Project 1: Analyzing the NYC Subway Dataset

Garrett Fox

MAY 11, 2015

Section 0.

References

- stackoverflow.com : basic python syntax examples
- python official website : basic python syntax examples
- http://openclassroom.stanford.edu/MainFolder/CoursePage.php?course=MachineLearning
- http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
- https://statistics.laerd.com/premium-sample/mwut/mann-whitney-test-in-spss-2.php

Section 1.

Statistical Test

1.1

- I used the **Mann-Whitney U** statistical test to analyze the data.
- I used a **one-tailed** p value.
- The null hypothesis was **that there is no difference in subway ridership while it is raining and while it is NOT raining, in other words…** H_0 is that the mean of the entries hourly value from the sample where rain=1 is equal to the mean of the entries hourly value from the sample where rain=0.
- The p-critical value used to determine statistical significance was **0.05.**

1.2

- This statistical test (Mann-Whitney U) is applicable to the dataset because….after a quick look at the histograms of the data from the 2 samples being compared (1. ridership on rainy days vs. 2. ridership on non-rainy days), it is clear that **the sample data does not resemble a normal distribution** (which is a crucial assumption for using a Welch's t-test) **so a non-parametric test like the Mann-Whitney U test is more appropriate to use in this situation** to try to determine a statistically significant difference between these non-normally distributed samples.
- I hypothesized that blank in the weather affects the subway ridership in this way…..The test is making the assumption that the the distribution of ridership in the two samples comes from a normal?? distribution.

1.3

My results from running the Mann-Whitney U test:

- p-value: **0.025**
- mean of sample 1 (# of entries on rainy days) = **1105.45**
- mean of sample 2 (# of entries on non_rainy days) = **1090.28**

1.4

- Using a p-critical value of 0.05 to determine statistical significance, the p-value obtained of 0.025 falls in the critical region, and so we can conclude that there IS a significant difference between the means of the # entries of two samples. So the slight increase in # of entries (an increase of 15 entries hourly) in the subway on RAINY days is most likely not just due to random chance from sampling error. I interpret this to mean that the rain does have some sort of effect in leading more people to use the subway, however, it's still unclear to me from this test whether or not other variables are just as important or more important in nudging people to use the subway more, for example, maybe on rainy days, the temperature is always a lot colder outside and the reason more people are riding the subway is really just to escape the cold air and ride the subway where it is a more comfortable warmer temperature due to people's body heat.

Section 2.

Linear Regression

2.1

I used the **gradient descent** approach to compute the coefficients *theta* for a certain set of features and produce predictions for ENTRIESn_hourly based off those features and thetas.

2.2

I used the following features (input variables) in the model for predicting entries hourly (or actually entries every 4 hours as the dataset would seem to indicate):

- **rain** : whether it was raining or not (0 or 1) at that location
- **precipi** : precipitation in inches at the time and location
- **Hour**: the hour of the timestamp, truncated from TIMEn
- **mintempi**: the minimum temperature in Fahrenheit on that day for that location
- **fog**: whether there was fog or not at that location
- **UNIT** (as dummy variable) : the data-collecting unit that is collecting entry and exit data, can collect from multiple banks of turnstiles at once, large subway stations can have more than one unit

---

2.3

Why did I select these features?

- I think the presence of rain, precipitation, fog, and mintempi because these features together can indicate physically unpleasant above-ground weather conditions that might encourage more people to choose the underground subway instead of above-ground transport options. Specifically, I chose **rain** and **precipitation** as features because people don't like to get wet. I chose **fog** because it can decrease visibility and cause above ground vehicle traffic to become slower and more congested. I chose **mintempi** because if the air is really cold, people might go down to the subway where it tends to be a bit warmer because of people's body heat and insulation from outside air. I chose **Hour** because more people might enter the subway at certain times of the day during rush hour to and from work, and less people might enter during down times of the day like 3am for example when most people are sleeping. I chose **UNIT** because certain turnstiles are going to get more riders coming in and out depending on their exact location in the city.
- All of these features, when included, improved the R^2 value (which is the proportion of variability in ENTRIESn_hourly explained by my set of features)

2.4

The coefficents theta (weights) of the features I got from the linear regression with gradient descent model are:

Feature	Theta(weight)
rain	-6.55
precipi	-9.72
Hour	468
mintempi	-88.5
fog	69.9

2.5

My model got a coefficient of determination $R^2 = 0.4649$

2.6

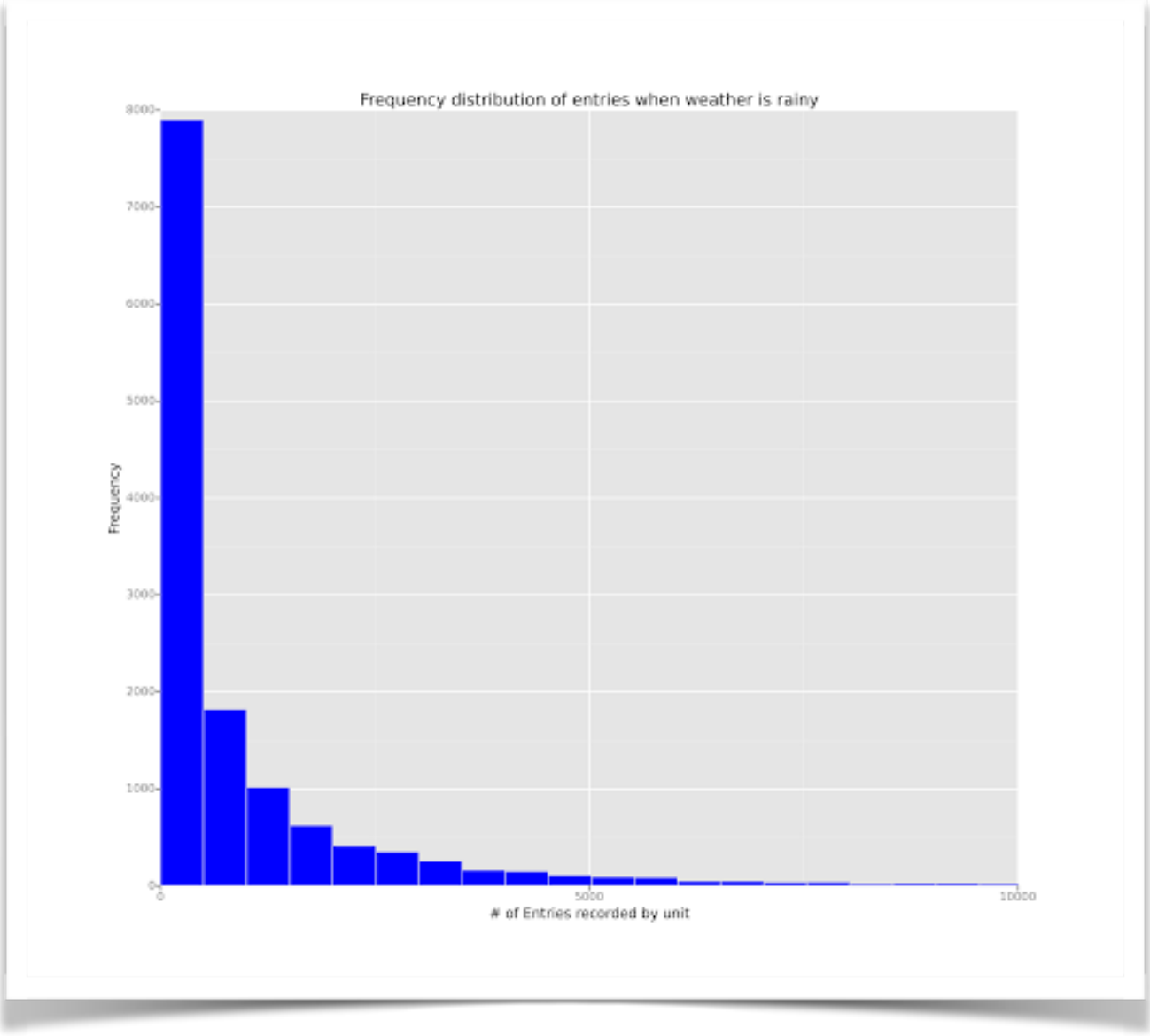
This means that about 46.5% of the variation in hourly entries on the subway can be attributed to the variation in the features included in my model. This means that rain, precipitation, fog, minimum temperature, unit location, and hour of day do appear to play a big role in determining how many people will enter the subway (measured every 4 hours), but we are still missing a lot of other important features in our model that would help explain the other 53.5% of the variation in hourly entries on the subway.

