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

Answer: **statistical test:** Mann-Whitney u-test (not a normal distribution) along with the mean of the two samples. **Two-tail test** given the fact that we don't know it the difference is greater or less than the average. **Null hypothesis:** There is no difference in Ridership on rainy days and ridership on days without rain. **P-critical** 0.025(two-tail),

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.

Answer: It's a non-parametric test; it doesn't assume the data is drawn from any particular probability distribution. The distributions of entries with and without rain are not normally distributed.

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

Answer: P-value: 0,012499956, rain: 1105.4464, without rain: 1090.2788

.

1.4 What is the significance and interpretation of these results?

Answer: The two samples come from the same population, and are critical at 0,012499956 which falls in the critical value zone established +-0.025 for the two tail test.

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:

1. **Gradient descent (as implemented in exercise 3.5)**
2. OLS using Statsmodels
3. Or something different?

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

Answer: Rain, precipi, maxtempi, Hour, UNIT. Yes rain is a dummy variables 0 represent the absence of rain and 1 the presence of rain and UNIT takes 0 or 1 if a given station holds a record(unit_R001 = 1 if R001 shows up on that particular record for the UNIT column).

2.3 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.

- Your reasons might be based on intuition. For example, response for fog might be: “I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often.”

- Your reasons might also be based on data exploration and experimentation, for example: “I used feature X because as soon as I included it in my model, it drastically improved my R^2 value.”

Answer: I used rain because it’s the most probable weather condition that would influence subway ridership and most common also(it can rain any time of the year.), maxtempi can indicate the temperature and what relationship temperature has on ridership. The hour variable helps pin-point the exact moment in which ridership changes and if the same hour with different weather conditions varies on ridership. Precipitation would be an interesting factor that might answer questions like, does the amount of rain influence ridership or is it just the fact that it rains without any variability in the amount being of importance. UNIT has a strong significance on the coefficient of determination output and makes sence to keep track of ridership per train station.

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

Hour: 412.067 **meantempi:** -52.441 **precipi:** 21.883

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

Answer: 0.463968977315, meaning that with this model we have explained 46% of the original data, and have a residual variability of 54%. But the R^2 value in this case is almost irrelevant, given the fact that whats being measured is the relationship between the predictors and response variable.

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 R^2 value?

Answer: I think this coefficient of determination shows a value closer to 0, making it not appropriate, alpha value and features should be changed in order to get a better R2 value.

