Data Exploration

R Markdown

Missing Values

```
##   release_date video_release_date imdb_url
## 1            9             100000       13
```

This suggests that we may be able to impute values for release_date, imdb_url and may have to remove all of the video_relase_date information.

Variable Features

Categorical Features (user_id, movie_id, rating, timestamp?,)

Distribution of Gender

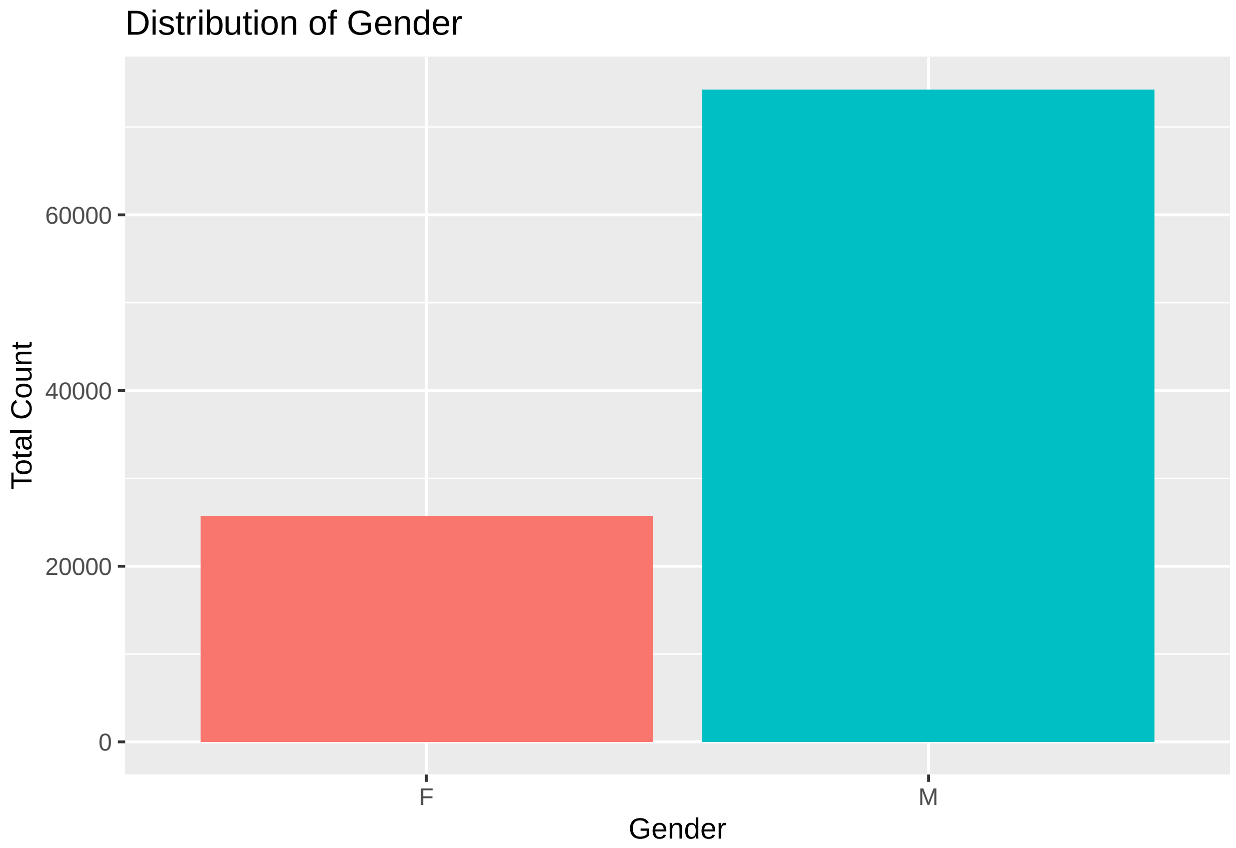
```r
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages -------------------------------------- tidyverse 1.3.0 --
```

```
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## v purrr   0.3.3
```

