# Problem Statement

Import dataset from the following link: [AirQuality Data Set](https://archive.ics.uci.edu/ml/datasets/Air+quality)

Perform the following written operations:

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

forecasturl = paste('https://archive.ics.uci.edu/ml/machine-learning-databases/00360/',

'AirQualityUCI.zip', sep='')

# create a temporary directory

td = tempdir()

# create the placeholder file

tf = tempfile(tmpdir=td, fileext=".zip")

# download into the placeholder file

download.file(forecasturl, tf)

# get the name of the first file in the zip archive

fname = unzip(tf, list=TRUE)$Name[1]

fname

# unzip the file to the temporary directory

unzip(tf, files=fname, exdir=td, overwrite=TRUE)

# fpath is the full path to the extracted file

fpath = file.path(td, fname)

fpath

airquality = read.csv(fpath,sep = ";")

View(airquality)

1. Create Univariate for all the columns.

#Univariate analysis is the simplest form of analyzing data. "Uni" means "one",

#so in other words your data has only one variable

#we can do univariate analysis by this command too

library(psych)

summary(airquality)

describe(airquality)

#or visually

library(purrr)

library(tidyr)

library(ggplot2)

airquality %>%

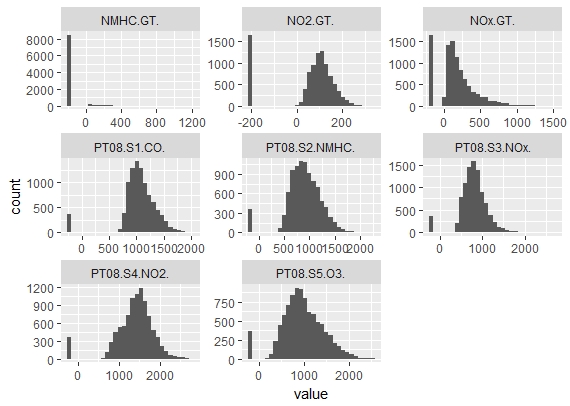
keep(is.numeric) %>%

gather() %>%

ggplot(aes(value)) +

facet\_wrap(~ key,scales = "free") +

geom\_histogram()



#or we can plot univariate individually for each variable

#hence plotting histogram

hist(airquality$PT08.S1.CO,xlab = "PT08.S1(CO)", ylab = "Frequency",main="Histogram of PT08.S1.CO",col="red")

hist(airquality$NMHC.GT,xlab = "NMHC(GT)", ylab = "Frequency",main="Histogram of NMHC.GT",col="blue")

hist(airquality$PT08.S2.NMHC,xlab = "PT08.S2(NMHC)", ylab = "Frequency",main="Histogram of PT08.S2.NMHC",col="yellow")

hist(airquality$NOx.GT ,xlab = "NOx(GT)", ylab = "Frequency",main="Histogram of NOx.GT",col="darkblue")

hist(airquality$PT08.S3.NOx,xlab = "PT08.S3(NOx)", ylab = "Frequency",main="Histogram of PT08.S3.NOx",col="pink")

hist(airquality$NO2.GT,xlab = "NO2(GT)", ylab = "Frequency",main="Histogram of NO2.GT",col="purple")

1. Check for missing values in all columns.

#with the help of summary function we can find which variable has how many NA value

#or check for missing values

summary(airquality)

#thus PT08.S1.CO.,NMHC.GT., PT08.S2.NMHC. , NOx.GT. , ...... NA=114 has missing values

1. Impute the missing values using appropriate methods.

#lets see the structure of airquality first

str(airquality)

library(mice)

md.pattern(airquality)

#visualizing

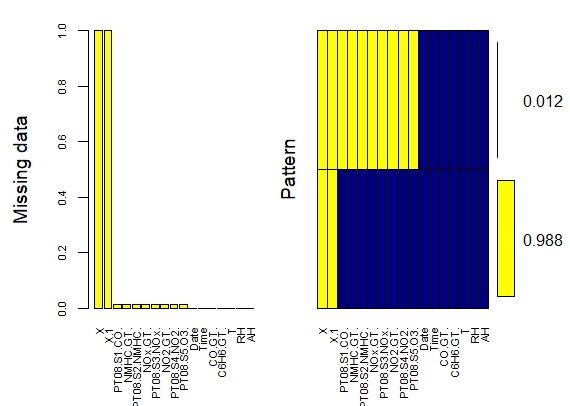
library(VIM)

mice\_plot <- aggr(airquality, col=c('navyblue','yellow'),

numbers=TRUE, sortVars=TRUE,

labels=names(airquality), cex.axis=.7,

gap=3, ylab=c("Missing data","Pattern"))



# In this case we are using predictive mean matching as imputation method

imputed\_Data <- mice(airquality, m=5, maxit = 50, method = 'pmm', seed = 500)

summary(imputed\_Data)

completeData <- complete(imputed\_Data)

View(completeData)

