LAPORAN TUGAS PROYEK

PRAKTIKUM STATISTIKA DASAR

KLASIFIKASI MENGGUNAKAN REGRESI LOGISTIK

Dosen Pengampu : Ronny Susetyoko S.Si., M.Si

Disusun Oleh :

Wiradika Nur Fadhillah

NRP 3321600028

D4 Sains Data Terapan

PROGRAM STUDI SAINS DATA TERAPAN

POLITEKNIK ELEKTRONIKA NEGERI SURABAYA

DESEMBER 2021

A. Pendahuluan

Sebuah perusahaan telah melakukan survey, dan dari survey tersebut diperlukan prediksi preferensi merek pelanggan yang hilang dari survei yang tidak lengkap. Dalam laporan berikut, akan disajikan analisis dari kedua sampel untuk melihat apakah mereka berguna untuk memprediksi preferensi merek pelanggan dan menemukan perbedaan di antara keduanya, serta prediksi untuk preferensi merek dalam survei yang tidak lengkap.

Ada beberapa indikator yang sangat mendukung pengambilan sampel data, dimana masing-masing dibuat stratifikasi berdasarkan tingkat pendidikan, wilayah dan mobil. Jenis sampel ini direkomendasikan untuk kampanye pemasaran bertarget, tetapi kami tidak dapat menghitung proporsi total semua pelanggan yang lebih memilih Sony/Acer.

Juga, ada tanda-tanda bahwa data mungkin telah dipalsukan, kualitasnya sangat dipertanyakan. Misalnya, pelanggan melaporkan gaji dan kredit mereka dengan presisi yang tidak masuk akal di dunia nyata. Misalnya gaji yang dilaporkan sebesar $ 113.236.3836. Tidak ada manusia yang akan melaporkan gaji mereka dengan 4 tempat desimal. Konversi nilai tukar dapat menambahkan tempat desimal, tetapi meskipun demikian, manusia cenderung melaporkan gaji atau tingkat kredit mereka dalam bilangan bulat: $110.000, $75.000, mungkin $42.000, tetapi tidak terlalu sering setepat $42.235.

B. Tujuan

Tujuan dari klasifikasi ini adalah untuk membangun model prediktif dan memilih model yang dapat memprediksi preferensi merek komputer konsumen dengan akurasi setidaknya 70% pada data uji. Sasaran ideal adalah model yang dapat memprediksi preferensi merek dengan tingkat kepastian minimal 90%.

C. Metodologi

Metodologi yang digunakan adalah Regresi logistik biner yang merupakan suatu metode analisis data yang digunakan untuk mencari hubungan antar variable respon (y) yang bersifat biner atau dikotomus dengan variable prediktor (x) yang bersifat polikotomus. Fungsi regresi logistik-nya dapat dituliskan sebagai berikut.

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}$$

dimana p = banyaknya variabel prediktor x_i

D. Sumber Data

Data yang digunakan pada praktikum kali ini bersumber dari website Kaggle dengan link https://www.kaggle.com/samanemami/market-research-survey.

Kumpulan data ini mencakup CompleteResponses.csv yang merupakan pemisahan training dan SurveyIncomplete.csv untuk pemisahan testing. Bagian training terdiri dari hampir 10.000 survei yang dijawab. Kunci survei ada di dalam survey_key.csv

E. Penetapan Variabel

Variabel pada data yakni Salary (Pendapatan), Age (Usia), Elevel (Pendidikan terakhir), Car (Mobil), Zipcode (Kode Pos), Credit (Kredit yang ada), dan Brand (Merek). Disini brand menjadi variable dependen sedangkan salary, age, car, zipcode, dan credit merupakan variable independen.

F. Pre Proccesing

Melakukan pemanggilan library

```
library(caret)
library(readr)
library(ggplot2)
library(arules)
library(Metrics)
library(textclean)
library(corrplot)
```

Melakukan import data dan mendeskripsikan tipe variable untuk kedua data menampilkan sampel data

```
```{r}
#data training
complete = read.csv("CompleteResponses.csv")
str(complete)
head(complete, 10)

#data testing
incomplete = read.csv("SurveyIncomplete.csv")
str(incomplete)
head(incomplete, 10)
```
```

```
'data.frame':   9898 obs. of  7 variables:
 $ salary : num  119807 106880 78021 63690 50874 ...
 $ age    : int  45 63 23 51 20 56 24 62 29 41 ...
 $ elevel : int  0 1 0 3 3 3 4 3 4 1 ...
 $ car    : int  14 11 15 6 14 14 8 3 17 5 ...
 $ zipcode: int  4 6 2 5 4 3 5 0 0 4 ...
 $ credit : num  442038 45007 48795 40889 352951 ...
 $ brand  : int  0 1 0 1 0 1 1 1 0 1 ...
'data.frame':   5000 obs. of  7 variables:
 $ salary : num  150000 82524 115647 141443 149211 ...
 $ age    : int  76 51 34 22 56 26 64 50 26 46 ...
 $ elevel : int  1 1 0 3 0 4 3 3 2 3 ...
 $ car    : int  3 8 10 18 5 12 1 9 3 18 ...
 $ zipcode: int  3 3 2 2 3 1 2 0 4 6 ...
 $ credit : num  377980 141658 360980 282736 215667 ...
 $ brand  : int  1 0 1 1 1 1 1 1 1 0 ...
```

complete

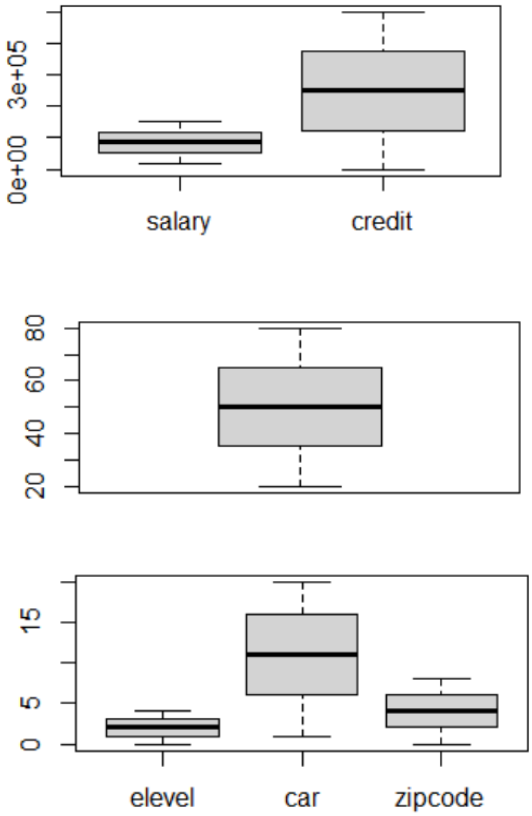
	salary <dbl>	age <int>	elevel <int>	car <int>	zipcode <int>	credit <dbl>	brand <int>
1	119806.54	45	0	14	4	442037.71	0
2	106880.48	63	1	11	6	45007.18	1
3	78020.75	23	0	15	2	48795.32	0
4	63689.94	51	3	6	5	40888.88	1
5	50873.62	20	3	14	4	352951.50	0
6	130812.74	56	3	14	3	135943.02	1
7	136459.34	24	4	8	5	80500.56	1
8	103866.90	62	3	3	0	359803.89	1
9	72298.80	29	4	17	0	276298.70	0
10	37803.33	41	1	5	4	493219.27	1

