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

Welch's T-Test research: http://en.wikipedia.org/wiki/Welch%27s_t_test

Section 1 References

Mann-Whitney Test: http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
SciPy Documentation http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
Forums: http://discussions.udacity.com/t/question-regarding-the-project/16311/2
Null Hypothesis:
http://www.socscistatistics.com/tests/mannwhitney/
https://statistics.laerd.com/premium-sample/mwut/mann-whitney-test-in-spss-2.php
1 tail versus 2 tail Tests:
http://www.graphpad.com/guides/prism/6/statistics/index.htm?stat_checklist_mannwhitney.htm
U-Critical Values:
http://psych.unl.edu/psycrs/handcomp/hcmann.PDF
http://www.statisticslectures.com/topics/mannwhitneyu/
http://psych.unl.edu/psycrs/handcomp/hcmann.PDF
http://stats.stackexchange.com/questions/116315/problem-with-mann-whitney-u-test-in-scipy
Z – Critical value at 1.96: http://en.wikipedia.org/wiki/1.96

Section 2 References

Step-Wise Regression: http://en.wikipedia.org/wiki/Stepwise_regression

Section 3 References:

GroupBy Function: http://pandas.pydata.org/pandas-docs/dev/groupby.html
X-Axis: http://stackoverflow.com/questions/23541497/is-there-a-way-to-plot-a-pandas-series-in-ggplot
Axis Breaks: http://docs.ggplot2.org/current/scale_continuous.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?

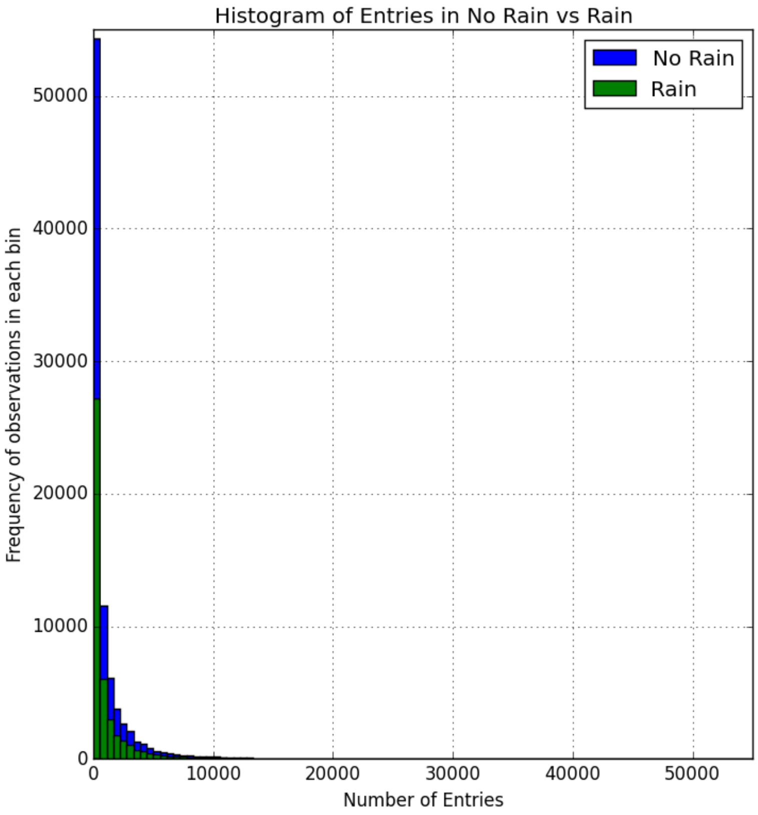
I used a Mann-Whitney U test with a Two-Tail P value in my test.