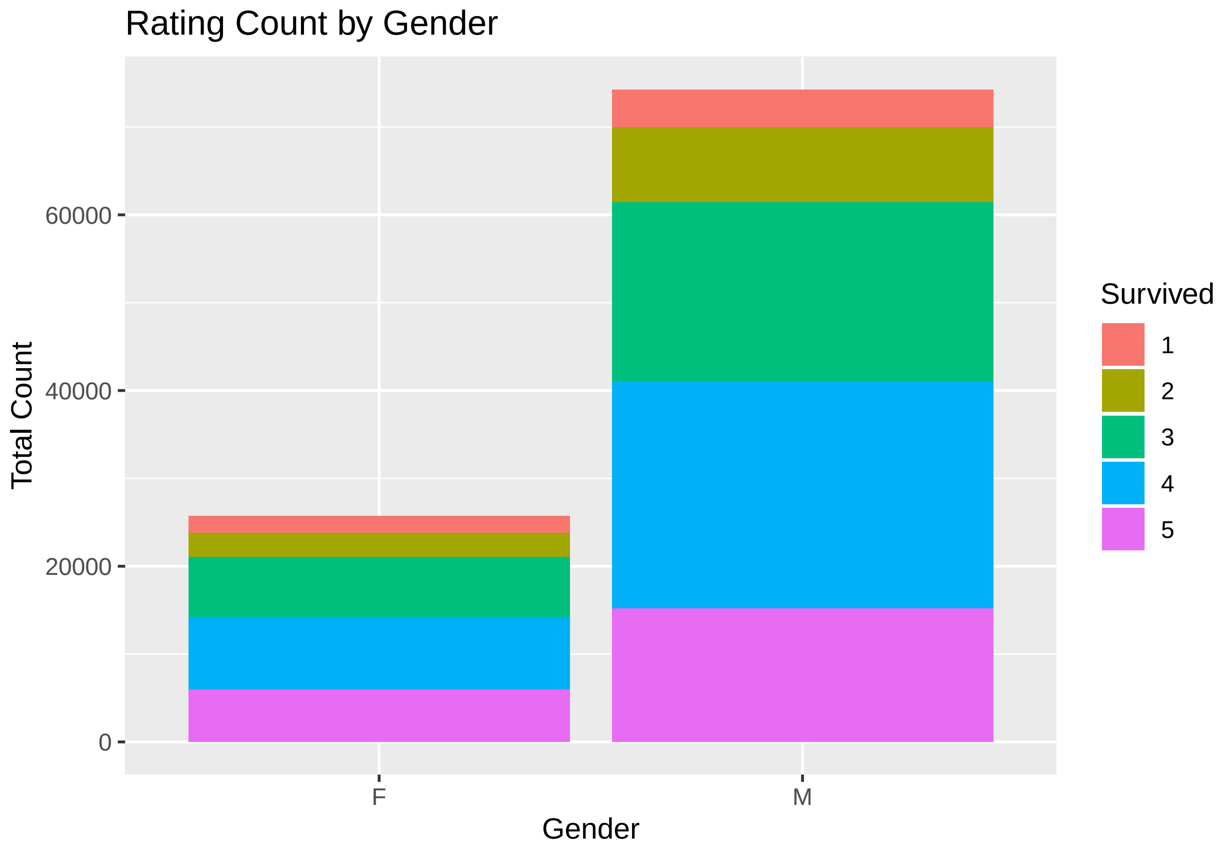
```
## -- Conflicts ----------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
# Distribution of gender
ggplot(ratings2, aes(x = factor(gender), fill = factor(gender))) +
  geom_bar( show.legend=FALSE) +
  xlab("Gender") +
  ylab("Total Count") +
  ggtitle("Distribution of Gender")
```



