Student	Alex Trejo
Date	June 23, 2015

Analyzing the NYC Subway Dataset

Section 0. References

- http://ggplot.yhathq.com/docs/index.html
- http://docs.ggplot2.org/0.9.3.1/index.html
- http://www.socsci.uci.edu/~schofer/2007soc8811/publicfiles.htm
- https://www.youtube.com/watch?v=dCG3VAfa11Y
- http://www.rincondepaco.com.mx/rincon/Inicio/Apuntes/Proyecto/archivos/Documentos/U_Mann.pdf
- https://github.com/yhat/ggplot/issues/33
- http://www.clockbackward.com/2009/06/18/ordinary-least-squares-linear-regression-flaws-problems-and-pitfalls/
- https://onlinecourses.science.psu.edu/stat501/node/29
- http://www.skymark.com/resources/tools/normal_test_plot.asp
- I know you demand specific links in sites like http://stackoverflow.com/, but I don't have the specific topics saved… :(

Section 1. Statistical Test

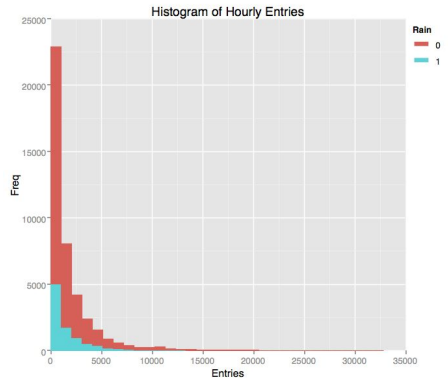
1.1 Which statistical test did you use to analyze the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis? What is your p-critical value?

I used the Mann-Whitney U test using the two-tail P value. The null hypothesis (H_0) is: "*There is no Ridership differences between Rainy and No Rainy days ($Me_1 = Me_2$)*" considering a critical p-value of 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.

Code:

```
datafile='improved-dataset/turnstile_weather_v2.csv'
df = pd.read_csv(datafile)
plot = ggplot(df, aes(x='ENTRIESn_hourly', color='rain', fill='rain')) + geom_histogram(binwidth=1000) + \
       ggtitle('Histogram of Hourly Entries') + \
       labs('Entries', 'Freq')
print plot
```



Conclusion:

Apparently, the two samples follow an exponential distribution, so we can not use an standard Welch's T test. We could use a Non-parametric Test like **Mann-Whitney's U Test**.

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.

Code:

```
datafile='improved-dataset/turnstile_weather_v2.csv'
df=pd.read_csv(datafile)
rain_df = df['ENTRIESn_hourly'][df['rain'] == 1]
no_rain_df = df['ENTRIESn_hourly'][df['rain'] == 0]
rain_mean = np.mean(rain_df)
no_rain_mean = np.mean(no_rain_df)
mannwhitneyu = sps.stats.mannwhitneyu(rain_df, no_rain_df)
print 'Rain Mean: ', rain_mean,' No Rain Mean: ',no_rain_mean
print 'Mann-Whithney Resuts: U=', mannwhitneyu[0],' p-value(two-tail)=',mannwhitneyu[1]*2
```

Results:

```
Rain Mean:  2028.19603547
No Rain Mean:  1845.53943866
Mann-Whithney Resuts: U= 153635120.5
p-value(two-tail)= 5.48213914249e-06
```

1.4 What is the significance and interpretation of these results?

At first sight, it seems that there is a difference between the two sample means (2028,2 vs. 1845,5). Looking at the results obtained with the Mann-Whitney U test I got a two-tailed p value of **5,48e-6**, lower than 0,05 (our p critical value) and I can reject the null hypothesis, so there is significant difference between the two samples. In other words, the Median of ridership in rainy days is higher than the Median of non rainy ones.

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:

I finally used the OLS approach using Statsmodels even though I worked a lot with the SGD approach using the ScikitLearn.

