Project – 1: Analyzing the NYC Subway Dataset

Section 0. References

1. Studied I Heart Stats from edx – Online
2. Meaning From Data – Professor Michel Starbird - Video
3. Codeacademy – Python online learning
4. Learn Python Hardway - Videos
5. LEARNING_PYTHON_FOR_DATA_ANALYSIS_AND_VISUALIZATION – Udemy - Bought the course

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?

The Mann-Whitney U-Test was used to compare the means of the two groups we were looking at: rainy days and non-rainy days.

Answering is looking for any significant difference whether positive or negative (ridership could be significantly higher or significantly lower on rainy days compared to non-rainy days) even though common intuition might say rainy days should have higher ridership (as people don't like walking in the rain).

A two-tail test was used; as the question being addressed is looking for any significant difference whether positive or negative (ridership could be significantly higher or significantly lower on rainy days compared to non-rainy days).

Null hypothesis - The two populations are the same, or simply put, that rain has no correlation with ridership

P critical value used was 0.05, or 5%

P value - 0.049999825586979442

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.

This particular test was used because the variances of two distributions are unequal (they have unequal standard deviations therefore unequal variances as well since standard deviation is just the square root of variance). The Mann-Whitney test is also appropriate in this case because the underlying distributions are not normally distributed (both the rainy and non-rainy distributions have a long tailed versus a bell curved shape) and it relaxes this assumption of normality as compared with other statistical tests such as Welch's t-test.

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.

Mean of the rainy days = 1105.4463767458733

Mean of the Non rainy days = 1090.278780151855

U = 1924409167.0

p = 0.049999825586979442

1.4 What is the significance and interpretation of these results?

From Mann-Whitney test, the result it seems that there is a statistically difference for the number of entries days with rain and when there is no rain as the p-value < 0.05 and the mean of the rainy days is higher than the non-rainy days. So we can expect an increased in ridership on rainy days. Other words, we are rejecting the Null Hypothesis

Section 2. Linear Regression

2.1 What approach did you use to compute the coefficients theta and produce prediction in your regression model:

Gradient Descent (GD) (as implemented in exercise 3.5)

OLS using Statsmodels

Or something different?

Both gradient descent (GD) and OLS models where used to run linear regression on the NYC subway dataset. Both models look for linear relationships between the features and the predicted values or NYC subway rides.

Gradient Decent was used to compute the coefficients of theta.

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

The following features were used as input variables:

-presence of rain

-precipitation levels

-mean temperature for the day

-weekday vs. weekend

-A set of dummy variables was also used to represent the data collection unit, and the time of the day.

2.3 Why are these features appropriate?

I have used the ['rain', 'precipi', 'Hour', 'meantempi']] as the core, because People using subway is more in the weekdays than the weekends. I also want to know if people are using more in rainy day or non-rainy day in what hours.
Others, like fog, precipi, etc…, also categorized as an extra feature as they are not heavily affecting the no of people using subway in those conditions.

The details of the different feature values are in attached in the programs.

After mixing and matching various features, these were the most relevant and important features based on their explicatory power and statistical significance. I had a bias for choosing the simplest model possible, without losing too much explicatory power or R^2.

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

```
Theta is:

-->Rain Coefficient:
          2.34565484e+01

-->Precipitation Coefficient:
          -4.37411689e+01

-->Mean Temperature Coefficient:
          -1.43751340e+02

-->Weekday Coefficient:
          4.41929749e+02

HOUR COEFFICIENTS:
          00: -8.50843807e+01
          04: -5.79076785e+01
          08:  2.43815118e+01
          12:  3.10171825e+00
          16:  2.88001619e+01
          20:  1.88589194e+03
```