incomplete

	salary <dbl>	age <int>	elevel <int>	car <int>	zipcode <int>	credit <dbl>	brand <int>
1	150000.00	76	1	3	3	377980.1	1
2	82523.84	51	1	8	3	141657.6	0
3	115646.64	34	0	10	2	360980.4	1
4	141443.39	22	3	18	2	282736.3	1
5	149211.27	56	0	5	3	215667.3	1
6	46202.25	26	4	12	1	150419.4	1
7	125821.24	64	3	1	2	173429.4	1
8	20141.14	50	3	9	0	447716.5	1
9	135261.85	26	2	3	4	223821.2	1
10	83273.93	46	3	18	6	213961.4	0

Mencari outlier

```
set.seed(203)
# Mencari outlier.
boxplot(complete[, c("salary", "credit")])
boxplot(complete[, c("age")])
boxplot(complete[, c("elevel", "car", "zipcode")])
```



Dapat dilihat bahwa tidak ada outlier. Maka selanjutnya mencari apakah ada nilai yang hilang.

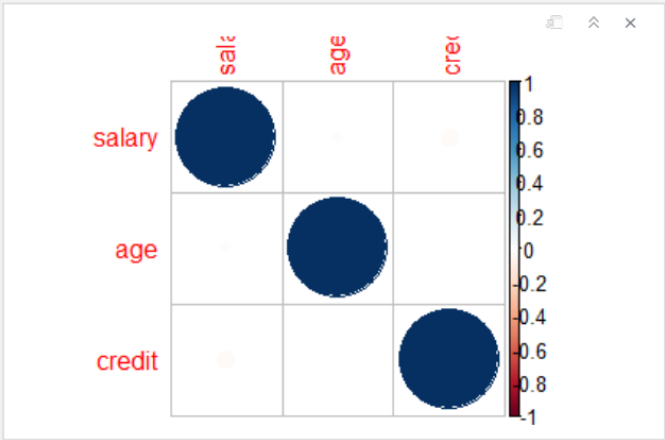
```{r}
sum(is.na(complete))
```

```
[1] 0
```

Dari data yang ada, diketahui bahwa tidak ada nilai yang hilang.

Untuk mencari korelasi antar variable, kita menggunakan plot matriks korelasi.

```{r}
corrplot(round(digits = 2, cor(complete[, c(1, 2, 6)], method = c("pearson", "kendall", "spearman"))))
```



Dapat dilihat bahwa tidak ada nilai yang hilang dan tidak ada outlier. Selanjutnya, tidak ada korelasi antara variabel numerik.

Mengubah data numerik menjadi faktor. Dan mengganti nama nilai agar lebih mudah dibaca dan dimengerti.

```{r}
#Mengubah data numerik menjadi faktor
complete$elevel <- as.factor(complete$elevel)
complete$car <- as.factor(complete$car)
complete$zipcode <- as.factor(complete$zipcode)
complete$brand <-as.factor(complete$brand)
```

```{r}
incomplete$elevel <- as.factor(incomplete$elevel)
incomplete$car <- as.factor(incomplete$car)
incomplete$zipcode <- as.factor(incomplete$zipcode)
incomplete$brand <-as.factor(incomplete$brand)
```

```{r}
#Mengubah nama
complete$elevel <- mgsub(x = complete$elevel, pattern = c(0, 1, 2, 3, 4), replacement = c("Less than HS",
    "HS", "College", "Degree", "Master's, Doc, others"))
complete$car <- mgsub(x = complete$car, pattern = c(1, 2, 3, 4,
    5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20), replacement = c("BMW",
    "Buick", "Cadillac", "Chevrolet", "Chrysler", "Dodge", "Ford", "Honda",
    "Hyundai", "Jeep", "Kia", "Lincoln", "Mazda", "Mercedes Benz", "Mitsubishi",
    "Nissan", "Ram", "Subaru", "Toyota", "None of the above"))
complete$zipcode <- mgsub(x = complete$zipcode, pattern = c(0,
    1, 2, 3, 4, 5, 6, 7, 8), replacement = c("New England", "Mid-Atlantic",
    "East North Central", "West North Central", "South Atlantic", "East South Central",
    "West South Central", "Mountain", "Pacific"))
complete$brand <- mgsub(pattern = c(0, 1), replacement = c("Acer", "Sony"),
    x = complete$brand)
```

```{r}
incomplete$elevel <- mgsub(x = incomplete$elevel, pattern = c(0, 1, 2, 3, 4), replacement = c("Less than HS",
    "HS", "College", "Degree", "Master's, Doc, others"))
incomplete$car <- mgsub(x = incomplete$car, pattern = c(1, 2, 3, 4,
    5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20), replacement = c("BMW",
    "Buick", "Cadillac", "Chevrolet", "Chrysler", "Dodge", "Ford", "Honda",
    "Hyundai", "Jeep", "Kia", "Lincoln", "Mazda", "Mercedes Benz", "Mitsubishi",
    "Nissan", "Ram", "Subaru", "Toyota", "None of the above"))
incomplete$zipcode <- mgsub(x = incomplete$zipcode, pattern = c(0,
    1, 2, 3, 4, 5, 6, 7, 8), replacement = c("New England", "Mid-Atlantic",
    "East North Central", "West North Central", "South Atlantic", "East South Central",
    "West South Central", "Mountain", "Pacific"))
incomplete$brand <- mgsub(pattern = c(0, 1), replacement = c("Acer", "Sony"),
    x = incomplete$brand)
```

Untuk persebaran data dengan variable brand pada survey yang lengkap, sebanyak 3744 pelanggan memilih Acer dan 6154 memilih Sony

```{r}
table(complete$brand, dnn= c("Brand"))
```

```
Brand
Acer Sony
3744 6154
```

Dapat dilihat pada tabel dibawah ini untuk brand Acer probabilitasnya 0.378 dan brand Sony 0.621

```{r}
prop.table(table(complete$brand, dnn = c("Brand")))
```

```
Brand
     Acer      Sony
0.3782582 0.6217418
```

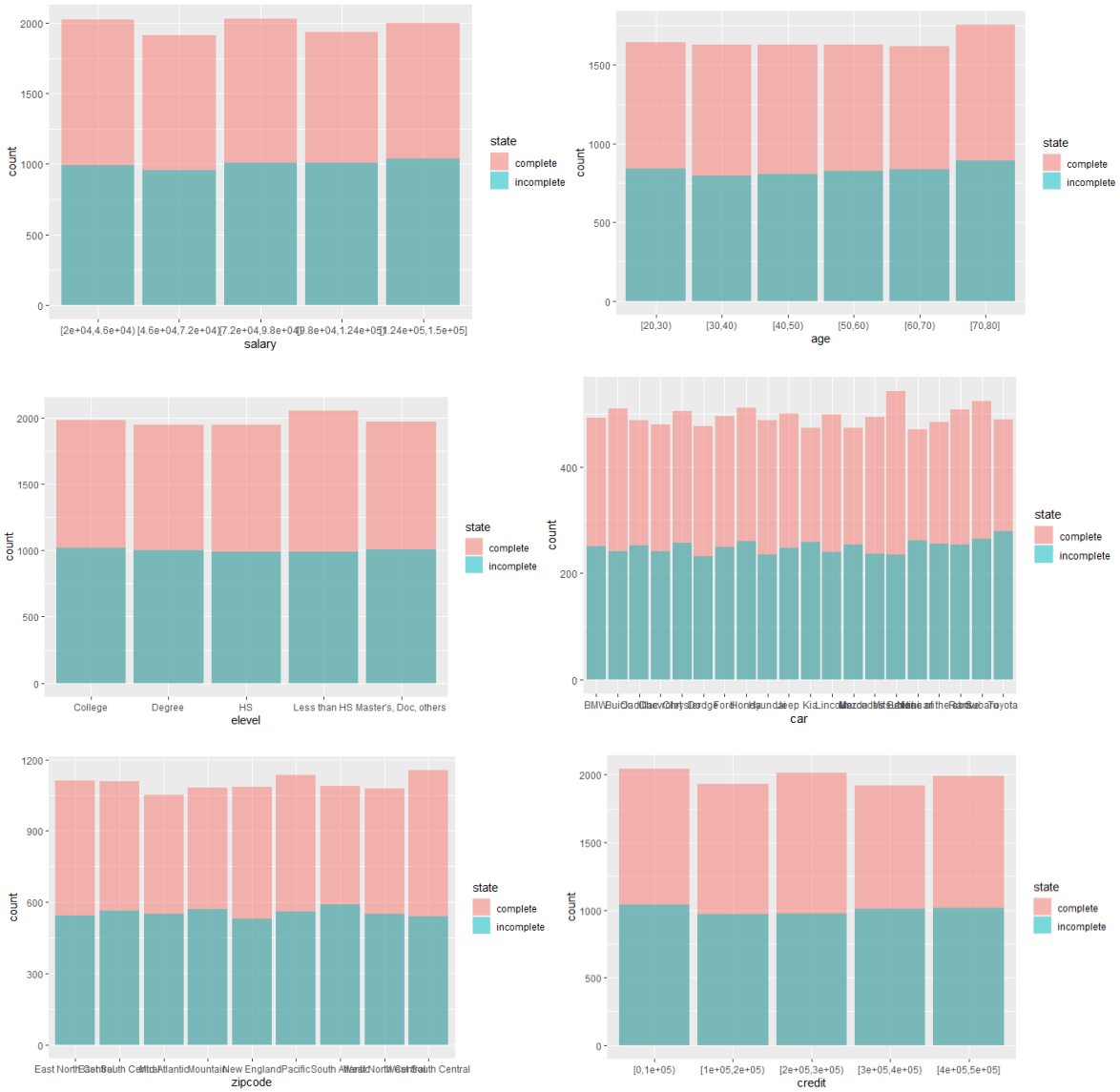
Selanjutnya mendiskritisasi variabel model untuk mendapatkan pengetahuan dan informasi tertentu tentang pola dan tren pelanggan. Di sini kita menyimpan data complete dan data incomplete tanpa diskritisasi, karena kita akan menggunakan kumpulan data ini untuk permodelan.

```{r}
complete_responses <- complete
incomplete_responses <- incomplete
# Diskritisasi
complete$salary <- discretize(complete$salary, method = "interval",
    breaks = 5)
incomplete$salary <- discretize(incomplete$salary, method = "interval",
    5)
complete$age <- discretize(complete$age, "interval", 6)
incomplete$age <- discretize(incomplete$age, method = "interval",
    breaks = 6)
complete$credit <- discretize(complete$credit, method = "interval",
    5)
incomplete$credit <- discretize(incomplete$credit, method = "interval",
    5)
```

Di sini, kita akan melihat apakah distribusi survei yang diselesaikan dan survei yang tidak diselesaikan serupa.

