Select Libraries

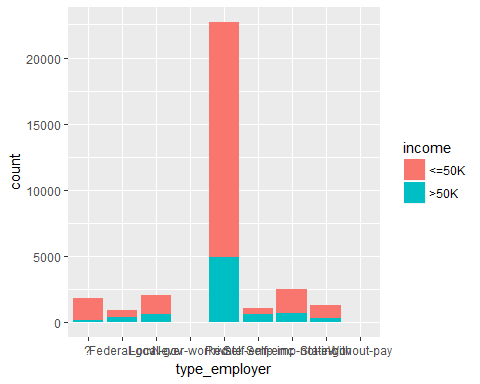
library(ggplot2)  
library(dplyr)

Load csv file

adult<-read.csv('adult\_sal.csv')  
  
  
adult<-select(adult,-X)  
  
table(adult$type\_employer)

##   
## ? Federal-gov Local-gov Never-worked   
## 1836 960 2093 7   
## Private Self-emp-inc Self-emp-not-inc State-gov   
## 22696 1116 2541 1298   
## Without-pay   
## 14

ggplot(adult,aes(type\_employer))+geom\_bar(aes(fill=income))

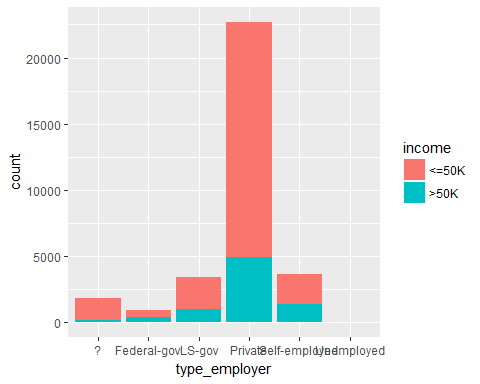


# Feature Engineering Combine employer.

employer<-function(job){  
   
 job<-as.character(job)  
 if(job=='Never-worked' | job=='Without-pay')  
 return('Unemployed')  
 else if(job=='Local-gov' | job=='State-gov')  
 return('LS-gov')  
 else if(job=='Self-emp-inc' | job=='Self-emp-not-inc')  
 return('Self-employed')  
 else  
 return(job)  
   
 }  
  
adult$type\_employer<-sapply(adult$type\_employer,employer)  
  
table(adult$type\_employer)

##   
## ? Federal-gov LS-gov Private Self-employed   
## 1836 960 3391 22696 3657   
## Unemployed   
## 21

ggplot(adult,aes(type\_employer))+geom\_bar(aes(fill=income))



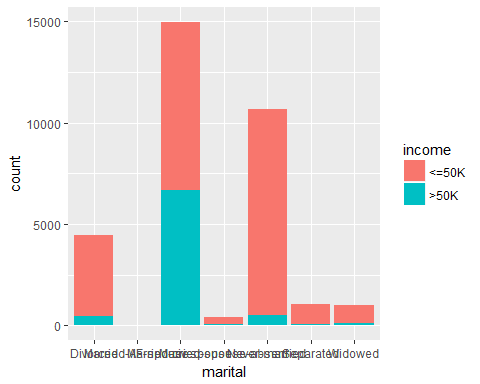
adult$type\_employer<-factor(adult$type\_employer)

# Feature Engineering Marital status

table(adult$marital)

##   
## Divorced Married-AF-spouse Married-civ-spouse   
## 4443 23 14976   
## Married-spouse-absent Never-married Separated   
## 418 10683 1025   
## Widowed   
## 993

ggplot(adult,aes(marital))+geom\_bar(aes(fill=income))



marital\_status<-function(status){  
 status<-as.character(status)  
 if(status=='Divorced' | status=='Separated' | status=='Widowed')  
 return('Not Married')  
 else if(status=='Never-married')  
 return(status)  
 else  
 return('Married')  
}  
  
adult$marital<-sapply(adult$marital,marital\_status)  
  
table(adult$marital)

##   
## Married Never-married Not Married   
## 15417 10683 6461

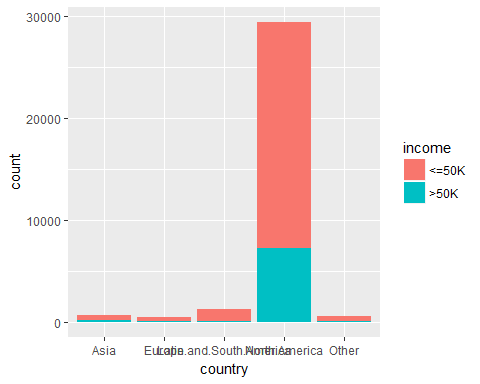
adult$marital<-factor(adult$marital)

Feature Engineering: Country

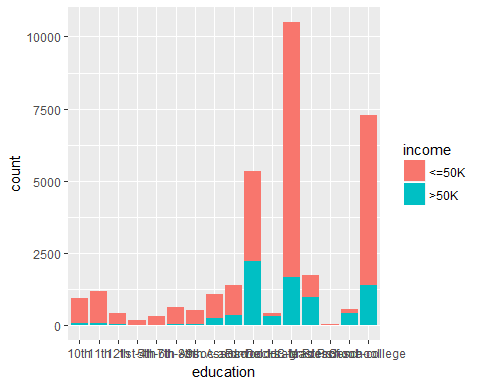
table(adult$country)

##   
## ? Cambodia   
## 583 19   
## Canada China   
## 121 75   
## Columbia Cuba   
## 59 95   
## Dominican-Republic Ecuador   
## 70 28   
## El-Salvador England   
## 106 90   
## France Germany   
## 29 137   
## Greece Guatemala   
## 29 64   
## Haiti Holand-Netherlands   
## 44 1   
## Honduras Hong   
## 13 20   
## Hungary India   
## 13 100   
## Iran Ireland   
## 43 24   
## Italy Jamaica   
## 73 81   
## Japan Laos   
## 62 18   
## Mexico Nicaragua   
## 643 34   
## Outlying-US(Guam-USVI-etc) Peru   
## 14 31   
## Philippines Poland   
## 198 60   
## Portugal Puerto-Rico   
## 37 114   
## Scotland South   
## 12 80   
## Taiwan Thailand   
## 51 18   
## Trinadad&Tobago United-States   
## 19 29170   
## Vietnam Yugoslavia   
## 67 16

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')   
 }  
}  
  
adult$country <- sapply(adult$country,group\_country)  
adult$country<-factor(adult$country)  
ggplot(adult,aes(country))+geom\_bar(aes(fill=income))

 ################################## Feature Engineering Education ####################################################

ggplot(adult,aes(education))+geom\_bar(aes(fill=income))



school<-c('10th','11th','12th','1st-4th','5th-6th','7th-8th','9th','Preschool')  
  
specialisation<- c('Bachelors','Doctorate','Masters','Prof-school')  
  
education<-function(edu){  
   
 if(edu %in% school)  
 return('school')  
 else if(edu %in% specialisation)  
 return('specialisation')  
 else  
 return('highschool')  
}  
  
adult$education<-sapply(adult$education,education)  
  
table(adult$education)

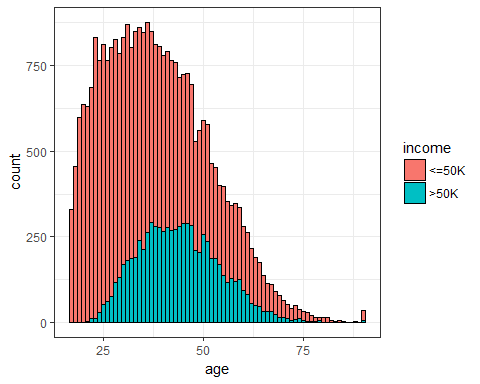
##   
## highschool school specialisation   
## 20241 4253 8067

Remove Missing Data

adult[adult=='?']<-NA  
  
adult<-na.omit(adult)

plots

ggplot(adult,aes(age))+geom\_histogram(aes(fill=income),color='black',binwidth = 1)+theme\_bw()

 #\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Logistic Regression \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

library(caTools)  
  
set.seed(101)  
  
sample<-sample.split(adult$income,SplitRatio = 0.7)  
  
train<-subset(adult,sample==T)  
test<-subset(adult,sample==F)  
  
model<-glm(income~.,family = binomial(link ="logit" ),data=train)

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

summary(model)

