Data Wrangling using R: Application to Netflix Data

Stephen Bernardo

2024-01-21

Package Loading

Load Lubridate

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

Data Loading

Load your dataset NetflixData.csv Download NetflixData.csv Download NetflixData.csvin R using read.csv or readr::read_csv(). Store the DataFrame to a new variable named netflix.

```
netflix <- read.csv('NetflixData.csv')
```

Create and Inspect Columns

1. Extract the value of year from Sys.time() and save it in a new variable named current_year.

```
netflix$current_year <- year(Sys.time())
```

- Inspect the class of current_year.

```
class(netflix$current_year)
```

```
## [1] "numeric"
```

- Convert it to a numeric data class.

```
netflix$current_year <-as.numeric(netflix$current_year)
```

2. Create a new column in the netflix DataFrame you created in question (1). Name this column time_since_release and assign it the value of the current year minus the release_year.

This variable will give you the number of years since the release of the tv show/movie.

```
netflix$time_since_release <- netflix$current_year - netflix$release_year
```

3. Create a new column named title_length. Assign it the value of number of characters in the title of each of the tv show/movie.

```
netflix$title_length <- nchar(netflix$title)
```

4. Inspect the class of each column in the dataframe. Instead of doing this one by one for each column, can you write a “for” loop over the columns of the dataframe that prints the class of each column? HINT: Inside the loop, autoprint is turned off.

```
for (column in names(netflix)) {
  column_class <- class(netflix[[column]])
  cat("Column '", column, "' has class: ", column_class, "\n")
}
```

```
## Column ' X ' has class:  integer
## Column ' type ' has class:  character
## Column ' title ' has class:  character
## Column ' country ' has class:  character
## Column ' date_added ' has class:  character
## Column ' release_year ' has class:  integer
## Column ' rating ' has class:  character
## Column ' duration_min_season ' has class:  integer
## Column ' sales ' has class:  numeric
## Column ' current_year ' has class:  numeric
## Column ' time_since_release ' has class:  numeric
## Column ' title_length ' has class:  integer
```

Data Summary Table

1. Compute the descriptive statistics (mean, median, min, max) for the variable you created time_since_release.

```
summary(netflix$time_since_release)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00    6.00    8.00   11.14   12.00   83.00
```

2. Can you compute the descriptive statistics for all the numerical variables in the data without repeating your code for each variable? HINT: You can use functions like lapply() or sapply() in R to do this.

```
#get all numeric columns
numeric_columns <- sapply(netflix, is.numeric)

#print statistics
print(lapply(netflix[, numeric_columns], summary))
```

```
## $X
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   233.5   466.0   466.0   698.5   931.0
##
## $release_year
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1942    2013    2017    2014    2019    2021
##
## $duration_min_season
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    3.00   90.00   71.07  107.00  312.00
##
## $sales
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.758  12.602  60.727  50.006  69.862 173.768
##
## $current_year
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2025    2025    2025    2025    2025    2025
##
## $time_since_release
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00    6.00    8.00   11.14   12.00   83.00
##
## $title_length
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   10.00   15.00   17.24   22.00   73.00
```

3. Can you create a function that takes two inputs: input_data (a dataframe) and id (a column name). Here are some helpful hints:

- The goal of the function is to compute the descriptive statistics (mean, median, min, max) for the variable in the column id of the dataframe input_data if the column is numeric or integer.
- Use the IF function inside the function to check if the class is integer or numeric Example code: if(class(input_data[,id]) == “numeric” | class(input_data[,id]) == “integer”)
- Create a DataFrame called summary within the loop function that has the required columns (variable, mean, median, min, max). Each column takes the appropriate value (e.g., mean(input_data[,id]) will give you the mean of column id in the input_data).
- The function should return the summary DataFrame. Example code: return(summary)
- Invoke the function you created for the variable time_since_release (Hint: use the column id to call the function). Save the output of the function to a new variable named output_data.

```
compute_summary_statistics <- function(input_data, id) {
  # Check if the column is numeric or integer
  if (class(input_data[, id]) == "numeric" | class(input_data[, id]) == "integer") {

    # Create a DataFrame to store summary statistics
    summary <- data.frame(
      variable = id,
      mean = mean(input_data[, id], na.rm = TRUE),
      median = median(input_data[, id], na.rm = TRUE),
      min = min(input_data[, id], na.rm = TRUE),
      max = max(input_data[, id], na.rm = TRUE)
    )

    # Return the summary DataFrame
    return(summary)
  } else {
    # If the column is not numeric or integer, print a message and return NULL
    cat("The specified column is not numeric or integer.\n")
    return(NULL)
  }
}
```

