Applying PCA to the 2019 “Monitoring the Future” Study

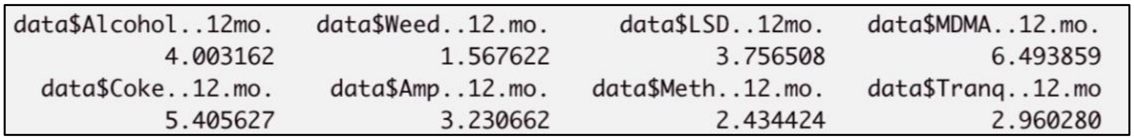
Project Stage 01

10 February 2023

Aretha Kassegnin, [redacted]

MAT 343 – Explorations In Data Analytics

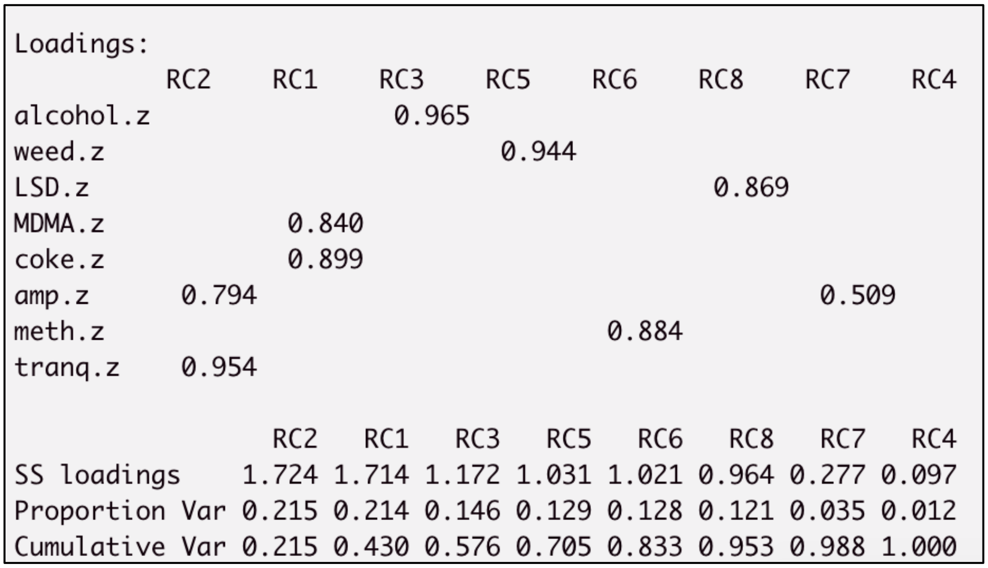
1. As shown in Figure 1.1, the VIF values for our dataset. We used the regression of the region on different types of substances; alcohol, cannabis, LSD, MDMA, cocaine, AMP, meth, and tranquilizer. However, we know if the VIF is greater than or equal to 5 then there is a moderate multicollinearity. Out of our 8 variables we only have 2 (MDMA and cocaine) variables that indicate moderate to strong multicollinearity.



```
data$Alcohol..12mo.   data$Weed..12.mo.     data$LSD..12mo.   data$MDMA..12.mo.
           4.003162            1.567622            3.756508            6.493859
  data$Coke..12.mo.    data$Amp..12.mo.   data$Meth..12.mo.   data$Tranq..12.mo
           5.405627            3.230662            2.434424            2.960280
```

Figure 1.1-- VIF Values for Variables

2. Figure 1.2 is the table of loadings for our data set.



```
Loadings:
          RC2    RC1    RC3    RC5    RC6    RC8    RC7    RC4
alcohol.z                0.965
weed.z                          0.944
LSD.z                                         0.869
MDMA.z            0.840
coke.z            0.899
amp.z      0.794                                     0.509
meth.z                                 0.884
tranq.z    0.954

                 RC2   RC1   RC3   RC5   RC6   RC8   RC7   RC4
SS loadings    1.724 1.714 1.172 1.031 1.021 0.964 0.277 0.097
Proportion Var 0.215 0.214 0.146 0.129 0.128 0.121 0.035 0.012
Cumulative Var 0.215 0.430 0.576 0.705 0.833 0.953 0.988 1.000
```

Figure 1.2 -Loading Values with data

a. The variables that showed up in the first three components are tranquilizes, amphetamines, cocaine, MDMA and alcohol.

b. The components from (a) are positively correlated because all the loadings were positive as they had values above 0, but were closer to 1 overall.

3. According to Figure 1.3 we need to remove 5 components.

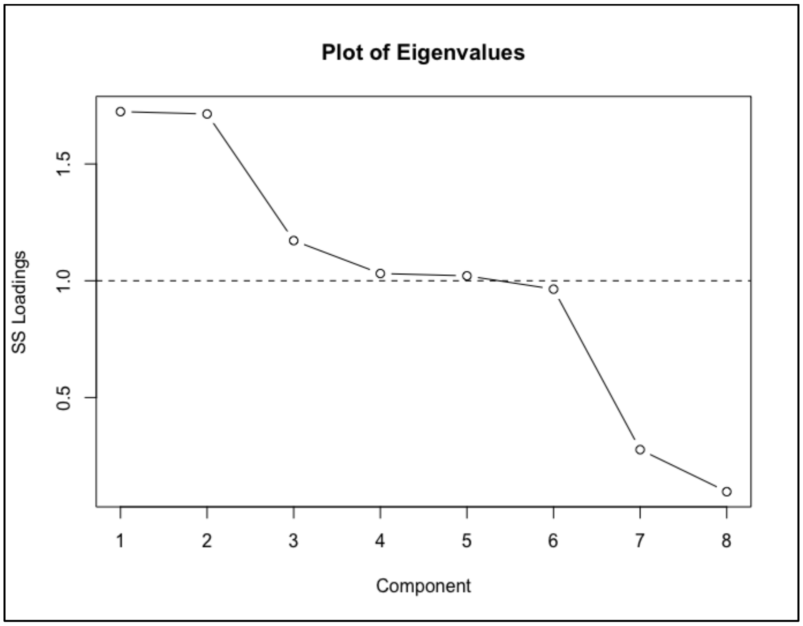
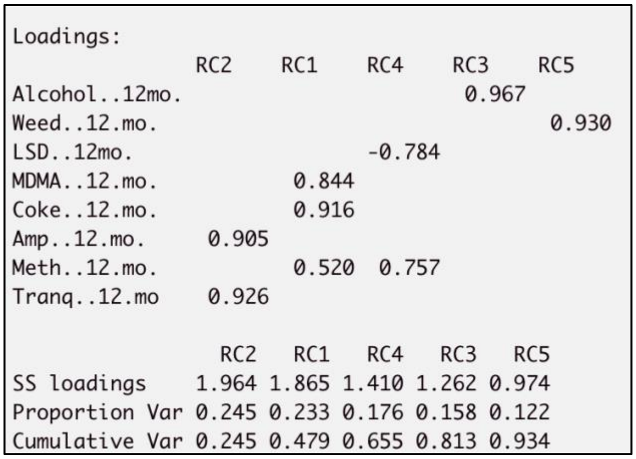


Figure 1.3 --Plot of Eigenvalues for data

4. When using Proportion of Variance Explained Criterion, if we wanted 70% of the variation, we should remove 4 components. If we want 90% then we should remove 2-3 variables.

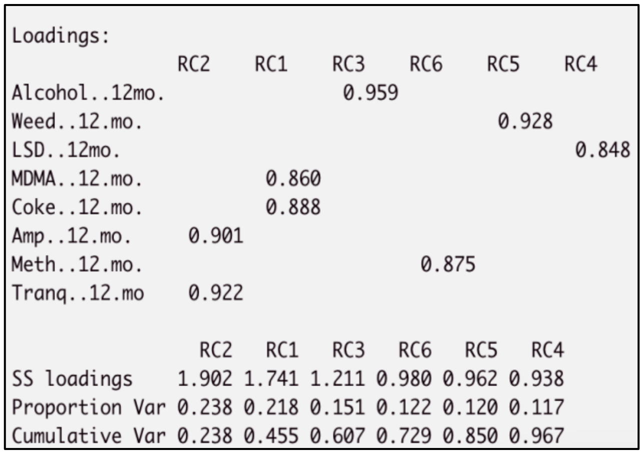
5.



```
Loadings:
                RC2    RC1    RC4    RC3    RC5
Alcohol..12mo.                        0.967
Weed..12.mo.                                 0.930
LSD..12mo.                   -0.784
MDMA..12.mo.           0.844
Coke..12.mo.           0.916
Amp..12.mo.     0.905
Meth..12.mo.           0.520  0.757
Tranq..12.mo    0.926

                  RC2   RC1   RC4   RC3   RC5
SS loadings     1.964 1.865 1.410 1.262 0.974
Proportion Var 0.245 0.233 0.176 0.158 0.122
Cumulative Var 0.245 0.479 0.655 0.813 0.934
```

