Edward Farrar
edward.farrar@ge.com
Due: 4 May 2015
Udacity Data Analyst Nanodegree

Short Answer Questions

Analyzing the NYC Subway Dataset

Intro to Data Science

Questions

Section 0. References

Dell, Inc. (2015) Statsoft Textbook [Multiple Regression: Residual Variance and R-Squared]. Retrieved from http://www.statsoft.com/Textbook/Multiple-Regression#cresidual.

ggplot from yhat. (2014) Docs [API Reference for syntax and graphing methods]. Retrieved from http://ggplot.yhathq.com/docs/index.html.

GitHub, Inc. [US]. (2015) possible fix for #376 [ggplot.py/setup.cfg code fix to display inline images with legends properly in IPython Notebook]. Retrieved from https://github.com/yhat/ggplot/commit/19e68e824782fda84e056baf8fcc88a8c361d9e3.

IP[y]: IPython Interactive Computing. (2015) The IPython notebook [Documentation]. Retrieved from http://ipython.org/ipython-doc/stable/notebook/index.html.

Lander, Jared P. *R for Everyone.* Upper Saddle River, NJ: Addison-Wesley, 2014.

Lutz, Mark. *Python Pocket Reference, Fifth Edition.* Sebastopol, CA: O'Reilly Media, Inc., 2014.

SciPy.org. (2015) Numpy and Scipy Documentation [API Reference for statistical method syntax]. Retrieved from http://docs.scipy.org/doc/.

Pandas 0.13.1 documentation. (2014) pandas: powerful Python datra analysis toolkit [API Reference for DataFrame, read_csv, to_csv, fillna, Computational tool, et al. syntax]. Retrieved from http://pandas.pydata.org/pandas-docs/version/0.13.1/.

Python Software Foundation. (2014) pandasql 0.6.2 [Documentation]. Retrieved from https://pypi.python.org/pypi/pandasql.

Python Software Foundation. (2015) python [Documentation for general syntax]. Retrieved from https://www.python.org/doc/.

SQLite. (2015) SQL As Understood by SQLite [Language Reference]. Retrieved from http://www.sqlite.org/lang.html.

StatsModels. (2013) StatsModels: Statistics in Pythom [Documentation for general syntax]. Retrieved from http://statsmodels.sourceforge.net/.

Wikibooks. (2015) LaTeX [Documentation for general syntax]. Retrieved from http://en.wikibooks.org/wiki/LaTeX.

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 choose to use the Mann-Whitney U Test to analyze the data. I used a two-tailed P value (doubling the result returned from `scipy.stats.mannwhitneyu`. My null hypothesis was that the population (number of riders) on rainy days and the population (number of riders) on non-rainy days are the same. The alternative hypothesis was that the two populations are not the same. I used a p-critical value <= 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.
Given that the distribution of ENTRIESn_hourly data are not normal, the Welch's T-test was inappropriate. The Mann-Whitney U Test is more appropriate for non-equal, non-normal sample sizes.

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]
Using the Original Data Set (turnstile_data_master_with_weather.csv) downloaded from https://www.dropbox.com/s/meyki2wl9xfa7yk/turnstile_data_master_with_weather.csv, I calculated the following values (see P1- Analyzing the NYC Subway Dataset - Statistical Test.pdf for the code):

```
With Rain Sample Size: 44104
Without Rain Sample Size: 87847

With Rain ENTRIESn_hourly Mean: 1105.44637675
Without Rain ENTRIESn_hourly Mean: 1090.27878015

The Mann-Whitney statistic = 1924409167.0
One-tailed p = 0.0193096344138
Two-tailed p = 0.0386192688276
Reject the null Hypothesis
```

1.4 What is the significance and interpretation of these results?
Based on the Mann-Whitney U Test results, the null hypothesis the population (number of riders) on rainy days and the population (number of riders) on non-rainy days are the same is rejected. I doubled the p-value return from the Mann-Whitney U Test in order to use a two-tailed p-value. The test showed there is a significant statistical difference between the two populations.

