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.

1. http://pandas.pydata.org/pandas-docs/stable/generated/pandas.get_dummies.html - to understand how to create dummy data
2. http://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.mannwhitneyu.html - for the MannWhitneyU test syntax
3. https://storage.googleapis.com/supplemental_media/udacityu/649959144/MannWhitneyUTest.pdf?GoogleAccessId=1069728276824-fdhtlb98k1m9qrmdgj4jgc7gjp2l1lsm@developer.gserviceaccount.com&Expires=1432263445&Signature=lR2iiPm347gZ080ezCRaSracD/Wi2PhKd2uXzIORN2TeimihvXRdBTBFLBqXqCmDJdy7aqeRtJActr7fLkGHTjukG/NcOuaE3XPlmZ58RAar06DsDB9OuL5rbQBQG%2BaPrsH9pivB6mMaDAlo%2Ba6m58MK3l/k/nkg5JOWbqLkM9U%3D – To understand the MannWhitney U test
4. http://blog.minitab.com/blog/adventures-in-statistics/how-high-should-r-squared-be-in-regression-analysis
5. http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-regression-analysis-results-p-values-and-coefficients
6. http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
7. http://blog.minitab.com/blog/adventures-in-statistics/why-you-need-to-check-your-residual-plots-for-regression-analysis
8. http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-a-regression-model-with-low-r-squared-and-low-p-values

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 to determine whether there would be significant difference in the subway ridership on the days it rained versus the days without rain. I used the two-tailed P value because we are only concerned with whether or not the two samples come from the same population. The null hypothesis is that the two samples (with and without rain) would be identical that is they come from the same population. That is:

H-nought: x = y (where x and y are the two samples)

H-Alternate: x < y or x > y

The P critical value was 0.05 (alpha level of 0.05).

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.

The Mann Whitney U test is applicable because we can't make any assumptions about the distribution of ridership in the two samples. Also the sample sizes could be significantly different.

1.2 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.

The mean of the two samples, the MannWhitneyU value and the P-value were as follows:

Mean of entries with rain:	1105.4463767458733,
Mean of entries without rain:	1090.278780151855,
MannWhitneyU:	1924409167.0,
P-value:	0.024999912793489721

1.4 What is the significance and interpretation of these results?

The P-value of 0.0499 (we multiple the P-value by 2 for a two-sided test) indicates that we should reject the null at a P-Critical value of 0.05. This indicates that there IS a statistically significant difference in the subway ridership when it is raining versus when it is not raining. It is worth noting that the P-value was very close to the threshold that I had set to determine whether or not the value is statistically significant.

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?

I used OLS using Statsmodels. This model gave me a better R^2 than the Linear Gradient Descent (which was .4662) for the same set of features.

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

In addition to the Dummy variables, I used ‘Hour’, ‘rain’, ‘day_week’, ‘fog’, ‘tempi’ and ‘wspdi’. I started with just the Dummy variables and kept adding features until there was hardly a noticeable change in the R^2 value and the adjusted R^2 value when more features were added.

The non-dummy features were normalized and an intercept was added before performing linear regression using OLS. I also had to delete one of dummy variables in order to remove the muticollinearity. I noticed that as the number of features was increased, the change to the adjusted R^2 was small. I settled on the final value of R^2 of 0.470 and an adjusted R^2 of 0.467 when adding additional features like tempi and wspdi did not change the R^2 and Adjusted R^2 values.

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.”

I tried various combinations of features (see above) to see how adding or removing them would change the R^2 and Adjusted R^2 values. I found that UNIT and Hour had the most impact but the other features also contributed slightly (about 5%) to the R^2 score. I also found that by normalizing the non-dummy features, the error regarding a potential multi-collinearity in the model went away. Also removing just a subset of the dummy variables from the features seemed to help improve the value of R^2.

