UDACITY

Nanodegree Project Evaluation

STUDENT Daniel Ruiz

COURSE Intro to Data Science

Hi Daniel,

Thank you for your project submission. I have just a few comments for you in the rubric below. The bullet points highlighted in blue need to be addressed for your project to meet specifications. Feel free to email dataanalyst-project@udacity.com if you have any questions. I look forward to your resubmission!

Charlotte and the Udacity Team

- Please answer Question 2.4, giving the coefficients of the non-dummy features from your linear regression model. These are the 'theta' values that are used to multiply the features when making predictions. See Lesson 3 of Intro to Data Science for a review of this topic.
- For each plot in Section 3, please provide a description giving some comment on insights into the data that the plot shows.

Click here to tell us whether this feedback was helpful.

Communication **Does Not Meet 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 not well-focused (e.g. stream of consciousness) or leave out important information (e.g. not fully answering the question).

  Action: In Question 1.1 you mention that you use a two-tailed test. For clarity, it would be best to be consistent and report and use your two-tailed p-value in 1.3 and 1.4. Do not switch between different kinds of tests during the hypothesis testing process.

  Action: Please clearly state the null hypothesis in Question 1.1. This should be a null hypothesis of 'no effect' or no difference between the two groups. For more information about the null hypothesis for a Mann-Whitney U-test, please see the Downloadable information from Lesson 3 of Intro to Data Science.

Comment: You should think about how you communicate with the reader in your final project. A question-and-answer format is appropriate, but writing your responses to each question in full sentences would be easier to read.

Comment: Consider your reader when reporting numerical results – how many significant figures are they likely to find useful?

Quality of Visualizations — Does Not Meet Specifications

- Plots depict relationships between two or more variables.
- Not all plots are of the appropriate type.
  Action: The plots in Section 3.2 are not suitable for the data. I don't think that the joining lines between different hours' bars should be included – they don't encode any real information about the dataset. I think (from experience) that the bars here show the maximum values of ENTRIESn_hourly for each hour – is this what you mean to show?
  You should consider a clearer way to summarise this data: perhaps a bar plot (geom_bar) of the mean/median or max values or a set of box or violin plots to summarise the distributions of ENTRIESn_hourly for each hour would be best.

  Comment: I think this is an interesting choice of data to visualise – it would be nice to be able to see how ENTRIESn_hourly changes at different times of day depending on whether it is raining or not.
- 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.
  Action: It is not clear what is being visualised in the plots in 3.2. You should state whether you are showing a mean/median/max/total per hour/day/for the whole month and whether this is for a single UNIT or a total for all UNITs. Don't forget to add some text regarding insights into the data.

Quality of Analysis — Does Not Meet Specifications

- The choice of statistical test type, features, and linear regression models are sometimes not appropriate based on the characteristics of the data.
  Action: It is mentioned in 2.2 that precipi was used as a dummy variable features in the linear model. Because precipi is a quantitative rather than a categorical variable, it is not clear how or why it was converted it to dummy variables. Please explain or correct your work.
- Statistical tests or linear regression models are not described thoroughly, or the reasons for choosing them are not clearly articulated.
  Action: In Question 1.1 you mention 0.02449 as a p-critical value – I think this might actually be a calculated p-value, rather than a p-critical. Please adjust your work.

  Action: It is correct that a Mann-Whitney U-test is a non-parametric test that does not make any assumption that your samples are drawn from any particular type of distribution. Can you explain why this is a useful property for the rainy/non-rainy ENTRIESn_hourly data? You might want to use your histograms from 3.1.

  Action: From looking at the R^2 value, it seems that you may have used dummy variables based on UNIT in the linear regression model. Please check your work and make sure all features are reported and justified in 2.2 and 2.3.

  Action: No justification for 'precipi' is given in 2.3 – why did you choose to use it in your model, as well as 'rain'?

Comment: Well done on your justification of 'rain' in 2.3: choosing features that are widely applicable to future predictions is a great idea. Using 'Hour' in order to look at the effect of other features holding the time of day fixed is also a good idea.

- The use and interpretation of statistical techniques are correct.
- Some conclusions are not correctly justified with data.

  Action: In Question 1.4, please use the p-value calculated (1.3) and the critical p-value chosen (1.1) to state your conclusion in terms of rejecting or failing to reject the null hypothesis.

  Action: The value of R^2 (given in 2.5) describes the proportion of the variance in the dataset that is explained by the model. I think that your model has a reasonable value of R^2, given the nature of the dataset. Take a look at these resources http://www.statsoft.com/Textbook/Multiple-Regression#cresidual and http://blog.minitab.com/blog/adventures-in-statistics/how-high-should-r-squared-be-in-regression-analysis on the topic, and give further details on the appropriateness of your model.

  Comment: I suspect (this is from my own experience as much as anything) that adjusting alpha will not affect R^2 too much, the model is probably close to being correctly fitted. To check, you could experiment with different alpha or check the R^2 using a different method (for example, statsmodels' OLS function) to fit the model with the same features.

  Action: In Section 4 you have some interesting ideas about conclusions to be drawn about the effect of rain on the number of subway riders. However, I don't think that these ideas are properly backed up with data – this would require further analysis, such as more specific statistical tests on subsets of the data, or further cleaning of the data to remove 0 values. It is not appropriate to draw conclusions solely from visualisations.
  You could perhaps start with your more general conclusion, covering the statistical test result from Section 1. You should definitely mention whether the null hypothesis from this test is rejected or not, and use the two means to help you interpret this finding. Once you have given your linear regression coefficients in 2.4, the coefficients of rain and precipi would be relevant information to support a conclusion. The R^2 value of the linear regression model doesn't really tell us anything about any apparent effect of rain on ridership, but could be useful if you want to back up the idea that those coefficients are really correct.
- No incorrect conclusions are drawn from the data.
- Shortcomings of the statistical tests or regression techniques used are not appropriately acknowledged.

  Comment: You are right to have concerns about the appearance of zero values of ENTRIESn_hourly in the dataset. It would be a good idea to dig into the data and see whether these are valid (after all, it is possible that no one enters a particular UNIT at a certain time of day) or not.
  Action: In 5.2, it is not the case that the Mann-Whitney U-test does not provide much output – a statistical test is an important statistical tool for inference, the output is the rejection (or not) of the null hypothesis. Descriptive statistics are always needed to properly understand a test result.
  Comment: Well done - you spotted some genuine issues with linear regression in general. Can you think of some ways to improve your analysis in this case?

PROJECT EVALUATION

Project Does Not Meet Specifications