Figure 1.5 Loading with 4 components removed



```
Loadings:
               RC2    RC1    RC3    RC6    RC5    RC4
Alcohol..12mo.                0.959
Weed..12.mo.                                0.928
LSD..12mo.                                         0.848
MDMA..12.mo.           0.860
Coke..12.mo.           0.888
Amp..12.mo.     0.901
Meth..12.mo.                         0.875
Tranq..12.mo    0.922

                 RC2   RC1   RC3   RC6   RC5   RC4
SS loadings    1.902 1.741 1.211 0.980 0.962 0.938
Proportion Var 0.238 0.218 0.151 0.122 0.120 0.117
Cumulative Var 0.238 0.455 0.607 0.729 0.850 0.967
```

Figure 1.6 Loading with 2 Components Removed

6. Figure 1.5 shows the loading with 4 components removed. Our target variable is the region on distinct types of substances. As you can see from the first component, RC2, the variables are AMP and tranquilizer which means this component covers common regions where these substances are used. The second component, RC1 shows the variables MDMA, coke and METH which means that RC1 must have at least 1 common region. Next, RC4 component has LSD and METH which is interesting because it was also in the last component, but it must mean that these 2 also share a common region. The fourth and fifth components RC3, only contains the variable alcohol and RC5 which Only contains the variable weed which means there are strong differences between the other components. Figure 1.6 shows the loading with only 2 Components revoked and there are some differences to Figure 1.5 which will be explained. RC2 is like Figure 1.5. RC1 component has MDMA and coke which means that they have a common region. The rest only have one variable each inside the component, RC3 contains alcohol, RC6 contains METH, RC5 contains weed and lastly RC4 contains LSD. Which means that 4 components cover different regions and do not share common pattern.

