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, 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 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.

I have mainly used:

1. Pandas documentation: http://pandas.pydata.org/pandas-docs/version/0.15.1/
2. ggplot docs: https://github.com/yhat/ggplot
3. Info on the Mann-Whitney U test: http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
4. http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-tutorial-and-examples; blogs on regression analysis

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 Mann-Whitney U-test. I set my p-critical value to be 0.05, which is a two-tailed p-value. The null hypothesis is that rain-days will have the same ridership numbers as non-rain-days.

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.

This test can be applied as it is nonparametric, ie. it does not assume an underlying probability distribution. Indeed from the data there is no apparent normal distribution.

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.

I got a higher average for rain days vs non-rainy hours, namely 1105 vs 1090. In terms of p-value, I got 0.04998 (two tailed).

1.4 What is the significance and interpretation of these results?
In view of p_crit = 0.05 this suggests we can reject the null and there would in fact be a difference in ridership between rainy and non-rainy hours.

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:

1. *Gradient descent (as implemented in exercise 3.5)*
2. *OLS using Statsmodels*
3. *Or something different?*

I used OLS, submitted in 3-8.

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 the following features: ‘rain’, ‘Hour’, ‘meanwindspdi’, ‘meantempi’. Also I used dummy features for the units.

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.”*

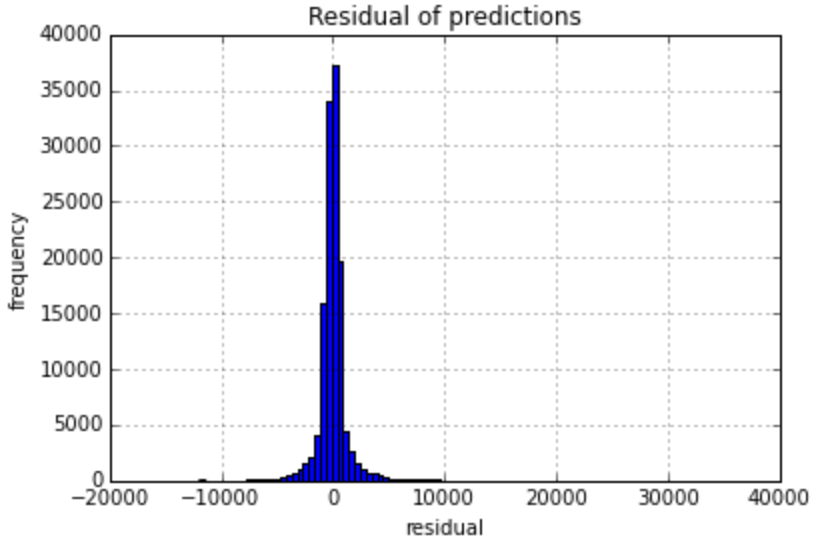
The reasons I choose these are largely intuitive: rain (for obvious reasons), and the meanwindspdi and meantempi for intuitive reasons: with more wind and lower temperatures, one would expect more ridership. The units really improved the R^2 and were not my first idea.

2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?
rain: 39.1
Hour: 62.3
meanwindspdi: 28.1
meantempi: -8.2

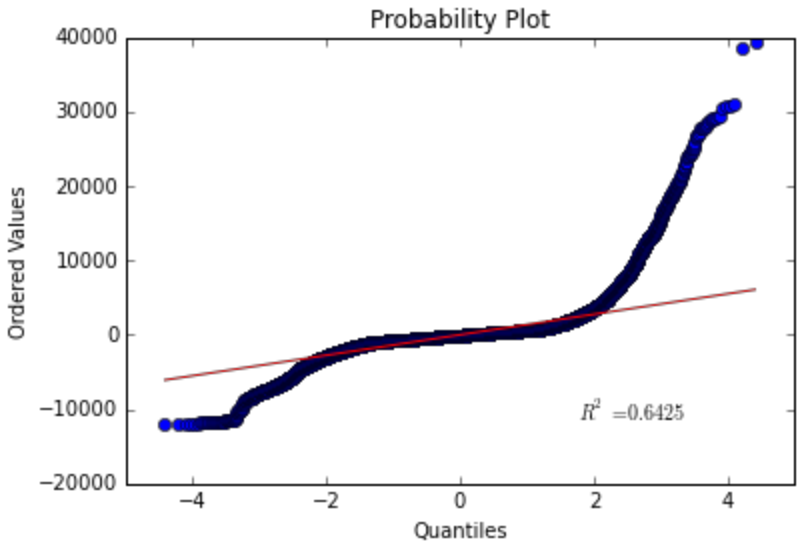
2.5 What is your model's R2 (coefficients of determination) value?
0.484

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 indicates that 48.2 % of the variance is explained by the model, which is a modest number. Therefore the question arises whether the linear model is appropriate. From the residual distribution we find that a long tail, with residual values up until around 40000 (see figure).



For an even more interesting picture, below is the QQ-plot (http://en.wikipedia.org/wiki/Q%E2%80%93Q_plot). This plot shows the quantile distribution of the residuals against a normal distribution. The S-shape of the curve shows how the residual distribution has notably fatter tails than the normal distribution; a linear regression therefore seems of limited value.



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.*

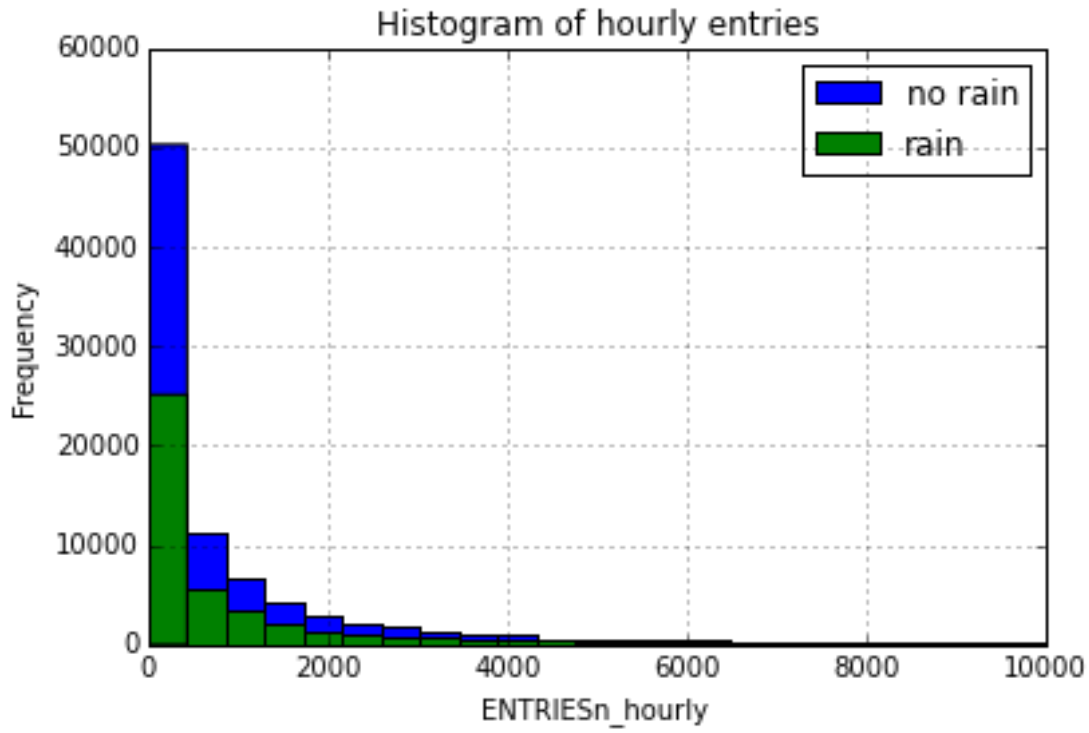


Figure 1: Frequency of hourly entries for rain and non-rain hours. The number of no-rain entries is about twice as large than the number of rain entries, which is reflected in the figure.

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*

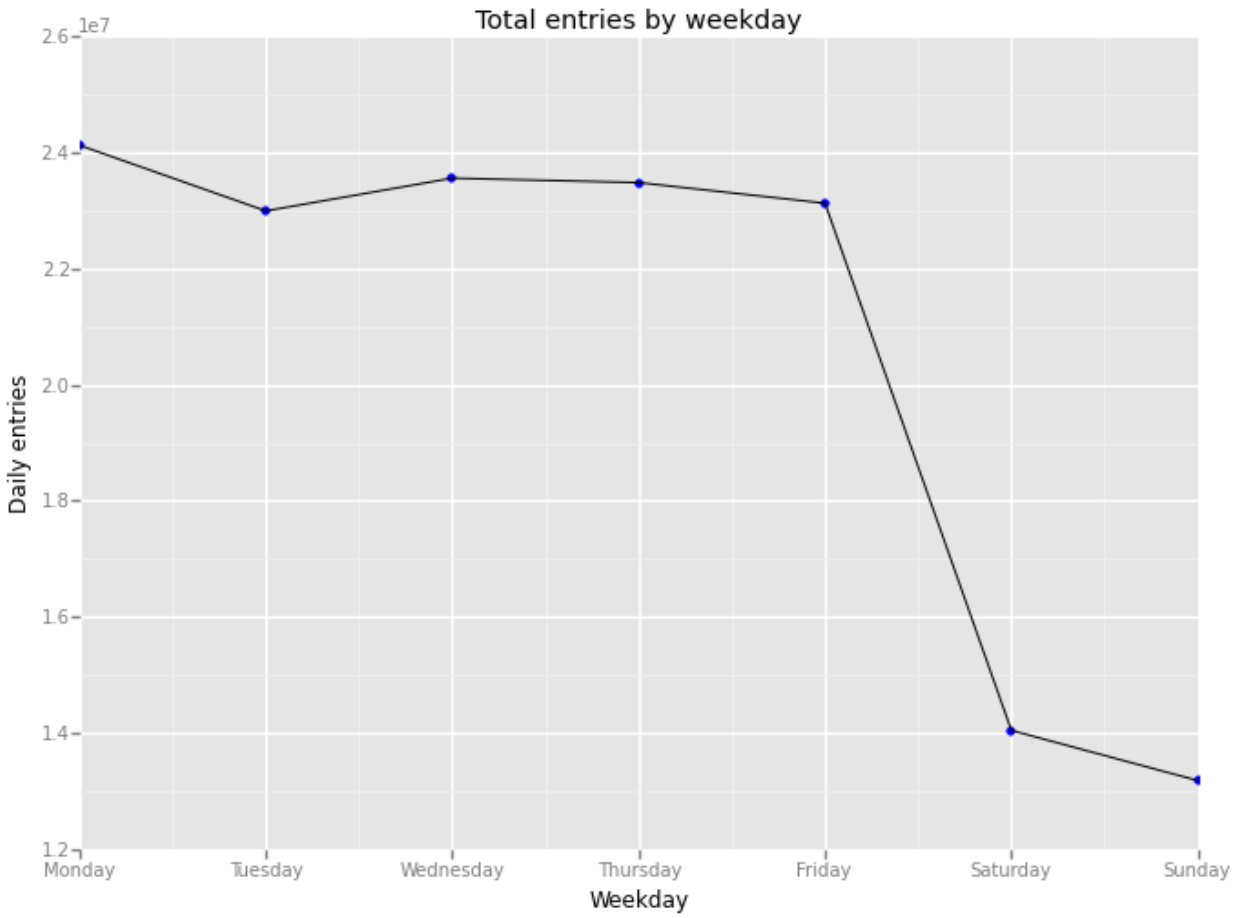


Figure 2: Total ridership per day of the week. Clearly on weekends the ridership drop; monday peaks, Tuesday is lowest of the business days.

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.

