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, 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.

1. **Obviously I used Udacity’s discussion board for reference.**
2. **I used pypi.python.org, especially to refer to the ggplot.**
3. **I used the statsmodel class from** [**http://statsmodels.sourceforge.net/0.5.0/generated/statsmodels.regression.linear\_model.OLS.html**](http://statsmodels.sourceforge.net/0.5.0/generated/statsmodels.regression.linear_model.OLS.html) **to calculate my r^2**
   1. **I used numpy mean and sum as part of my r^2 formula from** 
      1. [**http://docs.scipy.org/doc/numpy/reference/generated/numpy.mean.html**](http://docs.scipy.org/doc/numpy/reference/generated/numpy.mean.html)
      2. [**http://docs.scipy.org/doc/numpy/reference/generated/numpy.sum.html**](http://docs.scipy.org/doc/numpy/reference/generated/numpy.sum.html)

**Section 1. Statistical Test**

* 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 because the result of the histogram 3.1 indicates our data set is non-parameteric and test is appropriate to evaluate our hypothesis.**

**Our null hypothesis is…**

**….that distribution of people taking subway when it is raining….**

**… is not significantly different…**

**… to the distribution of people taking the subway when it is not raining (in other words, can we conclude that rain has no effect on people taking the subway?).**

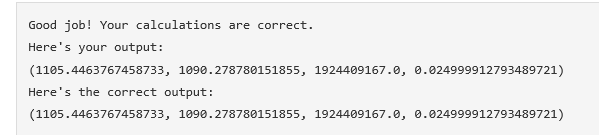
**Our alternative hypothesis would show that the two distributions are significantly different.**

**I choose a p-critical of 0.05**

* 1. 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.

**because the result of the histogram 3.1 indicates our data set is non-parameteric and test is appropriate to evaluate our hypothesis. The histogram below is showing that the more people take the subway whether it is raining or not, the less frequent it will happen.**

* 1. 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 of people taking subway when it is raining: 1105.45**

**Mean of people taking subway when it is not raining: 1090.28**

**U test = 1,924,409,167**

**P-value = 0.0249**

* 1. What is the significance and interpretation of these results?

**I chose a P – critical of 0.05 and the P-value the Mann Whitney U Test returned to me is 0.024999. But I would like to consider a Two tail test because we are trying to answer the question “Do people take the subway more or less when it is raining compared to when it is not?” So I double the value of 0.024999 giving me the P-value for a two tail is 0.04999; which is just slightly below my chosen P-critical value of 0.05. So therefore I have failed to reject our null hypothesis and conclude that the average number of people taking subway when it is raining is NOT significantly equal to the average number of people taking the subway when it is not 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:

1. OLS using Statsmodels or Scikit Learn
2. Gradient descent using Scikit Learn
3. Or something different?

**I used OLS using Statsmodels, see below**

**import numpy as np**

**import pandas**

**import statsmodels.api as sm**

**def linear\_regression(features, values):**

**features = sm.add\_constant(features)**

**model = sm.OLS(values,features)**

**results = model.fit()**

**intercept = results.params[0]**

**params = results.params[1:]**

**return intercept, params**

**def predictions(dataframe):**

**features = dataframe[['rain', 'precipi', 'Hour', 'meantempi']]**

**# Values**

**values = dataframe['ENTRIESn\_hourly']**

**# Perform linear regression**

**intercept, params = linear\_regression(features, values)**

**predictions = intercept + np.dot(features, params)**

**return predictions**

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

**Feature variables: *‘rain’,* *‘precipi’, ‘Hour’, and ‘meantempi’***

**Yes I used the dummy variable of ‘unit’ because different subway units may have an effect on how often people take that subway unit. Maybe a unit will go from an unpopulated area to an unpopulated area and another subway unit can go from a very populated area to an area of New York heavily concentrated with business buildings. Which therefore, the subway unit may have a very large effect on r^2 result.**

