Loading Data:

In [1]:

```
df = read.csv("toy_dataset.csv")
```

Gist of dataset:

In [2]:

```
head(df)
```

Number	Gender	Age	Income
1	Male	41	40367
2	Male	54	45084
3	Male	42	52483
4	Male	40	40941
5	Male	46	50289
6	Female	36	50786

Summary:

In [3]:

```
summary(df)

sum(is.na(df)) #Checking NULL Values
nrow(df)   #Calculating total number of observations
ncol(df)   #Calculating total number of attributes
```

```
     Number          Gender           Age            Income
 Min.   :     1   Female:66200   Min.   :25.00   Min.   :  -654
 1st Qu.: 37501   Male  :83800   1st Qu.:35.00   1st Qu.: 80868
 Median : 75001                  Median :45.00   Median : 93655
 Mean   : 75001                  Mean   :44.95   Mean   : 91253
 3rd Qu.:112500                  3rd Qu.:55.00   3rd Qu.:104519
 Max.   :150000                  Max.   :65.00   Max.   :177157
```

0

150000

4

Data Analysis:

Age and Gender Count:

In [4]:

```
count_age <- table(df$Age)
count_age
count_gender <- table(df$Gender)
count_gender
```

```
  25   26   27   28   29   30   31   32   33   34   35   36   37   38   39   40
1868 3918 3790 3798 3805 3731 3749 3759 3769 3710 3658 3780 3771 3734 3675 3740
  41   42   43   44   45   46   47   48   49   50   51   52   53   54   55   56
3862 3760 3862 3782 3670 3707 3773 3743 3762 3692 3729 3753 3724 3838 3687 3602
  57   58   59   60   61   62   63   64   65
3732 3762 3775 3582 3737 3684 3784 3879 1864
```

```
Female   Male
 66200  83800
```

Calculating overall Mean Income:

In [5]:

```
mean <- mean(df$Income)
mean
```

91252.7982733333

Calculating Mean Income according to the Age:

In [6]:

```
sort_age <- sort(unique(df$Age))

mean_income <- c()
for (i in sort_age){
  temp_df <- df[df$Age == i,]
  mean_income <- c(mean_income,mean(temp_df$Income))
}

df_ageincome <- data.frame("Age" = sort_age, "Mean_Income" = mean_income)
df_ageincome
```

Age	Mean_Income
25	91164.57
26	90883.18
27	91554.38
28	91829.50
29	90913.65
30	90957.50
31	91652.70
32	91318.67
33	91825.69
34	91266.26
35	91465.11
36	91192.68
37	91724.96
38	90827.78
39	91243.82
40	91431.08
41	90481.77
42	90897.55
43	90791.18
44	91244.34
45	91403.92
46	91672.58
47	91166.40
48	91392.80
49	90894.22
50	91768.77
51	90922.68
52	91341.80
53	91899.09
54	91892.17
55	91102.05
56	91143.79
57	91041.44
58	91035.45
59	90678.51
60	90657.31
61	91111.54
62	91940.22
63	90901.48
64	91731.44
65	90696.43

Calculating Sum of Income according to the Age:

In [7]:

```
sort_age <- sort(unique(df$Age))

sum_income <- c()
for (i in sort_age){
  temp_df <- df[df$Age == i,]
  sum_income <- c(sum_income,sum(temp_df$Income))
}

df_agesincome <- data.frame("Age" = sort_age, "Sum_Income" = sum_income)
df_agesincome
```

