**ANALISIS DATA MAHASISWA POLITEKNIK HARAPAN BERSAMA KOTA TEGAL MENGGUNAKAN METODE SIMPLE RANDOM SAMPLING (SRS)**

****

**DISUSUN OLEH:**

**NABILA ASSHAFA PUTRI (20090105)**

**TEKNIK INFORMATIKA**

**POLITEKNIK HARAPAN KOTA TEGAL**

**TAHUN 2021**

**BAB I**

**DATA**

Data atau informasi yang diperoleh melalui proses pengumpulan data menggunakan Simple Random Sampling (SRS) atau pengambilan data secara acak. Metode ini dapat dilakukan dg pengambilan sampel secara acak dari anggota dari populasi.

Metode SRS dalam pengumpulan data ini dilakukan dengan mengisi angket yang disebar luaskan pada mahasiswa PHB. Data yang diperoleh melalui proses pengumpulan angket, digunakan untuk menjawab masalah penelitian atau untuk menguji hipotesis penelitian yang diajukan. Pengumpulan data ini memakan waktu kurang lebih 7 hari.

Setelah proses pengumpulan data dilakukan dengan cermat dan lengkap, selanjutnya yang perlu dilakukan adalah langkah-langkah analisis data sebagai berikut :

1. Sebelum melakukan analisis data menggunakan teknik angket untuk mengecek akurasi dan ketelitian data secara pengamatan atau observasi visual. Data mentah yang diperoleh dari hasil pengumpulan data yang berupa angka-angka, tidak selalu siap untuk dianalisis. Selain itu harus melalui tahap-tahap tertentu. Pada proses pengumpulan data kemungkinan ada kesalahan dalam penulisan data atau angka-angka dan sebagainya.maka perlu dilakukan pembenahan (editing) data, dan pemberian kode agar lebih teliti. Terutama, data yang diperoleh dari hasil kuesioner, data-data tersebut perlu interpretasi lebih cermat.
2. Menetapkan data yang akan dianalisis. Ada beberapa jenis skala pengukuran data, yaitu nominal, ordinal, interval, dan rasio. Dengan menetapkan secara pasti dan cermat jenis skala pengukuran data tersebut maka akan dapat ditetapkan teknik analisis statistik yang paling sesuai dan tepat yang akan digunakan. Ketepatan menentukan teknik statistika yang akan digunakan tergantung dan berdasarkan pada tingkat atau skala pengukuran data yang akan dianalisis.
3. Mengetahui distribusi data variabel penelitian. Beberapa teknik analisis statistik menuntut variable yang akan dianalisis harus memenuhi persyaratan tentang distribusi data, yaitu harus berdistribusi normal. Bahkan beberapa teknik analisis statistik menuntut sifat-sifat varian skor-skor data variabel yang akan dianalisis.

Data yang ditemukan berbentuk kuantitatif dan kualitatif. Dalam data kualitatif/data kategori memiliki 3 jenis variable yaitu nominal, identifier, dan ordinal. Tetapi dalam data sampel yang dikumpulkan hanya terdapat 2 jenis yaitu nominal. Dalam data kuantitatif memiliki 2 jenis variable yaitu diskrit dan kontinyu. Dalam data sampel yang dikumpulkan hanya terdapat 1 jenis saja yaitu jenis kontinyu.

| Tipe data | Jenis | Skala variabel |
| --- | --- | --- |
| Kuantitatif | Kontinyu | * Usia * Lama Belajar * IPK |
| Kualitatif | Identifier | * Nama |
|  | Nominal | * Gender * Prodi * Kesesuaian * Pembiayaan * Kota Asal * Kos * kendaraan |

| Nama Variabel | Deskripsi | Keterangan |
| --- | --- | --- |
| Nama | Nama Panggilan | Character |
| Gender | Jenis kelamin  0. Perempuan  1. Laki-laki | Numeric [0,1] |
| Usia | Usia | Numeric |
| Prodi | Prodi  0. Teknik Informatika  1. ASP  2. Teknik Komputer  3. Teknik Mesin  4. Teknik Elektro  5. Akutansi  6. Farmasi  7. Kebidanan  8. Perhotelan  9. DKV | Numeric [0,1,2,3,4,5,6,7,8,9] |
| Kesesuaian | Kesesuaian Prodi  0. Ya  1. Tidak | Numeric [0,1] |
| Pembiayaan | Pembiayaan  0. Orangtua/Wali  1. Keduanya  2. Mandiri | Numeric [0,1,2] |
| Kota Asal | Kota Asal  0. Kota Tegal  1. Kab. Tegal  2. Kab. Brebes  3. Kab. Pemalang  4. Lainnya | Faktorial [0,1,2,3,4] |
| Kos | Kos  0. Ya  1. Tidak | Numeric [0,1] |
| Lama Belajar | Jam Belajar Harian | Numeric |
| Kendaraan | Kendaraan  0. Sepeda  1. Sepeda Motor  2. Mobil  3. Angkutan Umum | Numeric [0,1,2,3] |
| IPK | IPK | Numeric |

**BAB II**

**ANALISA**

1. TABEL KONTIGENSI

Menganalisis hubungan antar data kategori menggunakan table kontigensi

1. Temuan pertama 🡪 Hubungan antara jenis kelamin dan kos

Data studi dapat dilihat pada data.xlsx. Dalam kasus ini, kita akan menjawab pertanyaan: Apakah benar mahasiswa perempuan lebih yang ngekos?

Langkah awal, import terlebih dahulu datasetnya.

library(readxl)  
data <-read\_excel("C:/Users/Lenovo/Downloads/data.xlsx")  
View(data)

Variabel pertama gender merupakan variabel nominal yang mengidentikasi laki-laki atau perempuan. Variabel kedua kos merupakan variabel nominal yang mengidentikasi mahasiswa yang ngekos atau tidak.

Langkah kedua, buatlah variabel baru untuk menjelaskan variabel jenis kelamin dan kos. Jika jenis kelamin= 0 berarti laki-laki dan jika bernilai 1 maka perempuan. Jika kos = 0 berarti kos dan jika bernilai 1 berarti tidak kos. Gunakan fungsi dplyr.

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.1.2

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data =mutate(data, Jenis.kelamin =  
if\_else(data$`Jenis kelamin`==0 ,"pria","wanita"),  
kos =if\_else(data$Kos ==0,"kos","tidak kos"))

Langkah ketiga, membuat table frekuensi dan prosentase dari jenis kelamin dan kos

table(data$Jenis.kelamin)

##   
## pria wanita   
## 67 60

prop.table(table(data$Jenis.kelamin))\*100

##   
## pria wanita   
## 52.75591 47.24409

table(data$kos)

##   
## kos tidak kos   
## 30 97

prop.table(table(data$kos))\*100

##   
## kos tidak kos   
## 23.62205 76.37795

Terdapat sekitar 53% data mahasiswa PHB yang dikumpulkan berjenis kelamin laki-laki dan sekitar 47% mahasiswa berjenis perempuan. Dan terdapat sekitar 24% mahasiswa PHB yang kos dan sisanya sekitar 76% mahasiswa PHB yang tidak kos.

Melihat hubungan 2 variabel dengan menggunakan fungsi addmargins

addmargins(table(data$Jenis.kelamin,data$kos))

##   
## kos tidak kos Sum  
## pria 17 50 67  
## wanita 13 47 60  
## Sum 30 97 127

Dari 67 mahasiswa PHB yang berjenis kelamin laki-laki terdapat 50 yang memilih tidak kos dan dari 60 mahasiswa PHB berjenis kelamin perempuan terdapat 47 yang memilih tidak kos.

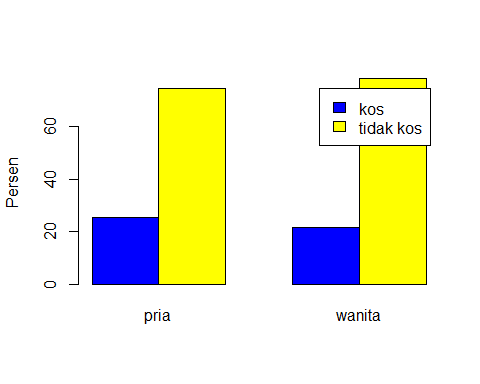
Untuk mengetahui persentase dari gambaran distribusi

prop.table(table(data$Jenis.kelamin,data$kos), margin =2)\*100

##   
## kos tidak kos  
## pria 56.66667 51.54639  
## wanita 43.33333 48.45361

Gunakan bar plot untuk mempermudah deskripsi tabel

barplot(prop.table(table(data$kos,  
 data$Jenis.kelamin),  
margin =2) \*100,beside =TRUE,  
legend.text =TRUE, ylab ="Persen", col =c("blue", "yellow")  
)



Berdasarkan graph yang terbentuk, terlihat bahwa pada data mahasiswa PHB, mahasiswa yang memilih untuk tidak kos sangat tinggi.

Kesimpulan, berdasarkan data diatas menunjukan jika anggapan bahwa mahasiswa yang ngekos itu banyak adalah perempuan adalah salah, karena berdasarkan graph diatas menunjukan bahwa mahasiswa pria lebih banyak yang ngekos daripada mahasiswa perempuan.