Appendix

```
data <- Data.Analytics...stage.1
region <- factor(data$Region, levels = c(1, 2, 3, 4),
             labels = c("Northeast", "Midwest", "South", "West"))
alcohol <- factor(data$Alcohol..12mo., levels = c(1, 2, 3, 4, 5, 6, 7),
              labels = c("None", "1-2x", "2-5x", "6-9x", "10-19x", "20-39x", "40+"))
weed <- factor(data$Weed..12.mo., levels = c(1, 2, 3, 4, 5, 6, 7),
            labels = c("None", "1-2x", "2-5x", "6-9x", "10-19x", "20-39x", "40+"))
LSD <- factor(data$LSD..12mo., levels = c(1, 2, 3, 4, 5, 6, 7),
           labels = c("None", "1-2x", "2-5x", "6-9x", "10-19x", "20-39x", "40+"))
mdma <- factor(data$MDMA..12.mo., levels = c(1, 2, 3, 4, 5, 6, 7),
            labels = c("None", "1-2x", "2-5x", "6-9x", "10-19x", "20-39x", "40+"))
coke <- factor(data$Coke..12.mo., levels = c(1, 2, 3, 4, 5, 6, 7),
            labels = c("None", "1-2x", "2-5x", "6-9x", "10-19x", "20-39x", "40+"))
amp <- factor(data$Amp..12.mo., levels = c(1, 2, 3, 4, 5, 6, 7),
           labels = c("None", "1-2x", "2-5x", "6-9x", "10-19x", "20-39x", "40+"))
meth <- factor(data$Meth..12.mo., levels = c(1, 2, 3, 4, 5, 6, 7),
            labels = c("None", "1-2x", "2-5x", "6-9x", "10-19x", "20-39x", "40+"))
tranq <- factor(data$Tranq..12.mo, levels = c(1, 2, 3, 4, 5, 6, 7),
             labels = c("None", "1-2x", "2-5x", "6-9x", "10-19x", "20-39x", "40+"))

names(data)
model03a <- lm(formula = data$Region ~
             data$Alcohol..12mo. + data$Weed..12.mo. + data$LSD..12mo. + data$MDMA..12.mo.
            +data$Coke..12.mo.+data$Amp..12.mo.+data$Meth..12.mo.+ data$Tranq..12.mo, data = data)
pairs(x = data[, c(2,3,4,5,6,7,8,9)], pch = 16)
library(car)
vif(model03a)

#Step 2 Partation data
library(caret)
set.seed(25)
inTrain <- createDataPartition(y = data$Region,
                        p = .75,
                        list = FALSE)

data.train <- data[ inTrain ,  ]
dim(data.train)[1]
dim(data)[1]
dim(data.train)[1]/dim(data)[1] #.75 so we are good
#testing data
data.test <- data[ -inTrain ,  ]
dim(data.test)[1]/dim(data)[1] #.249 so good
```

```
#bind everything together
data.train$trainortest <-
  rep("train", nrow(data.train))
names(data.train)
data.test$trainortest <-
  rep("test", nrow(data.test))
names(data.test)
data.all <- rbind(data.train, data.test)

#making sure the testing and training data look the same
boxplot(data.all$Region ~ (trainortest),
        data = data.all)
boxplot(data.all$Alcohol..12mo. ~ (trainortest),
        data = data.all)
boxplot(data.all$Weed..12.mo. ~ (trainortest),
        data = data.all)
boxplot(data.all$LSD..12mo. ~ (trainortest),
        data = data.all)
boxplot(data.all$MDMA..12.mo. ~ (trainortest),
        data = data.all)
boxplot(data.all$Coke..12.mo. ~ (trainortest),
        data = data.all)
boxplot(data.all$Amp..12.mo. ~ (trainortest),
        data = data.all)
boxplot(data.all$Meth..12.mo. ~ (trainortest),
        data = data.all)
boxplot(data.all$Tranq..12.mo ~ (trainortest),
        data = data.all)
#getting p.values for all
kruskal.test(data$Region ~ as.factor(trainortest),
             data = data.all)$p.value
#.909
kruskal.test(data$Alcohol..12mo. ~ as.factor(trainortest),
             data = data.all)$p.value
#.742
kruskal.test(data$Weed..12.mo. ~ as.factor(trainortest),
             data = data.all)$p.value
#.957
kruskal.test(data$LSD..12mo. ~ as.factor(trainortest),
             data = data.all)$p.value
#.952
kruskal.test(data$MDMA..12.mo. ~ as.factor(trainortest),
             data = data.all)$p.value
#614
```

```
kruskal.test(data$Coke..12.mo. ~ as.factor(trainortest),
            data = data.all)$p.value
#.723
kruskal.test(data$Amp..12.mo. ~ as.factor(trainortest),
            data = data.all)$p.value
#.744
kruskal.test(data$Meth..12.mo. ~ as.factor(trainortest),
            data = data.all)$p.value
#.911
kruskal.test(data$Tranq..12.mo ~ as.factor(trainortest),
            data = data.all)$p.value
#all p-values >.05
#PCA
# Correlation matrix
head(data.train)
y <- data.train$Region
X <- data.train[, c(2,3,4,5,6,7,8,9)]
X_test <-data.test[, c(2,3,4,5,6,7,8,9)]
head(X)
X_z <- as.data.frame(scale(X))
head(X_z)
cor(X_z)
round(cor(X_z), 3)

colnames(X_z) <- c("alcohol.z","weed.z","LSD.z","MDMA.z","coke.z","amp.z",
                  "meth.z","tranq.z")


model02 <- lm(formula = y ~
               alcohol.z + weed.z + LSD.z + MDMA.z
             +coke.z+amp.z+meth.z+ tranq.z, data = X_z)
library(car)
vif(model02)
library(psych)
pca01 <- principal(r = X_z, rotate = "varimax", nfactors = 8)

pca01$loadings
print(pca01$loadings, cutoff = 0.49)

#3
# Eigenvalues are in "SS Loadings"
ss.load <- c(1.724, 1.714, 1.172, 1.031, 1.021, 0.964 ,0.277, 0.097)
plot(ss.load, type = "b", main = "Plot of Eigenvalues",
     ylab = "SS Loadings", xlab = "Component"); abline(h = 1, lty =2)
```

```
#5 Rerun PCA
pca02 <- principal(r = X, rotate = "varimax", nfactors = 5)
print(pca02$loadings, cutoff = 0.5)
pca02_2 <- principal(r = X, rotate = "varimax", nfactors = 4)
print(pca02_2$loadings, cutoff = 0.5)
```