Project Write Up : Project 4, Group 11
Video Game Sales

Tyler Beringer, Jason Johnson, Brian Marowsky, Noah Stevens

Introduction and Purpose

For this project, we wanted to explore trends in sales reviews in regards to video game sales. Through the use of multiple data sets on video game sales and reviews ranging from 1980 to 2016 which contained trends of market sales per million units, as well as critic user reviews. We were able to evaluate trends throughout the years and find correlations and relationships between different values. With this data in hand, we were able to use our machine learning model to predict sales based on a number of inputs including platform and genre, better predicting which gaming platform and genre to look out for making future investments/interests.

The various aesthetics of our website and presentation reflect colors of a wide variety of gaming devices and platforms that can be found throughout our various data sets. With our data set being a high view composed of different platforms and models and not specific to one type or brand of gaming devices/genres,it was decided to incorporate mostly stock-type photos throughout our presentation and web application to better emphasize on the high level view that our data sets were presenting. Attached are links to both kaggle datasets we have used in the creation of our slide deck and web application. ** Insert here**

Data Cleaning

With using 2 data sets, part of our initial data cleaning process was to combine both data sets into 1 table structure in order to evaluate both sets together. Once combined, we evaluated the data types to make sure all categories were listed as their correct data type. Next step was to replace instances of ‘tbd’(to be determined) and replace them with NaN (Not a number). This in return allowed us to convert the column from a string to a float. With comparing both the df_reviews and df_games , each one had a

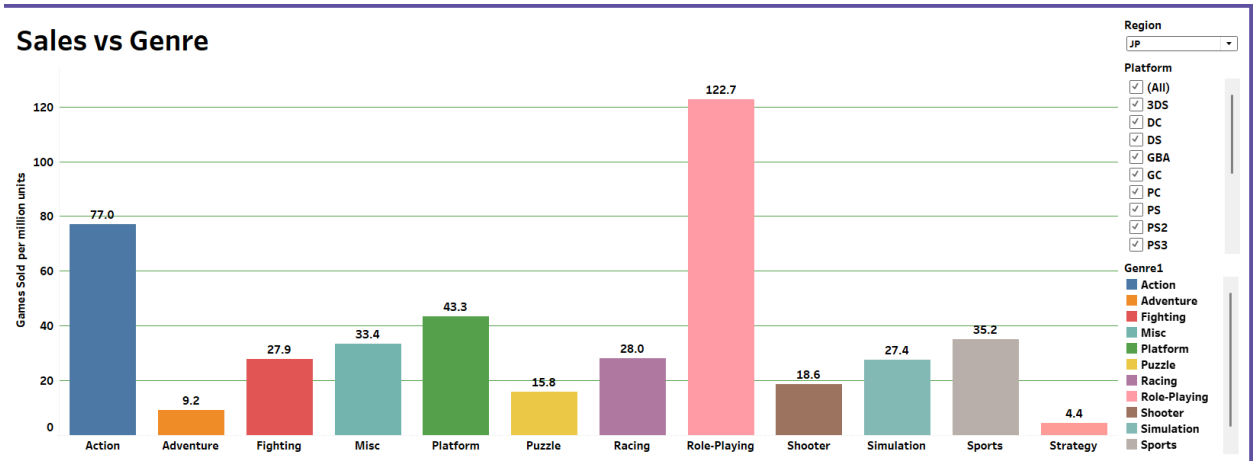
“platform” column. While both columns shared a lot of the same video game consoles. The df_game’s column had their abbreviation instead of the full name like in df_reviews. To fix this, we converted the full names to their respective abbreviations which allowed for a cohesive, uniformed platform column. With this done, we were finally able to merge both sets of data on their shared columns of “name”,”platform” as shown below and then saved as our clean data set “game_final_clean”

```
Data columns (total 15 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   name          16441 non-null  object
 1   platform      16441 non-null  object
 2   Year          16441 non-null  int64
 3   Genre         16441 non-null  object
 4   Publisher     16441 non-null  object
 5   NA            16441 non-null  float64
 6   EU            16441 non-null  float64
 7   JP            16441 non-null  float64
 8   Other         16441 non-null  float64
 9   Global        16441 non-null  float64
 10  Critic_Score  8299 non-null   float64
 11  Critic_Count  8007 non-null   float64
 12  User_Score    8863 non-null   float64
 13  Developer     9929 non-null   object
 14  Rating        9792 non-null   object
dtypes: float64(8), int64(1), object(6)
memory usage: 2.0+ MB
```

