01- Data Exploration

Directory and doc rules

```
knitr::opts_chunk$set(
  echo = TRUE,         # Display code chunks
  eval = TRUE,         # Evaluate code chunks
  warning = FALSE,     # Hide warnings
  message = FALSE,     # Hide messages
  fig.width = 20,       # Set plot width in inches
  fig.height = 9,      # Set plot height in inches
  fig.align = "center" # Align plots to the center
)
```

Load packages

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

Load data

Weight = mg
Length, width, height = mm
p450, SOD = activity/ (mg/protein)
Condition factor, economic factor = unitless

```
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")

#alldata has the site names, biomarkers, condition factor, average thickness and analyte data - each ro
alldata<- read.csv("/Users/cmantegna/Documents/WDFWmussels/data/alldata.csv")
```

fix zero's in the data frame and alldata frame

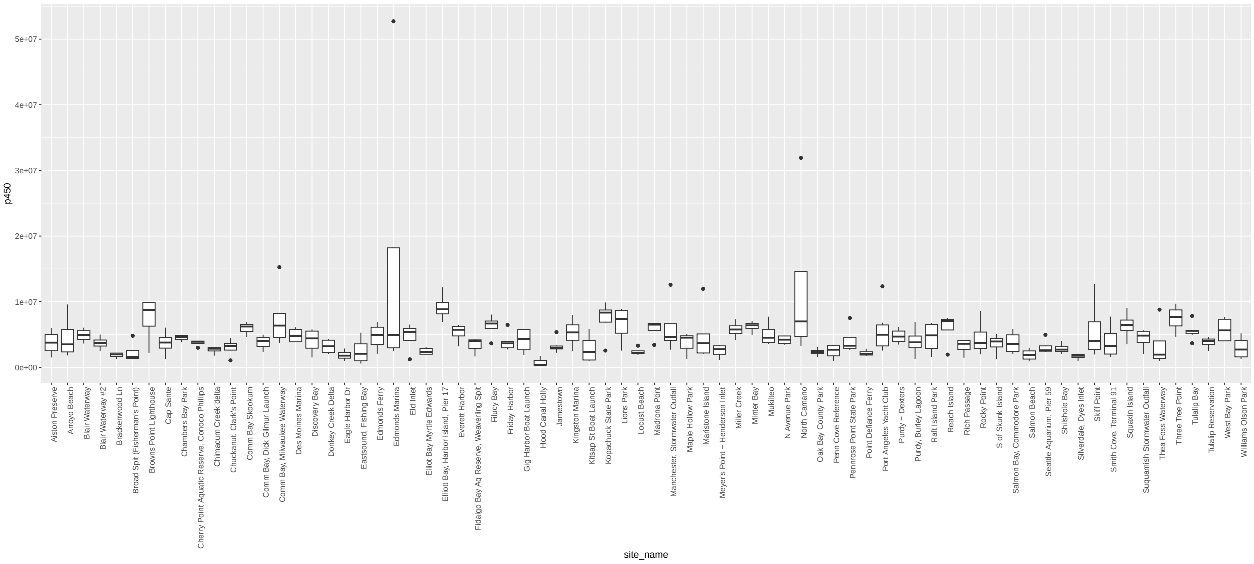
```
# Data contains 0's and must be adjusted in this order to preserve all usable data.

#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
alldata$SOD[alldata$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
alldata$p450[alldata$p450 <= 0] <- NA

#write.csv(alldata, "/Users/cmantegna/Documents/WDFWmussels/data/alldata.csv")
```

