PAC Report_Shanrong Zhou

Shanrong Zhou

2022-11-18

This report summarizes and illustrates the data analysis process that I've done for Predictive Analysis Competition Project.

Load Packages

Before I start doing anything, I loaded all the packages needed first, so that I won't accidentally load any packages for multiple times.

```
#data exploration and tidying
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(stringr)
library(caTools)
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(skimr)
library(ggcorrplot)
#intersect
library(base)
#create dummy variable
library(vtreat)
```

```
## Loading required package: wrapr
```

```
##
## Attaching package: 'wrapr'
```

```
## The following objects are masked from 'package:tidyr':
##
##     pack, unpack
```

```
## The following object is masked from 'package:dplyr':
##
##     coalesce
```

```
#regression tree
library(rpart); library(rpart.plot)
#bag ipred
library(ipred)
#random Forest
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
#ranger
library(ranger)
```

```
##
## Attaching package: 'ranger'
```

```
## The following object is masked from 'package:randomForest':
##
##     importance
```

```
#gbm
library(gbm)
```

```
## Loaded gbm 2.1.8.1
```

```
#xgboost
library(xgboost)
```

```
##
## Attaching package: 'xgboost'
```

```
## The following object is masked from 'package:dplyr':
##
##     slice
```

```
#box cox transformation and svm
library(e1071)
```

Data Exploration

Read Data

The first step is to set up the environment and import analysisData and scoringData.

```
setwd("/Users/alicezhou/Documents/Columbia/5200 Applied Analytics Frameworks & Methods I/Notes/PAC")
data = read.csv("analysisData.csv")
sdata = read.csv(file="scoringData.csv")
```

Variable Overview

Summary and comparison of two data sets are shown below. Here are some highlights and my thoughts for further analysis.

analysisData:

1. There are 19 variables, with 1 identifier (id) and 1 outcome / dependent variable (rating). The other variables can all be the potential predictors / independent variables to predictive models.
2. Outcome variable (rating) is continuous. Only models for continuous outcome variable should be considered, while logistics regression and classification trees can be excluded.
3. Predictors include categorical and continuous variables.
4. It's easier to explore the numeric and logical data types' characteristics by reading the overall summary. I can get a preliminary understanding, in terms of dispersion (standard deviation, interquartile range and range) for numeric variables and central tendency (mean) for both numeric and logical variables.
5. The character data type has more uniqueness and is less structured than the others.
    - song: It has 16,542 unique values, which is a similar amount to number of observations 19,485. Almost each song has its own name. This variable is too unique to be predictive. Therefore, I won't include it as the predictor for future modelling.
    - genre: It has only 2,937 unique values. However, every value may contain several genres of a song. I am going to remove the special characters, and transform it to new categorical variables, so that each genre type can be represented by a category individually. Besides, since it contains missing value, I will fill the missing value as 0 before data transformation.
    - performer: It has 6,687 unique values. Around 1/3 of songs have their unique performers. It's reasonable to be considered as a predictive variable now.

Comparison between analysisData and scoringData:

1. Compared to the variables of analysisData, only outcome variable (rating) is missed for scoringData. All the predictors are in same data type. We can use all the predictors in analysisData to train the predictive model, to predict rating of songs in scoringData.
2. There are significantly less rows for scoringData than analysisData. Some characters of genre and performer can be different from analysisData. Thus, I will only include the inter-sets of both data sets for these variables to data analysis.

```
str(data)
```

```
## 'data.frame':    19485 obs. of  19 variables:
##  $ id               : int  94500 64901 28440 19804 83560 16501 58033 67048 48848 95622 ...
##  $ performer        : chr  "Andy Williams" "Sandy Nelson" "Britney Spears" "Taylor Swift" ...
##  $ song             : chr  "......And Roses And Roses" "...And Then There Were Drums" "...Baby One More Time"
"...Ready For It?" ...
##  $ genre            : chr  "['adult standards', 'brill building pop', 'easy listening', 'mellow gold']" "['ro
ck-and-roll', 'space age pop', 'surf music']" "['dance pop', 'pop', 'post-teen pop']" "['pop', 'post-teen po
p']" ...
##  $ track_duration   : num  166106 172066 211066 208186 182080 ...
##  $ track_explicit   : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ danceability     : num  0.154 0.588 0.759 0.613 0.45 0.57 0.612 0.253 0.575 0.615 ...
##  $ energy           : num  0.185 0.672 0.699 0.764 0.294 0.629 0.542 0.232 0.434 0.497 ...
##  $ key              : int  5 11 0 2 7 9 5 0 5 7 ...
##  $ loudness         : num  -14.06 -17.28 -5.75 -6.51 -12.02 ...
##  $ mode             : int  1 0 0 1 1 0 1 1 1 1 ...
##  $ speechiness      : num  0.0315 0.0361 0.0307 0.136 0.0318 0.0331 0.0264 0.0318 0.0312 0.439 ...
##  $ acousticness     : num  0.911 0.00256 0.202 0.0527 0.832 0.593 0.0781 0.805 0.735 0.016 ...
##  $ instrumentalness: num  2.67e-04 7.45e-01 1.31e-04 0.00 3.53e-05 1.36e-04 0.00 1.80e-04 6.59e-05 0.00 ...
##  $ liveness         : num  0.112 0.145 0.443 0.197 0.108 0.77 0.0763 0.0939 0.105 0.312 ...
##  $ valence          : num  0.15 0.801 0.907 0.417 0.146 0.308 0.433 0.307 0.348 0.769 ...
##  $ tempo            : num  84 122 93 160 141 ...
##  $ time_signature   : int  4 4 4 4 4 4 4 3 4 3 ...
##  $ rating           : int  36 16 70 64 19 34 44 34 47 26 ...
```

```
skim(data)
```

Data summary

Name	data
Number of rows	19485
Number of columns	19
_____	
Column type frequency:	
character	3
logical	1
numeric	15
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
performer	0	1.00	1	113	0	6687	0
song	0	1.00	1	75	0	16542	0
genre	108	0.99	2	319	0	2937	0

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
track_explicit	0	1	0.12	FAL: 17203, TRU: 2282

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
id	0	1	50208.92	29047.37	3.00	24812.00	50487.00	75544.00	99999.00	▇▇▇▇▇
track_duration	0	1	220873.01	68749.76	29688.00	175173.00	214733.00	253306.00	3079157.00	▇▁▁▁▁
danceability	0	1	0.60	0.15	0.00	0.50	0.61	0.71	0.99	▁▂▅▇▂
energy	0	1	0.62	0.20	0.00	0.48	0.63	0.78	1.00	▁▃▅▇▅
key	0	1	5.23	3.56	0.00	2.00	5.00	8.00	11.00	▇▃▃▅▇
loudness	0	1	-8.67	3.61	-28.03	-11.04	-8.21	-5.86	2.29	▁▂▅▇▁
mode	0	1	0.73	0.44	0.00	0.00	1.00	1.00	1.00	▃▁▁▁▇
speechiness	0	1	0.07	0.08	0.00	0.03	0.04	0.07	0.92	▇▁▁▁▁
acousticness	0	1	0.29	0.28	0.00	0.05	0.19	0.51	0.99	▇▃▂▂▁
instrumentalness	0	1	0.03	0.14	0.00	0.00	0.00	0.00	0.98	▇▁▁▁▁
liveness	0	1	0.19	0.16	0.01	0.09	0.13	0.25	1.00	▇▃▁▁▁
valence	0	1	0.60	0.24	0.00	0.41	0.62	0.80	0.99	▂▅▇▇▇
tempo	0	1	120.24	27.92	0.00	99.08	119.00	136.39	241.01	▁▃▇▃▁
time_signature	0	1	3.93	0.32	0.00	4.00	4.00	4.00	5.00	▁▁▁▇▁
rating	0	1	36.69	16.55	0.00	24.00	36.00	50.00	91.00	▃▇▇▃▁

```
str(sdata)
```

```
## 'data.frame':    4844 obs. of  18 variables:
##  $ id              : int  50400 96747 1824 67597 86944 85423 5500 82675 5926 57666 ...
##  $ performer       : chr  "Paul Davis" "Luther Vandross" "The Olympics" "Maxine Nightingale" ...
##  $ song            : chr  "'65 Love Affair" "'Til My Baby Comes Home" "(Baby) Hully Gully" "(Bringing Out) T
he Girl In Me" ...
##  $ genre           : chr  "['album rock', 'bubblegum pop', 'country rock', 'folk rock', 'mellow gold', 'new
wave pop', 'soft rock', 'yacht rock']" "['funk', 'motown', 'neo soul', 'new jack swing', 'quiet storm', 'r&b',
'soul', 'urban contemporary']" "['doo-wop', 'rhythm and blues']" "['classic uk pop']" ...
##  $ track_duration  : num  219813 332226 127829 210973 180133 ...
##  $ track_explicit  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ danceability    : num  0.647 0.804 0.744 0.712 0.665 0.442 0.61 0.478 0.499 0.801 ...
##  $ energy          : num  0.686 0.714 0.47 0.753 0.552 0.717 0.377 0.298 0.249 0.875 ...
##  $ key             : int  2 11 7 9 6 9 9 2 5 7 ...
##  $ loudness        : num  -4.25 -6.71 -10.29 -7.67 -10.18 ...
##  $ mode            : int  0 0 1 1 1 0 1 1 1 1 ...
##  $ speechiness     : num  0.0274 0.183 0.035 0.0399 0.0307 0.0395 0.0935 0.0342 0.0422 0.0995 ...
##  $ acousticness    : num  0.432 0.0567 0.659 0.756 0.493 0.518 0.414 0.74 0.733 0.00366 ...
##  $ instrumentalness: num  6.19e-06 6.21e-06 0.00 1.58e-06 2.72e-02 2.62e-01 0.00 1.11e-05 0.00 4.25e-01 ...
##  $ liveness        : num  0.133 0.0253 0.256 0.169 0.421 0.184 0.0734 0.139 0.105 0.182 ...
##  $ valence         : num  0.952 0.802 0.908 0.953 0.899 0.376 0.519 0.376 0.539 0.637 ...
##  $ tempo           : num  155.7 139.7 116.2 116.2 96.9 ...
##  $ time_signature  : int  4 4 4 4 4 4 4 4 4 4 ...
```

```
skim(sdata)
```

Data summary

Name	sdata
Number of rows	4844
Number of columns	18

Column type frequency:	
character	3
logical	1
numeric	14
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
performer	0	1.00	2	72	0	2791	0
song	0	1.00	1	67	0	4604	0
genre	36	0.99	2	312	0	1702	0

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
track_explicit	0	1	0.13	FAL: 4213, TRU: 631

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
id	0	1	49953.08	28769.94	8.00	25172.75	49678.00	75322.75	99980.00	▇▇▇▇▇
track_duration	0	1	219882.00	63667.02	37013.00	174533.00	215152.50	252896.25	688453.00	▂▇▁▁▁
danceability	0	1	0.60	0.15	0.11	0.50	0.61	0.71	0.97	▁▃▇▇▂
energy	0	1	0.62	0.20	0.02	0.48	0.64	0.78	1.00	▁▃▇▇▅
key	0	1	5.24	3.56	0.00	2.00	5.00	8.00	11.00	▇▃▂▅▇
loudness	0	1	-8.63	3.55	-25.47	-11.02	-8.19	-5.86	-0.51	▁▂▅▇▃
mode	0	1	0.72	0.45	0.00	0.00	1.00	1.00	1.00	▃▁▁▁▇
speechiness	0	1	0.08	0.08	0.02	0.03	0.04	0.07	0.95	▇▁▁▁▁
acousticness	0	1	0.30	0.28	0.00	0.05	0.20	0.52	0.98	▇▃▂▂▁
instrumentalness	0	1	0.03	0.13	0.00	0.00	0.00	0.00	0.96	▇▁▁▁▁
liveness	0	1	0.19	0.15	0.01	0.09	0.13	0.24	0.99	▇▂▁▁▁
valence	0	1	0.60	0.24	0.04	0.42	0.63	0.80	0.98	▂▃▅▇▇
tempo	0	1	120.41	28.55	37.11	99.00	118.23	136.80	216.20	▁▃▇▃▁
time_signature	0	1	3.93	0.33	1.00	4.00	4.00	4.00	5.00	▁▁▁▇▁

Data Tidying

Encode Missing Data

```
data[is.na(data)] = 0
```

Data Parsing

For character type of data, since some variables contain observations with multiple delimited values, I am going to separate the values and places each one in its own row. Moreover, special character and unnecessary space have to be removed.

```
data1 <- data

#character data - performer
##seperate a collapsed column into multiple rows, so that each row only indicates 1 performer
data1 <- data1 %>%
  separate_rows(performer, sep = ",")
##remove special character
data1$performer <- gsub("[^[:alnum:] ]","",data1$performer)

#character data - genre
##seperate a collapsed column into multiple rows, so that each row only indicates 1 genre
data1 <- data1 %>%
  separate_rows(genre, sep = ",")
##remove special character
data1$genre <- gsub("[^[:alnum:] ]","",data1$genre)
##remove both leading and trailing whitespace
data1$genre <- trimws(data1$genre)

head(data1)
```

```
## # A tibble: 6 × 19
##      id performer song  genre track…¹ track…² dance…³ energy   key loudn…⁴  mode
##   <int> <chr>     <chr> <chr>   <dbl> <lgl>     <dbl>  <dbl> <int>   <dbl> <int>
## 1 94500 Andy Wil… ....… adul…  166106 FALSE     0.154  0.185     5   -14.1     1
## 2 94500 Andy Wil… ....… bril…  166106 FALSE     0.154  0.185     5   -14.1     1
## 3 94500 Andy Wil… ....… easy…  166106 FALSE     0.154  0.185     5   -14.1     1
## 4 94500 Andy Wil… ....… mell…  166106 FALSE     0.154  0.185     5   -14.1     1
## 5 64901 Sandy Ne… ...A… rock…  172066 FALSE     0.588  0.672    11   -17.3     0
## 6 64901 Sandy Ne… ...A… spac…  172066 FALSE     0.588  0.672    11   -17.3     0
## # … with 8 more variables: speechiness <dbl>, acousticness <dbl>,
## #   instrumentalness <dbl>, liveness <dbl>, valence <dbl>, tempo <dbl>,
## #   time_signature <int>, rating <int>, and abbreviated variable names
## #   ¹track_duration, ²track_explicit, ³danceability, ⁴loudness
```

As a result, we can see the observations are separated to multiple rows, and each row only indicates 1 genre and 1 performer. For example, the first observation (id: 94500), has been separated to 4 rows, because it represents 4 genres and 1 performer.

Data Transformation

It's hard to directly input logical and character data into models, so I am going to change them to either numeric or factor data type. Furthermore, to prepare future data analysis (eg. Principal Components Analysis), I will change integer to numeric data.

```
#data type transformation: logical to numeric
data1$track_explicit <- ifelse(data1$track_explicit==TRUE, 1, 0)
#data type transformation: character to factor
data1$performer <- as.factor(data1$performer)
data1$genre <- as.factor(data1$genre)
#data type transformation: integer to numeric
data1$rating <- as.numeric(data1$rating)
data1$key <- as.numeric(data1$key)
data1$mode <- as.numeric(data1$mode)
data1$time_signature <- as.numeric(data1$time_signature)

skim(data1)
```

```
## Warning in sorted_count(x): Variable contains value(s) of "" that have been
## converted to "empty".
```

Data summary

Name	data1
Number of rows	96654
Number of columns	19
_____	
Column type frequency:	
character	1
factor	2
numeric	16
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
song	0	1	1	75	0	16542	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
performer	0	1	FALSE	6869	Dra: 468, Gle: 441, Dio: 370, Jac: 320
genre	0	1	FALSE	996	mel: 2985, sof: 2830, adu: 2613, roc: 2462

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
id	0	1	50075.79	29019.08	3.00	24669.00	50317.00	75491.00	99999.00	
track_duration	0	1	224863.68	67397.37	29688.00	178906.00	219054.50	257686.25	3079157.00	
track_explicit	0	1	0.13	0.33	0.00	0.00	0.00	0.00	1.00	
danceability	0	1	0.60	0.15	0.00	0.50	0.61	0.71	0.99	
energy	0	1	0.62	0.20	0.00	0.48	0.64	0.78	1.00	
key	0	1	5.22	3.55	0.00	2.00	5.00	8.00	11.00	
loudness	0	1	-8.72	3.60	-28.03	-11.07	-8.27	-5.96	2.29	
mode	0	1	0.73	0.45	0.00	0.00	1.00	1.00	1.00	
speechiness	0	1	0.07	0.08	0.00	0.03	0.04	0.07	0.92	
acousticness	0	1	0.28	0.27	0.00	0.05	0.19	0.48	0.99	
instrumentalness	0	1	0.03	0.12	0.00	0.00	0.00	0.00	0.98	
liveness	0	1	0.19	0.16	0.01	0.09	0.13	0.25	1.00	
valence	0	1	0.61	0.24	0.00	0.43	0.63	0.80	0.99	
tempo	0	1	120.46	27.81	0.00	99.48	119.04	136.51	241.01	
time_signature	0	1	3.94	0.31	0.00	4.00	4.00	4.00	5.00	
rating	0	1	37.67	15.93	0.00	26.00	38.00	50.00	91.00	

Even though character data has been transformed to factor, it cannot be integrated over, summed over, or marginalized for further data analysis. Thus, I am going to create dummy variables so that these variables can be used in many machine learning models.

```
#create dummy variable - genre
data2 <- data1 %>%
    mutate(n=1) %>%
    pivot_wider(names_from=genre, values_from=n, names_prefix="genre_", values_fill=list(n=0))
#create dummy variable - performer
data2 <- data2 %>%
    mutate(n=1) %>%
    pivot_wider(names_from=performer, values_from=n, names_prefix="performer_", values_fill=list(n=0))
```

