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 utilized to analyze the NYC subway data. A two-tailed P value was used to test the null hypotheses that there is no difference between the distribution of hourly turnstile entries for rainy vs. non-rainy days. The p-critical value is .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.

Data analysis was conducted to determine whether the data is normally distributed. A statistical test such as Welch's t test could be utilized if the dataset is from a normal distribution.

The first step was to create a visualization of the distribution using a histogram of the frequency of hourly turnstile entries as shown in Section 3.1. The diagram clearly shows the data is not normally distributed for the rainy and non-rainy day datasets.

The second step was to perform the Shipiro-Wilk test for normality. The null hypothesis is the data is from a normal distribution. Below is the test statistic and p-value output for the rainy and non-rainy day datasets.

Rainy >>>
Shapiro-Wilk Test Statistic: 0.593882083893
p-Value: 0.0

Non-Rainy >>>
Shapiro-Wilk Test Statistic: 0.595618069172
p-Value: 0.0

Since the p-value for both datasets is < .05 the null hypothesis can be rejected that the data is normally distributed. Therefore, the Welch's t test is not applicable for the dataset.

Please note the Shapiro-Wilk test was also run on a random sample of 1000 from each dataset since Python displays warnings that the p-value may not be accurate for N > 5000. The difference was immaterial.

The Mann-Whitney U test is applicable for this dataset because it is a non-parametric test and does not assume the data is drawn from any particular probability 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.

>>>
Two-Tail p-Value: 5.48213914249e-06
Rainy Day Mean: 2028.19603547
Non-Rainy Mean: 1845.53943866

The reported p-value in the Mann-Whitney U test in Python is one-tail so multiplied by 2 for the two-tail hypothesis.

1.4 What is the significance and interpretation of these results?

Based on the p-critical value of .05, the null hypothesis is rejected. The interpretation of the results is that the distribution of the number of entries is statistically different between rainy and non-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?

The Statsmodels implementation of Ordinary Least Squares was used to compute the coefficients and predictions for ENTRIESn_hourly in the regression model. The OLS Regression Results classes for summary, params and rsquared were very useful experimenting with various independent and dummy variables and reviewing the impact on the model.

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

The input variables used in the model were rain, tempi and wspdi. The dummy variables used were UNIT and calculated Time Period from Hour.

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.

Some general assumptions were made about subway ridership that provided the basis for the features used in the model:

1. Weather conditions such as fog, rain and wind may increase ridership because people that might walk or ride a bike would use the subway instead.
2. Seasonality/temperatures may increase ridership particular in the winter and summer with extreme low or high temperatures.

3. Subway locations for highly traveled lines for commuters.
4. Hour of the day would have an impact on ridership because more people would use the subway during rush hour vs. weekends/evenings.

The first step was to analyze the various weather condition characteristics fog, rain, temperature and wind speed standalone from any other independent variables to determine if the null hypothesis could be rejected that each does not increase subway ridership.

Fog >>>
const: 1889.116150
fog: -257.135243
p-value: 0.076

Rain >>>
const: 1845.5394
rain: 182.6566
p-value: 0.000

Temperature >>>
const: -87.5296
tempi: 31.2837
p-value: 0.000

Wind >>>
const: 1632.4906
wspdi: 36.6778
p-value: 0.000

Based on the regression results, the null hypothesis was rejected for rain, temperature and wind speed. The null hypothesis could not be rejected for fog so it was not included in the model.

The next step was to add the categorical variables to the model by creating dummy variables for each distinct category for Unit and Time Period. The hour of the day really does not have a natural order for subway ridership. However, the hour were separated into the following 5 time periods:

1. Late Night < 6 am every day
2. Weekends 6 am – 12 am Saturday and Sunday
3. Evenings 8 pm – 12 am Monday – Friday
4. Midday 9 am – 3 pm Monday – Friday
5. Rush Hour 6 am – 9 am and 3 pm – 8 pm Monday – Friday

As shown in Section 3.2, a bar chart was created for turnstile entries by hour and time period. The dataset suggests the sample may not be representative of the population as there is only data for hours 0, 4, 8, 12, 16 and 20 or 4 hourly increments. In addition, the turnstile entries by time period is not consistent with the assumption that rush hour would have the most riders. Midday and evening have the highest representation that also suggests the sample may not be representative of the population or the hour data was grouped and the time periods were transposed with each other.

