**DATA SCIENCE**

**UJIAN TENGAH SEMESTER**



Disusun oleh:

Atikah Zakiyyah Sholiha

(1800018096)

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS TEKNOLOGI INDUSTRI**

**UNIVERSITAS AHMAD DAHLAN**

**APRIL 2020**

**Data Set Information:**

The dataset was downloaded from the UCI Machine Learning Repository.

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. The reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are munch more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

Two datasets were combined and few values were randomly removed.

1. Deskripsi Data (Score: 10)  
   Tuliskan deskripsi dari data yang Anda miliki (sebutkan atribut2nya, jumlah datanya, bagaimana data tsb diperoleh, siapa yang melakukan data collection, dll)

Jawab :

Dataset : Wine Quality Link : <https://www.kaggle.com/rajyellow46/wine-quality>

Informasi variabel yang ada :

* Type : Jenis Whine (white / red)

Input variable (berdasarkan physicochemical tests) :

* fixed acidity (keasaman tetap)
* volatile acidity (keasaman volatile)
* citric acid (asam sitrat)
* residual sugar (gula residu)
* chlorides (klorida)
* free sulfur dioxide (sulfur dioksida bebas)
* total sulfur dioxide (total sulfur dioksida)
* density (kepadatan)
* pH
* sulphates (sulfat)
* alcohol

Output variable (berdasarkan sensory data):

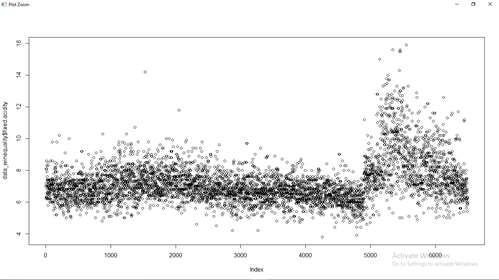
* quality (score antara 0 dan 10)

1. Visualisasi Data (Score: 40)  
   1) Tampilkan masing-masing atribut dengan plot and histogram

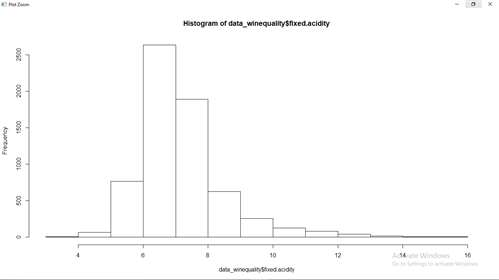
Jawab :

* Type : Jenis Whine (white / red)
* fixed acidity (keasaman tetap)

> plot(data\_winequality$fixed.acidity)

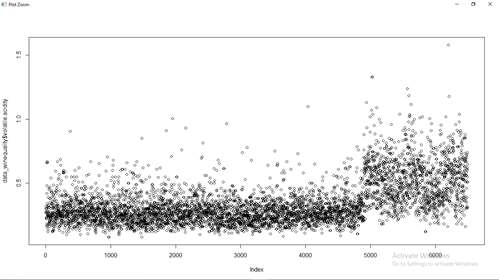


* hist(data\_winequality$fixed.acidity)

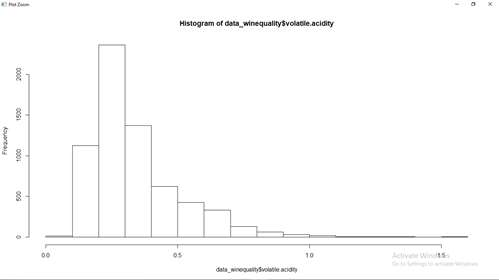


* volatile acidity (keasaman volatile

> plot(data\_winequality$volatile.acidity)

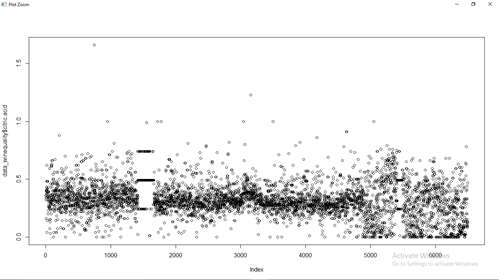


* hist(data\_winequality$volatile.acidity)

