Analyzing the NYC Subway Dataset

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, and 4 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

Please include a list of references you have used for this project. Please be specific - for example, instead of including a general website such as stackoverflow.com, try to include a specific topic from Stackoverflow that you have found useful.
Answer:
http://pandas.pydata.org/pandas-docs/stable/
https://github.com/ivanov/vim-ipython#id2
http://goo.gl/HBbvyy
http://docs.python.org/2/library/datetime.html
http://statsmodels.sourceforge.net/0.5.0/generated/statsmodels.regression.linear_model.OLS.html
http://scikit-learn.org/0.14/modules/generated/sklearn.linear_model.LinearRegression.html
http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
http://pandas.pydata.org/pandas-docs/stable/visualization.html#histograms
http://scikit-learn.org/0.14/modules/generated/sklearn.linear_model.SGDRegressor.html
http://ggplot.yhathq.com/docs/index.html
https://pypi.python.org/pypi/ggplot/
http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
http://en.wikipedia.org/wiki/Student%27s_t-test
http://www.statsdirect.com/help/default.htm#nonparametric_methods/mann_whitney.htm
https://statistics.laerd.com/premium-sample/mwut/mann-whitney-test-in-spss-2.php
https://github.com/yhat/ggplot/issues/376
http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.probplot.html
https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot

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:
Statistical test: Mann-Whitney U test
Two-tail P value

Null hypothesis: given two random samples of average numbers of subway entries per hour, the first calculated from population of subway entries on rainy days and the second calculated from the population of subway entries on non-rainy days, the probability that the first average will be higher than the second is 0.5.

Two tailed p critical: p = 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:
According to histogram of ENTRIESn_hourly, data are not normally distributed. There are enough observations to guarantee normality of the mean according to CLT. However t-test assumption is that the observations are drawn from the normal distribution. Therefore using t-test may produce wrong results in our case. Therefore we use Mann-Whitney U test that doesn't require the normality of the distributions.
Mann-Whitney U test is valid if the following assumptions are satisfied:

- Observations are independent of each other inside groups and between groups. We can assume this in our dataset, however the independence doesn't hold in reality, as we will discuss in conclusion of this document.
- Two groups have the same shape of the distribution. This is true according to histograms of ENTRIESn_hour on rainy days and non-rainy days.
- Dependent variable is continuous or ordinal. In our case we can assume ENTRIESn_hour is continuous.

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:
Two tailed p value: p = 0.0498
Mean with rain: 1105.44
Mean without rain: 1090.27

1.4 What is the significance and interpretation of these results?
Answer:
Given significance level of 5% the results are statistically significant and the average usage of subway on rainy days statistically differs from the usage on 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. OLS using Statsmodels or Scikit Learn
2. Gradient descent using Scikit Learn
3. Or something different?

Answer: OLS and Gradient descent. OLS was chosen to report results, because it gives a higher R^2.

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

Answer: Hour, minute, day_of_week
Dummy variables to represent UNIT were used

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."

Answer: Hour, minute, day_of_week and dummy variables for UNIT appear to be statistically significant in predicting the ridership. Other existing variables in the dataset are not statistically significant. This can be seen in two ways. Firstly, inclusion of Hour, minute, day_of_week and dummy variables into the model improves R^2 in the first to third digits of accuracy, while inclusion of any other variable does not. Secondly, the statistical model produces p-values for every predictor. It can be seen that only for the Hour, minute, day_of_week and dummy variables this p-value is less than 0.05 significance level and we can conclude they are significant.

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

Answer:
Intercept: 1147.48
Hour: 65.49
Day of Week: -84.36005743
Minute: -11.26721228

