460 CW Grp 10 - Exploratory analysis for Lubrizol

Alex Meehan

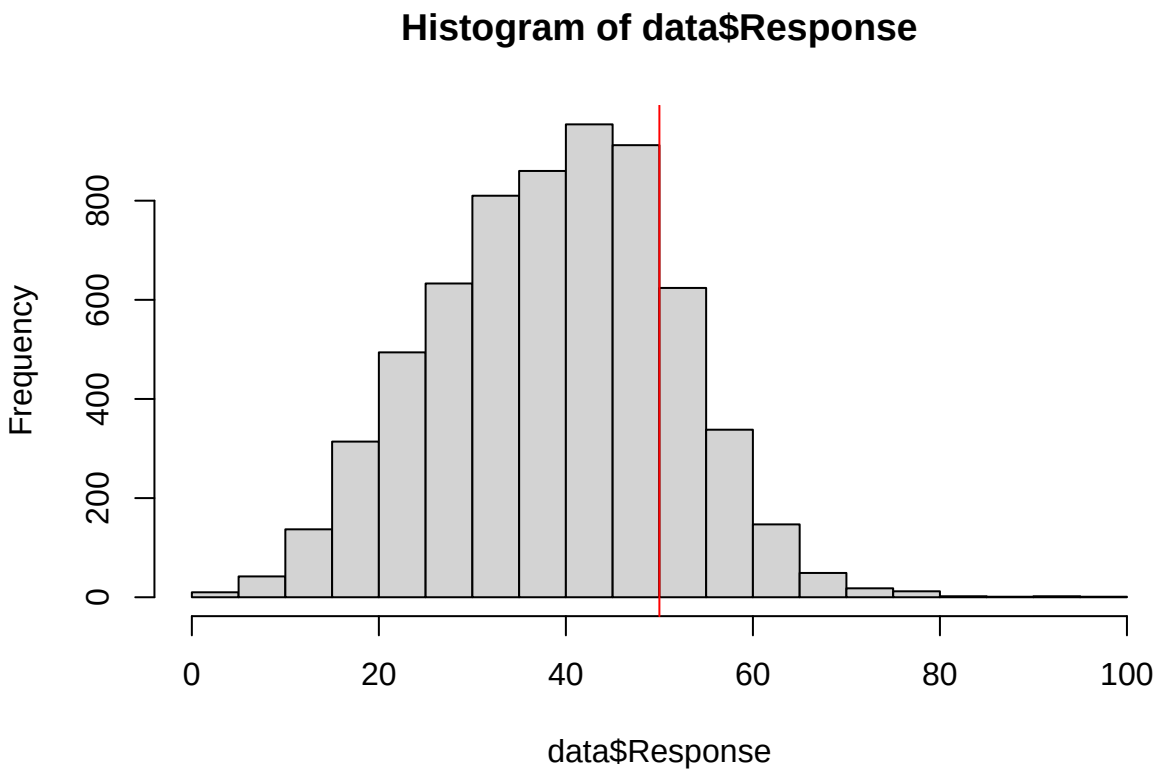
```
library(dplyr)
library(ggplot2)
```

import the data

```
data <- read.csv("../data/Data.csv", stringsAsFactors=FALSE)
data$LAB <- as.factor(data$LAB) # convert LAB to be a factor
data[is.na(data)] <- 0 # replace NAs with zero
```

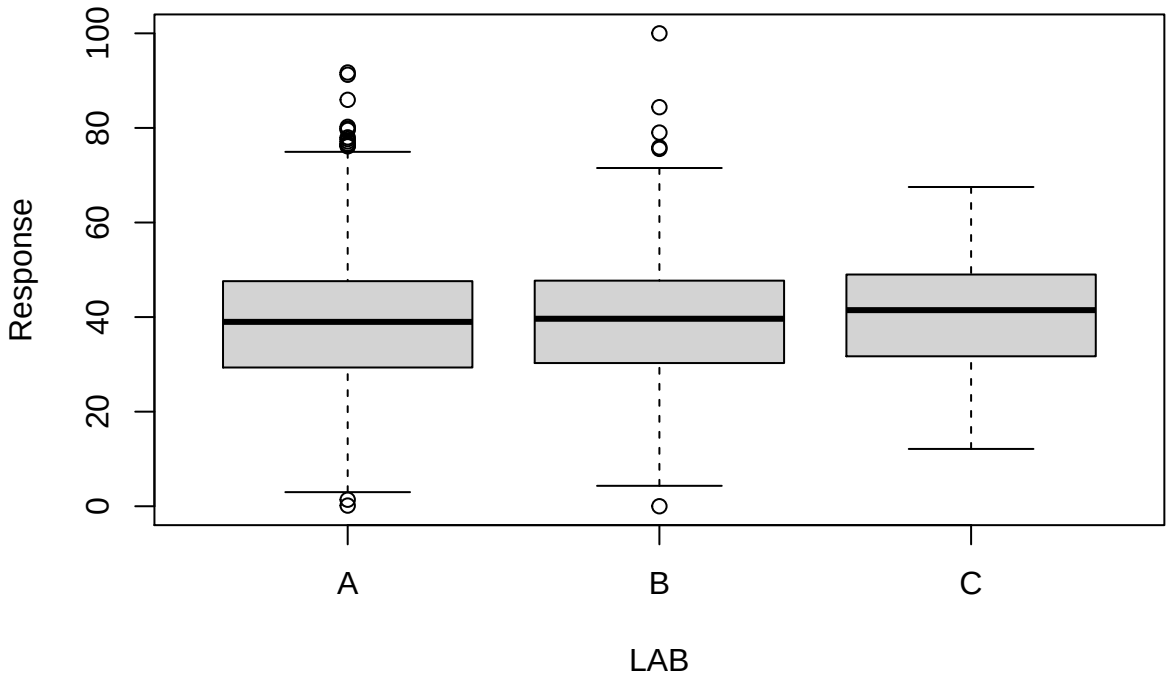
distribution of target variable ("Response")

```
hist(data$Response, breaks=20)
abline(v=50,col="red")
```



quick look at results by lab

```
boxplot(Response~LAB, data=data)
```




```
table(data$LAB)
```

```
##
##    A    B    C
## 4453 1591  316
```

find highest correlation sub-groups & meta-groups

```
corr_data <- data
corr_data$X <- NULL # drop identifier column
corr_data$LAB <- NULL # drop non-numeric LAB column
corr_data <- corr_data[1:nrow(corr_data),477:ncol(corr_data)]
# drop ingredients columns
corr_data<- scale(corr_data) # standardise the variables (for aggregation in meta-groups)
corr <- cor(corr_data, method="pearson") # calculates full corr matrix
corr <- corr[nrow(corr),1:ncol(corr)] # selects the last row (Response)
corr <- data.frame(colnames(corr_data), corr) #add back column names
corr <- corr[1:(nrow(corr)-1), ] # remove last "Response" row (correlation of 1.0)
corr <- corr[order(-abs(corr$corr)),]   # order correlation list from largest
colnames(corr) <- c("sub_grp", "pearson_corr")
library(stringr)
corr$meta_grp <- str_extract(corr$sub_grp, "^(.*?)_") # create meta-groups
corr_meta <- corr %>%
```

```
  group_by(meta_grp) %>%
  summarize(avg_corr = mean(pearson_corr))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
corr_meta <- corr_meta[order(-abs(corr_meta$avg_corr)),] # order correlation grp from largest
```