##   
## Call:  
## glm(formula = income ~ ., family = binomial(link = "logit"),   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.0462 -0.5186 -0.1964 -0.0080 3.7882   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -6.762e+00 4.614e-01 -14.656 < 2e-16 \*\*\*  
## age 2.589e-02 1.987e-03 13.035 < 2e-16 \*\*\*  
## type\_employerLS-gov -6.810e-01 1.262e-01 -5.397 6.78e-08 \*\*\*  
## type\_employerPrivate -4.422e-01 1.124e-01 -3.936 8.30e-05 \*\*\*  
## type\_employerSelf-employed -6.671e-01 1.240e-01 -5.378 7.52e-08 \*\*\*  
## type\_employerUnemployed -1.218e+01 1.355e+02 -0.090 0.928341   
## fnlwgt 5.304e-07 2.079e-07 2.551 0.010746 \*   
## educationschool -1.179e-01 1.347e-01 -0.876 0.381283   
## educationspecialisation 2.106e-01 1.008e-01 2.090 0.036578 \*   
## education\_num 2.384e-01 2.456e-02 9.706 < 2e-16 \*\*\*  
## maritalNever-married -1.240e+00 1.951e-01 -6.357 2.06e-10 \*\*\*  
## maritalNot Married -7.044e-01 1.954e-01 -3.605 0.000312 \*\*\*  
## occupationArmed-Forces -5.804e-01 1.823e+00 -0.318 0.750215   
## occupationCraft-repair 4.504e-02 9.450e-02 0.477 0.633656   
## occupationExec-managerial 7.712e-01 9.067e-02 8.506 < 2e-16 \*\*\*  
## occupationFarming-fishing -1.138e+00 1.622e-01 -7.013 2.34e-12 \*\*\*  
## occupationHandlers-cleaners -7.905e-01 1.724e-01 -4.585 4.54e-06 \*\*\*  
## occupationMachine-op-inspct -2.191e-01 1.198e-01 -1.830 0.067290 .   
## occupationOther-service -8.188e-01 1.385e-01 -5.913 3.35e-09 \*\*\*  
## occupationPriv-house-serv -3.536e+00 1.884e+00 -1.877 0.060505 .   
## occupationProf-specialty 5.364e-01 9.484e-02 5.656 1.55e-08 \*\*\*  
## occupationProtective-serv 6.011e-01 1.490e-01 4.036 5.44e-05 \*\*\*  
## occupationSales 2.847e-01 9.733e-02 2.925 0.003442 \*\*   
## occupationTech-support 6.827e-01 1.321e-01 5.169 2.36e-07 \*\*\*  
## occupationTransport-moving -1.167e-01 1.185e-01 -0.985 0.324464   
## relationshipNot-in-family -8.975e-01 1.916e-01 -4.684 2.81e-06 \*\*\*  
## relationshipOther-relative -1.147e+00 2.580e-01 -4.448 8.69e-06 \*\*\*  
## relationshipOwn-child -1.824e+00 2.363e-01 -7.717 1.19e-14 \*\*\*  
## relationshipUnmarried -1.065e+00 2.163e-01 -4.926 8.38e-07 \*\*\*  
## relationshipWife 1.459e+00 1.232e-01 11.843 < 2e-16 \*\*\*  
## raceAsian-Pac-Islander 6.064e-01 3.199e-01 1.896 0.058002 .   
## raceBlack 4.506e-01 2.842e-01 1.586 0.112837   
## raceOther 5.073e-02 4.211e-01 0.120 0.904125   
## raceWhite 6.532e-01 2.706e-01 2.414 0.015783 \*   
## sexMale 8.813e-01 9.338e-02 9.438 < 2e-16 \*\*\*  
## capital\_gain 3.123e-04 1.253e-05 24.933 < 2e-16 \*\*\*  
## capital\_loss 6.557e-04 4.557e-05 14.391 < 2e-16 \*\*\*  
## hr\_per\_week 2.939e-02 1.980e-03 14.845 < 2e-16 \*\*\*  
## countryEurope 1.109e-01 2.544e-01 0.436 0.663006   
## countryLatin.and.South.America -5.182e-01 2.555e-01 -2.028 0.042598 \*   
## countryNorth.America 5.868e-02 2.038e-01 0.288 0.773372   
## countryOther -3.572e-01 2.343e-01 -1.524 0.127484   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 24138 on 21502 degrees of freedom  
## Residual deviance: 14042 on 21461 degrees of freedom  
## AIC: 14126  
##   
## Number of Fisher Scoring iterations: 12

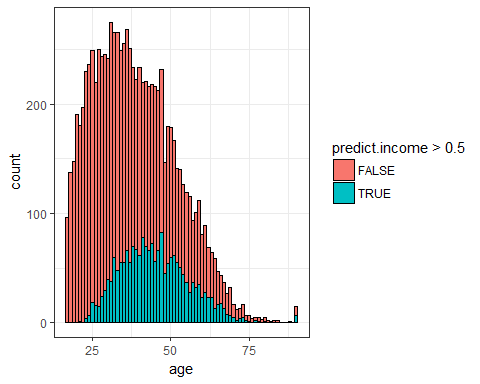
#step.model<-step(model)  
  
#summary(step.model)  
  
#predict.income<-predict(model,newdata=test,type='response')  
  
test$predict.income<-predict(model,newdata=test,type='response')  
  
  
table(test$income,test$predict.income>0.5)

##   
## FALSE TRUE  
## <=50K 6377 543  
## >50K 871 1424

accuracy<-(6372+1423)/(6372+1423+548+872)  
  
accuracy #84.6%

## [1] 0.8459034

ggplot(test,aes(age))+geom\_histogram(aes(fill=predict.income>0.5),color='black',binwidth = 1)+theme\_bw()

 #\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*