Section 3. Visualization

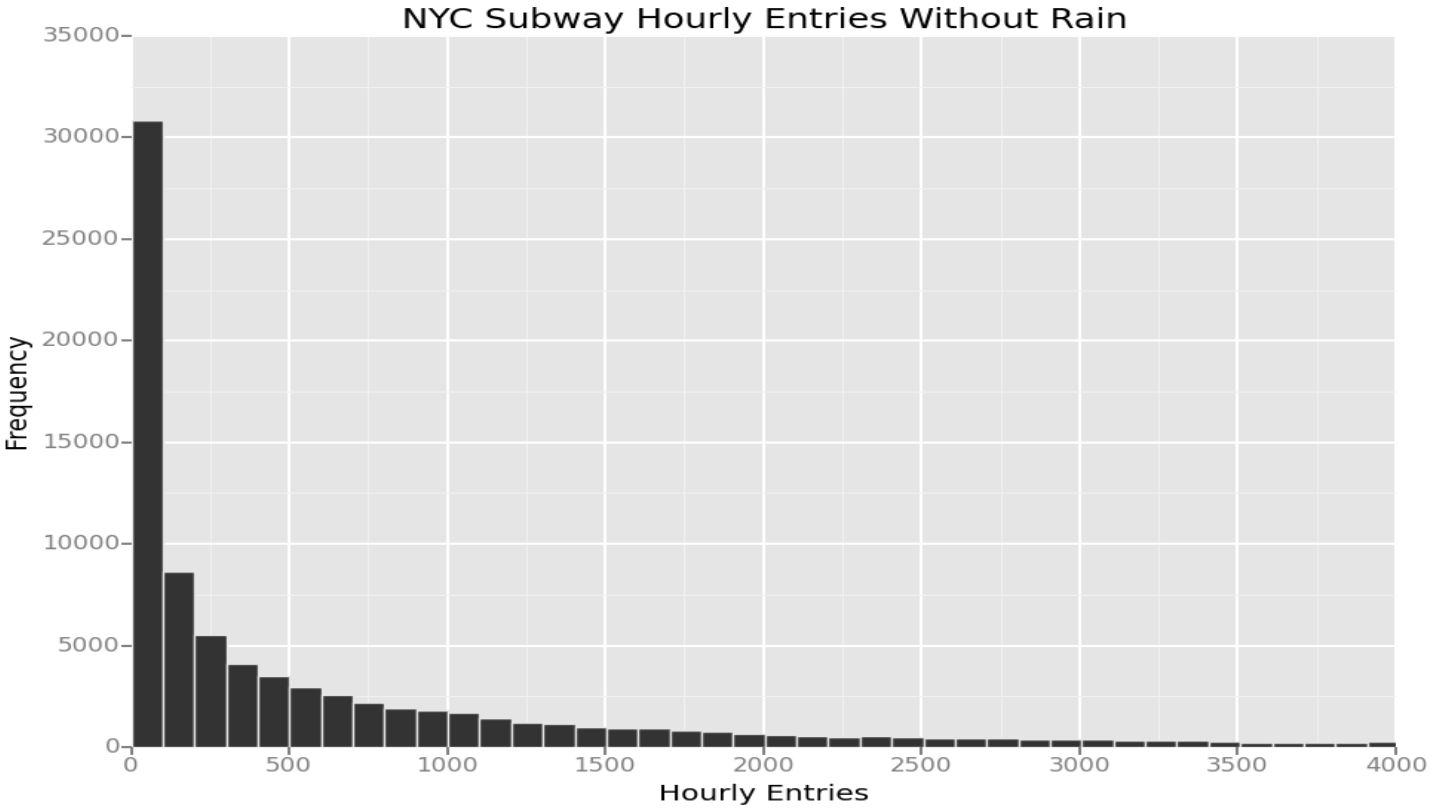
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

