Analyzing the NYC Subway Dataset

# 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?

The NYC subway data was analyzed with the Mann-Whitney U-Test. A one-tail P value was used. The P value returned by **scipy.stats.mannwhitneyu** is one-sided as noted here: <http://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.mannwhitneyu.html>

1. Why is this statistical test appropriate or applicable to the dataset?

The overarching question being asked within this course is whether subway ridership varies with the weather. The provided dataset allows hourly entries to MTA turnstiles to be spliced into two different samples, entries with and without rain. The Mann-Whitney U-Test tests the null hypothesis that the two samples being compared are derived from the same population. This null hypothesis allows us to test whether there is a statistically significant difference in ridership on rainy and non-rainy days (i.e., are the hourly entries derived from the same population). Furthermore, exploratory data analysis (namely the histogram generated in Problem 3.1) has shown that the data is not normally distributed. The Mann-Whitney U-Test does not assume normality of the data, making this test appropriate.

1. What results did you get from this statistical test? These should include the numerical values: p-values, as well as the means for each of the two samples under test.

The numerical results (rounded to three decimal places) of the Mann-Whitney U-Test are the following:

|  |  |
| --- | --- |
| Statistic | Value |
| Mean entries, with rain | 1105.446 |
| Mean entries, without rain | 1090.279 |
| U – Mann-Whitney test | 1924409167.000 |
| p-value | 0.025 |

1. What is the significance of these results?

With this small p-value, we reject the null hypothesis of the Mann-Whitney U-Test. In other words, the distribution of the number of entries is statistically different between rainy and non-rainy days.

# Section 2. Linear Regression

1. What approach did you use to compute the coefficients theta and produce prediction in your regression model:
   1. Gradient descent (as implemented in exercise 3.5)
   2. OLS using Statsmodels
   3. Or something different?

In my regression model I implemented **Ordinary Least Squares**. I implemented the normal equation directly, and did not use the Statsmodel package. I referenced the Machine Learning course taught by Stanford’s Andrew Ng. The lecture material can be found here: <https://d396qusza40orc.cloudfront.net/ml/docs/slides/Lecture4.pptx>

1. What features did you use in your model? Did you use any dummy variables as part of your features?

I used the following features directly from the provided dataset:

* rain
* fog
* Hour
* mintempi
* meantempi
* maxtempi

I created the following features by transforming data:

* weekday – The day of the week as an integer [0:6]
* precipi – The square root of the “precipi” field in the source dataset

I used dummy variables to tie the above chosen features to a given UNIT.

1. 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.

We proved earlier that hourly entries are statistically different among rainy and non-rainy days. As such, I chose features related to the rain (namely ‘rain’ and ‘precipi’). After “playing” with some polynomial features, I determined that I got the best fit by using the square root of ‘precipi’.

I chose to use mintempi, meantempi, and maxtempi as features as I believe temperature is a component of the weather that affects people’s decision making. A given temperature may affect how long and how much effort it takes to clothe in the morning. A person may simply choose to stay indoors due to discomfort with a given temperature.

Hour and weekday features were chosen as it is easily observed how ridership varies with time of day and day of week. A phenomenon supporting the Hour feature is the daily rush hours that mass transit systems and roadways exhibit and accommodate. An observation supporting the use of weekday is how the MTA train schedule varies between weekdays and weekends, with trains arriving with more infrequency on Saturday and Sunday.

1. What is your model’s R2 (coefficients of determination) value?

The R2 of my model is 0.489812134989.

