Taylor Imhof

Bellevue University | DSC 530

Final Project: Statistical Question Summarization

11/19/2021

**Project Overview**

For this final project, I simulated that I was hired by a new video game development studio for data analysis. Their primary goal was to understand which genre of video game has the highest likelihood of being successful. At our business understanding meeting, we also hashed out other objectives, such as the timeline for the project and available resources. Other questions that were brought up included which region should be targeted for marketing and which platform should be selected for their primary development efforts.

After meeting with the dev team and getting a feel for their goals, I searched for a useful dataset that could help answer their questions. I found a great dataset from Kaggle.com with tons of historical video game sale data (Kirubi, 2016). The data set included several categories of sales metrics, including North American and European sales and sales in the Japanese market. There was also an “*Other”* sales category for all other sales. In addition, there were also metrics tracking rating and count data for both users and critics. After reviewing the data, I came up with a hypothesis that I would be able to predict which genre would provide the game studio with the highest likelihood of being successful (measured in predicted global sales).

This project has served as a great test of my newfound data analytics skills gained throughout this course. From initial data import and cleaning to exploring data visually and statistically, I have a much greater appreciation for the basic processes involved in a data science project.

**Initial EDA (Import and Cleaning)**

The primary technology used to conduct EDA for this project is the Python programming language. The Anaconda distribution of Python comes packed with many powerful packages that help streamline data analysis efforts. These packages include Pandas, NumPy, SciPy, Matplotlib, and Seaborn.

Reading in the data was very straightforward. However, I quickly found that cleaning the data was going to take some time. There were tons of null values, created mostly where critic and user ratings were not included for the observations. I tried to find a method of imputed values for these to keep them in the final data set. However, I wound up opting to drop them altogether. Since the sample size was already quite large (~17,000 rows), I felt this did not impact my statistical analysis. A few column data types also needed to be converted, though this was easily accomplished via casting between types. Finally, a dataset subset based on only recent platforms (e.g., Xbox360, PS3) was created for more focused analysis.

**Data Understanding**

After getting a final, clean data set, the next step was to visualize the distributions of each of the variables/columns. For this, I opted to use histograms and boxplots from both the Matplotlib and Seaborn packages. I found the boxplot function for Seaborn to be particularly useful in identifying the existence of outliers, at least with my video game data frame. A nice feature of Seaborn’s histplot function was that I could include a kernel density estimation graph on top of the histograms.

After creating the histograms for my selected variables, I attempted to create a probability mass function (pmf). I will say that the methodology outlined in *Think Stats* was very hard to re-create in my project. I wound up using Seaborn’s barplot to see the PMF for release years of action games versus non-action games. Creating scatterplots was relatively easy, as Matplotlib’s scatter function is intuitive to implement. I chose to plot the relations between user/critic scores as well as other/global sales. I chose these as I felt they were going to be good features for predicting overall success. Seaborn’s regplot function plots a scatterplot and a regression line to the graph, making it easy to identify variable relationships.

After identifying positive visual relationships between the data, I tested my hypothesis using Pearson’s *r* for correlation testings. I found both relationships to be statistically significant (p < .005), and the correlation coefficients were also quite strong (particularly between total and global sales).

**Desired Variables**

After working through this project, I felt that since North America made up the vast majority of all sales observations, I thought it would be nice to have an additional column or data set focused on regional sales specific to the US and Canada. In addition, since these regions are the primary consumers of video games and have much higher populations, I feel that focusing on these specific regions would provide more useful insights. Perhaps I can revisit this project after finding another data set that contained such information in the future.

**Project Challenges**

*Think Stats* was a great resource for grasping foundational statistical concepts via Python code. However, since Downey provided most of his own modules rather than using industry-standard packages, it was difficult to follow his implementation when working on the final project. Thankfully, the Python community at large has an amazing capacity for detailed documentation. As a result, the official docs for all of the packages I used in this project were straightforward to follow (Google and StackOverflow also helped immensely).

Overall, I felt this project put my practical skills learned during this class to the test. It was also nice being able to bounce ideas off my classmates as we progressed through the course. I am looking forward to further developing my skill set with future projects and endeavors.

References

Kirubi, R. (2016). Video Game Sales with Ratings. *Kaggle.com*. Retrieved from <https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings>