Customer Segmentation

Wittawat Muangkot

2022-04-15

this paper will analyze mall customer dataset and do the customer segmentation
to find best custer for this dataset by refer elbow chart

First let start with import nescessary package using tiddyverse

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Next load dataset Mall_Customers into df variable and show head row of dataset for basic explore

```
df<- read_csv('Mall_Customers.xls')
```

```
## Rows: 200 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Gender
## dbl (4): CustomerID, Age, Annual Income (k$), Spending Score (1-100)
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(df)
```

```
## # A tibble: 6 x 5
##   CustomerID Gender   Age `Annual Income (k$)` `Spending Score (1-100)`
##        <dbl> <chr>  <dbl>                <dbl>                     <dbl>
## 1          1 Male      19                   15                        39
## 2          2 Male      21                   15                        81
## 3          3 Female    20                   16                         6
## 4          4 Female    23                   16                        77
## 5          5 Female    31                   17                        40
## 6          6 Female    22                   17                        76
```

Rename some columns name (Annual Income, Spending Score)

to easy spelling and remove columns customer ID dueto values with ununique and cannot use for build model

```
df<- df%>%
  mutate(Annual_Income=df$`Annual Income (k$)`,
         Spending_Score=df$`Spending Score (1-100)`)
df<-df[,c(2,3,6,7)]
```

Transform Gender columns into factor and change to binary values

```
df<- df%>%
  mutate(Gender= factor(Gender, levels = c("Male","Female"),labels= c("1","0")))
```

Let review dataset again and check missing values

```
glimpse(df)
```

```
## Rows: 200
## Columns: 4
## $ Gender         <fct> 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0, 1, 1~
## $ Age            <dbl> 19, 21, 20, 23, 31, 22, 35, 23, 64, 30, 67, 35, 58, 24,~
## $ Annual_Income  <dbl> 15, 15, 16, 16, 17, 17, 18, 18, 19, 19, 19, 19, 20, 20,~
## $ Spending_Score <dbl> 39, 81, 6, 77, 40, 76, 6, 94, 3, 72, 14, 99, 15, 77, 13~
```

```
print("check missing values")
```

```
## [1] "check missing values"
```

```
mean(complete.cases(df))
```

```
## [1] 1
```

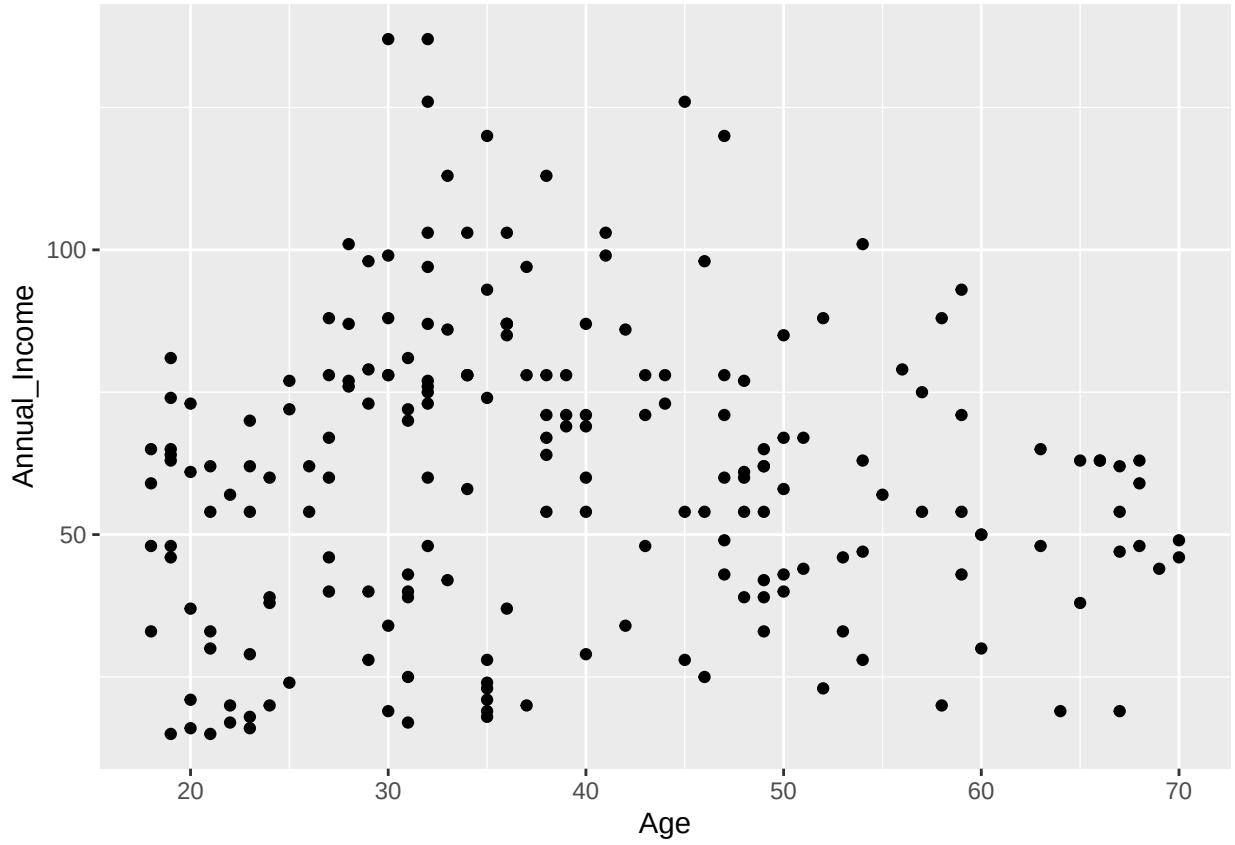
```
print("result 1 is No missing values in dataset")
```

```
## [1] "result 1 is No missing values in dataset"
```

