pset3

Bhargavi Ganesh

10/25/2019

```r
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts -------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

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

```r
library(seriation)
```

```
## Registered S3 method overwritten by 'seriation':
##   method         from
##   reorder.hclust gclus
```

```r
library(clustertend)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'

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

```r
library(mixtools)
```

```
## mixtools package, version 1.1.0, Released 2017-03-10
## This package is based upon work supported by the National Science Foundation under Grant No. SES-0518
```

```r
library(cluster)
library(factoextra)
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

```r
library(plotGMM)
load("/Users/bhargaviganesh/Documents/Problem-Set-3/State Leg Prof Data & Codebook/legprof-components.v
```

Question 2: Data Munging

```r
x_subset <- x %>%
  filter(sessid == "2009/10") %>%
  mutate_if(is.numeric, scale) %>%
  select(state, slength, t_slength, salary_real, expend) %>%
  drop_na()

x_scaled_subset <- x_subset %>%
  select(-state)

x_state_names <- x_subset[,1]
```

Question 3: Exploratory Data Analysis

```r
kable(summary(x_scaled_subset))
```

	slength.V1	t_slength.V1	salary_real.V1	expend.V1
	Min. :-1.331915	Min. :-1.282138	Min. :-1.133635	Min. :-0.782725
	1st Qu.:-0.615579	1st Qu.:-0.599190	1st Qu.:-0.734287	1st Qu.:-0.543377
	Median :-0.210107	Median :-0.238210	Median :-0.315876	Median :-0.244480
	Mean : 0.000000	Mean : 0.000000	Mean :-0.018539	Mean :-0.002027
	3rd Qu.: 0.171443	3rd Qu.: 0.133236	3rd Qu.: 0.436462	3rd Qu.:-0.024684
	Max. : 3.900711	Max. : 3.691295	Max. : 3.193723	Max. : 5.532723

The summary statistics above show that after scaling the variables such that the mean is 0 and the standard deviation is 1. The medians of all of the variables are below 0, suggesting that the distribution is skewed for all of the variables. The histograms below confirm this.

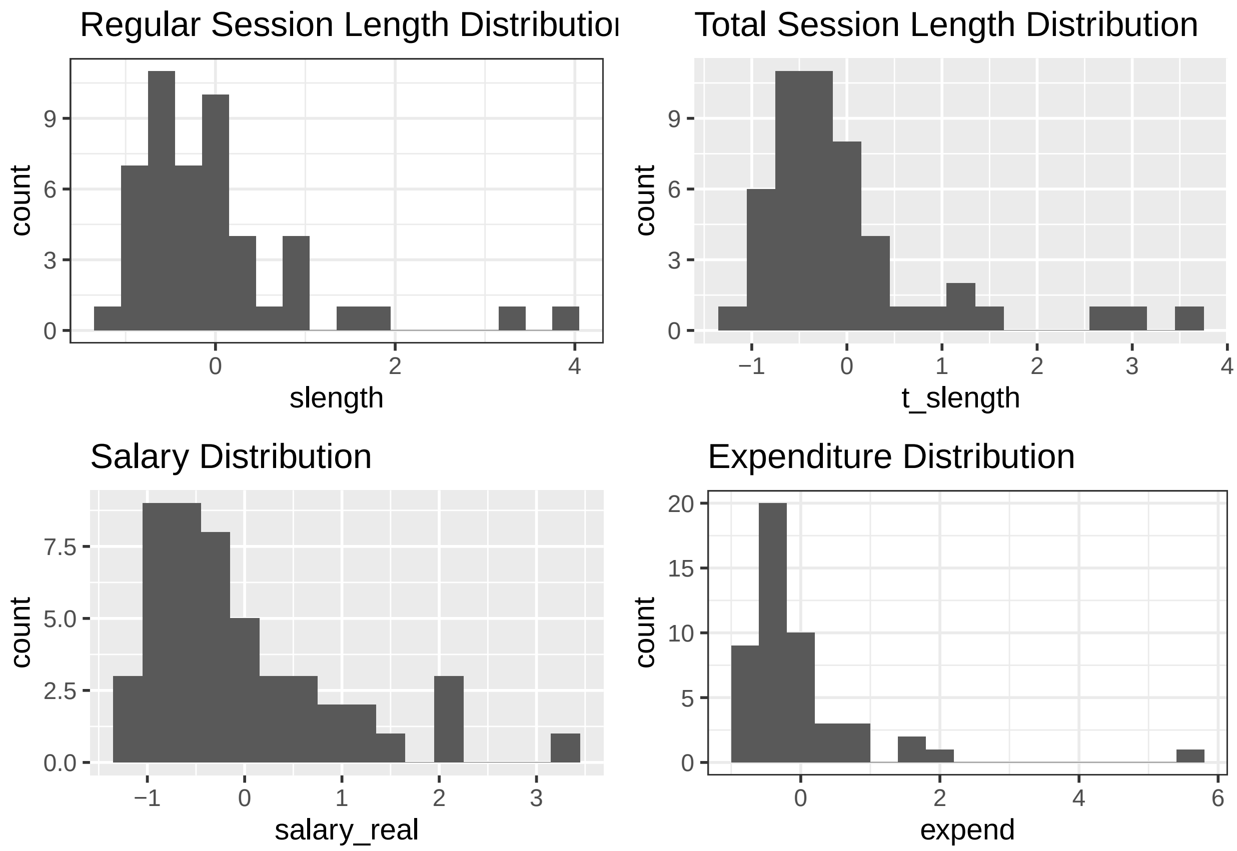
```r
hist_slength <- x_scaled_subset %>%
  ggplot() +
  geom_histogram(aes(x = slength), binwidth = 0.3) +
  labs(title = " Regular Session Length Distribution") +
  theme_bw()

hist_t_slength <- x_scaled_subset %>%
  ggplot() +
  geom_histogram(aes(x = t_slength), binwidth = 0.3) +
  labs(title = "Total Session Length Distribution")
  theme_bw()
```

