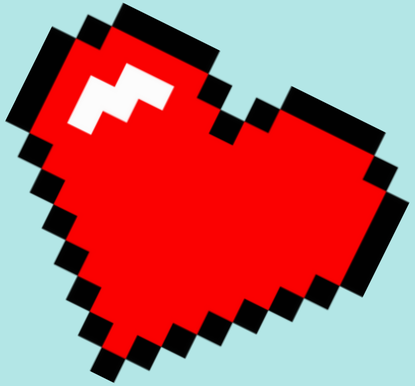
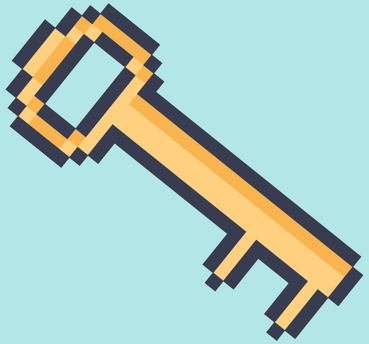
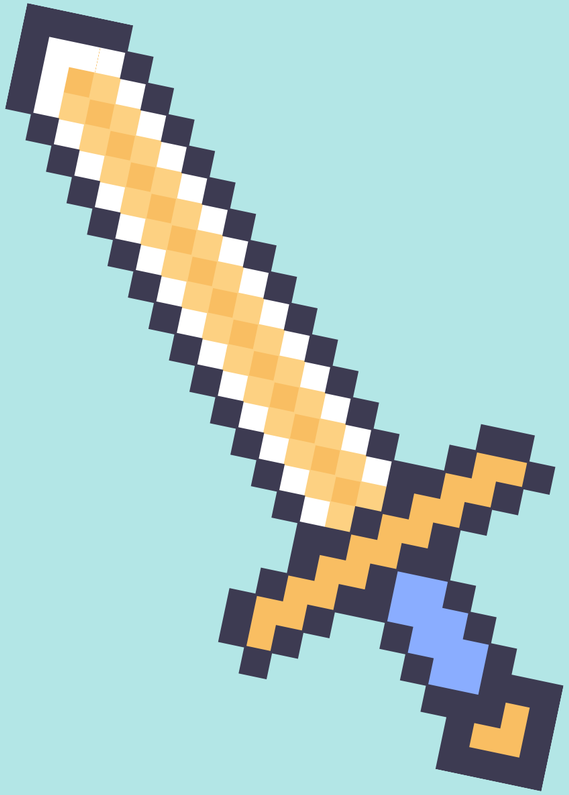
NATHAN NGUYEN





VIDEO GAME DATASET (2020)



MENU START

THE DATASET

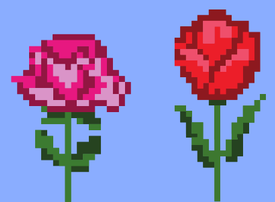
VARIABLES:

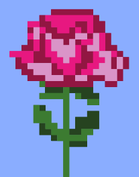
- title: name of video game title
- console: name of console that the game was first released on
- genre: the genre of the game
- publisher: name of game's publisher
- developer: name of game developer/developing company
- vg_score = the vgchartz critical score (out of 10)
- critic_score = the metacritic score (out of 10)
- user_score = the users score (out of 10)
- total_shipped = the total number of the game shipped in millions
- total_sales = the global sales in millions
- na_sales = the North American sales in millions
- jp_sales = the Japanese sales in millions
- pal_sales = the PAL(Phase Alternating Line region) sales in millions
- other_sales = other sales in millions
- release_date = the game's release date
- last_update = the last date that the data in the row was updated

