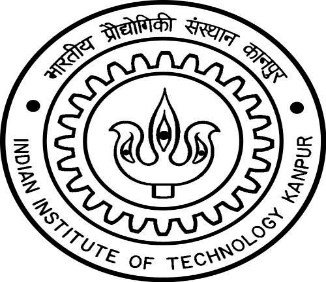
***PREDIcTIVE MODELLING foR adult census salary***



**2020-21**

***INDIAN INSTITUTE OF TECHNOLOGY, KANPUR***

**INSTRUCTOR:**

**DR. FAIZ HAMID**

**COMPLETED BY:**

**NIKHIL MUNAKHIYA  
201352**

**ACKNOWLEDGEMENT**

The success and final outcome of this project work required a lot of guidance and assistance from many people and we extremely fortunate to have got this all along the completion of our project work. Whatever we have done is only due to such guidance and assistance and I would like to thank all of them.

|  |  |
| --- | --- |
| \*\*\* | Introduction |
| 1) | About the Data |
| 2) | Data cleaning  2.1 Check Duplicate Data 2.2 Check Missing values 2.3 Transformed Data 2.4 Factorizing transformed data 2.5 Checking outliers 2.6 Checking Correlation of numeric data |
| 3) | Exploratory Data Analysis 3.1 Various plots |
| 4) | Model Fitting 4.1 Logistic Regression 4.2 Decision Tress |
| 5) | Conclusion |

I respect and thank **Dr.FAIZ HAMID, Department of Industrial and Management Engineering**.**, IIT Kanpur** for giving us an opportunity to do this project work and providing us all support and guidance which made us complete the project on time, I am extremely grateful to her for providing such a great support and guidance.

**THANK YOU**

Salary Prediction

**1)About the Data**

The dataset named Adult Census Income is available in kaggle and UCI repository. This data was extracted from the [1994 Census bureau database](http://www.census.gov/en.html) by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics).

**The prediction task is to determine whether a person makes over $50K a year or not.**

Target Variable - >50K, <=50K.  
  
1) age: continuous.  
2) workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.  
  
3) 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.  
  
4)marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.  
  
5)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.  
  
6)relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.  
  
7)race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.  
  
8)sex: Female, Male.

9)capital-gain: continuous.  
  
10)capital-loss: continuous.  
  
11)hours-per-week: continuous.  
  
12)native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

library(readr)

## Warning: package 'readr' was built under R version 4.0.5

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.0.5

##   
## 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(caTools)#for splitting

## Warning: package 'caTools' was built under R version 4.0.3

library(CatEncoders) # label & one hot encoding

## Warning: package 'CatEncoders' was built under R version 4.0.5

##   
## Attaching package: 'CatEncoders'

## The following object is masked from 'package:base':  
##   
## transform

library(lmtest)

## Warning: package 'lmtest' was built under R version 4.0.5

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.0.5

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

library(e1071)# For svr model

## Warning: package 'e1071' was built under R version 4.0.5

library(Metrics)

## Warning: package 'Metrics' was built under R version 4.0.5

library(caret)

## Warning: package 'caret' was built under R version 4.0.5

## Loading required package: lattice

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.0.5

##   
## Attaching package: 'caret'

## The following objects are masked from 'package:Metrics':  
##   
## precision, recall

library(car)

## Warning: package 'car' was built under R version 4.0.5

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.0.3

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

library(dummies)

## Warning: package 'dummies' was built under R version 4.0.3

## dummies-1.5.6 provided by Decision Patterns

library(superml)

## Warning: package 'superml' was built under R version 4.0.5

## Loading required package: R6

library(car)

dat <- read\_csv("census.csv")

##   
## -- Column specification --------------------------------------------------------  
## cols(  
## age = col\_double(),  
## workclass = col\_character(),  
## education = col\_character(),  
## maritalstatus = col\_character(),  
## occupation = col\_character(),  
## relationship = col\_character(),  
## race = col\_character(),  
## sex = col\_character(),  
## capitalgain = col\_double(),  
## capitalloss = col\_double(),  
## hoursperweek = col\_double(),  
## nativecountry = col\_character(),  
## over50k = col\_character()  
## )

head(dat,10)

