project

Jingning Zheng

2023-05-16

The question of this study is what factors generally influence the purchase of bicycles. The subjects of the research are mainly adults, mainly to understand some of the factors that influence people to buy bicycles.

The data comes from Kaggle, by collecting the background information of 1000 people and whether to buy a bicycle. The background information collected includes marital status, gender, age, number of children, number of vehicles, whether they own a house, etc.

First, we check the data to exclude rows containing NA values or null values. Then copy the response variable to form a new column and convert it to 1 and 0. The response variable of this data is whether to buy a bicycle. Finally delete the ID and response variables that were copied.

```
data1=read.csv("bike_buyers_clean.csv",header=T)
names(data1)
```

```
##  [1] "ID"               "Marital.Status"   "Gender"           "Income"
##  [5] "Children"         "Education"        "Occupation"       "Home.Owner"
##  [9] "Cars"             "Commute.Distance" "Region"           "Age"
## [13] "Purchased.Bike"
```

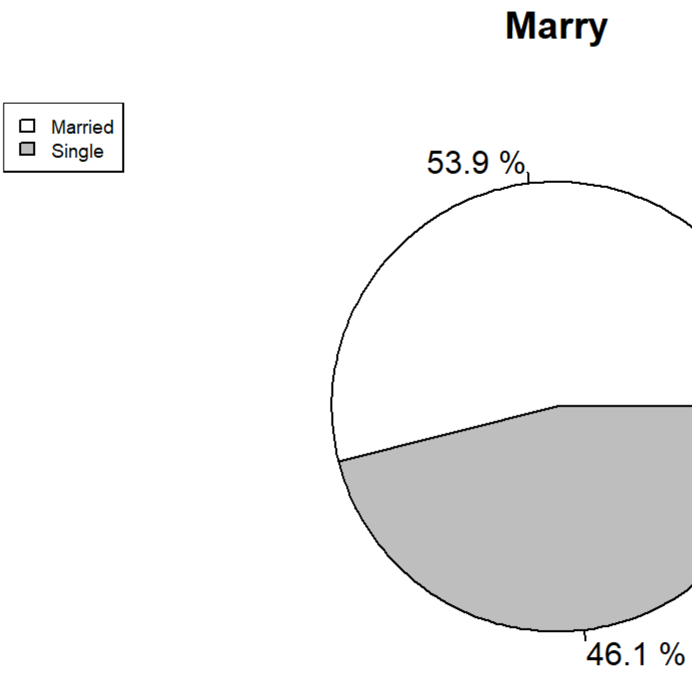
```
summary(data1)
```

```
##        ID        Marital.Status        Gender             Income      
##  Min.   :11000   Length:1000        Length:1000        Min.   : 10000  
##  1st Qu.:15291   Class :character   Class :character   1st Qu.: 30000  
##  Median :19744   Mode  :character   Mode  :character   Median : 60000  
##  Mean   :19966                                         Mean   : 56140  
##  3rd Qu.:24471                                         3rd Qu.: 70000  
##  Max.   :29447                                         Max.   :170000  
##     Children      Education          Occupation         Home.Owner       
##  Min.   :0.000   Length:1000        Length:1000        Length:1000       
##  1st Qu.:0.000   Class :character   Class :character   Class :character  
##  Median :2.000   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :1.908                                                           
##  3rd Qu.:3.000                                                           
##  Max.   :5.000                                                           
##       Cars       Commute.Distance      Region               Age       
##  Min.   :0.000   Length:1000        Length:1000        Min.   :25.00  
##  1st Qu.:1.000   Class :character   Class :character   1st Qu.:35.00  
##  Median :1.000   Mode  :character   Mode  :character   Median :43.00  
##  Mean   :1.452                                         Mean   :44.19  
##  3rd Qu.:2.000                                         3rd Qu.:52.00  
##  Max.   :4.000                                         Max.   :89.00  
##  Purchased.Bike    
##  Length:1000       
##  Class :character  
##  Mode  :character  
##                    
##                    
##                    
```

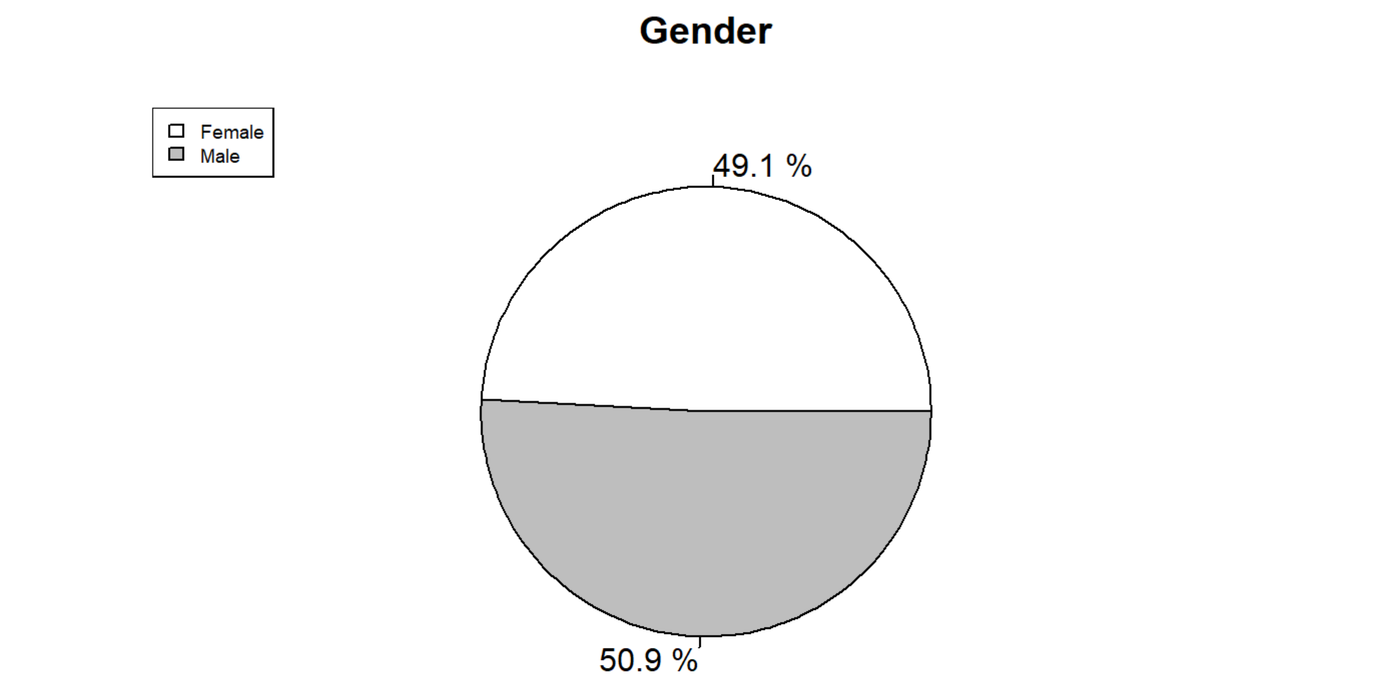
```
data1$Purchased=data1$Purchased.Bike
data1$Purchased=ifelse(data1$Purchased=="Yes",1,0)
data1=data1[,-c(1,13)]
```

Data Visualization

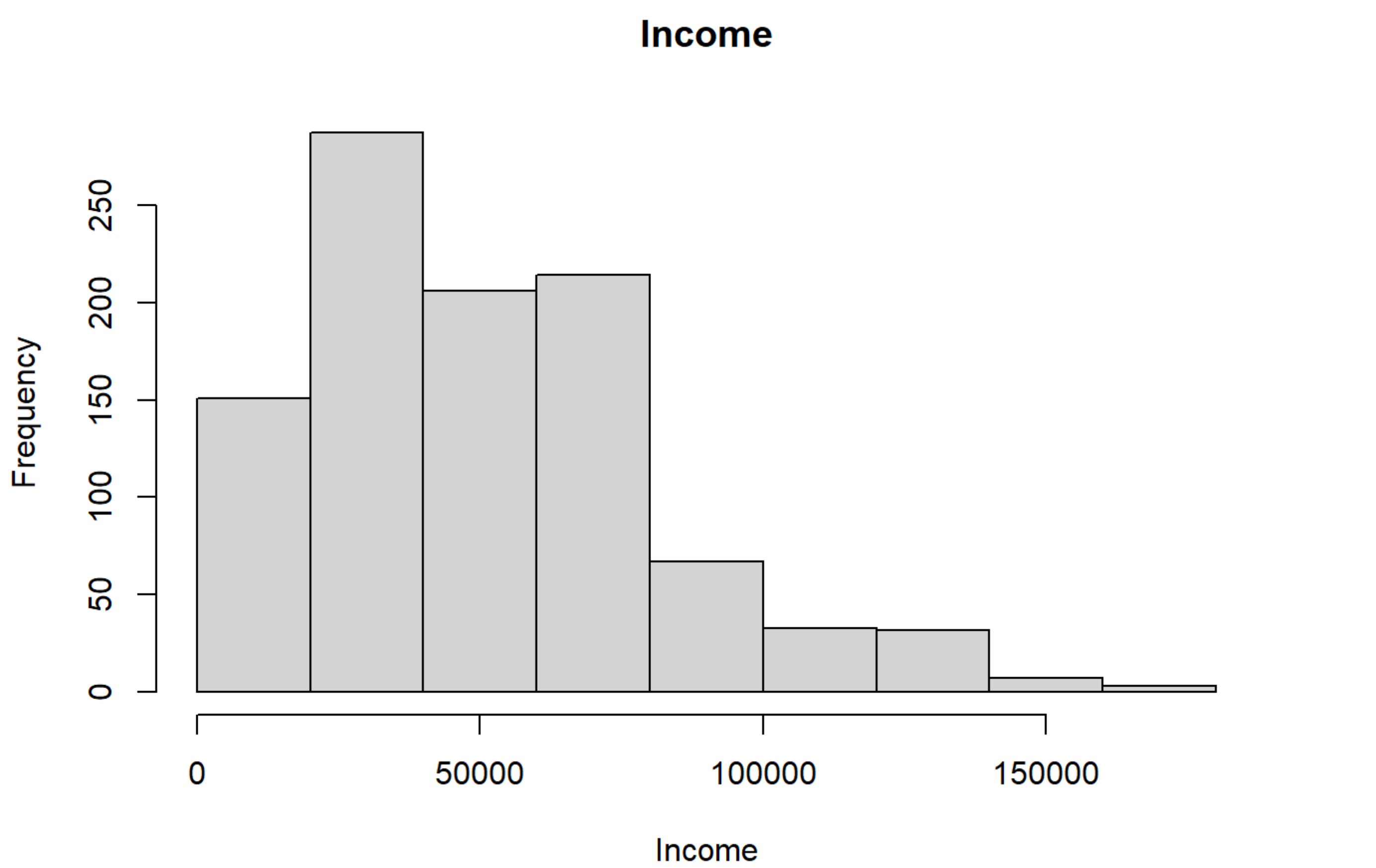
```
Marry = table(data1$Marital.Status)
piepercent<-paste(round(100*Marry/sum(Marry),2),"%")
pie(Marry,labels=piepercent,main="Marry",col=c("white","gray"))
legend("topleft",legend=c("Married","Single"),cex=0.6,fill=c("white","gray"))
```



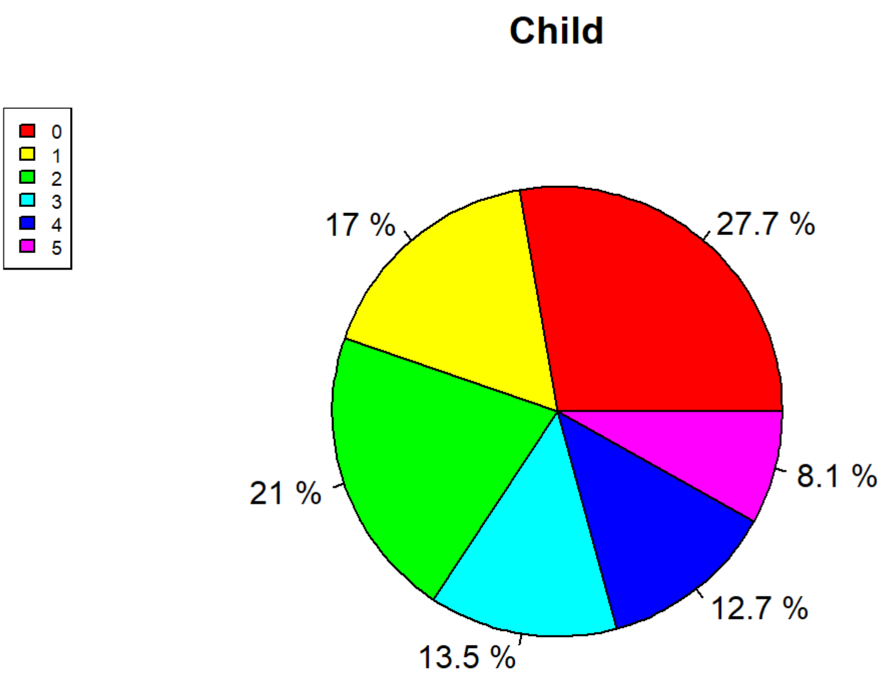

```
Gender = table(data1$Gender)
piepercent<-paste(round(100*Gender/sum(Gender),2),"%")
pie(Gender,labels=piepercent,main="Gender",col=c("white","gray"))
legend("topleft",legend=c("Female","Male"),cex=0.6,fill=c("white","gray"))
```



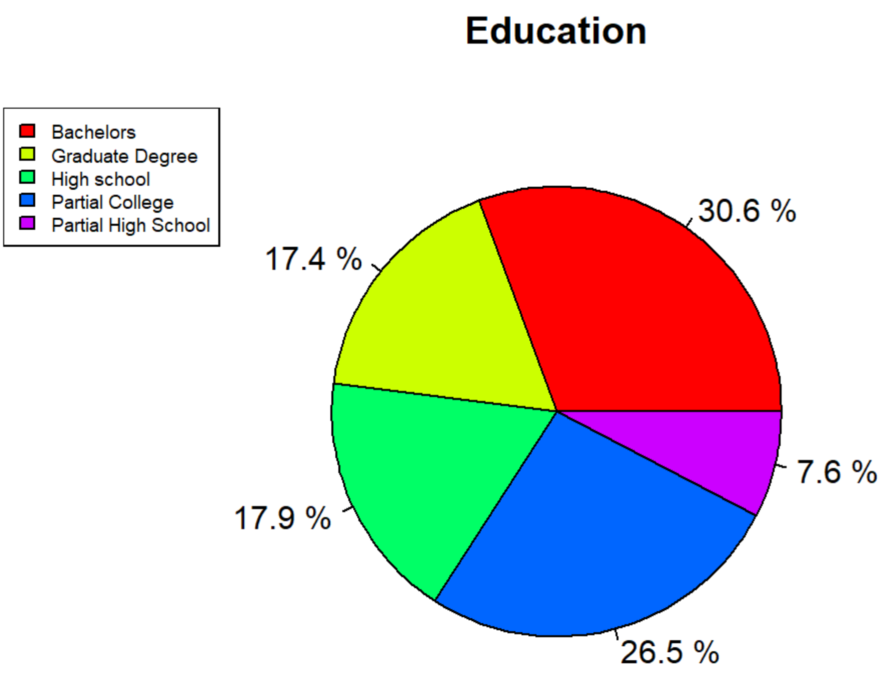

```
hist(data1$Income,xlab = "Income",main = "Income")
```



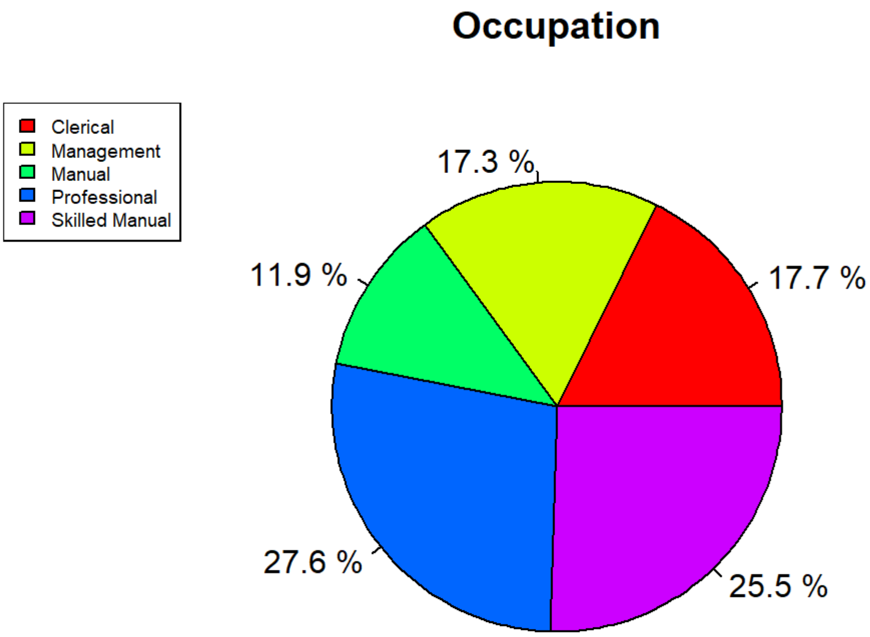
```
Child = table(data1$Children)
piepercent<-paste(round(100*Child/sum(Child),2),"%")
pie(Child,labels=piepercent,main="Child",col=rainbow(length(Child)))
legend("topleft",legend=c("0","1","2","3","4","5"),
       cex=0.6,fill=rainbow(length(Child)))
```



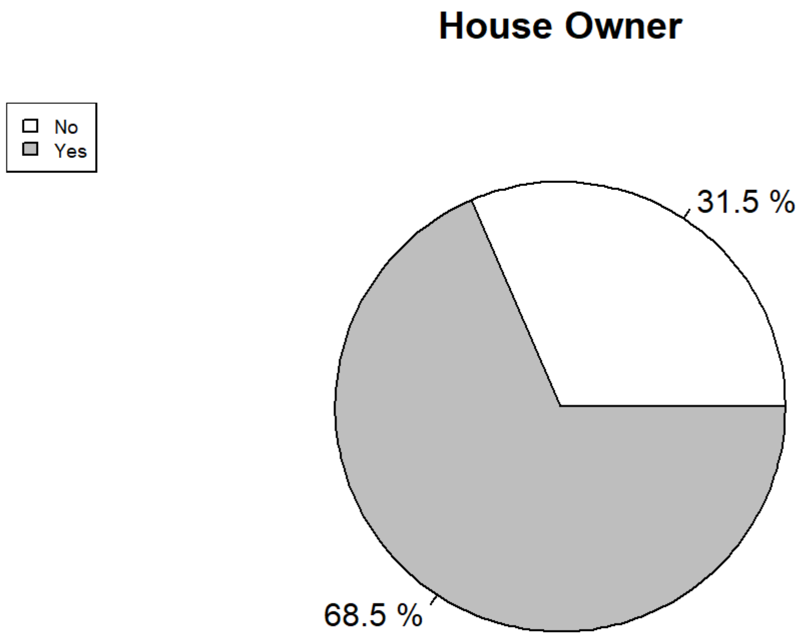

```
Edu = table(data1$Education)
piepercent<-paste(round(100*Edu/sum(Edu),2),"%")
pie(Edu,labels=piepercent,main="Education",col=rainbow(length(Edu)))
legend("topleft",legend=c("Bachelors","Graduate Degree","High school","Partial College","Partia
l High School"),
       cex=0.6,fill=rainbow(length(Edu)))
```



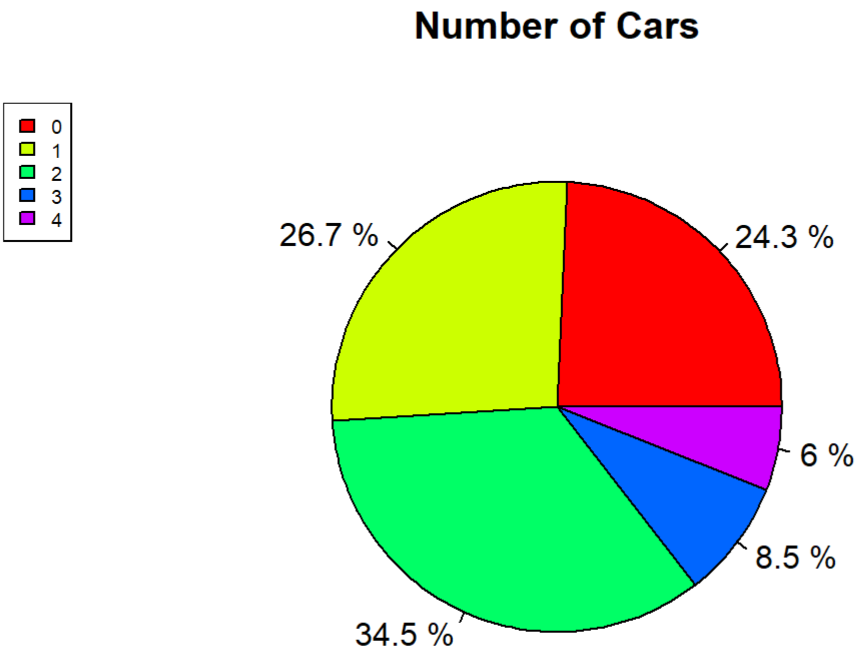

```
occ=table(data1$Occupation)
piepercent<-paste(round(100*occ/sum(occ),2),"%")
pie(occ,labels=piepercent,main="Occupation",col=rainbow(length(occ)))
legend("topleft",legend=c("Clerical","Management","Manual","Professional","Skilled Manual"),
       cex=0.6,fill=rainbow(length(occ)))
```



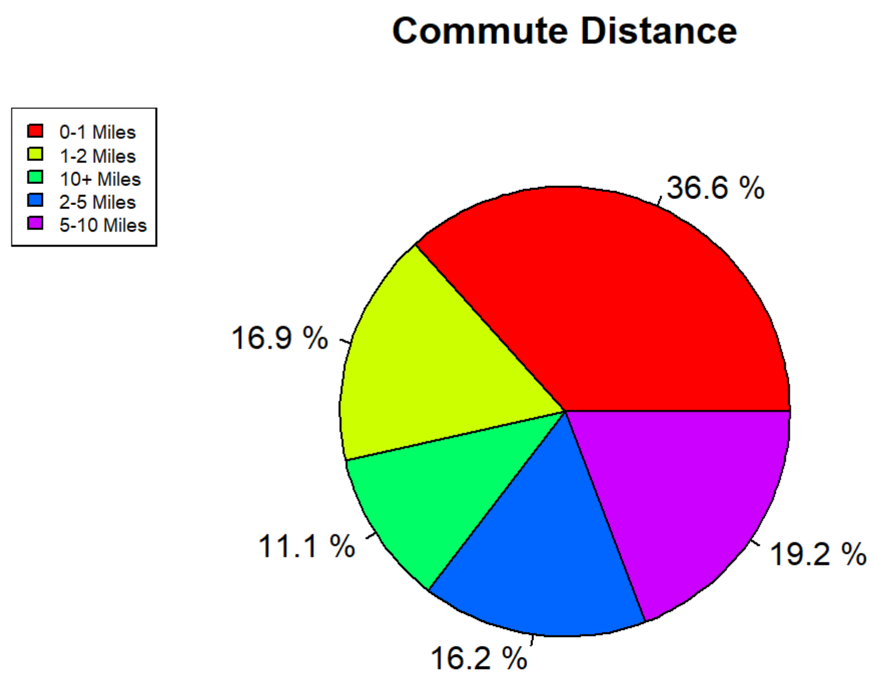

```
house = table(data1$Home.Owner)
piepercent<-paste(round(100*house/sum(house),2),"%")
pie(house,labels=piepercent,main="House Owner",col=c("white","gray"))
legend("topleft",legend=c("No","Yes"),cex=0.6,fill=c("white","gray"))
```



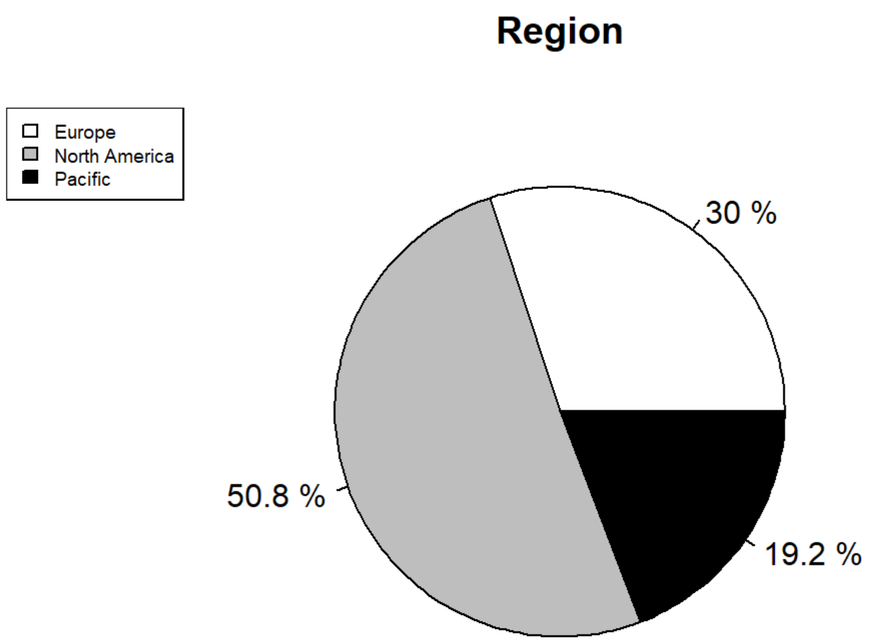

```
n.car=table(data1$Cars)
piepercent<-paste(round(100*n.car/sum(n.car),2),"%")
pie(n.car,labels=piepercent,main="Number of Cars",col=rainbow(length(n.car)))
legend("topleft",legend=c("0","1","2","3","4"),
       cex=0.6,fill=rainbow(length(n.car)))
```



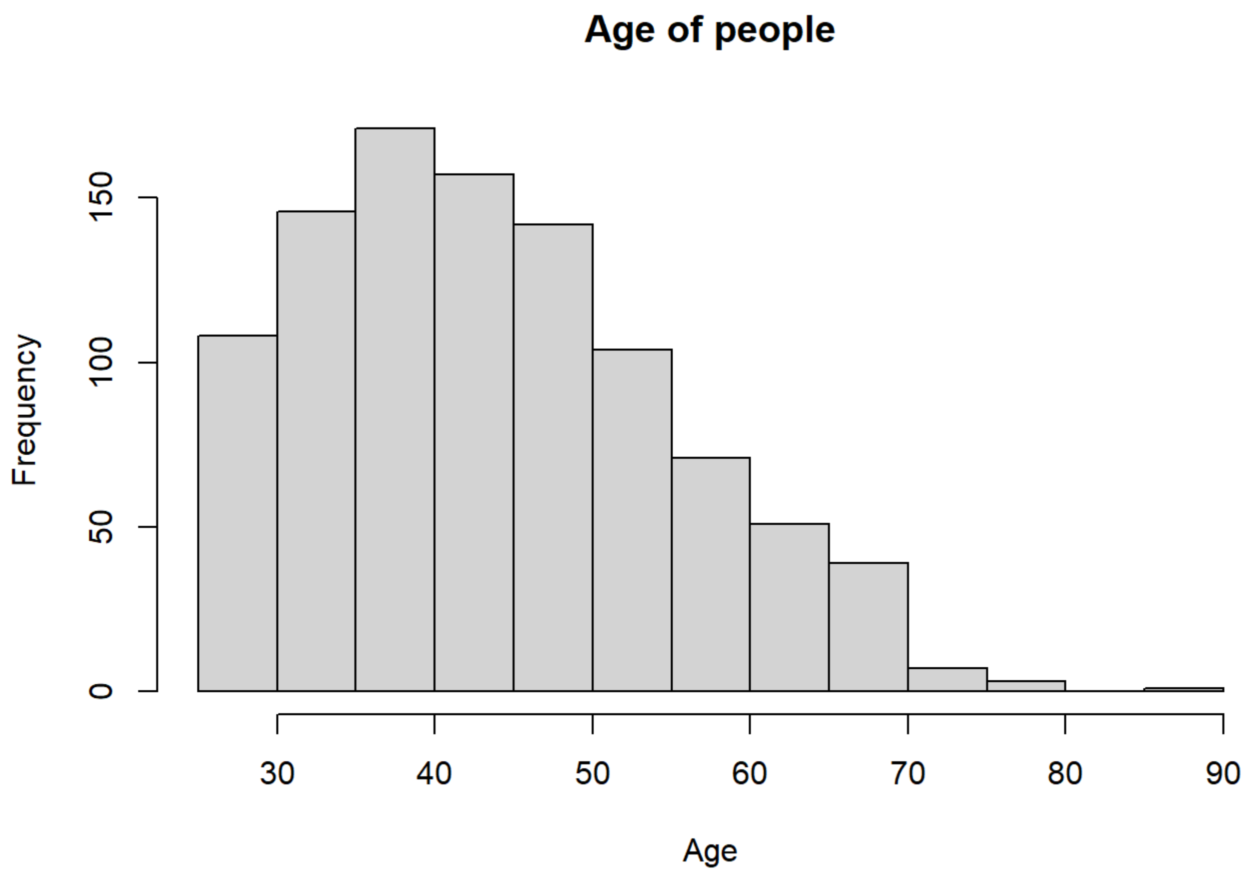

```
distance=table(data1$Commute.Distance)
piepercent<-paste(round(100*distance/sum(distance),2),"%")
pie(distance,labels=piepercent,main="Commute Distance",col=rainbow(length(distance)))
legend("topleft",legend=c("0-1 Miles","1-2 Miles","10+ Miles","2-5 Miles","5-10 Miles"),
       cex=0.6,fill=rainbow(length(distance)))
```



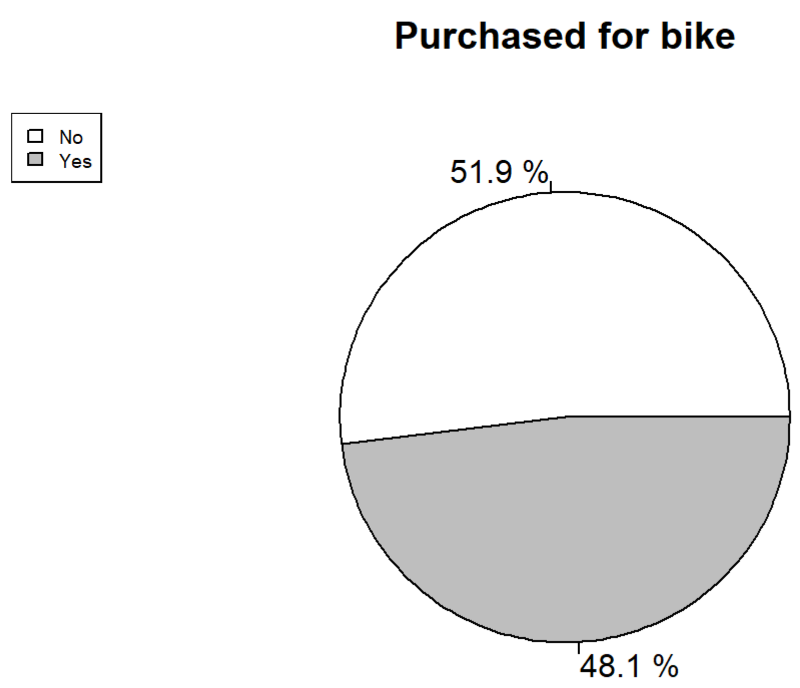

```
region=table(data1$Region)
piepercent<-paste(round(100*region/sum(region),2),"%")
pie(region,labels=piepercent,main="Region",col=c("white","gray","black"))
legend("topleft",legend=c("Europe","North America","Pacific"),
       cex=0.6,fill=c("white","gray","black"))
```




```
hist(data1$Age,xlab = "Age",main = "Age of people")
```



```
pay=table(data1$Purchased)
piepercent<-paste(round(100*pay/sum(pay),2),"%")
pie(pay,labels=piepercent,main="Purchased for bike",col=c("white","gray"))
legend("topleft",legend=c("No","Yes"),cex=0.6,fill=c("white","gray"))
```



Four models are fitted here, they are null model, full model, step model and select model. Among them, the step model uses both sides stepwise, and the select model is by deleting the non-significant variable in the full model.

```
#### null model
mod0=glm(Purchased~1,data=data1,family = binomial(link="logit"))
summary(mod0)
```

```
##
## Call:
## glm(formula = Purchased ~ 1, family = binomial(link = "logit"),
##     data = data1)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.145  -1.145  -1.145   1.210   1.210
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.07604    0.06329  -1.201     0.23
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1384.9  on 999  degrees of freedom
## Residual deviance: 1384.9  on 999  degrees of freedom
## AIC: 1386.9
##
## Number of Fisher Scoring iterations: 3
```

```
#### full model

modf=glm(Purchased~.,data=data1,family = binomial(link="logit"))
summary(modf)
```

```
##
## Call:
## glm(formula = Purchased ~ ., family = binomial(link = "logit"),
##     data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9875  -1.0264  -0.5319   1.0631   2.3132
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -3.359e-01  4.548e-01  -0.738 0.460214
## Marital.StatusSingle            6.540e-01  1.545e-01   4.232 2.32e-05 ***
## GenderMale                      2.186e-02  1.380e-01   0.158 0.874088
## Income                          1.343e-05  4.006e-06   3.351 0.000804 ***
## Children                       -1.255e-01  5.349e-02  -2.347 0.018943 *
## EducationGraduate Degree       -3.635e-01  2.223e-01  -1.635 0.101983
## EducationHigh School            1.709e-01  2.486e-01   0.687 0.491958
## EducationPartial College       -2.110e-01  2.155e-01  -0.979 0.327580
## EducationPartial High School -4.842e-01  3.613e-01  -1.340 0.180214
## OccupationManagement           -1.536e-01  4.136e-01  -0.371 0.710311
## OccupationManual               -3.921e-02  2.859e-01  -0.137 0.890934
## OccupationProfessional          4.124e-01  3.327e-01   1.240 0.215086
## OccupationSkilled Manual       -8.236e-02  2.683e-01  -0.307 0.758858
## Home.OwnerYes                   3.332e-01  1.688e-01   1.974 0.048338 *
## Cars                           -4.691e-01  9.124e-02  -5.142 2.72e-07 ***
## Commute.Distance1-2 Miles      -1.466e-01  2.113e-01  -0.694 0.487823
## Commute.Distance10+ Miles      -1.127e+00  2.972e-01  -3.793 0.000149 ***
## Commute.Distance2-5 Miles       2.187e-03  2.156e-01   0.010 0.991906
## Commute.Distance5-10 Miles     -7.018e-01  2.459e-01  -2.854 0.004312 **
## RegionNorth America            -1.313e-01  2.163e-01  -0.607 0.543918
## RegionPacific                   8.034e-01  2.436e-01   3.297 0.000976 ***
## Age                             3.000e-03  7.867e-03   0.381 0.702979
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1384.9  on 999  degrees of freedom
## Residual deviance: 1239.0  on 978  degrees of freedom
## AIC: 1283
##
## Number of Fisher Scoring iterations: 4
```

```
#### step model
mod1=step(modf,direction = "both",trace = F)
summary(mod1)
```

```
##
## Call:
## glm(formula = Purchased ~ Marital.Status + Income + Children +
##     Occupation + Home.Owner + Cars + Commute.Distance + Region,
##     family = binomial(link = "logit"), data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9454  -1.0439  -0.5495   1.0664   2.1384
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -2.921e-01  2.630e-01  -1.111 0.266729
## Marital.StatusSingle        6.259e-01  1.492e-01   4.195 2.73e-05 ***
## Income                      1.079e-05  3.784e-06   2.852 0.004350 **
## Children                   -1.204e-01  4.660e-02  -2.583 0.009789 **
## OccupationManagement        9.977e-02  3.498e-01   0.285 0.775499
## OccupationManual           -5.922e-02  2.634e-01  -0.225 0.822139
## OccupationProfessional      6.291e-01  3.097e-01   2.031 0.042206 *
## OccupationSkilled Manual    8.994e-02  2.538e-01   0.354 0.723073
## Home.OwnerYes               3.019e-01  1.646e-01   1.835 0.066580 .
## Cars                       -4.364e-01  7.798e-02  -5.596 2.20e-08 ***
## Commute.Distance1-2 Miles  -1.402e-01  2.044e-01  -0.686 0.492879
## Commute.Distance10+ Miles  -1.086e+00  2.873e-01  -3.781 0.000156 ***
## Commute.Distance2-5 Miles   6.032e-02  2.122e-01   0.284 0.776169
## Commute.Distance5-10 Miles -6.625e-01  2.220e-01  -2.984 0.002849 **
## RegionNorth America        -2.310e-01  2.109e-01  -1.095 0.273344
## RegionPacific               7.327e-01  2.392e-01   3.063 0.002188 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1384.9  on 999  degrees of freedom
## Residual deviance: 1247.0  on 984  degrees of freedom
## AIC: 1279
##
## Number of Fisher Scoring iterations: 4
```

```
### select model
mod2=glm(Purchased~Marital.Status+Income+Children+Cars+Commute.Distance+Region,
         data=data1,family = binomial(link="logit"))
summary(mod2)
```

```
##
## Call:
## glm(formula = Purchased ~ Marital.Status + Income + Children +
##     Cars + Commute.Distance + Region, family = binomial(link = "logit"),
##     data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0457  -1.0378  -0.6069   1.0492   2.2282
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -1.323e-01  1.829e-01  -0.723 0.469572
## Marital.StatusSingle       5.027e-01  1.383e-01   3.635 0.000278 ***
## Income                     1.411e-05  2.693e-06   5.240 1.61e-07 ***
## Children                  -9.436e-02  4.471e-02  -2.110 0.034820 *
## Cars                      -4.763e-01  7.550e-02  -6.309 2.81e-10 ***
## Commute.Distance1-2 Miles  -1.578e-01  2.035e-01  -0.775 0.438253
## Commute.Distance10+ Miles  -8.132e-01  2.674e-01  -3.041 0.002358 **
## Commute.Distance2-5 Miles   2.062e-01  2.061e-01   1.001 0.317001
## Commute.Distance5-10 Miles -5.180e-01  2.107e-01  -2.458 0.013953 *
## RegionNorth America        -1.397e-01  1.727e-01  -0.809 0.418669
## RegionPacific               7.769e-01  2.189e-01   3.548 0.000388 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1384.9  on 999  degrees of freedom
## Residual deviance: 1258.6  on 989  degrees of freedom
## AIC: 1280.6
##
## Number of Fisher Scoring iterations: 4
```

The selected model is the select model, and it is compared with the other three models by the Likelihood Ratio test. From the results, the model is better than the other three models.

```
#### compare
anova(mod0, mod2, test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Purchased ~ 1
## Model 2: Purchased ~ Marital.Status + Income + Children + Cars + Commute.Distance +
##     Region
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       999     1384.8
## 2       989     1258.6 10   126.28 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(modf, mod2, test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Purchased ~ Marital.Status + Gender + Income + Children + Education +
##     Occupation + Home.Owner + Cars + Commute.Distance + Region +
##     Age
## Model 2: Purchased ~ Marital.Status + Income + Children + Cars + Commute.Distance +
##     Region
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1       978     1239.0
## 2       989     1258.6 -11   -19.62  0.05082 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod1, mod2, test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Purchased ~ Marital.Status + Income + Children + Occupation +
##     Home.Owner + Cars + Commute.Distance + Region
## Model 2: Purchased ~ Marital.Status + Income + Children + Cars + Commute.Distance +
##     Region
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       984     1247.0
## 2       989     1258.6 -5   -11.54  0.04166 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#leave one out cross-validation
prop <- sum(data1$Purchased)/nrow(data1)
prop
```

```
## [1] 0.481
```

```
predicted <- as.numeric(fitted(mod1) > prop)
table1=xtabs(~ data1$Purchased + predicted)
table1
```

```
##                predicted
## data1$Purchased   0   1
##               0 338 181
##               1 167 314
```

```
acc = (table1[1,1]+table1[2,2])/sum(table1)
acc
```

```
## [1] 0.652
```

```
pihat <- vector(length=1000)
for (i in 1:1000) {
  pihat[i] <-
    predict(update(mod2, subset=-i),
            newdata=data1[i,], type="response")
}

yy <- as.numeric(data1$Purchased > 0)
yhat <- as.numeric(pihat >prop)
confusion <- table(yy, yhat)
confusion
```

```
##    yhat
## yy    0   1
##   0 326 193
##   1 168 313
```

```
acc = (confusion[1,1]+confusion[2,2])/sum(confusion)
acc
```

```
## [1] 0.639
```

From the verification results, after 1000 repeated tests, the accuracy of the model is 0.64, which is 64%. This means that the model is applied to the current data, and its accuracy is 64%.

```
#### K-fold cross validation
cv.binary(mod0)
```

```
##
## Fold:  7 9 6 3 8 4 1 10 2 5
## Internal estimate of accuracy = 0.519
## Cross-validation estimate of accuracy = 0.519
```

```
cv.binary(modf)
```

```
##
## Fold:  10 5 8 3 2 9 7 4 6 1
## Internal estimate of accuracy = 0.655
## Cross-validation estimate of accuracy = 0.633
```

```
cv.binary(mod1)
```

```
##
## Fold:  1 3 10 5 7 9 4 6 2 8
## Internal estimate of accuracy = 0.657
## Cross-validation estimate of accuracy = 0.646
```

```
cv.binary(mod2)
```

```
##
## Fold:  7 3 10 4 9 2 1 6 8 5
## Internal estimate of accuracy = 0.654
## Cross-validation estimate of accuracy = 0.643
```

```
cost<-function(r,pi=0) mean(abs(r-pi)>0.8)
out0=cv.glm(data1,mod2,cost,K=10)
names(out0)
```

```
## [1] "call"  "K"     "delta" "seed"
```

```
out1=cv.glm(data1,mod2,cost,K=10)
out2=cv.glm(data1,mod2,cost,K=10)
out3=cv.glm(data1,mod2,cost,K=10)

out0$delta
```

```
## [1] 0.0160 0.0152
```

```
out1$delta
```

```
## [1] 0.0140 0.0139
```

```
out2$delta
```

```
## [1] 0.0170 0.0164
```

```
out3$delta
```

```
## [1] 0.0160 0.0158
```

From the perspective of random folding, the select model has the highest accuracy among the four models. Then four k-fold cross validations are performed on the select model, and k is set to 10. The results obtained are all around 0.198.

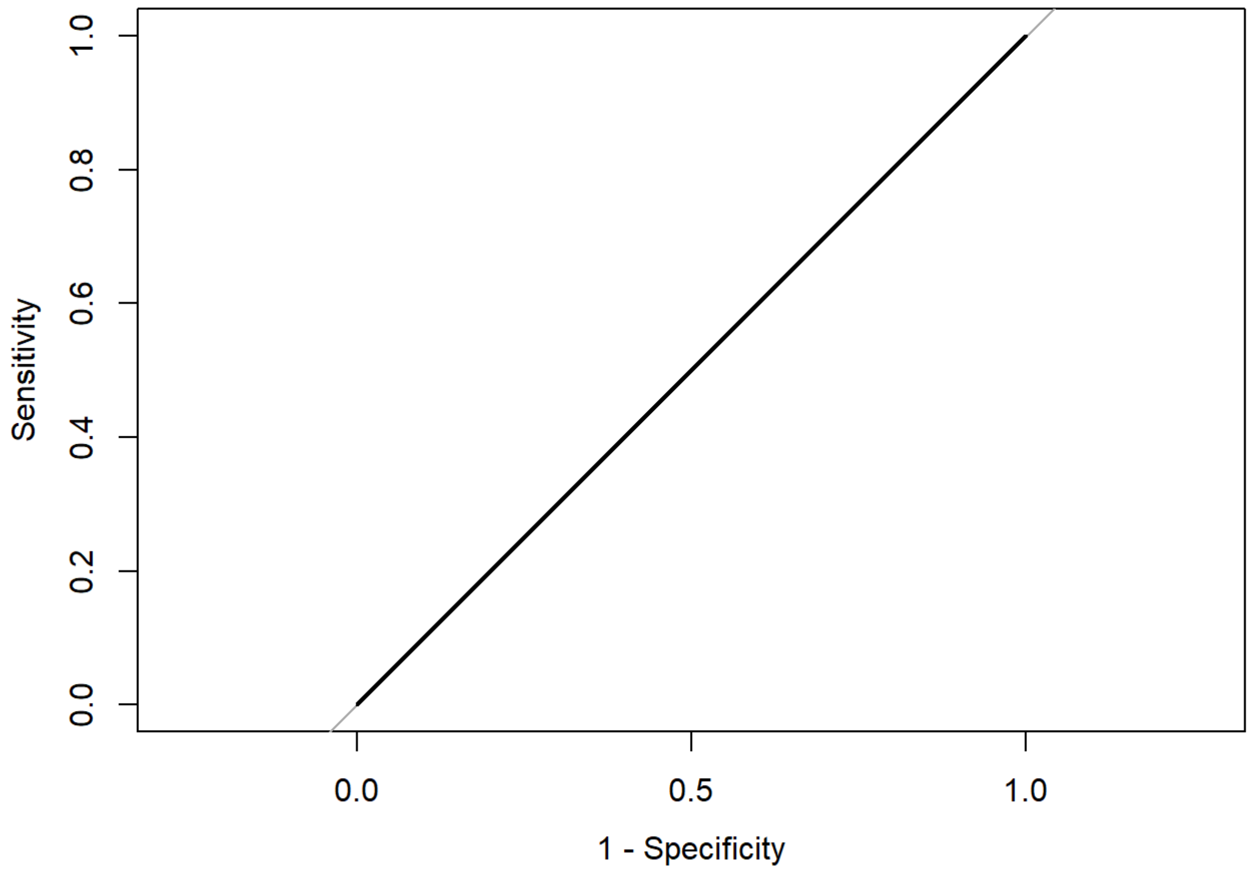
```
#### roc curve for each model
### null model

rocplot0 <- roc(Purchased ~ fitted(mod0), data=data1)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot.roc(rocplot0, legacy.axes=TRUE)
```



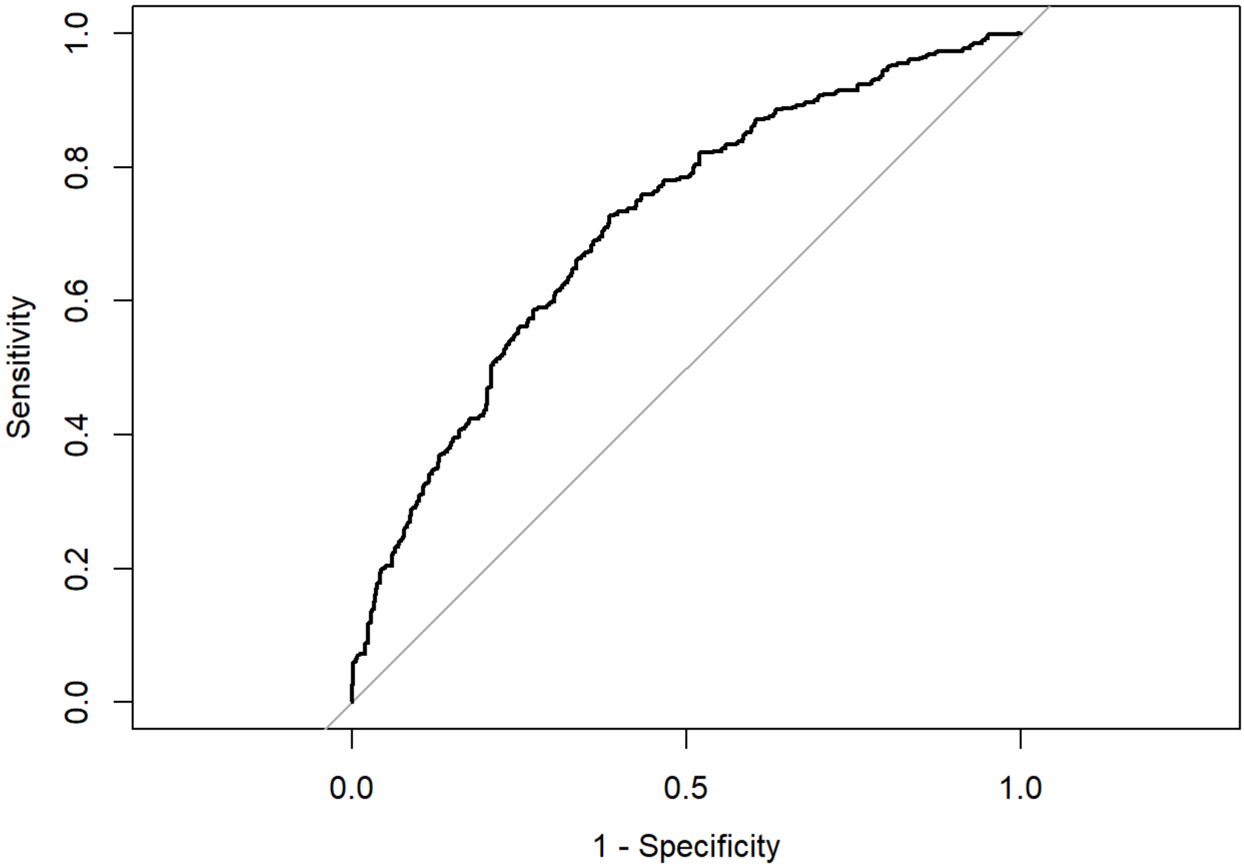

```
auc(rocplot0)
```

```
## Area under the curve: 0.5
```

```
### full model
rocplot1 <- roc(Purchased ~ fitted(modf), data=data1)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
plot.roc(rocplot1, legacy.axes=TRUE)
```



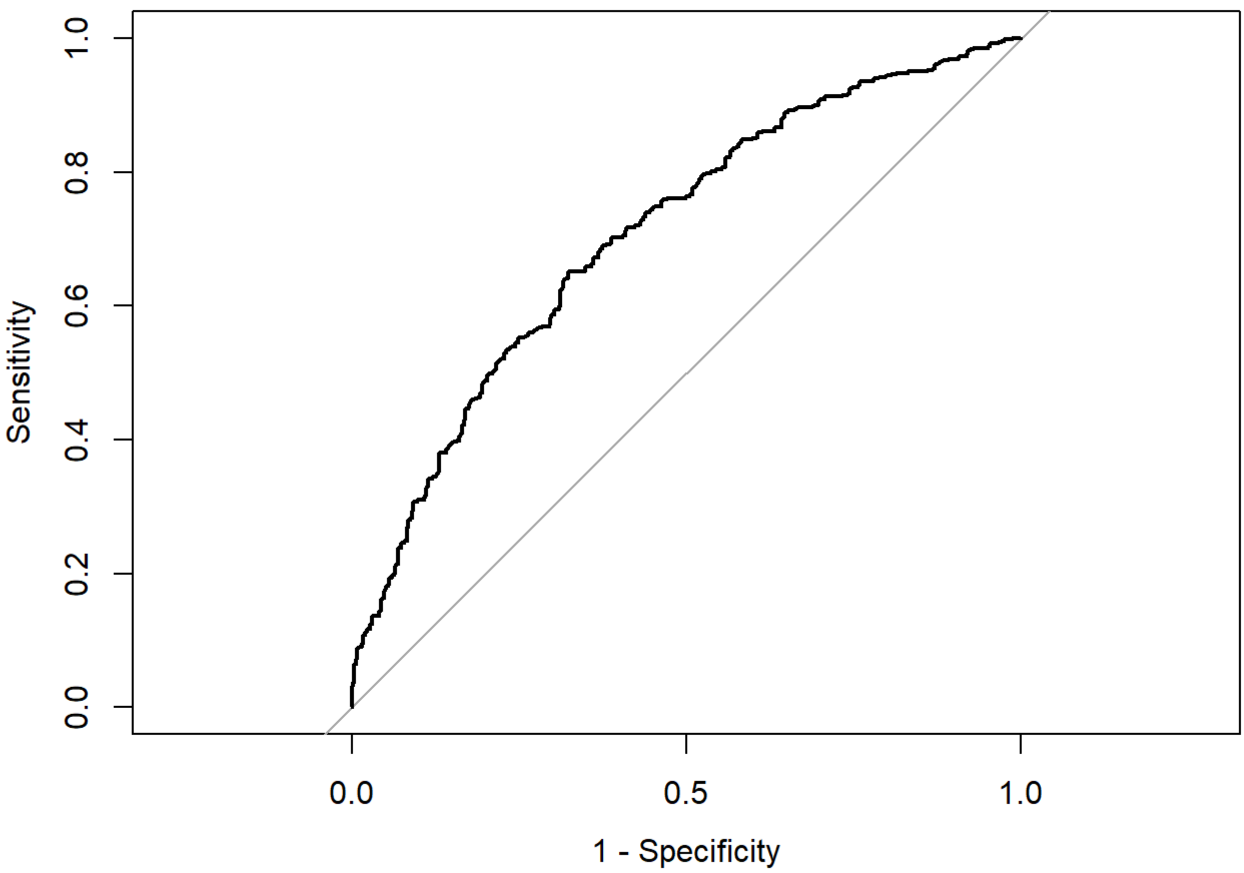

```
auc(rocplot1)
```

```
## Area under the curve: 0.7132
```

```
### step model
rocplot2 <- roc(Purchased ~ fitted(mod1), data=data1)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
plot.roc(rocplot2, legacy.axes=TRUE)
```