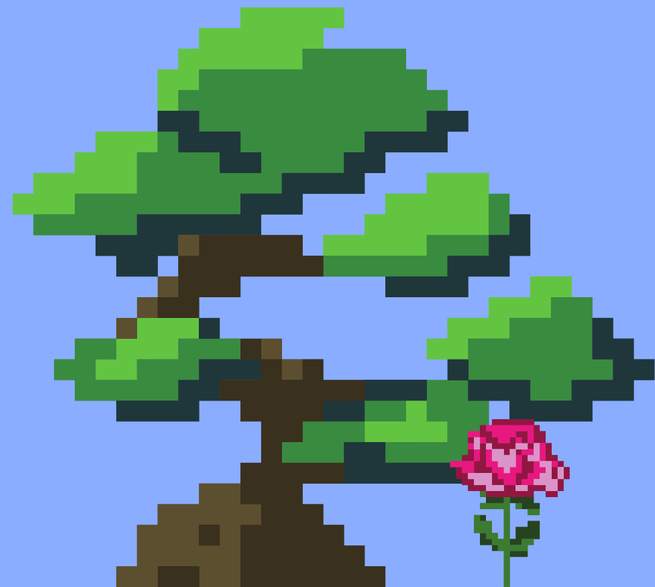
THE DATASET

CLEAN UP:

- Dropped total shipped
- Dropped total sales
- Dropped release date
- Dropped last update
- Dropped title
- Dropped publisher - messed with model
- Dropped developer - messed with model
- Dummied console
- Dummied genre





