data exploration

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.8 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.1  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(readxl)  
library(officer)

##   
## Attaching package: 'officer'  
##   
## The following object is masked from 'package:readxl':  
##   
## read\_xlsx

library(flextable)

##   
## Attaching package: 'flextable'  
##   
## The following object is masked from 'package:purrr':  
##   
## compose

library(extrafont)

## Registering fonts with R

#times new roman tables  
my\_ft\_theme <- function(ft, ...) {  
 # Remove vertical cell padding  
 ft <- padding(ft, padding.top = 0, padding.bottom = 0, part = "all")  
   
 # Change font to TNR 11  
 ft <- font(ft, fontname = "Times New Roman", part = "all")  
 ft <- fontsize(ft, part = "all", size = 12)  
 ft  
}

#times new roman ggplots  
#library(remotes)  
#remotes::install\_version("Rttf2pt1", version = "1.3.8")  
#font\_import(paths = NULL, recursive = TRUE, prompt = TRUE,pattern = NULL)  
#loadfonts(device = "win")

#load(file="headscan\_full")  
#headscan\_full <- read\_csv("C:\\Users\\19177\\OneDrive - Colostate\\Desktop\\Dissertation\\headscan\_dissertation\\headscan\_full.csv")  
#is.data.frame(headscan\_full)  
headscan\_full <- read\_excel("C:\\Users\\19177\\OneDrive - Colostate\\Desktop\\Dissertation\\headscan\_dissertation\\headscan\_full.xlsx")  
str(headscan\_full)

## tibble [2,017 × 33] (S3: tbl\_df/tbl/data.frame)  
## $ ID : chr [1:2017] "400-20201012-002" "400-20201012-003" "400-20201012-004" "400-20201012-005" ...  
## $ AA\_C : num [1:2017] 6.5 5.5 7 5.8 6.7 6 5.9 5.9 6.5 6.5 ...  
## $ BGl\_C : num [1:2017] 31.5 28.9 29.3 31.3 28.8 30.6 32 NA 30 27.7 ...  
## $ BiW\_C : num [1:2017] 13 12.7 14.3 14 13.7 13 14.1 13.8 14.3 15 ...  
## $ BiW\_L : num [1:2017] 11.5 10.8 12.1 10.9 10.4 10.6 10.9 11.1 11.3 11.6 ...  
## $ ChCh\_C : num [1:2017] 6.2 6.4 6.8 7 7 7 6.7 6.9 6.7 6.3 ...  
## $ GoSub\_C : num [1:2017] 9.3 9.3 11.5 9.3 10.3 10 7.9 10.6 8.5 10.2 ...  
## $ NRB\_L : num [1:2017] 1.7 1.8 1.9 2.1 1.9 1.4 1.7 1.8 1.6 1.7 ...  
## $ ProA\_L : num [1:2017] 2.8 2.5 3.1 2.3 2.8 2.8 2.6 2.7 3.2 2.8 ...  
## $ ProA\_C : num [1:2017] 3.1 2.7 3.3 2.7 3.1 2.9 2.7 2.9 3.4 3.1 ...  
## $ ProS\_C : num [1:2017] 1.8 2 1.4 1.3 2.2 2.2 1.9 1.4 2.6 2.4 ...  
## $ ProS\_L : num [1:2017] 1.7 1.8 1.4 1.3 2 2 1.8 1.2 2.4 2.2 ...  
## $ SelP\_C : num [1:2017] 4.2 4.1 5.1 4.5 4.7 4.8 4.6 4.2 4.7 4.4 ...  
## $ SelP\_L : num [1:2017] 4.2 4.1 5.1 4.4 4.7 4.8 4.6 4.1 4.6 4.4 ...  
## $ SelDH\_C : num [1:2017] 1.5 0.9 0.9 1.1 1.3 1.5 0.9 0.9 1.2 1.4 ...  
## $ SelM\_L : num [1:2017] 12.2 9.9 13 11.5 11.9 12.6 11.7 11.2 11.7 11.7 ...  
## $ SnasM\_C : num [1:2017] 8.2 5.5 8.4 7.4 7.3 8 7.8 7.6 6.4 7.5 ...  
## $ SmanM\_C : num [1:2017] 5.9 5.1 4.5 4.3 3.3 3.4 5.5 3.7 6.1 4.1 ...  
## $ SmanM\_L : num [1:2017] 5.5 5 4.5 4.2 3.3 3.4 5 3.6 5.9 4 ...  
## $ SnasM\_L : num [1:2017] 7.5 5.3 7.8 6.9 6.7 7.6 6.9 7.1 6.2 6.9 ...  
## $ TrHO\_C : num [1:2017] 17.9 16.3 16.9 16.6 15.9 16.2 16.9 NA 16.7 16.6 ...  
## $ TrEJ\_C : num [1:2017] 4 3.2 3.9 2.9 4.6 4.2 2.9 3.2 2.9 3.3 ...  
## $ TrGo\_C : num [1:2017] 8.4 5.7 7 6.1 6.8 7 7.5 6.1 6.7 6.4 ...  
## $ TrSel\_C : num [1:2017] 14.9 13.8 15 13.3 14 15.1 14 13.8 15.6 14.3 ...  
## $ TrSman\_C : num [1:2017] 17.7 14.5 17.8 14.7 15.7 16.4 14.9 15.9 15.1 16 ...  
## $ TrSnas\_C : num [1:2017] 16.3 14.2 16.7 14.5 15.2 15.7 14.8 14.9 15.7 NA ...  
## $ TrTr\_C : num [1:2017] 29.6 27.6 29.2 27.3 27.9 30 28.3 27.5 30.7 28.6 ...  
## $ TrTr\_L : num [1:2017] 15.5 14.1 15.6 14.9 14.6 14.6 14.7 15.1 15.7 14.4 ...  
## $ coder : chr [1:2017] "Kayna" "Kayna" "Kayna" "Kayna" ...  
## $ age : num [1:2017] 31 49 49 34 49 55 26 18 25 27 ...  
## $ gender : chr [1:2017] "Male" "Female" "Male" "Male" ...  
## $ race\_eth : chr [1:2017] "Black" "white" "white" "white" ...  
## $ age\_group: chr [1:2017] "18-36" "37-54" "37-54" "18-36" ...

