removing outliers

2022-07-26

library(readxl)  
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(writexl)  
library(extrafont)

## Registering fonts with R

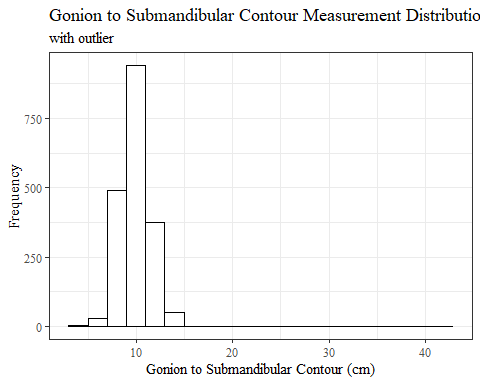
library(forcats)

headscan\_full<-read\_excel("C:\\Users\\19177\\OneDrive - Colostate\\Desktop\\Dissertation\\headscan\_dissertation\\headscan\_full.xlsx")  
str(headscan\_full)

## tibble [2,016 × 33] (S3: tbl\_df/tbl/data.frame)  
## $ ID : chr [1:2016] "400-20201012-002" "400-20201012-003" "400-20201012-004" "400-20201012-005" ...  
## $ AA\_C : num [1:2016] 6.5 5.5 7 5.8 6.7 6 5.9 5.9 6.5 6.5 ...  
## $ BGl\_C : num [1:2016] 31.5 28.9 29.3 31.3 28.8 30.6 32 NA 30 27.7 ...  
## $ BiW\_C : num [1:2016] 13 12.7 14.3 14 13.7 13 14.1 13.8 14.3 15 ...  
## $ BiW\_L : num [1:2016] 11.5 10.8 12.1 10.9 10.4 10.6 10.9 11.1 11.3 11.6 ...  
## $ ChCh\_C : num [1:2016] 6.2 6.4 6.8 7 7 7 6.7 6.9 6.7 6.3 ...  
## $ GoSub\_C : num [1:2016] 9.3 9.3 11.5 9.3 10.3 10 7.9 10.6 8.5 10.2 ...  
## $ NRB\_L : num [1:2016] 1.7 1.8 1.9 2.1 1.9 1.4 1.7 1.8 1.6 1.7 ...  
## $ ProA\_L : num [1:2016] 2.8 2.5 3.1 2.3 2.8 2.8 2.6 2.7 3.2 2.8 ...  
## $ ProA\_C : num [1:2016] 3.1 2.7 3.3 2.7 3.1 2.9 2.7 2.9 3.4 3.1 ...  
## $ ProS\_C : num [1:2016] 1.8 2 1.4 1.3 2.2 2.2 1.9 1.4 2.6 2.4 ...  
## $ ProS\_L : num [1:2016] 1.7 1.8 1.4 1.3 2 2 1.8 1.2 2.4 2.2 ...  
## $ SelP\_C : num [1:2016] 4.2 4.1 5.1 4.5 4.7 4.8 4.6 4.2 4.7 4.4 ...  
## $ SelP\_L : num [1:2016] 4.2 4.1 5.1 4.4 4.7 4.8 4.6 4.1 4.6 4.4 ...  
## $ SelDH\_C : num [1:2016] 1.5 0.9 0.9 1.1 1.3 1.5 0.9 0.9 1.2 1.4 ...  
## $ SelM\_L : num [1:2016] 12.2 9.9 13 11.5 11.9 12.6 11.7 11.2 11.7 11.7 ...  
## $ SnasM\_C : num [1:2016] 8.2 5.5 8.4 7.4 7.3 8 7.8 7.6 6.4 7.5 ...  
## $ SmanM\_C : num [1:2016] 5.9 5.1 4.5 4.3 3.3 3.4 5.5 3.7 6.1 4.1 ...  
## $ SmanM\_L : num [1:2016] 5.5 5 4.5 4.2 3.3 3.4 5 3.6 5.9 4 ...  
## $ SnasM\_L : num [1:2016] 7.5 5.3 7.8 6.9 6.7 7.6 6.9 7.1 6.2 6.9 ...  
## $ TrHO\_C : num [1:2016] 17.9 16.3 16.9 16.6 15.9 16.2 16.9 NA 16.7 16.6 ...  
## $ TrEJ\_C : num [1:2016] 4 3.2 3.9 2.9 4.6 4.2 2.9 3.2 2.9 3.3 ...  
## $ TrGo\_C : num [1:2016] 8.4 5.7 7 6.1 6.8 7 7.5 6.1 6.7 6.4 ...  
## $ TrSel\_C : num [1:2016] 14.9 13.8 15 13.3 14 15.1 14 13.8 15.6 14.3 ...  
## $ TrSman\_C : num [1:2016] 17.7 14.5 17.8 14.7 15.7 16.4 14.9 15.9 15.1 16 ...  
## $ TrSnas\_C : num [1:2016] 16.3 14.2 16.7 14.5 15.2 15.7 14.8 14.9 15.7 NA ...  
## $ TrTr\_C : num [1:2016] 29.6 27.6 29.2 27.3 27.9 30 28.3 27.5 30.7 28.6 ...  
## $ TrTr\_L : num [1:2016] 15.5 14.1 15.6 14.9 14.6 14.6 14.7 15.1 15.7 14.4 ...  
## $ coder : chr [1:2016] "Kayna" "Kayna" "Kayna" "Kayna" ...  
## $ age : num [1:2016] 31 49 49 34 49 55 26 18 25 27 ...  
## $ gender : chr [1:2016] "Male" "Female" "Male" "Male" ...  
## $ race\_eth : chr [1:2016] "Black" "white" "white" "white" ...  
## $ age\_group: chr [1:2016] "18-36" "37-54" "37-54" "18-36" ...