SCORE
01

HIGH-SCORE
0000

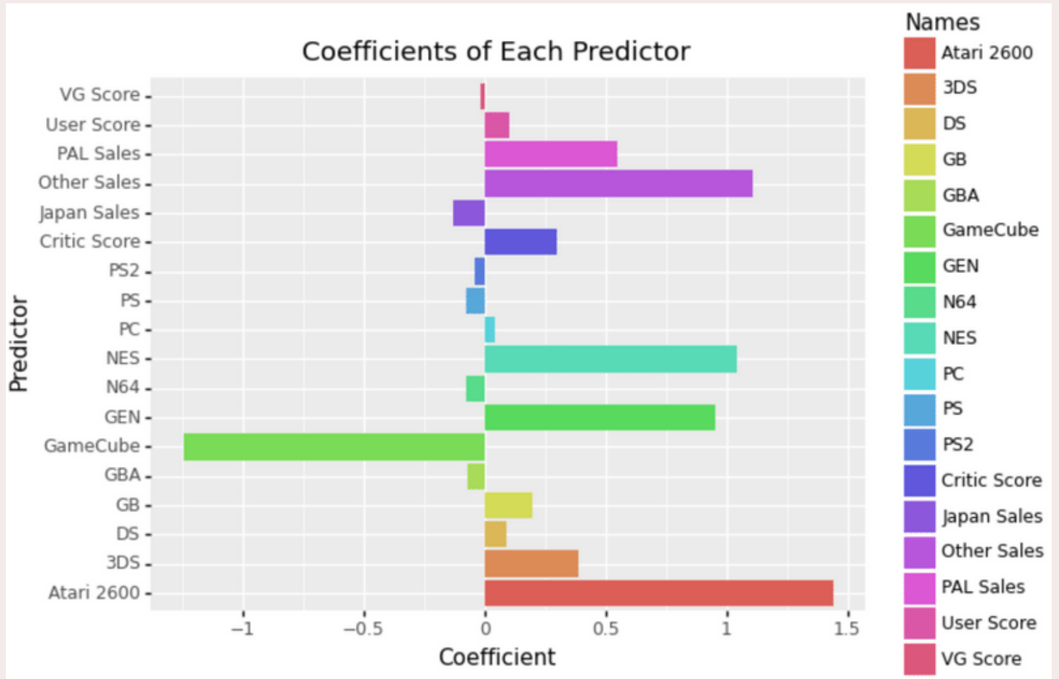
QUESTION 1

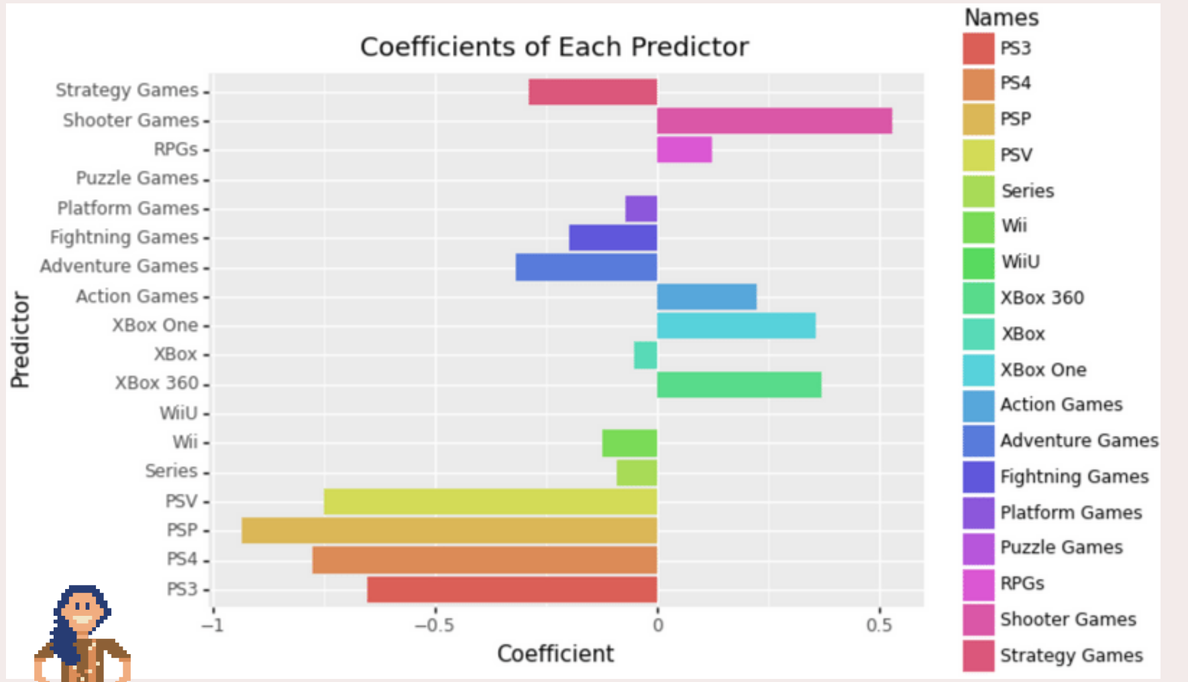
Looking at the coefficients, which variables (of console, genre, publisher, developer, vg score, critic score, user score) have the strongest relationship on North American sales?

METHOD

- 80/20 TTS
- Z-Scored continuous variables
- Fit Linear Model on Training Data
- MSE and R2 scores for both training and testing sets
  - MSE Train: ~0.6473
  - MSE Test: ~0.4636
  - R2 Train: ~0.7561
  - R2 Test: ~0.5697
- Model is somewhat fit with testing MSE lower than training MSE
- R2 is not the best but model has some accuracy

COEFFICIENTS CHART





STRONGEST COEFFICIENT: 1.440 (ATARI 2600)

