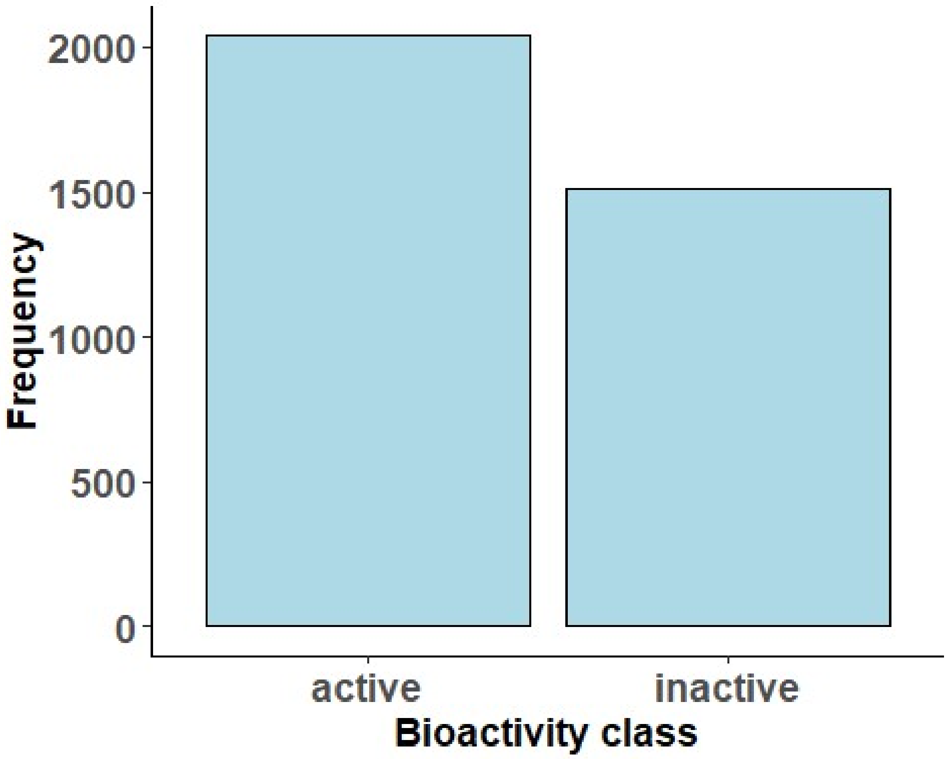
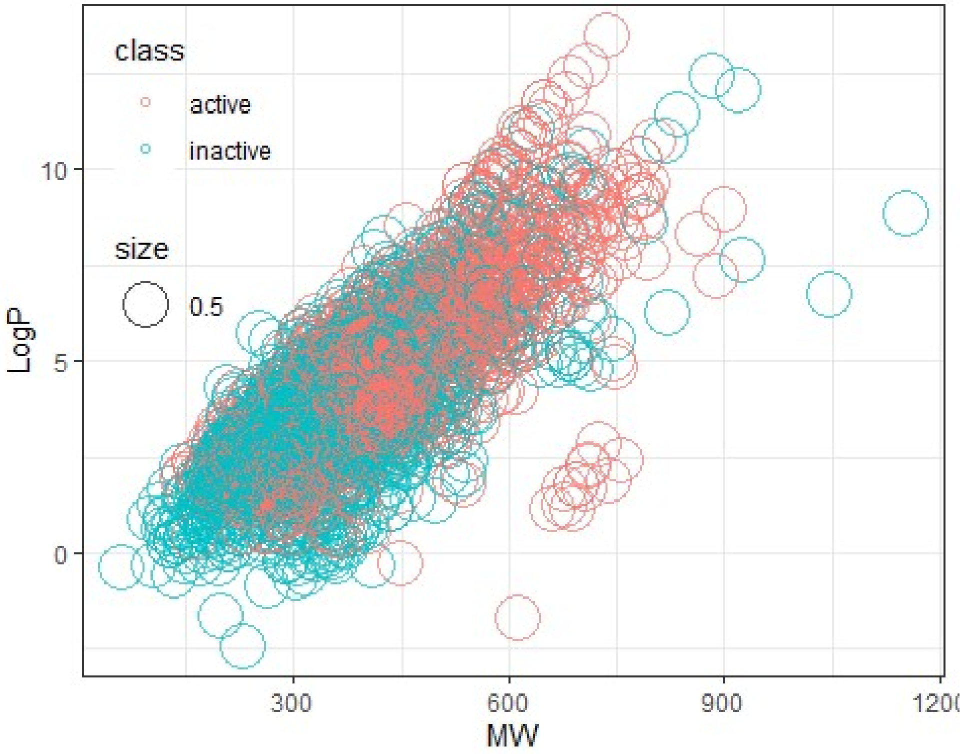
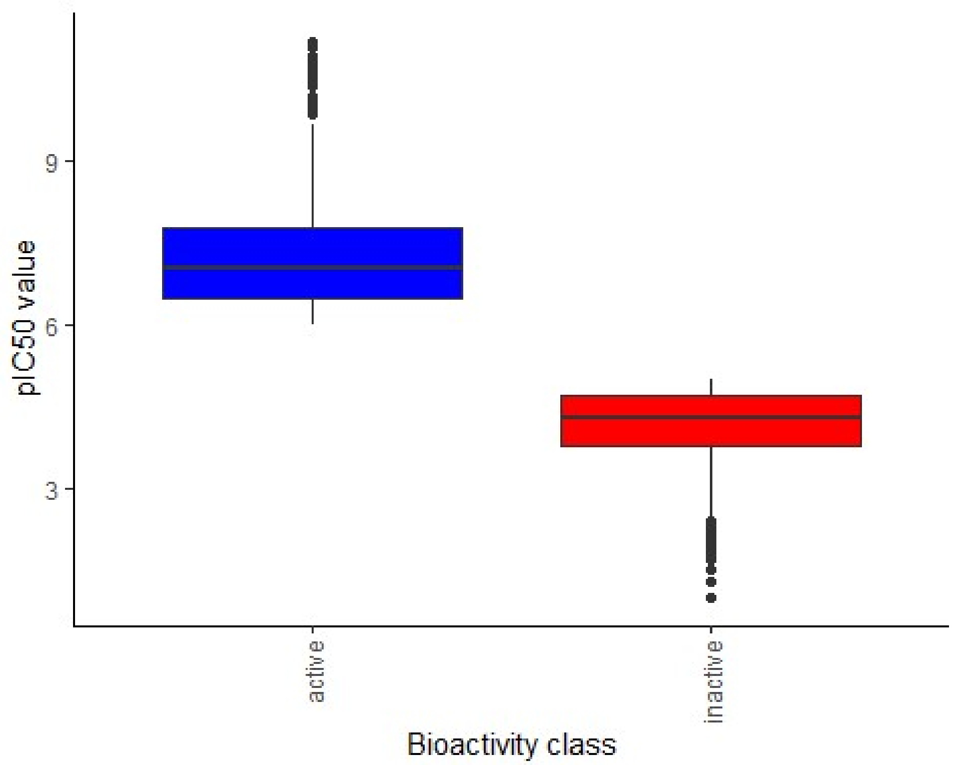
```
#loading the data
setwd("C:/Users/Icy/Documents/School-GMU/BINF 702/Project files")
Ach_data <-
read.csv("acetylcholinesterase_05_bioactivity_data_2class_pIC50.csv")

head(Ach_data)

##   X molecule_chembl_id                                  canonical_smiles
## 1 0       CHEMBL133897             CCOc1nn(-c2cccc(OCc3ccccc3)c2)c(=O)o1
## 2 1       CHEMBL336398        O=C(N1CCCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC1CC1
## 3 2       CHEMBL131588 CN(C(=O)n1nc(-c2ccc(Cl)cc2)nc1SCC(F)(F)F)c1ccccc1
## 4 3       CHEMBL130628     O=C(N1CCCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC(F)(F)F
## 5 4       CHEMBL130478         CSc1nc(-c2ccc(OC(F)(F)F)cc2)nn1C(=O)N(C)C
## 6 6       CHEMBL130098                CSc1nc(-c2ccc(Cl)cc2)nn1C(=O)N(C)C
##      class      MW   LogP NumHDonors NumHAcceptors    pIC50
## 1   active 312.325 2.8032          0             6 6.124939
## 2   active 376.913 4.5546          0             5 