Movie Data Analysis

```
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
data <- read.csv("movie.csv")
```

Step 1

First we will check if the design is balanced or not and to be able to find this we need to first check the distribution of participants in different genres

```
Action_F <- sum(data$Gender == "F" & data$Genre == "Action")
Comedy_F <- sum(data$Gender == "F" & data$Genre == "Comedy")
Drama_F <- sum(data$Gender == "F" & data$Genre == "Drama")



print("Female:")
```

```
## [1] "Female:"
```

```
print(paste("Action=",Action_F,"Comedy=",Comedy_F,"Drama=",Drama_F))
```

```
## [1] "Action= 39 Comedy= 33 Drama= 22"
```

```
Action_M <- sum(data$Gender == "M" & data$Genre == "Action")
Comedy_M <- sum(data$Gender == "M" & data$Genre == "Comedy")
Drama_M <- sum(data$Gender == "M" & data$Genre == "Drama")



print("Male:")
```

```
## [1] "Male:"
```

```
print(paste("Action=",Action_M,"Comedy=",Comedy_M,"Drama=",Drama_M))
```

```
## [1] "Action= 14 Comedy= 10 Drama= 19"
```

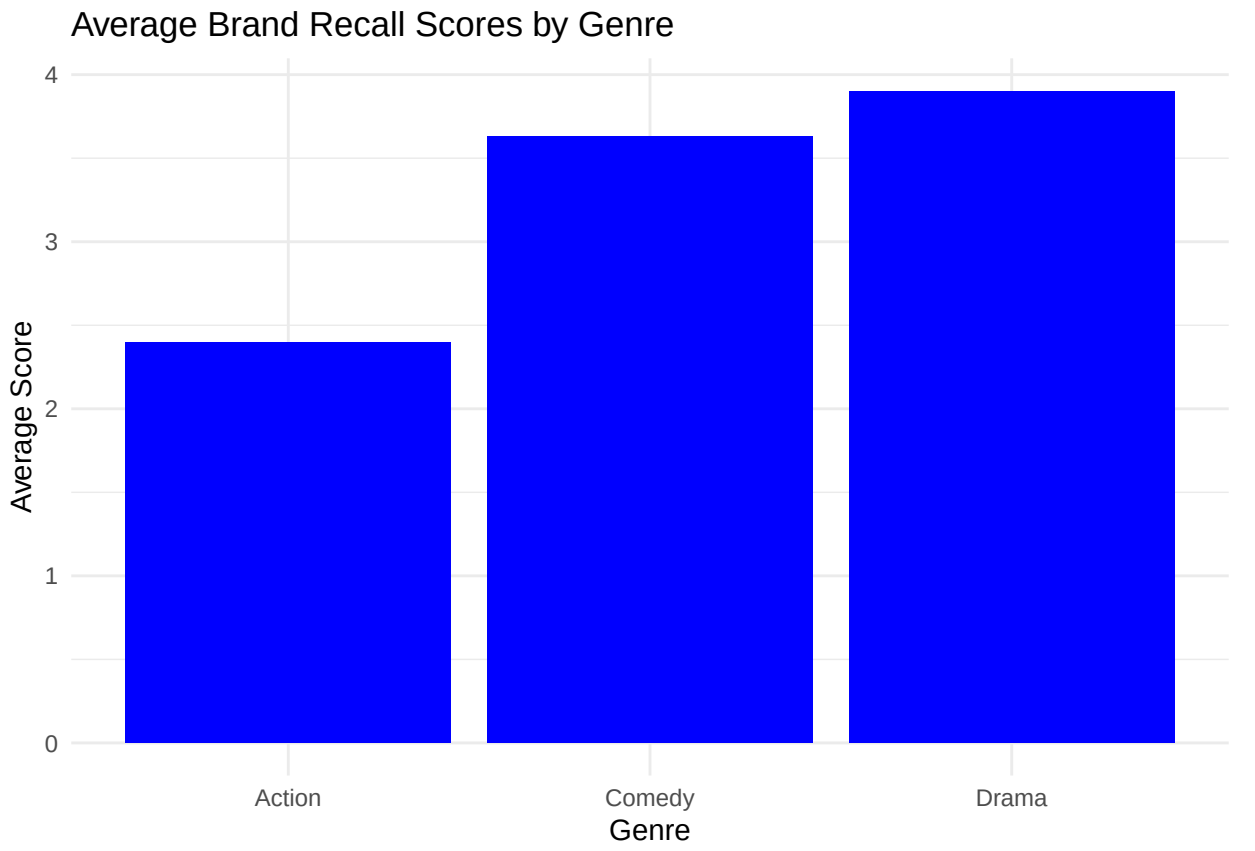
As it can be seen from the above numbers that the distribution of genres in not equal for both genders. As it can be seen there are more female participants in Action and Comedy Genre while in males there are more male participants in Drama genre so because of which the distribution is not balanced.

