Jeopardy Project All Seasons

Heather Deel

11/18/2019

## Libraries

library(readr)  
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(tidyr)  
library(stringr)  
library(ggplot2)  
library(forcats)  
library(RColorBrewer)

## Read in the data

seasons\_all <- read\_tsv("master\_season1-35.tsv")

## Parsed with column specification:  
## cols(  
## round = col\_double(),  
## value = col\_double(),  
## daily\_double = col\_character(),  
## category = col\_character(),  
## comments = col\_character(),  
## answer = col\_character(),  
## question = col\_character(),  
## air\_date = col\_date(format = ""),  
## notes = col\_character()  
## )

## Create new dataframe with all seasons

seasons\_all\_cat <- seasons\_all %>%  
 group\_by(category) %>%   
 count() %>%   
 arrange(desc(n))

## Clean up seasons data

seasons\_all\_clean <- seasons\_all %>%   
 separate(air\_date, sep="-", into = c("year", "month", "day")) %>%   
 select(-comments, -notes) %>%   
 mutate(category = str\_to\_lower(string = category)) %>%  
 mutate(answer = str\_to\_lower(string = answer)) %>%   
 mutate(question = str\_to\_lower(string = question)) %>%   
 filter(year != "2019")

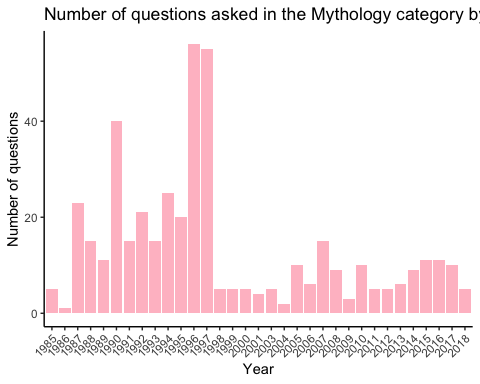
Question 2: How often does the mythology category appear over each year? Is there a trend in how often it occurs over time?

## Filter by mythology category and group by year

seasons\_all\_myth <- seasons\_all\_clean %>%   
 filter(category == "mythology") %>%   
 group\_by(year) %>%   
 count()

## Plot mythology category by year

seasons\_all\_myth %>%   
 ggplot(aes(x = year, y = n)) +  
 geom\_bar(stat = "identity", fill = "pink") +  
 theme\_classic() +  
 ggtitle("Number of questions asked in the Mythology category by season") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
 labs(x = "Year", y = "Number of questions")



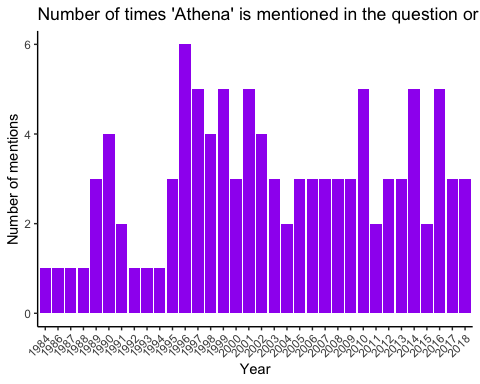
Question 3: How often does “Athena”, the greek goddess of war, appear over time? Is there a trend?

## Filter to “Athena” in question and answer columns and group by year

# note that with this code, it includes "Athena" not just in the mythology category - increases occurrences from 21 to 103  
  
seasons\_all\_athena <- seasons\_all\_clean %>%   
 filter(str\_detect(question, "athena") | str\_detect(answer, "athena")) %>%   
 group\_by(year) %>%   
 count()  
  
# create dataframe without grouping to see what other categories are there  
seasons\_all\_athena\_cat <- seasons\_all\_clean %>%   
 filter(str\_detect(question, "athena") | str\_detect(answer, "athena")) %>%   
 group\_by(category) %>%   
 count()

