BDM2053 – Big Data Algorithms and Statistics

Video Games Trend Analysis (Recommendation System)

April 20, 2023

Submitted to :
Prof. Dhruwal Shah

Submitted by Group 12 :
Auradee Castro
Bhumika Rajendra Babu
Miraj Sinya
Olivia Deguit
Roger Mais

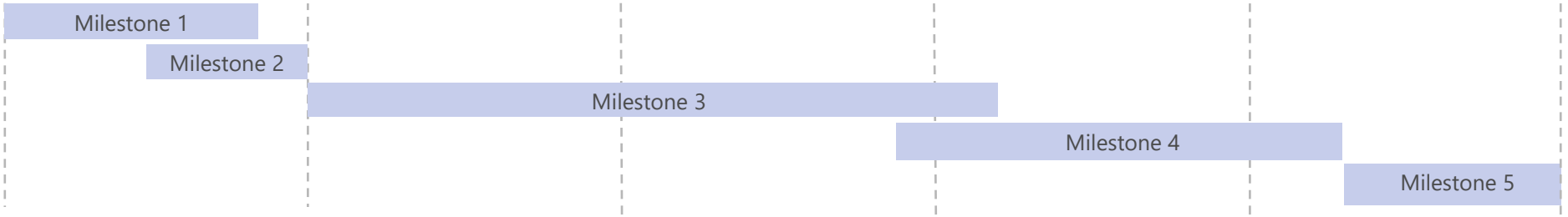
Case Background

- Client wants to venture into the Video Game Industry and has requested our assistance in understanding the market, and at the same time to develop a recommendation system for video game titles based on user preferences.
- Dataset containing information of video game such as sales (in units), genre, release year, platform, critic score and user scores, is accessible to the public on VGChartz, which is used for cross-validating the data collected from Kaggle.
- Deliverables to the client:
    a. **Recommendation Model** that provides suggestions for the video game titles to users based on the game title and platform *(optional)* as input
    b. **Market Analysis Report** that presents an analysis of the Video Game Industry trend using available data up to the year 2020
        - Regional Sales
        - Genre
        - Platform

Project Breakdown

Milestone 1 Project Details & Design	**Milestone 2** Data Acquisition	**Milestone 3** Data Analysis & Visualization	**Milestone 4** Model Creation & Evaluation	**Milestone 5** Project Presentation & Report
Activities				
• Identifying **Business Problem** • Breaking down business problem into **Key Questions** to be answered	• Sourcing relevant data to address the identified business problem • Cross-validating dataset • **Data Sources used:** kaggle VGChartz	• Exploratory Data Analysis • Evaluating Data Distribution • Data Preprocessing • Market Trend Analysis • **Tools used:** python	• Model development for video game recommendation • Assessment of user efficiency • **Tools used:** Power BI	• Preparation for final project report and presentation • **Tools used:** PDF

Timeline

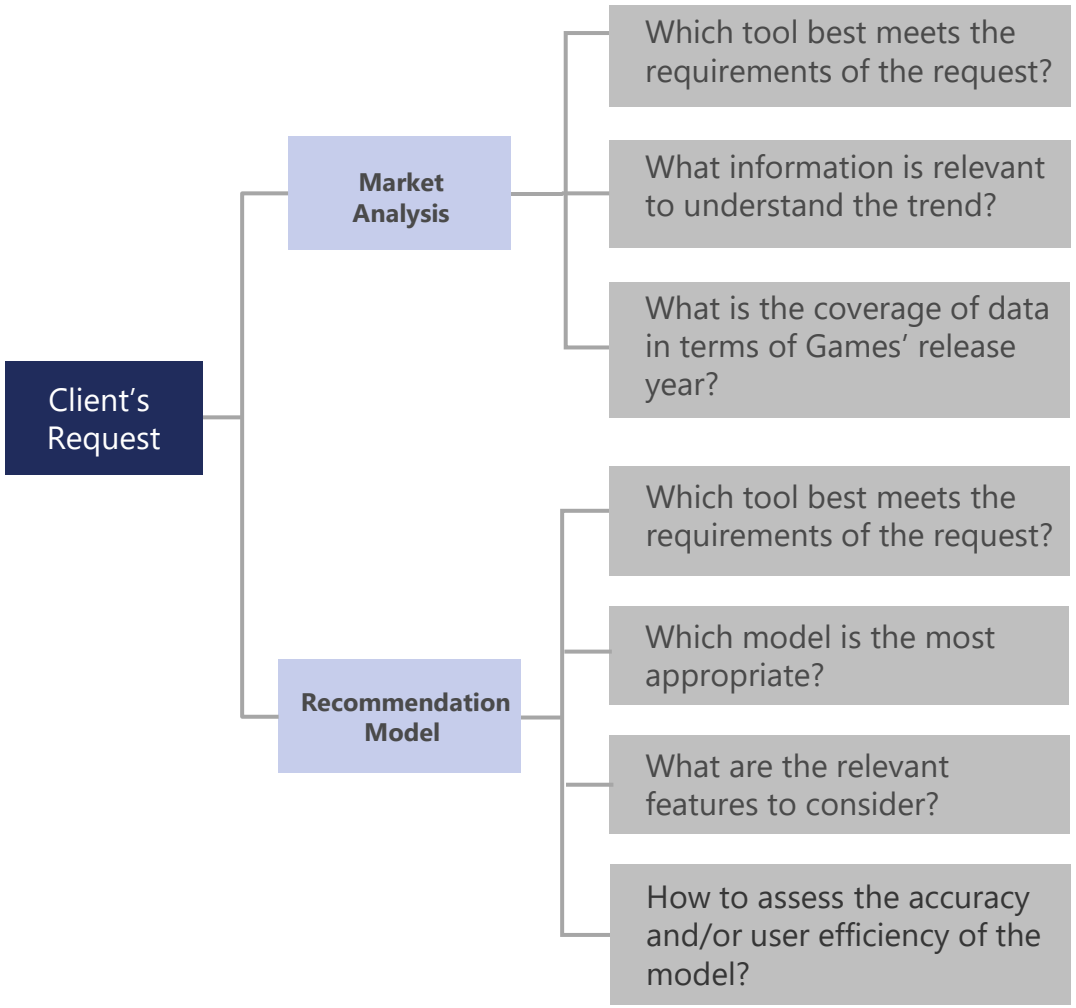


Project Schedule: March 16, 2023, to April 20, 2023

Project Details & Design

Breaking down the business problem into key questions

Key Questions



Complexities

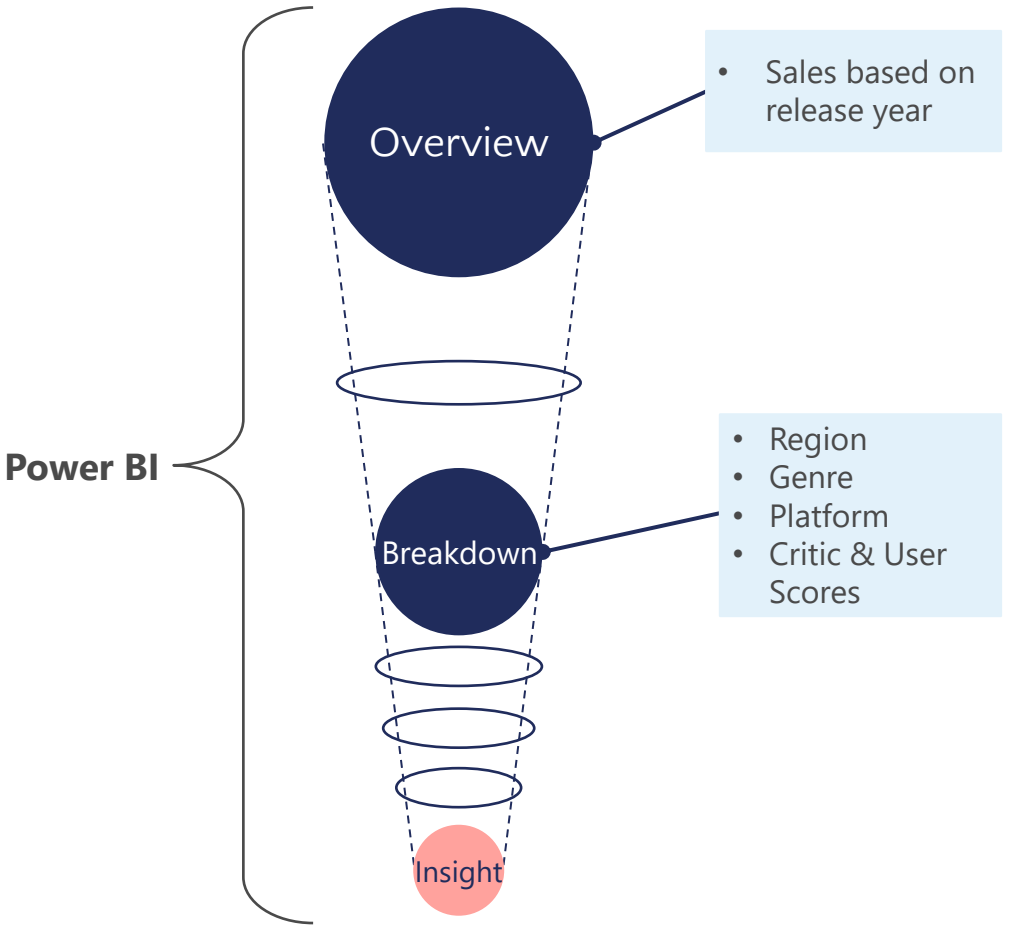
- Choosing the appropriate visualization tool, such as Excel, Python, or Power BI, that can handle the data size effectively
- Identifying relevant features to be taken into consideration to carry out Descriptive Analysis

