Introduction to Data Science

Final project: analyzing the NYC Subway Dataset

Maarten Mulder

1. Statistical test

To analyze to NYC subway data, I used the Mann-Whitney *U*-test, which tests whether the null hypothesis that two populations have the same distribution is true. I used a two-sided p-value instead of a one-sided p-value, because I do not know yet whether rain will decrease or increase influence ridership, if it has any impact at all. By using a two-sided p-value I test both possibilities instead of only testing whether the assumption that rain either increases or decreases ridership is correct.

The null hypothesis is: the distributions for entries on rainy days and for entries on non-rainy days are identical. If the Mann-Whitney *U*-test says that the null hypothesis should be rejected, this means that the two distributions are not identical.

My p-critical value is 5%. I took this number because a significance level of 5% when using a two-tailed test (corresponding to 2.5% for a one-tailed test) is often recommended. The interpretation of this value is: the probability that two datasets are in fact equal, but that the observed difference between the means of the samples is due to random sampling.

This statistical test is applicable because it is non-parametric, meaning that it does not make assumptions about the probability distribution in the samples being tested. In practice this means that this test can be applied when comparing any set of two samples. Many parametric tests assume that the analyzed data is normally distributed, which is not the case for the data we are analyzing (as was already shown during the exercises).

The test gives the following results (using the original data set):

Mann-Whitney *U*-statistic	1.9244 * 10^9
p-value (two-sided)	4.99998%
Mean of entries when it is raining	1105.45
Mean of entries when it is not raining	1090.28
Median of entries when it is raining	282
Median of entries when it is not raining	278

The calculated two-sided p-value of 4.99998% means that there is a 4.99998% probability that the mean of the two populations is in fact the same, but that due to random sampling, the means are different. Because this value is below the critical level of 5%, the conclusion is that **the distributions of the two populations are different**.

To see whether average ridership is higher for the “rainy” distribution or for the “non-rainy” distribution, we look at the mean and median values of the distributions. Both the mean and median values are higher when it is raining than when it is not raining. Based on these findings, we can conclude that **more people use the NYC subway when it is raining**.

2. Linear regression

In order to determine the coefficients theta for the linear regression model that predicts the number of hourly entries, I used the “gradient descent” method as described in exercise 3.5.

I used the following input variables in my model:

- Time of day (‘Hour’)
- Whether it is raining or not (‘rain’)
- Precipitation (‘precipi’)
- Mean temperature (‘meantempi’)
- Mean wind speed (‘meanwindspdi’)
- Whether it is foggy or not (‘fog’)

For the selection of input variables, I started using the four variables that were given as starting variables in exercise 3.5: rain, precipitation, time of day, and mean temperature. I decided to keep these variables in the model for the following reasons:

- I kept rain because I supposed that people who would otherwise walk a small distance might prefer to take the subway in case it rains.
- For the same reason I kept precipitation because I suppose that the stronger or longer it rains, the previously mentioned effect becomes stronger.
- By experience I know that subway usage depends on the time of day (mainly due to people commuting to and from work and most people sleeping during the night) so I kept time of day in my model as well.
- I also kept mean temperature because I supposed that people might be more likely to enter a heated subway when temperatures are very low.

I also used UNIT as a dummy variable because results might depend on the station considered (some stations might be more sensitive to changes in weather than others).

Using these four variables resulted in an R^2 of 46.40%. The R^2 measures the “goodness of fit” because it shows the percentage of the variability being modelled that is explained by the input variables of the model. The remainder is due to the residual variability.

I then improved my model in the following way:

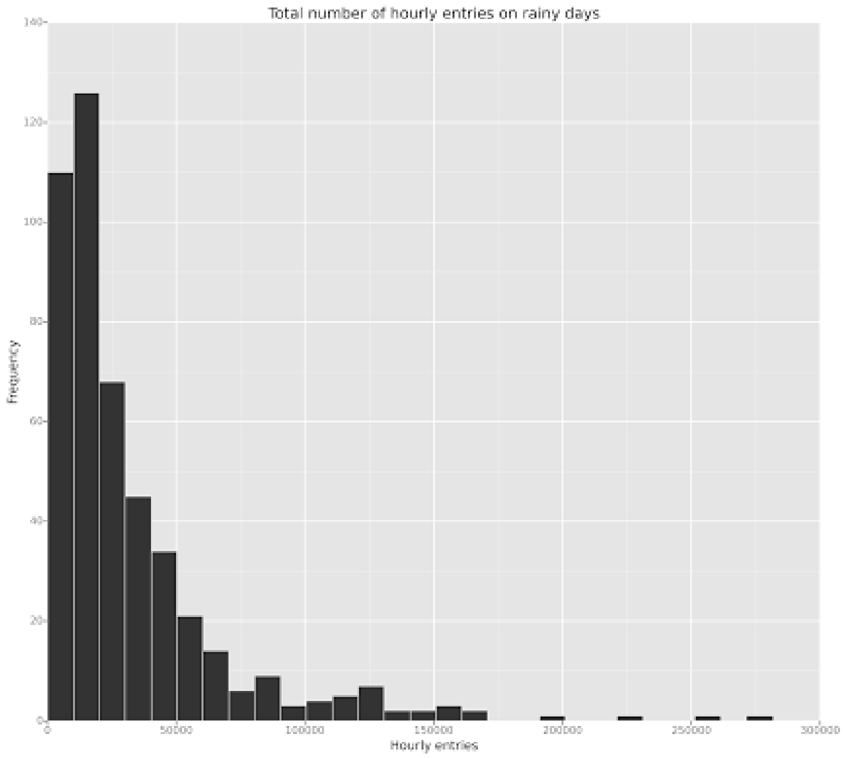
- I added the mean wind speed (meanwindspdi) because I supposed that people who would otherwise walk a short distance prefer to use the subway instead when there is heavy wind, both for comfort and safety reasons. By adding this variable, R^2 increased to 46.44%.
- For similar reasons I added fog as an input variable, resulting in an R^2 that improved to 46.51%.
- I wanted to use thunder because I supposed that many people will flee into underground subway stations when there is thunder and lightning, but unfortunately there is no subway data with thunder (all values of the “thunder” column are zero).

As it is stated in exercise 3.5 that the R^2 should be at least 20% and the R^2 of my linear regression model is 46.51%, this means that this linear regression model has a high goodness of fit. However, as the value is still far from 100%, the model is not excellent.

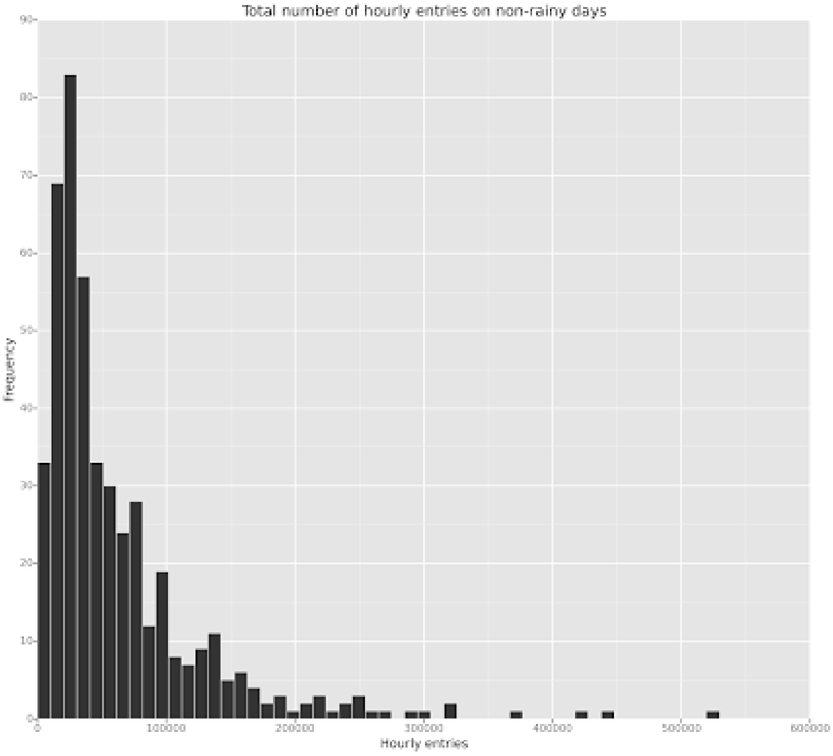
In short, I think this linear model is applicable, but it can still be improved.

