Stat6021\_Project2

Group 5

11/12/2019

Adult is the data set from 1994 Census database was offered by Barry Becker. (https://archive.ics.uci.edu/ml/datasets/adult). The Prediction task is to determine whether a person makes over $50K a year. Based on economic inflation (<http://www.in2013dollars.com/us/inflation/1994?amount=50000>), **$50K in 1994 is today (2019) worth $86K.** The reader can keep this in mind if the society structure is not changed. This work aims to predict whether or not an individual can earn $86K.

We start by data cleaning and mutating the data, since a number of the categorical variables have many classes. We broaden these classes and redefine new ones. In order to obtain the whole picture of this dataset, we make box plots grouped by income (>$50K, <$50K) first. The understanding the statistically significant predictors is important before we proceed machine learning.

## store data file with the variable name data  
## data cleaning  
## import library   
library(stringr)  
library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

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

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

library(ROCR)

## Loading required package: gplots

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

library(boot)  
library(extrafont)

## Registering fonts with R

library(ggthemes)  
data<-read.csv("adult.csv", header=FALSE ,sep=",", na.string = "?")  
#str\_replace(data, "-", ".")  
nr<-nrow(data)  
df<-data.frame(data)  
df = df[-1,] # row 1, sex has unwanted lable  
df[1, 1] <-39  
df<- na.omit(df)  
row.names(df) <- 1:nrow(df)  
data<-df

colnames(data)<-c("age","workclass", "fnlwgt", "education", "education\_num", "marital\_status", "occupation", "relationship", "race", "sex", "capital\_gain", "capital\_loss", "hours\_per\_week", "native\_country", "income")  
attach(data)  
#data  
   
#remoce missing data  
data <- na.omit(data)

is.numeric(age)

## [1] FALSE

age <-as.numeric(age)  
is.numeric(age)

## [1] TRUE

is.numeric(fnlwgt)

## [1] TRUE

is.numeric(education\_num)

## [1] TRUE

is.numeric(capital\_gain)

## [1] TRUE

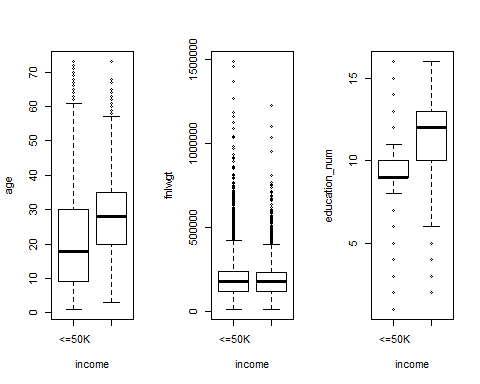
is.numeric(capital\_loss)

## [1] TRUE

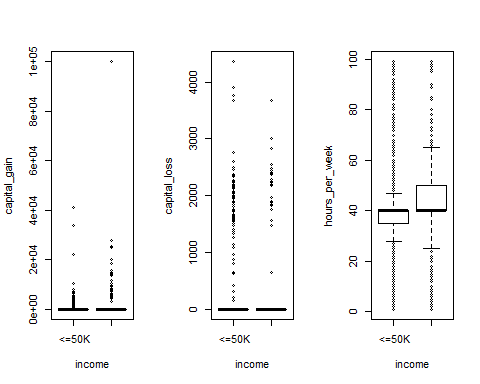
is.numeric(hours\_per\_week)

## [1] TRUE

#Use box plot to see each predictor vs. income  
##############  
par(mfrow=c(1,3))  
boxplot(age~income)  
boxplot(fnlwgt~income)  
  
###############  
boxplot(education\_num~income)

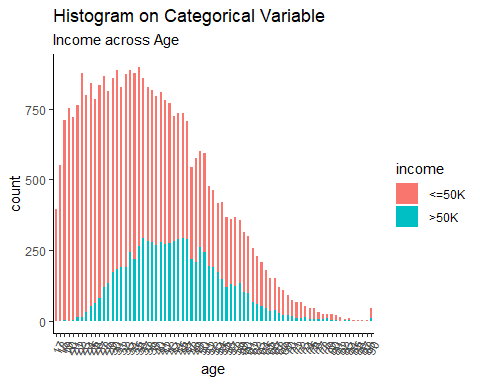


###############  
  
boxplot(capital\_gain~income)  
boxplot(capital\_loss~income)  
###############  
boxplot(hours\_per\_week~income)

