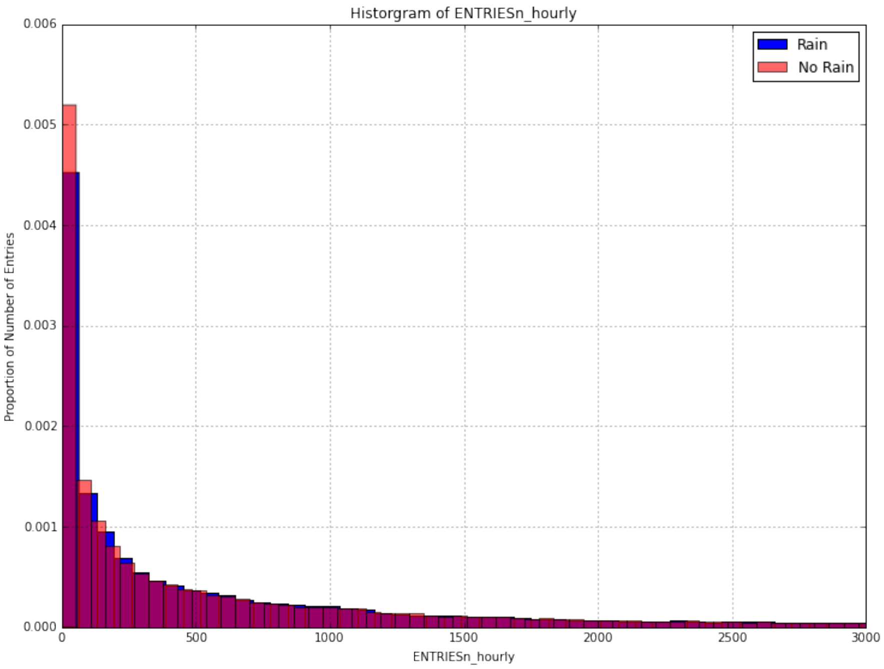
Section 1: Statistical Test

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?**
   We used the Mann-Whitney U-Test (scipy.stats.mannwhitneyu). According to SciPy.org the p-value is one-sided which means we need to multiple the results by 2 for a two-sided test. Two-sided results were reported below.

   What is the null hypothesis?
   The null hypothesis is that the two populations (rain/no-rain) are the same.

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 Mann-Whitney U test is used to compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed…" which is the case for the data investigated (see figure below). "The Mann-Whitney U test is often considered the nonparametric alternative to the independent t-test."

   - *Assumption #1*: Dependent variable should be measured at the ordinal or continuous level. (Number of hourly entries)
   - *Assumption #2*: Your independent variable should consist of two categorical, independent groups. (Subway riders during rain versus no rain)
   - *Assumption #3*: You should have independence of observations, which means that there is no relationship between the observations in each group or between the groups themselves. (Not sure this is 100% true since the riders in the rain may also be riding when there is no rain as well.)
   - *Assumption #4*: A Mann-Whitney U test can be used when your two variables are not normally distributed. However, in order to know how to interpret the results from a Mann-Whitney U test, you have to determine whether your two distributions (i.e., the distribution of scores for both groups of the independent variable; for example, 'males' and 'females' for the independent variable, 'gender') have the same shape. (Both are positively skewed in the test case of subway riders, see plot below.)



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.**

   Results from the statistical test (exercise 3.3):
   - Mean of hourly entries with rain: 1105.45
   - Mean of hourly entries without rain: 1090.28
   - U: 1,924,409,167
   - p: 0.05 (two-sided)

4. **What is the significance and interpretation of these results?**

   According to the results from exercises 3.3 and 3.4 the distribution of the number of entries is statistically different between rainy and non-rainy days. However, further investigation offline presented some additional facts that are discussed below.

   First, for any Mann-Whitney U test, the theoretical range of U is from 0 (complete separation between groups, null hypothesis (H0) most likely false and alternate hypothesis (H1) most likely true) to n1*n2 (little evidence in support of alternate hypothesis (H1)). In every test, we must determine whether the observed U supports the null or alternate hypothesis. Specifically, we need to determine a critical value of U such that if the observed value of U is less than or equal to the critical value, we reject H0 in favor of H1 and if the observed value of U exceeds the critical value we do not reject H0.

