Topic 1: Exercise 1

Daniel Alonso

November 15th, 2020

Importing the data

```
library(dplyr)
library(stringr)
library(PerformanceAnalytics)
library(foreach)
library(ggplot2)
library(reshape2)
library(MASS)
library(andrews)
```

```
d <- read.csv('../../datasets/Colleges.csv')
```

```
head(d)
```

```
##                                             X  Private Apps Accept Enroll
## 1 Abilene Christian University                Yes      1660   1232    721
## 2 Adelphi University                          Yes      2186   1924    512
## 3 Adrian College                              Yes      1428   1097    336
## 4 Agnes Scott College                         Yes       417    349    137
## 5 Alaska Pacific University                   Yes       193    146     55
## 6 Albertson College                           Yes       587    479    158
##   Top10perc Top25perc F.Undergrad P.Undergrad Outstate Room.Board Books
## 1        23        52        2885         537     7440       3300   450
## 2        16        29        2683        1227    12280       6450   750
## 3        22        50        1036          99    11250       3750   400
## 4        60        89         510          63    12960       5450   450
## 5        16        44         249         869     7560       4120   800
## 6        38        62         678          41    13500       3335   500
##   Personal PhD Terminal S.F.Ratio perc.alumni Expend Grad.Rate
## 1     2200  70       78      18.1          12   7041        60
## 2     1500  29       30      12.2          16  10527        56
## 3     1165  53       66      12.9          30   8735        54
## 4      875  92       97       7.7          37  19016        59
## 5     1500  76       72      11.9           2  10922        15
## 6      675  67       73       9.4          11   9727        55
```

1- Identify the type of all variables

```
foreach (col = d, nms = colnames(d)) %do% {
    print(str_interp("${nms} is of type ${typeof(col)}"))
}
```

represented as R type - numerical/categorical

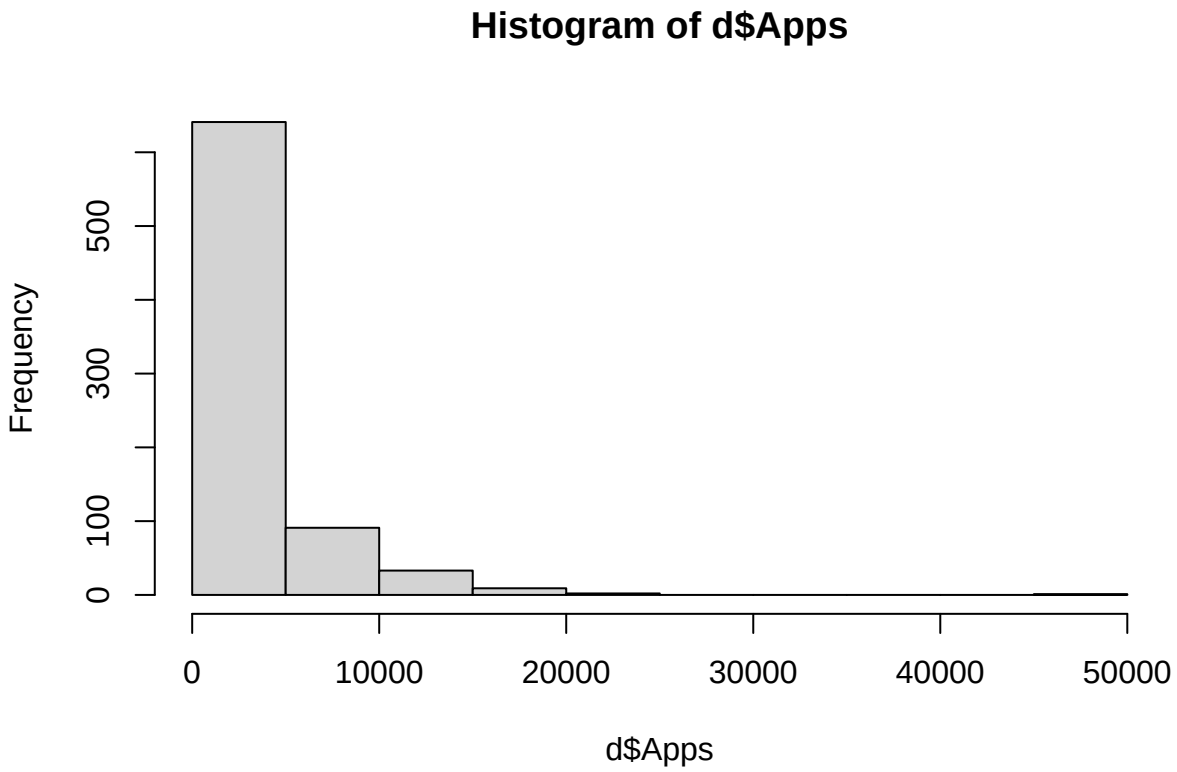
- X is of type character - categorical (id)
- Private is of type character - categorical binary
- Apps is of type integer - numerical discrete
- Accept is of type integer - numerical discrete
- Enroll is of type integer - numerical discrete
- Top10perc is of type integer - numerical discrete
- Top25perc is of type integer - numerical discrete
- F.Undergrad is of type integer - numerical discrete
- P.Undergrad is of type integer - numerical discrete
- Outstate is of type integer - numerical discrete
- Room.Board is of type integer - numerical discrete
- Books is of type integer - numerical discrete
- Personal is of type integer - numerical discrete
- PhD is of type integer - numerical discrete
- Terminal is of type integer - numerical discrete
- S.F.Ratio is of type double - numerical continuous
- perc.alumni is of type integer - numerical discrete
- Expend is of type integer - numerical discrete
- Grad.Rate is of type integer - numerical discrete

2- Perform a visual analysis of each quantitative variable. Then do so considering the variable *private*. Describe each plot.

Histograms

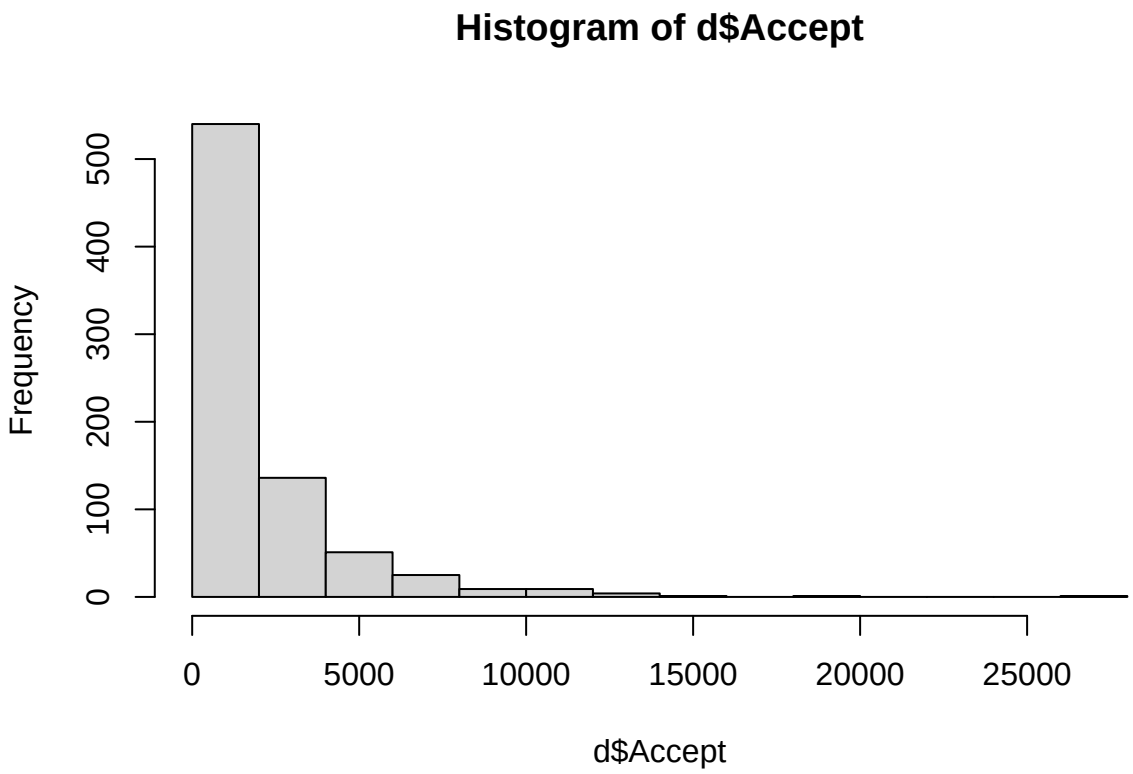
```
# setting plot sizes
options(repr.plot.width = 14, repr.plot.height = 8)
```

```
# Apps
hist(d$Apps)
```



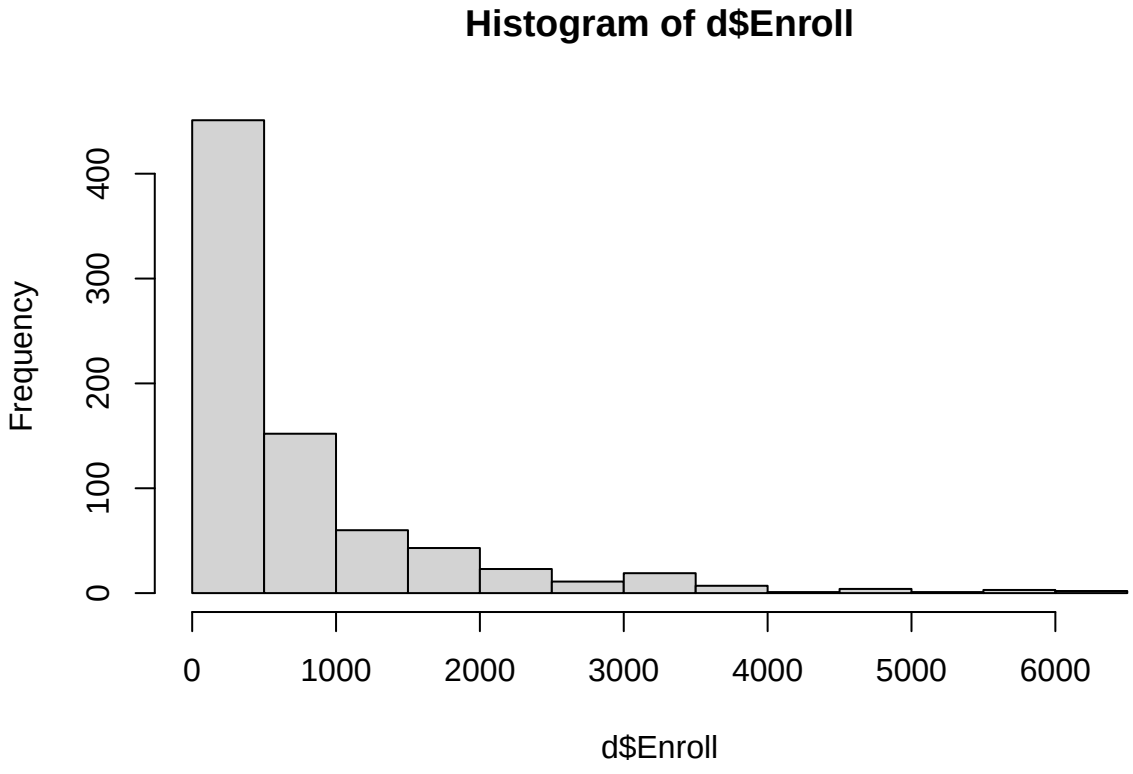
- Large right tail, higher frequency
- Long left tail
- Very right skewed
- Resembles gamma distribution

```
# Accept
hist(d$Accept)
```



