Webscraping IMBD Ratings of The Office Episodes

Code

The Office is an extremely popular American Sitcom. The show was based on a British show of the same name, and features a “mockumentary” style in which a crew documents the lives of ordinary American workers at a paper company. The show centers on the Regional Manager “Michael Scott” (Steve Carell). Michael is a clueless boss and is convinced that all his employees love him. The employees, however, are not amused by Michael’s absurd and sometimes offensive antics.

Steve Carell left the Office in Season 7. The show aired for 2 more seasons after his departure. Using data that I webscraped from IMBD I am interested in how Steve Carrells departure from the show effected the ratings.

Hide

```
rm(list = ls())

library(rvest)
library(ggplot2)
```

Using the rvest library, I scrape data from IMBD, extracting the episode name, year, and rating.

Hide

```
# Scraping Imbd ratings for "The Office" from the Web
url <- "https://www.imdb.com/search/title/?series=tt0386676&view=simple&count=250&sort=user_rati
ng,desc"
webpage <- read_html(url)

# Using Rvest and CSS selector extension I locate the node in the html and extract the text

# Get the rating
ratings <- html_elements(webpage, ".col-imdb-rating")
as.numeric(html_text(ratings, trim = TRUE)) #convert it to numeric data
```

```
  [1] 9.8 9.8 9.7 9.4 9.4 9.4 9.3 9.3 9.3 9.2 9.2 9.2 9.2 9.0 9.0 9.0 8.9 8.9 8.9 8.9 8.9 8.8 8.
7 8.7 8.7 8.7 8.7 8.7 8.7 8.7 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.5 8.5 8.5 8.5 8.
5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.4 8.4 8.4 8.4 8.4 8.4 8.3 8.3 8.3
 [64] 8.3 8.3 8.3 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.2 8.1 8.1 8.1 8.1 8.1 8.
1 8.1 8.1 8.1 8.1 8.1 8.1 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 7.
9 7.9 7.9 7.9 7.9 7.9 7.9 7.9 7.9 7.9 7.9 7.8 7.8 7.8 7.8 7.8 7.8
[127] 7.8 7.8 7.8 7.8 7.7 7.7 7.7 7.7 7.7 7.7 7.7 7.7 7.7 7.7 7.7 7.7 7.7 7.7 7.6 7.6 7.6 7.6 7.6 7.
6 7.6 7.6 7.6 7.6 7.6 7.5 7.5 7.5 7.5 7.5 7.5 7.5 7.5 7.5 7.4 7.4 7.4 7.4 7.4 7.4 7.3 7.3 7.3 7.
3 7.3 7.3 7.3 7.2 7.1 7.1 7.0 6.9 6.9 6.8 6.8 6.7 6.7 6.6 6.4
```

