Analyzing the NYC Subway Dataset short questions

Laurent BRINGUIER – January 2015 Cohort

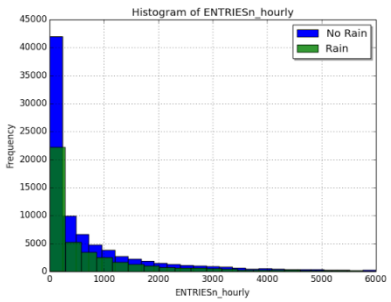
Section 1. Statistical Test

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?

We used the **Mann–Whitney *U* test** to determine whether or not the two distributions being compared are identical (in our case, the population of entries per hour on rainy days and the population of entries per hour on non-rainy days)**.**
Our **Null Hypothesis H0** is that the population of entries per hour on rainy days and the population of entries per hour on non-rainy days are identical.
The **Alternative Hypothesis H1** is that the two populations are not the identical.
As we consider that in case of H1, the deviation can be in either direction, we will use a two-tailed P value.
We will use **P<0.05 two-tailed.**

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.

We used the **Mann–Whitney *U* test** for a main reason:
Mann–Whitney *U* test is a non-parametric test and our data is not normally distributed (as shown on the below histogram) **.**
In addition to the above, both groups are independent and the distributions of both groups are equal under the null hypothesis.



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 of Rainy Days sample	Mean of Non Rainy Days sample	U	P-value
1105.446	1090.278	1924409167	0.04998
Note : as the p-value given by scipy.stats.mannwhitneyu (0.02499) is for a one-sided hypothesis, to get the two-sided p-value we multiply the returned p-value by 2 (=0.04998).			

1.4 What is the significance and interpretation of these results?

Previous results shows that : **P-value < 0.05 (two tailed)**
The Null Hypothesis H0 is rejected .
We can conclude that : **the population of "entries per hour on rainy days" and the population of "entries per hour on non-rainy days" are significantly different.**
In addition, we can also see in the results that the Mean of Rainy days sample is greater than the Mean of 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:

We used the **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?

We used the following input variables :
Precipitation, Rain, mean wind speed, mean temperature, mean dew point, max dew point, min dew point, min temperature, max temperature, max pressure, min pressure, mean pressure.

We used 2 dummy variables : **UNIT and Hour**

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.

We selected these features based on intuition and experimentation.
Firstly, we made the assumption that weather condition would influence the behavior of riders in two different ways :
- people would use the subway to protect themselves in case of bad conditions instead of walking outside : big rain , strong wind, low min temperature , etc…
-people would less use the subway if weather conditions gets better: min temperature increase, average temperature increase, etc..

Secondly, we used each feature one by one to check the influence on the model (increase of R2 value) and then try again with different combination of features.
These experimentations suggested for example that the Fog has a very limited effect, rain and precipitation and mean pressure a stronger one, mean wind speed a strong effect also.
As we were looking for the Best R2 results, we included also the features with small effect as long as they increased the coefficient of determination.

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

'precipi'	'rain'	'meanwindspdi'	'meantempi'	'meandewpti'	'maxdewpti'
3.85871913e+00	8.75191633e+00	4.72063927e+01	-2.07146964e+01	3.99620224e-02	4.01329225e+01
'mindewpti'	**'mintempi'**	**'maxtempi'**	**'maxpressurei'**	**'minpressurei'**	**'meanpressurei'**
-3.18612216e+01	-6.06410231e+01	1.98359909e+01	1.89775683e+00	-5.79328641e+01	9.23507882e+00