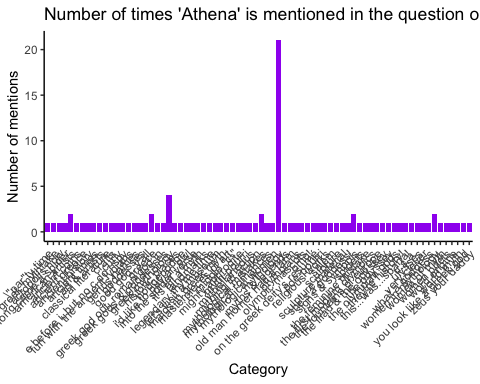
## Plot “Athena” by season

seasons\_all\_athena %>%   
 ggplot(aes(x = year, y = n)) +  
 geom\_bar(stat = "identity", fill = "purple") +  
 theme\_classic() +  
 ggtitle("Number of times 'Athena' is mentioned in the question or answer") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
 labs(x = "Year", y = "Number of mentions")



## Try plotting but by category

seasons\_all\_athena\_cat %>%   
 ggplot(aes(x = category, y = n)) +  
 geom\_bar(stat = "identity", fill = "purple") +  
 theme\_classic() +  
 ggtitle("Number of times 'Athena' is mentioned in the question or answer") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
 labs(x = "Category", y = "Number of mentions")



### Do it with multiple greek gods

## Filter to multiple greek gods in question and answer columns and group by year

greek\_gods <- c("athena", "zeus", "poseidon", "hades", "hera", "apollo", "demeter", "artemis", "aphrodite", "ares")  
  
seasons\_all\_greek <- seasons\_all\_clean %>%   
 filter(str\_detect(question, greek\_gods) | str\_detect(answer, greek\_gods))

## Warning in stri\_detect\_regex(string, pattern, negate = negate, opts\_regex  
## = opts(pattern)): longer object length is not a multiple of shorter object  
## length  
  
## Warning in stri\_detect\_regex(string, pattern, negate = negate, opts\_regex  
## = opts(pattern)): longer object length is not a multiple of shorter object  
## length

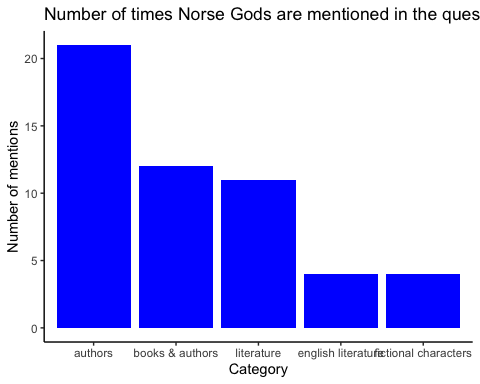
## Norse

norse\_gods <- c("odin", "loki", "thor", "njord", "heimdall", "frigg", "tyr", "baldur", "idun", "bragi")  
  
seasons\_all\_norse <- seasons\_all\_clean %>%   
 filter(str\_detect(question, norse\_gods) | str\_detect(answer, norse\_gods))

## Warning in stri\_detect\_regex(string, pattern, negate = negate, opts\_regex  
## = opts(pattern)): longer object length is not a multiple of shorter object  
## length  
  
## Warning in stri\_detect\_regex(string, pattern, negate = negate, opts\_regex  
## = opts(pattern)): longer object length is not a multiple of shorter object  
## length

## Graph norse gods by category

seasons\_all\_norse\_top5 <- seasons\_all\_norse %>%   
 filter(category == "authors" |  
 category == "books & authors" |  
 category == "literature" |   
 category == "english literature" |   
 category == "fictional characters") %>%   
 group\_by(category) %>%   
 count()  
   
  
seasons\_all\_norse\_top5 %>%  
 ggplot(aes(x = reorder(category, -n), y = n)) +  
 geom\_bar(stat = "identity", fill = "blue") +  
 theme\_classic() +  
 ggtitle("Number of times Norse Gods are mentioned in the question or answer") +  
 labs(x = "Category", y = "Number of mentions")



## Hindu

hindu\_gods <- c("vishnu", "brahma", "shiva", "lakshmi", "hanuman", "ganesha", "krishna", "kali", "rama", "saraswati")  
  
seasons\_all\_hindu <- seasons\_all\_clean %>%   
 filter(str\_detect(question, hindu\_gods) | str\_detect(answer, hindu\_gods))

## Warning in stri\_detect\_regex(string, pattern, negate = negate, opts\_regex  
## = opts(pattern)): longer object length is not a multiple of shorter object  
## length  
  
## Warning in stri\_detect\_regex(string, pattern, negate = negate, opts\_regex  
## = opts(pattern)): longer object length is not a multiple of shorter object  
## length

### Try adding column for god name

seasons\_all\_hindu\_name <- seasons\_all\_clean %>%   
 filter(str\_detect(question, hindu\_gods) | str\_detect(answer, hindu\_gods)) %>%   
 mutate(god\_name = stringr::str\_extract(question, "vishnu"))

## Warning in stri\_detect\_regex(string, pattern, negate = negate, opts\_regex  
## = opts(pattern)): longer object length is not a multiple of shorter object  
## length  
  
## Warning in stri\_detect\_regex(string, pattern, negate = negate, opts\_regex  
## = opts(pattern)): longer object length is not a multiple of shorter object  
## length

## Merga all 3 datasets

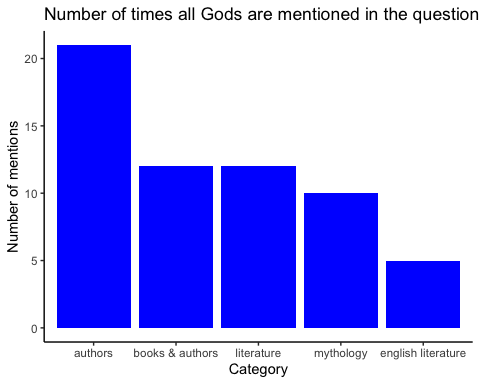
seasons\_all\_gods3 <- merge(merge(seasons\_all\_greek, seasons\_all\_norse, all = TRUE), seasons\_all\_hindu, all = TRUE)

## Group by category

seasons\_all\_gods\_cat <- seasons\_all\_gods3 %>%   
 group\_by(category) %>%   
 count()  
  
seasons\_all\_gods\_top5 <- seasons\_all\_gods\_cat %>%  
 filter(category == "authors" |  
 category == "books & authors" |  
 category == "literature" |   
 category == "mythology" |   
 category == "english literature")

## Now graph

seasons\_all\_gods\_top5 %>%  
 ggplot(aes(x = reorder(category, -n), y = n)) +  
 geom\_bar(stat = "identity", fill = "blue") +  
 theme\_classic() +  
 ggtitle("Number of times all Gods are mentioned in the question or answer") +  
 labs(x = "Category", y = "Number of mentions")



### will maybe try to facet by god or something for another interesting figure

### In depth analysis of gods and their prevalence in certain categories

### create columns for Zeus and Athena

seasons\_all\_gods3\_athenaQ <- seasons\_all\_gods3 %>%   
 mutate(god\_name = stringr::str\_extract(question, "athena"))  
  
seasons\_all\_gods3\_athenaA <- seasons\_all\_gods3 %>%   
 mutate(god\_name = stringr::str\_extract(answer, "athena"))  
  
seasons\_all\_gods3\_zeusQ <- seasons\_all\_gods3 %>%   
 mutate(god\_name = stringr::str\_extract(question, "zeus"))  
  
seasons\_all\_gods3\_zeusA <- seasons\_all\_gods3 %>%   
 mutate(god\_name = stringr::str\_extract(answer, "zeus"))  
  
seasons\_all\_gods\_ZAcat <- merge(merge(merge(seasons\_all\_gods3\_athenaQ, seasons\_all\_gods3\_athenaA, all = TRUE), seasons\_all\_gods3\_zeusQ, all = TRUE), seasons\_all\_gods3\_zeusA, all = TRUE)