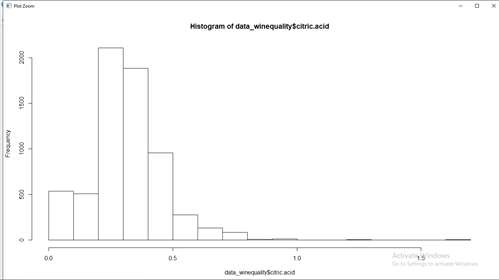


* citric acid (asam sitrat)

> plot(data\_winequality$citric.acid)

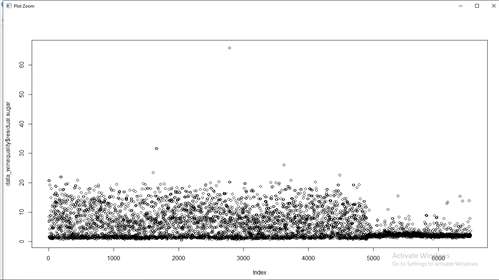


* hist(data\_winequality$citric.acid)

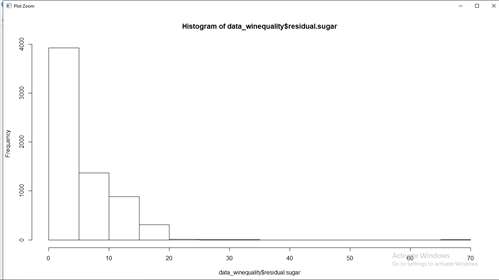


* residual sugar (gula residu)

> plot(data\_winequality$residual.sugar)

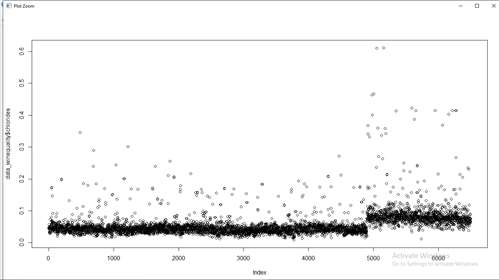


* hist(data\_winequality$residual.sugar)

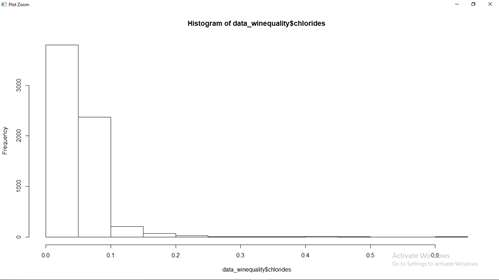


* chlorides (klorida)

> plot(data\_winequality$chlorides)

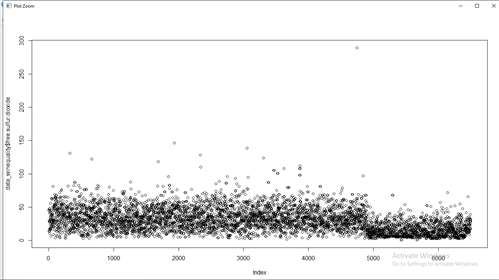


* hist(data\_winequality$chlorides)

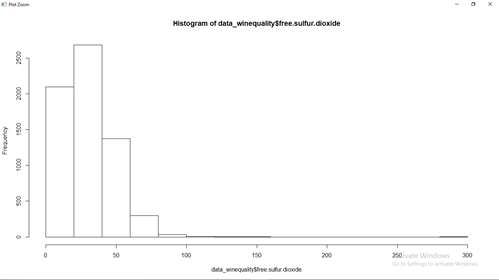


* free sulfur dioxide (sulfur dioksida bebas)

> plot(data\_winequality$free.sulfur.dioxide)

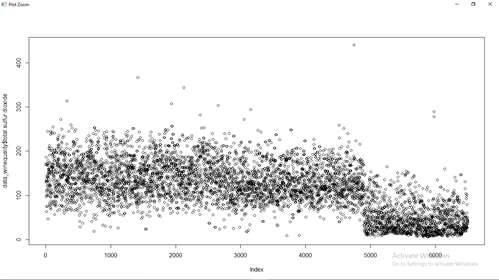


* hist(data\_winequality$free.sulfur.dioxide)