```
## List of 65
##  $ line                      :List of 6
##   ..$ colour       : chr "black"
##   ..$ size         : num 0.5
##   ..$ linetype     : num 1
##   ..$ lineend      : chr "butt"
##   ..$ arrow        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_line" "element"
##  $ rect                      :List of 5
##   ..$ fill         : chr "white"
##   ..$ colour       : chr "black"
##   ..$ size         : num 0.5
##   ..$ linetype     : num 1
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ text                      :List of 11
##   ..$ family       : chr ""
##   ..$ face         : chr "plain"
##   ..$ colour       : chr "black"
##   ..$ size         : num 11
##   ..$ hjust        : num 0.5
##   ..$ vjust        : num 0.5
##   ..$ angle        : num 0
##   ..$ lineheight   : num 0.9
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.title.x              :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 2.75pt 0pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.title.x.top          :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 0
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 2.75pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.title.y              :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 1
##   ..$ angle        : num 90
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 2.75pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.title.y.right        :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 0
##   ..$ angle        : num -90
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 0pt 2.75pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text                 :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : chr "grey30"
##   ..$ size         : 'rel' num 0.8
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text.x               :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 2.2pt 0pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text.x.top           :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 0
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 2.2pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text.y               :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 1
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 2.2pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text.y.right         :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 0
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 0pt 2.2pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.ticks                :List of 6
##   ..$ colour       : chr "grey20"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ lineend      : NULL
##   ..$ arrow        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_line" "element"
##  $ axis.ticks.length         : 'unit' num 2.75pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ axis.ticks.length.x       : NULL
##  $ axis.ticks.length.x.top   : NULL
##  $ axis.ticks.length.x.bottom: NULL
##  $ axis.ticks.length.y       : NULL
##  $ axis.ticks.length.y.left  : NULL
##  $ axis.ticks.length.y.right : NULL
##  $ axis.line                 : list()
##   ..- attr(*, "class")= chr [1:2] "element_blank" "element"
##  $ axis.line.x               : NULL
##  $ axis.line.y               : NULL
##  $ legend.background         :List of 5
##   ..$ fill         : NULL
##   ..$ colour       : logi NA
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ legend.margin             : 'margin' num [1:4] 5.5pt 5.5pt 5.5pt 5.5pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ legend.spacing            : 'unit' num 11pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ legend.spacing.x          : NULL
##  $ legend.spacing.y          : NULL
##  $ legend.key                :List of 5
##   ..$ fill         : chr "white"
##   ..$ colour       : logi NA
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ legend.key.size           : 'unit' num 1.2lines
##   ..- attr(*, "valid.unit")= int 3
##   ..- attr(*, "unit")= chr "lines"
##  $ legend.key.height         : NULL
##  $ legend.key.width          : NULL
##  $ legend.text               :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 0.8
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ legend.text.align         : NULL
##  $ legend.title              :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 0
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ legend.title.align        : NULL
##  $ legend.position           : chr "right"
##  $ legend.direction          : NULL
##  $ legend.justification      : chr "center"
##  $ legend.box                : NULL
##  $ legend.box.margin         : 'margin' num [1:4] 0cm 0cm 0cm 0cm
##   ..- attr(*, "valid.unit")= int 1
##   ..- attr(*, "unit")= chr "cm"
##  $ legend.box.background     : list()
##   ..- attr(*, "class")= chr [1:2] "element_blank" "element"
##  $ legend.box.spacing        : 'unit' num 11pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ panel.background          :List of 5
##   ..$ fill         : chr "white"
##   ..$ colour       : logi NA
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ panel.border              :List of 5
##   ..$ fill         : logi NA
##   ..$ colour       : chr "grey20"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ panel.spacing             : 'unit' num 5.5pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ panel.spacing.x           : NULL
##  $ panel.spacing.y           : NULL
##  $ panel.grid                :List of 6
##   ..$ colour       : chr "grey92"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ lineend      : NULL
##   ..$ arrow        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_line" "element"
##  $ panel.grid.minor          :List of 6
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 0.5
##   ..$ linetype     : NULL
##   ..$ lineend      : NULL
##   ..$ arrow        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_line" "element"
##  $ panel.ontop               : logi FALSE
##  $ plot.background           :List of 5
##   ..$ fill         : NULL
##   ..$ colour       : chr "white"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ plot.title                :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 1.2
##   ..$ hjust        : num 0
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 5.5pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ plot.subtitle             :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 0
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 5.5pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ plot.caption              :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 0.8
##   ..$ hjust        : num 1
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 5.5pt 0pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ plot.tag                  :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 1.2
##   ..$ hjust        : num 0.5
##   ..$ vjust        : num 0.5
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ plot.tag.position         : chr "topleft"
##  $ plot.margin               : 'margin' num [1:4] 5.5pt 5.5pt 5.5pt 5.5pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ strip.background          :List of 5
##   ..$ fill         : chr "grey85"
##   ..$ colour       : chr "grey20"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ strip.placement           : chr "inside"
##  $ strip.text                :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : chr "grey10"
##   ..$ size         : 'rel' num 0.8
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 4.4pt 4.4pt 4.4pt 4.4pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ strip.text.x              : NULL
##  $ strip.text.y              :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : num -90
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ strip.switch.pad.grid     : 'unit' num 2.75pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ strip.switch.pad.wrap     : 'unit' num 2.75pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi TRUE
##  - attr(*, "validate")= logi TRUE
```

```
hist_salary_real <- x_scaled_subset %>%
  ggplot() +
  geom_histogram(aes(x = salary_real), binwidth = 0.3) +
  labs(title = "Salary Distribution")
  theme_bw()
```

```
## List of 65
##  $ line                      :List of 6
##   ..$ colour       : chr "black"
##   ..$ size         : num 0.5
##   ..$ linetype     : num 1
##   ..$ lineend      : chr "butt"
##   ..$ arrow        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_line" "element"
##  $ rect                      :List of 5
##   ..$ fill         : chr "white"
##   ..$ colour       : chr "black"
##   ..$ size         : num 0.5
##   ..$ linetype     : num 1
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ text                      :List of 11
##   ..$ family       : chr ""
##   ..$ face         : chr "plain"
##   ..$ colour       : chr "black"
##   ..$ size         : num 11
##   ..$ hjust        : num 0.5
##   ..$ vjust        : num 0.5
##   ..$ angle        : num 0
##   ..$ lineheight   : num 0.9
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.title.x              :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 2.75pt 0pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.title.x.top          :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 0
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 2.75pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.title.y              :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 1
##   ..$ angle        : num 90
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 2.75pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.title.y.right        :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 0
##   ..$ angle        : num -90
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 0pt 2.75pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text                 :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : chr "grey30"
##   ..$ size         : 'rel' num 0.8
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text.x               :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 2.2pt 0pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text.x.top           :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 0
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 2.2pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text.y               :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 1
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 2.2pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text.y.right         :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 0
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 0pt 2.2pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.ticks                :List of 6
##   ..$ colour       : chr "grey20"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ lineend      : NULL
##   ..$ arrow        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_line" "element"
##  $ axis.ticks.length         : 'unit' num 2.75pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ axis.ticks.length.x       : NULL
##  $ axis.ticks.length.x.top   : NULL
##  $ axis.ticks.length.x.bottom: NULL
##  $ axis.ticks.length.y       : NULL
##  $ axis.ticks.length.y.left  : NULL
##  $ axis.ticks.length.y.right : NULL
##  $ axis.line                 : list()
##   ..- attr(*, "class")= chr [1:2] "element_blank" "element"
##  $ axis.line.x               : NULL
##  $ axis.line.y               : NULL
##  $ legend.background         :List of 5
##   ..$ fill         : NULL
##   ..$ colour       : logi NA
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ legend.margin             : 'margin' num [1:4] 5.5pt 5.5pt 5.5pt 5.5pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ legend.spacing            : 'unit' num 11pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ legend.spacing.x          : NULL
##  $ legend.spacing.y          : NULL
##  $ legend.key                :List of 5
##   ..$ fill         : chr "white"
##   ..$ colour       : logi NA
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ legend.key.size           : 'unit' num 1.2lines
##   ..- attr(*, "valid.unit")= int 3
##   ..- attr(*, "unit")= chr "lines"
##  $ legend.key.height         : NULL
##  $ legend.key.width          : NULL
##  $ legend.text               :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 0.8
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ legend.text.align         : NULL
##  $ legend.title              :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 0
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ legend.title.align        : NULL
##  $ legend.position           : chr "right"
##  $ legend.direction          : NULL
##  $ legend.justification      : chr "center"
##  $ legend.box                : NULL
##  $ legend.box.margin         : 'margin' num [1:4] 0cm 0cm 0cm 0cm
##   ..- attr(*, "valid.unit")= int 1
##   ..- attr(*, "unit")= chr "cm"
##  $ legend.box.background     : list()
##   ..- attr(*, "class")= chr [1:2] "element_blank" "element"
##  $ legend.box.spacing        : 'unit' num 11pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ panel.background          :List of 5
##   ..$ fill         : chr "white"
##   ..$ colour       : logi NA
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ panel.border              :List of 5
##   ..$ fill         : logi NA
##   ..$ colour       : chr "grey20"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ panel.spacing             : 'unit' num 5.5pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ panel.spacing.x           : NULL
##  $ panel.spacing.y           : NULL
##  $ panel.grid                :List of 6
##   ..$ colour       : chr "grey92"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ lineend      : NULL
##   ..$ arrow        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_line" "element"
##  $ panel.grid.minor          :List of 6
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 0.5
##   ..$ linetype     : NULL
##   ..$ lineend      : NULL
##   ..$ arrow        : logi FALSE
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_line" "element"
##  $ panel.ontop               : logi FALSE
##  $ plot.background           :List of 5
##   ..$ fill         : NULL
##   ..$ colour       : chr "white"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ plot.title                :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 1.2
##   ..$ hjust        : num 0
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 5.5pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ plot.subtitle             :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 0
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 0pt 0pt 5.5pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ plot.caption              :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 0.8
##   ..$ hjust        : num 1
##   ..$ vjust        : num 1
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 5.5pt 0pt 0pt 0pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ plot.tag                  :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : 'rel' num 1.2
##   ..$ hjust        : num 0.5
##   ..$ vjust        : num 0.5
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ plot.tag.position         : chr "topleft"
##  $ plot.margin               : 'margin' num [1:4] 5.5pt 5.5pt 5.5pt 5.5pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ strip.background          :List of 5
##   ..$ fill         : chr "grey85"
##   ..$ colour       : chr "grey20"
##   ..$ size         : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ strip.placement           : chr "inside"
##  $ strip.text                :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : chr "grey10"
##   ..$ size         : 'rel' num 0.8
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : 'margin' num [1:4] 4.4pt 4.4pt 4.4pt 4.4pt
##   .. ..- attr(*, "valid.unit")= int 8
##   .. ..- attr(*, "unit")= chr "pt"
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ strip.text.x              : NULL
##  $ strip.text.y              :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : num -90
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi TRUE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ strip.switch.pad.grid     : 'unit' num 2.75pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  $ strip.switch.pad.wrap     : 'unit' num 2.75pt
##   ..- attr(*, "valid.unit")= int 8
##   ..- attr(*, "unit")= chr "pt"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi TRUE
##  - attr(*, "validate")= logi TRUE
```

```
hist_expend <- x_scaled_subset %>%
  ggplot() +
  geom_histogram(aes(x = expend), binwidth = 0.4) +
  labs(title = "Expenditure Distribution") +
  theme_bw()

grid.arrange(hist_slength, hist_t_slength, hist_salary_real, hist_expend, ncol=2, nrow=2)
```



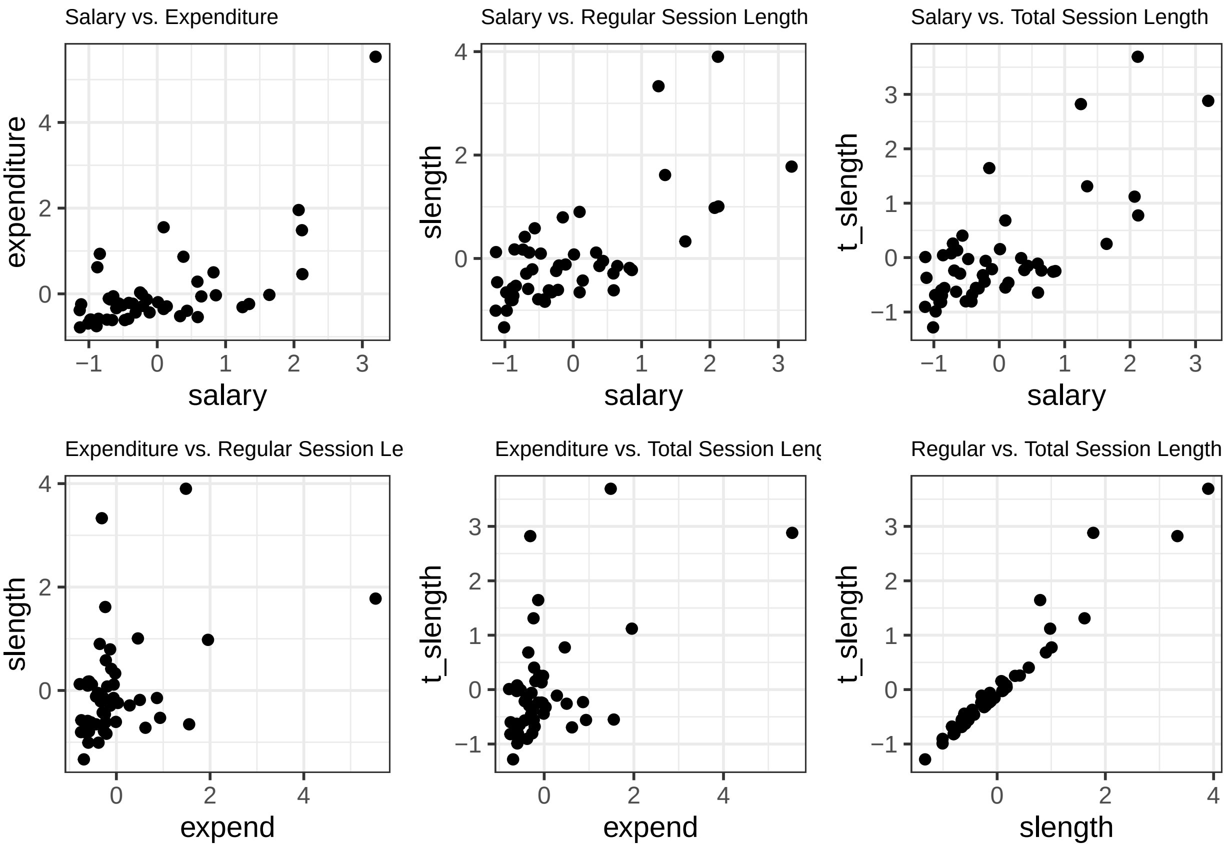
In addition to confirming the skewness of the data towards lower values, the histograms also demonstrate the presence of outliers for all of the variables.

Question 4: Diagnosing Clusterability

1) Visual diagnosis: Using scatterplots

```
scatterplot_salary_expend <- x_scaled_subset %>%
  ggplot() +
  geom_point(aes(x = salary_real, y = expend), stat = "identity") +
  labs(x = "salary", y = "expenditure", title = "Salary vs. Expenditure") +
```

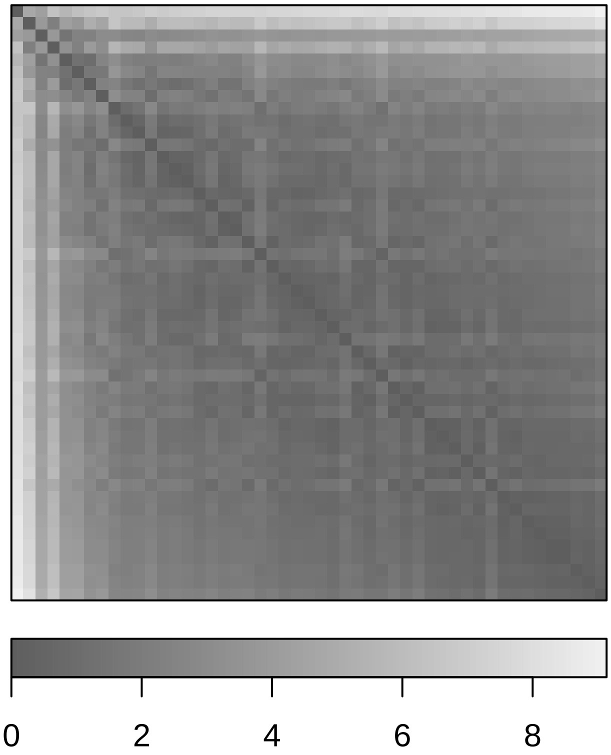
```r
  theme_bw() +
  theme(plot.title = element_text(size = 8))

scatterplot_salary_slength <- x_scaled_subset %>%
  ggplot() +
  geom_point(aes(x = salary_real, y = slength), stat = "identity") +
  labs(x = "salary", y = "slength", title = "Salary vs. Regular Session Length") +
  theme_bw() +
  theme(plot.title = element_text(size = 8))

scatterplot_salary_t_slength <- x_scaled_subset %>%
  ggplot() +
  geom_point(aes(x = salary_real, y = t_slength), stat = "identity") +
  labs(x = "salary", y = "t_slength", title = "Salary vs. Total Session Length") +
  theme_bw() +
  theme(plot.title = element_text(size = 8))

scatterplot_expend_slength <- x_scaled_subset %>%
  ggplot() +
  geom_point(aes(x = expend, y = slength), stat = "identity") +
  labs(x = "expend", y = "slength", title = "Expenditure vs. Regular Session Length") +
  theme_bw() +
  theme(plot.title = element_text(size = 7.5))

scatterplot_expend_t_slength <- x_scaled_subset %>%
  ggplot() +
  geom_point(aes(x = expend, y = t_slength), stat = "identity") +
  labs(x = "expend", y = "t_slength", title = "Expenditure vs. Total Session Length") +
  theme_bw() +
  theme(plot.title = element_text(size = 7.5))

scatterplot_slength_t_slength <- x_scaled_subset %>%
  ggplot() +
  geom_point(aes(x = slength, y = t_slength), stat = "identity") +
  labs(x = "slength", y = "t_slength", title = "Regular vs. Total Session Length") +
  theme_bw() +
  theme(plot.title = element_text(size = 8))

grid.arrange(scatterplot_salary_expend, scatterplot_salary_slength, scatterplot_salary_t_slength, scatte
scatterplot_slength_t_slength, ncol = 3, nrow = 2)
```



The scatterplots above show that all the variables appear to be positively associated with each other, with the strongest positive association between the regular session length and total session length. There appears to be potential for clusterability at the lower values and there is some seperation between low and high values that we can see in the plots.

2) ODI plot

```
dist_legal_profes <- dist(x_scaled_subset, method = "euclidean")
dissplot(dist_legal_profes)
```



Visual inspection of the ODI plot above does not provide clear evidence of strong clusters in the data. There does appear to be the possiblity for clustering, however, which I see from the faint outline of a dark square in the lefthand corner and righthand corner, which might suggest the presence of clusters.

Question 5: K Means Algorithm

```
set.seed(450)
kmeans <- kmeans(x_scaled_subset,
                 centers = 2,
                 nstart = 15)
```

After fitting the k-means algorithm above, we can see the cluster assignment by state below:

```
x_scaled_kmeans <- x_scaled_subset
x_scaled_kmeans$clusters <- as.factor(kmeans$cluster)
t <- as.table(kmeans$cluster)
t <- data.frame(t)
rownames(t) <- c(x_state_names)
colnames(t)[colnames(t)=="Freq"] <- "Assignment"
t$Var1 <- NULL
kable(t)
```

	Assignment
Alabama	1
Alaska	1
Arizona	1
Arkansas	1
California	2
Colorado	1
Connecticut	1
Delaware	1
Florida	1
Georgia	1
Hawaii	1
Idaho	1
Illinois	1
Indiana	1
Iowa	1
Kansas	1
Kentucky	1
Louisiana	1
Maine	1
Maryland	1
Massachusetts	2
Michigan	2
Minnesota	1
Mississippi	1
Missouri	1
Montana	1
Nebraska	1
Nevada	1
New Hampshire	1
New Jersey	1
New Mexico	1
New York	2
North Carolina	1
North Dakota	1
Ohio	2
Oklahoma	1
Oregon	1
Pennsylvania	2
Rhode Island	1
South Carolina	1
South Dakota	1
Tennessee	1
Texas	1
Utah	1
Vermont	1
Virginia	1
Washington	1
West Virginia	1
Wyoming	1

A summary of the clustering assignments can be seen below:

```
kmeans$centers
```

```
##      slength  t_slength salary_real     expend
## 1 -0.2932285 -0.2930275  -0.3023668 -0.2089249
## 2  2.1014710  2.1000302   2.0155567  1.4807388
```

```
kmeans$size
```

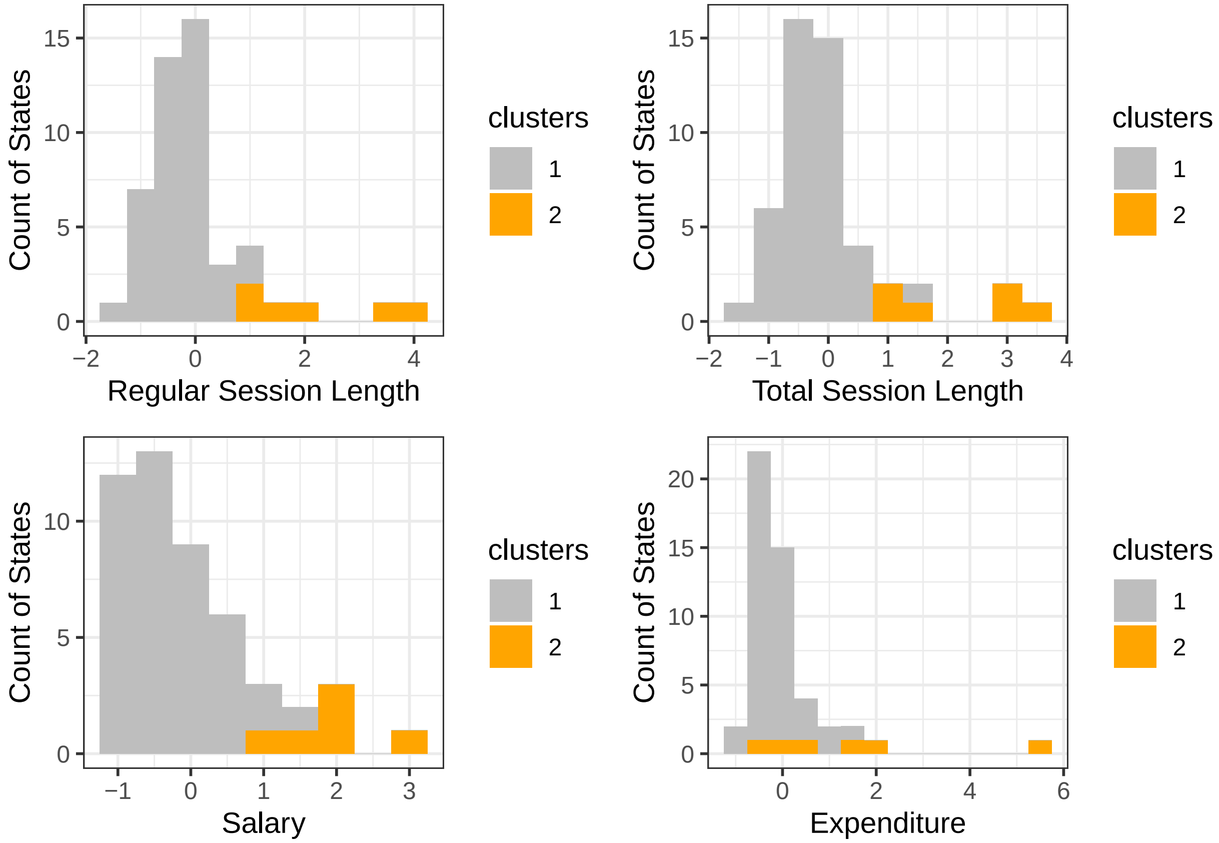
```
## [1] 43  6
```

As we can see above, cluster 1 was centered around -0.20 to -0.30 for all the variables, and cluster 2 was centered around 1.5 to 2.1. The majority of the states were assigned to cluster 1, with 43 states being assigned to this cluster, compared to just 6 states assigned to cluster 2.

The plots below show a visual description of the cluster assignment across the distribution for each variable. For regular session length, total session length, and salary, cluster 1 appears to be to the left of cluster 2. However, for expenditure, there doesn't seem to be such a clear division of the clusters geometrically.

