07- PCA + Pearson, SOD

Setup

```
knitr::opts_chunk$set(
  echo = TRUE,          # Display code chunks
  eval = TRUE,          # Evaluate code chunks
  warning = FALSE,      # Hide warnings
  message = FALSE,      # Hide messages
  fig.width = 8,        # Set plot width in inches
  fig.height = 5,       # Set plot height in inches
  fig.align = "center" # Align plots to the center
)
```

Load packages

```
library(tidyr)
library(tidyverse)
library(ggplot2)
library(vegan)
#library(tinytex)
```

Load data

Note:

For *data* the units are listed below. Weight = g
Length, width, height = mm
p450, SOD = activity/ (mg/protein)
Condition factor, economic factor = unitless
For *pah*, *indv*, and *allana* the units are ng/g
For *metal* the units are mg/kg

```
getwd()
```

```
## [1] "/Users/cmantegna/Documents/WDFWmussels/code"
```

```
#data has all sites, coordinates, p450, sod, condition factor, economic factor data
data<- read.csv("/Users/cmantegna/Documents/WDFWmussels/data/biomarkerfull.csv")

#pah has complete site values and different summed pah analyte whole tissue values
pah<- read.csv("/Users/cmantegna/Documents/Biomarker Data Analysis/sum_analytes.csv")

#indv has complete site values and individual named pah analyte whole tissue values
indv<- read.csv("/Users/cmantegna/Documents/Biomarker Data Analysis/individual_analytes.csv")

metal<- read.csv("/Users/cmantegna/Documents/Biomarker Data Analysis/metal.csv")
```

```
allana<- read.csv("/Users/cmantegna/Documents/Biomarker Data Analysis/allana.csv")
```

```
# Review data frame structure

#str(metal)
#str(allana)
#str(indv)
```

```
# Review basic data types and stats

#summary(data)
#summary(pah)
#summary(indv)
```

```
head(data)
```

```
##   latitude longitude       site_name site_number sample    p450   SOD
## 1 48.67938 -122.6301 Aiston Preserve          77    239 5965780 0.000
## 2 48.67938 -122.6301 Aiston Preserve          77    240 1508156 4.877
## 3 48.67938 -122.6301 Aiston Preserve          77    241 4674882 8.871
## 4 48.67938 -122.6301 Aiston Preserve          77    242 2861653 0.010
## 5 47.50161 -122.3859    Arroyo Beach          13    281 3448794 7.084
## 6 47.50161 -122.3859    Arroyo Beach          13    282 6485447 0.635
##   weight_initial length width height weight_final weight_change
## 1        11.6884   53.9 22.73  18.59       3.2826          8.41
## 2         10.833  53.49 23.92  18.36       3.4809          7.35
## 3        14.7041  55.99 27.79  19.57       4.7251          9.98
## 4        14.6121  58.55 28.38  19.55       4.4461         10.17
## 5        15.4756  58.14 26.11  20.16       4.6221         10.85
## 6        17.9501  60.43 27.56   22.3       6.1066         11.84
##   condition_factor avg_thickness economic_index
## 1           0.1560         0.700         0.0018
## 2           0.1374         0.790          0.002
## 3           0.1782         0.825          0.002
## 4           0.1737         0.930         0.0021
## 5           0.1866         0.920         0.0022
## 6           0.1959         0.965         0.0022
```

```
head(metal)
```

```
##   Latitude Longitude LabSampleID             SiteName LabSampleID.1
## 1 47.50159 -122.3858    L79603-1         Arroyo Beach      L79603-1
## 2 47.68203 -122.5067    L79603-2       Brackenwood Ln      L79603-2
## 3 47.29469 -122.5305    L79603-3         Salmon Beach      L79603-3
## 4 48.04887 -122.7711    L79603-4 Chimacum Creek delta      L79603-4
## 5 47.66141 -122.4989    L79603-5          Skiff Point      L79603-5
## 6 48.02655 -122.7509    L79603-6    S of Skunk Island      L79603-6
##        Analyte Qualifier Units PctSolids   DryValue
## 1 mercuryTotal         D mg/Kg      17.0 0.03600000
## 2 mercuryTotal         D mg/Kg      16.9 0.03745562
## 3 mercuryTotal         D mg/Kg      17.9 0.02379888
## 4 mercuryTotal         D mg/Kg      17.0 0.03264706
## 5 mercuryTotal         D mg/Kg      17.8 0.03932584
## 6 mercuryTotal         D mg/Kg      17.5 0.02868571
```

```r
head(allana)
```

```
##               SiteName Latitude Longitude      Analyte Qualifier Units
## 1         Arroyo Beach 47.50159 -122.3858 mercuryTotal         D mg/Kg
## 2       Brackenwood Ln 47.68203 -122.5067 mercuryTotal         D mg/Kg
## 3         Salmon Beach 47.29469 -122.5305 mercuryTotal         D mg/Kg
## 4 Chimacum Creek delta 48.04887 -122.7711 mercuryTotal         D mg/Kg
## 5          Skiff Point 47.66141 -122.4989 mercuryTotal         D mg/Kg
## 6    S of Skunk Island 48.02655 -122.7509 mercuryTotal         D mg/Kg
##   PctSolids   DryValue
## 1      17.0 0.03600000
## 2      16.9 0.03745562
## 3      17.9 0.02379888
## 4      17.0 0.03264706
## 5      17.8 0.03932584
## 6      17.5 0.02868571
```

