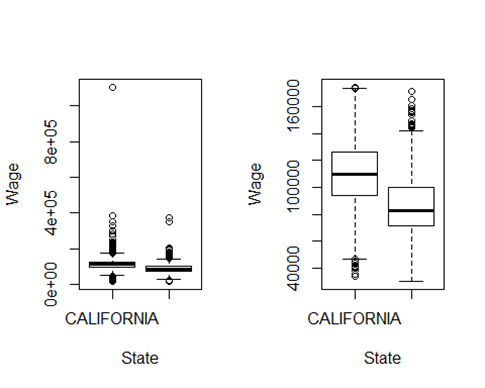
**Predictive Analysis**

**California & Texas wages PLOT**

* library(dplyr)
* library(ggplot2)
* dataset <- read.csv("c:/us\_perm\_visas.csv", header = TRUE, sep = ",",
* stringsAsFactors= FALSE)
* #Lets see sample data for the the dataset by running below command which selects every 5th row starting from 1st row.
* wages <- subset(dataset, wage\_offered\_unit\_of\_pay\_9089=="Year" & (job\_info\_work\_state =="CALIFORNIA" | job\_info\_work\_state =="TEXAS")& ( us\_economic\_sector=="IT" | us\_economic\_sector=="Finance"), select=c(wage\_offered\_from\_9089,job\_info\_work\_state,us\_economic\_sector))
* wages[wages==""] <- NA
* wages <- na.omit(wages)
* wages$wage\_offered\_from\_9089 <- as.numeric(gsub(",", "", wages$wage\_offered\_from\_9089))
* wages[, 'job\_info\_work\_state'] <- as.factor(wages[, 'job\_info\_work\_state'])
* par(mfrow=c(1,2))
* plot(wages$wage\_offered\_from\_9089 ~ wages$job\_info\_work\_state,xlab="State",ylab="Wage", title="State&Wage")
* wages <- wages[-which(wages$wage\_offered\_from\_9089 %in% boxplot.stats(wages$wage\_offered\_from\_9089)$out), ]
* plot(wages$wage\_offered\_from\_9089 ~ wages$job\_info\_work\_state,xlab="State",ylab="Wage", title="Boxpot - State&Wage")
* 

A graph shows that California wages higher than Texas.  These graphs show remove outliner’s importance since the second graph indicates clear information.

Fınance & IT PLOT

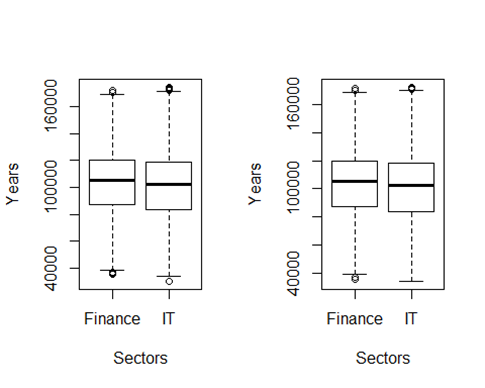
wages[, 'us\_economic\_sector'] <- as.factor(wages[, 'us\_economic\_sector'])

par(mfrow=c(1,2))

plot(wages$wage\_offered\_from\_9089 ~ wages$us\_economic\_sector,xlab="Sectors",ylab="Years", title="Years&Sector")

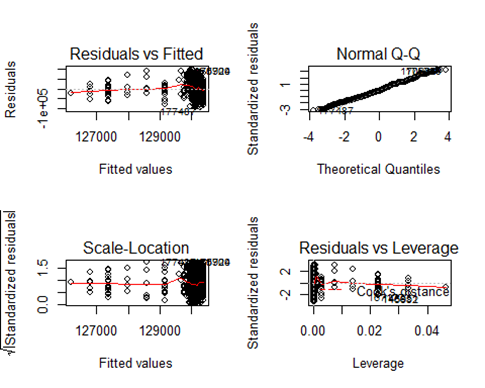
wages <- wages[-which(wages$wage\_offered\_from\_9089 %in% boxplot.stats(wages$wage\_offered\_from\_9089)$out), ]

plot(wages$wage\_offered\_from\_9089 ~ wages$us\_economic\_sector,xlab="Sectors",ylab="Years", title="Years&Sector-Boxplot")



The result suprises us becuase finance wages slightly higher than IT wages.

Salary & experıence prediction

wageYrs <- dataset[,c("job\_info\_alt\_cmb\_ed\_oth\_yrs","wage\_offer\_to\_9089")]  
wageYrs[wageYrs==""] <- NA  
wageYrs <- na.omit(wageYrs)  
wageYrs$wage\_offer\_to\_9089 <- as.numeric(gsub(",", "", wageYrs$wage\_offer\_to\_9089))  
#plot(wageYrs$wage\_offer\_to\_9089 ~ wageYrs$job\_info\_alt\_cmb\_ed\_oth\_yrs,xlab="Experience Years",ylab="Wages" )  
wageYrs <- wageYrs[-which(wageYrs$wage\_offer\_to\_9089 %in% boxplot.stats(wageYrs$wage\_offer\_to\_9089)$out), ]  
  
wageYrs.model <- lm(wage\_offer\_to\_9089 ~ job\_info\_alt\_cmb\_ed\_oth\_yrs, data=wageYrs)  
#abline(wageYrs.model)  
summary(wageYrs.model)  
##   
## Call:  
## lm(formula = wage\_offer\_to\_9089 ~ job\_info\_alt\_cmb\_ed\_oth\_yrs,   
##     data = wageYrs)  
##   
## Residuals:  
##    Min  1Q Median     3Q Max   
## -95280 -22130  -2830 24734 95020   
##   
## Coefficients:  
##                              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)                 130379.55 536.87 242.851 <2e-16 \*\*\*  
## job\_info\_alt\_cmb\_ed\_oth\_yrs    -49.94 82.26 -0.607 0.544      
## ---  
## Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 30320 on 6546 degrees of freedom  
## Multiple R-squared:  5.63e-05, Adjusted R-squared:  -9.645e-05   
## F-statistic: 0.3686 on 1 and 6546 DF,  p-value: 0.5438  
new.data <- data.frame(job\_info\_alt\_cmb\_ed\_oth\_yrs=c(10,15) )  
predict(wageYrs.model, new.data, interval="confidence")  
##        fit lwr      upr  
## 1 129880.1 128749.9 131010.3  
## 2 129630.4 127810.3 131450.5  
par(mfrow=c(2,2))  
plot(wageYrs.model)  


Since Pr(>|t|) is equal 0.554, the regression is insignificant. Normal Q-Q graph indicates dimensional line so it is successful.  Residuals vs Fitted graph’s line is not flat, so that shows insignificant importance.