**State\_level\_daily**

setwd("C:/Users/shravan19/Desktop/MBAproject/covid19") #setpath

state\_daily <- read.csv(file = "state\_level\_daily.csv",sep = ",") #loading data

library(dplyr)

library(ggplot2)

**exploring data**

head(state\_daily)

tail(state\_daily)

str(state\_daily)

**Missing values**

No missing values

**Removing negative values**

state\_daily <- filter(state\_daily, Confirmed > 0)

state\_daily <- filter(state\_daily, Death > 0)

state\_daily <- filter(state\_daily, Recovered > 0)

View(state\_daily)

**Histogram #use for numberic data**

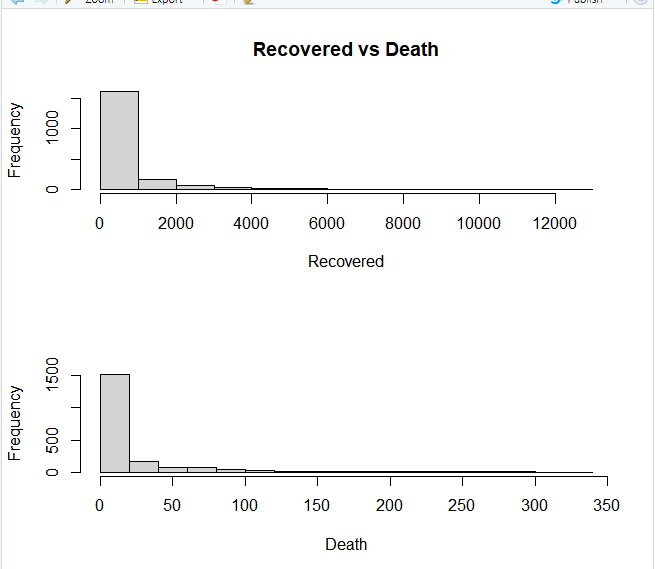
attach(plotdata)

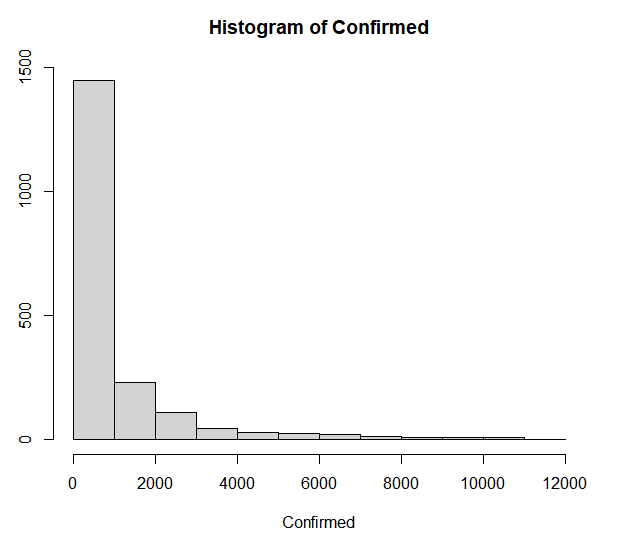
par(mfrow=c(2,1))

hist(Recovered,freq =TRUE)

hist(Death,freq =TRUE)

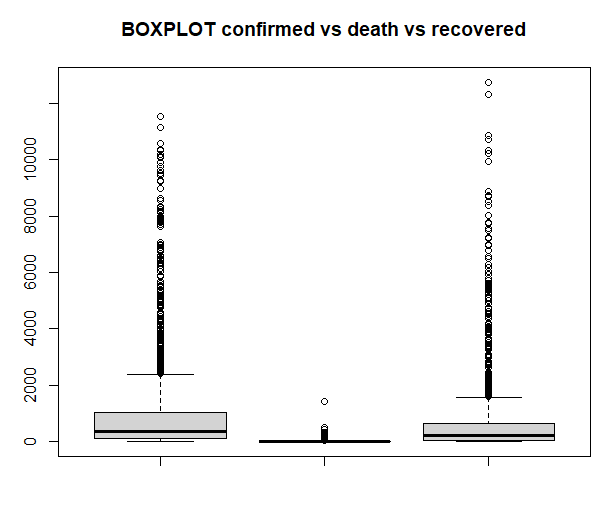
par(mfrow=c(1,1))





**BOXPLOT**

boxplot(state\_daily$Confirmed,state\_daily$Death,state\_daily$Recovered, main="BOXPLOT confirmed vs death vs recovered")



**Remove outlier**

xnew <- state\_daily

View(xnew)

check<- x[,-6] #to just check wheather there are more outlier or not

check<-check[,-8]

boxplot(check$Confirmed,check$Death,check$Recovered)

**To create a fix line to remove**

summary(xnew$Confirmed) #confirmed

IQR\_confirmed= 1034-125

upc=1034+1.5\*IQR\_confirmed #inter quartile range

upc

summary(xnew$Recovered)

IQR\_recover= 662-61

upc1=662+1.5\*IQR\_recover

upc1

summary(xnew$Death)