Data frame manipulations

Adjusting biomarker values for accurate stats

```r
# Data contains 0's and must be adjusted in this order to preserve all usable data.

#sod
#replace any SOD values at or below 0 with half of the lower detection limit of .005 (.005*.5). Lower d
data$SOD[data$SOD <= 0] <- 0.0025

#p450
#remove any p450 values that are 0 - those are true 0's not non-detectable. I am replacing with na so I
data$p450[data$p450 <= 0] <- NA
```

Data adjustment for analysis- SOD & p450

```r
#Average the

library(dplyr)

#simplifying the dataframe for joining with next steps
averaged_data <- data %>%
  group_by(site_number, latitude, longitude, site_name) %>%
  summarise(
    avg_p450 = mean(p450, na.rm = TRUE),
    avg_SOD = mean(SOD, na.rm = TRUE)
  ) %>%
  ungroup()  # Remove grouping for the new dataframe

print(averaged_data)
```

```
## # A tibble: 74 x 6
##    site_number latitude longitude site_name                avg_p450 avg_SOD
##          <int>    <dbl>     <dbl> <chr>                       <dbl>   <dbl>
##  1           1     48.1     -123. Port Angeles Yacht Club  5751355    7.39
##  2           2     48.0     -123. Jamestown                3263515   24.5
```

```
##  3           3     48.2     -123. Penn Cove Reference      2427656.  23.9
##  4           7     48.3     -123. North Camano            12290521    0.752
##  5           8     48.0     -123. Chimacum Creek delta     2641574.   2.19
##  6           9     48.0     -123. S of Skunk Island        3556923.  11.3
##  7          10     48.0     -123. Oak Bay County Park      2335145   19.8
##  8          11     48.0     -123. Maristone Island         4772561.   5.68
##  9          12     48.1     -123. Discovery Bay            4029898.   8.74
## 10          13     47.5     -122. Arroyo Beach             4480860.   8.83
## # i 64 more rows
```

```r
library(reshape2)
#merge data frames and reshape for input.
colnames(allana)[colnames(allana) == "SiteName"] <- "site_name"
merged_df <- merge(averaged_data, allana, by = c("site_name"), all.x = TRUE)

#reshape to get the analytes into their own columns with the DryValue as their values
reshaped_df <- dcast(merged_df, site_name + site_number +latitude + longitude + avg_p450 + avg_SOD ~ Ana

head(reshaped_df)
```

```
##                       site_name site_number latitude longitude avg_p450
## 1                Aiston Preserve          77 48.67938 -122.6301  3752618
## 2                   Arroyo Beach          13 47.50161 -122.3859  4480860
## 3                  Blair Waterway          41 47.27568 -122.4173  4879642
## 4                Blair Waterway #2          42 47.26324 -122.3857  3714918
## 5                  Brackenwood Ln          23 47.68234 -122.5064  1857012
## 6 Broad Spit (Fisherman's Point)          30 47.78184 -122.8347  2311731
##     avg_SOD  arsenic  cadmium   copper      lead mercuryTotal Sum40CBs
## 1  3.440125 7.245509 1.652695 4.940120 0.1772455   0.03305389 23.13322
## 2  8.832583 9.647059 1.952941 5.623529 0.2423529   0.03600000 34.81931
## 3  6.517750 8.114286 1.622857 5.828571 0.2554286   0.03205714 37.40489
## 4 10.796000 8.373494 1.704819 8.132530 0.1849398  -0.02360000 42.15557
## 5  9.835125 8.698225 1.857988 6.213018 0.2201183   0.03745562 29.49750
## 6  7.116250       NA       NA       NA        NA           NA       NA
##     SumBDEs  SumCHLDs   SumDDTs    SumHCHs    SumPAHs SumPAHs16
## 1 -1.095784 -1.095784  1.826307 -0.9131536   97.40305  31.65599
## 2  5.643129  3.121731  2.401332 -1.1406325  408.22636 168.09321
## 3 15.952084  4.345568  9.901294  0.8801150  715.09344 247.53234
## 4  9.164255 -1.710661 17.717559  0.9775205 1038.61553 299.36565
## 5  2.005830 -1.592865  2.182815 -1.2978901  389.36704 176.98502
## 6        NA        NA        NA         NA         NA        NA
##   SumPAHs42_DMNcorrected SumPAHsHMW SumPAHsLMW SumPCBs2x17      Zinc NA
## 1               97.40305   19.48061   79.13998    28.00338  77.24551 NA
## 2              408.22636  186.10320  216.11984    49.22730 104.11765 NA
## 3              715.09344  280.53666  456.55966    50.60661  98.85714 NA
## 4             1038.61553  403.22721  610.95031    54.98553  85.54217 NA
## 5              389.36704  176.98502  212.38202    44.24625  90.53254 NA
## 6                     NA         NA         NA          NA        NA NA
```