- Deciding which model, whether it's supervised or unsupervised, generated the most relevant results
- Identifying relevant features to be taken into consideration as input to develop the model
- Determining the evaluation criteria for assessing the accuracy and/or efficiency of the model

Data Analysis & Visualization: Market Analysis (1/4)

Power BI as the main tool for the Descriptive Market Analysis

Workflow



Details

Purpose

- Power BI leveraged for data visualization through interactive dashboards to represent insights from the dataset
- Power BI is a better tool than Excel in this scenario given its capability to effectively manage large datasets

Complexity

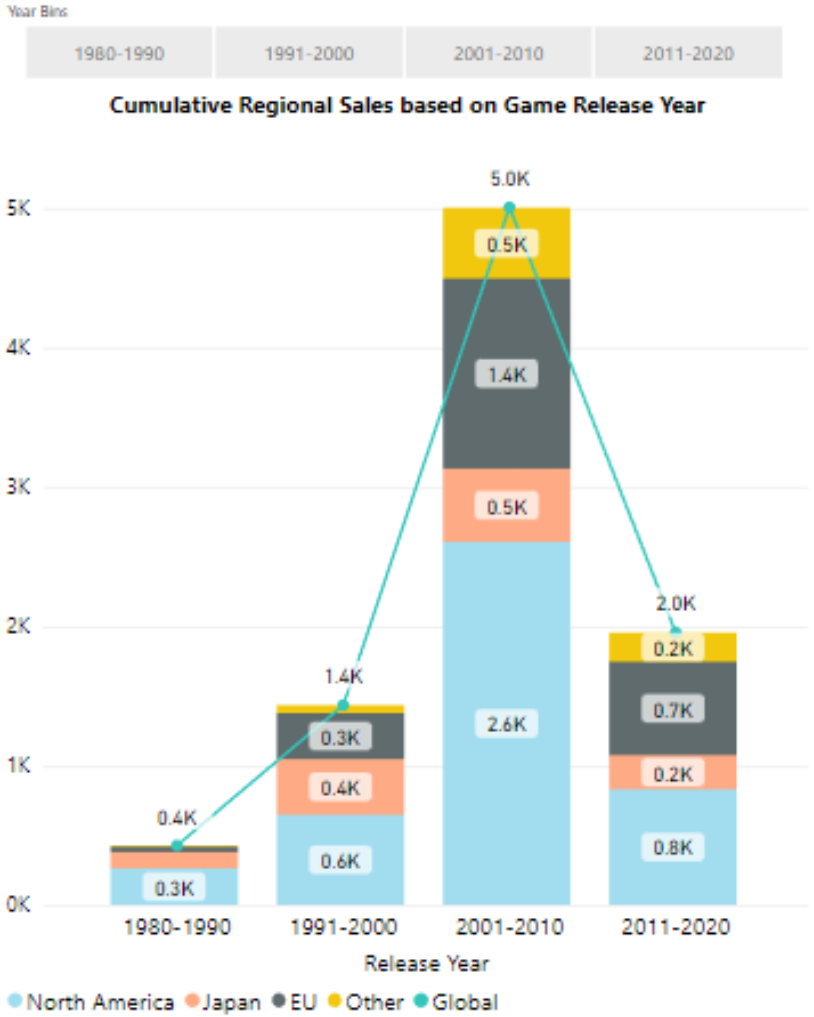
- Dataset included only the cumulative sales figures until the year 2020 creating a roadblock to analyze yearly sales growth trend

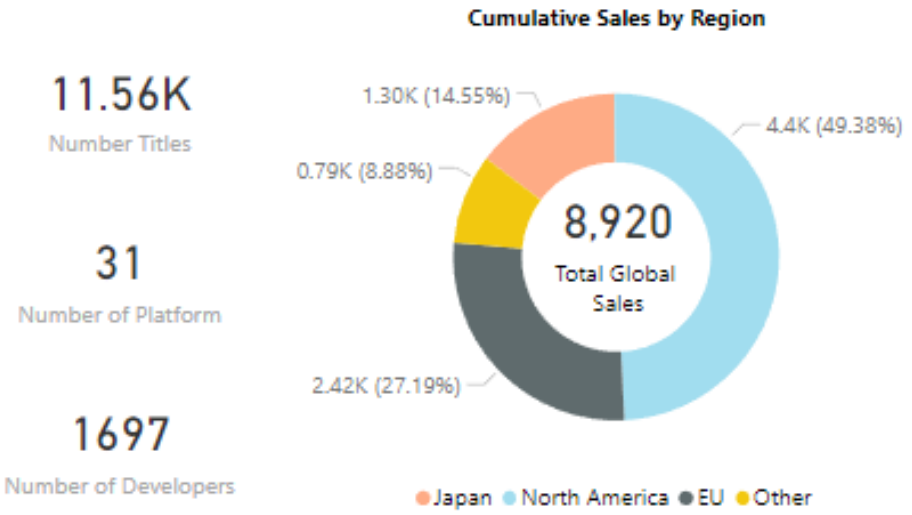
Resolution

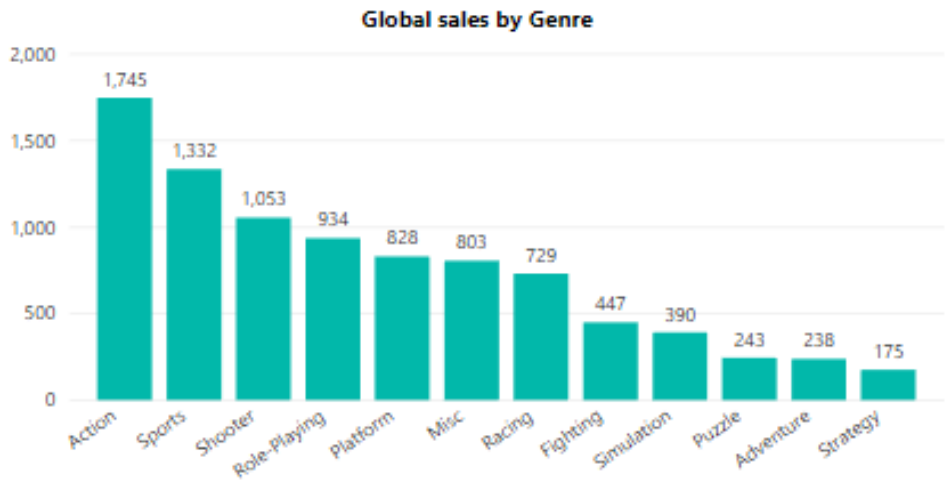
- Created 4 bins to group the data based on video game's release years spanning from 1980 to 2020
  - 1980 - 1990
  - 1991 - 2000
  - 2001 - 2010
  - 2011 - 2020

Data Analysis & Visualization: Market Analysis (2/4)

North America accounted for almost **half of the Global Sales** until 2020 with games released between 2001-2010 having the highest sales amounting to **2.6B units**