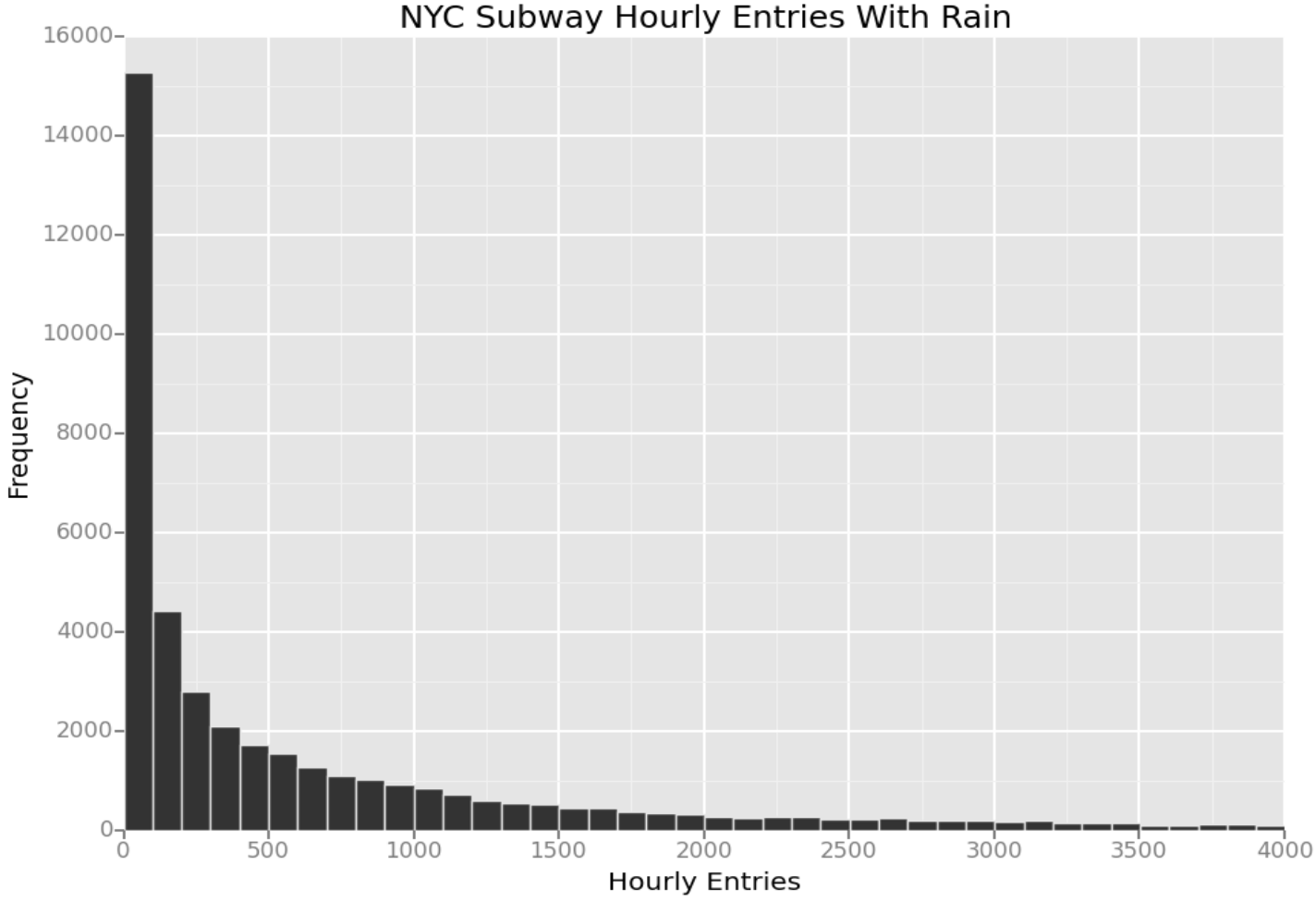
You can combine the two histograms in a single plot or you can use two separate plots.

If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.

