Praktikum 1 – Data Analitik  
**HUBUNGAN ANTARA SEBUAH VARIABEL KATEGORIK DENGAN SEBUAH VARIABEL NUMERIK**

Dalam praktikum ini akan dipelajari dan dipraktekkan bagaimana mengetahui adanya hubungan antar variabel dalam sebuah data multivariat atau data yang berisi lebih dari dua variabel. Variabel sendiri dapat diartikan sebagai sebuah konsep yang memiliki variasi nilai. Hubungan (relationship) antar variable dapat terjadi antara sebuah variabel kategorik dengan sebuah variabel kontinu, antar dua variabel kategorik serta antar dua variabel kontinu. Data latihan yang dipergunakan adalah data mtcars, airquality dan iris yang merupakan data latihan yang tersimpan di R.

|  |
| --- |
| > library(RMySQL)  > con = dbConnect(MySQL(), user = 'root', password = '', dbname =  + 'db\_iris', host = 'localhost')  > myQuery <- "select \* from iris;"  > df <- dbGetQuery(con, myQuery)  > View(df) |

**Hubungan antara sebuah variabel kategorik dengan sebuah variabel kontinu**

> by(df$Petal\_width, df$Species\_name, mean)

df$Species\_name: Setosa

[1] 0.246

---------------------------------------------------------------------------------------

df$Species\_name: Verginica

[1] 2.026

---------------------------------------------------------------------------------------

df$Species\_name: Versicolor

[1] 1.326

> by(df$Petal\_width, df$Species\_name, sd)

df$Species\_name: Setosa

[1] 0.1053856

---------------------------------------------------------------------------------------

df$Species\_name: Verginica

[1] 0.2746501

---------------------------------------------------------------------------------------

df$Species\_name: Versicolor

[1] 0.1977527

> by(df$Petal\_width, df$Species\_name, summary)

df$Species\_name: Setosa

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.100 0.200 0.200 0.246 0.300 0.600

---------------------------------------------------------------------------------------

df$Species\_name: Verginica

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.400 1.800 2.000 2.026 2.300 2.500

---------------------------------------------------------------------------------------

df$Species\_name: Versicolor

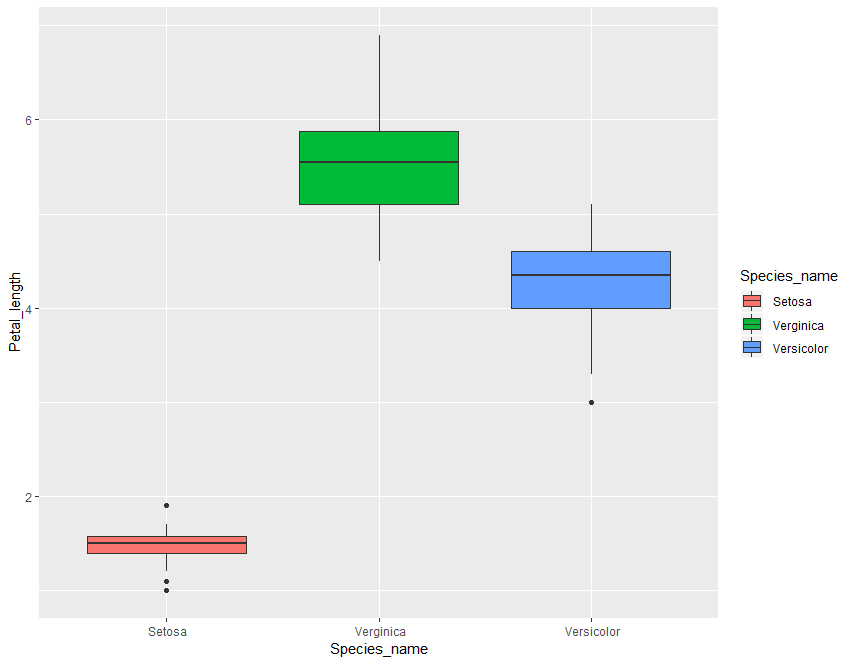
Min. 1st Qu. Median Mean 3rd Qu. Max.

1.000 1.200 1.300 1.326 1.500 1.800

**Metode visualisasi**

> library(ggplot2)

> qplot(Species\_name,Petal\_length, data=df, geom="boxplot", fill=Species\_name)



**Tugas 1 :**

Data: House Price Data

Data menunjukkan contoh harga rumah (dan karakteristik rumah yang terkait)  
untuk area metropolitan AS. Data antara lain menunjukkan identitas rumah, harga jual   
(dalam US $), luas (dalam kaki persegi), jumlah kamar tidur dan kamar mandi, jenis dinding serta lokasi rumah.  
Tujuan utama analisis dalam tugas 1 ini melihat adanya hubungan antara sebuah variabel numerik dengan sebuah variabel kategorik antara lain jenis dinding serta lokasi rumah. Misalnya apakah bahan dinding  
(yaitu, batu bata vs nonbrick) memiliki dampak signifikan pada harga? Atau, apakah penting  
lingkungan mana rumah itu berada? Jawaban atas pertanyaan-pertanyaan ini bisa membantu  
calon pembeli memutuskan berapa besar tawaran untuk rumah.

Gunakan analisis statistiska deskriptif baik yang bersifat numerik (rerata, simpangan baku, dll) serta beberapa metode grafikal (boxplot, density plot, histogram, dll). Berikan penjelasan untuk setiap output yang dihasilkan serta tulis R code yang digunakan. Tampilan grafik yang berwarna tentunya lebih baik.

Catatan : Kalimat yang dipergunakan untuk membuat penjelasan setiap output antar praktikan tidak boleh sama

|  |
| --- |
| > library(RMySQL)  > con = dbConnect(MySQL(), user = 'root', password = '', dbname =  + ‘direct\_marketing’, host = 'localhost')  > myQuery\_directmarketing <- "select \* from direct\_marketing;"  > df\_directmarketing <- dbGetQuery(con, myQuery\_directmarketing)  > View(df\_directmarketing) |

**Hubungan antara sebuah variabel kategorik (Age & Married) dengan sebuah variabel kontinu (Salary)**

> by(df\_directmarketing$Salary, df\_directmarketing$Age, mean)

df\_directmarketing$Age: Middle

[1] 72036.42

-----------------------------------------------------------------------

df\_directmarketing$Age: Old

[1] 56365.85

-----------------------------------------------------------------------

df\_directmarketing$Age: Young

[1] 27715.68

Catatan = Dari data diatas didapatkan rata-rata salary tertinggi pada Age=”middle” disusul oleh Age=”Old” dan terakhir dengan nilai rata-rata Salary terkecil adalah Age=”Young”.

> by(df\_directmarketing$Salary, df\_directmarketing$Married, mean)

df\_directmarketing$Married: Married

[1] 76696.41

-----------------------------------------------------------------------

df\_directmarketing$Married: Single

[1] 35345.98

Catatan = Dari data diatas didapatkan rata-rata salary tertinggi pada Married=’Married’.

> by(df\_directmarketing$Salary, df\_directmarketing$Age, sd)

df\_directmarketing$Age: Middle

[1] 24753.01

-----------------------------------------------------------------------

df\_directmarketing$Age: Old

[1] 30176.1

-----------------------------------------------------------------------

df\_directmarketing$Age: Young

[1] 16699.38

Catatan = Dari data diatas didapatkan simpangan baku salary tertinggi pada Age=”Old”. Ini menunjukkan bahwa usia “Old” memiliki rentang penghasilan yang jauh.

> by(df\_directmarketing$Salary, df\_directmarketing$Married, sd)

df\_directmarketing$Married: Married

[1] 26204.42

-----------------------------------------------------------------------

df\_directmarketing$Married: Single

[1] 18217.4

Catatan = Dari data diatas didapatkan simpangan baku salary tertinggi pada Married=”Married”. Ini menunjukkan bahwa usia Married=“Married” memiliki rentang penghasilan yang jauh.

> by(df\_directmarketing$Salary, df\_directmarketing$Age, summary)

df\_directmarketing$Age: Middle

Min. 1st Qu. Median Mean 3rd Qu. Max.

