Assignment 10.1

Name : Y Vasudev

Batch : DA with R , Excel and Tableau

Import dataset from the following link: AirQuality Data Set

Perform the following written operations:

1. Read the file in Zip format and get it into R.

airqu = paste('https://archive.ics.uci.edu/ml/machine-learning-databases/00360/', 'AirQualityUCI.zip', sep='')# create a temporary directorytd = tempdir()# create the placeholder filetf = tempfile(tmpdir=td, fileext=".zip")# download into the placeholder filedownload.file(airqu, tf)# get the name of the first file in the zip archivefname = unzip(tf, list=TRUE)$Name[1]fname# unzip the file to the temporary directoryunzip(tf, files=fname, exdir=td, overwrite=TRUE)# fpath is the full path to the extracted filefpath = file.path(td, fname)fpathd = read.csv(fpath,sep = ";")View(d)

2. Create Univariate for all the columns.

univariateTable(~date,time,data(airqu))

3. Check for missing values in all columns.

d1[is.na(PT08.S1.CO.)] = mean(PT08.S1.CO.)

d1[is.na(d1$PT08.S1.CO.)] = mean(d1$PT08.S1.CO.)

for(i in 1:ncol(airq)){

airqu[is.na(airq[,i]),i] <- mean(airqu[,i], na.rm = TRUE)

}

4. Impute the missing values using appropriate methods.

imp = impute(airqu, classes = list(integer = imputeMean(), factor = imputeMode()),

dummy.classes = "integer")

5. Create bi-variate analysis for all relationships.

library(psych)

pairs.panels( airq[,c(1,2,3,4,5,6)],

method = "pearson", # correlation method

hist.col = "red",

density = TRUE, # show density plots

ellipses = TRUE, # show correlation ellipses

lm=TRUE,

main ="Bivariate Scatter plots with Pearson Correlation & Histogram"

)

6. Test relevant hypothesis for valid relations.

7. Create cross tabulations with derived variables.

mytable<- xtabs(~x+y,data = airqu)

mytable

8. Check for trends and patterns in time series.

table(data$a,data$b)

plot.ts(datasetname)

souvenir\_decomp=decompose(souvenir\_ts)

plot(souvenir\_decomp)

9. Find out the most polluted time of the day and the name of the chemical compound.

**NOT YET COVERED IN THE SESSION S**

> head(airqu)

Date Time CO.GT. PT08.S1.CO. NMHC.GT. C6H6.GT. PT08.S2.NMHC.

1 2004-03-10 1899-12-30 18:00:00 2.6 1360.00 150 11.881723 1045.50

2 2004-03-10 1899-12-30 19:00:00 2.0 1292.25 112 9.397165 954.75

3 2004-03-10 1899-12-30 20:00:00 2.2 1402.00 88 8.997817 939.25

4 2004-03-10 1899-12-30 21:00:00 2.2 1375.50 80 9.228796 948.25

5 2004-03-10 1899-12-30 22:00:00 1.6 1272.25 51 6.518224 835.50

6 2004-03-10 1899-12-30 23:00:00 1.2 1197.00 38 4.741012 750.25

NOx.GT. PT08.S3.NOx. NO2.GT. PT08.S4.NO2. PT08.S5.O3. T RH AH

1 166 1056.25 113 1692.00 1267.50 13.600 48.875 0.7577538

2 103 1173.75 92 1558.75 972.25 13.300 47.700 0.7254874

3 131 1140.00 114 1554.50 1074.00 11.900 53.975 0.7502391

4 172 1092.00 122 1583.75 1203.25 11.000 60.000 0.7867125

5 131 1205.00 116 1490.00 1110.00 11.150 59.575 0.7887942

6 89 1336.50 96 1393.00 949.25 11.175 59.175 0.7847717

> tail(airqu)

Date Time CO.GT. PT08.S1.CO. NMHC.GT. C6H6.GT. PT08.S2.NMHC. NOx.GT. PT08.S3.NOx.

9466 <NA> <NA> NaN NaN NaN NaN NaN NaN NaN

9467 <NA> <NA> NaN NaN NaN NaN NaN NaN NaN

9468 <NA> <NA> NaN NaN NaN NaN NaN NaN NaN

9469 <NA> <NA> NaN NaN NaN NaN NaN NaN NaN

9470 <NA> <NA> NaN NaN NaN NaN NaN NaN NaN

9471 <NA> <NA> NaN NaN NaN NaN NaN NaN NaN

NO2.GT. PT08.S4.NO2. PT08.S5.O3. T RH AH

9466 NaN NaN NaN NaN NaN NaN

9467 NaN NaN NaN NaN NaN NaN

9468 NaN NaN NaN NaN NaN NaN

9469 NaN NaN NaN NaN NaN NaN

9470 NaN NaN NaN NaN NaN NaN

9471 NaN NaN NaN NaN NaN NaN

> names(airqu)

[1] "Date" "Time" "CO.GT." "PT08.S1.CO." "NMHC.GT."