1. What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model is appropriate for this dataset, given this R2 value?

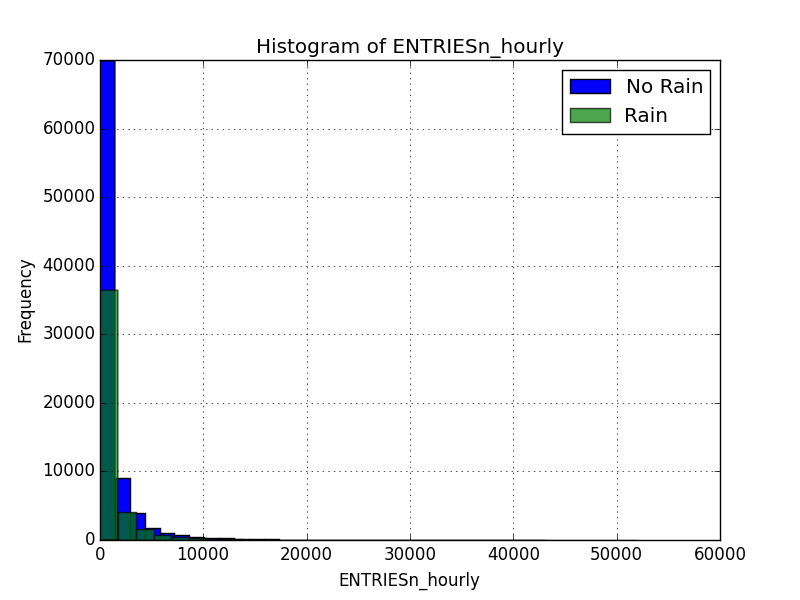
This R2 value is relatively low, meaning that the fit of this regression model is not very good. I do not believe that this linear model is appropriate given this low R2 value. Furthermore, the categorical nature of data in the given data set, such as ‘rain’, makes a linear model inappropriate.

# 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., scatterplots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

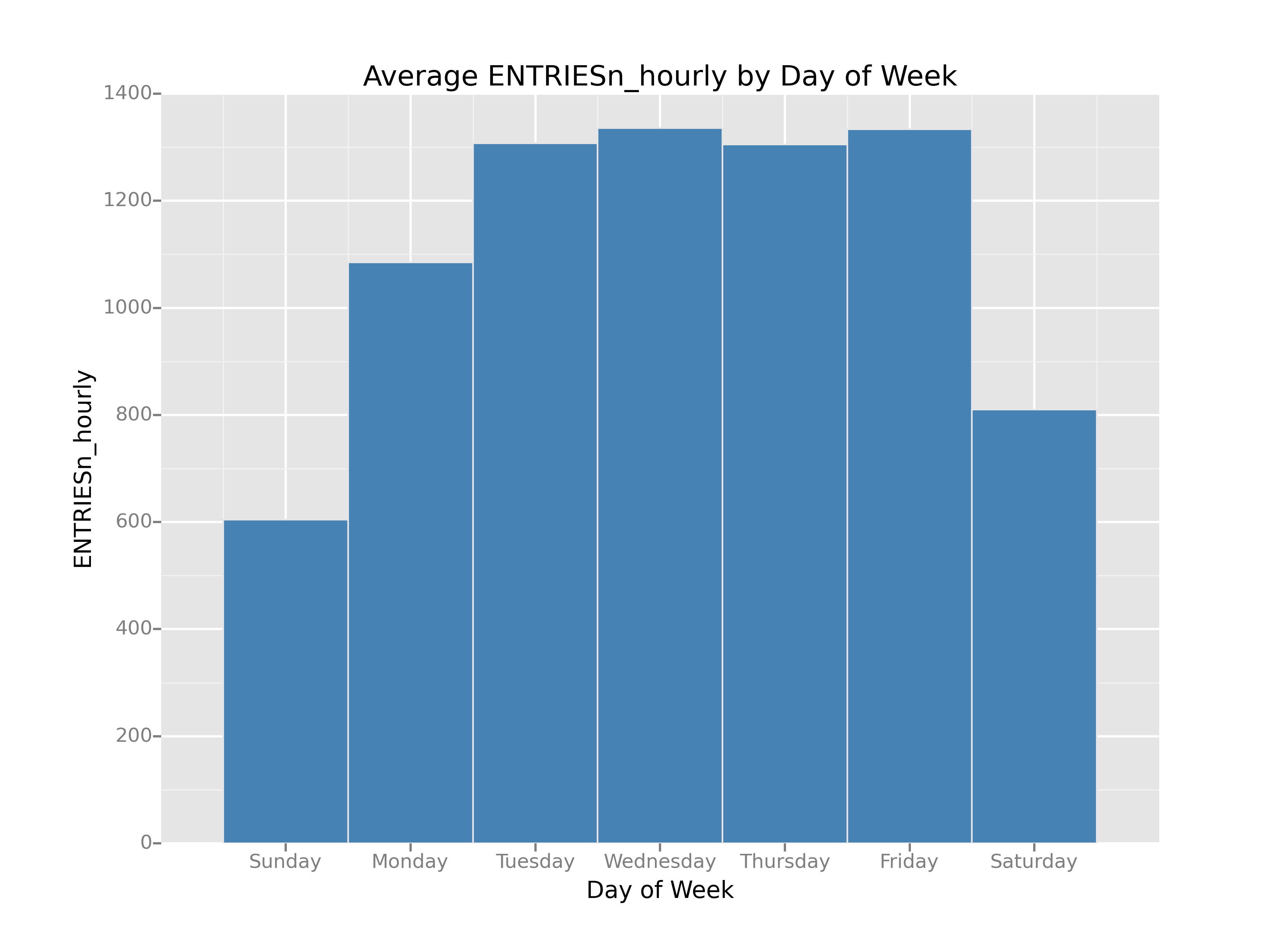
Remember to add appropriate titles and axes labels to your plots. Also please add a short description below each figure commenting on the key insights depicted in the figure.

