Introduction to Data Science

Problem Set 4

Alicia Helfrich and Allie Gleich

```
#\| label: rmarkdown-setup #\| echo: false #\| include: false

library(tidyverse)
library(knitr)
library(ggplot2)
library(janitor)
library(haven)
```

Set Up and Introduction

The following data set is pulled from the annual NORC General Social Survey for the year 2022. Each year, the survey collects a variety of information from a representative sample of the US population to understand attitudes and perceptions of a variety of social issues. We were interested in looking exploring reported happiness levels of individuals, and how this variable interacted with religion, employment status, and educational levels.

```
base_url <- "https://gss.norc.org/Documents/sas/"
week_url <- "GSS_sas.zip"
pulse_url <- paste0(base_url, week_url)

dir.create("data")

download.file(
pulse_url,
destfile = "data/GSS_sas.zip",
mode = "wb"
)
```

```r
zip_file <- "data/GSS_sas.zip"

unzip(zip_file, exdir= "data")

gss <- (read_sas("data/GSS_sas/gss7222_r1.sas7bdat")) %>%
  select(YEAR,ID,WRKSTAT,MARITAL,AGE,EDUC,SEX,RACE,BORN,
         GRANBORN,INCOME,PARTYID,NATAID,RELIG,HAPPY,CONEDUC,
         MYSKILLS,HISPANIC,BALLOT,ETHWORLD1,ETHWORLD2,ETHWORLD3,
         ETHWORLD4,ETHWORLD5,ETHWORLD6,ETHWORLD7,ETHWORLD8,ETHWORLD9
         ) %>%
    filter(YEAR == 2022)

color_palette <- c("#E41A1C", "#377EB8", "#4DAF4A")

color_palette_discrete <- c(1== "#E41A1C", 2 == "#377EB8",3 == "#4DAF4A")

clean_names(gss)
```

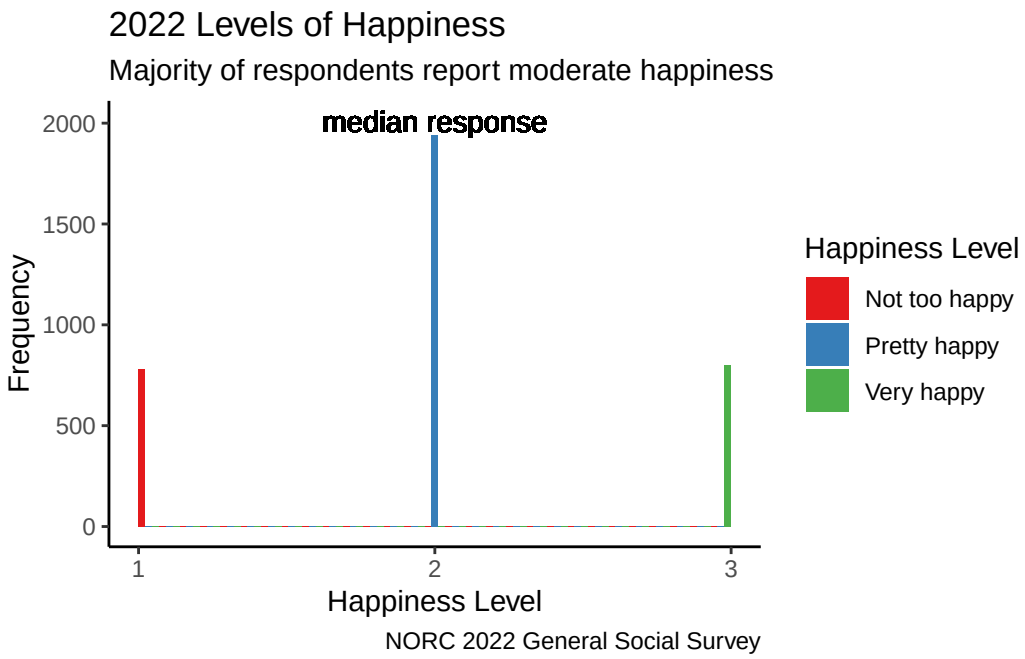
```
# A tibble: 3,544 x 28
    year    id wrkstat marital   age  educ   sex  race  born granborn income
   <dbl> <dbl>   <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>  <dbl>
 1  2022     1       1       3    72    16     2     1     1        4     12
 2  2022     2       5       1    80    18     1     1     1       NA     NA
 3  2022     3       1       3    57    12     2     1     1        1     12
 4  2022     4       3       5    23    16     2     1     1        0     12
 5  2022     5       8       5    62    14     1     1     1        2     12
 6  2022     6       1       5    27    12     1     1     1        0     12
 7  2022     7       2       5    20    12     2     3     1        4     12
 8  2022     8       1       1    47    16     1     1     1        4     12
 9  2022     9       1       5    31    12     2     1     1        0     11
10  2022    10       5       5    72    12     2    NA     1       NA      9
# i 3,534 more rows
# i 17 more variables: partyid <dbl>, nataid <dbl>, relig <dbl>, happy <dbl>,
#   coneduc <dbl>, myskills <dbl>, hispanic <dbl>, ballot <dbl>,
#   ethworld1 <dbl>, ethworld2 <dbl>, ethworld3 <dbl>, ethworld4 <dbl>,
#   ethworld5 <dbl>, ethworld6 <dbl>, ethworld7 <dbl>, ethworld8 <dbl>,
#   ethworld9 <dbl>
```

