Discovering Emerging Patterns in Clinical Research Data Using R

ABSTRACT

The purpose of our study is to create meaningful visualizations of de-identified data from screened participants using R. Base R can be utilized to build a graphics dashboard integrated into the Human Research Information System (HuRIS), the NIDA-IRP electronic health and records research system, that contains 20 years of clinical research data. The aim is to provide real-time resources to investigators, study physicians, counselors, and clinical research staff through relevant, comprehensive imagery. By providing users with visualization tools, the idea is to discover emerging patterns within the data that can bolster the efficiency of the screening process. Clinical trials for substance use disorders employ a broad outreach to find participants and receive hundreds of applications with very few people actually meeting the qualifications for a study. Geo-mapping previously qualified participants may provide insights to narrow the scope during outreach to participants and determine the best areas to target future clinical trial participants.

Discovering Emerging Patterns in Clinical Research Data with R

Leslie Cook BS, Jia-Ling Lin PhD, Mustapha Mezghanni MS

Biomedical Informatics Section, Intramural Research Program, National Institute on Drug Abuse, National Institutes of Health, Baltimore, MD

National Institute on Drug Abuse

Goals

Integrate new features into the NIDA-IRP electronic medical and research records systems to:

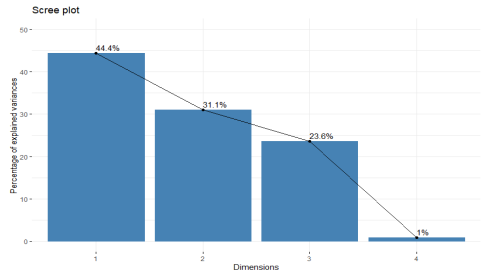
- ❑ explore and analyze clinical research data
- ❑ unsupervised learning methods, like Principal Component Analysis
- ❑ utilize R's powerful visualization libraries

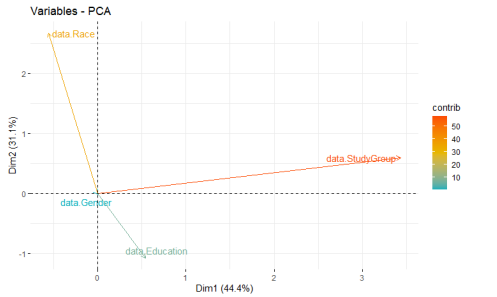
R Packages & Libraries

- ❑ ggplot2 : intuitive visualization library
- ❑ dplyr : data manipulation and transformation
- ❑ lubridate : date-time functions
- ❑ zipcodeR : U.S. Zip Codes functions
- ❑ maps : longitude and latitude data mapping
- ❑ factomineR & factoextra : functions for multivariate exploratory data analysis

Unsupervised Learning

Principal Component Analyses (PCA)