1. One visualization should be two histograms of ENTRIESn\_hourly for rainy days and non-rainy days. Remember to increase the number of bins on the histogram (by having larger number of bars, each with smaller width). The default bin width is not sufficient to capture the variability in the two samples.



The distribution of ENTRIESn\_hourly appears to not be normally distributed and skewed to the right on both rainy and non-rainy days. The mode appears to be within the smallest bin for both distributions. There are far fewer observations on rainy days than non-rainy days. Though it is difficult to discern from this visualization, but it appears that the distribution of ENTRIESn\_hourly on rainy days is shifted slightly right relative to the distribution on non-rainy days.

1. One visualization can be more freeform, some suggestions are:
   1. Ridership by time-of-day or day-of-week
   2. How ridership varies by subway station
   3. Which stations have more exits or entries at different times of day



The above bar chart shows the average hourly ridership by day of week. The sum of ENTRIESn\_hourly by day of week was divided by the count of rows for a given day of week (as each row represents an hour’s worth of data). The bar chart shows that the average hourly ridership is higher on weekdays than weekends, with Saturday seeing significantly higher ridership than Sunday. It appears that the average hourly ridership on Monday is significantly different than the rest of the weekdays. This may be due to a seasonal effect of Monday holidays. The given data set is a sample from May 2011. There is at least one major holiday that falls on Monday in the month of May: Memorial Day.

