Video Game Data Analysis

By Jasmine S. Gutierrez, William Ocampo, Stephen Lambert

Problem statement. What problem did you try to solve?

- We are trying to determine what predicts video game global sales.
- We are also observing patterns to see what makes a video game the most in sales or just in general to see the most popular genre of video games, seasons to buy games, and the consoles that are the most popular.

Data sources. Which dataset did you use to solve the problem? Specify the variables in the dataset and the size of the dataset.

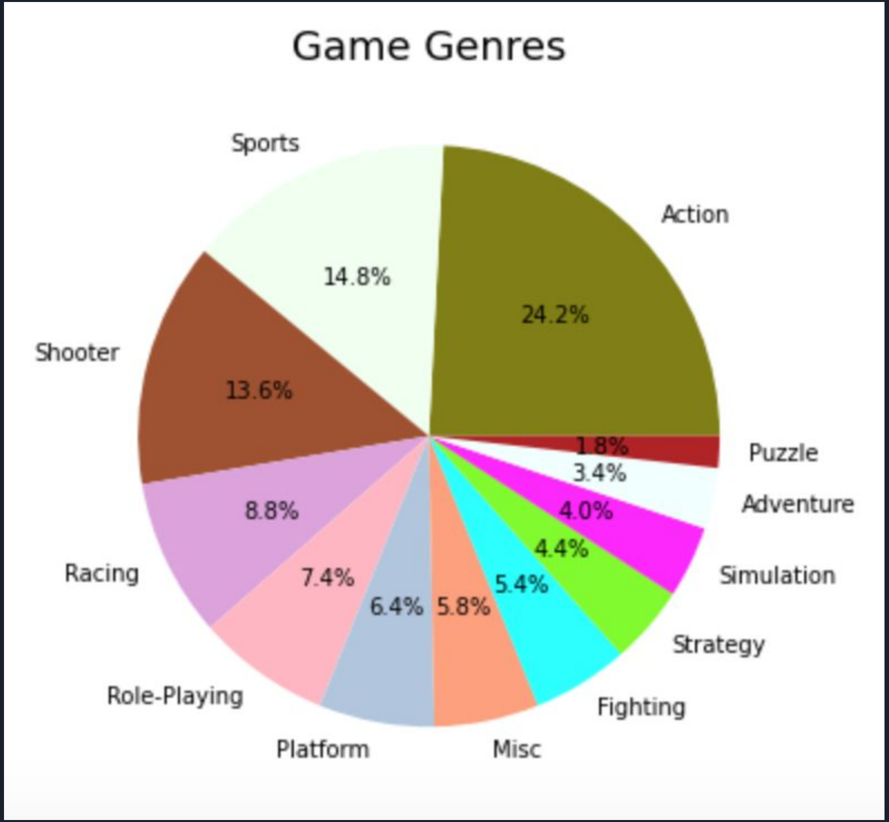
- The total size amount for the data set is 5733.
- We used 2 different data sets one from
  - Metacritic
    - Video game sales dataset containing info of video games as well
  - Kaggle dataset
    - Review rating of video games
  - The variables are called Rank, Name, Platform, Year, Genre, Publisher, NA_Sales,

