181058\_DSLab\_MID

Task 01

Create a Vector and save 6 values 1 ,2, 3,4,5,6 your id digits, Assign names to each instance and display length and type of vector.

ID <- c(1,8,1,0,5,8)  
names(ID) <- c("one","eight","one", "zero", "five","eight")  
# Can also do it like   
# id <- c("One"=1,"Eight"=8,"One"=1, "Zero"=0, "Five"=5,"Eight"=8)  
print(ID)

## one eight one zero five eight   
## 1 8 1 0 5 8

paste("LENGTH -> ", length(ID)) #alternative for concatenating

## [1] "LENGTH -> 6"

paste("Type -> ", typeof(ID))

## [1] "Type -> double"

Task 02

Create a matrix "YourName" and store 9 values row wise with 3 rows and name the row and column names. Multiply Complete Matrix with 100 and Display the output.

HamzaAli <- matrix(1:9, nrow = 3, byrow=TRUE)  
rownames(HamzaAli) <- c("R1","R2","R3")  
colnames(HamzaAli) <- c("C1","C2","C3")  
new\_mat <- HamzaAli \* 100  
print(HamzaAli)

## C1 C2 C3  
## R1 1 2 3  
## R2 4 5 6  
## R3 7 8 9

print(new\_mat)

## C1 C2 C3  
## R1 100 200 300  
## R2 400 500 600  
## R3 700 800 900

Task 03

Using Provided Dataset

df <- read.csv("DataSet\_MidExam\_C.csv")  
print("STRUCTURE")

## [1] "STRUCTURE"

print(str(df))

## 'data.frame': 1216 obs. of 7 variables:  
## $ Country...territory : Factor w/ 31 levels "Aruba","Barbados",..: 1 1 1 1 1 1 1 1 2 2 ...  
## $ Measure.Names : Factor w/ 2 levels "Confirmed","Suspected": 2 1 2 1 2 1 2 1 2 1 ...  
## $ Report.Epi.Week : int 1 1 2 2 3 3 4 4 1 1 ...  
## $ Year.of.Date : int 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 ...  
## $ Measure.Values : Factor w/ 113 levels "0","1","1,046",..: 1 1 1 1 1 1 1 1 1 56 ...  
## $ Laboratory.confirmed.cases: int 0 0 0 0 0 0 0 0 3 3 ...  
## $ Suspected.cases : Factor w/ 88 levels "0","1","1,046",..: 1 1 1 1 1 1 1 1 1 1 ...  
## NULL

print("SUMMARY")

## [1] "SUMMARY"

print(summary(df))

## Country...territory Measure.Names Report.Epi.Week  
## Barbados : 44 Confirmed:608 Min. : 1.0   
## Nicaragua : 44 Suspected:608 1st Qu.: 5.0   
## United States Virgin Islands : 44 Median :43.0   
## Costa Rica : 42 Mean :31.8   
## Saint Vincent and the Grenadines: 42 3rd Qu.:48.0   
## Sint Maarten : 42 Max. :52.0   
## (Other) :958 NA's :3   
## Year.of.Date Measure.Values Laboratory.confirmed.cases Suspected.cases  
## Min. :2015 0 :1022 Min. : 0.000 0 :1032   
## 1st Qu.:2015 1 : 17 1st Qu.: 0.000 3 : 6   
## Median :2015 4 : 13 Median : 0.000 115 : 4   
## Mean :2015 2 : 12 Mean : 4.689 2 : 4   
## 3rd Qu.:2016 3 : 8 3rd Qu.: 0.000 59 : 4   
## Max. :2016 10 : 4 Max. :526.000 1 : 2   
## (Other): 140 (Other): 164

#### Str vs Summary

According to the Output of Str and Summary,

1. Str provides the following:

* The total number of rows and columns, i.e. 1216 rows(objects) and 7 columns(variables)
* "$" gives description of each column(their factor(unique values) and int values itselfs)

1. As for summary,

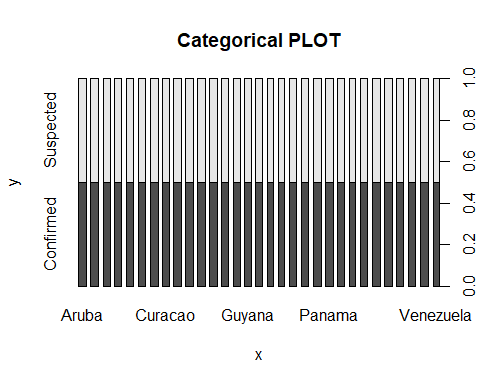
* It gives summary of head, 6 rows and others
* Each column would gives total number of distinct rows, for example, Barbados has 44 rows in Country...territory
* For Integer, it would gives out Min, max, media, median, number of NA's and basic math functions
* In short, it would give the summary of the dataset to give bried description of the provided datasets