For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that falls in this interval.

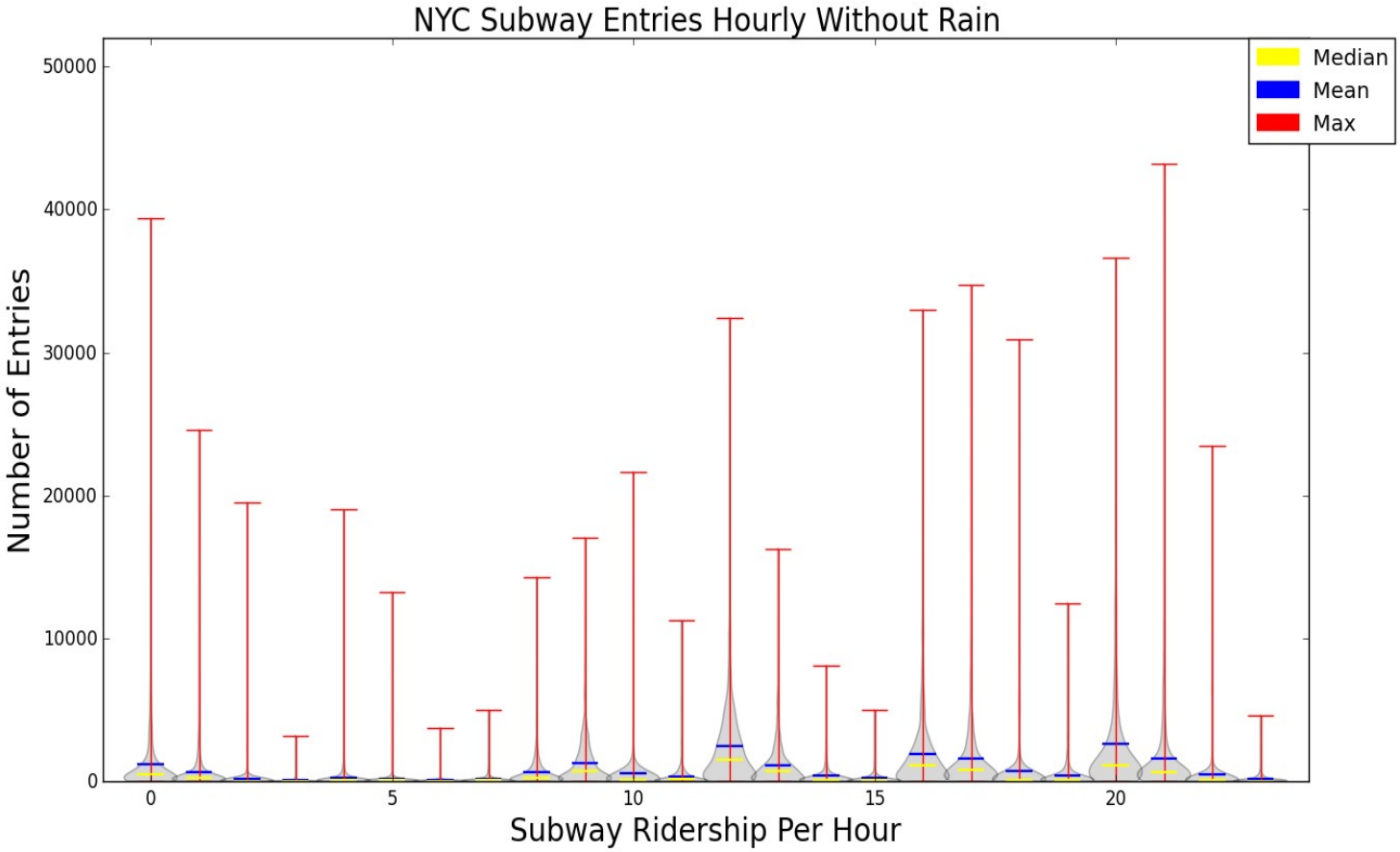
Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.


