exploring-demo-postimpute

2022-09-13

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(flextable)

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

library(writexl)  
library(extrafont)

## Registering fonts with R

library(forcats)

#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  
}

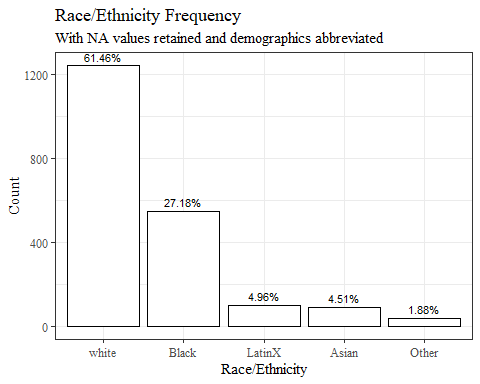
chosen\_withna<-read\_excel("C:\\Users\\19177\\OneDrive - Colostate\\Desktop\\Dissertation\\headscan\_dissertation\\chosen\_withna.xlsx")  
  
chosen\_withna$gender<-as.factor(chosen\_withna$gender)  
chosen\_withna$race\_eth<-as.factor(chosen\_withna$race\_eth)  
chosen\_withna$age\_group<-as.factor(chosen\_withna$age\_group)  
  
  
str(chosen\_withna)

## tibble [2,016 × 16] (S3: tbl\_df/tbl/data.frame)  
## $ ID : chr [1:2016] "400-20210129-009" "400-20201123-009" "400-20210203-001" "400-20210216-006" ...  
## $ AA\_C : num [1:2016] NA NA NA NA NA NA NA NA NA NA ...  
## $ BiW\_C : num [1:2016] NA NA NA NA NA NA NA NA NA NA ...  
## $ BiW\_L : num [1:2016] NA NA NA NA NA NA NA NA NA NA ...  
## $ GoSub\_C : num [1:2016] NA NA NA NA NA NA NA NA NA NA ...  
## $ NRB\_L : num [1:2016] NA NA NA NA NA NA NA NA NA NA ...  
## $ ProS\_L : num [1:2016] NA NA NA NA NA NA NA NA NA NA ...  
## $ SelP\_L : num [1:2016] NA NA NA NA NA NA NA NA NA NA ...  
## $ SelM\_L : num [1:2016] NA NA NA NA NA NA NA NA NA NA ...  
## $ SnasM\_C : num [1:2016] NA NA NA NA NA NA NA NA NA NA ...  
## $ TrSman\_C : num [1:2016] NA NA NA NA NA NA NA NA NA NA ...  
## $ TrTr\_C : num [1:2016] NA NA NA NA NA NA NA NA NA NA ...  
## $ TrTr\_L : num [1:2016] NA NA NA NA NA NA NA NA NA NA ...  
## $ gender : Factor w/ 4 levels "Female","Male",..: 2 2 2 2 1 1 2 1 2 1 ...  
## $ race\_eth : Factor w/ 8 levels "AIAN","Asian",..: 8 8 8 8 3 2 8 8 3 8 ...  
## $ age\_group: Factor w/ 3 levels "18-36","37-54",..: 2 1 1 2 2 2 3 3 2 2 ...

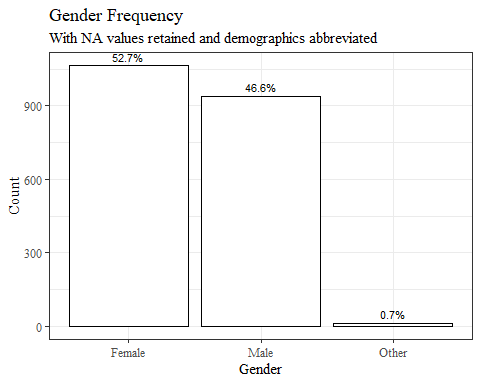
chosen\_withna1 <- chosen\_withna  
  
#race/eth  
chosen\_withna1$race\_eth <-   
 recode\_factor(chosen\_withna1$race\_eth, 'AIAN'= "Other",  
 'NHOPI' = "Other",  
 'PTNS' = "Other")  
  
#gender  
chosen\_withna1$gender <-   
 recode\_factor(chosen\_withna1$gender, 'Non-binary or Other'= "Other",  
 'Prefer not to say' = "Other")  
  
chosen\_withna1$gender[is.na(chosen\_withna1$gender)]="Other"  
  
summary(chosen\_withna1)