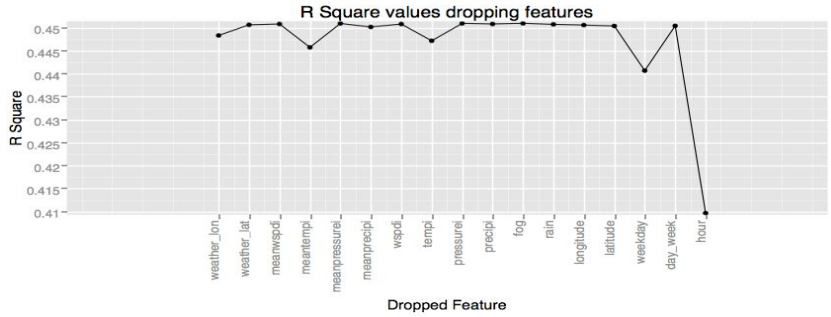
2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

The input variables are: *'fog', 'precipi', 'pressurei', 'tempi', 'wspdi', 'weather_lat'* and *'weather_lon'*, combined the dummy variables *'stations', 'conds', 'day_week' 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.

I will explain how I decided to use the features & Dummy vars described in question 2.2, recreating the process I followed, by trying to find the max R^2 by removing features sequentially:

- I started by deciding that 'Station' & 'Conds' had to be dummy var due to its discrete and qualitative characteristics.
- Once I had the initial model with 17 features & 2 dummy vars, I coded a little function to evaluate the R^2 for each model, dropping a feature each time, and the plotted results were:

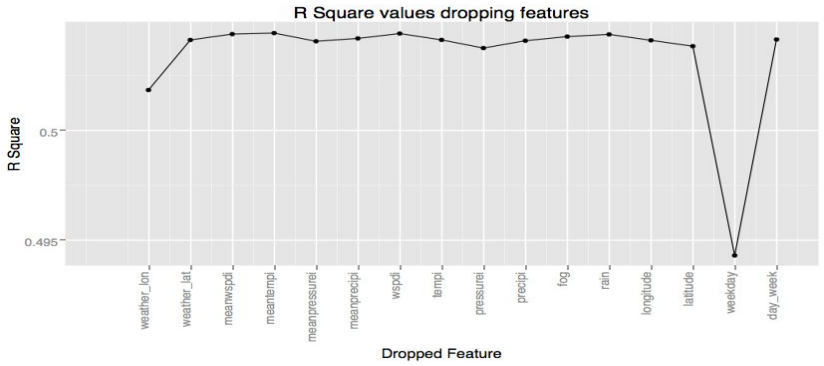


Comments:

We can see that the feature 'hour' causes a significant R2 drop until 0.40985 while the maximum R^2 value obtained is 0.451132. That's obvious, because the entries are Highly correlated with the 'hour' of the day. So 'rain' must be in the final model.

R^2_{max} = 0.451132083693
R^2_{min} = 0.409850113261

- After the results obtained, I was thinking in the 'Hour' feature, and I observed that is a discrete variable, taking values from [0, 4, 12, 16, 20], so I decided to convert 'hour' in a dummy variable and launch again the calculations, obtaining the following values - plot:



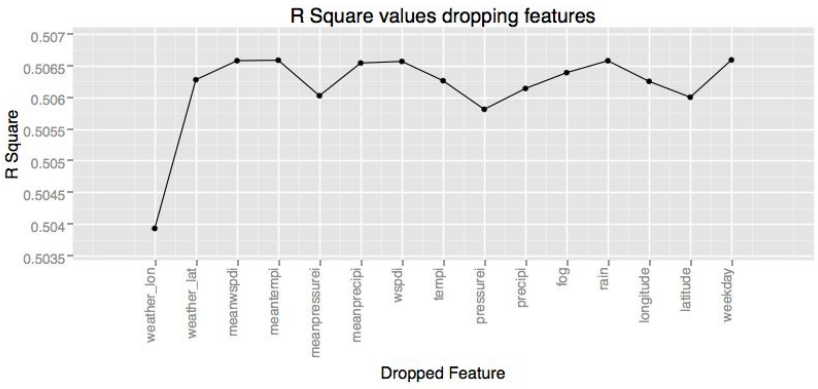
Comments:

First important thing is that the R^2 value has increased significatively, to a maximum value of 0.50446. That's a 10% versus last model, so the decision seems to be correct.

Another important thing, is that the feature 'weekday' causes a significant R^2 drop, if I remove it from the system.

