Movies

CPSC 392 Final Project

By: Ryan King

Question 1

Of the variables year, gross, votes, budget, runtime, and the various movie genres, which ones have the strongest relationship with a movie's IMDb score?

Variables Used

- Year
- Gross
- Votes
- Budget
- Runtime
- genre_Action
- genre_Adventure
- genre_Biography
- genre_Comedy
- genre _Crime
- genre_Drama
- genre_Family
- genre_Fantasy
- genre_Horror
- genre_Mystery
- genre_Romance
- genre_Sci-Fi
- genre_Thriller
- genre_Western

Steps

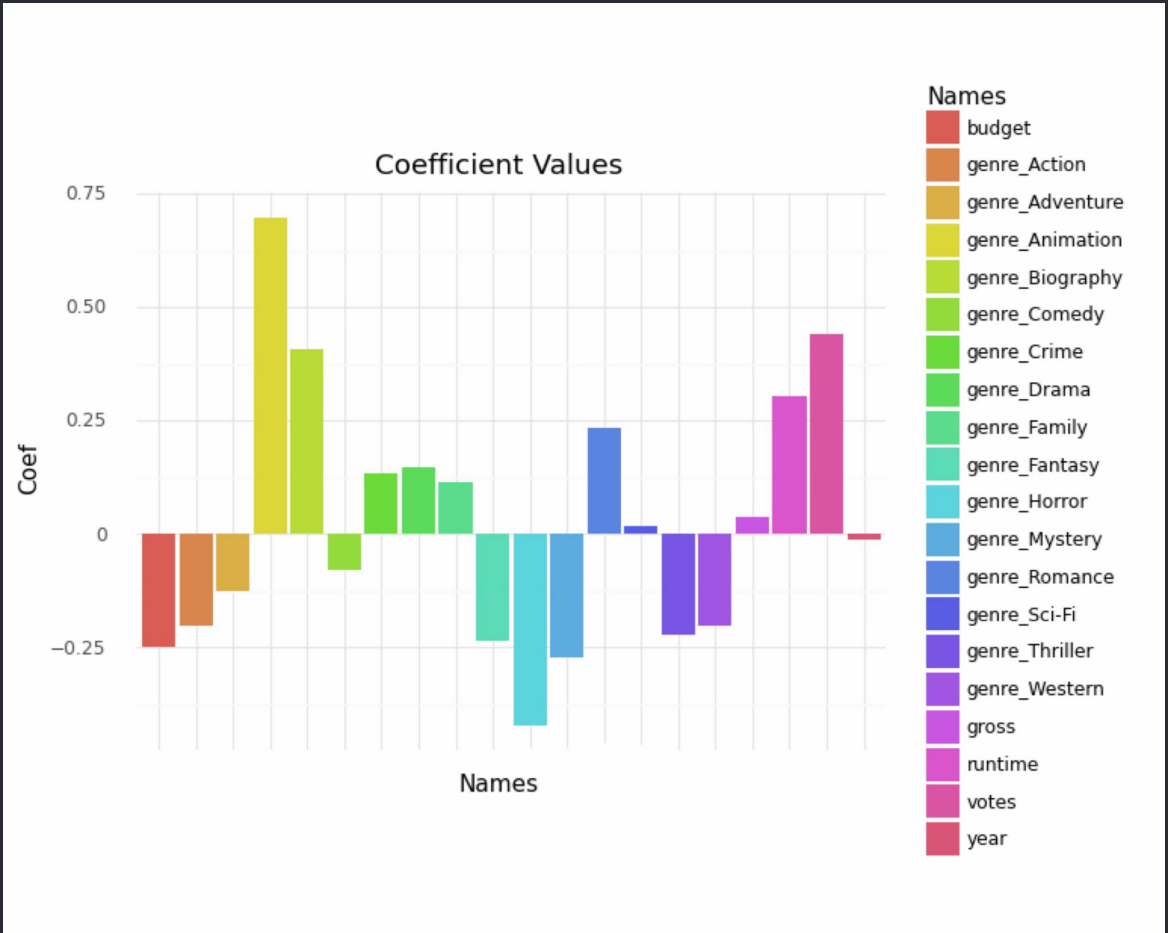
STEP 1 — Train/Test split with an 80/20 split