## ID AA\_C BiW\_C BiW\_L   
## Length:2016 Min. :44.00 Min. :101.0 Min. : 82.0   
## Class :character 1st Qu.:57.00 1st Qu.:124.0 1st Qu.:104.0   
## Mode :character Median :61.00 Median :133.0 Median :111.0   
## Mean :61.23 Mean :133.4 Mean :111.2   
## 3rd Qu.:65.00 3rd Qu.:141.0 3rd Qu.:118.0   
## Max. :87.00 Max. :187.0 Max. :152.0   
## NA's :40 NA's :37 NA's :35   
## GoSub\_C NRB\_L ProS\_L SelP\_L   
## Min. : 49.00 Min. : 3.00 Min. :12.00 Min. :30.00   
## 1st Qu.: 88.00 1st Qu.:15.00 1st Qu.:17.00 1st Qu.:42.00   
## Median : 98.00 Median :18.00 Median :19.00 Median :44.00   
## Mean : 98.75 Mean :17.95 Mean :19.12 Mean :44.53   
## 3rd Qu.:108.00 3rd Qu.:21.00 3rd Qu.:21.00 3rd Qu.:47.00   
## Max. :152.00 Max. :35.00 Max. :27.00 Max. :58.00   
## NA's :145 NA's :34 NA's :40 NA's :35   
## SelM\_L SnasM\_C TrSman\_C TrTr\_C   
## Min. : 84.0 Min. : 44.00 Min. :117.0 Min. :241.0   
## 1st Qu.:110.0 1st Qu.: 68.00 1st Qu.:143.0 1st Qu.:272.0   
## Median :116.0 Median : 75.00 Median :152.0 Median :282.0   
## Mean :116.3 Mean : 74.96 Mean :153.3 Mean :282.7   
## 3rd Qu.:123.0 3rd Qu.: 82.00 3rd Qu.:162.0 3rd Qu.:293.0   
## Max. :145.0 Max. :105.00 Max. :208.0 Max. :332.0   
## NA's :243 NA's :254 NA's :151 NA's :56   
## TrTr\_L gender race\_eth age\_group   
## Min. :127.0 Other : 14 Other : 38 18-36:991   
## 1st Qu.:141.0 Female:1063 Asian : 91 37-54:940   
## Median :146.0 Male : 939 Black : 548 55-72: 84   
## Mean :146.5 LatinX: 100 NA's : 1   
## 3rd Qu.:152.0 white :1239   
## Max. :173.0   
## NA's :52

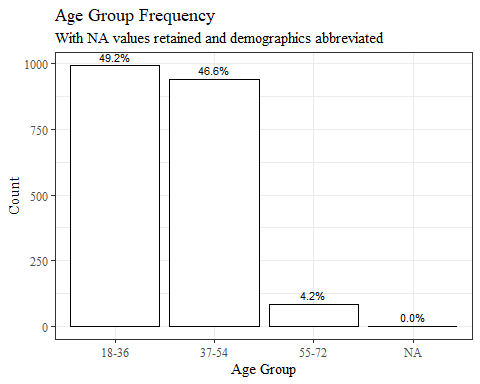
chosen\_withna1 %>%   
 ggplot(aes(x=fct\_infreq(race\_eth), label=scales::percent(prop.table(stat(count)))))+  
 geom\_bar(stat="count", color= "black", fill = "white")+  
 geom\_text(stat="count",  
 position= position\_dodge(0.9),  
 vjust = -0.5,   
 size = 3)+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Race/Ethnicity Frequency",  
 subtitle="With NA values retained and demographics abbreviated",  
 y="Count",  
 x="Race/Ethnicity")



chosen\_withna1 %>%   
 ggplot(aes(x=fct\_infreq(gender), label=scales::percent(prop.table(stat(count)))))+  
 geom\_bar(stat="count", color= "black", fill = "white")+  
 geom\_text(stat="count",  
 position= position\_dodge(0.9),  
 vjust = -0.5,   
 size = 3)+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Gender Frequency",  
 subtitle="With NA values retained and demographics abbreviated",  
 y="Count",  
 x="Gender")



chosen\_withna1 %>%   
 ggplot(aes(x=fct\_infreq(age\_group), label=scales::percent(prop.table(stat(count)))))+  
 geom\_bar(stat="count", color= "black", fill = "white")+  
 geom\_text(stat="count",  
 position= position\_dodge(0.9),  
 vjust = -0.5,   
 size = 3)+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Age Group Frequency",  
 subtitle="With NA values retained and demographics abbreviated",  
 y="Count",  
 x="Age Group")



