PES University, Bangalore Established under the Karnataka Act No. 16 of 2013

UE21CS342AA2 - Data Analytics

Worksheet 1a : Introduction to R and Exploratory Data Analysis

Richa Shahi - shahiricha2412@gmail.com , Abhay K Iyengar – abzee2002@gmail.com

Exploring Data with R

Prerequisites

This worksheet aims to develop your understanding of summary statistics and basic visualizations through a pragmatic approach.You can download the dataset from here.

Resources

- Check out this beautifully comprehensive resource for everything you need to get started with R.
- This online book provides guided explananations about visualizations in R using the ggplot2 library.

Use the following libraries and read the dataset:

```r
char_preds <- read.csv('movie_dataset.csv')
```

About the Data

To make this worksheet interesting for you all, we have picked this dataset from Kaggle which comprises of the Movies and the metadata associated with it collected using The Movie Database (TMDB). You can download the dataset from here. This dataset is the subset of this Kaggle dataset.

Data Dictionary

```
title - Name or Title of the movie.
budget - The budget of the movie in American Dollar(USD).
genres - The genres for  the entire movie.
id - The identifier for the movie in The Movie Database(TMDB).
original_language - The language associated with the original version of the
                    film.
popularity - Lifetime popularity score of a movie that is impacted by attributes
             like number of votes, number of views, etc.
release_date - The release date of the movie.
revenue - The revenue generated by the movie in American Dollar(USD).
runtime - The duration of the movie in minutes.
```

vote_average - The average of all the votes on the scale of 10.
vote_count - The number of votes for a movie.
director - The director of the movie.

Assignment Submission Format

The following problems are to be completed using the R programming language and should be submitted as a R markdown file (.rmd). Since the dataset is public and many of you students will have the same numerical answers, the grades are allocated on the analysis of the problems and personalized answers within the conclusion section.

Preliminary Guided Exercises

Make sure you have the R programming language installed on your system. It is also recommended to make sure RStudio, the popular IDE for R, is installed. RStudio provides a lot of useful functionality like R markdown, a script editor and GitHub integration. Use RStudio Projects as a great way of keeping each week's assignment work organized.

1. **Data Import**

   To import data from CSV files into a DataFrame:

```r
data <- read.csv('movie_dataset.csv', header=TRUE)
```

The header = TRUE argument specifies that the first row of your data contains the variable names. If thi is not the case you can specify header = FALSE (this is the default value so you can omit this argument entirely).

2. **Compact Summary**

   Use the str() function to return a compact and informative summary of the DataFrame.

```r
str(data)
```

```
## 'data.frame':    4041 obs. of  12 variables:
##  $ budget           : num  2.37e+08 3.00e+08 2.45e+08 2.50e+08 2.60e+08 2.58e+08 2.60e+08 2.80e+08 2
##  $ genres           : chr  "Action Adventure Fantasy Science-Fiction" "Adventure Fantasy Action" "Ac
##  $ id               : int  19995 285 206647 49026 49529 559 38757 99861 767 209112 ...
##  $ original_language: chr  "en" "en" "en" "en" ...
##  $ popularity       : num  150.4 139.1 107.4 112.3 43.9 ...
##  $ release_date     : chr  "10-12-2009" "19-05-2007" "26-10-2015" "16-07-2012" ...
##  $ revenue          : num  2.79e+09 9.61e+08 8.81e+08 1.08e+09 2.84e+08 ...
##  $ runtime          : num  162 169 148 165 132 139 100 141 153 151 ...
##  $ title            : chr  "Avatar" "Pirates of the Caribbean: At World's End" "Spectre" "The Dark Kr
##  $ vote_average     : num  7.2 6.9 6.3 7.6 6.1 5.9 7.4 7.3 7.4 5.7 ...
##  $ vote_count       : int  11800 4500 4466 9106 2124 3576 3330 6767 5293 7004 ...
##  $ director         : chr  "James Cameron" "Gore Verbinski" "Sam Mendes" "Christopher Nolan" ...
```

Here we see that data is a 'data.frame' object which contains 4041 rows and 12 variables (columns). Each the variables are listed along with their data class and the first 10 values.