Invoking function with the time_since_release column

```
output_data <- compute_summary_statistics(input_data = netflix, id = 'time_since_rele
ase')
```

Printing the output

```
print(output_data)
```

```
##             variable     mean median min max
## 1 time_since_release 11.13749      8   4  83
```

Data Analysis and Regressions

1. Split the data into two dataframes named tv_shows and movies. Hint: Use subsetting techniques in R using square brackets.

```
# Create a dataframe for TV shows
tv_shows <- netflix[netflix$type == "TV Show", ]

# Create a dataframe for Movies
movies <- netflix[netflix$type == "Movie", ]
```

2. Run two different regression models for tv_shows only, to examine the relationship between sales and other variables. The dependent variable is sales. Model 1: The independent variables are time_since_release, title_length, duration_min_season. Model 2: In addition to Model 1 variables, include country and rating in your regression.

Model 1

Regression

```
tv_show_model1 <- lm(sales ~ time_since_release + title_length + duration_min_season,
data = tv_shows)

# Summary of Model 1
summary(tv_show_model1)
```

```
##
## Call:
## lm(formula = sales ~ time_since_release + title_length + duration_min_season,
##     data = tv_shows)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1851 -0.8260  0.0220  0.8634  4.4232
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         10.360387   0.226318  45.778  < 2e-16 ***
## time_since_release   0.110952   0.013725   8.084 2.07e-14 ***
## title_length         0.003825   0.009590   0.399    0.690
## duration_min_season -0.041260   0.049880  -0.827    0.409
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.327 on 272 degrees of freedom
## Multiple R-squared:  0.2023, Adjusted R-squared:  0.1935
## F-statistic: 22.99 on 3 and 272 DF,  p-value: 2.705e-13
```

Interpretation

a. coefficients:

- The intercept is 10.473339.
- One unit increase in time_since_release will increase sales by 0.110952
- One unit increase intitle_length will increase sales by 0.003825
- One unit increase duration_min_season will decrease sales by -0.04126

b. p-values: The p-values show that only **time_since_release** is significant, since it is the only variable where p-values were less than the the p-value is less thanthe thresholds of 0, 0.001, 0.01, 0.05, and 0.1.

c. R-squared: The adjusted R-squared is **0.1935**, this means that only around 19.35% of the variability in sales is explained by the independent variables. This means that most of the variability in sales is unexplained by the model and there may be additional variables that explain the variability in sales.

