http://www.theonion.com/graphic/candidate-profile-donald-trump-50675

Section 0: References

Include a list of references you have used for this project. Please be specific – for example, instead of including a general website such as stackoverflow.com, try to include a specific topic from Stackoverflow that you have found useful.

http://work.thaslwanter.at/Stats/html/statsModels.html#linear-regression-analysis-with-python

http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf

https://storage.googleapis.com/supplemental_media/udacityu/4332539257/Linear%20Regression.pdf

http://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.mannwhitneyu.html

http://stats.stackexchange.com/questions/124995/how-do-i-interpret-the-p-value-returned-in-scipys-mann-whitney-u-test

https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php

http://www.chem.utoronto.ca/coursenotes/analsci/stats/ttest.html#3steps

http://www.chem.utoronto.ca/coursenotes/analsci/stats/12tailed.html

http://www.henry.k12.ga.us/ugh/apstat/chapternotes/7supplement.html

http://www.ats.ucla.edu/stat/mult_pkg/faq/general/tail_tests.htm

http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/tests-of-means/how-are-dependent-and-independent-samples-different/

http://www.ats.ucla.edu/stat/mult_pkg/faq/general/tail_tests.htm

https://storage.googleapis.com/supplemental_media/udacityu/649959144/MannWhitneyUTest.pdf

http://stattrek.com/regression/slope-test.aspx?Tutorial=AP

http://ggplot.yhathq.com/docs/index.html

http://ggplot.yhathq.com/docs/geom_histogram.html

http://stackoverflow.com/questions/23457513/why-python-ggplot-returns-name-aes-is-not-defined

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 two-tail P value? What is the null hypothesis? What is your p-critical value?

I used a SciPy Mann-Whitney U test to analyze the NYC subway data.

While I was trying to answer the question **whether subway ridership is higher during rainy days**, which requires a one-tail test, I decided to first use a two-tailed test in order to account for the possibility of missing an effect in the opposite direction.

M-W is used to test, for two populations with unknown distributions, if we draw randomly from each distribution, whether one distribution is more likely to generate a higher value than the other. For a two tailed test, the null and alternative hypothesis are as follows.

H_0: P(ENTRIESn_Hourly$_{rain}$> ENTRIESn_Hourly$_{clear}$) =0.5

H_0: P(ENTRIESn_Hourly$_{rain}$> ENTRIESn_Hourly$_{clear}$) != 0.5

To be conservative I used a p-critical value of .025. If the 2 * the SciPy M-W p value was less than the p-critical the next step was to perform the one sided test[1].

For the one tailed test

My null and alternative hypothesis were now

H_0: P(ENTRIESn_Hourly$_{rain}$> ENTRIESn_Hourly$_{clear}$) <=0.5

H_0: P(ENTRIESn_Hourly$_{rain}$> ENTRIESn_Hourly$_{clear}$) >0.5

In this case, we could reject the Null hypothesis if the SciPy M-W p value was less than the p-critical. This still left interpretation issues, for instance we still needed to know which of the distributions was greater than the other, so I also included the mean, and median values of each population in my output.

[1] The SciPy Mann-Whitney default p value is a one sided hypothesis, in order to get the two sided p value I needed to multiply the returned p-value by 2.

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 four assumptions that the data is “required” to pass in order for a Mann-Whitney U test to give a valid result are as follows.

Assumption #1: The dependent variable should be measured at the **ordinal** or **continuous level.**

Assumption #2: The independent variable should consist of **two categorical, independent groups**

Assumption #3: There should be **no relationship between the observations** in each group or between the groups themselves.

Assumption #4: The dependent variable is **not normally distributed** and their distributions should also have the same shape.

In the case of this experiment, ridership, the dependent variable, is measured at a **continuous level**[2], and its two associated distributions are **not normal** and have very **similar** shapes. This meets the requirements for assumption 1 and 4. Rain is a discrete random variable where we are interested to see if ridership, y=f (x='rain') is different than y=f(x='not rain'). Because rain is **independent** of ridership, whereas ridership may be dependent on rain, this takes care of assumption 2. Finally, since the probability of taking the train on a non-rainy day **does not affect** the probability of taking the train on a rainy day, we can say that there probably isn’t a relationship between the observations in each group. So hurdle for assumption 3 is also passed. Since all four hurdles are passed in regards to the assumptions, it can be stated that the Mann-Whitney u test is applicable to this dataset.