NYC Subway Hourly Entries Without Rain
Frequency
Hourly Entries
0
5000
10000
15000
20000
25000
30000
35000
0
500
1000
1500
2000
2500
3000
3500
4000

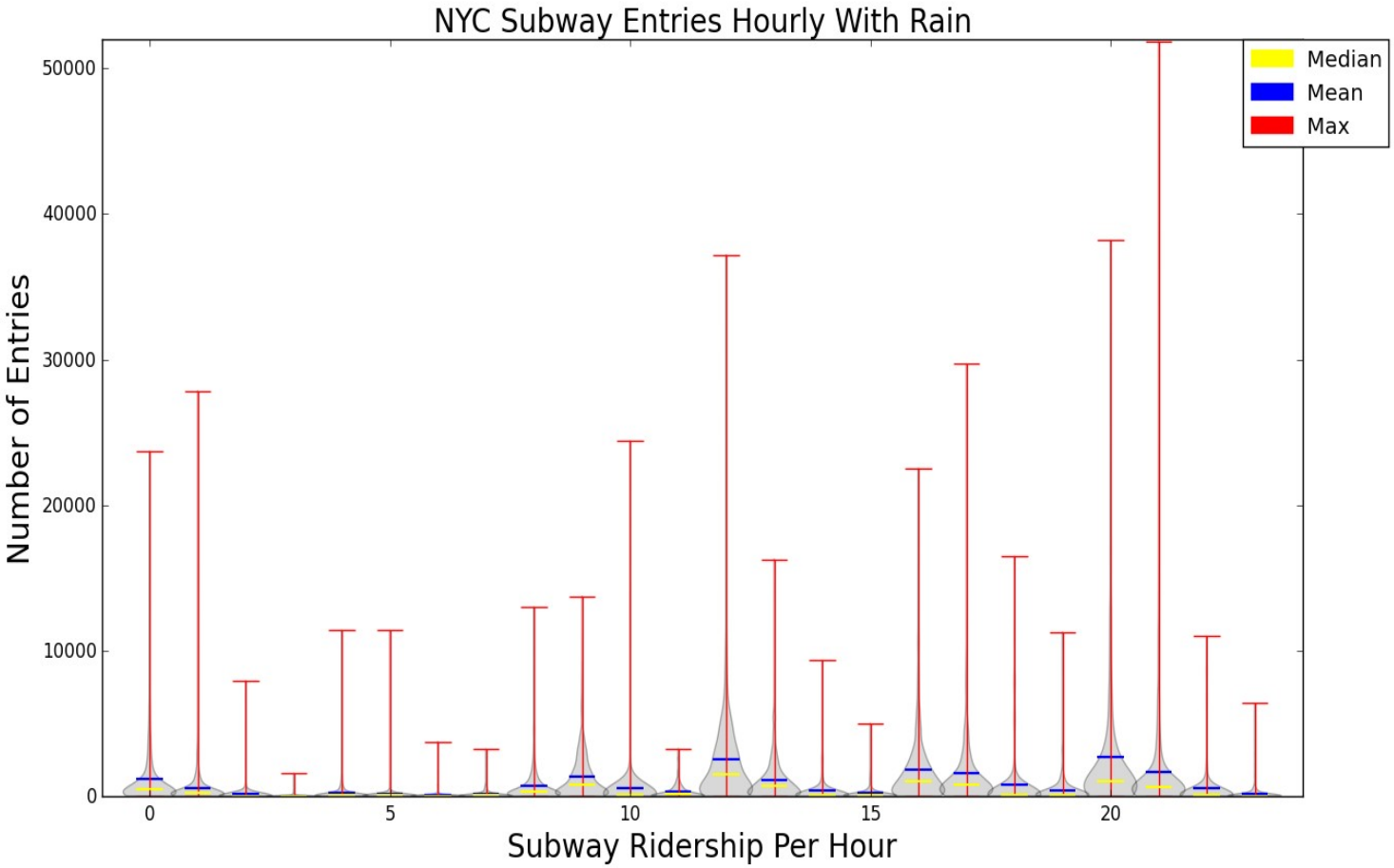

NYC Subway Hourly Entries With Rain
Frequency
Hourly Entries
0
2000
4000
6000
8000
10000
12000
14000
16000
0
500
1000
1500
2000
2500
3000
3500
4000

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.

