Project 2 – Alyse Keim

Section 0. References

Informational/Reference:

- https://en.wikipedia.org/wiki/Mann–Whitney_U_test
- https://docs.python.org/2/library/datetime.html
- http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.groupby.html
- http://statsmodels.sourceforge.net/devel/generated/statsmodels.regression.linear_model.OLS.html
- https://en.wikipedia.org/wiki/Coefficient_of_determination
- http://ggplot.yhathq.com

Coding examples:

- http://stackoverflow.com/questions/30222533/create-a-day-of-week-column-in-a-pandas-dataframe-using-python
- http://stackoverflow.com/questions/6871201/plot-two-histograms-at-the-same-time-with-matplotlib

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 am using a **Mann-Whitney U** test to analyze the data
- Because we are not looking for greater than or less then the different data sets, we will be using a **two tailed P** value, as this is analyzing *difference* instead.
- Hypothesis:
  - H_0: $\mu_I = \mu_C$ (OR $\mu_D = 0$)
  - Null hypothesis is that the population of subway riders is the same whether it is raining or not
- P Critical: .05 (5%)

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.

Looking at the data, I could not run a Welch's T test because even though it allows flexibility in assumptions like sample size and variance, it must be normally distributed and our data sets (both ran and no rain) were not normal. Therefore, I applied a nonparametric Mann-Whitney U test as it can be applied to unknown distributions. Additionally, it is meant to look for equality in populations of 2 samples and thus was perfect for our use case.

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: 1105.4463767458733
- Mean without rain: 1090.278780151855
- U statistic: 1924409167.0
- P value: 0.024999912793489721

1.4 What is the significance and interpretation of these results?

At first glance, we can see the mean without rain is 15 rides less than with rain, however, this does not tell us much. We must investigate further. The U statistic cannot exceed product of the sample sizes for the two samples. The U value is not half of this, the p value of .025, .05 for our two-tailed test, and thus is within our critical range .05, and thus we can, with 95% confidence, reject the null hypothesis that the ridership is the same with or without rain. The distribution of the number of entries into the NYC subway is significantly 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:

For my first linear regression model, I used the python library 'statsmodels' and the method of ordinary least squares to compute my coefficients theta.

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 both features and dummy variables. With features, I did many iterations to better understand both the data and accuracy of my model. The evolution of such is listed below. Additionally, I used the dummy variable 'UNIT' – as it was an important categorical parameter of the model. Our results could be seriously dampened if we did not keep track of location of subway (location is too big of factor for subway rider). This is exemplified below.

Without UNIT as dummy variable:

- Features: 'rain', 'precipi', 'Hour', 'meantempi
- R^2 = 0.0306

With UNIT as dummy variable:

1. Features: 'rain', 'precipi', 'Hour', 'meantempi
   - R^2: .4792

2. Features: rain, precipi, meanwindspdi, meantempi
   - R^2: 0.4436

3. Features: 'rain', 'precipi', 'Hour', 'mintempi'
   - R^2: .4796

4. Features: 'rain', 'precipi', 'meanwindspdi', 'Hour', 'mintempi'
   - R^2: .47998

5. Features: 'rain', 'precipi', 'fog', 'Hour', 'mintempi'
   - R^2: .48032

** Will use best iteration (5) for rest of questions.

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.

At first, I went with intuition and used rain, precipitant, hour, and mean temp. With these in place, I got a decent R^2 value, but I wanted to explore the data more. From there, I thought about how miserable the windspeed can be in the winter and thought maybe time of day wasn't as big of a factor. There I swapped the two – but my R^2 went down! This was a mere exploration of data that I wasn't sure about. Soon I used mintempi (due to my intuition) and tried both windspeed and hour.

The final iteration was interesting because I initially ignored fog completely – as that is the last factor I would ever consider personally for riding the subway, however, when I read the examples in this question I thought "why not." Therefore I went back and gave it a try. Looks like others feel differently about fog than I as it was the highest R^2 value (with the others held constant).

In general, however, I tried to keep all the features constant except one as I explored the data. This allowed me to compare 2 features against each other to decide which effected the model most.

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

For iteration 5 ('rain', 'precipi', 'fog', 'Hour', 'mintempi'):

- rain 9.459249
- precipi -36.465036
- fog 204.589120
- Hour 65.325722
- mintempi -15.328123