For large populations you can assume a normal distribution for U. This leads to the following assumptions for μ and σ.

$$\mu = n_1 n_2 / 2$$

$$\sigma^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

Given this information we can in turn calculate z-score using the following formula:

$$z = \frac{x_i - \mu}{\sigma}$$

We can also find the associated proportion from a pre-calculated z-table. For purposes of these tests we assumed p needs to be 0.05 or less for the alternate hypothesis to be true. The alternate hypothesis is the two populations (rain/no-rain) are different.

Looking this up on the z-table this results in a z-score of -1.6425 to demonstrate the alternate hypothesis is true (null hypothesis is false).

To calculate $U_{critical}$ we can use the formula below:

$$U_{critical} = z\sigma + \mu$$

Based on these formulas the results from looking at the population of entries when it rains versus the population when it does not the U value calculated is less than $U_{critical.}$ This demonstrates the validity of the Mann-Whitney U test and indicates the alternate hypothesis is true (95% confidence).

- U: 1,924,409,167
- $U_{critical}$: 1,926,481,293

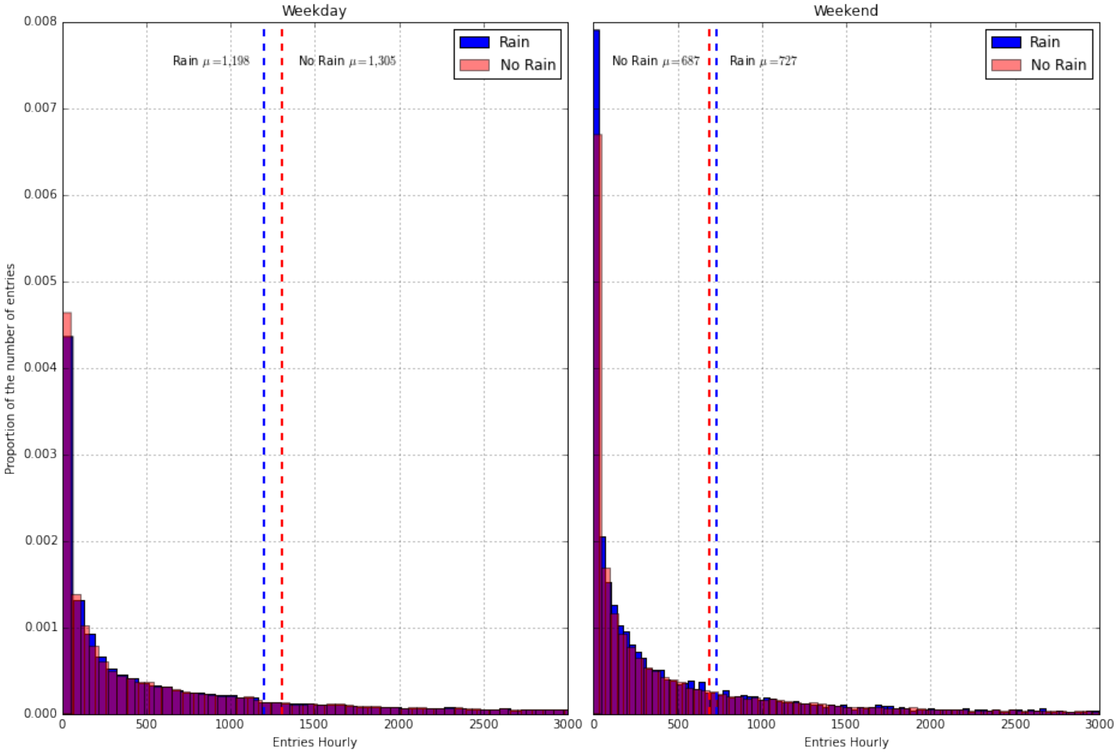
5. Additional observations

If you partition the data into weekday observations and weekend observations the results take on slightly different patterns. Notice the differences between weekday and weekend observations. Compare this with the observations taken from the class when considering the entire week (for a month's worth of data).

