PCA2\_imputed

2022-09-07

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(flextable)

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

library(readxl)  
library(extrafont)

## Registering fonts with R

library(fauxnaif) #na if in  
library(reshape2) #melt

##   
## Attaching package: 'reshape2'  
##   
## The following object is masked from 'package:tidyr':  
##   
## smiths

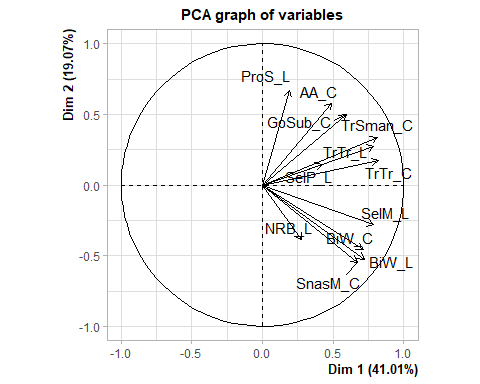
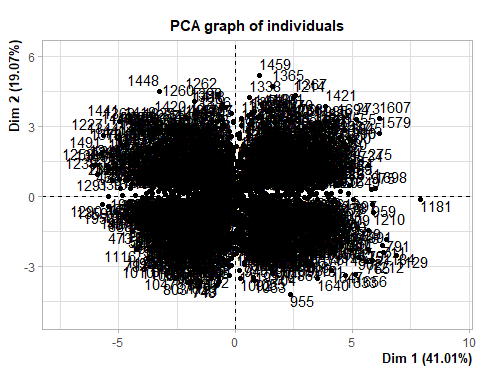
library(FactoMineR) #PCA

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

## 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 : 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" ...

#https://tem11010.github.io/Plotting-PCAs/  
  
imp\_selected1 <- imp\_selected %>% filter (AA\_C > 0, BiW\_C > 0, BiW\_L > 0, GoSub\_C > 0,   
 NRB\_L > 0, ProS\_L > 0, SelP\_L > 0, SelM\_L > 0,   
 SnasM\_C > 0, TrSman\_C > 0, TrTr\_C > 0, TrTr\_L > 0)  
  
# log price  
  
# center and scale the data  
for (i in 1:length(colnames(imp\_selected1))){  
   
 if (is.numeric(imp\_selected1[, i])==TRUE)  
   
 imp\_selected1[, i] <- as.numeric(scale(imp\_selected1[, i]))  
   
 else  
   
 imp\_selected1[, i] <- imp\_selected1[, i]  
   
}

pca1.1 <- PCA(imp\_selected1[ ,c("AA\_C", "BiW\_C", "BiW\_L", "GoSub\_C", "NRB\_L", "ProS\_L", "SelP\_L",   
 "SelM\_L", "SnasM\_C", "TrSman\_C", "TrTr\_C", "TrTr\_L")])



summary(pca1.1)

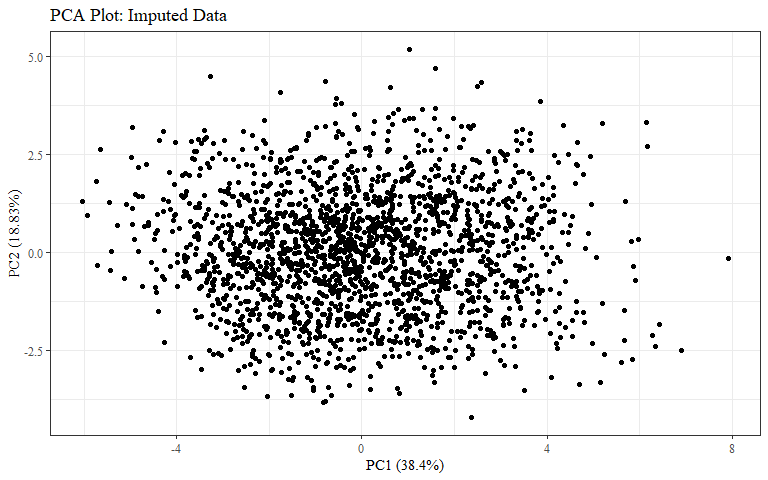
##   
## Call:  
## PCA(X = imp\_selected1[, c("AA\_C", "BiW\_C", "BiW\_L", "GoSub\_C",   
## "NRB\_L", "ProS\_L", "SelP\_L", "SelM\_L", "SnasM\_C", "TrSman\_C",   
## "TrTr\_C", "TrTr\_L")])   
##   
##   
## Eigenvalues  
## Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6 Dim.7  
## Variance 4.921 2.288 1.311 0.916 0.695 0.623 0.423  
## % of var. 41.007 19.068 10.926 7.637 5.790 5.194 3.526  
## Cumulative % of var. 41.007 60.075 71.000 78.638 84.427 89.621 93.147  
## Dim.8 Dim.9 Dim.10 Dim.11 Dim.12  
## Variance 0.336 0.250 0.110 0.076 0.051  
## % of var. 2.800 2.081 0.913 0.632 0.427  
## Cumulative % of var. 95.946 98.027 98.940 99.573 100.000  
##   
## Individuals (the 10 first)  
## Dist Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3 ctr  
## 1 | 2.744 | 1.599 0.026 0.340 | 0.020 0.000 0.000 | -0.891 0.030  
## 2 | 3.400 | -2.769 0.077 0.663 | 0.235 0.001 0.005 | -1.087 0.045  
## 3 | 4.358 | 3.397 0.116 0.608 | -0.353 0.003 0.007 | 0.103 0.000  
## 4 | 2.814 | -0.724 0.005 0.066 | -1.697 0.062 0.364 | -0.582 0.013  
## 5 | 1.524 | 0.263 0.001 0.030 | 0.832 0.015 0.298 | 0.675 0.017  
## 6 | 2.258 | 0.972 0.010 0.185 | 0.614 0.008 0.074 | 0.622 0.015  
## 7 | 1.726 | -0.349 0.001 0.041 | -0.955 0.020 0.306 | 0.803 0.024  
## 8 | 3.180 | -0.275 0.001 0.007 | -1.195 0.031 0.141 | -1.704 0.110  
## 9 | 3.476 | 1.061 0.011 0.093 | 1.424 0.044 0.168 | 0.962 0.035  
## 10 | 2.020 | 0.938 0.009 0.215 | 0.363 0.003 0.032 | 0.409 0.006  
## cos2   
## 1 0.106 |  
## 2 0.102 |  
## 3 0.001 |  
## 4 0.043 |  
## 5 0.196 |  
## 6 0.076 |  
## 7 0.216 |  
## 8 0.287 |  
## 9 0.077 |  
## 10 0.041 |  
##   
## Variables (the 10 first)  
## Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3 ctr cos2   
## AA\_C | 0.488 4.833 0.238 | 0.578 14.595 0.334 | 0.317 7.645 0.100 |  
## BiW\_C | 0.712 10.299 0.507 | -0.458 9.169 0.210 | 0.185 2.622 0.034 |  
## BiW\_L | 0.720 10.540 0.519 | -0.525 12.048 0.276 | -0.059 0.268 0.004 |  
## GoSub\_C | 0.591 7.101 0.349 | 0.499 10.900 0.249 | -0.442 14.921 0.196 |  
## NRB\_L | 0.277 1.559 0.077 | -0.388 6.574 0.150 | -0.275 5.763 0.076 |  
## ProS\_L | 0.192 0.751 0.037 | 0.670 19.628 0.449 | 0.276 5.822 0.076 |  
## SelP\_L | 0.421 3.609 0.178 | 0.148 0.959 0.022 | 0.744 42.172 0.553 |  
## SelM\_L | 0.786 12.566 0.618 | -0.281 3.444 0.079 | 0.312 7.401 0.097 |  
## SnasM\_C | 0.672 9.189 0.452 | -0.547 13.076 0.299 | 0.056 0.235 0.003 |  
## TrSman\_C | 0.812 13.395 0.659 | 0.339 5.027 0.115 | -0.328 8.185 0.107 |

