BDA - Project

Contents

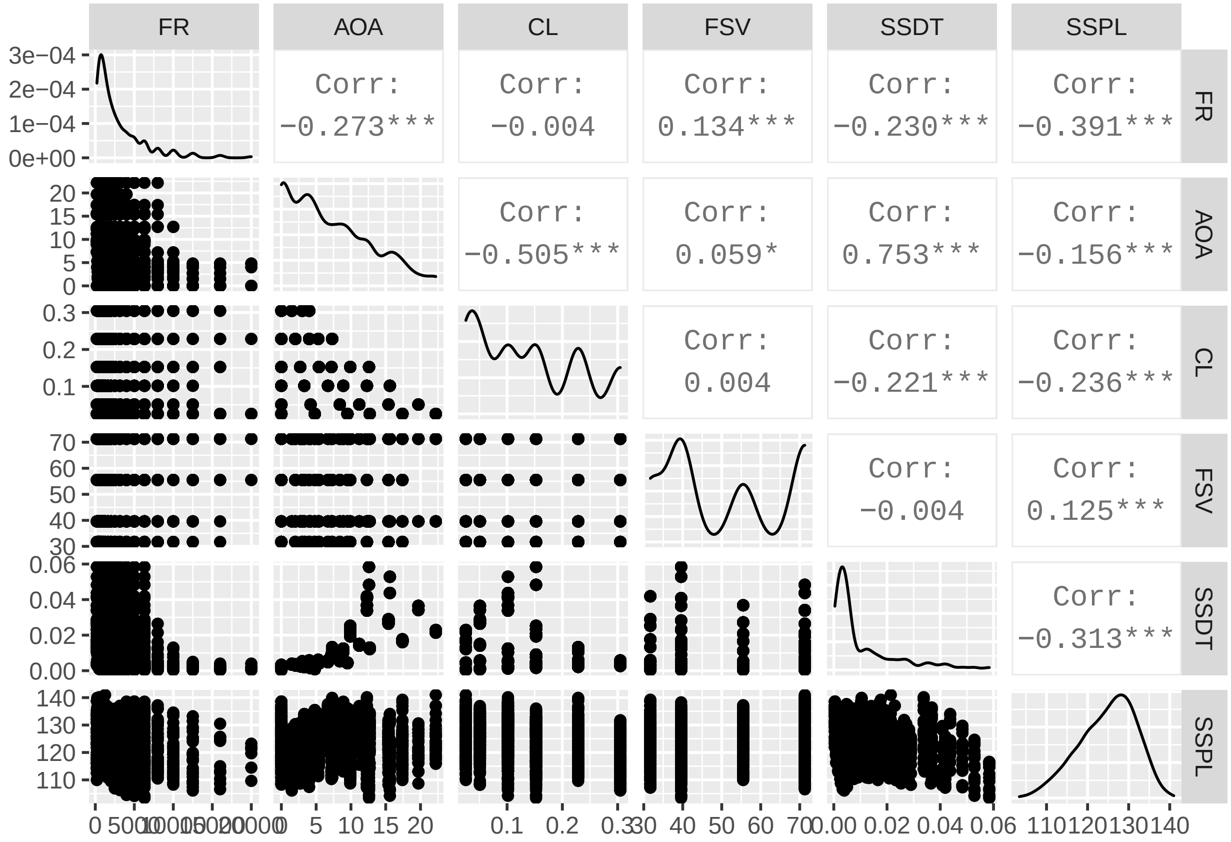
Data loading

```
airfoil <- read.csv("data/airfoil_self_noise.csv", sep=";")
colnames(airfoil) <- c("FR", "AOA", "CL", "FSV", "SSDT", "SSPL")
sample <- caTools::sample.split(airfoil, SplitRatio = 0.75)
train <- subset(airfoil, sample == TRUE)
test <- subset(airfoil, sample == FALSE)
print(airfoil[1:5,])
```