## # A tibble: 10 x 13  
## age workclass education maritalstatus occupation relationship race sex   
## <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <chr>  
## 1 39 State-gov Bachelors Never-married Adm-cleri~ Not-in-fami~ White Male   
## 2 50 Self-emp-~ Bachelors Married-civ-s~ Exec-mana~ Husband White Male   
## 3 38 Private HS-grad Divorced Handlers-~ Not-in-fami~ White Male   
## 4 53 Private 11th Married-civ-s~ Handlers-~ Husband Black Male   
## 5 28 Private Bachelors Married-civ-s~ Prof-spec~ Wife Black Fema~  
## 6 37 Private Masters Married-civ-s~ Exec-mana~ Wife White Fema~  
## 7 49 Private 9th Married-spous~ Other-ser~ Not-in-fami~ Black Fema~  
## 8 52 Self-emp-~ HS-grad Married-civ-s~ Exec-mana~ Husband White Male   
## 9 31 Private Masters Never-married Prof-spec~ Not-in-fami~ White Fema~  
## 10 42 Private Bachelors Married-civ-s~ Exec-mana~ Husband White Male   
## # ... with 5 more variables: capitalgain <dbl>, capitalloss <dbl>,  
## # hoursperweek <dbl>, nativecountry <chr>, over50k <chr>

sum(duplicated(dat))

## [1] 3462

**2)DATA CLEANING**

# 2.1We can see that we have 3462 duplicate data

**We remove it**.

**2.2Checking missing values**

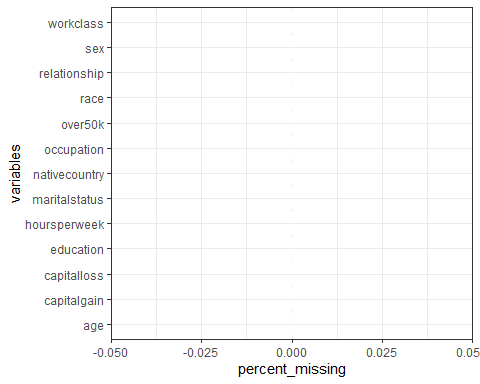
library(tidyr)

## Warning: package 'tidyr' was built under R version 4.0.5

missing\_data <- dat %>% summarise\_all(funs(sum(is.na(.))/n()))

## Warning: `funs()` was deprecated in dplyr 0.8.0.  
## Please use a list of either functions or lambdas:   
##   
## # Simple named list:   
## list(mean = mean, median = median)  
##   
## # Auto named with `tibble::lst()`:   
## tibble::lst(mean, median)  
##   
## # Using lambdas  
## list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))

missing\_data <- gather(missing\_data, key = "variables", value = "percent\_missing")  
ggplot(missing\_data, aes(x = reorder(variables, percent\_missing), y = percent\_missing)) +  
 geom\_bar(stat = "identity", fill = "#E7B800", aes(color = I('white')), size = 0.3)+  
 xlab('variables')+  
 coord\_flip()+   
 theme\_bw()



**# No missing values present in the data.**

table(dat$workclass)

##   
## ? Federal-gov Local-gov Never-worked   
## 1809 943 2067 7   
## Private Self-emp-inc Self-emp-not-inc State-gov   
## 22286 1074 2499 1279   
## Without-pay   
## 14

#Combined ‘Never-worked’ and ‘Without-pay’ into Unemployed as they meant the same

unemployed <- function(job){  
 job <- as.character(job)  
 if (job=='Never-worked' | job=='Without-pay'){  
 return('Unemployed')  
 }else{  
 return(job)  
 }  
}  
dat$workclass <- sapply(dat$workclass,unemployed)  
table(dat$workclass)

##   
## ? Federal-gov Local-gov Private   
## 1809 943 2067 22286   
## Self-emp-inc Self-emp-not-inc State-gov Unemployed   
## 1074 2499 1279 21

