Analyzing the NYC Subway Dataset

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.

https://github.com/lextoumbourou/study-notes/blob/master/ud617-intro-to-hadoop-and-mapreduce/lesson-3-mapreduce-code.md -- map reducer example

https://wiki.python.org/moin/ForLoop -- For Loop examples

http://stackoverflow.com/questions/19285014/python-name-error-name-not-defined -- key error research

https://docs.python.org/2/library/stdtypes.html -- python boolean discussion

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html -- Man Whitney info

http://docs.scipy.org/doc/numpy/reference/generated/numpy.mean.html --numpy mean

https://docs.python.org/2/library/tokenize.html -- tokenizing Python

https://pypi.python.org/pypi/ggplot/ -- ggplot

http://stackoverflow.com/questions/783897/truncating-floats-in-python -- truncating floats in python

http://www.statsoft.com/Textbook/Multiple-Regression#cresidual – residual variance and R-squared & interpreting the correlation coefficient R

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 Test, which returned a two-tailed p value of .05. This means that it is less than 5% likely that the null hypothesis is true. The null hypothesis is that the two populations are the same. Our test shows the probability that the result that the means are different due to chance is less than 5%.

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 statistical test is applicable to this data set. The test does not assume our data is drawn from any particular underlying probability distribution. The histogram used to explore our data plotted the hourly entries of our NYC subway dataset showed that the data did not appear to be Normally distributed.

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.

With Rain mean = `1105.44637675`
Without Rain mean = `1090.27878015`
P value = `0.0249999127935 (one-tailed)`
P value = `0.05 (two-tailed)`

1.4 What is the significance and interpretation of these results?

The means are different and there is less than a 5% chance that the difference is a random result.

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 used Gradient descent and Stasmodels

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

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

x = weather_turnstile[['Hour', 'maxpressurei','mindewpti','minpressurei','meandewpti','meanpressurei','fog','rain','meanwindspdi','mintempi','meantempi', 'maxtempi', 'precipi']] -- (second try)

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."
- 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."

I played with the data in excel and looked at various options that changed the mean. I also changed the variables in my program to see how the coefficients changed. I also expected that rain would be impactful based on prior Mann Whitney test.

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

Here are the coefficients from my Linear Regression.

	coef
Hour	55.8206
maxpressurei	-1520.4109
mindewpti	-13.3156
minpressurei	-2702.6010
meandewpti	27.0305
meanpressurei	4238.8598
fog	75.0209
rain	23.8486
meanwindspdi	55.8129
mintempi	-129.2457

```
meantempi       179.6198
maxtempi        -76.4883
precipi         -88.4573

Attempt 3: removed variables with potential collinearity.
Hour             55.7007
meanpressurei    21.3292
fog             324.3583
rain            -10.5728
meanwindspdi     52.6721
meantempi        -7.6669
precipi        -161.7852
```

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

```
r^2 three values are:
0.46430082406
0.554813069387 (includes possible collinearity, so thrown out) &
0.207
```

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 R2 value?

R-Square or the Coefficient of determination is used to evaluate our model fit. R-Squared shows the variability of the residual values around the regression line relative to the overall variability. The model's R-Squared coefficient of .4643 means that 53.6% residual variability is not explained by the model. This does not provide confidence that this linear model will appropriately predict ridership.

