Data 607 Homework 2

Josef Waples

9/4/2020

This project will make use of survey data from SurveyMonkey, an SQL relational database, and an R notebook.

Our goal is to collect survey response data concerning opinions on different Brad Pitt movies. We will compare our results to the IMDB database critic ranking of the same movies in order to see whether or not our classmates or the professional movie critics are bigger fans of Brad Pitt.

Our SurveyMonkey survey concerns the following movies. These six movies were randomly selected from a list of Brad Pitt movies found on Wikipedia at the following link: <https://en.wikipedia.org/wiki/Brad_Pitt_filmography> The following movies were selected because for some reason I can’t remember I recently had a conversation about Brad Pitt, who is in my opinion a good actor.

A River Runs Through It (1992) Moneyball (2011) Ad Astra (2019) World War Z (2013) The Curious Case of Benjamin Button (2008) Spy Game (2001)

The survey is available on SurveyMonkey at the following link. <https://www.surveymonkey.com/r/65KXG9V>

The survey scored these movies on a simple scale of 1 (not good) to 5 (extremely good) I specified that a sixth answer “I did not see this movie” in case the respondent had not seen this movie.

Here is a section of SQL code in order to retrieve the surveymonkey data from the SQL database. Full disclosure - The data came out in a nested / web format that I don’t think was exactly SQL format … so I"m not 100% certain if this would be reproducible.

Here is a section of SQL code in order to retrieve the surveymonkey data from the SQL database. Full disclosure - The data came out in a nested / web format that I don’t think was exactly SQL format … so I"m not 100% certain if this would be reproducible.

{r SQL}

CREATE TABLE data (id INT NOT NULL AUTO\_INCREMENT, A\_River\_Runs\_Through\_It VARCHAR(255) NOT NULL, Moneyball VARCHAR(255) NOT NULL, Ad\_Astra VARCHAR(255) NOT NULL, World\_War\_Z VARCHAR(255) NOT NULL, The\_Curious\_Case\_of\_Benjamin\_Button VARCHAR(255) NOT NULL, Spy\_Game INT NOT NULL, PRIMARY KEY (id)); LOAD DATA LOCAL INFILE ‘C:Server 8.0.csv’ INTO TABLE data FIELDS TERMINATED BY ‘,’ LINES TERMINATED BY ‘’ IGNORE 1 ROWS (id, A\_River\_Runs\_Through\_It, Moneyball, Ad\_Astra, World\_War\_Z, The\_Curious\_Case\_of\_Benjamin\_Button, Spy\_Game);

{r SQL meets R, echo=FALSE}

library(RMySQL)

connection <- dbConnect(MySQL(), user=‘root’, password=‘password’, dbname=‘movies’, host=‘lo calhost’) connection MySQLConnection:0,0

moviequery <- dbSendQuery(con, “SELECT \* FROM data”) dbFetch(moviequery)

moviequery <- dbSendQuery(con, “SELECT \* FROM data”) Class\_Movie\_Survey <- dbFetch(moviequery)

# Class\_Movie\_Survey <- read.table(file = ~/Downloads/Class\_Movie\_Survey.xlsx, sep = "\t", header=TRUE)  
library(readxl)  
  
Class\_Movie\_Survey <- read\_excel("~/Downloads/Class\_Movie\_Survey.xlsx")  
  
#taking a look at the data  
head(Class\_Movie\_Survey)

## # A tibble: 6 x 9  
## `What is your n… `What is your g… `What is your a… `A River Runs T… Moneyball  
## <chr> <chr> <chr> <chr> <chr>   
## 1 LeTicia Female 25-34 I haven't seen … 3   
## 2 Dan Male 25-34 I haven't seen … 4   
## 3 Magnus Male 25-34 3 5   
## 4 Omar Male 25-34 I haven't seen … 4   
## 5 Josef Male 25-34 5 I haven'…  
## 6 Bharani Male 25-34 3 I haven'…  
## # … with 4 more variables: `Ad Astra` <chr>, `World War Z` <chr>, `The Curious  
## # Case of Benjamin Button` <chr>, `Spy Game` <chr>

tail(Class\_Movie\_Survey)

## # A tibble: 6 x 9  
## `What is your n… `What is your g… `What is your a… `A River Runs T… Moneyball  
## <chr> <chr> <chr> <chr> <chr>   
## 1 LeTicia Female 25-34 I haven't seen … 3   
## 2 Dan Male 25-34 I haven't seen … 4   
## 3 Magnus Male 25-34 3 5   
## 4 Omar Male 25-34 I haven't seen … 4   
## 5 Josef Male 25-34 5 I haven'…  
## 6 Bharani Male 25-34 3 I haven'…  
## # … with 4 more variables: `Ad Astra` <chr>, `World War Z` <chr>, `The Curious  
## # Case of Benjamin Button` <chr>, `Spy Game` <chr>

length(Class\_Movie\_Survey) # 9 variables

## [1] 9

nrow(Class\_Movie\_Survey) # 6 rows (which equals 6 respondents)

