PCA

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(forcats)  
library(writexl)  
library(ggfortify) #autoplot  
library(scales) #percent()

##   
## Attaching package: 'scales'  
##   
## The following object is masked from 'package:purrr':  
##   
## discard  
##   
## The following object is masked from 'package:readr':  
##   
## col\_factor

#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  
}

#remove identifiers and demographic data  
  
headscan\_full1<-read\_excel("C:\\Users\\19177\\OneDrive - Colostate\\Desktop\\Dissertation\\headscan\_dissertation\\headscan\_full1.xlsx")  
  
  
  
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] 65 55 70 58 67 60 59 59 65 65 ...  
## $ BGl\_C : num [1:2016] 315 289 293 313 288 306 320 NA 300 277 ...  
## $ 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 ...  
## $ ChCh\_C : num [1:2016] 62 64 68 70 70 70 67 69 67 63 ...  
## $ 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 ...  
## $ ProA\_L : num [1:2016] 28 25 31 23 28 28 26 27 32 28 ...  
## $ ProA\_C : num [1:2016] 31 27 33 27 31 29 27 29 34 31 ...  
## $ ProS\_C : num [1:2016] 18 20 14 13 22 22 19 14 26 24 ...  
## $ ProS\_L : num [1:2016] 17 18 14 13 20 20 18 12 24 22 ...  
## $ SelP\_C : num [1:2016] 42 41 51 45 47 48 46 42 47 44 ...  
## $ SelP\_L : num [1:2016] 42 41 51 44 47 48 46 41 46 44 ...  
## $ SelDH\_C : num [1:2016] 15 9 9 11 13 15 9 9 12 14 ...  
## $ 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 ...  
## $ SmanM\_C : num [1:2016] 59 51 45 43 33 34 55 37 61 41 ...  
## $ SmanM\_L : num [1:2016] 55 50 45 42 33 34 50 36 59 40 ...  
## $ SnasM\_L : num [1:2016] 75 53 78 69 67 76 69 71 62 69 ...  
## $ TrHO\_C : num [1:2016] 179 163 169 166 159 162 169 NA 167 166 ...  
## $ TrEJ\_C : num [1:2016] 40 32 39 29 46 42 29 32 29 33 ...  
## $ TrGo\_C : num [1:2016] 84 57 70 61 68 70 75 61 67 64 ...  
## $ TrSel\_C : num [1:2016] 149 138 150 133 140 151 140 138 156 143 ...  
## $ TrSman\_C : num [1:2016] 177 145 178 147 157 164 149 159 151 160 ...  
## $ TrSnas\_C : num [1:2016] 163 142 167 145 152 157 148 149 157 NA ...  
## $ 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 ...  
## $ 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" ...

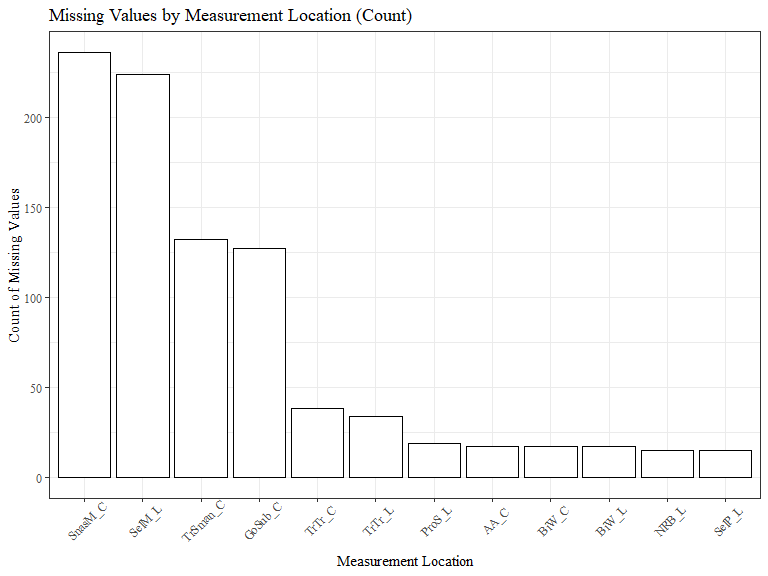
PCAdata\_num <- headscan\_full1[c(1,2,4,5,7,8,12,14,16,17,25,27,28)]  
  