-Unit coefficients are omitted here, because there are too many

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

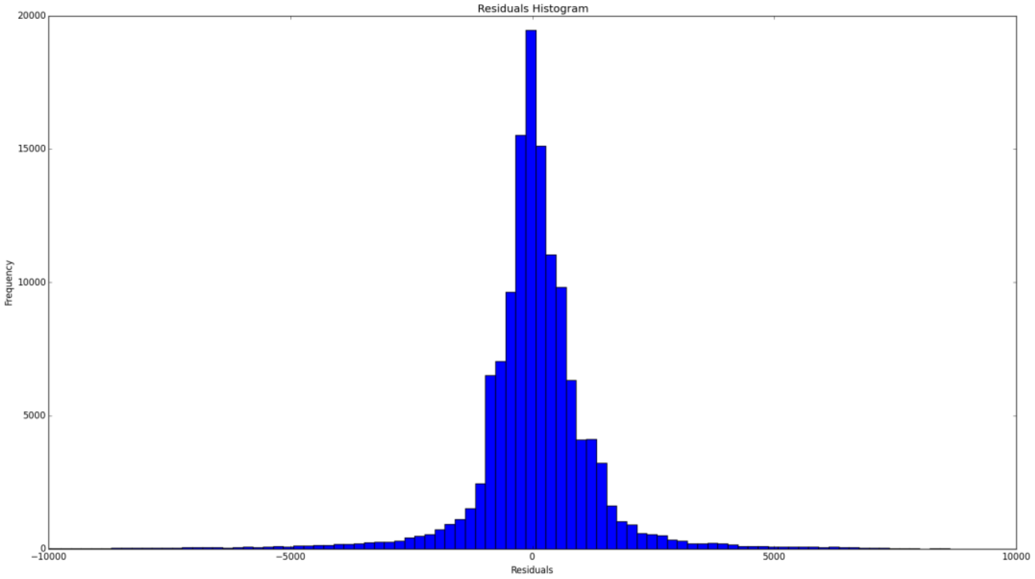
OLS Model r^2 value is 0.47924770782

Gradient descent GD

GD Model r^2 value is 0.501415456139

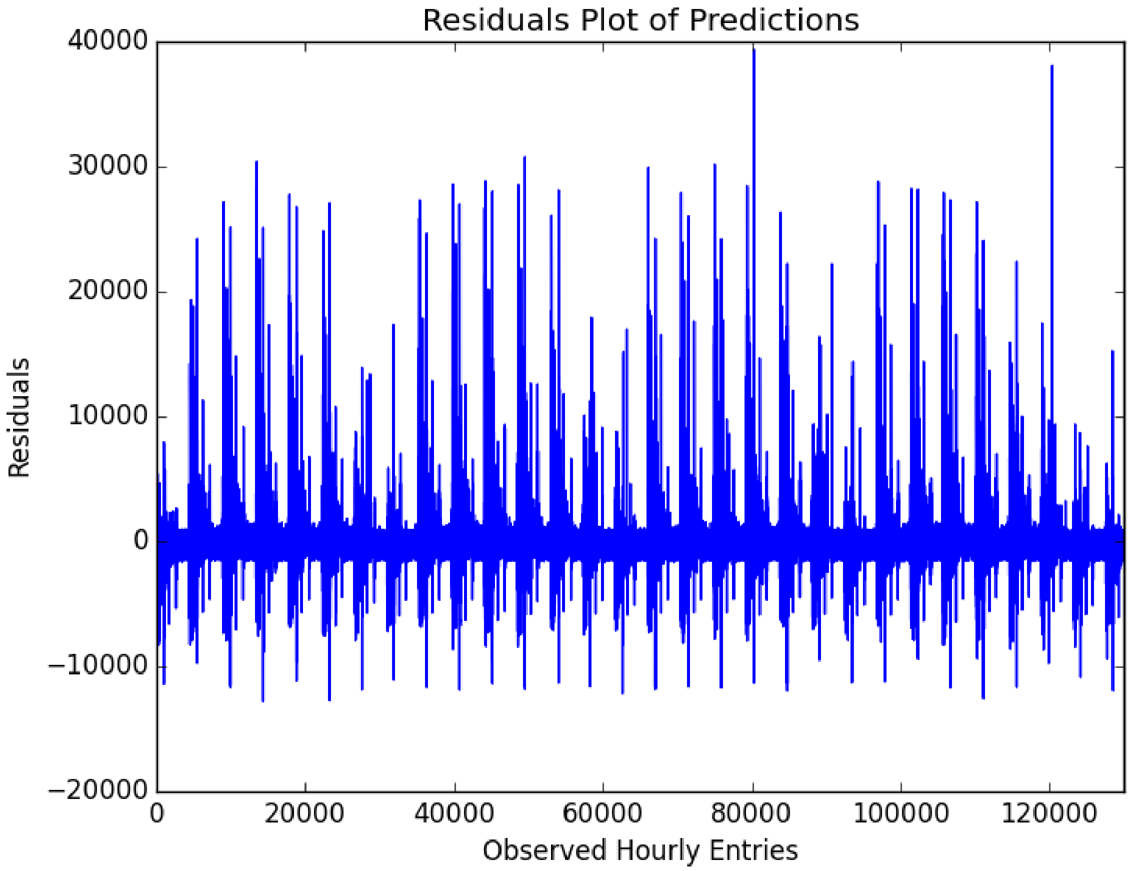
2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

