R Notebook

1. Data cleansing:

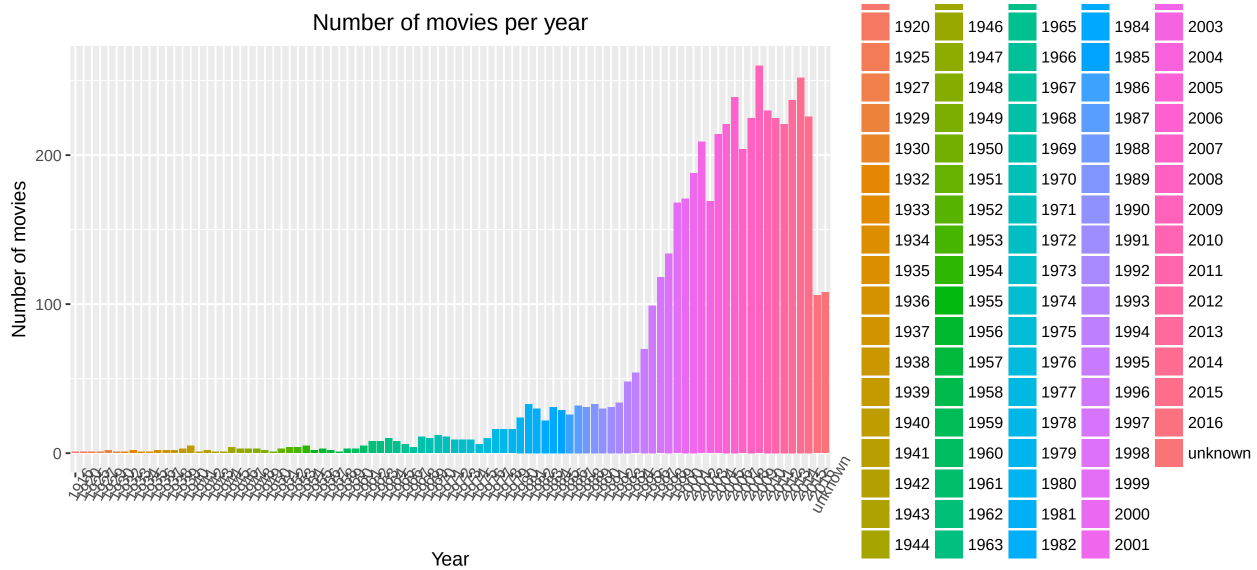
```
movies <- read.csv("IMDB_movie_dataset.csv", encoding = "UTF-8", stringsAsFactors = F)
movies$movie_title <- gsub("\u00A0", "", movies$movie_title)
head(movies)
str(movies)
```