imp\_selected1$pc1 <- pca1.1$ind$coord[, 1] # indexing the first column  
  
imp\_selected1$pc2 <- pca1.1$ind$coord[, 2] # indexing the second column

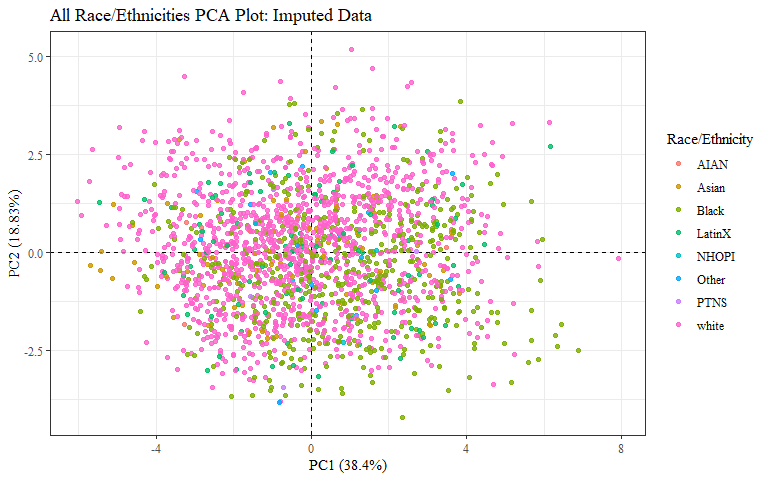
pca.vars1 <- pca1.1$var$coord %>% data.frame  
  
pca.vars1$vars1 <- rownames(pca.vars1)  
  
pca.vars.m1 <- melt(pca.vars1, id.vars = "vars1")

#circleFun <- function(center = c(0,0),diameter = 1, npoints = 100){  
 #r = diameter / 2  
 #tt <- seq(0,2\*pi,length.out = npoints)  
 #xx <- center[1] + r \* cos(tt)  
 #yy <- center[2] + r \* sin(tt)  
 #return(data.frame(x = xx, y = yy))  
#}  
  
#circ <- circleFun(c(0,0),2,npoints = 500)

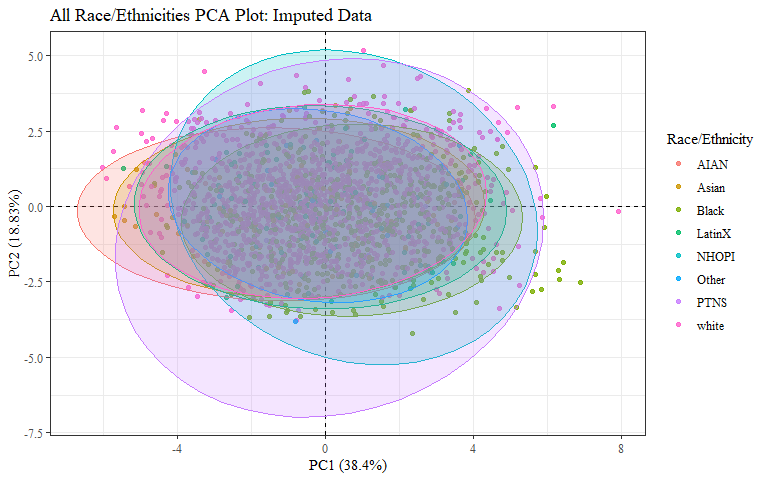
ggplot(data = imp\_selected1, aes(x = pc1, y = pc2)) +  
 geom\_point()+  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title="PCA Plot: Imputed Data")+  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))



#checking to make sure this matches plot in PCA.Rmd  
ggplot(data = imp\_selected1, aes(x = pc1, y = pc2, color = race\_eth)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title = "All Race/Ethnicities PCA Plot: Imputed Data",  
 color = "Race/Ethnicity")+  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))



ggplot(data = imp\_selected1, aes(x = pc1, y = pc2, color = race\_eth)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +   
 stat\_ellipse(geom="polygon", aes(fill = race\_eth),   
 alpha = 0.2,   
 show.legend = FALSE,  
 level = 0.95) +  
 #theme\_minimal() +  
 #theme(panel.grid = element\_blank(), panel.border = element\_rect(fill= "transparent"))+  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title = "All Race/Ethnicities PCA Plot: Imputed Data",  
 color = "Race/Ethnicity")+  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))