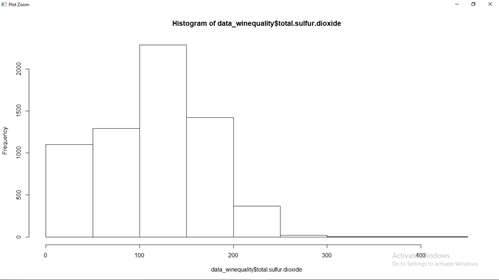


* total sulfur dioxide (total sulfur dioksida)

> plot(data\_winequality$total.sulfur.dioxide)

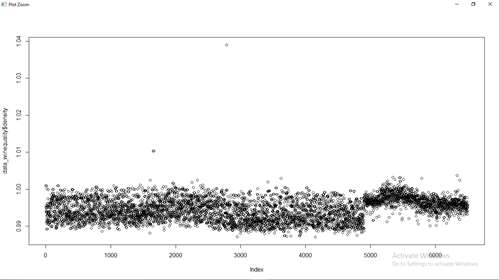


* hist(data\_winequality$total.sulfur.dioxide)

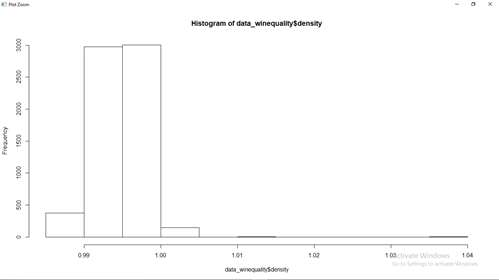


* density (kepadatan)

> plot(data\_winequality$density)

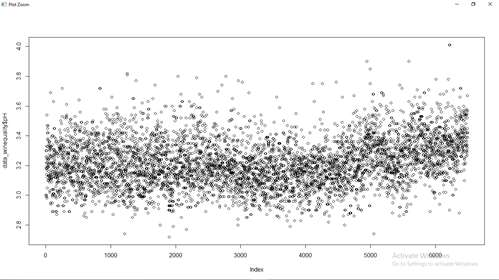


* hist(data\_winequality$density)

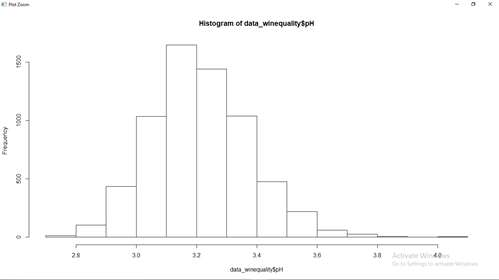


* pH

> plot(data\_winequality$pH)

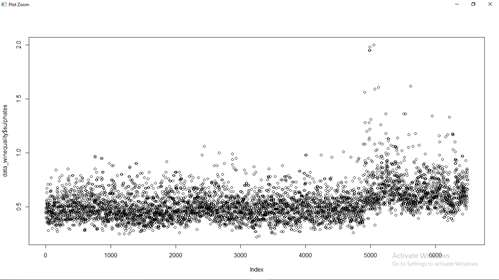


* hist(data\_winequality$pH)

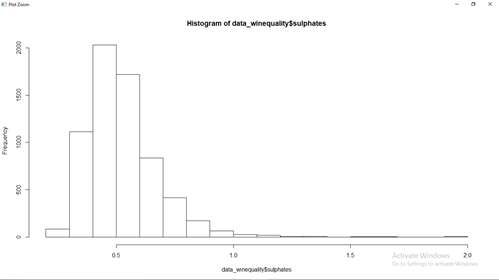


* sulphates (sulfat)

> plot(data\_winequality$sulphates)

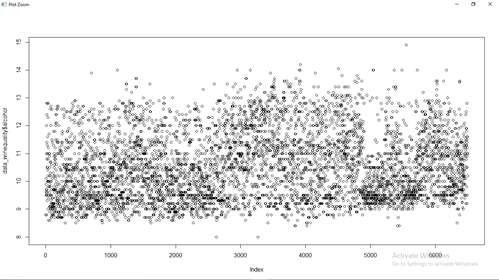


* hist(data\_winequality$sulphates)

