Office Project

Josh Kong

6/15/2020

```
library(tidytext)
library(schrute)
library(tidyverse)
```

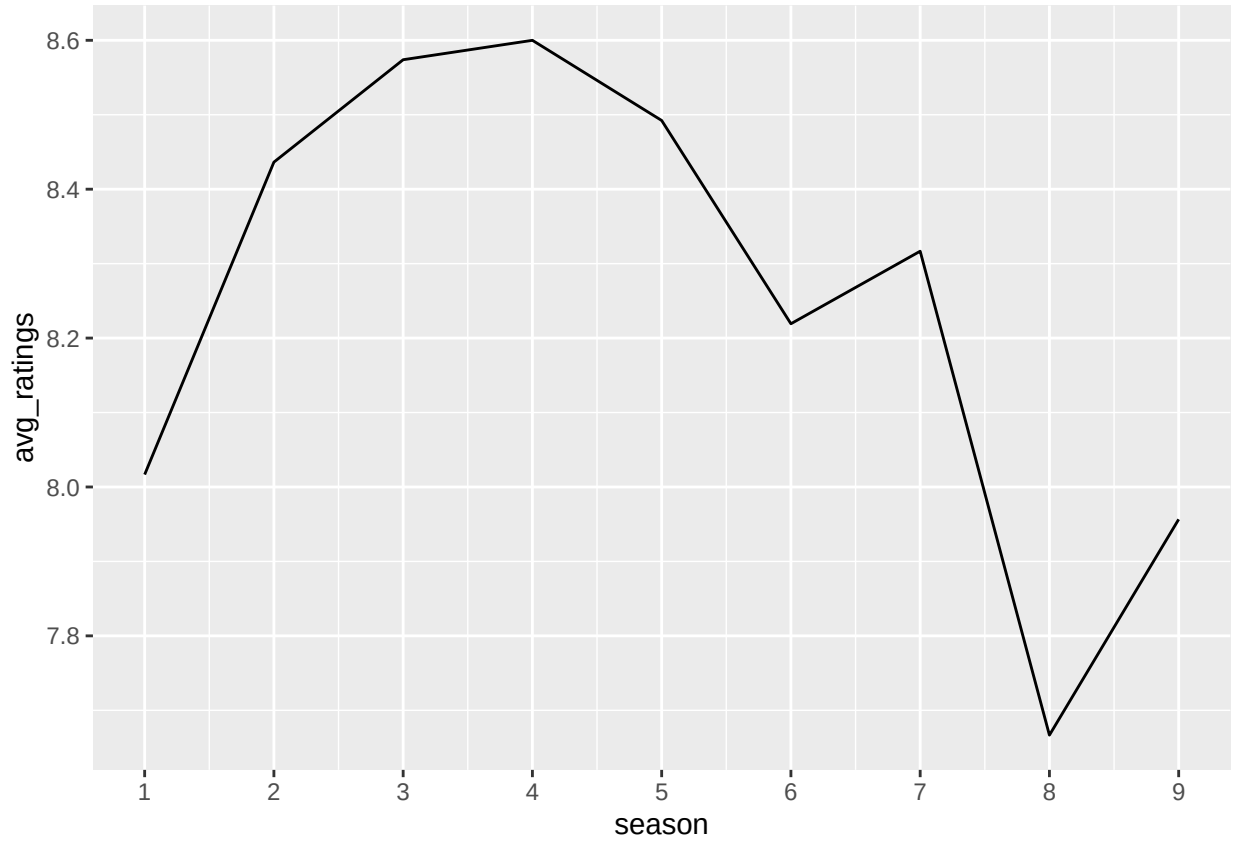
```
## -- Attaching packages ---------------------------------------------------------------------------- t

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------------------------- tidyver
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
office_transcripts <- as_tibble(theoffice) %>%
  mutate(season = as.integer(season),
         episode = as.integer(episode)) %>%
  mutate(character = str_remove_all(character, '"')) %>%
  mutate(name = str_to_lower(str_remove_all(episode_name, "\\.| \\(Part.*")))

office_ratings <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/
  mutate(name = str_to_lower(str_remove_all(title, "\\.| \\(Part.*|\\: Part.*")))
```