```
slength_clusters <- ggplot(x_scaled_kmeans, aes(slength, fill = clusters)) +
  geom_histogram(binwidth = 0.5) +
  theme_bw() +
  scale_fill_manual(values=c("gray", "orange")) +
  labs(x = "Regular Session Length",
       y = "Count of States")

t_slength_clusters <- ggplot(x_scaled_kmeans, aes(t_slength, fill = clusters)) +
  geom_histogram(binwidth = 0.5) +
  theme_bw() +
  scale_fill_manual(values=c("gray", "orange")) +
  labs(x = "Total Session Length",
       y = "Count of States")

salary_clusters <- ggplot(x_scaled_kmeans, aes(salary_real, fill = clusters)) +
  geom_histogram(binwidth = 0.5) +
  theme_bw() +
  scale_fill_manual(values=c("gray", "orange")) +
  labs(x = "Salary",
       y = "Count of States")

expenditure_clusters <- ggplot(x_scaled_kmeans, aes(expend, fill = clusters)) +
  geom_histogram(binwidth = 0.5) +
  theme_bw() +
  scale_fill_manual(values=c("gray", "orange")) +
  labs(x = "Expenditure",
       y = "Count of States")
```

```r
grid.arrange(slength_clusters, t_slength_clusters, salary_clusters, expenditure_clusters,
             ncol = 2, nrow = 2)
```



## Question 6: GMM Algorithm

```r
set.seed(450)
gmm1 <- mvnormalmixEM(x_scaled_subset, k=2)
```

```
## number of iterations= 11
```

Some of the results from the GMM output are displayed below. It took 11 iterations to establish convergence.

```r
gmm1$mu
```

```
## [[1]]
## [1] 2.156634 2.431846 1.694880 1.705799
##
## [[2]]
## [1] -0.2450724 -0.2763465 -0.2132464 -0.1960989
```

```
gmm1$sigma
```

```
## [[1]]
##            [,1]      [,2]      [,3]       [,4]
## [1,]  1.5611386 1.0197243 0.3432493 -0.3997010
## [2,]  1.0197243 0.8556104 0.3986101  0.3545036
## [3,]  0.3432493 0.3986101 1.2353043  2.0069944
## [4,] -0.3997010 0.3545036 2.0069944  4.4408608
##
## [[2]]
##            [,1]       [,2]      [,3]       [,4]
## [1,] 0.32491503 0.27954563 0.2379421 0.02504616
## [2,] 0.27954563 0.24528126 0.2100159 0.03137054
## [3,] 0.23794209 0.21001590 0.5825970 0.13823755
## [4,] 0.02504616 0.03137054 0.1382376 0.23965942
```

```
gmm1$lambda
```

```
## [1] 0.102041 0.897959
```

The table below shows each of the states and the posterior probability of being in either of the components. Using a threshold of 0.3, the column "components" assigns each State to either component 1 or component 2. 45 states were assigned to component 1, and the remaining states were assigned to component 2.

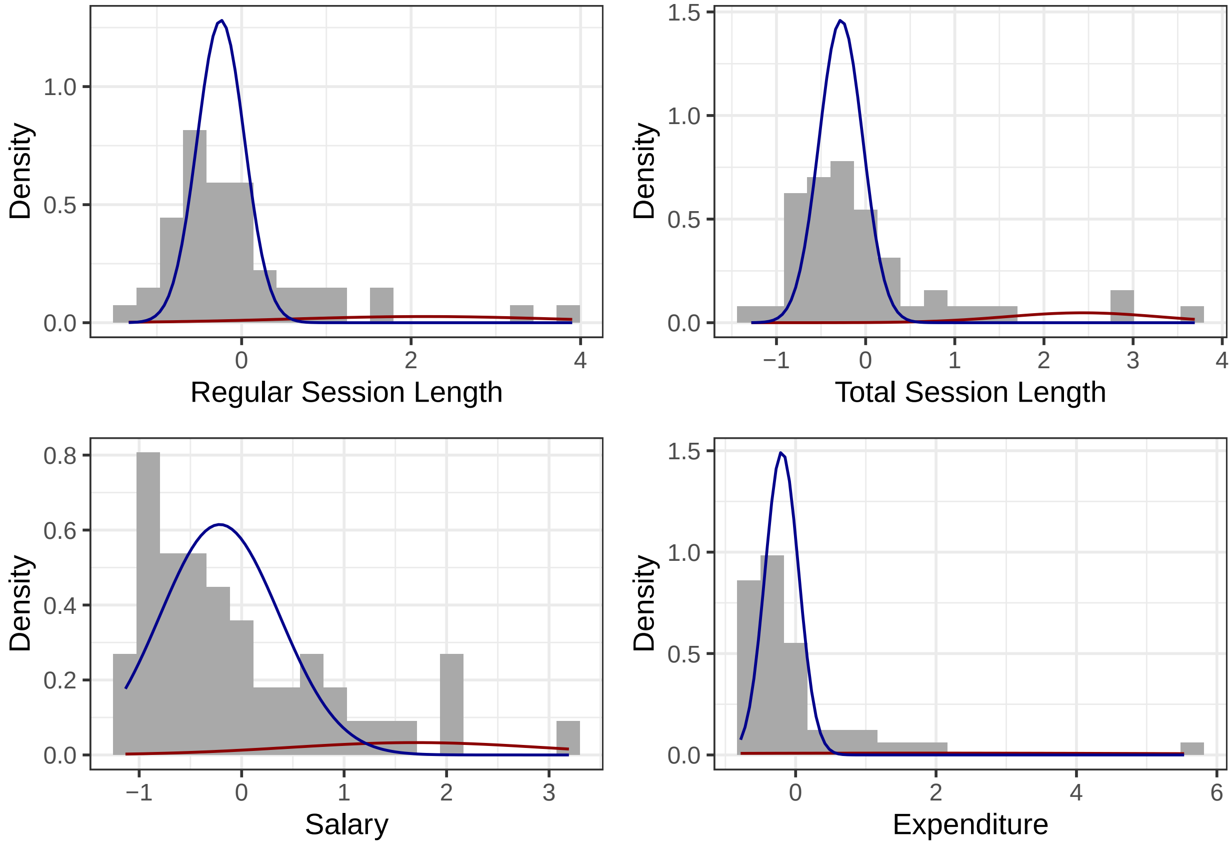
```
posterior <- data.frame(gmm1$posterior)
rownames(posterior) <- x_state_names
posterior$components <- ifelse(posterior$comp.1 < 0.3, 1, 2)
x_scaled_gmm <- x_scaled_subset
x_scaled_gmm$clusters <- posterior$components
kable(posterior)
```

	comp.1	comp.2	components
Alabama	0.0000000	1.0000000	1
Alaska	0.0000000	1.0000000	1
Arizona	1.0000000	0.0000000	2
Arkansas	0.0000000	1.0000000	1
California	1.0000000	0.0000000	2
Colorado	0.0000000	1.0000000	1
Connecticut	0.0000000	1.0000000	1
Delaware	0.0000076	0.9999924	1
Florida	0.0000000	1.0000000	1
Georgia	0.0000000	1.0000000	1
Hawaii	0.0000000	1.0000000	1
Idaho	0.0000000	1.0000000	1
Illinois	0.0000000	1.0000000	1
Indiana	0.0000000	1.0000000	1
Iowa	0.0000000	1.0000000	1
Kansas	0.0000000	1.0000000	1
Kentucky	0.0000000	1.0000000	1
Louisiana	0.0000000	1.0000000	1
Maine	0.0000000	1.0000000	1
Maryland	0.0000000	1.0000000	1
Massachusetts	0.9999999	0.0000001	2
Michigan	0.0000000	1.0000000	1
Minnesota	0.0000000	1.0000000	1
Mississippi	0.0000000	1.0000000	1
Missouri	0.0000000	1.0000000	1
Montana	0.0000000	1.0000000	1
Nebraska	0.0000000	1.0000000	1
Nevada	0.0000000	1.0000000	1
New Hampshire	0.0000000	1.0000000	1
New Jersey	0.0000000	1.0000000	1
New Mexico	0.0000000	1.0000000	1
New York	1.0000000	0.0000000	2
North Carolina	0.0000000	1.0000000	1
North Dakota	0.0000000	1.0000000	1
Ohio	0.0000000	1.0000000	1
Oklahoma	0.0000000	1.0000000	1
Oregon	0.0000000	1.0000000	1
Pennsylvania	0.9999993	0.0000007	2
Rhode Island	0.0000000	1.0000000	1
South Carolina	0.0000000	1.0000000	1
South Dakota	0.0000000	1.0000000	1
Tennessee	0.0000000	1.0000000	1
Texas	0.0000000	1.0000000	1
Utah	0.0000000	1.0000000	1
Vermont	0.0000000	1.0000000	1
Virginia	0.0000000	1.0000000	1
Washington	0.0000000	1.0000000	1
West Virginia	0.0000000	1.0000000	1
Wyoming	0.0000000	1.0000000	1

The plots below show the assignment clusters across the distributions of the variables. The dark red line represents cluster 2 while the blue line represents cluster 1.

