Analyzing the NYC Subway Dataset

Overview

This project consists of two parts. In Part 1 of the project, you should have completed the questions in Problem Sets 2, 3, and 4 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.

- reference for the used Mann-Whitney-U function: http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
- reference for Statsmodel OLS function: http://statsmodels.sourceforge.net/devel/generated/statsmodels.regression.linear_model.OLS.html
- reference for SGDRegressor from scilearn:http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html

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 wanted to test, if the use of subways in NYC is different when it is raining or when it is not raining. So my null hypothesis is that using the subway is the same for days it is raining or not:

μ_R = mean of people in population entering the subway per hour when it is raining
μ_N = mean of people in population entering the subway per hour when it is not raining

H_0: $\mu_R = \mu_N$

The alternative hypothesis is that there is a difference, that the means of the population for the two cases are not equal:

H_A: $\mu_R \neq \mu_N$

To test the null hypothesis the column ENTRIESn_hourly in the sample dataset is used. The dataset contains a "rain" feature and I split the dataset into two subsets, one containing the rainy days and one containing the dry days.

The test is a two tailed test because I am just interested if there is a difference in the subway use.

To verify that the mean for the sample datasets are not different by chance I use a 5% level of significance (2.5% on each tail). So the p-critical value is .025.

I used the Mann Whitney U Test to verify if the difference between subway use is significant.

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 size of the two sample datasets are different and I am not sure about their distribution. Therefore I use a test that is non-parametric and does not assume a normal distribution. Mann Whitney U is such a 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.

Executing the Mann-Whitney-U test on the two samples gives the following values:

U:	153635120.5
P:	.00000274
mean for rainy days:	2028.196
mean for dry days:	1845.539

1.4 What is the significance and interpretation of these results?

The algorithm used gives a one tailed p-value. For a two-tailed test I have to multiply it by 2. The p-critical value is .025 which is much greater than the 2*P returned. Therefore I have to reject the null hypothesis. The ridership of the population on rainy days is significantly different than on dry days. The means show that the ridership on rainy days is higher.

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:
OLS using Statsmodels or Scikit Learn
Gradient descent using Scikit Learn
Or something different?

I used the Statsmodel OLS least square algorithm and also scilearn SGDRegressor for gradient descent.

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

I tried several variants and achieved the best results with the following feature set:

['rain', 'meanprecipi', 'hour', 'meantempi','weekday'] + UNIT

I included dummy variables for the UNIT. Without them the prediction was very bad (R2 ~ .1).

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 included features that

a) are reflecting the weather conditions (rain, meanprecipii, meantempi)

and b) are reflecting the use of a subway station (UNIT, hour, weekday)

Including the UNIT (with dummy variables) dramatically improved the R2 value. The other features have just marginal effects.

I omitted to use other passenger related values as features (like ENTRIESn or EXITSn) as these values would not be available in a situation where prediction would be used.

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

For the OLS algorithm I have the following coefficients for the non-dummy features:

intercept:	778.224
rain	-39.604
meanprecipi	81.300
hour	123.404
meantempi	-13.843
weekday	982.720

For the SGDRegressor I have the following coefficients for the non-dummy features:

intercept:	920.591

rain	-35.692
meanprecipi	1210.517
hour	126.076
meantempi	-15.705
weekday	1019.225

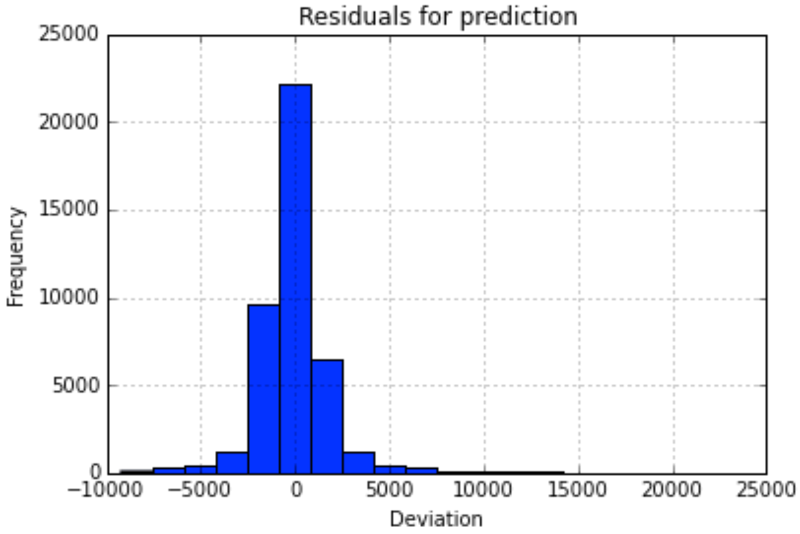
2.5 What is your model's R2 (coefficients of determination) value?