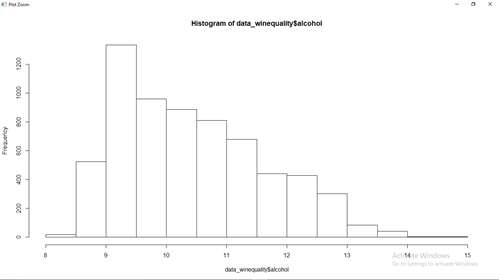


* alcohol

> plot(data\_winequality$alcohol)

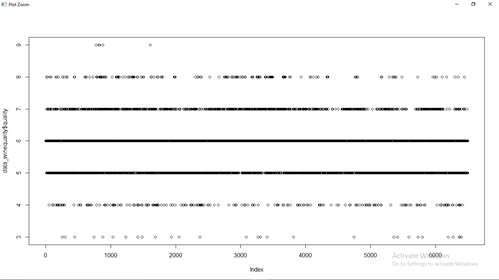


* hist(data\_winequality$alcohol)

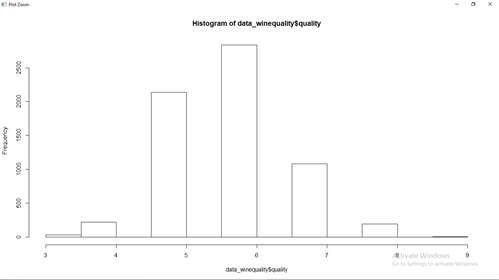


* quality (score antara 0 dan 10)

> plot(data\_winequality$quality)



* hist(data\_winequality$quality)



2) Visualisasikan hubungan nilai-nilai pada dua atribut yang berbeda.  
Lakukan hal ini pada minimal 10 pasang atribut. Sebutkan korelasi apa yang Anda dapatkan? (Korelasi positif/negatif)

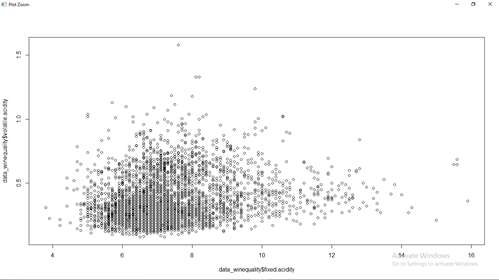
Jawab :

Korelasi positif : jika nilai yang dihasilkan cenderung meningkat (increase).

Korelasi negative : jika nilai yang dihasilkna cenderung menurun (decrease).

1. Visualisasi atribut fixed.acidity dengan volatile.acidity

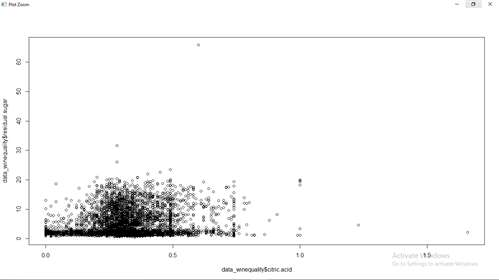
plot(data\_winequality$fixed.acidity,data\_winequality$volatile.acidity)



Korelasi yang didapatkan : positive

1. Visualisasi atribut citric.acid dengan residual.sugar

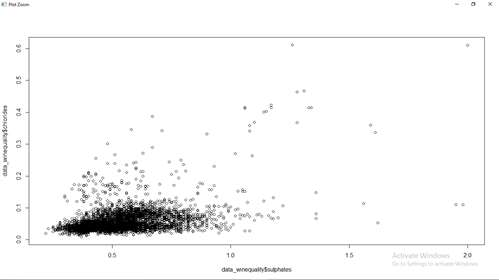
plot(data\_winequality$citric.acid,data\_winequality$residual.sugar)



Korelasi yang didapatkan : positive

1. Visualisasi atribut sulphates dengan chlorides

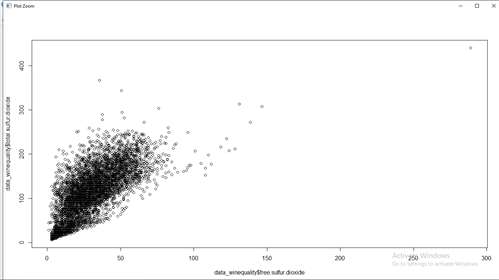
plot(data\_winequality$sulphates,data\_winequality$chlorides)