My overall conclusion is that there seems to be a tendency to have more ridership during rain, but the evidence is not strong enough to make a definitive statement. From the statistical test, we can reject the null hypothesis that there would be no difference between rain- and no-rain-ridership. From the regression analysis (OLS), the rain feature has a positive constant and therefore indicates at more ridership with rain. However, the confidence interval includes

0 and the p-value is 0.286 which suggests we cannot conclude rain has a positive correlation with ridership. Also in the regression, some features do much better. For one, the constant; but also the Hour-feature and most of the Units have much better confidence intervals and p-values; this suggests that if rain plays a role at all, it has a marginal contribution on top of the other features. Finally I found that varying the features (e.g. including 'fog' or not, replacing 'rain' with 'precipi' and more) leads to big changes in results for the rain- or other weather related coefficients (compared to the others).

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.*

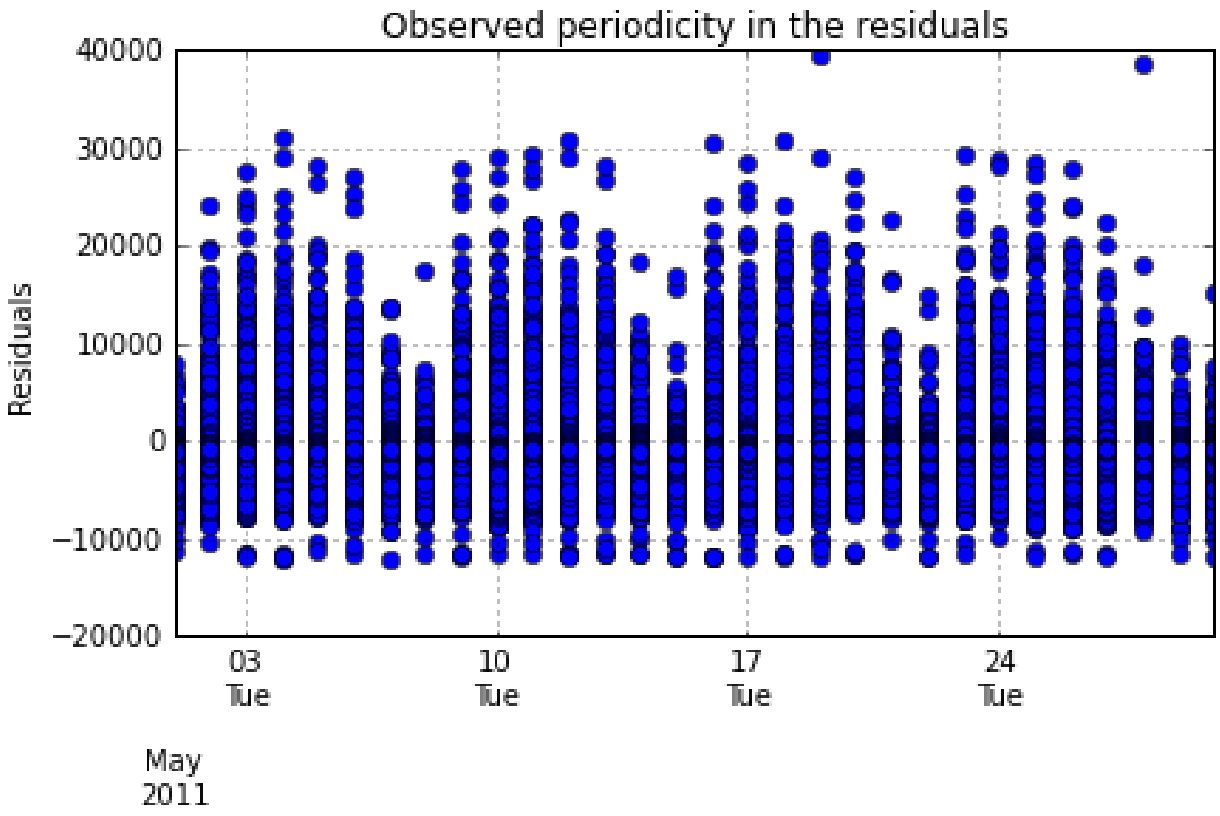
5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?