$R^2_{max} = 0.504461032931$
$R^2_{min} = 0.494332249612$

- Analyzing the data again, I observed that there are two variables 'weekday' and 'day_week' highly correlated ('weekday' is included in 'day_week'), so if I include one of them, I'll need to remove the other because is redundant information. Against the first conclusion of removing 'day_week', I decided to maintain it and remove 'weekday' because I consider that offers more information. Another important thing is that 'day_week' is, again, a discrete variable that takes values from [Monday - 0,1,2,3,4,5,6 - Sunday], so I can transform it into a Dummy var. To verify my theory, I included a new dummy var and maintained 'weekday' in the model. I expect to find that 'weekday' now has no effect to the R2 value if I remove it from the model. That's the result obtained:



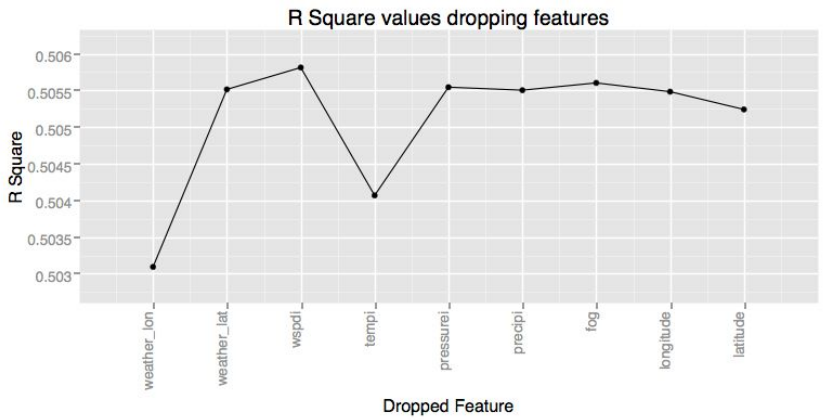
Comments:

And my theory was correct! Removing 'weekday' from the model doesn't affect the R^2 value.

Now another interesting fact appears… the apparent non importance of 'meanwspdi', 'meantempi', 'meanprecipi', 'wspdi' and 'rain'

$R^2_{max} = 0.506601409792$
$R^2_{min} = 0.50394448044$

- Now I can remove ‘weekday’ from the model and observe the remaining features... I decided two things, and plot the results again:
  - To use the ad-hoc values of precipitation, pressure, temperature and wind ('*precipi*', '*pressurei*', '*tempi*', '*wspdi*') instead of mean values, because the ad-hoc values are taken in certain hours [0, 4, 12, 16, 20] the same as the 'ENTRIESn_hourly', so I consider that this might be better to construct the model.
  - And remove ‘rain’ from the model, due its high correlation with ‘precipi’ & ‘conds’:

	Comments: Now I can observe that if I remove ‘wspdi’ from the model, the R^2 takes the max value, and if I remove ‘weather_lon’ or ‘tempi’ the the R^2 value drops. R^2_{max} = 0.505827670959 R^2_{min} = 0.503107810554

- At this point, we need to observe that the max & min R^2 values I’m obtaining are so close, so I need to decide which will be the definitive features to consider.

Finally, after evaluate different options, I decided to maintain ‘wspdi’ because, even though the R^2 value is better when removing it from the model, I thought that when it is very windy outside people might decide to use the subway more often. And I decided to remove ‘latitude’ & longitude’ because they are highly correlated with ‘station’.

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

Intercept = -2023060.72806 , W_{fog} = -8.30046521e+02, $W_{precipi}$ = -4.29197719e+03, $W_{pressurei}$ = -3.97411541e+02, W_{tempi} = -1.95389208e+01, W_{wspdi} = 3.28513880e+00, $W_{weather\_lat}$ = 1.06642980e+04, $W_{weather\_lon}$ = -3.25770873e+04

