Section 0. References

- http://people.duke.edu/~rnau/rsquared.htm
- http://connor-johnson.com/2014/02/18/linear-regression-with-python/
- http://statsmodels.sourceforge.net/devel/
- https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
- http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-regression-analysis-results-p-values-and-coefficients
- https://en.wikipedia.org/wiki/Coefficient_of_determination
- https://en.wikipedia.org/wiki/Goodness_of_fit
- http://graphpad.com/guides/prism/6/statistics/index.htm?one-tail_vs__two-tail_p_values.htm
- http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
- http://www.statsoft.com/Textbook/Multiple-Regression#cresidual
- http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.probplot.html
- http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/
- https://en.wikipedia.org/wiki/Multicollinearity

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?

> Test: Mann-Whitney U-Test, p= 0.05, two-tailed.
> I rejected the null hypothesis, which asserted that the populations of rainy days and non-rainy days were statistically the same.

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 allows us to test a null hypothesis stating the two populations being tested are the same. This test is applicable to the dataset, because the data we are testing is ordinal and we are trying to separate an entire population into two separate samples and determine if they're meaningfully/statistically different. It's preferable over a t-test, because it can be applied to non-normal distributions. We're using a two-tailed test instead of the scipy default of a one-tailed test because we don't want to assume that we know which population has the larger mean.

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	0.0499
Entries mean (rainy)	1105.45
Entries mean (not rainy)	1090.28

1.4 What is the significance and interpretation of these results?

These results tell me that these samples are definitely different from each other, and that the variable I used to split them (rain vs. no rain) has a meaningful impact on average ridership for the day. The p-value I got of 0.0499 definitely falls within the range of a p-critical value of 0.05, making the mean difference between them statistically significant.

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?

I used OLS using Statsmodels. I wasn't able to get my code for gradient descent working in Problem Set 3: Problem 8.

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 I selected were 'maxtempi', 'rain', 'Hour', and I added a column called 'day_of_week', which I computed in this way:

```
day_of_week = pd.to_datetime(dataframe['DATEn']).dt.dayofweek
```

I used dummy variables for UNIT as a part of my features.

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 selected these features by testing each one individually and seeing how they impacted my R2 value. I chose the values that seemed to have the highest impact and combined them together. I decided to also select a non-existing column for the day of the week as well, because I guessed that ridership would differ between the weekdays and the weekends.

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

maxtempi	-7.1832
rain	-59.3196
Hour	65.043
day_of_week	-86.0299

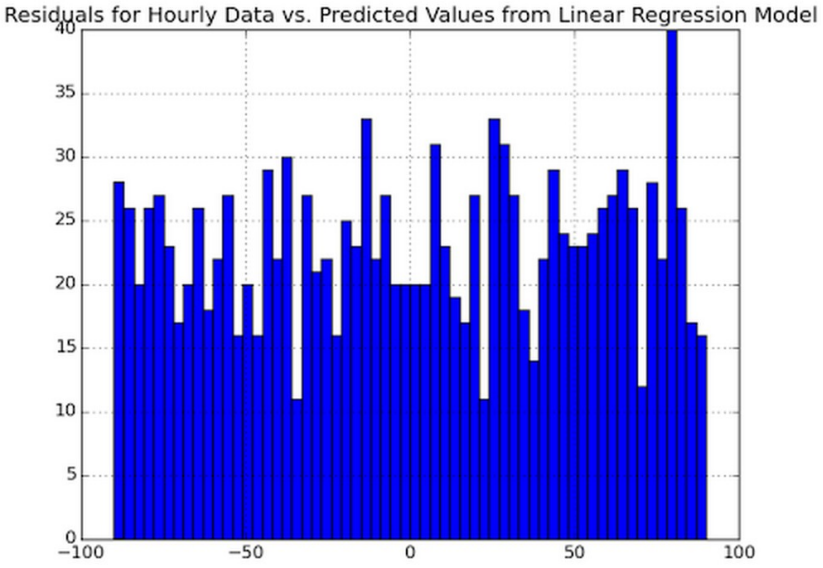
2.5 What is your model's R2 (coefficients of determination) value?

My R2 value is 0.4842.