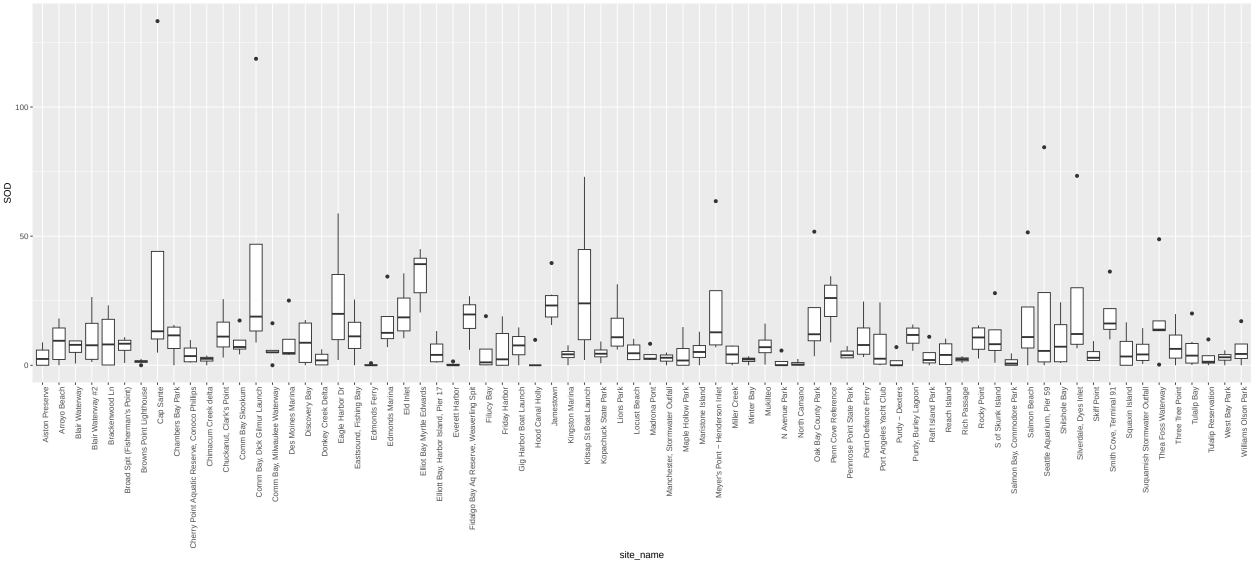
```
# check the data frame
#summary(alldata)
#str(alldata)
```

Boxplot of biomarker data, p450 and SOD

```
#p450
pplot<- ggplot(data, aes(x = site_name, y = p450)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

print(pplot)
```



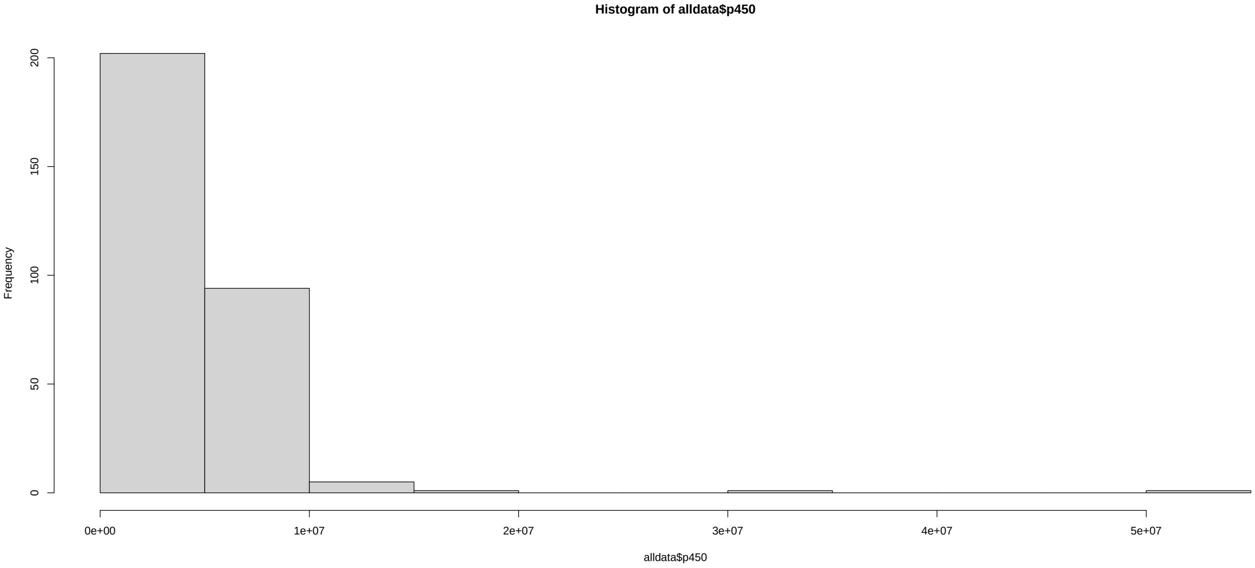
```
#SOD
splot<- ggplot(data, aes(x = site_name, y = SOD)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

print(splot)
```