R2 is the percentage of variance that is explained by the model. Higher is generally better, and this model's R2, over .50, is higher than 0.4. However, when determining goodness of fit, it is important to consider residuals as well.



This residuals histogram indicates that there rather long tails with several extremely large values. A very long tailed residuals histogram is an indicator that the linear model is not a great fit. As a result, I do not believe that this model fits the data well enough and I believe a different type of model might be a better choice for this data.

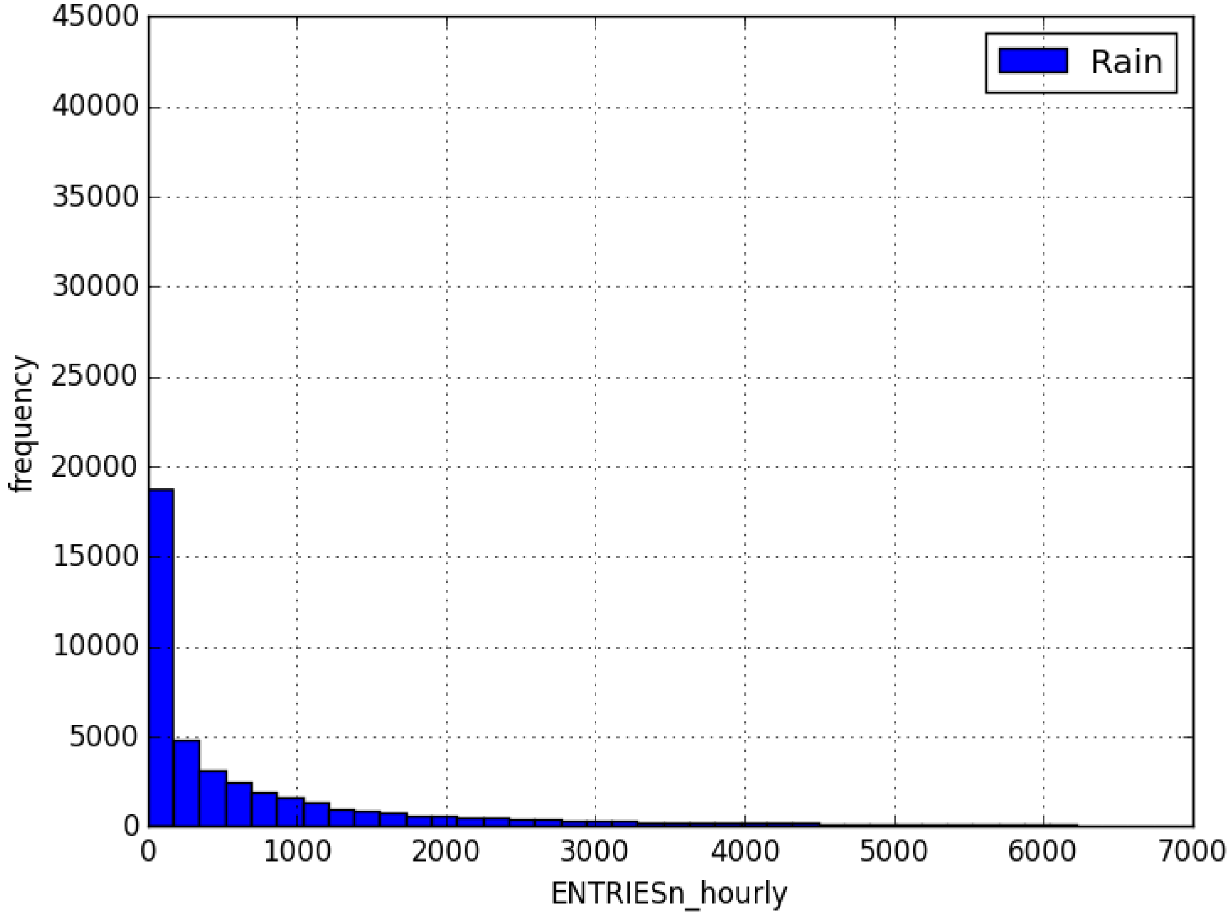
A residual plot would help determine whether the model is biased in an obvious way. In particular if there is a discernable pattern in the residuals plot, the model is biased.