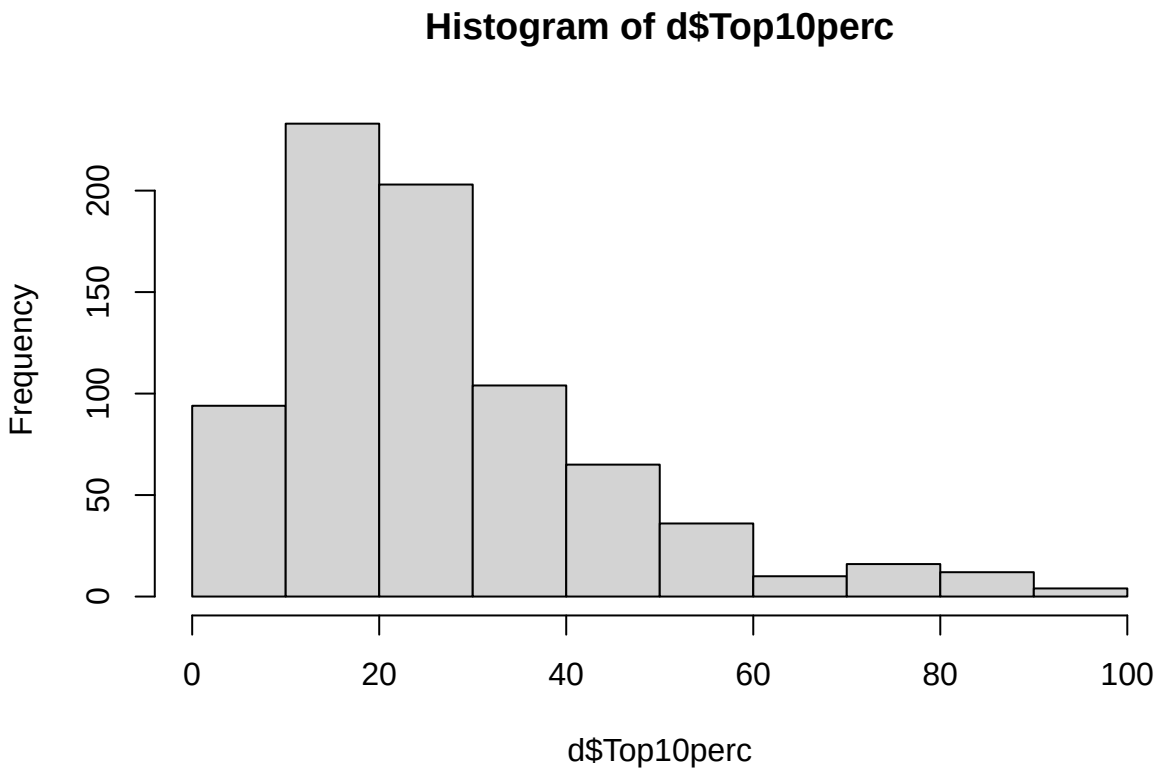
Pretty much the same as Apps but with a significantly shorter left tail. Probably because those students being accepted are just a subset of those that apply.

```
# Enroll
hist(d$Enroll)
```



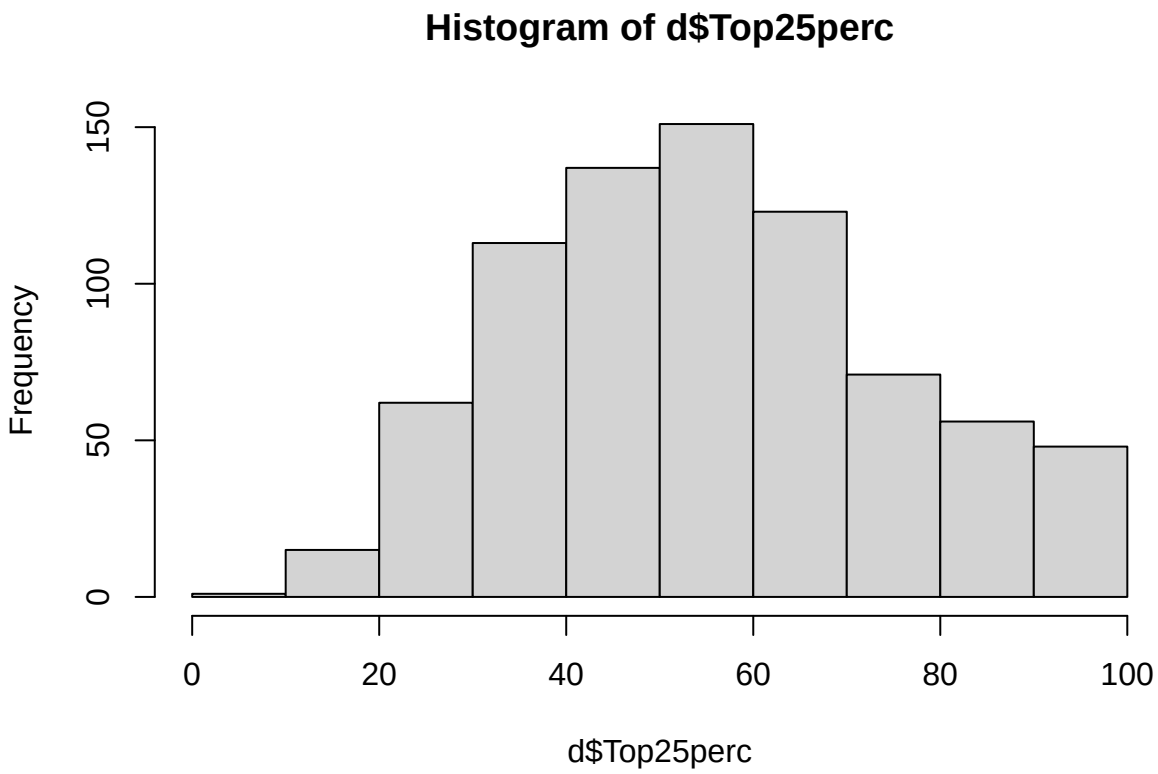
Enroll also tells the same story, but in this case our left tail is even smaller than the accept left tail. Maybe because significantly less students enroll than those that just get accepted

```
# Top10perc
hist(d$Top10perc)
```



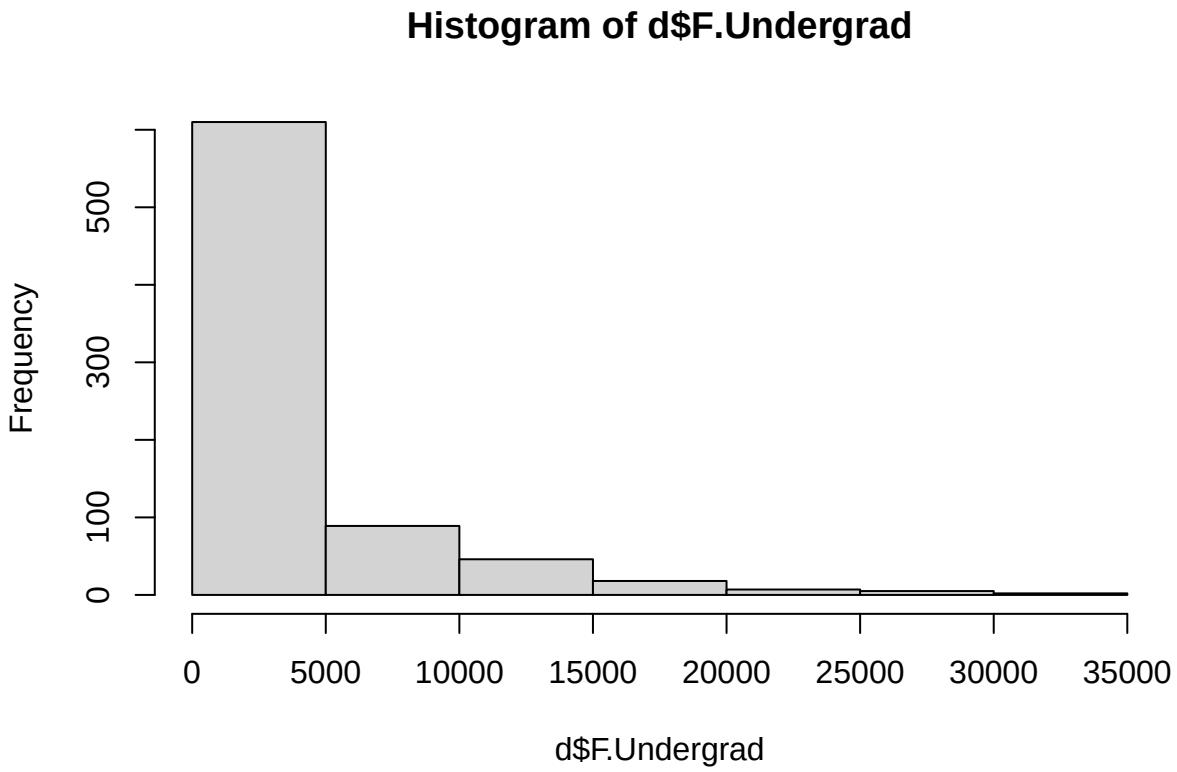
- long left tail
- short but high frequency right tail
- Resembles gaussian distribution
- most elements concentrate around 20

```
# Top25perc
hist(d$Top25perc)
```



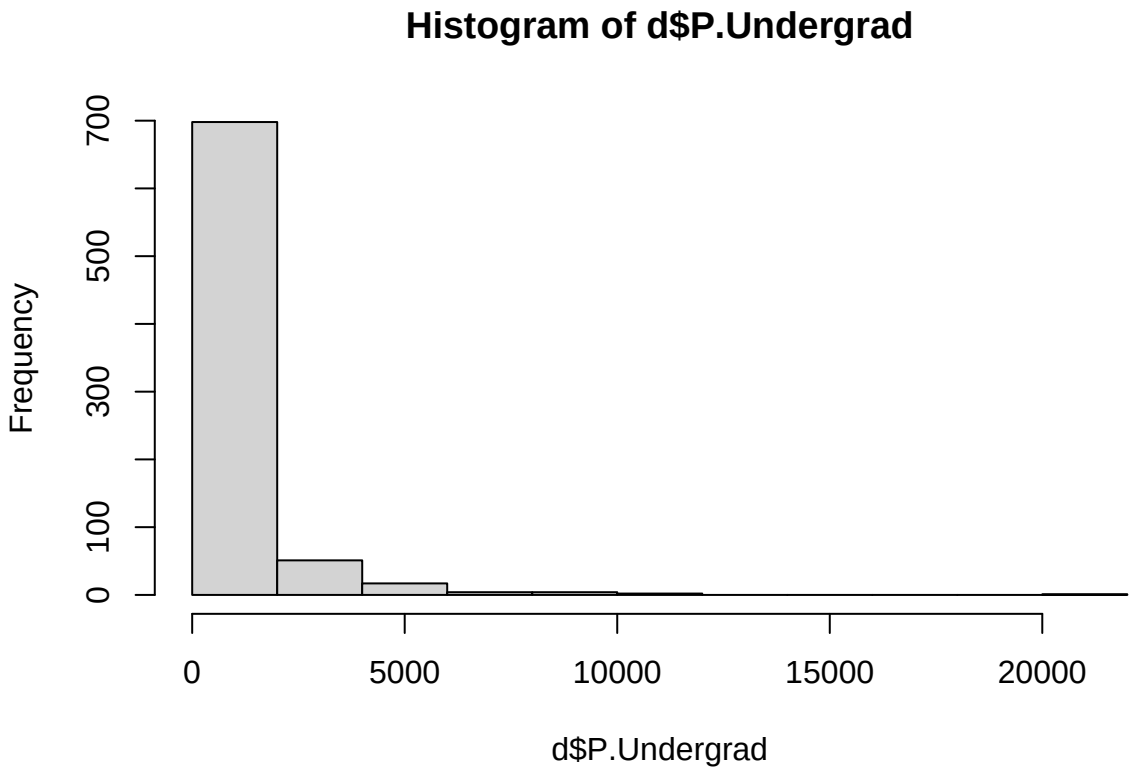
- mean is probably between 40-60
- small right tail
- large left tail
- Resembles gaussian distribution

```
# F.Undergrad
hist(d$F.Undergrad)
```



