Short Questions to Analyzing the NYC Subway Dataset
Analyzing the NYC Subway Dataset

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 0. References

Links or references used:

- *Ggplot from ŷhat. (n.d.). Retrieved April 6, 2015, from https://ggplot.yhathq.com/*
- *AnthroSpace. (n.d.). Retrieved April 6, 2015, from http://web.stanford.edu/~cengel/cgi-bin/anthrospace/ggplot-from-python-with-rpy2*
- *Limitations of regression analysis (n.d.). Retrieved April 6, 2015, from http://folk.uio.no/rnymoen/ECON4160_v09_Lect5v9.pdf *
- *Limitations of simple linear regression (n.d.). Retrieved April 6, 2015, from https://stat.duke.edu/courses/Fall00/sta103/lecture_notes/multregr.pdf*
- *Predictive Analytics Techniques. (n.d.). Retrieved April 6, 2015, from http://www.ftpress.com/articles/article.aspx?p=2248639&seqNum=5*
- *Zuur, A., Neno, E., Walker, N., Saveliev, A., & Smith, G. (n.d.). Limitations of Linear Regression Applied on Ecological Data. In Mixed Effects Models and Extensions in Ecology with R.*

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?

Answer:

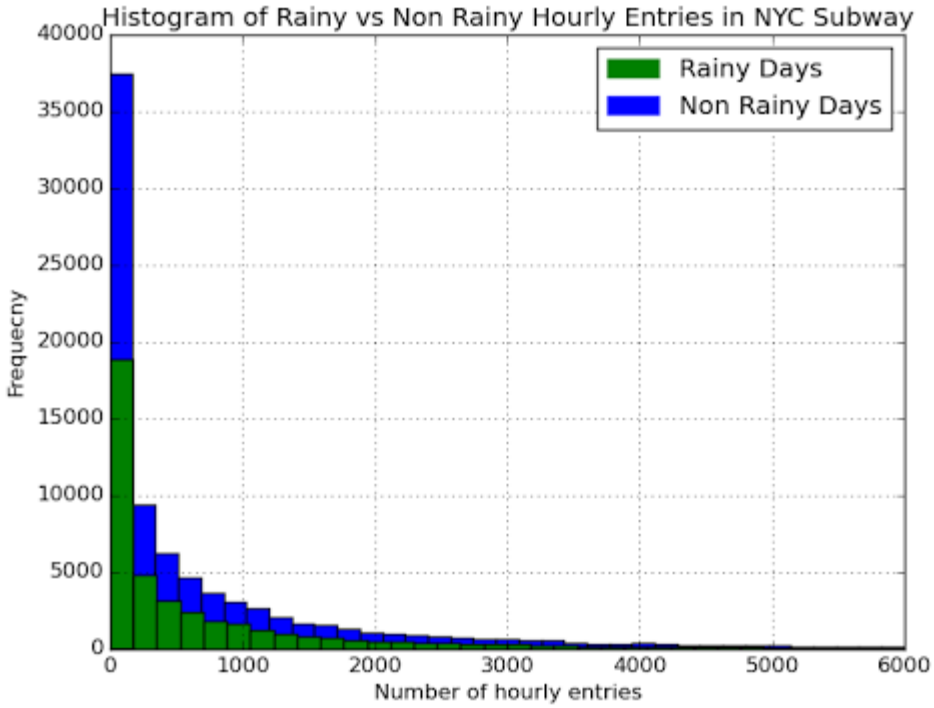
- *I used Mann-Whitney U test to anlayze the ridership from NYC subway dataset.*
- *The Python – scipy.stats.mannwhitneyu function gives a one sided p value as the result. I multiplied this by 2 and used the 2 tailed p value.*
- *The null hypothesis is – The distribution of subway ridership on rainy days is statistically similar to that on non-rainy days.*
- *I used a p-critical 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.

Answer:

- *The histogram plot below showing distribution of the number of hourly entries during and non-rainy days tells us that the **data comes from non-normal distribution**. This violates the*

basic assumptions of parametric tests like t test. Hence, parametric tests cannot be used.



