Assignment - 1

Session 10 – Correlations

**5. Problem Statement**

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.

Ans:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | | > library(readxl)  > AirQualityUCI <- read\_excel("E:/ACADGILD-DATA ANALYTICS/ppt/assignments/AirQualityUCI.xlsx")  > View(AirQualityUCI)  > Air <-AirQualityUCI  > dim(Air)  [1] 9357 15  > str(Air)  Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 9357 obs. of 15 variables:  $ Date : POSIXct, format: "2004-03-10" "2004-03-10" ...  $ Time : POSIXct, format: "1899-12-31 18:00:00" "1899-12-31 19:00:00" ...  $ 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 ...  > View(Air) | |  | | |  | | --- | | > | | |
|  |
|  |

1. Create Univariate for all the columns.

Ans:

> library(psych)

> describe(Air)

vars n mean sd median trimmed mad min max

Date 1 9357 NaN NA NA NaN NA Inf -Inf

Time 2 9357 NaN NA NA NaN NA Inf -Inf

CO(GT) 3 9357 -34.21 77.66 1.50 -18.41 1.48 -200 11.90

PT08.S1(CO) 4 9357 1048.87 329.82 1052.50 1069.72 218.19 -200 2039.75

NMHC(GT) 5 9357 -159.09 139.79 -200.00 -200.00 0.00 -200 1189.00

C6H6(GT) 6 9357 1.87 41.38 7.89 8.75 6.62 -200 63.74

PT08.S2(NMHC) 7 9357 894.48 342.32 894.50 907.06 288.37 -200 2214.00

NOx(GT) 8 9357 168.60 257.42 141.00 147.72 161.31 -200 1479.00

PT08.S3(NOx) 9 9357 794.87 321.98 794.25 799.84 238.70 -200 2682.75

NO2(GT) 10 9357 58.14 126.93 96.00 72.32 59.30 -200 339.70

PT08.S4(NO2) 11 9357 1391.36 467.19 1445.50 1426.54 349.15 -200 2775.00

PT08.S5(O3) 12 9357 974.95 456.92 942.00 972.05 403.64 -200 2522.75

T 13 9357 9.78 43.20 17.20 17.39 9.71 -200 44.60

RH 14 9357 39.48 51.22 48.55 48.04 20.65 -200 88.73

AH 15 9357 -6.84 38.98 0.98 0.99 0.45 -200 2.23

range skew kurtosis se

Date -Inf NA NA NA

Time -Inf NA NA NA

CO(GT) 211.90 -1.67 0.78 0.80

PT08.S1(CO) 2239.75 -1.72 5.83 3.41

NMHC(GT) 1389.00 4.07 18.85 1.45

C6H6(GT) 263.74 -4.51 19.17 0.43

PT08.S2(NMHC) 2414.00 -0.79 2.37 3.54

NOx(GT) 1679.00 0.82 1.50 2.66

PT08.S3(NOx) 2882.75 -0.38 3.10 3.33

NO2(GT) 539.70 -1.23 0.27 1.31

PT08.S4(NO2) 2975.00 -1.24 3.26 4.83

PT08.S5(O3) 2722.75 -0.03 0.64 4.72

T 244.60 -4.44 18.76 0.45

RH 288.73 -3.93 15.75 0.53

AH 202.23 -4.75 20.60 0.40

1. Check for missing values in all columns.

Ans:

> Air[Air ==-200] <-NA

> View(Air)

> library(VIM)

> aggr(Air, col=c('navyblue','yellow'),

+ numbers=TRUE, sortVars=TRUE,

+ labels=names(Air), cex.axis=.7,

+ gap=3, ylab=c("Missing data","Pattern")) # graphical presentation of NAs

Variables sorted by number of missings:

Variable Count

NMHC(GT) 0.9023191

CO(GT) 0.1798653

NO2(GT) 0.1754836

NOx(GT) 0.1751630

PT08.S1(CO) 0.0391151

C6H6(GT) 0.0391151

PT08.S2(NMHC) 0.0391151

PT08.S3(NOx) 0.0391151

PT08.S4(NO2) 0.0391151

PT08.S5(O3) 0.0391151

T 0.0391151

RH 0.0391151

AH 0.0391151

Date 0.0000000

Time 0.0000000

> sapply(Air, function(x) sum(is.na(x))) # count of NAs

Date Time CO(GT) PT08.S1(CO) NMHC(GT)

