Short Questions to Analyzing the NYC Subway Dataset

Analyzing the NYC Subway Dataset

Short Questions

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 U statistical test to analyze the original NYC subway data provided in the course. For this test, I used a two-tail P value to test if weather condition (sunny vs rainy) affects volume of ridership by either increasing or decreasing it.

The MannWhitney U test is a nonparametric test that can be used to test whether one distribution is more likely to generate a higher value than the other. In this case, I tested whether the distribution of rainy days generates a higher average (mean) ridership value than the distribution of non-rainy days using a p-critical value of 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.

The Mann-Whitney U test is applicable to the NYC Subway dataset because we make no assumptions about the distribution of each sample from the data. I found both datasets of rainy vs non-rainy hourly entries to be non-normal and wanted to determine if there was a statistical difference between the populations. According to the Wikipedia entry for the Mann-Whitney U test, the Mann-Whitney U test has greater efficiency than the t-test on non-normal distributions.

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.

The following results were obtained for the Mann-Whitney U test performed on two sample populations consisting of the number of hourly entries for rainy days vs non-rainy days. The sample for rainy days contained 44,104 records with a mean hourly entries found to be 1105.45 while the sample for non-rainy days contained 87,847 records with mean found to be 1090.28 for a difference of 15.17 between the two means.
The U statistic was 1924409167.0, with a one-tailed p-value of 0.0193 that was doubled to obtain the two-tailed p-value of 0.0386 which is less than the p-critical value of 0.050.

1.4 What is the significance and interpretation of these results?

Both the rainy and non-rainy populations were found to be non-normal with different mean hourly_entries values. Since the Mann-WhitneyU test resulted in a two-tailed p-value of 0.0386 less than the p-critical value of 0.050, we can reject the null hypothesis that the populations of rainy and non-rainy days are identical.

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 the gradient descent approach as implemented in the file "*Linear Regression_Subway data.py*" attached to this submission to perform a linear regression and compute the coefficients theta and produce prediction for the ENTRIESn_hourly field. I don't have time to explore OLS or something different right now.

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

My linear regression model used the following features from the data: *rain, meanwindspdi, Hour, maxtempi,* as well as dummy variables.

A quantity of 465 dummy variables was used by the prediction function. These dummy variables were created from the "Unit" variable automatically by the default code provided in exercise 3.5. The "Unit" variable represents the name of the remote unit that collects turnstile information. My previous submission incorrectly references a quantity of 552 dummy variables that was inferred from the last "Unit" named being "R552". However, I did not notice that there "Unit" names did not follow a consecutive numbering sequence.

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 tried using intuition to choose features and added 'thunder' but this caused the code to throw error messages so I decided to inspect 'thunder' and found it that it was equal to zero for every entry.

I then visually inspected the rest of the numeric variables in the dataset by plotting histograms and boxplots. I choose the features *rain, meanwindspdi, Hour*, and *maxtempi* because they seemed to show a lot of variation across the dataset. Intuitively

it makes sense that people may be inclined to ride the subway and stay off the streets during adverse weather conditions such as rainy and windy days. Lastly, plotting hours vs ridership seemed to show a strong relationship with what appears to be high ridership during what could be considered to be rush hour time frames.

I also ran the model using only rain + dummy features vs rain + *meanwindspdi, Hour, maxtempi* + the dummy features and compared the R^2 values. I found that the model with more features gave a better R^2 value of 0.458.

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

The coefficients, found by inspecting the final Theta array after running gradient descent, are tabulated below (rounded to two decimal points):

coefficient	Rain_coeff	meanwindspdi_coeff	Hour_coeff	maxtempi_coeff
value	2.03	55.71	464.50	-11.77

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

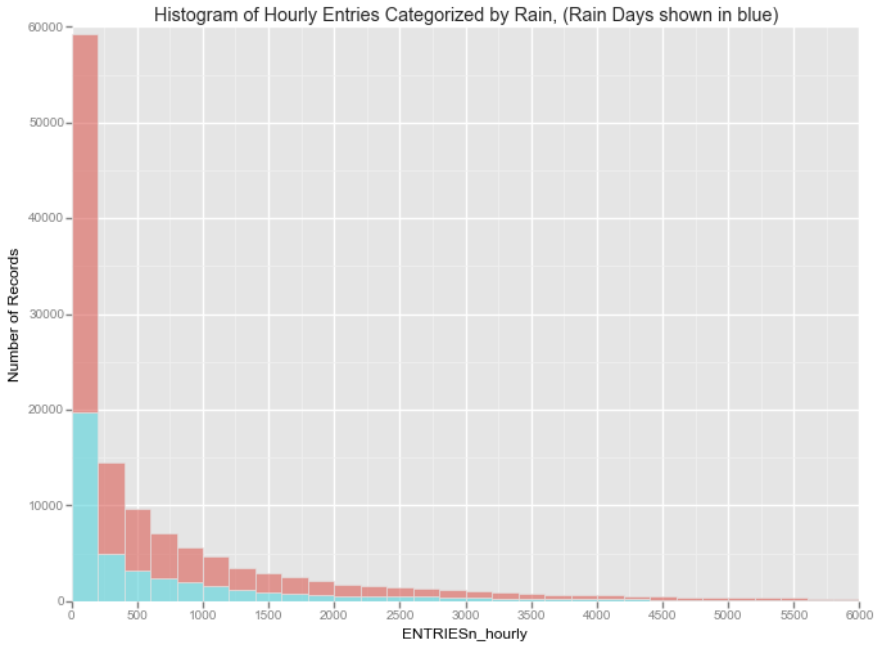
Using the features and coefficients above, the model's coefficient of determination R2 value is: 0.45822689882

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?