Korelasi yang didapatkan : positive

1. Visualisasi atribut free.sulfur.dioxide dengan total.sulfur.dioxide

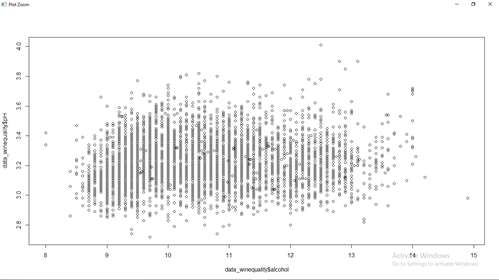
plot(data\_winequality$free.sulfur.dioxide,data\_winequality$total.sulfur.dioxide)



Korelasi yang didapatkan : positive

1. Visualisasi atribut alcohol dengan pH

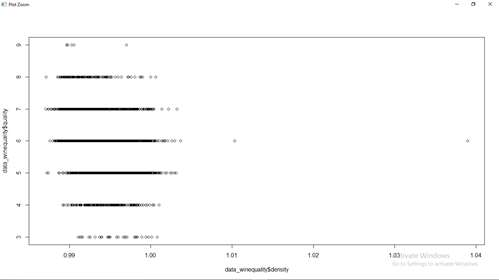
plot(data\_winequality$alcohol,data\_winequality$pH)



Korelasi yang didapatkan : negative

1. Visualisasi atribut fixed.density dengan quality

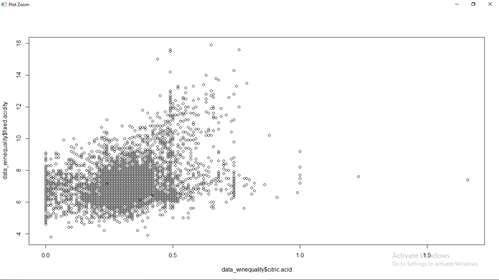
plot(data\_winequality$density,data\_winequality$quality)



Korelasi yang didapatkan : negative

1. Visualisasi atribut citric.acid dengan fixed.acidity

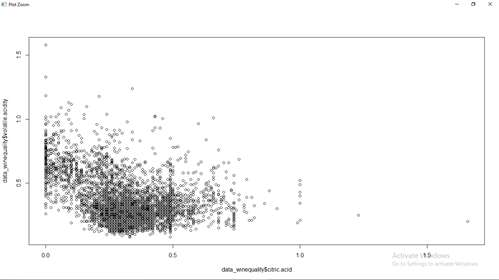
plot(data\_winequality$citric.acid,data\_winequality$fixed.acidity)



Korelasi yang dihasilkan : positive

1. Visualisasi atribut citric.acid dengan volatile.acidity

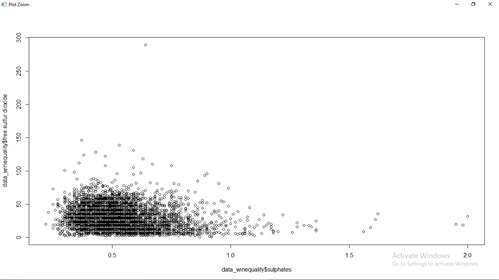
plot(data\_winequality$citric.acid,data\_winequality$volatile.acidity)



Korelasi yang didapatkan : negative

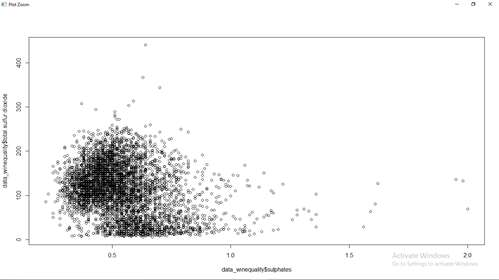
1. Visualisasi atribut sulphates dengan free.sulfur.dioxide

plot(data\_winequality$sulphates,data\_winequality$free.sulfur.dioxide)



