Analyzing the NYC Subway Dataset

by Swaroop Oggu in fulfillment of Udacity's Data Analyst Nanodegree, Project 1

Section 0. References

Stackoverflow.com

Pandas/scikit/scipy documentation

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?**

Mann-Whitney U-statistic test was used to analyze the NYC subway data. Two tail p value was used because assumption is that differences between datasets can go in both the directions i.e up or down and is not concluded or hypothesized that one data set will be larger. Null hypothesis is that both the data sets / samples are same and there is no significant difference. P-Critical value used was 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.**

Mann-Whitney U-statistic test was applicable in this case as the samples/data is not normally distributed, independent and random. The assumption the test making about the distribution of ridership in the two samples is that rain doesn't impact the ridership

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 With Rain	Mean Without Rain	pvalue	U-statistic
1105.4463767458733	1090.278780151855	0.024999912793489721*2 approximately 0.049 (Since Mann-whitney test returns a one sided p value it should be multiplied by 2 in this case as we are applying two tailed test)	1924409167.0

1.4 **What is the significance and interpretation of these results?**

The approximate two tailed p value 0.049 which satisfies the p critical value test of <=0.05 and high U-statistic test signify that the null hypothesis, which assumed ridership is independent of rainy vs non rainy day and distribution of both the data sets are same, can be rejected with 95% confidence

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. **OLS using Statsmodels or Scikit Learn**
2. **Gradient descent using Scikit Learn**
3. **Or something different?**

1,OLS using Statsmodels was used to determine the coefficients of theta and produce prediction

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 'rain', 'precipi', 'Hour', 'meantempi' along with UNIT() as dummy variable were used in the liner

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.**

The following features 'rain', 'precipi', 'Hour', 'meantempi' are more appropriate to be evaluated as logically rain, mean temperature and precipitation are related.

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

```
rain          29.464529
precipi       28.726380
Hour          65.334565
meantempi    -10.531825
```

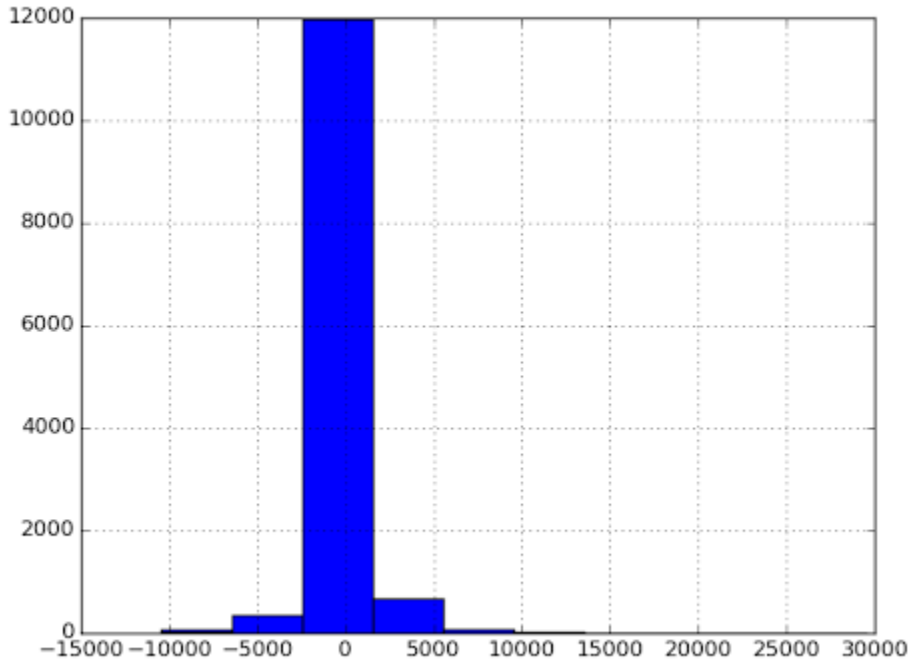
2.5 **What is your model's R2 (coefficients of determination) value?**

```
        R^2 Value result 0.47924770782
```

