ADULT DATASET
Data Visualisation
Aparajito Sengupta
IISWBM
PGDM – Business Analytics
2018 - 2019

Variable Description:

age: continuous.
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt: continuous.
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education-num: continuous.
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex: Female, Male.
capital-gain: continuous.
capital-loss: continuous.
hours-per-week: continuous.
native-country:

income: <= 50K , > 50K

From the dataset herewith enclosed (adult.txt) , following questions need to be answered:

LOADING OF THE TXT FILE

```
> adult <- read.table("C:/Users/Lenovo/Desktop/IISWBM BA/Assignment - Mid Term/Data_Vis
ualization_Assignment_Regular/adult.data.txt")

> head(adult)

   V1                 V2       V3         V4  V5                    V6                 V7
1 39,         State-gov,  77516, Bachelors, 13,        Never-married,      Adm-clerical,
2 50, Self-emp-not-inc,  83311, Bachelors, 13, Married-civ-spouse,   Exec-managerial,
3 38,           Private, 215646,   HS-grad,  9,             Divorced, Handlers-cleaners,
4 53,           Private, 234721,      11th,  7, Married-civ-spouse, Handlers-cleaners,
5 28,           Private, 338409, Bachelors, 13, Married-civ-spouse,    Prof-specialty,
6 37,           Private, 284582,   Masters, 14, Married-civ-spouse,   Exec-managerial,
             V8     V9    V10   V11 V12 V13           V14   V15
1 Not-in-family, White,   Male, 2174,  0, 40, United-States, <=50K
2       Husband, White,   Male,    0,  0, 13, United-States, <=50K
3 Not-in-family, White,   Male,    0,  0, 40, United-States, <=50K
4       Husband, Black,   Male,    0,  0, 40, United-States, <=50K
```

```
5          Wife, Black, Female,    0,  0, 40,          Cuba, <=50K
6          Wife, White, Female,    0,  0, 40, United-States, <=50K
```

Cleaning of Data

```
##It is needed to remove the ‘,’ and to maintain clear distance among variables for bet
ter readability.Thus using “sep” & “strip.white” functions.

> adult <- read.table("C:/Users/Lenovo/Desktop/IISWBM BA/Assignment - Mid Term/Data_Vis
ualization_Assignment_Regular/adult.data.txt", sep = ",", fill = FALSE, strip.white = T
RUE)

##Replacing variables with their appropriate Column Names,

> colnames(adult) <- c('age,','workclass','fnlwgt','Education','Education-Num','Marital
_Status','Occupation','Relationship','Race','Sex','Capital_Gain','Capital_Loss','Hours_
Per_Wk','Native_Country','Income')
```

Checking the data structure

```
> str(adult)
'data.frame':	32561 obs. of  15 variables:
 $ age,          : int  39 50 38 53 28 37 49 52 31 42 ...
 $ workclass     : Factor w/ 9 levels "?","Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...
 $ fnlwgt        : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159
449 ...
 $ Education     : Factor w/ 16 levels "10th","11th",..: 10 10 12 2 10 13 7 12 13 10 ..
.
 $ Education-Num : int  13 13 9 7 13 14 5 9 14 13 ...
 $ Marital_Status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 5 3 1 3 3 3 4
3 5 3 ...
 $ Occupation    : Factor w/ 15 levels "?","Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...
 $ Relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 1 2 1 6 6 2 1 2 1
...
 $ Race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
 $ Sex           : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 1 2 1 2 ...
 $ Capital_Gain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
 $ Capital_Loss  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Hours_Per_Wk  : int  40 13 40 40 40 40 16 45 50 40 ...
 $ Native_Country: Factor w/ 42 levels "?","Cambodia",..: 40 40 40 40 6 40 24 40 40 40
...
 $ Income        : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 2 2 ...
```

#Now, here is an interesting finding about this Adult dataset. Though, the response (dependent) variables can be considered as binary in nature but the majority of predictors (independent) are multilevel categorical variables.

\> ## Checking of the summary Statistics and the missing values

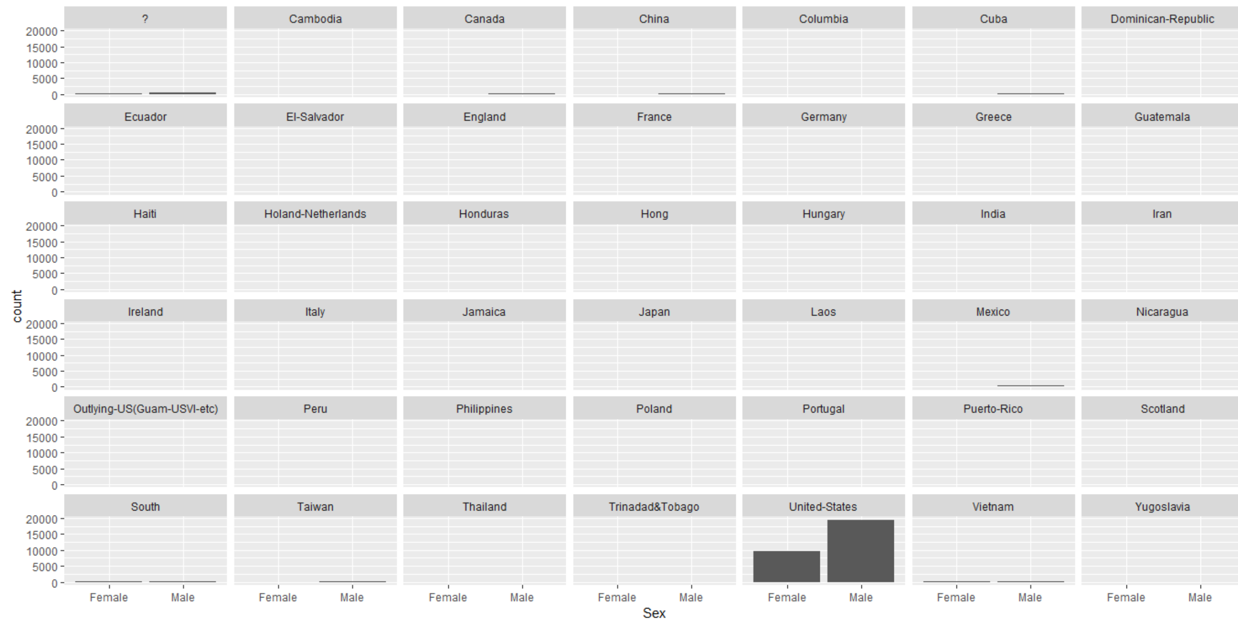
```
> summary(adult)
      age,                  workclass         fnlwgt                 Education
 Min.   :17.00   Private         :22696   Min.   :  12285   HS-grad     :10501
```

```
 1st Qu.:28.00   Self-emp-not-inc: 2541   1st Qu.: 117827   Some-college: 7291  
 Median :37.00   Local-gov       : 2093   Median : 178356   Bachelors   : 5355  
 Mean   :38.58   ?               : 1836   Mean   : 189778   Masters     : 1723  
 3rd Qu.:48.00   State-gov       : 1298   3rd Qu.: 237051   Assoc-voc   : 1382  
 Max.   :90.00   Self-emp-inc    : 1116   Max.   :1484705   11th        : 1175  
                 (Other)         :  981                     (Other)     : 5134  
 Education-Num                  Marital_Status            Occupation   
 Min.   : 1.00   Divorced             : 4443   Prof-specialty :4140  
 1st Qu.: 9.00   Married-AF-spouse    :   23   Craft-repair   :4099  
 Median :10.00   Married-civ-spouse   :14976   Exec-managerial:4066  
 Mean   :10.08   Married-spouse-absent:  418   Adm-clerical   :3770  
 3rd Qu.:12.00   Never-married        :10683   Sales          :3650  
 Max.   :16.00   Separated            : 1025   Other-service  :3295  
                 Widowed              :  993   (Other)        :9541  
         Relationship                   Race            Sex         Capital_Gain 
 Husband       :13193   Amer-Indian-Eskimo:  311   Female:10771   Min.   :    0  
 Not-in-family : 8305   Asian-Pac-Islander: 1039   Male  :21790   1st Qu.:    0  
 Other-relative:  981   Black             : 3124                  Median :    0  
 Own-child     : 5068   Other             :  271                  Mean   : 1078  
 Unmarried     : 3446   White             :27816                  3rd Qu.:    0  
 Wife          : 1568                                             Max.   :99999  

  Capital_Loss     Hours_Per_Wk         Native_Country    Income     
 Min.   :   0.0   Min.   : 1.00   United-States:29170   <=50K:24720  
 1st Qu.:   0.0   1st Qu.:40.00   Mexico       :  643   >50K : 7841  
 Median :   0.0   Median :40.00   ?            :  583                
 Mean   :  87.3   Mean   :40.44   Philippines  :  198                
 3rd Qu.:   0.0   3rd Qu.:45.00   Germany      :  137                
 Max.   :4356.0   Max.   :99.00   Canada       :  121                
                                  (Other)      : 1709                
# There is an unnamed category under ‘Work Class’ Variable that may be “Federal Govt” a
nd ‘Other’  comprises 2 class ‘Unemployed’ & ‘Without Pay’
```

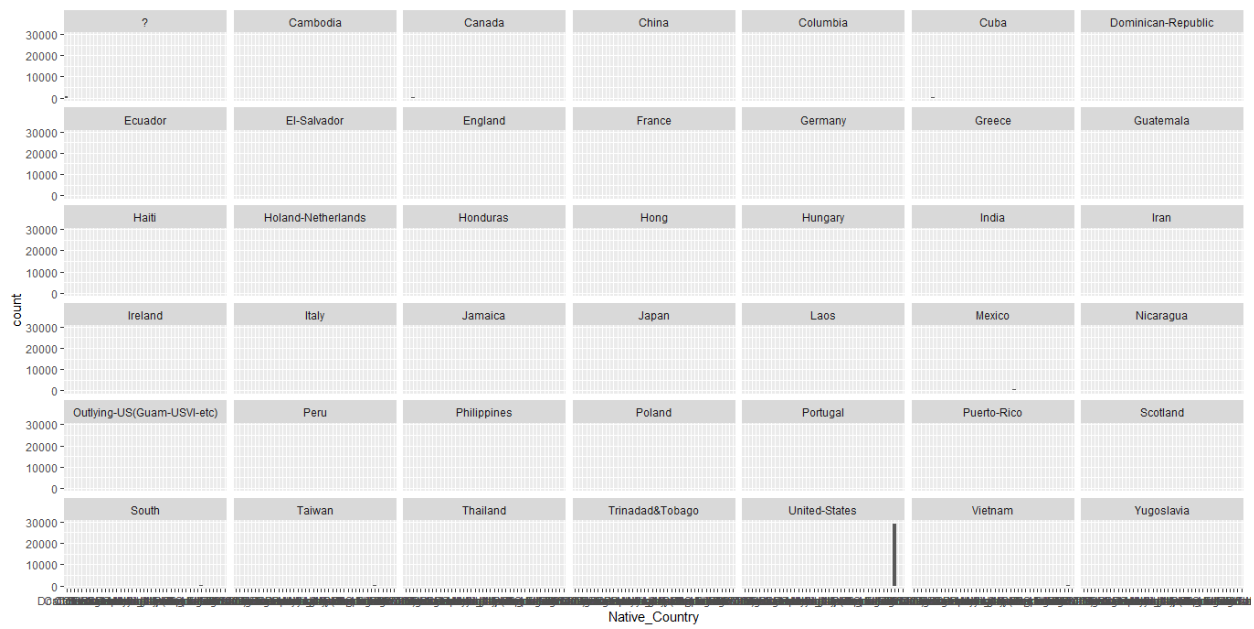
1.The distribution of respondents in terms of countries using a suitable diagram.

I want to see the respondents distribution in two different ways. Since the dynamics of the dataset is governed by respondents from United States I would like to visualize the response with and without United states. Also I want to see the gender distribution of respondents of all countries and then want to check the country wise total response without USA.