 Need a plot with just Asian, Black, LatinX, and white

imp\_selected2 <- imp\_selected1  
  
imp\_selected2$race\_eth <- as.factor(imp\_selected2$race\_eth)  
imp\_selected2$gender <- as.factor(imp\_selected2$gender)  
imp\_selected2$age\_group <- as.factor(imp\_selected2$age\_group)

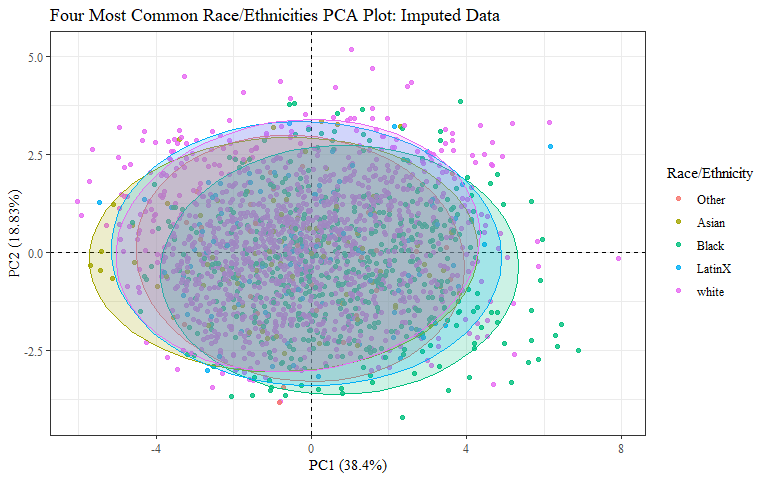
PCAdata\_race\_eth <- imp\_selected2  
  
PCAdata\_race\_eth$race\_eth <-   
 recode\_factor(PCAdata\_race\_eth$race\_eth, 'AIAN'= "Other",  
 'NHOPI' = "Other",  
 'PTNS' = "Other")  
  
PCAdata\_race\_eth\_nas <- imp\_selected1 %>%   
 mutate(race\_eth = na\_if\_in(race\_eth, c("AIAN", "Other", "NHOPI", "PTNS")))  
  
str(PCAdata\_race\_eth$race\_eth)

## Factor w/ 5 levels "Other","Asian",..: 3 5 5 5 5 5 3 3 5 5 ...

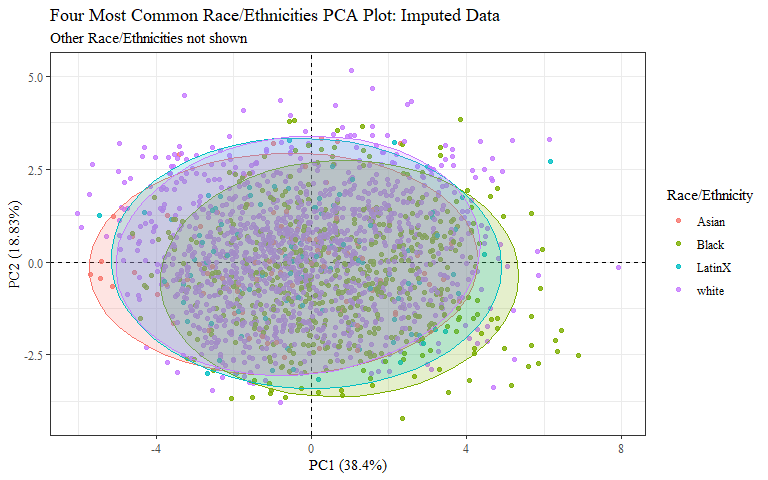
str(PCAdata\_race\_eth\_nas$race\_eth)

