ML_B_ex1

itamar_living in dream, Yishay is donkey

29 3 2022

here some intrduction

import the dataset to a local variable:

```
surveys <- read_csv("netflix-rotten-tomatoes-metacritic-imdb.csv")
```

```
## Rows: 15480 Columns: 29
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (21): Title, Genre, Tags, Languages, Series or Movie, Country Availabil...
## dbl   (7): Hidden Gem Score, IMDb Score, Rotten Tomatoes Score, Metacritic S...
## date  (1): Netflix Release Date
##
## 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.
```

Some information is already known to us just from calling the table. We know how many rows and columns there are, and we have some idea regarding the type of variables (text, numeric) but not much else.

A few commands to help us learn more about the structure of the data are shown here:

```
dim(surveys)
```

```
## [1] 15480    29
```

```
nrow(surveys)
```

```
## [1] 15480
```

```
ncol(surveys)
```

```
## [1] 29
```

```
head(surveys)
```

```
## # A tibble: 6 x 29
##   Title Genre Tags  Languages ‘Series or Mov~‘ ‘Hidden Gem Sc~‘ ‘Country Avail~‘
##   <chr> <chr> <chr> <chr>     <chr>                       <dbl> <chr>
## 1 Lets~ Crim~ Come~ Swedish,~ Series                        4.3 Thailand
## 2 HOW ~ Come~ Dram~ English   Movie                         7   Canada
## 3 Cent~ Dram~ Thri~ English   Movie                         6.4 Canada
## 4 ANNE+ Drama TV D~ Turkish   Series                        7.7 Belgium,Netherl~
## 5 Moxie Anim~ Soci~ English   Movie                         8.1 Lithuania,Polan~
## 6 The ~ Come~ Roma~ Thai      Movie                         8.6 Thailand
## # ... with 22 more variables: Runtime <chr>, Director <chr>, Writer <chr>,
## #   Actors <chr>, ‘View Rating‘ <chr>, ‘IMDb Score‘ <dbl>,
## #   ‘Rotten Tomatoes Score‘ <dbl>, ‘Metacritic Score‘ <dbl>,
## #   ‘Awards Received‘ <dbl>, ‘Awards Nominated For‘ <dbl>, Boxoffice <chr>,
## #   ‘Release Date‘ <chr>, ‘Netflix Release Date‘ <date>,
## #   ‘Production House‘ <chr>, ‘Netflix Link‘ <chr>, ‘IMDb Link‘ <chr>,
## #   Summary <chr>, ‘IMDb Votes‘ <dbl>, Image <chr>, Poster <chr>, ...
```

```
tail(surveys)
```

```
## # A tibble: 6 x 29
##   Title Genre Tags  Languages ‘Series or Mov~‘ ‘Hidden Gem Sc~‘ ‘Country Avail~‘
##   <chr> <chr> <chr> <chr>     <chr>                      <dbl> <chr>
## 1 Nijn~ <NA>  Kids~ <NA>      Series                       NA   Belgium,Netherl~
## 2 K-PO~ <NA>  TV D~ <NA>      Series                       NA   South Korea,Arg~
## 3 Drea~ <NA>  Anim~ <NA>      Series                       NA   Russia,Hong Kon~
## 4 Drea~ Anim~ TV C~ English   Series                        8.4 Belgium,Switzer~
## 5 Drea~ Anim~ TV C~ English   Series                        8.2 Belgium,Switzer~
## 6 Drea~ Anim~ TV C~ English   Series                        8.1 Belgium,Switzer~
## # ... with 22 more variables: Runtime <chr>, Director <chr>, Writer <chr>,
## #   Actors <chr>, ‘View Rating‘ <chr>, ‘IMDb Score‘ <dbl>,
## #   ‘Rotten Tomatoes Score‘ <dbl>, ‘Metacritic Score‘ <dbl>,
## #   ‘Awards Received‘ <dbl>, ‘Awards Nominated For‘ <dbl>, Boxoffice <chr>,
## #   ‘Release Date‘ <chr>, ‘Netflix Release Date‘ <date>,
## #   ‘Production House‘ <chr>, ‘Netflix Link‘ <chr>, ‘IMDb Link‘ <chr>,
## #   Summary <chr>, ‘IMDb Votes‘ <dbl>, Image <chr>, Poster <chr>, ...
```

```
names(surveys)
```

```
