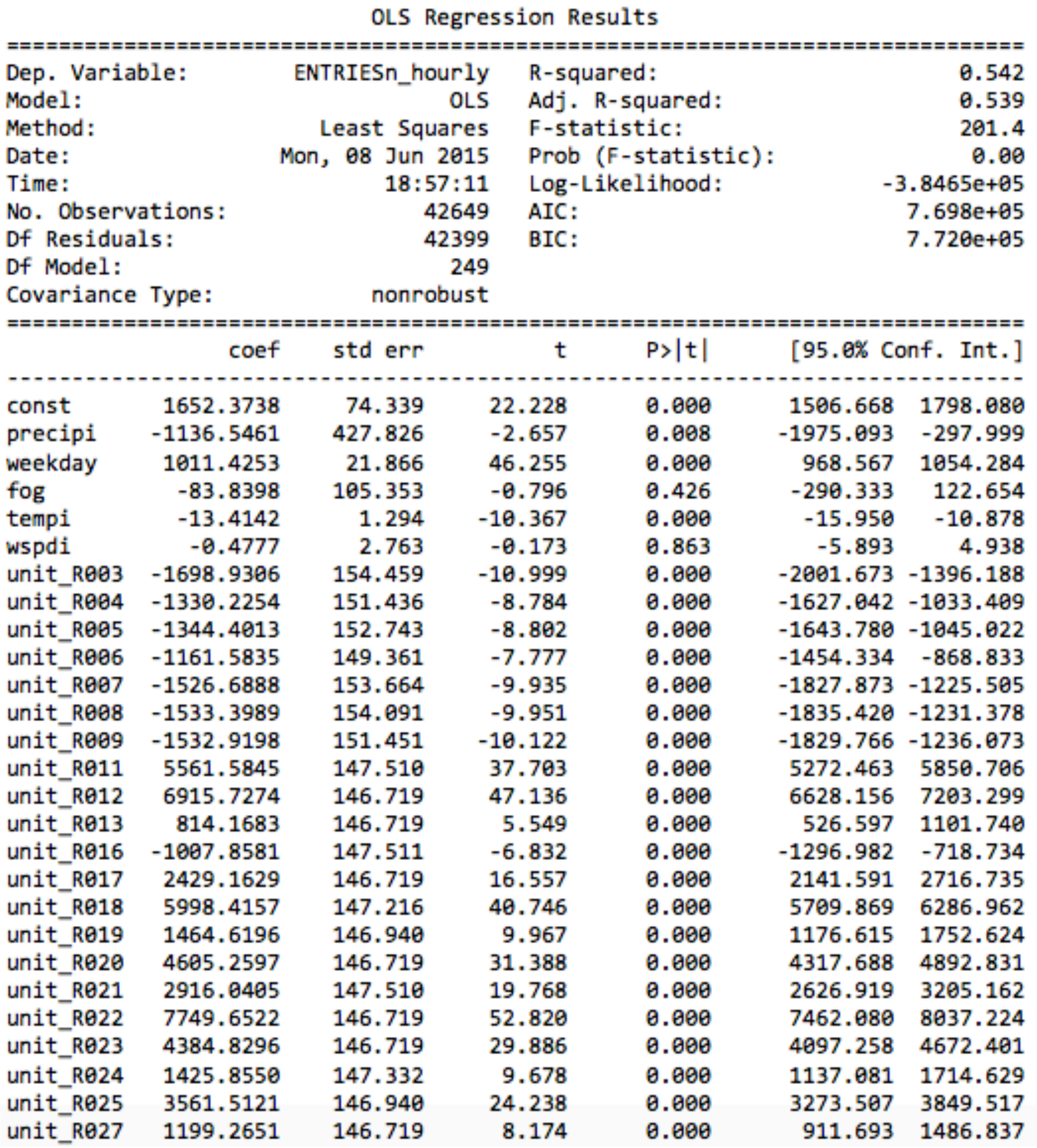
- precipi (precipitation)
- weekday (day of the week)
- fog
- tempi (temperature)
- wspdi (wind speed)

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 initially decided on including my features based on intuition, but in the end I excluded a number of those that did not have any significant impact on the model. Thus, I included precipitation as a more specific measure of the occurrence of rain. Fog, temperature, and wind speed are complementary weather features that should be of predictive value if rain affects ridership. Finally, controlling for the day of the week – in particular whether or not it is the weekend – should yield predictive power.