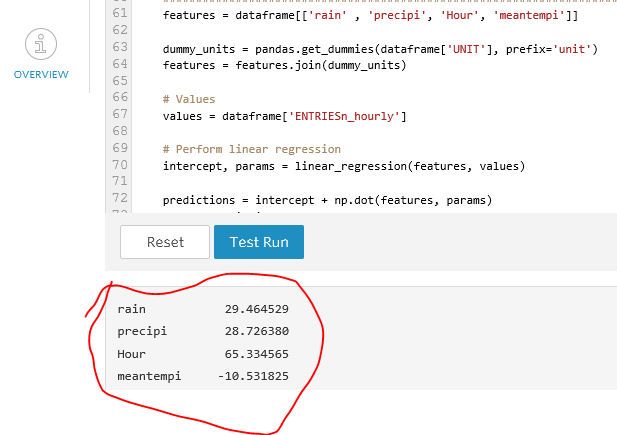
2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that

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.”
  + **‘rain’ is an obvious choice because it is a simple true/false indicator for rainy and non rainy days.**
  + **‘precipi’ is chosen because precipitation is obviously tells you if it is raining and how much. Maybe people wouldn’t take the train when it is just dripping but will most likely take the train when it is pouring very hard.**
  + **‘Hour’ is needed because the time of day affects the population using the subway. Around 12am to 5am, people are hardly going to places whether it rains or not.**
  + **‘meantempi’ is chosen because temperature is possible indication or rain. Usually when it gets cold it may mean it is about to rain because cloudly days blocks the sun thus becomes cold. Temperature can also be a variable that can affect our predictive model. For instance, maybe people tend to take the train when it is very cold than hot despite whether it is raining or not.**

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

**These are the parameters….**



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

**0.479**

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 R2value?

**An R^2 of 0.479 only explain nearly 48% of the variables of our regression model. What the means is that our Xs (in this case are the features of “rain”, “precipi”, “Hour”, and “meantempi”) has little or poor correlation with predicting the number of people taking the subway (“ENTRIESn\_hourly”). There is a weak tendency for Y to follow as X increases. What it looks like is that nearly 52% of our points are scattered far away from the true regression line. Ideally we would want to explain all or nearly all of our variables (probably at least 80%). So an R^2 of 0.479 indicates a poor fit to our regression model based on the features I have chosen.**

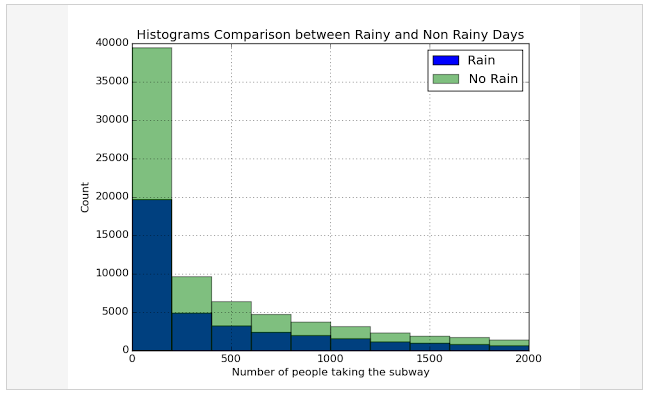
**Section 3. Visualization**

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.

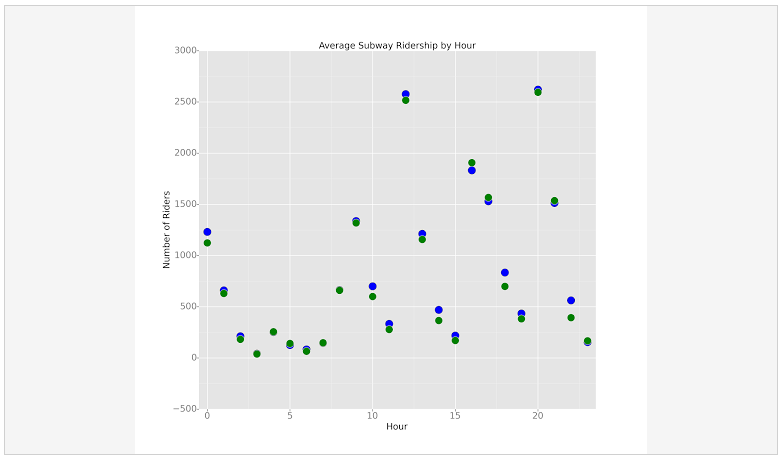
* You can combine the two histograms in a single plot or you can use two separate plots.
* If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
* For the histograms, 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.