Model 2

```
tv_show_model2 <- lm(sales ~ time_since_release + title_length + duration_min_season
+ country + rating, data = tv_shows)

# Summary of Model 2
summary(tv_show_model2)
```

```
##
## Call:
## lm(formula = sales ~ time_since_release + title_length + duration_min_season +
##     country + rating, data = tv_shows)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9568 -0.7011  0.0000  0.7414  4.6182
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.7719170  1.5018642   7.172 9.82e-12 ***
## time_since_release      0.1126505  0.0150598   7.480 1.52e-12 ***
## title_length            0.0006994  0.0105762   0.066   0.9473
## duration_min_season    -0.0335500  0.0545726  -0.615   0.5393
## countryArgentina       -0.5652755  1.1369133  -0.497   0.6195
## countryAustralia       -1.0935898  0.9126535  -1.198   0.2320
## countryBelgium         -0.6486392  1.4808798  -0.438   0.6618
## countryBrazil          -0.6559955  0.8645632  -0.759   0.4488
## countryCanada          -0.6366348  0.7412025  -0.859   0.3913
## countryChina           -1.4990808  0.9012412  -1.663   0.0976 .
## countryColombia        -2.1540295  1.4711609  -1.464   0.1445
## countryDenmark          0.4839906  1.4804721   0.327   0.7440
## countryFinland         -2.4408737  1.4818100  -1.647   0.1009
## countryFrance           0.4741722  0.7611345   0.623   0.5339
## countryGermany          0.5308010  0.9968998   0.532   0.5949
## countryIndia           -0.1617089  0.7393902  -0.219   0.8271
## countryIreland         -0.7294564  1.4835414  -0.492   0.6234
## countryIsrael          -0.1477560  1.4797765  -0.100   0.9205
## countryItaly           -0.0567180  1.1480545  -0.049   0.9606
## countryJapan           -0.0142749  0.7160378  -0.020   0.9841
## countryLebanon         -0.8891609  1.4716753  -0.604   0.5463
## countryMalaysia        -0.2137256  0.9966154  -0.214   0.8304
## countryMexico          -0.6511711  0.7643416  -0.852   0.3951
## countryNetherlands     -1.7962104  1.4964340  -1.200   0.2312
## countryNorway          -0.0868322  1.4910773  -0.058   0.9536
## countryPoland           1.2450583  1.1376601   1.094   0.2749
## countryRussia           1.1784119  1.5015316   0.785   0.4334
## countrySingapore       -0.6119740  1.1258819  -0.544   0.5873
## countrySouth Africa    -0.0741508  1.4795547  -0.050   0.9601
## countrySouth Korea      0.1383327  0.6960442   0.199   0.8426
## countrySpain           -1.4167888  0.8088916  -1.752   0.0812 .
## countrySweden          -1.8650753  1.1390221  -1.637   0.1029
## countryTaiwan          -0.2973155  0.7397119  -0.402   0.6881
## countryThailand        -0.6986747  0.9952929  -0.702   0.4834
## countryTurkey          -0.1748323  1.1262088  -0.155   0.8768
## countryUnited Kingdom  -0.2670385  0.6551171  -0.408   0.6839
## countryUnited States   -0.1993425  0.6248096  -0.319   0.7500
## ratingTV-14            -0.2484371  1.3597296  -0.183   0.8552
## ratingTV-G             -0.2017103  1.4090380  -0.143   0.8863
## ratingTV-MA             0.0514162  1.3580360   0.038   0.9698
## ratingTV-PG            -0.0865257  1.3627093  -0.063   0.9494
## ratingTV-Y             -0.2595788  1.3790049  -0.188   0.8509
## ratingTV-Y7            -0.5532761  1.4036684  -0.394   0.6938
## ratingTV-Y7-FV          1.0794285  1.9419844   0.556   0.5789
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.339 on 232 degrees of freedom
## Multiple R-squared:  0.3073, Adjusted R-squared:  0.1789
## F-statistic: 2.393 on 43 and 232 DF,  p-value: 1.835e-05
```

Interpretation:

a. coefficients:

- The intercept is 10.8845675
- One unit increase in time_since_release will increase sales by 0.1126505
- One unit increase intitle_length will increase sales by 0.003825
- One unit increase duration_min_season will increase sales by 0.0006994

Ratings - One unit increase in ratingTV-14 leads to a decrease in sales by 0.2484371 - One unit increase in ratingTV-G leads to a decrease in sales by 0.2017103 - One unit increase in ratingTV-MA leads to a increase in sales by 0.0514162 - One unit increase in ratingTV-PG leads to a decrease in sales by 0.0865257 - One unit increase in ratingTV-Y leads to a decrease in sales by 0.2595788 - One unit increase in ratingTV-Y7 leads to a decrease in sales by 0.5532761 - One unit increase in ratingTV-Y7-FV leads to a increase in sales by 1.0794285

