Project Report

December 14, 2024

1 Team 5 Project Report

Byrne,Jelovich,Shinn,Whitney,Youckton

1.1 Video Game Sales Analysis

DATA SOURCE **https://www.kaggle.com/datasets/gregorut/videogamesales?resource=download**
By December 13 you are required to submit your final project report as a .pdf file and the Jupyter notebook file (used for the coding in Python). The report should thoroughly describe your experiences and analysis, as though you were reporting the results of this project to a manager or supervisor. The following information should be contained in the report, as well as appropriate figures from the analysis incorporated into the text of your markdown file:

PROJECT PROPOSAL: For our project, we will be working with the video game sales dataset to examine video game genre popularity trends by decade, as well as how that popularity varies by region. The dataset we will be utilizing consists of 11 columns and 16,600 rows. The rows each represent a video game with sales greater than 100,000 copies and the columns consist of ranking of overall sales, the name of the video game, release platform, release year, game genre, game publisher, 4 columns representing sales in different regions of the world, and total global sales. Our intended research questions are as follows: How has video game genre popularity evolved by decade since 1980? How does genre popularity vary between different regions? We plan to import the dataset into Jupyter notebook and utilize the pandas, matplotlib, numpy and seaborn libraries to create histograms, box plots, and other useful visualizations to help observe trends and display our findings. Upon initially looking through the data, we noted that it is comprehensive in the sheer number of video games the data was collected on, the diversity of genres included, as well as the timespan of release dates.

1.2 • A description of the dataset and the reasons why you selected it for the final project.

1.2.1 *Description of dataset*

The video game sales dataset consists of 11 columns and over 16,500 rows. Each row represents a video game with sales greater than 100,000 copies and the columns consist of ranking of overall sales figures, game name, release platform, release year, game genre, game publisher, 4 columns representing sales in different regions of the world (North America, Europe, Japan, and 'Other' regions), and total global sales.

1.2.2 *Reason for selection*

Relevance: This dataset is highly relevant for analyzing trends in the video game industry, understanding the popularity of different genres and platforms, and identifying key factors that contribute to high sales.

Availability: The dataset is readily available on Kaggle and is well-documented, making it easy to use for analysis.

Interest: The video game industry is a dynamic and rapidly evolving field, making it an interesting subject for analysis. Additionally, the dataset's comprehensive nature allows for a wide range of analyses, from sales trends to market segmentation.

1.3 • A summary of any processing problems you identified with the data and key steps you took to overcome those issues.

Problems: Not all lines have year, so cannot group by decade - Clean data, remove lines altogether Year type is float, so cannot group categorically - Convert Year to String then extract first three characters to determine decade

Problem 1: Missing data for certain games (MAR-Missing at Random).

Solution: Our team used simple .isna and .sum functions to estimate missing values based on available data. For example, if Year data for a particular line was missing, we identified and then removed those lines before proceeding with the analysis. Result: "There are 271 missing values in the Year column."

Problem 2: In our analysis we wanted to explore patterns by decade, however, individual years were avialable in the dataset.

Solution: By dividing the year by 10 and then multiplying by 10, we were able to add an additional column for the decade of release.

Problem 3: Outliers in data that could skew the analysis.

Solution: Example 1; only 1 result in year 2020, this would account for an entirely new decade. We decided to remove this result with .dropna, so our analysis would be more relevant. Example 2; We determined there was an uneven amount of unique genres per decade through a .len() function. Since 1 of 4 decades were missing 1 genre, we considered this as acceptable to include and move forward with analysis.

1.4 • A statement on your 'Big Question' and the explanation why the dataset you chose is a good choice to answer it.

Q: How has video game genre popularity evolved by decade since 1980?

