Please look at revised sections as per the last evaluation

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 1. Statistical Test

1.1 Which statistical test did you use to analyse 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?

Ans:

We used Mann-Whitney U-Test. And we used the two-tail P value to check the following null hypothesis:

"*There is no statistical difference in distribution of hourly entries in NYC Subway between rainy and non-rainy days (p<=0.05)*"

The two tailed p-value we got for the Mann-Whitney U-Test was **0.498**. This exceed the null hypothesis declaration of P<=0.05, hence there is sufficient information to accept that there is no difference statistically in the hourly subway entries for rainy vs non-rainy days.

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.

Ans:

The main reason why this is applicable is because the data set is not having a normal distribution. While the t-test has the underlying assumptions of normality and equality of variance, the Mann-Whitney U-Test is its generalization, which has a underlying assumption that the samples under consideration are random and independent of each other, it also assumes that the sample is ordinal.

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.

Ans:

```
(1105.4463767458733, 1090.278780151855, 1924409167.0, 0.024999912793489721)
```

1 Mean of entries with rain: `1105.4463767458733`

2 Mean of entries without rain: `1090.278780151855`

3 Mann-Whitney U-statistic: `1924409167.0`

4 Mann-Whitney p-value: `0.024999912793489721`

1.4 What is the significance and interpretation of these results?

Ans:

Specifically the p value of 0.024 and in case of two-side as 0.048 is less than 0.05. Also, the mean between the two sets (rain vs without rain) are different. These indicate they are different statistically but the pattern of the distribution is not drastically different.

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?

Ans:

I used the Gradient Descent as well as OLS using Statsmodels

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

Ans:

The features used were: 'rain', 'precipi', 'Hour', 'meantempi', ‘meanwindspdi’ and calculated field of days of week. Yes, the Dummy variable was used to prefix the unit to the features.

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.”

Ans:

The features used by me were:

- 'rain', 'meantempi', 'meanwindspdi' : These features provide a weather perspective to ridership in the subway and its impact.
- 'Hour', ‘Daysofweek’ : The hour of the day and the day of the week impact from a overall peak ridership perspective.
- Adding the days of week improved my R square value.

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

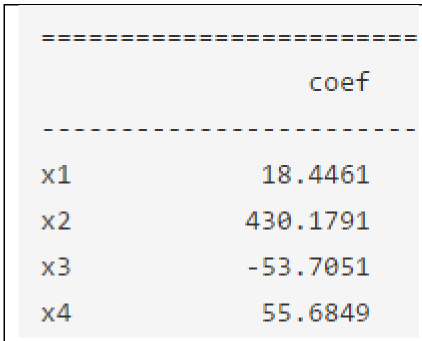
Ans:

In the problem 3.5 for the default liner regression with gradient decent the weight or alpha value was 0.1

In the OLS stats model I have implemented without alpha as with alpha on fit_regularized the execution as timing out.

In the OLS stats model the coefficients of the non-dummy features are:

Rain: x1, Hour:x2, meantempi:x3, meanwindspdi:x4



```
========================
                   coef
------------------------
x1              18.4461
x2             430.1791
x3             -53.7051
x4              55.6849
```

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

Ans:

In problem 3.5 I got a R^2 : 0.47009241754

In problem 3.8 I got a R^2 : 0.488989566113

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?

Ans:

The current R^2 value means that there is a 48.89% fit of the total variation in the data