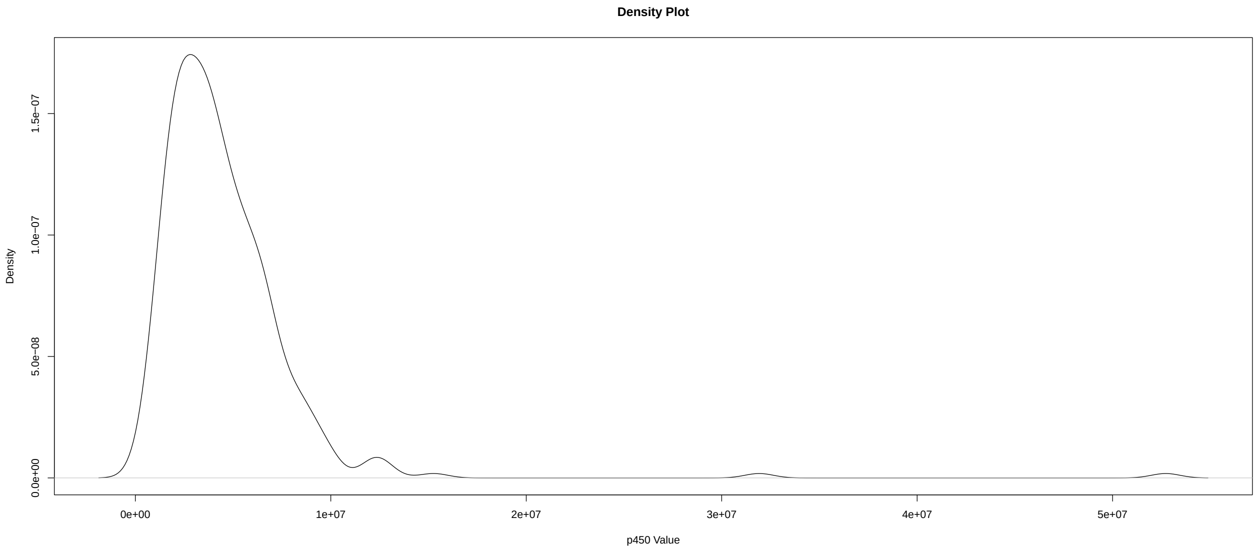
Histograms of p450, SOD, condition_factor and avg_thickness

Only avg_thickness looks normally distributed.

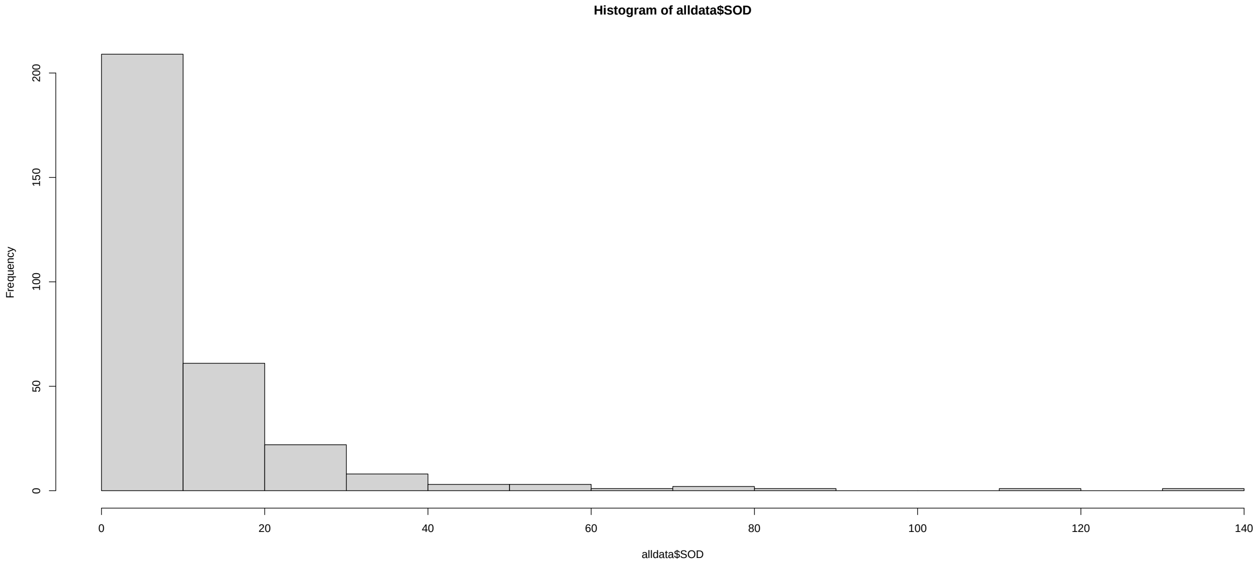
```
# p450 basic histogram + basic density plot
hist(alldata$p450)
```