- *Mann-Whiteny U test is applicable for non-parametric data i.e., data which is not drawn from any parametric distribution. In case of NYC dataset, the dataset does not come from normal distribution. Hence, Mann-Whitney U test is applicable.*
- *Also, the following assumptions for the NYC subway data hold true which are necessary for conducting Mann-Whitney U Test:*
    - *Observations from both groups are independent of each other*
    - *The dependent variable is either of continuous or ordinal type. In our case,* ***ENTRIESn_hourly is the dependent variable***
    - *One of the indepedent variables is of categorical type with two gropus. In our case, the independent variable is* ***'rain' which takes two values either 0 (no-rain) or 1 (rain)***
    - *In order to compare the means of two different groups, the distribution of scores for these two groups should have the same shape.*

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.

Answer:

- *Mean on rainy days: 1105.4463767458733*
- *Mean on non-rainy days: 1090.278780151855*
- *Mann Whitney U Test statistic : 1924409167.0*
- *One tailed P-value: 0.024999912793489721*
- *Two tailed P-value – 0.05*

1.4 What is the significance and interpretation of these results?

Answer: The Mann Whitney U statistical test resulted in a one sided p-value of 0.025. The two tailed p value is 0.05. This p-value is less than or equal to the critical value of 0.05 and we can reject the null hypothesis. Hence, we can conclude that the distribution of ridership on rainy days is significantly different from that on the 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?

Answer: I used Stochastic gradient descent method and gradient descent method. In case of stocastic gradient descent method, I used randomly selected training samples for calculating new set of parameter values.

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

Answer: I used the following features:

- *rain – indicating if it rained*
- *precipi – precipitation value*
- *Hour – time stamp*
- *Hour ^ 2 (square term)*
- *Hour ^3 (cubic term)*
- *Hour ^4 (order 4 term)*
- *meantempi – the mean temperature on the currenty day*
- *meanpressure – the mean pressure on the current day*
- *fog – presence of fog*
- *week of the day.*

Yes, I used the dummy features. I converted week of day to dummy_days before using them.

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."

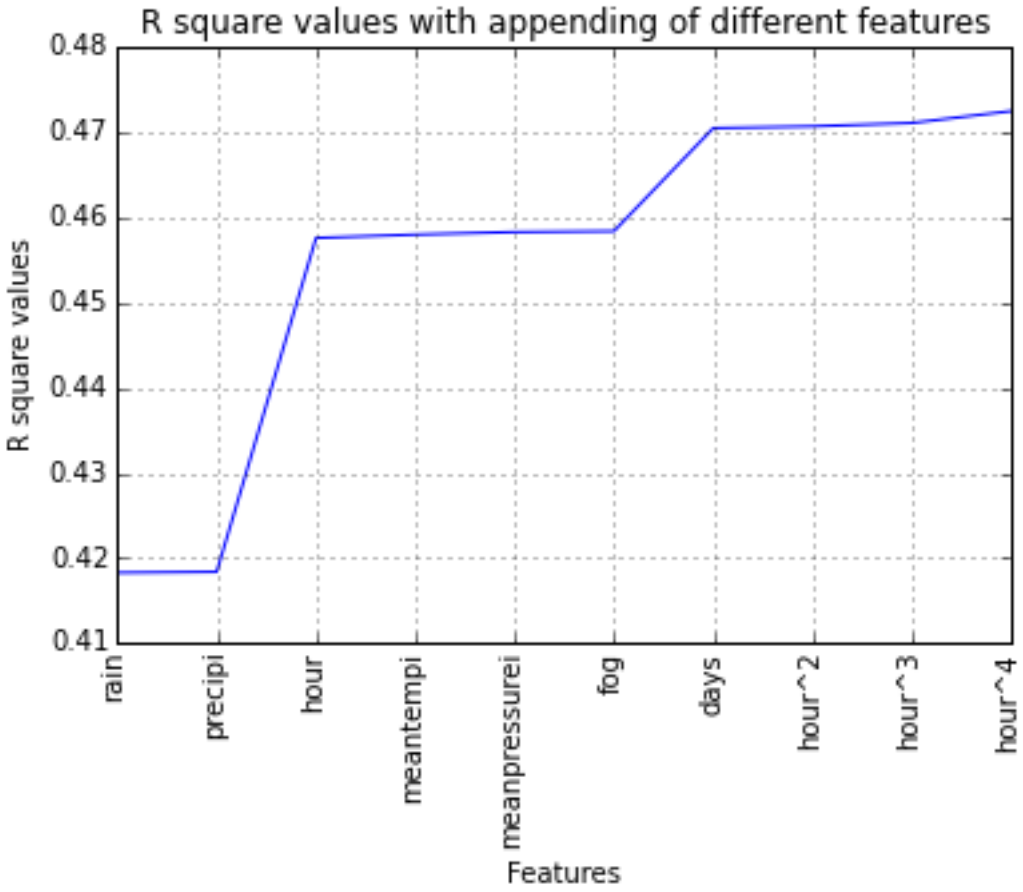
- Your reasons might also be based on data exploration and experimentation, for example: "I used feature X because as soon as I included it in my model, it drastically improved my R2 value."

