EDA Final Project using Wage Dataset

The project aims to perform Exploratory Data analysis on the Wage Dataset and find the predictors that play an important role in deciding the Hourly Wage in the model.

```
#importing all the libraries used in the project
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(summarytools)
```

```
## Registered S3 method overwritten by 'pryr':
##   method      from
##   print.bytes Rcpp
```

```
## For best results, restart R session and update pander using devtools:: or remotes::install_github('rapporter/pander')
```

```
library(vcd)
```

```
## Loading required package: grid
```

```
library(ggpubr)
library(fastDummies)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(moments)
```

Soliution 1: Load a dataset from a CSV file

```
#read data from csv file and store into dataframe
wage_dataframe <- read.csv(file = 'wage.csv')

#displaying first 6 elemenents of the data via head
head(wage_dataframe)
```

	married <dbl>	hourly_wage <dbl>	years_in_education <dbl>	years_in_employment <dbl>	num_dependents <dbl>	gender <chr>	race <chr>
1	1	3.24	12	2	3	female	white
2	0	3.00	11	0	2	male	white
3	1	6.00	8	28	0	male	white
4	1	5.30	12	2	1	male	white
5	1	8.75	16	8	0	male	white
6	0	11.25	18	7	0	male	white

6 rows

```
#converting married variable to factor before starting with analysis
wage_dataframe$married <- factor(wage_dataframe$married)
```

Solution 2: Display descriptive statistics about the dataset

```
#creating vector containing the column names for only numerical variables
descriptive_columns <- c("hourly_wage","years_in_education","years_in_employment","num_dependents")

#show descriptive stats for the mentioned column names only(numerical)
summary(wage_dataframe[descriptive_columns])
```

```
##   hourly_wage     years_in_education years_in_employment num_dependents
##  Min.   : 0.530   Min.   : 0.00      Min.   : 0.000      Min.   :0.000
##  1st Qu.: 3.350   1st Qu.:12.00      1st Qu.: 0.000      1st Qu.:0.000
##  Median : 4.670   Median :12.00      Median : 2.000      Median :1.000
##  Mean   : 5.918   Mean   :12.56      Mean   : 5.152      Mean   :1.044
##  3rd Qu.: 6.880   3rd Qu.:14.00      3rd Qu.: 7.000      3rd Qu.:2.000
##  Max.   :24.980   Max.   :18.00      Max.   :44.000      Max.   :6.000
##  NA's   :8        NA's   :3          NA's   :6           NA's   :5
```

From the result of the descriptive statistics, it is clear that maximun hourly wage is nearly 4 times its average while the average accumulates to 1.8 times first quartile or 25% of the hourly wage data. The range of number of dependents in our data is between 0 to 6 and mean of years in education is around 2.4 times the mean of years in employment.

```
#showing descriptive statistics for categorical variables only i.e married,race and gender

cross<-table(wage_dataframe$race,wage_dataframe$gender,wage_dataframe$married)
round(prop.table(cross,2)*100,digits=0)
```

```
## , ,  = 0
## 
##    
##               female male
##             0      1    0
##   nonwhite  0      7    3
##   white    25     40   28
## 
## , ,  = 1
## 
##    
##               female male
##             0      0    2
##   nonwhite  0      3    7
##   white    75     49   59
```

From the above table it is evident that white people have majority of population in race, there is not much difference in proportion of male and female and married have greater population than unmarried

Solution 4: Check if any records in the data have any missing values; handle the missing data as appropriate (interpolate missing values, delete records with missing values, etc).

```
#sapply function here takes dataframe as input and provides vector or matrix output.
#here it calculates the sum of NA'sfor all variables of the dataset
sapply(wage_dataframe, function(x) sum(is.na(x)))
```

```
##             married         hourly_wage  years_in_education years_in_employment 
##                   3                   8                   3                   6 
##      num_dependents              gender                race 
##                   5                   0                   0 
```