[6] "C6H6.GT." "PT08.S2.NMHC." "NOx.GT." "PT08.S3.NOx." "NO2.GT."

[11] "PT08.S4.NO2." "PT08.S5.O3." "T" "RH" "AH"

> summary(airqu)

Error in as.POSIXct.default(value) :

do not know how to convert 'value' to class “POSIXct”

> airqu$NMHC.GT.[is.na(airqu$NMHC.GT.)] <- 0.0000

> airqu$CO.GT.[is.na(airqu$CO.GT.)] <- 0.0000

> airqu$PT08.S1.CO.[is.na(airqu$PT08.S1.CO.)] <- 0.0000

> airqu$C6H6.GT.[is.na(airqu$C6H6.GT.)] <- 0.0000

> airqu$PT08.S2.NMHC.[is.na(airqu$PT08.S2.NMHC.)] <- 0.0000

> airqu$NOx.GT.[is.na(airqu$NOx.GT.)] <- 0.0000

> airqu$PT08.S3.NOx.[is.na(airqu$PT08.S3.NOx.)] <- 0.0000

> airqu$NO2.GT.[is.na(airqu$NO2.GT.)] <- 0.0000

> airqu$PT08.S4.NO2.[is.na(airqu$PT08.S4.NO2.)] <- 0.0000

> airqu$PT08.S5.O3.[is.na(airqu$PT08.S5.O3.)] <- 0.0000

> airqu$T[is.na(airqu$T)] <- 0.0000

> airqu$RH[is.na(airqu$RH)] <- 0.0000

> airqu$AH[is.na(airqu$AH)] <- 0.0000

> summary(airqu)

Error in as.POSIXct.default(value) :

do not know how to convert 'value' to class “POSIXct”

> class(airqu)

[1] "data.frame"

> str(airqu)

'data.frame': 9471 obs. of 15 variables:

$ Date :Class 'POSIXct' atomic [1:9471] 1.08e+09 1.08e+09 1.08e+09 1.08e+09 1.08e+09 ...

.. ..- attr(\*, "tzone")= chr "GMT"

$ Time :Class 'POSIXct' atomic [1:9471] -2.21e+09 -2.21e+09 -2.21e+09 -2.21e+09 -2.21e+09 ...

.. ..- attr(\*, "tzone")= chr "GMT"

$ CO.GT. : num 2.6 2 2.2 2.2 1.6 1.2 1.2 1 0.9 0.6 ...

$ PT08.S1.CO. : num 1360 1292 1402 1376 1272 ...

$ NMHC.GT. : num 150 112 88 80 51 38 31 31 24 19 ...

$ C6H6.GT. : num 11.88 9.4 9 9.23 6.52 ...

$ PT08.S2.NMHC.: num 1046 955 939 948 836 ...

$ NOx.GT. : num 166 103 131 172 131 89 62 62 45 -200 ...

$ PT08.S3.NOx. : num 1056 1174 1140 1092 1205 ...

$ NO2.GT. : num 113 92 114 122 116 96 77 76 60 -200 ...

$ PT08.S4.NO2. : num 1692 1559 1554 1584 1490 ...

$ PT08.S5.O3. : num 1268 972 1074 1203 1110 ...

$ T : num 13.6 13.3 11.9 11 11.2 ...

$ RH : num 48.9 47.7 54 60 59.6 ...

$ AH : num 0.758 0.725 0.75 0.787 0.789 ...

> head(airqu)

Date Time CO.GT. PT08.S1.CO. NMHC.GT. C6H6.GT. PT08.S2.NMHC.

1 2004-03-10 1899-12-30 18:00:00 2.6 1360.00 150 11.881723 1045.50

2 2004-03-10 1899-12-30 19:00:00 2.0 1292.25 112 9.397165 954.75

3 2004-03-10 1899-12-30 20:00:00 2.2 1402.00 88 8.997817 939.25

4 2004-03-10 1899-12-30 21:00:00 2.2 1375.50 80 9.228796 948.25

5 2004-03-10 1899-12-30 22:00:00 1.6 1272.25 51 6.518224 835.50

6 2004-03-10 1899-12-30 23:00:00 1.2 1197.00 38 4.741012 750.25

NOx.GT. PT08.S3.NOx. NO2.GT. PT08.S4.NO2. PT08.S5.O3. T RH AH

1 166 1056.25 113 1692.00 1267.50 13.600 48.875 0.7577538

2 103 1173.75 92 1558.75 972.25 13.300 47.700 0.7254874

3 131 1140.00 114 1554.50 1074.00 11.900 53.975 0.7502391

4 172 1092.00 122 1583.75 1203.25 11.000 60.000 0.7867125

5 131 1205.00 116 1490.00 1110.00 11.150 59.575 0.7887942

6 89 1336.50 96 1393.00 949.25 11.175 59.175 0.7847717