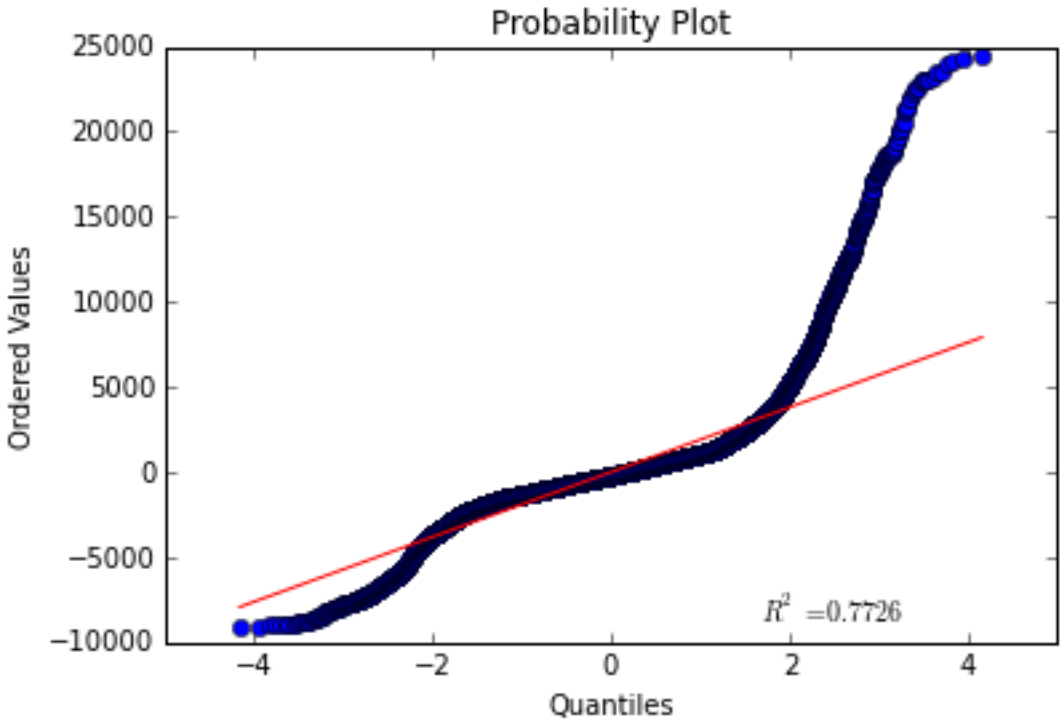
2.6 **What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?**

R2 value of 0.479 explains the variation by 47.9%. The goodness of fit can be well explained with residual plot where the original vs predicted diff is plotted. But also we should consider not over fitting the model which might lead to

huge diff between predicted and original values. Linear model might not be the best model, but it gives enough insight to know what to work on to improve.



To understand the long tails of the residual plot above. We need to figure out if the normal distribution is a better model for this data set, which can be achieved by a Normal Probability plot.