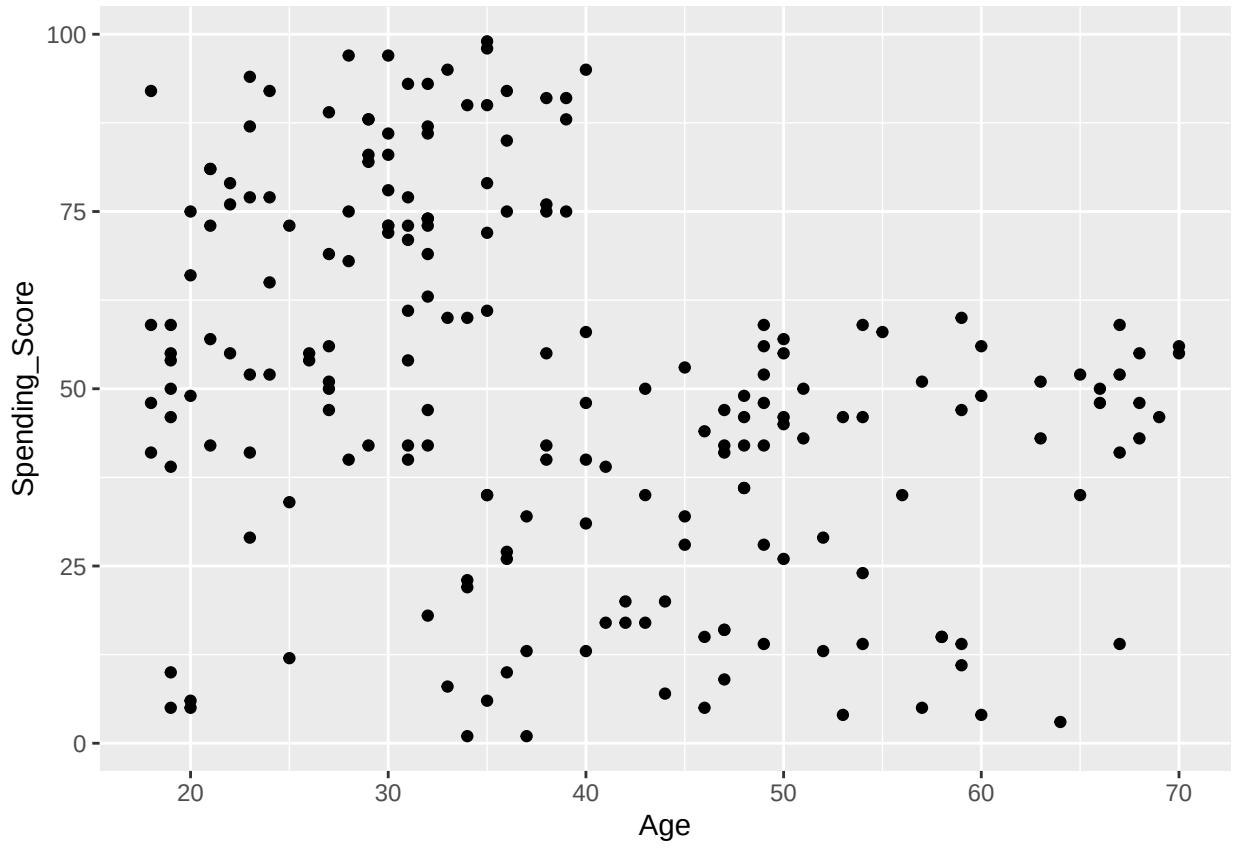
Explore data by visualization in ggplot

Let check data pattern by using scatter plot with 2 continuous value

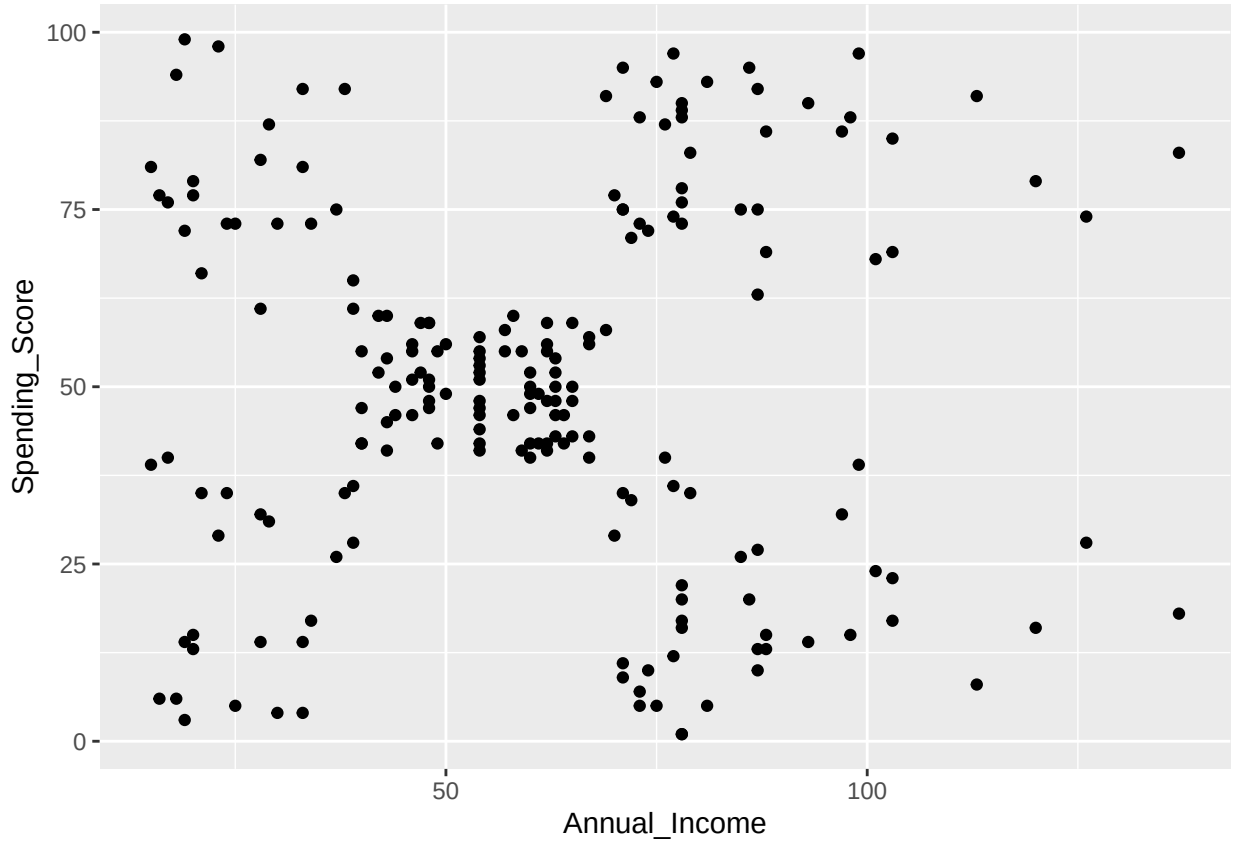
```
ggplot(df, aes(Age, Annual_Income))+
