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.

Welch's T test:

https://statistics.laerd.com/statistical-guides/independent-t-test-statistical-guide.php

http://en.wikipedia.org/wiki/Welch%27s_t_test

Mann-Whitney U test:

https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php

http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test

http://graphpad.com/guides/prism/6/statistics/index.htm?one-tail_vs__two-tail_p_values.htm

Dummy variables:

https://www.moresteam.com/whitepapers/download/dummy-variables.pdf

http://en.wikipedia.org/wiki/Dummy_variable_%28statistics%29

Coefficient of determination (R^2):

http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm

Normal Probability Plot:

http://en.wikipedia.org/wiki/Normal_probability_plot

http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm

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?*

The statistical test used to analyse the NYC subway data was the Mann-Whitney U-test. This test was used, more specifically, to compare the ridership of the subway during rainy and non-rainy days. The null hypothesis is that the two samples are drawn from the same population, or, in other terms, that the two distributions of "ENTRIESn_hourly" for rainy and non-rainy days are the same. Since no hypothesis was made about which sample would have larger or smaller values, the Mann-Whitney U-test was applied as a *two-tailed* test and the p-critical value (p*) was set beforehand to 0.05. The two-tailed p-value obtained from the test was also 0.05, thus satisfying the "≤ p*" criterion of statistical significance[1]. The null hypothesis was therefore rejected, i.e. the ridership of the subway is statistically different between rainy and non-rainy days with 95% confidence.

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.*

The reason for using the Mann-Whitney U-test is that the distributions of the two samples are non-gaussian (which makes the Welch's T-test *not* suitable in this case). Also, the distributions satisfy the requirements of the Mann-Whitney U-test. In particular:

- The dependent variable ("ENTRIESn_hourly") is ordinal, in the sense that one can always say, of any two observations, which is the greater.
- The independent variable consists of two categorical, independent groups (rainy and non-rainy).
- The observations within each sample and between the samples are independent.
- The distributions of both samples are equal (i.e. they have the same shape) under the null hypothesis.

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.*

Rainy days' mean = 1105.446

Non-rainy days' mean = 1090.279

(two-tailed) p-value = 0.05

[1] The two-tailed p-value was obtained by doubling the one-tailed p-value returned by default by the function `scipy.stats.mannwhitneyu`.

1.4 *What is the significance and interpretation of these results?*

Because the two-tailed p-value returned by the Mann-Whitney U-test is smaller than the critical value 0.05, we can conclude with 95% confidence that in rainy days the subway ridership is higher (by 1.4%) than in 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?*

I used gradient descent as implemented in exercise 3.5. In particular the learning rate α was set to 0.1 and the number of iterations to 75. The different features were also scaled by subtracting the corresponding means and dividing by the corresponding standard deviations.

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

The features used in the regression were: “rain”, “fog”, “precipi”, “meantempi”, “meanwindspdi”, “UNIT”, “Hour” and “Day” (the feature “Day” was added to the data-frame by extracting the day-of-week from “DATEn”).

The feature “UNIT” is obviously categorical and, as such, must be represented by a set of dummy variables in the regression. Regarding the features “Hour” and “Day”, although they both are (or can be) expressed by numbers, they do not follow the rules of numerical variables. For example the expression 5pm + 2am does not have any meaning[2]. For this reason they were also introduced in the regression model as two sets of dummy variables (a way to understand this choice is to consider the time-of-day as a more granular description of the categories ‘morning / afternoon / evening / night’).

The other two categorical features “rain” and “fog” were used as such in the regression model (for boolean indicators the dummy variable coincides with the feature itself and there is no need to use the function `pandas.get_dummies()`).

The other three features used (“precipi”, “meantempi” and “meanwindspdi”) are non-categorical.

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 […].*

[2] It is possible to derive one time-of-day from another by adding or subtracting time intervals (e.g. 5pm + 2h = 7pm) but two time-of-day cannot be added to or subtracted from each other. The same reasoning apply for day-of-week.