```r
slength_gmm <- ggplot(data.frame(x = gmm1$x[,1])) +
  geom_histogram(aes(x, ..density..), fill = "darkgray", bins= 20) +
  stat_function(geom = "line", fun = plot_mix_comps,
                args = list(gmm1$mu[[1]][1], gmm1$sigma[[1]][1, 1], lam = gmm1$lambda[1]),
                colour = "darkred") +
  stat_function(geom = "line", fun = plot_mix_comps,
                args = list(gmm1$mu[[2]][1], gmm1$sigma[[2]][2, 1], lam = gmm1$lambda[2]),
                colour = "darkblue") +
  xlab("Regular Session Length") +
  ylab("Density") +
  theme_bw()


t_slength_gmm <- ggplot(data.frame(x = gmm1$x[,2])) +
  geom_histogram(aes(x, ..density..), fill = "darkgray", bins = 20) +
  stat_function(geom = "line", fun = plot_mix_comps,
                args = list(gmm1$mu[[1]][2], gmm1$sigma[[1]][2, 2], lam = gmm1$lambda[1]),
                colour = "darkred") +
  stat_function(geom = "line", fun = plot_mix_comps,
                args = list(gmm1$mu[[2]][2], gmm1$sigma[[2]][2, 2], lam = gmm1$lambda[2]),
                colour = "darkblue") +
  xlab("Total Session Length") +
  ylab("Density") +
  theme_bw()


salary_gmm <-  ggplot(data.frame(x = gmm1$x[,3])) +
  geom_histogram(aes(x, ..density..), fill = "darkgray", bins = 20) +
  stat_function(geom = "line", fun = plot_mix_comps,
                args = list(gmm1$mu[[1]][3], gmm1$sigma[[1]][3, 3], lam = gmm1$lambda[1]),
                colour = "darkred") +
  stat_function(geom = "line", fun = plot_mix_comps,
                args = list(gmm1$mu[[2]][3], gmm1$sigma[[2]][3, 3], lam = gmm1$lambda[2]),
                colour = "darkblue") +
  xlab("Salary") +
  ylab("Density") +
  theme_bw()


expend_gmm <-  ggplot(data.frame(x = gmm1$x[,4])) +
  geom_histogram(aes(x, ..density..), fill = "darkgray", bins = 20) +
  stat_function(geom = "line", fun = plot_mix_comps,
                args = list(gmm1$mu[[1]][4], gmm1$sigma[[1]][4, 4], lam = gmm1$lambda[1]),
                colour = "darkred") +
  stat_function(geom = "line", fun = plot_mix_comps,
                args = list(gmm1$mu[[2]][4], gmm1$sigma[[2]][4, 4], lam = gmm1$lambda[2]),
                colour = "darkblue") +
  xlab("Expenditure") +
  ylab("Density") +
  theme_bw()
```

```r
grid.arrange(slength_gmm, t_slength_gmm, salary_gmm, expend_gmm, ncol = 2, nrow = 2)
```



From the plots above, it is possible to see that cluster 1 tended to be more normally distributed than cluster 2. This is likely due to the presence of outliers at higher values.

Question 7: PAM

The PAM algorithm below is similar to k-means but uses k-medoids instead.

```r
pam <- pam(x_scaled_subset, k=2)
pam$medoids
```

```
##       slength  t_slength salary_real     expend
## 39 -0.2101066 -0.2949443  -0.5984529 -0.3349137
## 22  1.0063116  0.7755062   2.1230894  0.4595596
```

```r
pam$clusinfo
```

```
##      size max_diss   av_diss diameter separation
## [1,]   42 2.239394 0.9580071 3.761115   1.080918
## [2,]    7 5.648779 2.4494864 6.503802   1.080918
```

As the summary above shows, the medoids for cluster 1 are centered around -0.21 to -0.60, whereas the medoids for cluster 2 are centered around 0.7756 to 2.123. There were 42 states assigned to cluster 1 compared 7 in cluster 2.

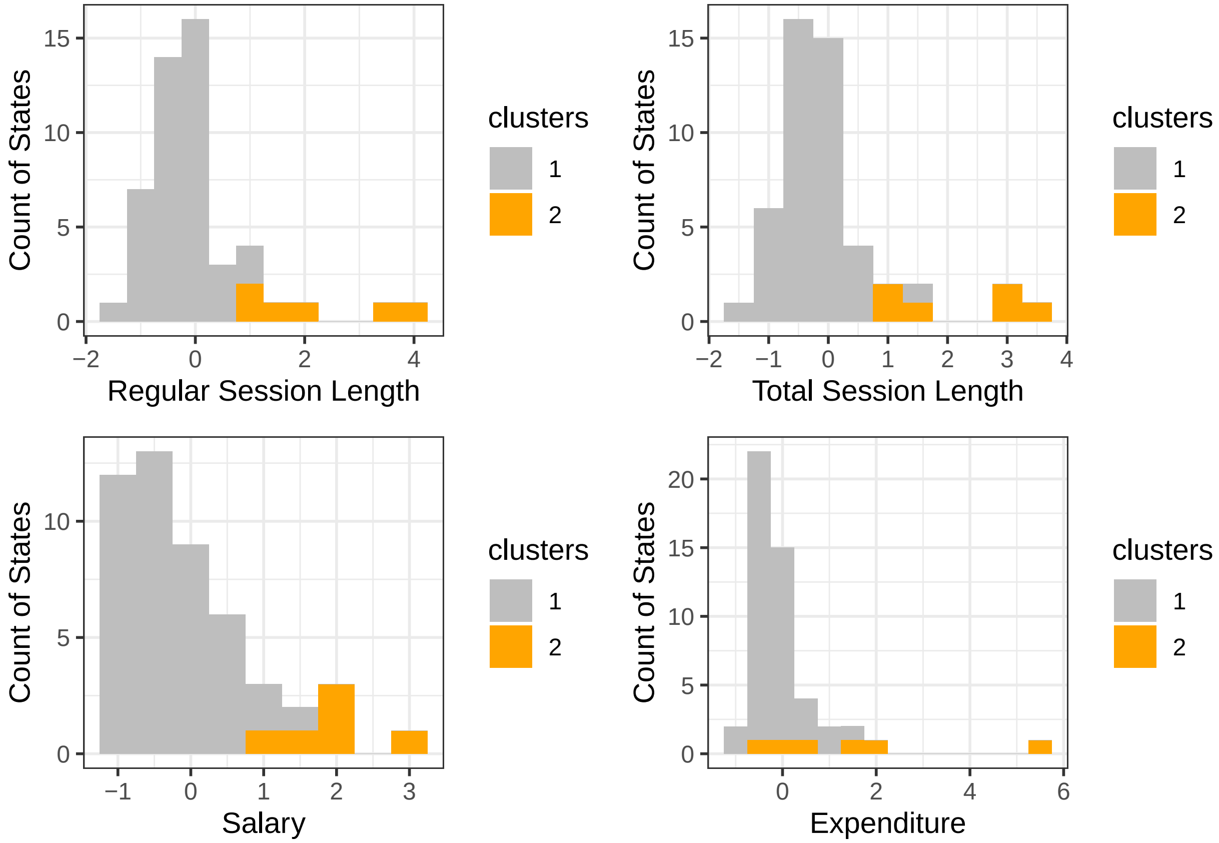
The table below shows the assignment of each of the states to each cluster.

