Project 1 – Analyzing the NYC Subway Dataset
Short questions
Helena van Eijk

Section 1. Statistical Test

1.1 Which statistical test did you use to analyze the NYC subway data? The Mann-Whitney U-Test is used to test if the average number of entries during rainy days is significantly different from the average number of entries during non-rainy days.
Did you use a one-tail or a two-tail P value? Two-tail
What is the null hypothesis? The two samples (rainy vs non-rainy) come from the same population and have identical distributions.
What is your p-critical value? 0.05. Between the two samples the difference in average number of entries is calculated and the probability that this difference would be found in case the null hypothesis is true. When this probability is smaller than 0.05, the null-hypothesis will be rejected.

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.
The test is non-parametric and therefore does not have any assumptions about the distribution, it only tests whether the two samples are likely to come from the same distribution. A non-parametric test is chosen because it is expected that the outcome variable is not normally distributed. This expectation is also confirmed by looking at the histogram; it is heavily skewed (see section 3). There are two assumptions that should be met. First, the assumption that all observations are independent from each other, which is true for the subway data. Furthermore, the responses are ordinal (of two observations you can say which is the greater), so also the second assumption is met.

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 with rain: 1105.446
Mean without rain: 1090.279
p-value: 2 * 0.0249999 = 0.0499998

1.4 What is the significance and interpretation of these results?
Under the assumption that the two sample sets are independent, we reject the null hypothesis with 95% certainty and state that the average number of riders is different on rainy days compared to non-rainy days.

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:
Gradient descent

2.2 What features (input variables) did you use in your model?
Rain, Precipi, Hour, Meantempi, UNIT

Did you use any dummy variables as part of your features? Yes, UNIT

2.3 Why did you select these features in your model?
Based on intuition and a bit trial and error. Trial and error in this case means that I have sometimes included extra parameters in the model, and test whether the fit of the model would go up. That is for instance why I included Precipi into the model. Logical thinking brought me to the other features of the final model.
For instance, Rain: I figured people ride the subway more when it rains outside as an alternative to riding the bike to work.
UNIT: I thought that the amount of people would differ per subway station, because some stations are probably in a more rural area than others.
Hour: the number of people that ride the subway will for sure differ per hour of the day. In the morning most people travel to work and in the afternoon they return back home, which will be reflected in the amount of people riding the subway.

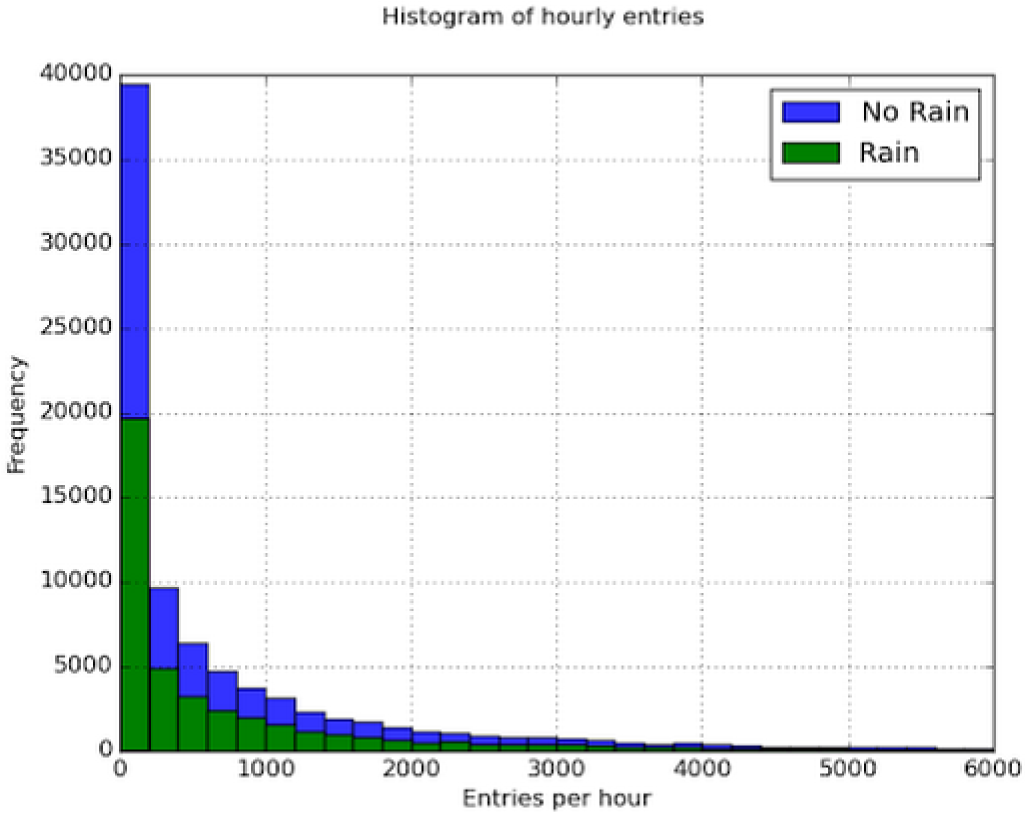
Meantempi: when the temperature goes up, people might ride the subway less, because they prefer taking the bike or walking over riding the subway.

