```r
# Loading the csv file 'movies_metadata.csv' from the
# folder in the drive, and storing it into the variable
# 'stage_1'
library("tidyverse")
library("readr")
library("stringr")
library("dplyr")
library("priceR")
library("tibble")
library("gridExtra")
library("lemon")

setwd(getwd())
stage_1 <- as_tibble(read_csv("Movies_DataSet/movies_metadata.csv",
    show_col_types = FALSE))

png("1_raw_data_preview.png", height = 50 * nrow(head(stage_1)),
    width = 200 * ncol(head(stage_1)))
grid.table(head(stage_1))

png("2_raw_data_summary.png", height = 50 * nrow(summary(stage_1)),
    width = 150 * ncol(summary(stage_1)))
grid.table(summary(stage_1))

head(stage_1)
```

```
## # A tibble: 6 x 24
##   adult belongs_to_colle~  budget genres homepage    id imdb_id original_langua~
##   <lgl> <chr>               <dbl> <chr>  <chr>    <dbl> <chr>   <chr>
## 1 FALSE {’id’: 10194, ’n~   3   e7 [{’id~ http://~   862 tt0114~ en
## 2 FALSE <NA>                6.5e7 [{’id~ <NA>      8844 tt0113~ en
## 3 FALSE {’id’: 119050, ’~   0     [{’id~ <NA>     15602 tt0113~ en
## 4 FALSE <NA>                1.6e7 [{’id~ <NA>     31357 tt0114~ en
## 5 FALSE {’id’: 96871, ’n~   0     [{’id~ <NA>     11862 tt0113~ en
## 6 FALSE <NA>                6   e7 [{’id~ <NA>       949 tt0113~ en
## # ... with 16 more variables: original_title <chr>, overview <chr>,
## #   popularity <dbl>, poster_path <chr>, production_companies <chr>,
## #   production_countries <chr>, release_date <date>, revenue <dbl>,
## #   runtime <dbl>, spoken_languages <chr>, status <chr>, tagline <chr>,
## #   title <chr>, video <lgl>, vote_average <dbl>, vote_count <dbl>
```

```r
# Dividing the list of columns into the ones that are to be
# kept and the ones to be reserved
to_keep_columns <- c("adult", "genres", "imdb_id", "popularity",
    "runtime", "vote_count", "production_countries", "original_language",
    "title")
drop_columns <- c("belongs_to_collection", "homepage", "id",
    "budget", "poster_path", "video", "tagline", "production_companies",
    "overview", "release_date", "revenue", "status", "original_title",
    "vote_average")
```

```r
### All the functions

# To check na values column wise
fun <- function(x) {
    tmp <- is.na(x)
    apply(tmp, 2, sum)
}

## Function to convert columns containing Dictionaries to
## List:
getAttribute <- function(vector) {
    vector <- as.vector(str_split(vector, regex("[\\[{'':,}\\]]"))[[1]])
    vector <- vector[!vector == "" & !vector == " "]
    vector <- as.vector(vector[which(vector == "name") + 1])
    return(toString(vector))
}

## Converts Currecy as per today's curr value:
convert_currency <- function(datum) {
    # retrives a list of currencies seen in datum
    curr_type = unique(str_sub(datum, 1, 4))

    for (curr in curr_type) {
        # Fetches the currency Valye using priceR package
        exch_rate = exchange_rate_latest(curr)
        conversion_value = as.double(exch_rate[exch_rate[1] ==
            "USD"])[2]
        # Retrieved values in data with curr currency
        sub_datum = datum[str_sub(datum, 1, 4) == curr]
        for (data in sub_datum) {
            ind = which(datum == data)
            value = as.double(str_sub(data, 5))
            res = as.integer(value * conversion_value)
            datum[ind] = res
        }
    }
    return(datum)
}

split_cols <- function(x, colname, df) {

    ncols <- NULL
    colm <- NULL
    ncols <- max(stringr::str_count(x, ", ")) + 1
    colm <- paste(colname, 1:ncols, sep = "_")

    df <- tidyr::separate(data = df, col = colname, sep = ", ",
        into = colm, remove = FALSE)
    unique_val_list <- data.frame(matrix(ncol = 1, nrow = 0))
    colnames(unique_val_list) <- colm[1]
    for (i in colm) {
        colnames(unique_val_list) <- i
        tmp <- as.data.frame(unique(df[, i]))
        colnames(tmp) <- i
        unique_val_list <- rbind(as.data.frame(unique_val_list),
            tmp)
    }

    unique_val_list <- as.data.frame(unique(unique_val_list))
    unique_val_list <- as.data.frame(na.omit(unique_val_list))

    for (i in 1:length(unique_val_list[, 1])) {
        df[unique_val_list[i, 1]] <- 0
    }

    for (i in 1:nrow(df)) {
        for (j in colm) {
            if (!is.na(df[i, j])) {
                k <- as.character(df[i, j])
                df[i, k] = 1
            }
        }
    }
    df <- select(df, -colm)
    # filename <- paste(filename, '.csv')
    # write.csv(unique_val_list,filename, row.names =
    # FALSE)
    return(df)
}
```

```
## Keeping necessary Columns only.
stage_1 <- stage_1[to_keep_columns]

png("3_stage_1_columns_filtered.png", height = 50 * nrow(head(stage_1)),
    width = 200 * ncol(head(stage_1)))
grid.table(head(stage_1))

png("4_stage_1_columns_filtered_summary.png", height = 50 * nrow(summary(stage_1)),
    width = 150 * ncol(summary(stage_1)))
grid.table(summary(stage_1))

head(stage_1)
```

```
## # A tibble: 6 x 9
##   adult genres        imdb_id  popularity runtime vote_count production_countri~
##   <lgl> <chr>         <chr>         <dbl>   <dbl>      <dbl> <chr>
## 1 FALSE [{’id’: 16, ~ tt01147~      21.9       81       5415 [{’iso_3166_1’: ’U~
## 2 FALSE [{’id’: 12, ~ tt01134~      17.0      104       2413 [{’iso_3166_1’: ’U~
## 3 FALSE [{’id’: 1074~ tt01132~      11.7      101         92 [{’iso_3166_1’: ’U~
## 4 FALSE [{’id’: 35, ~ tt01148~       3.86     127         34 [{’iso_3166_1’: ’U~
## 5 FALSE [{’id’: 35, ~ tt01130~       8.39     106        173 [{’iso_3166_1’: ’U~
## 6 FALSE [{’id’: 28, ~ tt01132~      17.9      170       1886 [{’iso_3166_1’: ’U~
## # ... with 2 more variables: original_language <chr>, title <chr>
```

```r
## Converting all Dictionary kinda Cols into Lists
stage_1$genres <- sapply(stage_1$genres, getAttribute, USE.NAMES = FALSE,
    simplify = "array")  # Genres Column
stage_1$production_countries <- sapply(stage_1$production_countries,
    getAttribute, USE.NAMES = FALSE, simplify = "array")

png("5_stage_1_post_dict_str_conversion.png", height = 50 * nrow(head(stage_1)),
    width = 200 * ncol(head(stage_1)))
grid.table(head(stage_1))

png("6_stage_1_post_dict_str_conversion_summary.png", height = 50 *
    nrow(summary(stage_1)), width = 150 * ncol(summary(stage_1)))
grid.table(summary(stage_1))

head(stage_1)
```

```
## # A tibble: 6 x 9
##   adult genres         imdb_id  popularity runtime vote_count production_countr~
##   <lgl> <chr>          <chr>         <dbl>   <dbl>      <dbl> <chr>
## 1 FALSE Animation, Co~ tt01147~      21.9       81       5415 United States of ~
## 2 FALSE Adventure, Fa~ tt01134~      17.0      104       2413 United States of ~
## 3 FALSE Romance, Come~ tt01132~      11.7      101         92 United States of ~
## 4 FALSE Comedy, Drama~ tt01148~       3.86     127         34 United States of ~
## 5 FALSE Comedy         tt01130~       8.39     106        173 United States of ~
## 6 FALSE Action, Crime~ tt01132~      17.9      170       1886 United States of ~
## # ... with 2 more variables: original_language <chr>, title <chr>
```

```r
# Replacing blank values with NA and then omitting the NAs.
stage_1 <- stage_1 %>%
    mutate(genres = ifelse(genres == "", NA, genres)) %>%
    mutate(production_countries = ifelse(production_countries ==
        "", NA, production_countries))

png("7_stage_1_post_blank_val_removal.png", height = 50 * nrow(head(stage_1)),
    width = 200 * ncol(head(stage_1)))
grid.table(head(stage_1))

png("8_stage_1_post_blank_val_removal_summary.png", height = 50 *
    nrow(summary(stage_1)), width = 150 * ncol(summary(stage_1)))
grid.table(summary(stage_1))

head(stage_1)
```

```
## # A tibble: 6 x 9
##   adult genres         imdb_id  popularity runtime vote_count production_countr~
##   <lgl> <chr>          <chr>         <dbl>   <dbl>      <dbl> <chr>
## 1 FALSE Animation, Co~ tt01147~      21.9       81       5415 United States of ~
## 2 FALSE Adventure, Fa~ tt01134~      17.0      104       2413 United States of ~
## 3 FALSE Romance, Come~ tt01132~      11.7      101         92 United States of ~
## 4 FALSE Comedy, Drama~ tt01148~       3.86     127         34 United States of ~
## 5 FALSE Comedy         tt01130~       8.39     106        173 United States of ~
## 6 FALSE Action, Crime~ tt01132~      17.9      170       1886 United States of ~
## # ... with 2 more variables: original_language <chr>, title <chr>
```

```r
# Joining the files movies metadata and IMDB movies.
IMDB_movies <- as_tibble(read_csv("IMDb movies.csv", show_col_types = FALSE))
IMDB_rating <- as_tibble(read_csv("IMDb ratings.csv", show_col_types = FALSE))
stage_1 <- dplyr::inner_join(stage_1, select(IMDB_movies, year,
    imdb_title_id, director, budget, worlwide_gross_income),
    by = c(imdb_id = "imdb_title_id"))
stage_1 <- dplyr::inner_join(stage_1, select(IMDB_rating, imdb_title_id,
    weighted_average_vote), by = c(imdb_id = "imdb_title_id"))
stage_1 <- na.omit(stage_1)

png("9_stage_1_post_merge.png", height = 50 * nrow(head(stage_1)),
    width = 200 * ncol(head(stage_1)))
grid.table(head(stage_1))

png("10_stage_1_post_merge.png", height = 50 * nrow(summary(stage_1)),
    width = 150 * ncol(summary(stage_1)))
grid.table(summary(stage_1))

head(stage_1)
```

```
## # A tibble: 6 x 14
##   adult genres         imdb_id  popularity runtime vote_count production_countr~
##   <lgl> <chr>          <chr>         <dbl>   <dbl>      <dbl> <chr>
## 1 FALSE Animation, Co~ tt01147~      21.9       81       5415 United States of ~
## 2 FALSE Adventure, Fa~ tt01134~      17.0      104       2413 United States of ~
## 3 FALSE Romance, Come~ tt01132~      11.7      101         92 United States of ~
## 4 FALSE Comedy, Drama~ tt01148~       3.86     127         34 United States of ~
## 5 FALSE Comedy         tt01130~       8.39     106        173 United States of ~
## 6 FALSE Action, Crime~ tt01132~      17.9      170       1886 United States of ~
## # ... with 7 more variables: original_language <chr>, title <chr>, year <dbl>,
## #   director <chr>, budget <chr>, worlwide_gross_income <chr>,
## #   weighted_average_vote <dbl>
```

```r
# Currency Conversion and Dollar Removal
stage_1$budget[!str_detect(stage_1$budget, "^\\$")] = convert_currency(stage_1$budget[!str_detect(stage_
    "^\\$")])  # Currency Conversion
```

```
## Daily GBP  exchange rate as at end of day 2021-12-14 GMT
## Daily EUR  exchange rate as at end of day 2021-12-13 GMT
## Daily CAD  exchange rate as at end of day 2021-12-13 GMT
## Daily FRF  exchange rate as at end of day 2021-12-14 GMT
## Daily DEM  exchange rate as at end of day 2021-12-14 GMT
## Daily AUD  exchange rate as at end of day 2021-12-14 GMT
## Daily DKK  exchange rate as at end of day 2021-12-14 GMT
## Daily JPY  exchange rate as at end of day 2021-12-14 GMT
## Daily HKD  exchange rate as at end of day 2021-12-14 GMT
## Daily RUR  exchange rate as at end of day 2021-12-14 GMT
## Daily ITL  exchange rate as at end of day 2021-12-14 GMT
## Daily ESP  exchange rate as at end of day 2021-12-13 GMT
## Daily BEF  exchange rate as at end of day 2021-12-14 GMT
## Daily SEK  exchange rate as at end of day 2021-12-14 GMT
## Daily INR  exchange rate as at end of day 2021-12-13 GMT
## Daily IEP  exchange rate as at end of day 2021-12-14 GMT
## Daily ATS  exchange rate as at end of day 2021-12-14 GMT
## Daily NOK  exchange rate as at end of day 2021-12-14 GMT
## Daily BRL  exchange rate as at end of day 2021-12-14 GMT
## Daily FIM  exchange rate as at end of day 2021-12-14 GMT
## Daily SGD  exchange rate as at end of day 2021-12-14 GMT
## Daily THB  exchange rate as at end of day 2021-12-14 GMT
## Daily NLG  exchange rate as at end of day 2021-12-14 GMT
## Daily CNY  exchange rate as at end of day 2021-12-14 GMT
## Daily HUF  exchange rate as at end of day 2021-12-14 GMT
## Daily CZK  exchange rate as at end of day 2021-12-14 GMT
## Daily PLN  exchange rate as at end of day 2021-12-14 GMT
## Daily KRW  exchange rate as at end of day 2021-12-13 GMT
## Daily CHF  exchange rate as at end of day 2021-12-14 GMT
## Daily ISK  exchange rate as at end of day 2021-12-14 GMT
## Daily EGP  exchange rate as at end of day 2021-12-14 GMT
## Daily BGL  exchange rate as at end of day 2021-12-14 GMT
## Daily TWD  exchange rate as at end of day 2021-12-14 GMT
## Daily MXN  exchange rate as at end of day 2021-12-14 GMT
## Daily LTL  exchange rate as at end of day 2021-12-13 GMT
## Daily NZD  exchange rate as at end of day 2021-12-14 GMT
## Daily ARS  exchange rate as at end of day 2021-12-14 GMT
## Daily VEB  exchange rate as at end of day 2021-12-14 GMT
## Daily NGN  exchange rate as at end of day 2021-12-14 GMT
## Daily LVL  exchange rate as at end of day 2021-12-14 GMT
## Daily ZAR  exchange rate as at end of day 2021-12-14 GMT
## Daily PKR  exchange rate as at end of day 2021-12-14 GMT
## Daily TRL  exchange rate as at end of day 2021-12-14 GMT
## Daily IDR  exchange rate as at end of day 2021-12-14 GMT
## Daily PHP  exchange rate as at end of day 2021-12-14 GMT
## Daily ILS  exchange rate as at end of day 2021-12-14 GMT
## Daily AMD  exchange rate as at end of day 2021-12-14 GMT
```

```
stage_1$worlwide_gross_income[!str_detect(stage_1$worlwide_gross_income,
    "^\\$")] = convert_currency(stage_1$worlwide_gross_income[!str_detect(stage_1$worlwide_gross_income
    "^\\$")])  # Currency Conversion
stage_1 = na.omit(stage_1)

stage_1$budget[str_detect(stage_1$budget, "^\\$")] = as.numeric(str_sub(stage_1$budget[str_detect(stage_
    "^\\$")], 3))  # Dollar removal
stage_1$worlwide_gross_income[str_detect(stage_1$worlwide_gross_income,
    "^\\$")] = as.numeric(str_sub(stage_1$worlwide_gross_income[str_detect(stage_1$worlwide_gross_income
    "^\\$")], 3))  #Dollar Removal
stage_1$budget = as.numeric(stage_1$budget)
stage_1$worlwide_gross_income = as.numeric(stage_1$worlwide_gross_income)

stage_1 = stage_1 %>%
    mutate(`hit/not` = ifelse(worlwide_gross_income/budget >
        1, 1, 0))

stage_1 <- na.omit(stage_1)

png("11_stage_1_post_cc.png", height = 50 * nrow(head(stage_1)),
    width = 200 * ncol(head(stage_1)))
grid.table(head(stage_1))
```

```r
png("12_stage_1_post_cc.png", height = 50 * nrow(summary(stage_1)),
    width = 150 * ncol(summary(stage_1)))
grid.table(summary(stage_1))

head(stage_1)
```

```
## # A tibble: 6 x 15
##   adult genres         imdb_id  popularity runtime vote_count production_countr~
##   <lgl> <chr>          <chr>         <dbl>   <dbl>      <dbl> <chr>
## 1 FALSE Animation, Co~ tt01147~      21.9       81       5415 United States of ~
## 2 FALSE Adventure, Fa~ tt01134~      17.0      104       2413 United States of ~
## 3 FALSE Romance, Come~ tt01132~      11.7      101         92 United States of ~
## 4 FALSE Comedy, Drama~ tt01148~       3.86     127         34 United States of ~
## 5 FALSE Comedy         tt01130~       8.39     106        173 United States of ~
## 6 FALSE Action, Crime~ tt01132~      17.9      170       1886 United States of ~
## # ... with 8 more variables: original_language <chr>, title <chr>, year <dbl>,
## #   director <chr>, budget <dbl>, worlwide_gross_income <dbl>,
## #   weighted_average_vote <dbl>, hit/not <dbl>
```

```r
## Calling the split_cols function to convert production
## companies into columns and sparse filling the cells
stage_1 <- as.data.frame(split_cols(stage_1$production_countries,
    "production_countries", stage_1))

## Calling the split_cols function to convert genres into
## columns and sparse filling the cells
stage_1 <- as.data.frame(split_cols(stage_1$genres, "genres",
    stage_1))

png("13_stage_1_post_pivoting.png", height = 50 * nrow(head(stage_1)),
    width = 200 * ncol(head(stage_1)))
grid.table(head(stage_1))

png("14_stage_1_post_pivoting.png", height = 50 * nrow(summary(stage_1)),
    width = 150 * ncol(summary(stage_1)))
grid.table(summary(stage_1))

head(stage_1)
```

```
##   adult                          genres   imdb_id popularity runtime vote_count
## 1 FALSE      Animation, Comedy, Family tt0114709  21.946943      81       5415
## 2 FALSE     Adventure, Fantasy, Family tt0113497  17.015539     104       2413
## 3 FALSE                Romance, Comedy tt0113228  11.712900     101         92
## 4 FALSE         Comedy, Drama, Romance tt0114885   3.859495     127         34
## 5 FALSE                         Comedy tt0113041   8.387519     106        173
## 6 FALSE Action, Crime, Drama, Thriller tt0113277  17.924927     170       1886
##       production_countries original_language                       title year
## 1 United States of America                en                   Toy Story 1995
## 2 United States of America                en                     Jumanji 1995
## 3 United States of America                en            Grumpier Old Men 1995
## 4 United States of America                en           Waiting to Exhale 1995
## 5 United States of America                en Father of the Bride Part II 1995
## 6 United States of America                en                        Heat 1995
##          director  budget worlwide_gross_income weighted_average_vote hit/not
## 1   John Lasseter 3.0e+07             404265438                   8.3       1
## 2    Joe Johnston 6.5e+07             262821940                   7.0       1
## 3   Howard Deutch 2.5e+07              71518503                   6.7       1
## 4 Forest Whitaker 1.6e+07              81452156                   5.9       1
## 5   Charles Shyer 3.0e+07              76594107                   6.1       1
## 6    Michael Mann 6.0e+07             187436818                   8.2       1
##   United States of America Germany United Kingdom France Italy Australia
## 1                        1       0              0      0     0         0
## 2                        1       0              0      0     0         0
## 3                        1       0              0      0     0         0
## 4                        1       0              0      0     0         0
## 5                        1       0              0      0     0         0
## 6                        1       0              0      0     0         0
##   Belgium Canada Iran Netherlands Hong Kong Japan Austria New Zealand Mexico
## 1       0      0    0           0         0     0       0           0      0
## 2       0      0    0           0         0     0       0           0      0
## 3       0      0    0           0         0     0       0           0      0
## 4       0      0    0           0         0     0       0           0      0
## 5       0      0    0           0         0     0       0           0      0
## 6       0      0    0           0         0     0       0           0      0
##   Taiwan Peru China South Africa Denmark Spain Serbia Sweden Czech Republic
## 1      0    0     0            0       0     0      0      0              0
## 2      0    0     0            0       0     0      0      0              0
## 3      0    0     0            0       0     0      0      0              0
## 4      0    0     0            0       0     0      0      0              0
## 5      0    0     0            0       0     0      0      0              0
## 6      0    0     0            0       0     0      0      0              0
##   Ireland Trinidad and Tobago Russia India Brazil Aruba Israel Luxembourg
## 1       0                   0      0     0      0     0      0          0
## 2       0                   0      0     0      0     0      0          0
## 3       0                   0      0     0      0     0      0          0
## 4       0                   0      0     0      0     0      0          0
## 5       0                   0      0     0      0     0      0          0
## 6       0                   0      0     0      0     0      0          0
##   Argentina Ecuador Bahamas Malaysia Switzerland Bulgaria Thailand Namibia
## 1         0       0       0        0           0        0        0       0
## 2         0       0       0        0           0        0        0       0
## 3         0       0       0        0           0        0        0       0
## 4         0       0       0        0           0        0        0       0
## 5         0       0       0        0           0        0        0       0
## 6         0       0       0        0           0        0        0       0
##   South Korea Norway Finland Afghanistan Iceland Romania Soviet Union Hungary
## 1           0      0       0           0       0       0            0       0
## 2           0      0       0           0       0       0            0       0
## 3           0      0       0           0       0       0            0       0
## 4           0      0       0           0       0       0            0       0
## 5           0      0       0           0       0       0            0       0
## 6           0      0       0           0       0       0            0       0
##   Chile Bhutan Poland Palestinian Territory Uruguay Turkey Morocco Algeria
## 1     0      0      0                     0       0      0       0       0
## 2     0      0      0                     0       0      0       0       0
## 3     0      0      0                     0       0      0       0       0
## 4     0      0      0                     0       0      0       0       0
## 5     0      0      0                     0       0      0       0       0
## 6     0      0      0                     0       0      0       0       0
##   Singapore Mongolia Bosnia and Herzegovina Mali Lebanon Kazakhstan Greece
## 1         0        0                      0    0       0          0      0
## 2         0        0                      0    0       0          0      0
## 3         0        0                      0    0       0          0      0
## 4         0        0                      0    0       0          0      0
## 5         0        0                      0    0       0          0      0
## 6         0        0                      0    0       0          0      0
##   United Arab Emirates Indonesia Egypt Slovenia Macedonia Estonia Portugal
## 1                    0         0     0        0         0       0        0
## 2                    0         0     0        0         0       0        0
## 3                    0         0     0        0         0       0        0
## 4                    0         0     0        0         0       0        0
## 5                    0         0     0        0         0       0        0
## 6                    0         0     0        0         0       0        0
##   Mauritania Cyprus Bangladesh Vietnam Lithuania Jordan Nigeria Philippines
## 1          0      0          0       0         0      0       0           0
## 2          0      0          0       0         0      0       0           0
## 3          0      0          0       0         0      0       0           0
## 4          0      0          0       0         0      0       0           0
## 5          0      0          0       0         0      0       0           0
## 6          0      0          0       0         0      0       0           0
##   Venezuela Pakistan Burkina Faso Latvia Cuba Malta Qatar Samoa Ukraine
## 1         0        0            0      0    0     0     0     0       0
## 2         0        0            0      0    0     0     0     0       0
## 3         0        0            0      0    0     0     0     0       0
## 4         0        0            0      0    0     0     0     0       0
## 5         0        0            0      0    0     0     0     0       0
## 6         0        0            0      0    0     0     0     0       0
##   Colombia Cambodia Panama Georgia Dominican Republic Azerbaijan Armenia
## 1        0        0      0       0                  0          0       0
## 2        0        0      0       0                  0          0       0
## 3        0        0      0       0                  0          0       0
## 4        0        0      0       0                  0          0       0
## 5        0        0      0       0                  0          0       0
## 6        0        0      0       0                  0          0       0
##   Botswana Croatia Costa Rica Ghana Tunisia Rwanda Angola Monaco Puerto Rico
## 1        0       0          0     0       0      0      0      0           0
## 2        0       0          0     0       0      0      0      0           0
## 3        0       0          0     0       0      0      0      0           0
## 4        0       0          0     0       0      0      0      0           0
## 5        0       0          0     0       0      0      0      0           0
## 6        0       0          0     0       0      0      0      0           0
##    "Lao People Slovakia Gibraltar Liechtenstein Chad Iraq Serbia and Montenegro
## 1            0        0         0             0    0    0                     0
## 2            0        0         0             0    0    0                     0
## 3            0        0         0             0    0    0                     0
## 4            0        0         0             0    0    0                     0
## 5            0        0         0             0    0    0                     0
## 6            0        0         0             0    0    0                     0
##   Paraguay Animation Adventure Romance Comedy Action History Drama Crime
## 1        0         1         0       0      1      0       0     0     0
## 2        0         0         1       0      0      0       0     0     0
## 3        0         0         0       1      1      0       0     0     0
## 4        0         0         0       1      1      0       0     1     0
## 5        0         0         0       0      1      0       0     0     0
## 6        0         0         0       0      0      1       0     1     1
##   Fantasy Science Fiction Music Horror Family Mystery Thriller Western War
## 1       0               0     0      0      1       0        0       0   0
## 2       1               0     0      0      1       0        0       0   0
## 3       0               0     0      0      0       0        0       0   0
## 4       0               0     0      0      0       0        0       0   0
## 5       0               0     0      0      0       0        0       0   0
## 6       0               0     0      0      0       0        1       0   0
##   Documentary TV Movie Foreign
## 1           0        0       0
## 2           0        0       0
## 3           0        0       0
## 4           0        0       0
## 5           0        0       0
## 6           0        0       0
```

```
## Converting the abbreviations into full forms for
## langauage column
lang_codes <- as_tibble(read_csv("language_codes_csv.csv", show_col_types = FALSE))
stage_1 <- dplyr::left_join(stage_1, lang_codes, by = c(original_language = "alpha2"),
    keep = FALSE)

png("15_stage_1_post_lang_codes.png", height = 50 * nrow(head(stage_1)),
    width = 200 * ncol(head(stage_1)))
grid.table(head(stage_1))

png("16_stage_1_post_lang_codes.png", height = 50 * nrow(summary(stage_1)),
    width = 150 * ncol(summary(stage_1)))
grid.table(summary(stage_1))

head(stage_1)
```

```
##   adult                         genres   imdb_id popularity runtime vote_count
## 1 FALSE      Animation, Comedy, Family tt0114709  21.946943      81       5415
## 2 FALSE     Adventure, Fantasy, Family tt0113497  17.015539     104       2413
## 3 FALSE                Romance, Comedy tt0113228  11.712900     101         92
## 4 FALSE         Comedy, Drama, Romance tt0114885   3.859495     127         34
## 5 FALSE                         Comedy tt0113041   8.387519     106        173
## 6 FALSE Action, Crime, Drama, Thriller tt0113277  17.924927     170       1886
##       production_countries original_language                         title year
## 1 United States of America                en                     Toy Story 1995
## 2 United States of America                en                       Jumanji 1995
## 3 United States of America                en              Grumpier Old Men 1995
## 4 United States of America                en             Waiting to Exhale 1995
## 5 United States of America                en Father of the Bride Part II 1995
## 6 United States of America                en                          Heat 1995
##          director  budget worlwide_gross_income weighted_average_vote hit/not
## 1   John Lasseter 3.0e+07             404265438                   8.3       1
## 2    Joe Johnston 6.5e+07             262821940                   7.0       1
## 3   Howard Deutch 2.5e+07              71518503                   6.7       1
## 4 Forest Whitaker 1.6e+07              81452156                   5.9       1
## 5   Charles Shyer 3.0e+07              76594107                   6.1       1
## 6    Michael Mann 6.0e+07             187436818                   8.2       1
##   United States of America Germany United Kingdom France Italy Australia
## 1                        1       0              0      0     0         0
## 2                        1       0              0      0     0         0
## 3                        1       0              0      0     0         0
## 4                        1       0              0      0     0         0
## 5                        1       0              0      0     0         0
## 6                        1       0              0      0     0         0
##   Belgium Canada Iran Netherlands Hong Kong Japan Austria New Zealand Mexico
## 1       0      0    0           0         0     0       0           0      0
## 2       0      0    0           0         0     0       0           0      0
## 3       0      0    0           0         0     0       0           0      0
## 4       0      0    0           0         0     0       0           0      0
## 5       0      0    0           0         0     0       0           0      0
## 6       0      0    0           0         0     0       0           0      0
##   Taiwan Peru China South Africa Denmark Spain Serbia Sweden Czech Republic
## 1      0    0     0            0       0     0      0      0              0
## 2      0    0     0            0       0     0      0      0              0
## 3      0    0     0            0       0     0      0      0              0
## 4      0    0     0            0       0     0      0      0              0
## 5      0    0     0            0       0     0      0      0              0
## 6      0    0     0            0       0     0      0      0              0
##   Ireland Trinidad and Tobago Russia India Brazil Aruba Israel Luxembourg
## 1       0                   0      0     0      0     0      0          0
## 2       0                   0      0     0      0     0      0          0
## 3       0                   0      0     0      0     0      0          0
## 4       0                   0      0     0      0     0      0          0
## 5       0                   0      0     0      0     0      0          0
## 6       0                   0      0     0      0     0      0          0
##   Argentina Ecuador Bahamas Malaysia Switzerland Bulgaria Thailand Namibia
## 1         0       0       0        0           0        0        0       0
## 2         0       0       0        0           0        0        0       0
## 3         0       0       0        0           0        0        0       0
## 4         0       0       0        0           0        0        0       0
## 5         0       0       0        0           0        0        0       0
## 6         0       0       0        0           0        0        0       0
##   South Korea Norway Finland Afghanistan Iceland Romania Soviet Union Hungary
## 1           0      0       0           0       0       0            0       0
## 2           0      0       0           0       0       0            0       0
## 3           0      0       0           0       0       0            0       0
## 4           0      0       0           0       0       0            0       0
## 5           0      0       0           0       0       0            0       0
## 6           0      0       0           0       0       0            0       0
##   Chile Bhutan Poland Palestinian Territory Uruguay Turkey Morocco Algeria
## 1     0      0      0                     0       0      0       0       0
## 2     0      0      0                     0       0      0       0       0
## 3     0      0      0                     0       0      0       0       0
## 4     0      0      0                     0       0      0       0       0
## 5     0      0      0                     0       0      0       0       0
## 6     0      0      0                     0       0      0       0       0
##   Singapore Mongolia Bosnia and Herzegovina Mali Lebanon Kazakhstan Greece
## 1         0        0                      0    0       0          0      0
## 2         0        0                      0    0       0          0      0
## 3         0        0                      0    0       0          0      0
## 4         0        0                      0    0       0          0      0
## 5         0        0                      0    0       0          0      0
## 6         0        0                      0    0       0          0      0
##   United Arab Emirates Indonesia Egypt Slovenia Macedonia Estonia Portugal
## 1                    0         0     0        0         0       0        0
## 2                    0         0     0        0         0       0        0
## 3                    0         0     0        0         0       0        0
## 4                    0         0     0        0         0       0        0
## 5                    0         0     0        0         0       0        0
## 6                    0         0     0        0         0       0        0
##   Mauritania Cyprus Bangladesh Vietnam Lithuania Jordan Nigeria Philippines
## 1          0      0          0       0         0      0       0           0
## 2          0      0          0       0         0      0       0           0
## 3          0      0          0       0         0      0       0           0
## 4          0      0          0       0         0      0       0           0
## 5          0      0          0       0         0      0       0           0
## 6          0      0          0       0         0      0       0           0
##   Venezuela Pakistan Burkina Faso Latvia Cuba Malta Qatar Samoa Ukraine
## 1         0        0            0      0    0     0     0     0       0
## 2         0        0            0      0    0     0     0     0       0
## 3         0        0            0      0    0     0     0     0       0
## 4         0        0            0      0    0     0     0     0       0
## 5         0        0            0      0    0     0     0     0       0
## 6         0        0            0      0    0     0     0     0       0
##   Colombia Cambodia Panama Georgia Dominican Republic Azerbaijan Armenia
## 1        0        0      0       0                  0          0       0
## 2        0        0      0       0                  0          0       0
## 3        0        0      0       0                  0          0       0
## 4        0        0      0       0                  0          0       0
## 5        0        0      0       0                  0          0       0
## 6        0        0      0       0                  0          0       0
##   Botswana Croatia Costa Rica Ghana Tunisia Rwanda Angola Monaco Puerto Rico
## 1        0       0          0     0       0      0      0      0           0
## 2        0       0          0     0       0      0      0      0           0
## 3        0       0          0     0       0      0      0      0           0
## 4        0       0          0     0       0      0      0      0           0
## 5        0       0          0     0       0      0      0      0           0
## 6        0       0          0     0       0      0      0      0           0
##    "Lao People Slovakia Gibraltar Liechtenstein Chad Iraq Serbia and Montenegro
## 1            0        0         0             0    0    0                     0
## 2            0        0         0             0    0    0                     0
## 3            0        0         0             0    0    0                     0
## 4            0        0         0             0    0    0                     0
## 5            0        0         0             0    0    0                     0
## 6            0        0         0             0    0    0                     0
##   Paraguay Animation Adventure Romance Comedy Action History Drama Crime
## 1        0         1         0       0      1      0       0     0     0
## 2        0         0         1       0      0      0       0     0     0
## 3        0         0         0       1      1      0       0     0     0
## 4        0         0         0       1      1      0       0     1     0
## 5        0         0         0       0      1      0       0     0     0
## 6        0         0         0       0      0      1       0     1     1
##   Fantasy Science Fiction Music Horror Family Mystery Thriller Western War
## 1       0               0     0      0      1       0        0       0   0
## 2       1               0     0      0      1       0        0       0   0
## 3       0               0     0      0      0       0        0       0   0
## 4       0               0     0      0      0       0        0       0   0
## 5       0               0     0      0      0       0        0       0   0
## 6       0               0     0      0      0       0        1       0   0
##   Documentary TV Movie Foreign English
## 1           0        0       0 English
## 2           0        0       0 English
## 3           0        0       0 English
## 4           0        0       0 English
## 5           0        0       0 English
## 6           0        0       0 English
```

```
## relocating the response variable to the last position
## column wise
stage_1 <- relocate(stage_1, `hit/not`, .after = last_col())

png("17_stage_1_post_relocation_n_final.png", height = 50 * nrow(head(stage_1)),
    width = 200 * ncol(head(stage_1)))
grid.table(head(stage_1))

png("18_stage_1_post_relocation_n_final", height = 50 * nrow(summary(stage_1)),
    width = 150 * ncol(summary(stage_1)))
grid.table(summary(stage_1))

head(stage_1)
```

```
##   adult                         genres   imdb_id popularity runtime vote_count
## 1 FALSE      Animation, Comedy, Family tt0114709  21.946943      81       5415
## 2 FALSE     Adventure, Fantasy, Family tt0113497  17.015539     104       2413
## 3 FALSE                Romance, Comedy tt0113228  11.712900     101         92
## 4 FALSE         Comedy, Drama, Romance tt0114885   3.859495     127         34
## 5 FALSE                         Comedy tt0113041   8.387519     106        173
## 6 FALSE Action, Crime, Drama, Thriller tt0113277  17.924927     170       1886
##       production_countries original_language                       title year
## 1 United States of America                en                   Toy Story 1995
## 2 United States of America                en                     Jumanji 1995
## 3 United States of America                en            Grumpier Old Men 1995
## 4 United States of America                en           Waiting to Exhale 1995
## 5 United States of America                en Father of the Bride Part II 1995
## 6 United States of America                en                        Heat 1995
##          director  budget worlwide_gross_income weighted_average_vote
## 1   John Lasseter 3.0e+07             404265438                   8.3
## 2    Joe Johnston 6.5e+07             262821940                   7.0
## 3   Howard Deutch 2.5e+07              71518503                   6.7
## 4 Forest Whitaker 1.6e+07              81452156                   5.9
## 5   Charles Shyer 3.0e+07              76594107                   6.1
## 6    Michael Mann 6.0e+07             187436818                   8.2
##   United States of America Germany United Kingdom France Italy Australia
## 1                        1       0              0      0     0         0
## 2                        1       0              0      0     0         0
## 3                        1       0              0      0     0         0
## 4                        1       0              0      0     0         0
## 5                        1       0              0      0     0         0
## 6                        1       0              0      0     0         0
##   Belgium Canada Iran Netherlands Hong Kong Japan Austria New Zealand Mexico
## 1       0      0    0           0         0     0       0           0      0
## 2       0      0    0           0         0     0       0           0      0
## 3       0      0    0           0         0     0       0           0      0
## 4       0      0    0           0         0     0       0           0      0
## 5       0      0    0           0         0     0       0           0      0
## 6       0      0    0           0         0     0       0           0      0
##   Taiwan Peru China South Africa Denmark Spain Serbia Sweden Czech Republic
## 1      0    0     0            0       0     0      0      0              0
## 2      0    0     0            0       0     0      0      0              0
## 3      0    0     0            0       0     0      0      0              0
## 4      0    0     0            0       0     0      0      0              0
## 5      0    0     0            0       0     0      0      0              0
## 6      0    0     0            0       0     0      0      0              0
##   Ireland Trinidad and Tobago Russia India Brazil Aruba Israel Luxembourg
## 1       0                   0      0     0      0     0      0          0
## 2       0                   0      0     0      0     0      0          0
## 3       0                   0      0     0      0     0      0          0
## 4       0                   0      0     0      0     0      0          0
## 5       0                   0      0     0      0     0      0          0
## 6       0                   0      0     0      0     0      0          0
##   Argentina Ecuador Bahamas Malaysia Switzerland Bulgaria Thailand Namibia
## 1         0       0       0        0           0        