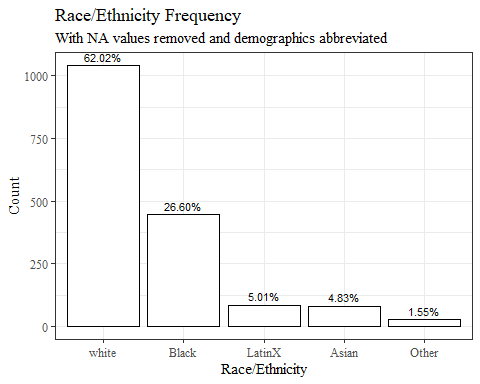
chosen\_nona<-read\_excel("C:\\Users\\19177\\OneDrive - Colostate\\Desktop\\Dissertation\\headscan\_dissertation\\chosen\_nona.xlsx")  
  
chosen\_nona$gender<-as.factor(chosen\_nona$gender)  
chosen\_nona$race\_eth<-as.factor(chosen\_nona$race\_eth)  
chosen\_nona$age\_group<-as.factor(chosen\_nona$age\_group)  
  
  
str(chosen\_nona)

## tibble [1,677 × 16] (S3: tbl\_df/tbl/data.frame)  
## $ ID : chr [1:1677] "400-20201012-002" "400-20201012-003" "400-20201012-004" "400-20201012-005" ...  
## $ AA\_C : num [1:1677] 65 55 70 58 67 60 59 59 65 65 ...  
## $ BiW\_C : num [1:1677] 130 127 143 140 137 130 141 138 143 150 ...  
## $ BiW\_L : num [1:1677] 115 108 121 109 104 106 109 111 113 116 ...  
## $ GoSub\_C : num [1:1677] 93 93 115 93 103 100 79 106 85 102 ...  
## $ NRB\_L : num [1:1677] 17 18 19 21 19 14 17 18 16 17 ...  
## $ ProS\_L : num [1:1677] 17 18 14 13 20 20 18 12 24 22 ...  
## $ SelP\_L : num [1:1677] 42 41 51 44 47 48 46 41 46 44 ...  
## $ SelM\_L : num [1:1677] 122 99 130 115 119 126 117 112 117 117 ...  
## $ SnasM\_C : num [1:1677] 82 55 84 74 73 80 78 76 64 75 ...  
## $ TrSman\_C : num [1:1677] 177 145 178 147 157 164 149 159 151 160 ...  
## $ TrTr\_C : num [1:1677] 296 276 292 273 279 300 283 275 307 286 ...  
## $ TrTr\_L : num [1:1677] 155 141 156 149 146 146 147 151 157 144 ...  
## $ gender : Factor w/ 4 levels "Female","Male",..: 2 1 2 2 2 2 2 2 2 2 ...  
## $ race\_eth : Factor w/ 8 levels "AIAN","Asian",..: 3 8 8 8 8 8 3 3 8 8 ...  
## $ age\_group: Factor w/ 3 levels "18-36","37-54",..: 1 2 2 1 2 3 1 1 1 1 ...

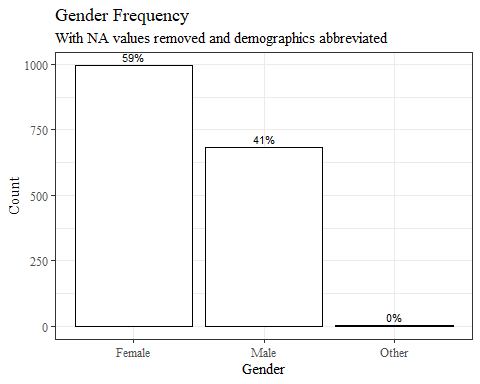
chosen\_nona1 <- chosen\_nona  
  
#race/eth  
chosen\_nona1$race\_eth <-   
 recode\_factor(chosen\_nona1$race\_eth, 'AIAN'= "Other",  
 'NHOPI' = "Other",  
 'PTNS' = "Other")  
  
#gender  
chosen\_nona1$gender <-   
 recode\_factor(chosen\_nona1$gender, 'Non-binary or Other'= "Other",  
 'Prefer not to say' = "Other")  
  
chosen\_nona1$gender[is.na(chosen\_nona1$gender)]="Other"  
  
summary(chosen\_nona1)