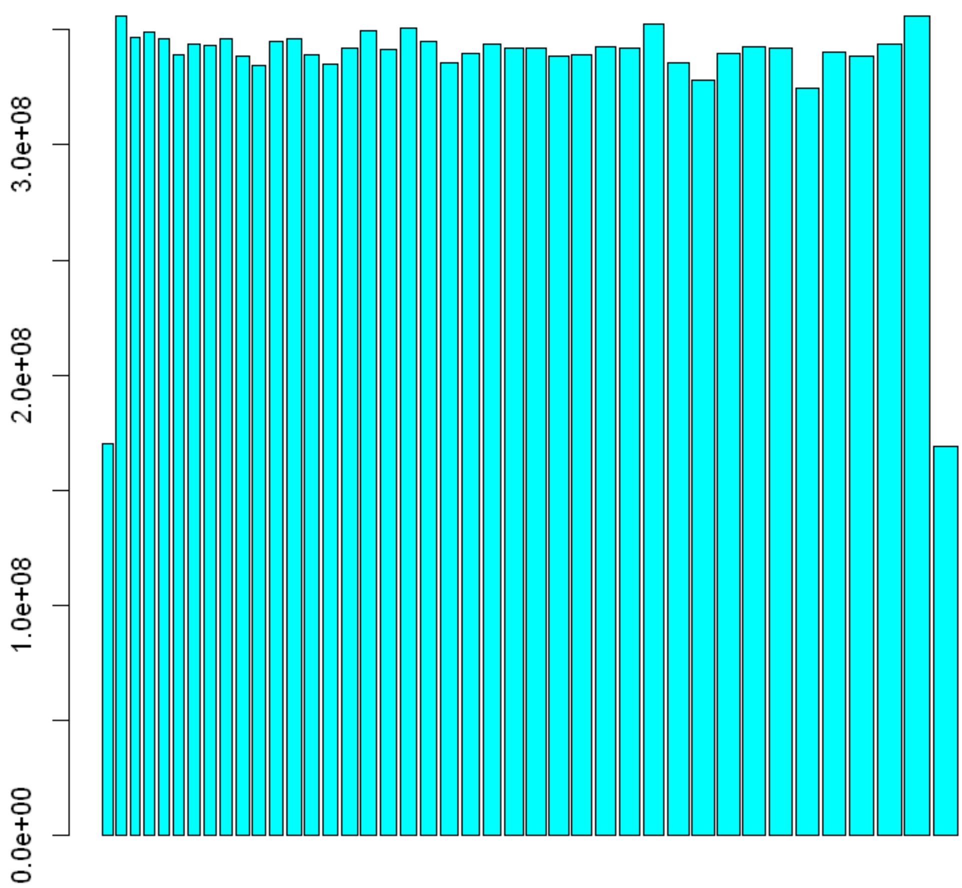
Age	Sum_Income
25	170295410
26	356080301
27	346991099
28	348768431
29	345926448
30	339362414
31	343605954
32	343266869
33	346091026
34	338597810
35	334579362
36	344708315
37	345894825
38	339150928
39	335321049
40	341952227
41	349440607
42	341774789
43	350635555
44	345086104
45	335452375
46	339830269
47	343970840
48	342083235
49	341944058
50	338810307
51	339050677
52	342805774
53	342232228
54	352682137
55	335893260
56	328299947
57	339766664
58	342475349
59	342311384
60	324734495
61	340483842
62	338707782
63	343971200
64	355826250
65	169058145

Visualization:

Plotting Sum of Income vs Age

In [8]:

```
barplot(df_agesincome$Sum_Income, df_agesincome$Age, col = "Cyan")
```



Calculating Mean Income according to the Gender:

In [9]:

```
unique_gender <- unique(df$Gender)

mean_income <- c()
for (i in unique_gender){
  temp_df <- df[df$Gender == i,]
  mean_income <- c(mean_income,mean(temp_df$Income))
}

df_genderincome <- data.frame("Gender" = unique_gender, "Mean_Income" = mean_income)
df_genderincome
```

Gender	Mean_Income
Male	95670.25
Female	85660.92

Calculating Sum of Income according to the Gender:

In [10]:

```
unique_gender <- unique(df$Gender)

sum_income <- c()
for (i in unique_gender){
  temp_df <- df[df$Gender == i,]
  sum_income <- c(sum_income,sum(temp_df$Income))
}

df_gendersincome <- data.frame("Gender" = unique_gender, "Sum_Income" = sum_income)
df_gendersincome
```