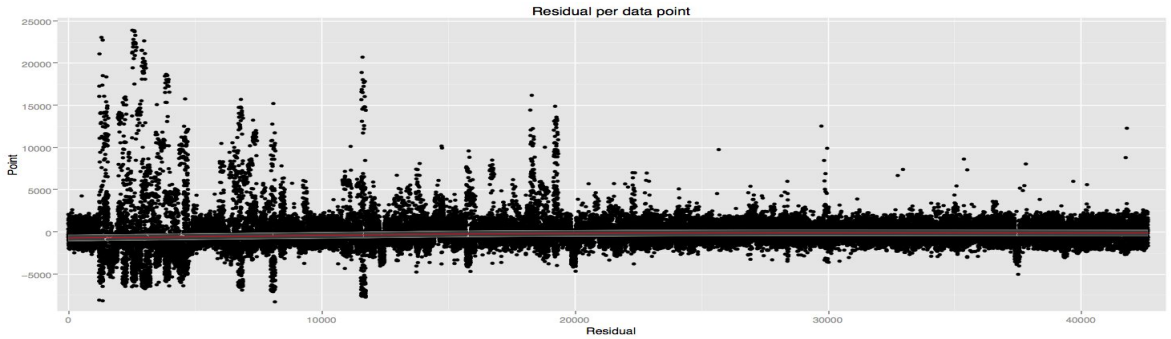
2.5 What is your model's R^2 (coefficients of determination) value?

The coefficient of determination obtained is **0.504706986021**

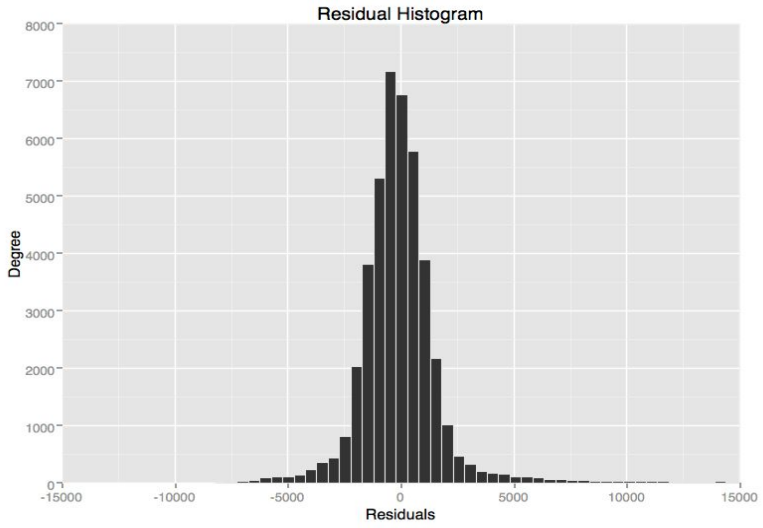
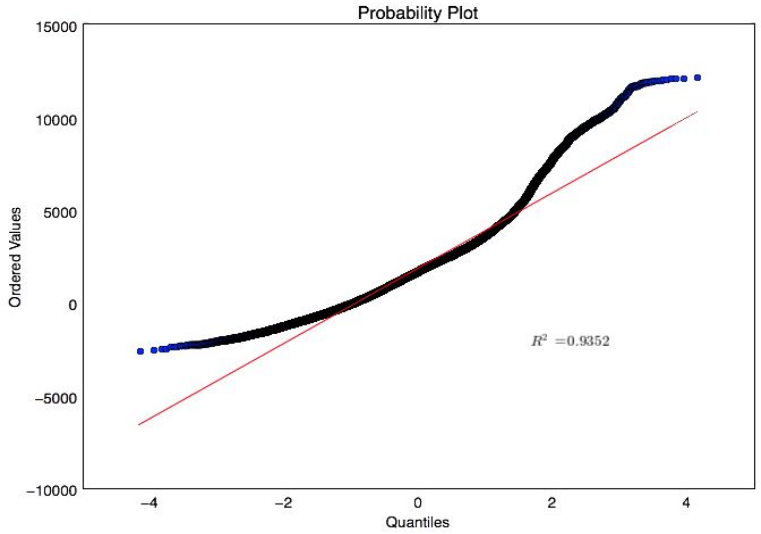
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

An R^2 of 0.5, means that 50% of the variation in ridership is 'due to' or is 'explained by' the selected variables ('*fog*', '*precipi*', '*pressurei*', '*tempi*', '*wspdi*', '*weather_lat*' and '*weather_lon*', combined the dummy variables '*stations*', '*conds*', '*day_week*' *and* '*hour*'). Taking into account that we are trying to explain a human behaviour in certain weather conditions and, as discovered here, "*Social scientists who are often trying to learn something about the huge variation in human behavior will tend to find it very hard to get r-squared values much above, say 25% or 30%"*, I can say that an R^2 of 0.5 is a very good fit value of my regression model.