Weather-related features

The five weather-related features used in the regression were chosen because they can be intuitively related to the subway ridership. For example:

- "rain" - When it is rainy it may be more convenient to wait for the train underground than for the bus under the rain.
- "precipi" - Same as for "rain". In this case, however, the feature provides us with a non-categorical and more granular description of how rainy the weather is.
- "fog" - When it is foggy it may be safer to use the subway than the car or the bus.
- "meanwindspdi" - In windy days one may want to use the subway to find some shelter while waiting for the train.
- "mintempi" - Similar reasoning can be applied when the temperature is too low.

As a general consideration the subway seems a better alternative than other public transportation means when one wants to find shelter from extreme weather.

Time- / location-related features

The features "UNIT", "Day" and "Hour" were included in the model (through dummy variables) so as to distinguish between the variability of the ridership due to the weather conditions and the variability due to the specific time and location in which the subway is accessed. Moreover including these three features results in a big-to-drastic increase in the coefficient of determination R^2 (with "UNIT" having the bigger effect, followed by "Hour" and then "Day").

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

The regression coefficients θ are shown in Table 1. Interestingly the coefficients of "rain" and "precipi" are negative. This *seems* to suggest that in rainy days the ridership decreases. This contradiction is only illusive and is explained in later sections.

Feature	$\boldsymbol{\theta}$
rain	-2.00e+01
precipi	-4.38e+01
fog	7.60e+01
meanwindspdi	2.62e+01
meantempi	-6.69e+01

Table 1 - Regression coefficients obtained for the non-dummy variables ("precipi", "meanwindspdi" and "meantempi") and the boolean indicators ("rain" and "precipi").

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

R^2 = 0.519

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 coefficient of determination R^2 represents the fraction of the variance that is explained by the fit. The closer R^2 to 1 the better the fit, as the model explains more and more of the data variability. In our case (R^2 = 0.519) about 52% of the variance in the data is accounted for by the regression model.

Whether this value is "high enough" to consider our regression model a good fit depends on the context in which the predictions are used. In general, the higher the concerns at stake the higher the value of R^2 that is considered acceptable. For safety or security concerns 0.519 is most likely not good enough. On the other hand, if the prediction are used only for semi-quantitative analysis (for example to identify some general trends in the subway ridership) the model can provide a sufficiently good fit.

To assess whether a regression model is adequate to fit a set of data it is good practice to also examine the residuals. A regression model is considered adequate if the residuals are normally distributed around zero, signifying that the model is not biased. Although the residuals do seem normally distributed around zero (Fig.1-a), the normal probability plot reveals that their extreme values substantially diverge from a gaussian distribution (Fig.1-b). In particular the normality of the residuals distribution is only satisfied for values that roughly fall between -2500 and 2500, where the probability plot can be approximated with a straight line. Even if the vast majority of the residuals (about 92%) is confined in that interval, the pronounced divergence of the extreme values from the straight line suggests that the distribution has very long tails, which is a reason to question the adequateness of our linear regression model.

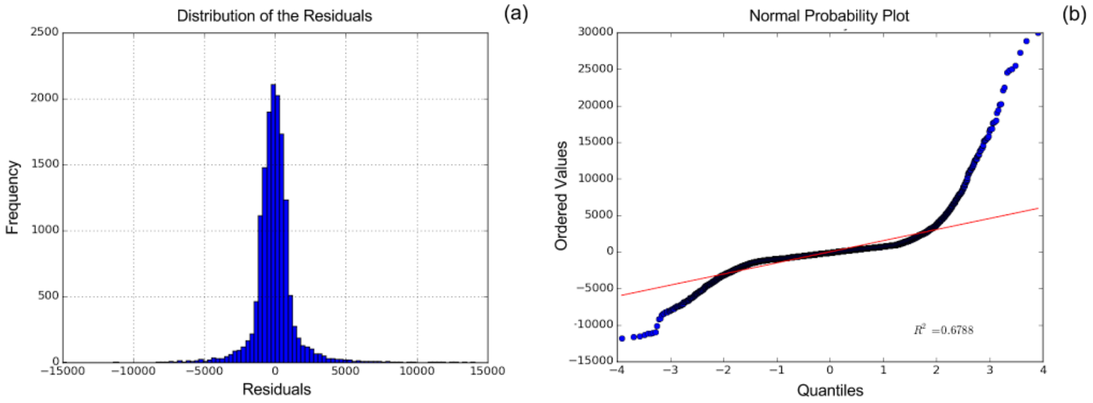

Figure 1 - **(a)** Distribution of the residuals for the linear regression model used to predict the subway ridership. **(b)** Normal probability plot of the residuals. The red line is the linear fit of the probability plot.

