Homework 1

Naga Kartheek Peddisetty, 50538422

02/13/2024

Question 1) Consider the “College” data in the ISLR2 package:

```
library(ISLR2)
library(ggplot2)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ lubridate 1.9.3     ✓ tibble    3.2.1
## ✓ purrr     1.0.2     ✓ tidyr     1.3.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
data(College)
head(College)
```

```
##                              Private Apps Accept Enroll Top10perc Top25perc
## Abilene Christian University     Yes 1660   1232    721        23        52
## Adelphi University               Yes 2186   1924    512        16        29
## Adrian College                   Yes 1428   1097    336        22        50
## Agnes Scott College              Yes  417    349    137        60        89
## Alaska Pacific University        Yes  193    146     55        16        44
## Albertson College                Yes  587    479    158        38        62
##                              F.Undergrad P.Undergrad Outstate Room.Board Books
## Abilene Christian University        2885         537     7440       3300   450
## Adelphi University                  2683        1227    12280       6450   750
## Adrian College                      1036          99    11250       3750   400
## Agnes Scott College                  510          63    12960       5450   450
## Alaska Pacific University            249         869     7560       4120   800
## Albertson College                    678          41    13500       3335   500
##                              Personal PhD Terminal S.F.Ratio perc.alumni Expend
## Abilene Christian University     2200  70       78      18.1          12   7041
## Adelphi University               1500  29       30      12.2          16  10527
## Adrian College                   1165  53       66      12.9          30   8735
## Agnes Scott College               875  92       97       7.7          37  19016
## Alaska Pacific University        1500  76       72      11.9           2  10922
## Albertson College                 675  67       73       9.4          11   9727
##                              Grad.Rate
## Abilene Christian University        60
## Adelphi University                  56
## Adrian College                      54
## Agnes Scott College                 59
## Alaska Pacific University           15
## Albertson College                   55
```

```
dim(College)
```

```
## [1] 777  18
```

```
str(College)
```

```
## 'data.frame':    777 obs. of  18 variables:
##  $ Private    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Apps       : num  1660 2186 1428 417 193 ...
##  $ Accept     : num  1232 1924 1097 349 146 ...
##  $ Enroll     : num  721 512 336 137 55 158 103 489 227 172 ...
##  $ Top10perc  : num  23 16 22 60 16 38 17 37 30 21 ...
##  $ Top25perc  : num  52 29 50 89 44 62 45 68 63 44 ...
##  $ F.Undergrad: num  2885 2683 1036 510 249 ...
##  $ P.Undergrad: num  537 1227 99 63 869 ...
##  $ Outstate   : num  7440 12280 11250 12960 7560 ...
##  $ Room.Board : num  3300 6450 3750 5450 4120 ...
##  $ Books      : num  450 750 400 450 800 500 500 450 300 660 ...
##  $ Personal   : num  2200 1500 1165 875 1500 ...
##  $ PhD        : num  70 29 53 92 76 67 90 89 79 40 ...
##  $ Terminal   : num  78 30 66 97 72 73 93 100 84 41 ...
##  $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
##  $ perc.alumni: num  12 16 30 37 2 11 26 37 23 15 ...
##  $ Expend     : num  7041 10527 8735 19016 10922 ...
##  $ Grad.Rate  : num  60 56 54 59 15 55 63 73 80 52 ...
```

```
colnames(College)
```

```
##  [1] "Private"     "Apps"        "Accept"      "Enroll"      "Top10perc"
##  [6] "Top25perc"   "F.Undergrad" "P.Undergrad" "Outstate"    "Room.Board"
## [11] "Books"       "Personal"    "PhD"         "Terminal"    "S.F.Ratio"
## [16] "perc.alumni" "Expend"      "Grad.Rate"
```

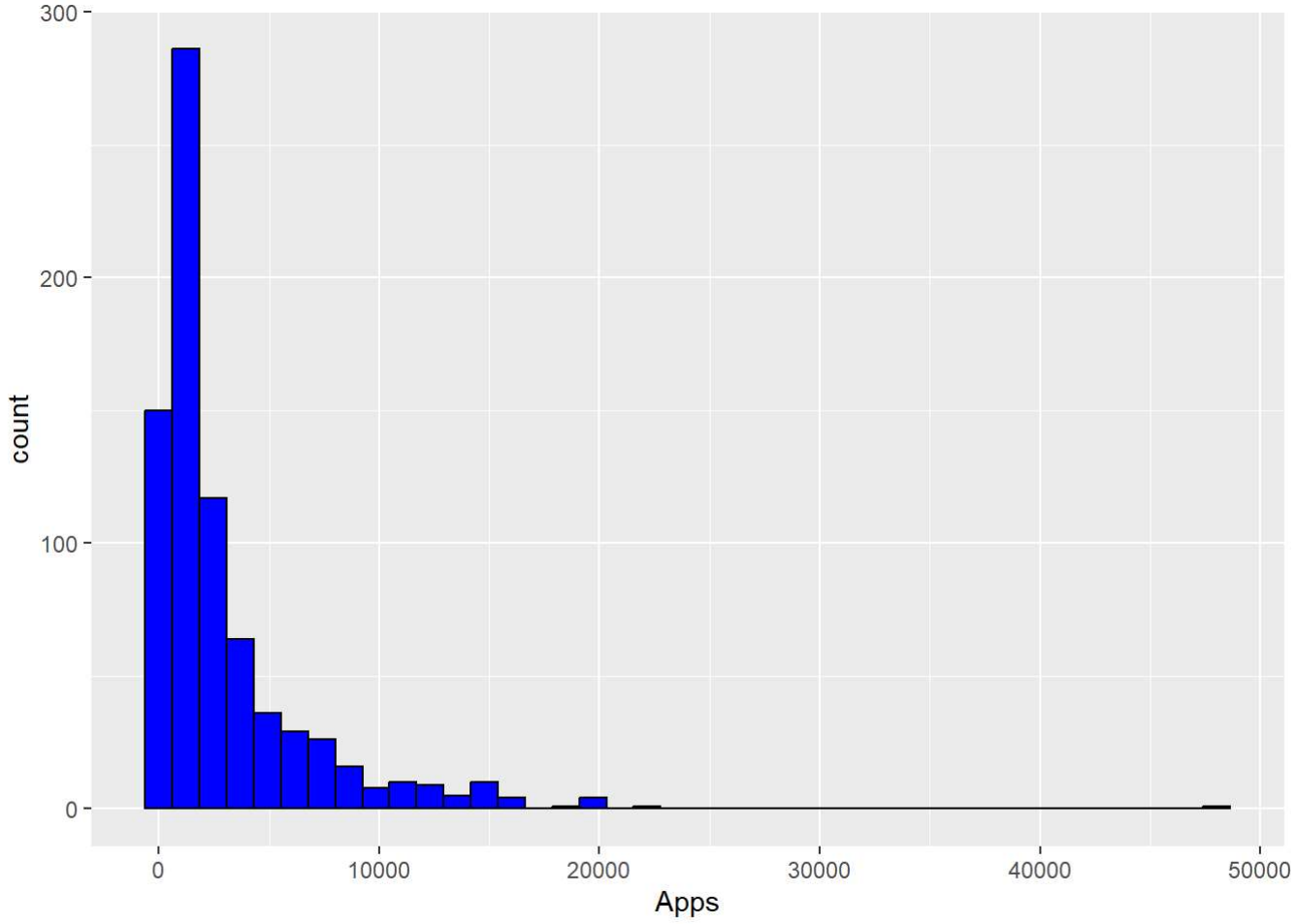
```
colSums(is.na(College))
```

```
##     Private        Apps      Accept      Enroll   Top10perc   Top25perc
##           0           0           0           0           0           0
## F.Undergrad P.Undergrad    Outstate  Room.Board       Books    Personal
##           0           0           0           0           0           0
##         PhD    Terminal   S.F.Ratio perc.alumni      Expend   Grad.Rate
##           0           0           0           0           0           0
```

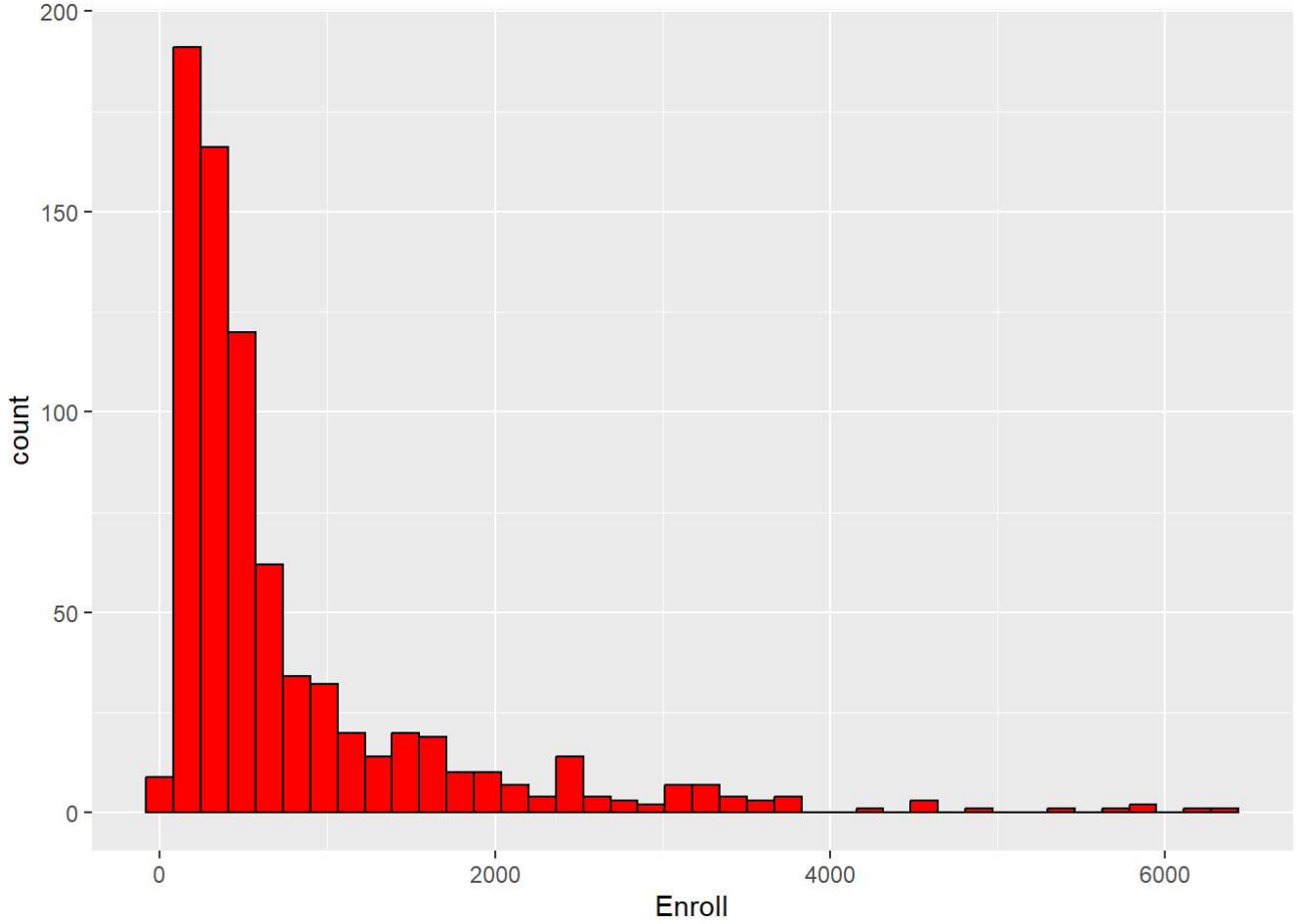
```
college <- College
```

a) Present some visualizations from your exploratory data analysis of this data such as pair plots and histograms? Do you think any scaling or transformation is required?

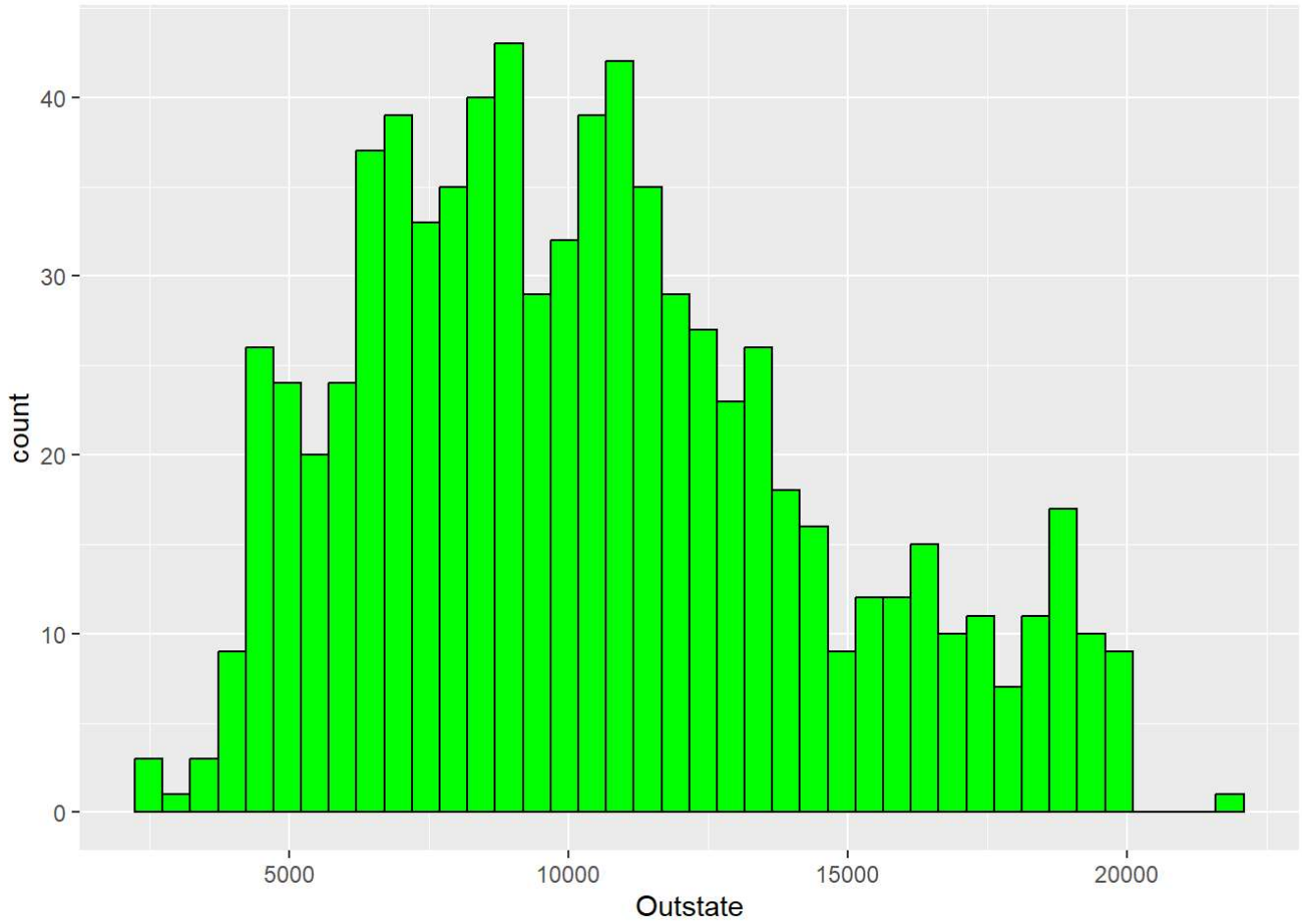
```
### Histograms

ggplot(data = college,aes(x=Apps)) +
  geom_histogram(bins = 40, color = 'black',fill = 'blue')
```



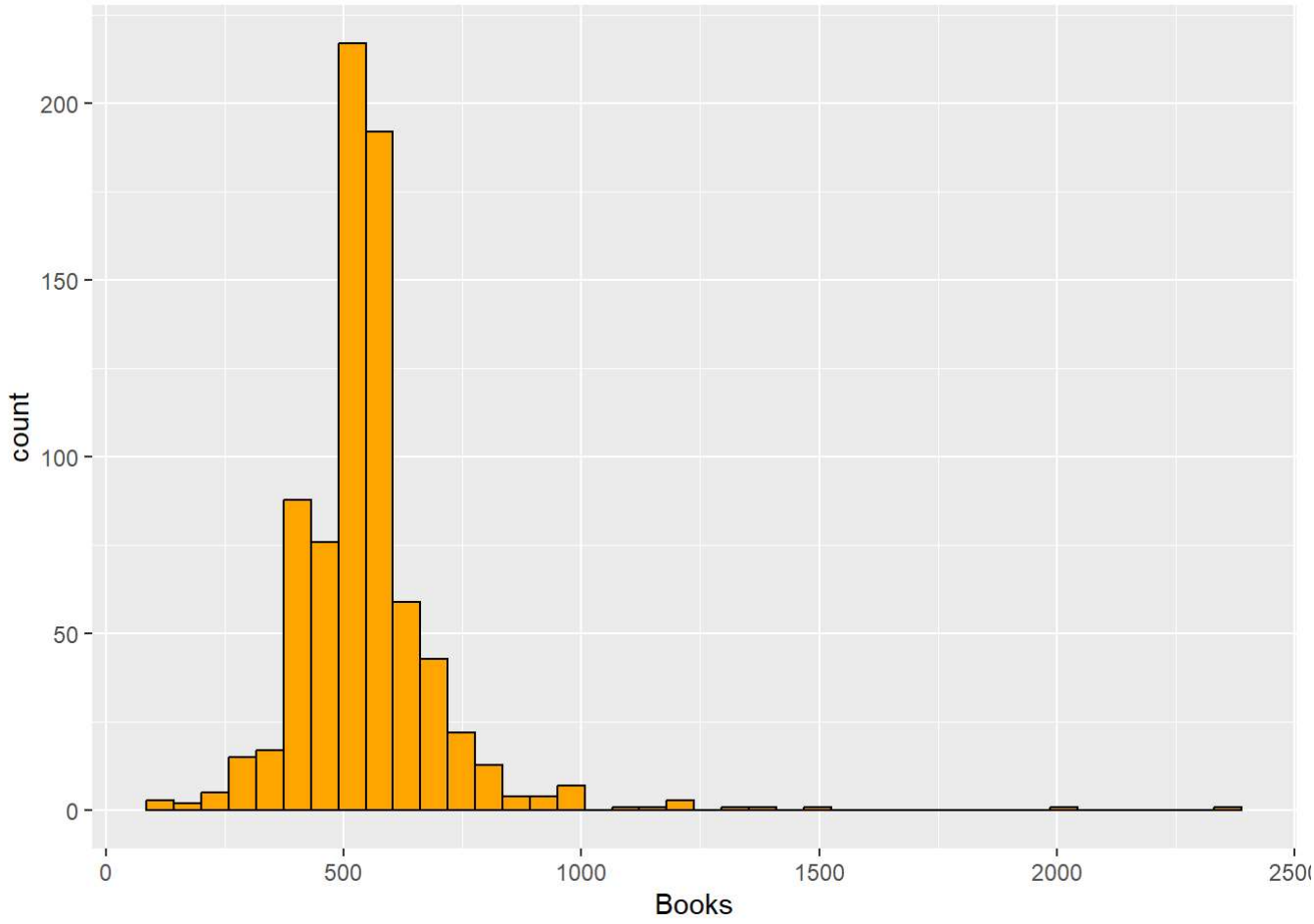

```
ggplot(data = college,aes(x=Enroll)) +
  geom_histogram(bins = 40, color = 'black',fill = 'red')
```



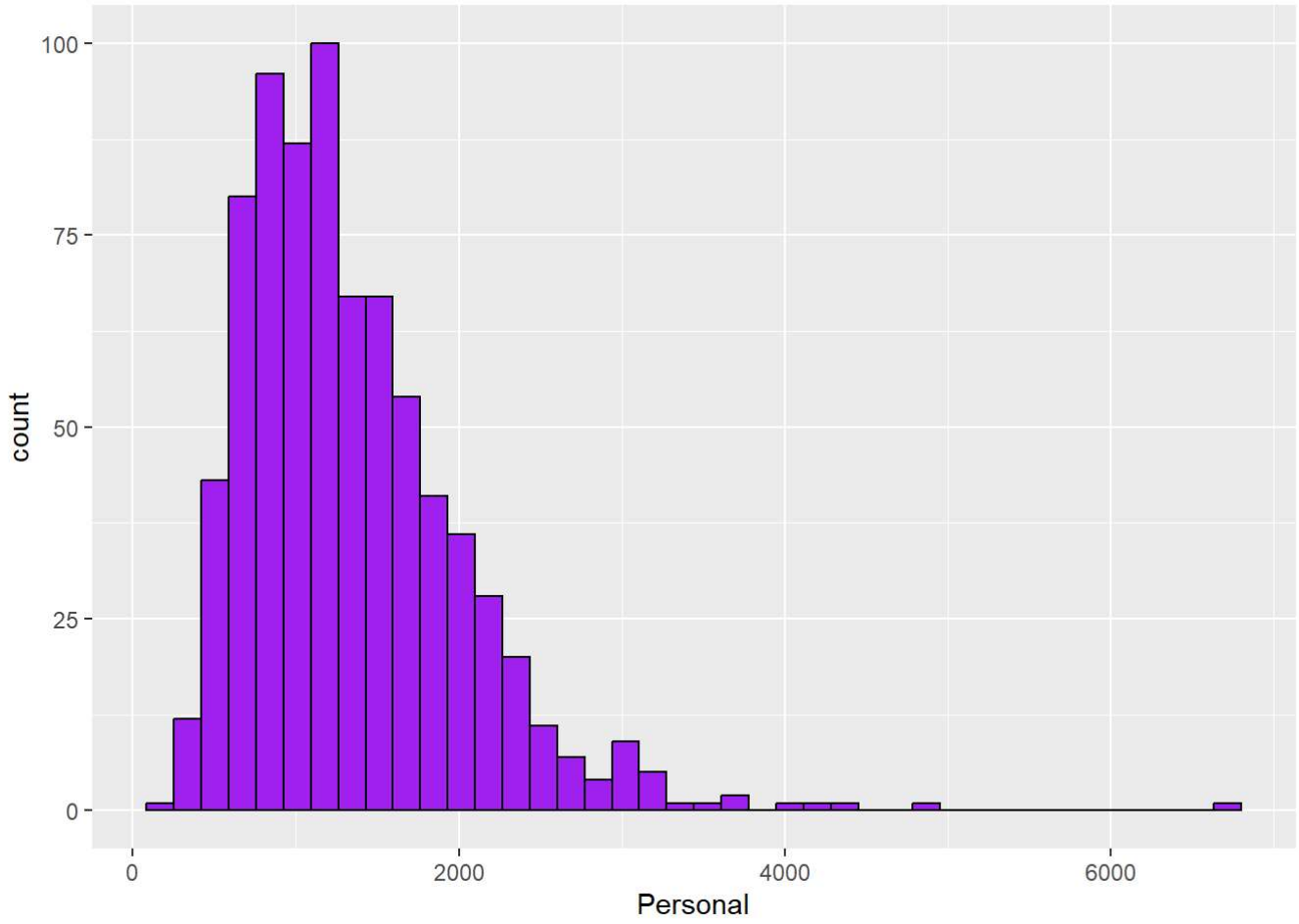

```
ggplot(data = college,aes(x=Outstate)) +
  geom_histogram(bins = 40, color = 'black',fill = 'green')
```



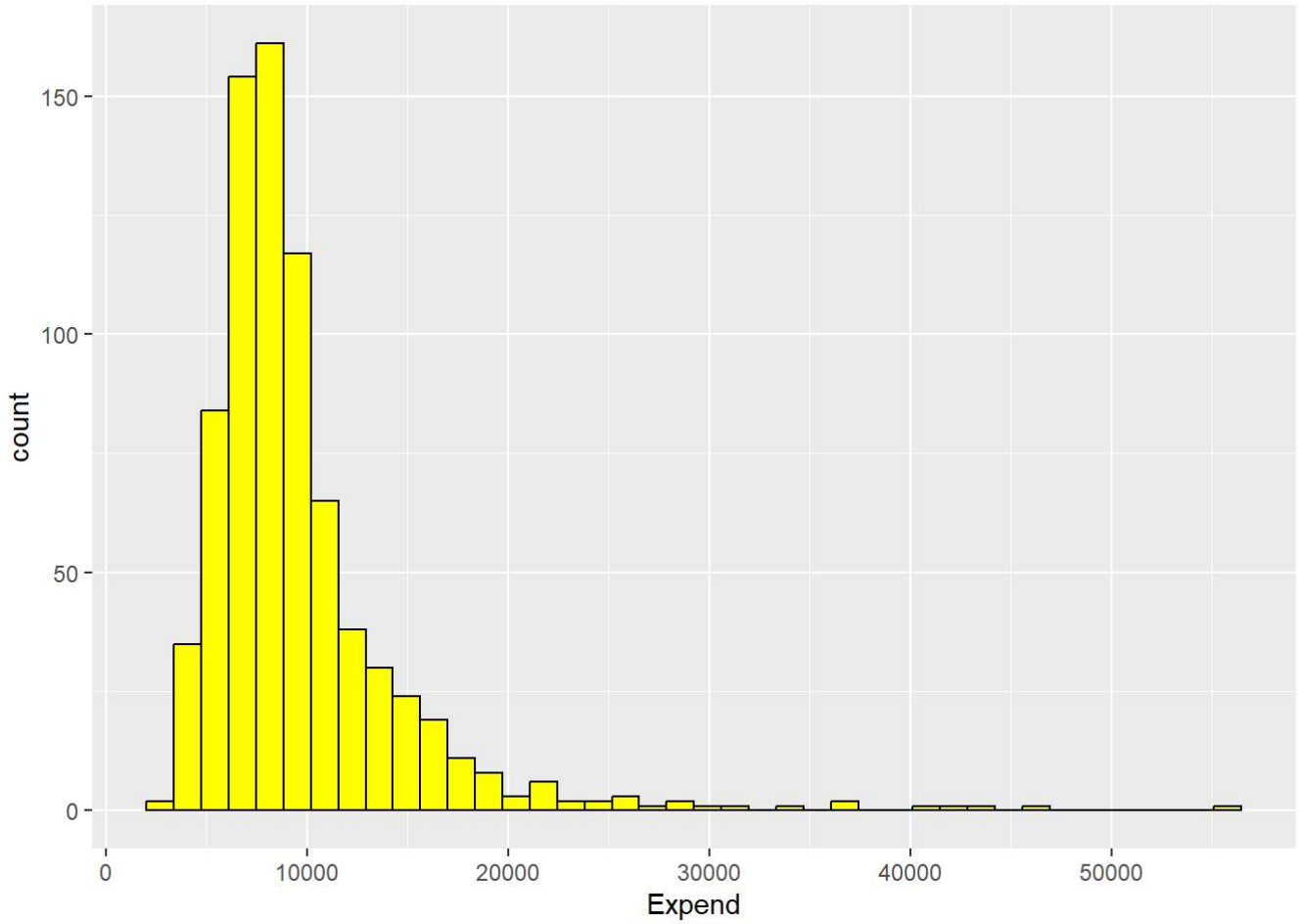

```
ggplot(data = college,aes(x=Books)) +
  geom_histogram(bins = 40, color = 'black',fill = 'orange')
```



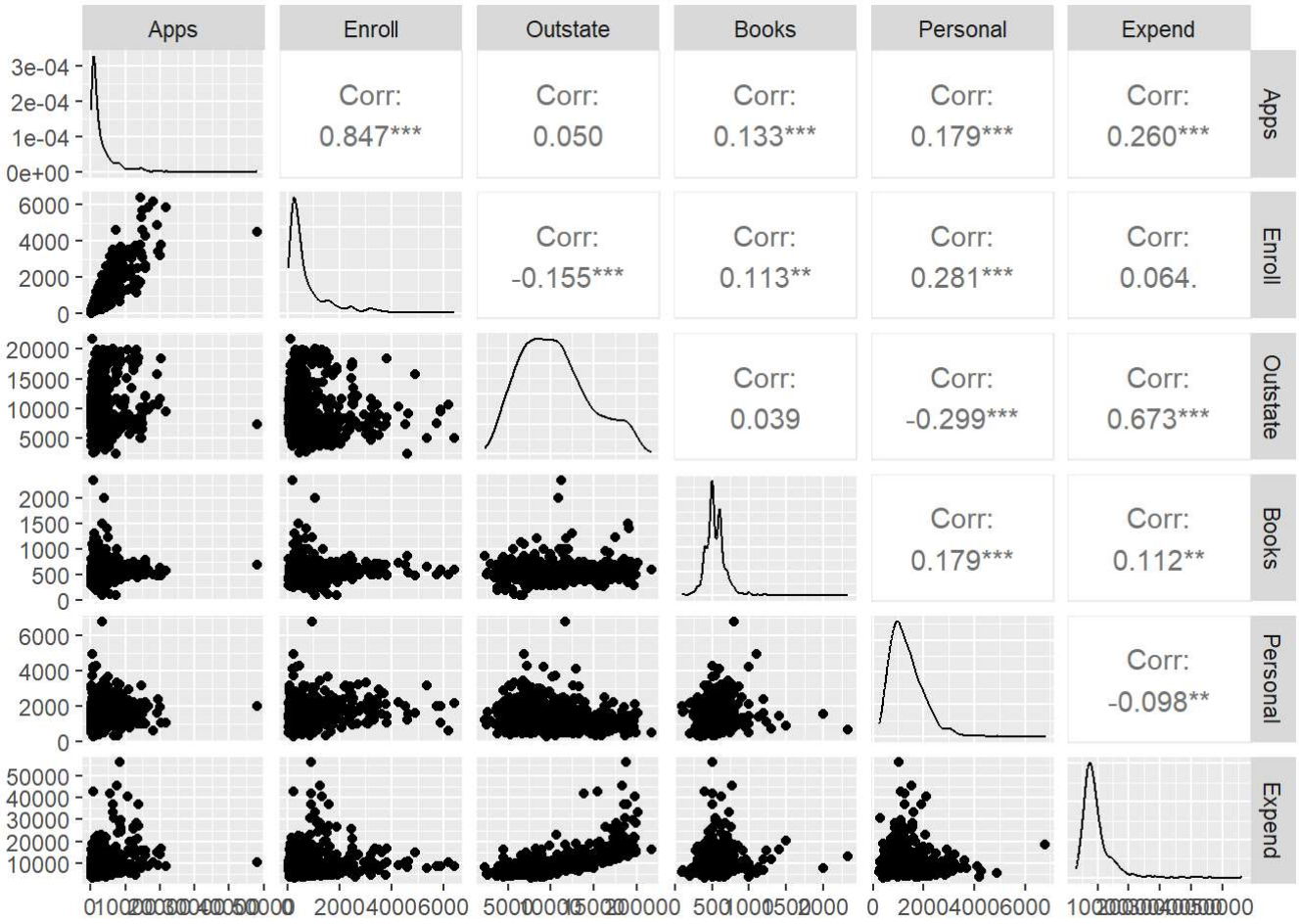

```
ggplot(data = college,aes(x=Personal)) +
  geom_histogram(bins = 40, color = 'black',fill = 'purple')
```




```
ggplot(data = college,aes(x=Expend)) +
  geom_histogram(bins = 40, color = 'black',fill = 'yellow')
```




```
### Creating pair plots of variables

ggpairs(college[ , c("Apps", "Enroll", "Outstate", "Books", "Personal", "Expend" )])
```

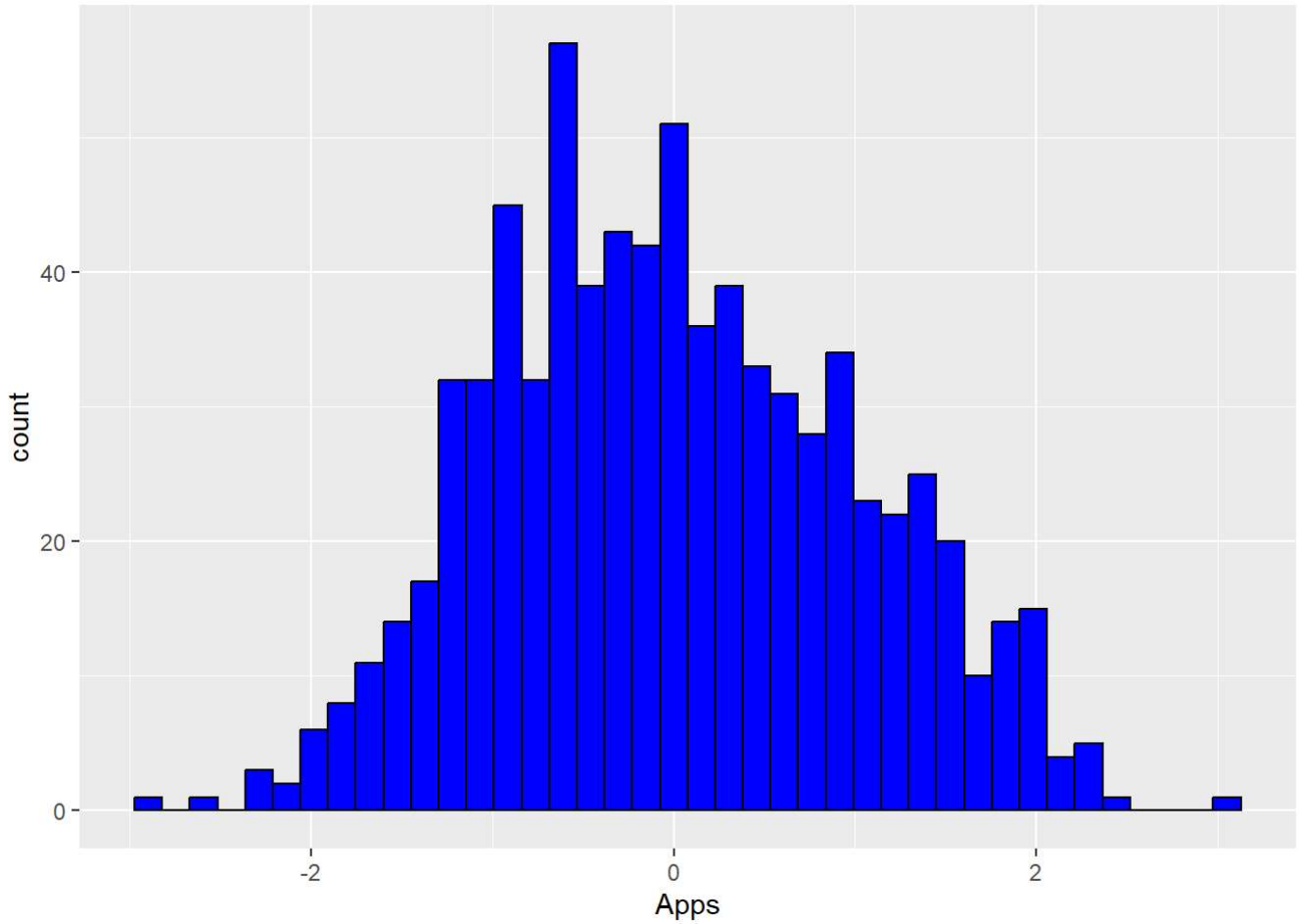


As seen above in the plots range is not uniformly distributed. So, yes the data requires scaling and log transformation.

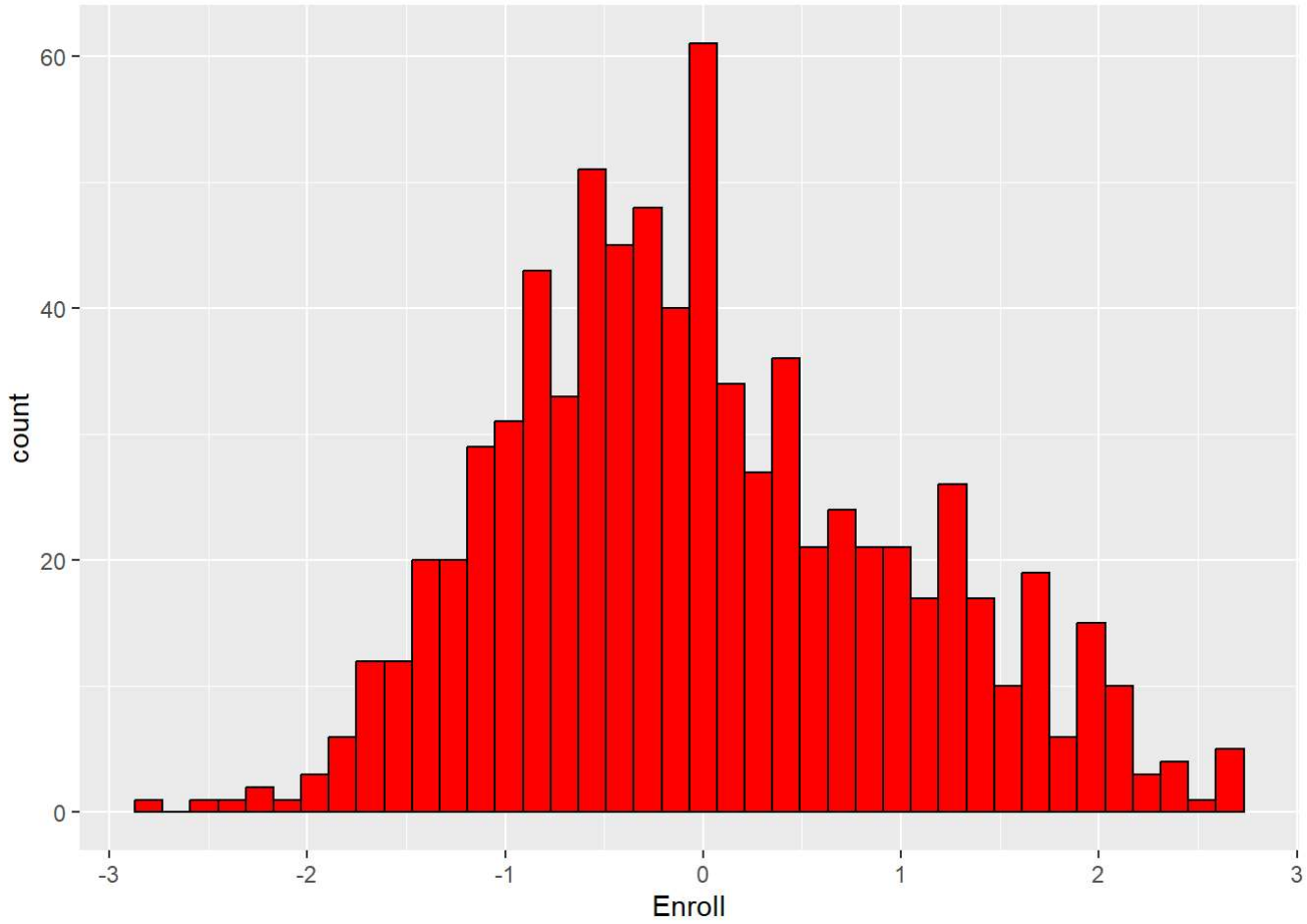
b) Scale the data appropriately (e.g., log transform) and present the visualizations in part A. Have any new relationships been revealed.

```
log_college <- scale(log(college[, c("Apps", "Enroll", "Outstate", "Books", "Personal", "Expen
d")]))

college$Apps <- log_college[, 1]
college$Enroll <- log_college[, 2]
college$Outstate <- log_college[, 3]
college$Books <- log_college[, 4]
college$Personal <- log_college[, 5]
college$Expend <- log_college[, 6]
```

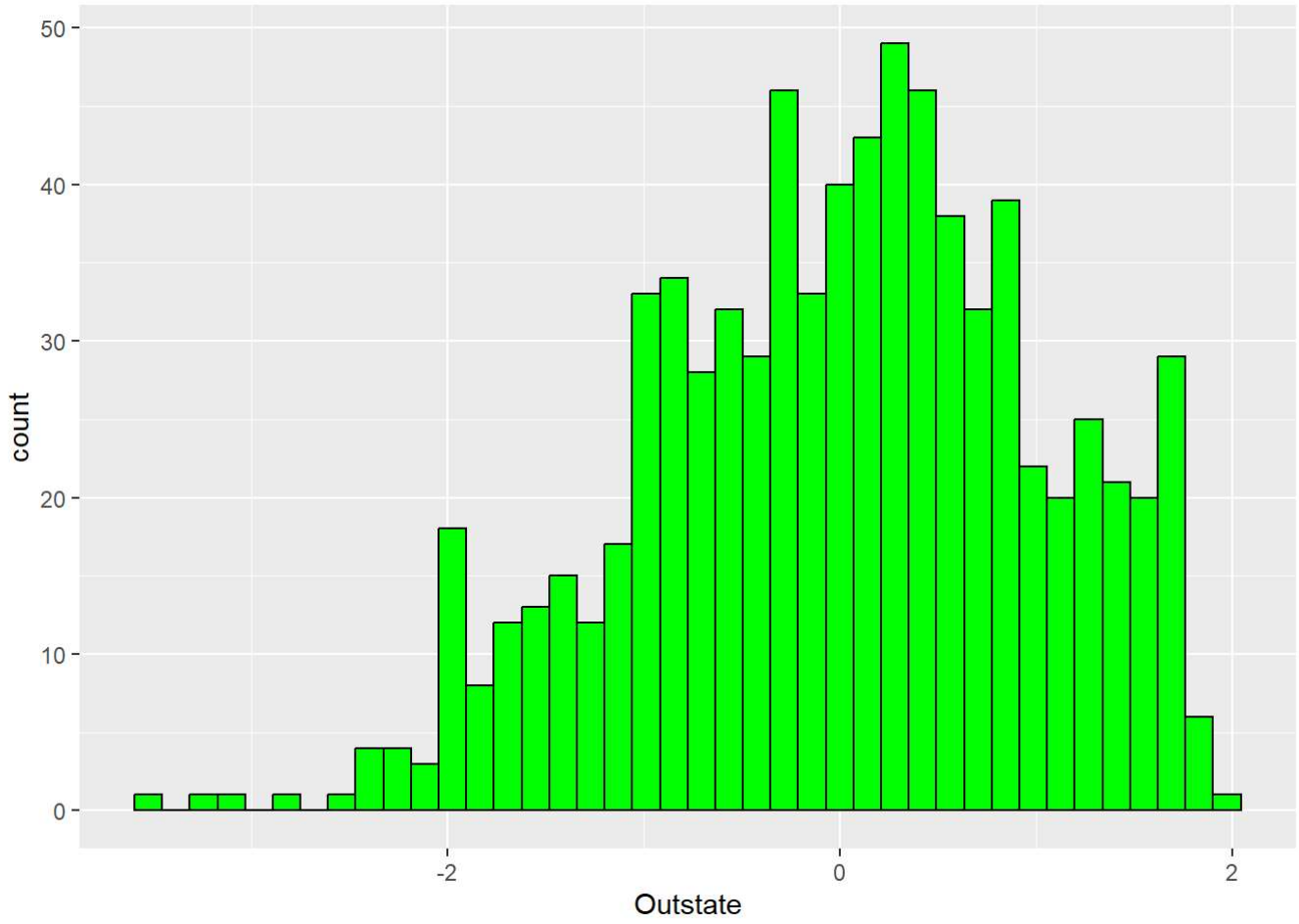
```
### Scaled and log transformed histograms

ggplot(data = college,aes(x=Apps)) +
  geom_histogram(bins = 40, color = 'black',fill = 'blue')
```



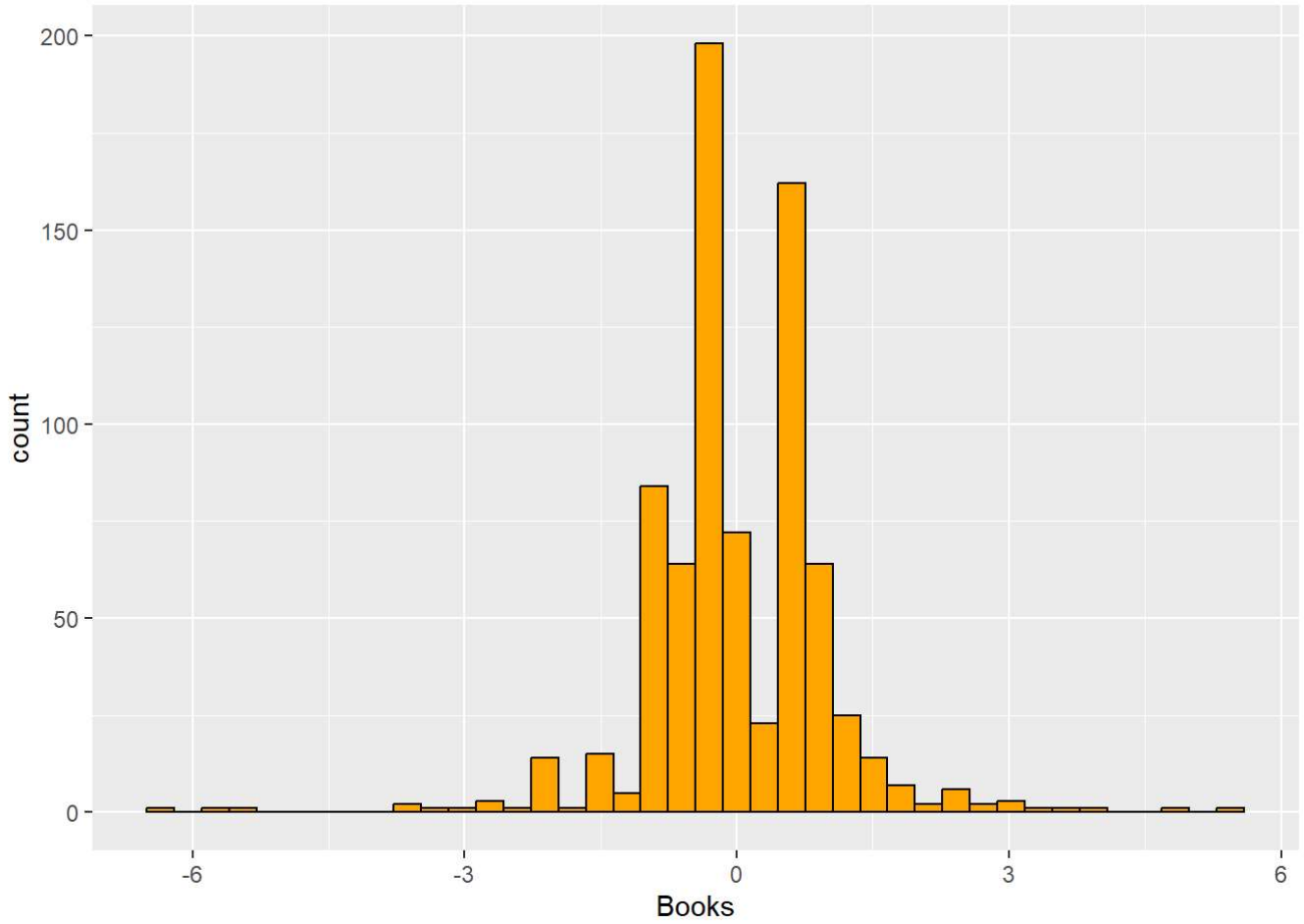

```
ggplot(data = college,aes(x=Enroll)) +
  geom_histogram(bins = 40, color = 'black',fill = 'red')
```



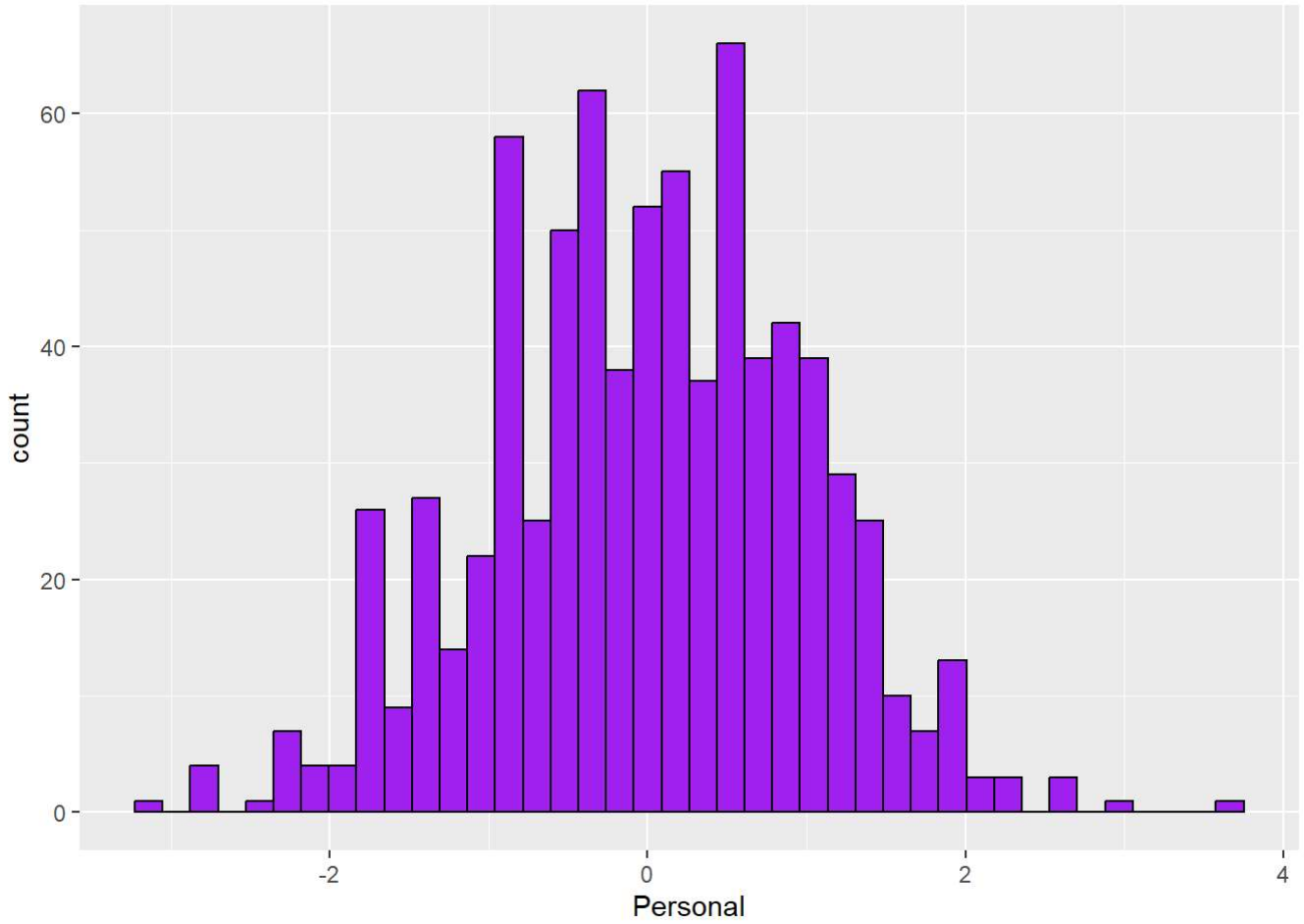

```
ggplot(data = college,aes(x=Outstate)) +
  geom_histogram(bins = 40, color = 'black',fill = 'green')
```



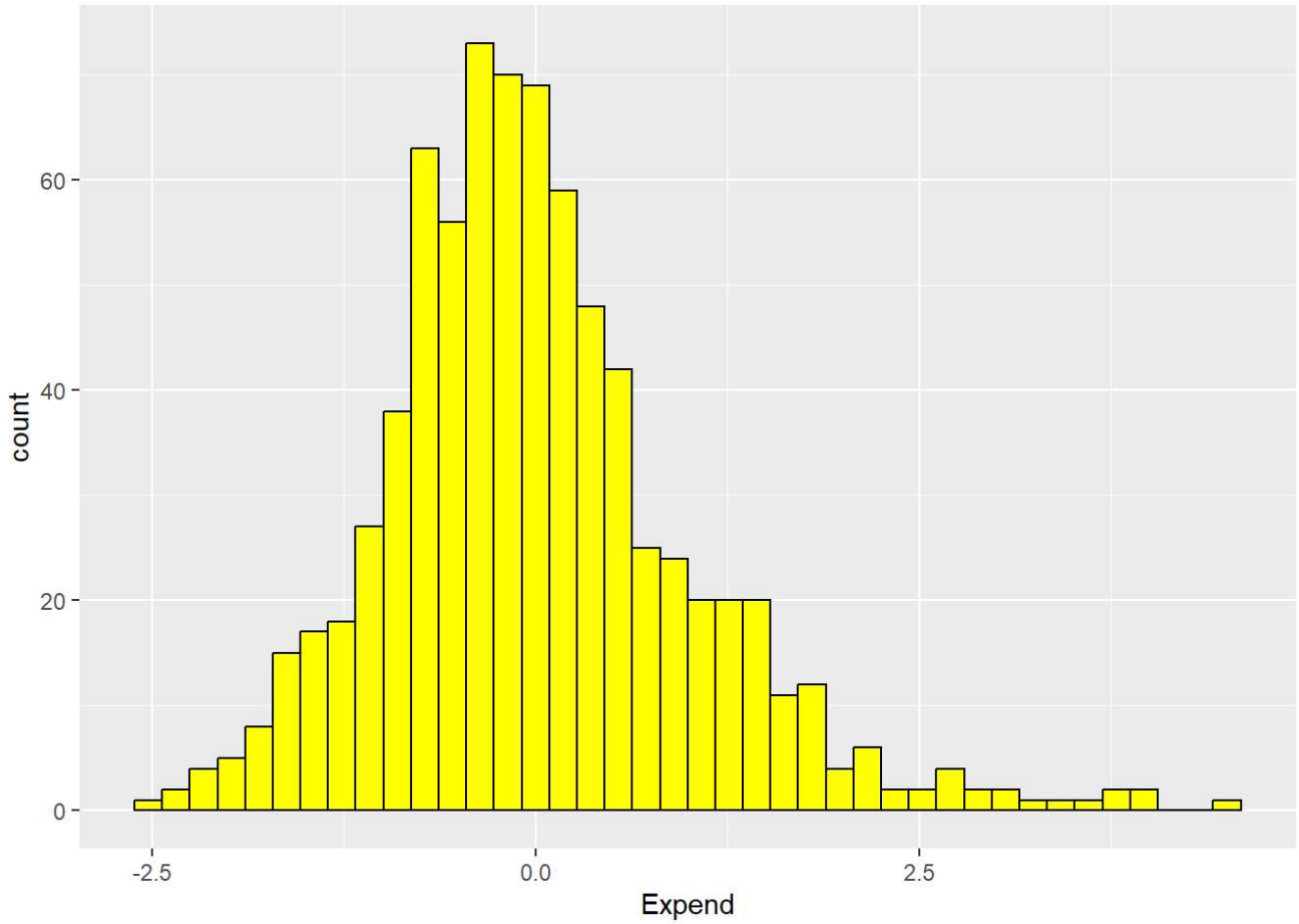

```
ggplot(data = college,aes(x=Books)) +
  geom_histogram(bins = 40, color = 'black',fill = 'orange')
```



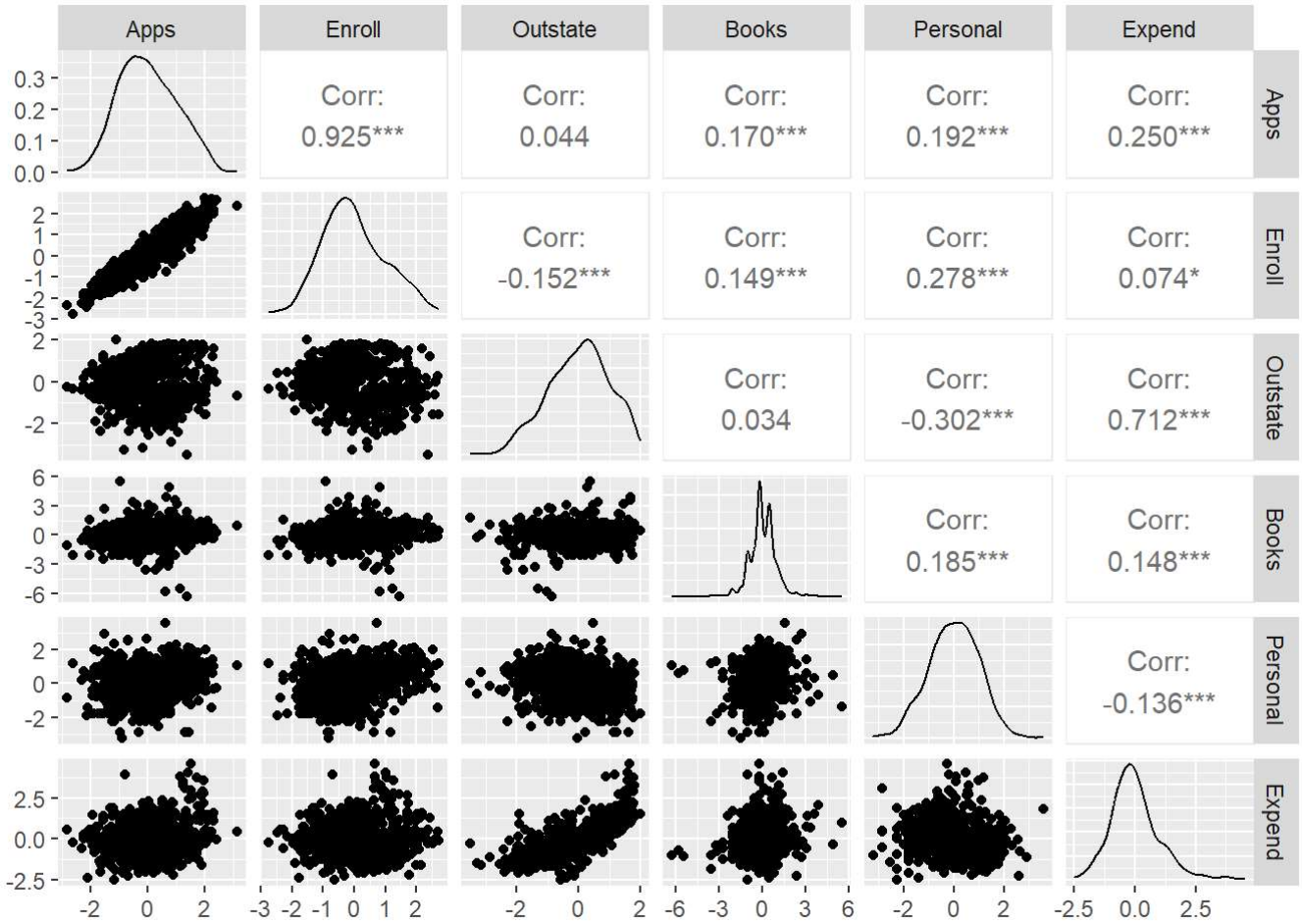

```
ggplot(data = college,aes(x=Personal)) +
  geom_histogram(bins = 40, color = 'black',fill = 'purple')
```




```
ggplot(data = college,aes(x=Expend)) +
  geom_histogram(bins = 40, color = 'black',fill = 'yellow')
```




```
### Scaled and log transformed pair plots
library(GGally)

ggpairs(college[ ,c("Apps", "Enroll", "Outstate", "Books", "Personal", "Expend" )])
```



From the histogram plots, we can now see that the distributions are approximately similar to Normal distribution.

There is positive correlation between number of applications received and number of new students enrolled.

There is positive correlation between number out of state tuition and instructional expenditure per student.

There is no correlation between Instructional expenditure per student and Number of applications received.

c) Subset the data into two data frames: “private” and “public”. Sort them alphabetically. Save them as tab deiminated *txt files. Be sure these are the only two objects saved in that file. Submit it with your assignment (only to Brightspace).

```
private <- college[college$Private == 'Yes',]
public <- college[college$Private == 'No',]

dim(private)
```

```
## [1] 565  18
```

```
dim(public)
```

```
## [1] 212  18
```

```
head(private)
```

```
##                              Private        Apps Accept      Enroll Top10perc
## Abilene Christian University     Yes -0.01119382   1232  0.42749001        23
## Adelphi University               Yes  0.24517114   1924  0.06848141        16
## Adrian College                   Yes -0.15140480   1097 -0.37327391        22
## Agnes Scott College              Yes -1.29786896    349 -1.31415564        60
## Alaska Pacific University        Yes -2.01539265    146 -2.27131164        16
## Albertson College                Yes -0.97939763    479 -1.16458649        38
##                              Top25perc F.Undergrad P.Undergrad   Outstate
## Abilene Christian University        52        2885         537 -0.6442963
## Adelphi University                  29        2683        1227  0.5928626
## Adrian College                      50        1036          99  0.3765792
## Agnes Scott College                 89         510          63  0.7259248
## Alaska Pacific University           44         249         869 -0.6047934
## Albertson College                   62         678          41  0.8267094
##                              Room.Board      Books    Personal PhD Terminal
## Abilene Christian University       3300 -0.6007606  1.25838362  70       78
## Adelphi University                 6450  1.2854137  0.46977813  29       30
## Adrian College                     3750 -1.0356631 -0.05063794  53       66
## Agnes Scott College                5450 -0.6007606 -0.64005007  92       97
## Alaska Pacific University          4120  1.5237159  0.46977813  76       72
## Albertson College                  3335 -0.2117271 -1.17440018  67       73
##                              S.F.Ratio perc.alumni      Expend Grad.Rate
## Abilene Christian University      18.1          12 -0.54122524        60
## Adelphi University                12.2          16  0.44314057        56
## Adrian College                    12.9          30 -0.01357532        54
## Agnes Scott College                7.7          37  1.89043578        59
## Alaska Pacific University         11.9           2  0.53329580        15
## Albertson College                  9.4          11  0.24969579        55
```

```
head(public)
```

-2 marks. Did not sorted.

```
##                                    Private        Apps Accept    Enroll
## Angelo State University                 No  0.69414112   2001 0.7872065
## Appalachian State University            No  1.36987502   4664 1.4492204
## Arizona State University Main campus    No  1.89190271  10308 2.1598465
## Arkansas Tech University                No  0.02942627   1729 0.7178677
## Auburn University-Main Campus           No  1.39933336   6791 1.9469389
## Bemidji State University                No -0.30723115    877 0.1359113
##                                    Top10perc Top25perc F.Undergrad
## Angelo State University                   24        54        4190
## Appalachian State University              20        63        9940
## Arizona State University Main campus      24        49       22593
## Arkansas Tech University                  12        52        3602
## Auburn University-Main Campus             25        57       16262
## Bemidji State University                  12        36        3796
##                                    P.Undergrad   Outstate Room.Board
## Angelo State University                   1512 -1.5621403       3592
## Appalachian State University              1035 -0.8641903       2540
## Arizona State University Main campus      7585 -0.6462881       4850
## Arkansas Tech University                   939 -2.5344771       2650
## Auburn University-Main Campus             1716 -1.0549236       3933
## Bemidji State University                   824 -1.9271282       2700
##                                         Books   Personal PhD Terminal
## Angelo State University             -0.2117271  1.0621339  60       62
## Appalachian State University        -6.3051524  1.0621339  83       96
## Arizona State University Main campus  1.0306642  1.1625959  88       93
## Arkansas Tech University            -0.6007606 -0.3651004  57       60
## Auburn University-Main Campus        0.4614777  0.9651692  85       91
## Bemidji State University             0.8134013  0.8451898  57       62
##                                    S.F.Ratio perc.alumni     Expend Grad.Rate
## Angelo State University                 23.1           5 -1.9190645        34
## Appalachian State University            18.3          14 -0.9930941        70
## Arizona State University Main campus    18.9           5 -1.5820451        48
## Arkansas Tech University                19.6           5 -1.5102474        48
## Auburn University-Main Campus           16.7          18 -0.6840050        69
## Bemidji State University                19.6          16 -2.0818283        46
```

```
#saving data frames as tab delimited files

write.table(private, file = "Private_Colleges.txt", sep = "\t", row.names = FALSE)
write.table(public, file = "Public_Colleges.txt", sep = "\t", row.names = FALSE)
```

d) Within each new data frame from part C, eliminate Universities that have less than the median number of HS students admitted from the top 25% of the class(“Top25perc”).

-2.5 marks. Public median is wrong.

```
private_median <- median(private$Top25perc)
public_median <- median(private$Top25perc)

private_df <- private[private$Top25perc >= private_median, ]
public_df <- public[public$Top25perc >= public_median, ]

dim(private_df)
```

```
## [1] 296  18
```

```
dim(public_df)
```

```
## [1] 89 18
```

```
head(private_df)
```

```
##                     Private        Apps Accept      Enroll Top10perc Top25perc
## Agnes Scott College     Yes -1.29786896    349 -1.31415564        60        89
## Albertson College       Yes -0.97939763    479 -1.16458649        38        62
## Albion College          Yes  0.11408459   1720  0.02027778        37        68
## Albright College        Yes -0.44849233    839 -0.78456007        30        63
## Alfred University       Yes  0.02835141   1425 -0.01683131        37        75
## Allegheny College       Yes  0.42514970   1900  0.00949898        44        77
##                     F.Undergrad P.Undergrad  Outstate Room.Board      Books
## Agnes Scott College         510          63 0.7259248       5450 -0.6007606
## Albertson College           678          41 0.8267094       3335 -0.2117271
## Albion College             1594          32 0.8931084       4826 -0.6007606
## Albright College            973         306 1.1828718       4400 -2.0979013
## Alfred University          1830         110 1.3293133       5406 -0.2117271
## Allegheny College          1707          44 1.4074359       4440 -1.0356631
##                       Personal PhD Terminal S.F.Ratio perc.alumni    Expend
## Agnes Scott College -0.6400501  92       97       7.7          37 1.8904358
## Albertson College   -1.1744002  67       73       9.4          11 0.2496958
## Albion College      -0.6997373  89      100      13.7          37 0.6567398
## Albright College    -1.7923347  79       84      11.3          23 0.6899647
## Alfred University   -1.4169230  82       88      11.3          31 0.5355357
## Allegheny College   -1.4169230  73       91       9.9          41 0.7040073
##                     Grad.Rate
## Agnes Scott College        59
## Albertson College          55
## Albion College             73
## Albright College           80
## Alfred University          73
## Allegheny College          76
```

```
head(public_df)
```

```
##                                     Private      Apps Accept    Enroll
## Appalachian State University             No 1.3698750   4664 1.4492204
## Auburn University-Main Campus            No 1.3993334   6791 1.9469389
## Bloomsburg Univ. of Pennsylvania         No 1.2984301   3028 0.7964559
## California Polytechnic-San Luis          No 1.4312331   3817 1.2957563
## California State University at Fresno    No 0.9258662   3294 1.1838441
## Central Washington University            No 0.4707252   2011 0.7778748
##                                     Top10perc Top25perc F.Undergrad
## Appalachian State University               20        63        9940
## Auburn University-Main Campus              25        57       16262
## Bloomsburg Univ. of Pennsylvania           15        55        5847
## California Polytechnic-San Luis            47        73       12911
## California State University at Fresno       5        60       13494
## Central Washington University               8        65        6507
##                                     P.Undergrad   Outstate Room.Board
## Appalachian State University               1035 -0.8641903       2540
## Auburn University-Main Campus              1716 -1.0549236       3933
## Bloomsburg Univ. of Pennsylvania            946 -0.5137467       2948
## California Polytechnic-San Luis            1404 -0.6642874       4877
## California State University at Fresno      1254 -0.5575685       4368
## Central Washington University              898 -0.7108905       3603
##                                         Books  Personal PhD Terminal
## Appalachian State University        -6.3051524 1.0621339  83       96
## Auburn University-Main Campus        0.4614777 0.9651692  85       91
## Bloomsburg Univ. of Pennsylvania    -0.2117271 0.7031291  66       68
## California Polytechnic-San Luis      0.5345970 1.1537524  72       81
## California State University at Fresno 0.4614777 0.9845033  90       90
## Central Washington University        0.7796805 0.3511161  67       89
##                                     S.F.Ratio perc.alumni      Expend
## Appalachian State University             18.3          14 -0.99309405
## Auburn University-Main Campus            16.7          18 -0.68400496
## Bloomsburg Univ. of Pennsylvania         18.0          19 -0.54122524
## California Polytechnic-San Luis          19.8          13 -0.09389363
## California State University at Fresno    21.2           8 -0.46356391
## Central Washington University           18.1           0 -0.76987763
##                                     Grad.Rate
## Appalachian State University               70
## Auburn University-Main Campus              69
## Bloomsburg Univ. of Pennsylvania           75
## California Polytechnic-San Luis            59
## California State University at Fresno      61
## Central Washington University              51
```

e) Create a new variable that categorizes graduation rate into “High” and “Low”, use a histogram or quantiles to determine how to create this variable. Append this variable to your “private” and “public” datasets.

```
min(college$Grad.Rate)
```

```
## [1] 10
```

```
max(college$Grad.Rate)
```

```
## [1] 118
```

```
median(college$Grad.Rate)
```

```
## [1] 65
```

```
quantile(college$Grad.Rate)
```

```
##   0%  25%  50%  75% 100%
##   10   53   65   78  118
```

```
q <- quantile(college$Grad.Rate)[[3]]

grad_rate_map <- function(rate) {
      if(rate <= q) {
         return("Low")
      }
      else {
         return("High")
      }
}
```

```
private_df$Grad.Rate.Category <- sapply(private_df$Grad.Rate, grad_rate_map)
public_df$Grad.Rate.Category <- sapply(public_df$Grad.Rate, grad_rate_map)
```

f) Create a “list structure” that contains your two datasets and save this to an *.RData file. Make sure that your file contains only the list structure. Submit this with your homework (only on Brightspace).

```
college_list <- list(private = private_df, public = public_df)

save(college_list, file = "college_list.RData")
```

Question 2) Consider the “groceries” data used in the Computational Lab.

```
library("arules")
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
##
## Attaching package: 'arules'
```

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

```
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
setwd("E:/Buffalo/files")

dats <- read.transactions("groceries.csv", sep = ",")
dats
```

```
## transactions in sparse format with
##  9835 transactions (rows) and
##  169 items (columns)
```

```
summary(dats)
```

```
## transactions as itemMatrix in sparse format with
##  9835 rows (elements/itemsets/transactions) and
##  169 columns (items) and a density of 0.02609146
##
## most frequent items:
##       whole milk other vegetables       rolls/buns             soda
##             2513             1903             1809             1715
##           yogurt          (Other)
##             1372            34055
##
## element (itemset/transaction) length distribution:
## sizes
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
## 2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55   46
##   17   18   19   20   21   22   23   24   26   27   28   29   32
##   29   14   14    9   11    4    6    1    1    1    1    3    1
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   4.409   6.000  32.000
##
## includes extended item information - examples:
##             labels
## 1 abrasive cleaner
## 2 artif. sweetener
## 3   baby cosmetics
```

```
head(toLongFormat(dats, decode = FALSE), n=10) #numeric assign. of item
```

```
##    TID item
## 1    1   30
## 2    1   89
## 3    1  119
## 4    1  133
## 5    2   34
## 6    2  158
## 7    2  168
## 8    3  167
## 9    4   39
## 10   4   92
```

```
head(toLongFormat(dats, decode = TRUE), n=10) #name assign. of item
```

```
##    TID                item
## 1    1        citrus fruit
## 2    1           margarine
## 3    1         ready soups
## 4    1 semi-finished bread
## 5    2              coffee
## 6    2      tropical fruit
## 7    2              yogurt
## 8    3          whole milk
## 9    4        cream cheese
## 10   4        meat spreads
```

```
# data by name and visualization
inspect(dats[1:5])
```

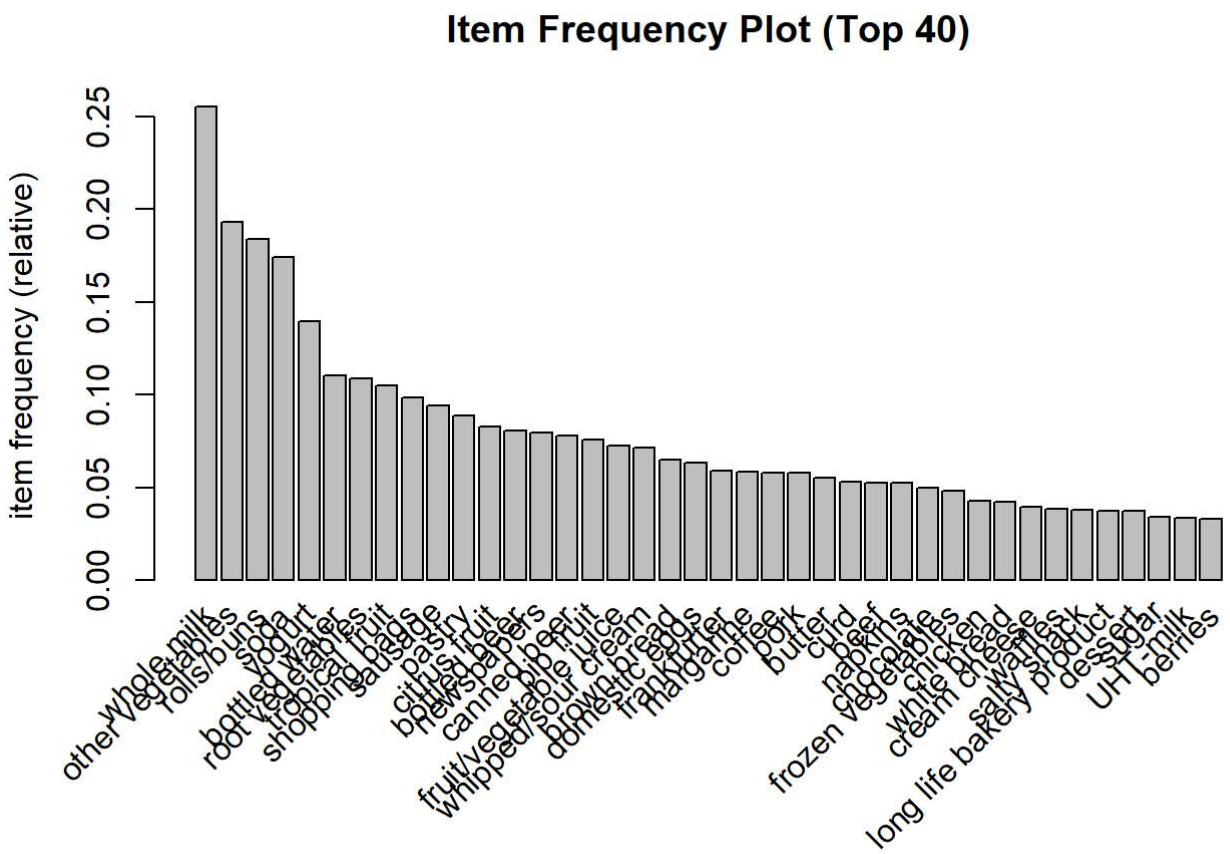
```
##     items
## [1] {citrus fruit,
##      margarine,
##      ready soups,
##      semi-finished bread}
## [2] {coffee,
##      tropical fruit,
##      yogurt}
## [3] {whole milk}
## [4] {cream cheese,
##      meat spreads,
##      pip fruit,
##      yogurt}
## [5] {condensed milk,
##      long life bakery product,
##      other vegetables,
##      whole milk}
```

a) Visualize the item frequency plot for the top 40 grocery items.

```
itemFrequency(dats[,1:5]) #support of different items
```

```
## abrasive cleaner artif. sweetener   baby cosmetics        baby food
##     0.0035587189     0.0032536858     0.0006100661     0.0001016777
##             bags
##     0.0004067107
```

```
itemFrequencyPlot(dats, topN = 40, main = "Item Frequency Plot (Top 40)")
```



b) Rank the top five rules with the highest “confidence”.

```
my_params <- list(support = .005, confidence = .01, minlen = 2, maxlen = 6)
my_rules <- apriori(dats, parameter = my_params)
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##        0.01    0.1    1 none FALSE            TRUE       5   0.005      2
##  maxlen target  ext
##       6  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 49
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
## sorting and recoding items ... [120 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [2050 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
my_rules
```

```
## set of 2050 rules
```

```
inspect(my_rules[1:3])
```

```
##     lhs             rhs                support     confidence coverage
## [1] {cake bar}   => {whole milk}       0.005592272 0.42307692 0.01321810
## [2] {whole milk} => {cake bar}         0.005592272 0.02188619 0.25551601
## [3] {dishes}     => {other vegetables} 0.005998983 0.34104046 0.01759024
##     lift     count
## [1] 1.655775 55
## [2] 1.655775 55
## [3] 1.762550 59
```

```
inspect(sort(my_rules, by = "confidence")[1:5])
```

```
##     lhs                    rhs              support confidence    coverage     lift count
## [1] {root vegetables,
##      tropical fruit,
##      yogurt}             => {whole milk} 0.005693950  0.7000000 0.008134215 2.739554    56
## [2] {other vegetables,
##      pip fruit,
##      root vegetables}    => {whole milk} 0.005490595  0.6750000 0.008134215 2.641713    54
## [3] {butter,
##      whipped/sour cream} => {whole milk} 0.006710727  0.6600000 0.010167768 2.583008    66
## [4] {pip fruit,
##      whipped/sour cream} => {whole milk} 0.005998983  0.6483516 0.009252669 2.537421    59
## [5] {butter,
##      yogurt}             => {whole milk} 0.009354347  0.6388889 0.014641586 2.500387    92
```

c) Rank the top ten rules with the highest “lift”.

```
inspect(sort(my_rules, by = "lift")[1:10])
```

```
##      lhs                     rhs                     support confidence   coverage     lift
count
## [1]  {ham}                => {white bread}        0.005083884 0.19531250 0.02602949 4.639851
50
## [2]  {white bread}        => {ham}                0.005083884 0.12077295 0.04209456 4.639851
50
## [3]  {citrus fruit,
##       other vegetables,
##       whole milk}         => {root vegetables}    0.005795628 0.44531250 0.01301474 4.085493
57
## [4]  {butter,
##       other vegetables}   => {whipped/sour cream} 0.005795628 0.28934010 0.02003050 4.036397
57
## [5]  {root vegetables}    => {herbs}              0.007015760 0.06436567 0.10899847 3.956477
69
## [6]  {herbs}              => {root vegetables}    0.007015760 0.43125000 0.01626843 3.956477
69
## [7]  {other vegetables,
##       root vegetables}    => {onions}             0.005693950 0.12017167 0.04738180 3.875044
56
## [8]  {citrus fruit,
##       pip fruit}          => {tropical fruit}     0.005592272 0.40441176 0.01382816 3.854060
55
## [9]  {berries}            => {whipped/sour cream} 0.009049314 0.27217125 0.03324860 3.796886
89
## [10] {whipped/sour cream} => {berries}            0.009049314 0.12624113 0.07168277 3.796886
89
```

d) Comment on the consistency and differences between B and C.

```
### Rules with high confidence (part b) indicate commonly co-purchased items. Customers who buy "root vegetables," "tropical fruit," and "yogurt" are highly likely (70% confidence) to also buy "whole milk".

### Rules with high lift (part c) indicate strong associations that are unexpected based on the frequency of individual items. For instance, "ham" and "white bread" have a lift of 4.639, sugge sting that customers who buy ham are 4.639 times more likely to buy white bread than if the purc hases were independent.

### The rules ranked by confidence have more modest lift values (in the 2-3 range), while the to p rules by lift have very high values (4+). So the lift metric surfaces the most disproportionat ely strong associations.

### There are some consistent associations between certain items in both sets of rules. For exam ple, "root vegetables" and "whipped/sour cream" appear in both top rule lists but with different consequents.
```

e) Your manager wants to develop “bundles” of items to increase their pastry sales. What recommendations can you make?

```
### Discounts like buy one pastry, get a discount on a coffee or combo deal pastry with jam or h oney for a special price.

### Bundle pastries with berries or whipped/sour cream, which also have high lift. These could b e used as pastry toppings to make a pastry bundle more appealing.

### Create a brunch or dessert bundle with pastries, tropical fruit, and yogurt. All three have strong rules connections.

### Consider bundles like a kid's bundle with pastries, candy, and fruit juice.
```

Question 3) Business Intelligence for RM, would like to initiate a study of the purchase behavior of customers who use the RM loyalty card (a card that customers scan at checkout to qualify for discounted prices). The use of the loyalty card allows RM to capture what is known as “point-of-sale” data, that is, a list of products purchased by customers as they check out of the market. David feels that better understanding of which products tend to be purchased together could lead to insights for better pricing and display strategies as well as a better understanding of sales and the potential impact of different levels of coupon discounts. This type of analysis is known as market basket analysis, as it is a study of what different customers have in their shopping baskets as they check out of the store.As a prototype study, David wants to investigate customer buying behavior with regard to bread, jelly, and peanut butter. RM’s Information Technology (IT) group, at David’s request, has provided a data set of purchases by 1000 customers over a one-week period. The data set is in the file MarketBasket, and it contains the following variables for each customer:- Bread—wheat, white, or none- Jelly—grape, strawberry, or none - Peanut butter—creamy, natural, or none The variables appear in the above order from left to right in the data set, where each row is a customer. For example, the first record of the data set is white grape none which means that customer 1 purchased white bread, grape jelly, and no peanut butter. The second record

is white strawberry none which means that customer 2 purchased white bread, strawberry jelly, and no peanut butter. The sixth record in the data set is none none none which means that the sixth customer did not purchase bread, jelly, or peanut butter.Other records are interpreted in a similar fashion.David would like you to do an initial study of the data to get a better understanding of RM customer behavior with regard to these three products

a) Create a Binary Transaction Matrix. Submit this matrix as a tab delimited text file (only to Brightspace).

```
library("readxl")

setwd("E:/Buffalo/files")

dat <- read_excel("E:/Buffalo/files/marketbasket.xlsx")
head(dat)
```

```
## # A tibble: 6 × 3
##   Bread Jelly      `Peanut Butter`
##   <chr> <chr>      <chr>
## 1 white grape      none
## 2 white strawberry none
## 3 none  none       none
## 4 white none       none
## 5 wheat none       none
## 6 none  none       none
```

```
dim(dat)
```

```
## [1] 1000    3
```

```
### Creating binary transaction matrix

bin_mat <- matrix(0, nrow = nrow(dat), ncol = 9)
colnames(bin_mat) <- c("wheat", "white", "none", "grape", "strawberry", "none", "creamy", "natur
al", "none")

# Fill the matrix based on data

for(i in 1:nrow(dat)){
  if(dat[i,1] == "wheat"){
    bin_mat[i,1] <- 1
  } else if(dat[i,1] == "white"){
    bin_mat[i,2] <- 1
  } else {
    bin_mat[i,3] <- 1
  }

  if(dat[i,2] == "grape"){
    bin_mat[i,4] <- 1
  } else if(dat[i,2] == "strawberry"){
    bin_mat[i,5] <- 1
  } else {
    bin_mat[i,6] <- 1
  }

  if(dat[i,3] == "creamy"){
    bin_mat[i,7] <- 1
  } else if(dat[i,3] == "natural"){
    bin_mat[i,8] <- 1
  } else {
    bin_mat[i,9] <- 1
  }
}

#saving binary transaction matrix as tab delimited file

write.table(bin_mat, file = "binary_transaction_matrix.txt", sep = "\t", row.names = FALSE)
```

b) David would like you to do an initial study of the data to get a better understanding of RM customer behavior about these three products. What can you advise?

```
### Look at the frequency and proportions of each product purchased to see the overall distribut
ion and popularity. This gives a basic understanding of customer behavior.

### Generate association rules to identify patterns and associations between products purchased
together. This can reveal if certain products are likely to be co-purchased.

### Analyze results over time to identify trends and seasonality in purchases of these products.
Buying behaviors may change over time.

### Perform market basket analysis to find common item sets purchased together. Frequent item se
ts can show which combinations of products customers tend to buy.

### Compare bread, jelly and peanut butter purchases to other product categories in the store. T
his can give a relative perspective on how prominent these items are in overall baskets.
```

-3 marks study of the data missing in part b. Utilize the apriori algorithm, as taught in the class and lab videos, to support your analysis.