Income Classification-Visualization

Geethika

#Loading the dataset

Analysing the dataset for various insights

```
incomeclassi<- read.csv("income_evaluation.csv")

dim(incomeclassi)
```

```
## [1] 32561    15
```

```
str(incomeclassi)
```

```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass     : chr  " State-gov" " Self-emp-not-inc" " Private" " Private" ...
##  $ fnlwgt        : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
##  $ education     : chr  " Bachelors" " Bachelors" " HS-grad" " 11th" ...
##  $ education.num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marital.status: chr  " Never-married" " Married-civ-spouse" " Divorced" " Married-civ-spouse" ...
##  $ occupation    : chr  " Adm-clerical" " Exec-managerial" " Handlers-cleaners" " Handlers-cleaners" ...
##  $ relationship  : chr  " Not-in-family" " Husband" " Not-in-family" " Husband" ...
##  $ race          : chr  " White" " White" " White" " Black" ...
##  $ sex           : chr  " Male" " Male" " Male" " Male" ...
##  $ 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.week: int  40 13 40 40 40 40 16 45 50 40 ...
##  $ native.country: chr  " United-States" " United-States" " United-States" " United-States" ...
##  $ income        : chr  " <=50K" " <=50K" " <=50K" " <=50K" ...
```

```
head(incomeclassi)
```

```
##   age         workclass fnlwgt  education education.num      marital.status
## 1  39         State-gov  77516  Bachelors            13       Never-married
## 2  50  Self-emp-not-inc  83311  Bachelors            13  Married-civ-spouse
## 3  38           Private 215646    HS-grad             9            Divorced
## 4  53           Private 234721       11th             7  Married-civ-spouse
## 5  28           Private 338409  Bachelors            13  Married-civ-spouse
## 6  37           Private 284582    Masters            14  Married-civ-spouse
##           occupation   relationship   race     sex capital.gain capital.loss
## 1       Adm-clerical  Not-in-family  White    Male         2174            0
```

```
## 2    Exec-managerial        Husband  White    Male            0            0
## 3  Handlers-cleaners  Not-in-family  White    Male            0            0
## 4  Handlers-cleaners        Husband  Black    Male            0            0
## 5     Prof-specialty           Wife  Black  Female            0            0
## 6    Exec-managerial           Wife  White  Female            0            0
##   hours.per.week native.country income
## 1             40  United-States  <=50K
## 2             13  United-States  <=50K
## 3             40  United-States  <=50K
## 4             40  United-States  <=50K
## 5             40           Cuba  <=50K
## 6             40  United-States  <=50K
```

```
summary(incomeclassi)
```

```
##       age          workclass            fnlwgt          education
##  Min.   :17.00   Length:32561       Min.   :  12285   Length:32561
##  1st Qu.:28.00   Class :character   1st Qu.: 117827   Class :character
##  Median :37.00   Mode  :character   Median : 178356   Mode  :character
##  Mean   :38.58                      Mean   : 189778
##  3rd Qu.:48.00                      3rd Qu.: 237051
##  Max.   :90.00                      Max.   :1484705
##  education.num   marital.status      occupation        relationship
##  Min.   : 1.00   Length:32561       Length:32561       Length:32561
##  1st Qu.: 9.00   Class :character   Class :character   Class :character
##  Median :10.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :10.08
##  3rd Qu.:12.00
##  Max.   :16.00
##      race               sex             capital.gain    capital.loss
##  Length:32561       Length:32561       Min.   :    0   Min.   :   0.0
##  Class :character   Class :character   1st Qu.:    0   1st Qu.:   0.0
##  Mode  :character   Mode  :character   Median :    0   Median :   0.0
##                                        Mean   : 1078   Mean   :  87.3
##                                        3rd Qu.:    0   3rd Qu.:   0.0
##                                        Max.   :99999   Max.   :4356.0
##  hours.per.week  native.country        income
##  Min.   : 1.00   Length:32561       Length:32561
##  1st Qu.:40.00   Class :character   Class :character
##  Median :40.00   Mode  :character   Mode  :character
##  Mean   :40.44
##  3rd Qu.:45.00
##  Max.   :99.00
```

```
# Cleaning the data

#Checking for 'NA' values and number of unique values for each variable.

sapply(incomeclassi,function(x) sum(is.na(x)))
```

```
##            age      workclass         fnlwgt      education  education.num
##              0              0              0              0              0
## marital.status     occupation   relationship           race            sex
```

```
##              0              0              0              0              0
##   capital.gain   capital.loss hours.per.week native.country         income
##              0              0              0              0              0
```

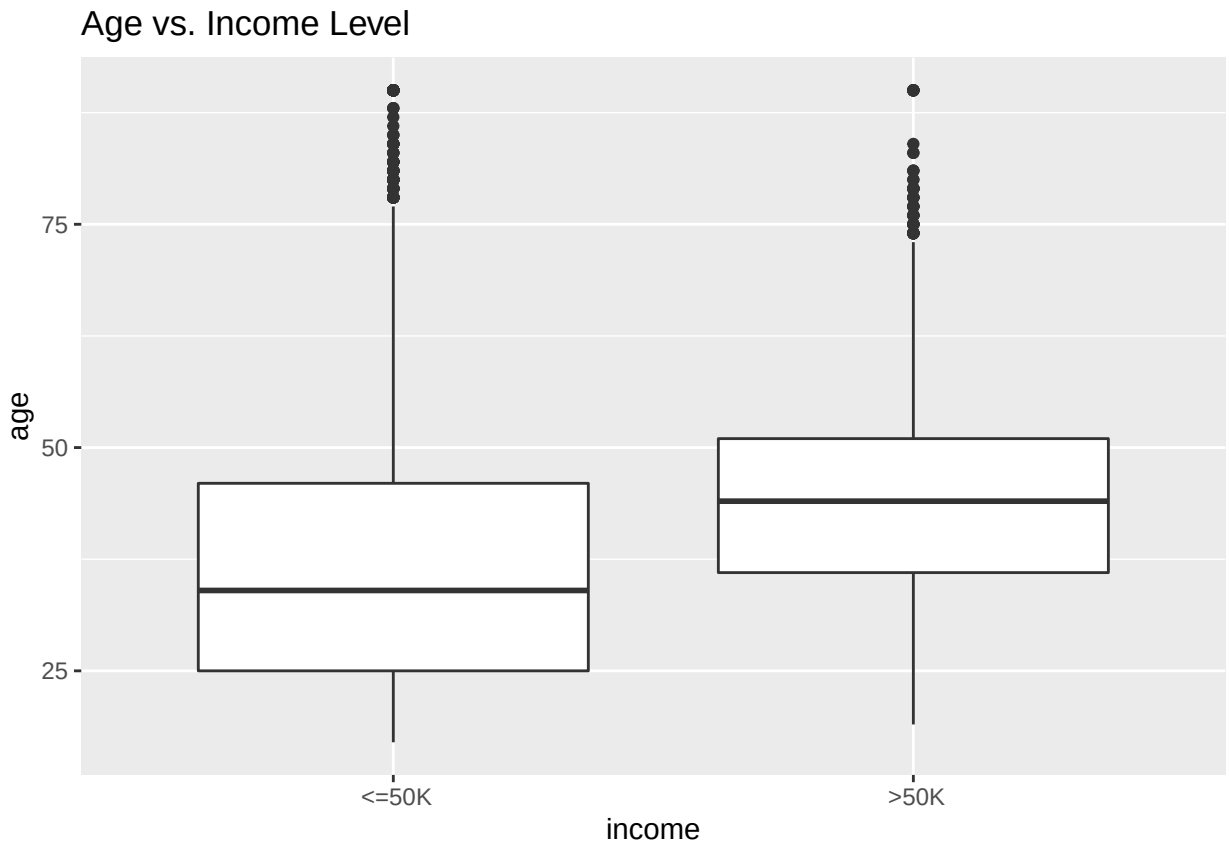
```
sapply(incomeclassi, function(x) length(unique(x)))
```

```
##            age      workclass         fnlwgt      education  education.num
##             73              9          21648             16             16
## marital.status     occupation   relationship           race            sex
##              7             15              6              5              2
##   capital.gain   capital.loss hours.per.week native.country         income
##            119             92             94             42              2
```

