Praktikum 2 – Data Analitik  
**HUBUNGAN ANTAR DUA VARIABEL KATEGORIK DAN**

**ANTAR DUA VARIABEL NUMERIK**

Dalam praktikum ini akan dipelajari dan dipraktekkan bagaimana mengetahui adanya hubungan antar dua variabel kategorik serta antar dua variabel numerik. Hubungan antar dua variabel kategori dapat diketahui menggunakan teknik tabulasi silang atau cross tabulation serta uji statistiknya menggunakan uji khi-kuadrat atau uji kebebasan. Sedangkan hubungan antar dua variabel numerik dapat dilihat menggunakan Teknik scatter plot serta koefisien korelasi. Nilai koefisien korelasi berkisar dari -1 sd 1.

**PRAKTIKUM**

**Hubungan antar dua variabel kategorik**

**Note : Setiap data harap dimasukkan ke dalam database terlebih dahulu. Kemudian akses melalui**

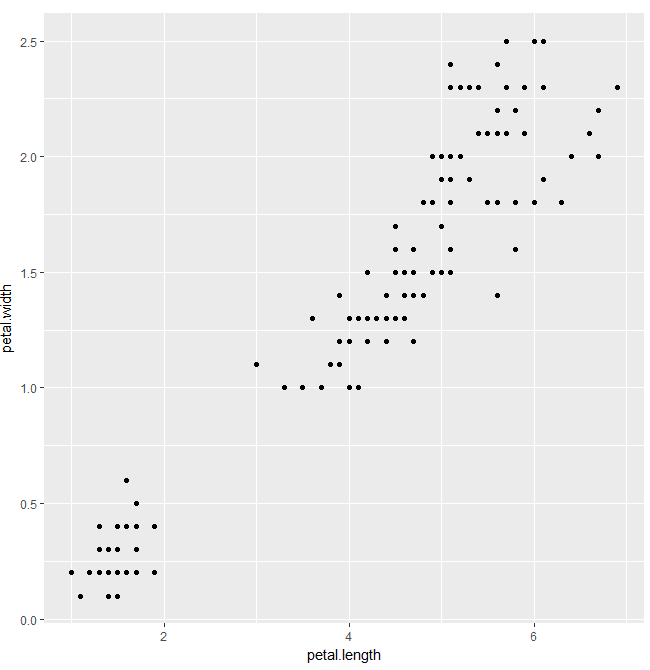
**database**

|  |
| --- |
| > data()  > View(Titanic)  > cross <- xtabs(Freq ~ Class+Sex, data=Titanic)  > cross  Sex  Class Male Female  1st 180 145  2nd 179 106  3rd 510 196  Crew 862 23  > prop.table(cross, 1)  Sex  Class Male Female  1st 0.5538462 0.4461538  2nd 0.6280702 0.3719298  3rd 0.7223796 0.2776204  > chisq.test(cross)  Pearson's Chi-squared test  data: cross  X-squared = 349.91, df = 3, p-value < 2.2e-16 |

|  |
| --- |
| Update R  > install.packages("installr")  > require(installr)  > updateR() |

**Hubungan antar dua variabel numerik**

|  |
| --- |
| > head(iris)  sepal.length sepal.width petal.length petal.width iris X X.1 X.2 X.3  1 5.1 3.5 1.4 0.2 Iris-setosa NA NA NA NA  2 4.9 3.0 1.4 0.2 Iris-setosa NA NA NA NA  3 4.7 3.2 1.3 0.2 Iris-setosa NA NA NA NA  4 4.6 3.1 1.5 0.2 Iris-setosa NA NA NA NA  5 5.0 3.6 1.4 0.2 Iris-setosa NA NA NA NA  6 5.4 3.9 1.7 0.4 Iris-setosa NA NA NA NA  > cor(iris$sepal.length, iris$sepal.width)  [1] -0.1093692  > cor(iris$petal.length, iris$petal.width)  [1] 0.9627571  > library(ggplot2)  > ggplot(iris, aes(x=petal.length, y=petal.width)) + geom\_point() |



**Scatter plot**

**TUGAS**

**Tugas 1:**

Data: House Price Data

Tujuan utama analisis dalam tugas 2 ini adalah melihat adanya hubungan antar dua variabel kategorik (misalnya bahan dinding dengan lokasi) serta hubungan antar dua variabel numerik (misalnya harga rumah dengan luas bangunan). Jawaban atas pertanyaan-pertanyaan ini bisa membantu  
calon pembeli memutuskan berapa besar tawaran untuk rumah.

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

**Hubungan antar dua variabel kategorik**

|  |
| --- |
| > library(RMySQL)  > con = dbConnect(MySQL(), user = 'root', password = '', dbname = 'houseprice', host = 'localhost')  > Query\_houseprice <- "select \* from houseprice\_table;"  > df\_houseprice <- dbGetQuery(con, Query\_houseprice)  > cross <- xtabs(Price ~ Neighborhood+Brick, data=df\_houseprice)  > cross  Brick  Neighborhood No Yes  East 3061500 2573900  North 4017600 829200  West 3409300 2803200  > prop.table(cross, 1)  Brick  Neighborhood No Yes  East 0.5432622 0.4567378  North 0.8289180 0.1710820  West 0.5487807 0.4512193  > chisq.test(cross)  Pearson's Chi-squared test  data: cross  X-squared = 1177900, df = 2, p-value < 2.2e-16 |

Dari hasil R diatas, dapat dilihat bahwa Neighborhood = North dan Brick=No, memiliki total Harga rumah paling besar,dengan nilai proportion terbesar. Sedangkan, Neighborhood = North dan Brick=Yes, memiliki total Harga rumah paling kecil,dengan nilai proportion terkecil. Dapat dilihat bahwa terdapat hubungan antara Neighborhood dengan Brick terkait dengan jumlah Price.

Dilihat dari keseluruhan Neighborhood, Jumlah Price dengan Brick = No memiliki total Price yang lebih besar ketimbang Neighborhood dengan Brick = Yes.

**Hubungan antar dua variabel numerik**

|  |
| --- |
| > head(df\_houseprice)  Price SqFt Bedrooms Bathrooms Offers Brick Neighborhood  1 114300 1790 2 2 2 No East  2 114200 2030 4 2 3 No East  3 114800 1740 3 2 1 No East  4 94700 1980 3 2 3 No East  5 119800 2130 3 3 3 No East  6 114600 1780 3 2 2 No North  > cor(df\_houseprice$Price, df\_houseprice$SqFt)  [1] 0.5529822  > library(ggplot2)  > ggplot(df\_houseprice, aes(x=SqFt, y=Price, color=Neighborhood)) + geom\_point() |

Dari grafik plot diatas maka didapatkan kesimpulan bahwa hubungan antar SqFt dan Price berbanding lurus, sehingga semakin tinggi nilai SqFt makan akan semakin tinggi pula nilai Price nya.

Nilai koefisein korelasi Price dan SqFt memiliki nilai=0.5529822

**Tugas 2:**

Cari contoh data sendiri. Lihat apakah di dalam datanya terdapat hubungan antar variable kategorik dan antar variable numerik. Buat tampilan grafiknya berwarna.

Data = DirectMarketing

**Hubungan antar dua variabel kategorik**

|  |
| --- |
| > library(RMySQL)  > con2 = dbConnect(MySQL(), user = 'root', password = '', dbname = 'direct\_marketing', host = 'localhost')  > myQuery\_directmarketing <- "select \* from direct\_marketing;"  > df\_directmarketing <- dbGetQuery(con2, myQuery\_directmarketing)  >  > cross2 <- xtabs(AmountSpent ~ Age+OwnHome, data=df\_directmarketing)  > cross2  OwnHome  Age Own Rent  Middle 534259 228600  Old 230908 62678  Young 31091 129234  > prop.table(cross2, 1)  OwnHome  Age Own Rent  Middle 0.7003378 0.2996622  Old 0.7865089 0.2134911  Young 0.1939248 0.8060752  > chisq.test(cross2)  Pearson's Chi-squared test  data: cross2  X-squared = 180090, df = 2, p-value < 2.2e-16 |

Dari hasil R diatas, dapat dilihat bahwa Age=Middle dan OwnHome=Own, memiliki total AmountSpent paling besar. Sedangkan, Age=Young dan OwnHome=Own, memiliki total AmountSpent paling kecil, Dapat dilihat bahwa terdapat factor penentu yaitu Age dan OwnHome terkait penentuan jumlah AmountSpent yang beragam. Jumlah AmountSpent pada kategori.

Dilihat dari keseluruhan Age, Jumlah AmountSpent dengan OwnHome=Own memiliki total Price yang lebih besar ketimbang OwnHome=Rent.

**Hubungan antar dua variabel numerik**

|  |
| --- |
| > head(df\_directmarketing)  Age Gender OwnHome Married Location Salary Children History Catalogs AmountSpent  1 Old Female Own Single Far 47500 0 High 6 755  2 Middle Male Rent Single Close 63600 0 High 6 1318  3 Young Female Rent Single Close 13500 0 Low 18 296  4 Middle Male Own Married Close 85600 1 High 18 2436  5 Middle Female Own Single Close 68400 0 High 12 1304  6 Young Male Own Married Close 30400 0 Low 6 495  > cor(df\_directmarketing$AmountSpent, df\_directmarketing$Salary)  [1] 0.6995957  > library(ggplot2)  > ggplot(df\_directmarketing, aes(x=Salary, y=AmountSpent, color=Age)) + geom\_point() |

Dapat disimpulkan hubungan antara dua variable numerik Salary dan AmounSpent memiliki hubungan berbanding lurus, semakin tinggi nilai salary, semakin tinggi AmounSpentnya. AmountSpent tinggi, didominasi oleh Age=Middle dan Old.

Nilai koefisein korelasi AMountSpent dan Salary memiliki nilai=0.6995957