Milestone 3 Report

Video Game Sales

Daniel Sternberg, ds44

Anbar Saleem, aas27

Raphael Roxas, rcr24

Nishant Joshi, nj26

Large dataset of video games released

Between 1980-2016

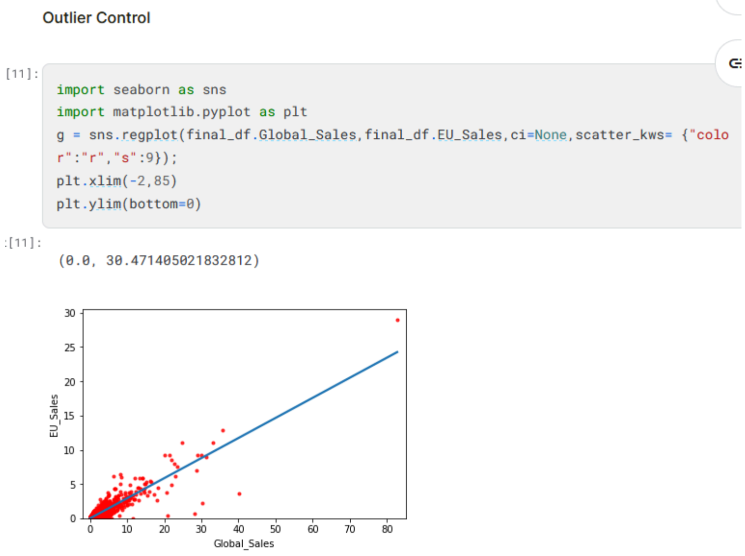
https://www.kaggle.com/datasets/gregorut/videogamesales/data

Based on the data set we have decided to work with, we plan on using supervised learning to predict the global sales of a video game based on their individual sales for their sales in North America, Europe, Japan, and Other (as labeled in the dataset). After looking over some exploratory data analysis on the data set, we found a few that gave us a stronger understanding of what this data set includes and how we can use it to create a model that would demonstrate our objective.

Linear Regression approach by Ömer Şenol:

https://www.kaggle.com/code/omersenol/linear-regression

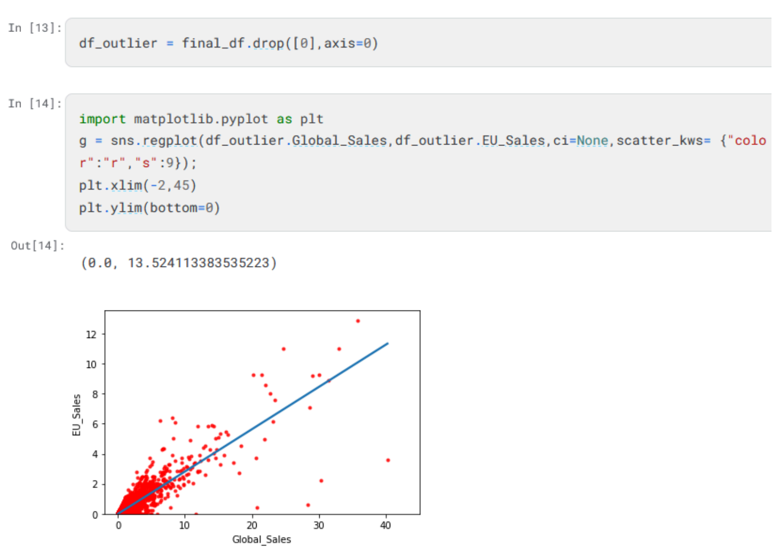
The author has his code commented to show his process through which he preprocesses the data to achieve a similar objective as ours: predict global sales given EU sales. Omer's code also showed visualizations of the graphs through different steps, making it clear to see what each line of code does when looking at the dataset. To achieve this he uses a linear regression approach using sklearn from the scikit learn module in python. As for dealing with processing data, Omer converts categorical variables to numerical ones and removes outliers/unnecessary data to ensure the graphs are more accurate.



Outlier Control

```
[11]:
import seaborn as sns
import matplotlib.pyplot as plt
g = sns.regplot(final_df.Global_Sales,final_df.EU_Sales,ci=None,scatter_kws= {"colo
r":"r","s":9});
plt.xlim(-2,85)
plt.ylim(bottom=0)
```

```
:[11]:
(0.0, 30.471405021832812)
```



```
In [13]:
df_outlier = final_df.drop([0],axis=0)
```

```
In [14]:
import matplotlib.pyplot as plt
g = sns.regplot(df_outlier.Global_Sales,df_outlier.EU_Sales,ci=None,scatter_kws= {"colo
r":"r","s":9});
plt.xlim(-2,45)
plt.ylim(bottom=0)
```

```
Out[14]:
(0.0, 13.524113383535223)
```

Analysis:
The positive slope of the regression line suggests a positive correlation between Global Sales and EU Sales, meaning as Global Sales increases, EU Sales tend to increase as well. The data points are fairly scattered around the regression line, which indicates variability that is not captured by the linear model. This could mean there are other factors affecting EU Sales that are not accounted for by Global Sales alone, or that the relationship is not perfectly linear. That is why Omer did a multiple regression model afterwards

What We Would Like To Extend To This Model:

- We would like to try…
    - Comparing the performance of simple linear regression models and multiple regression models based on evaluation metrics. And perform cross-validation to ensure the robustness of the models.
    - Considering polynomial regression and methods like Random Forests and Gradient Boosting. These techniques offer a path to better accommodate the non-linear interactions between various predictors and global sales.
    - Applying feature scaling using min-max normalization to ensure that all features contribute equally to the model and prevent features with larger magnitudes from dominating the model's training process.
    - Conducting exploratory data analysis and visualizing the relationships between global sales and individual regional sales (North America, Europe, Japan, Other) using scatter plots or correlation matrices to understand the data's distribution and identify any patterns.
    - Extending the model by integrating time series analysis, recognizing the potential impact of a game's release year on its sales performance. This extension acknowledges the evolving nature of the gaming industry and consumer preferences over time.