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.

http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
http://stattrek.com/statistics/dictionary.aspx?definition=coefficient_of_determination
http://statsmodels.sourceforge.net/0.5.0/generated/statsmodels.regression.linear_model.OLS.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 used the Mann-Whitney U test to analyze the NYC subway data, comparing the distribution of ridership for samples with rain vs. those without rain. I used a two-tail P value. My null hypothesis was that ridership would be the same with rain as without. My p-critical value was 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 a nonparametric test well-suited to comparing non-normal distributions. A brief examination of the histogram from section 3.1 is adequate to characterize both of our distributions as being non-normal, with a heavy skew to the right in both cases.

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 computed P value was *nearly* 0.05 (0.049999825586979442), with a mean of 1105.44 for samples with rain, and a mean of 1090.28 for samples without rain. Regarding the P value, the SciPy implementation of the Mann-Whitney U test provides a one-tail value by default. I doubled the computed value from Problem Set 3, Exercise 5 to get a two-tail value.

1.4 What is the significance and interpretation of these results?

I would classify these results as being statistically significant, with a P value just slightly less than the designated P-critical value. My interpretation is that the mean ridership varies between the two groups because more people ride the subway when it's raining.

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. Gradient descent (as implemented in exercise 3.5)
b. OLS using Statsmodels
c. Or something different?

I used OLS using Statsmodel to compute the coefficients and generate predictions.

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 continuous/numeric features:

rain, precipi, fog, mintempi, maxtempi, meanwindspdi

I used dummy variables to represent each unique value for UNIT.
I used dummy variables to represent each unique value for Hour.

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 R^2 value."

UNIT is a categorical variable which is effectively a unique location identifier. The dummy variables for UNIT were included because I think it's safe to assume that ridership will vary based on the location of a given subway stop. Stops will be surrounded by populations of varying densities, and may also be used more frequently at different times of day due to surrounding public facilities, etc.

Hour looks like a continuous variable at first glance, but I assumed that certain times of day would have radically different usage patterns than others (morning rush hour, lunch time, etc.). Because of this, I thought it would make more sense to give each hour its own input weight, thus I treated Hour as a categorical variable and created dummy variables to represent each distinct value. My assumption in this case was borne out by increased R^2 when treating Hour as categorical.

Inclusion of the remaining features was based on various assumptions made on my part, which are:

rain - that more people will tend to ride the subway when it rains
precipi - that greater amounts of precipitation will tend to increase ridership
fog - that the presence of fog, which reduces driving visibility, will tend to increase ridership
mintempi, maxtempi - that extreme temperatures in either direction will probably increase ridership
meanwindspdi - that higher winds will lead to increased ridership

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

The final weights for each of the non-dummy features were:

rain	**43.77854494**

precipi	**-72.99501577**
fog	**235.24826264**
mintempi	**-18.18357560**
maxtempi	**5.97974050**
meanwindspdi	**32.59278384**

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

My model's R^2 value was 0.523172699299.

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?

It means that my regression model fits, or is able to predict, about 52.3% of the variability in ridership. I think that's rather a low number in this context. Examining the residuals in Problem Set 3, Exercise 6, I observed a frequency distribution with long tails and a plot having obvious cyclical variability. This leads me to believe that a more complex, non-linear predictive model would do a better job.

