Group8 Draft

2023-03-14

0.1 Summary

```
dataset8 <- read_csv("dataset8.csv")
```

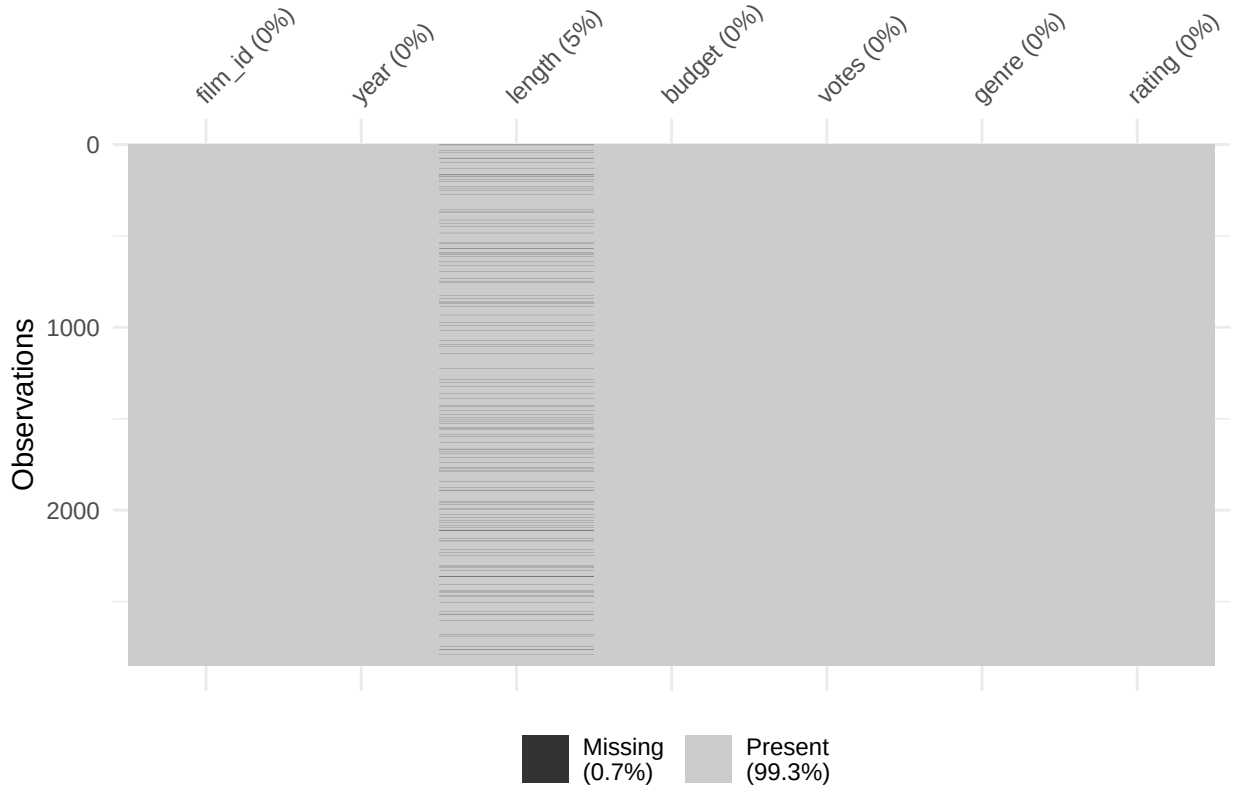
```
## Rows: 2847 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): genre
## dbl (6): film_id, year, length, budget, votes, rating
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
summary(dataset8)
```

```
##     film_id           year          length           budget
##  Min.   :   19   Min.   :1898   Min.   :  1.00   Min.   : 2.50
##  1st Qu.:14160   1st Qu.:1957   1st Qu.: 73.00   1st Qu.: 9.90
##  Median :29447   Median :1982   Median : 90.00   Median :11.90
##  Mean   :29181   Mean   :1976   Mean   : 82.22   Mean   :11.85
##  3rd Qu.:43982   3rd Qu.:1997   3rd Qu.:101.00   3rd Qu.:13.70
##  Max.   :58748   Max.   :2005   Max.   :480.00   Max.   :22.30
##                                 NA's   :131
##      votes           genre              rating
##  Min.   :     5   Length:2847        Min.   :0.800
##  1st Qu.:    11   Class :character   1st Qu.:3.700
##  Median :    29   Mode  :character   Median :4.600
##  Mean   :   657                      Mean   :5.342
##  3rd Qu.:   114                      3rd Qu.:7.700
##  Max.   :149494                      Max.   :9.200
##
```

