EDA

Final Project 01: EDA

Youth Risk Behavior Surveillance System (YRBSS), 2023

Exploratory data analysis

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2
-- 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 becon
```

```
library(Hmisc)
```

```


Attaching package: 'Hmisc'

The following objects are masked from 'package:dplyr':

    src, summarize
```

```
The following objects are masked from 'package:base':

    format.pval, units
```

```
library(naniar)

mdb.get('XXH2023_YRBS_Data.mdb', tables = TRUE)
```

```
[1] "XXHq"  "XXHqn"
```

```
# full data set will all item responses
data <- mdb.get('XXH2023_YRBS_Data.mdb', tables = "XXHq")
```

```
117 variables; Processing variable:1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 2
```

```
# already pre-processed to make each question dichotomous
data_qn<- mdb.get('XXH2023_YRBS_Data.mdb', tables = "XXHqn")
```

```
152 variables; Processing variable:1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 2
```

```
# see data dictionary (pdf included in project folder) for individual item responses and ques

vars <- data |> select(
  q1,  # age
  q2,  # sex
  q19, # forced sexual intercourse (lifetime)
  q20, # sexual violence (12 months)
  q21, # sexual dating violence (12 months)
  q22, # physical dating violence (12 months)
  q33, # current cigarette use (30 days)
  q36, # current electronic vapor use (30 days)
  q42, # current alcohol use (30 days)
  q48, # current marijuana use (30 days)
  q57, # age first sexual intercourse
  q58, # number sexual partners (lifetime)
  q59, # current sexual activity/number sexual partners (3 months)
  q60, # alcohol/drug use during sex (last time)
  q61, # condom use (last time)
  q62, # birth control method use (last time)
  q64, # sexual orientation
  q80, # social media use
  q81, # HIV testing (lifetime)
  q82, # STD testing (12 months)
  q84  # current mental health (30 days)
  )
```

What’s in my data?

```
str(data)
```

```
'data.frame':   20103 obs. of  117 variables:
 $ site    : 'labelled' chr  "XX" "XX" "XX" "XX" ...
  ..- attr(*, "label")= chr "site"
 $ raceeth : 'labelled' int  NA 5 5 5 5 5 5 5 8 3 ...
  ..- attr(*, "label")= chr "raceeth"
 $ q6orig  : 'labelled' chr  "505" "N N" "506" "N N" ...
  ..- attr(*, "label")= chr "q6orig"
 $ q7orig  : 'labelled' chr  "180" "233" "165" "105" ...
  ..- attr(*, "label")= chr "q7orig"
 $ record  : 'labelled' int  1 2 3 4 5 6 7 8 9 10 ...
  ..- attr(*, "label")= chr "record"
 $ orig.rec: 'labelled' logi  NA NA NA NA NA NA ...
  ..- attr(*, "label")= chr "orig_rec"
 $ q1      : 'labelled' int  3 4 5 6 3 5 6 4 4 6 ...
  ..- attr(*, "label")= chr "q1"
 $ q2      : 'labelled' int  1 2 2 1 2 2 2 1 2 1 ...
  ..- attr(*, "label")= chr "q2"
 $ q3      : 'labelled' int  1 1 3 2 1 1 3 1 1 3 ...
  ..- attr(*, "label")= chr "q3"
 $ q4      : 'labelled' int  NA 2 2 2 2 2 2 2 2 2 ...
  ..- attr(*, "label")= chr "q4"
 $ q5      : 'labelled' chr  "  C" "    E" "    E" "    E" ...
  ..- attr(*, "label")= chr "q5"
 $ q6      : 'labelled' num  1.65 NA 1.68 NA 1.85 1.8 1.83 1.52 1.65 1.63 ...
  ..- attr(*, "label")= chr "q6"
 $ q7      : 'labelled' num  81.7 NA 74.8 NA 56.7 ...
  ..- attr(*, "label")= chr "q7"
 $ q8      : 'labelled' int  4 5 5 4 5 4 5 3 5 4 ...
  ..- attr(*, "label")= chr "q8"
 $ q9      : 'labelled' int  4 1 3 1 1 1 1 1 1 1 ...
  ..- attr(*, "label")= chr "q9"
 $ q10     : 'labelled' int  1 1 2 2 1 1 1 1 1 2 ...
  ..- attr(*, "label")= chr "q10"
 $ q11     : 'labelled' int  1 1 3 8 1 1 1 1 1 NA ...
  ..- attr(*, "label")= chr "q11"
 $ q12     : 'labelled' int  1 1 1 1 1 1 1 1 1 1 ...
  ..- attr(*, "label")= chr "q12"
 $ q13     : 'labelled' int  1 1 1 1 1 1 5 1 1 1 ...
  ..- attr(*, "label")= chr "q13"
 $ q14     : 'labelled' int  3 1 1 3 1 5 1 1 1 1 ...
  ..- attr(*, "label")= chr "q14"
 $ q15     : 'labelled' int  1 1 1 1 1 2 1 1 1 1 ...
  ..- attr(*, "label")= chr "q15"
 $ q16     : 'labelled' int  1 2 1 1 3 1 1 1 2 1 ...
  ..- attr(*, "label")= chr "q16"
 $ q17     : 'labelled' int  1 2 1 1 1 1 1 1 2 1 ...
  ..- attr(*, "label")= chr "q17"
 $ q18     : 'labelled' int  2 2 2 2 2 1 2 2 2 2 ...
  ..- attr(*, "label")= chr "q18"
 $ q19     : 'labelled' int  2 2 2 1 2 2 2 2 2 2 ...
  ..- attr(*, "label")= chr "q19"
 $ q20     : 'labelled' int  1 1 1 1 1 1 1 1 1 1 ...
  ..- attr(*, "label")= chr "q20"
 $ q21     : 'labelled' int  2 2 2 2 2 2 2 2 2 2 ...
  ..- attr(*, "label")= chr "q21"
 $ q22     : 'labelled' int  2 2 2 2 1 2 2 2 2 2 ...
  ..- attr(*, "label")= chr "q22"
 $ q23     : 'labelled' int  1 3 1 1 1 5 1 1 1 3 ...
  ..- attr(*, "label")= chr "q23"
 $ q24     : 'labelled' int  2 1 2 2 2 2 2 2 2 2 ...
  ..- attr(*, "label")= chr "q24"
 $ q25     : 'labelled' int  2 2 2 2 2 2 2 2 2 2 ...
  ..- attr(*, "label")= chr "q25"
 $ q26     : 'labelled' int  1 2 1 1 1 2 1 2 1 1 ...
  ..- attr(*, "label")= chr "q26"
 $ q27     : 'labelled' int  2 2 2 2 2 2 2 2 1 NA ...
  ..- attr(*, "label")= chr "q27"
 $ q28     : 'labelled' int  2 2 2 2 1 2 2 2 2 1 ...
  ..- attr(*, "label")= chr "q28"
 $ q29     : 'labelled' int  1 1 1 1 1 1 1 1 2 1 ...
  ..- attr(*, "label")= chr "q29"
 $ q30     : 'labelled' int  1 1 1 1 1 1 1 1 3 1 ...
  ..- attr(*, "label")= chr "q30"
```