#histogram of all GoSub\_C values w/ outlier  
ggplot(data=headscan\_full, aes(x=GoSub\_C))+  
 geom\_bar(stat="bin", bins=20, color= "black", fill = "white")+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Gonion to Submandibular Contour Measurement Distribution",  
 subtitle= "with outlier",  
 y="Frequency",  
 x="Gonion to Submandibular Contour (cm)")

## Warning: Removed 127 rows containing non-finite values (stat\_bin).

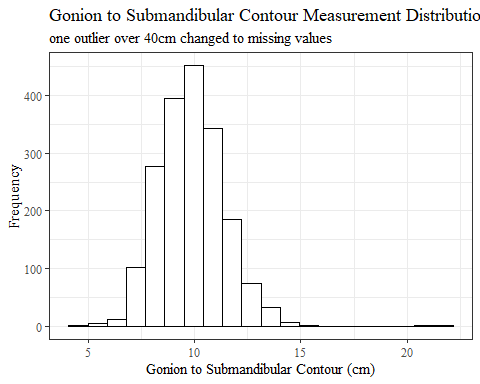


#REMOVING OUTLIER GoSubC  
  
max(headscan\_full$GoSub\_C, na.rm = TRUE)

## [1] 42.4

headscan\_full1 <- headscan\_full %>% mutate(GoSub\_C = replace(GoSub\_C, GoSub\_C>40, NA))  
  
#histogram of all GoSub\_C values  
ggplot(data=headscan\_full1, aes(x=GoSub\_C))+  
 geom\_bar(stat="bin", bins=20, color= "black", fill = "white")+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Gonion to Submandibular Contour Measurement Distribution",  
 subtitle = "one outlier over 40cm changed to missing values",  
 y="Frequency",  
 x="Gonion to Submandibular Contour (cm)")

## Warning: Removed 128 rows containing non-finite values (stat\_bin).



sum(is.na(headscan\_full$GoSub\_C))

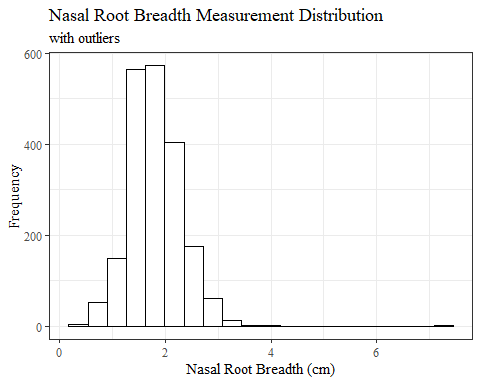
## [1] 127

sum(is.na(headscan\_full1$GoSub\_C))

## [1] 128

#histogram of all NRB\_L values w/ outlier  
ggplot(data=headscan\_full, aes(x=NRB\_L))+  
 geom\_bar(stat="bin", bins=20, color= "black", fill = "white")+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Nasal Root Breadth Measurement Distribution",  
 subtitle= "with outliers",  
 y="Frequency",  
 x="Nasal Root Breadth (cm)")

## Warning: Removed 15 rows containing non-finite values (stat\_bin).

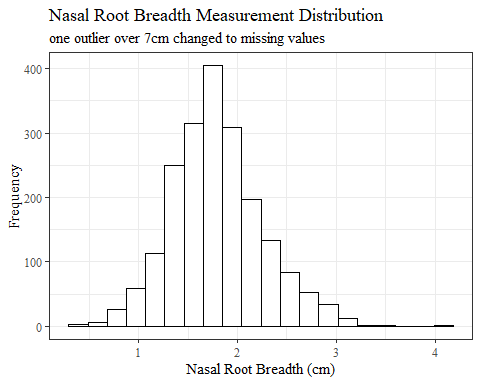


#REMOVING OUTLIER NRB\_L  
  
max(headscan\_full$NRB\_L, na.rm = TRUE)

## [1] 7.2

headscan\_full1 <- headscan\_full %>% mutate(NRB\_L = replace(NRB\_L, NRB\_L>7, NA))  
  
#histogram of all NRB\_L values  
ggplot(data=headscan\_full1, aes(x=NRB\_L))+  
 geom\_bar(stat="bin", bins=20, color= "black", fill = "white")+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Nasal Root Breadth Measurement Distribution",  
 subtitle = "one outlier over 7cm changed to missing values",  
 y="Frequency",  
 x="Nasal Root Breadth (cm)")

## Warning: Removed 16 rows containing non-finite values (stat\_bin).



sum(is.na(headscan\_full$NRB\_L))

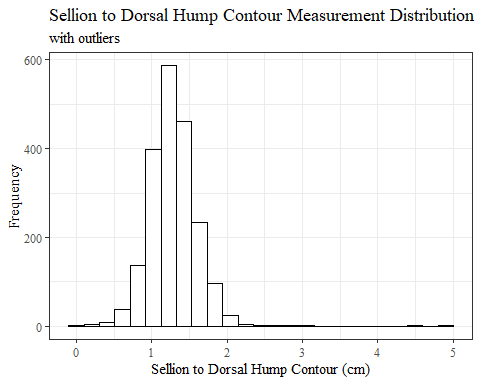
## [1] 15

sum(is.na(headscan\_full1$NRB\_L))

## [1] 16

#histogram of all SelDH\_C values w/ outlier  
ggplot(data=headscan\_full, aes(x=SelDH\_C ))+  
 geom\_bar(stat="bin", bins=25, color= "black", fill = "white")+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Sellion to Dorsal Hump Contour Measurement Distribution",  
 subtitle = "with outliers",  
 y="Frequency",  
 x="Sellion to Dorsal Hump Contour (cm)")

## Warning: Removed 14 rows containing non-finite values (stat\_bin).

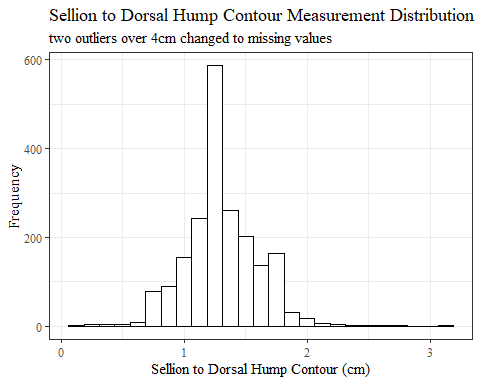


#REMOVING OUTLIER SelDH\_C  
  
max(headscan\_full$SelDH\_C, na.rm = TRUE)

