Section 0. References

List of references -

- http://www.cliffsnotes.com/math/statistics/principles-of-testing/one-and-twotailed-tests
- http://www.statisticshowto.com/what-does-it-mean-to-reject-the-null-hypothesis/
- http://fsweb.bainbridge.edu/dbyrd/statistics/hypothesistesting.htm
- http://www.graphpad.com/guides/prism/6/statistics/index.htm?stat_checklist_mannwhitney.htm
- http://matplotlib.org/api/pyplot_api.html#matplotlib.pyplot.hist
- http://changingminds.org/explanations/research/analysis/parametric_non-parametric.htm
- https://en.wikipedia.org/wiki/Dummy_variable_(statistics)
- Webcast on Multicollinearity
- Udacity Discussion forum
- http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-regression-analysis-results-p-values-and-coefficients
- http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
- Other minitab blog articles
- http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm
- http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/
- http://stats.stackexchange.com/questions/78644/significance-of-dummy-variables-in-regression
- https://github.com/yhat/ggplot
- http://docs.ggplot2.org/current/geom_histogram.html
- http://www.bertplot.com/visualization/?p=229
- http://www.clockbackward.com/2009/06/18/ordinary-least-squares-linear-regression-flaws-problems-and-pitfalls/
- http://www.ehow.com/info_8562780_disadvantages-linear-regression.html
- http://people.duke.edu/~rnau/testing.htm
- http://blog.cloudlychen.net/variance-regression-clustering-residual-and-variance/
- https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php

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 Mann-Whitney U test to analyze the ridership of NYC subway. I am trying to predict that -

- Ridership of NYC subway is dependent on rain.
- Ridership of NYC subway will increase when it is raining.

Since I am predicting that there is a difference in ridership (depending on the rain) and the difference is in one particular direction (more ridership when raining), therefore I have used a one-tail p-value.

Null hypothesis:
Null hypothesis is a statement that we try to reject by running the statistical tests.

In this case, null hypothesis is ‘Number of people riding the NYC subway will either decrease or remain same when it is raining, as compared to when it is not raining.’

Lets say,

μ_1 = Number of people riding the NYC subway when it is raining
μ_2 = Number of people riding the NYC subway when it is not raining

Null Hypothesis H0 : $\mu_1 \le \mu_2$
Alternate Hypothesis Ha : $\mu_1 > \mu_2$

P-critical value:
P-critical value is used to support or reject the null hypothesis. If p value is below p-critical value then reject the null hypothesis, and vice versa.

p-critical = 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.

Mann-Whitney U test is applicable due to the following assumptions -

- Comparing two random independent samples - Ridership when it’s raining, and ridership when it is not raining.
- One sample has larger value than the other.
- Dependent variable is measured at a continuous level.
- In Problem Set 3 Exercise 1, I plotted the histogram to show hourly entries when raining vs. when not raining. Distributions for both the groups have similar shape, but are not normally distributed (skewed distributions).
- As inferred from above, analyzing a non-parametric data set.

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.

- Mean of entries with rain - 2028.1960354720918,
- Mean of entries without rain - 1845.5394386644084,
- U value - 153635120.5
- p value - 2.7410695712437496e-06

1.4 What is the significance and interpretation of these results?

The Mann Whitney test doesn’t measure the difference between the two groups, but evaluates whether the two groups are significantly different from each other.

Mean of entries for both groups:
So, we can measure the difference by comparing the mean entries for both groups -
Mean of entries with rain > Mean of entries without rain

U-value:
We can study the U statistic to test the null hypothesis -

No. of records for rainy days N1 = 9585
No. of records for non-rainy days N2 = 33064
Normal approximated U-value = 9585 * 33064 / 2 = 158,459,220
U statistic (as measured by Mann Whitney test) = 153,635,120.5

Since the measured U value is closed to the normal approximated U value, it provides a strong indication against the null hypothesis.

p-value:
Finally, we will study the significance level. Since the p value is very low and below p-critical value (p-value < 0.05), therefore we can reject the null hypothesis.