*The **null hypothesis** for my test is:* H_0: the distributions of the two groups are equal
[In common parlance, *that there's no difference in ridership on rainy versus non-rainy days].*

The alternate hypothesis is: H_A: the medians of the two groups are not equal
The p-critical value is: 0.0499998 (2 X .024999912793489721)

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 is applicable because it **does not** rely on the assumption of 'normality'. In our subway data ridership (i.e. 'Entriesn') is not a normal distribution (see chart below). Although Mann-Whitney does not require an assumption of normality, in order to verify the Alternate Hypothesis that the **medians** are not equal, the Mann-Whitney **does** require an assumption that the underlying distributions are equal. (Referenced from this website:* https://statistics.laerd.com/premium-sample/mwut/mann-whitney-test-in-spss-2.php). One can see from the chart below that the underlying distributions are quite similar as both conform to a power-law distribution



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 Mann-Whitney I obtained were:

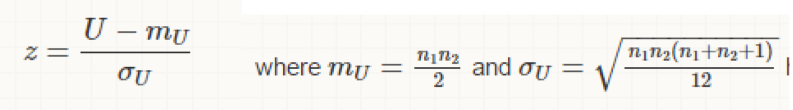
Mean ridership with rain:	*1105.4463767458733*
Mean ridership with OUT rain:	*1090.278780151855*
Mann-Whitney U value:	*1924409167.0*
P-Critical Value:	*0.024999912793489721*

Note because this is a 2 tail test we multiply P-Critical by 2 (0.049999825587)

1.4 What is the significance and interpretation of these results?

To interpret the results we must compare the U value to the U critical value. Since our sample size is well over 20 we can use Z-score as the critical value (formula below) [Referenced from this website: http://psych.unl.edu/psycrs/handcomp/hcmann.PDF]. Therefore at alpha = .05 our critical Z is ± 1.96. If our calculated Z is greater than 1.96 or less than -1.96 we will reject the null hypothesis. [Note - The precise value to 20 decimal places is 1.95996 39845 40054 23552.}

The calculated Z is as follows:



$$z = \frac{U - m_U}{\sigma_U} \quad \text{where } m_U = \frac{n_1 n_2}{2} \text{ and } \sigma_U = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}$$

Where, N_1 = Sample size (n) of ridership with Rain = 44104

N_2 = Sample size (n) of ridership without Rain = 87847

$$m_u = \frac{n_1 * n_2}{2} = \frac{44104 * 87847}{2} = 1937202044$$

$$\sigma_u = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}} = \sqrt{\frac{44104 * 87847\ (44104 + 87847 + 1)}{12}} = 6527093.331$$

$$Z = \frac{U - m_u}{\sigma_u} = \frac{1924409167.0 - 1937202044}{6527093.331} = -1.95996538601$$

Since our Z score (-1.959965) here is less than the precise Z critical value (-1.959963) we *reject the null hypothesis that these distributions are equal. In other words, we reject that the distribution of ridership is equal and we accept the alternative hypothesis the median ridership are, in fact, different on Rainy days versus Non-rainy days at an alpha of .05.* ***Given that mean ridership with rain is 1105 and mean ridership without rain is 1090 we could reasonably conclude that ridership with rain is greater than ridership without rain.***

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?

I implemented the OLS using Statsmodels

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

Input variable I used in my model were: maxpressurei, maxdewpti, mindewpti, minpressurei, meanpressurei, meanwindspdi, mintempi,

Dummy variables were used for the UNIT variable using the code:

dummy_units = pandas.get_dummies(dataframe['UNIT'], prefix='unit')

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.”

The rationale for my variable selection is based on Backward Step-wise Linear regression (reference from website: http://en.wikipedia.org/wiki/Stepwise_regression). In this methodology all variables are initially included and subsequently any variable with a T-Value of less than 1.53 is excluded. The threshold of 1.53 was developed using the Bonferroni point which reveals “how significant the best spurious variable should be based on chance alone. On a t-statistic scale, this occurs at about:

$\sqrt{2 \log p}$ *, where p is the number of predictors.”*

*After running the **first regression** (with all variables included), the T-value for each variable are as follows (Dummy variables for UNITS not included).*

VARIABLE	***T-Value***
Hour	*25.165129*
maxpressurei	*-1.998123*
maxdewpti	*1.595700*
mindewpti	*-0.987128*
minpressurei	*-2.241075*
meandewpti	*0.204202*
meanpressurei	*2.616349*
fog	*1.089965*
rain	*-0.536345*
meanwindspdi	*3.126026*
mintempi	*-1.170723*
meantempi	*0.452376*
maxtempi	*-0.343624*
precipi	*-0.509096*
thunder	*-0.966295*

