Tugas Data Science

Fhrezha Zeaneth

9/4/2020

Untuk melakukan analisis data set, yang harus dilakukan yaitu mendownload data set. Disini data set yang digunakan minimal 5000 data yaitu 50000 Sales Records.

Import Data set dari csv ke RStudio

```
`Dataset` <- read.csv("~/FILE SEKOLAH/Data Science/50000 Sales Records.csv",
sep=";")
```

RINGKASAN DATA

Terdapat sejumlah fungsi yang akan sering gunakan untuk mengecek dataset yang dianalisa. Fungsi-fungsi tersebut antara lain:

head(): mengecek n (default 6) observasi teratas. tail(): mengecek n (default 6) observasi terbawah. str(): mengecek struktur data atau jenis data pada masing-masing kolom. Jenis data yang ada pada R dapat berupa num (numerik), int (integer), Factor(factor), date (tanggal), dan chr (karakter atau string). summary(): ringkasan data.

Berikut adalah contoh penerapan fungsi-fungsi tersebut pada dataset 50000 Sales Records 1. Mengecek 10 observasi teratas

```
head(Dataset,10)

##                 Region   Country Item.Type Sales.Channel Order.Priority
## 1  Sub-Saharan Africa   Namibia Household       Offline              M
## 2              Europe   Iceland Baby Food        Online              H
## 3              Europe    Russia      Meat        Online              L
## 4              Europe  Moldova       Meat        Online              L
## 5              Europe     Malta    Cereal        Online              M
## 6                Asia Indonesia      Meat        Online              H
## 7  Sub-Saharan Africa  Djibouti Household        Online              M
## 8              Europe    Greece Household        Online              L
## 9  Sub-Saharan Africa  Cameroon Cosmetics       Offline              M
## 10 Sub-Saharan Africa   Nigeria Cosmetics        Online              C
##    Order.Date  Order.ID  Ship.Date Units.Sold Unit.Price Unit.Cost
## 1   8/31/2015 897751939 10/12/2015       3604  668.27.00 502.54.00
## 2  11/20/2010 599480426 01/09/2011       8435  255.28.00 159.42.00
## 3   6/22/2017 538911855  6/25/2017       4848     421.89    364.69
## 4   2/28/2012 459845054  3/20/2012       7225     421.89    364.69
## 5  08/12/2010 626391351  9/13/2010       1975     205.70 117.11.00
## 6   8/20/2010 472974574  8/27/2010       2542     421.89    364.69
## 7  02/03/2011 854331052 03/03/2011       4398  668.27.00 502.54.00
## 8  09/11/2015 895509612  9/26/2015         49  668.27.00 502.54.00
## 9   1/31/2014 241871583 02/04/2014       4031  437.20.00 263.33.00
## 10 11/21/2015 409090793 12/07/2015       7911  437.20.00 263.33.00
##    Total.Revenue Total.Cost Total.Profit
## 1     2408445.08 1811154.16    597290.92
## 2     2153286.80 1344707.70    808579.10
## 3     2045322.72 1768017.12    277305.60
## 4     3048155.25 2634885.25    413270.00
## 5      406257.50  231292.25    174965.25
## 6     1072444.38  927041.98    145402.40
## 7     2939051.46 2210170.92    728880.54
## 8       32745.23   24624.46      8120.77
## 9     1762353.20 1061483.23    700869.97
## 10    3458689.20 2083203.63   1375485.57
```

2. Mengecek 10 observasi terbawah

```
tail(Dataset,10)
```

```
##                                   Region            Country       Item.Type
## 49991               Sub-Saharan Africa         Seychelles   Personal Care
## 49992     Middle East and North Africa             Kuwait       Household
## 49993                           Europe             Sweden          Cereal
## 49994            Australia and Oceania          Australia       Beverages
## 49995                           Europe         San Marino   Personal Care
## 49996 Central America and the Caribbean Dominican Republic       Baby Food
## 49997 Central America and the Caribbean               Cuba Office Supplies
## 49998                             Asia            Vietnam   Personal Care
## 49999               Sub-Saharan Africa       Sierra Leone         Clothes
## 50000               Sub-Saharan Africa            Eritrea      Vegetables
##       Sales.Channel Order.Priority Order.Date  Order.ID  Ship.Date
Units.Sold
## 49991        Online              H 01/03/2014 742188107  2/14/2014
8809
## 49992        Online              L  6/17/2016 650651824  6/18/2016
1188
## 49993       Offline              L  1/15/2011 454923951  1/29/2011
961
## 49994       Offline              M 12/14/2014 770678317 12/15/2014
9695
## 49995        Online              C 03/06/2016 933901250  4/24/2016
988
## 49996       Offline              C 02/12/2015 350891578  3/15/2015
5470
## 49997        Online              M  3/28/2013 748260629 05/03/2013
5803
## 49998       Offline              L 11/11/2016 322932231 12/18/2016
1678
## 49999        Online              M  8/28/2010 492142713  9/21/2010
4820
## 50000       Offline              L 06/05/2014 943440902  6/30/2014
```

```
3983
##       Unit.Price Unit.Cost Total.Revenue Total.Cost Total.Profit
## 49991      81.73     56.67     719959.57  499206.03    220753.54
## 49992  668.27.00 502.54.00     793904.76  597017.52    196887.24
## 49993     205.70 117.11.00     197677.70  112542.71     85134.99
## 49994   47.45.00     31.79     460027.75  308204.05    151823.70
## 49995      81.73     56.67      80749.24   55989.96     24759.28
## 49996  255.28.00 159.42.00    1396381.60  872027.40    524354.20
## 49997  651.21.00    524.96    3778971.63 3046342.88    732628.75
## 49998      81.73     56.67     137142.94   95092.26     42050.68
## 49999  109.28.00     35.84     526729.60  172748.80    353980.80
## 50000  154.06.00     90.93     613620.98  362174.19    251446.79
```