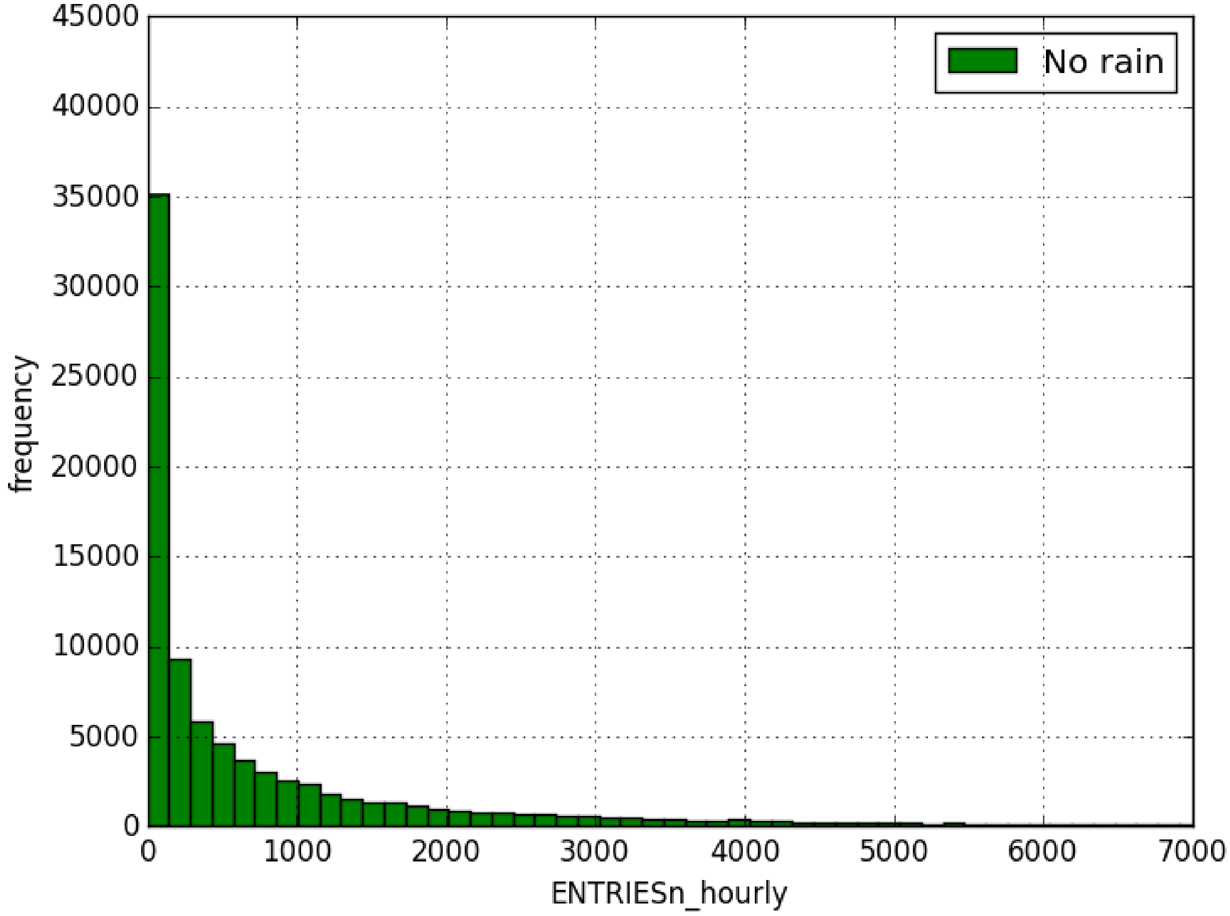


This plot plainly demonstrates that the linear model is missing underlying patterns in the data. Therefore, the goodness of fit for a linear model to this data is insufficient and we should use a different kind of model.

