01 Descriptive

Colin Linke

2024-05-08

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
##
## Attaching package: 'cowplot'
##
##
## The following object is masked from 'package:lubridate':
##
##     stamp
##
##
##
## Attaching package: 'patchwork'
##
##
## The following object is masked from 'package:cowplot':
##
##     align_plots
##
##
##
## Attaching package: 'nlme'
##
##
## The following object is masked from 'package:dplyr':
##
##     collapse
##
##
## Loading required package: Matrix
##
##
## Attaching package: 'Matrix'
##
##
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
##
##
## Attaching package: 'lme4'
##
##
## The following object is masked from 'package:nlme':
##
##     lmList
##
##
## Loading required package: carData
##
##
## Attaching package: 'car'
##
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
##
## The following object is masked from 'package:purrr':
##
##     some
##
##
## Use the command
##     lattice::trellis.par.set(effectsTheme())
##   to customize lattice options for effects plots.
## See ?effectsTheme for details.
##
##
## Attaching package: 'sjPlot'
##
##
## The following objects are masked from 'package:cowplot':
##
##     plot_grid, save_plot
##
##
##
## Attaching package: 'lmerTest'
##
##
## The following object is masked from 'package:lme4':
##
##     lmer
##
##
## The following object is masked from 'package:stats':
```

```
##
##    step
```

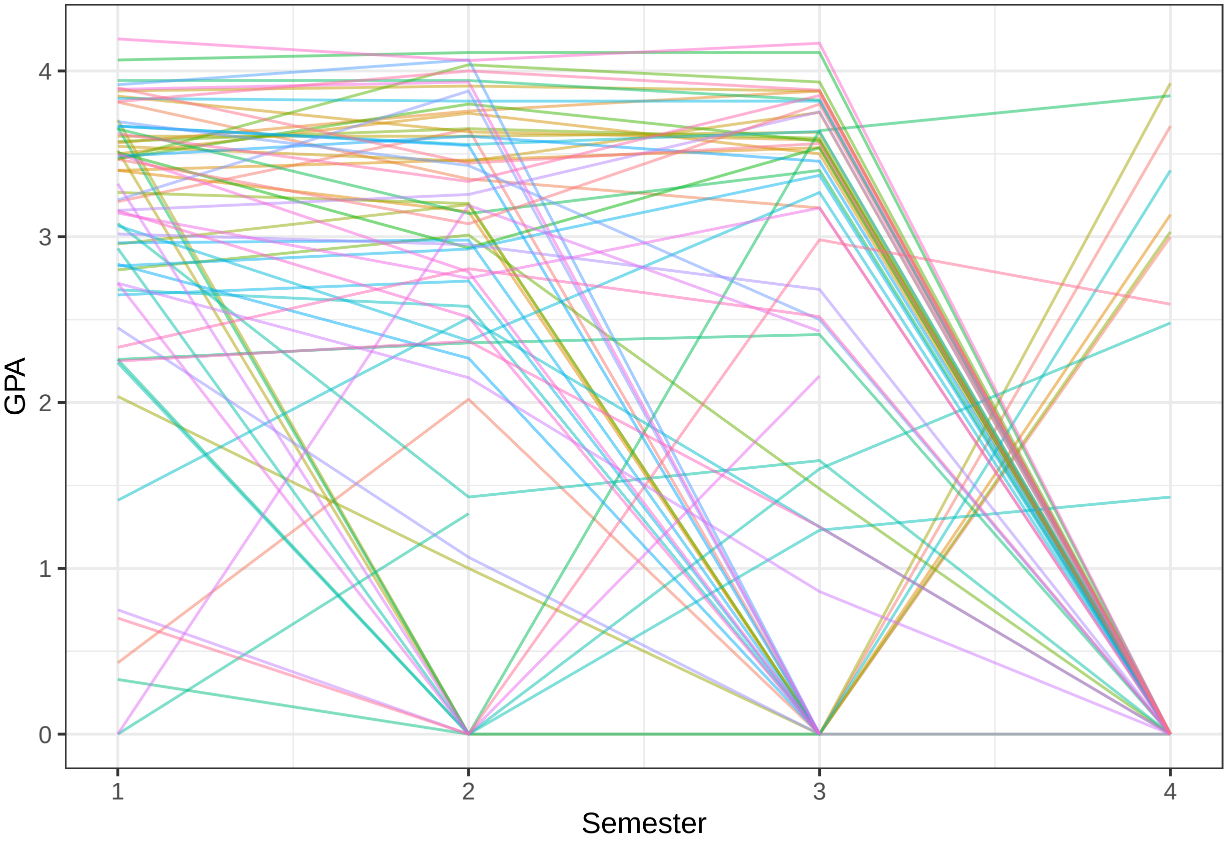
```
variables.paper.page7table <- c("Age1stround", "SATMath", "SATVerbal", "SATWriting", "GPA", "Parental_S
catvars.paper.page7table <- c("Sex", "Fager4_binary", "FH_binary")
paper.page7table <- CreateTableOne(data = (data.file.long %>% filter(Semester == 1)), vars = variables.p