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

The coefficient for only rain was as follows:

```
==============================================================================
            coef        std err         t           P>|t|        [95.0% Conf. Int.]
------------------------------------------------------------------------------
const     1845.5394     16.231        113.702       0.000        1813.726  1877.353
Rain       182.6566     34.238          5.335       0.000         115.549   249.765
==============================================================================
```

Note that there is no warning about potential multicollinearity in the above. Also, note that the P value indicates that the ‘rain’ feature (x1 above) is statistically significant. When we plot the fitted values against the observed values, however, we see that there is no linear relationship between the observed values and the fitted values.

The following coefficients were obtained for the rest of non-dummy variables:

```
==============================================================================
 Var        coef        std err        t              P>|t|          [95.0% Conf. Int.]
------------------------------------------------------------------------------
const      1886.5900    13.605         138.673        0.000          1859.925  1913.255
hour        877.6233    14.759         59.463         0.000          848.695   906.551
rain       92.2307      14.464         6.376          0.000          63.880   120.581
day_week  -307.3576     14.040         -21.892        0.000          -334.876  -279.840
fog         -68.2934     13.780        -4.956         0.000           -95.302   -41.285
tempi        35.8009    14.758          2.426         0.015          6.874    64.727
wspdi      -141.7691    14.652         -9.675         0.000          -170.488  -113.050
==============================================================================
```

The values for the coefficient with Dummy variables was as follows:

```
==============================================================================
 var         coef       std err        t              P>|t|          [95.0% Conf. Int.]
------------------------------------------------------------------------------
const      192.2517     166.461        1.155           0.248         -134.015   518.518
hour        844.3404    11.404         74.038          0.000         821.988   866.693
rain       60.9333      11.143         5.468          0.000          39.092    82.774
day_week  -285.6376     10.836         -26.359        0.000          -306.877  -264.398
fog         -61.4140    10.591         -5.799         0.000          -82.173   -40.655
tempi         24.4535   11.340          2.156         0.031            2.227    46.680
wspdi         27.0809   12.578         2.153          0.031          2.428    51.734
==============================================================================
```

The above shows that when dummy variables are introduced, the weights of the constant and other features / variables goes down.

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

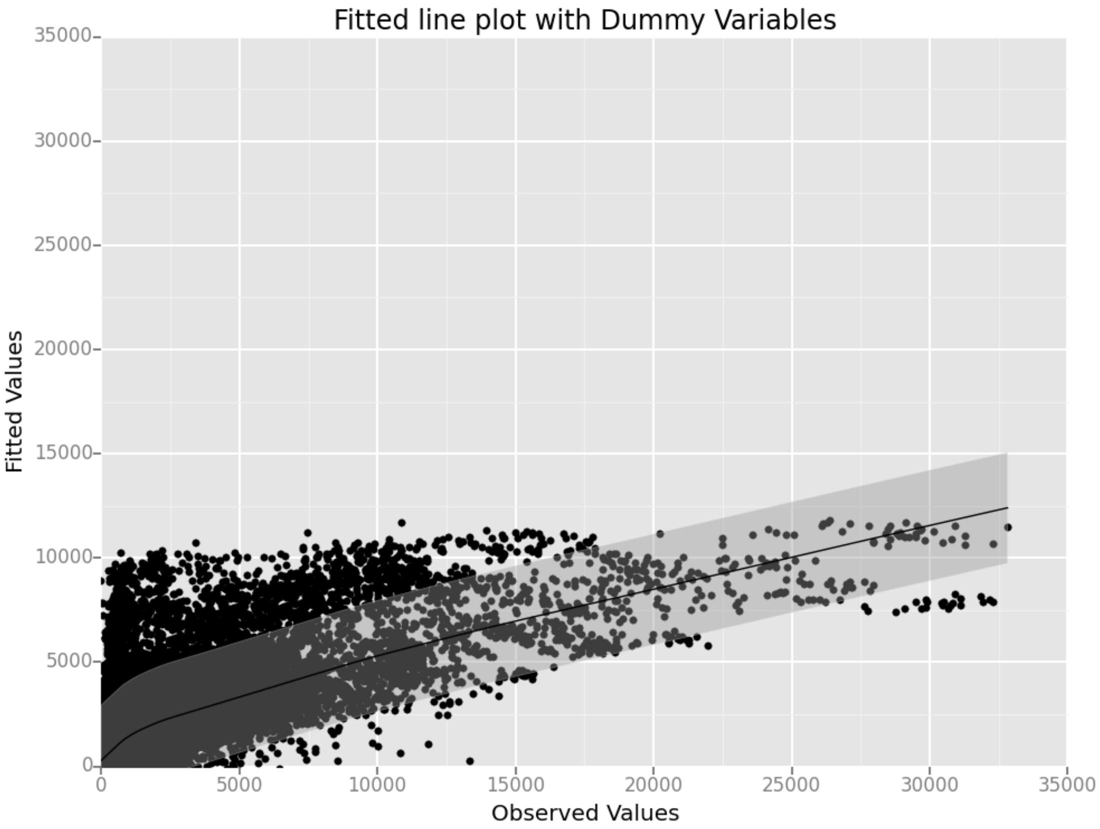
The highest R^2 score I got was .470. Here is the summary:

Features	Models	R-square	Adjusted R-square	multicollinearity
Dummy variables only	239	0.375	0.372	No
Dummy + rain	240	0.376	0.372	No
Dummy + rain+ hour	241	0.459	0.456	No
Dummy + rain+ hour+ day_week	242	0.469	0.466	No
Dummy + rain+ hour + day_week + fog	243	0.47	0.467	No
Dummy + rain+ day_week + hour +fog + tempi	244	0.47	0.467	No
Dummy + rain+ day_week + hour +fog + tempi + wspdi	244	0.47	0.467	No

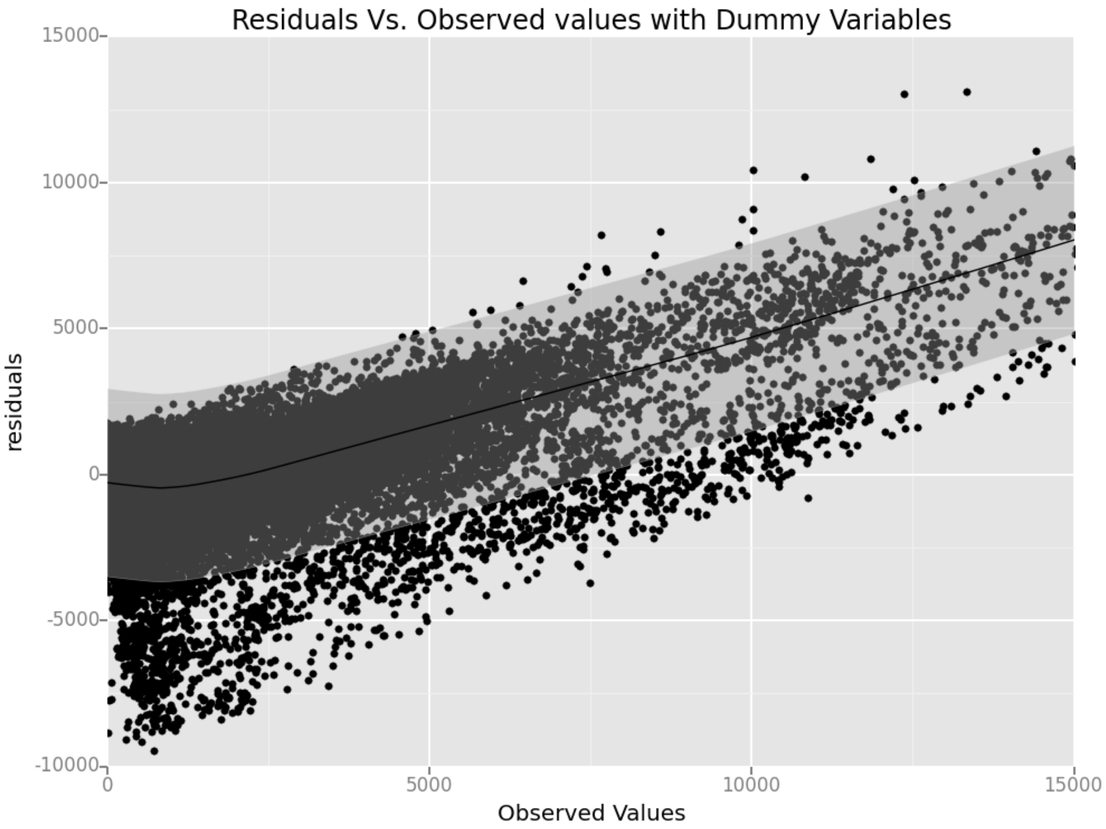
The final value of R^2 that I landed on was .470.