Another reason to be wary about a linear model is the presence of possible cyclic patterns in the residuals. Fig.2 shows the amplitude (and sign) of the residuals for the different datapoint ordered as they occur in the dataset (there are three nested sorting criteria; first "UNIT", then "DATEn"and last "TIMEn"). It is evident from the plot that there is a periodicity in the data, which makes the choice of a linear regression model less than ideal.

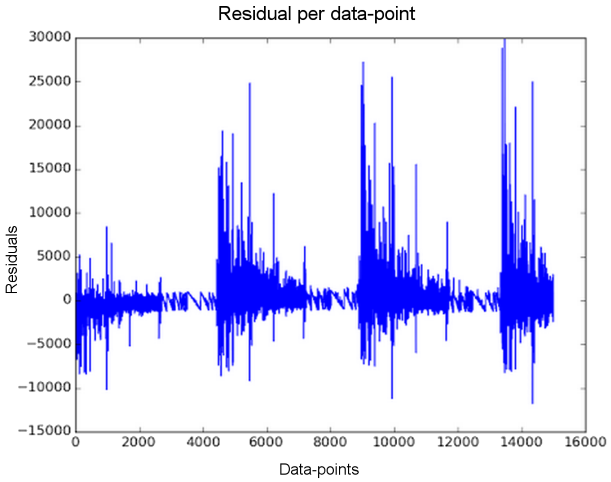

Figure 2 - Plot of residuals per data-point.

In conclusion the use of a linear regression model to predict the ridership of the NYC subway is not the best choice possible, and the design of a non-linear model seem at least advisable.

Section 3. Visualization

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

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. [...]*

Fig.3 shows the distributions of "ENTRIESn_hourly" for rainy and non-rainy days. The histograms are plotted in overlapping mode (not in stack). The size of the two samples substantially differ from each other. In particular, the difference in height between green and blue bars shows that the number of rainy days is roughly half the number of non-rainy days.

To visually assess whether the two distribution are different or not, the histograms of Fig.3 were normalized and plotted next to each other (see Fig.4). A by-eye comparison does not reveal any apparent difference between the two distributions. However the Mann-Whitney U-test showed that there is a statistically relevant difference between the two. In particular the average hourly-ridership increases by 1.4% in rainy days. Although this difference may not be practically relevant, the Mann-Whitney U-test tells us that it is *statistically* significant. This means that such difference is not ascribable to noise in our data but is associated with the weather condition with a confidence of 95%.

