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
Your submission is excellent; it is thorough and well argued. Section two is really outstanding. There are just a few statements I don't agree on in section 1

1. A clearer statement of the null hypothesis is needed.
2. The test is statistically conclusive.

An impressive work, clearly above average.

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

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.

MEETS SPECIFICATION

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 1.1 please note that the Null Hypothesis is not that: "subway ridership is the same when it is raining as when it is not" it is not clear which variables exactly are compared how they are compared (when comparing ridership are we comparing the ridership's sample means? the population means? or what else? please specify).
We know already what the sample means are, and we are trying to infer something about the population, so the null and alternative hypotheses should be concerned with inferring information regarding the population.
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.
Over the hypothesis statement and meaning: http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/basics/what-is-a-hypothesis-test/

Please clarify when stating further on: "hence we are performing a 2-tailed test." It is correct, but the rationale is not obvious.

In 1.4 it not correct to state that: "these two points alone are insufficient to indicate that rain is causing an increase in ridership". 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 information: http://stattrek.com/hypothesis-test/hypothesis-testing.aspx .

SPECIFICATION

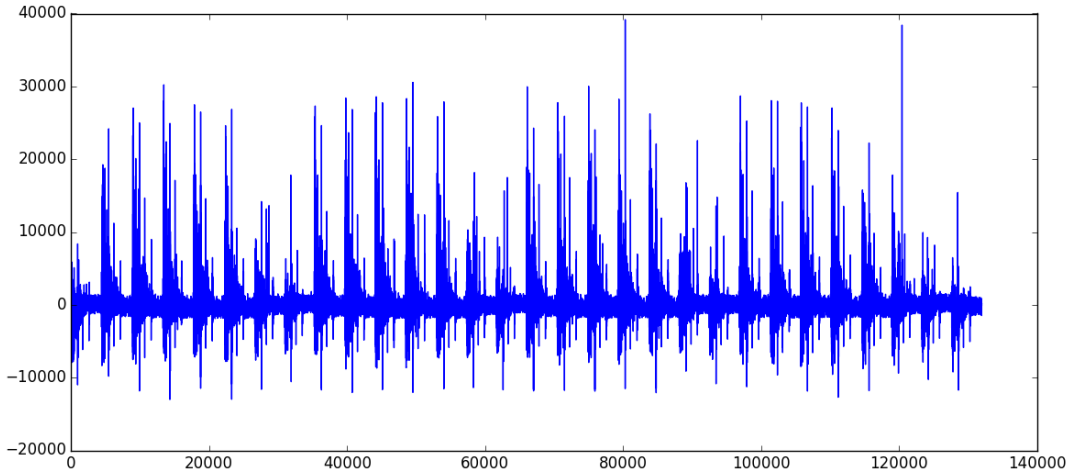
Statistical tests and linear regression models are described thoroughly, and the reasons for choosing them are articulated clearly.

MEETS SPECIFICATION

Reviewer Comments

An outstanding section 2: Regarding the excellent analysis of residuals the issue is that possibly a linear model cannot possibly capture the complexity of a non linear data generating process. That's why we have a non Gaussian residual curve. I'm attaching a plot of the residuals per data point on the small dataset. The fine grain will allow you to appreciate the cyclical pattern that normally emerges from a model with an R squared of 0.5. Your model is much more complex but I think the picture might be telling anyway.

Residuals per data point, the code would be really simple and could look like this:
import matplotlib.pyplot as plt
plt.plot(data - predictions)
plt.show()



SPECIFICATION

The use and interpretation of statistical techniques are correct.

MEETS SPECIFICATION

SPECIFICATION

All conclusions are correctly justified with data.

MEETS SPECIFICATION

How satisfied are you with this feedback?

Resubmit Project

SPECIFICATION

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

MEETS SPECIFICATION

Learn the best practices for revising and resubmitting your project.

Have a question about your review? Email us at review-support@udacity.com and include the link to this review.