7.000000
## 3 inactive 426.851 5.3574          0             5 4.301030
## 4   active 404.845 4.7069          0             5 6.522879
## 5   active 346.334 3.0953          0             6 6.096910
## 6   active 296.783 2.8501          0             5 7.000000

#test data for normality
Ach_norm<-rnorm(100)
shapiro.test(Ach_norm) # we cannot reject the hypothesis that the data is
normally distributed

##
##  Shapiro-Wilk normality test
##
## data:  Ach_norm
## W = 0.98641, p-value = 0.3993

#Create a frequency plot of the 2 bioactivity classes
library(rlang)

## Warning: package 'rlang' was built under R version 4.2.3

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.3

ggplot(Ach_data, aes(x = class)) +
  geom_bar(color = "black", fill = "lightblue") +
  labs(x = "Bioactivity class", y = "Frequency") +
  theme_classic() +
  theme(axis.text = element_text(size = 14, face = "bold"),
        axis.title = element_text(size = 14, face = "bold"),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        legend.position = "none")
```



```
#scatterplot of MW versus logP

ggplot(Ach_data, aes(x=MW, y=LogP, color=class, size=0.5)) +
  geom_point(alpha=0.7, shape=1, aes(fill=class), show.legend=TRUE,
stroke=0.5) +
  scale_size_continuous(range = c(2,10)) +
  labs(x = "MW", y = "LogP") +
  theme_bw() +
  theme(legend.position = c(0.02,0.98), legend.justification = c(0, 1),
legend.background = element_blank(), legend.key.width = unit(0.8,"cm"))
```



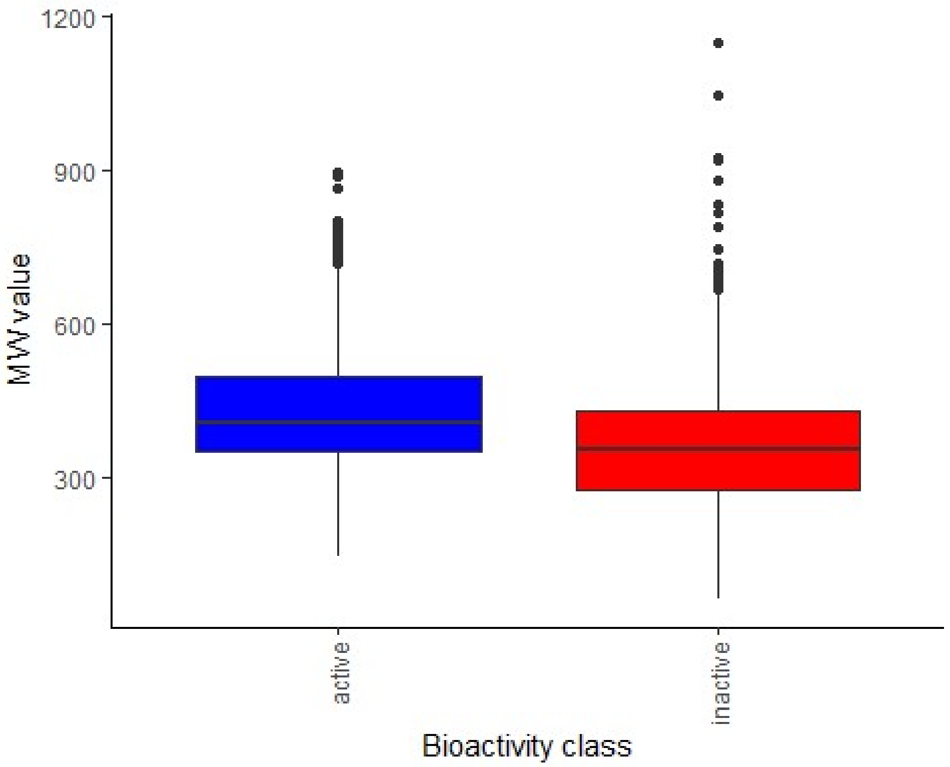
```
#box plots of pIC50 values versus bioactivity class

ggplot(data=Ach_data, aes(x=class, y=pIC50,fill=class)) +
  geom_boxplot() +
  scale_fill_manual(values=c("blue", "red")) +
  labs(x="Bioactivity class", y="pIC50 value") +
  theme_classic() +
  theme(axis.text.x = element_text(angle=90, vjust=0.5, hjust=1),
        legend.position="none")
```



```r
#mannwhitney U test. Non parametric stat significant tests

mannwhitney <- function(descriptor, verbose=FALSE) {
  # actives and inactives
  active <- Ach_data[Ach_data$class == "active", descriptor]
  inactive <- Ach_data[Ach_data$class == "inactive", descriptor]