Hide

```
# Getting the episode name
episode <- html_elements(webpage, ".unbold+ a")
html_text(episode)
```

```
  [1] "Finale"                              "Goodbye, Michael"                    "Stress Relie
f"                        "Niagara: Part 2"                     "A.A.R.M."
"Dinner Party"
  [7] "Garage Sale"                         "Threat Level Midnight"               "Casino Night"
"Niagara: Part 1"                     "Broke"                               "Goodbye, Toby"
 [13] "The Job"                             "Livin' the Dream"                    "Beach Games"
"The Injury"                          "Michael's Last Dundies"              "Classy Christmas"
 [19] "Company Picnic"                      "The Negotiation"                     "Gay Witch Hun
t"                        "Business School"                     "Gossip"
"Cafe Disco"
 [25] "Weight Loss"                         "Fun Run"                             "Women's Appre
ciation"                  "The Return"                          "A Benihana Christmas"
"Christmas Party"
 [31] "Search Committee"                    "Murder"                              "The Surplus"
"Heavy Competition"                   "Michael Scott Paper Company"         "The Duel"
 [37] "The Deposition"                      "Local Ad"                            "Safety Traini
ng"                       "The Merger"                          "Conflict Resolution"
"The Dundies"
 [43] "Dwight K. Schrute, (Acting) Manager" "The Lover"                           "Frame Toby"
"Golden Ticket"                       "Money"                               "Night Out"
 [49] "Product Recall"                      "Traveling Salesmen"                  "Back from Vac
ation"                    "Branch Closing"                      "The Client"
"Booze Cruise"
 [55] "The Search"                          "Happy Hour"                          "The Delivery:
Part 2"                 "Secret Santa"                        "Customer Survey"
"The Coup"
 [61] "PDA"                                 "The Delivery: Part 1"                "Branch Wars"
"Launch Party"                        "Cocktails"                           "Dwight's Speech"
 [67] "Dwight Christmas"                    "Nepotism"                            "Scott's Tots"
"Two Weeks"                           "New Boss"                            "Moroccan Christmas"
 [73] "The Convict"                         "Drug Testing"                        "Take Your Dau
ghter to Work Day"        "Valentine's Day"                     "The Fire"
"Office Olympics"
 [79] "E-Mail Surveillance"                 "Basketball"                          "Ultimatum"
"China"                               "Andy's Play"                         "Casual Friday"
 [85] "Dream Team"                          "Business Ethics"                     "Dunder Miffli
n Infinity"               "Did I Stutter?"                      "Phyllis' Wedding"
"The Secret"
 [91] "Sexual Harassment"                   "Diversity Day"                       "Garden Party"
"The Incentive"                       "Costume Contest"                     "Counseling"
 [97] "The Cover-Up"                        "Shareholder Meeting"                 "Koi Pond"
"The Meeting"                         "Business Trip"                       "Lecture Circuit: Pa
rt 2"
[103] "Lecture Circuit: Part 1"            "Survivor Man"                        "The Conventio
n"                        "Michael's Birthday"                  "The Fight"
"Performance Review"
[109] "Boys and Girls"                     "After Hours"                         "Pool Party"
"The List"                            "Manager and Salesman"                "Double Date"
[115] "Employee Transfer"                  "Crime Aid"                           "Prince Family
Paper"                  "Ben Franklin"                        "Initiation"
"Halloween"
[121] "Paper Airplane"                     "Stairmageddon"                       "Moving On"
```

```
"Whistleblower"                    "The Promotion"                    "Baby Shower"
[127] "Blood Drive"                      "Chair Model"                      "Grief Counsel
ing"                  "The Alliance"                     "Promos"
"Customer Loyalty"
[133] "The Target"                       "Test the Store"                   "Tallahassee"
"Trivia"                           "Christmas Wishes"                 "Viewing Party"
[139] "Sex Ed"                           "Body Language"                    "Secretary's D
ay"                   "Diwali"                           "The Carpet"
"Suit Warehouse"
[145] "The Boat"                         "Work Bus"                         "Last Day in F
lorida"               "Training Day"                     "The Sting"
"The Chump"
[151] "New Leads"                        "Job Fair"                         "Hot Girl"
"Health Care"                      "Free Family Portrait Studio"      "Special Project"
[157] "Mrs. California"                  "Pam's Replacement"                "The Seminar"
"WUPHF.com"                        "St. Patrick's Day"                "Sabre"
[163] "Mafia"                            "Lice"                             "The Whale"
"Turf War"                         "New Guys"                         "Doomsday"
[169] "The Inner Circle"                 "Junior Salesman"                  "The Farm"
"Jury Duty"                        "Spooked"                          "Todd Packer"
[175] "Christening"                      "Pilot"                            "Vandalism"
"Andy's Ancestry"                  "Lotto"                            "Couples Discount"
[181] "Fundraiser"                       "Roy's Wedding"                    "Here Comes Tr
eble"                 "Welcome Party"                    "Angry Andy"
"The Banker"
[187] "Gettysburg"                       "Get the Girl"
```