Data Visualization #1: Exploring Happiness Levels in 2022

```
gss$color_group <- cut(gss$HAPPY, breaks = c(.5,1.5,2.5,3.5),
                       labels = c("Not too happy","Pretty happy","Very happy"
                                  ))

color_palette_discrete <- c("Not too happy" = "#E41A1C",
                            "Pretty happy" = "#377EB8",
                            "Very happy" = "#4DAF4A")

ggplot(data = gss, aes(x = HAPPY, fill = color_group)) +
  geom_histogram(position = "dodge", width = 10.5, na.rm = TRUE) +  scale_x_continuous
  scale_y_continuous(limits = c(0, NA)) +
  scale_fill_manual(values = color_palette_discrete)+
  theme_classic() +
  geom_text(aes(x =2,y = 2010, label = "median response",
                show.legend = FALSE)) +
  labs(
    title = "2022 Levels of Happiness",
    subtitle = "Majority of respondents report moderate happiness",
    x = "Happiness Level",
    y = "Frequency",
    caption = "NORC 2022 General Social Survey",
    fill = "Happiness Level")
```



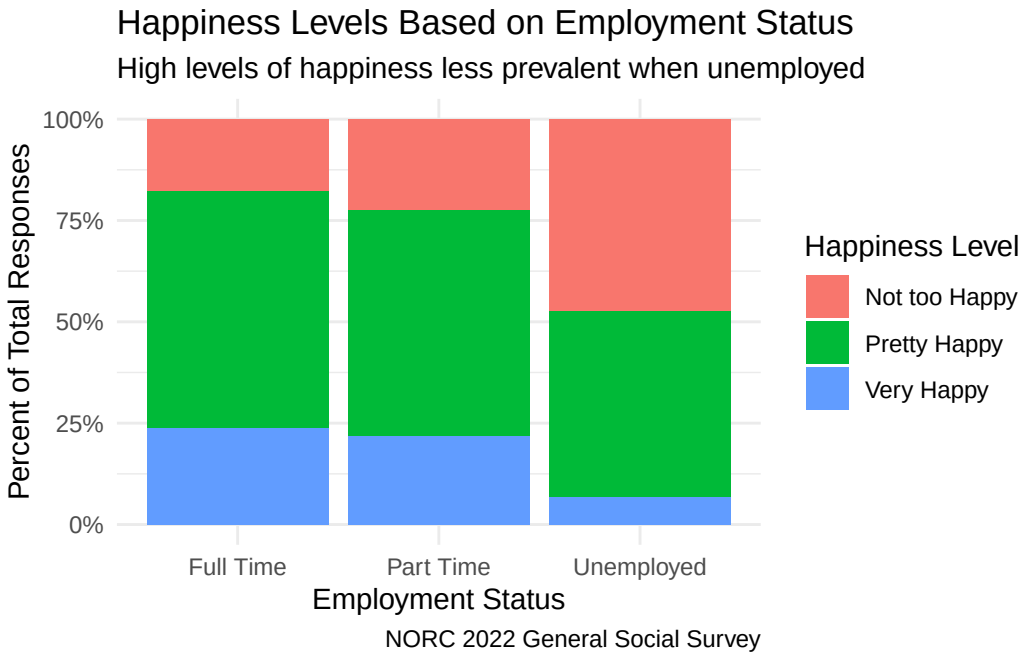
First, we sought to understand overall reported happiness levels. The graph shows that the majority of the respondents ranked their happiness at a 2 ("Pretty happy") out of 3 ("Very happy"). The least amount of people ranked themselves at 1, coded to represent "Not Too Happy", which would indicate that people are generally happy. To continue making public policy decisions that are targeted at improving people's overall happiness/wellbeing, we should further explore the factors that appear to have an association with happiness levels.

