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

[1] http://pbpython.com/visualization-tools-1.html
[2] http://stackoverflow.com/questions/28178457/plotting-average-values-using-python-and-ggplot
[3] http://ggplot.yhathq.com/docs/index.html
[4] http://en.wikipedia.org/wiki/One-_and_two-tailed_tests
[5] https://statistics.laerd.com/premium-sample/mwut/mann-whitney-test-in-spss-2.php
[6] http://stackoverflow.com/questions/6871201/plot-two-histograms-at-the-same-time-with-matplotlib
[7] http://stackoverflow.com/questions/265960/best-way-to-strip-punctuation-from-a-string-in-python
[8] http://people.duke.edu/~rnau/rsquared.htm
[9] http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
[10] http://blog.minitab.com/blog/adventures-in-statistics/why-you-need-to-check-your-residual-plots-for-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?

I used the Mann-Whitney U-Test, which is non-parametric, in a one-sided implementation. The null hypothesis of the test states that two groups are sampled from the same distributions or medians. The critical p-value is 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.

Before using the Mann-Whitney U-Test, there are some assumptions to consider that can be evaluated through some exploratory data analysis.
First, as can be seen from the figure below, the two samples do not seem normally distributed, then the Welch's t-test to compare means must be excluded.
Second, since the distributions of the independent variables have the same shape (see Figure in Section 3, Question 3.1), then we can interpret the results of the test as a difference of the medians.

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.

The results of the test are summarized in the following table.

Data	Value
Mean of Hourly Entries – Rain	1105.45
Mean of Hourly Entries – No Rain	1090.28
U value	1924409167.0
p-value	0.025

1.4 What is the significance and interpretation of these results?

Given the p-value, we can say that the probability of obtaining a sample data set as extreme as the observed data, given that the null hypothesis is true, is really small. Then we can reject the null hypothesis and state that there is a statistically significant difference between the two means.

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?

Gradient descent. I changed the values of alpha and number of iterations several times and in several combinations. The results shown in the next questions are got using the values 0.5 and 10, respectively. Such final values were chosen only for performance reasons, in fact the coefficient obtained and the convergent to a local minimum were the same as setting the two values to 0.1 and 100, respectively.

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

Here's the final list of features and dummy variables:

- Features: rain, precipi, hour, meantempi, fog, weekday
- Dummy: unit

It was important to keep track of the subway stop as a dummy variable because it has a significant impact on the usage of the service.
Moreover, a new feature was added to the dataset: weekday, which is set to 1 for working days (from Monday to Friday) and 0 for weekends (Saturday and Sunday)

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 final choice of the feature in the model was determined by a mix of intuition and experimentation. Rain, precipitation, hour, and mean temperature were maintained for two reasons: first they sound logical, second, experimenting and removing I did not get better results.
I also added fog because if it is foggy, commuters who usually drive might decide to use the subway for safety reasons.
Finally, through experimentation, I decided to create and add the weekday feature, because during weekdays people should use the subway more often than in the weekends (e.g. due to workers), as can be seen from the figure in Section 3, question 3.2. This variable slightly improved my R^2 value.
I also tried to include the single day of the week as a dummy variable in the model but it didn't improve the R^2 value.

Before getting to this model, I also tried to include only the rain feature, getting a lower R^2 value, but with a positive coefficient value.

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

The following table reports the list of coefficients for each feature in the final model. It can be observed that rain has a negative coefficient on the model, whereas the two greatest positive impacts are given by the day of the week and the time of day.

