Machine Learning Model for Real-life Predictions

Overview

- Machine learning models were trained on data gathered from Steamspy
- Determine the most popular features that well-selling games provide

Objective of Project

- The objective of this project is to train an appropriate machine learning model to be able to predict the relative success of a video game
- Based on characteristics such as:
  - Genre
  - Categories
  - Tags
  - Price

Data Pre-processing

- No missing values were found in the dataset
- One-hot encoding used to interpret data
- Variables were not scaled
- Removed outliers

Feature Selection

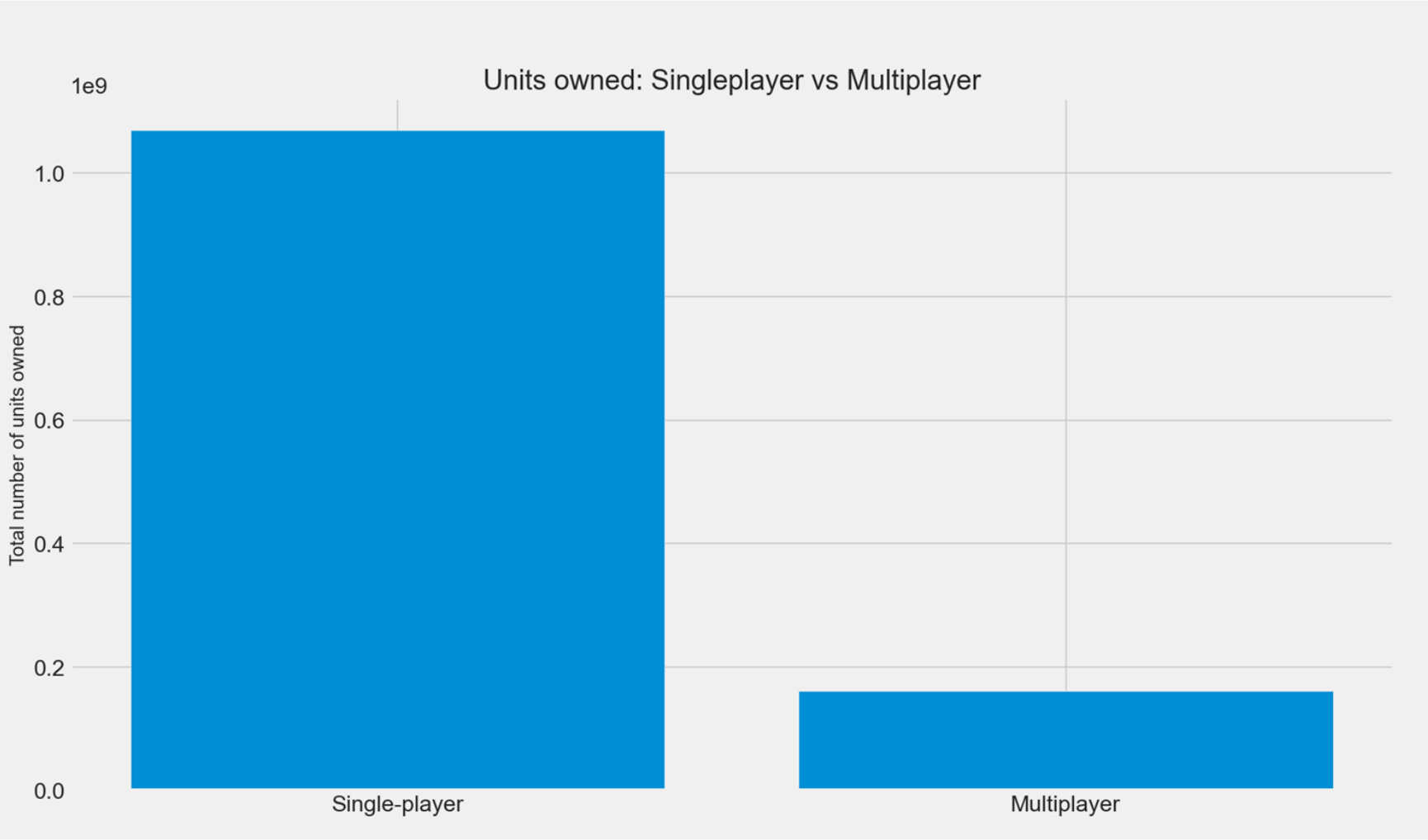
- Chose to prioritize the following columns:
  - Tags
  - Categories
  - Genre
  - Price
  - Positive ratings
  - Negative ratings.

