Hypothesis Testing

Data 101

Ryan Lee

Part A

Analyze Dataset

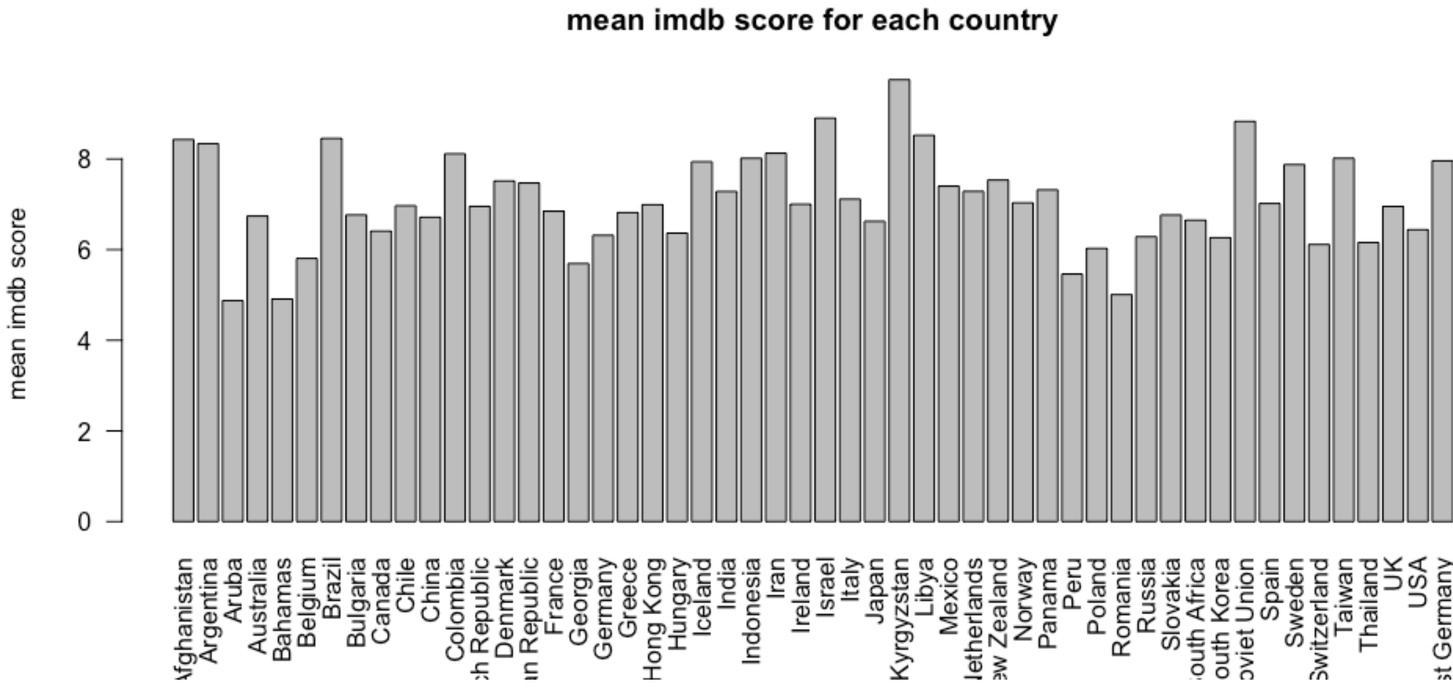
- I used the following code to analyze the data. This code allows you to check the average imdb score based on the categorical variable (country, content, Gross, Budget, genre)
- If there is a column that shows a significant difference compared to others, we would conduct hypothesis testing using them.

```
barplot(tapply(movies$imdb_score, movies$country,mean),ylab="mean imdb score",
        main="mean imdb score for each country",border="black",las=2)
tapply(movies$imdb_score,movies$content,mean)
tapply(movies$imdb_score,movies$Gross,mean)
tapply(movies$imdb_score,movies$Budget,mean)
tapply(movies$imdb_score,movies$genre,mean)
```

Part A

Data Analysis - Country

- We can see Kyrgyzstan and Israel's average score are higher than other countries.
- We can see Aruba, Bahamas and Romania's average score are lower than other countries.



Part A

Data Analysis - Other Variables

- In the content, R has highest average score and PG-13 has the lowest average score.
- In the Gross, Low Gross has highest average score and Medium Gross the lowest average score.
- In the Budget, Low Budget has highest average score and High Budget has lowest average score.
- In the genre, History has highest average score and Family has the lowest average score.

```
> tapply(movies$imdb_score,movies$content,mean)
       G       PG    PG-13        R
6.552577 6.301783 6.292623 6.795562
> tapply(movies$imdb_score,movies$Gross,mean)
    High      Low   Medium
6.552438 6.725589 6.347748
> tapply(movies$imdb_score,movies$Budget,mean)
    High      Low   Medium
6.136457 7.087669 6.439676
> tapply(movies$imdb_score,movies$genre,mean)
  Action   Comedy    Drama   Family  History   Sci-Fi
6.500382 6.426980 6.356172 5.908296 7.434874 6.341694
```