## chr [1:2016] "Black" "white" "white" "white" "white" "white" "Black" ...

ggplot(data = PCAdata\_race\_eth, aes(x = pc1, y = pc2, color = race\_eth)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +   
 stat\_ellipse(geom="polygon", aes(fill = race\_eth),   
 alpha = 0.2,   
 show.legend = FALSE,  
 level = 0.95) +  
 #theme\_minimal() +  
 #theme(panel.grid = element\_blank(), panel.border = element\_rect(fill= "transparent"))+  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title="Four Most Common Race/Ethnicities PCA Plot: Imputed Data",  
 color = "Race/Ethnicity")+  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))



PCAdata\_race\_eth\_nas %>%   
 drop\_na(race\_eth) %>%   
 ggplot(aes(x = pc1, y = pc2, color = race\_eth)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +   
 stat\_ellipse(geom="polygon", aes(fill = race\_eth),   
 alpha = 0.2,   
 show.legend = FALSE,  
 level = 0.95) +  
 #theme\_minimal() +  
 #theme(panel.grid = element\_blank(), panel.border = element\_rect(fill= "transparent"))+  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title="Four Most Common Race/Ethnicities PCA Plot: Imputed Data",  
 subtitle="Other Race/Ethnicities not shown",  
 color = "Race/Ethnicity")+  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))



Just Asian and Black?

PCAdata\_race\_eth1 <- PCAdata\_race\_eth  
  
PCAdata\_race\_eth1$race\_eth <-  
 recode\_factor(PCAdata\_race\_eth1$race\_eth, 'white'= "Other",  
 'LatinX' = "Other")  
  
PCAdata\_race\_eth\_nas1 <- PCAdata\_race\_eth\_nas %>%   
 mutate(race\_eth = na\_if\_in(race\_eth, c("white", "LatinX")))  
  
str(PCAdata\_race\_eth1$race\_eth)

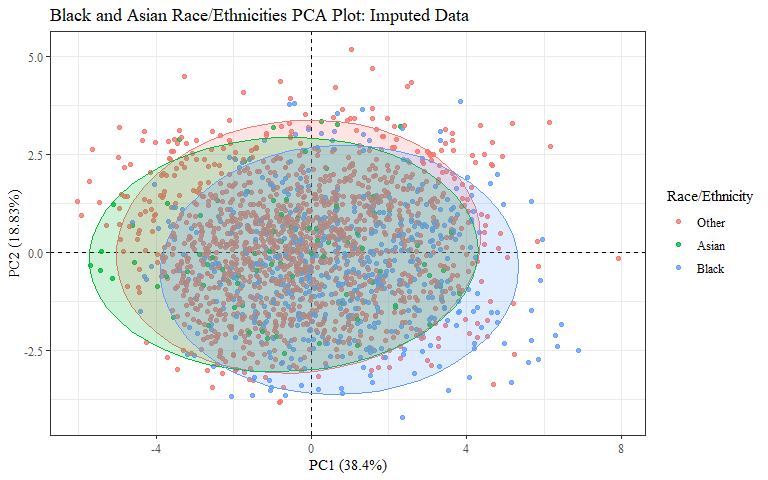
## Factor w/ 3 levels "Other","Asian",..: 3 1 1 1 1 1 3 3 1 1 ...

str(PCAdata\_race\_eth\_nas1$race\_eth)