Data frame for SOD GLM & PCA- All summed analytes and metals

```r
#create a table without the avg_p450 and NA column for SOD work
cols_to_keep <- colnames(reshaped_df)[!colnames(reshaped_df) %in% c("avg_p450", "NA")]
```

```r
sod_all <- reshaped_df[, cols_to_keep]

head(sod_all)
```

```
##                        site_name site_number latitude longitude   avg_SOD
## 1                Aiston Preserve          77 48.67938 -122.6301  3.440125
## 2                   Arroyo Beach          13 47.50161 -122.3859  8.832583
## 3                  Blair Waterway          41 47.27568 -122.4173  6.517750
## 4               Blair Waterway #2          42 47.26324 -122.3857 10.796000
## 5                  Brackenwood Ln          23 47.68234 -122.5064  9.835125
## 6 Broad Spit (Fisherman's Point)          30 47.78184 -122.8347  7.116250
##    arsenic  cadmium   copper      lead mercuryTotal Sum40CBs   SumBDEs
## 1 7.245509 1.652695 4.940120 0.1772455   0.03305389 23.13322 -1.095784
## 2 9.647059 1.952941 5.623529 0.2423529   0.03600000 34.81931  5.643129
## 3 8.114286 1.622857 5.828571 0.2554286   0.03205714 37.40489 15.952084
## 4 8.373494 1.704819 8.132530 0.1849398  -0.02360000 42.15557  9.164255
## 5 8.698225 1.857988 6.213018 0.2201183   0.03745562 29.49750  2.005830
## 6       NA       NA       NA        NA           NA       NA        NA
##    SumCHLDs   SumDDTs    SumHCHs    SumPAHs SumPAHs16 SumPAHs42_DMNcorrected
## 1 -1.095784  1.826307 -0.9131536   97.40305  31.65599               97.40305
## 2  3.121731  2.401332 -1.1406325  408.22636 168.09321              408.22636
## 3  4.345568  9.901294  0.8801150  715.09344 247.53234              715.09344
## 4 -1.710661 17.717559  0.9775205 1038.61553 299.36565             1038.61553
## 5 -1.592865  2.182815 -1.2978901  389.36704 176.98502              389.36704
## 6        NA        NA         NA         NA        NA                     NA
##   SumPAHsHMW SumPAHsLMW SumPCBs2x17      Zinc
## 1   19.48061   79.13998    28.00338  77.24551
## 2  186.10320  216.11984    49.22730 104.11765
## 3  280.53666  456.55966    50.60661  98.85714
## 4  403.22721  610.95031    54.98553  85.54217
## 5  176.98502  212.38202    44.24625  90.53254
## 6         NA         NA          NA        NA
```

```r
#create a table without the avg_SOD and NA column for p450 work
#cols_to_keep2 <- colnames(reshaped2_df)[!colnames(reshaped2_df) %in% c("avg_SOD", "arsenic", "cadmium"

#p450PAH <- reshaped2_df[, cols_to_keep2]
#p450PAH$plogdata <- plogdata

#print(p450PAH)
```