[1] P1- Analyzing the NYC Subway Dataset - Statistical Test.ipynb

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?
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?
I decided to experiment to find the best combination of features.
After much experimentation, I ended up using 5 features: rain, Hour, fog, mintempi and the UNIT dummy variables. I tested with and without dummy variables for the UNIT. Adding dummy variables improved the model.

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.
I began by testing just rain as the feature. The initial R^2 value was 0.061. I then proceeded to add various variables until I discovered a combination with the highest R^2 value:

Features	**R^2**
Rain	0.061
Rain, Hour	0.198
Rain, UNIT (dummies)	0.418
Rain, Fog, UNIT	0.418
Rain, Hour, UNIT	0.458
Rain, Hour, Mean DewPoint, UNIT	0.458
Rain, Hour, Mean Pressure, UNIT	0.458
Rain, Hour, Mean Wind Speed, UNIT	0.458
Rain, Hour, Min Temp, UNIT	0.458
Rain, Hour, Max Temp, UNIT	0.458
Rain, Hour, Precipitation, UNIT	0.458
Rain, Hour, Fog, Min Temp, UNIT	0.459
Rain, Hour, EXITs, UNIT	0.621
Rain, Hour, EXITs, Max Temp, Min Temp, Fog, UNIT	0.622
Rain, Hour, EXITs	0.637
Hour, EXITs	0.636

I initially thought that weather conditions that are easily perceived such as rain, fog, temperature, or precipitation would have the largest impact on the prediction. However, after experimentation, the results were a bit surprising. Many of the features had little impact individually. A combination of several features lead to a R^2 value of 0.459. Adding EXITn_hourly and dropping the UNIT dummy variables caused the R^2 value to jump as high as 0.637; however, if we are using this model to predict future riders, then EXIT data would be unknown. Therefore, as a result of the experimentation, I choose rain, Hour, fog, mintempi and the UNIT dummy variables as features.