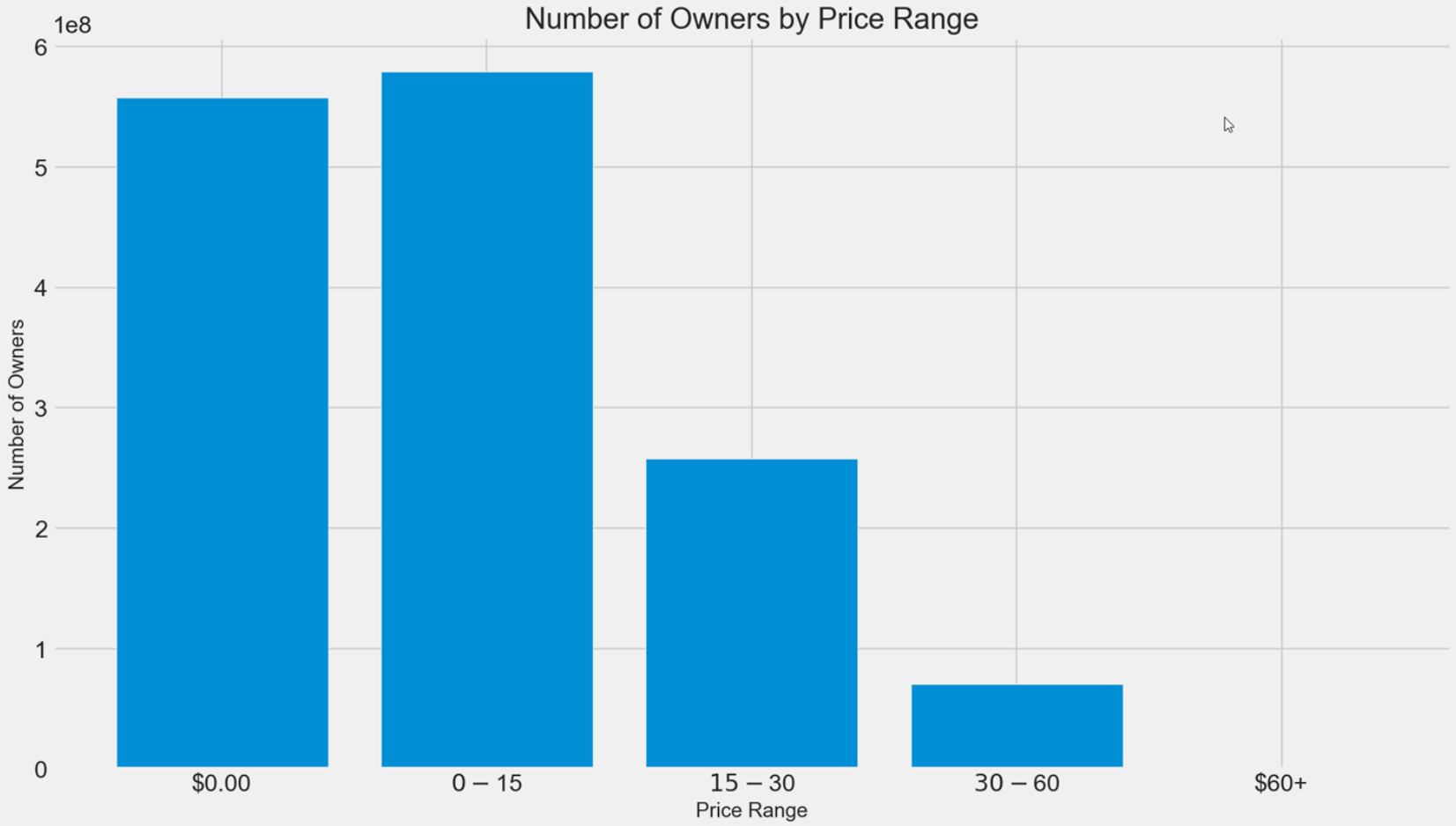
Feature Selection Continued

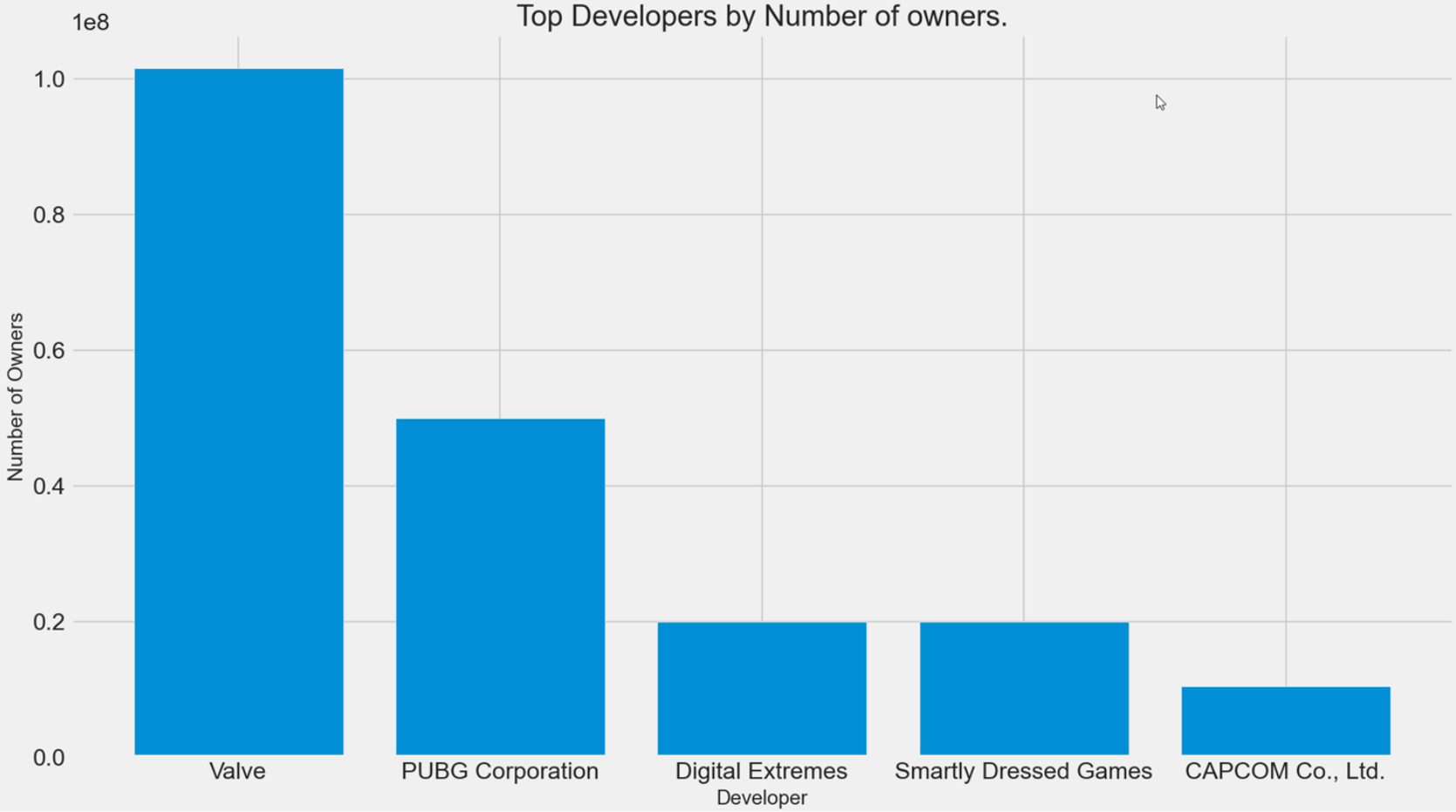
- Dropped achievements column
- Filtered out entries that were released before 2013
- Dropped entries that did not contain an ASCII value in the title

