UDACITY

Nanodegree Project Evaluation

STUDENT anant salame

COURSE Intro to Data Science

Hi anant,

Thank you for your project submission. On the whole, it is precise and clear - well done! I have 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 or post on the Discussion Forum if you have any questions. If you wish to change the capitalisation of your name on evaluation and certificates, you can adjust this under "Account" on the Udacity site.
I look forward to your resubmission!

Charlotte and the Udacity Team

Click here to tell us whether this feedback was helpful.

All questions must be fully answered in order to meet specifications. Currently the following questions are not answered in the project:

- Section 2, Question 6: Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?
- Section 3: Please add a short description below each figure commenting on the key insights depicted in the figure.

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.
  Well done on a concise and clear presentation of your analysis!

Quality of Visualizations — Does Not Meet Specifications

- Plots depict relationships between two or more variables.
- All plots are of the appropriate type.

- 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.

  The first visualisation is good, but the horizontal axis is cut off at 10000 entries. The largest value of ENTRIESn_hourly in the dataset is over 50000, so there is a 'long tail' of ENTRIESn_hourly not depicted in the plot. To ensure that the whole distribution is summarised, you should consider either adjusting the axes, providing an additional plot or including a note to explain how the data has been manipulated for this plot.

  The way that the second visualisation's bars have been labelled with the day of the week is good. However, the bars do appear in an unexpected order: for the plot to be easily interpretable, please adjust the order to the usual sequential order.

  It is not clear exactly what data is being represented in the second plot. Is this a sum, mean, median or other aggregation of ENTRIESn_hourly? Is this calculated per UNIT, per hour, per day or across the whole month?

  It would be clearer to use 'normal' numbers rather than scientific notation (1e7) on the vertical axis of this plot.

Quality of Analysis — Does Not Meet Specifications

- 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.

  This is beyond the scope of the course, but it may not be appropriate to include so many variables in the linear regression model. 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.

  The statement of the null hypothesis for the Mann-Whitney U test captures the general idea of the test. 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.
- Statistical tests or linear regression models are not described thoroughly, or the reasons for choosing them are not clearly articulated.

  In Section 1, Question 1, please report the p-critical value for the statistical test, not a p-value calculated during the test. The p-critical value is related to the significance level and is the threshold used to determine whether the observed U-statistic is statistically significant.

  It seems likely to me that you used dummy variables for UNIT as features for your linear regression model - it is difficult to achieve such an R^2 without them. Please make sure that all features are reported and justified in your report.

  It would be nice to see more specific justifications for the many features included in your linear regression model. What intuition led you to use them?

  The justification of the Mann-Whitney U-test is mostly correct -- it does not assume any particular distribution of the data. It is correct and important that the Mann-Whitney U-test requires two completely separate samples. However, the sense in which 'independence' is required is a little different to this. It means that it requires that all of the observations, from both groups to be independent of one another.

- Mistakes are made in use or interpretation of statistical techniques.
  Although you correctly choose to use a two-tailed test, the p-value reported is a one-tailed p-value. Please adjust the p-value calculated to reflect the two-tailed nature of the test. You may wish to read the SciPy documentation for the Mann-Whitney U-test for more information.

  In Section 2, Question 6, the R^2 value gives the proportion of the variation in the sample that can be explained by the fitted model, rather than simply by the features themselves. (For example, if there were an exact but more complex relationship between the chosen features and the dependent variable, this information would not be reflected in the R^2.)
- Some conclusions are not correctly justified with data.
  Well done for correctly interpreting the hypothesis test result and the coefficient of rain from the linear regression model in your conclusion in Section 4. Your reservations about the use of the linear regression coefficient are well-founded! You should also refer to descriptive statistics, such as the mean values of ENTRIESn_hourly for rainy and non-rainy days, in order to form a conclusion based on all of your analyses. You should attempt to form a coherent conclusion from your work in the rest of the report in this section.

  Your comment about the histograms seems to show a slightly incorrect interpretation of these plots. Whilst it is correct that the non-rainy day histogram "shows more values of Entries", this is simply a matter of overall frequency: here it is a reflection of the fact that there are more non-rainy days than rainy days in the dataset as a whole. This observation does not contribute to the answer about whether more people ride the NYC subway on a rainy day or a non-rainy day.
- No incorrect conclusions are drawn from the data.
  (See comments above about justification of conclusion.)
- Some shortcomings of the statistical tests or regression techniques used are appropriately acknowledged.
  Well done for coming up with some problems in the dataset and analysis!

  You mention that there are more non-rainy days than rainy days in the dataset, and say that this "creates bias". I'm not sure exactly what you mean here, but the Mann-Whitney U-test does take the two sample sizes into account, so there is no bias caused in the test result, at least.

  You are correct that for many UNITs, data is captured only at certain intervals, typically every four hours. This is not missing data but a variation in sampling interval.

PROJECT EVALUATION

Project Does Not Meet Specifications