## chr [1:2016] "Black" NA NA NA NA NA "Black" "Black" NA NA NA NA NA NA ...

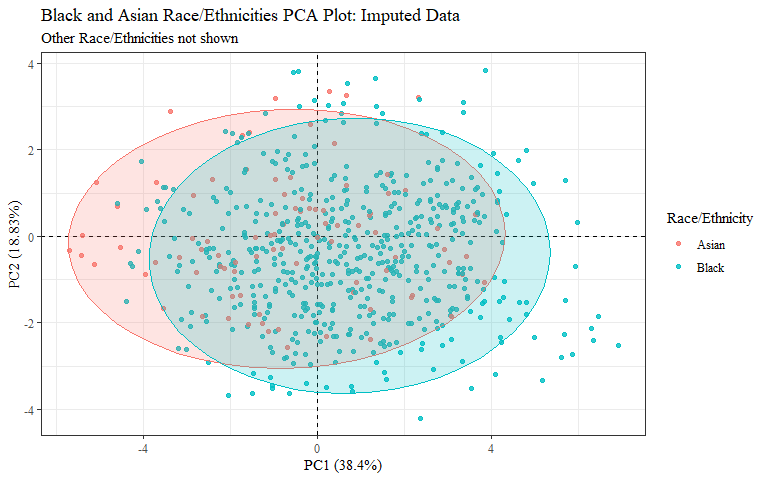
ggplot(data = PCAdata\_race\_eth1, aes(x = pc1, y = pc2, color = race\_eth)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +   
 stat\_ellipse(geom="polygon", aes(fill = race\_eth, na.rm=TRUE),   
 alpha = 0.2,   
 show.legend = FALSE,  
 level = 0.95) +  
 #theme\_minimal() +  
 #theme(panel.grid = element\_blank(), panel.border = element\_rect(fill= "transparent"))+  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title="Black and Asian Race/Ethnicities PCA Plot: Imputed Data",  
 color = "Race/Ethnicity")+  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))

## Warning: Ignoring unknown aesthetics: na.rm

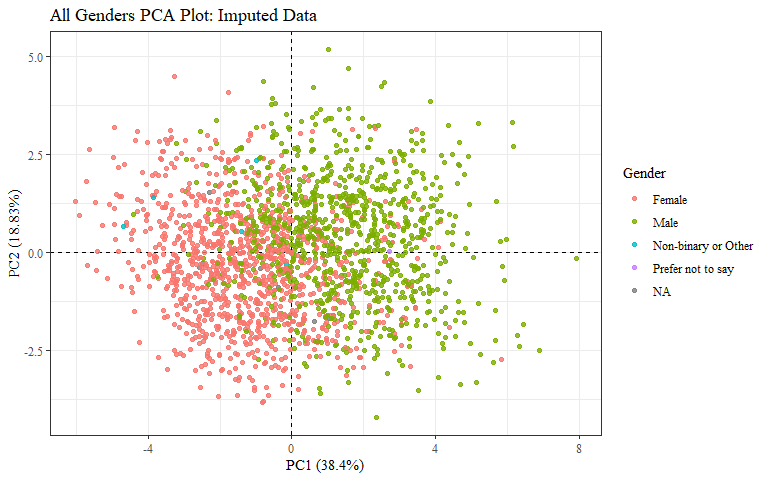


PCAdata\_race\_eth\_nas1 %>%   
 drop\_na(race\_eth) %>%   
 ggplot(aes(x = pc1, y = pc2, color = race\_eth)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +   
 stat\_ellipse(geom="polygon", aes(fill = race\_eth, na.rm=TRUE),   
 alpha = 0.2,   
 show.legend = FALSE,  
 level = 0.95) +  
 #theme\_minimal() +  
 #theme(panel.grid = element\_blank(), panel.border = element\_rect(fill= "transparent"))+  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title="Black and Asian Race/Ethnicities PCA Plot: Imputed Data",  
 subtitle = "Other Race/Ethnicities not shown",  
 color = "Race/Ethnicity")+  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))

