Analyzing the New York Subway Dataset

Student Notes | Code Review | Project Review

Does Not Meet Specifications

Communication

SPECIFICATION

Analysis done using methods learned in the course is explained in a way that would be understandable to a student who has completed the class.

MEETS SPECIFICATION

Reviewer Comments

Well Done!!

SPECIFICATION

The answers are a well-formed summary of the analyses and do not leave out important information (i.e. fully answering the question).

MEETS SPECIFICATION

Reviewer Comments

Well Done!!

Quality of Visualizations

SPECIFICATION

Plots depict relationships between two or more variables.

MEETS SPECIFICATION

Reviewer Comments

Well Done!!

SPECIFICATION

All plots and data are of the appropriate type.

MEETS SPECIFICATION

Reviewer Comments

Well Done!!

SPECIFICATION

All plots are appropriately labeled and titled. Plot is given an appropriate title. X-axis and y-axis are appropriately labeled. Visual cues (colors, size, etc) are easy to distinguish. It is clear what data are represented.

DOES NOT MEET SPECIFICATION

Reviewer Comments

In figure 3.1 "Distribution of NYC Subway ridership by weather" please notice that the bars of the non rainy entries are stacked above the bars of the rainy entries. This make a wrong impression about the real size of each population. I also notice that you used bar plot instead of histogram, which might cause this. consider using the {position="dodge"} parameter.

Optional : It looks like you may have wanted to change the color of the bars in your chart. To do this you can use the fill parameter instead of the color parameter. You can use the color parameter to select the color of the edges.

Quality of Analysis

SPECIFICATION

When using statistical tests and linear regression models, the choice of test type and features are always well justified based on the characteristics of the data.

MEETS SPECIFICATION

Reviewer Comments

Well done!! In section 1 you are correct, when distribution of the data is not specific, Mann Whitney U test is a good non parametric choice.

SPECIFICATION

Statistical tests and linear regression models are described thoroughly, and the reasons for choosing

How satisfied are you with this feedback? Resubmit Project

Reviewer Comments

Good job you explain clearly the reason that make you include each feature in the regression model. Optional: in your answer you indicate that R squared value was another consideration for the choice of features to the model, I would like to know more about this procedure.

SPECIFICATION

The use and interpretation of statistical techniques are correct.

DOES NOT MEET SPECIFICATION

Reviewer Comments

In Section 2.6: you are correct that high or low values of 'R squared' indicate if the model is poor or perfect respectively. You are also right that the value 0.45 is somewhere in the middle. Try to be more specific here, what R squared tells you about the relation between the predictors (features in the model) and the free variables (in this case Entries per hour). Here is a helpful resource on the topic: http://www.statsoft.com/Textbook/Multiple-Regression#cresidual
Here is another helpful blog post on R squared values: http://blog.minitab.com/blog/adventures-in-statistics/how-high-should-r-squared-be-in-regression-analysis

Excellent interpretation of the results from the statistical test, From the results, could you indicate the exact confidence level for the rejection of the null hypothesis?

Comment: currently, the coefficient for rain in your regression model is negative. One interpretation for this is that, holding all other variables constant, rain will have a negative effect on ridership. Since there are many variables included in the regression, it might be the case that some of the variables are highly correlated, which can cause your coefficient estimates to be unstable. It might be a good idea to build the regression model gradually, to ensure that highly correlated variables are not included together. For more information about multicollinearity, including information about the condition number, see the following Wikipedia article: http://en.wikipedia.org/wiki/Multicollinearity.

SPECIFICATION

All conclusions are correctly justified with data.

MEETS SPECIFICATION

Reviewer Comments

please read my comment below,

SPECIFICATION

No incorrect conclusions are drawn from the data.

DOES NOT MEET SPECIFICATION

Reviewer Comments

In the conclusion section you write "Based on the Mann-Whitney U-test, the distributions of rainy days and non-rainy days are different (mean hourly entries on rainy days is greater than non-rainy days in Section 1.3) ". This contradict your conclusion that "more people ride the NYC subway when it is not raining".

You are correct that "negative correlation cofficient of variable 'rain' " But please be careful the coefficients of a linear model cannot be interpreted as statistical test and the value of the coefficients depends on the contribution of other features in the model. Please read my comment regarding the multicollinearity above.

SPECIFICATION

Some shortcomings of the dataset and statistical tests or regression techniques used are appropriately acknowledged.

MEETS SPECIFICATION

Reviewer Comments

Well Done!!, please consider to include more shortcomings about the data set. Did you notice any outliers in the data set? Do you think the duration of data is enough to the analysis we are doing here?

Learn the best practices for revising and resubmitting your project.

Have a question about your review? Email us at review-support@udacity.com.