## ID AA\_C BiW\_C BiW\_L   
## Length:1677 Min. :44.00 Min. :101.0 Min. : 82.0   
## Class :character 1st Qu.:56.00 1st Qu.:123.0 1st Qu.:104.0   
## Mode :character Median :60.00 Median :131.0 Median :110.0   
## Mean :60.89 Mean :132.4 Mean :110.5   
## 3rd Qu.:65.00 3rd Qu.:141.0 3rd Qu.:117.0   
## Max. :87.00 Max. :187.0 Max. :148.0   
## GoSub\_C NRB\_L ProS\_L SelP\_L   
## Min. : 49.00 Min. : 3.00 Min. :12.00 Min. :31.00   
## 1st Qu.: 88.00 1st Qu.:15.00 1st Qu.:17.00 1st Qu.:42.00   
## Median : 97.00 Median :18.00 Median :19.00 Median :44.00   
## Mean : 97.87 Mean :17.86 Mean :19.09 Mean :44.33   
## 3rd Qu.:107.00 3rd Qu.:21.00 3rd Qu.:21.00 3rd Qu.:47.00   
## Max. :152.00 Max. :34.00 Max. :27.00 Max. :58.00   
## SelM\_L SnasM\_C TrSman\_C TrTr\_C   
## Min. : 84.0 Min. : 46.00 Min. :117.0 Min. :241.0   
## 1st Qu.:110.0 1st Qu.: 68.00 1st Qu.:142.0 1st Qu.:271.0   
## Median :116.0 Median : 75.00 Median :151.0 Median :281.0   
## Mean :116.3 Mean : 74.96 Mean :152.2 Mean :281.2   
## 3rd Qu.:123.0 3rd Qu.: 82.00 3rd Qu.:161.0 3rd Qu.:291.0   
## Max. :145.0 Max. :105.00 Max. :208.0 Max. :329.0   
## TrTr\_L gender race\_eth age\_group   
## Min. :127.0 Other : 2 Other : 26 18-36:826   
## 1st Qu.:140.0 Female:994 Asian : 81 37-54:777   
## Median :145.0 Male :681 Black : 446 55-72: 74   
## Mean :145.7 LatinX: 84   
## 3rd Qu.:151.0 white :1040   
## Max. :172.0

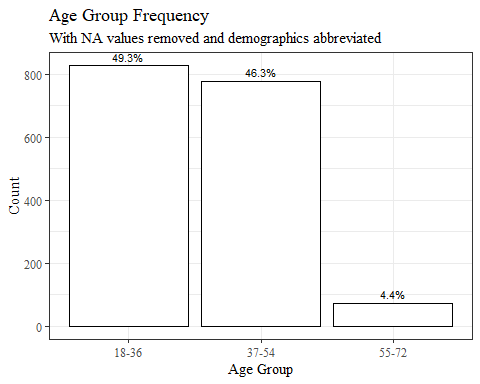
chosen\_nona1 %>%   
 ggplot(aes(x=fct\_infreq(race\_eth), label=scales::percent(prop.table(stat(count)))))+  
 geom\_bar(stat="count", color= "black", fill = "white")+  
 geom\_text(stat="count",  
 position= position\_dodge(0.9),  
 vjust = -0.5,   
 size = 3)+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Race/Ethnicity Frequency",  
 subtitle="With NA values removed and demographics abbreviated",  
 y="Count",  
 x="Race/Ethnicity")



chosen\_nona1 %>%   
 ggplot(aes(x=fct\_infreq(gender), label=scales::percent(prop.table(stat(count)))))+  
 geom\_bar(stat="count", color= "black", fill = "white")+  
 geom\_text(stat="count",  
 position= position\_dodge(0.9),  
 vjust = -0.5,   
 size = 3)+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Gender Frequency",  
 subtitle="With NA values removed and demographics abbreviated",  
 y="Count",  
 x="Gender")



chosen\_nona1 %>%   
 ggplot(aes(x=fct\_infreq(age\_group), label=scales::percent(prop.table(stat(count)))))+  
 geom\_bar(stat="count", color= "black", fill = "white")+  
 geom\_text(stat="count",  
 position= position\_dodge(0.9),  
 vjust = -0.5,   
 size = 3)+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Age Group Frequency",  
 subtitle="With NA values removed and demographics abbreviated",  
 y="Count",  
 x="Age Group")