## [1] 5

headscan\_full1 <- headscan\_full %>% mutate(SelDH\_C = replace(SelDH\_C, SelDH\_C>4, NA))  
  
#histogram of all SelDH\_C values  
ggplot(data=headscan\_full1, aes(x=SelDH\_C))+  
 geom\_bar(stat="bin", bins=25, color= "black", fill = "white")+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Sellion to Dorsal Hump Contour Measurement Distribution",  
 subtitle = "two outliers over 4cm changed to missing values",  
 y="Frequency",  
 x="Sellion to Dorsal Hump Contour (cm)")

## Warning: Removed 16 rows containing non-finite values (stat\_bin).



sum(is.na(headscan\_full$SelDH\_C))

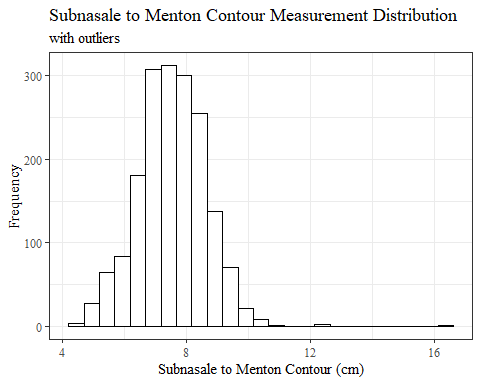
## [1] 14

sum(is.na(headscan\_full1$SelDH\_C))

## [1] 16

#histogram of all SnasM\_C values w/ outlier  
ggplot(data=headscan\_full, aes(x=SnasM\_C))+  
 geom\_bar(stat="bin", bins=25, color= "black", fill = "white")+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Subnasale to Menton Contour Measurement Distribution",  
 subtitle = "with outliers",  
 y="Frequency",  
 x="Subnasale to Menton Contour (cm)")

## Warning: Removed 236 rows containing non-finite values (stat\_bin).

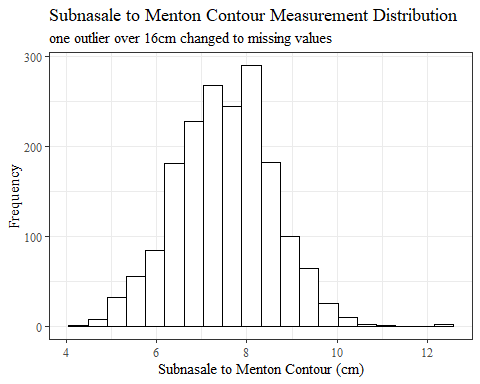


#REMOVING OUTLIER SnasM\_C  
  
max(headscan\_full$SnasM\_C, na.rm = TRUE)

## [1] 16.3

headscan\_full1 <- headscan\_full %>% mutate(SnasM\_C = replace(SnasM\_C, SnasM\_C>16, NA))  
  
#histogram of all SnasM\_C values  
ggplot(data=headscan\_full1, aes(x=SnasM\_C))+  
 geom\_bar(stat="bin", bins=20, color= "black", fill = "white")+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Subnasale to Menton Contour Measurement Distribution",  
 subtitle = "one outlier over 16cm changed to missing values",  
 y="Frequency",  
 x="Subnasale to Menton Contour (cm)")

## Warning: Removed 237 rows containing non-finite values (stat\_bin).



sum(is.na(headscan\_full$SnasM\_C))

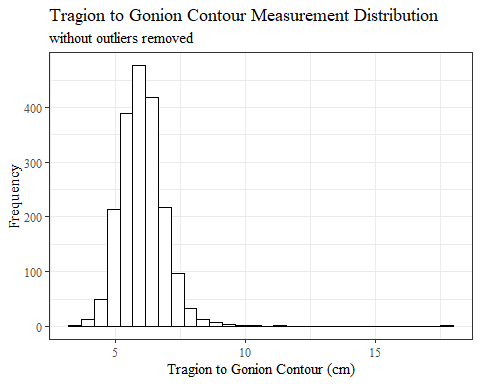
## [1] 236

sum(is.na(headscan\_full1$SnasM\_C))