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

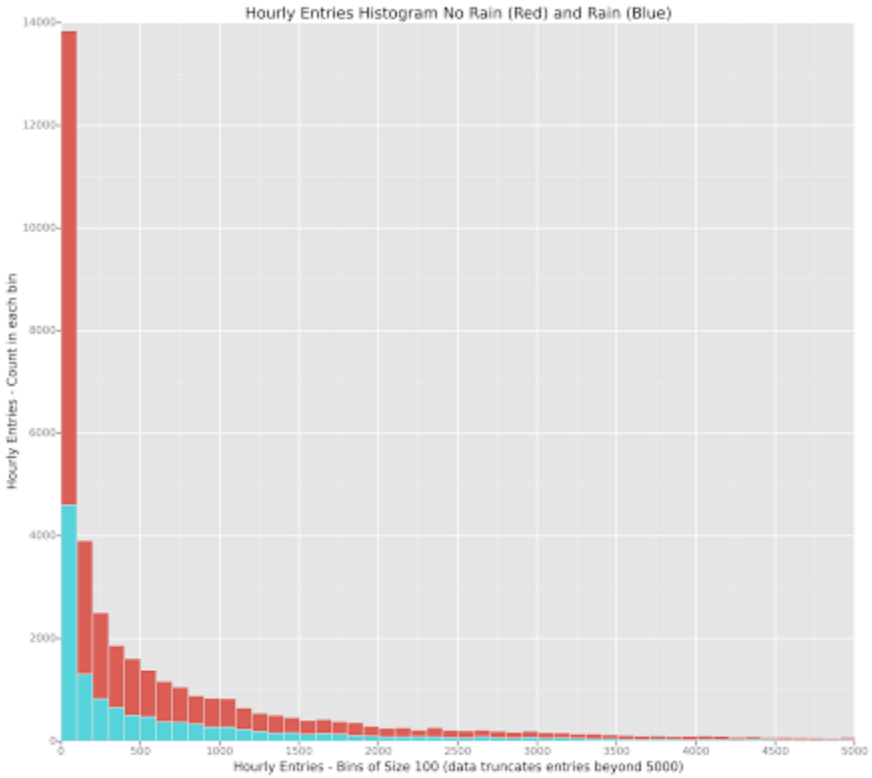
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

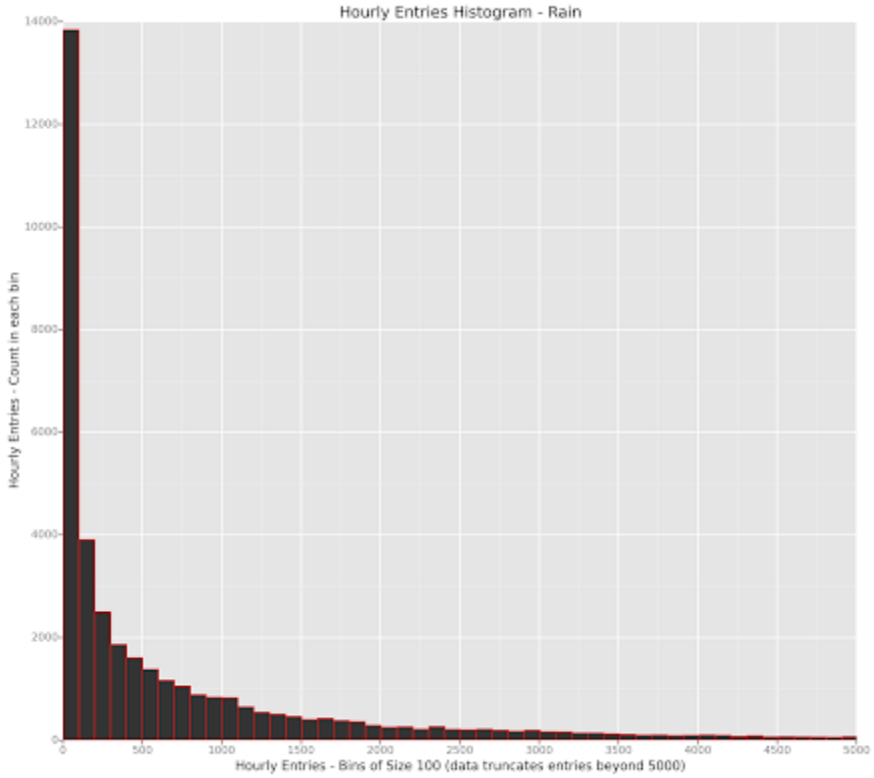
- You can combine the two histograms in a single plot or you can use two separate plots.

- 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.

