HW5

Raja Prabakaran

2024-06-08

a. Data gathering and integration

```r
bundesliga <- read.csv("Bundesliga_player.csv")
head(bundesliga)
```

```
##   X             name                       full_name age height
## 1 0     Manuel Neuer              Manuel Peter Neuer  37   1.93
## 2 1      Yann Sommer                                  34   1.83
## 3 2     Sven Ulreich                                  34   1.92
## 4 3  Johannes Schenk                                  20   1.91
## 5 4 Matthijs de Ligt                                  23   1.89
## 6 5  Dayot Upamecano Dayotchanculle Oswald Upamecano  24   1.86
##             nationality place_of_birth price max_price               position
## 1               Germany  Gelsenkirchen   7.0      45.0             Goalkeeper
## 2           Switzerland         Morges   5.0      13.0             Goalkeeper
## 3               Germany     Schorndorf   0.9       6.0             Goalkeeper
## 4               Germany    Schweinfurt   0.3       0.3             Goalkeeper
## 5           Netherlands     Leiderdorp  75.0      75.0 Defender - Centre-Back
## 6 France  Guinea-Bissau         Évreux  60.0      60.0 Defender - Centre-Back
##   shirt_nr  foot          club contract_expires joined_club
## 1        1 right Bayern Munich       2024-06-30  2011-07-01
## 2       27 right Bayern Munich       2025-06-30  2023-01-19
## 3       26 right Bayern Munich       2024-06-30  2021-07-01
## 4       35       Bayern Munich       2024-06-30  2022-07-01
## 5        4 right Bayern Munich       2027-06-30  2022-07-19
## 6        2 right Bayern Munich       2026-06-30  2021-07-05
##               player_agent outfitter
## 1          PRO Profil GmbH    adidas
## 2                Relatives      Puma
## 3 BMS Sportconsulting ...    adidas
## 4                   11WINS          
## 5          Rafaela Pimenta    adidas
## 6      Unique Sports Group      Nike
```

b. Data Exploration

##Summary Statistics

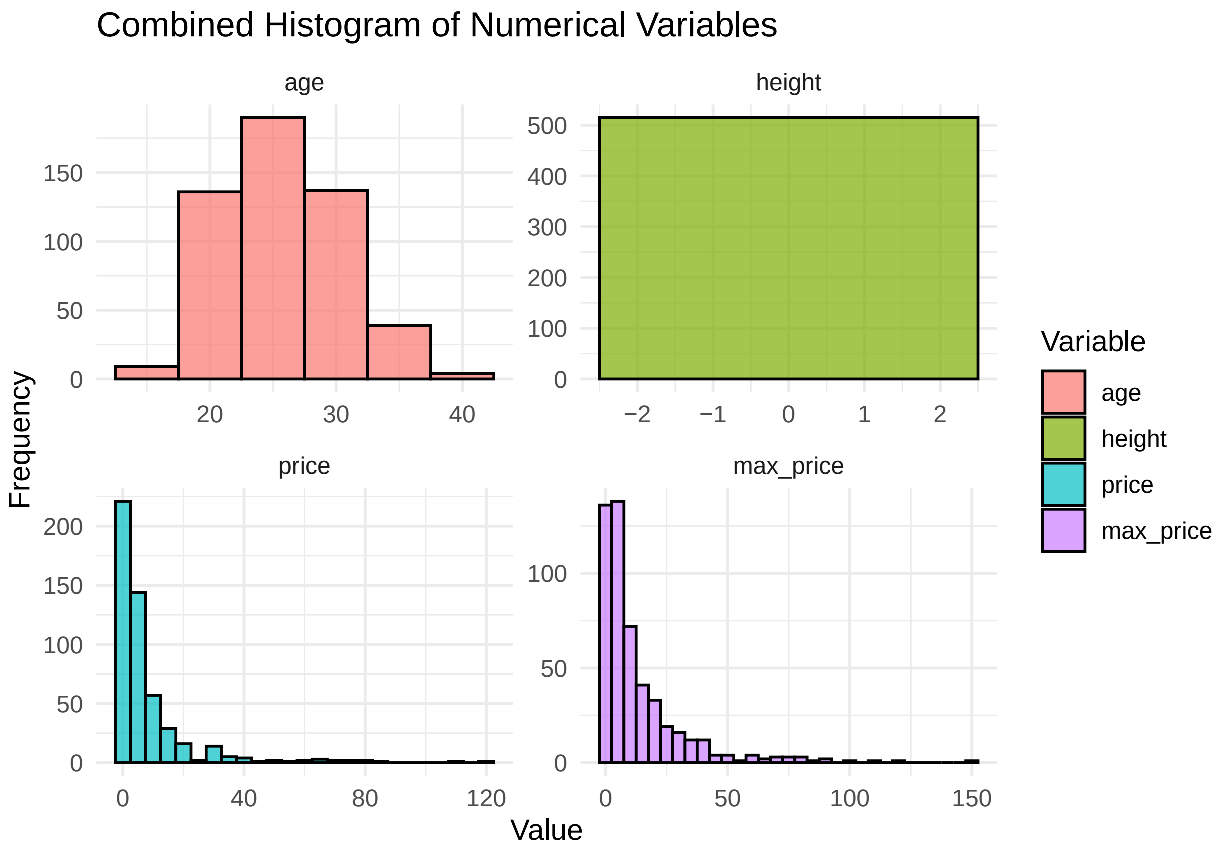
```r
summary(bundesliga)
```

```
##        X             name            full_name              age       
##  Min.   :  0.0   Length:515         Length:515         Min.   :17.00  
##  1st Qu.:128.5   Class :character   Class :character   1st Qu.:22.00  
##  Median :257.0   Mode  :character   Mode  :character   Median :25.00  
##  Mean   :257.0                                         Mean   :25.68  
##  3rd Qu.:385.5                                         3rd Qu.:29.00  
##  Max.   :514.0                                         Max.   :39.00  
##                                                                        
##      height       nationality        place_of_birth         price        
##  Min.   :1.680   Length:515         Length:515         Min.   :  0.025  
##  1st Qu.:1.800   Class :character   Class :character   1st Qu.:  1.200  
##  Median :1.850   Mode  :character   Mode  :character   Median :  3.500  
##  Mean   :1.848                                         Mean   :  8.483  
##  3rd Qu.:1.890                                         3rd Qu.:  9.000  
##  Max.   :2.000                                         Max.   :120.000  
##                                                        NA's   :5        
##    max_price         position            shirt_nr        foot          
##  Min.   :  0.10   Length:515         Min.   : 1.0   Length:515        
##  1st Qu.:  2.50   Class :character   1st Qu.: 9.0   Class :character  
##  Median :  7.00   Mode  :character   Median :20.0   Mode  :character  
##  Mean   : 13.51                      Mean   :19.8                     
##  3rd Qu.: 16.75                      3rd Qu.:29.0                     
##  Max.   :150.00                      Max.   :49.0                     
##  NA's   :5                                                            
##      club           contract_expires   joined_club        player_agent      
##  Length:515         Length:515         Length:515         Length:515        
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##   outfitter        
##  Length:515        
##  Class :character  
##  Mode  :character  
##                    
##                    
##                    
##                    
```

##Distributions of Key Variables

###Combined Histogram for Numerical Variables

```r
numerical_vars <- c("age", "height", "price", "max_price")

df_long <- melt(bundesliga, measure.vars = numerical_vars)

combined_histogram <- ggplot(df_long, aes(x = value, fill = variable)) +
  geom_histogram(binwidth = 5, color = "black", position = "identity", alpha = 0.7) +
  facet_wrap(~ variable, scales = "free") +
  labs(x = "Value", y = "Frequency", fill = "Variable") +
  ggtitle("Combined Histogram of Numerical Variables") +
  theme_minimal()

print(combined_histogram)
```



##Box Plots for Numerical Variables by Position

```
bundesliga$position <- as.factor(bundesliga$position)

plots <- lapply(numerical_vars, function(var) {
  plot_ly(data = bundesliga, x = bundesliga[[var]], y = bundesliga$position, type = "box",
          color = bundesliga$position, colors = "Set3",
          marker = list(line = list(color = 'rgb(0,0,0)', width = 1)),
          boxmean = TRUE) %>%
    layout(title = paste(var, "vs Position"),
           xaxis = list(title = var),
           yaxis = list(title = "Position"),
           template = "plotly_white")
})


plots
```

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

```
##
## [[2]]
##
## [[3]]
##
## [[4]]
```

c. Data Cleaning

##Fixing Typos and Cleaning Strings

```
bundesliga$nationality <- gsub("Â Â ", "", bundesliga$nationality)
bundesliga$place_of_birth <- gsub("Â Â ", "", bundesliga$place_of_birth)
```

##Changing Date Columns to Date Format

```
bundesliga$contract_expires <- as.Date(bundesliga$contract_expires, format="%d/%m/%Y")
bundesliga$joined_club <- as.Date(bundesliga$joined_club, format="%d/%m/%Y")
```

##Removing Unnecessary Columns

```
bundesliga <- bundesliga %>% select(-full_name, -player_agent,-place_of_birth, -contract_expires, -join
```

