Final_Exam

Khutso Ledwaba

2022-12-12

#Calling the dataset and libraries

```
library(ggplot2)
library(tidyverse)
library(caret)
library(caretEnsemble)
library(psych)
library(Amelia)
library(mice)
library(GGally)
library(rpart)
library(randomForest)
library("class")
library("factoextra")
library("ggpubr")
library("esquisse")
library("dplyr")


#call movies data file
Movies <- read.csv("netflix_titles.csv")
#omiting the NA's and selecting the main columns needed for the prediction
NewMovies <- Movies %>% select(1: 8)

#Studying the structure of the data
str(NewMovies)
```

```
## 'data.frame':    8807 obs. of  8 variables:
##  $ ShowID                          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ IMDB.rating                     : int  4 4 1 5 2 1 4 4 3 5 ...
##  $ US.Budget..in.millions.         : int  2623300 11242682 14647850 19792116 8566378 8051932 1140540 1
##  $ Genre                           : chr  "Comedy" "Horror" "Action" "Horror" ...
##  $ World.Wide.Box.office.gross     : int  471418365 1741631103 1959078633 1502221815 398505307 872602
##  $ Trailer.audience.views..weekly.: int  6652 84932 45810 81670 75333 82840 12784 38074 86802 31778
##  $ type                            : chr  "Movie" "TV Show" "TV Show" "TV Show" ...
##  $ title                           : chr  "Dick Johnson Is Dead" "Blood & Water" "Ganglands" "Jailbir
```

```
#Top five rows
head(NewMovies, n= 5)
```

```
##   ShowID IMDB.rating US.Budget..in.millions.  Genre World.Wide.Box.office.gross
```

```
## 1      1           4                2623300 Comedy                   471418365
## 2      2           4               11242682 Horror                  1741631103
## 3      3           1               14647850 Action                  1959078633
## 4      4           5               19792116 Horror                  1502221815
## 5      5           2                8566378 Action                   398505307
##   Trailer.audience.views..weekly.    type                title
## 1                            6652   Movie  Dick Johnson Is Dead
## 2                           84932 TV Show         Blood & Water
## 3                           45810 TV Show             Ganglands
## 4                           81670 TV Show Jailbirds New Orleans
## 5                           75333 TV Show          Kota Factory
```

```
#Bottom 5 rows
tail(NewMovies, n= 5)
```

```
##      ShowID IMDB.rating US.Budget..in.millions.   Genre
## 8803   8803           4                 3759298  Horror
## 8804   8804           1                 1971205 Romance
## 8805   8805           1                 9002993  Action
## 8806   8806           3                 7191738  Horror
## 8807   8807           3                 6830566  Action
##      World.Wide.Box.office.gross Trailer.audience.views..weekly.    type
## 8803                  1224111126                           80305   Movie
## 8804                   792492475                           80328 TV Show
## 8805                   637851571                           76948   Movie
## 8806                   367744732                           97676   Movie
## 8807                  1259065384                           60662   Movie
##            title
## 8803      Zodiac
## 8804 Zombie Dumb
## 8805  Zombieland
## 8806        Zoom
## 8807      Zubaan
```

#Show the number of movies and tv shows

```
table(NewMovies['type'])
```

```
## type
##   Movie TV Show
##    6131    2676
```

```
#Netflix has more movies than Tv shows
```