paper.page7table
```

```
##                               Stratified by Cluster_SEM1
##                                1st.cluster    2nd.cluster    3rd.cluster
##   n                              487            463            188
##   Age1stround (mean (SD))       18.32 (0.91)   18.30 (0.73)   18.30 (0.63)
##   SATMath (mean (SD))          541.05 (89.52) 554.98 (90.68) 554.24 (84.78)
##   SATVerbal (mean (SD))        530.63 (91.04) 541.56 (89.33) 541.24 (76.95)
##   SATWriting (mean (SD))       534.41 (90.45) 553.75 (92.03) 544.82 (83.87)
##   GPA (mean (SD))                3.10 (0.67)    3.04 (0.64)    2.71 (0.77)
##   Parental_SES (mean (SD))      12.55 (7.05)   10.23 (5.47)   10.24 (5.76)
##   STAI_SELF_Total (mean (SD))  40.14 (9.87)   39.23 (10.09)  41.46 (10.70)
##   BDI_SELF_Total (mean (SD))    3.33 (4.45)    3.13 (4.44)    4.24 (5.06)
##   Avg_Drinks_SEM1 (mean (SD))   0.40 (0.75)   29.29 (32.22)  54.54 (42.69)
##   Avg_MJ_SEM1 (mean (SD))       0.09 (0.40)    0.42 (0.72)   13.55 (8.13)
##   Sex (%)
##      female                      299 (61.4)     286 (61.8)      87 (46.3)
##      male                        186 (38.2)     173 (37.4)     100 (53.2)
##      NA                            2 ( 0.4)       4 ( 0.9)       1 ( 0.5)
##   Fager4_binary (%)
##      non smoker                  459 (94.3)     411 (88.8)     147 (78.2)
##      smoker                       19 ( 3.9)      42 ( 9.1)      38 (20.2)
##      NA                            9 ( 1.8)      10 ( 2.2)       3 ( 1.6)
##   FH_binary = positive (%)       109 (22.4)      98 (21.2)      49 (26.1)
##                               Stratified by Cluster_SEM1
##                                p      test
##   n
##   Age1stround (mean (SD))       0.896
##   SATMath (mean (SD))           0.049
##   SATVerbal (mean (SD))         0.146
##   SATWriting (mean (SD))        0.007
##   GPA (mean (SD))              <0.001
##   Parental_SES (mean (SD))     <0.001
##   STAI_SELF_Total (mean (SD))   0.039
##   BDI_SELF_Total (mean (SD))    0.017
##   Avg_Drinks_SEM1 (mean (SD))  <0.001
##   Avg_MJ_SEM1 (mean (SD))      <0.001
##   Sex (%)                       0.003
##      female
##      male
##      NA
##   Fager4_binary (%)            <0.001
##      non smoker
##      smoker
##      NA
##   FH_binary = positive (%)      0.397
```

Checking for the potential missing data

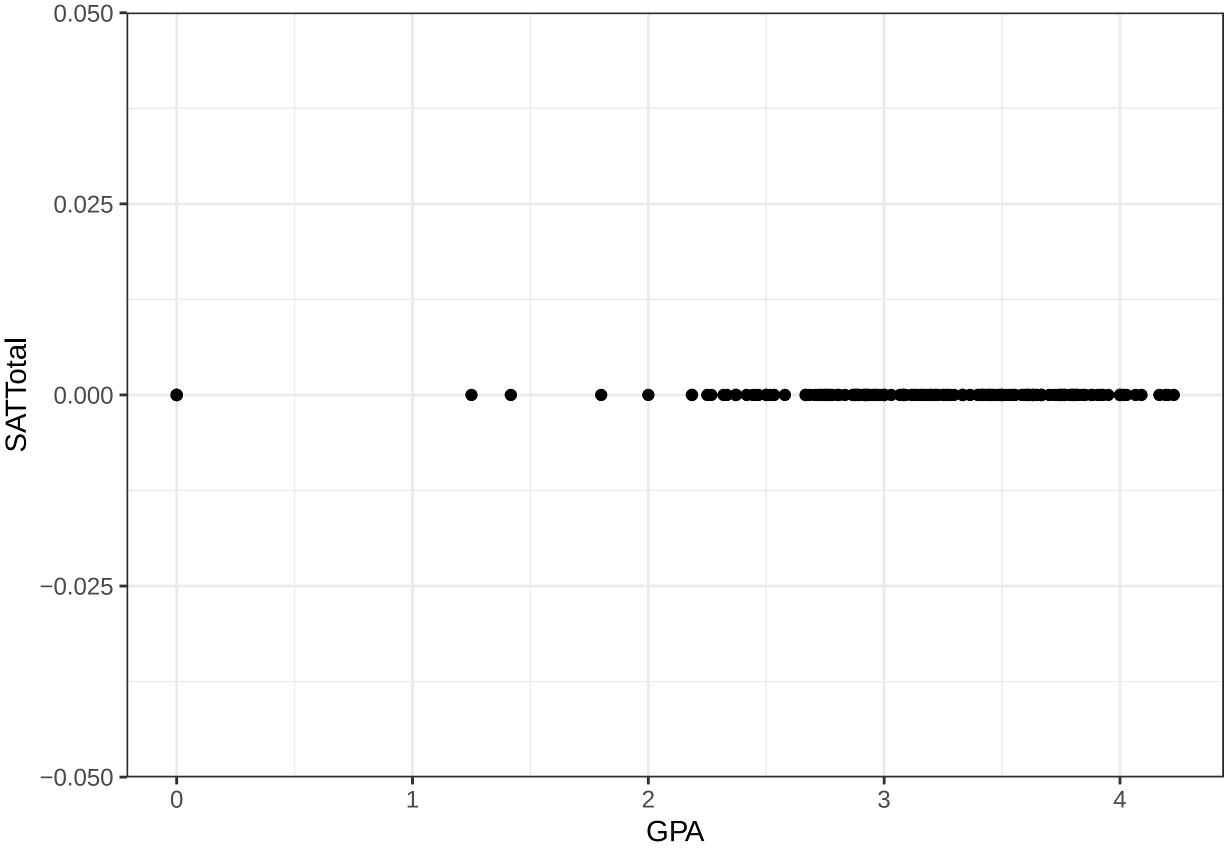
```r
plot.gpa0 <- ggplot(ind.gpa0, aes(x = Semester, y = GPA, group = BARCS_ID, col = BARCS_ID)) +
  geom_line(alpha = .5)+ theme_bw()
plot.gpa0 + theme(legend.position="none")
```