##  [1] "Title"                  "Genre"                  "Tags"
##  [4] "Languages"              "Series or Movie"        "Hidden Gem Score"
##  [7] "Country Availability"   "Runtime"                "Director"
## [10] "Writer"                 "Actors"                 "View Rating"
## [13] "IMDb Score"             "Rotten Tomatoes Score"  "Metacritic Score"
## [16] "Awards Received"        "Awards Nominated For"   "Boxoffice"
## [19] "Release Date"           "Netflix Release Date"   "Production House"
## [22] "Netflix Link"           "IMDb Link"              "Summary"
## [25] "IMDb Votes"             "Image"                  "Poster"
## [28] "TMDb Trailer"           "Trailer Site"
```

```
str(surveys)
```

```
## spec_tbl_df [15,480 x 29] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Title                : chr [1:15480] "Lets Fight Ghost" "HOW TO BUILD A GIRL" "Centigrade" "ANNE+
##  $ Genre                : chr [1:15480] "Crime, Drama, Fantasy, Horror, Romance" "Comedy" "Drama, Thi
##  $ Tags                 : chr [1:15480] "Comedy Programmes,Romantic TV Comedies,Horror Programmes,Tha
##  $ Languages            : chr [1:15480] "Swedish, Spanish" "English" "English" "Turkish" ...
##  $ Series or Movie      : chr [1:15480] "Series" "Movie" "Movie" "Series" ...
##  $ Hidden Gem Score     : num [1:15480] 4.3 7 6.4 7.7 8.1 8.6 8.7 6.9 8.3 5.3 ...
##  $ Country Availability : chr [1:15480] "Thailand" "Canada" "Canada" "Belgium,Netherlands" ...
##  $ Runtime              : chr [1:15480] "< 30 minutes" "1-2 hour" "1-2 hour" "< 30 minutes" ...
##  $ Director             : chr [1:15480] "Tomas Alfredson" "Coky Giedroyc" "Brendan Walsh" NA ...
##  $ Writer               : chr [1:15480] "John Ajvide Lindqvist" "Caitlin Moran" "Brendan Walsh, Dale
##  $ Actors               : chr [1:15480] "Kåre Hedebrant, Per Ragnar, Lina Leandersson, Henrik Dahl"
##  $ View Rating          : chr [1:15480] "R" "R" "Unrated" NA ...
##  $ IMDb Score           : num [1:15480] 7.9 5.8 4.3 6.5 6.3 7.4 7.5 3.9 6.7 6.6 ...
##  $ Rotten Tomatoes Score: num [1:15480] 98 79 NA NA NA NA NA NA NA NA ...
##  $ Metacritic Score     : num [1:15480] 82 69 46 NA NA NA NA NA NA NA ...
##  $ Awards Received      : num [1:15480] 74 1 NA 1 NA NA 2 NA 2 NA ...
##  $ Awards Nominated For : num [1:15480] 57 NA NA NA 4 NA 4 NA 1 NA ...
##  $ Boxoffice            : chr [1:15480] "$2,122,065" "$70,632" "$16,263" NA ...
##  $ Release Date         : chr [1:15480] "12 Dec 2008" "08 May 2020" "28 Aug 2020" "01 Oct 2016" ...
##  $ Netflix Release Date : Date[1:15480], format: "2021-03-04" "2021-03-04" ...
##  $ Production House     : chr [1:15480] "Canal+, Sandrew Metronome" "Film 4, Monumental Pictures, Li
##  $ Netflix Link         : chr [1:15480] "https://www.netflix.com/watch/81415947" "https://www.netfli
##  $ IMDb Link            : chr [1:15480] "https://www.imdb.com/title/tt1139797" "https://www.imdb.com/
##  $ Summary              : chr [1:15480] "A med student with a supernatural gift tries to cash in on
##  $ IMDb Votes           : num [1:15480] 205926 2838 1720 1147 63 ...
##  $ Image                : chr [1:15480] "https://occ-0-4708-64.1.nflxso.net/dnm/api/v6/evlCitJPPCVCr
##  $ Poster               : chr [1:15480] "https://m.media-amazon.com/images/M/MV5BOWM4NTY2NTMtZDZlZSO
##  $ TMDb Trailer         : chr [1:15480] NA "https://www.youtube.com/watch?v=eIbcxPy4okQ" "https://www
##  $ Trailer Site         : chr [1:15480] NA "YouTube" "YouTube" NA ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Title = col_character(),
##   ..   Genre = col_character(),
##   ..   Tags = col_character(),
##   ..   Languages = col_character(),
##   ..   `Series or Movie` = col_character(),
##   ..   `Hidden Gem Score` = col_double(),
##   ..   `Country Availability` = col_character(),
##   ..   Runtime = col_character(),
##   ..   Director = col_character(),
##   ..   