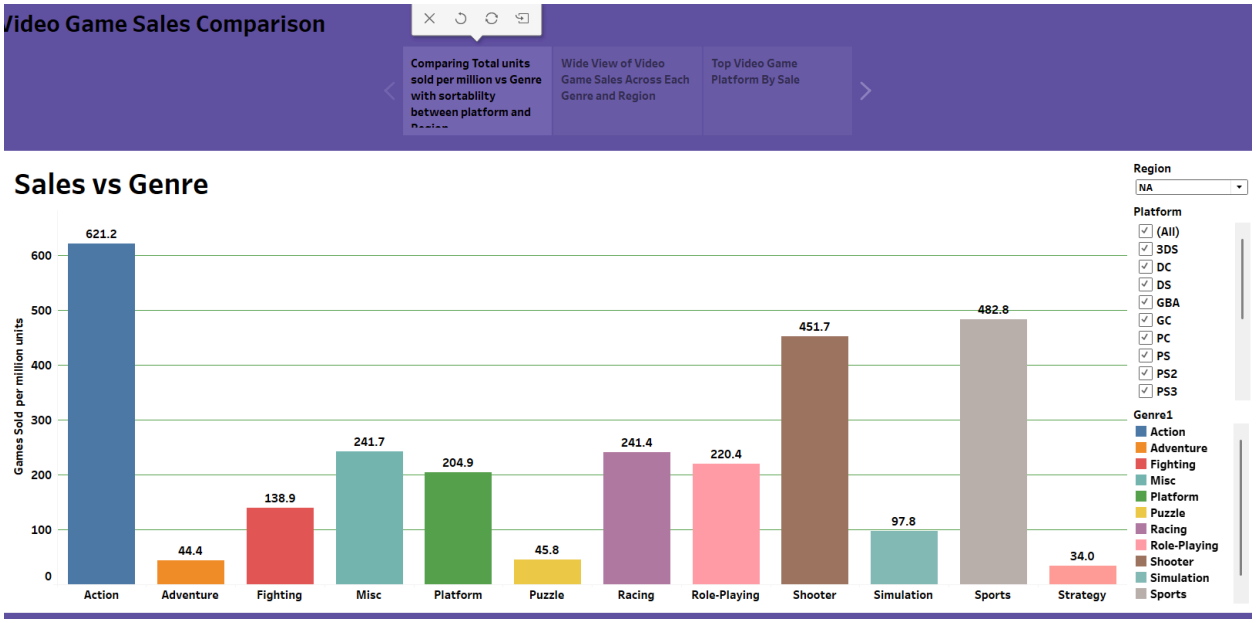
Tableau- Sales Data

We divided our task between the comparison of User scores(Critics) and Sales data to categories such as platform, Genre. Our data from kaggle shared that the “global sales” was a sum of the 3 different regions from around the world , EU,NA,JP. This provided an opportunity to compare the different regions in regard to categories like Genre and platform. For Example, it was found that in Japan (JP) Role play games had the most units sold in millions compared to North America (NA) which preferred action games over roleplay games.

Japan



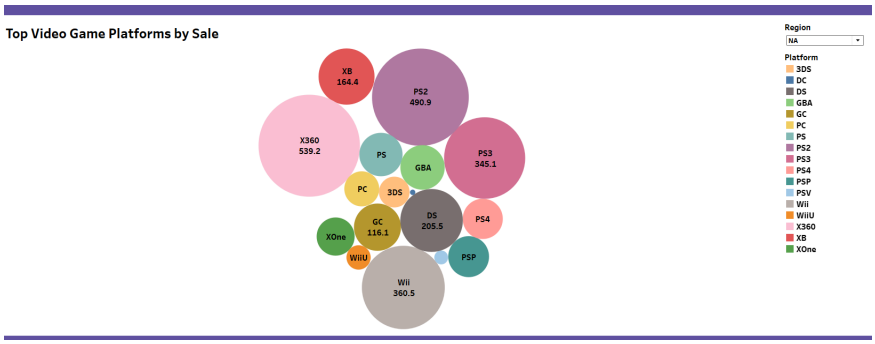
North America



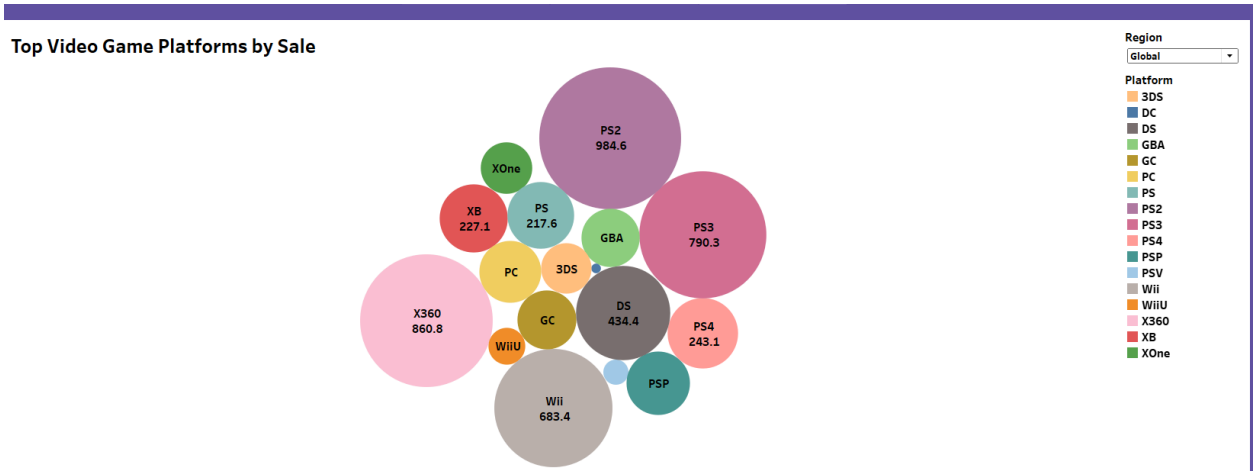
When looking at which video game platform had the most sales. Globally, PS2 was the winner overall while in North America, xbox 360 was on top.

