Part 1 :

A piece of paper with writing on it

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Part 2 :

1. Text

   Description automatically generated with low confidence

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9

2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9

3. month - month of the year: "jan" to "dec"

4. day - day of the week: "mon" to "sun"

5. FFMC - FFMC index from the FWI system: 18.7 to 96.20

6. DMC - DMC index from the FWI system: 1.1 to 291.3

7. DC - DC index from the FWI system: 7.9 to 860.6

8. ISI - ISI index from the FWI system: 0.0 to 56.10

9. temp - temperature in Celsius degrees: 2.2 to 33.30

10. RH - relative humidity in %: 15.0 to 100

11. wind - wind speed in km/h: 0.40 to 9.40

12. rain - outside rain in mm/m2 : 0.0 to 6.4

13. area - the burned area of the forest (in ha): 0.00 to 1090.84

Summary : this data include geography information, FWI system index, air information(such as humid rate, wind speed), weather information, affect range.

Missing value : row 510~517, they don’t have FFMC value.

Outliner : row 279, because it has strange value such as 854 index of FFMC that is lager 1.5IQR lage than Q3.

(b)

FFMC,DMC,DC and ISI can be derived to one group named fire Index

Rain and RH can be derived to one group named humid

(c)

According to this matrix,

DC and DMC ,Temp and DMC, Temp and DChave strong positive relation

RH and Temp have strong negative relation

Text

Description automatically generated

(d) corrplot

Chart, scatter chart, bubble chart

Description automatically generated

Part 3 :

(a)

Using function boxplot to observe if there’s any outliner in the data set, then exclude the outliner. Such as this data set, we can exclude the data that “total.sulfur.dioxide” above 250.

Chart, box and whisker chart

Description automatically generated

(b)

Print original dataset first, then use function “dataframe”[complete.case(“dataframe”)] to remove the row have none value. Using nrow(“dataframe”) can see how many rows in the dataset. New dataset’s row number – original dataset’s row number can get how many rows have been removed.

Code:

Graphical user interface, text, application

Description automatically generated

(c)

outliner

Observe whether dose the data exceed 1.5\*IQR from Q3 or below 1.5\*IQR from Q1(out of the box in boxplot). Then remove those data by considering the data size. Otherwise, it should remove by ratio, and start from the largest or smallest data.

Row:

"13" "29" "32" "34" "36" "60" "61" "62" "1469" "1581" "1582" "1583" "1584" "1585" "1586" "1587" "1588" "1589" "1590" "1591" "1592" "1593" "1594" "1595" "1596" "1597"

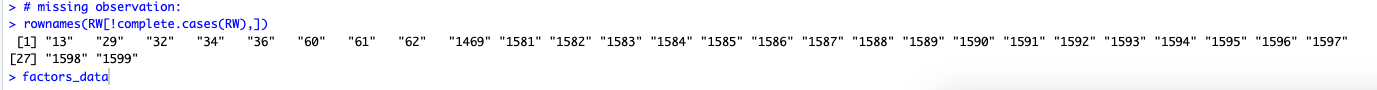
"1598" "1599"

Are missing value, those should be removed because it will biased parameters and results reduce the statistical strength of the analysis.

|  |  |  |
| --- | --- | --- |
|  | Outliners | Missing Observations |
| total.sulfur.dioxide | 55 Outliers, remove number is 278, 289. Because compare to other outliners, both are too large. |  |
| quality | 30 Outliers, remove number is77. Because compare to other outliners, this data is too large. |  |

Table

Description automatically generated



( d )

Both matrix including

1. fixed.acidity
2. volatile.acidity
3. citric.acid
4. residual.sugar
5. chlorides
6. free.sulfur.dioxide
7. total.sulfur.dioxide
8. quality

because they have wider dataset than other column.

Text

Description automatically generated with medium confidence

Table

Description automatically generated

R – code :

RW <- read.csv("~/Desktop/Assigments/2023\_DM/file/red-wine.csv", stringsAsFactors=TRUE)

RW = as.data.frame(RW)

RW

# add id column

# if row.names are strings

# df$idu <- row.names(df)

# if row numbers are integers (most likely!)

# RW$idu <- as.numeric(row.names(RW))

# a

boxplot(RW)

# b

RW\_rm\_miss=RW[complete.cases(RW),]

RW\_rm\_miss

check =nrow(RW) - nrow(RW\_rm\_miss)

check

#-------------------------------------------------------------------------------

# c

# outliner:

df = RW\_rm\_miss

for( i in c(1:ncol(df))){

print(paste("column : ",colnames(RW[i]) ))

check\_outliner(df,df[,c(i)])

}

check\_outliner = function(df,target){

outlier\_values <- boxplot.stats(target)$out # outlier values.

result = sort(c(outlier\_values))

print(result)

print(paste("number of outliners : ",length(result) ))

}

# missing observation:

rownames(RW[!complete.cases(RW),])

#-------------------------------------------------------------------------------

#d

df = RW\_rm\_miss

df\_scale = scale(df[,c(1:7,12)])

df\_scale

cor(df\_scale)

cov(df\_scale)