Data Visualization #2: Investigating a potential relationship between labor force status and happiness levels

```
labfor <-
  gss %>%
    select(WRKSTAT,HAPPY) %>%
  mutate(
    lab_stat = case_when(
      WRKSTAT == 1 ~ "Full Time",
      WRKSTAT == 2 ~ "Part Time",
      WRKSTAT == 3 ~ "Other",
      WRKSTAT == 4 ~ "Unemployed",
```

```
      WRKSTAT > 4 ~ "Out of Labor Force",
      WRKSTAT < 1 ~ "N/A",
    ),
    happy_text = case_when(
      HAPPY == 1 ~ "Very Happy",
      HAPPY == 2 ~ "Pretty Happy",
      HAPPY == 3 ~ "Not too Happy",
      HAPPY < 1 ~ "N/A"
    )
  ) %>%
  filter(lab_stat == "Full Time" | lab_stat == "Part Time" | lab_stat == "Unemployed",
         HAPPY != "N/A"
         )

ggplot(data = labfor,aes(x = lab_stat, fill = happy_text)) +
   geom_bar(position = "fill") +
  labs(title = "Happiness Levels Based on Employment Status",
       subtitle = "High levels of happiness less prevalent when unemployed",
       x = "Employment Status",
       y = "Percent of Total Responses",
      caption = "NORC 2022 General Social Survey" ) +
  scale_y_continuous(labels = scales::percent_format(scale = 100)) +
  scale_fill_discrete(name = "Happiness Level") +
  theme_minimal()
```



Jobs provide a source of purpose, direction, and financial stability for individuals, and therefore impacts general satisfaction. We hypothesized that individuals that are working full time or part time would have a higher proportion of individuals reporting moderate to high levels of happiness, compared to those that are unemployed. The above graph displays the proportion of respondents that reported that they were "Very Happy", "Pretty Happy", and "Not too Happy" based on their employment status at the time of the response. We can see that individuals working full time or part time reported similar levels of happiness. However, there is a noticeably lower proportion of individuals that reported being "very happy" and a visibly higher proportion of "Not too Happy" responses for those that were unemployed. This would appear to support our theory.