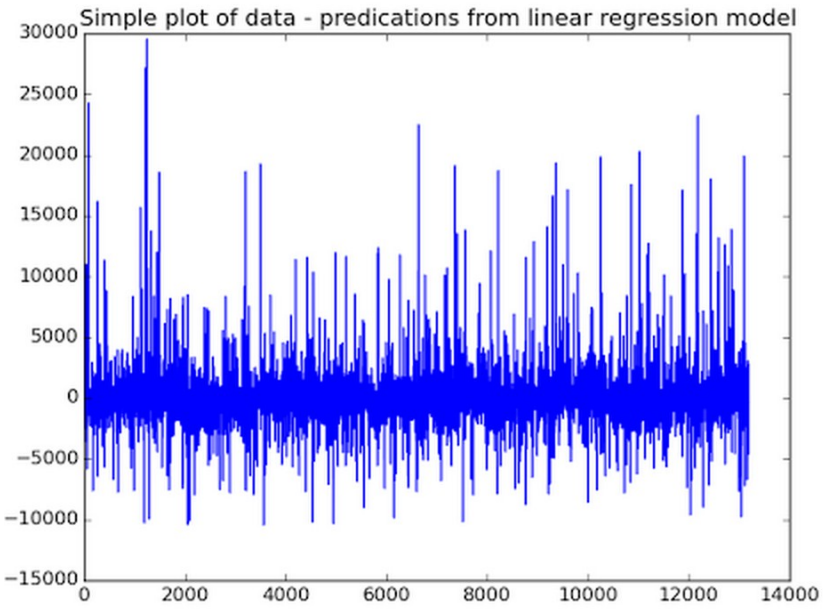
2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

R2 is a number between 0 and 1 that's supposed to magically tell me if the model I've generated fits the data well enough or not. Because it is 1 minus the ratio of residual variability, 0 means "this is awful, don't use it" and 1 means "maybe a little *too* good—make sure you're not making an errors because this might not make sense". An R2 value of 0.4842 means we have explained 48% of the original variability and are left with 52% of the variability of ridership being unaccounted for. The data itself is a little all over the place (see figure 3.2 especially), so finding an exact match closer to 1 I think would be a little unrealistic. Also people aren't machines and will ride or not ride for reasons beyond the scope of our weather data, so it's hard to account for all of their individual variability with information about just the weather.

I am not sure that a linear model to predict ridership is appropriate for this dataset because of the large degree of variability and how wide the range of variability is.



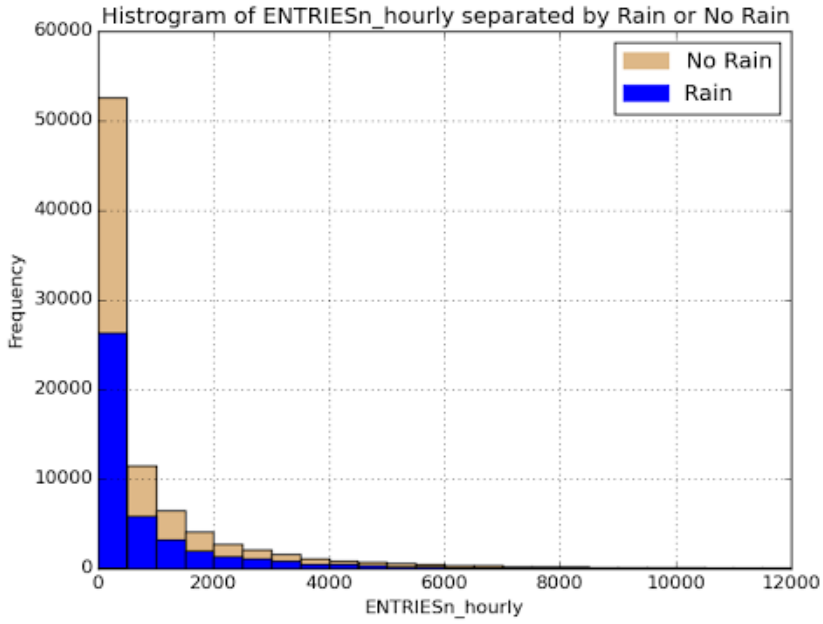
This histogram of the difference between the actual and predicated value is all over the place and very inconsistent. To help better understand this, I plotted just the differences of the data and residuals here:



The data is still not very consistent. It would seem that our data wouldn't really benefit as well from a linear model as it would perhaps from a non-linear model.