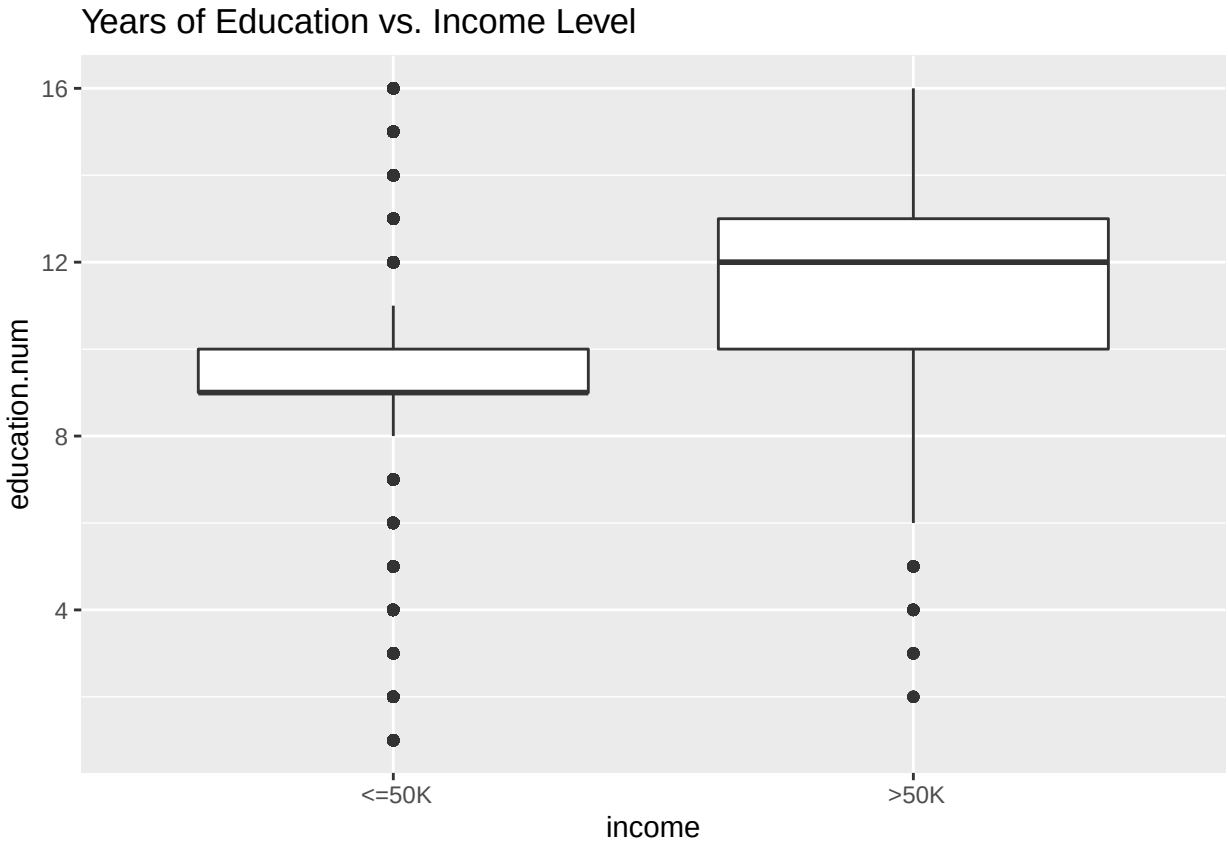
```
#missmap(incomeclassi, main = "Missing values vs observed")
table (complete.cases (incomeclassi))
```

```
##
##  TRUE
## 32561
```

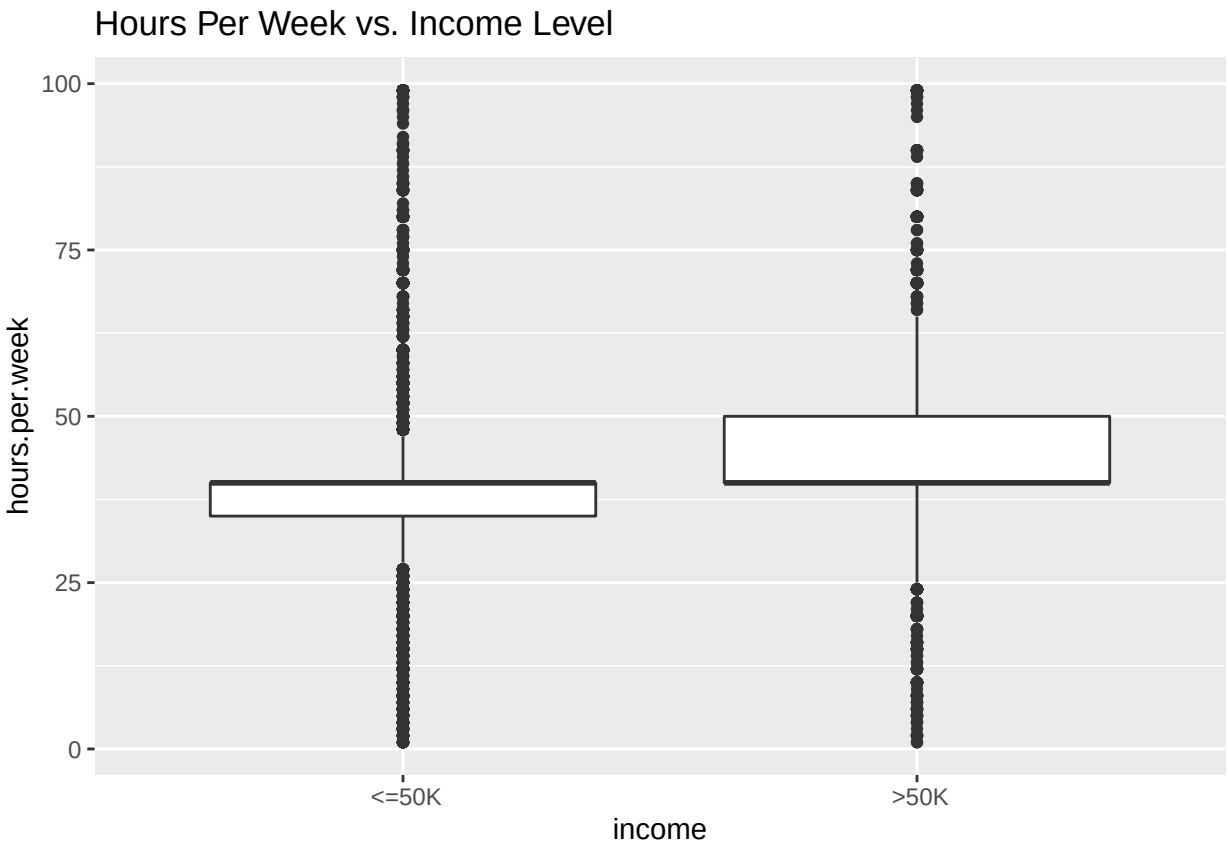
```
#converting the required features to factors

incomeclassi$education <- as.factor(incomeclassi$education)
incomeclassi$workclass<- as.factor(incomeclassi$workclass)
incomeclassi$occupation <- as.factor(incomeclassi$occupation)

#Beginning of exploratory data analysis

#Exploring DATA
#Viz 1
 ggplot(aes(x=income, y=age), data = incomeclassi) + geom_boxplot() +
     ggtitle('Age vs. Income Level')
```



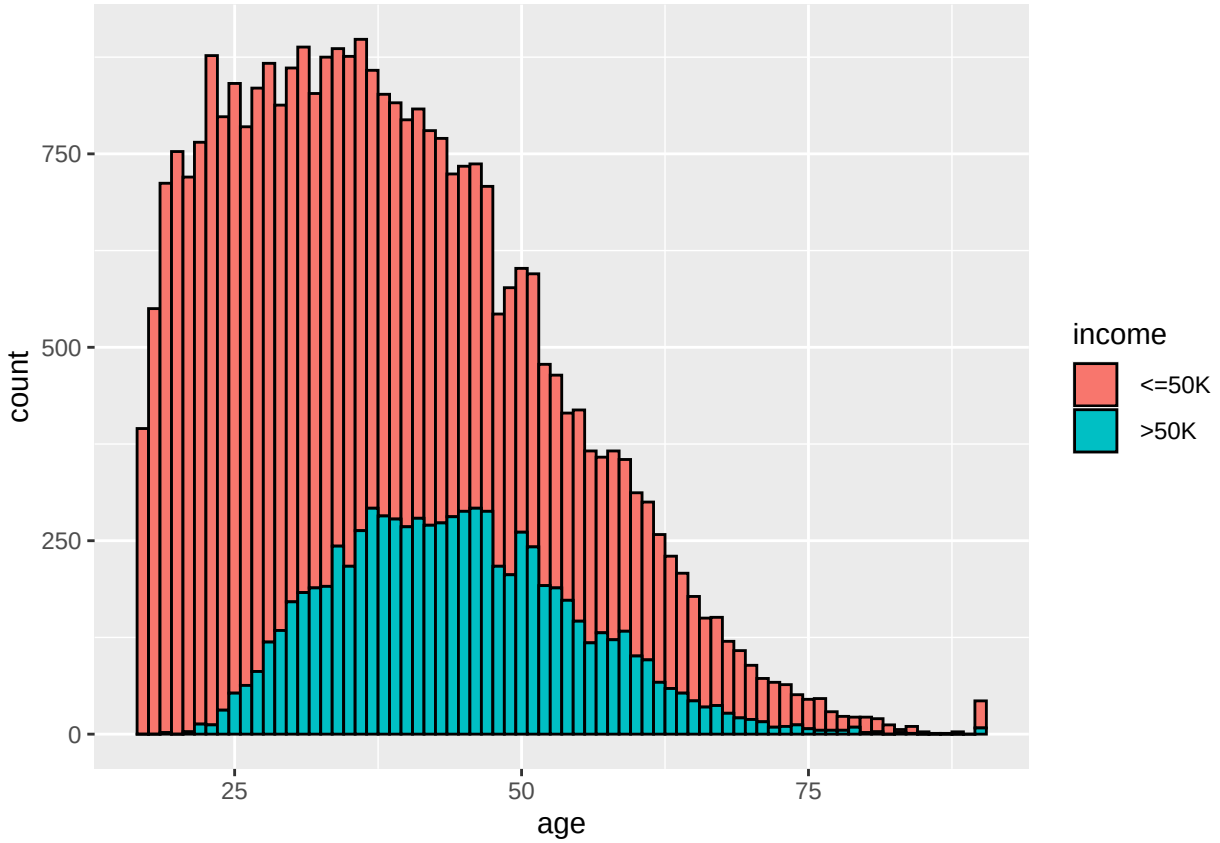

```
#Viz 2
ggplot(aes(x=income, y=education.num), data = incomeclassi) + geom_boxplot() +
     ggtitle('Years of Education vs. Income Level')
```




```
#Viz 3
ggplot(aes(x=income, y=hours.per.week), data = incomeclassi) + geom_boxplot() +
     ggtitle('Hours Per Week vs. Income Level')
```



