**Summary of Datasets**

All the data summary

N = 1846

Mean of X = 54.2656952958938

Variance of X = 279.324417290501

Standard Deviation of X = 16.7130014446987

Mean of Y = 47.8350992042866

Variance of Y = 720.802557659893

Standard Deviation of Y = 26.8477663439604

Correlation of X and Y = -0.0660189107021526

Dataset: a

N = 142

Mean of X = 54.2660997842958

Variance of X = 281.227028989269

Standard Deviation of X = 16.7698249540438

Mean of Y = 47.8347206249437

Variance of Y = 725.749775495945

Standard Deviation of Y = 26.9397434192671

Correlation of X and Y = -0.0641283521673983

Dataset: b

N = 142

Mean of X = 54.2687300223944

Variance of X = 281.207393188831

Standard Deviation of X = 16.7692394934544

Mean of Y = 47.8308231553028

Variance of Y = 725.533372314003

Standard Deviation of Y = 26.9357266899188

Correlation of X and Y = -0.0685863942410766

Dataset: c

N = 142

Mean of X = 54.2673197059859

Variance of X = 280.89802435686

Standard Deviation of X = 16.7600126598061

Mean of Y = 47.8377172672535

Variance of Y = 725.226843692268

Standard Deviation of Y = 26.9300360878382

Correlation of X and Y = -0.0683433564802556

Dataset: d

N = 142

Mean of X = 54.2632732394366

Variance of X = 281.069987591761

Standard Deviation of X = 16.7651420391168

Mean of Y = 47.8322528169014

Variance of Y = 725.515961004212

Standard Deviation of Y = 26.9354034869391

Correlation of X and Y = -0.0644718527009517

Dataset: e

N = 142

Mean of X = 54.2603034516901

Variance of X = 281.156953411424

Standard Deviation of X = 16.7677354884738

Mean of Y = 47.8398292090141

Variance of Y = 725.235215224891

Standard Deviation of Y = 26.9301915185335

Correlation of X and Y = -0.0603414419992176

Dataset: f

N = 142

Mean of X = 54.261441783169

Variance of X = 281.095332523976

Standard Deviation of X = 16.7658979038993

Mean of Y = 47.830251913662

Variance of Y = 725.756930772519

Standard Deviation of Y = 26.939876220438

Correlation of X and Y = -0.0617148379726301

Dataset: g

N = 142

Mean of X = 54.268805279507

Variance of X = 281.122363557963

Standard Deviation of X = 16.7667040159348

Mean of Y = 47.8354502040141

Variance of Y = 725.76349016083

Standard Deviation of Y = 26.939997961411

Correlation of X and Y = -0.0685042204941232

Dataset: h

N = 142

Mean of X = 54.267848823662

Variance of X = 281.124205610883

Standard Deviation of X = 16.766758947718

Mean of Y = 47.8358963311268

Variance of Y = 725.553748890474

Standard Deviation of Y = 26.93610493168

Correlation of X and Y = -0.068979735359512

Dataset: i

N = 142

Mean of X = 54.2658817854225

Variance of X = 281.194419895952

Standard Deviation of X = 16.7688526708285

Mean of Y = 47.8314956523239

Variance of Y = 725.688604796042

Standard Deviation of Y = 26.9386080708718

Correlation of X and Y = -0.0686092064182564

Dataset: j

N = 142

Mean of X = 54.2673411047887

Variance of X = 281.197993194393

Standard Deviation of X = 16.7689592161945

Mean of Y = 47.8395452253521

Variance of Y = 725.239694775897

Standard Deviation of Y = 26.9302746880884

Correlation of X and Y = -0.0629611002206543

Dataset: k

N = 142

Mean of X = 54.2699272309155

Variance of X = 281.231511825566

Standard Deviation of X = 16.7699586113254

Mean of Y = 47.8369879884085

Variance of Y = 725.63880888486

Standard Deviation of Y = 26.9376838069805

Correlation of X and Y = -0.0694455695935036

Dataset: l

N = 142

Mean of X = 54.2669163011972

Variance of X = 281.232887173399

Standard Deviation of X = 16.769999617573

Mean of Y = 47.8316019879718

Variance of Y = 725.650560268097

Standard Deviation of Y = 26.9379019277318

Correlation of X and Y = -0.0665752302046091

A screenshot of a social media post

Description generated with very high confidence

Figure . – Sequential Plot of complete Data set

A close up of text on a black background

Description generated with very high confidence

Figure – Sequential Plot of sub datasets (a-l)

A picture containing sky

Description generated with high confidence

Figure – Scatter Plot of complete dataset.

A close up of text on a white background

Description generated with high confidence

Figure – Scatter plot of Datasets (a-l).

**Observations**

The sequence plots show the scale of the dataset which is between 0 and 100 and the number of data points in the complete set is 1846. Each sub dataset has 146. The quantitative statistics show that the data that the values for mean, correlation, and standard deviation values are all equivalent for all and each dataset (a – l). The correlation statistic shows that there is no relationship between the two variables X, Y for datasets. In addition, the X, Y variables in each data sets have no linear relationship since the values are heavily scattered. This is also existing in the statistical summary of the complete dataset. However, when the visual dataset using scatter plots and sequential plots appears drastically different. There’s no linear relationship between the X and Y data points. The scatter plot interestingly shows various nonlinear patterns. These patterns form various shape groups with no noticeable outliers.