Hide

```
# Getting the year
year <- html_elements(webpage, ".unbold~ .text-muted")
html_text(year)
```

```
  [1] "(2013)" "(2011)" "(2009)" "(2009)" "(2013)" "(2008)" "(2011)" "(2011)" "(2006)" "(2009)"
"(2009)" "(2008)" "(2007)" "(2013)" "(2007)" "(2006)" "(2011)" "(2010)" "(2009)" "(2007)" "(200
6)" "(2007)" "(2009)" "(2009)" "(2008)" "(2007)" "(2007)" "(2007)"
 [29] "(2006)" "(2005)" "(2011)" "(2009)" "(2008)" "(2009)" "(2009)" "(2009)" "(2007)" "(2007)"
"(2007)" "(2006)" "(2006)" "(2005)" "(2011)" "(2009)" "(2008)" "(2009)" "(2007)" "(2008)" "(200
7)" "(2007)" "(2007)" "(2006)" "(2005)" "(2006)" "(2011)" "(2010)"
 [57] "(2010)" "(2009)" "(2008)" "(2006)" "(2011)" "(2010)" "(2007)" "(2007)" "(2007)" "(2006)"
"(2012)" "(2010)" "(2009)" "(2009)" "(2009)" "(2008)" "(2006)" "(2006)" "(2006)" "(2006)" "(200
5)" "(2005)" "(2005)" "(2005)" "(2011)" "(2010)" "(2010)" "(2009)"
 [85] "(2009)" "(2008)" "(2007)" "(2008)" "(2007)" "(2006)" "(2005)" "(2005)" "(2011)" "(2011)"
"(2010)" "(2010)" "(2010)" "(2009)" "(2009)" "(2009)" "(2008)" "(2009)" "(2009)" "(2007)" "(200
6)" "(2006)" "(2005)" "(2005)" "(2006)" "(2012)" "(2012)" "(2011)"
[113] "(2010)" "(2009)" "(2008)" "(2008)" "(2009)" "(2007)" "(2006)" "(2005)" "(2013)" "(2013)"
"(2013)" "(2010)" "(2009)" "(2008)" "(2009)" "(2008)" "(2006)" "(2005)" "(2013)" "(2013)" "(201
2)" "(2012)" "(2012)" "(2012)" "(2011)" "(2010)" "(2010)" "(2010)"
[141] "(2010)" "(2006)" "(2006)" "(2013)" "(2012)" "(2012)" "(2012)" "(2011)" "(2010)" "(2010)"
"(2010)" "(2008)" "(2005)" "(2005)" "(2012)" "(2012)" "(2011)" "(2011)" "(2011)" "(2010)" "(201
0)" "(2010)" "(2009)" "(2013)" "(2012)" "(2012)" "(2012)" "(2011)"
[169] "(2011)" "(2013)" "(2013)" "(2012)" "(2011)" "(2011)" "(2010)" "(2005)" "(2013)" "(2012)"
"(2011)" "(2013)" "(2012)" "(2012)" "(2012)" "(2012)" "(2012)" "(2010)" "(2011)" "(2012)"
```

I then put the scraped information in a dataframe suitable for visualization and analysis.

Hide

```
# Put the scraped elements in a new data frame
office_data <- data.frame(html_text(episode),
                          as.numeric(html_text(ratings, trim = TRUE)),
                          html_text(year))

# give sensible names to the dat frame
names(office_data) <- c("episode", "rating", "year")
```

IMBD site does not have info on whether Michael Scott was in a particular episode. Using my own knowledge of the show, I created a variable that tracks whether Michael was in the show. Michael was in every episode for the first 6 seasons. These seasons correspond to the years 2005 - 2010. Michael was in most of the episodes for season 7, and returned for the finale in season 9. Because these episodes do not cleanly correspond to years I created a short list of Michael Episodes that were not captured by the years.