To extract more specific data sets for further variable selection, I drop the variables (id & song) that has been excluded to data analysis.

```
#delete unused variables
data3 <- data2 %>% select(-id) %>% select(-song)

str(data3)
```

```
## tibble [19,485 × 7,880] (S3: tbl_df/tbl/data.frame)
##  $ track_duration                                                    : num [1:19485] 16
6106 172066 211066 208186 182080 ...
##  $ track_explicit                                                    : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ danceability                                                      : num [1:19485] 0.
154 0.588 0.759 0.613 0.45 0.57 0.612 0.253 0.575 0.615 ...
##  $ energy                                                            : num [1:19485] 0.
185 0.672 0.699 0.764 0.294 0.629 0.542 0.232 0.434 0.497 ...
##  $ key                                                               : num [1:19485] 5
11 0 2 7 9 5 0 5 7 ...
##  $ loudness                                                          : num [1:19485] -1
4.06 -17.28 -5.75 -6.51 -12.02 ...
##  $ mode                                                              : num [1:19485] 1
0 0 1 1 0 1 1 1 1 ...
##  $ speechiness                                                       : num [1:19485] 0.
0315 0.0361 0.0307 0.136 0.0318 0.0331 0.0264 0.0318 0.0312 0.439 ...
##  $ acousticness                                                      : num [1:19485] 0.
911 0.00256 0.202 0.0527 0.832 0.593 0.0781 0.805 0.735 0.016 ...
##  $ instrumentalness                                                  : num [1:19485] 2.
67e-04 7.45e-01 1.31e-04 0.00 3.53e-05 1.36e-04 0.00 1.80e-04 6.59e-05 0.00 ...
##  $ liveness                                                          : num [1:19485] 0.
112 0.145 0.443 0.197 0.108 0.77 0.0763 0.0939 0.105 0.312 ...
##  $ valence                                                           : num [1:19485] 0.
15 0.801 0.907 0.417 0.146 0.308 0.433 0.307 0.348 0.769 ...
##  $ tempo                                                             : num [1:19485] 84
122 93 160 141 ...
##  $ time_signature                                                    : num [1:19485] 4
4 4 4 4 4 4 3 4 3 ...
##  $ rating                                                            : num [1:19485] 36
16 70 64 19 34 44 34 47 26 ...
##  $ genre_adult standards                                             : num [1:19485] 1
0 0 0 0 0 0 0 0 0 ...
##  $ genre_brill building pop                                          : num [1:19485] 1
0 0 0 0 0 0 0 0 0 ...
##  $ genre_easy listening                                              : num [1:19485] 1
0 0 0 0 0 0 0 0 0 ...
##  $ genre_mellow gold                                                 : num [1:19485] 1
0 0 0 0 0 0 0 0 0 ...
##  $ genre_rockandroll                                                 : num [1:19485] 0
1 0 0 0 0 0 0 0 0 ...
##  $ genre_space age pop                                               : num [1:19485] 0
1 0 0 0 0 0 0 0 0 ...
##  $ genre_surf music                                                  : num [1:19485] 0
1 0 0 0 0 0 0 0 0 ...
##  $ genre_dance pop                                                   : num [1:19485] 0
0 1 0 0 0 0 0 0 0 ...
##  $ genre_pop                                                         : num [1:19485] 0
0 1 1 0 0 0 0 1 0 ...
##  $ genre_postteen pop                                                : num [1:19485] 0
0 1 1 0 0 0 0 1 0 ...
##  $ genre_country                                                     : num [1:19485] 0
0 0 0 1 1 0 0 0 0 ...
##  $ genre_country dawn                                                : num [1:19485] 0
0 0 0 1 0 0 0 0 0 ...
##  $ genre_nashville sound                                             : num [1:19485] 0
0 0 0 1 0 0 0 0 0 ...
##  $ genre_australian country                                          : num [1:19485] 0
0 0 0 0 1 0 0 0 0 ...
##  $ genre_contemporary country                                        : num [1:19485] 0
0 0 0 0 1 0 0 0 0 ...
##  $ genre_country road                                                : num [1:19485] 0
0 0 0 0 1 0 0 0 0 ...
##  $ genre_funk                                                        : num [1:19485] 0
```

```
0 0 0 0 0 1 0 0 0 ...
##  $ genre_neo soul                                                    : num [1:19485] 0
0 0 0 0 0 1 0 0 0 ...
##  $ genre_new jack swing                                              : num [1:19485] 0
0 0 0 0 0 1 0 0 0 ...
##  $ genre_quiet storm                                                 : num [1:19485] 0
0 0 0 0 0 1 0 0 0 ...
##  $ genre_rb                                                          : num [1:19485] 0
0 0 0 0 0 1 0 0 0 ...
##  $ genre_urban contemporary                                          : num [1:19485] 0
0 0 0 0 0 1 0 0 0 ...
##  $ genre_blues rock                                                  : num [1:19485] 0
0 0 0 0 0 0 1 0 0 ...
##  $ genre_garage rock                                                 : num [1:19485] 0
0 0 0 0 0 0 1 0 0 ...
##  $ genre_modern blues rock                                           : num [1:19485] 0
0 0 0 0 0 0 1 0 0 ...
##  $ genre_neopsychedelic                                              : num [1:19485] 0
0 0 0 0 0 0 1 0 0 ...
##  $ genre_nu gaze                                                     : num [1:19485] 0
0 0 0 0 0 0 1 0 0 ...
##  $ genre_punk blues                                                  : num [1:19485] 0
0 0 0 0 0 0 1 0 0 ...
##  $ genre_                                                            : num [1:19485] 0
0 0 0 0 0 0 0 0 1 ...
##  $ genre_folk                                                        : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_folk rock                                                   : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_singersongwriter                                            : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_soft rock                                                   : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_yacht rock                                                  : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_bubblegum pop                                               : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_lounge                                                      : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_rockabilly                                                  : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_sunshine pop                                                : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian pop                                                : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_boston rock                                                 : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_dance rock                                                  : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_new romantic                                                : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_new wave                                                    : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_new wave pop                                                : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_album rock                                                  : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_synthpop                                                    : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic soul                                                : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic country pop                                         : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_outlaw country                                              : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
```

```
##  $ genre_texas country                                                    : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_disco                                                            : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_postdisco                                                        : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_motown                                                           : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_soul                                                             : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_southern soul                                                    : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_doowop                                                           : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_rhythm and blues                                                 : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_art rock                                                         : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic rock                                                     : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_hard rock                                                        : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_metal                                                            : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_psychedelic rock                                                 : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_rock                                                             : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_chicago soul                                                     : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_blues                                                            : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_electric blues                                                   : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_jazz blues                                                       : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_louisiana blues                                                  : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_new orleans blues                                                : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_piano blues                                                      : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_roots rock                                                       : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian singersongwriter                                        : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic canadian rock                                            : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_heartland rock                                                   : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_permanent wave                                                   : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_country rock                                                     : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_southern rock                                                    : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_christmas instrumental                                           : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_philly soul                                                      : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_british invasion                                                 : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_protopunk                                                        : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_contemporary vocal jazz                                          : num [1:19485] 0
```

```
0 0 0 0 0 0 0 0 0 ...
##  $ genre_vocal jazz                                                                  : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##  $ genre_deep adult standards                                                        : num [1:19485] 0
0 0 0 0 0 0 0 0 0 ...
##   [list output truncated]
```

Now, the analysisData contains same rows of observation as original, meanwhile it contains the dummy variables derived by performer and genre. Each dummy variable represents a category of either performer or genre.

I am going to perform the same data tidying process for scoringData, before we check the inter-sets of variables/columns between two data sets.

```
#Encode Missing Data
sdata[is.na(sdata)] = 0

#Data Parsing
sdata1 <- sdata
#character data - performer
##seperate a collapsed column into multiple rows, so that each row only indicates 1 performer
sdata1 <- sdata1 %>%
  separate_rows(performer, sep = ",")
##remove special character
sdata1$performer <- gsub("[^[:alnum:] ]","",sdata1$performer)

#character data - genre
##seperate a collapsed column into multiple rows, so that each row only indicates 1 genre
sdata1 <- sdata1 %>%
  separate_rows(genre, sep = ",")
##remove special character
sdata1$genre <- gsub("[^[:alnum:] ]","",sdata1$genre)
##remove both leading and trailing whitespace
sdata1$genre <- trimws(sdata1$genre)

#Data Transformation
#data type transformation: logical to numeric
sdata1$track_explicit <- ifelse(sdata1$track_explicit==TRUE, 1, 0)
#data type transformation: character to factor
sdata1$performer <- as.factor(sdata1$performer)
sdata1$genre <- as.factor(sdata1$genre)
#data type transformation: integer to numeric
sdata1$key <- as.numeric(sdata1$key)
sdata1$mode <- as.numeric(sdata1$mode)
sdata1$time_signature <- as.numeric(sdata1$time_signature)

#create dummy variables - genre
sdata2 <- sdata1 %>%
    mutate(n=1) %>%
    pivot_wider(names_from=genre, values_from=n, names_prefix="genre_", values_fill=list(n=0))
#create dummy variables - performer
sdata2 <- sdata2 %>%
    mutate(n=1) %>%
    pivot_wider(names_from=performer, values_from=n, names_prefix="performer_", values_fill=list(n=0))

#delete unused variables
sdata3 <- sdata2 %>% select(-id) %>% select(-song)

str(sdata3)
```

```
## tibble [4,844 × 3,500] (S3: tbl_df/tbl/data.frame)
##  $ track_duration                                                    : num [1:4844] 219813 3
32226 127829 210973 180133 ...
##  $ track_explicit                                                    : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ danceability                                                      : num [1:4844] 0.647 0.
804 0.744 0.712 0.665 0.442 0.61 0.478 0.499 0.801 ...
##  $ energy                                                            : num [1:4844] 0.686 0.
714 0.47 0.753 0.552 0.717 0.377 0.298 0.249 0.875 ...
##  $ key                                                               : num [1:4844] 2 11 7 9
6 9 9 2 5 7 ...
##  $ loudness                                                          : num [1:4844] -4.25 -
6.71 -10.29 -7.67 -10.18 ...
##  $ mode                                                              : num [1:4844] 0 0 1 1
1 0 1 1 1 1 ...
##  $ speechiness                                                       : num [1:4844] 0.0274
0.183 0.035 0.0399 0.0307 0.0395 0.0935 0.0342 0.0422 0.0995 ...
##  $ acousticness                                                      : num [1:4844] 0.432 0.
0567 0.659 0.756 0.493 0.518 0.414 0.74 0.733 0.00366 ...
##  $ instrumentalness                                                  : num [1:4844] 6.19e-06
6.21e-06 0.00 1.58e-06 2.72e-02 2.62e-01 0.00 1.11e-05 0.00 4.25e-01 ...
##  $ liveness                                                          : num [1:4844] 0.133 0.
0253 0.256 0.169 0.421 0.184 0.0734 0.139 0.105 0.182 ...
##  $ valence                                                           : num [1:4844] 0.952 0.
802 0.908 0.953 0.899 0.376 0.519 0.376 0.539 0.637 ...
##  $ tempo                                                             : num [1:4844] 155.7 13
9.7 116.2 116.2 96.9 ...
##  $ time_signature                                                    : num [1:4844] 4 4 4 4
4 4 4 4 4 4 ...
##  $ genre_album rock                                                  : num [1:4844] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_bubblegum pop                                               : num [1:4844] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_country rock                                                : num [1:4844] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_folk rock                                                   : num [1:4844] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_mellow gold                                                 : num [1:4844] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_new wave pop                                                : num [1:4844] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_soft rock                                                   : num [1:4844] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_yacht rock                                                  : num [1:4844] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_funk                                                        : num [1:4844] 0 1 0 0
1 0 0 0 0 0 ...
##  $ genre_motown                                                      : num [1:4844] 0 1 0 0
1 0 0 0 0 0 ...
##  $ genre_neo soul                                                    : num [1:4844] 0 1 0 0
0 0 0 0 0 0 ...
##  $ genre_new jack swing                                              : num [1:4844] 0 1 0 0
0 0 0 0 0 0 ...
##  $ genre_quiet storm                                                 : num [1:4844] 0 1 0 0
1 0 0 0 0 0 ...
##  $ genre_rb                                                          : num [1:4844] 0 1 0 0
0 0 0 0 0 0 ...
##  $ genre_soul                                                        : num [1:4844] 0 1 0 0
1 0 0 0 0 0 ...
##  $ genre_urban contemporary                                          : num [1:4844] 0 1 0 0
0 0 0 0 0 0 ...
##  $ genre_doowop                                                      : num [1:4844] 0 0 1 0
0 0 0 0 1 0 ...
##  $ genre_rhythm and blues                                            : num [1:4844] 0 0 1 0
```

```
0 0 0 0 1 0 ...
##  $ genre_classic uk pop                                                    : num [1:4844] 0 0 0 1
0 0 0 0 0 0 ...
##  $ genre_chicago soul                                                      : num [1:4844] 0 0 0 0
1 0 0 0 0 0 ...
##  $ genre_classic soul                                                      : num [1:4844] 0 0 0 0
1 0 0 0 0 0 ...
##  $ genre_disco                                                             : num [1:4844] 0 0 0 0
1 0 0 0 0 0 ...
##  $ genre_southern soul                                                     : num [1:4844] 0 0 0 0
1 0 0 0 0 0 ...
##  $ genre_classic garage rock                                               : num [1:4844] 0 0 0 0
0 1 0 0 0 0 ...
##  $ genre_                                                                  : num [1:4844] 0 0 0 0
0 0 1 0 0 0 ...
##  $ genre_east coast hip hop                                                : num [1:4844] 0 0 0 0
0 0 0 1 0 0 ...
##  $ genre_hip hop                                                           : num [1:4844] 0 0 0 0
0 0 0 1 0 0 ...
##  $ genre_pop rap                                                           : num [1:4844] 0 0 0 0
0 0 0 1 0 0 ...
##  $ genre_rap                                                               : num [1:4844] 0 0 0 0
0 0 0 1 0 0 ...
##  $ genre_southern hip hop                                                  : num [1:4844] 0 0 0 0
0 0 0 1 0 0 ...
##  $ genre_adult standards                                                   : num [1:4844] 0 0 0 0
0 0 0 0 1 0 ...
##  $ genre_brill building pop                                                : num [1:4844] 0 0 0 0
0 0 0 0 1 0 ...
##  $ genre_rockandroll                                                       : num [1:4844] 0 0 0 0
0 0 0 0 1 0 ...
##  $ genre_vocal harmony group                                               : num [1:4844] 0 0 0 0
0 0 0 0 1 0 ...
##  $ genre_new jersey rap                                                    : num [1:4844] 0 0 0 0
0 0 0 0 0 1 ...
##  $ genre_deep adult standards                                              : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_outlaw country                                                    : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_redneck                                                           : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_northern soul                                                     : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_glee club                                                         : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_hollywood                                                         : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_postteen pop                                                      : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_classic country pop                                               : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_country                                                           : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_country gospel                                                    : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_lounge                                                            : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_rockabilly                                                        : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_modern country rock                                               : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_classic rock                                                      : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_hard rock                                                         : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
```

```
##  $ genre_metal                                                    : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_rock                                                     : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_easy listening                                           : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_british folk                                             : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_folk                                                     : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_roots rock                                               : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_singersongwriter                                         : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_jazz blues                                               : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_jazz saxophone                                           : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_soul jazz                                                : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_classic girl group                                       : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_vocal jazz                                               : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_beach music                                              : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_philly soul                                              : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_freakbeat                                                : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_protopunk                                                : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_psychedelic rock                                         : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_british invasion                                         : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_merseybeat                                               : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_dance pop                                                : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_pop                                                      : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_postdisco                                                : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_g funk                                                   : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_gangster rap                                             : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_west coast rap                                           : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_trap                                                     : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_pittsburgh rap                                           : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_melodic rap                                              : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_glam metal                                               : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_indie pop rap                                            : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_conscious hip hop                                        : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_nc hip hop                                               : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_baton rouge rap                                          : num [1:4844] 0 0 0 0
```

```
0 0 0 0 0 0 ...
##  $ genre_atl hip hop                                                                 : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_dirty south rap                                                             : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##   [list output truncated]
```