Writer = col_character(),
##   ..   Actors = col_character(),
##   ..   `View Rating` = col_character(),
##   ..   `IMDb Score` = col_double(),
##   ..   `Rotten Tomatoes Score` = col_double(),
##   ..   `Metacritic Score` = col_double(),
##   ..   `Awards Received` = col_double(),
##   ..   `Awards Nominated For` = col_double(),
##   ..   Boxoffice = col_character(),
##   ..   `Release Date` = col_character(),
##   ..   `Netflix Release Date` = col_date(format = ""),
##   ..   `Production House` = col_character(),
##   ..   `Netflix Link` = col_character(),
##   ..   `IMDb Link` = col_character(),
##   ..   Summary = col_character(),
##   ..   `IMDb Votes` = col_double(),
##   ..   Image = col_character(),
##   ..   Poster = col_character(),
##   ..   ‘TMDb Trailer‘ = col_character(),
##   ..   ‘Trailer Site‘ = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
summary(surveys)
```

```
##     Title              Genre               Tags            Languages
##  Length:15480       Length:15480       Length:15480       Length:15480
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Series or Movie    Hidden Gem Score Country Availability   Runtime
##  Length:15480       Min.   :0.600    Length:15480         Length:15480
##  Class :character   1st Qu.:3.800    Class :character     Class :character
##  Mode  :character   Median :6.800    Mode  :character     Mode  :character
##                     Mean   :5.938
##                     3rd Qu.:7.900
##                     Max.   :9.800
##                     NA's   :2101
##    Director            Writer             Actors          View Rating
##  Length:15480       Length:15480       Length:15480       Length:15480
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    IMDb Score    Rotten Tomatoes Score Metacritic Score Awards Received
##  Min.   :1.000   Min.   :  0.00        Min.   :  5.00   Min.   :  1.000
##  1st Qu.:5.800   1st Qu.: 38.00        1st Qu.: 44.00   1st Qu.:  1.000
##  Median :6.600   Median : 64.00        Median : 57.00   Median :  3.000
##  Mean   :6.496   Mean   : 59.52        Mean   : 56.81   Mean   :  8.764
##  3rd Qu.:7.300   3rd Qu.: 83.00        3rd Qu.: 70.00   3rd Qu.:  8.000
##  Max.   :9.700   Max.   :100.00        Max.   :100.00   Max.   :300.000
##  NA's   :2099    NA's   :9098          NA's   :11144    NA's   :9405
##  Awards Nominated For  Boxoffice         Release Date
##  Min.   :  1.00       Length:15480       Length:15480
##  1st Qu.:  2.00       Class :character   Class :character
##  Median :  5.00       Mode  :character   Mode  :character
##  Mean   : 13.98
##  3rd Qu.: 12.00
##  Max.   :386.00
##  NA's   :7819
##  Netflix Release Date Production House   Netflix Link        IMDb Link
##  Min.   :2015-04-14   Length:15480       Length:15480       Length:15480
##  1st Qu.:2016-08-09   Class :character   Class :character   Class :character
##  Median :2018-10-05   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2018-05-18
##  3rd Qu.:2020-03-18
##  Max.   :2021-03-04
```

```
##
##    Summary            IMDb Votes          Image              Poster
##  Length:15480       Min.   :      5.0   Length:15480       Length:15480
##  Class :character   1st Qu.:    403.5   Class :character   Class :character
##  Mode  :character   Median :   2322.0   Mode  :character   Mode  :character
##                     Mean   :  42728.4
##                     3rd Qu.:  20890.5
##                     Max.   :2354197.0
##                     NA's   :2101
##  TMDb Trailer       Trailer Site
##  Length:15480       Length:15480
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

