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

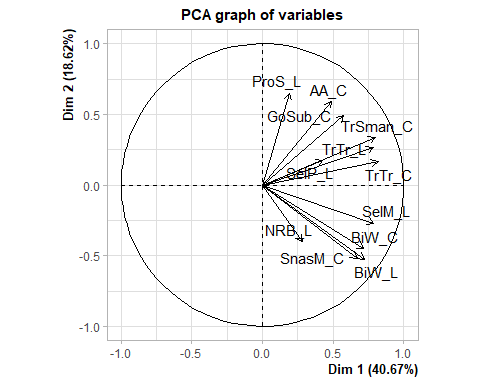
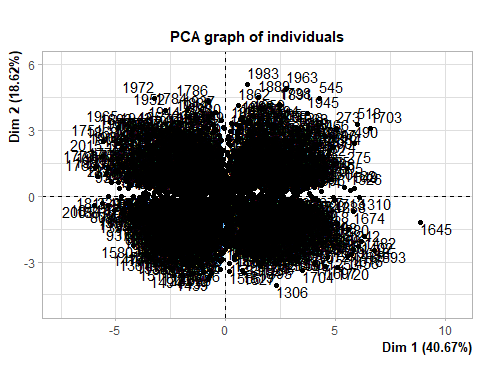
library(FactoMineR) #PCA

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

## tibble [2,016 × 15] (S3: tbl\_df/tbl/data.frame)  
## $ 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)

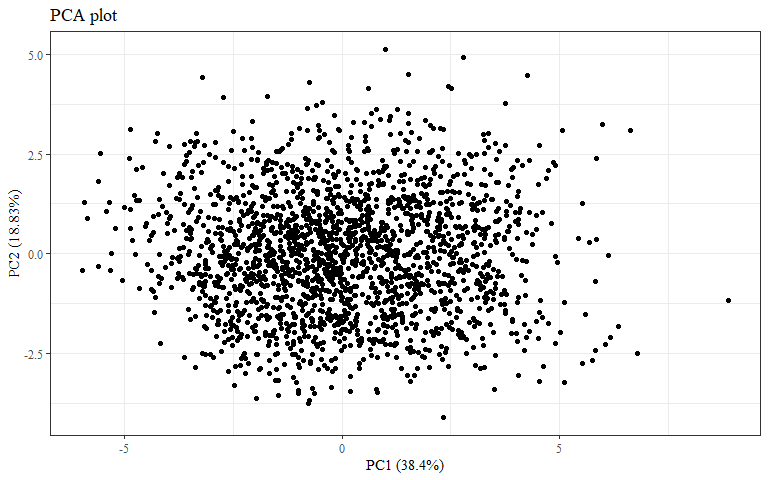
##   
## Call:  
## PCA(X = imp\_selected1[, c("AA\_C", "BiW\_C", "BiW\_L", "GoSub\_C",   
## "NRB\_L", "ProS\_L", "SelP\_L", "SelM\_L", "SnasM\_C", "TrSman\_C",   
## "TrTr\_C", "TrTr\_L")])   
##   
##   
## Eigenvalues  
## Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6 Dim.7  
## Variance 4.880 2.235 1.253 0.923 0.710 0.654 0.429  
## % of var. 40.665 18.622 10.443 7.693 5.919 5.449 3.579  
## Cumulative % of var. 40.665 59.288 69.730 77.423 83.343 88.791 92.370  
## Dim.8 Dim.9 Dim.10 Dim.11 Dim.12  
## Variance 0.356 0.250 0.152 0.089 0.069  
## % of var. 2.969 2.082 1.265 0.741 0.573  
## Cumulative % of var. 95.338 97.421 98.686 99.427 100.000  
##   
## Individuals (the 10 first)  
## Dist Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3 ctr  
## 1 | 2.680 | 1.560 0.025 0.339 | 0.042 0.000 0.000 | -0.867 0.030  
## 2 | 3.332 | -2.723 0.075 0.668 | 0.174 0.001 0.003 | -1.044 0.043  
## 3 | 4.238 | 3.329 0.113 0.617 | -0.259 0.001 0.004 | 0.143 0.001  
## 4 | 2.700 | -0.702 0.005 0.068 | -1.625 0.059 0.362 | -0.498 0.010  
## 5 | 1.498 | 0.245 0.001 0.027 | 0.830 0.015 0.307 | 0.628 0.016  
## 6 | 2.217 | 0.936 0.009 0.178 | 0.626 0.009 0.080 | 0.606 0.015  
## 7 | 1.680 | -0.332 0.001 0.039 | -0.908 0.018 0.292 | 0.813 0.026  
## 8 | 3.040 | -0.279 0.001 0.008 | -1.135 0.029 0.139 | -1.588 0.100  
## 9 | 3.380 | 1.062 0.011 0.099 | 1.350 0.040 0.159 | 0.867 0.030  
## 10 | 1.957 | 0.911 0.008 0.217 | 0.328 0.002 0.028 | 0.359 0.005  
## cos2   
## 1 0.105 |  
## 2 0.098 |  
## 3 0.001 |  
## 4 0.034 |  
## 5 0.176 |  
## 6 0.075 |  
## 7 0.234 |  
## 8 0.273 |  
## 9 0.066 |  
## 10 0.034 |  
##   
## Variables (the 10 first)  
## Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3 ctr cos2   
## AA\_C | 0.488 4.880 0.238 | 0.592 15.664 0.350 | 0.287 6.592 0.083 |  
## BiW\_C | 0.715 10.481 0.511 | -0.450 9.054 0.202 | 0.182 2.631 0.033 |  
## BiW\_L | 0.719 10.605 0.518 | -0.527 12.423 0.278 | -0.051 0.209 0.003 |  
## GoSub\_C | 0.576 6.796 0.332 | 0.489 10.706 0.239 | -0.442 15.571 0.195 |  
## NRB\_L | 0.280 1.604 0.078 | -0.402 7.239 0.162 | -0.276 6.081 0.076 |  
## ProS\_L | 0.191 0.745 0.036 | 0.650 18.898 0.422 | 0.244 4.768 0.060 |  
## SelP\_L | 0.427 3.731 0.182 | 0.173 1.346 0.030 | 0.738 43.508 0.545 |  
## SelM\_L | 0.788 12.731 0.621 | -0.270 3.267 0.073 | 0.296 7.009 0.088 |  
## SnasM\_C | 0.669 9.177 0.448 | -0.520 12.117 0.271 | 0.063 0.321 0.004 |  
## TrSman\_C | 0.797 13.014 0.635 | 0.333 4.971 0.111 | -0.323 8.327 0.104 |

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

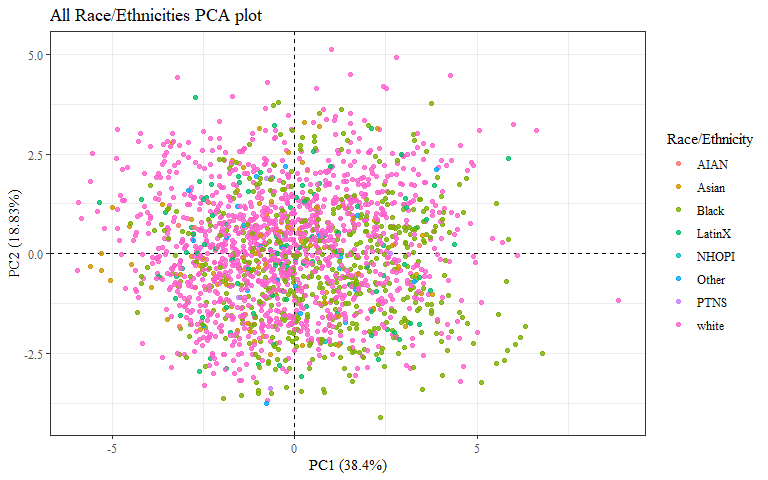
pca.vars1 <- pca1.1$var$coord %>% data.frame  
  
pca.vars1$vars1 <- rownames(pca.vars1)  
  
pca.vars.m1 <- melt(pca.vars1, id.vars = "vars1")