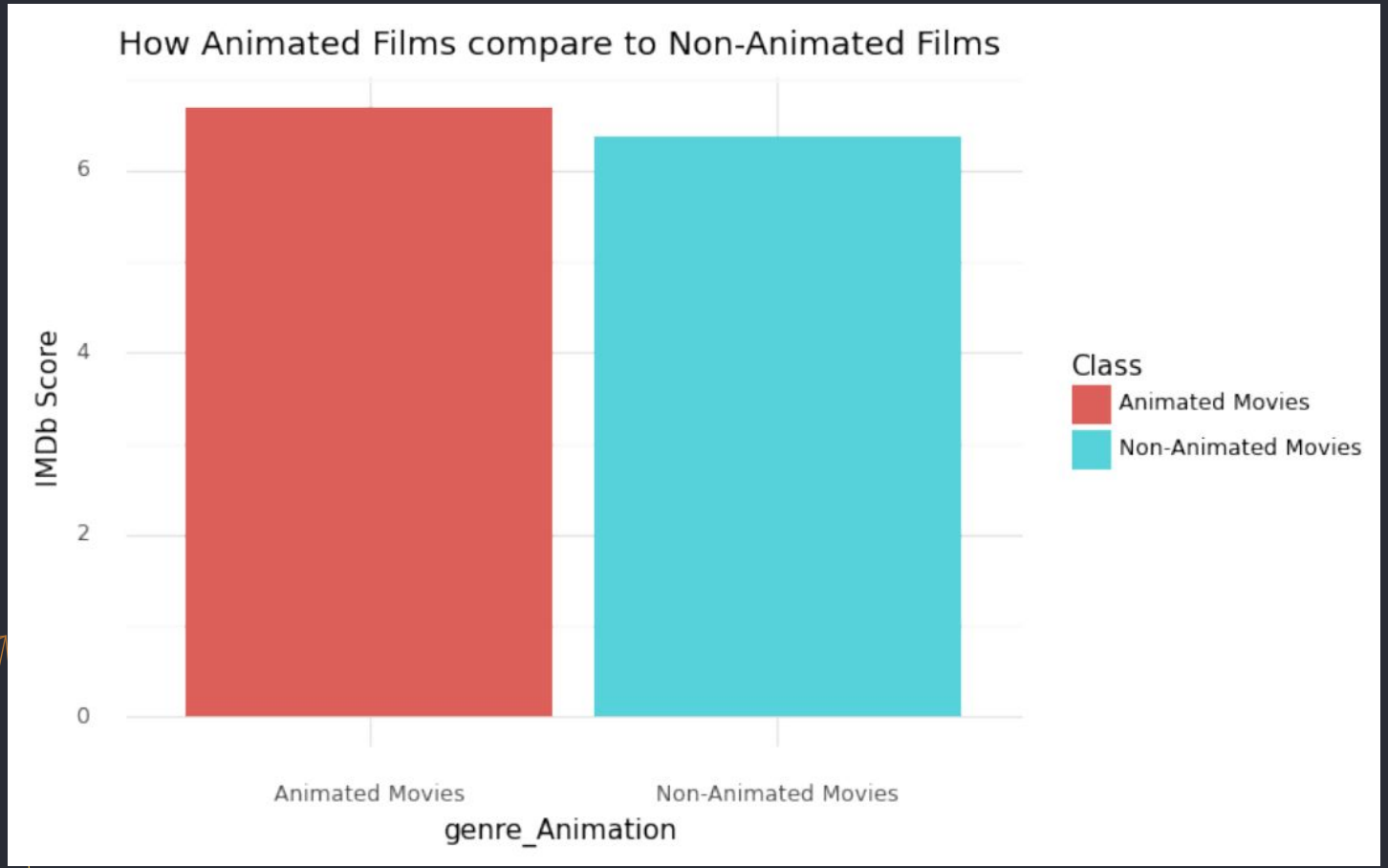
STEP 2 — Z-Scored variables

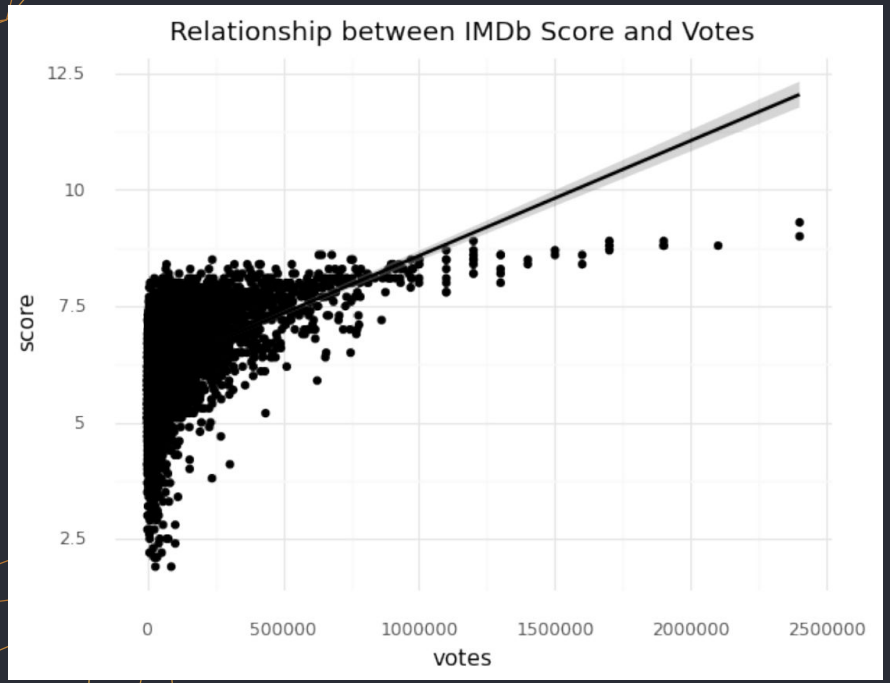
STEP 3 — Fit linear regression model

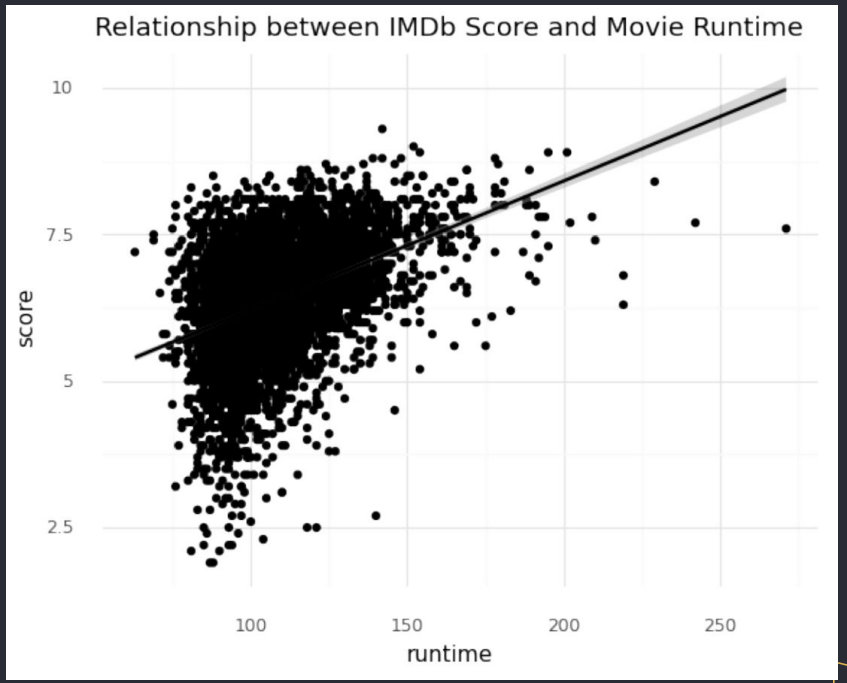
STEP 4 — Stored all variable coefficients in a dataframe

STEP 5 — Created a plot to display those coefficients


Coefficient Values
0.75
0.50
0.25
0
−0.25
Coef
Names
Names
budget
genre_Action
genre_Adventure
genre_Animation
genre_Biography
genre_Comedy
genre_Crime
genre_Drama
genre_Family
genre_Fantasy
genre_Horror
genre_Mystery
genre_Romance
genre_Sci-Fi
genre_Thriller
genre_Western
gross
runtime
votes
year


How Animated Films compare to Non-Animated Films
IMDb Score
6
4
2
0
Animated Movies
Non-Animated Movies
genre_Animation
Class