```
$ q31     : 'labelled' int  2 2 1 2 1 2 1 2 2 2 ...
 ..- attr(*, "label")= chr "q31"
$ q32     : 'labelled' int  1 1 5 1 5 1 6 1 1 1 ...
 ..- attr(*, "label")= chr "q32"
$ q33     : 'labelled' int  1 1 1 1 1 1 1 1 1 1 ...
 ..- attr(*, "label")= chr "q33"
$ q34     : 'labelled' int  1 1 1 1 1 1 1 1 1 1 ...
 ..- attr(*, "label")= chr "q34"
$ q35     : 'labelled' int  2 2 1 1 1 1 1 2 1 2 ...
 ..- attr(*, "label")= chr "q35"
$ q36     : 'labelled' int  1 1 2 7 1 1 7 1 1 1 ...
 ..- attr(*, "label")= chr "q36"
$ q37     : 'labelled' int  1 1 2 8 1 1 8 1 1 1 ...
 ..- attr(*, "label")= chr "q37"
$ q38     : 'labelled' int  1 1 1 1 3 1 1 1 1 1 ...
 ..- attr(*, "label")= chr "q38"
$ q39     : 'labelled' int  1 1 1 1 1 1 1 1 1 1 ...
 ..- attr(*, "label")= chr "q39"
$ q40     : 'labelled' int  1 1 3 3 3 1 3 1 1 1 ...
 ..- attr(*, "label")= chr "q40"
$ q41     : 'labelled' int  1 1 6 5 5 2 5 1 5 1 ...
 ..- attr(*, "label")= chr "q41"
$ q42     : 'labelled' int  1 1 1 2 2 1 2 1 1 1 ...
 ..- attr(*, "label")= chr "q42"
$ q43     : 'labelled' int  1 1 1 1 1 1 2 1 1 1 ...
 ..- attr(*, "label")= chr "q43"
$ q44     : 'labelled' int  1 1 1 2 2 1 6 1 1 1 ...
 ..- attr(*, "label")= chr "q44"
$ q45     : 'labelled' int  1 1 1 8 2 1 8 1 1 1 ...
 ..- attr(*, "label")= chr "q45"
$ q46     : 'labelled' int  1 1 1 4 1 1 2 1 1 1 ...
 ..- attr(*, "label")= chr "q46"
$ q47     : 'labelled' int  1 1 1 5 1 1 6 1 1 1 ...
 ..- attr(*, "label")= chr "q47"
$ q48     : 'labelled' int  1 1 1 2 1 1 1 1 1 1 ...
 ..- attr(*, "label")= chr "q48"
$ q49     : 'labelled' int  2 1 1 1 1 1 1 1 1 1 ...
 ..- attr(*, "label")= chr "q49"
$ q50     : 'labelled' int  1 1 1 1 1 1 1 1 1 1 ...
 ..- attr(*, "label")= chr "q50"
$ q51     : 'labelled' int  1 1 2 2 1 1 1 1 1 1 ...
 ..- attr(*, "label")= chr "q51"
$ q52     : 'labelled' int  1 1 1 1 1 1 1 1 1 1 ...
```

