Lab Notebook

Grace Acton

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(reticulate)
library(knitr)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

Research Questions

What histories are prioritized in National History Bowl questions, and how is this related to high school history curricula?

To what extent are NHB questions overrepresenting/underrepresenting women, the global south, and military history?

Project Data

This project analyzes thirty National History Bowl packets, each consisting of 56 to 61 questions. These sets represent two academic years: the first year available (2015-2016) and last year available (2021-2022). To glean a representative sample of each year, the first 5 packets from the C, A, and Nationals sets were chosen. B-sets were omitted due to their frequently similar difficulty level to C-sets.

Each question was copied from the PDF packet document available on the IAC website and read for OCR errors. During the transcription process, I read each question for women's names, using Google and Wikipedia

to investigate unfamiliar names. I want to note that gender is a complicated, highly personal identity, and identifying women in history necessitates making many assumptions. I want to acknowledge that transgender, nonbinary, Two Spirit, and other people outside the western gender binary have always existed, whether our historical record includes them or not, and it is impossible to say whether I have correctly identified the gender of all individuals named in this data set.

Secondary to this important acknowledgement, I want to make clear my intent behind two additional metadata categories, `is_fictional` and `is_myth`. Many questions that are **about** a male author mention female characters from their works; I think it is important to differentiate between women historical figures, and fictional women. However, there are also female and feminine figures in many mythological and religious systems, who I don't feel should be categorized the same way as literary characters. For example, there are questions about several Greek and Roman goddesses; even if, say, Athena does not feel "real" to me, she did to the people who worshiped her. As such, religious and mythological figures are separated from literary characters in my metadata scheme. If a named woman is a literary figure, then `is_fictional = TRUE`, whereas if she is a mythological or religious entity, `is_myth = TRUE`.

Lastly, there is the issue of distinguishing questions which mention a woman from questions **about** a woman – that is, questions for which the answer is a named woman. To facilitate the separation of these two categories, the field `is_answer` will have value `TRUE` when a woman is the answer to the question, but remain `FALSE` if she is merely mentioned in a question whose answer is a different historical topic.

Below is a data dictionary for the dataframes `nhbb`, and `nhbb_clean`

```
data_dict <- read.csv("data_dict_nhbb.csv")
kable(data_dict) %>%
  kable_styling(latex_options = "striped",
                full_width = T) %>%
  column_spec(3, width = "3in")
```

Variable	Type	Description	Acceptable.Values
ID	character	A unique identifier for each question block, generated from the year, set, tournament round, and lightning round numbers.	
year	integer	The year the set was played in.	4-digit year
set	integer	The difficulty level of the set. Options are: C, A, or Nationals	c, a, nationals
tournament_round	integer	The round within a tournament that the set was played during. Will always be between 1 and 5, inclusive.	1, 2, 3, 4, 5
game_round	integer	Round number within the game, sometimes called a "quarter". Will always be between 1 and 4, inclusive.	1, 2, 3, 4
lightning_round	integer	If the question was played in round 3, indicates which of the 3 possible lightning rounds the question was part of.	1, 2, 3
question_num	integer	Number of the question within its quarter.	1 - 10
type	character	What format is the question?	tossup, tossup + bonus, or lightning
lightning_lead	character	The lead-in phrase for a lightning round.	
q_text	character	Text of the question.	
a_text	character	Answer to the question.	
bonus_q_text	character	If question is of type "tossup + bonus", this variable contains the text of the bonus question.	
bonus_a_text	character	If question is of type "tossup + bonus", this variable contains the text of the answer to the bonus question.	
woman_named	boolean	Is a woman mentioned by name in the text of the question, answer, bonus question, or bonus answer?	TRUE, FALSE
is_fictional	boolean	Are any of the women named a fictional character?	TRUE, FALSE
is_myth	boolean	Are any of the women named a mythological or religious figure?	TRUE, FALSE
is_answer	boolean	Is a named woman the answer to the question?	TRUE, FALSE
names	character	Names of women mentioned in the question, separated by	

Data Import and Cleaning

Although there has already been an element of human oversight of the questions – namely, my reading the questions as I transcribed them in order to detect OCR errors and identify womens names – some formatting elements of National History Bowl questions are easier to remove or rectify in an automated fashion. Many questions contain pronunciation guides, which aren't necessary to this analysis. Additionally, special characters are used in fourth quarter questions to identify the cut-off points between the 30-point, 20-point, and 10-point portions of the questions. As an exemplar of the complex formatting of a fourth quarter question, I provide Question 5 from Quarter 4 of Round 2 of the 2022 National Championship:

This country resettled Jewish refugees from Nazi Germany in the town of Sosúa [[soh-SOO-ah]] after the Évian Conference. During one massacre in this country, the fate of the victims was determined by how they pronounced the term "perejil' ' [[peh-reh-HEEL]] (+) when soldiers held up a certain herb. One leader of this country had several thousands migrants from a neighboring country killed in the Parsley Massacre and was nicknamed"El (*) Jefe" [[HE-feh]]. For ten points, name this Caribbean country that was ruled for over thirty years by Rafael Trujillo [[troo-HEE-yoh]] from Santo Domingo.