Animated Movies
Non-Animated Movies


Relationship between IMDb Score and Votes
score
votes
12.5
10
7.5
5
2.5
0
500000
1000000
1500000
2000000
2500000


Relationship between IMDb Score and Movie Runtime
score
runtime
10
7.5
5
2.5
100
150
200
250

Question 2

When comparing a model using PCA on all the continuous variables other than score, in the dataset and retaining enough PCs to keep 90% of the variance, to a model using all the continuous variables other than score, what is the difference in the mean squared error when predicting the IMDb score of a movie?

Variables Used

- Year
- Gross
- Votes
- Budget
- Runtime

Steps

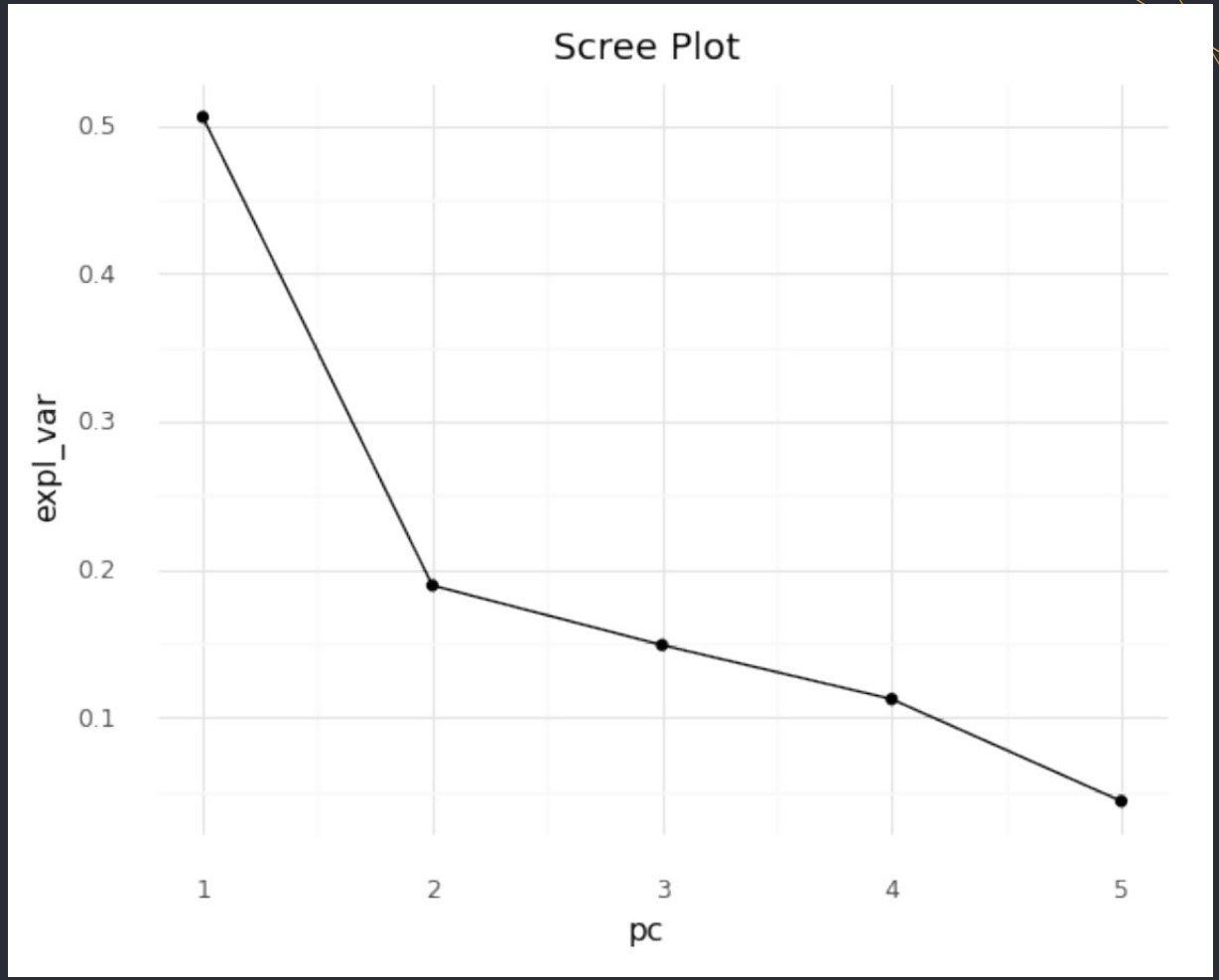
STEP 1 — Created original model to compare PCA model to

