VISUALIZATION ASSIGNMENT

Prepared by:

Team Data Insighters

Sreekar Bethu

Balacoumarane Vetrivel

Mohammed Ismail Khan

Meghna Vinay Amin

Table of Contents

Task-1

Task-2

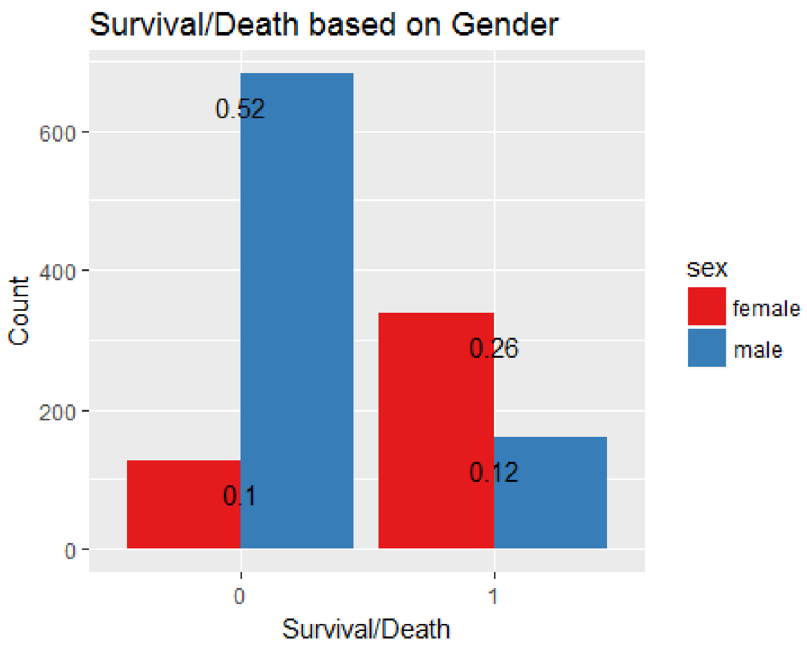
Code

TASK-1

1. Gender on survival

There are no spaces or missing values in Gender, hence all the observations are considered.

Graphical Representation:



Inferences:

1) Proportion of males who died is higher than the proportion of females who died.
2) Proportion of females survived is more than double the proportion of males survived.

Numbers from dataset given are as below:

Number of people died: 809 Females: 466

Number of people survived: 500 Males: 843

As per the calculation below are the approximate results:

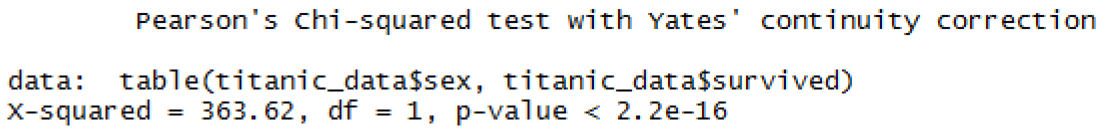
Gender	Survived	Died
Male	161	682
Female	339	127

Hypothesis testing:

Null Hypothesis (H_o) - The Gender doesn't have any impact on the survival i.e. proportion of a males surviving= proportion of a females surviving.

Alternate Hypothesis (H_1) – The Gender has an impact on the survival of the passenger i.e. proportion of males surviving is not equal to proportion of females surviving.

Chi-square test is performed in Rstudio:



```
	Pearson's Chi-squared test with Yates' continuity correction

data:  table(titanic_data$sex, titanic_data$survived)
X-squared = 363.62, df = 1, p-value < 2.2e-16
```

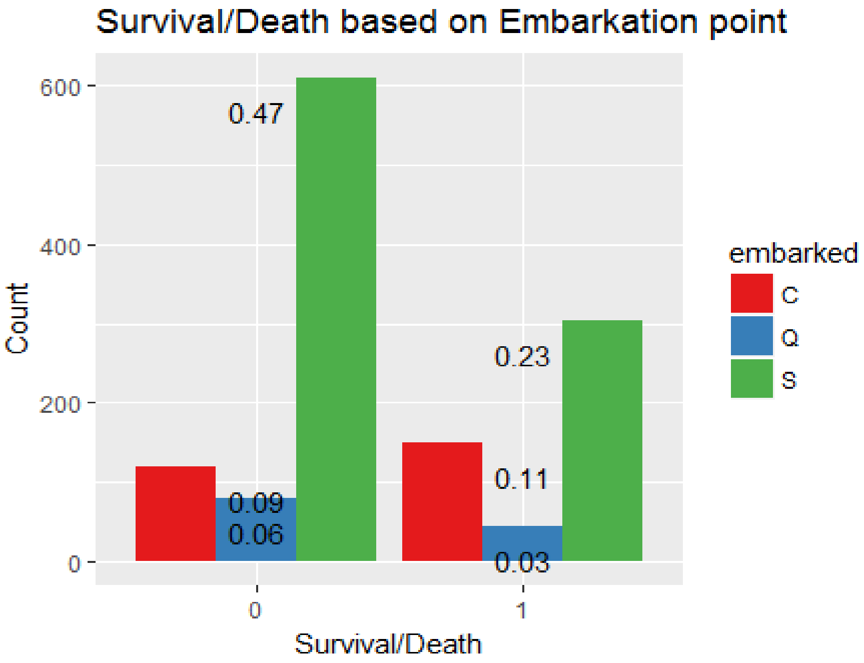
Since p- value is less than 0.05, we fail to accept our null Hypothesis.

Conclusion: Accepting the Alternate Hypothesis (H_1) i.e. the gender has an impact on the survival of the passenger.

2. Embarkation points on survival

Two spaces, that are present in embarkation point of survival are assumed to be of less importance and hence ignored.

Graphical Representation:



Inferences:

1. Proportion of people died is higher in S than C and Q.
2. Proportion of people survived is more in C than S and Q.

Numbers from dataset given are as below:

Number of people died: 809
Number of people survived: 500

Number of passengers in S: 914
Number of passengers in C: 270
Number of passengers in Q: 123

As per the calculation below are the approximate results:

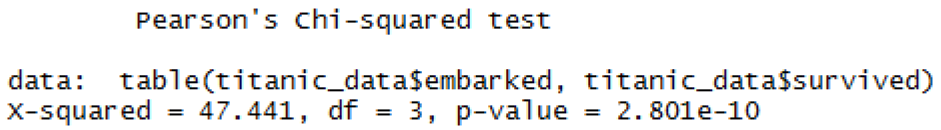
Embarkation Point	Survived	Died
S	304	610
C	150	120
Q	44	79

Hypothesis testing:

Null Hypothesis (H_0) - The Embarkation point doesn't have any impact on the survival i.e. proportion of people survived is equal in all the embarkation points.

Alternate Hypothesis (H_1) – The Embarkation point has impact on the survival i.e. proportion of people survived is more in the people from Embarkation Point C.