Section 3.

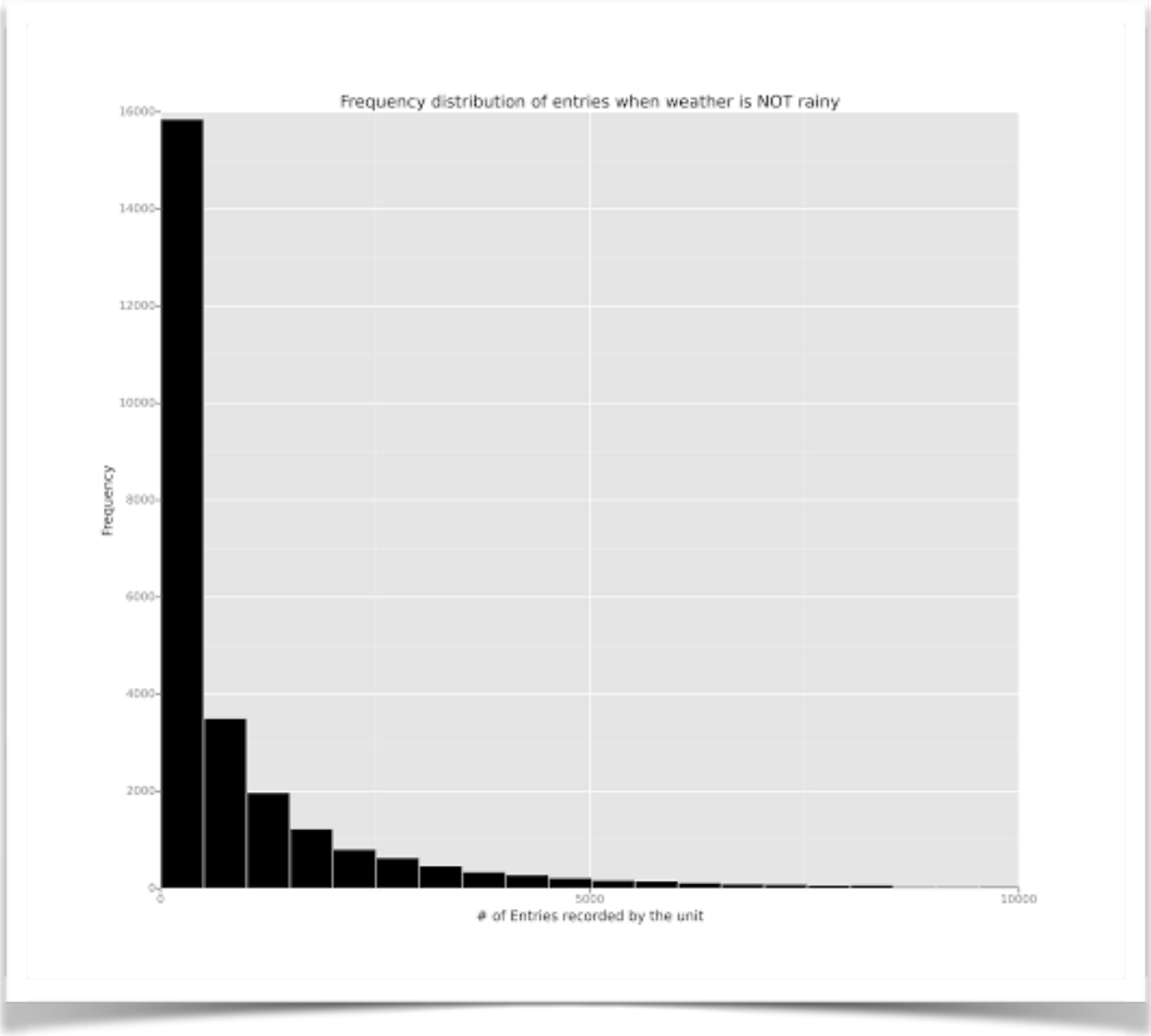
Visualization

3.1

Histogram 1: ENTRIESn_hourly while it is raining

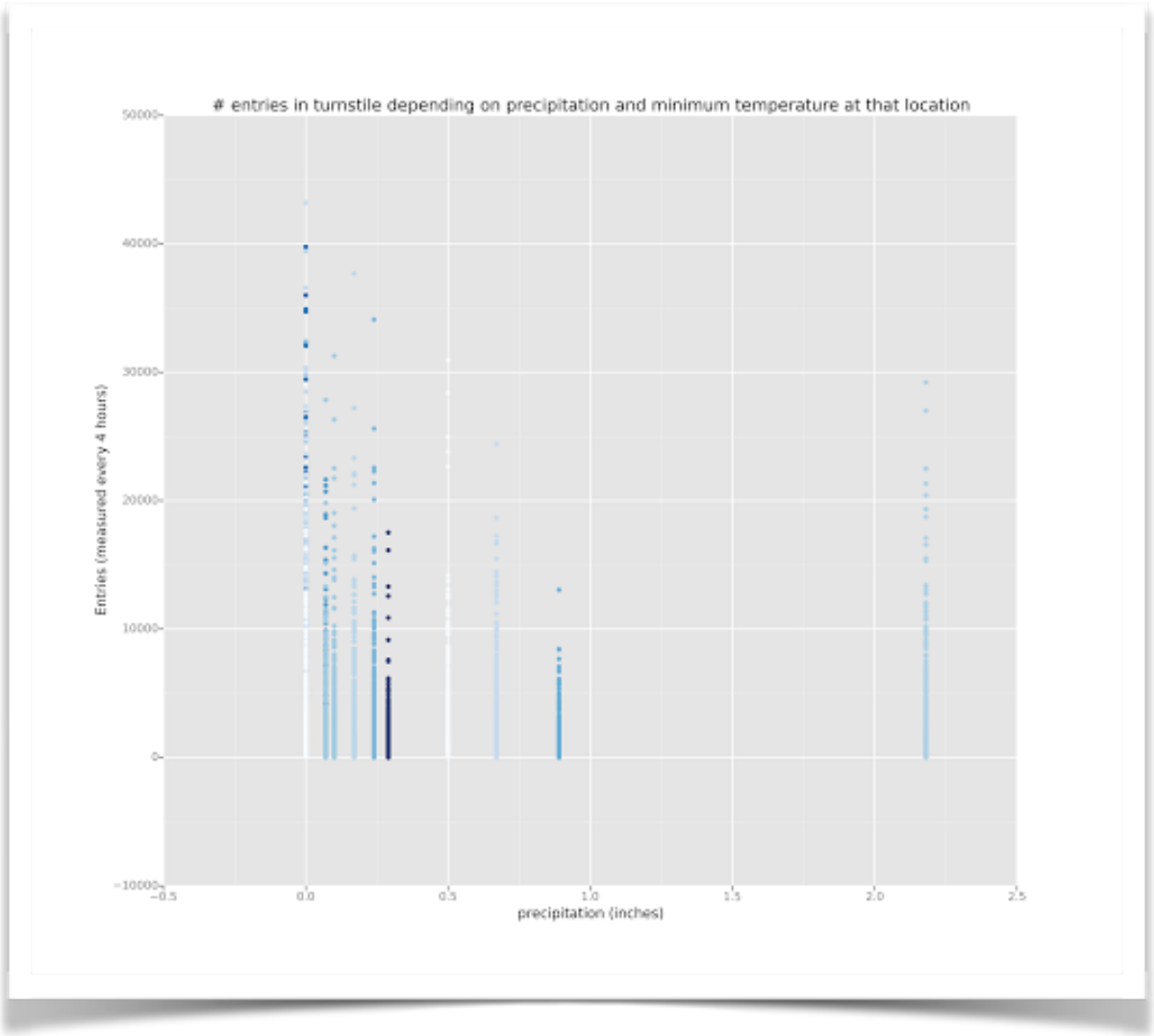


Histogram 2: ENTRIESn_hourly while it is NOT raining



3.2

My freeform visualization shows that, besides the data to the far right of the graph, the trend is that as precipitation increases, subway entries are decreasing. The blue color gradient is signifying lower minimum temperature(light blue) and higher minimum temperature(dark blue). It would appear that most of the darkest blue data points have relatively low numbers of entries, which appears to show that during warmer days, less people enter the subway.