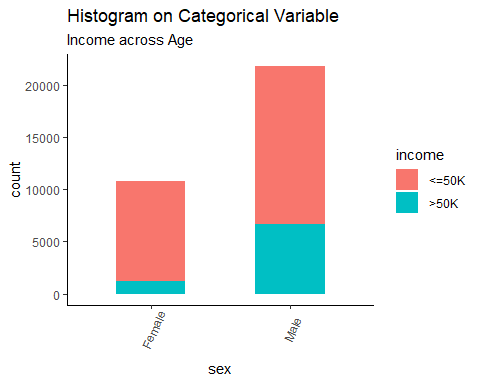


###############

theme\_set(theme\_classic())  
  
# Histogram on a Categorical variable  
g <- ggplot(data, aes(age))  
g + geom\_bar(aes(fill=income), width = 0.5) +   
 theme(axis.text.x = element\_text(angle=65, vjust=0.6)) +   
 labs(title="Histogram on Categorical Variable",   
 subtitle="Income across Age")



theme\_set(theme\_classic())  
  
# Histogram on a Categorical variable  
g <- ggplot(data, aes(sex))  
g + geom\_bar(aes(fill=income), width = 0.5) +   
 theme(axis.text.x = element\_text(angle=65, vjust=0.6)) +   
 labs(title="Histogram on Categorical Variable",   
 subtitle="Income across Age")



data$education <- trimws(data$education)

summary(data$education)

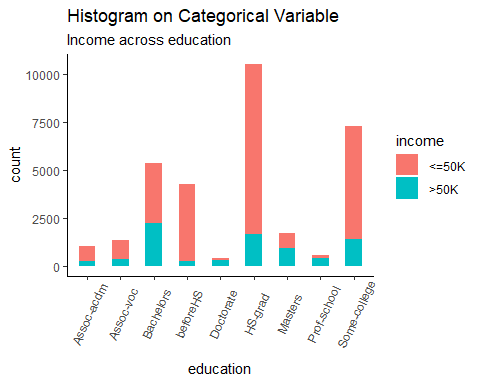
## Length Class Mode   
## 32560 character character

###combine high school below or 12th together   
data$education <-gsub('^12th', 'beforeHS', data$education)  
data$education <-gsub('^10th', 'beforeHS', data$education)  
data$education <-gsub('^11th', 'beforeHS', data$education)  
data$education <-gsub('^1st-4th', 'beforeHS', data$education)  
data$education <-gsub('^5th-6th', 'beforeHS', data$education)  
data$education <-gsub('^7th-8th', 'beforeHS', data$education)  
data$education <-gsub('^9th', 'beforeHS', data$education)  
data$education <-gsub('^Preschool', 'beforeHS', data$education)  
data$education<-as.factor(data$education)

summary(data$education)

## Assoc-acdm Assoc-voc Bachelors beforeHS Doctorate   
## 1067 1382 5354 4253 413   
## HS-grad Masters Prof-school Some-college   
## 10501 1723 576 7291

theme\_set(theme\_classic())  
  
# Histogram on a Categorical variable  
g <- ggplot(data, aes(education))  
g + geom\_bar(aes(fill=income), width = 0.5) +   
 theme(axis.text.x = element\_text(angle=65, vjust=0.6)) +   
 labs(title="Histogram on Categorical Variable",   
 subtitle="Income across education")



summary(data$workclass)

## ? Federal-gov Local-gov Never-worked   
## 1836 960 2093 7   
## Private Self-emp-inc Self-emp-not-inc State-gov   
## 22696 1116 2541 1297   
## Without-pay   
## 14

data$workclass <- trimws(data$workclass)

levels(data$workclass)[1] <- 'Unknown'  
  
# combine into Sele-Employed job  
data$workclass <- gsub('^Self-emp-inc', 'Self-Employed', data$workclass)  
data$workclass <- gsub('^Self-emp-not-inc', 'Self-Employed', data$workclass)  
  
# combine into Other/Unknown  
data$workclass <- gsub('^Never-worked', 'Other', data$workclass)  
data$workclass <- gsub('^Without-pay', 'Other', data$workclass)  
data$workclass <- gsub('^Other', 'Others', data$workclass)  
data$workclass <- gsub('^Unknown', 'Other', data$workclass)  
  
# combine into Government job  
data$workclass <- gsub('^Federal-gov', 'Government', data$workclass)  
data$workclass <- gsub('^Local-gov', 'Government', data$workclass)  
data$workclass <- gsub('^State-gov', 'Government', data$workclass)   
   
  
data$workclass <- as.factor(data$workclass)