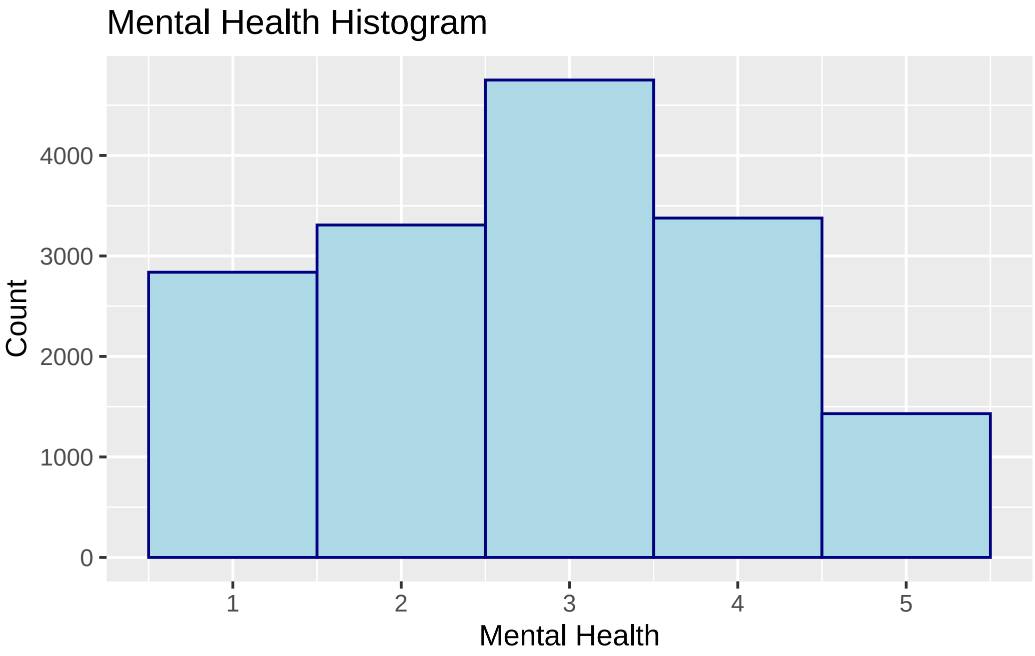
```
 ..- attr(*, "label")= chr "q52"
$ q53     : 'labelled' int  NA 1 1 1 1 1 1 1 1 1 ...
 ..- attr(*, "label")= chr "q53"
$ q54     : 'labelled' int  1 1 1 1 1 1 1 1 1 1 ...
 ..- attr(*, "label")= chr "q54"
$ q55     : 'labelled' int  1 1 1 1 1 1 1 1 1 1 ...
 ..- attr(*, "label")= chr "q55"
$ q56     : 'labelled' int  2 2 1 1 1 1 1 2 1 1 ...
 ..- attr(*, "label")= chr "q56"
$ q57     : 'labelled' int  1 1 7 5 5 3 7 1 6 7 ...
 ..- attr(*, "label")= chr "q57"
$ q58     : 'labelled' int  1 1 4 2 2 3 2 1 3 2 ...
 ..- attr(*, "label")= chr "q58"
$ q59     : 'labelled' int  1 1 3 3 3 2 3 1 2 3 ...
 ..- attr(*, "label")= chr "q59"
$ q60     : 'labelled' int  1 1 3 3 3 3 3 1 3 3 ...
 ..- attr(*, "label")= chr "q60"
$ q61     : 'labelled' int  1 1 2 3 2 3 2 1 3 2 ...
 ..- attr(*, "label")= chr "q61"
$ q62     : 'labelled' int  1 1 3 3 4 2 6 1 1 3 ...
 ..- attr(*, "label")= chr "q62"
$ q63     : 'labelled' int  1 1 4 3 2 2 2 1 3 3 ...
 ..- attr(*, "label")= chr "q63"
$ q64     : 'labelled' int  2 1 3 1 1 1 1 3 3 3 ...
 ..- attr(*, "label")= chr "q64"
$ q65     : 'labelled' int  1 1 1 1 1 1 1 1 1 1 ...
 ..- attr(*, "label")= chr "q65"
$ q66     : 'labelled' int  2 4 4 3 3 3 3 4 3 3 ...
 ..- attr(*, "label")= chr "q66"
$ q67     : 'labelled' int  1 1 1 2 3 2 4 1 2 2 ...
 ..- attr(*, "label")= chr "q67"
$ q68     : 'labelled' int  2 1 3 6 3 6 1 1 2 7 ...
 ..- attr(*, "label")= chr "q68"
$ q69     : 'labelled' int  3 1 2 6 3 3 2 1 2 6 ...
 ..- attr(*, "label")= chr "q69"
$ q70     : 'labelled' int  1 1 1 1 4 1 1 1 1 4 ...
 ..- attr(*, "label")= chr "q70"
$ q71     : 'labelled' int  1 1 2 1 2 1 2 1 2 7 ...
 ..- attr(*, "label")= chr "q71"
$ q72     : 'labelled' int  1 1 1 1 1 1 1 1 1 4 ...
 ..- attr(*, "label")= chr "q72"
$ q73     : 'labelled' int  1 1 2 1 2 1 3 4 1 4 ...
 ..- attr(*, "label")= chr "q73"
```

```
$ q74     : 'labelled' int  2 2 4 5 3 7 5 4 2 6 ...
 ..- attr(*, "label")= chr "q74"
$ q75     : 'labelled' int  4 5 8 1 3 8 1 2 4 5 ...
 ..- attr(*, "label")= chr "q75"
$ q76     : 'labelled' int  1 5 8 3 8 6 8 1 1 1 ...
 ..- attr(*, "label")= chr "q76"
$ q77     : 'labelled' int  2 6 4 6 1 6 1 1 1 1 ...
 ..- attr(*, "label")= chr "q77"
$ q78     : 'labelled' int  2 2 1 2 1 3 1 2 3 1 ...
 ..- attr(*, "label")= chr "q78"
$ q79     : 'labelled' int  1 1 1 1 2 2 2 1 1 2 ...
 ..- attr(*, "label")= chr "q79"
$ q80     : 'labelled' int  6 4 8 8 6 8 8 6 6 6 ...
 ..- attr(*, "label")= chr "q80"
$ q81     : 'labelled' int  2 2 2 2 2 2 2 2 2 3 ...
 ..- attr(*, "label")= chr "q81"
$ q82     : 'labelled' int  2 2 2 2 2 2 2 2 2 3 ...
 ..- attr(*, "label")= chr "q82"
$ q83     : 'labelled' int  1 2 1 1 1 1 1 5 2 2 ...
 ..- attr(*, "label")= chr "q83"
$ q84     : 'labelled' int  1 3 2 3 3 NA 3 2 3 1 ...
 ..- attr(*, "label")= chr "q84"
$ q85     : 'labelled' int  3 5 1 4 3 2 2 1 1 2 ...
 ..- attr(*, "label")= chr "q85"
$ q86     : 'labelled' int  1 1 1 1 1 1 7 1 1 1 ...
 ..- attr(*, "label")= chr "q86"
$ q87     : 'labelled' int  NA 4 3 3 4 5 3 7 4 7 ...
 ..- attr(*, "label")= chr "q87"
$ q88     : 'labelled' int  2 2 2 1 2 2 2 2 2 2 ...
 ..- attr(*, "label")= chr "q88"
$ q89     : 'labelled' int  NA 2 3 1 1 3 4 2 1 2 ...
 ..- attr(*, "label")= chr "q89"
$ q90     : 'labelled' int  1 2 2 1 1 4 2 1 1 1 ...
 ..- attr(*, "label")= chr "q90"
$ q91     : 'labelled' int  1 2 2 1 1 1 1 1 1 1 ...
 ..- attr(*, "label")= chr "q91"
$ q92     : 'labelled' int  2 1 1 1 1 1 1 1 1 1 ...
 ..- attr(*, "label")= chr "q92"
$ q93     : 'labelled' int  1 1 1 1 1 1 1 1 1 1 ...
 ..- attr(*, "label")= chr "q93"
 [list output truncated]
```

Each of the questions asked in the National High School Youth Risk Behavior Survey (YRBS) has its own column in the dataset. Each row represents a participant in the survey, so their responses associated to each of questions is a single observation. The dataset mostly consists of integers that correlate to each of the possible multiple choice answers. So, if for question 1 (q1), when asked how old they were, the observation is 1. That would mean that person responded with A), they were 12 years or younger.

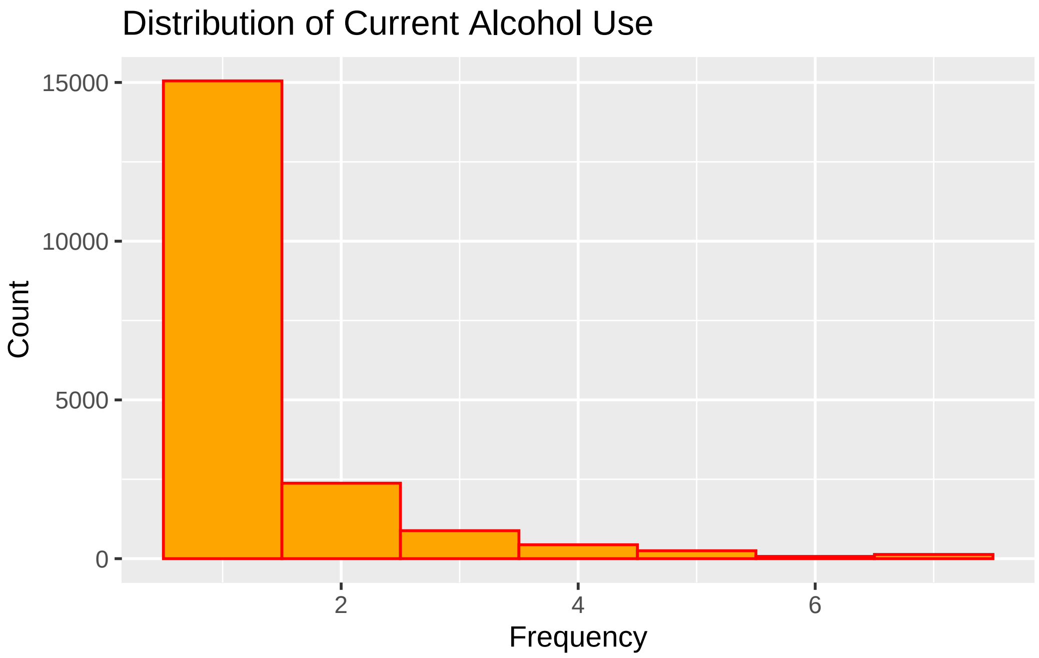
Data Variation