Feature Selection 1

Variable Inter-set

From the summary of data3 and sdata3, we can see the numbers of variable are different (7,880 vs. 3,500). The reason is that dummy variables derived from the values of performer and genre can be different between two data sets. To make sure the variables to be analysed are aligned in both data sets, I am going to take their inter-sets of the column names to be the predictors.

```
variable1 <- intersect(names(data3), names(sdata3))
data4 <- data3[c(variable1,"rating")] %>% relocate(rating, .before=track_duration)
sdata4 <- sdata3[variable1]

#tidy up the column names, especially the dummy variables derived from the value of performer and genre
data4 <- clean_names(data4)
sdata4 <- clean_names(sdata4)

str(data4)
```

```
## tibble [19,485 × 2,387] (S3: tbl_df/tbl/data.frame)
##  $ rating                                                                : num [1:19485] 36 16 7
0 64 19 34 44 34 47 26 ...
##  $ track_duration                                                        : num [1:19485] 166106
172066 211066 208186 182080 ...
##  $ track_explicit                                                        : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ danceability                                                          : num [1:19485] 0.154
0.588 0.759 0.613 0.45 0.57 0.612 0.253 0.575 0.615 ...
##  $ energy                                                                : num [1:19485] 0.185
0.672 0.699 0.764 0.294 0.629 0.542 0.232 0.434 0.497 ...
##  $ key                                                                   : num [1:19485] 5 11 0
2 7 9 5 0 5 7 ...
##  $ loudness                                                              : num [1:19485] -14.06
-17.28 -5.75 -6.51 -12.02 ...
##  $ mode                                                                  : num [1:19485] 1 0 0 1
1 0 1 1 1 1 ...
##  $ speechiness                                                           : num [1:19485] 0.0315
0.0361 0.0307 0.136 0.0318 0.0331 0.0264 0.0318 0.0312 0.439 ...
##  $ acousticness                                                          : num [1:19485] 0.911
0.00256 0.202 0.0527 0.832 0.593 0.0781 0.805 0.735 0.016 ...
##  $ instrumentalness                                                      : num [1:19485] 2.67e-0
4 7.45e-01 1.31e-04 0.00 3.53e-05 1.36e-04 0.00 1.80e-04 6.59e-05 0.00 ...
##  $ liveness                                                              : num [1:19485] 0.112
0.145 0.443 0.197 0.108 0.77 0.0763 0.0939 0.105 0.312 ...
##  $ valence                                                               : num [1:19485] 0.15 0.
801 0.907 0.417 0.146 0.308 0.433 0.307 0.348 0.769 ...
##  $ tempo                                                                 : num [1:19485] 84 122
93 160 141 ...
##  $ time_signature                                                        : num [1:19485] 4 4 4 4
4 4 4 3 4 3 ...
##  $ genre_adult_standards                                                 : num [1:19485] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_brill_building_pop                                              : num [1:19485] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_easy_listening                                                  : num [1:19485] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_mellow_gold                                                     : num [1:19485] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_rockandroll                                                     : num [1:19485] 0 1 0 0
0 0 0 0 0 0 ...
##  $ genre_space_age_pop                                                   : num [1:19485] 0 1 0 0
0 0 0 0 0 0 ...
##  $ genre_surf_music                                                      : num [1:19485] 0 1 0 0
0 0 0 0 0 0 ...
##  $ genre_dance_pop                                                       : num [1:19485] 0 0 1 0
0 0 0 0 0 0 ...
##  $ genre_pop                                                             : num [1:19485] 0 0 1 1
0 0 0 0 1 0 ...
##  $ genre_postteen_pop                                                    : num [1:19485] 0 0 1 1
0 0 0 0 1 0 ...
##  $ genre_country                                                         : num [1:19485] 0 0 0 0
1 1 0 0 0 0 ...
##  $ genre_country_dawn                                                    : num [1:19485] 0 0 0 0
1 0 0 0 0 0 ...
##  $ genre_nashville_sound                                                 : num [1:19485] 0 0 0 0
1 0 0 0 0 0 ...
##  $ genre_australian_country                                              : num [1:19485] 0 0 0 0
0 1 0 0 0 0 ...
##  $ genre_contemporary_country                                            : num [1:19485] 0 0 0 0
0 1 0 0 0 0 ...
##  $ genre_country_road                                                    : num [1:19485] 0 0 0 0
0 1 0 0 0 0 ...
##  $ genre_funk                                                            : num [1:19485] 0 0 0 0
```

```
0 0 1 0 0 0 ...
##  $ genre_neo_soul                                                    : num [1:19485] 0 0 0 0
0 0 1 0 0 0 ...
##  $ genre_new_jack_swing                                              : num [1:19485] 0 0 0 0
0 0 1 0 0 0 ...
##  $ genre_quiet_storm                                                 : num [1:19485] 0 0 0 0
0 0 1 0 0 0 ...
##  $ genre_rb                                                          : num [1:19485] 0 0 0 0
0 0 1 0 0 0 ...
##  $ genre_urban_contemporary                                          : num [1:19485] 0 0 0 0
0 0 1 0 0 0 ...
##  $ genre_blues_rock                                                  : num [1:19485] 0 0 0 0
0 0 0 1 0 0 ...
##  $ genre_garage_rock                                                 : num [1:19485] 0 0 0 0
0 0 0 1 0 0 ...
##  $ genre_modern_blues_rock                                           : num [1:19485] 0 0 0 0
0 0 0 1 0 0 ...
##  $ genre_neopsychedelic                                              : num [1:19485] 0 0 0 0
0 0 0 1 0 0 ...
##  $ genre_nu_gaze                                                     : num [1:19485] 0 0 0 0
0 0 0 1 0 0 ...
##  $ genre_punk_blues                                                  : num [1:19485] 0 0 0 0
0 0 0 1 0 0 ...
##  $ genre                                                             : num [1:19485] 0 0 0 0
0 0 0 0 0 1 ...
##  $ genre_folk                                                        : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_folk_rock                                                   : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_singersongwriter                                            : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_soft_rock                                                   : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_yacht_rock                                                  : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_bubblegum_pop                                               : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_lounge                                                      : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_rockabilly                                                  : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_sunshine_pop                                                : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_canadian_pop                                                : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_boston_rock                                                 : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_dance_rock                                                  : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_new_romantic                                                : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_new_wave                                                    : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_new_wave_pop                                                : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_album_rock                                                  : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_synthpop                                                    : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_classic_soul                                                : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_classic_country_pop                                         : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_outlaw_country                                              : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
```

```
##  $ genre_texas_country                                : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_disco                                        : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_postdisco                                    : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_motown                                       : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_soul                                         : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_southern_soul                                : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_doowop                                       : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_rhythm_and_blues                             : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_art_rock                                     : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_classic_rock                                 : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_hard_rock                                    : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_metal                                        : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_psychedelic_rock                             : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_rock                                         : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_chicago_soul                                 : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_blues                                        : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_electric_blues                               : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_jazz_blues                                   : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_louisiana_blues                              : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_new_orleans_blues                            : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_piano_blues                                  : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_roots_rock                                   : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_canadian_singersongwriter                    : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_classic_canadian_rock                        : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_heartland_rock                               : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_permanent_wave                               : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_country_rock                                 : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_southern_rock                                : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_christmas_instrumental                       : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_philly_soul                                  : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_british_invasion                             : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_protopunk                                    : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_vocal_jazz                                   : num [1:19485] 0 0 0 0
```

```
0 0 0 0 0 0 ...
##  $ genre_deep_adult_standards                                                    : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_alternative_metal                                                       : num [1:19485] 0 0 0 0
0 0 0 0 0 0 ...
##   [list output truncated]
```

```
str(sdata4)
```

```
## tibble [4,844 × 2,386] (S3: tbl_df/tbl/data.frame)
##  $ track_duration                                                    : num [1:4844] 219813 3
32226 127829 210973 180133 ...
##  $ track_explicit                                                    : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ danceability                                                      : num [1:4844] 0.647 0.
804 0.744 0.712 0.665 0.442 0.61 0.478 0.499 0.801 ...
##  $ energy                                                            : num [1:4844] 0.686 0.
714 0.47 0.753 0.552 0.717 0.377 0.298 0.249 0.875 ...
##  $ key                                                               : num [1:4844] 2 11 7 9
6 9 9 2 5 7 ...
##  $ loudness                                                          : num [1:4844] -4.25 -
6.71 -10.29 -7.67 -10.18 ...
##  $ mode                                                              : num [1:4844] 0 0 1 1
1 0 1 1 1 1 ...
##  $ speechiness                                                       : num [1:4844] 0.0274
0.183 0.035 0.0399 0.0307 0.0395 0.0935 0.0342 0.0422 0.0995 ...
##  $ acousticness                                                      : num [1:4844] 0.432 0.
0567 0.659 0.756 0.493 0.518 0.414 0.74 0.733 0.00366 ...
##  $ instrumentalness                                                  : num [1:4844] 6.19e-06
6.21e-06 0.00 1.58e-06 2.72e-02 2.62e-01 0.00 1.11e-05 0.00 4.25e-01 ...
##  $ liveness                                                          : num [1:4844] 0.133 0.
0253 0.256 0.169 0.421 0.184 0.0734 0.139 0.105 0.182 ...
##  $ valence                                                           : num [1:4844] 0.952 0.
802 0.908 0.953 0.899 0.376 0.519 0.376 0.539 0.637 ...
##  $ tempo                                                             : num [1:4844] 155.7 13
9.7 116.2 116.2 96.9 ...
##  $ time_signature                                                    : num [1:4844] 4 4 4 4
4 4 4 4 4 4 ...
##  $ genre_adult_standards                                             : num [1:4844] 0 0 0 0
0 0 0 0 1 0 ...
##  $ genre_brill_building_pop                                          : num [1:4844] 0 0 0 0
0 0 0 0 1 0 ...
##  $ genre_easy_listening                                              : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_mellow_gold                                                 : num [1:4844] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_rockandroll                                                 : num [1:4844] 0 0 0 0
0 0 0 0 1 0 ...
##  $ genre_space_age_pop                                               : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_surf_music                                                  : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_dance_pop                                                   : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_pop                                                         : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_postteen_pop                                                : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_country                                                     : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_country_dawn                                                : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_nashville_sound                                             : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_australian_country                                          : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_contemporary_country                                        : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_country_road                                                : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_funk                                                        : num [1:4844] 0 1 0 0
1 0 0 0 0 0 ...
##  $ genre_neo_soul                                                    : num [1:4844] 0 1 0 0
```

```
0 0 0 0 0 0 ...
##  $ genre_new_jack_swing                                                : num [1:4844] 0 1 0 0
0 0 0 0 0 0 ...
##  $ genre_quiet_storm                                                   : num [1:4844] 0 1 0 0
1 0 0 0 0 0 ...
##  $ genre_rb                                                            : num [1:4844] 0 1 0 0
0 0 0 0 0 0 ...
##  $ genre_urban_contemporary                                            : num [1:4844] 0 1 0 0
0 0 0 0 0 0 ...
##  $ genre_blues_rock                                                    : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_garage_rock                                                   : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_modern_blues_rock                                             : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_neopsychedelic                                                : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_nu_gaze                                                       : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_punk_blues                                                    : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre                                                               : num [1:4844] 0 0 0 0
0 0 1 0 0 0 ...
##  $ genre_folk                                                          : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_folk_rock                                                     : num [1:4844] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_singersongwriter                                              : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_soft_rock                                                     : num [1:4844] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_yacht_rock                                                    : num [1:4844] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_bubblegum_pop                                                 : num [1:4844] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_lounge                                                        : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_rockabilly                                                    : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_sunshine_pop                                                  : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_canadian_pop                                                  : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_boston_rock                                                   : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_dance_rock                                                    : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_new_romantic                                                  : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_new_wave                                                      : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_new_wave_pop                                                  : num [1:4844] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_album_rock                                                    : num [1:4844] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_synthpop                                                      : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_classic_soul                                                  : num [1:4844] 0 0 0 0
1 0 0 0 0 0 ...
##  $ genre_classic_country_pop                                           : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_outlaw_country                                                : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_texas_country                                                 : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
```

```
##  $ genre_disco                                                    : num [1:4844] 0 0 0 0
1 0 0 0 0 0 ...
##  $ genre_postdisco                                                : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_motown                                                   : num [1:4844] 0 1 0 0
1 0 0 0 0 0 ...
##  $ genre_soul                                                     : num [1:4844] 0 1 0 0
1 0 0 0 0 0 ...
##  $ genre_southern_soul                                            : num [1:4844] 0 0 0 0
1 0 0 0 0 0 ...
##  $ genre_doowop                                                   : num [1:4844] 0 0 1 0
0 0 0 0 1 0 ...
##  $ genre_rhythm_and_blues                                         : num [1:4844] 0 0 1 0
0 0 0 0 1 0 ...
##  $ genre_art_rock                                                 : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_classic_rock                                             : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_hard_rock                                                : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_metal                                                    : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_psychedelic_rock                                         : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_rock                                                     : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_chicago_soul                                             : num [1:4844] 0 0 0 0
1 0 0 0 0 0 ...
##  $ genre_blues                                                    : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_electric_blues                                           : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_jazz_blues                                               : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_louisiana_blues                                          : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_new_orleans_blues                                        : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_piano_blues                                              : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_roots_rock                                               : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_canadian_singersongwriter                                : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_classic_canadian_rock                                    : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_heartland_rock                                           : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_permanent_wave                                           : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_country_rock                                             : num [1:4844] 1 0 0 0
0 0 0 0 0 0 ...
##  $ genre_southern_rock                                            : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_christmas_instrumental                                   : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_philly_soul                                              : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_british_invasion                                         : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_protopunk                                                : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_vocal_jazz                                               : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_deep_adult_standards                                     : num [1:4844] 0 0 0 0
```

```
0 0 0 0 0 0 ...
##  $ genre_alternative_metal                                                    : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##  $ genre_canadian_metal                                                       : num [1:4844] 0 0 0 0
0 0 0 0 0 0 ...
##   [list output truncated]
```

Remove Near Zero Variance

The variables with zero variance or near zero variance have little predictable value, which should be removed from predictors.

```
removeZeroVar <- data4[, sapply(data4, var) != 0]
removenearZeroVar <- nearZeroVar(removeZeroVar, names=TRUE, freqCut = 9999/1, uniqueCut = 1)
data5 <- data4[, setdiff(names(data4), removenearZeroVar)]

s_removeZeroVar <- sdata4[, sapply(sdata4, var) != 0]
s_removenearZeroVar <- nearZeroVar(s_removeZeroVar, names=TRUE, freqCut = 9999/1, uniqueCut = 1)
sdata5 <- sdata4[, setdiff(names(sdata4), s_removenearZeroVar)]

#repeat variable inter-set after removing near zero variance
variable2 <- intersect(names(data5), names(sdata5))
data6 <- data5[c(variable2,"rating")] %>% relocate(rating, .before=track_duration)
sdata6 <- sdata5[variable2]

str(data6)
```

```
## tibble [19,485 × 1,961] (S3: tbl_df/tbl/data.frame)
##  $ rating                                                : num [1:19485] 36 16 70 64 19 34 44 34 47 26
...
##  $ track_duration                                        : num [1:19485] 166106 172066 211066 208186 1
82080 ...
##  $ track_explicit                                        : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ danceability                                          : num [1:19485] 0.154 0.588 0.759 0.613 0.45
0.57 0.612 0.253 0.575 0.615 ...
##  $ energy                                                : num [1:19485] 0.185 0.672 0.699 0.764 0.294
0.629 0.542 0.232 0.434 0.497 ...
##  $ key                                                   : num [1:19485] 5 11 0 2 7 9 5 0 5 7 ...
##  $ loudness                                              : num [1:19485] -14.06 -17.28 -5.75 -6.51 -1
2.02 ...
##  $ mode                                                  : num [1:19485] 1 0 0 1 1 0 1 1 1 1 ...
##  $ speechiness                                           : num [1:19485] 0.0315 0.0361 0.0307 0.136 0.
0318 0.0331 0.0264 0.0318 0.0312 0.439 ...
##  $ acousticness                                          : num [1:19485] 0.911 0.00256 0.202 0.0527 0.
832 0.593 0.0781 0.805 0.735 0.016 ...
##  $ instrumentalness                                      : num [1:19485] 2.67e-04 7.45e-01 1.31e-04 0.
00 3.53e-05 1.36e-04 0.00 1.80e-04 6.59e-05 0.00 ...
##  $ liveness                                              : num [1:19485] 0.112 0.145 0.443 0.197 0.108
0.77 0.0763 0.0939 0.105 0.312 ...
##  $ valence                                               : num [1:19485] 0.15 0.801 0.907 0.417 0.146
0.308 0.433 0.307 0.348 0.769 ...
##  $ tempo                                                 : num [1:19485] 84 122 93 160 141 ...
##  $ time_signature                                        : num [1:19485] 4 4 4 4 4 4 4 3 4 3 ...
##  $ genre_adult_standards                                 : num [1:19485] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_brill_building_pop                              : num [1:19485] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_easy_listening                                  : num [1:19485] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_mellow_gold                                     : num [1:19485] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rockandroll                                     : num [1:19485] 0 1 0 0 0 0 0 0 0 0 ...
##  $ genre_space_age_pop                                   : num [1:19485] 0 1 0 0 0 0 0 0 0 0 ...
##  $ genre_surf_music                                      : num [1:19485] 0 1 0 0 0 0 0 0 0 0 ...
##  $ genre_dance_pop                                       : num [1:19485] 0 0 1 0 0 0 0 0 0 0 ...
##  $ genre_pop                                             : num [1:19485] 0 0 1 1 0 0 0 0 1 0 ...
##  $ genre_postteen_pop                                    : num [1:19485] 0 0 1 1 0 0 0 0 1 0 ...
##  $ genre_country                                         : num [1:19485] 0 0 0 0 1 1 0 0 0 0 ...
##  $ genre_country_dawn                                    : num [1:19485] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_nashville_sound                                 : num [1:19485] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_australian_country                              : num [1:19485] 0 0 0 0 0 1 0 0 0 0 ...
##  $ genre_contemporary_country                            : num [1:19485] 0 0 0 0 0 1 0 0 0 0 ...
##  $ genre_country_road                                    : num [1:19485] 0 0 0 0 0 1 0 0 0 0 ...
##  $ genre_funk                                            : num [1:19485] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_neo_soul                                        : num [1:19485] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_new_jack_swing                                  : num [1:19485] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_quiet_storm                                     : num [1:19485] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_rb                                              : num [1:19485] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_urban_contemporary                              : num [1:19485] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_blues_rock                                      : num [1:19485] 0 0 0 0 0 0 0 1 0 0 ...
##  $ genre_garage_rock                                     : num [1:19485] 0 0 0 0 0 0 0 1 0 0 ...
##  $ genre_modern_blues_rock                               : num [1:19485] 0 0 0 0 0 0 0 1 0 0 ...
##  $ genre_neopsychedelic                                  : num [1:19485] 0 0 0 0 0 0 0 1 0 0 ...
##  $ genre_nu_gaze                                         : num [1:19485] 0 0 0 0 0 0 0 1 0 0 ...
##  $ genre_punk_blues                                      : num [1:19485] 0 0 0 0 0 0 0 1 0 0 ...
##  $ genre                                                 : num [1:19485] 0 0 0 0 0 0 0 0 0 1 ...
##  $ genre_folk                                            : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_folk_rock                                       : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_singersongwriter                                : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_soft_rock                                       : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_yacht_rock                                      : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_bubblegum_pop                                   : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_lounge                                          : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rockabilly                                      : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_sunshine_pop                                    : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ genre_canadian_pop                                  : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_boston_rock                                   : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_dance_rock                                    : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_romantic                                  : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_wave                                      : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_wave_pop                                  : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_album_rock                                    : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_synthpop                                      : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_soul                                  : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_country_pop                           : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_outlaw_country                                : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_texas_country                                 : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_disco                                         : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_postdisco                                     : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_motown                                        : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_soul                                          : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_southern_soul                                 : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_doowop                                        : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rhythm_and_blues                              : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_art_rock                                      : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_rock                                  : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_hard_rock                                     : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_metal                                         : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_psychedelic_rock                              : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rock                                          : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_chicago_soul                                  : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_blues                                         : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_electric_blues                                : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_jazz_blues                                    : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_louisiana_blues                               : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_orleans_blues                             : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_piano_blues                                   : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_roots_rock                                    : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_singersongwriter                     : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_canadian_rock                         : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_heartland_rock                                : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_permanent_wave                                : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country_rock                                  : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_southern_rock                                 : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_christmas_instrumental                        : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_philly_soul                                   : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_british_invasion                              : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_protopunk                                     : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_vocal_jazz                                    : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_deep_adult_standards                          : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_alternative_metal                             : num [1:19485] 0 0 0 0 0 0 0 0 0 0 ...
##   [list output truncated]
```

```
str(sdata6)
```

```
## tibble [4,844 × 1,960] (S3: tbl_df/tbl/data.frame)
##  $ track_duration                                                : num [1:4844] 219813 332226 127829 210973 18
0133 ...
##  $ track_explicit                                                : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ danceability                                                  : num [1:4844] 0.647 0.804 0.744 0.712 0.665
0.442 0.61 0.478 0.499 0.801 ...
##  $ energy                                                        : num [1:4844] 0.686 0.714 0.47 0.753 0.552
0.717 0.377 0.298 0.249 0.875 ...
##  $ key                                                           : num [1:4844] 2 11 7 9 6 9 9 2 5 7 ...
##  $ loudness                                                      : num [1:4844] -4.25 -6.71 -10.29 -7.67 -10.1
8 ...
##  $ mode                                                          : num [1:4844] 0 0 1 1 1 0 1 1 1 1 ...
##  $ speechiness                                                   : num [1:4844] 0.0274 0.183 0.035 0.0399 0.03
07 0.0395 0.0935 0.0342 0.0422 0.0995 ...
##  $ acousticness                                                  : num [1:4844] 0.432 0.0567 0.659 0.756 0.493
0.518 0.414 0.74 0.733 0.00366 ...
##  $ instrumentalness                                              : num [1:4844] 6.19e-06 6.21e-06 0.00 1.58e-0
6 2.72e-02 2.62e-01 0.00 1.11e-05 0.00 4.25e-01 ...
##  $ liveness                                                      : num [1:4844] 0.133 0.0253 0.256 0.169 0.421
0.184 0.0734 0.139 0.105 0.182 ...
##  $ valence                                                       : num [1:4844] 0.952 0.802 0.908 0.953 0.899
0.376 0.519 0.376 0.539 0.637 ...
##  $ tempo                                                         : num [1:4844] 155.7 139.7 116.2 116.2 96.9
...
##  $ time_signature                                                : num [1:4844] 4 4 4 4 4 4 4 4 4 4 ...
##  $ genre_adult_standards                                         : num [1:4844] 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_brill_building_pop                                      : num [1:4844] 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_easy_listening                                          : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_mellow_gold                                             : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rockandroll                                             : num [1:4844] 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_space_age_pop                                           : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_surf_music                                              : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_dance_pop                                               : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_pop                                                     : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_postteen_pop                                            : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country                                                 : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country_dawn                                            : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_nashville_sound                                         : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_australian_country                                      : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_contemporary_country                                    : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country_road                                            : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_funk                                                    : num [1:4844] 0 1 0 0 1 0 0 0 0 0 ...
##  $ genre_neo_soul                                                : num [1:4844] 0 1 0 0 0 0 0 0 0 0 ...
##  $ genre_new_jack_swing                                          : num [1:4844] 0 1 0 0 0 0 0 0 0 0 ...
##  $ genre_quiet_storm                                             : num [1:4844] 0 1 0 0 1 0 0 0 0 0 ...
##  $ genre_rb                                                      : num [1:4844] 0 1 0 0 0 0 0 0 0 0 ...
##  $ genre_urban_contemporary                                      : num [1:4844] 0 1 0 0 0 0 0 0 0 0 ...
##  $ genre_blues_rock                                              : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_garage_rock                                             : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_modern_blues_rock                                       : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_neopsychedelic                                          : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_nu_gaze                                                 : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_punk_blues                                              : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre                                                         : num [1:4844] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_folk                                                    : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_folk_rock                                               : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_singersongwriter                                        : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_soft_rock                                               : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_yacht_rock                                              : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_bubblegum_pop                                           : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_lounge                                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rockabilly                                              : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_sunshine_pop                                            : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_pop                                            : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ genre_boston_rock                                : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_dance_rock                                 : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_romantic                               : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_wave                                   : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_wave_pop                               : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_album_rock                                 : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_synthpop                                   : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_soul                               : num [1:4844] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_classic_country_pop                        : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_outlaw_country                             : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_texas_country                              : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_disco                                      : num [1:4844] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_postdisco                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_motown                                     : num [1:4844] 0 1 0 0 1 0 0 0 0 0 ...
##  $ genre_soul                                       : num [1:4844] 0 1 0 0 1 0 0 0 0 0 ...
##  $ genre_southern_soul                              : num [1:4844] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_doowop                                     : num [1:4844] 0 0 1 0 0 0 0 0 1 0 ...
##  $ genre_rhythm_and_blues                           : num [1:4844] 0 0 1 0 0 0 0 0 1 0 ...
##  $ genre_art_rock                                   : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_rock                               : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_hard_rock                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_metal                                      : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_psychedelic_rock                           : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rock                                       : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_chicago_soul                               : num [1:4844] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_blues                                      : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_electric_blues                             : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_jazz_blues                                 : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_louisiana_blues                            : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_orleans_blues                          : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_piano_blues                                : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_roots_rock                                 : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_singersongwriter                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_canadian_rock                      : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_heartland_rock                             : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_permanent_wave                             : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country_rock                               : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_southern_rock                              : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_christmas_instrumental                     : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_philly_soul                                : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_british_invasion                           : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_protopunk                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_vocal_jazz                                 : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_deep_adult_standards                       : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_alternative_metal                          : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_metal                             : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##   [list output truncated]
```