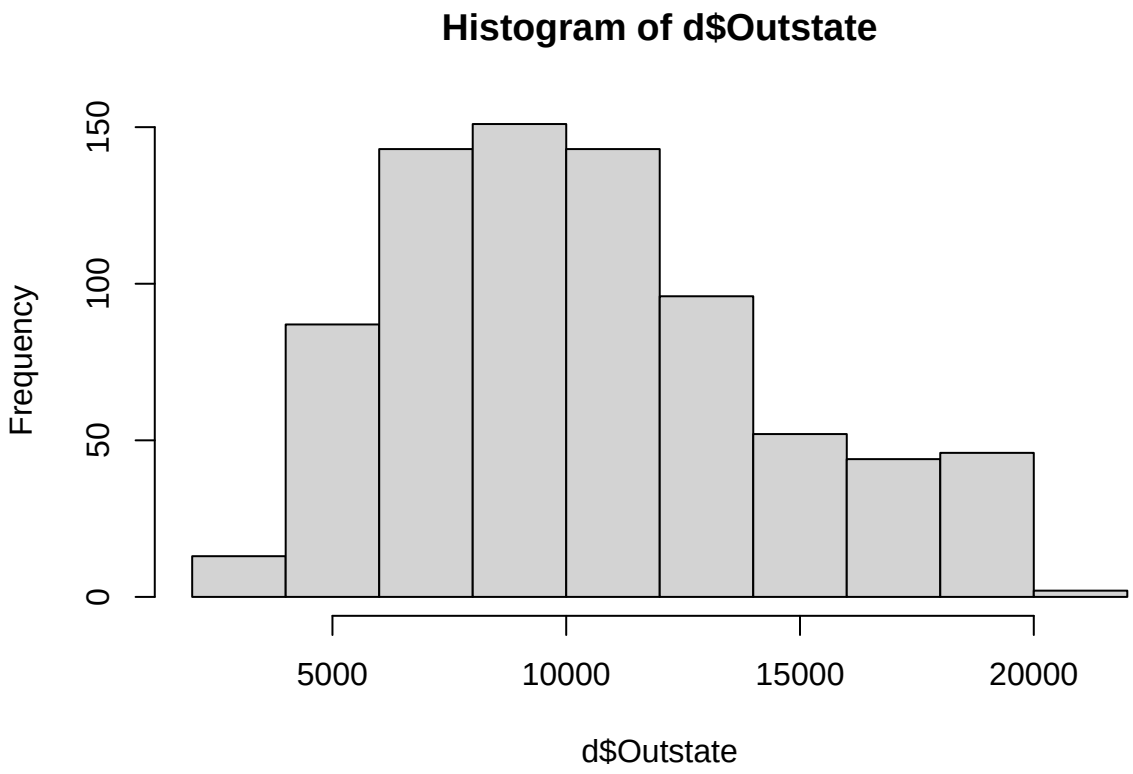
- long left tail
- most of the universities between 0-5000
- significantly less data at 5000+
- Resembles gamma distribution

```
# P.Undergrad
hist(d$P.Undergrad)
```



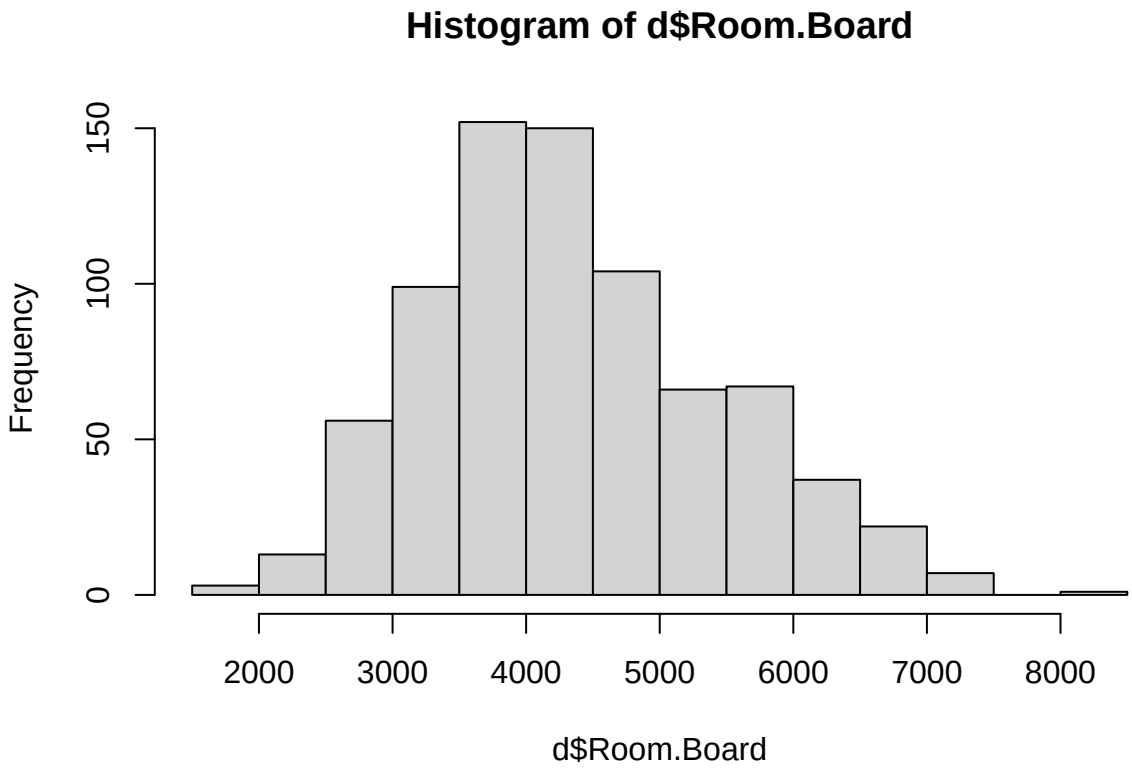
Similar to F.Undergrad we have a quite long right tail, this one resembles Apps more, and similar to F.Undergrad we also have a large frequency right tail and possibly more than 90% of the data between 0 and 5000

```
# Outstate
hist(d$Outstate)
```



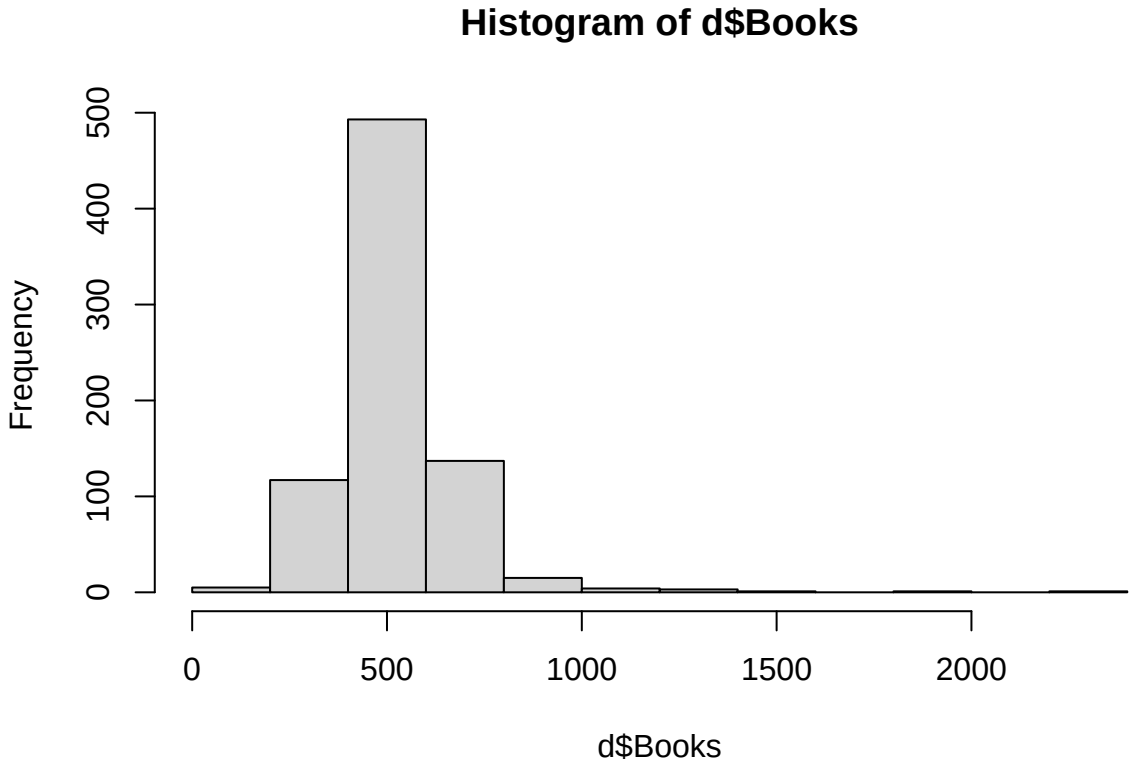
- Mean around 10000
- relatively heavy left tail but not very long
- light right tail
- resembles gaussian distribution

```
# Room.Board
hist(d$Room.Board)
```



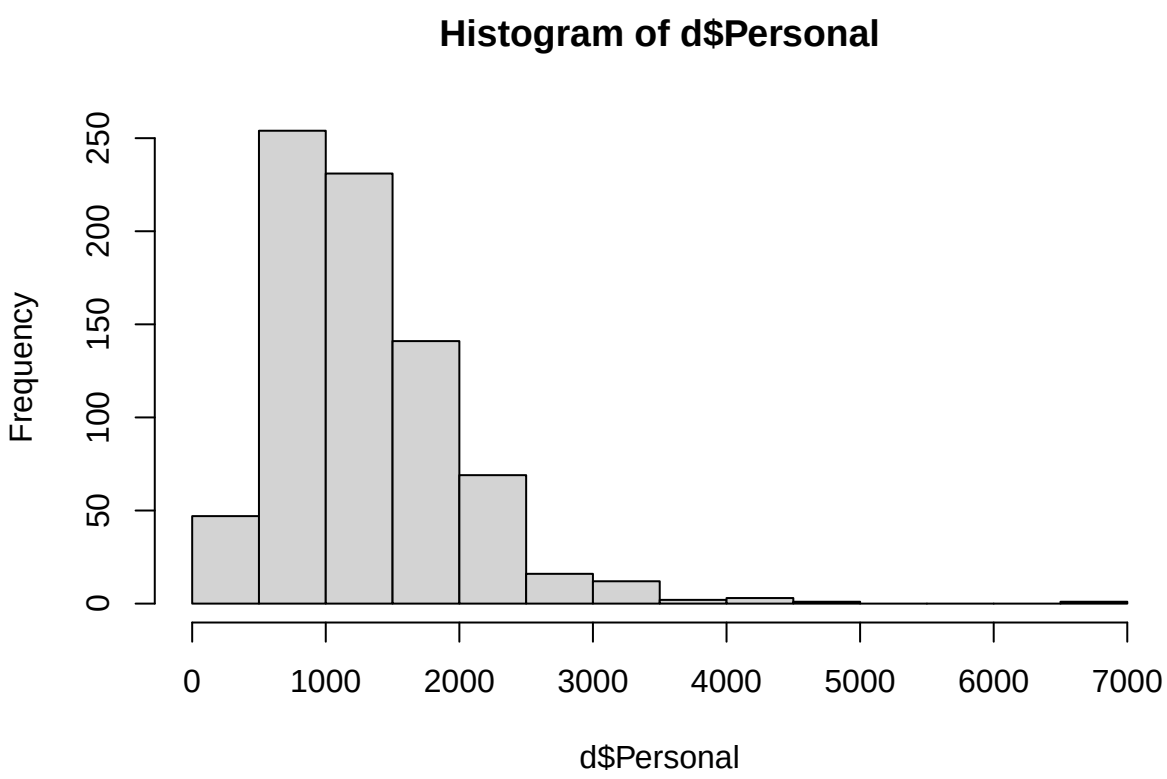
- resembles a slightly right skewed gaussian distribution
- relatively long left tail
- mean between 3500-4500

```
# Books
hist(d$Books)
```



