Analyzing the NYC Subway Dataset

Short 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, 4, and 5 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 1. Statistical Test**

1.1 Which statistical test did you use to analyse the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis?

**Welch’s t-test**

I used a two-tail t-test. (See the SubwayDatat-test.py code.)

The null hypothesis is there is no significant difference between the hourly entries with rain and hourly entries without rain. That is, the sample means of hourly entries from the two sample sets (“with rain” and “without rain”) are equal.

Since I had already performed the **Mann-Whitney U test** as part of Problem Set 3, I will also interpret the results.

The null hypothesis is that the two samples (hourly entries “with rain” and “without rain”) come from the same population or in other words, there is no difference in the distributions of the two groups.

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 first question in Problem Set 3 asks me to plot two histogram charts and, from eyeballing the chart, decide whether the data is normally distributed and whether I can use the Welch’s t-test on the data. My initial instinct was the data isn’t normally distributed, but rather skewed to the left. Later, Problem Set 3 asks me to calculate the Mann-Whitney U test, a non-parametric test, so I felt even more certain the data wasn’t normally distributed.

However, for this project, I decided to run the Shapiro-Wilk test that the instructor mentioned briefly to determine whether the sample of data conforms to a normal distribution. (See the script SubwayDataShapiroWilk.py for code.)

**Shapiro-Wilk test**

The results were:

UserWarning: p-value may not be accurate for N > 5000.

warnings.warn("p-value may not be accurate for N > 5000.")

[(0.5938820838928223, 0.0), (0.5956180691719055, 0.0)]

I researched how to interpret this result (<http://mvpprograms.com/help/mvpspc/distributions/NormalityTestingGuidelines> and <http://www.graphpad.com/support/faqid/959/> were good sources). Some sources said the Shapiro-Wilk test is for best for 3-5000 sample size. Because the turnstile data is a very large sample (N > 5000), it seems like the t-test and ANOVA will be robust enough to handle this data set, so I should be OK running a t-test with the subway data:

“**So how useful are normality tests?**

Not very. Normality tests are less useful than some people guess. With small samples, the normality tests don't have much power to detect nongaussian distributions. With large samples, it doesn't matter so much if data are nongaussian, since the t tests and ANOVA are fairly robust to violations of this standard.” ([GraphPad](http://www.graphpad.com/support/faqid/959/))

While the data doesn’t look very Gaussian to me based on eyeballing the charts, from the above quote, I concluded that using a t-test on the data would still be a robust test to determine whether we should retain or reject the null hypothesis because the data set is so large. It is also possible that the histogram is only displaying the right half of the distribution, in which case, the curve does look convincingly normally distributed.

And to be explicit, here are my assumptions:

* Running a t-test assumes that the full population follows a normal distribution and that samples taken from the full population will map approximately to a t-distribution.
* I also used the Welch’s t-test instead of the Student’s t-test because there is a big difference in the sample sizes of the two data sets (“with rain” N = 9585, “without rain” N = 33064) so the Welch’s t-test will take this into account better since it’s likely the two samples have unequal variance.
* I’m taking the p-value directly and not dividing by two because I’m running a two-tailed t-test. If I cared only about one direction, I would divide by two for one tail.

However, as mentioned earlier, since I already conducted the Mann-Whitney U test, I will interpret those results. Based on eyeballing the data, another argument could be made that the sample size of hourly entries “without rain” is much bigger than hourly entries “with rain” so a nonparametric test could provide more accurate interpretation of the data.

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.