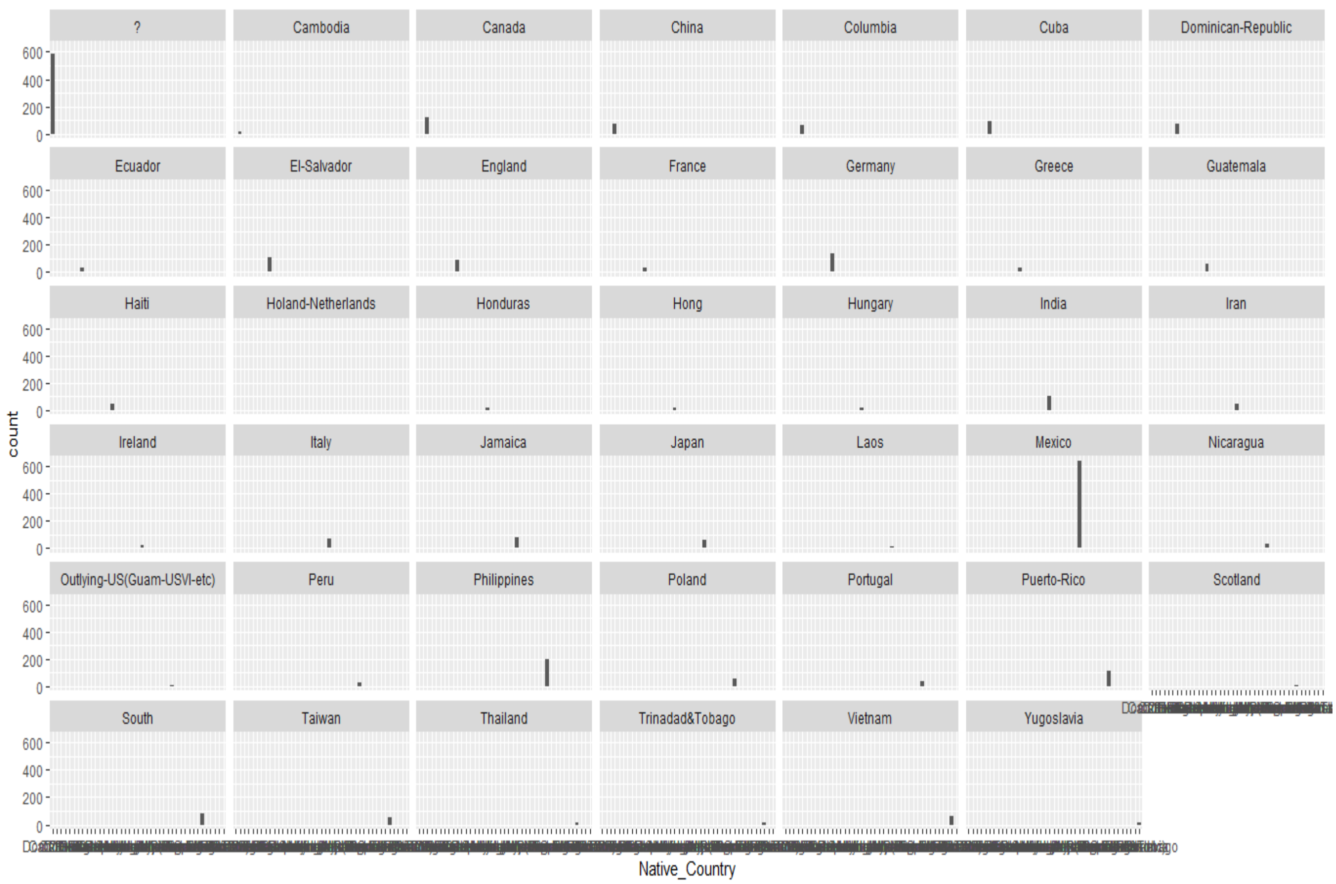
```
ggplot(adult, aes(Native_Country)) +geom_bar()+facet_wrap(~ Native_Country)

#All Countries wrt. Gender distribution
```



#All countries total (This plot reflects United States as an outlier respondents that obscure most of the minion countries.



#And want to see the respondents distribution without the USA



The distribution of respondents with respect to age (converting the age into categorical variable) & country.

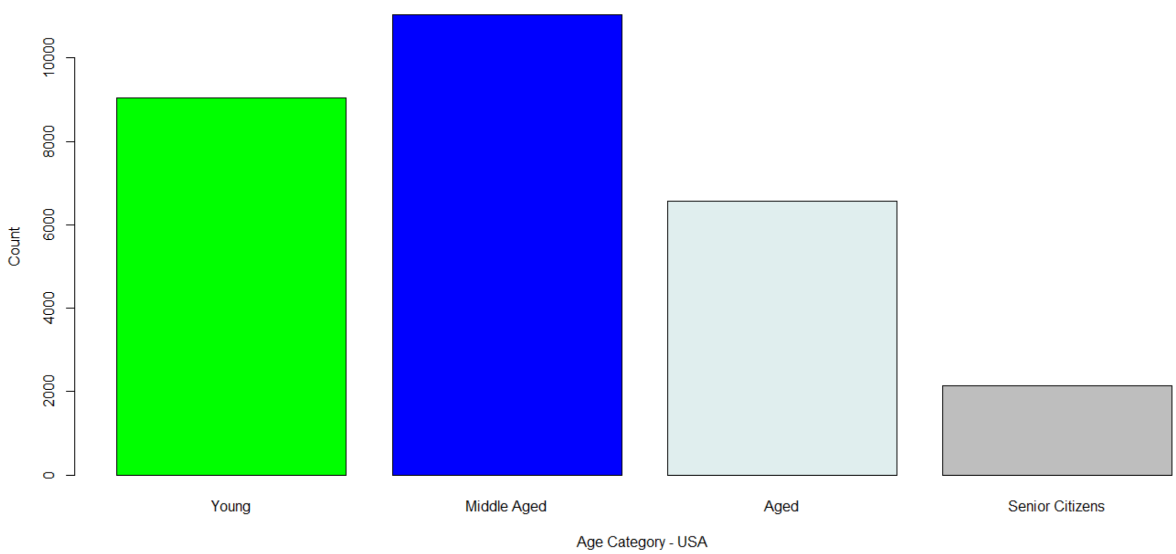
```
library(ggplot2)

windows()


age_cat <- cut(adult$`age`, breaks = c(17,30,45,60,90),label = c("Young","Middle Aged","Aged","Senior Citizens"))

adult <- cbind(adult,age_cat)

windows(10,10)

subset(adult, adult$Native_Country == "United-States" ) -> pu

subset(adult, adult$Native_Country != "United-States" ) -> pi

barplot(table(pu$age_cat),col = c("green","blue","azure2","grey"),xlab = "Age Category - USA ",ylab = "Count" )
```

#Again as above I want to see the age category wise distribution of other countries separately excluding USA for a better visibility and USA alone.

#Here is the graph that shows the respondents age category wise distribution. We can see the respondents frequency is highest from middle aged group followed by the young ,then aged people and senior citizens.

It is not clear whether the survey was conducted taking uniform sample size even in the USA or not . But if the sample size was same, probably this graph is also a reflection of respondents willingness to participate in a survey like this as well. Clearly with increasing age the willingness is diminishing. Or else this is a biased sampling.

#USA Graph:

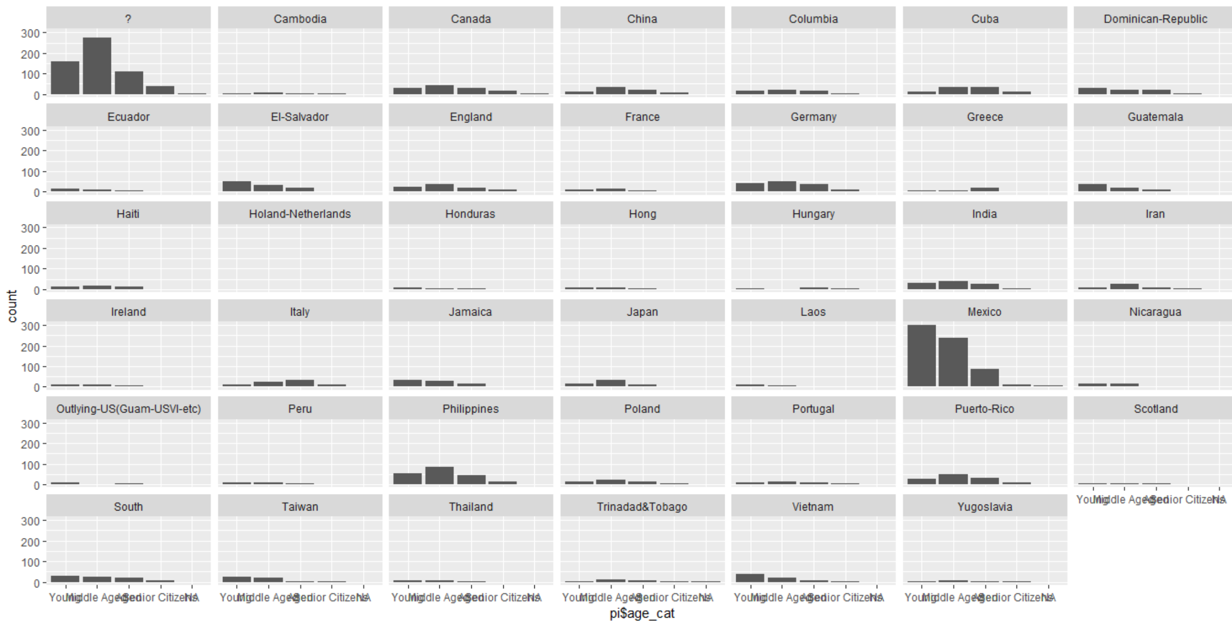


From the other countries :

windows(10,10)

ggplot(pi, aes(pi$age_cat)) +geom_bar()+facet_wrap(~ Native_Country)

We can see from the plot that most of the countries respondent age group wise following the same trend as usa.,i.e, Middle age group dominates followed by young and the aged group and lastly the sr.citizens but this trend is not seen in Mexico,Jamaika,South,Guatmala & Taiwan where the young respondents dominates.



The distribution of respondents among gender and occupation

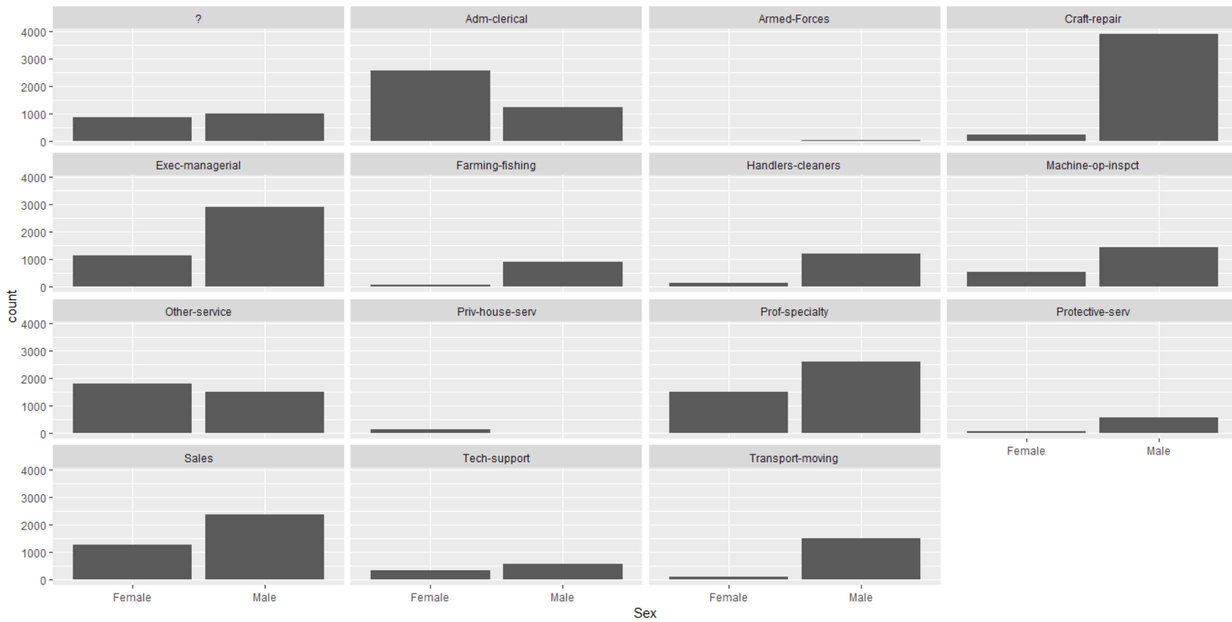
ggplot(adult, aes(Sex)) +geom_bar()+facet_wrap(~ Occupation)

#From the below plot it is visible that women predominates only in Admin-Clerical ,Private house service and Other-Service sectors else it is mainly male dominated distribution.

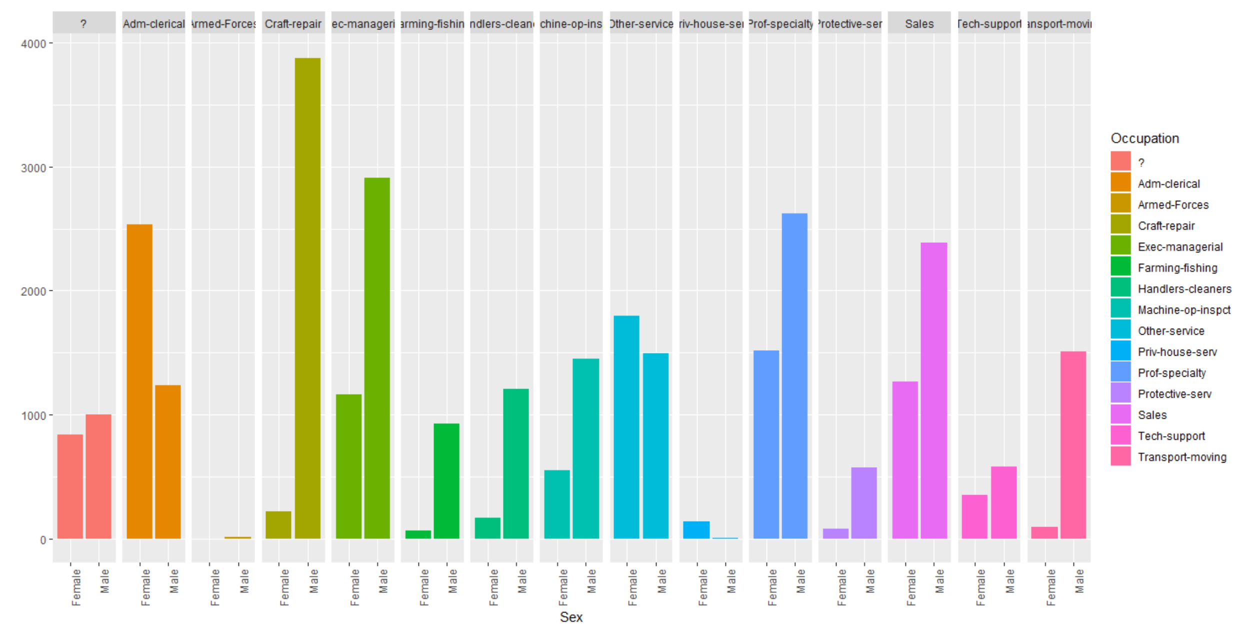
#We talk about gender equality but across nations (including USA) the reality is different. It is surprising to find that there is no presence of women in the armed force section.