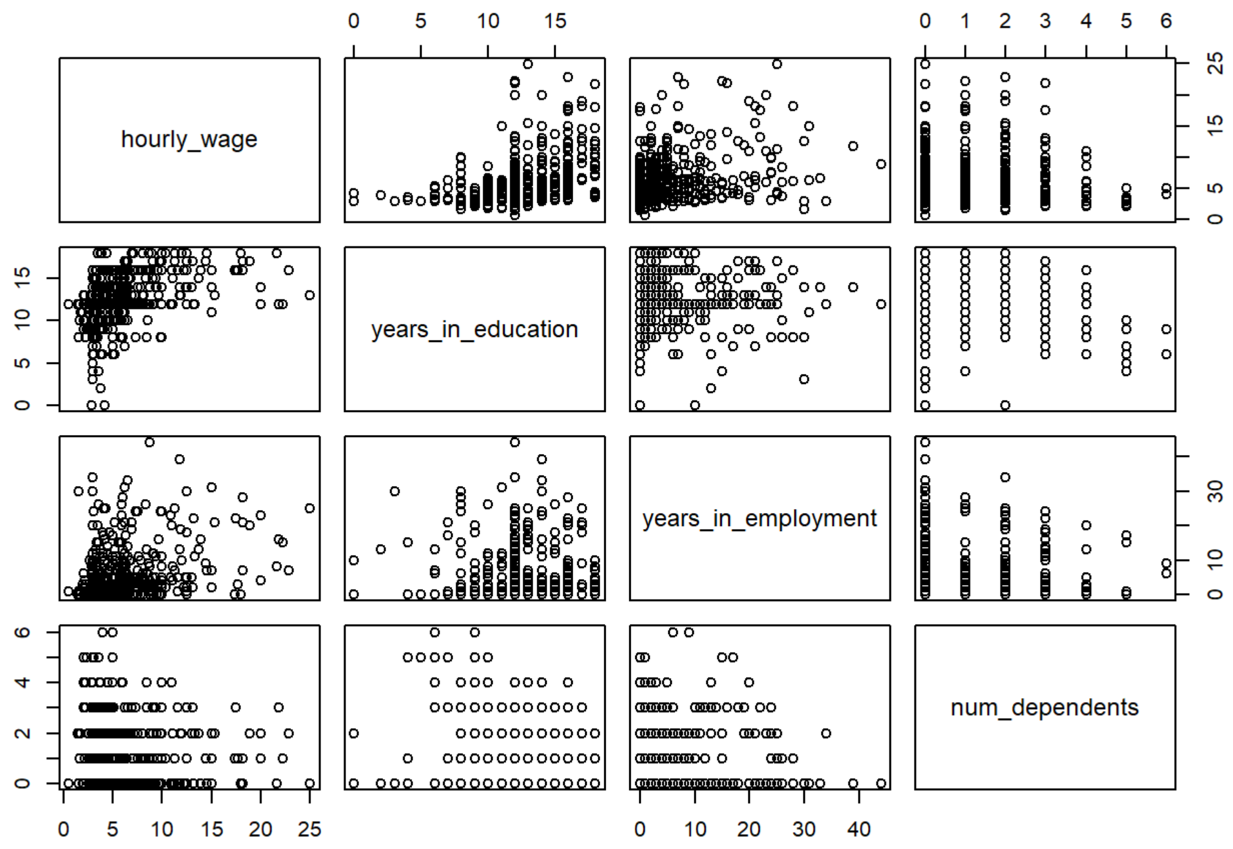
Once NA's are detected, we need to perform data preparation and manuplation for smooth analysis

```
#removing the records in dataset which are having no values for married column
wage_dataframe= wage_dataframe[!is.na(wage_dataframe$married),]

#DisplayPrep function definition
DisplayPrep = function(data,var,round_rum=0)
{
#storing the average of column(acc to var value) for the records having no value for that particular column.
data[is.na(data[,var]),var]= mean(data[,var],na.rm=T)

#rounding the specified column to particular number of digits according to round_rum value
data[,var]= round(data[,var], round_rum)

return(data[,var]) #returning values to be assigned to variable of calling function
}



#here DisplayPrep function passes dataframe name as first argument, column name as second argument and round value as third
 argumment(to round off the particular column to mentioned no. of digits)

return_val=DisplayPrep(wage_dataframe,"hourly_wage",2)
wage_dataframe$hourly_wage<-return_val            #assigning values to hourly_wage column of dataframe

return_val=DisplayPrep(wage_dataframe,"years_in_education",0)
wage_dataframe$years_in_education<-return_val      #assigning values to years_in_education column of dataframe

return_val=DisplayPrep(wage_dataframe,"years_in_employment",0)
wage_dataframe$years_in_employment<-return_val    #assigning values to years_in_employment column of dataframe

return_val=DisplayPrep(wage_dataframe,"num_dependents",0)
wage_dataframe$num_dependents<-return_val          #assigning values to num_dependents column of dataframe


#removing the records in dataset which are having no values for gender column
wage_dataframe= wage_dataframe[!(wage_dataframe$gender==""),]
#removing the records in dataset which are having no values for race column
wage_dataframe= wage_dataframe[!(wage_dataframe$race==""),]

#resultant data after data cleaning
summary(wage_dataframe)
```

```
##  married  hourly_wage     years_in_education years_in_employment num_dependents
##  0:199   Min.   : 0.530   Min.   : 0.00      Min.   : 0.000      Min.   :0.00  
##  1:309   1st Qu.: 3.393   1st Qu.:12.00      1st Qu.: 0.000      1st Qu.:0.00  
##          Median : 4.750   Median :12.00      Median : 2.000      Median :1.00  
##          Mean   : 5.934   Mean   :12.57      Mean   : 5.187      Mean   :1.03  
##          3rd Qu.: 6.880   3rd Qu.:14.00      3rd Qu.: 7.000      3rd Qu.:2.00  
##          Max.   :24.980   Max.   :18.00      Max.   :44.000      Max.   :6.00  
##     gender              race          
##  Length:508         Length:508        
##  Class :character   Class :character  
##  Mode  :character   Mode  :character  
## 
## 
## 
```

Solution 3:Build a graph visualizing (some of) the numerical variables of the dataset

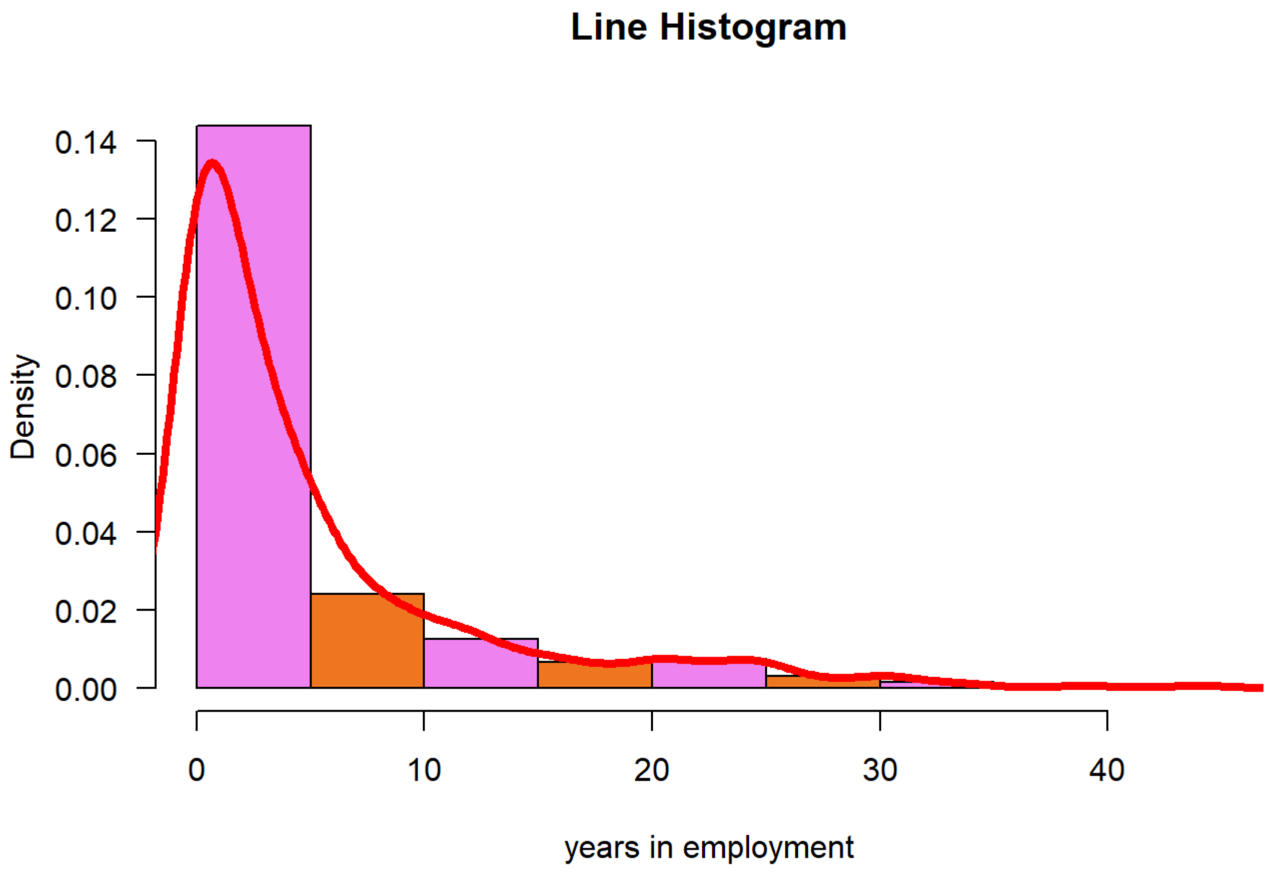
```
 pairs(~hourly_wage+ years_in_education + years_in_employment + num_dependents, data = wage_dataframe)
```



From the above visualization graph it is evident that positive linear relationship exists between hourly wage and years of education. However, that data is randomly distributed for years in employment and hourly wage.Also, the Hourly wage is inversely proportional to number of dependents(houly wage decreases with increase in number of dependents and vice versa)

Solution5: Display the distribution of (some of) numerical variables as histograms. Provide verbal comments on the graph.

```
#plotting histogram
hist(wage_dataframe$years_in_employment,freq = FALSE,col=c("violet", "chocolate2"), xlab ="years in employment", las =1, mai
n="Line Histogram")

#lines plot the density for years_in_employment on histogram shown by red line
lines(density(wage_dataframe$years_in_employment), lwd = 4, col = "red")
```

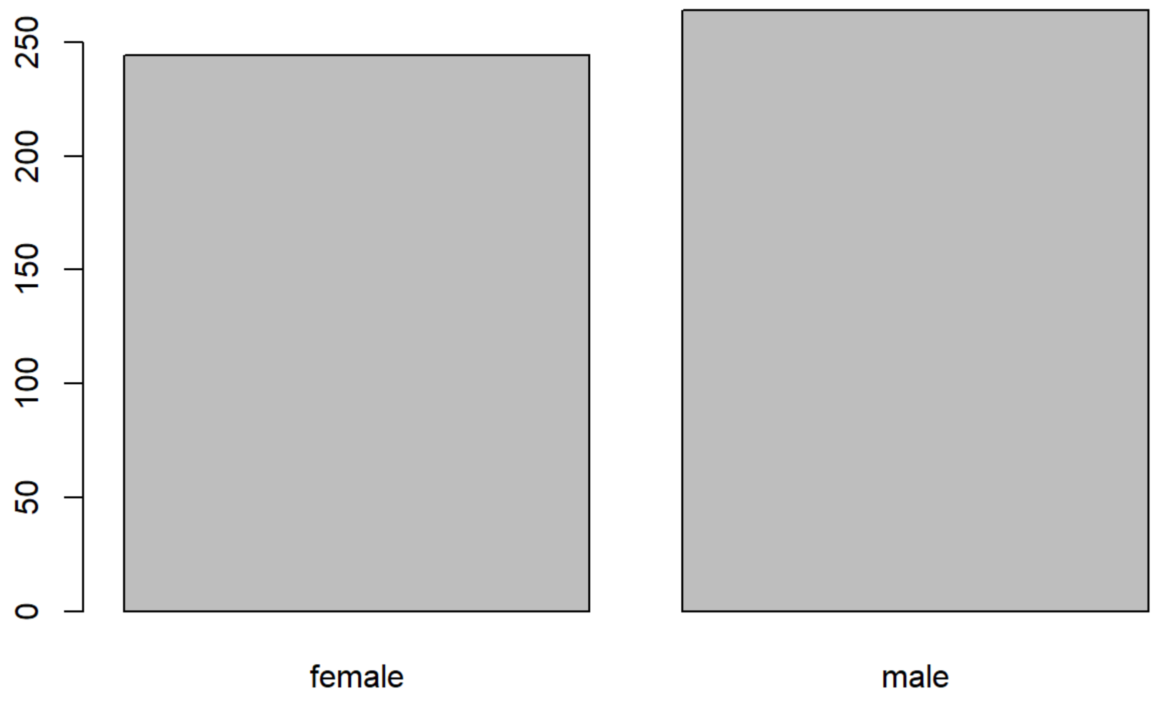


The above histogram displays the height as an years in employment on x-axis and density is plotted on the y-axis. The density representation depicts the rate of change of years_in_employment. Also, the histogram is positively skewed and not following normal distribution for years in employment variable.

Solution 6:Display unique values of a categorical variable

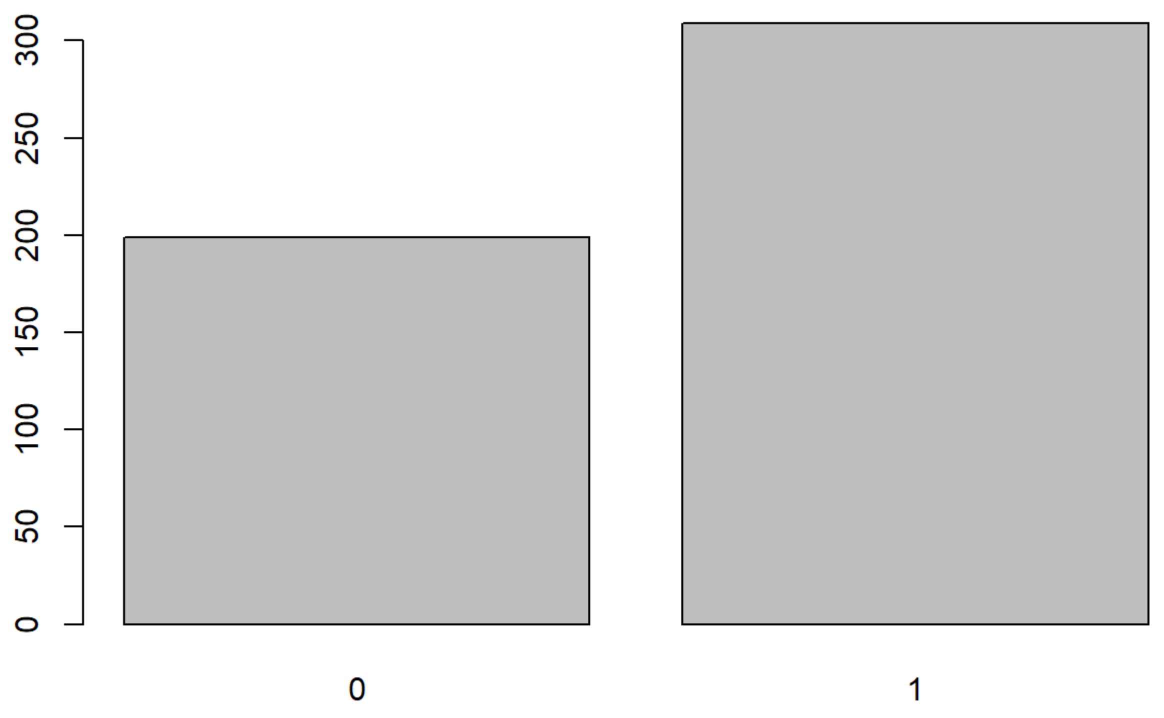
```
DisplayUnique = function(data,var)
{
  t= table(data[,var])

  cat("The unique values for categorical variable", var, "is \n")

  #printing the unique values for particular categorical column
  print(rownames(t))

  #plotting the barplot to depict the occurances of unique values of particular categorical variable
  barplot(t)
}

#display unique values for gender column
DisplayUnique(wage_dataframe,"gender")
```

```
## The unique values for categorical variable gender is
## [1] "female" "male"
```



