IT 497 Exam 2 Question 3

Agyemang Eric

# Question 3.   
library(tidyverse)

## -- Attaching packages --------------------------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.3 v dplyr 1.0.2  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(readr)  
  
# a. Read in the following Horror Movie data:  
horror\_movies <-read\_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2019/2019-10-22/horror\_movies.csv")

## Parsed with column specification:  
## cols(  
## title = col\_character(),  
## genres = col\_character(),  
## release\_date = col\_character(),  
## release\_country = col\_character(),  
## movie\_rating = col\_character(),  
## review\_rating = col\_double(),  
## movie\_run\_time = col\_character(),  
## plot = col\_character(),  
## cast = col\_character(),  
## language = col\_character(),  
## filming\_locations = col\_character(),  
## budget = col\_character()  
## )

# b. Examine the data (look at head and tail)  
# Head  
head(horror\_movies)

## # A tibble: 6 x 12  
## title genres release\_date release\_country movie\_rating review\_rating  
## <chr> <chr> <chr> <chr> <chr> <dbl>  
## 1 Gut ~ Drama~ 26-Oct-12 USA <NA> 3.9  
## 2 The ~ Horror 13-Jan-17 USA <NA> NA   
## 3 Slee~ Horror 21-Oct-17 Canada <NA> NA   
## 4 Trea~ Comed~ 23-Apr-13 USA NOT RATED 3.7  
## 5 Infi~ Crime~ 10-Apr-15 USA <NA> 5.8  
## 6 In E~ Horro~ 2017 UK <NA> NA   
## # ... with 6 more variables: movie\_run\_time <chr>, plot <chr>, cast <chr>,  
## # language <chr>, filming\_locations <chr>, budget <chr>

# Tail  
tail(horror\_movies)

## # A tibble: 6 x 12  
## title genres release\_date release\_country movie\_rating review\_rating  
## <chr> <chr> <chr> <chr> <chr> <dbl>  
## 1 13 S~ Horro~ 11-Apr-14 Poland R 6.3  
## 2 Vict~ Drama~ 25-Nov-15 USA PG-13 6   
## 3 The ~ Horror 9-Oct-15 USA UNRATED 4.6  
## 4 Talo~ Horror 13-Oct-17 USA <NA> 4.6  
## 5 Bloo~ Actio~ 17-Mar-12 Japan R 3   
## 6 13 C~ Drama~ 15-Apr-16 USA <NA> 5.1  
## # ... with 6 more variables: movie\_run\_time <chr>, plot <chr>, cast <chr>,  
## # language <chr>, filming\_locations <chr>, budget <chr>

