#display dataframeView(df)

#size dim(df)

#check datatypesstr(df)

#check for missing valuescolSums(is.na(df))

#drop null valuesdf <- na.omit(df) #it will delete all rows

df <- read.csv("C:/Users/Andre/Downloads/Data Setinc5000 Company List\_2014.csv")

names(df)

df <- df[ ,!(names(df) %in% drops)]

View(df)

#drop null values method-3df <- read.csv("C:/Users/Andre/Downloads/Data Setinc5000 Company List\_2014.csv")df <- df[ ,!(names(df) %in% c("X\_input"))]

View(df)

#check duplicatesdf <- df[!duplicated(df$id),]

#round of to 2 decimal places#df$growth <- round(df$growth,digit=2)

View(df)

#checking outliersinstall.packages("tidyverse") #one time processlibrary(ggplot2)

#drawing boxplot and labeling outliersggplot(df, aes(x=revenue, y=growth)) + geom\_boxplot(outlier.colour = "red", outlier.shape = 1)+ scale\_x\_continuous(labels = scales::comma)+coord\_cartesian(ylim = c(0, 1000))

#Calcualting IQR growthQ1\_growth <- quantile(df$growth,0.25)Q3\_growth <- quantile(df$growth,0.75)

IQR\_growth <- Q3\_growth - Q1\_growth

#using IQR functionIQR\_growth\_new <- IQR(df$growth) #calculate using IQR

#Calculate IQR for revenue columnIQR\_revenue <- IQR(df$revenue)

Q1\_revenue <- quantile(df$revenue,0.25)Q3\_revenue <- quantile(df$revenue,0.75)

#removing outliersno\_outliers <- subset(df, df$growth> (Q1\_growth - 1.5\*IQR\_growth) & df$growth< (Q3\_growth + 1.5\*IQR\_growth))

no\_outliers <- subset(no\_outliers, no\_outliers$revenue> (Q1\_revenue - 1.5\*IQR\_revenue) & no\_outliers$revenue< (Q3\_revenue + 1.5\*IQR\_revenue)

no\_outliers <- subset(no\_outliers, no\_outliers$revenue> (Q1\_revenue - 1.5\*IQR\_revenue) & no\_outliers$revenue< (Q3\_revenue + 1.5\*IQR\_revenue))View(df)

ggplot(no\_outliers, aes(x=rank, y=workers)) + geom\_point()+ scale\_y\_continuous(labels=scales::comma)+ coord\_cartesian(ylim = c(0, 300))

ggplot(no\_outliers, aes(x=yrs\_on\_list)) + geom\_bar()+coord\_cartesian(ylim = c(0, 1500), xlim=c(0,15))

write.csv(no\_outliers, "cleaned\_df.csv" ) IQR\_revenue <- IQR(df$revenue)

plot(mtcars$mpg, mtcars$hp, main="Scatter Plot", xlab="MPG", ylab="Horsepower", col="blue", pch=19)

barplot(mtcars$mpg, names.arg=mtcars$carb, main="Bar Plot", xlab="Carburetors", ylab="MPG")

x <- seq(0, 2\*pi, length.out=100)y <- sin(x)plot(x, y, type="l", main="Line Plot", xlab="x", ylab="sin(x)", col="red")

x <- seq(0, 2\*pi, length.out=100)y <- cos(x)plot(x, y, type="l", main="Line Plot", xlab="x", ylab="cos(x)", col="blue")

#iris dataset data("iris")

hist(iris$Sepal.Width, main="Histogram",xlab="Sepal Length", ylab="Frequency", col="orange")

boxplot(Sepal.Length ~ Species, data=iris,main="Box Plot", xlab="Species", ylab="Sepal Length")

boxplot(Sepal.Width ~ Species, data=iris, main="Box Plot", xlab="Species", ylab="Sepal Width")

#summary statistics

summary( no\_outliers)

no\_outliers<- subset(no\_outliers, no\_outliers$revenue>(q1\_revenue +1.5\* IQR\_revenue))

# Divariate analysis

ggplot(no\_outliers, aes(x=rank,y=workers))+

geom\_point()+scale\_y\_continous(labels=scales::comma)+

cood\_cartesian(ylim= c(0,30))

ggplot (no\_outliers, aes(x=yrs\_on\_list))+

geom\_bar()+coord\_cartesian(ylim= c(0,1500), xlim-c(0,15))

ggplot(no\_outliers, aes (x =yrs\_on\_list)+

geom\_bar()+

#export cleaned data

write.csv(no\_outliers,"cleaned\_df.csv")