Movie Data Correlation

Konrad Gawel

8/11/2021

Introduction

The following is a quick case study which aims to explore movie data and find correlations between variables. Data can be found at https://www.kaggle.com/danielgrijalvas/movies

Load necessary libraries for data cleaning, exploration and visualization

```
library(readr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v dplyr   1.0.7
## v tibble  3.1.3     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stringr)
library(ggplot2)
library(janitor)
```

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(lubridate)
```

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

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

```
library(ggcorrplot)
```

Import data

```
movies_df_original = read_csv("movies.csv")
```

```
## Rows: 7668 Columns: 15
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): name, rating, genre, released, director, writer, star, country, com...
## dbl (6): year, score, votes, budget, gross, runtime
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Take a look and explore the data set

```
head(movies_df_original)
```

```
## # A tibble: 6 x 15
##   name   rating genre   year released score  votes director writer star  country
##   <chr>  <chr>  <chr>  <dbl> <chr>    <dbl>  <dbl> <chr>    <chr>  <chr> <chr>
## 1 The S~ R      Drama   1980 June 13~   8.4 9.27e5 Stanley~ Steph~ Jack~ United~
## 2 The B~ R      Adven~  1980 July 2,~   5.8 6.5 e4 Randal ~ Henry~ Broo~ United~
## 3 Star ~ PG     Action  1980 June 20~   8.7 1.2 e6 Irvin K~ Leigh~ Mark~ United~
## 4 Airpl~ PG     Comedy  1980 July 2,~   7.7 2.21e5 Jim Abr~ Jim A~ Robe~ United~
## 5 Caddy~ R      Comedy  1980 July 25~   7.3 1.08e5 Harold ~ Brian~ Chev~ United~
## 6 Frida~ R      Horror  1980 May 9, ~   6.4 1.23e5 Sean S.~ Victo~ Bets~ United~
## # ... with 4 more variables: budget <dbl>, gross <dbl>, company <chr>,
## #   runtime <dbl>
```

```
glimpse(movies_df_original)
```

```
## Rows: 7,668
## Columns: 15
## $ name     <chr> "The Shining", "The Blue Lagoon", "Star Wars: Episode V - The~
## $ rating   <chr> "R", "R", "PG", "PG", "R", "R", "R", "R", "PG", "R", "PG", "P~
## $ genre    <chr> "Drama", "Adventure", "Action", "Comedy", "Comedy", "Horror",~
## $ year     <dbl> 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1~
## $ released <chr> "June 13, 1980 (United States)", "July 2, 1980 (United States~
## $ score    <dbl> 8.4, 5.8, 8.7, 7.7, 7.3, 6.4, 7.9, 8.2, 6.8, 7.0, 6.1, 7.3, 5~
```

```
## $ votes    <dbl> 927000, 65000, 1200000, 221000, 108000, 123000, 188000, 33000~
## $ director <chr> "Stanley Kubrick", "Randal Kleiser", "Irvin Kershner", "Jim A~
## $ writer   <chr> "Stephen King", "Henry De Vere Stacpoole", "Leigh Brackett", ~
## $ star     <chr> "Jack Nicholson", "Brooke Shields", "Mark Hamill", "Robert Ha~
## $ country  <chr> "United Kingdom", "United States", "United States", "United S~
## $ budget   <dbl> 1.9e+07, 4.5e+06, 1.8e+07, 3.5e+06, 6.0e+06, 5.5e+05, 2.7e+07~
## $ gross    <dbl> 46998772, 58853106, 538375067, 83453539, 39846344, 39754601, ~
## $ company  <chr> "Warner Bros.", "Columbia Pictures", "Lucasfilm", "Paramount ~
## $ runtime  <dbl> 146, 104, 124, 88, 98, 95, 133, 129, 127, 100, 116, 109, 114,~
```

The "released" column and "year" column seem to differ for certain movies. Need to extract out only the "year" portion of the "released" column for consistency. Also arrange data frame by highest grossing movies

```
movies_df <- movies_df_original %>%
  mutate(released_year = (str_extract(released,"\\d{4}"))) %>%
  arrange(desc(gross))
```