I then remove the variable with the smallest absolute value that is below 1.53 *(above this would be ‘*meandewpti*’.) I then re-run the model until all t-values that remain exceed the 1.53 threshold*

*The variables eliminated (in order) are: ‘*meandewpti’, ‘maxtempi’, ‘precipi’, ‘rain’, ‘fog’, ‘meantempi’, ‘thunder’

The variables that remain and their corresponding t-values are:

VARIABLE	***T-Value***
Hour	*25.191941*
maxpressurei	*-2.226894*
maxdewpti	*3.910068*
mindewpti	*-2.759943*
minpressurei	*-4.040406*
meanpressurei	*3.597480*
meanwindspdi	*2.970189*
mintempi	*-4.261653*

The surprising elimination of rain and precipitation may be explained by the fact that rain and precipitation are likely closely correlated to barometric pressure. That is, barometric pressure is a very good indicator of in climate weather (i.e. wind, thunderstorms, snow, etc.).

Equally surprising might be the inclusion of dew point as a critical variable. However, dew point is also closely related to precipitation (see website: http://ww2010.atmos.uiuc.edu/%28Gh%29/guides/maps/sfcobs/dwp.rxml) since dew points are a measure of how much moisture is in the air. Including dew point may reduce the explanatory value of rain and precipitation since dew point is a more comprehensive measure of moisture (i.e. measures moisture with relative humidity less than 100%).   Interestingly, if one were to continue the investigation into other variables: relative humidity *might be an interesting variable to consider. Relative humidity can be inferred by examining the degree of closeness between actual temperature and dew point*

An additional interesting variable that was kept in our analysis is mintempi. Also surprising is that meantempi or maxteampi were excluded*. A possible explanation for this is that riders are particularly sensitive to cold (and not so sensitive to hot). If this theory is correct mintempi would be the appropriate measure to explain this sensitivity (not so with meantemp or max temp)*

Lastly, unsurprisingly, Hour is one of the best predictors of ridership with one of the highest t-values (25)

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

The coefficient of the non-dummy feature in my linear regression are as follows:

VARIABLE	**Coefficients**
Hour	62.496224
maxpressurei	-1325.185856
maxdewpti	26.578563
mindewpti	-14.069428
minpressurei	-1903.051070
meanpressurei	3121.357180
meanwindspdi	27.755674
mintempi	-23.338327

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

The R2 of my model is .486818

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?

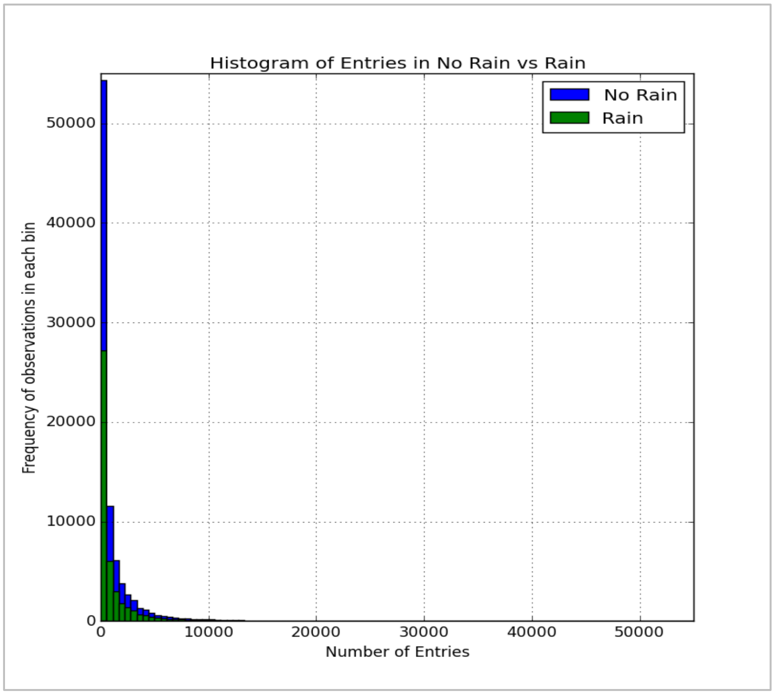
The R2 value here of .48 reveals the model explains 48% of the total variation in ridership using our input variables ('Hour', 'maxpressurei', etc.). The 48% indicates our model is certainly not perfect (a perfect score being 1.0). However, given that complexity of the systems we are measuring (i.e. social and weather systems) explaining 48% of the variation of seems like a very useful model for prediction. Furthermore, our relatively high T-Values for each variable (all above 2.2) and an F Value of 19 also support this conclusion.

