A Quick Look into Bike Buyers Dataset

Florencia

2023-09-24

```
#import csv
bike_buyers <- read.csv("E:/Bike Buyers/bike_buyers.csv")
```

1.Basic data characteristics

```
dim(bike_buyers)
```

```
## [1] 1000   13
```

EXPLANATION

The dim function returns the dimension of the bike_buyers dataset. it shows that the bike_buyers dataset has 1000 rows and 13 columns (1000 instances and 13 attributes)

```
str(bike_buyers)
```

```
## 'data.frame':    1000 obs. of  13 variables:
##  $ ï..ID           : int  12496 24107 14177 24381 25597 13507 27974 19364 22155 19280 ...
##  $ Marital.Status  : chr  "Married" "Married" "Married" "Single" ...
##  $ Gender          : chr  "Female" "Male" "Male" "" ...
##  $ Income          : int  40000 30000 80000 70000 30000 10000 160000 40000 20000 NA ...
##  $ Children        : int  1 3 5 0 0 2 2 1 2 2 ...
##  $ Education       : chr  "Bachelors" "Partial College" "Partial College" "Bachelors" ...
##  $ Occupation      : chr  "Skilled Manual" "Clerical" "Professional" "Professional" ...
##  $ Home.Owner      : chr  "Yes" "Yes" "No" "Yes" ...
##  $ Cars            : int  0 1 2 1 0 0 4 0 2 1 ...
##  $ Commute.Distance: chr  "0-1 Miles" "0-1 Miles" "2-5 Miles" "5-10 Miles" ...
##  $ Region          : chr  "Europe" "Europe" "Europe" "Pacific" ...
##  $ Age             : int  42 43 60 41 36 50 33 43 58 NA ...
##  $ Purchased.Bike  : chr  "No" "No" "No" "Yes" ...
```

```
writeLines("\n")
```

```
sapply(bike_buyers, class)
```

```
##           ï..ID   Marital.Status           Gender           Income
##       "integer"      "character"      "character"        "integer"
##        Children        Education       Occupation       Home.Owner
##       "integer"      "character"      "character"      "character"
##            Cars Commute.Distance           Region              Age
##       "integer"      "character"      "character"        "integer"
##   Purchased.Bike
##     "character"
```

EXPLANATION

there are two different data types in the bike_buyers dataset which are integer (ID, Income, Children, Cars, Age) and character (Marital.Status, Gender, Education, Occupation, Home.Owner, Commute Distance, Region, Purchased.Bike)

```
BasicSummary <- function(df, dgts = 3){
  m <- ncol(df)
varNames <- colnames(df)
varType <- vector("character",m)
topLevel <- vector("character",m)
topCount <- vector("numeric",m)
missCount <- vector("numeric",m)
levels <- vector("numeric", m)

for (i in 1:m){
x <- df[,i]
varType[i] <- class(x)
xtab <- table(x, useNA = "ifany")
levels[i] <- length(xtab)
nums <- as.numeric(xtab)
maxnum <- max(nums)
topCount[i] <- maxnum
maxIndex <- which.max(nums)
lvls <- names(xtab)
topLevel[i] <- lvls[maxIndex]
missIndex <- which((is.na(x)) | (x == "") | (x == " "))
missCount[i] <- length(missIndex)
}
n <- nrow(df)
topFrac <- round(topCount/n, digits = dgts)
missFrac <- round(missCount/n, digits = dgts)
## #
summaryFrame <- data.frame(variable = varNames, type = varType,
 levels = levels, topLevel = topLevel,
 topCount = topCount, topFrac = topFrac,
 missFreq = missCount, missFrac = missFrac)
 return(summaryFrame)
 }

BasicSummary(bike_buyers)
```

```
##            variable      type levels      topLevel topCount topFrac missFreq
## 1             ï..ID   integer   1000         11000        1   0.001        0
## 2    Marital.Status character      3       Married      535   0.535        7
## 3            Gender character      3          Male      500   0.500       11
## 4            Income   integer     17         60000      165   0.165        6
## 5          Children   integer      7             0      274   0.274        8
## 6         Education character      5     Bachelors      306   0.306        0
## 7        Occupation character      5  Professional      276   0.276        0
## 8        Home.Owner character      3           Yes      682   0.682        4
## 9              Cars   integer      6             2      342   0.342        9
## 10 Commute.Distance character      5     0-1 Miles      366   0.366        0
## 11           Region character      3 North America      508   0.508        0
## 12              Age   integer     54            40       40   0.040        8
## 13   Purchased.Bike character      2            No      519   0.519        0
##    missFrac
## 1     0.000
## 2     0.007
## 3     0.011
## 4     0.006
## 5     0.008
## 6     0.000
## 7     0.000
## 8     0.004
## 9     0.009
## 10    0.000
## 11    0.000
## 12    0.008
## 13    0.000
```

EXPLANATION

It is clear that all the variables have clear and simple explanatory names which are not difficult to understand and it describes the data in the dataset. From the 7 variables in the dataset, 5 of them were integer, and the rest is character. It can be seen that the integer variables has more levels than the character variables.

2. Summary Statistics

```
summary(bike_buyers)
```

```
##      ï..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   : 56268
##  3rd Qu.:24471                                         3rd Qu.: 70000
##  Max.   :29447                                         Max.   :170000
##                                                        NA's   :6
##     Children     Education          Occupation         Home.Owner
##  Min.   :0.00   Length:1000        Length:1000        Length:1000
##  1st Qu.:0.00   Class :character   Class :character   Class :character
##  Median :2.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1.91
##  3rd Qu.:3.00
##  Max.   :5.00
##  NA's   :8
##       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   :1.455                                         Mean   :44.18
##  3rd Qu.:2.000                                         3rd Qu.:52.00
##  Max.   :4.000                                         Max.   :89.00
##  NA's   :9                                             NA's   :8
##  Purchased.Bike
##  Length:1000
##  Class :character
##  Mode  :character
##
##
##
##
```

```
writeLines("Mean:")
```

```
## Mean:
```

```
sapply(bike_buyers[, c(4,12)], mean, na.rm=TRUE)
```

```
##      Income         Age
## 56267.60563    44.18145
```

```
writeLines("\nDescription:")
```

```
##
## Description:
```

```
sapply(bike_buyers[, c(4,12)], quantile, na.rm=TRUE)
```

```
##      Income Age
## 0%    10000  25
## 25%   30000  35
## 50%   60000  43
## 75%   70000  52
## 100% 170000  89
```

```
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 4.1.3
```

```
## Loading required package: lattice
```

```
## Warning: package 'lattice' was built under R version 4.1.3
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
describe(bike_buyers)
```

```
## bike_buyers
##
##  13  Variables      1000  Observations
## ---------------------------------------------------------------------------------
## ï..ID
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0     1000        1    19966     6176    11781    12627
##      .25      .50      .75      .90      .95
##    15291    19744    24471    27544    28413
##
## lowest : 11000 11047 11061 11090 11116, highest: 29337 29355 29380 29424 29447
## ---------------------------------------------------------------------------------
## Marital.Status
##        n  missing distinct
##      993        7        2
##
## Value      Married  Single
## Frequency      535     458
## Proportion   0.539   0.461
## ---------------------------------------------------------------------------------
## Gender
##        n  missing distinct
##      989       11        2
##
## Value      Female   Male
## Frequency     489    500
## Proportion  0.494  0.506
## ---------------------------------------------------------------------------------
## Income
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      994        6       16    0.986    56268    34273    10000    20000
##      .25      .50      .75      .90      .95
##    30000    60000    70000   100000   120000
##
## lowest :  10000  20000  30000  40000  50000, highest: 120000 130000 150000 160000 170000
##
## Value       10000  20000  30000  40000  50000  60000  70000  80000  90000
## Frequency      73     74    134    153     40    165    123     90     38
## Proportion  0.073  0.074  0.135  0.154  0.040  0.166  0.124  0.091  0.038
##
## Value      100000 110000 120000 130000 150000 160000 170000
## Frequency      29     16     17     32      4      3      3
## Proportion  0.029  0.016  0.017  0.032  0.004  0.003  0.003
## ---------------------------------------------------------------------------------
## Children
##        n  missing distinct     Info     Mean      Gmd
##      992        8        6     0.96     1.91    1.827
##
## lowest : 0 1 2 3 4, highest: 1 2 3 4 5
##
## Value          0     1     2     3     4     5
## Frequency    274   169   209   133   126    81
## Proportion 0.276 0.170 0.211 0.134 0.127 0.082
## ---------------------------------------------------------------------------------
## Education
```

```
##        n  missing distinct 
##     1000        0        5 
## 
## lowest : Bachelors           Graduate Degree     High School         Partial College     P
artial High School
## highest: Bachelors           Graduate Degree     High School         Partial College     P
artial High School
## 
## Value                Bachelors     Graduate Degree         High School
## Frequency                  306                 174                 179
## Proportion               0.306               0.174               0.179
## 
## Value          Partial College Partial High School
## Frequency                  265                  76
## Proportion               0.265               0.076
## --------------------------------------------------------------------------------
## Occupation 
##        n  missing distinct 
##     1000        0        5 
## 
## lowest : Clerical       Management     Manual         Professional   Skilled Manual
## highest: Clerical       Management     Manual         Professional   Skilled Manual
## 
## Value            Clerical     Management         Manual   Professional
## Frequency             177            173            119            276
## Proportion          0.177          0.173          0.119          0.276
## 
## Value      Skilled Manual
## Frequency             255
## Proportion          0.255
## --------------------------------------------------------------------------------
## Home.Owner 
##        n  missing distinct 
##      996        4        2 
## 
## Value         No   Yes
## Frequency    314   682
## Proportion 0.315 0.685
## --------------------------------------------------------------------------------
## Cars 
##        n  missing distinct     Info     Mean      Gmd 
##      991        9        5    0.925    1.455    1.226 
## 
## lowest : 0 1 2 3 4, highest: 0 1 2 3 4
## 
## Value          0     1     2     3     4
## Frequency    238   267   342    85    59
## Proportion 0.240 0.269 0.345 0.086 0.060
## --------------------------------------------------------------------------------
## Commute.Distance 
##        n  missing distinct 
##     1000        0        5 
## 
## lowest : 0-1 Miles  1-2 Miles  10+ Miles  2-5 Miles  5-10 Miles
## highest: 0-1 Miles  1-2 Miles  10+ Miles  2-5 Miles  5-10 Miles
## 
```

```
## Value       0-1 Miles  1-2 Miles  10+ Miles  2-5 Miles 5-10 Miles
## Frequency         366        169        111        162        192
## Proportion      0.366      0.169      0.111      0.162      0.192
## ---------------------------------------------------------------------------
## Region
##        n  missing distinct
##     1000        0        3
##
## Value             Europe North America       Pacific
## Frequency            300           508           192
## Proportion         0.300         0.508         0.192
## ---------------------------------------------------------------------------
## Age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      992        8       53    0.999    44.18    12.85    28.00    30.00
##      .25      .50      .75      .90      .95
##    35.00    43.00    52.00    60.90    65.45
##
## lowest : 25 26 27 28 29, highest: 73 74 78 80 89
## ---------------------------------------------------------------------------
## Purchased.Bike
##        n  missing distinct
##     1000        0        2
##
## Value         No   Yes
## Frequency    519   481
## Proportion 0.519 0.481
## ---------------------------------------------------------------------------
```

EXPLANATION

Missing values were found in the bike_buyers dataset, 7 missing values in Marital.Status variable, 11 in Gender variable, 6 in Income variable, 8 missing values iin CHildren variable, 4 in Home.Owner variable, 9 in Cars variables, and 8 missing values in the Age variable.

```
bike_buyers[, c(2,3,6:8, 10, 11, 13)] <- lapply(bike_buyers[, c(2,3,6:8, 10, 11, 13)], as.fac
tor)
```

3. Data anomalies

```
ThreeSigma <- function(x, t = 3){

 mu <- mean(x, na.rm = TRUE)
 sig <- sd(x, na.rm = TRUE)
 if (sig == 0){
 message("All non-missing x-values are identical")
}
 up <- mu + t * sig
 down <- mu - t * sig
 out <- list(up = up, down = down)
 return(out)
 }

Hampel <- function(x, t = 3){

 mu <- median(x, na.rm = TRUE)
 sig <- mad(x, na.rm = TRUE)
 if (sig == 0){
 message("Hampel identifer implosion: MAD scale estimate is zero")
 }
 up <- mu + t * sig
 down <- mu - t * sig
 out <- list(up = up, down = down)
 return(out)
 }

BoxplotRule<- function(x, t = 1.5){

 xL <- quantile(x, na.rm = TRUE, probs = 0.25, names = FALSE)
 xU <- quantile(x, na.rm = TRUE, probs = 0.75, names = FALSE)
 Q <- xU - xL
 if (Q == 0){message("Boxplot rule implosion: interquartile distance is zero")
 }
 up <- xU + t * Q
 down <- xU - t * Q
 out <- list(up = up, down = down)
 return(out)
}

ExtractDetails <- function(x, down, up){

 outClass <- rep("N", length(x))
 indexLo <- which(x < down)
 indexHi <- which(x > up)
 outClass[indexLo] <- "L"
 outClass[indexHi] <- "U"
 index <- union(indexLo, indexHi)
 values <- x[index]
 outClass <- outClass[index]
 nOut <- length(index)
 maxNom <- max(x[which(x <= up)])
 minNom <- min(x[which(x >= down)])
 outList <- list(nOut = nOut, lowLim = down,
```

```
 upLim = up, minNom = minNom,
 maxNom = maxNom, index = index,
 values = values,
 outClass = outClass)
 return(outList)
 }
```

```
FindOutliers <- function(x, t3 = 3, tH = 3, tb = 1.5){
 threeLims <- ThreeSigma(x, t = t3)
 HampLims <- Hampel(x, t = tH)
 boxLims <- BoxplotRule(x, t = tb)

 n <- length(x)
 nMiss <- length(which(is.na(x)))

 threeList <- ExtractDetails(x, threeLims$down, threeLims$up)
 HampList <- ExtractDetails(x, HampLims$down, HampLims$up)
 boxList <- ExtractDetails(x, boxLims$down, boxLims$up)

 sumFrame <- data.frame(method = "ThreeSigma", n = n,
 nMiss = nMiss, nOut = threeList$nOut,
 lowLim = threeList$lowLim,
 upLim = threeList$upLim,
 minNom = threeList$minNom,
 maxNom = threeList$maxNom)
 upFrame <- data.frame(method = "Hampel", n = n,
 nMiss = nMiss, nOut = HampList$nOut,
 lowLim = HampList$lowLim,
 upLim = HampList$upLim,
 minNom = HampList$minNom,
 maxNom = HampList$maxNom)
 sumFrame <- rbind.data.frame(sumFrame, upFrame)
 upFrame <- data.frame(method = "BoxplotRule", n = n,
 nMiss = nMiss, nOut = boxList$nOut,
 lowLim = boxList$lowLim,
 upLim = boxList$upLim,
 minNom = boxList$minNom,
 maxNom = boxList$maxNom)
 sumFrame <- rbind.data.frame(sumFrame, upFrame)

 threeFrame <- data.frame(index = threeList$index,
 values = threeList$values,
 type = threeList$outClass)
 HampFrame <- data.frame(index = HampList$index,
 values = HampList$values,
 type = HampList$outClass)
 boxFrame <- data.frame(index = boxList$index,
 values = boxList$values,
 type = boxList$outClass)
 outList <- list(summary = sumFrame, threeSigma = threeFrame,
 Hampel = HampFrame, boxplotRule = boxFrame)
 return(outList)
}
```

```
FullSummary <- FindOutliers(bike_buyers$Income)
FullSummary$summary
```

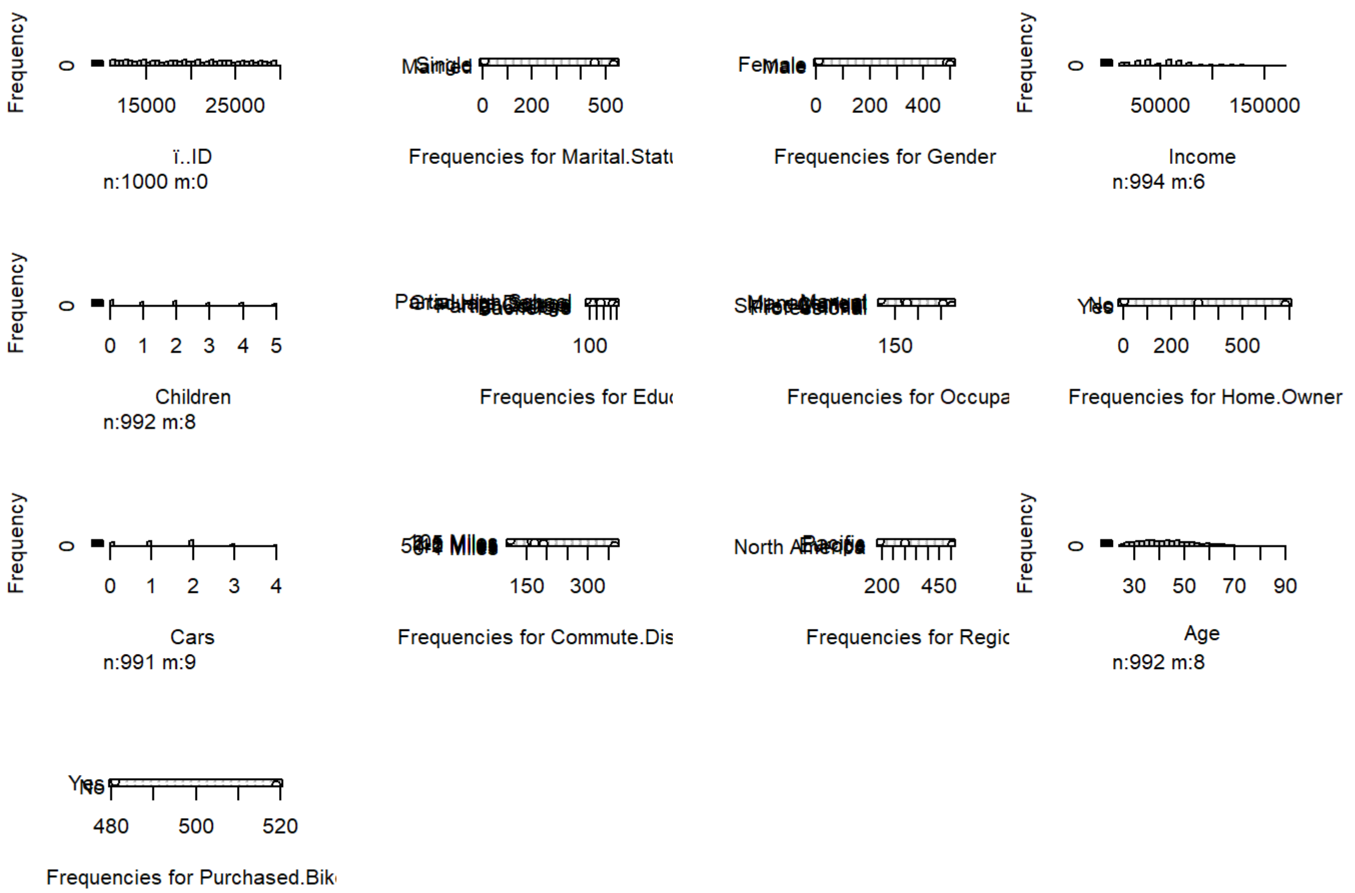
```
##        method    n nMiss nOut    lowLim    upLim minNom maxNom
## 1  ThreeSigma 1000     6   10 -36935.85 149471.1  10000 130000
## 2      Hampel 1000     6   10 -28956.00 148956.0  10000 130000
## 3 BoxplotRule 1000     6   10  10000.00 130000.0  10000 130000
```

EXPLANATION From these three method of finding the outliers, three of them detect the same amount of the outliers which is 10 outliers. For the upper and lower limit, the BoxplotRule has the lowest upper and lower outlier limit among the three of them, but it does’nt give that big/ much difference The lower and upper limits of the non-outlying data values of the three rule has the same value

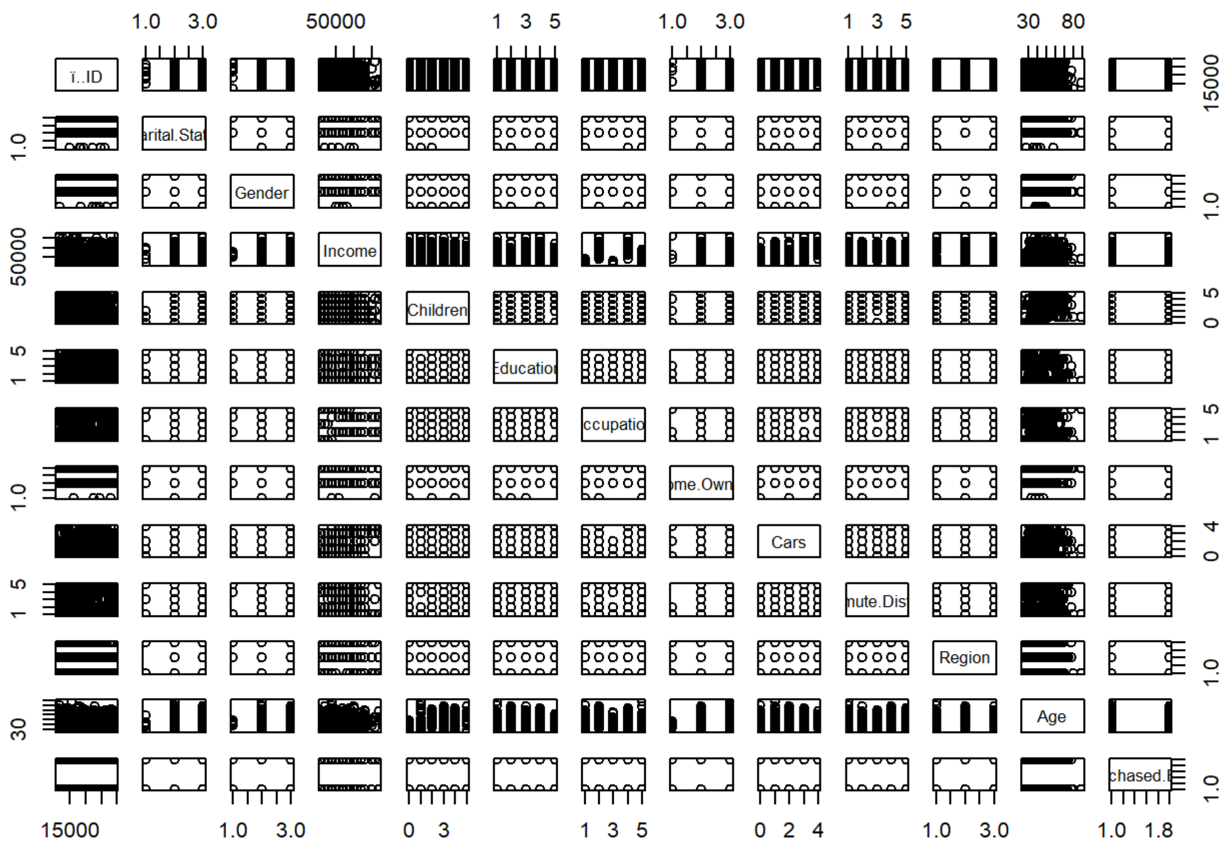
```
rcorr(as.matrix(bike_buyers[c(1,4,5, 9, 12)]), type = "spearman")
```

```
##          ï..ID Income Children Cars   Age
## ï..ID     1.00  -0.06    -0.02 0.03 -0.05
## Income   -0.06   1.00     0.29 0.33  0.20
## Children -0.02   0.29     1.00 0.28  0.60
## Cars      0.03   0.33     0.28 1.00  0.22
## Age      -0.05   0.20     0.60 0.22  1.00
##
## n
##          ï..ID Income Children Cars Age
## ï..ID     1000    994      992  991 992
## Income     994    994      986  985 987
## Children   992    986      992  983 985
## Cars       991    985      983  991 983
## Age        992    987      985  983 992
##
## P
##          ï..ID  Income Children Cars   Age
## ï..ID           0.0593 0.5023   0.3593 0.1239
## Income   0.0593        0.0000   0.0000 0.0000
## Children 0.5023 0.0000          0.0000 0.0000
## Cars     0.3593 0.0000 0.0000          0.0000
## Age      0.1239 0.0000 0.0000   0.0000
```

```
hist.data.frame(bike_buyers)
```




```
plot(bike_buyers)
```



```
Table <- table(bike_buyers$Income, bike_buyers$Purchased.Bike, bike_buyers$Gender)
print(Table)
```

```
## , ,  = 
## 
##          
##          No Yes
##   10000   0   0
##   20000   0   0
##   30000   0   0
##   40000   0   0
##   50000   0   1
##   60000   2   1
##   70000   2   1
##   80000   4   0
##   90000   0   0
##   100000  0   0
##   110000  0   0
##   120000  0   0
##   130000  0   0
##   150000  0   0
##   160000  0   0
##   170000  0   0
## 
## , ,  = Female
## 
##          
##          No Yes
##   10000  25  17
##   20000  20  20
##   30000  41  26
##   40000  33  41
##   50000   9   9
##   60000  34  41
##   70000  29  34
##   80000  26  18
##   90000   9  10
##   100000  9   3
##   110000  4   2
##   120000  2   5
##   130000  9   8
##   150000  0   1
##   160000  0   1
##   170000  0   1
## 
## , ,  = Male
## 
##          
##          No Yes
##   10000  20  11
##   20000  23  11
##   30000  40  27
##   40000  31  48
##   50000  11  10
##   60000  48  39
##   70000  27  30
##   80000  26  16
##   90000   5  14
```

```
##   100000  9   8
##   110000  4   6
##   120000  6   4
##   130000  8   7
##   150000  1   2
##   160000  0   2
##   170000  2   0
```

```
Table <- table(bike_buyers$Age, bike_buyers$Marital.Status)
print(Table)
```

```
##
##          Married Single
##   25  0       2      3
##   26  0       5     11
##   27  0      10     13
##   28  1       7     14
##   29  0       5     11
##   30  0       8     18
##   31  0       5     20
##   32  0      18     15
##   33  0       8     13
##   34  0      16     15
##   35  1      16     18
##   36  0      16     21
##   37  0      15     17
##   38  0      13     24
##   39  1       8     13
##   40  1      23     16
##   41  0      14     14
##   42  0      18     16
##   43  1      19     16
##   44  0      16     11
##   45  0      20     11
##   46  0      19      8
##   47  0      23     16
##   48  0      21      8
##   49  0      15      8
##   50  0      13     10
##   51  0      13      9
##   52  0      13     12
##   53  0      14     10
##   54  0      12      4
##   55  0      14      3
##   56  0      11      5
##   57  0       4      4
##   58  1       7      4
##   59  0      14      6
##   60  0       8      6
##   61  0       7      2
##   62  0       5      8
##   63  0       6      3
##   64  0      10      0
##   65  0       6      3
##   66  0      10      4
##   67  0       5      5
##   68  0       1      2
##   69  0       7      1
##   70  0       4      0
##   71  0       1      0
##   72  0       1      0
##   73  0       2      2
##   74  0       0      1
##   78  0       1      1
##   80  0       1      0
##   89  0       1      0
```

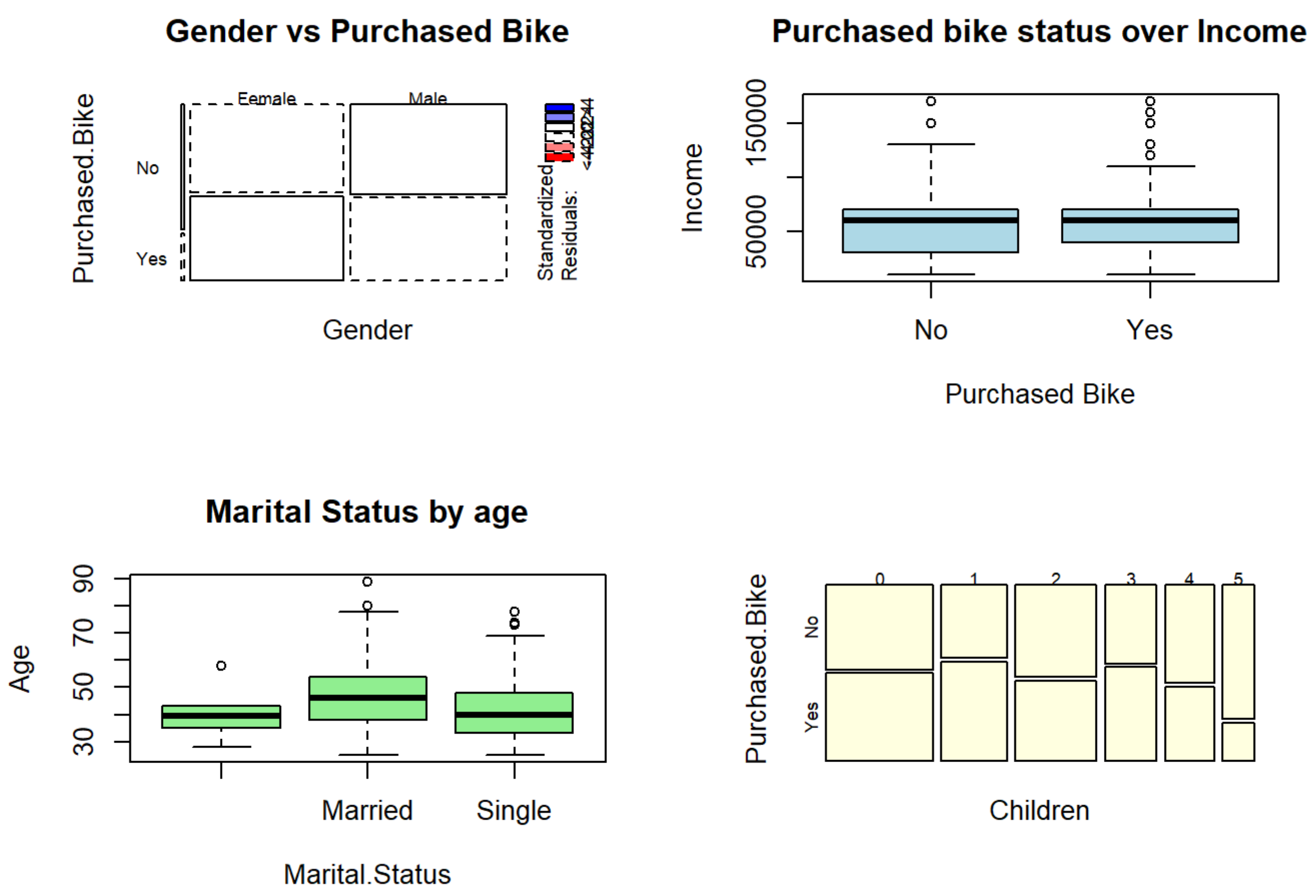
```
matrix <- layout( matrix(c(1,2,3,4), nrow=2, byrow=TRUE) )

mosaicplot(Gender~Purchased.Bike,
           data = bike_buyers,
           main = "Gender vs Purchased Bike",
           col = "pink",
           las=1,
           shade = TRUE)

boxplot(Income~Purchased.Bike,
        data = bike_buyers,
        xlab = "Purchased Bike",
        main = "Purchased bike status over Income",
        col = "lightblue")

boxplot(Age~Marital.Status,
           data = bike_buyers,
        main = "Marital Status by age",
        col = "lightgreen")

mosaicplot(Children~Purchased.Bike,
           data = bike_buyers,
           main = "",
           col = "lightyellow")
```



EXPLANATION

The first plot in the upper left tells us that the amount of bike purchased by the female and male gender has not much difference

In the second plot (the upper right), the income of the buyers doesn’t really affect the purchased bike, so people with higher income will not be guaranteed to buy the bike.

The third plot (lower left), indicates that most people with high age have the status of being married and for the last plot, people with range 0-4 children tend to purchased bike rather than people with 5 children which is the most amount of children in the dataset.