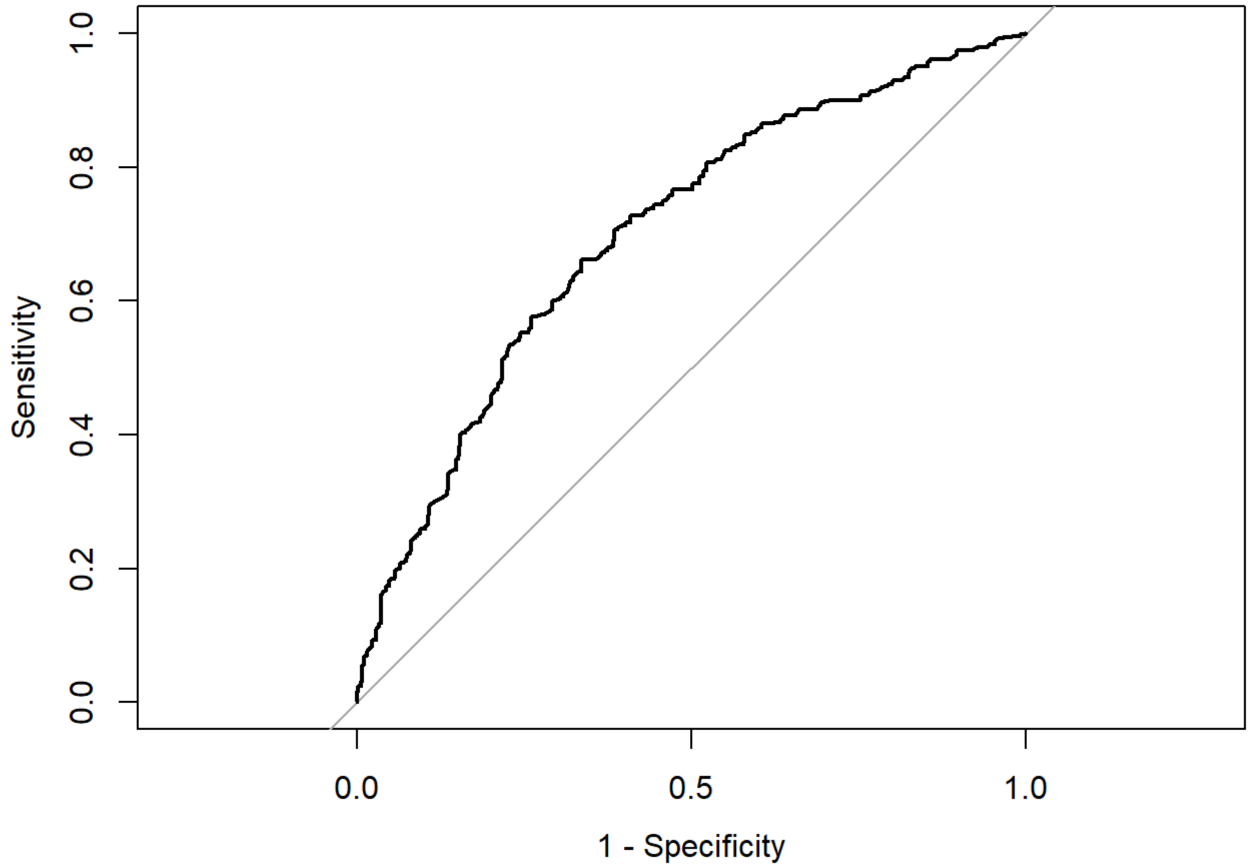

```
auc(rocplot2)
```

```
## Area under the curve: 0.7076
```

```
### select model
rocplot2 <- roc(Purchased ~ fitted(mod2), data=data1)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

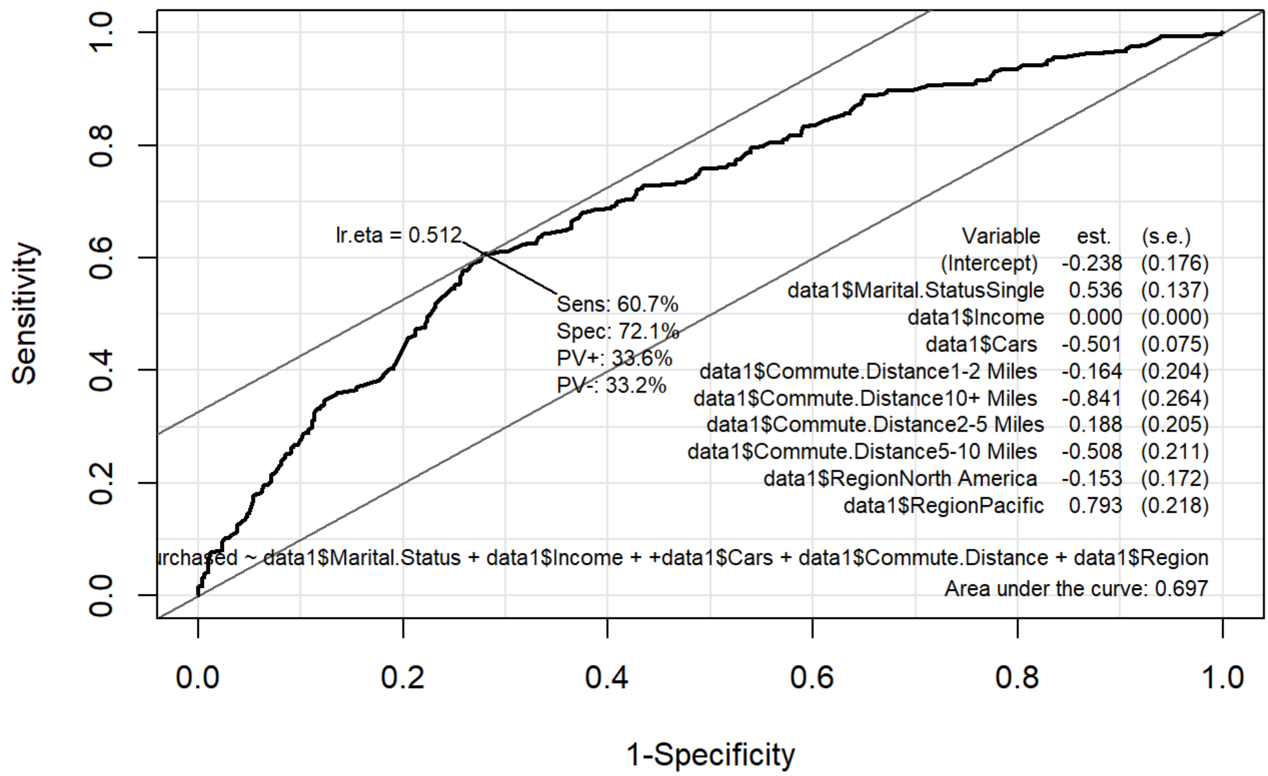
```
plot.roc(rocplot2, legacy.axes=TRUE)
```




```
auc(rocplot2)
```

```
## Area under the curve: 0.7014
```

```
ROC(form=data1$Purchased~data1$Marital.Status+data1$Income+
      +data1$Cars+data1$Commute.Distance+data1$Region,plot="ROC")
```



From the roc curve graph, Sensitivity = 60.7% means that the ratio of observed number of people who purchased a bike to predicted number of people who purchased a bike is 0.607. Specificity = 72.1% means that the ratio of the observed number of people who have not purchased a bike to the predicted number of people who have not purchased a bike is 0.721. AUC=0.697 means that The applicability of the model to the current data is 69.7%

```
#### correlation

cor(data1$Purchased, fitted(modf))
```

```
## [1] 0.3707286
```

```
cor(data1$Purchased, fitted(mod1))
```

```
## [1] 0.3605808
```

```
cor(data1$Purchased, fitted(mod2))
```

```
## [1] 0.3488597
```

From the perspective of correlation, the correlations of full model, step model and select model are all lower than 4.0, which indicates that the positive correlation between variables is relatively weak, and the correlation of select model is the lowest among the three models.

From the results, the select model is more in line with the data, and the accuracy of the model is medium. For this data, we can know that for people to buy bicycles, marital status, whether they have a car, communication distance and region are the main factors affecting them, among which marital status is a positive influence, whether they have a car and communication distance are negative influences.