  geom_point()
```




```
ggplot(df, aes(Age, Spending_Score))+
  geom_point()
```

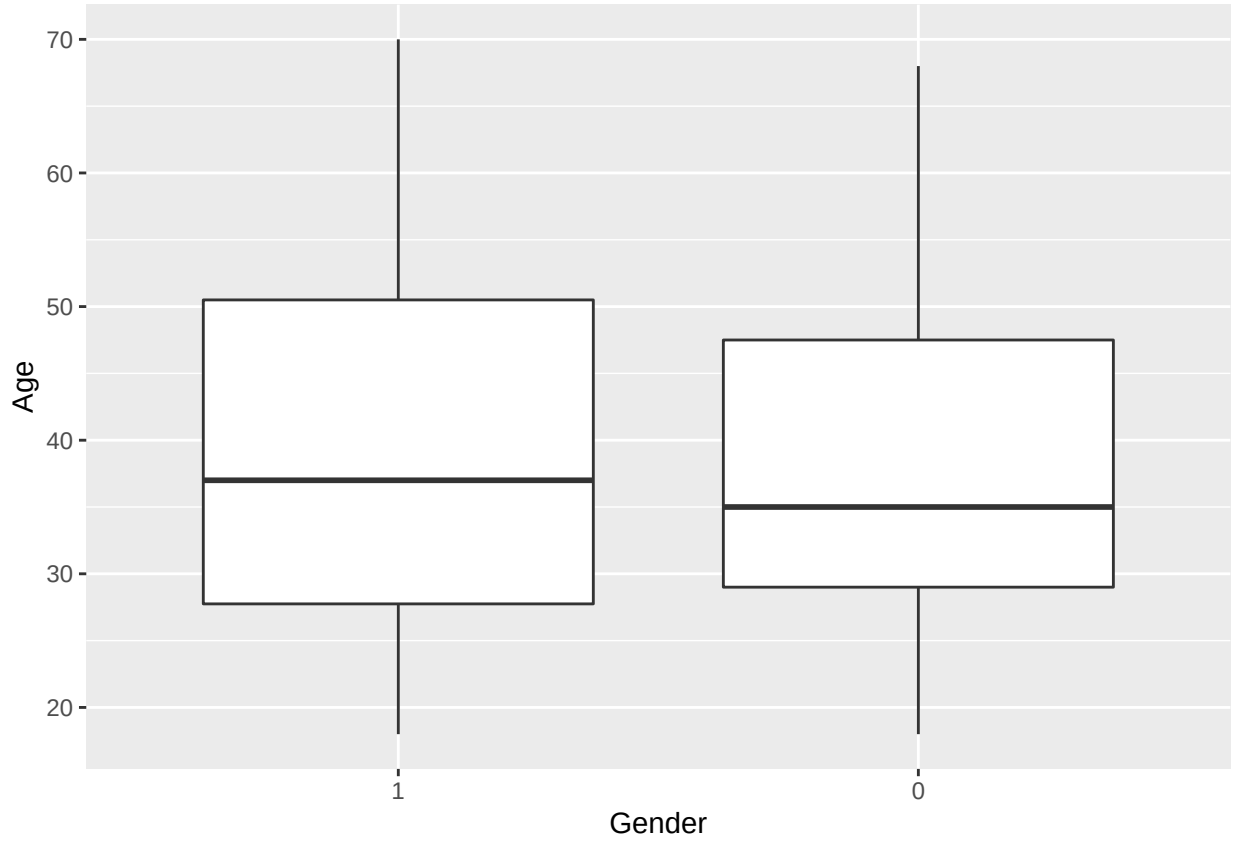



```
ggplot(df, aes(Annual_Income, Spending_Score))+
  geom_point()
```