```
surveys_fixed <- surveys %>% separate(`Release Date`, c("day", "month", "year"), sep = " ", convert = TR
surveys_fixed$`Netflix Release Date` <- as.Date(surveys$`Netflix Release Date`)
surveys_fixed$Boxoffice = gsub("\\$", "", surveys_fixed$Boxoffice)
surveys_fixed$Boxoffice = gsub("\\,", "", surveys_fixed$Boxoffice)
surveys_fixed$Boxoffice = as.integer(surveys_fixed$Boxoffice)


surveys_selected <- select(surveys, where(is.numeric))
surveys_selected <- select(surveys, !starts_with("Link"))

surveys_selected <- surveys %>% filter( grepl("Action", Genre) & `Series or Movie` == "Movie")

surveys %>% group_by(Runtime) %>%  tally(sort = TRUE)
```

```
## # A tibble: 5 x 2
##   Runtime          n
##   <chr>        <int>
## 1 1-2 hour      9121
## 2 < 30 minutes  3996
## 3 > 2 hrs       2028
## 4 30-60 mins     334
## 5 <NA>             1
```

```
# surveys$Runtime[surveys$Runtime == "1-2 hour"] <- "1.5"
```

Data Variation

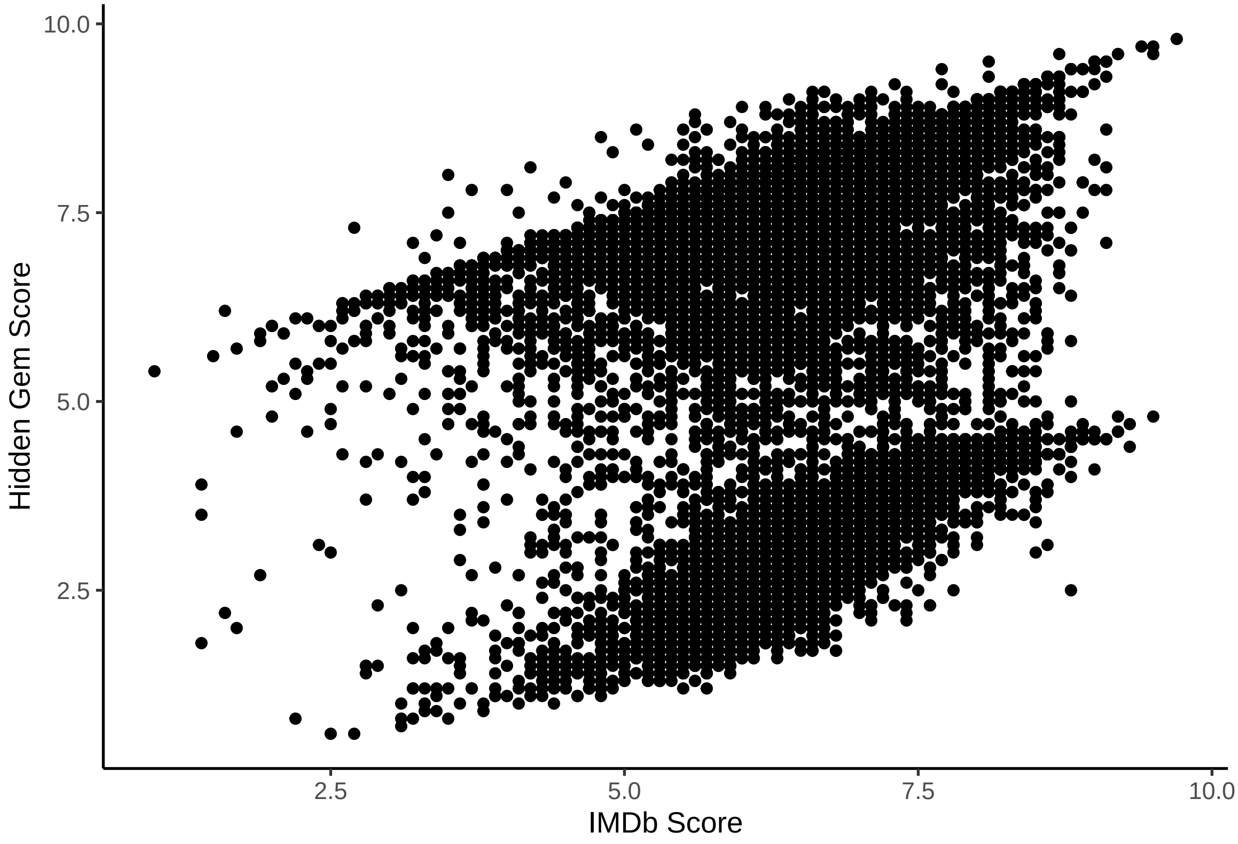
One important fact to know about our data is its variation, can you think of reasons why?

hidden gems

IMDb Score, Rotten Tomatoes Score, Metacritic Score

```
# plotting a scatter plot
ggplot(data = surveys,  aes(y = `Hidden Gem Score`, x= `IMDb Score`)) +
  geom_point() + theme_classic()
```