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?

The R^2 score of .470 indicates that 47% of the variability in the data can be attributed to the features chosen. Lets look at the Residual plot and the Fitted line plot for this linear regression:



The graph above shows that the while there seems to be a linear relationship between the observed and fitted values, the fitted values seem to be falling far short of the observed values. So the model is inadequate and is missing some predictors that might be causing the variability missing in the fitted values.



The Residual plot above reflects the analysis of the fitted line plot above. For the lower end of the spectrum of observed values, the residuals are systematically low while for the higher values the residuals are systematically high.

This indicates that the predictors in my model are not adequate to predict the values accurately.

Section 3. Visualization

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

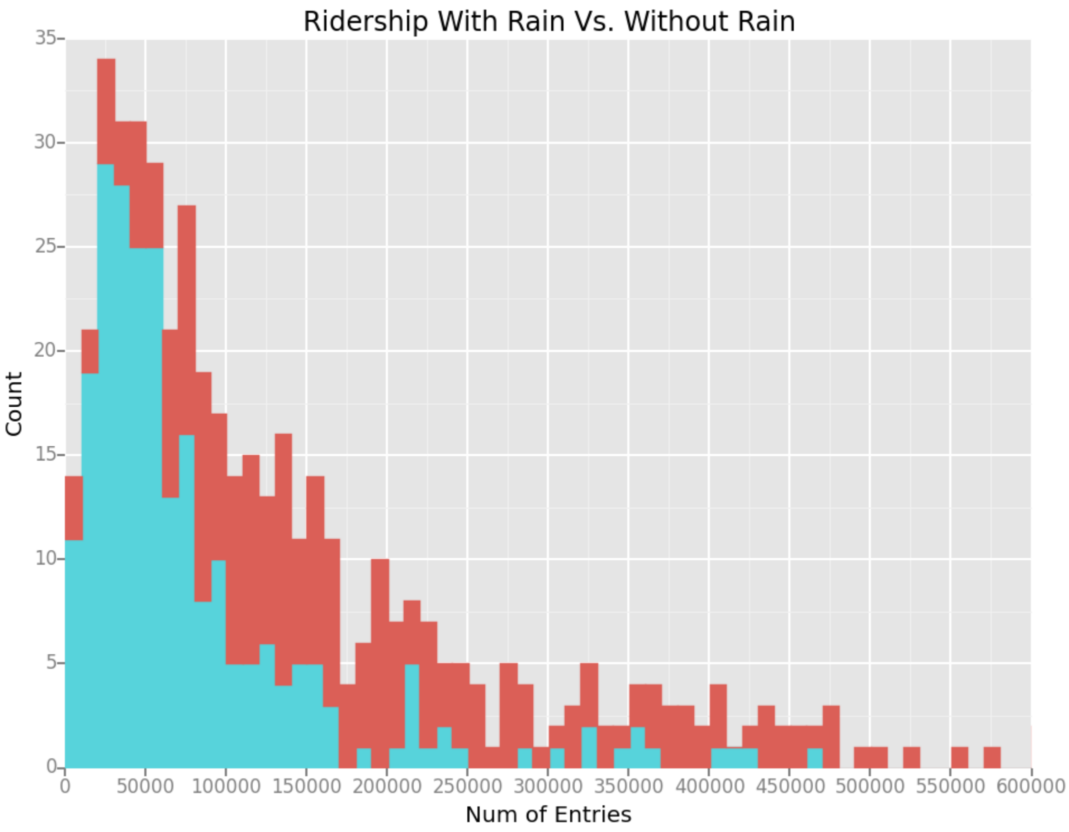
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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.