Split Data

I split data5 into train and test sample, so that I can build the model with train sample and check the model performance with test sample later.

```
set.seed(617)
split=sample.split(data5$rating, SplitRatio = 0.7)
train=data6[split,]
test=data6[!split,]
```

To be safe, I remove zero variance and tidy up the column names again after splitting the data.

```
train <- train[ , which(apply(train, 2, var) != 0)]
test <- test[ , which(apply(test, 2, var) != 0)]
train <- clean_names(train)
test <- clean_names(test)
str(train)
```

```
## tibble [13,638 × 1,930] (S3: tbl_df/tbl/data.frame)
##  $ rating                                                : num [1:13638] 36 70 64 19 34 44 34 47 30 55
...
##  $ track_duration                                        : num [1:13638] 166106 211066 208186 182080 3
31466 ...
##  $ track_explicit                                        : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ danceability                                          : num [1:13638] 0.154 0.759 0.613 0.45 0.57
0.612 0.253 0.575 0.34 0.518 ...
##  $ energy                                                : num [1:13638] 0.185 0.699 0.764 0.294 0.629
0.542 0.232 0.434 0.948 0.432 ...
##  $ key                                                   : num [1:13638] 5 0 2 7 9 5 0 5 9 10 ...
##  $ loudness                                              : num [1:13638] -14.06 -5.75 -6.51 -12.02 -7.
61 ...
##  $ mode                                                  : num [1:13638] 1 0 1 1 0 1 1 1 1 0 ...
##  $ speechiness                                           : num [1:13638] 0.0315 0.0307 0.136 0.0318 0.
0331 0.0264 0.0318 0.0312 0.137 0.0459 ...
##  $ acousticness                                          : num [1:13638] 0.911 0.202 0.0527 0.832 0.59
3 0.0781 0.805 0.735 0.0941 0.401 ...
##  $ instrumentalness                                      : num [1:13638] 2.67e-04 1.31e-04 0.00 3.53e-
05 1.36e-04 0.00 1.80e-04 6.59e-05 9.07e-04 0.00 ...
##  $ liveness                                              : num [1:13638] 0.112 0.443 0.197 0.108 0.77
0.0763 0.0939 0.105 0.867 0.299 ...
##  $ valence                                               : num [1:13638] 0.15 0.907 0.417 0.146 0.308
0.433 0.307 0.348 0.604 0.701 ...
##  $ tempo                                                 : num [1:13638] 84 93 160 141 128 ...
##  $ time_signature                                        : num [1:13638] 4 4 4 4 4 4 3 4 4 4 ...
##  $ genre_adult_standards                                 : num [1:13638] 1 0 0 0 0 0 0 0 1 1 ...
##  $ genre_brill_building_pop                              : num [1:13638] 1 0 0 0 0 0 0 0 1 1 ...
##  $ genre_easy_listening                                  : num [1:13638] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_mellow_gold                                     : num [1:13638] 1 0 0 0 0 0 0 0 1 1 ...
##  $ genre_rockandroll                                     : num [1:13638] 0 0 0 0 0 0 0 0 0 1 ...
##  $ genre_space_age_pop                                   : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_surf_music                                      : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_dance_pop                                       : num [1:13638] 0 1 0 0 0 0 0 0 0 0 ...
##  $ genre_pop                                             : num [1:13638] 0 1 1 0 0 0 0 1 0 0 ...
##  $ genre_postteen_pop                                    : num [1:13638] 0 1 1 0 0 0 0 1 0 0 ...
##  $ genre_country                                         : num [1:13638] 0 0 0 1 1 0 0 0 0 0 ...
##  $ genre_country_dawn                                    : num [1:13638] 0 0 0 1 0 0 0 0 0 0 ...
##  $ genre_nashville_sound                                 : num [1:13638] 0 0 0 1 0 0 0 0 0 0 ...
##  $ genre_australian_country                              : num [1:13638] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_contemporary_country                            : num [1:13638] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_country_road                                    : num [1:13638] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_funk                                            : num [1:13638] 0 0 0 0 0 1 0 0 0 0 ...
##  $ genre_neo_soul                                        : num [1:13638] 0 0 0 0 0 1 0 0 0 0 ...
##  $ genre_new_jack_swing                                  : num [1:13638] 0 0 0 0 0 1 0 0 0 0 ...
##  $ genre_quiet_storm                                     : num [1:13638] 0 0 0 0 0 1 0 0 0 0 ...
##  $ genre_rb                                              : num [1:13638] 0 0 0 0 0 1 0 0 0 0 ...
##  $ genre_urban_contemporary                              : num [1:13638] 0 0 0 0 0 1 0 0 0 0 ...
##  $ genre_blues_rock                                      : num [1:13638] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_garage_rock                                     : num [1:13638] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_modern_blues_rock                               : num [1:13638] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_neopsychedelic                                  : num [1:13638] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_nu_gaze                                         : num [1:13638] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_punk_blues                                      : num [1:13638] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre                                                 : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_folk                                            : num [1:13638] 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_folk_rock                                       : num [1:13638] 0 0 0 0 0 0 0 0 1 1 ...
##  $ genre_singersongwriter                                : num [1:13638] 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_soft_rock                                       : num [1:13638] 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_yacht_rock                                      : num [1:13638] 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_bubblegum_pop                                   : num [1:13638] 0 0 0 0 0 0 0 0 0 1 ...
##  $ genre_lounge                                          : num [1:13638] 0 0 0 0 0 0 0 0 0 1 ...
##  $ genre_rockabilly                                      : num [1:13638] 0 0 0 0 0 0 0 0 0 1 ...
##  $ genre_sunshine_pop                                    : num [1:13638] 0 0 0 0 0 0 0 0 0 1 ...
```

```
##  $ genre_canadian_pop                                    : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_boston_rock                                     : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_dance_rock                                      : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_romantic                                    : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_wave                                        : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_wave_pop                                    : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_album_rock                                      : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_synthpop                                        : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_soul                                    : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_country_pop                             : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_outlaw_country                                  : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_texas_country                                   : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_disco                                           : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_postdisco                                       : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_motown                                          : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_soul                                            : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_southern_soul                                   : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_doowop                                          : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rhythm_and_blues                                : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_art_rock                                        : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_rock                                    : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_hard_rock                                       : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_metal                                           : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_psychedelic_rock                                : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rock                                            : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_chicago_soul                                    : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_blues                                           : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_electric_blues                                  : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_jazz_blues                                      : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_louisiana_blues                                 : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_orleans_blues                               : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_piano_blues                                     : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_roots_rock                                      : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_singersongwriter                       : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_canadian_rock                           : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_heartland_rock                                  : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_permanent_wave                                  : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country_rock                                    : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_southern_rock                                   : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_christmas_instrumental                          : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_philly_soul                                     : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_british_invasion                                : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_protopunk                                       : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_vocal_jazz                                      : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_deep_adult_standards                            : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_alternative_metal                               : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##   [list output truncated]
```

```
str(test)
```

```
## tibble [5,847 × 1,695] (S3: tbl_df/tbl/data.frame)
##  $ rating                                                : num [1:5847] 16 26 15 30 62 49 51 40 8 30
...
##  $ track_duration                                        : num [1:5847] 172066 242106 230693 141666 39
4133 ...
##  $ track_explicit                                        : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ danceability                                          : num [1:5847] 0.588 0.615 0.661 0.544 0.519
0.605 0.637 0.718 0.44 0.617 ...
##  $ energy                                                : num [1:5847] 0.672 0.497 0.763 0.725 0.372
0.545 0.528 0.755 0.427 0.197 ...
##  $ key                                                   : num [1:5847] 11 7 6 1 1 0 4 2 0 9 ...
##  $ loudness                                              : num [1:5847] -17.28 -11.91 -12.2 -7.05 -12.
63 ...
##  $ mode                                                  : num [1:5847] 0 1 0 0 1 1 1 1 1 1 ...
##  $ speechiness                                           : num [1:5847] 0.0361 0.439 0.0326 0.0759 0.0
284 0.0297 0.0328 0.0534 0.0254 0.0267 ...
##  $ acousticness                                          : num [1:5847] 0.00256 0.016 0.0205 0.0605 0.
0816 0.0902 0.744 0.281 0.563 0.777 ...
##  $ instrumentalness                                      : num [1:5847] 7.45e-01 0.00 3.07e-04 0.00 4.
67e-05 2.44e-01 0.00 1.18e-05 0.00 0.00 ...
##  $ liveness                                              : num [1:5847] 0.145 0.312 0.382 0.112 0.0652
0.149 0.114 0.0614 0.295 0.133 ...
##  $ valence                                               : num [1:5847] 0.801 0.769 0.905 0.746 0.298
0.865 0.702 0.832 0.335 0.384 ...
##  $ tempo                                                 : num [1:5847] 122 194 150 142 131 ...
##  $ time_signature                                        : num [1:5847] 4 3 4 4 4 4 4 4 3 4 ...
##  $ genre_adult_standards                                 : num [1:5847] 0 0 0 1 0 0 1 0 0 0 ...
##  $ genre_brill_building_pop                              : num [1:5847] 0 0 0 1 0 0 1 0 0 1 ...
##  $ genre_easy_listening                                  : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_mellow_gold                                     : num [1:5847] 0 0 0 0 1 0 1 0 0 0 ...
##  $ genre_rockandroll                                     : num [1:5847] 1 0 0 1 0 0 0 0 0 0 ...
##  $ genre_space_age_pop                                   : num [1:5847] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_surf_music                                      : num [1:5847] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_dance_pop                                       : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_pop                                             : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_postteen_pop                                    : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country                                         : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country_dawn                                    : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_nashville_sound                                 : num [1:5847] 0 0 0 0 0 0 0 0 0 1 ...
##  $ genre_australian_country                              : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_contemporary_country                            : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country_road                                    : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_funk                                            : num [1:5847] 0 0 1 0 0 0 0 0 1 0 ...
##  $ genre_neo_soul                                        : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_jack_swing                                  : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_quiet_storm                                     : num [1:5847] 0 0 1 0 0 0 0 0 1 0 ...
##  $ genre_rb                                              : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_urban_contemporary                              : num [1:5847] 0 0 1 0 0 0 0 0 0 0 ...
##  $ genre_blues_rock                                      : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_garage_rock                                     : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_modern_blues_rock                               : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_nu_gaze                                         : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_punk_blues                                      : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre                                                 : num [1:5847] 0 1 0 0 0 0 0 1 0 0 ...
##  $ genre_folk                                            : num [1:5847] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_folk_rock                                       : num [1:5847] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_singersongwriter                                : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_soft_rock                                       : num [1:5847] 0 0 0 0 1 0 1 0 0 0 ...
##  $ genre_yacht_rock                                      : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_bubblegum_pop                                   : num [1:5847] 0 0 0 1 0 0 1 0 0 0 ...
##  $ genre_lounge                                          : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rockabilly                                      : num [1:5847] 0 0 0 1 0 0 0 0 0 0 ...
##  $ genre_sunshine_pop                                    : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_pop                                    : num [1:5847] 0 0 0 0 1 0 0 0 0 0 ...
```

```
##  $ genre_boston_rock                                  : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_dance_rock                                   : num [1:5847] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_new_romantic                                 : num [1:5847] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_new_wave                                     : num [1:5847] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_new_wave_pop                                 : num [1:5847] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_album_rock                                   : num [1:5847] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_synthpop                                     : num [1:5847] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_classic_soul                                 : num [1:5847] 0 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_classic_country_pop                          : num [1:5847] 0 0 0 0 0 0 0 1 0 0 ...
##  $ genre_outlaw_country                               : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_texas_country                                : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_disco                                        : num [1:5847] 0 0 1 0 0 0 0 0 0 0 ...
##  $ genre_postdisco                                    : num [1:5847] 0 0 1 0 0 0 0 0 0 0 ...
##  $ genre_motown                                       : num [1:5847] 0 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_soul                                         : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_southern_soul                                : num [1:5847] 0 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_doowop                                       : num [1:5847] 0 0 0 1 0 0 0 0 0 0 ...
##  $ genre_rhythm_and_blues                             : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_art_rock                                     : num [1:5847] 0 0 0 0 0 1 0 0 0 0 ...
##  $ genre_classic_rock                                 : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_hard_rock                                    : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_metal                                        : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_psychedelic_rock                             : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rock                                         : num [1:5847] 0 0 0 0 1 1 0 0 0 0 ...
##  $ genre_chicago_soul                                 : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_blues                                        : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_electric_blues                               : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_jazz_blues                                   : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_louisiana_blues                              : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_orleans_blues                            : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_piano_blues                                  : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_roots_rock                                   : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_singersongwriter                    : num [1:5847] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_classic_canadian_rock                        : num [1:5847] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_heartland_rock                               : num [1:5847] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_permanent_wave                               : num [1:5847] 0 0 0 0 0 1 0 0 0 0 ...
##  $ genre_country_rock                                 : num [1:5847] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_southern_rock                                : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_christmas_instrumental                       : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_philly_soul                                  : num [1:5847] 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_british_invasion                             : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_protopunk                                    : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_vocal_jazz                                   : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_deep_adult_standards                         : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_alternative_metal                            : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_metal                               : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##   [list output truncated]
```