Results μ and from the Mann_Whitney U test for different experiments using May 2011 Subway Data combined with weather information.

	μ of hourly entries with rain	μ of hourly entries w/o rain	U < $U_{critical}$	p <= 0.05	$U_{critical}$
Entire week	1105	1090	✓ 1,924,409,167	✓ 0.05	1,926,481,293
Weekdays	1198	1305	✓ 985,531,036	✓ 6.374E-15	1,009,932,347
Weekend	727	686	✓ 129,024,196	✓ 0.0006	130,677,630

In terms of a visual representation this is also depicted in the plots below. Again notice the mean for ridership when it rains on weekends versus during the week (note this pattern actually reverses when you look at dataset v2. See section 5)



So, what does this translate to? It does appear there is a difference in ridership (small) for rainy versus no rain days when looking at the total riders for the month of May 2011. Additional exploration as well as more data may provide a more definitive conclusion.

Section 2: Linear Regression

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 (see item 6+ at the end of the section)
   - Or something different?

2. **What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?**
   - Features used: [rain, fog, Hour, mintempi, maxtempi, meanpressurei, meanwindspdi]
   - Dummy_features: Units

3. **Why did you select these features in your model?**

   Weather associated parameters seemed to have an impact on the R^2 value. The assumption was that deviations from "nice" weather would increase ridership. The features that best represented this seem to be rain (yes/no), fog (yes/no), min and max temperature, mean pressure (pressure variations are also associated with different types of weather) and wind speed (also another sign of bad weather).

   These features plotted against the hour/time of day should show differences in the number of subway riders.

   After running various experiments it seems min and max temperature, as well as wind speed and pressure had the largest effect on increasing R^2. This again correlates with the hypothesis that bad weather does affect subway ridership.

4. **What is your model's R^2 (coefficients of determination) value?**
   R^2 as calculated via R^2 formula from Exercise 3.7 and data/predictions provided by Exercise 3.5
   - 0.4575 ← from using the hour
   - 0.4576 ← from using rain (yes or no) and the hour.
   - 0.4577 ← from using rain (yes or no), fog (yes or no), and the hour.
   - 0.4588 ← adding max and min temperatures.
   - 0.4594 ← adding mean pressure and mean wind speed.

   R^2 as calculated in Exercise 3.5
   - 0.4742 ← final as calculated in Exercise 3.5.

5. 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?

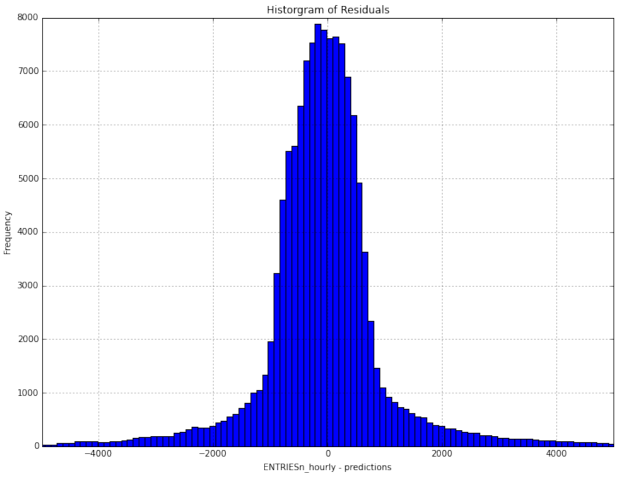
When interpreting R^2 the values range between 0 and 1 with numbers closer to 1 being better. In my particular case both the values calculated for R^2 (using Exercise 3.5 and Exercise 3.7 with my data) were below 50% or 0.5. For cases involving human behavior this does not necessarily indicate the model is bad. The recommendation from exercise 3.5 was to obtain a R^2 of 0.20 or greater which was accomplished.