Section 4.

Conclusion

4.1
More people ride the subway when it is raining.

4.2
I arrived at this conclusion by analyzing the results from running the Mann-Whitney U test on the dataset, as well as a linear regression model.

Analyzing the results of Mann-Whitney U test:

- mean of sample 1 (# of entries on *rainy* days) = **1105.45**
- mean of sample 2 (# of entries on *non-rainy* days) = **1090.28**
  - So the mean for # of entries on rainy days is **15.17** higher than that of non-rainy days. That is, we observed an increase in the number of entries on rainy days.
- p-value: **0.025**
  - If we use a p-critical value of 0.05, the p-value of 0.025 falls within the critical region and therefore we determine that the observed increase of 15.17 is statistically significant.

Analyzing the results from linear regression:

- When the following features (rain, precipi, Hour, mintempi, UNIT) were included in the regression, an R^2 value of 0.4649 was obtained. This means that according to this dataset, about 46.5% of the variation in entries on the subway can be attributed to the variation in the 5 features included in my regression model. This means that rain, precipitation, fog, minimum temperature, unit location, and hour of day do appear to play a substantial role in influencing how many people will enter the subway (measured every 4 hours), but we are still missing a lot of other important features in our model that would help explain the other 53.5% of the variation in hourly entries on the subway. Also, when I remove the two features rain and precept from the regression model, and only use (Hour, mintempi, UNIT) an R^2 value of 0.464875 is obtained, which is barely less than the 0.4649 obtained when rain and precipi were included as features. This would seem to indicate that presence of rain and amount of precipitation may have a tiny effect on how many people enter the subway, but it doesn't seem to be a very large effect, if it exists at all. Of course, this could be due to shortcomings in the dataset and the methods of analysis, which I will now discuss.

Section 5.

Reflection

5.1 Potential shortcomings:

1. Shortcomings in the dataset:
   - For example, the amount of precipitation could be recorded more precisely at each UNIT or Station. It seems to be a very rough measure of precipitation in inches because there are only 3 different values in the entire dataset .89, .5, and 0 inches. I suppose the NYC subway would have to install their own precipitation measuring devices to do that though.
   - It would be helpful to have a dataset that included more rainy days in it, so our models would have more real data to work with. What happens on only 2 rainy days (the days where .89 and .5 inches fell) may not really tell us much about larger trends.
   - It would be helpful to know the intensity of the rain at the time the rain reading was recorded, whether it was a light drizzle vs. a heavy downpour.
2. Shortcomings in the methods of analysis:
   - Shortcomings of the Mann Whitney U test
     - It is a non-parametric test, so it has a lower power (less likely to find a difference) than a parametric test, but due to the non-normality of our data distribution, we have to accept this sacrifice in statistical power.
   - Shortcomings of the linear regression model used (gradient descent)
     - I could not seem to improve the accuracy of the model's predictions (R squared) beyond .4649, partly due to imperfections in the dataset, but also perhaps because there are other important variables (features) out there that need to be measured that weren't included in the model.
     - As far as weaknesses in gradient descent as a method, as opposed to other optimization algorithms, here are a couple opinions I found searching the web:
     - From Quora: "As far as optimization algorithms go, gradient descent is at best a passable one (the fact is, it has a relatively poor convergence rate, and can potentially zigzag near the optimum, although line searches and other techniques can help alleviate such problems). However, in many big data applications, the bottleneck tends to be the problem size rather than the algorithm used for optimization, so when viewed in this light, gradient descent can be a decent option…"
     - From Wikipedia: Limitations - "For some of the above examples, gradient descent is relatively slow close to the minimum: technically, its asymptotic rate of convergence is inferior to many other methods. For poorly conditioned convex problems, gradient descent increasingly 'zigzags' as the gradients point nearly orthogonally to the shortest direction to a minimum point."