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.

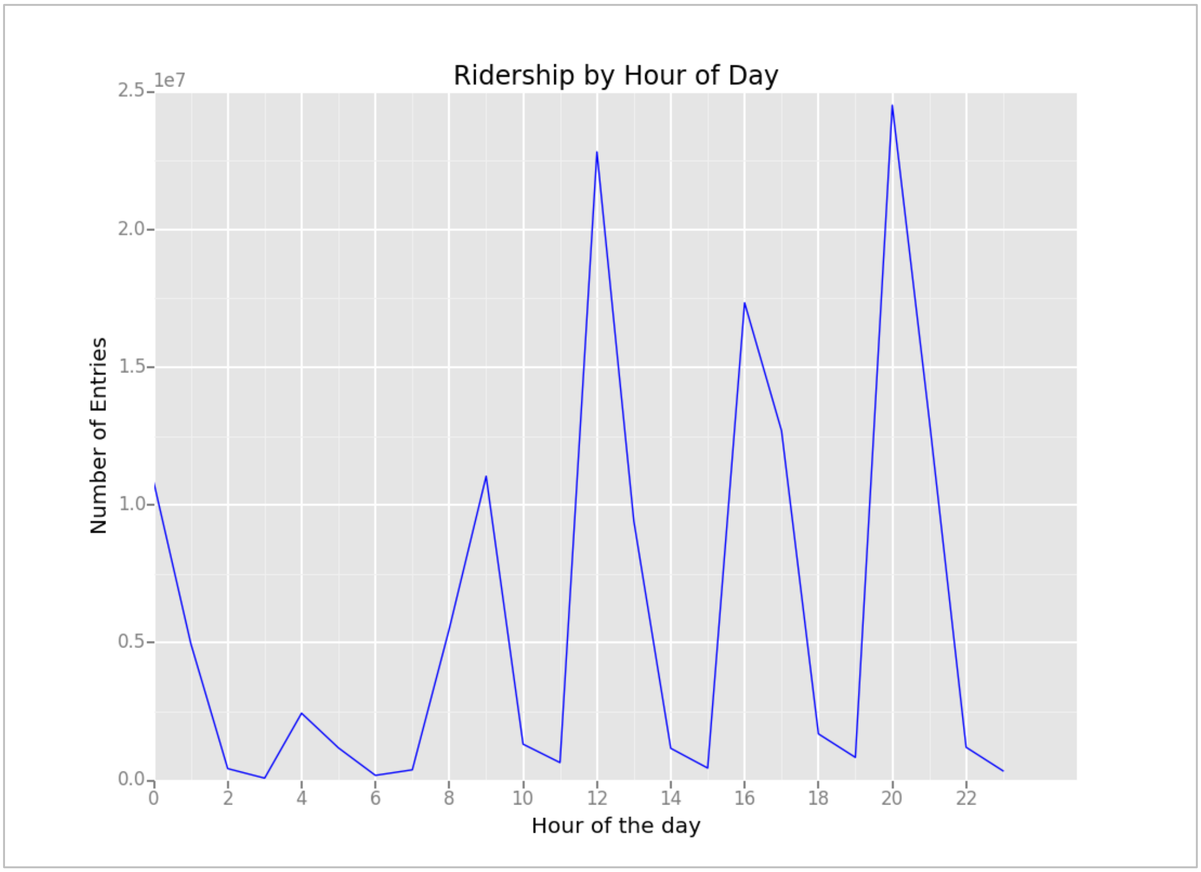


Comments: *This histogram shows that the distributions of ridership for Rainy days and Non-Rainy day both follow a Power-Law distribution with the vast majority of observations of Entries near zero, and few observations with extremely large number of entries (as many as 55,000). It also displays that there are more occurrences of No-Rain than Rain.*