```
x_scaled_pam <- x_scaled_subset
x_scaled_pam$clusters <- pam$clustering
t <- as.table(pam$clustering)
t <- data.frame(t)
rownames(t) <- c(x_state_names)
colnames(t)[colnames(t)=="Freq"] <- "Assignment"
t$Var1 <- NULL
kable(t)
```

	Assignment
Alabama	1
Alaska	1
Arizona	1
Arkansas	1
California	2
Colorado	1
Connecticut	1
Delaware	1
Florida	1
Georgia	1
Hawaii	1
Idaho	1
Illinois	2
Indiana	1
Iowa	1
Kansas	1
Kentucky	1
Louisiana	1
Maine	1
Maryland	1
Massachusetts	2
Michigan	2
Minnesota	1
Mississippi	1
Missouri	1
Montana	1
Nebraska	1
Nevada	1
New Hampshire	1
New Jersey	1
New Mexico	1
New York	2
North Carolina	1
North Dakota	1
Ohio	2
Oklahoma	1
Oregon	1
Pennsylvania	2
Rhode Island	1
South Carolina	1
South Dakota	1
Tennessee	1
Texas	1
Utah	1
Vermont	1
Virginia	1
Washington	1
West Virginia	1
Wyoming	1

Below, I show a distribution plot for each of the variables (similarly to what was done for the k-means model). The trends and analysis are similar to that of k-means.

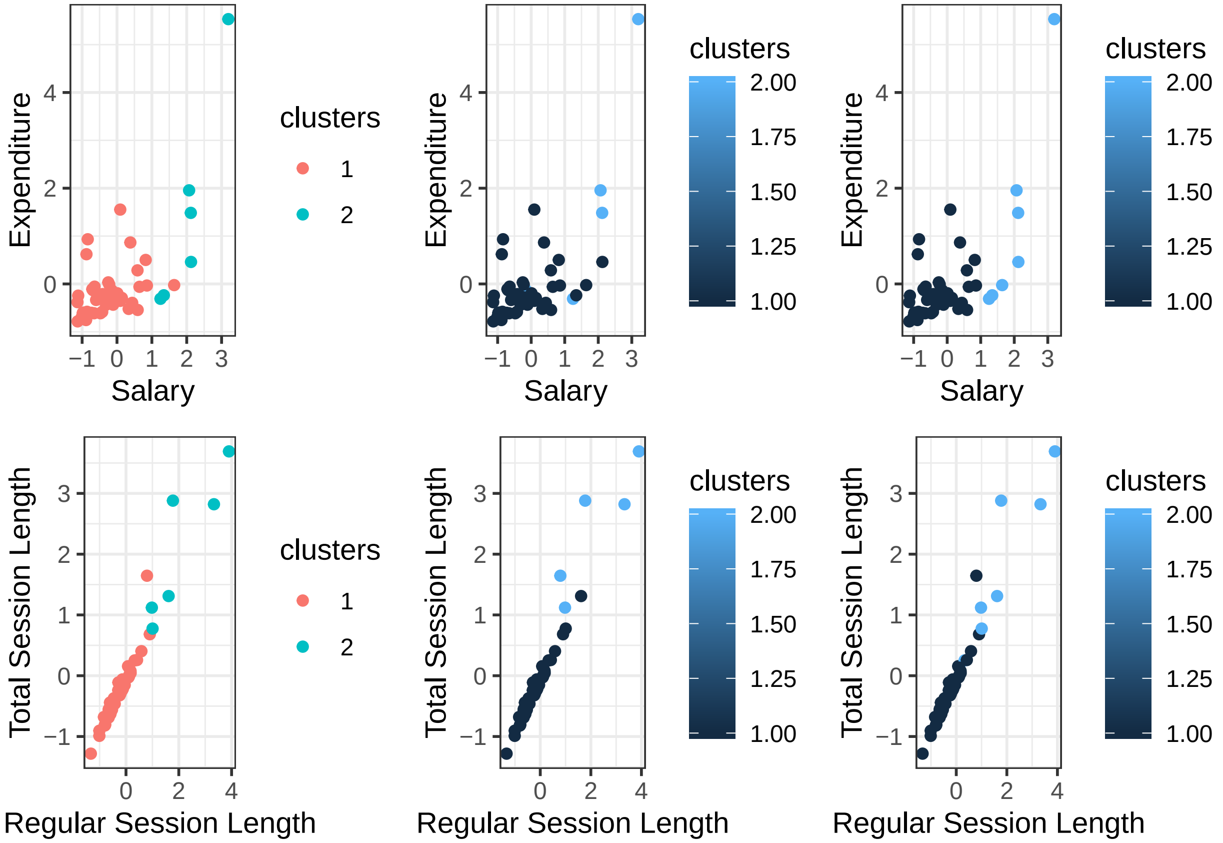
```
slength_pam <- ggplot(x_scaled_pam, aes(slength, fill = clusters)) +
  geom_histogram(binwidth = 0.5) +
  theme_bw() +
  scale_fill_manual(values=c("gray", "orange")) +
  labs(x = "Regular Session Length",
       y = "Count of States")

t_slength_pam <- ggplot(x_scaled_pam, aes(t_slength, fill = clusters)) +
  geom_histogram(binwidth = 0.5) +
  theme_bw() +
  scale_fill_manual(values=c("gray", "orange")) +
  labs(x = "Total Session Length",
       y = "Count of States")

salary_pam <- ggplot(x_scaled_pam, aes(salary_real, fill = clusters)) +
  geom_histogram(binwidth = 0.5) +
  theme_bw() +
  scale_fill_manual(values=c("gray", "orange")) +
  labs(x = "Salary",
       y = "Count of States")

expenditure_pam <- ggplot(x_scaled_pam, aes(expend, fill = clusters)) +
  geom_histogram(binwidth = 0.5) +
  theme_bw() +
  scale_fill_manual(values=c("gray", "orange")) +
  labs(x = "Expenditure",
       y = "Count of States")

grid.arrange(slength_clusters, t_slength_clusters, salary_clusters, expenditure_clusters,
             ncol = 2, nrow = 2)
```



Question 9: Comparing Outputs of 3 Models

To compare the 3 models, I plotted the cluster assignments for expenditure versus salary and regular session length versus total session length. Visually, it is possible to see the density of clusters at the lower values and higher values, across different methods.