```
# demographic data variation
ggplot(data = data) +
  geom_histogram(mapping = aes(x= q84), bins = 5, color = "navy", fill = "lightblue")+
  labs(x = "Mental Health",
       y = "Count",
       title = "Mental Health Histogram")
```



Frequency of poor mental health (1 - Never, 2 - Rarely, 3 - Sometimes, 4 - Most of the time, 5 - Always) looks normally distributed, perhaps with a slight right skew.

```
# first drink
ggplot(data = data) +
  geom_histogram(mapping = aes(x= q42), bins = 7, color = "red", fill = "orange")+
  labs(x = "Frequency",
       y = "Count",
       title = "Distribution of Current Alcohol Use")
```



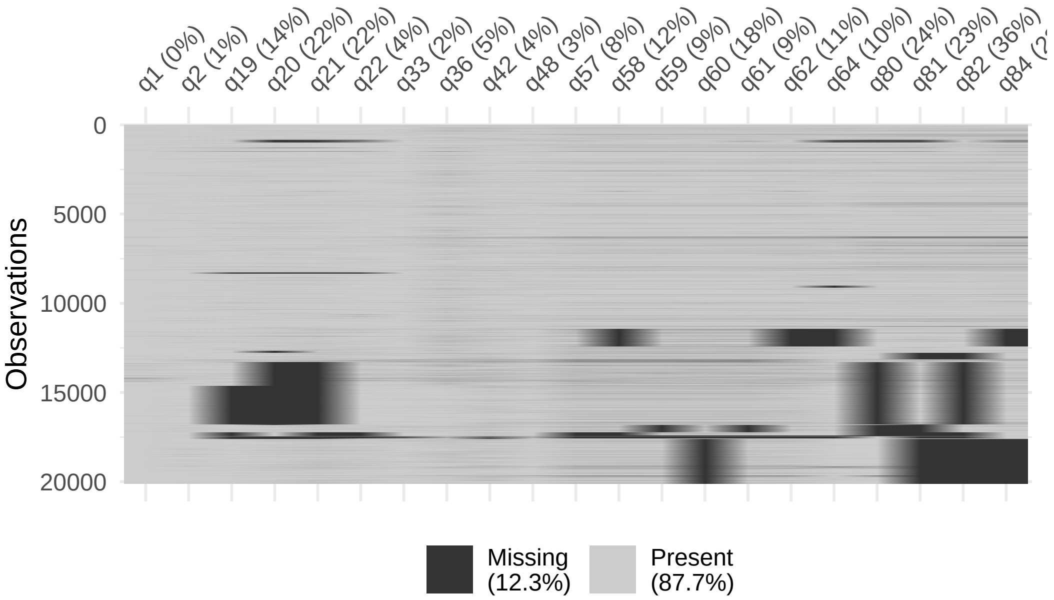
However, for distribution of current alcohol use there was dramatic right skew as the vast majority responding to the survey haven't consumed alcohol. This is unsurprising given that alcohol use is a lower frequency behavior, let along significant/frequent use.

Missing Data