#SOD Pearson- summed analytes + metals

```r
#get the column names from sod_all so I don't have to individually type each one

all_columns <- names(sod_all)

# Remove the columns you don't want to include in the model
excluded_columns <- c('latitude', 'longitude', 'site_name', 'site_number')
independent_columns <- all_columns[!all_columns %in% excluded_columns]

# Enclose each column name in backticks to handle special characters
independent_columns <- sapply(independent_columns, function(x) paste0("`", x, "`"))

# Create a string representing the formula
```

```r
formula_str <- paste("avg_SOD ~", paste(independent_columns, collapse = " + "))

# Convert the string to a formula object
formula <- as.formula(formula_str)

#SODall_glm<- glm(formula, data = sod_all, family = poisson())
#print(summary(SODall_glm))
```

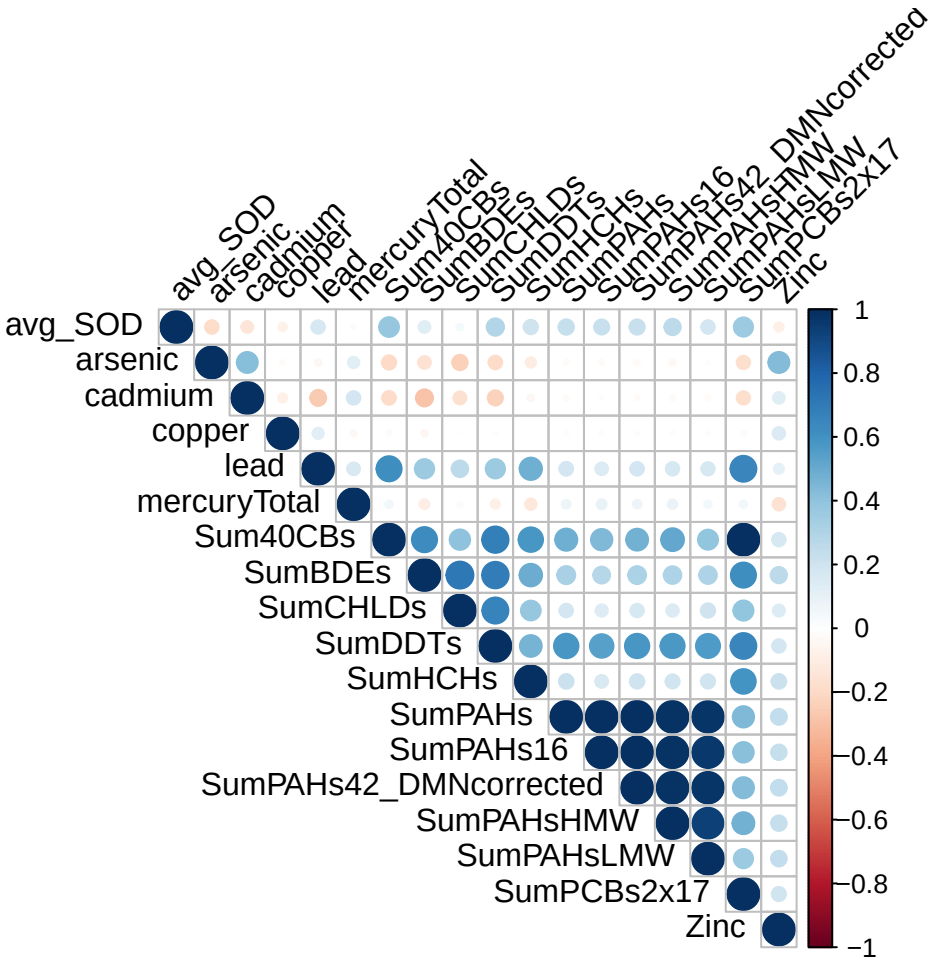
```r
library(corrplot)

# Extract variable names from the formula
variables <- all.vars(formula)

# Subset the dataframe 'sod_all' using the extracted variables
subset_data <- sod_all[, variables]

# Compute Pearson correlation for each pair of variables
correlation_results <- cor(subset_data, method = "pearson", use = "complete.obs")

# View the correlation matrix
print(correlation_results)
```

```
##                            avg_SOD     arsenic     cadmium       copper
## avg_SOD                 1.00000000 -0.18109708 -0.13679429 -0.071886723
## arsenic                -0.18109708  1.00000000  0.42964277 -0.024474635
## cadmium                -0.13679429  0.42964277  1.00000000 -0.080496601
## copper                 -0.07188672 -0.02447463 -0.08049660  1.000000000
## lead                    0.17909043 -0.04530496 -0.25205192  0.126068698
## mercuryTotal           -0.02066344  0.12808725  0.18107692 -0.037083547
## Sum40CBs                0.38620940 -0.19933462 -0.18703454  0.026498886
## SumBDEs                 0.13709532 -0.15984494 -0.28670855 -0.041398174
## SumCHLDs                0.04655850 -0.23355789 -0.16565064  0.009537200
## SumDDTs                 0.29487850 -0.18349868 -0.22608907 -0.011444606
## SumHCHs                 0.20381451 -0.10503545 -0.04081879  0.006241142
## SumPAHs                 0.23381015 -0.02848968 -0.02744569 -0.013884628
## SumPAHs16               0.22471733 -0.02857410 -0.01190872 -0.017910276
## SumPAHs42_DMNcorrected  0.22880768 -0.02736677 -0.02408440 -0.014481562
## SumPAHsHMW              0.26252622 -0.03805331 -0.02969452 -0.014786304
## SumPAHsLMW              0.18100595 -0.01410298 -0.02141851 -0.014497708
## SumPCBs2x17             0.36205716 -0.17571353 -0.17543567  0.023121946
## Zinc                   -0.08616938  0.43419382  0.13501307  0.153751640
##                                