3. Visualization

Histogram of total number of hourly entries on rainy days (using a bin size of 10,000):



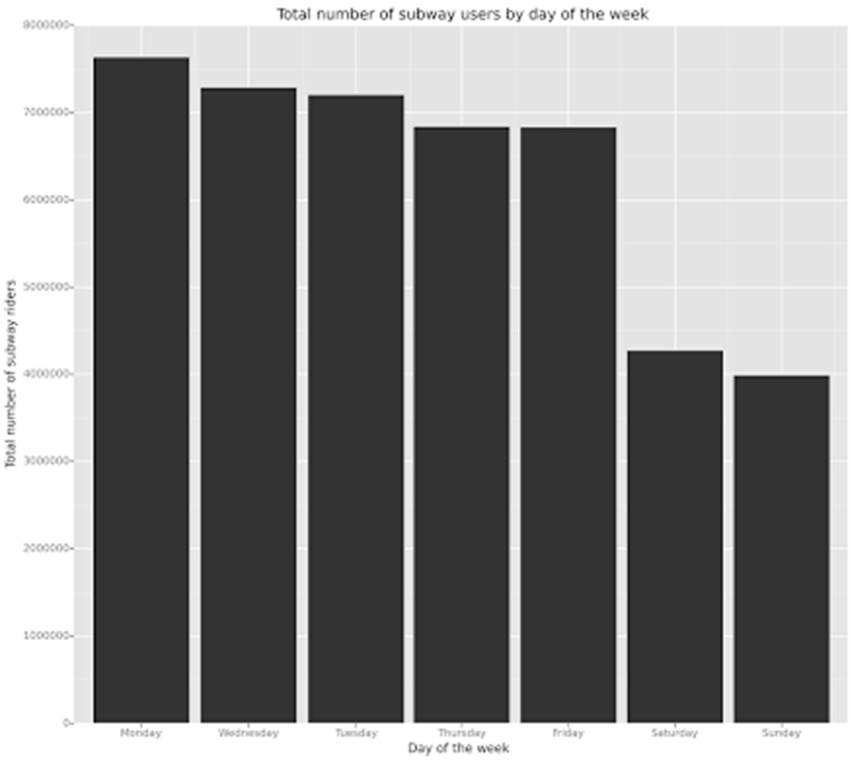
Histogram of total number of hourly entries on non-rainy days (using a bin size of 10,000):



Key insights from the two histograms:

- The data is clearly not linearly distributed.
- Most hourly entries are below 50,000, both for rainy and non-rainy days.
- There are much more entries for non-rainy days then for rainy days in the dataset considered (as shown by the higher average value of the bars in the non-rainy-days histogram as well as the higher number of bars)

Bar plot of the total number of subway users (based on the sum of hourly entries) by day of the week:



Key insights from this chart:

- In the weekends people use the subway less often. This is intuitive because a part of subway users are commuters, who often work only on weekdays.
- Monday is the busiest day of the week.

4. Conclusion

Based on my analysis of the NYC subway data, my conclusion is that **more people ride the NYC subway when it is raining**.

I based this conclusion on the following two insights that I gained from my analysis:

- The Mann-Whitney *U*-test tells me that the distributions for entries on rainy days and for entries on non-rainy days are not identical. Combined with my finding that both the mean and median are higher for the “non-rainy” distribution

- In the linear regression model, the ‘rain’ coefficient is positive, meaning that when it is raining there is an additional positive value added to the predicted number of riders.

5. Reflection

Potential shortcomings of my analysis method and/or the dataset are the following:

- The linear regression model has an important limitation that is already stated in the name: it assumes that there is a *linear* dependence between the input variables and the output variable. It is hard to prove that there is a linear dependence between the weather-based input variables (like for example the mean temperature) and the output variable (the number of subway riders).

- The data was apparently gathered in May. Because ridership behavior as a function of whether it rains or not might depend on the month of the year, it would be more useful to have a dataset that represents a full year (or possibly multiple years).

I would like to share the following insights on the data set:

- The dataset is large but not too large that it cannot be opened in Excel or a text editor, which is useful for getting a “feel” for the data structure.

- The general quality of the data set is good (no missing data, no need to merge several files possibly being in different formats, etc.)