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

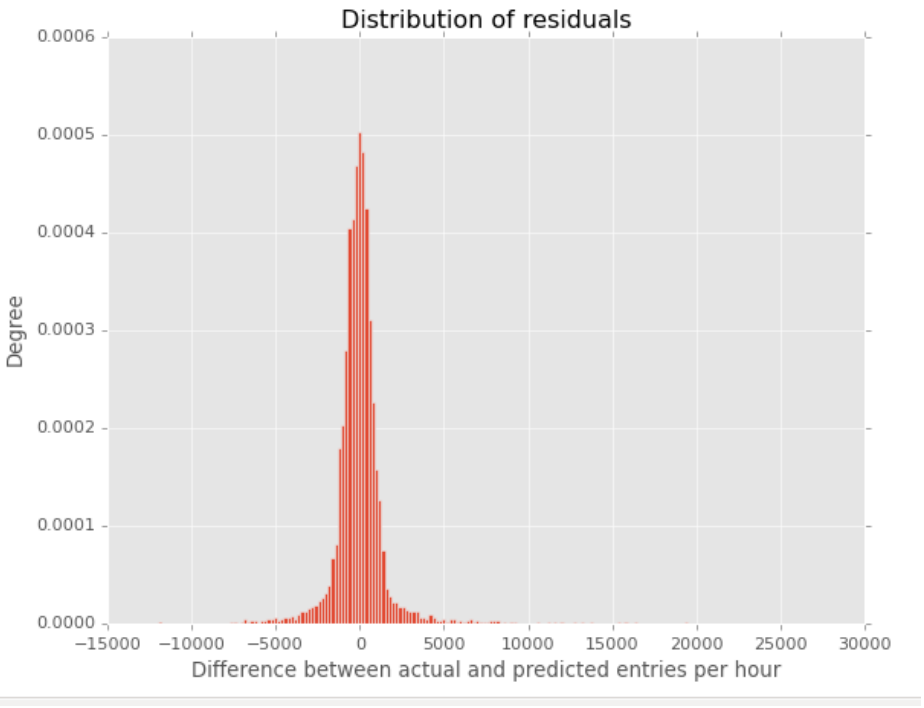
Answer: 0.4868

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?

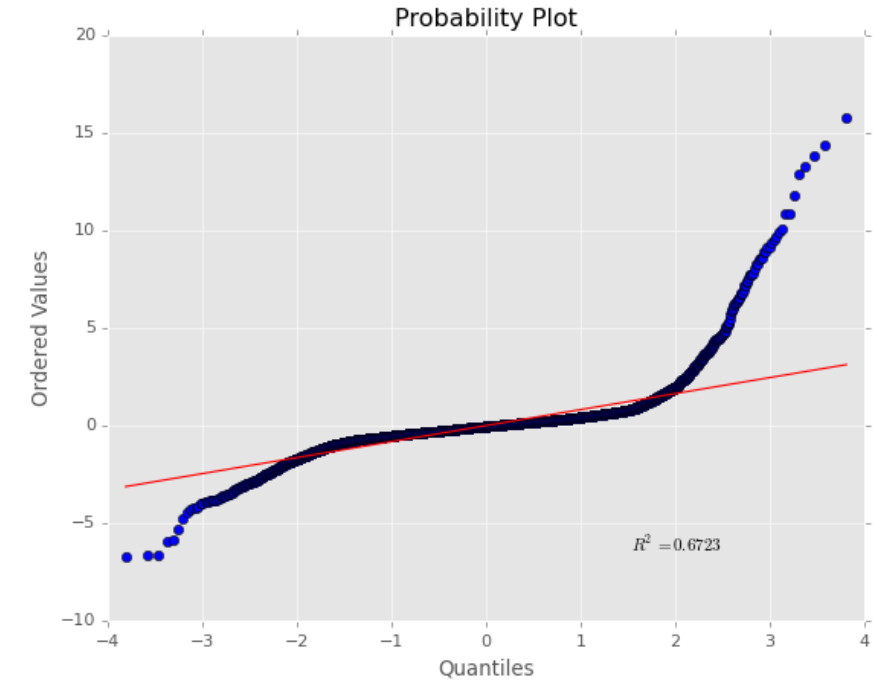
Answer: R^2 tells us that only 48.68% of ridership is explained by the selected predictors (Hour, minute, day_of_week and UNIT). One thought to improve the model is to split a dataset by stations and apply the linear model for each station, if we have enough observations for each station. Or to avoid many models, we could group stations into similar clusters before application of the model. This would reduce the dimensionality of data and we could get a better fit. We could also use non-linear transformations, for

example by calculating Hour^2, Hour^3, because the number of entries dependency on the hour of the day is non-linear.