# c. Examine the structure (str) of the data  
str(horror\_movies)

## tibble [3,328 x 12] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ title : chr [1:3328] "Gut (2012)" "The Haunting of Mia Moss (2017)" "Sleepwalking (2017)" "Treasure Chest of Horrors II (2013)" ...  
## $ genres : chr [1:3328] "Drama| Horror| Thriller" "Horror" "Horror" "Comedy| Horror| Thriller" ...  
## $ release\_date : chr [1:3328] "26-Oct-12" "13-Jan-17" "21-Oct-17" "23-Apr-13" ...  
## $ release\_country : chr [1:3328] "USA" "USA" "Canada" "USA" ...  
## $ movie\_rating : chr [1:3328] NA NA NA "NOT RATED" ...  
## $ review\_rating : num [1:3328] 3.9 NA NA 3.7 5.8 NA 5.1 6.5 4.6 5.4 ...  
## $ movie\_run\_time : chr [1:3328] "91 min" NA NA "82 min" ...  
## $ plot : chr [1:3328] "Directed by Elias. With Jason Vail, Nicholas Wilder, Sarah Schoofs, Kirstianna Mueller. Family man Tom has seen"| \_\_truncated\_\_ "Directed by Jake Zelch. With Nicola Fiore, Brinke Stevens, Curtis Carnahan, Jake Zelch." "Directed by David Briggs. With Alysia Topol, Anthony Makela, Kelsi Ashley, Patrick J. Carew." "Directed by M. Kelley, Shawn C. Phillips, Alex Powers. With Veronica Ricci, Nicholas Adam Clark, James Cullen B"| \_\_truncated\_\_ ...  
## $ cast : chr [1:3328] "Jason Vail|Nicholas Wilder|Sarah Schoofs|Kirstianna Mueller|Kaitlyn Mueller|Angie Bullaro|Ria Burns-Wilder|Mile"| \_\_truncated\_\_ "Nicola Fiore|Brinke Stevens|Curtis Carnahan|Jake Zelch|Haley Bordelon|Asalee Biagioli|Emily Christina|Erin Felt"| \_\_truncated\_\_ "Alysia Topol|Anthony Makela|Kelsi Ashley|Patrick J. Carew|Chiv Chivaul|Starlotte Dresen|Melanie Durette|Emilie "| \_\_truncated\_\_ "Veronica Ricci|Nicholas Adam Clark|James Cullen Bressack|Nick Waugh|Shawn C. Phillips|Brent Buser|Miles Dougal|"| \_\_truncated\_\_ ...  
## $ language : chr [1:3328] "English" "English" "English" "English" ...  
## $ filming\_locations: chr [1:3328] "New York, USA" NA "Sudbury, Ontario, Canada" "Baltimore, Maryland, USA" ...  
## $ budget : chr [1:3328] NA "$30,000" NA NA ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. title = col\_character(),  
## .. genres = col\_character(),  
## .. release\_date = col\_character(),  
## .. release\_country = col\_character(),  
## .. movie\_rating = col\_character(),  
## .. review\_rating = col\_double(),  
## .. movie\_run\_time = col\_character(),  
## .. plot = col\_character(),  
## .. cast = col\_character(),  
## .. language = col\_character(),  
## .. filming\_locations = col\_character(),  
## .. budget = col\_character()  
## .. )

# d. Use select to obtain the columns title, release\_country, review\_rating, and budget.  
  
horror\_movies\_1 <- select(horror\_movies, title, release\_country, review\_rating, budget)  
horror\_movies\_1

## # A tibble: 3,328 x 4  
## title release\_country review\_rating budget   
## <chr> <chr> <dbl> <chr>   
## 1 Gut (2012) USA 3.9 <NA>   
## 2 The Haunting of Mia Moss (2017) USA NA $30,000   
## 3 Sleepwalking (2017) Canada NA <NA>   
## 4 Treasure Chest of Horrors II (2013) USA 3.7 <NA>   
## 5 Infidus (2015) USA 5.8 <NA>   
## 6 In Extremis (2017) UK NA $3,400,000  
## 7 Ghostlight (2013) USA 5.1 <NA>   
## 8 Parasyte: Part 2 (2015) Japan 6.5 <NA>   
## 9 Stranger in the House (2015) Spain 4.6 <NA>   
## 10 Tutak Tutak Tutiya (2016) India 5.4 <NA>   
## # ... with 3,318 more rows

# e. Omit all other columns  
horror\_movies\_2<- select(horror\_movies, -c(genres, release\_date, movie\_rating, movie\_run\_time, plot, cast, language, filming\_locations))  
horror\_movies\_2

## # A tibble: 3,328 x 4  
## title release\_country review\_rating budget   
## <chr> <chr> <dbl> <chr>   
## 1 Gut (2012) USA 3.9 <NA>   
## 2 The Haunting of Mia Moss (2017) USA NA $30,000   
## 3 Sleepwalking (2017) Canada NA <NA>   
## 4 Treasure Chest of Horrors II (2013) USA 3.7 <NA>   
## 5 Infidus (2015) USA 5.8 <NA>   
## 6 In Extremis (2017) UK NA $3,400,000  
## 7 Ghostlight (2013) USA 5.1 <NA>   
## 8 Parasyte: Part 2 (2015) Japan 6.5 <NA>   
## 9 Stranger in the House (2015) Spain 4.6 <NA>   
## 10 Tutak Tutak Tutiya (2016) India 5.4 <NA>   
## # ... with 3,318 more rows

