ANALYZING THE NYC SUBWAY DATASET

Final project

1 FEBRUARY 2015

ERNI DURDEVIC

Short Questions

Overview

This project consists of two parts. In Part 1 of the project, you should have completed the questions in Problem Sets 2, 3, 4, and 5 in the Introduction to Data Science course.

This document addresses part 2 of the project. Please use this document as a template and answer the following questions to explain your reasoning and conclusion behind your work in the problem sets. You will attach a document with your answers to these questions as part of your final project submission.

Section 1. Statistical Test

1.1 Which statistical test did you use to analyse 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 the **Mann-Whitney U-Test** because the subway entries per hour did not have a normal distribution. I used a **two-tail** P value, because the null hypothesis was "**The ridership in rainy and non-rainy days are the same**". The p-critical value was **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.

The **Mann-Whitney U-Test** is a nonparametric test; it does not assume any particular distribution for the two samples.

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.

The mean on rainy days was = **1105.446**

The mean on non-rainy days was = **1090.279**

The one-tail p-value was = **0.02499991**

The two-tail p value was = **0.04999982**

1.4 What is the significance and interpretation of these results?

The significance level is set to 0.05. Since the obtained p-value is slightly lower, we can reject the null hypothesis. In other words, we can be 95% confident that there is a difference between rainy and 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:

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

I used both Gradient descent and statsmodels implementation of OLS (Ordinary Least Squares).

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 dummy variables for 'UNIT' variable and 'rain', 'precipi', 'Hour', 'meantempi', 'meanpressurei', 'fog' without dummies.

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."

The first attempt was based on intuition, I thought that 'UNIT', 'rain', 'hour' and 'fog' would have an influence on the usage of the subway. I then started to run various tests with the 'UNIT' dummy and one variable at the time. I took the variables that positively influenced the R^2 value and put them all together.

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

The coefficients are:

```
-2.03351469e+01 -> rain

-7.40799943e+00 -> precipi

 4.67925760e+02 -> Hour

-7.55496277e+01 -> meantempi

-3.91541750e+01 -> meanpressurei

 5.53910361e+01 -> fog
```

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

R^2 = 0.464728300213

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?

The R^2 value is quite small, the linear model is not working very well. In this case, the linear model works almost as good as a simple mean.

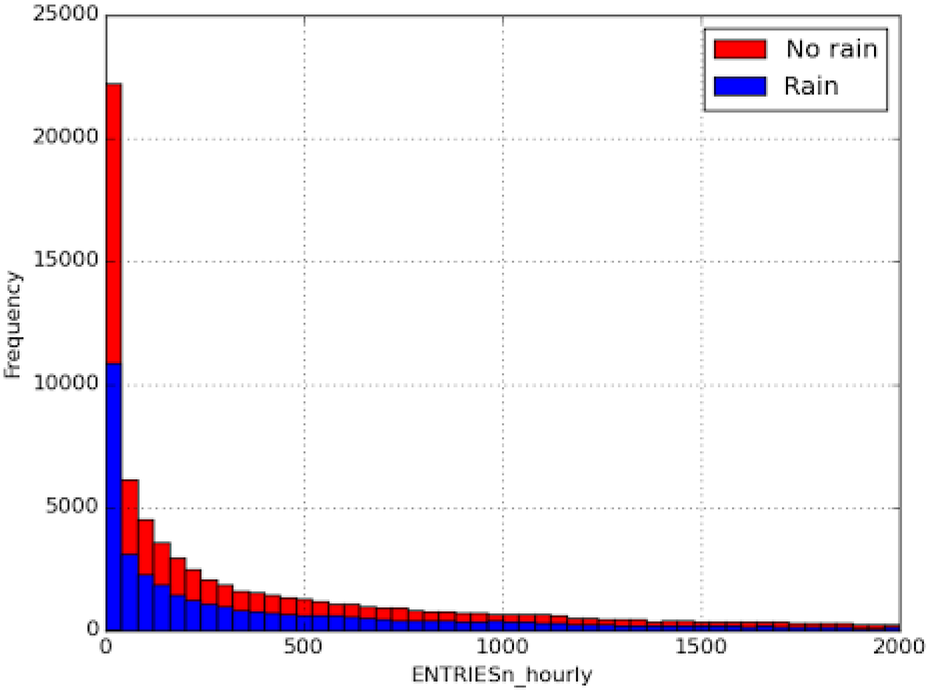
Section 3. Visualization

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., scatter plots, 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.

