Name - Punit Mehndiratta

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, 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

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 - References used -

Reference 1 - Reindexing timeseries from dtype to datetime-dtype
stackoverflow.com/questions/13654699/reindexing-pandas-timeseries-from-object-dtype-to-datetime-dtype

Reference 2- Python strptime and strftime behaviour

https://docs.python.org/2/library/datetime.html#strftime-strptime-behavior

Reference 3 - String formatting and strip punctuation , formatting in data frames

stackoverflow.com/questions/265960/best-way-to-strip-punctuation-from-a-string-in-python

stackoverflow.com/questions/16729483/converting-strings-to-floats-in-a-dataframe

Reference 4 - ggplot documentation

docs.ggplot2.org/current/geom_histogram.html
ggplot.yhathq.com/docs/geom_histogram.html

Reference 5- Logging tutorial Python

https://docs.python.org/2/howto/logging.html#logging-basic-tutorial

Reference 6 - OLS in Python

statsmodels.sourceforge.net/devel/examples/notebooks/generated/ols.html

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 -
In order to analyze the number of entries at subway Vs weather(rain or no rain) , we performed Mann-Whitney U test, since the distribution was not normal.
Mann-Whitney U test provided two tailed p value of p =0.025 which is lower than or equal to our p-critical value (p<= 0.025 one sided p critical or p<= 0.05 two sided p-critical value) and number of entries were significant i.e. > 20 hence the null hypothesis is considered not to be true.

The Null hypothesis suggest that there is no statistical difference between Number of Hourly entries when it rain or it doesn't rain.

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-
Since the distribution is not normal (based on Histogram plot for the data as well as results of Shapiro-Wilk test shows it not to be a normal or probability density function), we can't perform the Welch's t-test, hence we performed Mann-Whitney U test which is applicable for testing two populations with unknown distribution.
Here we assume that the number of Hourly entries at one particular time interval were independent of number of Hourly entries at another time interval (i.e. the two data did not influence each other hence allowing us to Use Mann-Whitney Test)

Also the number of entries are significant (> 20) hence the Mann-Whitney U test provides a good statistical test. It showed that there is statistical difference between subway ridership on Rainy Vs Non rainy days.

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 of Hourly entries when Rain =1105.4463767458733
Mean of Hourly entries when not raining = 1090.278780151855,
u value for the data =1924409167.0
p value = 0.024999912793489721 (which is <= p critical value of 0.025 for two sided test)

1.4 What is the significance and interpretation of these results?

Answer -

The Test showed that there is statistical and significant difference between subway ridership on Rainy Vs Non rainy days. Hence null hypothesis is not true.

On average there is more ridership during rainy days rather than 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 - Gradient Descent

Also i have used OLS using statsmodel

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

Answer -
The features(input variables) used for the analysis are
‘rain’ (whether it rained or not)
’percipi’ (precipitation at that time)
‘Hour’ (hour of the day)
‘mintempi’ (minimum temperature)
‘fog’ (foggy or not)
‘meanwindspdi’ (mean wind speed)

The dummy variables used is ‘UNIT’ (representing different stations).

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 -

Based on Intuition (ex- fog and rain as well as cold temperature cause people to use subway more often generally) and PCA i decided to try certain variables and check whether it improved my R2 value which it did in many cases.Adding ‘fog’ . Hour, mintempi and meanwindspdi to rain, percipi and Hour variables improved the R2 value to 0.4655. (using gradient descent)

Using OLS the value is 0.485 (somewhat increased) but the no. of observation are also 10000.

2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?
Answer -
The weights of non-dummy features were represented as below obtained using OLS
(by removing unit from the feature and calculating the coefficients)

Feature	Coef
rain	-1.3881
precipi	-164.1548
Hour	55.6398
mintempi	-10.7569
meanwindspdi	50.0675
fog	335.9967

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

Answer -

Using Gradient descent	R2 value =	0.465488999034
Using OLS summary report	R-squared:	0.485

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 -
Since the R2 value is quite high (almost close to 0.5) , this shows the or coefficient of determination(R2) is a fairly good and our model of using gradient descent (or OLS) is reasonable to explain the impact of the variables we chose (fog, rain , minimum temp , Hour of day etc.) on the ridership of NYC subway.