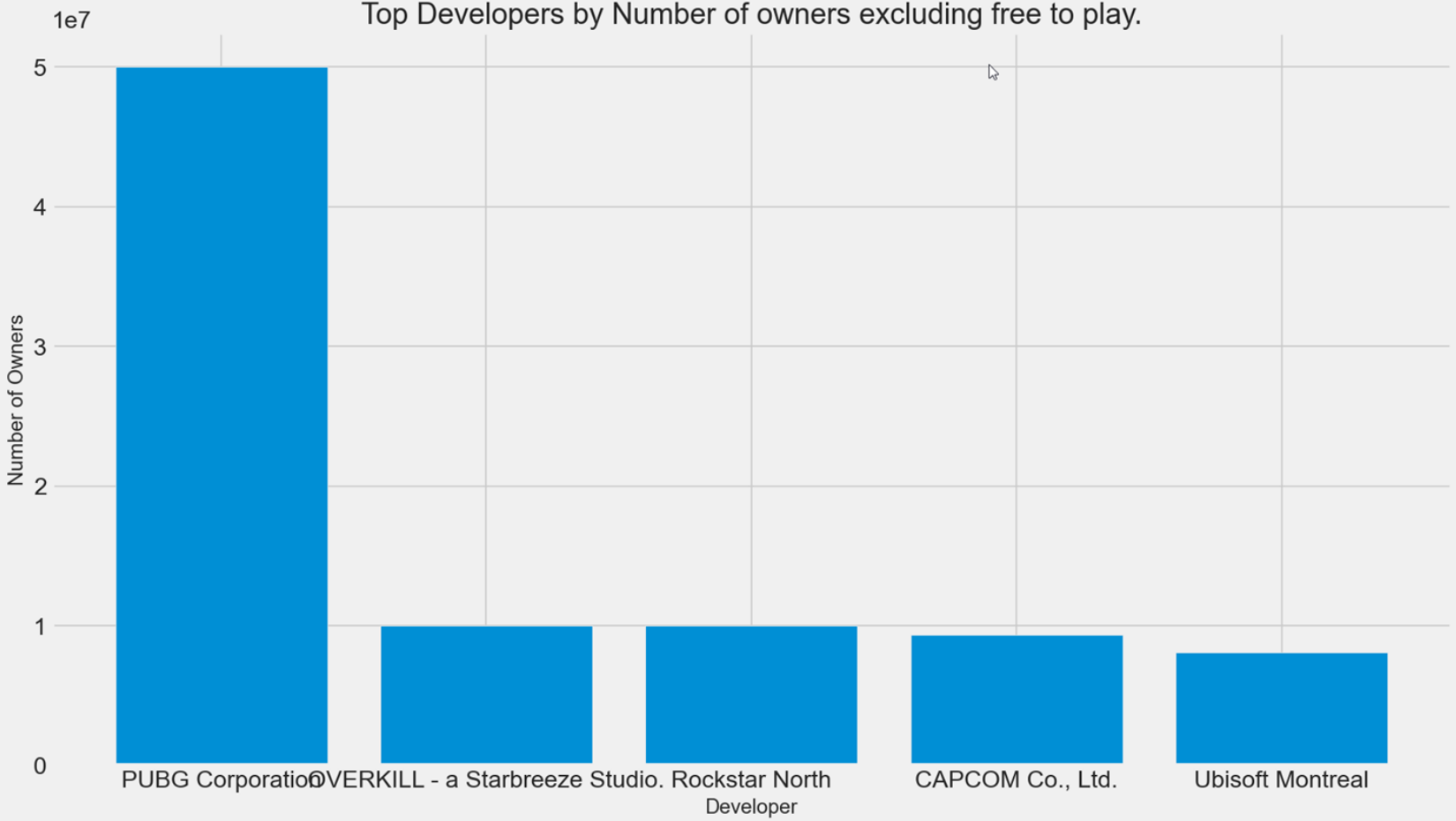
Data Visualization

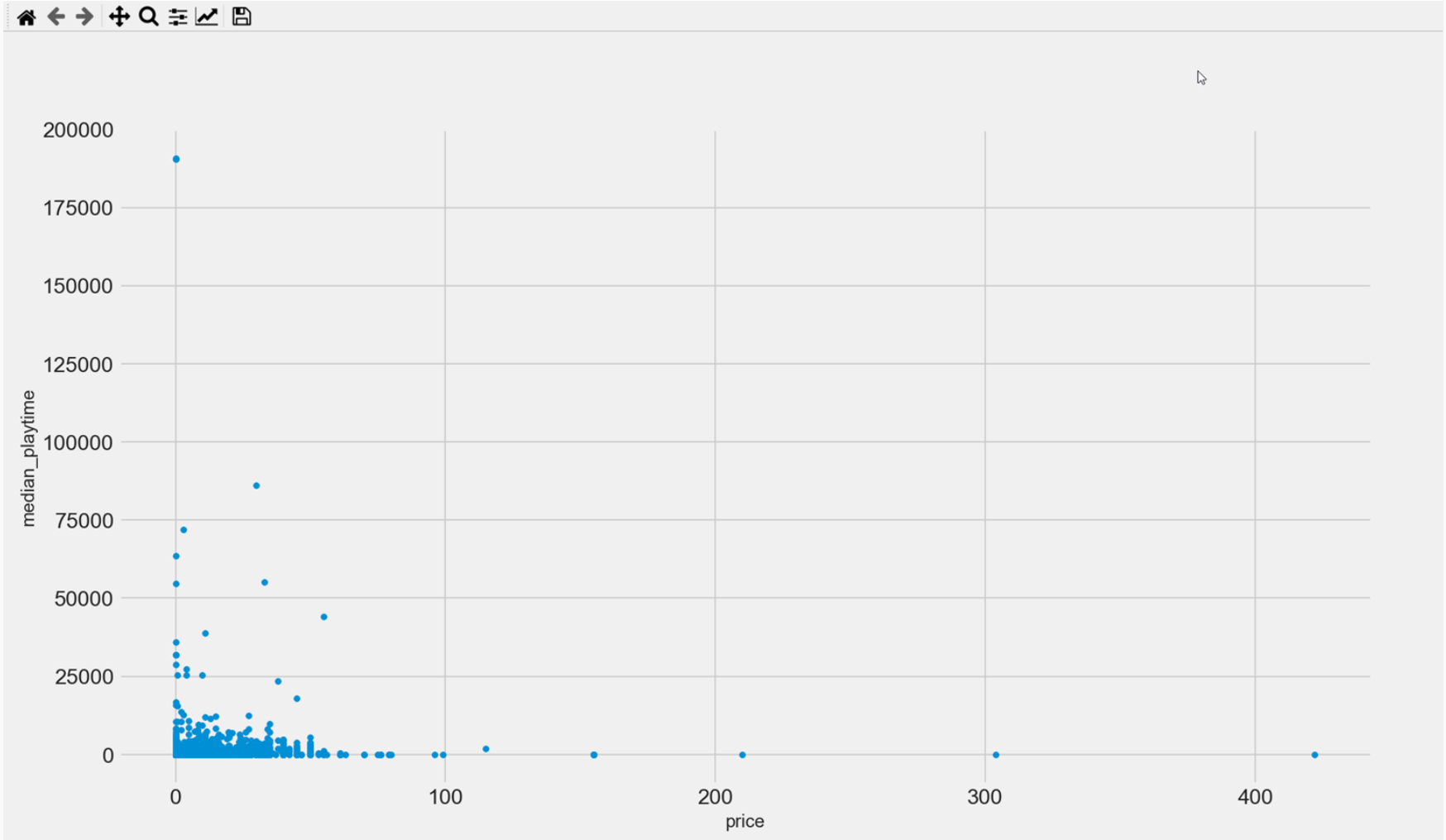
- Process of creating graphs and charts to help you understand the patterns and trends in the data.
- Helps to identify outliers and anomalies in the data.


Units owned: Singleplayer vs Multiplayer
1e9
1.0
0.8
0.6
0.4
0.2
0.0
Total number of units owned
Single-player
Multiplayer


Number of Owners by Price Range
1e8
6
5
4
3
2
1
0
Number of Owners
$0.00
0 – 15
15 – 30
30 – 60
$60+
Price Range


Top Developers by Number of owners.
1e8
1.0
0.8
0.6
0.4
0.2
0.0
Number of Owners
Valve
PUBG Corporation
Digital Extremes
Smartly Dressed Games
CAPCOM Co., Ltd.
Developer


Top Developers by Number of owners excluding free to play.
1e7
5
4
3
2
1
0
Number of Owners
PUBG Corporation
OVERKILL - a Starbreeze Studio.
Rockstar North
CAPCOM Co., Ltd.
Ubisoft Montreal
Developer


200000
175000
150000
125000
100000
75000
50000
25000
0
median_playtime
0
100
200
300
400
price

Platform and Machine Configuration Used

- A Visual Studio Code environment was used with the Anaconda distribution. Every python script was run on a local machine using the environment.