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

Feature	Coefficient
rain	-26.9686
Hour	67.3968
fog	117.2483
mintempi	-11.3658

2.5 What is your model's R2 (coefficients of determination) value?
The R^2 value is 0.459.[2]

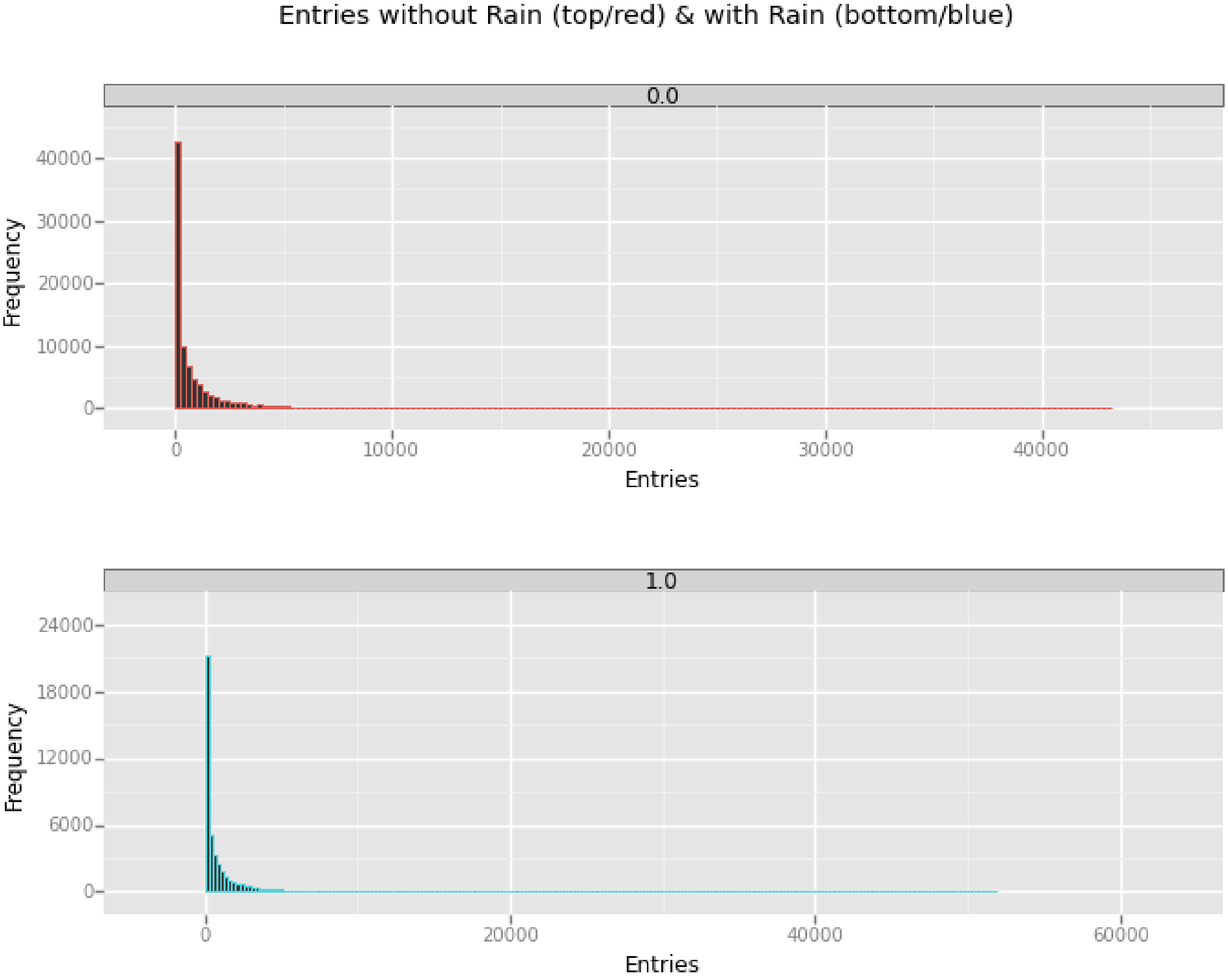
2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?
R^2, also known as the Coefficient of determination, is 1 minus the ratio of variability. This means that when the variability is small, the predictions from the model are good. At the extremes, no relationship between the actual and predicted results would mean that the variability is 1 thus making the R^2 value 0 (meaning the model can't predict any actual value). The other extreme is when there is perfect relationship in which the variability would be 0 thus making the R2 value 1 (meaning that the model perfectly predicts every actual value). In our case, the R2 value is 0.459 which means that the model has accounted for approximately 46% of the variability in the data and we are left with about 54% variability. This indicates the model is not very good at predicting actual values. Additionally, a residuals plot shows a cyclical pattern which also indicates a non-linear model may be more appropriate.

As a result, I do not think the OLS model is sufficient for predicting ridership for this dataset.