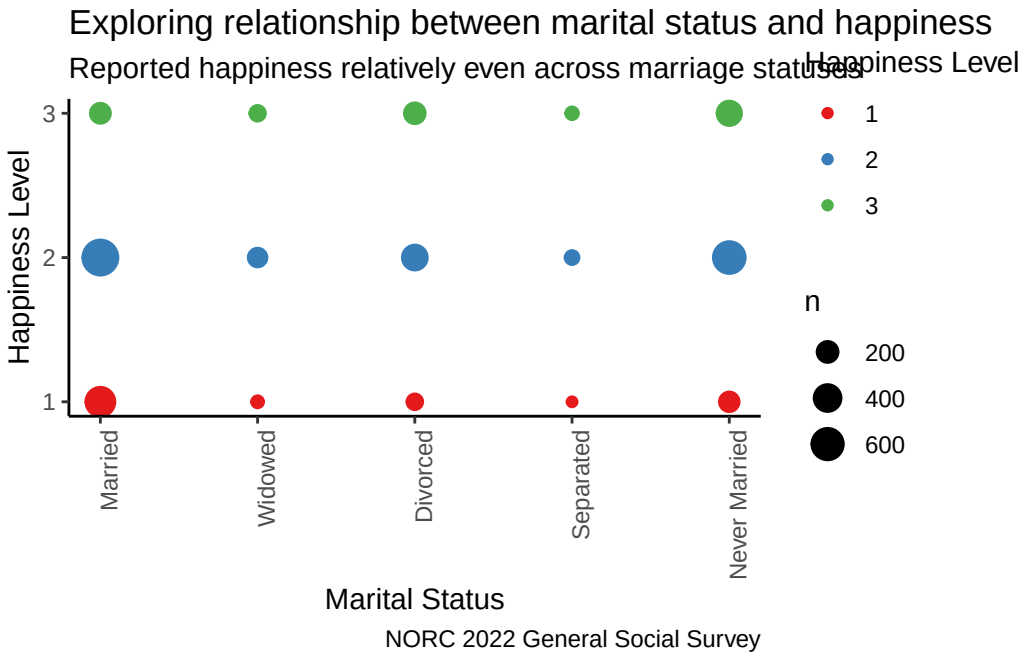
Data Visualization #3: Exploring how religious identity relates to reported happiness levels

```
custom_labels <- c("1" = "Not too happy", "2" = "Somewhat happy", "3" = "Very happy")

gss_marital_happy <- gss %>%
  group_by(HAPPY, MARITAL) %>%
  tally()
```

```r
gss_marital_happy %>% filter(!is.na(n)) %>%

ggplot(aes(x = MARITAL, y = HAPPY)) +
  geom_point(aes(size = n, color = as.factor(HAPPY))) +
  scale_x_continuous(breaks = (seq(1, 5, by = 1)),
                     labels=c("Married",
                              "Widowed",
                              "Divorced",
                              "Separated",
                              "Never Married")) +
  scale_y_discrete(labels = c("1" = "Not too happy",
                              "2" = "Somewhat happy",
                              "3" = "Very happy")) +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  scale_y_continuous(breaks = (seq(1,3, by = 1))) +
  labs(
    title = "Exploring relationship between marital status and happiness",
    subtitle = "Reported happiness relatively even across marriage statuses",
      "How Marital Status Impacts Happiness",
     x = "Marital Status",
       y = "Happiness Level",
      caption = "NORC 2022 General Social Survey",
     fill = "Happiness Level") +
    scale_color_manual( name = "Happiness Level",
      values = c("1" = "#E41A1C", "2" = "#377EB8", "3" = "#4DAF4A"))
```



The above graph shows that for all marital statuses, the majority of respondents ranked their happiness at a 2, or "pretty happy". This aligns with Data Visualization #1, which shows that the median happiness level in the General Social Survey is a 2. An interesting finding, however, is that the plot above reveals that for all marital statuses, except for "married", the second most selected happiness rating was a 3, meaning very happy. For those in the "married" category, the second highest ranked happiness rating was 1, being very unhappy. This reveals that there may be a slight negative correlation between being married and being happy. All in all, there are no strong correlations above, which is good news for public policy makers, who don't need to focus on marital status when trying to improve the well being of their constituents.