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

The individual coefficients or weights can be seen below. Given that the summary output for the final regression model was quite long (mostly due to the many dummy variables), I only show a partial output. However, it contains the most important summary statistics as well as the core info for the main features. The data for the parameters can be found in the second column under 'coef'. (The complete output is available upon request.)



```
                            OLS Regression Results
==============================================================================
Dep. Variable:          ENTRIESn_hourly   R-squared:                       0.542
Model:                              OLS   Adj. R-squared:                  0.539
Method:                   Least Squares   F-statistic:                     201.4
Date:                  Mon, 08 Jun 2015   Prob (F-statistic):               0.00
Time:                          18:57:11   Log-Likelihood:            -3.8465e+05
No. Observations:                 42649   AIC:                         7.698e+05
Df Residuals:                     42399   BIC:                         7.720e+05
Df Model:                           249
Covariance Type:              nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const       1652.3738     74.339     22.228      0.000      1506.668  1798.080
precipi    -1136.5461    427.826     -2.657      0.008     -1975.093  -297.999
weekday     1011.4253     21.866     46.255      0.000       968.567  1054.284
fog          -83.8398    105.353     -0.796      0.426      -290.333   122.654
tempi        -13.4142      1.294    -10.367      0.000       -15.950   -10.878
wspdi         -0.4777      2.763     -0.173      0.863        -5.893     4.938
unit_R003  -1698.9306    154.459    -10.999      0.000     -2001.673 -1396.188
unit_R004  -1330.2254    151.436     -8.784      0.000     -1627.042 -1033.409
unit_R005  -1344.4013    152.743     -8.802      0.000     -1643.780 -1045.022
unit_R006  -1161.5835    149.361     -7.777      0.000     -1454.334  -868.833
unit_R007  -1526.6888    153.664     -9.935      0.000     -1827.873 -1225.505
unit_R008  -1533.3989    154.091     -9.951      0.000     -1835.420 -1231.378
unit_R009  -1532.9198    151.451    -10.122      0.000     -1829.766 -1236.073
unit_R011   5561.5845    147.510     37.703      0.000      5272.463  5850.706
unit_R012   6915.7274    146.719     47.136      0.000      6628.156  7203.299
unit_R013    814.1683    146.719      5.549      0.000       526.597  1101.740
unit_R016  -1007.8581    147.511     -6.832      0.000     -1296.982  -718.734
unit_R017   2429.1629    146.719     16.557      0.000      2141.591  2716.735
unit_R018   5998.4157    147.216     40.746      0.000      5709.869  6286.962
unit_R019   1464.6196    146.940      9.967      0.000      1176.615  1752.624
unit_R020   4605.2597    146.719     31.388      0.000      4317.688  4892.831
unit_R021   2916.0405    147.510     19.768      0.000      2626.919  3205.162
unit_R022   7749.6522    146.719     52.820      0.000      7462.080  8037.224
unit_R023   4384.8296    146.719     29.886      0.000      4097.258  4672.401
unit_R024   1425.8550    147.332      9.678      0.000      1137.081  1714.629
unit_R025   3561.5121    146.940     24.238      0.000      3273.507  3849.517
unit_R027   1199.2651    146.719      8.174      0.000       911.693  1486.837
```

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

As can be seen from the output, the value of the coefficient of determination is about 0.54.

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?

