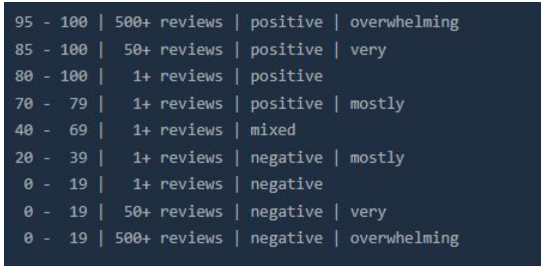
# **Unlocking Game Success:** *Can We Predict Game Reviews from Key Attributes?*

*Ananya Achan, Elaine Keung, Mia Kobayashi, Francisco Munoz*

# Research Question

Can we predict the review of a game based on its features such as price, genres, publishers, and other relevant features?

**Interest**: Guidance for developers to create popular and well received games with respect to metadata (writing descriptions), OS support, discount timeline and strategy, etc based on “successful” games, while exploring consumer habits in game preference.



```
95 - 100 | 500+ reviews | positive | overwhelming
85 - 100 |  50+ reviews | positive | very
80 - 100 |   1+ reviews | positive
70 -  79 |   1+ reviews | positive | mostly
40 -  69 |   1+ reviews | mixed
20 -  39 |   1+ reviews | negative | mostly
 0 -  19 |   1+ reviews | negative
 0 -  19 |  50+ reviews | negative | very
 0 -  19 | 500+ reviews | negative | overwhelming
```

# Data

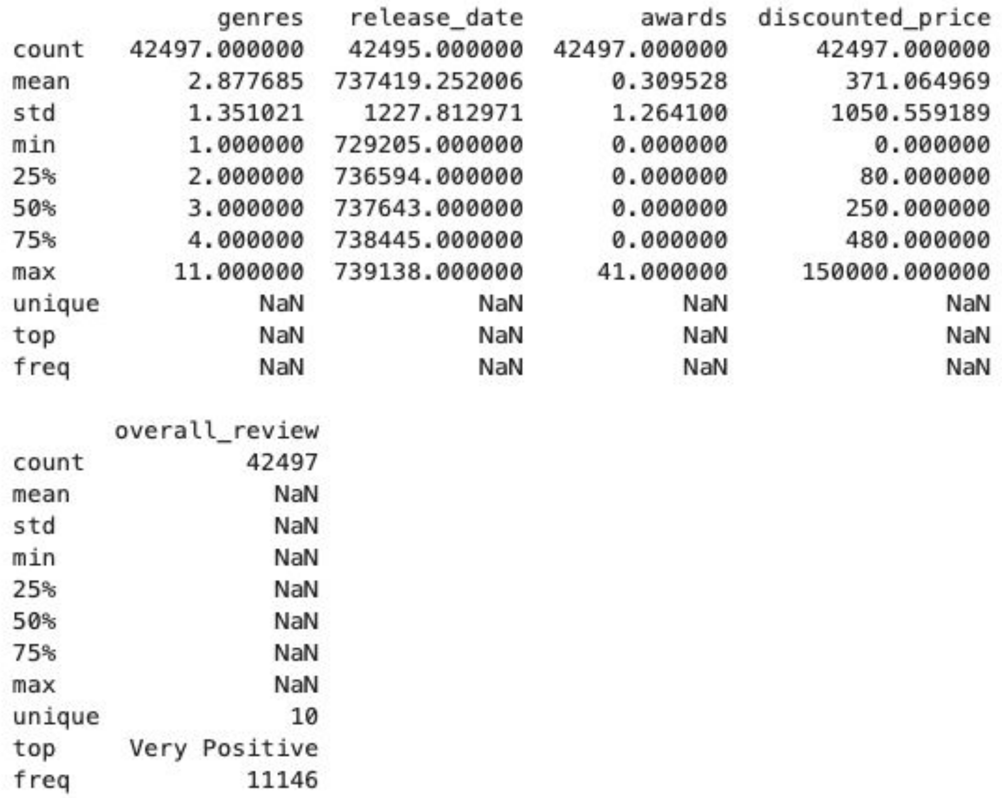
Date: 19 May 2024

Source: Kaggle. Steam Store: a site dedicated to “playing, discussing, and creating games”