In general the R2 value gives a measure of how well the regression line represents the data, with an R2 = 1.0 meaning a perfect fit. The R2 value explains the percentage of the original variability between the model and the data.

Since my model gives an R2 = 0.46, I interpret this as the model being able to explain 46% of the variability between the model and the data, leaving 54% residual variability. It appears that this model would be moderately appropriate for predicting ridership as more than 40% of variability can be explained but it leaves room for improvement to seek out additional variables that have stronger relationship to ridership volume. A lower R2 value also indicates that the model predictions would have a higher error and be of lower precision than a model producing a higher R2 value.

Section 3. Visualization



Histogram used to compare differences in distribution between Rain and Non Rain data. The Rain data is shown in blue, Non-rain in red, with binwidth=200. The x-axis scale for ENTRIESn_hourly is truncated at 6,000 to improve visibility. The long tail extends to 51,800.

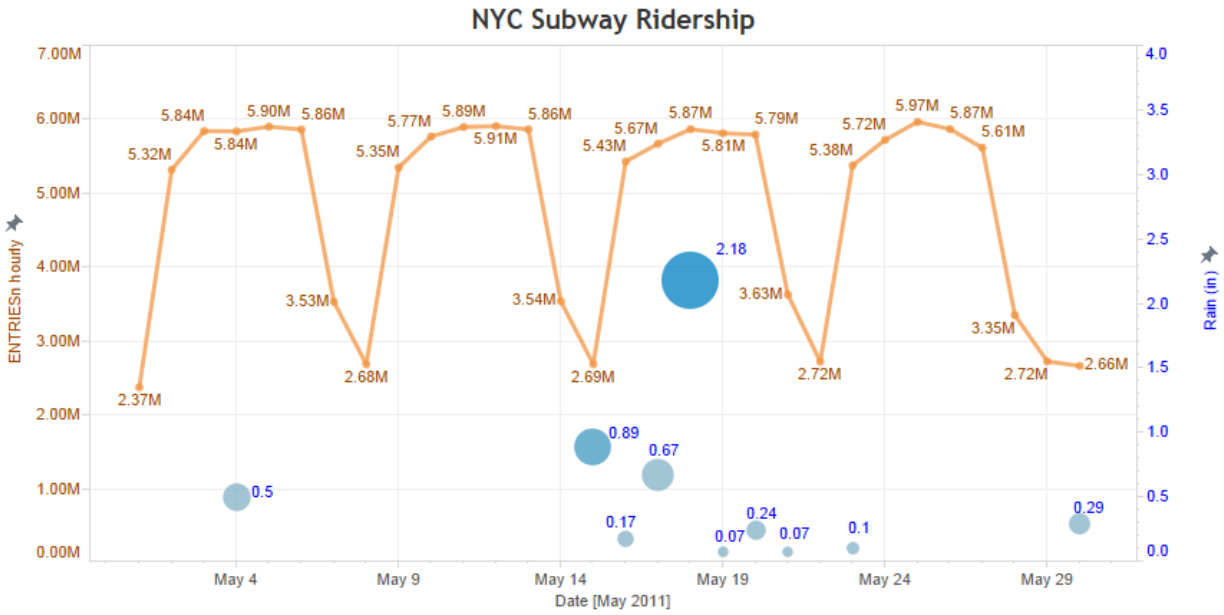


Figure 2. Combined line graph and scatter plot. The orange line shows the total hourly entries by day on the left axis while the blue dots indicate the total rain precipitation in inches by day. The size of the blue dots is scaled to the value on the right axis.

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?

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.

My analysis and interpretation of the NYC Subway data indicates that average entries are indeed higher on rainy days. The higher average entries on rainy days was found to be statistically significant using the Mann-Whitney U statistical test and so it is reasonable to conclude that more people do ride the NYC subway when it is raining. Although the coefficient of determination for the linear regression model for total ridership indicated a weak fit to the data, the rain coefficient in the linear model was found to be 2.03 indicating an increased ridership on during rainy days.

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, Linear regression model, Statistical test.

Several potential shortcomings were observed while conducting this analysis. The relationship between the Unit turnstile variable and actual Station turnstiles in dataset is not defined. It is not clear if turnstile units could be aggregated by location or if Units with low ridership could be excluded to improve the analysis.

Secondly, the relationship between ridership and rain days does not appear linear, and is in fact a time series and it is not clear how a linear model can be applied to time series data. Furthermore, multiple variables were used in the linear regression model that were chosen simply by visual inspection of their distribution. It is not clear if the chosen variables provide predictive value to the model other than the fact the R2 value increased when they were included compared to using the rain variable alone.

Lastly, all the dummy variables were used as-is, as provided by the skeleton code within exercise 3.5 and no work was performed to determine their suitability for use in the model.

Future analysis on this dataset should consider performing statistical tests on all the variables used as well as consider reducing the number of dummy variables.