Model Planning

- Different machine learning regression models were used when training on the dataset.
  - Linear Discriminant Analysis
  - Logistic Regression
  - K-Nearest Neighbour
  - Decision Trees
  - Lasso Regression
  - Gaussian Naive Bayes

Model Training

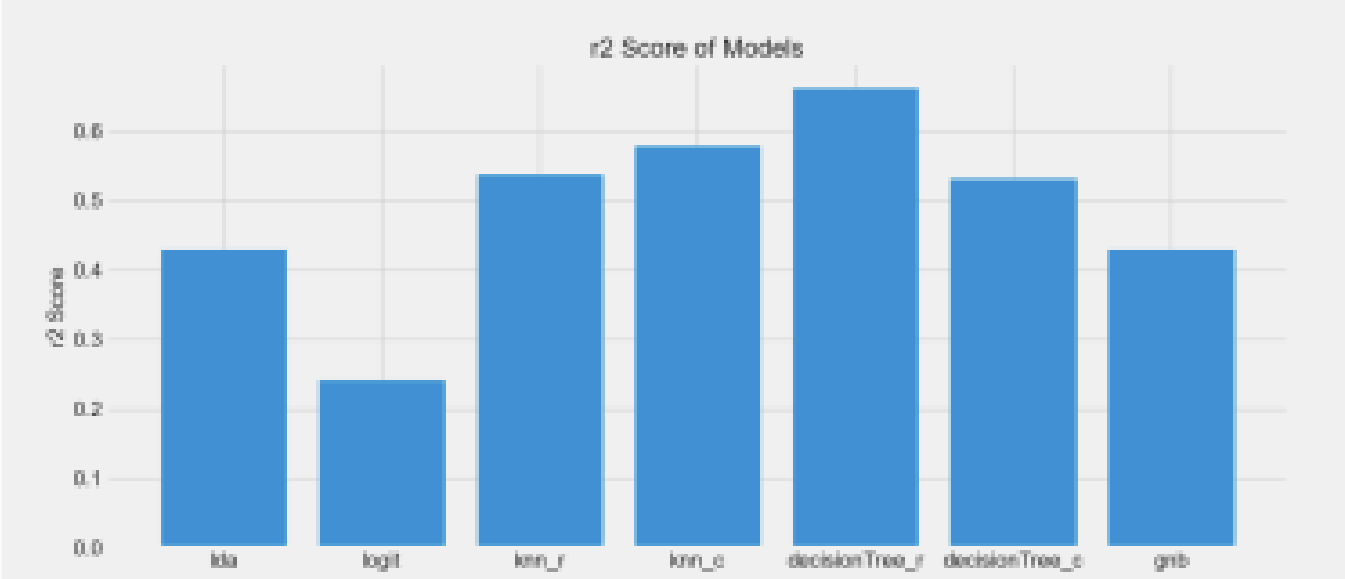
- Used sklearn
- Models were fitted and trained to predict for the “owners” metric which is an estimated number of the users that own the game on the platform.