The histogram produced was as follows:



- Blue represent Number of Entries when it was raining
- Red represents Number of Entries without rain.

The Number of Entries for each Unit were grouped together and summed for with and without rain (See the code below), so the numbers on the X-axis represent the sum of entries each station for the month of May, while the numbers on the Y axis represent the frequency of occurrence of those summed values.

We see from the histogram that there **is** a noticeable difference in the ridership when it rains vs. when it does not rain across all stations. The graph shows that the **ridership decreases when it is raining**. It is worth noting that the dataset had more rows for when it was not raining than for when it was raining (the distribution was almost 2:1), which would explain why the red bars are wider. Also, the aggregation of the number of entries by stations, caused the total entries without rain to be much larger for some stations than the number of entries with rain. I limited the x-axis to 6000000 in order to remove the outliers for the entries without rain.

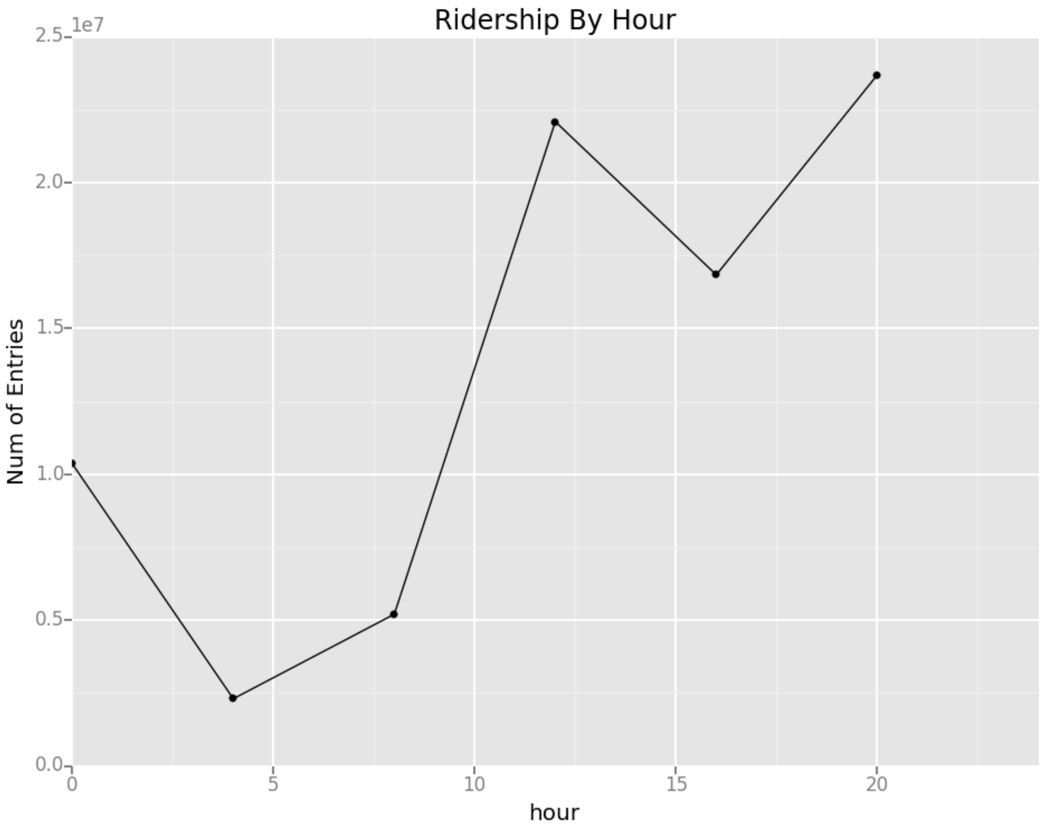
```
    entries_df = turnstile_weather[['UNIT','ENTRIESn_hourly', 'rain']]
    entries_df_summed = entries_df.groupby(['UNIT', 'rain']).sum()

    entries_df_summed.reset_index(inplace=True)
#    print entries_df_summed.max()
    plot = ggplot(entries_df_summed, aes(x='ENTRIESn_hourly', fill='rain',
color='rain') ) +\
                 geom_histogram(binwidth=10000, position='identity' ) +\
                 xlab('Num of Entries') + ylab('Count') + xlim(0, 600000) +\
                 ggtitle('Ridership With Rain Vs. Without Rain')
```

