Jatim Camp#4

Workshop Business Analytic Intelligence And Machine Learning With Rstudio

26 – 27 April 2019, At Hotel Cleo Surabaya

Profile

Amri Muhaimin
Pendidikan S1 Statistika Institiute Teknologi Sepuluh Nopember
Organisasi Professional Statistics (Statistika ITS) 2016-2017
Digitalent Scholarship Batch 1 by Kominfo
Staff Internal DSI Jatim 2019

A Wildan Al Azis
Pendidikan S1 Statistika Institiute Teknologi Sepuluh Nopember
Staff Internal DSI Jatim 2019
Data Scientist At Fineoz Jakarta

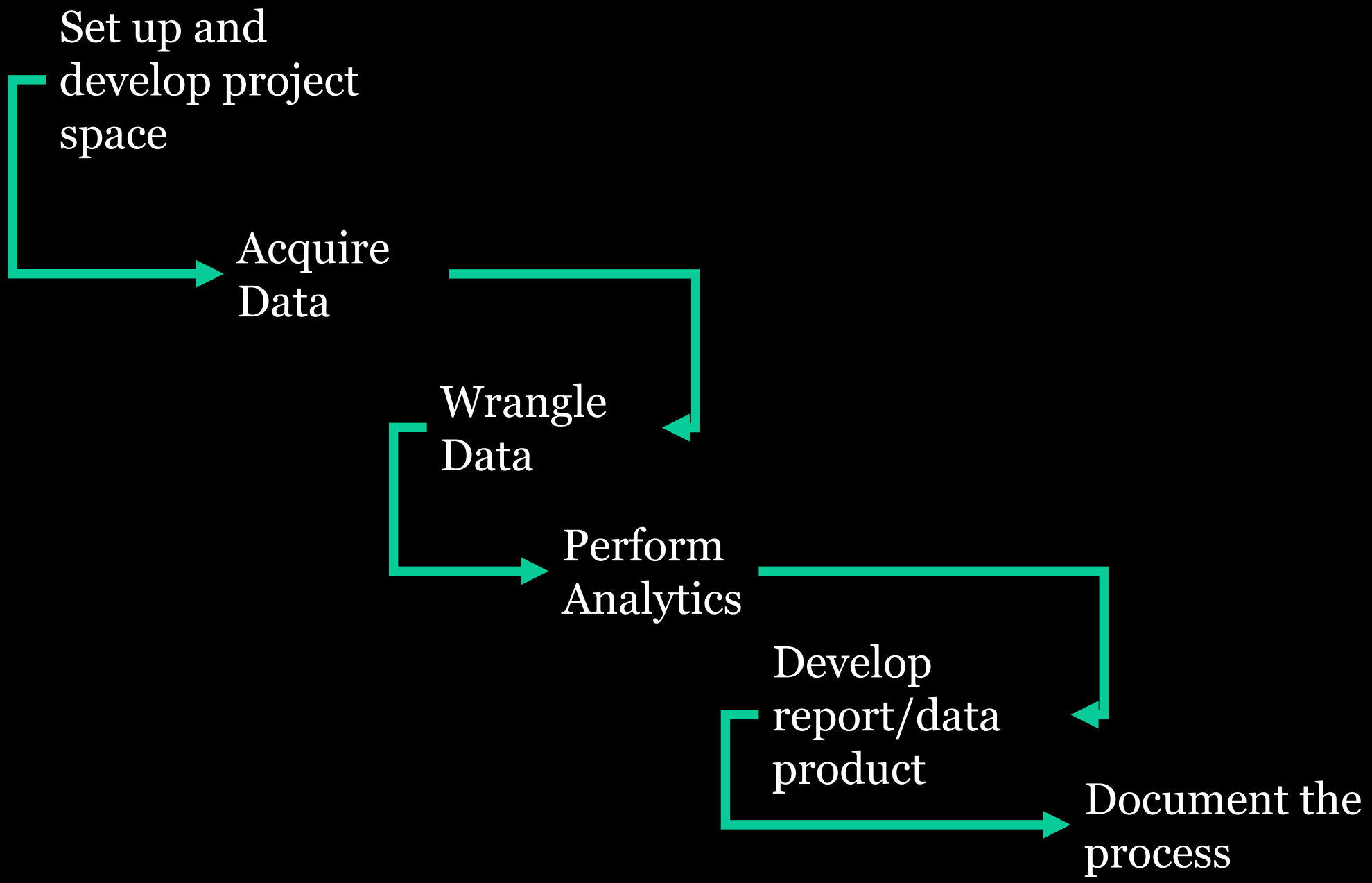
Outline

Day 1

1. Preview BI
2. Getting, Collecting Data
3. Cleaning and Preparing Data
4. Visualize Data
5. Machine Learning

Day 2

1. Deep Down Machine Learning
    - Feature Selection
    - Oversampling
    - Model Comparison


Set up and develop project space
Acquire Data
Wrangle Data
Perform Analytics
Develop report/data product
Document the process

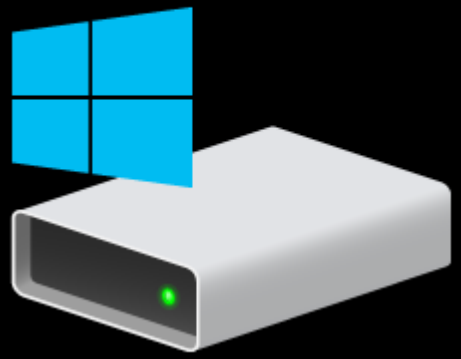
Sebelum Dimulai

```
library(twitteR)
library(dplyr)
library(zoo)
library(ggplot2)
library(reshape)
library(VIM)
library(Hmisc)
library(mice)
library(datasets)
library(gganimate)
```

```
library(gganimate)
library(gapminder)
library(tidyr)
library(graphics)
library(RCurl)
library(cowplot)
library(forecast)
library(caret)
library(DMwR)
library(MASS)
library(caTools)
```

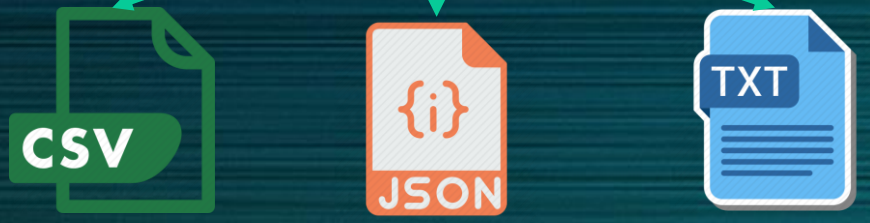
Internet

1. Crawling
2. Scraping



Local Disk

Beberapa Contoh Format Data



read.csv fromJson read.table

Melihat Directory

```
getwd()
"C:/Users/user/Documents"
```

Mengimpor Data Ke R#1

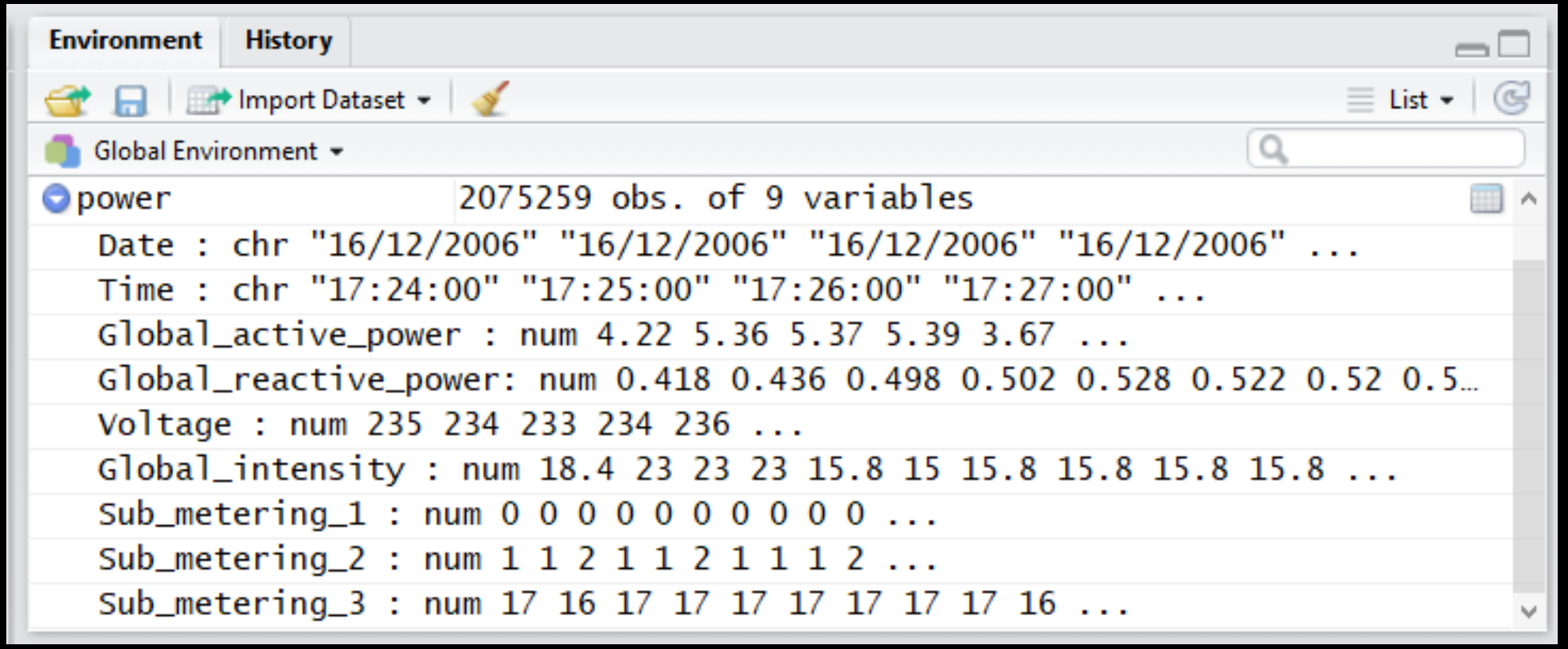
```
#download
download.file("http://archive.ics.uci.edu/ml/machine-learning-
databases/00235/household_power_consumption.zip",
              destfile = "C:/Users/user/Documents/household_power_consumption.zip")

#unzip
unzip("C:/Users/user/Documents/household_power_consumption.zip")

#read to R
power = read.table("C:/Users/user/Documents/household_power_consumption.txt",
sep=";", header=T, na.strings=c("?",""), stringsAsFactors=FALSE)
```

Melakukan download dari internet, melakukan unzip (jika data dalam format rar). Terakhir mengimport data tersebut ke dalam R.

Mengimpor Data Ke R#1

View(power)

	Date	Time	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_mete
1	16/12/2006	17:24:00	4.216	0.418	234.84	18.4	0	1	
2	16/12/2006	17:25:00	5.360	0.436	233.63	23.0	0	1	
3	16/12/2006	17:26:00	5.374	0.498	233.29	23.0	0	2	
4	16/12/2006	17:27:00	5.388	0.502	233.74	23.0	0	1	
5	16/12/2006	17:28:00	3.666	0.528	235.68	15.8	0	1	
6	16/12/2006	17:29:00	3.520	0.522	235.02	15.0	0	2	
7	16/12/2006	17:30:00	3.702	0.520	235.09	15.8	0	1	
8	16/12/2006	17:31:00	3.700	0.520	235.22	15.8	0	1	
9	16/12/2006	17:32:00	3.668	0.510	233.99	15.8	0	1	
10	16/12/2006	17:33:00	3.662	0.510	233.86	15.8	0	2	
11	16/12/2006	17:34:00	4.448	0.498	232.86	19.6	0	1	
12	16/12/2006	17:35:00	5.412	0.470	232.78	23.2	0	1	
13	16/12/2006	17:36:00	5.224	0.478	232.99	22.4	0	1	
14	16/12/2006	17:37:00	5.268	0.398	232.91	22.6	0	2	
15	16/12/2006	17:38:00	4.054	0.422	235.24	17.6	0	1	
16	16/12/2006	17:39:00	3.384	0.282	237.14	14.2	0	0	
17	16/12/2006	17:40:00	3.270	0.152	236.73	13.8	0	0	


Environment
History
Import Dataset
List
Global Environment
power 2075259 obs. of 9 variables
Date : chr "16/12/2006" "16/12/2006" "16/12/2006" "16/12/2006" ...
Time : chr "17:24:00" "17:25:00" "17:26:00" "17:27:00" ...
Global_active_power : num 4.22 5.36 5.37 5.39 3.67 ...
Global_reactive_power: num 0.418 0.436 0.498 0.502 0.528 0.522 0.52 0.5...
Voltage : num 235 234 233 234 236 ...
Global_intensity : num 18.4 23 23 23 15.8 15 15.8 15.8 15.8 15.8 ...
Sub_metering_1 : num 0 0 0 0 0 0 0 0 0 0 ...
Sub_metering_2 : num 1 1 2 1 1 2 1 1 1 2 ...
Sub_metering_3 : num 17 16 17 17 17 17 17 17 17 16 ...

Mengimpor Data Ke R#2

```
telco = read.csv("TelcoChurn.csv",sep = “;",header=T)
```

Jika data sudah berada dalam local disk dan sesuai format yang dapat dibaca oleh R, mengimport data ke dalam R tidak begitu rumit. Tapi ingat, perhatikan separator dalam format data tersebut.

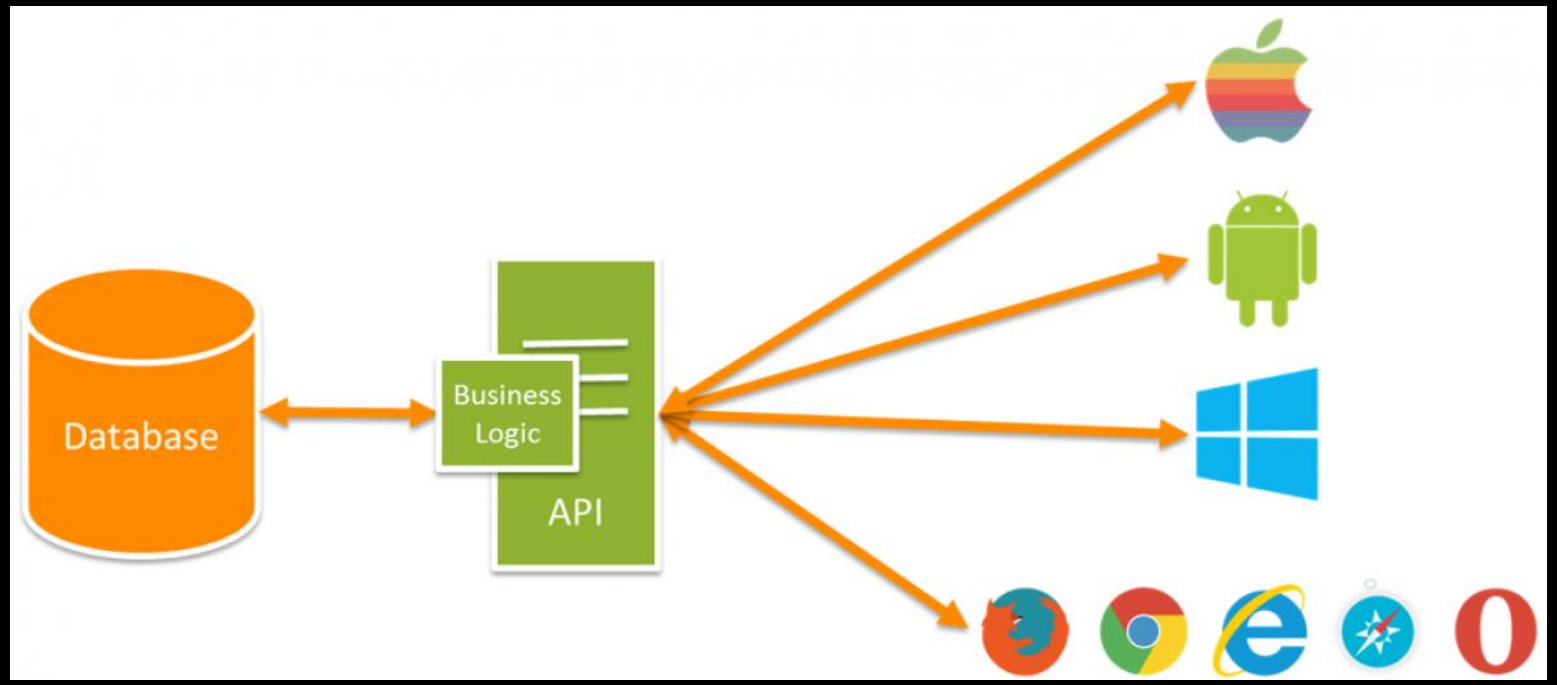
Mengimpor Data Ke R#2

View(telco)

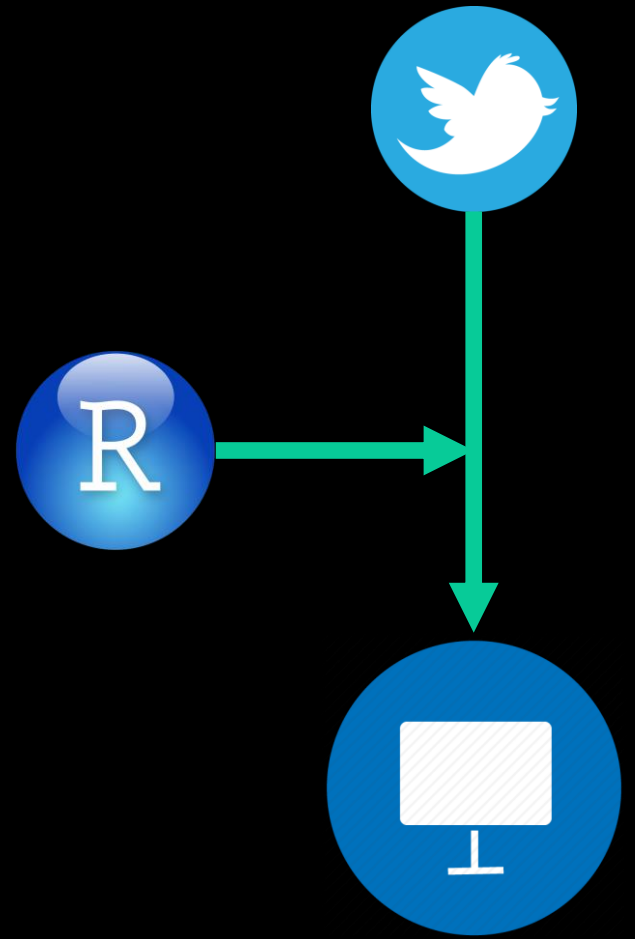
	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity
1	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No
2	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes
3	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes
4	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes
5	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No
6	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No
7	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No
8	6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes
9	7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No
10	6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes
11	9763-GRSKD	Male	0	Yes	Yes	13	Yes	No	DSL	Yes
12	7469-LKBCI	Male	0	No	No	16	Yes	No	No	No internet service
13	8091-TTVAX	Male	0	Yes	No	58	Yes	Yes	Fiber optic	No
14	0280-XJGEX	Male	0	No	No	49	Yes	Yes	Fiber optic	No
15	5129-JLPIS	Male	0	No	No	25	Yes	No	Fiber optic	Yes
16	3655-SNQYZ	Female	0	Yes	Yes	69	Yes	Yes	Fiber optic	Yes
17	8191-XWSZG	Female	0	No	No	52	Yes	No	No	No internet service

Crawling Data From Internet

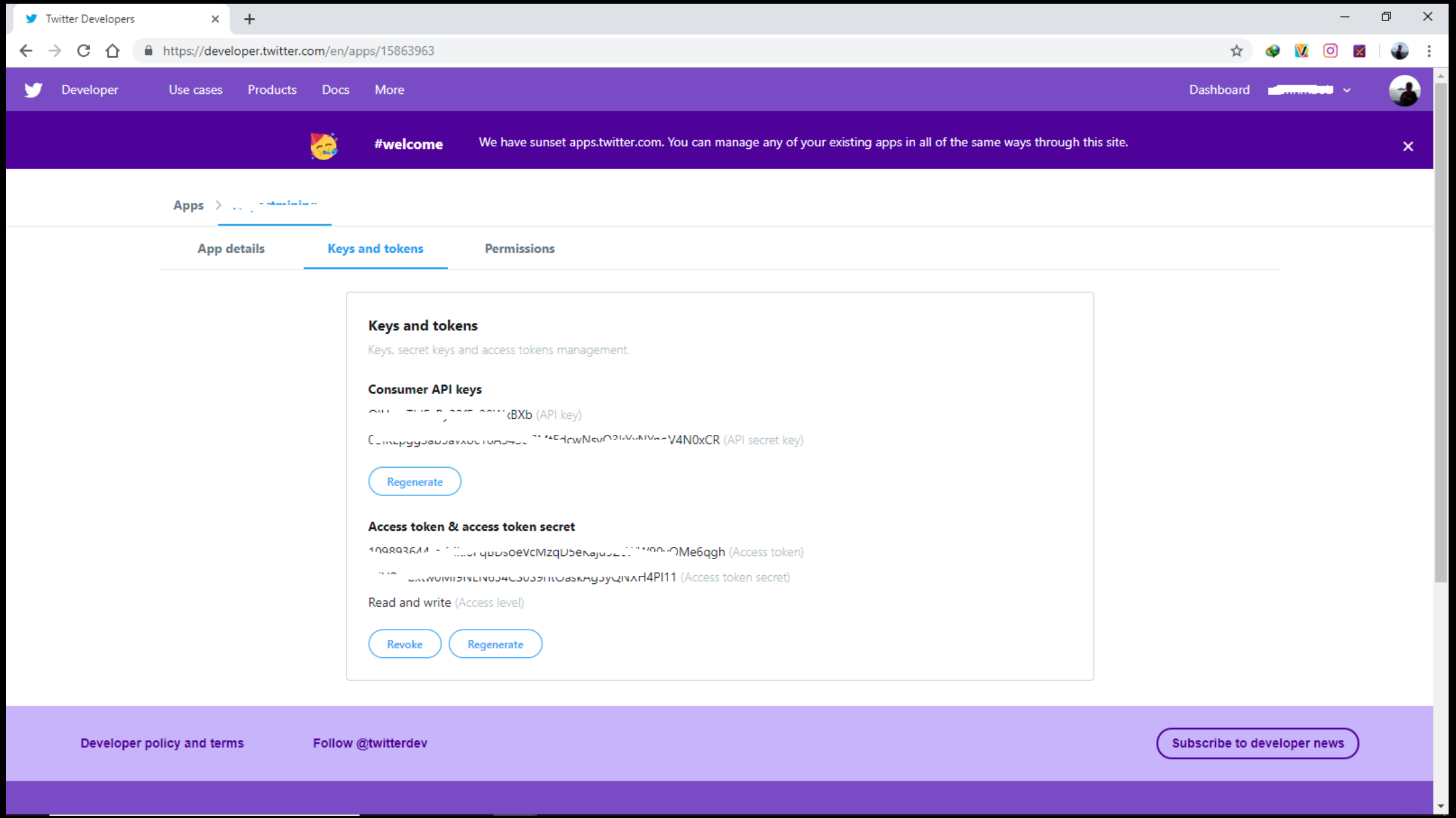
- Crawling adalah aplikasi script program untuk melakukan scan kesemua halaman di internet dan dibuatkan index untuk data yang dicarinya



Skema crawling



Contoh Mengambil Tweet dari Twitter



Collect Tweet From Twitter

```
#Definisikan API
consumer_key = 'OlHsruThlEzRyXXXXXXX'
consumer_secret = '0BlKLpggSab5avx0c16A5XXXXXXXXXXXXXXXXXXXXXXXXX'
access_token = ‘XXXXXXXX-erMkl8PqBDsoeVcMzqD5eKaju9ZcWW90vOMe6qgh'
access_secret = ‘XXXXXXXXXXXXXXXEN634CS0S9rItOaskAg3yQNXH4Pl11'
#Menseetup
setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)
#Mulai Mencari Tweet
tweets = searchTwitter("#AvengersEndgame", n = 1000, lang = "en")
#merubah tweet kedalam bentuk data frame
tweets = twListToDF(tweets)
```

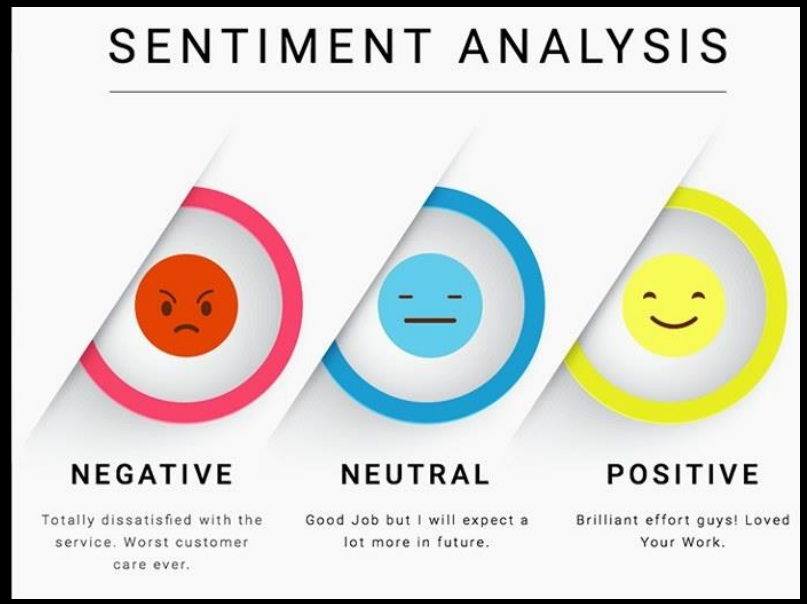
View(tweets)

	text	favorited	favoriteCount	replyToSN	created	truncated	repl
1	RT @RoyMwangi10: #AvengersEndgame if you know you kn...	FALSE	0	NA	2019-04-26 04:15:19	FALSE	NA
2	RT @elisaafaberi: one taught me love, one taught me patien...	FALSE	0	NA	2019-04-26 04:15:19	FALSE	NA
3	Less than an hour till #AvengersEndgame	FALSE	0	NA	2019-04-26 04:15:19	FALSE	NA
4	RT @elisaafaberi: one taught me love, one taught me patien...	FALSE	0	NA	2019-04-26 04:15:19	FALSE	NA
5	RT @roywoodjr: Question for current/former movie employ...	FALSE	0	NA	2019-04-26 04:15:19	FALSE	NA
6	RT @ebuyhouseinc: https://t.co/yRzG4TgMcd is here to revo...	FALSE	0	NA	2019-04-26 04:15:19	FALSE	NA
7	RT @TheFlippist: <ed><U+00A0><U+00BD><ed><U+00B4...	FALSE	0	NA	2019-04-26 04:15:19	FALSE	NA
8	I have more to say than this but its aaaaaaaaall spoilers, so t...	FALSE	0	NA	2019-04-26 04:15:18	FALSE	NA
9	Oh man. #AvengersEndgame was just fabulous, amazing &a...	FALSE	0	NA	2019-04-26 04:15:18	TRUE	NA
10	It's an epic film, @genmillscereal! #AlreadyBeen to see @Ma...	FALSE	0	NA	2019-04-26 04:15:18	FALSE	NA
11	RT @Istrmendigoria_: My new alarm #AvengersEndgame htt...	FALSE	0	NA	2019-04-26 04:15:17	FALSE	NA
12	@RobertDowneyJr you made us cry. #AvengersEndgame	FALSE	0	RobertDowneyJr	2019-04-26 04:15:17	FALSE	NA
13	RT @elisaafaberi: one taught me love, one taught me patien...	FALSE	0	NA	2019-04-26 04:15:17	FALSE	NA
14	RT @_PVRCinemas: #Retweet if you are going to see #Aveng...	FALSE	0	NA	2019-04-26 04:15:17	FALSE	NA
15	#AvengersEndgame is the greatest film of all time.	FALSE	0	NA	2019-04-26 04:15:16	FALSE	NA
16	My god. I HAVE NO WORDS The <ed><U+00A0><U+00BD>...	FALSE	4	NA	2019-04-26 04:15:16	FALSE	NA
17	RT @Avengers: We're in the endgame now. Download the hr...	FALSE	0	NA	2019-04-26 04:15:16	FALSE	NA

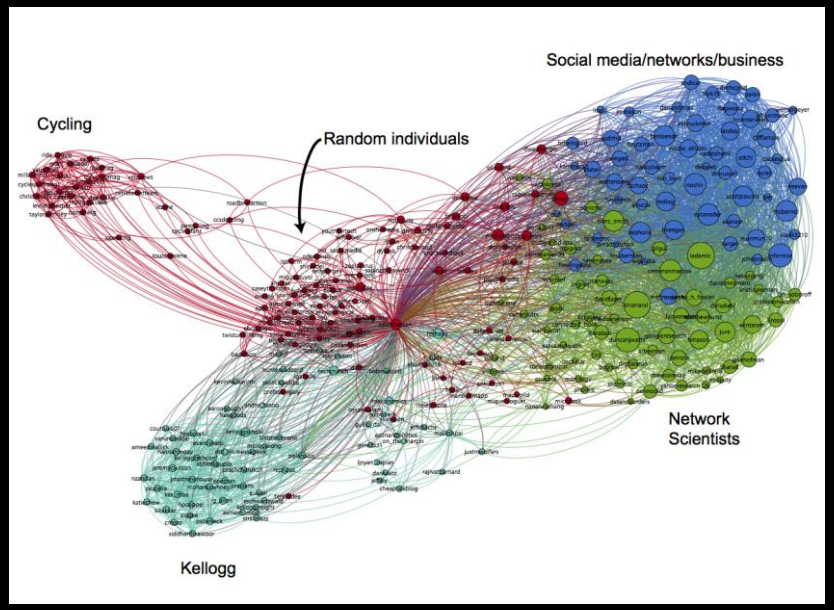
Contoh Teks Analisis

Wordcloud

Sentiment Analysis



Social Network Analysis



Cleaning & Preparing Data

Memahami Data

Dalam melakukan cleaning dan preparing, pahami tentang data terlebih dahulu. Misal mengetahui tentang skala datanya, serta tujuan data tersebut ada.