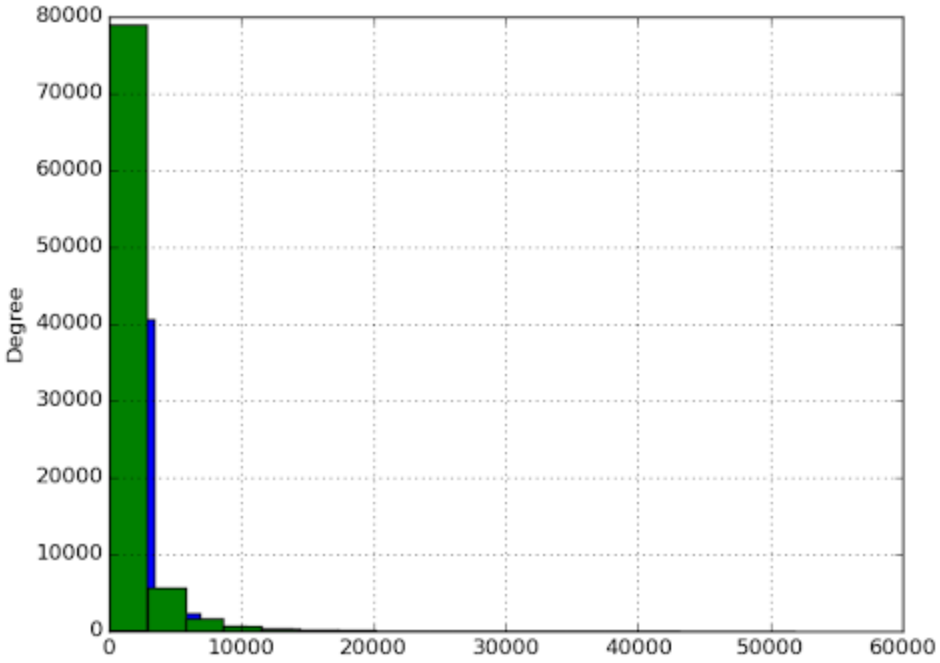
Section 3. Visualization

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.

Answer -

Histogram for Ridership when it Rains Vs Not raining (binsize =15)



ENTRIESn_hourly - X axis
Frequency - Y axis
Blue is representing the ridership with ‘no rain’
Green plot represents ridership when its raining.

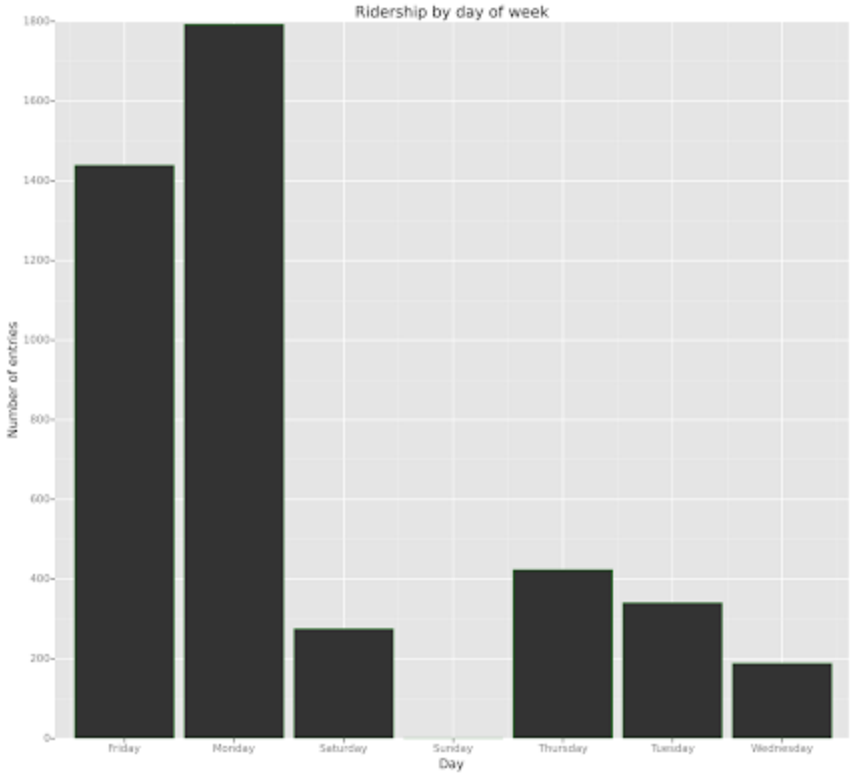
The above figure shows more ridership when raining Vs Not raining. Also it shows that the ridership plot is not a normal distribution.