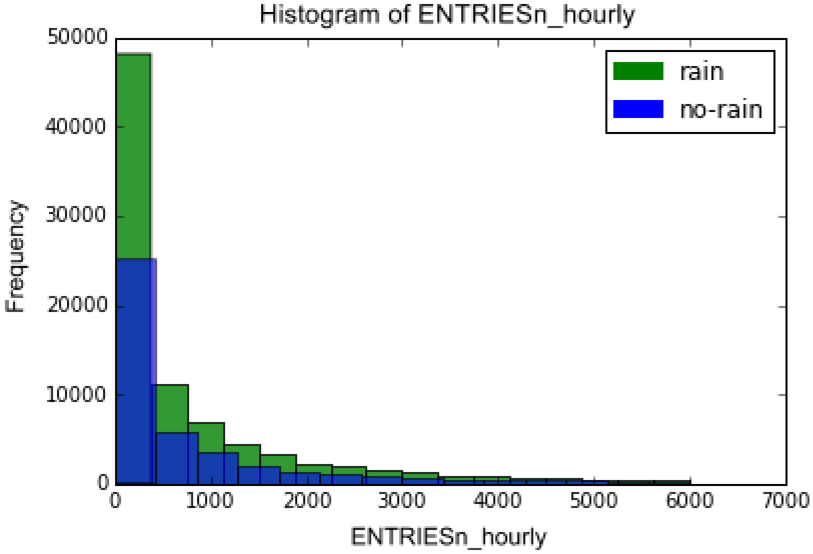

Figure 3 - Distributions of ENTRIESn_hourly for rainy and non-rainy days. The histograms are plotted in overlapping mode (not in stack).

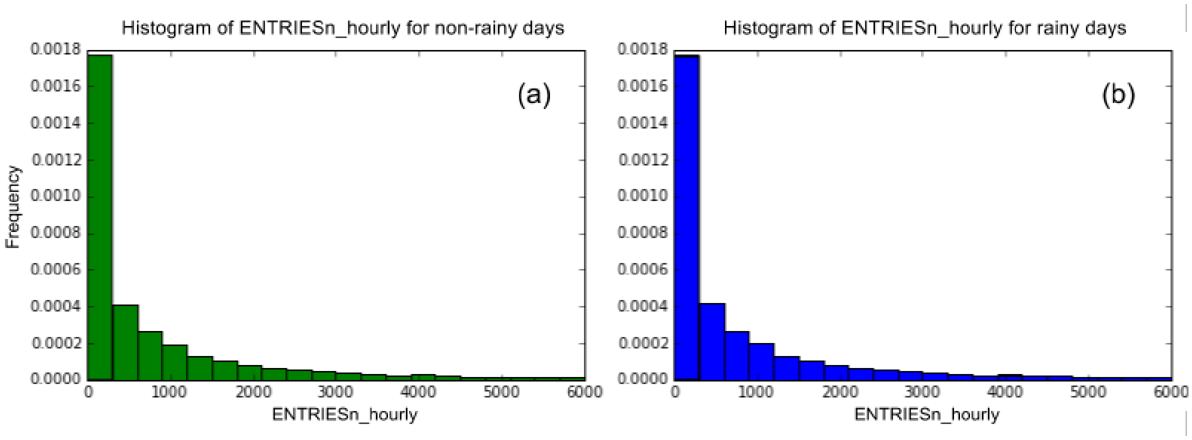

Figure 4 - Normalized distributions of ENTRIESn_hourly for non-rainy **(a)** and rainy days **(b)**.

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 the Ridership by time-of-day and the ridership by day-of-week.*

Fig.5 and Fig.6 show the ridership by time-of-day and the ridership by day-of-week respectively. There are five pronounced peaks in Fig.5 corresponding to the hours of higher ridership. The peaks at 8-9 am and 4-5 pm may reflect the typical commuters' trip to and from work. The other peaks may reflect the movements of people working at different shifts than the traditional 9-to-5 (see peaks at noon and 8 pm).