Step 2

Constructing two different preliminary graphs that investigate different features of the data.

First graph will be a bar chart

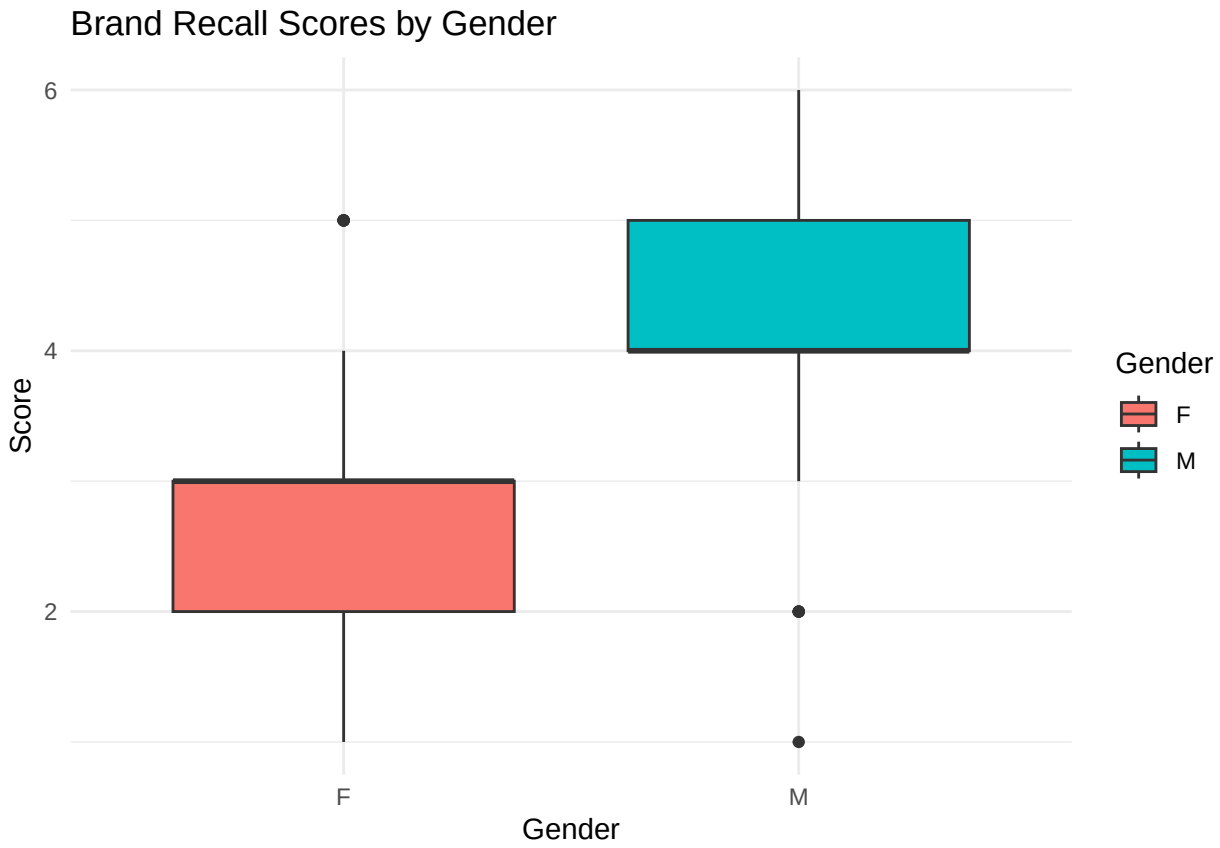
```
avg_scores <- aggregate(Score ~ Genre, data, mean)

ggplot(avg_scores, aes(x = Genre, y = Score)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(x = "Genre", y = "Average Score", title = "Average Brand Recall Scores by Genre") +
  theme_minimal()
```