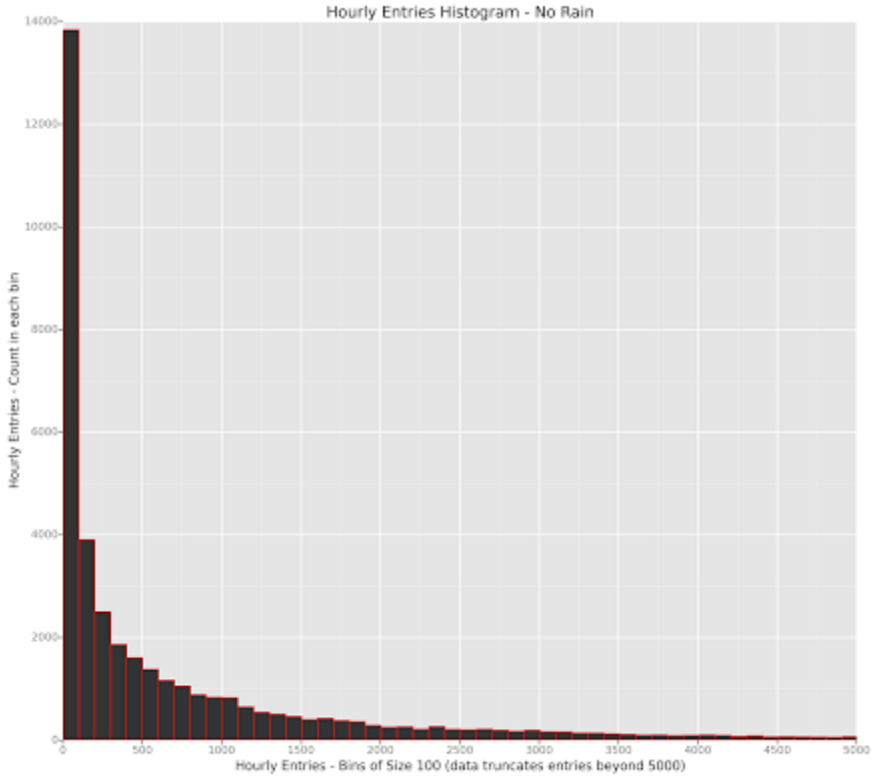
- Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.



This chart shows rider counts of hourly Entries put into Size 100 bins. The shifts in the proportional size comparing the no rain bars to the rain bars provide the interesting insight. Starting at the left where the entries are fewer, the ratio is more than double no rain compared to rain. As you shift to the higher value bins starting around 500 the ratio lessens. This visual comparison shows that the higher quantity bins have proportionally more rain. This generally shows that higher volume ridership times tend to be proportionally more rainy.



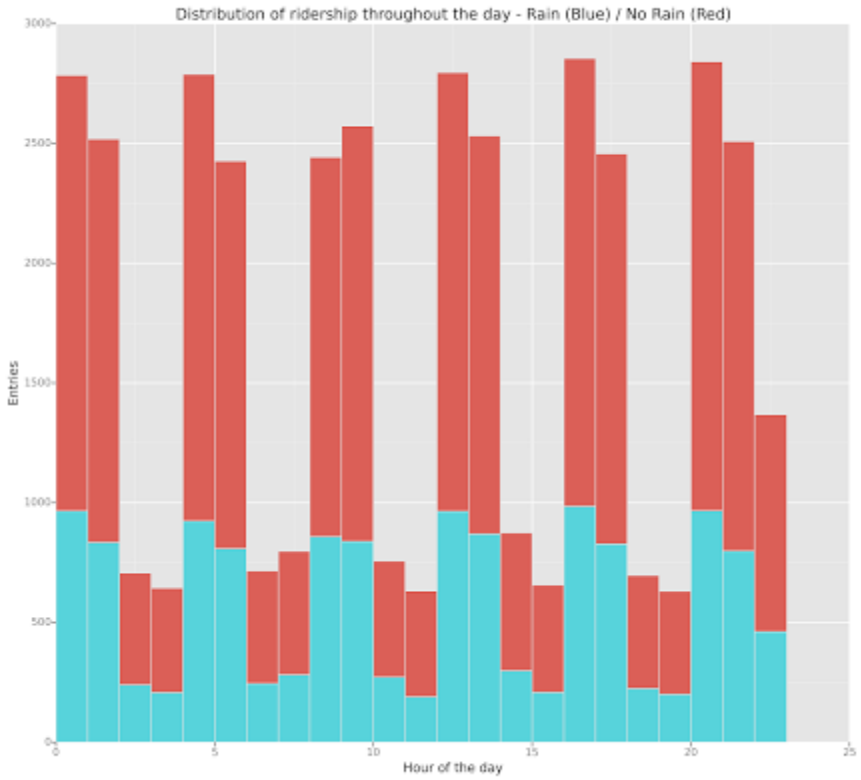
This histogram shows hourly Entries into the NYC subway separated into bins of 100. This reflects the same data as the first chart while allowing us to focus specifically on rainy time periods.