```
## Warning: Removed 2101 rows containing missing values (geom_point).
```



```
head(surveys$`Hidden Gem Score`)
```

```
## [1] 4.3 7.0 6.4 7.7 8.1 8.6
```

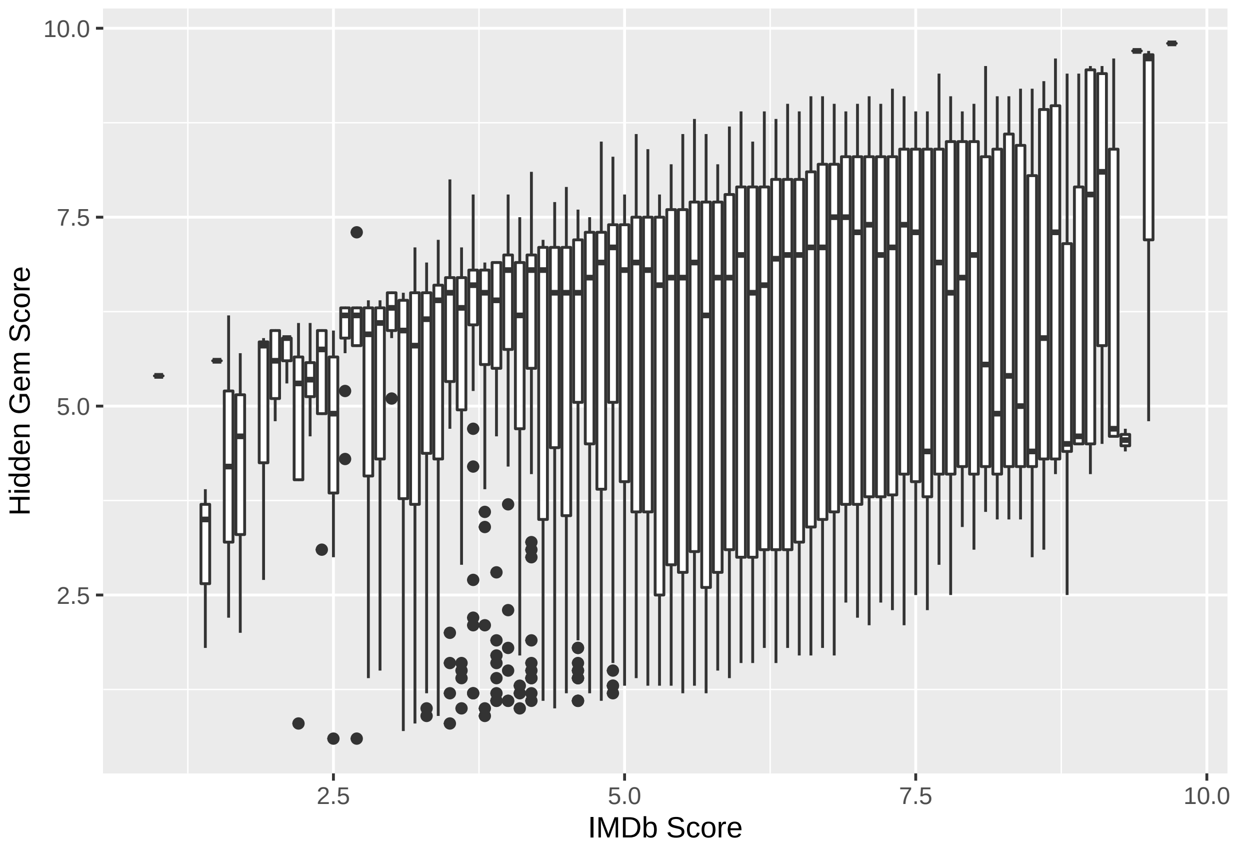
```
head(surveys$`IMDb Score`)
```