One problem with using R^2 as a measure of model validity is that adding more variables into the model can always increase it. To some extent this behavior was seen when adding features such as "meandewpti". The increase in R^2 associated with adding this feature was 0.0001. It was subsequently removed with minimal change.

Another method to determine the "goodness" associated with the model is to look at plots of the residuals. "Residuals can be thought of as elements of variation unexplained by the fitted model. Since this is a form of error, the same general assumptions apply to the group of residuals that we typically use for errors in general: *one expects them to be (roughly) normal and (approximately) independently distributed with a mean of 0 and some constant variance*."

Stated another way, if the residuals appear to behave randomly, it suggests that the model fits the data well. On the other hand, if non-random structure is evident in the residuals, it is a clear sign that the model fits the data poorly.

Looking at the histogram of the residuals for the model used the errors are roughly normal with the bulk of the errors occurring in the -2500 to 2500 range with the greatest frequency ~ -1000 to 1000.

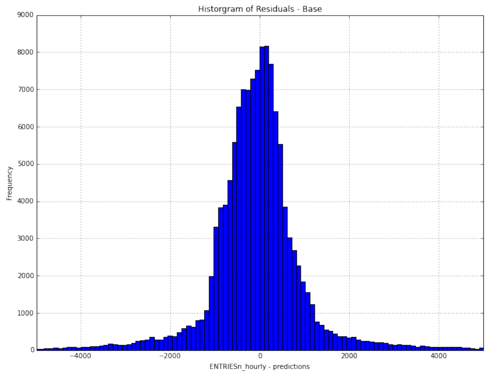


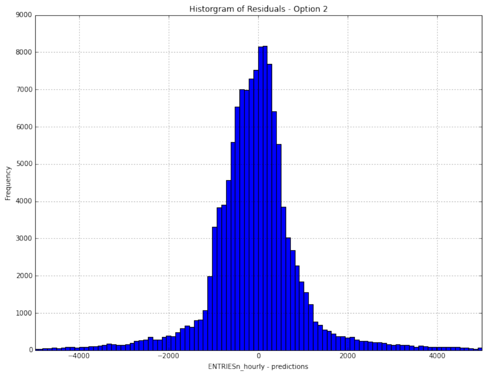
6. Using other methods to find the coefficients theta and produce predictions for ENTRIESn_hourly in a regression model – OLS

In addition to using gradient descent I also explored using OLS by using Statsmodels. I looked at the same feature combinations as what was done for the gradient descent case as shown in the table below. I also explored several other options. From the weather based tests the values seemed to peek at ~0.46. Doing some additional tests with day of the week and date yielded better results (0.46 – 0.47). If the actual number of exits where known this resulted in the model with the best R^2.

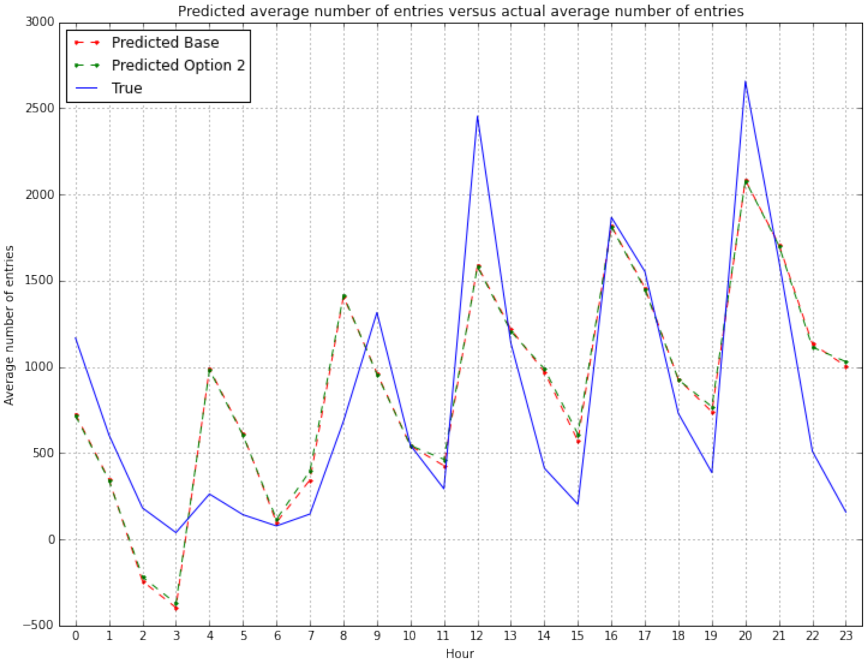
Test	Features	R^2 Gradient	R^2 OLS
Weather based tests			
1	Hour + Unit	0.458	n/a
2	Hour + rain + UNIT	0.458	0.458
3	Hour + rain + fog + UNIT	0.458	0.458
4	Hour + rain + mintempi + maxtempi + UNIT	0.459	0.459
5	Hour + rain + mintempi + maxtempi + meanpressurei + meanwindspdi + UNIT	0.459	0.459
Additional tests			
6	Hour + rain + EXITs_hourly + UNIT	n/a	0.621
7	Hour + rain + dayofweek + UNIT	n/a	0.463
8	Hour + rain + DATEn + UNIT	n/a	0.471