# f. Obtain the movies released in the USA with a movie rating of either PG-13 or R (Omit movies with other ratings)  
library(dplyr)  
  
movie\_rating\_select <- c("PG-13", "R")  
horror\_movies\_3 <- filter(horror\_movies, release\_country == "USA")%>%filter(movie\_rating%in%movie\_rating\_select)  
horror\_movies\_3

## # A tibble: 368 x 12  
## title genres release\_date release\_country movie\_rating review\_rating  
## <chr> <chr> <chr> <chr> <chr> <dbl>  
## 1 Sept~ Horror 15-Aug-14 USA R 4.1  
## 2 Holl~ Comed~ 19-Apr-13 USA R 6.9  
## 3 Para~ Horro~ 26-Jun-12 USA R 2.2  
## 4 Exit~ Actio~ 5-Nov-13 USA R 3.4  
## 5 Stri~ Comed~ 14-May-13 USA R 4.9  
## 6 Blac~ Actio~ 14-Jul-12 USA PG-13 4.6  
## 7 Dark~ Horror 18-Mar-13 USA R 3.3  
## 8 The ~ Horro~ 1-Feb-13 USA PG-13 3.5  
## 9 Nazi~ Horro~ 5-May-15 USA R 2.5  
## 10 The ~ Horro~ 5-Feb-16 USA R 4.3  
## # ... with 358 more rows, and 6 more variables: movie\_run\_time <chr>,  
## # plot <chr>, cast <chr>, language <chr>, filming\_locations <chr>,  
## # budget <chr>

# g. Omit any rows with missing data in the columns title, release\_country, review\_rating, or budget  
  
good <- complete.cases(horror\_movies\_3)  
  
horror\_movies\_4 <- horror\_movies\_3[good, ]  
horror\_movies\_4

## # A tibble: 174 x 12  
## title genres release\_date release\_country movie\_rating review\_rating  
## <chr> <chr> <chr> <chr> <chr> <dbl>  
## 1 Holl~ Comed~ 19-Apr-13 USA R 6.9  
## 2 Nazi~ Horro~ 5-May-15 USA R 2.5  
## 3 Zomb~ Horro~ 20-Apr-12 USA R 3.2  
## 4 Snow~ Horror 16-Mar-12 USA PG-13 2.2  
## 5 Self~ Horror 17-Sep-13 USA R 3.2  
## 6 Alph~ Actio~ 1-Sep-13 USA R 3.4  
## 7 Disc~ Horror 12-Sep-14 USA R 2.1  
## 8 Kill~ Horro~ 5-Nov-13 USA R 3.1  
## 9 You'~ Horro~ 10-Jul-12 USA R 3.1  
## 10 Dark~ Fanta~ 9-Aug-16 USA R 4.3  
## # ... with 164 more rows, and 6 more variables: movie\_run\_time <chr>,  
## # plot <chr>, cast <chr>, language <chr>, filming\_locations <chr>,  
## # budget <chr>

# h. Omit any rows with budget in any currency other than US Dollars (this is probably the most difficult part of the question).  
  
horror\_movies\_5 <- subset(horror\_movies\_4, startsWith(as.character(budget), "$"))  
horror\_movies\_5

## # A tibble: 165 x 12  
## title genres release\_date release\_country movie\_rating review\_rating  
## <chr> <chr> <chr> <chr> <chr> <dbl>  
## 1 Holl~ Comed~ 19-Apr-13 USA R 6.9  
## 2 Zomb~ Horro~ 20-Apr-12 USA R 3.2  
## 3 Snow~ Horror 16-Mar-12 USA PG-13 2.2  
## 4 Self~ Horror 17-Sep-13 USA R 3.2  
## 5 Alph~ Actio~ 1-Sep-13 USA R 3.4  
## 6 Disc~ Horror 12-Sep-14 USA R 2.1  
## 7 Kill~ Horro~ 5-Nov-13 USA R 3.1  
## 8 You'~ Horro~ 10-Jul-12 USA R 3.1  
## 9 The ~ Drama~ 27-Jan-12 USA R 3.9  
## 10 Evil~ Fanta~ 5-Apr-13 USA R 6.5  
## # ... with 155 more rows, and 6 more variables: movie\_run\_time <chr>,  
## # plot <chr>, cast <chr>, language <chr>, filming\_locations <chr>,  
## # budget <chr>