```
## Warning: Removed 24 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



```r
ggplot(ind.sat0, aes(y = SATTotal, x = GPA)) + geom_point() + theme_bw()
```

```
## Warning: Removed 4 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



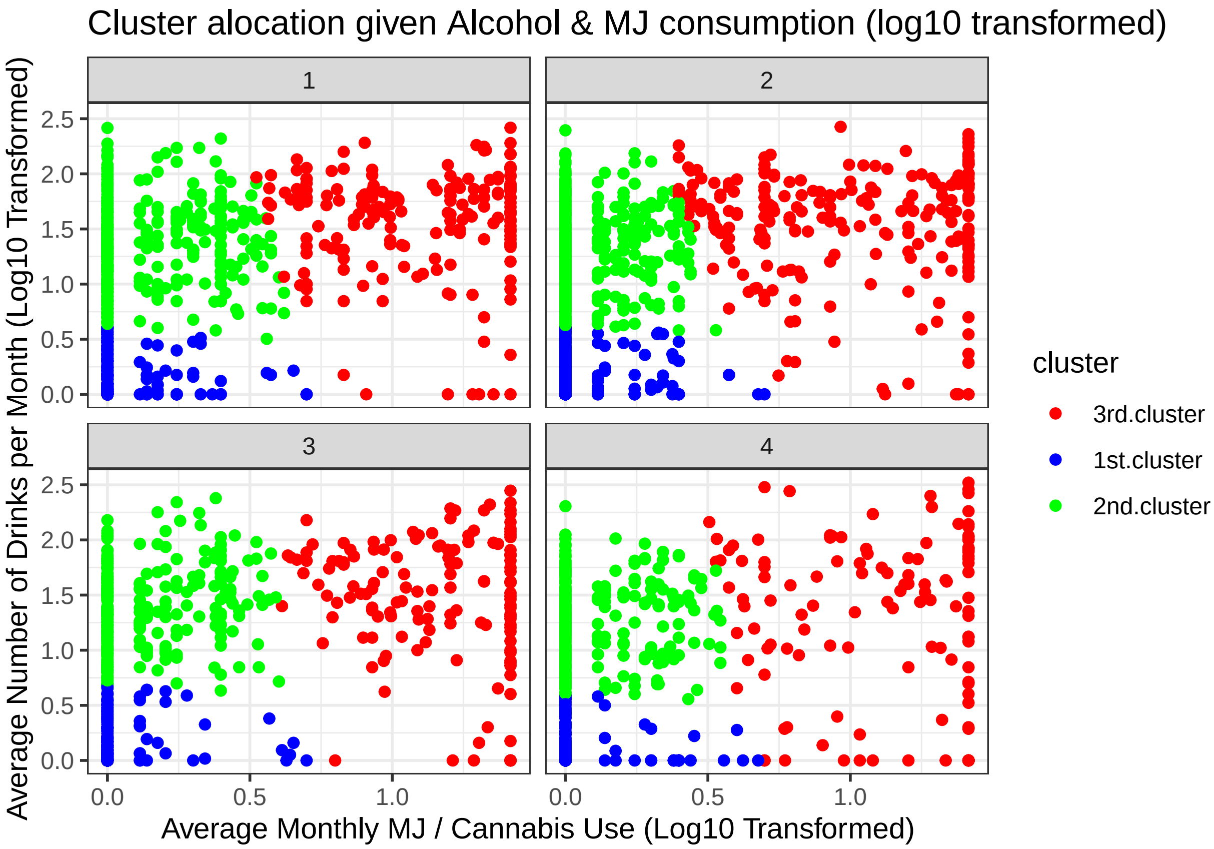
```r
dim(ind.sat0 %>% filter(GPA == 0))
```