Korelasi yang didapatkan : negative

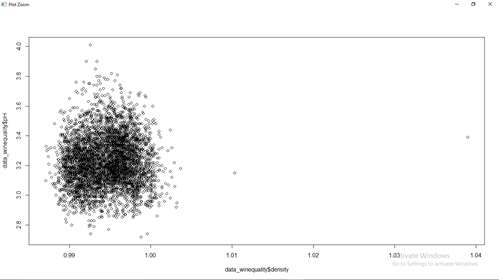
1. Visualisasi atribut sulphates dengan total.sulfur.dioxide

plot(data\_winequality$sulphates,data\_winequality$total.sulfur.dioxide)

Korelasi yang didapatkan : negative

1. Visualisasi atribut density dengan pH

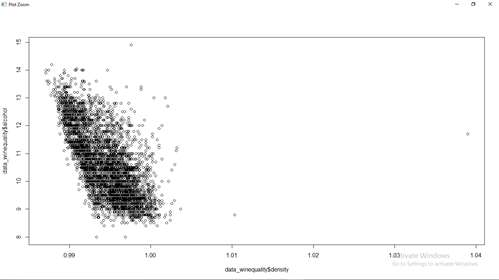
plot(data\_winequality$density,data\_winequality$pH)



Korelasi : positive

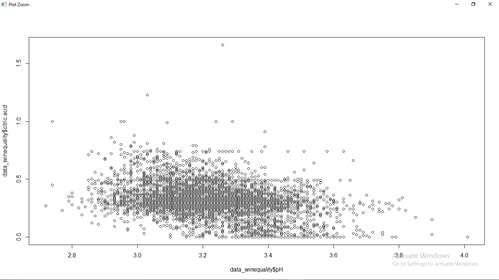
1. Visualisasi atribut density dengan alcohol

plot(data\_winequality$density,data\_winequality$alcohol)



Korelasi : negative

1. plot(data\_winequality$pH,data\_winequality$citric.acid)



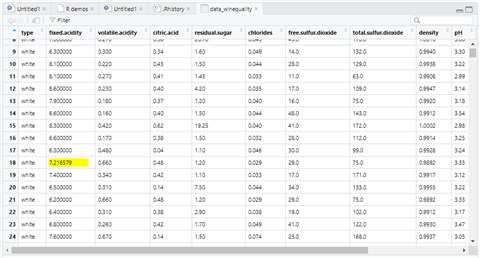
Korelasi : negative

1. Preprocessing data (Score: 50)Lakukan data preprosesing pada tiap atribut yang Anda miliki.  
   1) Jika data yang Anda miliki memiliki nilai atribut yang kosong (missing value), maka isilah nilai yang kosong tersebut dengan mean atau modus (tergantung tipe data yang Anda miliki)

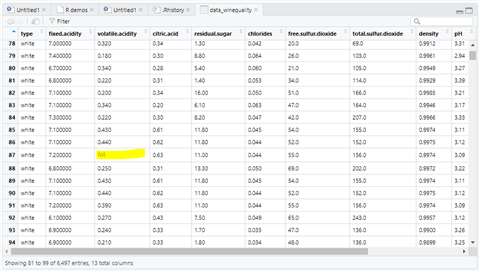
Jawab :

1. atribut fixed.acidity

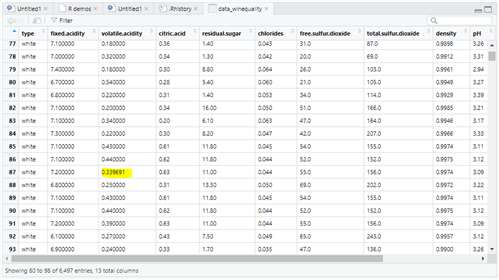
data\_winequality$fixed.acidity = ifelse(is.na(data\_winequality$fixed.acidity),ave(data\_winequality$fixed.acidity, FUN = function(x) mean(x, na.rm = 'TRUE')),data\_winequality$fixed.acidity)