#There is significant male dominance in Craft Repairing sector, High end Executive Managerial Section , Sales, Transport moving sectors. The job that demands physical labor are still chosen by men.

#If this data is a true representation of the entire demography then there exists a strong gender inequality across occupations.



#Here attaching another pattern of looking at the same plot



The distribution of respondents for marital status, race & education using suitable Diagram

```
# round(prop.table(table(adult$Education,adult$Race,adult$Marital_Status),2),2)
#Before plotting the tabled data I wanted to check is there any significant distributio
n found in a tabular form. It’s a lengthy table but I am capturing only the excerpt of
my key findings,
```

, , = Divorced

	Amer-Indian-Eskimo	Asian-Pac-Islander	Black
10th	0.02	0.00	0.01
11th	0.01	0.00	0.01
Assoc-acdm	0.01	0.01	0.01
Assoc-voc	0.02	0.00	0.01
Bachelors	0.01	0.01	0.02
HS-grad	0.06	0.03	0.05
Masters	0.01	0.00	0.01
Some-college	0.05	0.02	0.04

	Other	White
Bachelors	0.01	0.02
Doctorate	0.00	0.00
HS-grad	0.04	0.05
Masters	0.00	0.01
Some-college	0.01	0.03

, , = Married-AF-spouse #No significant contribution because of the biased sampling.

, , = Married-civ-spouse

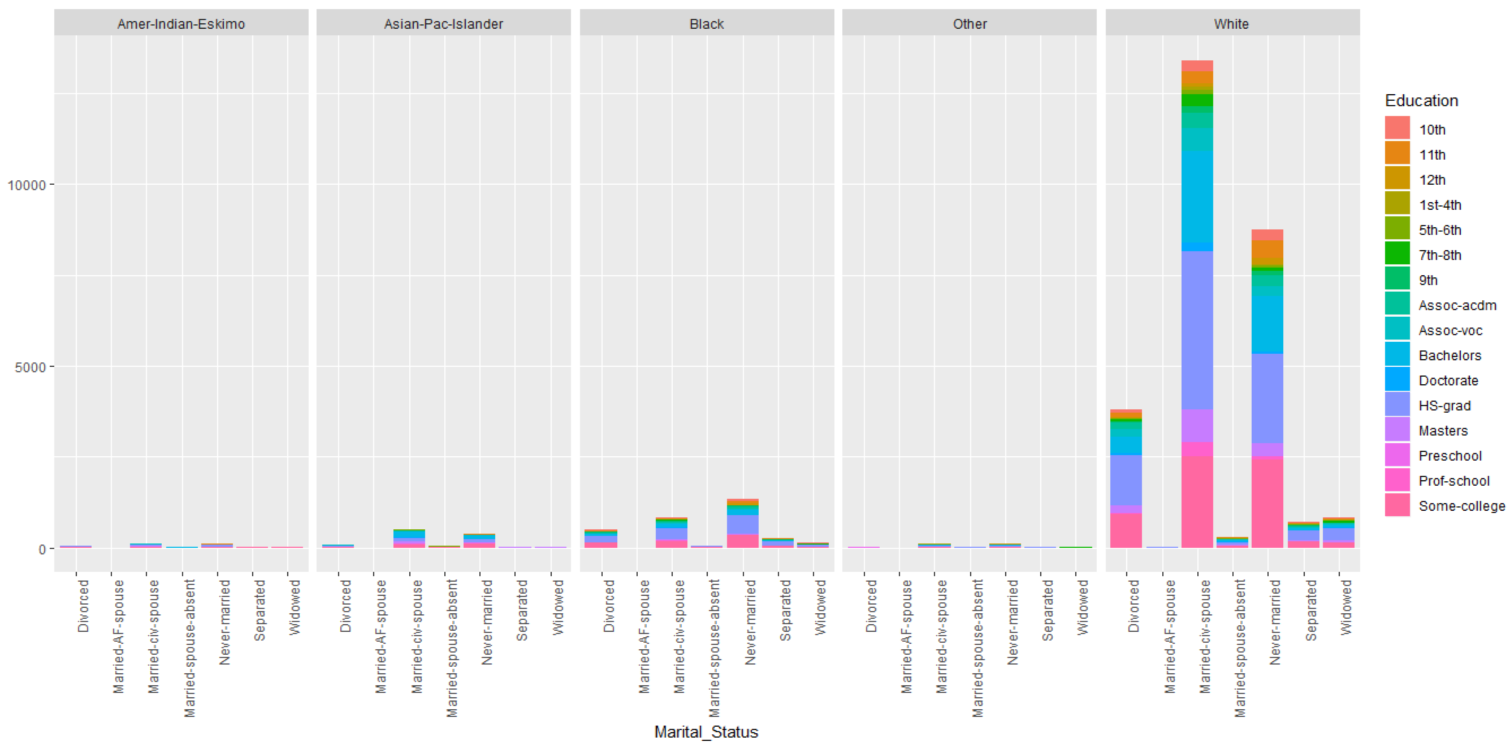
	Amer-Indian-Eskimo	Asian-Pac-Islander	Black
10th	0.02	0.00	0.01
11th	0.01	0.01	0.01
5th-6th	0.01	0.01	0.00
7th-8th	0.01	0.01	0.01
9th	0.00	0.00	0.01
Assoc-acdm	0.01	0.01	0.01
Assoc-voc	0.02	0.01	0.01
Bachelors	0.03	0.14	0.03
Doctorate	0.00	0.03	0.00
HS-grad	0.15	0.10	0.10
Masters	0.01	0.06	0.01
Prof-school	0.01	0.03	0.00
Some-college	0.10	0.07	0.06

	Other	White
10th	0.01	0.01
11th	0.01	0.01
12th	0.02	0.00
1st-4th	0.01	0.00
5th-6th	0.03	0.01
7th-8th	0.03	0.01
9th	0.01	0.01
Assoc-acdm	0.01	0.01
Assoc-voc	0.01	0.02
Bachelors	0.05	0.09
Doctorate	0.00	0.01
HS-grad	0.11	0.16
Masters	0.01	0.03
Preschool	0.00	0.00
Prof-school	0.01	0.01
Some-college	0.06	0.09

```
, ,  = Married-spouse-absent #Insignificant distribution (below 1%)
, ,  = Never-married


               Amer-Indian-Eskimo Asian-Pac-Islander Black
  10th                       0.01               0.00  0.02
  11th                       0.03               0.01  0.03
  12th                       0.01               0.00  0.01
  1st-4th                    0.01               0.00  0.00
  9th                        0.01               0.00  0.01
  Assoc-acdm                 0.00               0.01  0.01
  Assoc-voc                  0.02               0.02  0.01
  Bachelors                  0.03               0.11  0.05
  HS-grad                    0.14               0.07  0.16
  Masters                    0.00               0.02  0.01
  Prof-school                0.00               0.01  0.00
  Some-college               0.07               0.11  0.11

               Other White
  10th          0.02  0.01
  11th          0.01  0.02
  12th          0.02  0.01
  1st-4th       0.01  0.00
  5th-6th       0.00  0.00
  7th-8th       0.02  0.00
  9th           0.01  0.00
  Assoc-acdm    0.01  0.01
  Assoc-voc     0.01  0.01
  Bachelors     0.04  0.05
  Doctorate     0.00  0.00
  HS-grad       0.10  0.09
  Masters       0.01  0.01
  Preschool     0.00  0.00
  Prof-school   0.00  0.00
  Some-college  0.10  0.09

, ,  = Separated


               Amer-Indian-Eskimo Asian-Pac-Islander Black

  Bachelors                  0.00               0.01  0.00
  HS-grad                    0.01               0.01  0.04
  Some-college               0.02               0.00  0.02

               Other White
  5th-6th       0.01  0.00
  9th           0.01  0.00
  HS-grad       0.01  0.01
  Some-college  0.01  0.01
```

Do higher skillsets like sales, technical-support, transport prof, armed forces guarantee a high income?

I explored it by plotting occupation against income levels. As shown below and it is evident that acquiring a high skill does not guarantee an increment in income. The workers with a low skill set like craft-repair, maintenance services, cleaner, private house security earn more as compared to those with the higher skill sets.



Here are some extremely important list of the above observations from the graph and the table :

#This distribution is clearly divided into majorly two halves. 12th & below 12th standard , Some College attendees (and drop outs) and Graduates. Across race this is prominent.

#Married dropouts from the college is highest among whites followed by never married and divorced. .#Those who have completed 10th standard most likely to complete 12th standard as well.

Masters pursuant are more from the married category.

#Asian Pac Islander and Blacks both have highest percentage of HS-Graduates from both married and never married categories.

#In white divorce category, back divorce category and in Asian Pac Islander categories the HS-Grad frequencies are pre dominant.

Exploring the relationship among age, income category & education.

relation <- cbind(adult$`age,`,adult$Income,adult$Education) class(relation)

colnames(relation) <- c("age","Income","Education")

cor(relation)

Output :

```
                 age      Income    Education
age        1.00000000 0.23403710 -0.01050828
Income     0.23403710 1.00000000  0.07931661
Education -0.01050828 0.07931661  1.00000000
```

#Clearly visible age has a positive correlation with Income and education has a –ve correlation with age.

If we consider Income to be the dependent variable on age and education as independent variables and build a regression model

lm(relation$Income ~ relation$age + relation$Education)

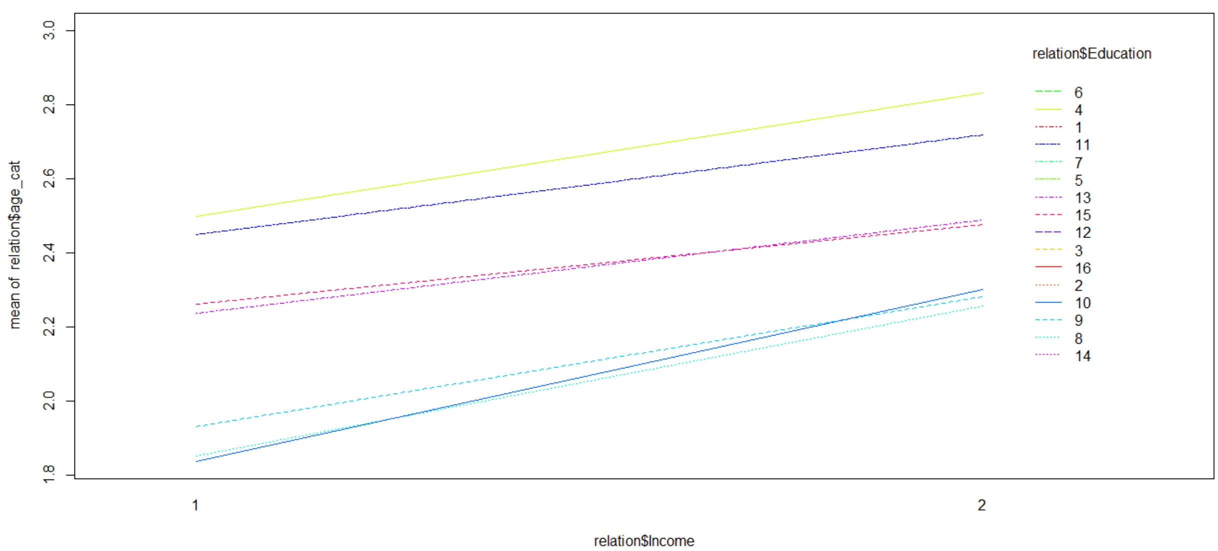
```
Call:
lm(formula = relation$Income ~ relation$age + relation$Education)

Coefficients:
       (Intercept)        relation$age  relation$Education
          0.854640            0.007363            0.009035
```

Which gives us the following equation,

Income = .855 + .0073 Age + .0090 Education

#The factor load is high on Education than that of age.

Let us visualize the same with plotting quick interaction plot with these three variables,



It is quite visible that Income increases with age (Though unusual outliers are there for couple of cases where Sr.Citizen's are earning way above average and way above than their own group) and if it is supported by the higher education then the earning increases more. We can see that from the above calculated correlation coefficient among Age and Income (above) also but one surprising finding is that higher education beyond graduation does not ensure a proportionate high income though.

The Income graph suddenly becomes nonchalant with respect to higher education.

Possibly many other socio economic factors like , gender, race, country, area of expertise, Political,Economic factors, position , social security etc would have interplaying to determine the Income of an individual.Those taken into consideration would provide us a better picture.