**#Combine State and Local gov jobs into a category called SL-gov and combine self-employed jobs into a category called self-emp**

group\_emp <- function(job){  
 if (job=='Local-gov' | job=='State-gov'){  
 return('SL-gov')  
 }else if (job=='Self-emp-inc' | job=='Self-emp-not-inc'){  
 return('self-emp')  
 }else{  
 return(job)  
 }  
}  
dat$workclass <- sapply(dat$workclass,group\_emp)  
table(dat$workclass)

##   
## ? Federal-gov Private self-emp SL-gov Unemployed   
## 1809 943 22286 3573 3346 21

table(dat$maritalstatus)

##   
## Divorced Married-AF-spouse Married-civ-spouse   
## 4394 23 14692   
## Married-spouse-absent Never-married Separated   
## 397 10488 1005   
## Widowed   
## 979

**#No need of transforming martial status**

Asia <- c('China','Hong','India','Iran','Cambodia','Japan', 'Laos' ,  
 'Philippines' ,'Vietnam' ,'Taiwan', 'Thailand')  
  
North.America <- c('Canada','United-States','Puerto-Rico' )  
  
Europe <- c('England' ,'France', 'Germany' ,'Greece','Holand-Netherlands','Hungary',  
 'Ireland','Italy','Poland','Portugal','Scotland','Yugoslavia')  
  
Latin.and.South.America <- c('Columbia','Cuba','Dominican-Republic','Ecuador',  
 'El-Salvador','Guatemala','Haiti','Honduras',  
 'Mexico','Nicaragua','Outlying-US(Guam-USVI-etc)','Peru',  
 'Jamaica','Trinadad&Tobago')  
Other <- c('South')  
  
group\_country <- function(ctry){  
 if (ctry %in% Asia){  
 return('Asia')  
 }else if (ctry %in% North.America){  
 return('North.America')  
 }else if (ctry %in% Europe){  
 return('Europe')  
 }else if (ctry %in% Latin.and.South.America){  
 return('Latin.and.South.America')  
 }else{  
 return('Other')   
 }  
}  
  
dat$nativecountry <- sapply(dat$nativecountry,group\_country)  
table(dat$nativecountry)

##   
## Asia Europe Latin.and.South.America   
## 671 521 1301   
## North.America Other   
## 29405 80

head(dat,10)

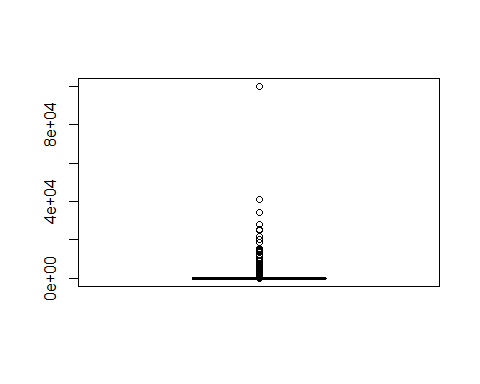
## # A tibble: 10 x 13  
## age workclass education maritalstatus occupation relationship race sex   
## <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <chr>  
## 1 39 SL-gov Bachelors Never-married Adm-cleric~ Not-in-fami~ White Male   
## 2 50 self-emp Bachelors Married-civ-s~ Exec-manag~ Husband White Male   
## 3 38 Private HS-grad Divorced Handlers-c~ Not-in-fami~ White Male   
## 4 53 Private 11th Married-civ-s~ Handlers-c~ Husband Black Male   
## 5 28 Private Bachelors Married-civ-s~ Prof-speci~ Wife Black Fema~  
## 6 37 Private Masters Married-civ-s~ Exec-manag~ Wife White Fema~  
## 7 49 Private 9th Married-spous~ Other-serv~ Not-in-fami~ Black Fema~  
## 8 52 self-emp HS-grad Married-civ-s~ Exec-manag~ Husband White Male   
## 9 31 Private Masters Never-married Prof-speci~ Not-in-fami~ White Fema~  
## 10 42 Private Bachelors Married-civ-s~ Exec-manag~ Husband White Male   
## # ... with 5 more variables: capitalgain <dbl>, capitalloss <dbl>,  
## # hoursperweek <dbl>, nativecountry <chr>, over50k <chr>