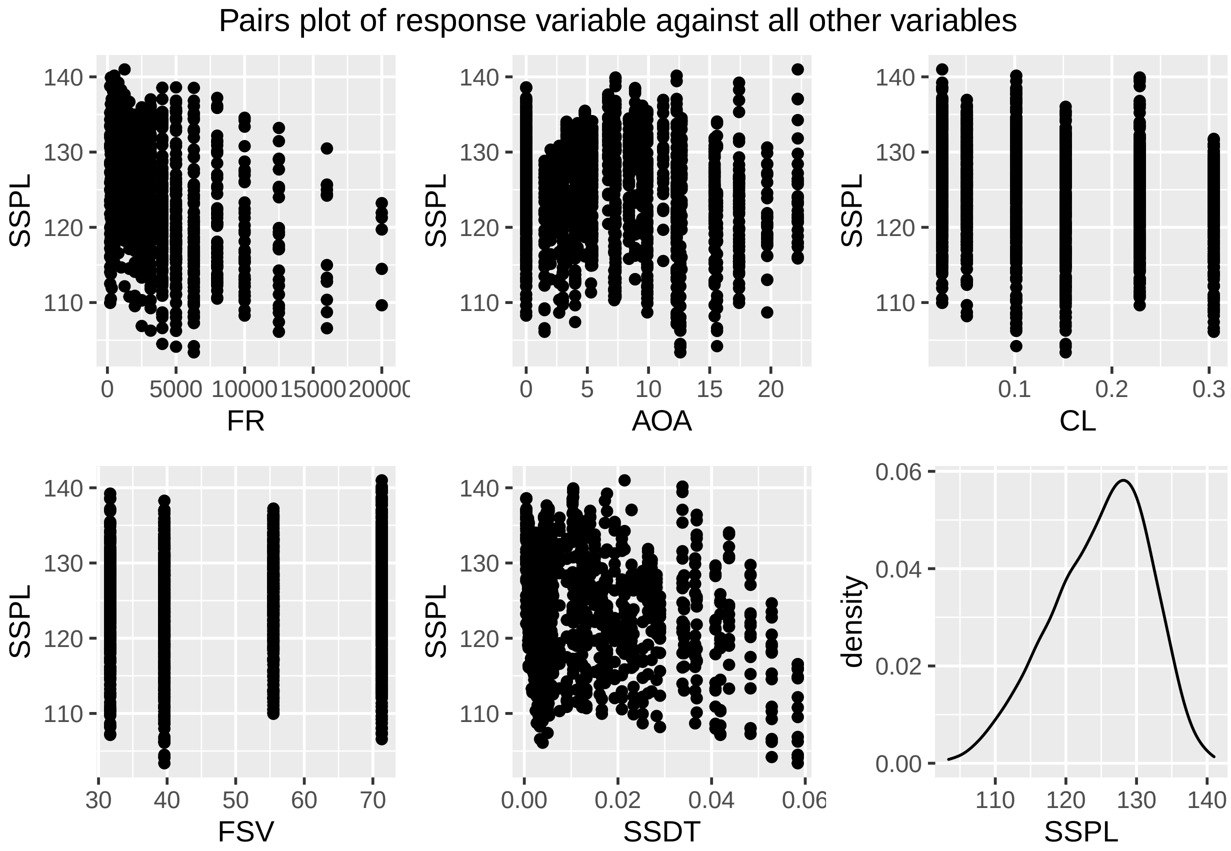
```
##     FR AOA     CL  FSV       SSDT    SSPL
## 1  800   0 0.3048 71.3 0.00266337 126.201
## 2 1000   0 0.3048 71.3 0.00266337 125.201
## 3 1250   0 0.3048 71.3 0.00266337 125.951
## 4 1600   0 0.3048 71.3 0.00266337 127.591
## 5 2000   0 0.3048 71.3 0.00266337 127.461
```

