PROPOSAL

Video Game Sales Exploratory Data Analysis

By
Reem Alsuqaih
Njood AL Qahtani

Abstract:

Who doesn't like video games? Video games have become a favorite pastime for most of us. Video games are a billion-dollar business and have been for many years. Now imagine being able to analyze the sales trends for the popular Nintendo games and many more such platforms. How cool would it be!

Design

This project is one of the T5 Data Science Boot Camp requirements. Data provided by Kaggle & data. world have been used in this project. In this module, we will be laying the foundation for our analysis by processing and exploring a large amount of data on video game sales. The dataset contains information regarding the sales of video games across various regions like North America, Europe, Japan and globally, while also giving information regarding the Names, Publishers and Platforms. This dataset has been made available thanks to Kaggle which is the home for many such datasets and competitions.
Get the data here.

You are a data scientist in the gaming industry and want to analyze the sales pattern for different features. You want to investigate different features and approach the game producer with a thorough report. You want to find the answers to the following questions:

1. Which region has performed the best in terms of sales?
2. What are the top 10 games currently making the most sales globally?
3. What are the top games for different regions?
4. Are there any games with release year older than 2000 that are still making high sales? What are they?
5. What are the top gaming genres that are making high sales?
6. Does the publisher have any impact on the regional sales?
7. Is there any region that has out-performed global average sales?

Learning Objectives

After this module, we should be able to do the following:

1- Use panda's library to analyze different features of the dataset, which includes,

2- Read the dataset.

3- Use functions like df. query and df. group by.

4- Analyze the sales for different regions with respect to various features.

5- Use plotly library to visualize the given results

6- Plot graphs like bar graphs and pie charts

Understanding the dataset

This dataset contains a list of video games with sales greater than 100,000 copies. In this data set there are 11 columns. Their names and data types as follows:

Fields include

- Rank - Ranking of overall sales
- Name - The games name
- Platform - Platform of the games release (i.e. PC, PS4, etc.)
- Year - Year of the game's release
- Genre - Genre of the game
- Publisher - Publisher of the game
- NA_Sales - Sales in North America (in millions)
- EU_Sales - Sales in Europe (in millions)
- JP_Sales - Sales in Japan (in millions)
- Other_Sales - Sales in the rest of the world (in millions)
- Global_Sales - Total worldwide sales.

Algorithms

- Cleaning the data and remove null values
- Transform the text data in sales to numbers.

Tools

- Pandas for data manipulation
- Scikit-learn for modeling
- re for clean data
- Matplotlib for plotting
- Plotty for visualizations
- World cloud

Communication

- The slides will be provided here, Feel free to any pull requests.