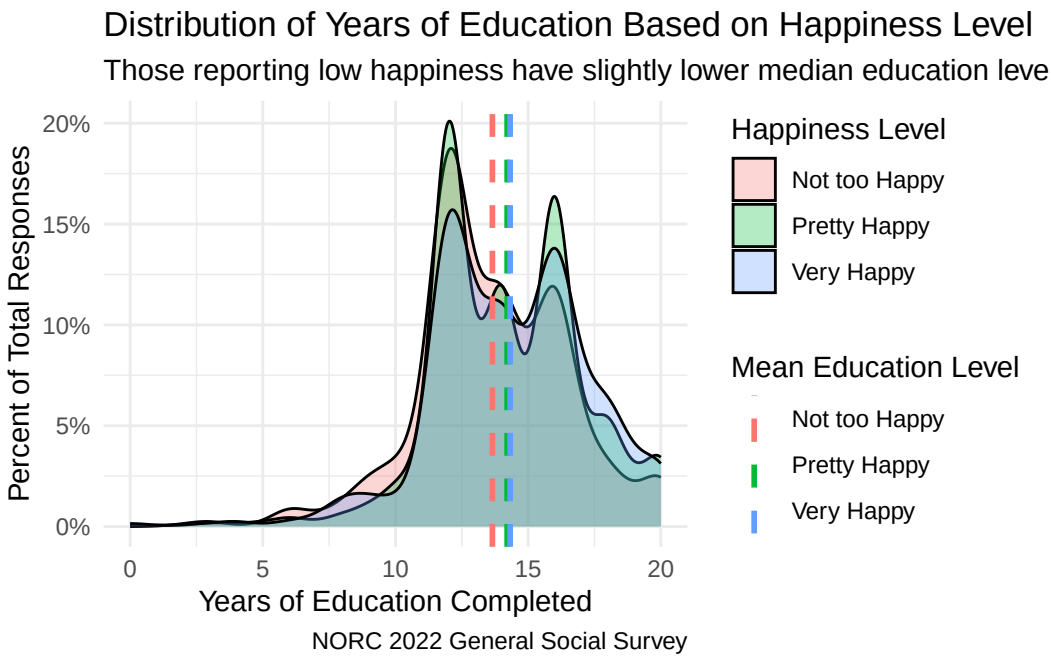
Data Visualization #4: Examining how education level interacts with reported happiness levels

```
#establishing new education data frame
educ_in <-
  gss %>%
    select(EDUC,HAPPY) %>%
  mutate(
```

```
    happy_text = case_when(
      HAPPY == 1 ~ "Very Happy",
      HAPPY == 2 ~ "Pretty Happy",
      HAPPY == 3 ~ "Not too Happy",
      HAPPY < 1 ~ "N/A"
    )
  ) %>%
  filter(happy_text != "N/A")

#creating a data frame containing mean years of education by happiness level
mu_educ <-
  educ_in %>%
    group_by(happy_text) %>%
    summarize(mean_educ =
                mean(EDUC, na.rm = TRUE)
              )

#creating an density plot with mean lines
ggplot(data = educ_in, aes(x = EDUC, fill = happy_text))+
  geom_density(alpha = .3) +
  geom_vline(data = mu_educ, aes(xintercept = mean_educ, color = happy_text), linetype
    scale_y_continuous(labels = scales::percent_format(scale = 100)) +
    labs(title = "Distribution of Years of Education Based on Happiness Level",
         subtitle = "Those reporting low happiness have slightly lower median educatio
         x = "Years of Education Completed",
         y = "Percent of Total Responses",
         color = "Mean Education Level",
         caption = "NORC 2022 General Social Survey") +
    scale_fill_discrete(name = "Happiness Level")+
  theme_minimal()
```



We hypothesize that individuals that higher levels of education could be associated with higher levels of happiness. The graph above shows a density plot of of individuals at each year of education, disaggregated by reported happiness level. The mean education level of individuals at each happiness level is shown by the vertical dashed lines. As shown, individuals that reported that they were "Pretty happy" or "Very happy" had similar mean educational levels, with a visible difference from individuals that reported they were "Not too happy".This could indicate a potential positive correlation between educational level and happiness, but additional investigation should be done to determine that the statistical significance of this relationship, if at all.