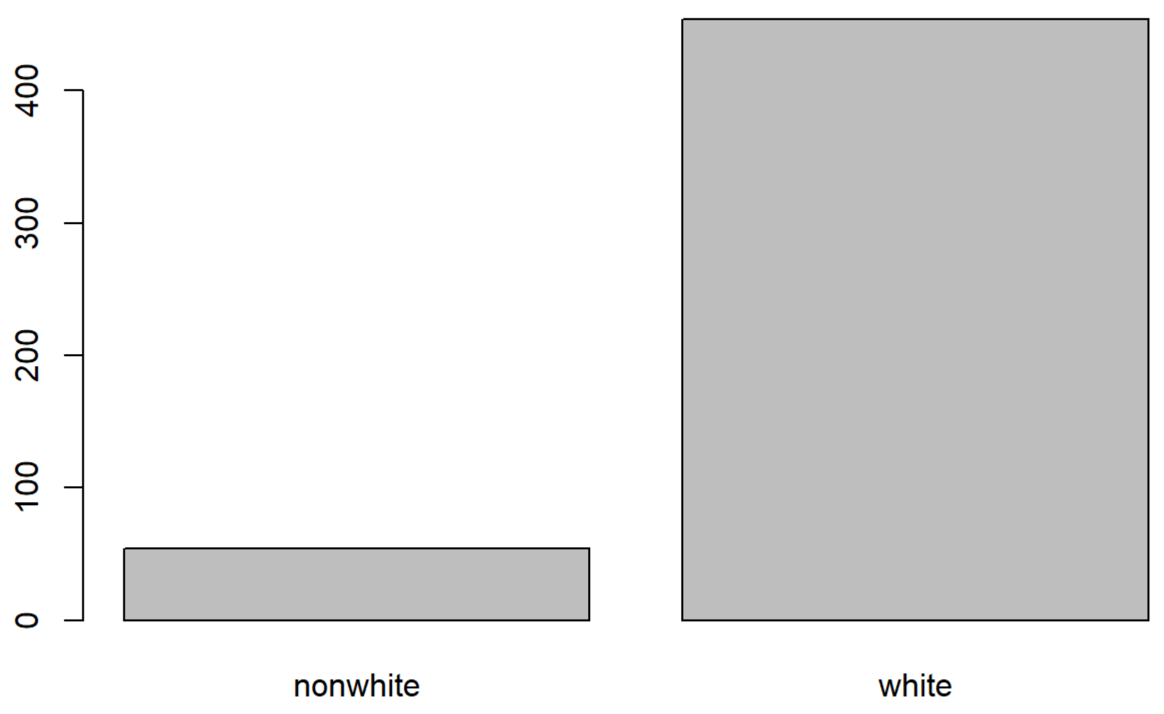
```
#display unique values for married column
DisplayUnique(wage_dataframe,"married")
```

```
## The unique values for categorical variable married is
## [1] "0" "1"
```



```
#display unique values for race column
DisplayUnique(wage_dataframe,"race")
```

```
## The unique values for categorical variable race is
## [1] "nonwhite" "white"
```



solution 7: Build a contingency table of two potentially related categorical variables. Conduct a statistical test of the independence between the variables. Provide verbal comments on the output.

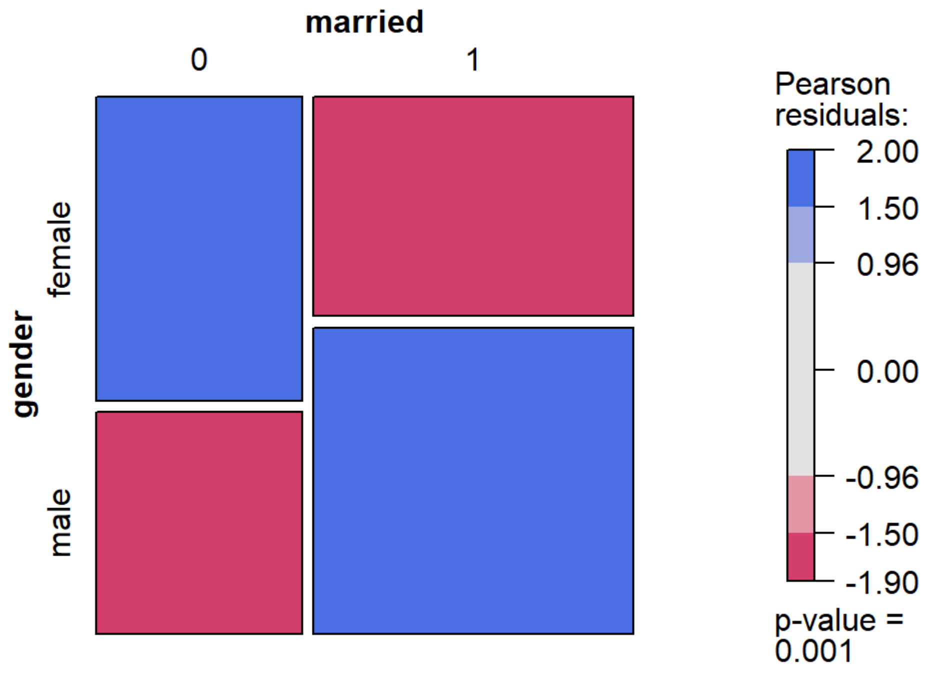
```
#ctable in summarytools package helped to build the contingency table between the 2 categorical variable in my data which is
married and gender
#totals is set to false to avoisd calculation of rows and column values and chisq is set to true to calculate the test to ch
eck the independence between 2 variables
with(wage_dataframe,
     print(ctable(x = married, y = gender, prop = 'n',
                  totals = FALSE, headings = FALSE,chisq = TRUE),
          method = "render"))
```

	gender	
married	**female**	**male**
0	115	84
1	129	180
X^2 *= 11.8442* ***df*** *= 1* ***p*** *= .0006*		

Generated by summarytools (https://github.com/dcomtois/summarytools) 0.9.8 (R (https://www.r-project.org/) version 4.0.3)
2021-01-04

The p-value of chi square test less than 0.05 so we reject the null hypothesis of independence between the two variables. In our context, this indicates that married and gender are dependent on each other and significant relationship exists between the two variables. Also, married males have highest population than all the 4 groups and unmarried males have lowest. Married female have higher proportion than unmarried females.

```
#Mosiac graph depicts the visula representation between married and gender where each tile reflects the cell frequency and c
olor reflects the statistical significance
vcd::mosaic( ~ married + gender, data = wage_dataframe, gp=shading_max, split_vertical=T)
```



From the Mosiac graph it is evident that Males who are married have highest majority than all the combinations and females have higher proprtion of being unmaried than males.