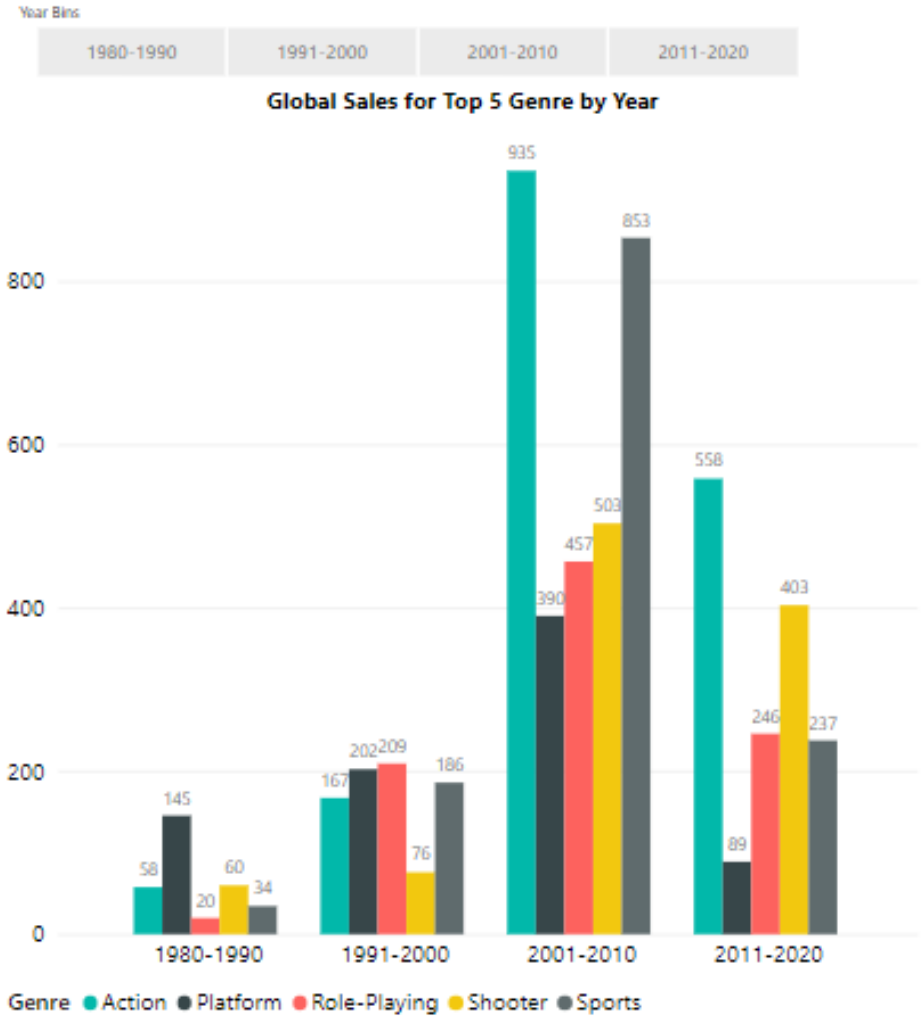


Filters

* Based on data till 2020
Sales Figures represented in Millions of Units (5.0K is equivalent to 5 Billion)
Others: Rest of the world, i.e., Africa, Asia excluding Japan, North America & Europe excluding the E.U.

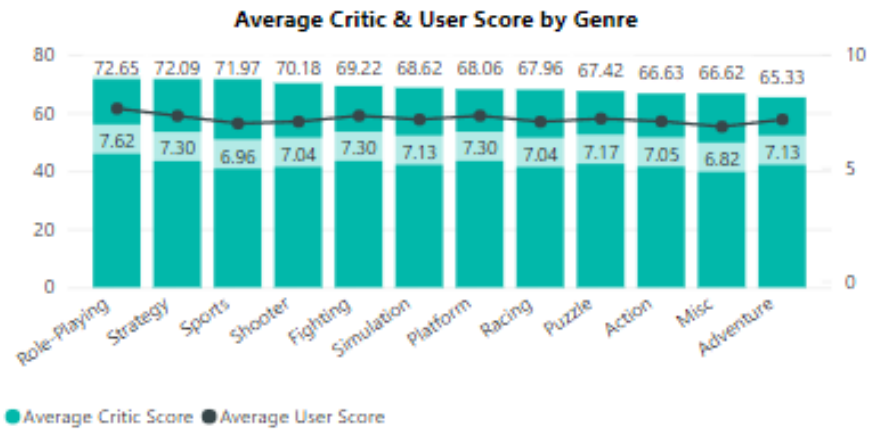
Data Analysis & Visualization: Market Analysis (3/4)

Transition from 2D graphics to 3D graphics in the late 90s resulted in **exponential growth** of the Action genre from **167M** in the 90s to **935M** in the 2000s



Count of Game Titles by Genre

Genre	1980-1990	1991-2000	2001-2010	2011-2020	Total
Action	68	179	1022	696	**1933**
Sports	22	349	854	160	**1366**
Misc	8	127	927	271	**1318**
Role-Playing	11	192	721	341	**1214**
Adventure	2	107	656	303	**1054**
Shooter	30	145	505	149	**814**
Racing	10	209	481	74	**761**
Simulation	4	95	520	111	**718**
Fighting	4	187	333	87	**606**
Strategy		131	358	96	**580**
Platform	32	135	358	71	**579**
Puzzle	17	74	342	63	**491**
Total	**208**	**1930**	**7077**	**2421**	**11429**



Filters

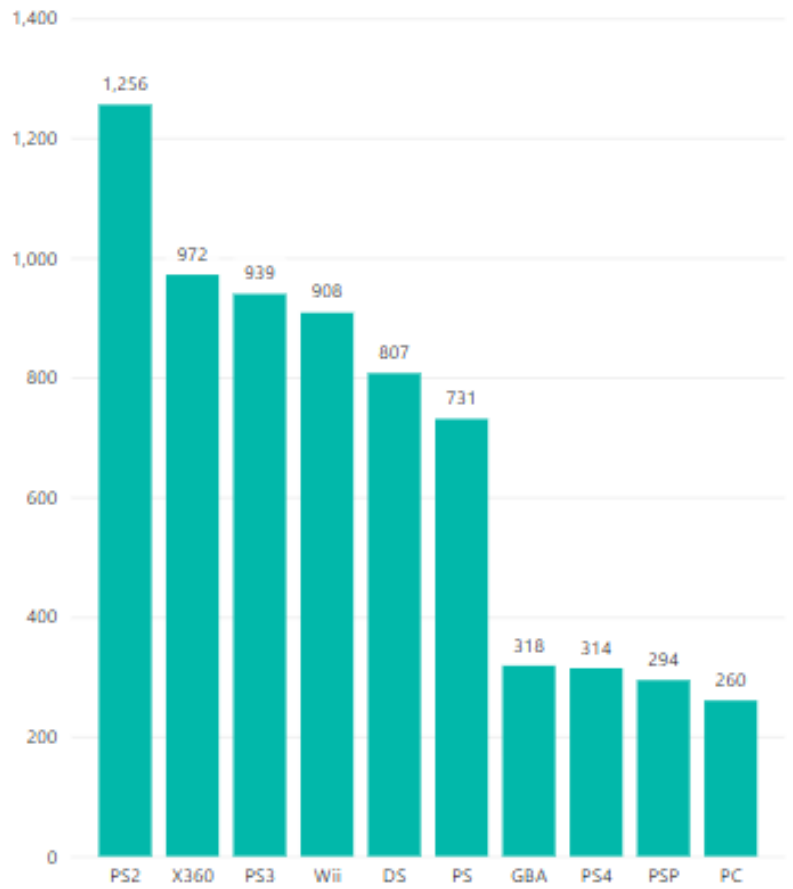
* Based on data till 2020
Sales Figures represented in Millions of Units (5.0K is equivalent to 5 Billion)

Data Analysis & Visualization: Market Analysis (4/4)

Dawn of Gaming Consoles with **extensive graphics** and **3D visualization capabilities** in the **late 90s** contributed to proliferation of console sales making arcade games obsolete