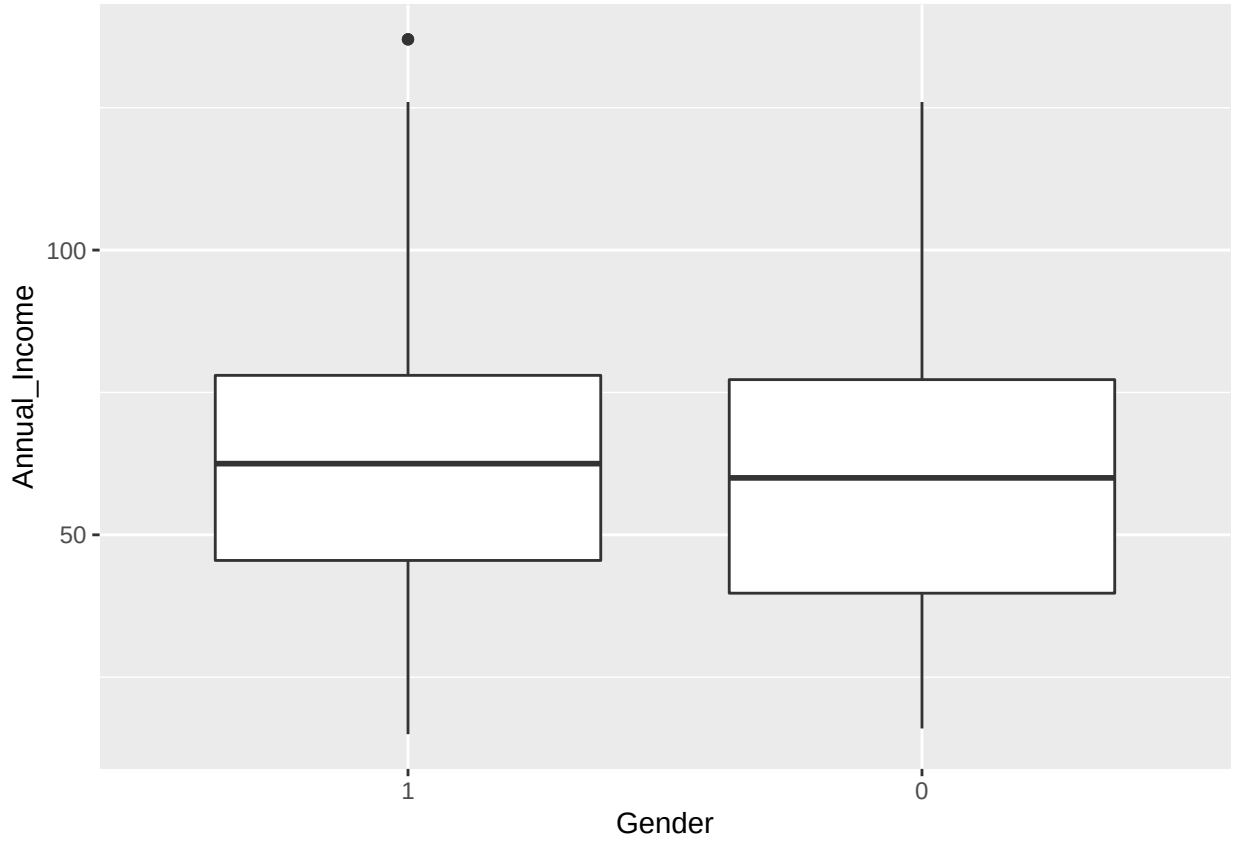


As we can see some pattern of group from Annual_Income and Spending_Score Next let compare group of continuous value by descrete values using boxplot chart

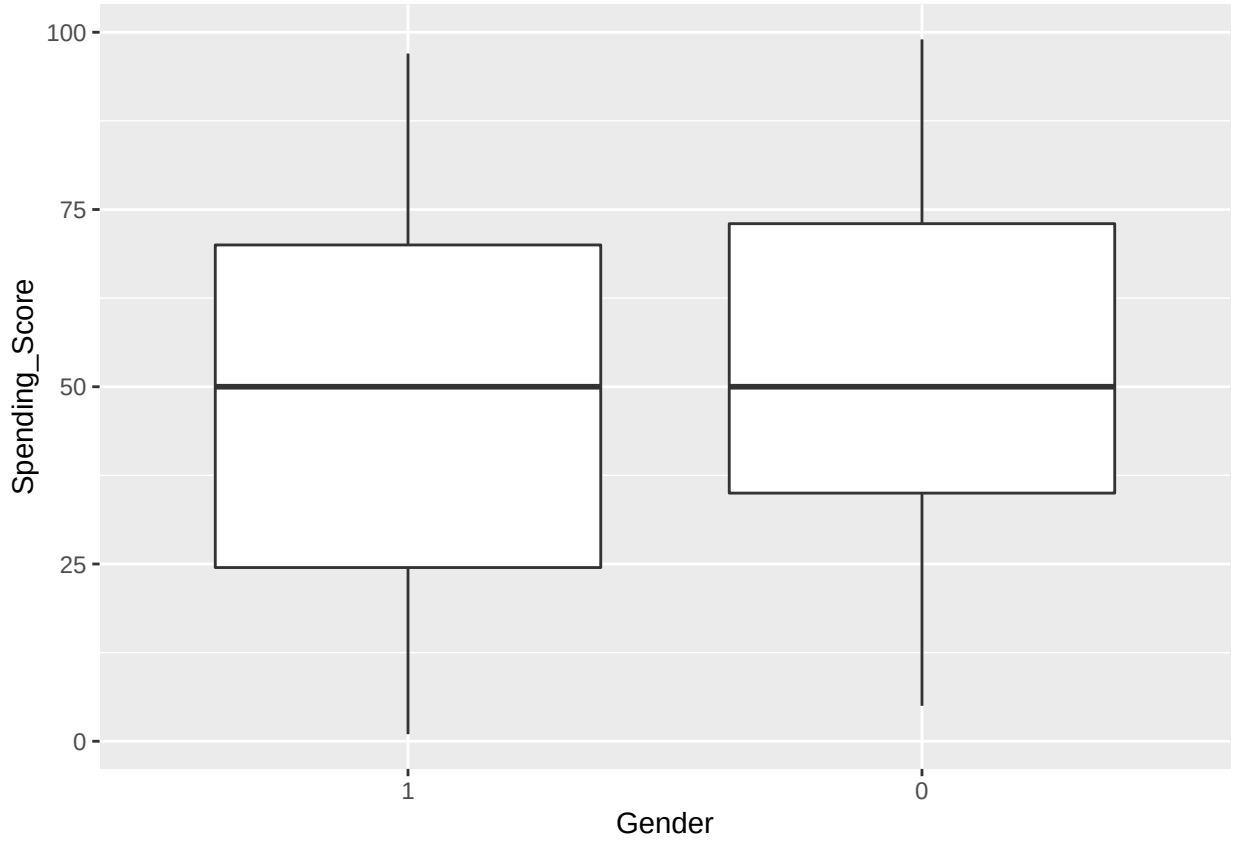
```
ggplot(df, aes(Gender, Age))+
  geom_boxplot()
```




```
ggplot(df, aes(Gender, Annual_Income))+
  geom_boxplot()
```



