Analyzing the NYC Subway Dataset

Udacity Nanodegree Project 1

Coursework:

- Intro to Data Science

By: Joe Nyzio

Analyzing the NYC Dataset

Udacity Nanodegree

By Joe Nyzio

visit: http://nbviewer.ipython.org/gist/JoeNyzio/fbcfbe777248c092b900 for the project section code on ipython notebook viewer.

Section 1 Statistical Test.

1.1 Which statistical test did you use to analyse 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?

Statistical Test

I analyzed the NYC subway data using a python implementation of a Mann Whitney U test.

Null Hypothesis

Ho = The difference between NYC subway ridership on rainy and non rainy days is likely due to random variation within the sample sets. They are equal.

Alternative Hypothesis

H1 = The difference between NYC subway ridership on rainy and non rainy days is not due to random variation within the sample sets. They are not equal.

P-Critical

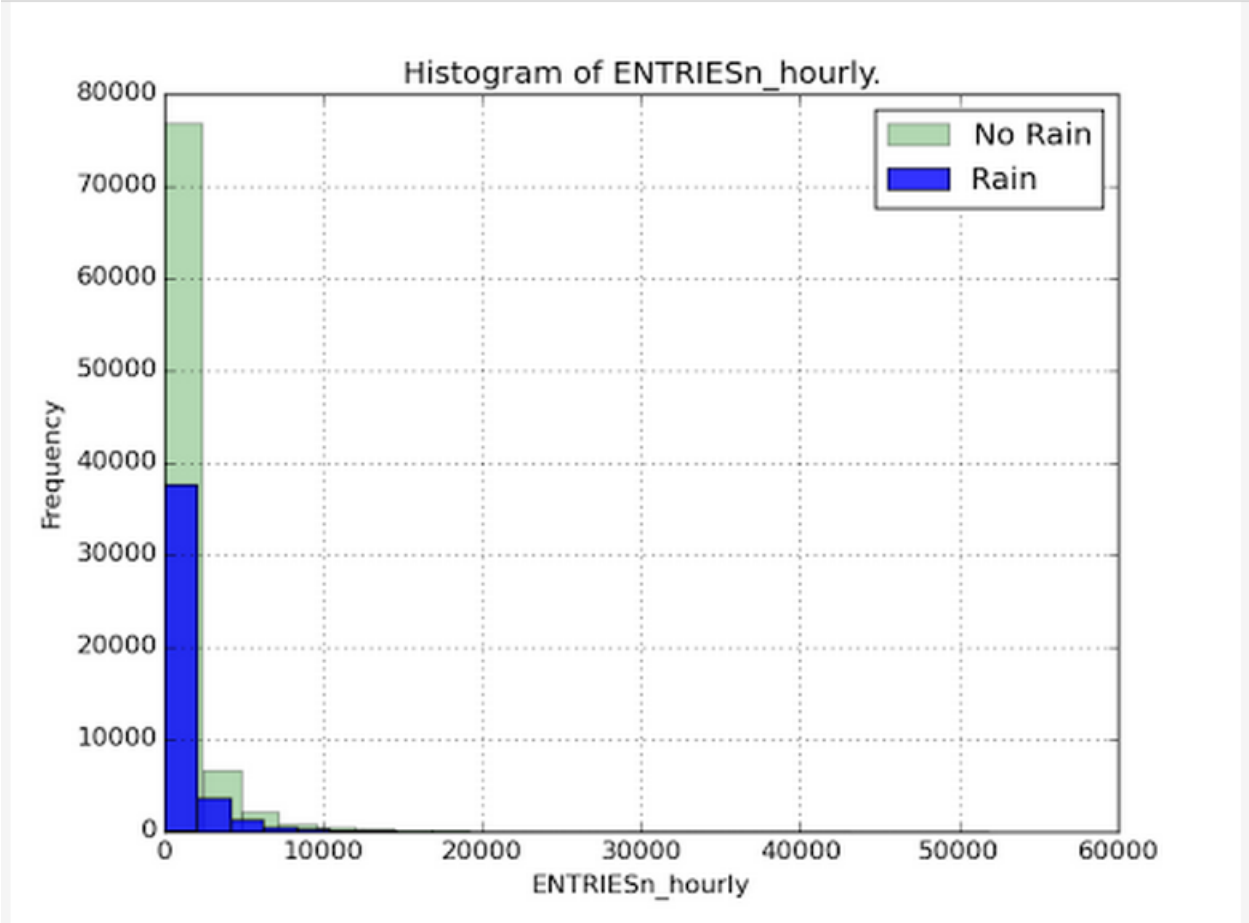
P-critical = .05.

