Analyzing the NYC Subway Dataset – Bram Stynen – NAND April 15

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

SQL reference:

- http://www.1keydata.com/sql/sql-syntax.html

Converting date and time:

- https://docs.python.org/2/library/datetime.html
- http://stackoverflow.com/questions/12070193/why-is-datetime-strptime-not-working-in-this-simple-example

fillna():

- http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.fillna.html

Histograms:

- http://pandas.pydata.org/pandas-docs/stable/visualization.html#histograms

Nonparametric test:

- http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
- http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html

OLS:

- http://statsmodels.sourceforge.net/devel/

Gradien descent:

- http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html
- http://scikit-learn.org/stable/modules/sgd.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?

Mann-Whitney-U test
Two-tail p value (we do not assume directionality)
Null hypothesis H_0 = [Based upon entry frequency within the time period given, NYC subway ridership levels do not differ significantly between rainy and non-rainy days with threshold p-value 0.05.]

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 is required because the data has a long right-side tail. Therefore, the distribution is not normally distributed. Although not tested, the null hypothesis in the Shapiro-Wilk test would most likely be rejected. The nonparametric Mann-Whitney-U test does not require normal distributions so we opt for this test. At least 20 observations are required for each samples but we have many more than 20.

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.

Mean ridership on rainy days: 1105.45
Mean ridership on non-rainy days: 1090.28
U: 1924409167
p-value: 0.04999983 (two times the regular output p value which is for one-sided tests)

1.4 What is the significance and interpretation of these results?

The mean ridership on rainy days vs. non-rainy days is significantly different at threshold p-value of 0.05. There are significantly more NYC subway riders on rainy days vs non-rainy days. It is clear we have to conclude that people are more likely to take the subway on rainy days than non-rainy days. However, the small difference in mean ridership suggests that rain may not be a major determinant in predicting ridership.

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?

The ordinary least squares method and linear regression with gradient descent. I will discuss the OLS analysis. I'm including the code here because the website wouldn't save my submissions on 3.5 and 3.8.

Code for OLS:

```
   Y = values
   X = features
   X = sm.add_constant(X)
   model = sm.OLS(Y,X)
   results = model.fit()
   params = results.params[1:(len(X)+1)]
   intercept = results.params[0]
```

Code for SGD:

```
   clf = SGDRegressor()
   clf.fit(features, values)
   intercept = clf.intercept_
   params = clf.coef_
```

Code for implementation of days of the week in OLS model:

```
   dataframe['DATEn'] = pandas.to_datetime(dataframe['DATEn'])
  # Create new column to identify the day of the week
   x = []
   for i in range(len(dataframe ['DATEn'])):
      x.append(dataframe ['DATEn'][i].weekday())
```

```
dataframe['Week'] = x

dummy_units = pandas.get_dummies(dataframe['Week'])
features = features.join(dummy_units)
```

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

Units, hour of the day, day of the week and mean temperature. The units and days of the week were integrated as dummy variables.

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?

The selection was mostly based on intuition. One can expect to have highly variable traffic intensity depending on the subway line geographical positions (therefore, units were implemented). Next, traffic intensity can be expected to be much higher during peak hours, i.e. mornings and evenings on weekdays. This explains why days of the week and hours were also taken as features. We can expect weather to be a (minor) factor explaining why we also add mean temperature. The rain feature improves predictive accuracy by a small fraction but it's so small that Occam's razor comes into place.
A quick test reveals that addition of weekdays dummy variables increases R2 from 0.47 to 0.491 and unit identity is the major contributor to prediction accuracy. Intuition may help a bit with getting started on a predictive model, but a more ideal situation would be to systematically include and exclude available features, keep track of the predictive power and then use a deconvolution approach to combine the most interesting features to get the optimized model setup. Before starting this, correlation between predictors should be tested as well to remove non-independent variables.

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

1, and 0 for null values.

2.5 What is your model?s R2 (coefficients of determination) value?
0.491

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 predictive power of this model if fairly weak. If highly accurate predictions are required, inclusion of more features, even potentially from outside of our available database, would be necessary.