```
##   color     director_name num_critic_for_reviews duration
## 1 Color     James Cameron                    723      178
## 2 Color    Gore Verbinski                    302      169
## 3 Color        Sam Mendes                    602      148
## 4 Color Christopher Nolan                    813      164
## 5             Doug Walker                     NA       NA
## 6 Color    Andrew Stanton                    462      132
##   director_facebook_likes actor_3_facebook_likes     actor_2_name
## 1                       0                    855 Joel David Moore
## 2                     563                   1000    Orlando Bloom
## 3                       0                    161     Rory Kinnear
## 4                   22000                  23000   Christian Bale
## 5                     131                     NA       Rob Walker
## 6                     475                    530  Samantha Morton
##   actor_1_facebook_likes     gross                          genres
## 1                   1000 760505847 Action|Adventure|Fantasy|Sci-Fi
## 2                  40000 309404152        Action|Adventure|Fantasy
## 3                  11000 200074175       Action|Adventure|Thriller
## 4                  27000 448130642                 Action|Thriller
## 5                    131        NA                     Documentary
## 6                    640  73058679         Action|Adventure|Sci-Fi
##      actor_1_name                                            movie_title
## 1     CCH Pounder                                                 Avatar
## 2     Johnny Depp               Pirates of the Caribbean: At World's End
## 3 Christoph Waltz                                                Spectre
## 4       Tom Hardy                                  The Dark Knight Rises
## 5     Doug Walker Star Wars: Episode VII - The Force Awakens
## 6    Daryl Sabara                                            John Carter
##   num_voted_users cast_total_facebook_likes         actor_3_name
## 1          886204                      4834            Wes Studi
## 2          471220                     48350       Jack Davenport
## 3          275868                     11700     Stephanie Sigman
## 4         1144337                    106759 Joseph Gordon-Levitt
## 5               8                       143
## 6          212204                      1873         Polly Walker
##   facenumber_in_poster
## 1                    0
## 2                    0
## 3                    1
## 4                    0
## 5                    0
## 6                    1
##                                                       plot_keywords
## 1                           avatar|future|marine|native|paraplegic
## 2     goddess|marriage ceremony|marriage proposal|pirate|singapore
## 3                              bomb|espionage|sequel|spy|terrorist
## 4 deception|imprisonment|lawlessness|police officer|terrorist plot
## 5
## 6               alien|american civil war|male nipple|mars|princess
##                                         movie_imdb_link
## 1 http://www.imdb.com/title/tt0499549/?ref_=fn_tt_tt_1
## 2 http://www.imdb.com/title/tt0449088/?ref_=fn_tt_tt_1
## 3 http://www.imdb.com/title/tt2379713/?ref_=fn_tt_tt_1
## 4 http://www.imdb.com/title/tt1345836/?ref_=fn_tt_tt_1
## 5 http://www.imdb.com/title/tt5289954/?ref_=fn_tt_tt_1
## 6 http://www.imdb.com/title/tt0401729/?ref_=fn_tt_tt_1
##   num_user_for_reviews language country content_rating    budget
## 1                 3054  English     USA          PG-13 237000000
## 2                 1238  English     USA          PG-13 300000000
## 3                  994  English      UK          PG-13 245000000
## 4                 2701  English     USA          PG-13 250000000
## 5                   