SCORE
02

QUESTION 2

HIGH-SCORE
0000

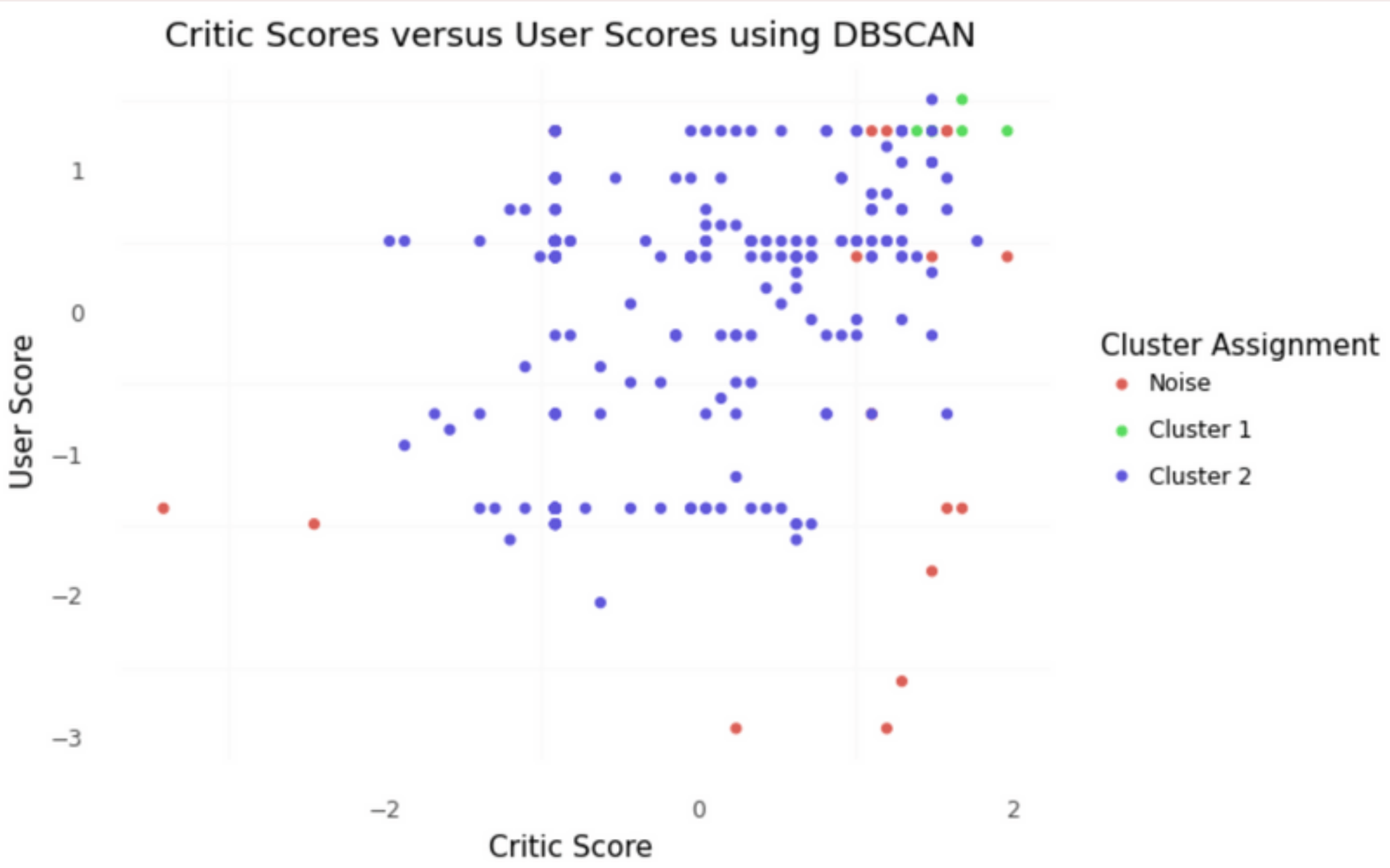
When considering critic score, user score, and North American sales, what clusters emerge and what characterize those clusters?

METHOD

- The **DBSCAN** model does not make any assumptions about the shape of the clusters.
- Good for data with noise, **Bad** for overlapping data points
- For the Video Game dataset, **DBSCAN** is the best since the dataset has somewhat noticeable noise points and the other clustering methods seem worse for this data
- EPS chosen using elbow method
- Average silhouette score: ~0.5062
    - Not the best but also not the worst