```
## [1] 12  4
```

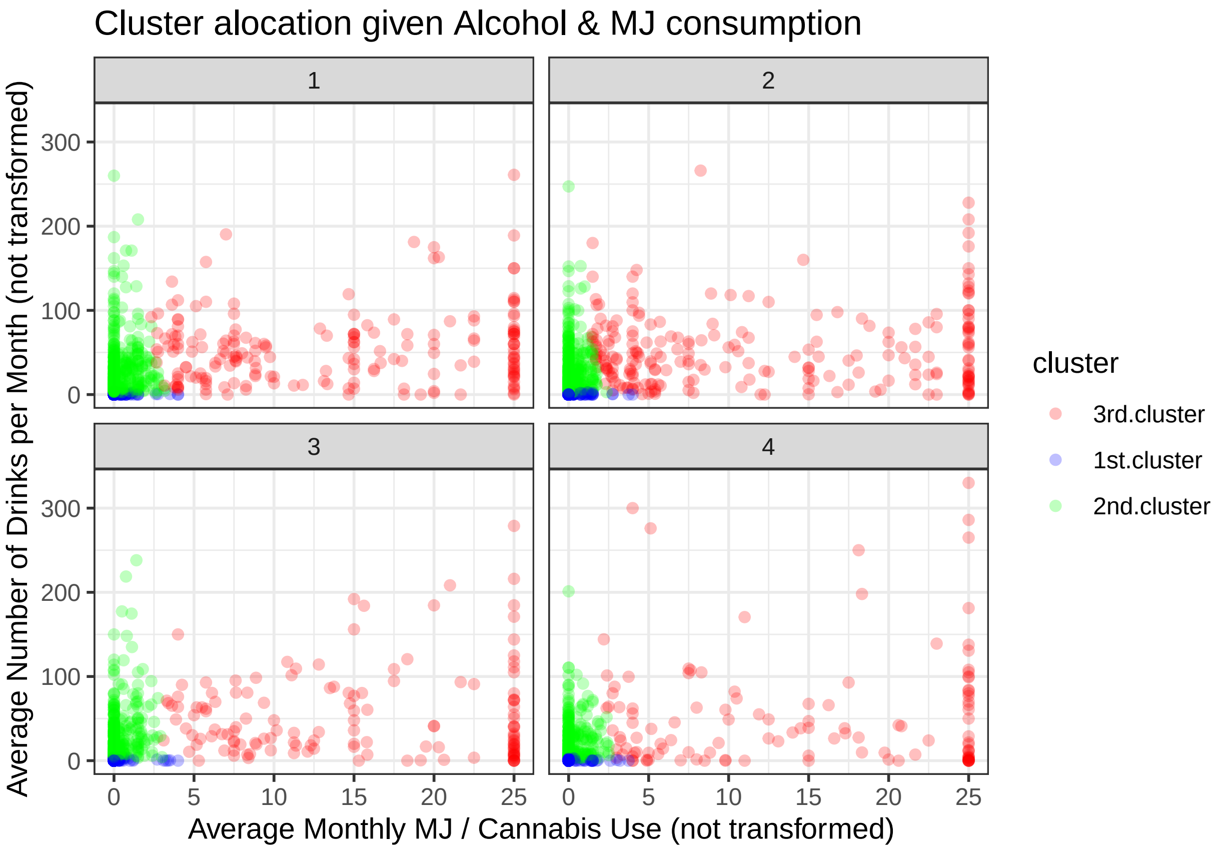
Considering that almost all students have non zero GPA data before and after GPA = 0, it seems likely that a lot students were wrongly assigned 0 for NAs here. Additionally, a completed SAT can not have a point total of zero. It is also extremely likely here that 0 entries here mean that the SAT are missing data, as most students also have a nonzero GPA. Zero entries are subsequently imputed as NAs to avoid falsifying the later estimates and means (etc.)

The following plots show the the composition of the three different cluster given the alcohol and marijuana consumption choices. The first plot is a direct replication of the plot on the bottom of page 5, while the second plot represent the cluster alocation given the untransformed alcohol and marijuana consumptions. In the third plot, the untransformed variables are show, however only values up to 12 average monthly consumed alcoholic beverages & 10 times avarage MJ consumptions are being displayed in order for the separation between cluster 1 and 2 to become visible apparent.

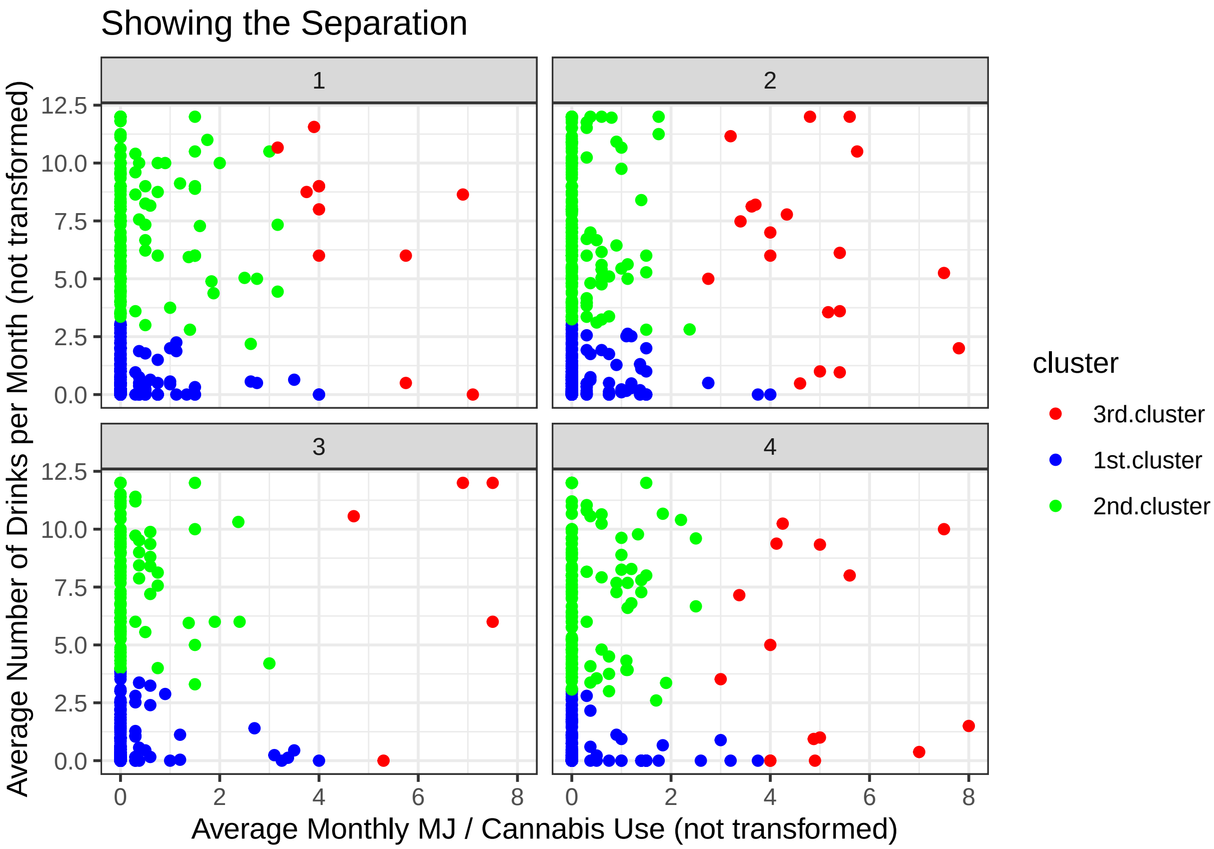
```r
data.file.long <- data.file.long %>% mutate(Cluster_current = as.factor(Cluster_current),
                                  BARCS_ID = as.factor(BARCS_ID))
cluster.colors <- c('1st.cluster' = "blue", '2nd.cluster' = "green", '3rd.cluster' = "red")
cluster.title <- "Cluster"
plot.page5 <- ggplot(data.file.long,
                     aes(x = LOG_Avg_MJ_current, y = LOG_Avg_Drinks_current, col = Cluster_current)) +
  geom_point() +
  xlab("Average Monthly MJ / Cannabis Use (Log10 Transformed)") +
  ylab("Average Number of Drinks per Month (Log10 Transformed)") +
  labs(colour = "cluster") +
  ggtitle("Cluster allocation given Alcohol & MJ consumption (log10 transformed)") +
  scale_colour_manual(values = cluster.colors, na.translate = FALSE) + theme_bw() +
  facet_wrap(~ Semester)