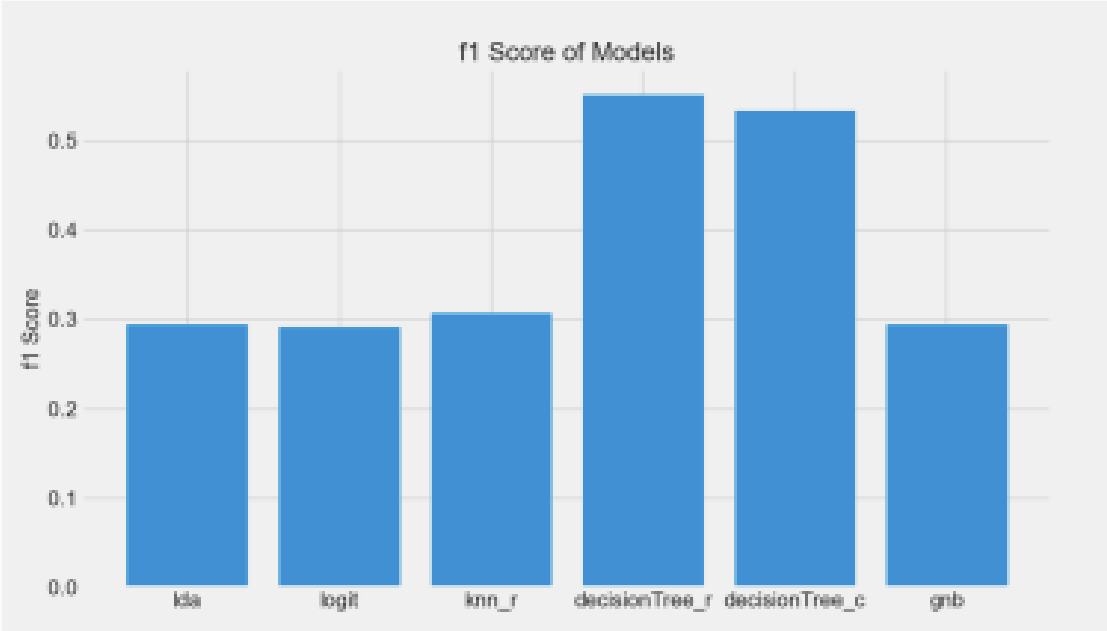
Model Evaluation

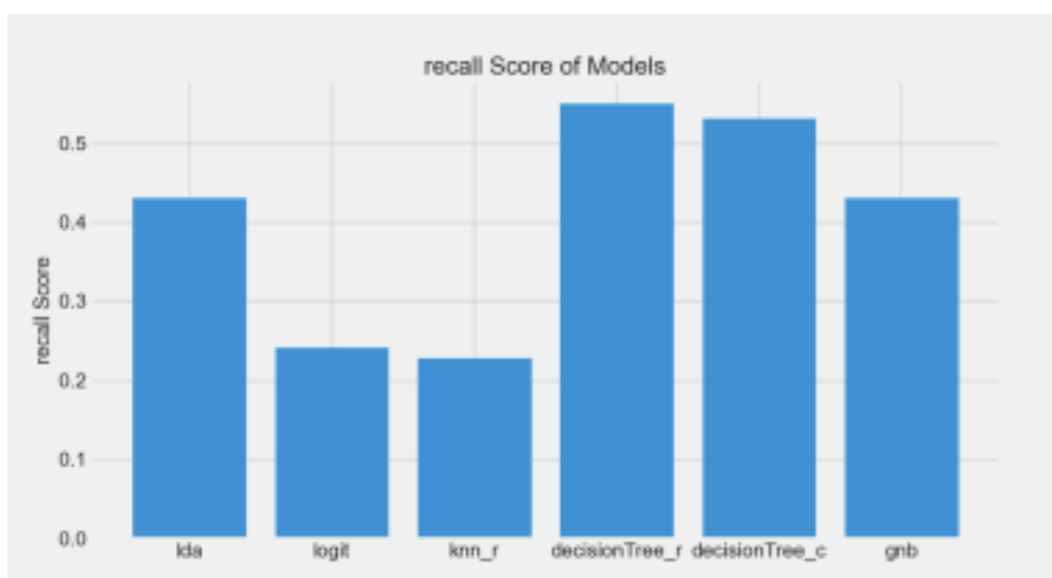
- Model evaluation was done using 4 statistics:
  - Accuracy / r2 score
  - F1 score
  - Precision
  - Recall
- Model optimization was done during the data processing and feature selection phases.