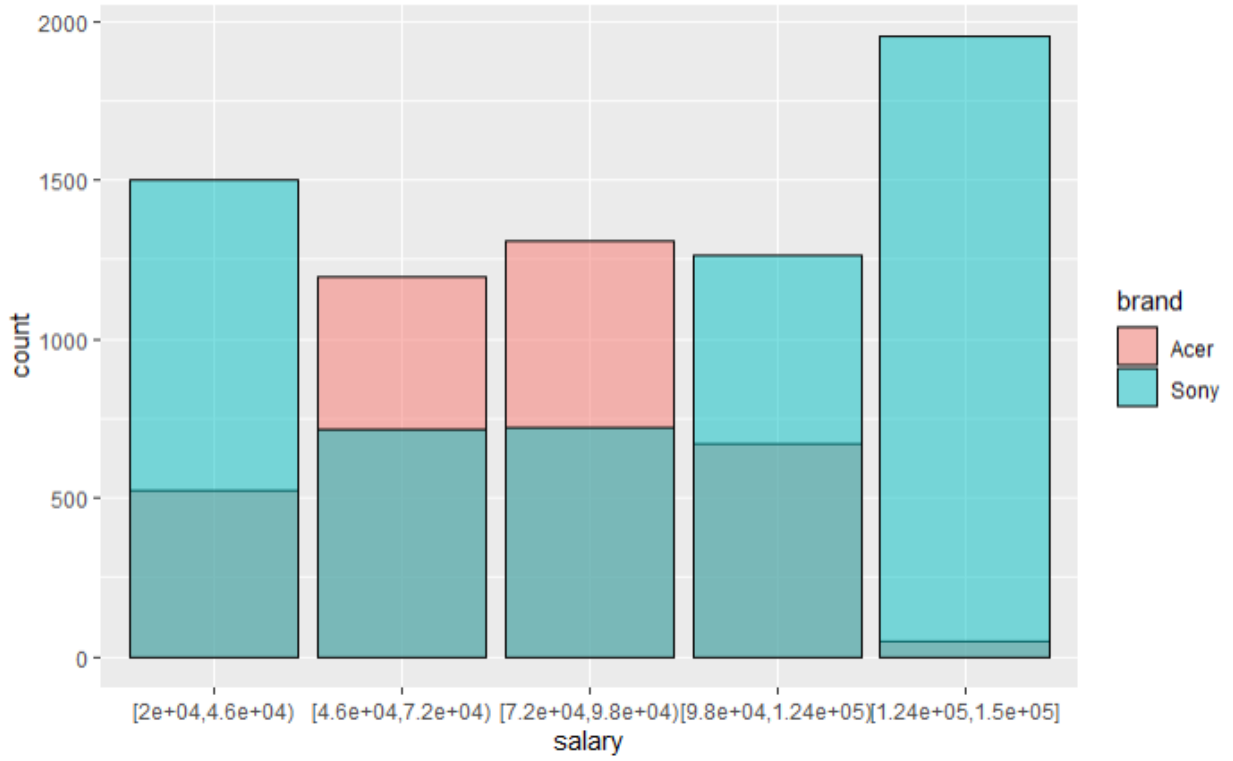
```{r}
complete$state <- "complete"
incomplete$state <- "incomplete"
whole_responses <- rbind(complete, incomplete)

for (i in c("salary", "age", "elevel", "car", "zipcode", "credit")) {
    print(ggplot(data = whole_responses, aes_string(i, fill = "state")) + geom_bar(alpha = 0.5,
        position = "identity"))
}
```

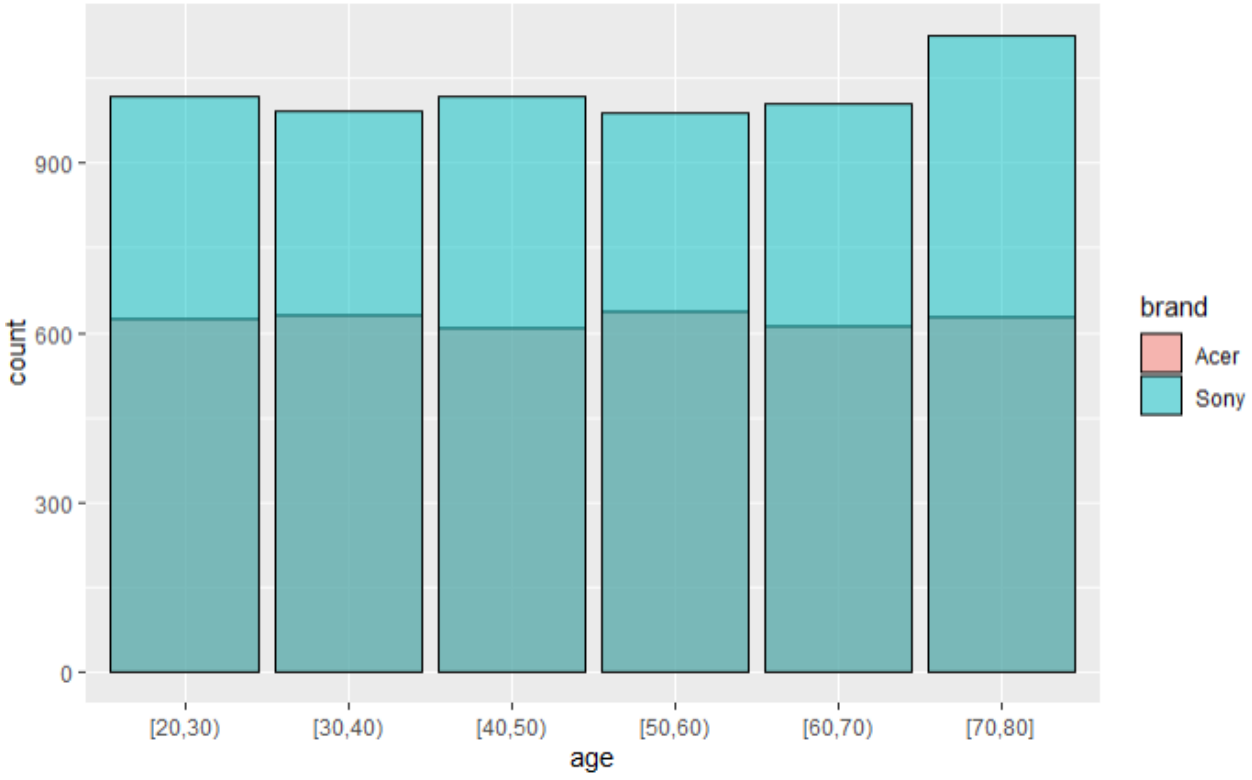


Kita bisa melihat, distribusi di kedua dataset cukup mirip, yang berarti bahwa model berdasarkan survei lengkap yang digunakan dalam survei tidak lengkap harus konsisten.