Chi-square test is performed in Rstudio:



```
                Pearson's Chi-squared test

data:  table(titanic_data$embarked, titanic_data$survived)
X-squared = 47.441, df = 3, p-value = 2.801e-10
```

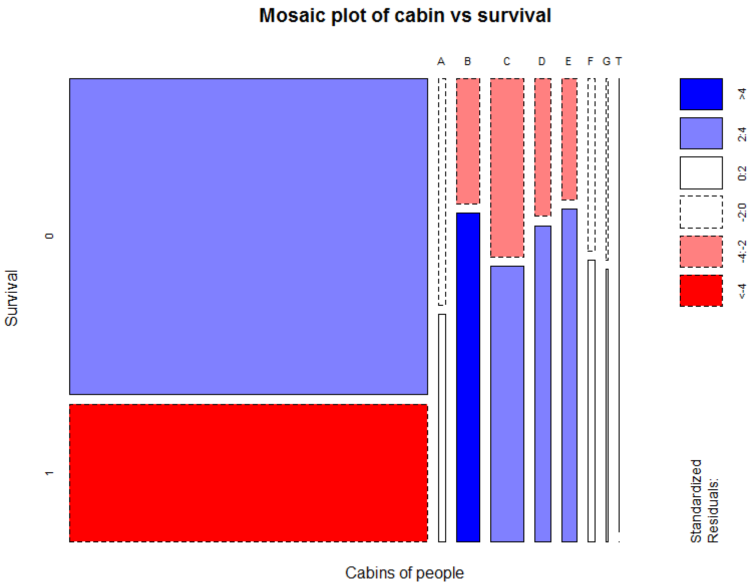
Since p- value is less than 0.05, we fail to accept our null Hypothesis.

Conclusion: Accepting the Alternate Hypothesis (H_1) i.e. the embarkation point has an impact on the survival of the passenger.

3. Cabin on survival

From the data given, there are 1014 observations with SPACE in the cabin column. We assumed those observations as people without cabin. Remaining 295 observations have 187 levels in the column. Values in the cabin column start with an Alphabet followed by a number. We assumed the observations starting with same Alphabet belongs to a group and categorized the cabin column into 9 levels as (People with no cabin, A, B, C, D, E, F, G, T)

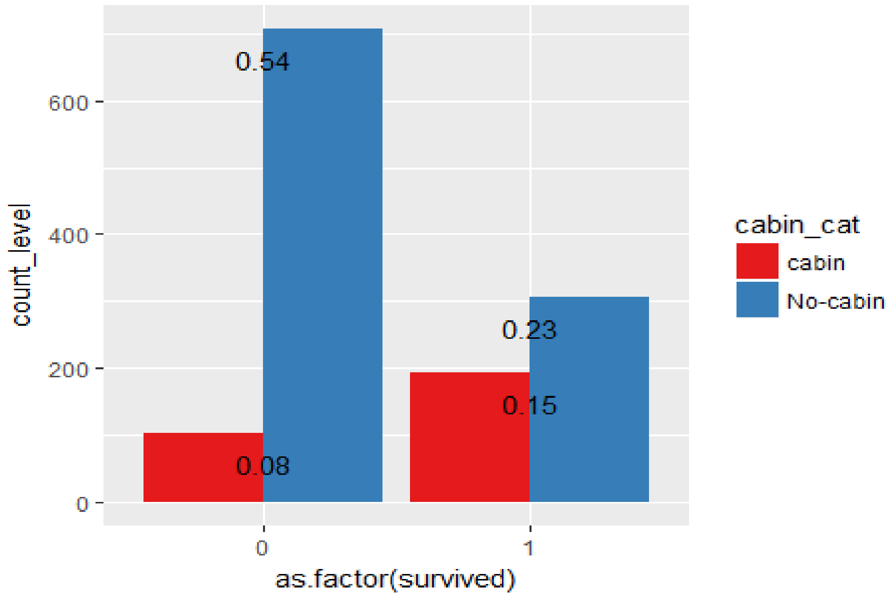
Graphical Representation:



Inferences:

1) The proportion of death is higher in the people without cabin.
2) The proportion of survival is higher in the people of cabins (A, B, C, D and E) of higher level.

Graphical Representation of cabin and No-cabin:



Numbers from dataset given are as below:

Number of people died: 809

Number of people survived: 500

Number of passengers with no cabin: 1014

Number of passengers with cabin: 295

As per the calculation below are the approximate results:

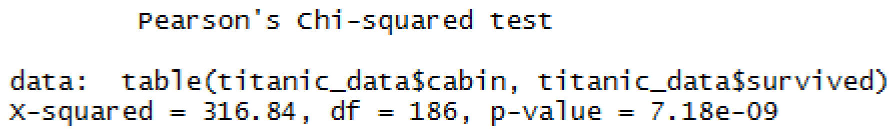
	Survived	Died
Cabin	193	102
No Cabin	307	707

Hypothesis testing:

Null Hypothesis (H_o) - The Cabin doesn't have any impact on the survival i.e. survival proportion of people with cabin is equal to survival proportion of people with no cabin.

Alternate Hypothesis (H_1) – The Cabin has an impact on the survival i.e. survival proportion of people with cabin is not equal to survival proportion of people with no cabin

Chi-square test is performed in Rstudio:



```
	Pearson's Chi-squared test

data:  table(titanic_data$cabin, titanic_data$survived)
X-squared = 316.84, df = 186, p-value = 7.18e-09
```

Since p- value is less than 0.05, we fail to accept our null Hypothesis.

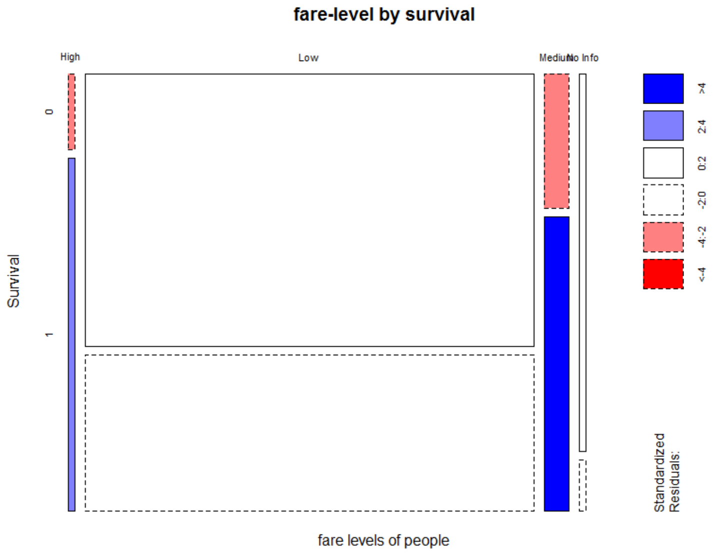
Conclusion: Accepting the Alternate Hypothesis (H_1) i.e. the Cabin has an impact on the survival of the passenger.

4. Fare levels on survival:

Fares of the passengers are categorized into 4 levels as mentioned below.

If ‘fare = 0’ then it is categorized as 'No Info'
If ‘fare > 0’ & ‘fare < 100’ and then it is categorized as ‘Low'
If ‘fare > 100’ & ‘fare < 250’ and then it is categorized as ‘Medium'
If ‘fare > 250’ and then it is categorized as ‘High’

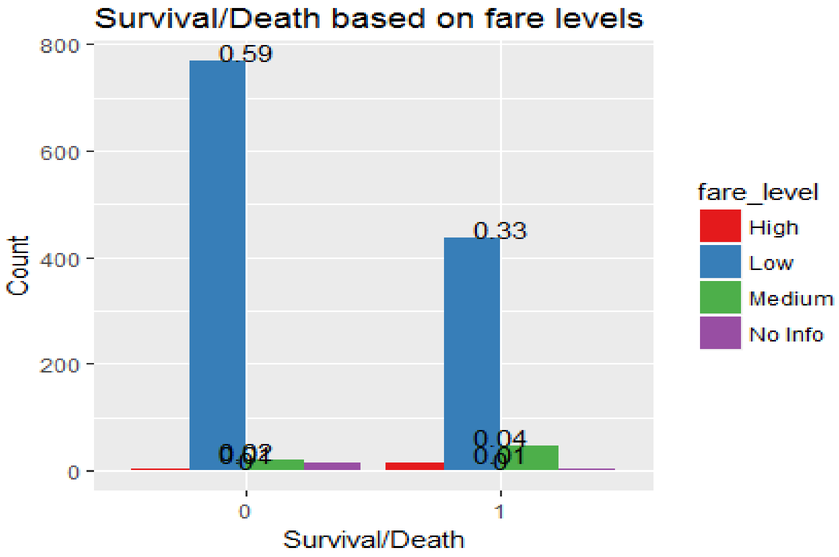
Graphical Representation:



Inference:

1. The proportion of people survived is high in High and Medium fare levels as they would have had taken the higher cabins. These people may have been given higher preference to be lead to the safety boats and hence there is a higher survival rate in these fare levels.
2. The proportion of people dead is higher in low fare levels.

Bar-Graphical Representation:



Numbers from dataset given are as below:

Number of people died: 809	Number of passengers with low fair level: 1207
Number of people survived: 500	Number of passengers with medium fair level: 67
	Number of passengers with high fair level: 17
	Number of passengers with no fair level: 17

As per the calculation below are the approximate results:

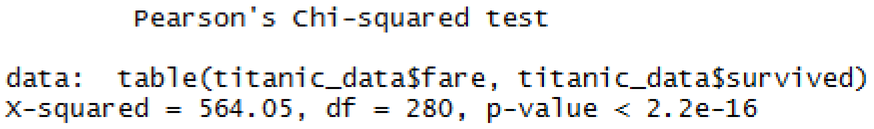
Fair Level	Survived	Died
Low	438	769
Medium	46	21
High	14	3
No Fair	2	15

Hypothesis testing:

Null Hypothesis (H_o) - The fare level doesn't have any impact on the survival i.e. survival proportion of people will be equal in all fare levels.

Alternate Hypothesis (H_1) – The fare level has an impact on the survival i.e. survival proportion of people will not be equal in all fare levels.

Chi-square test is performed in Rstudio:



```
	Pearson's Chi-squared test

data:  table(titanic_data$fare, titanic_data$survived)
X-squared = 564.05, df = 280, p-value < 2.2e-16
```

Since p- value is less than 0.05, we fail to accept our null Hypothesis.

Conclusion: Accepting the Alternate Hypothesis (H_1) i.e. the fare level has an impact on the survival of the passenger.

TASK-2

5. Data Set:

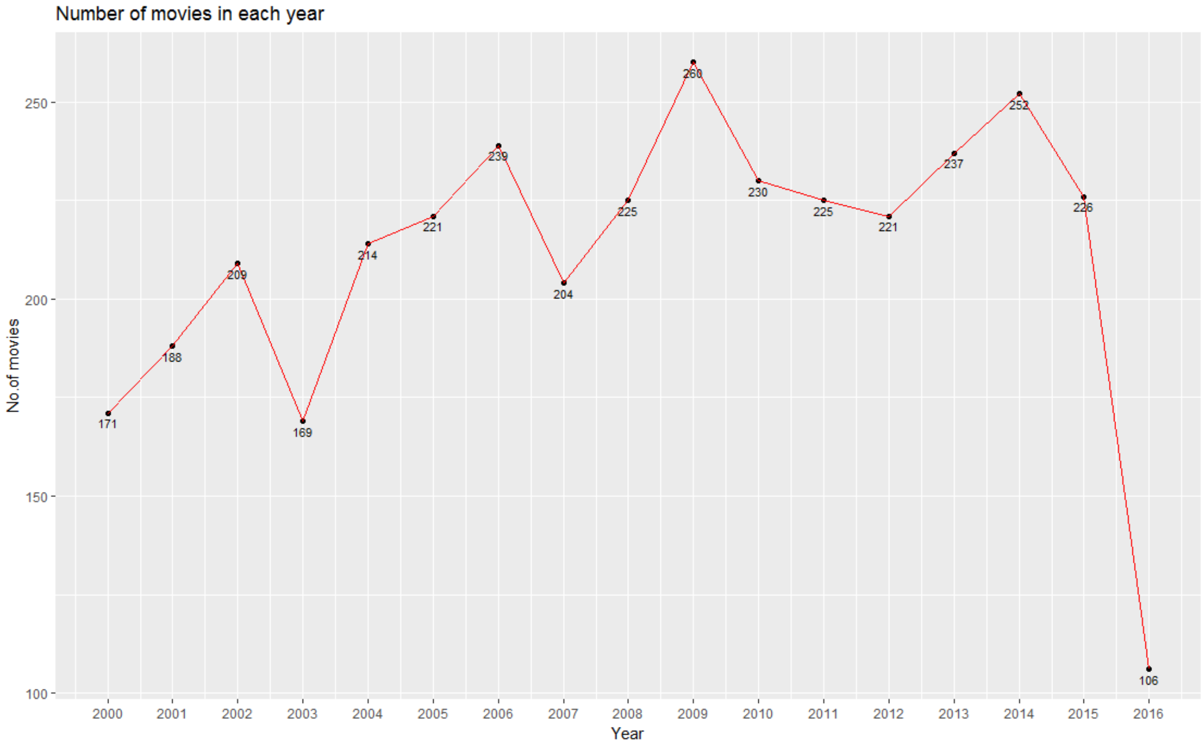
IMDB dataset of the movies released from 2000 -2016 across the globe has been considered for this assignment. The dataset contains 3597 records and the fields are given below:

Variable	Description	Type	Possible Values
color	Type of movie-Black or White or colour	Factor	2-Black & White, Colour
movie_title	Movie title	Factor	String
duration	Duration Of film	int	
director_name	Name of Director	Factor	String
director_facebook_likes	Number of Facebook likes of director	int	
actor_1_name	Name of actor 1	Factor	String
actor_1_facebook_likes	Number of Facebook likes for actor 1	int	
actor_2_name	Name of Actor_2	Factor	String
actor_2_facebook_likes	Number of Facebook likes for actor 2	int	
actor_3_name	Name of actor 3	Factor	String
actor_3_facebook_likes	Number of Facebook likes for actor	int	
gross	Gross Earning Of film	int	
gross_in_million	Gross Earning Of film in millions	int	
genres	Genres Of film	Factor	String
num_critic_for_reviews	Number of Critical Reviews for film	int	
num_voted_users	Number of Voted users of film	int	
cast_total_facebook_likes	Total facebook likes of Film	int	
facenumber_in_poster	Number of facenumber on poster	int	
plot_keywords	Keywords of Plot	Factor	String
num_user_for_reviews	Number of User Reviews for film	int	
language	Language Of film	Factor	String
country	Country Of film	Factor	String
content_rating	Number of Content rating for	Factor	String
budget	Budget Of film	num	
Budget_in_million	Budget Of film in millions	num	
title_year	The year the film was released	int	
imdb_score	IMDB score of film	num	
aspect_ratio		num	
movie_facebook_likes	Number of Facebook likes of movie	int	

The initial analysis showed that the dataset had missing values along with a few duplicate records. The missing rows have been filtered and the rows with complete information have been used.
Total number of complete records = 2766
Total number of incomplete records = 831.