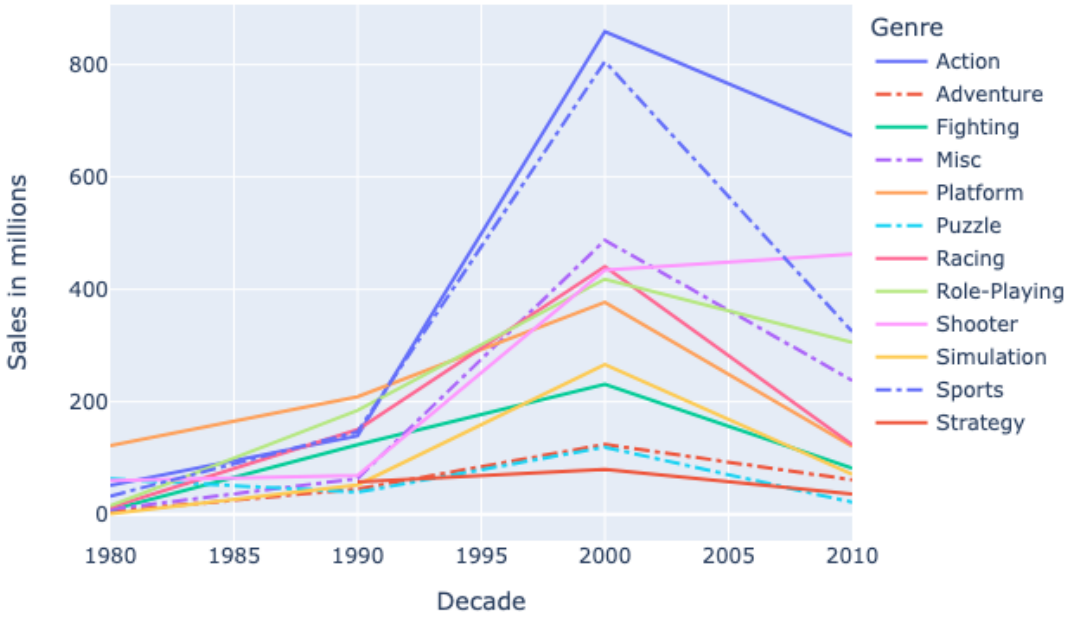
Q: Shooter genre trending higher in 2010s; how does it vary between different regions?

1.5 • A detailed description of the results of your exploratory analysis and what preliminary conclusions you were able to draw based on this analysis.

We set out to form an analysis on regional video game sales from 1980's through 2010's. The genres followed similar trends in growth and peaks; however, in the 2010's, one genre trended higher vs. lower. Although the "Action" genre was still the highest grossing genre, the "Shooter" genre was the only genre consistently increasing in sales. When further analyzing the "Shooter" genre, we found that North America and Europe were the largest consumers, with North America's sales approximately double that of Europe. Japan was the region for lowest sales in "Shooter" genre

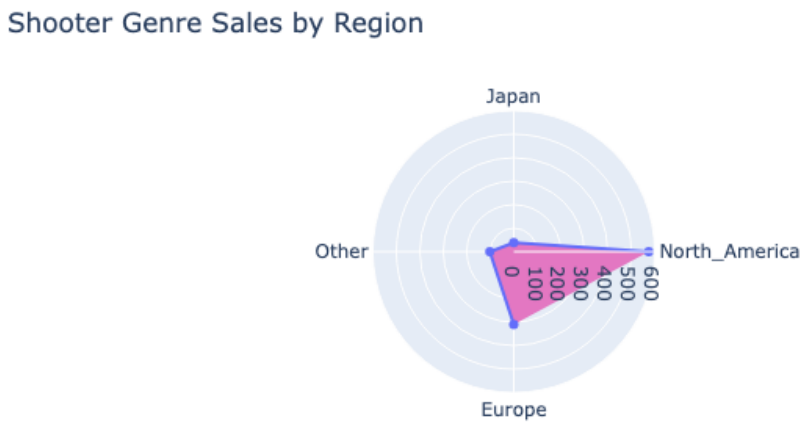
[8]:
```
from IPython import display
display.Image("./newplot (1).png")
```

[8]:



[9]:
```
from IPython import display
display.Image("./newplot.png")
```

[9]:



1.6 • The overview of the methodology for your analysis of the big question (the pros and cons of that method(s), any alternative methods you considered).

The strategy we used in our exploratory analysis, was to produce several different models for interpreting the data. We started by creating some simple tables to group data or to examine sets of columns side-by-side. Once we had a clear view of how to proceed, we started producing simple charts for visualizing the data. By producing so many tables and graphs, we determined that several offered little to no insignt. With twelve genres, accross four regions and nearly 40 years, grouping by decade was a good way to see trends. In some cases, graphs that used the 'Year' column produced better looking results.

1.7 • Your final conclusions based on your analysis.

The conclusion we were able to draw off of this analysis was that, in the 1980's, arcade like games were more popular, while in the 1990's they improved console systems and made gaming from home much better. What also made the popularity differ was that in the future decades there seems to have been a bigger play base which brought online gaming with friends in the 2000's and 2010's. The conclusion with the "Shooter" genre was that it may have increased in the 2010s but the popularity in different regions was varied, it was popular in the North America and Europe regions, but not in the Japan region it would show limited growth.

1.8 • Prospective of the potential future research in this area: a short sketch of any additional analyses that you would have liked to carry out and any additional data that would have been needed in order to extend your analysis.

One area of interest for the future would have to involve video game popularity for the whole decade of 2020. Described previously, each decades most popular genre determined by sales has changed. It would be fascinating to see if this trend continues into the 2020 decade. The analysis to determine popularity would involve another pivot table, as well as a line or bar chart.

We could also use the data for the first 5 years of the 2020's and perform a regression model to determine the trend for the second half of the decade.

[ ]: