Tugas Data Science

Dio Cahyo Saputra

4/9/2020

#### Memanggil packages yang dibutuhkan

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.3

library(here)

## Warning: package 'here' was built under R version 3.6.3

## here() starts at D:/Daigaku/praktikum2020

### read.csv digunakan untuk memanggil file avocado.csv yang akan digunakan lalu fungsi here berisi posisi alamat directori sekarang

avocado\_df = read.csv(here("data-raw","avocado.csv"))

### summary untuk mengetahui nilai seperti mean median max min pada setiap variable

summary(avocado\_df)

## X Date AveragePrice Total.Volume   
## Min. : 0.00 1/1/2017 : 108 Min. :0.440 Min. : 85   
## 1st Qu.:10.00 1/10/2016: 108 1st Qu.:1.100 1st Qu.: 10839   
## Median :24.00 1/11/2015: 108 Median :1.370 Median : 107377   
## Mean :24.23 1/14/2018: 108 Mean :1.406 Mean : 850644   
## 3rd Qu.:38.00 1/15/2017: 108 3rd Qu.:1.660 3rd Qu.: 432962   
## Max. :52.00 1/17/2016: 108 Max. :3.250 Max. :62505647   
## (Other) :17601   
## X4046 X4225 X4770 Total.Bags   
## Min. : 0 Min. : 0 Min. : 0 Min. : 0   
## 1st Qu.: 854 1st Qu.: 3009 1st Qu.: 0 1st Qu.: 5089   
## Median : 8645 Median : 29061 Median : 185 Median : 39744   
## Mean : 293008 Mean : 295155 Mean : 22840 Mean : 239639   
## 3rd Qu.: 111020 3rd Qu.: 150207 3rd Qu.: 6243 3rd Qu.: 110783   
## Max. :22743616 Max. :20470573 Max. :2546439 Max. :19373134   
##   
## Small.Bags Large.Bags XLarge.Bags type   
## Min. : 0 Min. : 0 Min. : 0.0 conventional:9126   
## 1st Qu.: 2849 1st Qu.: 127 1st Qu.: 0.0 organic :9123   
## Median : 26363 Median : 2648 Median : 0.0   
## Mean : 182195 Mean : 54338 Mean : 3106.4   
## 3rd Qu.: 83338 3rd Qu.: 22029 3rd Qu.: 132.5   
## Max. :13384587 Max. :5719097 Max. :551693.7   
##   
## year region   
## Min. :2015 Albany : 338   
## 1st Qu.:2015 Atlanta : 338   
## Median :2016 BaltimoreWashington: 338   
## Mean :2016 Boise : 338   
## 3rd Qu.:2017 Boston : 338   
## Max. :2018 BuffaloRochester : 338   
## (Other) :16221

#### untuk mengetahui detail dari data.frame avocado\_df salah satunya tipe data pada setiap variable dan beberapa isi data pada avocado\_df

str(avocado\_df)

## 'data.frame': 18249 obs. of 14 variables:  
## $ X : int 0 1 2 3 4 5 6 7 8 9 ...  
## $ Date : Factor w/ 169 levels "1/1/2017","1/10/2016",..: 54 51 48 58 42 39 36 45 33 27 ...  
## $ AveragePrice: num 1.33 1.35 0.93 1.08 1.28 1.26 0.99 0.98 1.02 1.07 ...  
## $ Total.Volume: num 64237 54877 118220 78992 51040 ...  
## $ X4046 : num 1037 674 795 1132 941 ...  
## $ X4225 : num 54455 44639 109150 71976 43838 ...  
## $ X4770 : num 48.2 58.3 130.5 72.6 75.8 ...  
## $ Total.Bags : num 8697 9506 8145 5811 6184 ...  
## $ Small.Bags : num 8604 9408 8042 5677 5986 ...  
## $ Large.Bags : num 93.2 97.5 103.1 133.8 197.7 ...  
## $ XLarge.Bags : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ type : Factor w/ 2 levels "conventional",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ year : int 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...  
## $ region : Factor w/ 54 levels "Albany","Atlanta",..: 1 1 1 1 1 1 1 1 1 1 ...

### quantile untuk mengetahui nilai max min quantile 1 median quantile 3 pada 1 variable saja

quantile(avocado\_df$Total.Volume)

## 0% 25% 50% 75% 100%   
## 84.56 10838.58 107376.76 432962.29 62505646.52

### median untuk mengetahui nilai median/nilai tengah pada 1 variable saja

median(avocado\_df$Total.Volume)

## [1] 107376.8

### min untuk mengetahui nilai minimal pada 1 variable saja

min(avocado\_df$Total.Volume)

## [1] 84.56

### max untuk mengetahui nilai maximum pada 1 variable saja

max(avocado\_df$Total.Volume)

## [1] 62505647

### mean untuk mengetahui nilai rata-rata pada 1 variable saja

mean(avocado\_df$Total.Volume)

## [1] 850644

### ggplot untuk membuat sebuah plot, disini membuat beberapa plot berdasarkan tahun

ggplot(avocado\_df, aes(x = AveragePrice, y = Total.Bags/1000000)) +  
 geom\_point(aes(colour = type), alpha = 0.2) +  
 facet\_wrap(~year, scales = "free")