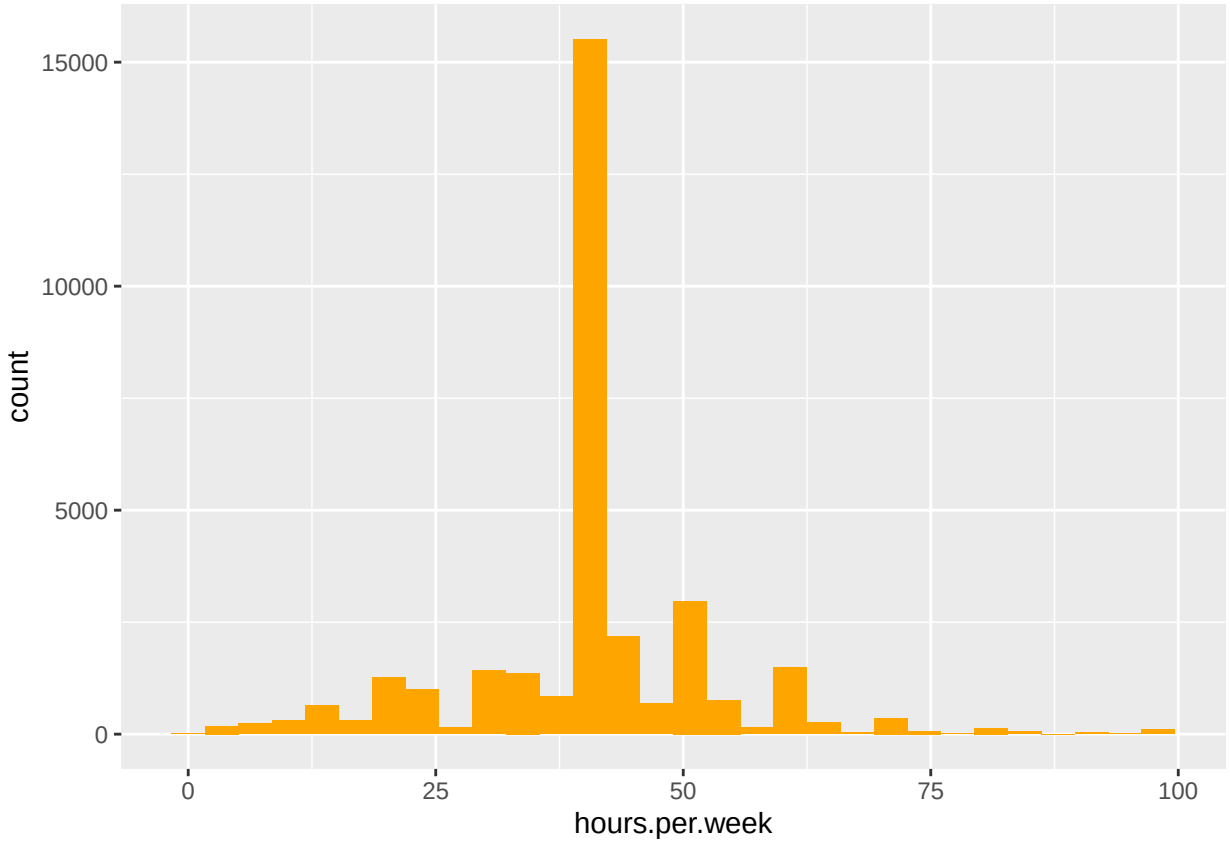

```
#viz 4
ggplot(incomeclassi, aes(age)) + geom_histogram(aes(fill = income), color = "black",
                                                binwidth = 1)
```




```
#Here the coloring is indicative of percentage. From this plot we can see that the percentage of people
```

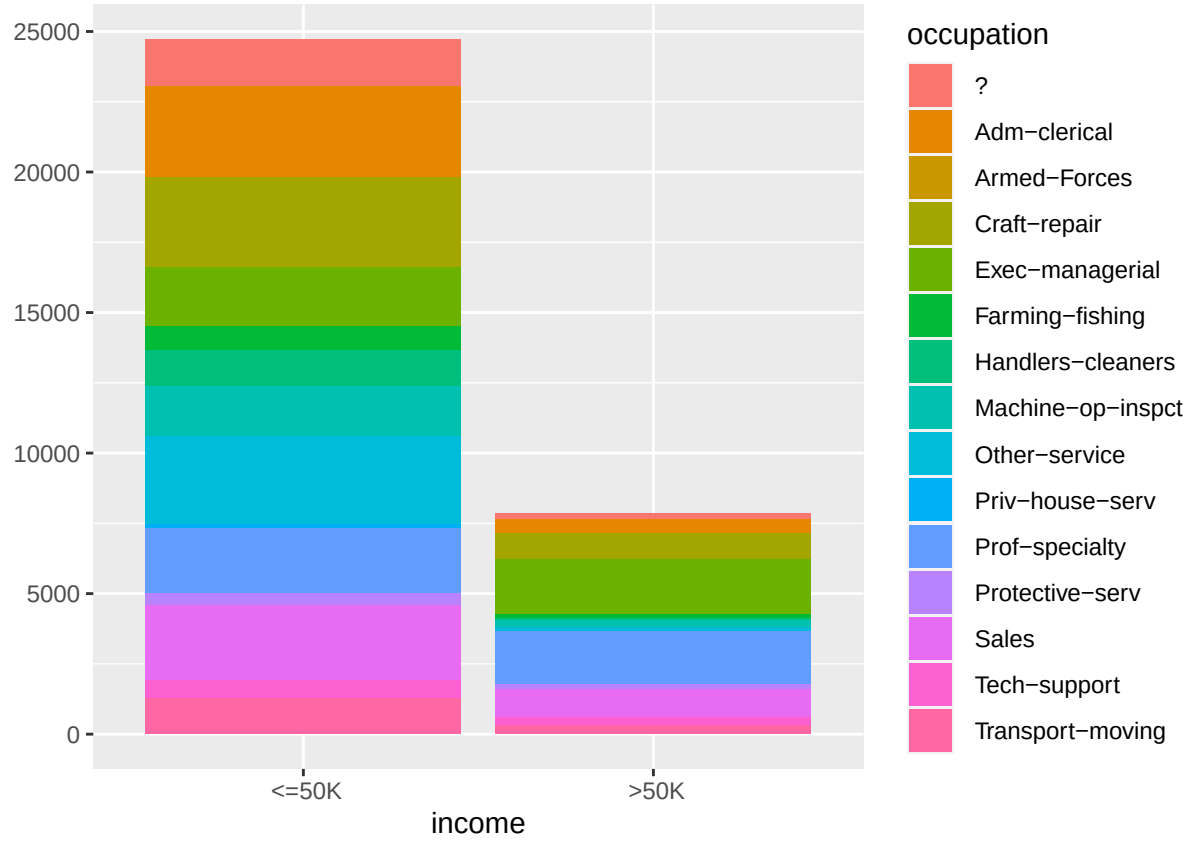
```
#viz 5

ggplot(incomeclassi, aes(hours.per.week)) + geom_histogram(fill = 'orange')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#It is clear that the highest frequency of hours.per.week occurs at 40.
```

```
#Viz 6

#qplot for different occupations
qplot (income, data = incomeclassi, fill = occupation)
```




```
#creating the required variabled
#creating the required variables
by_workclass <- incomeclassi %>% group_by(occupation, income) %>%
  summarise(n=n())
```

```
## `summarise()` has grouped output by 'occupation'. You can override using the `.groups` argument.
```

```
by_education <- incomeclassi %>% group_by(education, income) %>%
  summarise(n=n())
```

```
## `summarise()` has grouped output by 'education'. You can override using the `.groups` argument.
```

```
by_education$education <- ordered(by_education$education,
levels = c(' Preschool', ' 1st-4th', ' 5th-6th',' 7th-8th', ' 9th',
' 10th', ' 11th',' 12th', ' HS-grad', ' Prof-school',' Assoc-acdm',
' Assoc-voc', ' Some-college', ' Bachelors', ' Masters', 'Doctorate'))

by_education <- by_education %>% drop_na()

by_marital <- incomeclassi %>% group_by(marital.status, income) %>%
  summarise(n=n())
```

```
## `summarise()` has grouped output by 'marital.status'. You can override using the `.groups` argument.
```

```
by_occupation <- incomeclassi %>% group_by(occupation, income) %>% summarise(n=n())
```

```
## `summarise()` has grouped output by 'occupation'. You can override using the `.groups` argument.
```

```
by_relationship <- incomeclassi %>% group_by(relationship, income) %>% summarise(n=n())
```

```
## `summarise()` has grouped output by 'relationship'. You can override using the `.groups` argument.
```

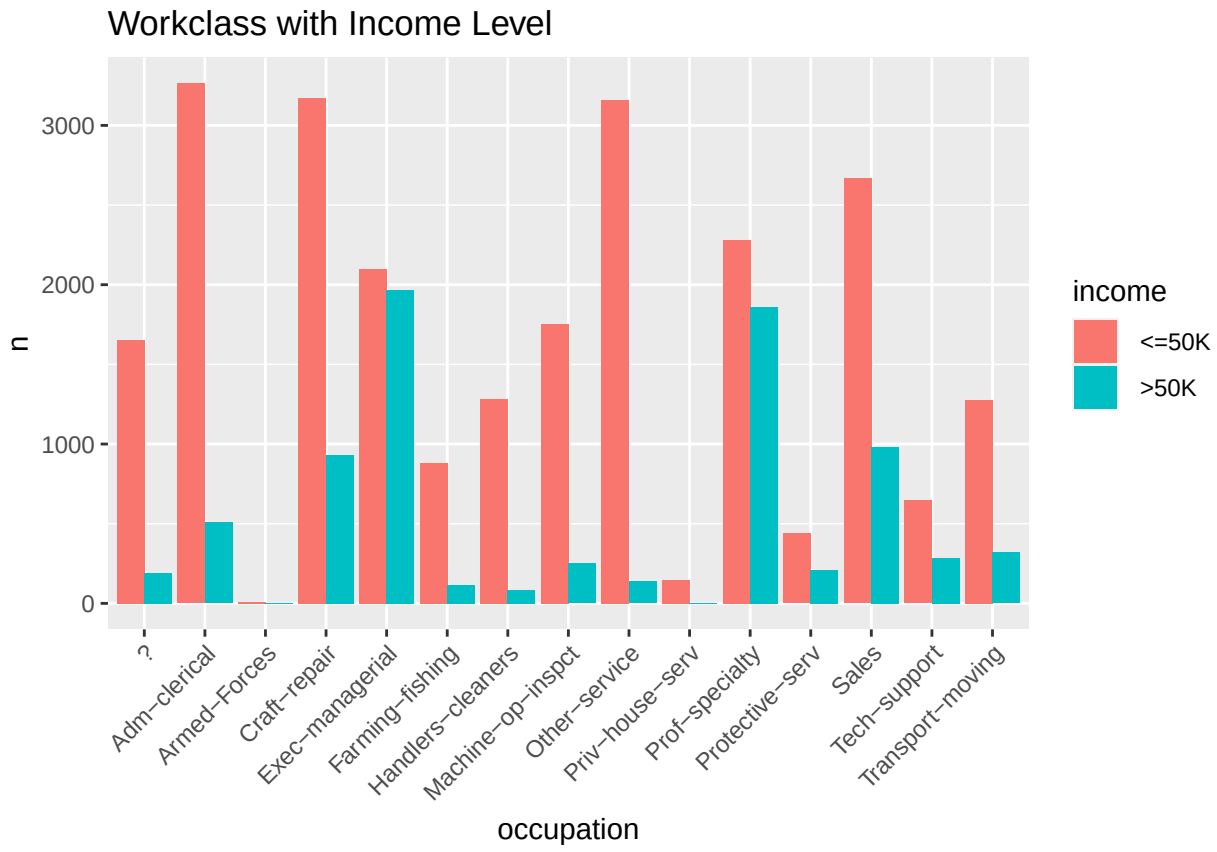
```
by_race <- incomeclassi %>% group_by(race, income) %>% summarise(n=n())
```

```
## `summarise()` has grouped output by 'race'. You can override using the `.groups` argument.
```

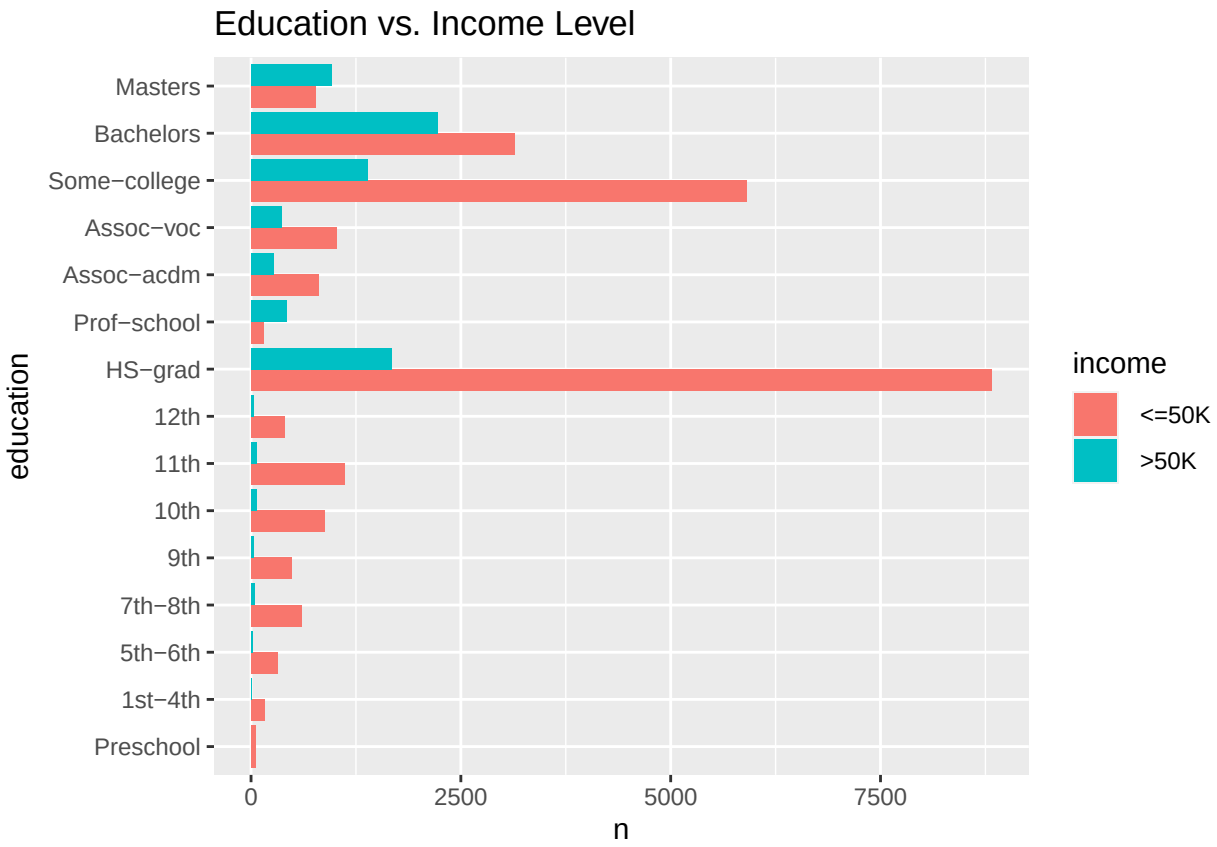
```
by_sex <- incomeclassi %>% group_by(sex, income) %>% summarise(n=n())
```

```
## `summarise()` has grouped output by 'sex'. You can override using the `.groups` argument.
```