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

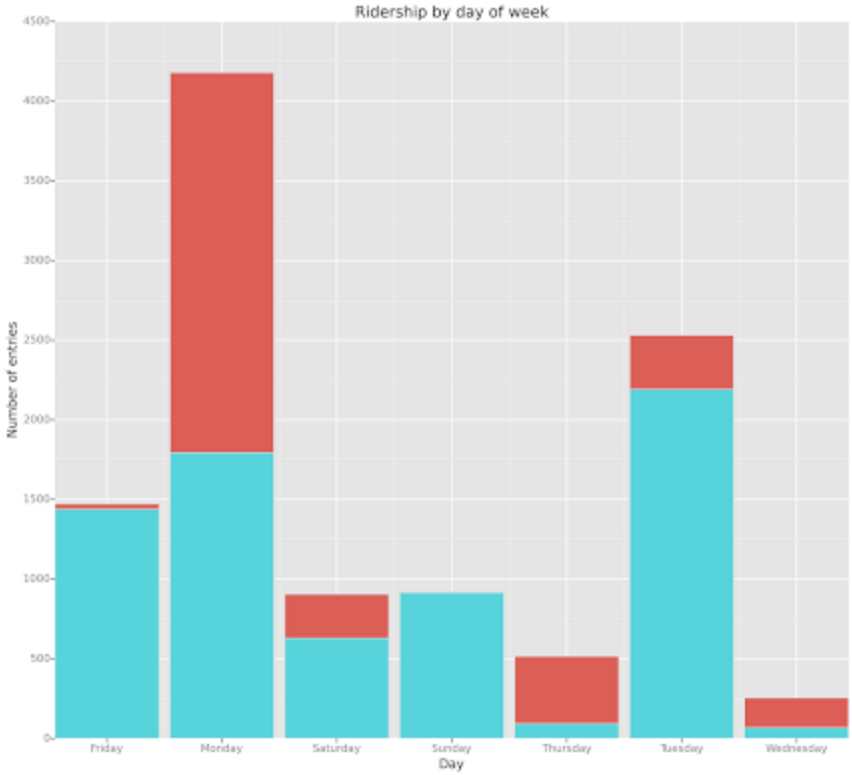
Answer - Below are many visualization for ridership by day of the week , or day of the week with variability of Rain or ridership by units , or ridership by Hour of the day.

Ridership by Day of the week - Histogram



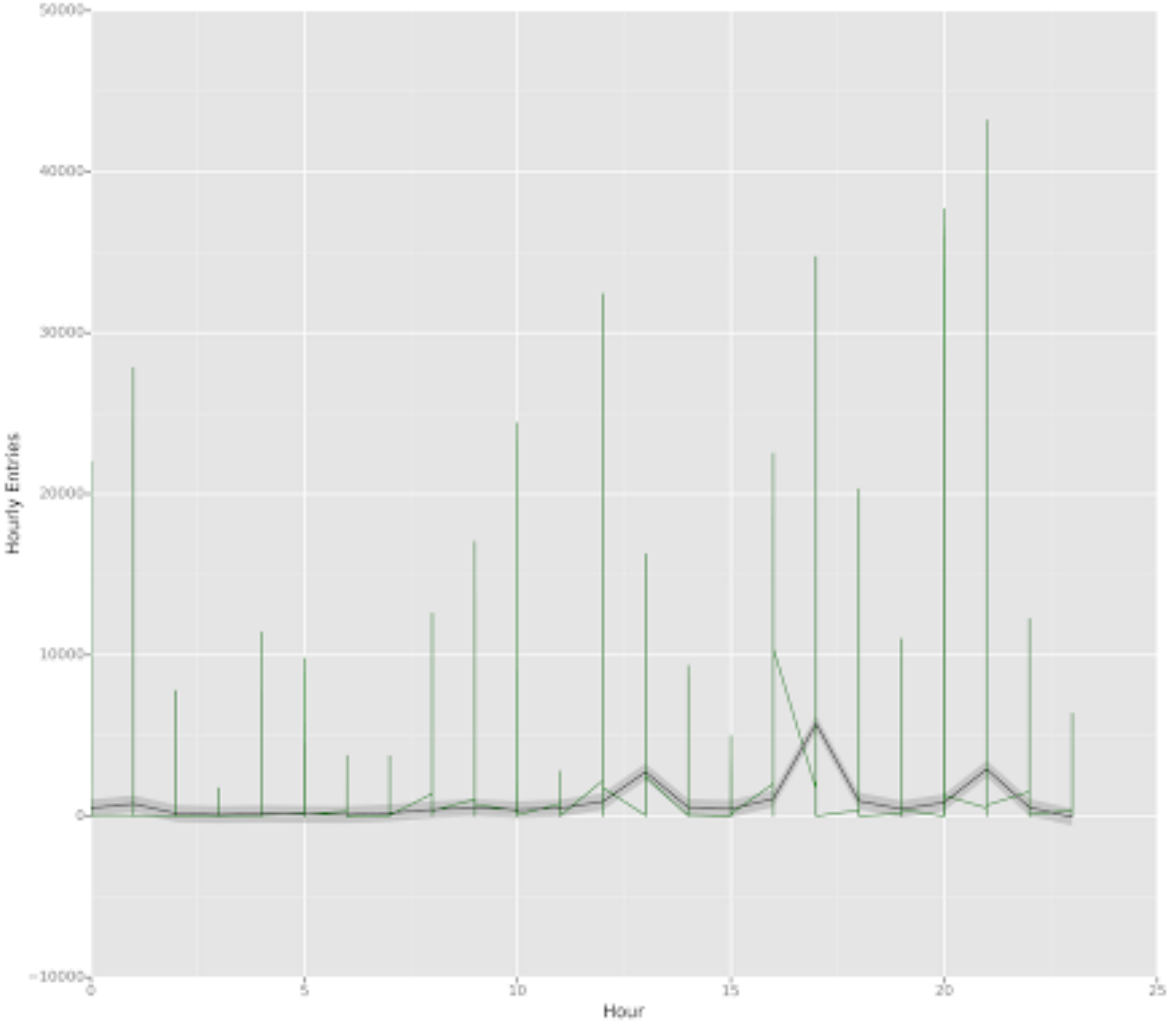
As we can see based on above visualization , the busiest days are Monday , while the least busy day is Sunday.

