R Notebook

Code

#Introduction

Data

Code

Code

Gender <int>	Earning <chr>	Age <int>
2	12000	19
1	32933.333333	32
2	21991.666667	19
1	33333.333333	27
2	43000	58
1	V	64
2	29166.666667	40
1	28000	55
1	V	21
2	29166.666667	44

1-10 of 5,128 rows Previous **1** 2 3 4 5 6 … 100 Next

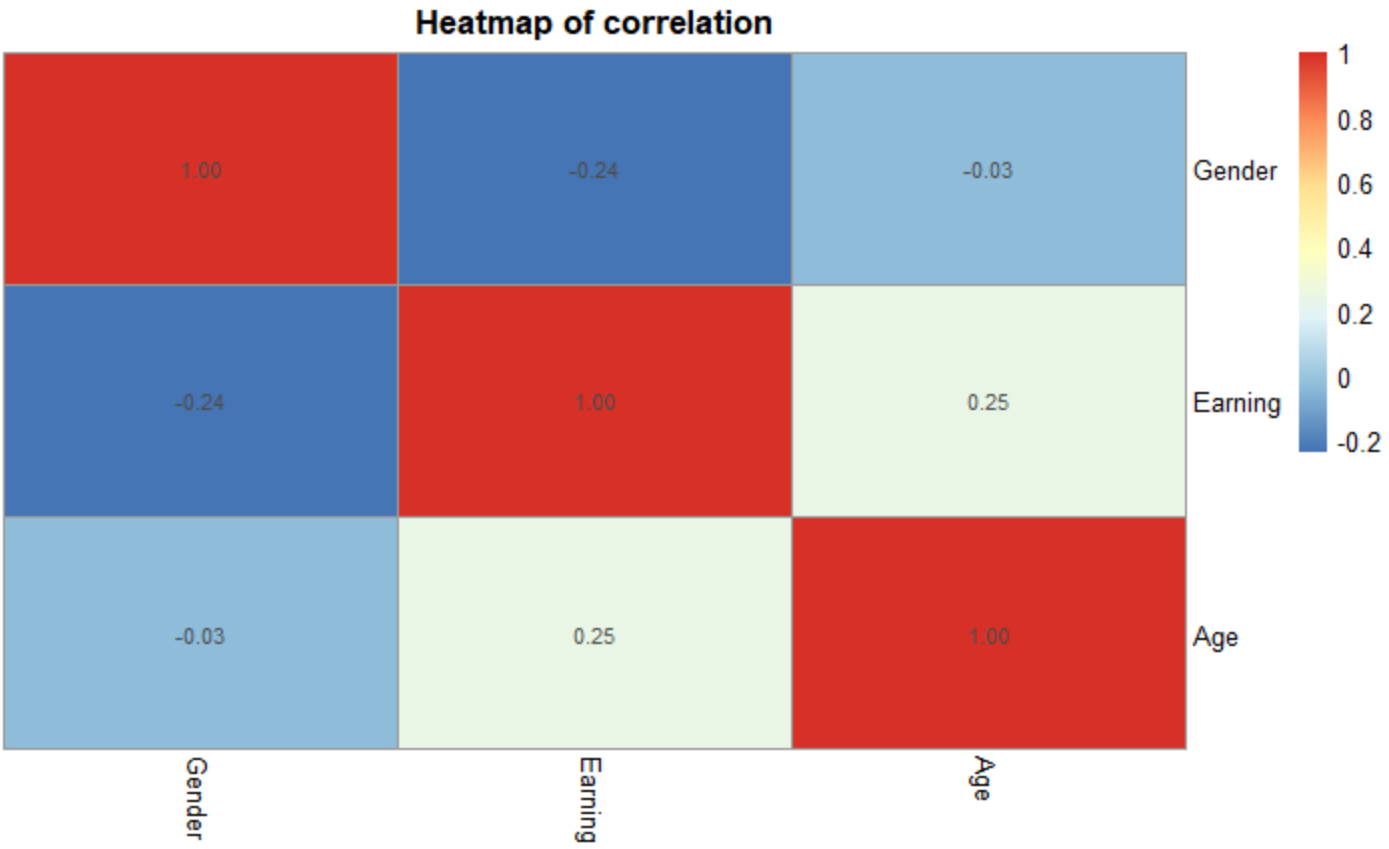
Code

```
'data.frame':   5128 obs. of  3 variables:
 $ Gender : int  2 1 2 1 2 1 2 1 1 2 ...
 $ Earning: num  12000 32933 21992 33333 43000 ...
 $ Age    : int  19 32 19 27 58 64 40 55 21 44 ...
```

Code

Heatmap represent the correlation between the variables. Regression model has assumption that independent variable don't have high correlation between them.

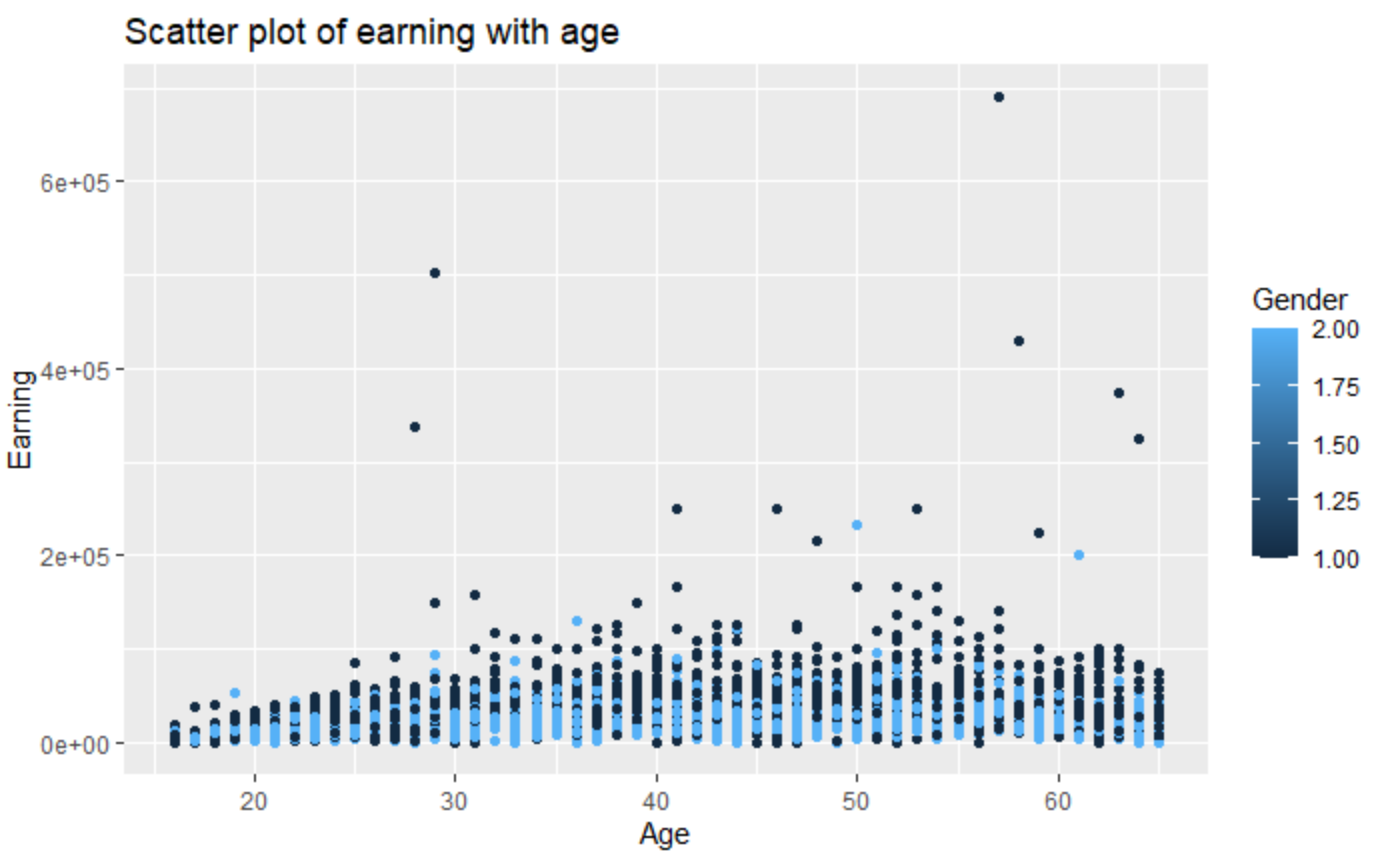
Code



Code

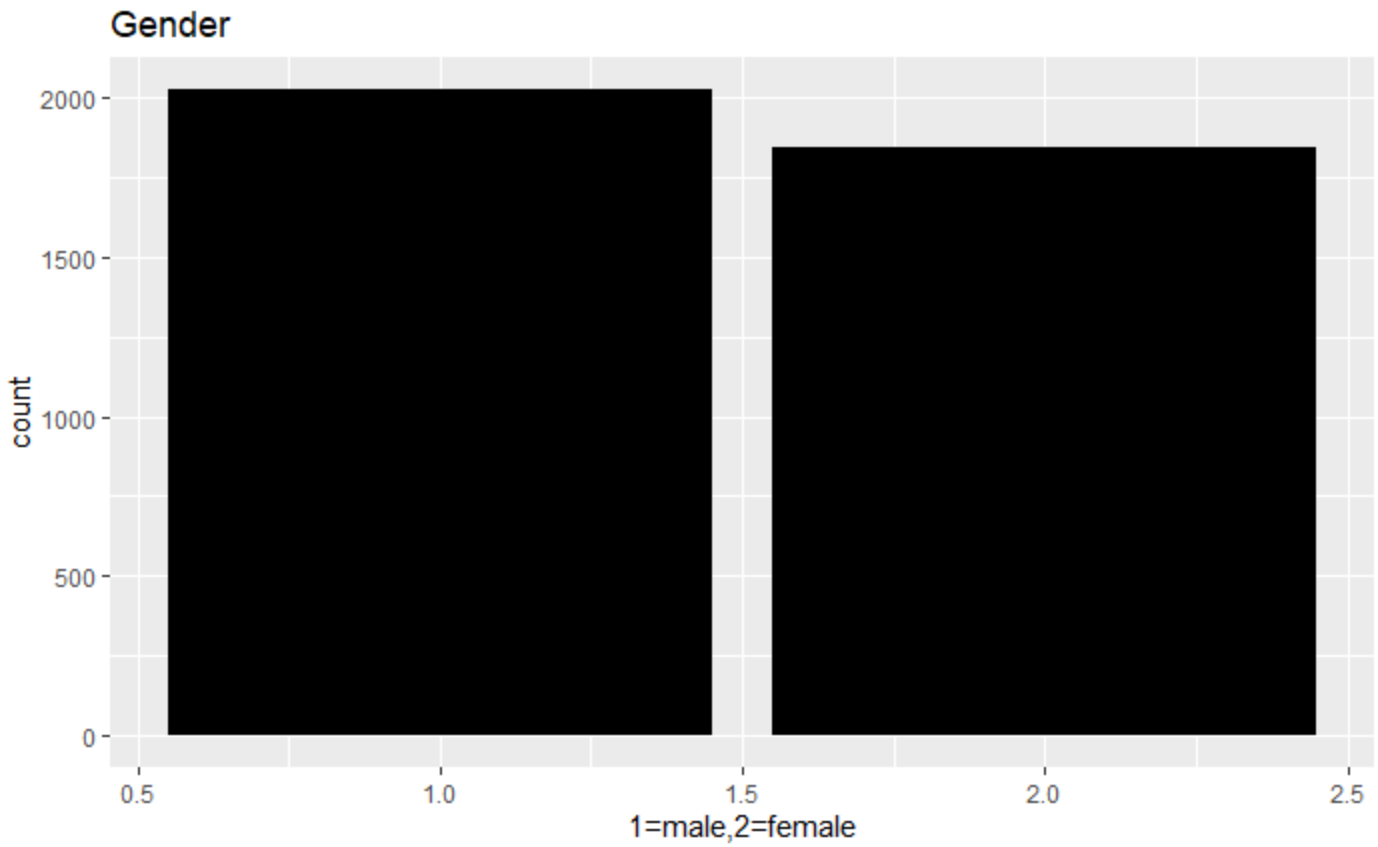
```
     Gender         Earning           Age       
 Min.   :1.000   Min.   :     0   Min.   :16.00  
 1st Qu.:1.000   1st Qu.: 22142   1st Qu.:30.00  
 Median :1.000   Median : 32917   Median :41.00  
 Mean   :1.476   Mean   : 35290   Mean   :40.49  
 3rd Qu.:2.000   3rd Qu.: 41667   3rd Qu.:51.00  
 Max.   :2.000   Max.   :690000   Max.   :65.00  
```