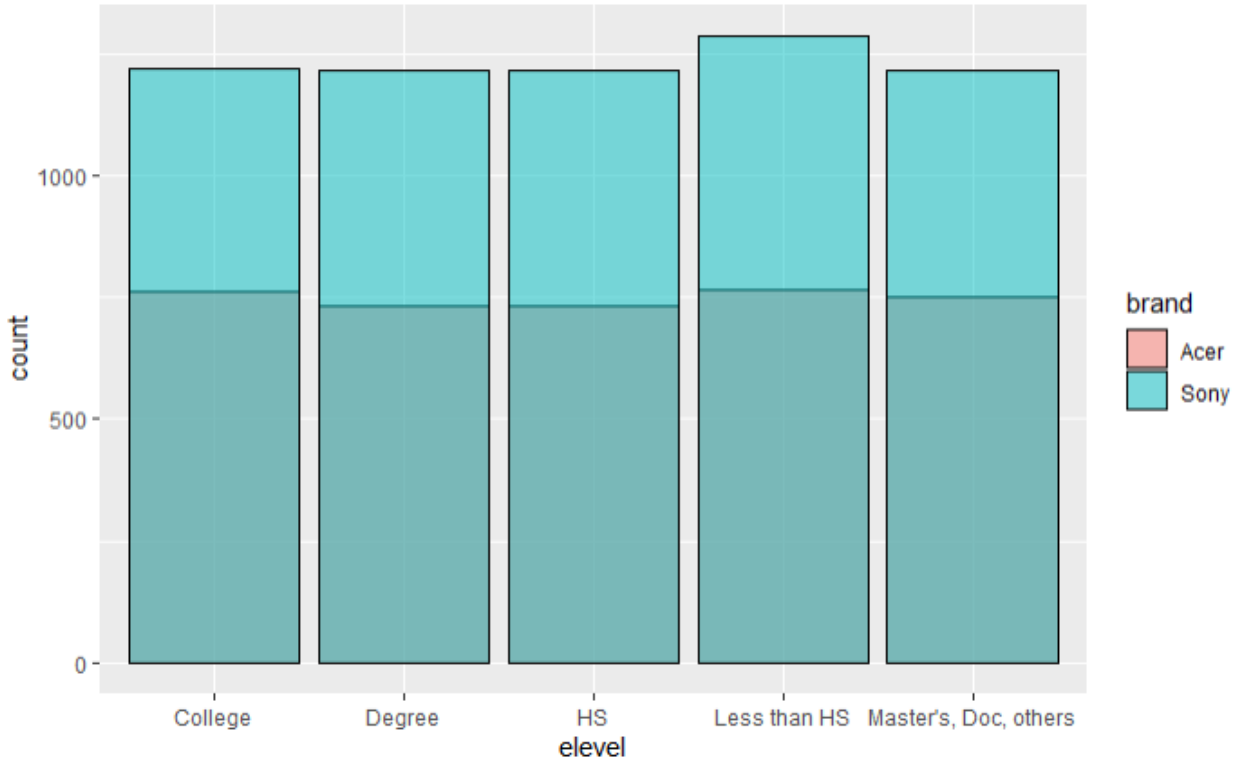
```{r}
ggplot(data = complete, mapping = aes(x = salary, fill = brand)) + geom_bar(alpha = 0.5,
    position = "identity", color = "black", bins = 20, )
```



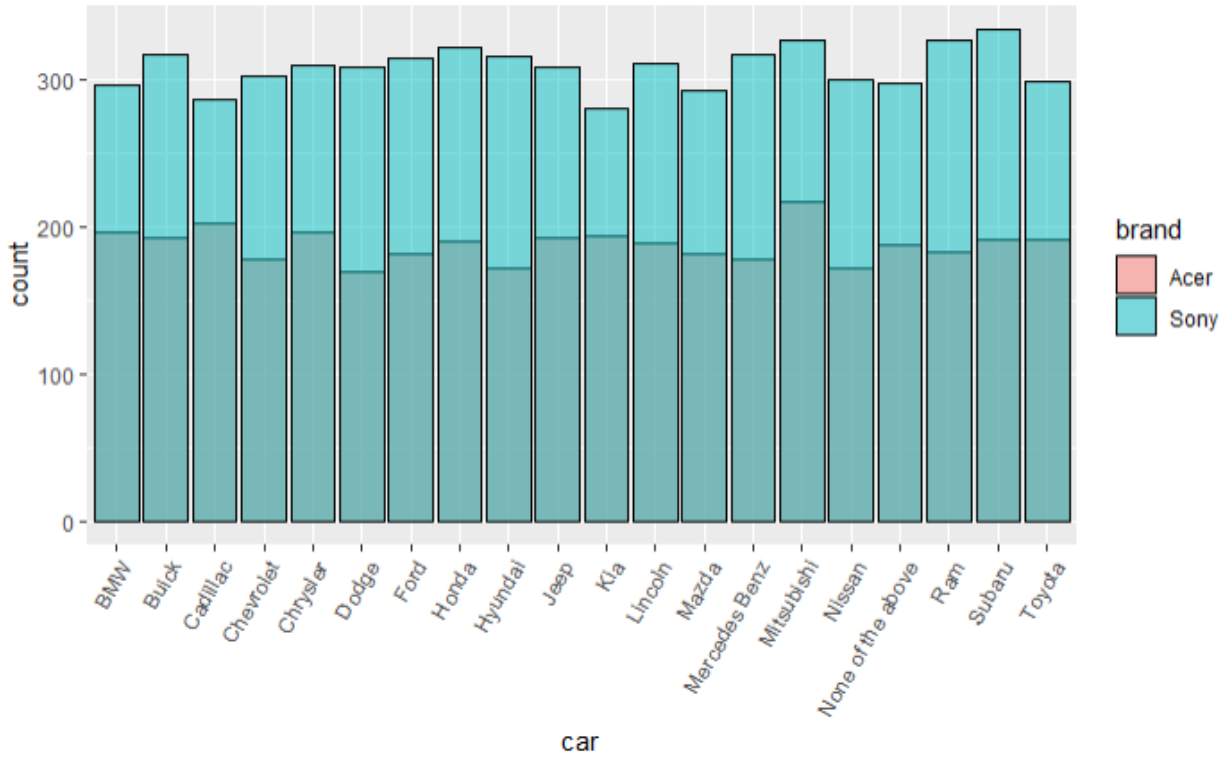

```r
ggplot(data = complete, mapping = aes(x = age, fill = brand)) + geom_bar(alpha = 0.5,
    position = "identity", color = "black", bins = 20)
```