#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)

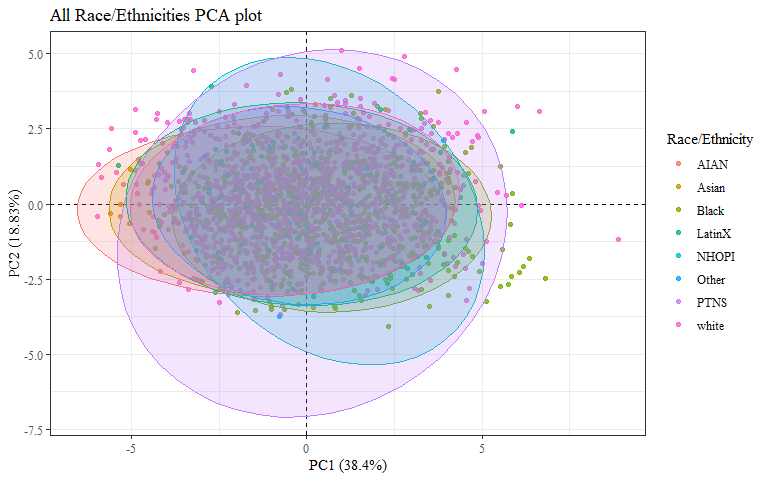
ggplot(data = imp\_selected1, aes(x = pc1, y = pc2)) +  
 geom\_point()+  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title="PCA plot")+  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))



#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",  
 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",  
 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)

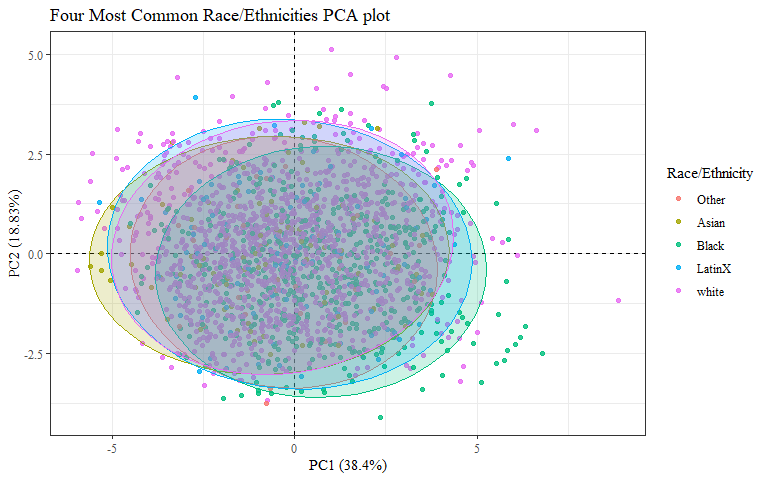
PCAdata\_race\_eth <- imp\_selected2  
  
PCAdata\_race\_eth$race\_eth <-   
 recode\_factor(PCAdata\_race\_eth$race\_eth, 'AIAN'= "Other",  
 'NHOPI' = "Other",  
 'PTNS' = "Other")  
  
PCAdata\_race\_eth\_nas <- imp\_selected1 %>%   
 mutate(race\_eth = na\_if\_in(race\_eth, c("AIAN", "Other", "NHOPI",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