PCAdata\_full <- headscan\_full1[c(1,2,4,5,7,8,12,14,16,17,25,27,28,31,32,33)]

measureNAs <- read\_excel("C:\\Users\\19177\\OneDrive - Colostate\\Desktop\\Dissertation\\headscan\_dissertation\\measureNAs.xlsx")  
  
PCA\_NAs <- slice(measureNAs, c(1,3,4,6,7,11,13,15,16,24,26,27))  
  
#Size 12 Table TNR  
flextable(PCA\_NAs) %>%  
 my\_ft\_theme()%>%   
 bold(part = "header") %>%   
 set\_caption("NA values for PCA Measurement Locations, total 891 NA values") %>%   
 autofit() %>%   
 set\_header\_labels(values = list(measure\_name = "Measurement Location",  
 measureNAprops = "Proportion of NA values",  
 measureNAsums = "Count of NA values"))

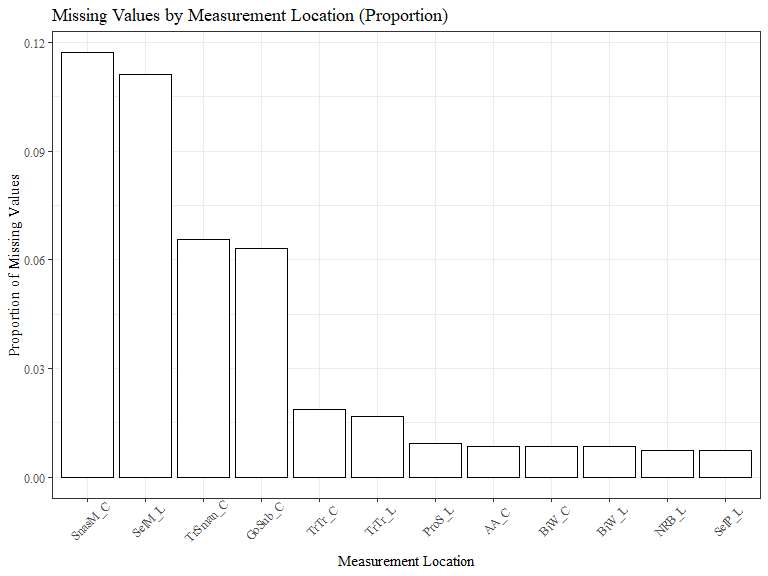
**Table** : NA values for PCA Measurement Locations, total 891 NA values

| **Measurement Location** | **Proportion of NA values** | **Count of NA values** |
| --- | --- | --- |
| AA\_C | 0.0084 | 17 |
| BiW\_C | 0.0084 | 17 |
| BiW\_L | 0.0084 | 17 |
| GoSub\_C | 0.0630 | 127 |
| NRB\_L | 0.0074 | 15 |
| ProS\_L | 0.0094 | 19 |
| SelP\_L | 0.0074 | 15 |
| SelM\_L | 0.1111 | 224 |
| SnasM\_C | 0.1171 | 236 |
| TrSman\_C | 0.0655 | 132 |
| TrTr\_C | 0.0188 | 38 |
| TrTr\_L | 0.0169 | 34 |

PCA\_NAs$measure\_name <- fct\_reorder(PCA\_NAs$measure\_name, PCA\_NAs$measureNAsums, .desc=TRUE)  
  
#bar chart with counts  
ggplot(data=PCA\_NAs, aes(x=measure\_name, y=measureNAsums))+  
 geom\_bar(stat= "identity", color= "black", fill = "white")+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 theme(axis.text.x = element\_text(angle = 45, vjust=0.7))+  
 labs(title="Missing Values by Measurement Location (Count)",  
 y="Count of Missing Values",  
 x="Measurement Location")



#bar chart with proportions  
ggplot(data=PCA\_NAs, aes(x=measure\_name, y=measureNAprops))+  
 geom\_bar(stat= "identity", color= "black", fill = "white")+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 theme(axis.text.x = element\_text(angle = 45, vjust=0.7))+  
 labs(title="Missing Values by Measurement Location (Proportion)",  
 y="Proportion of Missing Values",  
 x="Measurement Location")



PCAdata\_num <- PCAdata\_num %>% drop\_na()  
  
str(PCAdata\_num)

## tibble [1,707 × 13] (S3: tbl\_df/tbl/data.frame)  
## $ ID : chr [1:1707] "400-20201012-002" "400-20201012-003" "400-20201012-004" "400-20201012-005" ...  
## $ 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 ...

PCAdata\_full <- left\_join(PCAdata\_num, PCAdata\_full, by= "ID")  
  
PCAdata\_full <- PCAdata\_full %>%   
 rename(AA\_C = "AA\_C.x",  
 BiW\_C = "BiW\_C.x",  
 BiW\_L = "BiW\_L.x",  
 GoSub\_C = "GoSub\_C.x",  
 NRB\_L = "NRB\_L.x",  
 ProS\_L = "ProS\_L.x",  
 SelP\_L = "SelP\_L.x",  
 SelM\_L = "SelM\_L.x",  
 SnasM\_C = "SnasM\_C.x",  
 TrSman\_C = "TrSman\_C.x",  
 TrTr\_C = "TrTr\_C.x",  
 TrTr\_L = "TrTr\_L.x")  
  
PCAdata\_full <- PCAdata\_full[-c(14:25)]

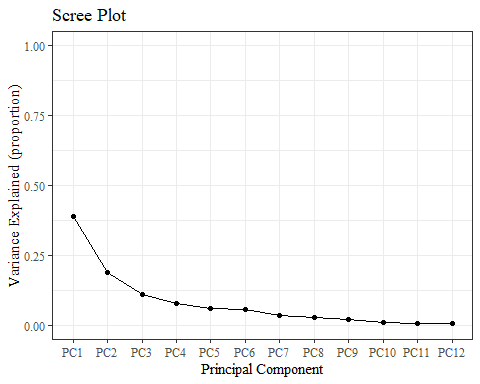
PCAdata\_num <- PCAdata\_num[-c(1)]  
PCAdata\_full <- PCAdata\_full[-c(1)]

#https://www.statology.org/principal-components-analysis-in-r/  
  
#calculate principal components  
pca\_res <- prcomp(PCAdata\_num, scale=TRUE)  
  
#reverse the signs (R calculates eigenvectors in negative direction)  
pca\_res$rotation <- -1\*pca\_res$rotation  
  
#display PCs  
pca\_res$rotation

## PC1 PC2 PC3 PC4 PC5 PC6  
## AA\_C 0.22576070 -0.3837106 -0.27416365 -0.02282620 0.3642372 -0.14431296  
## BiW\_C 0.32225991 0.3141000 -0.13621038 0.33864508 0.2261677 0.21092242  
## BiW\_L 0.32391471 0.3644642 0.06976369 0.30325502 0.1958235 0.18065673  
## GoSub\_C 0.25719890 -0.3265418 0.39845599 -0.15172254 -0.3534331 0.08878022  
## NRB\_L 0.13294498 0.2763527 0.20662977 -0.61857373 0.4867618 -0.42838253  
## ProS\_L 0.09494484 -0.4062407 -0.26355561 -0.30257756 0.3514326 0.61448874  
## SelP\_L 0.19482656 -0.1131699 -0.63508917 0.04362454 -0.2488195 -0.44535392  
## SelM\_L 0.35309926 0.1852597 -0.28557238 -0.30025346 -0.3460723 0.05858366  
## SnasM\_C 0.29684583 0.3520322 -0.06741584 -0.31198996 -0.2306142 0.26485559  
## TrSman\_C 0.36326916 -0.2245596 0.30656575 -0.06771101 -0.1829567 -0.01224794  
## TrTr\_C 0.37202099 -0.1186274 0.14897626 0.19845266 0.1167115 -0.19689280  
## TrTr\_L 0.35538955 -0.1941301 0.15772836 0.24953415 0.1025714 -0.14906804  
## PC7 PC8 PC9 PC10 PC11  
## AA\_C 0.3912006 -0.63598058 0.029571191 -0.012017633 0.03041627  
## BiW\_C -0.2997093 -0.21302780 -0.003846428 -0.002119206 0.36336495  
## BiW\_L -0.2076260 -0.10064001 0.014213192 -0.093935191 -0.34572851  
## GoSub\_C -0.3020771 -0.28097674 0.039178606 -0.581404530 0.02466210  
## NRB\_L -0.2214636 0.04456400 -0.056141929 -0.060516481 0.01722892  
## ProS\_L -0.2102236 0.33451185 0.014645784 -0.007291929 -0.08576819  
## SelP\_L -0.3905179 0.09749183 -0.018400597 -0.066343117 -0.33173542  
## SelM\_L 0.1282476 0.11077120 -0.002591739 0.029365834 0.61458518  
## SnasM\_C 0.4472883 -0.06905681 -0.046185954 -0.046089088 -0.49114562  
## TrSman\_C -0.2079414 -0.11803227 0.001587085 0.786710706 -0.09460723  
## TrTr\_C 0.2506982 0.41622676 0.692854637 -0.096576867 0.00724528  
## TrTr\_L 0.2333014 0.36533176 -0.715177494 -0.116841046 0.03550978  
## PC12  
## AA\_C -0.11473161  
## BiW\_C 0.54895905  
## BiW\_L -0.64237044  
## GoSub\_C 0.04756513  
## NRB\_L 0.04509741  
## ProS\_L -0.01104097  
## SelP\_L 0.08333821  
## SelM\_L -0.37255367  
## SnasM\_C 0.33529815  
## TrSman\_C 0.01949312  
## TrTr\_C 0.08967036  
## TrTr\_L 0.04284398

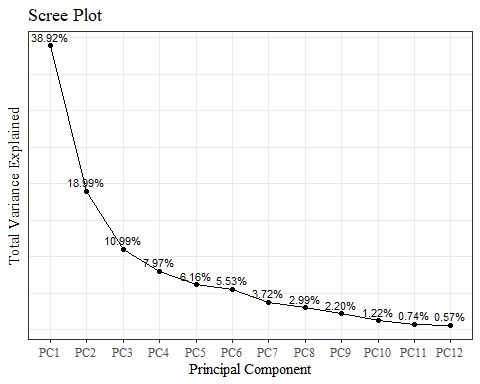
var\_explained = pca\_res$sdev^2 / sum(pca\_res$sdev^2)  
  
var\_explained\_total <- var\_explained

qplot(c(1:12), var\_explained) +   
 geom\_line() +   
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))+  
 xlab("Principal Component") +   
 ylab("Variance Explained (proportion)") +  
 ggtitle("Scree Plot") +  
 ylim(0, 1) +  
 scale\_x\_discrete(limits=c("PC1","PC2","PC3","PC4","PC5","PC6","PC7","PC8","PC9","PC10","PC11","PC12"))



var\_explained\_data <- data.frame(var\_explained)  
  
var\_explained\_data <- var\_explained\_data %>%   
 rename(v\_e = "var\_explained")  
  
  
var\_explained\_data$vep <- var\_explained\_data$v\_e  
  
var\_explained\_data <- rownames\_to\_column(var\_explained\_data, "PC\_num")  
  
var\_explained\_data$vep <- percent(var\_explained\_data$vep, accuracy=0.01)  
  
var\_explained\_data$PC\_num <- as.factor(var\_explained\_data$PC\_num)  
  
var\_explained\_data$PC\_num <-   
 recode\_factor(var\_explained\_data$PC\_num,   
 '1'= "PC1",  
 '2'= "PC2",  
 '3'= "PC3",  
 '4'= "PC4",  
 '5'= "PC5",  
 '6'= "PC6",  
 '7'= "PC7",  
 '8'= "PC8",  
 '9'= "PC9",  
 '10'= "PC10",  
 '11'= "PC11",  
 '12'= "PC12")