One tail or two tail?

Test type = One tailed t test.

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 distribution below shows that we have a non normal distribution so it will be useful to use a test that does not make any assumptions of the distribution. The Mann-Whitney U is a non parametric test that makes no assumptions of a 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.

```
Good job! Your calculations are correct.
Here's your output:
(1105.4463767458733, 1090.278780151855, 1924409167.0, 0.024999912793489721)
```

with_rain_mean = 1105.446
without_rain_mean = 1090.279
U = 1924409167.0
p = .0249

1.4 What is the significance and interpretation of these results?

The first 2 outputs are representative mean values of the ENTRIESn_hourly on days that have been grouped as either being days with_rain or without_rain.

The Mann Whitney U statistic is 1924409167.0 and has been used to calculate the p-value.

We need to determine if the p-value allows us to accept or reject the null hypothesis. My p-critical was set to .05 and our p-value is .0249.

(P-value = .0249) < (P-critical = .05)

Reject Null. Accept the alternative hypothesis.

The Mann Whitney U test calculated a p-value extreme enough to reject the null and accept the alternative hypothesis. This tells us that the difference between these data sets are statistically significant. There is a significant difference in subway ridership between rainy and non rainy days.

with_rain_mean > without_rain_mean

Now we know that more people are likely to ride the subway when it is raining.

My conclusion

There is a statistically significant increase in NYC subway ridership on rainy days.

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:

a - I used Gradient descent as implemented in exercise 3.5.

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

My final model included 'Hour' as a feature. 'UNIT', and 'DATEn' were also included as dummy variables and joined to my features list. Alpha was set to .12 over 100 iterations. The following is a numerical summary of how I came to this conclusion.

I started with a base model that included only 'UNIT' and 'rain' as features.

```
Your r^2 value is 0.425122760487
```

I began to add features and recorded the ones that increased r^2.

UNIT', 'rain', 'DATEn'

```
Your r^2 value is 0.440176747695
```

'UNIT', 'rain', 'DATEn', 'Hour'

```
Your r^2 value is 0.478296205124
```

'UNIT', 'rain', 'DATEn', 'Hour', 'meantempi'

```
Your r^2 value is 0.478296228616
```

To check if 'rain' was providing a significant contribution to the model I removed it to see the effect it had on r^2.

UNIT, DATEn, Hour, meantempi

```
Your r^2 value is 0.478296228482
```

r^2 with ‘rain’ - r^2 without ‘rain’ =

0.478296205124 - 0.478296228482 = .000000000134

‘Rain’ as a feature provides a .000000000134 increase in the predictive value of r^2.

‘Rain’ does not seem to be relevant to the success of this model.

Other features I tried were similarly irrelevant. I decided to only keep the features that lead to significant improvement in the accuracy of the model. In the end these features were ‘UNIT’, ‘DATEn’ and ‘Hour’.

‘UNIT’, ‘DATEn’ and ‘Hour’

```
Your r^2 value is 0.47829620519
```

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 started off putting features in randomly to test their relation to the r^2 value. I didn’t like being so dependent on the computer for providing insight so I tried rebuilding the question in a way I could understand outside of numbers and calculations. I decided to approach the model as a behavioral crowd psychology problem. If I could determine which features were most likely to cause predictable patterns in behavior I should be able to more accurately predict when they would ride the subway. This idea didn’t work so I decided to build the model from scratch using only features that lead to significant improvements.

Everything I added to the model provided about the same amount of improvement. I ended up with features I felt I could justify really well that actually weren’t providing any more use than any of the other features. I decided to strip down the model to its essentials and try a lightweight approach.