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

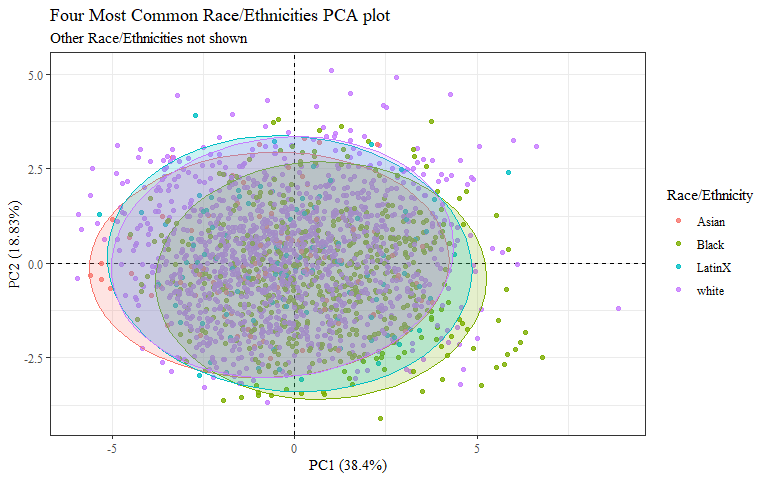
str(PCAdata\_race\_eth\_nas$race\_eth)

## 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",  
 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",  
 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)

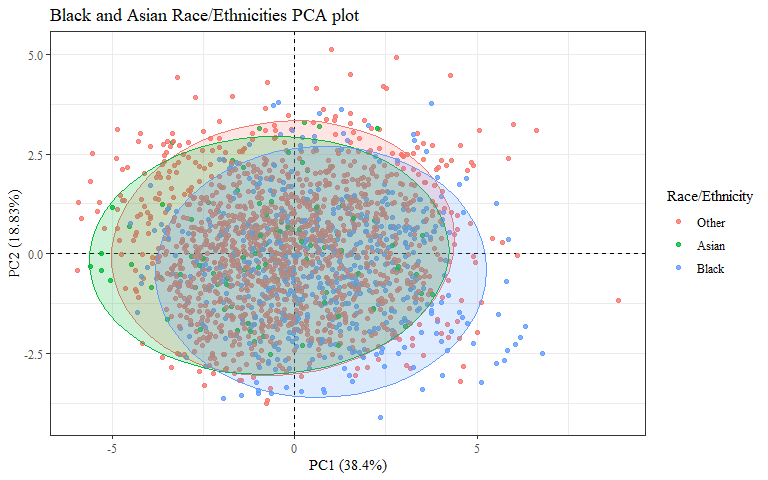
## Factor w/ 3 levels "Other","Asian",..: 3 1 1 1 1 1 3 3 1 1 ...

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

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

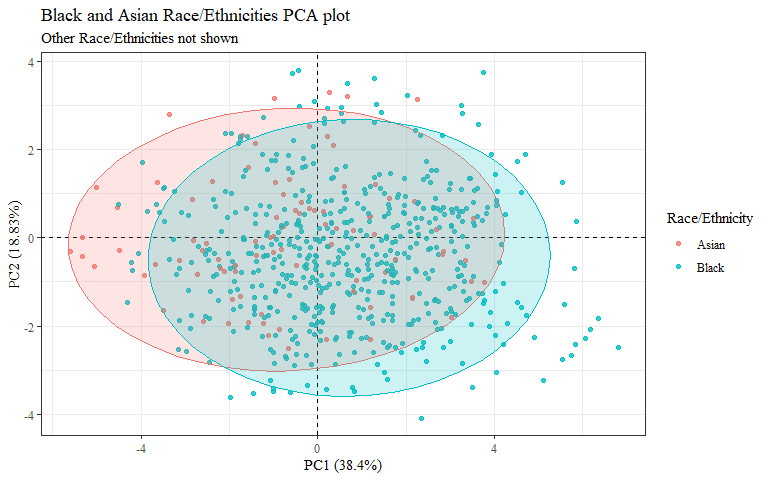
ggplot(data = PCAdata\_race\_eth1, aes(x = pc1, y = pc2, color = race\_eth)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +   
 stat\_ellipse(geom="polygon", aes(fill = race\_eth, na.rm=TRUE),   
 alpha = 0.2,   
 show.legend = FALSE,  
 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",  
 color = "Race/Ethnicity")+  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))