  # compare samples
  res <- wilcox.test(active, inactive)

  # interpret
  alpha <- 0.05
  if (res$p.value > alpha) {
    interpretation <- "Same distribution (fail to reject H0)"
  } else {
    interpretation <- "Different distribution (reject H0)"
  }

  # print results
  if (verbose) {
    cat(paste("Descriptor:", descriptor, "\n"))
    cat(paste("Statistics:", res$statistic, "\n"))
    cat(paste("p-value:", res$p.value, "\n"))
    cat(paste("alpha:", alpha, "\n"))
    cat(paste("Interpretation:", interpretation, "\n"))
  }
  # return results
  return(data.frame(Descriptor=descriptor, Statistics=res$statistic,
                    p=res$p.value, alpha=alpha,
                    Interpretation=interpretation))
}


mannwhitney("pIC50")
```

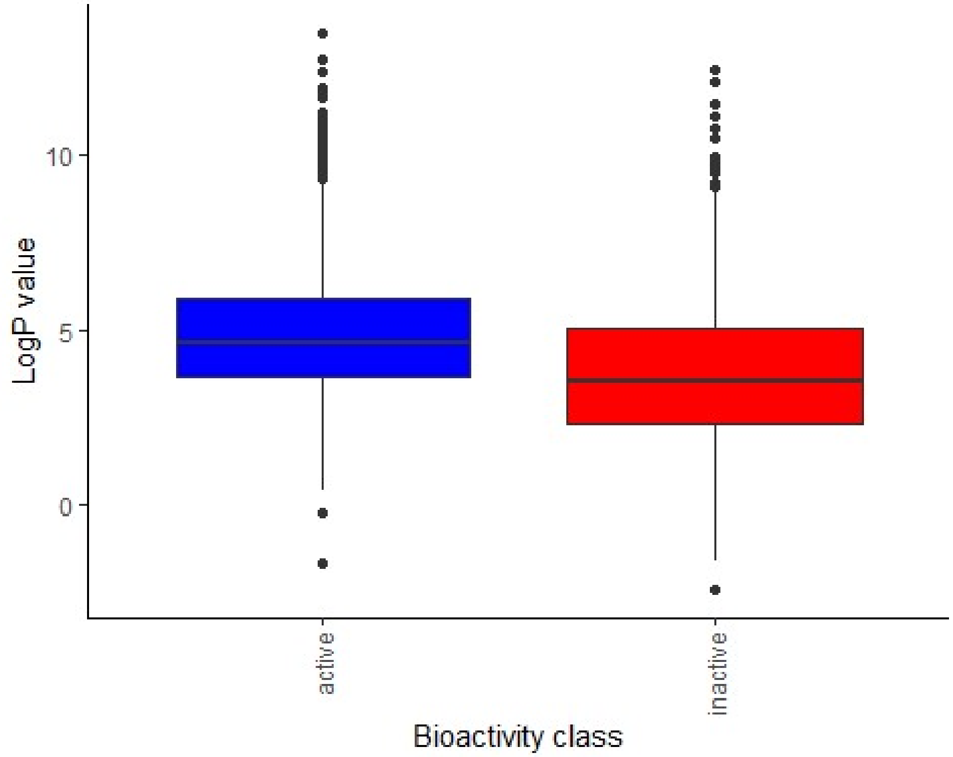
```
##   Descriptor Statistics p alpha                     Interpretation
## W      pIC50    3078890 0  0.05 Different distribution (reject H0)
```

```
#MW boxplot versus bioactivity class
ggplot(data=Ach_data, aes(x=class, y=MW,fill=class)) +
  geom_boxplot() +
  scale_fill_manual(values=c("blue", "red")) +
