PCA2

2022-08-16

#run this with SA3\_noout data

library(FactoMineR)  
library(factoextra)

## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(tidyverse)

## ── Attaching packages  
## ───────────────────────────────────────  
## tidyverse 1.3.2 ──

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

library(MASS)

##   
## Attaching package: 'MASS'  
##   
## The following object is masked from 'package:dplyr':  
##   
## select

library(reshape2)

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

library(cowplot)  
library(flextable)

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

library(readxl)  
library(extrafont)

## Registering fonts with R

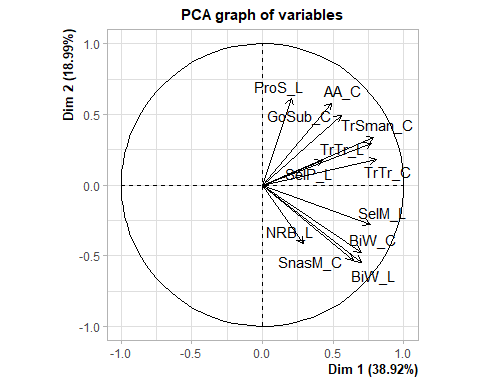
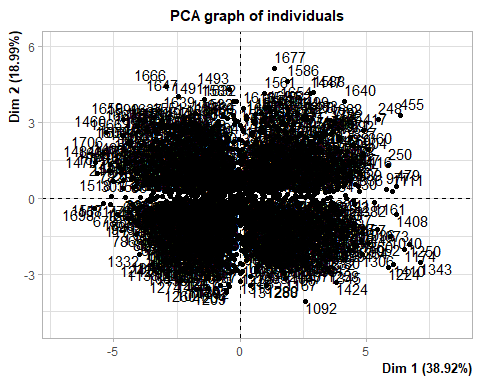
library(forcats)  
library(writexl)  
library(ggfortify)  
library(fauxnaif)

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

## tibble [1,707 × 15] (S3: tbl\_df/tbl/data.frame)  
## $ AA\_C : num [1:1707] 65 55 70 58 67 60 59 59 65 65 ...  
## $ BiW\_C : num [1:1707] 130 127 143 140 137 130 141 138 143 150 ...  
## $ BiW\_L : num [1:1707] 115 108 121 109 104 106 109 111 113 116 ...  
## $ GoSub\_C : num [1:1707] 93 93 115 93 103 100 79 106 85 102 ...  
## $ NRB\_L : num [1:1707] 17 18 19 21 19 14 17 18 16 17 ...  
## $ ProS\_L : num [1:1707] 17 18 14 13 20 20 18 12 24 22 ...  
## $ SelP\_L : num [1:1707] 42 41 51 44 47 48 46 41 46 44 ...  
## $ SelM\_L : num [1:1707] 122 99 130 115 119 126 117 112 117 117 ...  
## $ SnasM\_C : num [1:1707] 82 55 84 74 73 80 78 76 64 75 ...  
## $ TrSman\_C : num [1:1707] 177 145 178 147 157 164 149 159 151 160 ...  
## $ TrTr\_C : num [1:1707] 296 276 292 273 279 300 283 275 307 286 ...  
## $ TrTr\_L : num [1:1707] 155 141 156 149 146 146 147 151 157 144 ...  
## $ gender : chr [1:1707] "Male" "Female" "Male" "Male" ...  
## $ race\_eth : chr [1:1707] "Black" "white" "white" "white" ...  
## $ age\_group: chr [1:1707] "18-36" "37-54" "37-54" "18-36" ...

#https://tem11010.github.io/Plotting-PCAs/  
  
PCAdata\_full1 <- PCAdata\_full %>% 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(PCAdata\_full1))){  
   
 if (is.numeric(PCAdata\_full1[, i])==TRUE)  
   
 PCAdata\_full1[, i] <- as.numeric(scale(PCAdata\_full1[, i]))  
   
 else  
   
 PCAdata\_full1[, i] <- PCAdata\_full1[, i]  
   
}

pca1 <- PCA(PCAdata\_full1[ ,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)

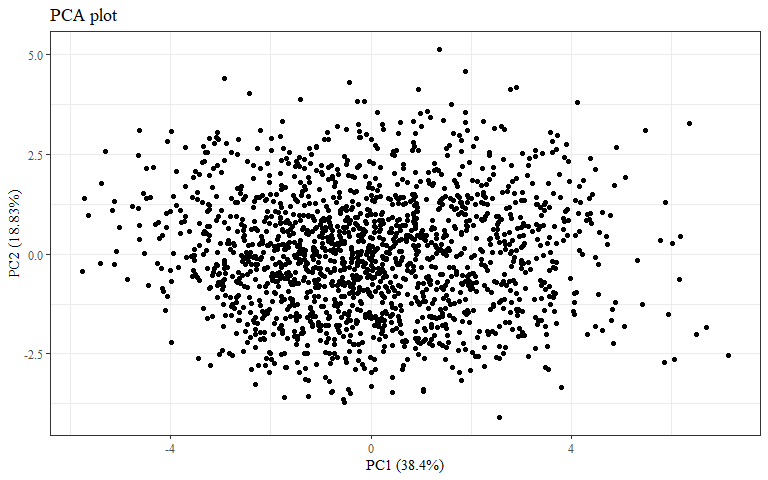
##   
## Call:  
## PCA(X = PCAdata\_full1[, c("AA\_C", "BiW\_C", "BiW\_L", "GoSub\_C",   
## "NRB\_L", "ProS\_L", "SelP\_L", "SelM\_L", "SnasM\_C", "TrSman\_C",   
## "TrTr\_C", "TrTr\_L")])   
##   
##   
## Eigenvalues  
## Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6 Dim.7  
## Variance 4.670 2.279 1.318 0.956 0.739 0.664 0.447  
## % of var. 38.921 18.989 10.985 7.970 6.161 5.533 3.723  
## Cumulative % of var. 38.921 57.909 68.895 76.865 83.026 88.559 92.282  
## Dim.8 Dim.9 Dim.10 Dim.11 Dim.12  
## Variance 0.359 0.264 0.146 0.089 0.068  
## % of var. 2.995 2.197 1.218 0.742 0.566  
## Cumulative % of var. 95.277 97.474 98.693 99.434 100.000  
##   
## Individuals (the 10 first)  
## Dist Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3 ctr  
## 1 | 2.901 | 1.855 0.043 0.409 | 0.132 0.000 0.002 | -0.928 0.038  
## 2 | 3.104 | -2.450 0.075 0.623 | 0.213 0.001 0.005 | -1.098 0.054  
## 3 | 4.489 | 3.624 0.165 0.652 | -0.173 0.001 0.001 | -0.013 0.000  
## 4 | 2.631 | -0.456 0.003 0.030 | -1.537 0.061 0.341 | -0.586 0.015  
## 5 | 1.581 | 0.526 0.003 0.111 | 0.835 0.018 0.279 | 0.643 0.018  
## 6 | 2.382 | 1.215 0.019 0.260 | 0.677 0.012 0.081 | 0.567 0.014  
## 7 | 1.651 | -0.072 0.000 0.002 | -0.877 0.020 0.282 | 0.771 0.026  
## 8 | 3.022 | -0.027 0.000 0.000 | -1.011 0.026 0.112 | -1.734 0.134  
## 9 | 3.498 | 1.389 0.024 0.158 | 1.336 0.046 0.146 | 0.895 0.036  
## 10 | 2.100 | 1.202 0.018 0.327 | 0.282 0.002 0.018 | 0.331 0.005  
## cos2   
## 1 0.102 |  
## 2 0.125 |  
## 3 0.000 |  
## 4 0.050 |  
## 5 0.166 |  
## 6 0.057 |  
## 7 0.218 |  
## 8 0.329 |  
## 9 0.066 |  
## 10 0.025 |  
##   
## Variables (the 10 first)  
## Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3 ctr cos2   
## AA\_C | 0.488 5.097 0.238 | 0.579 14.723 0.335 | 0.315 7.517 0.099 |  
## BiW\_C | 0.696 10.385 0.485 | -0.474 9.866 0.225 | 0.156 1.855 0.024 |  
## BiW\_L | 0.700 10.492 0.490 | -0.550 13.283 0.303 | -0.080 0.487 0.006 |  
## GoSub\_C | 0.556 6.615 0.309 | 0.493 10.663 0.243 | -0.457 15.877 0.209 |  
## NRB\_L | 0.287 1.767 0.083 | -0.417 7.637 0.174 | -0.237 4.270 0.056 |  
## ProS\_L | 0.205 0.901 0.042 | 0.613 16.503 0.376 | 0.303 6.946 0.092 |  
## SelP\_L | 0.421 3.796 0.177 | 0.171 1.281 0.029 | 0.729 40.334 0.532 |  
## SelM\_L | 0.763 12.468 0.582 | -0.280 3.432 0.078 | 0.328 8.155 0.108 |  
## SnasM\_C | 0.642 8.812 0.412 | -0.531 12.393 0.282 | 0.077 0.454 0.006 |  
## TrSman\_C | 0.785 13.196 0.616 | 0.339 5.043 0.115 | -0.352 9.398 0.124 |

PCAdata\_full1$pc1 <- pca1$ind$coord[, 1] # indexing the first column  
  
PCAdata\_full1$pc2 <- pca1$ind$coord[, 2] # indexing the second column

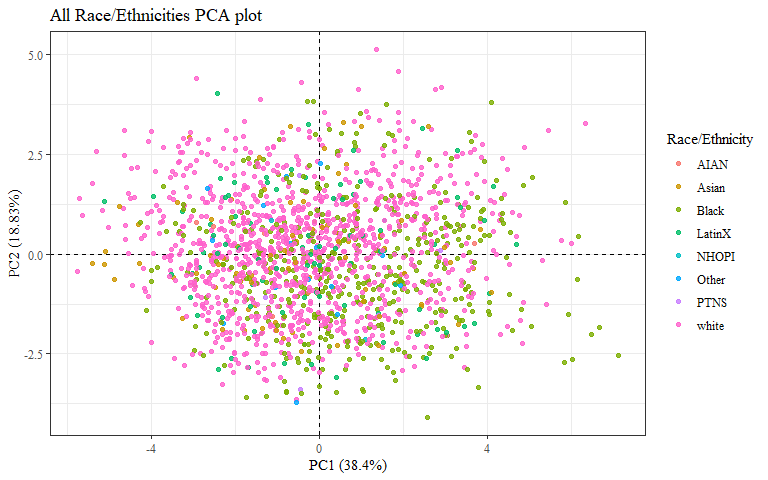
pca.vars <- pca1$var$coord %>% data.frame  
  
pca.vars$vars <- rownames(pca.vars)  
  
pca.vars.m <- melt(pca.vars, id.vars = "vars")

#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 = PCAdata\_full1, 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"))