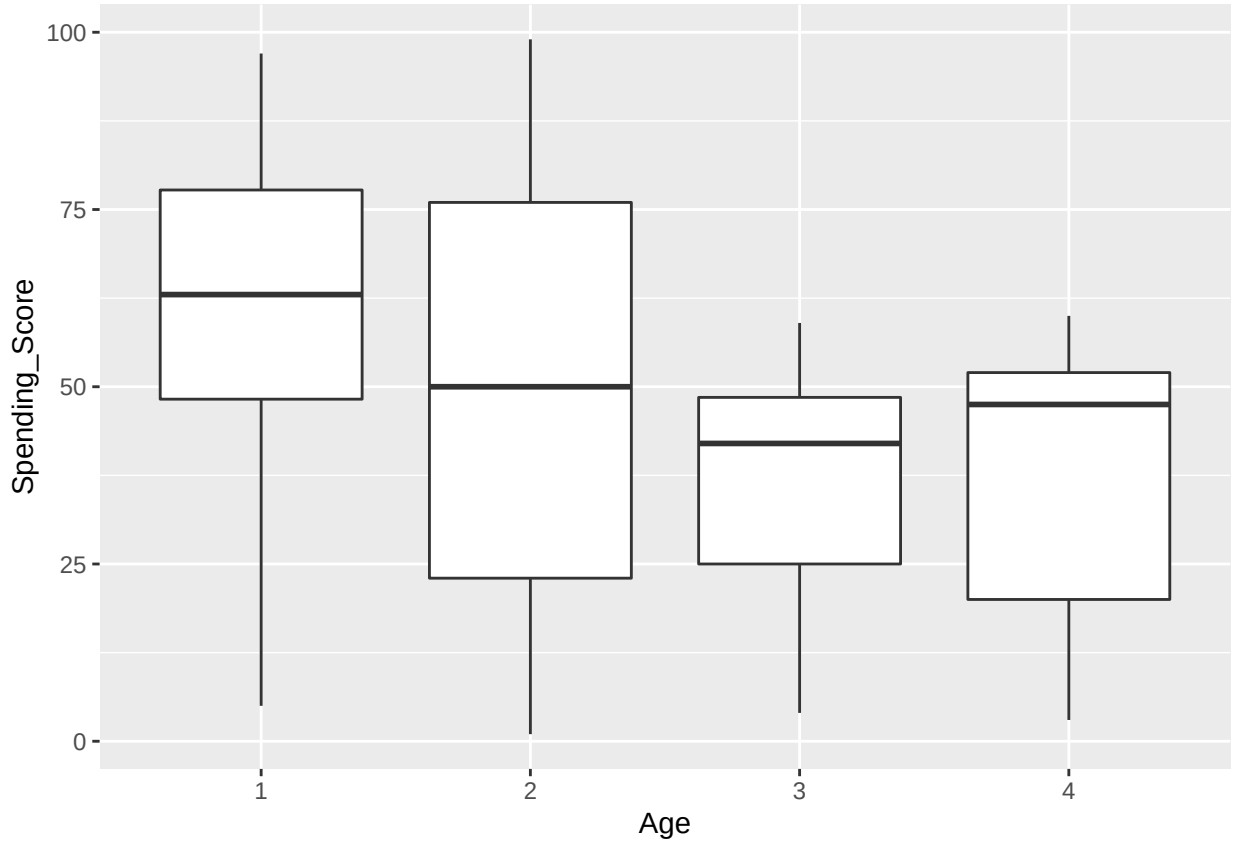

```
ggplot(df, aes(Gender, Spending_Score))+
  geom_boxplot()
```



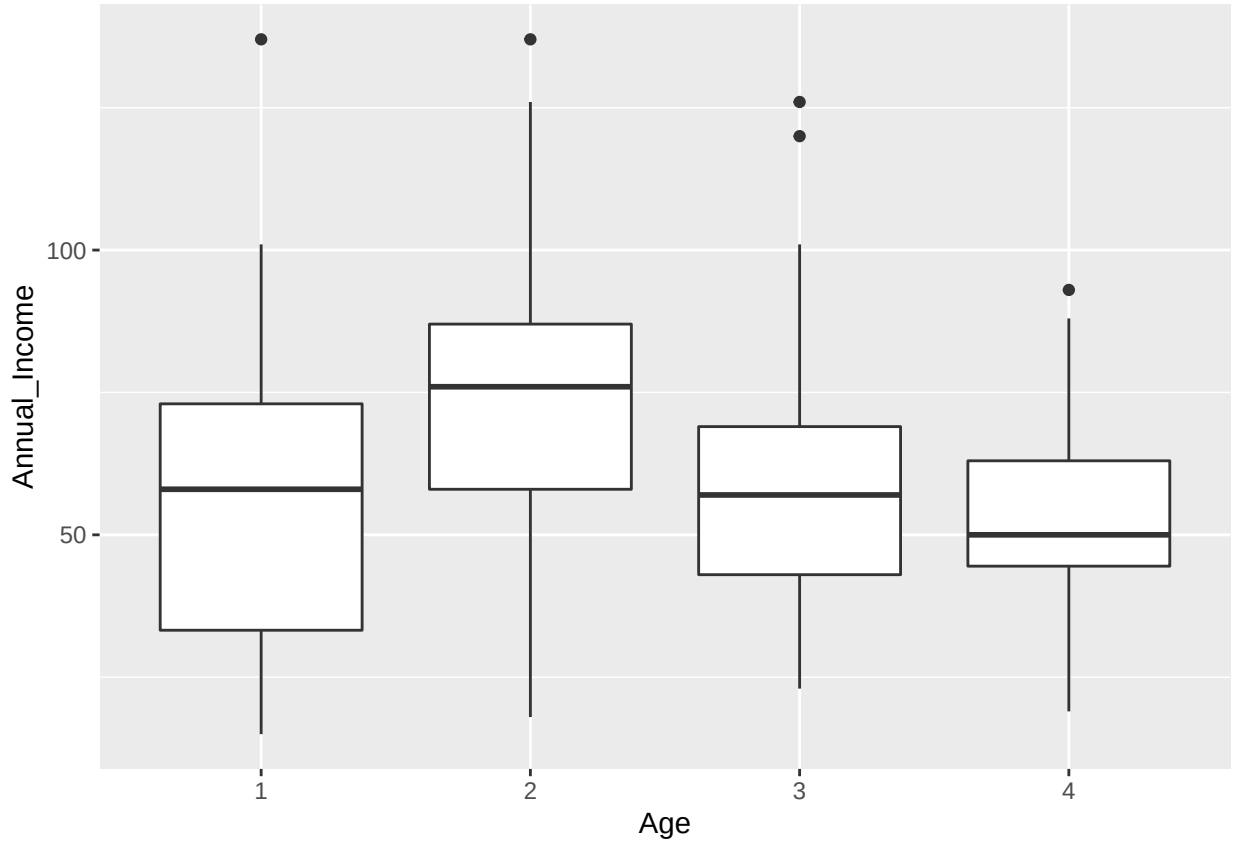
After we compare in values in group of gender, there are no significant difference in each group and we will next descrete age columns for more explore data

```
df<- df%>%
  mutate(Age= cut(df$Age, breaks = 4, labels = FALSE ))
df<- df%>%
  mutate_if(is.integer,as.factor)

ggplot(df, aes(Age, Spending_Score))+
  geom_boxplot()
```




```
ggplot(df, aes(Age, Annual_Income))+
  geom_boxplot()
```