#### Finding the Categorical and Numerical attributes

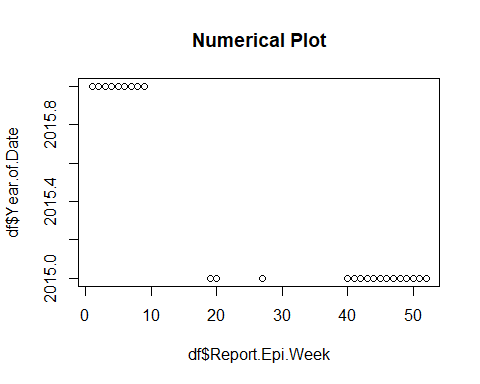
* Any int value gives numerical attributes
* Any factor of limited values (M/F, Confirmed/Suspected) gives Categorical attribues Now, looking at summary, we can draw following conclusion

1. **Country...territory**, **Measure.Names**, **Measure.Values**, **Suspected.cases** has distinct values(factor), so it's categorical attributes. **Year.of.Data** gives 2 values (2015,2016), so it can be also categorical
2. **Report.Epi.Week**, **Year.of.Date**, **Laboratory.Confirmed.cases**, **Suspected.cases** are all numerical attributes

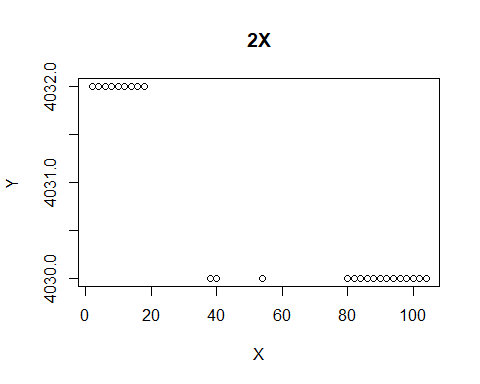
df <- read.csv("DataSet\_MidExam\_C.csv")  
plot(main = "Categorical PLOT", df$Country...territory, df$Measure.Names)



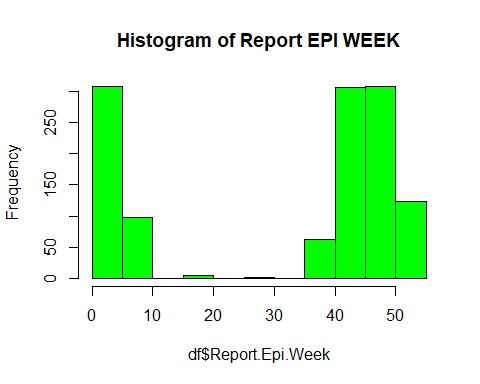
plot(main = "Numerical Plot", df$Report.Epi.Week, df$Year.of.Date)



plot(main = "2X", (df$Report.Epi.Week)\*2, (df$Year.of.Date)\*2, xlab = "X", ylab="Y")



hist(df$Report.Epi.Week,main = "Histogram of Report EPI WEEK", col="green", breaks = 10)



Task 04

#### Top & Last5 entries of Provided Dataset

print("Top 5 values")

## [1] "Top 5 values"

print(head(df, n = 5))

## Country...territory Measure.Names Report.Epi.Week Year.of.Date  
## 1 Aruba Suspected 1 2016  
## 2 Aruba Confirmed 1 2016  
## 3 Aruba Suspected 2 2016  
## 4 Aruba Confirmed 2 2016  
## 5 Aruba Suspected 3 2016  
## Measure.Values Laboratory.confirmed.cases Suspected.cases  
## 1 0 0 0  
## 2 0 0 0  
## 3 0 0 0  
## 4 0 0 0  
## 5 0 0 0

print("Last 5 values")

## [1] "Last 5 values"

print(tail(df, n = 5))