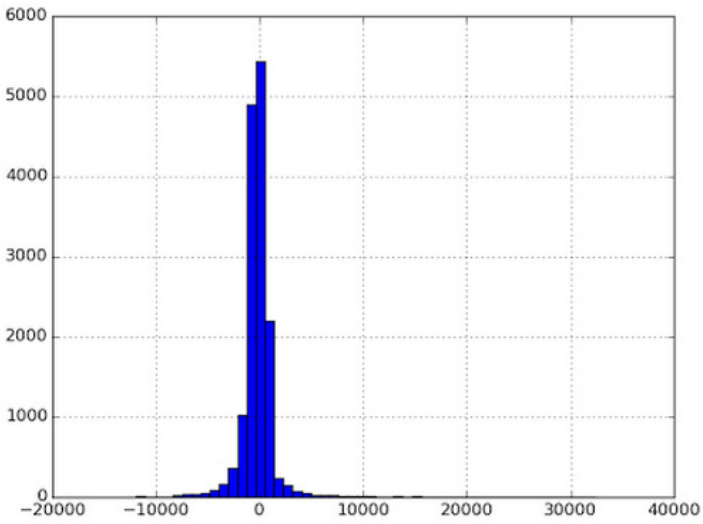
Feature	Coefficient
rain	-4.43070591e+01
precipi	-8.35132017e+00
hour	4.67203731e+02
meantempi	-5.79944819e+01
fog	4.90215003e+01
weekday	2.51257439e+02

2.5 What is your model's R^2 (coefficients of determination) value?

0.475231975828

2.6 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?

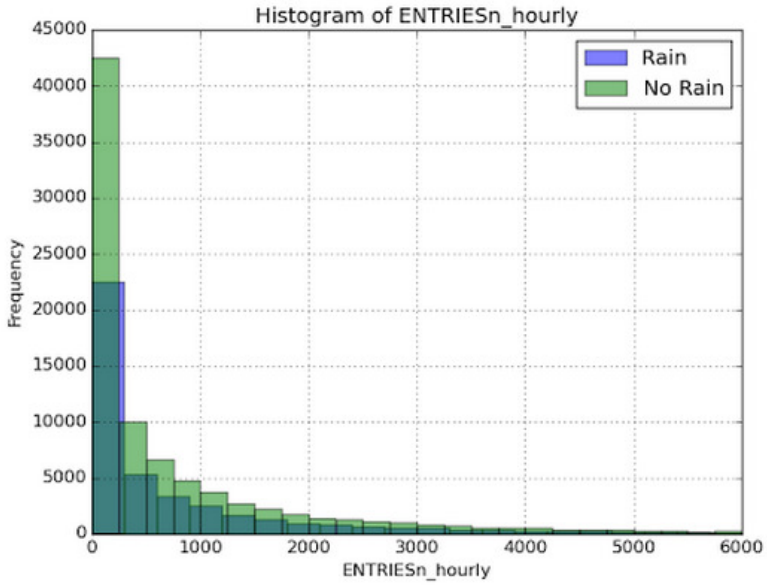
R^2 value is the portion of total variance explained by the model and practically means how well the data fit the model. Looking at the residual plot (see the figure below) it seems that it satisfies the assumptions of normality and independent distribution with a mean of 0. This means that the model is well behaved. Anyway, given the R^2 value, it only describes the 47.5% of the variance. The appropriateness of the model depends on its usage. If it is intended to be used for the subway service level management it could be considered useful.



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.

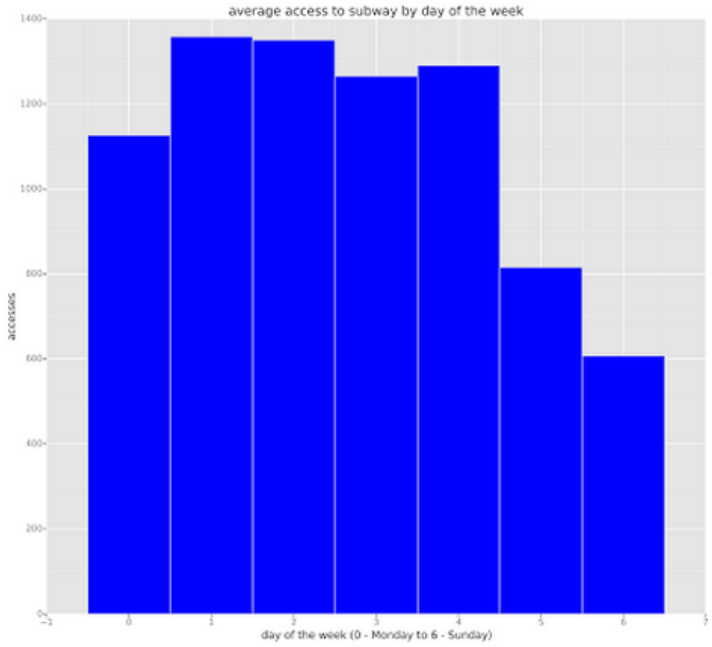
The following figure reports a histogram containing the number of occurrences of accesses for both rainy and non-rainy days. As stated in Section 1, it is clear that the two distributions are not normally distributed. However, it is necessary to consider that non-rainy days are fewer than rainy days then this figure alone does not hold statistical value.