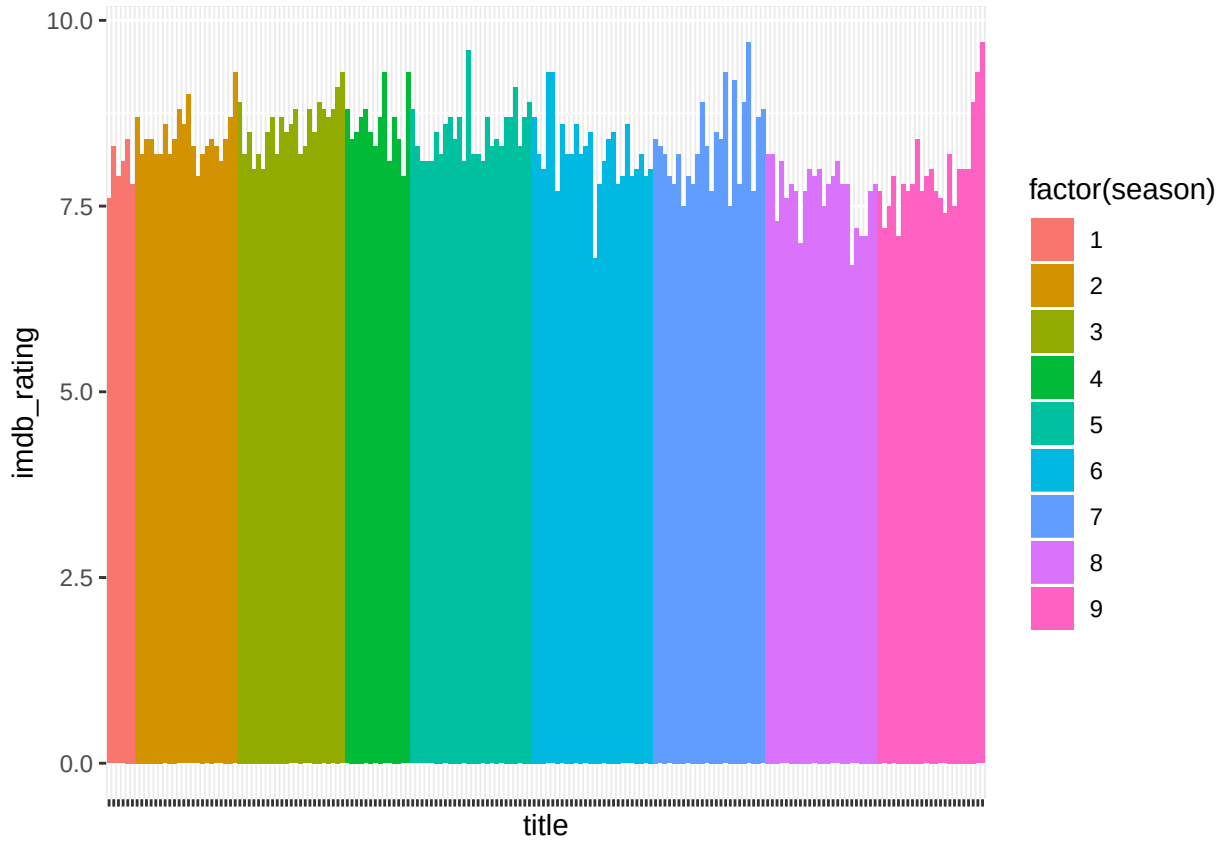
```
## Parsed with column specification:
## cols(
##   season = col_double(),
##   episode = col_double(),
##   title = col_character(),
##   imdb_rating = col_double(),
##   total_votes = col_double(),
##   air_date = col_date(format = "")
## )
```

Ratings of the Office by Episode and Season

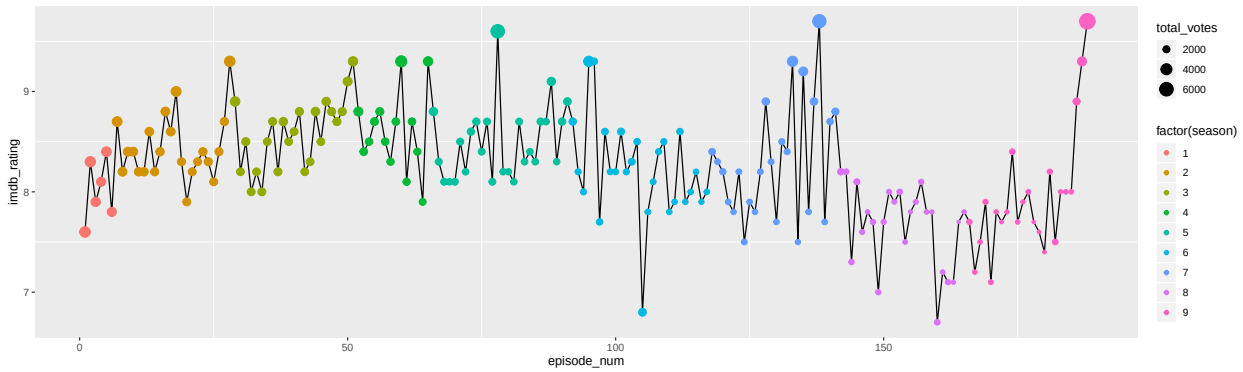
```r
# Looking at the office ratings per season
office_ratings %>%
  group_by(season) %>%
  summarise(avg_ratings = mean(imdb_rating)) %>%
  ggplot(aes(season,avg_ratings)) + geom_line() +
  scale_x_continuous(breaks=1:9)
```




```r
# Looking at the office ratings per episode using a bar graph

office_ratings %>%
  mutate(title = fct_inorder(title)) %>%
  ggplot(aes(title,imdb_rating, fill = factor(season)))+
  geom_col()+
  theme(axis.text.x = element_blank())
```