Some dummy variables can be lost from removing zero variance. To make sure the variables to be analysed are aligned in both train and test samples, I am going to take their inter-sets of the column names to be the predictors.

```
variable3 <- intersect(names(train), names(test))
train <- train[c(variable3)] %>% relocate(rating, .before=track_duration)
test <- test[c(variable3)] %>% relocate(rating, .before=track_duration)

#same treatment to scoringData
sdata6 <- clean_names(sdata6)
sdata6 <- sdata6[variable3[-1]]
#prepare a set of scoringData with id, to prepare the final step of result extraction
sdata6_id <- clean_names(sdata2)[c(variable3[-1],"id")]

str(train)
```

```
## tibble [13,638 × 1,664] (S3: tbl_df/tbl/data.frame)
##  $ rating                                                : num [1:13638] 36 70 64 19 34 44 34 47 30 55
...
##  $ track_duration                                        : num [1:13638] 166106 211066 208186 182080 3
31466 ...
##  $ track_explicit                                        : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ danceability                                          : num [1:13638] 0.154 0.759 0.613 0.45 0.57
0.612 0.253 0.575 0.34 0.518 ...
##  $ energy                                                : num [1:13638] 0.185 0.699 0.764 0.294 0.629
0.542 0.232 0.434 0.948 0.432 ...
##  $ key                                                   : num [1:13638] 5 0 2 7 9 5 0 5 9 10 ...
##  $ loudness                                              : num [1:13638] -14.06 -5.75 -6.51 -12.02 -7.
61 ...
##  $ mode                                                  : num [1:13638] 1 0 1 1 0 1 1 1 1 0 ...
##  $ speechiness                                           : num [1:13638] 0.0315 0.0307 0.136 0.0318 0.
0331 0.0264 0.0318 0.0312 0.137 0.0459 ...
##  $ acousticness                                          : num [1:13638] 0.911 0.202 0.0527 0.832 0.59
3 0.0781 0.805 0.735 0.0941 0.401 ...
##  $ instrumentalness                                      : num [1:13638] 2.67e-04 1.31e-04 0.00 3.53e-
05 1.36e-04 0.00 1.80e-04 6.59e-05 9.07e-04 0.00 ...
##  $ liveness                                              : num [1:13638] 0.112 0.443 0.197 0.108 0.77
0.0763 0.0939 0.105 0.867 0.299 ...
##  $ valence                                               : num [1:13638] 0.15 0.907 0.417 0.146 0.308
0.433 0.307 0.348 0.604 0.701 ...
##  $ tempo                                                 : num [1:13638] 84 93 160 141 128 ...
##  $ time_signature                                        : num [1:13638] 4 4 4 4 4 4 3 4 4 4 ...
##  $ genre_adult_standards                                 : num [1:13638] 1 0 0 0 0 0 0 0 1 1 ...
##  $ genre_brill_building_pop                              : num [1:13638] 1 0 0 0 0 0 0 0 1 1 ...
##  $ genre_easy_listening                                  : num [1:13638] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_mellow_gold                                     : num [1:13638] 1 0 0 0 0 0 0 0 1 1 ...
##  $ genre_rockandroll                                     : num [1:13638] 0 0 0 0 0 0 0 0 0 1 ...
##  $ genre_space_age_pop                                   : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_surf_music                                      : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_dance_pop                                       : num [1:13638] 0 1 0 0 0 0 0 0 0 0 ...
##  $ genre_pop                                             : num [1:13638] 0 1 1 0 0 0 0 1 0 0 ...
##  $ genre_postteen_pop                                    : num [1:13638] 0 1 1 0 0 0 0 1 0 0 ...
##  $ genre_country                                         : num [1:13638] 0 0 0 1 1 0 0 0 0 0 ...
##  $ genre_country_dawn                                    : num [1:13638] 0 0 0 1 0 0 0 0 0 0 ...
##  $ genre_nashville_sound                                 : num [1:13638] 0 0 0 1 0 0 0 0 0 0 ...
##  $ genre_australian_country                              : num [1:13638] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_contemporary_country                            : num [1:13638] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_country_road                                    : num [1:13638] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_funk                                            : num [1:13638] 0 0 0 0 0 1 0 0 0 0 ...
##  $ genre_neo_soul                                        : num [1:13638] 0 0 0 0 0 1 0 0 0 0 ...
##  $ genre_new_jack_swing                                  : num [1:13638] 0 0 0 0 0 1 0 0 0 0 ...
##  $ genre_quiet_storm                                     : num [1:13638] 0 0 0 0 0 1 0 0 0 0 ...
##  $ genre_rb                                              : num [1:13638] 0 0 0 0 0 1 0 0 0 0 ...
##  $ genre_urban_contemporary                              : num [1:13638] 0 0 0 0 0 1 0 0 0 0 ...
##  $ genre_blues_rock                                      : num [1:13638] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_garage_rock                                     : num [1:13638] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_modern_blues_rock                               : num [1:13638] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_nu_gaze                                         : num [1:13638] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_punk_blues                                      : num [1:13638] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre                                                 : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_folk                                            : num [1:13638] 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_folk_rock                                       : num [1:13638] 0 0 0 0 0 0 0 0 1 1 ...
##  $ genre_singersongwriter                                : num [1:13638] 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_soft_rock                                       : num [1:13638] 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_yacht_rock                                      : num [1:13638] 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_bubblegum_pop                                   : num [1:13638] 0 0 0 0 0 0 0 0 0 1 ...
##  $ genre_lounge                                          : num [1:13638] 0 0 0 0 0 0 0 0 0 1 ...
##  $ genre_rockabilly                                      : num [1:13638] 0 0 0 0 0 0 0 0 0 1 ...
##  $ genre_sunshine_pop                                    : num [1:13638] 0 0 0 0 0 0 0 0 0 1 ...
##  $ genre_canadian_pop                                    : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ genre_boston_rock                                  : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_dance_rock                                   : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_romantic                                 : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_wave                                     : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_wave_pop                                 : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_album_rock                                   : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_synthpop                                     : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_soul                                 : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_country_pop                          : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_outlaw_country                               : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_texas_country                                : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_disco                                        : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_postdisco                                    : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_motown                                       : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_soul                                         : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_southern_soul                                : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_doowop                                       : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rhythm_and_blues                             : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_art_rock                                     : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_rock                                 : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_hard_rock                                    : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_metal                                        : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_psychedelic_rock                             : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rock                                         : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_chicago_soul                                 : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_blues                                        : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_electric_blues                               : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_jazz_blues                                   : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_louisiana_blues                              : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_orleans_blues                            : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_piano_blues                                  : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_roots_rock                                   : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_singersongwriter                    : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_canadian_rock                        : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_heartland_rock                               : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_permanent_wave                               : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country_rock                                 : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_southern_rock                                : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_christmas_instrumental                       : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_philly_soul                                  : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_british_invasion                             : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_protopunk                                    : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_vocal_jazz                                   : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_deep_adult_standards                         : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_alternative_metal                            : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_metal                               : num [1:13638] 0 0 0 0 0 0 0 0 0 0 ...
##   [list output truncated]
```

```
str(test)
```

```
## tibble [5,847 × 1,664] (S3: tbl_df/tbl/data.frame)
##  $ rating                                                : num [1:5847] 16 26 15 30 62 49 51 40 8 30
...
##  $ track_duration                                        : num [1:5847] 172066 242106 230693 141666 39
4133 ...
##  $ track_explicit                                        : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ danceability                                          : num [1:5847] 0.588 0.615 0.661 0.544 0.519
0.605 0.637 0.718 0.44 0.617 ...
##  $ energy                                                : num [1:5847] 0.672 0.497 0.763 0.725 0.372
0.545 0.528 0.755 0.427 0.197 ...
##  $ key                                                   : num [1:5847] 11 7 6 1 1 0 4 2 0 9 ...
##  $ loudness                                              : num [1:5847] -17.28 -11.91 -12.2 -7.05 -12.
63 ...
##  $ mode                                                  : num [1:5847] 0 1 0 0 1 1 1 1 1 1 ...
##  $ speechiness                                           : num [1:5847] 0.0361 0.439 0.0326 0.0759 0.0
284 0.0297 0.0328 0.0534 0.0254 0.0267 ...
##  $ acousticness                                          : num [1:5847] 0.00256 0.016 0.0205 0.0605 0.
0816 0.0902 0.744 0.281 0.563 0.777 ...
##  $ instrumentalness                                      : num [1:5847] 7.45e-01 0.00 3.07e-04 0.00 4.
67e-05 2.44e-01 0.00 1.18e-05 0.00 0.00 ...
##  $ liveness                                              : num [1:5847] 0.145 0.312 0.382 0.112 0.0652
0.149 0.114 0.0614 0.295 0.133 ...
##  $ valence                                               : num [1:5847] 0.801 0.769 0.905 0.746 0.298
0.865 0.702 0.832 0.335 0.384 ...
##  $ tempo                                                 : num [1:5847] 122 194 150 142 131 ...
##  $ time_signature                                        : num [1:5847] 4 3 4 4 4 4 4 4 3 4 ...
##  $ genre_adult_standards                                 : num [1:5847] 0 0 0 1 0 0 1 0 0 0 ...
##  $ genre_brill_building_pop                              : num [1:5847] 0 0 0 1 0 0 1 0 0 1 ...
##  $ genre_easy_listening                                  : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_mellow_gold                                     : num [1:5847] 0 0 0 0 1 0 1 0 0 0 ...
##  $ genre_rockandroll                                     : num [1:5847] 1 0 0 1 0 0 0 0 0 0 ...
##  $ genre_space_age_pop                                   : num [1:5847] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_surf_music                                      : num [1:5847] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_dance_pop                                       : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_pop                                             : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_postteen_pop                                    : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country                                         : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country_dawn                                    : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_nashville_sound                                 : num [1:5847] 0 0 0 0 0 0 0 0 0 1 ...
##  $ genre_australian_country                              : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_contemporary_country                            : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country_road                                    : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_funk                                            : num [1:5847] 0 0 1 0 0 0 0 0 1 0 ...
##  $ genre_neo_soul                                        : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_jack_swing                                  : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_quiet_storm                                     : num [1:5847] 0 0 1 0 0 0 0 0 1 0 ...
##  $ genre_rb                                              : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_urban_contemporary                              : num [1:5847] 0 0 1 0 0 0 0 0 0 0 ...
##  $ genre_blues_rock                                      : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_garage_rock                                     : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_modern_blues_rock                               : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_nu_gaze                                         : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_punk_blues                                      : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre                                                 : num [1:5847] 0 1 0 0 0 0 0 1 0 0 ...
##  $ genre_folk                                            : num [1:5847] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_folk_rock                                       : num [1:5847] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_singersongwriter                                : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_soft_rock                                       : num [1:5847] 0 0 0 0 1 0 1 0 0 0 ...
##  $ genre_yacht_rock                                      : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_bubblegum_pop                                   : num [1:5847] 0 0 0 1 0 0 1 0 0 0 ...
##  $ genre_lounge                                          : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rockabilly                                      : num [1:5847] 0 0 0 1 0 0 0 0 0 0 ...
##  $ genre_sunshine_pop                                    : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_pop                                    : num [1:5847] 0 0 0 0 1 0 0 0 0 0 ...
```

```
##  $ genre_boston_rock                                  : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_dance_rock                                   : num [1:5847] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_new_romantic                                 : num [1:5847] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_new_wave                                     : num [1:5847] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_new_wave_pop                                 : num [1:5847] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_album_rock                                   : num [1:5847] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_synthpop                                     : num [1:5847] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_classic_soul                                 : num [1:5847] 0 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_classic_country_pop                          : num [1:5847] 0 0 0 0 0 0 0 1 0 0 ...
##  $ genre_outlaw_country                               : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_texas_country                                : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_disco                                        : num [1:5847] 0 0 1 0 0 0 0 0 0 0 ...
##  $ genre_postdisco                                    : num [1:5847] 0 0 1 0 0 0 0 0 0 0 ...
##  $ genre_motown                                       : num [1:5847] 0 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_soul                                         : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_southern_soul                                : num [1:5847] 0 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_doowop                                       : num [1:5847] 0 0 0 1 0 0 0 0 0 0 ...
##  $ genre_rhythm_and_blues                             : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_art_rock                                     : num [1:5847] 0 0 0 0 0 1 0 0 0 0 ...
##  $ genre_classic_rock                                 : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_hard_rock                                    : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_metal                                        : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_psychedelic_rock                             : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rock                                         : num [1:5847] 0 0 0 0 1 1 0 0 0 0 ...
##  $ genre_chicago_soul                                 : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_blues                                        : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_electric_blues                               : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_jazz_blues                                   : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_louisiana_blues                              : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_orleans_blues                            : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_piano_blues                                  : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_roots_rock                                   : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_singersongwriter                    : num [1:5847] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_classic_canadian_rock                        : num [1:5847] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_heartland_rock                               : num [1:5847] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_permanent_wave                               : num [1:5847] 0 0 0 0 0 1 0 0 0 0 ...
##  $ genre_country_rock                                 : num [1:5847] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_southern_rock                                : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_christmas_instrumental                       : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_philly_soul                                  : num [1:5847] 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_british_invasion                             : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_protopunk                                    : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_vocal_jazz                                   : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_deep_adult_standards                         : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_alternative_metal                            : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_metal                               : num [1:5847] 0 0 0 0 0 0 0 0 0 0 ...
##   [list output truncated]
```

```
str(sdata6)
```

```
## tibble [4,844 × 1,663] (S3: tbl_df/tbl/data.frame)
##  $ track_duration                                    : num [1:4844] 219813 332226 127829 210973 18
0133 ...
##  $ track_explicit                                    : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ danceability                                      : num [1:4844] 0.647 0.804 0.744 0.712 0.665
0.442 0.61 0.478 0.499 0.801 ...
##  $ energy                                            : num [1:4844] 0.686 0.714 0.47 0.753 0.552
0.717 0.377 0.298 0.249 0.875 ...
##  $ key                                               : num [1:4844] 2 11 7 9 6 9 9 2 5 7 ...
##  $ loudness                                          : num [1:4844] -4.25 -6.71 -10.29 -7.67 -10.1
8 ...
##  $ mode                                              : num [1:4844] 0 0 1 1 1 0 1 1 1 1 ...
##  $ speechiness                                       : num [1:4844] 0.0274 0.183 0.035 0.0399 0.03
07 0.0395 0.0935 0.0342 0.0422 0.0995 ...
##  $ acousticness                                      : num [1:4844] 0.432 0.0567 0.659 0.756 0.493
0.518 0.414 0.74 0.733 0.00366 ...
##  $ instrumentalness                                  : num [1:4844] 6.19e-06 6.21e-06 0.00 1.58e-0
6 2.72e-02 2.62e-01 0.00 1.11e-05 0.00 4.25e-01 ...
##  $ liveness                                          : num [1:4844] 0.133 0.0253 0.256 0.169 0.421
0.184 0.0734 0.139 0.105 0.182 ...
##  $ valence                                           : num [1:4844] 0.952 0.802 0.908 0.953 0.899
0.376 0.519 0.376 0.539 0.637 ...
##  $ tempo                                             : num [1:4844] 155.7 139.7 116.2 116.2 96.9
...
##  $ time_signature                                    : num [1:4844] 4 4 4 4 4 4 4 4 4 4 ...
##  $ genre_adult_standards                             : num [1:4844] 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_brill_building_pop                          : num [1:4844] 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_easy_listening                              : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_mellow_gold                                 : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rockandroll                                 : num [1:4844] 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_space_age_pop                               : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_surf_music                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_dance_pop                                   : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_pop                                         : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_postteen_pop                                : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country                                     : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country_dawn                                : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_nashville_sound                             : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_australian_country                          : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_contemporary_country                        : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country_road                                : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_funk                                        : num [1:4844] 0 1 0 0 1 0 0 0 0 0 ...
##  $ genre_neo_soul                                    : num [1:4844] 0 1 0 0 0 0 0 0 0 0 ...
##  $ genre_new_jack_swing                              : num [1:4844] 0 1 0 0 0 0 0 0 0 0 ...
##  $ genre_quiet_storm                                 : num [1:4844] 0 1 0 0 1 0 0 0 0 0 ...
##  $ genre_rb                                          : num [1:4844] 0 1 0 0 0 0 0 0 0 0 ...
##  $ genre_urban_contemporary                          : num [1:4844] 0 1 0 0 0 0 0 0 0 0 ...
##  $ genre_blues_rock                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_garage_rock                                 : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_modern_blues_rock                           : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_nu_gaze                                     : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_punk_blues                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre                                             : num [1:4844] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_folk                                        : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_folk_rock                                   : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_singersongwriter                            : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_soft_rock                                   : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_yacht_rock                                  : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_bubblegum_pop                               : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_lounge                                      : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rockabilly                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_sunshine_pop                                : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_pop                                : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_boston_rock                                 : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ genre_dance_rock                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_romantic                                : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_wave                                    : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_wave_pop                                : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_album_rock                                  : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_synthpop                                    : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_soul                                : num [1:4844] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_classic_country_pop                         : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_outlaw_country                              : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_texas_country                               : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_disco                                       : num [1:4844] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_postdisco                                   : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_motown                                      : num [1:4844] 0 1 0 0 1 0 0 0 0 0 ...
##  $ genre_soul                                        : num [1:4844] 0 1 0 0 1 0 0 0 0 0 ...
##  $ genre_southern_soul                               : num [1:4844] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_doowop                                      : num [1:4844] 0 0 1 0 0 0 0 0 1 0 ...
##  $ genre_rhythm_and_blues                            : num [1:4844] 0 0 1 0 0 0 0 0 1 0 ...
##  $ genre_art_rock                                    : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_rock                                : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_hard_rock                                   : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_metal                                       : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_psychedelic_rock                            : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rock                                        : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_chicago_soul                                : num [1:4844] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_blues                                       : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_electric_blues                              : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_jazz_blues                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_louisiana_blues                             : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_orleans_blues                           : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_piano_blues                                 : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_roots_rock                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_singersongwriter                   : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_canadian_rock                       : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_heartland_rock                              : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_permanent_wave                              : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country_rock                                : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_southern_rock                               : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_christmas_instrumental                      : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_philly_soul                                 : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_british_invasion                            : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_protopunk                                   : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_vocal_jazz                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_deep_adult_standards                        : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_alternative_metal                           : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_metal                              : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_rock                               : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##   [list output truncated]
```

```
str(sdata6_id)
```

```
## tibble [4,844 × 1,664] (S3: tbl_df/tbl/data.frame)
##  $ track_duration                                    : num [1:4844] 219813 332226 127829 210973 18
0133 ...
##  $ track_explicit                                    : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ danceability                                      : num [1:4844] 0.647 0.804 0.744 0.712 0.665
0.442 0.61 0.478 0.499 0.801 ...
##  $ energy                                            : num [1:4844] 0.686 0.714 0.47 0.753 0.552
0.717 0.377 0.298 0.249 0.875 ...
##  $ key                                               : num [1:4844] 2 11 7 9 6 9 9 2 5 7 ...
##  $ loudness                                          : num [1:4844] -4.25 -6.71 -10.29 -7.67 -10.1
8 ...
##  $ mode                                              : num [1:4844] 0 0 1 1 1 0 1 1 1 1 ...
##  $ speechiness                                       : num [1:4844] 0.0274 0.183 0.035 0.0399 0.03
07 0.0395 0.0935 0.0342 0.0422 0.0995 ...
##  $ acousticness                                      : num [1:4844] 0.432 0.0567 0.659 0.756 0.493
0.518 0.414 0.74 0.733 0.00366 ...
##  $ instrumentalness                                  : num [1:4844] 6.19e-06 6.21e-06 0.00 1.58e-0
6 2.72e-02 2.62e-01 0.00 1.11e-05 0.00 4.25e-01 ...
##  $ liveness                                          : num [1:4844] 0.133 0.0253 0.256 0.169 0.421
0.184 0.0734 0.139 0.105 0.182 ...
##  $ valence                                           : num [1:4844] 0.952 0.802 0.908 0.953 0.899
0.376 0.519 0.376 0.539 0.637 ...
##  $ tempo                                             : num [1:4844] 155.7 139.7 116.2 116.2 96.9
...
##  $ time_signature                                    : num [1:4844] 4 4 4 4 4 4 4 4 4 4 ...
##  $ genre_adult_standards                             : num [1:4844] 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_brill_building_pop                          : num [1:4844] 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_easy_listening                              : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_mellow_gold                                 : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rockandroll                                 : num [1:4844] 0 0 0 0 0 0 0 0 1 0 ...
##  $ genre_space_age_pop                               : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_surf_music                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_dance_pop                                   : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_pop                                         : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_postteen_pop                                : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country                                     : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country_dawn                                : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_nashville_sound                             : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_australian_country                          : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_contemporary_country                        : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country_road                                : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_funk                                        : num [1:4844] 0 1 0 0 1 0 0 0 0 0 ...
##  $ genre_neo_soul                                    : num [1:4844] 0 1 0 0 0 0 0 0 0 0 ...
##  $ genre_new_jack_swing                              : num [1:4844] 0 1 0 0 0 0 0 0 0 0 ...
##  $ genre_quiet_storm                                 : num [1:4844] 0 1 0 0 1 0 0 0 0 0 ...
##  $ genre_rb                                          : num [1:4844] 0 1 0 0 0 0 0 0 0 0 ...
##  $ genre_urban_contemporary                          : num [1:4844] 0 1 0 0 0 0 0 0 0 0 ...
##  $ genre_blues_rock                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_garage_rock                                 : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_modern_blues_rock                           : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_nu_gaze                                     : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_punk_blues                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre                                             : num [1:4844] 0 0 0 0 0 0 1 0 0 0 ...
##  $ genre_folk                                        : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_folk_rock                                   : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_singersongwriter                            : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_soft_rock                                   : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_yacht_rock                                  : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_bubblegum_pop                               : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_lounge                                      : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rockabilly                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_sunshine_pop                                : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_pop                                : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_boston_rock                                 : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ genre_dance_rock                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_romantic                                : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_wave                                    : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_wave_pop                                : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_album_rock                                  : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_synthpop                                    : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_soul                                : num [1:4844] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_classic_country_pop                         : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_outlaw_country                              : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_texas_country                               : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_disco                                       : num [1:4844] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_postdisco                                   : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_motown                                      : num [1:4844] 0 1 0 0 1 0 0 0 0 0 ...
##  $ genre_soul                                        : num [1:4844] 0 1 0 0 1 0 0 0 0 0 ...
##  $ genre_southern_soul                               : num [1:4844] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_doowop                                      : num [1:4844] 0 0 1 0 0 0 0 0 1 0 ...
##  $ genre_rhythm_and_blues                            : num [1:4844] 0 0 1 0 0 0 0 0 1 0 ...
##  $ genre_art_rock                                    : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_rock                                : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_hard_rock                                   : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_metal                                       : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_psychedelic_rock                            : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_rock                                        : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_chicago_soul                                : num [1:4844] 0 0 0 0 1 0 0 0 0 0 ...
##  $ genre_blues                                       : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_electric_blues                              : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_jazz_blues                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_louisiana_blues                             : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_new_orleans_blues                           : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_piano_blues                                 : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_roots_rock                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_singersongwriter                   : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_classic_canadian_rock                       : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_heartland_rock                              : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_permanent_wave                              : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_country_rock                                : num [1:4844] 1 0 0 0 0 0 0 0 0 0 ...
##  $ genre_southern_rock                               : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_christmas_instrumental                      : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_philly_soul                                 : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_british_invasion                            : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_protopunk                                   : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_vocal_jazz                                  : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_deep_adult_standards                        : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_alternative_metal                           : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_metal                              : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##  $ genre_canadian_rock                               : num [1:4844] 0 0 0 0 0 0 0 0 0 0 ...
##   [list output truncated]
```

Feature Selection 2

Considering the large amount of predictors (1,664), it's heavily time-consuming for subset selection or skrinkage. It's also too complicated if we are going to use filter methods, and check the relencancy and non-redundancy among cariables one by one.