Rows: 42,496

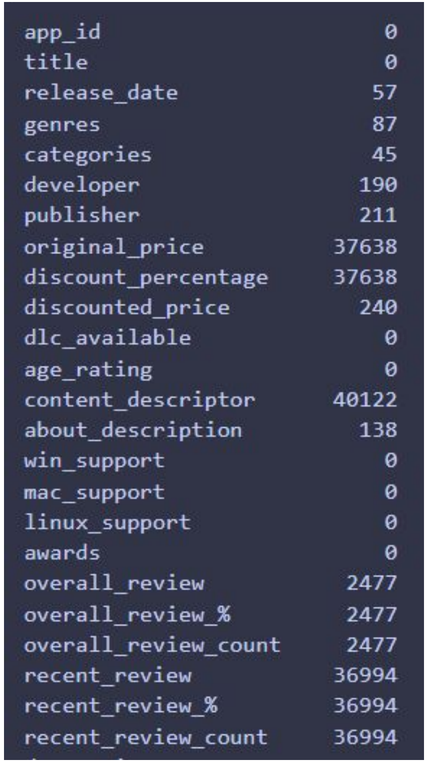
Columns: 24

# Summary of Data



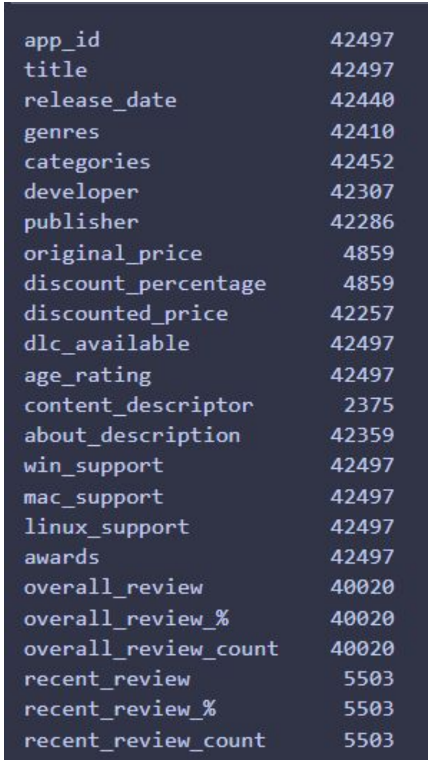
```
             genres   release_date        awards  discounted_price
count  42497.000000   42495.000000  42497.000000      42497.000000
mean       2.877685  737419.252006      0.309528        371.064969
std        1.351021    1227.812971      1.264100       1050.559189
min        1.000000  729205.000000      0.000000          0.000000
25%        2.000000  736594.000000      0.000000         80.000000
50%        3.000000  737643.000000      0.000000        250.000000
75%        4.000000  738445.000000      0.000000        480.000000
max       11.000000  739138.000000     41.000000     150000.000000
unique          NaN            NaN           NaN               NaN
top             NaN            NaN           NaN               NaN
freq            NaN            NaN           NaN               NaN

       overall_review
count           42497
mean              NaN
std               NaN
min               NaN
25%               NaN
50%               NaN
75%               NaN
max               NaN
unique             10
top     Very Positive
freq            11146
```

Total number of null values in column



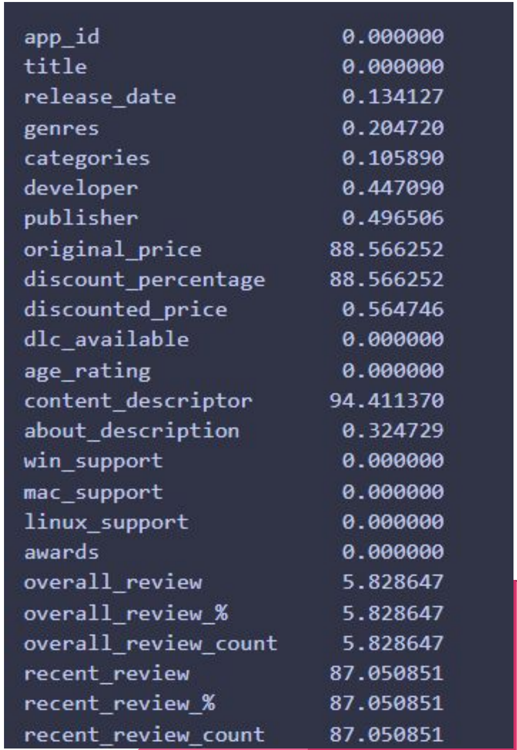
```
app_id                      0
title                       0
release_date               57
genres                     87
categories                 45
developer                 190
publisher                 211
original_price          37638
discount_percentage     37638
discounted_price          240
dlc_available               0
age_rating                  0
content_descriptor      40122
about_description         138
win_support                 0
mac_support                 0
linux_support               0
awards                      0
overall_review           2477
overall_review_%         2477
overall_review_count     2477
recent_review           36994
recent_review_%         36994
recent_review_count     36994
```