Section 3. Visualization

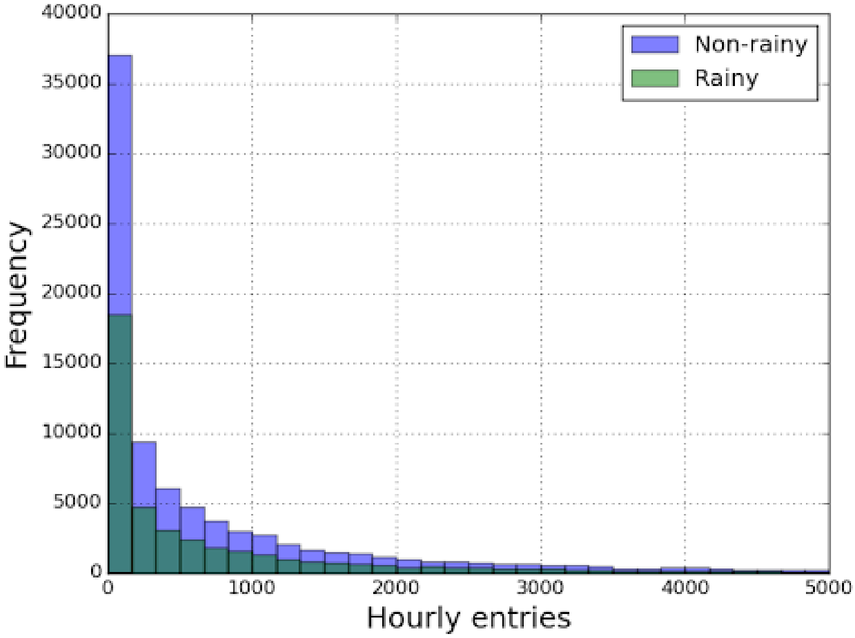

Figure 1. NYC subway ridership. Histogram of hourly entries on rainy and non-rainy days.

Figure 1 shows that the hourly entries in the NYC subway is not normally distributed. It is clear there are peak hours where many customers take the subway while the majority of the time, the number of hourly entries are relatively low (long right tail). We can also observe that there is less data for rainy days (note the bars are not stacked).

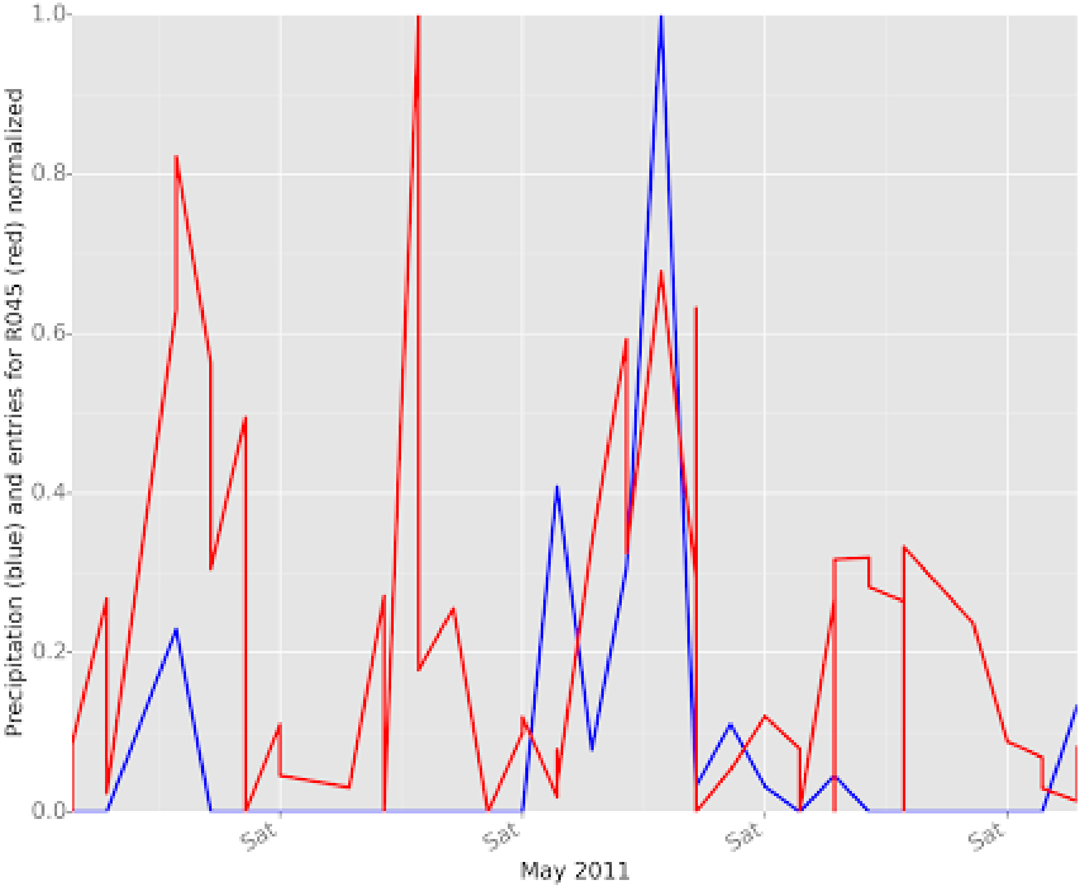

Figure 2. Ridership for unit R045 (red) and precipitation for May 2011

Figure 2 shows how on weekdays, there is more subway traffic (more customers) than during the weekend. The data suggests that this trend is more significant than the correlation between precipitation and ridership. Figure 3 (below) confirms the trend that during the weekends, ridership is decreased:

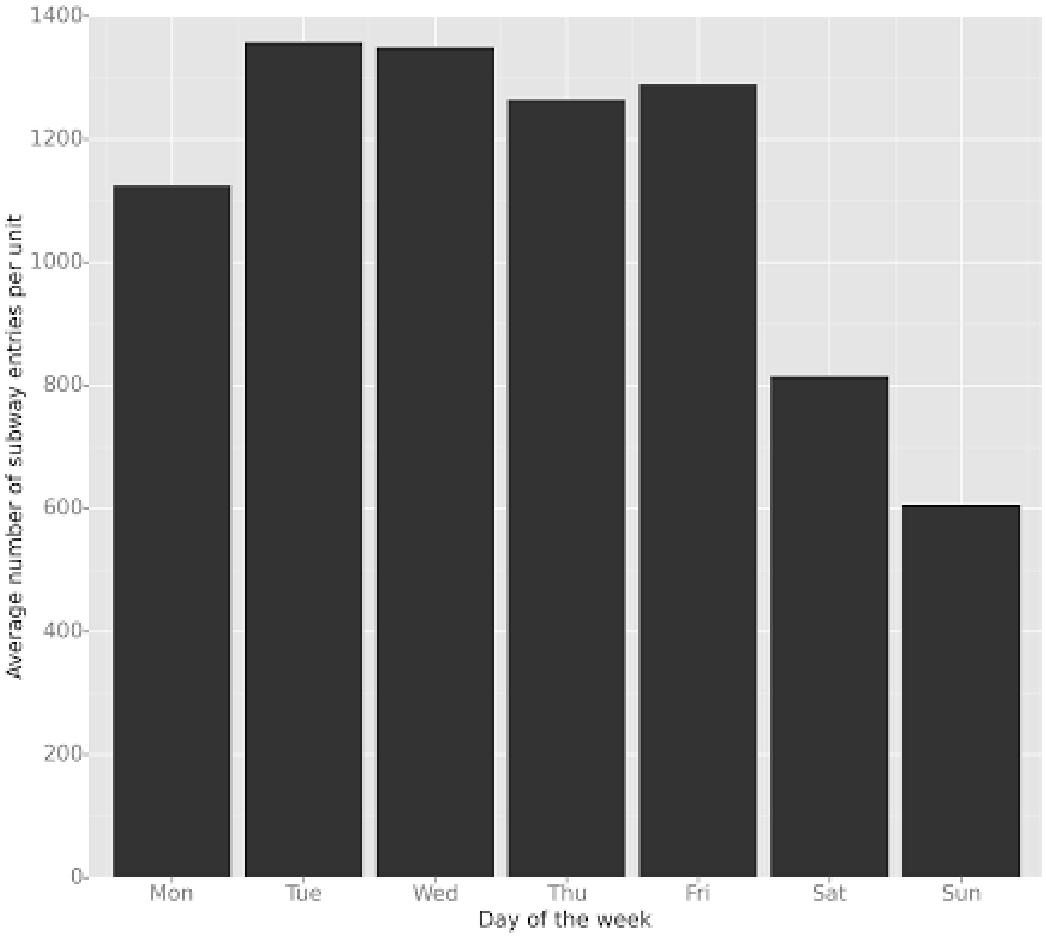

Figure 3. Ridership per day of the week.

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?

More people ride the NYC subway when it is raining. The mean ridership per subway unit is 1105 on rainy days and 1090 on days without rain. The difference is significant (p-value < 0.05) using a Mann-Whitney-U test. Despite being significant, the difference is almost negligible.

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.

The Mann-Whitney-U test proves that there is a statistically significant difference (p-value < 0.05) between ridership on rainy days vs. non-rainy days. However, the predictive power of the 'rain' variable to estimate ridership (as hourly entries) is low. The R2 value only increases by less than 10^{-4} when the rain variable is incorporated in a model that contains the unit, day of the week, hour of the day and mean temperature features. Other factors, such as the unit identity or time are more important predictors of ridership, relative to weather factors.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

The dataset is limited to the month of May. To understand better the effect of weather, it would be better to have data from throughout the year.

Several optimizations could help the analysis. First, an explanatory data analysis on the features might help to eliminate variables whose information is already contained in other variables. Second, a conversion of using the hours of the day + days of the week into hours of the week might improve the predictive power significantly. By taking the hours of the week as a variable, you take into account both the day and time of the day simultaneously, separating the mornings and evenings of the weekdays with the mornings and evenings of the weekend. In this case, you would need to provide a categorical variable to hours of the week, splitting mornings and evenings of weekdays from other times of the week and then use dummy variables or perform k-nearest neighbor analysis. This may greatly improve the outcome. Third, transformations of datasets before linear regression may be necessary. For example, it could benefit the analysis to log-transform the predicted values to get to a dataset closer to a normal distribution. Finally, we are lacking a test set so additional data would help to test our prediction quality or we would have to set aside a part of the training set to use as a test set.

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

/