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.

myfile <- read.csv(zip.file.extract("~/files/test.csv", "myzip.zip"))

1. Create Univariate for all the columns.

> data()

> ?airquality

> a = airquality

> names(a)

> summary(a)

> boxplot(a)

3. Check for missing values in all columns.

install.packages("VIM")

install.packages("naniar")

install.packages("missMDA")

install.packages("Amelia")

install.packages("mice")

install.packages("missForest")

install.packages("FactoMineR")

install.packages("Tidyverse")

ozo <- read.table("ozoneNA.csv", header = TRUE,

sep=",", row.names = 1)

WindDirection <- ozo[, 12]

don <- ozo[, 1:11] *#### keep the continuous variables*

summary(don)

head(don)

dim(don)

**library**(VIM)

**library**(FactoMineR)

**library**(missMDA

dim(na.omit(don))

4. Impute the missing values using appropriate methods.

**library**(missMDA)

?estim\_ncpPCA

?imputePCA

nb <- estim\_ncpPCA(don,method.cv = "Kfold", verbose = FALSE) *# estimate the number of components from incomplete data*

*#(available methods include GCV to approximate CV)*

nb$ncp *#2*

5. Create bi-variate analysis for all relationships.

> pairs(a,panel=panel.smooth)

> var(a) > var(a,na.rm=TRUE) > cor(a,use=”complete.obs”)

6. Test relevant hypothesis for valid relations.

require("datasets")

data("airquality")

str(airquality)

head(airquality)

ggpairs(airquality, columns=1:3, title=" air quality ")

fit\_1 <- lm(Volume ~ Girth, data = airquality)

summary(fit\_1)

7. Create cross tabulations with derived variables.

complete.tab <- descr:::CreateNewTab(ct)

class(complete.tab)

complete.tab

8. Check for trends and patterns in time series.

> str(airquality)

> head(airquality)

plot(airquality)

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

download.EPA <- [**function**](http://inside-r.org/r-doc/base/function)(year, property = [**c**](http://inside-r.org/r-doc/base/c)("ozone","so2","co","no2","pm25.frm","pm25","pm10","wind","temp","pressure","dewpoint","hap","voc","lead"), type=[**c**](http://inside-r.org/r-doc/base/c)("hourly","daily","annual")){  
**if**(property=="ozone"){PROP="44201"}  
**if**(property=="so2"){PROP="42401"}  
**if**(property=="co"){PROP="42101"}  
**if**(property=="no2"){PROP="42602"}  
   
**if**(property=="pm25.frm"){PROP="88101"}  
**if**(property=="pm25"){PROP="88502"}  
**if**(property=="pm10"){PROP="81102"}  
   
**if**(property=="wind"){PROP="WIND"}  
**if**(property=="temp"){PROP="TEMP"}  
**if**(property=="pressure"){PROP="PRESS"}  
**if**(property=="dewpoint"){PROP="RH\_DP"}  
**if**(property=="hap"){PROP="HAPS"}  
**if**(property=="voc"){PROP="VOCS"}  
**if**(property=="lead"){PROP="lead"}  
   
URL <- paste0("<http://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/>",type,"\_",PROP,"\_",year,".zip")  
[**download.file**](http://inside-r.org/r-doc/utils/download.file)(URL,destfile=paste0(type,"\_",PROP,"\_",year,".zip"))  
[**unzip**](http://inside-r.org/r-doc/utils/unzip)(paste0(type,"\_",PROP,"\_",year,".zip"),exdir=paste0([**getwd**](http://inside-r.org/r-doc/base/getwd)()))  
[**read.table**](http://inside-r.org/r-doc/utils/read.table)(paste0(type,"\_",PROP,"\_",year,".csv"),sep=",",header=T)  
}

[**data**](http://inside-r.org/r-doc/utils/data) <- download.EPA(year=2013,property="ozone",type="daily")