For the OLS algorithm R2 is .482. For SGDRegressor I got .463 using 10 iterations

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?

Both algorithms give R2 below .5. R2 is an indicator of how well the model fits the data and therefore the resulting model is explaining the data not very well. It explains less than 50% of the variation in the data. I think that the model is not describing the reality good enough to use it predicting the subway ridership.

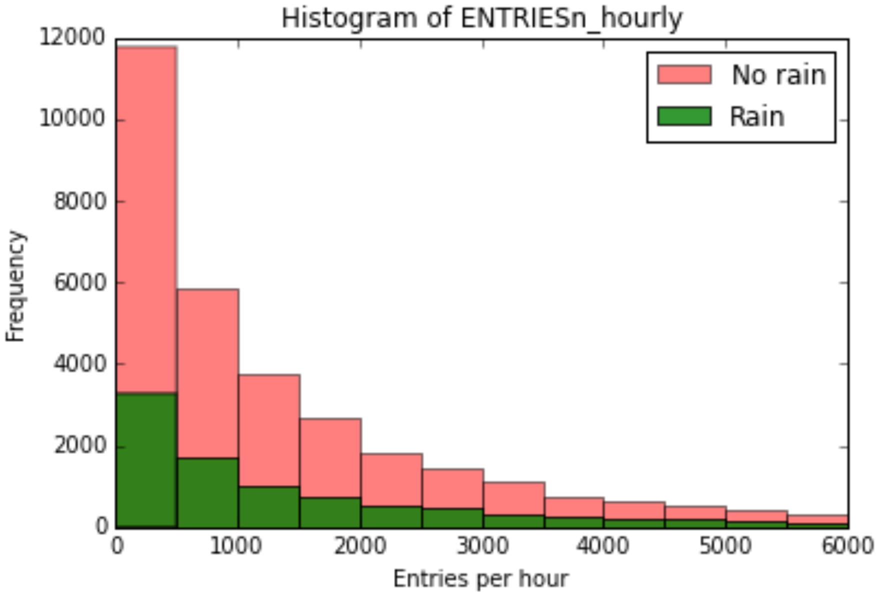
The histogram below show the large deviation between predicted and actual values. Below 50% of differences are around 0 and there are larger tails also.



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.

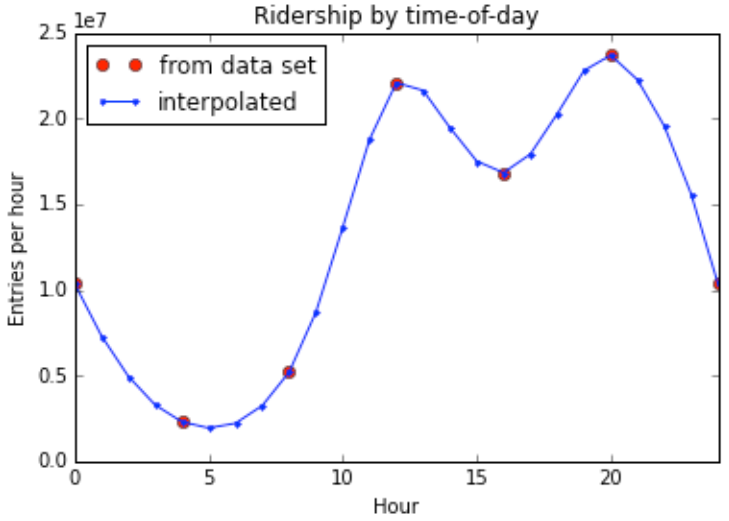
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.
You can combine the two histograms in a single plot or you can use two separate plots.
If you decide to use to two separate plots for the two histograms, please ensure 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.



