

Final Project Summary
David Suffolk
August 9th, 2019

Statistical/Hypothetical Questions

In my final project, I looked into whether an app that was paid for had higher customer satisfaction than apps that were free. The app marketplace offers multiple free apps or free versions of paid apps. Therefore, the hypothesis is that if a customer decides to purchase an app, they are more likely to be satisfied with it and provide a higher rating.

Outcome of your EDA

Through my EDA, I was able to find some support for the theory that paid apps receive higher ratings than free apps but it was not consistent or significant. When reviewing the visualization of the PMF, there was a slight increase in ratings for paid apps over free apps. The CDF also showed that a paid app had a small chance of getting a slightly higher rating. When the coefficient was calculated, a low to moderate positive relationship between ratings and app type was present. However, when the null hypothesis of no difference between the means of the ratings for both app types was tested, there was no statistical result to adjust the null hypothesis.

What do you feel was missed during the analysis?

I was able to find a slight improvement in ratings for paid apps against free apps but was unable to fully determine where and when that occurred. While I looked at installs, number of reviews, and content ratings, there was no additional context added to the understanding in this particular analysis.

The data also showed that a significant portion of the ratings given by users are 4 and 5. While I approached the ratings on its full scale, it may have been useful to just investigate the apps with ratings between 4 and 5 as that magnification may have provided more context.

Were there any variables you felt could have helped in the analysis?

The original dataset included category and genre data and I think this information could have given more context to any differences in the app marketplace for free and paid apps. Furthermore, it would have been interesting to look into the details of the paid apps and see if there was a certain price bracket that impacted reviews (ex: apps for \$10 vs \$1).

Were there any assumptions made you felt were incorrect?

I think my assumptions were on the right track but the analysis did not provide enough strength to the argument that paid apps have more customer satisfaction than free apps.

What challenges did you face, what did you not fully understand?

My biggest challenges were with formatting and cleaning the data. For example, I assigned numerical values to variables so that I could perform the statistical analysis and I was able to work with this. However, it made the presentation of the data less clear.