Importing raw data and merging

2023-10-04

Importing all the raw data

Import datasets from Imdb

```
# Load tidyverse
library(tidyverse)
# Define the URL of the first Imdb dataset
url_basics <- "https://datasets.imdbws.com/title.basics.tsv.gz"
# Define the column types
col_types = cols(
  tconst = col_character(),
  titleType = col_character(),
  primaryTitle = col_character(),
  originalTitle = col_character(),
  startYear = col_double(),
  endYear = col_double(),
  runtimeMinutes = col_double(),
  genres = col_character(),
  isAdult = col_logical()
)
# Download and import the dataset
title_basics_data <- read_delim(url_basics, delim = "\t", col_names = TRUE, col_types = col_types)
# Look at the first few rows of the dataset
View(title_basics_data)

# Define the URL of the second Imdb dataset
url_crew <- "https://datasets.imdbws.com/title.crew.tsv.gz"
# Define the column types and download and import the dataset
imdb_crew_data <- read_delim(url_crew, delim = "\t", col_names = TRUE,
                             col_types = cols(
                               tconst = col_character(),
                               directors = col_character(),
                               writers = col_character()
                             )
)
# Look at the first few rows of the dataset
head(imdb_crew_data)
```

```
## # A tibble: 6 x 3
##   tconst    directors writers
##   <chr>     <chr>     <chr>
## 1 tt0000001 nm0005690 "\\N"
## 2 tt0000002 nm0721526 "\\N"
## 3 tt0000003 nm0721526 "\\N"
## 4 tt0000004 nm0721526 "\\N"
```

```
## 5 tt0000005 nm0005690 "\\N"
## 6 tt0000006 nm0005690 "\\N"
```

```
# Define the URL of the third Imdb dataset
url_name_basics <- "https://datasets.imdbws.com/name.basics.tsv.gz"
# Define the column types and download and import the dataset
col_types <- cols(
  nconst = col_character(),
  primaryName = col_character(),
  birthYear = col_integer(),
  deathYear = col_character(),
  primaryProfession = col_character(),
  knownForTitles = col_character()
)
# Download and import the dataset
imdb_name_basics <- read_delim(url_name_basics, delim = "\t", col_types = col_types)
# Look at the first few rows of the dataset
head(imdb_name_basics)
```

```
## # A tibble: 6 x 6
##   nconst    primaryName     birthYear deathYear primaryProfession knownForTitles
##   <chr>     <chr>               <int> <chr>     <chr>              <chr>
## 1 nm0000001 Fred Astaire         1899 "1987"    soundtrack,actor~ tt0053137,tt0~
## 2 nm0000002 Lauren Bacall        1924 "2014"    actress,soundtra~ tt0038355,tt0~
## 3 nm0000003 Brigitte Bardot      1934 "\\N"     actress,soundtra~ tt0049189,tt0~
## 4 nm0000004 John Belushi         1949 "1982"    actor,soundtrack~ tt0077975,tt0~
## 5 nm0000005 Ingmar Bergman       1918 "2007"    writer,director,~ tt0050986,tt0~
## 6 nm0000006 Ingrid Bergman       1915 "1982"    actress,soundtra~ tt0036855,tt0~
```

Merging Imbd datsets

```
library(dplyr)
# Merge imdb_basics_data and imdb_crew_data based on the variable 'tconst'
imdb_basics_crew <- left_join(title_basics_data, imdb_crew_data, by="tconst")

# Rename the variable 'nconsts' to 'directors' since it's an array of nconsts
imdb_basics_crew <- imdb_basics_crew %>%
  rename(director_identifier = directors)

print(imdb_basics_crew)
```

```
## # A tibble: 10,225,586 x 11
##    tconst    titleType primaryTitle      originalTitle isAdult startYear endYear
##    <chr>     <chr>     <chr>             <chr>         <lgl>       <dbl>   <dbl>
##  1 tt0000001 short     Carmencita        Carmencita    FALSE        1894      NA
##  2 tt0000002 short     Le clown et ses ~ Le clown et ~ FALSE        1892      NA
##  3 tt0000003 short     Pauvre Pierrot    Pauvre Pierr~ FALSE        1892      NA
##  4 tt0000004 short     Un bon bock       Un bon bock   FALSE        1892      NA
##  5 tt0000005 short     Blacksmith Scene  Blacksmith S~ FALSE        1893      NA
##  6 tt0000006 short     Chinese Opium Den Chinese Opiu~ FALSE        1894      NA
##  7 tt0000007 short     Corbett and Cour~ Corbett and ~ FALSE        1894      NA
##  8 tt0000008 short     Edison Kinetosco~ Edison Kinet~ FALSE        1894      NA
##  9 tt0000009 movie     Miss Jerry        Miss Jerry    FALSE        1894      NA
## 10 tt0000010 short     Leaving the Fact~ La sortie de~ FALSE        1895      NA
```

```
## # i 10,225,576 more rows
## # i 4 more variables: runtimeMinutes <dbl>, genres <chr>,
## #   director_identifier <chr>, writers <chr>
```

```
# Merge the dataframe that was created above with the imdb_name_basics data
imdb_basics_crew_name <- left_join(imdb_basics_crew, imdb_name_basics, by = c ("director_identifier" =

# Delete all variables that we will not use in the project
imdb_basics_crew_name <- imdb_basics_crew_name %>%
  select(-endYear, -writers, -birthYear, -deathYear)

# See all unique values of the titleType
unique_values <- unique(imdb_basics_crew_name$titleType)

print(unique_values)
```

```
##  [1] "short"        "movie"        "tvShort"      "tvMovie"      "tvSeries"
##  [6] "tvEpisode"    "tvMiniSeries" "tvSpecial"    "video"        "videoGame"
## [11] "tvPilot"
```

```
# Select only the data that contains titleType movie
filtered_data <- imdb_basics_crew_name %>% filter(titleType == "movie")
```

Merging Imbd datasets with Boxofficemojo dataset

We will merge the Imdb datasets with the boxofficemojo dataset for data regarding the opening gross and theaters the movie was played in.

```
# Import Boxofficemojo data
openingweekend_unique <- read.csv("../../gen/data-preparation/weekend_data.csv")

# Merge the datasets by using a left join
merged_dataset <- openingweekend_unique %>% left_join(filtered_data, by = c("name" = "primaryTitle"))

# Find duplicated ride_ids
merged_dataset %>%
  count(name) %>%
  filter(n > 1)
```

```
##                                     name  n
## 1                                     21  4
## 2                          A Christmas Carol 21
## 3                 A Nightmare on Elm Street  3
## 4                             A Quiet Place  2
## 5                            A Star Is Born  8
## 6                           About Last Night  2
## 7                                   Aladdin  6
## 8                        Alice in Wonderland 12
## 9                          American Gangster  2
## 10                         American Reunion  2
## 11                                Apollo 13  2
## 12                               Armageddon  6
## 13                                  Arrival  4
## 14                                   Avatar  4
## 15                                   Barbie  3
## 16                                   Batman  2
```

```
## 17                    Beauty and the Beast 14
## 18                         Bedtime Stories  3
## 19                                 Beowulf  3
## 20                               Big Daddy  3
## 21                             Birds of Prey  6
## 22                             Black Panther  2
## 23                               Black Widow 11
## 24                                     Bolt  2
## 25                                   Brave  4
## 26                               Bride Wars  2
## 27                              Bridesmaids  2
## 28                             Bullet Train  2
## 29                                 Candyman  4
## 30                             Casino Royale  2
## 31                                 Cast Away  3
## 32                      Catch Me If You Can  4
## 33                         Charlie's Angels  2
## 34                    Cheaper by the Dozen  3
## 35                                 Chronicle  2
## 36                               Cinderella 23
## 37                      Clash of the Titans  2
## 38                                   Click  7
## 39                                    Coco  4
## 40                                   Congo  2
## 41                                Contagion  3
## 42                               Contraband  4
## 43                                   Creed  4
## 44                              Daddy's Home  2
## 45                                 Daredevil  2
## 46                               Date Night  2
## 47                          Dawn of the Dead  2
## 48                                 Dear John  3
## 49                     Diary of a Wimpy Kid  2
## 50    Diary of a Wimpy Kid: Rodrick Rules  2
## 51                               Dick Tracy  3
## 52                                 Dinosaur  5
## 53                               Divergent  2
## 54                           Doctor Dolittle  2
## 55                             Don't Breathe  2
## 56                           Double Jeopardy  2
## 57                                    Dumbo  2
## 58                                     Dune  2
## 59                                  Dunkirk  2
## 60                          Edge of Tomorrow  2
## 61                                 Elemental  3
## 62                                     Elvis  3
## 63                                  Elysium  4
## 64                                Enchanted  3
## 65                                     Epic  3
## 66                                   Eraser  2
## 67                                 Face/Off  2
## 68                           Fantastic Four  6
## 69                                   Flight  7
## 70                              Fool's Gold  8
```

```
## 71                          Four Brothers  2
## 72                          Freaky Friday  2
## 73                        Friday the 13th  4
## 74                           Funny People  3
## 75                                   Fury 14
## 76                               Garfield  2
## 77                               Get Hard  3
## 78                           Ghostbusters  2
## 79                              Gladiator  4
## 80                                  Glass 12
## 81                               Godzilla 10
## 82                     Gone in 60 Seconds  2
## 83                               Good Boys  3
## 84                                 Gravity  5
## 85                               Grown Ups  2
## 86                               Hairspray  2
## 87                               Halloween 12
## 88                                Hannibal  4
## 89                          Happy Death Day  2
## 90                          Haunted Mansion  3
## 91                                  Hellboy  2
## 92                                 Hercules  8
## 93                            Hide and Seek 23
## 94                                   Hitch  2
## 95                                    Home 59
## 96                                     Hop  2
## 97                       Hotel Transylvania  2
## 98                     Hotel Transylvania 2  2
## 99          How the Grinch Stole Christmas  2
## 100               How to Train Your Dragon  2
## 101                               Hustlers  3
## 102                                 Ice Age  4
## 103                               Immortals  2
## 104                               Inception  2
## 105                         Independence Day  4
## 106                               Inside Man  2
## 107                               Inside Out 18
## 108                         Inspector Gadget  2
## 109                           Into the Woods  5
## 110                                 Iron Man  3
## 111                                       It  3
## 112                         It's Complicated  9
## 113                            Jack and Jill  4
## 114                                   Joker 12
## 115                                  Jumper  3
## 116                            Jurassic Park  2
## 117                                King Kong  6
## 118                               Knives Out  2
## 119                          Les Misérables 14
## 120                               Lights Out 14
## 121                           Lilo & Stitch  2
## 122                              Little Man  4
## 123                                   Logan  2
## 124                                    Lucy  4
```

```