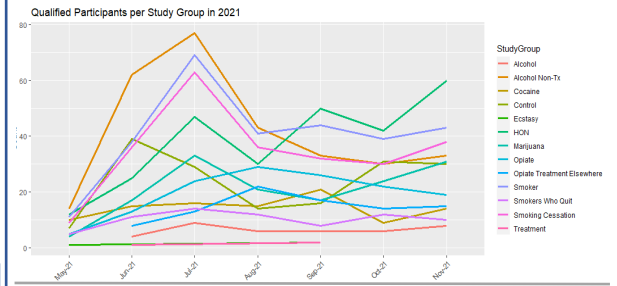


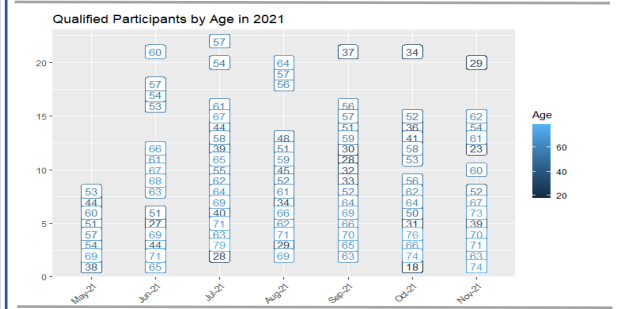


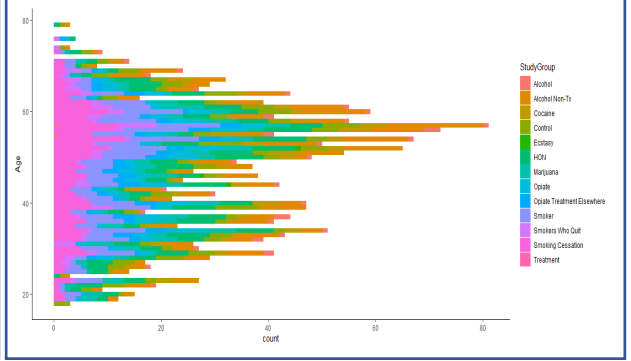
Dataset Preview

Age	Gender	Race	Education	Zip	ScreeningDate	StudyGroup
60	Female	B/AA - Black or African American	H.S. Diploma	21213	2021-05-11	Smoking Cessation
48	Male	W - White	Some College	21204	2021-05-13	Alcohol Non-Tx
35	Male	W - White	Some College	21817	2021-05-13	Alcohol Non-Tx
26	Female	W - White	Some High School	21817	2021-05-13	Cocaine
35	Male	W - White	Some College	21817	2021-05-13	Cocaine
48	Male	W - White	Some College	21204	2021-05-13	HON
26	Female	W - White	Some High School	21817	2021-05-13	HON
35	Male	W - White	Some College	21817	2021-05-13	HON
35	Male	W - White	Some College	21817	2021-05-13	Marijuana
26	Female	W - White	Some High School	21817	2021-05-13	Opiate
35	Male	W - White	Some College	21817	2021-05-13	Opiate

