Richard McMahon

AXA Life Invest Group

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.

References used

Problem set 2
Item 1
This was not difficult. I did look at the reference provided in https://dev.mysql.com/doc/refman/5.1/en/counting-rows.html

Item 2
I did not use any references here.

Item 3
Similarly here I used the detail provided in the exercise itself but no more than this

Item 4
As for items 1, 2, and 3

Item 5
This was a particularly tricky problem. I did find the detail on the Discussion Forum very helpful to complete this part of the problem. The link to this is copied here. http://discussions.udacity.com/t/problem-set-2-part-5-includes-answer/15134/4. I appreciate it includes a clear pointer to the result but the way it was laid out was very helpful for me.

Item 6

Problem set 3

Item3
http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html

Problem set 4

Item 1

For the first graphic/visualisation I produced I found the following link useful to add x and y labels to the plot that I produced.

http://stackoverflow.com/questions/21487329/add-x-and-y-labels-to-a-pandas-plot

I also found helpful the insight from the discussion forum

http://stackoverflow.com/questions/22635110/sorting-the-order-of-bars-in-pandas-matplotlib-bar-plots

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?

Response

The statistical test used to analyse the NYC subway data was the Mann Whitney U test.

I used a two-tail P value.

The null hypothesis is that 2 populations are the same. In this case the null hypothesis is that hourly NYC subway entries are the same on days that rain as on days that do not rain.

The p – critical value is 5%.

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.

Response

This test was applicable to the dataset as the ‘entries’ data does not appear to be normally distributed. This was clear from the initial histogram plotted showing hourly entries by days that rain and days that do not rain.

The null hypothesis considered by the test is that the ridership in the two populations/samples (days that rain and days that do not rain) is the same. The test does not assume our data is drawn from any particular or specific underling probability distribution.

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.

Response

The results obtained are the following

Average hourly entries on days that rain = 1105.4463767458733
Average hourly entries on days that do not rain = 1090.278780151855
Mann Whitney U test statistic = 1924409167.0
P value = 0.024999912793489721 (one sided)
P value = 0.4999982558697940 (two sided)

Note: The Mann Whitney U test returns a one sided p value and given we are considering a two sided test we need to double it.

1.4 What is the significance and interpretation of these results?

Response

The results conclude based on a p critical of 5% that the null hypothesis that the two populations are the same is not true i.e. the analysis rejects the null hypothesis.
The p value from the test (doubled) is less than 5% (but only just!) and hence the conclusion / interpretation of these results.

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?

Response

I implemented the requirements in exercise 3.5 – gradient descent.

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

Response

The features used were ‘rain’ and ‘hour’.

‘UNIT’ or station was used as a dummy variable.

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.”

Response

I used ‘rain’ in particular as a predictor as the Mann Whitney U test did indicate that whether it rains or not possibly results in different levels of subway ridership i.e. the null hypothesis that entries were the same on days it rains and days it does not rain was rejected using a p critical of 5%. ‘Rain’ based on this detail is clearly a determining factor in the level of ridership and hence makes sense to include in the regression analysis as a predictor variable.

‘Rain’ may not however be a determinant of the distribution of ridership. The histogram plotted would indicate clearly that the distribution of ridership on days it rains and on days it does not rain is non normal. Also the shape of the histogram including both ‘rain’ and ‘non rain’ NYC ridership together would indicate that the same fundamental distribution applies to both (with possibly different parameters attaching to this distribution).

‘Hour’ and the use of this as a feature or predictor variable was more intuitive – ridership would intuitively be greater at different times of day e.g. morning time when people are making their way to work and evening time when they are going home.

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

Response

The coefficients of the non-dummy variables are `-1.50863832e+02  -1.51550728e+02`

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

Response

The R2 value is `0.463247669262`

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?

Response

The R2 value calculated is above the minimum threshold specified of 0.4.

The R2 is a statistic to evaluate model fit. The closer the R2 value is to 1 the better fit the model in terms of predicting subway ridership.

The R2 computed here would indicate that the model could possibly be improved by either taking a different approach or by adding additional features/predictor variables to the regression function. Ideally the R2 would increase as additional relevant features are added to the function.

If there is no relationship between NYC ridership and the factors ‘rain’ and ‘hour’ then the R2 produced in this model would be 0.

If ‘rain’ and ‘hour’ were the only and sole determinants of NYC ridership then the R2 would be 1.

In this case the R2 value produced by this model is approximately 0.46. This effectively means that 54% of the variability or movement in the response variable (NYC ridership) is not explained by the features (predictor variables in the model) that are captured within this particular model (‘rain’ and ‘hour’). Another way of saying this is that less than half of the variability in NYC ridership is explained by ‘rain’ and ‘hour’.