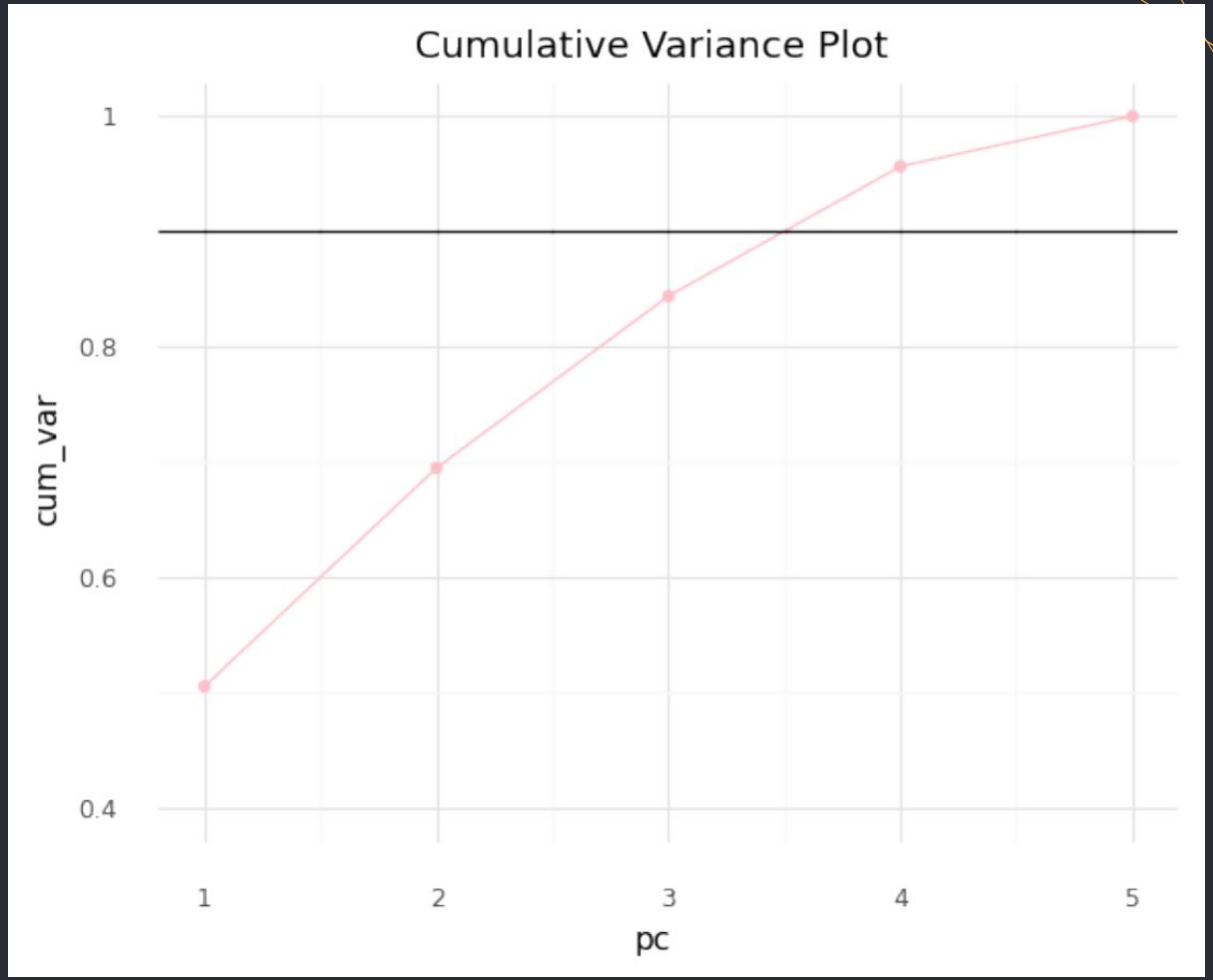
STEP 2 — Calculated MSE values of original model