[2] P1- Analyzing the NYC Subway Dataset - Statistical Test.ipynb

Section 3. Visualization

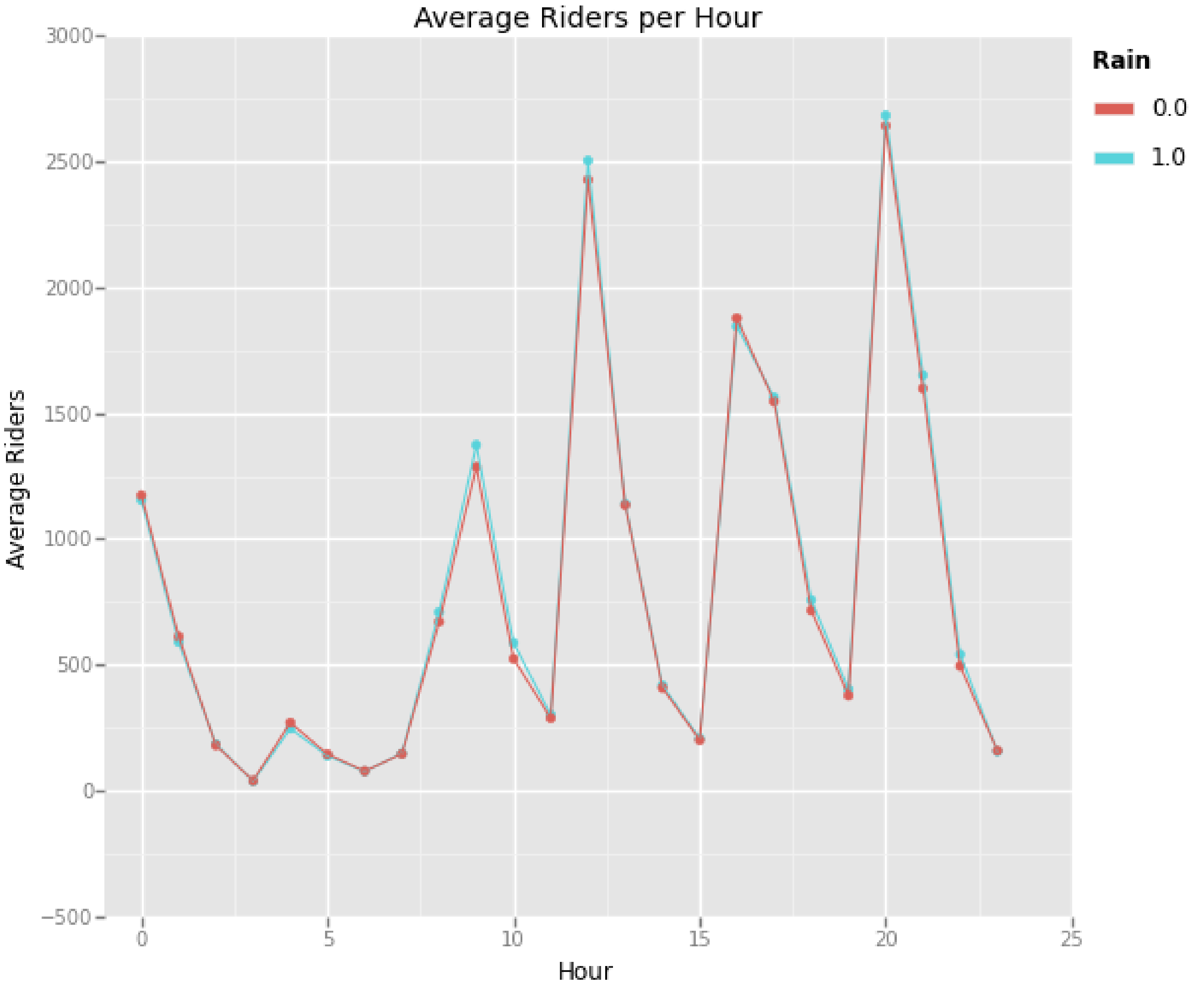
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.[3]



The primary insight from the histograms is that the data is not normal. As a result, statistical tests such as Welch's T-test, which assumes normalized data, cannot be used. A more appropriate test is the Mann-Whitney U Test and was used for this project.

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

[3] P1- Analyzing the NYC Subway Dataset - Visualization.ipynb



By averaging the ENTRIESn_hourly data and grouping it by Hour and rain, we can clearly see that the average ridership is slightly greater when it's raining over nearly all hours of the day. This supports the results of the statistical test that showed there was statistically more riders on the subway when it was raining.