  labs(x="Bioactivity class", y="MW value") +
  theme_classic() +
  theme(axis.text.x = element_text(angle=90, vjust=0.5, hjust=1),
        legend.position="none")
```

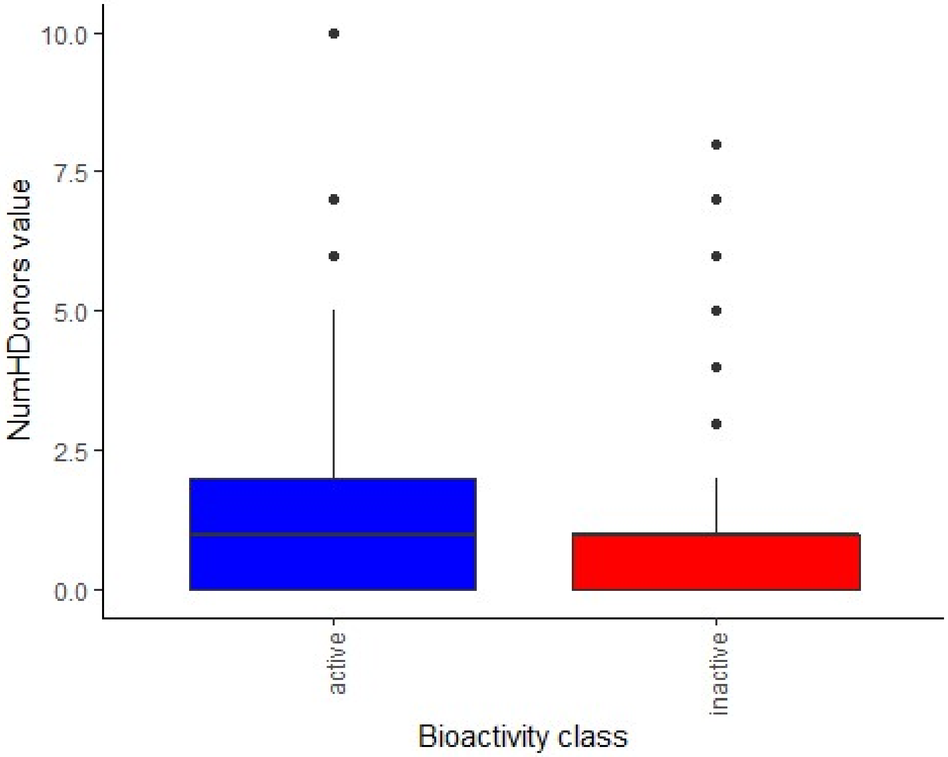


```
#mannwhitney test in MW

mannwhitney("MW")
```

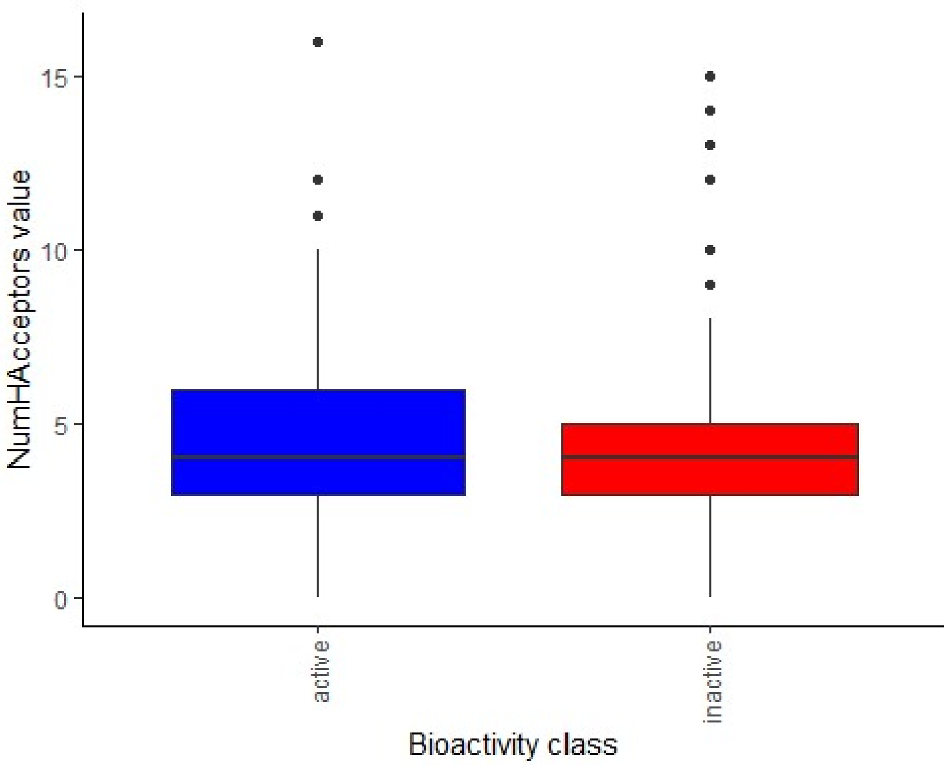
```
##   Descriptor Statistics            p alpha
Interpretation
## W         MW    2020111 4.144511e-57  0.05 Different distribution (reject
H0)

#boxplot of logP versus bioactivity class
ggplot(data=Ach_data, aes(x=class, y=LogP,fill=class)) +
  geom_boxplot() +
  scale_fill_manual(values=c("blue", "red")) +
  labs(x="Bioactivity class", y="LogP value") +
  theme_classic() +
  theme(axis.text.x = element_text(angle=90, vjust=0.5, hjust=1),
        legend.position="none")
```



```
#mannwhitney test in LogP

mannwhitney("LogP")

##   Descriptor Statistics            p alpha
Interpretation
## W       LogP    2036990 4.637335e-61  0.05 Different distribution (reject
H0)

#Now looking at NumH donors versus bioactivity class
ggplot(data=Ach_data, aes(x=class, y=NumHDonors,fill=class)) +
  geom_boxplot() +
  scale_fill_manual(values=c("blue", "red")) +
  labs(x="Bioactivity class", y="NumHDonors value") +
  theme_classic() +
  theme(axis.text.x = element_text(angle=90, vjust=0.5, hjust=1),
        legend.position="none")
```



```
#mannwhitney test on NumHdonors

mannwhitney("NumHDonors")

##   Descriptor Statistics            p alpha
Interpretation
## W NumHDonors    1717885 5.040191e-10  0.05 Different distribution (reject
H0)

#NumHAcceptors
ggplot(data=Ach_data, aes(x=class, y=NumHAcceptors,fill=class)) +
  geom_boxplot() +
  scale_fill_manual(values=c("blue", "red")) +
  labs(x="Bioactivity class", y="NumHAcceptors value") +
  theme_classic() +
  theme(axis.text.x = element_text(angle=90, vjust=0.5, hjust=1),
        legend.position="none")
```



```
#mannwhitney test on NumHAcceptors
mannwhitney("NumHAcceptors")

##      Descriptor Statistics            p alpha
## W NumHAcceptors    1671318 8.850162e-06  0.05
##                       Interpretation
## W Different distribution (reject H0)
```