This question has special characters "(+)" and "(*)", as well as double-bracketed pronunciation guides, all of which are superfluous to the text necessary for my analysis. The script `data_prep.R` uses the `stringr` package to remove these additional characters, leaving just the plaintext of the questions.

Additionally, the `data_prep.R` script uses data about the packets and questions to generate a unique identifier for each question and answer. This identifier is ispired by a standard museum object labeling format, which takes the form of `year.accession.box.object`. In this case, I have amended the format to be `year.set.tournament_round.quarter.question_number`. *Quarter 3 questions have an additional digit indicated which thematic lightning round option the question was part of.*

When questions and answers are written to text files, the ID number is followed by an underscore and an additional identifier: `_q` for question, `_a` for answer, `_bonus_q` for quarter 2 bonus questions, or `_bonus_a` for answers to quarter 2 bonus questions. Creating separate files for questions and answers allow them to be analyzed separately, and embedding these identifiers into the file names will allow me to easily separate questions from answers, while having a shared ID number maintains the connection between a question and its corresponding answer.

```
# use data prep script as source - script available in GitHub
source("data_prep.R")

kable(head(nhbb_clean %>% select(ID, q_text, a_text))) %>%
  kable_styling(latex_options = "striped",
                full_width = T) %>%
  column_spec(2, width = "6in")
```

ID	q_text	a_text
2014.c.1.1.1	This man's opinion in Barron v. Baltimore refused to apply Bill of Rights protections to state governments. He was the target of Andrew Jackson's derisive remark "now let him enforce it," and he asserted his Court's power in a case about "midnight appointments." For 10 points, name this Chief Justice whose ruling in Marbury v. Madison asserted the power of judicial review.	John Marshall
2014.c.1.1.2	Pope Leo I convinced Attila the Hun not to perform this kind of action. This kind of action was done by Gauls following the 390 BC Battle of the Allia. In 410 AD, Alaric led an army of Visigoths in performing this action, which St. Jerome identified as the effective end of the Empire. For 10 points, identify this action in which the "eternal city" is pillaged.	sack of Rome
2014.c.1.1.3	The formation of this institution triggered the Key West Agreement on its proper usage, which was negotiated on its behalf by its first secretary, Stuart Symington. This organization conducted Project Blue Book, a study of UFOs, and it operates Area 51. For 10 points, name this branch of the US military which operates B-52s, F-16s, and other fighter jets.	United States Air Force
2014.c.1.1.4	In 1970, the concluding game of this event featured the dramatic entrance of an injured Willis Reed. This event included the "flu game" in 1997 and an upset in 2011 that put a championship in the hands of Jason Terry, Shawn Marion, and Dirk Nowitski. For 10 points, name this annual June event, which awards the Larry O'Brien trophy to teams such as the Bulls, Heat, or Lakers.	NBA Finals
2014.c.1.1.5	This country's banks, which had been largely taken over by Russian organized criminals by 2013, underwent a savings account "haircut." This country's "enosis" movement propelled Archbishop Makarios to power, sparking an invasion that still divides this island. For 10	Cyprus

Statistical Analysis #1: Women's Inclusion/Exclusion

1. For how many questions is the answer a woman?

First, we need to know the total number of questions in the corpus. This is not so simple as knowing the number of rows in the data frame, because bonus questions are included in the same row as the tossup they were written to complement. So, to get the total number of questions, I need to know the number of tossups and the number of bonuses, and add them together.

```
num_bonus <- nrow(nhbb_clean %>% filter(bonus_q_text != "")) # number of bonuses
num_tossup <- nrow(nhbb_clean %>% filter(q_text != ""))

total_qs <- num_bonus + num_tossup
total_qs
```

```
## [1] 1774
```

Next I will count the number of rows which I've tagged as having a woman as the answer/

```
woman_answers <- nhbb_clean %>%
  filter(is_answer == T)
nrow(woman_answers)
```

```
## [1] 43
```

43 out of 1774 questions have a woman as the answer, or 2.42 percent.