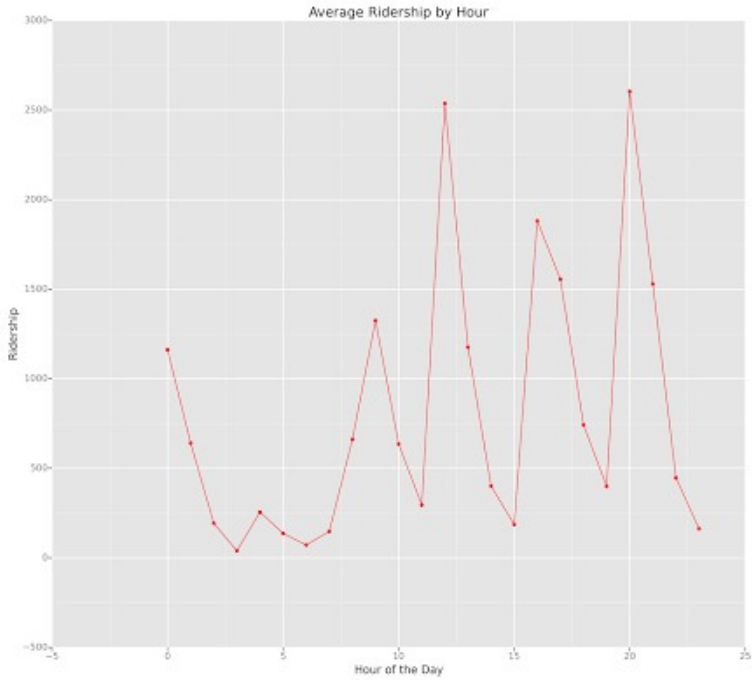
Section 3. Visualization

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



This visualization helped really drive home the difference between rainy days and non-rainy days: in every single instance, no rain has a higher ridership than rain.

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.



This visualization helped to show me that ridership by hour wasn't as smooth of a curve was I anticipated it would be. It's actually really jumpy and all over the place and kind of inconsistent. My guess was that ridership would be highest in the morning and in the evening (like traffic on the roads), but that's definitely not the case here.

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?

(See 4.2.)

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.

More people ride the NYC subway when it's raining. When I separated the entire ridership population into two separate samples (one sample for ridership on rainy days, and one sample for ridership on non-rainy days) and performed a two-tailed Mann-Whitney U-Test to compare them, I was able to determine that the two samples were indeed statistically different, with a p-critical value of 0.05. My p value came out 0.0499.

After that, I created a linear regression test to determine if I could accurately create a model to represent the data using a single line/equation. I was able to create one with an R2 value of 0.4842. Because this value isn't close to 0, it indicates that there is likely a correlation between these values and making a hypothesis about ridership for the day.

Because the entries mean for rainy days is higher than the entries mean for non-rainy days and I'm able to create a model using linear regression with a decent R2 value, I feel I can safely predict based on the data and my statistical analysis that NYC subway ridership is higher on rainy days.

Section 5. Reflection

5.1 There are some shortcoming in the analysis made here. First, the linear regression model doesn't seem to be the best model to fit this data. If you look at the graphs in 2.6, you can see that the residuals are very inconsistent which implies that the linear regression model isn't the best option for this kind of data. A non-linear model might be better suited to this application.
Also, the dataset that was given for this project only covers for one month of the year—May 1-May 31. Having a data set that covers at least a few years would be more accurate because it would help smooth over localized and seasonally-related things (huge snow storm predicted but didn't hit, big sports games, terror threats, sickness outbreaks, etc...).
The data points given in the dataset are also really similar (all the temperature options, precipitation/rain, windspeed/mean windspeed, etc...), which puts us at danger of multicollinearity. Even though we have a lot of options to draw from, we need to be careful to only choose the one of each potential set that might make us susceptible to multicollinearity.