Only ‘UNIT’, ‘DATEn’, and ‘Hour’ improved the model by any more than the 10th decimal place. I couldn’t justify a situation where providing such little improvement would be worth it so I left the features out.

FINAL FEATURES

‘UNIT’, ‘DATEn’, ‘Hour’

2.4 What are the coefficients for the non-dummy features in your linear regression model?

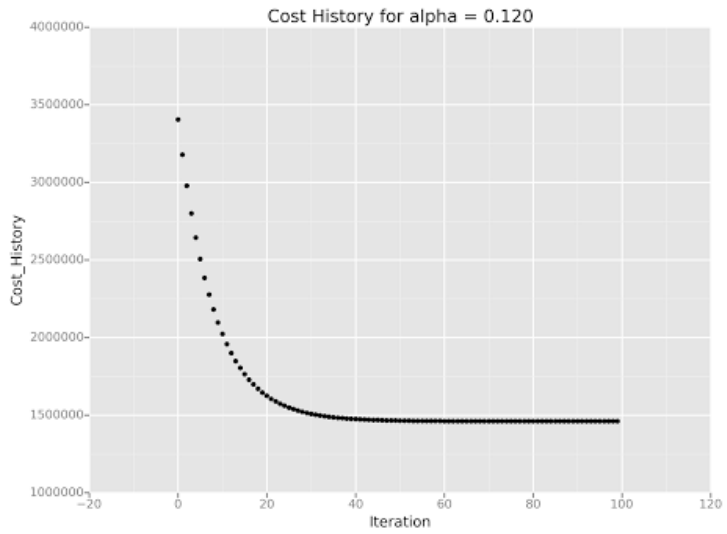
The coefficients for non-dummy features in my linear regression model are listed as the following array.

```
Out[12]: (array([ 2822.48966799,  3090.80803219,  3359.12639638, ...,   516.11677998,
                   516.11677998,   516.11677998]), <ggplot: (278625061)>)
```

2.5 What is your model’s R^2 value?

‘UNIT’, ‘DATEn’, ‘Hour’

```
Your r^2 value is 0.47829620519
The plot of your cost history is shown below.
```