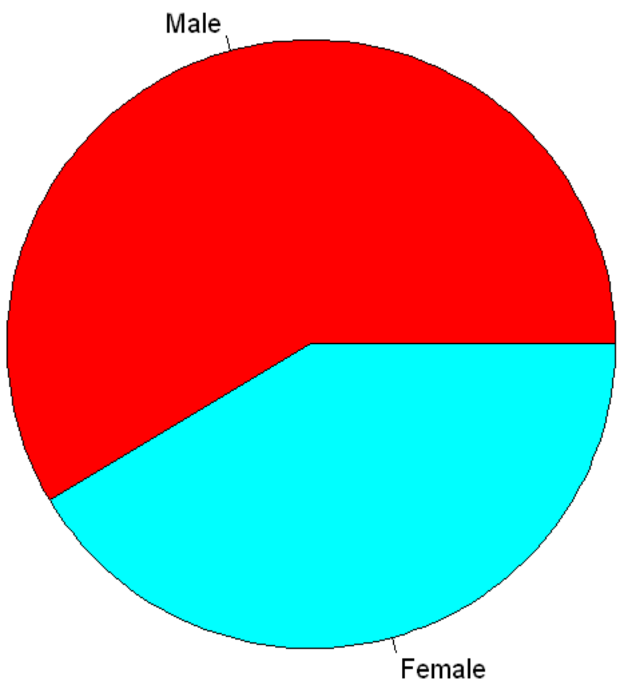
Gender	Sum_Income
Male	8017166715
Female	5670753026

Visualization:

Plotting Sum of Income vs Gender

In [11]:

```
pie(df_gendersincome$Sum_Income, df_gendersincome$Gender, col = rainbow(2))
```



Random sampling according to the Gender:

In [12]:

```
df_male <- df[df$Gender == "Male", ]
df_female <- df[df$Gender == "Female", ]
```

In [13]:

```
sample_male <- df_male[sample(nrow(df_male), 30, replace = FALSE), ]
sample_female <- df_female[sample(nrow(df_female), 30, replace = FALSE), ]
```

Assumptions:

1.Simple random sample, that the data is collected from a representative, randomly selected portion of the total population.

2.When plotted, results in a normal distribution, bell-shaped distribution curve.

3.Sample size should be less than 30. Here we take sample of 30.

4.Homogeneity of variance. Homogeneous, or equal, variance exists.

5.Here we predetermine Level of Significance as 5% or 0.05.

Assumption 2:To check whether our data sets are normally distributed or not.

To check whether data is normally distributed or not we use Shapiro-Wilk test.

If the p-value is greater than the level of significance i.e. 0.05 we can assume that the given data is normally distributed.

After checking that data is normally distibuted we plot it's graph to see whether it is bell shaped distibution curve or not using dnorm(x,mean(x),sd(x)).

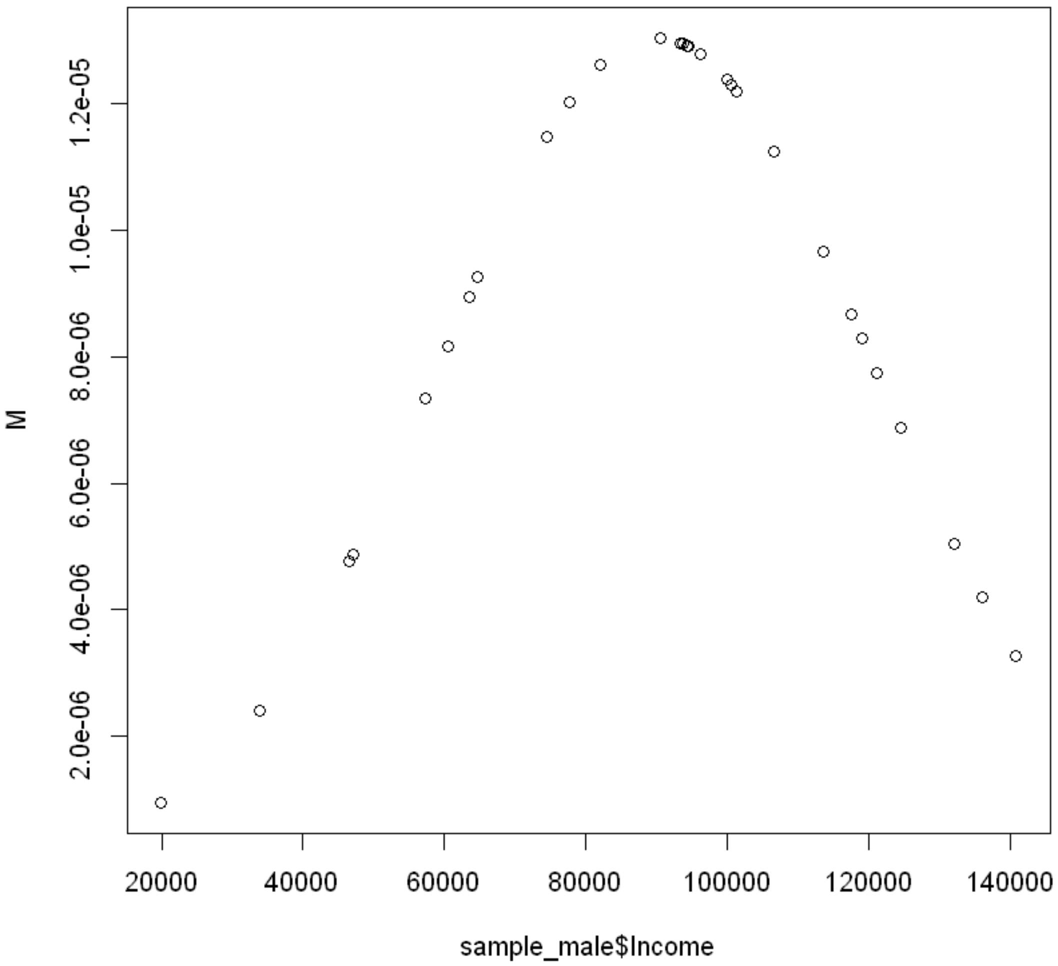
For Male:

In [14]:

```
shapiro.test(sample_male$Income)
M<-dnorm(sample_male$Income,mean(sample_male$Income),sd(sample_male$Income))
plot(sample_male$Income,M)
```

```
	Shapiro-Wilk normality test

data:  sample_male$Income
W = 0.96968, p-value = 0.5304
```



Conclusion:

Here we conclude that our data for Male is normally distributed. As the p-value is greater than level of significance i.e. 0.05 and the graph shows the bell shaped distribution.

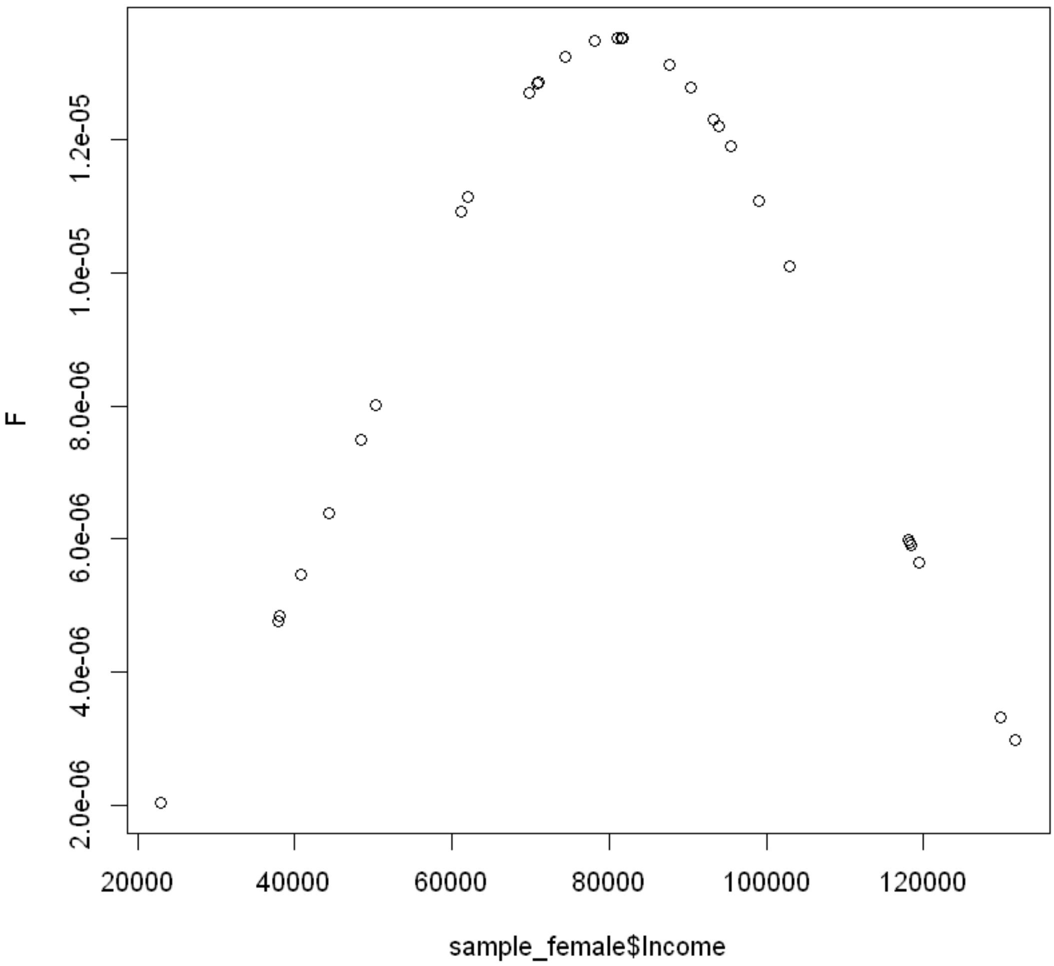
For Female:

In [15]:

```
shapiro.test(sample_female$Income)
F<-dnorm(sample_female$Income,mean(sample_female$Income),sd(sample_female$Income))
plot(sample_female$Income,F)
```

```
        Shapiro-Wilk normality test

data:  sample_female$Income
W = 0.971, p-value = 0.5668
```



Conclusion:

Here we conclude that our data for Female is normally distributed. As the p-value is greater than level of significance i.e. 0.05 and the graph shows the bell shaped distribution.

Assumption 4:To check whether homogenity is present in data or not.

To check whether the variances are homogeneous i.e. equal or not we use F-test. If the p-value is greater than level of significance i.e. 0.05, we can assume that the variances of the two variables are equal.

In [56]:

```
var.test(sample_male$Income,sample_female$Income)
```

```
        F test to compare two variances

data:  sample_male$Income and sample_female$Income
F = 1.0758, num df = 29, denom df = 29, p-value = 0.8454
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5120272 2.2601822
sample estimates:
ratio of variances
          1.075767
```

Conclusion:

Homogenity is present in the given data sets. As the p-value of f-test is greater than the legel of significance i.e. 0.05.

t-test:

Setting of Hypothesis:

Null Hypothesis, Ho:There is no significant difference between the mean income of Male and Female i.e. no relation exists between the mean income of Male and Female.

Alternative Hypothesis, H1:There is a significant difference between the mean income of Male and Female i.e. some relation exists between the mean income of Male and Female.

Test Statistic:

Method 1: From Scratch

Calculating the mean of the samples:

In [17]:

```
mean_Male <- mean(sample_male$Income)
print(mean_Male)
mean_Female <- mean(sample_female$Income)
print(mean_Female)
```

```
[1] 90028.3
[1] 80429.2
```

Calculating the standard deviation of the samples:

In [18]:

```
sd_Male <- sd(sample_male$Income)
print(sd_Male)
sd_Female <-sd(sample_female$Income)
print(sd_Female)
```

```
[1] 30577.81
[1] 29481.34
```

Calculating t-value:

In [20]:

```
t_stat<-(mean_Male - mean_Female) / sqrt((sd_Male^2 / length(sample_male$Income)) + (sd_Female^2 / length(sample_female$Income)))
print(t_stat)
```

```
[1] 1.237812
```

Method 2: Direct Method

In [21]:

```
t <- t.test(sample_male$Income, sample_female$Income, alternate = "two.sided")
t
```

```
        Welch Two Sample t-test

data:  sample_male$Income and sample_female$Income
t = 1.2378, df = 57.923, p-value = 0.2208
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5924.447 25122.647
sample estimates:
mean of x mean of y
  90028.3   80429.2
```

Calculating p-value:

In [22]:

```
p_value_t <- t$p.value
p_value_t
```

0.220779851359298

Level of Significance:

Here we predetermine level of significance as 5% i.e. 0.05.

Degree of freedom:

df=n1+n2-2

where,

```
n1=Sample size of Male.

n2=Sample size of Female.
```

In [23]:

```
df<-length(sample_male$Income)+length(sample_female$Income)-2
print(df)
```

```
[1] 58
```

Critical value:

Tabulated value of t at 58 degrees of freedom at 5% level of significance is 1.672.