Similar to what was done for the gradient decent model the residuals for the OLS models (1 & 8) were plotted in a histogram. Looking at the histogram of the residuals for the models the errors seen are roughly normal with the bulk of the errors occurring in the -2500 to 2500 range with the greatest frequency ~ -1500 to 1500 which provides some confidence in the validity of the model.





After the calculations for R^2 were completed for the various models and residuals reviewed, plots for the predicted values and the actual values were developed. Two options were chosen to plot. The first was the base model (2) which used hour, rain and unit number as the features. The second was the model that added date as another feature (8). Since the number of predicted and actual values exceeded 130,000 each the mean for each hour was calculated and used in the plots instead. Results from the comparison of predicted to actual are shown below. In general both OLS models tend to overstate minimums and understate maximums when compared to the actual values measured. To develop a better model other features may need to be identified or an alternate approach for linear regression chosen.

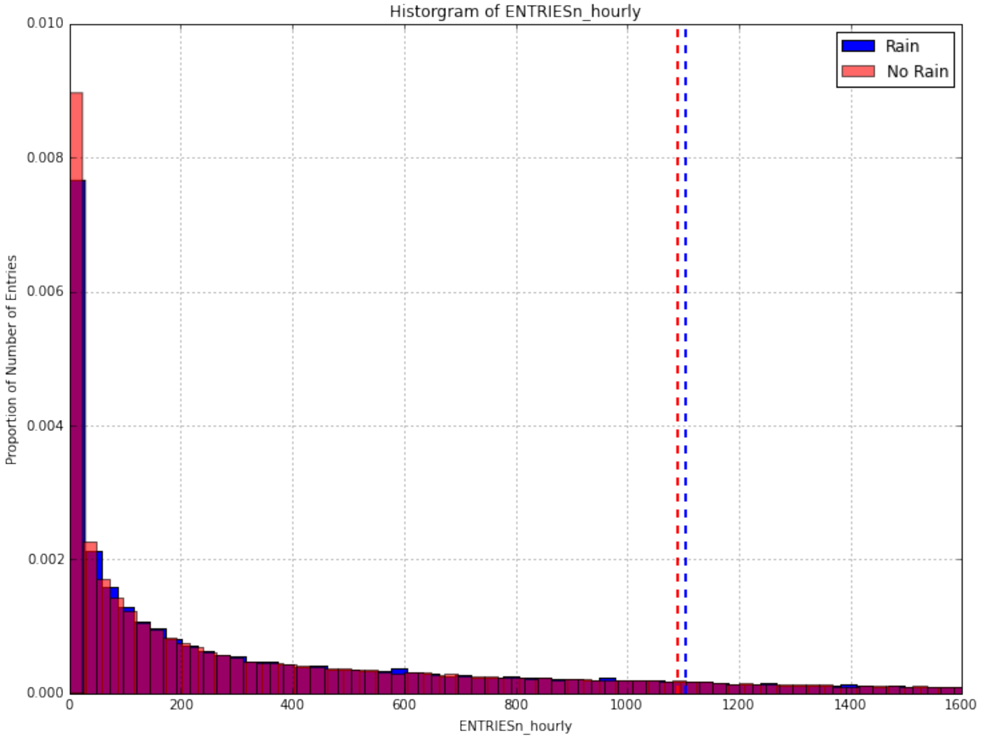


Section 3: Visualization

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.