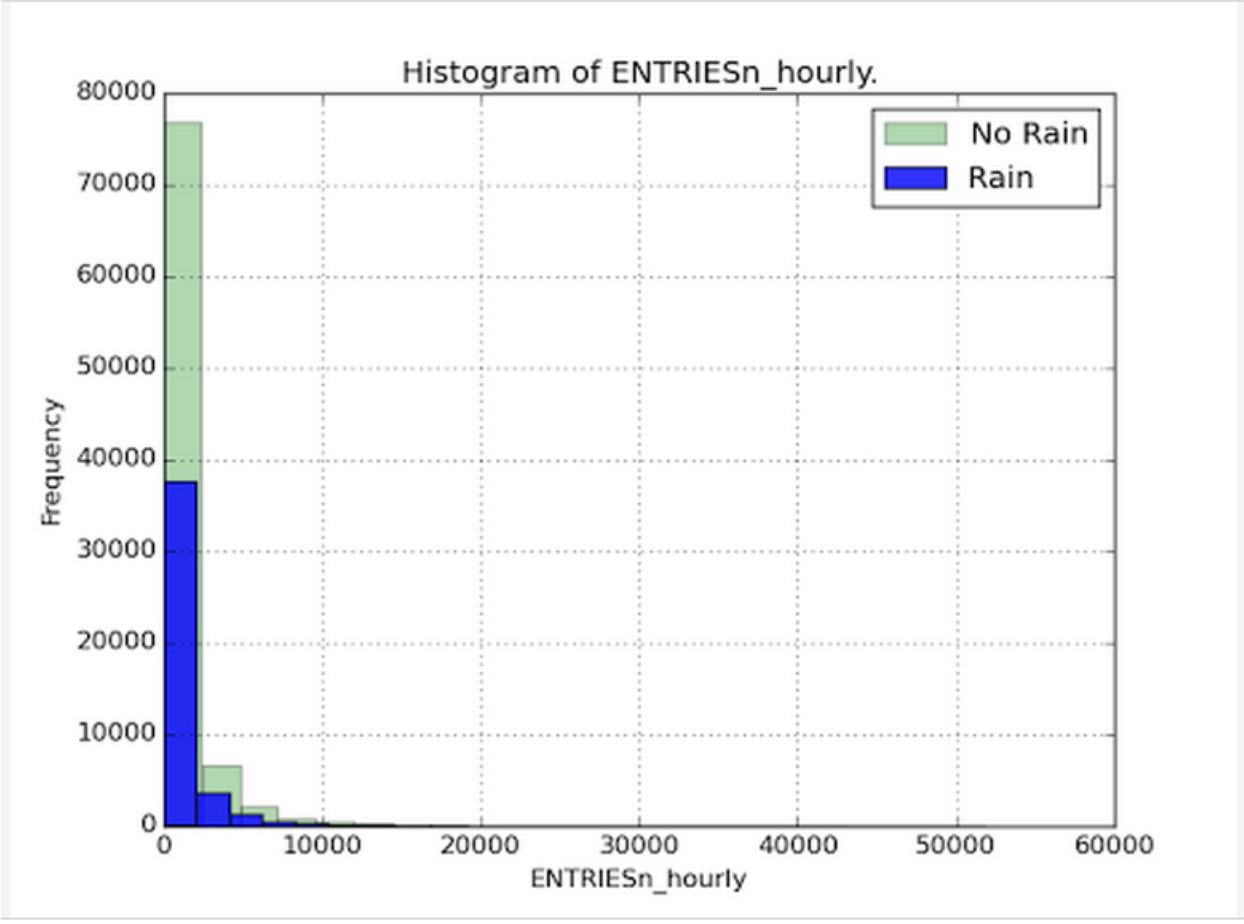


2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

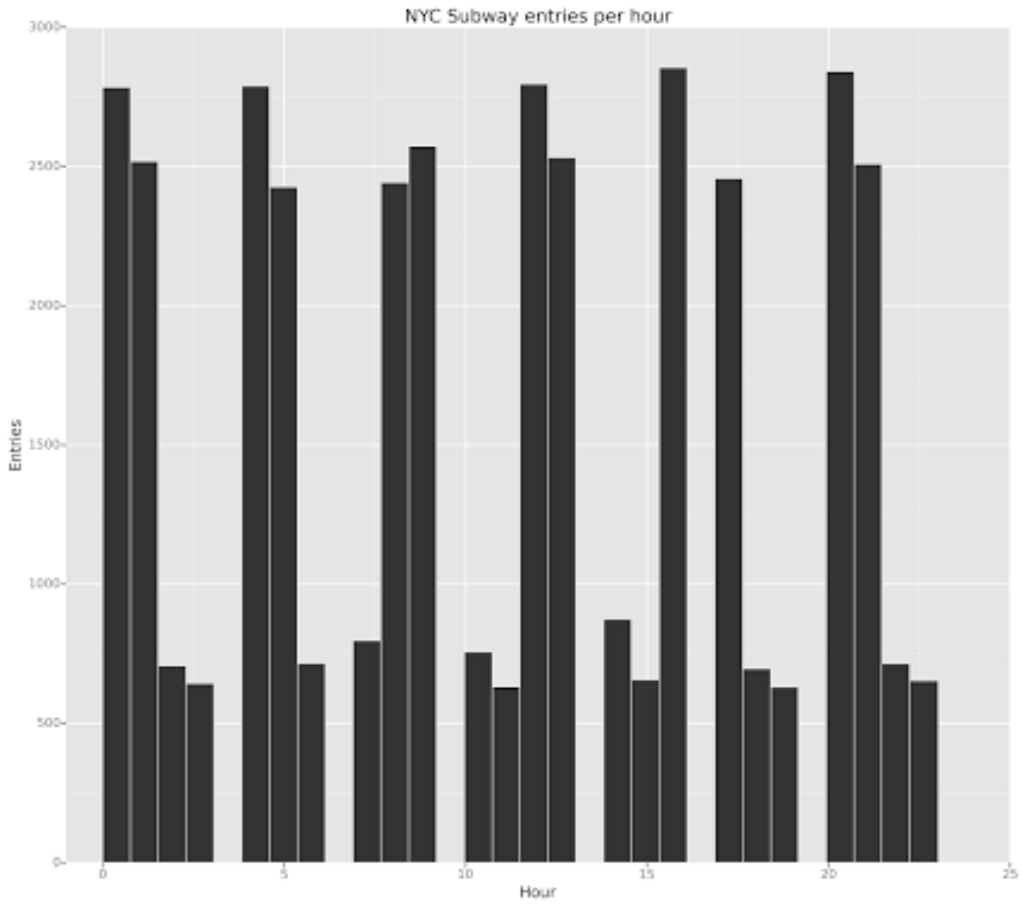
This r^2 means that my model explains about %47.8+ of the variability within the dataset. It would be nice to have a higher r^2 value but this isn't bad considering its trying to model human behaviors. I think this linear model and r^2 value is appropriate for predicting the dataset.