```
## [1] 7.9 5.8 4.3 6.5 6.3 7.4
```

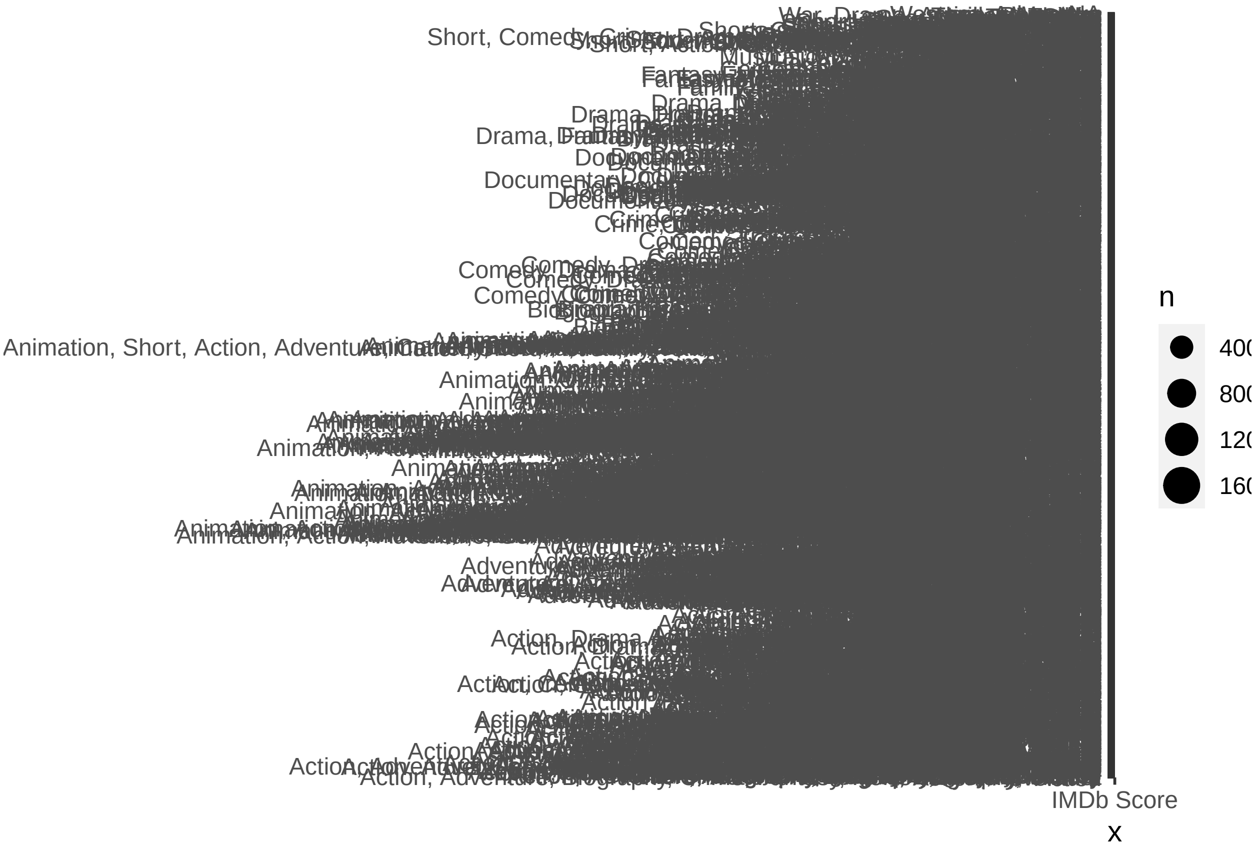
```
# creating boxplot
ggplot(data = surveys, mapping = aes(x = `IMDb Score`, y = `Hidden Gem Score`)) +
  geom_boxplot(mapping = aes(group = cut_width(`IMDb Score`, 0.1)))
```

```
## Warning: Removed 2099 rows containing missing values (stat_boxplot).
```

```
## Warning: Removed 2 rows containing non-finite values (stat_boxplot).
```

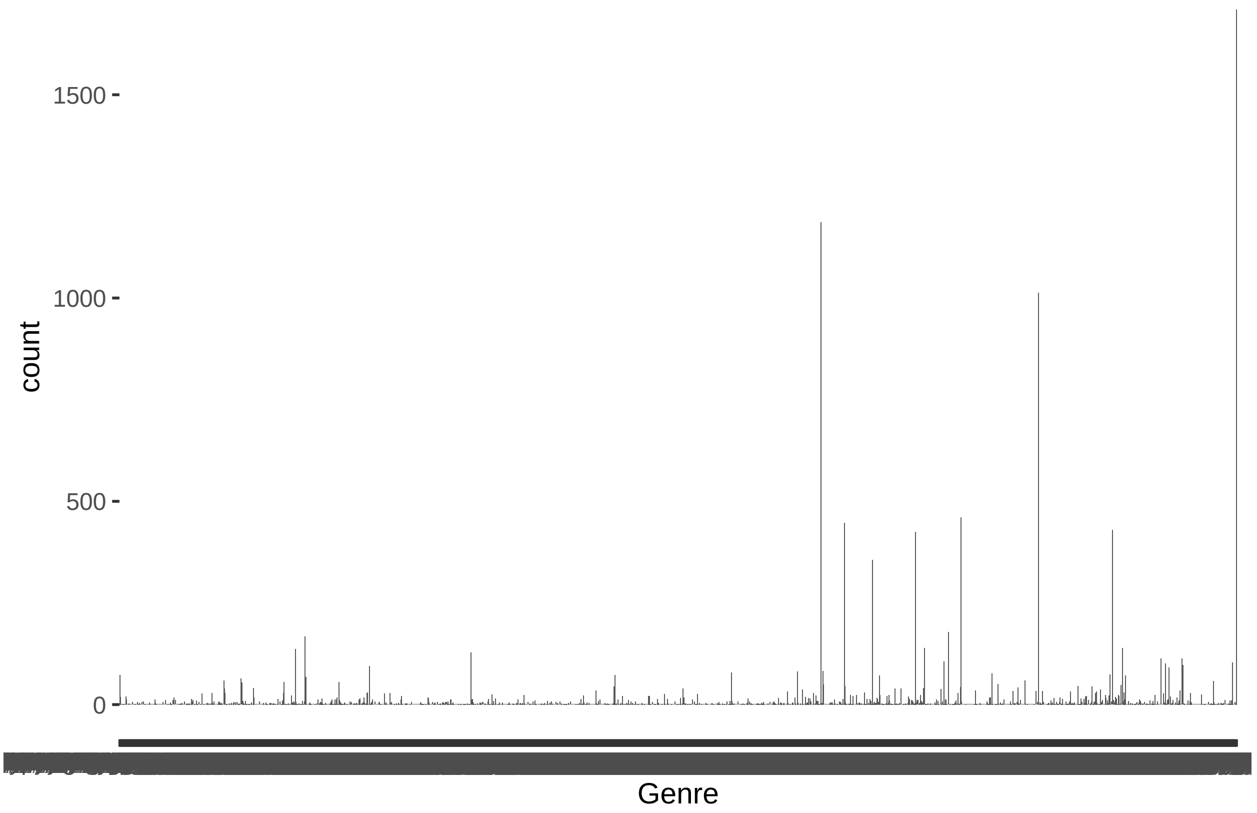