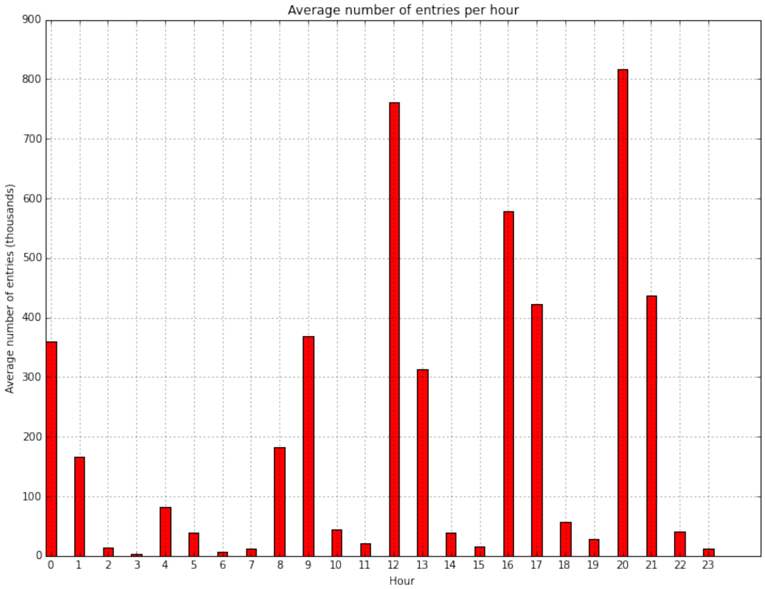
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

1. **Histogram of NYC subway ridership for rainy versus non-rainy days taken during the month of May 2011**

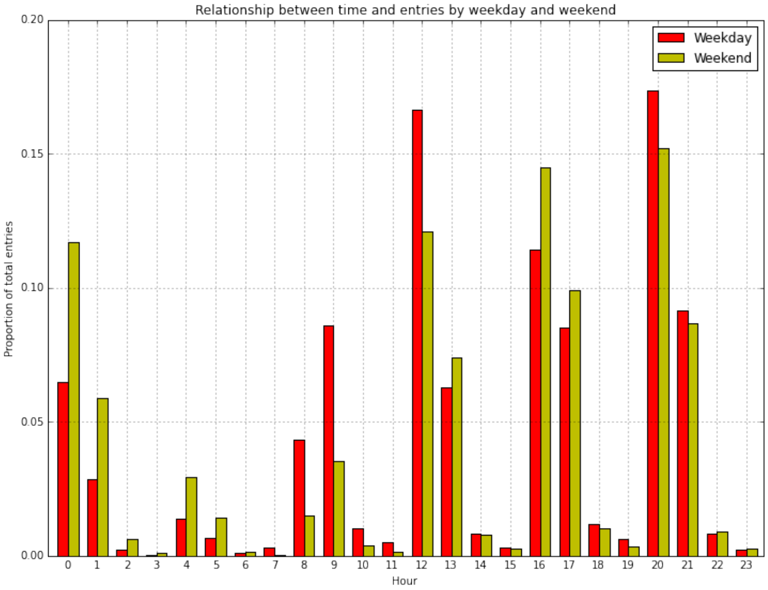


The histogram of NYC subway ridership shows there is an increase in entries during days where it rained. Despite having fewer overall riders (~50% of the number of entries for no rain days) the histograms shown above depicts an increase in the mean for the entries with rain versus no rain.

