Correlation_R

Olamide Emida

2022-07-19

correlation with R

Data set Name : Movies.csv
Data source: Click here

Tasks

1. Do budgets on the movies affect the revenue generated from the movies?
2. Do movies' scores affect the revenue generated from the movies?
3. Do movies' rating affect the revenue generated from the movies?
4. What other relationships can be shown?

Setting up my R environment by loading the following libraries

```
library(tidyverse)
```

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

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

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

```
library(dplyr)
library(ggplot2)
library(here)
```

```
## here() starts at C:/Users/HP/Documents/Python Practices
```

```
library(ggcorrplot)
require(scales)
```

```
## Loading required package: scales
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor
```

Import movies dataset

```
movies_df = 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.
```

Scrutinizing the data

View the data

```
View(movies_df)
```

The data type of each column

```
glimpse(movies_df)
```

```
## 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 shape of the data

```
dim(movies_df)
```

```
## [1] 7668   15
```

The 15 columns of the data

```
colnames(movies_df)
```

```
##  [1] "name"     "rating"   "genre"    "year"     "released" "score"
##  [7] "votes"    "director" "writer"   "star"     "country"  "budget"
## [13] "gross"    "company"  "runtime"
```

Data Cleaning

remove rows with null values

rows with null values

```
movies_df %>%
  filter(if_any(everything(),is.na))
```

```
## # A tibble: 2,247 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 Fame    R      Drama  1980 May 16,~   6.6 21000 Alan Pa~ Chris~ Eddi~ United~
##  2 Stir C~ R      Come~  1980 Decembe~   6.8 26000 Sidney ~ Bruce~ Gene~ United~
##  3 Urban ~ PG     Drama  1980 June 6,~   6.4 14000 James B~ Aaron~ John~ United~
##  4 Altere~ R      Horr~  1980 Decembe~   6.9 33000 Ken Rus~ Paddy~ Will~ United~
##  5 Little~ R      Come~  1980 March 2~   6.5  5100 Ron Max~ Kimi ~ Tatu~ United~
##  6 Raise ~ PG     Acti~  1980 August ~   5     4100 Jerry J~ Adam ~ Jaso~ United~
##  7 My Bod~ PG     Come~  1980 Septemb~   7.1  8900 Tony Bi~ Alan ~ Chri~ United~
##  8 Prom N~ R      Horr~  1980 July 18~   5.4 16000 Paul Ly~ Willi~ Lesl~ Canada
##  9 Smokey~ PG     Acti~  1980 August ~   5.3 15000 Hal Nee~ Hal N~ Burt~ United~
## 10 Seems ~ PG     Come~  1980 Decembe~   6.7  9100 Jay San~ Neil ~ Gold~ United~
## # ... with 2,237 more rows, and 4 more variables: budget <dbl>, gross <dbl>,
## #   company <chr>, runtime <dbl>
```

drop rows with null values and return shape of the new data

```
new_movies_df <- na.omit(movies_df)
dim(new_movies_df)
```

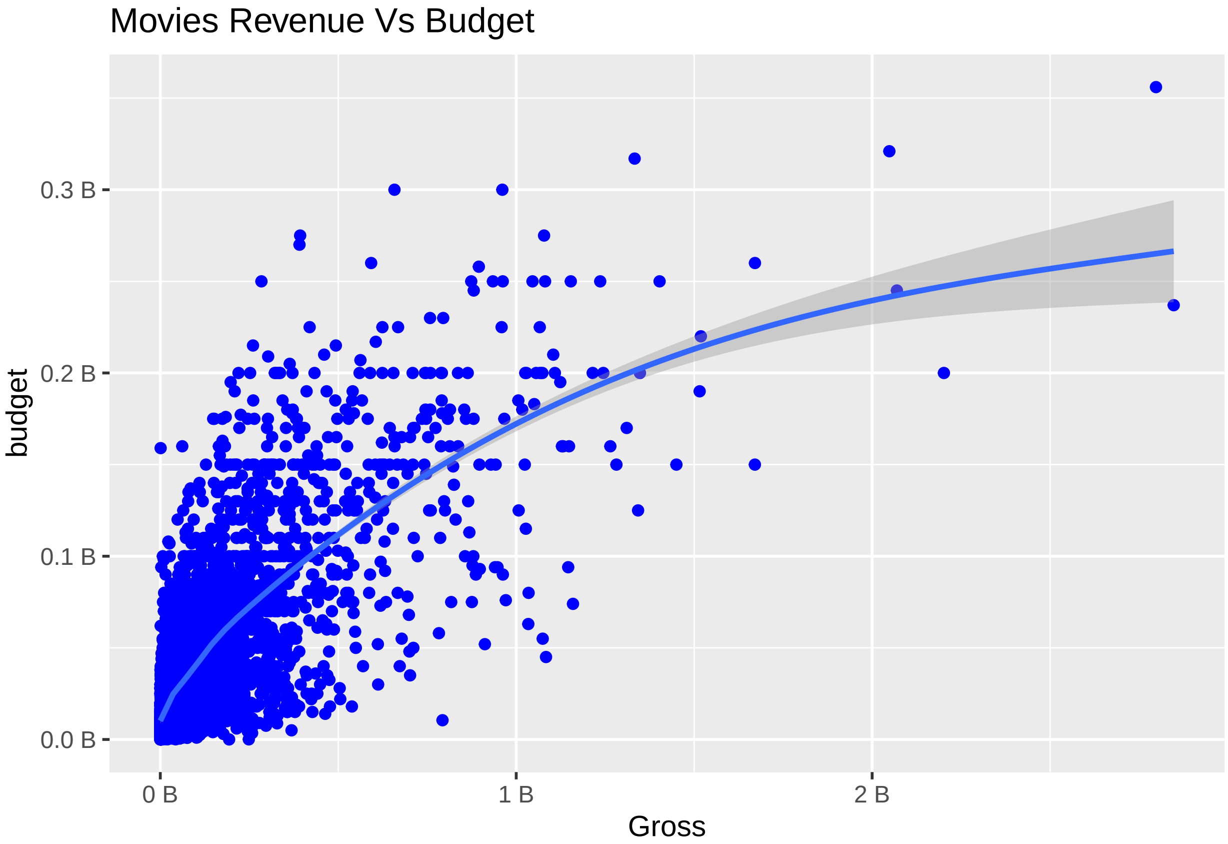
```
## [1] 5421   15
```

Data Visualizations

Task 1: Movies Budget Vs Revenue