##Removing Missing Values

```
bundesliga <- na.omit(bundesliga)
```

##Checking for Missing Values and Summary Statistics After Cleaning

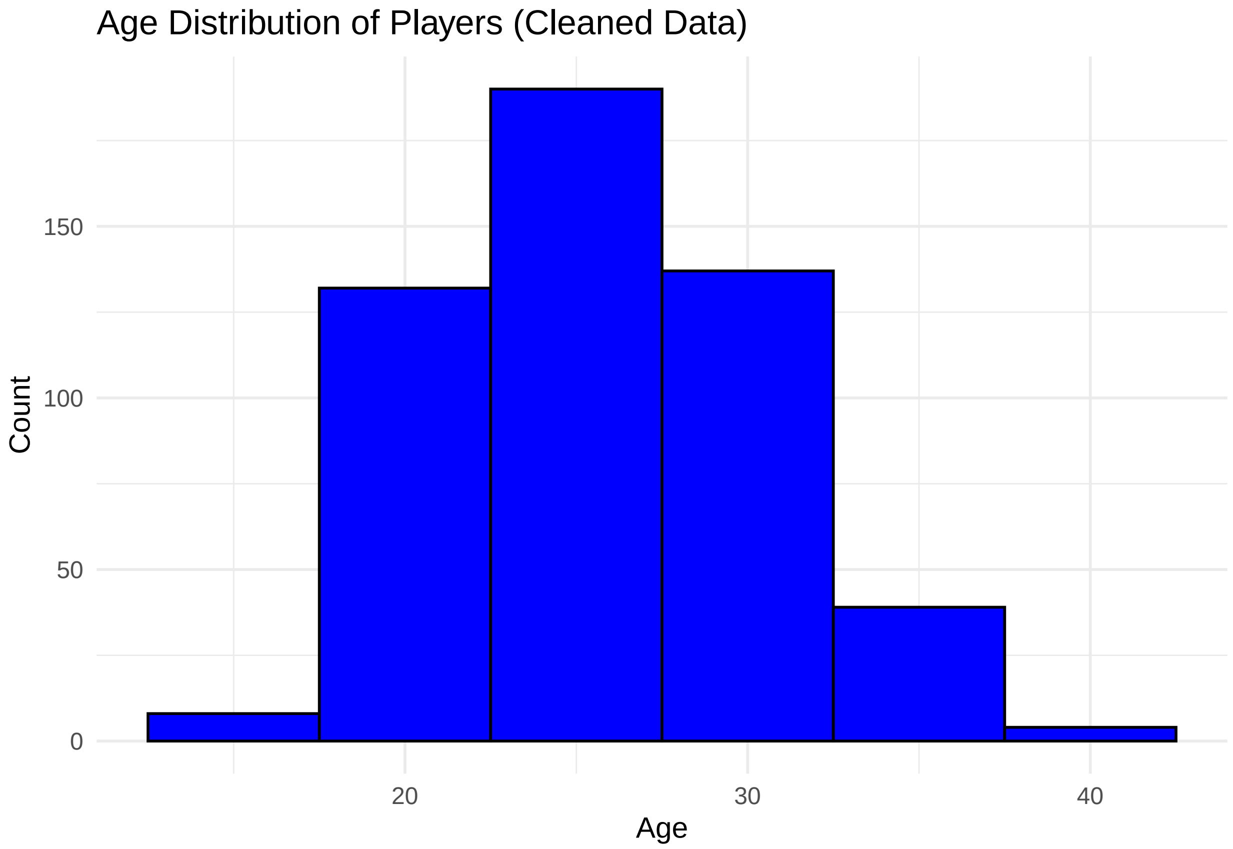
###Cleaning

```
summary(bundesliga)
```

```
##        X             name                age            height
##  Min.   :  0.0   Length:510         Min.   :17.00   Min.   :1.680
##  1st Qu.:128.2   Class :character   1st Qu.:22.00   1st Qu.:1.800
##  Median :257.5   Mode  :character   Median :25.00   Median :1.850
##  Mean   :257.5                      Mean   :25.76   Mean   :1.847
##  3rd Qu.:386.8                      3rd Qu.:29.00   3rd Qu.:1.890
##  Max.   :514.0                      Max.   :39.00   Max.   :2.000
##
##  nationality            price            max_price
##  Length:510         Min.   :  0.025   Min.   :  0.10
##  Class :character   1st Qu.:  1.200   1st Qu.:  2.50
##  Mode  :character   Median :  3.500   Median :  7.00
##                     Mean   :  8.483   Mean   : 13.51
##                     3rd Qu.:  9.000   3rd Qu.: 16.75
##                     Max.   :120.000   Max.   :150.00
##
##                         position      shirt_nr        foot
##  Defender - Centre-Back     : 87   Min.   : 1.0   Length:510
##  Attack - Centre-Forward    : 72   1st Qu.: 9.0   Class :character
##  Goalkeeper                 : 68   Median :19.5   Mode  :character
##  midfield - Central Midfield: 56   Mean   :19.7
##  Defender - Right-Back      : 43   3rd Qu.:29.0
##  Defender - Left-Back       : 40   Max.   :49.0
##  (Other)                    :144
##      club              outfitter
##  Length:510         Length:510
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

###Visualizing Clean Data

```
ggplot(bundesliga, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  theme_minimal() +
  labs(title = "Age Distribution of Players (Cleaned Data)", x = "Age", y = "Count")
```



d. Data Preprocessing

##Normalization

```
bundesliga[numerical_vars] <- scale(bundesliga[numerical_vars])
```

##Creating Dummy Variables for Categorical Columns

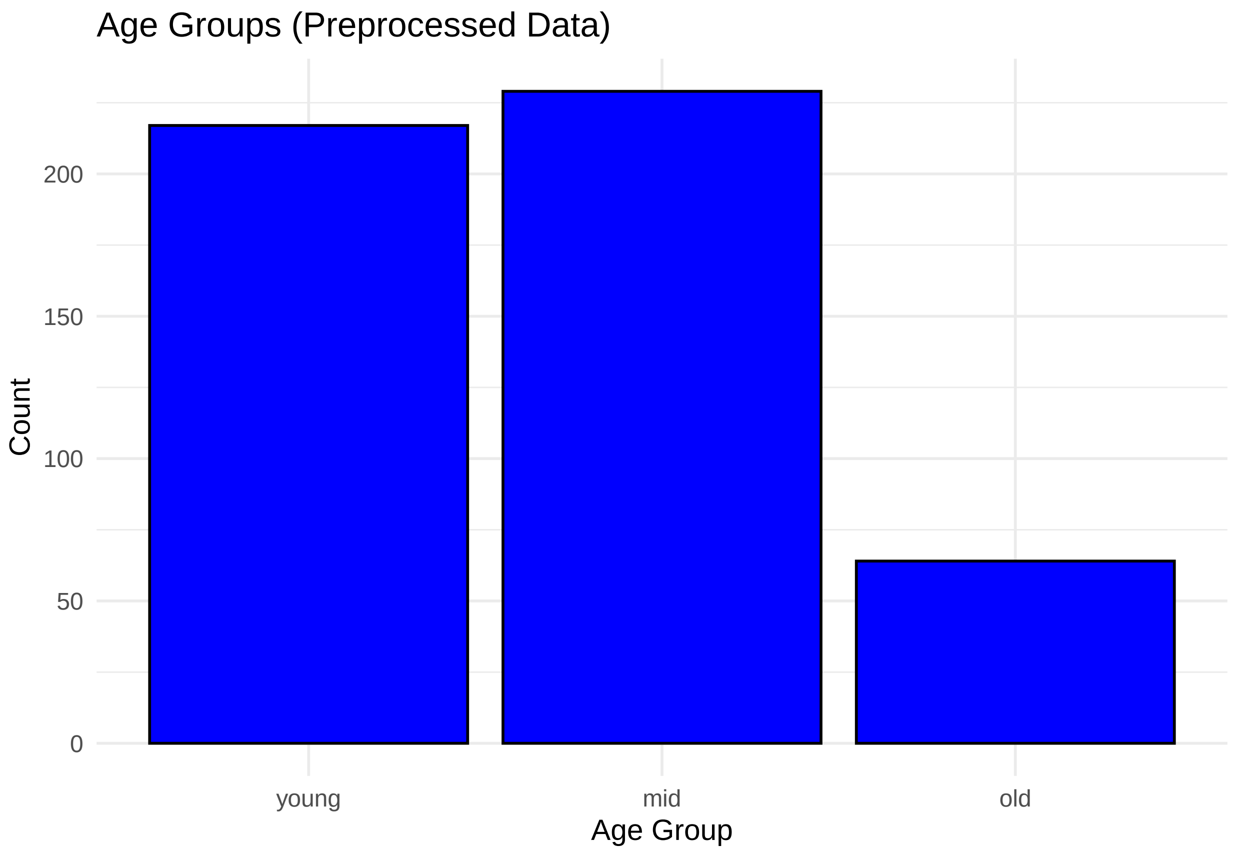
```
categorical_cols <- c('position', 'nationality', 'club')

bundesliga <- bundesliga %>%
  mutate(across(all_of(categorical_cols), as.factor)) %>%
  model.matrix(~.-1, data=.) %>%
  as.data.frame()
```

##Binning 'Age' into Categories: 'young', 'mid', 'old'

```
bundesliga$age_group <- cut(bundesliga$age, breaks = 3, labels = c("young", "mid", "old"))
```

##Visualizing Preprocessed Data