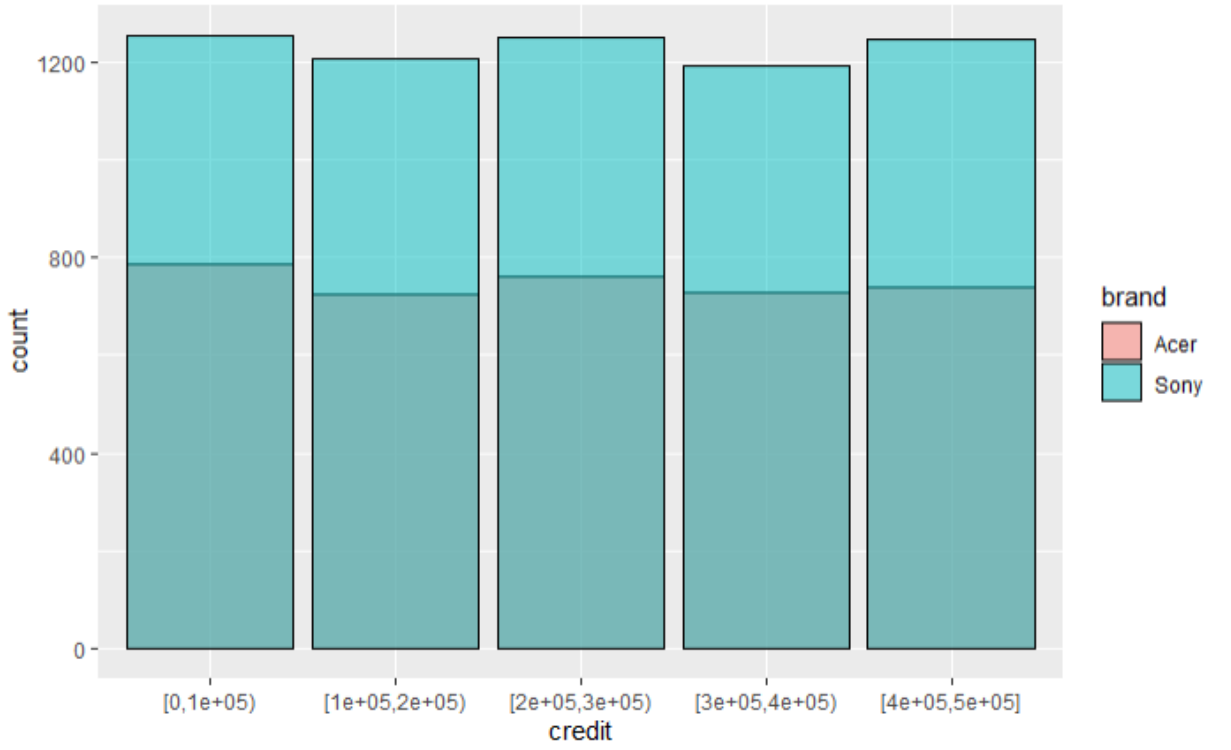
```r
ggplot(data = complete, mapping = aes(x = elevel, fill = brand)) + geom_bar(alpha = 0.5,
    position = "identity", color = "black", bins = 20)
```




```r
ggplot(data = complete, mapping = aes(x = car, fill = brand)) + geom_bar(alpha = 0.5,
    position = "identity", color = "black", bins = 20) + theme(axis.text.x = element_text(angle = 60,
    hjust = 1))
```




```r
ggplot(data = complete, mapping = aes(x = credit, fill = brand)) + geom_bar(alpha = 0.5,
    position = "identity", color = "black", bins = 20)
```



Dalam bagan ini, kita dapat mengamati bahwa survei ini bertingkat, karena setiap grup di setiap fitur terwakili secara setara.