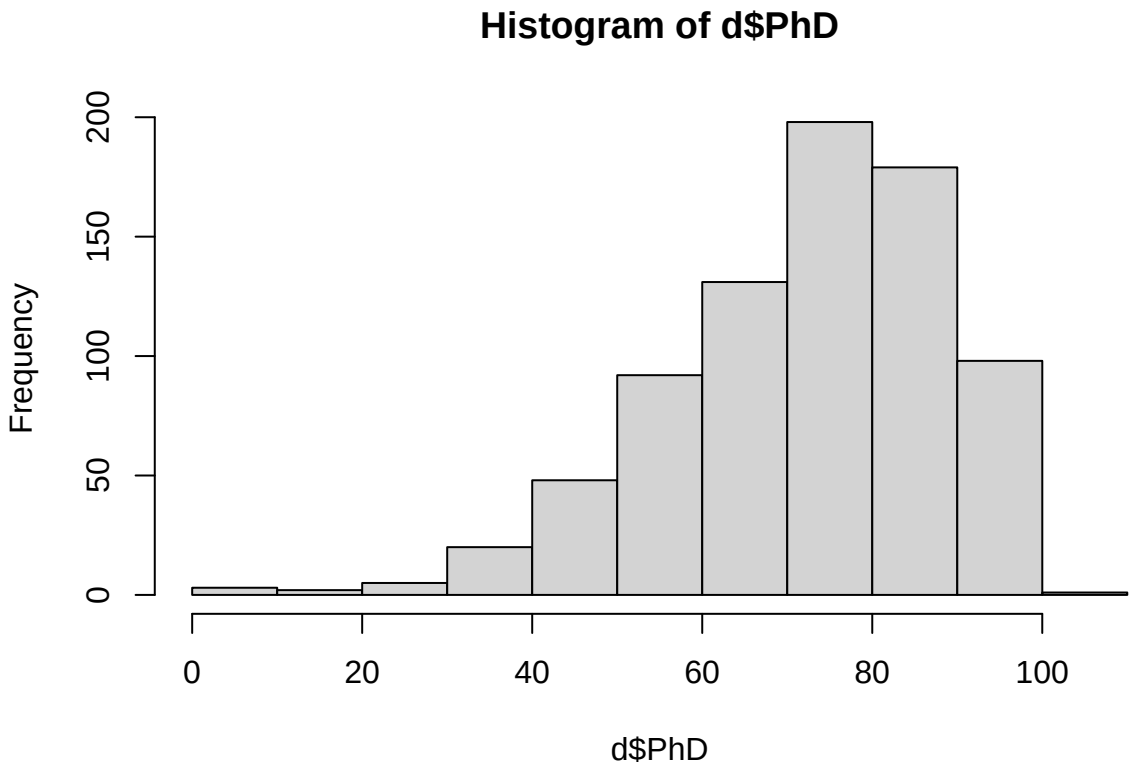
- long left tail
- very right skewed
- mean most probably around 500
- nearly all values between 400-750

```
# Personal
hist(d$Personal)
```



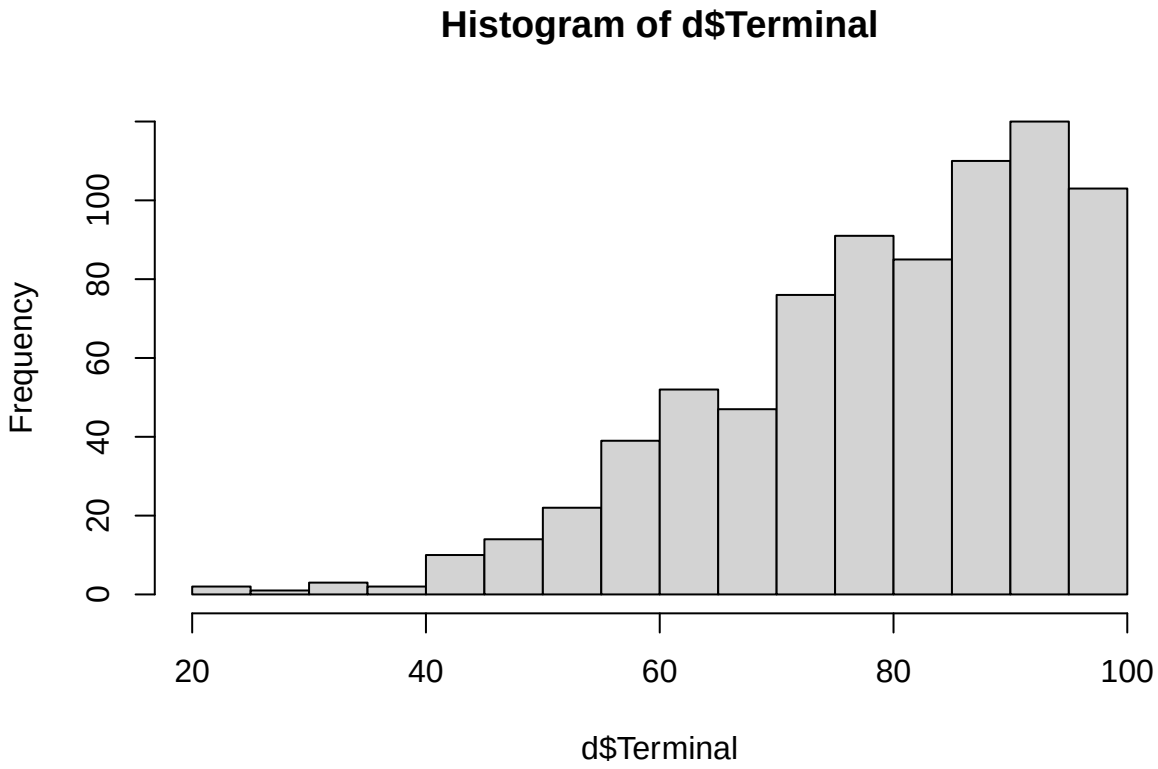
- very right skewed
- long left tail
- resembles a gamma distribution

```
# PhD
hist(d$PhD)
```



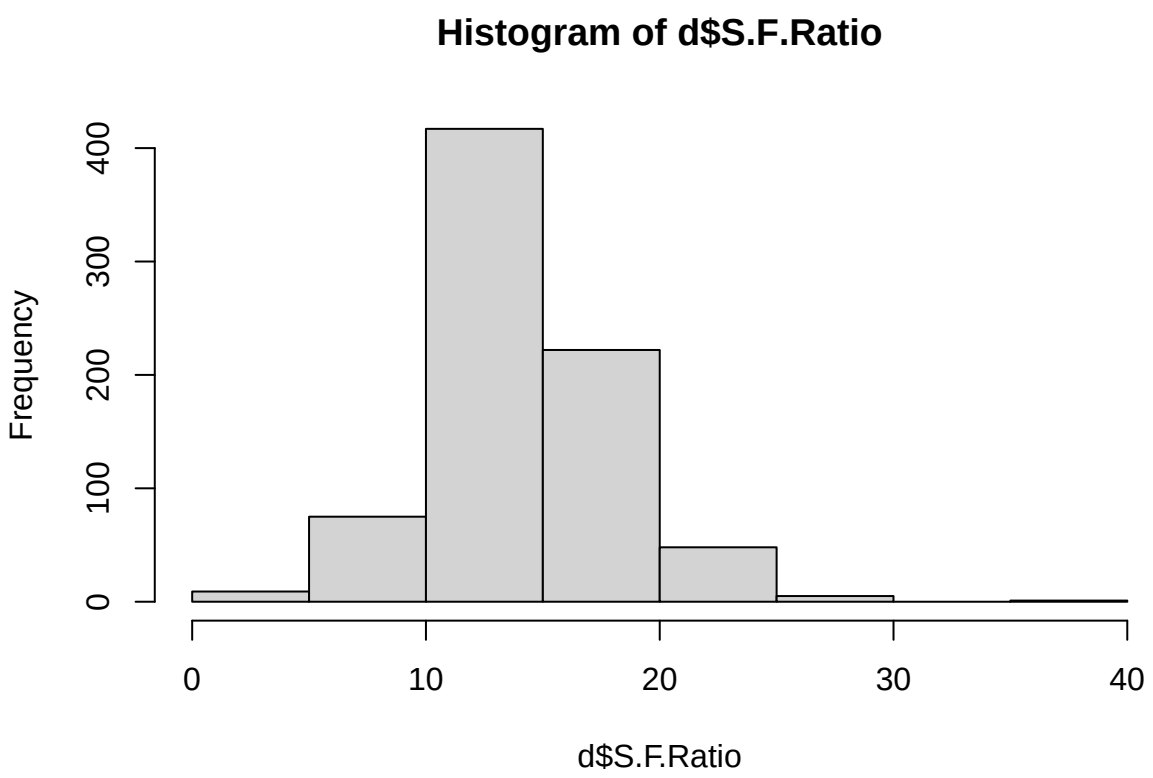
- long and light right tail
- very left skewed

```
# Terminal
hist(d$Terminal)
```



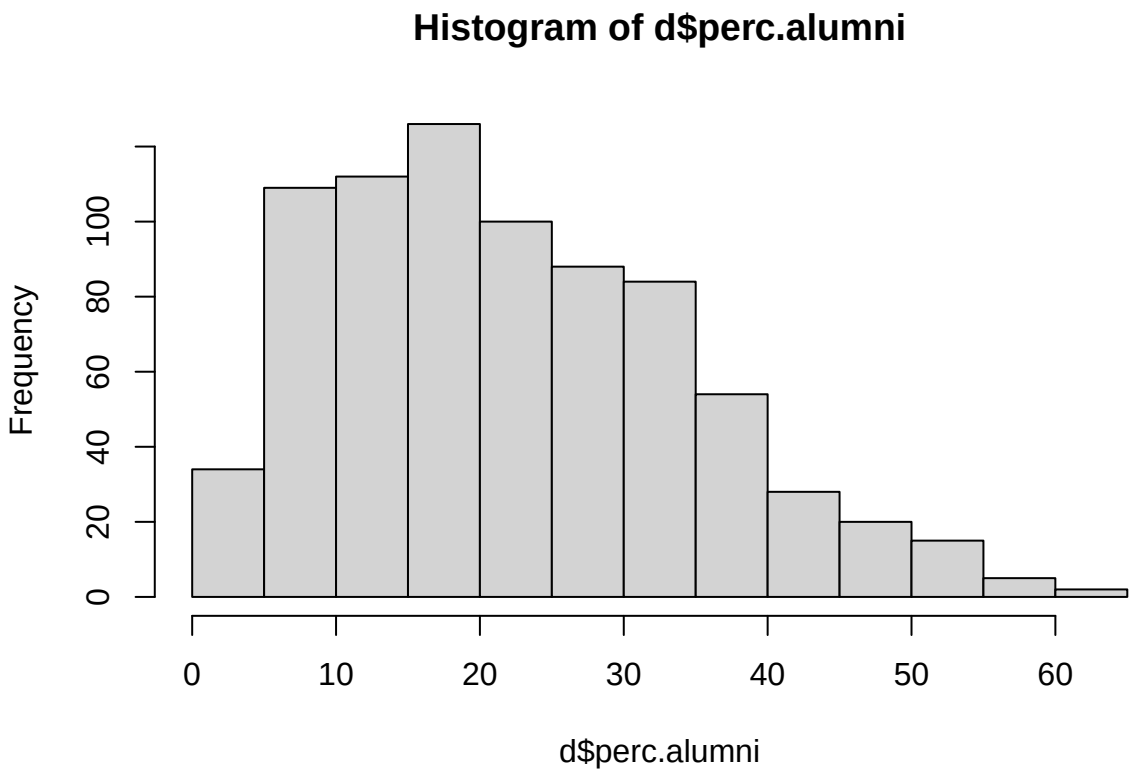
- very left skewed
- light and long right tail
- frequency increases with value (the higher the more common perhaps?)

```
# S.F.Ratio
hist(d$S.F.Ratio)
```