**T-test**

t-test statistic = **5.0428827476194309**

p-value = **4.6414024316324798e-07**

mean of hourly entries with rain = **2028.19603547**

mean of hourly entries without rain = **1845.53943866**

**Mann-Whitney U test**

Mann-Whitney U test statistic = **1924409167.0**

p-value = **0.024999912793489721**

mean of hourly entries with rain = **1105.4463767458733**

mean of hourly entries without rain = **1090.278780151855**

1.4 What is the significance and interpretation of these results?

The p-value of the two-tailed Welch’s t-test is less than 0.05 (alpha level), which suggests that there is a statistically significant difference between the means of hourly entries “with rain” and “without rain” and therefore we can reject the null hypothesis. In plainer English, it means that it’s likely that there is a difference between the hourly entries when it is raining versus when it is not. So people are may be riding the subway more when it is not raining.

The Mann-Whitney U test is much harder to interpret. The U-statistic makes little sense out of context (see the Piazza discussion [@92](https://piazza.com/class/i2ddoj5wy8i6j5?cid=92)). The p-value suggests we can reject the null hypothesis, and therefore conclude that there is a statistically significant difference between the distribution of hourly entries “with rain” and “without rain”. The descriptive stats on hourly entries in the subway data show:

ENTRIESn\_hourly

count 42649.000000

mean 1886.589955

std 2952.385585

min 0.000000

25% 274.000000

50% 905.000000

75% 2255.000000

max 32814.000000

**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:

Gradient descent (as implemented in exercise 3.5)

OLS using Statsmodels

Or something different?

For exercise 3.5, I stuck with Gradient descent. For the Optional Regression exercise, I used the ordinary least squares model from Statsmodel, and I also tried some polynomials by squaring a few variables in the features set.

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

During my first pass at building the model, I used the following features:

Hour, rain, meantempi, precipi.

But the R^2 for that model was really terrible.

In an attempt at improving the model, I used the following features:

Hour, rain, meantempi, meanwindspdi, precipi. Hour (squared), meantempi (squared), precipi (squared).

I also used **UNIT** as the dummy variable in both linear models.

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 with the four initial features in the first linear model (Hour, rain, meantempi, precipi) but without UNIT as a dummy variable. The R-squared value, approximately 0.03, was very terrible with this model. Subsequently, I tried to improve the model by adding features based on my intuition.

I added meanwindspdi because I thought that people were more likely to ride the subway when it is really windy outside because it is colder or they don’t want to feel discomfort from a heavy wind blowing. Adding meanwindspdi improved the R-squared value a slight amount but not by much.

Then I tried other features such as adding fog and thunder but they didn’t improve my R-squared much. Furthermore, I didn’t want to superficially increase my goodness of fit just because I was adding more features since that takes me closer to the sample dataset but doesn’t tell me if my model is improving.

I also tried taking some features, squaring them, and then adding them to my model. I did this with Hour, meantempi, and precipi because I had gotten a good R-squared value from them in my first linear model. Adding these squared features increased my R-squared value by a bit but it was still under 10%.

Eventually I realized that when I add UNIT as a dummy variable, the R-squared value increases drastically – from 3% to 48%. Therefore, I added UNIT as a dummy variable to my model and ended up with the feature set described in 2.2.

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

According to the Udacity online grader, the R-squared value for the first linear regression model (exercise 3.5) is:

0.461129068126 (or 46.1%).

For the second linear model, optional exercise 3.8:

0.48561137181 (or 48.6%).

This R-squared value varies based on the dataset that is randomly selected to perform the testing.

2.5 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?

The R-squared value of 0.486 or 48.6% means that approximately 48.6% of the variance in the original subway data set can be accounted for by my linear regression model. In other words, I can explain about 48.6% of the variance in the labels is explained by the linear regression of the features I chose.

It is difficult to determine what a good value for R-squared is. [This article](http://people.duke.edu/~rnau/rsquared.htm) from a Professor at Duke University goes into depth about how to determine whether an R^2 value is appropriate for the model: his conclusion is that it depends. He writes that there are many factors to consider beyond just the R-squared such as an adjusted R-squared value and standard error of the regression. He also wrote about the importance of considering whether there are “intuitively obvious relationships” and what the stakes are; for example, determining how good a linear model is for predicting the effectiveness of a new drug has significantly different stakes than using a linear model to predict subway ridership.