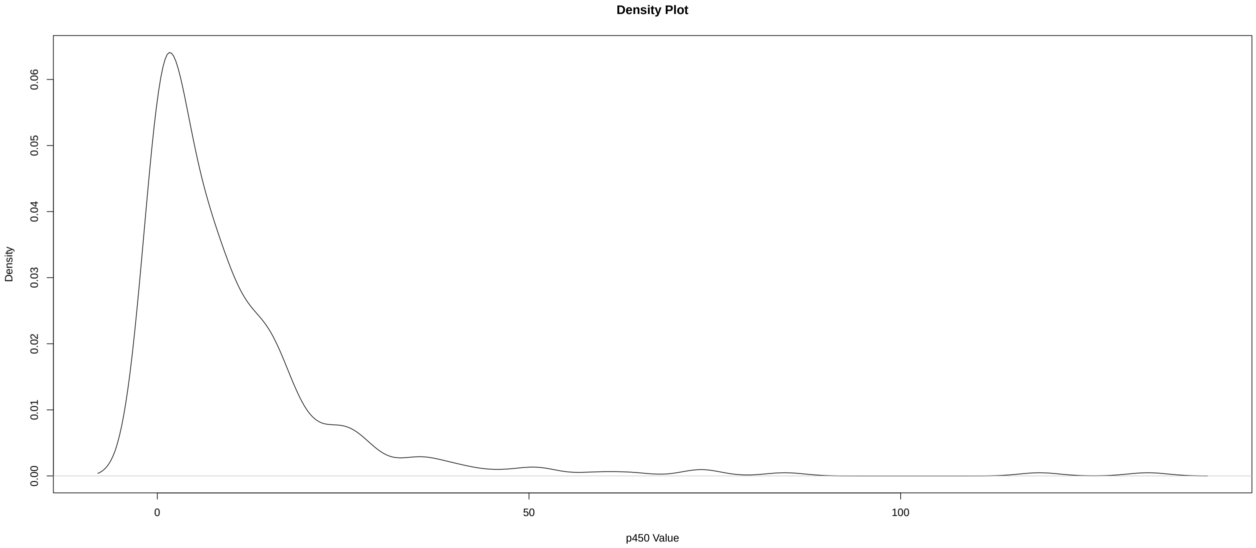
```
plot(density(alldata$p450, na.rm= TRUE), main="Density Plot", xlab="p450 Value")
```