Number of values in column



```
app_id                  42497
title                   42497
release_date            42440
genres                  42410
categories              42452
developer               42307
publisher               42286
original_price           4859
discount_percentage      4859
discounted_price        42257
dlc_available           42497
age_rating              42497
content_descriptor       2375
about_description       42359
win_support             42497
mac_support             42497
linux_support           42497
awards                  42497
overall_review          40020
overall_review_%        40020
overall_review_count    40020
recent_review            5503
recent_review_%          5503
recent_review_count      5503
```

Percentage of null values in column



```
app_id                   0.000000
title                    0.000000
release_date             0.134127
genres                   0.204720
categories               0.105890
developer                0.447090
publisher                0.496506
original_price          88.566252
discount_percentage     88.566252
discounted_price         0.564746
dlc_available            0.000000
age_rating               0.000000
content_descriptor      94.411370
about_description        0.324729
win_support              0.000000
mac_support              0.000000
linux_support            0.000000
awards                   0.000000
overall_review           5.828647
overall_review_%         5.828647
overall_review_count     5.828647
recent_review           87.050851
recent_review_%         87.050851
recent_review_count     87.050851
```

# Features of Interest

**overall_review**: The overall review category classification based on the rating score of the game (target)

**genres:** List of genres the game belongs to

**release_date**: Date the game was published

**awards**: Number of awards the game has received

**discounted_price:** Price after the discount (as of the time the data was scraped)

# Additional Features

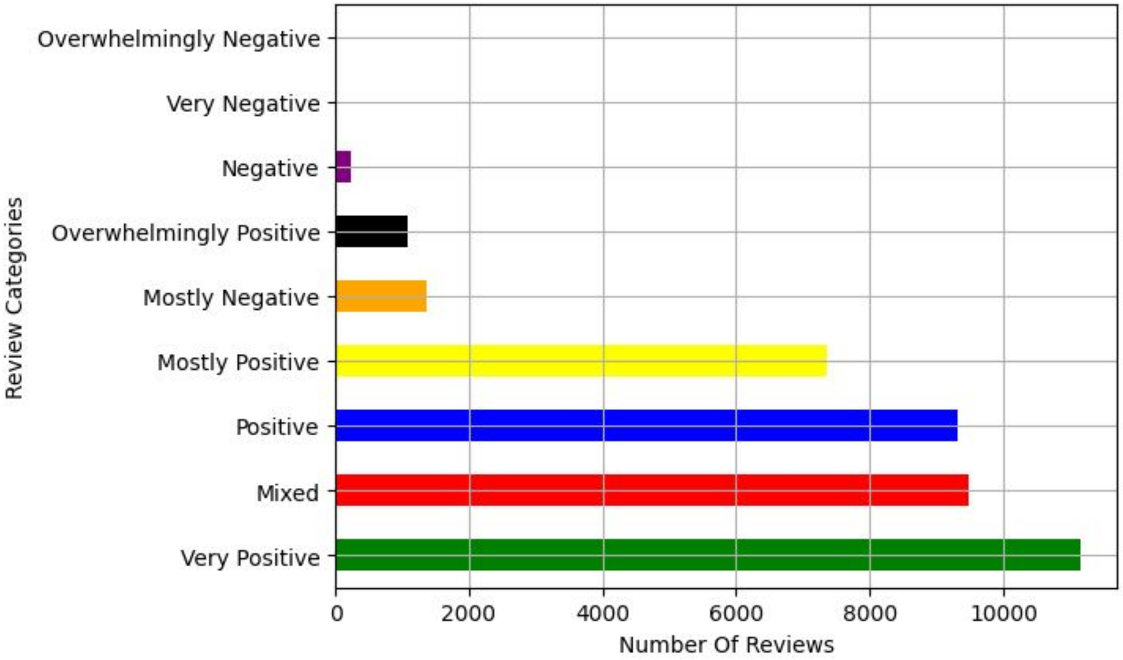
**categories:** Types of game content / features