Decision:

Interpretation using t-value:

In [24]:

```
if (t_stat > 1.672){
    print("Here we Reject Null Hypothesis as calculated t-value is greater than the tabulated t-value i.e. there is some relation exist be
} else{
    print("Here we Accept Null Hypothesis as calculated t-value is less than the tabulated t-value i.e. there is no relation exist between
}
```

```
[1] "Here we Accept Null Hypothesis as calculated t-value is less than the tabulated t-value i.e. there is no relation exis
t between the mean income of Male and Female or we can say our data is insignificant"
```

Interpretation using p-value:

In [25]:

```
if (p_value_t < 0.05){
    print("Here we Reject Null Hypothesis as calculated t-value is greater than the tabulated t-value i.e. there is some relation exist be
} else{
    print("Here we Accept Null Hypothesis as calculated t-value is less than the tabulated t-value i.e. there is no relation exist between
}
```

```
[1] "Here we Accept Null Hypothesis as calculated t-value is less than the tabulated t-value i.e. there is no relation exis
t between the mean income of Male and Female or we can say our data is insignificant"
```

F-test:

Setting of Hypothesis:

Null Hypothesis, Ho:There is no significant difference between the variance of income of Male and Female i.e. no relation exists between the variance of income of Male and Female.

Alternative Hypothesis, H1:There is a significant difference between the variance of income of Male and Female i.e. some relation exists between the variance of income of Male and Female.

Test Statistic:

Method 1: From Scratch

Calculating the mean of the samples:

In [26]:

```
mean_Male <- mean(sample_male$Income)
print(mean_Male)
mean_Female <- mean(sample_female$Income)
print(mean_Female)
```

```
[1] 90028.3
[1] 80429.2
```

Calculating the variance of the samples:

In [27]:

```
var_Male <- var(sample_male$Income)
print(var_Male)
var_Female <- var(sample_female$Income)
print(var_Female)
```

```
[1] 935002192
[1] 869149274
```

Calculating F-value:

In [28]:

```
F_stat <- var_Male / var_Female
F_stat
```

1.07576709840175

Method 2: Direct Method:

In [29]:

```
var.test(sample_male$Income, sample_female$Income)
```

```
	F test to compare two variances

data:  sample_male$Income and sample_female$Income
F = 1.0758, num df = 29, denom df = 29, p-value = 0.8454
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5120272 2.2601822
sample estimates:
ratio of variances 
          1.075767 
```

Calculating p-value:

In [43]:

```
df1 <- length(sample_male$Income) - 1
df2 <- length(sample_female$Income) - 1

# Calculate the p-value
p_value_F <- pf(F_stat, df1, df2, lower.tail = FALSE)
p_value_F
```

0.422719599756794

Level of Significance:

Here we predetermine level of significance as 5% i.e. 0.05.

Critical Value:

Tabulated value of F at 29,29 degrees of freedom at 5% level of significance is 1.85.

Decision:

Interpretation using F-value:

In [57]:

```
if (F_stat > 1.672){
    print("Here we Reject Null Hypothesis as calculated t-value is greater than the tabulated t-value i.e. there is some relation exist be
} else{
    print("Here we Accept Null Hypothesis as calculated t-value is less than the tabulated t-value i.e. there is no relation exist betweer
}
```

```
[1] "Here we Accept Null Hypothesis as calculated t-value is less than the tabulated t-value i.e. there is no relation exis
t between the variance of income of Male and Female or we can say our data is insignificant"
```

Interpretation using p-value:

In [45]:

```
if (p_value_F < 0.05){
    print("Here we Reject Null Hypothesis as calculated t-value is greater than the tabulated t-value i.e. there is some relation exist be
} else{
    print("Here we Accept Null Hypothesis as calculated t-value is less than the tabulated t-value i.e. there is no relation exist betweer
}
```

```
[1] "Here we Accept Null Hypothesis as calculated t-value is less than the tabulated t-value i.e. there is no relation exis
t between the mean variance of Male and Female or we can say our data is insignificant"
```

Chi-square test:

Setting of Hypothesis:

Null Hypothesis, Ho:There is no significant difference between the Age and Gender of a customer i.e. no relation exists between the Age and Gender of a customer.

Alternative Hypothesis, H1:There is a significant difference between the Age and Gender of a customer i.e. some relation exists between the Age and Gender of a customer.