1. Create bi-variate analysis for all relationships.

library(psych)

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

method = "pearson", # correlation method

hist.col = "red",

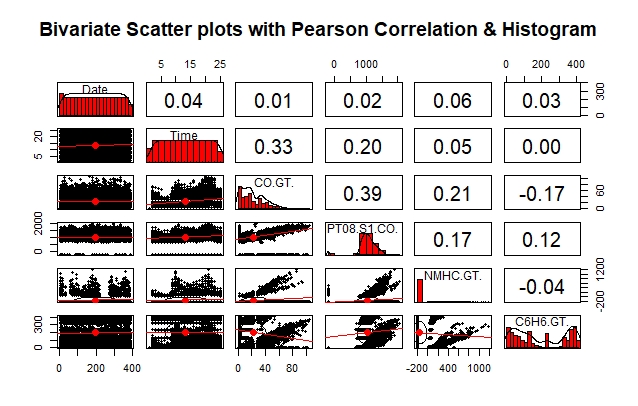
density = TRUE, # show density plots

ellipses = TRUE, # show correlation ellipses

lm=TRUE,

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

)



1. Test relevant hypothesis for valid relations.

#Using builtin dataset (airquality)

#lets see the structure first

str(airquality)

#we do paired test for continous variables

#some of test are as follows

#define the null hypothesis

#Ho: Mean of first variable - Mean of 2 variable is equal to 0

#Ha: Mean of first variable - Mean of 2 variable is not equal to 0

t.test(x=airquality$Ozone, y=airquality$Solar.R ,alternative = "two.sided",mu=0 ,paired = TRUE)

t.test(x=airquality$Temp, y=airquality$Wind ,alternative = "two.sided",mu=0 ,paired = TRUE)

t.test(x=airquality$Ozone, y=airquality$Temp ,alternative = "two.sided",mu=0 ,paired = TRUE)

t.test(x=airquality$Day, y=airquality$Solar.R ,alternative = "two.sided",mu=0 ,paired = TRUE)

> t.test(x=airquality$Ozone, y=airquality$Solar.R ,alternative = "two.sided",mu=0 ,paired = TRUE)

Paired t-test

data: airquality$Ozone and airquality$Solar.R

t = -17.593, df = 110, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-158.7772 -126.6282

sample estimates:

mean of the differences

-142.7027

> t.test(x=airquality$Temp, y=airquality$Wind ,alternative = "two.sided",mu=0 ,paired = TRUE)

Paired t-test

data: airquality$Temp and airquality$Wind

t = 72.978, df = 152, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

66.08593 69.76374

sample estimates:

mean of the differences

67.92484

> t.test(x=airquality$Ozone, y=airquality$Temp ,alternative = "two.sided",mu=0 ,paired = TRUE)

Paired t-test

data: airquality$Ozone and airquality$Temp

t = -14.14, df = 115, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-40.74819 -30.73457

sample estimates:

mean of the differences

-35.74138

> t.test(x=airquality$Day, y=airquality$Solar.R ,alternative = "two.sided",mu=0 ,paired = TRUE)

Paired t-test

data: airquality$Day and airquality$Solar.R

t = -22.353, df = 145, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-184.8230 -154.7934

sample estimates:

mean of the differences

-169.8082

#as p value of this test is <0.05 we reject the null hypo

#and accept the alternative hypothesis which says there

#Mean of 1 variable - Mean of 2 variable is not equal to 0

#thus this are some test that we performed

1. Create cross tabulations with derived variables.

#we are using Builtin data "airquality"

attach(airquality)

unique(Wind)

unique(Temp)

#derived variables of wind and temp

x<- cut(Wind,quantile(Wind))

x<- cut(Wind,breaks = seq(1,21,3),labels = c("wind1","wind2","wind3","wind4","wind5","wind6"))

y<- cut(Temp,quantile(Temp))

y<- cut(Temp,breaks = seq(55,100,9),labels = c("temp1","temp2","temp3","temp4","temp5"))

table(x,y)

#or like this using xtabs function

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

mytable

#crosstabulate

library(gmodels)

CrossTable(x,y)