Country - One unit increase in countryArgentina leads to a decrease in sales by 0.5652755 - One unit increase in countryAustralia leads to a decrease in sales by 1.0935898 - One unit increase in countryBelgium leads to a decrease in sales by 0.6486392 - One unit increase in countryBrazil leads to a decrease in sales by 0.6559955 - One unit increase in countryCanada leads to a decrease in sales by 0.6366348 - One unit increase in countryChina leads to a decrease in sales by 1.4990808 - One unit increase in countryColombia leads to a decrease in sales by 2.1540295 - One unit increase in countryDenmark leads to a increase in sales by 0.4839906 - One unit increase in countryFinland leads to a decrease in sales by 2.4408737 - One unit increase in countryFrance leads to a increase in sales by 0.4741722 - One unit increase in countryGermany leads to a increase in sales by 0.530801 - One unit increase in countryIndia leads to a decrease in sales by 0.1617089 - One unit increase in countryIreland leads to a decrease in sales by 0.7294564 - One unit increase in countryIsrael leads to a decrease in sales by 0.147756 - One unit increase in countryItaly leads to a decrease in sales by 0.056718 - One unit increase in countryJapan leads to a decrease in sales by 0.0142749 - One unit increase in countryLebanon leads to a decrease in sales by 0.8891609 - One unit increase in countryMalaysia leads to a decrease in sales by 0.2137256 - One unit increase in countryMexico leads to a decrease in sales by 0.6511711 - One unit increase in countryNetherlands leads to a decrease in sales by 1.7962104 - One unit increase in countryNorway leads to a decrease in sales by 0.0868322 - One unit increase in countryPoland leads to a increase in sales by 1.2450583 - One unit increase in countryRussia leads to a increase in sales by 1.1784119 - One unit increase in countrySingapore leads to a decrease in sales by 0.611974 - One unit increase in countrySouth Africa leads to a decrease in sales by 0.0741508 - One unit increase in countrySouth Korea leads to a increase in sales by 0.1383327 - One unit increase in countrySpain leads to a decrease in sales by 1.4167888 - One unit increase in countrySweden leads to a decrease in sales by 1.8650753 - One unit increase in countryTaiwan leads to a decrease in sales by 0.2973155 - One unit increase in countryThailand leads to a decrease in sales by 0.6986747 - One unit increase in countryTurkey leads to a decrease in sales by 0.1748323 - One unit increase in countryUnited Kingdom leads to a decrease in sales by 0.2670385 - One unit increase in countryUnited States leads to a decrease in sales by 0.1993425

b. p-values: The p-values show that

- **time_since_release** is significant since p-values were less than the thresholds of 0, 0.001, 0.01, 0.05, and 0.1.
- **countryChina** and **countrySpain** are significant if we were use a threshold were p-values are less than 0.1.

- The rest of the variables are not significant.

c. R-squared: The adjusted R-squared is **0.1789**, this means that only around 17.89% of the variability in sales is explained by the independent variables. This means that most of the variability in sales is unexplained by the model and there may be additional variables that explain the variability in sales. In addition, this may mean that country and rating did not help in further explaining the variability in sales, as determined by the p values.

3. Run two different regression models for movies only, to examine the relationship between sales and other variables. The dependent variable is sales. Model 1: The independent variables are time_since_release, title_length, duration_min_season. Model 2: In addition to Model 1 variables, include country and rating in your regression. HINT: Remember these are factor variables? Report and interpret the regression results (coefficients, p-values, R-squared) within your .Rmd/.html file.

Model 1

```
movies_model1 <- lm(sales ~ time_since_release + title_length + duration_min_season,
data = movies)

# Summary of Model 1
summary(movies_model1)
```

```
##
## Call:
## lm(formula = sales ~ time_since_release + title_length + duration_min_season,
##     data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1131 -1.0228  0.0509  0.9879  3.7867
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         17.736653   0.249579  71.066   <2e-16 ***
## time_since_release  -0.095526   0.005733 -16.664   <2e-16 ***
## title_length        -0.009509   0.005580  -1.704   0.0889 .
## duration_min_season  0.497798   0.001964 253.509   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.417 on 651 degrees of freedom
## Multiple R-squared:  0.991,  Adjusted R-squared:  0.991
## F-statistic: 2.388e+04 on 3 and 651 DF,  p-value: < 2.2e-16
```

Interpretation:

a. coefficients:

- The intercept is 17.641127.
- One unit increase in time_since_release will decrease sales by 0.095526.
- One unit increase intitle_length will decrease sales by 0.009509.
- One unit increase duration_min_season will increase sales by 0.497798.

b. p-values: The p-values show that only **time_since_release** and **duration_min_season** are significant, since the p-values of these variables were less than the the p-value is less thanthe thresholds of 0, 0.001, 0.01, 0.05, and 0.1.

c. R-squared: The adjusted R-squared is **0.991**, this means that only around 99.10% of the variability in sales is explained by the independent variables. This means that only a little variability in sales is unexplained by the model.