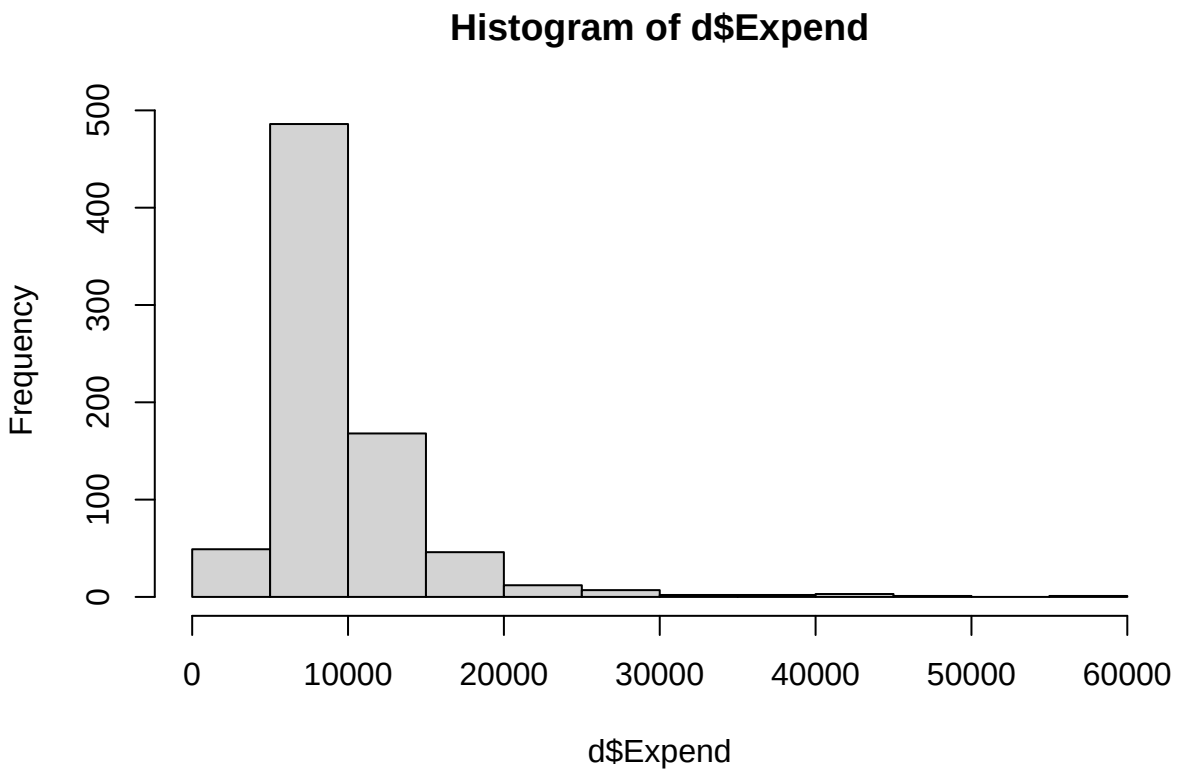
- most values between 10-20
- light and long left tail
- light right tail
- resembles gaussian distribution

```
# perc.alumni
hist(d$perc.alumni)
```



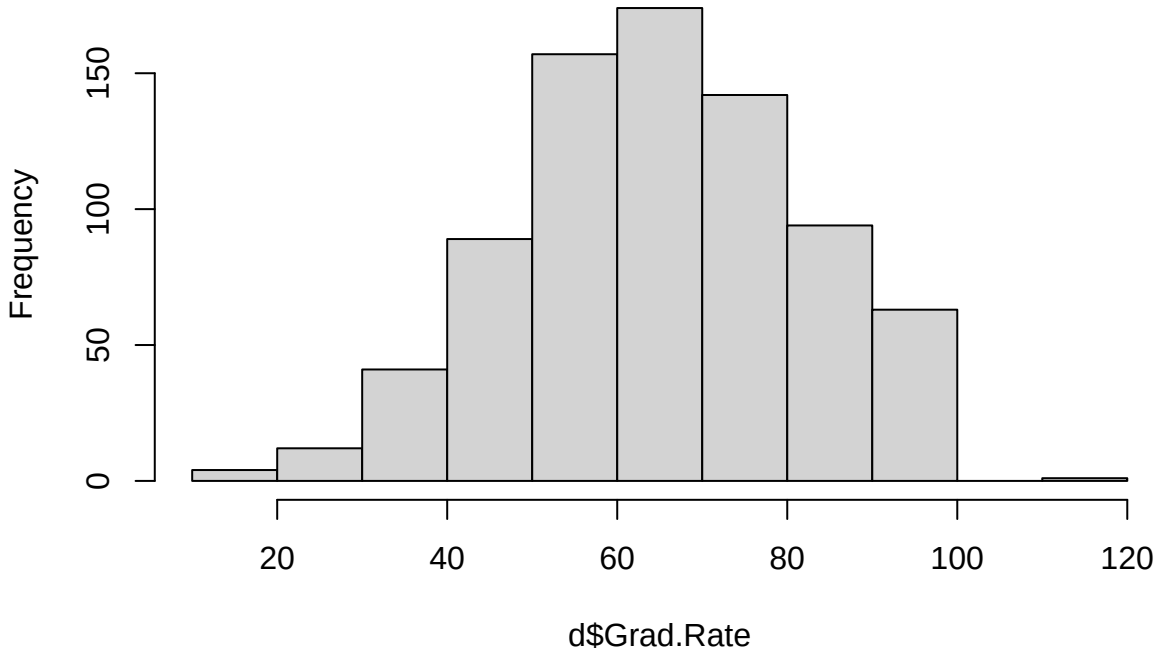
- right skewed
- long and light left tail
- relatively flat around the highest frequency part

```
# Expend
hist(d$Expend)
```



- very right skewed
- very long left tail
- most values around 10000
- resembles gamma distribution

```
# Grad.Rate
hist(d$Grad.Rate)
```



- quite resembles a gaussian distribution with mean between 50-70
- light leftmost bin frequency

Boxplots

```
# setting plot sizes
options(repr.plot.width = 18, repr.plot.height = 10)
```

```
# numeric columns
cols = colnames(d)[colnames(d) != 'X' & colnames(d) != 'Private']

# maximum per column
sapply(d %>% dplyr::select(cols), max, na.rm = TRUE)
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(cols)` instead of `cols` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

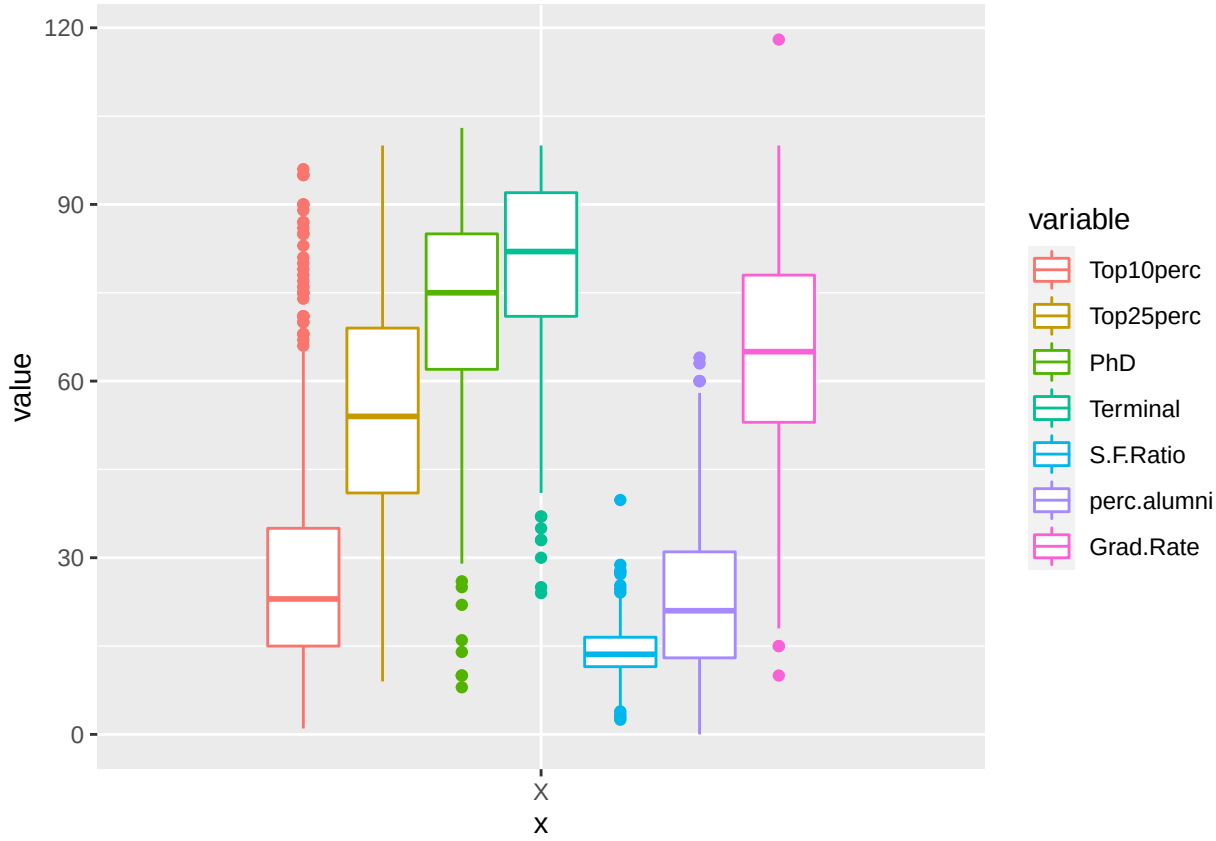
```
##        Apps      Accept      Enroll   Top10perc   Top25perc F.Undergrad
##     48094.0     26330.0      6392.0        96.0       100.0     31643.0
## P.Undergrad    Outstate  Room.Board       Books    Personal         PhD
##     21836.0     21700.0      8124.0      2340.0      6800.0       103.0
##    Terminal   S.F.Ratio perc.alumni      Expend   Grad.Rate
##       100.0        39.8        64.0     56233.0       118.0
```

```
# low val cols
low_val_c = c('Top10perc','Top25perc','PhD','Terminal','S.F.Ratio','perc.alumni','Grad.Rate')
low_val = melt(d, id.vars='X', measure.vars=low_val_c)

# med val cols
```

```
med_val_c = c('Enroll','Room.Board','Books','Personal')
med_val = melt(d, id.vars='X', measure.vars=med_val_c)

# high val cols
high_val_c = c('Apps','Accept','F.Undergrad','P.Undergrad','Outstate','Expend')
high_val = melt(d, id.vars='X', measure.vars=high_val_c)
```

Variables with low values (maximum under 120)

```
low_val %>%
    ggplot(aes(x="X", y=value)) +
    geom_boxplot(aes(color=variable))
```