> [**str**](http://inside-r.org/r-doc/utils/str)([**data**](http://inside-r.org/r-doc/utils/data))  
'data.frame': 390491 obs. of 28 variables:  
 $ State.Code : int 1 1 1 1 1 1 1 1 1 1 ...  
 $ County.Code : int 3 3 3 3 3 3 3 3 3 3 ...  
 $ Site.Num : int 10 10 10 10 10 10 10 10 10 10 ...  
 $ Parameter.Code : int 44201 44201 44201 44201 44201 44201 44201 44201 44201 44201 ...  
 $ POC : int 1 1 1 1 1 1 1 1 1 1 ...  
 $ Latitude : num 30.5 30.5 30.5 30.5 30.5 ...  
 $ Longitude : num -87.9 -87.9 -87.9 -87.9 -87.9 ...  
 $ Datum : Factor w/ 4 [**levels**](http://inside-r.org/r-doc/base/levels) "NAD27","NAD83",..: 2 2 2 2 2 2 2 2 2 2 ...  
 $ Parameter.Name : Factor w/ 1 level "Ozone": 1 1 1 1 1 1 1 1 1 1 ...  
 $ Sample.Duration : Factor w/ 1 level "8-HR RUN AVG BEGIN HOUR": 1 1 1 1 1 1 1 1 1 1 ...  
 $ Pollutant.Standard : Factor w/ 1 level "Ozone 8-Hour 2008": 1 1 1 1 1 1 1 1 1 1 ...  
 $ Date.Local : Factor w/ 365 [**levels**](http://inside-r.org/r-doc/base/levels) "2013-01-01","2013-01-02",..: 59 60 61 62 63 64 65 66 67 68 ...  
 $ Units.of.Measure : Factor w/ 1 level "Parts per million": 1 1 1 1 1 1 1 1 1 1 ...  
 $ Event.Type : Factor w/ 3 [**levels**](http://inside-r.org/r-doc/base/levels) "Excluded","Included",..: 3 3 3 3 3 3 3 3 3 3 ...  
 $ Observation.Count : int 1 24 24 24 24 24 24 24 24 24 ...  
 $ Observation.Percent: num 4 100 100 100 100 100 100 100 100 100 ...  
 $ Arithmetic.Mean : num 0.03 0.0364 0.0344 0.0288 0.0345 ...  
 $ X1st.Max.Value : num 0.03 0.044 0.036 0.042 0.045 0.045 0.045 0.048 0.057 0.059 ...  
 $ X1st.Max.Hour : int 23 10 18 10 9 10 11 12 12 10 ...  
 $ AQI : int 25 37 31 36 38 38 38 41 48 50 ...  
 $ Method.Name : Factor w/ 1 level " - ": 1 1 1 1 1 1 1 1 1 1 ...  
 $ Local.Site.Name : Factor w/ 1182 [**levels**](http://inside-r.org/r-doc/base/levels) ""," 201 CLINTON ROAD, JACKSON",..: 353 353 353 353 353 353 353 353 353 353 ...  
 $ Address : Factor w/ 1313 [**levels**](http://inside-r.org/r-doc/base/levels) " Edgewood Chemical Biological Center (APG), Waehli Road",..: 907 907 907 907 907 907 907 907 907 907 ...  
 $ State.Name : Factor w/ 53 [**levels**](http://inside-r.org/r-doc/base/levels) "Alabama","Alaska",..: 1 1 1 1 1 1 1 1 1 1 ...  
 $ County.Name : Factor w/ 631 [**levels**](http://inside-r.org/r-doc/base/levels) "Abbeville","Ada",..: 32 32 32 32 32 32 32 32 32 32 ...  
 $ City.Name : Factor w/ 735 [**levels**](http://inside-r.org/r-doc/base/levels) "Adams","Air Force Academy",..: 221 221 221 221 221 221 221 221 221 221 ...  
 $ CBSA.Name : Factor w/ 414 [**levels**](http://inside-r.org/r-doc/base/levels) "","Adrian, MI",..: 94 94 94 94 94 94 94 94 94 94 ...  
 $ Date.of.Last.Change: Factor w/ 169 [**levels**](http://inside-r.org/r-doc/base/levels) "2013-05-17","2013-07-01",..: 125 125 125 125 125 125 125 125 125 125 ...