lead mercuryTotal    Sum40CBs     SumBDEs
## avg_SOD                 0.17909043  -0.02066344  0.38620940  0.13709532
## arsenic                -0.04530496   0.12808725 -0.19933462 -0.15984494
## cadmium                -0.25205192   0.18107692 -0.18703454 -0.28670855
## copper                  0.12606870  -0.03708355  0.02649889 -0.04139817
## lead                    1.00000000   0.16161533  0.61893341  0.36781393
## mercuryTotal            0.16161533   1.00000000  0.05533165 -0.09540215
## Sum40CBs                0.61893341   0.05533165  1.00000000  0.62046842
## SumBDEs                 0.36781393  -0.09540215  0.62046842  1.00000000
## SumCHLDs                0.26944325   0.02901457  0.40878409  0.71573142
## SumDDTs                 0.36657185  -0.08564286  0.68161682  0.69101363
## SumHCHs                 0.48918284  -0.12941072  0.58861450  0.49906532
## SumPAHs                 0.18793861   0.07652008  0.48116105  0.32040720
```

```
## SumPAHs16               0.15666432   0.09638492  0.44798700  0.28113248
## SumPAHs42_DMNcorrected  0.18359108   0.07712182  0.47028260  0.31501120
## SumPAHsHMW              0.17912920   0.08863539  0.51042358  0.30627078
## SumPAHsLMW              0.17600615   0.05354115  0.39419431  0.30794301
## SumPCBs2x17             0.65795381   0.05539238  0.99469715  0.61980432
## Zinc                    0.10012336  -0.15507712  0.17851321  0.26134075
##                           SumCHLDs     SumDDTs      SumHCHs     SumPAHs
## avg_SOD                 0.04655850  0.29487850  0.203814508  0.23381015
## arsenic                -0.23355789 -0.18349868 -0.105035453 -0.02848968
## cadmium                -0.16565064 -0.22608907 -0.040818791 -0.02744569
## copper                  0.00953720 -0.01144461  0.006241142 -0.01388463
## lead                    0.26944325  0.36657185  0.489182843  0.18793861
## mercuryTotal            0.02901457 -0.08564286 -0.129410720  0.07652008
## Sum40CBs                0.40878409  0.68161682  0.588614504  0.48116105
## SumBDEs                 0.71573142  0.69101363  0.499065325  0.32040720
## SumCHLDs                1.00000000  0.66786488  0.384878707  0.18064671
## SumDDTs                 0.66786488  1.00000000  0.465467166  0.58823991
## SumHCHs                 0.38487871  0.46546717  1.000000000  0.21162856
## SumPAHs                 0.18064671  0.58823991  0.211628564  1.00000000
## SumPAHs16               0.14798578  0.54458486  0.162357968  0.99718621
## SumPAHs42_DMNcorrected  0.17873956  0.58166849  0.206042981  0.99975415
## SumPAHsHMW              0.15072869  0.57419176  0.194937175  0.98951135
## SumPAHsLMW              0.20376620  0.56765210  0.198470280  0.97443069
## SumPCBs2x17             