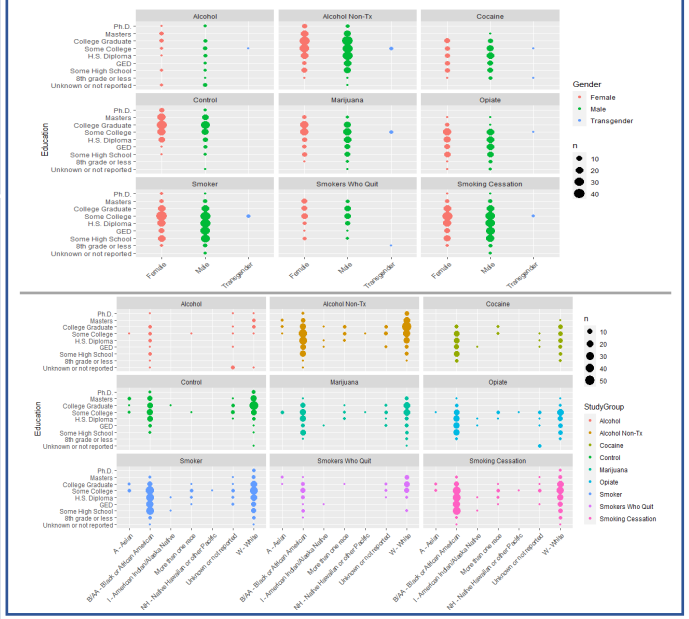
Visualizations



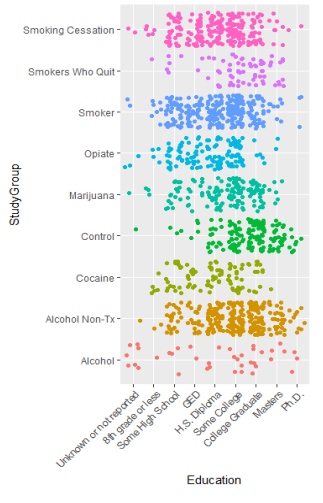


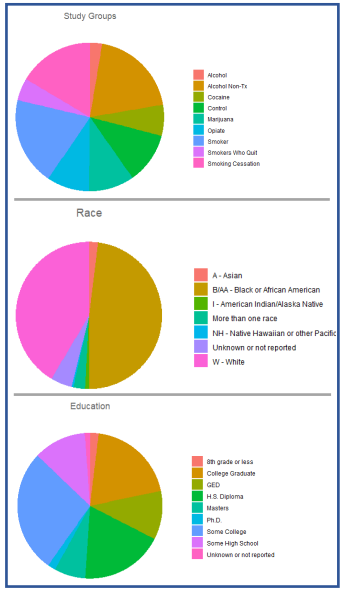


Visualizations



Visualizations

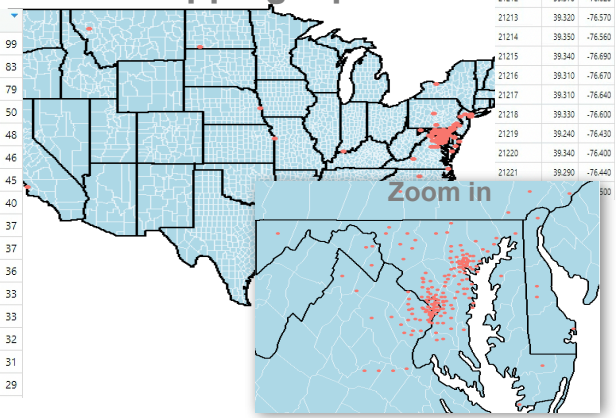


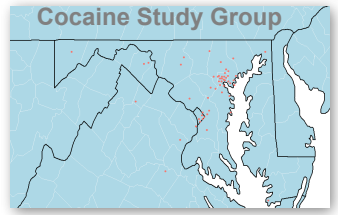


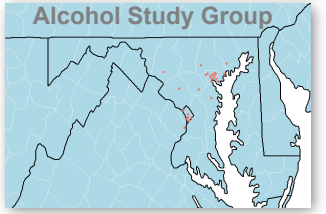
Geo-Mapping Zip Codes

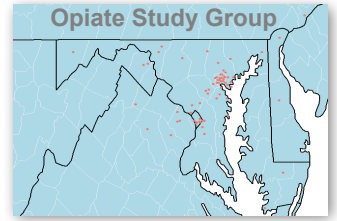
Zip	n
21213	99
21215	83
21218	79
21206	50
21217	48
21216	46
21224	45
21202	40
21222	37
21223	37
21201	36
21207	33
21234	33
21205	32
21229	31
21221	29

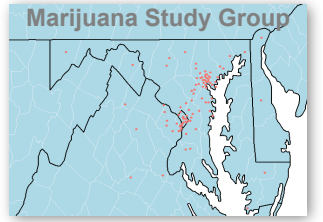
zipcode	lat	lng
21212	39.370	-76.620
21213	39.320	-76.570
21214	39.350	-76.560
21215	39.340	-76.690
21216	39.310	-76.670
21217	39.310	-76.640
21218	39.330	-76.600
21219	39.240	-76.430
21220	39.340	-76.400
21221	39.290	-76.440











Conclusion

This study demonstrates the feasibility of discovering emerging patterns in clinical research data through meaningful visualizations and analyses with R. Future directions are to:

- ❑ Build a visual and analysis library for investigators
- ❑ Integrate analysis tools into NIDA research systems
- ❑ Automate Interface between electronic medical and research records systems and analytical tools

Dimensionality Reduction of a Dataset using the Built-In R Function : prcomp()

Leslie Cook BS, Jia-Ling Lin PhD, Mustapha Mezghanni MS

Biomedical Informatics Section, Intramural Research Program, National Institute on Drug Abuse, National Institutes of Health, Baltimore, MD

National Institute on Drug Abuse

Principle Component Analyses

- Technique for preprocessing and dimensionality reduction on large datasets
- Extracts core features of the data without the need for human intervention
- First step for reshaping and reducing the dimensions of a dataset for machine learning and prediction models.
- Computation of linear combinations returns an ordered rank of the variances, with the most captured in the first PC