## Warning: Ignoring unknown aesthetics: na.rm

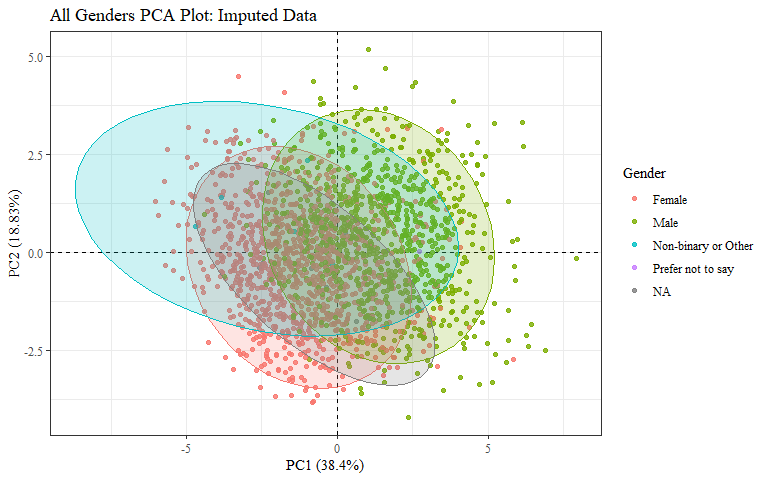


#checking to make sure this matches plot in PCA.Rmd  
ggplot(data = imp\_selected1, aes(x = pc1, y = pc2, color = gender)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title = "All Genders PCA Plot: Imputed Data",  
 color= "Gender")+  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))



ggplot(data = imp\_selected1, aes(x = pc1, y = pc2, color = gender)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +   
 stat\_ellipse(geom="polygon", aes(fill = gender),   
 alpha = 0.2,   
 show.legend = FALSE,  
 level = 0.95) +  
 #theme\_minimal() +  
 #theme(panel.grid = element\_blank(), panel.border = element\_rect(fill= "transparent"))+  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title="All Genders PCA Plot: Imputed Data",  
 color= "Gender")+  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))

## Too few points to calculate an ellipse

 Need plot with just Male and Female

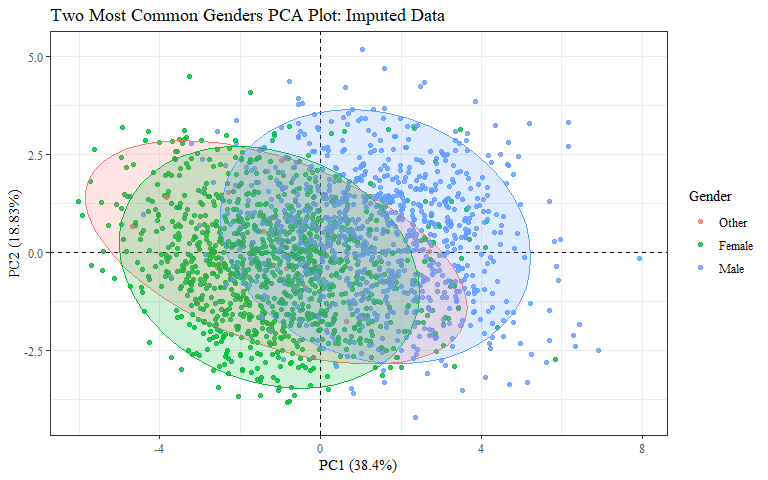
PCAdata\_gender <- imp\_selected2  
  
PCAdata\_gender$gender <-   
 recode\_factor(PCAdata\_gender$gender, 'Non-binary or Other'= "Other",  
 'Prefer not to say' = "Other")  
  
PCAdata\_gender$gender[is.na(PCAdata\_gender$gender)]="Other"  
  
PCAdata\_gender\_nas <- imp\_selected1 %>%   
 mutate(gender = na\_if\_in(gender, c("Non-binary or Other", "Prefer not to say")))  
  
str(PCAdata\_gender$gender)