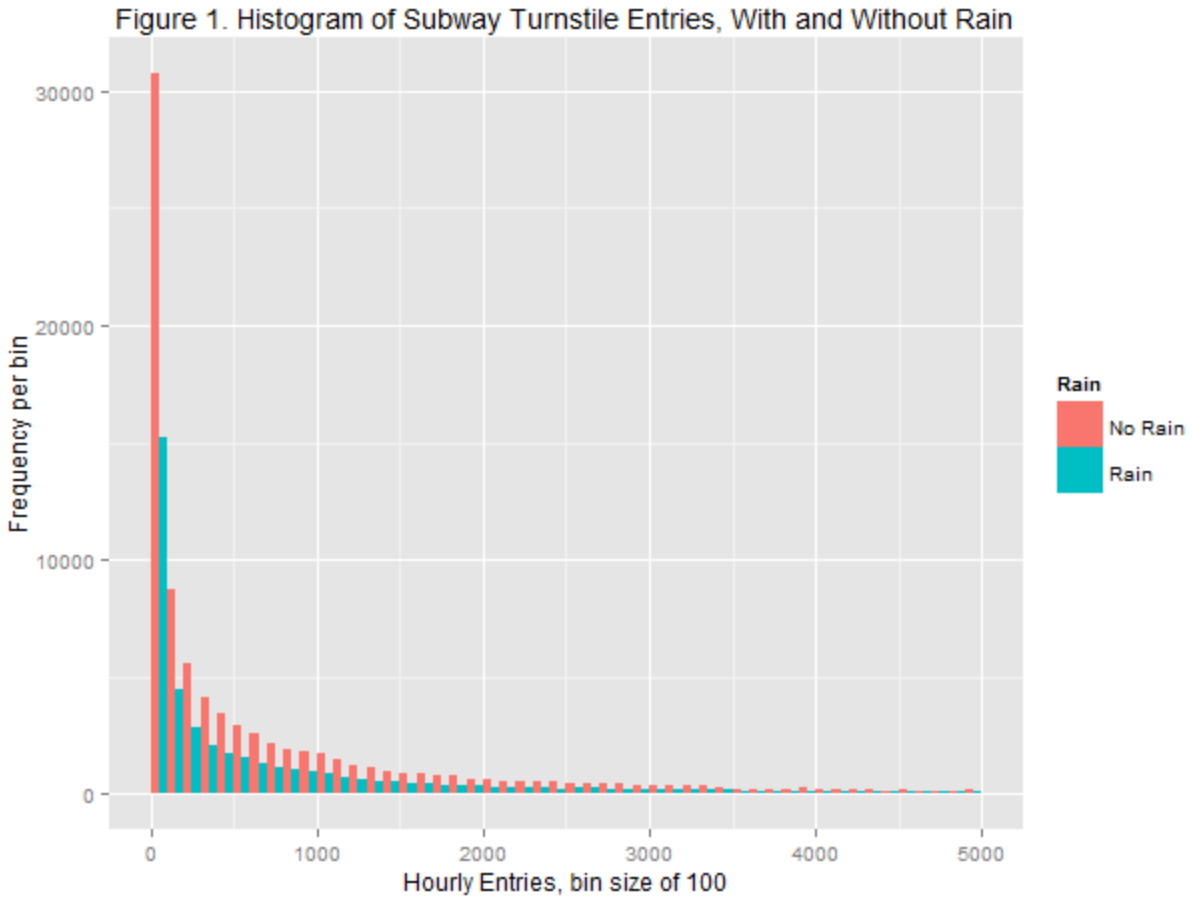
Section 3. Visualization

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.

In Figure 1, we get a sense of the distributions for turnstile entries with and without rain. There is an apparent abundance of riders in samples without rain, but this should not be taken to indicate that average ridership is higher when it's not raining. Rather, there are more entries in the dataset without rain (87,847 entries) than there are with rain (44,104 entries). Both distributions are heavily right skewed, as noted in section 1.2, and have long tails. Figure 1 clips both tails at 5000 entries for the sake of readability.