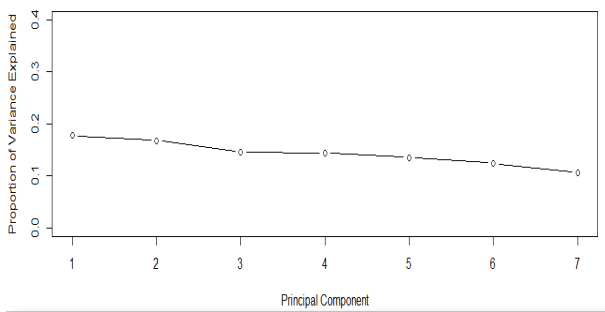
Program to compute PCA

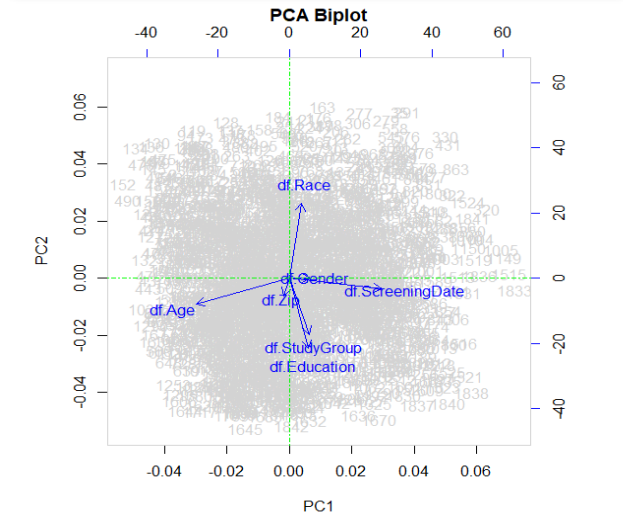
```
##############################################
pr.factor_df <- prcomp(factor_df,
                       scale = FALSE,
                       center = TRUE,
                       retx = T)
##############################################
########### calculate the variance #############
pr.var <- pr.factor_df$sdev^2
########## calculate the "least squares" #######
pve <- pr.var / sum(pr.var)
########### create the scree plot ##############
plot(pve, xlab = "Principal Component",
     ylab = "Proportion of Variance Explained",
     ylim = c(0, .6), type = "b",
     main = "Scree Plot")
abline(h=0.44, col="green", lty=4)
abline(v=1.00, col="green", lty=4)
##############################################
############ calculate the mean ################
colMeans(factor_df)
####### calculate the standard deviation #######
apply(factor_df, 2, sd)
###############  Scaled PCA model ##############
pr.with.scaling <- prcomp(factor_df,
                          scale = T,
                          center = T)
############# create the biplot ################
biplot(pr.with.scaling,
       col=c("light gray", "blue"),
       expand = .7, main = "PCA Biplot")
abline(h=0, col="green", lty=4)
abline(v=0, col="green", lty=4)
##############################################
```

PCA – 7 variables

```
##############################################
factor_df <- data.frame(df$Gender, df$Race,
                        df$Education, df$StudyGroup,
                        df$Age,df$Zip, df$ScreeningDate)
##############################################
```