```
ggplot(new_movies_df, aes(x = gross, y = budget))+
  geom_point(color = "Blue")+
  geom_smooth()+
  labs(title = "Movies Revenue Vs Budget", x = "Gross", Y = "Budget")+
  scale_x_continuous(labels = unit_format(unit = "B", scale = 1e-9))+
  scale_y_continuous(labels = unit_format(unit = "B", scale = 1e-9))
```

```
## ‘geom_smooth()‘ using method = ’gam’ and formula ’y ~ s(x, bs = "cs")’
```

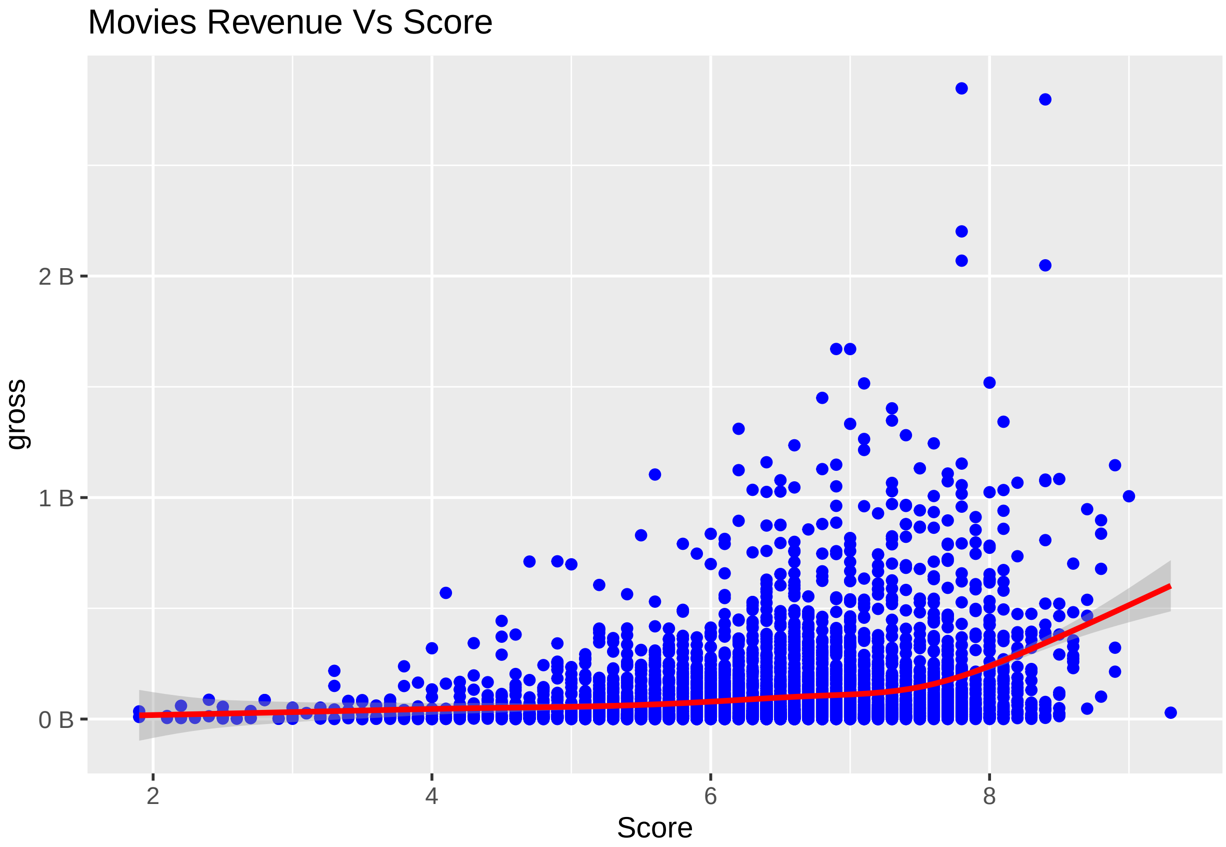


There is a positive relationship between movies revenue and budget: as budget increases, revenue also increases

Task 2: Movies Score Vs Revenue