```
ggplot(bundesliga, aes(x = age_group)) +
  geom_bar(fill = "blue", color = "black") +
  theme_minimal() +
  labs(title = "Age Groups (Preprocessed Data)", x = "Age Group", y = "Count")
```



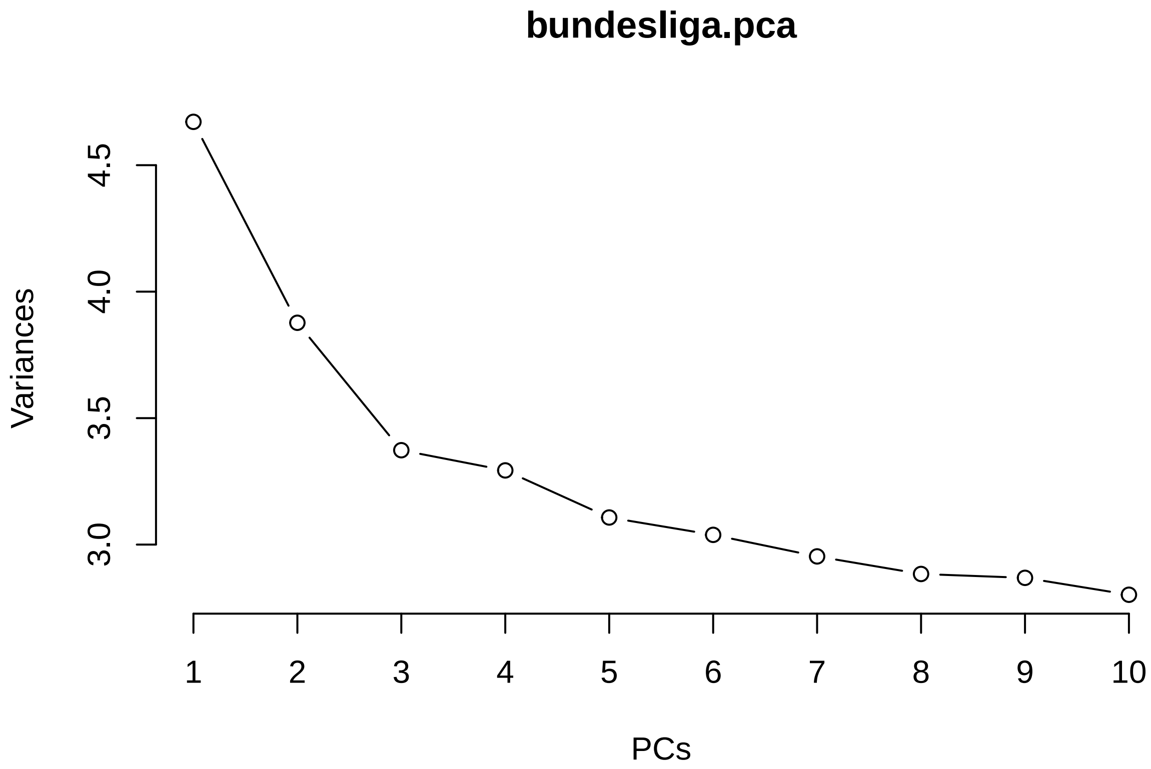
e. Clustering

##PCA and Clustering

```
numeric_cols <- sapply(bundesliga, is.numeric)
bundesliga_numeric <- bundesliga[, numeric_cols]

bundesliga.pca <- prcomp(bundesliga_numeric, center = TRUE, scale. = TRUE)
summary(bundesliga.pca)
```