**R Project Source Code**

##############

# functions

#--------------------------------------------------------

# name: \_subset()

# desc: design to pass variable args to subset function

#--------------------------------------------------------

mysubset <- function(dataset, ssubset) {

return (subset(dataset, eval(parse(text=ssubset))))

}

#--------------------------------------------------------

# name: compute\_data()

# desc: computes mean, variance, standard deviation

#--------------------------------------------------------

compute\_data <- function(dataset) {

# N

N = nrow(dataset)

# x

mn\_x <- mean(dataset$x)

vr\_x <- var(dataset$x)

std\_x <- sqrt(vr\_x)

# y

mn\_y <- mean(dataset$y)

vr\_y <- var(dataset$y)

std\_y <- sqrt(vr\_y)

# correlation

# cor\_xy <- cor.test(dataset$x, dataset$y, method = "pearson")

cor\_xy <- cor(dataset$x, dataset$y)

# print a summary of x

result <- ""

result <- paste(result,"N = ",N, '\n', sep="")

result <- paste(result,"Mean of X = ",mn\_x, '\n', sep="")

result <- paste(result,"Variance of X = ",vr\_x, '\n', sep="")

result <- paste(result,"Standard Deviation of X = ",std\_x, '\n', sep="")

# print a summary of y

result <- paste(result,"Mean of Y = ",mn\_y, '\n', sep="")

result <- paste(result,"Variance of Y = ",vr\_y, '\n', sep="")

result <- paste(result,"Standard Deviation of Y = ",std\_y, '\n', sep="")

# print correlation of x and y

result <- paste(result,"Correlation of X and Y = ",cor\_xy,"\n",sep="")

return (result)

}

#----------------------------------------------------

# name: make\_sequential\_plot

# desc:

#----------------------------------------------------

make\_sequential\_plot <- function(input, a) {

plot(ts(input$x),ylab = "X", main = a)

plot(ts(input$y),ylab = "Y", main = a)

}

#----------------------------------------------------

# name: make\_scatter\_plot

# desc:

#----------------------------------------------------

make\_scatter\_plot <- function(input, a) {

# Plot the chart for cars with weight between 2.5 to 5 and mileage between 15 and 30.

plot(y = input$y,

x = input$x,

xlab = "X",

ylab = "Y",

xlim = c(0,100),

ylim = c(0,100),

main = a,

abline(lm(input$x~input$y))

)

}

#################

# main

path <- "C:\\Users\\developer\\Desktop\\cezdev2\\libs\\c3dclassessdk\\cnotes\\cmodels\\cdatamining\\hw0"

datafile <- paste(path, "w1taskCSV.csv", sep="\\")

# 1. load the dataset

ds <- read.csv(datafile, header=T)

# 2. calculate the mean, standard deviation, correlation

output <- ""

output <- paste("All the data summary\n", sep="")

output <- paste(output,compute\_data(ds),"\n",sep="")

x = c('a','b','c','d','e','f','g','h','i','j','k','l')

n = length(x)

for(i in 1:n) {

output <- paste(output, "Dataset: ", x[i], "\n", sep="")

q <- paste("dataset==","'",x[i],"'",sep="")

ss <- mysubset(ds, q)

output <- paste(output, compute\_data(ss), "\n", sep="")

}

file <- paste("summary.txt",sep="")

datafile <- paste(path, file, sep="\\")

fileConn<-file(datafile)

writeLines(output, fileConn)

close(fileConn)

# 3. sequential plots

file <- paste("sequential-plot-all.jpg",sep="")

datafile <- paste(path, file, sep="\\")

png(file = datafile)

par(mfrow=c(2,1))

make\_sequential\_plot(ds, "all")

dev.off()

file <- paste("sequential-plot.jpg",sep="")

datafile <- paste(path, file, sep="\\")

png(file = datafile)

par(mfrow=c(3,4))

x = c('a','b','c','d','e','f','g','h','i','j','k','l')

n = length(x)

for(i in 1:n) {

q <- paste("dataset==","'",x[i],"'",sep="")

ss <- mysubset(ds, q)

make\_sequential\_plot(ss, x[i])

}

dev.off()

# 4. scatter plots

file <- paste("scatter-plot-all.jpg",sep="")

datafile <- paste(path, file, sep="\\")

png(file = datafile)

par(mfrow=c(2,1))

make\_scatter\_plot(ds, "all")

dev.off()

file <- paste("scatter-plot.jpg",sep="")

datafile <- paste(path, file, sep="\\")

png(file = datafile)

par(mfrow=c(3,4))

x = c('a','b','c','d','e','f','g','h','i','j','k','l')

n = length(x)

for(i in 1:n) {

q <- paste("dataset==","'",x[i],"'",sep="")

ss <- mysubset(ds, q)

make\_scatter\_plot(ss, x[i])

}

dev.off()