Adding the calculated time period categorical variable had a slightly better R^2 at .522 vs. .519 for hour. The coefficient for temperature inverted to negative using hour as well. Therefore, time period was used as dummy variables in addition to unit.

Research was performed on the multicollinearity warning in the model results. The warning appears to be related to inclusion of dummy variables for every category for unit and time period. Dropping various dummy columns to avoid the inclusion of a category for every dummy column did not resolve the warning due to the high volume of units.

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

Coefficients >>>
constant: 1561.4537
rain: 28.6133
tempi: 3.9507
wspdi: 18.2972

2.5 What is your model's R^2 (coefficients of determination) value?

R^2 >>>
Dep. Variable: ENTRIESn_hourly
R-squared: 0.522
Model: OLS
No. Observations: 42649

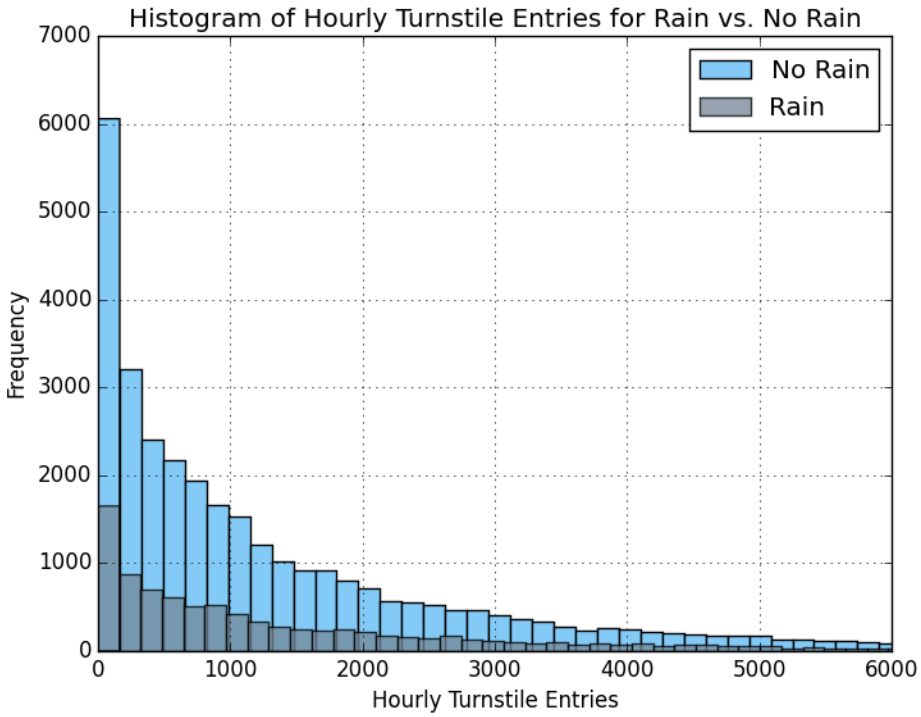
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

The interpretation of the R^2 value is that 52.2% of the variation in the ENTRIESn_hourly dependent variable is explained by the model. The higher the R^2 value the better the fit. This suggests the model is good but most likely impacted by the limitations in the dataset for seasonality weather impact and inconsistencies in the time period representation.