**2.4 Factorize the newly created column**

dat$workclass <- sapply(dat$workclass,factor)  
dat$nativecountry <- sapply(dat$`nativecountry`,factor)  
dat$maritalstatus <- sapply(dat$`maritalstatus`,factor)  
dat$occupation <- sapply(dat$occupation,factor)  
dat$relationship <- sapply(dat$relationship,factor)  
dat$over50k <- sapply(dat$over50k,factor)  
dat$education<- sapply(dat$education,factor)  
dat$sex<- sapply(dat$sex,factor)  
dat$race<- sapply(dat$race,factor)

**2.5 CHECKING OUTLIERS**

boxplot(dat$capitalgain)



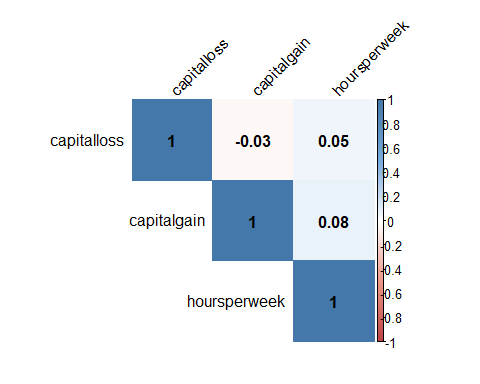
Q <- quantile(dat$capitalgain,probs = c(.25,.75),na.rm = TRUE)  
iqr <- IQR(dat$capitalgain)  
up <- Q[2]+1.5\*iqr # Upper Range   
low<- Q[1]-1.5\*iqr # Lower Range???  
eliminated<- subset(dat,dat$capitalgain > (Q[1] - 1.5\*iqr) & dat$capitalgain < (Q[2]+1.5\*iqr))   
eliminated

## # A tibble: 0 x 13  
## # ... with 13 variables: age <dbl>, workclass <fct>, education <fct>,  
## # maritalstatus <fct>, occupation <fct>, relationship <fct>, race <fct>,  
## # sex <fct>, capitalgain <dbl>, capitalloss <dbl>, hoursperweek <dbl>,  
## # nativecountry <fct>, over50k <fct>

**#No outliers present**

**2.6 Correlation for variables with numeric values**

#corrletaion  
M <- cor(dat[,c(9:11)])  
  
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))  
corrplot::corrplot(M, method="color", col=col(200),   
 type="upper", order="hclust",   
 addCoef.col = "black", # Add coefficient of correlation  
 tl.col="black", tl.srt=45, #Text label color and rotation  
)

 ##

**3) Exploratory Data Analysis And Chi-Square Test**

**1) Against Native Country**

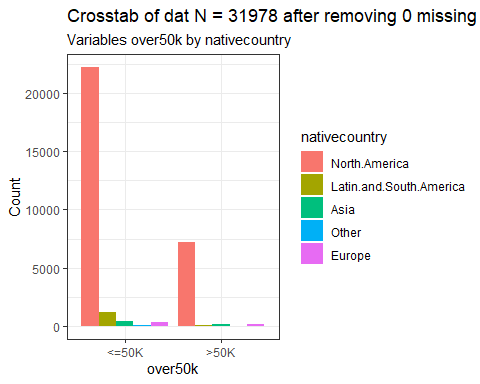
library(CGPfunctions)