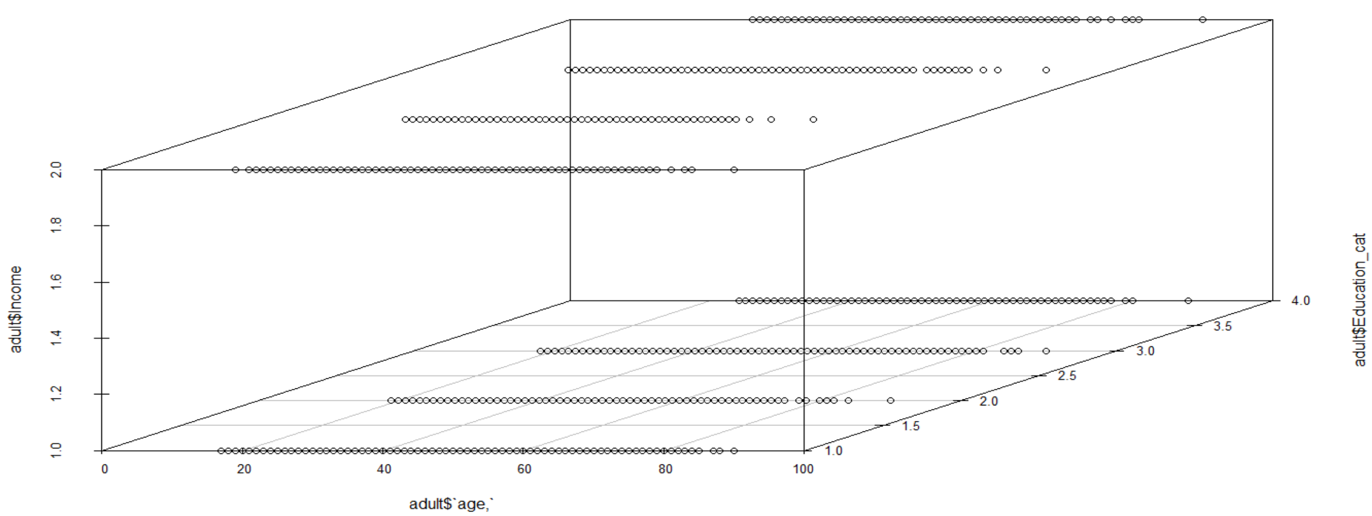
#I am also interested to see education strata wise ,age wise income distribution among the respondents. I thus ,would like to draw a coplot with these variables. I further clubbed the much segregated education into 4 major groups like "school","HS","Grad","Higher" to check the income distribution wrt. their age,

summary(adult$`Education-Num`)

Education_cat <- cut(adult$`Education-Num`, breaks = c(0,10,12,14,16),labels = c("school","HS","Grad","Higher"))

adult <- cbind(adult,Education_cat)

```
scatterplot3d::scatterplot3d((adult$Income ~ adult$`age,`| adult$Education_cat)
```



#We can clearly see the age group & stratified income distribution(>50k &<50k) with respect to 4 Education Category that we have just defined.

#With basic education the income is stagnant after a certain point of time

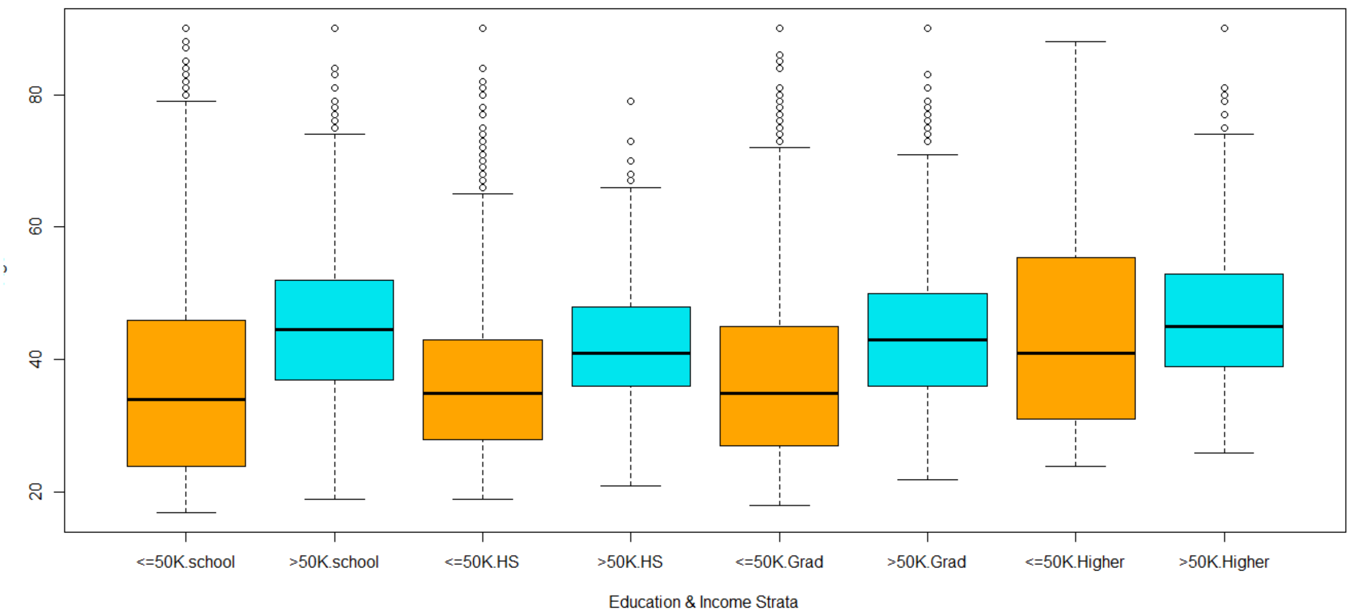
#Majority of the participants are from lower education side having lower income

Higher the education the more stagnant is the income. Even there are majority from higher education group who are earning same and even less than the other groups

This can be further established if we boxplot Income wrt. Age, Education & Income, windows(10,10)

par(mfrow= c(1,2))

boxplot(adult$`age,`~ adult$Income + adult$Education_cat,legend= T,col = c("orange","turquoise2"), xlab = "Education & Income Strata", ylab = "Age")



#The median value of the school goers (<= 50k income) is much lower than that of >=50k group. In both the cases the outvalues are there. This is due to the ~9th -~10th standard contribution in the technical fields. The <7th standards earns much less that the average.

#The median value difference of <=50 income group and the median value for its generic group of income group >=50k are there but in both the cases 3rd quartile contribution is high than their lower counter part. This is possibly these group earns through business or being contributors as important labour force.

#If we compare grads students 3rd quartile value from <=50k income group with 3rd quartile value of from the same income group from school the finding is quite annoying. The 3rd quartile value of the latter group is slightly higher than the former group.

This is possibly for the lower education upper crust group would have got more time to work and earn than that of the their grad counter part.

#The school pass out >= 50 k income group surpass >=50k graduate in both their median and 3rd quartile value.

#Highly educated individuals are surprisingly earning less (We can make out that the number of graduates earning >50K are more than the high school or upper-primary school educated.

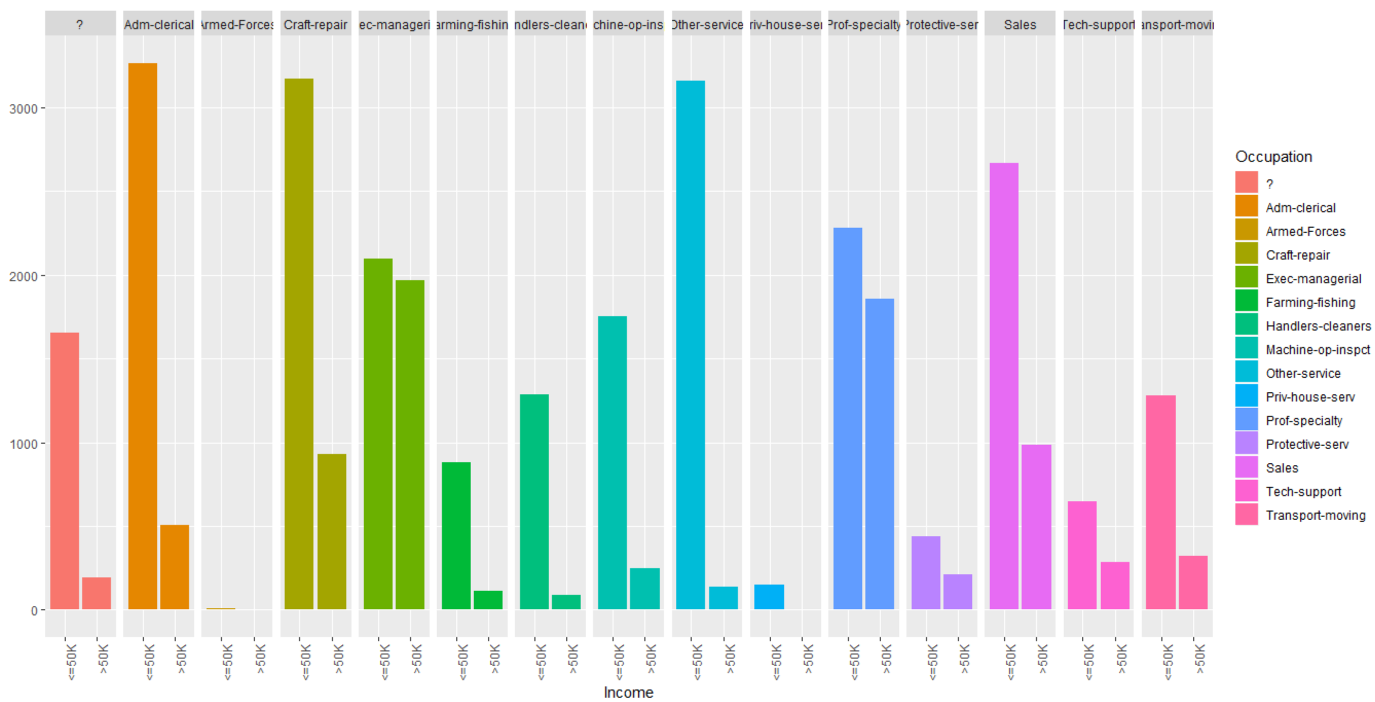
However, we also notice that they are certainly higher in number when compared to master's or Phd holders It is also unfortunate to know that there are roughly 10% of people (*n=94*) with doctorate degrees working in low-skilled jobs and they are earning even less than 50K per annum.) than all the other group strengths reflecting probably the un availability of jobs wrt their merits and skills but from the same education group only the over 50k earners are surpassing all the other groups with respect to their 1st quartile,median and 3rd quartile values.

adulta <30 earn <=50k and those >= 45 hail from >= 50k income strata

Exploring the relationship among income category & occupation.Does the relationship change with race & country?

Let's check out ,

```
> qplot(data = adult,Income, fill = Occupation) + facet_grid ( ~ Occupation)+ theme(axi
s.text.x = element_text(angle = 90, hjust = 1))
```



#If the dataset is true representation of the population then the earning disparities exist across 15 Occupation classes.

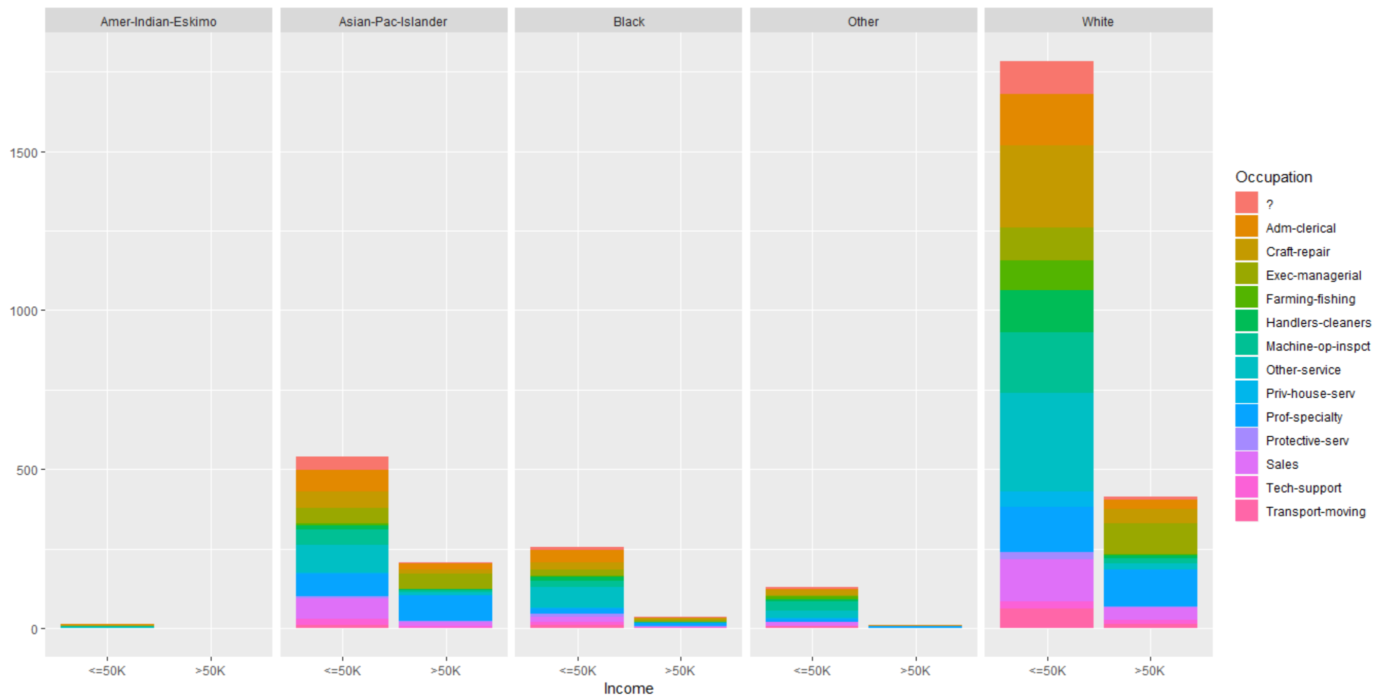
#The data from Armed forced is insufficient enough to be considered.

#The gap is minimum in Exec- Managerial category followed by Specialty professionals.(where the education and experience would have played major role but we cannot conclude anything without checking the same further).

#The class difference is maximum in other services, Handlers & Cleaners and in Admn- Clerical section. There are significant difference exit in Sales profession & Transport moving sectors as well.

<u>Plotting Income Distribution wrt. Occupation & Race</u>

qplot (data = pi,Income, fill = Occupation) + facet_grid (. ~ Race)+ theme(axis.text.x = element_text(angle = 90, hjust = 1)) # For rest of the countries



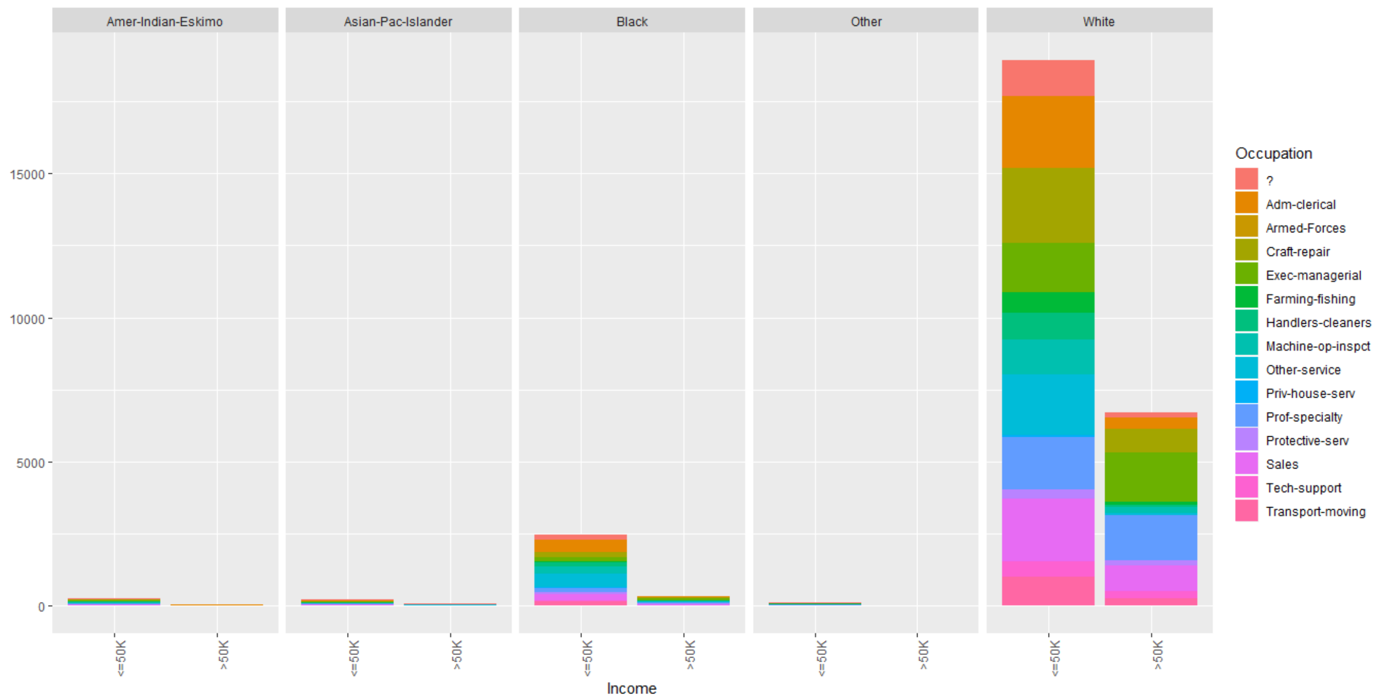
#The racial discrimination in earning is rampant . I have thus segregated the other countries from the USA. Still we can see White earns obnoxiously more than the backs Asian Islander or than the other race.

#Prof Specialty is still being the lucrative high income potential category followed by Management Executive and Tech Support system.

#But in these over 50K specialty category also Whites are paid way more than their black, other and Is lander counterpart.

#For USA

qplot (data = pu,Income, fill = Occupation) + facet_grid (. ~ Race)+ theme(axis.text.x = element_text(angle = 90, hjust = 1))



#This Racial discrimination is prominent if we check Black vs White specialty Income sectors alone. Same skill set yet black earns <50k in sales, Admn –Official, and Exc- Managerial sectors.

#The most unfortunate sector is that in Private House Service sector also Black laborers are more but paid least compared to the White labour counterpart.

Plotting Income Distribution wrt. Occupation & Native Country

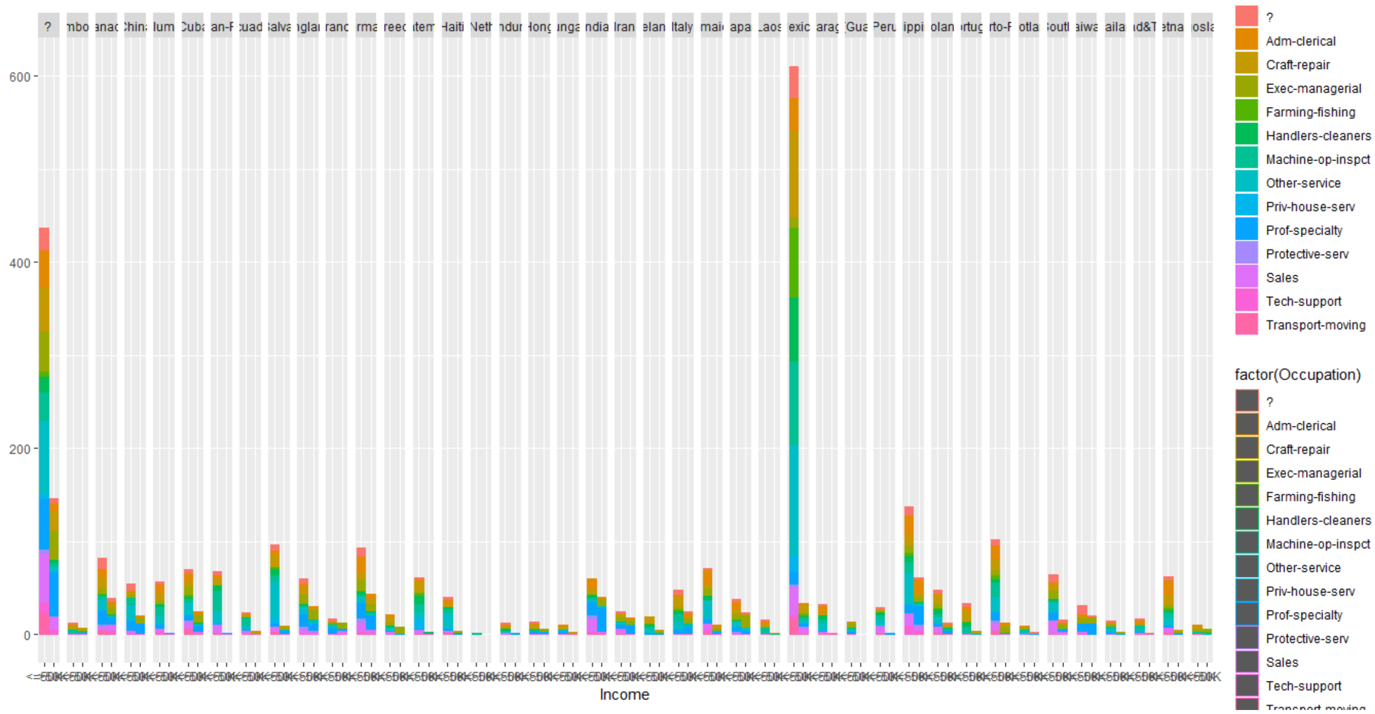
#For Other Countries

```
qplot(data = pi,Income, fill = Occupation, color = factor(Occupation)) + facet_grid (. ~ Native_Country)
```

#Though the number of participants are extremely less from the other countries but still it is visible that the world is clearly divided into two halves with respect to earning though catering to the same category of occupations.

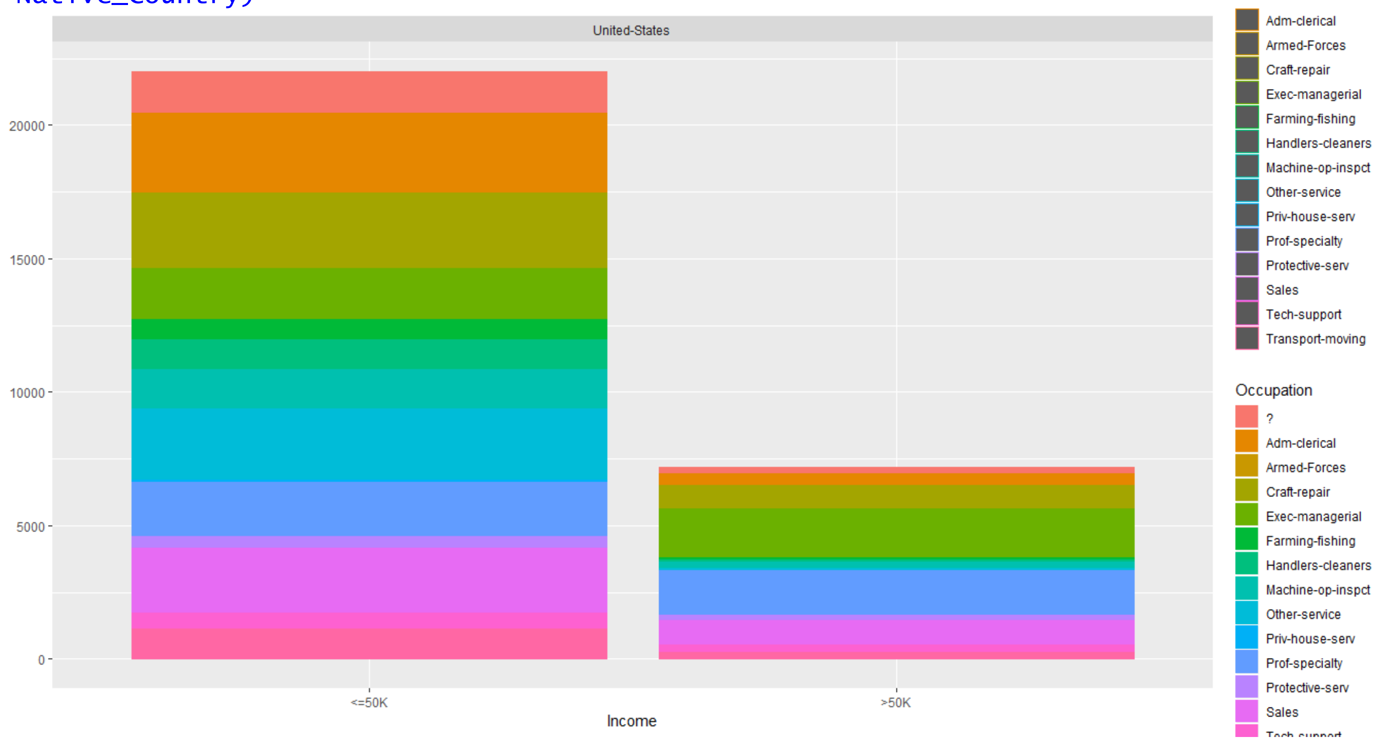
#We did not consider the USA respondents in this graph still it is seen Asia, Africa, South East Asian Countries , even South American developing countries are earning <50k / annum .The obvious currency convertibility ratio could be fetched as a plausible solution but we need to remember that this data domain is USA only and this data is collected from (Extraction was done by Barry Becker from the 1994 Census database) Census # Please refer https://archive.ics.uci.edu/ml/datasets/adult.

Thus the discrimination is reflecting the Income disparity prevailing in the US soil only.



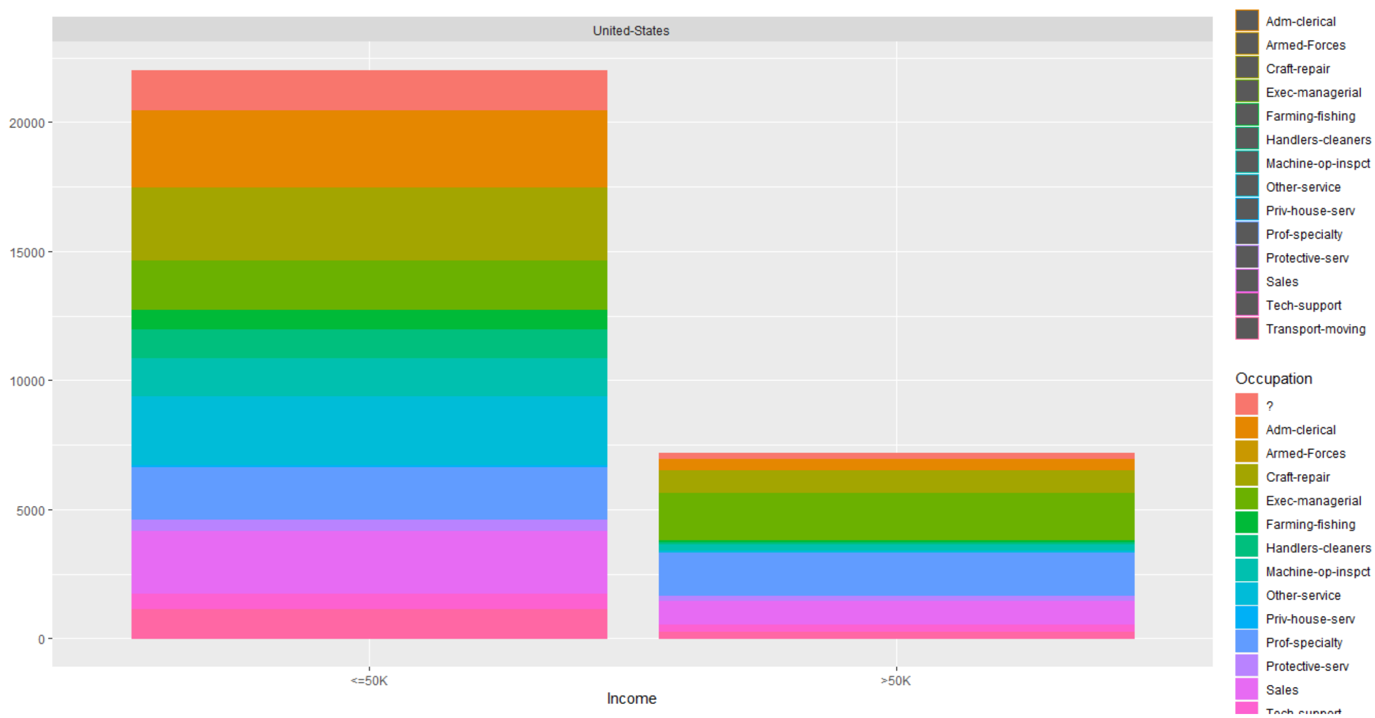
#For the USA

```
qplot(data = pu,Income, fill = Occupation, color = factor(Occupation)) + facet_grid (. ~ Native_Country)
```



Even for the US the Income discrimination is huge. Around 70% of the respondents belong to <50k Income category.

#The category of high potential income sectors remain the same as per our above findings. Since the USA natives are over 80% the USA finding alone has become the represntation of the adult dataset being the significant contributor to all the variables inside the dataset.



Let's Check if the following variables are worth studying : -

The objective is to find out the following factors influencing INCOME as dependent variable –

a. Marital Status
b. Relationship
c. Capital Gain / Capital Loss   8 + 7 + 5 + 5 = 25 Marks

Before I go for a visual mapping let us explore the Mathematical contribution of the variables, Marital Status, Relationship and Capital Gain/Loss as independent predictors of dependent variable Income.

a) table(adult$Marital_Status,adult$Income)

```
                         <=50K  >50K
  Divorced                3980   463
  Married-AF-spouse         13    10
```

```
  Married-civ-spouse     8284  6692
  Married-spouse-absent   384    34
  Never-married         10192   491
  Separated               959    66
  Widowed                 908    85
```

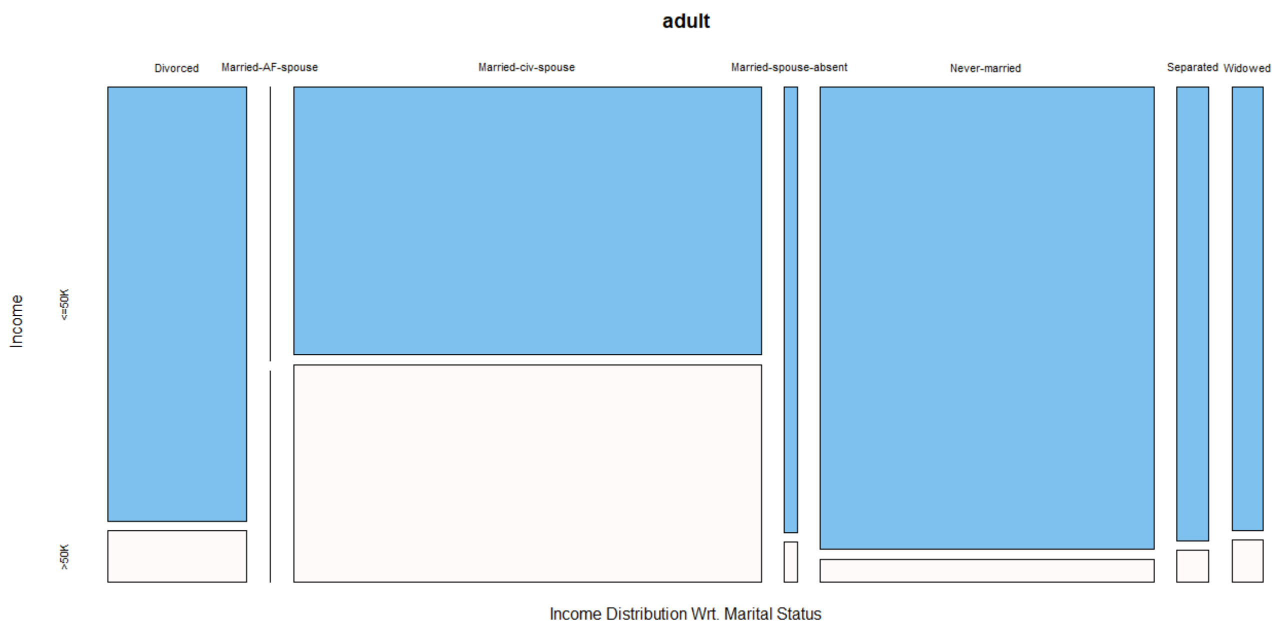
#From the above table I would like to conduct Chi Square test to check if these 2 variables are related.

chisq.test(table(adult$Marital_Status,adult$Income))

```
chisq.test(table(adult$Marital_Status,adult$Income))

        Pearson's Chi-squared test

data:  table(adult$Marital_Status, adult$Income)
X-squared = 6517.7, df = 6, p-value < 2.2e-16
```

#The result reveals that the p value is << than alpha = 5% thus I reject the null hypothesis that there is no difference between the means and conclude accepting the alternative hypothesis that significant difference between means exist among the groups and marital status play an important role in Income status too.

```
> mosaicplot(~ Marital_Status + Income , data = adult,color = c("skyblue2","snow1"), xl
ab = "Income Distribution Wrt. Marital Status")
```



#Here listing some interesting observations from the above plot :

Divorced, Married but spouse absent, Never Married, Separated and Widowed are almost collectively below strata people with <= 50k annual income.

#This is probably the joint income of the married people contributing to their Income status (>50k)

Married section is a big fat section with maximum participants where >50 group is slightly lower than <=50 in strengths.

#The next big group is never married group and they have least numbers of > 50k/anm participants.

#The 3rd big group is Divorced with fairly <=50/anm Income strengths.

b) #Relationship

#Let us check by conducting Pearson's chisquare test assuming equal means from the various groups in H0. The alternative Hypothesis Ha to be there is significant difference among the means.

```
> table(adult$Relationship,adult$Income)

                 <=50K >50K
  Husband         7275 5918
  Not-in-family   7449  856
  Other-relative   944   37
  Own-child       5001   67
  Unmarried       3228  218
  Wife             823  745
```

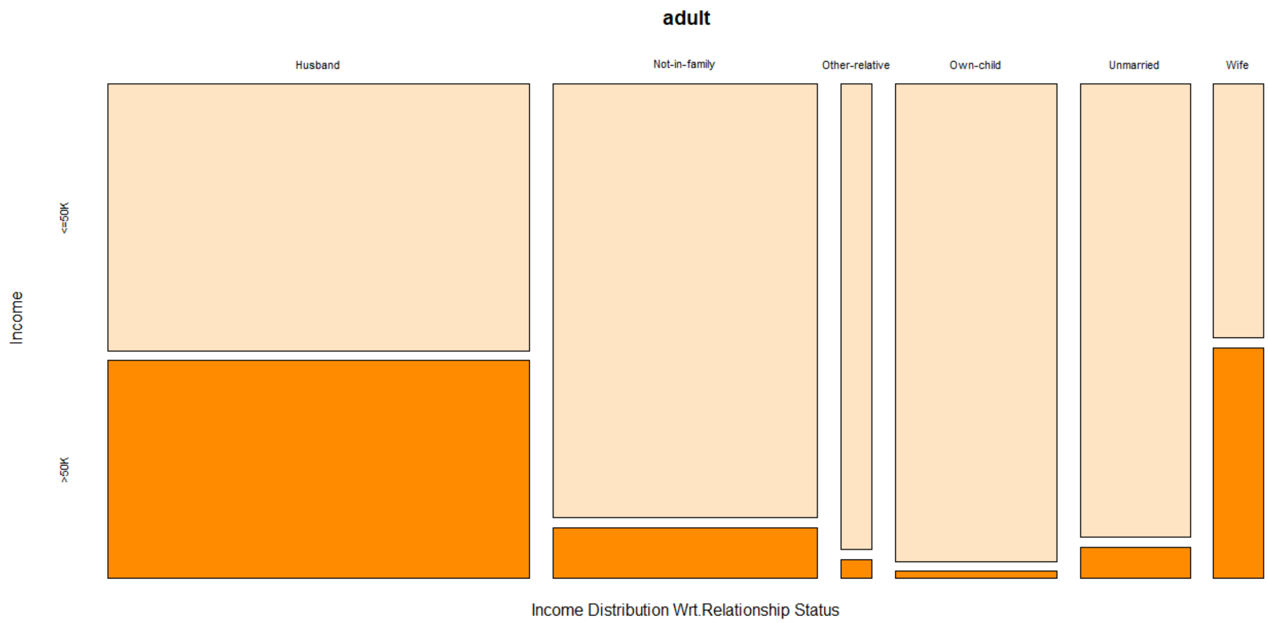
chisq.test(table(adult$Relationship,adult$Income))

```
Pearson's Chi-squared test

data:  table(adult$Relationship, adult$Income)
X-squared = 6699.1, df = 5, p-value < 2.2e-16
```

#Since p value of the Chsq test is << alpha = 5% we would reject the null hypothesis and accept there are significant difference between various relationship status wrt. their mean income.In another word, Relationship status play significant role for Income distribution.

Again want to plot a Mosaic Plot :

mosaicplot(~ Relationship + Income , data = adult,color = c("bisque","darkorange"), xlab = "Income Distribution Wrt.Relationship Status")



This is again an interesting revelation "

#Wives are richest as >50K/Anm Income group. They are almost 50% of their generic respondents community.

#Husbands are the second richest community with >50k /anm Income .They are holistically the maximum respondent group as well. But the <=50k Husband group is dominating in over all Husband group.

#The next largest respondents are not having any family and the rich from this group are significantly low in percentage.

Those who own a child are least richest group from all the Relationship groups.The maximum percentage of the respondents are <=50K/Anm category.

#Unmarried riches are only few (way below in percentage) than their <=50 /anm counterpart.

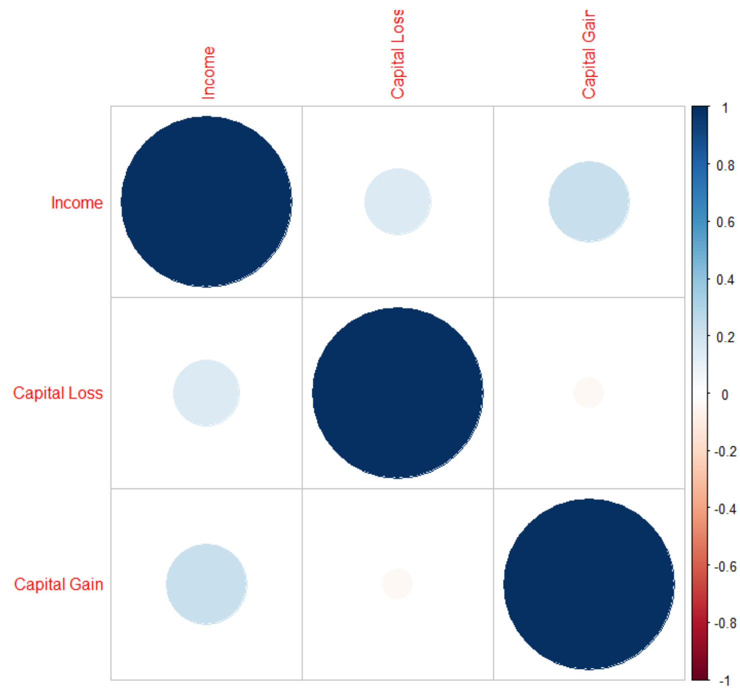
#The other relative section is also dominated by the <=50k Income group.

c) Income and Capital Loss/Capital Gain

library(corrplot)

```
inccap <- cbind(adult$Income,adult$Capital_Loss,adult$Capital_Gain)
```

```
> colnames(inccap) <- c("Income","Capital Loss","Capital Gain")
> corrplot(cor(inccap))
```



The correlation plot reveals that there is a significant relationship existing between Income and Capital Gain and for Capital Loss the relationship is strong but having relatively lesser strength than Capital Gain.

I wanted to see the mathematical relationship among these three variables and thus would like to run the cor function to know the coefficients,

```
> cor(inccap)
                Income Capital Loss Capital Gain
Income       1.0000000   0.15052631   0.22332882
Capital Loss 0.1505263   1.00000000  -0.03161506
Capital Gain 0.2233288  -0.03161506   1.00000000
```

#Clearly reflecting and explaining the above plot only. Capital Loss and Capital gain has no correlation between them.

Examining the relationship between income category, education (number), age (number) for various occupation and comment on the findings.