Melihat Struktur Data Frame Power

str(power)

```
'data.frame':   2075259 obs. of  9 variables:
 $ Date                  : chr  "16/12/2006" "16/12/2006" "16/12/2006" "16/12/2006" ...
 $ Time                  : chr  "17:24:00" "17:25:00" "17:26:00" "17:27:00" ...
 $ Global_active_power   : num  4.22 5.36 5.37 5.39 3.67 ...
 $ Global_reactive_power: num  0.418 0.436 0.498 0.502 0.528 0.522 0.52 0.52 0.51 0.51 ...
 $ Voltage               : num  235 234 233 234 236 ...
 $ Global_intensity      : num  18.4 23 23 23 15.8 15 15.8 15.8 15.8 15.8 ...
 $ Sub_metering_1        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Sub_metering_2        : num  1 1 2 1 1 2 1 1 1 2 ...
 $ Sub_metering_3        : num  17 16 17 17 17 17 17 17 17 16 ...
```

Menampilkan Beberapa Data Bagian Awal

head(power)

	Date	Time	Global_active_power	Global_reactive_power	Voltage	Global_intensity
1	16/12/2006	17:24:00	4.216	0.418	234.84	18.4
2	16/12/2006	17:25:00	5.360	0.436	233.63	23.0
3	16/12/2006	17:26:00	5.374	0.498	233.29	23.0
4	16/12/2006	17:27:00	5.388	0.502	233.74	23.0
5	16/12/2006	17:28:00	3.666	0.528	235.68	15.8
6	16/12/2006	17:29:00	3.520	0.522	235.02	15.0

	Sub_metering_1	Sub_metering_2	Sub_metering_3
1	0	1	17
2	0	1	16
3	0	2	17
4	0	1	17
5	0	1	17
6	0	2	17

Menampilkan Beberapa Data Bagian Akhir

tail(power)

	Date	Time	Global_active_power	Global_reactive_power	Voltage
2075254	26/11/2010	20:57:00	0.946	0	240.33
2075255	26/11/2010	20:58:00	0.946	0	240.43
2075256	26/11/2010	20:59:00	0.944	0	240.00
2075257	26/11/2010	21:00:00	0.938	0	239.82
2075258	26/11/2010	21:01:00	0.934	0	239.70
2075259	26/11/2010	21:02:00	0.932	0	239.55

	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3
2075254	4.0	0	0	0
2075255	4.0	0	0	0
2075256	4.0	0	0	0
2075257	3.8	0	0	0
2075258	3.8	0	0	0
2075259	3.8	0	0	0

Deskripsi Data Power

- Data pemakaian listrik rumah tangga selama 47 bulan, yang dikumpulkan oleh Electricite De France. Feature yang terdiri di data tersebut.
  1. Date format dd/mm/yy
  2. Time hh:mm:ss
  3. Global_active_power, rata-rata pemakaian aktif (kilowatts)
  4. Global_reactive_power, rata-rata pemakaian reactive (kilowatts)
  5. Voltage, rata-rata perubahan voltage (volts)
  6. Global_intensity, intesitas pemakaian listrik (ampere)
  7. Sub_metering_1
  8. Sub_metering_2
  9. Sub_metering_3

Melihat Struktur Data Frame Telco

```
'data.frame':   7043 obs. of  21 variables:
 $ customerID      : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963 2565 5536
 6512 6552 1003 4771 5605 4535 ...
 $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
 $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
 $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...
 $ MultipleLines   : Factor w/ 3 levels "No","No phone service",..: 2 1 1 2 1 3 3 2 3 1 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...
 $ OnlineSecurity  : Factor w/ 3 levels "No","No internet service",..: 1 3 3 3 1 1 1 3 1 3 .
..
 $ OnlineBackup    : Factor w/ 3 levels "No","No internet service",..: 3 1 3 1 1 1 3 1 1 3 .
..
 $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3 1 1 3 1 .
..
 $ TechSupport     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1 1 1 3 1 .
..
 $ StreamingTV     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 3 1 3 1 .
..
 $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 1 1 3 1 .
..
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1
...
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
```

Menampilkan Beberapa Data Bagian Awal

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
1	7590-VHVEG	Female	0	Yes	No	1	No
2	5575-GNVDE	Male	0	No	No	34	Yes
3	3668-QPYBK	Male	0	No	No	2	Yes
4	7795-CFOCW	Male	0	No	No	45	No
5	9237-HQITU	Female	0	No	No	2	Yes
6	9305-CDSKC	Female	0	No	No	8	Yes

	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection
1	No phone service	DSL	No	Yes	No
2	No	DSL	Yes	No	Yes
3	No	DSL	Yes	Yes	No
4	No phone service	DSL	Yes	No	Yes
5	No	Fiber optic	No	No	No
6	Yes	Fiber optic	No	No	Yes

	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling
1	No	No	No	Month-to-month	Yes
2	No	No	No	One year	No
3	No	No	No	Month-to-month	Yes
4	Yes	No	No	One year	No
5	No	No	No	Month-to-month	Yes
6	No	Yes	Yes	Month-to-month	Yes

	PaymentMethod	MonthlyCharges	TotalCharges	Churn
1	Electronic check	29.85	29.85	No
2	Mailed check	56.95	1889.50	No
3	Mailed check	53.85	108.15	Yes
4	Bank transfer (automatic)	42.30	1840.75	No
5	Electronic check	70.70	151.65	Yes
6	Electronic check	99.65	820.50	Yes

Menampilkan Beberapa Data Bagian Akhir

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
7038	2569-WGERO	Female	0	No	No	72	Yes
7039	6840-RESVB	Male	0	Yes	Yes	24	Yes
7040	2234-XADUH	Female	0	Yes	Yes	72	Yes
7041	4801-JZAZL	Female	0	Yes	Yes	11	No
7042	8361-LTMKD	Male	1	Yes	No	4	Yes
7043	3186-AJIEK	Male	0	No	No	66	Yes

	MultipleLines	InternetService	OnlineSecurity	OnlineBackup
7038	No	No	No internet service	No internet service
7039	Yes	DSL	Yes	No
7040	Yes	Fiber optic	No	Yes
7041	No phone service	DSL	Yes	No
7042	Yes	Fiber optic	No	No
7043	No	Fiber optic	Yes	No

	DeviceProtection	TechSupport	StreamingTV	StreamingMovies
7038	No internet service	No internet service	No internet service	No internet service
7039	Yes	Yes	Yes	Yes
7040	Yes	No	Yes	Yes
7041	No	No	No	No
7042	No	No	No	No
7043	Yes	Yes	Yes	Yes

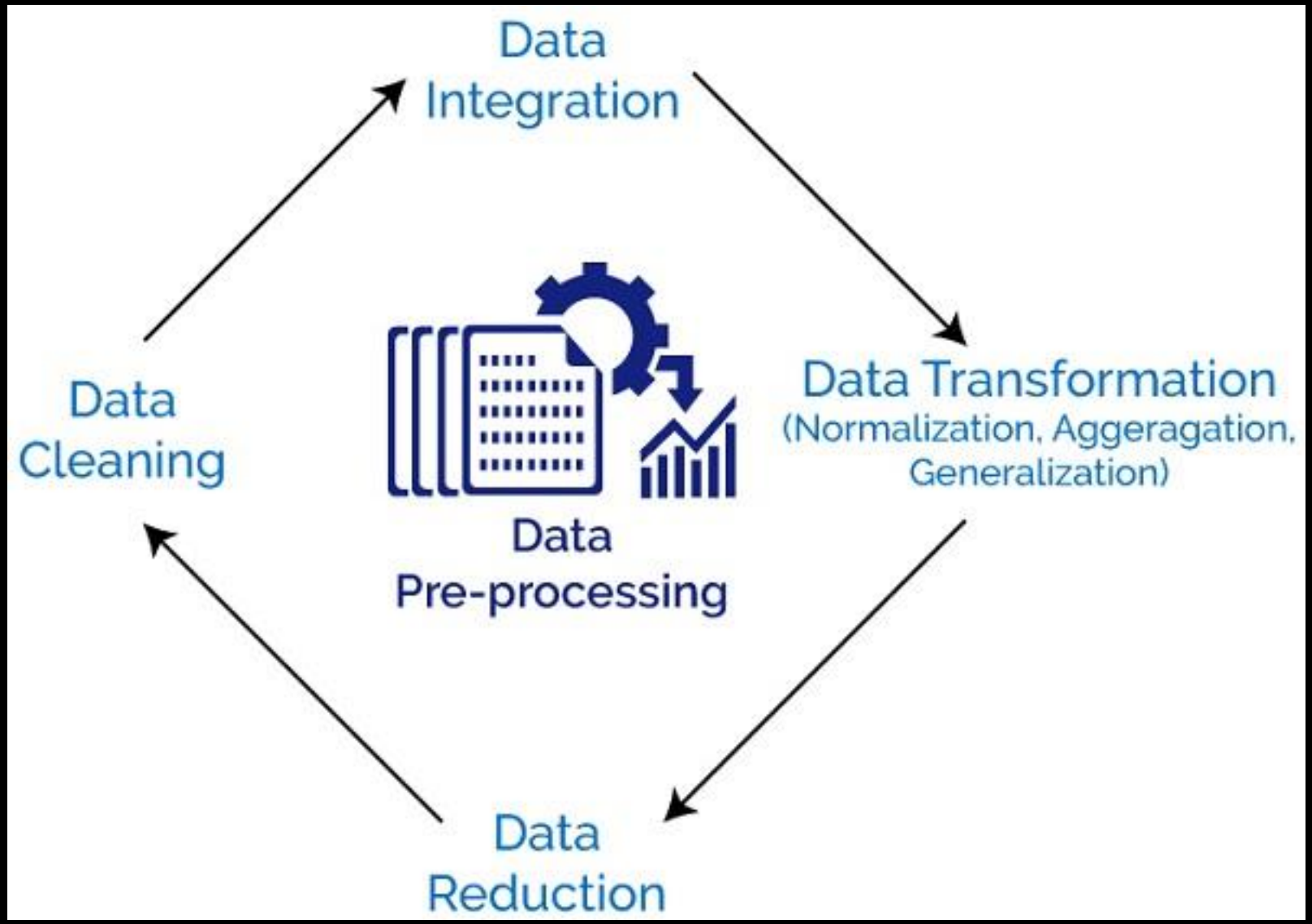
	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges
7038	Two year	Yes	Bank transfer (automatic)	21.15
7039	One year	Yes	Mailed check	84.80
7040	One year	Yes	Credit card (automatic)	103.20
7041	Month-to-month	Yes	Electronic check	29.60
7042	Month-to-month	Yes	Mailed check	74.40
7043	Two year	Yes	Bank transfer (automatic)	105.65

	TotalCharges	Churn
7038	1419.40	No
7039	1990.50	No
7040	7362.90	No
7041	346.45	No
7042	306.60	Yes
7043	6844.50	No

Deskripsi Data Telco

- The data set includes information about:
- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

Cleaning & Preparing Data



Periksa Type Data pada Data Frame

Data type you want	Function
Character / String	`as.character(OBJECT, ...)`
Factor / Category	`as.factor(OBJECT, ...)`
Numeric / Double	`as.numeric(OBJECT, ...)`
Integer	`as.integer(OBJECT, ...)`
Date	`as.Date(OBJECT, format="yyyy-mm-dd", ...)`
Datetime	`as.POSIXct(OBJECT, tz="CURRENT TIME ZONE", ...)`

Data Power

```
'data.frame':   2075259 obs. of  9 variables:
 $ Date                 : chr  "16/12/2006" "16/12/2006" "16/12/2006" "16/12/2006" ...
 $ Time                 : chr  "17:24:00" "17:25:00" "17:26:00" "17:27:00" ...
 $ Global_active_power  : num  4.22 5.36 5.37 5.39 3.67 ...
 $ Global_reactive_power: num  0.418 0.436 0.498 0.502 0.528 0.522 0.52 0.52 0.51 0.51 ...
 $ Voltage              : num  235 234 233 234 236 ...
 $ Global_intensity     : num  18.4 23 23 23 15.8 15 15.8 15.8 15.8 15.8 ...
 $ Sub_metering_1       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Sub_metering_2       : num  1 1 2 1 1 2 1 1 1 2 ...
 $ Sub_metering_3       : num  17 16 17 17 17 17 17 17 17 16 ...
```

Solusi Dirubah kedalam type data Date

Data Power

```
#Merubah char ke Date
power$Date = as.Date(power$Date,format="%d/%m/%Y")
#Menambah kolom DateTime
power$Datetime = as.POSIXct(paste(power$Date, power$Time))
#Menambah kolom Month
power$Month = format(power$Date,"%Y-%m")
#Tampilkan Strukturnya
str(power)
```

```
'data.frame':   2075259 obs. of  11 variables:
 $ Date                 : Date, format: "2006-12-16" "2006-12-16" ...
 $ Time                 : chr  "17:24:00" "17:25:00" "17:26:00" "17:27:00" ...
 $ Global_active_power  : num  4.22 5.36 5.37 5.39 3.67 ...
 $ Global_reactive_power: num  0.418 0.436 0.498 0.502 0.528 0.522 0.52 0.52 0.51 0.51 ...
 $ Voltage              : num  235 234 233 234 236 ...
 $ Global_intensity     : num  18.4 23 23 23 15.8 15 15.8 15.8 15.8 15.8 ...
 $ Sub_metering_1       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Sub_metering_2       : num  1 1 2 1 1 2 1 1 1 2 ...
 $ Sub_metering_3       : num  17 16 17 17 17 17 17 17 17 16 ...
 $ Datetime             : POSIXct, format: "2006-12-16 17:24:00" "2006-12-16 17:25:00" ...
 $ Month                : chr  "2006-12" "2006-12" "2006-12" "2006-12" ...
```

Data Telco

```
'data.frame':   7043 obs. of  21 variables:
 $ customerID      : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963 2565 5536
 6512 6552 1003 4771 5605 4535 ...
 $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
 $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
 $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...
 $ MultipleLines   : Factor w/ 3 levels "No","No phone service",..: 2 1 1 2 1 3 3 2 3 1 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...
 $ OnlineSecurity  : Factor w/ 3 levels "No","No internet service",..: 1 3 3 3 1 1 1 3 1 3 .
..
 $ OnlineBackup    : Factor w/ 3 levels "No","No internet service",..: 3 1 3 1 1 1 3 1 1 3 .
..
 $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3 1 1 3 1 .
..
 $ TechSupport     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1 1 1 3 1 .
..
 $ StreamingTV     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 3 1 3 1 .
..
 $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 1 1 3 1 .
..
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1
...
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
```

Data Telco

Seniorcitizen merupakan data kategorikal namun masih dianggap numeric, padahal nilai pada kolom tersebut adalah ‘0’ dan ‘1’, yang berarti 1 terdapat penduduk senior dan 0 tidak terdapat , sehingga perlu dirubah kedalam factor

#Solusi
telco$************ = as.factor(telco$************)
#Tampilkan struktur data baru

```
'data.frame':   7043 obs. of  21 variables:
 $ customerID      : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963 2565 5536
 6512 6552 1003 4771 5605 4535 ...
 $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
 $ SeniorCitizen   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

Missing Values

- In statistic[illegible]s, occur when no da[illegible] an observa[illegible]urrence and can ha[illegible]ons that can be dra[illegible]

F1	F2	F3	F4	F5	Class
Good	20	5	7	Old	Normal
Good	Missing	8	8	Old	Normal
Good	15	10	10	Old	Normal
Good	50	10	10	Old	Normal
Good	70	10	10	Old	Abnormal
Bad	20	5	7	Old	Abnormal
Good	20	5	80	Old	Abnormal
Good	85	100	100	Old	Abnormal
Good	20	100	Missing	Old	Abnormal
Good	24	6	8.4	Old	Normal
Good	12	9.6	9.6	Old	Normal
Good	18	12	12	Old	Normal
Good	60	12	12	Old	Normal
Good	84	Missing	12	Old	Abnormal
Bad	24	6	8.4	Old	Abnormal
Good	24	6	96	Old	Abnormal
Good	102	120	120	Old	Abnormal
Good	24	120	72	Old	Abnormal

TABLE 2 Comparison of Imputation Techniques for Missing Data

Imputation Method	*Advantages*	*Disadvantages*	*Best Used When:*
		Imputation Using Only Valid Data	
Complete Data	• Simplest to implement • Default for many statistical programs	• Most affected by nonrandom processes • Greatest reduction in sample size • Lowers statistical power	• Large sample size • Strong relationships among variables • Low levels of missing data
All Available Data	• Maximizes use of valid data • Results in largest sample size possible without replacing values	• Varying sample sizes for every imputation • Can generate "out of range" values for correlations and eigenvalues	• Relatively low levels of missing data • Moderate relationships among variables
		Imputation Using Known Replacement Values	
Case Substitution	• Provides realistic replacement values (i.e., another actual observation) rather than calculated values	• Must have additional cases not in the original sample • Must define similarity measure to identify replacement case	• Additional cases are available • Able to identify appropriate replacement cases
Hot and Cold Deck Imputation	• Replaces missing data with actual values from the most similar case or best known value	• Must define suitably similar cases or appropriate external values	• Established replacement values are known, or • Missing data process indicates variables upon which to base similarity
		Imputation by Calculating Replacement Values	
Mean Substitution	• Easily implemented • Provides all cases with complete information	• Reduces variance of the distribution • Distorts distribution of the data • Depresses observed correlations	• Relatively low levels of missing data • Relatively strong relationships among variables
Regression Imputation	• Employs actual relationships among the variables • Replacement values calculated based on an observation's own values on other variables • Unique set of predictors can be used for each variable with missing data	• Reinforces existing relationships and reduces generalizability • Must have sufficient relationships among variables to generate valid predicted values • Understates variance unless error term added to replacement value • Replacement values may be "out of range"	• Moderate to high levels of missing data • Relationships sufficiently established so as to not impact generalizability • Software availability
		Model-Based Methods for MAR Missing Data Processes	
Model-Based Methods	• Accommodates both nonrandom and random missing data processes • Best representation of original distribution of values with least bias	• Complex model specification by researcher • Requires specialized software • Typically not available directly in software programs (except EM method in SPSS)	• Only method that can accommodate nonrandom missing data processes • High levels of missing data require least biased method to ensure generalizability

Missing Values

```
#Buat Data Missing
nilai <- data.frame(45, 53, NA,76,91,82,NA,65)

#Hitung rata-ratanya
mean(nilai)
```

```
[1] NA
Warning message:
In mean.default(nilai) : argument is not numeric or logical: returning NA
```

```
#Solusi
mean(nilai, na.rm=TRUE)
```

Imputasi Dengan Median

```
#Buat Data Missing
nilai <- c(45,53,NA,76,91,82,NA,65)
#Hitung rata-ratanya
mean(nilai)
#Solusi
mean(nilai,na.rm=TRUE)
#impute missing value
#Hitung nilai untuk imputasi misal median
me = median(nilai,na.rm = TRUE)
#Ganti NA dengan nilai tertentu
nilai[is.na(nilai)] = 999
nilai
#Ganti 999 dengan median
nilai[nilai==999]=me
mean(nilai)
```

Missing Values Power

#Cara cepat untuk melihat apakah ada missing value di data tersebut

#Menggunakan summary, yang berarti melihat deskripsi keseluruhan data tersebut

summary(power)

```
     Date                Time           Global_active_power Global_reactive_power
Min.   :2006-12-16   Length:2075259     Min.   : 0.076      Min.   :0.000
1st Qu.:2007-12-12   Class :character   1st Qu.: 0.308      1st Qu.:0.048
Median :2008-12-06   Mode  :character   Median : 0.602      Median :0.100
Mean   :2008-12-05                      Mean   : 1.092      Mean   :0.124
3rd Qu.:2009-12-01                      3rd Qu.: 1.528      3rd Qu.:0.194
Max.   :2010-11-26                      Max.   :11.122      Max.   :1.390
                                        NA's   :25979       NA's   :25979
   Voltage      Global_intensity Sub_metering_1   Sub_metering_2   Sub_metering_3
Min.   :223.2   Min.   : 0.200   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
1st Qu.:239.0   1st Qu.: 1.400   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
Median :241.0   Median : 2.600   Median : 0.000   Median : 0.000   Median : 1.000
Mean   :240.8   Mean   : 4.628   Mean   : 1.122   Mean   : 1.299   Mean   : 6.458
3rd Qu.:242.9   3rd Qu.: 6.400   3rd Qu.: 0.000   3rd Qu.: 1.000   3rd Qu.:17.000
Max.   :254.2   Max.   :48.400   Max.   :88.000   Max.   :80.000   Max.   :31.000
NA's   :25979   NA's   :25979    NA's   :25979    NA's   :25979    NA's   :25979
   Datetime                         Month
Min.   :2006-12-16 17:24:00   Length:2075259
1st Qu.:2007-12-12 00:18:30   Class :character
Median :2008-12-06 07:13:00   Mode  :character
Mean   :2008-12-06 07:13:00
3rd Qu.:2009-12-01 14:07:30
Max.   :2010-11-26 21:02:00
```

Imputasi Missing Values Global_Active_Power

```
#Lihat Nilai Maxnya
max(power$Global_active_power,na.rm = T)
#Hitung Nilai Median
medianpower = median(power$Global_active_power,na.rm=T)
#Rubah Nilai NA kedalam nilai diatas Max bebas
power$Global_active_power[is.na(power$Global_active_power)] = 999
#Rubah Nilai bebas tadi menjadi nilai median
power$Global_active_power[power$Global_active_power==999] =
medianpower
```

Imputasi Missing Values Global_reactive_Power

```
#Reactive Power
#Lihat Nilai Maxnya Global reactive power
max(power$Global_reactive_power,na.rm = T)
#Hitung Nilai Median
medianrepower = median(power$Global_reactive_power,na.rm=T)
#Rubah Nilai NA kedalam nilai diatas Max bebas
power$Global_reactive_power[is.na(power$Global_reactive_power)] =
999
#Rubah Nilai bebas tadi menjadi nilai median
power$Global_reactive_power[power$Global_reactive_power==999] =
medianrepower
```

Imputasi Missing Values Global_voltage

```
#Voltage
#Lihat Nilai Maxnya
max(power$Voltage,na.rm = T)
#Hitung Nilai Median
medianvoltage = median(power$Voltage,na.rm=T)
#Rubah Nilai NA kedalam nilai diatas Max bebas
power$Voltage[is.na(power$Voltage)] = 999
#Rubah Nilai bebas tadi menjadi nilai median
power$Voltage[power$Voltage==999] = medianvoltage
```

Imputasi Missing Values Global_Intensity

```
#Intensity
#Lihat Nilai Maxnya
***(power$*************,na.rm = T)
#Hitung Nilai Median
medianintensity = ******(power$Global_intensity,na.rm=T)
#Rubah Nilai NA kedalam nilai diatas Max bebas
power$Global_intensity[is.na(power$Global_intensity)] = 999
#Rubah Nilai bebas tadi menjadi nilai median
power$Global_intensity[power$Global_intensity==999] = medianintensity
```

Imputasi Missing Values Sub_metering_1

```
#Sub_Metering_1
#Lihat Nilai Maxnya
max(power$Sub_metering_1,na.rm = T)
#Hitung Nilai Mode
modemetering1 = mode(power$Sub_metering_1)
#Rubah Nilai NA kedalam nilai diatas Max bebas
power$Sub_metering_1[is.na(power$Sub_metering_1)] = 999
#Rubah Nilai bebas tadi menjadi nilai median
power$Sub_metering_1[power$Sub_metering_1==999] = modemetering1
```

Imputasi Missing Values Sub_metering_2

```
#Sub_Metering_2
#Lihat Nilai Maxnya
max(power$Sub_metering_2,na.rm = T)
#Hitung Nilai Mode
modemetering2 = mode(power$Sub_metering_2)
#Rubah Nilai NA kedalam nilai diatas Max bebas
power$Sub_metering_2[is.na(power$Sub_metering_2)] = 999
#Rubah Nilai bebas tadi menjadi nilai median
power$Sub_metering_2[power$Sub_metering_2==999] = modemetering2
```

Imputasi Missing Values Sub_metering_3

```
#Sub_Metering_3
#Lihat Nilai Maxnya
max(power$Sub_metering_3,na.rm = T)
#Hitung Nilai Mode
modemetering3 = mode(power$Sub_metering_3)
#Rubah Nilai NA kedalam nilai diatas Max bebas
power$Sub_metering_3[is.na(power$Sub_metering_3)] = 999
#Rubah Nilai bebas tadi menjadi nilai mode
power$Sub_metering_3[power$Sub_metering_3==999] = modemetering3
```

Imputasi Missing Values Power

#cek kembali dengan summary
summary(power)

```
     Date                Time           Global_active_power Global_reactive_power
Min.   :2006-12-16   Length:2075259     Min.   : 0.076      Min.   :0.000
1st Qu.:2007-12-12   Class :character   1st Qu.: 0.310      1st Qu.:0.048
Median :2008-12-06   Mode  :character   Median : 0.602      Median :0.100
Mean   :2008-12-05                      Mean   : 1.085      Mean   :0.124
3rd Qu.:2009-12-01                      3rd Qu.: 1.520      3rd Qu.:0.194
Max.   :2010-11-26                      Max.   :11.122      Max.   :1.390
                                                            NA's   :25979
   Voltage      Global_intensity Sub_metering_1   Sub_metering_2   Sub_metering_3
Min.   :223.2   Min.   : 0.200   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
1st Qu.:239.0   1st Qu.: 1.400   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
Median :241.0   Median : 2.600   Median : 0.000   Median : 0.000   Median : 1.000
Mean   :240.8   Mean   : 4.628   Mean   : 1.122   Mean   : 1.299   Mean   : 6.458
3rd Qu.:242.9   3rd Qu.: 6.400   3rd Qu.: 0.000   3rd Qu.: 1.000   3rd Qu.:17.000
Max.   :254.2   Max.   :48.400   Max.   :88.000   Max.   :80.000   Max.   :31.000
NA's   :25979   NA's   :25979    NA's   :25979    NA's   :25979    NA's   :25979
   Datetime                       Month
Min.   :2006-12-16 17:24:00   Length:2075259
1st Qu.:2007-12-12 00:18:30   Class :character
Median :2008-12-06 07:13:00   Mode  :character
Mean   :2008-12-06 07:13:00
3rd Qu.:2009-12-01 14:07:30
Max.   :2010-11-26 21:02:00
```

Missing Values Telco

summary(is.na(telco))

```
customerID        gender          SeniorCitizen    Partner         Dependents
Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
FALSE:7043      FALSE:7043      FALSE:7043      FALSE:7043      FALSE:7043

  tenure        PhoneService    MultipleLines   InternetService OnlineSecurity
Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
FALSE:7043      FALSE:7043      FALSE:7043      FALSE:7043      FALSE:7043

OnlineBackup    DeviceProtection TechSupport    StreamingTV     StreamingMovies
Mode :logical   Mode :logical    Mode :logical  Mode :logical   Mode :logical
FALSE:7043      FALSE:7043       FALSE:7043     FALSE:7043      FALSE:7043

 Contract       PaperlessBilling PaymentMethod  MonthlyCharges  TotalCharges
Mode :logical   Mode :logical    Mode :logical  Mode :logical   Mode :logical
FALSE:7043      FALSE:7043       FALSE:7043     FALSE:7043      FALSE:7032
                                                                TRUE :11

  Churn
Mode :logical
FALSE:7043
```

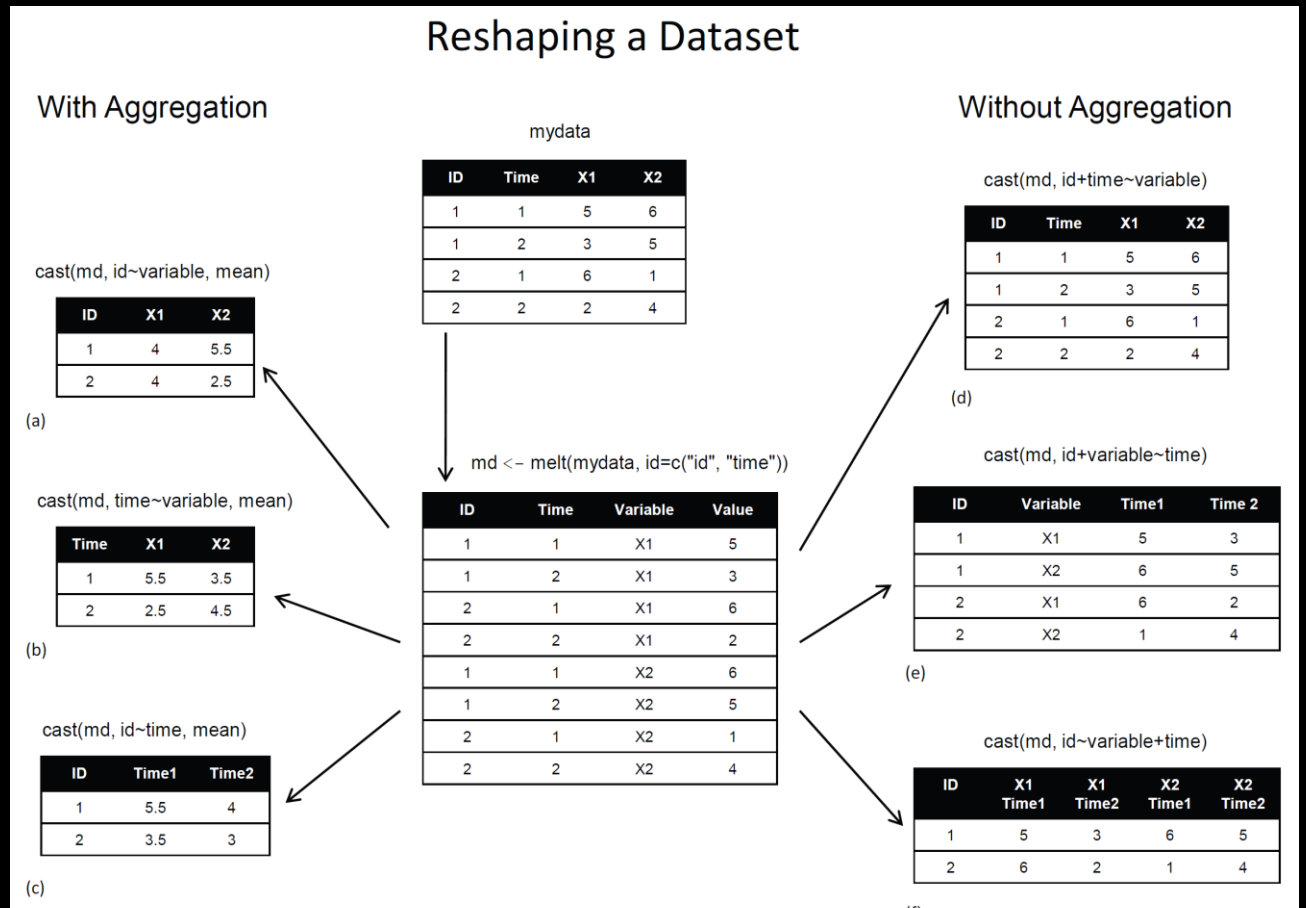
Impute Missing Values Telco

```
#Lihat Nilai Maxnya
max(*****$TotalCharges,na.rm = TRUE)
#Hitung Nilai Median
mediantelco = median(telco$**********,na.rm=TRUE)
#Rubah Nilai NA kedalam nilai diatas Max bebas
telco$********** [is.na(*****$TotalCharges)] = 99999
#Rubah Nilai bebas tadi menjadi nilai median
telco$TotalCharges[telco$TotalCharges == 99999] = mediantelco
#cek kembali
summary(is.na(telco))
```

```
customerID       gender          SeniorCitizen    Partner         Dependents
Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
FALSE:7043      FALSE:7043      FALSE:7043      FALSE:7043      FALSE:7043
  tenure        PhoneService    MultipleLines   InternetService OnlineSecurity
Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
FALSE:7043      FALSE:7043      FALSE:7043      FALSE:7043      FALSE:7043
OnlineBackup    DeviceProtection TechSupport    StreamingTV     StreamingMovies
Mode :logical   Mode :logical    Mode :logical  Mode :logical   Mode :logical
FALSE:7043      FALSE:7043       FALSE:7043     FALSE:7043      FALSE:7043
 Contract       PaperlessBilling PaymentMethod  MonthlyCharges  TotalCharges
Mode :logical   Mode :logical    Mode :logical  Mode :logical   Mode :logical
FALSE:7043      FALSE:7043       FALSE:7043     FALSE:7043      FALSE:7043
  Churn
Mode :logical
FALSE:7043
```

Data Reshaping

Merubah bentuk data untuk mempermudah analisa, seperti melakukan aggregate, membentuk rata-rata, melakukan transpose dan sejenisnya



Reshaping Data Telco

```
#Membentuk grup berdasarkan Bulan
power_group = group_by(power,Month)
#tampilkan head data
head(power_group)
```

```
# A tibble: 6 x 11
# Groups:   Month [1]
  Date       Time     Global_active_power Global_reactive_power Voltage Global_intensity
  <date>     <chr>                  <dbl>                 <dbl>   <dbl>            <dbl>
1 2006-12-16 17:24:00                4.22                 0.418    235.             18.4
2 2006-12-16 17:25:00                5.36                 0.436    234.             23
3 2006-12-16 17:26:00                5.37                 0.498    233.             23
4 2006-12-16 17:27:00                5.39                 0.502    234.             23
5 2006-12-16 17:28:00                3.67                 0.528    236.             15.8
6 2006-12-16 17:29:00                3.52                 0.522    235.             15
# ... with 5 more variables: Sub_metering_1 <dbl>, Sub_metering_2 <dbl>,
#   Sub_metering_3 <dbl>, Datetime <dttm>, Month <chr>
```

Data Reshaping Power

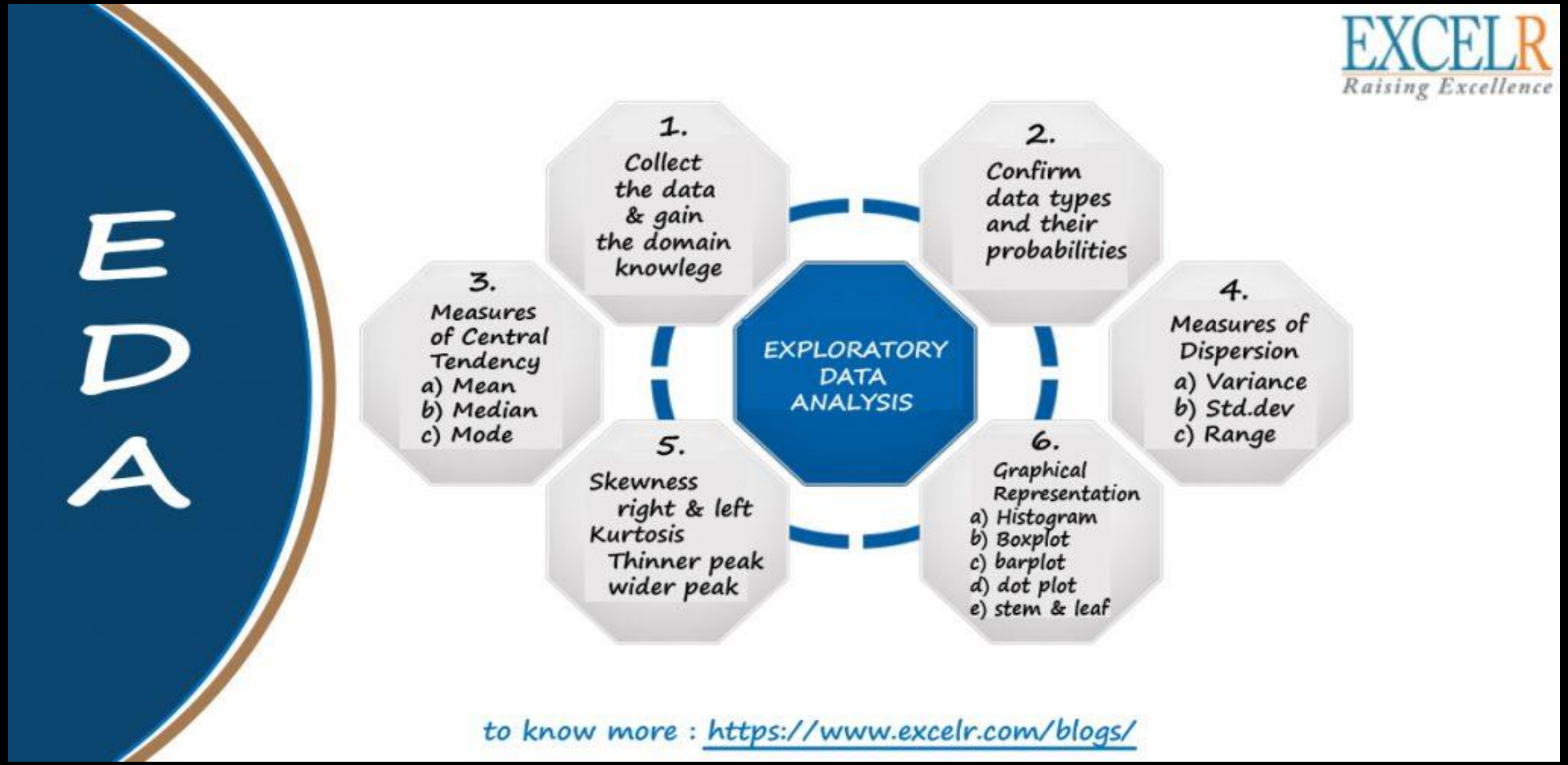
```
#Membentuk grup berdasarkan Bulan
power_group = group_by(power,Month)
#Membentuk data perbulan berisi maksimum pemakaian dan total pemakaian
power_monthly = summarize((power_group),Max_Demand_kW =
max(Global_active_power),
                          Total_use_kWh = sum(Global_active_power)/60)
#menghapus partial month dari data frame
power_monthly = power_monthly[2:47,]
#mengkonversi month ke data
power_monthly$Month = as.Date(paste0(power_monthly$Month,"-01"))
```

Data Reshaping Power

```
#cek data yang sudah direshape
head(power_monthly)
```

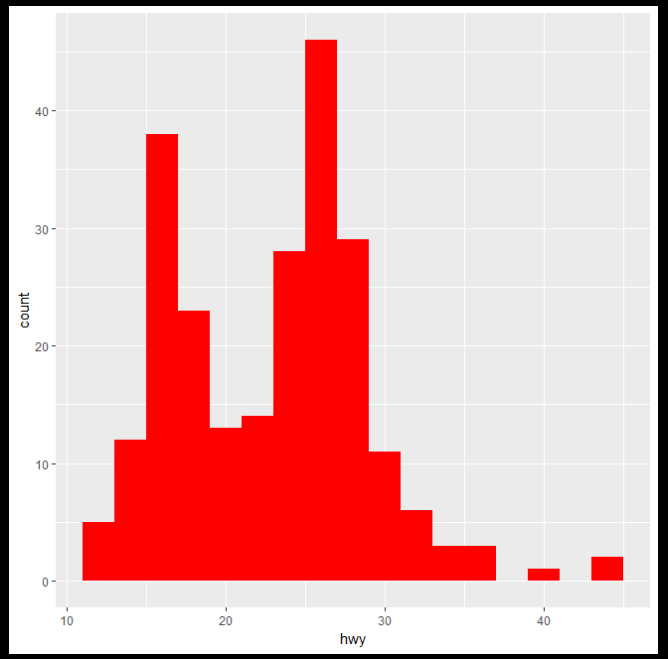
```
# A tibble: 6 x 3
  Month      Max_Demand_kW Total_use_kWh
  <date>             <dbl>         <dbl>
1 2007-01-01          9.27         1150.
2 2007-02-01          9.41          942.
3 2007-03-01         10.7           981.
4 2007-04-01          8.16          624.
5 2007-05-01          7.67          733.
6 2007-06-01          7.61          595.
```

Data Visualization

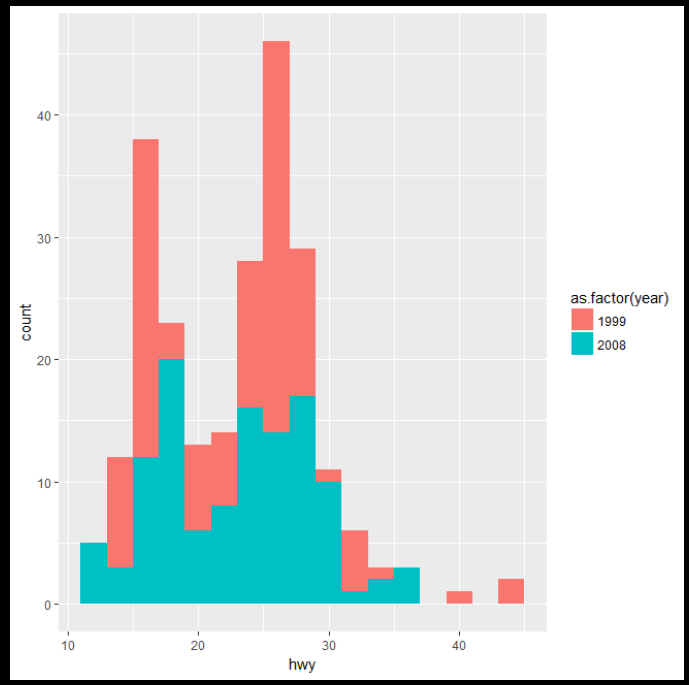


Histogram

```
ggplot(mpg,aes(hwy))+geom_
histogram (binwidth
=2,fill="red")
```

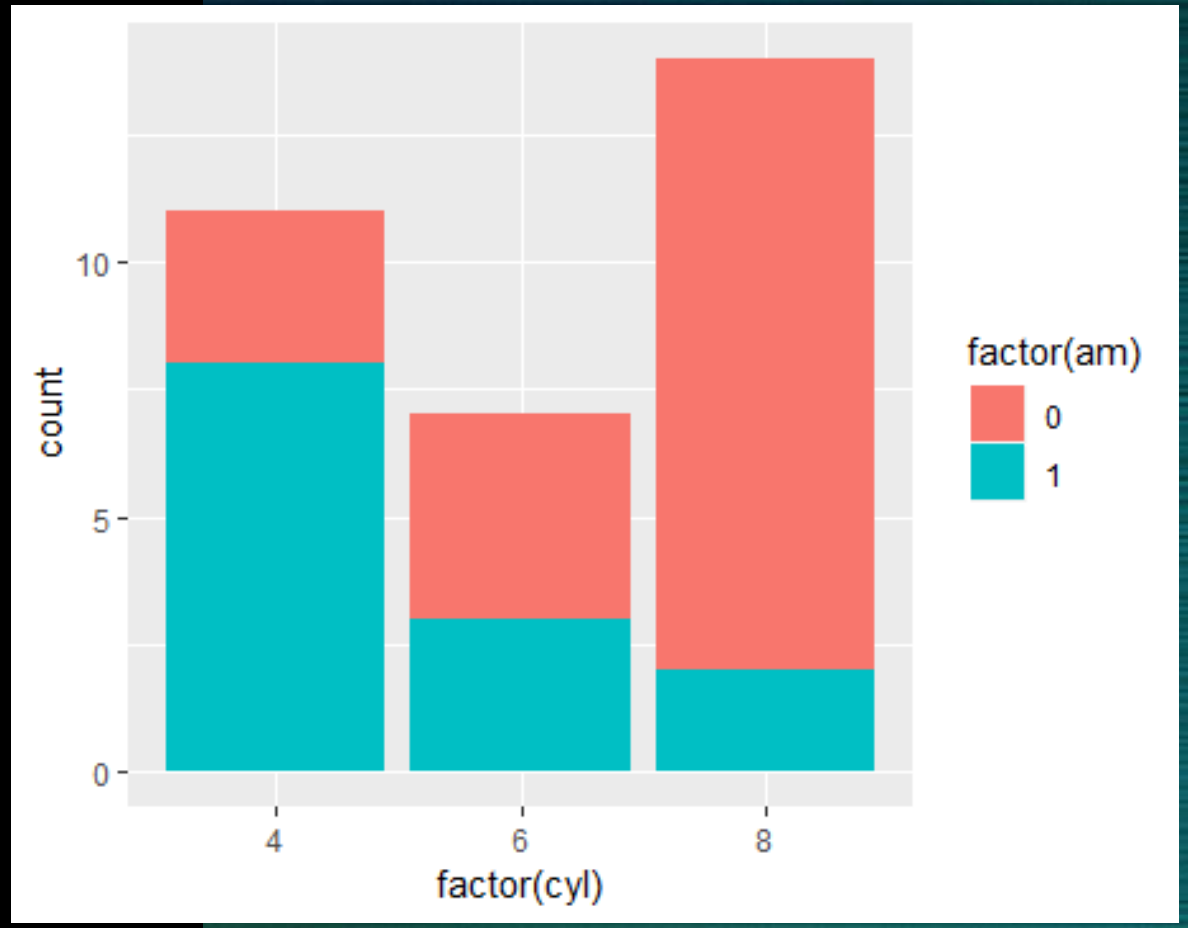



```
ggplot(mpg,aes(hwy,fill=as.factor
(year)))+geom_histogram(binwidth
=2)
```



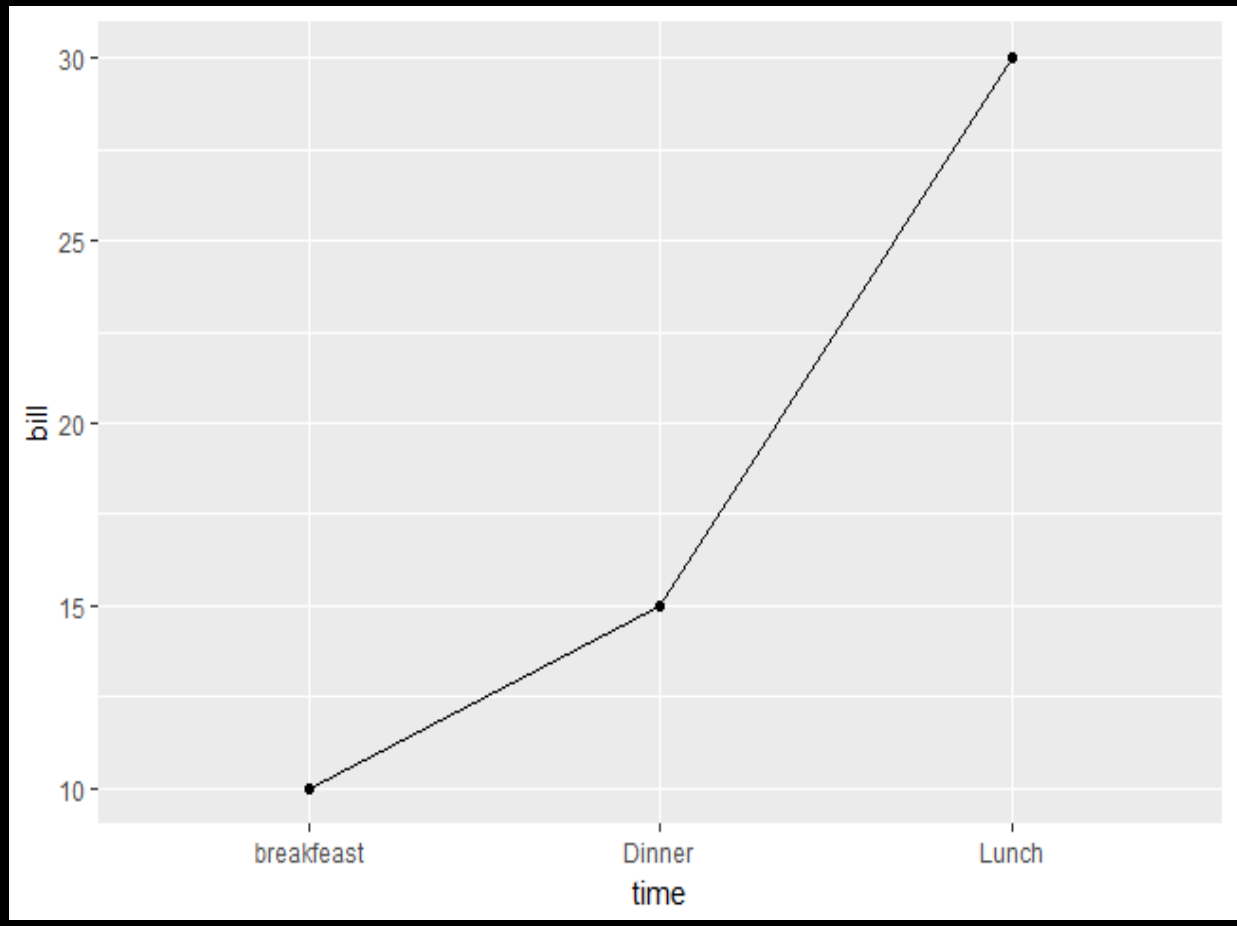
Barplot

```
ggplot(mtcars,aes(x=factor(cyl),fill =
factor(am)))+geom_bar()
```



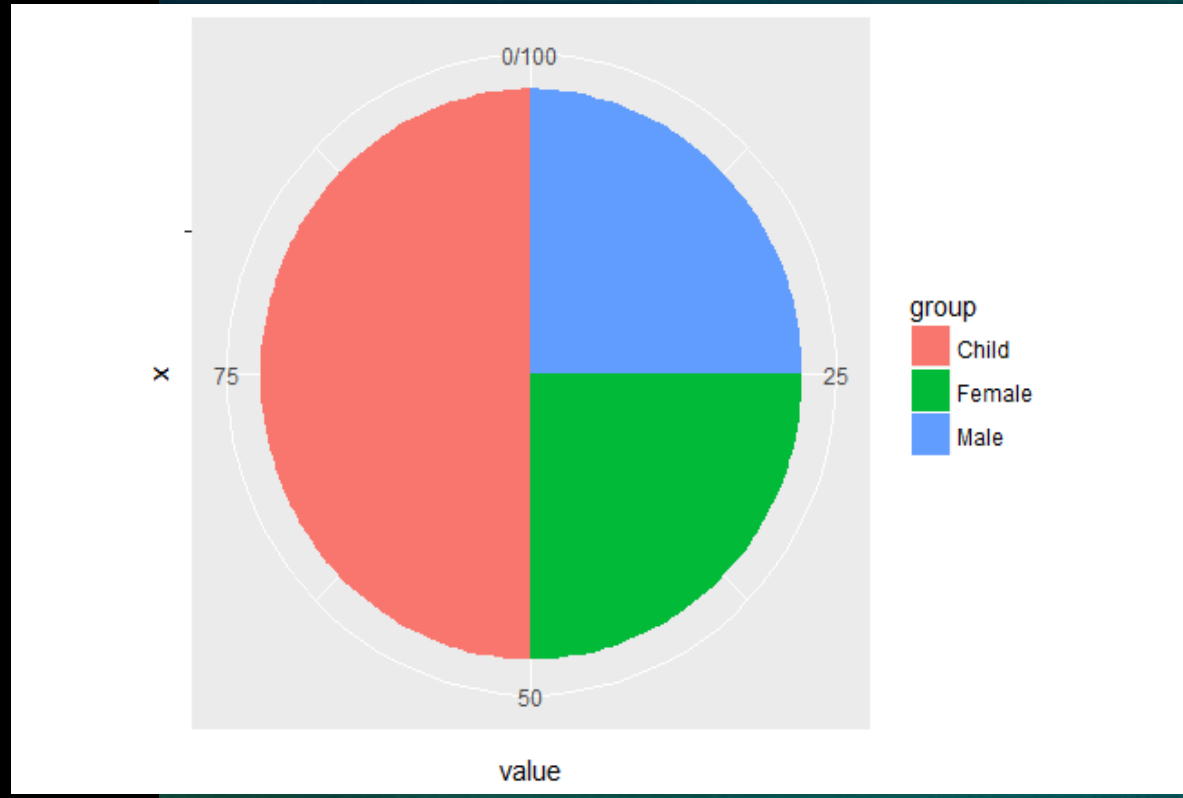
Line Plots

```
df = data.frame(time=c("breakfeast",
"Lunch", "Dinner"),
                bill=c(10, 30, 15))
ggplot(data=df, aes(x=time, y=bill,
group=1)) + geom_line()+ geom_point()
```



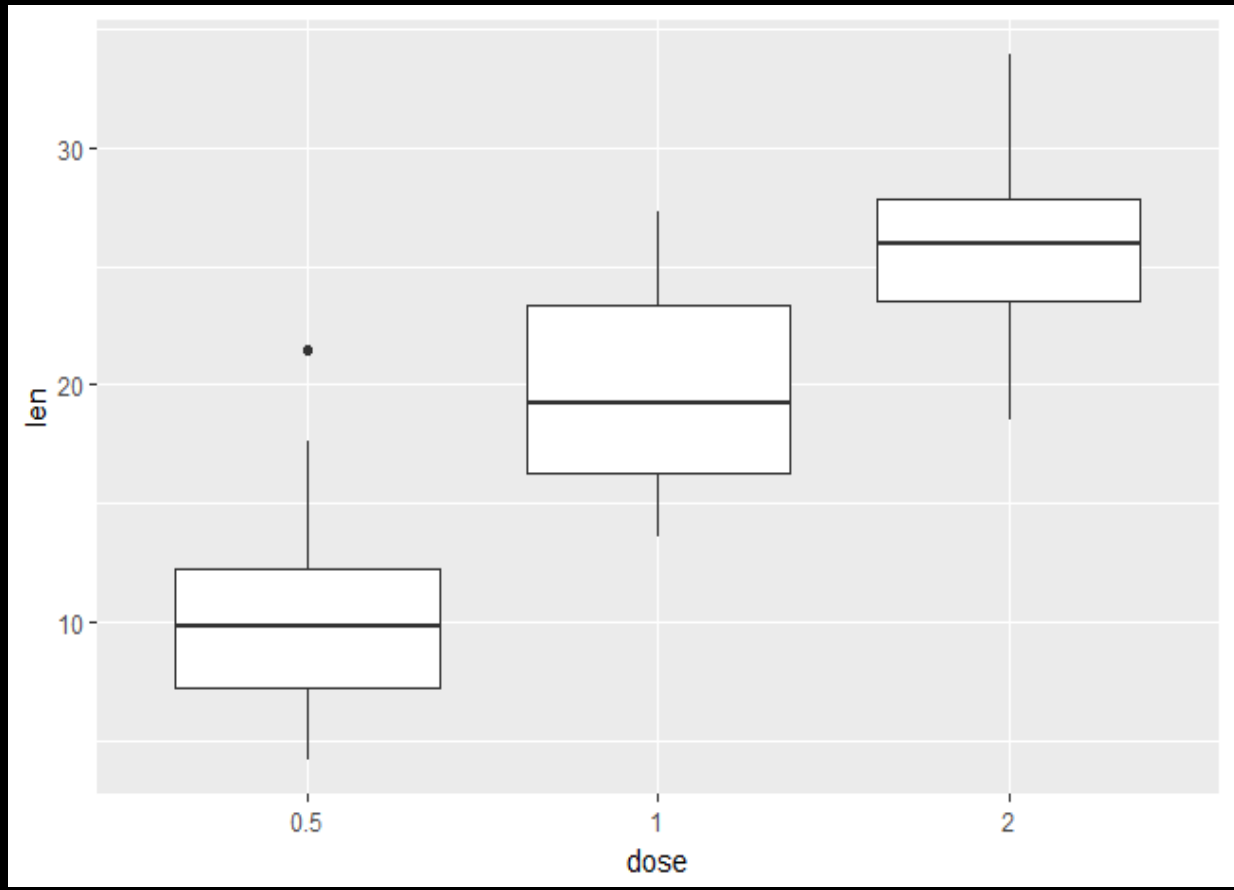
Pie Chart

```
#pie chart dengan data sendiri
df = data.frame(group = c("Male", "Female",
"Child"),value = c(25, 25, 50))
#Buat barplot definisikan sebagai misal bp
bp = ggplot(df, aes(x="", y=value, fill=group))+
geom_bar(width = 1, stat = "identity")
#buat piechartnya
bp + coord_polar("y", start=0)
```



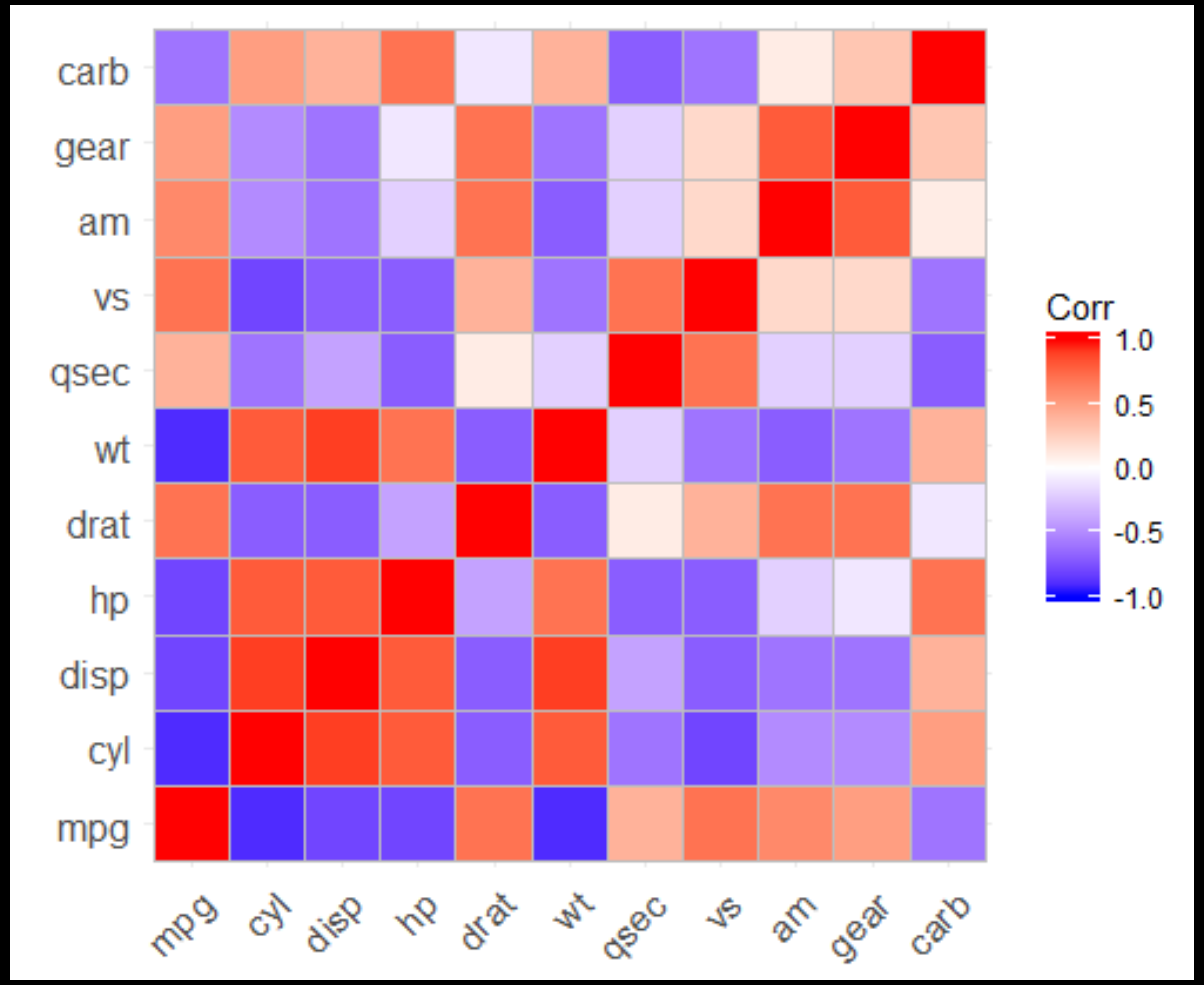
Box Plots

