Assignment 2

For this Assignment, you will use the attached dataset. The goal of this assignment is to practice visualization tools in R.

library(ggplot2)  
library(plyr)  
library(foreach)

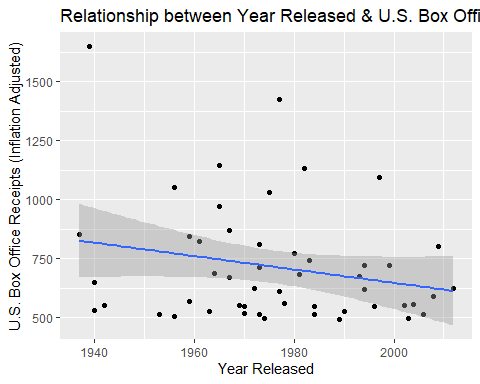
1. Load the data as a data frame. (1 point)

data <- read.csv("Top50Movies.csv")  
  
# Explicitly picking up all the numeric columns  
sub\_data = data[,c(3,4,5,8,9,10)]  
  
# Removes , in the values and converts them to numeric values. Replaces -- with NA  
convertToNumeric = function(col\_data){  
 col\_data = as.array(col\_data)  
 new\_col\_data = aaply(.data = col\_data, .margins = c(1), .fun = function(value){  
 value = gsub("[,]","", value )  
 value = gsub("[--]",NA, value )  
 value = as.numeric(value)  
 })  
}  
  
# Calling the convertToNumeric for all the needed columns  
budgetInfAdj = convertToNumeric(col\_data = sub\_data$Budget..Inflation.Adjusted.Millions...)  
worldInfAdj = convertToNumeric(col\_data = sub\_data$World.Box.Office.Receipts..Inflation.Adjusted.Millions...)  
usInfAdj = convertToNumeric(col\_data = sub\_data$U.S..Box.Office.Receipts..Inflation.Adjusted.Millions...)  
budgetNonInfAdj = convertToNumeric(col\_data = sub\_data$Budget..Non.Inflation.Adjusted.Millions...)  
worldNonInfAdj = convertToNumeric(col\_data = sub\_data$World.Box.Office.Receipts..Non.Inflation.Adjusted.Millions...)  
usNonInfAdj = convertToNumeric(col\_data = sub\_data$U.S..Box.Office.Receipts..Non.Inflation.Adjusted.Millions...)  
  
  
# Replacing the old non-numeric columns with transformed numeric columns  
keep <- names(data)  
keep <- keep[-c(3,4,5,8,9,10)]  
new\_data <- subset(data, select = keep)  
new\_data = cbind(new\_data, budgetInfAdj, worldInfAdj, usInfAdj, budgetNonInfAdj, worldNonInfAdj, usNonInfAdj)  
head(new\_data)

## Title Year.Released Rating Genre budgetInfAdj  
## 1 Gone With the Wind 1939 G Drama 13  
## 2 Star Wars 1977 PG SciFi/Fantasy 20  
## 3 The Sound of Music 1965 G Musical NA  
## 4 E.T. 1982 PG SciFi/Fantasy NA  
## 5 Titanic 1997 PG-13 Drama 100  
## 6 The Ten Commandments 1956 G Drama 184  
## worldInfAdj usInfAdj budgetNonInfAdj worldNonInfAdj usNonInfAdj  
## 1 3242 1650 3 391 199  
## 2 2468 1426 11 798 461  
## 3 1145 1145 NA 163 163  
## 4 1970 1132 NA 757 435  
## 5 3636 1096 200 2185 659  
## 6 1053 1053 14 80 80

1. Explore the relationship between the year released and inflation-adjusted U.S. box office receipts. Which plotting tool should you use for this purpose? Explain your findings. (2 points)

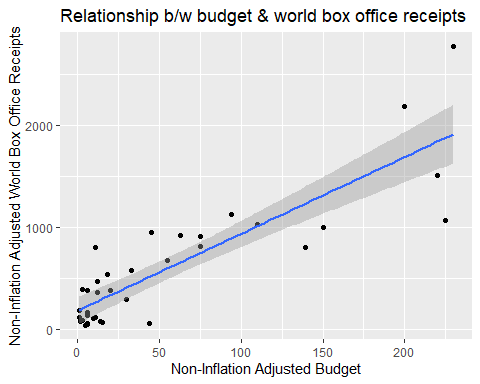
ggplot(data = new\_data, mapping = aes(x = Year.Released, y = usInfAdj)) +  
 # geom\_line(na.rm = TRUE)  
 geom\_point(na.rm = TRUE) +  
 geom\_smooth(method = "lm") +  
 labs(title="Relationship between Year Released & U.S. Box Office Receipts", x="Year Released", y="U.S. Box Office Receipts (Inflation Adjusted)")



# Scatter plot with trend line is one of the appropriate plots that can be used. Through the graph we can understand that there is a weak negative linear relationship between "Inflation adjusted U.S. Box office Receipts" and time. This means, the inflation-adjusted U.S. Box Office Receipts slightly declined with time.

1. Explore the relationship between noninflation-adjusted budget and noninflation-adjusted world box office receipts. Explain your findings. (2 points)

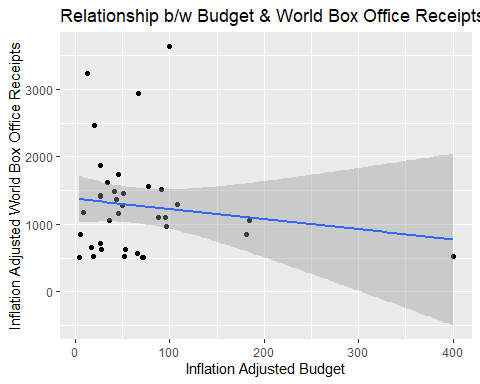
ggplot(data = new\_data, mapping = aes(x = budgetNonInfAdj, y = worldNonInfAdj)) +  
 geom\_point(na.rm = TRUE) +  
 geom\_smooth(method="lm", na.rm=TRUE) +  
 labs(title="Relationship b/w budget & world box office receipts", x="Non-Inflation Adjusted Budget", y="Non-Inflation Adjusted World Box Office Receipts")



# Through the graph we can understand that there is a strong positive linear relationship between Non-Inflation-Adjusted World Box office Receipts and Non-Inflation adjusted Budget. This means, the non-inflation-adjusted World Box Office Receipts increased linearly with increase in non-inflation-adjusted budget.

1. Explore the relationship between inflation-adjusted budget and inflation-adjusted world box office receipts. Explain your findings. (2 points)

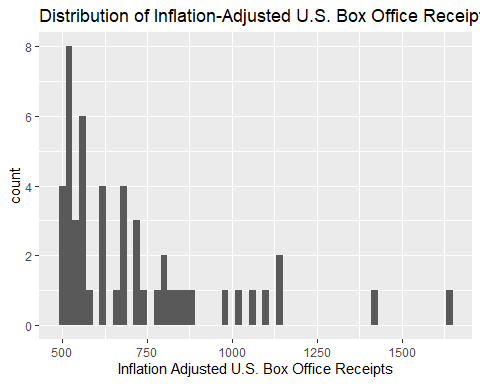
ggplot(data = new\_data, mapping = aes(x = budgetInfAdj, y = worldInfAdj)) +  
 geom\_point(na.rm = TRUE) +  
 geom\_smooth(method="lm", na.rm=TRUE) +  
 labs(title="Relationship b/w Budget & World Box Office Receipts", x="Inflation Adjusted Budget", y="Inflation Adjusted World Box Office Receipts")



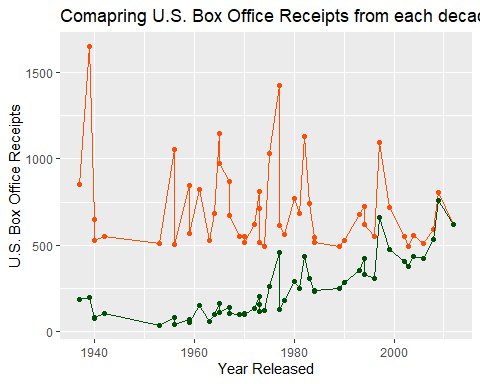
# Through the graph we can understand that there is No linear relationship (or a weak negative linear relationship) between "Inflation-Adjusted World Box office Receipts" and Inflation-Adjusted Budget. This means, the Inflation-Adjusted World Box Office Receipts remained almost same even though inflation-adjusted budget increased.

1. Use a proper plotting tool to how the distribution of inflation-adjusted U.S. box office receipts. (2 points)

ggplot(data = new\_data, mapping = aes(x = usInfAdj)) +  
 geom\_histogram(na.rm = TRUE, binwidth = 20) +  
 labs(title="Distribution of Inflation-Adjusted U.S. Box Office Receipts", x="Inflation Adjusted U.S. Box Office Receipts")

 6. Compare U.S. Box Office Receipts from each decade. Which plotting tool can be used here? Explain your findings. ( 2 points)

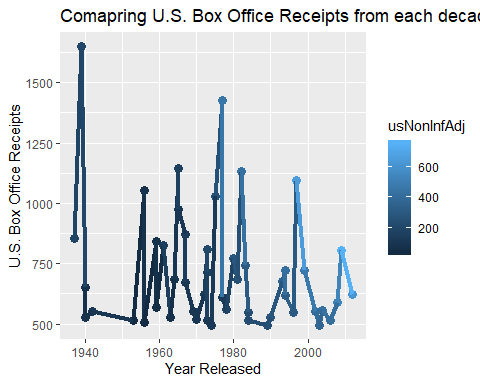
ggplot(data = new\_data, mapping = aes(x = Year.Released)) +  
 geom\_line( mapping = aes(y = usInfAdj), color = "#FC4E07", na.rm = TRUE) +  
 geom\_point(mapping = aes(y = usInfAdj), color = "#FC4E07", na.rm = TRUE) +  
 geom\_line( mapping = aes(y = usNonInfAdj), color = "#004E07", na.rm = TRUE) +  
 geom\_point(mapping = aes(y = usNonInfAdj), color = "#004E07", na.rm = TRUE) +  
 labs(title="Comapring U.S. Box Office Receipts from each decade", x="Year Released", y = "U.S. Box Office Receipts")



# Orange Graph depicts Inflation-Adjusted U.S Box Office Receipts  
  
# Decade-wise Observations  
# 1940 - 1950 - Overall Decreased  
# 1950 - 1960 - Remained Same  
# 1960 - 1970 - Overall Slight Increase (Incresaed & Decreased)  
# 1970 - 1980 - Overall Same (Incresaed & Decreased)  
# 1980 - 1990 - Overall Decreased   
# 1990 - 2000 - Overall Increased   
# 2000 - 2010 - Overall Increased  
# 2010 - 2020 - Decresaed  
  
# Green Graph depicts Non-Inflation-Adjusted U.S Box Office Receipts  
  
# Decade-wise Observations  
# 1940 - 1950 - Remained almost same  
# 1950 - 1960 - Remained almost same  
# 1960 - 1970 - Slightly Increased / Remained almost same  
# 1970 - 1980 - Overall Increased  
# 1980 - 1990 - Remained almost same (Increased and Decreased)  
# 1990 - 2000 - Overall Incresed  
# 2000 - 2010 - Overall Incresed  
# 2010 - 2020 - Decreased

1. Show the temporal pattern of U.S. Box Office Receipts. Which plotting tool can be used here. Explain your findings. (2 points)

ggplot(data = new\_data, mapping = aes(x = Year.Released, y = usInfAdj, col = usNonInfAdj)) +  
 geom\_point(size = 3) +   
 geom\_line(size = 1.5) +   
 labs(title="Comapring U.S. Box Office Receipts from each decade", x="Year Released", y = "U.S. Box Office Receipts")

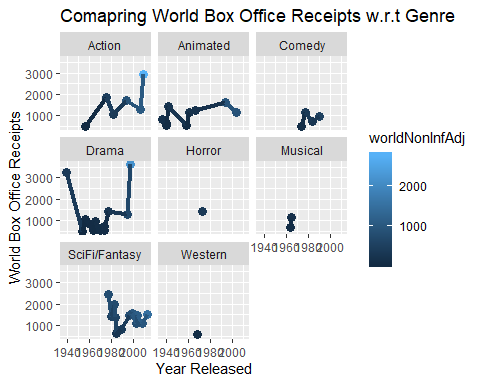


# Temporal Patttern suggests that even though Non-Inflation-Adjusted U.S. Box Office Receipts suggested good increase, Inflation-adjusted U.S. Box Office Receipts suggested a clear decline in receipts.

1. Explore the data yourself. Share an interesting finding that has not been explored above. (2 points)

ggplot(data = new\_data, mapping = aes(x = Year.Released, y = worldInfAdj, col = worldNonInfAdj)) +  
 geom\_point(size = 3) +   
 geom\_line(size = 1.5) +   
 labs(title="Comapring World Box Office Receipts w.r.t Genre", x="Year Released", y = "World Box Office Receipts") +   
 facet\_wrap(~Genre)

## geom\_path: Each group consists of only one observation. Do you need to  
## adjust the group aesthetic?  
## geom\_path: Each group consists of only one observation. Do you need to  
## adjust the group aesthetic?



# Observing how each Genre impacted the world receipts over decades, we can understand that:  
# 1. Action : increased and in trend  
# 2. Animated : Remained same and in trend  
# 3. Comedy : Increased but after 1990, not many comedy movies  
# 4. Drama : Declined in the middle but overall same. Not many drama movies made after 2000  
# 5. Horror : Only one observation and not in trend  
# 6. Musical : Only two observations and not in trend  
# 7. SciFi/Fantasy : Declined initially later picked up. Overall declined and in trend  
# 8. Western : Only one observation