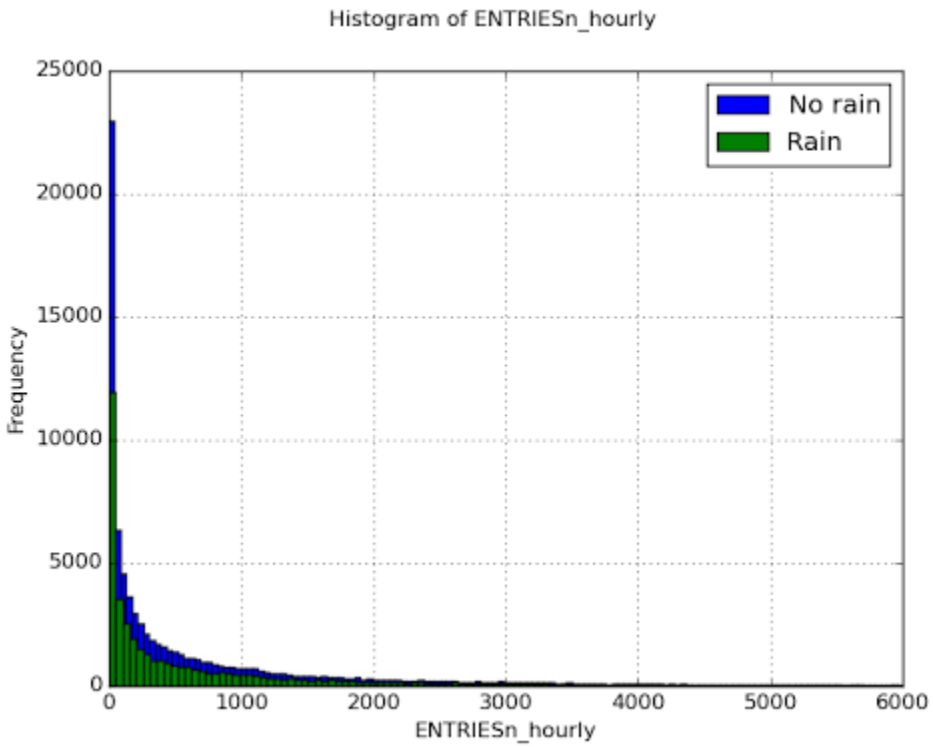


For data with long tails relative to the normal distribution, the non-linearity of the normal probability plot can show up in two ways. First, the middle of the data may show an S-like pattern and then for long tails, **the first few points show increasing departure from the fitted line *below* the line and last few points show increasing departure from the fitted line *above* the line**

From the above plot we reasonably conclude that the normal distribution can be improved upon as a model for the NYC rider data set (turnstile-weater).

Section 3. Visualization

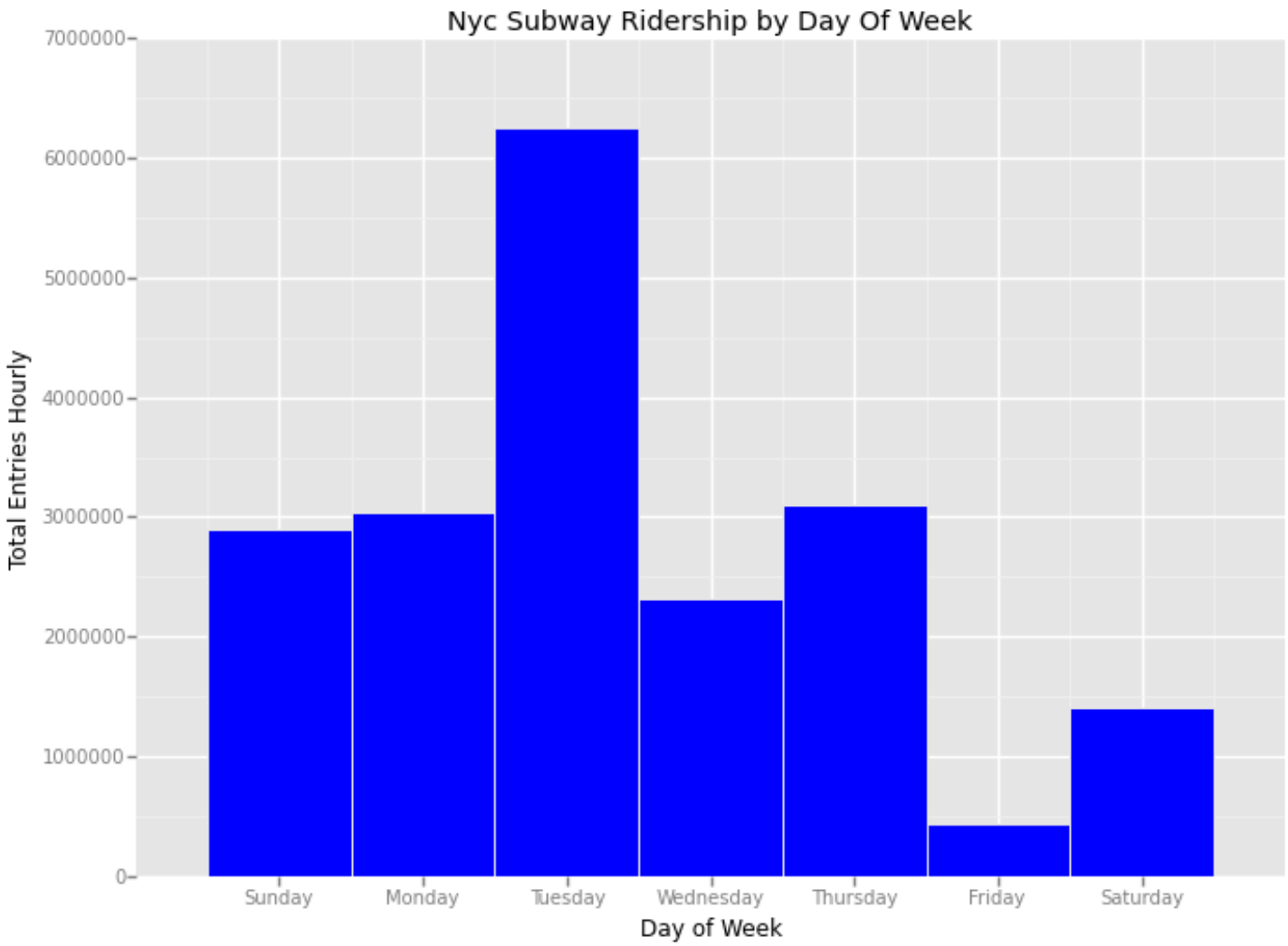
3.1 **One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.**