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?
Based on my analysis, I conclude that more people ride the subway when it is raining. Using the Mann-Whitney U Test, I was able to determine with a p-value of 0.0387 that there is a statistically higher probability of people riding the subway when it's raining versus when it is not raining. Additionally, the test is supported by the fact that the mean number of riders is slightly higher on rainy days (1105) versus non-rainy days (1090). Further, the visualization, Average Riders per Hour (shown above), shows that on rainy days, the average number of riders is slightly higher than on days without rain. The plot reinforces the statistical conclusion that more people ride the subway when it's 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 used several analyses to come to this conclusion. The first challenge was to determine if the ENTRIESn_hourly data was normal or not. Unfortunately the Shapiro-Wilk test is unreliable for sample populations over 5000. The sample population for rainy and non-rainy days was 87847 and 44104, respectively. So, I decided to create a rough histogram of the data. It was immediately apparent that the data was non-normal based on the histogram of the subdivided data. This led me to use the Mann-Whitney U Test. The null hypothesis was that the ridership (ENTRIESn_hourly) were the same on rainy and non-rainy days. Since the Mann-Whitney U Test is used to test whether a distribution (dataset) is more likely to generate a higher value than the other. In other words, what is the likelihood that picking a number from the rainy day set will be higher than picking a number from the non-rainy day set. The test gave a p-value of 0.387. Any value under 0.05 (p-critical) indicates that the null hypothesis (that the number of riders are the same on rainy and non-rainy days) was to be rejected. Therefore, there is a significant probability that more people ride on rainy days.

The linear regression also confirmed that rain impacted the ability to predict ridership. While the two factors that influences ridership the most were the hour of the day and the number of exits from the subway, these two features alone resulted in a lower R2 value (0.636) than when rain was introduced to the model (0.637). While slight, this does indicate that rain can be used to better predict subway ridership than when rain is ignored.

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.

There are several shortcomings to the dataset. First, the data only covers a one month period of one year (May 2011) and there were 20 non-rainy days and 10 rainy days. Expanding the dataset to include multiple months and years may have lead to better analysis. However, care should be taken not to include too much data. For instance, the question was, “does rain affect ridership?”, so including winter months may not be appropriate since it rarely rains during the winter months in New York City. Another shortcoming of the dataset was the inclusion of thunder information. Thunder was not recorded for any day, so it was useless as a feature for the linear regression model. I attempted to use the “improved” dataset, but the Mann-Whitney U Test returned 'nan' for the p-value. I was unable to determine why and eventually used the standard dataset. In particular, the condition field in the improved data would have been nice to include in the linear regression model as a potential feature.

As for the statistical test and linear regression model, I thought the Mann-Whitney U Test was appropriate. I would have liked to have been able to use the Shapiro-Wilk Test to test whether or not the data was normal, but I was unsure if sampling the dataset down to 5000 entries was appropriate or acceptable in order to use the test. In the end, I visually determined the data was non-normal using a histogram. The linear regression model, OLS, was the most interesting part of the project; however, I was hoping the model would have produced a much better R^2 value. My intuition on what features would impact the model were completely wrong. Fortunately, I experimented enough to determine the best combination of relevant features to get the R^2 value as high as possible (0.637). I feel more data (quantity and features) would have allowed for a better model. Additionally, plotting the residuals showed a very cyclical pattern which indicates that a non-linear regression may be more suitable.

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

When I plotted the Average Riders per Hour, I was focused on whether the rainy day average would show up as higher than the non-rainy day averages across the recorded hours. When it did, I was relieved because I felt the statistics and linear regression were at least on the right track. When I looked a second time, a second bit of information emerged. I noticed that the higher averages were not during the expected rush hour periods (7-9AM and 4-6PM), but rather at the noon and 8PM hours. Not knowing the culture of New Yorkers very well, I find it interesting that the typical lunch and after dinner hours have more subway riders. If the sociological makeup of the ridership was known, this could reveal interesting facts about lower income people (those unlikely to have other transportation in NYC such as taxis or a personal car) and their work patterns (i.e. shift workers coming and going to work, etc.). Overall, the dataset was sufficient to teach the concepts of Intro to Data Science, but are not sufficient to drawn “real-world” conclusions given the limited data available.