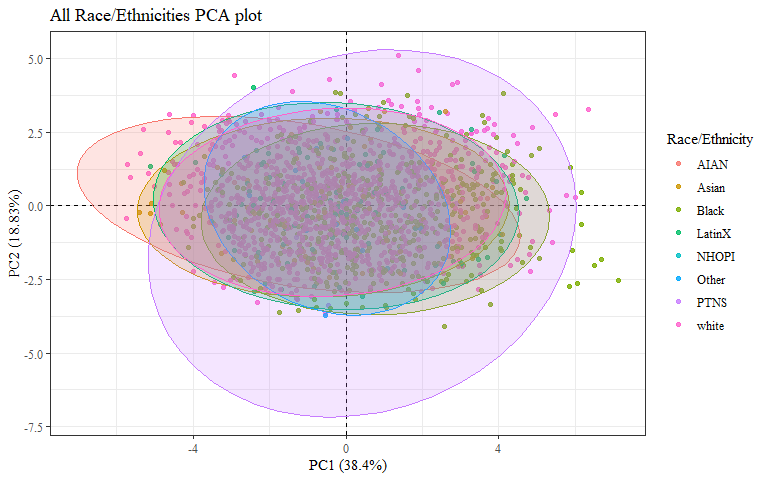


#checking to make sure this matches plot in PCA.Rmd  
ggplot(data = PCAdata\_full1, aes(x = pc1, y = pc2, color = race\_eth)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title = "All Race/Ethnicities PCA plot",  
 color = "Race/Ethnicity")+  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))



ggplot(data = PCAdata\_full1, 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"))

## Too few points to calculate an ellipse

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

PCAdata\_full2 <- PCAdata\_full1  
  
PCAdata\_full2$race\_eth <- as.factor(PCAdata\_full2$race\_eth)  
PCAdata\_full2$gender <- as.factor(PCAdata\_full2$gender)  
PCAdata\_full2$age\_group <- as.factor(PCAdata\_full2$age\_group)

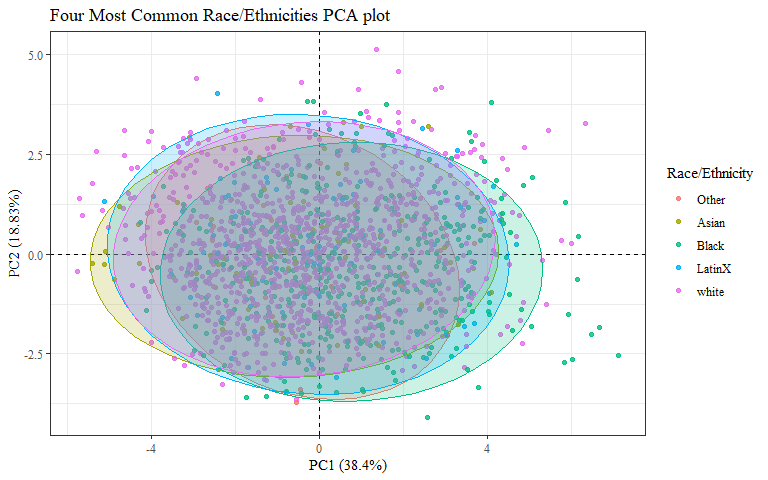
PCAdata\_race\_eth <- PCAdata\_full2  
  
PCAdata\_race\_eth$race\_eth <-   
 recode\_factor(PCAdata\_race\_eth$race\_eth, 'AIAN'= "Other",  
 'NHOPI' = "Other",  
 'PTNS' = "Other")  
  
PCAdata\_race\_eth\_nas <- PCAdata\_full1 %>%   
 mutate(race\_eth = na\_if\_in(race\_eth, c("AIAN", "Other", "NHOPI", "PTNS")))  
  
str(PCAdata\_race\_eth$race\_eth)

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

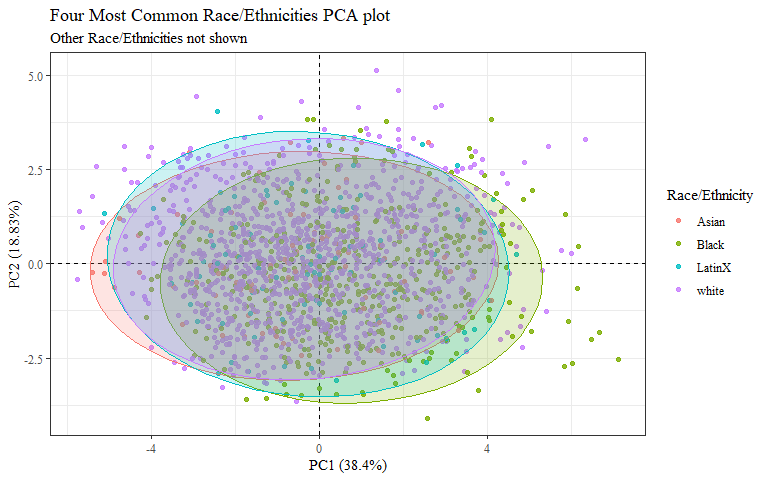
str(PCAdata\_race\_eth\_nas$race\_eth)