Hence there are in this case definitely other factors that impact rider ship on the NYC subway outside of whether it rains or not and outside of what time is day is in question.

These conclusions in relation to the R2 value (the model really only goes some of the way) are borne out by the analysis of residuals that is highlighted in my response to 5.1 below.

Section 3. Visualization

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

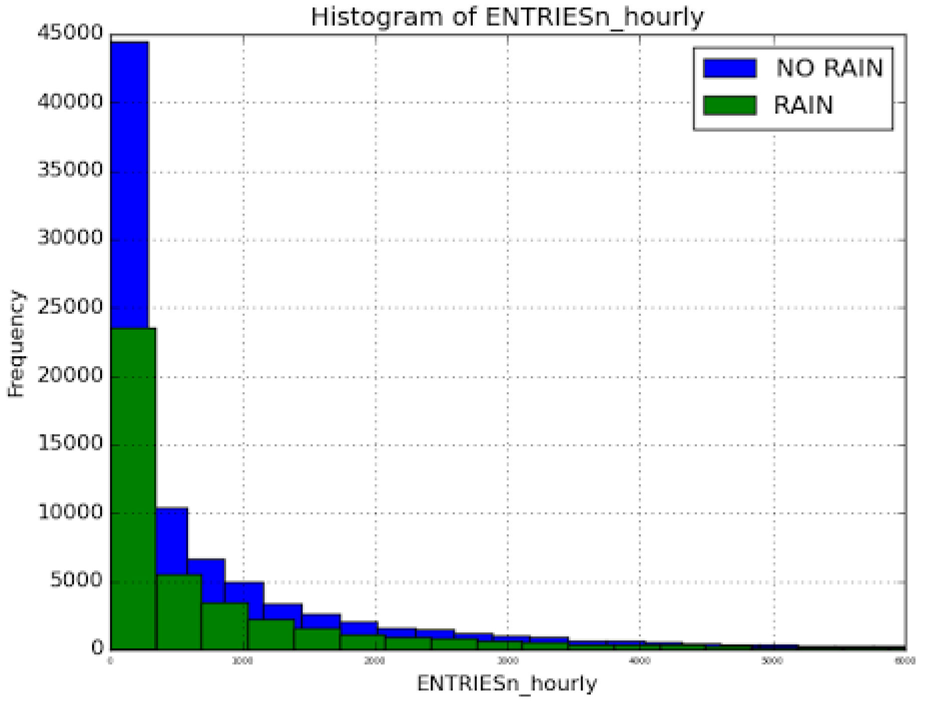
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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.

Response

First visualisation is copied below



Comments

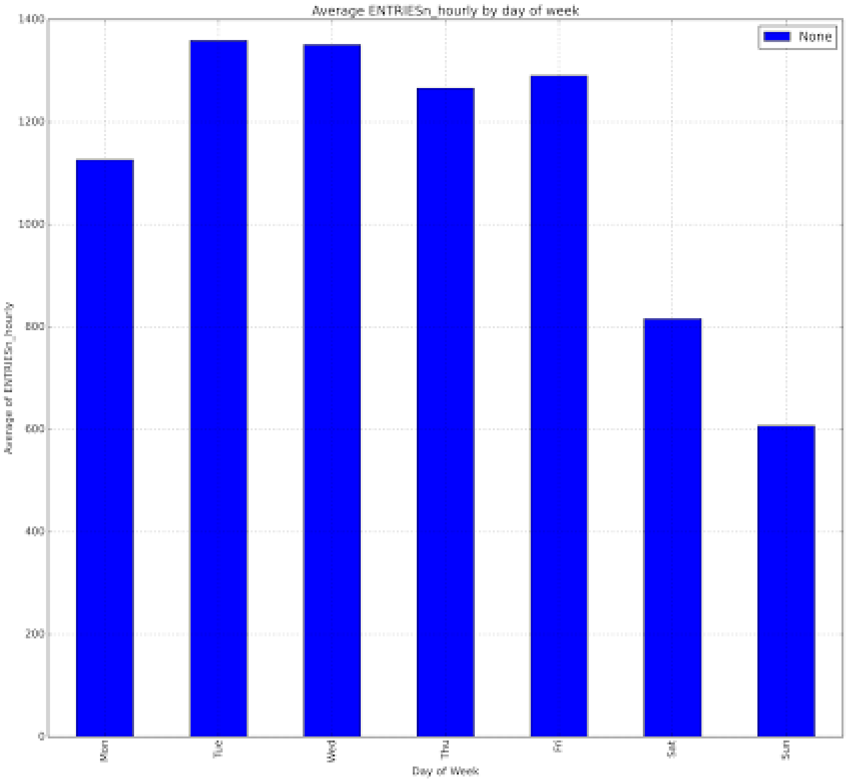
1. This shows that hourly turnstile entries either when it is raining or not raining are not normal in terms of underlying probability distribution
2. As a consequence statistical tests such as Welch's t-test would not be appropriate for this data in terms of tools for analysis.