2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?
Rain: 2.92398062e+00
Precipi: 1.46526720e+01
Hour: 4.67708502e+02
Meantempi: -6.22179395e+01

2.5 What is your model's R2 (coefficients of determination) value?
0.483218679844

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-squared is a measure for the part of the variability that is explained by the statistical model. It tells you how well the regression line approximates the real data points. You can interpret it as the proportion of total variation of outcomes explained by the model, so I would say that with a R-squared value of 0.48 you could say that the model is predicting ridership reasonably, but not great.

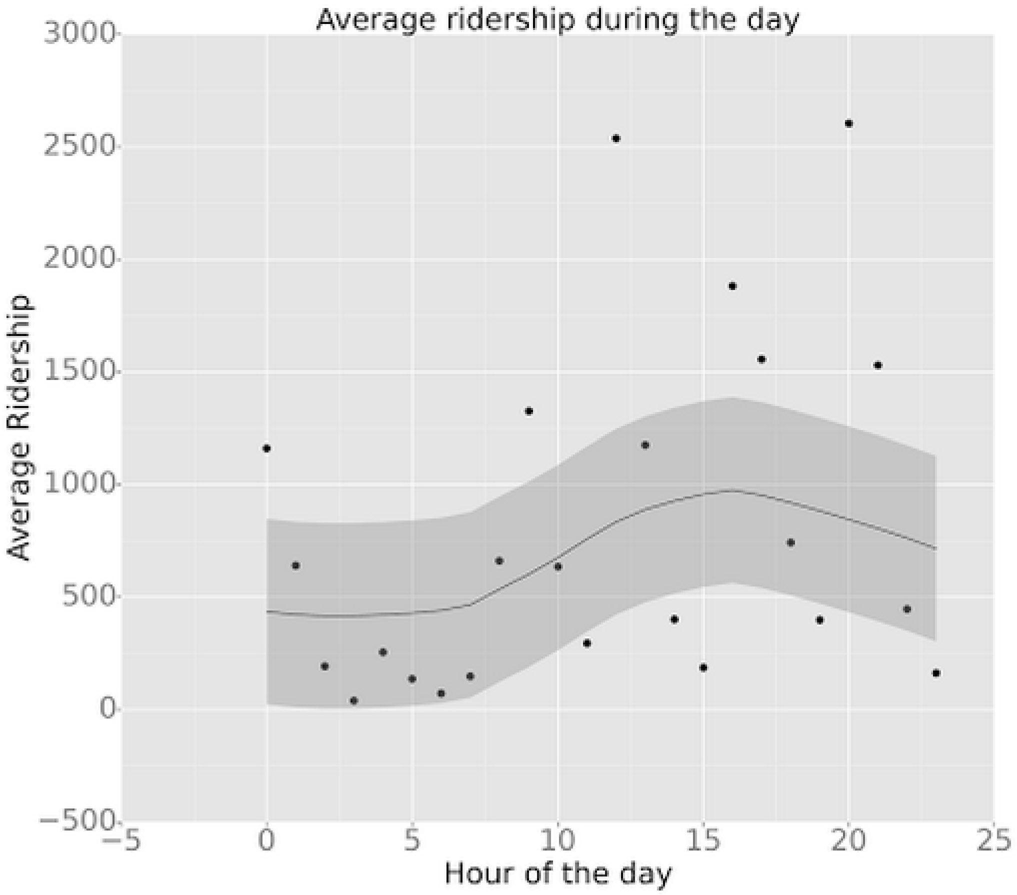
Section 3. Visualization



Comments: this histogram shows the distribution of hourly entries on rainy days versus non-rainy days. This histogram is merely displayed to compare the distributions, not to draw any conclusions on whether or not the average number of entries is different for these two groups.

One may notice that the set for non-rainy days is larger than the set for rainy days. Both histograms are heavily leaning towards zero, indicating that there are a lot of hours that have recorded very low entries per hour. Final note: the x-axis is truncated at 6000, so outliers (sometimes even exceeding 50000) are not displayed.

Besides this first plot, I decided to focus on the ridership during the day. For every hour of the day you see the average hours of ridership per station. You see peaks in ridership around 9am, 12am, 4pm and 8pm. It is quite fluctuating during the day, but if you add a smoothener, you see an overall increase in the afternoon.



Used python code:

```
plt.figure()
   turnstile_weather['ENTRIESn_hourly'][(turnstile_weather['rain'] == 0) &
(turnstile_weather['ENTRIESn_hourly'] <= 6000)].plot(kind = 'hist', bins = 30, alpha = 0.8, label = 'No Rain')
   turnstile_weather['ENTRIESn_hourly'][(turnstile_weather['rain'] == 1) &
(turnstile_weather['ENTRIESn_hourly'] <= 6000)].plot(kind = 'hist', bins = 30, label = 'Rain')
   plt.ylabel("Frequency")
   plt.suptitle("Histogram of ENTRIESn_hourly")
   plt.legend()
   return plt

AvgHour = turnstile_weather[['Hour','ENTRIESn_hourly']].groupby('Hour',as_index=False).mean();

   plot = ggplot(AvgHour,aes(x='Hour',y='ENTRIESn_hourly'))+geom_point(size = 60)+stat_smooth() +
ylab("Average Ridership") + xlab("Hour of the day") + ggtitle("Average ridership during the day") +
theme(text=element_text(size=30))