**From this histogram, it is showing that we have more data of people taking the subway when it is not raining compared to when it is raining (because the green bars are usually higher than the blue bars). It is also telling us that the frequency of people taking the subway decreases as the more the people take the subway. And last, this tells us our data will not be normal and will require a non parametric tests (Mann Whitney U Test) if we want to perform further analysis.**

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



**This is scatter plot chart shows the average number of people riding the subway in every typical hours for each weather situation (raining and not raining).**

**The dots in blue are the value of the average people taking the subway in a given hour when it is raining**

**The dots in green are the value of the average people taking the subway in a given hour when it is NOT raining**

**The green and blues does for each given hour are not far away apart from each other. In fact, if the plot of the average number of people when it is raining and the average number of people when it is not raining looks almost identical. Therefore, we conclude visually that there is no significance with people taking the subway when it is raining or not.**

**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?

**From our analysis, we do not believe that people ride the NYC subway more (or less) whether it is raining or not.**

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.

**From question 1.4, I concluded the hypothesis that rain does not have an effect on the people taking the subway through the performed Mann Whitney Test, the scatter plot of number of people per both rain and non rainy days, and the coefficient of determination.**

* **The Man Whitney Test shows the P-Value of 0.04999 is much less than my proposed P-Critical of 0.05 thus rejecting the null hypothesis that the two distributions (people taking the subway when it is raining vs not) are not too much different under rainy conditions.**
* **The scatter plot showing hardly a difference visually between the plot of people taking the subway when it is raining and the plot of people taking the subway when it is not raining.**
* **And my coefficient of determination returning an R^2 value of 0.48; a poor goodness of fit. Also we c an note the parameters of non dummy variables. We have a parameter for “rain” of 29.465 with a Y intercept**

**of 1,539 (number of people using the subway). We only have two values for rain: 1 and 0 (either raining or not raining). So a X value of 1 indicates adds 29.495 people to the 1,568 making a Y change of 1,597.5 per subway unit. A value of 1,597.5 when it is raining is not that much different than 1,568 so therefore the condition of whether it is raining or not does not have an influence on the number of people taking the subway. There are other features that I included that will also influence our Y value and those are precipitation, Hour, and average temperature. But the factor of rain itself, which is the main question of this project, hardly has an impact on predicting the Y value ( number of people taking the subway).**

**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.

**One bias the found in our data set is using “ENTRIEs\_hour” meaning the number of people entering into the subway in a given hour of the day. The reason this is bias is because the time of day people using the subway has a large influence on our data. Some people can possibly take the subway more during mornings and evenings where people have to commute between work and home but hardly anyone would take the subway at night because most are already asleep. Therefore, this is an important factor that needs to be considered and filtered out. A hypothesis question that would probably be better is if we measure the number of people taking the subway, at a given hour, only during the weekday, and only during a certain season. Like for example, “Do more people take the subway when it is raining every weekday between 6AM and 7AM during summer”.**

**I am pointing out hour because there are too many reasons why people take the subway at a given hour besides the fact that it is raining.**

**I am pointing out weekday because probably they are the days people are on a normal routine. Whereas weekends can be chaotic probably do to exciting events that people go out of the way to attend using the subway.**

**I am pointing out season (I said summer) because in the winter it is usually more snowing and hailing where as there can probably be more rain in the other 3 seasons. Not to also include that people like to tend to walk during summers and take the subway during winter when it is very cold outside.**

**So hypothesizing if people taking the subway when it is raining when ONLY given a set of specific conditions is the best way to ask.**

**Unfortunately, our data set is too small to answer any specific hypothesis like the example I gave. One thing wrong with our data is that it is only during the period of May. So making any conclusions from data derived only during the month of May will not give any good returned conclusion what it may be year round.**

**Regarding the linear regression: A linear regression is not appropriate for our case as we are measuring if the number of people take the subway more when it is either “raining” or “Not raining”. Using two discrete measures of raining and not raining is not appropriate for linear regression and continuous values like amount of precipitation is more appropriate. So a linear regression is appropriate for measuring the number of people taking the subway in a given hour as the percentage of precipitation increases rather than measuring the number of people taking the subway in a given hour when it is just raining or not raining.**

**Overall, my conclusion with answering the question: “Do more people take the subway when it is raining” when given the data set to try to answer is not a good question to ask at all. It needs to be broke down more to filter out bias such as the example question I said: “Do more people take the subway when it is raining every weekday between 6AM and 7AM during summer?”.**

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