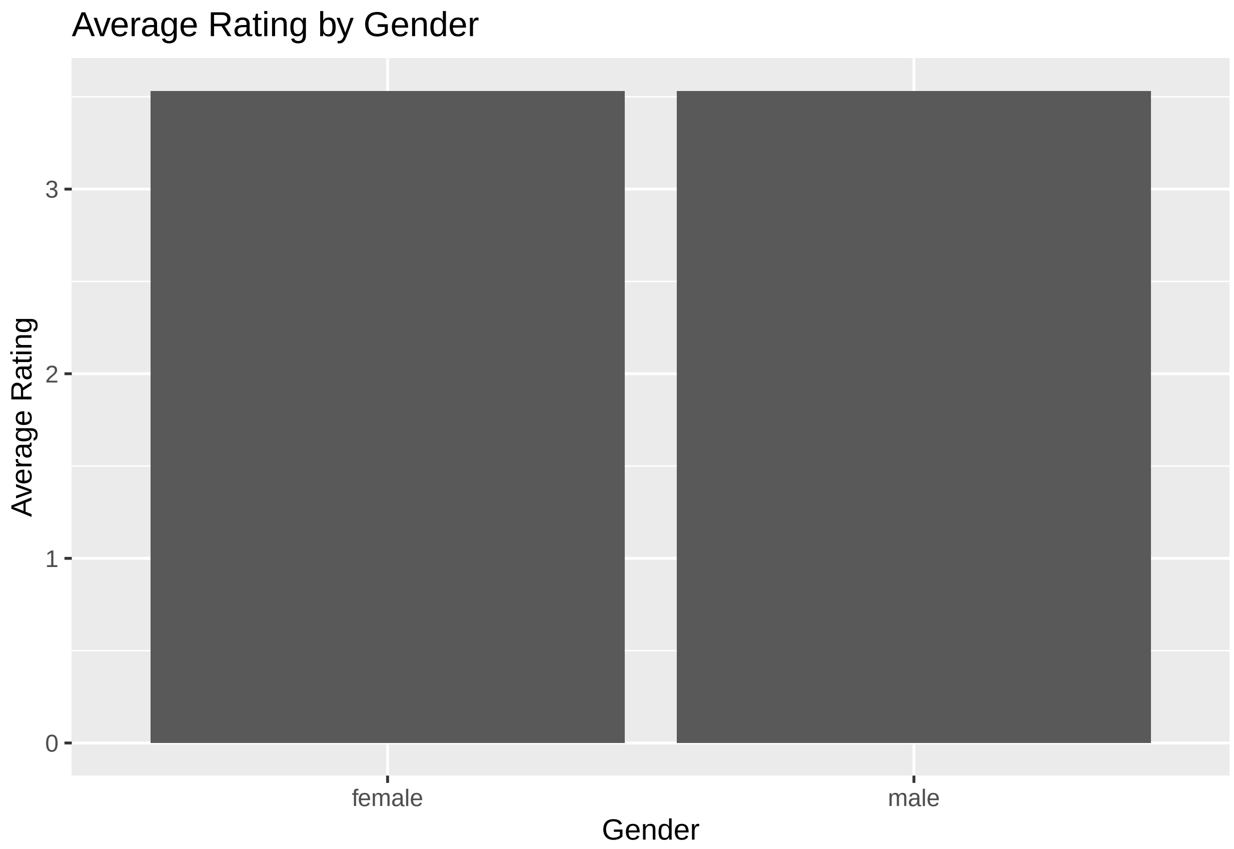
Ratings by Gender

```
# Ratings by Gender
ggplot(subset(ratings2, !is.na(gender)), aes(x = gender, fill = as.factor(rating))) +
  geom_bar() +
  ggtitle("Rating Count by Gender") +
  xlab("Gender") +
  ylab("Total Count") +
  labs(fill = "Survived")
```



Average Rating by Gender

```
male_mean <- ratings2 %>% filter(gender=='M') %>% pull(rating) %>% mean
female_mean <- ratings2 %>% filter(gender=='F') %>% pull(rating) %>% mean
mean_gender <- c(male_mean, female_mean)
gender <- c("male","female")
mean_gender_df <- data.frame(gender, mean_gender)
ggplot(mean_gender_df, aes(x=gender, y=mean_gender)) +
  geom_bar(stat="identity") +
  ggtitle("Average Rating by Gender") +
  xlab("Gender") +
  ylab("Average Rating")
```