To answer the question, I think this linear model is appropriate to predict ridership but it could be improved. The importance of the R-squared value depends on how much variance my features need to account for. If I expect a lot of randomness in my data, then a low R-squared value is acceptable. However, if I expect my features should explain a great deal of the data’s variation and I want a high degree of accuracy in my predictions, then a high R-squared is necessary. 48.6% isn’t a small R-squared value, nor is it a large one. And this second linear regression model’s R-squared value is already an improvement over the first linear model in exercise 3.5 which had an R-squared value of 46.1%. Therefore, I think this linear model is appropriate but I would expect to continue attempting to improve the linear model by testing other features to see if I could create a better linear model with a higher R-squared value.

**Section 3. Visualization**

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

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.

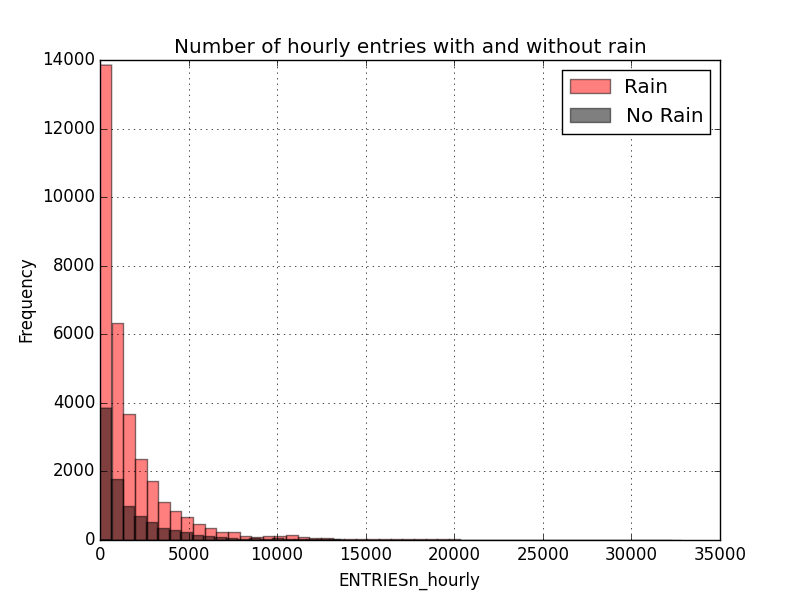
3.1 One visualization should contain two histograms: one of ENTRIESn\_hourly for rainy days and one of ENTRIESn\_hourly for non-rainy days.

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, you might have one interval (along the x-axis) with values from 0 to 1000. The height of the bar for this interval will then represent the number of records (rows in our data) that have ENTRIESn\_hourly that fall into 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.



Description:

The plot above with two histograms shows:

1. The pattern of hourly entries is consistent whether there is rain or no rain because the histograms follow the same shape. This means that there is not a noticeable drop off in the number of subway riders when it is not raining, otherwise we would see the No Rain histogram skew close to zero, and it would not tail off like the Rain histogram
2. The number of observations for hourly entries with rain is much greater (5 times) than the number of observations for hourly entries without rain.

See code for visualization (Subway2Matplotlib.py).

3.2 One visualization can be more freeform. Some suggestions are:

Ridership by time-of-day or day-of-week

Which stations have more exits or entries at different times of day



Description:

The plot above with the bar chart shows:

1. The average hourly entries for subway turnstiles vary based on the day of the week. Wednesday, Thursday, and Friday have the highest averages with Thursday having the greatest average hourly entries.
2. There is a big drop off in the amount of subway ridership during the weekends, with Sunday having the lowest average hourly entries, less than half of the average hourly entries on Thursday. So the day with the least traffic for riding the subway is Sunday (though this plot doesn’t account for different stations).
3. The range of average hourly entries from Monday to Sunday is approximately 1050 (Sunday) to 2300 (Thursday).

