Section 1

1. **Two-tailed Mann-Whitney's U-Test** was used to analyse the NYC subway data. We wanted to know if raining would cause number of hourly entries to decrease or even increase, thus **two-tailed P-value testing** was used. The null hypothesis here is **"Raining does not have an effect to number of hourly entries."** while alternate hypothesis is **"Raining does increase / decrease the number of hourly entries."**. P-critical value used here is **0.05**.
2. Mann-Whitney's U-Test is applicable to the test since, unlike Welch's t-test, **it can be used for non-normalised data[1]** [1], in other words it does not assume the distribution of ridership is normal.
3. The result I got from this statistical test was:
    a. p-value (two-tailed) = **0.05 (rounded up)**
    b. Mean of ridership when raining: **1105.45**
    c. Mean of ridership when not raining: **1090.28**
4. Since the p-value is **0.05 (rounded up)**, which is **smaller than p-critical value 0.05**, we can conclude that raining does cause a change in subway ridership (i.e. we can reject the null hypothesis). The change is of positive direction, as suggested by **mean of ridership when raining > when not raining**.

Section 2

1. **Linear regression with gradient descent** and in optional training 3.8 **OLS with stat models** were used to compute the coefficient theta and produce prediction.
2. After trying out adding all available features, I decided to go back on using the four original features: **'rain', 'precipi', 'Hour', 'meantempi'**, and in addition, dummy variables are added as part of the features; They are made out of **UNIT** field of given data.
3. By adding all features, r-squared increased from **0.464** to **0.466**. As they did not drastically improve the prediction, I reverted to using the original four features.
4. 2.92398062e+00, 1.46526720e+01, 4.67708502e+02, -6.22179395e+01
5. 0.463968815042
6. This is an interesting question. The thing is we can't really use R^2 to find out if our model is really a good fit for our dataset, simply because we do not have a way to find out what a good or bad R^2 values are.

---

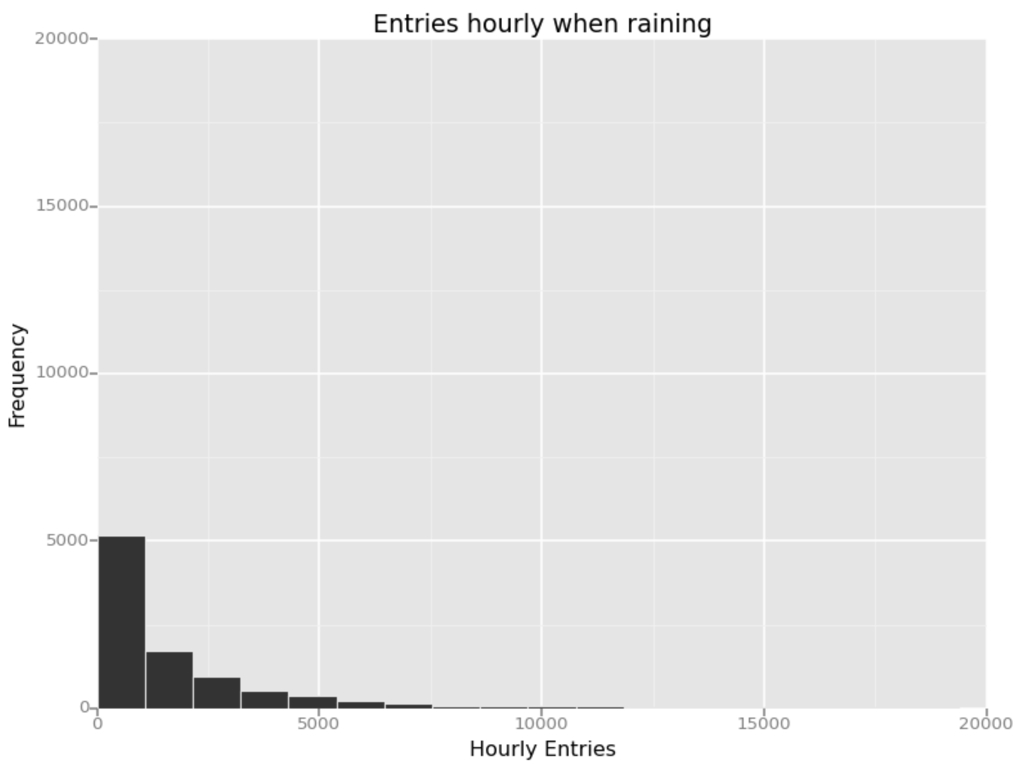
[1] Quoted from a paper by Fay, Michael P.; Proschan, Michael A.:

"In light of the difficulty in distinguishing between normality and those t-distributions with moderate sample sizes, and in light of the relative efficiency results that showed that the WMW is asymptotically more powerful for t-distributions with degrees of freedom less than 18, it seems that in general the WMW test will often be asymptotically more powerful than the t-test for real high quality data."