## [1] 6

str(Class\_Movie\_Survey) # looking at the structure, we see all variables are character types

## tibble [6 × 9] (S3: tbl\_df/tbl/data.frame)  
## $ What is your name? : chr [1:6] "LeTicia" "Dan" "Magnus" "Omar" ...  
## $ What is your gender? : chr [1:6] "Female" "Male" "Male" "Male" ...  
## $ What is your age? : chr [1:6] "25-34" "25-34" "25-34" "25-34" ...  
## $ A River Runs Throught It : chr [1:6] "I haven't seen this movie" "I haven't seen this movie" "3" "I haven't seen this movie" ...  
## $ Moneyball : chr [1:6] "3" "4" "5" "4" ...  
## $ Ad Astra : chr [1:6] "4" "4" "4" "I haven't seen this movie" ...  
## $ World War Z : chr [1:6] "5" "3" "I haven't seen this movie" "4" ...  
## $ The Curious Case of Benjamin Button: chr [1:6] "5" "I haven't seen this movie" "I haven't seen this movie" "4" ...  
## $ Spy Game : chr [1:6] "I haven't seen this movie" "I haven't seen this movie" "3" "I haven't seen this movie" ...

# i think this is because the titles of the movies (header column) didn't have underscores  
  
Class\_Movie\_Survey$`A River Runs Throught It` <- as.numeric(Class\_Movie\_Survey$`A River Runs Throught It`)

## Warning: NAs introduced by coercion

Class\_Movie\_Survey$Moneyball <- as.numeric(Class\_Movie\_Survey$Moneyball)

## Warning: NAs introduced by coercion

Class\_Movie\_Survey$`Ad Astra` <- as.numeric(Class\_Movie\_Survey$`Ad Astra`)

## Warning: NAs introduced by coercion

Class\_Movie\_Survey$`World War Z` <- as.numeric(Class\_Movie\_Survey$`World War Z`)

## Warning: NAs introduced by coercion

Class\_Movie\_Survey$`The Curious Case of Benjamin Button` <- as.numeric(Class\_Movie\_Survey$`The Curious Case of Benjamin Button`)

## Warning: NAs introduced by coercion

Class\_Movie\_Survey$`Spy Game` <- as.numeric(Class\_Movie\_Survey$`Spy Game`)

## Warning: NAs introduced by coercion

# I am able to force coreciion the types to numeric  
  
# are there any missing values here?  
any(is.na(Class\_Movie\_Survey)) # TRUE

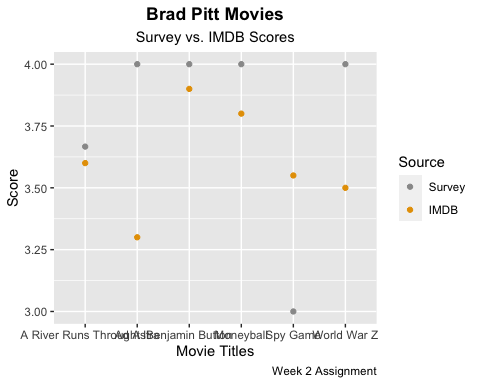
## [1] TRUE

# OK I see that the "I didn't see this movie" can turned into an NA value  
  
# how many NA values are there?  
sum(is.na(Class\_Movie\_Survey)) # there are 14 NA values

## [1] 14

# create a new column that is the sum of the NAs for each respondent  
Class\_Movie\_Survey$total\_na\_count\_respondent <- apply(Class\_Movie\_Survey, 1, function(x) sum(is.na(x)))  
  
  
# Find the average survey score for each movie remembering to remove NA / "I haven't seen this movie" values before creating an average   
  
Average\_Score\_A\_River\_Runs\_Throught\_It <- sum(Class\_Movie\_Survey$`A River Runs Throught It`,na.rm=TRUE) / (nrow(Class\_Movie\_Survey) - sum(is.na(Class\_Movie\_Survey$`A River Runs Throught It`)))  
  
Average\_Score\_Moneyball <- sum(Class\_Movie\_Survey$Moneyball,na.rm=TRUE ) / (nrow(Class\_Movie\_Survey) - sum(is.na(Class\_Movie\_Survey$Moneyball)))  
  
Average\_Score\_Ad\_Astra <- sum(Class\_Movie\_Survey$`Ad Astra`,na.rm=TRUE) / (nrow(Class\_Movie\_Survey) - sum(is.na(Class\_Movie\_Survey$`Ad Astra`)))  
  