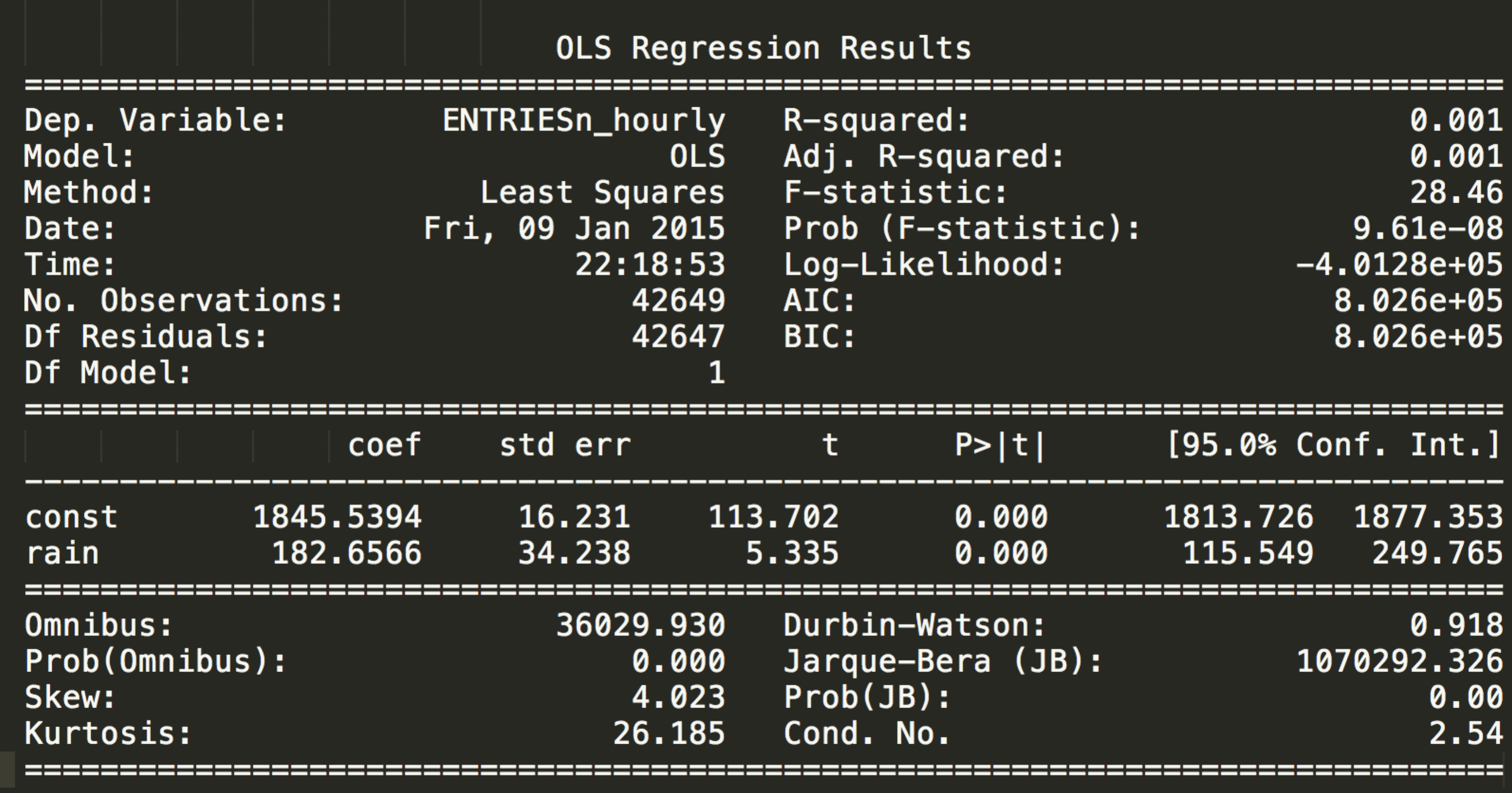
See code for visualization (Subway2MatplotlibFreeform.py).