1. Temuan kedua 🡪 Hubungan antara prodi dan kendaraan

Data studi dapat dilihat pada data.xlsx. Dalam kasus ini, kita akan menjawab pertanyaan: prodi manakah yang mahasiswanya paling banyak menggunakan motor?

Langkah awal, import terlebih dahulu datasetnya.

library(readxl)  
data <-read\_excel("C:/Users/Lenovo/Downloads/data.xlsx")

Variabel pertama prodi merupakan variabel nominal yang mengidentifikasi prodi-prodi apa saja yang ada di Politeknik Harapan Bersama yaitu ada Teknik Informatika, ASP, Teknik Komputer, Teknik Mesin, Teknik Elektro, Akuntansi, Farmasi, Kebidanan, DKV, Perhotelan . Variabel kedua kendaraan merupakan variabel nominal yang mengidentifikasi kendaraan apa saja yang dipakai mahasiswa selama kegiatan di kampus yaitu ada Sepeda Motor,Mobil, dan Angkutan Umum.

Langkah kedua, buatlah variabel baru untuk menjelaskan variabel Prodi dan Kendaraan.

|  |  |
| --- | --- |
| prodi | 0 = Teknik Informatika  1 = ASP  2 = Teknik Komputer  3 = Teknik Mesin  4 = Teknik Elektro  5 = Akutansi  6 = Farmasi  7 = Kebidanan  8 = Perhotelan  9 = DKV |
| kendaraan | 1 = sepeda motor  2 = mobil  3 = angkutan umum |

Gunakan fungsi dplyr.

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.1.2

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data =mutate(data, jurusan =  
case\_when(  
 data$Prodi ==0~"informatika",  
 data$Prodi ==1~"ASP",  
 data$Prodi ==2~"T.komputer",  
 data$Prodi ==3~"T.mesin",  
 data$Prodi ==4~"T.elektro",  
 data$Prodi ==5~"Akuntansi",  
 data$Prodi ==6~"Farmasi",  
 data$Prodi ==7~"Kebidanan",  
 data$Prodi ==8~"PerhoteLAN",  
 data$Prodi ==9~"DKV"  
 ),  
kendaraan =case\_when(  
 data$Kendaraan ==1~"SEPEDA MOTOR",  
 data$Kendaraan ==2~"MOBIL",  
 data$Kendaraan ==3~"ANGKUTAN UMUM"  
 )  
)

Langkah ketiga, penulis membuat tabel frekuensi dan prosentase untuk melihat berapa banyak mahasiswa yang mengambil prodi. Gunakan variabel prodi.

table(data$jurusan)

##   
## Akuntansi ASP DKV Farmasi informatika Kebidanan   
## 5 21 4 5 33 5   
## PerhoteLAN T.elektro T.komputer T.mesin   
## 5 8 22 19

prop.table(table(data$jurusan))\*100

##   
## Akuntansi ASP DKV Farmasi informatika Kebidanan   
## 3.937008 16.535433 3.149606 3.937008 25.984252 3.937008   
## PerhoteLAN T.elektro T.komputer T.mesin   
## 3.937008 6.299213 17.322835 14.960630

table(data$kendaraan)

##   
## ANGKUTAN UMUM MOBIL SEPEDA MOTOR   
## 18 2 107

prop.table(table(data$kendaraan))\*100

##   
## ANGKUTAN UMUM MOBIL SEPEDA MOTOR   
## 14.173228 1.574803 84.251969

Diperoleh data terdapat yang menggunakan angkutan umum ada 18,mahasiswa yang menggunakan mobil terdapat 2 mahasiswa ,mahasiswa yang menggunakan sepeda motor terdapat 107 mahasiswa. Dengan persentase keseluruhan mahasiswa yang menggunakan angkutan umum yaitu 14%, mahasiswa yang menggunakan mobil yaitu 1%,mahasiswa yang menggunakan sepeda motor yaitu 84%.

Langkah keempat, buatlah tabel kontingensi menggunakan fungsi add margins untuk menunjukan distribusi **Prodi** dan **Kendaraan**. Tujuan dari tabel kontingensi ini untuk melihat hubungan Prodi dengan kendaraan.

addmargins(table(data$jurusan,data$kendaraan))

##   
## ANGKUTAN UMUM MOBIL SEPEDA MOTOR Sum  
## Akuntansi 3 0 2 5  
## ASP 3 1 17 21  
## DKV 0 0 4 4  
## Farmasi 1 0 4 5  
## informatika 2 1 30 33  
## Kebidanan 0 0 5 5  
## PerhoteLAN 0 0 5 5  
## T.elektro 1 0 7 8  
## T.komputer 5 0 17 22  
## T.mesin 3 0 16 19  
## Sum 18 2 107 127

Dari data diatas diperoleh diketahui bahwa yang bawa sepeda motor dalam prodi Teknik Informatika ada 33 orang, ASP ada 17 orang, Teknik Komputer ada 17 orang, Teknik Mesin ada 16 orang, Teknik Elektro ada 8 orang, Akuntansi ada 2 orang, Farmasi ada 4 orang, Kebidanan ada 4 orang, DKV ada 4 orang, Perhotelan ada 5 orang . Menghitung distribusi kondisional Prodi dan Kendaraan akan memberi gambaran yang lebih jelas tentang hubungan antara kedua variabel.

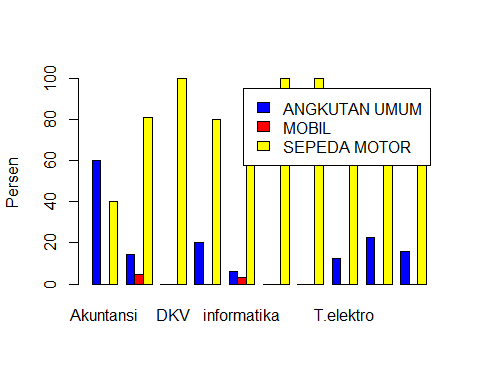
Selanjutnya, agar lebih mudah untuk memberikan gambaran distribusinya, rubah dalam bentuk prosentase. Margin menunjukan urutan prosentase. margin= 2 artinya prosentase berdasarkan urutan kedua (dalam hal ini kendaraan) sehingga diperoleh prosentase.

prop.table(table(data$jurusan,data$kendaraan), margin =2)\*100

##   
## ANGKUTAN UMUM MOBIL SEPEDA MOTOR  
## Akuntansi 16.666667 0.000000 1.869159  
## ASP 16.666667 50.000000 15.887850  
## DKV 0.000000 0.000000 3.738318  
## Farmasi 5.555556 0.000000 3.738318  
## informatika 11.111111 50.000000 28.037383  
## Kebidanan 0.000000 0.000000 4.672897  
## PerhoteLAN 0.000000 0.000000 4.672897  
## T.elektro 5.555556 0.000000 6.542056  
## T.komputer 27.777778 0.000000 15.887850  
## T.mesin 16.666667 0.000000 14.953271

Kemudian, agar lebih mudah dalam mendeskripsikan tabel, gunakan barplot

barplot(prop.table(table(data$kendaraan,  
 data$jurusan),  
margin =2) \*100,beside =TRUE,  
legend.text =TRUE, ylab ="Persen", col =c("blue", "red", "yellow")  
)



Kesimpulan, berdasarkan data diatas menunjukan bahwa prodi yang mahasiswanya paling banyak menggunakan motor yaitu prodi Teknik informatika dengan prosentase sekitar 28%.

1. Temuan ketiga 🡪 Hubungan antara jenis kelamin dan biaya

Data studi dapat dilihat pada data.xlsx. Dalam kasus ini, kita akan menjawab pertanyaan: Apakah benar mahasiswa laki-laki lebih banyak yang membiayai kuliahnya sendiri dari pada perempuan?

Langkah awal, import terlebih dahulu datasetnya.

library(readxl)  
data <-read\_excel("C:/Users/Lenovo/Downloads/data.xlsx")  
View(data)

Variabel pertama jenis kelamin merupakan variabel nominal yang mengidentikasi laki-laki atau perempuan. Variabel kedua pembiayaan merupakan variabel nominal yang mengidentikasi mahasiswa yang membiayai kuliahnya mandiri, orangtua atau keduanya.

Langkah kedua, buatlah variabel baru untuk menjelaskan variabel jenis kelamin dan biaya. Jika jenis kelamin= 0 berarti laki-laki dan jika bernilai 1 maka perempuan. Jika pembiayaan = 0 berarti orangtua dan jika bernilai 1 berarti keduanya dan jika bernilai 2 berarti mandiri, dan jika 3 berarti orangtua/wali.

Gunakann fungsi dplyr.