Answer: Following are the reasons for choosing the features in my model:

1. *Weather Indicators: The weather conditions may directly impact the behaviour of a traveler. In case of bad weather, a person may prefer to take a subway than drive out using his own vehicle or take the road transport.*
    - *For instance, in case of rainy days, people would not prefer to drive their own vehicles due to various reasons such as inconvenience, risk, etc. On such rainy days, people may prefer to take the subway. Hence, I have used rain feature.*
    - *Similarly, precipitation indicates the amount of rainfall received. Heavier the rainfall, the chances might be more that a person takes the subway. In case of minor drizzles, a person may not mind taking out his own vehicle. Hence, I have used precipitation.*
    - *Temperature and pressure are similar indicators that might affect the ridership. Higher outside temperature may cause a person to take a more peaceful journey in subway which is air-conditioned (I understand that this is a strong assumption I am making. The point is, in case of higher outside temperatures may increase the fuel cost due to fuel inefficiency if the person is using self transport).*
    - *Fog is another important feature as the presence of fog directly impacts road travel. It would be difficult to drive on roads in case of fog which may increase the subway ridership.*
2. *Time indicators:*
    - *Time of day: This is a very significant feature as peak time (office hours) will have higher traffic which leads to more ridership. Also, a plot of the hourly entries v/s time of day indicates that there is a relationship between the two.*
    - *There seems to be a non linear relationship between time of day and hourly entries. Hence, I have used higher powers of hour (upto 4 degrees).*
    - *Day of week: The number of people traveling in general through any mode of transport would be more during weekdays as they have to attend offices and their businesses. Hence, this would be another important feature.*
3. *It can be seen from the table and graph below that the r-squared value keeps increasing as I append the above mentioned features. This shows that as the feature set is appended with various features explained above, the percentage variation explained in the dependent variable increased. This justifies the use of the features.*

Feature Name	***R Square Value***
rain	*0.4183*
rain + precipi	*0.4184*
rain + precipi + hour	*0.4576*
rain + precipi + hour + meantempi	*0.4580*
rain + precipi + hour + meantempi + meanpressurei	*0.4583*

rain + precipi + hour + meantempi + meanpressurei + fog	*0.4584*
rain + precipi + hour + meantempi + meanpressurei + fog + days	*0.4705*
rain + precipi + hour + meantempi + meanpressurei + fog + days + hour^2	*0.4707*
rain + precipi + hour + meantempi + meanpressurei + fog + days + hour^2 + hour^3	*0.4711*
rain + precipi + hour + meantempi + meanpressurei + fog + days + hour^2 + hour^3 + hour^4	*0.4725*



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

Answer:

Following are the coefficients of the non dummy features using gradient descent method:

Feature	***Coefficient value***
Rain	*-34.61*
Precipi	*-25.51*
Hour	*315.21*
Meantempi	*-54.42*
Meanpressurei	*-26.43*
Fog	*36.33*
Hour^2	*268.77*
Hour^3	*41.35*
Hour^4	*-176.04*
Day Of Week 0	*-2.5*
Day Of Week 1	*59.45*
Day Of Week 2	*88.91*
Day Of Week 3	*62.72*
Day Of Week 4	*69.38*
Day Of Week 5	*-84.2*
Day Of Week 6	*-174.49*
Intercept term	*1095.34*

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

Answer: R square value using Gradient Descent Method: 0.4725

R square value using Stochastic Gradient Descent Method: 0.4648

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?