NA                                        NA
## 6                  738  English     USA          PG-13 263700000
##   title_year actor_2_facebook_likes imdb_score aspect_ratio
## 1       2009                    936        7.9         1.78
## 2       2007                   5000        7.1         2.35
## 3       2015                    393        6.8         2.35
## 4       2012                  23000        8.5         2.35
## 5         NA                     12        7.1           NA
## 6       2012                    632        6.6         2.35
##   movie_facebook_likes
## 1                33000
## 2                    0
## 3                85000
## 4               164000
## 5                    0
## 6                24000
## 'data.frame':    5043 obs. of  28 variables:
##  $ color                    : chr  "Color" "Color" "Color" "Color" ...
##  $ director_name            : chr  "James Cameron" "Gore Verbinski" "Sam Mendes" "Christopher Nolan"
##  $ num_critic_for_reviews   : int  723 302 602 813 NA 462 392 324 635 375 ...
##  $ duration                 : int  178 169 148 164 NA 132 156 100 141 153 ...
##  $ director_facebook_likes  : int  0 563 0 22000 131 475 0 15 0 282 ...
##  $ actor_3_facebook_likes   : int  855 1000 161 23000 NA 530 4000 284 19000 10000 ...
##  $ actor_2_name             : chr  "Joel David Moore" "Orlando Bloom" "Rory Kinnear" "Christian Bale
##  $ actor_1_facebook_likes   : int  1000 40000 11000 27000 131 640 24000 799 26000 25000 ...
##  $ gross                    : int  760505847 309404152 200074175 448130642 NA 73058679 336530303 2008
##  $ genres                   : chr  "Action|Adventure|Fantasy|Sci-Fi" "Action|Adventure|Fantasy" "Acti
##  $ actor_1_name             : chr  "CCH Pounder" "Johnny Depp" "Christoph Waltz" "Tom Hardy" ...
##  $ movie_title              : chr  "Avatar" "Pirates of the Caribbean: At World's End" "Spectre" "The
##  $ num_voted_users          : int  886204 471220 275868 1144337 8 212204 383056 294810 462669 321795
##  $ cast_total_facebook_likes: int  4834 48350 11700 106759 143 1873 46055 2036 92000 58753 ...
##  $ actor_3_name             : chr  "Wes Studi" "Jack Davenport" "Stephanie Sigman" "Joseph Gordon-Lev
##  $ facenumber_in_poster     : int  0 0 1 0 0 1 0 1 4 3 ...
##  $ plot_keywords            : chr  "avatar|future|marine|native|paraplegic" "goddess|marriage ceremon
##  $ movie_imdb_link          : chr  "http://www.imdb.com/title/tt0499549/?ref_=fn_tt_tt_1" "http://www
##  $ num_user_for_reviews     : int  3054 1238 994 2701 NA 738 1902 387 1117 973 ...
##  $ language                 : chr  "English" "English" "English" "English" ...
##  $ country                  : chr  "USA" "USA" "UK" "USA" ...
##  $ content_rating          : chr  "PG-13" "PG-13" "PG-13" "PG-13" ...
##  $ budget                  : num  2.37e+08 3.00e+08 2.45e+08 2.50e+08 NA ...
##  $ title_year              : int  2009 2007 2015 2012 NA 2012 2007 2010 2015 2009 ...
##  $ actor_2_facebook_likes  : int  936 5000 393 23000 12 632 11000 553 21000 11000 ...
##  $ imdb_score              : num  7.9 7.1 6.8 8.5 7.1 6.6 6.2 7.8 7.5 7.5 ...
##  $ aspect_ratio            : num  1.78 2.35 2.35 2.35 NA 2.35 2.35 1.85 2.35 2.35 ...
##  $ movie_facebook_likes    : int  33000 0 85000 164000 0 24000 0 29000 118000 10000 ...
```

```
movies$title_year[is.na(movies$title_year)] <- "unknown"
movies$title_year <- as.factor(movies$title_year)
```

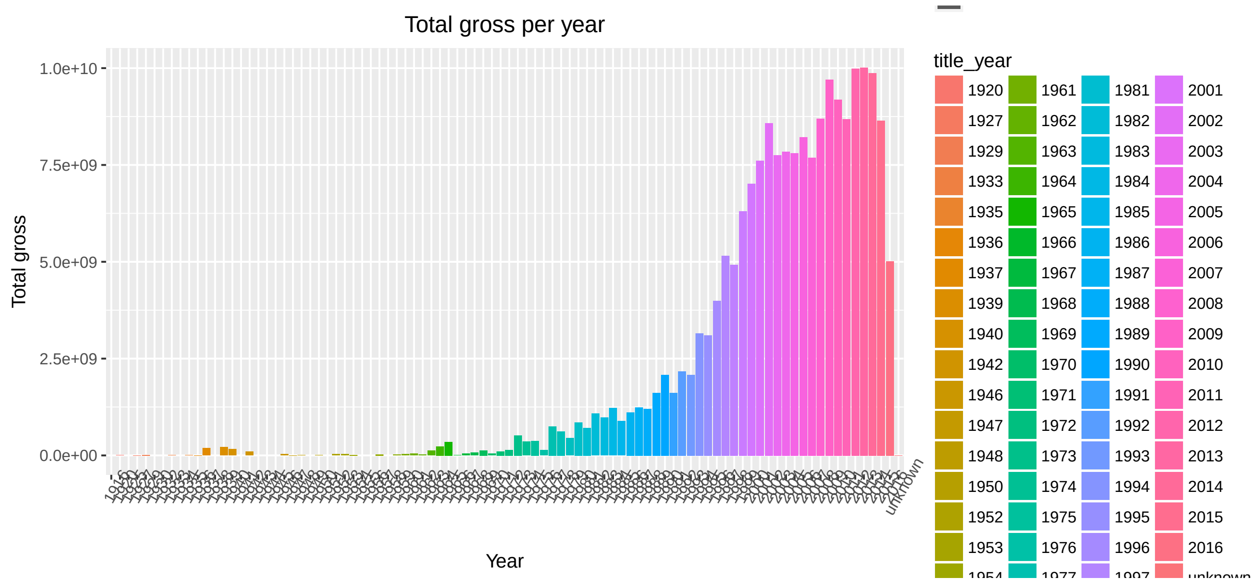
2. Data visualization:

```
ggplot(movies, aes(x = title_year, fill = title_year)) + geom_bar() + theme(axis.text.x = element_text(a
```

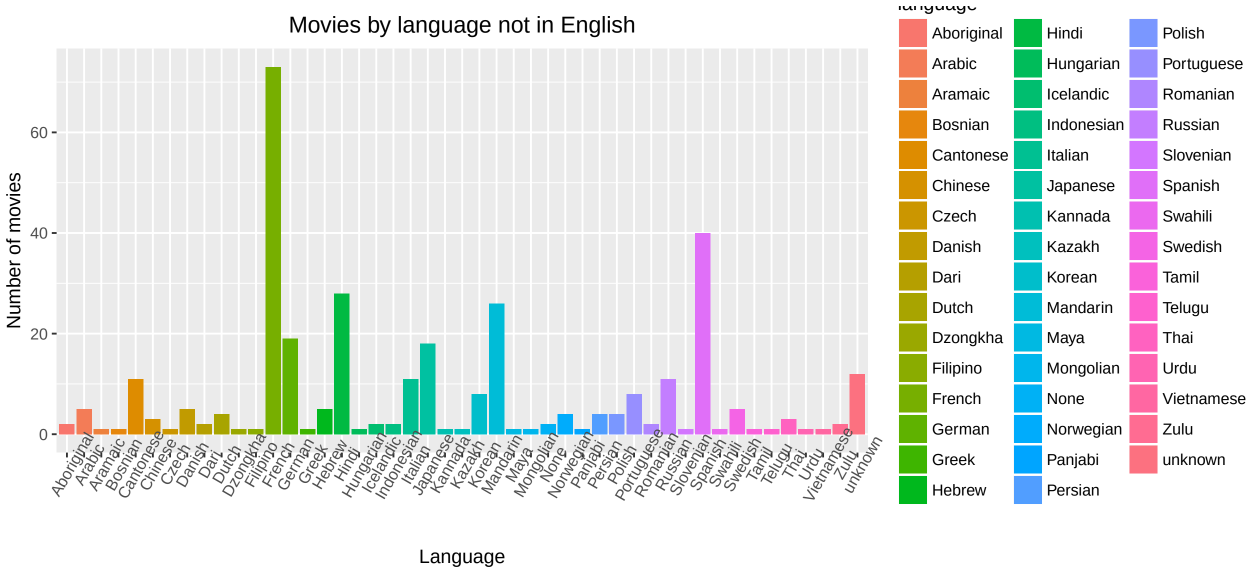


```
ggplot(movies, aes(x = title_year, y = gross, fill = title_year)) + geom_bar(stat = "sum") + theme(axis
```