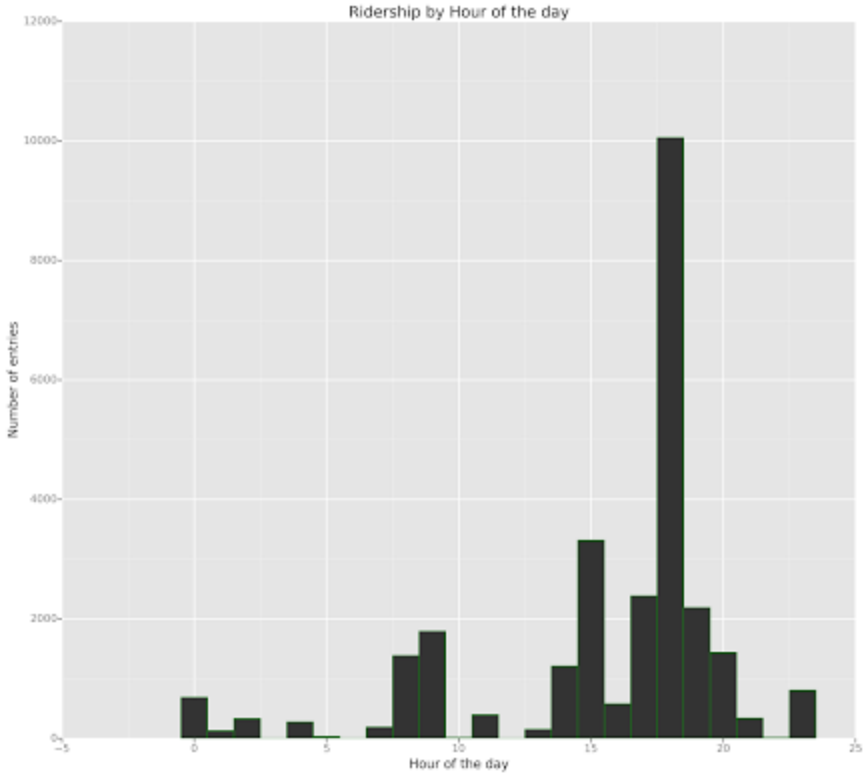
Ridership by Day of the week - with filling variable ‘rain’



Here we see that though more people take subway during rain' - Blue color, still on Monday we have people taking subway more (even when it was not raining).
While on weekend , we have less people taking subway and mostly when its raining.