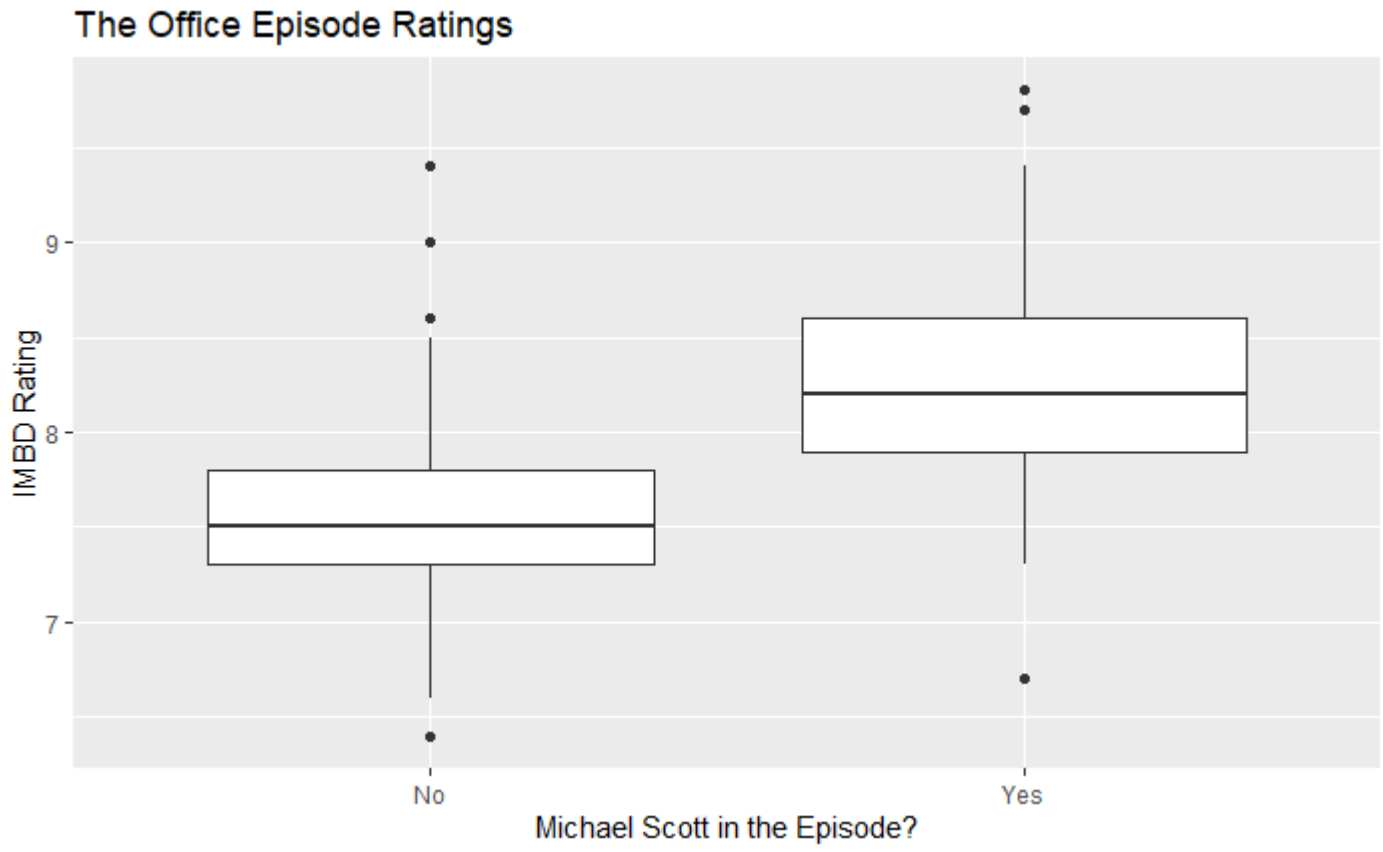
Hide

```
# Create a variable tracking if the character "Michael Scott" was in the episode

office_data$michael_scott <- "No"

# Create a character vector for the years and episodes the Micheal was in
michael_scott_years <- c("(2005)", "(2006)", "(2007)", "(2008)", "(2009)", "(2010)")
michael_scott_episodes <- c("Finale", "Goodbye, Michael", "Garage Sale", "Threat Level Midnigh
t",
                            "Michael's Last Dundies", "The Search", "PDA", "Ultimatum",
                            "Training Day", "The Seminar", "Todd Packer")


# I index the data checking to see if the row was in one of these character vectors
# and change the Michael Scott status to "Yes" accordingly
office_data$michael_scott[office_data$year %in% michael_scott_years] <- "Yes"
office_data$michael_scott[office_data$episode %in% michael_scott_episodes] <- "Yes"
```

First I am interested in the overall rating based on Michael Scott's character. Statistical analysis reveals that episodes with Michael in them had an average rating of 8.26 and that episodes without had an average of 7.55. This is nearly a 1 point difference on a scale of 10, and is statistically significant $p < 0.001$.

Hide

```
ggplot(data = office_data, mapping = aes(x = michael_scott, y = rating)) +
  geom_boxplot() +
  xlab("Michael Scott in the Episode?") +
  ylab("IMBD Rating") +
  ggtitle("The Office Episode Ratings")
```