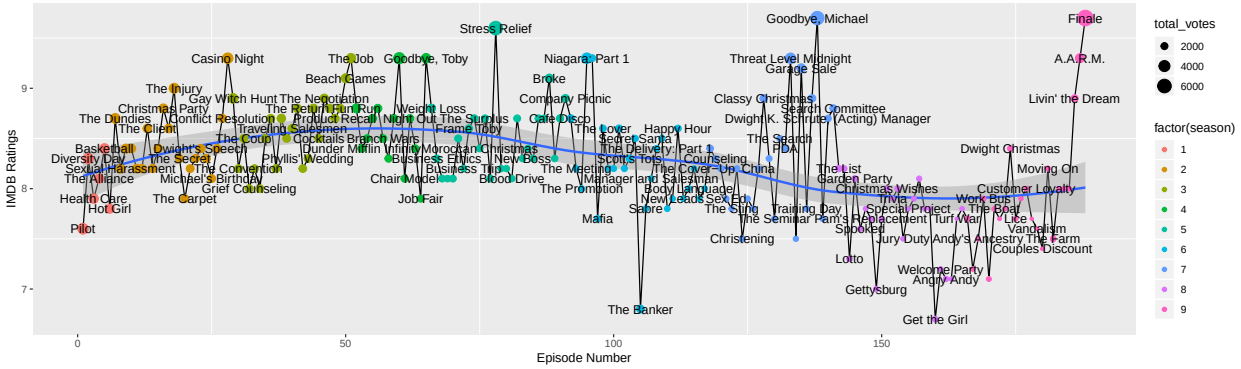

```
# Looking at the office ratings using a line graph
office_ratings %>%
  mutate(title = fct_inorder(title), episode_num = row_number()) %>%
  ggplot(aes(episode_num,imdb_rating)) +
  geom_line(group = 1) +
  geom_point(aes(color = factor(season),size = total_votes)) +
  theme(panel.grid.major.x = element_blank(),
        panel.grid.major.y = element_blank())
```