3. Mengecek struktur Dataset

```
str(Dataset)
```

```
## 'data.frame':    50000 obs. of  14 variables:
##  $ Region        : Factor w/ 7 levels "Asia","Australia and Oceania",..: 7
4 4 4 4 1 7 4 7 7 ...
##  $ Country       : Factor w/ 185 levels "Afghanistan",..: 112 68 133 105
100 70 41 59 26 119 ...
##  $ Item.Type     : Factor w/ 12 levels "Baby Food","Beverages",..: 7 1 8 8
3 8 7 7 5 5 ...
##  $ Sales.Channel : Factor w/ 2 levels "Offline","Online": 1 2 2 2 2 2 2 2
1 2 ...
##  $ Order.Priority: Factor w/ 4 levels "C","H","L","M": 4 2 3 3 4 2 4 3 4 1
...
##  $ Order.Date    : Factor w/ 2766 levels "01/01/2010","01/01/2011",..:
2639 1343 2294 1759 750 2557 114 832 989 1355 ...
##  $ Order.ID      : int  897751939 599480426 538911855 459845054 626391351
472974574 854331052 895509612 241871583 409090793 ...
##  $ Ship.Date     : Factor w/ 2811 levels "01/01/2011","01/01/2012",..:
1095 64 2338 1845 2683 2643 208 2782 123 1487 ...
##  $ Units.Sold    : int  3604 8435 4848 7225 1975 2542 4398 49 4031 7911
...
##  $ Unit.Price    : Factor w/ 12 levels "09.33","109.28.00",..: 11 6 7 7 5
7 11 11 8 8 ...
##  $ Unit.Cost     : Factor w/ 12 levels "0,313888889",..: 8 3 7 7 2 7 8 8 4
4 ...
##  $ Total.Revenue : Factor w/ 41172 levels "10001.76","1000113.40",..:
14447 12625 11833 18346 24311 1181 17717 19721 9423 20779 ...
##  $ Total.Cost    : Factor w/ 41154 levels "0,886111111",..: 8888 4210 8488
16246 13359 39957 12435 14709 832 11343 ...
##  $ Total.Profit  : Factor w/ 41163 levels "0,670833333",..: 31383 37214
17119 23345 10360 7015 35318 37297 34553 5982 ...
```

4. Meringkas Data

```
summary(Dataset)
```

```
##                                 Region
##  Asia                             : 7348
##  Australia and Oceania            : 4017
##  Central America and the Caribbean: 5451
##  Europe                           :12841
##  Middle East and North Africa     : 6128
##  North America                    : 1099
##  Sub-Saharan Africa               :13116
##                              Country              Item.Type
Sales.Channel
##  Trinidad and Tobago             :  321   Fruits       : 4221
Offline:24966
##  Guinea                          :  318   Meat         : 4221   Online
:25034
##  Cape Verde                      :  315   Cosmetics    : 4193
##  Maldives                        :  311   Vegetables   : 4191
##  Finland                         :  310   Personal Care: 4186
##  Democratic Republic of the Congo:  308   Beverages    : 4173
##  (Other)                         :48117   (Other)      :24815
##  Order.Priority      Order.Date       Order.ID             Ship.Date
##  C:12446        1/21/2017 :   34   Min.   :100013196   7/16/2014 :   35
##  H:12471        4/14/2013 :   32   1st Qu.:324007046   12/28/2012:   34
##  L:12588        05/03/2011:   31   Median :550422394   12/08/2014:   33
##  M:12495        12/29/2014:   31   Mean   :549733027   10/06/2011:   32
##                 2/24/2010 :   31   3rd Qu.:776782381   10/10/2010:   32
##                 5/28/2017 :   31   Max.   :999999463   11/17/2013:   32
##                 (Other)   :49810                       (Other)   :49802
##    Units.Sold        Unit.Price          Unit.Cost        Total.Revenue
##  Min.   :    1   09.33    : 4221   0,313888889: 4221   939431.90 :    7
##  1st Qu.: 2498   421.89   : 4221   364.69     : 4221   1415041.10:    5
##  Median : 5018   437.20.00: 4193   263.33.00  : 4193   142444.90 :    5
##  Mean   : 5000   154.06.00: 4191   90.93      : 4191   46110.63  :    5
##  3rd Qu.: 7493   81.73    : 4186   56.67      : 4186   738563.64 :    5
##  Max.   :10000   47.45.00 : 4173   31.79      : 4173   8465.73   :    5
##                  (Other)  :24815   (Other)    :24815   (Other)   :49968
##      Total.Cost        Total.Profit
##  534841.37:    7   404590.53 :    7
##  34675.26 :    5   11435.37  :    5
##  435918.42:    5   1641.25.00:    5
##  541709.16:    5   28768.88  :    5
##  624980.16:    5   302645.22 :    5
##  65057.16 :    5   325732.28 :    5
##  (Other)  :49968   (Other)   :49968
```