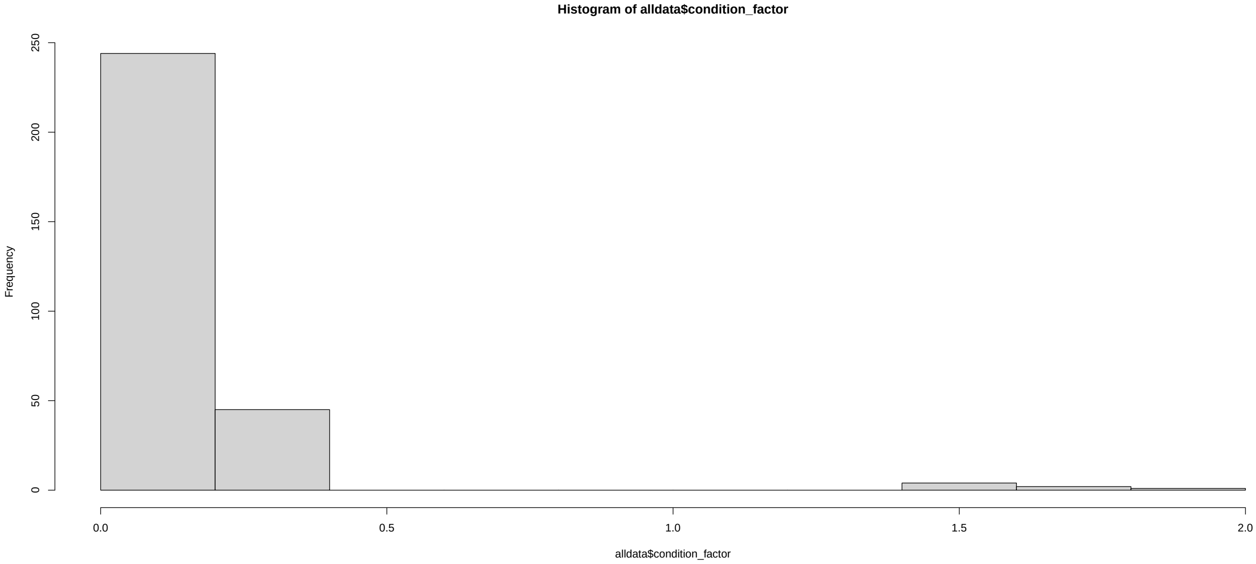
```
# SOD basic histogram + basic density plot
hist(alldata$SOD)
```



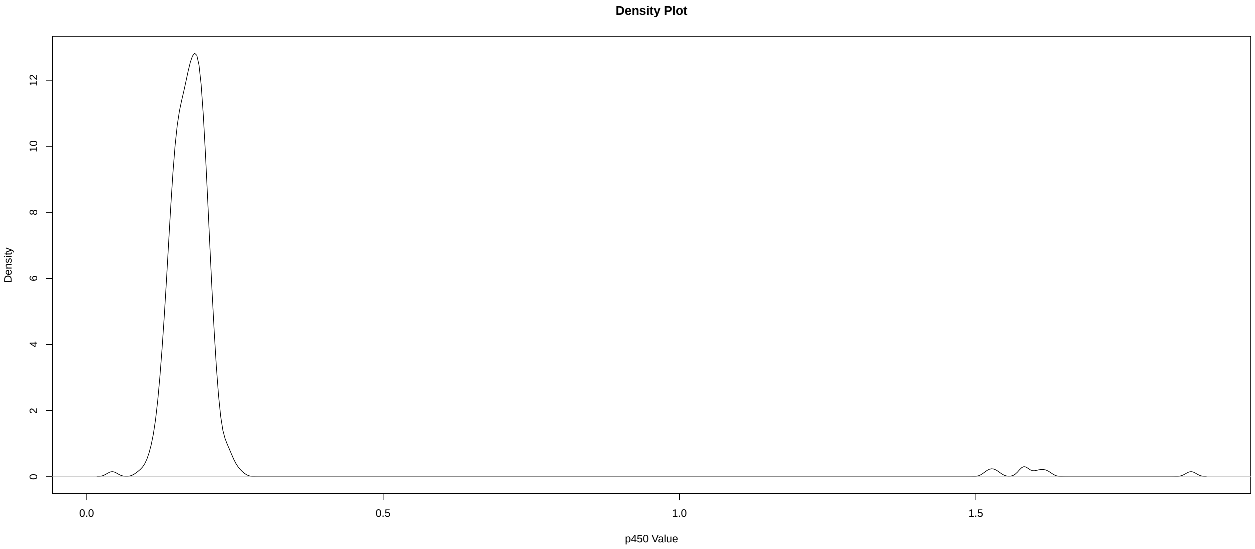
```
plot(density(alldata$SOD, na.rm= TRUE), main="Density Plot", xlab="p450 Value")
```