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.1.2

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data =mutate(data, jenis.kelamin =case\_when(  
 data$`Jenis kelamin`==0~"pria",  
 data$`Jenis kelamin`==1~"wanita"  
 ),  
biaya =case\_when(  
 data$Pembiayaan ==0~"orangtua/wali",  
 data$Pembiayaan ==1~"keduanya",  
 data$Pembiayaan ==2~"mandiri"  
 )  
)

Langkah ketiga, membuat table frekuensi dan prosentase dari jenis kelamin dan pembiayaan

table(data$jenis.kelamin)

##   
## pria wanita   
## 67 60

prop.table(table(data$jenis.kelamin))\*100

##   
## pria wanita   
## 52.75591 47.24409

table(data$biaya)

##   
## keduanya mandiri orangtua/wali   
## 38 19 70

prop.table(table(data$biaya))\*100

##   
## keduanya mandiri orangtua/wali   
## 29.92126 14.96063 55.11811

Terdapat sekitar 53% data mahasiswa PHB yang dikumpulkan berjenis kelamin laki-laki dan sekitar 47% mahasiswa berjenis perempuan. Dan terdapat sekitar 30% mahasiswa PHB yang membiayai studi menggunakan biaya keduanya, dan 14% mahasiswa PHB yang mengunakan pembiayaan mandiri dan terdapat 55% pembiayaan studi berasal dari orangtua/wali.

Langkah keeampat , melihat hubungan 2 variabel dengan menggunakan fungsi add margins

addmargins(table(data$jenis.kelamin,data$biaya))

##   
## keduanya mandiri orangtua/wali Sum  
## pria 26 12 29 67  
## wanita 12 7 41 60  
## Sum 38 19 70 127

Hasilnya dari 67 mahasiswa PHB yang berjenis kelamin laki-laki terdapat 26 orang yang pembiayaan studi berasal dari keduanya, 12 orang mendiri, dan 29 orang biayanya berasal dari orangtua/wali. Dari 60 mahasiswa PHB yang berjenis kelamin perempuan terdapat 12 orang yang biayanya berasal dari keduanya, 19 orang mandiri, dan 41 orang biayanya berasal dari orangtua/ wali.

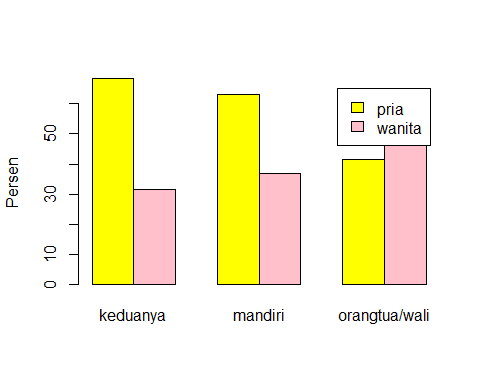
Langkah kelima, Untuk mengetahui prosentase dari gambaran distribusi, gunakan barplot untuk mempermudah deskripsitabel

prop.table(table(data$jenis.kelamin,data$biaya), margin =2)\*100

##   
## keduanya mandiri orangtua/wali  
## pria 68.42105 63.15789 41.42857  
## wanita 31.57895 36.84211 58.57143

Mendeskripsikan table dalam bentuk graph berwarna

barplot(prop.table(table(data$jenis.kelamin,  
 data$biaya),  
margin =2) \*100,beside =TRUE,  
legend.text =TRUE, ylab ="Persen", col =c( "yellow", "pink")  
)



Kesimpulan, berdasarkan data diatas menunjukan jika anggapan bahwa mahasiswa laki-laki lebih banyak yang membiayai kuliahnya sendiri secara mandiri adalah benar.

1. REGRESI LINEAR

Menentukan predictor mana yang paling potensial untuk menentukan IPK.

1. Temuan pertama 🡪 Hubungan antara IPK dan Jam belajar

Impor data terlebih dahulu

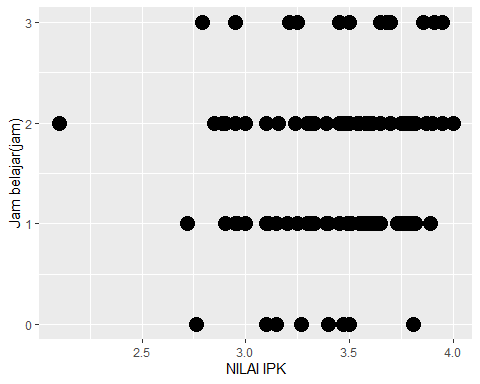
library(readxl)  
data<-read\_excel("C:/Users/Lenovo/Downloads/data.xlsx")  
data$IPK <-as.numeric(data$IPK)

Buatlah scatterplot untuk prediktor manakah yang paling potensial untuk menentukan IPK menggunakan fungsi plot. Karena kita ingin memprediksi apakah lama belajar akan berpengaruh terhadap IPK, maka variabel Nilai IPK sebagai sumbu x dan Jam belajar sebagai sumbu y.

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.1.2

ggplot(data, aes(x=IPK, y=`Jam belajar`)) +geom\_point(size=5)+  
labs(x="NILAI IPK", y="Jam belajar(jam)")



Untuk memberikan keterangan pada legenda, buatlah variabel baru yang menunjukan kategori waktu belajar itu cepat,sedang atau lama. Jika cepat bernilai 1, sedang bernilai 3 dan lama bernilai 3 berdasarkan variabel Jam belajar.

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.1.2

##   
## Attaching package: 'dplyr'

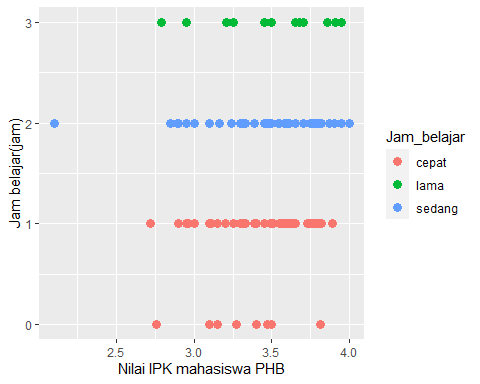
## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data =mutate(data, Jam\_belajar=case\_when(  
 data$`Jam belajar`<=1~"cepat ",  
 data$`Jam belajar`<=2~"sedang ",  
 data$`Jam belajar`>=2~"lama "  
),  
nilai =case\_when(  
 data$IPK <=3~"grade C ",  
 data$IPK <=3.5~"grade B ",  
 data$IPK >=3.5~"grade A "))

Buat scatterplot dengan legenda dan berilah warna berdasarkan waktu belajar.

ggplot(data, aes(x=IPK, y=`Jam belajar`, color = Jam\_belajar)) +  
geom\_point(size=3)+  
labs(x="Nilai IPK mahasiswa PHB", y="Jam belajar(jam)")



Dengan melihat scatterplot, terlihat bahwa jam belajar mahasiswa tidak terlalu perngaruh terhadap nilai IPK.

Buat dataset baru

databaru =filter(data, `Jam belajar`<3)  
databaru

## # A tibble: 112 x 13  
## Nama `Jenis kelamin` Usia Prodi `Kesesuaian prodi` Pembiayaan `Kota asal`  
## <chr><dbl><dbl><dbl><dbl><dbl><dbl>  
## 1 Fabil 0 20 2 1 0 0  
## 2 Amel 1 19 2 0 1 0  
## 3 Widia 1 21 3 1 0 1  
## 4 Sinta 1 20 7 0 0 1  
## 5 Ami 1 20 2 1 0 0  
## 6 Taufik 0 19 1 0 0 1  
## 7 Ziky 0 22 1 1 0 0  
## 8 Silvia 1 20 1 1 2 1  
## 9 Gilang 0 20 4 1 2 1  
## 10 Putri 0 22 3 0 0 0  
## # ... with 102 more rows, and 6 more variables: Kos <dbl>, Jam belajar <dbl>,  
## # Kendaraan <dbl>, IPK <dbl>, Jam\_belajar <chr>, nilai <chr>

Kedua variabel (jam belajar dan IPK ) bersifat kuantitaif, scatterplot tidak menunjukan bentuk non-linear yang kuat, dan outlier sudah dihapus. Dengan demikian, syarat untuk melakukan analisa regresi linear terpenuhi.

Pertama, carilah nilai korelasi antara Jam belajar dan IPK menggunakan fungsi cor

cor(databaru$IPK, databaru$`Jam belajar`)

## [1] 0.1212045

Korelasi antara Jam belajar dan IPK adalah sekitar 0.121, yang menunjukkan hubungan linier positif yang cukup.