```
# missing data for complete data
summary(vars)
```

```
       q1              q2             q19             q20       
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:4.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000  
 Median :5.000   Median :2.000   Median :2.000   Median :1.000  
 Mean   :4.893   Mean   :1.504   Mean   :1.904   Mean   :1.229  
 3rd Qu.:6.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:1.000  
 Max.   :7.000   Max.   :2.000   Max.   :2.000   Max.   :5.000  
 NA's   :98      NA's   :158     NA's   :2801    NA's   :4351  
      q21             q22             q33             q36       
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000  
 Median :2.000   Median :2.000   Median :1.000   Median :1.000  
 Mean   :1.975   Mean   :1.729   Mean   :1.103   Mean   :1.654  
 3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:1.000  
 Max.   :6.000   Max.   :6.000   Max.   :7.000   Max.   :7.000  
 NA's   :4504    NA's   :837     NA's   :364     NA's   :1012   
      q42             q48             q57             q58       
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000  
 Median :1.000   Median :1.000   Median :1.000   Median :1.000  
 Mean   :1.396   Mean   :1.474   Mean   :2.509   Mean   :1.714  
 3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:5.000   3rd Qu.:2.000  
 Max.   :7.000   Max.   :6.000   Max.   :8.000   Max.   :7.000  
 NA's   :901     NA's   :503     NA's   :1530    NA's   :2458   
      q59             q60             q61             q62              q64       
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000  
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:1.000  
 Median :1.000   Median :1.000   Median :1.000   Median :1.00   Median :1.000  
 Mean   :1.617   Mean   :1.607   Mean   :1.464   Mean   :1.92   Mean   :1.707  
 3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:3.00   3rd Qu.:2.000  
 Max.   :8.000   Max.   :3.000   Max.   :3.000   Max.   :8.00   Max.   :6.000  
 NA's   :1764    NA's   :3655    NA's   :1791    NA's   :2159   NA's   :2003   
      q80             q81             q82             q84       
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:6.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000  
 Median :6.000   Median :2.000   Median :2.000   Median :3.000  
 Mean   :6.049   Mean   :2.082   Mean   :2.006   Mean   :2.825  
 3rd Qu.:8.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:4.000  
 Max.   :8.000   Max.   :3.000   Max.   :3.000   Max.   :5.000  
 NA's   :4900    NA's   :4584    NA's   :7325    NA's   :4398   
```

```
vis_miss(vars, warn_large_data = FALSE)
```



There is a noticeable amount of missing data (12.3%), which is clearly evident in a missing data visualization. Importantly, missing data does not seem to be at random, with specific observations having large portions of missing data. Implications of missing data and steps for corrective action are outlined in the project description.

Covariation

```r
# take numeric data out
numeric_data <- vars[sapply(vars, is.numeric)]
corr_matrix <- cor(numeric_data, use = "complete.obs", method = "pearson")
print(corr_matrix)
```

```
               q1           q2         q19         q20         q21         q22
q1   1.000000000  0.043289131 -0.05266049  0.01156255  0.01835779  0.08799288
q2   0.043289131  1.000000000  0.18348539 -0.17688813 -0.12063521 -0.04174763
q19 -0.052660492  0.183485387  1.00000000 -0.44721327 -0.31282691 -0.23789477
q20  0.011562550 -0.176888133 -0.44721327  1.00000000  0.64060740  0.29375441
q21  0.018357791 -0.120635207 -0.31282691  0.64060740  1.00000000  0.33978955
q22  0.087992878 -0.041747633 -0.23789477  0.29375441  0.33978955  1.00000000
q33  0.047377692  0.029717170 -0.08637008  0.13426348  0.13523305  0.16734432
q36  0.102978674 -0.066709582 -0.21723392  0.19195904  0.14762353  0.27060021
```