Solution 8. Retrieve a subset of the data based on two or more criteria and present descriptive statistics on the subset. Provide verbal comments on the output.

```
#built subset to group the dataset by race and filter according to two conditions i,e gender to be equal to female and numbe
r of dependents to be greater than 3
wage_dataframe_subset= wage_dataframe %>% group_by(race) %>% filter(gender=="female",num_dependents>3)


summary(wage_dataframe_subset) #display descriptive statistics
```

```
##  married  hourly_wage     years_in_education years_in_employment num_dependents
##  0:7     Min.   : 2.000   Min.   : 5.00      Min.   :0.00        Min.   :4.0   
##  1:3     1st Qu.: 2.458   1st Qu.: 9.25      1st Qu.:0.00        1st Qu.:4.0   
##          Median : 3.000   Median :10.00      Median :0.50        Median :4.0   
##          Mean   : 3.859   Mean   : 9.80      Mean   :1.40        Mean   :4.4   
##          3rd Qu.: 3.740   3rd Qu.:12.00      3rd Qu.:1.75        3rd Qu.:5.0   
##          Max.   :10.000   Max.   :12.00      Max.   :5.00        Max.   :5.0   
##     gender              race          
##  Length:10          Length:10         
##  Class :character   Class :character  
##  Mode  :character   Mode  :character  
##                                       
##                                       
## 
```

THe summary statistics shows that with number of dependents being greater than 3, the mean of the same is 4.4 and mean of hourly wage is 3.859pounds and mean of year_in_education is 7 times the mean of years_in_employment

Solution 9. Conduct a statistical test of the significance of the difference between the means of two subsets of the data. Provide verbal comments

```
t.test(formula = hourly_wage ~ race, #independent sample T test for Hourly wage and race
       data = wage_dataframe)
```

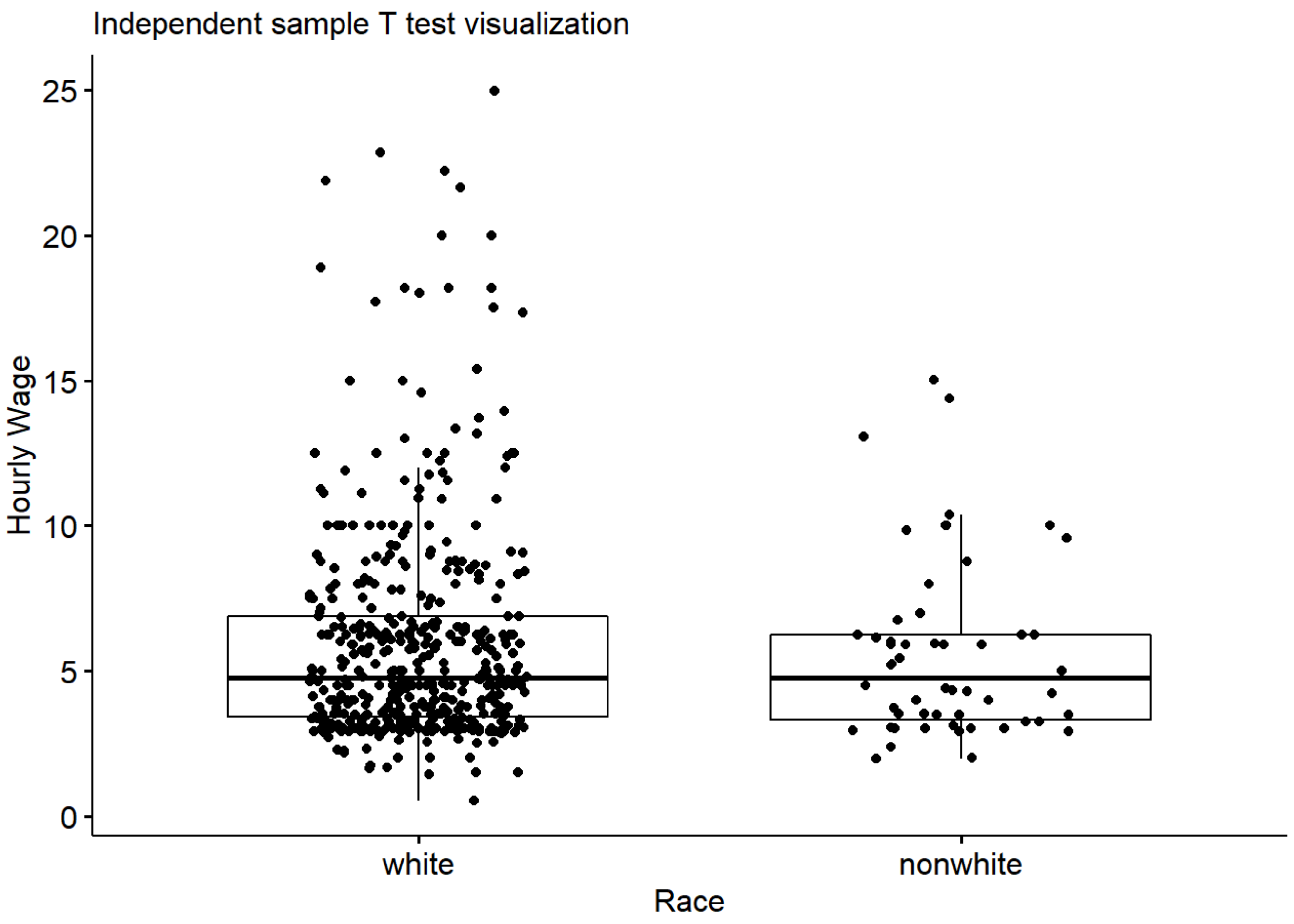
```
## 
##  Welch Two Sample t-test
## 
## data:  hourly_wage by race
## t = -0.70891, df = 73.064, p-value = 0.4806
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.231315  0.585175
## sample estimates:
## mean in group nonwhite    mean in group white 
##               5.645741               5.968811
```

Independent sample T test is conducted to test the significance of difference of mean according to subset of mean hourly wage for race group white vs mean hourly wage for race group non white.