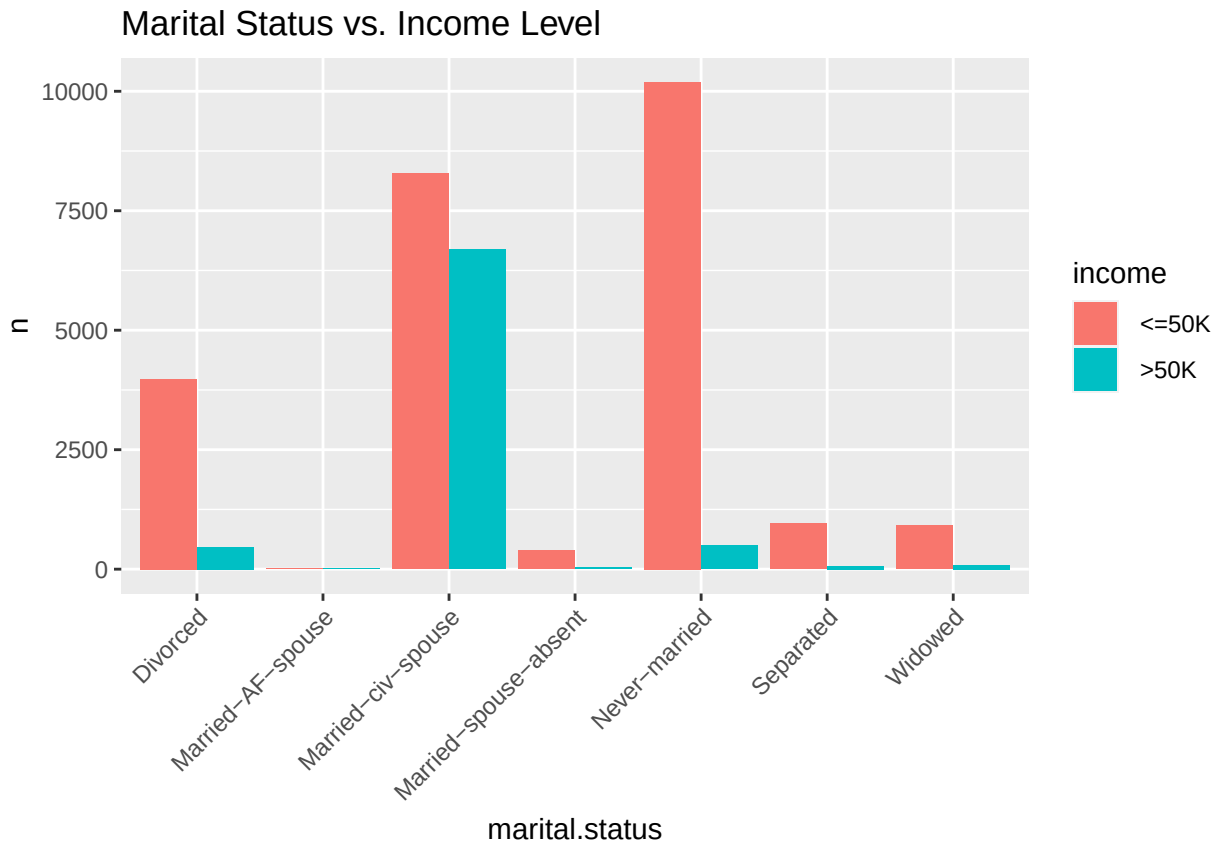
```
#viz 7

ggplot(aes(x=occupation, y=n, fill=income), data=by_workclass) + geom_bar(stat = 'identity', position =
```



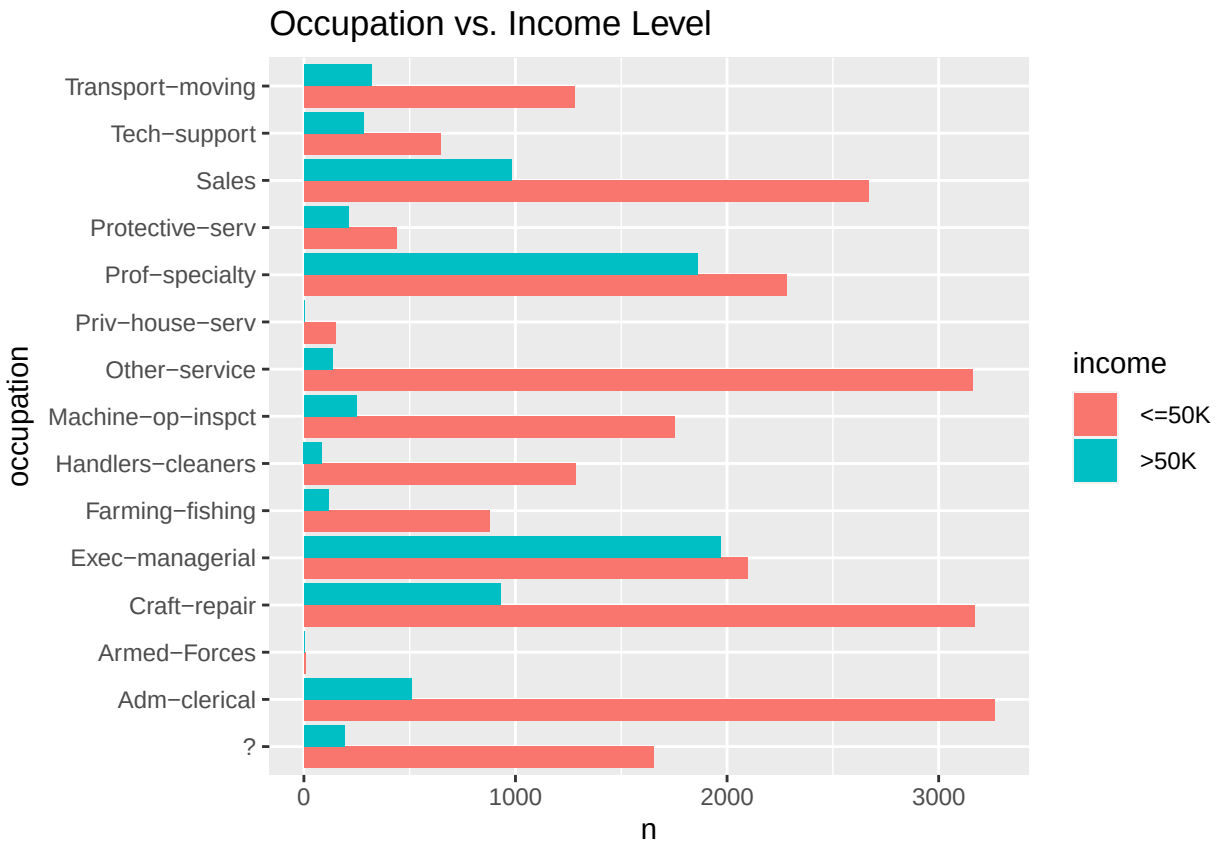
```
#viz 8

ggplot(aes(x=education, y=n, fill=income), data=by_education) + geom_bar(stat = 'identity', position =
```



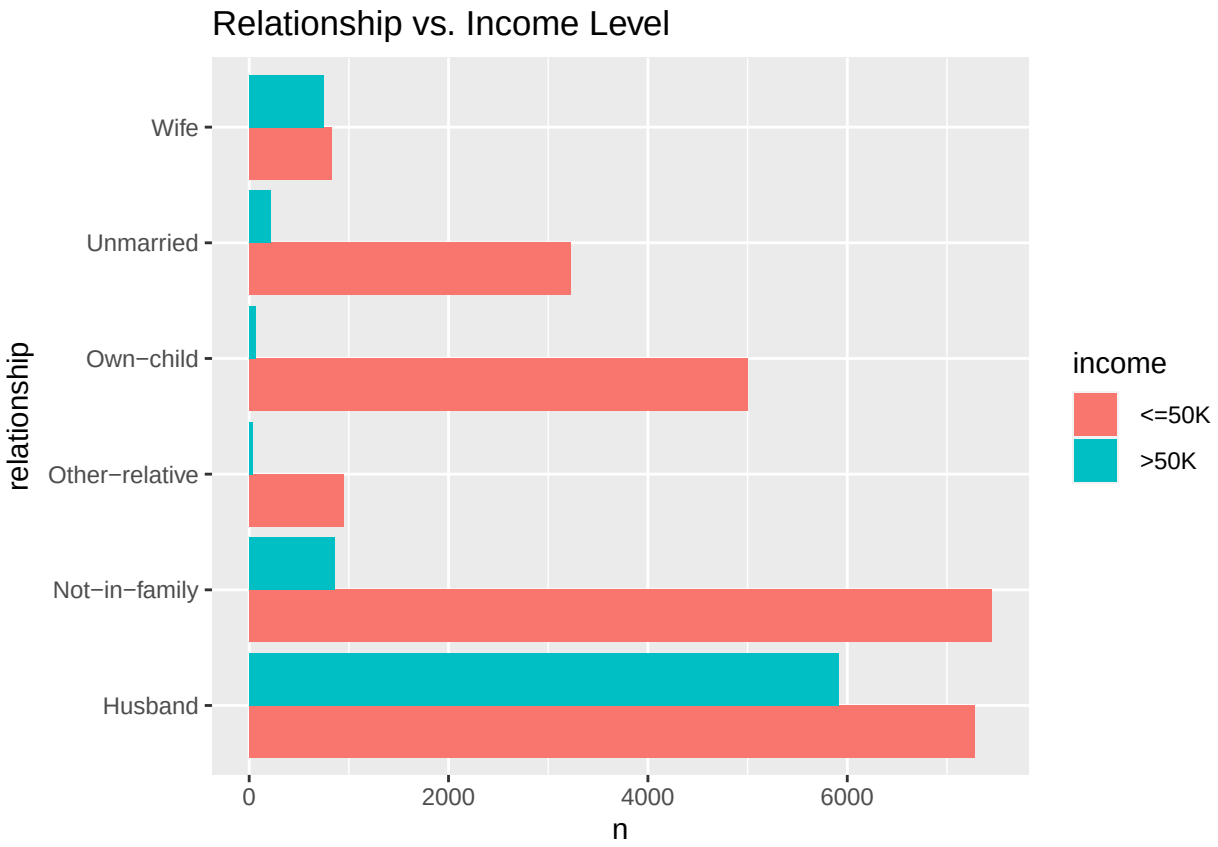

```
#viz 9

ggplot(aes(x=marital.status, y=n, fill=income), data=by_marital) + geom_bar(stat = 'identity', position=
```