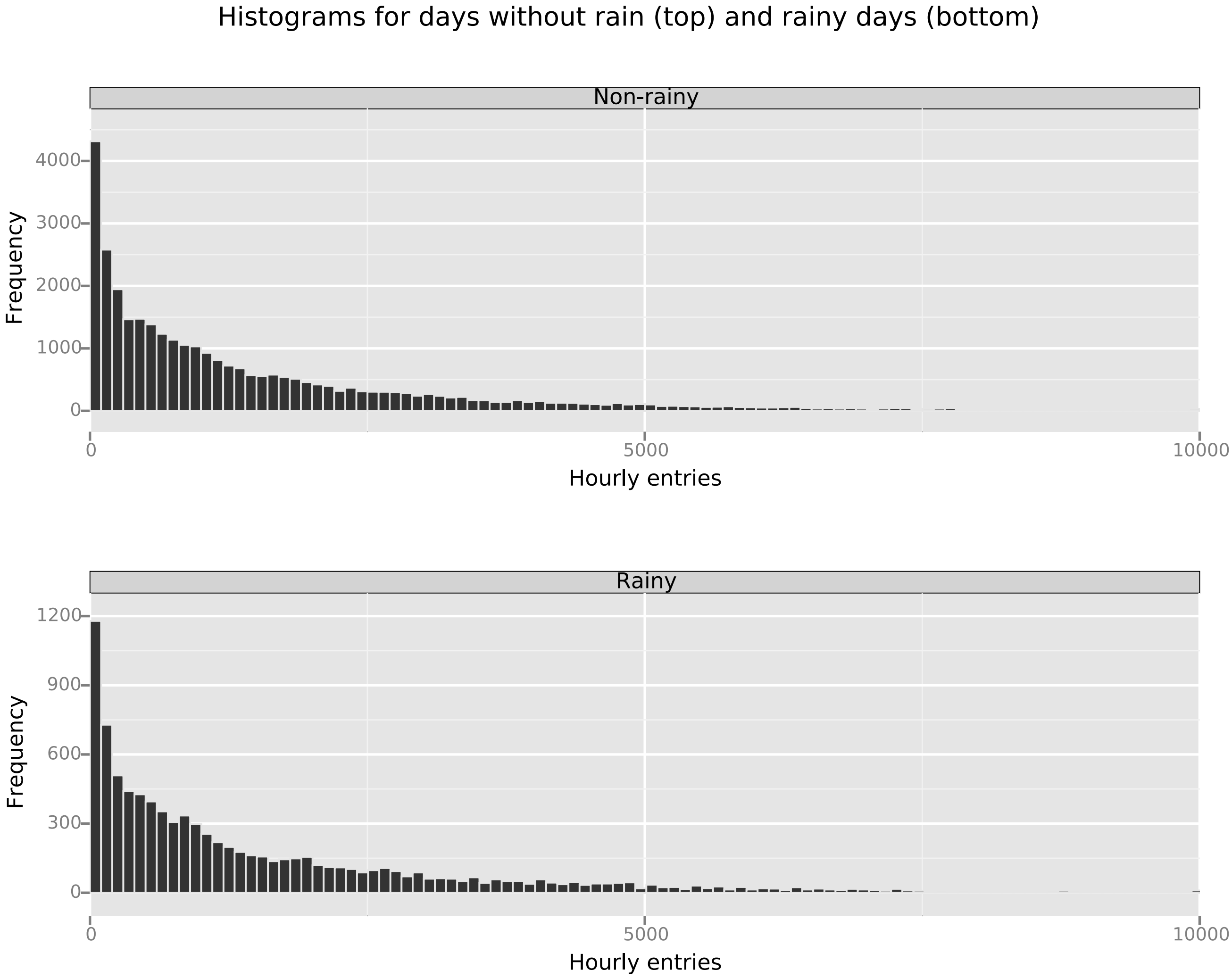
The R^2 value of 0.54 tells us that about 54 percent of the variation in ridership (measured in hourly entries) can be accounted for by our knowledge of the other variables included. This suggests that the linear model is relatively appropriate as a tool to predict ridership, but it can certainly be improved as almost 45 percent of the variability hinges on other factors not included in the model.

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.

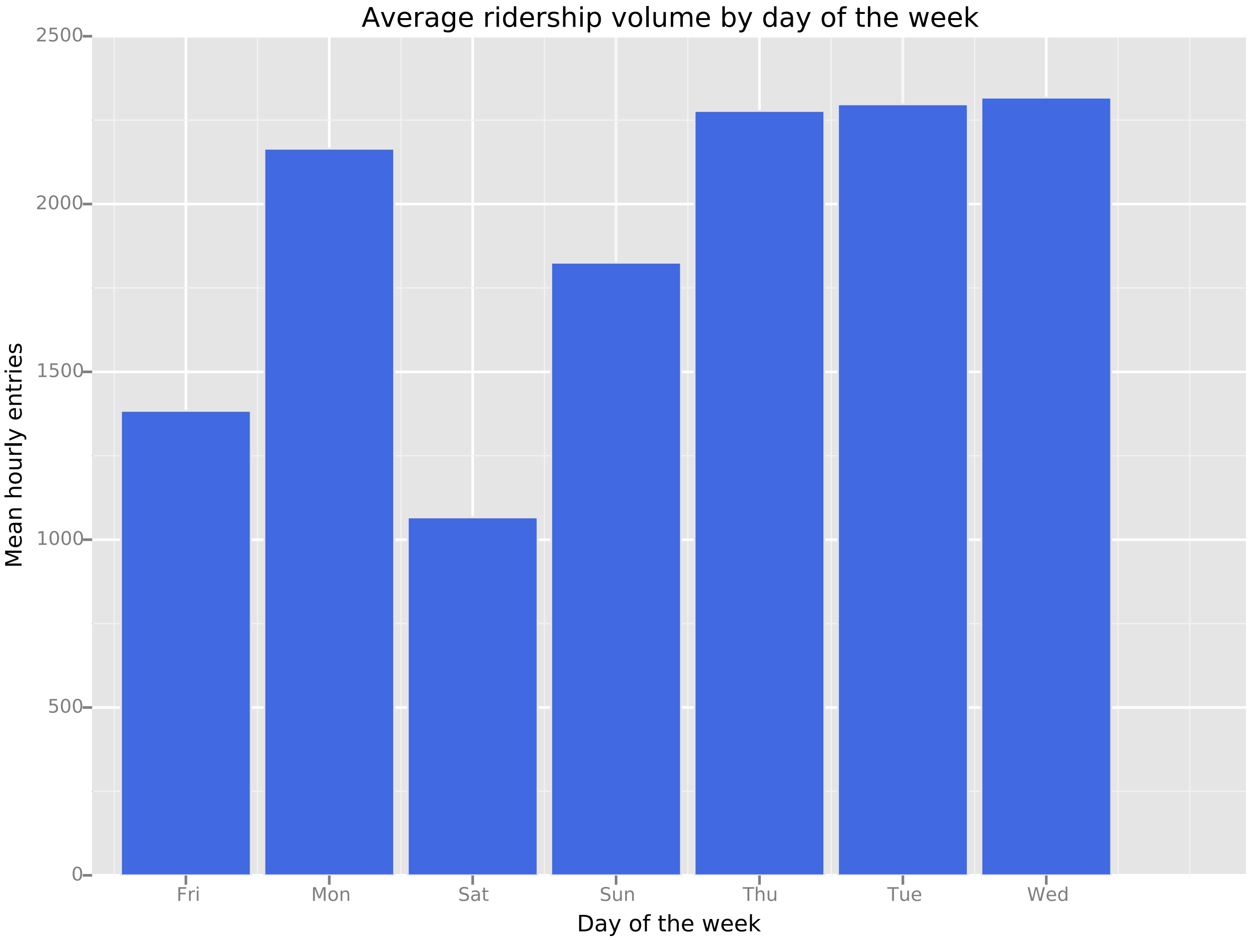
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

