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?

I used Mann-Whitney U test for analyzing NYC subway data. I used two-tail P value, therefore making no assumptions on direction of difference. The null hypothesis of this test is that the distributions of tested two groups are equal. My p-critical value is 0.05.

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.

This test is applicable because it is a non-parametric test, meaning that it makes no assumptions on probability distributions of tested data. Common parametric tests cannot be used here because the data does not follow normal probability distributions.

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.

Mean hourly entries with rain: 2028
Mean hourly entries without rain: 1846
p-value: 5.4e-06

1.4 What is the significance and interpretation of these results?

It can be concluded that the distribution of the number of entries is different between rainy and not rainy periods. This finding is statistically significant because received p-value is much smaller compared to the p-critical value.

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:

I used gradient descent.

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

I used Unit and Hour data to predict hourly entries. Units were used as dummy variables.

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.

I included hour data because I expected that time of the day affects how much subway is used. Since different stations are used with greatly varying intensity I thought that adding this information could also improve my model. Weather data was omitted because it did not improve the model.

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

Theta coefficient for Hour was 468

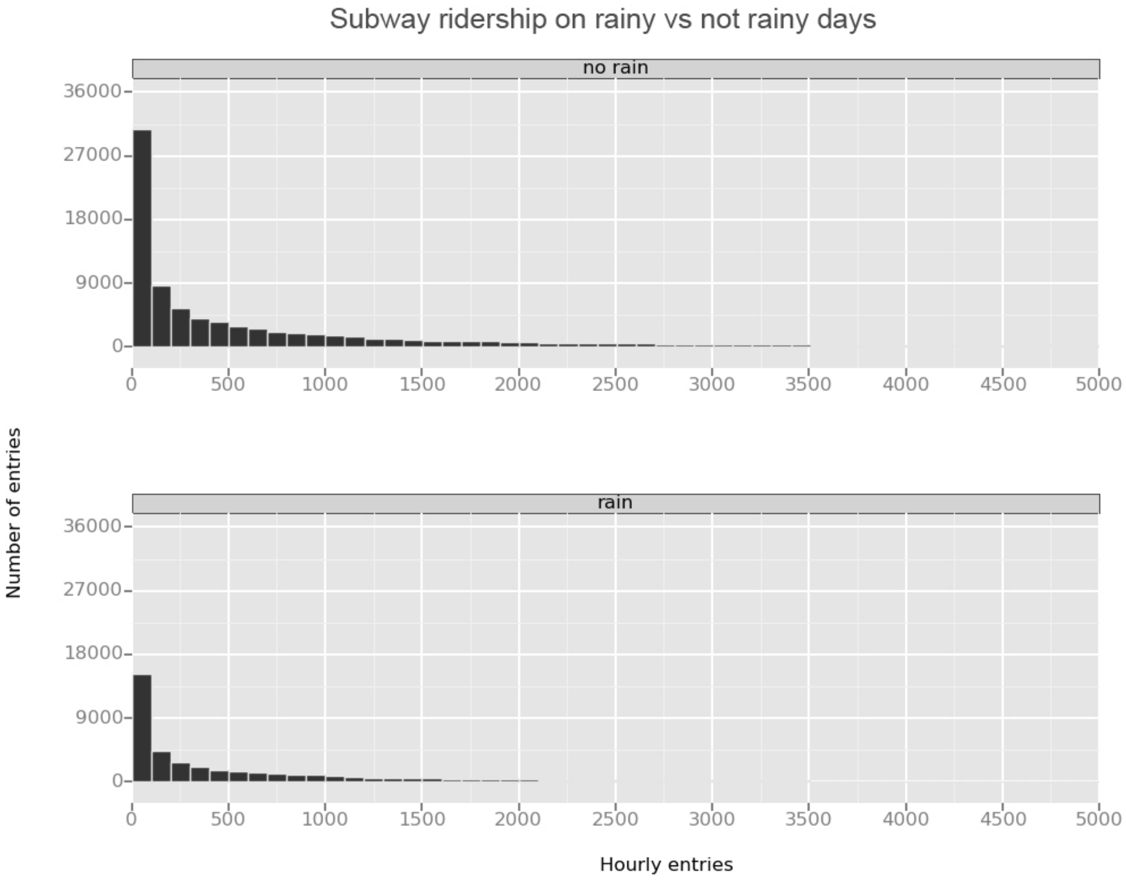
2.5 What is your model's R^2 (coefficients of determination) value?

$R^2 = 0.463$

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?

R^2 value of approximately 0.5 means that there is a weak correlation between input variables (Units and Hour) and predicted values (hourly entries). Units and time data cannot be used to predict hourly entries precisely, but their values describe around half of variability in this category.