0 0 1683 366 8443

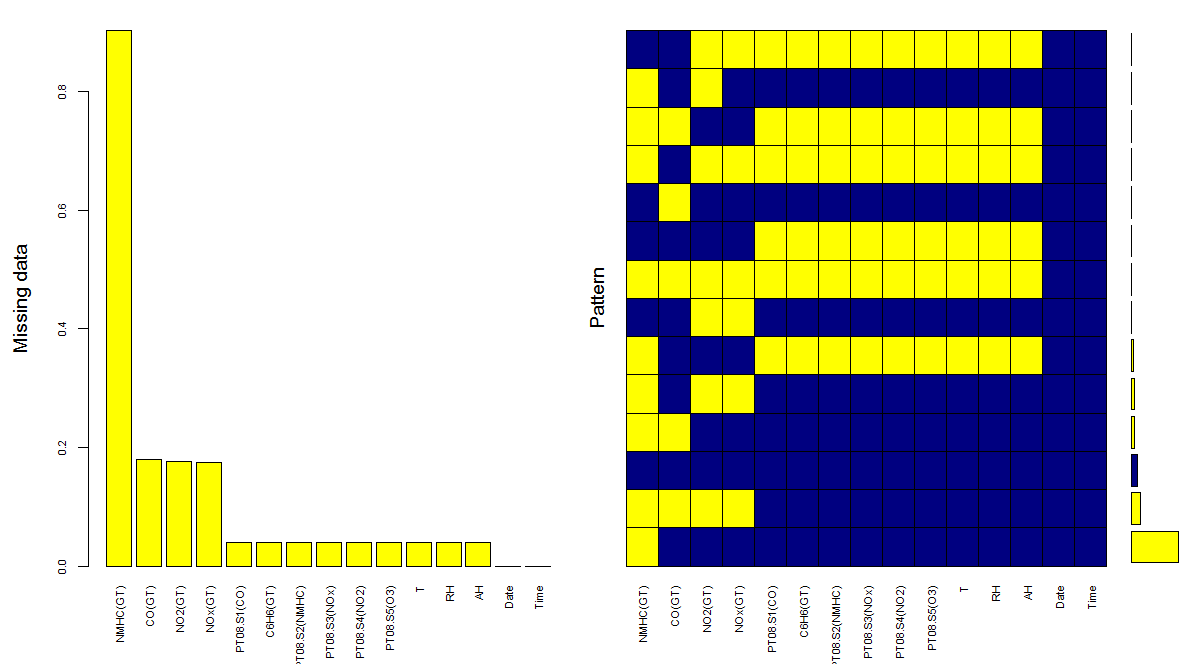
C6H6(GT) PT08.S2(NMHC) NOx(GT) PT08.S3(NOx) NO2(GT)

366 366 1639 366 1642

PT08.S4(NO2) PT08.S5(O3) T RH AH

366 366 366 366 366

> Air$`NMHC(GT)` <-NULL



1. Impute the missing values using appropriate methods.

Ans:

|  |
| --- |
| > names(Air)  [1] "Date" "Time" "CO(GT)" "PT08.S1(CO)"  [5] "C6H6(GT)" "PT08.S2(NMHC)" "NOx(GT)" "PT08.S3(NOx)"  [9] "NO2(GT)" "PT08.S4(NO2)" "PT08.S5(O3)" "T"  [13] "RH" "AH"  > Air$Date1 <-as.numeric(as.Date(Air$Date))  > library(mice)  > impute <- mice(Air[,-c(1,2)], m=5, maxit=5, method='cart', seed=100) # impute missing values  iter imp variable  1 1 CO(GT)  > summary(impute)  Class: mids  Number of multiple imputations: 5  Imputation methods:  name Measure geo\_type\_name geo\_entity\_id  "" "" "" ""  geo\_entity\_name year\_description data\_valuemessage X  "" "" "" ""  PredictorMatrix:  name Measure geo\_type\_name geo\_entity\_id geo\_entity\_name  name 0 1 0 1 0  Measure 1 0 1 1 1  geo\_type\_name 1 1 0 1 1  geo\_entity\_id 1 1 1 0 1  geo\_entity\_name 1 1 1 1 0  year\_description 1 1 1 1 1  year\_description data\_valuemessage X  name 1 1 0  Measure 1 1 0  geo\_type\_name 1 1 0  geo\_entity\_id 1 1 0  geo\_entity\_name 1 1 0  year\_description 0 1 0  Number of logged events: 1  it im dep meth out  1 0 0 constant X  > complete <- complete(impute) # replaces the NAs with imputed values  > str(complete)  'data.frame': 2769 obs. of 8 variables:  $ name : Factor w/ 21 levels "Air Toxics Concentrations- Average Benzene Concentrations",..: 1 1 1 1 1 1 1 1 1 1 ...  $ Measure : Factor w/ 9 levels "Average Concentration",..: 1 1 1 1 1 1 1 1 1 1 ...  $ geo\_type\_name : Factor w/ 4 levels "Borough","CD",..: 1 1 1 1 1 3 4 4 4 4 ...  $ geo\_entity\_id : int 1 2 3 4 5 1 101 102 103 104 ...  $ geo\_entity\_name : Factor w/ 99 levels "Astoria","Battery Park City",..: 12 13 55 68 81 61 51 62 38 65 ...  $ year\_description : Factor w/ 57 levels " Bath Beach",..: 52 52 52 52 52 52 52 52 52 52 ...  $ data\_valuemessage: Factor w/ 694 levels " Farragut"," Fordham",..: 224 224 420 31 28 225 225 224 223 223 ...  $ X : logi NA NA NA NA NA NA ...  > sapply(complete, function(x) sum(is.na(x))) # check missing values  name Measure geo\_type\_name geo\_entity\_id  0 0 0 0  geo\_entity\_name year\_description data\_valuemessage X  0 0 0 2769 |
|  |
| |  | | --- | | > | |