However, dimension reduction (Principal Components Analysis) would be an efficient feature selection approach here. With Principal Components Analysis ("PCA"), 1,664 predictors will be reduced to a smaller number (e.g., 70%) of components based on a measure of similarity (e.g., correlation). I am going to use the reduced number of components to predict the the outcome instead of the original set of predictors.

Principal Components Analysis

I did a mistake here. I didn't include brackets while indicating the reduced number of components (1664*0.7) for PCA. Although PCA generates linear combinations of original 1664 predictors for predictive modelling, the number of predictors have not been reduced so that it cannot successfully capture 70% of variance. I think the following modelling process could be affected as well.

I will include the correct code in the last part of this report, which is what I learned this time and what I would like to change if I can re-do the whole data analysis process again.

```
trainPredictors = train[,-1]
pca = prcomp(trainPredictors,scale. = T)
train_components = data.frame(rating = train$rating, cbind(pca$x[,1:1664*0.7]))

testPredictors = test[,-1]
test_pca = predict(pca,newdata=testPredictors)
test_components = data.frame(rating = test$rating, cbind(test_pca[,1:1664*0.7]))

#scoringdata
sdata6_pca = predict(pca,newdata=sdata6)
sdata6_components = data.frame(sdata6_pca[,1:1664*0.7])

str(train_components)
```

```
## 'data.frame':    13638 obs. of  1664 variables:
##  $ rating  : num  36 70 64 19 34 44 34 47 30 55 ...
##  $ PC1     : num  -1.8592 1.7006 1.3401 -0.0335 0.9193 ...
##  $ PC2     : num  2.827 -0.57 -0.57 1.146 0.147 ...
##  $ PC2.1   : num  2.827 -0.57 -0.57 1.146 0.147 ...
##  $ PC3     : num  -0.693 0.967 0.988 1.043 2.46 ...
##  $ PC4     : num  3.049 -0.19 0.353 2.178 1.882 ...
##  $ PC4.1   : num  3.049 -0.19 0.353 2.178 1.882 ...
##  $ PC5     : num  -0.167 -0.824 -0.911 -1.748 -3.594 ...
##  $ PC6     : num  0.0342 -0.8325 -0.5826 -0.5704 -1.1054 ...
##  $ PC7     : num  -0.759 0.996 0.848 1.332 3.289 ...
##  $ PC7.1   : num  -0.759 0.996 0.848 1.332 3.289 ...
##  $ PC8     : num  -0.1046 -0.0297 -0.0856 -0.1552 -0.4854 ...
##  $ PC9     : num  -1.287 0.292 0.361 0.69 2.2 ...
##  $ PC9.1   : num  -1.287 0.292 0.361 0.69 2.2 ...
##  $ PC10    : num  0.196 1.011 0.487 -0.512 -1.437 ...
##  $ PC11    : num  1.016 1.715 0.933 -0.843 -2.872 ...
##  $ PC11.1  : num  1.016 1.715 0.933 -0.843 -2.872 ...
##  $ PC12    : num  0.0109 1.4133 0.9287 -1.4183 -3.5119 ...
##  $ PC13    : num  0.5942 0.3156 0.0914 0.9925 1.0326 ...
##  $ PC14    : num  0.0185 -0.6803 -0.4546 0.8672 1.6325 ...
##  $ PC14.1  : num  0.0185 -0.6803 -0.4546 0.8672 1.6325 ...
##  $ PC15    : num  0.587 0.68 0.608 -0.546 -0.49 ...
##  $ PC16    : num  0.0425 0.8317 0.3756 0.5739 0.2454 ...
##  $ PC16.1  : num  0.0425 0.8317 0.3756 0.5739 0.2454 ...
##  $ PC17    : num  -0.2779 0.2445 -0.0877 -0.0126 2.1888 ...
##  $ PC18    : num  -0.4618 -0.2497 -0.2581 0.0618 1.2131 ...
##  $ PC18.1  : num  -0.4618 -0.2497 -0.2581 0.0618 1.2131 ...
##  $ PC19    : num  0.25607 -0.04965 0.09355 0.19171 0.00803 ...
##  $ PC20    : num  0.1016 -0.02 0.0096 0.0915 0.1455 ...
##  $ PC21    : num  -0.547 0.271 0.32 -0.487 -0.772 ...
##  $ PC21.1  : num  -0.547 0.271 0.32 -0.487 -0.772 ...
##  $ PC22    : num  0.357 0.405 0.391 0.095 -0.246 ...
##  $ PC23    : num  -0.947 1.907 1.696 -0.489 -1.162 ...
##  $ PC23.1  : num  -0.947 1.907 1.696 -0.489 -1.162 ...
##  $ PC24    : num  1.7081 0.0919 0.2845 0.4879 -0.6741 ...
##  $ PC25    : num  -0.9363 -0.2507 -0.313 -0.2383 0.0953 ...
##  $ PC25.1  : num  -0.9363 -0.2507 -0.313 -0.2383 0.0953 ...
##  $ PC26    : num  1.9855 -0.0614 0.2478 0.6597 -1.1524 ...
##  $ PC27    : num  -0.1675 0.1956 0.2394 0.0373 -0.3959 ...
##  $ PC28    : num  -1.196 -0.5386 -0.7753 -0.4742 0.0113 ...
##  $ PC28.1  : num  -1.196 -0.5386 -0.7753 -0.4742 0.0113 ...
##  $ PC29    : num  -0.2165 0.9854 1.0732 -0.0691 -1.1348 ...
##  $ PC30    : num  -0.1785 -0.205 -0.142 0.0179 -0.033 ...
##  $ PC30.1  : num  -0.1785 -0.205 -0.142 0.0179 -0.033 ...
##  $ PC31    : num  0.0724 -0.7284 -0.5649 0.2889 0.9089 ...
##  $ PC32    : num  2.511 -0.463 -0.332 0.786 0.4 ...
##  $ PC32.1  : num  2.511 -0.463 -0.332 0.786 0.4 ...
##  $ PC33    : num  -2.0877 0.2791 -0.0339 -0.473 -0.062 ...
##  $ PC34    : num  1.39 -1.697 -1.274 0.254 0.52 ...
##  $ PC35    : num  0.3239 0.3024 0.2973 0.0274 -0.0261 ...
##  $ PC35.1  : num  0.3239 0.3024 0.2973 0.0274 -0.0261 ...
##  $ PC36    : num  -0.1549 -0.7221 -0.6302 -0.0731 0.2762 ...
##  $ PC37    : num  -0.3118 0.3074 0.3744 0.3637 0.0537 ...
##  $ PC37.1  : num  -0.3118 0.3074 0.3744 0.3637 0.0537 ...
##  $ PC38    : num  0.4986 0.5858 0.5743 0.3054 0.0955 ...
##  $ PC39    : num  -0.638 0.999 0.345 -1.493 1.022 ...
##  $ PC39.1  : num  -0.638 0.999 0.345 -1.493 1.022 ...
##  $ PC40    : num  -0.512 1.133 1.01 -1.001 0.449 ...
##  $ PC41    : num  -0.2518 0.1138 0.0871 0.0934 -0.2201 ...
##  $ PC42    : num  0.192 -0.356 -0.432 0.944 -0.309 ...
##  $ PC42.1  : num  0.192 -0.356 -0.432 0.944 -0.309 ...
##  $ PC43    : num  -0.15543 -0.04387 0.00373 -0.69002 -0.00278 ...
##  $ PC44    : num  -0.448 -0.225 -0.199 0.874 -0.106 ...
```

```
##  $ PC44.1  : num  -0.448 -0.225 -0.199 0.874 -0.106 ...
##  $ PC45    : num  -1.4496 -0.4196 -0.3349 0.0189 -0.5455 ...
##  $ PC46    : num  0.8471 -0.0219 -0.1424 -0.0504 0.1622 ...
##  $ PC46.1  : num  0.8471 -0.0219 -0.1424 -0.0504 0.1622 ...
##  $ PC47    : num  0.594 -0.652 -0.397 -0.831 0.705 ...
##  $ PC48    : num  0.333 1 0.585 0.272 -0.157 ...
##  $ PC49    : num  -0.00741 0.4377 0.20293 0.40344 -0.28689 ...
##  $ PC49.1  : num  -0.00741 0.4377 0.20293 0.40344 -0.28689 ...
##  $ PC50    : num  0.351 -0.694 -0.461 -0.564 0.621 ...
##  $ PC51    : num  -0.2553 0.098 -0.0357 0.0663 -0.4465 ...
##  $ PC51.1  : num  -0.2553 0.098 -0.0357 0.0663 -0.4465 ...
##  $ PC52    : num  -0.1673 -0.3716 -0.1626 0.2762 -0.0169 ...
##  $ PC53    : num  0.741 0.428 0.35 -0.375 0.159 ...
##  $ PC53.1  : num  0.741 0.428 0.35 -0.375 0.159 ...
##  $ PC54    : num  -0.41591 -0.22016 -0.00426 0.47541 0.20739 ...
##  $ PC55    : num  0.0382 -0.2742 -0.1203 -1.8235 1.5838 ...
##  $ PC56    : num  -0.306 0.257 0.173 0.575 -0.729 ...
##  $ PC56.1  : num  -0.306 0.257 0.173 0.575 -0.729 ...
##  $ PC57    : num  -0.281 0.655 0.398 -0.763 -0.251 ...
##  $ PC58    : num  0.2813 -0.2973 -0.0806 0.5422 0.131 ...
##  $ PC58.1  : num  0.2813 -0.2973 -0.0806 0.5422 0.131 ...
##  $ PC59    : num  0.0144 0.5941 0.2724 -0.0546 -0.1046 ...
##  $ PC60    : num  -0.174 -0.769 -0.586 -0.726 1.025 ...
##  $ PC60.1  : num  -0.174 -0.769 -0.586 -0.726 1.025 ...
##  $ PC61    : num  1.933 0.2427 0.4085 0.114 0.0976 ...
##  $ PC62    : num  -0.822 0.762 0.482 0.815 -1.543 ...
##  $ PC62.1  : num  -0.822 0.762 0.482 0.815 -1.543 ...
##  $ PC63    : num  1.038 0.423 0.259 -0.653 0.197 ...
##  $ PC64    : num  0.53 -0.163 -0.193 0.708 0.403 ...
##  $ PC65    : num  0.925 -1.314 -1.002 0.963 -0.658 ...
##  $ PC65.1  : num  0.925 -1.314 -1.002 0.963 -0.658 ...
##  $ PC66    : num  -0.3794 0.1974 0.1355 -0.0368 -0.1525 ...
##  $ PC67    : num  1.033 0.39 0.62 -0.79 0.887 ...
##  $ PC67.1  : num  1.033 0.39 0.62 -0.79 0.887 ...
##  $ PC68    : num  -1.203 0.202 0.11 -0.322 -0.283 ...
##  $ PC69    : num  0.255 -0.458 -0.296 0.33 -0.615 ...
##   [list output truncated]
```

```
str(test_components)
```

```
## 'data.frame':    5847 obs. of  1664 variables:
##  $ rating  : num  16 26 15 30 62 49 51 40 8 30 ...
##  $ PC1     : num  -0.167 0.936 1.286 -1.129 -4.268 ...
##  $ PC2     : num  1.28 -0.07 2.69 3.22 -3.22 ...
##  $ PC2.1   : num  1.28 -0.07 2.69 3.22 -3.22 ...
##  $ PC3     : num  0.2392 0.0661 -0.2797 -1.2293 -0.7576 ...
##  $ PC4     : num  1.214 0.506 -4.987 4.241 -1.137 ...
##  $ PC4.1   : num  1.214 0.506 -4.987 4.241 -1.137 ...
##  $ PC5     : num  -0.382 -0.212 1.035 -0.25 0.472 ...
##  $ PC6     : num  0.0926 0.0688 0.0367 0.4249 -0.0929 ...
##  $ PC7     : num  -0.1801 0.0794 0.5658 -1.452 0.3975 ...
##  $ PC7.1   : num  -0.1801 0.0794 0.5658 -1.452 0.3975 ...
##  $ PC8     : num  -0.0197 0.0129 0.192 -0.4891 0.0979 ...
##  $ PC9     : num  -0.7994 -0.0759 1.298 -2.5902 1.5242 ...
##  $ PC9.1   : num  -0.7994 -0.0759 1.298 -2.5902 1.5242 ...
##  $ PC10    : num  -0.12 -0.307 1.271 1.84 -0.218 ...
##  $ PC11    : num  0.3788 -0.2643 -0.0999 0.5925 0.2449 ...
##  $ PC11.1  : num  0.3788 -0.2643 -0.0999 0.5925 0.2449 ...
##  $ PC12    : num  0.3454 -0.035 -0.0293 0.6367 -0.2364 ...
##  $ PC13    : num  -0.507 -0.304 1.487 -0.952 0.735 ...
##  $ PC14    : num  -0.4 -0.125 0.328 -1.024 0.274 ...
##  $ PC14.1  : num  -0.4 -0.125 0.328 -1.024 0.274 ...
##  $ PC15    : num  0.373 -0.1262 -0.3449 1.7965 -0.0863 ...
##  $ PC16    : num  0.1919 -0.0528 -0.6519 -0.4471 -1.1119 ...
##  $ PC16.1  : num  0.1919 -0.0528 -0.6519 -0.4471 -1.1119 ...
##  $ PC17    : num  -0.0608 -0.347 0.4374 1.4745 -0.9523 ...
##  $ PC18    : num  -0.541 -0.278 0.245 -0.391 -0.351 ...
##  $ PC18.1  : num  -0.541 -0.278 0.245 -0.391 -0.351 ...
##  $ PC19    : num  0.045 0.0777 -0.2595 0.3648 -0.3385 ...
##  $ PC20    : num  -0.1142 -0.0811 -0.1012 0.1323 0.0273 ...
##  $ PC21    : num  -0.0575 -0.0354 -0.1423 -0.2914 0.5555 ...
##  $ PC21.1  : num  -0.0575 -0.0354 -0.1423 -0.2914 0.5555 ...
##  $ PC22    : num  -0.1523 0.0858 -0.4084 -0.7797 1.6893 ...
##  $ PC23    : num  0.354 0.402 -1.352 -0.408 -0.304 ...
##  $ PC23.1  : num  0.354 0.402 -1.352 -0.408 -0.304 ...
##  $ PC24    : num  -0.197 0.275 -0.368 -0.243 3.172 ...
##  $ PC25    : num  1.164 0.265 0.438 1.327 -2.963 ...
##  $ PC25.1  : num  1.164 0.265 0.438 1.327 -2.963 ...
##  $ PC26    : num  -0.751 0.37 0.107 -2.839 4.453 ...
##  $ PC27    : num  0.869 0.319 -0.105 0.251 -1.012 ...
##  $ PC28    : num  0.437 -0.246 1.478 1.523 -1.99 ...
##  $ PC28.1  : num  0.437 -0.246 1.478 1.523 -1.99 ...
##  $ PC29    : num  1.3009 0.5583 -0.0322 0.7669 -0.7013 ...
##  $ PC30    : num  0.3653 0.0528 0.3032 0.8575 -1.6545 ...
##  $ PC30.1  : num  0.3653 0.0528 0.3032 0.8575 -1.6545 ...
##  $ PC31    : num  -0.679 -0.51 0.063 0.918 -0.167 ...
##  $ PC32    : num  0.164 0.194 -0.1 -1.162 -1.98 ...
##  $ PC32.1  : num  0.164 0.194 -0.1 -1.162 -1.98 ...
##  $ PC33    : num  0.178 -0.202 0.268 1.028 1.513 ...
##  $ PC34    : num  -0.3861 0.0408 1.0106 -0.4966 -0.9268 ...
##  $ PC35    : num  -0.21369 0.00447 -0.3235 -0.10398 -0.22548 ...
##  $ PC35.1  : num  -0.21369 0.00447 -0.3235 -0.10398 -0.22548 ...
##  $ PC36    : num  0.6468 0.0647 1.033 0.5777 5 ...
##  $ PC37    : num  -0.55 -0.146 -2.323 0.37 2.016 ...
##  $ PC37.1  : num  -0.55 -0.146 -2.323 0.37 2.016 ...
##  $ PC38    : num  -0.652 -0.055 0.988 -1.51 -5.204 ...
##  $ PC39    : num  -2.182 -0.976 -1.787 0.907 0.604 ...
##  $ PC39.1  : num  -2.182 -0.976 -1.787 0.907 0.604 ...
##  $ PC40    : num  -1.861 -0.582 1.502 0.509 -0.864 ...
##  $ PC41    : num  0.682 0.132 -1.241 -1.996 -3.942 ...
##  $ PC42    : num  0.78 0.174 -1.457 0.612 1.898 ...
##  $ PC42.1  : num  0.78 0.174 -1.457 0.612 1.898 ...
##  $ PC43    : num  0.0787 0.065 1.2388 -0.497 2.9639 ...
##  $ PC44    : num  -0.589 0.23 -1.645 0.108 -2.564 ...
```

```
##  $ PC44.1  : num  -0.589 0.23 -1.645 0.108 -2.564 ...
##  $ PC45    : num  -0.556 0.274 -0.228 -0.554 1.931 ...
##  $ PC46    : num  0.121 -0.49 -0.994 1.334 1.826 ...
##  $ PC46.1  : num  0.121 -0.49 -0.994 1.334 1.826 ...
##  $ PC47    : num  0.5769 -0.2043 -0.0167 1.0592 4.2408 ...
##  $ PC48    : num  -0.818 -0.513 -0.625 0.414 -3.632 ...
##  $ PC49    : num  -0.977 -0.274 -1.036 -0.107 1.306 ...
##  $ PC49.1  : num  -0.977 -0.274 -1.036 -0.107 1.306 ...
##  $ PC50    : num  0.2384 -0.0289 0.0236 0.0189 2.4136 ...
##  $ PC51    : num  -0.281 -0.265 0.413 0.382 0.824 ...
##  $ PC51.1  : num  -0.281 -0.265 0.413 0.382 0.824 ...
##  $ PC52    : num  0.0994 -0.118 0.9531 0.0651 -2.2572 ...
##  $ PC53    : num  0.2028 0.2768 -0.0725 -0.5891 -1.9934 ...
##  $ PC53.1  : num  0.2028 0.2768 -0.0725 -0.5891 -1.9934 ...
##  $ PC54    : num  0.3969 0.0543 0.3106 0.3092 -1.6132 ...
##  $ PC55    : num  -0.1486 -0.0476 -0.9878 0.9303 -3.8938 ...
##  $ PC56    : num  -0.704 -0.161 0.587 0.228 -0.162 ...
##  $ PC56.1  : num  -0.704 -0.161 0.587 0.228 -0.162 ...
##  $ PC57    : num  0.703 0.337 -0.987 1.007 1.745 ...
##  $ PC58    : num  -0.359 -0.044 -1.22 -0.264 -0.111 ...
##  $ PC58.1  : num  -0.359 -0.044 -1.22 -0.264 -0.111 ...
##  $ PC59    : num  0.269 -0.101 0.664 -0.46 1.373 ...
##  $ PC60    : num  0.4024 0.0162 -0.5287 1.5174 -0.2541 ...
##  $ PC60.1  : num  0.4024 0.0162 -0.5287 1.5174 -0.2541 ...
##  $ PC61    : num  -0.658 0.233 -0.872 -1.786 -1.24 ...
##  $ PC62    : num  -0.32 -0.355 -0.267 1.253 -4.276 ...
##  $ PC62.1  : num  -0.32 -0.355 -0.267 1.253 -4.276 ...
##  $ PC63    : num  1.569 0.263 0.981 1.789 2.642 ...
##  $ PC64    : num  -2.177 -0.886 -1.439 -0.202 4.093 ...
##  $ PC65    : num  -0.613 -0.112 -1.379 -1.173 -5.909 ...
##  $ PC65.1  : num  -0.613 -0.112 -1.379 -1.173 -5.909 ...
##  $ PC66    : num  1.2407 0.4445 1.1498 1.0546 0.0171 ...
##  $ PC67    : num  0.273 0.398 -0.293 -2.163 0.468 ...
##  $ PC67.1  : num  0.273 0.398 -0.293 -2.163 0.468 ...
##  $ PC68    : num  -0.0577 0.1398 0.4837 1.5351 1.682 ...
##  $ PC69    : num  0.837 -0.14 1.041 0.512 -2.505 ...
##   [list output truncated]
```

```
str(sdata6_components)
```

```
## 'data.frame':    4844 obs. of  1663 variables:
##  $ PC1     : num  -4.25256 2.97624 0.00695 -0.22294 0.43314 ...
##  $ PC2     : num  -1.859 2.73 2.22 0.338 7.14 ...
##  $ PC2.1   : num  -1.859 2.73 2.22 0.338 7.14 ...
##  $ PC3     : num  -1.0616 -1.1452 -0.0977 0.1097 -1.1206 ...
##  $ PC4     : num  -0.182 -6.566 1.531 0.644 -6.032 ...
##  $ PC4.1   : num  -0.182 -6.566 1.531 0.644 -6.032 ...
##  $ PC5     : num  0.235 1.799 -0.319 -0.329 1.959 ...
##  $ PC6     : num  -0.3027 -0.2987 0.0796 -0.1285 0.8184 ...
##  $ PC7     : num  0.366 0.581 -0.3 0.186 -0.495 ...
##  $ PC7.1   : num  0.366 0.581 -0.3 0.186 -0.495 ...
##  $ PC8     : num  0.08049 0.20436 -0.08548 -0.00493 0.06555 ...
##  $ PC9     : num  -0.894 2.481 -0.904 -0.256 1.236 ...
##  $ PC9.1   : num  -0.894 2.481 -0.904 -0.256 1.236 ...
##  $ PC10    : num  0.834 2.877 0.497 0.202 0.64 ...
##  $ PC11    : num  0.958 1.455 -0.436 0.155 -3.142 ...
##  $ PC11.1  : num  0.958 1.455 -0.436 0.155 -3.142 ...
##  $ PC12    : num  -1.2736 1.3078 0.1293 -0.0466 -1.0395 ...
##  $ PC13    : num  0.3416 4.2028 -0.9346 0.0461 -0.8438 ...
##  $ PC14    : num  0.1937 -0.3602 -0.4231 -0.0653 0.855 ...
##  $ PC14.1  : num  0.1937 -0.3602 -0.4231 -0.0653 0.855 ...
##  $ PC15    : num  -0.331 0.99 0.387 -0.151 -1.318 ...
##  $ PC16    : num  -0.4906 1.087 0.1784 -0.0408 -1.7369 ...
##  $ PC16.1  : num  -0.4906 1.087 0.1784 -0.0408 -1.7369 ...
##  $ PC17    : num  0.3547 1.6079 0.1159 0.0704 -0.41 ...
##  $ PC18    : num  0.4245 0.5379 -0.4229 0.0786 0.2665 ...
##  $ PC18.1  : num  0.4245 0.5379 -0.4229 0.0786 0.2665 ...
##  $ PC19    : num  -0.7895 -0.0516 0.2633 -0.1138 0.1817 ...
##  $ PC20    : num  -0.00202 0.03649 -0.03097 -0.07385 -0.00285 ...
##  $ PC21    : num  -0.0479 -0.85 -0.09 0.0679 0.6116 ...
##  $ PC21.1  : num  -0.0479 -0.85 -0.09 0.0679 0.6116 ...
##  $ PC22    : num  0.8881 -1.6743 -0.169 -0.0607 0.4427 ...
##  $ PC23    : num  -1.014 -3.276 0.796 0.211 1.642 ...
##  $ PC23.1  : num  -1.014 -3.276 0.796 0.211 1.642 ...
##  $ PC24    : num  1.386 -0.401 0.184 -0.573 -0.185 ...
##  $ PC25    : num  -0.883 0.956 0.943 0.27 -0.463 ...
##  $ PC25.1  : num  -0.883 0.956 0.943 0.27 -0.463 ...
##  $ PC26    : num  1.0909 -0.929 -1.5143 0.0864 -0.107 ...
##  $ PC27    : num  -0.8407 -0.4119 0.7051 0.2333 -0.0934 ...
##  $ PC28    : num  2.556 0.651 -0.395 0.348 -0.213 ...
##  $ PC28.1  : num  2.556 0.651 -0.395 0.348 -0.213 ...
##  $ PC29    : num  -0.375 -1.482 0.638 0.438 0.318 ...
##  $ PC30    : num  0.3014 -0.0178 0.0329 0.1593 0.1446 ...
##  $ PC30.1  : num  0.3014 -0.0178 0.0329 0.1593 0.1446 ...
##  $ PC31    : num  -0.0843 1.2116 -0.2218 -0.2627 0.6936 ...
##  $ PC32    : num  -1.45541 0.11663 -0.70845 0.00592 -0.25941 ...
##  $ PC32.1  : num  -1.45541 0.11663 -0.70845 0.00592 -0.25941 ...
##  $ PC33    : num  1.3292 0.00665 0.5635 0.07116 0.06191 ...
##  $ PC34    : num  -0.7296 1.9238 -0.9293 -0.0936 -0.6393 ...
##  $ PC35    : num  -0.1218 -0.2895 -0.0845 -0.0609 0.16 ...
##  $ PC35.1  : num  -0.1218 -0.2895 -0.0845 -0.0609 0.16 ...
##  $ PC36    : num  -0.562 0.5139 -0.1182 -0.0321 -0.4646 ...
##  $ PC37    : num  -1.034 0.522 1.317 -0.232 1.043 ...
##  $ PC37.1  : num  -1.034 0.522 1.317 -0.232 1.043 ...
##  $ PC38    : num  0.0943 -0.5729 -1.6526 0.3332 -0.3347 ...
##  $ PC39    : num  -0.469 -0.879 -0.517 -0.508 1.395 ...
##  $ PC39.1  : num  -0.469 -0.879 -0.517 -0.508 1.395 ...
##  $ PC40    : num  1.5798 -0.203 -1.1428 -0.0903 0.4722 ...
##  $ PC41    : num  0.9505 0.0799 -2.1914 0.3962 -1.0455 ...
##  $ PC42    : num  0.354 -0.1283 1.7675 0.0555 0.1325 ...
##  $ PC42.1  : num  0.354 -0.1283 1.7675 0.0555 0.1325 ...
##  $ PC43    : num  -0.6 0.399 -1.262 0.153 -0.557 ...
##  $ PC44    : num  0.337 -0.108 2.413 0.164 1.256 ...
##  $ PC44.1  : num  0.337 -0.108 2.413 0.164 1.256 ...
```

```
##  $ PC45    : num  0.228 1.085 0.752 0.108 0.735 ...
##  $ PC46    : num  0.27085 0.00554 0.41401 -0.28865 -0.14683 ...
##  $ PC46.1  : num  0.27085 0.00554 0.41401 -0.28865 -0.14683 ...
##  $ PC47    : num  -1.123 1.21 0.219 -0.27 -0.676 ...
##  $ PC48    : num  1.283 -1.059 -0.259 -0.16 0.53 ...
##  $ PC49    : num  1.214 -0.901 -0.592 -0.152 0.172 ...
##  $ PC49.1  : num  1.214 -0.901 -0.592 -0.152 0.172 ...
##  $ PC50    : num  -0.754 0.44 0.2 -0.154 -0.158 ...
##  $ PC51    : num  0.3562 -0.1101 -0.5611 0.0735 -0.0942 ...
##  $ PC51.1  : num  0.3562 -0.1101 -0.5611 0.0735 -0.0942 ...
##  $ PC52    : num  -0.9046 0.5262 -0.2872 0.1204 -0.0144 ...
##  $ PC53    : num  -0.0727 -0.6014 -0.0355 0.0751 0.2268 ...
##  $ PC53.1  : num  -0.0727 -0.6014 -0.0355 0.0751 0.2268 ...
##  $ PC54    : num  0.12488 0.00595 -0.5956 0.18596 -0.86125 ...
##  $ PC55    : num  1.857 0.726 0.831 -0.45 -0.15 ...
##  $ PC56    : num  0.176 0.5161 -0.5767 0.3297 -0.0172 ...
##  $ PC56.1  : num  0.176 0.5161 -0.5767 0.3297 -0.0172 ...
##  $ PC57    : num  1.1945 -0.2624 1.0673 0.0782 -0.4461 ...
##  $ PC58    : num  0.4791 -0.5532 -0.2797 -0.1121 -0.0782 ...
##  $ PC58.1  : num  0.4791 -0.5532 -0.2797 -0.1121 -0.0782 ...
##  $ PC59    : num  -1.0997 0.0937 0.3974 -0.1478 0.5703 ...
##  $ PC60    : num  1.2833 1.5028 0.6461 0.0802 -1.2729 ...
##  $ PC60.1  : num  1.2833 1.5028 0.6461 0.0802 -1.2729 ...
##  $ PC61    : num  -0.286 -0.239 -0.397 -0.575 0.482 ...
##  $ PC62    : num  0.506 0.3929 0.2521 0.0634 -0.284 ...
##  $ PC62.1  : num  0.506 0.3929 0.2521 0.0634 -0.284 ...
##  $ PC63    : num  -1.15 -0.519 0.617 -0.161 -1.341 ...
##  $ PC64    : num  -2.016 0.334 -1.027 -0.467 0.4 ...
##  $ PC65    : num  0.109 0.661 -0.289 -0.304 0.801 ...
##  $ PC65.1  : num  0.109 0.661 -0.289 -0.304 0.801 ...
##  $ PC66    : num  0.72 -0.129 1.126 -0.316 -0.746 ...
##  $ PC67    : num  0.181 -0.198 -1.076 0.739 0.955 ...
##  $ PC67.1  : num  0.181 -0.198 -1.076 0.739 0.955 ...
##  $ PC68    : num  0.462 0.465 0.364 0.208 -1.078 ...
##  $ PC69    : num  0.8104 0.4086 0.3244 -0.4462 -0.0676 ...
##  $ PC70    : num  -1.18245 -0.3473 -0.21923 0.00857 0.28346 ...
##   [list output truncated]
```