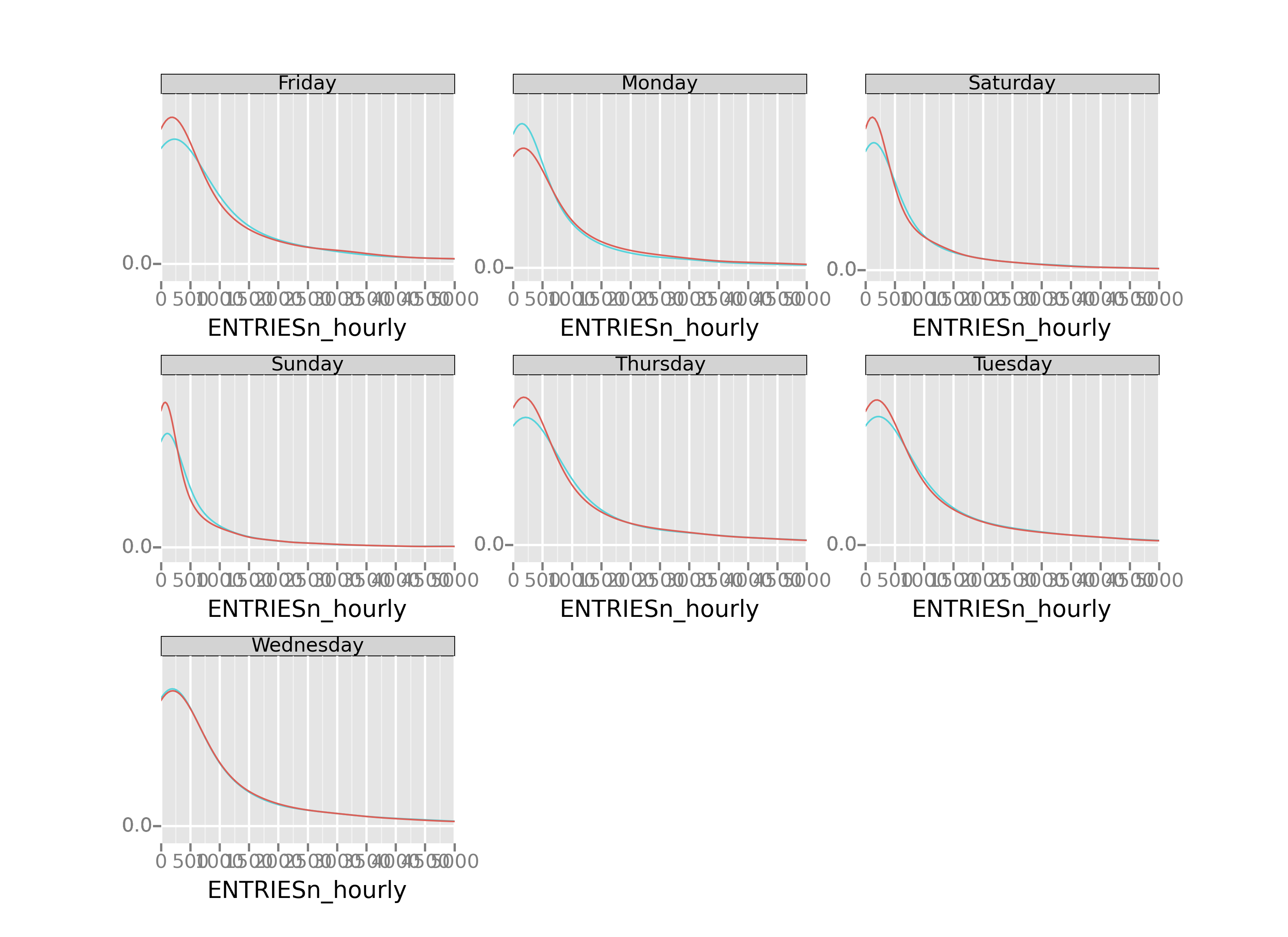
# Section 4. Conclusion

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 my analysis, I believe that more people ride the NYC subway when it is raining.

1. What analyses lead you to this conclusion?

As seen in Section 1 Question 3, the mean of ENTRIESn\_hourly is greater for hours with rain than without (1,105 vs. 1,090). Additionally, a Mann-Whitney U Test shows that the ENTRIESn\_hourly sample with rain appears to be drawn from a different distribution (or population) than the ENTRIESn\_hourly sample from hours without rain. In Section 3 Question 1 a histogram of ENTRIESn\_hourly with and without rain is shown. From this visualization, it appears that the distribution of ENTRIESn\_hourly on days with rain has a fatter right tail. In other words, this fatter tail implies higher ridership, with higher frequencies of ENTRIESn\_hourly in larger bins.

****

This faceted grid shows the density functions of ENTRIESn\_hourly with and without rain for each day of the week. Currently legends do not work within ggplot facets. Red represents without rain, and blue represents with rain. It appears for every day, except Monday, that the distribution of ENTRIESn\_hourly has a fatter right tail when there is rain.

# Section 5. Reflection

1. Please discuss potential shortcomings of the data set and the methods of your analysis.

The data set provided contains only one month of MTA data. This smaller data set is subject to effects of seasonality, as the time of year may also affect ridership. Additionally, the Month of May contains a Monday holiday, which appears to have an affect on the data in the visualizations in Section 3 Question 2 and Section 4. The largest shortcoming I see with this data set is that it appears that the MTA component of the data is produced on an hourly basis, but it is joined to daily weather data. For example, if it rained at any point in a given day, every hour of that day will reflect that it rained. This prevents a truly granular analysis of how the weather can affect ridership within a day.