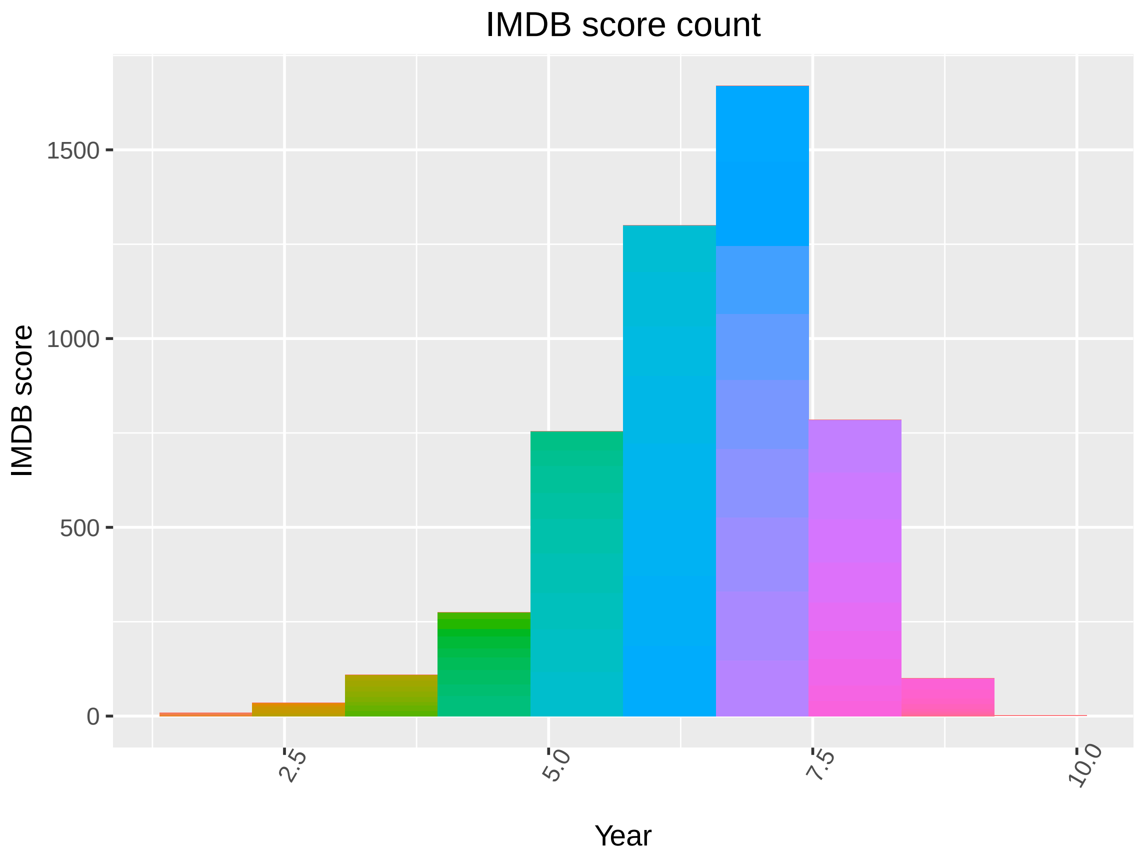
```
## Warning: Removed 884 rows containing non-finite values (stat_sum).
```



```
movies$language[movies$language == ""] <- "unknown"
movies$language <- as.factor(movies$language)
ggplot(movies[which(movies$language != "English"), ], aes(x = language, fill = language)) + geom_bar() +
```

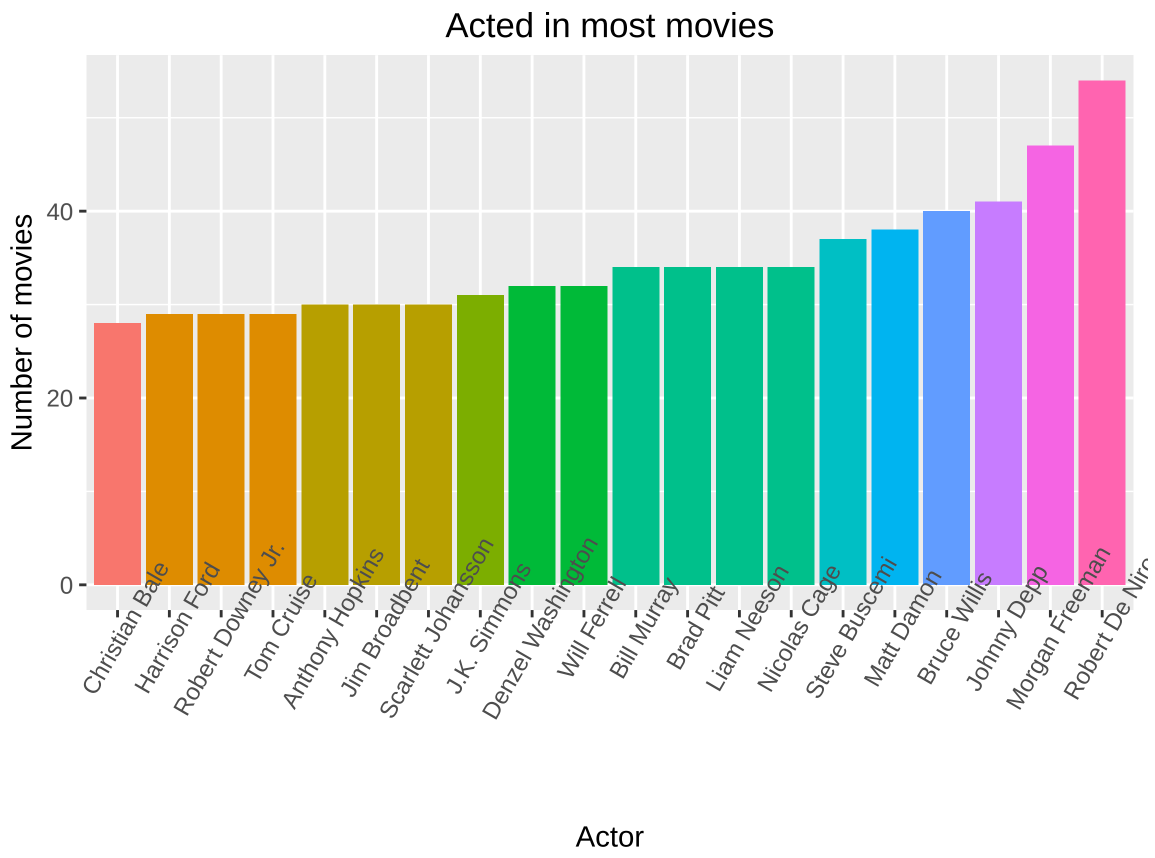


```
ggplot(movies, aes(x = imdb_score, fill = cut(imdb_score, 100))) + geom_histogram(bins = 10, show.legen
```



```
actors <- c(movies$actor_1_name, movies$actor_2_name, movies$actor_3_name)
actors <- actors[nchar(actors) > 1]
actors <- factor(actors)
actors <- reorder(actors,actors,FUN=length)
actors2 <- data.frame(sort(table(actors), decreasing = T)[1:20])


ggplot(actors2, aes(x = reorder(actors2$actors, actors2$Freq), y = Freq, fill = cut(Freq, 100))) + them
```


Acted in most movies
Number of movies
Christian Bale
Harrison Ford
Robert Downey Jr.
Tom Cruise
Anthony Hopkins
Jim Broadbent
Scarlett Johansson
J.K. Simmons
Denzel Washington
Will Ferrell
Bill Murray
Brad Pitt
Liam Neeson
Nicolas Cage
Steve Buscemi
Matt Damon
Bruce Willis
Johnny Depp
Morgan Freeman
Robert De Nirc
Actor