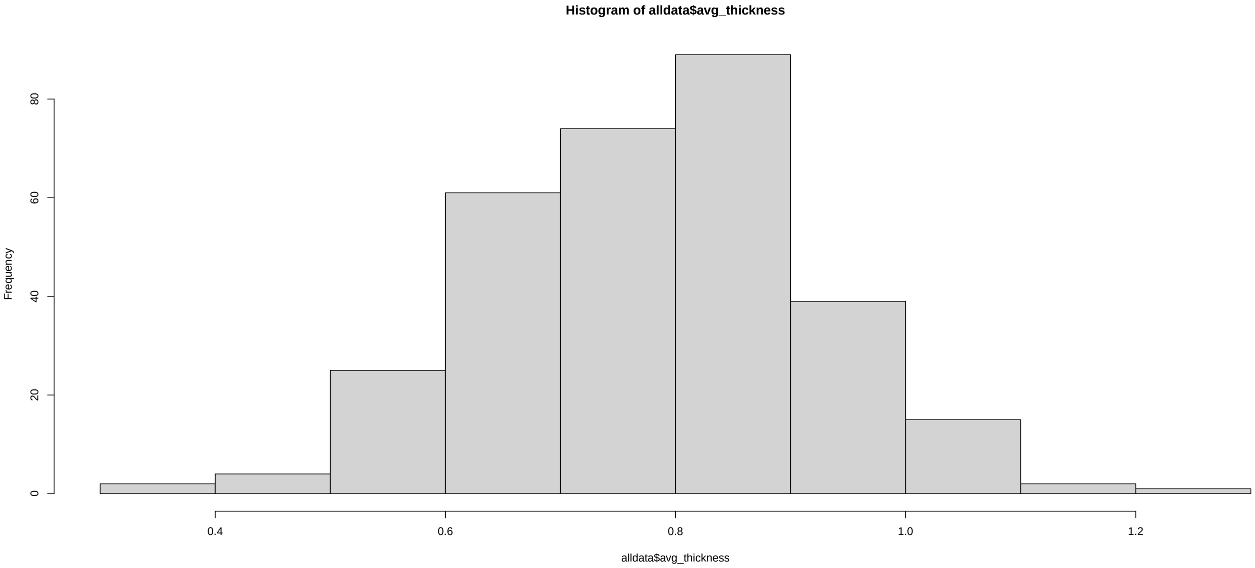
```r
# Condition_factor basic histogram + basic density plot
hist(alldata$condition_factor)
```



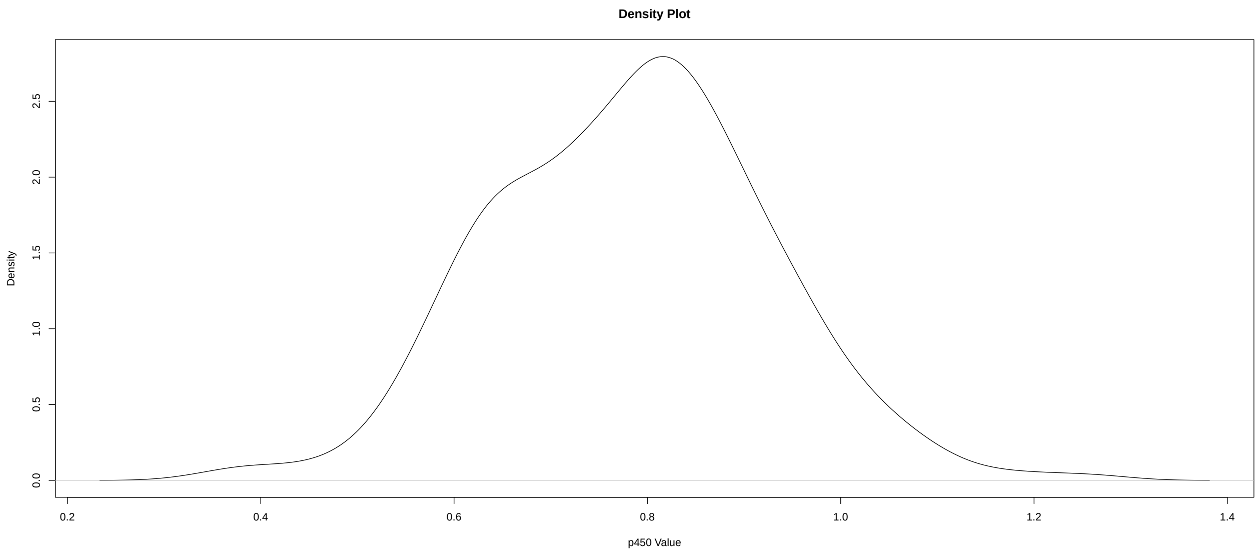
```r
plot(density(alldata$condition_factor, na.rm= TRUE), main="Density Plot", xlab="p450 Value")
```



```r
# Avg_thickness basic histogram + basic density plot
hist(alldata$avg_thickness)
```



```r
plot(density(alldata$avg_thickness, na.rm= TRUE), main="Density Plot", xlab="p450 Value")
```