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 results achieved from the Mann-Whitney Test were as follows

ENTRIESn_hourly with Rain Median: 939
ENTRIESn_hourly with No Rain Median: 893
ENTRIESn_hourly with Rain Mean: 2028.196035
ENTRIESn_hourly with No Rain Mean: 1845.539439

[2] Even though ENTRIESn_hourly takes integer values, the large range of values taken and the even, interval nature of values makes it a continuous-type data

U Statistic: 153635120.5
One-tailed p value: 2.74106957124e-06
Two-tailed p value: 5.48213914249e-06

1.4 What is the significance and interpretation of these results?

Two-Tailed Test:
5.48213914249e-06 < .025 indicates that we should reject the null hypothesis that the two populations share the same distribution.
One-Tailed Test:
2.74106957124e-06<. 025 indicates that we should the null hypothesis that turnstile ridership on days with rain isn't higher than on days without rain.

The fact that our descriptive statistics for rainy day ridership were larger than the descriptive statistics for clear day ridership, gives additional support to our hypothesis that rain increases subway ridership. These statistics also clarify that the one tailed test M-W p value we reached earlier was referring to the rainy population being larger distribution.

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:

I performed multiple regression using statsmodels in order to compute the coefficients for theta and produce my prediction for ENTRIESn_hourly.

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

For the features of my model, I decided to include *rain* and *weekday* as numerical variables, in addition to *hour* and *station* as dummy variables. I decided against including a constant variable because its p value was high[3] and it created issues with my condition number. More importantly, however, it stands to reason that if the coefficients of the included features went to 0(specifically if no riders went in any of the stations) then it fair to assume ridership would also be 0

2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that the select features will contribute to the predictive power of your model

I used a combination of intuition and data exploration when selecting the features for my regression model. First, I included ***rain*** because this was the scenario I

[3] Again a high p value means we should not reject the null hypothesis that the coefficient for the constant should be 0.

wanted to test against. While the results that came back seemed imply significance for the variable, the R^2 was at 0, which meant that the model itself wasn't really that great[4]. I added location at the ***station*** level next, since I believed that the proximity and relevance of different turnstile would have a large impact on subway ridership. This assumption was proven accurate as the R^2 jumped up to *0.521*. I added ***weekday*** next because I knew from experience that more people rode the subway on weekdays due to obligations for work than they did on weekends. I reasoned that were it to rain on a weekday, it would skew the numbers to make it look like more people rode the subway on rainy days. While R^2 jumped again, this time to *0.536,* I noticed unsurprisingly that the coefficient attached to the rain variable had gotten smaller. In addition, the p value for rain was now very high. This change in significance implied that there was indeed some interaction going on between the variables. I wanted to see if I could improve the model even furthers so I added ***hour***. As expected this didn't produce a large change rain's assigned coefficient[5], however, with a new R^2 of .637 the model was dramatically more accurate. This seemed relevant given that rain's p value was still very high[6]. While I played around with adding additional variables to the model, none of them made R^2 much higher. In the cases R^2 did go up, the effect was so small that I still left the variable out, due to risk of over fitting.

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

The coefficients of the non-dummy features were
Rain: 16.8156
Weekday: 451.9305

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

The model's final R^2 value was 0.637.

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 R2 value?

R^2 measures how well the observed outcomes are replicated by the model. In our case our model explains about 63.7% of the outcome. While I'm sure that tests exist which could improve the accuracy of our model even further, this final R^2 is actually

[4] The null hypothesis for each variable in the regression model is that the coefficient should be equal to 0. In this instance the assigned p value for rain was 0, which implies that the null hypothesis should be rejected.

[5] While the coefficients for the model should change when adding any new feature, I did not expect the change to be dramatic in this instance, due to the fact that the 0 to 1 value for rain measured at the day level. To give an example of this, if it had only rained for one hour some day at some location, all the remaining hours would have been assigned a 1 as well even if there had been no other rain.

