Movie revenue analysis and visulisation

2023-10-25

#Visulisation of Ticket Revenue The objective of our data visualization project was to analyze and present the movie ticket revenue along with their respective genres. By visually representing this information, we aimed to uncover insights and patterns that would help stakeholders understand the revenue distribution across different movie genres.

The focus of this research is to explore the relationship between movie ticket revenue and movie genres. Understanding how ticket revenue varies across different genres is crucial for the film industry, as it can inform decision-making processes related to film financing, marketing strategies, and audience targeting. By analyzing a comprehensive dataset of movie ticket revenue, we aim to investigate whether certain genres consistently generate higher revenue compared to others. This research is essential to provide insights into the financial performance of different genres and help industry professionals make informed decisions regarding film production, distribution, and marketing efforts.

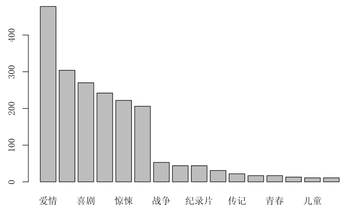
```
setwd("/Users/davidsmacbook/Desktop/PKU/R")
dat = read.csv("movies.csv", header = TRUE, fileEncoding = "GBK") #Read Data
par(family = "simsun")


dat$类型 = as.factor(dat$类型) #Convert to factor
summary(dat)
```

```
##     电影名称             总票房          类型
##  Length:1985        Min.   :     0   爱情   :478
##  Class :character   1st Qu.:   151   剧情   :304
##  Mode  :character   Median :   734   喜剧   :270
##                     Mean   : 10086   动画   :242
##                     3rd Qu.:  4752   惊悚   :222
##                     Max.   :567927   动作   :206
##                     NA's   :14       (Other):263
```

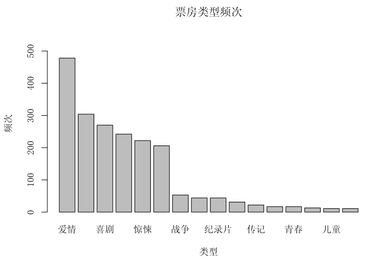
```
str(dat) # Basic Overview
```

```
## 'data.frame':    1985 obs. of  3 variables:
##  $ 电影名称: chr  "哪吒之魔童降世" "�浪浪地球" "我和我的祖国" "中国机长" ...
##  $ 总票房  : int  497286 465592 266138 253041 220293 171785 129731 101386 87235 79551 ...
##  $ 类型    : Factor w/ 16 levels "传记","儿童",..: 5 14 3 3 7 7 4 3 3 4 ...
```

```
# Bar Chart
freq = sort(table(dat$类型), decreasing = T)
barplot(freq)
```




```
# Solving display problem of the genre
par(family = "simsun")
bar = barplot(freq, ylab = "频次",ylim = c(0,550),xlab = "类型", main = "票房类型频次")
```

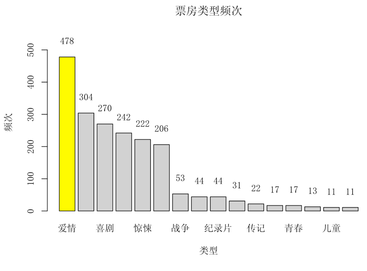



```
# Create the color vector with default color for all columns
colors <- rep("lightgray", length(freq))

# Specify the index of the column to color
col_index <- 1

# Set the color for the specified column
colors[col_index] <- "Yellow"
bar = barplot(freq, ylab = "频次",ylim = c(0,550),
              xlab = "类型", main = "票房类型频次",col = colors)
text(bar,freq+50,freq,)


#Spine Plot
median_revenue <- median(dat$总票房, na.rm = TRUE) # Calculate the median
print(median_revenue)
```

```
## [1] 734
```

```
# Create a new categorical variable based on median
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.3     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

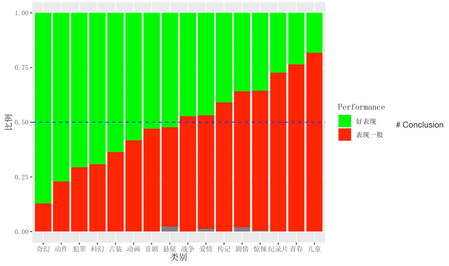



```
dat <- dat %>%
  mutate(performance = ifelse(总票房 > median_revenue, "好表现", "表现一般"))
head(dat)
```

	电影名称 <chr>	总票房 <int>	类型 <fct>	performance <chr>
1	哪吒之魔童降世	497286	动画	好表现
2	逆浪地球	465592	科幻	好表现
3	我和我的祖国	266138	剧情	好表现
4	中国机长	253041	剧情	好表现
5	疯狂的外星人	220293	喜剧	好表现
6	飞驰人生	171785	喜剧	好表现

6 rows

```
library(ggplot2)

# Spine plot
spine_plot <- ggplot(dat, aes(reorder(类型, -ifelse(performance == "好表现" & !is.na(performance), 1, 0)), fill = performance)) +
  geom_bar(position = "fill") +
  labs(title = "电影类别表现分布",
       x = "类别",
       y = "比例",
       fill = "Performance") +
  scale_fill_manual(values = c("好表现" = "green", "表现一般" = "red")) + theme(text = element_text(family = "simsun")) + geom_hline(yintercept = 0.5, linetype = "dashed", color = "blue")


# Display the spine plot
print(spine_plot)
```



The spine plot display the revenue performance of different genres. We can see just about half of the genres perform better. Surprisingly 奇幻 genre does extremly well comparing to other genres, while in 儿童, only about 25% of them perform well comparing to other genres.

Due to the limitation of the dataset, the graph only present the absolute performance by classify them solely based on their revenue. We do not include the cost of the production to have a fair analysis. Foreseeable, some higher cost genres result in higher absolute profit. Therefore, the graph is heavily influenced by the cost structure of different genres and their scale of investment.

In conclusion, the next step we do is to gather information about the cost structure of each genre. By doing so, we can understand which genre is the best in profitability. Moreover, we could also start analysis regarding rating of the movie and investigate the relationship between cost effective, popularity and market preference within different genres.