Section 3. Visualization

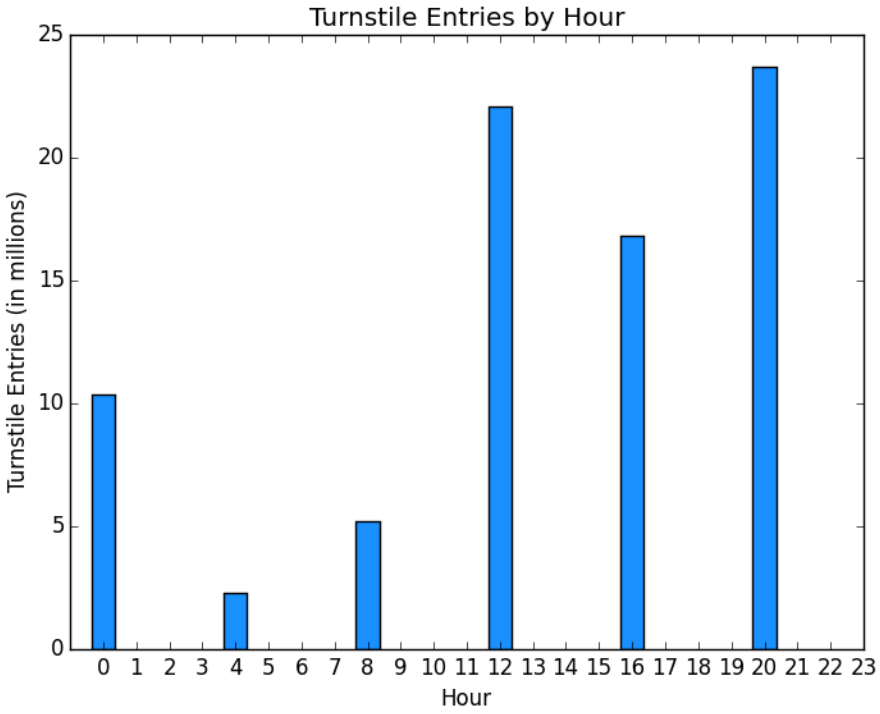
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

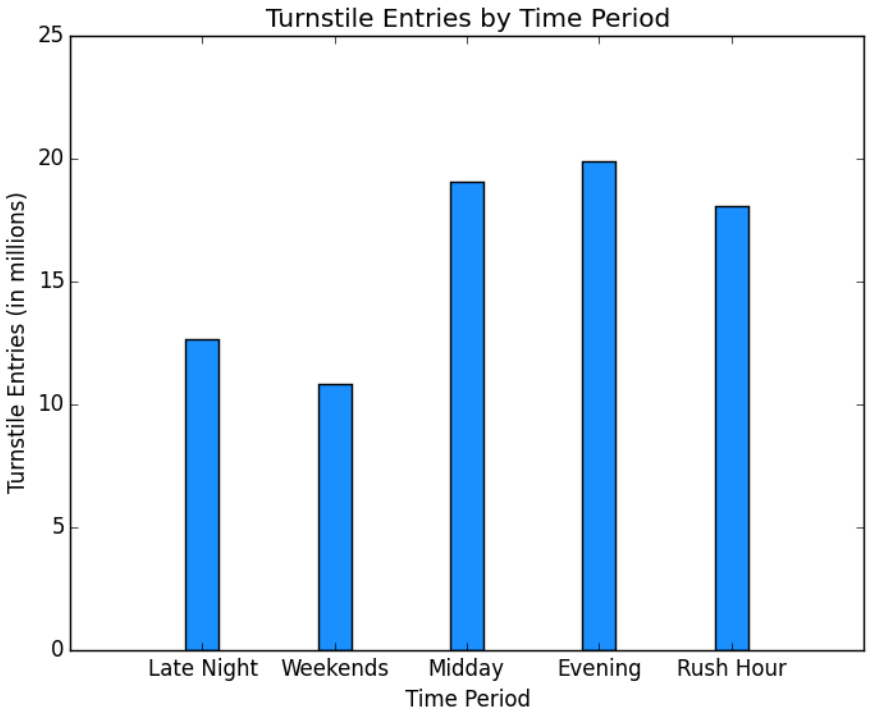
The following visualization is a histogram of hourly turnstile entries for rainy and non-rainy days on the same plot.