```
## Importance of components:
##                            PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.16132 1.96901 1.83662 1.81475 1.76272 1.74313 1.71851
## Proportion of Variance 0.00657 0.00545 0.00474 0.00463 0.00437 0.00427 0.00415
## Cumulative Proportion  0.00657 0.01202 0.01677 0.02140 0.02577 0.03004 0.03420
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.69820 1.69369 1.67385 1.66673 1.66665 1.65908 1.65756
## Proportion of Variance 0.00406 0.00403 0.00394 0.00391 0.00391 0.00387 0.00386
## Cumulative Proportion  0.03825 0.04229 0.04623 0.05014 0.05404 0.05791 0.06178
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     1.65621 1.64443 1.64222 1.63545 1.62419 1.61703 1.61045
## Proportion of Variance 0.00386 0.00380 0.00379 0.00376 0.00371 0.00368 0.00365
## Cumulative Proportion  0.06564 0.06944 0.07323 0.07699 0.08070 0.08438 0.08803
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     1.60638 1.60120 1.59469 1.59249 1.58376 1.57615 1.57271
## Proportion of Variance 0.00363 0.00361 0.00358 0.00357 0.00353 0.00349 0.00348
## Cumulative Proportion  0.09166 0.09526 0.09884 0.10241 0.10594 0.10943 0.11291
##                           PC29    PC30    PC31    PC32    PC33    PC34    PC35
## Standard deviation     1.56619 1.56021 1.55920 1.54483 1.54102 1.53848 1.53307
## Proportion of Variance 0.00345 0.00342 0.00342 0.00336 0.00334 0.00333 0.00331
## Cumulative Proportion  0.11636 0.11978 0.12320 0.12656 0.12990 0.13323 0.13653
##                           PC36    PC37    PC38    PC39    PC40   PC41    PC42
## Standard deviation     1.52192 1.51527 1.50549 1.50061 1.49192 1.4849 1.47971
## Proportion of Variance 0.00326 0.00323 0.00319 0.00317 0.00313 0.0031 0.00308
## Cumulative Proportion  0.13979 0.14302 0.14621 0.14937 0.15251 0.1556 0.15869
##                           PC43    PC44    PC45    PC46    PC47    PC48    PC49
## Standard deviation     1.46933 1.44868 1.43144 1.42606 1.42337 1.42101 1.42072
## Proportion of Variance 0.00304 0.00295 0.00288 0.00286 0.00285 0.00284 0.00284
## Cumulative Proportion  0.16172 0.16467 0.16756 0.17042 0.17327 0.17611 0.17894
##                           PC50    PC51    PC52    PC53    PC54    PC55    PC56
## Standard deviation     1.42034 1.41933 1.41913 1.41891 1.41870 1.41855 1.41836
## Proportion of Variance 0.00284 0.00283 0.00283 0.00283 0.00283 0.00283 0.00283
## Cumulative Proportion  0.18178 0.18462 0.18745 0.19028 0.19311 0.19594 0.19877
##                           PC57    PC58    PC59    PC60    PC61    PC62    PC63
## Standard deviation     1.41819 1.41792 1.41786 1.41777 1.41763 1.41763 1.41755
## Proportion of Variance 0.00283 0.00283 0.00283 0.00283 0.00283 0.00283 0.00283
## Cumulative Proportion  0.20160 0.20443 0.20725 0.21008 0.21291 0.21573 0.21856
##                           PC64    PC65    PC66    PC67    PC68    PC69    PC70
## Standard deviation     1.41743 1.41722 1.41719 1.41715 1.41713 1.41701 1.41697
## Proportion of Variance 0.00283 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282
## Cumulative Proportion  0.22139 0.22421 0.22704 0.22986 0.23269 0.23551 0.23833
##                           PC71    PC72    PC73    PC74    PC75    PC76    PC77
## Standard deviation     1.41688 1.41682 1.41677 1.41674 1.41666 1.41662 1.41653
## Proportion of Variance 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282
## Cumulative Proportion  0.24116 0.24398 0.24680 0.24963 0.25245 0.25527 0.25809
##                           PC78    PC79    PC80    PC81    PC82    PC83    PC84
## Standard deviation     1.41651 1.41649 1.41639 1.41633 1.41632 1.41629 1.41626
## Proportion of Variance 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282
## Cumulative Proportion  0.26092 0.26374 0.26656 0.26938 0.27220 0.27502 0.27784
##                           PC85    PC86    PC87    PC88    PC89    PC90    PC91
## Standard deviation     1.41622 1.41619 1.41616 1.41614 1.41608 1.41607 1.41607
## Proportion of Variance 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282
## Cumulative Proportion  0.28067 0.28349 0.28631 0.28913 0.29195 0.29477 0.29759
##                           PC92    PC93    PC94    PC95    PC96    PC97    PC98
## Standard deviation     1.41601 1.41599 1.41598 1.41596 1.41594 1.41591 1.41588
## Proportion of Variance 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282
## Cumulative Proportion  0.30041 0.30323 0.30605 0.30887 0.31169 0.31451 0.31733
##                           PC99   PC100   PC101   PC102   PC103   PC104   PC105
## Standard deviation     1.41587 1.41584 1.41580 1.41577 1.41573 1.41560 1.41560
## Proportion of Variance 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282
## Cumulative Proportion  0.32015 0.32297 0.32579 0.32860 0.33142 0.33424 0.33706
##                          PC106   PC107   PC108   PC109   PC110   PC111   PC112
## Standard deviation     1.41560 1.41560 1.41560 1.41560 1.41560 1.41560 1.41560
## Proportion of Variance 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282
## Cumulative Proportion  0.33988 0.34270 0.34552 0.34833 0.35115 0.35397 0.35679
##                          PC113   PC114   PC115   PC116   PC117   PC118   PC119
## Standard deviation     1.41560 1.41560 1.41560 1.41560 1.41560 1.41560 1.41560
## Proportion of Variance 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282
## Cumulative Proportion  0.35961 0.36243 0.36525 0.36806 0.37088 0.37370 0.37652
##                          PC120   PC121   PC122   PC123   PC124   PC125   PC126
## Standard deviation     1.41560 1.41560 1.41560 1.41560 1.41560 1.41560 1.41560
## Proportion of Variance 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282
## Cumulative Proportion  0.37934 0.38216 0.38497 0.38779 0.39061 0.39343 0.39625
##                          PC127   PC128   PC129   PC130   PC131   PC132   PC133
## Standard deviation     1.41560 1.41560 1.41560 1.41560 1.41560 1.41560 1.41560
## Proportion of Variance 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282
## Cumulative Proportion  0.39907 0.40189 0.40470 0.40752 0.41034 0.41316 0.41598
##                          PC134   PC135   PC136   PC137   PC138   PC139   PC140
## Standard deviation     1.41560 1.41560 1.41560 1.41560 1.41560 1.41560 1.41560
## Proportion of Variance 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282
## Cumulative Proportion  0.41880 0.42161 0.42443 0.42725 0.43007 0.43289 0.43571
##                          PC141   PC142   PC143   PC144   PC145   PC146   PC147
## Standard deviation     1.41560 1.41560 1.41560 1.41560 1.41560 1.41560 1.41560
## Proportion of Variance 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282 0.00282
## Cumulative Proportion  0.43853 0.44134 0.44416 0.44698 0.44980 0.45262 0.45544
##                          PC148   PC149   PC150   PC151   PC152   PC153   PC154
## Standard deviation     1.41560 1.41560 1.41560 1.41560 1.39668 1.34360 1.32378
## Proportion of Variance 0.00282 0.00282 0.00282 0.00282 0.00274 0.00254 0.00246
## Cumulative Proportion  0.45825 0.46107 0.46389 0.46671 0.46945 0.47199 0.47446
##                          PC155   PC156   PC157   PC158   PC159   PC160   PC161
## Standard deviation     1.31205 1.29999 1.29862 1.29355 1.28979 1.28598 1.27650
## Proportion of Variance 0.00242 0.00238 0.00237 0.00235 0.00234 0.00233 0.00229
## Cumulative Proportion  0.47688 0.47926 0.48163 0.48398 0.48632 0.48865 0.49094
##                          PC162   PC163   PC164   PC165  PC166   PC167   PC168
## Standard deviation     1.27068 1.25995 1.25588 1.25241 1.2497 1.24333 1.23534
## Proportion of Variance 0.00227 0.00223 0.00222 0.00221 0.0022 0.00217 0.00215
## Cumulative Proportion  0.49321 0.49544 0.49766 0.49987 0.5021 0.50424 0.50638
##                          PC169  PC170   PC171   PC172   PC173   PC174   PC175
## Standard deviation     1.22913 1.2212 1.21398 1.21321 1.20703 1.20102 1.19399
## Proportion of Variance 0.00212 0.0021 0.00207 0.00207 0.00205 0.00203 0.00201
## Cumulative Proportion  0.50851 0.5106 0.51268 0.51475 0.51680 0.51883 0.52083
##                         PC176   PC177   PC178   PC179   PC180   PC181   PC182
## Standard deviation     1.1911 1.18686 1.18265 1.18165 1.17581 1.16498 1.15797
## Proportion of Variance 0.0020 0.00198 0.00197 0.00196 0.00194 0.00191 0.00189
## Cumulative Proportion  0.5228 0.52481 0.52678 0.52874 0.53068 0.53259 0.53448
##                          PC183   PC184   PC185   PC186   PC187   PC188   PC189
## Standard deviation     1.15401 1.15083 1.14173 1.13725 1.13435 1.12777 1.12432
## Proportion of Variance 0.00187 0.00186 0.00183 0.00182 0.00181 0.00179 0.00178
## Cumulative Proportion  0.53635 0.53821 0.54005 0.54187 0.54368 0.54547 0.54724
##                          PC190   PC191  PC192  PC193   PC194   PC195   PC196
## Standard deviation     1.11835 1.10962 1.0998 1.0985 1.08910 1.04184 1.02642
## Proportion of Variance 0.00176 0.00173 0.0017 0.0017 0.00167 0.00153 0.00148
## Cumulative Proportion  0.54900 0.55073 0.5524 0.5541 0.55580 0.55733 0.55881
##                          PC197   PC198   PC199   PC200   PC201   PC202   PC203
## Standard deviation     1.02458 1.01181 1.00199 1.00099 1.00098 1.00098 1.00098
## Proportion of Variance 0.00148 0.00144 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.56029 0.56173 0.56314 0.56455 0.56596 0.56737 0.56877
##                          PC204   PC205   PC206   PC207   PC208   PC209   PC210
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.57018 0.57159 0.57300 0.57441 0.57582 0.57723 0.57864
##                          PC211   PC212   PC213   PC214   PC215   PC216   PC217
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.58005 0.58146 0.58287 0.58428 0.58569 0.58709 0.58850
##                          PC218   PC219   PC220   PC221   PC222   PC223   PC224
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.58991 0.59132 0.59273 0.59414 0.59555 0.59696 0.59837
##                          PC225   PC226   PC227   PC228   PC229   PC230   PC231
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.59978 0.60119 0.60260 0.60401 0.60541 0.60682 0.60823
##                          PC232   PC233   PC234   PC235   PC236   PC237   PC238
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.60964 0.61105 0.61246 0.61387 0.61528 0.61669 0.61810
##                          PC239   PC240   PC241   PC242   PC243   PC244   PC245
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.61951 0.62092 0.62233 0.62373 0.62514 0.62655 0.62796
##                          PC246   PC247   PC248   PC249   PC250   PC251   PC252
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.62937 0.63078 0.63219 0.63360 0.63501 0.63642 0.63783
##                          PC253   PC254   PC255   PC256   PC257   PC258   PC259
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.63924 0.64065 0.64205 0.64346 0.64487 0.64628 0.64769
##                          PC260   PC261   PC262   PC263   PC264   PC265   PC266
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.64910 0.65051 0.65192 0.65333 0.65474 0.65615 0.65756
##                          PC267   PC268   PC269   PC270   PC271   PC272   PC273
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.65897 0.66037 0.66178 0.66319 0.66460 0.66601 0.66742
##                          PC274   PC275   PC276   PC277   PC278   PC279   PC280
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.66883 0.67024 0.67165 0.67306 0.67447 0.67588 0.67729
##                          PC281   PC282   PC283   PC284   PC285   PC286   PC287
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.67869 0.68010 0.68151 0.68292 0.68433 0.68574 0.68715
##                          PC288   PC289   PC290   PC291   PC292   PC293   PC294
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.68856 0.68997 0.69138 0.69279 0.69420 0.69561 0.69701
##                          PC295   PC296   PC297   PC298   PC299   PC300   PC301
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.69842 0.69983 0.70124 0.70265 0.70406 0.70547 0.70688
##                          PC302   PC303   PC304   PC305   PC306   PC307   PC308
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.70829 0.70970 0.71111 0.71252 0.71393 0.71533 0.71674
##                          PC309   PC310   PC311   PC312   PC313   PC314   PC315
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.71815 0.71956 0.72097 0.72238 0.72379 0.72520 0.72661
##                          PC316   PC317   PC318   PC319   PC320   PC321   PC322
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.72802 0.72943 0.73084 0.73225 0.73365 0.73506 0.73647
##                          PC323   PC324   PC325   PC326   PC327   PC328   PC329
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.73788 0.73929 0.74070 0.74211 0.74352 0.74493 0.74634
##                          PC330   PC331   PC332   PC333   PC334   PC335   PC336
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.74775 0.74916 0.75057 0.75197 0.75338 0.75479 0.75620
##                          PC337   PC338   PC339   PC340   PC341   PC342   PC343
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.75761 0.75902 0.76043 0.76184 0.76325 0.76466 0.76607
##                          PC344   PC345   PC346   PC347   PC348   PC349   PC350
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.76748 0.76889 0.77030 0.77170 0.77311 0.77452 0.77593
##                          PC351   PC352   PC353   PC354   PC355   PC356   PC357
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.77734 0.77875 0.78016 0.78157 0.78298 0.78439 0.78580
##                          PC358   PC359   PC360   PC361   PC362   PC363   PC364
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.78721 0.78862 0.79002 0.79143 0.79284 0.79425 0.79566
##                          PC365   PC366   PC367   PC368   PC369   PC370   PC371
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.79707 0.79848 0.79989 0.80130 0.80271 0.80412 0.80553
##                          PC372   PC373   PC374   PC375   PC376   PC377   PC378
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.80694 0.80834 0.80975 0.81116 0.81257 0.81398 0.81539
##                          PC379   PC380   PC381   PC382   PC383   PC384   PC385
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.81680 0.81821 0.81962 0.82103 0.82244 0.82385 0.82526
##                          PC386   PC387   PC388   PC389   PC390   PC391   PC392
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.82666 0.82807 0.82948 0.83089 0.83230 0.83371 0.83512
##                          PC393   PC394   PC395   PC396   PC397   PC398   PC399
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.83653 0.83794 0.83935 0.84076 0.84217 0.84358 0.84498
##                          PC400   PC401   PC402   PC403   PC404   PC405   PC406
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.84639 0.84780 0.84921 0.85062 0.85203 0.85344 0.85485
##                          PC407   PC408   PC409   PC410   PC411   PC412   PC413
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.85626 0.85767 0.85908 0.86049 0.86190 0.86330 0.86471
##                          PC414   PC415   PC416   PC417   PC418   PC419   PC420
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.86612 0.86753 0.86894 0.87035 0.87176 0.87317 0.87458
##                          PC421   PC422   PC423   PC424   PC425   PC426   PC427
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.87599 0.87740 0.87881 0.88022 0.88162 0.88303 0.88444
##                          PC428   PC429   PC430   PC431   PC432   PC433   PC434
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.88585 0.88726 0.88867 0.89008 0.89149 0.89290 0.89431
##                          PC435   PC436   PC437   PC438   PC439   PC440   PC441
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.89572 0.89713 0.89854 0.89994 0.90135 0.90276 0.90417
##                          PC442   PC443   PC444   PC445   PC446   PC447   PC448
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.90558 0.90699 0.90840 0.90981 0.91122 0.91263 0.91404
##                          PC449   PC450   PC451   PC452   PC453   PC454   PC455
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.91545 0.91686 0.91826 0.91967 0.92108 0.92249 0.92390
##                          PC456   PC457   PC458   PC459   PC460   PC461   PC462
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.92531 0.92672 0.92813 0.92954 0.93095 0.93236 0.93377
##                          PC463   PC464   PC465   PC466   PC467   PC468   PC469
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.93518 0.93658 0.93799 0.93940 0.94081 0.94222 0.94363
##                          PC470   PC471   PC472   PC473   PC474   PC475   PC476
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.94504 0.94645 0.94786 0.94927 0.95068 0.95209 0.95350
##                          PC477   PC478   PC479   PC480   PC481   PC482   PC483
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.95490 0.95631 0.95772 0.95913 0.96054 0.96195 0.96336
##                          PC484   PC485   PC486   PC487   PC488   PC489   PC490
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.96477 0.96618 0.96759 0.96900 0.97041 0.97182 0.97322
##                          PC491   PC492   PC493   PC494   PC495   PC496   PC497
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.97463 0.97604 0.97745 0.97886 0.98027 0.98168 0.98309
##                          PC498   PC499   PC500   PC501   PC502   PC503   PC504
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 1.00098 1.00098
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141 0.00141
## Cumulative Proportion  0.98450 0.98591 0.98732 0.98873 0.99014 0.99154 0.99295
##                          PC505   PC506   PC507   PC508   PC509     PC510
## Standard deviation     1.00098 1.00098 1.00098 1.00098 1.00098 3.499e-15
## Proportion of Variance 0.00141 0.00141 0.00141 0.00141 0.00141 0.000e+00
## Cumulative Proportion  0.99436 0.99577 0.99718 0.99859 1.00000 1.000e+00
```