```
ggplot(new_movies_df, aes(x = score, y = gross))+
  geom_point(color = "Blue")+
  geom_smooth(color = "Red")+
  labs(title = "Movies Revenue Vs Score", x = "Score", Y = "Gross")+
  scale_y_continuous(labels = unit_format(unit = "B", scale = 1e-9))
```

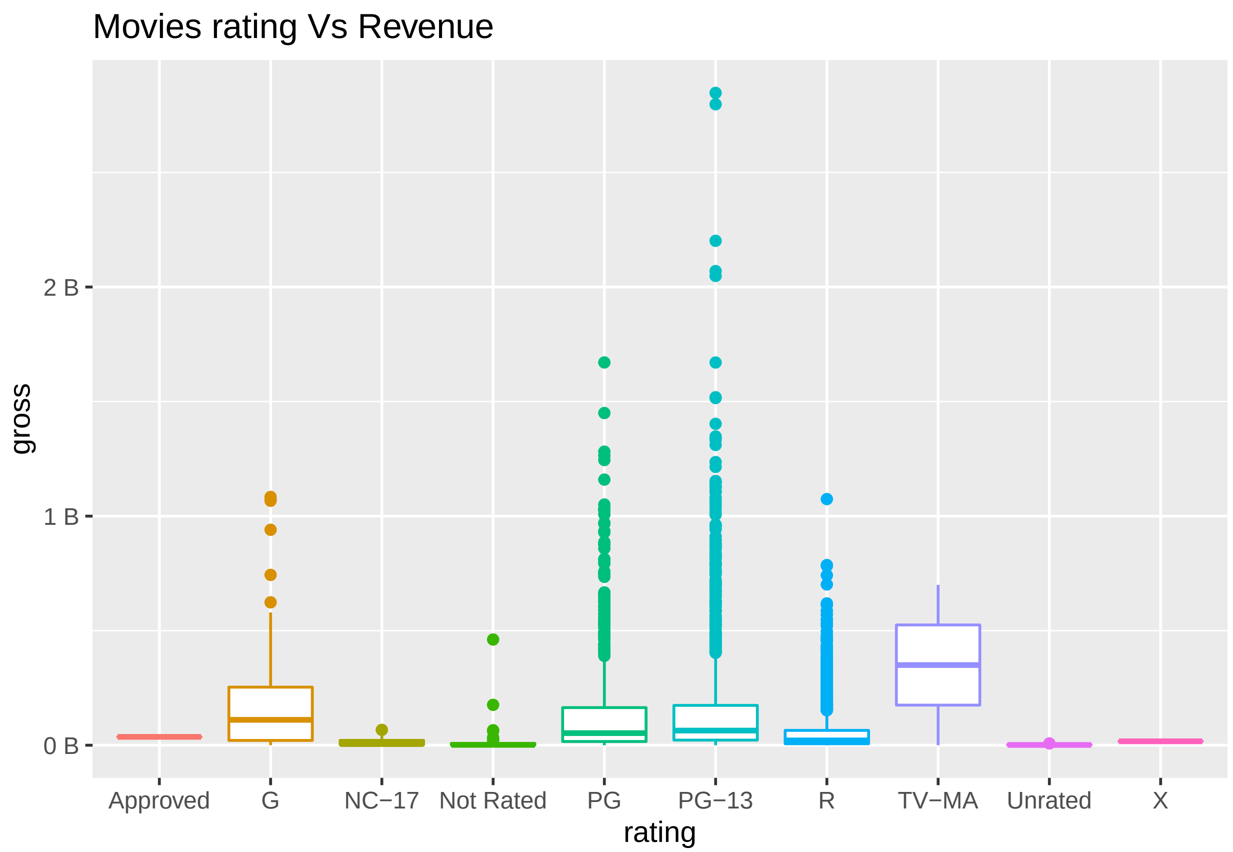
```
## ‘geom_smooth()‘ using method = ’gam’ and formula ’y ~ s(x, bs = "cs")’
```



There is a negative relationship between movies revenue and score

Task 3: Movies rating Vs Revenue