   return plot
```

Section 4. Conclusion

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?
More people tend to ride the subway when it is raining.

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.
Based on the statistical test you could tell that there was a significant difference between the two samples and that it is not probable that the two samples come from the same distribution. The test is a two-way test so you need visualization to determine which group has a higher number of people riding the subway compared to the other group. From the histogram you see that the blue group (is the no rain group) has a higher number of low hourly entries. The distribution looks heavier on the left. This would suggest that this group has less hourly entries than the other. However, the group rain has fewer samples than the no rain group, so the plots can only be used to get a picture about the distribution. The Mann-Whitney U-Test however also provides us with the sample mean for both groups and you see that the mean for the rain group is higher. This leads to the conclusion that significantly more people ride the subway when it is raining compared to when it is not raining.

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:
In general a disadvantage of the Mann-Whitney U-Test is that is has a lower power than parametric tests. The results about the difference between two groups could be displayed a bit weaker than actually is the case. However, since you do not know the distribution of the population, it is an understandable choice.

With regard of the regression model and the choice of using gradient descent as a cost function you can think about some general disadvantages. It could require a lot of iterations before you come to convergence because you have no sense of the appropriate direction of the size of the step to take. In this particular case, I do think the estimate was converging quite fast. However, you never know if your estimate is converging towards a local minimum, whereas the actual minimum is in a completely different area.

The data itself also has some shortcomings, for instance the fact that a lot of parameters (about the weather) only have one measurement per day, while the outcome variable (ridership) is measured in hourly entries. The true effect of rain is possibly significantly higher if the amount of rain was known per hour of the day.

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?
Not something particularly interesting.