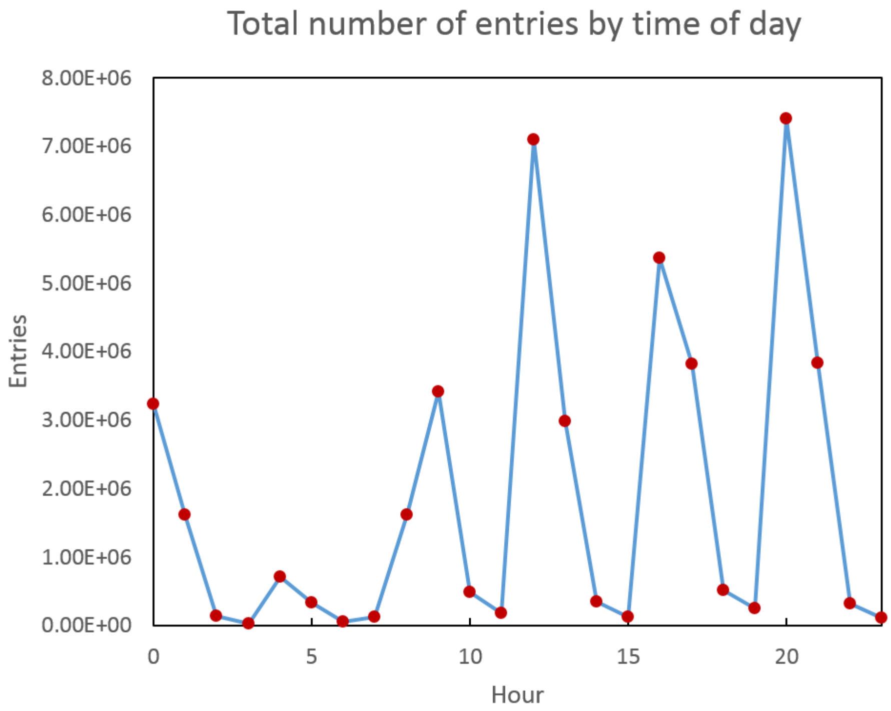

Figure 5 - Number of total entries by the time of the day.

As a general consideration, the number of people entering the subway must be equal to the number of people exiting it[3]. This means that the area subtended by one or more of the peaks must be roughly equal to the area subtended by the remaining peaks. This seems indeed to be the case. For example the areas subtended by the two peaks at noon and 8 pm is roughly the same, suggesting - but not proving! - that they mainly reflect the outward and back trips of the same commuters. Also the areas subtended by the peak at 4-5 pm is roughly the same as the sum of the areas subtended by the peaks at 9am and midnight. This also seems to suggest that those people taking the subway at 4-5 pm are the same who will take it again (or took it before) either at midnight or 8 am.

Fig.6 shows a very clear trend in the ridership of the subway along the week. In particular people resort to the subway substantially more during the working days than during the weekend. This suggests that the subway is mainly used to commute to/from work.

[3] This should be true with a very good approximation in the time-span of a month, assuming the dataset comprises all the NY subway stations.

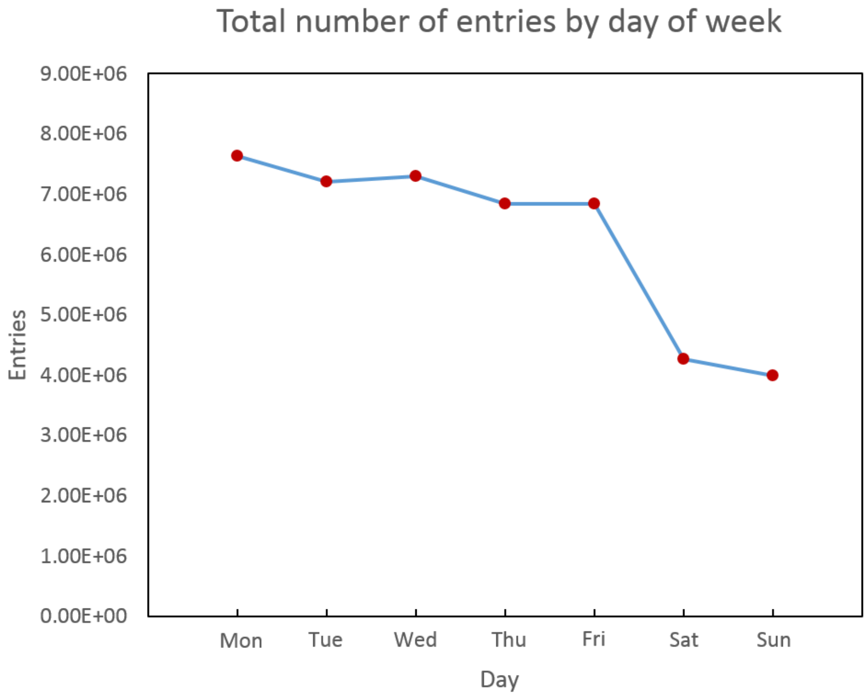

Figure 6 - Number of total entries by the day of the week.

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?*

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 result of the Mann-Whitney U-test shows that there is a small but statistically significant difference in the subway ridership between rainy and non-rainy days. More specifically in rainy days the average number of entries per hour increases by 1.4%. Although this difference may not be practically relevant, it is statistically significant with a confidence of 95%. On the other hand the regression value for the features "rain" and "precipi" is negative (see Table 1), equivocally suggesting that the ridership decreases when it rains.

The reason for this apparent contradiction lies in the fact that the regression coefficient is an estimate of how much the dependent variable changes per unit-change of the corresponding feature **provided that all the other features remain constant**. In other words the regression model tells us that when all the other factors remain unchanged (i.e. for the same day-of-week, time-of-day, unit, mean temperature, wind speed and so forth) the readership decreases when it rains. In reality, however, rain brings changes in other weather factors as well. For example, in rainy days the mean temperature could be lower and/or there may be higher chances of fog. Consequently the increase of 1.4% in the ridership in rainy days, as revealed by the Mann-Whitney U-test, is probably not ascribable to the rain *per se*, but rather to the *ensemble of other concomitant factors that tend to occur in rainy days*.

On this note it is worth mentioning that both distributions of 'ENTRIESn_hourly' for rainy and non-rainy days shown in Fig.3-4 include the variability carried by all the features other than "rain". In line with this reasoning, when performing the linear regression with less features (for example only with "rain" and "unit") the regression coefficient for "rain" becomes positive. This happens because the contribution of the other features to the ridership variability becomes accounted for by "rain".

All that said, it is important to keep in mind that the residual analysis previously discussed suggests that a linear regression model is not the best tool to make prediction about the NYC subway ridership.

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: Dataset, Analysis, such as the linear regression model or statistical test.*

Possible shortcomings in the dataset.

The dataset covers a time-span of a month (may 2011), but the impact of the features on the ridership may be different in different times of the year. For example the temperature may play a bigger role in winter than it does in spring, and its effect could even reverse in summer (in hot summer-days people may want to use the subway because the air in the tunnels is cooler, and in winter for the opposite reason). A dataset covering one year of subway ridership would be more informative about trends that are not specific of any specific month or season. Month/season-specific trends could still be extracted though.

Possible shortcomings in the regression model.

As already discussed in previous sections, a linear regression model is not the best choice to fit the NYC subway dataset and make prediction about the ridership. The cyclic pattern observed in the residuals reveals the presence of non-linear dependencies between the ridership and (some of) the features. The sorting criteria of the dataset entries - ordered by "UNIT", "DATAn" and "TIMEn" - suggest that the cyclic pattern is due to spatio-temporal factors. This could also be inferred by the plots in Fig.5-6, where the ridership is obviously dependent on the time-related features. It could be argued that this issue may be *partly* addressed by including the sorting features in the regression. By doing so one would separate - although only to a certain extent - the variability of the ridership due to the non-linear dependency upon such features from the variability due to the dependency upon the other features. Indeed, by including dummy variables for "UNIT", "Day" (proxy for "DATEn") and "Hour" (proxy for "TIMEn") the coefficient of determination R^2 increased drastically. This, however, does not solve the issue completely as the regression coefficients of *all* the features are nevertheless affected by the cyclic patterns contained in the dataset.

Even in the absence of this cyclic pattern, non-linearity issues may emerge from other inherent characteristics of the phenomenon under study. The quasi-linear dependency that the ridership may have on some variables could “break” for extreme values of the features. For example, the subway has a finite capacity and cannot serve more than a given number of customers, no matter the weather conditions.

Another possible shortcoming is that linear regression models do not reveal the relationship that different features may have with each other. In the context of this project, it is not easy for example to find out what are the features that, together with ‘rain’, make the ridership increase in rainy days.

Finally linear regression models are sensitive to outliers. For example, the subway readership could drastically increase during a one- or two-day strike of the bus drivers. This would be accounted for in the model even if there is no relation between the peak in the subway ridership and any of the other factors considered in the dataset.

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

All my insights were described in my answers to the previous questions.