Hide

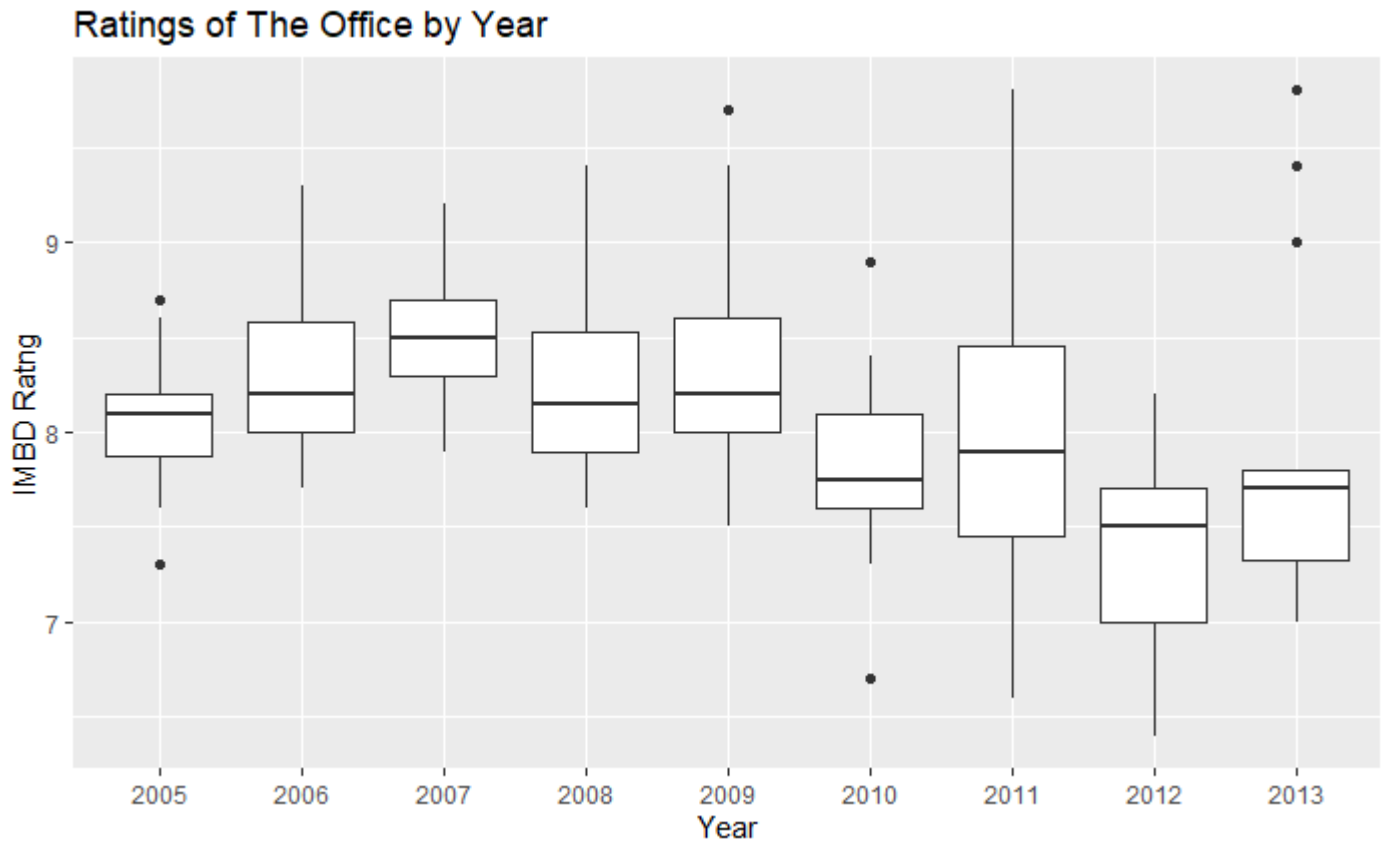
```
t.test(office_data$rating ~ office_data$michael_scott)
```

```
    Welch Two Sample t-test

data:  office_data$rating by office_data$michael_scott
t = -7.6924, df = 81.083, p-value = 3.007e-11
alternative hypothesis: true difference in means between group No and group Yes is not equal to
0
95 percent confidence interval:
 -0.8959149 -0.5276939
sample estimates:
 mean in group No mean in group Yes
         7.555102          8.266906
```

I was also interested in how the show fared overtime. Many fans - myself included - consider seasons 2-4 to be the golden era of The office. Season 8, the first season without Michael, is often considered to be a lowpoint in the series.

Hide

```
ggplot(data = office_data, mapping = aes(x = year, y = rating)) +
  geom_boxplot() +
  scale_x_discrete(labels = c("2005", "2006", "2007", "2008", "2009", "2010", "2011", "2012", "2
013"), name = "Year") +
  ylab("IMBD Ratng") +
  ggtitle("Ratings of The Office by Year")
```



Although the years do not perfectly correspond to the seasons (Most season span over two calendar years) this graph reflects the traditional wisdom. Years 2006 - 2009, which encompass seasons 2-5 experience the highest average ratings. Meanwhile year 2012, which had most of season 8 episodes, is a low point on the graph.

While The Office endend on a strong note - the finale was one of the highest rated episodes - the show did suffer from the departure of it’s main character. Using web-scraping methods we can see that the episodes and years without Michael Scott had significantly lower ratings on average.