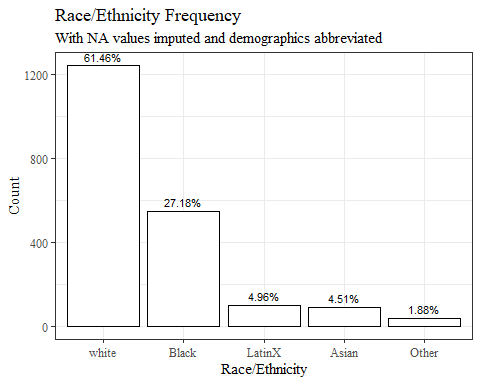
chosen\_imputed<-read\_excel("C:\\Users\\19177\\OneDrive - Colostate\\Desktop\\Dissertation\\headscan\_dissertation\\chosen\_imputed.xlsx")  
  
chosen\_imputed$gender<-as.factor(chosen\_imputed$gender)  
chosen\_imputed$race\_eth<-as.factor(chosen\_imputed$race\_eth)  
chosen\_imputed$age\_group<-as.factor(chosen\_imputed$age\_group)  
  
  
str(chosen\_imputed)

## tibble [2,016 × 16] (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] 65 55 70 58 67 60 59 59 65 65 ...  
## $ BiW\_C : num [1:2016] 130 127 143 140 137 130 141 138 143 150 ...  
## $ BiW\_L : num [1:2016] 115 108 121 109 104 106 109 111 113 116 ...  
## $ GoSub\_C : num [1:2016] 93 93 115 93 103 100 79 106 85 102 ...  
## $ NRB\_L : num [1:2016] 17 18 19 21 19 14 17 18 16 17 ...  
## $ ProS\_L : num [1:2016] 17 18 14 13 20 20 18 12 24 22 ...  
## $ SelP\_L : num [1:2016] 42 41 51 44 47 48 46 41 46 44 ...  
## $ SelM\_L : num [1:2016] 122 99 130 115 119 126 117 112 117 117 ...  
## $ SnasM\_C : num [1:2016] 82 55 84 74 73 80 78 76 64 75 ...  
## $ TrSman\_C : num [1:2016] 177 145 178 147 157 164 149 159 151 160 ...  
## $ TrTr\_C : num [1:2016] 296 276 292 273 279 300 283 275 307 286 ...  
## $ TrTr\_L : num [1:2016] 155 141 156 149 146 146 147 151 157 144 ...  
## $ gender : Factor w/ 4 levels "Female","Male",..: 2 1 2 2 2 2 2 2 2 2 ...  
## $ race\_eth : Factor w/ 8 levels "AIAN","Asian",..: 3 8 8 8 8 8 3 3 8 8 ...  
## $ age\_group: Factor w/ 3 levels "18-36","37-54",..: 1 2 2 1 2 3 1 1 1 1 ...

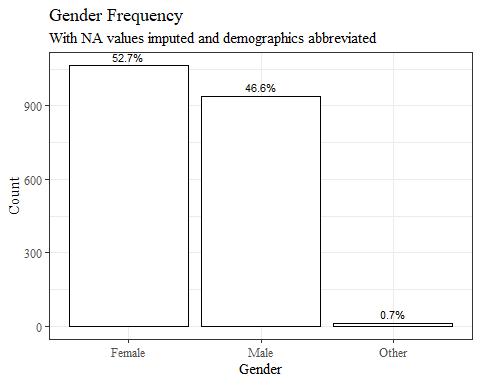
chosen\_imputed1 <- chosen\_imputed  
  
#race/eth  
chosen\_imputed1$race\_eth <-   
 recode\_factor(chosen\_imputed1$race\_eth, 'AIAN'= "Other",  
 'NHOPI' = "Other",  
 'PTNS' = "Other")  
  
#gender  
chosen\_imputed1$gender <-   
 recode\_factor(chosen\_imputed1$gender, 'Non-binary or Other'= "Other",  
 'Prefer not to say' = "Other")  
  
chosen\_imputed1$gender[is.na(chosen\_imputed1$gender)]="Other"  
  
summary(chosen\_imputed1)