**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 versus when it is not raining?

The findings from the above analyses suggest there is a slight correlation between subway ridership and rain. I used the statsmodel library to calculate the linear OLS model. I used rain and added a constant as the feature set (the constant was to offset any biases in the data). Here are the results:



Based on the linear model and the R-squared value, which is very small, I concluded there is a relationship between the variables rain and subway ridership but I don’t think it is strong enough to conclude that more people ride the subway when it is raining compared to when it is not.

4.2 What analyses lead you to this conclusion?

To summarize the findings in previous sections, I performed analyses including the following: Welch’s t-test, linear modeling with consideration of R-squared values, visual “eyeballing” test with a histogram plot. The Welch’s t-test suggests that there is a statistically significant difference between the mean of the hourly entries with rain versus without rain. This conclusion isn’t robust enough to predict whether more people ride the subway when it is raining. The linear model that used the Hour, rain, meantempi, meanwindspdi, and precipi features suggests that these factors explained almost 50% of the variance in the labels, which means that a combination of time, rain, mean temperature, wind speed, and how much rain potentially predict whether more people ride the subway. But the linear regression model that uses only rain and a constant as the features suggests there isn’t enough predictive power to conclude that rain causes more people to ride the subway. Additionally, there are some shortcomings in the data discussed below that cause me to hesitate to conclude that rain predicts subway ridership.

**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 data set and the methods   
of your analysis.

To answer the question of whether more people ride the subway when it is raining versus when it is not, there needs to be more data in the dataset. The turnstile data only covers May 2011. I plotted on a line chart the number of hours in a day when rain was recorded at a certain unit (R003). It looks like there were only three rain storms whereas most of the month remained rain free.



(See ConclusionLineDateOnlyOneUnit.py)

Therefore, it is inaccurate to conclude whether more people ride the subway when it is raining considering there is not enough data to control for seasonal weather.

The shortcomings of the analyses revolve around the models and assumptions. The linear regression model assumes that each feature is acting independently. My models don’t take into account how the features are correlated with each other and I didn’t perform any analysis to determine covariance or to control for it. It’s intuitive to think that when it is raining, the mean temperature might be lower. But which factor is really the one causing more people to ride the subway (if either of them have an effect)?

Another shortcoming of the way I performed linear regression is the more features I add the closer my model will fit with the dataset. If I used all the features in the data to create the model and calculate the R-squared value, the R-squared value would be very strong. However, this isn’t a good way to perform predictions: adding more features may result in a higher R-squared value, but the model would be a good fit only for my data set because I have created an spurious fit. This artificially tricks me into thinking I have a good model but my data may not be representative of the true population. A better way is to have two separate datasets: a training and a test set to determine the robustness and predictive power in my model.

Additionally because there are only three rain events, there is a random chance that one or more of those rain storms coincided with higher subway ridership. For example, it could have rained during commuting hours, and thus produce a false positive correlation between rain and higher subway ridership.

One way I could improve my linear model is to make Hour a dummy variable. Intuitively, the time of day should have an effect on subway ridership – with more people riding during the day/waking hours. However, using the value of Hour as a feature doesn’t really help my model, i.e. because Hour doubles (from 4 pm to 8 pm) does not mean that the hourly entries will double. But adding it as a feature suggests this kind of linear relationship. Therefore adding Hour as a dummy variable would likely strengthen my model and increase my R-squared value.

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

The ENTRIESn\_hourly column was not calculating the entries per hour but the difference between the number of entries at the current point in time and last time the entries was measured. Usually the common interval was 4 hours between measurements. Therefore, it would be more accurate to divide the ENTRIESn\_hourly column by the number of hours that has elapsed so that conclusions drawn from this number more accurately reflect hourly entries .