0.39797013  0.65410761  0.596529010  0.44634293
## Zinc                    0.14415937  0.18184344  0.212938547  0.24922279
##                          SumPAHs16 SumPAHs42_DMNcorrected  SumPAHsHMW
## avg_SOD                 0.22471733             0.22880768  0.26252622
## arsenic                -0.02857410            -0.02736677 -0.03805331
## cadmium                -0.01190872            -0.02408440 -0.02969452
## copper                 -0.01791028            -0.01448156 -0.01478630
## lead                    0.15666432             0.18359108  0.17912920
## mercuryTotal            0.09638492             0.07712182  0.08863539
## Sum40CBs                0.44798700             0.47028260  0.51042358
## SumBDEs                 0.28113248             0.31501120  0.30627078
## SumCHLDs                0.14798578             0.17873956  0.15072869
## SumDDTs                 0.54458486             0.58166849  0.57419176
## SumHCHs                 0.16235797             0.20604298  0.19493717
## SumPAHs                 0.99718621             0.99975415  0.98951135
## SumPAHs16               1.00000000             0.99755901  0.98920448
## SumPAHs42_DMNcorrected  0.99755901             1.00000000  0.98671078
## SumPAHsHMW              0.98920448             0.98671078  1.00000000
## SumPAHsLMW              0.96998307             0.97845527  0.93306600
## SumPCBs2x17             0.41341467             0.43593960  0.47305939
## Zinc                    0.23466720             0.24785194  0.23798835
##                         SumPAHsLMW SumPCBs2x17        Zinc
## avg_SOD                 0.18100595  0.36205716 -0.08616938
## arsenic                -0.01410298 -0.17571353  0.43419382
## cadmium                -0.02141851 -0.17543567  0.13501307
## copper                 -0.01449771  0.02312195  0.15375164
## lead                    0.17600615  0.65795381  0.10012336
## mercuryTotal            0.05354115  0.05539238 -0.15507712
## Sum40CBs                0.39419431  0.99469715  0.17851321
## SumBDEs                 0.30794301  0.61980432  0.26134075
## SumCHLDs                0.20376620  0.39797013  0.14415937
```

```
## SumDDTs                  0.56765210  0.65410761  0.18184344
## SumHCHs                  0.19847028  0.59652901  0.21293855
## SumPAHs                  0.97443069  0.44634293  0.24922279
## SumPAHs16                0.96998307  0.41341467  0.23466720
## SumPAHs42_DMNcorrected   0.97845527  0.43593960  0.24785194
## SumPAHsHMW               0.93306600  0.47305939  0.23798835
## SumPAHsLMW               1.00000000  0.36435140  0.24095207
## SumPCBs2x17              0.36435140  1.00000000  0.19847293
## Zinc                     0.24095207  0.19847293  1.00000000
```