## [1] 237

#histogram of all TrGo\_C values  
ggplot(data=headscan\_full, aes(x=TrGo\_C))+  
 geom\_bar(stat="bin", bins=30, color= "black", fill = "white")+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Tragion to Gonion Contour Measurement Distribution",  
 subtitle = "without outliers removed",  
 y="Frequency",  
 x="Tragion to Gonion Contour (cm)")

## Warning: Removed 80 rows containing non-finite values (stat\_bin).

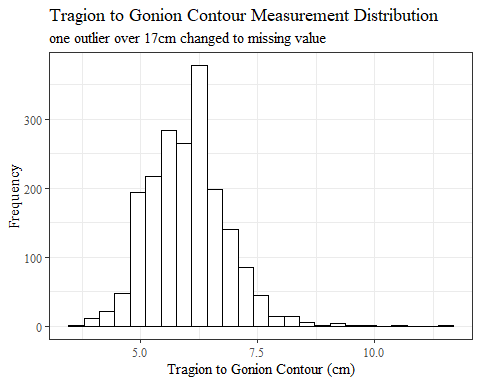


#REMOVING OUTLIER TrGo\_C  
  
max(headscan\_full$TrGo\_C, na.rm = TRUE)

## [1] 17.8

headscan\_full1 <- headscan\_full %>% mutate(TrGo\_C = replace(TrGo\_C, TrGo\_C>17, NA))  
  
#histogram of all TrGo\_C values  
ggplot(data=headscan\_full1, aes(x=TrGo\_C))+  
 geom\_bar(stat="bin", bins=25, color= "black", fill = "white")+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Tragion to Gonion Contour Measurement Distribution",  
 subtitle = "one outlier over 17cm changed to missing value",  
 y="Frequency",  
 x="Tragion to Gonion Contour (cm)")

## Warning: Removed 81 rows containing non-finite values (stat\_bin).



sum(is.na(headscan\_full$TrGo\_C))

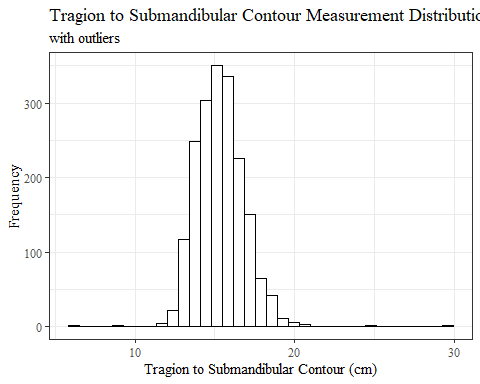
## [1] 80

sum(is.na(headscan\_full1$TrGo\_C))

## [1] 81

#histogram of all TrSman\_C values  
ggplot(data=headscan\_full, aes(x=TrSman\_C))+  
 geom\_bar(stat="bin", bins=35, color= "black", fill = "white")+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Tragion to Submandibular Contour Measurement Distribution",  
 subtitle = "with outliers",  
 y="Frequency",  
 x="Tragion to Submandibular Contour (cm)")

## Warning: Removed 132 rows containing non-finite values (stat\_bin).

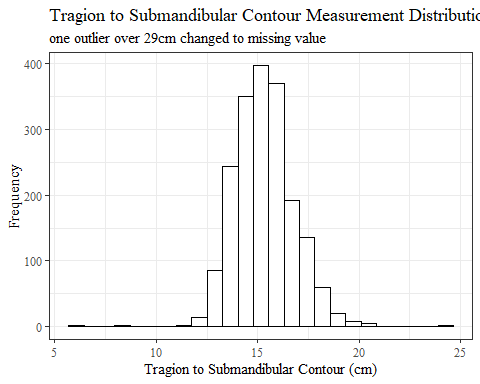


#REMOVING OUTLIER TrSman\_C  
  
max(headscan\_full$TrSman\_C, na.rm = TRUE)

## [1] 29.8

headscan\_full1 <- headscan\_full %>% mutate(TrSman\_C = replace(TrSman\_C, TrSman\_C>29, NA))  
  