This histogram shows hourly Entries into the NYC subway separated into bins of 100. This reflects the same data as the first chart while allowing us to focus specifically on non-rainy or dry time periods.

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:



This scatter diagram shows ridership for each hour of the day for both rainy and non-rainy time periods. Both rainy and non-rainy ridership is clearly affected by the time of day. Clearly the specific hour relates to ridership in a non-linear fashion.

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.

I conclude that more people ride the NYC subway when it's raining than when it's not for this dataset. This is supported by several analyses. The Mann Whitney test showing that the mean difference between the two data sets is not likely to be due to random variation. The likelihood that the mean difference was random shown by the p value is only 2.499%. Given the ridership during rain showed a higher mean `1105.446` compared with the ridership without rain of

1090.279. In addition, the Linear regression model showed a theta or coefficient figure of 23.8486, which is a significant enough positive number to show that rain does increase ridership.

While it's clear that in this data set that rainy time periods have more ridership than other time periods, the difference is not large and the conclusion is not strong. The difference between the two means is 15.167 per hour or only 1.39% more ridership when it rains. Additionally, the Linear Regression theta coefficient of 23.8486 while positive is not large relative to some of the other coeficients. Temperature, fog, wind speed, and hour of day all have a greater influence on ridership according to the Linear Regression model with larger magnitude coefficients. Their influence is roughly 2.3 to 7.5 times more impactful.

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.

While the data analysis shows that the mean ridership during rainy time periods is more than non-rainy time periods, we should not assert that in general rain days have more ridership for the NYC subway. The time period associated with our particular dataset is only one month. Relative to the lifetime of the NYC subway, this is an extremely limited example. In addition, there could be other anomalies with the data that are influencing the results. For example, the ten rainy days during this thirty-one day time period include eight week days and Memorial Day. I would anticipate increased weekday ridership. This disproportionate mix of data could influence ridership.

The analysis methods also contain potential error areas. The linear regression model has limitations. The underlying assumption is that the variable relationship is linear, which is not confirmed when there is a R-Squared value of .555. A visual plot or histogram of the residuals shows a distribution that is not random or bell-shaped, which indicates that the model structure is not sound and requires additional variables to achieve a better model. The model did not include what intuitively appropriate variables such as week days versus weekend or holidays. Those that commute can inherently relate to the idea that there would be more riders during commute times than non-commute times. Other relevant variables could also be missing. While the linear regression model predicts that rain increases ridership, the coefficient of determination is .555, which is not close to 1, which could lead us to believe that our model leaves out many important variables that impact ridership. Our statistical models strongly agree that the means differ on this dataset; however, those results are dependent on the dataset used. While our dataset is almost 132K strong, the time breadth of the data is not large. This limits our ability to rely on the statistical test. While our analysis of the dataset leads us to conclude that ridership

increases when it rains, a broader look at the breadth of the data and uncertainties introduced via the statistical tests lead us to want to expand our data analysis.

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

The requirement for accurate predictions increases as we use the data to commit fiscal resources. This may lead us to expand our dataset and variables. We may want to expand our methodologies and toolsets to simulations or a hybrid. The early indicators from investigating the impact of rain on ridership lead us to believe that ridership increases with rain. This may become suitable for a map reduce application if we wanted to investigate larger questions that depend on the ridership under different conditions. Clearly if we were concerned about fiscal investment based on our models, we would expand our investigation to other variables and methods.