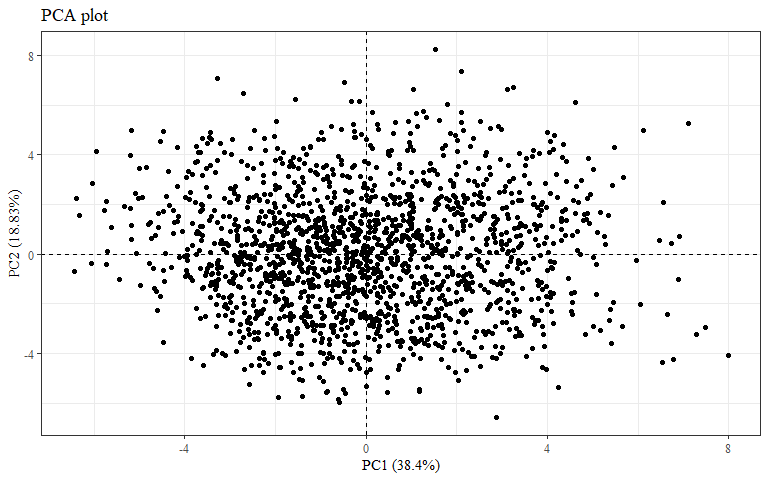
ggplot(data=var\_explained\_data, aes(x=PC\_num, y=v\_e, group=1)) +   
 geom\_line() +  
 geom\_point() +   
 geom\_text(aes(label=vep),  
 position= position\_dodge(0.9),  
 vjust = -0.5,   
 size = 3)+  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))+  
 theme(axis.text.y=element\_blank(),  
 axis.ticks.y=element\_blank()   
 )+  
 labs(title="Scree Plot",  
 x="Principal Component",  
 y="Total Variance Explained")



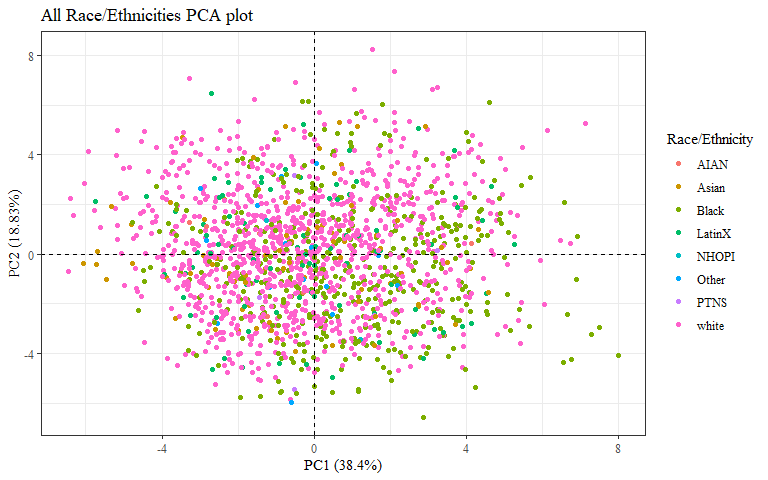
#making plot axis match the ggplot versions in PCA2.Rmd!!!   
#compare to PCAdata\_full1 pc1 and pc2 columns (should be same sign, here times 100)  
pca\_res$x <- pca\_res$x \* 100  
pca\_res$x[,1] <- pca\_res$x[,1] \* -1  
#pca\_res$x <- pca\_res$x[,2] \* 100  
  