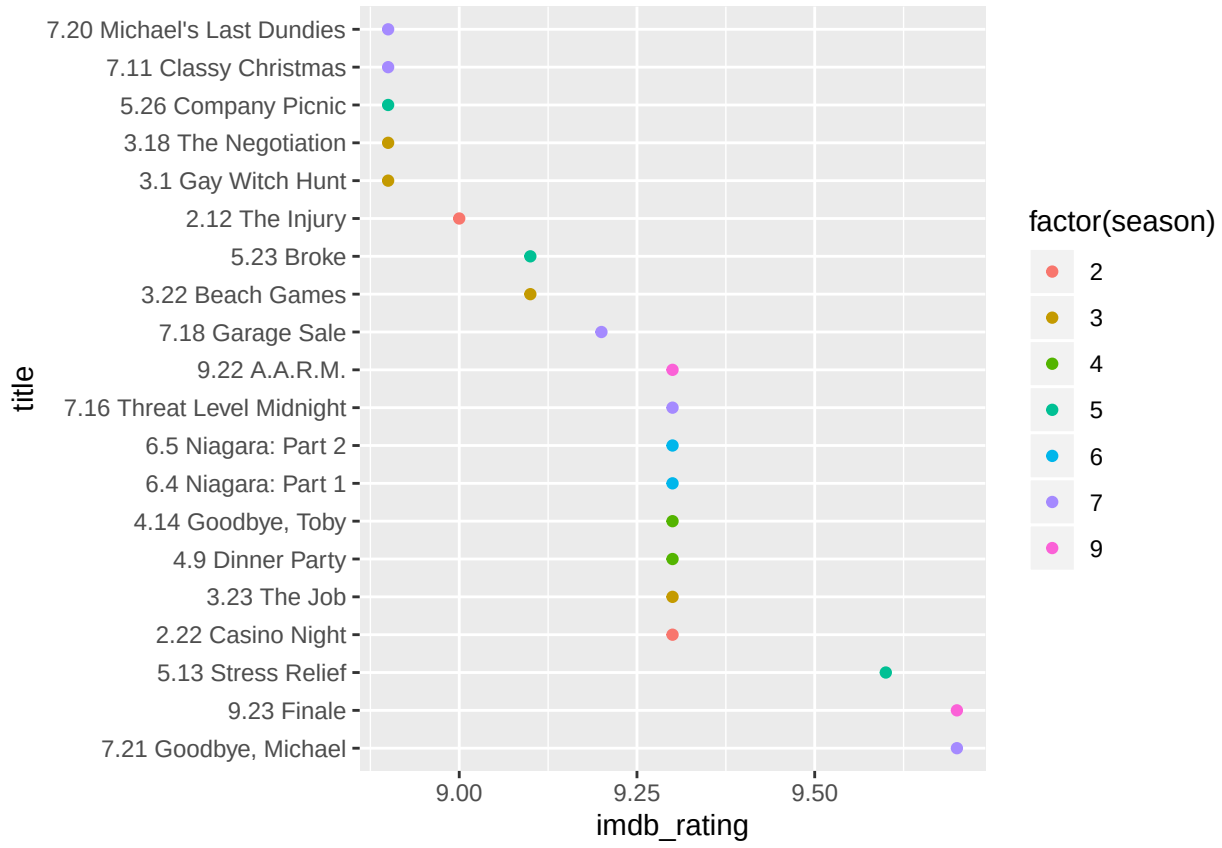

```
#this plot shows the trend that the office ratings follow along with the episode names
office_ratings %>%
  mutate(title = fct_inorder(title), episode_num = row_number()) %>%
  ggplot(aes(episode_num,imdb_rating)) +
```

```
  geom_line(group = 1) +
  geom_smooth() +
  geom_point(aes(color = factor(season),size = total_votes)) +
  geom_text(aes(label = title),check_overlap = TRUE)+
  theme(panel.grid.major.x = element_blank(),
        panel.grid.major.y = element_blank())+
  labs(x = "Episode Number", y = "IMDB Ratings")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
#looking at the top 20 rated office episodes
office_ratings %>%
  arrange(desc(imdb_rating)) %>%
  mutate(title = paste0(season,".",episode," ",title), title = fct_inorder(title)) %>%
  head(20) %>%
  ggplot(aes(title,imdb_rating)) + geom_point(aes(color = factor(season))) + coord_flip()
```



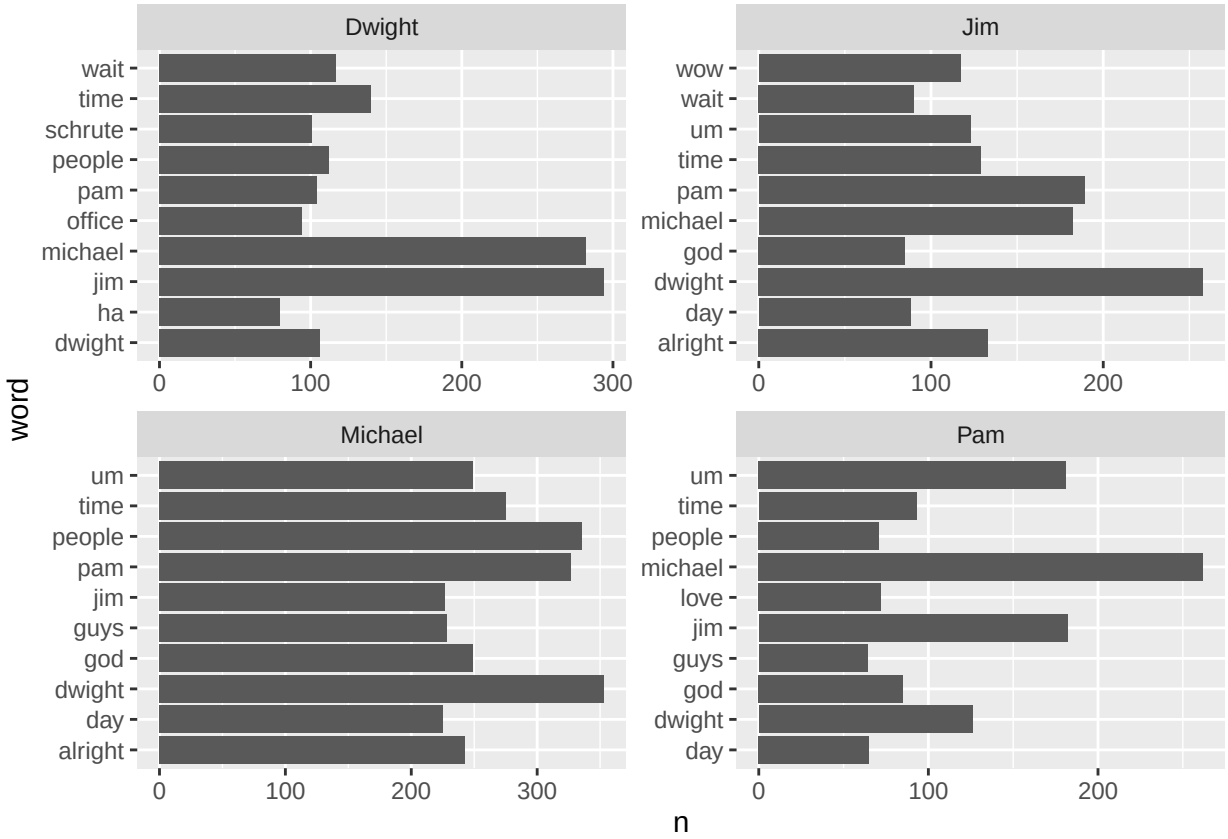

```
#NOTE, no episodes from season 1 or 8 are in the top 20 episodes
```