```
q42  0.122541945 -0.019832343 -0.16164781  0.17271540  0.12611752  0.23953625
q48  0.107186838 -0.018268222 -0.18331643  0.15092136  0.12501591  0.21236707
q57  0.345259368  0.009691592 -0.16147668  0.11317051  0.10752976  0.32492519
q58  0.248174649  0.033683541 -0.27912406  0.19851714  0.16658013  0.33699225
q59  0.259971316  0.014415880 -0.24386555  0.19913098  0.19019546  0.38587573
q60  0.269648251  0.018651688 -0.22008578  0.14671880  0.12705852  0.35105833
q61  0.260398006 -0.016004990 -0.27319596  0.19813929  0.16357330  0.36867080
q62  0.231962888  0.026462104 -0.19864691  0.14710241  0.13924876  0.31322774
q64 -0.003214449 -0.215079029 -0.12193217  0.11936215  0.06613979 -0.01314926
q80  0.030829901 -0.113979701 -0.05004099  0.06229103  0.03897358  0.13262835
q81 -0.106773520  0.029433706  0.06402879 -0.04451474 -0.02717716 -0.07882149
q82 -0.083390860  0.072585261  0.06630414 -0.04592658 -0.06062538 -0.08471064
q84  0.015470561 -0.315465596 -0.21375873  0.23949728  0.14836363  0.12691559
            q33         q36          q42         q48          q57         q58
q1   0.04737769  0.10297867  0.122541945  0.10718684  0.345259368  0.24817465
q2   0.02971717 -0.06670958 -0.019832343 -0.01826822  0.009691592  0.03368354
q19 -0.08637008 -0.21723392 -0.161647815 -0.18331643 -0.161476682 -0.27912406
q20  0.13426348  0.19195904  0.172715401  0.15092136  0.113170512  0.19851714
q21  0.13523305  0.14762353  0.126117515  0.12501591  0.107529761  0.16658013
q22  0.16734432  0.27060021  0.239536249  0.21236707  0.324925194  0.33699225
q33  1.00000000  0.35961029  0.322483363  0.31540781  0.093725399  0.22056688
q36  0.35961029  1.00000000  0.499468612  0.60480937  0.266951009  0.39688131
q42  0.32248336  0.49946861  1.000000000  0.42372694  0.229140192  0.36169901
q48  0.31540781  0.60480937  0.423726942  1.00000000  0.231519612  0.36193479
q57  0.09372540  0.26695101  0.229140192  0.23151961  1.000000000  0.61260084
q58  0.22056688  0.39688131  0.361699012  0.36193479  0.612600844  1.00000000
q59  0.21595256  0.37567371  0.355733635  0.33387009  0.793528736  0.80919811
q60  0.09569002  0.26757299  0.216810052  0.20906687  0.918037552  0.69719273
q61  0.15988716  0.36416252  0.293791988  0.33010712  0.839431084  0.72404315
q62  0.11467330  0.27397928  0.225478233  0.24092732  0.771820926  0.61258767
q64  0.03896317  0.03598482  0.002373658  0.06276252 -0.052794657 -0.01225541
q80  0.03779096  0.14572381  0.131165239  0.10975589  0.138938593  0.12376998
q81 -0.07160469 -0.10302084 -0.074945932 -0.09446938 -0.116478487 -0.15959412
q82 -0.05452084 -0.10066865 -0.080750265 -0.09558210 -0.132702227 -0.16693502
q84  0.09199763  0.18037463  0.137417815  0.17709805  0.071714938  0.08646163
            q59         q60         q61         q62          q64         q80
q1   0.25997132  0.26964825  0.26039801  0.23196289 -0.003214449  0.03082990
q2   0.01441588  0.01865169 -0.01600499  0.02646210 -0.215079029 -0.11397970
q19 -0.24386555 -0.22008578 -0.27319596 -0.19864691 -0.121932168 -0.05004099
q20  0.19913098  0.14671880  0.19813929  0.14710241  0.119362145  0.06229103
q21  0.19019546  0.12705852  0.16357330  0.13924876  0.066139785  0.03897358
q22  0.38587573  0.35105833  0.36867080  0.31322774 -0.013149264  0.13262835
q33  0.21595256  0.09569002  0.15988716  0.11467330  0.038963174  0.03779096
```