```
# quantiles
sapply(d %>% dplyr::select(low_val_c), quantile, na.rm = TRUE)
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(low_val_c)` instead of `low_val_c` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
##      Top10perc Top25perc PhD Terminal S.F.Ratio perc.alumni Grad.Rate
## 0%           1         9   8       24       2.5           0        10
## 25%         15        41  62       71      11.5          13        53
## 50%         23        54  75       82      13.6          21        65
## 75%         35        69  85       92      16.5          31        78
## 100%        96       100 103      100      39.8          64       118
```

For the low value variables we can notice a few things:

- Top10perc

In our boxplot we can see that the median for this variable is 23, the maximum is 96 and the minimum is 1 with the 25% and 75% percentiles sitting at 15 and 35 respectively.

The range for this variable is of 95 and the IQR (interquartile range) is equal to 20.

- Top25perc

In our boxplot we can see that the median for this variable is 54, the maximum is 100 and the minimum is 9 with the 25% and 75% percentiles sitting at 41 and 69 respectively.

The range for this variable is of 91 and the IQR is equal to 28

- PhD

In our boxplot we can see that the median for this variable is 75, the maximum is 103 and the minimum is 8 with the 25% and 75% percentiles sitting at 62 and 85 respectively.

The range for this variable is of 95 and the IQR is equal to 23

- Terminal

In our boxplot we can see that the median for this variable is 82, the maximum is 100 and the minimum is 24 with the 25% and 75% percentiles sitting at 71 and 92 respectively.

The range for this variable is of 76 and the IQR is equal to 21

- S.F.Ratio

In our boxplot we can see that the median for this variable is 13.6, the maximum is 39.8 and the minimum is 2.5 with the 25% and 75% percentiles sitting at 11.5 and 16.5 respectively.

The range for this variable is of 37.3 and the IQR is equal to 5

- perc.alumni

In our boxplot we can see that the median for this variable is 21, the maximum is 64 and the minimum is 0 with the 25% and 75% percentiles sitting at 13 and 31 respectively.

The range for this variable is of 64 and the IQR is equal to 18

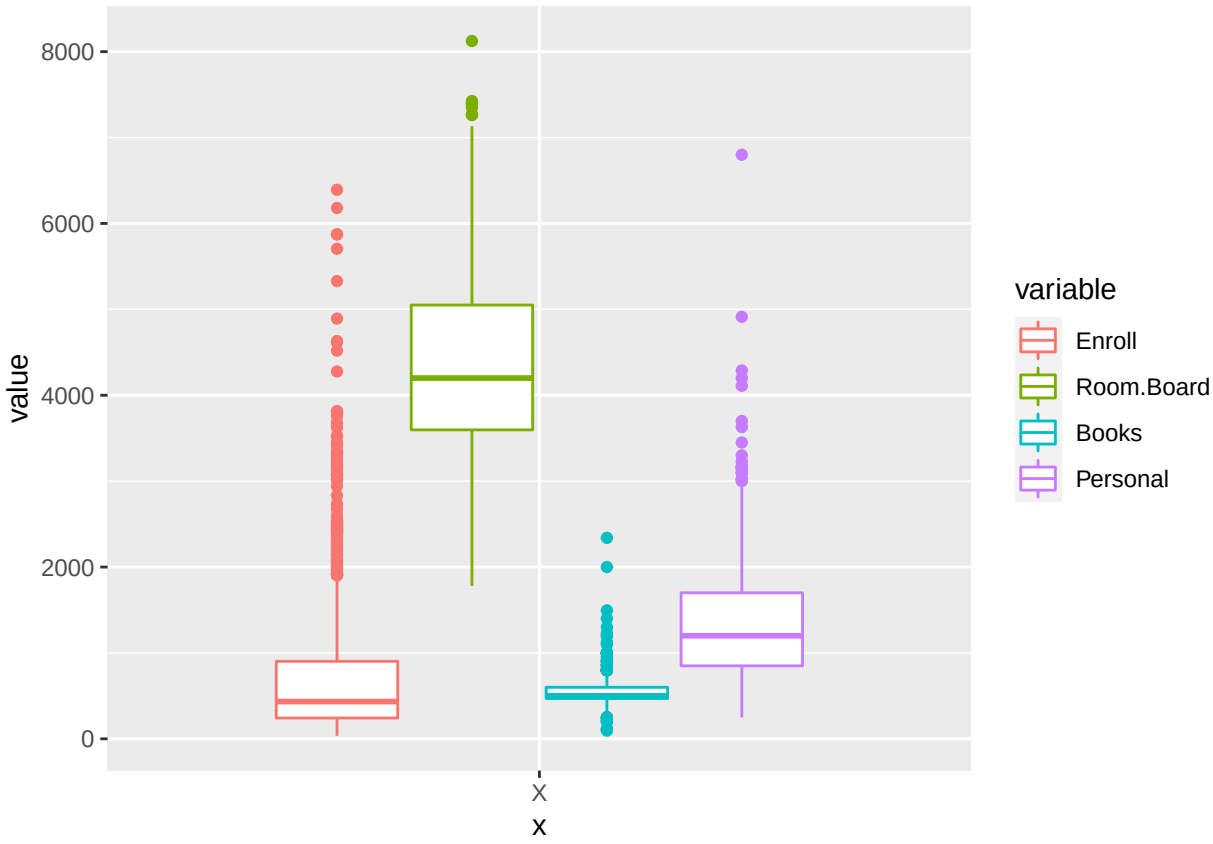
- Grad.Rate

In our boxplot we can see that the median for this variable is 65, the maximum is 118 and the minimum is 10 with the 25% and 75% percentiles sitting at 53 and 78 respectively.

The range for this variable is of 108 and the IQR is equal to 25

Variables with medium values (maximum less than or equal to 8124 and not in low values list)

```
med_val %>%
    ggplot(aes(x="X", y=value)) +
    geom_boxplot(aes(color=variable))
```




```
# quantiles
sapply(d %>% dplyr::select(med_val_c), quantile, na.rm = TRUE)
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(med_val_c)` instead of `med_val_c` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
##      Enroll Room.Board Books Personal
## 0%       35       1780    96      250
## 25%     242       3597   470      850
## 50%     434       4200   500     1200
## 75%     902       5050   600     1700
## 100%   6392       8124  2340     6800
```

For the medium value variables we can notice a few things:

- Enroll

In our boxplot we can see that the median for this variable is 494, the maximum is 6392 and the minimum is 35 with the 25% and 75% percentiles sitting at 242 and 902 respectively.

The range for this variable is of 6357 and the IQR (interquartile range) is equal to 660.

- Room.Board

In our boxplot we can see that the median for this variable is 4200, the maximum is 8124 and the minimum is 1780 with the 25% and 75% percentiles sitting at 3597 and 5050 respectively.

The range for this variable is of 6344 and the IQR is equal to 1453.

- Books

In our boxplot we can see that the median for this variable is 500, the maximum is 2340 and the minimum is 96 with the 25% and 75% percentiles sitting at 470 and 600 respectively.

The range for this variable is of 2244 and the IQR is equal to 130

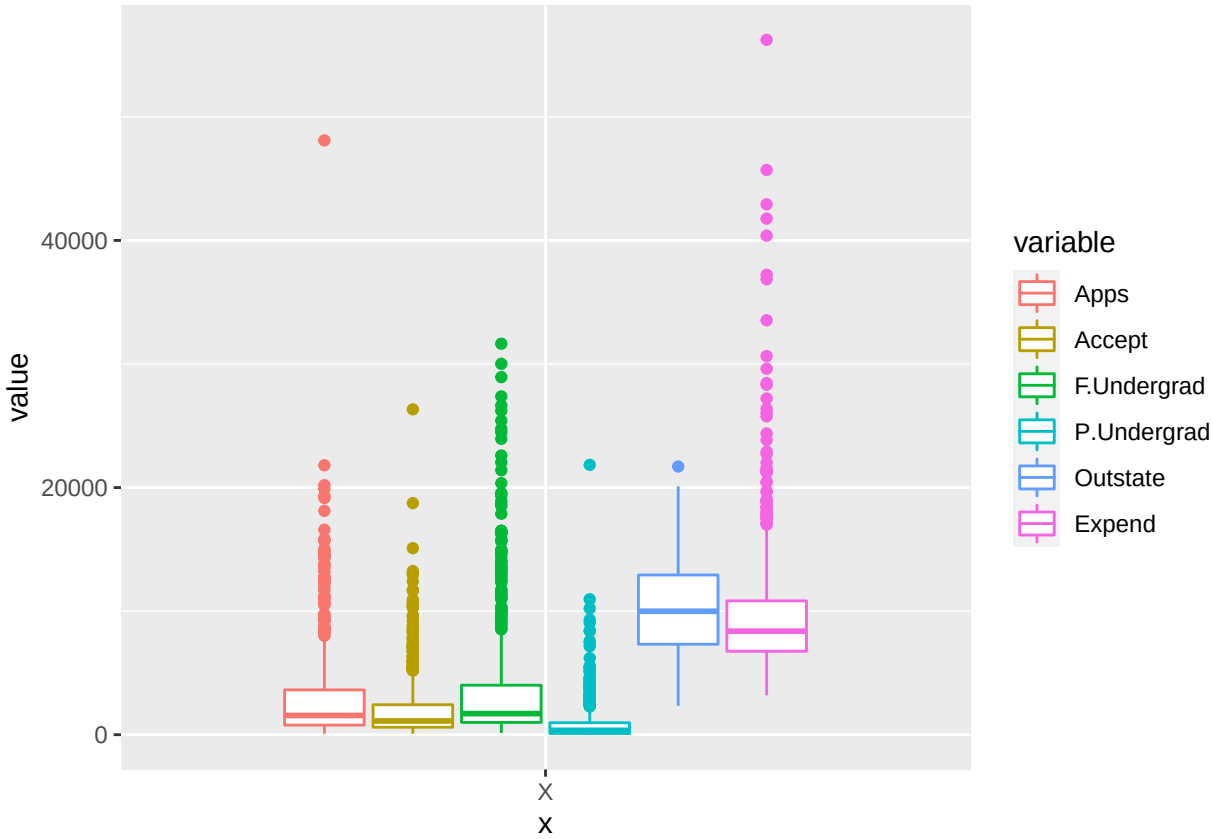
- Personal

In our boxplot we can see that the median for this variable is 1200, the maximum is 6800 and the minimum is 250 with the 25% and 75% percentiles sitting at 850 and 1700 respectively.

The range for this variable is of 6550 and the IQR is equal to 850

Variables with high values (maximums above 8124)

```
high_val %>%
    ggplot(aes(x="X", y=value)) +
    geom_boxplot(aes(color=variable))
```




```
# quantiles
sapply(d %>% dplyr::select(high_val_c), quantile, na.rm = TRUE)
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(high_val_c)` instead of `high_val_c` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
##       Apps Accept F.Undergrad P.Undergrad Outstate Expend
## 0%      81     72         139           1     2340   3186
## 25%    776    604         992          95     7320   6751
## 50%   1558   1110        1707         353     9990   8377
## 75%   3624   2424        4005         967    12925  10830
```