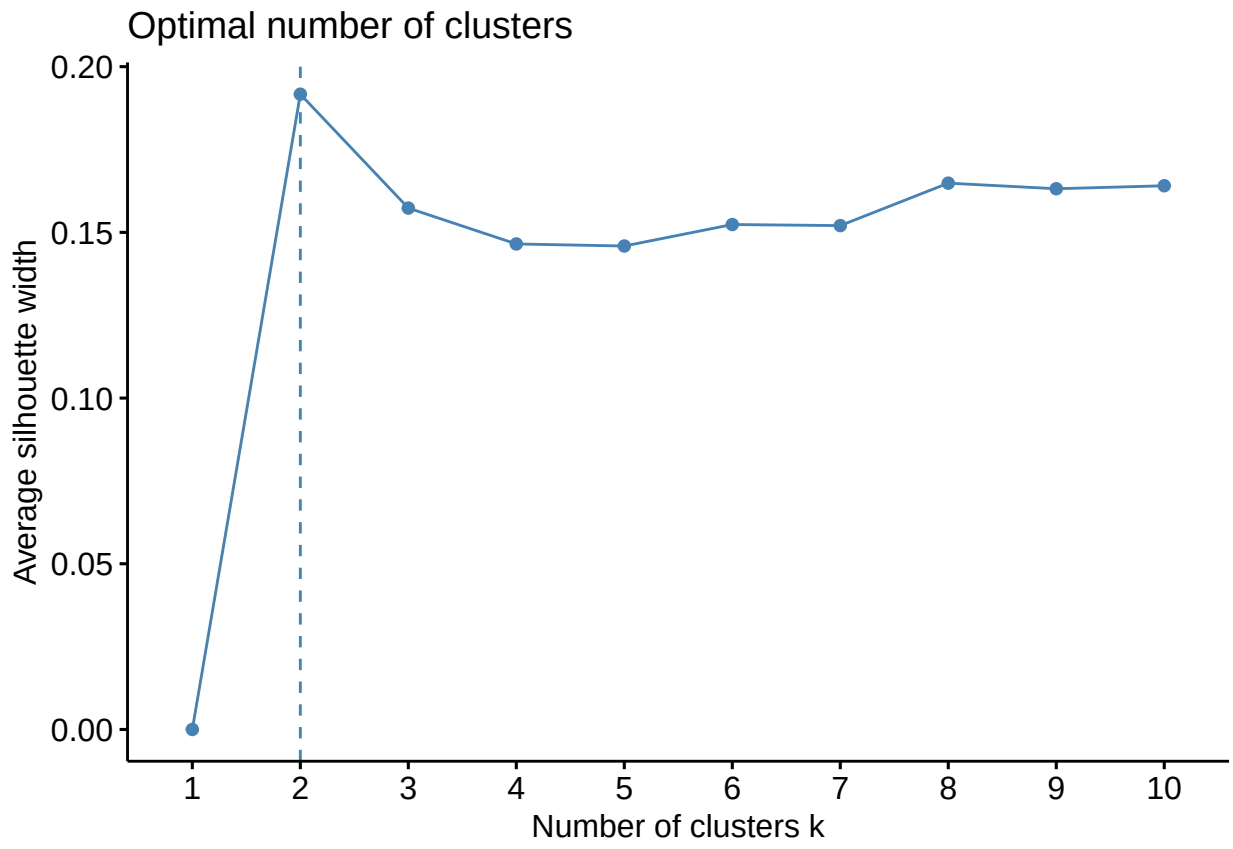
#Partioning the data into training and test

```
NM <- createDataPartition(NewMovies$Trailer.audience.views..weekly.,p=0.7,list=F)

Movies_Train <- NewMovies[NM,]
Movies_Test <- NewMovies[-NM,]

NM_Normal <- preProcess(Movies_Train[,-c(4,7:8)],method="range")
```

```
NM_Train <- predict(NM_Normal, Movies_Train)
NM_Test <- predict(NM_Normal,Movies_Test)
```

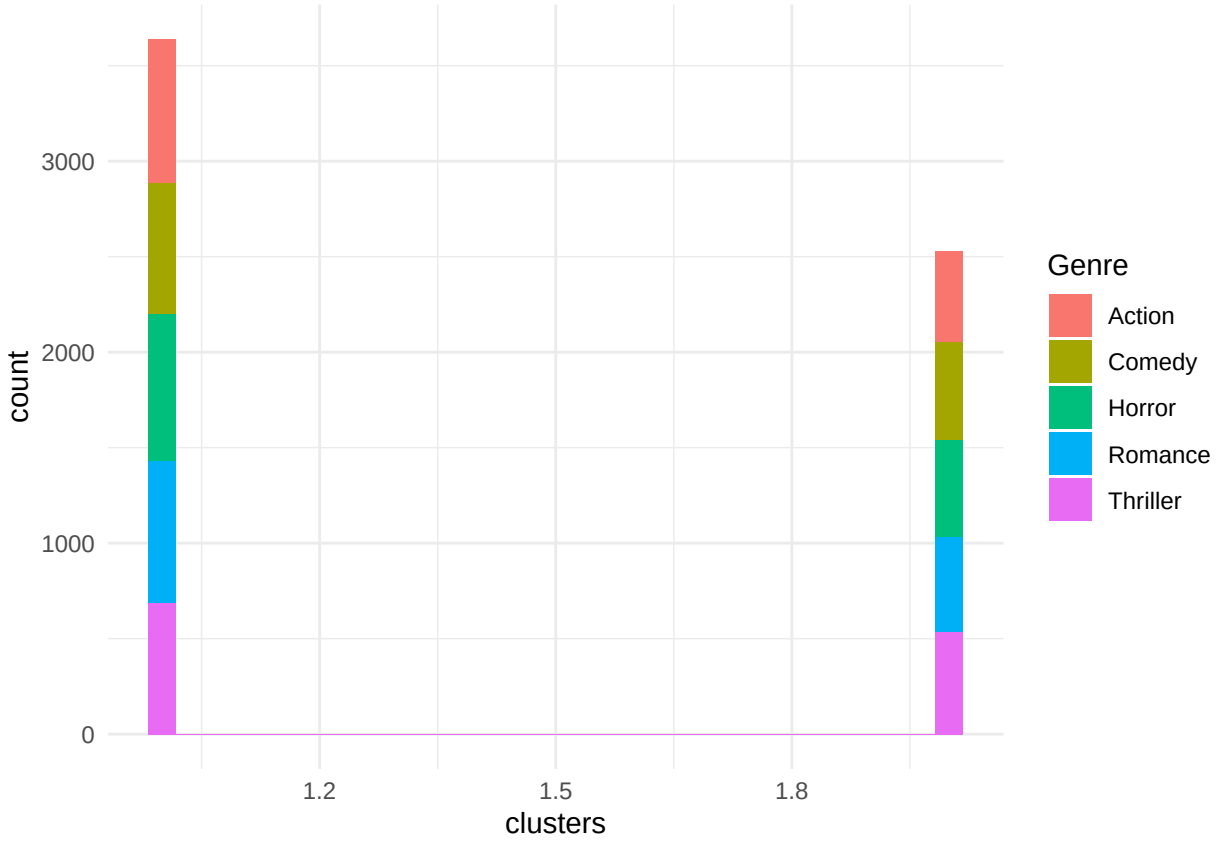
#Finding the optimal k value

```
Movies_plot <- fviz_nbclust(NM_Train[,-c(4,7:8)],kmeans,method="silhouette")
Movies_plot
```