25300 51900 68450 72036 90425 140700

-----------------------------------------------------------------------

df\_directmarketing$Age: Old

Min. 1st Qu. Median Mean 3rd Qu. Max.

10100 33000 54600 56366 74800 168800

-----------------------------------------------------------------------

df\_directmarketing$Age: Young

Min. 1st Qu. Median Mean 3rd Qu. Max.

10200 15050 21400 27716 35400 80700

Catatan = Dari data diatas dapat dilihat review data dari ketiga Age. s

> by(df\_directmarketing$Salary, df\_directmarketing$Married, summary)

df\_directmarketing$Married: Married

Min. 1st Qu. Median Mean 3rd Qu. Max.

10100 60025 76700 76696 96375 168800

-----------------------------------------------------------------------

df\_directmarketing$Married: Single

Min. 1st Qu. Median Mean 3rd Qu. Max.

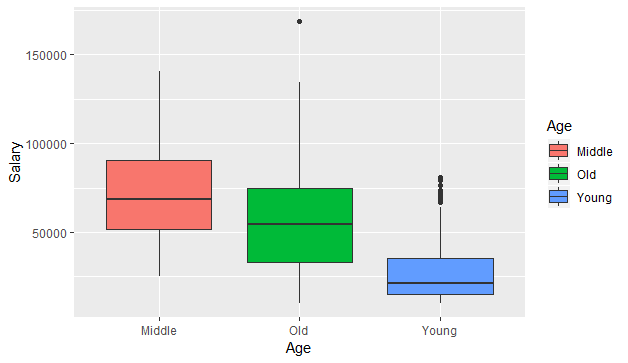
10200 18250 33100 35346 49975 82800

Catatan = Dari data diatas dapat dilihat review data dari kedua jenis status Married.

**Metode visualisasi**

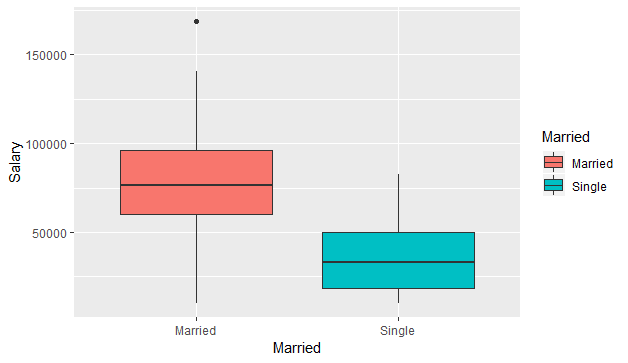
> library(ggplot2)

> qplot(Age,Salary, data=df\_directmarketing, geom="boxplot", fill=Age)



Catatan: dari tampilan grafik boxplot Hubungan Age dan salary diatas dapat dilihat bahwa Rentang Usia Age=”Middle” memiliki Salary yang paling tinggi. Disusul oleh rentang usia Age=”old” dengan rentang salary ditengah. Dan Age=”Young” yang memiliki salary paling rendah.

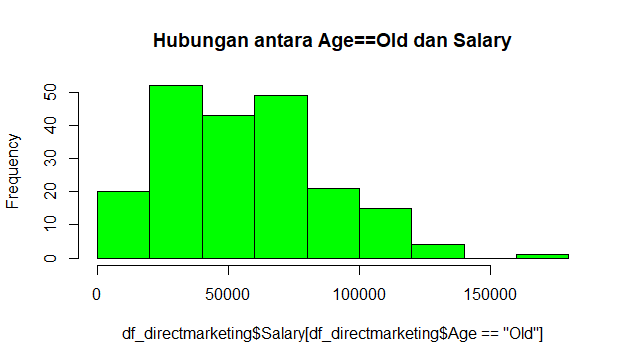
> qplot(Married,Salary, data=df\_directmarketing, geom="boxplot", fill=Married)



Catatan: dari tampilan grafik boxplot Hubungan status Married dan salary diatas dapat dilihat bahwa Married=”Married” memiliki Salary yang lebih tinggi dibanding Married=”Single” yang memiliki salary paling rendah. Terdapat nilai outlier pada Married=Married pada titik salary >150000.