The correct version of PCA should be like this, whose following results would be shown in the last part of this report.

```
#trainPredictors = train[,-1]
#pca = prcomp(trainPredictors,scale. = T)
#train_components = data.frame(rating = train$rating, cbind(pca$x[,1:(1664*0.7)]))

#testPredictors = test[,-1]
#test_pca = predict(pca,newdata=testPredictors)
#test_components = data.frame(rating = test$rating, cbind(test_pca[,1:(1664*0.7)]))

##scoringdata
#sdata6_pca = predict(pca,newdata=sdata5)
#sdata6_components = data.frame(sdata5_pca[,1:(1664*0.7)])

#str(train_components)
#str(test_components)
#str(sdata6_components)
```

Data Analysis - Modeling

Now, everything is ready for predictive modeling. Some models usually have higher flexibility and accuracy (eg. Bagging, Boosting, Random Forest, Support Vector Machine), while the other models have higher interpretability (eg. Linear Regression). Since our goal is to improve the predictive accuracy, I prefer the former.

I will still try various models with default model parameters first. Each predictive model's RMSE (root-mean-square error) will be calculated to measure the model's accuracy for prediction. I will compare these models' RMSE, and then pick the models with lowest RMSE (best accuracy) for further parameters turning.

Multiple Regression

Linear Multiple Regression gets a 15.25 test RMSE.

```
lm = lm(rating~.,train_components)

pred_train_lm=predict(lm)
rmse_train_lm=sqrt(mean((pred_train_lm-train_components$rating)^2)); rmse_train_lm
```

```
## [1] 13.85866
```

```
pred_test_lm=predict(lm, newdata=test_components)
```

```
## Warning in predict.lm(lm, newdata = test_components): prediction from a rank-
## deficient fit may be misleading
```

```
rmse_test_lm=sqrt(mean((pred_test_lm-test_components$rating)^2)); rmse_test_lm
```

```
## [1] 15.24716
```

Regression Tree

Regression Tree gets a 15.33 test RMSE.

```
model_tree = rpart(rating~., data=train_components, method = 'anova')

pred_train_tree=predict(model_tree)
rmse_train_tree=sqrt(mean((pred_train_tree-train_components$rating)^2)); rmse_train_tree
```

```
## [1] 15.30466
```

```
pred_test_tree=predict(model_tree, newdata=test_components)
rmse_test_tree=sqrt(mean((pred_test_tree-test_components$rating)^2)); rmse_test_tree
```

```
## [1] 15.3337
```

Random Forest

Bagging (Random Forest) gets a 14.79 test RMSE.

```
set.seed(1031)
rf = randomForest(rating~.,
                  data=train_components,
                  mtry = 12,
                  ntree = 1000)
pred_train_rf = predict(rf)
rmse_train_rf = sqrt(mean((pred_train_rf - train_components$rating)^2)); rmse_train_rf
```

```
## [1] 14.84409
```

```
pred_test_rf = predict(rf, newdata=test_components)
rmse_test_rf = sqrt(mean((pred_test_rf - test_components$rating)^2)); rmse_test_rf
```

```
## [1] 14.79053
```

Ranger

Random Forest (Ranger) gets a 14.75 test RMSE.

```
set.seed(1031)
cv_forest_ranger = ranger(rating ~ .,
                          data=train_components,
                          num.trees = 1000)
```

```
## Growing trees.. Progress: 32%. Estimated remaining time: 1 minute, 7 seconds.
## Growing trees.. Progress: 61%. Estimated remaining time: 38 seconds.
## Growing trees.. Progress: 91%. Estimated remaining time: 8 seconds.
```

```
#test ranger rmse
pred_train = predict(cv_forest_ranger, data = train_components, num.trees = 1000)
rmse_train_cv_forest_ranger = sqrt(mean((pred_train$predictions - train_components$rating)^2)); rmse_train_cv_f
orest_ranger
```

```
## [1] 5.859258
```

```
pred_test = predict(cv_forest_ranger, data = test_components, num.trees = 1000)
rmse_test_cv_forest_ranger = sqrt(mean((pred_test$predictions - test_components$rating)^2)); rmse_test_cv_fores
t_ranger
```

```
## [1] 14.75273
```

XGBoost

Boosting (XGBoost) gets a 15.85 test RMSE.

```
xgboost = xgboost(data=as.matrix(train_components[,-1]),
                  label = train_components$rating,
                  nrounds=10000,
                  verbose = 0,
                  early_stopping_rounds = 100)
xgboost$best_iteration
```

```
## [1] 4656
```

```
#test xgboost rmse
pred_train = predict(xgboost,
               newdata=as.matrix(train_components[,-1]))
rmse_train_xgboost = sqrt(mean((pred_train - train_components$rating)^2)); rmse_train_xgboost
```

```
## [1] 0.3667121
```

```
pred_test = predict(xgboost,
               newdata=as.matrix(test_components[,-1]))
rmse_test_xgboost = sqrt(mean((pred_test - test_components$rating)^2)); rmse_test_xgboost
```

```
## [1] 15.85062
```

gbm

Boosting (gbm) gets a 14.66 test RMSE.

```
set.seed(1031)
cvboost = gbm(rating ~ .,
              data=train_components,
              distribution="gaussian",
              n.trees=500)

#test gbm rmse
pred_train = predict(cvboost, n.trees = 500)
rmse_train_cv_boost = sqrt(mean((pred_train - train_components$rating)^2)); rmse_train_cv_boost
```

```
## [1] 14.05392
```

```
pred_test = predict(cvboost, newdata = test_components, n.trees = 500)
rmse_test_cv_boost = sqrt(mean((pred_test - test_components$rating)^2)); rmse_test_cv_boost
```

```
## [1] 14.65646
```

Radial SVM

Support Vector Machine (Radial) gets a 15.13 test RMSE.

```
svmRadial = svm(rating~.,data = train_components,kernel='radial')
summary(svmRadial)
```

```
##
## Call:
## svm(formula = rating ~ ., data = train_components, kernel = "radial")
##
##
## Parameters:
##    SVM-Type:  eps-regression
##  SVM-Kernel:  radial
##        cost:  1
##       gamma:  0.0006013229
##     epsilon:  0.1
##
##
## Number of Support Vectors:  12549
```

```
#test svm rmse
pred_train = predict(svmRadial)
rmse_train_svm = sqrt(mean((pred_train - train_components$rating)^2)); rmse_train_svm
```

```
## [1] 13.66127
```

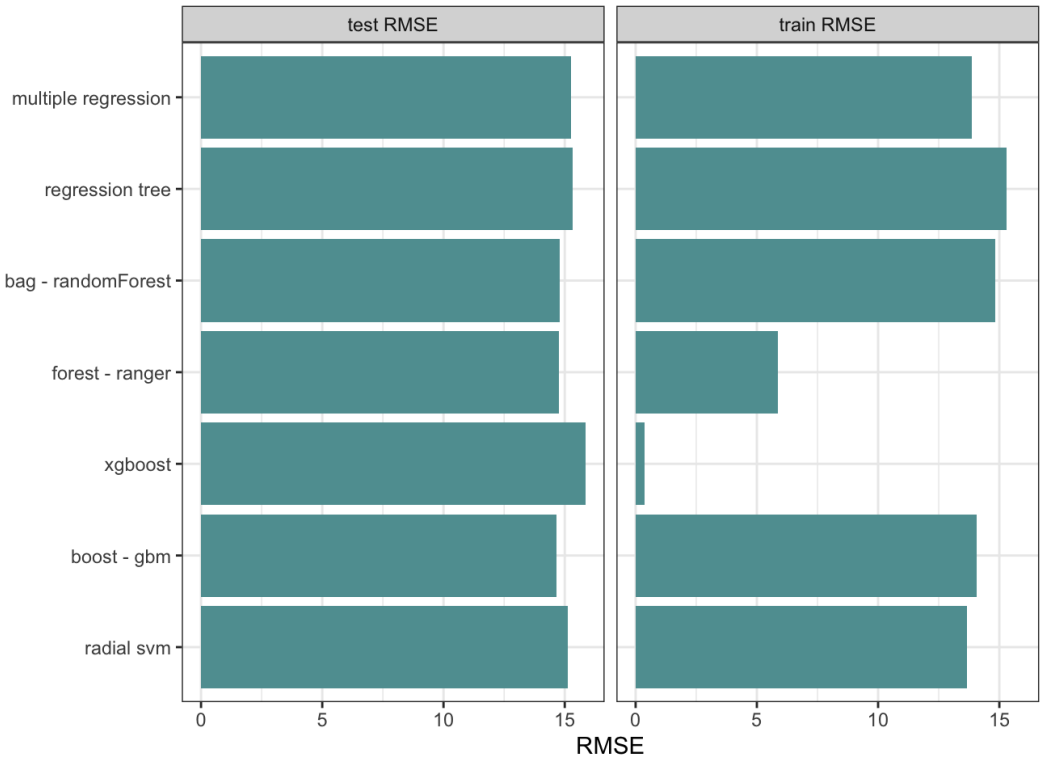
```
pred_test = predict(svmRadial,newdata=test_components)
rmse_test_svm = sqrt(mean((pred_test - test_components$rating)^2)); rmse_test_svm
```

```
## [1] 15.12644
```

Comparison

Among all the models built above, I am going to compare their RMSE with visualization approach.

```
data.frame(
  id = 1:7,
  model = c('multiple regression','regression tree','bag - randomForest','forest - ranger', 'xgboost','boost - gbm','radial svm' ),
  rmse_train = c(rmse_train_lm, rmse_train_tree, rmse_train_rf, rmse_train_cv_forest_ranger, rmse_train_xgboost, rmse_train_cv_boost, rmse_train_svm),
  rmse = c(rmse_test_lm, rmse_test_tree, rmse_test_rf, rmse_test_cv_forest_ranger, rmse_test_xgboost, rmse_test_cv_boost, rmse_test_svm))%>%
  rename('train RMSE' = rmse_train, 'test RMSE' = rmse)%>%
  pivot_longer(cols = 3:4,names_to = 'Sample', values_to = 'RMSE')%>%
  ggplot(aes(x=reorder(model,desc(id)), y = RMSE))+
  geom_col(fill = 'cadetblue')+
  xlab('')+
  coord_flip()+
  theme_bw()+
  facet_wrap(~Sample)
```



Interestingly, eventhough XGBoost get a very low train RMSE (0.37), the test RMSE is the highest (15.85). XGBoost derives the most extreme over-fitting issue here.