```
## 100% 48094  26330       31643       21836    21700  56233
```

For the high value variables we can notice a few things:

- Apps

In our boxplot we can see that the median for this variable is 1558, the maximum is 48094 and the minimum is 81 with the 25% and 75% percentiles sitting at 776 and 3624 respectively.

The range for this variable is of 48013 and the IQR (interquartile range) is equal to 2848.

- Accept

In our boxplot we can see that the median for this variable is 1110, the maximum is 26330 and the minimum is 72 with the 25% and 75% percentiles sitting at 604 and 2424 respectively.

The range for this variable is of 26258 and the IQR is equal to 1820.

- F.Undergrad

In our boxplot we can see that the median for this variable is 1707, the maximum is 31643 and the minimum is 139 with the 25% and 75% percentiles sitting at 992 and 4005 respectively.

The range for this variable is of 31504 and the IQR is equal to 3013

- P.Undergrad

In our boxplot we can see that the median for this variable is 353, the maximum is 21836 and the minimum is 1 with the 25% and 75% percentiles sitting at 95 and 967 respectively.

The range for this variable is of 21835 and the IQR is equal to 872

- Outstate

In our boxplot we can see that the median for this variable is 9990, the maximum is 21700 and the minimum is 2340 with the 25% and 75% percentiles sitting at 7320 and 12925 respectively.

The range for this variable is of 19360 and the IQR is equal to 5605

- Expend

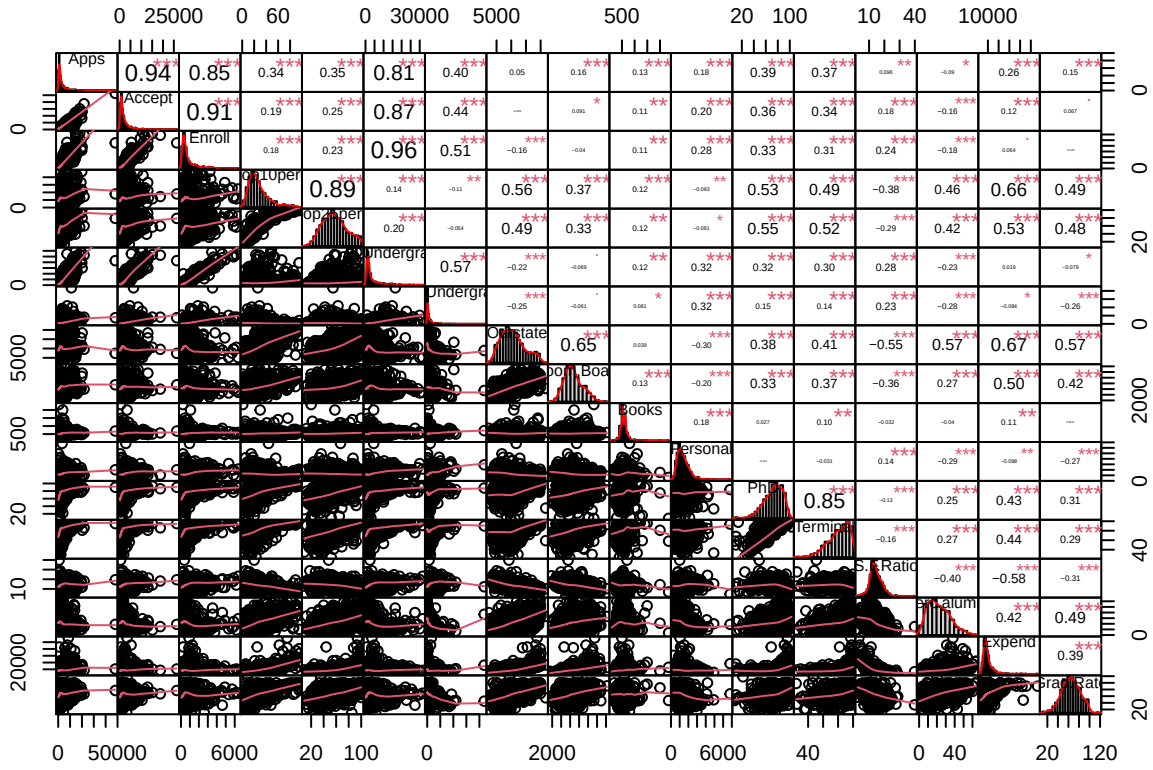
In our boxplot we can see that the median for this variable is 8377, the maximum is 56233 and the minimum is 3186 with the 25% and 75% percentiles sitting at 6751 and 10830 respectively.

The range for this variable is of 53047 and the IQR is equal to 4079

3- Perform a visual analysis of all quantitative variables together. Then, Perform a visual analysis of each of the quantitative variables taking into account the variable Private.

Correlation and scatterplots for all variables

```
pa <- d %>% dplyr::select(cols)
chart.Correlation(pa, histogram=TRUE, pch=19, method="pearson")
```



Here we can see a plot with all the numerical variables, a histogram and density plot in the main diagonal. To the bottom of the diagonal we see scatterplots for each pair of variables and a line that goes through the scatterplot. Above the main diagonal we see all the correlation coefficients for each pair of variables.

We see there are some highly correlated variables and some others which have negligible correlation.

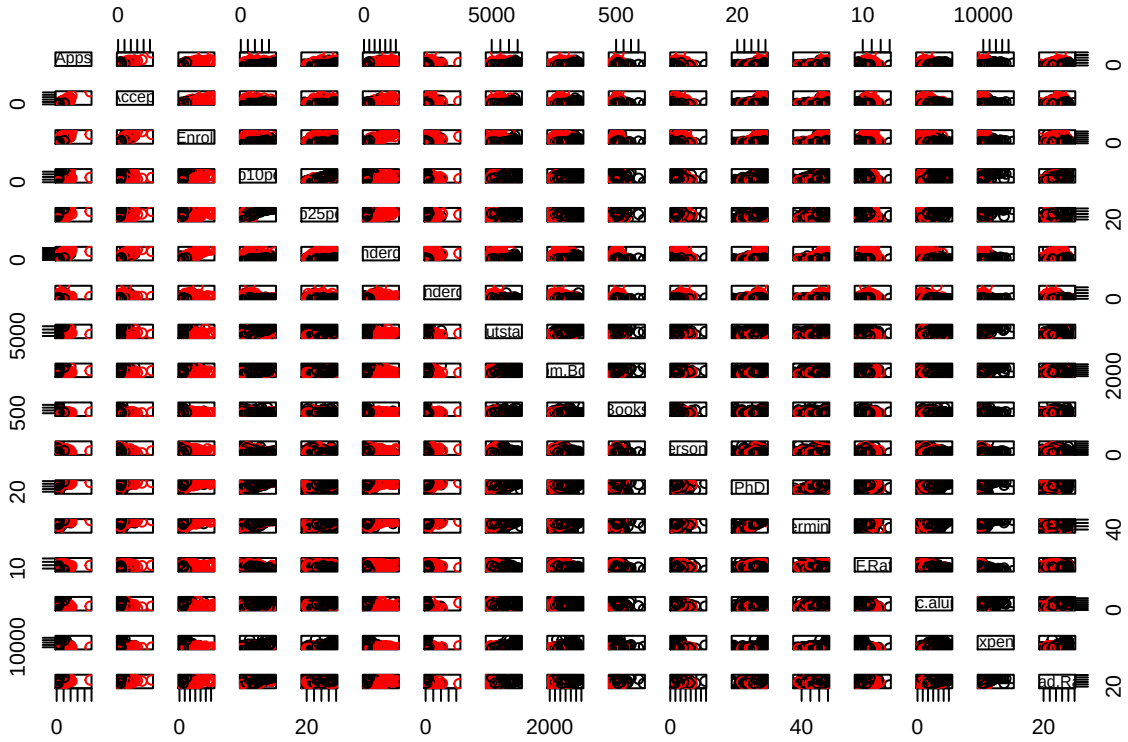
All correlation coefficients are Pearson correlation coefficients.

The variables with the highest correlation are the following:

- F.Undergrad vs Enroll with very high correlation of 0.96
- Apps vs Accept with a similarly high correlation coefficient of 0.94
- Accept vs Enroll also have a very high correlation of 0.91
- Top10perc vs Top25perc have a correlation of 0.89
- F.Undergrad vs Accept have a correlation of 0.
- Both Apps vs Enroll and Phd vs Terminal have a correlation of 0.85

```
# setting colors
color_1 <- "black"
color_2 <- "red"
colors <- as.numeric(as.factor(d$Private)) - 1
i <- 1
for (num in colors) {
    colors[i] = ifelse(num==1, color_1, color_2)
    i <- i + 1
}
```

```
pa <- d %>% dplyr::select(cols)
pairs(pa,pch=1,col=colors)
```



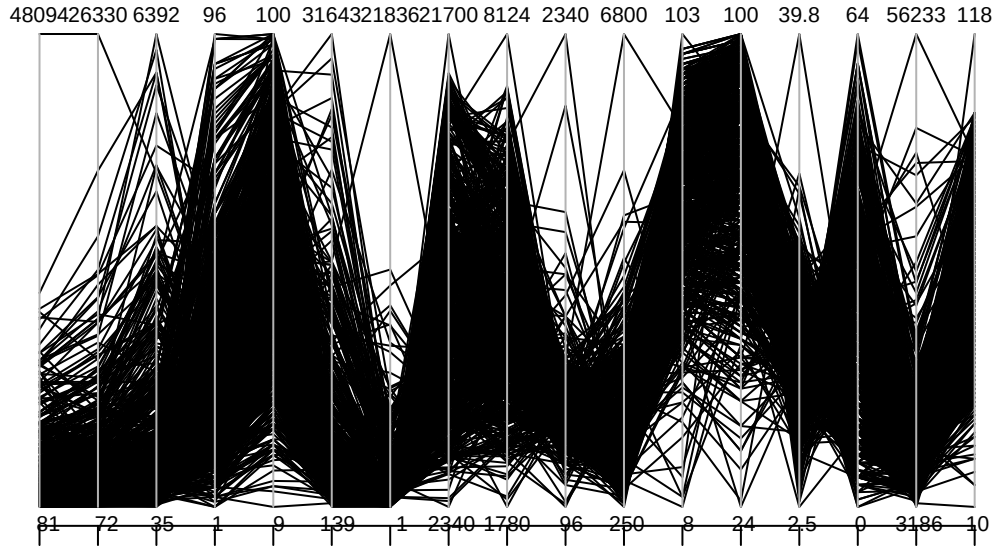
In most of these scatter plots we see an interesting pattern, where for varibles plotted versus Top10perc, Top20perc, PhD and Terminal there seems to be no clear divide between each one ofg the groups, except when compared to variables that significantly diffentiate both groups.

The clear differentiator variables seem to be P.Undergrad, F.Undergrad, S.F.Ratio, Apps, Accept, Enrool and outstate.

And sure, I could definitely agree that the size of each plot might be overestimating such division in groups, but some of these variables do show a stark difference in values.

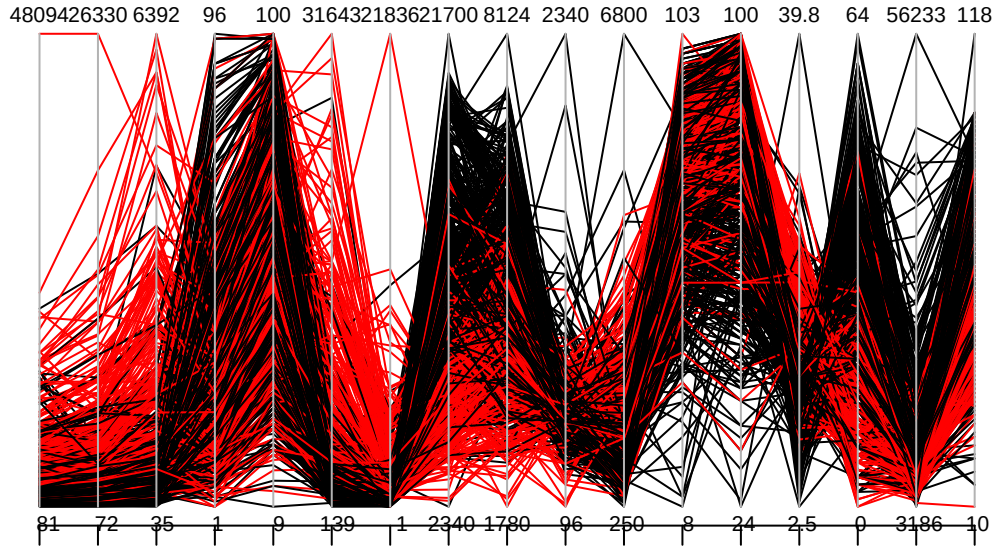
PCP

```
parcoord(pa,var.label=TRUE)
```



We can basically just see the spread of the data here, the only thing I take from it is that apps and accept have really very few outliers in terms of quantity, but they're significantly larger than the rest. Shows how skewed these variables are.

```
parcoord(pa,var.label=TRUE, col=colors)
```



We can see that the more extreme variables of Apps, Accept, P.Undergrad, F.Undergrad seem to be affected by whether it's private or not.

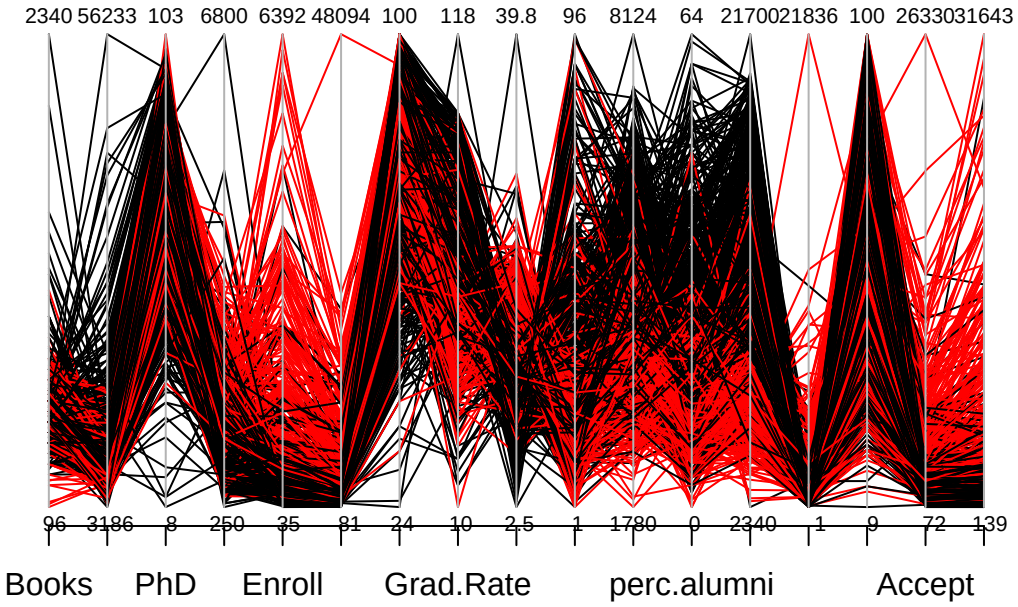
In the case of outstate too, but instead of being the highest values, the values where it's private correspond to the lowest values.

For Grad.Rate it also seem somewhat divided and in a similar way for perc.alumni.

For Top10perc, Top25perc, PhD or Terminal there seems to be no noticeable difference made by whether it's private or not.

Randomizing the order of the variables we get a different but perhaps better PCP.

```
cols <- sample(cols)
pa <- pa %>% dplyr::select(cols)
parcoord(pa,var.label=TRUE, col=colors)
```



Here we can more clearly see that for S.F.Ratio there is a stark difference between both groups, where the group labeled as Private seems to have the higher values.

For the rest of the variables we get much of the same we explained earlier.

If we were to line up variables with similarly shaped histograms, we could get an interesting PCP, let's try that.

We will exclude Top10perc and Top25perc as these do not contribute too much to the plot

```
# right skewed variables
right_skewed <- c('Apps','Accept','P.Undergrad','F.Undergrad','Personal','Expend','Books','perc.alumni'
right_skewed <- d %>% dplyr::select(right_skewed)
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(right_skewed)` instead of `right_skewed` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
# normal-like variables
normal_like <- c('Outstate','Room.Board','perc.alumni','Grad.Rate','S.F.Ratio')
normal_like <- d %>% dplyr::select(normal_like)
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(normal_like)` instead of `normal_like` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

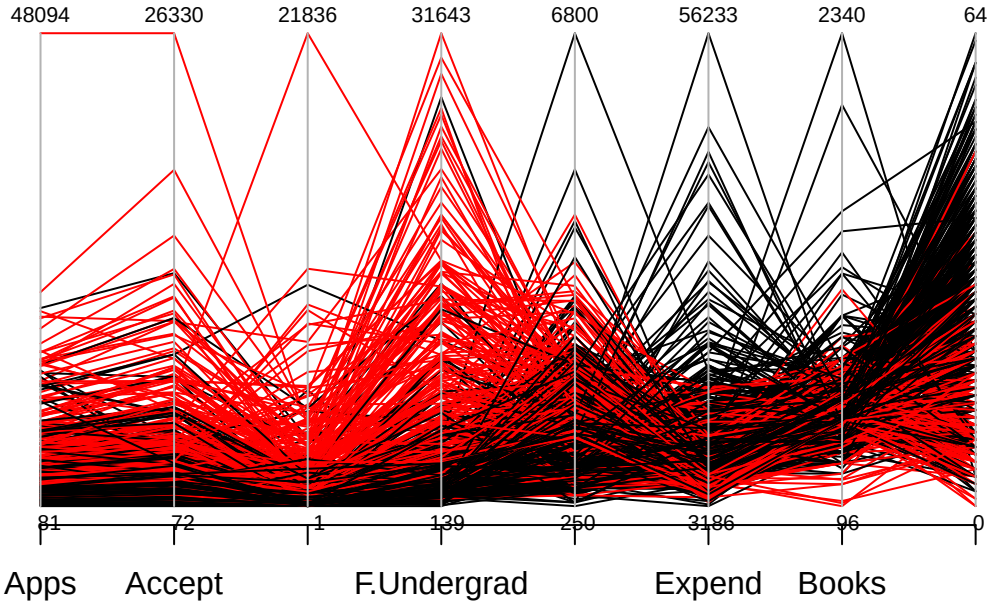
```
# left skewed variables
left_skewed <- c('PhD','Terminal')
left_skewed <- d %>% dplyr::select(left_skewed)
```

```
## Note: Using an external vector in selections is ambiguous.
```

```
## i Use `all_of(left_skewed)` instead of `left_skewed` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

right skewed variables PCP

```
parcoord(right_skewed,var.label=TRUE, col=colors)
```

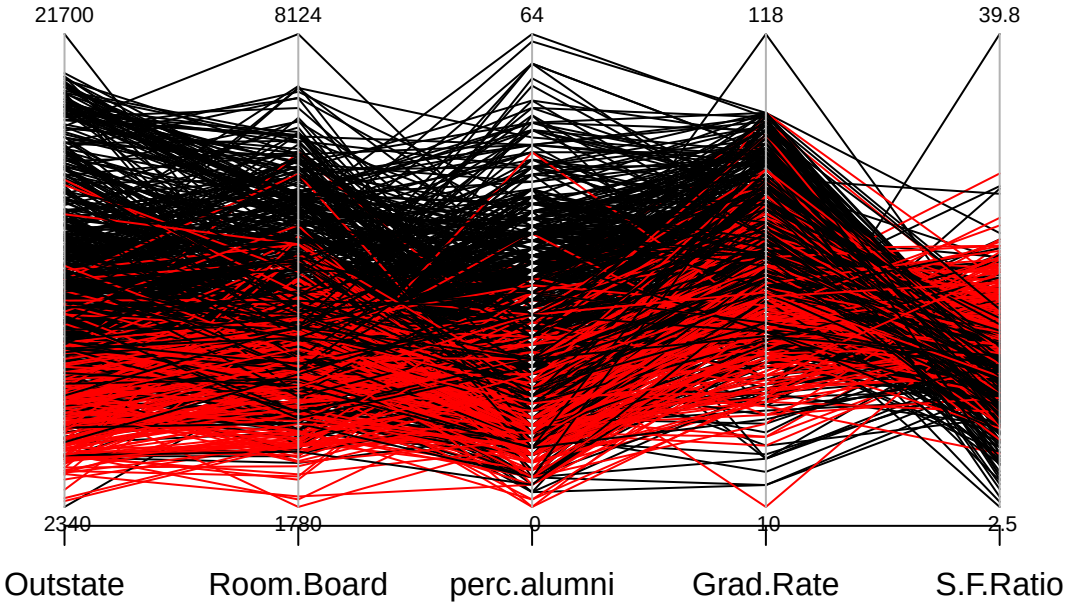


After grouping similarly shaped variables the plots are somewhat clearer. Confirming most of our previous observations.

The general idea is that for these variables, pretty much everything but perc.alumni, Books and Expend follow a pattern where the higher-than-usual variables pretty much belong completely to entries of Private colleges.

normal-like variables PCP

```
parcoord(normal_like,var.label=TRUE, col=colors)
```

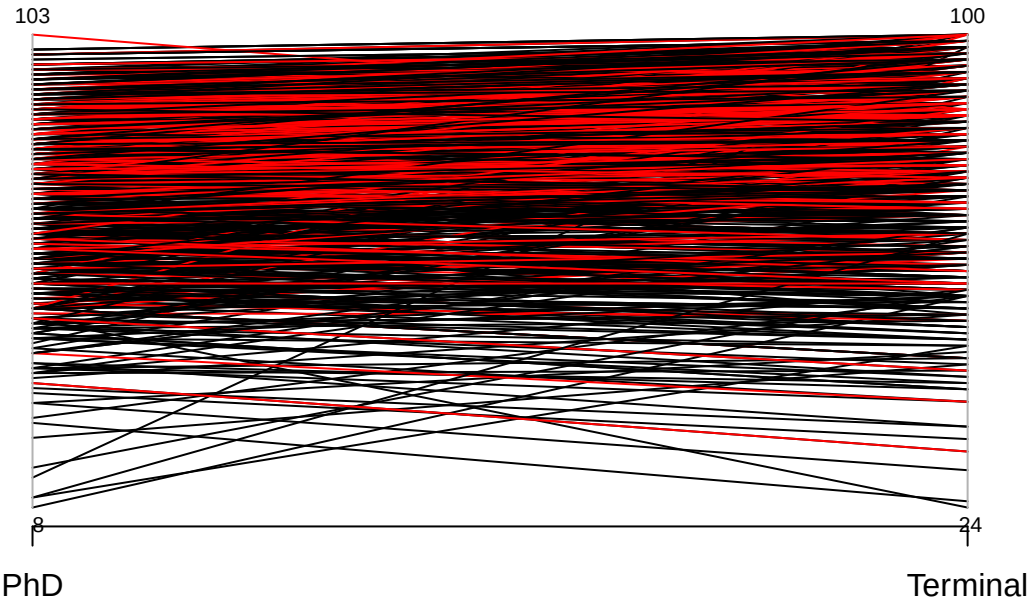


I particularly like this one. These variables look significantly clearer here than in the first PCP.

The first observations still hold though. For pretty much all these variables there's a significant divide in the same direction except for S.F.Ratio where that pattern is inverted.

left skewed variables PCP

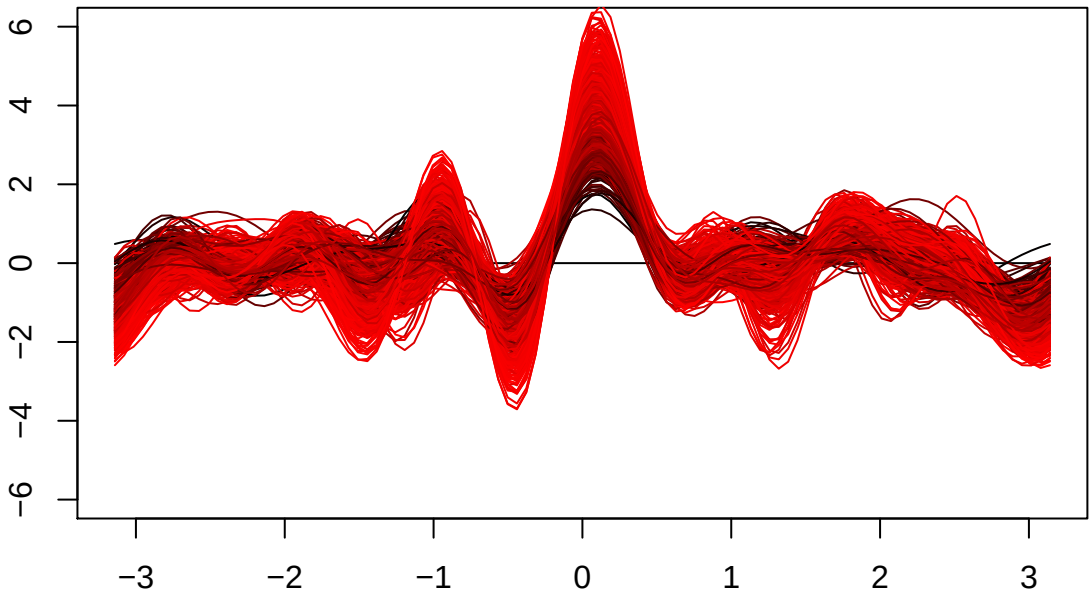
```
parcoord(left_skewed,var.label=TRUE, col=colors)
```



being private or not makes no difference here, therefore not very interesting.

Andrews' plot

```
par(mfrow=c(1,1))
andrews(as.data.frame(cbind(pa,as.factor(d$Private))),clr=7,ymax=6)
```



The Andrews' plot makes it difficult to appreciate a significant difference between both groups in my opinion. Perhaps I might need to read up more on it but it doesn't tell me too much about the values other than around the middle. Where the we have the 0 and around -2, where the groups seem more divided than everywhere else, taking different directions in the dataset.