The above histogram shows that subway entries for rainy days vs non rainy days is not normally distributed. But this doesn't drive the conclusion that non rainy days has the highest ridership as the samples for rainy days are significantly less than the non rainy days

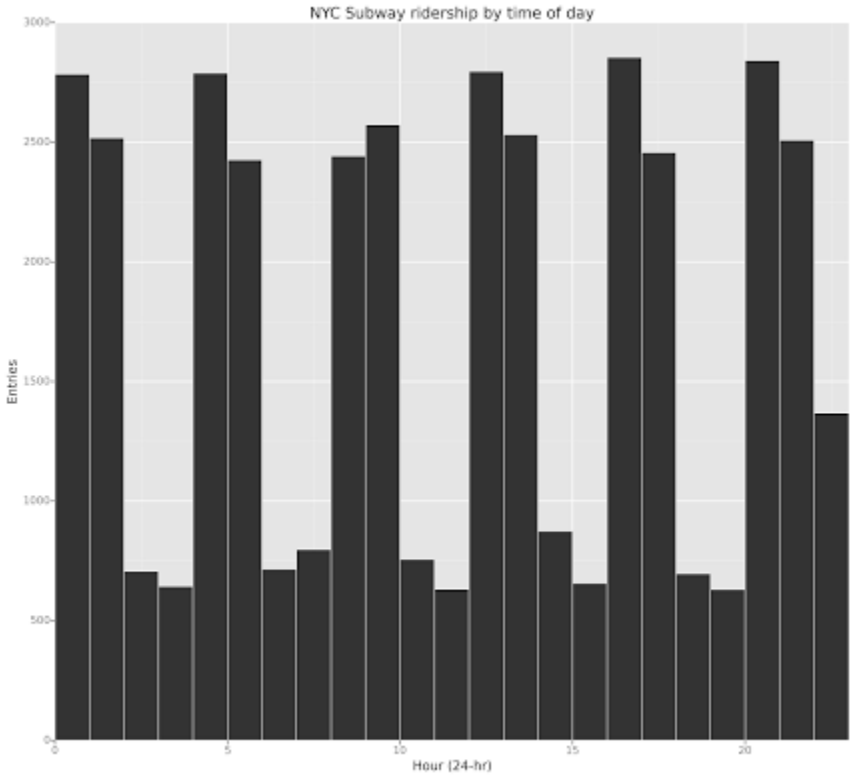
3.2 **One visualization can be more freeform you should feel free to implement something that we discussed in class or attempt to implement something more advanced**

Ridership by Day of Week



The figure clearly depicts that on **TUESDAYS** the ridership is high.

Ridership by Hour of Day



The figure clearly depicts that for every five hours there is surge in the ridership.

Section 4. Conclusion

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?**

Considering Mann-Whitney-U-Test p value results we can certainly predict that there will be a high ridership on NYC subways during rainy days. Even though the plots signify that ridership on non-rainy days is more which is due to less no of samples from rainy days we can count on the affirmative result of two data sets being statistically different from the Mann-Whitney test

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.**

Linear regression R^2 value result provided a variance of 47.9% which is also the coefficient of determination of the outcome by 47.9%. The positive weight of the rain feature for linear regression indicated that rain is an important feature that drives the ridership. Along with the linear regression results the statistical test Mann-Whitney-U-Test also quantitatively confirmed that two data sets are statistically different

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**.

Improved Data set with more sample data paired with normalization on additional features Eg UNIT would have resulted in a better confidence level to confirm the assertion of rainy day ridership. Also certain factors like Holidays, are not taken in to account, If it's raining and is not a working day, the ridership might not be a match to working day non rainy day ridership

It might not be possible to derive a linear relationship between most of the features in the data set. Would have been insightful to work on other regression/nonlinear models to see if there are any new findings

Statistical Test has enough information to lead us on to the model to be used

5.1 **Do you have any other insight about the dataset that you would like to share with us?**

Additional data sets for training and testing to get more accurate predictions.