## Country...territory Measure.Names Report.Epi.Week Year.of.Date  
## 1212 Venezuela Confirmed 50 2015  
## 1213 Venezuela Suspected 51 2015  
## 1214 Venezuela Confirmed 51 2015  
## 1215 Venezuela Suspected 52 2015  
## 1216 Venezuela Confirmed 52 2015  
## Measure.Values Laboratory.confirmed.cases Suspected.cases  
## 1212 0 0 0  
## 1213 0 0 0  
## 1214 0 0 0  
## 1215 0 0 0  
## 1216 0 0 0

print("Summary")

## [1] "Summary"

print(summary(df))

## Country...territory Measure.Names Report.Epi.Week  
## Barbados : 44 Confirmed:608 Min. : 1.0   
## Nicaragua : 44 Suspected:608 1st Qu.: 5.0   
## United States Virgin Islands : 44 Median :43.0   
## Costa Rica : 42 Mean :31.8   
## Saint Vincent and the Grenadines: 42 3rd Qu.:48.0   
## Sint Maarten : 42 Max. :52.0   
## (Other) :958 NA's :3   
## Year.of.Date Measure.Values Laboratory.confirmed.cases Suspected.cases  
## Min. :2015 0 :1022 Min. : 0.000 0 :1032   
## 1st Qu.:2015 1 : 17 1st Qu.: 0.000 3 : 6   
## Median :2015 4 : 13 Median : 0.000 115 : 4   
## Mean :2015 2 : 12 Mean : 4.689 2 : 4   
## 3rd Qu.:2016 3 : 8 3rd Qu.: 0.000 59 : 4   
## Max. :2016 10 : 4 Max. :526.000 1 : 2   
## (Other): 140 (Other): 164

print("5 Line Summary")

## [1] "5 Line Summary"

fivenum(df$Report.Epi.Week)

## [1] 1 5 43 48 52

fivenum(df$Year.of.Date)

## [1] 2015 2015 2015 2016 2016

fivenum(df$Laboratory.confirmed.cases)

## [1] 0 0 0 0 526

#### Difference Between Summary and 5Num summary

1. As described by its name, summary gives detailed view of the datasets, even **factors** and **string**
2. Factor only works with **numerical attributes** as it gives provides

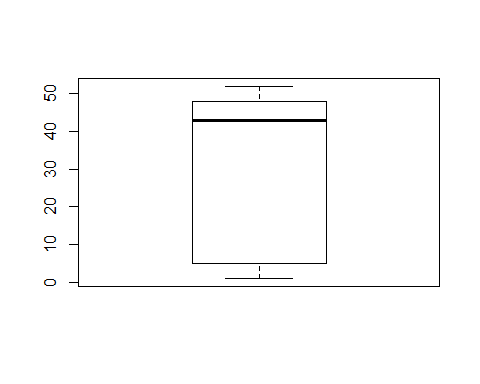
* Minimum
* First Quartile
* Median
* Third Quartile
* Maximum

#### BoxPlot and Outliers

print("BoxPlot")

## [1] "BoxPlot"

boxplot(df$Report.Epi.Week)

 According to the Boxplot, \* Lowest line gives min values = 1 \* 2nd lower will gives 1st quartile = 5 \* bolded line is the median of column = 43 \* Above that gives third quartilse = 48 \* Top values give Maximun values = 52 All the values below minimum and above maximum are referred as outliers (1 > Outliers > 52)

Task 05

#### Encode Categorical to Numerical

df <- read.csv("DataSet\_MidExam\_C.csv")  
df$Country...territory <- factor(df$Country...territory, labels= 1:31)  
df$Measure.Names <- factor(df$Measure.Names, labels = c(1,2))  
df$Measure.Values <- factor(df$Measure.Values, labels = 1:113)  
df$Suspected.cases <- factor(df$Suspected.cases, labels = 1:88)

#### Missing values with mean

According to summary, we have NA at Report.Epi.Week

df$Report.Epi.Week <- ifelse(is.na(df$Report.Epi.Week),  
 ave(df$Report.Epi.Week, FUN = function(x) mean(x, na.rm = TRUE)), df$Report.Epi.Week)  
#print(df$Report.Epi.Week)  
#for conforming  
#is.na(df$Report.Epi.Week)

#### Split to 80/20

library(caTools)

## Warning: package 'caTools' was built under R version 3.3.3

set.seed(123)  
split = sample.split(df$Measure.Values, SplitRatio = 0.8)  
trainingset <- subset(df, split == TRUE)  
testset <- subset(df, split == FALSE)  
paste("Training Set", dim(trainingset))

## [1] "Training Set 994" "Training Set 7"

paste("Test set", dim(testset))

## [1] "Test set 222" "Test set 7"