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?

I used OLS using Statsmodels for implementing linear regression.

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

Features (input variables) used in the model -

- rain
- hour
- weekday
- meantempi

Dummy variables used in the model -

- UNIT
- Conds

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.

Features (input variables) used in the model -

- rain - I used rain because I thought that if it is raining then more people would prefer to take the subway, instead of road transportation.
- hour - I used hour because I assumed that during some rush hours (like office hours in morning and evening) subways are more crowded.

- weekday - I used weekday because subways are often more crowded on working days as compared to weekends. I believe that many people use subway as the mode of transportation to reach office.
- meantempi - I used this variable because it improved the R^2 value a little bit, and there was no negative affect on condition number or p-values for other features.

Dummy variables used in the model -

- UNIT - I used UNIT because I thought that UNIT could be a good categorical parameter that can help in predicting the entry counts based on the historical data of remote units.
- conds - I used conds because I thought that weather conditions could influence people's decision to ride on the subway.

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

Coefficients of non-dummy features are -

hour	855.712308
weekday	424.876608
rain	37.347645
meantempi	-143.167552

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

R^2 (coefficients of determination) = 0.486361261136

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?

R-squared:
After watching lecture videos and reading articles on R-squared, I understand that R-squared is used to measure how closely the real data fits to the regression line, and it ranges from 0 to 1. The closer it is towards 1, the better the regression model. However, in few areas of study where you try to predict the human behavior, the prediction becomes tough and R-squared is close to 0.5. But, even a low r-squared model can provide useful predictions if we use statistically significant features to draw conclusions.

In my regression model, I am trying to predict the ridership of NYC subway based on the external events (such as weather, day, time, location, etc.), therefore I am expecting a comparatively low R-squared value as it involves predicting human reaction to these events. My regression model gives an R-squared value of 0.486, which means that it explains 48% of the variation.

However, along with R-squared value, I need to take into account other statistics measures (such as p-value, confidence interval, f-test, condition number, residual plot) to assess the goodness-of-fit of my regression model.

Condition number: 22.2
Condition number is not perfect (as in 1 or close to 1), but it is acceptable (less than 30).

I removed the multicollinearity to get an acceptable condition number.

P-value, Confidence Interval, and F-test:

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
const	1886.5900	10.276	183.583	0.000	1866.448	1906.732
hour	855.7123	10.402	82.268	0.000	835.325	876.100
weekday	424.8766	10.478	40.548	0.000	404.339	445.414
rain	37.3476	13.418	2.783	0.005	11.047	63.648
meantempi	-143.1676	11.929	-12.001	0.000	-166.549	-119.786