## Warning: Ignoring unknown aesthetics: na.rm

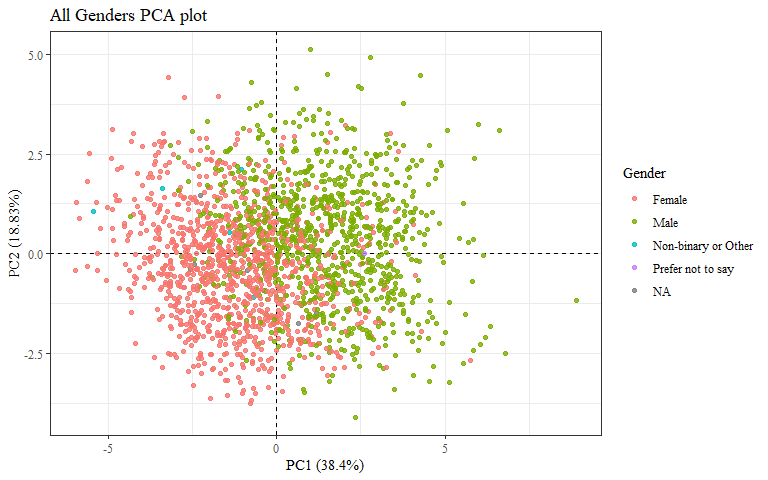


PCAdata\_race\_eth\_nas1 %>%   
 drop\_na(race\_eth) %>%   
 ggplot(aes(x = pc1, y = pc2, color = race\_eth)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +   
 stat\_ellipse(geom="polygon", aes(fill = race\_eth, na.rm=TRUE),   
 alpha = 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",  
 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

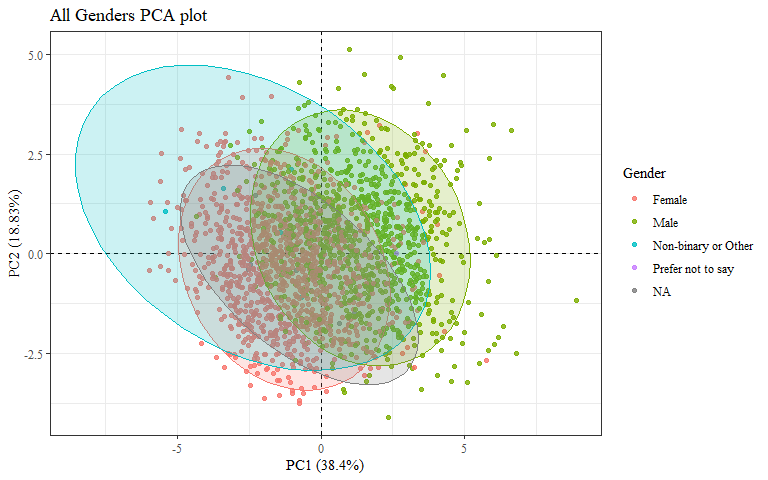


#checking to make sure this matches plot in PCA.Rmd  
ggplot(data = imp\_selected1, aes(x = pc1, y = pc2, color = gender)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title = "All Genders PCA plot",  
 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",  
 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