Looking at the office transcripts

```
#use %in% if you want to do == for a vector!!!
blacklist <- c("yeah", "hey", "uh", "gonna")
blacklist_characters <- c("Everyone", "All", "Both", "Guy", "Girl", "Group")

transcript_words <- office_transcripts %>%
  group_by(character) %>%
  filter(n()>=100,
         n_distinct(episode_name) > 2) %>%
  ungroup() %>%
  select(-text_w_direction) %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words, by = "word") %>%
  filter(!word %in% blacklist,
         !character %in% blacklist_characters)


character_words <- transcript_words %>%
  count(character, word, sort = TRUE)

#looking at some of the most common words said by major characters
```

```
character_words %>%
  filter(character %in% c("Michael", "Jim", "Pam", "Dwight")) %>%
  group_by(character) %>%
  top_n(10) %>%
  ungroup() %>%
  ggplot(aes(word,n)) +geom_col()+ coord_flip() + facet_wrap(~character,scales = "free")
```

```
## Selecting by n
```



What are some of the things that affect the ratings of an episode?

Some of the things that we can observe are

- Season
- Director
- Writers
- Lines per character in the episode

```
#begin by joining the datasets of office transcripts and office ratings

#trying to see if there is a relationship between popular characters and ratings
#character must have at least 50 lines and must appear in at least 5 episodes to qualify
ratings_summarized <- office_ratings %>%
```

```
  group_by(name) %>%
  summarize(imdb_rating = mean(imdb_rating))

character_lines_ratings <- office_transcripts %>%
  filter(!character %in% blacklist_characters) %>%
  count(character, name) %>%
  group_by(character) %>%
  filter(sum(n) >= 50,
         n() >= 5) %>%
  inner_join(ratings_summarized, by = "name")

character_lines_ratings %>%
  summarize(avg_rating = mean(imdb_rating),
            nb_episodes = n()) %>%
  arrange(desc(avg_rating))
```

```
## # A tibble: 39 x 3
##    character     avg_rating nb_episodes
##    <chr>              <dbl>       <int>
##  1 Carol               8.7            6
##  2 Charles             8.62           6
##  3 Karen               8.6           25
##  4 Holly               8.57          16
##  5 Jan                 8.48          37
##  6 Michael             8.42         130
##  7 David Wallace       8.42          16
##  8 David               8.38          27
##  9 Roy                 8.36          28
## 10 Josh                8.34           7
## # ... with 29 more rows
```