Model 2

Regression:

```
movies_model2 <- lm(sales ~ time_since_release + title_length + duration_min_season +
rating + country, data = movies)

# Summary of Model 2
summary(movies_model2)
```

```
##
## Call:
## lm(formula = sales ~ time_since_release + title_length + duration_min_season +
##     rating + country, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0528 -0.8745  0.0077  0.9113  3.9765
##
## Coefficients: (1 not defined because of singularities)
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           21.366005   2.082521  10.260   <2e-16 ***
## time_since_release    -0.096582   0.006152 -15.698   <2e-16 ***
## title_length          -0.013723   0.006014  -2.282   0.0229 *
## duration_min_season    0.500260   0.002405 208.006   <2e-16 ***
## ratingNR              -2.860948   1.485287  -1.926   0.0546 .
## ratingPG              -3.656433   1.434041  -2.550   0.0110 *
## ratingPG-13           -3.310065   1.430281  -2.314   0.0210 *
## ratingR               -3.548990   1.422994  -2.494   0.0129 *
## ratingTV-14           -3.234192   1.422158  -2.274   0.0233 *
## ratingTV-G            -3.041819   1.455434  -2.090   0.0370 *
## ratingTV-MA           -3.068279   1.418863  -2.162   0.0310 *
## ratingTV-PG           -3.084497   1.427958  -2.160   0.0312 *
## ratingTV-Y            -3.127097   1.500857  -2.084   0.0376 *
## ratingTV-Y7           -3.363143   1.473114  -2.283   0.0228 *
## ratingTV-Y7-FV        -3.704739   2.496485  -1.484   0.1383
## countryArgentina      -0.036944   1.722311  -0.021   0.9829
## countryAustralia      -0.031592   1.584831  -0.020   0.9841
## countryAustria        -0.480890   1.813013  -0.265   0.7909
## countryBelgium        -3.921930   2.112446  -1.857   0.0639 .
## countryBrazil         -0.379702   1.567392  -0.242   0.8087
## countryBulgaria       -0.790722   2.067051  -0.383   0.7022
## countryCambodia       -1.433127   2.070775  -0.692   0.4892
## countryCanada         -0.349329   1.521657  -0.230   0.8185
## countryChile           0.404646   2.073227   0.195   0.8453
## countryChina          -1.529484   1.676535  -0.912   0.3620
## countryColombia       -1.931321   2.073636  -0.931   0.3520
## countryDenmark               NA         NA      NA       NA
## countryEgypt          -0.213866   1.570051  -0.136   0.8917
## countryFinland        -3.179871   2.070654  -1.536   0.1252
## countryFrance         -0.462372   1.540853  -0.300   0.7642
## countryGermany        -0.794812   1.616394  -0.492   0.6231
## countryGhana          -1.836005   1.815533  -1.011   0.3123
## countryGreece          0.969841   2.064729   0.470   0.6387
## countryHong Kong      -0.260197   1.576185  -0.165   0.8689
## countryIndia          -0.761111   1.528607  -0.498   0.6187
## countryIndonesia      -1.148243   1.576239  -0.728   0.4666
## countryIreland        -0.916516   1.717097  -0.534   0.5937
## countryIsrael         -1.837470   1.716832  -1.070   0.2849
## countryItaly          -0.590844   1.601708  -0.369   0.7123
## countryJapan          -0.677441   1.558546  -0.435   0.6640
## countryKuwait         -2.054212   2.077266  -0.989   0.3231
## countryLebanon        -0.149511   2.065620  -0.072   0.9423
## countryMalaysia       -1.607421   1.818178  -0.884   0.3770
## countryMexico         -0.759188   1.573322  -0.483   0.6296
```

```
## countryNamibia          0.435290   2.073009   0.210   0.8338
## countryNetherlands     -0.859335   1.678664  -0.512   0.6089
## countryNew Zealand     -2.045396   1.809385  -1.130   0.2588
## countryNigeria         -1.126068   1.572059  -0.716   0.4741
## countryNorway          -0.507200   1.721121  -0.295   0.7683
## countryPakistan         0.179190   1.625634   0.110   0.9123
## countryPhilippines     -0.762897   1.589362  -0.480   0.6314
## countrySaudi Arabia    -1.387413   1.619457  -0.857   0.3919
## countrySerbia           0.555125   2.067971   0.268   0.7885
## countrySingapore       -1.251333   2.070877  -0.604   0.5459
## countrySlovenia        -0.119018   2.067313  -0.058   0.9541
## countrySouth Africa    -0.745710   1.669704  -0.447   0.6553
## countrySouth Korea     -0.648960   1.610003  -0.403   0.6870
## countrySpain            0.245309   1.562820   0.157   0.8753
## countrySweden          -2.162021   2.070486  -1.044   0.2968
## countryTaiwan          -1.881120   1.816024  -1.036   0.3007
## countryThailand        -2.017988   1.812355  -1.113   0.2660
## countryTurkey          -1.453623   1.573467  -0.924   0.3559
## countryUnited Kingdom -0.412159   1.523608  -0.271   0.7869
## countryUnited States  -0.407710   1.505265  -0.271   0.7866
## countryUruguay         -0.709834   2.073396  -0.342   0.7322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.41 on 591 degrees of freedom
## Multiple R-squared:  0.9919, Adjusted R-squared:  0.991
## F-statistic:  1148 on 63 and 591 DF,  p-value: < 2.2e-16
```