What we can do though is to make a residual plot of our model vs data, where basically we plot the distance each actual data point to our model, to find out whether the errors happen at random (i.e. stochastics error) - which means the model first well, or if the errors have pattern in it, which is a sign that our model causes systematic errors, which is bad. Unfortunately we are not being taught on how to plot residuals at this course yet, so I won't overwhelm my submission with that, but more information is available in the footnote.[2]

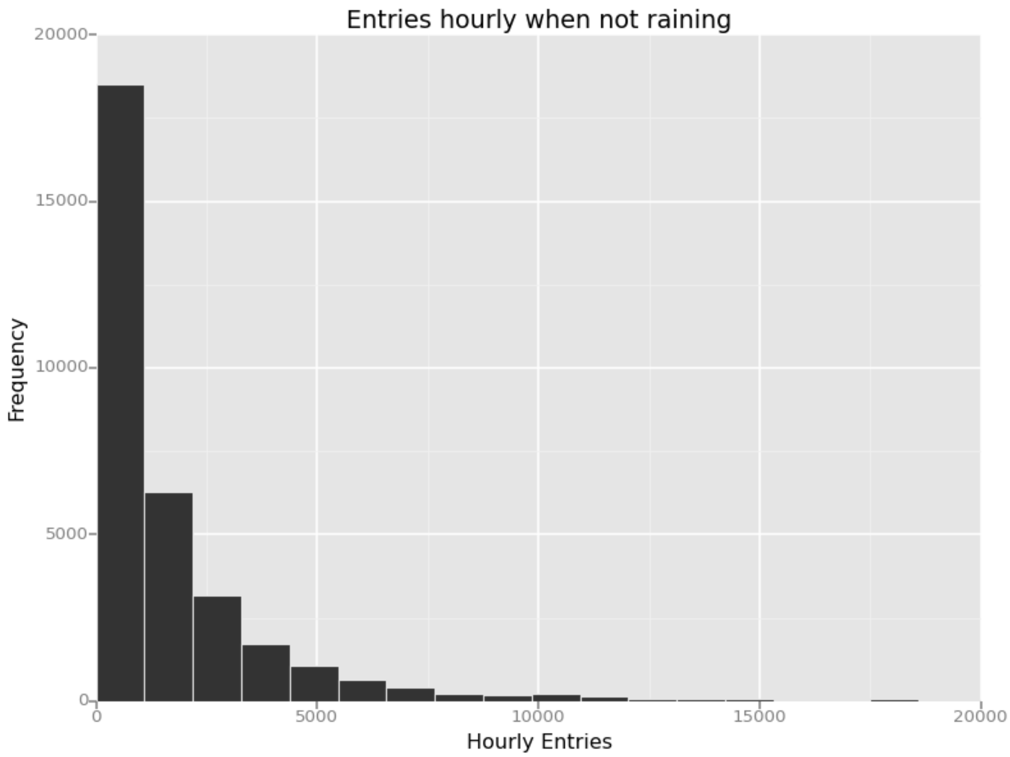
Section 3

1. Following are the histograms:

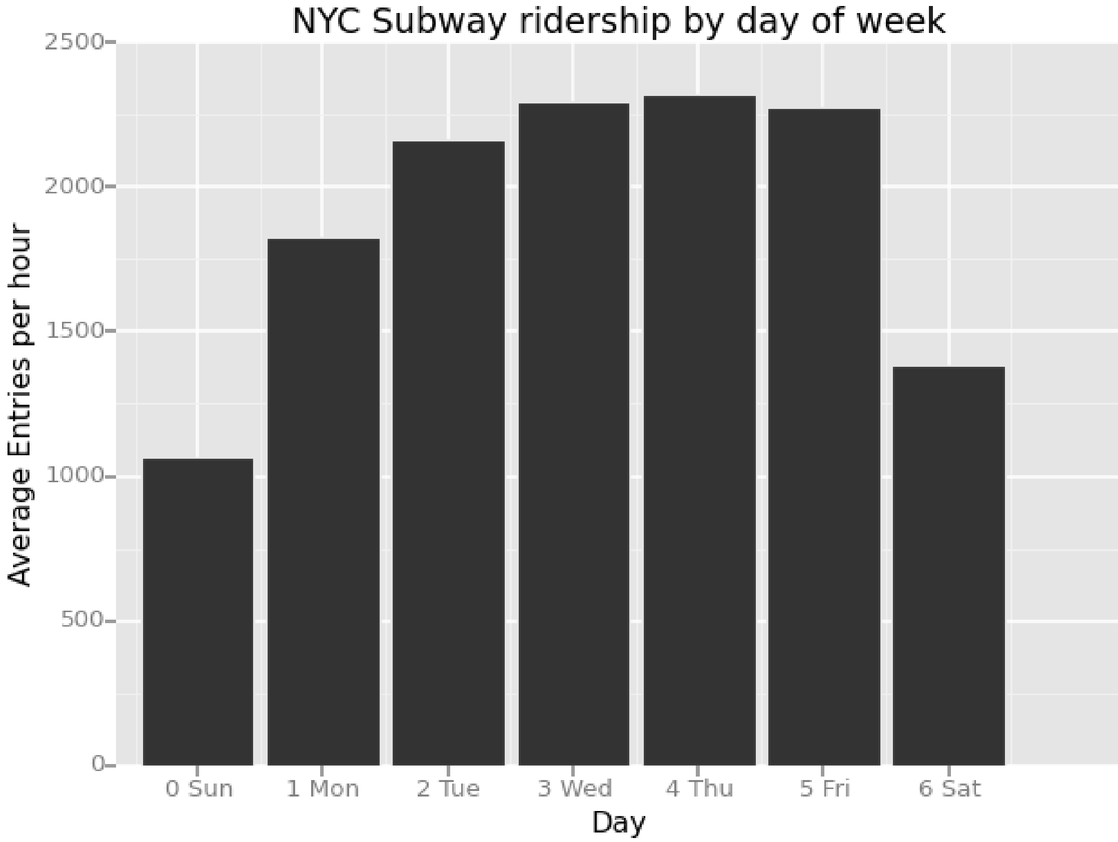


[2] This article explains that you can't use R^2 as a trusted source to validate your model: http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

And this article shows how you could use residual plot instead to validate the model. It is a very interesting read: http://blog.minitab.com/blog/adventures-in-statistics/why-you-need-to-check-your-residual-plots-for-regression-analysis



2. Entries by days of week:



Section 4

1. People seemed to ride NYC subway when it is raining, as shown in the analysis provided in section 1, where mean of ridership when raining is **1105.45**, larger than the mean of ridership when not raining (**1090.28**). But even with that information we still could not conclude the hypothesis that people do ride NYC subway more when it is raining is correct. We then conducted a Mann-Whitney's U-test to ensure if that hypothesis is indeed likely to be true.

   From the test we get that the p-value is **positive 0.05 (rounded up)**, which is **smaller than p-critical value 0.05**, and thus we can conclude that **null hypothesis can be rejected**, in other words raining does cause a change in subway ridership. The change is of positive direction, as suggested by **mean of ridership when raining > when not raining**.

2. From statistical test the explanation is as described above. From linear regression, we found out that the weight for **rain** feature was **2.92398062e+00**. Since this is a positive value, we can be sure that rain does positively affect ridership i.e. all other things being equal, when it is raining we can expect ridership to be larger.

Section 5

1.1. Unfortunately the dataset does not account for more demographic information like how many people are there around the area of each UNIT, and the age of that population. That demographic information could be useful since one can always expect that number of ridership within a certain area is larger when that area contains more population of people old enough to ride trains.