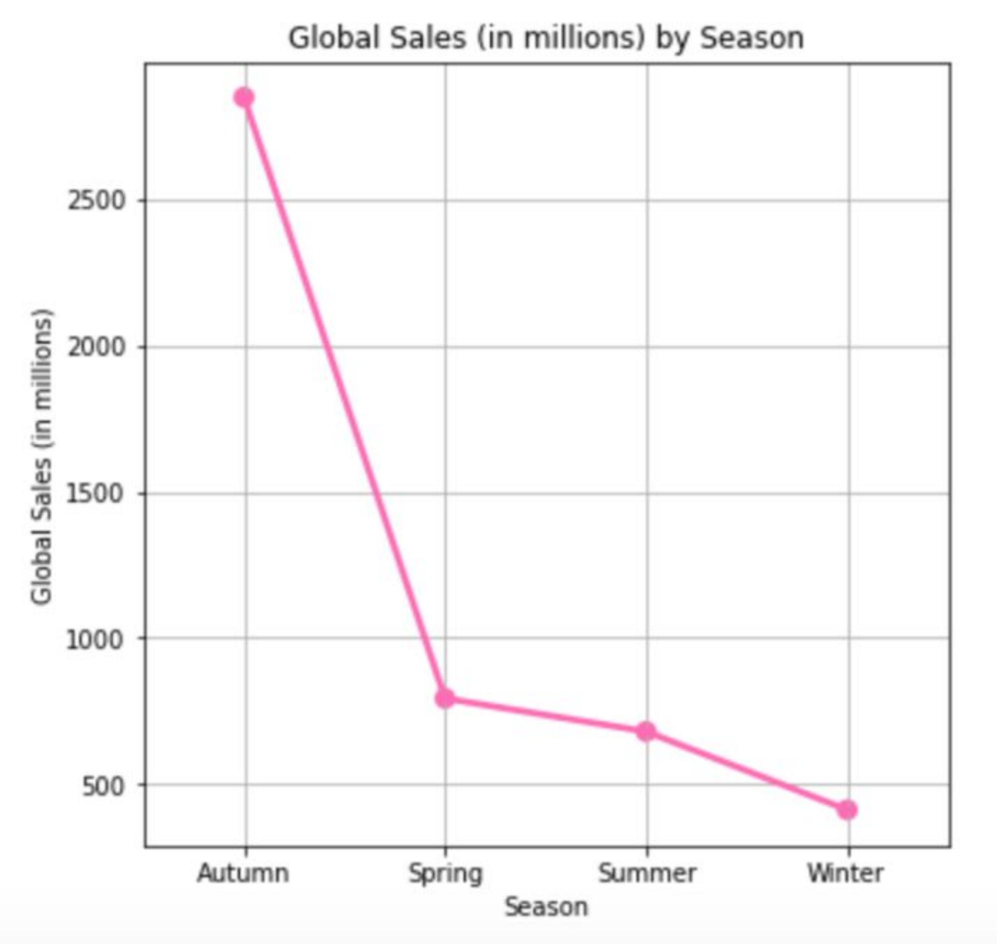
    EU_Sales, JP_Sales, Other Sales, and Global_Sales

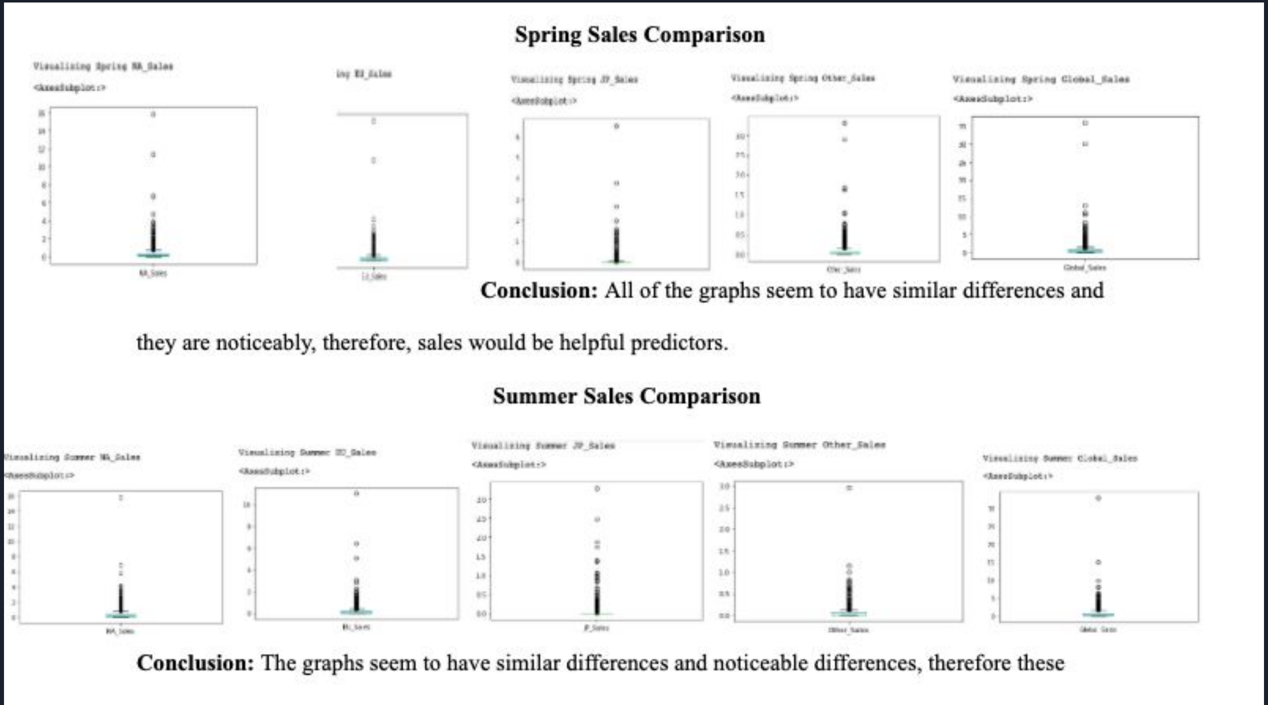
Brief description of data science solution. Which models did you build to solve the problem? Explain how you prepared the data and specify the regression/classification/clustering techniques that you used to build the models.