Year Bins

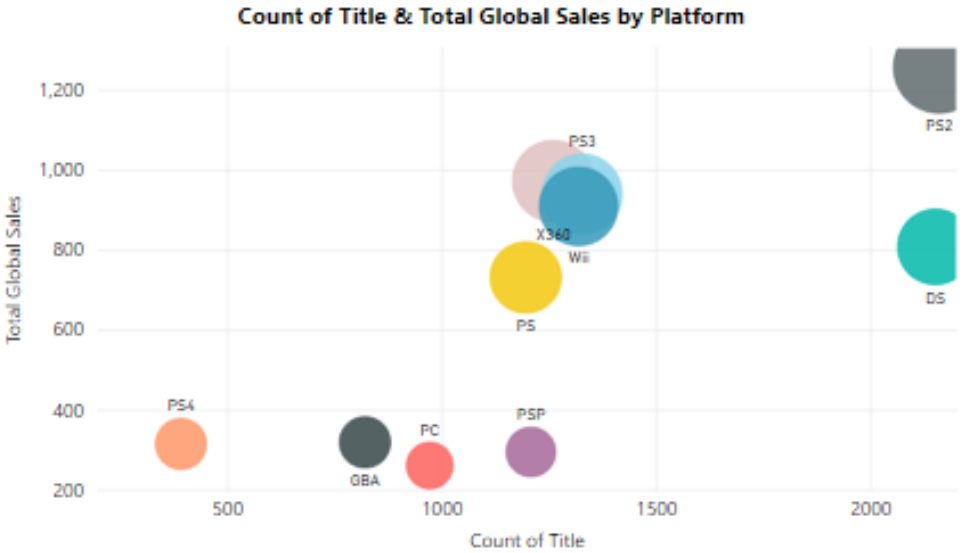
1980-1990	1991-2000	2001-2010	2011-2020



Number of Developers in Top 10 platforms

Platform	1980-1990	1991-2000	2001-2010	2011-2020	Total
DS	1		490	64	**518**
PS2		44	457	5	**465**
Wii			393	93	**432**
X360			311	230	**423**
PS3			245	264	**402**
PSP			262	21	**267**
GBA		1	206		**206**
PS4				176	**176**
3DS				144	**144**
PS		102	40		**125**
Total	**1**	**128**	**1228**	**510**	**1466**

- Sony's "Play Station" Gaming Console series **dominated** the Global Game Sales for **3 consecutive decades**
- PS2 console developed in the late 90s having a sales run until 2013 had the **largest number of game titles** 1,256 & the **highest Global Game Sales** until 2020 of **1.25B** units



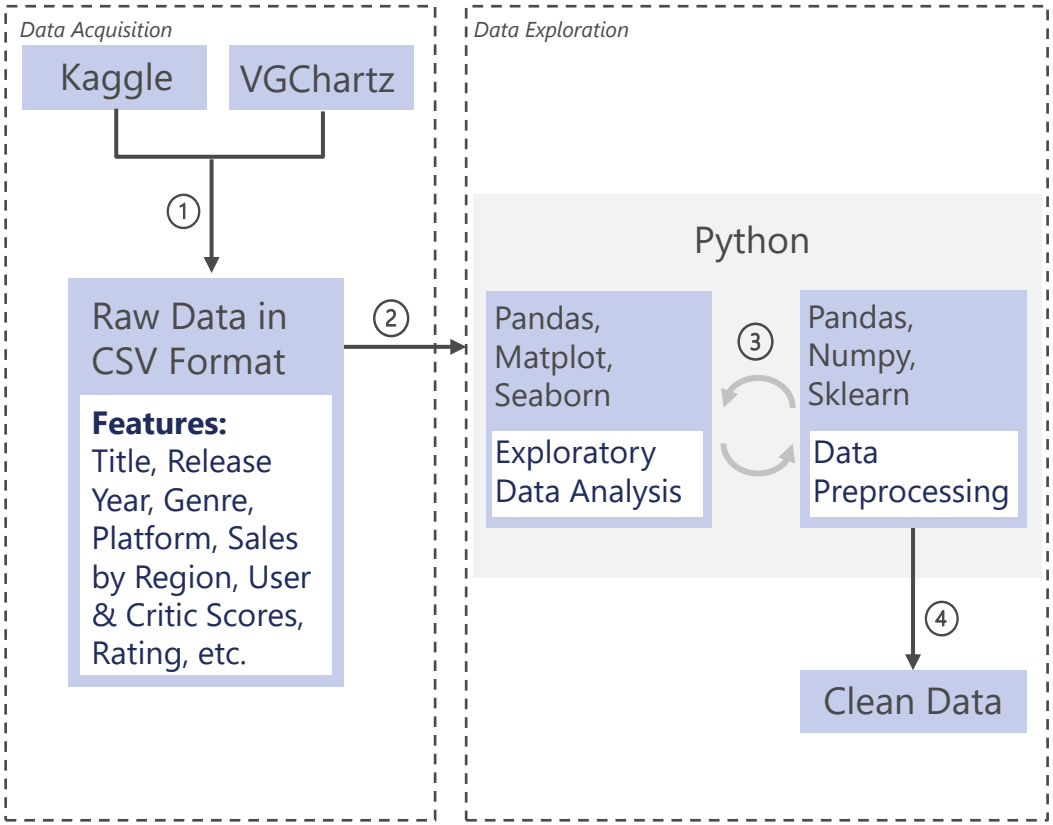
Filters

* Based on data till 2020
Sales Figures represented in Millions of Units (5.0K is equivalent to 5 Billion)

Data Acquisition & Exploration

Python as the main tool for data exploration and data preprocessing

Data Flow



Details

Description

① Sourcing dataset from Kaggle that met the requirements of the project

② Loading CSV data file into Python using Pandas

③ Conducting exploratory data analysis and data preprocessing

④ Getting clean data for model development

Complexity

- Acquiring dataset with relevant features to answer the key questions
- Handling missing data from the dataset

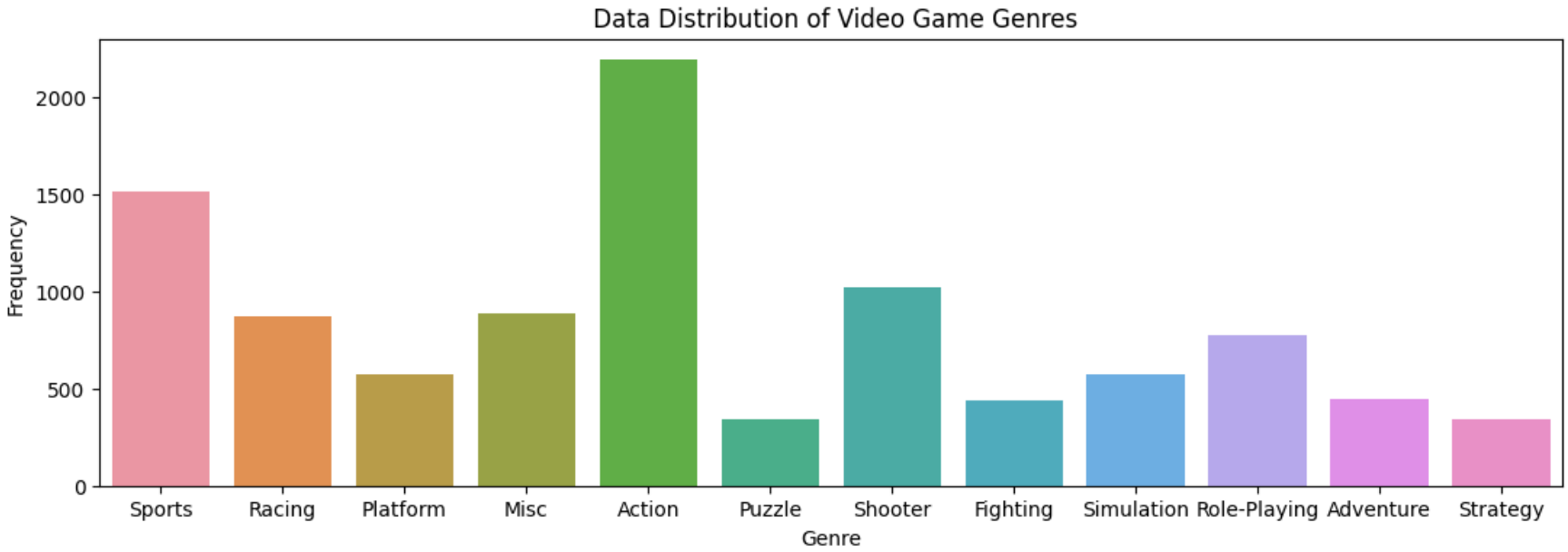
Resolution

- Dropped unnecessary records with missing values
- Replaced *null* values with the imputed data from the respective columns

Exploratory Data Analysis & Preprocessing (1/4)

