Analyzing the NYC Subway Dataset

Short 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 1. Statistical Test**

1.1 Which statistical test did you use to analyse the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis?

The first question in Problem Set 3 asks me to plot two histogram charts and, from eyeballing the chart, decide whether the data is normally distributed and whether I can use the Welch’s t-test on the data. My initial instinct was the data isn’t normally distributed, but rather skewed to the left. Later, the problem set has me calculate the Mann-Whitney U test, a non-parametric test, so I felt even more certain the data wasn’t normally distributed.

However, for this project, I decided to run the Shapiro-Wilk test that Dave mentioned briefly to determine whether the sample of data conform to a normal distribution. (See the script SubwayDataShapiroWilk.py for code.)

The results were:

UserWarning: p-value may not be accurate for N > 5000.

warnings.warn("p-value may not be accurate for N > 5000.")

[(0.5938820838928223, 0.0), (0.5956180691719055, 0.0)]

I researched how to interpret this research (<http://mvpprograms.com/help/mvpspc/distributions/NormalityTestingGuidelines> and <http://www.graphpad.com/support/faqid/959/> were good sources). Some sources said the Shapiro-Wilk test is for best for 3-5000 sample size. Because the turnstile data is a very large sample (N > 5000), it seems like the t-test and ANOVA will be robust enough to handle this data set, so I should be OK running a t-test with the subway data:

“**So how useful are normality tests?**

Not very. Normality tests are less useful than some people guess. With small samples, the normality tests don't have much power to detect nongaussian distributions. With large samples, it doesn't matter so much if data are nongaussian, since the t tests and ANOVA are fairly robust to violations of this standard.” (<http://www.graphpad.com/support/faqid/959/>)

In this case, I decided to run a t-test on the subway data to compare the two samples entries hourly with rain and without rain as well as interpret the Mann-Whitney U test I had already performed as part of Problem Set 3.

**T-test**

- null hypothesis = two samples came from the same population

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.

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.

1.4 What is the significance and interpretation of these results?

**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:

Gradient descent (as implemented in exercise 3.5)

OLS using Statsmodels

Or something different?

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

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.”

2.4 What is your model’s R2 (coefficients of determination) value?

2.5 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?

**Section 3. Visualization**

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

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, you might have one interval (along the x-axis) with values from 0 to 1000. The height of the bar for this interval will then represent the number of records (rows in our data) that have ENTRIESn\_hourly that fall into 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. Some suggestions are:

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

Which stations have more exits or entries at different times of day

**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 versus when it is not raining?

4.2 What analyses lead you to this conclusion?

**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 data set and the methods   
of your analysis.

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