INTERPRETATION The p-value of the test is 0.4806, which is greater than the significance level alpha = 0.05. We can conclude that group white average Hourly wage is not significantly different from group non-white average Hourly wage

t is the t-test statistic value (t = -0.70891), df is the degrees of freedom (df= 73.064), p-value is the significance level of the t-test (p-value = 0.4806). conf.int is the confidence interval of the means difference at 95% (conf.int = [-1.231315, 0.585175]); sample estimates is the mean value of the sample (mean =5.645741,5.968811).

```
#using ggboxplot to visualize significance between houry wage mean for both the groups of race
bxp <- ggboxplot(
  wage_dataframe, x = "race", y = "hourly_wage",
  ylab = "Hourly Wage", xlab = "Race", add="jitter"
  )
bxp +
  labs(subtitle = "Independent sample T test visualization")
```



Solution 10: Create pivot tables, i.e., create a table that groups the data by a certain categorical variable and displays summaries for each categorical variable. Provide verbal comments.

```
#Grouping dataframe by race and gender and created new column mean_hourly_wage to store mean of hourly wage
#and count to store the no of occurances
wage_dataframe %>% group_by(race,gender) %>% summarise(Mean_houry_wage= mean(hourly_wage),count=n())
```

```
## `summarise()` regrouping output by 'race' (override with `.groups` argument)
```

race	gender	Mean_houry_wage	count
<chr>	<chr>	<dbl>	<int>
nonwhite	female	4.485600	25
nonwhite	male	6.645862	29
white	female	4.662374	219
white	male	7.186298	235

4 rows

The above table depicts that white males have highest majority as compared to all 4 groups. Also the mean of white males is highest while non-white females is the lowest

```
#grouping dataframe by married and filter according to race equal to white and created new column mean_hourly_wage to store
 mean of hourly wage
wage_dataframe %>% group_by(married) %>% filter(race=="white") %>% summarize(Mean_houry_wage= mean(hourly_wage))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

married	Mean_houry_wage
<fct>	<dbl>
0	4.940405
1	6.601957

2 rows

This depicts that average of hourly wage for married whites is more than unmarried whites.

Solution 11: Implement a linear regression model and interpret its output.

```
#before implementing linear regression we need to create dummy variables for the categorical columns and storing in new data
frame-wage_df
wage_df = dummy_cols(wage_dataframe, select_columns = c('gender','race','married'))

#since n-1 dummy variables are need to represent n categorical variable

#deleting the columns which are not needed
wage_df$married<-NULL
wage_df$gender<-NULL
wage_df$race<-NULL


wage_df$married_1<-NULL
wage_df$race_white<-NULL
wage_df$gender_male<-NULL

str(wage_df)
```

```
## 'data.frame':    508 obs. of  7 variables:
##  $ hourly_wage        : num  3.24 3 6 5.3 8.75 ...
##  $ years_in_education : num  12 11 8 12 16 18 12 12 17 16 ...
##  $ years_in_employment: num  2 0 28 2 8 7 3 4 21 2 ...
##  $ num_dependents     : num  3 2 0 1 0 0 0 2 0 0 ...
##  $ gender_female      : int  1 0 0 0 0 0 1 1 0 1 ...
##  $ race_nonwhite      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ married_0          : int  0 1 0 0 0 1 1 1 0 1 ...
```

```
#implementing linear model
#here hourly_wage is dependent variable and the other after ~ are independent variable
model = lm(hourly_wage ~married_0 + race_nonwhite + gender_female + num_dependents + years_in_employment+years_in_education,
data=wage_df)



summary(model)  #showing the model summary
```

```
##
## Call:
## lm(formula = hourly_wage ~ married_0 + race_nonwhite + gender_female +
##     num_dependents + years_in_employment + years_in_education,
##     data = wage_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6080 -1.7905 -0.5754  1.0525 14.8461
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.45041    0.74236  -0.607   0.5443
## married_0           -0.70078    0.28815  -2.432   0.0154 *
## race_nonwhite        0.03850    0.43365   0.089   0.9293
## gender_female       -1.71540    0.27309  -6.281 7.29e-10 ***
## num_dependents       0.12625    0.11089   1.139   0.2555
## years_in_employment  0.15208    0.01910   7.963 1.13e-14 ***
## years_in_education   0.52171    0.04967  10.505  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.985 on 501 degrees of freedom
## Multiple R-squared:  0.3542, Adjusted R-squared:  0.3465
## F-statistic: 45.81 on 6 and 501 DF,  p-value: < 2.2e-16
```

From the above summary table, it is evident that race_nonwhite has highest p value. so we need to remove it and perform the linear regression again

```
model = lm(hourly_wage ~married_0 - race_nonwhite + gender_female + num_dependents + years_in_employment+years_in_education,
data=wage_df)

summary(model)
```

```
##
## Call:
## lm(formula = hourly_wage ~ married_0 - race_nonwhite + gender_female +
##     num_dependents + years_in_employment + years_in_education,
##     data = wage_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6130 -1.7896 -0.5745  1.0784 14.8426
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.44432    0.73844  -0.602   0.5476
## married_0           -0.69879    0.28699  -2.435   0.0152 *
## gender_female       -1.71607    0.27272  -6.292 6.81e-10 ***
## num_dependents       0.12704    0.11042   1.150   0.2505
## years_in_employment  0.15212    0.01907   7.975 1.04e-14 ***
## years_in_education   0.52143    0.04952  10.530  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.982 on 502 degrees of freedom
## Multiple R-squared:  0.3542, Adjusted R-squared:  0.3478
## F-statistic: 55.07 on 5 and 502 DF,  p-value: < 2.2e-16
```

From the above summary table, it is evident that num_dependents has highest p value. so we need to remove it and perform the linear regression again

```
  model = lm(hourly_wage ~married_0 - race_nonwhite + gender_female - num_dependents + years_in_employment+years_in_educatio
n, data=wage_df)