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

Response

Second visualisation is shown below



This is a bar chart showing the average of ENTRIESn_hourly by day of week i.e. Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday.

The chart could be the starting point of an analysis that may want to explore whether there is a difference in activity depending on what week day is in question or maybe an analysis of working days v's weekend days.

The chart could be used to formulate a null hypothesis in this way.

Obviously the statistical analysis that may ensue is not complete as yet and is not part of the data visualisation project required here.

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?

Response

More people do appear to ride the NYC subway when it is raining compared to when it is not. This is borne out in this analysis in a number of key areas. The first is the histogram produced above that compares entry levels to the NYC subway when it rains v's when it does not rain. The histogram would indicate that the null hypothesis (which is subsequently tested) that ridership levels are equal is possibly not true. The histogram would also indicate that a non-parametric test should be performed to test this null hypothesis as the data is clearly not normal. The Mann Whitney U test would reject the Null Hypothesis that ridership when it rains is the same as ridership when it does not rain.

However the histogram has essentially the same shape for ridership on days when it rains v's days when it does not rain. Hence 'Rain' may not be a determining factor in the underlying distribution of NYC ridership but more a driver of level.

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.

Response

The analysis that leads to this conclusion is the following

- The Mann Whitney statistical U test would reject the null hypothesis that hourly entries when it rains v's when it does not rain are from the same population
- Using linear regression with gradient descent and using 'rain' as a predictor variable or feature does show that 'rain' is a determining factor in the level of NYC ridership. Other factors such as average temperature were also considered and these do not appear to be determining factors in the level of 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:

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

Response

Data shortcomings

1. The analysis project took some time in particular in relation to the visualisation piece due to the format of the data and some issues I had grappling with the code. I appreciate this is my issue as such but it took from the underlying analysis I was trying to perform – in the case of the visualisation I wanted to look at week day as a possible factor for analysis on NYC ridership. I wanted to look at other items beyond 'rain' and I did struggle in terms of visualisation to complete this piece.
2. The data is only taken over a 1 month period which is May 2011. To draw firm conclusions in relation to ridership data over a longer period or inclusion of additional months would be of value.
3. One positive in this regard is that the data is not concentrated on a number of particular hours of the day over a 24 hour period. For example the ridership data is not all taken from 7 am to 9 am in the morning and from 5.30 pm to 6.30 pm in the evening. There is a mix of data over the full 24 hour period
4. Similarly there is also a mix of data over all days of the week which is a further positive

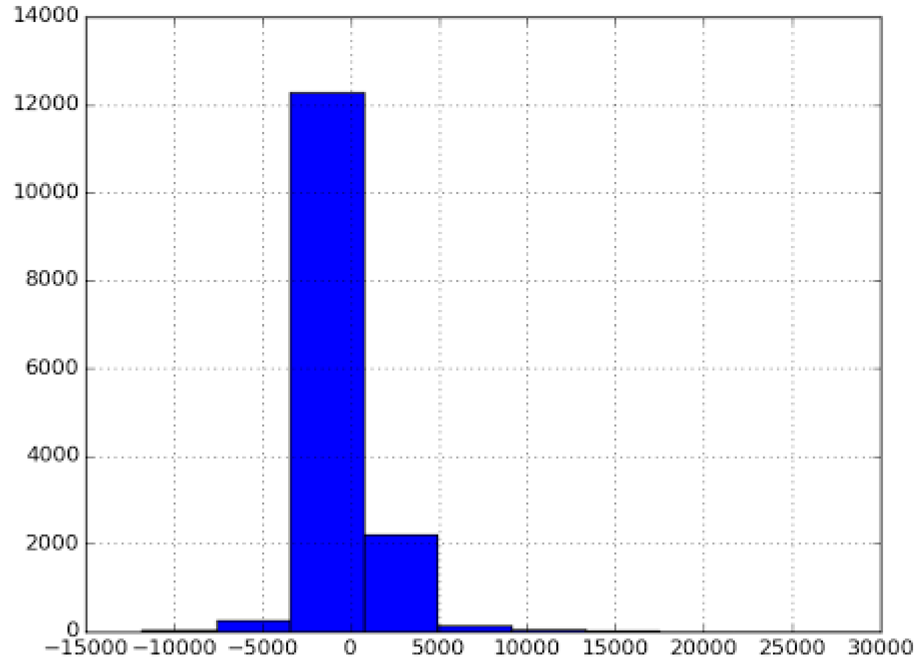
Analysis

1. The linear regression model as formulated could be a better fit to the data. The R2 value of 0.46 while possibly acceptable could be higher.
2. Other variables within the data set could be considered in more such as weather related items - temperature for example. This has been done to a light extent in that a simple scatter plot was developed illustrating hourly entries and mean temperature. While this temperature related analysis is