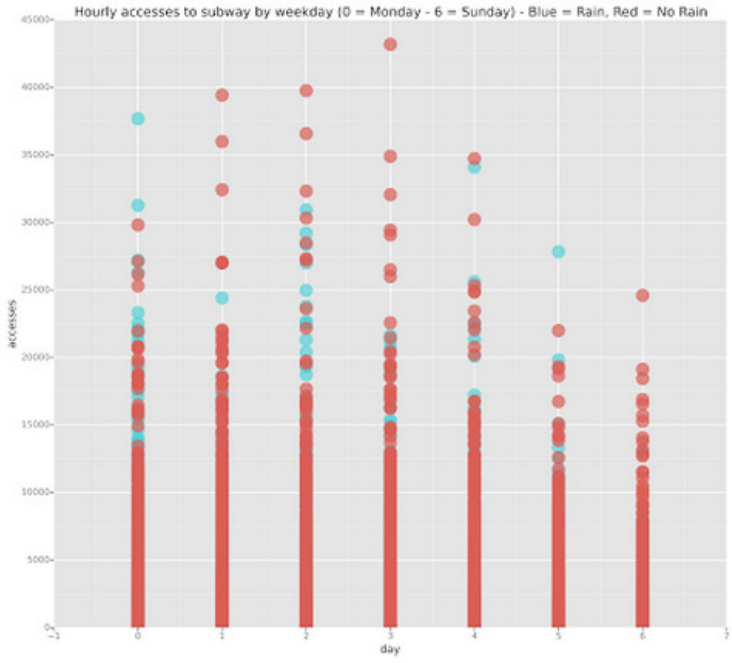
3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:

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

The following figure reports the average number of ridership by day of the week. The average number is more significant than the total sum because each month can have a different number of weekdays (e.g. in May 2011 there were 5 Mondays, but only 4 Saturdays). This figure confirms that subway is more used during weekdays than in the weekend, probably due to commuters.



Another visualization, based on the following figure, shows the hourly number of entries by weekday and also by rain condition (blue is rain, while red is no rain). It is clear that the number of rainy days and hours is much smaller than non-rainy ones, and that lower level of ridership on all days of the week are associated to non-rainy hours. Moreover, as per the previous figure, an important impact to the ridership is based on the weekday.



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?

Yes, according to the results of the Mann-Whitney U-Test, there's a small but significant difference that leads to the interpretation that more people ride when it's 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.

The conclusion in mainly driven by the statistical tests. Whereas, using linear regression it is not possible to determine a clear impact due to weather condition in general, and rain in particular, on the subway ridership. In fact, weather conditions seem to explain less than half of the phenomenon and some other factors seem to have an important impact, such as the single subway stop, the time of the day and the day of the week.

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.

For what concerns potential pitfalls deriving from the dataset, it should be considered that the phenomenon is too complex to be studied with a dataset including only one month. In fact, weather conditions are not comparable in such a short time. Moreover, we do not have any information about possible exogenous variables that could have impacted our analysis, such as important events in the city or subway service maintenance.

For what concerns the methods used some attention must be focused on some points: (i) the assumption of linearity of the phenomenon, (ii) weather variables could be subject to multicollinearity, and (iii) statistical tests should have been conducted on grouped data to eliminate noise coming from less used subway stops or night hours.

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

I would perform a principal component analysis to reduce and combine the number of weather features.
Some more inputs are given by the combination of fog and rain performed through mapreduce. The average access differences by weather condition combinations, shown in the table below, if statistically proven, could reveal an even more important contribution of the fog than the rain in subway ridership.

Weather condition	Average accesses
No fog-rain	1098.95330076
No fog-no rain	1078.54679697
fog-no rain	1315.57980681
fog-rain	1115.13151799