summary(model)
```

```
##
## Call:
## lm(formula = hourly_wage ~ married_0 - race_nonwhite + gender_female -
##     num_dependents + years_in_employment + years_in_education,
##     data = wage_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4827 -1.7643 -0.5605  0.9925 14.7266
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.11660    0.68151  -0.171  0.86422
## married_0           -0.76079    0.28197  -2.698  0.00721 **
## gender_female       -1.70827    0.27272  -6.264 8.07e-10 ***
## years_in_employment  0.15063    0.01904   7.913 1.62e-14 ***
## years_in_education   0.50802    0.04814  10.552  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.983 on 503 degrees of freedom
## Multiple R-squared:  0.3525, Adjusted R-squared:  0.3474
## F-statistic: 68.47 on 4 and 503 DF,  p-value: < 2.2e-16
```

The above table obtained is the most parsimonious model as all the independent variables are significantly related to the dependent variable i.e Hourly wage as the p values for all is less than 0.05.

CONCLUSION: 1. Variable married_0 and gender_female are inversely proportional to hourly wage For Instance: if no. of unmarried person is increased by 1 then houly wage will decrease by 0.76079 pounds. If no. of female person is increased by 1 then houly wage will decrease by 1.70827 pounds.

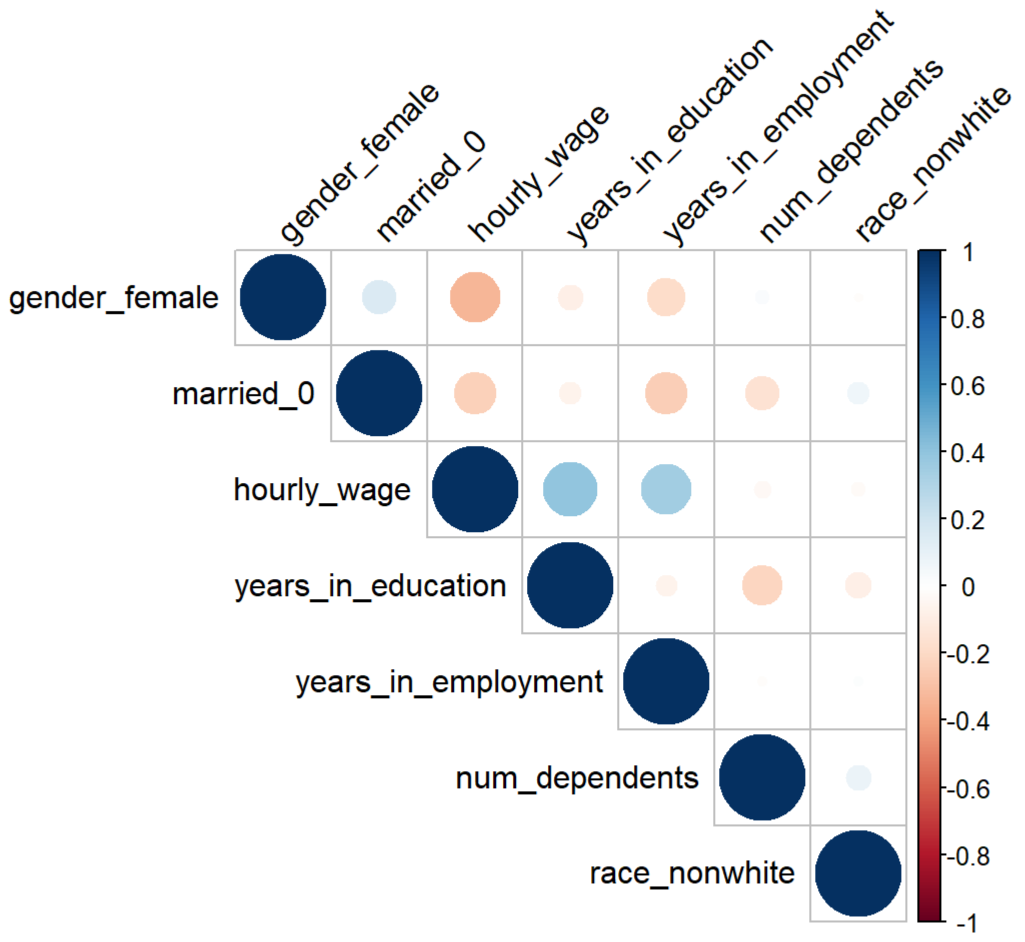
2. Variable years_in_employment and years_in_education are directly proportional to hourly wage
   For Instance: if no. of years in employment is increased by 1 then houly wage will increase by 0.15063 pounds. If no. of years in education is increased by 1 then houly wage will increase by 0.50802 pounds.

```
 #to check the coorelation between the dependent and all the independent variables

res= cor(wage_df, use="complete.obs", method="pearson")

#rounding off  correlation matrix values to 2 decimal places and displaying the results
round(cor(wage_df),2)
```

```
##                     hourly_wage years_in_education years_in_employment
## hourly_wage                1.00               0.39                0.34
## years_in_education         0.39               1.00               -0.06
## years_in_employment        0.34              -0.06                1.00
## num_dependents            -0.04              -0.22               -0.01
## gender_female             -0.34              -0.09               -0.19
## race_nonwhite             -0.03              -0.09                0.01
## married_0                 -0.23              -0.07               -0.24
##                     num_dependents gender_female race_nonwhite married_0
## hourly_wage                  -0.04         -0.34         -0.03     -0.23
## years_in_education           -0.22         -0.09         -0.09     -0.07
## years_in_employment          -0.01         -0.19          0.01     -0.24
## num_dependents                1.00          0.03          0.08     -0.15
## gender_female                 0.03          1.00         -0.01      0.16
## race_nonwhite                 0.08         -0.01          1.00      0.06
## married_0                    -0.15          0.16          0.06      1.00
```

From the coorelation analysis, it can be inferred that no independent variable is highly correlated with dependent variable i.e hourly wage

```
#visualization for correlation
corrplot(res, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



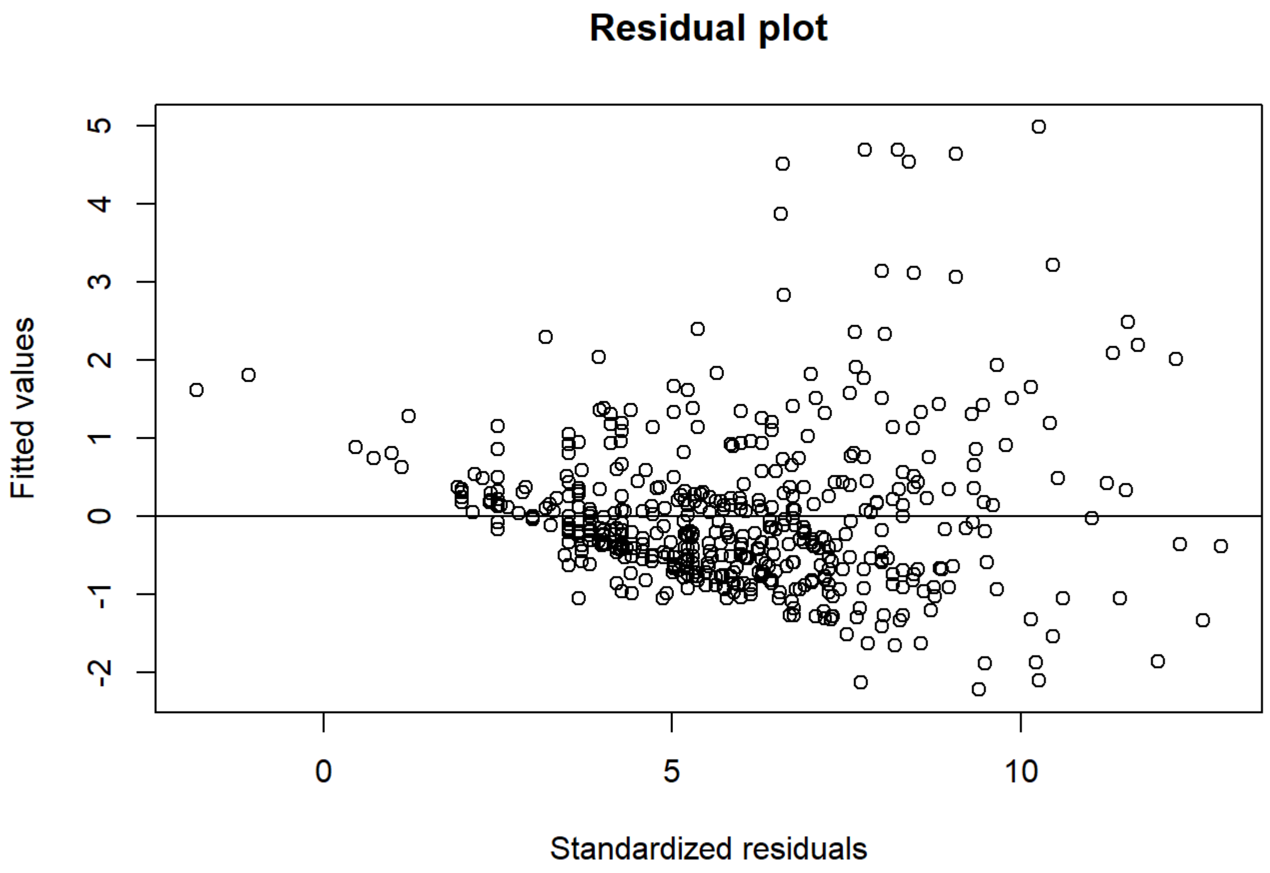
Coefficients on the variables. The coefficients has been estimated as per the most parsimonious model. Our model is thus described by the line:

hourly_wage = -0.45041-0.70078*married_0+0.03850*race_nonwhite-1.71540*gender_female+0.12625*num_dependents+0.15208*years_in_employment + 0.52171* years_in_education

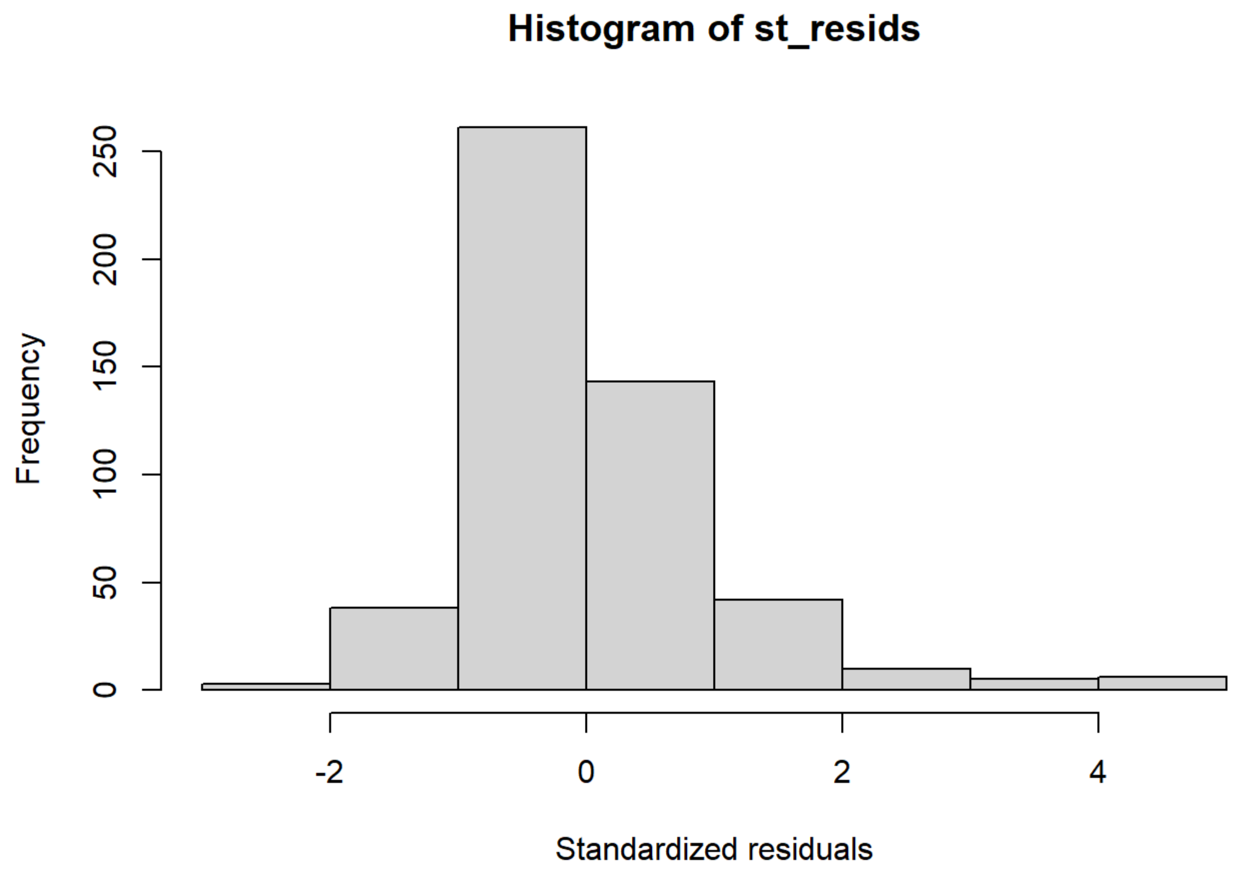
RESIDUAL ANALYSIS

we can plot the standardized residuals and their histogram to confirm if the the assumptions of normality of the distribution of residuals and of the zero mean of residuals are valid with this model.

```
st_resids = rstandard(model)

plot(x=model$fitted.values, y=st_resids, abline(h=0), xlab="Standardized residuals", ylab="Fitted values", main = "Residual
 plot")
```



```
hist(st_resids, xlab="Standardized residuals", breaks=10)
```



The scatterplot and the histogram suggest the residuals are not equally distributed around 0 and are doesnt follow normal distribution.

```
jarque.test(st_resids) #performing Jarque-Bera test
```

```
##
##  Jarque-Bera Normality Test
##
## data:  st_resids
## JB = 857.76, p-value < 2.2e-16
## alternative hypothesis: greater
```

From Jarque-Bera test indicates that the residuals are not normally distributed as the p value is below the 0.05 significance level.

CONCLUSION To conclude, the model is not at all a good choice to estimate the hourly wage according to independent variables of the parsimonious model as the adjusted R square value is 0.3474 which means that model is able to explain only 34.7% of variance for dependent variable i.e Hourly wage which is explained by independent variables-married_0, gender_female, years_in_employment and years_education

Also it can be summarized that by removal of some of the non-significant variables, the model quality has improved slightly from 0.3465(adj R square for initial model) to 0.3474(adj. R square for parsimonius model).

The multicolinearity does not exists in our model