suppressWarnings(print(plot.page5))
```



```
plot.page5.nottransformed <- ggplot(data.file.long,
                          aes(x = Avg_MJ_current, y = Avg_Drinks_current, col = Cluster_current)) +
  geom_point(alpha = 0.25) + #ylim(0, 50) +
  xlab("Average Monthly MJ / Cannabis Use (not transformed)") +
  ylab("Average Number of Drinks per Month (not transformed)") +
  labs(colour = "cluster") +
  ggtitle("Cluster allocation given Alcohol & MJ consumption") +
  scale_colour_manual(values = cluster.colors, na.translate = FALSE) + theme_bw() +
  facet_wrap(~ Semester)

suppressWarnings(print(plot.page5.nottransformed))
```



```
plot.page5.nottransformed.focused <- ggplot(data.file.long,
                        aes(x = Avg_MJ_current, y = Avg_Drinks_current, col = Cluster_current)) +
  geom_point(alpha = 1) + ylim(0, 12) + xlim(0, 8) +
  xlab("Average Monthly MJ / Cannabis Use (not transformed)") +
  ylab("Average Number of Drinks per Month (not transformed)") +
  labs(colour = "cluster") +
  ggtitle("Showing the Separation") +
  scale_colour_manual(values = cluster.colors, na.translate = FALSE) + theme_bw() +
  facet_wrap(~ as.factor(Semester))