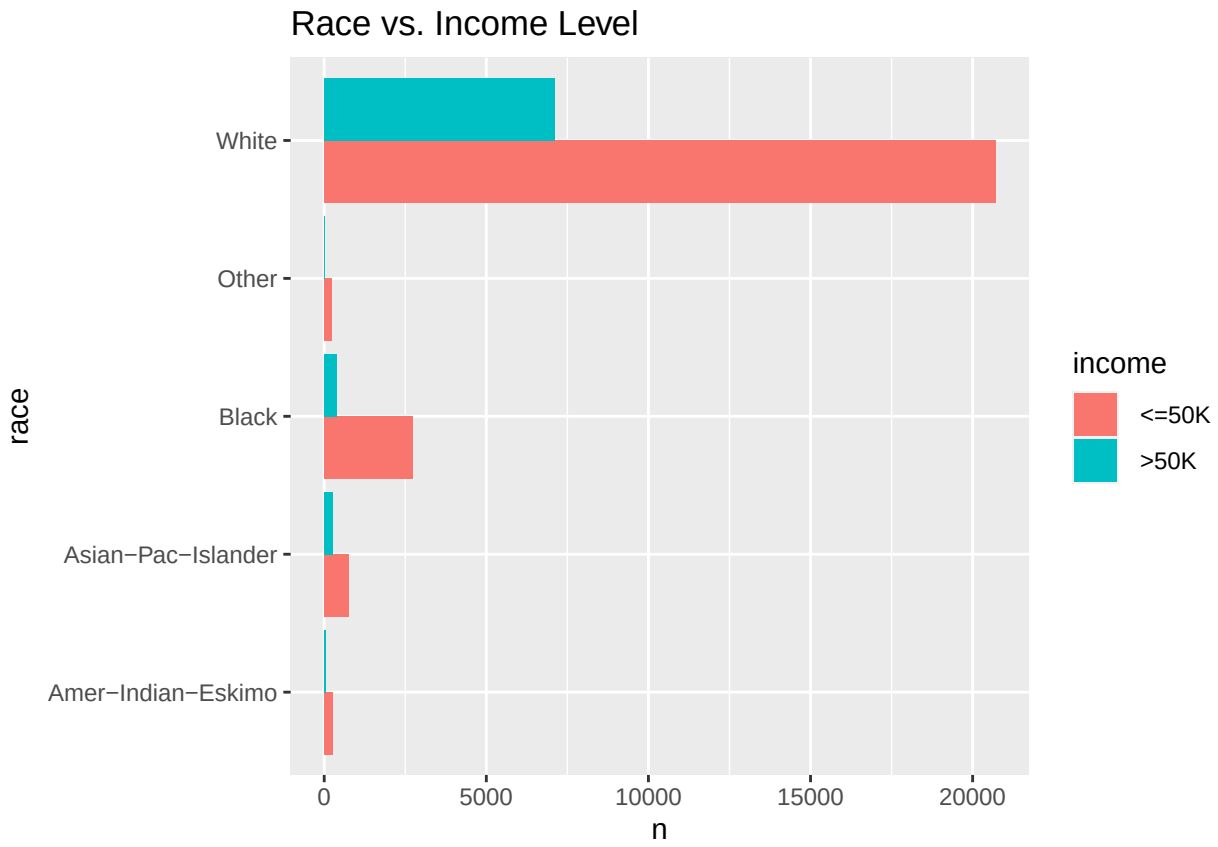

```
#viz 10

ggplot(aes(x=occupation, y=n, fill=income), data=by_occupation) + geom_bar(stat = 'identity', position=
```



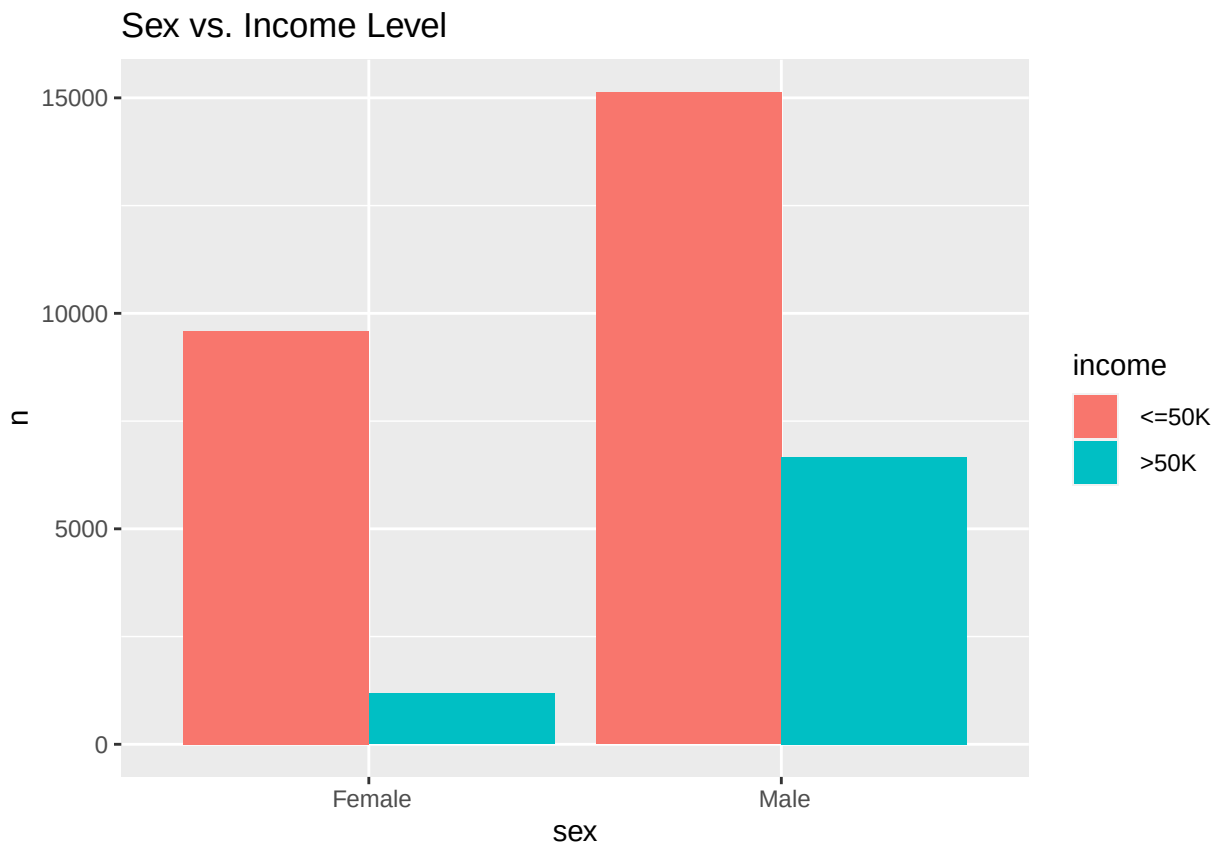

```
#viz 11

ggplot(aes(x=relationship, y=n, fill=income), data=by_relationship) + geom_bar(stat = 'identity', positi
```



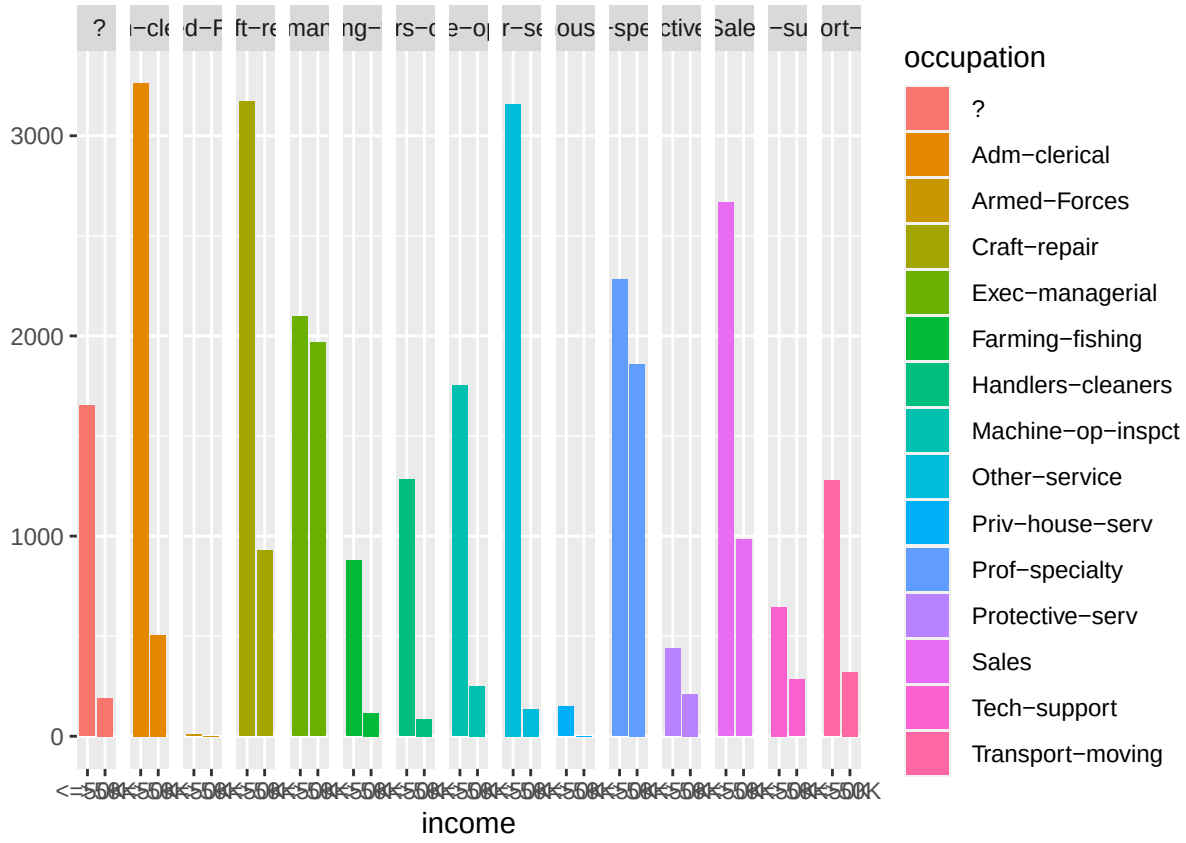

```
#viz 12

ggplot(aes(x=race, y=n, fill=income), data=by_race) + geom_bar(stat = 'identity', position = position_d
```