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

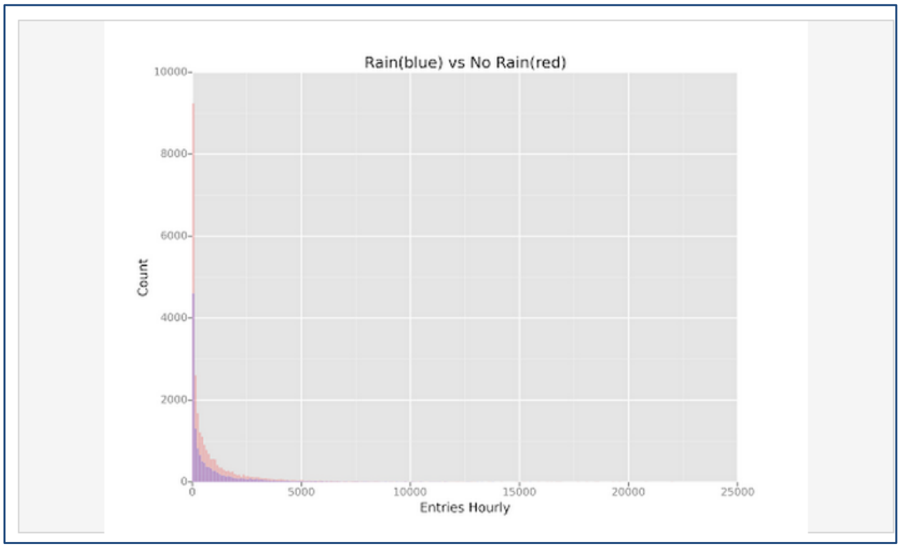
Remember to add appropriate titles and axes labels to your

plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

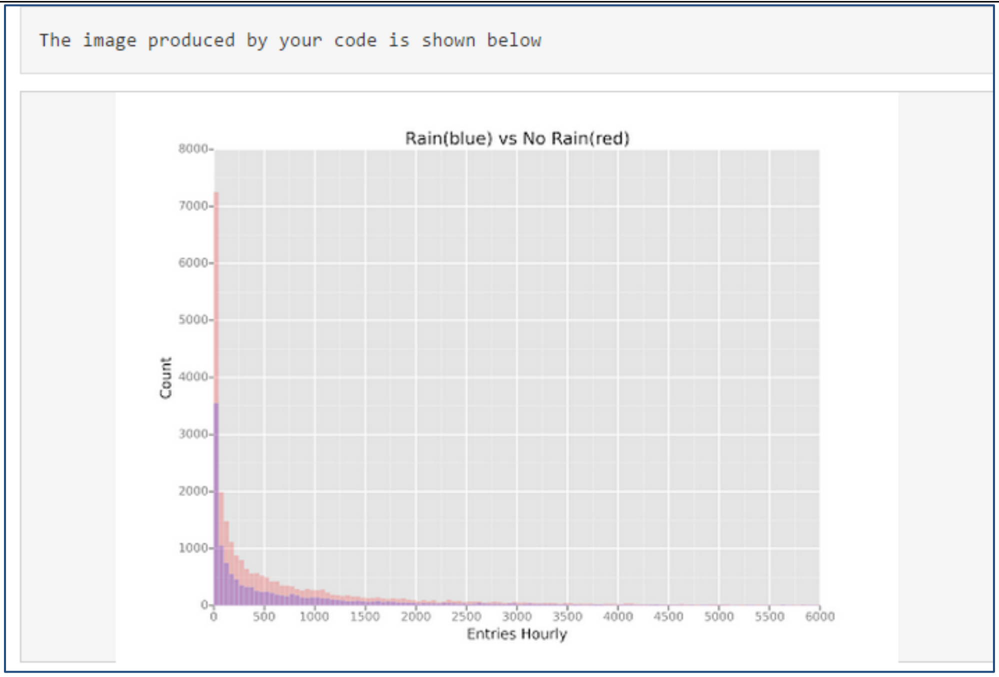
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

- 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.