```r
screeplot(bundesliga.pca, type = "l") + title(xlab = "PCs")
```

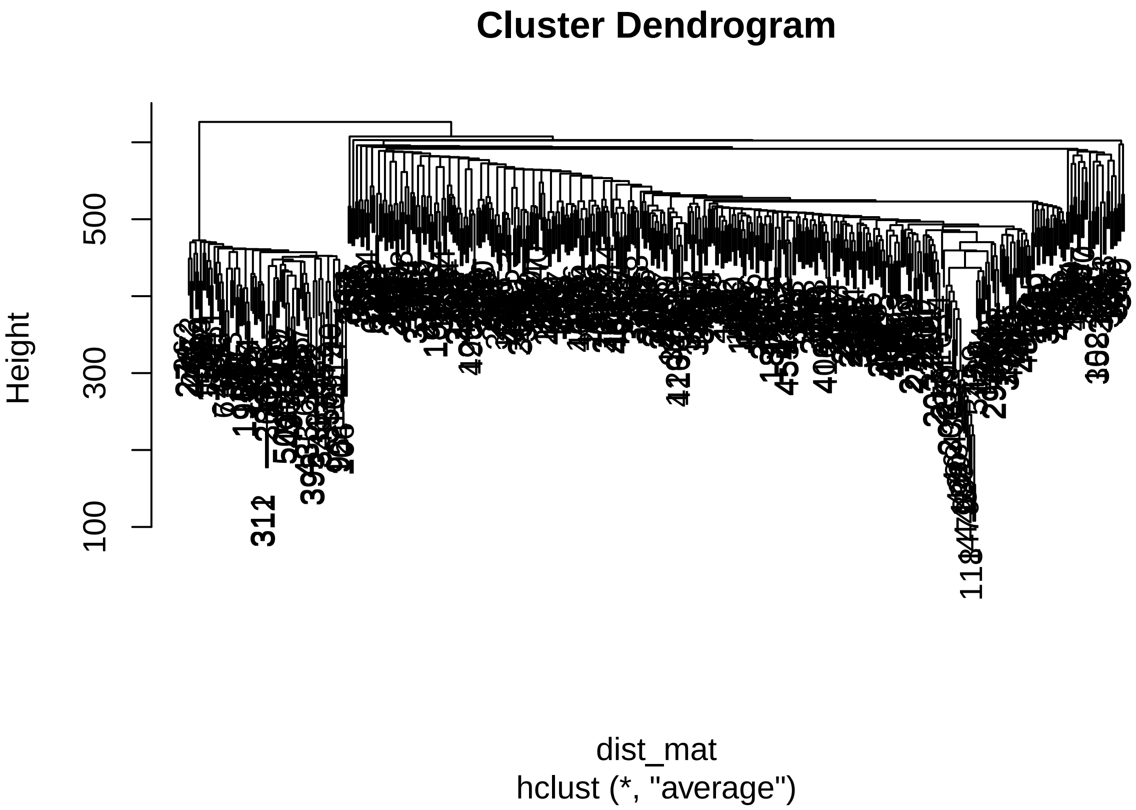