After we descrete age columns, we notice that there are some difference in Annual_Income ane Spending_score by each age grouop

Fit data to Kmean model by using varies k to find best k

K -> 2,3,4. . . ,10

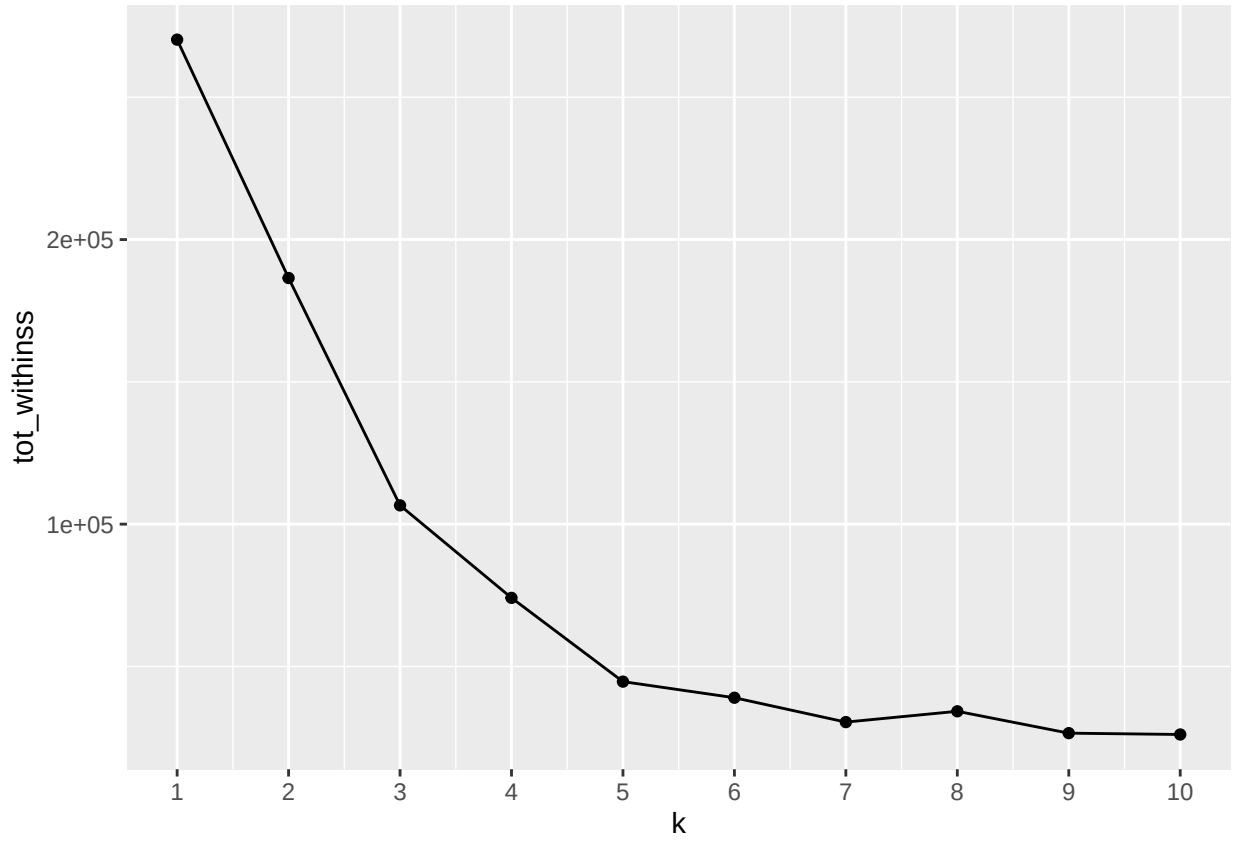
```
set.seed(42)
tot_withinss <- map_dbl(1:10,  function(k){
  model <- kmeans(x = df, centers = k)
  model$tot.withinss
})
```

we will using elbow chart to select the best k

```
elbow_df <- data.frame(
  k = 1:10,
  tot_withinss = tot_withinss
)
```