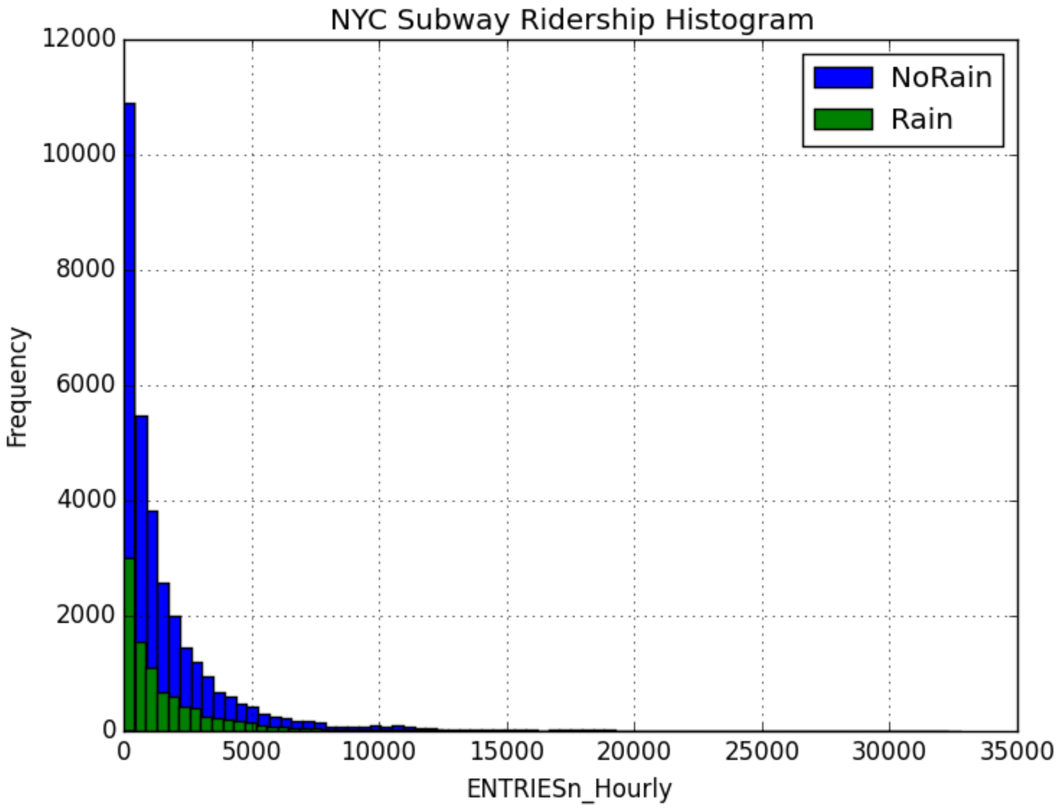
[6] Rain's p ended up with a final value of .104

pretty high and is more than large enough to be significant. This implies that this linear model was appropriate for this dataset. Another point to be noted is that the condition number for the regression model is pretty low at 37.2. This implies that the model does a pretty good job at handling new data, and isn't just limited to explaining away the data we already have.

Section 3: Visualization

Please include two visualizations that show the relationship between two or more variables in the NYC subway data. 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.