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5    PC6    PC7
Standard deviation     1.1145 1.0851 1.0085 1.0027 0.9728 0.9331 0.8608
Proportion of Variance 0.1775 0.1682 0.1453 0.1436 0.1352 0.1244 0.1058
Cumulative Proportion  0.1775 0.3457 0.4909 0.6346 0.7698 0.8942 1.0000
```

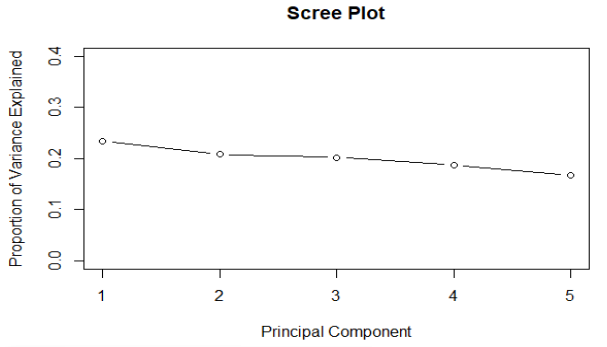


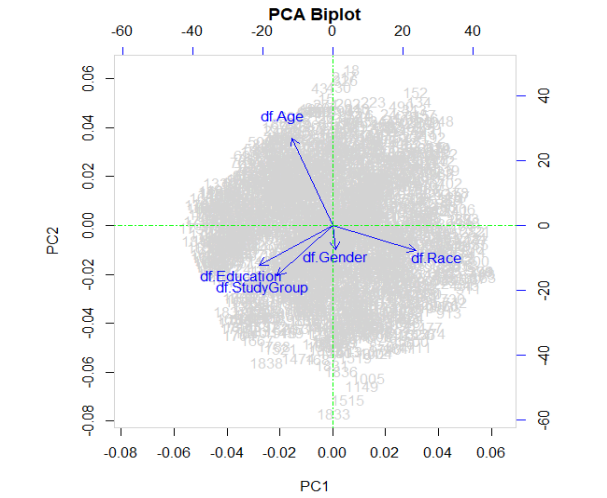


PCA – 5 variables

```
##############################################
factor_df <- data.frame(df$Gender, df$Race,
                        df$Education,
                        df$StudyGroup, df$Age)
##############################################
```

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5
Standard deviation     1.0837 1.0207 1.0052 0.9677 0.9148
Proportion of Variance 0.2349 0.2084 0.2021 0.1873 0.1674
Cumulative Proportion  0.2349 0.4432 0.6453 0.8326 1.0000
```

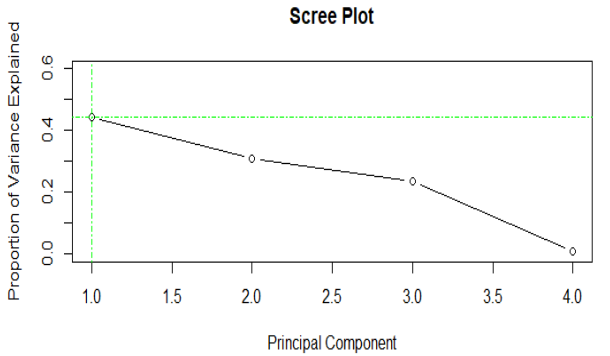


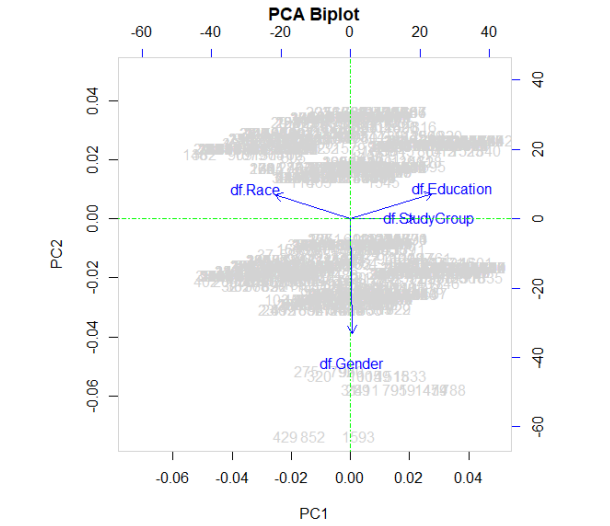


PCA – 4 variables

```
##############################################
factor_df <- data.frame(df$Gender,
                        df$Race,
                        df$Education,
                        df$StudyGroup)
##############################################
```

```
Importance of components:
                          PC1    PC2    PC3     PC4
Standard deviation     3.5202 2.9457 2.5694 0.51551
Proportion of Variance 0.4436 0.3106 0.2363 0.00951
Cumulative Proportion  0.4436 0.7542 0.9905 1.00000
```