## Warning: package 'CGPfunctions' was built under R version 4.0.5

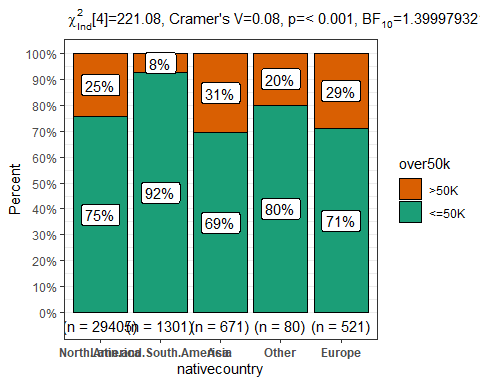
## Registered S3 methods overwritten by 'lme4':  
## method from  
## cooks.distance.influence.merMod car   
## influence.merMod car   
## dfbeta.influence.merMod car   
## dfbetas.influence.merMod car

# plotting and checking independence usingc chi.square test   
PlotXTabs(dat,nativecountry,over50k,plottype = "side")

## Plotted dataset dat variables nativecountry by over50k

 #we see in both the incomes most data of people are from north america

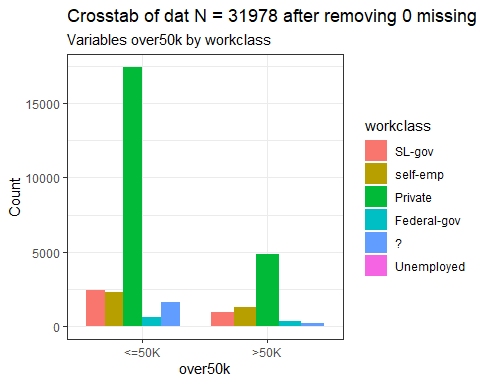
PlotXTabs2(dat,nativecountry,over50k)



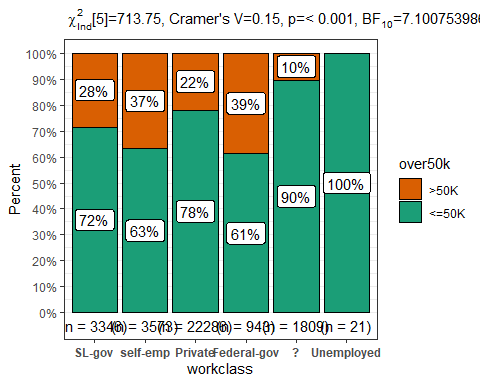
**2) Against Workclass**

PlotXTabs(dat,workclass,over50k,plottype = "side")

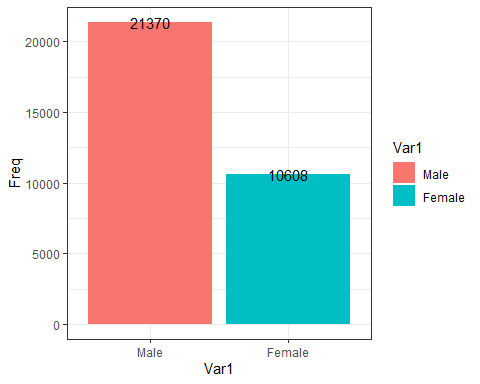
## Plotted dataset dat variables workclass by over50k

 #We can see that most of the people in both salary works in Private

PlotXTabs2(dat,workclass,over50k)



sex1 <- as.data.frame(table(dat$sex))  
library(ggplot2)  
#simple male female plot  
ggplot(data =sex1,aes(x=Var1,y=Freq,fill=Var1))+  
 geom\_bar(stat="identity")+geom\_text(label=sex1$Freq)



**# Majority of earning people are male ## Setting Hypothesis**

**#H0: Two variable are independant**

**#H1; Two var are dependent**

**#If we get Chi calculated less than Chi tabulated we reject the null hypothesis.**

df <- data.frame(table(dat$sex,dat$race,dat$over50k))  
  
table <-xtabs(~dat$race+dat$over50k,dat)   
as.data.frame(table)