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

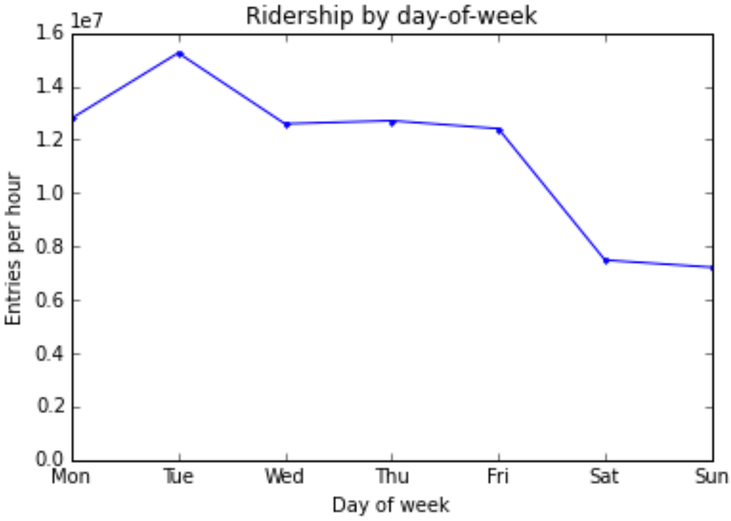
3.2.1 Ridership by time of day

The following pictures shows how many people entered subway stations per hour. The data set has only aggregated values for 4 hours and these values are shown in the graph as red points. To give an impression of how the use of the subway might change over a day an interpolated line through the points is also shown (using Steinman interpolation).



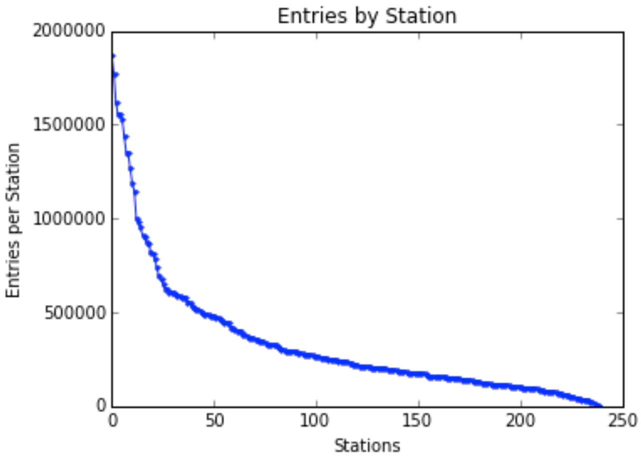
3.2.2 Ridership by day of week

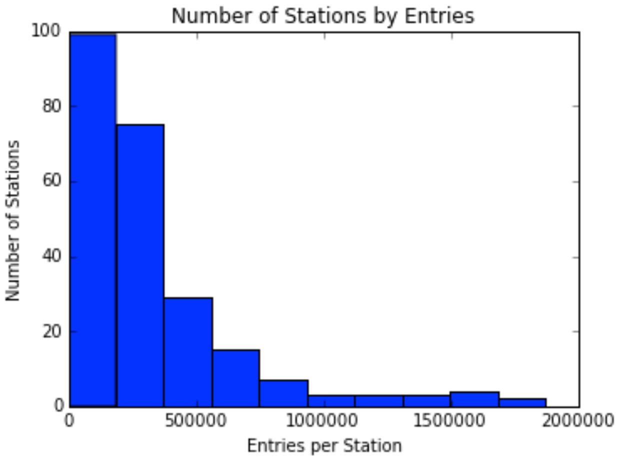
The following picture show how many people entered subway stations per day.



3.2.3 Entries per Subway station

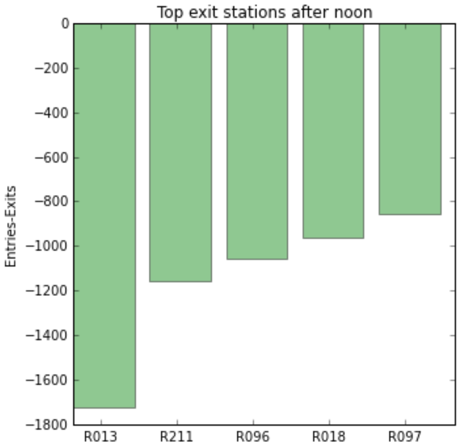
To give an impression how different the subway stations are used the following pictures show the number of entries per station. In the left graph the subway stations are ordered by the number of entries. In the right graphic the bar show, how many subway stations exists for a range of entries.

