1.

a. Prediction
b. Classification
c. Optimization
d. Unsupervised Learning

2. Based on problem category.

3. Feature selection

4.

a. df <- read.csv('IMDB_data.csv')[-2,]
b. unique.genre <- unique(df2$Genre)
c. The above produces a result in vector hence I'm converting the same to a dataframe:
unique.genre.df <- data.frame(unique.genre)
d. sort(unique.genre.df$unique.genre)
e. Initially not possible as it contains empty data points.
Steps are:

1. Convert the empty cells to NA's

```
df$imdbRating <- as.numeric(as.character(df$imdbRating))

df$imdbVotes <- as.numeric(as.character(df$imdbVotes))
```

2. Calculate the mean:

```
df.new <- df$imdbVotes[is.na(df$imdbVotes)] = mean(df$imdbVotes,na.rm = T)
#Load the data once again
df <- read.csv('IMDB_data.csv')[-2,]
#Convert the empty cells to NA's
df$imdbRating <- as.numeric(as.character(df$imdbRating))

df$imdbVotes <- as.numeric(as.character(df$imdbVotes))
```

3. Calculate the median:

```
df.new <- df$imdbVotes[is.na(df$imdbVotes)]median(df$imdbVotes,na       .rm = T)
```

4. Comparing the values, doesn't seem to be fitting good hence proceeding with Knn Imputation:

```
#Load the data once again
imdb.df <- read.csv('IMDB_data.csv')[-2,]
#Convert the empty cells to NA's
```

```
imdb.df$imdbRating <- as.numeric(as.character(imdb.df$imdbRating))

imdb.df$imdbVotes <- as.numeric(as.character(imdb.df$imdbVotes))
```

5. Imputing the missing values using Knn:

```
imdb.df <- knnImputation(imdb.df)
```

Comparing the values imputed with other methods seems like knn imputed values are most accurate and hence proceeding with next step.

6.
```
sq.df.ra.vo <- (imdb.df$imdbRating - imdb.df$imdbVotes)
  sq.df.ra.vo <- data.frame(sq.df.ra.vo)

 sq.df.ra.vo <- sqrt(abs(sq.df.ra.vo))
```

Above throws error since negative values are obtained, hence using
“abs” function to obtain the absolute value along with square rootfunc

7.
```
sq.df.ra.vo <- sqrt(abs(sq.df.ra.vo))
imdb.df <- cbind(imdb.df, sq.df.ra.vo)
```

5. Secondary data refers to data which is available for public and was collected by someone else. Used when situation arises like primary data cannot be obtained at all.
   Example: Census.

6.
```
data <- data$imdbVotes[is.na(data$imdbVotes)] = mean(data$imdbVotes,na.rm = T)
```

7. The data which we obtain and clean is not standardized between a range and may vary sometimes and we wouldn’t be able to feed the same into model as the model’s performance will decrease hence we go for scaling where we obtain the data in a specific range and we feed it into model.

8. Outlier’s are the abnormal data point in the data set. There are several ways to remove them.

   1. Remove them using Boxplot, Grubb’s test or use Outlier package, I,e use these to detect the outliers first and then remove them from dataset.
   2. Convert them to NA’s and impute them using statistical technique or algorithm.

9.
```
data <- data[complete.cases(data),]
```

or

```
data <- na.omit(data)
```

10. An inner join returns only the rows in which the left table have matching keys in the right table..
    A left outer join returns all rows from the left table, and any rows with matching keys
    From right table.