Check for any duplicate rows

```
movies_df %>%
  get_dupes(name, rating, genre, released)
```

```
## No duplicate combinations found of: name, rating, genre, released
```

```
## # A tibble: 0 x 17
## # ... with 17 variables: name <chr>, rating <chr>, genre <chr>, released <chr>,
## #   dupe_count <int>, year <dbl>, score <dbl>, votes <dbl>, director <chr>,
## #   writer <chr>, star <chr>, country <chr>, budget <dbl>, gross <dbl>,
## #   company <chr>, runtime <dbl>, released_year <chr>
```

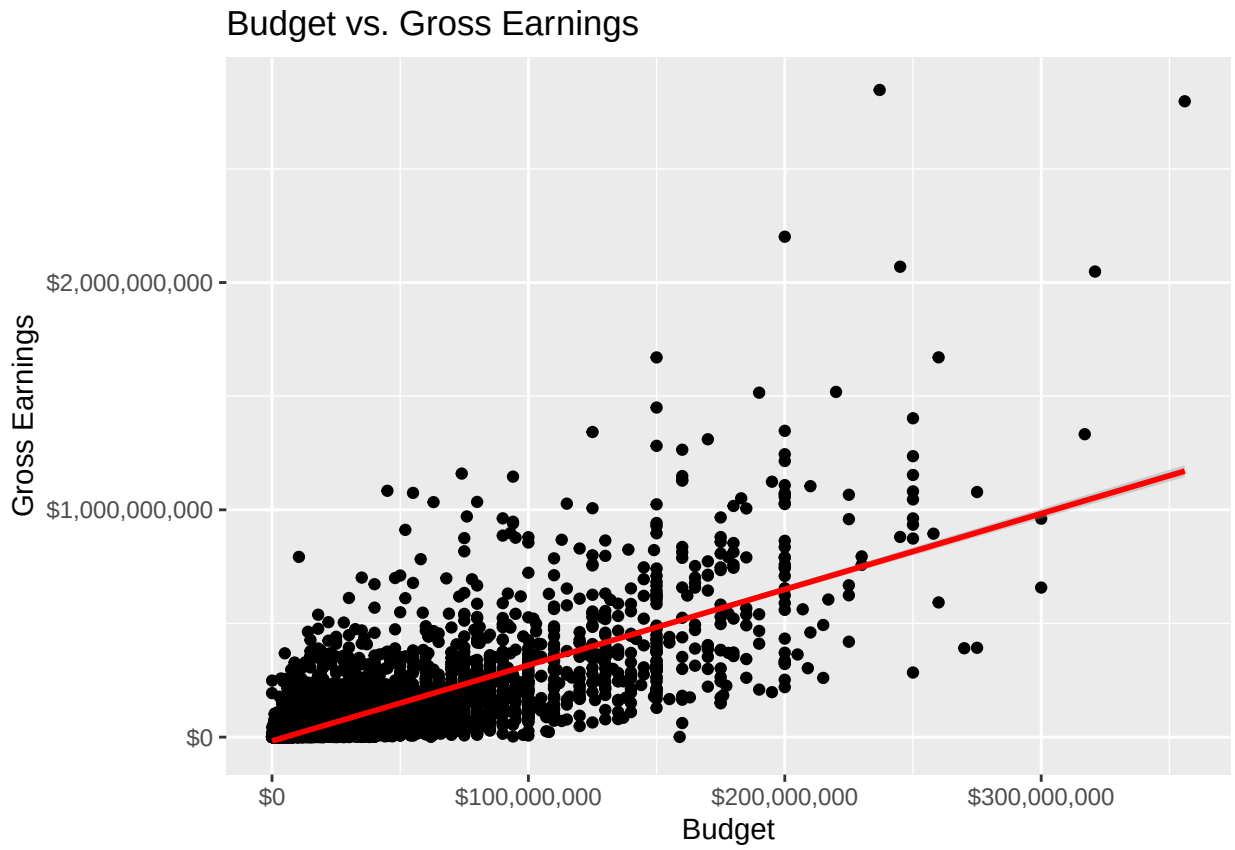
Make scatterplot of numeric variables to see potential correlation with gross revenue

```
ggplot(movies_df, aes(x = budget, y = gross)) +
  geom_point() +
  geom_smooth(method = lm, color = "red") +
  ggtitle("Budget vs. Gross Earnings") +
  labs(x = "Budget", y = "Gross Earnings") +
  scale_x_continuous(labels = scales::dollar_format()) +
  scale_y_continuous(labels = scales::dollar_format())
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2232 rows containing non-finite values (stat_smooth).
```