- For the data exploration portion of this assignment, our group explored our datasets by calculating the mean of each of the seasons (spring, summer, autumn, winter) for the variables NA_Sales, EU_Sales, JP_Sales, Other_Sales, and Global Sales.
- Then we performed a hypothesis test, using f_oneway ANOVA, this allowed us to determine if multiple variables have the same distributions, or not. Lastly, we plotted our results to explore the outliers, and determine the most valuable variables.
- For linear analysis, we tried to determine what helps drive global sales. We came to the conclusion that NA_Sales, EU_Sales, JP_Sales, Other_Sales, Platform_PS2, Genre_Simulation and Genre_Action help drive game sales.
- For clustering, we tried to determine which categories are the best to observe and test. We used single linkage clustering and Hierarchical, K-means, and DBSCAN for Platforms, Genre, and Seasons.

Results. What were the results of your models? Use tables and figures to visualize your results.




Global Sales (in millions) by Season
Global Sales (in millions)
2500
2000
1500
1000
500
Autumn
Spring
Summer
Winter
Season



Spring Sales Comparison

```
Visualizing Spring NA_Sales
<AxesSubplot:>
```

```
ing EU_Sales
```

```
Visualizing Spring JP_Sales
<AxesSubplot:>
```

```
Visualizing Spring Other_Sales
<AxesSubplot:>
```

```
Visualizing Spring Global_Sales
<AxesSubplot:>
```

Conclusion: All of the graphs seem to have similar differences and they are noticeably, therefore, sales would be helpful predictors.

Summer Sales Comparison

```
Visualizing Summer NA_Sales
<AxesSubplot:>
```

```
Visualizing Summer EU_Sales
<AxesSubplot:>
```

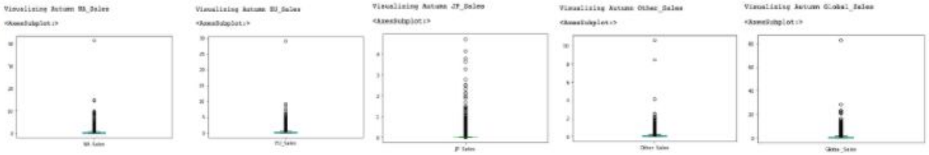
```
Visualizing Summer JP_Sales
<AxesSubplot:>
```

```
Visualizing Summer Other_Sales
<AxesSubplot:>
```

```
Visualizing Summer Global_Sales
<AxesSubplot:>
```

Conclusion: The graphs seem to have similar differences and noticeable differences, therefore these

Autumn Sales Comparison



```
Visualizing Autumn NA_Sales
<AxesSubplot:>
```

```
Visualizing Autumn EU_Sales
<AxesSubplot:>
```

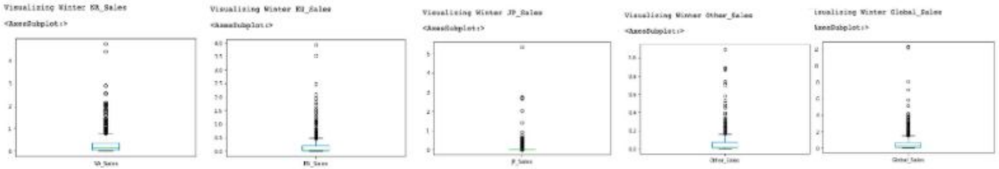
```
Visualizing Autumn JP_Sales
<AxesSubplot:>
```

```
Visualizing Autumn Other_Sales
<AxesSubplot:>
```

```
Visualizing Autumn Global_Sales
<AxesSubplot:>
```

Conclusion: The graphs have clear noticeable differences, therefore, the sales variables continue to be good predictors.