CLUSTERING PLOT 1

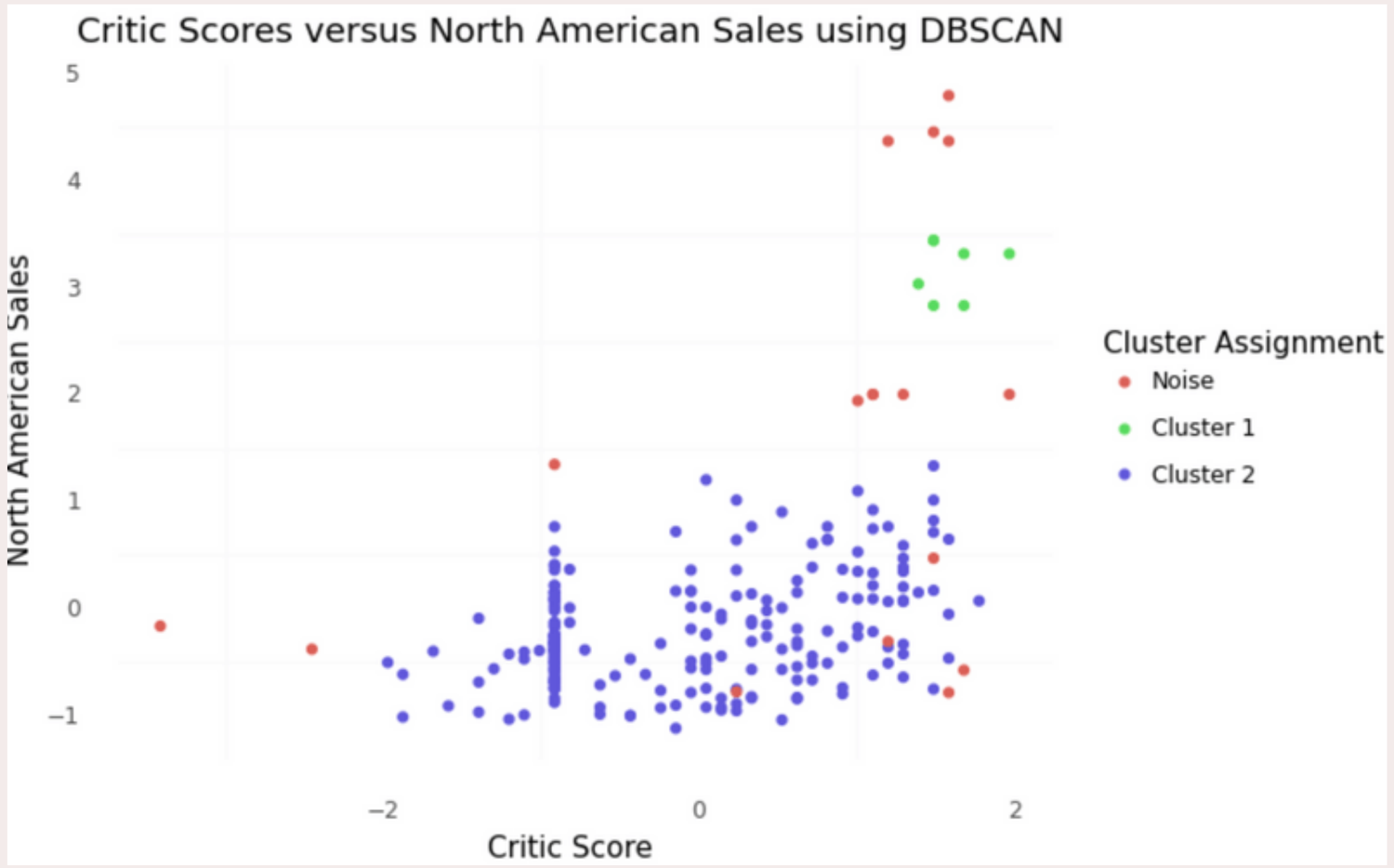
Cluster One appears to be games that are very popular with everyone and critic alike.



Cluster Two appears to be games that are somewhat popular with critics but more popular with everyone