Exploratory data visualization

```
pairs <- ggpairs(airfoil)
pairs
```

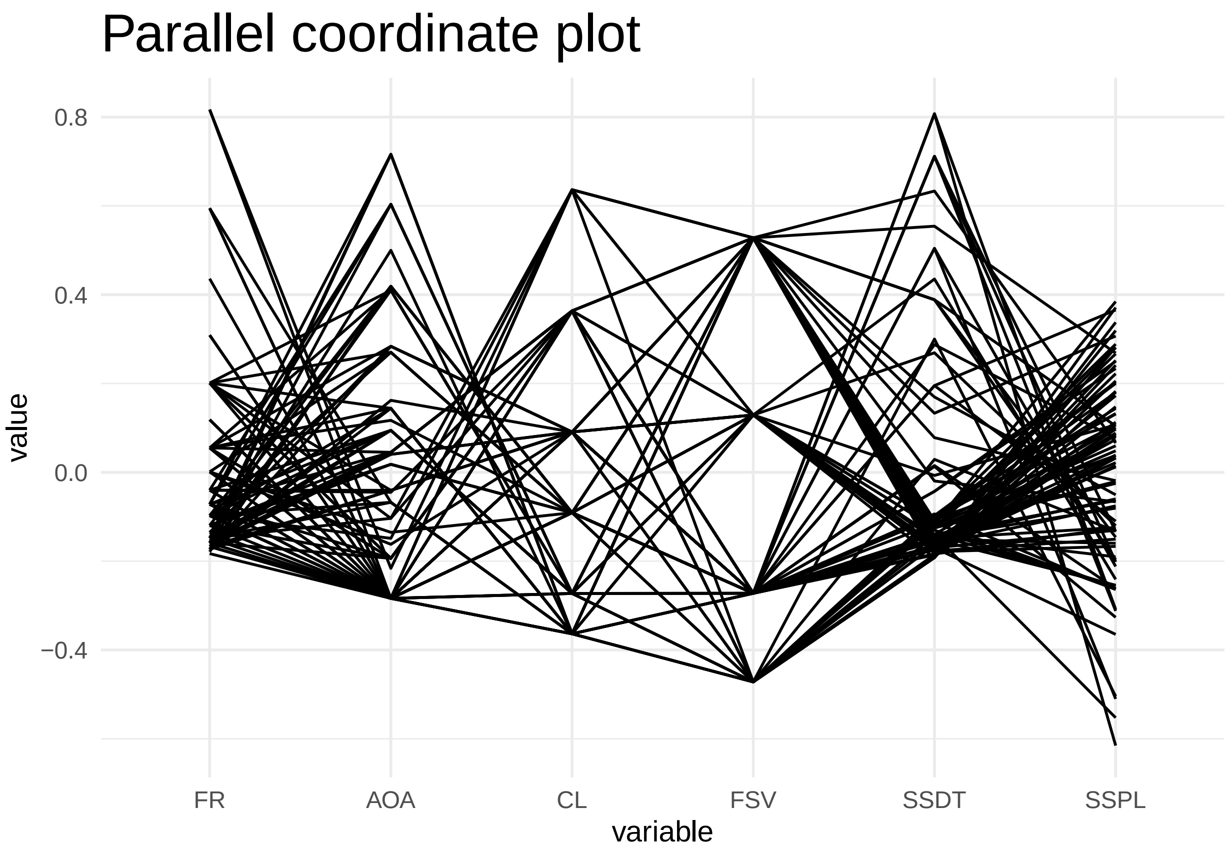


```
gridExtra::grid.arrange(pairs[6,1], pairs[6,2], pairs[6,3],
                        pairs[6,4], pairs[6,5], pairs[6,6], nrow=2,
                        top="Pairs plot of response variable against all other variables")
```