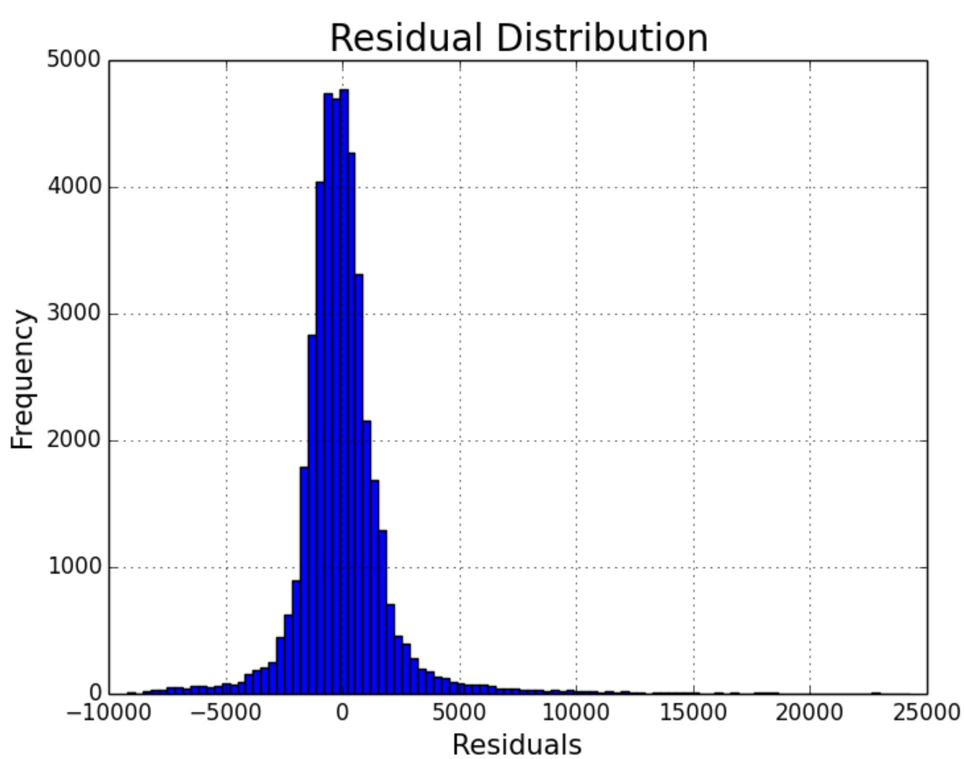
The low p values (< 0.05) and confidence intervals suggest that non-dummy variables are statistically significant i.e. if we change the predictor variable then response variable (ridership) will change.

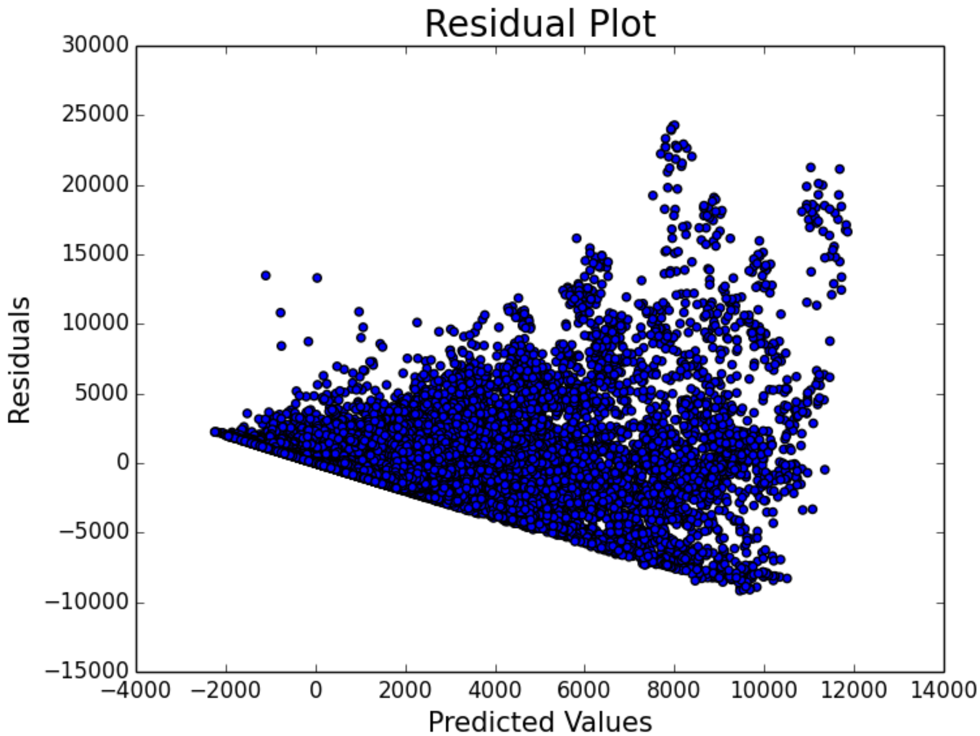
However, there are some dummy variables with high p values (> 0.05). Now, as I understand, there might exist few dummy variables that are not statistically significant and have no effect on the response variable.