Average\_Score\_World\_War\_Z <- sum(Class\_Movie\_Survey$`World War Z`,na.rm=TRUE) / (nrow(Class\_Movie\_Survey) - sum(is.na(Class\_Movie\_Survey$`World War Z`)))  
  
Average\_Score\_The\_Curious\_Case\_of\_Benjamin\_Button <- sum(Class\_Movie\_Survey$`The Curious Case of Benjamin Button`,na.rm=TRUE) / (nrow(Class\_Movie\_Survey) - sum(is.na(Class\_Movie\_Survey$`The Curious Case of Benjamin Button`)))  
  
Average\_Score\_Spy\_Game <- sum(Class\_Movie\_Survey$`Spy Game`,na.rm=TRUE) / (nrow(Class\_Movie\_Survey) - sum(is.na(Class\_Movie\_Survey$`Spy Game`)))  
  
# Create a data frame for our movies + survey scores  
SurveyScoreDateFrame <- data.frame("Movies" = c("A River Runs Throught It", "Moneyball", "Ad Astra", "World War Z", "Benjamin Button", "Spy Game"), "Average Survey Score" = c(Average\_Score\_A\_River\_Runs\_Throught\_It, Average\_Score\_Moneyball, Average\_Score\_Ad\_Astra, Average\_Score\_World\_War\_Z, Average\_Score\_The\_Curious\_Case\_of\_Benjamin\_Button, Average\_Score\_Spy\_Game))  
   
# Find the IMDB score for each movie (using google searches and creating variables)  
IMDB\_Score\_A\_River\_Runs\_Throught\_It <- 7.2  
IMDB\_Score\_Moneyball <- 7.6  
IMDB\_Score\_Ad\_Astra <- 6.6  
IMDB\_Score\_World\_War\_Z <- 7  
IMDB\_Score\_The\_Curious\_Case\_of\_Benjamin\_Button <- 7.8  
IMDB\_Score\_Spy\_Game <- 7.1  
  
  
# Divide each IMDB score by 2 because we scored our movies on a 5-point scale  
IMDB\_Score\_A\_River\_Runs\_Throught\_It <- IMDB\_Score\_A\_River\_Runs\_Throught\_It / 2  
IMDB\_Score\_Moneyball <- IMDB\_Score\_Moneyball / 2  
IMDB\_Score\_Ad\_Astra <- IMDB\_Score\_Ad\_Astra / 2  
IMDB\_Score\_World\_War\_Z <- IMDB\_Score\_World\_War\_Z / 2  
IMDB\_Score\_The\_Curious\_Case\_of\_Benjamin\_Button <- IMDB\_Score\_The\_Curious\_Case\_of\_Benjamin\_Button / 2  
IMDB\_Score\_Spy\_Game <- IMDB\_Score\_Spy\_Game / 2  
  
  
# Create a vecotr of IMDB Scores  
IMDB\_Scores\_Vector <- c(IMDB\_Score\_A\_River\_Runs\_Throught\_It, IMDB\_Score\_Moneyball, IMDB\_Score\_Ad\_Astra, IMDB\_Score\_World\_War\_Z, IMDB\_Score\_The\_Curious\_Case\_of\_Benjamin\_Button, IMDB\_Score\_Spy\_Game)  
   
SurveyScoreDateFrame$IMDB\_Scores <- IMDB\_Scores\_Vector  
  
# in order to graph we need a long data frame instead of a wide one  
  
library(tidyr)  
  
long\_data\_SurveyScoreDateFrame <- gather(SurveyScoreDateFrame, SurveyScoreDateFrame.Average.Survey.Score, IMDB\_Scores, -Movies)  
  
colnames(long\_data\_SurveyScoreDateFrame) <- c("Movies", "Measure", "Score")

library(ggplot2)  
  
ggplot(data = long\_data\_SurveyScoreDateFrame, aes(x = Movies, y = Score, color = Measure)) +   
 geom\_point() +  
 labs(title="Brad Pitt Movies", subtitle = "Survey vs. IMDB Scores") +  
 theme(plot.title = element\_text(hjust = 0.5, face = "bold")) +  
 theme(plot.subtitle = element\_text(hjust = 0.5)) +  
 labs(caption = "Week 2 Assignment") +  
 xlab("Movie Titles") +  
 # scale\_colour\_manual(values = c(Average.Survey.Score = 'turquoise',  
 # IMDB\_Scores = 'tan1')) +  
 scale\_color\_manual(values=c("#999999", "#E69F00"),   
 name="Source",  
 labels=c("Survey", "IMDB"))



ylab("Score")

## $y  
## [1] "Score"  
##   
## attr(,"class")  
## [1] "labels"

# scale\_y\_discrete(limits=c("1","2","3", "4", "5"))

My conclusion here is that the class survey graders are either more generous graders of movies in generally, or else our class is made up of bigger fans of Brad Pitt than our movie critics.