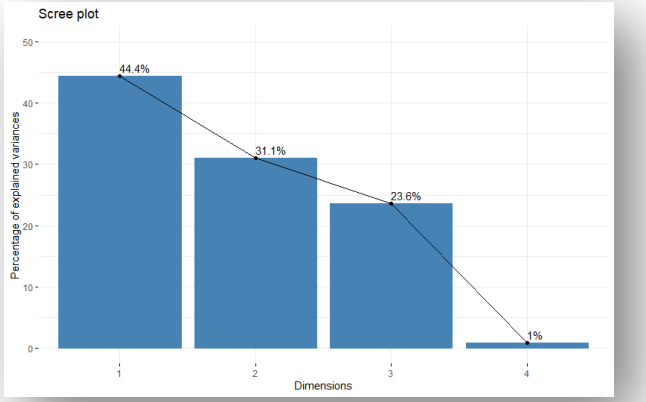
Advanced Graphics of PCA with FactoMineR & factoextra Library in R

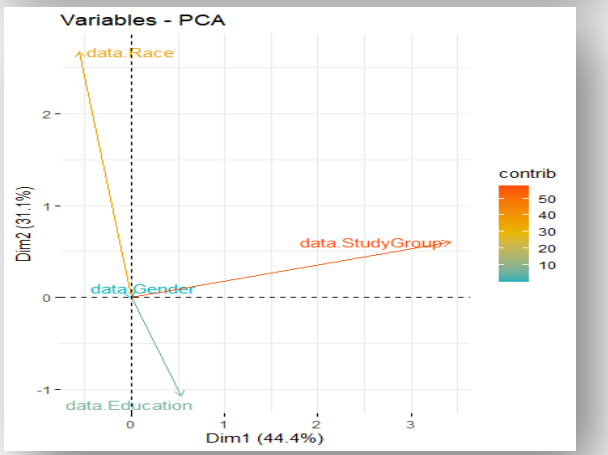
Leslie Cook BS, Jia-Ling Lin PhD, Mustapha Mezghanni MS

Biomedical Informatics Section, Intramural Research Program, National Institute on Drug Abuse, National Institutes of Health, Baltimore, MD

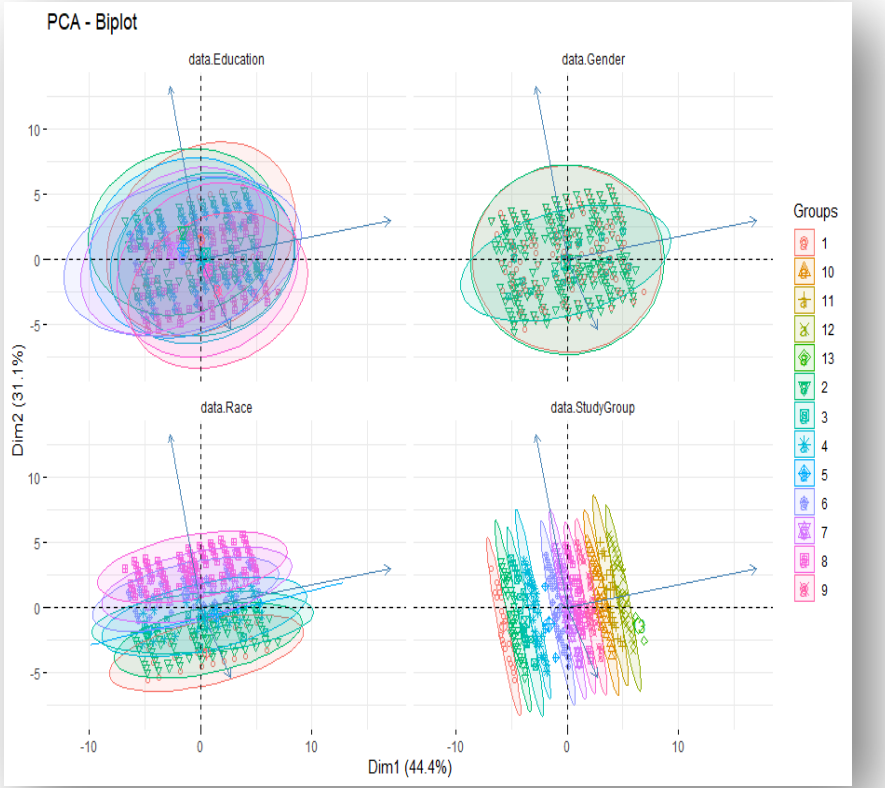
National Institute on Drug Abuse

```
#####################################################
df <- data.frame(data$Gender,data$Race,
                 data$Education, data$StudyGroup)
################ compute the PCA ####################
res.pca <- PCA(df,  graph = FALSE, scale = FALSE)
############# create the scree plot #################
fviz_screeplot(res.pca,
               addlabels = TRUE,
               ylim = c(0, 50))
#####################################################
### biplot to color variables base on contribution ###
fviz_pca_var(res.pca, col.var="contrib",
     gradient.cols = c("#00AFBB","#E7B800","#FC4E07"),
     repel = TRUE)
#####################################################
```