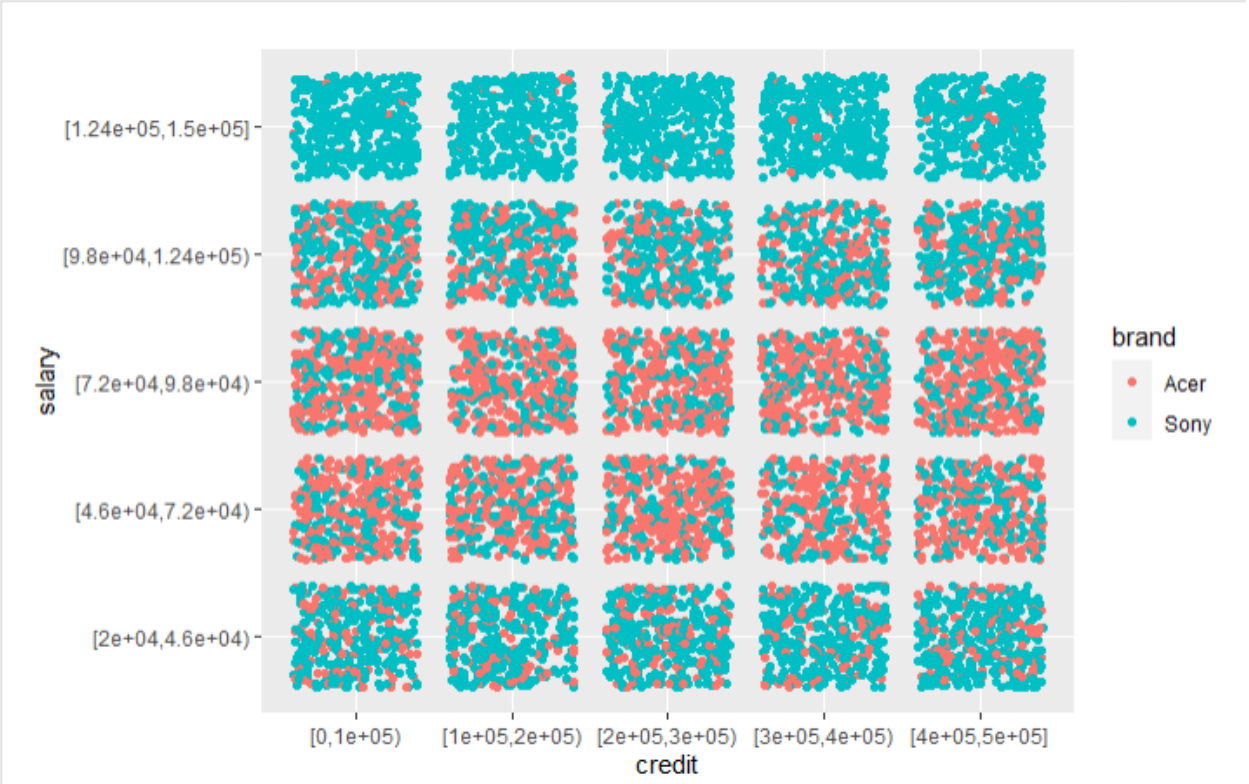
Scatterplot antar variabel

a. Antara credit dan age terhadap brand

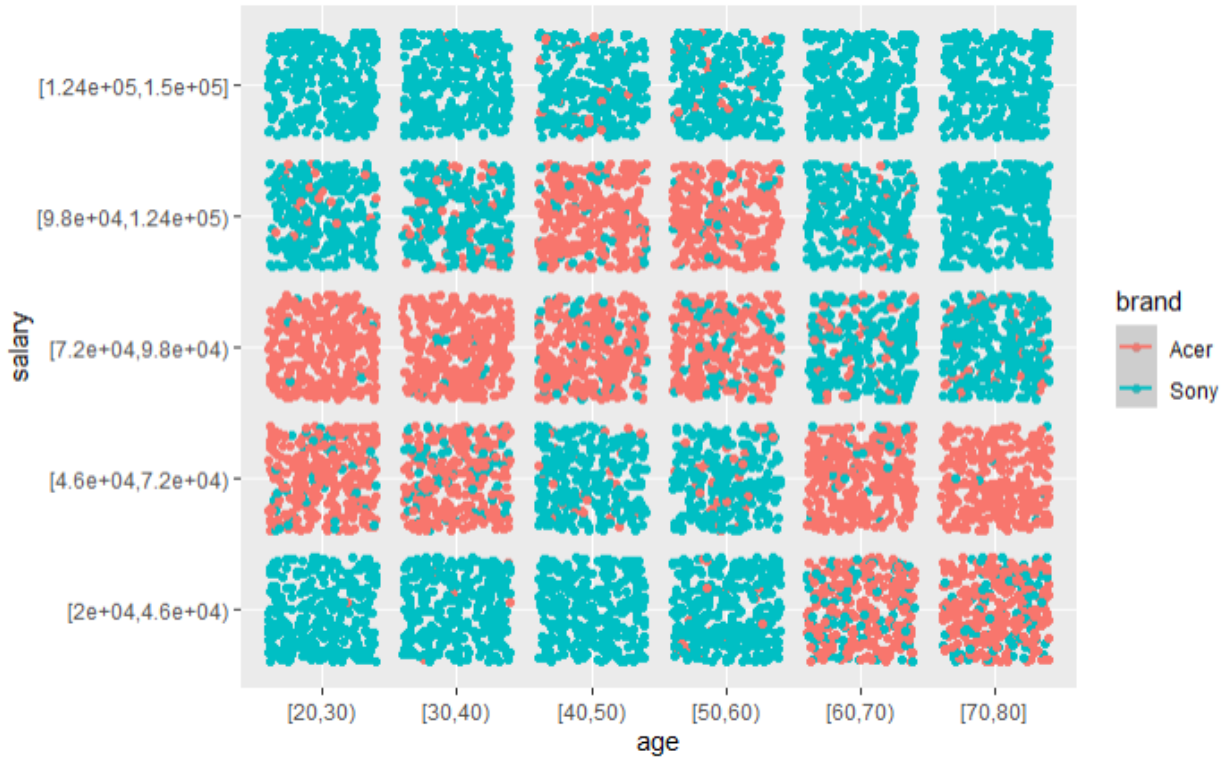
```r
ggplot(data = complete, mapping = aes(x = age, y = credit, color = brand)) +
    geom_jitter()
```

b. Antara salary dan credit terhadap brand

```{r}
ggplot(data = complete, mapping = aes(x = credit, y = salary, color = brand)) +
    geom_jitter()
```




```{r}
ggplot(data = complete, mapping = aes(x = age, y = salary, color = brand)) +
    geom_jitter() + geom_smooth()
```



Kita dapat mengamati pola yang jelas dalam diagram sebaran usia vs gaji. Konsumen antara 20 dan 40 tahun yang memiliki gaji antara 46000 dan 98000 cenderung membeli Acer, dan sisanya Sony. Mereka yang berusia antara 40 dan 60 tahun dan memiliki gaji antara 72000 dan 124000 juga lebih mungkin untuk membeli Acer. Terakhir, konsumen yang berusia antara 60 hingga 80 tahun cenderung membeli Acer jika gajinya antara 20000 hingga 72000.

G. Permodelan

```r
set.seed(123)
sel_compl_responses <- complete[, c(1, 2, 7)]
# We build the new train and test sets.
training_index <- createDataPartition(y = sel_compl_responses$brand, p = 0.75,
    list = FALSE)
strainSet <- sel_compl_responses[training_index, ]
stestSet <- sel_compl_responses[-training_index, ]
strainSet$brand <- as.factor(strainSet$brand)

model.reglog <- glm(brand~.,data=strainSet, family="binomial")
summary(model.reglog)
```

```
Call:
glm(formula = brand ~ ., family = "binomial", data = strainSet)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7195  -0.9435   0.2261   0.8839   1.4940

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                1.07119    0.08416  12.727  < 2e-16 ***
salary[4.6e+04,7.2e+04)   -1.53130    0.07993 -19.157  < 2e-16 ***
salary[7.2e+04,9.8e+04)   -1.67195    0.07949 -21.035  < 2e-16 ***
salary[9.8e+04,1.24e+05) -0.43692    0.08020  -5.448  5.1e-08 ***
salary[1.24e+05,1.5e+05]  2.61395    0.17645  14.814  < 2e-16 ***
age[30,40)                -0.03841    0.09436  -0.407    0.684
age[40,50)                -0.01238    0.09437  -0.131    0.896
age[50,60)                -0.11847    0.09382  -1.263    0.207
age[60,70)                -0.03117    0.09488  -0.329    0.743
age[70,80]                 0.10409    0.09281   1.122    0.262
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9847.1  on 7423  degrees of freedom
Residual deviance: 7846.8  on 7414  degrees of freedom
AIC: 7866.8

Number of Fisher Scoring iterations: 6
```

```{r}
(exp(coef(model.reglog)))
```

```
              (Intercept)  salary[4.6e+04,7.2e+04)  salary[7.2e+04,9.8e+04)
                2.9188583                0.2162552                0.1878796
 salary[9.8e+04,1.24e+05) salary[1.24e+05,1.5e+05]               age[30,40)
                0.6460208               13.6528785                0.9623159
               age[40,50)               age[50,60)               age[60,70)
                0.9876934                0.8882772                0.9693079
               age[70,80]
                1.1097046
```

Dari pengujian odd-ratio, diketahui bahwa berdasarkan variable salary dan age pelanggan lebih banyak cenderung untuk memilih merek Sony daripada Acer

```{r}
pred1 <- predict(model.reglog, stestSet, type="response")
predicted1 <- round(pred1)
tab1 <- table(Predicted = predicted1, Reference = stestSet$brand)
tab1
```

```
         Reference
Predicted Acer Sony
        0  639  357
        1  297 1181
```