```
# Variance between two categorical variables
ggplot(data = surveys) +
  geom_count(mapping = aes(x = "IMDb Score", y = Genre))
```



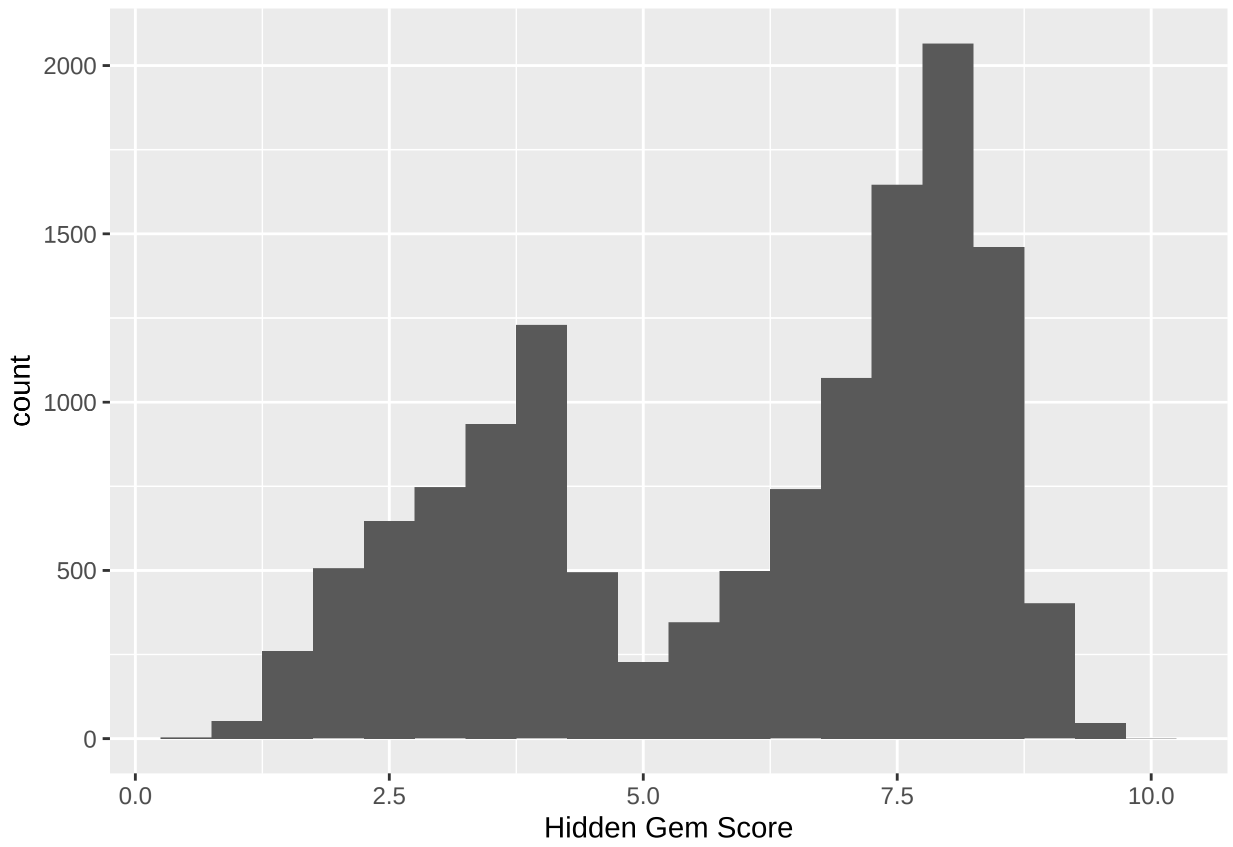
Data Visualisation - It is not working on this dataset!!