#pca\_res$x

#<https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_pca.html>

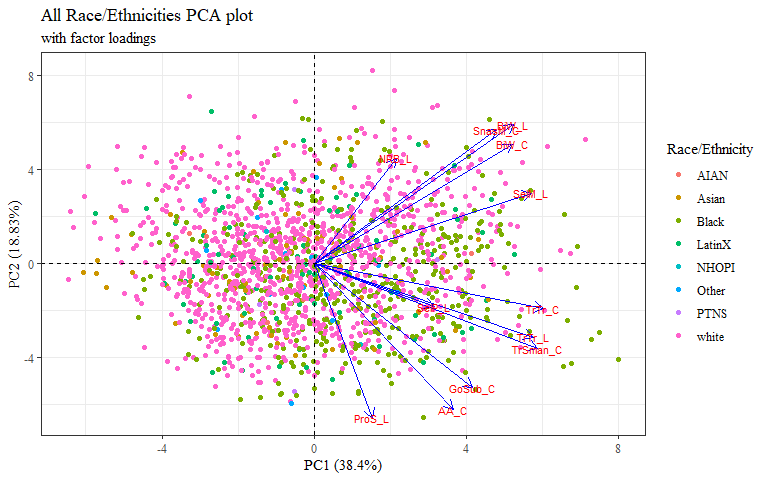
autoplot(pca\_res, data=PCAdata\_full)+  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))+   
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title="PCA plot")



autoplot(pca\_res, data=PCAdata\_full, colour="race\_eth")+  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman")) +  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title = "All Race/Ethnicities PCA plot",  
 color = "Race/Ethnicity")

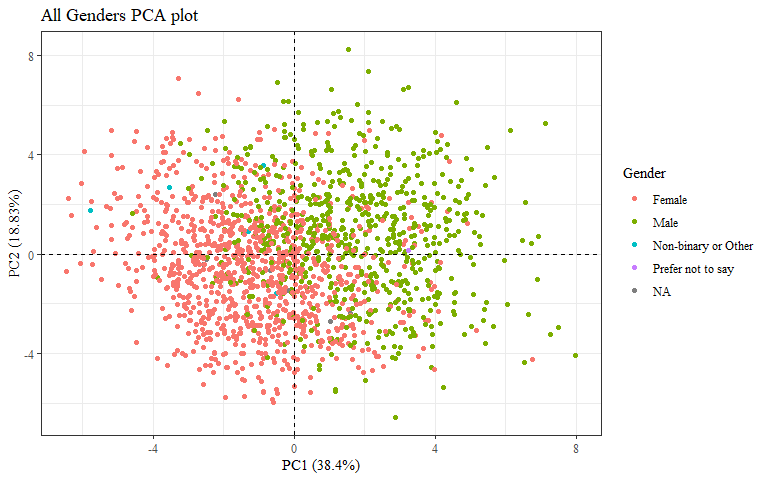


autoplot(pca\_res, data=PCAdata\_full, colour="race\_eth",  
 loadings = TRUE, loadings.colour = 'blue',  
 loadings.label = TRUE, loadings.label.size = 3)+  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))+  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title = "All Race/Ethnicities PCA plot",  
 subtitle = "with factor loadings",  
 color = "Race/Ethnicity")

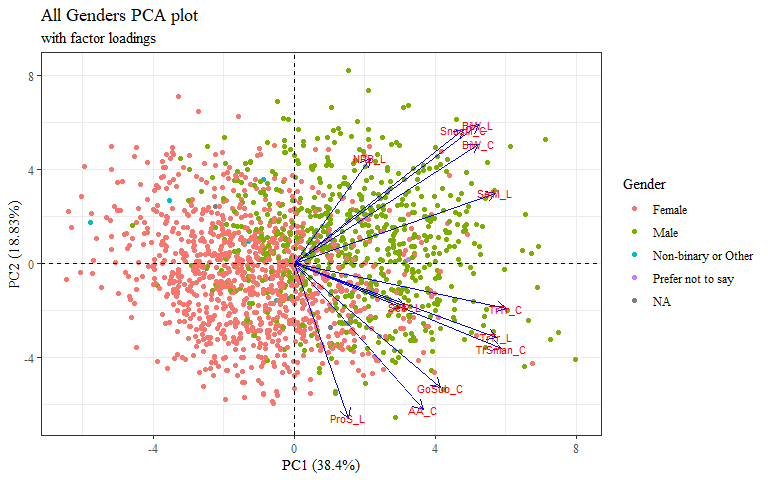
 arrows close together=higher correlation