Answer: An R2 value of 0.4725 (using gradient descent method) means using the current set of features used to calculate the value of theta, 47.25% of the variation in the output can be explained. This also means that there is lot more variation that is not being captured by the features that are used.

This model cannot be used if precise prediction is required. However, in order to understand and predict the trend, this model should be sufficient.

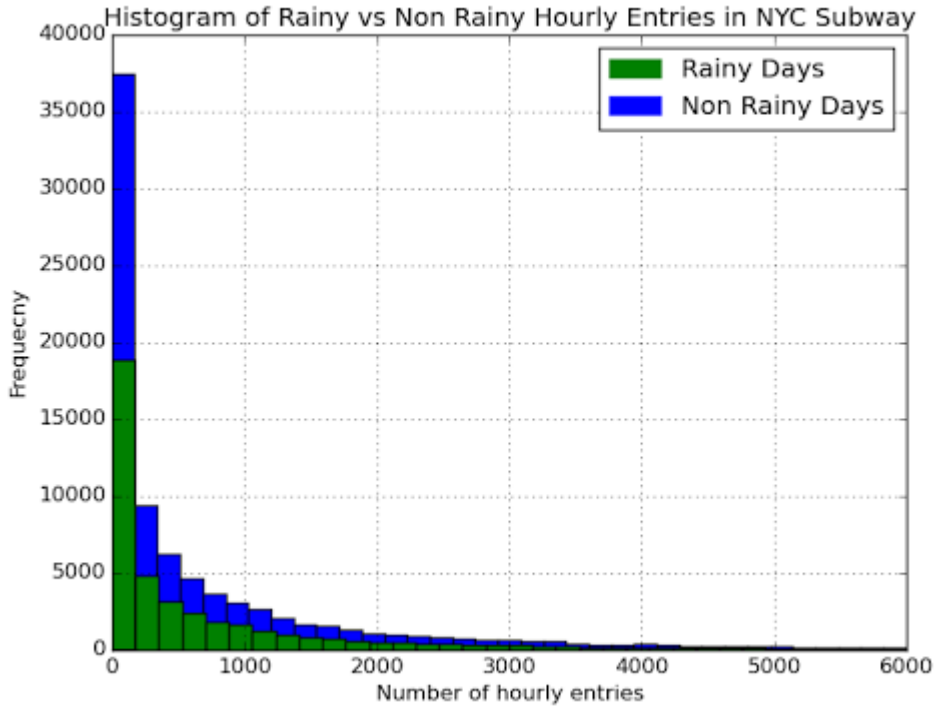
Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

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.