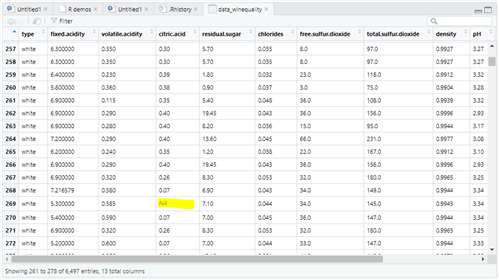
1. atribut volatile.acidity



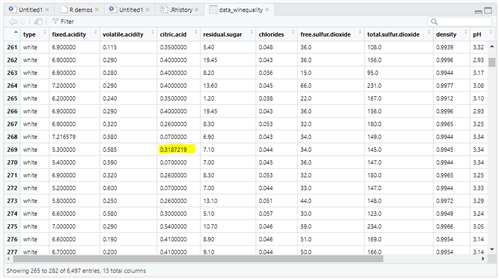
data\_winequality$volatile.acidity = ifelse(is.na(data\_winequality$volatile.acidity),ave(data\_winequality$volatile.acidity, FUN = function(x) mean(x, na.rm = 'TRUE')),data\_winequality$volatile.acidity)



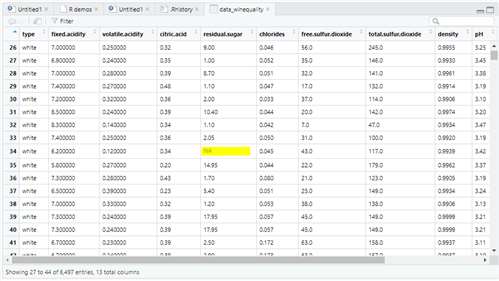
1. atribut citric.acid



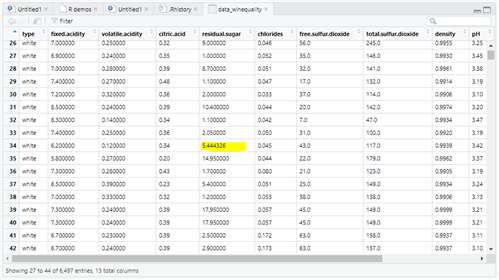
data\_winequality$citric.acid = ifelse(is.na(data\_winequality$citric.acid),ave(data\_winequality$citric.acid, FUN = function(x) mean(x, na.rm = 'TRUE')),data\_winequality$citric.acid)



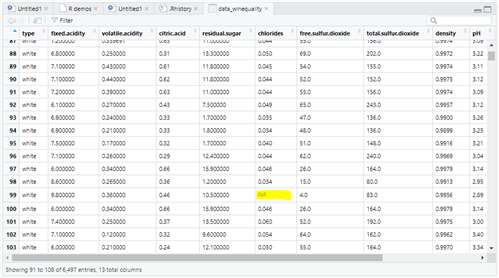
1. atribut residu.sugar



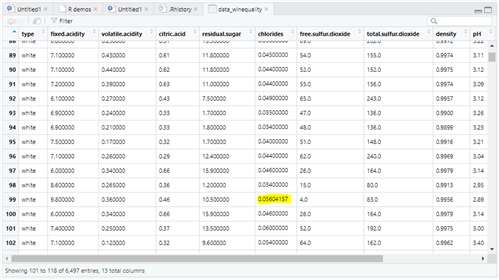
data\_winequality$residual.sugar = ifelse(is.na(data\_winequality$residual.sugar),ave(data\_winequality$residual.sugar, FUN = function(x) mean(x, na.rm = 'TRUE')),data\_winequality$residual.sugar)



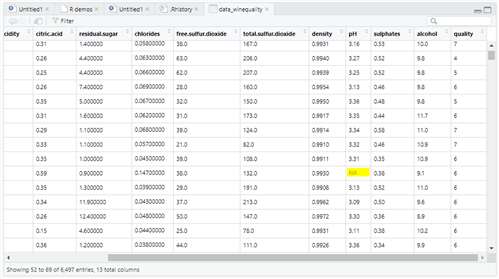
1. atribut chlorides



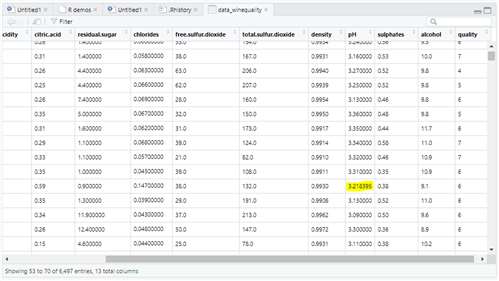
data\_winequality$chlorides = ifelse(is.na(data\_winequality$chlorides),ave(data\_winequality$chlorides, FUN = function(x) mean(x, na.rm = 'TRUE')),data\_winequality$chlorides)