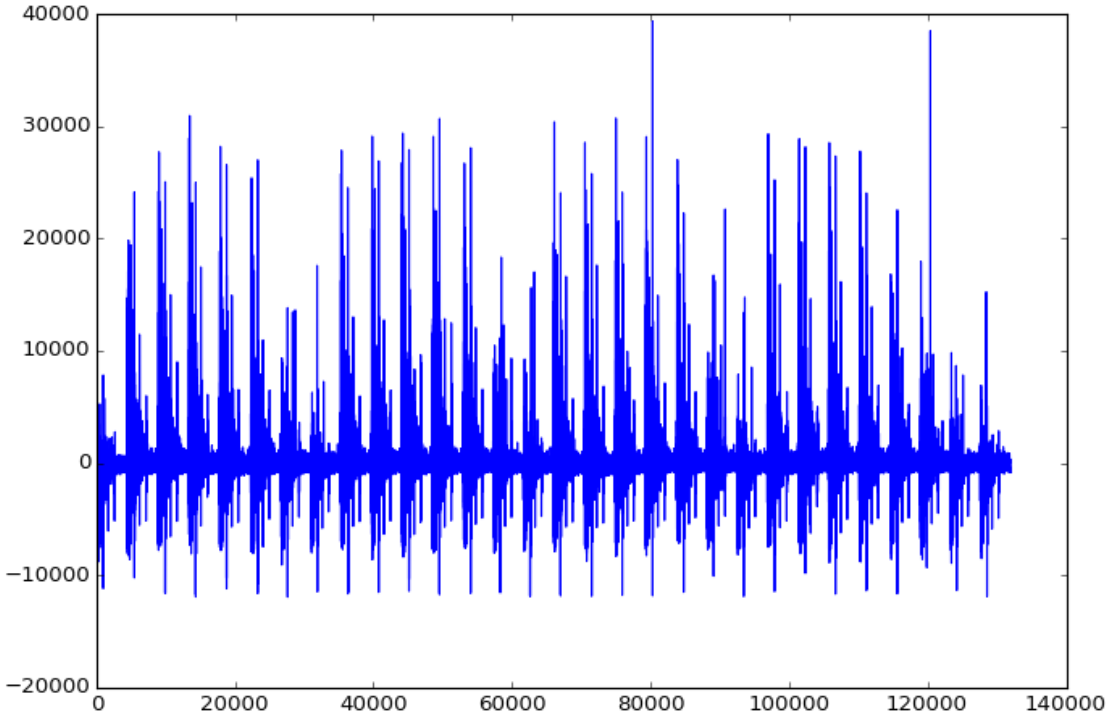
Data:

There are a couple of things that I did not dive into, which might lead to more insight:

a. The database history spans just a month which is quite limited. Especially the winter months might reveal quite different patterns, when rain in combination with colder temperatures might lead to different ridership behavior.
b. Is the data complete for all units (turnstiles);
c. There should obviously be some correlations between the different weather variables, especially 'rain' and 'precipi'; probably also between the 'pressure'- and 'temp'-type variables. Looking at this might lead to more insights into the reliability of the data.

Another point can be highlighted by plotting the residuals as a function of time (Fig 3). There clearly is some periodicity in the residuals over weekdays, where the residuals are smaller in weekends. However zooming in further in Fig 4: here the x-axis is less descriptive, as it just counts the number of hourly entries over a month, however therefore has hourly resolution. In this case we can see that there is a big difference in residuals over the course of each day. This kind of periodicity suggests we should look into nonlinear modeling as well, or break up the entire set into periods and do piecewise linear regression.


Observed periodicity in the residuals
Residuals
40000
30000
20000
10000
0
−10000
−20000
03
Tue
10
Tue
17
Tue
24
Tue
May
2011



Analysis:
I refer back to section 2.6 above which shows that linear regression has a limited validity, as suggested by the fat tails, and to the above discussion on periodicity of residuals.
I think it would be interesting to do regression analysis for individual units to see how the analysis would come out. After all it may be that, for example, high-traffic units have a stronger relationship which partly gets averaged out when summing over all units. Also it would be interesting to perform regression only for certain periods of a weekday, like the peak hours from Monday-Friday. Finally we could play with other types of regression, as there are many more such as isotonic regression which are available in the scikits.learn module.