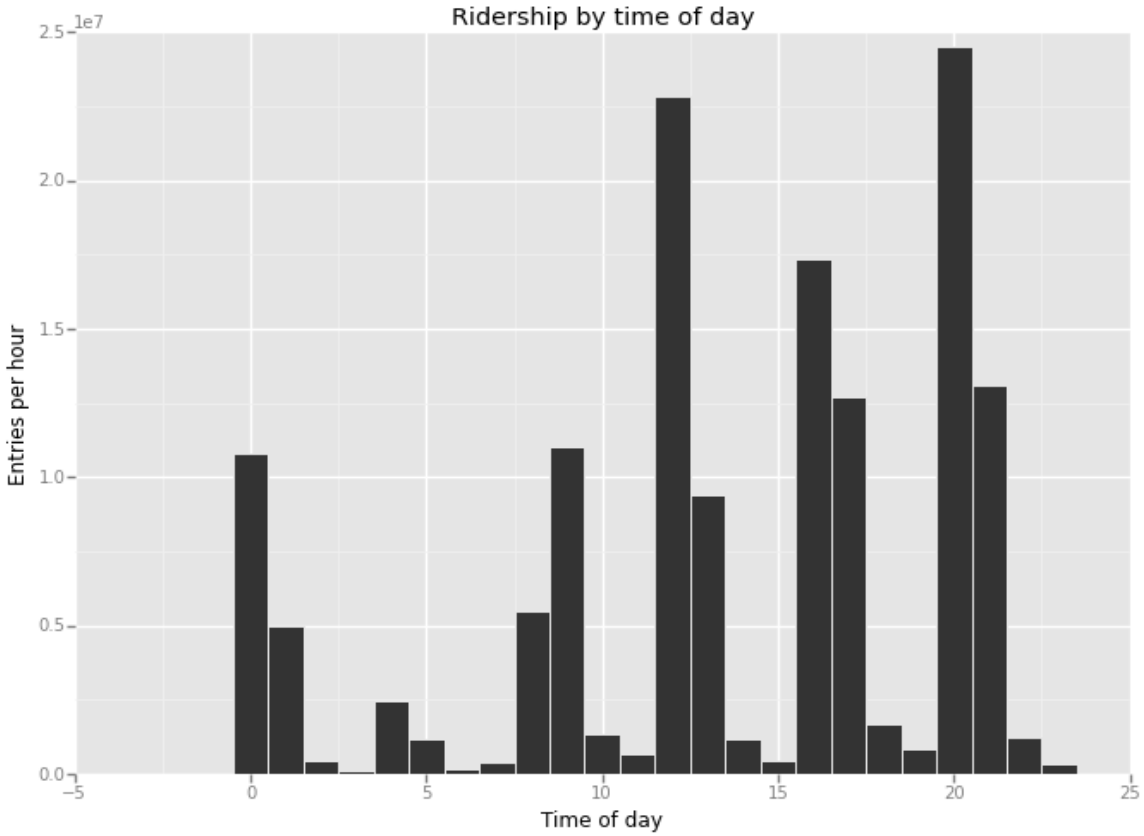


Note: The entries per hour > 6000 have been clipped off for better visualization

Comment: It can be observed from the plot above that the data for non-rainy days is more than that for rainy days. The distribution of ridership for rainy and non-rainy days look similar.

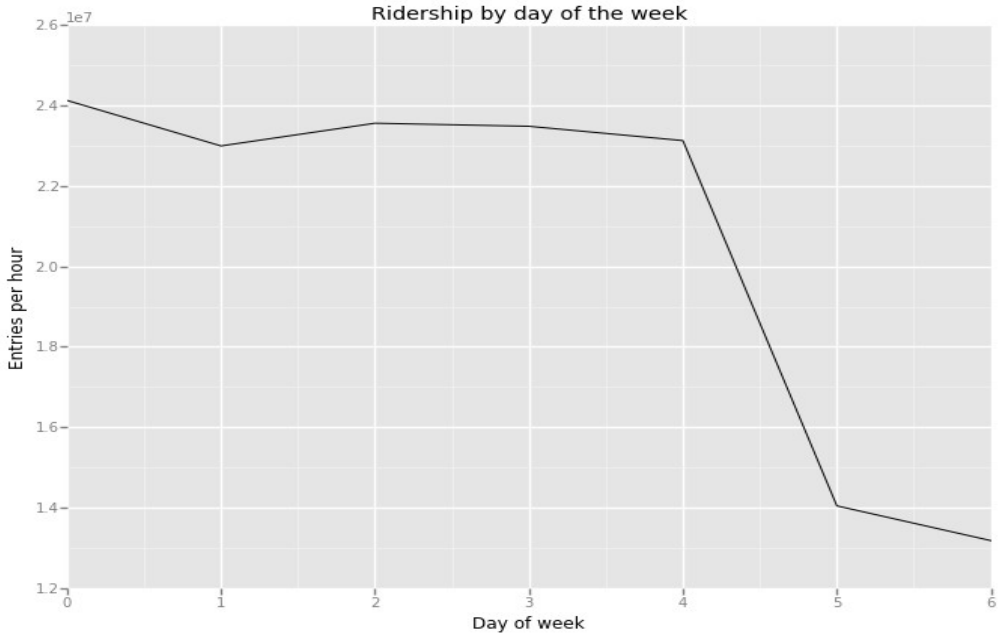
3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:

- Ridership by time-of-day



Comment: The entries per hour varies non-linearly with the time of day. The peaks at different times of day (9, 12, 16, 20) indicate peak hours during which the traffic is very high.