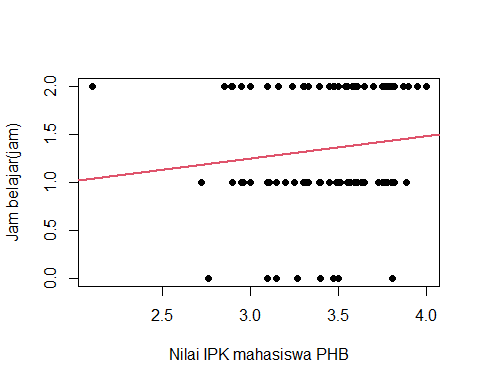
Kedua, carilah persamaan regresi yang menghubungkan antara Jam belajar dan IPK. Interpretasikan kemiringan dan intersep garis menggunakan fungsi lm

lm(`Jam belajar`~IPK, databaru)

##   
## Call:  
## lm(formula = `Jam belajar` ~ IPK, data = databaru)  
##   
## Coefficients:  
## (Intercept) IPK   
## 0.5493 0.2339

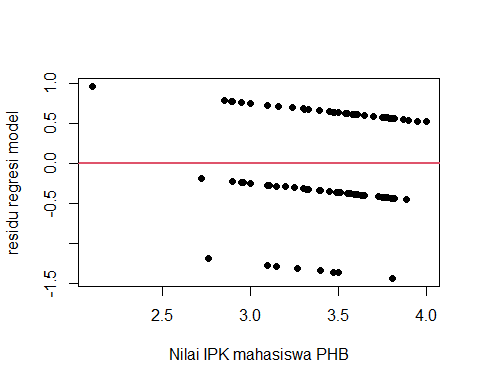
Ketiga, buatlah scatterplot untuk menunjukkan hubungan komposisi antara Jam belajar dan IPK, dan beri keterangan dengan garis regresi dan korelasi.

plot(databaru$IPK, databaru$`Jam belajar`, pch=16,  
xlab ="Nilai IPK mahasiswa PHB",  
ylab ="Jam belajar(jam)")  
abline(lm(databaru$`Jam belajar`~ databaru$IPK), col=2, lwd=2)  
text(3.4, 40, paste("y = 0.5493+0.2339x"))



Keempat, buatlah plot residu untuk menunjukan jarak titik data dengan garis linear (model linear).

model =lm(`Jam belajar`~ IPK, databaru)  
residu =resid(model)  
plot(databaru$IPK, residu, pch=16,  
xlab ="Nilai IPK mahasiswa PHB",  
ylab ="residu regresi model")  
abline(h=0, col=2, lwd=2)



Fungsi Berdasarkan plot residu, terlihat tidak ada oulier data. Kita dapat melihat nilai residu dengan menggunakan fungsi summary

summary(model)

##   
## Call:  
## lm(formula = `Jam belajar` ~ IPK, data = databaru)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.4406 -0.3921 -0.2816 0.6062 0.9594   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.5493 0.6336 0.867 0.388  
## IPK 0.2339 0.1827 1.281 0.203  
##   
## Residual standard error: 0.6258 on 110 degrees of freedom  
## Multiple R-squared: 0.01469, Adjusted R-squared: 0.005733   
## F-statistic: 1.64 on 1 and 110 DF, p-value: 0.203

Dengan model regresi linear yang terbentuk, kita dapat melakukan prediksi. Misal, berapakah lama belajar jika IPK 3.85?

nilaibaru =data.frame(IPK=3.85)  
predict(model, nilaibaru)

## 1   
## 1.449991

Berdasarkan model, lama belajar mahasiswa jika IPK 3.85 adalah sekitar 1,4 jam .

Kelima, carilah nilai determinasi untuk mengetahui apakah lama belajar dapat dijadikan prediktor yang baik untuk menentukan nilai IPK.

Dengan menggunakan fungsi summary diperoleh nilai determinasinya 0,01469. Hal ini menunjukan bahwa 1.4% variabilitas pengaruh Jam belajar terhadap IPK .

Setelah dibuat scatterplot, residual plot dan menghilangkan outlier, sangat memungkinkan untuk memodelkan antara hubungan komposisi antara Jam belajar dan IPK. Namun, karena nilai varibilitas jam belajar hanya 1.4% yang termasuk rendah, maka variable lama belajar tidak dapat dijadikan predictor yang baik untuk menentukan nilai IPK.

1. Temuan kedua🡪 Hubungan antara IPK dan usia

Langkah pertama, mengimport data

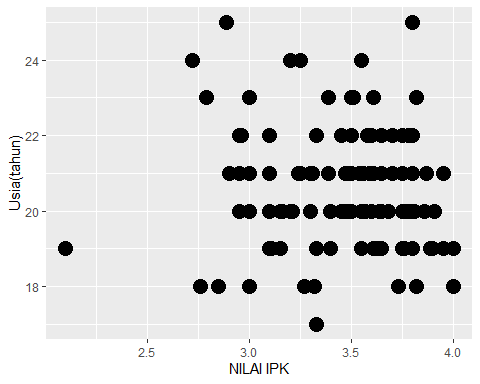
library(readxl)  
data <- read\_excel("C:/Users/Lenovo/Downloads/data.xlsx")  
data$IPK <- as.numeric(data$IPK)

Buatlah scatterplot untuk prediktor manakah yang paling potensial untuk menentukan IPK menggunakan fungsi plot. Karena kita ingin memprediksi apakah usia akan berpengaruh terhadap IPK, maka variabel Nilai IPK sebagai sumbu x dan usia sebagai sumbu y.

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.1.2

ggplot(data, aes(x=IPK, y=Usia)) + geom\_point(size=5)+  
 labs(x="NILAI IPK", y="Usia(tahun)")



Untuk memberikan keterangan pada legenda, buatlah variabel baru yang menunjukan kategori usia itu muda, pertengah, tua

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.1.2

##   
## Attaching package: 'dplyr'

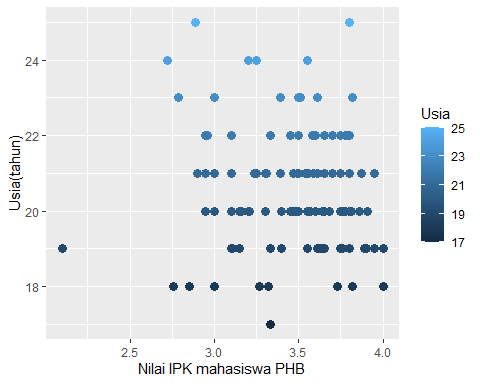
## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data = mutate(data, usia = case\_when(  
 data$`Usia` <= 20 ~ "usia muda",  
 data$`Usia` <= 25 ~ "usia pertengahan",  
 data$`Usia` >= 25 ~ "usia tua"  
),  
nilai = case\_when(  
 data$IPK <= 3 ~ "grade C ",  
 data$IPK <= 3.5 ~ "grade B",  
 data$IPK >= 3.5 ~ "grade A"  
)  
)

Buat scatterplot dengan legenda dan berilah warna berdasarkan usia.

ggplot(data, aes(x=IPK, y=`Usia`, color = Usia)) +  
 geom\_point(size=3)+  
 labs(x="Nilai IPK mahasiswa PHB", y="Usia(tahun)")



Dengan melihat scatterplot, terlihat bahwa usia mahasiswa tidak terlalu perngaruh terhadap nilai IPK.

Membuat data baru

databaru = filter(data, `Usia` < 25)  
databaru

## # A tibble: 125 x 13  
## Nama `Jenis kelamin` Usia Prodi `Kesesuaian prodi` Pembiayaan `Kota asal`  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Fabil 0 20 2 1 0 0  
## 2 Amel 1 19 2 0 1 0  
## 3 Widia 1 21 3 1 0 1  
## 4 Sinta 1 20 7 0 0 1  
## 5 Ami 1 20 2 1 0 0  
## 6 Taufik 0 19 1 0 0 1  
## 7 Ziky 0 22 1 1 0 0  
## 8 Silvia 1 20 1 1 2 1  
## 9 Gilang 0 20 4 1 2 1  
## 10 Putri 0 22 3 0 0 0  
## # ... with 115 more rows, and 6 more variables: Kos <dbl>, Jam belajar <dbl>,  
## # Kendaraan <dbl>, IPK <dbl>, usia <chr>, nilai <chr>

Kedua variabel (Usia dan Nilai IPK) bersifat kuantitaif, scatterplot tidak menunjukan bentuk non-linear yang kuat, dan outlier sudah dihapus. Dengan demikian, syarat untuk melakukan analisa regresi linear terpenuhi.

Pertama, Carilah nilai korelasi antara hubungan usia dengan nilai IPK mahasiswa PHB menggunakan fungsi cor

cor(databaru$IPK, databaru$`Usia`)

## [1] -0.07827798

Korelasi antara hubungan usia dengan nilai IPK mahasiswa PHB adalah sekitar -0.08, yang menunjukkan hubungan linier negatif yang cukup kuat.

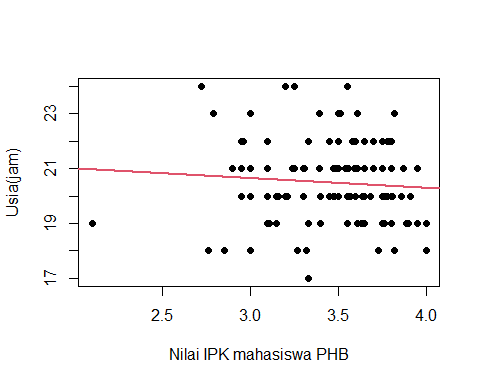
Kedua, carilah persamaan regresi yang menghubungkan usia dan nilai IPK mahasiswa PHB dihasilkan. Interpretasikan kemiringan dan intersep garis menggunakan fungsi lm

lm(`Usia`~IPK, databaru)

##   
## Call:  
## lm(formula = Usia ~ IPK, data = databaru)  
##   
## Coefficients:  
## (Intercept) IPK   
## 21.7257 -0.3529

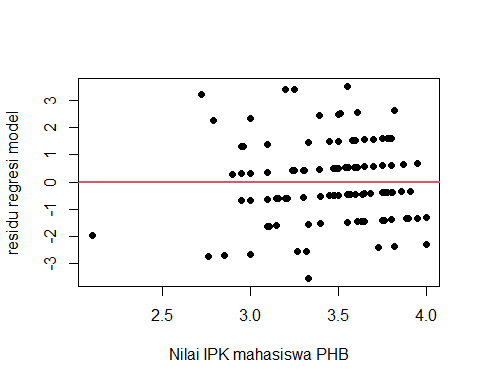
Ketiga, buatlah scatterplot untuk menunjukkan hubungan usia dan nilai IPK mahasiswa PHB dan beri keterangan dengan garis regresi dan korelasi.