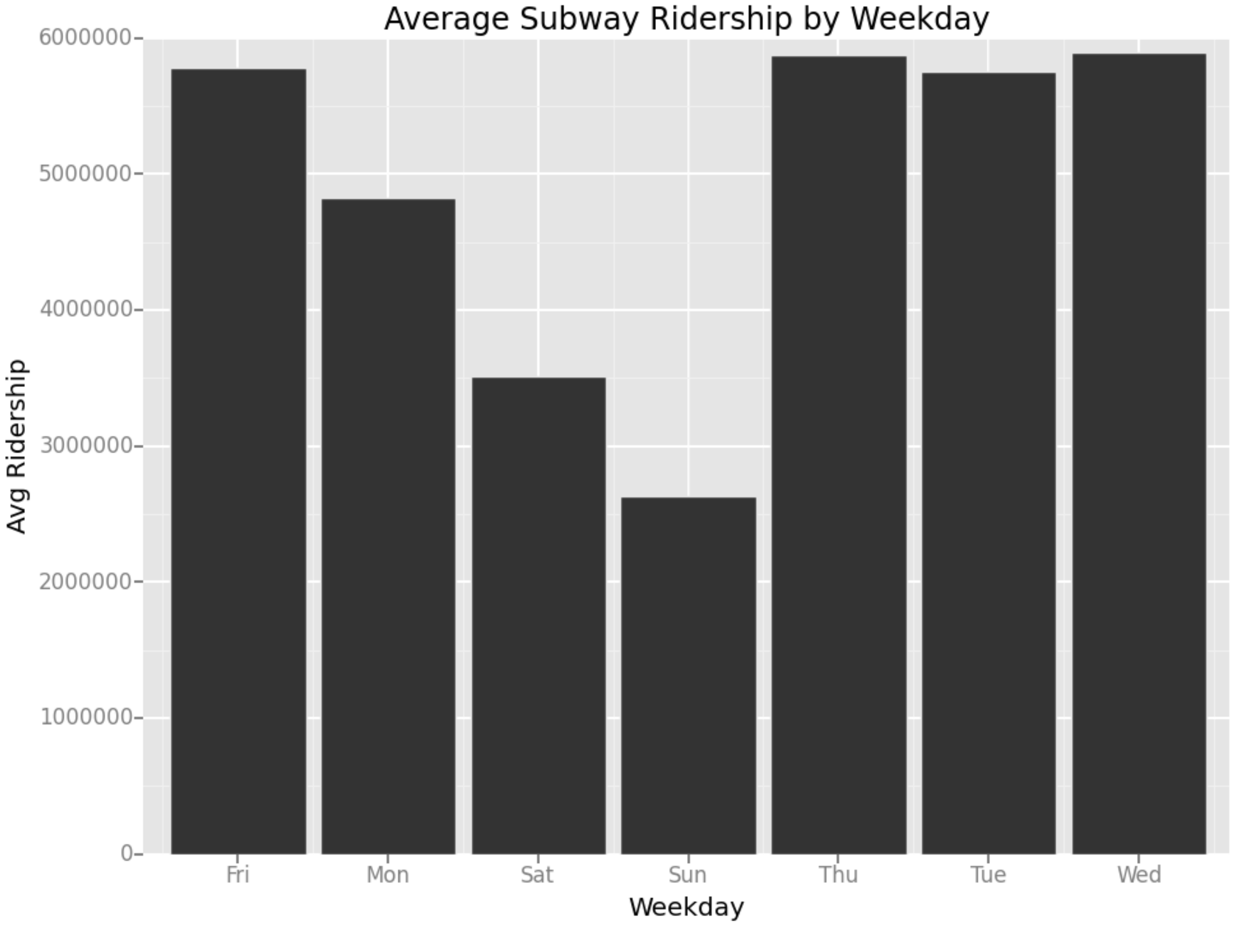
Section 3. Visualization