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

I plotted the graph of ridership by time of day. The result was as follows:



The plot above shows that when mapped across all stations, the ridership peaks around 1 PM and 8 between Noon and 8 PM. Please note the above graph was produced using the improved data set on my local machine.

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?

The histogram clearly shows that there are more riders when it is not raining than when it is raining when the data is aggregated over all the stations. This appears to be in conflict with the mean values of entries when it rains vs. when it does not rain in the sample provided, as was seen in Problem Set 3.3 (mean of entries with rain = 1105.45 and mean of entries without rain = 1090.28), but that might be due to outliers in the sample data set. The data in the histogram was aggregated over all stations and mapped by frequency of occurrence of a given number of entries, thus removing any station specific variability.

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 MannWhiteyU statistic was used to determine that there *IS* a statistically significant difference in the ridership when it is raining or when it is not raining. Although the P-value was almost equal to the P-critical value that I used for my threshold, it still proved that the difference was significant.

When I used the OLS method for linear regression, I used rain as one of the features to predict the value. The coefficient for ‘rain’ was 60.9333 (with the dummy variables included) and the P value was 0.000 which indicates that the feature is statistically significant. However, when I ran the linear regression with only ‘rain’ as the feature, the R^2 was very low (0.001) which indicates that while the ridership might change when it rains or not, as a predictor, it is much less significant than some of the other 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.

The linear regression I ran was only on the variables listed in the dataset - I made no attempt to change / correlate any of the variables. For example, day_week was used as is. As was pointed out in one of the discussions in the forum by Charlie, the day_week column has cyclical values and by using it as it, I am trying to fit it into a linear model, which might cause problems. So even if is a strong predictor of ridership, the fitted values might be off because of the way I have used the variable. I would like to try to convert that variable to a dummy variable as discussed in the forum and see how that impacts the results.

Also, the dataset seems to be biased towards more data for the days without rains vs. for days on which it rains. This would skew the mean for the number of riders on the days without rain if taken across all days of the week and across all stations. Assuming ridership is higher on weekdays than on weekends, if we had more days with rain on weekdays than on weekends, then the ridership on days without rain might look lower on average than days with rain.

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

The weather related columns (rain, temperature, fog, etc.) are all highly correlated (the temperature might be lower on the days with rain or there will be fog on the days that the temperature is low). Using them together as predictors in a linear model is likely to cause problems with mutilcollinearity which makes the model unstable. The model might benefit from using fewer of the weather related variables but by adding other information like holidays or events in the city (like a ball game which would draw people out to use public transit rather than trying to find parking and paying for it) etc.