Now looking at directors and writers

```
director_writer_features <- office_transcripts %>%
  distinct(name, director, writer) %>%
  gather(type, value, director, writer) %>%
  separate_rows(value, sep = ";") %>%
  unite(feature, type, value, sep = ": ") %>%
  group_by(feature) %>%
  filter(n() >= 3) %>%
  mutate(value = 1) %>%
  ungroup()

character_line_features <- character_lines_ratings %>%
  ungroup() %>%
  transmute(name, feature = character, value = log2(n))

season_features = office_ratings %>%
  distinct(name, season) %>%
  transmute(name, feature = paste("season:", season), value = 1)
features <- bind_rows(director_writer_features,
                      character_line_features,
                      season_features) %>%
```

```
  semi_join(office_ratings, by = "name") %>%
  semi_join(office_transcripts, by = "name")

episode_feature_matrix <- features %>%
  cast_sparse(name, feature, value)
ratings <- ratings_summarized$imdb_rating[match(rownames(episode_feature_matrix), ratings_summarized$nam
ratings_summarized$name
```

```
##   [1] "a benihana christmas"              "aarm"
##   [3] "after hours"                       "andy's ancestry"
##   [5] "andy's play"                       "angry andy"
##   [7] "baby shower"                       "back from vacation"
##   [9] "basketball"                        "beach games"
##  [11] "ben franklin"                      "blood drive"
##  [13] "body language"                     "booze cruise"
##  [15] "boys and girls"                    "branch closing"
##  [17] "branch wars"                       "broke"
##  [19] "business ethics"                   "business school"
##  [21] "business trip"                     "cafe disco"
##  [23] "casino night"                      "casual friday"
##  [25] "chair model"                       "china"
##  [27] "christening"                       "christmas party"
##  [29] "christmas wishes"                  "classy christmas"
##  [31] "cocktails"                         "company picnic"
##  [33] "conflict resolution"               "costume contest"
##  [35] "counseling"                        "couples discount"
##  [37] "crime aid"                         "customer loyalty"
##  [39] "customer survey"                   "did i stutter?"
##  [41] "dinner party"                      "diversity day"
##  [43] "diwali"                            "doomsday"
##  [45] "double date"                       "dream team"
##  [47] "drug testing"                      "dunder mifflin infinity"
##  [49] "dwight christmas"                  "dwight k schrute, (acting) manager"
##  [51] "dwight's speech"                   "e-mail surveillance"
##  [53] "employee transfer"                 "finale"
##  [55] "frame toby"                        "free family portrait studio"
##  [57] "fun run"                           "fundraiser"
##  [59] "garage sale"                       "garden party"
##  [61] "gay witch hunt"                    "get the girl"
##  [63] "gettysburg"                        "golden ticket"
##  [65] "goodbye, michael"                  "goodbye, toby"
##  [67] "gossip"                            "grief counseling"
##  [69] "halloween"                         "happy hour"
##  [71] "health care"                       "heavy competition"
##  [73] "here comes treble"                 "hot girl"
##  [75] "initiation"                        "job fair"
##  [77] "junior salesman"                   "jury duty"
##  [79] "koi pond"                          "last day in florida"
##  [81] "launch party"                      "lecture circuit"
##  [83] "lice"                              "livin' the dream"
##  [85] "local ad"                          "lotto"
##  [87] "mafia"                             "manager and salesman"
##  [89] "michael scott paper company"       "michael's birthday"
```

```