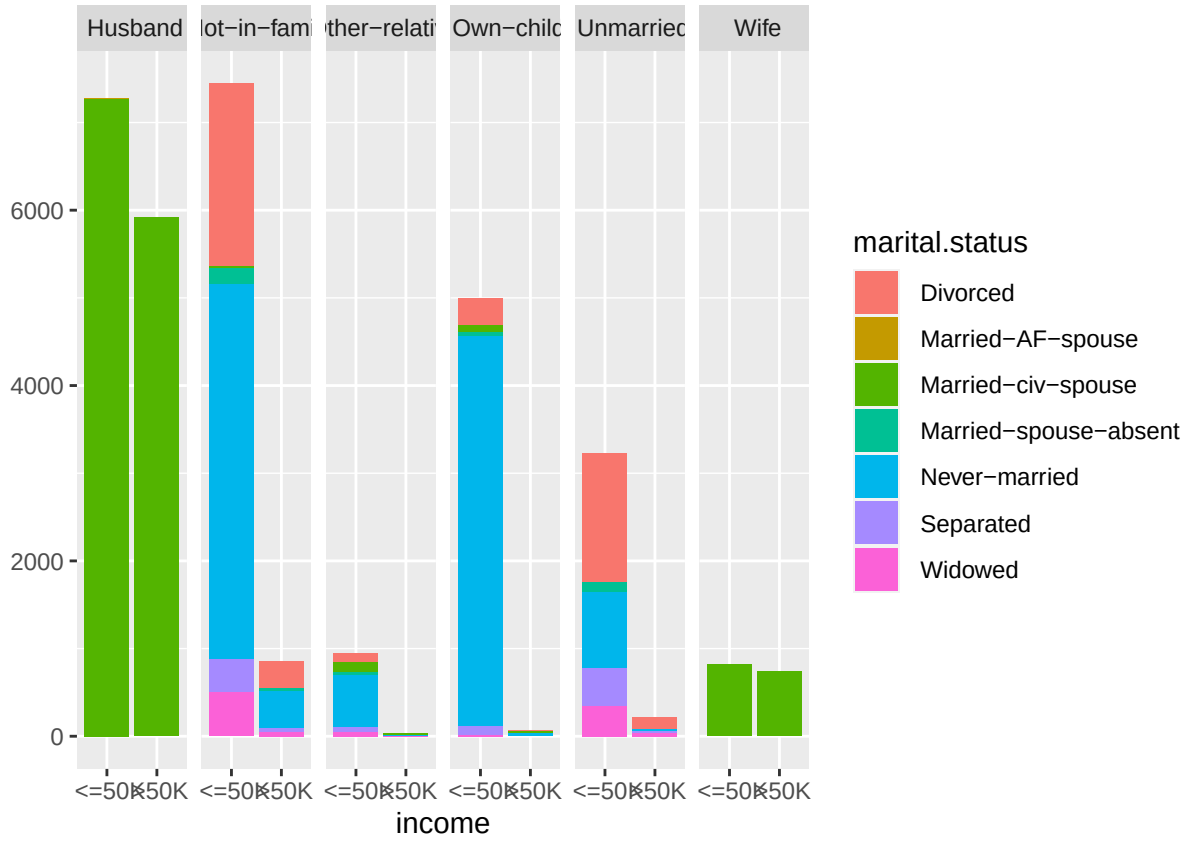

```
#viz 13

ggplot(aes(x=sex, y=n, fill=income), data=by_sex) + geom_bar(stat = 'identity', position = position_dod
```



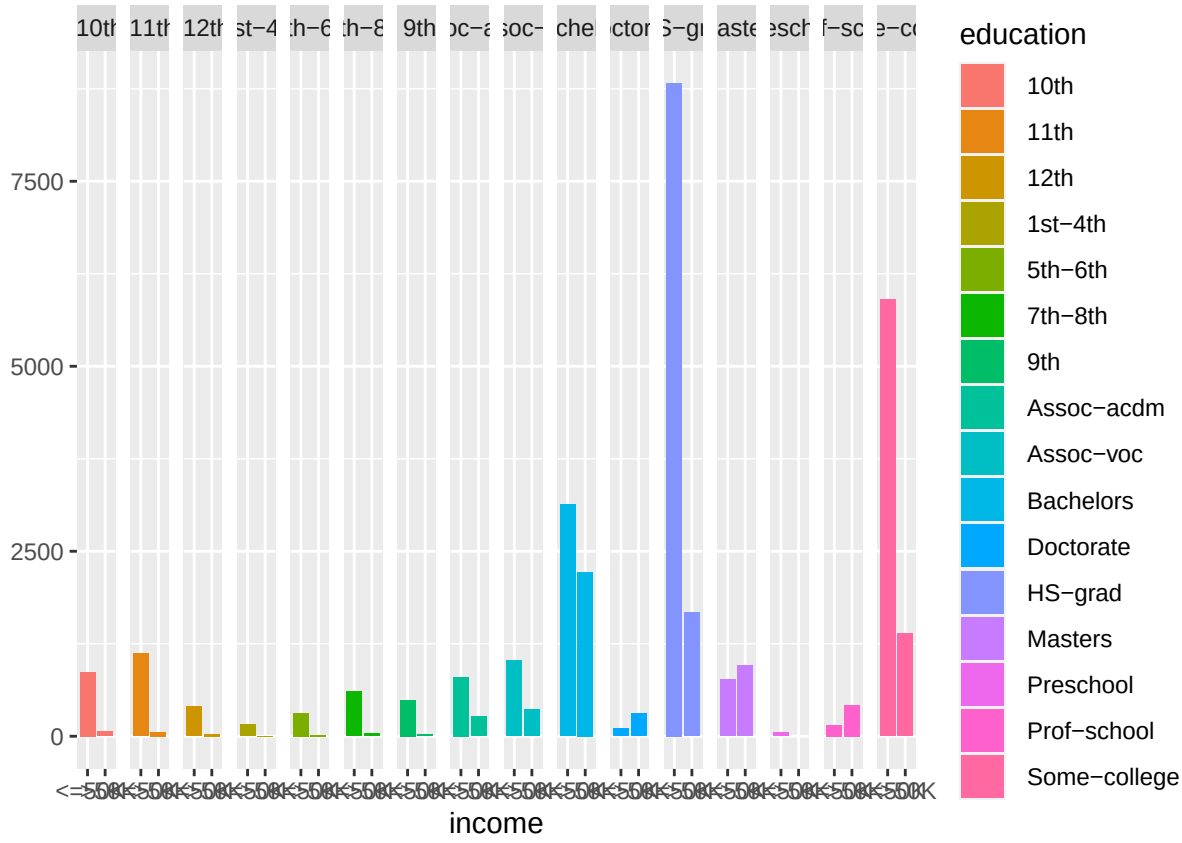

```
#creating different qplots

qplot (income, data = incomeclassi, fill = occupation) + facet_grid (. ~ occupation)
```




```
qplot (income, data = incomeclassi, fill = marital.status) + facet_grid (. ~ relationship)
```