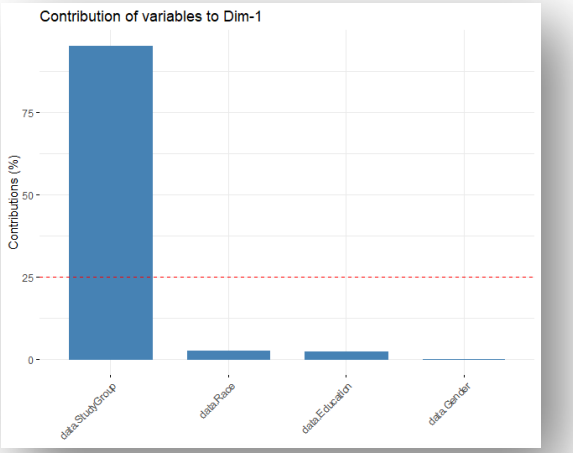


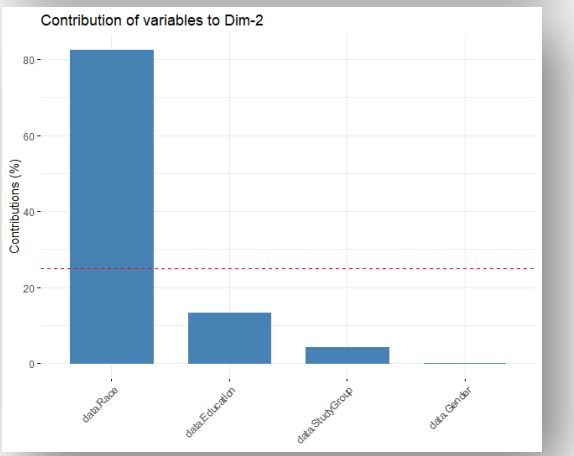



```
#####################################################
df <- data.frame(data$StudyGroup,
                 data$Race,
                 data$Education,
                 data$Gender)
#####################################################
######### Compute PCA on the data set #############
df.pca <- PCA(df, graph =TRUE, scale=FALSE)
###### create facet biplot with ellipses #########
fviz_pca(df.pca,
         label = FALSE,      # hide labels
         habillage = df,     # color by groups
         addEllipses = TRUE)# concentrate ellipses
#####################################################
```




```
#####################################################
df <- data.frame(data$Gender,data$Race,
                 data$Education, data$StudyGroup)
#####################################################
####### contributions of variables to PC1 ########
fviz_contrib(res.pca,
             choice = "var",
             axes = 1, top = 10)
####### contributions of variables to PC2 ########
fviz_contrib(res.pca,
             choice = "var",
             axes = 2, top = 10)
#####################################################
```





Acknowledgements

This work was supported by the Summer Internship Program by NIDA, IRP.

A special thank you to Mustapha Mezghanni and Dr. Jia-Ling Lin for their time and invaluable mentorship throughout the summer. Thank you to the BIS staff and to all the keynote speakers and workshop organizers for the IRP. Shout out to Rolanda Morris, Dr. Stephen Heishman, Christie Brannock, and Mark Forster, the MVP's of the virtual internship experience!