Plot the elbow plot by ggplot

```
ggplot(elbow_df, aes(x = k, y = tot_withinss)) +
  geom_line() + geom_point()+
  scale_x_continuous(breaks = 1:10)
```



notice that the best k on the elbow chart is 5

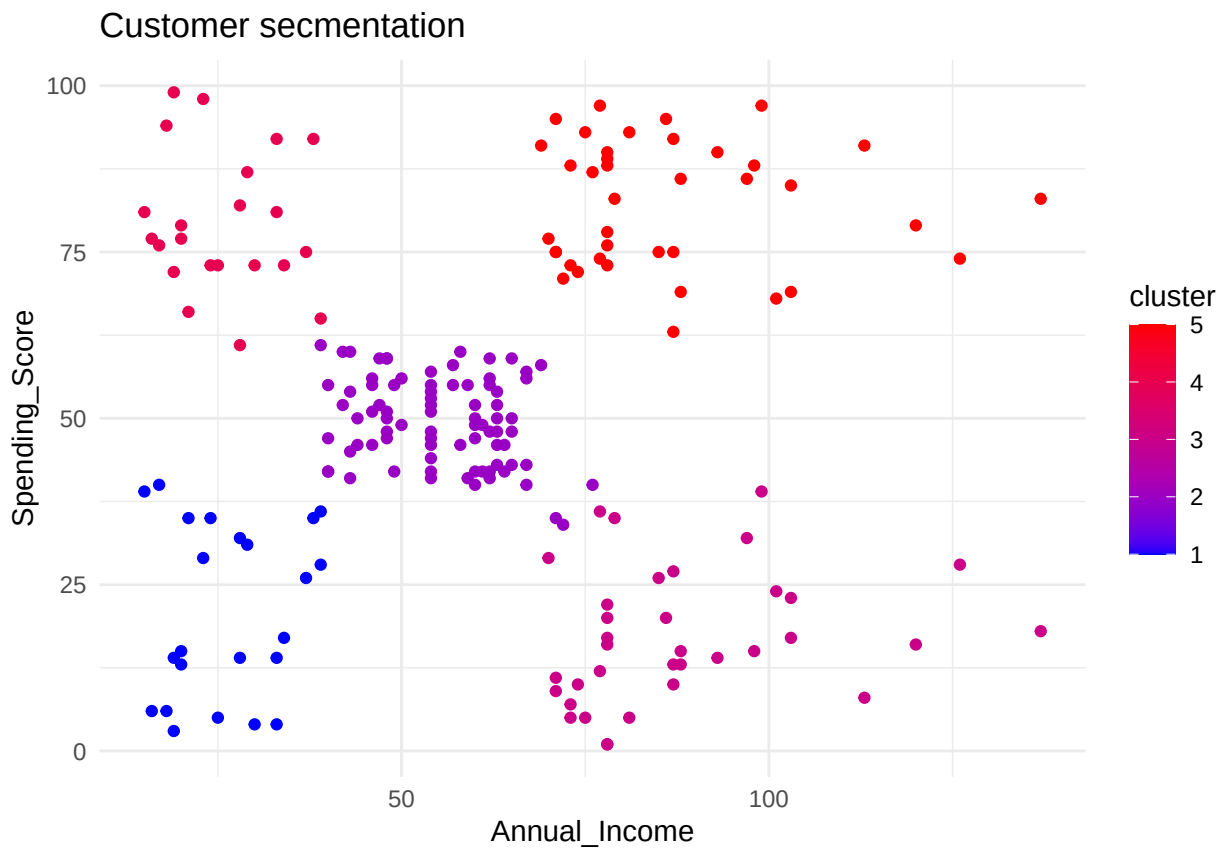
Select the best k for this model (k=5) and set seed to control random data

```
set.seed(42)
model<- kmeans(df, centers = 5)

df$cluster<-model$cluster
```

after completed fit data to model, we'll show result by plot scatter plot x=Annual_Income y=Spending_score and show difference cluster by customer segment

```
ggplot(df, aes(Annual_Income, Spending_Score, col=cluster))+
  geom_point()+
  theme_minimal()+
  scale_color_gradient(low="blue", high="red")+
  labs(title='Customer secmentation',
       x='Annual_Income',
       y='Spending_Score')
```



##Conclusion we separate group of data Annual_Income and Spring_Score by using each color for each cluster and notice that in k = 5, it can represent best in each group