age\_sumstats <- headscan\_full %>%   
 summarise(n = n(),  
 min = min(age, na.rm = TRUE),  
 max = max(age, na.rm = TRUE),  
 mean = mean(age, na.rm = TRUE),  
 sd = sd(age, na.rm = TRUE),  
 se = sd/sqrt(n),  
 percent5th = quantile(age, 0.05, na.rm=TRUE),  
 percent25th = quantile(age, 0.25, na.rm=TRUE),  
 percent50th = quantile(age, 0.50, na.rm=TRUE),  
 percent75th = quantile(age, 0.75, na.rm=TRUE),  
 percent95th = quantile(age, 0.95, na.rm=TRUE),  
 na = sum(is.na(age)))  
  
age\_sumstats <- age\_sumstats %>%   
 mutate(across(where(is.numeric), round, 2))  
  
#HTML Table   
#age\_sumstats %>%   
 #kbl(caption = "Age SumStats") %>%   
 #kable\_styling(bootstrap\_options = c("striped", "hover", "condensed"), full\_width = TRUE)  
  
#Size 12 Table TNR  
flextable(age\_sumstats) %>%  
 my\_ft\_theme()%>%   
 bold(part = "header") %>%   
 set\_caption("Age SumStats")

**Table** : Age SumStats

| **n** | **min** | **max** | **mean** | **sd** | **se** | **percent5th** | **percent25th** | **percent50th** | **percent75th** | **percent95th** | **na** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2,017 | 18 | 72 | 36.39 | 11.51 | 0.26 | 19 | 26 | 37 | 46 | 54 | 1 |

#Autofit Width Table TNR  
flextable(age\_sumstats) %>%  
 my\_ft\_theme()%>%   
 bold(part = "header") %>%   
 set\_caption("Age SumStats") %>%   
 autofit() %>%   
 fit\_to\_width(7.5)

**Table** : Age SumStats

| **n** | **min** | **max** | **mean** | **sd** | **se** | **percent5th** | **percent25th** | **percent50th** | **percent75th** | **percent95th** | **na** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2,017 | 18 | 72 | 36.39 | 11.51 | 0.26 | 19 | 26 | 37 | 46 | 54 | 1 |

raceage\_sumstats <- headscan\_full %>%   
 group\_by(race\_eth) %>%   
 summarise(n = n(),  
 min = min(age, na.rm = TRUE),  
 max = max(age, na.rm = TRUE),  
 mean = mean(age, na.rm = TRUE),  
 sd = sd(age, na.rm = TRUE),  
 se = sd/sqrt(n),  
 percent5th = quantile(age, 0.05, na.rm=TRUE),  
 percent25th = quantile(age, 0.25, na.rm=TRUE),  
 percent50th = quantile(age, 0.50, na.rm=TRUE),  
 percent75th = quantile(age, 0.75, na.rm=TRUE),  
 percent95th = quantile(age, 0.95, na.rm=TRUE),  
 na = sum(is.na(age)))  
  
raceage\_sumstats <- raceage\_sumstats %>%   
 mutate(across(where(is.numeric), round, 2))  
   
#HTML Table   
#raceage\_sumstats %>%   
 #kbl(caption = "Age SumStats by Race/Ethnicity") %>%   
 #kable\_styling(bootstrap\_options = c("striped", "hover", "condensed"), full\_width = TRUE)  
  
#Size 12 Table TNR  
flextable(raceage\_sumstats) %>%  
 my\_ft\_theme()%>%   
 bold(part = "header") %>%   
 set\_caption("Age SumStats by Race/Ethnicity") %>%   
 set\_header\_labels(values = list(race\_eth = "Race/Ethnicity"))

**Table** : Age SumStats by Race/Ethnicity

| **Race/Ethnicity** | **n** | **min** | **max** | **mean** | **sd** | **se** | **percent5th** | **percent25th** | **percent50th** | **percent75th** | **percent95th** | **na** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AIAN | 8 | 27 | 56 | 43.25 | 11.78 | 4.17 | 27.70 | 33.50 | 45.5 | 53.00 | 56.00 | 0 |
| Asian | 91 | 18 | 56 | 33.23 | 11.76 | 1.23 | 18.00 | 21.50 | 31.0 | 42.50 | 54.00 | 0 |
| Black | 548 | 18 | 71 | 37.92 | 10.79 | 0.46 | 21.00 | 29.00 | 39.0 | 47.00 | 54.00 | 0 |
| LatinX | 100 | 18 | 55 | 34.63 | 11.93 | 1.19 | 19.00 | 23.00 | 34.0 | 44.50 | 53.10 | 1 |
| NHOPI | 4 | 19 | 40 | 27.00 | 9.76 | 4.88 | 19.15 | 19.75 | 24.5 | 31.75 | 38.35 | 0 |
| Other | 21 | 20 | 72 | 37.48 | 14.75 | 3.22 | 20.00 | 25.00 | 33.0 | 51.00 | 55.00 | 0 |
| PTNS | 5 | 29 | 40 | 36.60 | 4.72 | 2.11 | 30.20 | 35.00 | 39.0 | 40.00 | 40.00 | 0 |
| white | 1,240 | 18 | 62 | 36.05 | 11.64 | 0.33 | 19.00 | 25.00 | 36.0 | 46.00 | 54.00 | 0 |

#Autofit Width Table TNR  
flextable(raceage\_sumstats) %>%  
 my\_ft\_theme()%>%   
 bold(part = "header") %>%   
 set\_caption("Age SumStats by Race/Ethnicity") %>%   
 fit\_to\_width(7.5) %>%   
 set\_header\_labels(values = list(race\_eth = "Race/Ethnicity"))

**Table** : Age SumStats by Race/Ethnicity

| **Race/Ethnicity** | **n** | **min** | **max** | **mean** | **sd** | **se** | **percent5th** | **percent25th** | **percent50th** | **percent75th** | **percent95th** | **na** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AIAN | 8 | 27 | 56 | 43.25 | 11.78 | 4.17 | 27.70 | 33.50 | 45.5 | 53.00 | 56.00 | 0 |
| Asian | 91 | 18 | 56 | 33.23 | 11.76 | 1.23 | 18.00 | 21.50 | 31.0 | 42.50 | 54.00 | 0 |
| Black | 548 | 18 | 71 | 37.92 | 10.79 | 0.46 | 21.00 | 29.00 | 39.0 | 47.00 | 54.00 | 0 |
| LatinX | 100 | 18 | 55 | 34.63 | 11.93 | 1.19 | 19.00 | 23.00 | 34.0 | 44.50 | 53.10 | 1 |
| NHOPI | 4 | 19 | 40 | 27.00 | 9.76 | 4.88 | 19.15 | 19.75 | 24.5 | 31.75 | 38.35 | 0 |
| Other | 21 | 20 | 72 | 37.48 | 14.75 | 3.22 | 20.00 | 25.00 | 33.0 | 51.00 | 55.00 | 0 |
| PTNS | 5 | 29 | 40 | 36.60 | 4.72 | 2.11 | 30.20 | 35.00 | 39.0 | 40.00 | 40.00 | 0 |
| white | 1,240 | 18 | 62 | 36.05 | 11.64 | 0.33 | 19.00 | 25.00 | 36.0 | 46.00 | 54.00 | 0 |

genderage\_sumstats <- headscan\_full %>%   
 group\_by(gender) %>%   
 summarise(n = n(),  
 min = min(age, na.rm = TRUE),  
 max = max(age, na.rm = TRUE),  
 mean = mean(age, na.rm = TRUE),  
 sd = sd(age, na.rm = TRUE),  
 se = sd/sqrt(n),  
 percent5th = quantile(age, 0.05, na.rm=TRUE),  
 percent25th = quantile(age, 0.25, na.rm=TRUE),  
 percent50th = quantile(age, 0.50, na.rm=TRUE),  
 percent75th = quantile(age, 0.75, na.rm=TRUE),  
 percent95th = quantile(age, 0.95, na.rm=TRUE),  
 na = sum(is.na(age)))  
  
genderage\_sumstats <- genderage\_sumstats %>%   
 mutate(across(where(is.numeric), round, 2))  
  
#HTML Table  
#genderage\_sumstats %>%   
 #kbl(caption = "Age SumStats by Gender") %>%   
 #kable\_styling(bootstrap\_options = c("striped", "hover", "condensed"), full\_width = TRUE)  
  
#Size 12 Table TNR  
flextable(genderage\_sumstats) %>%  
 my\_ft\_theme()%>%   
 bold(part = "header") %>%   
 set\_caption("Age SumStats by Gender") %>%   
 set\_header\_labels(values = list(gender = "Gender"))

**Table** : Age SumStats by Gender

| **Gender** | **n** | **min** | **max** | **mean** | **sd** | **se** | **percent5th** | **percent25th** | **percent50th** | **percent75th** | **percent95th** | **na** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Female | 1,064 | 18 | 71 | 36.75 | 11.57 | 0.35 | 20.00 | 26.0 | 38.0 | 47.00 | 54.0 | 1 |
| Male | 939 | 18 | 72 | 35.97 | 11.48 | 0.37 | 19.00 | 26.0 | 35.0 | 46.00 | 54.1 | 0 |
| Non-binary or Other | 5 | 29 | 43 | 34.00 | 5.48 | 2.45 | 29.40 | 31.0 | 32.0 | 35.00 | 41.4 | 0 |
| Prefer not to say | 1 | 39 | 39 | 39.00 |  |  | 39.00 | 39.0 | 39.0 | 39.00 | 39.0 | 0 |
|  | 8 | 22 | 49 | 39.12 | 9.79 | 3.46 | 24.45 | 33.5 | 41.5 | 46.75 | 49.0 | 0 |

#Autofit Width Table TNR  
flextable(genderage\_sumstats) %>%  
 my\_ft\_theme()%>%   
 bold(part = "header") %>%   
 set\_caption("Age SumStats by Gender") %>%   
 fit\_to\_width(7.5) %>%   
 set\_header\_labels(values = list(gender = "Gender"))

**Table** : Age SumStats by Gender

| **Gender** | **n** | **min** | **max** | **mean** | **sd** | **se** | **percent5th** | **percent25th** | **percent50th** | **percent75th** | **percent95th** | **na** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Female | 1,064 | 18 | 71 | 36.75 | 11.57 | 0.35 | 20.00 | 26.0 | 38.0 | 47.00 | 54.0 | 1 |
| Male | 939 | 18 | 72 | 35.97 | 11.48 | 0.37 | 19.00 | 26.0 | 35.0 | 46.00 | 54.1 | 0 |
| Non-binary or Other | 5 | 29 | 43 | 34.00 | 5.48 | 2.45 | 29.40 | 31.0 | 32.0 | 35.00 | 41.4 | 0 |
| Prefer not to say | 1 | 39 | 39 | 39.00 |  |  | 39.00 | 39.0 | 39.0 | 39.00 | 39.0 | 0 |
|  | 8 | 22 | 49 | 39.12 | 9.79 | 3.46 | 24.45 | 33.5 | 41.5 | 46.75 | 49.0 | 0 |

agegroup\_sumstats <- headscan\_full %>%   
 group\_by(age\_group) %>%   
 summarise(n = n(),  
 min = min(age, na.rm = TRUE),  
 max = max(age, na.rm = TRUE),  
 mean = mean(age, na.rm = TRUE),  
 sd = sd(age, na.rm = TRUE),  
 se = sd/sqrt(n),  
 percent5th = quantile(age, 0.05, na.rm=TRUE),  
 percent25th = quantile(age, 0.25, na.rm=TRUE),  
 percent50th = quantile(age, 0.50, na.rm=TRUE),  
 percent75th = quantile(age, 0.75, na.rm=TRUE),  
 percent95th = quantile(age, 0.95, na.rm=TRUE),  
 na = sum(is.na(age)))

## Warning in min(age, na.rm = TRUE): no non-missing arguments to min; returning  
## Inf

## Warning in max(age, na.rm = TRUE): no non-missing arguments to max; returning  
## -Inf

agegroup\_sumstats <- agegroup\_sumstats %>%   
 mutate(across(where(is.numeric), round, 2))  
   
#HTML Table   
#agegroup\_sumstats %>%   
 #kbl(caption = "Age SumStats by Age Group") %>%   
 #kable\_styling(bootstrap\_options = c("striped", "hover", "condensed"), full\_width = TRUE)  
  
#Size 12 Table TNR  
flextable(agegroup\_sumstats) %>%  
 my\_ft\_theme()%>%   
 bold(part = "header") %>%   
 set\_caption("Age SumStats by Age Group")%>%   
 set\_header\_labels(values = list(age\_group = "Age Group"))