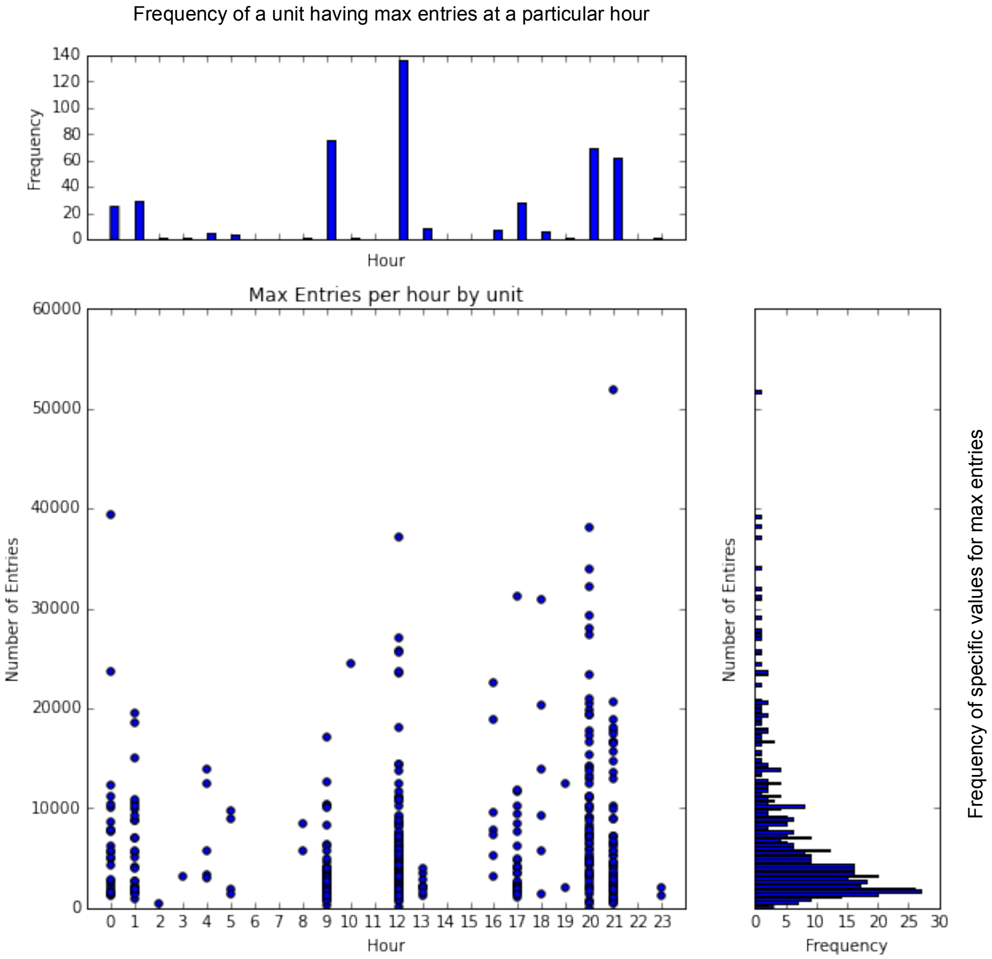
2. Average ridership by time of day during the month of May, 2011



The plot above provides a visualization of the average number of entries per hour gathered over a 1 month time period. From looking at the bars in the red plot patterns in ridership can be seen. Morning rush hour, lunchtime, evening rush hour and two peaks associated with perhaps a combination of people going/returning for entertainment and traveling to and from work.



The plot above also reinforces the hypothesis that some of the late night/early morning entries are entertainment related. See the plot above that breaks down the proportion of entries per hour by weekday and weekend views.



The three plots above provide different visualizations of the number of entries based on maximum entries per hour by unit. These plots also mirror the previous entries per hour plots above. The hours with the largest number of maximum entries per unit align well with the highest entry times per hour shown previously. The plot on the right provides some additional information by indicating that the maximum unit entries seem to be clustered below 10,000 entries per hour. The spikes in maximum entries up to 50,000+ could be attributed to events or other special occasions. More exploration would be required to confirm this however.