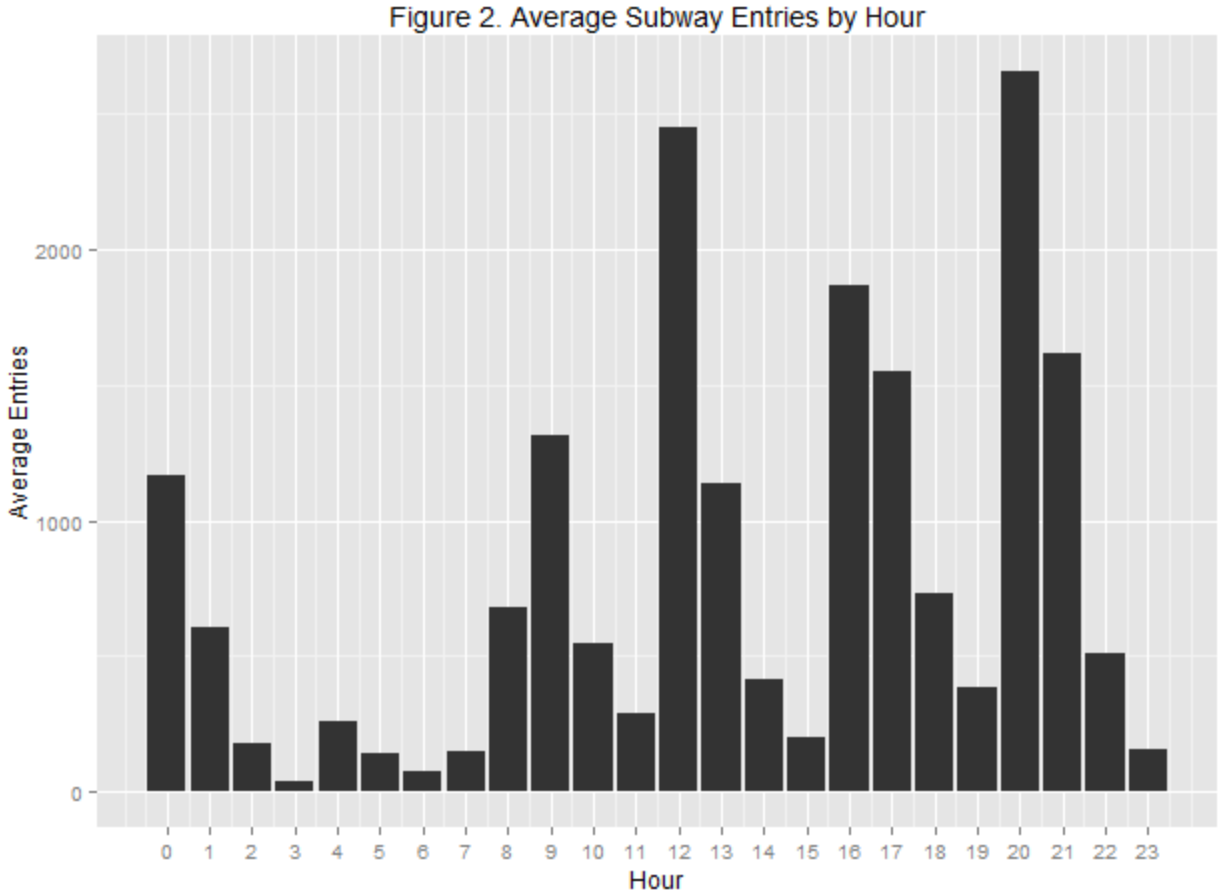


3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or

attempt to implement something more advanced if you'd like. Some suggestions are:

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

In Figure 2 we substantiate our assumption from section 2.3 that Hour is well interpreted as a categorical variable. There are significant spikes in ridership from one hour to the next, having no easily discernible pattern. A linear weight applied to Hour would fail to capture that variability and would lead to a poorer fit in our model, as was born out during model tuning.



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?

I think it's clear that more people ride the subway when it's raining. The effect is not as pronounced as I had expected, but the various analyses performed definitely indicate that there is an effect. I was surprised in particular by the relatively small difference between the mean of turnstile entries for rows with rain vs. those without rain.

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 first analytical technique to provide support for my conclusion was the Mann-Whitney U test, which yielded a P value just below my designated P-critical value of 0.05 when comparing ridership with rain vs. ridership without rain. The positive coefficient for the 'rain' feature (43.77854494) in my linear regression model also supported my conclusion. However, as with the unexpectedly small difference in mean of entries for the two samples, I was expecting a larger coefficient here.

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.

Dataset - I think the most obvious shortcomings of the dataset are its seasonal scope and its limited size. The entire dataset is from May of 2011. For a lot of people, spring is a nice time to be outside, which might impact ridership regardless of, or in spite of, the presence of rain.

There could also be a variety of socio-political events at work, relative to that particular date range. A much larger dataset, preferably one spanning multiple years rather than a single month, would provide a more robust analysis in the face of unknown factors affecting ridership. It would allow us to smooth out variability based on seasonal trends and one-time events, which would improve the analysis considerably.

Analysis - The statistical tests used seem adequate to the task of establishing correlation between rain and increased ridership in the dataset provided. I was a little confused by the coefficients yielded by the linear regression model I used. The contrasting positive and negative values yielded for 'rain' and 'precipi', respectively, were particularly interesting. I expected positive coefficients for both, and am not entirely sure why *the presence of rain* would increase ridership, while *greater amounts of rain* would appear to decrease it. I think that a non-linear model with additional features would provide better results, as noted in section 2.6.