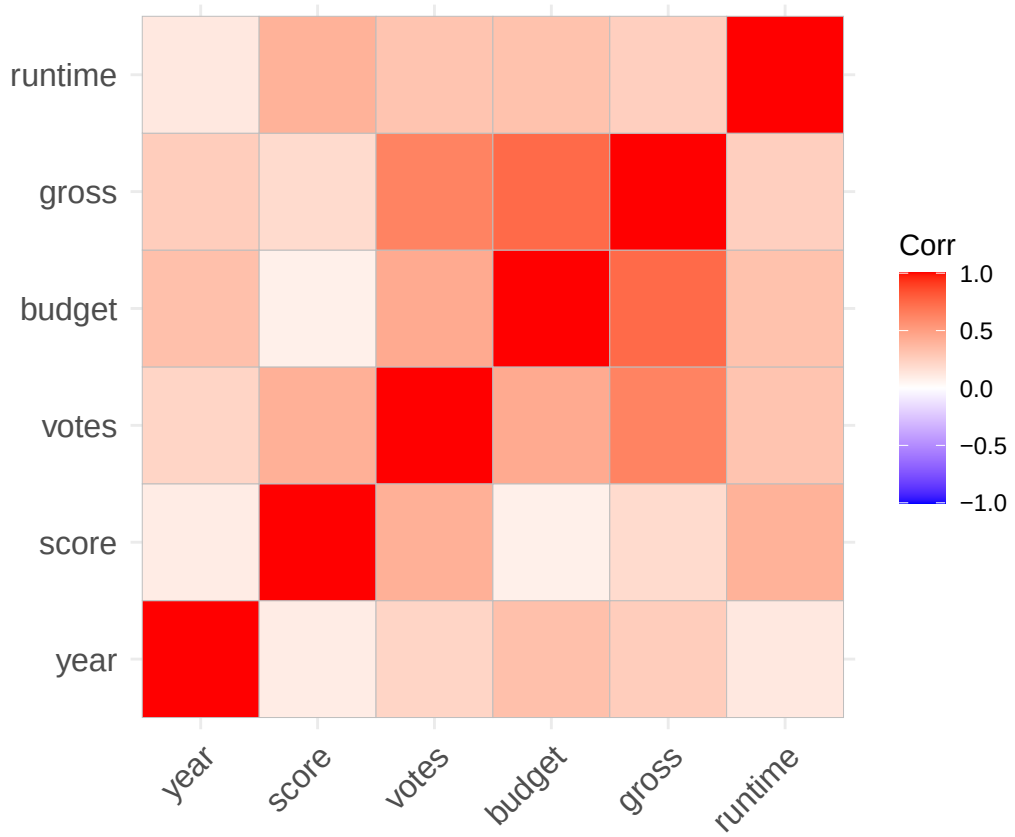
```
## Warning: Removed 2232 rows containing missing values (geom_point).
```



Calculate correlation of numerical variables and show as heat map. It appears that budget and votes have the highest correlation to gross earnings.

```
correlation_matrix <-  cor(movies_df[sapply(movies_df, is.numeric)], use = "pairwise.complete.obs")

ggcorrplot(correlation_matrix)
```


runtime
gross
budget
votes
score
year
year
score
votes
budget
gross
runtime
Corr
1.0
0.5
0.0
−0.5
−1.0