**Table** : Age SumStats by Age Group

| **Age Group** | **n** | **min** | **max** | **mean** | **sd** | **se** | **percent5th** | **percent25th** | **percent50th** | **percent75th** | **percent95th** | **na** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 18-36 | 992 | 18 | 36 | 26.19 | 5.60 | 0.18 | 18.00 | 21 | 26 | 31 | 35.0 | 0 |
| 37-54 | 940 | 37 | 54 | 45.41 | 4.99 | 0.16 | 37.95 | 41 | 45 | 50 | 53.0 | 0 |
| 55-72 | 84 | 55 | 72 | 55.96 | 2.94 | 0.32 | 55.00 | 55 | 55 | 55 | 61.4 | 0 |
|  | 1 | Inf | -Inf |  |  |  |  |  |  |  |  | 1 |

#Autofit Width Table TNR  
flextable(agegroup\_sumstats) %>%  
 my\_ft\_theme()%>%   
 bold(part = "header") %>%   
 set\_caption("Age SumStats by Age Group") %>%   
 fit\_to\_width(7.5) %>%  
 set\_header\_labels(values = list(age\_group = "Age Group"))

**Table** : Age SumStats by Age Group

| **Age Group** | **n** | **min** | **max** | **mean** | **sd** | **se** | **percent5th** | **percent25th** | **percent50th** | **percent75th** | **percent95th** | **na** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 18-36 | 992 | 18 | 36 | 26.19 | 5.60 | 0.18 | 18.00 | 21 | 26 | 31 | 35.0 | 0 |
| 37-54 | 940 | 37 | 54 | 45.41 | 4.99 | 0.16 | 37.95 | 41 | 45 | 50 | 53.0 | 0 |
| 55-72 | 84 | 55 | 72 | 55.96 | 2.94 | 0.32 | 55.00 | 55 | 55 | 55 | 61.4 | 0 |
|  | 1 | Inf | -Inf |  |  |  |  |  |  |  |  | 1 |

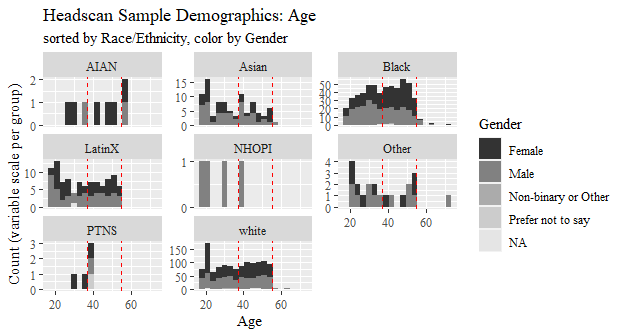
#explore race\_eth, frequency of each

#explore gender, frequency of each

PLOTS

ggplot(data=headscan\_full, aes(x= age, fill=gender))+  
 theme(text=element\_text(family= "Times New Roman"))+  
 scale\_fill\_grey(na.value="gray90")+  
 geom\_histogram(binwidth = 3)+  
 labs(title="Headscan Sample Demographics: Age",  
 subtitle="sorted by Race/Ethnicity, color by Gender",  
 y="Count (variable scale per group)",  
 x="Age",  
 fill="Gender")+  
 facet\_wrap(~race\_eth, scales="free\_y")+  
 scale\_y\_continuous(breaks = function(x) unique(floor(pretty(seq(0, (max(x) + 1) \* 1.1)))))+  
 geom\_vline(xintercept = 37, color="red", linetype= "dashed")+  
 geom\_vline(xintercept = 55, color="red", linetype= "dashed")

## Warning: Removed 1 rows containing non-finite values (stat\_bin).



#ggplot(data=headscan\_full)+  
 #geom\_histogram(binwidth=2, aes(x=age, fill=race\_eth))+  
 #scale\_fill\_grey()+  
 #labs(title= "Demographics of Headscan Sample",  
 #subtitle = "Data collected by Human Solutions",  
 #fill= "Ethnicity",  
 #y="Count",  
 #x="Age")

#ggplot(data=headscan\_full)+  
 #geom\_boxplot(aes(x=race\_eth, y=age, fill=gender))+  
 #scale\_fill\_grey()+  
 #theme(axis.text.x = element\_text(angle = 45))