```
ToothGrowth$dose =
as.factor(ToothGrowth$dose)
ggplot(ToothGrowth, aes(x = dose,y = len)) +
geom_boxplot()
```

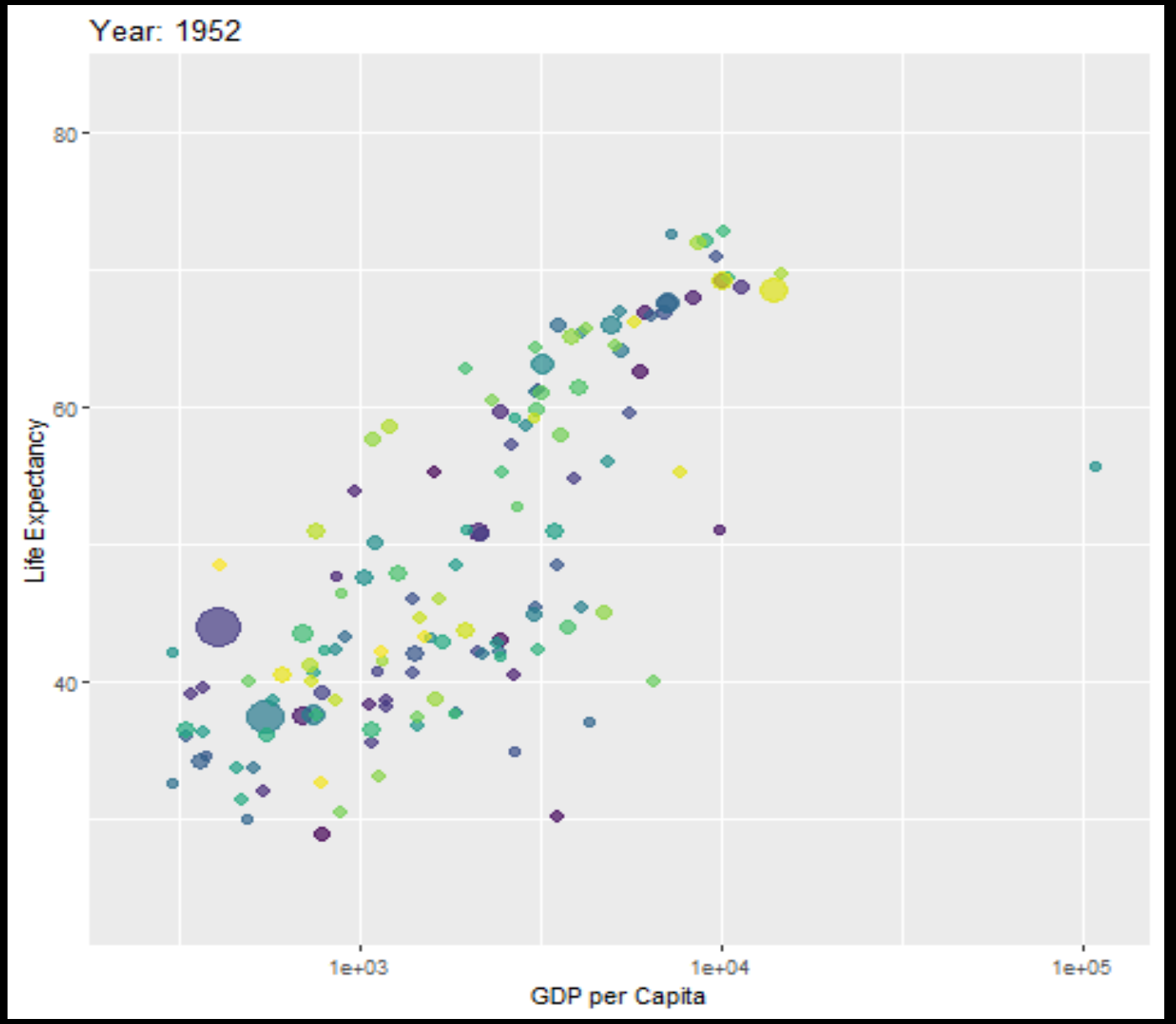


Plot Korelasi

```
#Plot korelasi dengan data mtcars
#menghitung matrix korelasi
corr = round(cor(mtcars),1)
#membuat plot korelasi
ggcorrplot(corr)
```

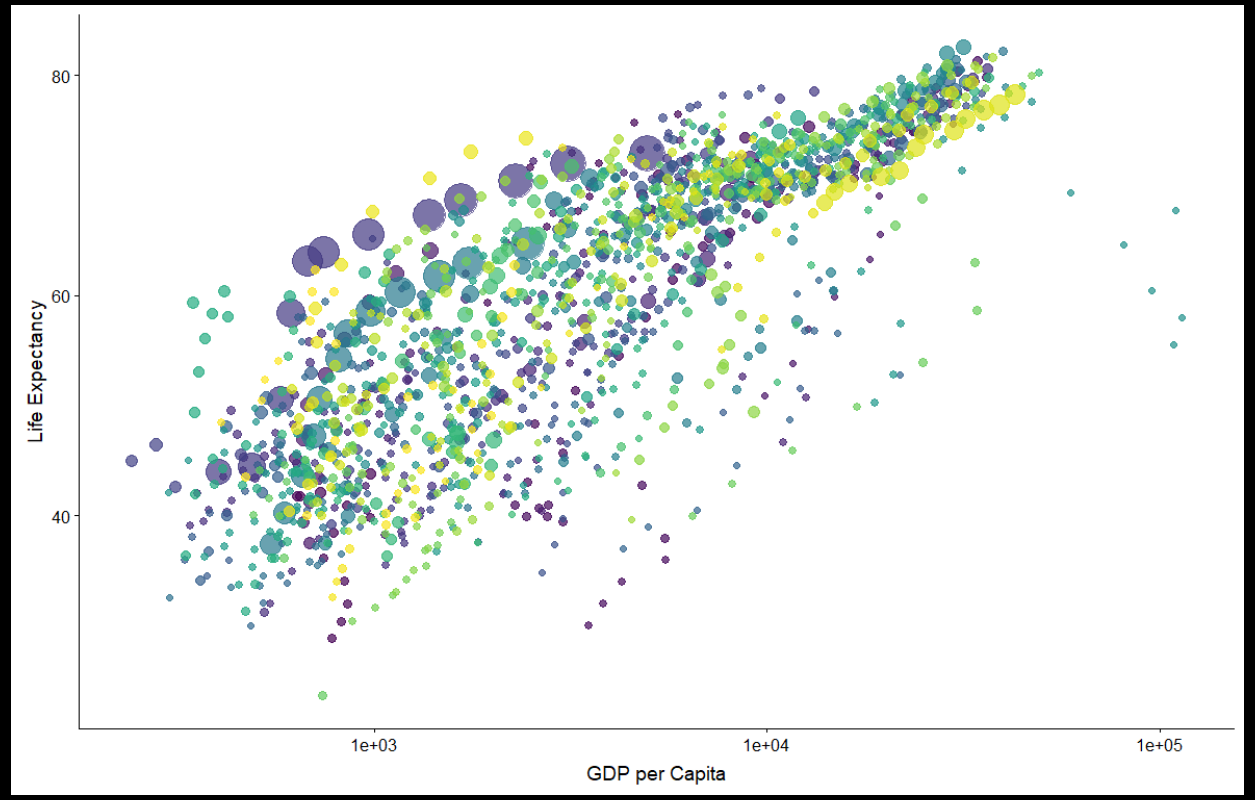


Plot Bergerak



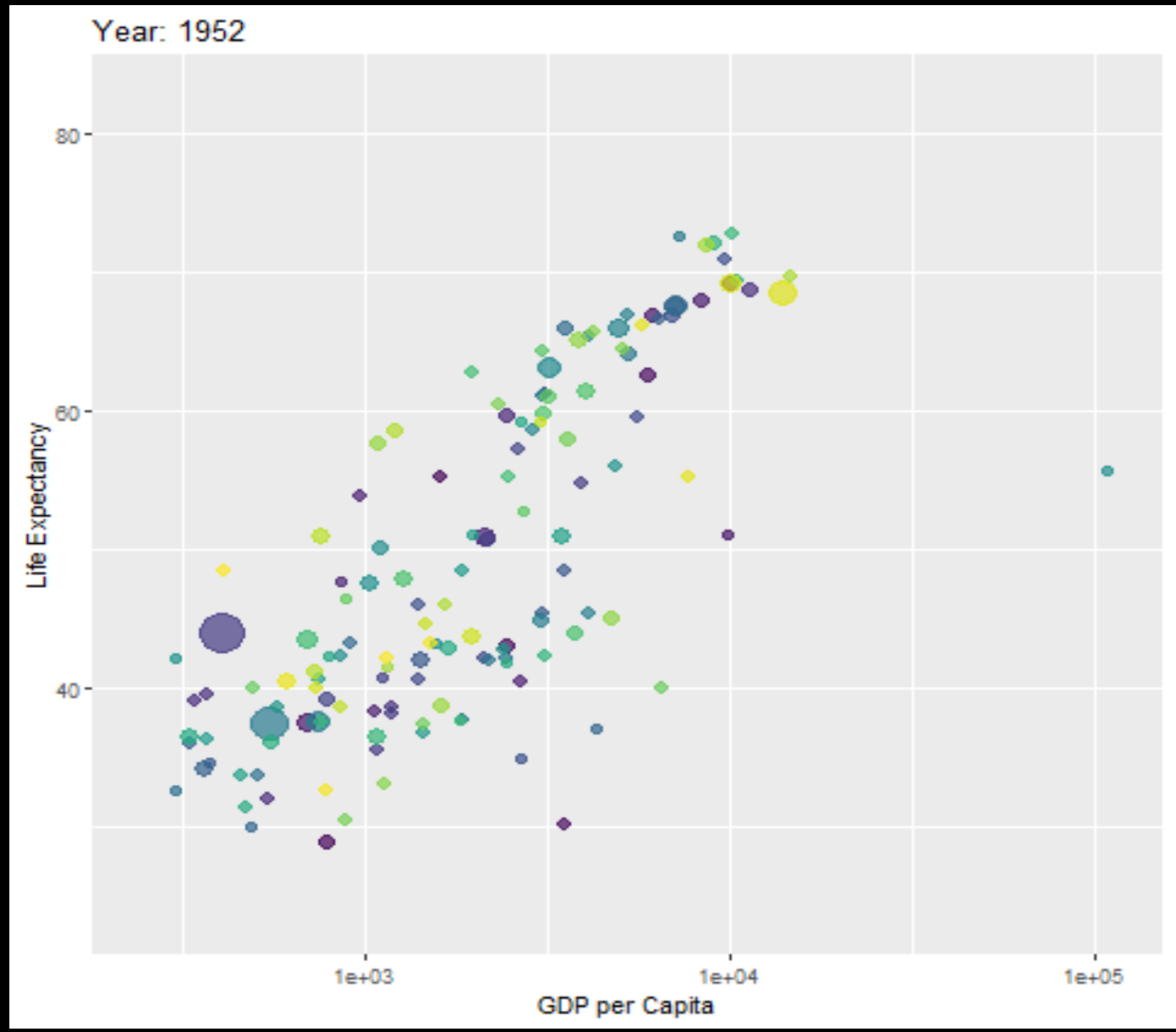
Plot Bergerak

```
#Pertama buat plot biasa definisikan misal
p
p <- ggplot(gapminder,

aes(x=gdpPercap,y=lifeExp,size=pop,colou
r=country)) +
  geom_point(show.legend=F,alpha=0.7)+
  scale_color_viridis_d()+
  scale_size(range=c(2,12)) +
  scale_x_log10()+
  labs(x="GDP per Capita",y="Life
Expectancy")
#Tampilkan P
p
```



Plot Bergerak

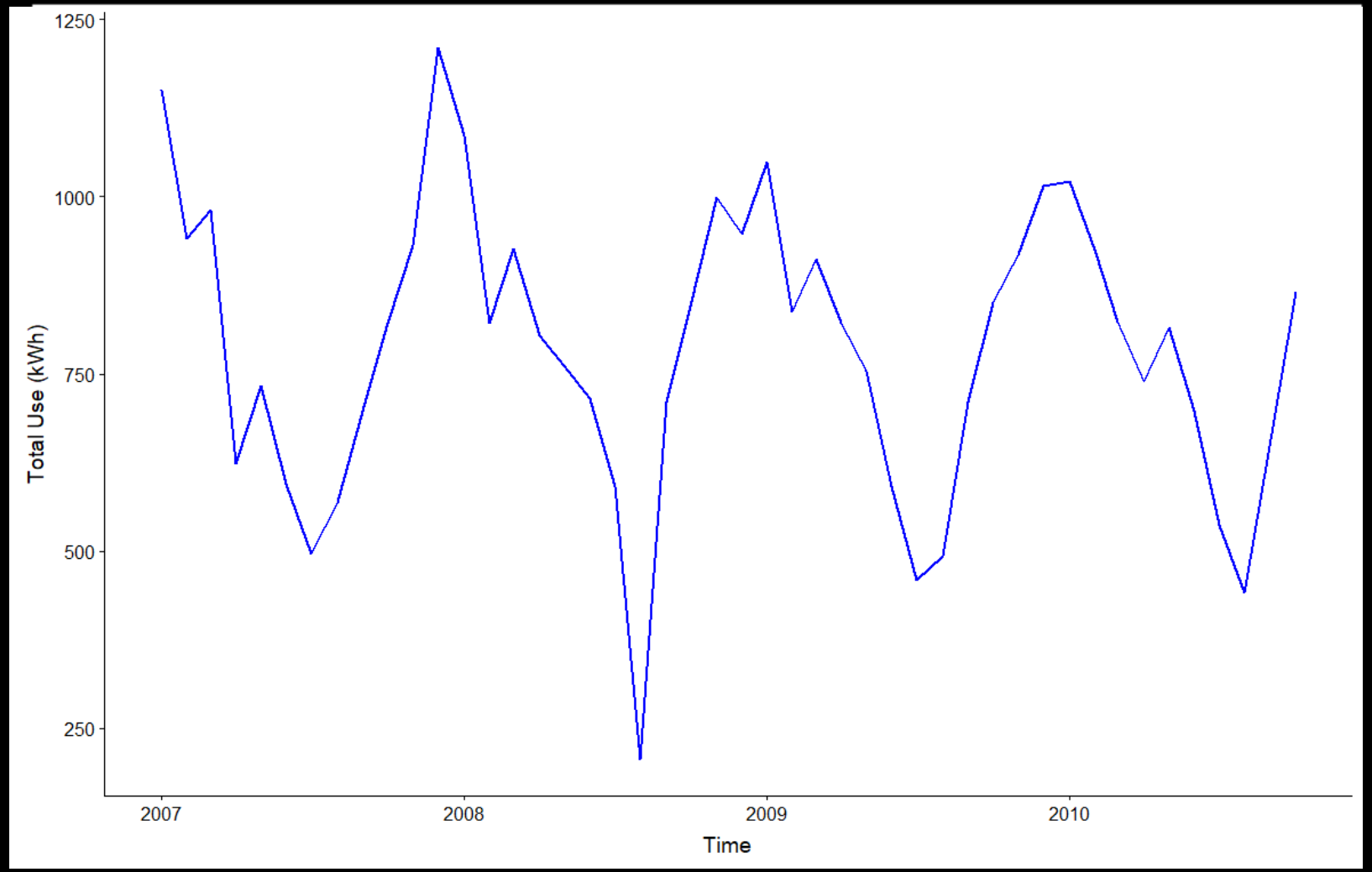
```
#Buat p bergerak
p + transition_time(year) +
  labs(title="Year: {frame_time}")
```



Data Visualisasi Untuk Power

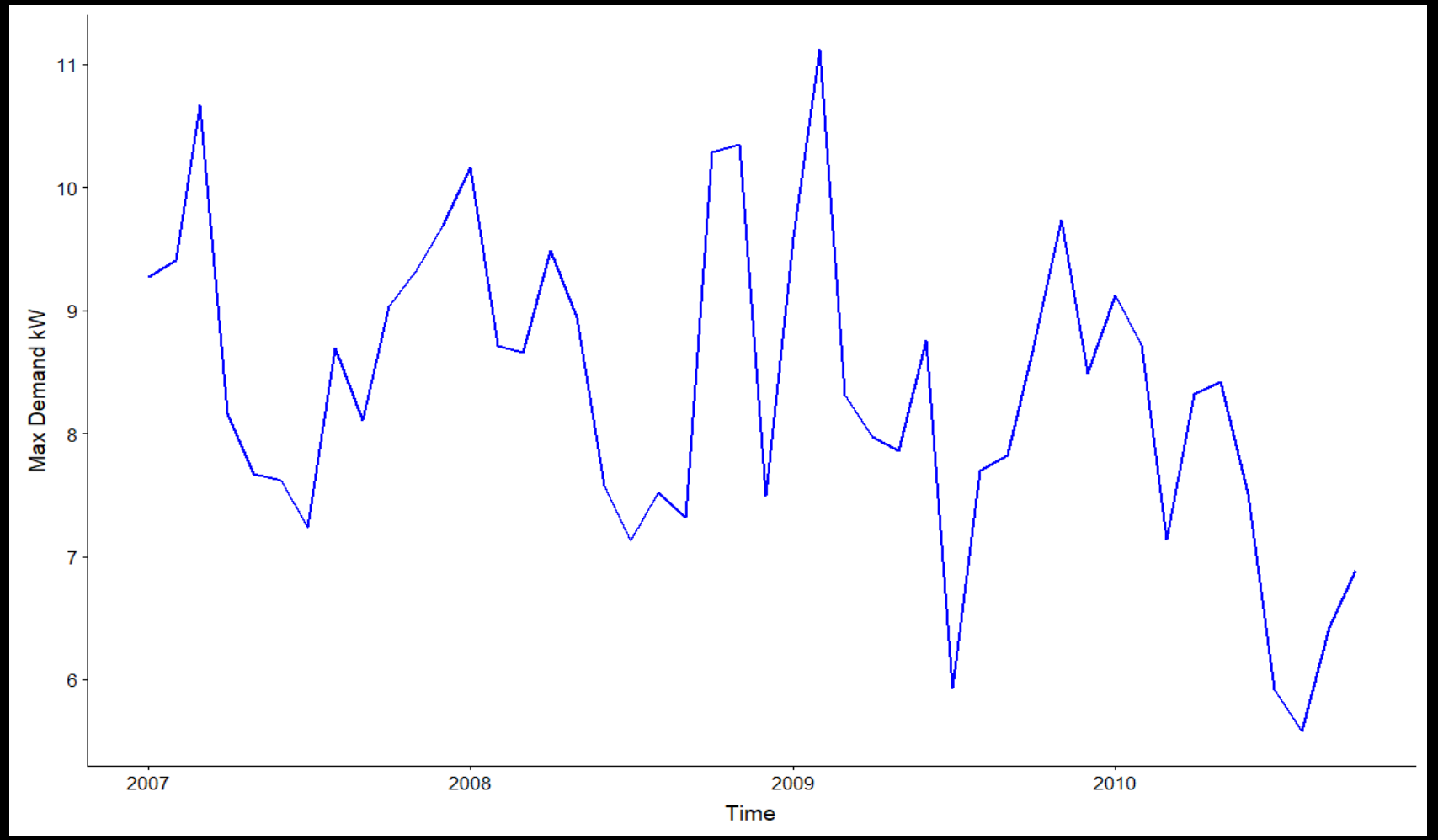
#Time Series Plot Total Use kWh

```
ggplot(power_monthly,aes(Month, Total_use_kWh)) +
  geom_line(col="blue",lwd=1) + labs(y="Total Use (kWh)",x="Time")
```



Data Visualisasi Untuk Power

```
#Time Series Plot Max Demand
ggplot(power_monthly,aes(Month, **********)) +
  geom_line(col="blue",lwd=1) + labs(y="Max Demand kW",x="Time")
```



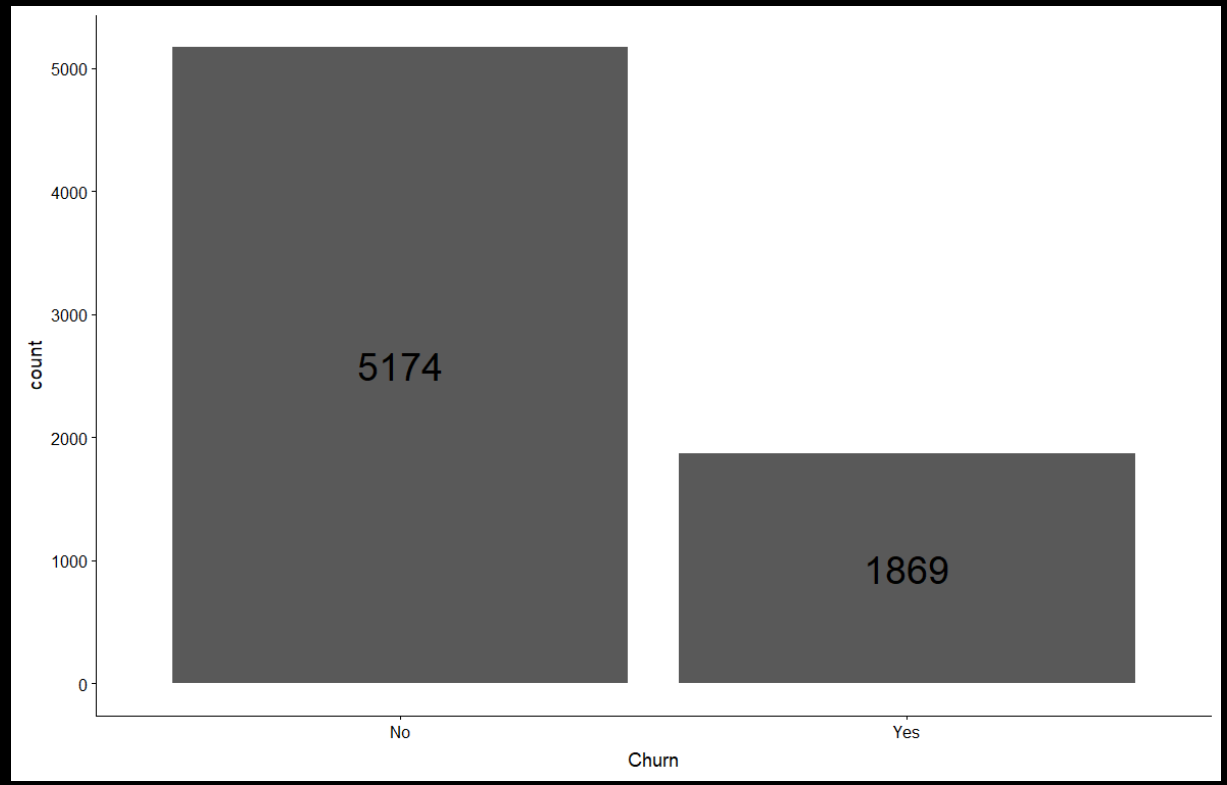
Kesimpulan Berdasarkan Visualisasi Power

1. Data pemakaian listrik penduduk perancis selama tahun 2007 hingga 2010 mempunyai pola musiman, pola musiman yang terbentuk adalah selalu rendah di pertengahan tahun dan selalu tinggi di awal tahun. Hal tersebut mungkin desebabkan adanya intervensi libur musim panas sehingga menyebabkan penduduk meninggalkan rumah dan mengurangi pemakaian listrik. Sedangkan tinggi diawal tahun intervensi karena natal dan tahun baru.
2. Maksimal Pemakaian listrik sedikit fluktuatif, namun selalu tinggi diakhir hingga awal tahun. Dugaan awal adalah karena intervensi dari perayaan natal dan tahun baru.

Data Visualisasi Untuk Telco

#1.Tampilkan jumlah pelanggan yang beralih dan tidak

```
ggplot(telco, aes(x = Churn)) + geom_bar() +
  geom_text(aes(label=..count..),stat="count",position = position_stack(0.5),size=10)
```



#2.Tampilkan proporsinya

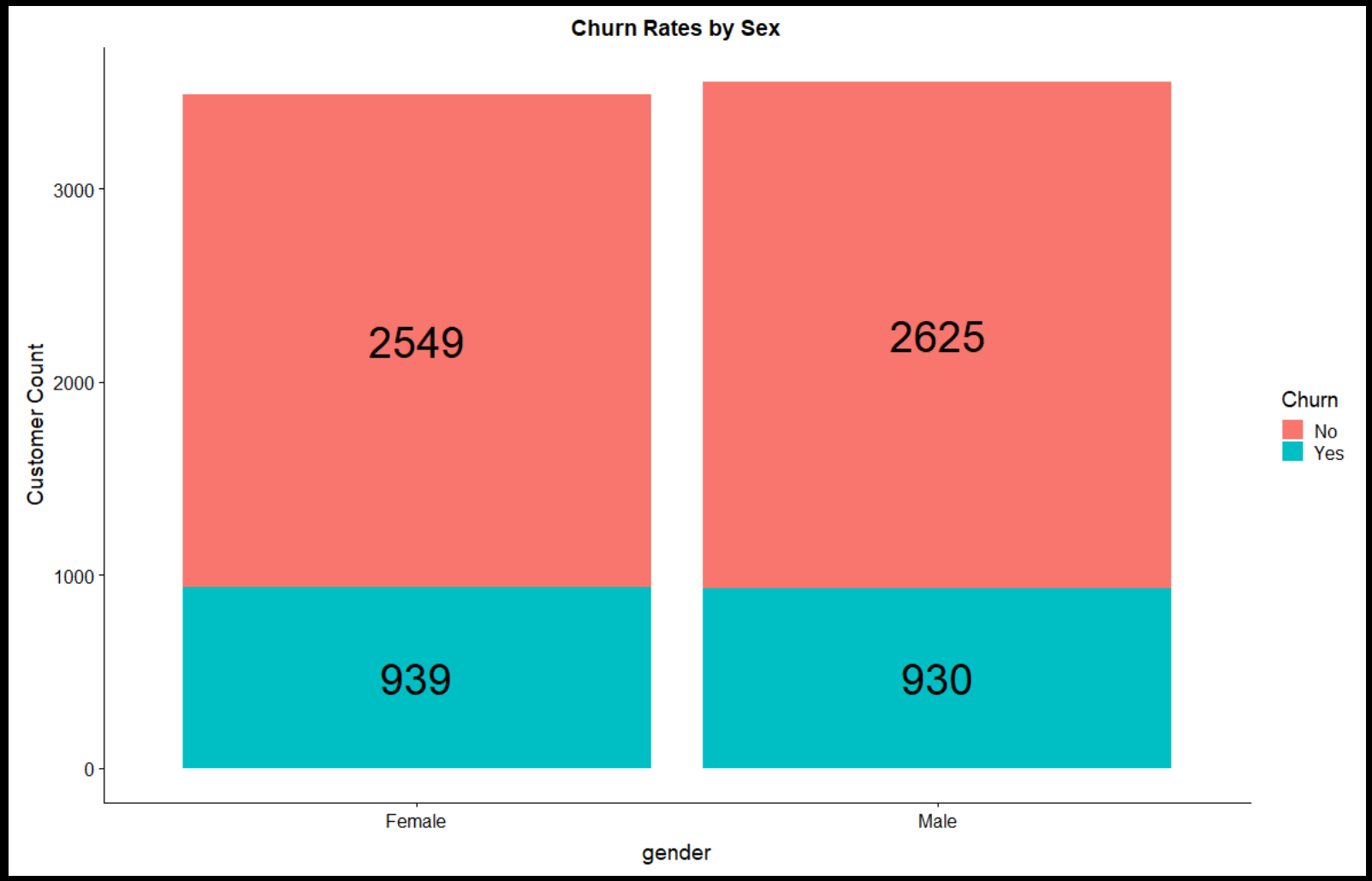
```
prop.table(table(telco$Churn))
```

```
       No       Yes
0.7346301 0.2653699
>
```

Data Visualisasi Untuk Telco

#3.Hubungan Churn Rate dengan Jenis Kelamin

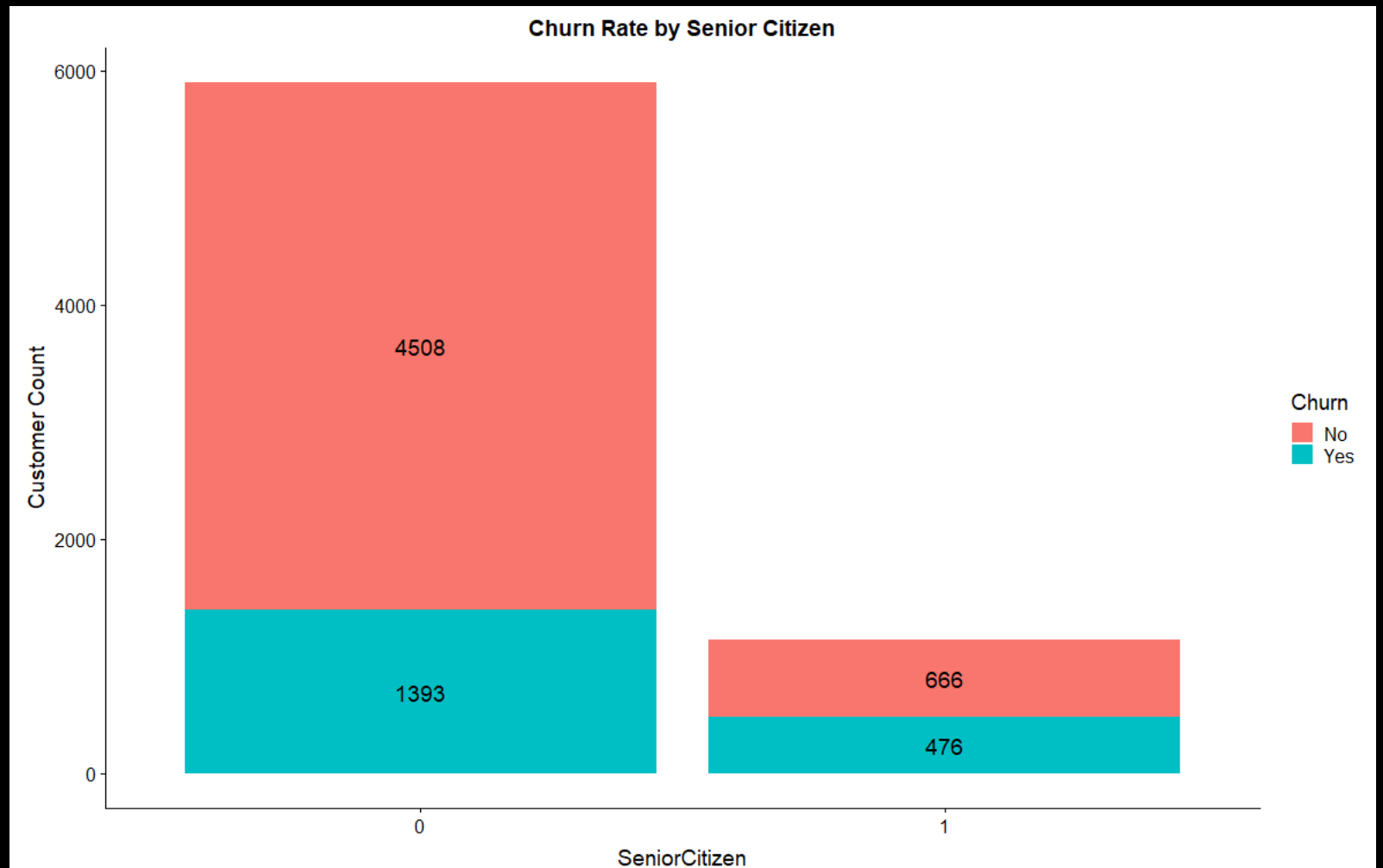
```
ggplot(telco, aes(x = gender, fill = Churn)) + geom_bar() +
  geom_text(aes(label=..count..),stat="count",position=position_stack(0.5),size=10) +
  labs(y = "Customer Count",title = "Churn Rates by Sex")
```



Data Visualisasi Untuk Telco

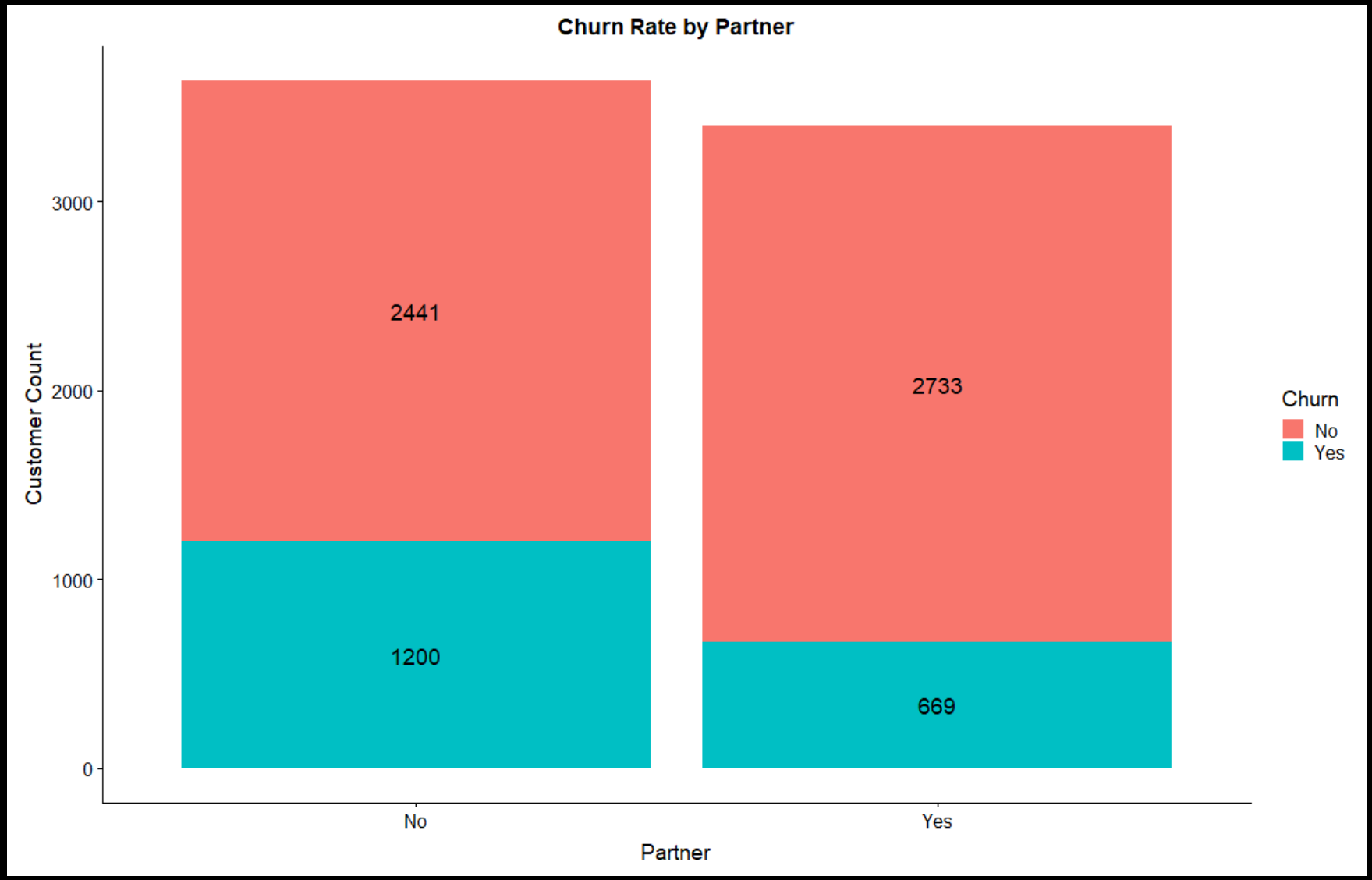
#4.Hubungan Churn Rate dengan Senior Citizen

```
ggplot(telco,aes(x=**************,fill=Churn)) + ****_***() +
  geom_text(aes(label=..count..),stat='count',position=position_stack(0.5),size=5)+
  labs(y="Customer Count",title="Churn Rate by Senior Citizen",size=5)
```



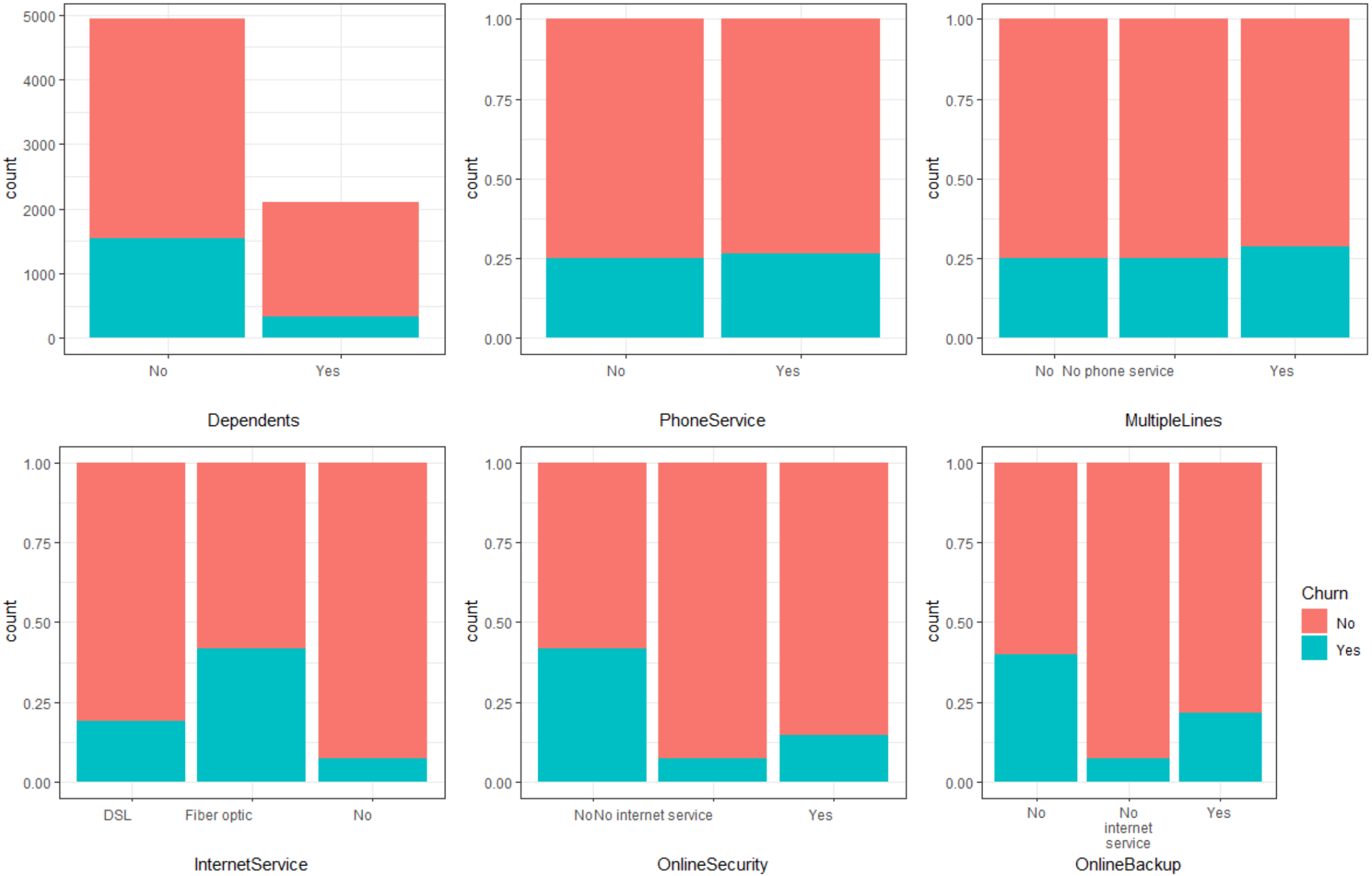
Data Visualisasi Telco

#5.Tulis Code Untuk Menampilkan Hubungan Churn Rate dengan Partner



Data Visualisasi Telco

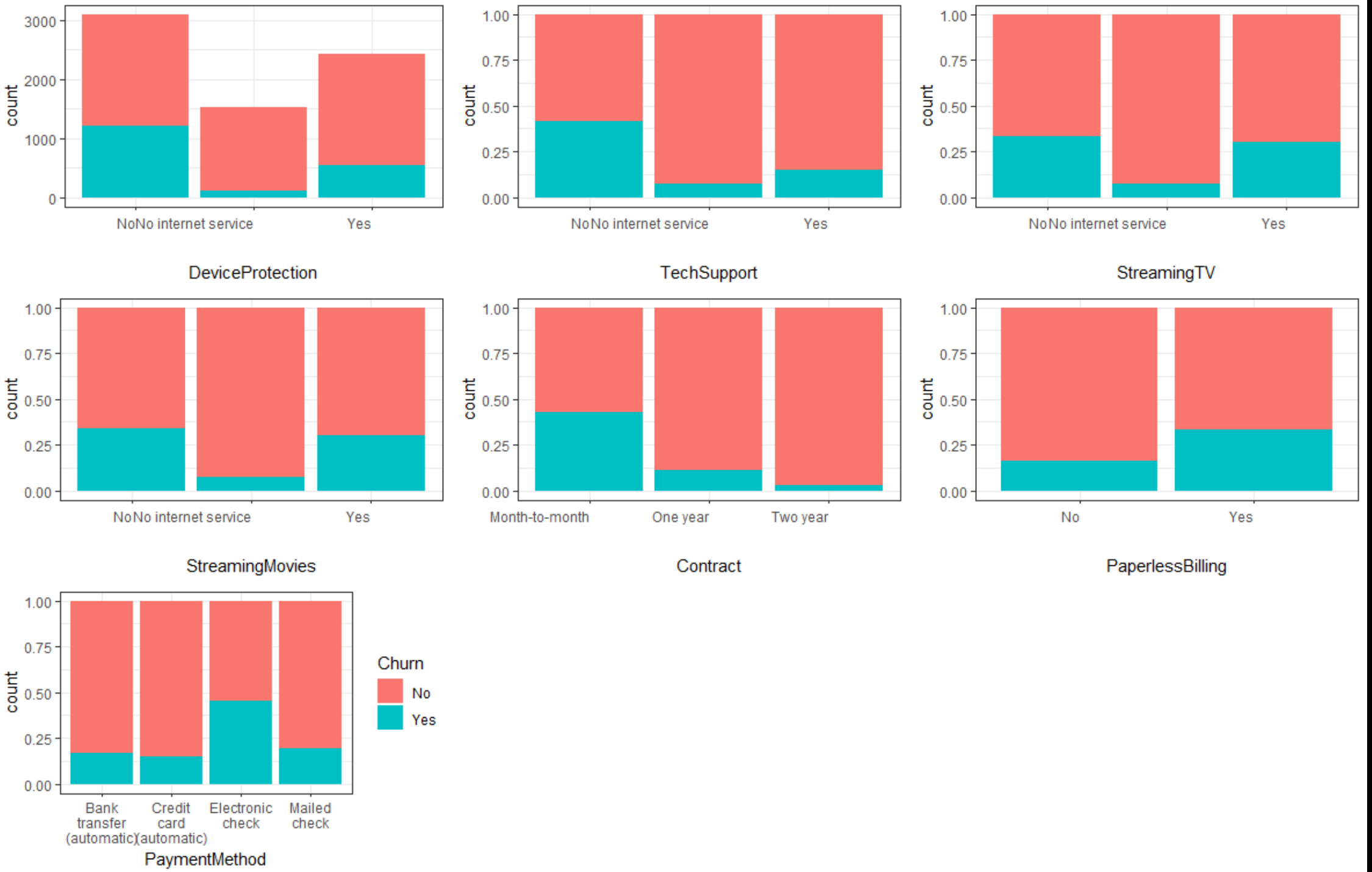
```
#6.Multiplot
#definisikan theme1 untuk merapikan label plot nanti
theme1 = theme_bw()+
  theme(axis.text.x = element_text(angle = 0, hjust = 1, vjust = 0.5),legend.position="none")
#atur grid
options(repr.plot.width = 12, repr.plot.height = 8)
#bentuk plot
plot_grid(ggplot(telco, aes(x=Dependents,fill=Churn))+ geom_bar()+theme1,
          ggplot(telco, aes(x=PhoneService,fill=Churn))+ geom_bar(position = 'fill')+theme1,
          ggplot(telco, aes(x=MultipleLines,fill=Churn))+ geom_bar(position = 'fill')+theme1,
          ggplot(telco, aes(x=InternetService,fill=Churn))+ geom_bar(position = 'fill')+theme1,
          ggplot(telco, aes(x=OnlineSecurity,fill=Churn))+ geom_bar(position = 'fill')+theme1,
          ggplot(telco, aes(x=OnlineBackup,fill=Churn))+ geom_bar(position = 'fill')+theme_bw()+
          scale_x_discrete(labels = function(x) str_wrap(x, width = 10)),align = "h")
```


count
5000
4000
3000
2000
1000
0
No
Yes
Dependents
1.00
0.75
0.50
0.25
0.00
No
Yes
PhoneService
No
No phone service
Yes
MultipleLines
DSL
Fiber optic
No
InternetService
No
No internet service
Yes
OnlineSecurity
No
No internet service
Yes
OnlineBackup
Churn
No
Yes

Data Visualisasi Telco

#7.Multiplot Variabel yang tersisa

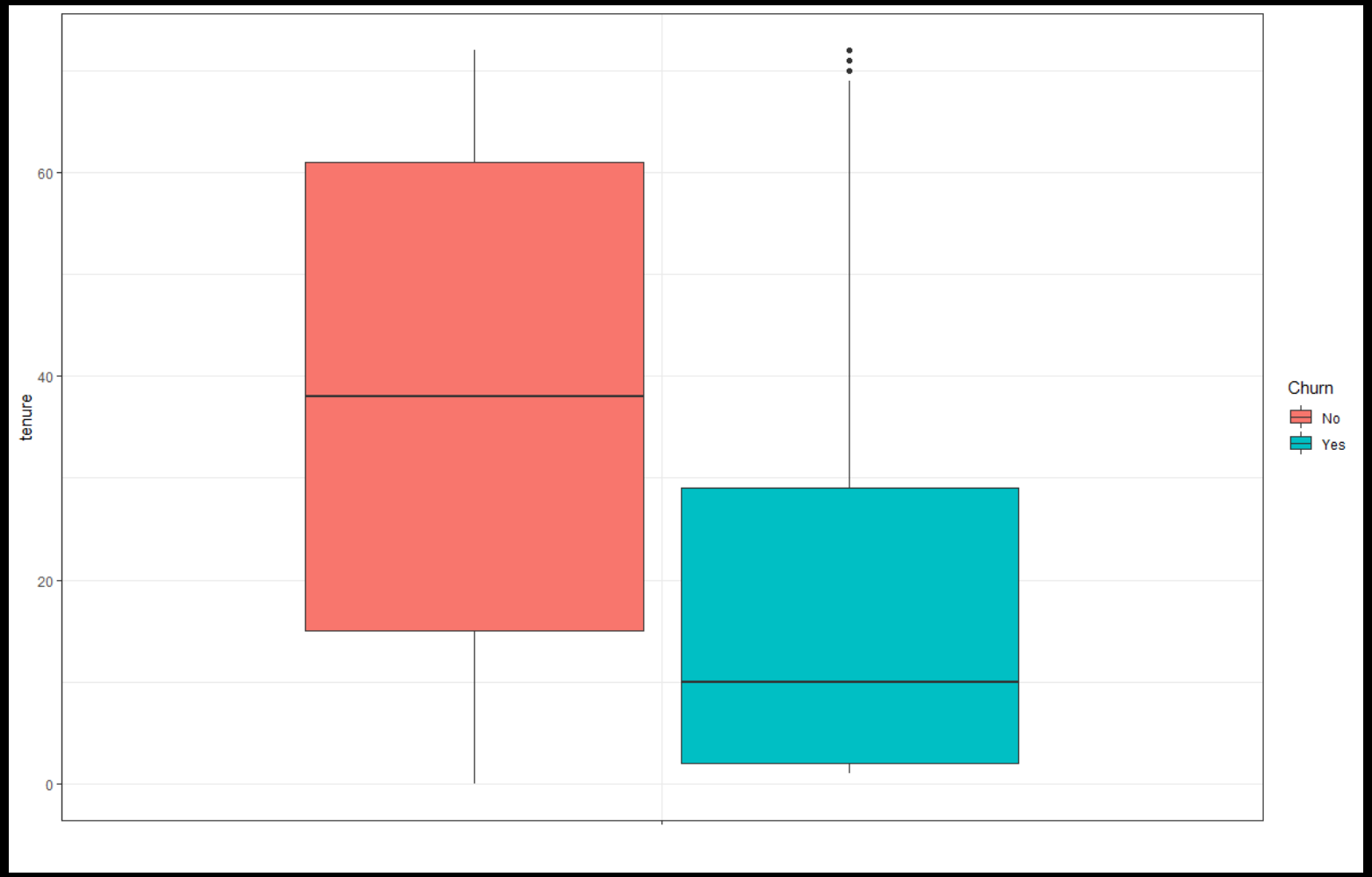
```
plot_grid(ggplot(telco, aes(x=DeviceProtection,fill=Churn))+ geom_bar()+theme1,
          ggplot(telco, aes(x=TechSupport,fill=Churn))+ geom_bar(position = 'fill')+theme1,
          ggplot(telco, aes(x=StreamingTV,fill=Churn))+ geom_bar(position = 'fill')+theme1,
          ggplot(telco, aes(x=StreamingMovies,fill=Churn))+ geom_bar(position = 'fill')+theme1,
          ggplot(telco, aes(x=Contract,fill=Churn))+ geom_bar(position = 'fill')+theme1,
          ggplot(telco, aes(x=PaperlessBilling,fill=Churn))+ geom_bar(position = 'fill')+theme1,
          ggplot(telco, aes(x=PaymentMethod,fill=Churn))+ geom_bar(position = 'fill')+theme_bw()+
          scale_x_discrete(labels = function(x) str_wrap(x, width = 10)),align = "h")
```


count
3000
2000
1000
0
NoNo internet service
Yes
DeviceProtection
1.00
0.75
0.50
0.25
0.00
TechSupport
StreamingTV
StreamingMovies
Month-to-month
One year
Two year
Contract
No
PaperlessBilling
Bank transfer (automatic)
Credit card (automatic)
Electronic check
Mailed check
PaymentMethod
Churn
No
Yes

Data Visualisasi Telco

#8.Hubungan Churn Rate Dengan Tenure

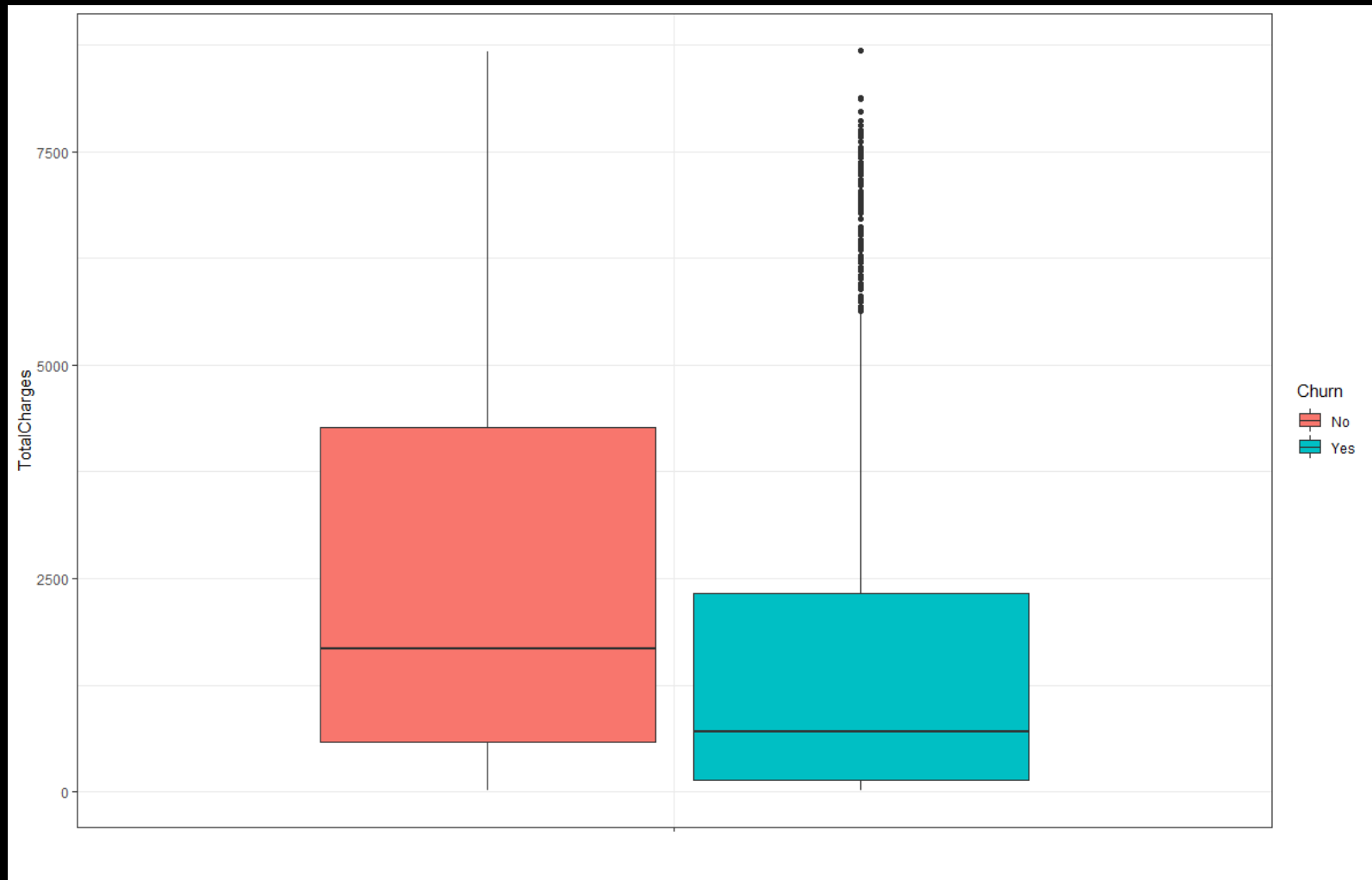
```
ggplot(telco, aes(y= tenure, x = "", fill = Churn)) + geom_boxplot()+
theme_bw()+ xlab(" ")
```



Data Visualisasi Telco

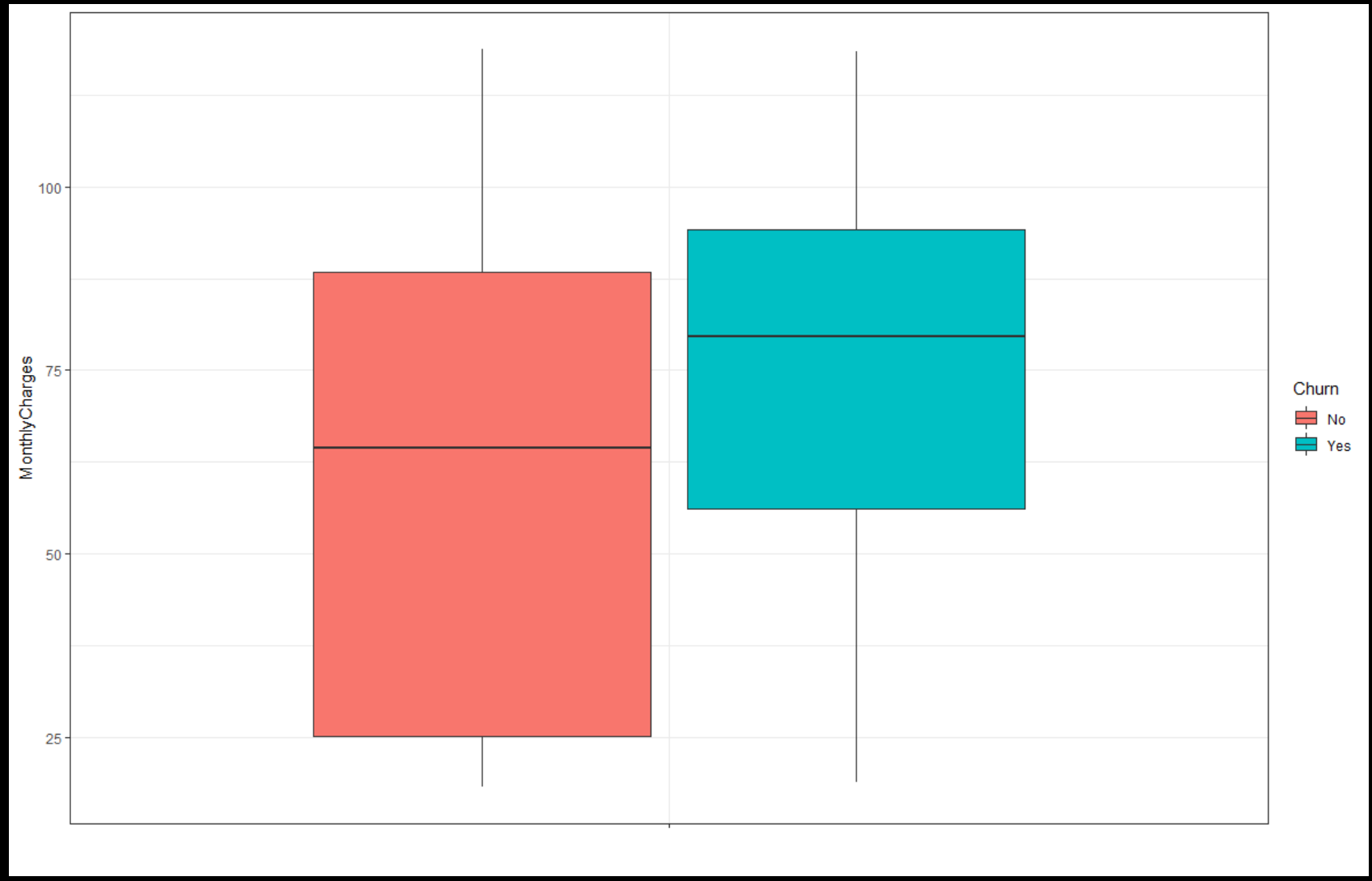
#9.Hubungan Churn Rate Dengan Total Charges