|  |
| --- |
| > #we are using Builtin data "airquality"  > attach(airquality)  > unique(Wind)  [1] 7.4 8.0 12.6 11.5 14.3 14.9 8.6 13.8 20.1 6.9 9.7 9.2 10.9 13.2 12.0 18.4 16.6 5.7  [19] 16.1 20.7 10.3 6.3 1.7 4.6 4.1 5.1 4.0 15.5 3.4 2.3 2.8  > unique(Temp)  [1] 67 72 74 62 56 66 65 59 61 69 68 58 64 57 73 81 79 76 78 84 85 82 87 90 93 92 80 77 75 83  [31] 88 89 91 86 97 94 96 71 63 70  > #derived variables of wind and temp  > x<- cut(Wind,quantile(Wind))  > x<- cut(Wind,breaks = seq(1,21,3),labels = c("wind1","wind2","wind3","wind4","wind5","wind6"))  > y<- cut(Temp,quantile(Temp))  > y<- cut(Temp,breaks = seq(55,100,9),labels = c("temp1","temp2","temp3","temp4","temp5"))  > table(x,y)  y  x temp1 temp2 temp3 temp4 temp5  wind1 0 0 2 1 2  wind2 0 1 11 10 6  wind3 4 9 18 14 3  wind4 4 11 17 8 1  wind5 4 4 13 3 0  wind6 3 2 0 0 0  > #or like this using xtabs function  > mytable<- xtabs(~x+y,data = airquality)  > mytable  y  x temp1 temp2 temp3 temp4 temp5  wind1 0 0 2 1 2  wind2 0 1 11 10 6  wind3 4 9 18 14 3  wind4 4 11 17 8 1  wind5 4 4 13 3 0  wind6 3 2 0 0 0  > library(gmodels)  > CrossTable(x,y)    Cell Contents  |-------------------------|  | N |  | Chi-square contribution |  | N / Row Total |  | N / Col Total |  | N / Table Total |  |-------------------------|    Total Observations in Table: 151    | y  x | temp1 | temp2 | temp3 | temp4 | temp5 | Row Total |  -------------|-----------|-----------|-----------|-----------|-----------|-----------|  wind1 | 0 | 0 | 2 | 1 | 2 | 5 |  | 0.497 | 0.894 | 0.000 | 0.031 | 6.464 | |  | 0.000 | 0.000 | 0.400 | 0.200 | 0.400 | 0.033 |  | 0.000 | 0.000 | 0.033 | 0.028 | 0.167 | |  | 0.000 | 0.000 | 0.013 | 0.007 | 0.013 | |  -------------|-----------|-----------|-----------|-----------|-----------|-----------|  wind2 | 0 | 1 | 11 | 10 | 6 | 28 |  | 2.781 | 3.206 | 0.009 | 1.656 | 6.404 | |  | 0.000 | 0.036 | 0.393 | 0.357 | 0.214 | 0.185 |  | 0.000 | 0.037 | 0.180 | 0.278 | 0.500 | |  | 0.000 | 0.007 | 0.073 | 0.066 | 0.040 | |  -------------|-----------|-----------|-----------|-----------|-----------|-----------|  wind3 | 4 | 9 | 18 | 14 | 3 | 48 |  | 0.124 | 0.020 | 0.100 | 0.571 | 0.174 | |  | 0.083 | 0.188 | 0.375 | 0.292 | 0.062 | 0.318 |  | 0.267 | 0.333 | 0.295 | 0.389 | 0.250 | |  | 0.026 | 0.060 | 0.119 | 0.093 | 0.020 | |  -------------|-----------|-----------|-----------|-----------|-----------|-----------|  wind4 | 4 | 11 | 17 | 8 | 1 | 41 |  | 0.001 | 1.836 | 0.012 | 0.322 | 1.565 | |  | 0.098 | 0.268 | 0.415 | 0.195 | 0.024 | 0.272 |  | 0.267 | 0.407 | 0.279 | 0.222 | 0.083 | |  | 0.026 | 0.073 | 0.113 | 0.053 | 0.007 | |  -------------|-----------|-----------|-----------|-----------|-----------|-----------|  wind5 | 4 | 4 | 13 | 3 | 0 | 24 |  | 1.095 | 0.020 | 1.126 | 1.295 | 1.907 | |  | 0.167 | 0.167 | 0.542 | 0.125 | 0.000 | 0.159 |  | 0.267 | 0.148 | 0.213 | 0.083 | 0.000 | |  | 0.026 | 0.026 | 0.086 | 0.020 | 0.000 | |  -------------|-----------|-----------|-----------|-----------|-----------|-----------|  wind6 | 3 | 2 | 0 | 0 | 0 | 5 |  | 12.617 | 1.368 | 2.020 | 1.192 | 0.397 | |  | 0.600 | 0.400 | 0.000 | 0.000 | 0.000 | 0.033 |  | 0.200 | 0.074 | 0.000 | 0.000 | 0.000 | |  | 0.020 | 0.013 | 0.000 | 0.000 | 0.000 | |  -------------|-----------|-----------|-----------|-----------|-----------|-----------|  Column Total | 15 | 27 | 61 | 36 | 12 | 151 |  | 0.099 | 0.179 | 0.404 | 0.238 | 0.079 | |  -------------|-----------|-----------|-----------|-----------|-----------|-----------| |
|  |
| |  | | --- | | > | |

1. Check for trends and patterns in time series.

#since this topics are not yet teach in class

#i'm not able to do it yet

#as they are yet to covered in coming classes

#thats why i left it off

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

#i'm cutting this off

#same for this

#since this topics are not yet teach in class

#i'm not able to do it yet

#as they are yet to covered in coming classes

#thats why i left it off