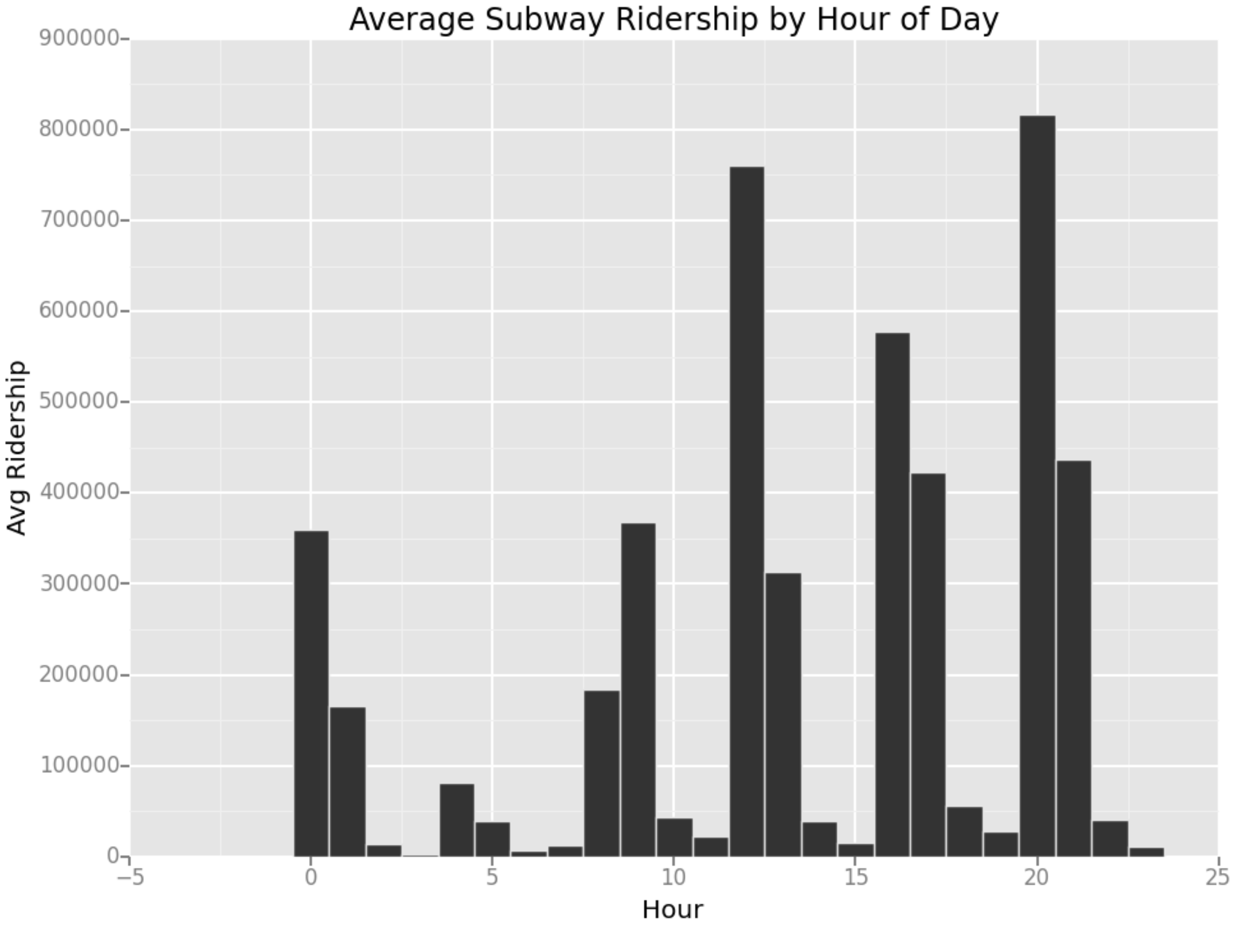
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