Shapiro-Wilkes test for normality

Only avg_thickness is normally distributed.

```
shapiro.test(alldata$p450)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  alldata$p450
## W = 0.55896, p-value < 2.2e-16
```

```
shapiro.test(alldata$SOD)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  alldata$SOD
## W = 0.61784, p-value < 2.2e-16
```

```
shapiro.test(alldata$condition_factor)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  alldata$condition_factor
## W = 0.23346, p-value < 2.2e-16
```

```
shapiro.test(alldata$avg_thickness)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  alldata$avg_thickness
## W = 0.99674, p-value = 0.7814
```

Kruskal-Wallis, p450

p450 and site have a statistically significant relationship, p< 0.0000008077

```r
kruskal.test(p450 ~ site_name, data = alldata)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  p450 by site_name
## Kruskal-Wallis chi-squared = 137.23, df = 73, p-value = 8.077e-06
```

```r
kruskal.test(p450 ~ SOD, data = alldata)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  p450 by SOD
## Kruskal-Wallis chi-squared = 252.32, df = 254, p-value = 0.5179
```

```r
kruskal.test(p450 ~ condition_factor, data = alldata)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  p450 by condition_factor
## Kruskal-Wallis chi-squared = 259.49, df = 255, p-value = 0.4102
```

```r
kruskal.test(p450 ~ avg_thickness, data = alldata)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  p450 by avg_thickness
## Kruskal-Wallis chi-squared = 106.49, df = 110, p-value = 0.577
```

Kruskal-Wallis, SOD

SOD and site have a statistically significant relationship, p< 0.0000005669

```r
kruskal.test(SOD ~ site_name, data = alldata)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  SOD by site_name
## Kruskal-Wallis chi-squared = 138.64, df = 73, p-value = 5.669e-06
```

```r
kruskal.test(SOD ~ p450, data = alldata)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  SOD by p450
## Kruskal-Wallis chi-squared = 303, df = 303, p-value = 0.4892
```

```r
kruskal.test(SOD ~ condition_factor, data = alldata)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  SOD by condition_factor
## Kruskal-Wallis chi-squared = 250.73, df = 259, p-value = 0.6323
```

```r
kruskal.test(SOD ~ avg_thickness, data = alldata)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  SOD by avg_thickness
## Kruskal-Wallis chi-squared = 107.24, df = 110, p-value = 0.5567
```

Post hoc test. Kruskal-Wallac Multiple Comparisons

p450 and site_name show no true differences despite K-W test result.

SOD and site_name show a true site difference between Elliott Bay, Mrytle Edwards and Hood Canal only.

```r
library(pgirmess)

mc_p450<- as.data.frame(kruskalmc(p450 ~ site_name, data = alldata, method = "bonferroni"))
mc_SOD<- as.data.frame(kruskalmc(SOD ~ site_name, data = alldata, method = "bonferroni"))
```