6. Question 1: What is the pattern of total movies released per year?

- Line Graph is plotted as below;
- Y-axis: Number of movies
- X-axis: Years; in increasing order to identify the pattern of total number of movies released
- Total number of movies produced is displayed at the tip of each bar



Inferences:

1. Year 2009 has most number (260) of movies released and Year 2016 has the least number (106) of movies.
2. On analysing the trends, it can be inferred that the data set might be generated in the mid of 2016 and so it doesn't include all the movies released in that year (106).
3. From the graph plotted, we can infer that total number of movies produced has a gradual increase.

7. Question 2:
Is the gross of each movie affected by IMDB rating?

Categorize the IMDB ratings as below:

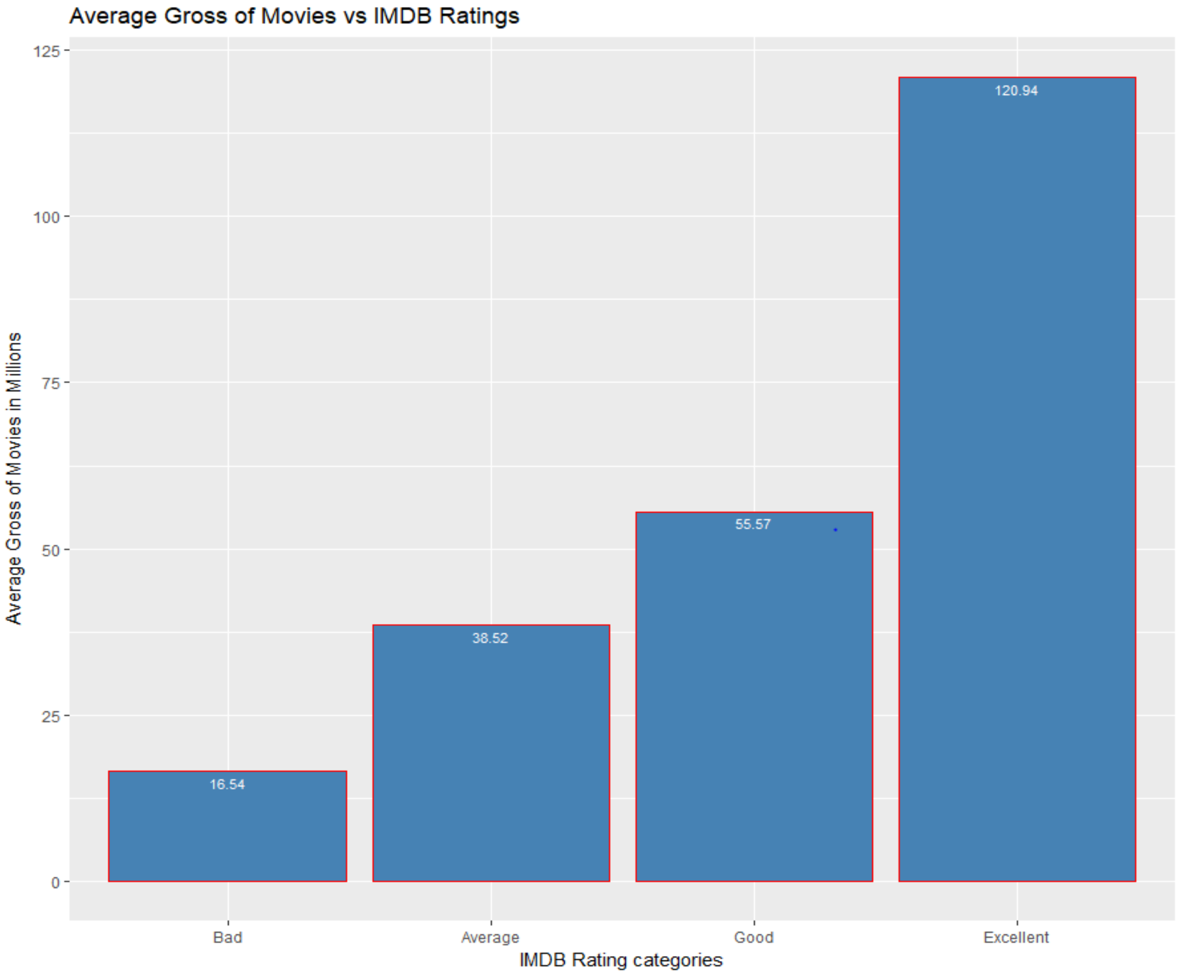
If ‘imdb_score >= 1’ & ‘imdb_score < 3’ then it is categorized as 'Bad'
If ‘imdb_score >= 3’ & ‘imdb_score < 6’ then it is categorized as 'Average'
If ‘imdb_score >= 6’ & ‘imdb_score < 8’ then it is categorized as 'Good'
If ‘imdb_score >= 8’ then it is categorized as 'Excellent'

After categorizing the IMDB score, bar graph is plotted for averages of gross for each category of IMDB scores.

- Bar Graph is plotted as below;
- Y-axis: Avg. Gross in million
- X-axis: Categories of IMDB score
- Graph is sorted in the ascending order of Avg. Gross in million