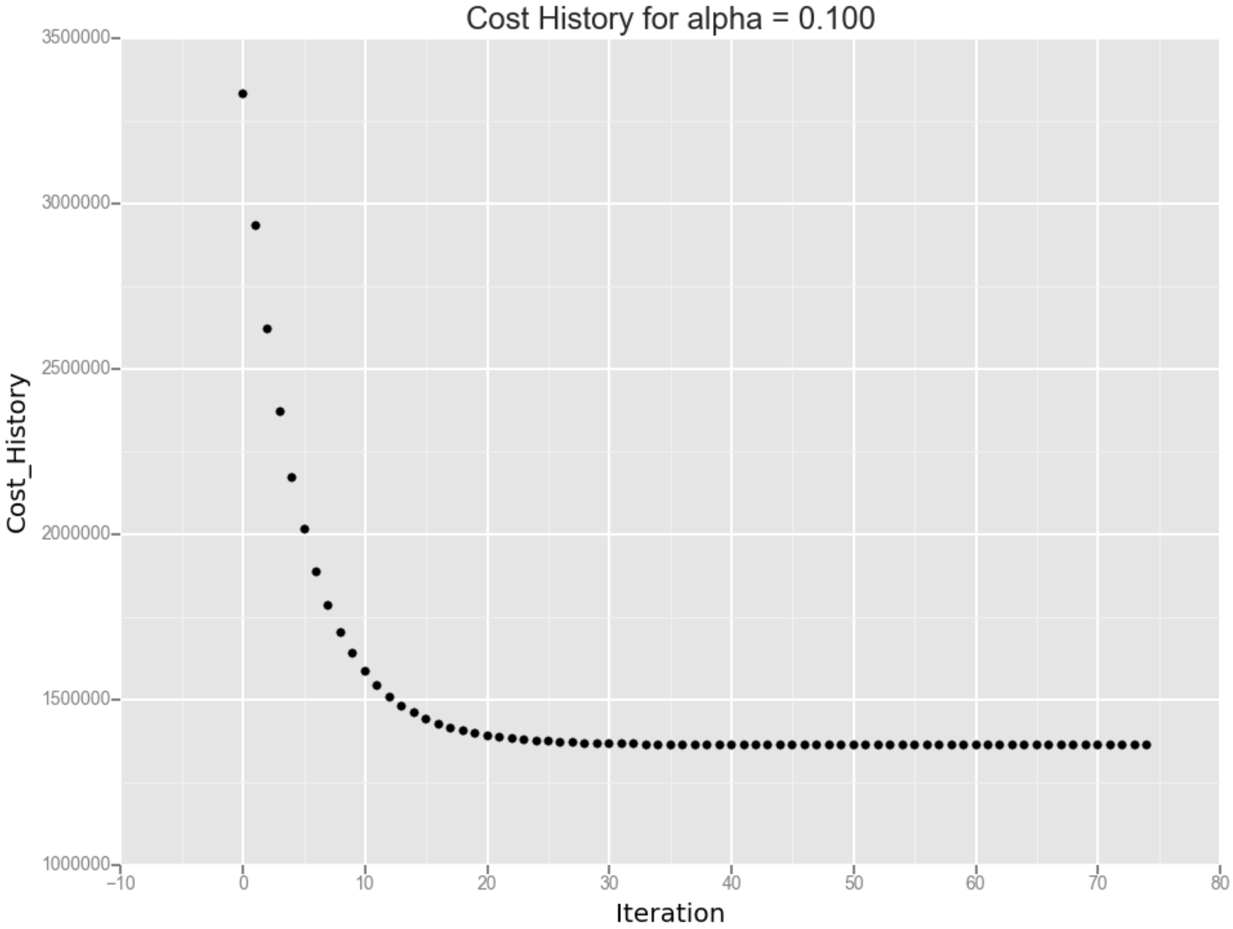



No rain
frequency
ENTRIESn_hourly
0
5000
10000
15000
20000
25000
30000
35000
40000
45000
0
1000
2000
3000
4000
5000
6000
7000