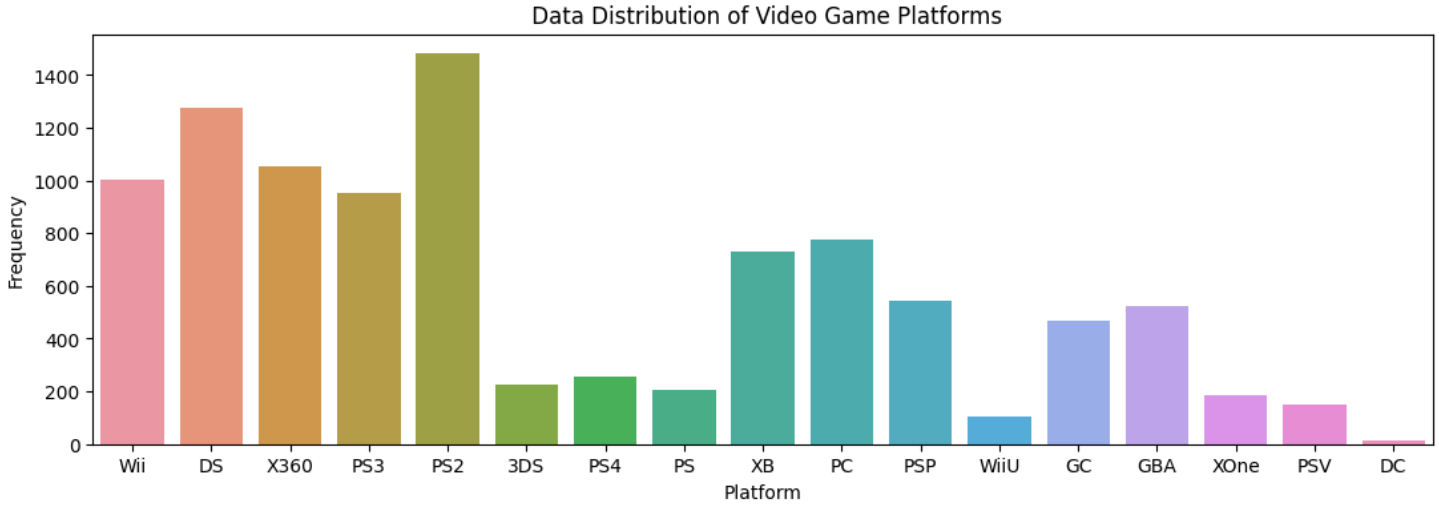
- Checked missing values for each relevant feature in the dataset
  - Missing-data imputation on critic scores and user scores

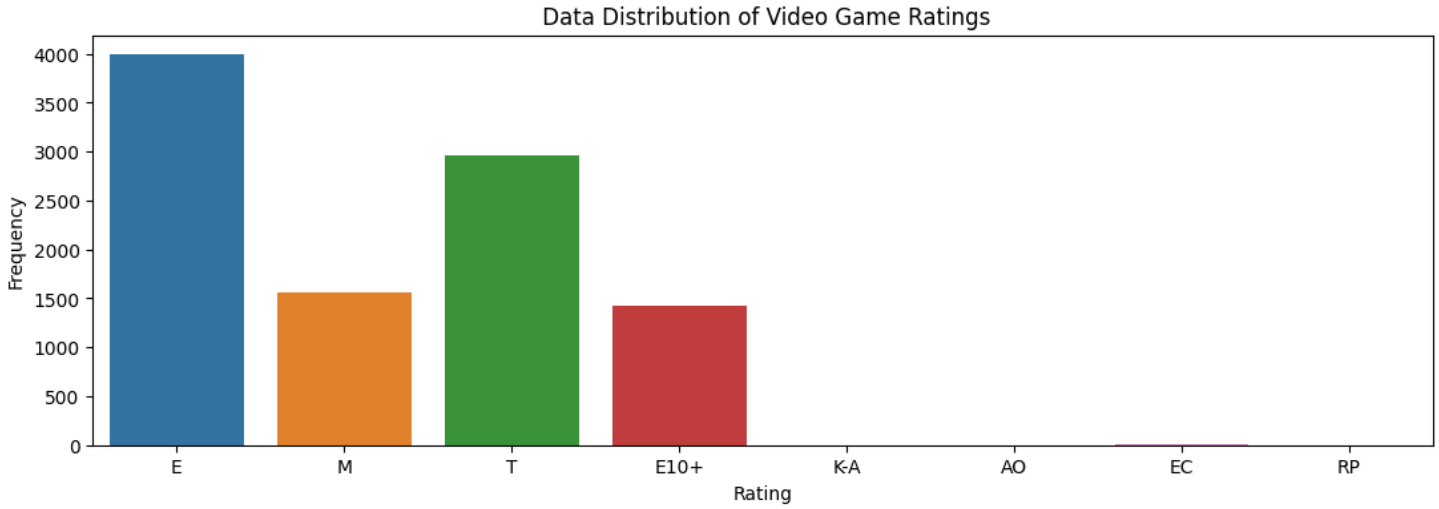
    The imputed data is the mean value of the respective feature within a particular genre, e.g., the average of all user scores under the 'Action' category.
  - Removal of missing data for other features
- Analyzed data distribution for each feature by leveraging Matplotlib & Seaborn packages



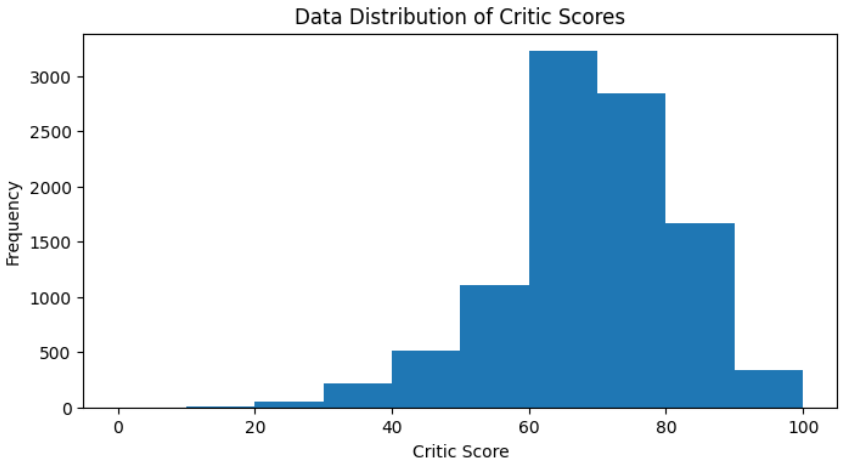
Exploratory Data Analysis & Preprocessing (2/4)

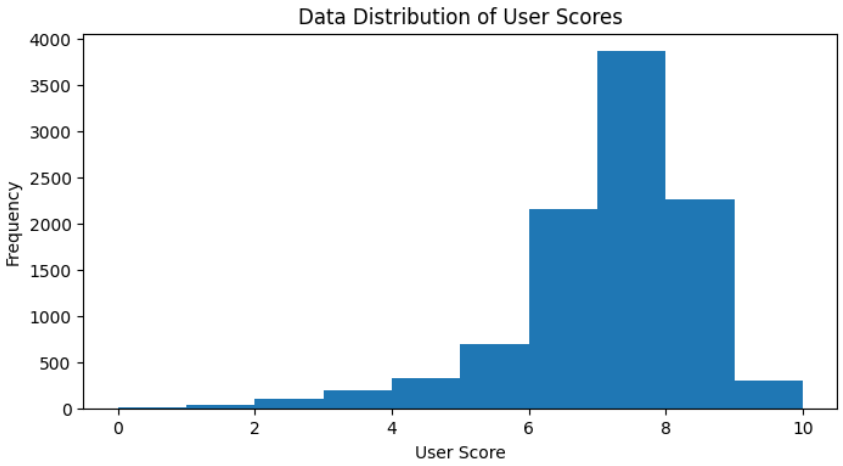
There is a scarcity of data available for certain platforms such as DC, and ratings such as 'K-A', 'AO', 'EC' and 'RP'. If a game falls under these platforms or ratings, the recommendations might also include games from other platforms or ratings that share similar features based on various factors used in distance calculation.