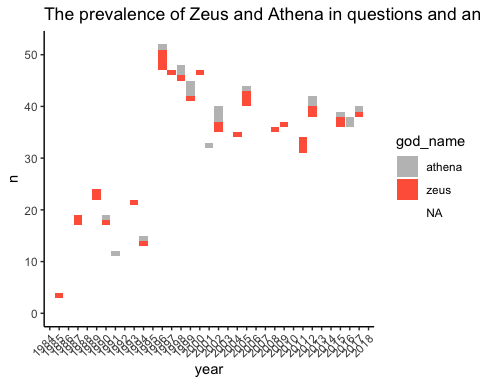
### Fill by god (Zeus or Athena) to look at their prevalences over time

seasons\_all\_gods\_ZAcat\_group <- seasons\_all\_gods\_ZAcat %>%  
 group\_by(category, year, god\_name) %>%   
 count()

## Releveling

# make it so it says 'other'  
seasons\_all\_gods\_ZAcat$god\_name <- as.character(seasons\_all\_gods\_ZAcat$god\_name)  
seasons\_all\_gods\_ZAcat$god\_name[is.na(seasons\_all\_gods\_ZAcat$god\_name)] <- "other"  
  
seasons\_all\_gods\_ZAcat$god\_name <- as.factor(seasons\_all\_gods\_ZAcat$god\_name)  
  
seasons\_all\_gods\_ZAcat$god\_name <- relevel(seasons\_all\_gods\_ZAcat$god\_name, "other")

color\_pal <- c("grey75", "tomato1", "royalblue1")  
  
seasons\_all\_gods\_ZAcat\_group %>%  
 ggplot(aes(x = year, y = n, fill = god\_name)) +  
 scale\_fill\_manual(values = color\_pal) +  
 geom\_bar(position = "stack", stat = "identity") +  
 theme\_classic() +  
 ggtitle("The prevalence of Zeus and Athena in questions and answers over time") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))

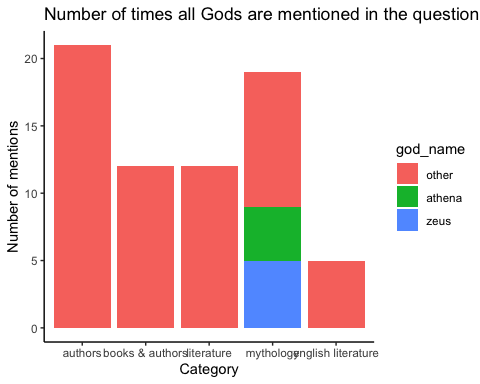


### note that this is testing the prevalence of Zeus and Athena out of all 15 tested gods from all categories

### Then filter to the top 5 categories (authors, books & authors, literature, mythology, english literature, and fill by god)

seasons\_all\_gods\_ZAcat\_top5 <- seasons\_all\_gods\_ZAcat %>%  
 filter(category == "authors" |  
 category == "books & authors" |  
 category == "literature" |   
 category == "mythology" |   
 category == "english literature")  
  
seasons\_all\_gods\_ZAcat\_top5\_group <- seasons\_all\_gods\_ZAcat\_top5 %>%   
 group\_by(category, year, god\_name) %>%   
 count()

seasons\_all\_gods\_ZAcat\_top5\_group %>%  
 ggplot(aes(x = reorder(category, -n), y = n, fill = god\_name)) +  
 geom\_bar(stat = "identity") +  
 theme\_classic() +  
 ggtitle("Number of times all Gods are mentioned in the question or answer") +  
 labs(x = "Category", y = "Number of mentions")



### use forcats to change 'NA' to 'other and reorder so that Zeus and Athena are on the bottom of the chart