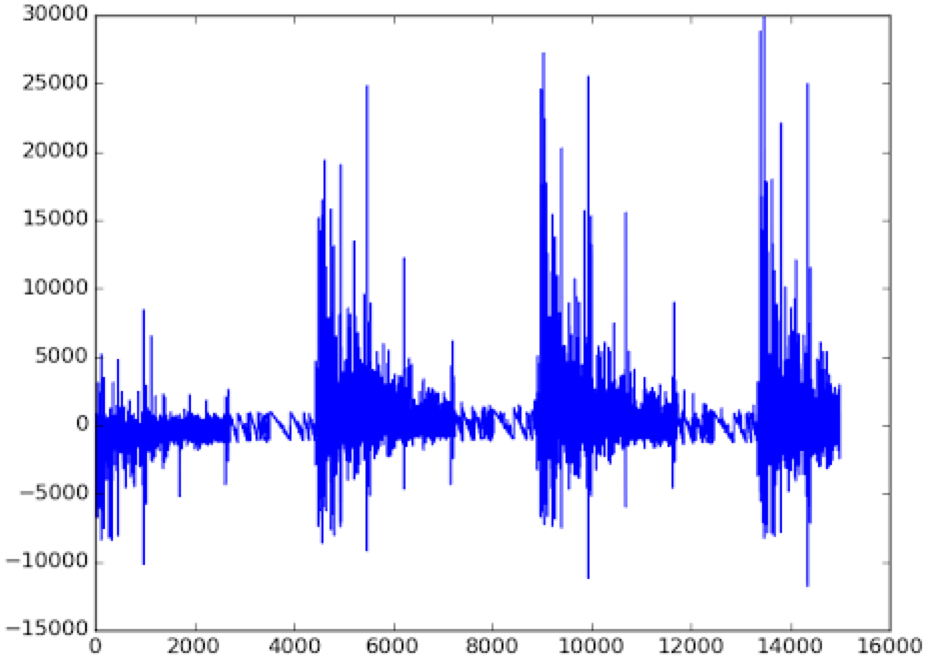
relatively simple it does not provide any definite or concrete clues in relation to how temperature may impact on ridership

3. The analysis could also be further enhanced to consider ridership not only by days that rain and days that do not rain but by day of week e.g. whether (intuitively one would expect this to be the same) ridership varies from week days to weekends or possible more detailed whether there are any significant differences between ridership on different individual week days.
4. However the key factor in determining the quality of the linear model determined is analysis of the residuals from the model i.e. actual data points v's predicted data points. An initial residual analysis is performed in question 3.6 where a histogram of the residuals is considered. This is shown below with 'Frequency' on the y axis and residuals on the x axis.

   There is clearly some bias in the residuals – they are more skewed in a positive direction - which may indicate that a linear model is not the best fit for this analysis and that an alternative should be explored.



   In addition an actual plot of the residuals shown below demonstrates a cyclical pattern which would cast doubt on the linear model chosen and would push the analysis towards considering an alternative. Residuals plot shown below.



5. Lastly the actual coefficients themselves in my regression model merit some comment. The two predictor variables are ‘Rain’ and ‘Hour of day’. In regression with a single independent variable – in this case NYC ridership – the coefficient tells us how much the dependent variable (ridership in this case) is expected to increase (if the coefficient is positive) or decrease (if the coefficient is negative) when the predictor variable increases by 1. In particular in the case of whether NYC ridership is a function of whether it rains or not a coefficient of -150 may indicate the model is not a good fit for the analysis being conducted. This -150 would indicate that as rain increases (in this case rain is either 0 or 1 so is binary in nature) NYC ridership decreases. This may not be the most sensible outcome and would further indicate that the model can be improved.

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

I have two insights that are worth exploring further

1. Mean temperature does not appear to be a significant driver of NYC ridership – more investigation needed
2. Day of week does appear to be a factor so this also merits more analysis – maybe one for another course !!