As it can be seen in the graph that Drama and Comedy has comparatively higher average scores as compare to Action genre. This information can be helpful for the business and can guide them to select the genres with higher average which ultimately can be more effective for product placement.

Second plot can be a box plot where Score is distributed by Genre.

```
ggplot(data, aes(x = Gender, y = Score, fill = Gender)) +
  geom_boxplot() +
  labs(x = "Gender", y = "Score", title = "Brand Recall Scores by Gender") +
  theme_minimal()
```



The above graph helps us gain insights that whether the gender has any impact on the brand scores and it can ultimately help the business to change their advertising strategies depending on the target audience's gender.

Step 3

Mathematical model for this situation, with all appropriate parameters.

$$Y = \beta_0 + \beta_1 \cdot \text{Gender} + \beta_2 \cdot \text{Genre} + \beta_3 \cdot \text{Gender} \cdot \text{Genre} + \varepsilon$$

$$Y = \text{Brand Recall Score}$$
$$\beta_0 = \text{Intercept}$$
$$\beta_1 = \text{Effect of Gender}$$
$$\beta_2 = \text{Effect of Genre}$$
$$\beta_3 = \text{Combine effect of Gender and Genre}$$
$$\varepsilon = Error$$

Step 4

Now we analyze the data to study the effect of Gender and Genre on the brand recall Score. These conclusions are only required to be at the qualitative level and can be based off the outcomes of the hypothesis tests we conduct in this part and the preliminary plots in step 2. We do not need to statistically examine the multiple comparisons between contrasts and interactions.

Null Hypothesis

Null Hypothesis: No effect of Gender or Genre on the brand recall Score. Alternative Hypothesis: Effect of either Gender or Genre on the brand recall Score.

Analysis

Now we conduct two way Analysis of Variance(ANOVA) so that we can understand the effect of Gender and Genre on the brand recall Score.

```
model <- aov(Score ~ Gender + Genre + Gender:Genre, data = data)
summary(model)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## Gender          1  71.58   71.58  79.804 3.28e-15 ***
## Genre           2  50.36   25.18  28.070 7.15e-11 ***
## Gender:Genre    2  15.08    7.54   8.405 0.000368 ***
## Residuals     131 117.51    0.90
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As it can be seen from the above ANOVA table that Gender and Genre both have their p-values smaller than the typical significance level of 0.05 Gender has the p-value of 3.28e-15, Genre has the p-value of 7.15e-11 and p-value of the combined effect of Gender and Genre is also less than 0.05 which is 0.000368.We know that if the p-value is smaller then 0.05 then we can reject the null hypothesis because when p-value is small which suggests that it is very unlikely that the observed differences are due to some random chance. So based on this knowledge we can reject our null hypothesis that there is no effect of Gender or Genre on the brand recall Score and We can conclude that Gender and Genre both have effect on the brand recall score individually and combined as well.We can also suggest that the combined effect of the Genre and Gender is different from what was expected specially after looking at their individual effects.

Step 5

Conclusion

Based on the analysis we can say that the type of movie genre plays an important role in letting people remember and recognize a brand and specifically as we have seen in the preliminary plots as well that drama category had strong effect on the score. So if the businesses want people to remember their brand. They should consider putting their brand in dramatic movies. Because normally people get emotionally attached to the characters and story of drama movies. Due to this emotional connection the brand stands out and the chances of people remembering it and recognizing it later increases.

Also as we saw from the hypothesis testing that Gender also plays significant role on the brand recall Score. So the business also have to understand regarding the audience which they want to reach with their product. By knowing this they can see that which drama movies would be best suited for their product which eventually helps their brand to get in front of right people who will most likely remember and recognize it.