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

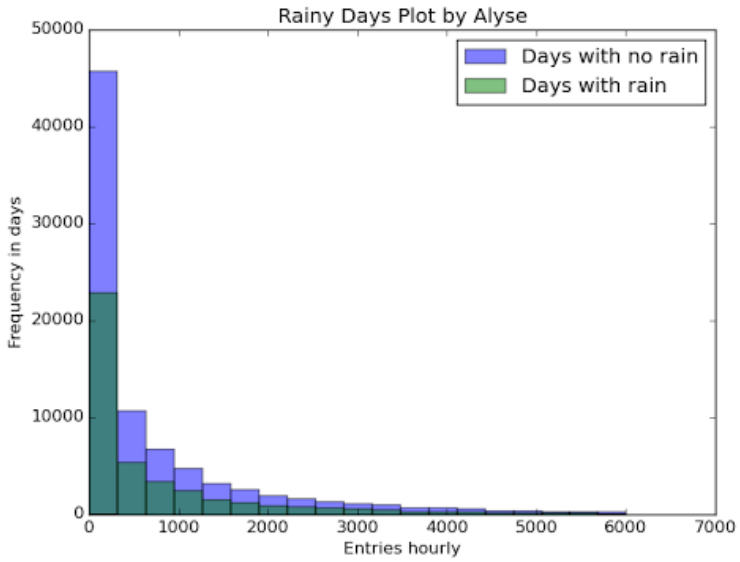
R^2 = **.48032**

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 R^2 is essentially the ratio of explained variance to total variance and thus is indicative of how much variance can be explained. In my example, the features rain, precipitation, fog, hour, and minimum temperature explains 48.03% of the variance in the model. While in most cases this is extremely small and definitely insufficient, I can accept this for now for our use case (although I would like more computational power to test more parameters!) Because our use case is in extremely uncontrolled environment, explaining all the variance would be impossible. We would need to enhance this with many other data sources (like event data, station shut downs, seasonal tourism, paramedic schedules) to help explain the anomalies. Therefore, for the data we have and the computational power we have, I think this linear model is appropriate.

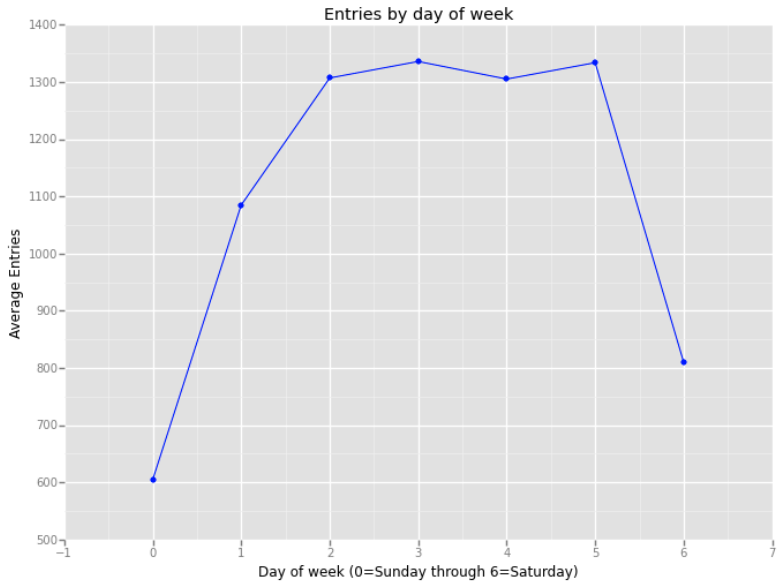
Section 3. Visualization

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



This visualization is overlaying histograms of our data. The blue represents the total entries hourly to the subway over all days without rain. Additionally, the green represents entries hourly with rain. While it is a great way to visualize the data, this is a sum rather than an average, and thus, the visualization does not tell us much about the relationship between the two subsets of data (as the total days without rain is greater than the total rainy days). It is created via madplotlib.pyplot module in python.

3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like.



This visualization is a line plot is showing the average entries to the subway per day of week. We can see that more people ride the subway during the week rather than the weekends with the peak ridership happening on Wednesday and Friday. This plot was created via the ggplot python module.

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?

Yes, base off the statistical evidence of the Mann-Whitney U test, more people ride the subway when it is raining. This is confirmed through the p value that falls within the critical value and U statistic. Additionally, from our deeper investigation, we can start to identify the leading aspects of rainy weather that causes people to go underground like precipitation, fog, and miniature temperature values.

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.

The analysis that led me to my conclusion was continuous throughout. First, I observed that mean of the data with rain was 1.3% higher than without. This aligned with my intuition and gave me motive to investigate further. From the Mann-Whitney U test, I found a u statistic of 1924409167 below half and a p value of .025, which doubled for the 2 tailed test gives us .05. This is within our critical range, and thus, we are 95% confident that the population distribution of rain is significantly different than the population without. Additionally, I was able to account for 48% of variance (r^2), which means we are able to begin understanding why our model is the way it is and exactly what factors seem to affect it the most.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

While the data set and model was great for our purposes of learning, in a real life analysis, there are many shortcomings. First off is in the data. The most immediate factor that presents itself is the time range. The data is taken from one month of data, 4 years ago. With such a small sample set, 1 month within 1 season of mild weather we do not get an accurate representation of variance in weather that could be captured over a longer period of time (I would like to see the data for snow!). Additionally, this was 4 years ago, and thus trends that hold true back then might not hold true today due to deterioration of subways, abandonment of subway stations, evolution of residents, etc. While size and sampling time make a difference, the dataset also lacks evolution. With modeling, it is important to have an evolving relationship with your data. After the initial dig down, you find important aspects and you build on them. For example, if hour was a huge contributor to a higher R^2, I may go back to the collection point and ask more specific timestamps. Similar, a Boolean 'yes' or 'no' to precipitation might not be enough – what about inches or ferocity of rain?

Not only would does the data pose it's shortcomings, but the models do as well. Some of these include the static nature of the data. Once a model is trained, no new data will be taken into account. Even though we could recalibrate the model as time goes on, it will not evolve with new information that may become the norm. Additionally, the models lack scalability. As I already experienced already with the computational power, models such as these lack the ability to handle massive amounts of data. Therefore we must sample, pick and choose in the correct, making assumptions that may not be correct. Some of this is being combatted with the evolution of technology, however, there are still many unknowns.

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?

Great, interesting dataset to learn on! Very applicable, however, I may only think this because I live in NYC and can easily relate ☺