ANALISA STATISTIK

Fungsi-fungsi yang dapat digunakan untuk melakukan analisis statistika deskriptif adalah sebagai berikut:

mean() : menghitung nilai rata-rata variabel numerik. sd() : menghitung simpangan baku variabel numerik. var() : menghitung varians variabel numerik. median() : menghitung median suatu variabel numerik. range() : memperoleh nilai minimum dan maksimum suatu variabel numerik. IQR() : memperoleh nilai jarak antar kuartil. quantile() : memperoleh kuantil variabel numerik.

Berikut adalah contoh penerapan fungsi-fungsi tersebut: 1. Menghitung rata-rata Units Sold pada Dataset 50000 Sales Records

```
mean(Dataset$Units.Sold, na.rm = TRUE)

## [1] 4999.619
```

2. Menghitung median Units Sold pada Dataset 50000 Sales Records

```
median(Dataset$Units.Sold, na.rm = TRUE)

## [1] 5017.5
```

3. Menghitung simpangan baku Unit SOld pada Dataset 50000 Sales Records

```
sd(Dataset$Units.Sold, na.rm = TRUE)

## [1] 2884.335
```

4. Menghitung Varians Units Sold pada Dataset 50000 Sales Records

```
var(Dataset$Units.Sold, na.rm = TRUE)

## [1] 8319389
```

5. Menghitung range Units Sold pada Dataset 50000 Sales Records

```
range(Dataset$Units.Sold, na.rm = TRUE)

## [1]     1 10000
```

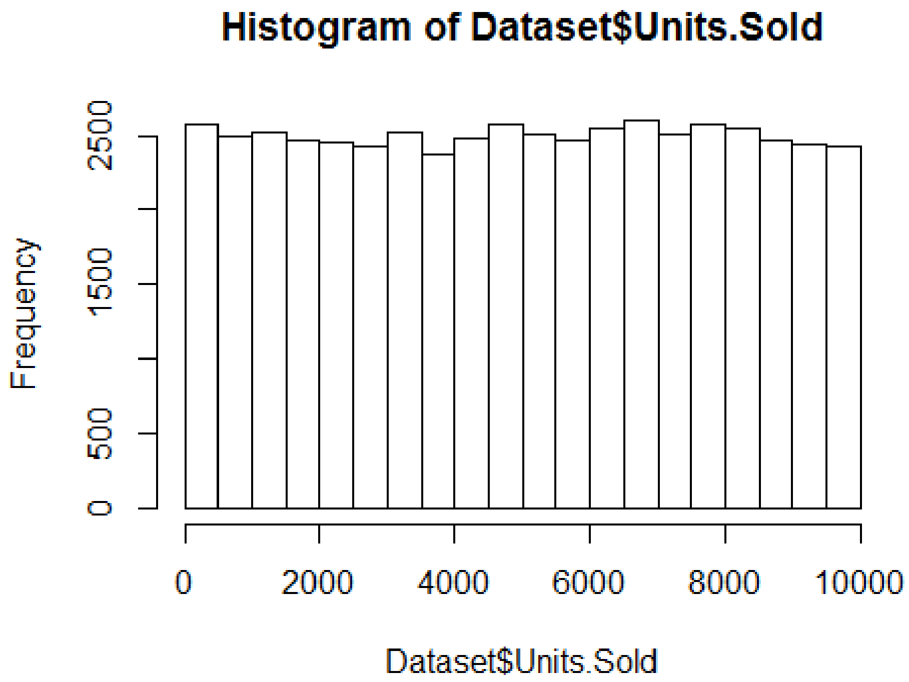
Visualisasi

Histogram

Fungsi hist() dapat digunakan untuk membuat histogram pada R. Secara sederhana fungsi tersebut didefinisikan sebagai berikut:

hist(x, breaks="Sturges") Catatan: x: vektor numerik breaks: breakpoints antar sel histogram.

```
hist(Dataset$Units.Sold)
```



Untuk analisis kolom lain, bisa melihat dari percobaan analisis diatas.