The following histograms compare ridership measured in hourly entries across days without rain and 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:

I decided to create a bar plot showing the average hourly entries by day of the week:



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.

At this point, it seems like there is indeed a statistical difference between the average ridership volume during times when it is raining compared to when it is not. Rain seems to drive more people toward using the subway – at least in NYC. The mean difference in hourly entries was a robust finding, which both the results from the Mann-Whitney U test and the Welch's t-test confirmed.

However, it is curious to note that the coefficient for precipitation in the linear regression was consistently negative, which would suggest a negative relationship, i.e., more precipitation implies *lower* ridership volume. This goes against the findings of the statistical test. Yet it may well be that there are underlying non-linear patterns that are not appropriately captured by the linear regression model. In addition to that, most of the variance is actually explained by the different units, and not by the external weather conditions. This is no surprise, as public transit use tends to follow regular patterns, and particular routes are traveled much more often than others. Thus, the ratio across units will be fairly stable, and will mostly vary by the time of day and day of the week (cf. second plot) – much more so than variation due to weather patterns. Nonetheless, the negative coefficient for precipitation in the model is an issue that should be further explored.

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.*

There are obvious limitations when it comes to the data set. While the data is complete for the month of May 2011, it still only captures information for a single month. We would at least need to further investigate whether this particular month was relatively representative before generalizing our findings. Maybe the weather was relatively uncommon, or other factors influenced ridership numbers, e.g., large-scale events, strikes, construction projects, etc. Moreover, more data is (almost) always better, therefore we would like to compare this particular month to other months. Ideally, we would have data for an entire year, thereby controlling for seasonal patterns due to holidays, regular events, economic cycles, etc. Still, having this much data for the entire month is quite impressive, and we should be able to glean relevant insights from it.

Beyond the available data, it is also important to note that the statistical tools used here were relatively limited. As alluded to above, there may well be non-linear patterns in the data that we have not been able to account for. Linear regression assumes a linear relationship between the variables, but this may not be appropriate. At this point, we have not tested whether adding non-linear elements to the (overall still linear) regression model (e.g., splines) might improve it. In addition to that, non-linear machine learning techniques would very likely lead to a significantly higher prediction accuracy since they can better account for non-linearity. Yet this would come at the price of being less (easily) interpretable. In light of this, the current model seems appropriate as a first cut.

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

Not at this point.

Section 6. 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.

- http://stackoverflow.com/questions/22179119/normality-test-of-a-distribution-in-python
- http://en.wikipedia.org/wiki/Student%27s_t-test
- http://en.wikipedia.org/wiki/Coefficient_of_determination
- Field, Andy. Discovering statistics using IBM SPSS statistics. Sage, 2013.
- https://github.com/yhat/ggplot/issues/315