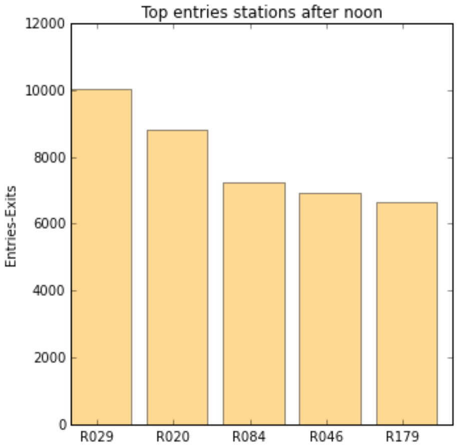


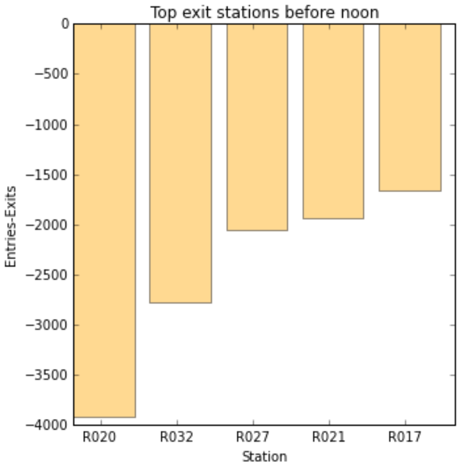


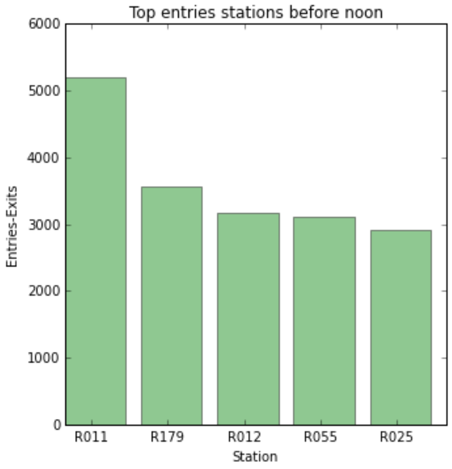
3.2.4 Entries and Exits compared

The next graphics show the top 5 subway stations for entries and exists before noon and after noon. This allows to compare which subway stations are used more often in specific time ranges. Subway station R020 is a good example that is used often for exits in the morning and for entries after noon. It is the subway station 47th-50th St. - Rockefeller Center.









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?

Analysis of the NYC subway data shows that there is a significant difference of ridership between days it is raining and days it is not raining. When it is raining, on average 9% more people use the subway.

Although the influence of the rain is significant the prediction model for the ridership shows only a negligible dependence on the fact it is raining or not. But overall, the prediction model was not well suited to predict the ridership with high quality.

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.

To test for the difference in ridership on rainy days compared to days it is not raining I used a Mann-Whitney-U test on two appropriate samples of the data. The resulting p-value of the tests showed that there is a significant difference in the two samples and the mean of the ridership at rainy days was higher.

Creating a linear model to predict the ridership was not very successful. Although several combinations of features from the dataset were tested I was not able to find a model with an R2 of more than 50%. Trying various feature sets showed that the dependency on the UNIT (the subway station) was really high and the dependency on the fact is is raining was rather low. Without UNIT the R2 was juest around 10%.

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:
Dataset, Analysis, such as the linear regression model or statistical test.

Testing the difference of ridership on rainy days compared to days it is not raining was successful using the Mann-Whitney-U test.

But I was not able to find a reasonable feature set to predict the ridership that has a R2 of more than 50%. Reason for that might be that the linear model I used is not suitable to predict the ridership what could also be seen by the residuals histogram in 2.6.

The dataset itself is just for a single month (May 2011). This and the fact that the hourly data is averaged over 4 hours limits the conclusions that could be drawn from this data. Information about subway ridership could also be improved by integrating additional facts like public holidays or school holidays.

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