Test Statistic:

Creating contingency table:

In [46]:

```
Male_young <- nrow(df_male[df_male$Age <= 45, ])
Female_young <- nrow(df_female[df_female$Age <= 45, ])
Male_old <- nrow(df_male[df_male$Age > 45, ])
Female_old <-nrow(df_female[df_female$Age > 45, ])

contingency <- data.frame("Male" = c(Male_young, Male_old), "Female" = c(Female_young, Female_old), row.names = c("Age<=45", "Age>45"))
contingency
```

	Male	Female
Age<=45	43198	33993
Age>45	40602	32207

Method 1: From Scratch

Calculating sum:

In [47]:

```
row_totals <- rowSums(contingency)
col_totals <- colSums(contingency)
total <- sum(contingency)

row_totals
col_totals
total
```

Age<=45
77191
Age>45
72809

Male
83800
Female
66200

150000

Calculating Exected Frequencies:

In [48]:

```
expected <- outer(row_totals, col_totals) / total
expected
```

	Male	Female
Age<=45	43124.04	34066.96
Age>45	40675.96	32133.04

Applying Chi-square test:

In [49]:

```
chi_stat <- sum((contingency - expected)^2 / expected)
chi_stat
```

0.59214695343731

Method 2: Direct Method

In [50]:

```
chi <-  chisq.test(contingency)
chi
```

```
	Pearson's Chi-squared test with Yates' continuity correction

data:  contingency
X-squared = 0.58417, df = 1, p-value = 0.4447
```

Calculating p_value:

In [51]:

```
df_chi <- (nrow(contingency) - 1) * (ncol(contingency) - 1)
p_value_chi <- 1 - pchisq(chi_stat, df_chi)
p_value_chi
```

0.441590103979592

Level of Significance:

Here we predetermine level of significance as 5% i.e. 0.05.

Degrees of Freedom:

df = (r-1)(c-1) where,

r=Total number of rows.

c=Total number of coloumns.

In [52]:

```
df_chi <- (nrow(contingency) - 1) * (ncol(contingency) - 1)
df_chi
```

1

Critical Value:

Tabulated value of chi-square at 1 degrees of freedom at 5% level of significance is 3.84.

Decision:

Interpratation using chi_square value:

In [53]:

```
if (chi_stat > 3.84){
    print("Here we Reject Null Hypothesis as calculated t-value is greater than the tabulated t-value i.e. there is some relation exist be
} else{
    print("Here we Accept Null Hypothesis as calculated t-value is less than the tabulated t-value i.e. there is no relation exist betweer
}
```

```
[1] "Here we Accept Null Hypothesis as calculated t-value is less than the tabulated t-value i.e. there is no relation exis
t between the Age and Gender of a customer or we can say our data is insignificant"
```

Interpretation using p-value:

In [54]:

```
if (p_value_chi < 0.05){
    print("Here we Reject Null Hypothesis as calculated t-value is greater than the tabulated t-value i.e. there is some relation exist be
} else{
    print("Here we Accept Null Hypothesis as calculated t-value is less than the tabulated t-value i.e. there is no relation exist betweer
}
```

```
[1] "Here we Accept Null Hypothesis as calculated t-value is less than the tabulated t-value i.e. there is no relation exis
t between the Age and Gender of a customer or we can say our data is insignificant"
```

Conclusion:

In [58]:

```
if(p_value_t < 0.05){
    print("From t-test it is cleary shown that the there is some relations exist between the mean income of male and female i.e. our data
} else{
        print("From t-test it is cleary shown that the there is no relations exist between the mean income of male and female i.e. our dat
}

if(p_value_F < 0.05){
    print("From F-test it is cleary shown that the there is some relations exist between the variance of income of male and female i.e. ou
} else{
        print("From F-test it is cleary shown that the there is no relations exist between the variance of income of male and female i.e.
}

if(p_value_chi < 0.05){
    print("From chi-square test it is cleary shown that the there is some relations exist between the Age and Gender of a customer i.e. ou
} else{
        print("From chi-square test it is cleary shown that the there is no relations exist between the Age and Gender of a customer i.e.
}
```

```
[1] "From t-test it is cleary shown that the there is no relations exist between the mean income of male and female i.e. ou
r data is insignificant"
[1] "From F-test it is cleary shown that the there is no relations exist between the variance of income of male and female
i.e. our data is insignificant"
[1] "From chi-square test it is cleary shown that the there is no relations exist between the Age and Gender of a customer
i.e. our data is insignificant"
```

Overall Conclusion:

Our data is significant.