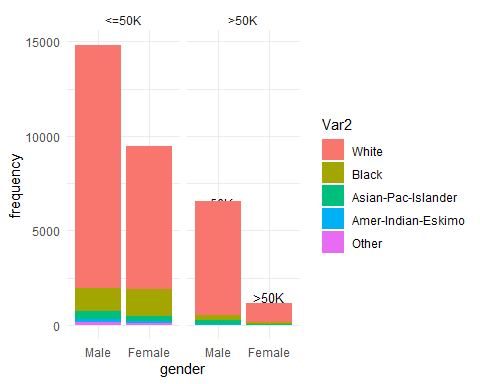
## dat.race dat.over50k Freq  
## 1 White <=50K 20420  
## 2 Black <=50K 2654  
## 3 Asian-Pac-Islander <=50K 703  
## 4 Amer-Indian-Eskimo <=50K 275  
## 5 Other <=50K 231  
## 6 White >50K 7010  
## 7 Black >50K 374  
## 8 Asian-Pac-Islander >50K 253  
## 9 Amer-Indian-Eskimo >50K 36  
## 10 Other >50K 22

chisq.test(table)

##   
## Pearson's Chi-squared test  
##   
## data: table  
## X-squared = 323, df = 4, p-value < 2.2e-16

**We reject null hypothesis**

ggplot(df, aes(x =Var1, y =Freq,fill=Var2))+geom\_bar(stat = "identity")+  
 geom\_bar(stat="identity", fill="steelblue")+  
 geom\_text(aes(label=Var3), vjust=-0.3, size=3.5)+  
 theme\_minimal()+geom\_bar(aes(fill =Var2), stat="identity")+facet\_wrap(~Var3)+  
 xlab("gender")+ylab("frequency")



**# The majority of people are white.**

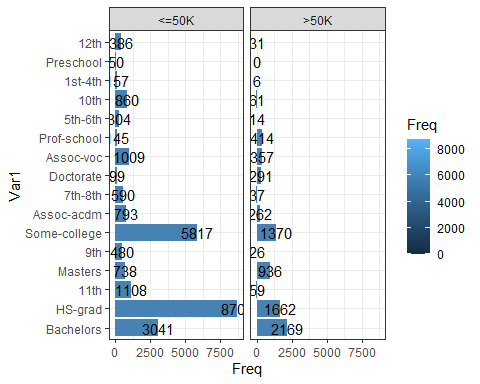
**##education and salary**

df2 <- data.frame((table(dat$education,dat$over50k)))  
  
table2 <- xtabs(~dat$education+dat$over50k,dat)  
chisq.test(table2)

##   
## Pearson's Chi-squared test  
##   
## data: table2  
## X-squared = 4367.3, df = 15, p-value < 2.2e-16

**#again reject null hypothesis.**

ggplot(df2, aes( y=Var1, x =Freq,fill=Freq))+geom\_bar(stat = "identity")+  
 geom\_bar(stat="identity", fill="steelblue")+  
 facet\_wrap(~Var2)+geom\_text(aes(label=Freq))



**#Hs graduates are the highest earners that earns less than 50K #In more than 50K Bachelors earns most**

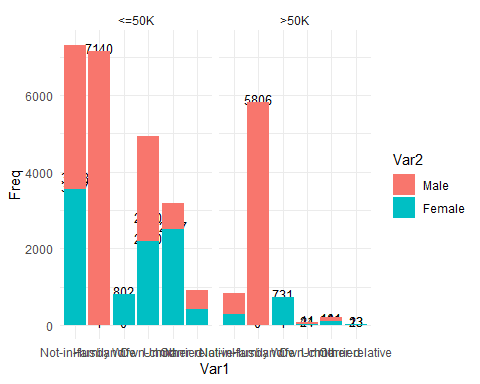
**##Relationship and salary**

df3 <- data.frame(table(dat$relationship,dat$sex,dat$over50k))  
  
table3 <- xtabs(~dat$relationship+dat$over50k,dat)  
chisq.test(table3)

##   
## Pearson's Chi-squared test  
##   
## data: table3  
## X-squared = 6588.3, df = 5, p-value < 2.2e-16

#again reject null hypothesis

ggplot(df3, aes( x=Var1, y=Freq,fill=Var2))+geom\_bar(stat = "identity")+  
 geom\_bar(stat="identity", fill="steelblue")+  
 geom\_text(aes(label=Freq),vjust=.18, size=3.5)+  
 theme\_minimal()+geom\_bar(aes(fill =Var2), stat="identity")+facet\_wrap(~Var3)



**#Mostly Husband are the one earning. there is female husband that earns less than 50K**

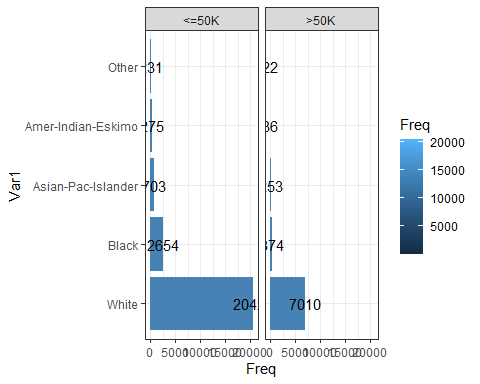
**#race& salary**

df5 <- data.frame(table(dat$race,dat$over50k))  
table5 <- xtabs(~dat$race+dat$over50k,dat)  
chisq.test(table5)

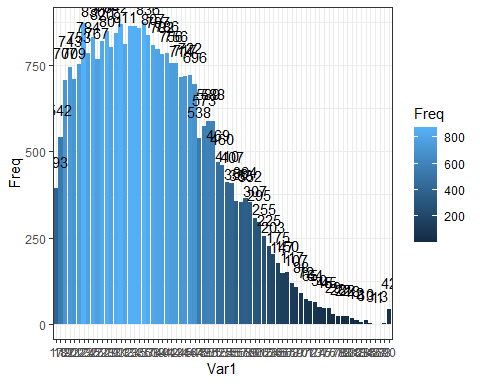
##   
## Pearson's Chi-squared test  
##   
## data: table5  
## X-squared = 323, df = 4, p-value < 2.2e-16

#Again Reject

ggplot(df5, aes( y=Var1, x =Freq,fill=Freq))+geom\_bar(stat = "identity")+  
 geom\_bar(stat="identity", fill="steelblue")+  
 facet\_wrap(~Var2)+geom\_text(aes(label=Freq))

 #Mostly people earning are white ##Age Plot

df6 <- data.frame(table(dat$age))  
ggplot(df6,aes(x=Var1,y=Freq,fill=Freq))+geom\_bar(stat="identity")+geom\_text(aes(label=Freq),vjust=-1.8)



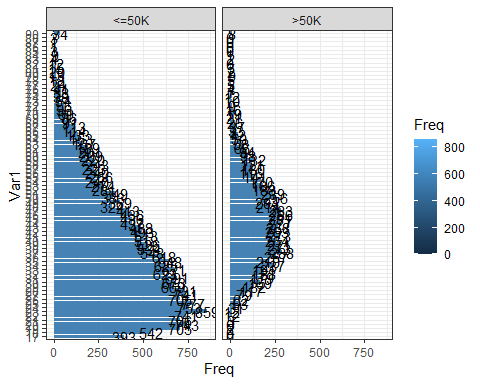
df7 <- data.frame(table(dat$age,dat$over50k))  
table6 <- xtabs(~dat$age+dat$over50k,dat)  
chisq.test(table6)

## Warning in chisq.test(table6): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: table6  
## X-squared = 3445.5, df = 72, p-value < 2.2e-16

Again reject

ggplot(df7, aes( y=Var1, x =Freq,fill=Freq))+geom\_bar(stat = "identity")+  
 geom\_bar(stat="identity", fill="steelblue")+  
 facet\_wrap(~Var2)+geom\_text(aes(label=Freq))



**##ENDING WITH THE EDA PART**

Now moving ahead

**4) Model Fitting**

## splitting data into train and test

split <- sample.split(dat$over50k,.90)  
train <- subset(dat,split==TRUE)  
test <- subset(dat,split==FALSE)  
dim(train)