STEP 3 — Produced Scree and Cumulative Variance plots to get PC values

STEP 4 — Created PCA models using PC values

STEP 5 — Compared MSE values of all models


Scree Plot
0.5
0.4
0.3
0.2
0.1
expl_var
1
2
3
4
5
pc


Cumulative Variance Plot
1
0.8
0.6
0.4
cum_var
1
2
3
4
5
pc

MSE Values Produced

Original Model

- Train MSE: 0.611
- Test MSE: 0.619

2 PCs Model

- Train MSE: 0.865
- Test MSE: 0.842

4 PCs Model

- Train MSE: 0.654
- Test MSE: 0.683

Question 3

When considering the variables movie gross, score, and budget, what clusters are shown, and describe what those clusters mean for those groups of movies?

Variables Used

- Gross
- Budget
- Score

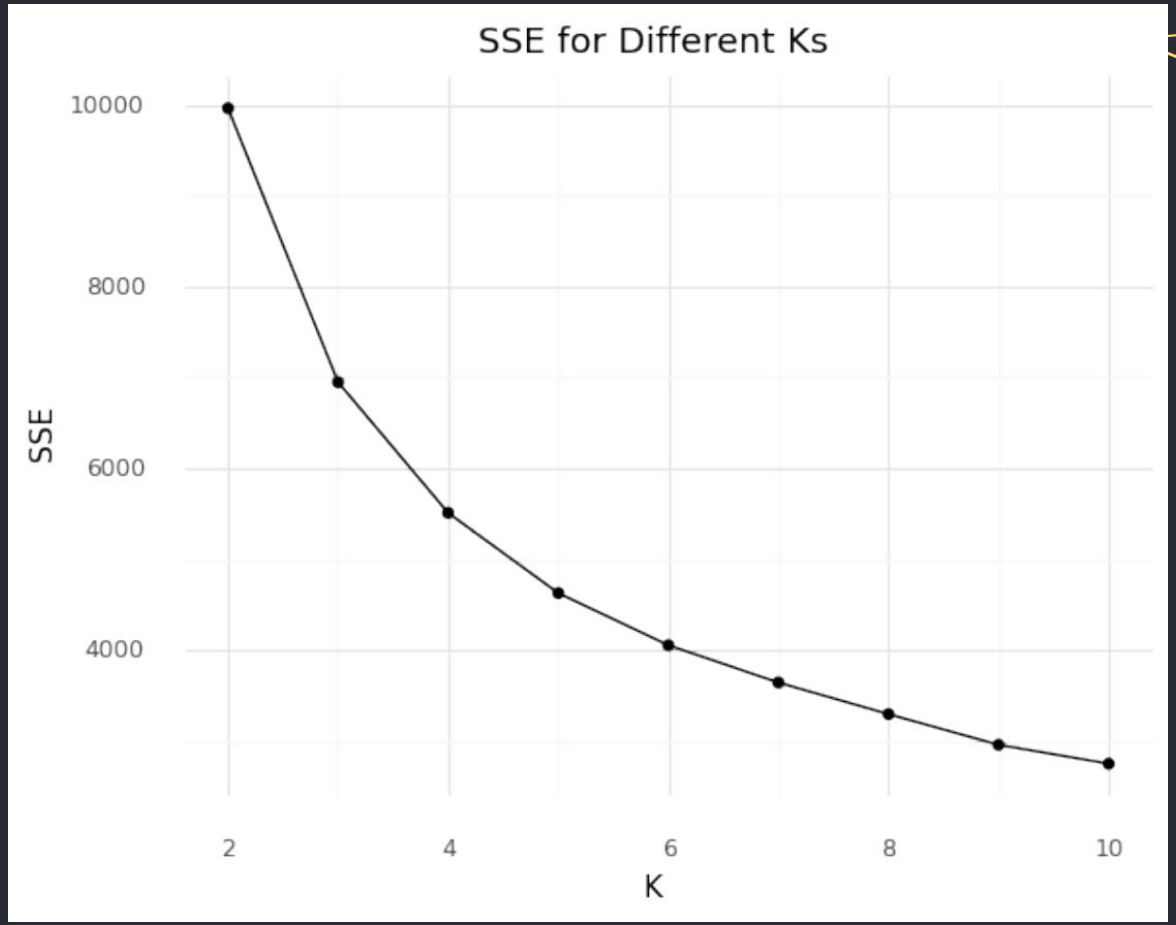
Steps

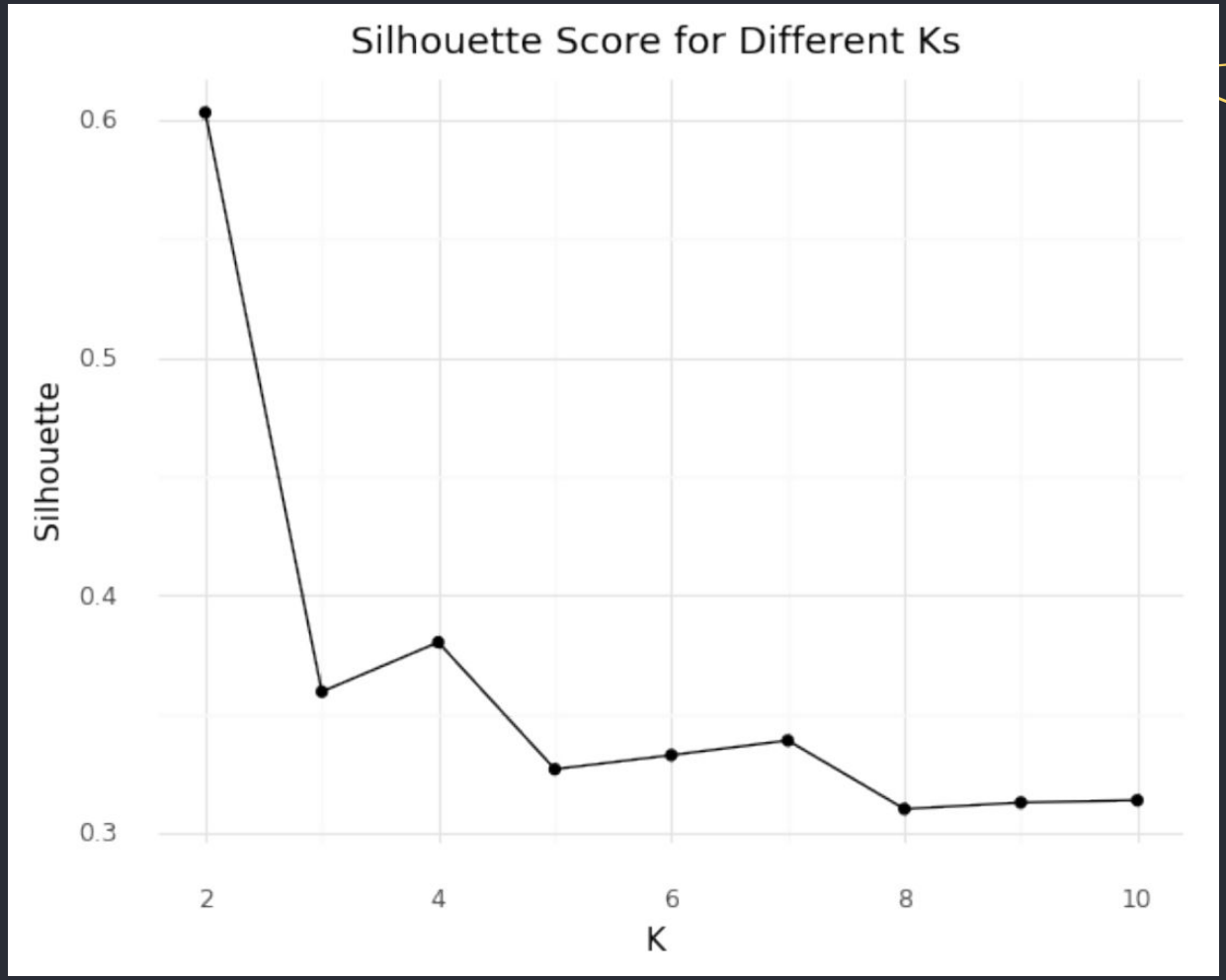
STEP 1 — Created a graph of the Sum of Squared Errors for different K values