#histogram of all TrSman\_C values  
ggplot(data=headscan\_full1, aes(x=TrSman\_C))+  
 geom\_bar(stat="bin", bins=25, color= "black", fill = "white")+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Tragion to Submandibular Contour Measurement Distribution",  
 subtitle = "one outlier over 29cm changed to missing value",  
 y="Frequency",  
 x="Tragion to Submandibular Contour (cm)")

## Warning: Removed 133 rows containing non-finite values (stat\_bin).



sum(is.na(headscan\_full$TrSman\_C))

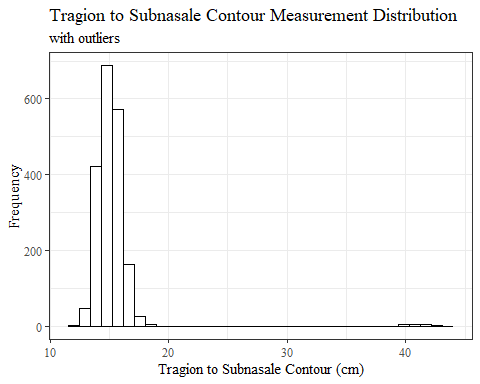
## [1] 132

sum(is.na(headscan\_full1$TrSman\_C))

## [1] 133

#histogram of all TrSnas\_C values  
ggplot(data=headscan\_full, aes(x=TrSnas\_C))+  
 geom\_bar(stat="bin", bins=35, color= "black", fill = "white")+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Tragion to Subnasale Contour Measurement Distribution",  
 subtitle = "with outliers",  
 y="Frequency",  
 x="Tragion to Subnasale Contour (cm)")

## Warning: Removed 71 rows containing non-finite values (stat\_bin).

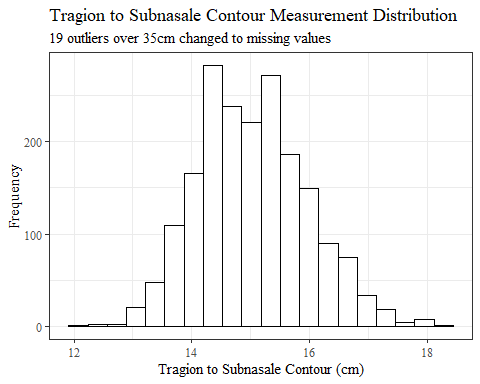


#REMOVING OUTLIER TrSnas\_C  
  
max(headscan\_full$TrSnas\_C, na.rm = TRUE)

## [1] 43.7

headscan\_full1 <- headscan\_full %>% mutate(TrSnas\_C = replace(TrSnas\_C, TrSnas\_C>35, NA))  
  
#histogram of all TrSnas\_C values  
ggplot(data=headscan\_full1, aes(x=TrSnas\_C))+  
 geom\_bar(stat="bin", bins=20, color= "black", fill = "white")+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Tragion to Subnasale Contour Measurement Distribution",  
 subtitle = "19 outliers over 35cm changed to missing values",  
 y="Frequency",  
 x="Tragion to Subnasale Contour (cm)")

## Warning: Removed 90 rows containing non-finite values (stat\_bin).



sum(is.na(headscan\_full$TrSnas\_C))

## [1] 71

sum(is.na(headscan\_full1$TrSnas\_C))

## [1] 90

str(headscan\_full1)