3. **Summary Statistics**

   To access the data in any of the variables (columns) in our data frame we can use the $ notation. Indexing in R starts at 1, which means the first element is at index 1. Access the first 10 values of the title column:

```
data$title[1:10]
```

```
##  [1] "Avatar"
##  [2] "Pirates of the Caribbean: At World’s End"
##  [3] "Spectre"
##  [4] "The Dark Knight Rises"
##  [5] "John Carter"
##  [6] "Spider-Man 3"
##  [7] "Tangled"
##  [8] "Avengers: Age of Ultron"
##  [9] "Harry Potter and the Half-Blood Prince"
## [10] "Batman v Superman: Dawn of Justice"
```

```
We can assign a column to another variable and calculate a mean of a numeric variable or get a summary
a variable using the summary() function.
```

```
movie_names <- data$title
summary(movie_names)
```

```
##    Length     Class      Mode 
##      4041 character character
```

```
movie_budget <- data$budget
mean(movie_budget)
```

```
## [1] 32853716
```

```
summary(movie_budget)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
##         0   3000000  18000000  32853716  45000000 380000000
```

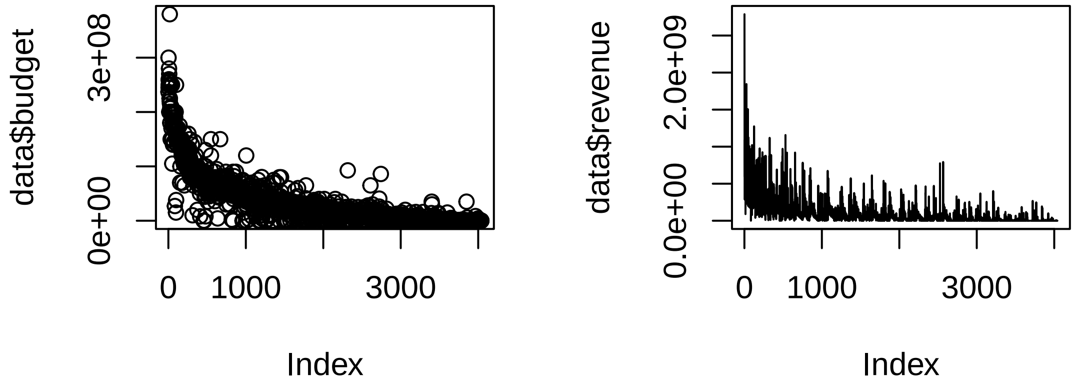
Notice how the behavior of the summary function changes with different types of variables. Let’s now try to explore how we can visualize our data.

4. **Scatter Plots and Line Plots**

   The most common high level function used to produce plots in R is the plot function.

```
par(mfrow = c(1,2)) # To plot different plots in the same row

plot(data$budget, type="p") # scatter plot
plot(data$revenue, type="l") # line plot
```



The horizontal axis in these scatter plots represents the index or the row number of data point.

5. **Sorting a data frame**

   To sort a dataframe with respect to a column we can use the order() function. Let us sort the dataframe to get the top 10 highest grossing movies.

```
sorted_data <- data[order(data$revenue, decreasing = TRUE), ] # To sort in descending order

# The head function is used to get the first 10 rows
top_10_rows <- head(sorted_data, n = 10)
```

6. **Column Transformation**

   Highest Revenue might not be the right indicator for a successful movie. So lets plot the ROI (Return on Investment for all movies)

   ROI = Net Return/Cost of Investment

```
data$ROI = data$revenue / data$budget

# Print the first 5 rows with their title and ROI
data[1:5, c("title", "ROI")]
```

```
##                                       title       ROI
## 1                                    Avatar 11.763566
## 2 Pirates of the Caribbean: At World's End  3.203333
## 3                                   Spectre  3.594590
## 4                     The Dark Knight Rises  4.339756
## 5                               John Carter  1.092843
```

Next, you can sort the data frame with respect to ROI to get movies with highest returns.

7. **Data Pre-processing**

   A lot of times real-world datasets are not curated and cleaned. Values are not stored in proper formats and hence requires cleaning and appropriate transformation before the data is suitable for analysis. In our case we see that the genre is stored as a string. Lets us split the string to get all genre labels.

```r
# Convert the space-separated string of genres to a list of genres for each movie
data$genres <- strsplit(data$genres, " ")

# Extract the individual genres and count their occurrences
label_counts <- table(unlist(data$genres)) # label_counts will be a data frame with "Var1" and "Freq"

# Sort the counts in descending order
label_counts <- sort(label_counts, decreasing = TRUE)
```