```
ggplot(telco, aes(y= TotalCharges, x = "", fill = Churn)) + geom_boxplot()+
theme_bw()+ xlab(" ")
```



Data Visualisasi Telco

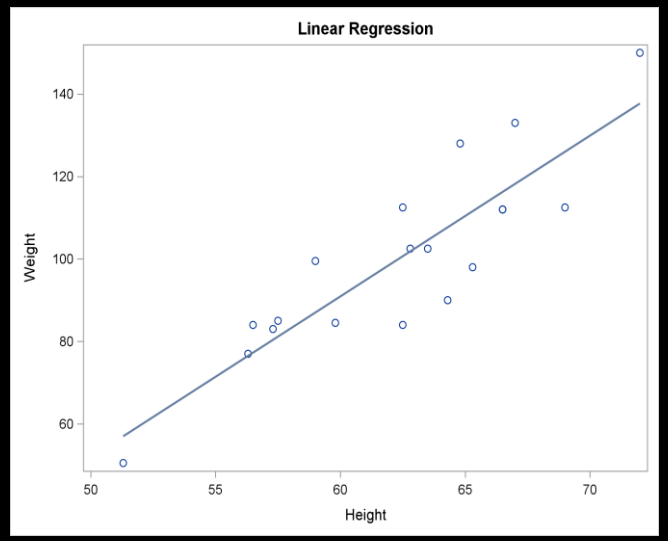
#10.Tulis Kode untuk menampilkan hubungan churn rate dengan monthly charge



Kesimpulan Berdasarkan Visualisasi Telco

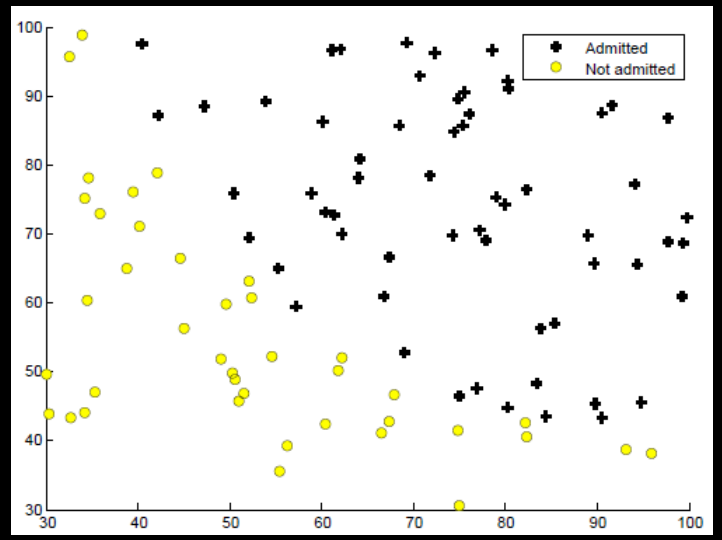
1. Data tidak balance, pelanggan yang tidak churn rate lebih banyak dibandingkan pelanggan yang churn rate
2. Tidak memiliki senior citizen cenderung tidak churn rate.
3. Pelanggan yang tidak memiliki kebergantungan mempunyai kecenderungan untuk churn rate yang lebih tinggi.
4. Contract month-to-month memiliki kecenderungan lebih tinggi.
5. Median pelanggan untuk churn rate adalah 10 bulan.
6. Total tagihan yang tinggi atau mahal membuat pelanggan untuk churn rate

Analytics



Regresi

Klasifikasi



Machine Learning

Umumnya mesin learning dibagi kedalam 4 bagian, yakni supervised learning, unsupervised learning, reinforcement learning, dan asosiasi rule.
Regresi adalah metode yang digunakan untuk memprediksi berdasarkan data yang telah ada. Hal tersebut membuat regresi mempunyai syarat bahwa variabel Y nya bertipe kontinu.
Sedangkan klasifikasi adalah melakukan prediksi berupa kategorikal data berdasarkan data yang telah ada.

Data Structure

Observations	Y	X_1	X_2	.	.	.	X_k
1	Y_1	X_{11}	X_{12}	.	.	.	X_{1k}
2	Y_2	X_{21}	X_{22}	.	.	.	X_{2k}
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
n	Y_n	X_{n1}	X_{n2}	.	.	.	X_{nk}

Classificatin Case

Y: Nominal / Ordinal

X: Nominal / Ordinal / Interval / Rasio

Regression Case

Y: Interval / Rasio

X: Nominal / Ordinal / Interval / Rasio

The Logistic Model

- In any regression problem, the key quantity is the mean value of outcome variable, given values of some predictor variables. It can be called “conditional mean”.

- In linear regression, This conditional mean can be expressed by

$$E(Y|\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}$$

- It is possible to take any values of $\boldsymbol{x}$ in range $(-\infty, \infty)$

The Logistic Model

$$y = \begin{cases} 0 & \text{if not default} \\ 1 & \text{if default} \end{cases}$$

$$x = \text{ balance}$$

- In binary response variable we can take :

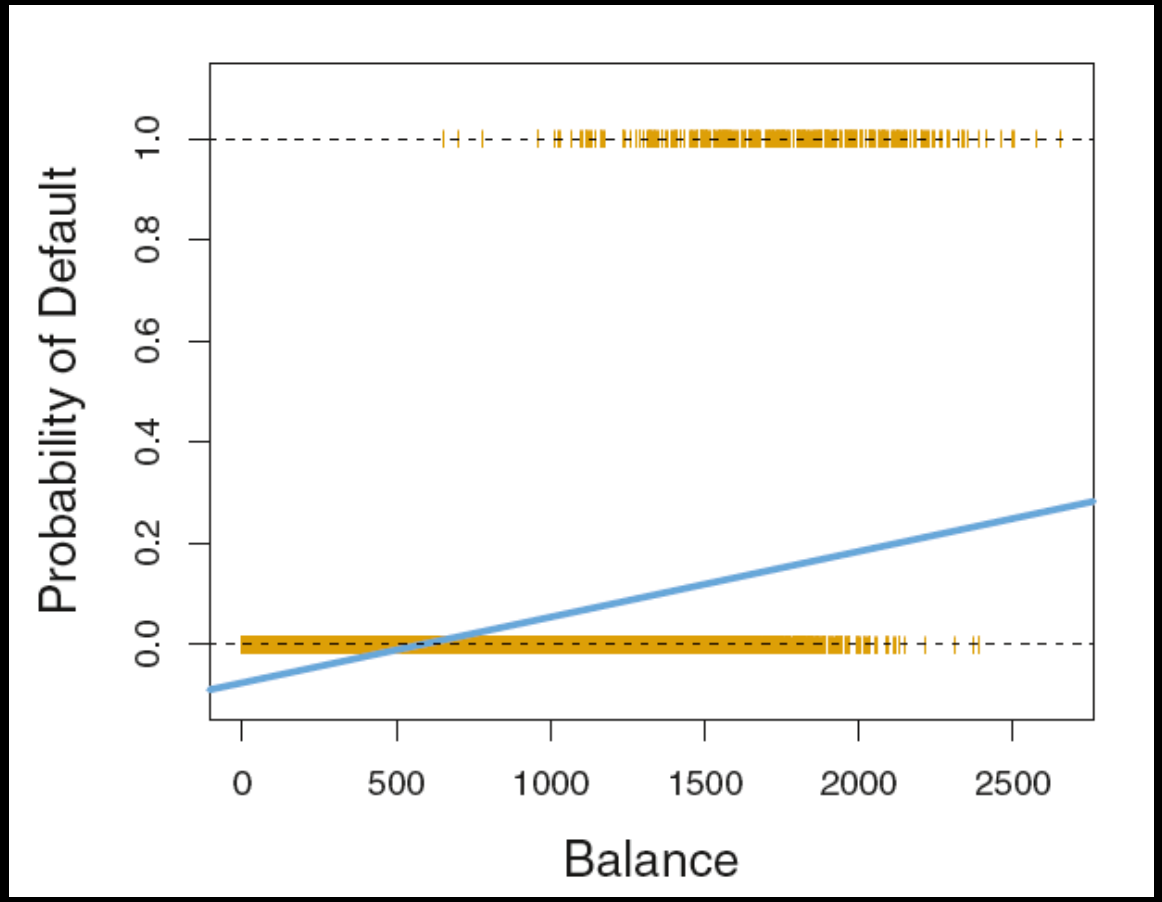
$$E(Y|x) = p(x) = P(Y = 1|x)$$

- And if we apply the linear regression model, we get :

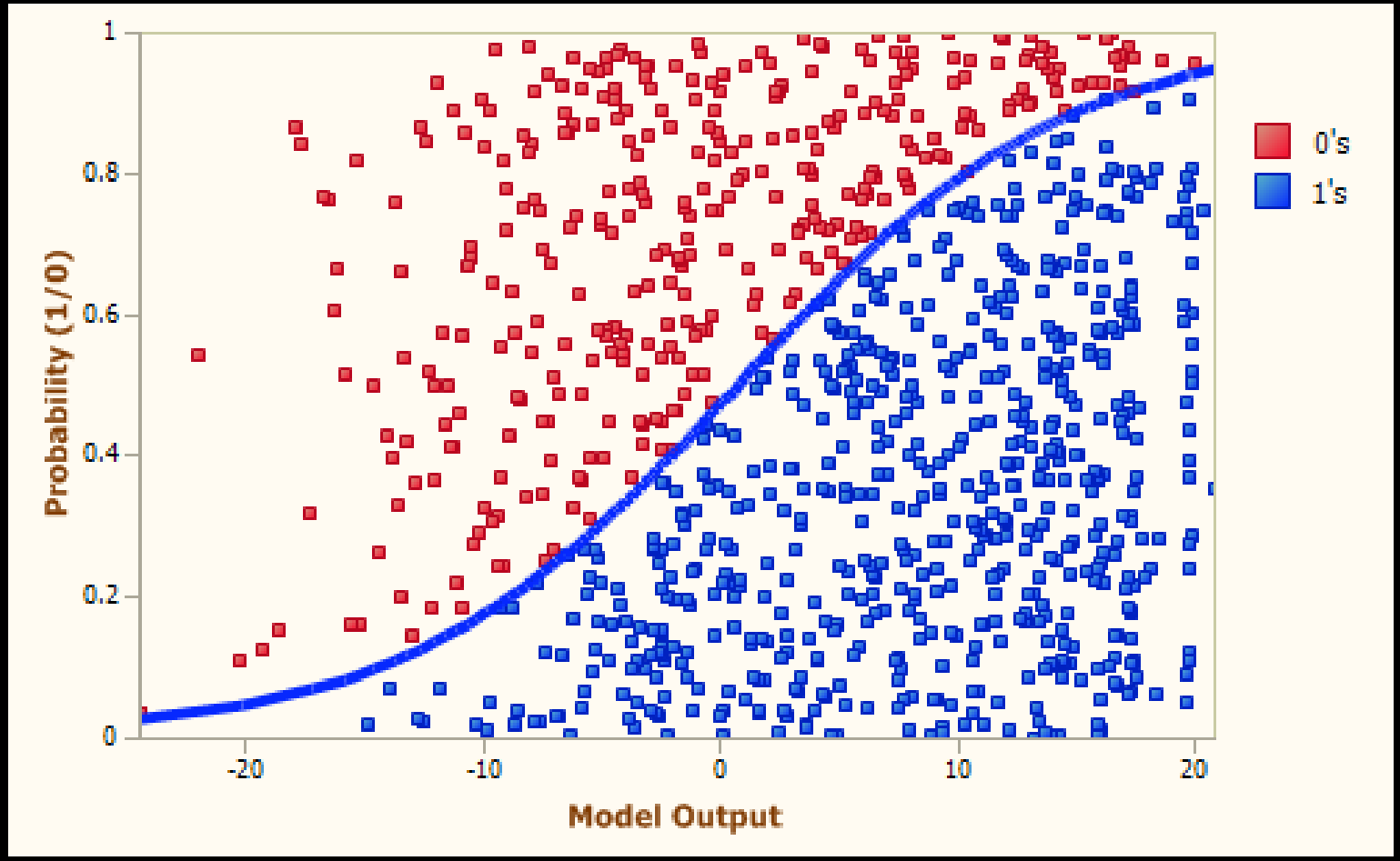
$$p(x) = \beta_0 + \beta x$$

The Logistic Model

- For very large (or small) balance, we will get values bigger than 1 (or smaller than 0)



Picture from: Intoduction to Statistical Learning


1
0.8
0.6
0.4
0.2
0
Probability (1/0)
-20
-10
0
10
20
Model Output
0's
1's

The Logistic Model

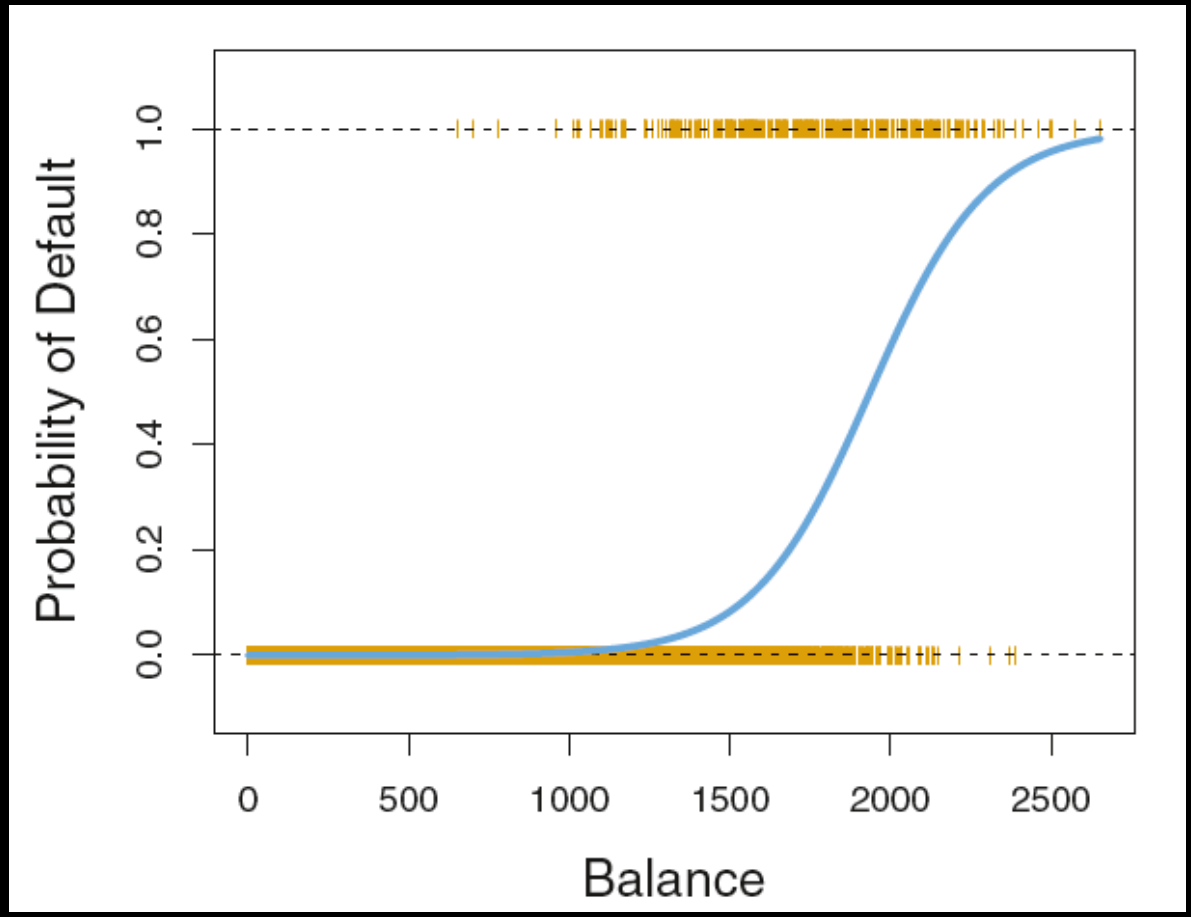
Logistic function gives the output of $p(X)$ between 0 and 1 :

$$p(X) = \frac{e^{\beta_0+\beta x}}{1 + e^{\beta_0+\beta x}}$$

or equivalent with

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta x}$$

$$log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta x x$$



Picture from: Intoduction to Statistical Learning

The Logistic Model

Logistic function gives the output of $p(X)$ between 0 and 1 :

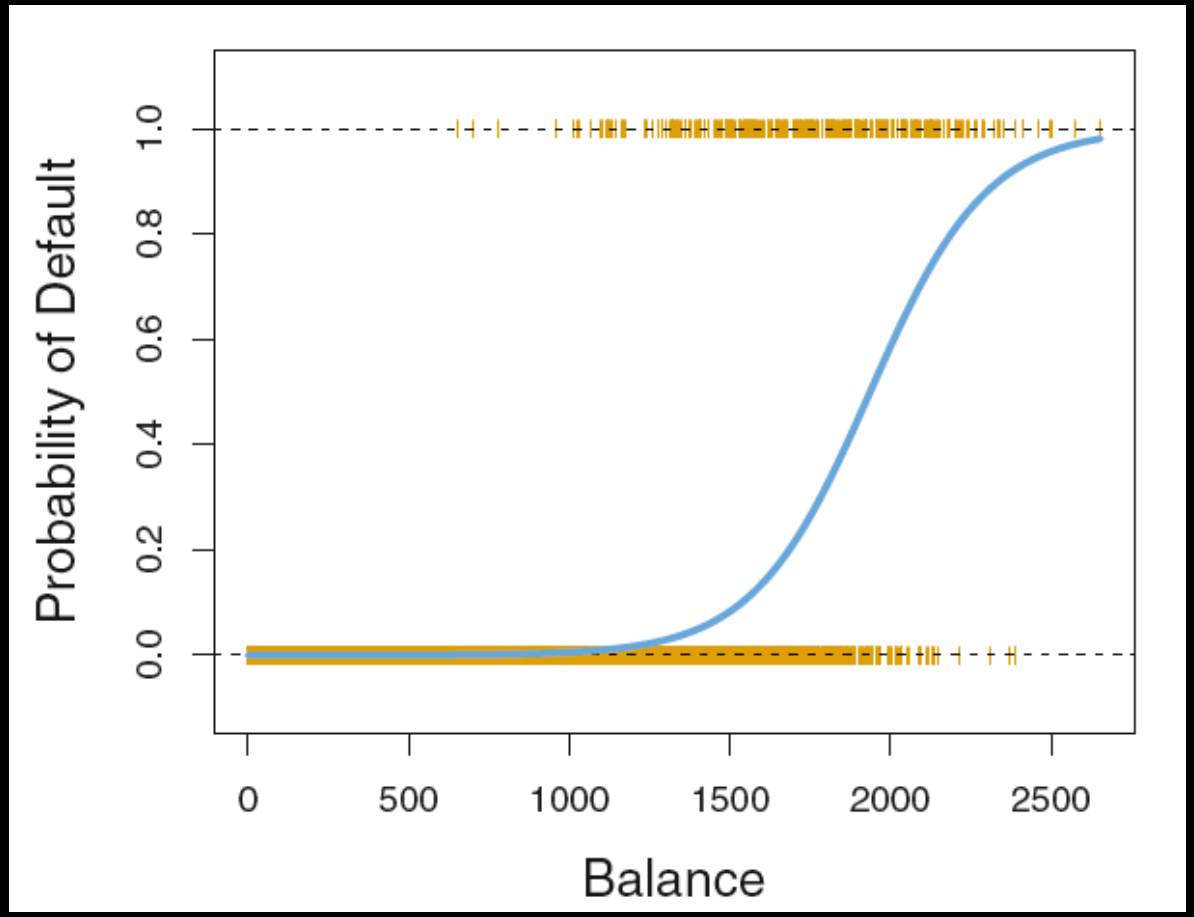
$$p(X) = \frac{e^{\beta_0+\beta x}}{1 + e^{\beta_0+\beta x}}$$

or equivalent with

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta x}$$

$$log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta x x$$

log(odds) or logit



Picture from: Intoduction to Statistical Learning

Multiple Logistic Regression

Consider the log-odds function:

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta x$$

If there are more than 1 variables, we can formulate the function:

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Or equivalent with

$$log\left(\frac{p(X)}{1-p(X)}\right) = \sum_{i=0}^{p} \beta_i x_i$$

where $x_0 = 1$

Estimating Parameter

Maksimum Likelihood Estimation:

- Since y is binary, $y \sim Bernouli(\theta)$, i.e

$$p(y|\theta) = \theta^{y}(1-\theta)^{1-y}, \qquad \text{for } y = \{0,1\} \text{ and } 0 \leq \theta \leq 1$$

- The likelihood function:

$$L(\theta|\boldsymbol{y}) = \prod_{i=1}^{N} p(y_i|\theta_i) = \prod_{i=1}^{N} {\theta_i}^{y_i}(1-\theta_i)^{1-y_i}$$

Estimating Parameter

- The log-likelihood function:

$$\log L(\theta|y) = \ell(\theta) = \sum_{i=1}^{N} y_i \log \theta_i + (1 - y_i) \log(1 - \theta_i)$$

Estimating Parameter

Remember that

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = \sum_{i=0}^{p} \beta_i x_i \,,$$

where $x_0 = 1$

and

$$\theta_i = P(y_i = 1 | \boldsymbol{x}_i, \boldsymbol{\beta}) = p(\boldsymbol{x}_i) = \frac{exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)}{1 + exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)}$$

Estimating Parameter

The log-likelihood function can be formulated by :

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left(y_i \log p(\boldsymbol{x}_i) + (1 - y_i) \log\left(1 - p(\boldsymbol{x}_i)\right) \right)$$

$$= \sum_{i=1}^{N} \left(y_i \log\left(\frac{p(\boldsymbol{x}_i)}{1 - p(\boldsymbol{x}_i)} \right) - \log\left(\frac{1}{1 - p(\boldsymbol{x}_i)} \right) \right)$$

$$= \sum_{i=1}^{N} \left(y_i \left(\sum_{j=0}^{p} \beta_j x_{ij} \right) - \log\left(1 + \exp\left(\sum_{j=0}^{p} \beta_j x_{ij} \right) \right) \right)$$

Estimating Parameter

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left(y_i \left(\sum_{j=0}^{p} \beta_j x_{ij} \right) - \log \left(1 + \exp \left(\sum_{j=0}^{p} \beta_j x_{ij} \right) \right) \right)$$

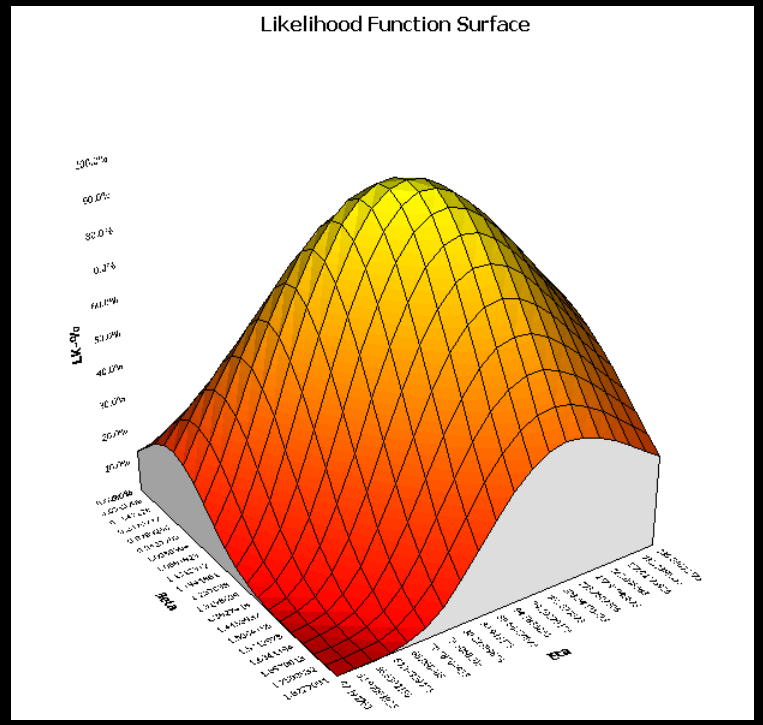
Estimating Parameter

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left(y_i \left(\sum_{j=0}^{p} \beta_j x_{ij} \right) - \log \left(1 + \exp \left(\sum_{j=0}^{p} \beta_j x_{ij} \right) \right) \right)$$

There's no closed-form can solve this maximization problem

Stochastic Gradient Descent

- Stochastic Gradient Descent → Minimization Problem
- We want to maximize the likelihood function



$$\text{Negative Log}-\text{Likelihood} = -\ell(\boldsymbol{\beta})$$

$$= \sum_{i=1}^{N} \left(\log\left(1 + \exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)\right) \right.$$

$$\left. + y_i \left(\sum_{j=0}^{p} \beta_j x_{ij}\right)\right)$$

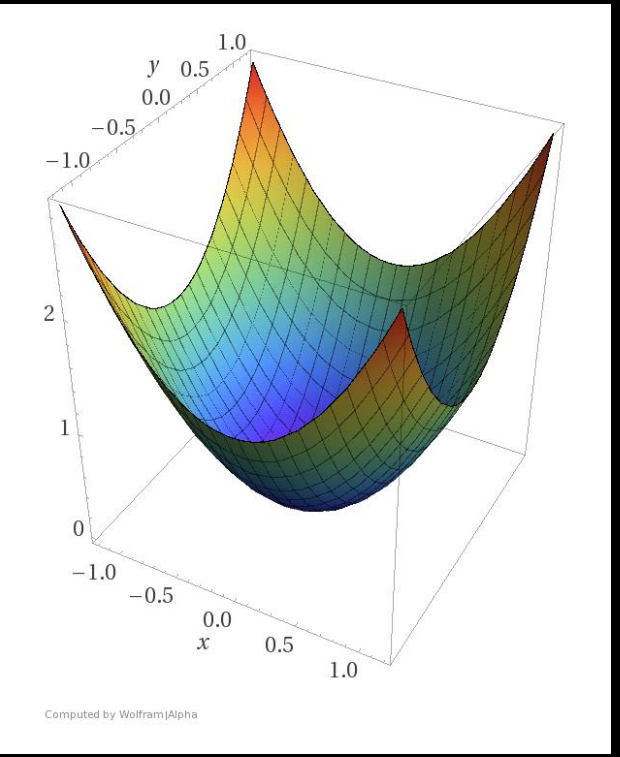
Stochastic Gradient Descent

- Initialize a learning rate and random small parameter value of $\boldsymbol{\beta}$
- Update Rule:

$$\boldsymbol{\beta}^{(t+1)} \leftarrow \boldsymbol{\beta}^{(t)} - \Delta\boldsymbol{\beta}$$
$$\Delta\boldsymbol{\beta} = \eta\nabla_{\boldsymbol{\beta}}\left(-\ell(\boldsymbol{\beta})\right)$$
$$= -\eta\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta})$$

- Gradient:

$$\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}) = \left[\frac{\partial\ell(\boldsymbol{\beta})}{\partial\beta_0}, \frac{\partial\ell(\boldsymbol{\beta})}{\partial\beta_1}, \ldots, \frac{\partial\ell(\boldsymbol{\beta})}{\partial\beta_p}\right]^{\boldsymbol{T}}$$



Stochastic Gradient Descent

- Initialize a learning rate and random small parameter value of $\boldsymbol{\beta}$
- Update Rule:

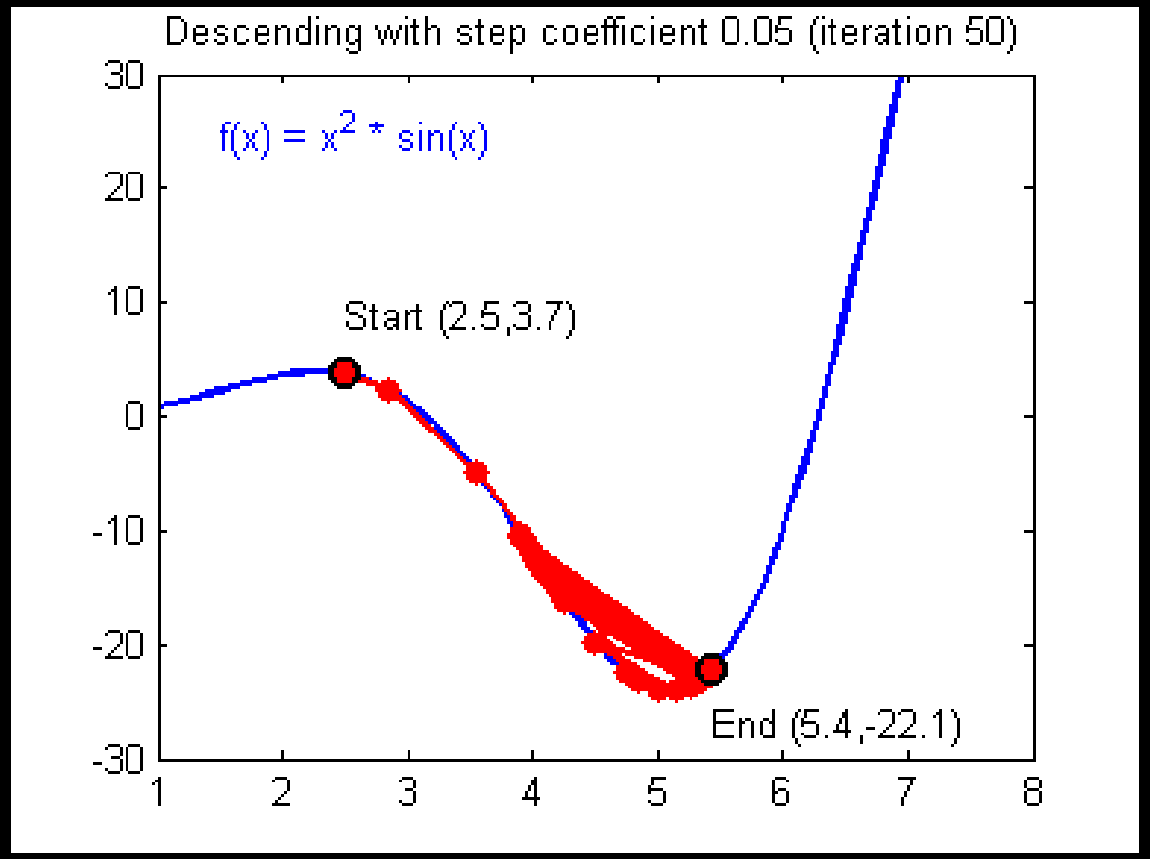
$$\boldsymbol{\beta}^{(t+1)} \leftarrow \boldsymbol{\beta}^{(t)} - \Delta\boldsymbol{\beta}$$
$$\Delta\boldsymbol{\beta} = \eta \nabla_{\boldsymbol{\beta}} \left(-\ell(\boldsymbol{\beta})\right)$$
$$= -\eta \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$$

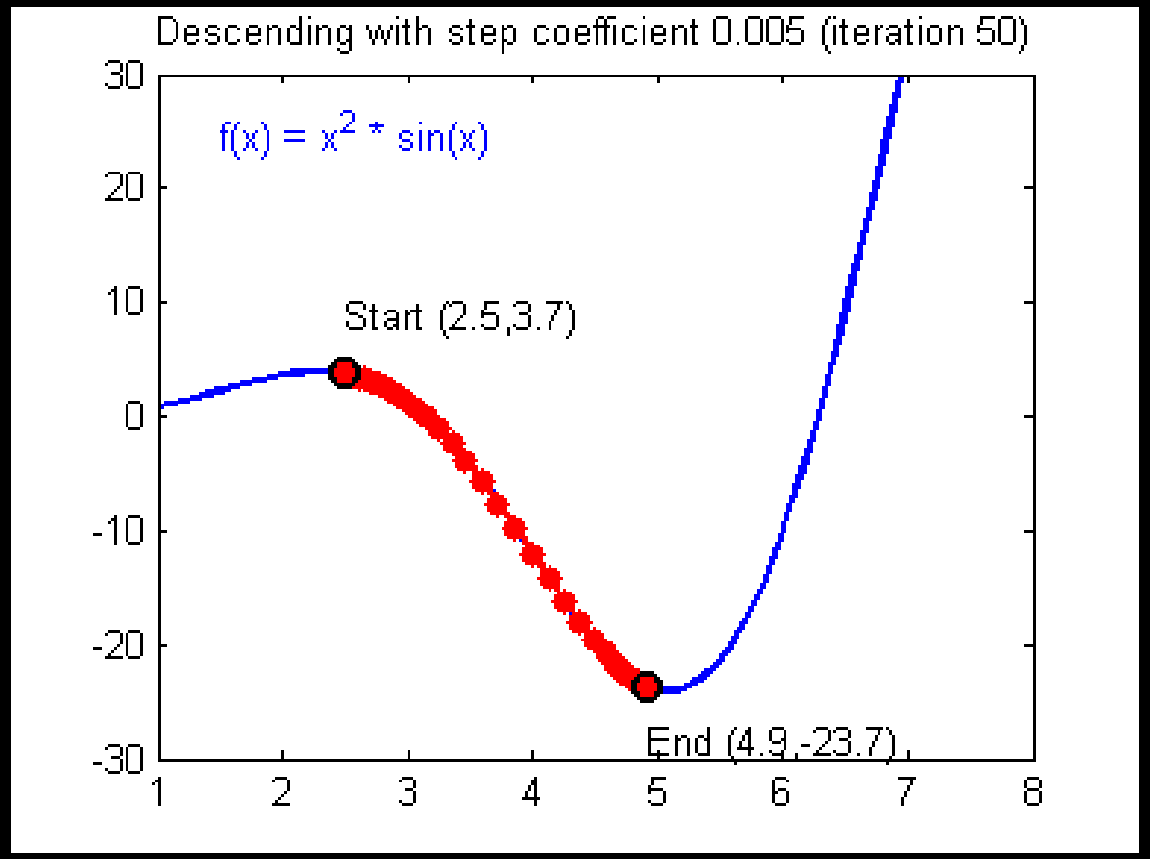
Learning Rate

- Gradient:

$$\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \left[\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_0}, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_1}, \ldots, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_p}\right]^{\boldsymbol{T}}$$

Stochastic Gradient Descent





Stochastic Gradient Descent

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_0} = \frac{\partial \left(\sum_{i=1}^{N} \left(y_i \left(\sum_{j=0}^{p} \beta_i x_{ij} \right) - \log \left(1 + \exp\left(\sum_{j=0}^{p} \beta_i x_{ij} \right) \right) \right) \right)}{\partial \beta_0}$$

$$= \sum_{i=1}^{N} \left(y_i - \frac{\exp\left(\sum_{j=0}^{p} \beta_i x_{ij} \right)}{1 + \exp\left(\sum_{j=0}^{p} \beta_i x_{ij} \right)} \right)$$

$$= \sum_{i=1}^{N} \left(y_i - P(y_i = 1 | \boldsymbol{x}_i, \boldsymbol{\beta}) \right)$$

Stochastic Gradient Descent

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_1} = \sum_{i=1}^{N} x_{i1}(y_i - P(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}))$$

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_2} = \sum_{i=1}^{N} x_{i2}(y_i - P(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}))$$

$$\vdots$$

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_p} = \sum_{i=1}^{N} x_{ip}(y_i - P(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}))$$

Stochastic Gradient Descent

- Update Rule:

$$\begin{pmatrix} \beta_0^{(t+1)} \\ \beta_1^{(t+1)} \\ \vdots \\ \beta_p^{(t+1)} \end{pmatrix} \leftarrow \begin{pmatrix} \beta_0^{(t)} \\ \beta_1^{(t)} \\ \vdots \\ \beta_p^{(t+} \end{pmatrix} + \eta \begin{pmatrix} \sum_{i=1}^{N} \left(y_i - P(y_i = 1 | \boldsymbol{x}_i, \boldsymbol{\beta}^{(t)})\right) \\ \sum_{i=1}^{N} x_{i1} \left(y_i - P(y_i = 1 | \boldsymbol{x}_i, \boldsymbol{\beta}^{(t)})\right) \\ \vdots \\ \sum_{i=1}^{N} x_{ip} \left(y_i - P(y_i = 1 | \boldsymbol{x}_i, \boldsymbol{\beta}^{(t)})\right) \end{pmatrix}$$

Stochastic Gradient Descent

For each observation $\boldsymbol{x}_i$: Update $\boldsymbol{\beta}$

$$\begin{pmatrix} {\beta_0}^{(t+1)} \\ {\beta_1}^{(t+1)} \\ \vdots \\ {\beta_p}^{(t+1)} \end{pmatrix} \leftarrow \begin{pmatrix} {\beta_0}^{(t)} \\ {\beta_1}^{(t)} \\ \vdots \\ {\beta_p}^{(t)} \end{pmatrix} + \eta \begin{pmatrix} \left(y_i - P\left(y_i = 1 | \boldsymbol{x}_i, \boldsymbol{\beta}^{(t)}\right)\right) \\ x_{i1}\left(y_i - P\left(y_i = 1 | \boldsymbol{x}_i, \boldsymbol{\beta}^{(t)}\right)\right) \\ \vdots \\ x_{ip}\left(y_i - P\left(y_i = 1 | \boldsymbol{x}_i, \boldsymbol{\beta}^{(t)}\right)\right) \end{pmatrix}$$

Do until converge: $\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)} \approx \mathbf{0}$ or $\rightarrow$ epoch

SGD Algorithm

1. Initialize learning rate and random values of parameter $\boldsymbol{\beta}$
2. For each observation do:

$$\begin{pmatrix} {\beta_0}^{(t+1)} \\ {\beta_1}^{(t+1)} \\ \vdots \\ {\beta_p}^{(t+1)} \end{pmatrix} \leftarrow \begin{pmatrix} {\beta_0}^{(t)} \\ {\beta_1}^{(t)} \\ \vdots \\ {\beta_p}^{(t)} \end{pmatrix} + \eta \begin{pmatrix} \left(y_i - P\left(y_i = 1 | \boldsymbol{x}_i, \boldsymbol{\beta}^{(t)}\right)\right) \\ x_{i1}\left(y_i - P\left(y_i = 1 | \boldsymbol{x}_i, \boldsymbol{\beta}^{(t)}\right)\right) \\ \vdots \\ x_{ip}\left(y_i - P\left(y_i = 1 | \boldsymbol{x}_i, \boldsymbol{\beta}^{(t)}\right)\right) \end{pmatrix}$$

3. Keep updating beta until converge, or until defined epoch. Do until converge: $\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)} \approx \mathbf{0}$ or $\rightarrow$ epoch

Evaluation

- After we get the value of $\boldsymbol{\beta}$, we can calculate:

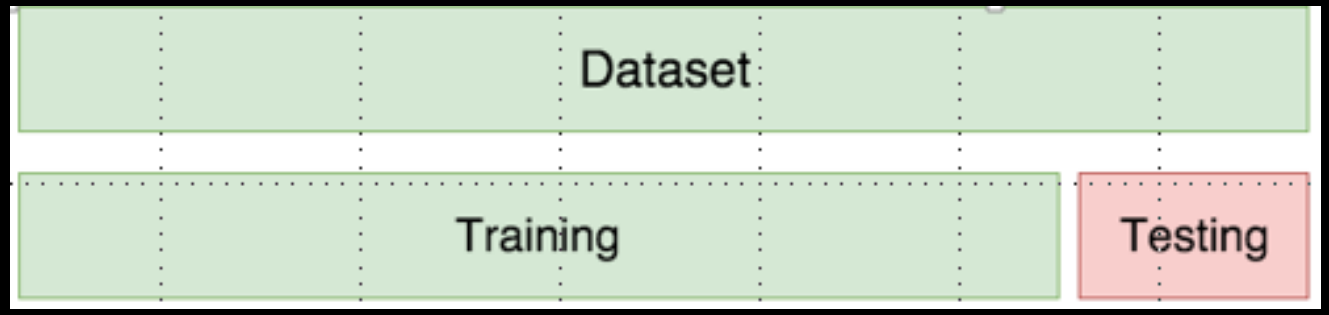
$$p(\boldsymbol{x}_i) = \frac{exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)}{1 + exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)}$$

where $0 < p(\boldsymbol{x}_i) < 1$

- Our prediction can be calculated by: $\hat{y}_i = \begin{cases} 0, & p(\boldsymbol{x}_i) < 0.5 \\ 1, & p(\boldsymbol{x}_i) > 0.5 \end{cases}$

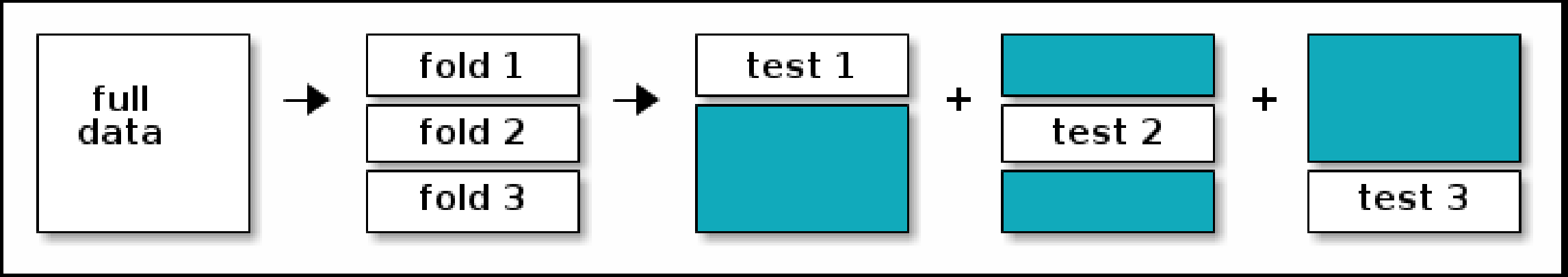
Model Performance

- In Sample Out Sample; Membagi data kedalam data testing dan training. Data training digunakan untuk membangun model kemudian data testing digunakan untuk mengevaluasi kebaikan model. Untuk regresi biasanya menggunakan **RMSE** sedangkan klasifikasi menggunakan **akurasi prediksi**.



Cross Validation

- Membagi data kedalam beberapa bagian (secara umum >2) secara proporsional, kemudian menggunakan masing-masing bagian sebagai data training dan testing dan dilakukan sebanyak jumlah bagian tersebut.



Evaluation methods for classification

Confusion Matrix		Reference	
		Positive	Negative
Prediction	Positive	TP	FP
	Negative	FN	TN

$$Accuracy = \frac{\#correct}{\#predictions} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

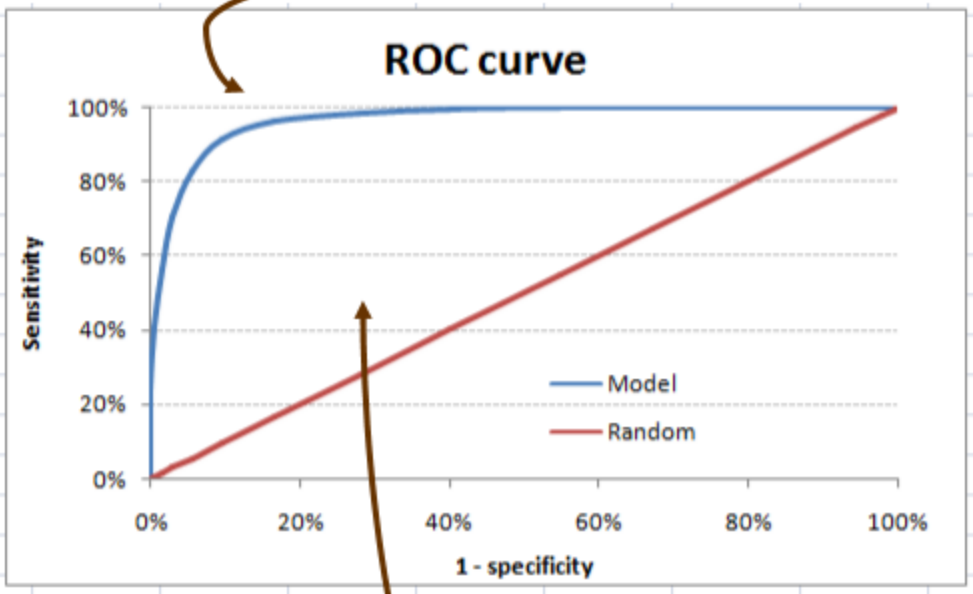
$$Specificity = \frac{TN}{TN + FN}$$

How good at avoiding false alarms

$$Recall = Sensitivity = \frac{TP}{TP + FN}$$

How good it is at detecting positives

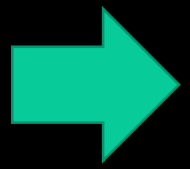
Receiver Operating Characteristic curve



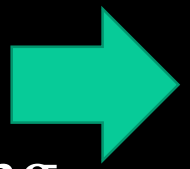
Area under the curve (AUC)

Machine Learning Secara Singkat

Membagi Data Ke Dalam Training dan Testing



Membangun Model Berdasarkan hasil cleaning dan preparing



Mengevaluasi Model, Jika Baik Gunakan/Simpulkan, Jika Kurang Baik Ulangi Membangun Model dengan Merubah parameter atau mengeksplor data lagi

Hipotesis Testing

H_0 : $B_i = 0$ (Variabel Prediktor Tidak Berpengaruh Signifikan Terhadap Respon
H_1 : $B_i \neq 0$ (Variabel Prediktor Berpengaruh Signifikan Terhadap Respon)
Taraf Signifikan : Tentukan α, missal 5% = 0.05
Statistik Uji : p-value
Daerah Kritis : Tolak H_0 jika p-value kurang dari 0.05

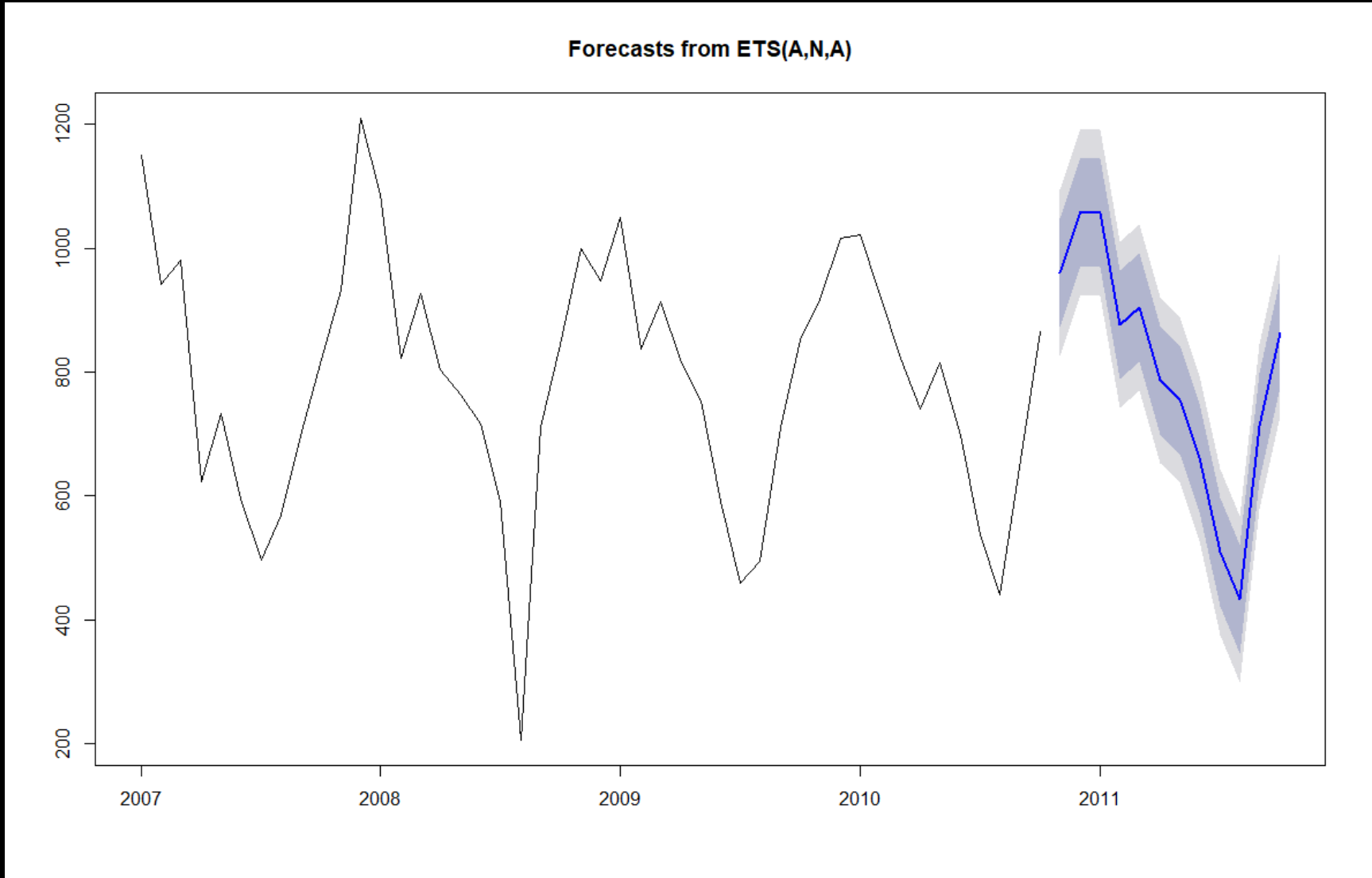
Analytics power Metode Forecast