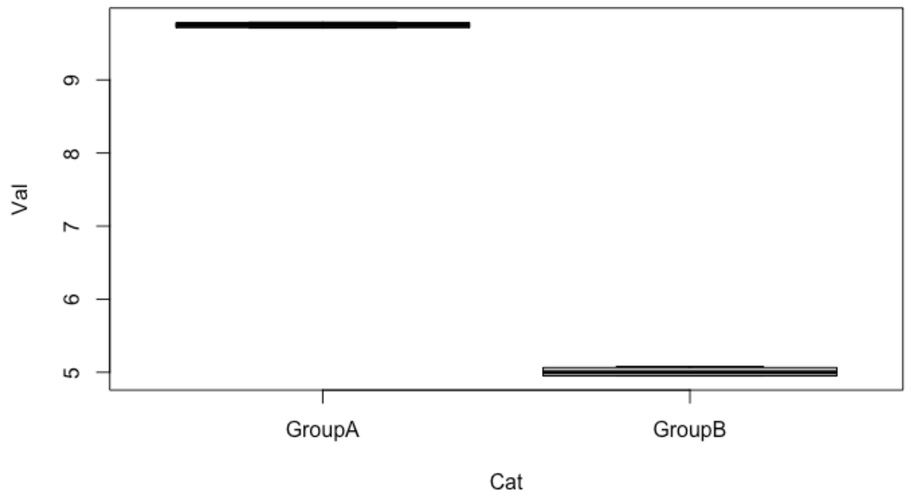
Part A

1st statement - "Kyrgyzstan Low Budget movies have higher imdb score than Romania Medium Budget movies."

- Kyrgyzstan and Low Budget both had high average imdb score in slide 3.
- Romania and Medium Budget both had low average imdb score in slide 4.
- Calculated p-value is lower than 0.05, so we can reject null hypothesis.
- In the box plot, Low budget Kyrgyzstan movies were totally higher.

```
#Hypothesis test 1
A <- movies[movies$Budget == 'Low' & movies$country =='Kyrgyzstan',]
B <- movies[movies$Budget == 'Medium' & movies$country == 'Romania',]
Cat1 <- rep("GroupA",nrow(A))
Cat2 <- rep("GroupB",nrow(B))
Cat <- c(Cat1,Cat2)
Val <- c(A[,c("imdb_score")],B[,c("imdb_score")])

d <- data.frame(Cat,Val)
boxplot(Val~Cat,data=d)
```



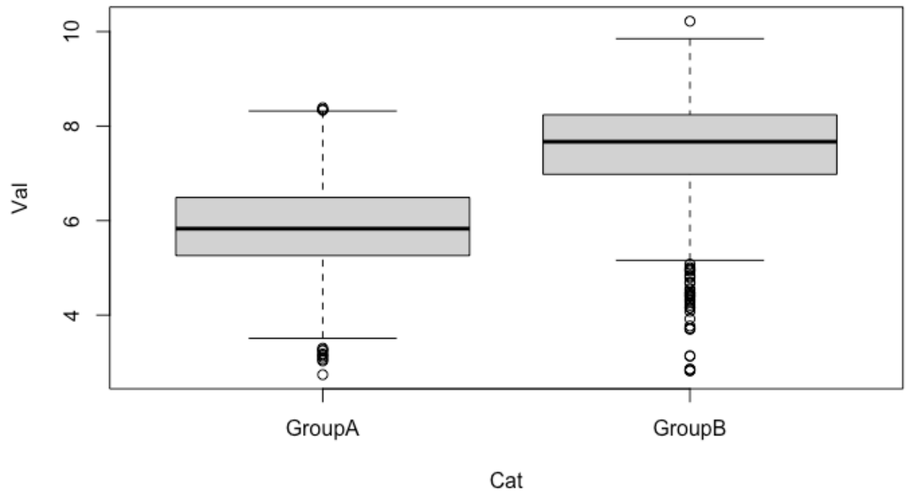
Part A

Statement 2-“R rated History genre movies have higher imdb score than PG-13 rated Family movies”

- PG-13 contents and Family genre had low average imdb score in slide 4.
- R contents and History genre had high average imdb score in slide 4.
- In the plot, we can see group B(R rated History movies) has higher average score than PG-13 family movies.

```r
#Hypothesis test 2
A <- movies[movies$content=="PG-13" & movies$genre=='Family',]
B <- movies[movies$content=='R' & movies$genre=="History",]
Cat1 <- rep("GroupA",nrow(A))
Cat2 <- rep("GroupB",nrow(B))
Cat <- c(Cat1,Cat2)
Val <- c(A[,c("imdb_score")],B[,c("imdb_score")])

d <- data.frame(Cat,Val)
boxplot(Val~Cat,data=d)
```