plot(databaru$IPK, databaru$`Usia`, pch=16,  
 xlab = "Nilai IPK mahasiswa PHB",  
 ylab = "Usia(jam)")  
abline(lm(databaru$`Usia` ~ databaru$IPK), col=2, lwd=2)  
text(3.4, 40, paste("y = 21.8135-0.3657x"))



Keempat, buatlah plot residu untuk menunjukan jarak titik data dengan garis linear (model linear).

model = lm(`Usia` ~ IPK, databaru)  
residu = resid(model)  
plot(databaru$IPK, residu, pch=16,  
 xlab = "Nilai IPK mahasiswa PHB",  
 ylab = "residu regresi model")  
abline(h=0, col=2, lwd=2)



Berdasarkan plot residu, terlihat tidak ada oulier data. Kita dapat melihat nilai residu dengan menggunakan fungsi summary.

summary(model)

##   
## Call:  
## lm(formula = Usia ~ IPK, data = databaru)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.5506 -0.6847 -0.3848 0.6399 3.5270   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 21.7257 1.4090 15.419 <2e-16 \*\*\*  
## IPK -0.3529 0.4052 -0.871 0.386   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.458 on 123 degrees of freedom  
## Multiple R-squared: 0.006127, Adjusted R-squared: -0.001953   
## F-statistic: 0.7583 on 1 and 123 DF, p-value: 0.3856

Dengan model regresi linear yang terbentuk, kita dapat melakukan prediksi. Misal, berapakah usia yang didapat pada nilai IPK 3.85 ?

nilaibaru = data.frame(IPK=3.85)  
predict(model, nilaibaru)

## 1   
## 20.36711

Berdasarkan model, rata-rata nilai IPK 3.85 yang diperoleh mahasiswa PHB adalah sekitar 20 tahun.

Kelima, carilah nilai determinasi untuk mengetahui apakah usia dapat dijadikan prediktor yang baik untuk menentukan nilai IPK.

Dengan menggunakan fungsi summary diperoleh nilai determinasinya 0,006127. Hal ini menunjukan bahwa 0.6% variabilitas pengaruh usia terhadap IPK.

Setelah dibuat scatterplot, residual plot dan menghilangkan outlier, sangat memungkinkan untuk memodelkan hubungan komposisi antara usia dan nilai IPK mahasiswa PHB dengan model regresi linier. Namun, karena nilai varibiitas usia hanya 0.6% yang termasuk rendah, maka variable usia tidak dapat dijadikan predictor yang baik untuk menentukan nilai IPK.

1. REGRESI LINEAR SEDERHANA

Langkah awal, import data terlebih dahulu

library(readxl)  
data <-read\_excel("C:/Users/Lenovo/Downloads/data.xlsx")  
data$IPK <-as.numeric(data$IPK)

1. Statistika deskriptif

Statistika deskriptif yang akan digunakan adalah nilai minimum, quartil 1, median, rata-rata, quartil 3, dan nilai maksimum. Berikut hasil analisa statistika deskriptif pada data nilai IPK dan usia:

summary(data)

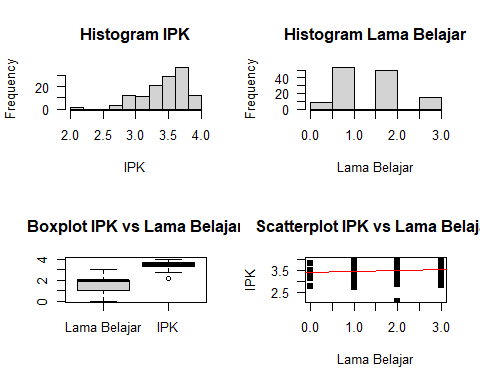
## Nama Jenis kelamin Usia Prodi   
## Length:127 Min. :0.0000 Min. :17.00 Min. :0.00   
## Class :character 1st Qu.:0.0000 1st Qu.:20.00 1st Qu.:0.00   
## Mode :character Median :0.0000 Median :20.00 Median :2.00   
## Mean :0.4724 Mean :20.72 Mean :2.52   
## 3rd Qu.:1.0000 3rd Qu.:21.50 3rd Qu.:3.50   
## Max. :1.0000 Max. :43.00 Max. :9.00   
## Kesesuaian prodi Pembiayaan Kota asal Kos   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:1.0000   
## Median :1.0000 Median :0.0000 Median :0.0000 Median :1.0000   
## Mean :0.5669 Mean :0.5984 Mean :0.7953 Mean :0.7638   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :2.0000 Max. :4.0000 Max. :1.0000   
## Jam belajar Kendaraan IPK   
## Min. :0.000 Min. :1.000 Min. :2.100   
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:3.250   
## Median :2.000 Median :1.000 Median :3.500   
## Mean :1.551 Mean :1.299 Mean :3.460   
## 3rd Qu.:2.000 3rd Qu.:1.000 3rd Qu.:3.715   
## Max. :3.000 Max. :3.000 Max. :4.000

Dari statistika deskriptif diatas dapat diketahui bahwa rata-rata IPK mahasiswa Statistika PHB sebesar 3,46 dengan nilai IPK terbesar adalah 4,0 dan nilai IPK terkecil adalah 2,10. Median nilai IPK mahasiswa Statistika PHB sebesar 3,50. Rata-rata lama belajar mahasiswa Statistika PHB sebesar 1,22 jam, dengan lama belajar terbesar adalah 4 jam dan lama belajar terkecil adalah 0 jam. Median lama belajar mahasiswa statistika PHB adalah 1 jam.

1. Visualisasi data

Visualisasi data yang akan digunakan adalah histogram, scatterplot, dan boxplot. Berikut hasil visualisasi data pada data nilai IPK dan lama belajar: summary(data)

par(mfrow=c(2,2))  
hist(data$`IPK`,main="Histogram IPK",xlab="IPK", las=1)  
hist(data$`Jam belajar`, main ="Histogram Lama Belajar",  
xlab="Lama Belajar", las=1)  
boxplot(data$`Jam belajar`, data$`IPK`,  
main="Boxplot IPK vs Lama Belajar",  
names=c("Lama Belajar","IPK"))  
plot(data$`Jam belajar`, data$`IPK`, main="Scatterplot IPK vs Lama Belajar",  
xlab="Lama Belajar",  
ylab="IPK", pch=15,las=1)  
abline(lm(data$`IPK`~data$`Jam belajar`),col="red")



Dari visualisasi diatas dapat diketahui bahwa histogram IPK memiliki bentuk tangga yang mengindikasikan bahwa secara visual nilai IPK mengikuti disribusi normal, sedangkan histogram lama belajar memiliki bentuk yang skewness ke kanan yang artinya ada nilai yang besar namun kebanyakan nilainya kecil. Hasil scatterplot menunjukkan bahwa terdapat sedikit korelasi positif yang kuat antara variabel Lama Belajar dengan variabel IPK. Jika waktu Lama Belajar meningkat maka nilai IPK juga meningkat, begitu juga sebaliknya jika waktu Lama Belajar menurun maka nilai IPK juga akan menurun. Hasil boxplot dapat diketahui bawah terdapat data outlier pada variabel IPK dan terdapat data outlier juga pada variabel lama belajar.

1. Analisis korelasi

waktu\_belajar = data$`Jam belajar`  
IPK=data$`IPK`  
cor.test(waktu\_belajar, IPK)

##   
## Pearson's product-moment correlation  
##   
## data: waktu\_belajar and IPK  
## t = 1.4004, df = 125, p-value = 0.1639  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.05103845 0.29216994  
## sample estimates:  
## cor   
## 0.1242813

Nilai korelasi sebesar 0.15230571 menunjukkan bahwa terdapat sedikit korelasi positif antara variabel lama belajar dengan variabel IPK. Jika waktu lama belajar meningkat maka nilai IPK juga meningkat, begitu juga sebaliknya jika waktu lama belajar menurun maka nilai IPK juga akan menurun.