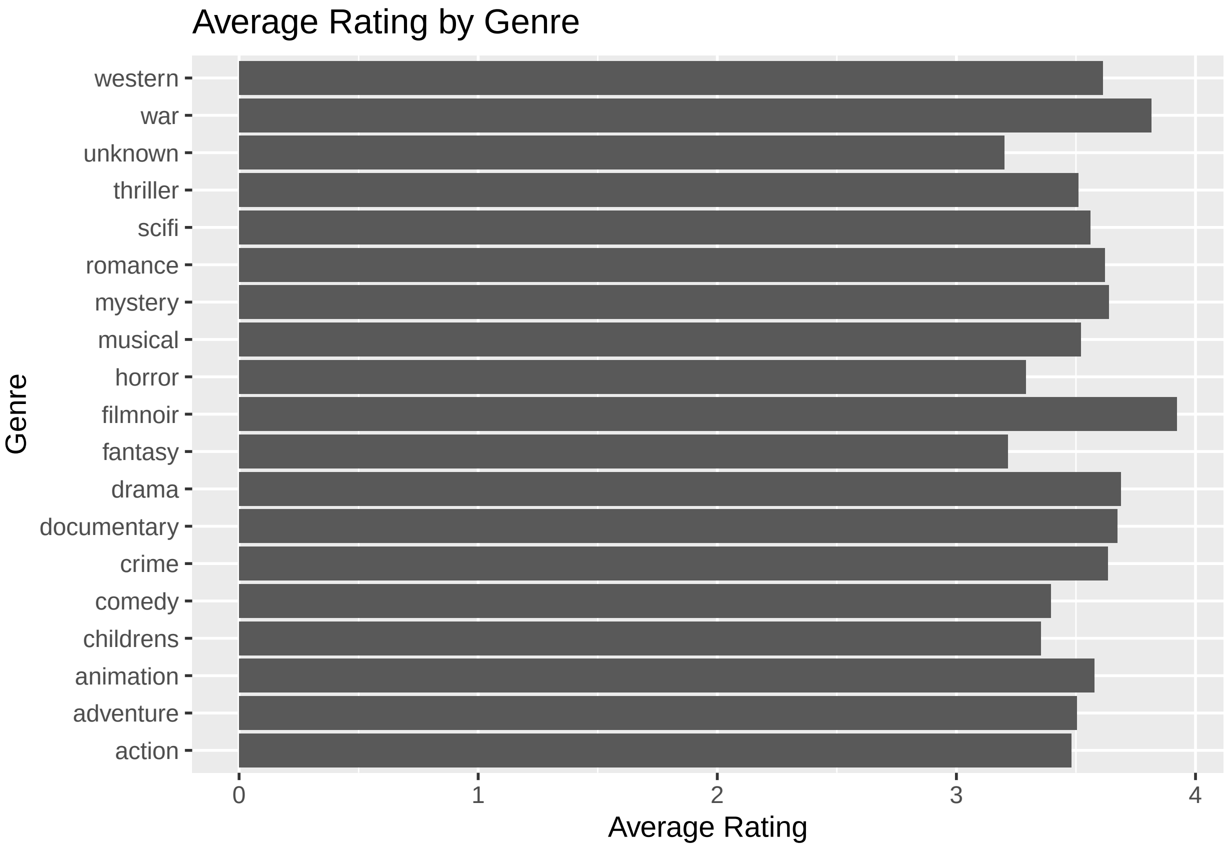


Average Rating by Genre

```
# ratings <- merge(ratings, movie_info, by.x='movie_id', by.y='movie_id',all.x=TRUE, all.y=TRUE)
# ratings <- merge(ratings, user_info, by.x='user_id', by.y='user_id',all.x=TRUE, all.y=TRUE)
# ratings2 <- ratings

# convert genres to factor
genres <- ratings2[,9:27]
for(i in 1:ncol(genres)) {
  genres[,i] <- as.factor(genres[,i])
}
ratings2[,9:27] <- genres
genres_rating <- cbind(genres, ratings2[,3])
colnames(genres_rating)[20] <- "rating"

# show average rating by genre
mean <- rep(0,ncol(genres_rating)-1)
for(i in 1:(ncol(genres_rating)-1)) {
  mean[i] <- genres_rating %>% filter(genres_rating[[i]] == 1) %>% pull(rating) %>% mean
}
genres <- names(genres)
df <- data.frame(genres, mean)
ggplot(df, aes(x=genres, y=mean)) +
  geom_bar(stat="identity") +
  coord_flip() +
  xlab("Genre") +
  ylab("Average Rating") +
  ggtitle("Average Rating by Genre")
```


Average Rating by Genre
Genre
western
war
unknown
thriller
scifi
romance
mystery
musical
horror
filmnoir
fantasy
drama
documentary
crime
comedy
childrens
animation
adventure
action
0
1
2
3
4
Average Rating