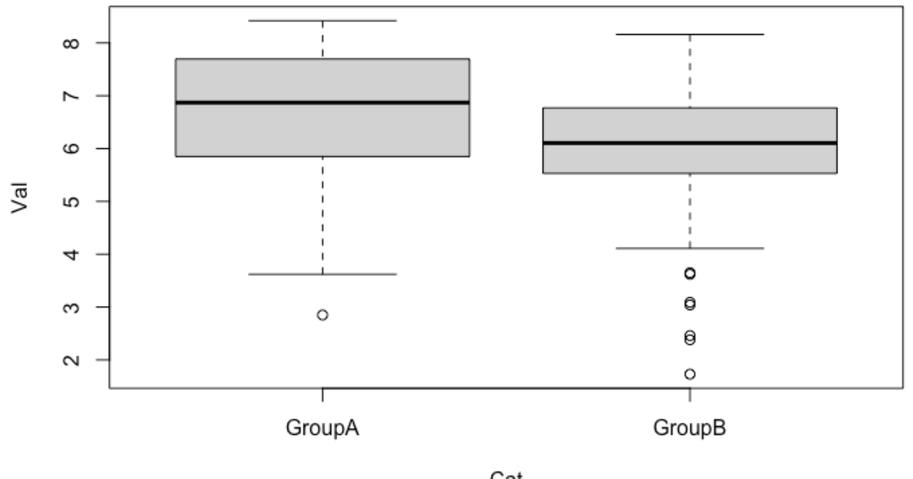
Part A

Statement 3-“G rated Sci-fi movies have higher imdb score than R rated Family movies”

- G rated had lower average imdb score than R rated in slide 4.
- But, Family genre had significantly lower average imdb score than Sci-fi genre, so I tried this statement.
- In the plot, we can see group A(G,Sci-fi) is generally higher than Group B(R, Family).

```
#Hypothesis test 3
A <- movies[movies$content == "G" & movies$genre=='Sci-Fi',]
B <- movies[movies$content == 'R' & movies$genre == 'Family',]
Cat1 <- rep("GroupA",nrow(A))
Cat2 <- rep("GroupB",nrow(B))
Cat <- c(Cat1,Cat2)
Val <- c(A[,c("imdb_score")],B[,c("imdb_score")])

d <- data.frame(Cat,Val)
boxplot(Val~Cat,data=d)
```



Part B

A - “G rated Sci-fi movies have higher imdb score than R rated Family movies”

- Null hypothesis - “G rated Sci-fi movies have same imdb score as R rated Family movies.”
- The p-value is 0.001, which is lower than 0.05. So, we can reject the null hypothesis with A.

```
#Hypothesis test 3
A <- movies[movies$content == "G" & movies$genre=='Sci-Fi',]
B <- movies[movies$content == 'R' & movies$genre == 'Family',]
Cat1 <- rep("GroupA",nrow(A))
Cat2 <- rep("GroupB",nrow(B))
Cat <- c(Cat1,Cat2)
Val <- c(A[,c("imdb_score")],B[,c("imdb_score")])

d <- data.frame(Cat,Val)
PermutationTestSecond::Permutation(d,"Cat","Val",1000,"GroupA","GroupB")
```

```
> PermutationTestSecond::Permutation(d,"Cat","Val",1000,"GroupA","GroupB")
[1] 0.001
```

Part B

B-"Low Budget Comedy movies have higher imdb score than High Budget Drama movies"

- Null hypothesis - "Low Budget Comedy movies have same imdb score as High Budget Drama movies"
- 0.021 < 0.05 . It is bigger p-value than A, but we can still reject null hypothesis with this p-value.

```
#Hypothesis test 4
A <- movies[movies$Budget == "Low" & movies$genre=='Comedy',]
B <- movies[movies$Budget == 'High' & movies$genre == 'Drama',]
Cat1 <- rep("GroupA",nrow(A))
Cat2 <- rep("GroupB",nrow(B))
Cat <- c(Cat1,Cat2)
Val <- c(A[,c("imdb_score")],B[,c("imdb_score")])

d <- data.frame(Cat,Val)
PermutationTestSecond::Permutation(d,"Cat","Val",1000,"GroupA","GroupB")
```

```
> PermutationTestSecond::Permutation(d,"Cat","Val",1000,"GroupA","GroupB")
[1] 0.021
```

Part B

C-“PG-13 rated Drama movies have higher imdb score than PG rated Sci-Fi movies”

- Null hypothesis - “PG-13 rated Drama movies have same imdb score as PG rated Sci-Fi movies”
- 0.106 > 0.05. So, we cannot reject null hypothesis with this alternative hypothesis.

```
#Hypothesis test 5
A <- movies[movies$content == "PG-13" & movies$genre=='Drama',]
B <- movies[movies$content == 'PG' & movies$genre == 'Sci-Fi',]
Cat1 <- rep("GroupA",nrow(A))
Cat2 <- rep("GroupB",nrow(B))
Cat <- c(Cat1,Cat2)
Val <- c(A[,c("imdb_score")],B[,c("imdb_score")])

d <- data.frame(Cat,Val)
PermutationTestSecond::Permutation(d,"Cat","Val",1000,"GroupA","GroupB")
```

```
> PermutationTestSecond::Permutation(d,"Cat","Val",1000,"GroupA","GroupB")
[1] 0.106
```