## tibble [2,016 × 33] (S3: tbl\_df/tbl/data.frame)  
## $ ID : chr [1:2016] "400-20201012-002" "400-20201012-003" "400-20201012-004" "400-20201012-005" ...  
## $ AA\_C : num [1:2016] 6.5 5.5 7 5.8 6.7 6 5.9 5.9 6.5 6.5 ...  
## $ BGl\_C : num [1:2016] 31.5 28.9 29.3 31.3 28.8 30.6 32 NA 30 27.7 ...  
## $ BiW\_C : num [1:2016] 13 12.7 14.3 14 13.7 13 14.1 13.8 14.3 15 ...  
## $ BiW\_L : num [1:2016] 11.5 10.8 12.1 10.9 10.4 10.6 10.9 11.1 11.3 11.6 ...  
## $ ChCh\_C : num [1:2016] 6.2 6.4 6.8 7 7 7 6.7 6.9 6.7 6.3 ...  
## $ GoSub\_C : num [1:2016] 9.3 9.3 11.5 9.3 10.3 10 7.9 10.6 8.5 10.2 ...  
## $ NRB\_L : num [1:2016] 1.7 1.8 1.9 2.1 1.9 1.4 1.7 1.8 1.6 1.7 ...  
## $ ProA\_L : num [1:2016] 2.8 2.5 3.1 2.3 2.8 2.8 2.6 2.7 3.2 2.8 ...  
## $ ProA\_C : num [1:2016] 3.1 2.7 3.3 2.7 3.1 2.9 2.7 2.9 3.4 3.1 ...  
## $ ProS\_C : num [1:2016] 1.8 2 1.4 1.3 2.2 2.2 1.9 1.4 2.6 2.4 ...  
## $ ProS\_L : num [1:2016] 1.7 1.8 1.4 1.3 2 2 1.8 1.2 2.4 2.2 ...  
## $ SelP\_C : num [1:2016] 4.2 4.1 5.1 4.5 4.7 4.8 4.6 4.2 4.7 4.4 ...  
## $ SelP\_L : num [1:2016] 4.2 4.1 5.1 4.4 4.7 4.8 4.6 4.1 4.6 4.4 ...  
## $ SelDH\_C : num [1:2016] 1.5 0.9 0.9 1.1 1.3 1.5 0.9 0.9 1.2 1.4 ...  
## $ SelM\_L : num [1:2016] 12.2 9.9 13 11.5 11.9 12.6 11.7 11.2 11.7 11.7 ...  
## $ SnasM\_C : num [1:2016] 8.2 5.5 8.4 7.4 7.3 8 7.8 7.6 6.4 7.5 ...  
## $ SmanM\_C : num [1:2016] 5.9 5.1 4.5 4.3 3.3 3.4 5.5 3.7 6.1 4.1 ...  
## $ SmanM\_L : num [1:2016] 5.5 5 4.5 4.2 3.3 3.4 5 3.6 5.9 4 ...  
## $ SnasM\_L : num [1:2016] 7.5 5.3 7.8 6.9 6.7 7.6 6.9 7.1 6.2 6.9 ...  
## $ TrHO\_C : num [1:2016] 17.9 16.3 16.9 16.6 15.9 16.2 16.9 NA 16.7 16.6 ...  
## $ TrEJ\_C : num [1:2016] 4 3.2 3.9 2.9 4.6 4.2 2.9 3.2 2.9 3.3 ...  
## $ TrGo\_C : num [1:2016] 8.4 5.7 7 6.1 6.8 7 7.5 6.1 6.7 6.4 ...  
## $ TrSel\_C : num [1:2016] 14.9 13.8 15 13.3 14 15.1 14 13.8 15.6 14.3 ...  
## $ TrSman\_C : num [1:2016] 17.7 14.5 17.8 14.7 15.7 16.4 14.9 15.9 15.1 16 ...  
## $ TrSnas\_C : num [1:2016] 16.3 14.2 16.7 14.5 15.2 15.7 14.8 14.9 15.7 NA ...  
## $ TrTr\_C : num [1:2016] 29.6 27.6 29.2 27.3 27.9 30 28.3 27.5 30.7 28.6 ...  
## $ TrTr\_L : num [1:2016] 15.5 14.1 15.6 14.9 14.6 14.6 14.7 15.1 15.7 14.4 ...  
## $ coder : chr [1:2016] "Kayna" "Kayna" "Kayna" "Kayna" ...  
## $ age : num [1:2016] 31 49 49 34 49 55 26 18 25 27 ...  
## $ gender : chr [1:2016] "Male" "Female" "Male" "Male" ...  
## $ race\_eth : chr [1:2016] "Black" "white" "white" "white" ...  
## $ age\_group: chr [1:2016] "18-36" "37-54" "37-54" "18-36" ...

write\_xlsx(headscan\_full1, "C:\\Users\\19177\\OneDrive - Colostate\\Desktop\\Dissertation\\headscan\_dissertation\\headscan\_full1.xlsx")