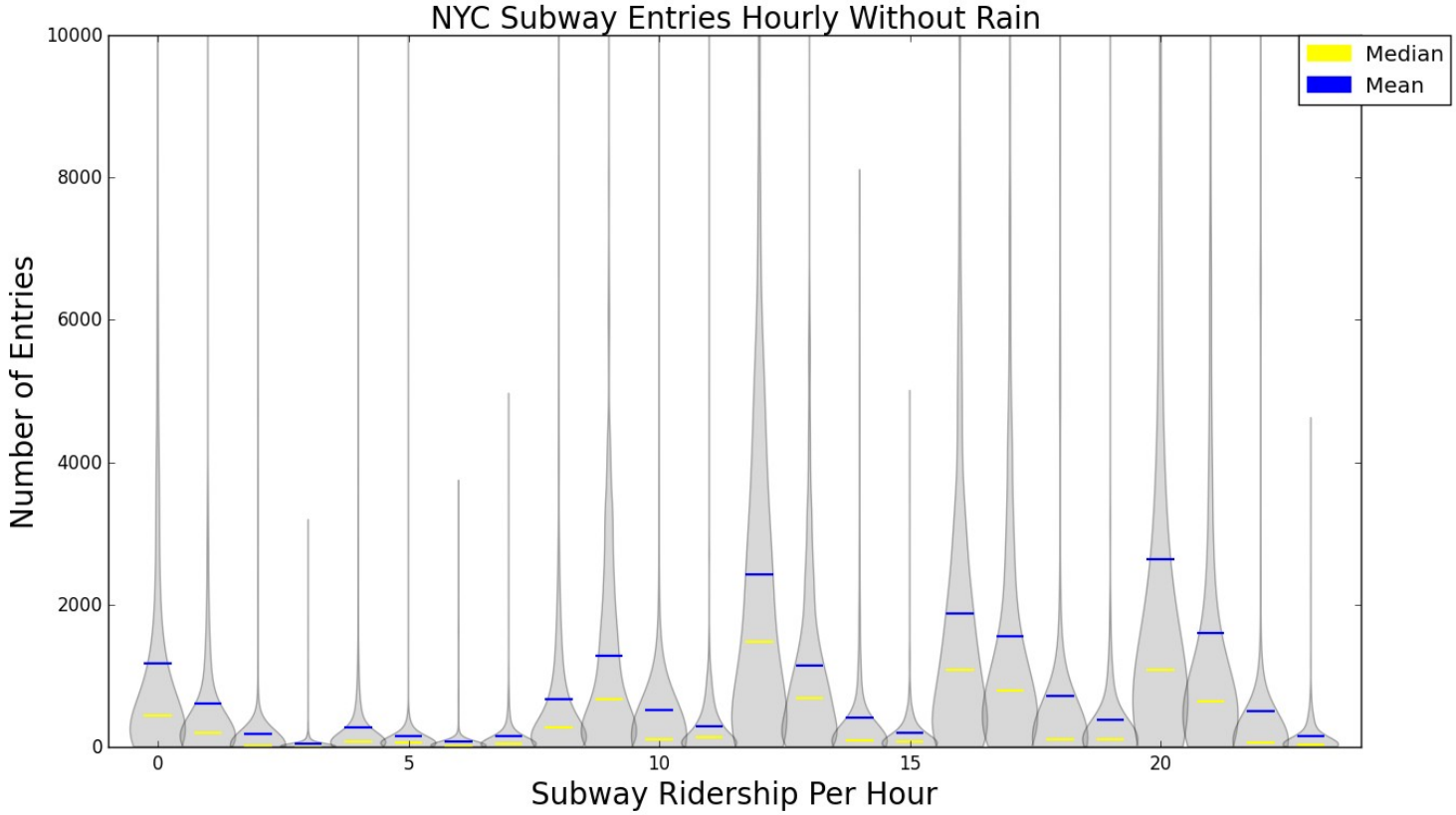
Description: In this graphs I intend to show the difference between entries in the NYC Subway in rainy and normal days. The goal it's to detect variability in each hour of the day.