```
compare_1 <- ggplot(x_scaled_kmeans, aes(x = salary_real, y = expend, col = clusters)) +
  geom_point() +
  theme_bw() +
  scale_fill_manual(values=c("gray", "orange")) +
  labs(x = "Salary",
       y = "Expenditure")

compare_2 <- ggplot(x_scaled_gmm, aes(x = salary_real, y = expend, col = clusters)) +
  geom_point() +
  theme_bw() +
  scale_fill_manual(values=c("gray", "orange")) +
  labs(x = "Salary",
       y = "Expenditure")

compare_3 <- ggplot(x_scaled_pam, aes(x = salary_real, y = expend, col = clusters)) +
  geom_point() +
  theme_bw() +
  scale_fill_manual(values=c("gray", "orange")) +
  labs(x = "Salary",
       y = "Expenditure")

compare_4 <- ggplot(x_scaled_kmeans, aes(x = slength, y = t_slength, col = clusters)) +
  geom_point() +
  theme_bw() +
  scale_fill_manual(values=c("gray", "orange")) +
  labs(x = "Regular Session Length",
       y = "Total Session Length")

compare_5 <- ggplot(x_scaled_gmm, aes(x = slength, y = t_slength, col = clusters)) +
  geom_point() +
  theme_bw() +
  scale_fill_manual(values=c("gray", "orange")) +
  labs(x = "Regular Session Length",
       y = "Total Session Length")

compare_6 <- ggplot(x_scaled_pam, aes(x = slength, y = t_slength, col = clusters)) +
  geom_point() +
  theme_bw() +
  scale_fill_manual(values=c("gray", "orange")) +
  labs(x = "Regular Session Length",
       y = "Total Session Length")

grid.arrange(compare_1, compare_2, compare_3, compare_4, compare_5, compare_6, ncol = 3, nrow = 2)
```

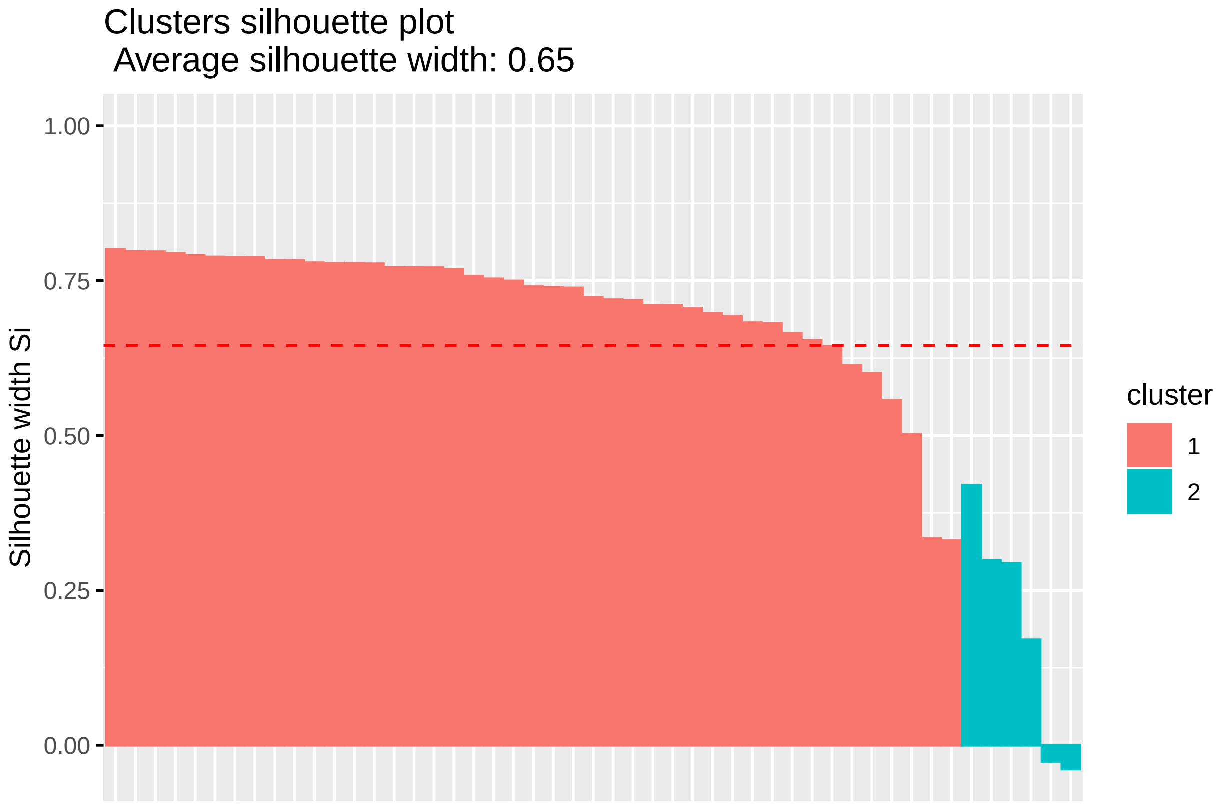


Question 10: Internal Validation

Please see silhouette plots below

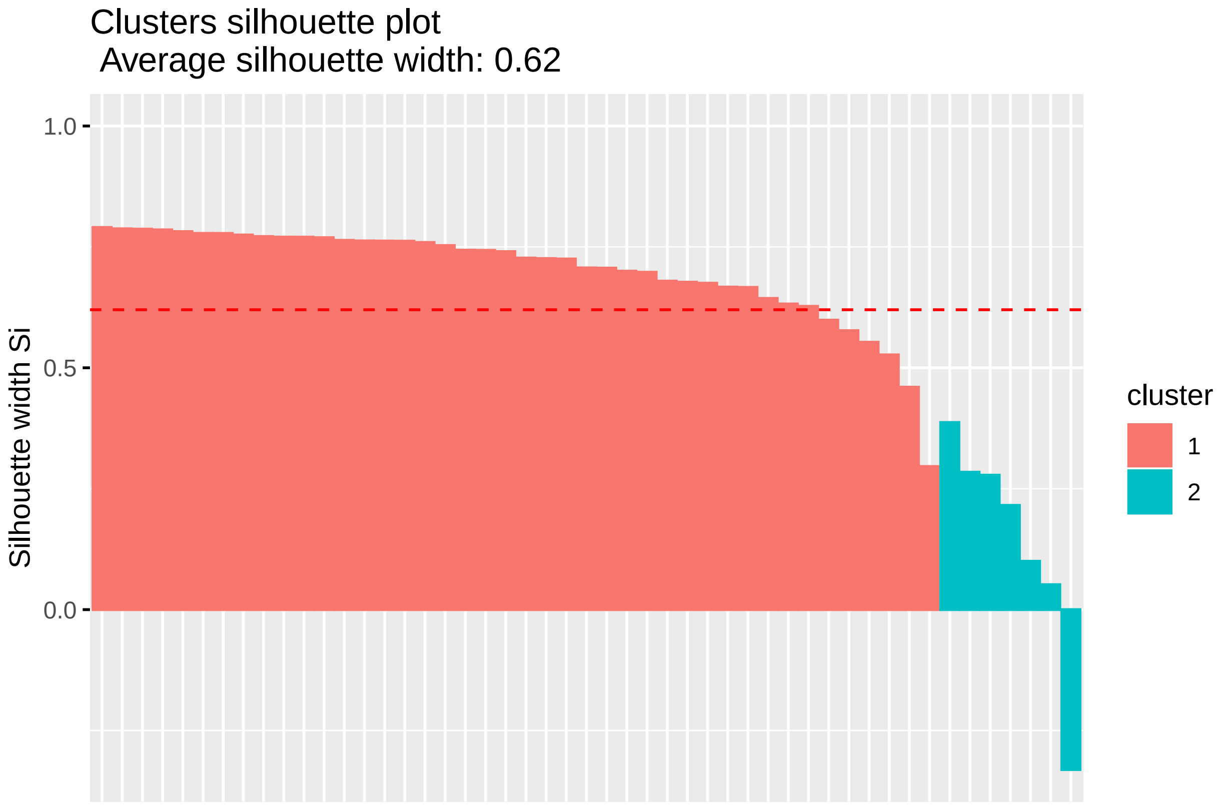
```
sil1 <- silhouette(kmeans$cluster, dist(x_scaled_subset))
fviz_silhouette(sil1)
```

```
##   cluster size ave.sil.width
## 1       1   43          0.71
## 2       2    6          0.19
```



```
sil2 <- silhouette(pam)
fviz_silhouette(sil2)
```

```
##   cluster size ave.sil.width
## 1       1   42          0.70
## 2       2    7          0.14
```



For the purposes of internal validation, I created silhouette plots to compare the silhouette width of PAM versus K-means. I was not sure how to do this for GMM. It is possible to see that the average silhouette width was slightly higher for k-means versus PAM. Based on this, I would say that the kmeans approach is more optimal than PAM.