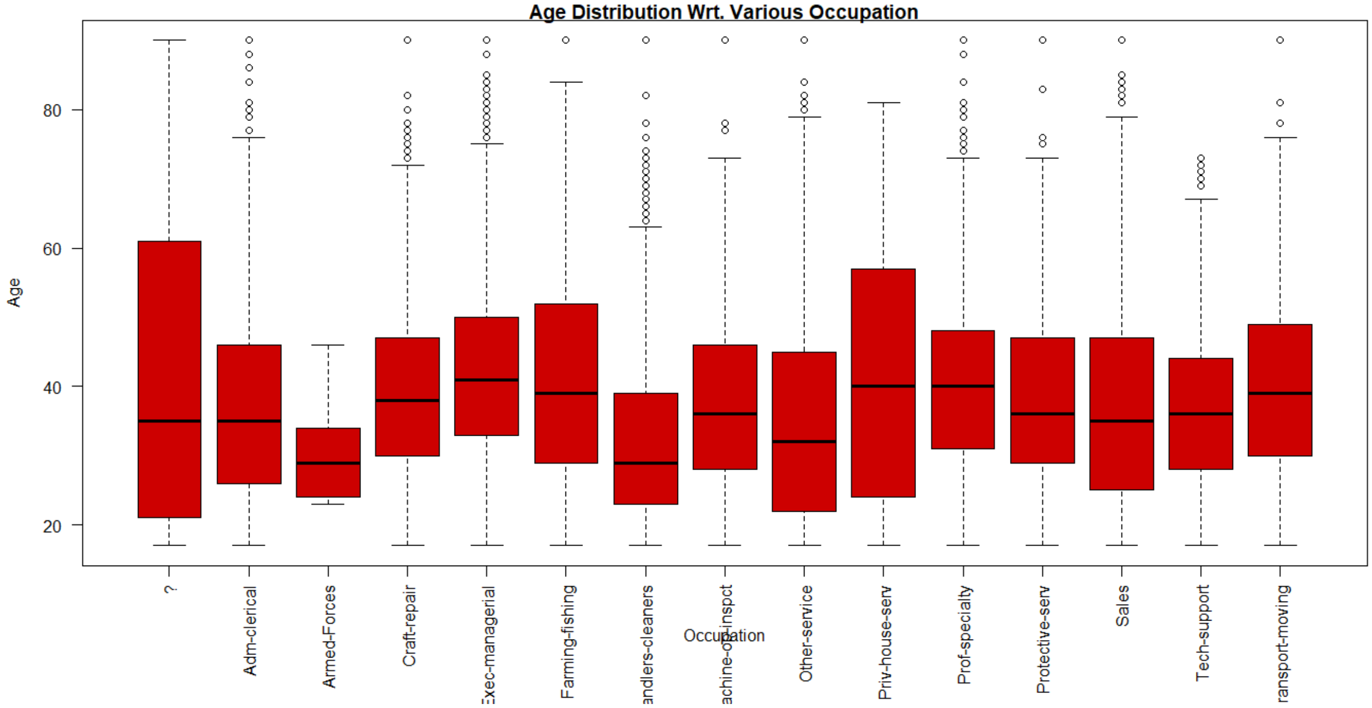
For better visualization :-

Let us check age distribution across Occupation

windows(10,10)

par(mar=c(7,5,1,1))

```
> boxplot(adult$`age,` ~ adult$Occupation,main = "Age Distribution Wrt. Various Occupat
ion",xlab = "Occupation", ylab = "Age", col = "red3",las = 2)
```



Age being a continuous variable has a significant impact on Income distribution that we have seen earlier as well. Now thi graph is representing the range of Age that a particular job can accommodate, Median and 3rd quartile of age in each sector and the pareto between the jobs wrt age.

It is visible that "?" sector is accommodating maximum number of age group people with a median value of 35-36 roughly which is the median value of Adm – Clerical, Prospective Services, Sales, Tech Support,Machine Inspect .

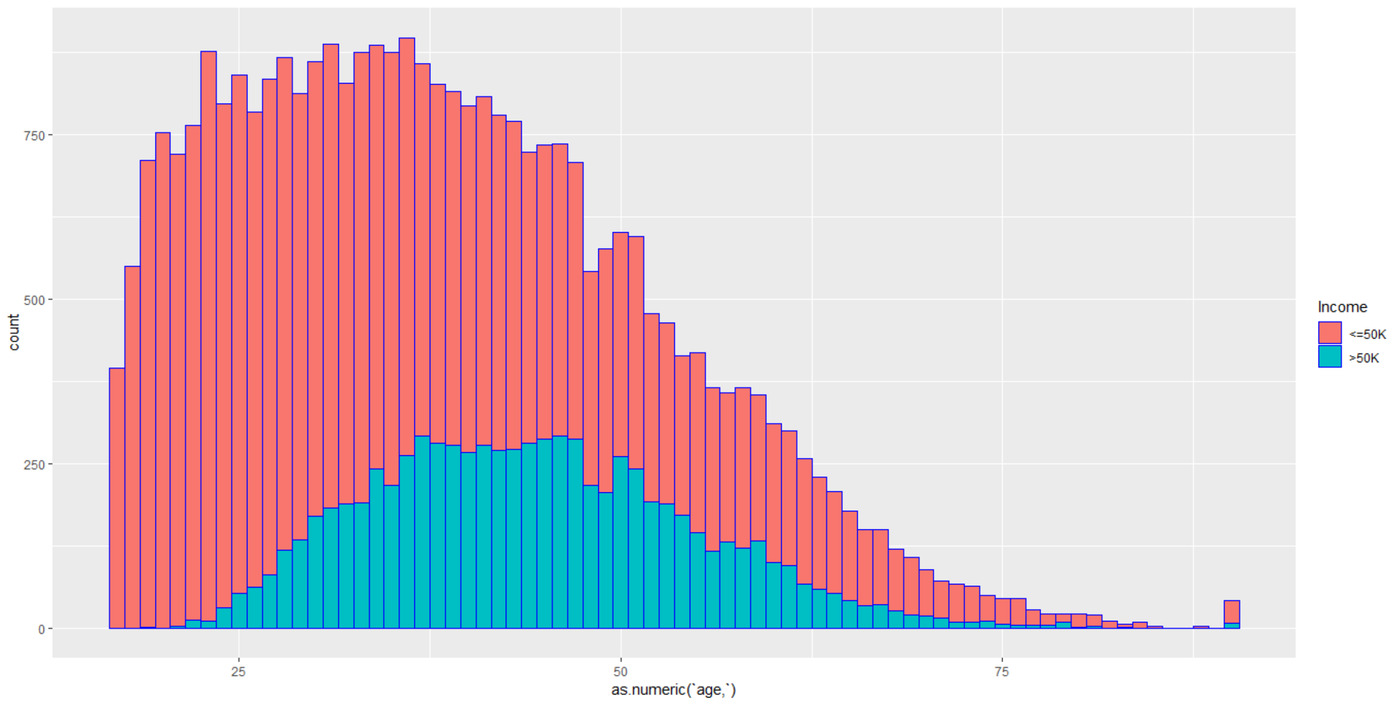
Armed forces has a very specified and defined equi distant quartiles which is possibly reflecting the strict Armed force joining and retiring criteria. This section is not having any outliers as well.

Cleaners category has maximum number of outliers . Though the median value is 25 yrs but people from >62 to >8o years are also available in this sector . This is possibly due to choosing this job as a leaving measure by the other wise jobless, unemployable elderly people.

Median value of Executive Managerial, Private House Service, Prof Specialty and Transport moving are nearly same ,41 years. The 3rd quartile value for Private House Service sector ends at the age of 58 with no outliers.

Craft repair, Prof Specialty, Tech support sectors are showing outliers reflecting that skilled worker are in demand irrespective of their age bracket.

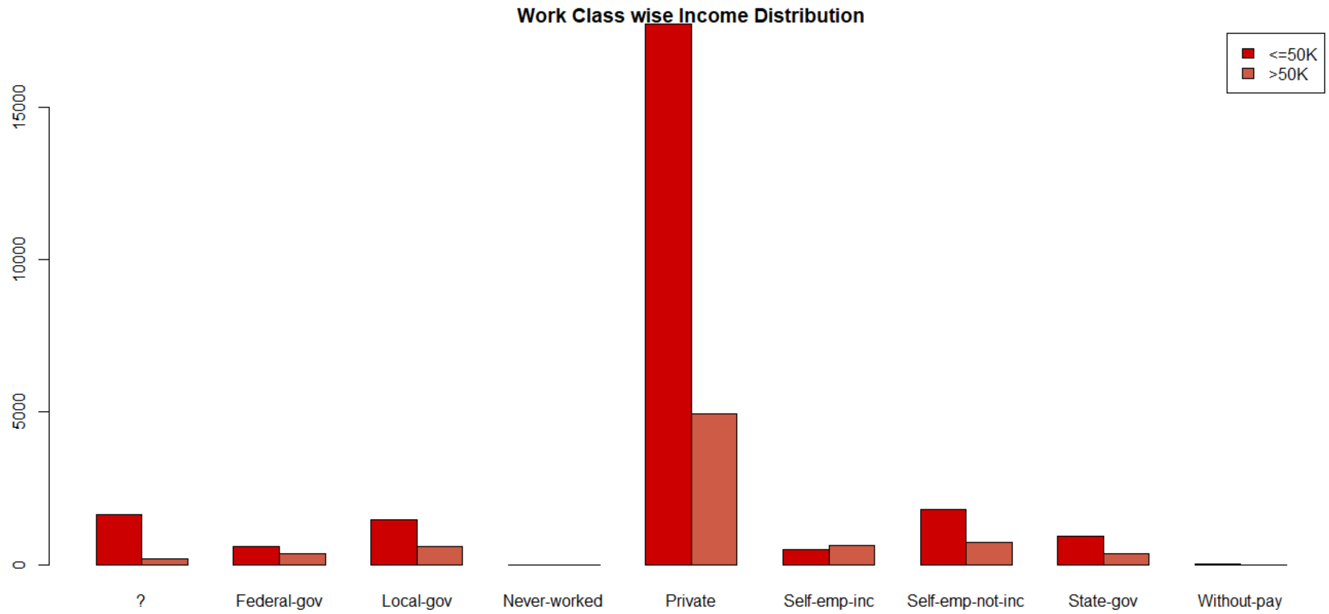
#Wanted to Explore Income Category distribution wrt. Age.



It is visible from the above plot that most of the respondents are from <50$k category. The people who belong from >50$k per annum are basically at their midcareer levels (33 to 52).

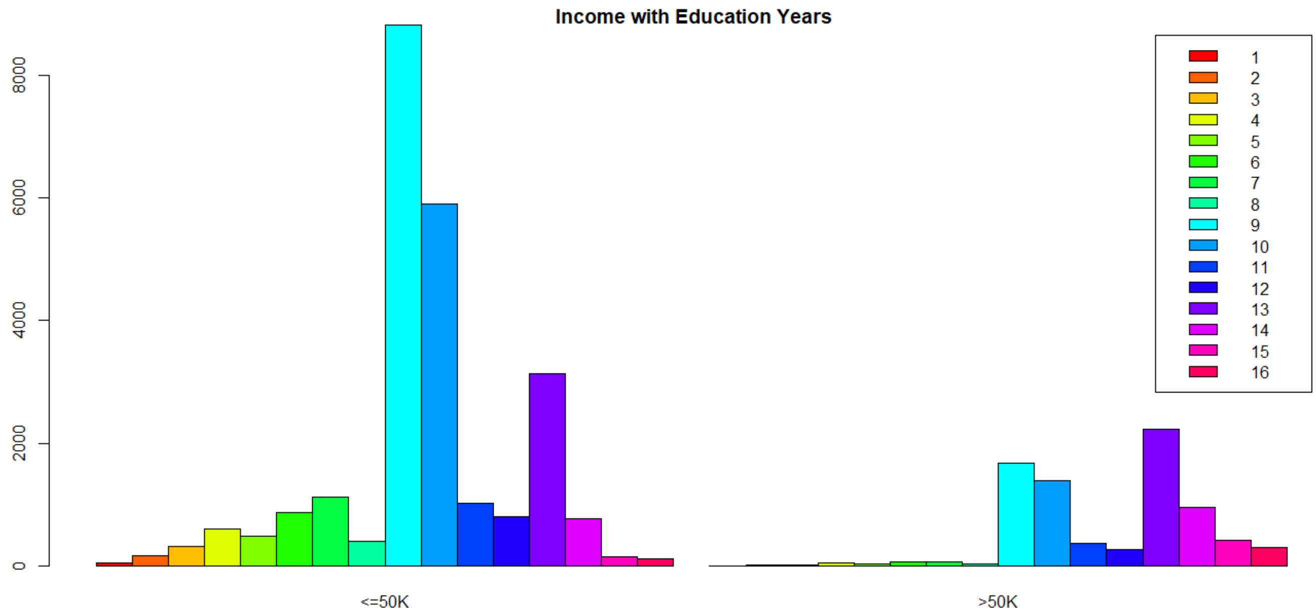
Let us see the table of the Work Class & the Income distribution :

```
> table(adult$Income,adult$workclass)

            ? Federal-gov Local-gov Never-worked Private
  <=50K  1645         589      1476            7   17733
  >50K    191         371       617            0    4963

       Self-emp-inc Self-emp-not-inc State-gov Without-pay
  <=50K          494             1817       945          14
  >50K           622              724       353           0


> barplot(table(adult$Income,adult$workclass),legend = T, col = c("red3","coral3"),besi
de = T, main = "Work Class wise Income Distribution")
#Only “self employed inc” section earns >50k$/annum out numbering its <50k$/annum count
er part.
# >50k$/annum and <50k$/annum earners both belong from Private sector.
# Federal(4:3) & Local Govt.(2:1) both have more <50k$/annum than that of >50k$/annum b
ut the difference is not that significantly high.

> count <- table(adult[adult$workclass == 'Government',]$Income)["<=50K"]
> count <- c(count, table(adult[adult$workclass == 'Government',]$Income)[">50K"])
> count <- c(count, table(adult[adult$workclass == 'Other/Unknown',]$Income)["<=50K"])
> count <- c(count, table(adult[adult$workclass == 'Other/Unknown',]$Income)[">50K"])
> count <- c(count, table(adult[adult$workclass == 'Private',]$Income)["<=50K"])
> count <- c(count, table(adult[adult$workclass == 'Private',]$Income)[">50K"])
> count <- c(count, table(adult[adult$workclass == 'Self-Employed',]$Income)["<=50K"])
> count <- c(count, table(adult[adult$workclass == 'Self-Employed',]$Income)[">50K"])
> count <- as.numeric(count)
```



Let us check the Year of Education & Income Group:

```
> windows(10,10)
> par(mar=c(7,5,1,1))
> table(adult$`Education-Num`,adult$Income)

     <=50K >50K
  1     51    0
  2    162    6
  3    317   16
  4    606   40
  5    487   27
  6    871   62
  7   1115   60
  8    400   33
  9   8826 1675
  10  5904 1387
  11  1021  361
  12   802  265
  13  3134 2221
  14   764  959
  15   153  423
  16   107  306
> barplot(table(adult$`Education-Num`,adult$Income), col= rainbow(16),beside = T, main
= " Income with Education Years",legend = T)
```




```
> chisq.test(table(adult$`Education-Num`,adult$Income))

        Pearson's Chi-squared test

data:  table(adult$`Education-Num`, adult$Income)
X-squared = 4429.7, df = 15, p-value < 2.2e-16
```

#The low p value < alpha @5% signifies we should reject the null hypothesis which states that the mean income among 16 years of education are same.