##  [91] "michael's last dundies"            "money"
##  [93] "moroccan christmas"                "moving on"
##  [95] "mrs california"                    "murder"
##  [97] "nepotism"                          "new boss"
##  [99] "new guys"                          "new leads"
## [101] "niagara"                           "night out"
## [103] "office olympics"                   "pam's replacement"
## [105] "paper airplane"                    "pda"
## [107] "performance review"                "phyllis' wedding"
## [109] "pilot"                             "pool party"
## [111] "prince family paper"               "product recall"
## [113] "promos"                            "roy's wedding"
## [115] "sabre"                             "safety training"
## [117] "scott's tots"                      "search committee"
## [119] "secret santa"                      "secretary's day"
## [121] "sex ed"                            "sexual harassment"
## [123] "shareholder meeting"               "special project"
## [125] "spooked"                           "st patrick's day"
## [127] "stairmageddon"                     "stress relief"
## [129] "suit warehouse"                    "survivor man"
## [131] "take your daughter to work day"    "tallahassee"
## [133] "test the store"                    "the alliance"
## [135] "the banker"                        "the boat"
## [137] "the carpet"                        "the chump"
## [139] "the client"                        "the convention"
## [141] "the convict"                       "the coup"
## [143] "the cover-up"                      "the delivery"
## [145] "the deposition"                    "the duel"
## [147] "the dundies"                       "the farm"
## [149] "the fight"                         "the fire"
## [151] "the incentive"                     "the injury"
## [153] "the inner circle"                  "the job"
## [155] "the list"                          "the lover"
## [157] "the meeting"                       "the merger"
## [159] "the negotiation"                   "the promotion"
## [161] "the return"                        "the search"
## [163] "the secret"                        "the seminar"
## [165] "the sting"                         "the surplus"
## [167] "the target"                        "the whale"
## [169] "threat level midnight"             "todd packer"
## [171] "training day"                      "traveling salesmen"
## [173] "trivia"                            "turf war"
## [175] "two weeks"                         "ultimatum"
## [177] "valentine's day"                   "vandalism"
## [179] "viewing party"                     "weight loss"
## [181] "welcome party"                     "whistleblower"
## [183] "women's appreciation"              "work bus"
## [185] "wuphfcom"
```

```
ratings
```

```
##   [1] 7.60 8.30 7.90 8.40 8.70 8.20 8.40 8.40 8.20 8.20 8.20 8.80 8.60 8.30 8.40
##  [16] 8.70 9.30 8.90 8.20 8.50 8.20 8.50 8.70 8.20 8.70 8.60 8.80 8.20 8.30 8.90
##  [31] 8.80 8.70 8.80 9.10 9.30 8.80 8.50 8.30 9.30 8.10 8.70 8.40 7.90 9.30 8.80
```

```
##  [46] 8.30 8.10 8.10 8.10 8.20 8.70 8.40 9.60 8.20 8.10 8.70 8.30 8.40 8.30 8.70
##  [61] 9.10 8.30 8.70 8.90 8.70 8.20 8.00 9.30 7.70 8.60 8.20 8.30 8.50 6.80 7.80
##  [76] 8.45 7.80 7.90 8.60 7.90 7.90 8.00 8.40 8.30 7.80 7.90 8.20 8.90 8.30 7.70
##  [91] 8.40 9.30 7.50 9.20 7.80 9.70 7.70 8.70 8.80 8.20 8.20 7.30 8.10 7.60 7.80
## [106] 7.70 7.00 7.90 8.00 7.80 7.90 7.80 7.80 6.70 7.10 7.80 7.20 7.50 7.80 7.80
## [121] 8.40 7.90 7.70 7.40 7.50 8.00 8.00 8.90 9.30 9.70 8.10 7.80 8.60 9.00 8.30
## [136] 7.90 8.40 8.30 8.10 8.00 8.00 8.50 8.80 8.50 8.40 8.70 8.80 8.50 8.60 8.70
## [151] 8.10 8.60 8.20 8.20 8.00 8.20 8.20 7.80 8.50 8.90 8.00 7.50 8.10 7.20 7.10
## [166] 7.70 7.90 7.10 7.70 7.60 8.00 7.50 8.50 8.20 7.70 8.00 7.70 8.70
```

```
ratings_summarized
```

```
## # A tibble: 185 x 2
##    name                  imdb_rating
##    <chr>                       <dbl>
##  1 a benihana christmas          8.7
##  2 aarm                          9.3
##  3 after hours                   8.1
##  4 andy's ancestry               7.5
##  5 andy's play                   8.2
##  6 angry andy                    7.1
##  7 baby shower                   8.1
##  8 back from vacation            8.5
##  9 basketball                    8.4
## 10 beach games                   9.1
## # ... with 175 more rows
```