0.2 Missing data

```
vis_miss(dataset8)
```



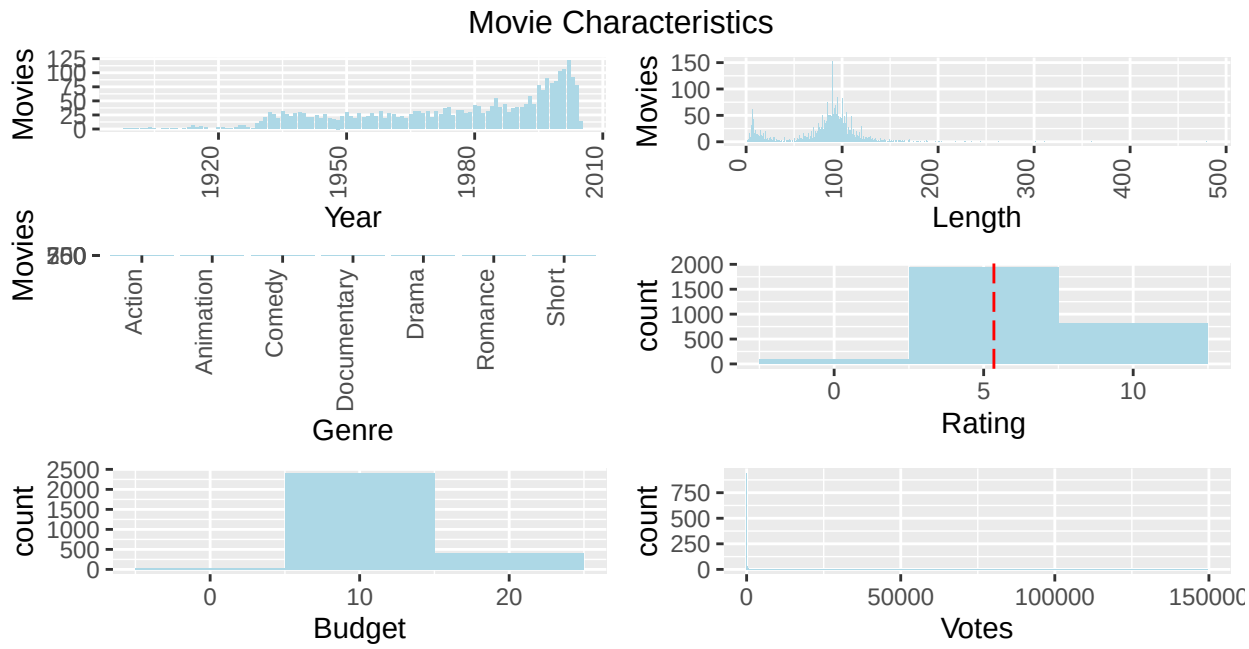
Plots

IMDB movie Characteristics

```
p1 <- ggplot(data=dataset8, aes(x=genre)) +
       geom_bar(fill="lightblue") +
       xlab("Genre") +
       ylab("Movies") +
       theme(axis.text.x=element_text(angle=90, hjust=1, vjust=0))
p2 <- ggplot(data=dataset8, aes(x=year)) +
       geom_bar(fill="lightblue") +
       xlab("Year") +
       ylab("Movies") +
       theme(axis.text.x=element_text(angle=90, hjust=1, vjust=0))
p3<- ggplot(data=dataset8, aes(x=length)) +
  geom_bar(fill="lightblue") +
  xlab("Length") +
  ylab("Movies") +
  theme(axis.text.x=element_text(angle=90, hjust=1, vjust=0))
p4 <- ggplot(data=dataset8, aes(x=rating)) +
  geom_histogram(binwidth=5, fill="lightblue") +
  geom_vline(xintercept=mean(dataset8$rating), colour='red', linetype='longdash') +
  geom_text(label='Mean', x=55, y=60, hjust='center', size=3) +
  xlab("Rating")
p5 <- ggplot(data=dataset8, aes(x=budget)) +
  geom_histogram(binwidth=10, fill="lightblue") +
  xlab("Budget")
```

```
p6 <- ggplot(data=dataset8, aes(x=votes)) +
  geom_histogram(binwidth=10, fill="lightblue") +
  xlab("Votes")
grid.arrange(p2, p3, p1, p4, p5, p6, nrow=4,
             top="Movie Characteristics")
```

```
## Warning: Removed 131 rows containing non-finite values (`stat_count()`).
```



