**STUDENT** Karen Thurston

**COURSE** Intro to Data Science v. 2

Hi Charlotte,

Thank you for your detailed analysis of my work. It has helped me to better understand the statistical tests. I viewed all of lesson 3 again and made a chart for myself of the statistical tests covered in that lesson to know when it is appropriate to use each. The most important thing I learned is that the tests come with their own predetermined NULL hypothesis—I had somehow missed this and thought that I could choose what I wanted to the NULL hypothesis to be, which led to the incorrect conclusions about the data in my analysis.

I have corrected all the errors you have highlighted in blue, but believe there is a misunderstanding about some of my work. I thought it would be easiest to make notes within the body of your Project Report, so you will see my notes in ***green italic text***.

I hope my explanations and changes to my project submission now fully meet the requirements, but if not, I am looking forward to additional guidance to improve.

Karen

Hi Karen,

Thank you for your resubmission! You’ve made some useful improvements to your work. The linear regression section is still missing a few details. There is an important misconception about the null hypothesis of your statistical test which has consequences on the conclusions drawn in Sections 1 and 4. I have comments about this and other aspects of the project in the rubric. The bullet points highlighted in blue need to be addressed for your project to meet specifications. Feel free to email dataaanalyst-project@udacity.com or post on the Discussion Forum if you have any questions. I look forward to your resubmission!

Charlotte and the Udacity Team

**Communication Meets Specifications**

● 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.

● The answers are a well-formed summary of the analyses.

In most industries, a data analyst will be required to write short summaries of a detailed

analysis. A long report might simply be discarded by busy colleagues so sensible editing is an

important skill to learn. Thank you for the improvements you’ve made! Some discussion (for

example, of how the Mann-Whitney U-test works) could be summarised to shorten the

submission. Including the code in an appendix rather than the main project is appropriate -

thank you!

**Quality of Visualizations Does Not Meet Specifications**

● Plots depict relationships between two or more variables.

● All plots are of an appropriate type.

Well done for the improvements to your plots!

● Some plots are not appropriately labeled and titled or visual cues are not always easy to

distinguish. It is not clear what data are represented.

Please reduce the binwidth on Figure 3.1a - this is currently too wide to properly summarise

the lower values of the data. Including a larger version of the plot would make it more easy to

read. ***FIXED***

Figures 3.1a and 3.1b would be clearer if coloured bars were used, rather than only outlinesthis

would allow the reader to read the frequency for each bin. I think this can be achieved

with histtype = bar. (By the way, I am not sure that the code included for Figures 3.1a is

actually the code used to generated Figure 3.1a) ***FIXED***

I think it was a very good idea to include Figure 3.1b to show more detail of the histogram,

well done! ***THANK YOU***

I believe there is a mistake in labelling days on Figure 3.2: I think that Monday corresponds to

0 and so on. Please check your work and correct if necessary. ***FIXED***

It is not clear what data is represented in Figure 3.2 - are these total values for a single week in

May, a mean/median daily total, or something else? **They were totals.** Are these aggregated across all UNITs? **Yes**

Please note that it would not be appropriate to simply take a sum across all days in the month

- there are 30 days in the ‘original’ dataset, and 31 in the ‘improved dataset’, so there are some

weekdays that occur 5 times in the month, and some that occur 4 times. This could easily

mislead a reader and should either be clearly documented or (preferably) an average taken

instead of a total across the month. ***FIXED, average used, I agree it was not appropriate to use the totals.***

It would be a good idea to find a clearer way to annotate the vertical axis on Figure 3.2, rather

than using scientific notation (1e7). ***FIXED***

If possible, try to modify the bars in Figure 3.2 so that they are evenly spaced. ***FIXED***

**Quality of Analysis Does Not Meet Specifications**

● The choice of statistical test type, features, and linear regression models are appropriate based on

the characteristics of the data.

● Statistical tests or linear regression models are not described thoroughly, or the reasons for

choosing them are not clearly articulated.

The null hypothesis in 1.1 is stated incorrectly. The null hypothesis for the Mann-Whitney

U-test is consistent with the two distributions being the same. See the first page

‘downloadables’ material on Mann-Whitney U-test for more information. This leads to

incorrect interpretation of the p-value, the U-statistic and the differences between means and

medians in light of this result. Please carefully work through your analysis in Section 1 and

Section 4 using the correct null hypothesis to avoid mistakes. ***FIXED***

In 1.1 you state that you use a one-tailed p-value, but in 1.4 you use a two-tailed p-value for

your statistical test. I think the mistake is in 1.1. The reason given for using a one-tailed

p-value is not correct here - this part should probably fit into 1.2, as it looks like a reason why

a Mann-Whitney U-test was used. **FIXED**

Please explain why you used 'fog' and 'UNIT's in your feature set (question 2.3). What made

you think that these might help to predict the number of riders? **SEE EXPLANATION IN MY PROJECT**

In Question 2.4, please report which feature relates to which coefficient for each of the

non-dummy feature, so that you can interpret these coefficients. **FIXED**

● Mistakes are made in use or interpretation of statistical techniques.

Well done for your work in learning about the U-statistic. It is quite difficult to interpret this

value ‘by eye’, which is why we use the p-value to understand how unlikely our reported

U-statistic should be. The maximum possible value of U (corresponding to a p-value of 0.5) is

half of the product of the two sample sizes, not their product as you state. (This is because the

U reported is the minimum of the two summed ranks from the two samples.) **Please note my original explanation in my project: “the value of U should be about one half the maximum if the null hypothesis is true. My U value is 1,924,409,167 which is just under one half of the maximum”**

Even though our U-value is seemingly close to the maximum possible U, the value of U (under the assumption of the correct null hypothesis) is approximately normally distributed with variance defined depending on the sample sizes and it is using this normal distribution we find that that a

result this extreme is unlikely (p = 0.038) under the null hypothesis. **I BELIEVE I WAS CORRECT IN MY ORIGINAL STATEMENT ABOUT the U-STATISTIC (see above quote from my project). I have also highlighted the relevant text in red in my project report.**

● Some conclusions are not correctly justified with data.

In Question 2.6 please state whether you think the linear model is appropriate to predict

ridership for this dataset. This should be discussed in terms of the sources of data, the aim of

the model and the performance of the model (R^2 value and perhaps residuals). **DONE**

In Section 4, you should use the Mann-Whitney U-test result supported by descriptive

statistics about the two samples and any relevant linear regression coefficients to support

your conclusion. It would be a good idea to restate the numerical values you use and the

criteria you use to interpret them. (For example, comparing the p-value to the critical value or

comparing descriptive statistics to one another.)

● Some of the conclusions drawn are incorrect.

This follows from problems with null hypothesis stated above. **FIXED**

● Some shortcomings of the statistical tests or regression techniques used are appropriately

acknowledged.

Can you think of any shortcomings of your statistical test or linear regression techniques?

PROJECT EVALUATION

**Project Does Not Meet Specifications**