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

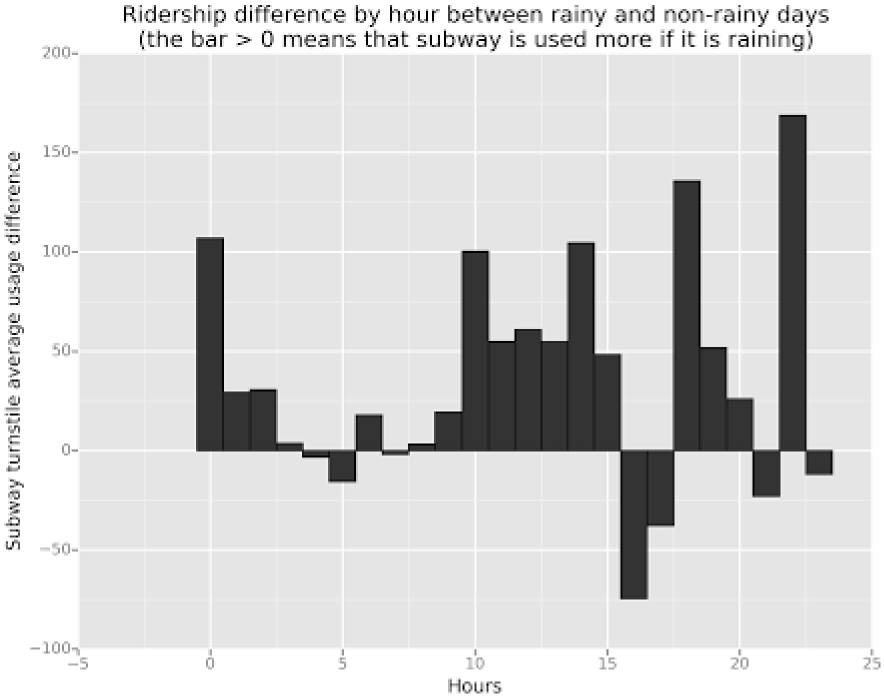
- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms, 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.



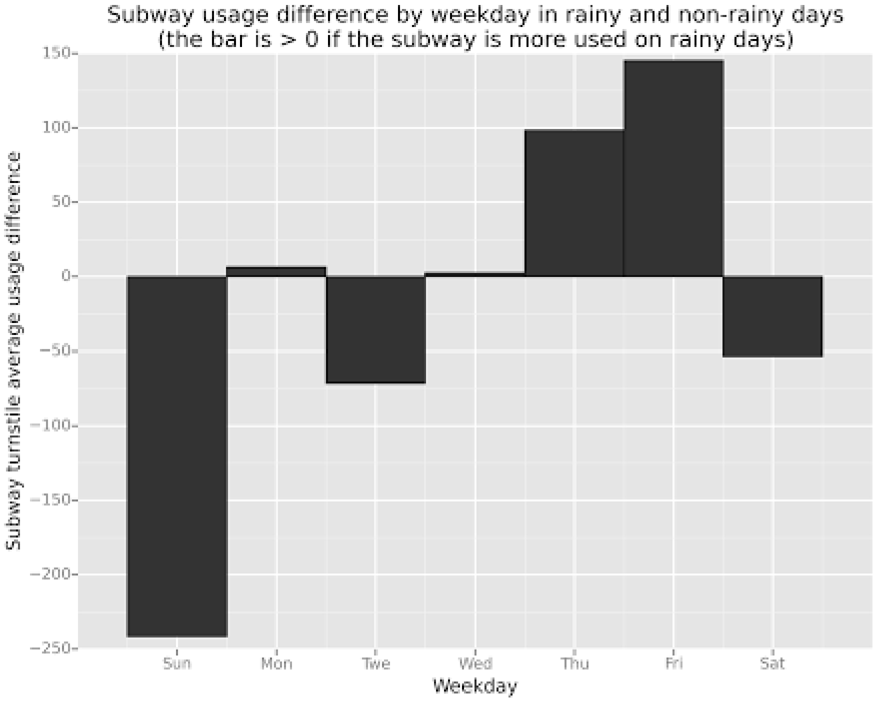
The graph shows that there are much more turnstile with fewer entrance on non-rainy days.

3.2 One visualization can be more freeform. Some suggestions are:

- Ridership by time-of-day
- Ridership by day-of-week



The graph shows the average increase/decrease of subway usage during the day due to the rain. The graph suggest that from 1 AM until 10 AM the rain does not affected subway usage, this could be because by night and for the way to work people tend to use or not use the subway in any case. The spikes at 6 PM, 10 PM and midnight suggest that people care more about taking the subway when it is raining especially for going out and coming back in the evening. The countertrend at 4 PM and 5 PM suggest that if it is raining people tend to delay their trip for a couple of hours if they can.



The usage by weekday suggests that rain causes a sensible drop in subway usage on Sunday.

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?

On average people use more the subway when it is raining, nonetheless there are some exceptions on some days of the week and hours of the day.

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.

From the Mann-Whitney U-Test we can be 95% confident that there is a difference between rainy and non-rainy days. Considering the means of usage on rainy and non-rainy days, and taking the one-tail p-value 0.02499991 (because in this case we are testing ne null hypothesis "The ridership is not increased in rainy days") from the previous test we can also be 97.5% confident that people use more the subway on rainy days.

The 'rain' coefficient in the linear model found with the Gradient descent method of the exercise above is misleading, because it is negative even though the rain influences positively the ridership. By running the gradient descent method only with the 'rain' feature, we obtain a positive value for rain coefficient (25.59), confirming that the rain influences positively the ridership.

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

The Dataset has a big difference between the total entries and total exits in the subway, and we do not know if the missing exits are concentrated in some particular stations or they are uniform in the whole net. This makes difficult to extract conclusions for entries and exits per station. The dataset (the one passed in the exercises, not the downloadable improved one) did not have the station name, so it was not possible to analyse the data by station.

By looking at data (for example the above histograms of ridership by hour and by day of the week), we can easily see that the ridership cannot be described by a linear model. A non-linear model would probably be better, leading to a higher R^2.

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?

Not more then what I already shared in the questions above…