Should this be a more serious project, latitude and longitude can be used to find out the location and we should be able to find number of population in the radius of several hundred meters around each UNIT position. We could do that by combining the data of this dataset with another dataset containing population per area[3].

1.2. Linear regression is not a good tool to model this dataset, as shown by the low Coefficient of determination value. Polynomial regression would likely yield a better fit to the data.

Personally, in addition to demographic information as stated above I would enter the features into a Support Vector Machine or Artificial Neural Networks to predict whether a combination of features may result in a ridership or no ridership, and maybe use Bayesian Network to get more understanding of the features used in this dataset.

A similar approach has been used to predict bus ridership, where they used Bayesian inference technique, namely Markov Chain Monte Carlo (Monte Carlo is just a fancy name for “random”) for learning the parameters, and then use Gaussian Process Model to train the system. [2]

2. I hope we’d get to learn more about Bayesian Network later in this Nanodegree program!

References

[1] Fay, Michael P.; Proschan, Michael A. (2010). "Wilcoxon–Mann–Whitney or *t*-test? On assumptions for hypothesis tests and multiple interpretations of decision rules". *Statistics Surveys* **4**: 1–39. doi:10.1214/09-SS051. MR 2595125. PMC 2857732. PMID 20414472.

[2] Sourav Bhattacharya , Santi Phithakkitnukoon , Petteri Nurmi , Arto Klami , Marco Veloso , Carlos Bento, Gaussian process-based predictive modeling for bus ridership, Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication, September 08-12, 2013, Zurich, Switzerland doi:10.1145/2494091.2497349

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

[4] Jim Frost. (2015). http://blog.minitab.com/blog/adventures-in-statistics/how-high-should-r-squared-be-in-regression-analysis

---

[3] Sites like http://factfinder.census.gov/, http://www.data.gov/, and https://www.openstreetmap.org/ may provide the required demographic dataset to aid our research.

[5] Jim Frost. (2013). http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
[6] Jim Frost. (2012). http://blog.minitab.com/blog/adventures-in-statistics/why-you-need-to-check-your-residual-plots-for-regression-analysis