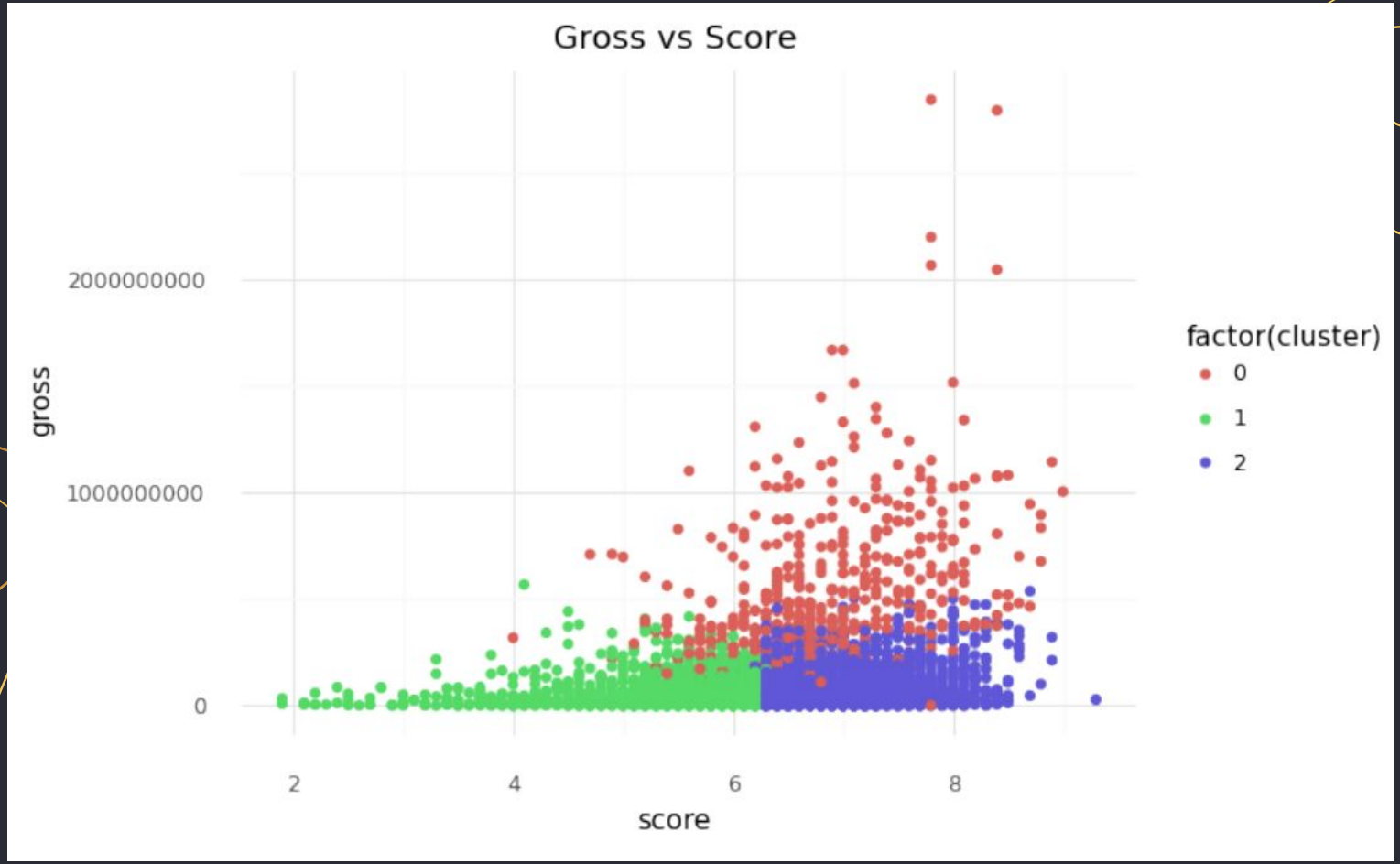
STEP 2 — Created a graph for Silhouette Scores for different K values

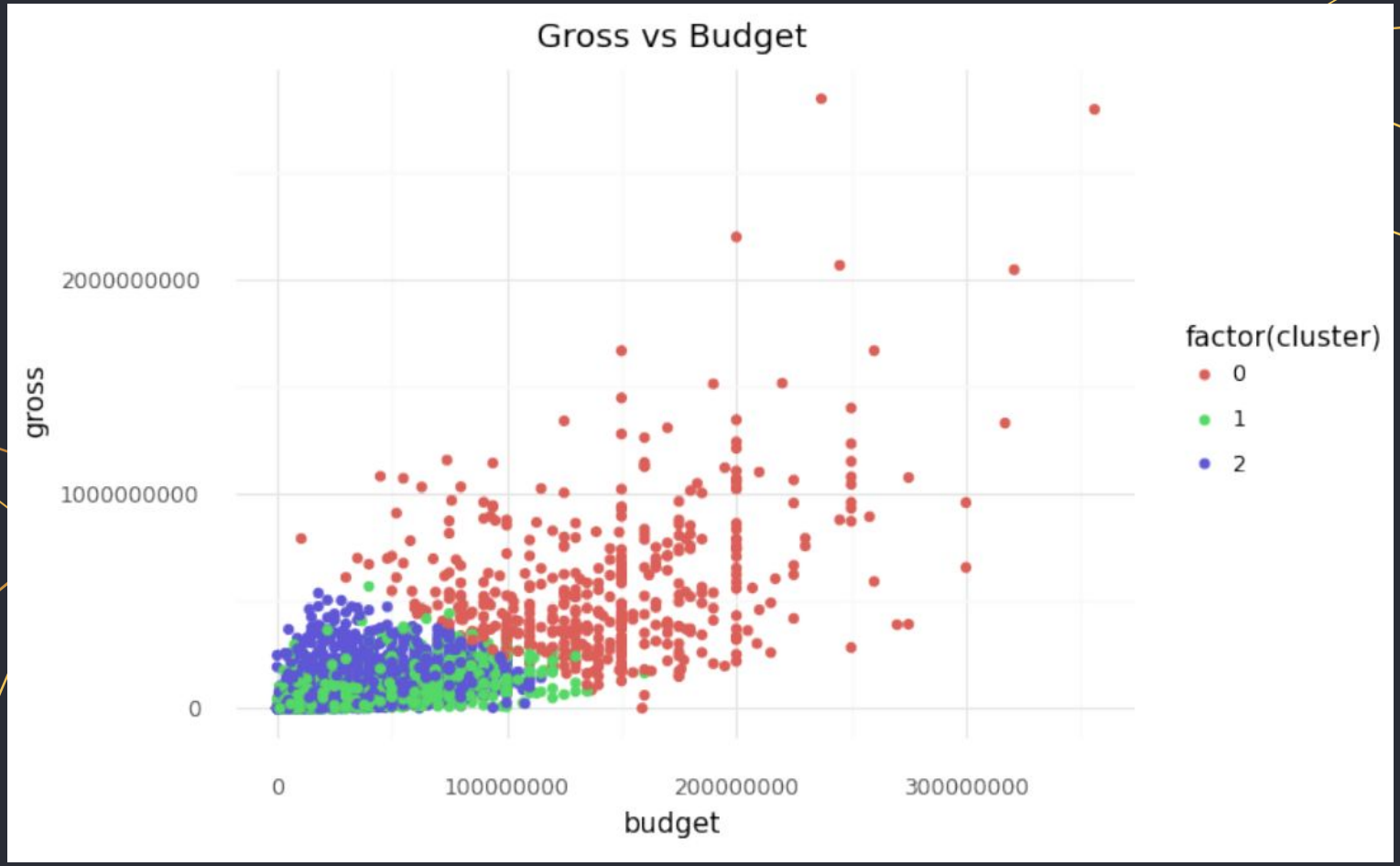
STEP 3 — Used graphs to determine there are 3 clusters