# i. Omit any rows with budgets less than $500,000  
horror\_movies\_5$budget<-parse\_number(horror\_movies\_5$budget)  
horror\_movies\_5$budget <-as.numeric(horror\_movies\_5$budget)  
horror\_movies\_6 <- horror\_movies\_5[!(horror\_movies\_5$budget < 500000), ]  
horror\_movies\_6

## # A tibble: 156 x 12  
## title genres release\_date release\_country movie\_rating review\_rating  
## <chr> <chr> <chr> <chr> <chr> <dbl>  
## 1 Holl~ Comed~ 19-Apr-13 USA R 6.9  
## 2 Snow~ Horror 16-Mar-12 USA PG-13 2.2  
## 3 Self~ Horror 17-Sep-13 USA R 3.2  
## 4 Kill~ Horro~ 5-Nov-13 USA R 3.1  
## 5 You'~ Horro~ 10-Jul-12 USA R 3.1  
## 6 The ~ Drama~ 27-Jan-12 USA R 3.9  
## 7 Evil~ Fanta~ 5-Apr-13 USA R 6.5  
## 8 Hans~ Actio~ 25-Jan-13 USA R 6.1  
## 9 Livi~ Horror 2013 USA PG-13 5.5  
## 10 Crow~ Horro~ 7-Sep-12 USA R 3.8  
## # ... with 146 more rows, and 6 more variables: movie\_run\_time <chr>,  
## # plot <chr>, cast <chr>, language <chr>, filming\_locations <chr>,  
## # budget <dbl>

# j. Convert movie\_run\_time to numeric. You can do this any way you like, but may want to use parse\_number from the readr package.  
horror\_movies\_6$movie\_run\_time<-parse\_number(horror\_movies\_6$movie\_run\_time)  
horror\_movies\_6

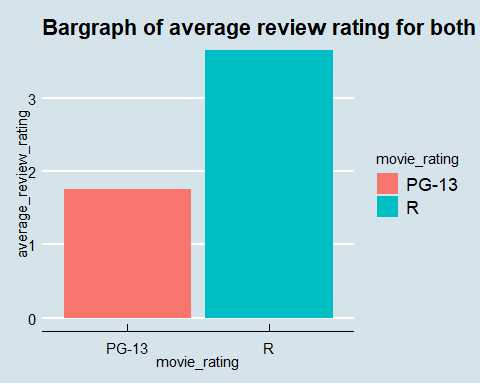
## # A tibble: 156 x 12  
## title genres release\_date release\_country movie\_rating review\_rating  
## <chr> <chr> <chr> <chr> <chr> <dbl>  
## 1 Holl~ Comed~ 19-Apr-13 USA R 6.9  
## 2 Snow~ Horror 16-Mar-12 USA PG-13 2.2  
## 3 Self~ Horror 17-Sep-13 USA R 3.2  
## 4 Kill~ Horro~ 5-Nov-13 USA R 3.1  
## 5 You'~ Horro~ 10-Jul-12 USA R 3.1  
## 6 The ~ Drama~ 27-Jan-12 USA R 3.9  
## 7 Evil~ Fanta~ 5-Apr-13 USA R 6.5  
## 8 Hans~ Actio~ 25-Jan-13 USA R 6.1  
## 9 Livi~ Horror 2013 USA PG-13 5.5  
## 10 Crow~ Horro~ 7-Sep-12 USA R 3.8  
## # ... with 146 more rows, and 6 more variables: movie\_run\_time <dbl>,  
## # plot <chr>, cast <chr>, language <chr>, filming\_locations <chr>,  
## # budget <dbl>