Regarding the appropriateness of this linear model, I've done a residual analysis, by plotting the Residuals histogram & the Residuals per data point. If the residuals histogram follows a normal distribution and the residuals per point don't follow any special pattern, we can conclude that our linear model is appropriate for this dataset. Let's take a look...



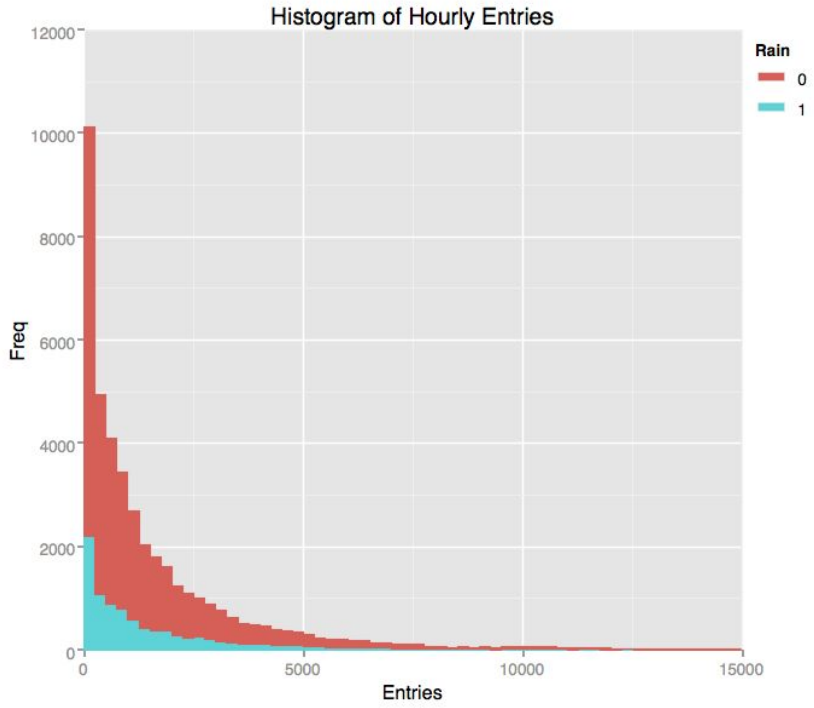
The residual per data point seems to follow a time pattern that a trend to reduce residuals as the time responses increases.

<table>
<tr>
<td> 
</td>
<td> 
</td>
</tr>
<tr>
<td>The residuals Histogram seems to have normal distribution, but if we observe it closely, it has long tails, which suggests that there are some very large residuals a reason to question our linear regression model.</td>
<td>The probability plot confirms what I’ve seen in residuals histogram, a Righ Skew, because it bends up and to the left of the normal line that indicates a long tail to the right. (See Residuals Histogram the values over 10000)</td>
</tr>
<tr>
<td colspan="2">Summarizing:

The model, at first sight, seems appropriate for this dataset, but the residual analysis reveals poor fitting, so I think I will need to review the model and take a look to the dataset to find possibles outliers, and if it’s possible try to use a larger set (Instead of using one month only, use one year or more).</td>
</tr>
</table>

Section 3. Visualization

3.1 ENTRIESn_hourly for rainy vs non-rainy days.



Comments:

As commented in the first section, in this plot we can appreciate the difference between hourly entries in rainy and non-rainy days. (The x scale was truncated to 15000)