1. Create bi-variate analysis for all relationships.

Ans:

> library(psych)

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

+ method = "pearson", # correlation method

+ hist.col = "red",

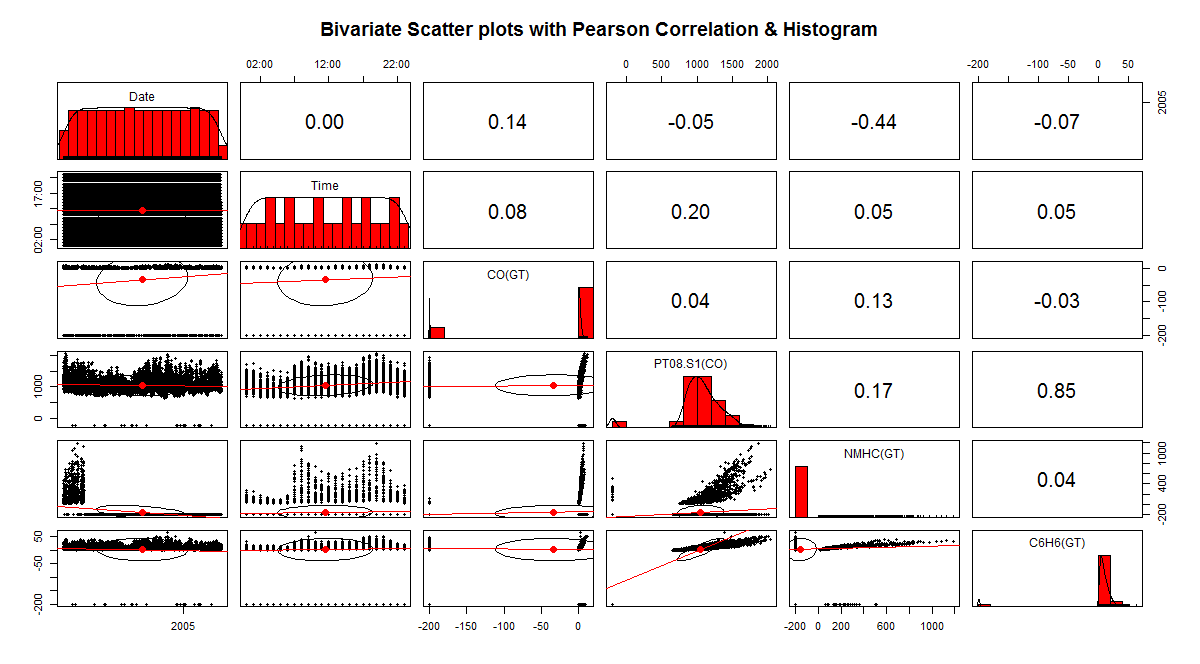
+ density = TRUE, # show density plots

+ ellipses = TRUE, # show correlation ellipses

+ lm=TRUE,

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

+ )



1. Test relevant hypothesis for valid relations.

Ans:

> str(AirQualityUCI)

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 9357 obs. of 15 variables:

$ Date : POSIXct, format: "2004-03-10" "2004-03-10" ...

$ Time : POSIXct, format: "1899-12-31 18:00:00" "1899-12-31 19:00:00" ...

$ 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 ...

> 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

1. Create cross tabulations with derived variables.

Ans:

|  |
| --- |
| > attach(airquality)  The following objects are masked from airquality (pos = 3):  Day, Month, Ozone, Solar.R, Temp, Wind  > 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  [16] 18.4 16.6 5.7 16.1 20.7 10.3 6.3 1.7 4.6 4.1 5.1 4.0 15.5 3.4 2.3  [31] 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  [27] 80 77 75 83 88 89 91 86 97 94 96 71 63 70  > 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  > 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.

Ans:

# Not covered in class

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

Ans:

# Not covered in class