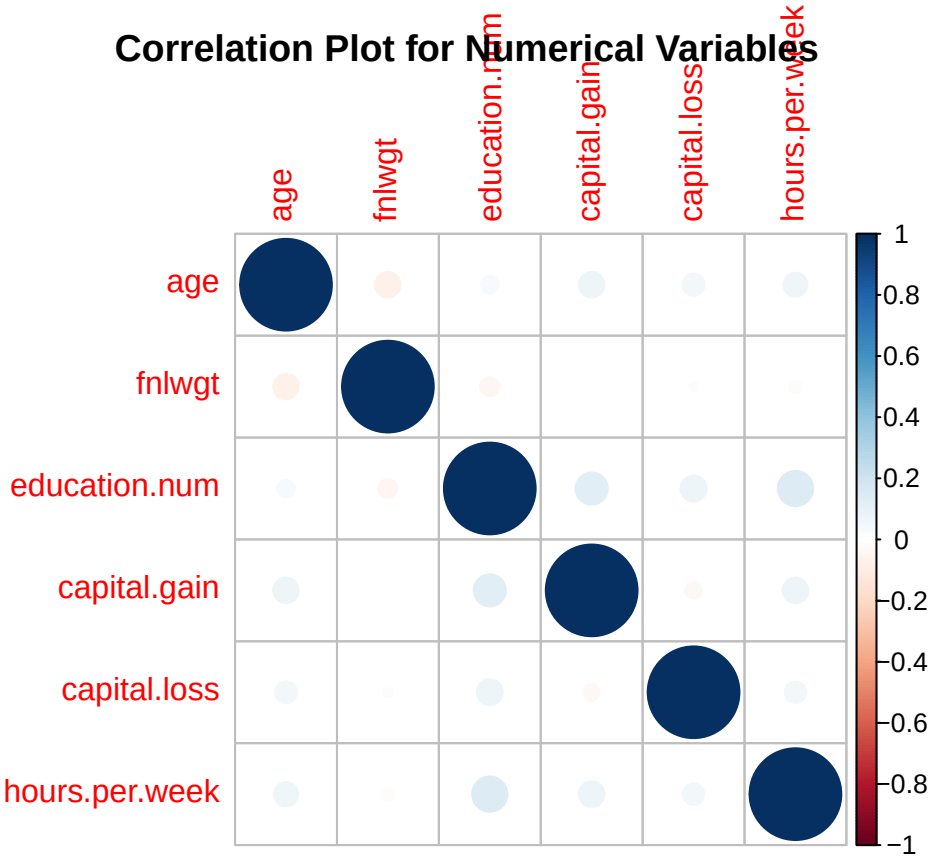

```
qplot (income, data = incomeclassi, fill = education) +
  facet_grid (. ~ education)
```




```
#Trying to Normalize the data and then do correlation plots

normalize <- function(x) { (x - min(x)) / (max(x) - min(x))}

incomeclassi$age<-normalize(incomeclassi$age)

incomeclassi$education.num<-normalize(incomeclassi$education.num)

incomeclassi$hours.per.week<-normalize(incomeclassi$hours.per.week)
incomeclassi$fnlwgt<-normalize(incomeclassi$fnlwgt)

incomeclassi$capital.gain<-normalize(incomeclassi$capital.gain)

incomeclassi$capital.loss<-normalize(incomeclassi$capital.loss)

#loading the correlation plot library
library(corrplot)
```

```
## corrplot 0.92 loaded
```

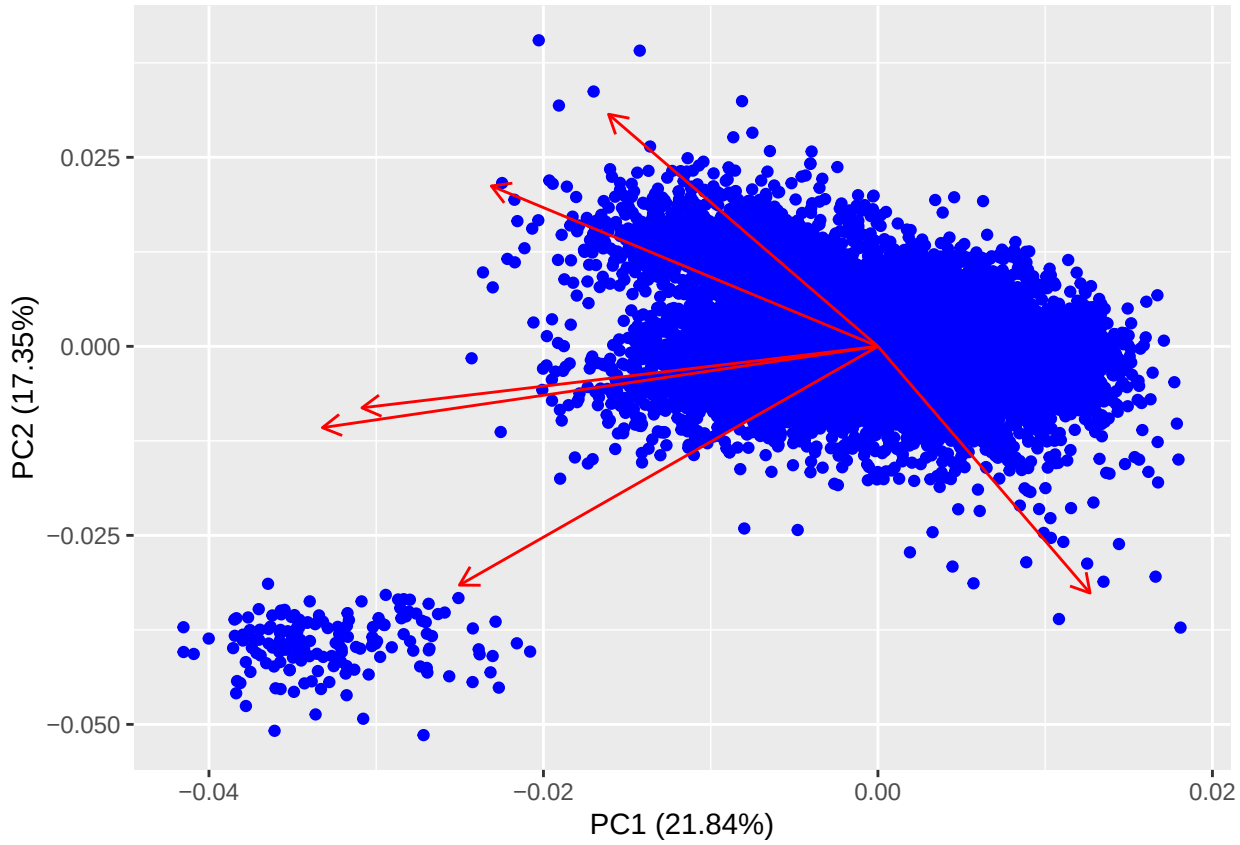
```
# Plotting the Correlation plot

par(mar=c(5.1,4.1,4.1,2.1))  ## Restore plot margins
 numeric.var <- sapply(incomeclassi, is.numeric) ## Find numerical variables
 corr.matrix <- cor(incomeclassi[,numeric.var])
 corrplot(corr.matrix, main="\n\nCorrelation Plot for Numerical Variables")
```




```
## Explore Numerical Variable
## Correlation between numerical variables
numeric.var <- sapply(incomeclassi, is.numeric)
## Calculate the correlation matrix
inc.corr <- cor(incomeclassi[,numeric.var])
inc.corr
```

```
##                      age        fnlwgt education.num  capital.gain
## age           1.00000000 -0.0766458679    0.03652719  0.0776744982
## fnlwgt       -0.07664587  1.0000000000   -0.04319463  0.0004318858
## education.num  0.03652719 -0.0431946327    1.00000000  0.1226301147
## capital.gain   0.07767450  0.0004318858    0.12263011  1.0000000000
## capital.loss   0.05777454 -0.0102517117    0.07992296 -0.0316150630
## hours.per.week 0.06875571 -0.0187684906    0.14812273  0.0784086154
##                capital.loss hours.per.week
## age              0.05777454     0.06875571
## fnlwgt          -0.01025171    -0.01876849
## education.num    0.07992296     0.14812273
## capital.gain    -0.03161506     0.07840862
## capital.loss     1.00000000     0.05425636
## hours.per.week   0.05425636     1.00000000
```

```r
#Data Reduction and Transformation
# Performing PCA on the data
# Perform Scree Plot and Parallel Analysis


incomeclassi.pca<-prcomp(incomeclassi[,c(1,3,5,11,12,13)],center=TRUE, scale.=TRUE)

library(ggfortify)

autoplot(incomeclassi.pca,  colour = 'blue', loadings = TRUE)
```




```r
summary(incomeclassi.pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5    PC6
## Standard deviation     1.1448 1.0203 1.0093 0.9705 0.9415 0.8953
## Proportion of Variance 0.2184 0.1735 0.1698 0.1570 0.1477 0.1336
## Cumulative Proportion  0.2184 0.3919 0.5617 0.7187 0.8664 1.0000
```

```r
str(incomeclassi.pca)
```

```
## List of 5
##  $ sdev    : num [1:6] 1.145 1.02 1.009 0.97 0.942 ...
##  $ rotation: num [1:6, 1:6] -0.383 0.21 -0.551 -0.415 -0.267 ...
```

```
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:6] "age" "fnlwgt" "education.num" "capital.gain" ...
##   .. ..$ : chr [1:6] "PC1" "PC2" "PC3" "PC4" ...
##  $ center  : Named num [1:6] 0.2956 0.1205 0.6054 0.0108 0.02 ...
##   ..- attr(*, "names")= chr [1:6] "age" "fnlwgt" "education.num" "capital.gain" ...
##  $ scale   : Named num [1:6] 0.1869 0.0717 0.1715 0.0739 0.0925 ...
##   ..- attr(*, "names")= chr [1:6] "age" "fnlwgt" "education.num" "capital.gain" ...
##  $ x       : num [1:32561, 1:6] -0.8462 0.0972 0.4358 0.4805 0.1051 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:6] "PC1" "PC2" "PC3" "PC4" ...
##  - attr(*, "class")= chr "prcomp"
```