We should accept the alternative hypothesis and conclude that there is significant effect on income distribution across years of education.

But there lies an anomaly which is reflecting from the above graph.

Chisquare test does not tell us why :

#The drop outs are highest after 9 -10 years of education. This is the two classes that represents maximum number of respondents from <50k$/anm income group.

#Surprisingly this two years 9-10 are the 2^{nd} and 3^{rd} largest contributors of >50k$/anm group too.

#The 13 years of education are the highest bracket for > 50k$/anm income group where as the same years of education is the 3^{rd} largest frequency from <=50k$/anm income group.

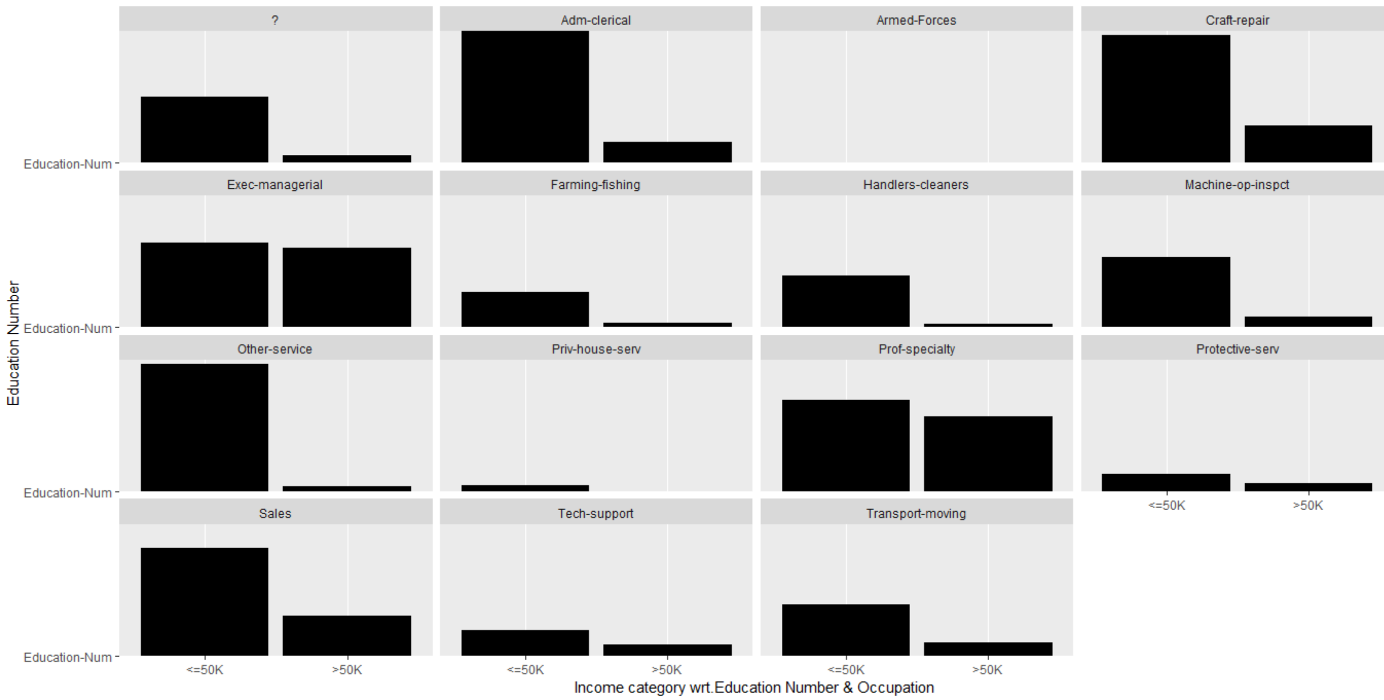
#Above 13 years of education does not ensure an earning >50k$/anm though majority of >14 yrs of education earns >50k$ /anm than their fellow <50k$/anm counter part.

<=8 years of education prominently ensures an earning <50k$/anm though little exceptions are there. Few of them do earn >50k$/anm.

Here is the category wise income distribution. This is self-explanatory which we have already discussed above in various forums.

library(ggplot2)

ggplot(adult, aes(x= Income,y='Education-Num')) + geom_bar(stat = "identity", fill = c("73"))+facet_wrap(~Occupation)+ylab("Education Number") + xlab("Income category wrt.Education Number & Occupation ")



#Observations:

We don’t know if the data is representative or not. We have seen previously a racial and gender discrimination during our previous analysis. We have seen there is an absence of uniformity from the various respondent classes but if we consider this dataset as representative dataset then there are following significant observations that are worth mentioning:

#All the sectors have wages discriminations but the same is least in Exec- Managerial sector followed by Prof-specialty and Tech Support sectors.

#Wages discrimination is highest in Other Services followed by Admn-Clerical , Handlers and Cleaners sectors.

If the job is specialty or technical where professional expertise play a pivotal role the differences of the two income group are less but the works those are labor intensive have high discriminations.

Finally I wanted to check the interaction between Categorical Variable Income wrt. Continuous variables Age & Number of Years of Education and wanted to know if they are plotted what the graph would appear to be.

Here is the finding :

plot(lm(adult$Income ~ adult$`age,` + adult$`Education-Num`)) windows(10,10)

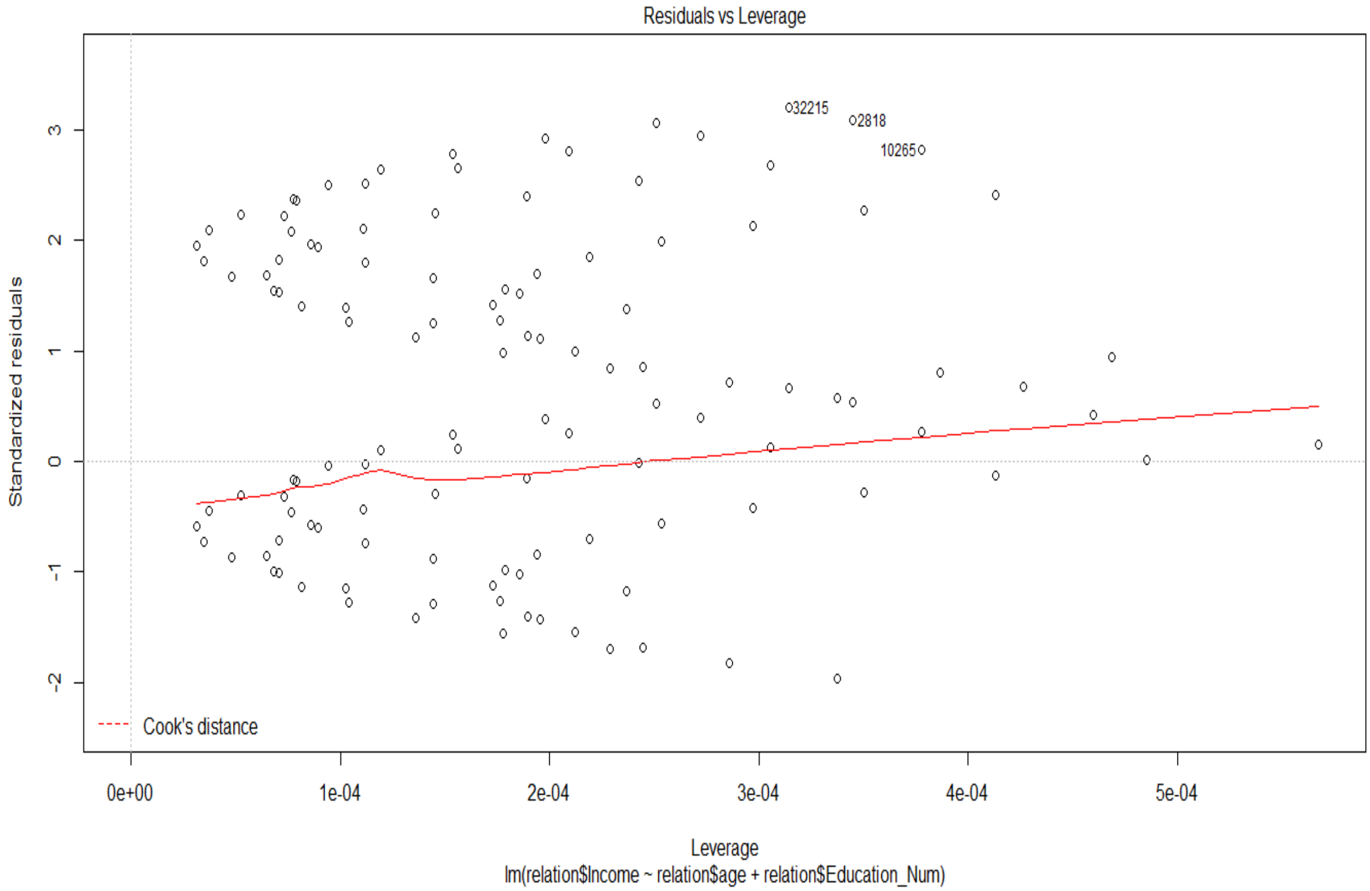
```
> lm(formula = relation$Income ~ relation$age + relation$Education_Num)

Call:
lm(formula = relation$Income ~ relation$age + relation$Education_Num)

Coefficients:
           (Intercept)            relation$age  relation$Education_Num
               0.46607                 0.10495                 0.05552
```

Thus Income = .46607 + .10495*Age + .05552*Education_Num

#But the above model should not to be considered as good predictor of Income as there are other important variables like , fnlwgt, marital status missed in the above model (I adhered to the Q - 4 which does not allow me to talk about fnlwgt and marital status here. But in the end I will explore these two variables too and will build my model) .But Age and Education Number having least p values indicate they are the two most important variables that are essential for income prediction.

Here is our graph,

```
> plot(lm(formula = relation$Income ~ relation$age + relation$Education_Num))
```



Now for academic interest only I want to build my Income predictor model (I called prcomp function to run a PCA to identify the Principal Components followed by plotting the Scree plot. I then run a regression analysis with all the independent variable fixing Income to be predicted and had zeroed upon Age, Education_Num, fnlwgt and marital status Cap_Gain &Cap_Loss as most significant predictors)

```
> summary(lm(as.numeric(adult$Income) ~ adult$`age,` + adult$`Education-Num`+adult$fnlw
gt + as.numeric(adult$Marital_Status) + adult$Capital_Gain + adult$Capital_Loss))

Call:
lm(formula = as.numeric(adult$Income) ~ adult$`age,` + adult$`Education-Num` + 
    adult$fnlwgt + as.numeric(adult$Marital_Status) + adult$Capital_Gain + 
    adult$Capital_Loss)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9830 -0.2574 -0.1134  0.1436  1.2628 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      6.333e-01  1.339e-02  47.288  < 2e-16 ***
adult$`age,`                     5.397e-03  1.604e-04  33.650  < 2e-16 ***
```

```
adult$`Education-Num`              4.830e-02  8.258e-04  58.486  < 2e-16 ***
adult$fnlwgt                       8.456e-08  1.993e-08   4.243 2.21e-05 ***
as.numeric(adult$Marital_Status) -3.473e-02  1.446e-03 -24.015  < 2e-16 ***
adult$Capital_Gain                 1.000e-05  2.870e-07  34.850  < 2e-16 ***
adult$Capital_Loss                 1.261e-04  5.230e-06  24.111  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3781 on 32554 degrees of freedom
Multiple R-squared:  0.2183,   Adjusted R-squared:  0.2182
F-statistic:  1516 on 6 and 32554 DF,  p-value: < 2.2e-16
```

All the p values are << than alpha @5% indicating we should reject the null hypothesis and accept that these variables are having significant role in income prediction.

Here is the final graph :

```
> plot(lm(as.numeric(adult$Income) ~ adult$`age,` + adult$`Education-Num`+adult$fnlwgt
+ as.numeric(adult$Marital_Status) + adult$Capital_Gain + adult$Capital_Loss)
+ )
```