## 125                          Madagascar  3
## 126                               Mama 17
## 127                        Man of Steel  3
## 128                         Man on Fire  3
## 129                      Meet the Parents  2
## 130                       Mission to Mars  2
## 131                                Moana  4
## 132                          Mortal Kombat  4
## 133                      Mr. & Mrs. Smith  2
## 134                                Mulan  3
## 135          Murder on the Orient Express  2
## 136              My Best Friend's Wedding  4
## 137                   My Bloody Valentine  2
## 138                     National Treasure  2
## 139                              Neighbors 11
## 140                          Night School  3
## 141                          No Good Deed  5
## 142                        No Time to Die  2
## 143                                  Noah  5
## 144                               Non-Stop  3
## 145                        Now You See Me  2
## 146                               Oblivion  9
## 147                               Obsessed  9
## 148                        Ocean's Eleven  2
## 149                           Open Season  5
## 150                          Panic Room  2
## 151                              Payback 15
## 152                          Pet Sematary  2
## 153                         Pete's Dragon  2
## 154                          Peter Rabbit  2
## 155                                Pixels  2
## 156                     Planet of the Apes  3
## 157                           Poltergeist  4
## 158                         Power Rangers  2
## 159                             Predators  3
## 160                          Project X  7
## 161                          Prometheus  3
## 162                        Public Enemies  2
## 163                         Puss in Boots  5
## 164                               Rampage  9
## 165                                 Rango  2
## 166                                Ransom 11
## 167                          Red Dragon  2
## 168                          Ride Along  4
## 169                                 Rio  5
## 170                          Robin Hood 12
## 171          Robin Hood: Prince of Thieves  2
## 172                             RoboCop  2
## 173                              Robots  2
## 174                           Rocketman  4
## 175                           Rush Hour  3
## 176                            S.W.A.T.  2
## 177                          Safe Haven  8
## 178                          Safe House  6
```

```
## 179                              Salt  8
## 180                       Scary Movie  2
## 181                            Scream  9
## 182                             Shaft  3
## 183                     Sherlock Holmes  5
## 184                            Sin City  3
## 185                              Sing  2
## 186                         Skyscraper  4
## 187                      Sleepy Hollow  3
## 188                             Smile 21
## 189                       Son of God  2
## 190                          Spectre  2
## 191                             Split 14
## 192                               Spy  5
## 193                  Starship Troopers  2
## 194                             Storks  2
## 195                      Suicide Squad  2
## 196                             Taken  5
## 197                           Tangled  5
## 198                            Tarzan  2
## 199                               Ted  2
## 200         Teenage Mutant Ninja Turtles  4
## 201                     The Accountant  4
## 202                  The Addams Family  4
## 203              The Amityville Horror  2
## 204                      The Avengers  4
## 205                      The Bad Guys  4
## 206                          The Boss 12
## 207                      The Break-Up  2
## 208                        The Butler  4
## 209               The Call of the Wild  3
## 210                      The Campaign  5
## 211                 The Cat in the Hat  2
## 212       The Day the Earth Stood Still  2
## 213                  The Devil Inside  3
## 214                   The Expendables  4
## 215            The Fast and the Furious  2
## 216                         The Firm  2
## 217                        The Flash  2
## 218               The Forbidden Kingdom  2
## 219                      The Fugitive 13
## 220              The General's Daughter  3
## 221               The Girl on the Train  6
## 222                    The Great Gatsby  5
## 223                    The Green Hornet  3
## 224                          The Grudge  6
## 225                      The Hangover  2
## 226                     The Happening  2
## 227                      The Haunting  6
## 228                          The Heat  3
## 229                    The Interpreter  4
## 230                  The Invisible Man  9
## 231                    The Jungle Book  6
## 232                     The Karate Kid  2
```

```
## 233                       The Last Samurai  3
## 234 The League of Extraordinary Gentlemen  3
## 235                          The Lion King  3
## 236                     The Little Mermaid 11
## 237                        The Lone Ranger  3
## 238                       The Longest Yard  2
## 239                          The Lost City  8
## 240                          The Lucky One  2
## 241                  The Magnificent Seven  3
## 242                               The Mask  7
## 243                      The Mask of Zorro  2
## 244                               The Mummy 16
## 245                                The Nun  6
## 246                     The Nutty Professor  3
## 247                         The Other Woman 14
## 248                             The Patriot  7
## 249                        The Perfect Guy  2
## 250                           The Proposal  5
## 251                              The Purge  2
## 252                     The Purge: Anarchy  2
## 253                               The Rock  4
## 254                        The Sixth Sense  3
## 255                     The Stepford Wives  2
## 256                      The Time Machine  4
## 257                               The Town  4
## 258                       The Ugly Truth  2
## 259                          The Village  7
## 260                            The Visit 20
## 261                              The Vow  5
## 262                           The Wolfman  2
## 263                         The Wolverine  2
## 264                                 Thor  3
## 265                               Titanic  8
## 266                          Tomorrowland  2
## 267                          Total Recall  6
## 268                             True Grit  2
## 269                             True Lies  2
## 270                                 Turbo  3
## 271                              Twilight 12
## 272                               Twister  2
## 273                          Unbreakable  7
## 274                             Unbroken  6
## 275                            Uncharted  4
## 276                           Underworld  9
## 277                              Unknown 10
## 278                          Unstoppable 15
## 279                                   Us  5
## 280                       Valentine's Day  6
## 281                           Van Helsing  4
## 282                                Venom  7
## 283                               Wanted 16
## 284                      War of the Worlds  6
## 285                      What Lies Beneath  3
## 286                        What Women Want  4
```

```
## 287                 When a Stranger Calls  2
## 288                           White Noise  6
## 289                                Wonder  7
## 290                          Wonder Woman  5
```

```
# Remove full and partial duplicates
merged_dataset_unique <- merged_dataset %>%
# Only based on ride_id instead of all cols
  distinct(name, .keep_all = TRUE)

# Find duplicated ride_ids in bike_share_rides_unique
merged_dataset_unique %>%
# Count the number of occurrences of each ride_id
count(name) %>%
# Filter for rows with a count > 1
filter(n > 1)
```

```
## [1] name n
## <0 rows> (or 0-length row.names)
```

Converting merged_dataset_unique into a csv file

We will convert the dataset into a csv file to avoid that importing takes very long each time, as it is a very big dataset.

```
csv_bestand <- "merged_dataset_unique.csv"
folder <- "../../gen/data-preparation"

folder_path <- file.path(folder, csv_bestand)

write_csv2(merged_dataset_unique, file = folder_path)

cat("The dataframe has been saved as", folder_path, "\n")
```

```
## The dataframe has been saved as ../../gen/data-preparation/merged_dataset_unique.csv
```