0        0       0
## 2         0       0       0        0           0        0        0       0
## 3         0       0       0        0           0        0        0       0
## 4         0       0       0        0           0        0        0       0
## 5         0       0       0        0           0        0        0       0
## 6         0       0       0        0           0        0        0       0
##   South Korea Norway Finland Afghanistan Iceland Romania Soviet Union Hungary
## 1           0      0       0           0       0       0            0       0
## 2           0      0       0           0       0       0            0       0
## 3           0      0       0           0       0       0            0       0
## 4           0      0       0           0       0       0            0       0
## 5           0      0       0           0       0       0            0       0
## 6           0      0       0           0       0       0            0       0
##   Chile Bhutan Poland Palestinian Territory Uruguay Turkey Morocco Algeria
## 1     0      0      0                     0       0      0       0       0
## 2     0      0      0                     0       0      0       0       0
## 3     0      0      0                     0       0      0       0       0
## 4     0      0      0                     0       0      0       0       0
## 5     0      0      0                     0       0      0       0       0
## 6     0      0      0                     0       0      0       0       0
##   Singapore Mongolia Bosnia and Herzegovina Mali Lebanon Kazakhstan Greece
## 1         0        0                      0    0       0          0      0
## 2         0        0                      0    0       0          0      0
## 3         0        0                      0    0       0          0      0
## 4         0        0                      0    0       0          0      0
## 5         0        0                      0    0       0          0      0
## 6         0        0                      0    0       0          0      0
##   United Arab Emirates Indonesia Egypt Slovenia Macedonia Estonia Portugal
## 1                    0         0     0        0         0       0        0
## 2                    0         0     0        0         0       0        0
## 3                    0         0     0        0         0       0        0
## 4                    0         0     0        0         0       0        0
## 5                    0         0     0        0         0       0        0
## 6                    0         0     0        0         0       0        0
##   Mauritania Cyprus Bangladesh Vietnam Lithuania Jordan Nigeria Philippines
## 1          0      0          0       0         0      0       0           0
## 2          0      0          0       0         0      0       0           0
## 3          0      0          0       0         0      0       0           0
## 4          0      0          0       0         0      0       0           0
## 5          0      0          0       0         0      0       0           0
## 6          0      0          0       0         0      0       0           0
##   Venezuela Pakistan Burkina Faso Latvia Cuba Malta Qatar Samoa Ukraine
## 1         0        0            0      0    0     0     0     0       0
## 2         0        0            0      0    0     0     0     0       0
## 3         0        0            0      0    0     0     0     0       0
## 4         0        0            0      0    0     0     0     0       0
## 5         0        0            0      0    0     0     0     0       0
## 6         0        0            0      0    0     0     0     0       0
##   Colombia Cambodia Panama Georgia Dominican Republic Azerbaijan Armenia
## 1        0        0      0       0                  0          0       0
## 2        0        0      0       0                  0          0       0
## 3        0        0      0       0                  0          0       0
## 4        0        0      0       0                  0          0       0
## 5        0        0      0       0                  0          0       0
## 6        0        0      0       0                  0          0       0
##   Botswana Croatia Costa Rica Ghana Tunisia Rwanda Angola Monaco Puerto Rico
## 1        0       0          0     0       0      0      0      0           0
## 2        0       0          0     0       0      0      0      0           0
## 3        0       0          0     0       0      0      0      0           0
## 4        0       0          0     0       0      0      0      0           0
## 5        0       0          0     0       0      0      0      0           0
## 6        0       0          0     0       0      0      0      0           0
##    "Lao People Slovakia Gibraltar Liechtenstein Chad Iraq Serbia and Montenegro
## 1            0        0         0             0    0    0                     0
## 2            0        0         0             0    0    0                     0
## 3            0        0         0             0    0    0                     0
## 4            0        0         0             0    0    0                     0
## 5            0        0         0             0    0    0                     0
## 6            0        0         0             0    0    0                     0
##   Paraguay Animation Adventure Romance Comedy Action History Drama Crime
## 1        0         1         0       0      1      0       0     0     0
## 2        0         0         1       0      0      0       0     0     0
## 3        0         0         0       1      1      0       0     0     0
## 4        0         0         0       1      1      0       0     1     0
## 5        0         0         0       0      1      0       0     0     0
## 6        0         0         0       0      0      1       0     1     1
##   Fantasy Science Fiction Music Horror Family Mystery Thriller Western War
## 1       0               0     0      0      1       0        0       0   0
## 2       1               0     0      0      1       0        0       0   0
## 3       0               0     0      0      0       0        0       0   0
## 4       0               0     0      0      0       0        0       0   0
## 5       0               0     0      0      0       0        0       0   0
## 6       0               0     0      0      0       0        1       0   0
##   Documentary TV Movie Foreign English hit/not
## 1           0        0       0 English       1
## 2           0        0       0 English       1
## 3           0        0       0 English       1
## 4           0        0       0 English       1
## 5           0        0       0 English       1
## 6           0        0       0 English       1
```