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 figure below illustrates **ridership by hour of day**. The graph shows the expected peaks at 9am, 12noon, and evening (4-5pm). Interestingly, there is also a large spike at 8pm and midnight. A speculative guess is that this traffic corresponds to late-working commuters and dinner-goers (at the 8pm peak) and those coming home from bars at midnight.*



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?

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.

As stated in Section 2…we have rejected our null hypothesis that ridership distributions are equal for rain and no-rain. And we have accepted the alternative hypothesis the medians for ridership are, in fact, different on Rainy days versus Non-rainy days at an alpha of .05. ***Given that mean ridership with rain is 1105 and mean ridership without rain is 1090 we could reasonably conclude that ridership with rain is greater than ridership without rain.***

Additionally, as stated in Section 3, rain was ***NOT*** *a variable I chose to keep in my final regression model. This was based on the fact that the T-Value did not meet the required threshold of 1.53 for inclusion. As mentioned earlier this* ***does not mean that rain is a poor indicator of rider****, but rather that rain is likely correlated with other variables (like barometric pressure) and that when included barometric pressure is a better predictor of ridership (removing the variable rain).*

As a result I would rely on the results from Section 2 to conclude that ridership does vary with rain.

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,

   There are some underlying problems in the data set. First, the data covers a 30 day period in May of 2011. There is likely some sampling risk in making conclusion from such a short period of time. For instance there is likely a great deal of seasonality in ridership (i.e. December ridership is very different from May). If we were to use this set to predict ridership in say December we could be at risk of making some very false conclusions.

   *Another risk in the dataset is that each individual Units do not appear to provide data for every hour in the day. Ideally we would prefer each Unit provide a reading **for every hour**.*

   *In very related vein, if we cannot get a "Entriesn" reading for **every** hour we would at least want to read the data at consistent intervals across all machines (i.e. all reading are at 12M, 4am, 8am, etc.). This would reduce the risk of a single machine over-influencing a specific hour. However in our data set this does not appear to be the case. Some machines pull at 4am, some at 5am. Given that ridership is heavily influenced by time and that some turnstiles have enormous traffic, this leads to the conclusion that there is risk that our sample may not accurately represent the population.*

2. Analysis, such as the linear regression model or statistical test.

   The development of my linear regression model was based on backward step-wise regression, where all variables are included initially and subsequently removing variables without proper explanatory value. A risk in this methodology is that we may 'over fit' or model to the data. However, I have attempted to reduce this risk by choosing a rather stringent threshold for the t-value (1.53) and applying the requirement across all variables. A more vigorous analysis might have broken the dataset in a 70/30 split for training, and testing of the model. This would further reduce the risk of over fitting.

   Another assumption of my model is that each of the variables have a linear relationship with "Entries" however it is possible that some of the variables are non-linear in their influence of ridership. Further analysis would need to be executed on the distribution of the residual plots to search for any underlying 'under fitting' of the model

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

Examination of the coefficients in my regression highlight some interesting takeaways:

- *As Wind speed increases Entries decrease (a coefficient of* 27.755674)
- *There is some interesting Co-Efficient related to pressure:*
  - *maxpressurei -1325.185856*
  - *minpressurei -1903.051070*
  - *meanpressurei 3121.357180*

  *These coefficients indicate that a higher **Mean** pressure leads to greater ridership. This conclusion is intuitive as high pressure tends to yield more stable weather. However, both Min pressure and Max Pressure are inversely related to ridership (i.e. lower max or min pressure would predict less ridership). A lower min pressure is perhaps intuitive (lower pressure usually coincides with storms). However more investigation would be needed to develop a theory behind why a lower Max pressure would yield less ridership*.