Webscraping

Elizabeth Marge

3/17/2021

## Question 1

library(rvest)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyverse)

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

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.5 v stringr 1.4.0  
## v tidyr 1.1.2 v forcats 0.5.0  
## v readr 1.3.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x readr::guess\_encoding() masks rvest::guess\_encoding()  
## x dplyr::lag() masks stats::lag()

library(repurrrsive)  
library(listviewer)

Link <- "https://www.imdb.com/best-of/highest-rated-movies-of-2019/ls091392448/"  
page = read\_html(Link)  
MovieTitles2019 = page%>%   
 html\_nodes(".lister-item-header a")%>%  
 html\_text()  
  
Movieratings2019 = page%>%   
 html\_nodes(".ratings-imdb-rating strong")%>%  
 html\_text()  
  
moviesdataframe = data.frame(MovieTitles2019, Movieratings2019)  
moviesdataframe

## MovieTitles2019 Movieratings2019  
## 1 Joker 8.4  
## 2 Avengers: Endgame 8.4  
## 3 Parasite 8.6  
## 4 Once Upon a Time... In Hollywood 7.6  
## 5 Toy Story 4 7.8  
## 6 Spider-Man: Far from Home 7.5  
## 7 Cold War 7.6  
## 8 John Wick: Chapter 3 - Parabellum 7.4  
## 9 How to Train Your Dragon: The Hidden World 7.5  
## 10 El Camino: A Breaking Bad Movie 7.3  
## 11 Rocketman 7.3  
## 12 Alita: Battle Angel 7.3  
## 13 Midsommar 7.1  
## 14 Booksmart 7.2  
## 15 Shazam! 7.0  
## 16 Fighting with My Family 7.1  
## 17 Aladdin 6.9  
## 18 Ad Astra 6.5  
## 19 The Lion King 6.9  
## 20 The Dirt 7.0

## Question 2

parse\_number(Movieratings2019)

## [1] 8.4 8.4 8.6 7.6 7.8 7.5 7.6 7.4 7.5 7.3 7.3 7.3 7.1 7.2 7.0 7.1 6.9 6.5 6.9  
## [20] 7.0

parse\_number(Movieratings2019) -> y  
y

## [1] 8.4 8.4 8.6 7.6 7.8 7.5 7.6 7.4 7.5 7.3 7.3 7.3 7.1 7.2 7.0 7.1 6.9 6.5 6.9  
## [20] 7.0

summary(y)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 6.500 7.075 7.300 7.420 7.600 8.600

## a) Min - 6.5  
## b) Max- 8.6  
## c) Q1- 7.075  
## d) Median - 7.300  
## e) Q3- 7.600  
## f) Mean 7.420

## Question 3

boxplot(y)