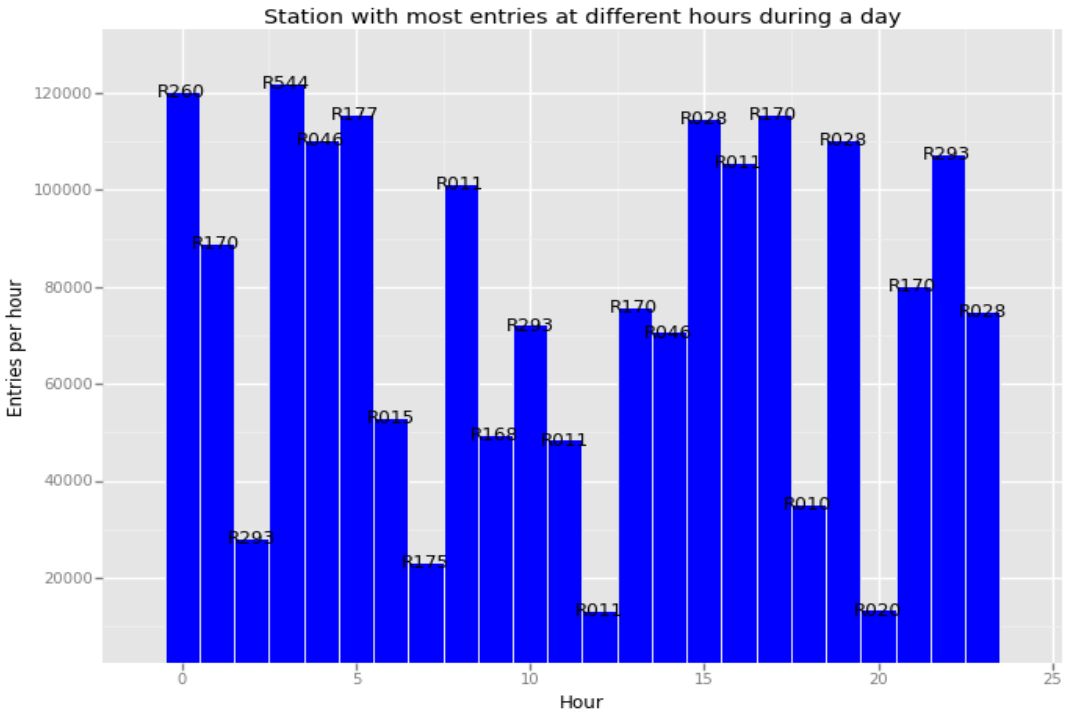
- Ridership by day-of-week



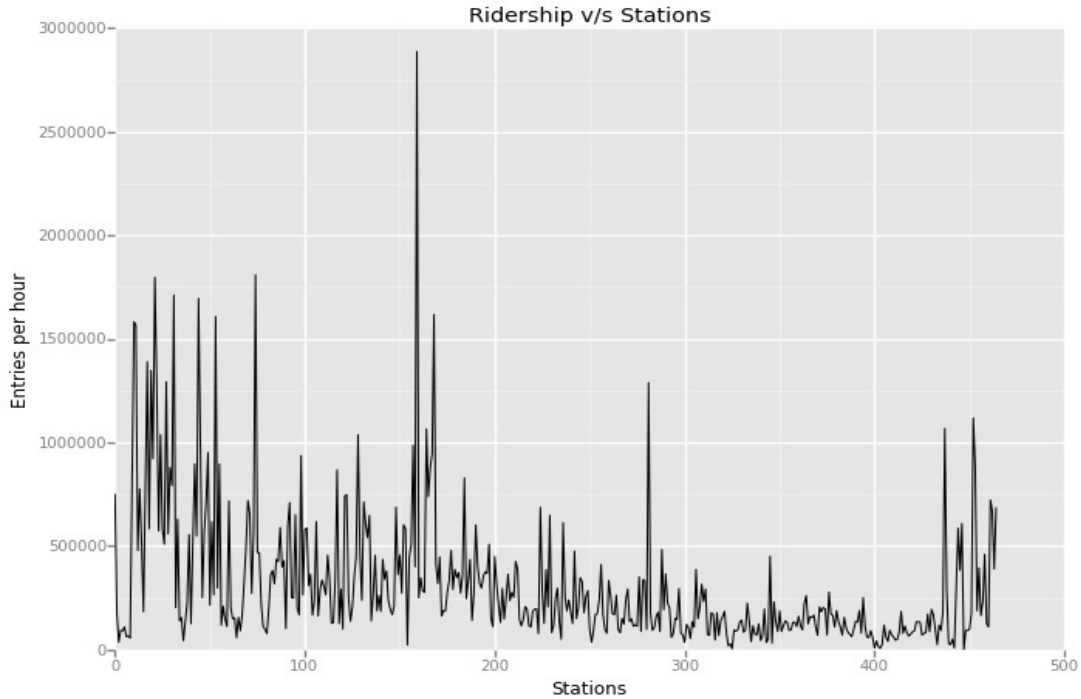
Comment: It can be seen that the ridership is low during weekends compared to weekdays. This might be due to the reason that most of the offices/businesses are closed during weekends.