1. Analisis Regresi

Model\_Regresi =summary(lm(IPK~waktu\_belajar))  
Model\_Regresi

##   
## Call:  
## lm(formula = IPK ~ waktu\_belajar)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.3831 -0.1826 0.0679 0.2424 0.5169   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.38106 0.06346 53.28 <2e-16 \*\*\*  
## waktu\_belajar 0.05104 0.03645 1.40 0.164   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3248 on 125 degrees of freedom  
## Multiple R-squared: 0.01545, Adjusted R-squared: 0.007569   
## F-statistic: 1.961 on 1 and 125 DF, p-value: 0.1639

Nilai varibilitas yang didapat diatas yakni 0.01545 atau sekitar 1.5%. Jadi dapat disimpulkan bahwa variabel lama belajar tidak terlalu berpengaruh terhadap nilai IPK

**BAB III**

**KODE**

1. TABEL KONTIGENSI
2. Hubungan variable jenis kelaminn dan kos

* Fungsi untuk mengimport data (excel)

library(readxl)  
data <- read\_excel("C:/Users/Lenovo/Downloads/data.xlsx")  
View(data)

* Fungsi untuk mendeskripsikan variable baru antar data kategori

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.1.2

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data = mutate(data, Jenis.kelamin =  
 if\_else(data$`Jenis kelamin` == 0 ,"pria","wanita"),  
 kos = if\_else(data$Kos == 0,"kos","tidak kos"))

* Fungsi untuk mengetahui frekuensi dan prosentase antar data kategori

table(data$Jenis.kelamin)

##   
## pria wanita   
## 67 60

prop.table(table(data$Jenis.kelamin))\*100

##   
## pria wanita   
## 52.75591 47.24409

table(data$kos)

##   
## kos tidak kos   
## 30 97

prop.table(table(data$kos))\*100

##   
## kos tidak kos   
## 23.62205 76.37795

* Fungsi untuk mengetahui distribusi antar data kategori

addmargins(table(data$Jenis.kelamin,data$kos))

##   
## kos tidak kos Sum  
## pria 17 50 67  
## wanita 13 47 60  
## Sum 30 97 127

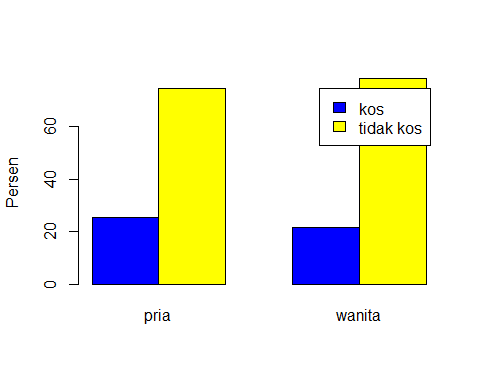
* Fungsi untuk mengetahui distribusi bentuk prosentase

prop.table(table(data$Jenis.kelamin,data$kos), margin = 2)\*100

##   
## kos tidak kos  
## pria 56.66667 51.54639  
## wanita 43.33333 48.45361

* Fungsi untuk mendeskripsikan table dalam bentuk graph berwarna

barplot(prop.table(table(data$kos,  
 data$Jenis.kelamin),  
 margin = 2) \*100,beside = TRUE,  
 legend.text = TRUE, ylab = "Persen", col = c("blue", "yellow")  
)



1. Hubungan variabel jenis kelamin dan biaya

* Fungsi untuk mengimport data (excel)

library(readxl)  
data <- read\_excel("C:/Users/Lenovo/Downloads/data.xlsx")  
View(data)

* Fungsi untuk mendeskripsikan variable baru antar data kategori

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.1.2

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data = mutate(data, jenis.kelamin = case\_when(  
 data$`Jenis kelamin` == 0 ~ "pria",  
 data$`Jenis kelamin` == 1 ~"wanita"  
 ),  
 biaya = case\_when(  
 data$Pembiayaan == 0 ~ "orangtua/wali",  
 data$Pembiayaan == 1 ~ "keduanya",  
 data$Pembiayaan == 2 ~ "mandiri"  
 )  
)

* Fungsi untuk mengetahui frekuensi dan prosentase antar data kategori

table(data$jenis.kelamin)

##   
## pria wanita   
## 67 60

prop.table(table(data$jenis.kelamin))\*100

##   
## pria wanita   
## 52.75591 47.24409

table(data$biaya)

##   
## keduanya mandiri orangtua/wali   
## 38 19 70

prop.table(table(data$biaya))\*100

##   
## keduanya mandiri orangtua/wali   
## 29.92126 14.96063 55.11811

* Fungsi untuk mengetahui distribusi antar data kategori

addmargins(table(data$jenis.kelamin,data$biaya))

##   
## keduanya mandiri orangtua/wali Sum  
## pria 26 12 29 67  
## wanita 12 7 41 60  
## Sum 38 19 70 127

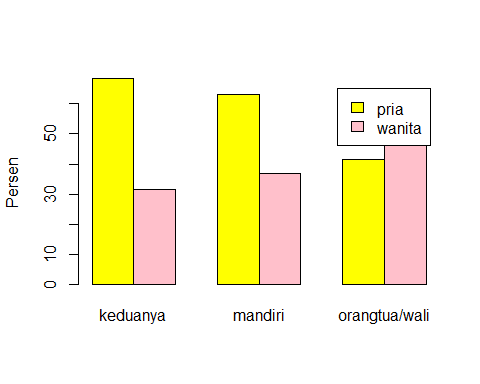
* Fungsi untuk mengetahui distribusi bentuk prosentase

prop.table(table(data$jenis.kelamin,data$biaya), margin = 2)\*100

##   
## keduanya mandiri orangtua/wali  
## pria 68.42105 63.15789 41.42857  
## wanita 31.57895 36.84211 58.57143

* Fungsi untuk mendeskripsikan table dalam bentuk graph berwarna

barplot(prop.table(table(data$jenis.kelamin,  
 data$biaya),  
 margin = 2) \*100,beside = TRUE,  
 legend.text = TRUE, ylab = "Persen", col = c( "yellow", "pink")  
)



1. Hubungan prodi dengan kendaraan

* Fungsi untuk mengimport data (excel)

library(readxl)  
data <- read\_excel("C:/Users/Lenovo/Downloads/data.xlsx")

* Fungsi untuk mendeskripsikan variable baru antar data kategori

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.1.2

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data = mutate(data, jurusan =  
 case\_when(  
 data$Prodi == 0 ~ "informatika",  
 data$Prodi == 1 ~ "ASP",  
 data$Prodi == 2 ~ "T.komputer",  
 data$Prodi == 3 ~ "T.mesin",  
 data$Prodi == 4 ~ "T.elektro",  
 data$Prodi == 5 ~ "Akuntansi",  
 data$Prodi == 6 ~ "Farmasi",  
 data$Prodi == 7 ~ "Kebidanan",  
 data$Prodi == 8 ~ "PerhoteLAN",  
 data$Prodi == 9 ~ "DKV"  
 ),  
 kendaraan = case\_when(  
 data$Kendaraan == 1 ~ "SEPEDA MOTOR",  
 data$Kendaraan == 2 ~ "MOBIL",  
 data$Kendaraan == 3 ~ "ANGKUTAN UMUM"  
 )  
)

* Fungsi untuk mengetahui frekuensi dan prosentase antar data kategori

table(data$jurusan)

##   
## Akuntansi ASP DKV Farmasi informatika Kebidanan   
## 5 21 4 5 33 5   
## PerhoteLAN T.elektro T.komputer T.mesin   
## 5 8 22 19

prop.table(table(data$jurusan))\*100

##   
## Akuntansi ASP DKV Farmasi informatika Kebidanan   
## 3.937008 16.535433 3.149606 3.937008 25.984252 3.937008   
## PerhoteLAN T.elektro T.komputer T.mesin   
## 3.937008 6.299213 17.322835 14.960630

table(data$kendaraan)

##   
## ANGKUTAN UMUM MOBIL SEPEDA MOTOR   
## 18 2 107

prop.table(table(data$kendaraan))\*100

##   
## ANGKUTAN UMUM MOBIL SEPEDA MOTOR   
## 14.173228 1.574803 84.251969

* Fungsi untuk mengetahui distribusi antar data kategori

addmargins(table(data$jurusan,data$kendaraan))

##   
## ANGKUTAN UMUM MOBIL SEPEDA MOTOR Sum  
## Akuntansi 3 0 2 5  
## ASP 3 1 17 21  
## DKV 0 0 4 4  
## Farmasi 1 0 4 5  
## informatika 2 1 30 33  
## Kebidanan 0 0 5 5  
## PerhoteLAN 0 0 5 5  
## T.elektro 1 0 7 8  
## T.komputer 5 0 17 22  
## T.mesin 3 0 16 19  
## Sum 18 2 107 127

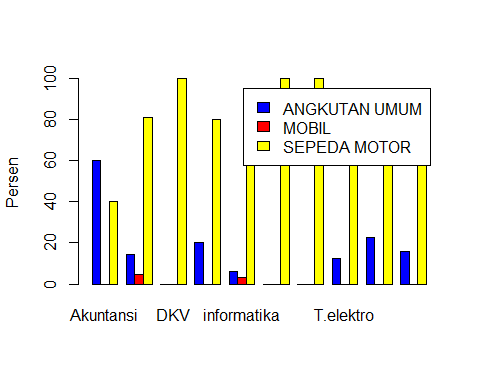
* Fungsi untuk mengetahui distribusi bentuk prosentase

prop.table(table(data$jurusan,data$kendaraan), margin = 2)\*100

##   
## ANGKUTAN UMUM MOBIL SEPEDA MOTOR  
## Akuntansi 16.666667 0.000000 1.869159  
## ASP 16.666667 50.000000 15.887850  
## DKV 0.000000 0.000000 3.738318  
## Farmasi 5.555556 0.000000 3.738318  
## informatika 11.111111 50.000000 28.037383  
## Kebidanan 0.000000 0.000000 4.672897  
## PerhoteLAN 0.000000 0.000000 4.672897  
## T.elektro 5.555556 0.000000 6.542056  
## T.komputer 27.777778 0.000000 15.887850  
## T.mesin 16.666667 0.000000 14.953271