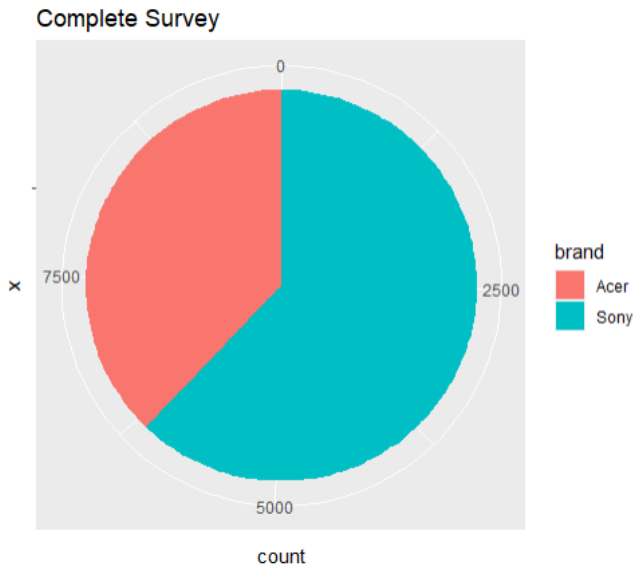
Hasil prediksi menunjukkan bahwa berdasarkan data uji, dari 2474 observasi terdapat 639 observasi categori Acer. Model belum berhasil memprediksi dengan tepat karena sebanyak 357 orang yang diprediksi memilih Acer ternyata memilih Sony. Serta terdapat 1181 observasi categori Sony dan model juga belum berhasil memprediksi dengan tepat karena sebanyak 297 orang yang diprediksi memilih Sony ternyata memilih Acer.

```{r}
# Creating a dataframe of observed and predicted data
act_pred1 <- data.frame(observed = stestSet$brand, predicted =
                          factor(predicted1))
#% accuracy benar
sum(diag(tab1))/sum(tab1)
```

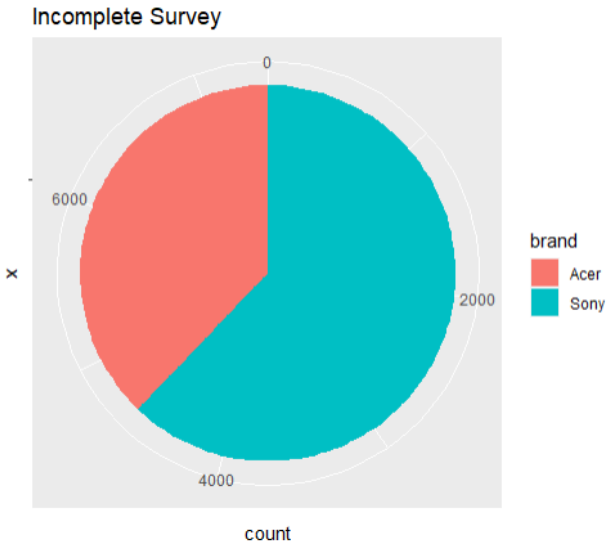
```
[1] 0.7356508
```

Tingkat akurasi benar untuk model yang telah dibuat sebesar 73,56 % yang bisa dibilang masih kurang akurat.

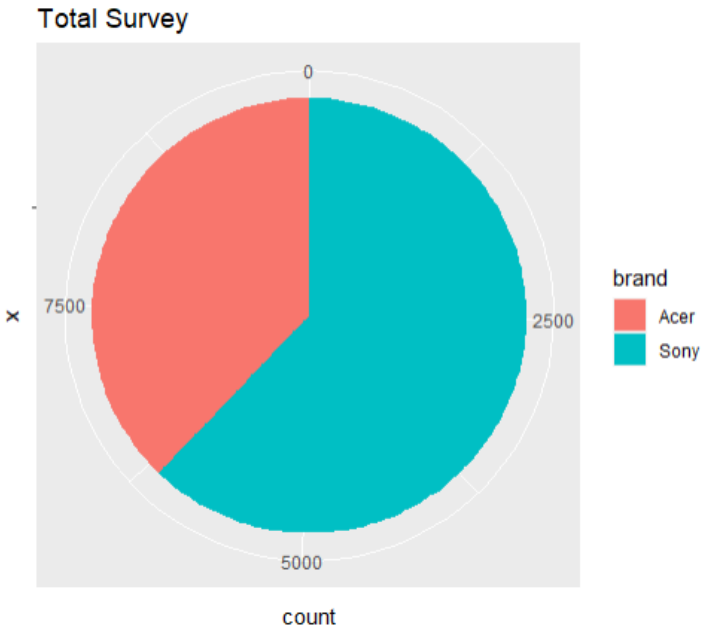
```{r}
ggplot(complete_responses, aes(x = "", fill = brand)) + geom_bar() + coord_polar(theta = "y") +
    ggtitle("Complete Survey")
```




```{r}
ggplot(predicted_responses, aes(x = "", fill = brand)) + geom_bar() + coord_polar(theta = "y") +
    ggtitle("Incomplete Survey")
```



```r
ggplot(rbind(complete_responses, predicted_responses$brand), aes(x = "", fill = brand)) +
    geom_bar() + coord_polar(theta = "y") + ggtitle("Total Survey")
```



Kita dapat mengamati bahwa distribusi Acer dan Sony sangat mirip dalam survei lengkap dan survei tidak lengkap. Hal ini diharapkan karena profil pelanggan dan distribusinya sangat mirip, hampir sama. Jadi total survei menunjukkan bahwa merek pilihan untuk pelanggan adalah Sony, seperti yang diharapkan.

H. Kesimpulan

- Pelanggan lebih memilih Sony daripada Acer. Jadi, jika harus memilih di antara satu merek, akan diprioritaskan merek Sony.
- **Pelanggan yang lebih memilih Acer** adalah
    - Antara 20 dan 40 tahun, dengan gaji antara 50.000 dan 100.000.
    - Antara 40 dan 60 tahun, dengan gaji antara 80.000 dan 120.000.
    - Antara 60 dan 80 tahun, dengan gaji kurang dari 80.000.

    Pengguna lainnya lebih memilih Sony.

I. Evaluasi

Tingkat akurasi dari permodelan dengan menggunakan metode regresi logistik biner ini kurang akurat, sehingga memungkinkan untuk adanya model lain yang lebih akurat dalam penetuan prediksi.