To access the fit of the linear regression, we can plot the residuals:

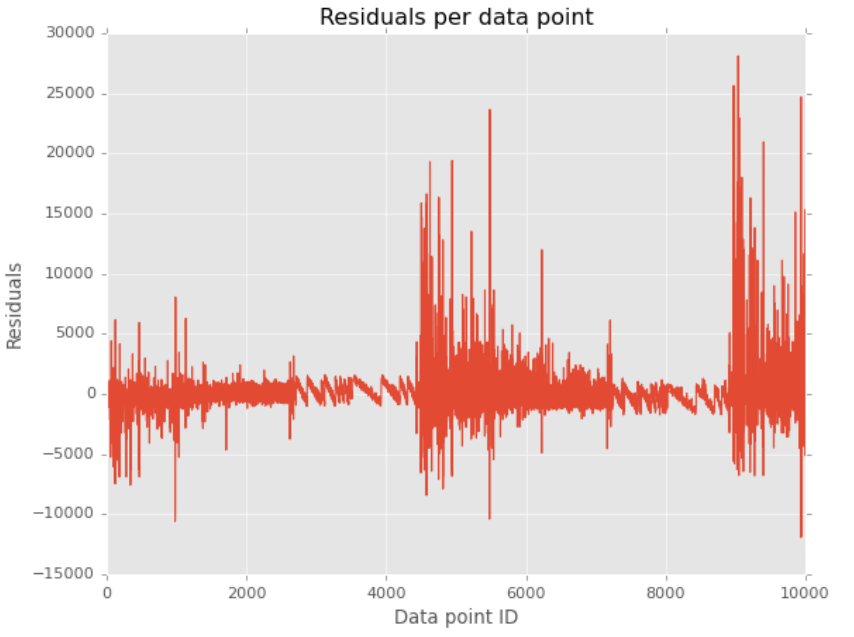


From this graph we can see that there present long tails, which are also non-symmetric. To examine in more detail, here is a Q-Q plot for normalized distribution of residuals (x-axis) compared to the normal distribution (y-axis):



This suggests some residuals are large and do not obey normal distribution, so the fit of the linear model can be questioned.

Finally, let's plot the residuals per data point:



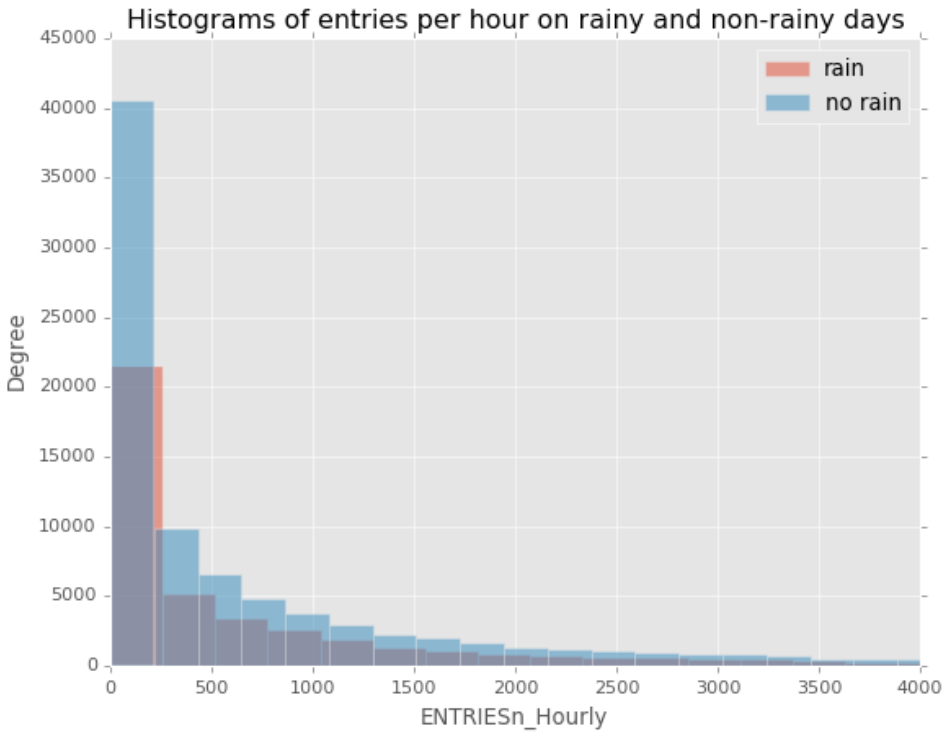
Irregularity and oscillations are observed here. Linear model wasn't able to deal with it.

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.