* Fungsi untuk mendeskripsikan table dalam bentuk graph berwarna

barplot(prop.table(table(data$kendaraan,  
 data$jurusan),  
 margin = 2) \*100,beside = TRUE,  
 legend.text = TRUE, ylab = "Persen", col = c("blue", "red", "yellow")  
)



1. REGRESI LINEAR
2. Hubungan lama belajar dengan IPK

* Fungsi untuk import data

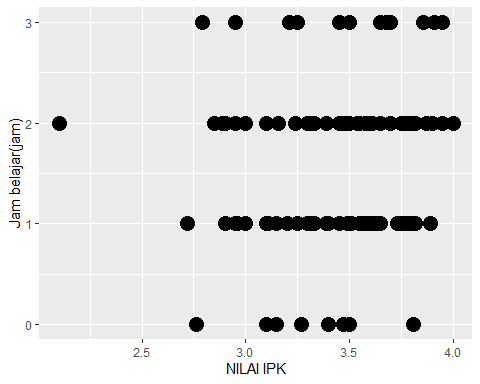
library(readxl)  
data <- read\_excel("C:/Users/Lenovo/Downloads/data.xlsx")  
data$IPK <- as.numeric(data$IPK)

* Fungsi untuk membuat scatterplot hubungan IPK dengan jam belajar

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.1.2

ggplot(data, aes(x=IPK, y=`Jam belajar`)) + geom\_point(size=5)+  
 labs(x="NILAI IPK", y="Jam belajar(jam)")



* Fungsi untuk mendeskripsikan variable baru antar data kategori

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.1.2

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data = mutate(data, Jam\_belajar= case\_when(  
 data$`Jam belajar` <= 1 ~ "cepat ",  
 data$`Jam belajar` <= 2 ~ "sedang ",  
 data$`Jam belajar` >= 2 ~ "lama "  
),  
nilai = case\_when(  
 data$IPK <= 3 ~ "grade C ",  
 data$IPK <= 3.5 ~ "grade B ",  
 data$IPK >= 3.5 ~ "grade A "))

* Fungsi untuk melihat frekuensi data

table(data$Jam\_belajar)

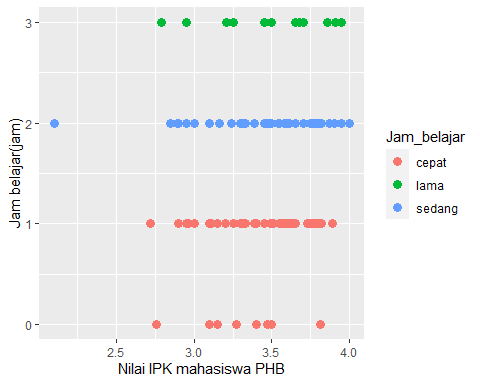
##   
## cepat lama sedang   
## 63 15 49

table(data$nilai)

##   
## grade A grade B grade C   
## 62 49 16

* Fungsi untuk membuat grafik plot(scatterplot)

ggplot(data, aes(x=IPK, y=`Jam belajar`, color = Jam\_belajar)) +  
 geom\_point(size=3)+  
 labs(x="Nilai IPK mahasiswa PHB", y="Jam belajar(jam)")



* Fungsi untuk membuat data yang baru dan menghapus data apabila data outlier

databaru = filter(data, `Jam belajar` < 3)  
databaru

## # A tibble: 112 x 13  
## Nama `Jenis kelamin` Usia Prodi `Kesesuaian prodi` Pembiayaan `Kota asal`  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Fabil 0 20 2 1 0 0  
## 2 Amel 1 19 2 0 1 0  
## 3 Widia 1 21 3 1 0 1  
## 4 Sinta 1 20 7 0 0 1  
## 5 Ami 1 20 2 1 0 0  
## 6 Taufik 0 19 1 0 0 1  
## 7 Ziky 0 22 1 1 0 0  
## 8 Silvia 1 20 1 1 2 1  
## 9 Gilang 0 20 4 1 2 1  
## 10 Putri 0 22 3 0 0 0  
## # ... with 102 more rows, and 6 more variables: Kos <dbl>, Jam belajar <dbl>,  
## # Kendaraan <dbl>, IPK <dbl>, Jam\_belajar <chr>, nilai <chr>

* Fungsi untuk mengetahui nilai korelasi antar variable jam belajar dan IPK

cor(databaru$IPK, databaru$`Jam belajar`)

## [1] 0.1212045

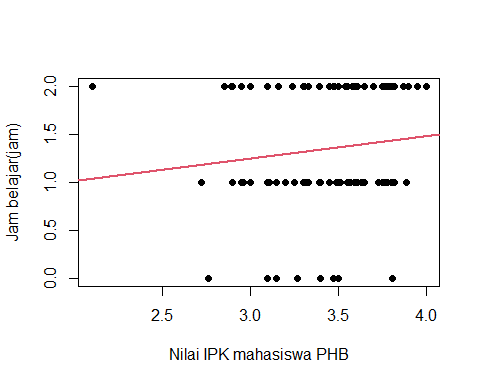
* Fungsi untuk mencari persamaan regresi

lm(`Jam belajar`~IPK, databaru)

##   
## Call:  
## lm(formula = `Jam belajar` ~ IPK, data = databaru)  
##   
## Coefficients:  
## (Intercept) IPK   
## 0.5493 0.2339

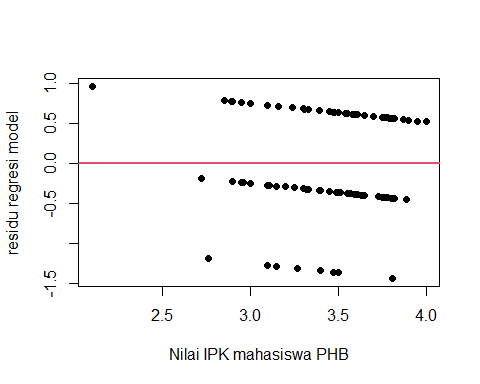
* Fungsi untuk membuat scatterplot dengan garis regresi dan korelasi

plot(databaru$IPK, databaru$`Jam belajar`, pch=16,  
 xlab = "Nilai IPK mahasiswa PHB",  
 ylab = "Jam belajar(jam)")  
abline(lm(databaru$`Jam belajar` ~ databaru$IPK), col=2, lwd=2)  
text(3.4, 40, paste("y = 0.5493+0.2339x"))



* Fungsi untuk membuat plot residu(tidak ada outlier) dengan garis linear

model = lm(`Jam belajar` ~ IPK, databaru)  
residu = resid(model)  
plot(databaru$IPK, residu, pch=16,  
 xlab = "Nilai IPK mahasiswa PHB",  
 ylab = "residu regresi model")  
abline(h=0, col=2, lwd=2)



* Fungsi untuk melihat nilai plot residu

summary(model)

##   
## Call:  
## lm(formula = `Jam belajar` ~ IPK, data = databaru)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.4406 -0.3921 -0.2816 0.6062 0.9594   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.5493 0.6336 0.867 0.388  
## IPK 0.2339 0.1827 1.281 0.203  
##   
## Residual standard error: 0.6258 on 110 degrees of freedom  
## Multiple R-squared: 0.01469, Adjusted R-squared: 0.005733   
## F-statistic: 1.64 on 1 and 110 DF, p-value: 0.203

* Fungsi untuk memprediksi data(co: hubungan jam belajar dengan IPK)

nilaibaru = data.frame(IPK=3.85)  
predict(model, nilaibaru)

## 1   
## 1.449991

1. Hubungan usia dengan IPK

* Fungsi untuk mengimport data

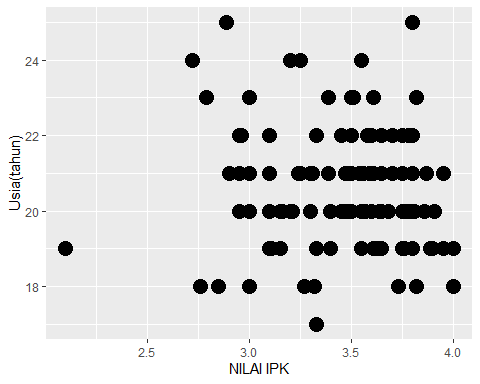
library(readxl)  
data <- read\_excel("C:/Users/Lenovo/Downloads/data.xlsx")  
data$IPK <- as.numeric(data$IPK)

* Fungsi untuk membuat scatterplot hubungan IPK dengan usia

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.1.2

ggplot(data, aes(x=IPK, y=Usia)) + geom\_point(size=5)+  
 labs(x="NILAI IPK", y="Usia(tahun)")