```
#Analytic
#Power
#Melakukan agregate
total_use_ts = ts(power_monthly$Total_use_kWh, start=c(2007,1), frequency=12)
#Meramalkan
total_use_fc = forecast(total_use_ts, h=12)
#Melihat hasil  model yang terbentuk
summary(total_use_fc)
```

```
Model Information:
ETS(A,N,A)

Call:
 ets(y = object, lambda = lambda, allow.multiplicative.trend = allow.multiplicative.trend)

  Smoothing parameters:
    alpha = 0.0032
    gamma = 1e-04

  Initial states:
    l = 798.8483
    s=260.2981 162.3073 65.5043 -86.6345 -365.1369 -288.3135
           -136.6612 -42.2095 -11.2411 105.6832 77.8047 258.5992

  sigma:  68.2047

     AIC     AICc      BIC
594.5887 610.5887 622.0184

Error measures:
                    ME    RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set -4.76736 68.2047 50.09697 -2.623833 8.405142 0.6017987 -0.07380123

Forecasts:
         Point Forecast    Lo 80     Hi 80    Lo 95     Hi 95
Nov 2010       960.4598 873.0520 1047.8677 826.7811 1094.1386
Dec 2010      1058.4539 971.0456 1145.8621 924.7744 1192.1333
Jan 2011      1056.7635 969.3548 1144.1723 923.0835 1190.4436
Feb 2011       875.9629 788.5538  963.3721 742.2822 1009.6437
Mar 2011       903.8420 816.4324  991.2516 770.1606 1037.5234
Apr 2011       786.8980 699.4880  874.3081 653.2160  920.5801
May 2011       755.9506 668.5402  843.3611 622.2679  889.6334
Jun 2011       661.4900 574.0791  748.9009 527.8066  795.1734
Jul 2011       509.8472 422.4359  597.2585 376.1631  643.5313
Aug 2011       433.0165 345.6048  520.4283 299.3318  566.7013
Sep 2011       711.5142 624.1020  798.9264 577.8287  845.1996
Oct 2011       863.6528 776.2401  951.0654 729.9667  997.3389
```

Plot Hasil Forecast Power

```
#Menampilkan plot
plot(total_use_fc)
```



Analytic Telco Logistik Regression

```
#Analytic Telco
#Jangan Lupa Menghilangkan variabel ID
telco = telco[,-1]
#Logistik Regression Model Awal
modelawal = glm(Churn~.,family=binomial,data=telco)
#Lihat Hasil Model
summary(modelawal)
```

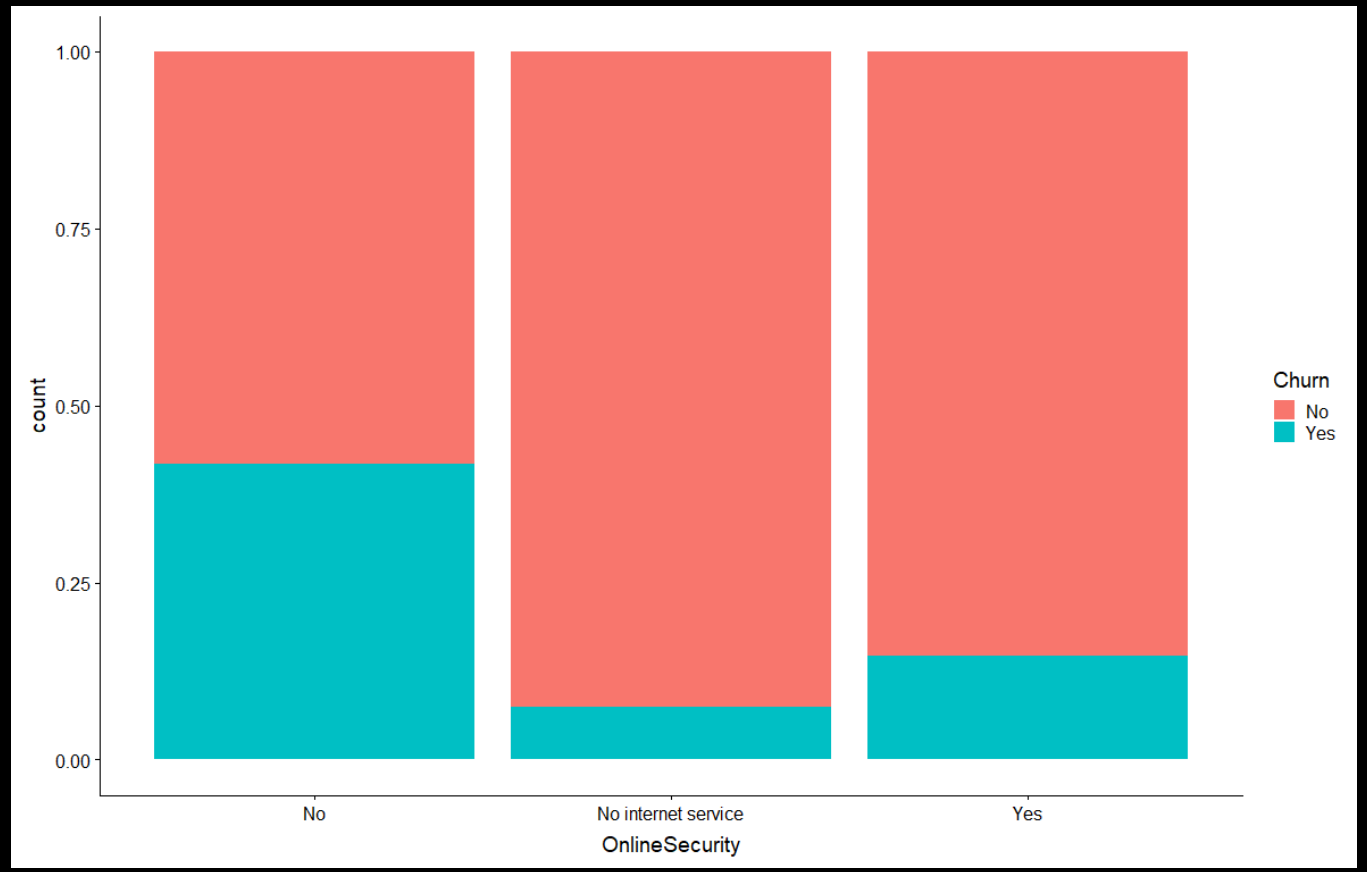
```
Call:
glm(formula = Churn ~ ., family = binomial(link = "logit"), data = telcodf)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9179  -0.6781  -0.2850   0.7269   3.4255

Coefficients: (7 not defined because of singularities)
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            1.151e+00  8.146e-01   1.413  0.15762
genderMale                            -2.188e-02  6.479e-02  -0.338  0.73557
SeniorCitizen1                         2.151e-01  8.452e-02   2.545  0.01094 *
PartnerYes                            -2.692e-03  7.779e-02  -0.035  0.97239
DependentsYes                         -1.538e-01  8.971e-02  -1.714  0.08653 .
tenure                                -5.941e-02  6.156e-03  -9.650  < 2e-16 ***
PhoneServiceYes                        1.798e-01  6.479e-01   0.278  0.78138
MultipleLinesNo phone service                 NA         NA      NA       NA
MultipleLinesYes                       4.469e-01  1.771e-01   2.524  0.01160 *
InternetServiceFiber optic             1.753e+00  7.976e-01   2.198  0.02796 *
InternetServiceNo                     -1.791e+00  8.066e-01  -2.220  0.02639 *
OnlineSecurityNo internet service             NA         NA      NA       NA
OnlineSecurityYes                     -2.055e-01  1.786e-01  -1.150  0.25004
OnlineBackupNo internet service               NA         NA      NA       NA
OnlineBackupYes                        2.579e-02  1.752e-01   0.147  0.88298
DeviceProtectionNo internet service           NA         NA      NA       NA
DeviceProtectionYes                    1.477e-01  1.763e-01   0.838  0.40219
TechSupportNo internet service                NA         NA      NA       NA
TechSupportYes                        -1.789e-01  1.805e-01  -0.991  0.32150
StreamingTVNo internet service                NA         NA      NA       NA
StreamingTVYes                         5.912e-01  3.261e-01   1.813  0.06987 .
StreamingMoviesNo internet service            NA         NA      NA       NA
StreamingMoviesYes                     6.038e-01  3.264e-01   1.850  0.06433 .
ContractOne year                      -6.671e-01  1.075e-01  -6.208 5.38e-10 ***
ContractTwo year                      -1.390e+00  1.758e-01  -7.905 2.68e-15 ***
PaperlessBillingYes                    3.418e-01  7.447e-02   4.591 4.42e-06 ***
PaymentMethodCredit card (automatic)  -8.649e-02  1.141e-01  -0.758  0.44835
PaymentMethodElectronic check          3.057e-01  9.448e-02   3.236  0.00121 **
PaymentMethodMailed check             -5.666e-02  1.148e-01  -0.493  0.62176
MonthlyCharges                        -4.037e-02  3.173e-02  -1.272  0.20326
TotalCharges                           3.184e-04  7.009e-05   4.543 5.54e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8150.1  on 7042  degrees of freedom
Residual deviance: 5829.3  on 7019  degrees of freedom
AIC: 5877.3