The following block of code creates a dataframe consisting only of questions which have a woman as the answer.

```
# filter nhbb_clean to only questions that have a woman as the answer
woman_answers <- nhbb_clean %>%
  filter(is_answer == T)

# establish a new df for writing these questions to

woman_qs <- as.data.frame(matrix(ncol = 4, nrow = nrow(woman_answers)))
colnames(woman_qs) <- c("ID", "question", "answer", "type")

# write only questions for which the woman is the answer to the new df
for (i in 1:nrow(woman_answers)){
  if(woman_answers$a_text[i] %in% woman_answers$names[i] == TRUE |
     woman_answers$names[i] %in% woman_answers$a_text[i] == TRUE |
     grepl(woman_answers$a_text[i], woman_answers$names[i]) == TRUE |
     grepl(woman_answers$names[i], woman_answers$a_text[i]) == TRUE) {
    woman_qs$ID[i] <- woman_answers$ID[i]
    woman_qs$question[i] <- woman_answers$q_text[i]
    woman_qs$answer[i] <- woman_answers$a_text[i]
    woman_qs$type[i] <- "tossup/lightning"
  } else if(woman_answers$bonus_a_text[i] %in% woman_answers$names[i] == TRUE |
            woman_answers$names[i] %in% woman_answers$bonus_a_text[i] == TRUE |
            grepl(woman_answers$bonus_a_text[i], woman_answers$names[i]) == TRUE |
```

```
            grepl(woman_answers$names[i], woman_answers$bonus_a_text[i]) == TRUE){
    woman_qs$ID[i] <- woman_answers$ID[i]
    woman_qs$question[i] <- woman_answers$bonus_q_text[i]
    woman_qs$answer[i] <- woman_answers$bonus_a_text[i]
    woman_qs$type[i] <- "bonus"
  }
}

# fix the "type" column and include lightning lead-ins to question text

for (j in 1:nrow(woman_qs)) {
  if(grepl("lightning", woman_qs$ID[j]) == TRUE){
    woman_qs$type[j] <- "lightning"
    woman_qs$question[j] <- paste(woman_answers$lightning_lead[j], woman_answers$q_text[j], sep = "")
  } else if(woman_qs$type[j] == "tossup/lightning" &
            grepl("lightning", woman_qs$ID[j]) == FALSE){
    woman_qs$type[j] <- "tossup"
  }
}

# print as kable table
kable(woman_qs) %>%
  kable_styling(full_width = T) %>%
  column_spec(1, width = "15%") %>%
  column_spec(2, width = "50%")
```

ID	question	answer	type
2014.c.3.1.9	This author fictionalized Vice-President Aaron Burr in her novel The Minister's Wooing and depicted fugitives in the Great Dismal Swamp in Dred. She wrote a "key to" her major novel, which led Abraham Lincoln to describe her as "the little lady who started this big war." For 10 points, name this author who created Simon Legree in the anti-slavery book Uncle Tom's Cabin.	Harriet Beecher Stowe	tossup
2014.c.3.3.lightning.2.7	In the 1990s, who or what was the...Five-member "girl group" whose song "Wannabe" led the British pop culture revival?	The Spice Girls	lightning
2014.c.5.1.10	This military leader was declared a martyr by Pope Callixtus III, a generation after this saint's capture at Compiègne. Pierre Cauchon led the court which condemned her to be burned at the stake in 1431 following her relief of the siege of Orléans. For 10 points, name this teenage French patriot of the Hundred Years War.	Joan of Arc	tossup
2014.c.5.2.8	This goddess, who was born from the left eye of Izanagi, is worshipped at a "grand shrine" that is torn down and rebuilt every twenty years. She is believed to be the ancestor of the "tenno," who sits on the Chrysanthemum Throne. For 10 points, name this sun goddess, the sister of Susanowo, and the patroness of the Imperial house of Japan.	Amaterasu	tossup
2014.c.5.3.lightning.3.3	The Spanish Armada was... Sent against what Queen of England?	Elizabeth I	lightning

2. How many questions mention a woman by name?

To answer this question, I can again use my metadata tags and the `dplyr::filter()` function.

```
woman_named <- nhbb_clean %>%
  filter(woman_named == T)

# number of questions that contain a named woman
nrow(woman_named)
```

```
## [1] 207
```