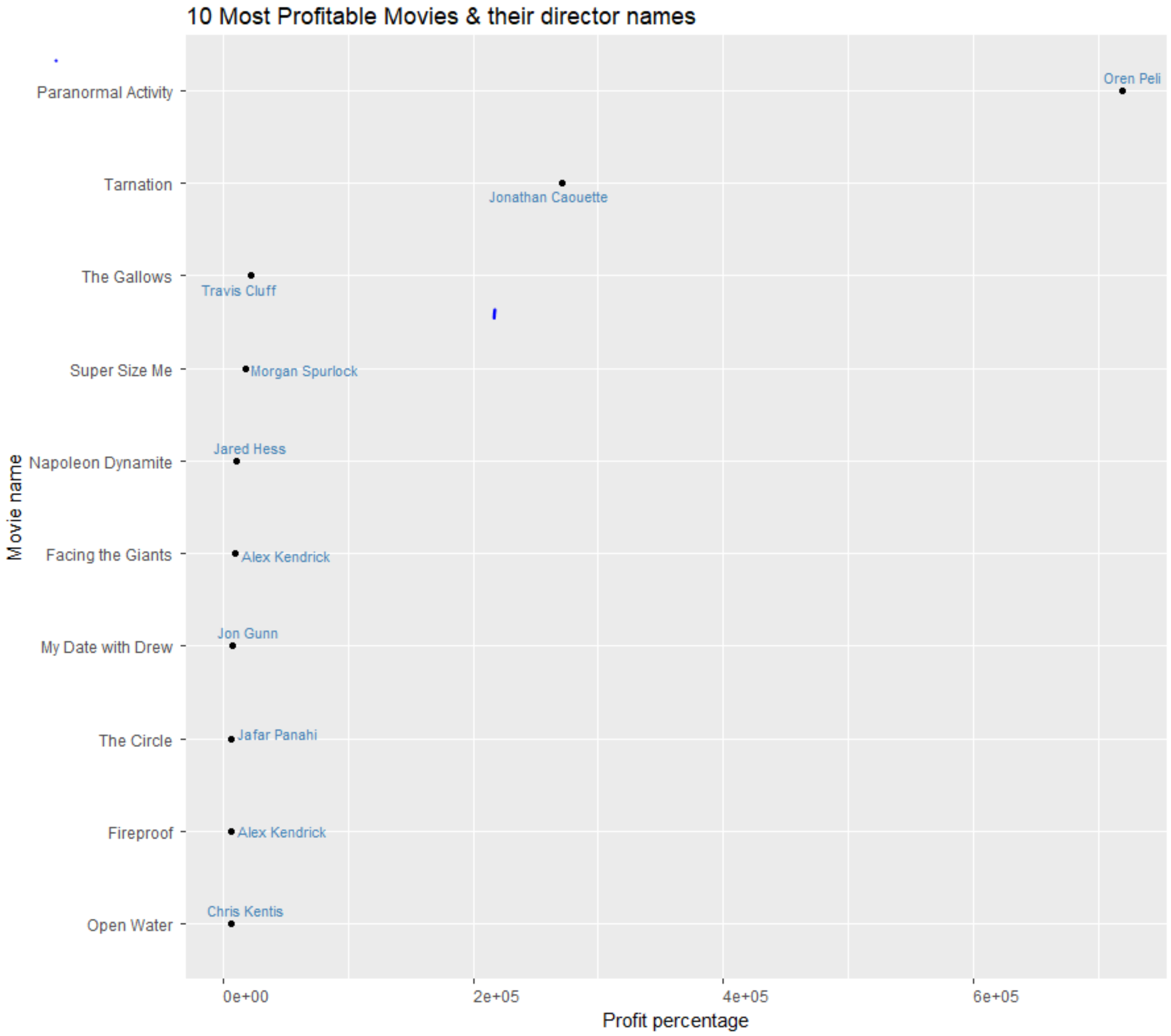
Inferences:

1. Average gross of movies is increasing along with IMDB rating Categories.
2. From the plot, it can be inferred that a higher IMDB rating is directly proportional to the Gross revenue.



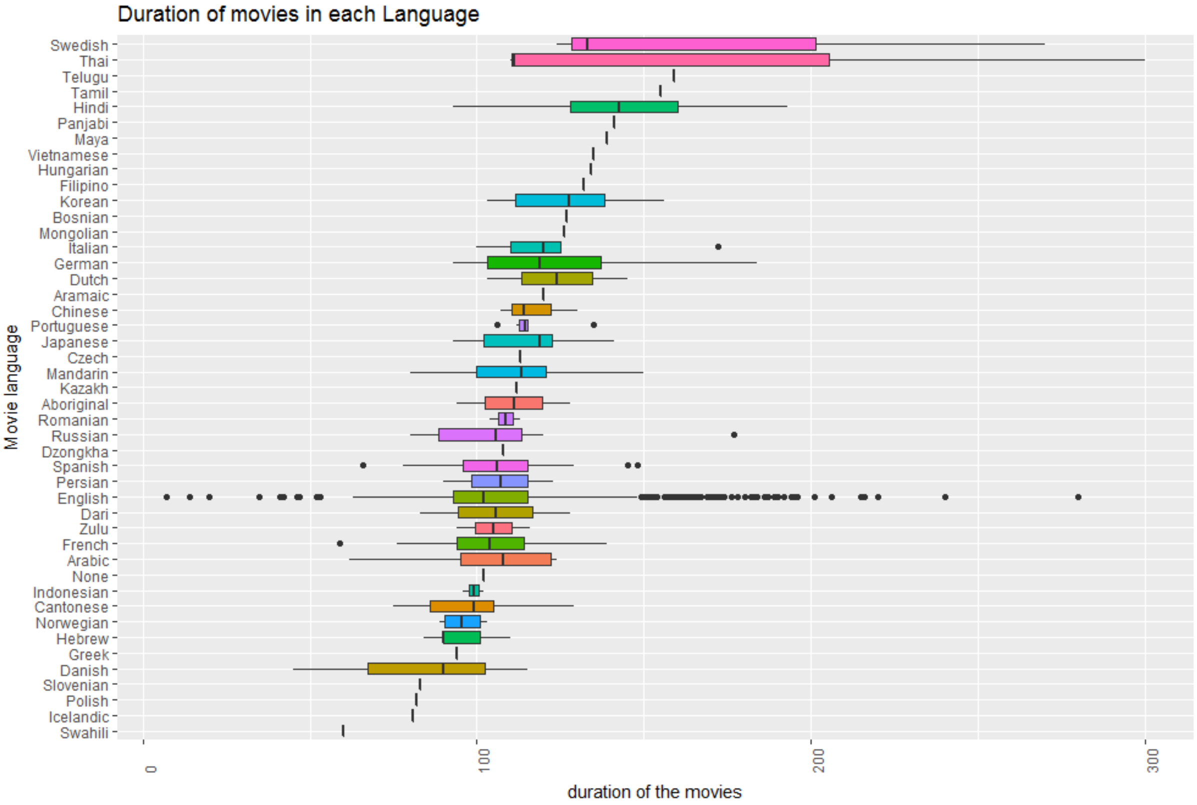
8. Question 3: Which are the top 10 profitable movies with their respective movie directors?

- Y-axis: Top 10 movie names
- X-axis: Movie profit percentages. Points represent the respective director's name
- Profit percentage of movies 'Paranormal Activity' and 'Tarnation' is too high and hence the scaling of other movies is improper.



9. Question 4:
What is the duration of movies across each language?

- Box-plot is used to understand the spread of duration of movies in each language.
- Y-axis: Languages
- X-axis: Duration



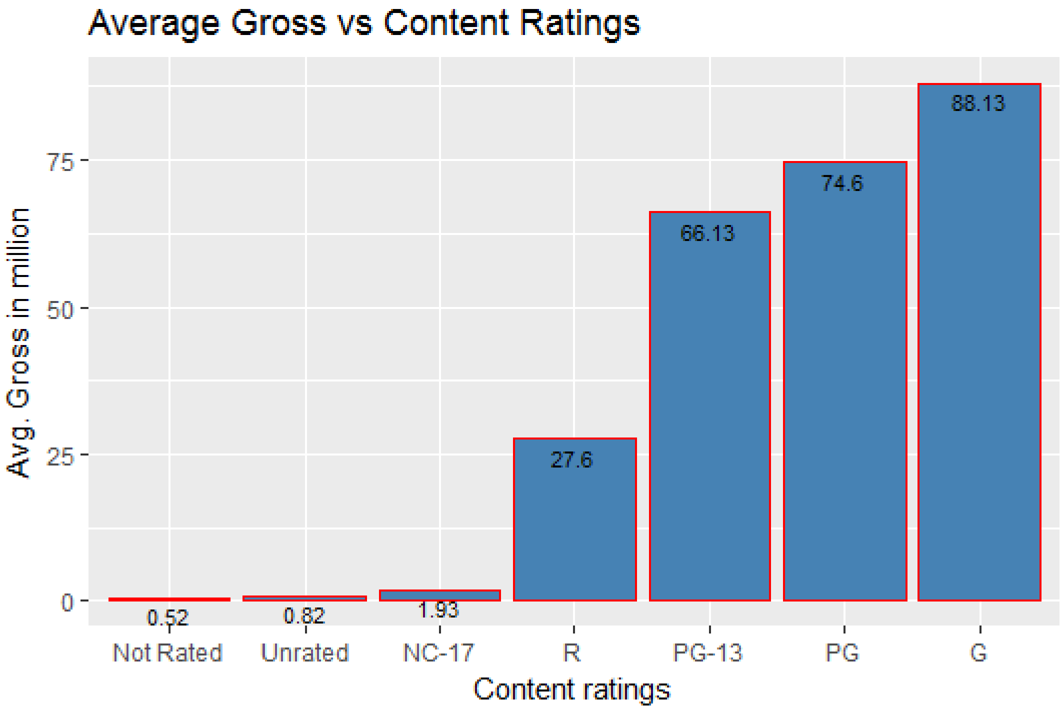
Inferences:

1. Median of Swedish & Thai language is less but the spread is wider, number of movies in Telugu & Tamil is only 1.
2. From the top, duration of Hindi movies is wider, and the median of Duration is higher compared to other languages.
3. Considering all languages, duration of movies is wider in English language and 50% of English movies have duration lesser than 120 Min.
4. Inter-Quartile range of most number of languages lies between 50 & 100.

10. Question 5:

Movies with which content ratings have performed well?

- Bar Graph is plotted as below;
- Y-axis: Avg. Gross of movies
- X-axis: Content rating. Avg. Gross of each rating is displayed at the tip of each bar.
- Graph is sorted in the ascending order of Avg. Gross in million

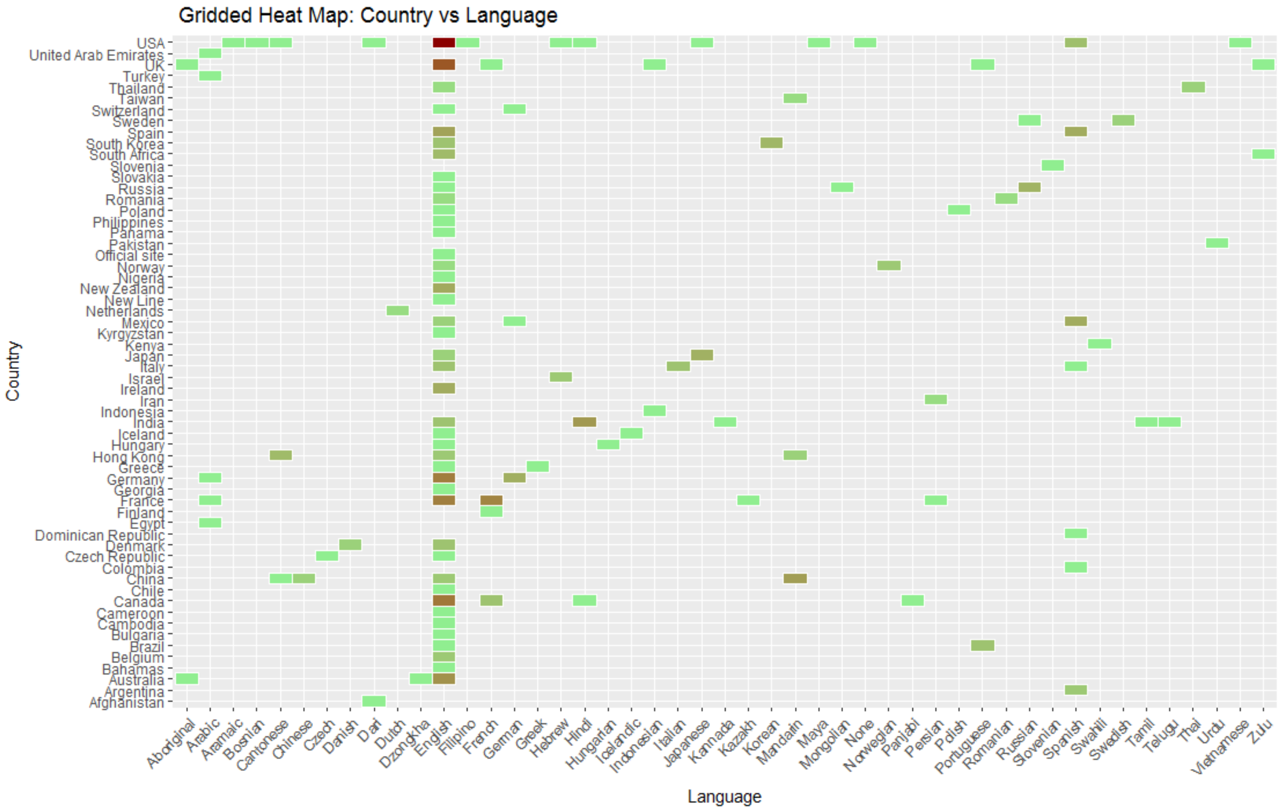


Inferences:

1. The graph shows that movies with G content rating has highest average gross earnings as the movies can be viewed by children also. A visit to the movies for children will mostly have the adults also accompanying them. Hence increased number of tickets means increased sales and subsequently higher gross.
2. The movies with rating R would mean that the movie can only be viewed by a restricted audience (people only above 17 years age) and hence a dip in the number of tickets being sold. This is reflected by the small gross earning.
3. Thus, it can be inferred that the content rating of the movie affects the average gross earnings.