```
ggplot(data = new_movies_df)+
  geom_boxplot(mapping = aes(x = rating, y = gross, color = rating))+
  labs(title = 'Movies rating Vs Revenue')+
  theme(legend.position = "none")+
  scale_y_continuous(labels = unit_format(unit = "B", scale = 1e-9))
```



As median and max value for each category is close to 0, there is no relationship between movies revenue and rating

Task 4: Other relationships in the data

Create a new data set for numeric features

```
numeric_features = new_movies_df %>%
  select(year,score,votes,budget,gross,runtime)
head(numeric_features)
```

```
## # A tibble: 6 x 6
##    year score   votes   budget     gross runtime
##   <dbl> <dbl>   <dbl>    <dbl>     <dbl>   <dbl>
## 1  1980   8.4  927000 19000000  46998772     146
## 2  1980   5.8   65000  4500000  58853106     104
## 3  1980   8.7 1200000 18000000 538375067     124
## 4  1980   7.7  221000  3500000  83453539      88
## 5  1980   7.3  108000  6000000  39846344      98
## 6  1980   6.4  123000   550000  39754601      95
```

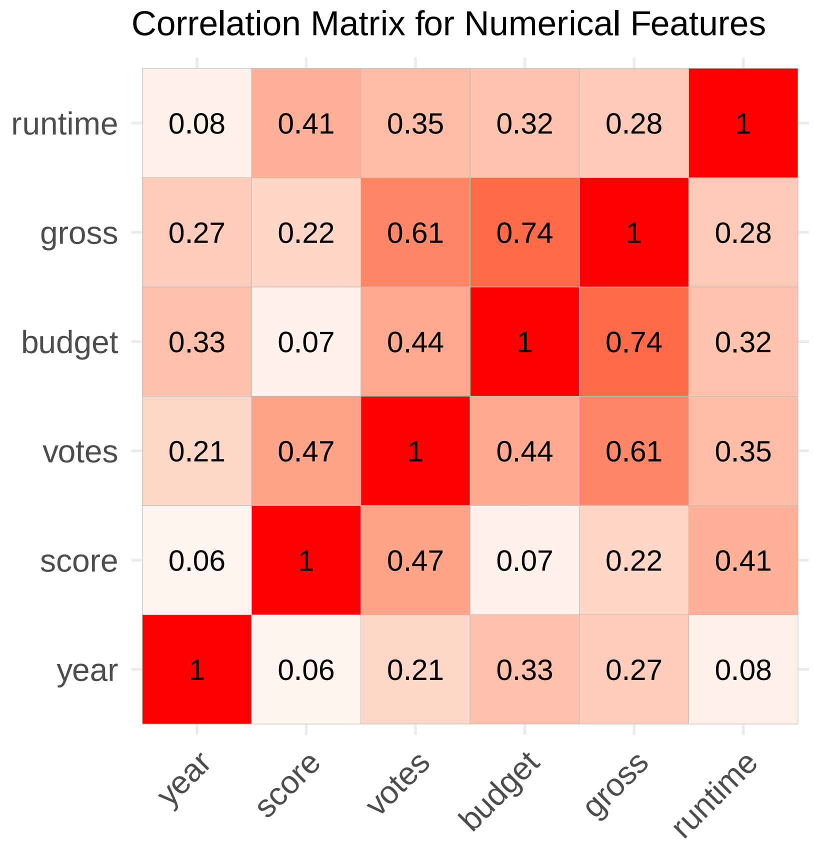
Correlation matrix for the numerical features

```
movie_correlation = round(cor(numeric_features),2)
head(movie_correlation)
```

```
##         year score votes budget gross runtime
## year    1.00  0.06  0.21   0.33  0.27    0.08
## score   0.06  1.00  0.47   0.07  0.22    0.41
## votes   0.21  0.47  1.00   0.44  0.61    0.35
## budget  0.33  0.07  0.44   1.00  0.74    0.32
## gross   0.27  0.22  0.61   0.74  1.00    0.28
## runtime 0.08  0.41  0.35   0.32  0.28    1.00
```

Showing the correlation as a heatmap

```
ggcorrplot(movie_correlation, lab = TRUE, show.legend = FALSE, title = "Correlation Matrix for Numerical
```



Conclusion

- There is a positive relationship between movies revenue and budget: as budget increases, revenue also increases
- There is a negative relationship between movies revenue and score.
- There is no relationship between movies revenue and rating