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

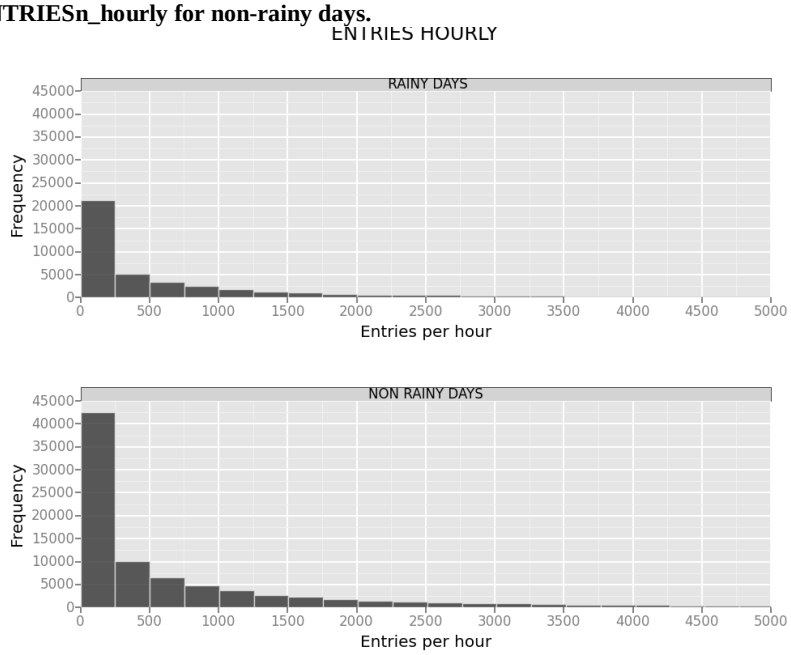
Our model's **R2 = 0.5067**

2.6 What does this R2 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 R2 value?

The closer R2 is to 1 the better is the model.
We strongly believe that this model can be improved. However, to predict human activity or behavior , a squared values around 50% is in our opinion acceptable.
At least, for our Analysis, this model shown a relatively high correlation between some weather features and ridership. The positive weight for the rain and precipitation suggest a correlation between the presence of Rain and an increase of the ridership per hour.

Section 3. Visualization

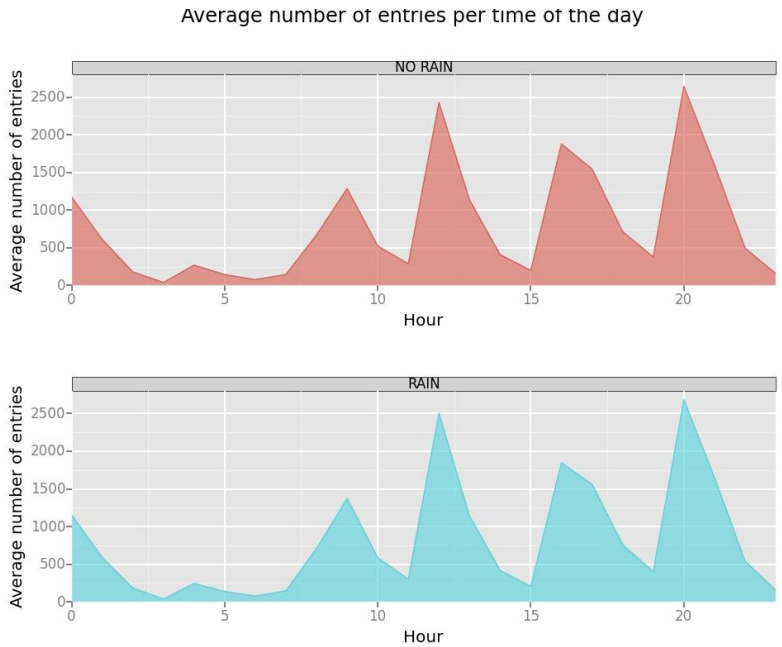
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.



The frequency of Entries per hour for rainy days (figure 1) and non rainy days (figure 2) for the NYC subway. plotted using ggplot2 with bins = 250.

Despite both samples seems to follow a similar type of distribution, they are not normally distributed.
Note : "Rainy Days" sample has less records than "Non Rainy Days" sample. The x-axis has been truncated at 5,000 cutting off outliers in the long tail which extends beyond 40,000.