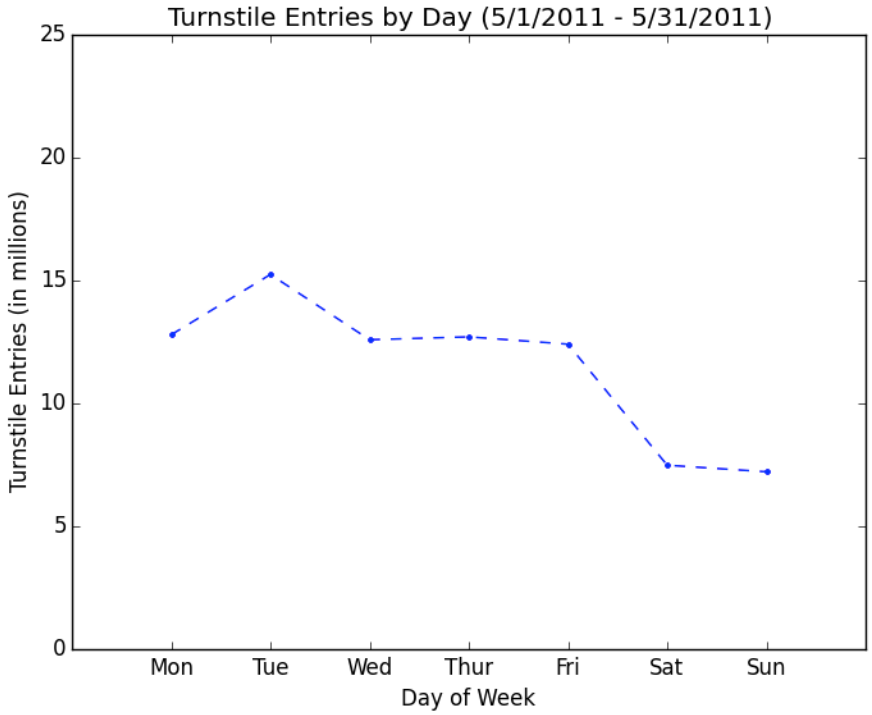
3.2 One visualization can be more freeform.

The following 2 visualizations are bar charts of turnstile entries by hour and the custom categorical time period.

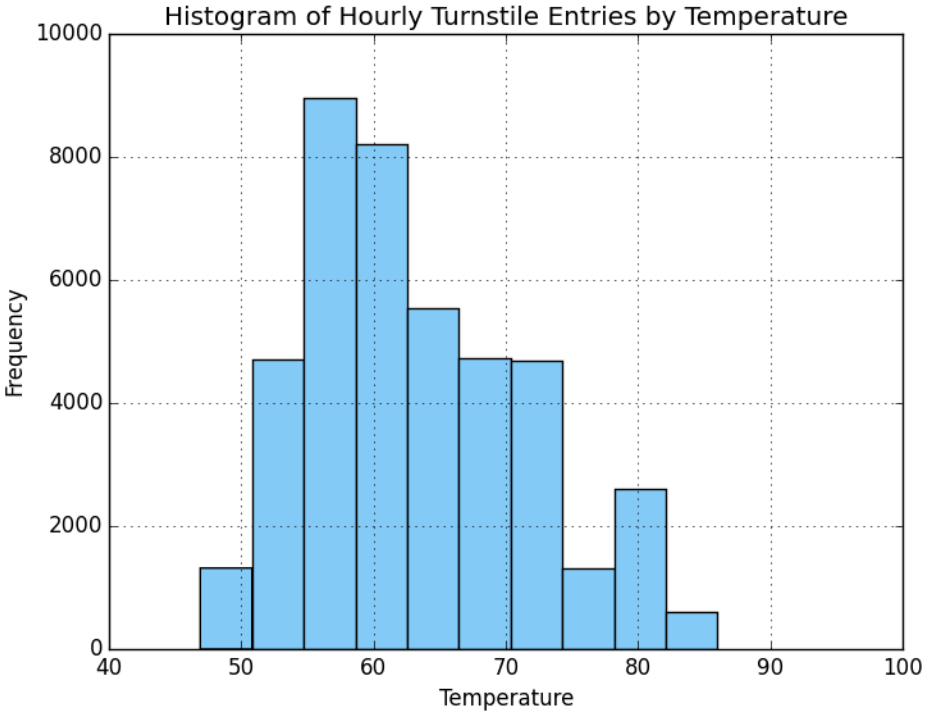




The following visualization is a line chart by day of the week to show higher ridership for weekdays and demonstrating that the dataset was limited to May 2011.



