CS510 - Explorations in Data Science
Andrew Wilson
Austen Hsiao

Midpoint Report - Game Sales and Visualization

Midpoint Meeting at 11:30 AM on 7/22

Project Objective and Approach Update:

Originally our aim was to answer several questions about video game sales and do so through the use of visualizations on game sales datasets. What we found though, was that it is actually extremely difficult to find various sources for this data unless we are willing to pay for an industry report. Due to that we were limited to working with a dataset primarily sourced from a single location (vgchartz). This is less than ideal, so we decided to pivot the project objective a little to still utilize this dataset, but focus more on the visualization piece.

We now want the main focus of the project to be on a comparison between using Python for visualizations versus using a fully powered business intelligence tool, in this case, Tableau. This allows us to use the dataset we already have (with known validation issues) and continue our data science exploration. With this change as well, we want to tighten up the number of ‘questions’ we’re aiming to answer to a more reasonable 3 to 5. That will more easily enable a one-to-one comparison of the visualization tools.

From the project plan, some of the queries we originally sought to explore included:

- Publishers with lowest review scores
- Most popular genres by sales
- Platforms with best user scores

Since our focus is now on visualization, we can tweak these queries into graphable datasets or time series:

- Average critic review scores by publisher
- Sales by genre (by platform?)
- User scores by platform

Within some of these datasets, we can approach them in multiple ways—for example, “user scores by platform” can be graphed as average user scores for all games vs. platform, or as individual user scores for video games that exist across multiple platforms.

Team Structure Update:

This has also caused us to switch up our team structure a bit. Before we had planned to pair program on our endeavors, now we have a nice divide where one of us can be the primary in exploring Python, the other exploring Tableau. The plan is to still tackle the questions and

CS510 - Explorations in Data Science
Andrew Wilson
Austen Hsiao
presentation creation together, along with cross training and updates along the way of what is difficult in one tool versus the other.

Project Milestone Update:

Milestone	Completed?
Main dataset identified	✅
Additional sources identified for corroboration	✅
Visualization techniques/tools explored and main tools chosen	✅
*Research presentation	✅
*Original questions trimmed to key 3	✅
*3 visualizations completed in Python	In progress…
*3 visualizations completed in Tableau	In progress…
Final presentation assembled	

**These milestones were modified or added from the original project plan*

Our main dataset was sourced from kaggle.com and contains approximately 50,000 entries. As mentioned in the objective update, we had difficulty finding alternative data sources that weren't scrapped from vgchartz. A good portion of time was spent trying to find corroborating and/or alternative sources to validate the kaggle dataset. While we were able to spot-check multiple values within the data, there did not seem to be an overall alternative to the set as a whole.

We also spent a good deal of time looking at the BI software landscape, before settling on Tableau as the tool of choice. Not only did it repeatedly come up as one of the more intuitive tools available, they also offered a student license. Additionally Python options were explored and we decided to go with Seaborn and plotly.py as our libraries to use. It is very likely that our final visualizations will be based on Seaborn, but we wanted to include plotly.py in our exploration as one focuses on statistical 2D visualizations, but the other can do 3D and beyond (in case that is needed).

We think we're about where we expected at this midpoint. After some scope massaging in our midpoint meeting, we now believe we are equipped to finish out this project. We only ended up needing to do a small pivot and now can lean into the visualizations themselves.