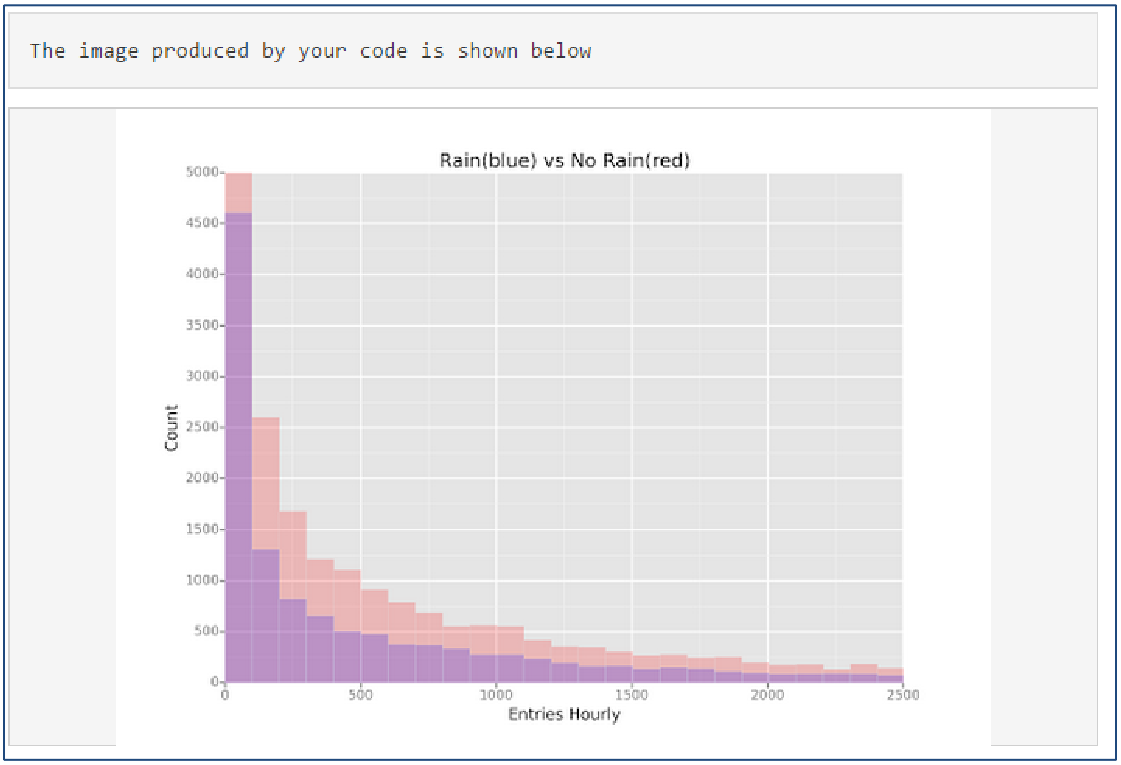


Xlim to 25000, Bin size 100



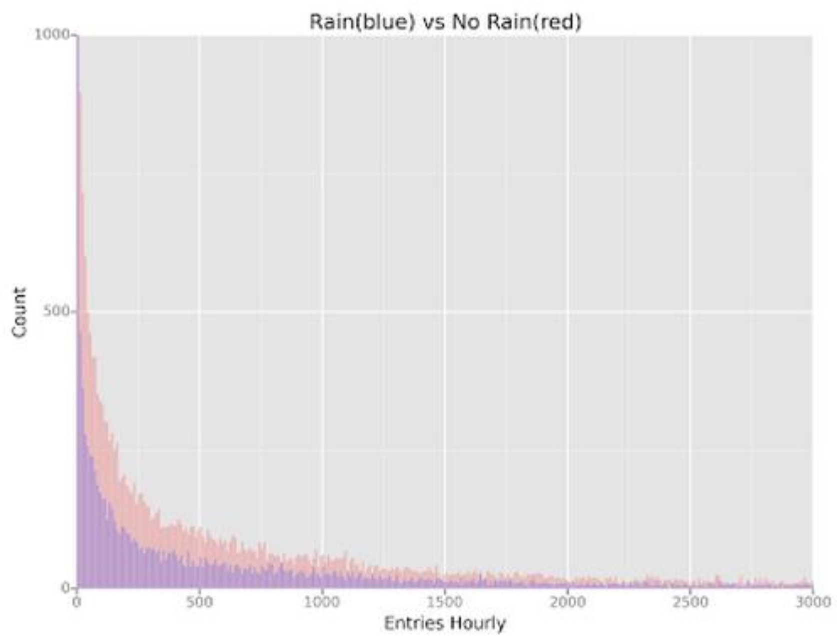
```
The image produced by your code is shown below
```

X limit to 6000, bin size 50 red is “No rain” and front purple is “Rain”



```
The image produced by your code is shown below
```

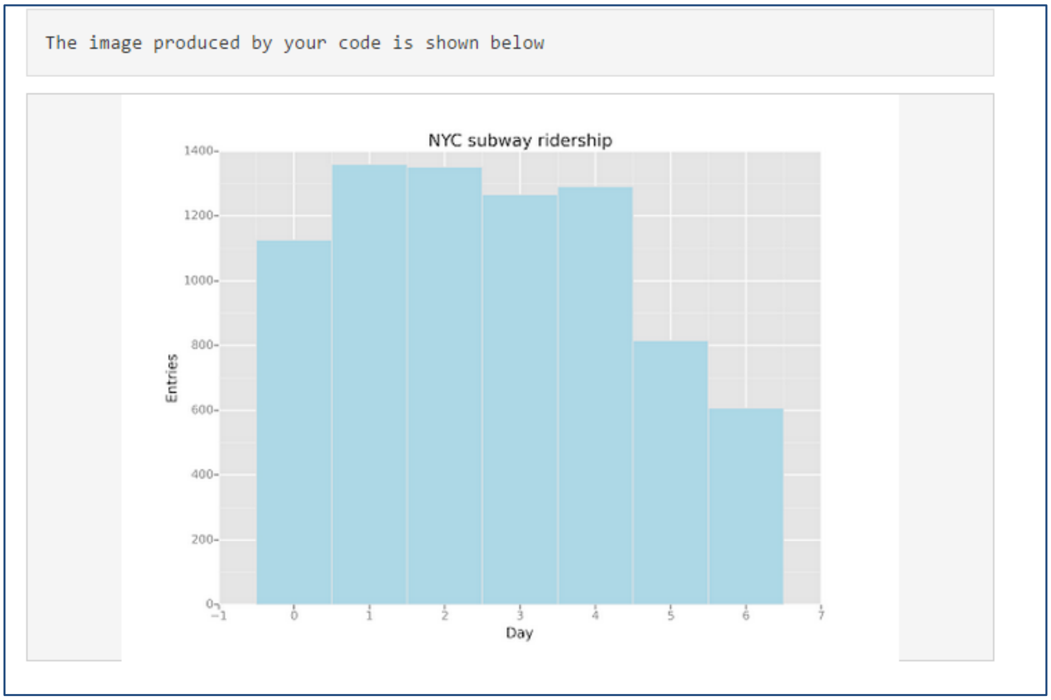
X limit 2500 and Y Limit 5000, bin size 50 red is “No rain” and front purple is “Rain”



Bin Size 10 and limited X and Y.

3.2 One visualization can be more freeform. Some suggestions are:

- Ridership by time-of-day or day-of-week
- Which stations have more exits or entries at different times of day



```
The image produced by your code is shown below
```

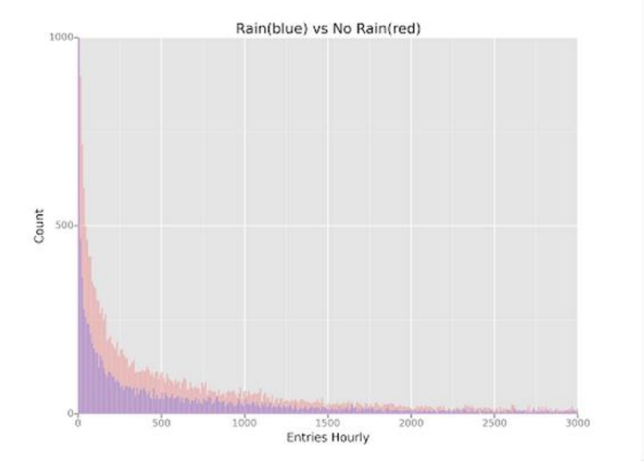
Mean Ridership by the day of the week.