CLUSTERING PLOT 2

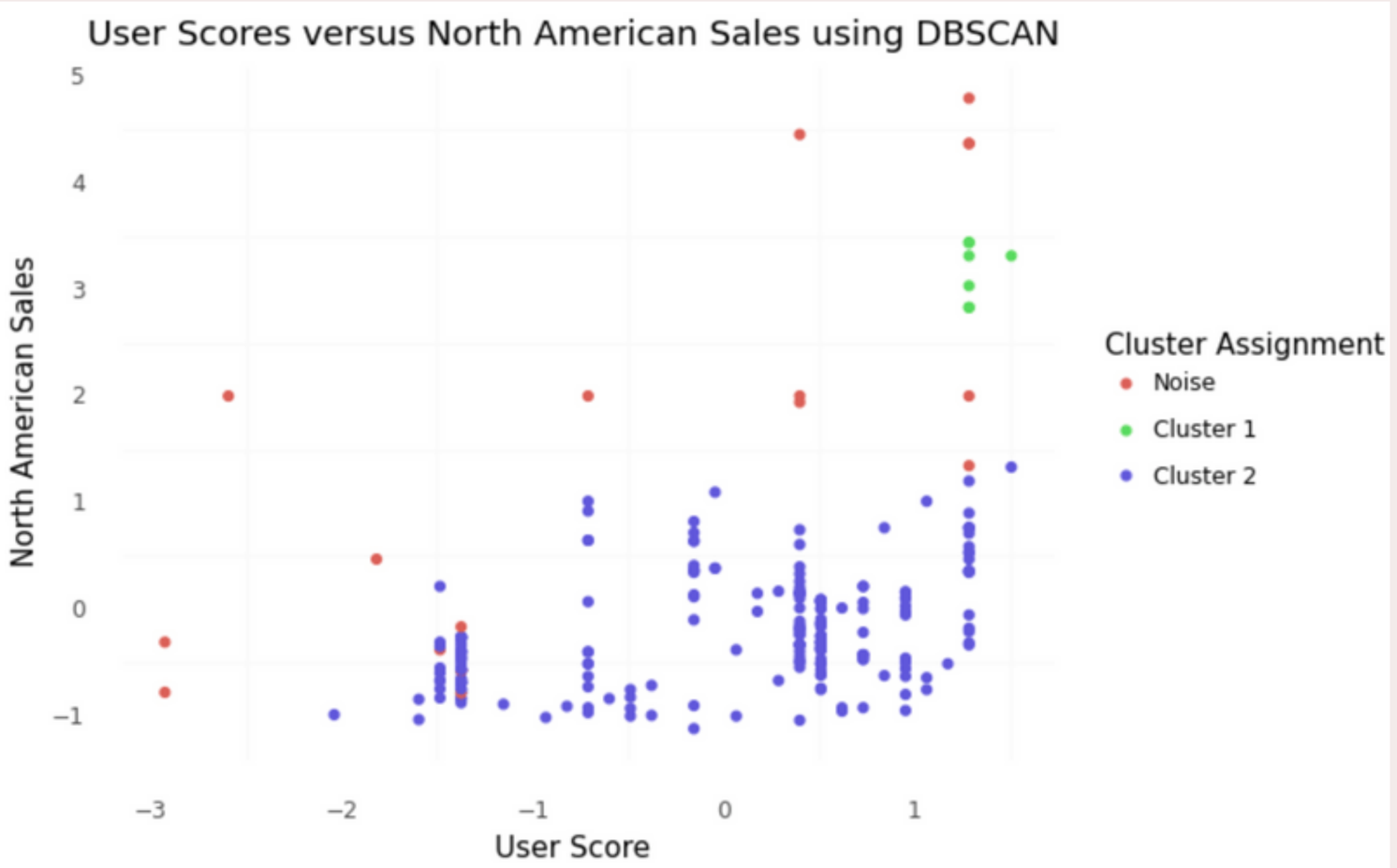
Cluster One appears to be games that are very popular with critics and had high sales in North America



Cluster Two appears to be games that are somewhat popular with critics and had average to somewhat high sales

CLUSTERING PLOT 3

Cluster One appears to be games that are very popular with general audiences and had high sales in North America



Cluster Two appears to be games that are more popular with general audiences but had average to somewhat high sales