```
## integer(0)
```

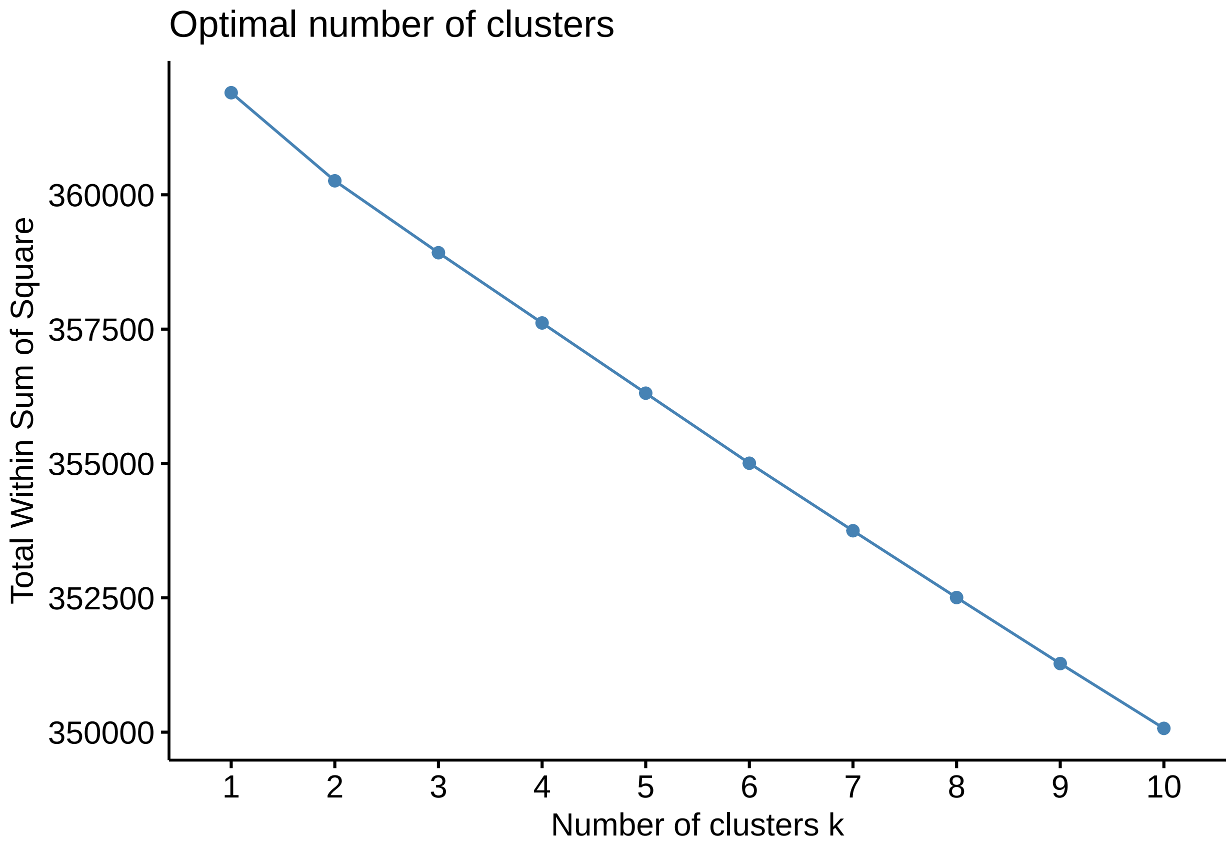
```
bundesliga_pca_df <- as.data.frame(bundesliga.pca$x)
```

Hierarchical Clustering

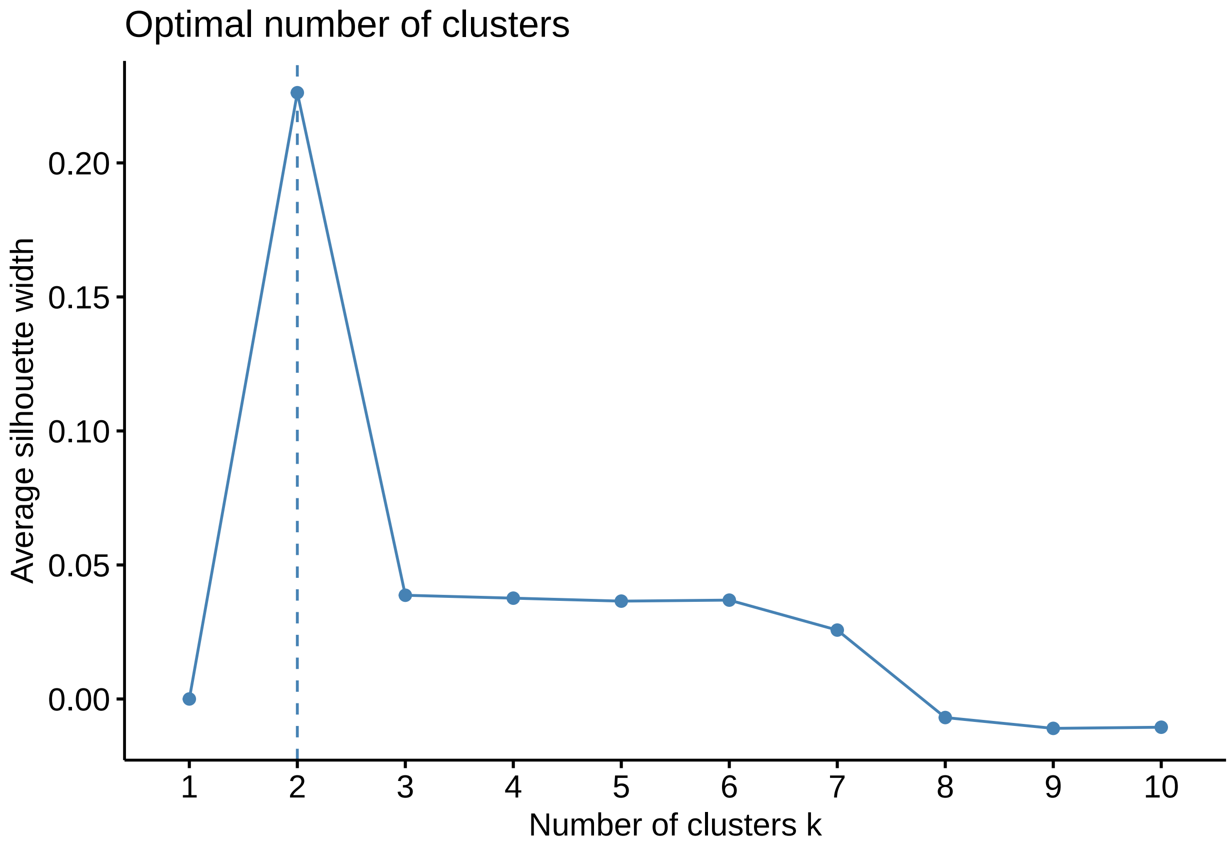
```
dist_mat <- dist(bundesliga_pca_df, method = 'manhattan')
hfit <- hclust(dist_mat, method = 'average')
plot(hfit)
```



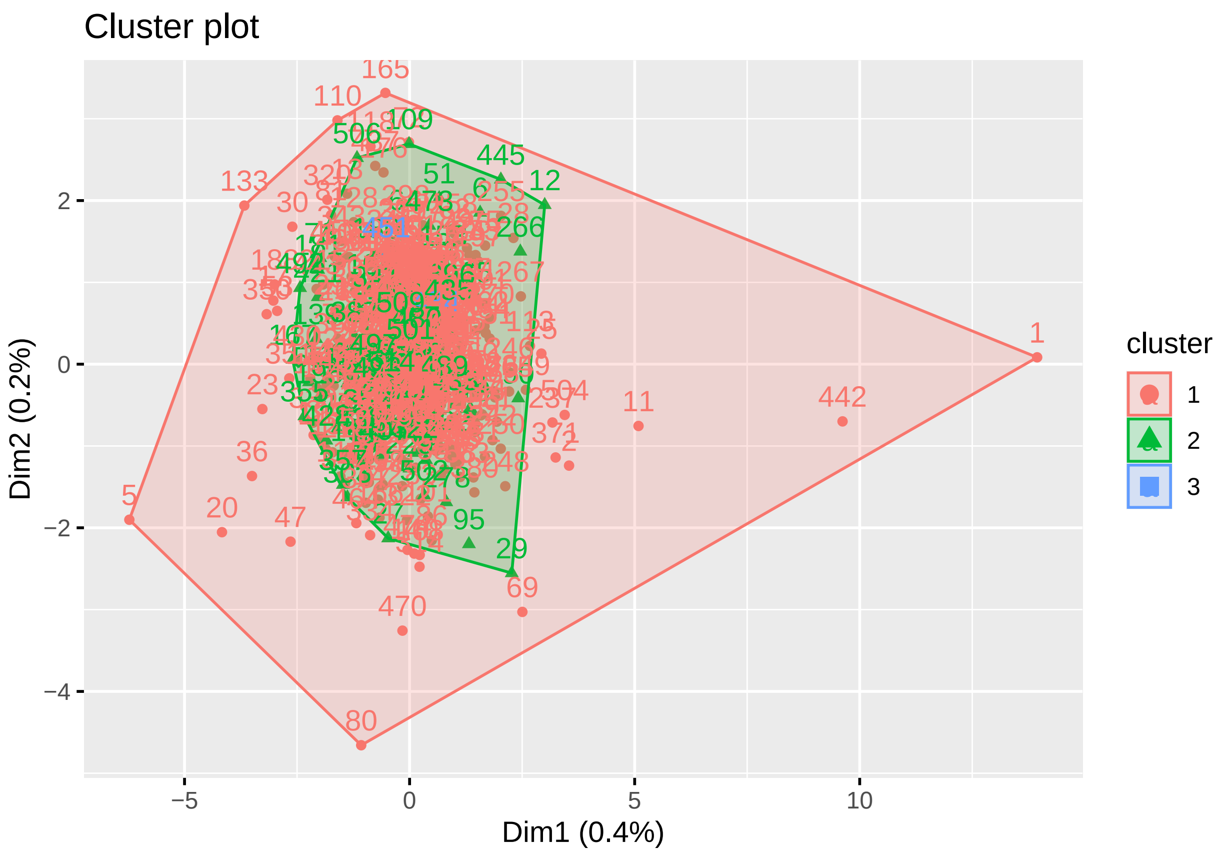
```r
fviz_nbclust(bundesliga_pca_df, FUN = hcut, method = "wss")
```



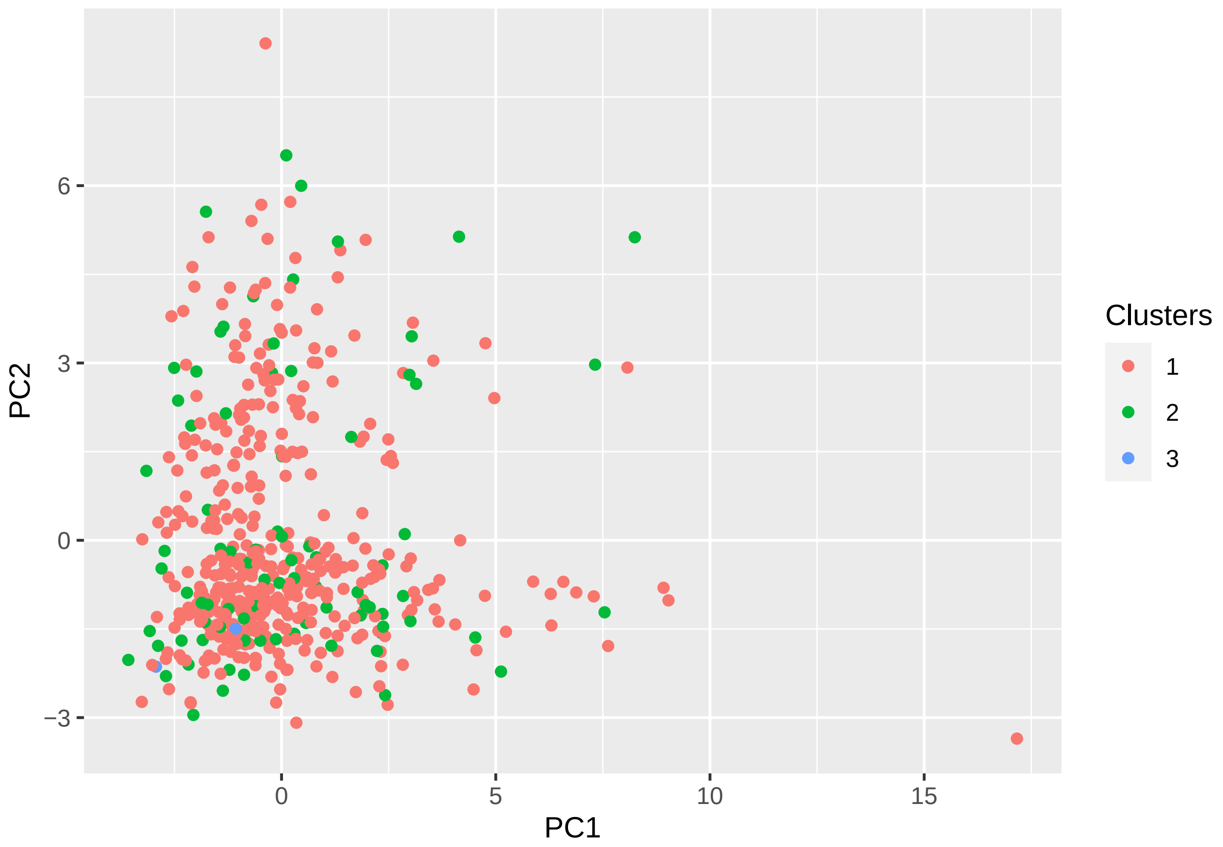
```
fviz_nbclust(bundesliga_pca_df, FUN = hcut, method = "silhouette")
```



