Sources used:

http://blog.yhathq.com/

www.wikipedia.com

http://stackoverflow.com/

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html

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?

The statistical test I used was the Mann-Whitney U test, due to the non-normal distribution. The Mann-Whitney U test method I used in Python was for a one-tail P-value. The null hypothesis is that the two data sets are from the same population. I choose .05 for a two tailed test. I also calculated the mean of both samples. To clarify from the documentation for the Python function, the difference between a one sided test and a two sided test for a Mann-Whitney U test, is that a two sided test is gotten by multiplying the value of the one-sided test by two.

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 statistical test largely relies on the independence of the observed. The Mann-Whitney U test does not require a particular distribution to be effective, and is better than the t-test on non-normal distributions as the t-test assumes normal distributions and thus is not apt for analyzing them. However, it is also less informative than the t-test. The t-test provides whether the result is over or under the expected mean of the distribution, whereas the Mann-Whitney U test simply provides that the two distributions are not identical.

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.

Here are the results I got from the statistical test.

Mean with rain: 1105.4463767458733

Mean without rain: 1090.278780151855

U: 1924409167.0

p: 0.024999912793489721

The p-value for my purposes is thus .05 if rounded to two decimal points, but in actuality less than .05.

1.4 What is the significance and interpretation of these results?

Given the results of the tests, we have reason to think that more people use the subway when it is raining, with a difference is statistically significant at a .05 p value for a two-sided test, and our knowledge that the mean of rainy days is higher than the mean of non-rainy days. The .05 p-value for the Mann-Whitney U test tells us that the populations are unlikely to be similar.

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:

Gradient descent (as implemented in exercise 3.5)

OLS using Statsmodels

Or something different?

I used OLS using Statsmodels.

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

I used rain, precipi, Hour, and meantempi. I also used the dummy variable of unit.

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. Your reasons might be based on intuition. For example, response for fog might be: “I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often.”

Your reasons might also be based on data exploration and experimentation, for example: “I used feature X because as soon as I included it in my model, it drastically improved my R2 value.”

Rain, precipitation, hour, and mean temperature all would be very likely to cause changes in the number of users. People's behavior is likely to respond to rain, as if it rains, it is plausible people will choose subways over other forms of transportation that leave them more exposed. If the rain is heavier, it is likelier that the impact of the rain will be stronger. People will use the subway more or less dependent upon the hour of the day, as there are hours with heavier usage. People will also respond to the temperature, as colder days are likely to see heavier subway use.

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

The coefficients I got were these:

2.94645287e+01 for rain, 2.87263803e+01 for precipi, 6.53345653e+01 for hour, and

-1.05318249e+01 for meantempi

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

The model's R2 is approximately .48 when I round it to two places.

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?

The R2 value means that my model covers a bit less than half of the variance. It means that the model does not provide a very good prediction, but that it does have a substantial amount of predictive power. So, it could be used as a heuristic, but about 50% of what is seen is not explained. In a case like this, it is unlikely an R2 of 1 will be found though without overfitting.

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

Remember to add appropriate titles and axes labels to your

plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

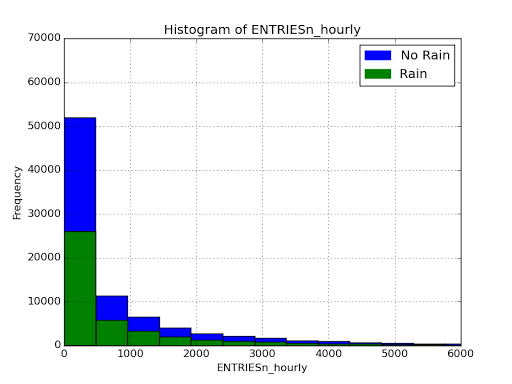
3.1 One visualization should contain two histograms: one of ENTRIESn\_hourly for rainy days and one of ENTRIESn\_hourly for non-rainy days.

You can combine the two histograms in a single plot or you can use two separate plots.

If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.

For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn\_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn\_hourly that falls in this interval.

Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.

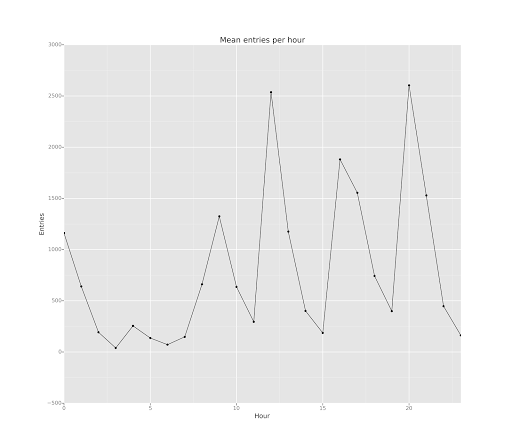


The key insight of the histogram is that the numbers are not normally distributed, and that the numbers for one sample are larger than the numbers for the other. Comparison, however, is limited by the fact that raw entrants is not as valuable as a measure of average entrants for time-period.

3.2 One visualization can be more freeform. Some suggestions are:

Ridership by time-of-day

Ridership by day-of-week



The second chart is on mean entries by hour. It is a line chart tracking the average entries by hour to detect trends across stations in usage, so that way peaks and valleys can be identified. A major insight is how jagged the distribution of usage is.

Section 4. Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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?

More people ride the NYC subway when it is raining. The two distributions are not the same, and rain is associated with more riders, and on average rainy days have more riders than non-rainy days.

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.

Using the Mann Whitney U test, it is clear that the two distributions are not identical to each other. Additionally, the mean of the distribution for rainy days is higher than the distribution for non-rainy days, which does suggest that rainy days will generally have more riders especially given a significant sample size. Also, when running a linear regression, the coefficient for rain is positive, meaning that more rain is associated with more ridership, and as well the coefficient for precipitation is also positive which means that the more rain the more riders.

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.

The dataset in this case only covers one month of data. Some biases may exist because it might not be the case that every month has the same patterns. Rain in December might end up causing people to stay home instead of riding, even though the rain in May causes people to ride more frequently. Additionally, a large proportion of the data collected comes from a few stations, and this can be found by summing up the count of records for each Unit, as the standard deviation of these count is 779.5553 and so is greater than the mean 283.7656, and the mean is greater than the median of 182 by a large proportion, and about 10% of the data collected comes from station R549, out of 464 other stations. This sort of distribution suggests that the data collected is not a fair sampling of the entire population.

The analysis also has some limitations, as the R squared measure was rather low for the linear regression, even though it included other factors besides the rain, which means that our findings on the relationship between rain and ridership are not responsible for much of the variation. Also, at a p-value of .05 exactly, the chance that the distributions are different is actually only 1/20, which is a substantial margin for error and barely makes the cut-off point.