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. Some suggestions are:



Hour	No Rain	Rain
0	1174,01	1155,01
1	612,01	590,98
2	180,23	185,16
3	39,55	37,38
4	270,40	246,25
5	144,63	138,47
6	77,69	76,66
7	145,73	147,55
8	670,67	709,43
9	1286,78	1373,75
10	**524,11**	**587,17**
11	287,81	301,27
12	2428,30	2504,67
13	1134,99	1140,04
14	408,78	418,57
15	200,96	207,04
16	1877,93	1845,86
17	1547,05	1563,34
18	715,91	758,09
19	378,11	402,70
20	2642,37	2684,29
21	1598,72	1651,61
22	**495,32**	**541,03**
23	158,86	156,26
TOTAL	19000,92	19422,59

This plot is showing the average number of entries per time of the day for all the stations on non rainy days and rainy days. We can see that the Rainy Day graph shows some higher values (10h, 22h..) and a higher Total (+422 entries)

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?

Concluding that “more people ride the NYC subway because of the Rain” would be a too simplistic and probably wrong way to say it.
However, based on our analysis we can definitely say that **there is a significant correlation between the rainy day factor and the chance to have a higher number of entries.**

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.

The statistical Analysis :

1 - The Mann-Whitney test suggests that there is a significant difference in ridership per hour between rainy days and non rainy days.
2 - The comparison of the two sample mean suggest that the ridership during rainy days is slightly greater than during non rainy days.

Conclusion of the statistical Analysis : There is a significant difference between ridership per hour on rainy day and on non rainy day and it seems to be an increase.

The visualization tends to show the same aspect.

The linear regression:

The positive weight related to the rain feature suggests that there is a positive correlation between the rain and the increase of ridership per hour.
Precipitation weight suggests also that the increase of the amount of rain during the day increases the ridership.

All these evidences supported that in the period of May 2011, there was a greater chance to find more people riding the NYC subway on Rainy days than on Non Rainy days.

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.**

We would like to highlight several shortcomings of our analysis on the dataset and method used:

- The sample size : the dataset consisted only in 30 days including 10 rainy days and 20 non rainy days. A bigger dataset may lead to different conclusions.
- The granularity : entries are given per Hour but the rain is given per Day. With this dataset we have to assume that during a rainy day, it rain all day long which is not the reality. It would more accurate to run the analysis on the "rainy Hours" versus "non rainy Hours" to isolate the impact of the rain factor.
- We are not using the location of the station (assuming it rains everywhere).
- Our Analysis is performed on a dataset mixing many variables impacting the number of entries, for example:
    - Station : The behavior of riders going to a touristic station can be very different from the behavior of riders going to the station close to their workplace or house.
    - Hour of the day : different hours of the day shows different behavior also.
    - Day of the week : each day of the week has a particular entry / hour pattern.
    - Particular date : Some dates can be special events date bringing also a different behavior
    - Other weather conditions : heavy rain with strong wind and cold temperature can have a different influence than a light rain with no wind and warm temperature.

In order to improve the isolation of the rain factor it would be useful to have a dataset providing enough records to pick a significant sample as below :
Same hour , same day of the week , same station, same type of condition (range of temperature, pressure, wind, etc) and a factor rain that varies.

Gradient descent was a good approach to be able to experiment the model and add or remove features. However, as the number of features is not so important, we could have used an analytical method instead or as a complement.

Sources

- **YouTube, (2015).** ***2.7 - Gradient Descent For Linear Regression - [Machine Learning] By Andrew Ng*****.** [online] Available at: https://www.youtube.com/watch?v=ZgXjKa0ChDw&index=11&list=PLLH73N9cB21V_O2JqILVX557BST2cqJw4 [Accessed 10 Apr. 2015].
- **Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of Fit?. (2015). [Blog]** ***The Minitab Blog*****.** Available at: http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit [Accessed 1 Mar. 2015].
- **Wikipedia, (2015).** ***Mann–Whitney U test*****.** [online] Available at: http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test [Accessed 10 Apr. 2015].