Code



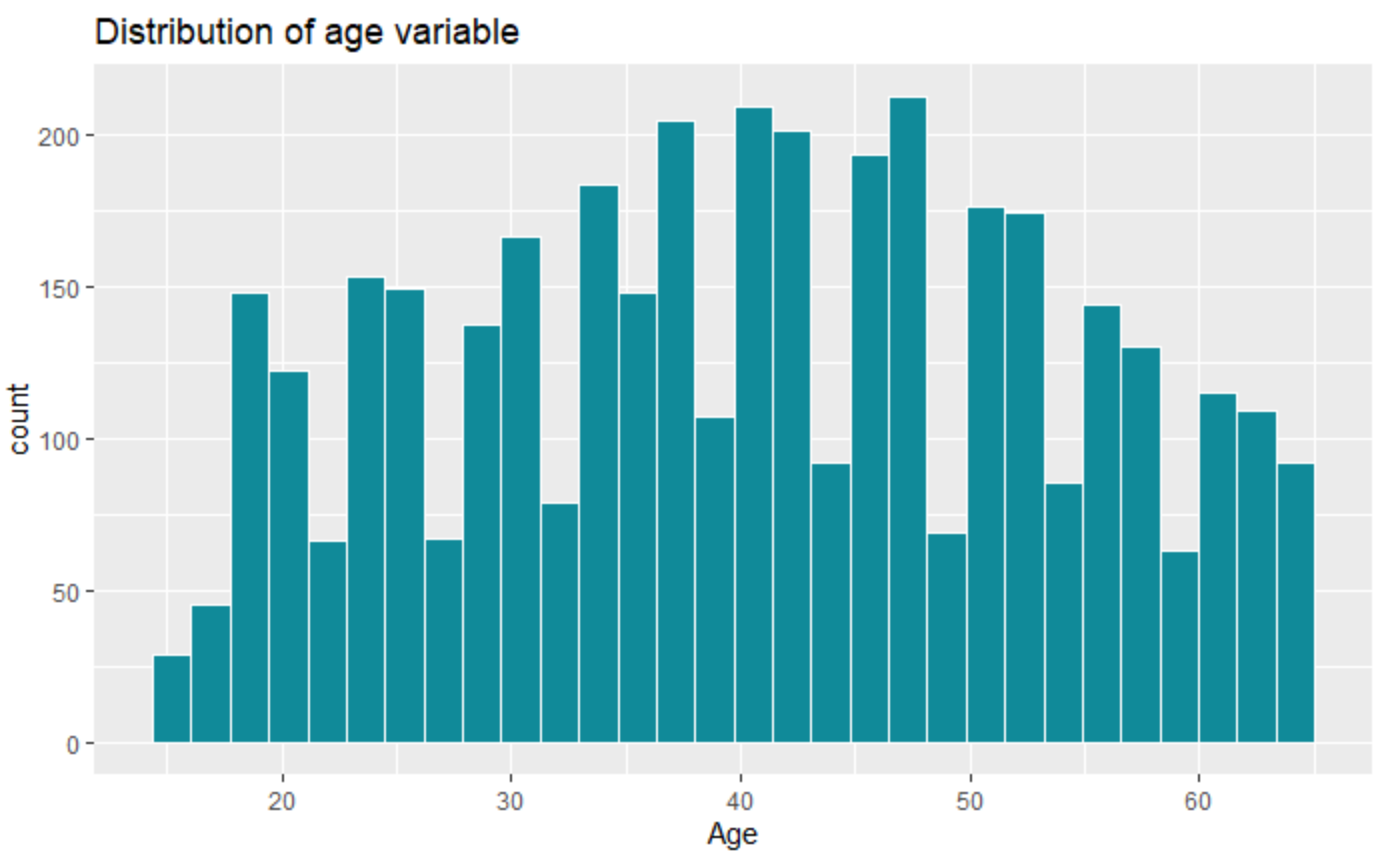
Above plot represent the scatter plot of age variable with earning variable and color coges represent the age of the participants. black dots represent the data for male candidates and blue color represent the data for female participants.

Code



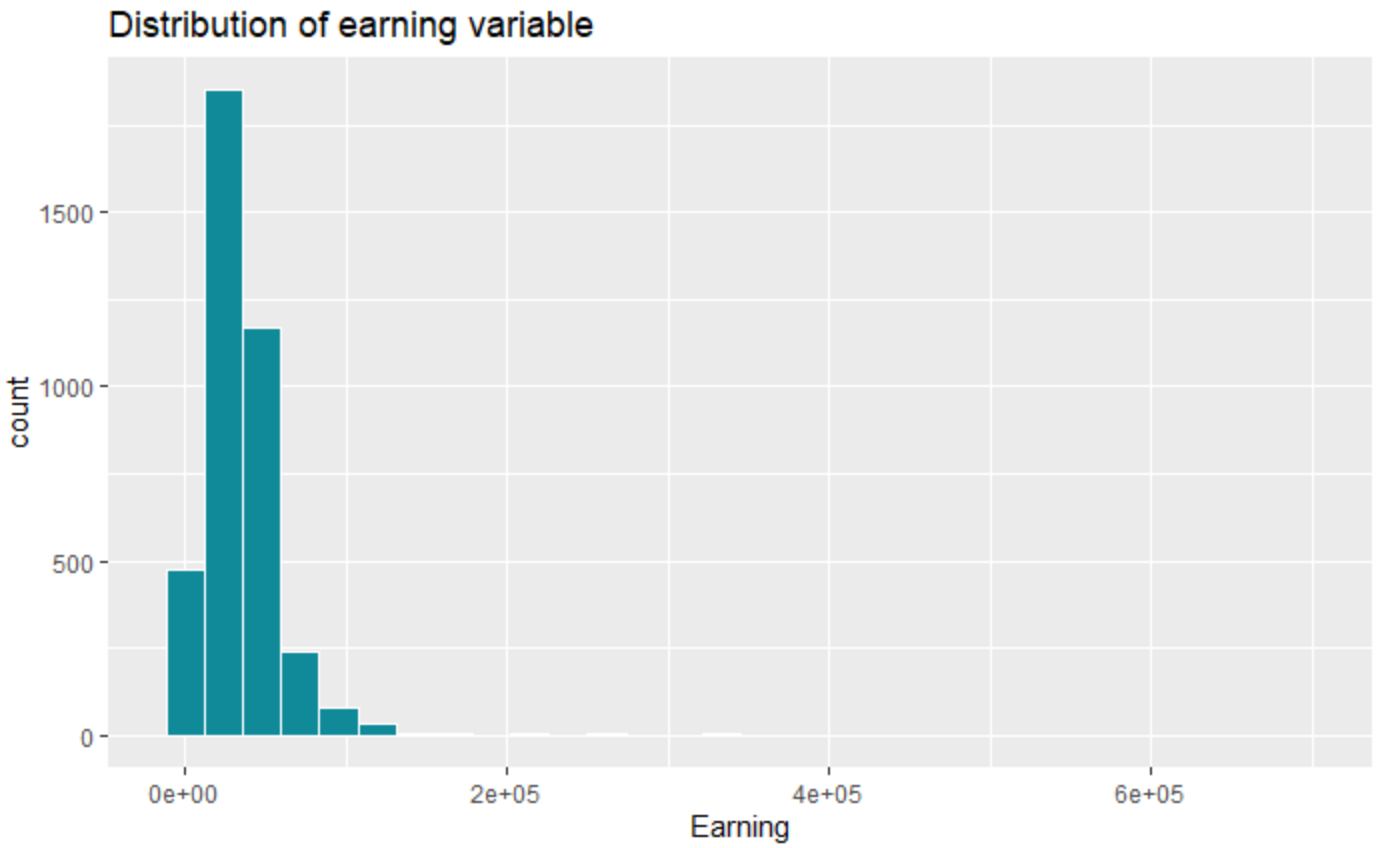
Above bar chart represent the count of the male and females in the data.

Code



This histogram represent the distribution of the age variable.

Code



This histogram represent the distribution of the earning variable. # Results

Code

```
Call:
lm(formula = Earning ~ Gender, data = dff)

Residuals:
   Min     1Q Median     3Q    Max
-41571 -12404  -2404   7864 648429

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  41571.0      603.2   68.92   <2e-16 ***
Gender2     -13185.5      874.0  -15.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27140 on 3865 degrees of freedom
Multiple R-squared:  0.05562,   Adjusted R-squared:  0.05537
F-statistic: 227.6 on 1 and 3865 DF,  p-value: < 2.2e-16
```

Code

```
[1] 15122.21
```

Code

```
[1] 27136.14
```

our model shows that gender is statistically significant variable in the model. Now if we take a look at the coefficient value of the Gender, we can see it has a negative sign with it. Which shows that female has 13185 less income than the males when all other covariates are constant. R square value is showing that gender is able explain only 5% variance in the earning variable. The model has mean absolute error of 15122.21 and root mean square error is 27136.14 which will help us to compare the model performance with next model.

Code

```
Call:
lm(formula = Earning ~ ., data = dff)

Residuals:
   Min     1Q Median     3Q    Max
-53341 -12184  -2662   7250 640227

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  20714.98    1454.11   14.25   <2e-16 ***
Gender2     -12739.01     848.05  -15.02   <2e-16 ***
Age            509.78      32.54   15.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26320 on 3864 degrees of freedom
Multiple R-squared:  0.112, Adjusted R-squared:  0.1116
F-statistic: 243.7 on 2 and 3864 DF,  p-value: < 2.2e-16
```

Code

```
[1] 14412.85
```

Code

```
[1] 26313.31
```

our model represents that both the independent variables are statistically significant at significance level of 0.05. Now if we take a look at the coefficient value of the Gender, we can see it has a negative sign with it. Which shows that female has 12739 less income than the males when all other covariates are constant. Age has a postive coefficent that explains that with the increase in age earning also increase and it makes sense. With a one unit increase in age earning will be 509.78 more according to our mode when all other factors are constant. R square value is showing that independent variables are able explain only 11.16 % variance in the earning variable. The model has mean absolute error of 14412.85 and root mean square error is 26313.31 which clearly shows that this model is performing much better than the pervious one.