Interpretation

a. coefficients:

- The intercept is 21.269423
- One unit increase in time_since_release leads to a decrease in sales by 0.096582
- One unit increase in title_length leads to a decrease in sales by 0.013723
- One unit increase in duration_min_season leads to a increase in sales by 0.50026

Ratings - One unit increase in ratingNR leads to a decrease in sales by 2.860948 - One unit increase in ratingPG leads to a decrease in sales by 3.656433 - One unit increase in ratingPG-13 leads to a decrease in sales by 3.310065 - One unit increase in ratingR leads to a decrease in sales by 3.54899 - One unit increase in ratingTV-14 leads to a decrease in sales by 3.234192 - One unit increase in ratingTV-G leads to a decrease in sales by 3.041819 - One unit increase in ratingTV-MA leads to a decrease in sales by 3.068279 - One unit increase in ratingTV-PG leads to a decrease in sales by 3.084497 - One unit increase in ratingTV-Y leads to a decrease in sales by 3.127097 - One unit increase in ratingTV-Y7 leads to a decrease in sales by 3.363143 - One unit increase in ratingTV-Y7-FV leads to a decrease in sales by 3.704739

Country - One unit increase in countryArgentina leads to a decrease in sales by 0.036944 - One unit increase in countryAustralia leads to a decrease in sales by 0.031592 - One unit increase in countryAustria leads to a decrease in sales by 0.48089 - One unit increase in countryBelgium leads to a decrease in sales by 3.92193 - One unit increase in countryBrazil leads to a decrease in sales by 0.379702 - One unit increase in countryBulgaria leads to a decrease in sales by 0.790722 - One unit increase in countryCambodia leads to a decrease in sales by 1.433127 - One unit increase in countryCanada leads to a decrease in sales by 0.349329 - One unit increase in countryChile leads to a increase in sales by 0.404646 - One unit increase in countryChina leads to a decrease in sales by 1.529484 - One unit increase in countryColombia leads to a decrease in sales by 1.931321 - There is no coefficient identified for countryDenmark - One unit increase in countryEgypt leads to a decrease in sales by 0.213866 - One unit increase in countryFinland leads to a

decrease in sales by 3.179871 - One unit increase in countryFrance leads to a decrease in sales by 0.462372 - One unit increase in countryGermany leads to a decrease in sales by 0.794812 - One unit increase in countryGhana leads to a decrease in sales by 1.836005 - One unit increase in countryGreece leads to a increase in sales by 0.969841 - One unit increase in countryHong Kong leads to a decrease in sales by 0.260197 - One unit increase in countryIndia leads to a decrease in sales by 0.761111 - One unit increase in countryIndonesia leads to a decrease in sales by 1.148243 - One unit increase in countryIreland leads to a decrease in sales by 0.916516 - One unit increase in countryIsrael leads to a decrease in sales by 1.83747 - One unit increase in countryItaly leads to a decrease in sales by 0.590844 - One unit increase in countryJapan leads to a decrease in sales by 0.677441 - One unit increase in countryKuwait leads to a decrease in sales by 2.054212 - One unit increase in countryLebanon leads to a decrease in sales by 0.149511 - One unit increase in countryMalaysia leads to a decrease in sales by 1.607421 - One unit increase in countryMexico leads to a decrease in sales by 0.759188 - One unit increase in countryNamibia leads to a increase in sales by 0.43529 - One unit increase in countryNetherlands leads to a decrease in sales by 0.859335 - One unit increase in countryNew Zealand leads to a decrease in sales by 2.045396 - One unit increase in countryNigeria leads to a decrease in sales by 1.126068 - One unit increase in countryNorway leads to a decrease in sales by 0.5072 - One unit increase in countryPakistan leads to a increase in sales by 0.17919 - One unit increase in countryPhilippines leads to a decrease in sales by 0.762897 - One unit increase in countrySaudi Arabia leads to a decrease in sales by 1.387413 - One unit increase in countrySerbia leads to a increase in sales by 0.555125 - One unit increase in countrySingapore leads to a decrease in sales by 1.251333 - One unit increase in countrySlovenia leads to a decrease in sales by 0.119018 - One unit increase in countrySouth Africa leads to a decrease in sales by 0.74571 - One unit increase in countrySouth Korea leads to a decrease in sales by 0.64896 - One unit increase in countrySpain leads to a increase in sales by 0.245309 - One unit increase in countrySweden leads to a decrease in sales by 2.162021 - One unit increase in countryTaiwan leads to a decrease in sales by 1.88112 - One unit increase in countryThailand leads to a decrease in sales by 2.017988 - One unit increase in countryTurkey leads to a decrease in sales by 1.453623 - One unit increase in countryUnited Kingdom leads to a decrease in sales by 0.412159 - One unit increase in countryUnited States leads to a decrease in sales by 0.40771 - One unit increase in countryUruguay leads to a decrease in sales by 0.709834