## ID AA\_C BiW\_C BiW\_L   
## Length:2016 Min. :44.00 Min. :101.0 Min. : 82.0   
## Class :character 1st Qu.:57.00 1st Qu.:124.0 1st Qu.:105.0   
## Mode :character Median :61.00 Median :133.0 Median :111.0   
## Mean :61.22 Mean :133.4 Mean :111.2   
## 3rd Qu.:65.00 3rd Qu.:141.0 3rd Qu.:117.0   
## Max. :87.00 Max. :187.0 Max. :152.0   
## GoSub\_C NRB\_L ProS\_L SelP\_L SelM\_L   
## Min. : 49.00 Min. : 3.00 Min. :12.00 Min. :30.00 Min. : 84   
## 1st Qu.: 89.00 1st Qu.:15.00 1st Qu.:17.00 1st Qu.:42.00 1st Qu.:111   
## Median : 99.00 Median :18.00 Median :19.00 Median :44.00 Median :117   
## Mean : 99.17 Mean :17.95 Mean :19.12 Mean :44.53 Mean :117   
## 3rd Qu.:109.00 3rd Qu.:21.00 3rd Qu.:21.00 3rd Qu.:47.00 3rd Qu.:123   
## Max. :152.00 Max. :35.00 Max. :27.00 Max. :58.00 Max. :145   
## SnasM\_C TrSman\_C TrTr\_C TrTr\_L gender   
## Min. : 44.00 Min. :117.0 Min. :241.0 Min. :127.0 Other : 14   
## 1st Qu.: 69.00 1st Qu.:144.0 1st Qu.:273.0 1st Qu.:141.0 Female:1063   
## Median : 76.00 Median :154.0 Median :283.0 Median :146.0 Male : 939   
## Mean : 75.62 Mean :153.8 Mean :282.7 Mean :146.5   
## 3rd Qu.: 83.00 3rd Qu.:162.0 3rd Qu.:292.0 3rd Qu.:152.0   
## Max. :105.00 Max. :208.0 Max. :332.0 Max. :173.0   
## race\_eth age\_group   
## Other : 38 18-36:991   
## Asian : 91 37-54:940   
## Black : 548 55-72: 84   
## LatinX: 100 NA's : 1   
## white :1239   
##

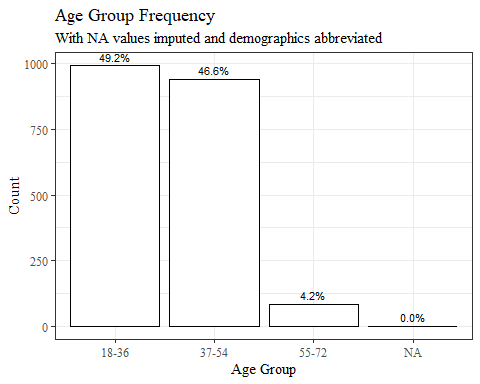
chosen\_imputed1 %>%   
 ggplot(aes(x=fct\_infreq(race\_eth), label=scales::percent(prop.table(stat(count)))))+  
 geom\_bar(stat="count", color= "black", fill = "white")+  
 geom\_text(stat="count",  
 position= position\_dodge(0.9),  
 vjust = -0.5,   
 size = 3)+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Race/Ethnicity Frequency",  
 subtitle="With NA values imputed and demographics abbreviated",  
 y="Count",  
 x="Race/Ethnicity")



chosen\_imputed1 %>%   
 ggplot(aes(x=fct\_infreq(gender), label=scales::percent(prop.table(stat(count)))))+  
 geom\_bar(stat="count", color= "black", fill = "white")+  
 geom\_text(stat="count",  
 position= position\_dodge(0.9),  
 vjust = -0.5,   
 size = 3)+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Gender Frequency",  
 subtitle="With NA values imputed and demographics abbreviated",  
 y="Count",  
 x="Gender")



chosen\_imputed1 %>%   
 ggplot(aes(x=fct\_infreq(age\_group), label=scales::percent(prop.table(stat(count)))))+  
 geom\_bar(stat="count", color= "black", fill = "white")+  
 geom\_text(stat="count",  
 position= position\_dodge(0.9),  
 vjust = -0.5,   
 size = 3)+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Age Group Frequency",  
 subtitle="With NA values imputed and demographics abbreviated",  
 y="Count",  
 x="Age Group")



#write\_xlsx(chosen\_withna1, "C:\\Users\\19177\\OneDrive - Colostate\\Desktop\\Dissertation\\headscan\_dissertation\\chosen\_withna1.xlsx")  
#(chosen\_nona1, "C:\\Users\\19177\\OneDrive - Colostate\\Desktop\\Dissertation\\headscan\_dissertation\\chosen\_nona1.xlsx")  
#write\_xlsx(chosen\_imputed1, "C:\\Users\\19177\\OneDrive - Colostate\\Desktop\\Dissertation\\headscan\_dissertation\\chosen\_imputed1.xlsx")