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




Average Subway Ridership by Hour of Day
Avg Ridership
900000
800000
700000
600000
500000
400000
300000
200000
100000
0
−5
0
5
10
15
20
25
Hour



The above bar chart shows the average hourly ridership by day of week. The sum of ENTRIESn_hourly by day of week was divided by the count of rows for a given day of week (as each row represents an hour's worth of data). The bar chart shows that the average hourly ridership is higher on weekdays than weekends, with Saturday seeing significantly higher ridership than Sunday. It appears that the average hourly ridership on Monday is significantly different than the rest of the weekdays. This may be due to a seasonal effect of Monday holidays. The given data set is a sample from May 2011. There is at least one major holiday that falls on Monday in the month of May: Memorial Day.

Section 4. Conclusion

4.1 Do more people ride the NYC subway when it is raining or when it is not raining?

Based on this analysis, I believe that more people ride the NYC subway when it is raining.

4.2 Rationale

As we have seen the numerical results of the Mann-Whitney U-Test (Section 1.3) the mean of ENTRIESn_hourly is greater for hours with rain than without (1,105 vs. 1,090), approximately.

Additionally, (Based on Average Subway Ridership by Weekday and Average Subway Ridership by Hour of Day) it shows that the ENTRIESn_hourly sample with rain appears to be different population than the sample from without rain.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

a. Dataset,
b. Analysis, such as the linear regression model or statistical test.

1. The data set provided contains only one month of MTA data.
2. This smaller data set is subject to effects of seasonality, as the time of year may also affect ridership.
3. We are not considering different seasons of the weather, like winter, summer and spring. If we are not collecting data, for the rainy days, in any of the above other seasons, before concluding the hypothesis.
4. Additionally, the Month of May contains a Monday holiday, which appears to have an effect on the data in the visualizations in Section 3.2 and Section 4.
5. For example, if it rained at any point in a given day, every hour of that day will reflect that it rained. This prevents a truly granular analysis of how the weather can affect ridership within a day.
6. Current linear model is good for the currently available data set. But if we need to do the ANOVA Test, we may need to have more details like how many men and woman travelled in a rainy days Vs Non-rainy Days for peak and non-peak hours, station wise, etc…

One immediate big concern pops up while exploring the data was that there were markedly more entries than there were exits. Some of the logical explanations could be that there were miscounts, or some turnstiles/stations were not included in the data set. Presumably, this would have had an equivalent effect on both rain and no-rain data sets, so for the purposes of this study, it likely had little to no effect.

A combination of increased sample size (larger data set) and normalization by location/turnstile ID could have potentially increased the confidence of both the Mann-Whitney U test and the linear regression model. As we saw from examining the 'UNIT' column, ridership varied greatly. Simply put, some stations and turnstiles were naturally more active than others. The Mann-Whitney U test did not take this into account, and only looked at the subway entry distributions for rain and no-rain. Examining how the same stations at the same day and time varied by rain could have increased the fidelity of the test.

The linear regression model was adequate for the purpose of the study, but could certainly have been improved. It's possible that the region of study had a linear relationship, but it is still an assumption and simplification. Considering the extreme, subway ridership certainly has an asymptotic limit; only so many riders can get on the subways! As mentioned in Section 2.6, the inclusion of more features or polynomial combinations could have increased the accuracy of the model. Given more data, it would have also been appropriate to split the data into a training data set (~70%), a cross-validation data set (~20%), and a testing set (~10%). This could have illuminated any errors with high variance, high bias, and any over/under-fitting.

P.S: I would like to move towards to data science, even though I am working database side for many years. I am not very familiar with the Statistics part and I am still learning from different sources, as I mentioned in section 0. Please guide me, if this report has issues or anything to be corrected.