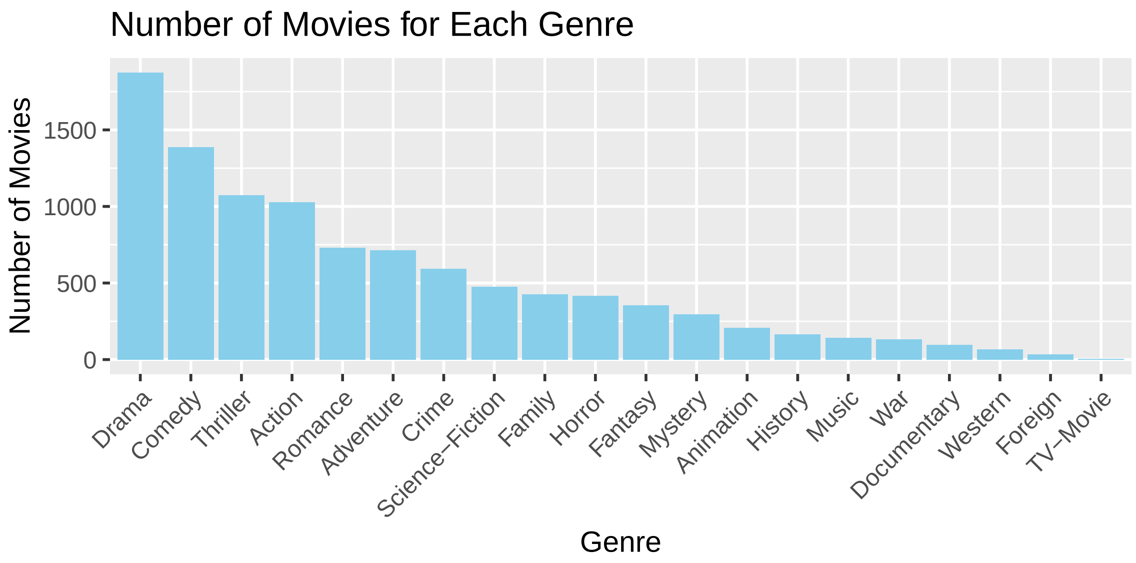
8. **Using the ggplot2 Library**

```r
  # Load the ggplot2 library
  library(ggplot2)

  # Make sure you have label_counts data frame with "Var1" and "Freq" columns

  # Convert the table object to a data frame
  label_counts_df <- as.data.frame(label_counts)

  # Plot the bar chart using ggplot2
  ggplot(label_counts_df, aes(x = Var1, y = Freq)) +
    geom_bar(stat = "identity", fill = "skyblue") +
    labs(x = "Genre", y = "Number of Movies", title = "Number of Movies for Each Genre") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Nearly half of the movies in the dataset are tagged as Drama. Also note that one movie could belong to multiple genres.

Problems

The problems for this part of the worksheet are:

- *Problem 1*
- *Problem 2*
- *Problem 3*
- *Problem 4*
- *Problem 5*
- *Problem 6*

Problems

Problem 1

Get the summary statistics (mean, median, min, max, 1st quartile, 3rd quartile and standard deviation). Calculate these only for the numerical columns. What can you determine from the summary statistics? What summary statistics can be useful for categorical columns? Classify all the variables/columns into their types of data attributes (nominal, ordinal, interval, ratio).

Problem 2

Investigate the data set for missing values. Also classify the missingness as MCAR, MAR or MNAR. Recommend ways to replace missing values in the dataset and apply them for revenue, budget and runtime columns.

Hint: Make sure to capture data from both, missing values in numeric fields and empty strings in descriptive fields. Convert all missing placeholders to type NA. Look at the distribution of the dataset to classify the type of missing values.

Problem 3

Analyze the spread of the data set along years. How number of movie releases have changed over the years ?

Problem 4

Create a horizontal box plot using the column “runtime”. What inferences can you make from this box and whisker plot? Comment on the skew of the runtime field (visual inspection is enough).

Problem 5

Analyze the top 20 titles with highest budget, revenue and ROI. Plot a horizontal bar graph for all three metrics in each case. What analysis can you make by looking at these graphs? What kind of movies attracts the highest investments and do they promise a better ROI ?

Problem 6

Put yourself in the shoes of a production house. You want to produce the next big blockbuster. Plot the ROI, revenue and budget across genres to finalize the genre of your upcoming movie as you did in the previous problem. Elaborate your answers with proper explanation. Since one movie can fall in multiple genre categories, you are free to choose a combination. You can also understand how the popularity of different genres have changed along the years. Do provide a nice name to your movie and your dream cast ;)