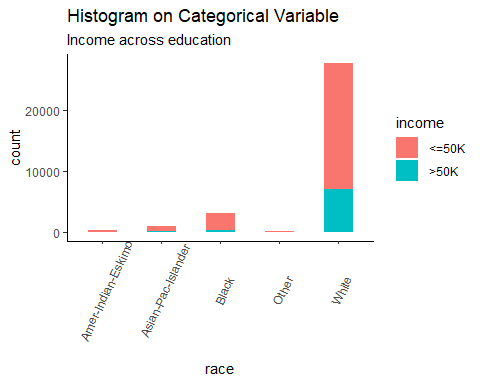
data <- na.omit(data)

data$workclass <- gsub('[[:punct:]]', 'Other', data$workclass)  
data$workclass <- as.factor(data$workclass)

summary(data$workclass)

## Government Other Others Private   
## 4350 1836 21 22696   
## SelfOtherEmployed   
## 3657

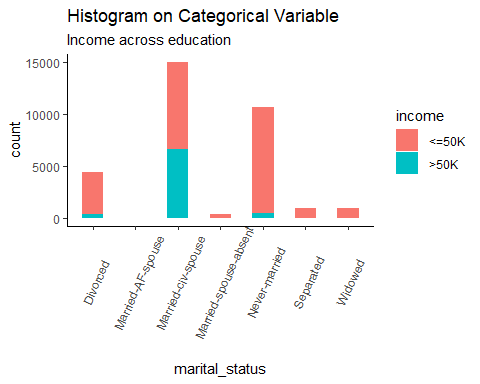
theme\_set(theme\_classic())  
  
# Histogram on a Categorical variable  
g <- ggplot(data, aes(race))  
g + geom\_bar(aes(fill=income), width = 0.5) +   
 theme(axis.text.x = element\_text(angle=65, vjust=0.6)) +   
 labs(title="Histogram on Categorical Variable",   
 subtitle="Income across education")



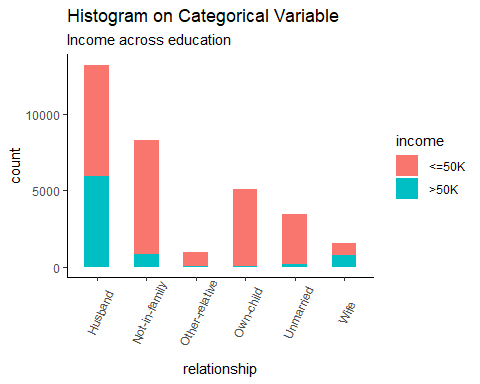
summary(data$marital\_status)

## Divorced Married-AF-spouse Married-civ-spouse   
## 4443 23 14976   
## Married-spouse-absent Never-married Separated   
## 418 10682 1025   
## Widowed   
## 993

theme\_set(theme\_classic())  
  
# Histogram on a Categorical variable  
g <- ggplot(data, aes(marital\_status))  
g + geom\_bar(aes(fill=income), width = 0.5) +   
 theme(axis.text.x = element\_text(angle=65, vjust=0.6)) +   
 labs(title="Histogram on Categorical Variable",   
 subtitle="Income across education")



theme\_set(theme\_classic())  
  
# Histogram on a Categorical variable  
g <- ggplot(data, aes(relationship))  
g + geom\_bar(aes(fill=income), width = 0.5) +   
 theme(axis.text.x = element\_text(angle=65, vjust=0.6)) +   
 labs(title="Histogram on Categorical Variable",   
 subtitle="Income across education")



data$native\_country <- trimws(data$native\_country)   
summary(data$native\_country)

## Length Class Mode   
## 32560 character character

#Need to delete Outlying-US(Guam-USVI-etc)   
data$native\_country <- as.factor(data$native\_country)

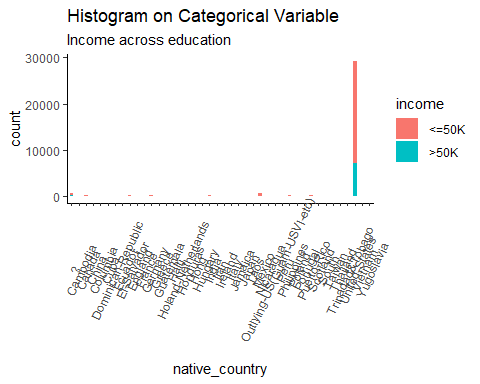
summary(data$native\_country)

## ? 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 29169   
## Vietnam Yugoslavia   
## 67 16

data <- na.omit(data)

data$native\_country <- as.factor(data$native\_country)

theme\_set(theme\_classic())  
  
# Histogram on a Categorical variable  
g <- ggplot(data, aes(native\_country))  
g + geom\_bar(aes(fill=income), width = 0.5) +   
 theme(axis.text.x = element\_text(angle=65, vjust=0.6)) +   
 labs(title="Histogram on Categorical Variable",   
 subtitle="Income across education")

 ```