For example - I am using conds as dummy variable. There are few conds (fog, mist, haze, partly clouded) that have high p-values. In this case, I will not use them to predict ridership, but I will still keep them in my model to perform analysis and draw conclusions.

Since I have few non-significant dummy variables, I also used the F-test to evaluate the overall significance of my regression model. P-value for the F-test is 0.00, and it indicates that I have a statistically significant regression model.

Residual plot:





From the above charts, we can see -

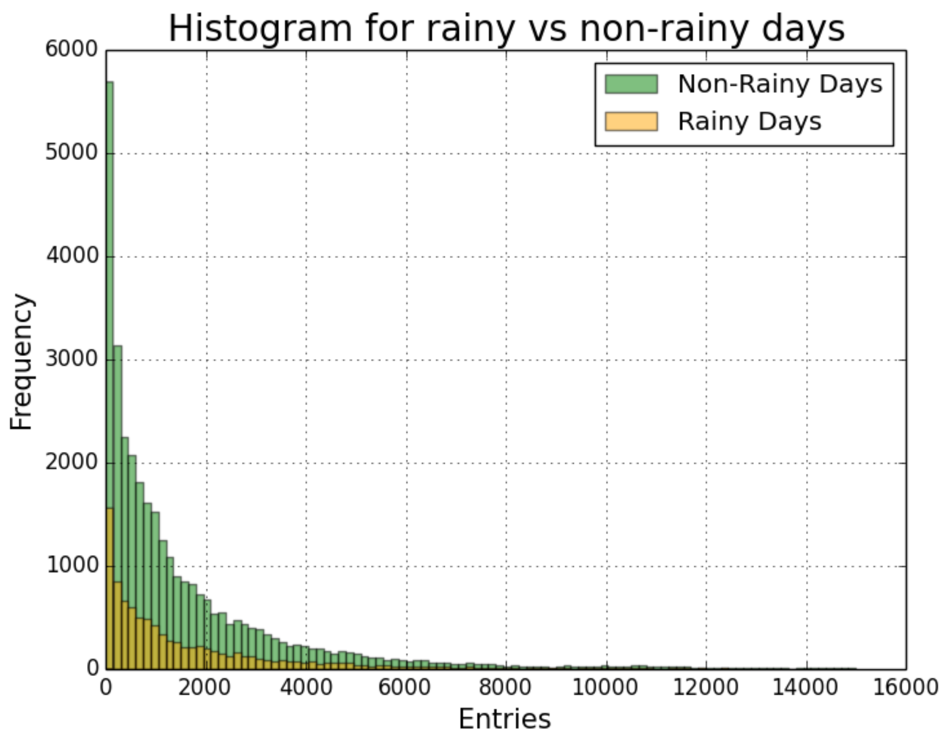
- Residuals are uniformly distributed.
- Residuals are scattered around 0, sometimes it has positive values and sometimes negative.
- I can see a pattern in my residual plot. As explained in few of the articles mentioned in Section 0, as the size of the predicted value increases, the residuals may become more scattered. Thus, giving a funnel type shape to the plotted residuals. This phenomenon is known as heteroscedasticity, and it makes our model less efficient. Heteroscedasticity is prominent in my residual plot. More about this in Section 5.

Conclusion:
After analyzing all the statistics measure, I conclude that even though my overall regression model looks statistically significant, but low R-squared value and residual plot indicates room for improvement. Therefore, this model is not yet a good fit to make accurate predictions.

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.



