|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| Katy – R and Power BI project  Data Technician – Manchester Cohort April 2023 |
| |  |  |  | | --- | --- | --- | | Katy | 4/6/23 | R studio and Power BI | |

Contents

[The aim for this Project 2](#_Toc131674099)

[Step 1: Initial Exploratory Analysis 2](#_Toc131674100)

[Step 2: Clean the data 2](#_Toc131674101)

[Step 2.1: Outlier removal 3](#_Toc131674102)

[Step 3: Exploratory Data Analysis 4](#_Toc131674103)

[Step 4: Export data 5](#_Toc131674104)

[Step 5: Create Power BI Dashboard 5](#_Toc131674105)

[My reflection on this assignment. 9](#_Toc131674106)

## The aim for this Project

To analyze the performance of Hollywood movies using data which includes

Title, genre, studio, profitability and ratings for movies released 2007 -2011.

### Step 1: Initial Exploratory Analysis

Within this step, I loaded data, looked at the data and load the library. Once that was completed, I then imported the library and

checked the data types.

The R language I used for these commands in Step 1 are:

|  |  |
| --- | --- |
| Load data | Df<-read.csv(<https://public.tableau.com/app/sample-data/HollywoodsMostProfitableStories.csv>) |
| Take a look at the data | View (df) |
| Load the library | install.packages(“tidyverse”) |
| Import library | Install.packages(“tidyverse”) |
| Check the data types | Str (df) |

### Step 2: Clean the data

During this step, I checked for missing values, dropped missing values and I checked to ensure that the rows had been removed.I checked for

duplicates, rounded off the values to 2 places and then viewed the dataframe.

The R language I used for these commands in Step 2 are:

|  |  |
| --- | --- |
| Check for missing values | colSums(is.na(df)) |
| Drop missing values | df <- na.omit(df) |
| Check to make sure that the rows  have been removed | colSums(is.na(df)) |
| Check for duplicates | dim(df[duplicated(df$Film),])[1] |
| Round off values to 2 places | df$Profitability <-round(df$Profitability ,digit=2) |
| View(df) | dim(df) |

### Step 2.1: Outlier removal

Within this step, I checked for outliers using a boxplot by using the command:

* library(ggplot2)

I then created a boxplot that highlights the outliers, using this command:

* ggplot(df, aes(x=Profitability, y=Worldwide.Gross)) + geom\_boxplot(outlier.colour = "red", outlier.shape = 1)+ scale\_x\_continuous(labels = scales::comma)+coord\_cartesian(ylim = c(0, 1000))

This is what i successfully outputted, within R.

Chart

Description automatically generated

The next stage within this step was to Remove outliers in ‘Profitability’ with the following commands:

* Q1 <- quantile(df$Profitability, .25)
* Q3 <- quantile(df$Profitability, .75)
* IQR <- IQR(df$Profitability)
* no\_outliers <- subset(df, df$Profitability> (Q1 - 1.5\*IQR) & df$Profitability< (Q3 + 1.5\*IQR))
* dim(no\_outliers)

Finally, I removed the outliers in ‘Worldwide.Gross’ using:

* Q1 <- quantile(no\_outliers$Worldwide.Gross, .25)
* Q3 <- quantile(no\_outliers$Worldwide.Gross, .75)
* IQR <- IQR(no\_outliers$Worldwide.Gross)
* df1 <- subset(no\_outliers, no\_outliers$Worldwide.Gross> (Q1 - 1.5\*IQR) & no\_outliers$Worldwide.Gross< (Q3 + 1.5\*IQR))
* dim(df1)

### Step 3: Exploratory Data Analysis

I summarised the statistics/Univariates the Analysis by using:

* summary(df1)

Chart, scatter chart

Description automatically generatedIt was then time to produce a scatterplot. I used the command:

* ggplot(df1, aes

(x=Lead.Studio, y=Rotten.

Tomatoes..)) + geom\_point

()+ scale\_y\_continuous

(labels = scales::comma)

+coord\_cartesian(ylim =

c(0, 110))+theme(axis.text.x

= element\_text

(angle = 90))

This is was was outputted:

Finally, i produced a barchart within R using the command:

* ggplot(df1, aes(x=Year)) + geom\_bar()

Chart, bar chart

Description automatically generated

### Step 4: Export data

To export the data, I ran this command:

* write.csv(df1,”clean\_df.csv”)

### Step 5: Create Power BI Dashboard

I was given some advice and tips on what the client wanted to see. They

would prefer to have the visuals blue, green and brown – which reflects

their company colours. I was also asked to provide specific visuals as

follows:

1. The average Rotten Tomatoes ratings of each genre:

Chart, bar chart

Description automatically generated

1. The number of movies produced per year:

Chart, pie chart

Description automatically generated

1. The audience score for each film:

A screenshot of a computer

Description automatically generated with medium confidence

1. The profitability per studio:

Chart

Description automatically generated

1. The worldwide gross per genre:

Diagram, pie chart

Description automatically generated

I also took it upon myself to make extra visuals, which were not requested by the client.

I chose to visualise the Profitability by genre and Year:

Chart, waterfall chart

Description automatically generated

I also choose to visualise the Average audience score on Rotten Tomatoes between 2007-2011:

Text

Description automatically generated

Here is a visual of my Dashboard – as a whole:

Graphical user interface, application

Description automatically generated

### My reflection on this assignment.

Throughout my Data Bootcamp, I have come to realise I enjoy profusly Excel, SQL and Power BI. I have found using R studio very interesting and it has unlocked another snippet of a Data tool. I feel as though I have excelled from my first visual and I am very proud to present this assignment to you. I hope you enjoyed reading it as much as I have enjoyed compiling it.