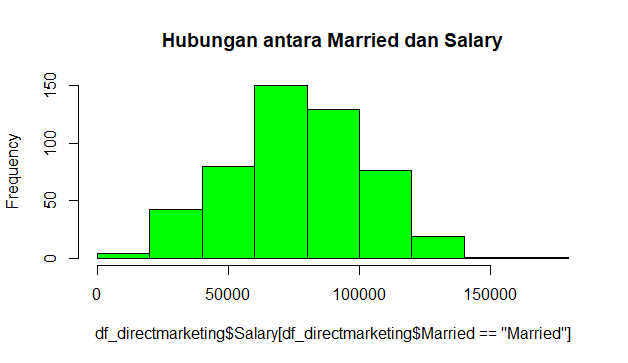
**Visualisasi Data dengan Histogram**

> hist(df\_directmarketing$Salary[df\_directmarketing$Age=="Old"], main="Hubungan antara Age dan Salary",col = c("green"))



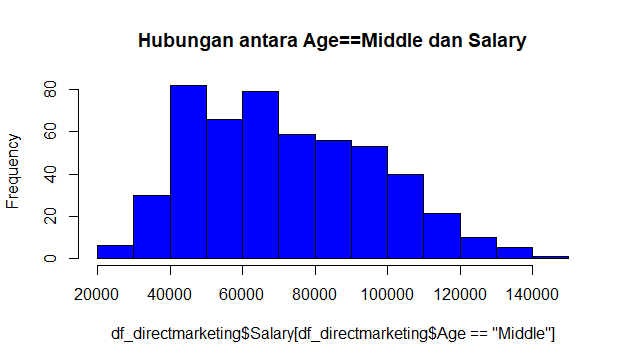
Catatan: dari tampilan grafik Histogram Hubungan Age=Old dan salary diatas dapat dilihat bahwa rentang usia Age=”Old” memiliki Salary yang ‘rata tengah’ sedikit beragam dengan 3 nilai salary utama yang mencolok dan simpangan baku yang besar. Nilai Modus atau paling banyak muncul pada frek=50 dan antara ± 20000-40000. Terdapat sedikit frekuensi dengan rentang salary tertinggi bernilai >150000.

> hist(df\_directmarketing$Salary[df\_directmarketing$Married=="Married"], main="Hubungan antara Married dan Salary",col = c("green"))



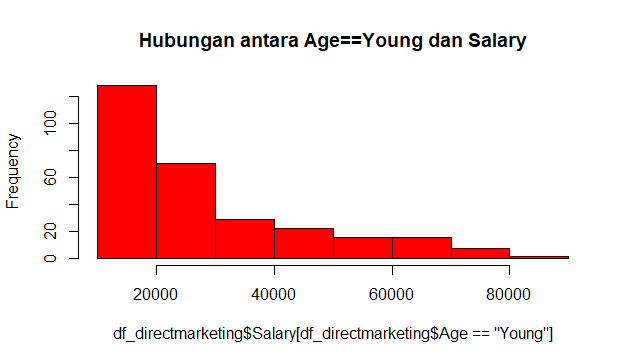
Catatan: dari tampilan grafik Histogram Hubungan Married=Married dan salary diatas dapat dilihat bahwa rentang usia Married=”Married” memiliki Salary yang ‘rata tengah’ sedikit beragam dengan 2 nilai salary utama yang mencolok dan simpangan baku yang besar. Nilai Modus atau paling banyak muncul pada frek=150 dan antara ± 60000-80000.

hist(df\_directmarketing$Salary[df\_directmarketing$Age=="Middle"], main="Hubungan antara Age==Middle dan Salary",col = c("blue"))



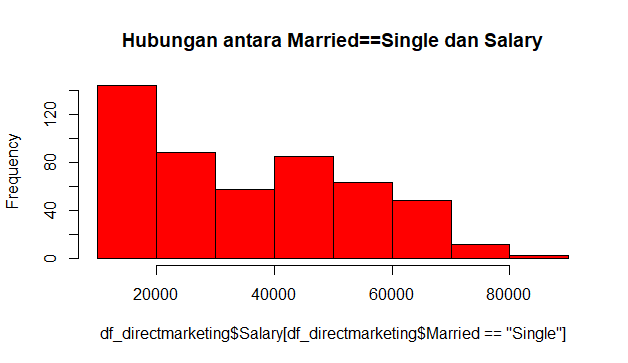
Catatan: dari tampilan grafik histogram “Hubungan Age==Middle dan salary” diatas dapat dilihat bahwa rentang usia Age=”Middle” memiliki Salary yang rata tengah yang menunjukkan salary kurang beragam dan cenderung homogen, dengan simpangan baku yang kecil dan nilai Modus atau paling banyak muncul pada frek=80 dan antara ± 40000-50000.

> hist(df\_directmarketing$Salary[df\_directmarketing$Age=="Young"], main="Hubungan antara Age==Young dan Salary",col = c("red"))



Catatan: dari tampilan grafik histogram “Hubungan Age==Young dan salary” diatas dapat dilihat bahwa rentang usia Age=”Young” memiliki Salary yang signifikan pada frekuensi dan menunjukkan rentang salary kecil. Terdapat nilai Modus atau paling banyak muncul pada frek>140 dan antara ± 10000-20000.

> hist(df\_directmarketing$Salary[df\_directmarketing$Married=="Single"], main="Hubungan antara Married==Single dan Salary",col = c("red"))

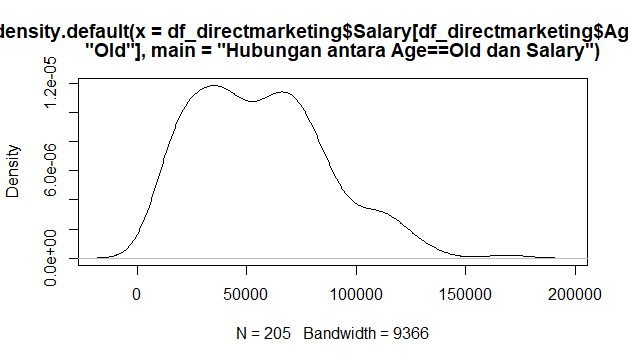


Catatan: dari tampilan grafik histogram “Hubungan Married==Single dan salary” diatas dapat dilihat bahwa rentang usia Married=”Single” memiliki Salary yang rata kiri yang menunjukkan salary rendah dengan simpangan baku yang kecil dan nilai Modus atau paling banyak muncul pada frek=160 dan antara ± 0-20000.

**Visualisai Data dengan Density Plot**

> density\_old=density(df\_directmarketing$Salary[df\_directmarketing$Age=="Old"])

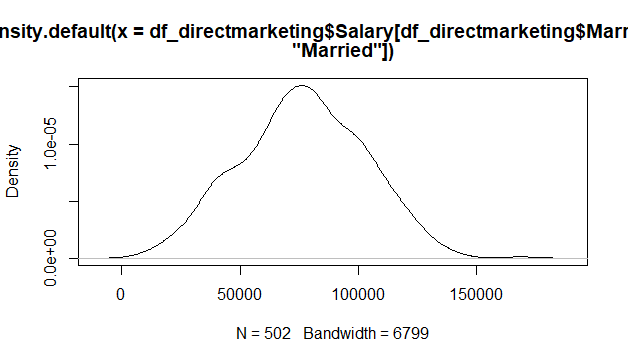
> plot(density\_old)



Catatan: dari tampilan grafik density Hubungan Age=Old dan salary diatas dapat dilihat bahwa rentang usia Age=”Old” memiliki Salary yang ‘rata tengah’ sedikit beragam dengan puncak nilai salary utama yang menonjol antara 0-100000. Nilai Modus atau paling banyak muncul pada frek=50 dan antara ± 20000-40000. Terdapat sedikit frekuensi dengan rentang salary tertinggi bernilai >150000.