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?

Based on the difference in the sample size for Rain vs No Rain the histograms it does not seems like the Rain has major impact on the ridership of the subway.



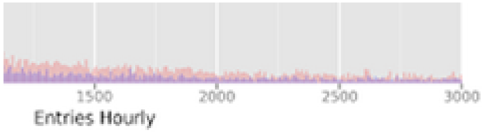
Also to note is that when the entries are really high, the impact of rain is not seen at all as shown below:

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 NYC subway has no impact of the rain on ridership, although there is a higher values on the x axis, but the sample sizes being different negates it. This is more of an indication that when people have to travel (like office hours) they will go in NYC subway irrespective of rain or no rain, but during other times it is not very conclusive.

This can be seen in the chart when there are high entries impact of rain is not there.



The use of rain as a feature in the value R2 has low impact. This is shown by:

Features Used	**R2 Value**
'rain', 'Hour', 'meantempi', 'meanwindspdi'	`r^2 value is 0.464421896347`
'Hour', 'meantempi', 'meanwindspdi'	`r^2 value is 0.464404278873`
'rain', 'Hour', 'meanwindspdi'	`r^2 value is 0.4641396228`
'rain', 'meanwindspdi'	`r^2 value is 0.425918664698`

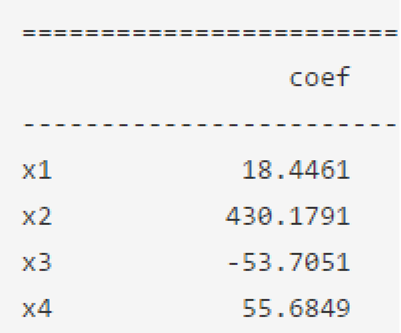
'Hour'	r^2 value is 0.463175379772
‘DayofWeek’ Calculated field dataframe['DayofWeek'] = pandas.to_datetime(dataframe['DATEn'].astype(str),format='%Y-%m-%d').dt.dayofweek	r^2 value is 0.43013769058
'DayofWeek', 'Hour'	r^2 value is 0.468440165223

If you see above table expect the hour of the day other aspects do not have a considerable impact on the R2 value.

Although this also makes us understand that only going with mean values may give us the wrong impression.

The coefficients of the non-dummy features also emphasizes on the impact of each feature shown below:

Rain: x1, Hour:x2, meantempi:x3, meanwindspdi:x4



```
========================
                  coef
------------------------
x1             18.4461
x2            430.1791
x3            -53.7051
x4             55.6849
```

Except the hour of the day other features do not have a strong impact.

Also, with only 48.89% fit, the linear regression model also says there no extremely strong factor that impact the increase in ridership during rain.

The other factor that has come out is that weekends have lower subway ridership than weekdays. And also certain hours of the day there is a peak in the sub way ridership, shown by higher R2 above compared to other features.

But factors like temperature do not have a very high impact on the ridership shown by lower R2 value in table above.

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. Linear regression model,
3. Statistical test.

In the dataset we had the relation sip between a lot of factors and the ridership was not exact linear although for a few it can be a fit like to know the impact of rain or the time of the day etc. But when multiple factors apply in a non-linear manner this model does not accurately predict. This was clear by the 48% fit we got from the Gradient Descent as well the OLS model. When there are multiple factors and nonlinear impact, which is the case in most of the practical scenarios, these models may not work well.

Although for a set of sample distributions Mann-Whitney U-test helps a lot in understanding the distribution when compared to the T-Test as this is nonparametric. If we take the example of rain vs no rain as well this dataset is a bit unique where overall with rain the ridership is quite high but there are some aspects like when the ridership is medium the rain has a impact.

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

Although I felt this was a very good exercise in touch the basics and multiple tips of the icebergs, it gives us an initial toolset for analysis of the structured data.

We would need to better understand how to discover unknown factors which could have impact on the analysis we are looking at, may be we have not seen it yet, excited to see what lies ahead with exploring data etc