Analyzing the New York Subway Dataset

Does Not Meet Specifications

Student Notes | Code Review | Project Review

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

Your submission is quite good but some more work is needed on the concepts, the definitions and the conclusions. I have left as many comments as possible to try helping you and guiding you through the process. Please invest some more time in perfecting your work so that we can build the best possible project for you to showcase in your portfolio. Keep up your good work!

SPECIFICATION

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

MEETS SPECIFICATION

Quality of Visualizations

SPECIFICATION

Plots depict relationships between two or more variables.

MEETS SPECIFICATION

SPECIFICATION

All plots and data are of the appropriate type.

MEETS SPECIFICATION

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

Please add a short description below each figure commenting on the key insights depicted in the figure, this is a mandatory requirement of section 3.

Please label the X axis with the appropriate day names instead of using numbers to improve plot readability.

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.

DOES NOT MEET SPECIFICATION

Reviewer Comments

In this case it might be a good idea to use a two-tailed statistical test. Picking a one-tailed test means that we assume in advance (before we collect the data) that rain will not be associated with lower ridership, which is a very strong assumption. For more information, see the following: http://graphpad.com/guides/prism/6/statistics/index.htm?one-tail_vs__two-tail_p_va lues.htm
By default the Scipy Mann Whitney U returns a one tailed p value. In order to get the two tailed value you would need to double it. For more information you can refer to: http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html

Optional: Please note that the Null Hypothesis is not that: “there is no statistical difference in hourly entries between rainy and non-rainy days” in fact the purpose of statistical tests is to compare distributions of two variables and some related statistics. In this particular case we are assessing the chance that a randomly selected value from the population with the larger mean rank is greater than a randomly selected value from the other population. Please note that an exact statement of the null hypothesis can be found in the downloadables from Lesson 3. The downloadable notes about the Mann-Whitney U test can be accessed by clicking on the appropriate link below the video window of any of the Lesson 3 videos.

Good job in providing proper arguments when justifying the usage of the Mann Whitney U test. Well done!

SPECIFICATION

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

them are articulated clearly.

DOES NOT MEET SPECIFICATION

Reviewer Comments

Question 2.4 requires the coefficients or θ values of each of the non-dummy features in the fitted linear regression model - this will need to be reported from your OLS code. Please provide these values. These coefficients ('theta') are discussed in Lesson 3 of Intro to Data Science.

In 2.6 in order to meet specification it would be necessary to discuss more thoroughly the meaning of the R squared and the result obtained by the model. A generic statement is not sufficient. What is the meaning of the R squared? How would you assess the performance of the R squared you obtained? For more information about R^2 value, please see this webpage: http://www.statsoft.com/Textbook/Multiple-Regression#cresidual.

Please note that the second part of question 2.6 requires to state whether a linear model is appropriate for this dataset: To assess whether a linear regression model is adequate to fit a set of data it is good practice to examine the distribution of the residuals, the difference between the predicted and the actual values. Please note that the histogram of the residuals has long tails, which suggests that there are some very large residuals a reason to question our linear regression model. You can check the normality of residuals visually or, to have a clearer perspective, you could use a probability plot such as: http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.probplot.html. More information is available here (QQ plot is a type of probability plot): http://en.wikipedia.org/wiki/Q%E2%80%93Q_plot http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.probplot.html.

SPECIFICATION

The use and interpretation of statistical techniques are correct.

MEETS SPECIFICATION

SPECIFICATION

All conclusions are correctly justified with data.

DOES NOT MEET SPECIFICATION

Reviewer Comments

The statistical test performed actually confirmed that there is a difference: the two tailed P value is less than the chosen P critical value. We reject the null hypothesis at the 95% confidence level that the distributions are the same and accept the alternate hypothesis that they are different. So the means are actually different with statistical significance and more people ride the subway when it rains. For more

How satisfied are you with this feedback?

Resubmit Project

No incorrect conclusions are drawn from the data.

MEETS SPECIFICATION

SPECIFICATION

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

MEETS SPECIFICATION

Reviewer Comments

Optionals:

1. You could explore more thoroughly the limitedness of the linear model and its implications as you successfully did for the dataset. You could simply extend to 5.1 the reasoning regarding the residuals I proposed in my comment in 2.6. In addition it might be interesting to plot the residual per data point, some interesting patterns might emerge to help understand why a linear model might not be the best choice for this problem. By merely plotting the difference between predictions and actual values (residuals) you will, most likely, see that the residuals follow a cyclical pattern. If so, that might prove that some non-linearity in the data should be addressed by designing a non linear model. The code is really simple and looks like this: import matplotlib.pyplot as pltplt.plot(data - predictions) plt.show()
2. What about the database itself, do you think it covers a long enough time span?
3. Because there are many variables included in the dataset that might be very closely related, such as minimum, mean and maximum temperature, it may be difficult to disentangle the effects of such similar features and we may run the risk of problems with collinearity, which can cause some linear regression algorithms to give incorrect results.
   http://en.wikipedia.org/wiki/Multicollinearity

Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

Watch Video (3:01)

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

Blog
News & Media
Developer API

UDACITY

About
Jobs
Contact Us
Legal