Machine Learning: Ridge Regression

Want to see the effect of a character, season, writer, and director on the rating of an office episode

```
library(glmnet)
```

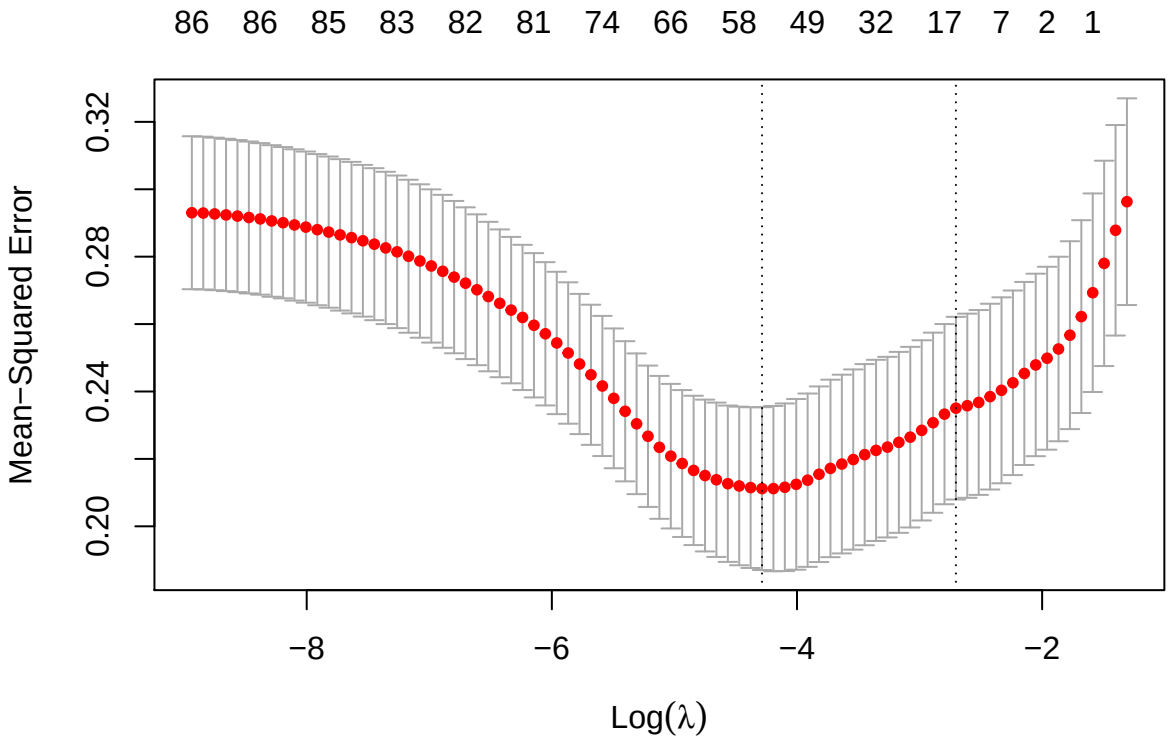
```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

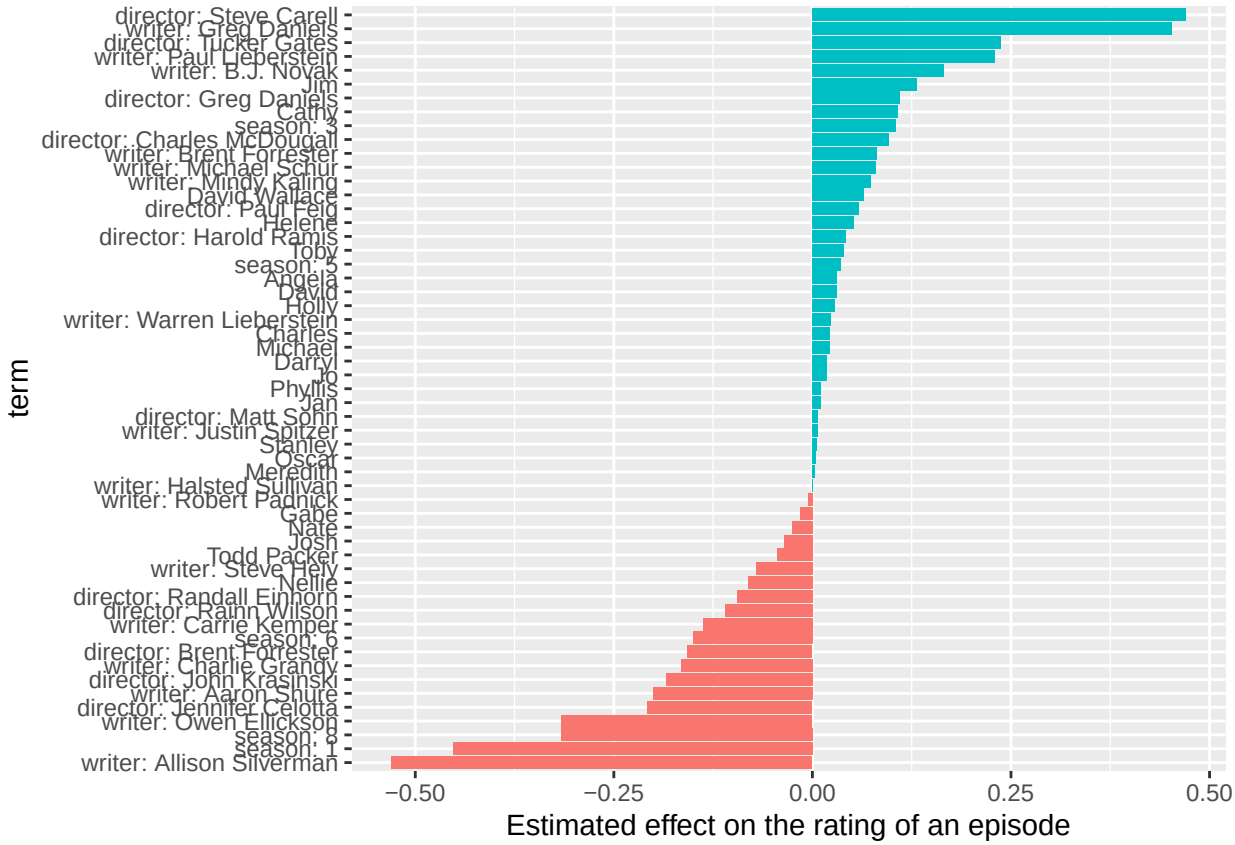
```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 3.0-2
```

```
library(broom)
mod <- cv.glmnet(episode_feature_matrix, ratings)
plot(mod)
```




```
tidy(mod$glmnet.fit) %>%
  filter(lambda == mod$lambda.min,
         term != "(Intercept)") %>%
  mutate(term = fct_reorder(term, estimate)) %>%
  ggplot(aes(term, estimate, fill = estimate > 0)) +
  geom_col() +
  coord_flip() +
  labs(y = "Estimated effect on the rating of an episode") +
  theme(legend.position = "none")
```



Looking at the graph, it appears that when Steve Carell is a director, or Greg Daniels is the writer, the episode tends to do very well.

It appears that when Allison Silverman is a writer, or the episode is in season 1, the episode tends to do poorly.