> density\_married=density(df\_directmarketing$Salary[df\_directmarketing$Married=="Married"])

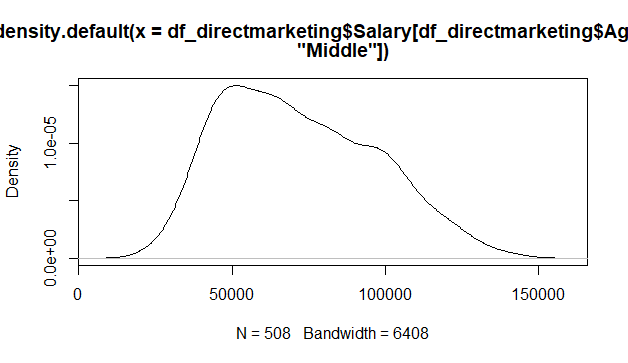
> plot(density\_married)



Catatan: dari tampilan grafik density gram Hubungan Married=Married dan salary diatas dapat dilihat bahwa rentang usia Married=”Married” memiliki Salary yang ‘rata tengah’ sedikit beragam dengan 2 nilai salary utama yang mencolok dan simpangan baku yang besar. Nilai Modus atau paling banyak muncul pada frek=150 dan antara ± 60000-80000.

> density\_middle=density(df\_directmarketing$Salary[df\_directmarketing$Age=="Middle"])

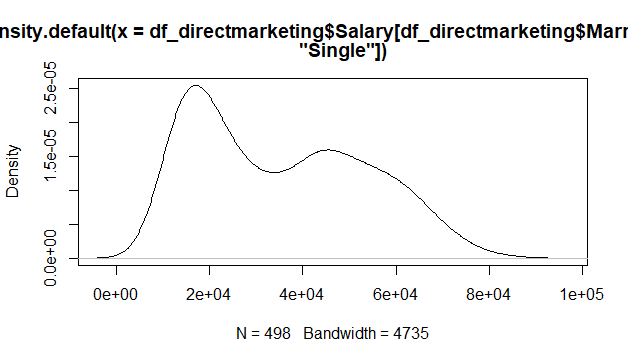
> plot(density\_middle)



Catatan: dari tampilan grafik density “Hubungan Age==Middle dan salary” diatas dapat dilihat bahwa rentang usia Age=”Middle” memiliki Salary yang rata tengah yang menunjukkan salary kurang beragam dan cenderung homogen antara ±30000-120000 , dengan nilai Modus atau paling banyak muncul pada frek=80 dan antara ± 40000-50000.

> density\_single=density(df\_directmarketing$Salary[df\_directmarketing$Married=="Single"])

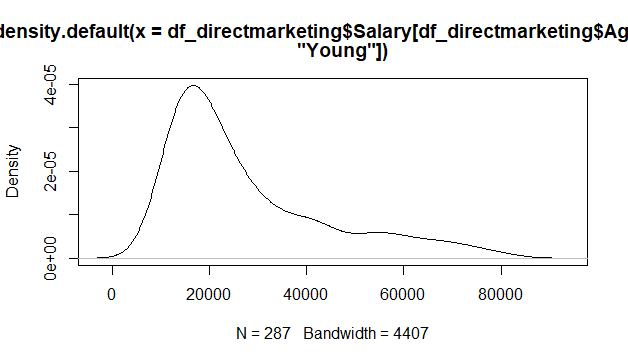
> plot(density\_single)



Catatan: dari tampilan grafik density gram “Hubungan Married==Single dan salary” diatas dapat dilihat bahwa rentang usia Married=”Single” memiliki Salary yang rata kiri yang menunjukkan salary rendah dengan simpangan baku yang kecil dan nilai Modus atau paling banyak muncul pada frek=160 dan antara ± 0-20000.

> density\_young=density(df\_directmarketing$Salary[df\_directmarketing$Age=="Young"])

> plot(density\_young)



Catatan: dari tampilan grafik density “Hubungan Age==Young dan salary” diatas dapat dilihat bahwa rentang usia Age=”Young” memiliki Salary yang signifikan yang menunjukkan terdapat rentang salary yang ekstrim antara ±10000 - 30000. Terdapat nilai Modus atau paling banyak muncul pada frek>140 dan antara ±10000-20000.