-

- Stations with most entries during time of day



- Ridership at different stations



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?

Answer: After looking at the statistical tests peformed on the subway dataset, we may interpret that more people 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.

Answer: Following are the reasons for my interpretation:

1. The Mann Whitney U test on the NYC subway dataset for distribution of ridership on rainy and non-rainy days resulted in:

p-value: 0.05 (2 sided p value) which is less than or equal to p-critical value of 0.05. Hence, we can reject the null hypothesis and interpret the results of the test that the ridership during rainy days is stiatistically different from that on the non-rainy days.

Since the mean ridership on rainy days (1105) is greater than that on the non-rainy days (1090) we can conclude that more people the subway when it is raining.

2. Also, when we perform linear regression to predict the hourly entries, using rain as a feature resulted in an increased R square value of 0.4183 compared to 0.4091 when rain is not used. This explains that using rain as a feature increases the amount of variation predicted by the independent variables.

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.

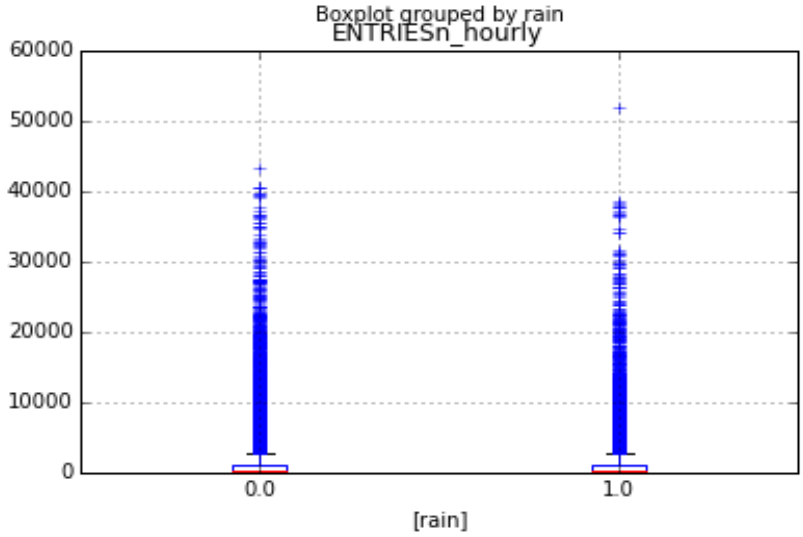
Answer:

- *Considering some of the independent variables like rain, precipi, fog, day of the week, time during the day, etc, most of these variables are binary variables. Logistic regression would be a better method to apply to such variables because the cost function used in linear regression does not suit for binary predictors.*
- *While we perform linear regression on the given dataset, we have to keep in mind the assumptions that need to hold good before we apply linear regression:*
    1. *Linear relationship between the dependent and independent variables*
    2. *Normality of data*
    3. *Homogenity of data*
    4. *Independence*
- *Some of these assumptions don't hold good incase of our dataset. For instance, we are using variables like rain and mean temperature which are not independent. That is, there is some*

correlation between these two variables. This may violate the assumption of independence.

It is difficult to establish the normality of the data in this case. And even if the data is normally distributed, the variance for different variables is not the same. Hence, homogenity of data does not hold good in many cases.

- *After looking at the scatter plots between the dependent variable and different independent variables, it does not seem that some of these variables share a linear relationship with the dependent variable. Hence, linear regression may not be suitable for the current dataset.*
- *Another important shortcoming of linear regression is that it is very sensitive to outliers. It can be seen from the box plot bekow that there are many outliers for ENTRIESn_hourly in case of rainy and non-rainy days. These outliers can cause in incorrect fitting of the data.*



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