## chr [1:1707] "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)

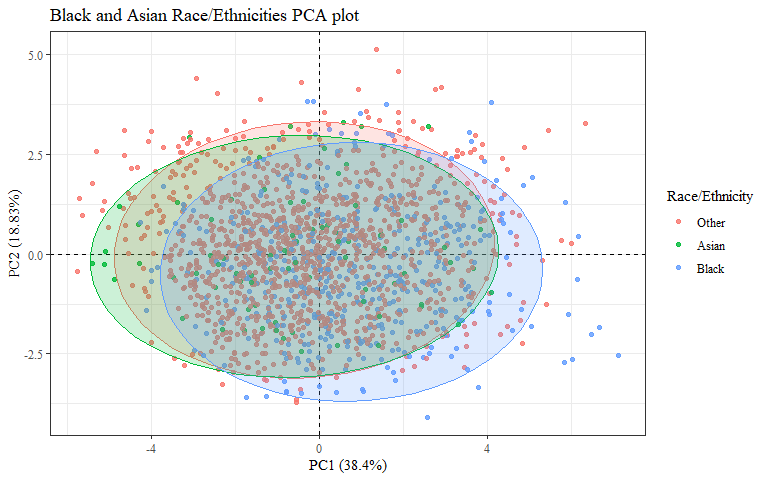
## Factor w/ 3 levels "Other","Asian",..: 3 1 1 1 1 1 3 3 1 1 ...

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

## chr [1:1707] "Black" NA NA NA NA NA "Black" "Black" NA NA NA NA NA NA ...

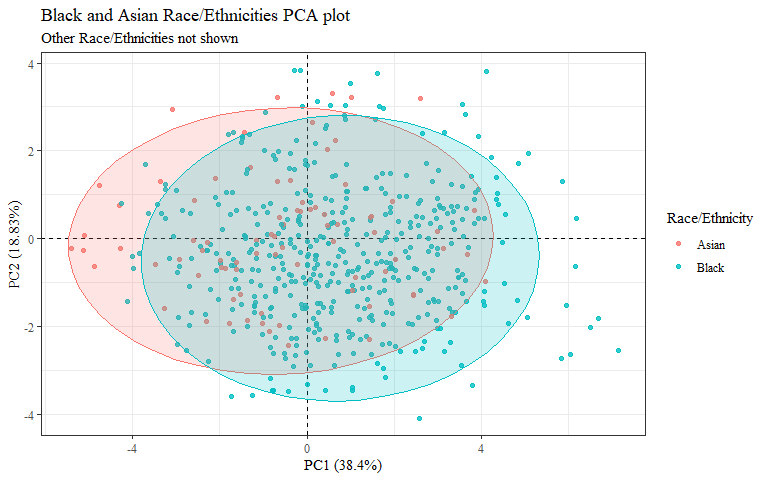
ggplot(data = PCAdata\_race\_eth1, aes(x = pc1, y = pc2, color = race\_eth)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +   
 stat\_ellipse(geom="polygon", aes(fill = race\_eth, na.rm=TRUE),   
 alpha = 0.2,   
 show.legend = FALSE,  
 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