```
ggplot(data = surveys) +
  geom_bar(mapping = aes(x = Genre))
```

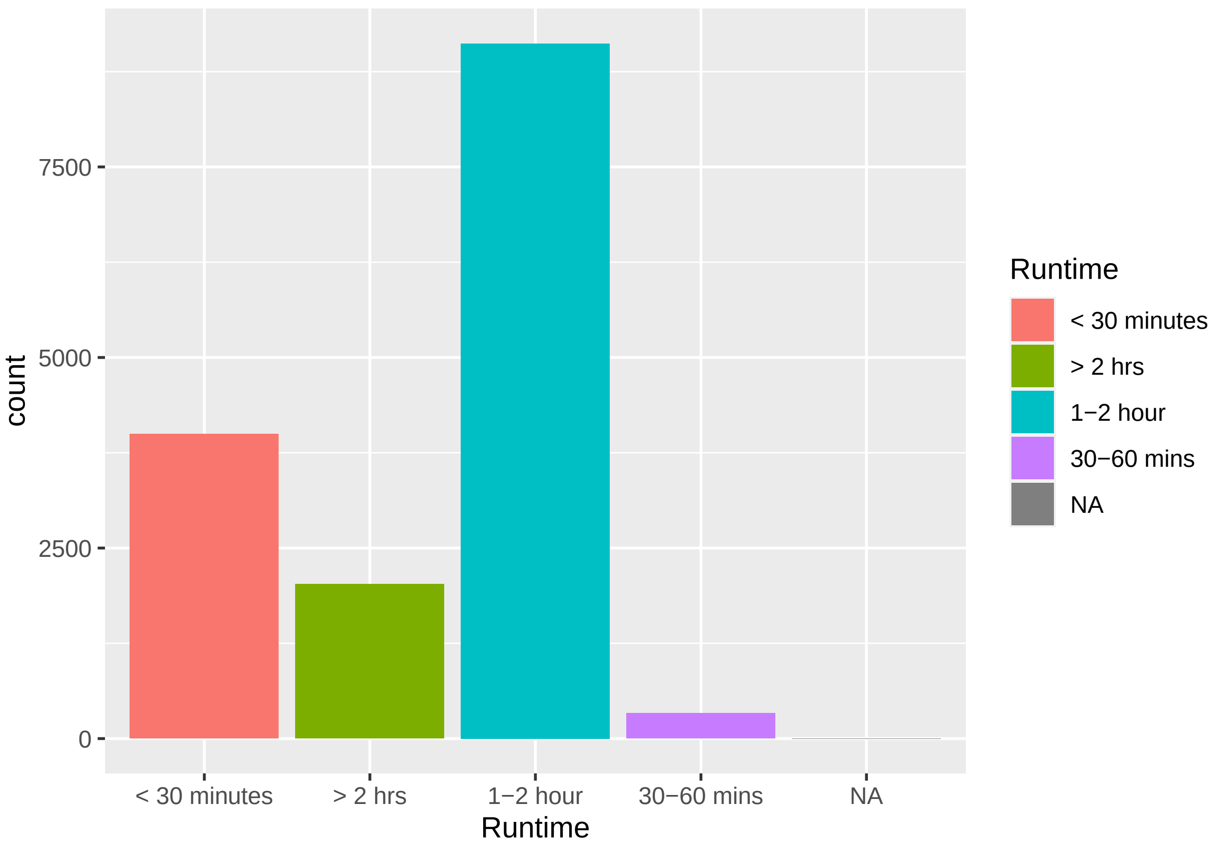


```
surveys$`Hidden Gem Score` <- as.numeric(surveys$`Hidden Gem Score`)
ggplot(data = surveys) +
  geom_histogram(mapping = aes(x = `Hidden Gem Score`), binwidth = 0.5)
```

```
## Warning: Removed 2101 rows containing non-finite values (stat_bin).
```



```r
ggplot(data = surveys) +
  geom_bar(mapping = aes(x = Runtime, fill = Runtime))
```


count
7500
5000
2500
0
< 30 minutes
> 2 hrs
1−2 hour
30−60 mins
NA
Runtime
Runtime
< 30 minutes
> 2 hrs
1−2 hour
30−60 mins
NA