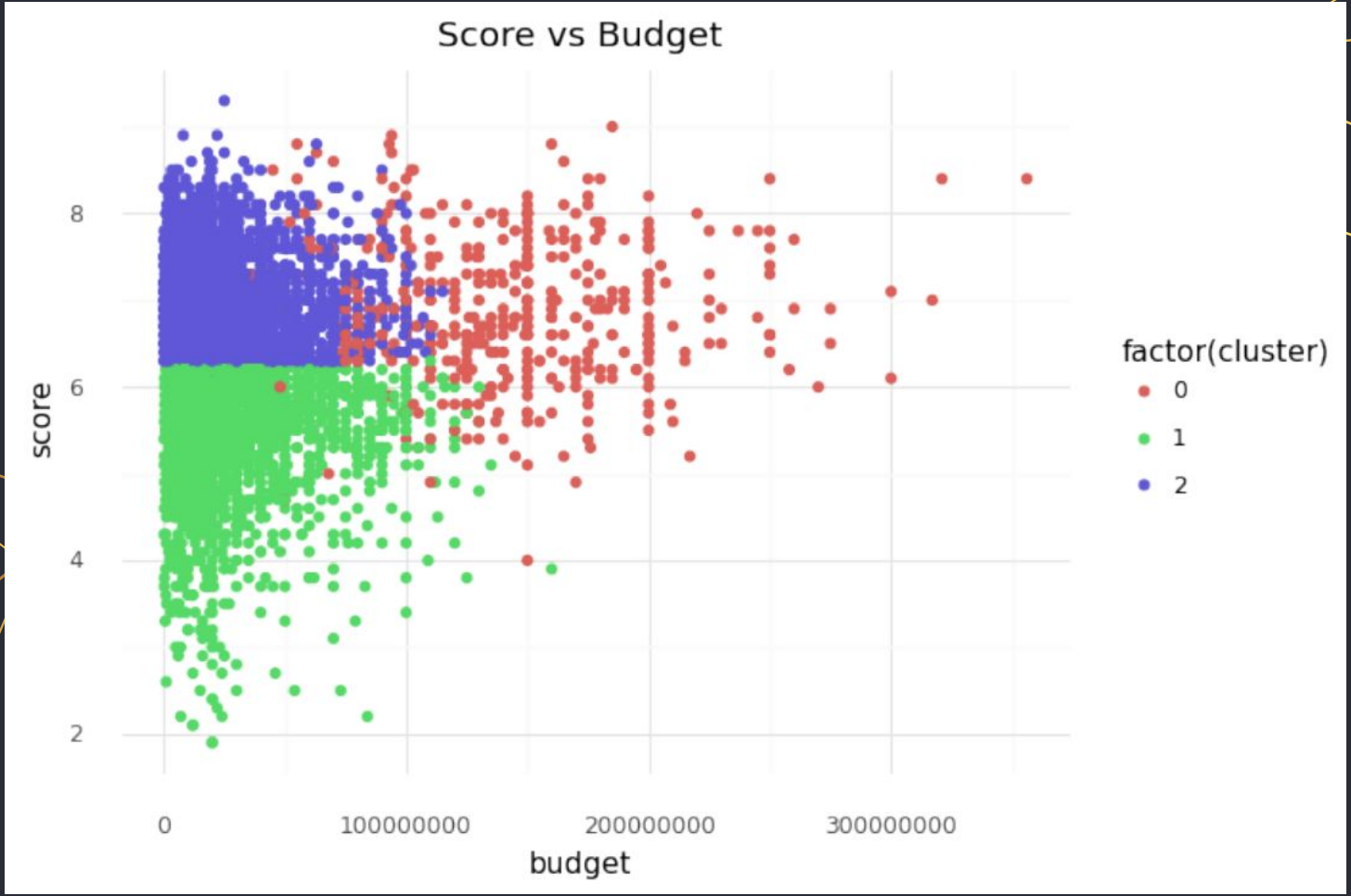
STEP 4 — Produced plots to show the clusters within the different variables relationships


SSE for Different Ks
10000
8000
6000
4000
SSE
2
4
6
8
10
K


Silhouette Score for Different Ks
Silhouette
0.6
0.5
0.4
0.3
2
4
6
8
10
K


Gross vs Score
gross
2000000000
1000000000
0
2
4
6
8
score
factor(cluster)
0
1
2


Gross vs Budget
gross
2000000000
1000000000
0
0
100000000
200000000
300000000
budget
factor(cluster)
0
1
2


Score vs Budget
score
8
6
4
2
0
100000000
200000000
300000000
budget
factor(cluster)
0
1
2

Thank you!