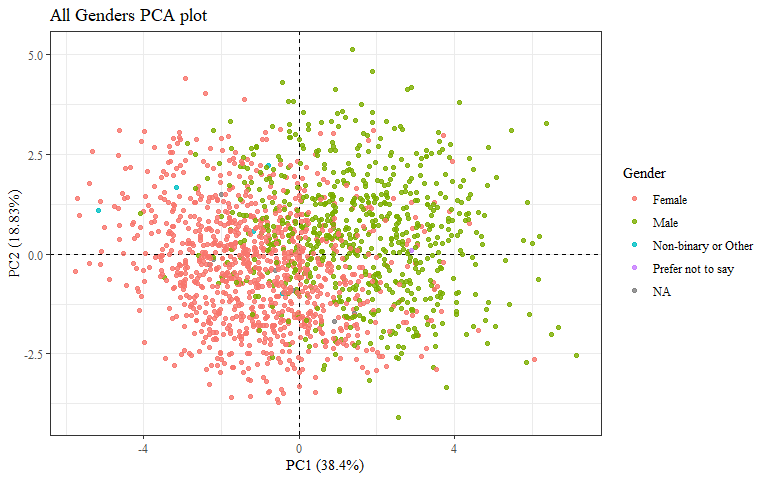


PCAdata\_race\_eth\_nas1 %>%   
 drop\_na(race\_eth) %>%   
 ggplot(aes(x = pc1, y = pc2, color = race\_eth)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +   
 stat\_ellipse(geom="polygon", aes(fill = race\_eth, na.rm=TRUE),   
 alpha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

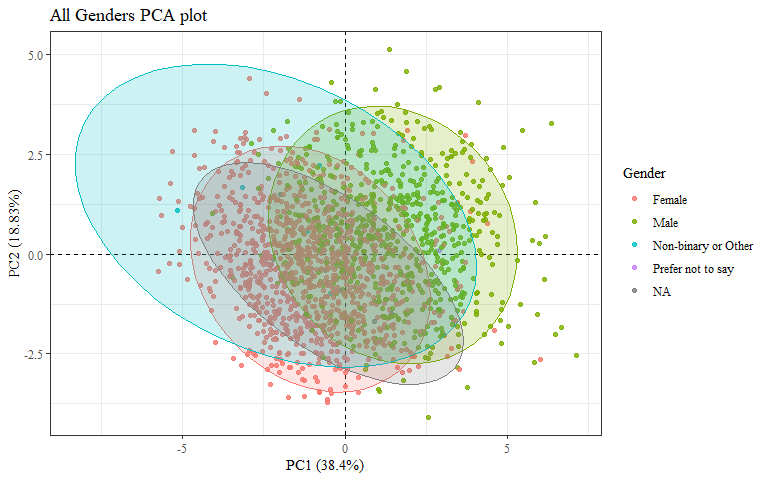


#checking to make sure this matches plot in PCA.Rmd  
ggplot(data = PCAdata\_full1, aes(x = pc1, y = pc2, color = gender)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title = "All Genders PCA plot",  
 color= "Gender")+  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))



ggplot(data = PCAdata\_full1, 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

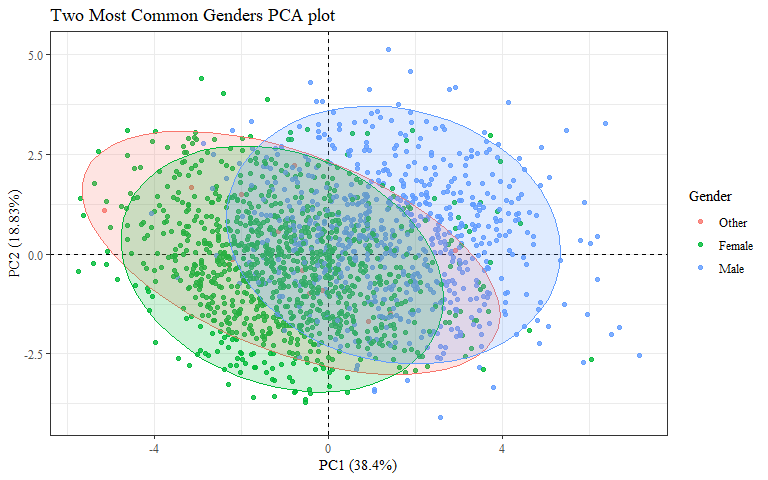
PCAdata\_gender <- PCAdata\_full2  
  
PCAdata\_gender$gender <-   
 recode\_factor(PCAdata\_gender$gender, 'Non-binary or Other'= "Other",  
 'Prefer not to say' = "Other")  
  