COULD add ellipses (like in PCA2) to this graph with factor loadings.. but cannot add loadings to plot in PCA2 with ellipses

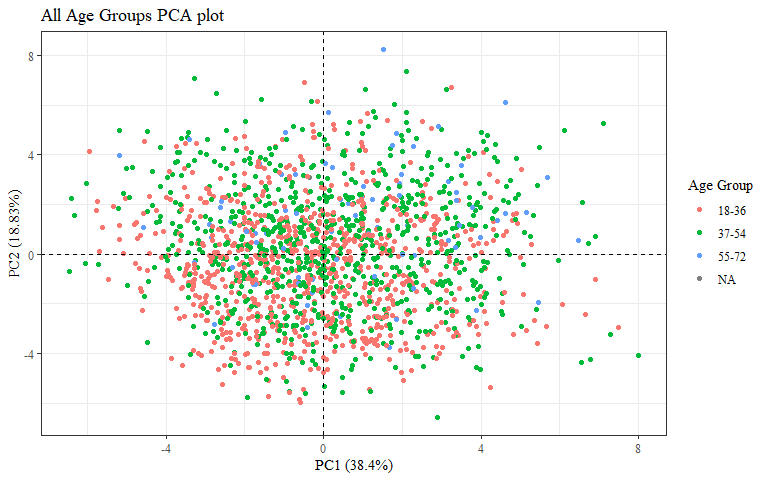
autoplot(pca\_res, data=PCAdata\_full, colour="gender")+  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))+  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title = "All Genders PCA plot",  
 color = "Gender")



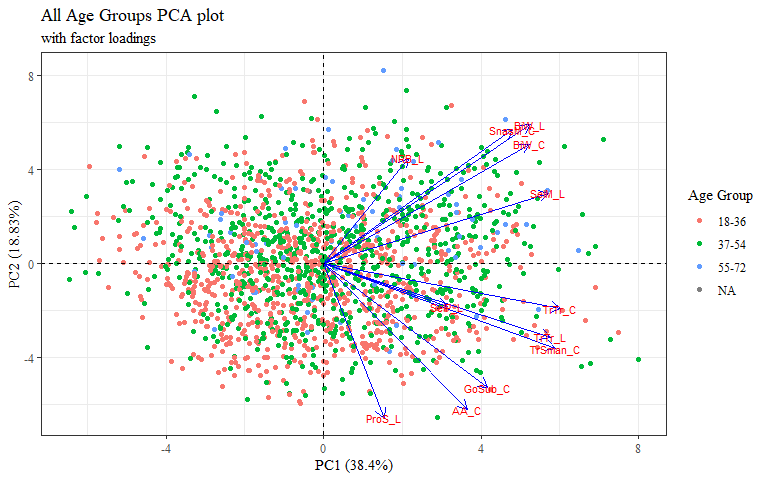
autoplot(pca\_res, data=PCAdata\_full, colour="gender",  
 loadings = TRUE, loadings.colour = 'blue',  
 loadings.label = TRUE, loadings.label.size = 3)+  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))+  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title = "All Genders PCA plot",  
 subtitle = "with factor loadings",  
 color = "Gender")



autoplot(pca\_res, data=PCAdata\_full, colour="age\_group")+  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))+  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title = "All Age Groups PCA plot",  
 color = "Age Group")



autoplot(pca\_res, data=PCAdata\_full, colour="age\_group",  
 loadings = TRUE, loadings.colour = 'blue',  
 loadings.label = TRUE, loadings.label.size = 3)+  
 geom\_hline(yintercept = 0, lty = 2) +  
 geom\_vline(xintercept = 0, lty = 2) +  
 theme\_bw() + theme(text=element\_text(family= "Times New Roman"))+  
 labs(y="PC2 (18.83%)",  
 x="PC1 (38.4%)",  
 title = "All Age Groups PCA plot",  
 subtitle = "with factor loadings",  
 color = "Age Group")



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