Section 4: Conclusion

1. **From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining versus when it is not raining?**
   Based on the analysis conducted as part of this course and the associated project it does appear that on average slightly more people do ride the subway when it rains. If the results are categorized into riders during the week and during weekends it appears that the weekend riders drive the general results (more riders when it rains). Perhaps this is due to weekend ridership being associated more with pleasure/leisure trips versus work related? When presented with alternatives more riders appear to use the subway when it rains. Additional investigation as well as more data is required to figure out the specific patterns associated with NYC subway ridership.

2. **What analyses lead you to this conclusion?**
   The tests employing the Mann-Whitney U test as well as calculations of the mean for rain and no rain subway ridership provided most of the support for the conclusion. Visualizations of the proportion of riders for the various days and weather conditions also conveyed the point in an easier to consume and understand manner.

Section 5: Reflection

1. **Please discuss potential shortcomings of the data set and the methods of your analysis.**
   I think one of the shortcomings associated with the data was the amount available. As mentioned previously, the data used corresponded to the month of May in 2011. Ideally more data (longer time period) would be beneficial to evaluating patterns in ridership. Adding information about events and other happenings would also be useful to help explain spikes or dramatic decreases. At present it's hard to definitively say the affect of rain.

2. **(Optional) Do you have any other insight about the dataset that you would like to share with us?**
   I think additional exploration sub-dividing the data will lead to other insights associated with the subway and weather data. For instance, the area I explored associated with ridership by weekday and weekend for rainy and no rain days. There are still several patterns present that are worth exploring (e.g. mean of rain and no rain for weekdays, weekends)

   Using the additional dataset (v2) several other interesting facts were seen. For instance the pattern associated with rain/no rain for weekdays and weekends seems to have reversed. There is also a bigger delta between rain/no rain ridership for the entire week. This is based on fewer observations overall however.

	μ of hourly entries with rain	μ of hourly entries w/o rain
Entire week	2028	1846
Weekdays	2228	2134
Weekend	1092	1226

   So my original statement above having more/additional data may help clarify ridership patterns.

Section 6: References

1. “Linear Regression” *Linear Regression — Statsmodels 0.7.0 Documentation*. 14 Dec. 2014. Web. 11 Jan. 2015.
2. "Quickstart." *Quickstart — Patsy 0.3.0-dev Documentation*. Nathaniel J. Smith, 15 Apr. 2013. Web. 11 Jan. 2015.
3. "Pandas: Powerful Python Data Analysis Toolkit" *Pandas: Powerful Python Data Analysis Toolkit — Pandas 0.15.2 Documentation*. Web. 10 Jan. 2015.
4. "NumPy" *NumPy — Numpy*. Web. 13 Jan. 2015.
5. "Mann–Whitney U Test." *Wikipedia*. Wikimedia Foundation. Web. 9 Dec. 2014.
6. "Linear Regression." *Wikipedia*. Wikimedia Foundation. Web. 2 Jan. 2015.
7. "Ordinary Least Squares." *Wikipedia*. Wikimedia Foundation. Web. 11 Jan. 2015.
8. "Statistical Help." *StatsDirect*. StatsDirect Limited. Web. 2 Jan. 2015.
9. "Why You Need to Check Your Residual Plots for Regression Analysis: Or, To Err Is Human, To Err Randomly Is Statistically Divine | Minitab." *Why You Need to Check Your Residual Plots for Regression Analysis: Or, To Err Is Human, To Err Randomly Is Statistically Divine | Minitab*. Web. 13 Dec. 2014.
10. "5.2.4. Are the Model Residuals Well-behaved?" *5.2.4. Are the Model Residuals Well-behaved?* Web. 12 Dec. 2014.
11. "Introduction." *Matplotlib: Python Plotting — Matplotlib 1.4.2 Documentation*. Web. 18 Dec. 2014.
12. "Ggplot from ŷhat." *Ggplot*. Web. 17 Dec. 2014.