## [1] 28781 13

dim(test)

## [1] 3197 13

na\_cols <- colSums(is.na(train)) / nrow(train)  
na\_cols <- names(na\_cols[which(na\_cols > 0.9)])  
  
cat\_cols <- names(train)[sapply(train, is.character)]  
  
for(c in cat\_cols){  
 lbl <- LabelEncoder$new()  
 lbl$fit(c(train[[c]], test[[c]]))  
 train[[c]] <- lbl$transform(train[[c]])  
 test[[c]] <- lbl$transform(test[[c]])  
}  
dim(train)

## [1] 28781 13

## 4.1) Logistic regression

glm\_model = glm(over50k ~ ., family = binomial(logit), data = train)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

test$predicted.income = predict(glm\_model, newdata=test)  
set.seed(1)  
test$income\_class <- ifelse(test$predicted.income > 0.5, ">50K","<=50K")  
glm\_con <-confusionMatrix(as.factor( test$income\_class),as.factor( test$over50k))  
glm\_con

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction <=50K >50K  
## <=50K 2328 379  
## >50K 100 390  
##   
## Accuracy : 0.8502   
## 95% CI : (0.8373, 0.8624)  
## No Information Rate : 0.7595   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.5319   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9588   
## Specificity : 0.5072   
## Pos Pred Value : 0.8600   
## Neg Pred Value : 0.7959   
## Prevalence : 0.7595   
## Detection Rate : 0.7282   
## Detection Prevalence : 0.8467   
## Balanced Accuracy : 0.7330   
##   
## 'Positive' Class : <=50K   
##

# We get the Metrics

**Accuracy : 0.8445 Higher the better**

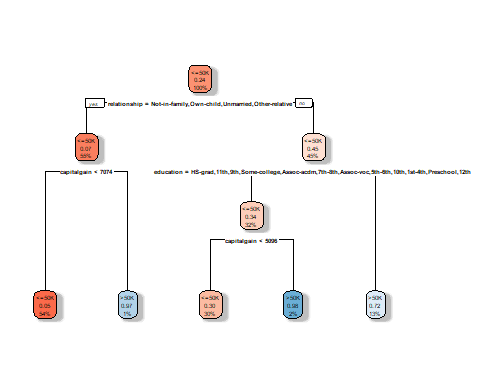
**Sensitivity : 0.9642 Higher the better  
Specificity : 0.4668 Lower the better**

**4.2) Decision Tree**

library(rpart)  
library(rpart.plot)

## Warning: package 'rpart.plot' was built under R version 4.0.5

tree\_model <- rpart(over50k ~ ., train, method = "class")  
all\_probs <- predict(tree\_model, test, type = "prob")  
rpart.plot(tree\_model, box.palette="RdBu", shadow.col="gray")



test$income\_class <- ifelse(all\_probs[,1]>0.5,"<=50K",">50K")  
dt\_con <- confusionMatrix(as.factor(test$income\_class),test$over50k)  
dt\_con

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction <=50K >50K  
## <=50K 2307 344  
## >50K 121 425  
##   
## Accuracy : 0.8546   
## 95% CI : (0.8418, 0.8666)  
## No Information Rate : 0.7595   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.5581   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9502   
## Specificity : 0.5527   
## Pos Pred Value : 0.8702   
## Neg Pred Value : 0.7784   
## Prevalence : 0.7595   
## Detection Rate : 0.7216   
## Detection Prevalence : 0.8292   
## Balanced Accuracy : 0.7514   
##   
## 'Positive' Class : <=50K   
##

#**For Decision Tree we get following metrics**

**Accuracy : 0.842   
Sensitivity : 0.9498  
Specificity : 0.5020**

**5) Conclusion**

**1)We can see that Decision Tree have higher accuracy then Logistic regression.**

**2)We can see that Maritial statues, Working Hours and Sex really matters if you want to earn more than 50K per year. In contrary, the working class is not that important. Generally speaking, you will get equal opportunity if you work hard enough, no matter what kinds of job are you doing.**