We can clearly see that Random Forest, Ranger, and gbm models work the best (get the lowest test RMSE) for test samples. Therefore, we pick these three models for further parameters tuning.

Data Analysis - Model Tuning

I turned several times for Random Forest, Ranger, and gbm models. The best results and the critical steps for the iterative (decreasing test RMSE and improving predictive accuracy of models) process are shown below.

Random Forest - Tuned Model Parameters

Tuned Bagging (Random Forest) gets a 14.75 test RMSE.

```
set.seed(1031)
bag_rf = randomForest(rating~.,
                  data=train_components,
                  mtry = 36,
                  ntree = 1000)
pred_train_rf = predict(bag_rf)
rmse_train_rf = sqrt(mean((pred_train_rf - train_components$rating)^2)); rmse_train_rf
```

```
## [1] 14.81577
```

```
pred_test_rf = predict(bag_rf, newdata=test_components)
rmse_test_rf = sqrt(mean((pred_test_rf - test_components$rating)^2)); rmse_test_rf
```

```
## [1] 14.75382
```

Ranger - Tuned Model Parameters

Tuned Random Forest (Ranger) gets a 14.65 test RMSE.

```
set.seed(627143)
cv_forest_ranger = ranger(rating ~ .,
                          data=train_components,
                          num.trees = 1000,
                          mtry=36,
                          min.node.size = 30,
                          splitrule = "extratrees")
```

```
## Growing trees.. Progress: 61%. Estimated remaining time: 19 seconds.
```

```
#test ranger rmse
pred_train = predict(cv_forest_ranger, data = train_components, num.trees = 1000)
rmse_train_cv_forest_ranger = sqrt(mean((pred_train$predictions - train_components$rating)^2)); rmse_train_cv_f
orest_ranger
```

```
## [1] 11.6772
```

```
pred_test = predict(cv_forest_ranger, data = test_components, num.trees = 1000)
rmse_test_cv_forest_ranger = sqrt(mean((pred_test$predictions - test_components$rating)^2)); rmse_test_cv_fores
t_ranger
```

```
## [1] 14.64986
```

gbm - Tuned Model Parameters

Tuned Boosting (gbm) gets a 14.53 test RMSE.

```
set.seed(1031)
cvboost = gbm(rating ~ .,
              data=train_components,
              distribution="gaussian",
              n.trees=1000,
              interaction.depth=10,
              shrinkage=0.02,
              n.minobsinnode = 1)

#test gbm rmse
pred_train = predict(cvboost, n.trees = 1000)
rmse_train_cv_boost = sqrt(mean((pred_train - train_components$rating)^2)); rmse_train_cv_boost
```

```
## [1] 12.39256
```

```
pred_test = predict(cvboost, newdata = test_components, n.trees = 1000)
rmse_test_cv_boost = sqrt(mean((pred_test - test_components$rating)^2)); rmse_test_cv_boost
```

```
## [1] 14.52983
```

gbm - Changed PCA Parameters

I changed PCA to higher variation (80%) of predictor selection, and the tuned Boosting (gbm) gets a 14.51 test RMSE.

```
#feature selection - pca changed to 80%
#analysisdata
trainPredictors = train[,-1]
pca = prcomp(trainPredictors,scale. = T)
train_components = data.frame(rating = train$rating, cbind(pca$x[,1:1664*0.8]))

testPredictors = test[,-1]
test_pca = predict(pca,newdata=testPredictors)
test_components = data.frame(rating = test$rating, cbind(test_pca[,1:1664*0.8]))

#scoringdata
sdata6_pca = predict(pca,newdata=sdata6)
sdata6_components = data.frame(sdata6_pca[,1:1664*0.8])


#gbm
set.seed(1031)
cvboost = gbm(rating ~ .,
              data=train_components,
              distribution="gaussian",
              n.trees=1000,
              interaction.depth=10,
              shrinkage=0.02,
              n.minobsinnode = 1)

#test gbm rmse
pred_train = predict(cvboost, n.trees = 1000)
rmse_train_cv_boost = sqrt(mean((pred_train - train_components$rating)^2)); rmse_train_cv_boost
```

```
## [1] 12.14556
```

```
pred_test = predict(cvboost, newdata = test_components, n.trees = 1000)
rmse_test_cv_boost = sqrt(mean((pred_test - test_components$rating)^2)); rmse_test_cv_boost
```

```
## [1] 14.50807
```

gbm - Changed Seed

I changed PCA to higher variation (80%) of predictor selection, and then changed seed before training the model. Furthermore, I tuned the model parameters again.

Tuned Boosting (gbm) gets a 14.44 test RMSE, which is the best accuracy I get from the models.

```
set.seed(627143)
cvboost = gbm(rating ~ .,
              data=train_components,
              distribution="gaussian",
              n.trees=500,
              interaction.depth=12,
              shrinkage=0.013,
              n.minobsinnode = 1)

#test gbm rmse
pred_train = predict(cvboost, n.trees = 500)
rmse_train_cv_boost = sqrt(mean((pred_train - train_components$rating)^2)); rmse_train_cv_boost
```

```
## [1] 13.3457
```

```
pred_test = predict(cvboost, newdata = test_components, n.trees = 500)
rmse_test_cv_boost = sqrt(mean((pred_test - test_components$rating)^2)); rmse_test_cv_boost
```

```
## [1] 14.43914
```

Prediction

Finally, I fit this model into scoringData, to generate the prediction data set for final submission.

```
pred = predict(cvboost, newdata = sdata6_components, n.trees = 500)
submissionFile=data.frame(id=sdata6_id$id,rating=pred)
write.csv(submissionFile, file="submission_17_20.csv", row.names=F)
```

Discussion about Mistake (PCA)

Although I made a mistake during the competition, I want to re-do the Principal Components Analysis correctly and see what will be the final results.

```
trainPredictors = train[,-1]
pca = prcomp(trainPredictors,scale. = T)
train_components = data.frame(rating = train$rating, cbind(pca$x[,1:(1664*0.8)]))

testPredictors = test[,-1]
test_pca = predict(pca,newdata=testPredictors)
test_components = data.frame(rating = test$rating, cbind(test_pca[,1:(1664*0.8)]))
str(train_components)
```

```
## 'data.frame':    13638 obs. of  1332 variables:
##  $ rating: num  36 70 64 19 34 44 34 47 30 55 ...
##  $ PC1   : num  -1.8592 1.7006 1.3401 -0.0335 0.9193 ...
##  $ PC2   : num  2.827 -0.57 -0.57 1.146 0.147 ...
##  $ PC3   : num  -0.693 0.967 0.988 1.043 2.46 ...
##  $ PC4   : num  3.049 -0.19 0.353 2.178 1.882 ...
##  $ PC5   : num  -0.167 -0.824 -0.911 -1.748 -3.594 ...
##  $ PC6   : num  0.0342 -0.8325 -0.5826 -0.5704 -1.1054 ...
##  $ PC7   : num  -0.759 0.996 0.848 1.332 3.289 ...
##  $ PC8   : num  -0.1046 -0.0297 -0.0856 -0.1552 -0.4854 ...
##  $ PC9   : num  -1.287 0.292 0.361 0.69 2.2 ...
##  $ PC10  : num  0.196 1.011 0.487 -0.512 -1.437 ...
##  $ PC11  : num  1.016 1.715 0.933 -0.843 -2.872 ...
##  $ PC12  : num  0.0109 1.4133 0.9287 -1.4183 -3.5119 ...
##  $ PC13  : num  0.5942 0.3156 0.0914 0.9925 1.0326 ...
##  $ PC14  : num  0.0185 -0.6803 -0.4546 0.8672 1.6325 ...
##  $ PC15  : num  0.587 0.68 0.608 -0.546 -0.49 ...
##  $ PC16  : num  0.0425 0.8317 0.3756 0.5739 0.2454 ...
##  $ PC17  : num  -0.2779 0.2445 -0.0877 -0.0126 2.1888 ...
##  $ PC18  : num  -0.4618 -0.2497 -0.2581 0.0618 1.2131 ...
##  $ PC19  : num  0.25607 -0.04965 0.09355 0.19171 0.00803 ...
##  $ PC20  : num  0.1016 -0.02 0.0096 0.0915 0.1455 ...
##  $ PC21  : num  -0.547 0.271 0.32 -0.487 -0.772 ...
##  $ PC22  : num  0.357 0.405 0.391 0.095 -0.246 ...
##  $ PC23  : num  -0.947 1.907 1.696 -0.489 -1.162 ...
##  $ PC24  : num  1.7081 0.0919 0.2845 0.4879 -0.6741 ...
##  $ PC25  : num  -0.9363 -0.2507 -0.313 -0.2383 0.0953 ...
##  $ PC26  : num  1.9855 -0.0614 0.2478 0.6597 -1.1524 ...
##  $ PC27  : num  -0.1675 0.1956 0.2394 0.0373 -0.3959 ...
##  $ PC28  : num  -1.196 -0.5386 -0.7753 -0.4742 0.0113 ...
##  $ PC29  : num  -0.2165 0.9854 1.0732 -0.0691 -1.1348 ...
##  $ PC30  : num  -0.1785 -0.205 -0.142 0.0179 -0.033 ...
##  $ PC31  : num  0.0724 -0.7284 -0.5649 0.2889 0.9089 ...
##  $ PC32  : num  2.511 -0.463 -0.332 0.786 0.4 ...
##  $ PC33  : num  -2.0877 0.2791 -0.0339 -0.473 -0.062 ...
##  $ PC34  : num  1.39 -1.697 -1.274 0.254 0.52 ...
##  $ PC35  : num  0.3239 0.3024 0.2973 0.0274 -0.0261 ...
##  $ PC36  : num  -0.1549 -0.7221 -0.6302 -0.0731 0.2762 ...
##  $ PC37  : num  -0.3118 0.3074 0.3744 0.3637 0.0537 ...
##  $ PC38  : num  0.4986 0.5858 0.5743 0.3054 0.0955 ...
##  $ PC39  : num  -0.638 0.999 0.345 -1.493 1.022 ...
##  $ PC40  : num  -0.512 1.133 1.01 -1.001 0.449 ...
##  $ PC41  : num  -0.2518 0.1138 0.0871 0.0934 -0.2201 ...
##  $ PC42  : num  0.192 -0.356 -0.432 0.944 -0.309 ...
##  $ PC43  : num  -0.15543 -0.04387 0.00373 -0.69002 -0.00278 ...
##  $ PC44  : num  -0.448 -0.225 -0.199 0.874 -0.106 ...
##  $ PC45  : num  -1.4496 -0.4196 -0.3349 0.0189 -0.5455 ...
##  $ PC46  : num  0.8471 -0.0219 -0.1424 -0.0504 0.1622 ...
##  $ PC47  : num  0.594 -0.652 -0.397 -0.831 0.705 ...
##  $ PC48  : num  0.333 1 0.585 0.272 -0.157 ...
##  $ PC49  : num  -0.00741 0.4377 0.20293 0.40344 -0.28689 ...
##  $ PC50  : num  0.351 -0.694 -0.461 -0.564 0.621 ...
##  $ PC51  : num  -0.2553 0.098 -0.0357 0.0663 -0.4465 ...
##  $ PC52  : num  -0.1673 -0.3716 -0.1626 0.2762 -0.0169 ...
##  $ PC53  : num  0.741 0.428 0.35 -0.375 0.159 ...
##  $ PC54  : num  -0.41591 -0.22016 -0.00426 0.47541 0.20739 ...
##  $ PC55  : num  0.0382 -0.2742 -0.1203 -1.8235 1.5838 ...
##  $ PC56  : num  -0.306 0.257 0.173 0.575 -0.729 ...
##  $ PC57  : num  -0.281 0.655 0.398 -0.763 -0.251 ...
##  $ PC58  : num  0.2813 -0.2973 -0.0806 0.5422 0.131 ...
##  $ PC59  : num  0.0144 0.5941 0.2724 -0.0546 -0.1046 ...
##  $ PC60  : num  -0.174 -0.769 -0.586 -0.726 1.025 ...
##  $ PC61  : num  1.933 0.2427 0.4085 0.114 0.0976 ...
##  $ PC62  : num  -0.822 0.762 0.482 0.815 -1.543 ...
```

```
##  $ PC63  : num  1.038 0.423 0.259 -0.653 0.197 ...
##  $ PC64  : num  0.53 -0.163 -0.193 0.708 0.403 ...
##  $ PC65  : num  0.925 -1.314 -1.002 0.963 -0.658 ...
##  $ PC66  : num  -0.3794 0.1974 0.1355 -0.0368 -0.1525 ...
##  $ PC67  : num  1.033 0.39 0.62 -0.79 0.887 ...
##  $ PC68  : num  -1.203 0.202 0.11 -0.322 -0.283 ...
##  $ PC69  : num  0.255 -0.458 -0.296 0.33 -0.615 ...
##  $ PC70  : num  1.175 -0.663 -0.308 1.051 0.692 ...
##  $ PC71  : num  0.143 0.136 -0.189 -1.008 0.37 ...
##  $ PC72  : num  -0.5969 0.1403 0.0455 -0.5538 0.061 ...
##  $ PC73  : num  -1.3249 0.2513 0.0588 0.7217 -0.5228 ...
##  $ PC74  : num  0.2486 0.3259 -0.0462 0.3912 -0.1207 ...
##  $ PC75  : num  -0.0417 0.2659 0.2658 -0.2175 0.6431 ...
##  $ PC76  : num  -0.64 0.135 -0.11 1.042 -0.742 ...
##  $ PC77  : num  -1.1806 -0.047 -0.0523 0.2187 -0.2662 ...
##  $ PC78  : num  -0.20501 0.00996 0.02038 -0.24466 -1.23561 ...
##  $ PC79  : num  0.416 -0.456 -0.33 -0.097 0.358 ...
##  $ PC80  : num  0.459 -0.692 -0.419 0.57 0.721 ...
##  $ PC81  : num  -0.8881 1.468 0.8268 0.0277 -0.9964 ...
##  $ PC82  : num  0.331 0.319 0.34 1.213 1.82 ...
##  $ PC83  : num  0.931 0.555 0.394 -0.881 0.347 ...
##  $ PC84  : num  0.565 -0.496 -0.307 -0.465 0.887 ...
##  $ PC85  : num  0.477 0.3749 -0.0964 -1.2816 -0.6137 ...
##  $ PC86  : num  -0.137 0.022 -0.046 -0.583 -1.062 ...
##  $ PC87  : num  0.987 0.15 0.21 0.883 1.295 ...
##  $ PC88  : num  0.0628 -0.0108 -0.2665 -0.5979 -0.3226 ...
##  $ PC89  : num  0.29 -0.243 -0.216 0.493 0.736 ...
##  $ PC90  : num  0.317 0.139 0.0857 0.1787 0.7394 ...
##  $ PC91  : num  0.0615 0.1275 0.1615 0.824 1.3633 ...
##  $ PC92  : num  0.2151 -0.2436 -0.0365 0.3731 -0.2237 ...
##  $ PC93  : num  0.501 0.163 0.0951 -0.5237 -1.2581 ...
##  $ PC94  : num  0.188 0.106 0.124 0.812 1.821 ...
##  $ PC95  : num  -0.1634 0.1332 0.0489 0.1145 1.1988 ...
##  $ PC96  : num  -0.3363 -0.0156 -0.0495 -0.8292 -0.8046 ...
##  $ PC97  : num  0.1761 0.1489 0.0858 0.5842 3.3701 ...
##  $ PC98  : num  -1.023 0.283 0.122 1.316 11.083 ...
##   [list output truncated]
```

```
#scoringdata
sdata6_pca = predict(pca,newdata=sdata6)
sdata6_components = data.frame(sdata6_pca[,1:(1664*0.8)])

set.seed(627143)
cvboost = gbm(rating ~ .,
              data=train_components,
              distribution="gaussian",
              n.trees=500,
              interaction.depth=12,
              shrinkage=0.013,
              n.minobsinnode = 1)

#test gbm rmse
pred_train = predict(cvboost, n.trees = 500)
rmse_train_cv_boost = sqrt(mean((pred_train - train_components$rating)^2)); rmse_train_cv_boost
```

```
## [1] 13.3457
```

```
pred_test = predict(cvboost, newdata = test_components, n.trees = 500)
rmse_test_cv_boost = sqrt(mean((pred_test - test_components$rating)^2)); rmse_test_cv_boost
```

```
## [1] 14.43914
```

```
pred = predict(cvboost, newdata = sdata6_components, n.trees = 500)
submissionFile_redo=data.frame(id=sdata6_id$id,rating=pred)
library(diffdf)
diffdf(submissionFile_redo, submissionFile)
```

```
## No issues were found!
```

However, as we can see from the results, there's no difference on model accuracy and final prediction result between more (1664) or less (1165) predictors generated from Principal Components Analysis. As long as the measure of similarity has been considered and the linear combinations has been generated by PCA. The results for predictive models will be the same.