#Using Kmeans to form the clusters

```
Movies_k <- kmeans(NM_Train[,-c(4,7:8)],centers=2,nstart=25)

Movies_Train$clusters <- Movies_k$cluster

#esquisser()
ggplot(Movies_Train) +
  aes(x = clusters, fill = Genre) +
  geom_histogram(bins = 30L) +
  scale_fill_hue(direction = 1) +
  theme_minimal()
```




```
#This plot tells us two pieces of key information:
#Both clusters represent a strong interest in the students love for action and thriller genre movies. A
#This graph also displays that cluster 1 contains the most interest in each Genre with Action being the
```

#clustering of the different factors and interpretation of the data

```
Movies_Train %>% select(IMDB.rating,clusters) %>% group_by(clusters,IMDB.rating) %>% count()
```

```
## # A tibble: 5 x 3
## # Groups:   clusters, IMDB.rating [5]
##   clusters IMDB.rating     n
##      <int>       <int> <int>
## 1        1           3  1172
## 2        1           4  1243
## 3        1           5  1223
## 4        2           1  1260
## 5        2           2  1269
```

```
Movies_Train %>% select(Genre,clusters) %>% group_by(clusters,Genre) %>% count()
```

```
## # A tibble: 10 x 3
## # Groups:   clusters, Genre [10]
##    clusters Genre        n
##       <int> <chr>    <int>
```

```
##  1        1 Action     755
##  2        1 Comedy     685
##  3        1 Horror     769
##  4        1 Romance    743
##  5        1 Thriller   686
##  6        2 Action     478
##  7        2 Comedy     512
##  8        2 Horror     511
##  9        2 Romance    493
## 10        2 Thriller   535
```

```
#Interpretation of the data

#Based on the findings of the data, The IMDB ratings are highest in the cluster one group which is indi

#The university can base their approach on which movies to stream not only on the genre of well liked m
```

#Aggretion of the clusters to draw final conclusion

```
aggregate(Movies_Train[,-c(4,7:8)], by=list(Movies_Train$clusters),FUN="median")
```

```
##   Group.1 ShowID IMDB.rating US.Budget..in.millions.
## 1       1   4396           4                10280289
## 2       2   4446           2                10775995
##   World.Wide.Box.office.gross Trailer.audience.views..weekly. clusters
## 1                  1058599579                          49527.5        1
## 2                  1067518434                          49740.0        2
```

#Conclusion

```
#From this data the university will be able to gain a general sense of what movies/shows they can relea
#Just because a movie has a high trailer view, it doesn't mean the movie will be watched.
#Multiple factors go into a student's movie choice. What is the genre of the film? How well is it rated
#Another factor is the budget and box office sales. Cluster 2 movies had the highest budget, but they w

#Based on the data if the university had to pick a genre, Thriller would be the best option because it
```