However, some of these may be questions for which the *answer* is a woman, but not mention a woman by name in the question text. It would be more accurate to filter out those questions which have a woman as the answer, and only add them back in if they contain another woman's name.

```
# establish a new df for writing these questions to

woman_named_df <- as.data.frame(matrix(ncol = 4, nrow = nrow(nhbb_clean)))
colnames(woman_named_df) <- c("ID", "names", "answer", "type")

# write questions to this new df
nhbb_clean <- nhbb_clean %>%
  filter(is.na(woman_named) == FALSE)

for (i in 1:nrow(nhbb_clean)) {
  if(nhbb_clean$woman_named[i] == TRUE & nhbb_clean$is_answer[i] == FALSE){
      woman_named_df$ID[i] <- nhbb_clean$ID[i]
      woman_named_df$names[i] <- nhbb_clean$names[i]
      woman_named_df$answer[i] <- nhbb_clean$a_text[i]
      woman_named_df$type[i] <- nhbb_clean$type[i]
  } else if(nhbb_clean$woman_named[i] == TRUE & nhbb_clean$is_answer[i] == TRUE) {
      names_vec <- as.list(strsplit(nhbb_clean$names[i], ", ")[[1]])
      if(length(names_vec) > 1) {
        woman_named_df$ID[i] <- nhbb_clean$ID[i]
        woman_named_df$names[i] <- nhbb_clean$names[i]
        woman_named_df$answer[i] <- nhbb_clean$a_text[i]
        woman_named_df$type[i] <- nhbb_clean$type[i]
      }
    }
}

woman_named_df <- woman_named_df %>%
  filter(is.na(names) == FALSE) %>%
  filter(names != answer)
```

So, of those 1774 total questions, 10.03 percent mention a woman by name.

3. Has the proportion of questions and answers that mention women changed over time?

Let's get the IDs of questions that name a woman or have a woman answer, and use them to filter the original NHBB data.

```
woman_ids <- unique(c(woman_answers$ID, woman_named_df$ID))

nhbb_women <- nhbb_clean %>%
  filter(ID %in% woman_ids) %>%
  mutate(q_or_a = if_else(is_answer == TRUE, "answer", "question"))
```

```
women_year <- nhbb_women %>%
  group_by(year) %>%
  count(name = "women")

total_year <- nhbb_clean %>%
  group_by(year) %>%
  count(name = "total")

years <- left_join(women_year, total_year, by = "year")
years <- years %>%
  mutate(proportion = round(women/total, digits = 4))

years
```

```
## # A tibble: 4 x 4
## # Groups:   year [4]
##    year women total proportion
##   <int> <int> <int>      <dbl>
## 1  2014    29   255      0.114
## 2  2015    77   494      0.156
## 3  2021    65   509      0.128
## 4  2022    36   255      0.141
```

Using a Chi-Squared test will let me see if the proportion of questions that feature women is likely to be related to the year the questions were written.

```
# variables: year and proportion
# null hypothesis: the proportion of questions that feature women is unrelated to the year the set was

chisq_data <- as.data.frame(years)
rownames(chisq_data) <- chisq_data$year
chisq_data <- chisq_data %>%
  select(-proportion) %>%
  mutate(no_women = total-women) %>%
  select(-total, -year)


chisq <- chisq.test(chisq_data)
chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  chisq_data
## X-squared = 3.0691, df = 3, p-value = 0.3811
```

```
chisq$observed
```

```
##      women no_women
## 2014    29      226
## 2015    77      417
## 2021    65      444
## 2022    36      219
```

```
chisq$expected
```

```
##         women no_women
## 2014 34.88764 220.1124
## 2015 67.58625 426.4137
## 2021 69.63847 439.3615
## 2022 34.88764 220.1124
```

The Chi-squared test of independence produced a p-value of 0.3811, meaning that any difference in the proportion of questions about women from year to year is likely due to randomness. Thus, the proportion of questions about women is probably *not* related to the year the questions were written.

4. Is the proportion of questions about women related to the difficulty level of the set?

This is the same sort of analytical approach I applied to the years. I'll once again conduct a chi-squared test to see if these two variables are independent of each other.

```
women_set <- nhbb_women %>%
  group_by(set) %>%
  count(name = "women")

total_set <- nhbb_clean %>%
  group_by(set) %>%
  count(name = "total")

sets <- left_join(women_set, total_set, by = "set")
sets <- sets %>%
  mutate(proportion = round(women/total, digits = 4))

sets
```

```
## # A tibble: 3 x 4
## # Groups:   set [3]
##   set       women total proportion
##   <chr>     <int> <int>      <dbl>
## 1 a            67   504      0.133
## 2 c            61   510      0.120
## 3 nationals    79   499      0.158
```

```
# variables: year and proportion
# null hypothesis: the proportion of questions that feature women is unrelated to the year the set was
```

```r
chisq_data <- as.data.frame(sets)
rownames(chisq_data) <- chisq_data$set
chisq_data <- chisq_data %>%
  select(-proportion) %>%
  mutate(no_women = total-women) %>%
  select(-total, -set)


chisq <- chisq.test(chisq_data)
chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  chisq_data
## X-squared = 3.2963, df = 2, p-value = 0.1924
```

```r
chisq$observed
```

```
##           women no_women
## a            67      437
## c            61      449
## nationals    79      420
```

```r
chisq$expected
```

```
##              women no_women
## a         68.95440 435.0456
## c         69.77528 440.2247
## nationals 68.27032 430.7297
```

Once again, these two variables are likely not related.