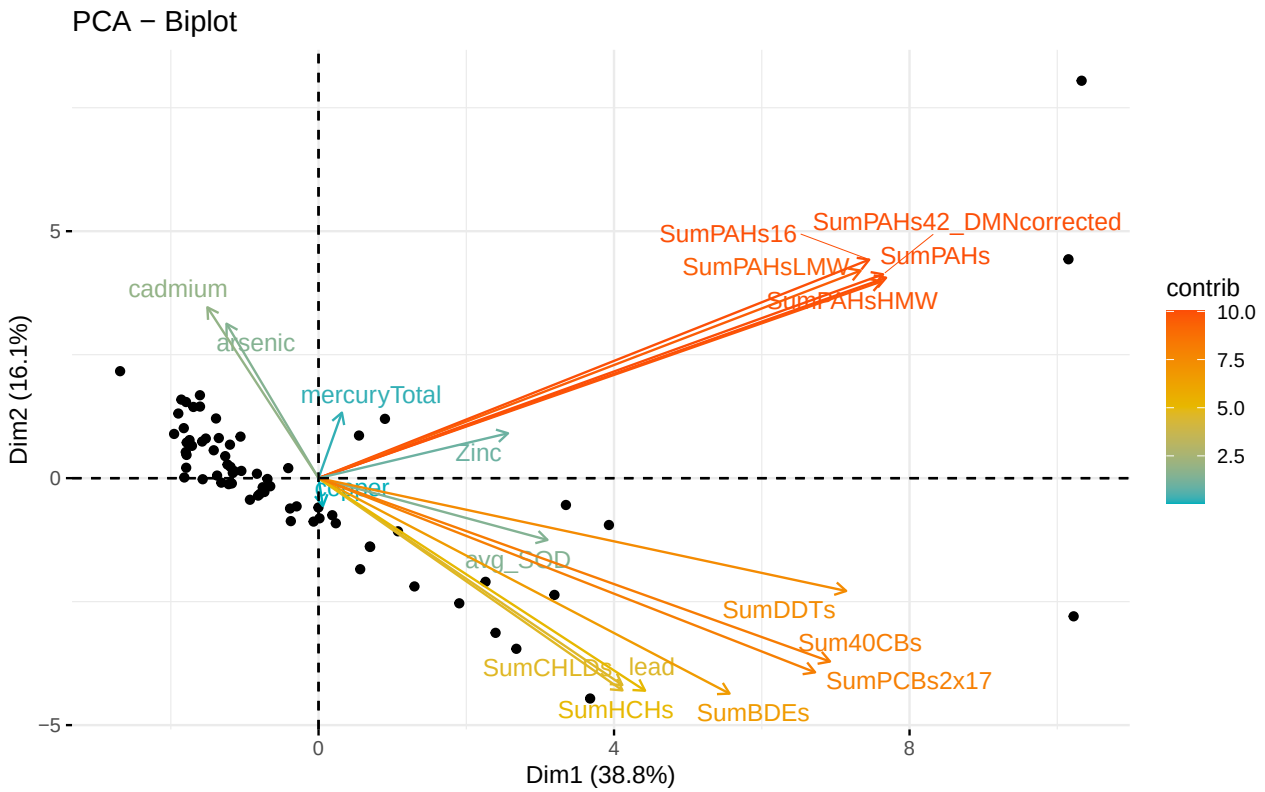
```
corrplot(correlation_results, method = "circle", type = "upper", tl.col = "black", tl.srt = 45)
```



SOD PCA - all analytes + metals

```
# PCA Plot with biomarkers
#install.packages("FactoMineR")
#install.packages("factoextra")
library('FactoMineR')
library("factoextra")


# Remove NAs from the dataset
df_clean <- na.omit(sod_all)
```

```r
# Selecting the relevant variables for PCA
pca_data <- df_clean[, c("avg_SOD", "arsenic", "cadmium", "copper", "lead", "mercuryTotal", "Sum40CBs",

# Performing PCA
pca_res <- PCA(pca_data, scale.unit = TRUE, graph = FALSE)

# Plotting the PCA
pcaplot<- fviz_pca_biplot(pca_res, label = "var", col.var = "contrib",
                gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
                repel = TRUE)  # Avoid text overlapping (slow if many points)



print(pcaplot)
```




```r
#ggsave(plot=pcaplot, filename="/Users/cmantegna/Documents/WDFWmussels/output/pca.png", width=15, heigh
```