11. Question 6:
Which are the prominent languages in each country?

- Heat map is produced as below;
- Y-axis: Language
- X-axis: Country



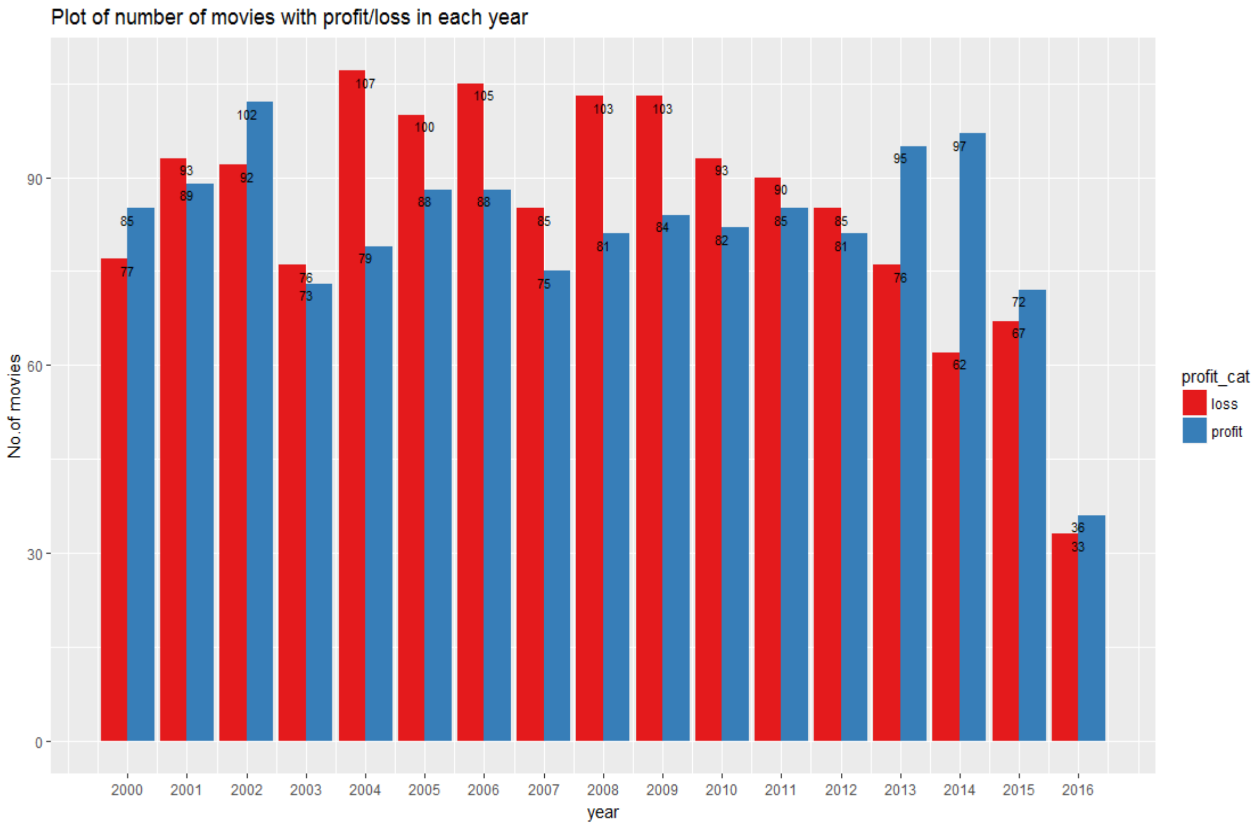
Inferences:

1. English is widely used language across many countries.
2. USA is the country where movies are produced in many languages.
3. Prominent language for movies in India is Hindi.

12. Question 7:

Comparing the profit and loss of movies in the recent years

- Bar Graph is plotted as below;
- Y-axis: Number of movies
- X-axis: Year
- Number of movies with profit & loss are distinguished by the two different colours of bars.
- X-axis is arranged in the sequence of years to identify the pattern of total number of movies released over the years.



Inferences:

1. Movies incurred substantial losses during the years 2004 to 2011
2. In the recent years, starting 2013 number of movies that made profit is higher than number of movies under loss.