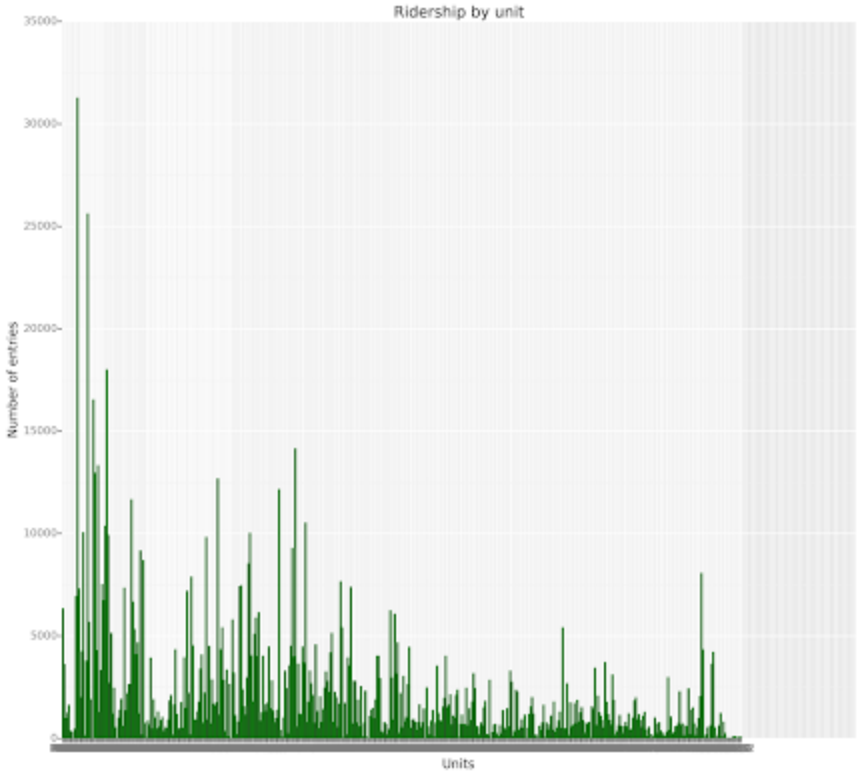
Ridership by Hour of the Day - Histogram and Lineplot


50000
40000
30000
20000
10000
0
−10000
Hourly Entries
0
5
10
15
20
25
Hour



Based on ridership plot above , we can see that more people use subway in the evening hours in between 5:30 - 6:30 pm time frame. While the subway ridership is minimum during certain time intervals (late night , early morning ,noon)

Ridership by Units (stations)



Above we can see that certain Units (stations) have higher ridership than other units.

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 the subway when its raining than when its not.

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 -

Here we have used the standard set of data.
Based on the statistical analysis(Mann-Whitney U test) of hourly entries (ridership) for the hourly interval when its raining versus when its not raining we can see that the Mann Whitney test provides us with a p value (standard data set) of 0.0249 which is less than the p critical value. Because of this we reject the null hypothesis that there is no statistical difference between ridership when its raining versus not raining. Also we can see that the average ridership (U mean value) is higher(1105.446) when its raining versus when its not raining. (1090.27).

Based on histogram of the data of ridership when its raining versus when its not also shows that more people rides the subway when it rains than when it does not rain.

Based on regression analysis(using gradient descent method) with normalized features and using the features which impact the ridership the most - i.e. rain , fog , time of the day , unit , minimum temperature , mean wind speed , we can see that our coefficient of determination is quite good (almost 0.5) which shows that the above factors which includes rain can be used to predict fairly the ridership in the NYC subway.

The above analysis shows that NYC subway ridership is higher when its raining Vs when its not raining.

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 -

<u>The shortcoming of dataset (using original dataset , as that was used in the analysis)</u>,

- The dataset used contain information of data in May 2011 which is one of the heaviest rainy season in NYC.
- Also there are some units which have very heavy ridership in NYC than the others.

The combination of above 2 facts makes the dataset skewed.

A better dataset would be if we can have a dataset where we can measure the effect of rain or weather against the particular station ridership rather than combining the results of less busy and more busy stations.
Also the dataset would have been better if we checked the data for a month also when it does not rain that much.

The shortcoming of analysis method (using original dataset)-

The statistical analysis method
The linear regression method used by me (gradient descent) provides a fairly good model but i am unable to improve the R2 value more than 0.46.

Through using OLS i find the R2 value increased to 0.485

As per graphical observation, i can see that day of the week also impact the ridership , but i was not able to include that in the features list . It would have improved the prediction model.

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

During the OLS model analysis of subway data , i find the warnings related to multicollinearity issues , which are probably caused by dummy variables Unit and Hour.