IQR\_death= 18-2

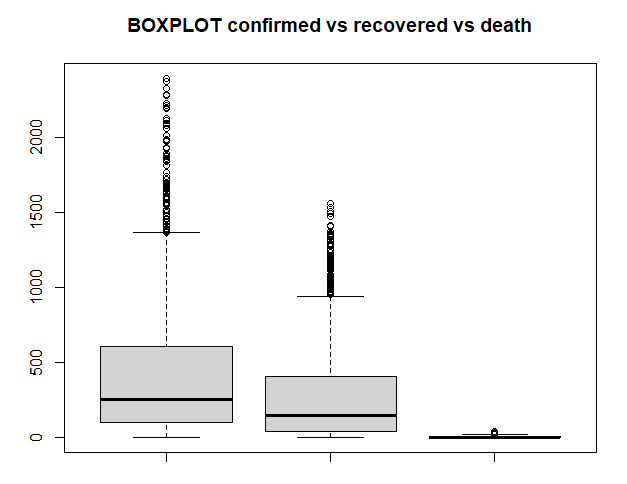
upc3=18+1.5\*IQR\_death

upc3

xnew1=subset(xnew,Confirmed<=2397.5 & Death<=42 & Recovered<=1563.5)

boxplot with no outliers

boxplot(xnew1$Confirmed,xnew1$Recovered,xnew1$Death)



**GGNEWPLOT**

setwd("C:/Users/shravan19/Desktop/MBAproject/covid19")

state\_update <- read.csv(file = "state\_update.csv",header = T,sep = ",")

View(state\_update)

attach(state\_update)

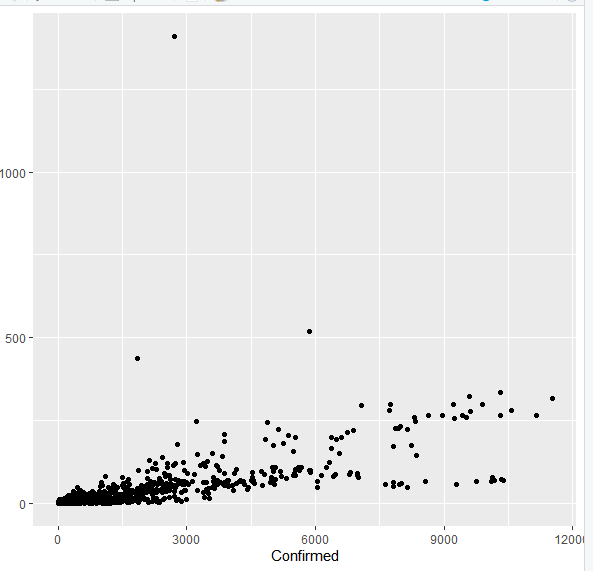
library(ggplot2)

ggplot(state\_update, mapping = aes(x=Confirmed, y=Death)) + geom\_point()

ggplot(state\_update, mapping = aes(x= Confirmed, y= Death, color = Recovered)) + geom\_point(alpha = .8,size = 3 ) + geom\_smooth(method = "lm",se = FALSE,size = 1.5)

**GGPLOT**

ggplot(state\_daily, mapping = aes(x= Confirmed, y= Death)) + geom\_point()



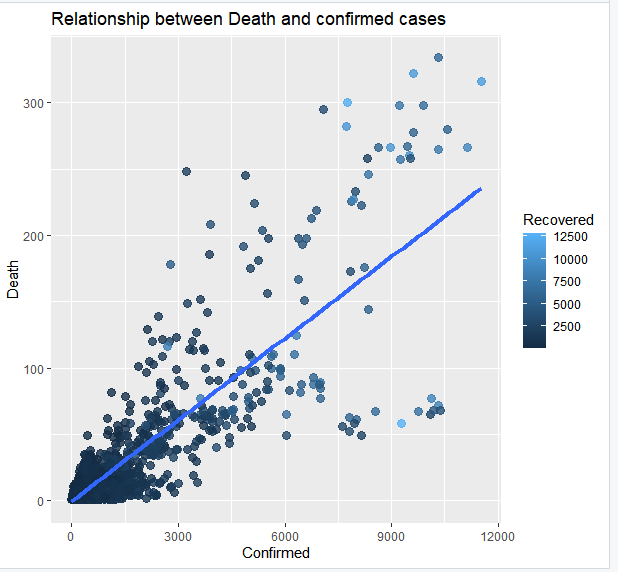
Removing outlier

plotdata <- filter(state\_daily,Death < 400 )

#ggplot

ggplot(plotdata, mapping = aes(x= Confirmed, y= Death, color = Recovered)) + geom\_point(alpha = .8,size = 3 ) + geom\_smooth(method = "lm",se = FALSE,size = 1.5)

#se = confidence #Alpha= transperancy , lm = linear regression line , linear model



Here we can see at the start phase there were too many confirmed and death cases as time passes confirmed cases start decreasing slowly and the no of death rate is also decreased. here the major role is the recovery rate is more.