```r
h3 <- cutree(hfit, k = 3)
fviz_cluster(list(data = bundesliga_pca_df, cluster = h3))
```

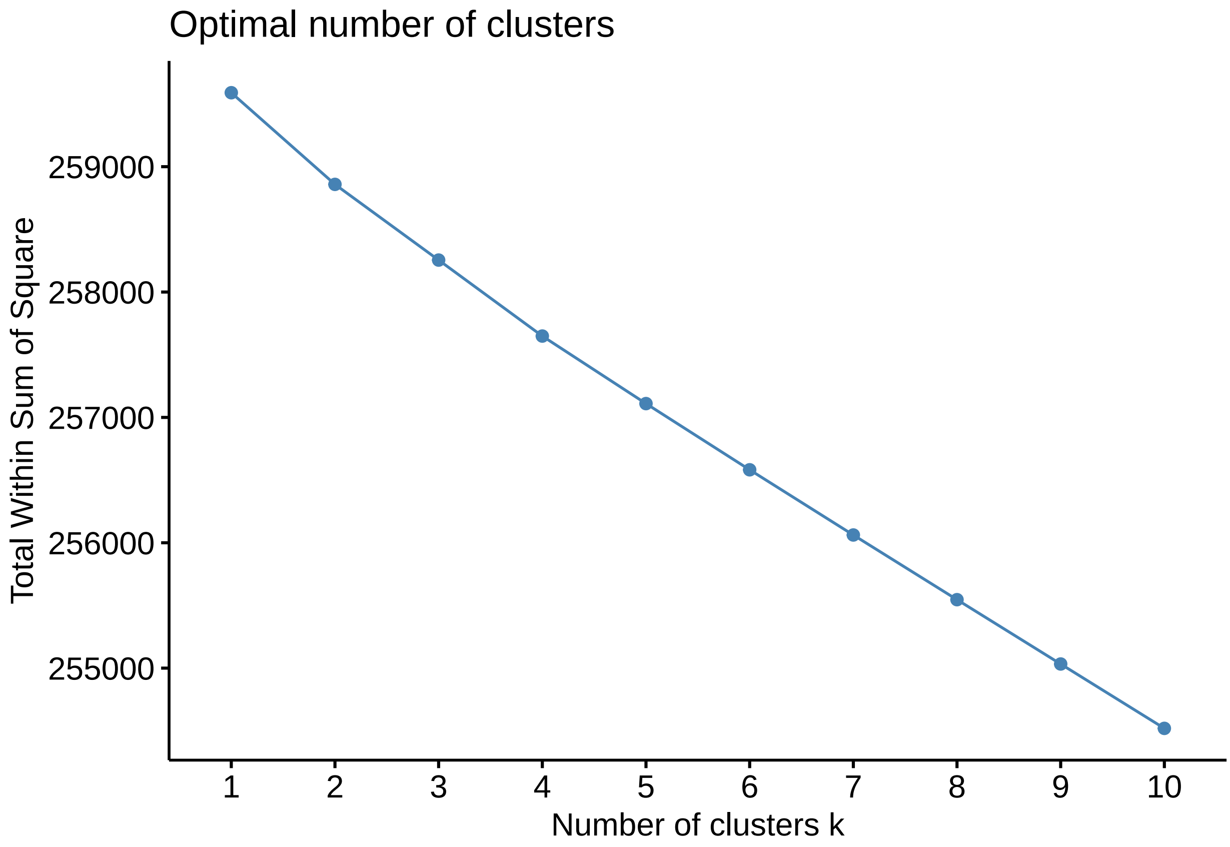


```
bundesliga_pca_df$Clusters <- as.factor(h3)

ggplot(data = bundesliga_pca_df, aes(x = PC1, y = PC2, col = Clusters)) +
  geom_point()
```

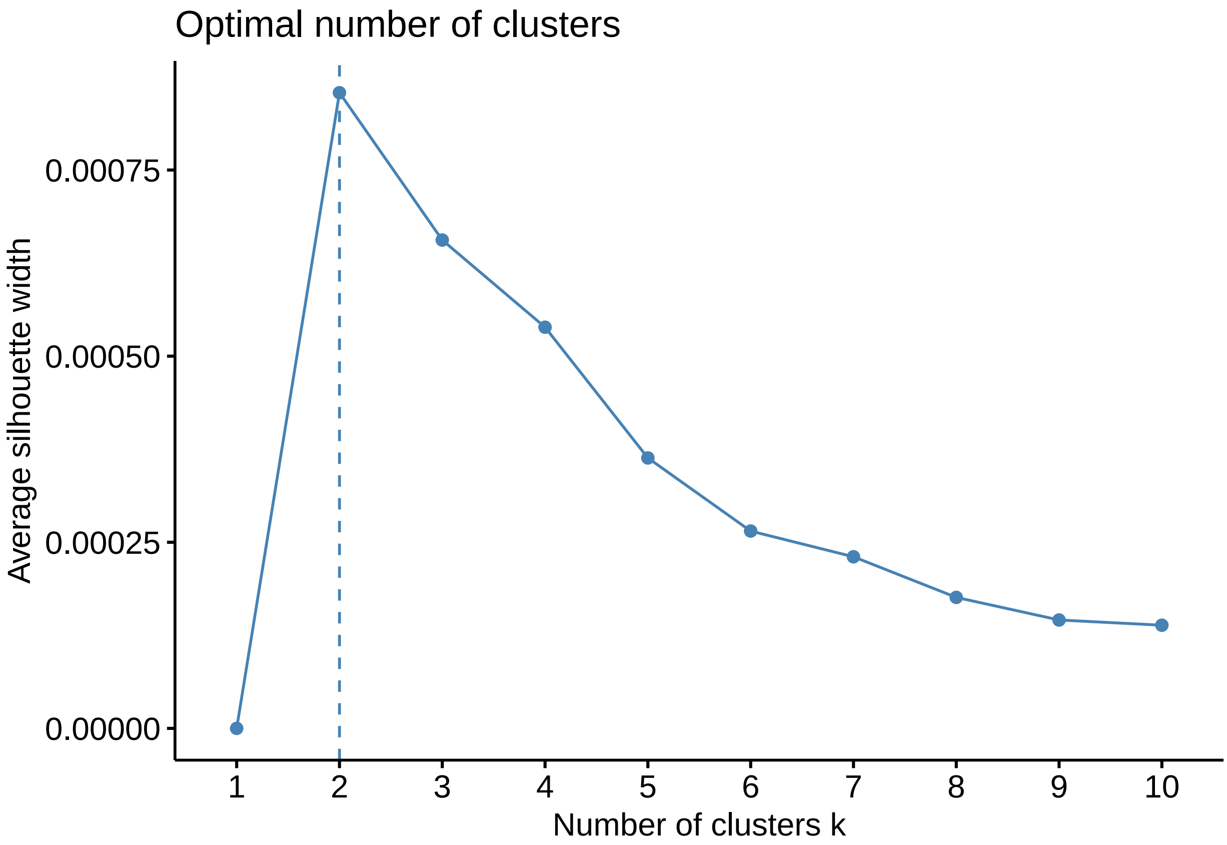


K-Means Clustering

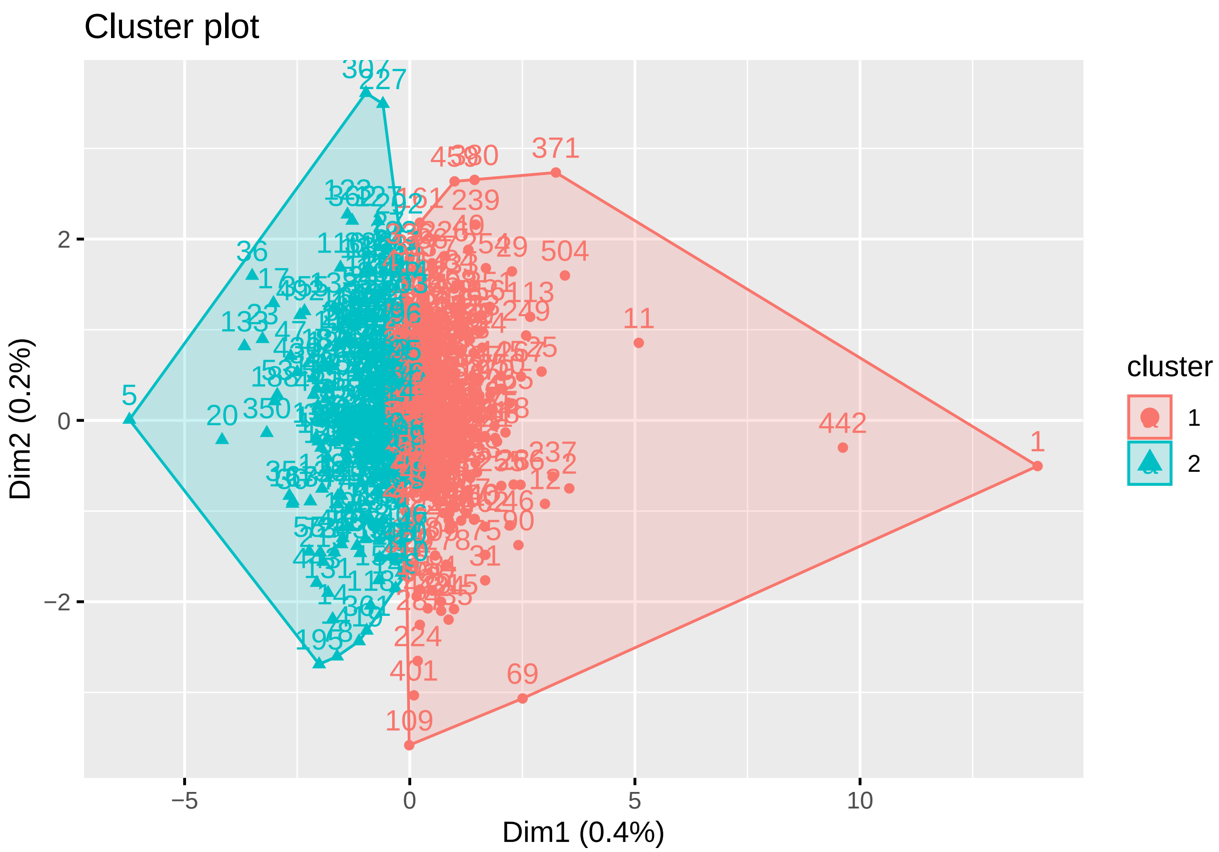
```
preproc <- preProcess(bundesliga_pca_df, method = c("center", "scale"))
bundesliga_normalized <- predict(preproc, bundesliga_pca_df)