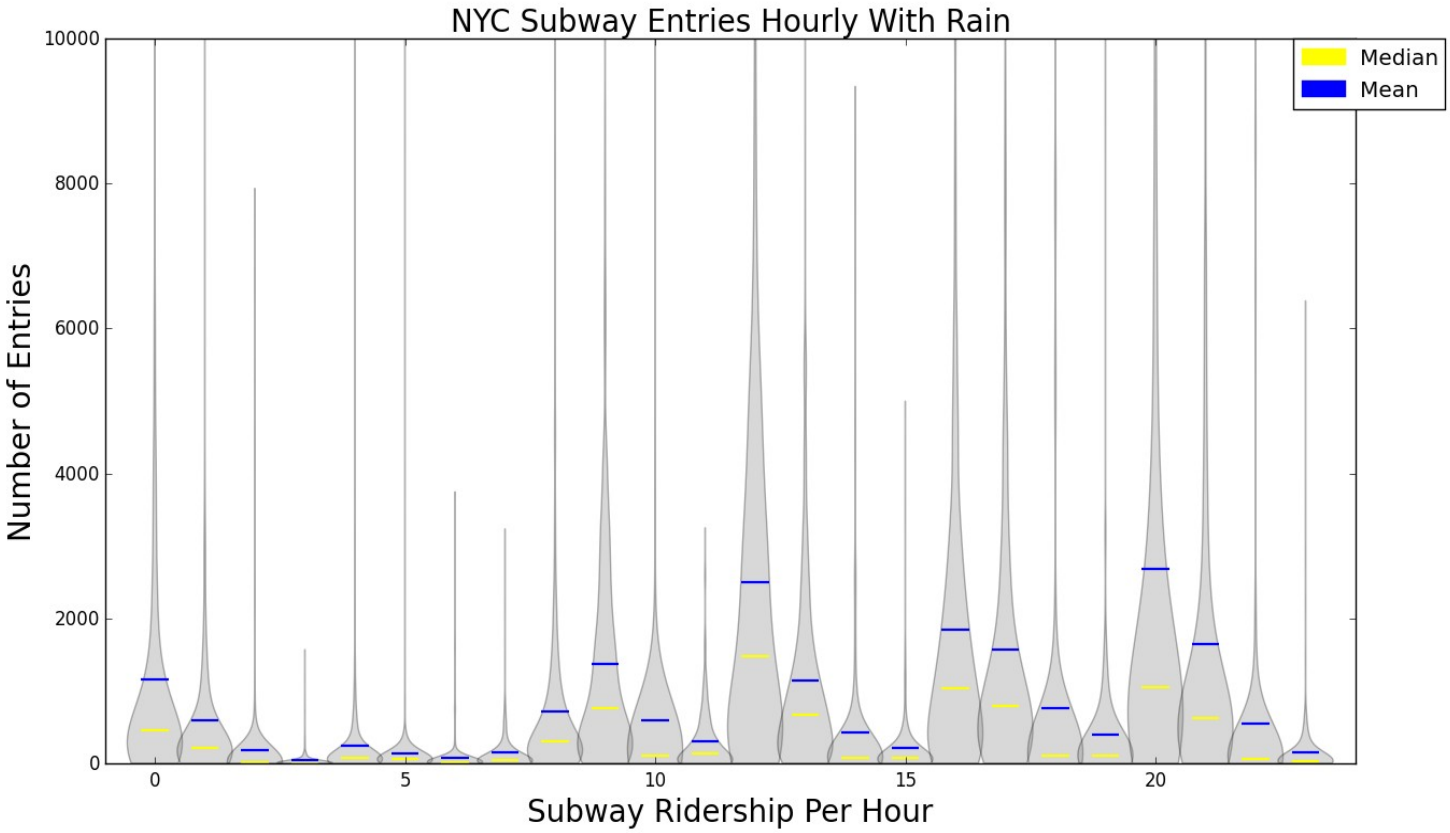
This plot shows ridership on a day without rain, showing the max values for each hour.



This plot shows ridership on a day with rain, showing the max values for each hour.



This plot shows ridership on a day without rain, showing the mean and median with more detail.



This plot shows ridership on a day with rain, showing the mean and median with more detail.

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?

Answer: From my analysis I can gather a couple of conclusions, first the null hypothesis was rejected, because there is a significant difference between the mean ridership on rainy vs non rainy days. Despite the fact that the average on rainy days is higher than the one on days without rain it's hard to make a prediction on whether ridership its truly higher given that the number of records in each category is very unbalance and the higher number of 0 entries on the days without rain brings down the average, but my gathering and interpretation leads me into believing that ridership for the NYC subway is higher on rainy days (Rainy days mean:1105.4464 – No rain days mean:1090.2788).

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

Answer: Mann-Whitney u-test (not a normal distribution) along with the mean of the two samples. P-value: 0,012499956, rain: 1105.4464, without rain: 1090.2788. This shows higher ridership average on rainy days compared to and example of days without rain from the same distribution. R^2 (coefficients of determination) value 0.463968977315 demonstrates a working model that has room for improvement.

Section 5. Reflection

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1. Dataset,

   Answer: The dataset has a couple of potential shortcoming. First, there appears to be many measing values from the Entries_hourly column that have been filled in with random zeros. Second, there are many different types of value representantions, dates, float, ints and even strings which could make it hard to work on some occasions.

2. Analysis, such as the linear regression model or statistical test.

   Answer: The Mann-Whitney Test it's a non-parametric test that lest us bypass knowledge on the distribution of the data, but comes at a price of lower power, meaning, if there is a significant difference between the two samples it will be harder to find than in a parametric test. The linear regression model will only allow insights into linear relationships and right now might be wrongly influenced by outliers.

Bibliography

Linear regression:

source: University of duke

url: https://stat.duke.edu/courses/Fall00/sta103/lecture_notes/multregr.pdf

---

Mann-Whitney U Test:

source: Laerd Statitics

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

---

Ggplot

source Yhat

url:https://ggplot.yhathq.com/