```
q36  0.37567371  0.26757299  0.36416252  0.27397928  0.035984820  0.14572381
q42  0.35573364  0.21681005  0.29379199  0.22547823  0.002373658  0.13116524
q48  0.33387009  0.20906687  0.33010712  0.24092732  0.062762522  0.10975589
q57  0.79352874  0.91803755  0.83943108  0.77182093 -0.052794657  0.13893859
q58  0.80919811  0.69719273  0.72404315  0.61258767 -0.012255410  0.12376998
q59  1.00000000  0.82756661  0.81877831  0.71952759 -0.022858457  0.13816433
q60  0.82756661  1.00000000  0.88459307  0.79664031 -0.039446212  0.13764495
q61  0.81877831  0.88459307  1.00000000  0.74951966 -0.010068125  0.14343270
q62  0.71952759  0.79664031  0.74951966  1.00000000 -0.052240397  0.11569188
q64 -0.02285846 -0.03944621 -0.01006813 -0.05224040  1.000000000 -0.02859077
q80  0.13816433  0.13764495  0.14343270  0.11569188 -0.028590771  1.00000000
q81 -0.13724962 -0.11781406 -0.12940316 -0.11425347  0.002334760 -0.05484249
q82 -0.15403660 -0.12677757 -0.14128092 -0.10769066  0.022408033 -0.05766697
q84  0.09920760  0.08289895  0.11818952  0.07317028  0.216477896  0.11422085
            q81         q82         q84
q1  -0.10677352 -0.08339086  0.01547056
q2   0.02943371  0.07258526 -0.31546560
q19  0.06402879  0.06630414 -0.21375873
q20 -0.04451474 -0.04592658  0.23949728
q21 -0.02717716 -0.06062538  0.14836363
q22 -0.07882149 -0.08471064  0.12691559
q33 -0.07160469 -0.05452084  0.09199763
q36 -0.10302084 -0.10066865  0.18037463
q42 -0.07494593 -0.08075027  0.13741781
q48 -0.09446938 -0.09558210  0.17709805
q57 -0.11647849 -0.13270223  0.07171494
q58 -0.15959412 -0.16693502  0.08646163
q59 -0.13724962 -0.15403660  0.09920760
q60 -0.11781406 -0.12677757  0.08289895
q61 -0.12940316 -0.14128092  0.11818952
q62 -0.11425347 -0.10769066  0.07317028
q64  0.00233476  0.02240803  0.21647790
q80 -0.05484249 -0.05766697  0.11422085
q81  1.00000000  0.47601039 -0.02231173
q82  0.47601039  1.00000000 -0.04120638
q84 -0.02231173 -0.04120638  1.00000000
```

```
library(reshape2)
```

```
Attaching package: 'reshape2'
```

```
The following object is masked from 'package:tidyr':

    smiths
```

```r
melted_corr <- melt(corr_matrix)

# find strongly correlated pairs
strong_corrs <- subset(melted_corr, Var1 != Var2 & (value > 0.7 | value < -0.7))

# get rid of duplicate pairs
strong_corrs_unique <- strong_corrs[!duplicated(t(apply(strong_corrs[ , 1:2], 1, sort))), ]

# View the results
print(strong_corrs_unique)
```

```
    Var1 Var2     value
223  q59  q57 0.7935287
224  q60  q57 0.9180376
225  q61  q57 0.8394311
226  q62  q57 0.7718209
244  q59  q58 0.8091981
246  q61  q58 0.7240432
266  q60  q59 0.8275666
267  q61  q59 0.8187783
268  q62  q59 0.7195276
288  q61  q60 0.8845931
289  q62  q60 0.7966403
310  q62  q61 0.7495197
```

```r
nrow(strong_corrs_unique)
```

```
[1] 12
```

There are 12 pairs of correlated variables.

Project description

This project is focused on methodology, rather than attempting to answer a specific question about the data itself. Instead, the learning outcomes for this project are centered around gaining intuition for different unsupervised classification models and their appropriate uses. The core component of this project is building multiple types of models and comparing the strengths and appropriateness of their classification models. The models included will be primarily unsupervised (K-means clustering, hierarchical clustering), but will also include a supervised model (decision tree). For this project to successful, pre-processing of data and careful selection of predictor variables is essential. Based on the initial exploratory data analysis conducted here, a potential problem is the amount of missing data in the data set. Missing data analysis (e.g., MCAR, MAR, MNAR) and imputation are important components of model selection, and does impact both the types of models fit as well as the kinds of conclusions that can be drawn. However, evaluating and compensating for missing data like the type seen in this data set is beyond the scope of this project both due to time constraints. For this project, the data set will be significantly reduced using list wise deletion with the understanding that this is a dramatic oversimplification of the data and results would likely vary if other methods were used to deal ith missing data.