Number of Fisher Scoring iterations: 6
```

MODEL BELUM CUKUP BAIK, NEXT STEP

- Cleaning the Categorical features
- Standardising Continuous features
- Creating derived features
- Creating dummy variables for factor variables
- Handling Imbalanced
- Creating the final dataset
- Splitting the data into train and validation set.
- Model evaluation

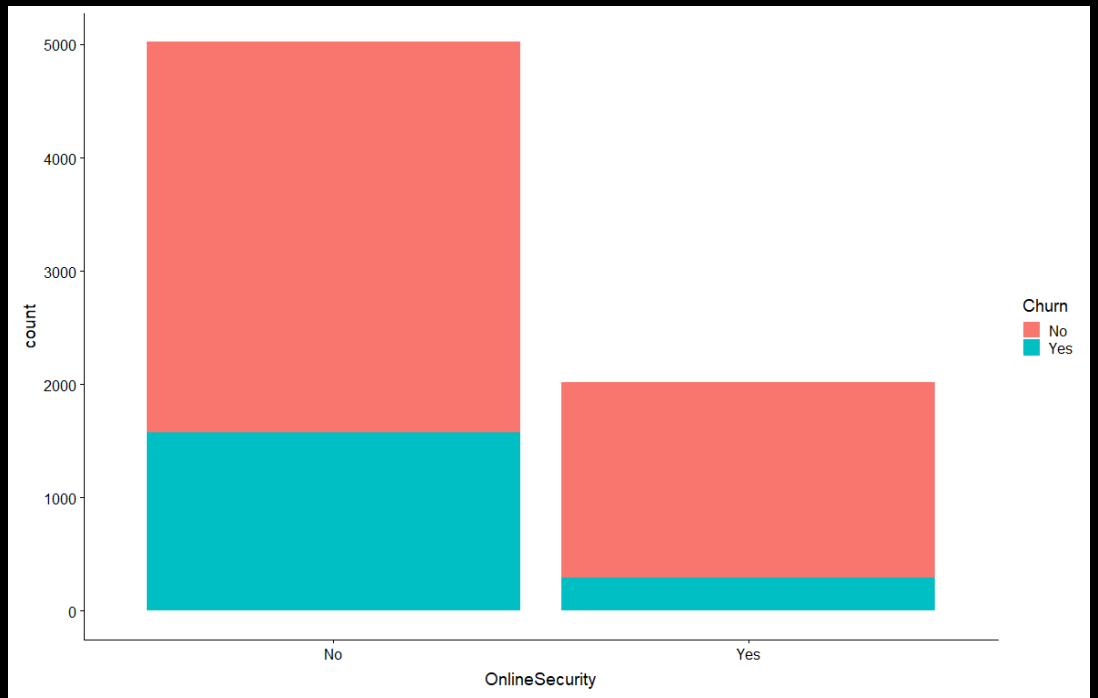
Cleaning the Categorical features

Mengoptimalkan variabel kategori, seperti pada bagian online security, kelas No internet service dan No terpisah, Padahal keduanya memiliki arti yang sama.



Cleaning Categorical Feature

```
#Membersihkan Kategorikal Feature
#Merubah No internet service menjadi no
telco = data.frame(lapply(telco,function(x){gsub("No internet service","No",x)}))
#Merubah No phone service menjadi No
telco = data.frame(lapply(telco,function(x){gsub("No phone service","No",x)}))
#Cek Apakah sudah terganti Gunakan salah satu saja, misal online security
ggplot(telco,aes(x=OnlineSecurity,fill=Churn))+geom_bar()
```



Standardising Continuous features

Mentransformasi Data (kontinu) kedalam skala yang sama. Pada kasus ini variabel yang akan distandarkan adalah tenure, monthly charge dan total charges.

Standardizing

```
#Standardizing
#Memisah Feature/Variabel Numeric kedalam dataframe baru
numeric_feat = c("tenure","MonthlyCharges","TotalCharges")
#Menggabungkan Numeric Feat kedalam data telco
telco[numeric_feat] = sapply(telco[numeric_feat], as.numeric)
#Membuat Dataframe telco_int
telco_int = telco[,c("tenure", "MonthlyCharges", "TotalCharges")]
#Melakukan standardize
telco_int = data.frame(scale(telco_int))
#tampilkan Telco_int
head(telco_int)
```

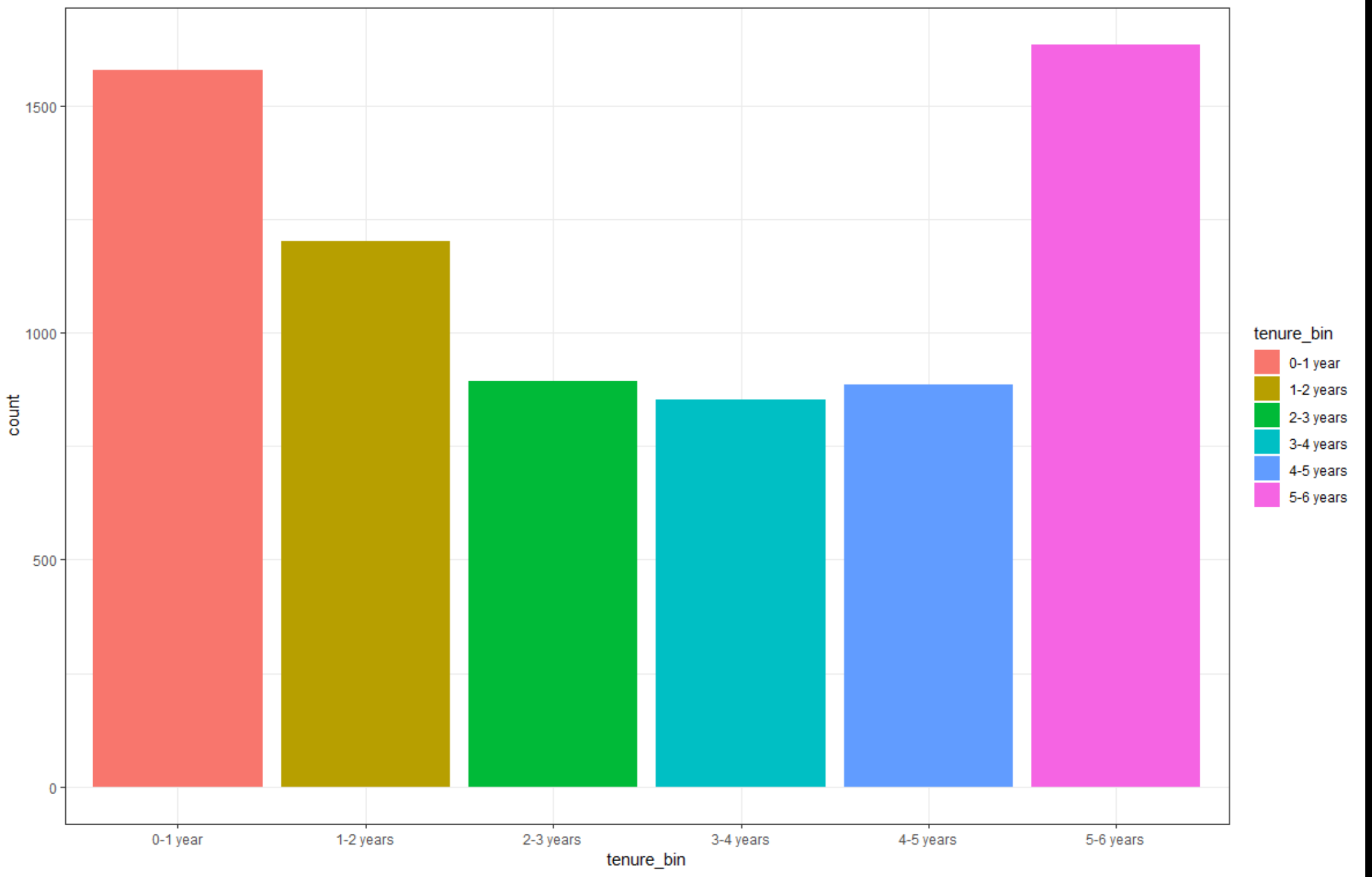
```
      tenure MonthlyCharges TotalCharges
1 -1.4229430     -0.72355882   -0.3994534
2 -0.2948291      0.04037956   -0.9500269
3 -0.9633411     -0.09266589   -1.6442053
4  0.2065548     -0.45746793   -0.9850008
5 -0.9633411      0.53608115   -1.2367067
6  1.5017967      1.70559356    1.5076825
```

Creating Derived Feature

Membuat derived feature adalah membuat feature baru berdasarkan feature yang ada. Pada kasus ini akan dibuat feature baru yang berasal dari feature tenure.

Creating derived features

```
#Membuat Feature Baru Berdasarkan Feature yang ada
telco = mutate(telco,tenure_bin = tenure)
telco$tenure_bin[telco$tenure_bin >=0 & telco$tenure_bin <= 12] <- '0-1 year'
telco$tenure_bin[telco$tenure_bin > 12 & telco$tenure_bin <= 24] <- '1-2 years'
telco$tenure_bin[telco$tenure_bin > 24 & telco$tenure_bin <= 36] <- '2-3 years'
telco$tenure_bin[telco$tenure_bin > 36 & telco$tenure_bin <= 48] <- '3-4 years'
telco$tenure_bin[telco$tenure_bin > 48 & telco$tenure_bin <= 60] <- '4-5 years'
telco$tenure_bin[telco$tenure_bin > 60 & telco$tenure_bin <= 73] <- '5-6 years'
telco$tenure_bin <- as.factor(telco$tenure_bin)
#Tampilkan Barchart Variabel Baru
ggplot(telco, aes(tenure_bin, fill = tenure_bin)) + geom_bar()+ theme_bw()
```


1500
1000
500
0
count
0-1 year
1-2 years
2-3 years
3-4 years
4-5 years
5-6 years
tenure_bin
tenure_bin
0-1 year
1-2 years
2-3 years
3-4 years
4-5 years
5-6 years

Creating dummy variables for factor variables

Membentuk variabel dummy hanya dilakukan pada kategorikal data saja. Tujuannya adalah mengatasi singularitas atau kombinasi linear.

Creating dummy variables for factor variables

```
#Membuat Dummy Variabel
#Buat data frame untuk variabel yang kategori
telco_cat <- telco[,-c(5,18,19)]
#Buat Variabel Dummynya
dummy<- data.frame(sapply(telco_cat,function(x) data.frame(model.matrix(~x-1,data
=telco_cat))[,-1]))
#Tampilkan
View(dummy)
```

	gender	SeniorCitizen	Partner	Dependents	PhoneService	MultipleLines	InternetService.xFiber.optic	InternetService.xNo	OnlineSecu
1	0	0	1	0	0	0	0	0	
2	1	0	0	0	1	0	0	0	
3	1	0	0	0	1	0	0	0	
4	1	0	0	0	0	0	0	0	
5	0	0	0	0	1	0	1	0	
6	0	0	0	0	1	1	1	0	
7	1	0	0	1	1	1	1	0	
8	0	0	0	0	0	0	0	0	

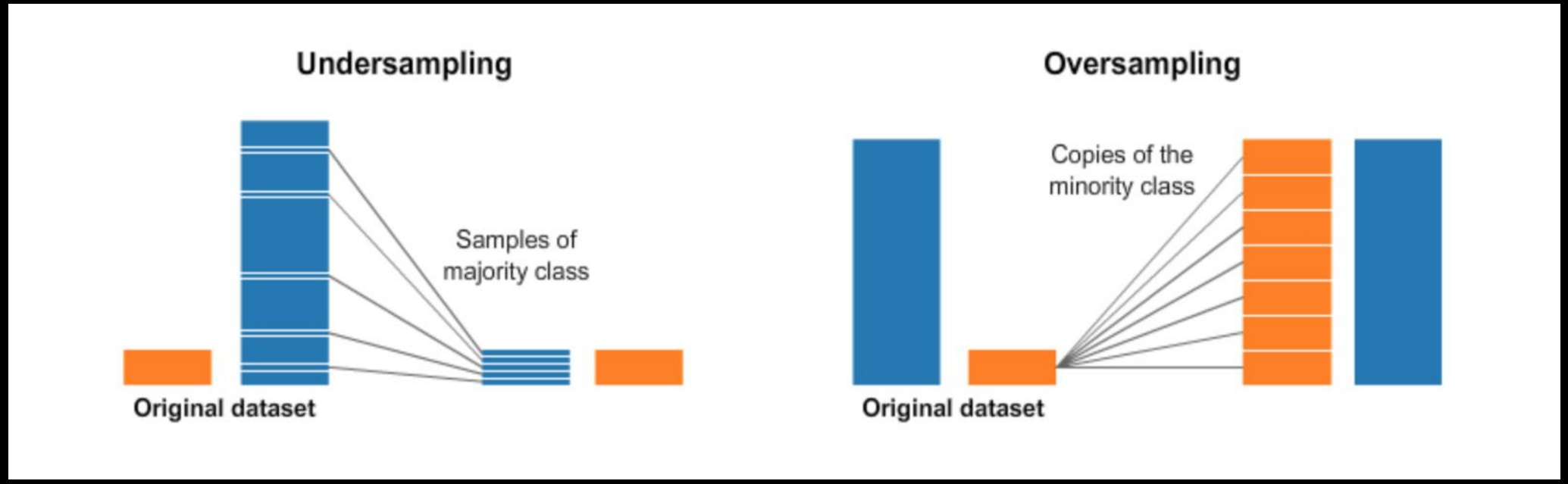
Memerging Data Final

```
telco_final = cbind (telco_int,dummy)
telco_final$Churn = as.factor(telco_final$Churn)
str(telco_final)
```

```
'data.frame':   7043 obs. of  29 variables:
 $ tenure                                : num  -1.423 -0.295 -0.963 0.207 -0.963 ...
 $ MonthlyCharges                        : num  -0.7236 0.0404 -0.0927 -0.4575 0.5361 ...
 $ TotalCharges                          : num  -0.399 -0.95 -1.644 -0.985 -1.237 ...
 $ gender                                : num  0 1 1 1 0 0 1 0 0 1 ...
 $ SeniorCitizen                         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Partner                               : num  1 0 0 0 0 0 0 0 1 0 ...
 $ Dependents                            : num  0 0 0 0 0 0 1 0 0 1 ...
 $ PhoneService                          : num  0 1 1 0 1 1 1 0 1 1 ...
 $ MultipleLines                         : num  0 0 0 0 0 1 1 0 1 0 ...
 $ InternetService.xFiber.optic          : num  0 0 0 0 1 1 1 0 1 0 ...
 $ InternetService.xNo                   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ OnlineSecurity                        : num  0 1 1 1 0 0 0 1 0 1 ...
 $ OnlineBackup                          : num  1 0 1 0 0 0 1 0 0 1 ...
 $ DeviceProtection                      : num  0 1 0 1 0 1 0 0 1 0 ...
 $ TechSupport                           : num  0 0 0 1 0 0 0 0 1 0 ...
 $ StreamingTV                           : num  0 0 0 0 0 1 1 0 1 0 ...
 $ StreamingMovies                       : num  0 0 0 0 0 1 0 0 1 0 ...
 $ Contract.xOne.year                    : num  0 1 0 1 0 0 0 0 0 1 ...
 $ Contract.xTwo.year                    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ PaperlessBilling                      : num  1 0 1 0 1 1 1 0 1 0 ...
 $ PaymentMethod.xCredit.card..automatic.: num  0 0 0 0 0 0 1 0 0 0 ...
 $ PaymentMethod.xElectronic.check       : num  1 0 0 0 1 1 0 0 1 0 ...
 $ PaymentMethod.xMailed.check           : num  0 1 1 0 0 0 0 1 0 0 ...
 $ Churn                                 : Factor w/ 2 levels "0","1": 1 1 2 1 2 2 1 1 2 1 .
..
 $ tenure_bin.x1.2.years                 : num  0 0 1 0 1 0 1 0 1 0 ...
 $ tenure_bin.x2.3.years                 : num  0 1 0 0 0 0 0 0 0 0 ...
 $ tenure_bin.x3.4.years                 : num  0 0 0 1 0 0 0 0 0 0 ...
 $ tenure_bin.x4.5.years                 : num  0 0 0 0 0 0 0 0 0 1 ...
 $ tenure_bin.x5.6.years                 : num  0 0 0 0 0 1 0 0 0 0 ...
```

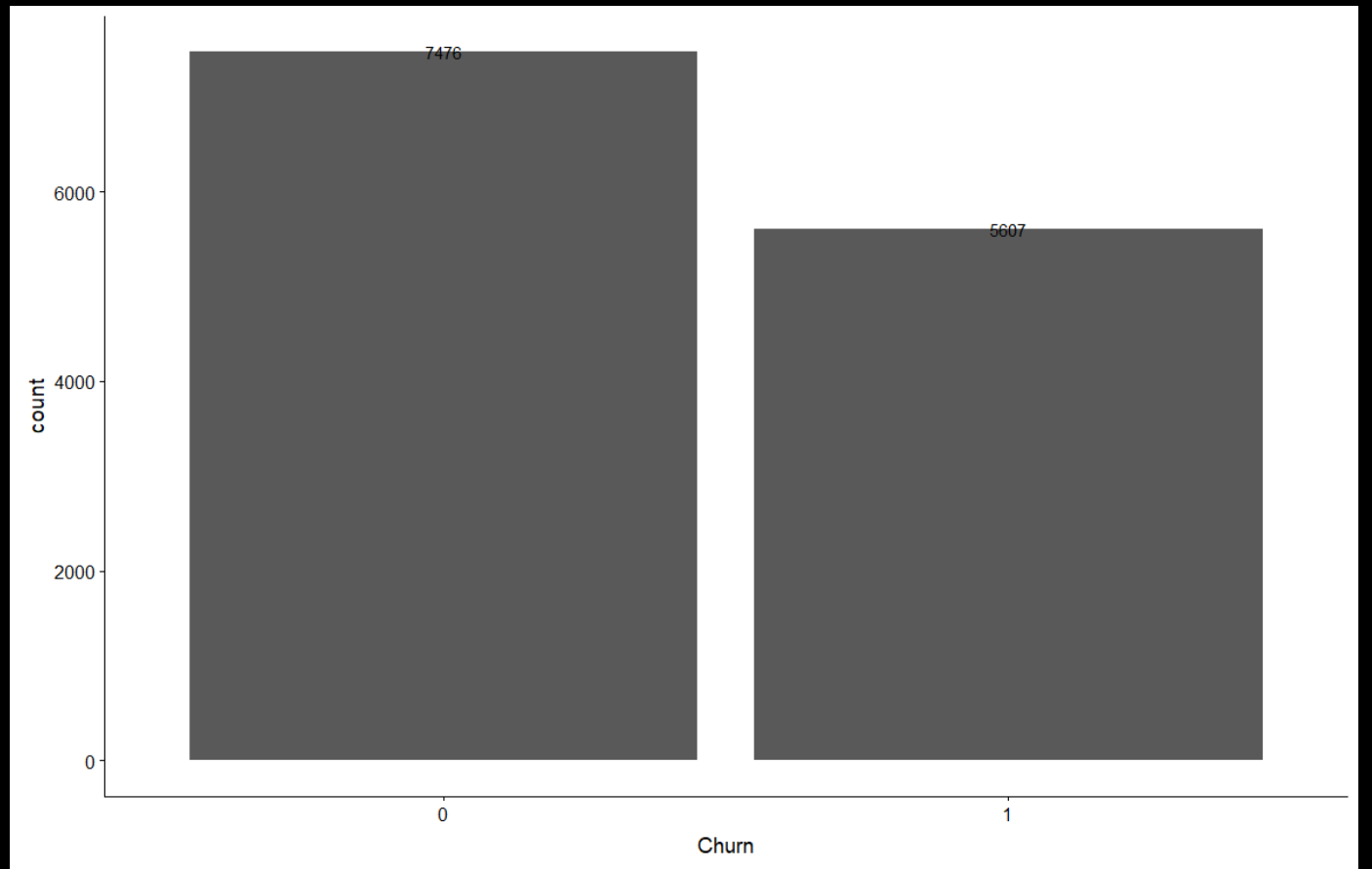
Handling Imbalanced With Oversampling/Undersampling

Oversampling adalah membuat data baru berdasarkan data yang telah ada untuk menyamakan proporsi dengan data yang majoritas, sedangkan undersampling adalah sebaliknya.



Oversampling Undersampling

```
#Oversampling Undersampling
telco_final = SMOTE(Churn~.,telco_final,k=5)
#tampilkan barplotnya
ggplot(telco_final, aes(x = Churn)) +
  geom_bar()+geom_text(aes(label=..count..),stat='count')
```



Splitting the data into train and validation set.

```
#Bagi Data Training dan Testing
indices = sample.split(telco_final$Churn, SplitRatio = 0.8)
train = telco_final[indices,]
test = telco_final[!(indices),]
```

Model 1

```
#Buat Model 1
model1 <- glm(*****~.,train,family = binomial)
summary(model1)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.53418	0.25573	-2.089	0.036721	*
tenure	-0.81276	0.17250	-4.712	2.46e-06	***
MonthlyCharges	-0.13605	0.03092	-4.400	1.08e-05	***
TotalCharges	0.08502	0.03008	2.826	0.004713	**
gender	-0.02414	0.05093	-0.474	0.635541	
SeniorCitizen	0.01628	0.06828	0.238	0.811545	
Partner	-0.18140	0.06121	-2.963	0.003043	**
Dependents	-0.18628	0.07073	-2.634	0.008450	**
PhoneService	-0.36872	0.10350	-3.562	0.000368	***
MultipleLines	0.09114	0.06157	1.480	0.138790	
InternetService.xFiber.optic	0.99121	0.07615	13.017	< 2e-16	***
InternetService.xNo	-1.21446	0.10739	-11.309	< 2e-16	***
OnlineSecurity	-0.56788	0.06607	-8.595	< 2e-16	***
OnlineBackup	-0.36010	0.05982	-6.020	1.74e-09	***
DeviceProtection	-0.20713	0.06277	-3.300	0.000967	***
TechSupport	-0.52645	0.06801	-7.741	9.85e-15	***
StreamingTV	0.28136	0.06575	4.279	1.88e-05	***
StreamingMovies	0.16730	0.06459	2.590	0.009590	**
Contract.xOne.year	-1.25388	0.08066	-15.544	< 2e-16	***
Contract.xTwo.year	-2.21393	0.12779	-17.324	< 2e-16	***
PaperlessBilling	0.33102	0.05873	5.636	1.74e-08	***
PaymentMethod.xCredit.card..automatic.	-0.03003	0.08909	-0.337	0.736027	
PaymentMethod.xElectronic.check	0.48504	0.07444	6.516	7.22e-11	***
PaymentMethod.xMailed.check	0.30970	0.08952	3.460	0.000541	***
tenure_bin.x1.2.years	-0.02170	0.12214	-0.178	0.859009	
tenure_bin.x2.3.years	0.30620	0.21016	1.457	0.145117	
tenure_bin.x3.4.years	0.48747	0.29062	1.677	0.093479	.
tenure_bin.x4.5.years	0.81332	0.37468	2.171	0.029952	*
tenure_bin.x5.6.years	1.41106	0.47058	2.999	0.002713	**

Feature Selection

Masih Terdapat Variabel yang tidak signifikan, Selanjutnya dilakukan pemilihan feature. Pemilihan feature adalah memilih feature yang terbaik atau yang terpenting yang digunakan dalam model. Banyak sekali metode dalam pemilihan feature, pada kasus ini, kita menggunakan stepAIC atau stepwise yang dikoreksi menggunakan nilai AIC. Nilai AIC adalah nilai kebaikan model dimana feature didalam model tersebut berbeda dengan model yang akan dibandingkan.

Feature Selection

```
#Pemilihan Feature dengan stepaic
model2 <- stepAIC(model1,direction="both")
```

```
Churn ~ tenure + MonthlyCharges + TotalCharges + Partner +
Dependents + PhoneService + MultipleLines +
InternetService.xFiber.optic + InternetService.xNo +
OnlineSecurity + OnlineBackup + DeviceProtection + TechSupport +
StreamingTV + StreamingMovies + Contract.xOne.year +
Contract.xTwo.year + PaperlessBilling +
PaymentMethod.xElectronic.check + PaymentMethod.xMailed.check +
tenure_bin.x2.3.years + tenure_bin.x3.4.years +
tenure_bin.x4.5.years + tenure_bin.x5.6.years
```

Bentuk Model 2

```
#Model 2
model2 <- glm(Churn ~ tenure + MonthlyCharges + TotalCharges + Partner + Dependents +
          PhoneService + MultipleLines + InternetService.xFiber.optic +
          InternetService.xNo + OnlineSecurity + OnlineBackup + DeviceProtection +
          TechSupport + StreamingTV + StreamingMovies + Contract.xOne.year +
          Contract.xTwo.year + PaperlessBilling + PaymentMethod.xElectronic.check +
          PaymentMethod.xMailed.check + tenure_bin.x2.3.years + tenure_bin.x3.4.years +
          tenure_bin.x4.5.years + tenure_bin.x5.6.years,train,family=binomial)
summary(model2)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.59275	0.15775	-3.757	0.000172	***
tenure	-0.83771	0.10678	-7.845	4.32e-15	***
MonthlyCharges	-0.13616	0.03085	-4.413	1.02e-05	***
TotalCharges	0.08518	0.02989	2.850	0.004370	**
Partner	-0.17961	0.06061	-2.964	0.003041	**
Dependents	-0.18871	0.06926	-2.725	0.006438	**
PhoneService	-0.37075	0.10319	-3.593	0.000327	***
MultipleLines	0.09194	0.06142	1.497	0.134385	
InternetService.xFiber.optic	0.99388	0.07575	13.120	< 2e-16	***
InternetService.xNo	-1.21591	0.10722	-11.340	< 2e-16	***
OnlineSecurity	-0.56784	0.06600	-8.603	< 2e-16	***
OnlineBackup	-0.36054	0.05977	-6.032	1.62e-09	***
DeviceProtection	-0.20727	0.06275	-3.303	0.000957	***
TechSupport	-0.52746	0.06781	-7.779	7.32e-15	***
StreamingTV	0.28050	0.06571	4.269	1.96e-05	***
StreamingMovies	0.16807	0.06452	2.605	0.009184	**
Contract.xOne.year	-1.25624	0.08048	-15.609	< 2e-16	***
Contract.xTwo.year	-2.21632	0.12761	-17.368	< 2e-16	***
PaperlessBilling	0.33098	0.05867	5.642	1.68e-08	***
PaymentMethod.xElectronic.check	0.49967	0.06031	8.286	< 2e-16	***
PaymentMethod.xMailed.check	0.32405	0.07785	4.162	3.15e-05	***
tenure_bin.x2.3.years	0.33760	0.12002	2.813	0.004911	**
tenure_bin.x3.4.years	0.53097	0.16713	3.177	0.001488	**
tenure_bin.x4.5.years	0.86947	0.21914	3.968	7.26e-05	***
tenure_bin.x5.6.years	1.48083	0.27566	5.372	7.79e-08	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Bentuk Model 3

```
#Model 3
model3 <- glm(Churn ~ tenure + MonthlyCharges + TotalCharges + Partner + Dependents +
              PhoneService  + InternetService.xFiber.optic +
              InternetService.xNo + OnlineSecurity + OnlineBackup + DeviceProtection +
              TechSupport + StreamingTV + StreamingMovies + Contract.xOne.year +
              Contract.xTwo.year + PaperlessBilling + PaymentMethod.xElectronic.check +
              PaymentMethod.xMailed.check + tenure_bin.x2.3.years + tenure_bin.x3.4.years +
              tenure_bin.x4.5.years + tenure_bin.x5.6.years,train,family='binomial')
summary(model3)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.60561	0.15745	-3.846	0.000120	***
tenure	-0.83383	0.10672	-7.813	5.58e-15	***
MonthlyCharges	-0.13944	0.03075	-4.534	5.78e-06	***
TotalCharges	0.08495	0.02988	2.843	0.004474	**
Partner	-0.17399	0.06047	-2.877	0.004008	**
Dependents	-0.19354	0.06919	-2.797	0.005152	**
PhoneService	-0.33844	0.10083	-3.356	0.000789	***
InternetService.xFiber.optic	1.01721	0.07426	13.698	< 2e-16	***
InternetService.xNo	-1.22825	0.10691	-11.489	< 2e-16	***
OnlineSecurity	-0.56766	0.06600	-8.601	< 2e-16	***
OnlineBackup	-0.35400	0.05960	-5.939	2.86e-09	***
DeviceProtection	-0.20543	0.06273	-3.275	0.001058	**
TechSupport	-0.53042	0.06777	-7.827	5.01e-15	***
StreamingTV	0.28251	0.06570	4.300	1.71e-05	***
StreamingMovies	0.17397	0.06439	2.702	0.006895	**
Contract.xOne.year	-1.25475	0.08045	-15.596	< 2e-16	***
Contract.xTwo.year	-2.20687	0.12741	-17.321	< 2e-16	***
PaperlessBilling	0.33619	0.05859	5.738	9.60e-09	***
PaymentMethod.xElectronic.check	0.49878	0.06029	8.273	< 2e-16	***
PaymentMethod.xMailed.check	0.31952	0.07778	4.108	3.99e-05	***
tenure_bin.x2.3.years	0.34537	0.11983	2.882	0.003951	**
tenure_bin.x3.4.years	0.54062	0.16695	3.238	0.001203	**
tenure_bin.x4.5.years	0.87329	0.21905	3.987	6.70e-05	***
tenure_bin.x5.6.years	1.48242	0.27559	5.379	7.49e-08	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Model 3 sudah baik, sekarang lakukan perhitungan akurasi

Hitung Ketepatan Akurasi Dengan Model final /Model 3

```
#Model Final
final_model = model3
#Hitung Prediksi atau Yhat
pred = predict(final_model,type='response',newdata = test[,-24])
#buat feature pred pada data test
test$prob = pred
#Lakukan prediksi berdasarkan cutoff 0.5
pred_churn = factor(ifelse(pred>=0.50,"Yes","No"))
#Buat data actual churn
actual_churn = factor(ifelse(test$Churn==1,"Yes","No"))
#buat table akurasi/confusion matrix
t = table(pred_churn,actual_churn)
#tampilkan table
t
```

```
           actual_churn
pred_churn   No  Yes
       No  1169  262
       Yes  326  859
```

Confusion Matrix

```
#Hitung Cutoff
cutoff_churn = factor(ifelse(pred >=0.50, "Yes",
"No"))
#buat confusion matrix
con_final =
confusionMatrix(cutoff_churn,actual_churn,positive
= 'Yes')
#tampilkan confusion matrix
con_final
```

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1169  262
       Yes  326  859

               Accuracy : 0.7752
                 95% CI : (0.7587, 0.7911)
    No Information Rate : 0.5715
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5443
 Mcnemar's Test P-Value : 0.009375

            Sensitivity : 0.7663
            Specificity : 0.7819
         Pos Pred Value : 0.7249
         Neg Pred Value : 0.8169
             Prevalence : 0.4285
         Detection Rate : 0.3284
   Detection Prevalence : 0.4530
      Balanced Accuracy : 0.7741

       'Positive' Class : Yes
```