North America

Platform



Global Platform



This information shared above would be most helpful in determining future prospects with sales not only globally but in their respective regional markets.

Tableau- Critic Score
**Insert here **

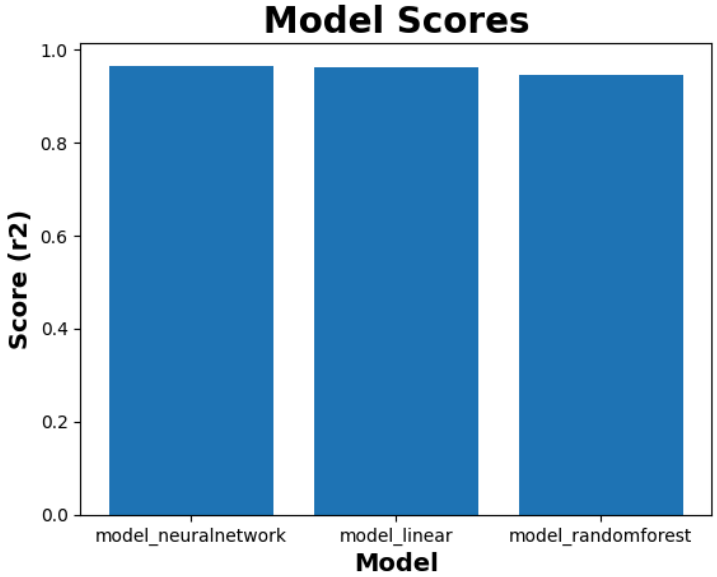
Both pages were added to the application by using Java script attachments of the URL from the Tableau public website where both models were published.
https://public.tableau.com/app/profile/jason.johnson3075/viz/VideoGameScores_17440655448690/Story1
https://public.tableau.com/app/profile/brian.marowsky/viz/Game_Sales_17440653955190/Story1

Machine Learning model page

With our Data cleaned and prepared into one data set , we were able to begin work on our machine learning model. We ended up testing three models—Neural Network, Linear Regression, and Random Forest—to predict **Global Sales**. Performance was measured using the R^2 score, where **1 means perfect accuracy**. All models comfortably beat the goal of **0.80**, showing they're strong predictors.

- **Neural Network**: R^2 = **0.96**
- **Linear Regression**: R^2 = **0.96**
- **Random Forest**: R^2 = **0.94**

The **Neural Network and Linear models performed slightly better** than the Random Forest, but overall, all three did a solid job predicting sales.



About us/ Work cited page /Report

Our about us page of our application included the information and picture of each group member who worked on the project. The Work Cited page

includes any sources we used to create our project. Report page includes an embedded copy of this report.

Biases, limitations and Conclusion

When it comes to bias some limitations some aspects we notice when creating our data set were as follows the data sets very limited scope of the market in general the main attribute of the measure of sales being in per million units sold rather than actual dollar amounts of video games being sold another limitation is with the advent of modern day subscription platforms for video games Microsoft's Xbox game pass where measurement of success is based on the number of downloads rather than a monetary cost this small being only a portion of market shares does play a role in the overall market of video games which is something to do need to be considered in the future data sets being created on these matters along this similar view of content being viewed and sold another aspect that was noticed in comparing video game sales by title is was the categorization of DLC or downloadable content which are additional software updates to games already released usually available for an additional charge. A form of sales would be accurate to describe DLC content; it would be important to differentiate the content outside of the category of standalone video games as most DLC content is an addition to a game already released. Some future work in projects that could be undertaken with further data collection could include training a machine learning model that predicts global sales without other sales data from regions taking into account as mentioned our data sets had a sum of the three regions of our market to obtain the global sales index so removing the regional variables would create a more accurate and wider representation of the data. With the inclusion of the above suggestions as well as more thorough global ratings price points for sales the future of our model and

application would only become better trained better accuracy and the better predictor of future development for video games around the world.

Work Cited

Images

-https://www.pexels.com/search/video%20game/

- metacritic logo, wink emoji, earth, sales

Data sets-https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings

https://www.kaggle.com/datasets/deepcontractor/top-video-games-19952021-metacritic

https://www.kaggle.com/code/maxkliment/video-games-predicting-global-sales

https://github.com/martabaker/ds_project_4_group_02