## Factor w/ 3 levels "Other","Female",..: 3 2 3 3 3 3 3 3 3 3 ...

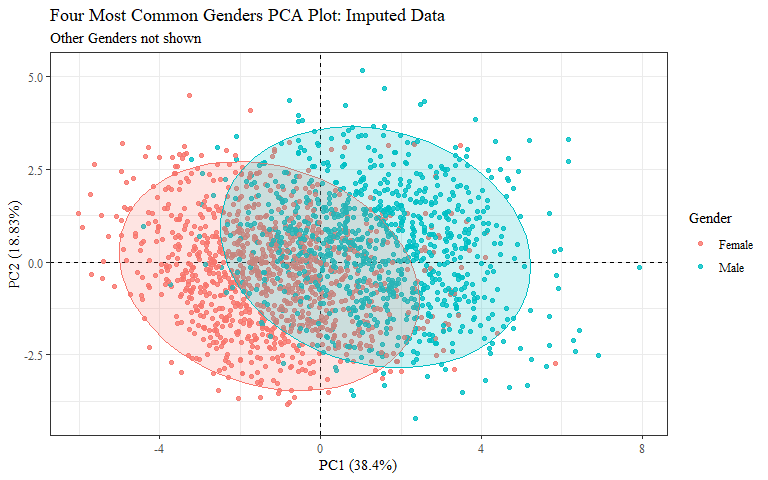
str(PCAdata\_gender\_nas$gender)

## chr [1:2016] "Male" "Female" "Male" "Male" "Male" "Male" "Male" "Male" ...

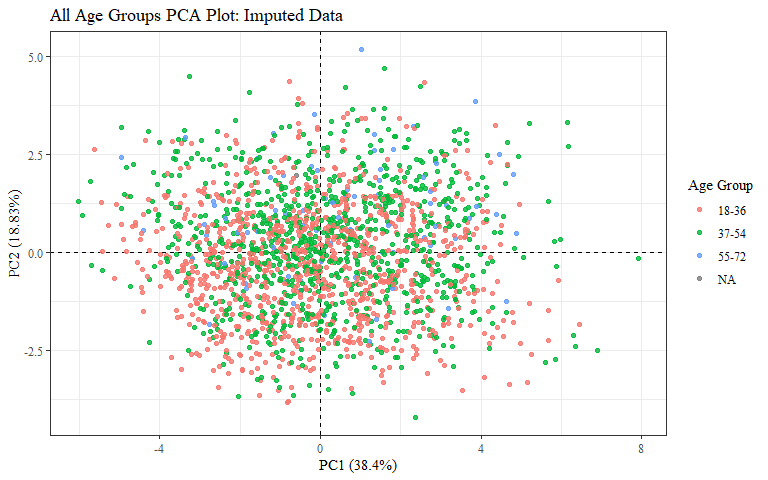
ggplot(data = PCAdata\_gender, aes(x = pc1, y = pc2, color = gender)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +   
 stat\_ellipse(geom="polygon", aes(fill = gender),   
 alpha = 0.2,   
 show.legend = FALSE,  
 level = 0.95) +  
 #theme\_minimal() +  
 #theme(panel.grid = element\_blank(), panel.border = element\_rect(fill= "transparent"))+  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title="Two Most Common Genders PCA Plot: Imputed Data",  
 color = "Gender")+  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))



PCAdata\_gender\_nas %>%   
 drop\_na(gender) %>%   
 ggplot(aes(x = pc1, y = pc2, color = gender)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +   
 stat\_ellipse(geom="polygon", aes(fill = gender),   
 alpha = 0.2,   
 show.legend = FALSE,  
 level = 0.95) +  
 #theme\_minimal() +  
 #theme(panel.grid = element\_blank(), panel.border = element\_rect(fill= "transparent"))+  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title="Four Most Common Genders PCA Plot: Imputed Data",  
 subtitle="Other Genders not shown",  
 color = "Gender")+  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))



#checking to make sure this matches plot in PCA.Rmd  
ggplot(data = imp\_selected1, aes(x = pc1, y = pc2, color = age\_group)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title = "All Age Groups PCA Plot: Imputed Data",  
 color= "Age Group")+  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))



ggplot(data = imp\_selected1, aes(x = pc1, y = pc2, color = age\_group)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +   
 stat\_ellipse(geom="polygon", aes(fill = age\_group),   
 alpha = 0.2,   
 show.legend = FALSE,  
 level = 0.95) +  
 #theme\_minimal() +  
 #theme(panel.grid = element\_blank(), panel.border = element\_rect(fill= "transparent"))+  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title = "All Age Groups PCA Plot: Imputed Data",  
 color= "Age Group")+  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))

## Too few points to calculate an ellipse