SCORE
03

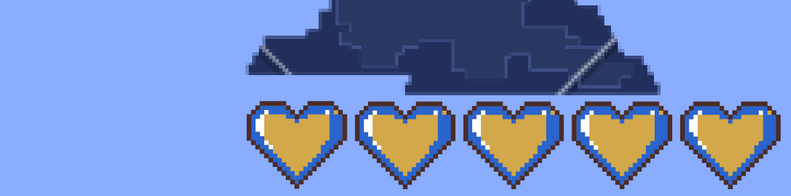
QUESTION 3

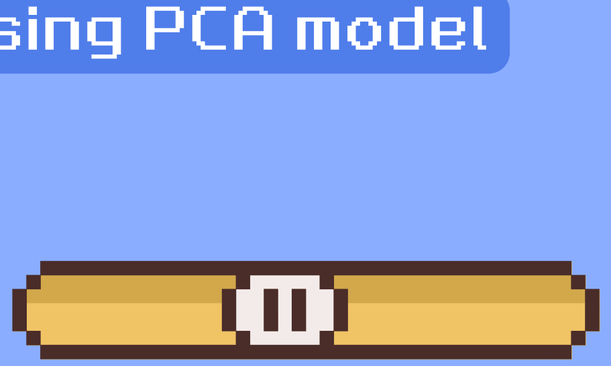
HIGH-SCORE
0000

When comparing a model using PCA on all the continuous variables(except critic score) in the dataset and retaining enough PCs to keep 85% of the variance, to a model using all the continuous variables(except critic score), how much of a difference is there in mean squared error when predicting critic score?