* Fungsi untuk mendeskripsikan variable baru antar data kategori

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.1.2

##   
## Attaching package: 'dplyr'

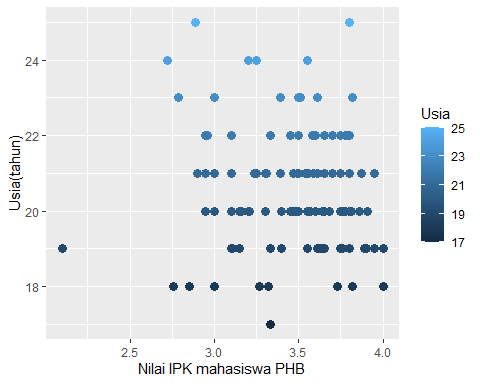
## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data = mutate(data, usia = case\_when(  
 data$`Usia` <= 20 ~ "usia muda",  
 data$`Usia` <= 25 ~ "usia pertengahan",  
 data$`Usia` >= 25 ~ "usia tua"  
),  
nilai = case\_when(  
 data$IPK <= 3 ~ "grade C ",  
 data$IPK <= 3.5 ~ "grade B",  
 data$IPK >= 3.5 ~ "grade A"  
)  
)

* Fungsi untuk membuat grafik plot(scatterplot)

ggplot(data, aes(x=IPK, y=`Usia`, color = Usia)) +  
 geom\_point(size=3)+  
 labs(x="Nilai IPK mahasiswa PHB", y="Usia(tahun)")



* Fungsi untuk membuat data yang baru dan menghapus data apabila data outlier

databaru = filter(data, `Usia` < 25)  
databaru

## # A tibble: 125 x 13  
## Nama `Jenis kelamin` Usia Prodi `Kesesuaian prodi` Pembiayaan `Kota asal`  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Fabil 0 20 2 1 0 0  
## 2 Amel 1 19 2 0 1 0  
## 3 Widia 1 21 3 1 0 1  
## 4 Sinta 1 20 7 0 0 1  
## 5 Ami 1 20 2 1 0 0  
## 6 Taufik 0 19 1 0 0 1  
## 7 Ziky 0 22 1 1 0 0  
## 8 Silvia 1 20 1 1 2 1  
## 9 Gilang 0 20 4 1 2 1  
## 10 Putri 0 22 3 0 0 0  
## # ... with 115 more rows, and 6 more variables: Kos <dbl>, Jam belajar <dbl>,  
## # Kendaraan <dbl>, IPK <dbl>, usia <chr>, nilai <chr>

* Fungsi untuk mengetahui nilai korelasi antar variable

cor(databaru$IPK, databaru$`Usia`)

## [1] -0.07827798

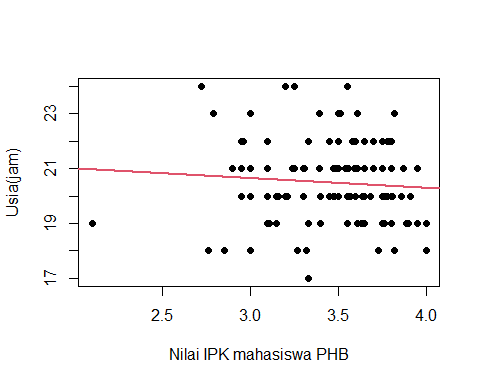
* Fungsi untuk mencari persamaan regresi

lm(`Usia`~IPK, databaru)

##   
## Call:  
## lm(formula = Usia ~ IPK, data = databaru)  
##   
## Coefficients:  
## (Intercept) IPK   
## 21.7257 -0.3529

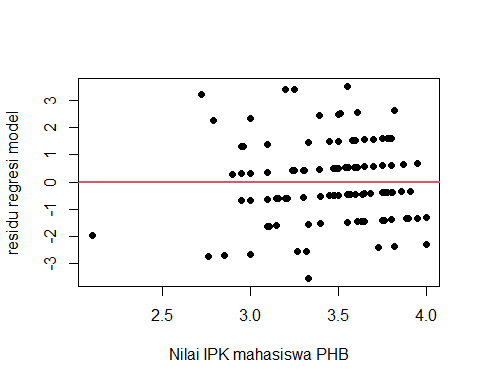
* Fungsi untuk membuat scatterplot dengan garis regresi dan korelasi

plot(databaru$IPK, databaru$`Usia`, pch=16,  
 xlab = "Nilai IPK mahasiswa PHB",  
 ylab = "Usia(jam)")  
abline(lm(databaru$`Usia` ~ databaru$IPK), col=2, lwd=2)  
text(3.4, 40, paste("y = 21.8135-0.3657x"))



* Fungsi untuk membuat plot residu(tidak ada outlier) dengan garis linear

model = lm(`Usia` ~ IPK, databaru)  
residu = resid(model)  
plot(databaru$IPK, residu, pch=16,  
 xlab = "Nilai IPK mahasiswa PHB",  
 ylab = "residu regresi model")  
abline(h=0, col=2, lwd=2)



* Fungsi untuk melihat nilai plot residu

summary(model)

##   
## Call:  
## lm(formula = Usia ~ IPK, data = databaru)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.5506 -0.6847 -0.3848 0.6399 3.5270   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 21.7257 1.4090 15.419 <2e-16 \*\*\*  
## IPK -0.3529 0.4052 -0.871 0.386   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.458 on 123 degrees of freedom  
## Multiple R-squared: 0.006127, Adjusted R-squared: -0.001953   
## F-statistic: 0.7583 on 1 and 123 DF, p-value: 0.3856

* Fungsi untuk memprediksi data(co: hubungan usia dengan IPK)

nilaibaru = data.frame(IPK=3.85)  
predict(model, nilaibaru)

## 1   
## 20.36711

## REGRES LINEAR SEDERHANA

* Fungsi untuk import data (excel)

library(readxl)  
data <-read\_excel("C:/Users/Lenovo/Downloads/data.xlsx")  
data$IPK <-as.numeric(data$IPK)

* Fungsi untuk mendeskripsikan data

summary(data)

## Nama Jenis kelamin Usia Prodi   
## Length:127 Min. :0.0000 Min. :17.00 Min. :0.00   
## Class :character 1st Qu.:0.0000 1st Qu.:20.00 1st Qu.:0.00   
## Mode :character Median :0.0000 Median :20.00 Median :2.00   
## Mean :0.4724 Mean :20.72 Mean :2.52   
## 3rd Qu.:1.0000 3rd Qu.:21.50 3rd Qu.:3.50   
## Max. :1.0000 Max. :43.00 Max. :9.00   
## Kesesuaian prodi Pembiayaan Kota asal Kos   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:1.0000   
## Median :1.0000 Median :0.0000 Median :0.0000 Median :1.0000   
## Mean :0.5669 Mean :0.5984 Mean :0.7953 Mean :0.7638   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :2.0000 Max. :4.0000 Max. :1.0000   
## Jam belajar Kendaraan IPK   
## Min. :0.000 Min. :1.000 Min. :2.100   
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:3.250   
## Median :2.000 Median :1.000 Median :3.500   
## Mean :1.551 Mean :1.299 Mean :3.460   
## 3rd Qu.:2.000 3rd Qu.:1.000 3rd Qu.:3.715   
## Max. :3.000 Max. :3.000 Max. :4.000

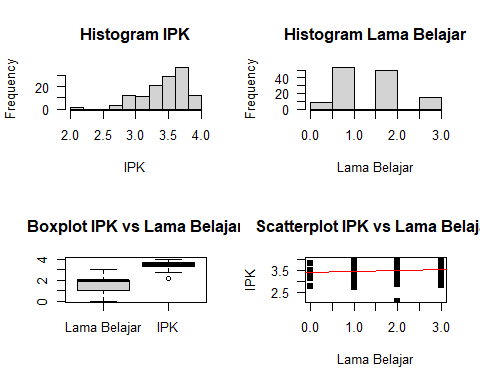
* Fungsi untuk memvisualisasikan data

par(mfrow=c(2,2))  
hist(data$`IPK`,main="Histogram IPK",xlab="IPK", las=1)

hist(data$`Jam belajar`, main ="Histogram Lama Belajar",  
 xlab="Lama Belajar", las=1)

boxplot(data$`Jam belajar`, data$`IPK`,  
main="Boxplot IPK vs Lama Belajar",  
names=c("Lama Belajar","IPK"))

plot(data$`Jam belajar`, data$`IPK`, main="Scatterplot IPK vs Lama Belajar",  
xlab="Lama Belajar",  
ylab="IPK", pch=15,las=1)  
abline(lm(data$`IPK`~data$`Jam belajar`),col="red")



* Fungsi untuk mengetahui nilai korelasi antar hubungan

waktu\_belajar = data$`Jam belajar`  
IPK=data$`IPK`  
cor.test(waktu\_belajar, IPK)

##   
## Pearson's product-moment correlation  
##   
## data: waktu\_belajar and IPK  
## t = 1.4004, df = 125, p-value = 0.1639  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.05103845 0.29216994  
## sample estimates:  
## cor   
## 0.1242813

* Fungsi untuk melihat nilai variabilitas antar hubungan data

Model\_Regresi =summary(lm(IPK~waktu\_belajar))  
Model\_Regresi

##   
## Call:  
## lm(formula = IPK ~ waktu\_belajar)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.3831 -0.1826 0.0679 0.2424 0.5169   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.38106 0.06346 53.28 <2e-16 \*\*\*  
## waktu\_belajar 0.05104 0.03645 1.40 0.164   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3248 on 125 degrees of freedom  
## Multiple R-squared: 0.01545, Adjusted R-squared: 0.007569   
## F-statistic: 1.961 on 1 and 125 DF, p-value: 0.1639