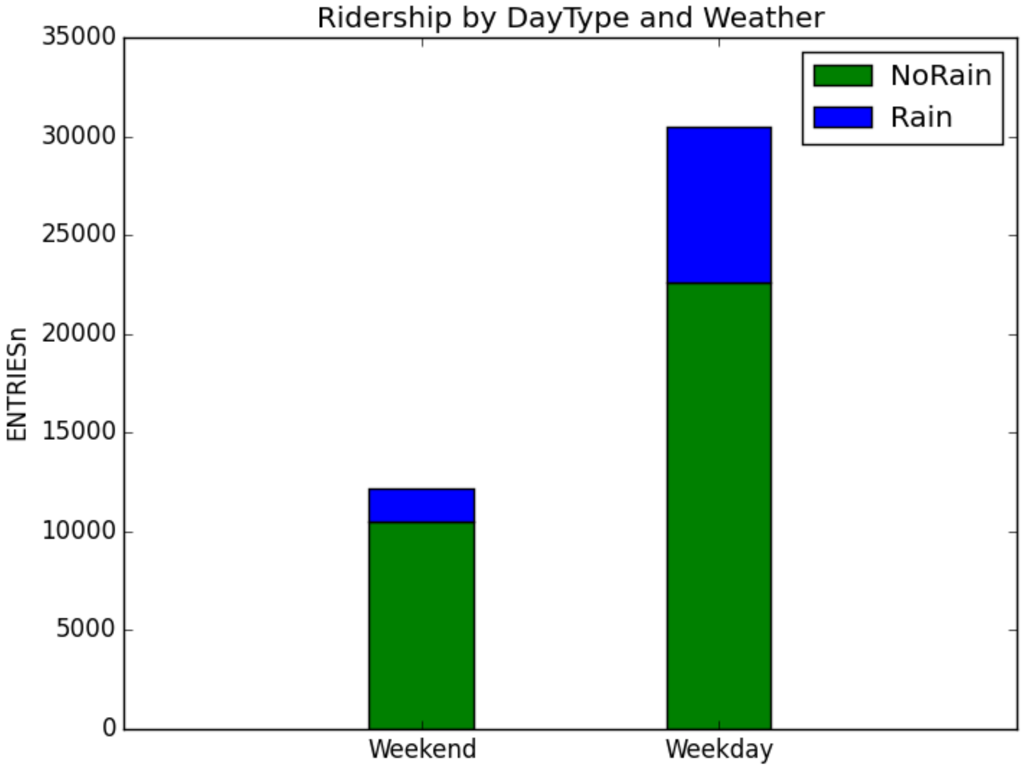
3.1 **One** visualization should **contain two histograms**: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.



From the figure above we see that while the scale is different due to lower total volume[7], the frequency distribution of hourly ridership follows roughly the same shape.

[7] There are fewer days with rain there are days without rain, and thus less total rainy day riders.

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. Some suggestions are:



The figure above visualizes the interaction effect that, I believe, based on my analysis of the data, is skewing our results. The main takeaway is that rain seems to occur more often on weekdays than on weekends. This makes sense given that rain is a completely random effect and the odds of it occurring on one the five weekdays are greater than it occurring on one of the two weekends. However, since our model shows, with a statistical degree of certainty, that more people ride the subway on weekdays than on weekends, it makes sense it is this feature could be inflating our "Rainy" distribution.

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?

My analysis lead me to conclude that based on the results of the mann-whitney U test, and of the regression model, we cannot conclude that rain has ANY effect on subway turnstile ridership.

4.2 What analysis leads you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.

While the Mann-Whitney U test indicated, at 97.5% confidence level, that we had a greater probability of pulling a higher value out of the “Rain” population than we did for the “Not Rain” population, the regression model showed that this higher likelihood is more likely due to the interaction effect of rain occurring on more weekdays than weekends. According to our regression model, a little over 450 more people ride the subway on the weekday than they do on the weekend. While this could be due to a variety of factors, it makes sense given riders obligation to commute to and from work on weekdays. In any case, the weekday coefficient was assigned a p value of 0, while the rain coefficient, which, at 16.8, was much lower, had a high value of .104. This means that while for work, we can reject the null hypothesis that the coefficient should be equal to 0; we cannot make this same statement for rain. The fact that we were able to attain the, relatively high, confidence coefficient (R^2) of 0.637 brings additional weight to this interpretation.

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.

Least squares regression can perform very badly when some points in the training data have excessively large or small values for the dependent variable compared to the rest of the training data. In the case of our subway turnstile data, we actually have a large number of extremely high ridership values, which can be interpreted as outliers. Another potential shortcoming in our model/dataset is that by definition, out regression model assumes for linearity’s. In the real world a lot of strong relationships exist that aren’t necessarily linear. In the case of our subway data, time is an example of a variable that doesn’t follow a linear trend. Subway ridership increases as some points in the day, and decreases at others. Even within hourly groups, there could be instances were ridership increases dramatically for 30 minutes during rush hour, but then falls to nothing, as the number of people of needing to use the subway at that time is used up. This can also occur multiple times during the day, with different highs and lows. While this issue is partly mitigated by grouping time of day into different blocks, it still represents a problem for the model.

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

There seemed to be a large number of extreme values in the dataset. I didn’t know if this was due to how the data was logged (meaning it was tracking error), or if we actually had such a large number of riders pass through the turnstiles on certain hours. It would have probably made the model more accurate if we removed these outliers, but without context I decided that could be risky so chose not to do this.