b. p-values: The p-values show that

- only **time_since_release** and **duration_min_season** are significant, since the p-values of these variables were less than the the p-value is less than the thresholds of 0, 0.001, 0.01, 0.05, and 0.1.
- **title_length** is significant when compared to the p-value thresholds of 0.05 and 0.01
- For ratings, ratingNR is significant if the p-value threshold used is 0.01. In addition, ratingNR, ratingPG, ratingPG-13, ratingR, ratingTV-14, ratingTV-G, ratingTV-MA, ratingTV-PG, ratingTV-Y, ratingTV-Y7 are significant under a p-value threshold of 0.05 and 0.1.
- **countryBelgium** is significant if we were use a threshold were p-values are less than 0.1.
- The rest of the variables are not significant.

c. R-squared: The adjusted R-squared is **0.1935**, this means that only around 19.35% of the variability in sales is explained by the independent variables. This means that most of the variability in sales is unexplained by the model and there may be additional variables that explain the variability in sales.

4. What are the differences you observe in your results for the regression outputs for tv_shows and movies? Which variables are significant?

Model 1

a. coefficients:

- The intercept for TV (10.473339) is less than that for movies (17.641127)
- For time_since_release, one unit increase will mean an increase sales by 0.110952 for TV shows while it will mean decrease sales by -0.095526 for movies

- For title_length, one unit increase will mean increase sales by 0.003825 for TV shows while it will mean decrease sales by -0.009509 for movies. For duration_min_season, one unit increase will mean decrease sales by -0.04126 for TV shows while it will mean increase sales by 0.497798 for movies.

b. p-values: Only time_since_release is significant for TV shows. O model1_movies <- lm(sales ~ time_since_release + title_length + duration_min_season, data = movies)

c, R-squared: The adjusted R-squared for TV shows (0.1935) is less than the R-squared for movies (0.991), which means that Model 1 can explain more of the variability for movies than in TV shows.

Model 2

a. coefficients:

- The intercept for TV (10.8845675) is less than that for movies (21.269423)
- For time_since_release, one unit increase will mean an increase sales by 0.1126505 for TV shows while it will mean decrease sales by -0.096582 for movies
- For title_length, one unit increase will mean increase sales by 0.003825 for TV shows while it will mean decrease sales by -0.013723 for movies. For duration_min_season, one unit increase will mean increase sales by -0.0006994 for TV shows while it will mean increase sales by 0.50026 for movies.

b. p-values:

- Only me_since_release is significant for TV shows. On the other hand, time_since_release and duration_min_season are significant.
- title_length is not significant for TV shows. On the other hand, title_length is significant for movies when compared to the p-value thresholds of 0.05 and 0.01
- ratings are not significant for TV shows. On the other hand, for movies, ratingNR is significant if the p-value threshold used is 0.01. In addition, for movies, In addition, ratingNR, ratingPG, ratingPG-13, ratingR, ratingTV-14, ratingTV-G, ratingTV-MA, ratingTV-PG, ratingTV-Y, ratingTV-Y7 are significant under a p-value threshold of 0.05 and 0.1.
- for country, **countryChina** and **countrySpain** were significant for tv shows if p-values are less than 0.1 in threshold. One the other hand, for movies, **countryBelgium** is significant if we were use a threshold were p-values are less than 0.1.

c. R-squared: The adjusted R-squared for TV shows (0.1789) is less than the R-squared for movies (0.1935), which means that Model 2 can explain more of the variability for movies than in TV shows.