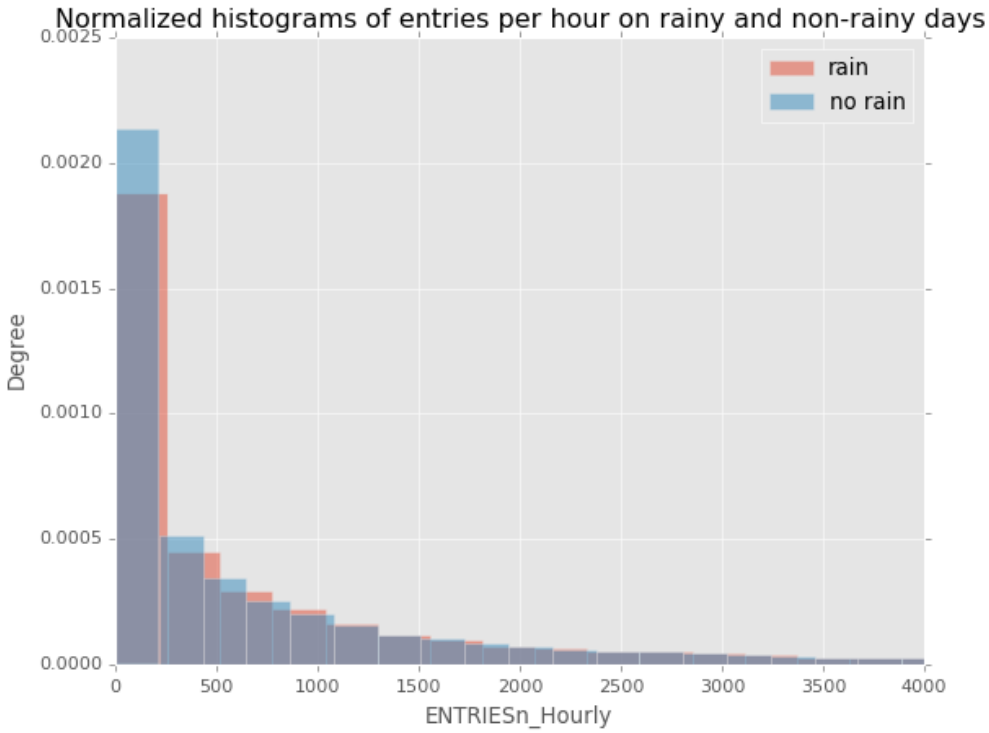
- Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.



Key insights:

- Distribution of hourly subway entries is not Gaussian, rather looks like Exponential. This fact influences the statistical test choice for comparing averages.

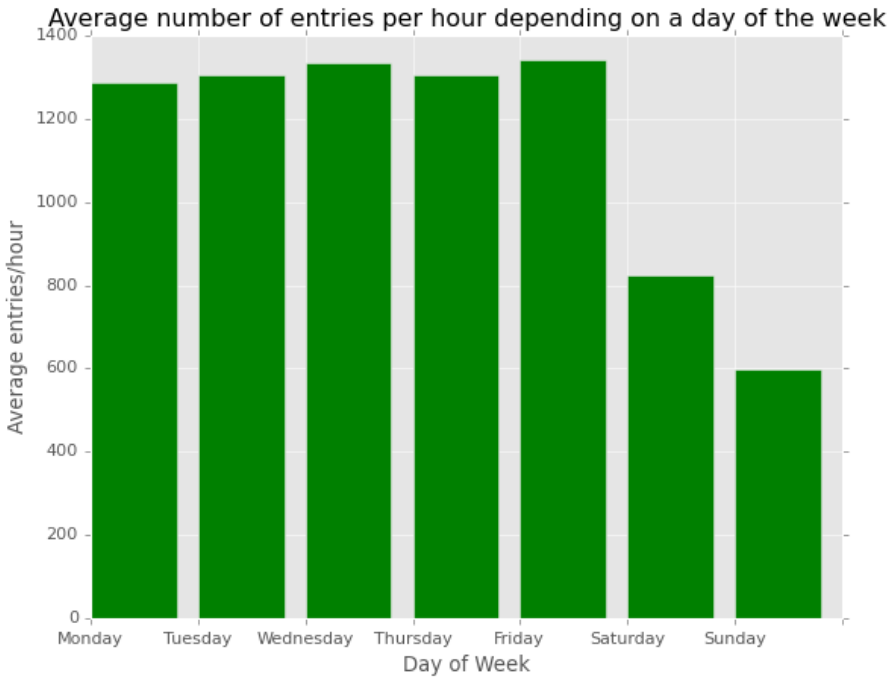
- Graph suggests there are more observations from non-rainy days compared to rainy days. In fact there are about twice as many observations for non-rainy days. Therefore to get more information, it is good to see the normalized histogram:



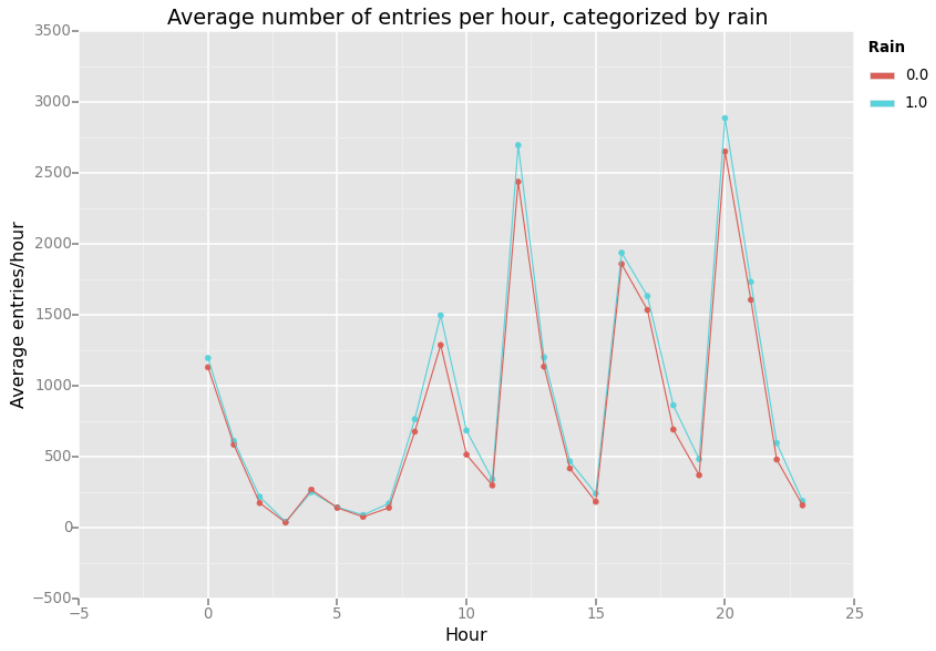
- From this histogram we can see that the average of the entries per hour on non-rainy days is lower than on rainy days. If we assume both distributions are exponential, than 'non-rainy' one has a larger lambda and since the expectation is 1/lambda, it follows that the expectation is smaller. However the difference between averages is apparently small.

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
- Ridership by day-of-week



Key insights: Average number of entries per hour is lower on weekends.



Key insights:

- Average number of entries per hour is lower on weekends
- Average number of entries per hour is lower on non-rainy days
- There are certain times during the day when the peaks of subway entries are observed

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: More people ride when it is raining. As the Mann-Whitney U-test has shown, the difference is statistically significant. We can also see this from the figure above.

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: The Mann-Whitney U-test has shown that the difference is statistically significant. The linear regression doesn't show the same, but it is a limitation of the method. For example, look at the figure above. Assume we fitted a regression line having one variable only, 'Hour'. If we don't take into account higher order predictors such as 'Hour squared', the graph of our prediction would look indeed like a line. The point is that this line would be almost the same if we ran the linear regression on data with rain or on data without rain, because the variability of ridership in respect to 'rain' given 'Hour' is negligible compared to the variability in respect to 'Hour' given 'rain'.

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: The dataset contains information about date, time and weather. However there are other factors that could influence the ridership on rainy days vs. no-rainy days. For example, on rainy days the traffic may have more jams and more people would use subway. Therefore we can't say that rain caused more people to take subway. We can only say that people use subway more on rainy days. And even the last statement may be not accurate given this dataset. We can see that it contains data collected for one month only, from 2011-05-01 till 2011-05-30. Besides the fact that it contains at least twice as many non-rainy records compared to rainy records, it could well happened that during this month there was a construction work going. Because of that construction it could be difficult to get to work on time by car and people preferred subway. This would contribute bias to data and therefore bias for our conclusions.

The linear regression model is one of the simplest models available. It is often enough to use it and get good results. However in our case the results of linear regression are not

satisfactory. The fit of the model was analyzed in Section 2.6 by investigating residuals. The analysis has shown that we didn't 'learn' much by using the linear model.

Though 'linear' in the method name refers not to the data, but to coefficients, even if we introduce new predictors into the equation by non-linear transformation of existing, we may not reach the desired accuracy. On the other hand other methods, which deal more efficiently with non-linear data, could potentially achieve better accuracy.

As for the statistical method, we assume that observations are independent of each other. However the independence doesn't hold. For example, if more people enter the subway at 8 AM in the morning to get to work, it is likely more people will enter subway at 6 PM in the evening to come back home.

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

Answer: Only some trivial insights, such as 'ridership is less on weekends' and 'ridership has peaks at certain hours'.