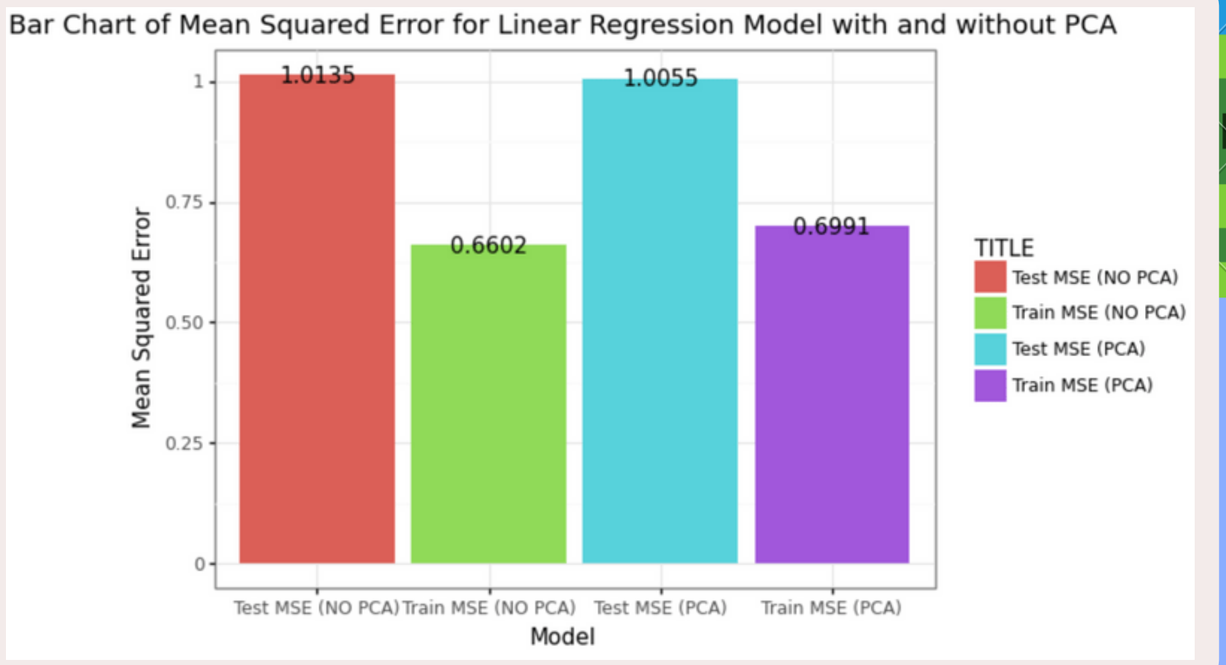
METHOD

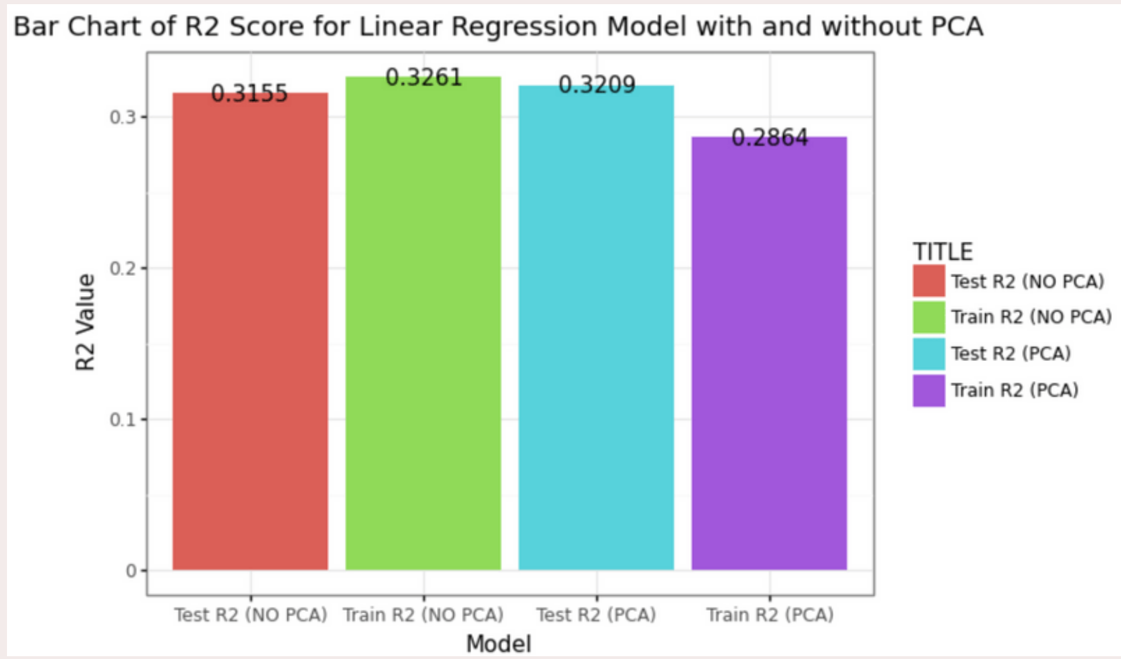
- 80/20 TTS
- Z-Scored continuous variables
- Fitted training set on Linear Model without PCA
- MSE and R2 scores for both training and testing set
  - Training R2: ~0.3261
  - Test R2: ~0.3155
  - Train MSE: ~0.6602
  - Test MSE: ~1.0135
- Model appears to be somewhat fit and not extremely accurate
- Linear PCA model with 3 components (from SCREE plot)
- MSE and R2 for both training and testing set using PCA model
  - Training R2: ~0.2864
  - Test R2: ~0.3209
  - Train MSE: ~0.6991
  - Test MSE: ~1.0055





MSE & R2 CHARTS





Difference of ~0.1 for Test
Difference of ~0.3 for Train

Difference of ~0.1 for Test
Difference of ~0.3 for Train

SUMMARY

- Question 1
  - ATARI 2600 console is the strongest predictor
  - Unusual since the ATARI is outdated
  - Possibly only include consoles/games from the year 2000 and forward
- Question 2
  - PLOT 1 shows that user and critic scores are fairly similar and in the same cluster
  - PLOT 2 shows that higher critic scores and higher North American sales are clustered together
  - PLOT 3 shows that higher user scores and higher North American sales are clustered together
  - Logically makes sense, if the general audience and critic enjoyed a game more, then sales should be higher for that game
- Question 3
  - When comparing a model using PCA on all the continuous variables(except critic score) in the dataset and retaining enough PCs to keep 85% of the variance, to a model using all the continuous variables(except critic score),there is a small difference (~0.1-~0.3) in mean squared error when predicting critic score


THANK YOU!
PLAY AGAIN?
YES
NO