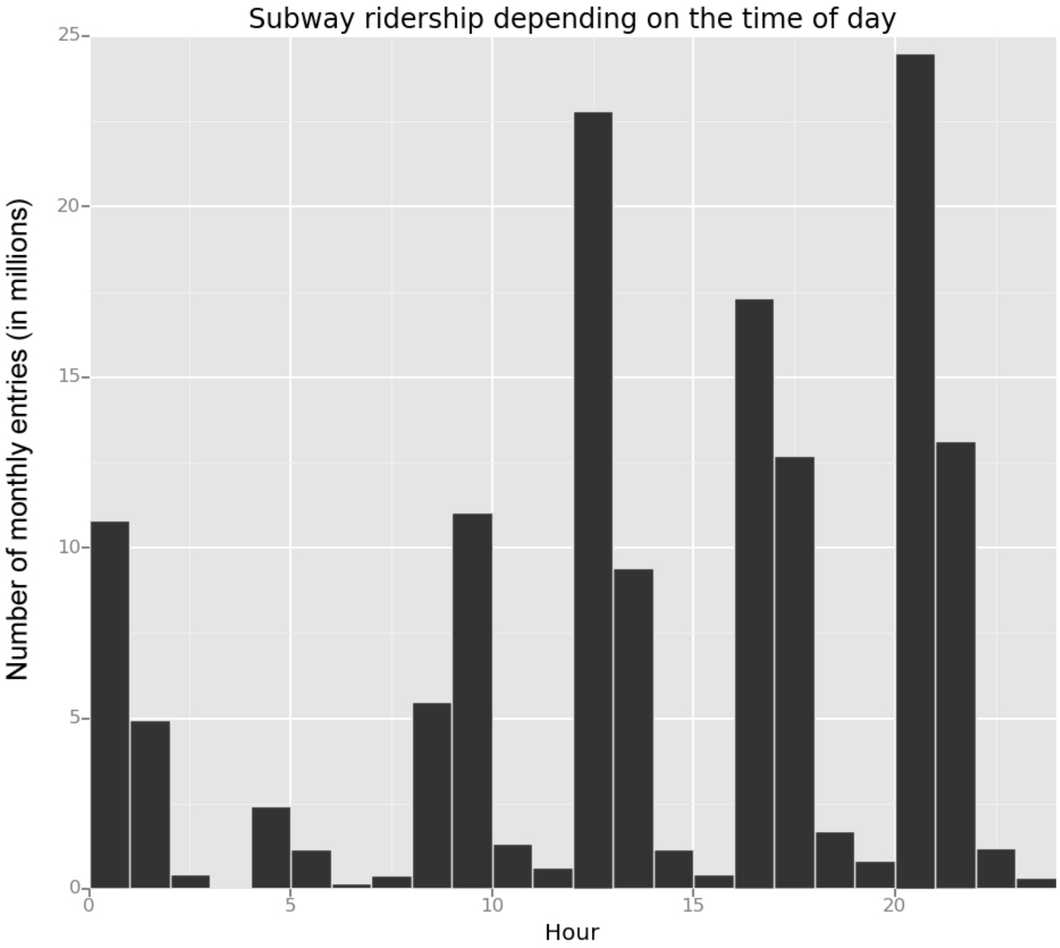
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.



Key insights of this figure are:

1. Distribution of ridership looks similar between rainy and not rainy days.
2. There are more data about not rainy days than rainy days

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.



Key insights of this figure are:

1. Subway usage heavily depends on the time of day
2. There are five peak usage times.

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?

I can conclude from my analysis that more people ride the subway when it is raining compared to when it is not. However, difference is not very big and weather (rain included) cannot be used to predict how much subway is used.

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.

Following analysis led me to my conclusions. First, mean hourly entries with rain are 10.9% higher (2048 vs 1846) compared to without rain. Comparison of ridership on rainy vs non rainy days with Mann-Whitney U test shows that there is a statistically significant difference in distributions of these two samples. Rain data cannot be used to predict hourly entries because adding rain data to linear regression model does not improve value of coefficient of determination (R^2). It means that although raining affects the subway ridership, other factors are causing most of the variability in hourly entries values. One of these factors is recording unit, which can describe considerable part of variability in hourly entries (R^2=0,42).

5.1 Please discuss potential shortcomings of the methods of your analysis, including: Dataset, Analysis, such as the linear regression model or statistical test.

Some Hour values are overrepresented compared to others in the given NYC subway dataset. It could potentially affect the outcome of my rain analysis, if there are differences how rain influences subway ridership during different times of day. For example if there are times of day when rain does not influence subway ridership and these entries are underrepresented, then i will overestimate rains effect.

My regression analysis only looks for linear relationships between features and values. For example adding hour value to regression model increased coefficient of determination by only 0.03. However, time of day clearly influences subway ridership (Figure 3.2), but my model did not get any better because the change is not linear.