**about_description**: A textual description about the game

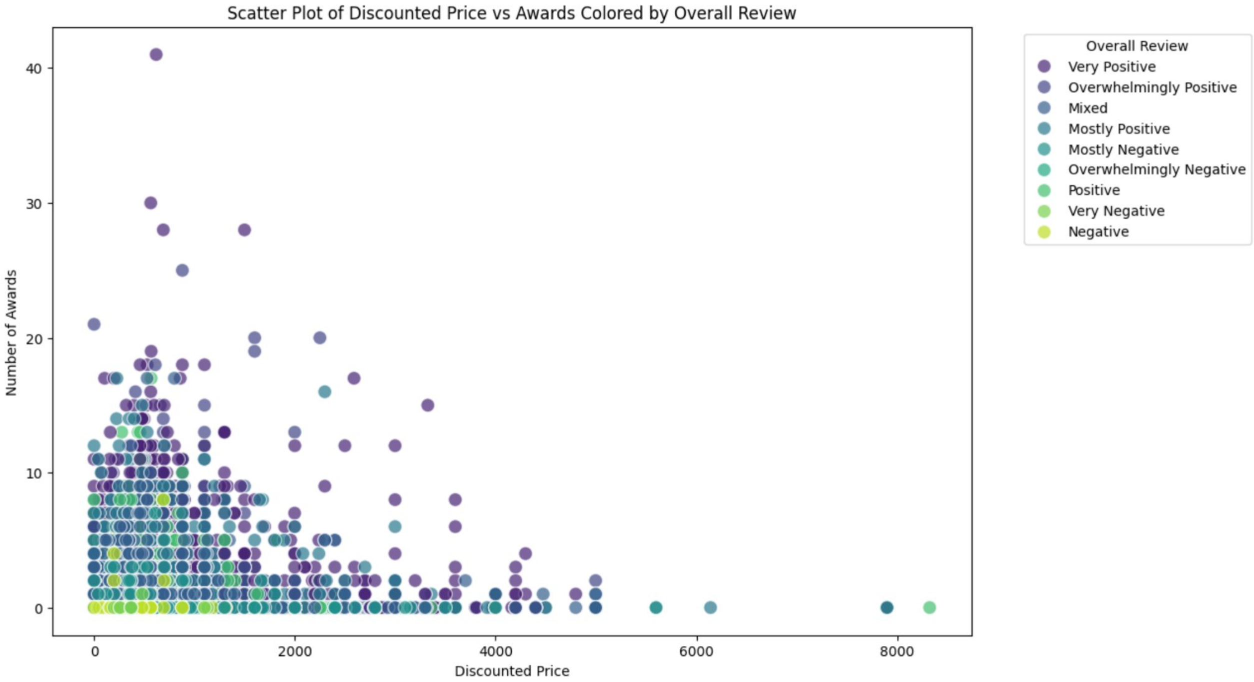
**mac_support**: Binary indicators of platform support
(25% of the games support mac)

**developer:** The company or individual who developed the game

# Outcome Variable: Review



# Relationship between Price, Awards and Review



# Baseline Prediction Model

Multi-class Logistic Regression

- Simple, interpretable
- Probabilistic outputs
- Feature importance

# Prediction Algorithms

- **Decision Trees**
  Are intuitive and easy to interpret, allowing us to visualize how different features impact the final decision, which facilitates explaining the results to non-technical users.

- **XGBoost**
  An optimized version of decision trees, is highly effective in classification due to its ability to handle large datasets and enhance performance through boosting techniques. We will assign a higher penalty to misclassifications of the minority class using scale weights parameter.

- **Neural networks + hidden layers**
  Powerful for capturing complex non-linear relationships and patterns in the data. Their ability to model intrinsic complexities is essential in a domain as variable as video game preferences, where feature interactions can be highly sophisticated.

# Performance Metrics

- Accuracy: Measures the ratio of correctly predicted instances to the total instances, can be misleading with class imbalance.
- Precision, Recall, and F1-Score
- Confusion Matrix
- Balanced Accuracy
- ROC-AUC and Precision-Recall AUC

# Questions?