Code

13. R Code

```
## VISUALIZATION ASSIGNMENT ##
## Team: Data Insighters
## Members:
#  Sreekar Bethu
#  Balacoumarane Vetrivel
#  Mohammed Ismail Khan
#  Meghna Vinay Amin
#  Misha Singh

################## TASK - 1 ##################

# load libraries
pacman :: p_load(tidyverse, reshape2, readxl, jsonlite, corrplot, XLConnect, magrittr, ggrepel)

# Read data into r from CSV file
titanic_data <- read.csv("titanic3.csv",header=TRUE)

# Get column names
names(titanic_data)

# View the summary of data
summary(titanic_data)

# Get details of each column
str(titanic_data)

## Task - 1

############################ 1 Gender on Survival ############################

# Get the class of Gender
cols_to_change = c('sex')
sapply(titanic_data[cols_to_change], class)

# Plot the distribution of Gender on Survival
titanic_data %>%
  group_by(sex,survived) %>%
  summarise(count_level = n(),
          percentage = n()/nrow(titanic_data)) %>%
  ggplot(aes(x= as.factor(survived),y=count_level,fill=sex)) +
  geom_bar(stat='identity',position='dodge') + xlab("Survival/Death") + ylab("Count") +
  ggtitle("Survival/Death based on Gender") +
  geom_text(aes(label=round(percentage,2)),vjust=2) + scale_fill_brewer(palette = "Set1")

# Chi square test for Gender on Survival
chisq.test(table(titanic_data$sex, titanic_data$survived))

############################ 2 Embarked on Survival ############################
```

```
# Get the class of Embarkation
cols_to_change = c('embarked')
sapply(titanic_data[cols_to_change], class)

# Two observations with NA are removed and distribution of Embarktion on Survival is plotted
titanic_data %>%
  subset(embarked != "") %>%
  group_by(embarked,survived) %>%
  filter(!is.na(embarked)) %>%
  summarise(count_level = n(),
            percentage = n()/nrow(titanic_data)) %>%
  ggplot(aes(x= as.factor(survived),y=count_level,fill=embarked)) +
  geom_bar(stat='identity',position='dodge') + xlab("Survival/Death") + ylab("Count") +
  ggtitle("Survival/Death based on Embarkation point") +
  geom_text(aes(label=round(percentage,2)),vjust=2) + scale_fill_brewer(palette = "Set1")

# Chi square test for embarkation point on Survival
chisq.test(table(titanic_data$embarked, titanic_data$survived))

############################ 3 Cabin on Survival ############################

# Generate new variable with first letter of Cabin
titanic_data$cabin_first <- substr(titanic_data$cabin,1,1)
# Check the class of new variable
cols_to_change = c('cabin_first')
sapply(titanic_data[cols_to_change], class)
# Change the new variable class to factor
titanic_data[cols_to_change] = lapply(titanic_data[cols_to_change], factor)
sapply(titanic_data[cols_to_change], class)

# Mosaic plot of different cabins on survival
mosaicplot(table(titanic_data$cabin_first, titanic_data$survived),
           xlab='Cabins of people',ylab='Survival',
           main='Mosaic plot of cabin vs survival', shade=T, color = T)

# Generate a new variable to check if passenger is having cabin or not
titanic_data$cabin_cat[titanic_data$cabin_first==""] <- 'No-cabin'
titanic_data$cabin_cat[titanic_data$cabin_first!=""] <- 'cabin'

# distribution of cabin on Survival is plotted
titanic_data %>%
  group_by(cabin_cat,survived) %>%
  summarise(count_level = n(),
            percentage = n()/nrow(titanic_data)) %>%
  ggplot(aes(x= as.factor(survived),y=count_level,fill=cabin_cat)) +
  geom_bar(stat='identity',position='dodge') + xlab("Survival/Death") + ylab("Count") +
  ggtitle("Survival/Death based on Cabin") +
  geom_text(aes(label=round(percentage,2)),vjust=2) + scale_fill_brewer(palette = "Set1")
```

```
# Chi square test of cabin on survival
chisq.test(table(titanic_data$cabin, titanic_data$survived))

############################## 4 Fare on Survival ##############################

# Split the fare level into following categories
titanic_data$fare_level[titanic_data$fare == 0] = 'No Info'
titanic_data$fare_level[titanic_data$fare < 100 & titanic_data$fare > 0] = 'Low'
titanic_data$fare_level[titanic_data$fare < 250 & titanic_data$fare >= 100] = 'Medium'
titanic_data$fare_level[titanic_data$fare >= 250] = 'High'

# Check the class of new variable
cols_to_change = c('fare_level')
sapply(titanic_data[cols_to_change], class)
# Check the new variable class to fator
titanic_data[cols_to_change] = lapply(titanic_data[cols_to_change], factor)
sapply(titanic_data[cols_to_change], class)

# Mosaic plot of different fare levels on survival
mosaicplot(table(titanic_data$fare_level, titanic_data$survived),
          xlab='fare levels of people',ylab='Survival',
          main='fare-level by survival', shade=TRUE)

# bar plot representation of different fare levels on survival
titanic_data %>%
  group_by(fare_level,survived) %>%
  filter(!is.na(fare_level)) %>%
  summarise(count_level = n(),
           percentage = n()/nrow(titanic_data)) %>%
  ggplot(aes(x= as.factor(survived),y=count_level,fill=fare_level)) +
  geom_bar(stat='identity',position='dodge') + xlab("Survival/Death") + ylab("Count") +
  ggtitle("Survival/Death based on fare levels") +
  geom_text(aes(label=round(percentage,2)),vjust=0) + scale_fill_brewer(palette = "Set1")

# Chi square test
chisq.test(table(titanic_data$fare, titanic_data$survived))

################# END OF TASK-1 #################

################### TASK - 2 ###################

# Read IMDB data into R from Excel
IMDB2000 <- read.csv("IMDB_2000.csv",header=TRUE)

# To get the summary of the data loaded and string details of data.
summary(IMDB2000)
str(IMDB2000)

# To find number of missing and non-missing rows
missing = IMDB2000 %>%
  filter(!complete.cases(.))
```

```
nrow(missing)
non_missing = IMDB2000 %>%
  filter(complete.cases(.))
nrow(non_missing)

  # Q1.          What is the pattern of total movies released per year?

  # Plot
  IMDB2000 %>%
  subset(title_year != "") %>%
  group_by(title_year) %>%
  filter(!is.na(title_year)) %>%
  summarise(count_level = n()) %>%
  ggplot(aes(x= title_year,y=count_level)) +
  geom_point() + geom_line(color='red') +
  ggtitle("Number of movies in each year") + xlab("Year") + ylab("No.of movies") +
  scale_x_continuous(breaks = seq(2000,2016,1)) +
  geom_text(aes(label=count_level),vjust=1.5,color="black",size=3)

################################################################################

# Q2.          Is the gross of each movie affected by IMDB rating?

# Categorize IMDB Score
IMDB2000$imdb_cat[IMDB2000$imdb_score >= 1 & IMDB2000$imdb_score < 3] = 'Bad'
IMDB2000$imdb_cat[IMDB2000$imdb_score >= 3 & IMDB2000$imdb_score < 6] = 'Average'
IMDB2000$imdb_cat[IMDB2000$imdb_score >= 6 & IMDB2000$imdb_score < 8] = 'Good'
IMDB2000$imdb_cat[IMDB2000$imdb_score >= 8] = 'Excellent'

# Check the class of new variable
cols_to_change = c('imdb_cat')
sapply(IMDB2000[cols_to_change], class)
# Change the new variable class to factor
IMDB2000[cols_to_change] = lapply(IMDB2000[cols_to_change], factor)
sapply(IMDB2000[cols_to_change], class)

# Plot
IMDB2000 %>%
  filter(!is.na(gross)) %>%
  group_by(imdb_cat) %>%
  summarise(avg_gross = mean((gross))/1000000) %>%
  ggplot(aes(x=reorder(imdb_cat, avg_gross),y=avg_gross)) +
  geom_bar(stat='identity',fill='steelblue',color='red') +
  xlab("IMDB Rating categories") +
  ylab("Average Gross of Movies in Millions") +
  ggtitle("Average Gross of Movies vs IMDB Ratings") +
  geom_text(aes(label=round(avg_gross, digits=2)),vjust=1.5,color="white",size=3)
```

```
##################################################################################

## Q3.        Which are the top 10 profitable movies with their respective movie directors?

 # Plot
 IMDB2000 %>%
 subset(movie_title != "") %>%
 subset(director_name != "") %>%
 filter(!is.na(movie_title)) %>%
 filter(!is.na(director_name)) %>%
 filter(!is.na(gross)) %>%
 filter(!is.na(budget)) %>%
 mutate(profit = ((gross - budget)/budget) * 100)%>%
 top_n(10, profit) %>%
 ggplot(aes(x=reorder(movie_title, profit), y=profit)) +
 geom_point() +
 geom_text_repel(aes(label = director_name),size=3,colour="steelblue") +
 xlab("Movie name") +
 ylab("Profit percentage") +
 ggtitle("10 Most Profitable Movies & their director names") +
 coord_flip()

##################################################################################

 # Q4.        What is the duration of movies across each language?

 # Plot
 IMDB2000 %>%
 subset(language != "") %>%
 subset(duration != "") %>%
 filter(!is.na(language)) %>%
 filter(!is.na(duration)) %>%
 group_by(language,duration) %>%
 ggplot(aes(x=reorder(language, duration) ,y=duration, fill=language)) +
 geom_boxplot() +
 xlab("Movie language") + ylab("duration of the movies") +
 ggtitle("Duration of movies in each Language") + guides(fill=FALSE) +
 coord_flip()

##################################################################################

 # Q5.        Movies with which content ratings have performed well?

 # Plot
 IMDB2000 %>%
 subset(content_rating != "") %>%
 filter(!is.na(gross)) %>%
 group_by(content_rating) %>%
 filter(!is.na(content_rating)) %>%
 summarise(avg_gross = mean(gross)/1000000) %>%
 ggplot(aes(x=reorder(content_rating, avg_gross),y=avg_gross)) +
 geom_bar(stat='identity',fill='steelblue',color = 'red') + xlab("Content ratings") +
 ylab("Avg. Gross in million") +
 ggtitle("Average Gross vs Content Ratings") +
 geom_text(aes(label=round(avg_gross,digits = 2)),vjust=1.5,color="Black",size=3)

##################################################################################


# Q6          Which are the prominent languages in each country?


# Plot
IMDB2000 %>%
 subset(country != "") %>%
 subset(language != "") %>%
 group_by(country,language) %>%
 filter(!is.na(country)) %>%
 filter(!is.na(language)) %>%
 summarise(Movie_Count=n()) %>%
 ggplot(aes(language,country))+
 geom_tile(aes(fill=log(Movie_Count)),colour="whit e")+
 scale_fill_gradient(low="lightgreen",high = "darkred")+
 xlab("Language")+
 ylab("Country")+
 ggtitle(" Gridded Heat Map: Country vs Language")+
 theme(axis.text.x=element_text(angle=45, hjust=1)) +  guides(fill=FALSE)

##################################################################################

 # Q7          Comparing the profit and loss of movies in the recent years

 IMDB2000 %>%
 filter(!is.na(gross)) %>%
 filter(!is.na(budget)) %>%
 subset(title_year != "") %>%
 mutate(profit = (gross - budget)) %>%
 mutate(profit_cat = ifelse(profit>0,'profit','loss')) %>%
 group_by(title_year,profit_cat) %>%
 filter(!is.na(title_year)) %>%
 summarise(count_level = n()) %>%
 ggplot(aes(x= title_year,y=count_level,fill=profit_cat)) +
 geom_bar(stat='identity',position='dodge') + xlab("year") + ylab("No.of movies") +
 geom_text(aes(label=count_level),vjust=1.5,color="black",size=3) +
 ggtitle("Plot of number of movies with profit/loss in each year") +
 scale_x_continuous(breaks = seq(2000,2016,1)) +
 scale_fill_brewer(palette = "Set1")
```