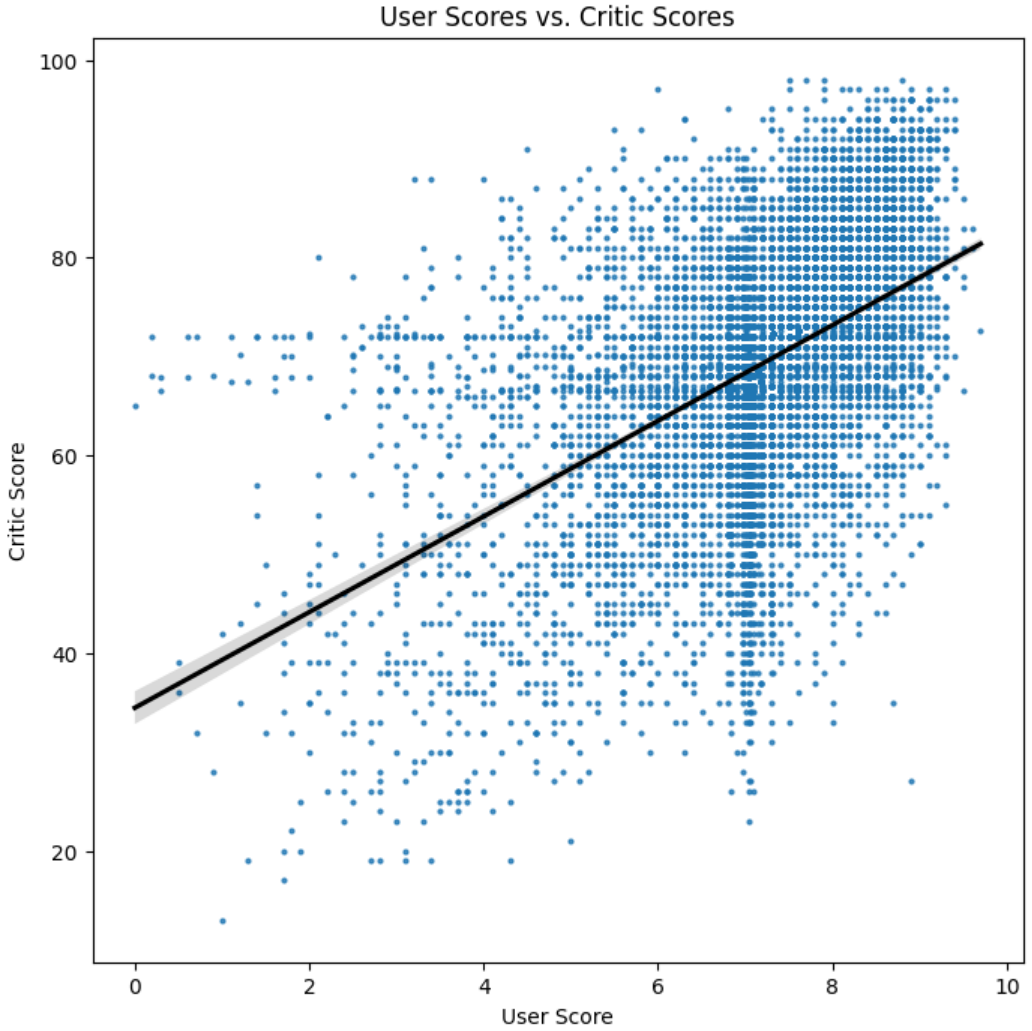




Exploratory Data Analysis & Preprocessing (3/4)







Exploratory Data Analysis & Preprocessing (4/4)

- Converted categorical features to dummy indicators where the value 0 (representing No) or 1 (representing Yes)
- Transformed numerical features to a standardized form for features to have similar scale

Raw Data

Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User_Score	User_Count	Developer	Rating
Wii	2006.0	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	76.0	51.0	8	322.0	Nintendo	E
NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	NaN	NaN	NaN	NaN	NaN	NaN
Wii	2008.0	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52	82.0	73.0	8.3	709.0	Nintendo	E
Wii	2009.0	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	80.0	73.0	8	192.0	Nintendo	E
GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	NaN	NaN	NaN	NaN	NaN	NaN

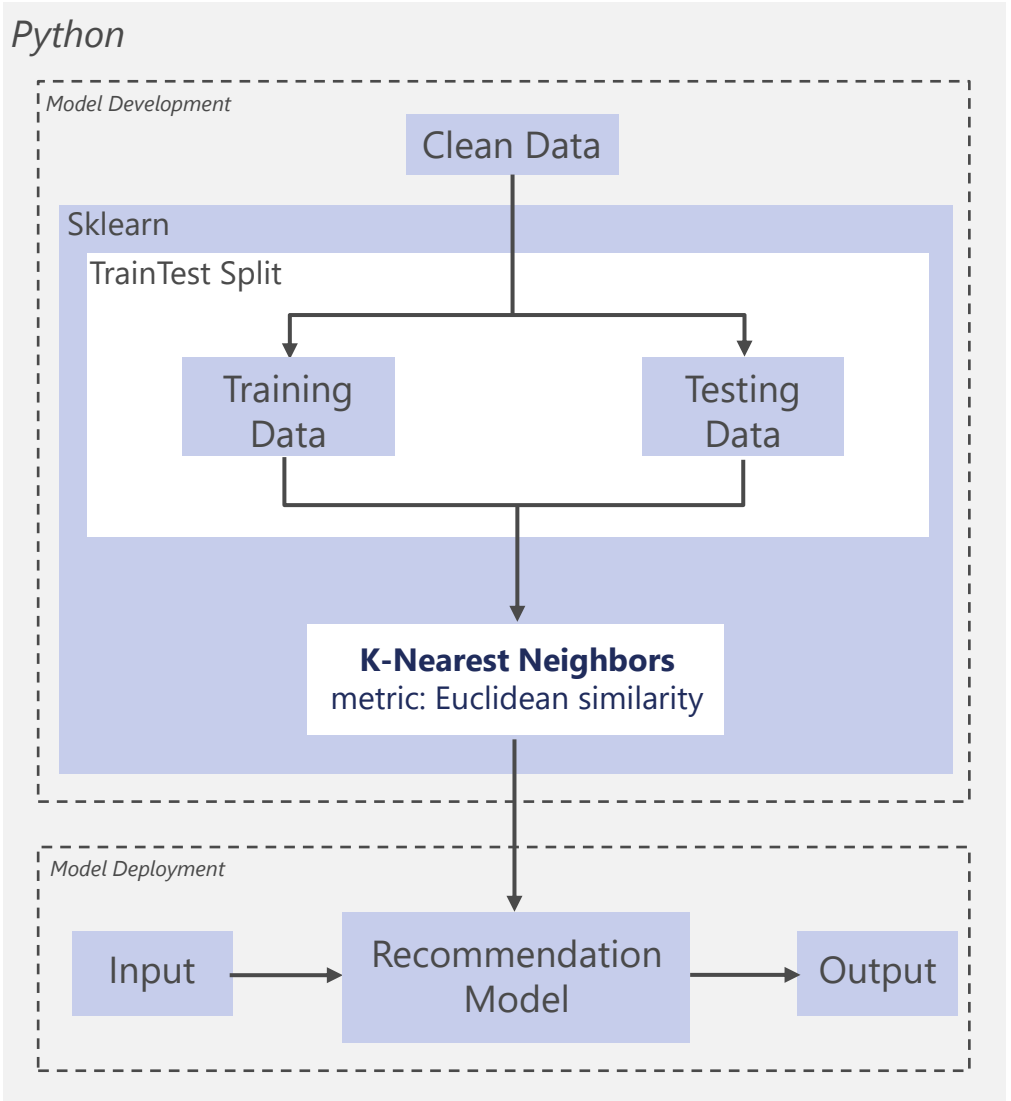
Data Preprocessing

Clean Data

	Critic_Score	User_Score	Platform_3DS	Platform_DC	Platform_DS	Platform_GBA	Platform_GC	Platform_PC	Platform_PS	Platform_PS2	...	Genre_Sports	Genre_Strategy
0	0.567605	0.683282	-0.15314	-0.037537	-0.3832	-0.235302	-0.222413	-0.290432	-0.146119	-0.418178	...	2.365115	-0.187809
1	0.885224	0.683282	-0.15314	-0.037537	-0.3832	-0.235302	-0.222413	-0.290432	-0.146119	-0.418178	...	2.365115	-0.187809
2	0.885224	0.453538	-0.15314	-0.037537	-0.3832	-0.235302	-0.222413	-0.290432	-0.146119	-0.418178	...	2.365115	-0.187809
3	0.885224	0.223794	-0.15314	-0.037537	-0.3832	-0.235302	-0.222413	-0.290432	-0.146119	-0.418178	...	2.365115	-0.187809
4	1.044034	-2.150224	-0.15314	-0.037537	-0.3832	-0.235302	-0.222413	-0.290432	-0.146119	-0.418178	...	2.365115	-0.187809

Model Creation: Recommendation Model (KNN)

Model Architecture



Details

Model

- K-Nearest Neighbor (KNN) from Python scikit-learn is used to create the model
- KNN is a supervised machine learning algorithm that employs distance calculation to determine the similarity between data points
- Value of K determines the number of neighboring data points to be considered during classification or regression