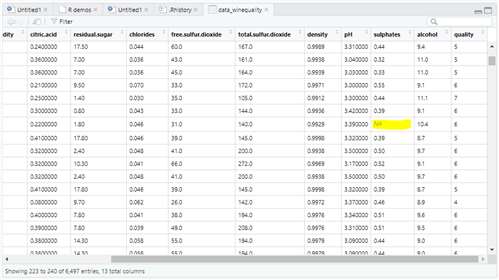
1. atribut pH



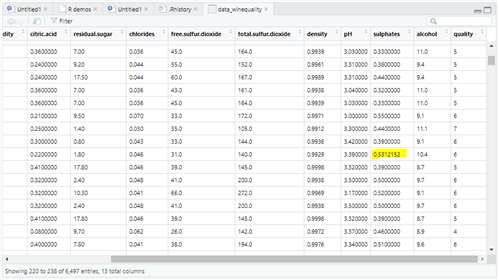
data\_winequality$pH = ifelse(is.na(data\_winequality$pH),ave(data\_winequality$pH, FUN = function(x) mean(x, na.rm = 'TRUE')),data\_winequality$pH)



1. atribut sulphates



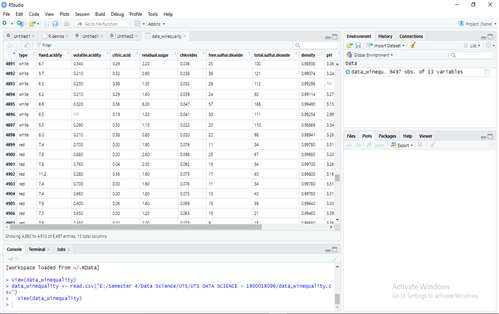
data\_winequality$sulphates = ifelse(is.na(data\_winequality$sulphates),ave(data\_winequality$sulphates, FUN = function(x) mean(x, na.rm = 'TRUE')),data\_winequality$sulphates)



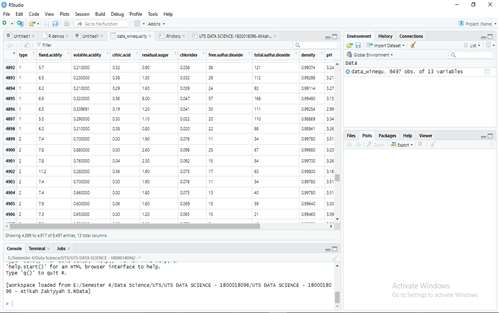
2) Jika data yang Anda miliki memiliki nilai yang tidak normal (anomali) maka gantilah nilai anomali tersebut dengan nilai terkecil atau terbesar dari atribut tsb (binning by boundary)

Jawab :

Dalam dataset yang saya ambil tidak terdapat data anomaly atau outlier



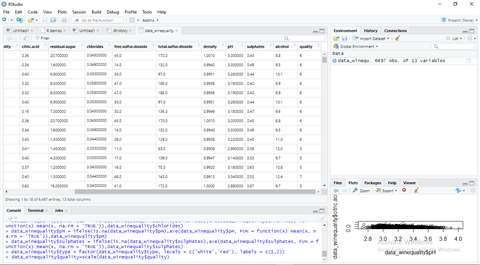
data\_winequality$type = factor(data\_winequality$type, levels = c('white','red'), labels = c(1,2))



3) Pilihlah salah satu atribut yang isinya real numbers pada Data yang Anda miliki, lakukan normalisasi pada atribut tersebut Gunakan Scaling pada R

Jawab :

Normalisasi atribut quality dengan menggunakan scaling



data\_winequality$quality=scale(data\_winequality$quality)