# Clean Data #

library(dplyr)

library(ggplot2)

library(stats)

setwd("C:/Users/shravan19/Desktop/MBAproject/covid19/changes")

state\_update <- read.csv(file = "state\_update.csv",header = T,sep = ",")

#Perform EDA Exploratory data analysis #

View(state\_update)

attach(state\_update)

# DATA SUMMARISATION #

summary(state\_update) #find mean median mode

# DATA VISUALISATION #

#Histogram for continous data

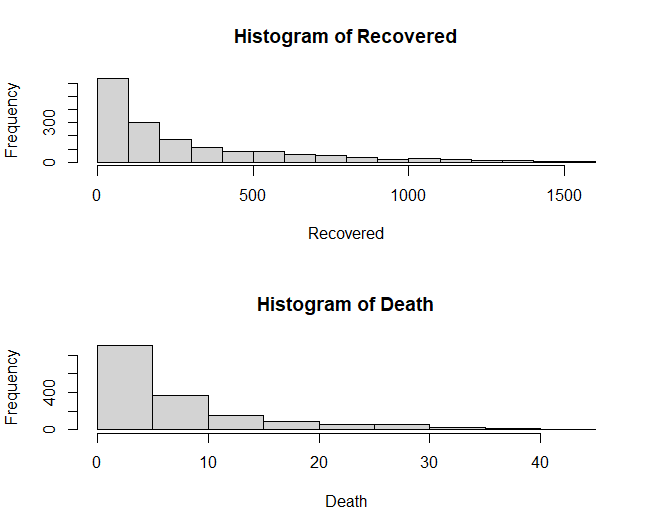
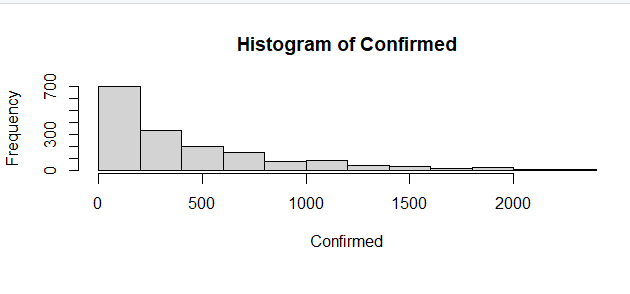
par(mfrow=c(2,1))

hist(Recovered,freq =TRUE)

hist(Death,freq =TRUE)

hist(Confirmed,freq = TRUE)

par(mfrow=c(1,1))

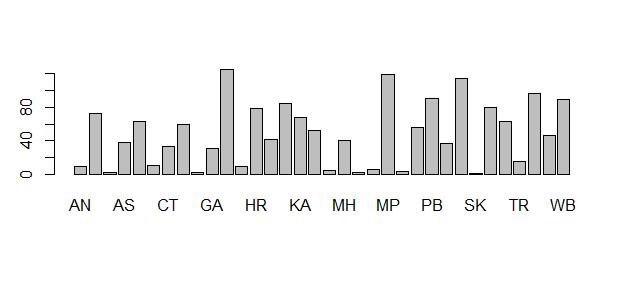


# Bar plot for categorical data

table(State\_name) #check unique values

table(State)

barplot(table(State))

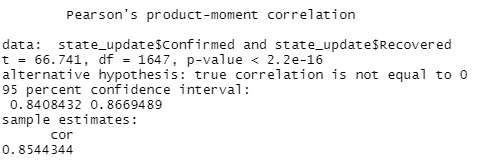


#linear regression model#

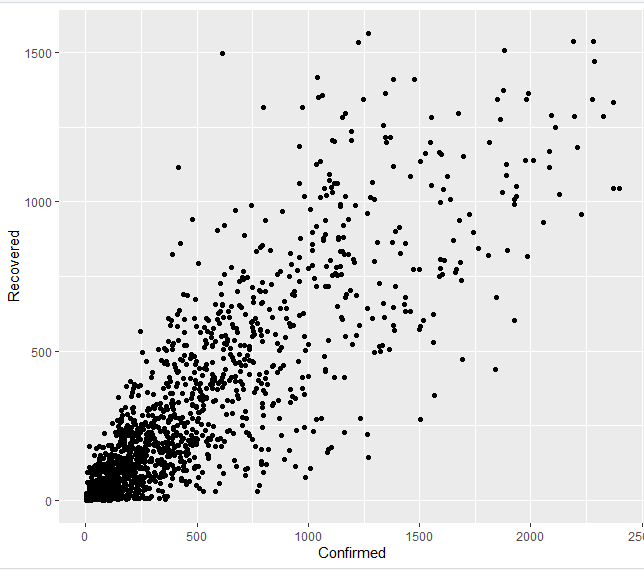
#prediction model#

# cor.test use for numeric #

cor.test(state\_update$Confirmed, state\_update$Recovered) #to see the confidence and mostly used by numeric values



ggplot(state\_update, mapping = aes(x=Confirmed, y=Recovered)) + geom\_point()



ggplot(state\_update, mapping = aes(x= Confirmed, y= Recovered, color = Death))+ geom\_point(alpha = .8,size = 3 ) + geom\_smooth(method = "lm",se = FALSE,size = 1.5)