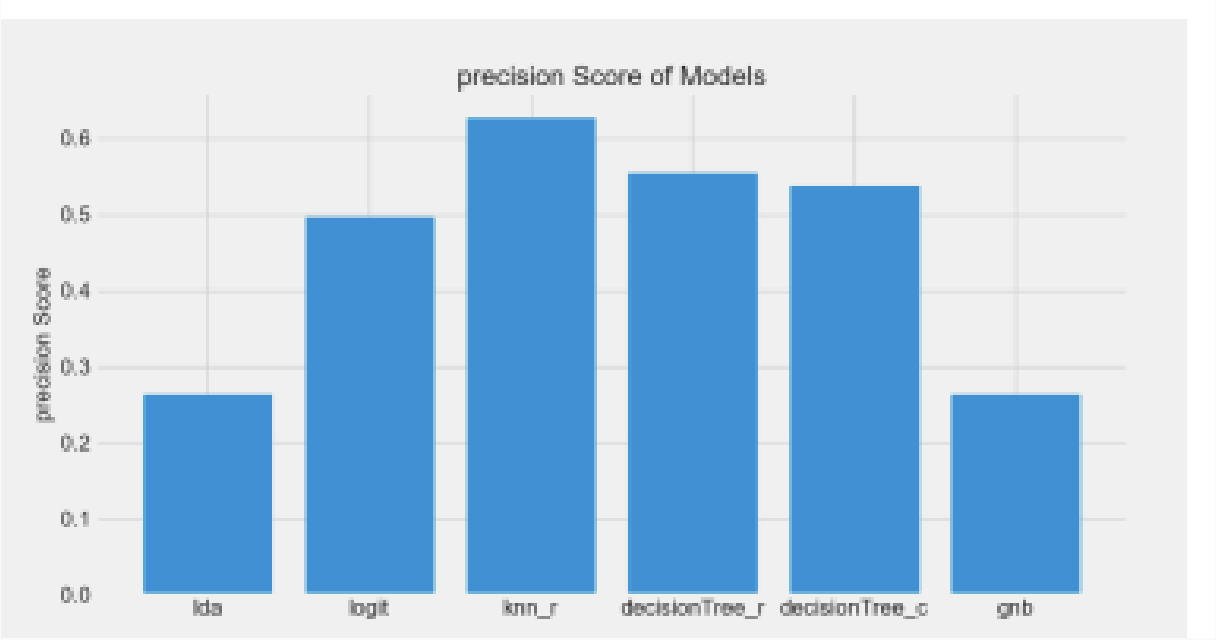
Final Model Building

- Comparison graphs were plotted after all model evaluation was done


r2 Score of Models
r2 Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
lda
logit
knn_r
knn_c
decisionTree_r
decisionTree_c
gnb


f1 Score of Models
f1 Score
0.0
0.1
0.2
0.3
0.4
0.5
lda
logit
knn_r
decisionTree_r
decisionTree_c
gnb


recall Score of Models
recall Score
0.0
0.1
0.2
0.3
0.4
0.5
lda
logit
knn_r
decisionTree_r
decisionTree_c
gnb


precision Score of Models
precision Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
lda
logit
knn_r
decisionTree_r
decisionTree_c
gnb

Result Table

Model	R2 / Accuracy	F1	Recall	Precision
Logistic Regression	0.2421	0.2918	0.2421	0.4976
LDA	0.4307	0.2949	0.4307	0.2642
KNN Regressor	0.5386	0.3071	0.2282	0.6281
KNN Classifier	0.5786	0.5603	0.5786	0.5520
DT Regressor	0.6623	0.5525	0.5501	0.5562
DT Classifier	0.5320	0.5345	0.5320	0.5383
Gaussian NB	0.4307	0.2949	0.4307	0.2642
Mean	0.4879	0.4051	0.4275	0.4715
Std	0.1254	0.1249	0.1325	0.1359

Interpretation of Results:

- Decision tree model yields the best results of performance metrics.
- The final model did not achieve the specified accuracy
  - Number of owners was an estimated value between a broad range.
  - An accurate value of sales may provide better performance metrics.
- Strong decision tree results may suggest:
  - Non-linear relationships between variables
  - Mixture of both categorical and numerical data.

Sensitivity Analysis

- Various attempts were made at changing certain aspects of the process in order to obtain the best performance metrics:
  - Drop entries based on their release date
  - Experiment with the value of SelectKBest

Future Research

- More emphasis should be put on obtaining accurate sales figures.
- An alternative solution could be to replace the number of owners column with a sales column.