Section 3 Visualization

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



3.2 One visualization can be more freeform.



Histogram of the NYC subway entries per hour.

Section 4 Conclusion

My analysis concludes that more people ride the subway when it is raining but that rain is not a significant feature in forming a model to predict subway ridership.

Results from Mann Whitney U

```
Good job! Your calculations are correct.
Here's your output:
(1105.4463767458733, 1090.278780151855, 1924409167.0, 0.024999912793489721)
```

with_rain_mean = 1105.446
without_rain_mean = 1090.279
U = 1924409167.0
p = .0249

(P-value = .0249) < (P-critical = .05)

with_rain_mean - without_rain_mean = 15.167

Our Mann Whitney U test has calculated the (with_rain_mean, without_rain_mean, U, p) values. This tells us that 15 more people, on average, will ride the subway on rainy days as opposed to non rainy days. Our p-value of .0249 exceeds the p-critical value of .05. This lets us reject the null and accept the alternative hypothesis.

Ho = LOSER!

Ho = The difference between NYC subway ridership on rainy and non rainy days is likely due to random variation within the sample sets. They are equal.

H1 = Choose alternative

H1 = The difference between NYC subway ridership on rainy and non rainy days is not due to random variation within the sample sets. They are not equal.

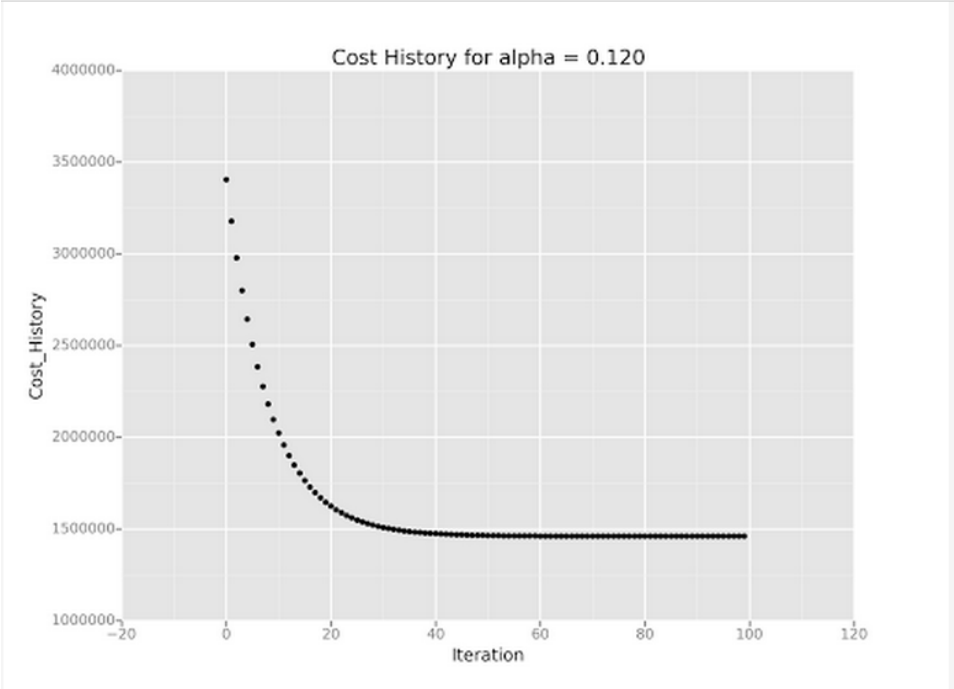
FINAL CONCLUSION

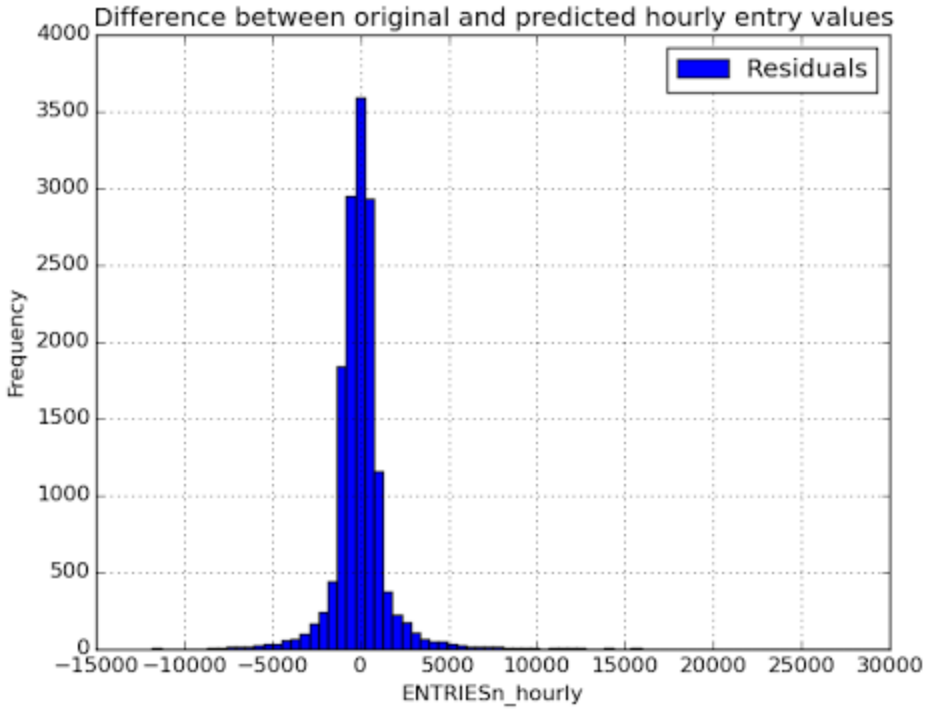
There are a significantly higher number of people that choose to ride the subway on rainy days in NYC.

Linear regression

My r^2 value was calculated using linear regression including ‘Hour’, as a feature, UNIT, DATEn, as dummy variables added to the features, and an alpha of .12 over 100 iterations.

```
Your r^2 value is 0.47829620519
The plot of your cost history is shown below.
```





I did not include ‘rain’ when computing r^2. Though the Mann Whitney U test showed significance, I have no reason to believe that it is a great fit for the predictive model. There was no significant improvement to the model when ‘rain’ or any other features were included.

FINAL MODEL

‘Hour’, ‘UNIT’, ‘DATEn’
alpha = .12
iterations = 100

Section 5 Reflection

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**

Dataset:

I think the dataset could benefit from a few extra features to help determine what would currently appear to be random variations in the ridership. We could include events like sports games or important events. We could include lists of scheduled maintenance or the closing of other types of transportation due to weather.

Analysis:

The linear regression model I came up with has very few features that are relevant to the primary purpose of the dataset. I only used ‘UNIT’, ‘DATEn’, and ‘Hour’ as features. This helps me make predictions about the subway ridership but has not brought me any closer to relating that subway ridership to the weather. I did not check to see the relationship between the variables I used. There may be multicollinearity within the features causing redundancy in the model.