numeric_cols <- sapply(bundesliga_normalized, is.numeric)
bundesliga_normalized_numeric <- bundesliga_normalized[, numeric_cols]

fviz_nbclust(bundesliga_normalized_numeric, kmeans, method = "wss")
```



```
fviz_nbclust(bundesliga_normalized_numeric, kmeans, method = "silhouette")
```



```
fit_kmeans <- kmeans(bundesliga_normalized_numeric, centers = 2, nstart = 25)

fviz_cluster(fit_kmeans, data = bundesliga_normalized_numeric)
```



f. Classification

##Creating the Target Variable and Splitting the Data

```
set.seed(123)

bundesliga$age_group <- as.factor(bundesliga$age_group)

index <- createDataPartition(y = bundesliga$age_group, p = 0.7, list = FALSE)
train_data <- bundesliga[index,]
test_data <- bundesliga[-index,]

predictors <- sapply(train_data, is.numeric) | sapply(train_data, is.factor)
train_data <- train_data[, predictors]
test_data <- test_data[, predictors]

names(train_data) <- make.names(names(train_data))
names(test_data) <- make.names(names(test_data))
```

##Training the k-NN Model

```
train_control <- trainControl(method = "cv", number = 10)

knn_model <- train(age_group ~ ., data = train_data, method = "knn", trControl = train_control, tuneLeng
print(knn_model)
```

```
## k-Nearest Neighbors
##
## 358 samples
## 711 predictors
##   3 classes: ’young’, ’mid’, ’old’
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 322, 322, 322, 321, 323, 322, ...
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    5  0.4632518  0.06926053
##    7  0.4800815  0.08423095
##    9  0.5075547  0.12794725
##   11  0.5161218  0.14377039
##   13  0.4962849  0.10413494
##   15  0.4969820  0.10331296
##   17  0.5219863  0.14801505
##   19  0.5161175  0.13499342
##   21  0.4965144  0.09920748
##   23  0.4967439  0.10049915
##   25  0.5080888  0.11999935
##   27  0.4941248  0.09504042
##   29  0.4769863  0.06443056
##   31  0.4743629  0.06014139
##   33  0.4661840  0.04594871
##   35  0.4797726  0.06971625
##   37  0.4936572  0.09467006
##   39  0.4545174  0.02474169
##   41  0.4602402  0.03367573
##   43  0.4632518  0.03873998
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 17.
```

```
tree_model <- train(age_group ~ ., data = train_data, method = "rpart", trControl = train_control)

print(tree_model)
```

```
## CART
##
## 358 samples
## 711 predictors
##   3 classes: ’young’, ’mid’, ’old’
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 322, 323, 322, 321, 322, 323, ...
## Resampling results across tuning parameters:
```

```
##
##   cp         Accuracy   Kappa
##   0.0000000  1.0000000  1.0000000
##   0.2284264  0.9313063  0.8789116
##   0.7715736  0.6613857  0.3789116
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.
```

g. Evaluation

#Confusion Matrix

```
train_data$age_binary <- ifelse(train_data$age_group == "young", "young", "not_young")
test_data$age_binary <- ifelse(test_data$age_group == "young", "young", "not_young")

train_data$age_binary <- factor(train_data$age_binary, levels = c("young", "not_young"))
test_data$age_binary <- factor(test_data$age_binary, levels = c("young", "not_young"))

tree_model_binary <- train(age_binary ~ ., data = train_data, method = "rpart", trControl = train_contro

tree_predictions_binary <- predict(tree_model_binary, newdata = test_data)

tree_predictions_binary <- factor(tree_predictions_binary, levels = c("young", "not_young"))

tree_conf_matrix_binary <- confusionMatrix(tree_predictions_binary, test_data$age_binary)
print(tree_conf_matrix_binary)
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  young not_young
##   young        65         0
##   not_young     0        87
##
##                Accuracy : 1
##                  95% CI : (0.976, 1)
##     No Information Rate : 0.5724
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
##             Sensitivity : 1.0000
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 1.0000
##              Prevalence : 0.4276
##          Detection Rate : 0.4276
##    Detection Prevalence : 0.4276
##       Balanced Accuracy : 1.0000
```

```
##
##        'Positive' Class : young
##
```

##Precision and Recal

```
cm_values <- as.numeric(tree_conf_matrix_binary$table)
true_negative <- cm_values[1]
false_positive <- cm_values[2]
false_negative <- cm_values[3]
true_positive <- cm_values[4]

precision <- true_positive / (true_positive + false_positive)
recall <- true_positive / (true_positive + false_negative)

precision
```

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

```
recall
```

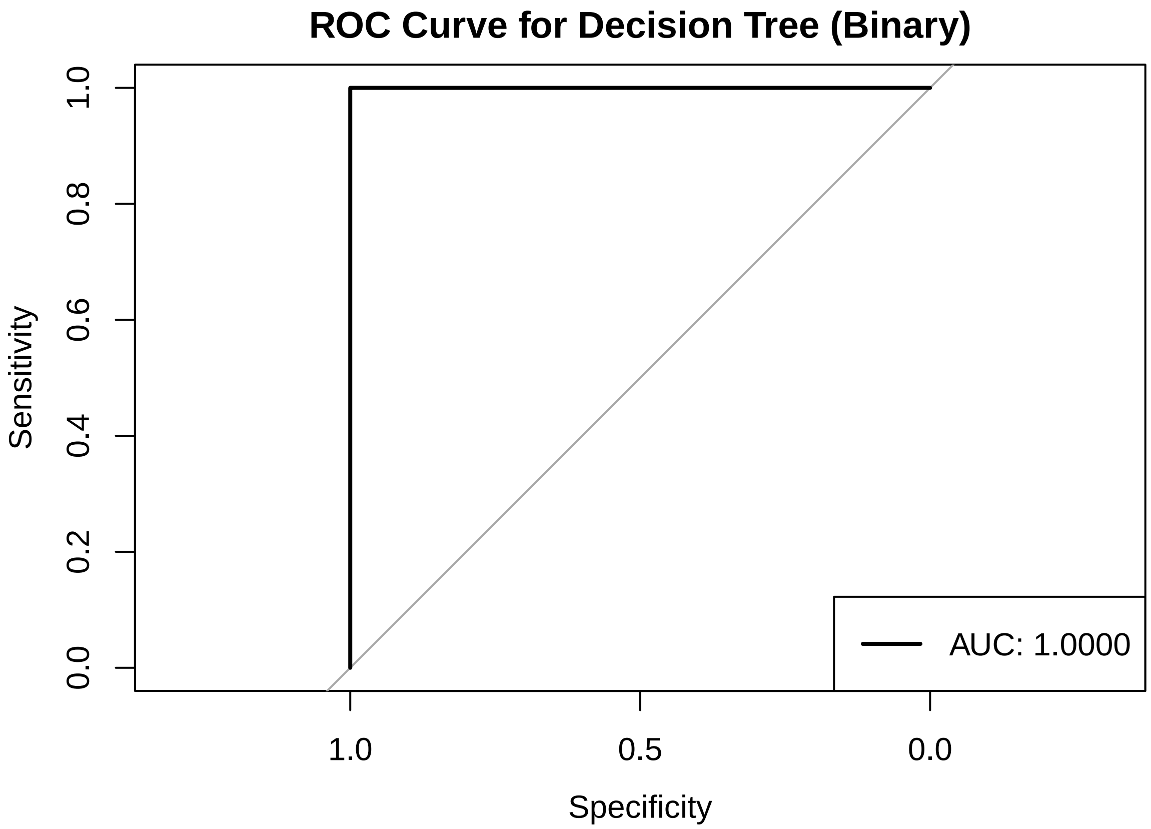
```
## [1] 1
```

##ROC Curve

```
tree_pred_prob_binary <- predict(tree_model_binary, newdata = test_data, type = "prob")
head(tree_pred_prob_binary)
```

```
##    young not_young
## 5      1         0
## 6      1         0
## 7      0         1
## 11     0         1
## 14     0         1
## 15     0         1
```

```
roc_tree_binary <- roc(response = test_data$age_binary, predictor = tree_pred_prob_binary[, "young"])

plot(roc_tree_binary, main = "ROC Curve for Decision Tree (Binary)")
legend("bottomright", legend = c("AUC: 1.0000"), lwd = 2)
```



h. Report

- Refer pdf

i. Reflection

- Refer pdf