Complexity

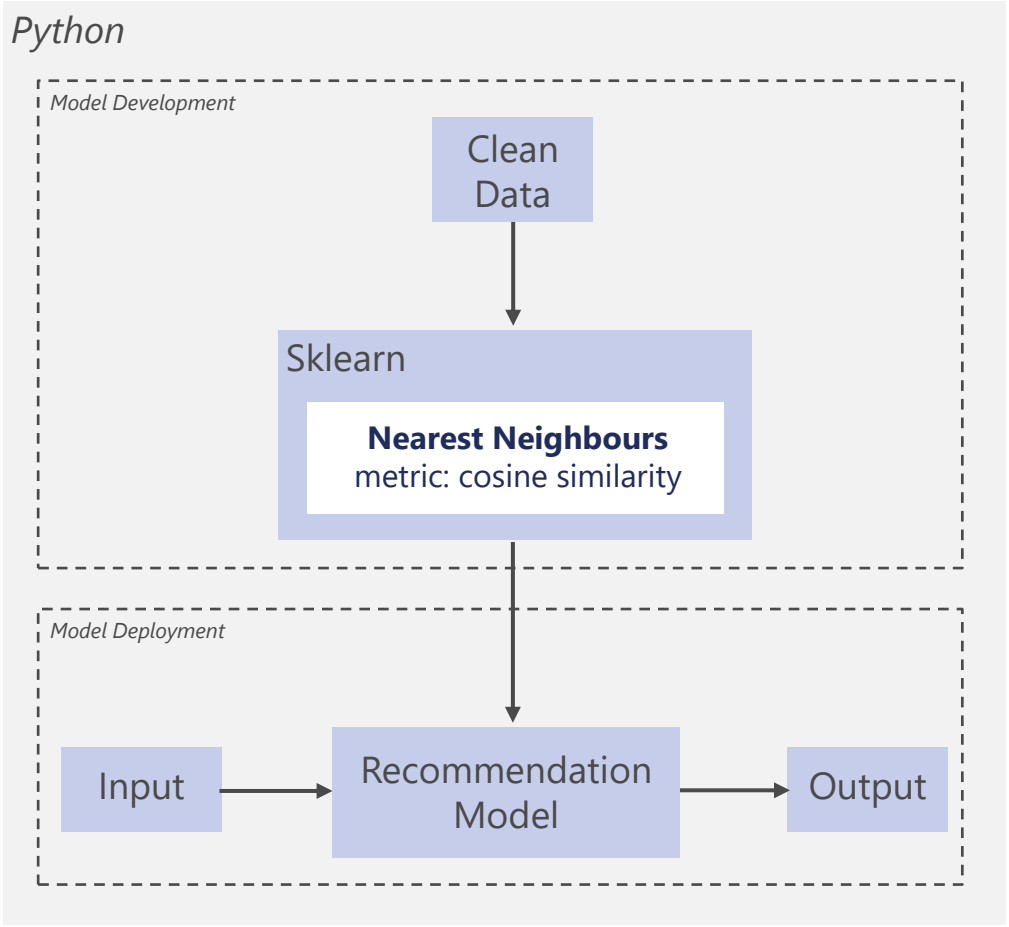
- Complexity on the user input requirement is high, as the user needs to provide all 10 features that the model is trained on for the recommendations to be generated

Resolution

- **Selecting a different model in order to get a better user experience with less input values**

Model Creation: Recommendation Model (NN)

Model Architecture



Details

Model

- Nearest Neighbor (NN) from Python scikit-learn is used to build the model
- NN is an unsupervised machine learning algorithm that uses distance computation to calculate data point's similarity
- Cosine similarity is utilized for distance metric of NN

Complexity

- Games that are available on multiple platforms receive more than ten results for recommendations

Resolution

- Added an optional parameter platform to limit the results
- For multiple platforms: Merge all outcomes, arrange them in ascending order, and select the first ten with the least distance calculation

Model Evaluation (KNN and NN Comparison)

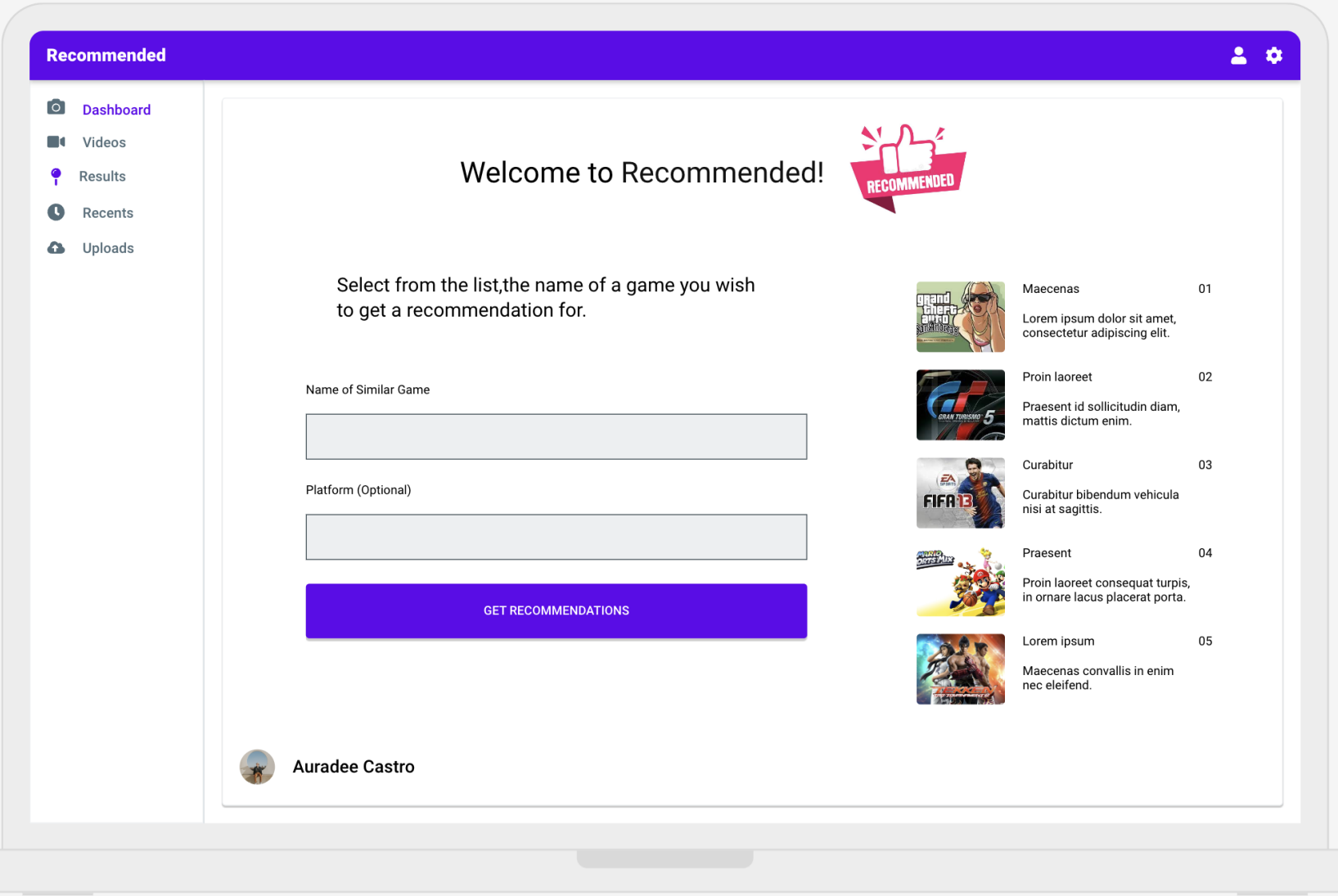
	K-Nearest Neighbors (Supervised Algorithm)	Nearest Neighbor (Unsupervised Algorithm) ✓
Usage	• Better choice if the dataset is large and complex, and the accuracy of the recommendations is crucial	• More suitable for a smaller and simpler dataset where a lightweight algorithm is preferred • No training is required
User Efficiency	• Input is the set of features that was used in the model training to properly classify the game and get recommendations	• Input is the game title and platform *(optional)* – the recommended games have already been pre-calculated for each game

Conclusion:

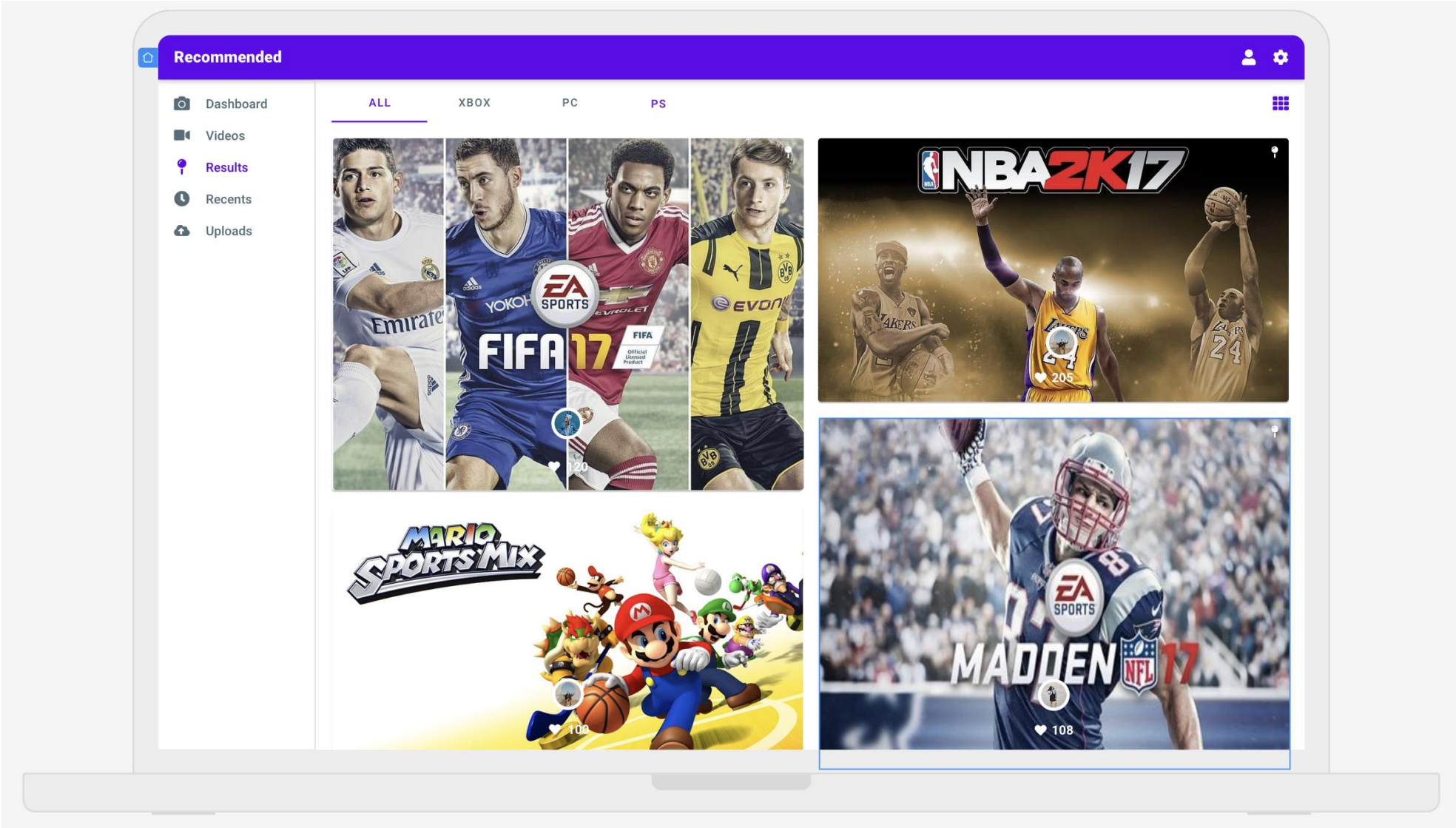
- Nearest Neighbor algorithm was identified as a **more appropriate choice** for the recommendation model aimed at providing video game suggestions based on the title and platform *(optional)* as input.
- The **simplicity and straightforwardness** of the input features were the determining factors, making a full-fledged classification algorithm unnecessary.

*

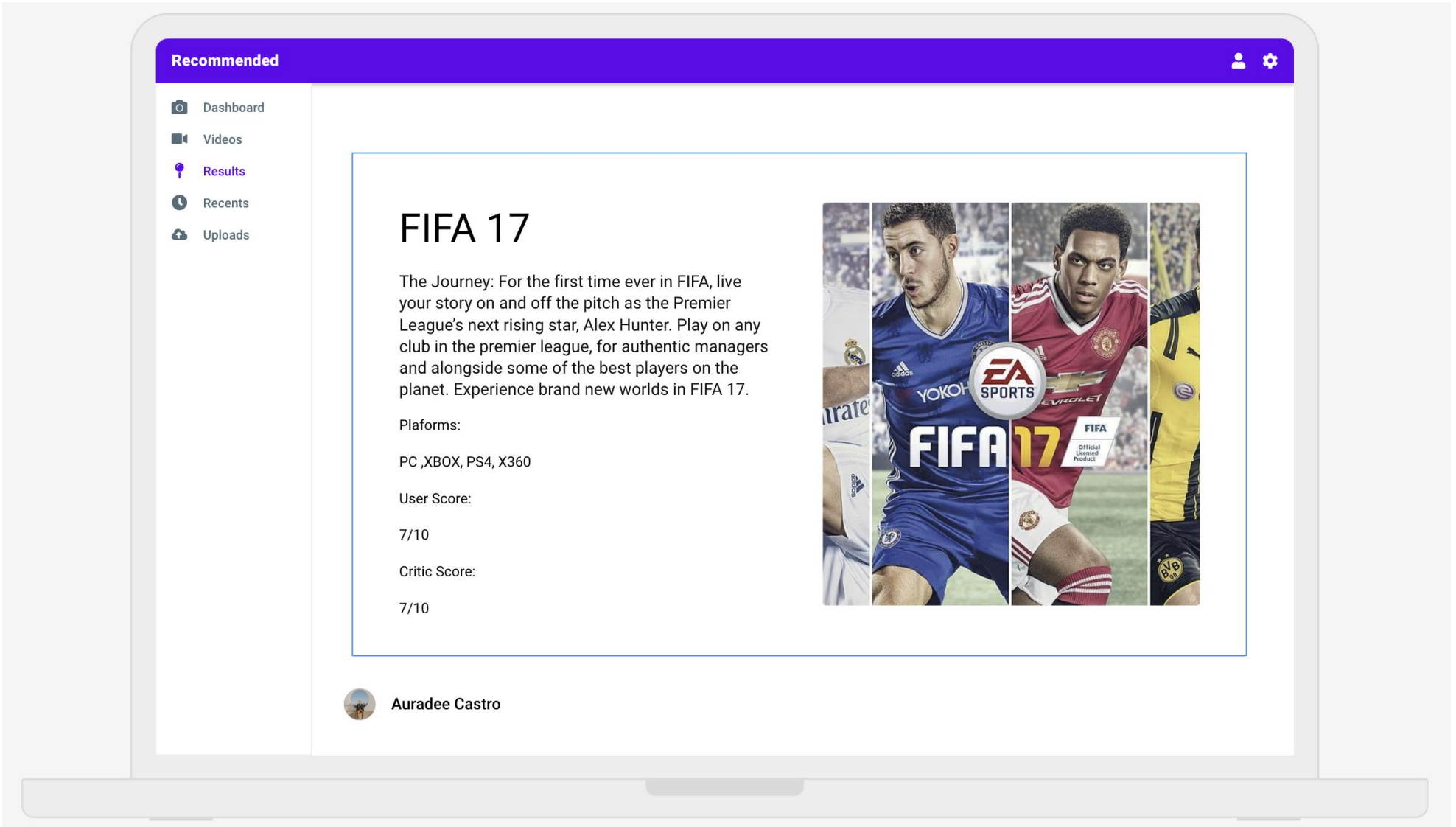
Video Game Recommendation System (UI Prototype)



Video Game Recommendation System (UI Prototype)



Video Game Recommendation System (UI Prototype)



Conclusion

- Sourced the data from Kaggle then preprocessed it
- Conducted Market Trend Analysis by leveraging Power BI
- Moved forward with the EDA (Exploratory Data Analysis)
- Researched appropriate algorithm for the recommendation model
- Selected Nearest Neighbor unsupervised learning model as more appropriate model for the recommendation system
- Developed a UI prototype for the recommendation model for demo purposes


Lambton
College

References

- Kirubi, Rush. (2017). Video Game Sales with Ratings. *Kaggle*. https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings
- *VGChartz*. https://www.vgchartz.com/gamedb/
- Kharwal, Aman. (2021, January 17). Book Recommendation System. *The Clever Programmer*. https://thecleverprogrammer.com/2021/01/17/book-recommendation-system/
- Vijay, Harsh. (2020 April 11). Recommendation System using KNN. *Auriga*. https://www.aurigait.com/blog/recommendation-system-using-knn/
- Makwana, Aman. (2022, December 26). Understanding Recommendation System and KNN with Project – Book Recommendation System. *Medium*. https://aman-makwana101932.medium.com/understanding-recommendation-system-and-knn-with-project-book-recommendation-system-c648e47ff4f6
- Varun. (2020 September 27). Cosine similarity: How does it measure the similarity, Maths behind and usage in Python. *Toward Data Science*. https://towardsdatascience.com/cosine-similarity-how-does-it-measure-the-similarity-maths-behind-and-usage-in-python-50ad30aad7db