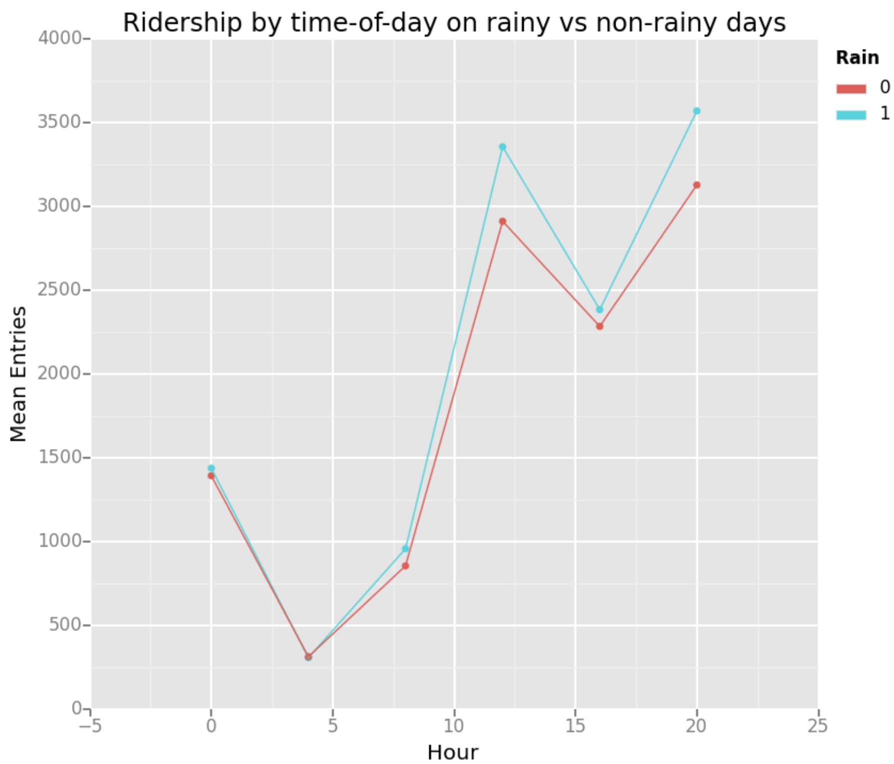
Key insights:

- This histogram depicts the distribution of ridership on rainy days and non-rainy days.
- Distribution is not normal, however it has a similar shape for both cases.
- Looking at the chart, it seems like the number of riders on non-rainy days are more than the rainy days, but we should consider the sample size for both groups before drawing any inferences.
    - No. of records for non-rainy days = 33064
    - No. of records for rainy days = 9585

  As we can see, there is a huge difference in the sample sizes, and therefore it will be wrong to draw any conclusions without performing any statistical tests. Therefore, we will only use this histogram to study the distribution of the data set.

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:

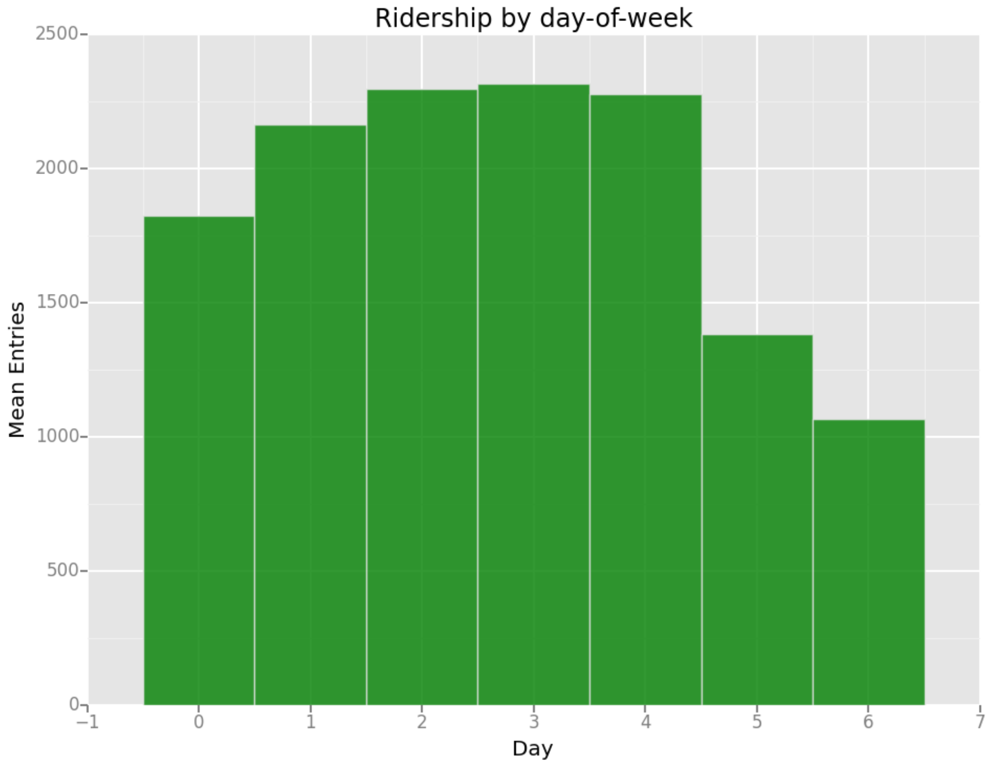
- Ridership by time-of-day on rainy vs non-rainy days



Key insights:

- This chart depicts how rain influences the ridership at different times of day (0 for 'No Rain' and 1 for 'Rain').
- Mean hourly entries at different times of day are significantly more when it is raining.
- More people ride the subway between 8am-12 pm and 4pm-8 pm, irrespective of the rain.

- Ridership by day-of-week



Key insights:

- This chart depicts the mean hourly entries on different days of week (0 for Monday and 6 for Sunday)
- More people ride the subway on weekdays when compared to weekend.
- Mean hourly entries for Monday is significantly less as compared to other weekdays.
- Mean hourly entries for Sunday is significantly less as compared to Saturday.

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?

Mann-Whitney test provides enough evidence to reject the null hypothesis, and mean of entries when it is raining is more than when it is not raining.

Therefore, we can infer from the test that the difference between the two groups is statistically significant, and more people ride the NYC subway when it is raining as compared to when it is not raining for the given dataset.

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.

Mann Whitney U test –

Responses of the two groups were compared using Mann-Whitney U test, and the difference between two groups were found to be significant ($U= 153635120$, $p < 0.05$ one tailed). The mean entries for rainy days and non-rainy days are 2028 and 1845 respectively. Mean entries increased by 185 for rainy days.

Linear Regression Model –

In the current regression model, the coefficient for rain is +37 (approx.) and $p < 0.05$, which makes rain a statistically significant feature in predicting the ridership. The +ve coefficient indicates that on rainy days the response variable (i.e. ridership) will increase. However, as concluded in Section 3, linear regression model is not a good fit and therefore we cannot use it to make these predictions.

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

Shortcomings of dataset:

- Dataset includes only one-month data, May 2011. If the prediction model is trained on only one-month data, it can make biased predictions. For example, number of riders might decrease in June/July as the schools are closed, but the predictions are based on May data that doesn't consider summer holidays. Therefore, we need more varied set of training data.
- Dataset doesn't take into account public holidays.
- Dataset doesn't consider external events that might reduce the ridership, other than weather conditions. For example, some subway station might not be fully functioning due to maintenance works.

2. Analysis, such as the linear regression model or statistical test.

Shortcomings of linear regression model:

- Residual plot depicts heteroscedasticity, and as I understand, we need to transform the predicted values such as using log, square root, etc. But, I am still new to statistics, and don't know much about the transformation process. If you have any good reading material, please share. Thanks!
- I was unable to find the linear relationship between hour and ENTRIESn_hourly, meantempi and ENTRIESn_hourly.
- Linear regression model might not be a good fit to predict human behavior.