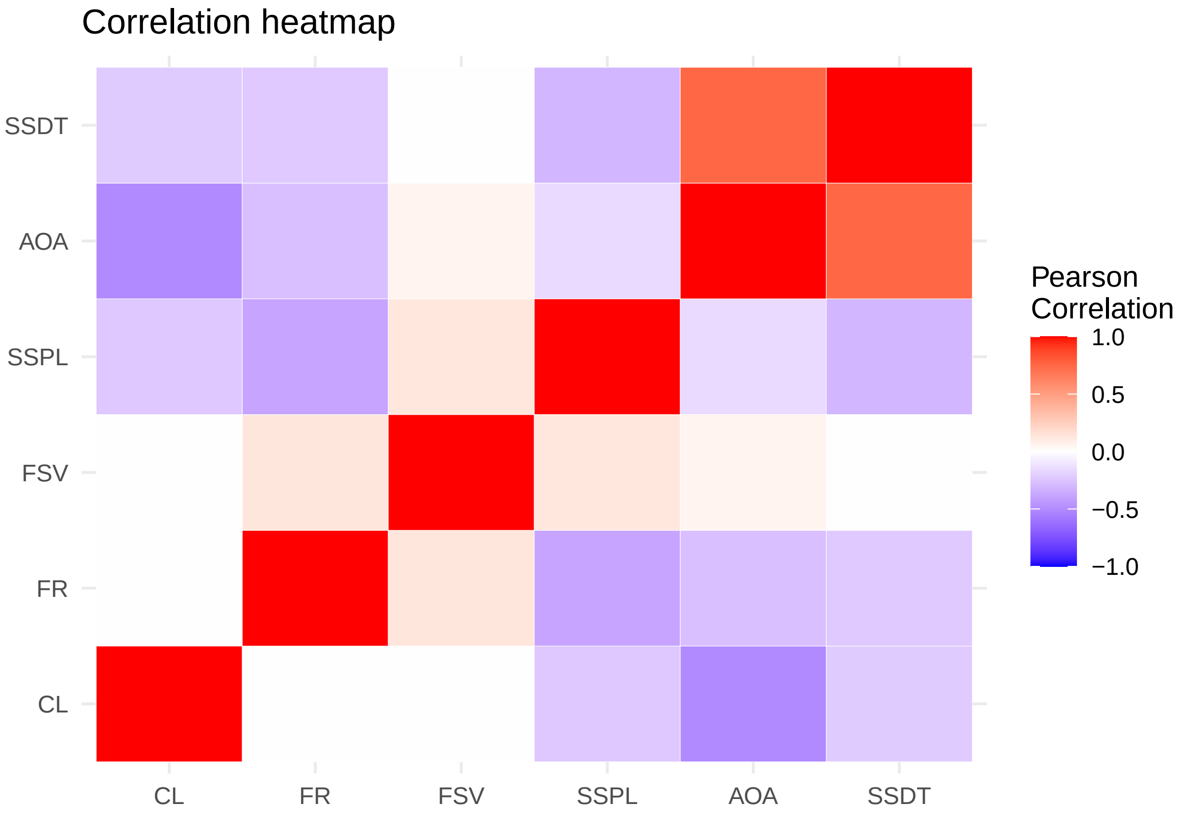
Pairs plot of response variable against all other variables

```
ggparcoord(airfoil[sample(order(airfoil),500),],
           title = "Parallel coordinate plot",
           scale="center") +
  scale_color_viridis(discrete = T) +
  theme_minimal() +
  theme(
    legend.position="none",
    plot.title = element_text(size=20)
  )
```



```
reorder_cormat <- function(data) {
  cormat <- cor(data)
  # Use correlation between variables as distance
  dd <- as.dist((1-cormat)/2)
  hc <- hclust(dd)
  cormat <-cormat[hc$order, hc$order]
}

ggplot(data = reshape2::melt(reorder_cormat(airfoil)), aes(x=Var1, y=Var2, fill=value)) +
  geom_tile(color="white") +
  xlab("") + ylab("") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
    name="Pearson\nCorrelation") +
  theme_minimal()+
  labs(title="Correlation heatmap")
```


Correlation heatmap
SSDT
AOA
SSPL
FSV
FR
CL
CL
FR
FSV
SSPL
AOA
SSDT
Pearson Correlation
1.0
0.5
0.0
−0.5
−1.0