# Analysis

#Load data:

df<- read.csv("/cloud/project/HollywoodsMostProfitableStories.csv")

#Take a look at the data:

View(hmp)

#Load library:

install.packages("tidyverse")

#Import library

library(tidyverse)

# Check data types:

str(hmp)

# Check for missing values:

colSums(is.na(hmp))

#Drop missing values

hmp<-na.omit(hmp)

# check to make sure that the rows have been removed

colSums(is.na(hmp))

dim(hmp)

#Check for duplicates

hmp <- hmp[!duplicated(hmp$Film), ]

dim(hmp)

#round off values to 2 places

hmp$Profitability <- round(hmp$Profitability ,digit=2)

hmp$Worldwide.Gross <- round(hmp$Worldwide.Gross ,digit=2)

dim(hmp)

#Check for outliers using a boxplot

library(ggplot2)

#Create a boxplot that highlights the outliers

ggplot(hmp,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))

#Remove outliers in 'Profitability'

Q1 <- quantile(hmp$Profitability, .25)

Q3 <- quantile(hmp$Profitability, .75)

IQR <- IQR(hmp$Profitability)

no\_outliers <- subset(hmp, hmp$Profitability> (Q1 - 1.5\*IQR) & hmp$Profitability< (Q3 + 1.5\*IQR))

dim(no\_outliers)

ggplot(no\_outliers,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))

# Remove outliers in 'Worldwide.Gross'

Q1 <- quantile(no\_outliers$Worldwide.Gross, .25)

Q3 <- quantile(no\_outliers$Worldwide.Gross, .75)

IQR <- IQR(no\_outliers$Worldwide.Gross)

hmp1 <- subset(no\_outliers, no\_outliers$Worldwide.Gross> (Q1 - 1.5\*IQR) & no\_outliers$Worldwide.Gross< (Q3 + 1.5\*IQR))

dim(hmp1)

#Summary Statistics/Univariate Analysis:

summary(hmp1)

#bivariate analysis

#scatterplot

ggplot(hmp1, 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))

#bar chart

ggplot(hmp1, aes(x=Year)) + geom\_bar()

#Export clean data

write.csv(hmp1, "clean\_hmp.csv")