# k1. Using a bargraph (or histogram), show the average review rating for both PG-13 and R movies  
  
# Using the bargraph we first find the average review ratings  
horror\_movies\_7 <- mutate(horror\_movies\_6, average\_review\_rating = (review\_rating/156))

library(ggthemes)

## Warning: package 'ggthemes' was built under R version 4.0.3

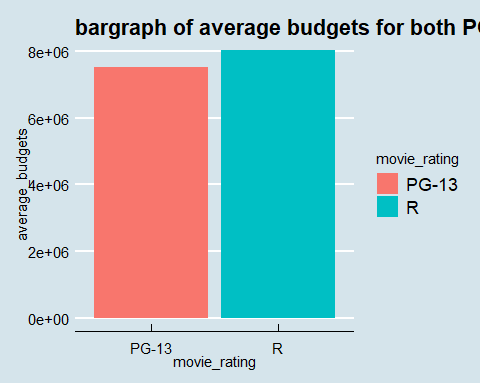
p<-ggplot(data = horror\_movies\_7, aes(x = movie\_rating, y = average\_review\_rating, fill = movie\_rating)) + geom\_bar(stat="identity") + theme\_economist() + ggtitle("Bargraph of average review rating for both PG-13 and R movies")+ theme(legend.position = "right")  
  
p



# K2. Write a two sentence explanation of the graph.  
# This is a bargraph showing the movie ratings (PG-13 movies on the left, and R movies on the right) on the x-axis ploted against the average review rating on the y-axis.   
  
# From this graph, the R movies have the highest average review rating which is almsot twice that of PG-13 movies, while the PG-13 movies is the least average review rating. the legend of this graph is shown at the right.

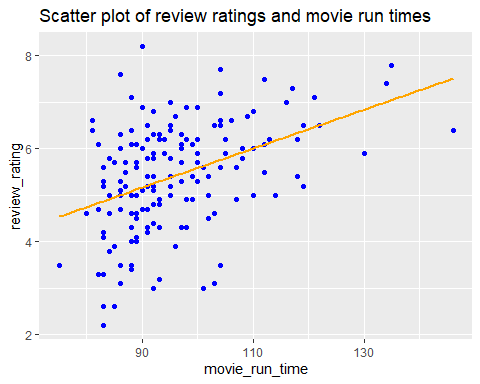
# l1. Using a bargraph (or histogram), show the average budgets for both PG-13 and R movies  
  
# Using the bargraph we first find the average budgets  
horror\_movies\_8 <- mutate(horror\_movies\_7, average\_budgets = (budget/156))

library(ggthemes)  
  
p2<-ggplot(data = horror\_movies\_8, aes(x = movie\_rating, y = average\_budgets, fill = movie\_rating)) + geom\_bar(stat="identity") + theme\_economist() + ggtitle("bargraph of average budgets for both PG-13 and R movies") + theme(legend.position = "right")  
  
p2



# l2. Write a two sentence explanation of the graph.  
  
# This is a bargraph showing the average budget for both PG-13 movies displayed on the left and R movies on the right where both movie rating are plotted on the x-axis and avreage budgets on the y-axis.  
  
# Here, we see that the R movies have a slightly higher budget than that of the PG-13 movies. The legend of this graph is shown at the right position of the graph.

# m1. Find the scatter plot of the review ratings and movie run times.  
  
ggplot(horror\_movies\_8, aes(x = movie\_run\_time, y = review\_rating)) + geom\_point(color="blue") + geom\_smooth(method = lm, formula = y ~ x, se = FALSE, color="orange", fill="blue") + ggtitle("Scatter plot of review ratings and movie run times")



# m2. Does the plot reveal any relationship between the variables? Write a two sentence explanation of the graph.  
  
# No. This plot does not reveal a relationship between the variables.  
  
# From this graph, we can see that the movie run times are plotted on the x-axis and are plotted against the review ratings on the y-axis, with the data points scatterd without a particular patten.   
  
# We can also see that some data points fall close to the line of best fit while most of the points do not fall close to the line. This shows that there is a considereably high vriablility between the data points and the line. These show no patten of relationship existing between the two variables.