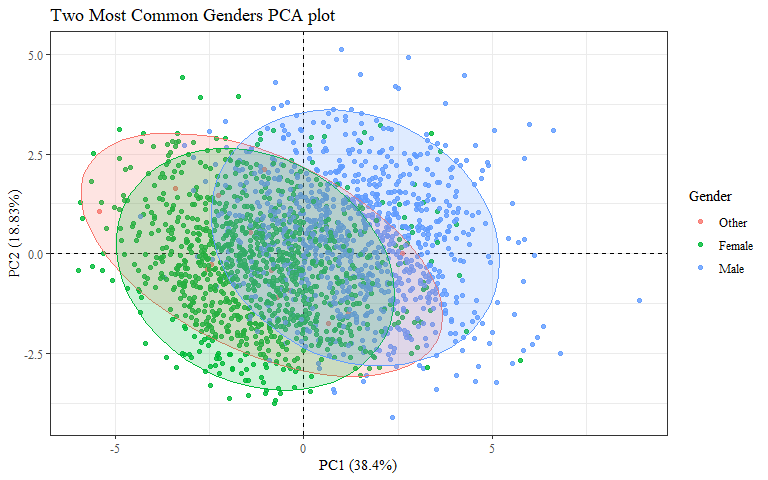
PCAdata\_gender <- imp\_selected2  
  
PCAdata\_gender$gender <-   
 recode\_factor(PCAdata\_gender$gender, 'Non-binary or Other'= "Other",  
 'Prefer not to say' = "Other")  
  
PCAdata\_gender$gender[is.na(PCAdata\_gender$gender)]="Other"  
  
PCAdata\_gender\_nas <- imp\_selected1 %>%   
 mutate(gender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

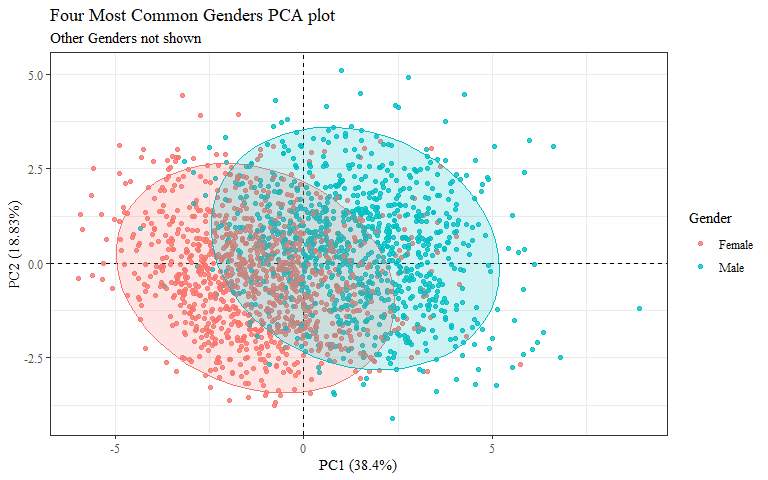
str(PCAdata\_gender\_nas$gender)

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

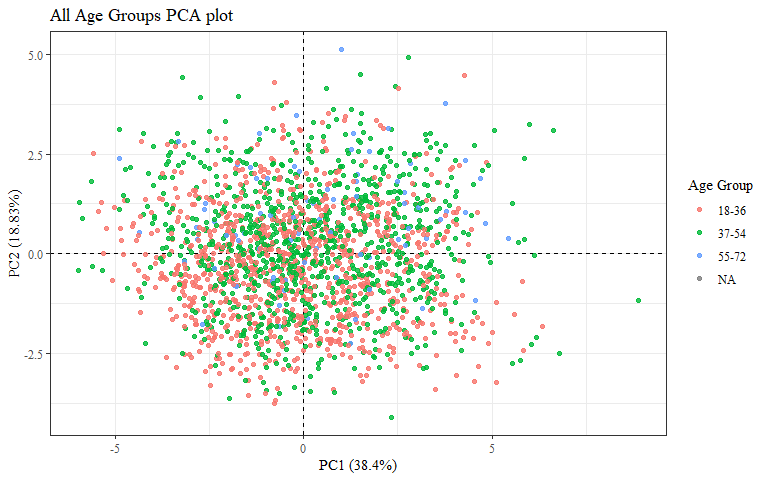
ggplot(data = PCAdata\_gender, aes(x = pc1, y = pc2, color = gender)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +   
 stat\_ellipse(geom="polygon", aes(fill = gender),   
 alpha = 0.2,   
 show.legend = FALSE,  
 level = 0.95) +  
 #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",  
 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",  
 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",  
 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",  
 color= "Age Group")+  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))

## Too few points to calculate an ellipse