The following visualization is a histogram chart for temperature to show the relatively mild temperatures for the period of time represented in the dataset.



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?

As shown in Section 3.1, the distribution of the number of turnstile entries has a similar shape but the frequency of lower entries is lower when it is raining. This suggests that more people take the subway when it is raining. However, this visualization could be skewed based on the representation of entries for rainy days. The number of records for rain vs. no rain is 9585 and 33064 respectively. The number of rainy days is well represented in the dataset.

In conclusion based on analysis and interpretation of the data, one can conclude that more people ride the NYC subway when it is 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.

As discussed in Section 1.3, the Mann-Whitney U statistical test indicates that the distribution is statistically different between rainy and non-rainy days. The means indicate on average 183 more people ride the subway when it is raining. The regression model for the rain independent variable has a positive coefficient of 28.6133. This indicates the amount of the hourly entries dependent variable is expected to increase when the independent variable increases by 1.

The statistical test and regression model both confirm that rain has a positive relationship on subway ridership.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including: Dataset, Analysis, such as the linear regression model or statistical test.

There are several potential shortcomings with the analysis of the subway ridership data:

1. Dataset is limited to the month of May neutralizing the seasonality effect of most of the weather condition variables.
2. Dataset does not have samples for all hours of the day so ridership for peak subway time periods may not be properly represented.
3. Inclusion of dummy variables for 240 categorical units introduces multicollinearity in the regression model.

The amount of precipitation is very low for rainy days with the 25% and 50% quadrants having no measurable amount of precipitation for some records in the dataset.

Precipitation >>>
count: 9585.000000
mean: 0.020520
std: 0.051408
min: 0.000000
25%: 0.000000
50%: 0.000000
75%: 0.010000
max: 0.300000

Temperatures were relatively mild. The mean temperature was 63.10 with min and max of 46.9 and 86 respectively.

Temperature >>>
count: 42649.000000
mean: 63.103780
std: 8.455597
min: 46.900000
25%: 57.000000
50%: 61.000000
75%: 69.100000
max: 86.000000

The regression model had small coefficients for rain and tempi. The dataset should be expanded to include a sample from all months to provide better insight in the impact of precipitation and temperature on subway ridership.

As shown in in Section 3.2, the dataset was not representative of all hours in the day. Time period groups of late night, weekends, midday, evening and rush hour should have revealed insights into peak ridership timeframes. The data suggests the hours may have been grouped and thus limiting the ability to identify the time periods more accurately.

The inclusion of the dummy variables for unit introduced multicollinearity in the regression model. One recommendation is to drop a dummy variable from the model with the presence of a constant but with the amount of units this was not successful. This could impact the prediction model accuracy on the dependent variable.

Based on these potential issues and the R^2 value of .522 it is not recommended to utilize the regression model to predict NYC subway ridership.

References:

http://en.wikipedia.org/wiki/List_of_New_York_City_Subway_services
https://www.ied.edu.hk/apfslt/v11_issue2/inel/page6.htm
http://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.mannwhitneyu.html
https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php
http://www.statisticslectures.com/topics/mannwhitneyu/
http://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.shapiro.html
http://www.datarobot.com/blog/multiple-regression-using-statsmodels/
http://dss.princeton.edu/online_help/analysis/interpreting_regression.htm
http://stats.stackexchange.com/questions/70899/what-correlation-makes-a-matrix-singular-and-what-are-implications-of-singularit
http://en.wikipedia.org/wiki/Multicollinearity#Remedies_for_multicollinearity
http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
http://blog.yhathq.com/posts/logistic-regression-and-python.html