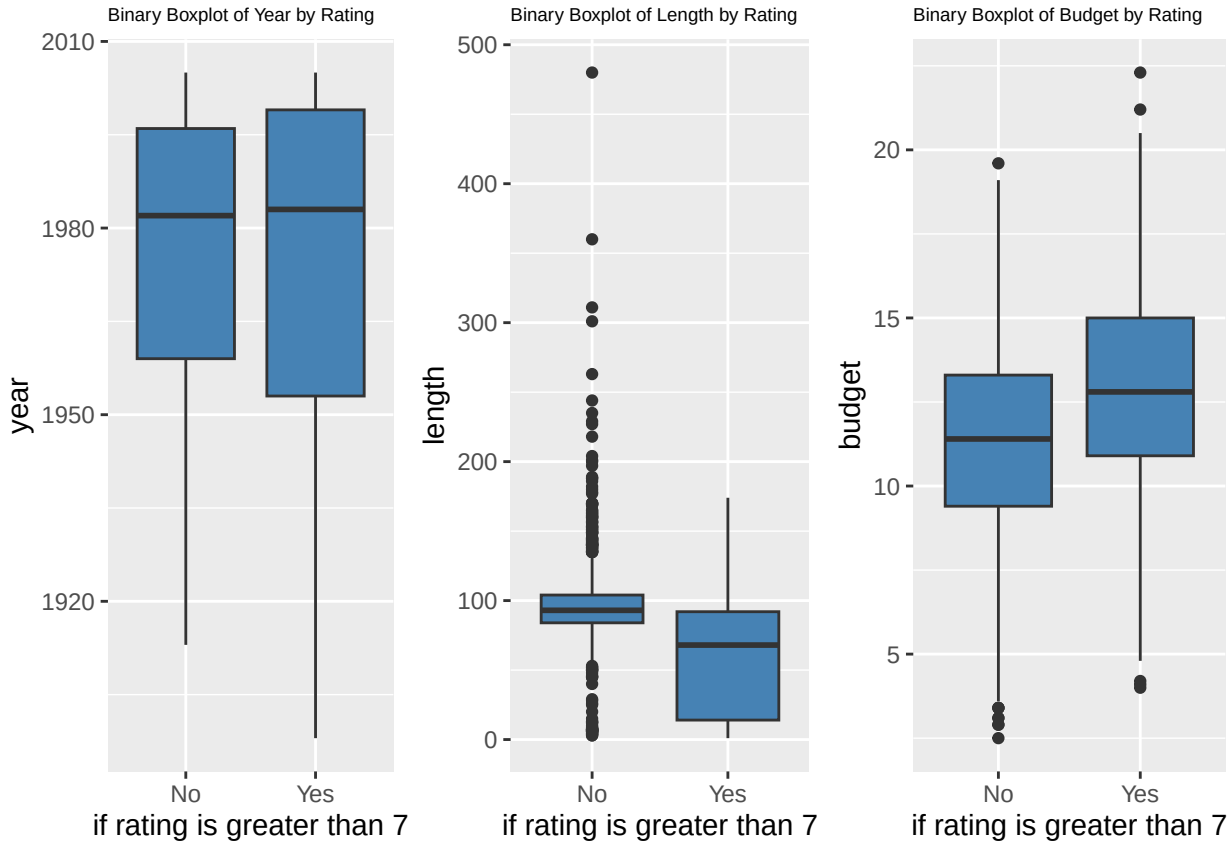
Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
imdb_group_08 <- dataset8 %>%
  mutate(Rating_above_7 = ifelse(rating > 7,"Yes","No"))
# Convert chr to factor
imdb_group_08$Rating_above_7 <- as.factor(imdb_group_08$Rating_above_7)
imdb_group_08$genre <- as.factor(imdb_group_08$genre)

p7 <- ggplot(data = imdb_group_08, mapping = aes(x = Rating_above_7, y = year)) +
  geom_boxplot(fill = "steelblue") +
  labs(x = "if rating is greater than 7", y = "year",
       title = "Binary Boxplot of Year by Rating")  +
  theme(plot.title = element_text(size=7))
p8 <- ggplot(data = imdb_group_08, mapping = aes(x = Rating_above_7, y = length)) +
  geom_boxplot(fill = "steelblue") +
  labs(x = "if rating is greater than 7", y = "length",
```

```
       title = "Binary Boxplot of Length by Rating")  +
  theme(plot.title = element_text(size=7))
p9 <- ggplot(data = imdb_group_08, mapping = aes(x = Rating_above_7, y = budget)) +
  geom_boxplot(fill = "steelblue") +
  labs(x = "if rating is greater than 7", y = "budget",
       title = "Binary Boxplot of Budget by Rating")  +
  theme(plot.title = element_text(size=7))
grid.arrange(p7, p8, p9, ncol=3)
```

```
## Warning: Removed 131 rows containing non-finite values (`stat_boxplot()`).
```



Creating an additional column that indicate yes or no if rating is greater than 7

##Modeling predictor included, by rating variable, using generalized liner model fit model if binomial, 0<y<1, the variabale rating not in.

```
model<-glm(rating~year+length+budget+votes,data=dataset8)
#summary(model)
```