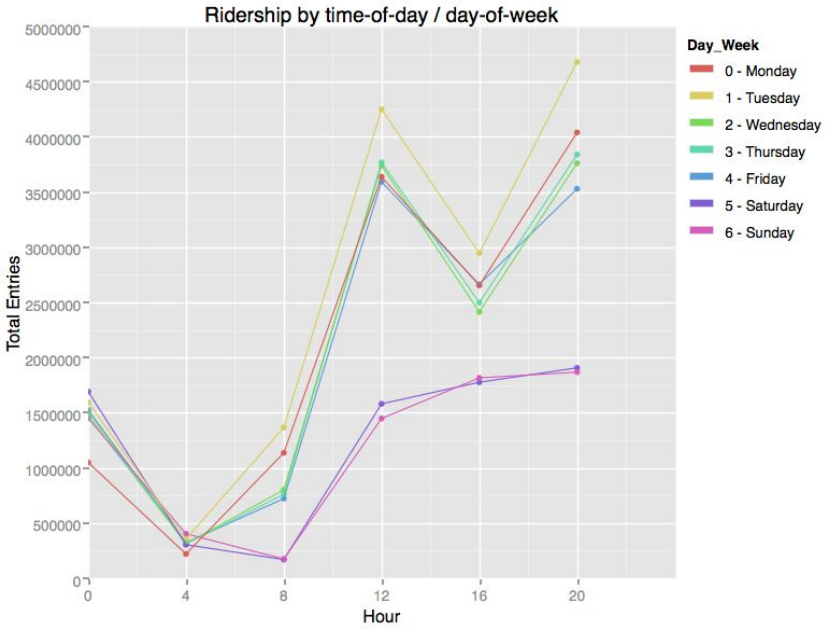
The key insights depicted are:

1. The distributions seems to be exponential.
2. The frequency decreases quicker in no rainy days than in rainy ones.

Code:

```
datafile='improved-dataset/turnstile_weather_v2.csv'
df = pd.read_csv(datafile)
ggplot(df, aes(x='ENTRIESn_hourly', color='rain', fill='rain')) +
geom_histogram(binwidth=250) + ggtitle('Histogram of Hourly
Entries') + xlim(low=0, high=15000) + labs('Entries', 'Freq')
```

3.2 Ridership by time-of-day / day-of-week



Comments:

In this plot I want to show the ridership by time of the day along with the day of week thanks to the color classification done by ggplot.

The key insights depicted are:

1. Week-days (Mon-Fri) the ridership is higher than weekends.
1. The curve shapes are also different among weekdays and weekends.
2. At 16h in weekdays the ridership slows down (people are at work) while in weekends is stable between 12h & 24h.
3. The most part of the ridership along the day is clearly concentrated between 8h and 20h.

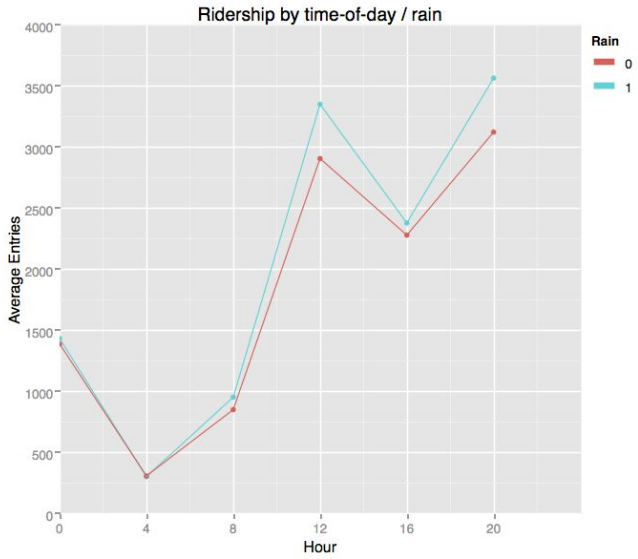
Code:

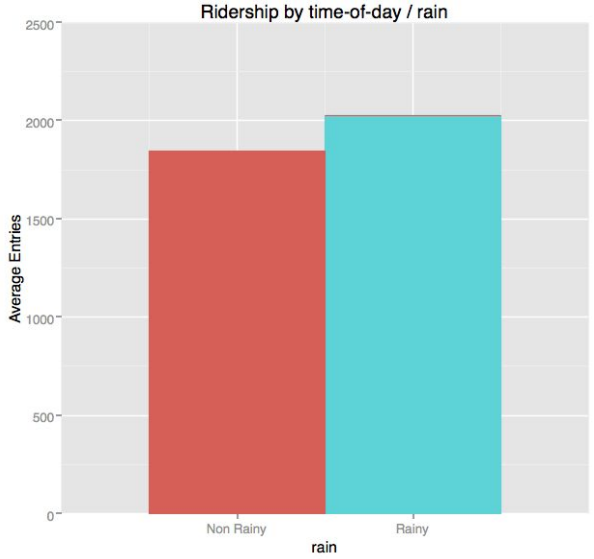
```
datafile='improved-dataset/turnstile_weather_v2.csv'
df = pd.read_csv(datafile)
days_list = ['0 - Monday', '1 - Tuesday', '2 - Wednesday', '3 -
Thursday', '4 - Friday', '5 - Saturday', '6 - Sunday']
days = dict(zip((0,1,2,3,4,5,6),days_list))
df['day_week'] = df['day_week'].map(days)
df_by_Day_Hour = df.groupby(['hour',
'day_week']).aggregate(np.sum).reset_index()
print df_by_Day_Hour
plot = ggplot(aes(x='hour', y='ENTRIESn_hourly', color='day_week',
order='day_week'), data=df_by_Day_Hour) + geom_point() +
ggtitle('Ridership by time-of-day / day-of-week') +
labs('Hour','Total Entries') + scale_y_continuous(limits = (0,
5000000)) + scale_x_continuous(breaks=(0, 4, 8, 12, 16, 20), limits
= (0, 24)) + geom_line()
```

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?

Yes, they do. In the next question I explain how I achieve this conclusion, but essentially, the analysis and experimentation done, demonstrates that the two samples have different Median values and leads me to accept that there is an evidence that the people use more the NYC subway when it's raining than when is not raining. From the data exploration, we can obtain some statistics and graphs very useful that give us some idea if our hypothesis is or not correct, but we need to confirm them with the analysis. As an example let's take a look at two plots showing us the Hours of the day vs. the average of hourly entries, or the total average of hourly entries, in rainy and non rainy days:





These two charts show clearly that apparently on rainy days more people ride the NYC subway than the non-rainy ones.
(`Rain Mean:  2028.19603547 No Rain Mean:  1845.53943866)`

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.

As seen in the Section 1, after explore the data, I discover that the samples didn't follow a normal distribution so I couldn't use a parametric standard test. I needed a non parametric test to compare two independent samples. I decided to use the Mann-Whitney U Test with a null hypothesis like "*There is no Ridership differences between Rainy and No Rainy days ($Me_1 = Me_2$)*" and considering a critical p-value of 0.05. If the two-tailed p value obtained with the test is lower or equal than 0.05, I can reject the null hypothesis. After code the functions using the tools shown at class, I got a two-tailed p value of **5,48e-6**, lower than 0,05, so I can reject the null hypothesis and conclude that there is significant difference between the two samples.

As experimented in section 2, If I drop from the model the variables related with the rain, like 'rain, 'precipi' and 'conds', the coefficient of determination suffers an important fall, so this values must be considered, but I need to count that they are correlated. After experiment with the model with each value separately (trying to achieve the best R^2 value), I decided to don't use the 'rain' feature, include the 'precipi' feature, and use the 'conds' variable as a dummy one.

Section 5. Reflection.

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

I discovered and proved some things in the OLS method applied in my analysis:

- In reality most systems are not linear and OLS method attempt to fit a “plane” through n dimensional data sets. In this case is better to use SGD .
- As the number of independent variables in a regression model increases, its R^2 will always go up (not because the model is better) and can cause serious calculations difficulties. We need to determinate & select only those useful features.
- A training point that has a dependent value that differs a lot from the rest of the data will have a disproportionately large effect on the resulting model, so I think that is a good practice to preprocess the data with an outlier detection algorithm that attempts either to remove outliers altogether or de-emphasize them by giving them less weight than other points when constructing the linear regression model.
- When a subset of the independent variables fed to it are significantly correlated to each other, the method lead to poor predictions. So we need to study the independent variables and its correlation before use them.

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

In relation to the data set, I’m aware that due to a memory & CPU limitation, we need to use an “small” data set but, as commented in Question 2.6, If we want to obtain a good model, we would need to work with at least a year of data, and include the month as variable and maybe more weather variables, like snow.