PCAdata\_gender$gender[is.na(PCAdata\_gender$gender)]="Other"  
  
PCAdata\_gender\_nas <- PCAdata\_full1 %>%   
 mutate(gender = 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 ...

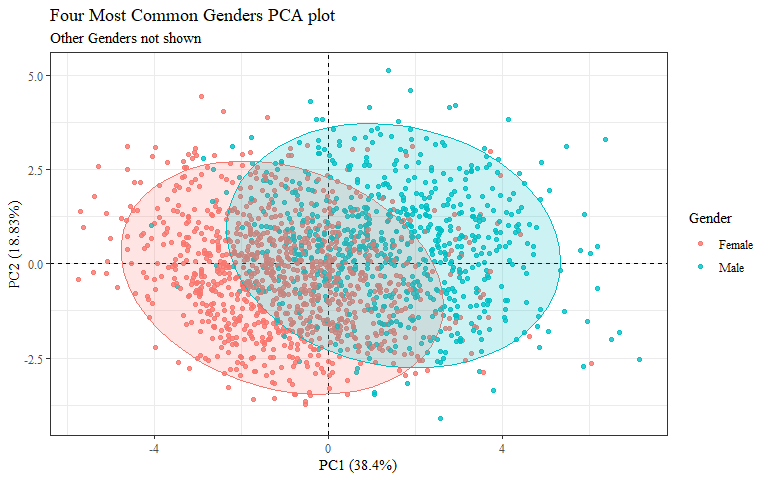
str(PCAdata\_gender\_nas$gender)

## chr [1:1707] "Male" "Female" "Male" "Male" "Male" "Male" "Male" "Male" ...

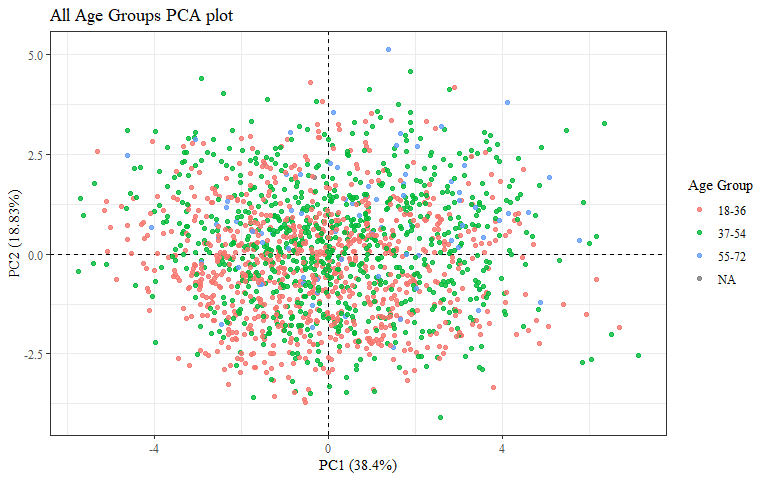
ggplot(data = PCAdata\_gender, aes(x = pc1, y = pc2, color = gender)) +  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 geom\_point(alpha = 0.8) +   
 stat\_ellipse(geom="polygon", aes(fill = gender),   
 alpha = 0.2,   
 show.legend = FALSE,  
 level = 0.95) +  
 #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 = PCAdata\_full1, 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 = PCAdata\_full1, 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