Winter Sales Comparison



```
Visualizing Winter NA_Sales
<AxesSubplot:>
```

```
Visualizing Winter EU_Sales
<AxesSubplot:>
```

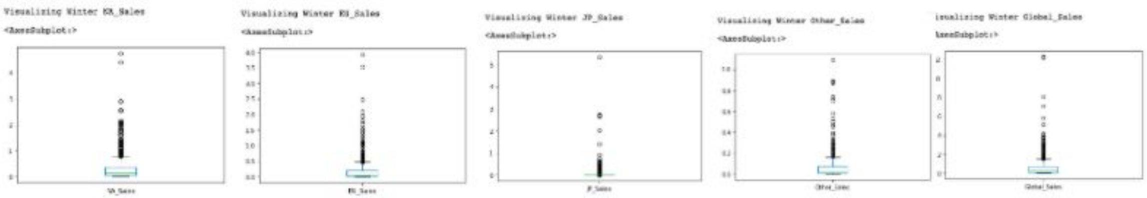
```
Visualizing Winter JP_Sales
<AxesSubplot:>
```

```
Visualizing Winter Other_Sales
<AxesSubplot:>
```

```
isualizing Winter Global_Sales
AxesSubplot:>
```

Conclusion: The variables continue to have noticeable differences, therefore, we conclude that the sales variables are good predictors.

Winter Sales Comparison



```
Visualizing Winter NA_Sales
<AxesSubplot:>
```

```
Visualizing Winter EU_Sales
<AxesSubplot:>
```

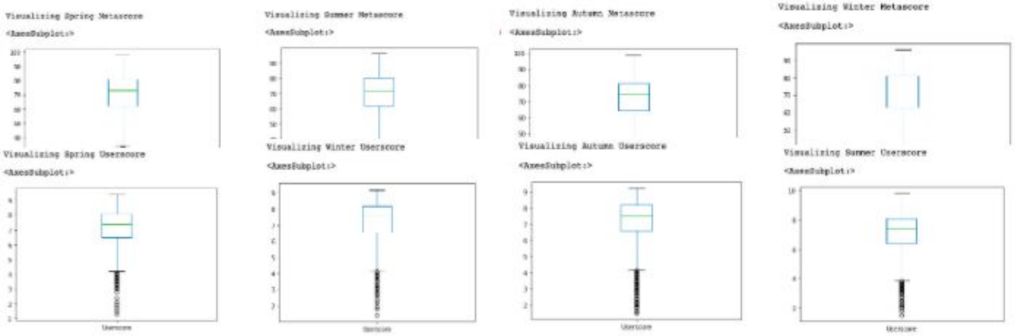
```
Visualizing Winter JP_Sales
<AxesSubplot:>
```

```
Visualizing Winter Other_Sales
<AxesSubplot:>
```

```
isualizing Winter Global_Sales
AxesSubplot:>
```

Conclusion: The variables continue to have noticeable differences, therefore, we conclude that the sales variables are good predictors.

Metascores and User Score Comparisons



```
Visualizing Spring Metascore
<AxesSubplot:>
```

```
Visualizing Summer Metascore
<AxesSubplot:>
```

```
Visualizing Autumn Metascore
<AxesSubplot:>
```

```
Visualizing Winter Metascore
<AxesSubplot:>
```

```
Visualizing Spring Userscore
<AxesSubplot:>
```

```
Visualizing Winter Userscore
<AxesSubplot:>
```

```
Visualizing Autumn Userscore
<AxesSubplot:>
```

```
Visualizing Summer Userscore
<AxesSubplot:>
```

Conclusion: The metascore and user score outlier differences are barely noticeable, therefore, metascore and user score would not be good predictors.

Conclusions. What conclusions can you make from your results?

- The most important variables for data analysis are NA_Sales, EU_Sales, JP_Sales, Other_Sales, Platform_PS2, Genre_Action, Genre_Simulation.
- Metascores and user scores would not be good predictors.
- The results from clustering showed that Platform and Genre produces the best results, which makes sense because certain platforms and genres tend to be more popular than others.