suppressWarnings(print(plot.page5.nottransformed.focused))
```



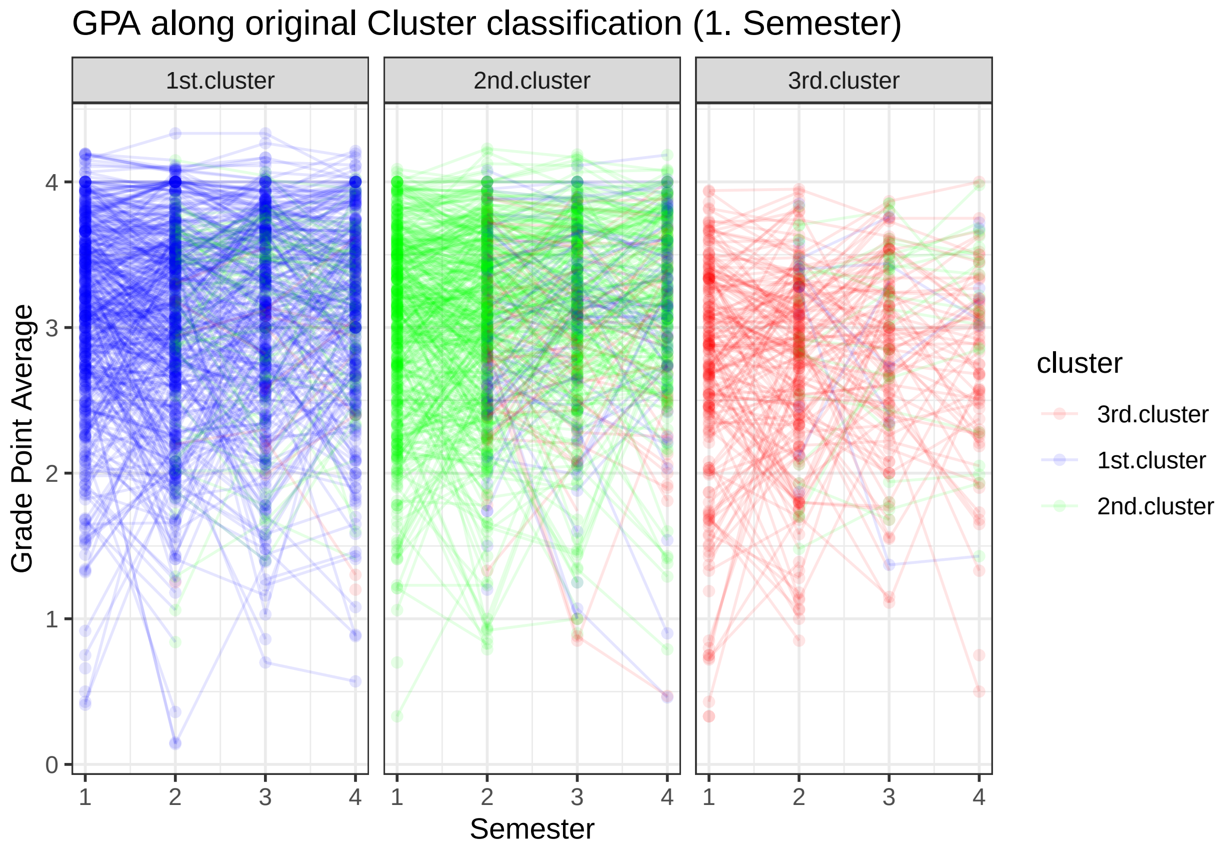
```
#cowplot::plot_grid(plot.page5, plot.page5.nottransformed, nrow = 2) ##just not good
```

It is noteworthy that the separation between Cluster 1 and 2 is around 3 monthly average alcohol beverages, though the hyperplane of the separation varies across the semester.

```
gpa.spaghetti <- ggplot(data.file.long %>% filter(!is.na(Cluster_SEM1)), aes(x = Semester, y = GPA, grou
  geom_point(alpha = 0.1) +
  geom_line(alpha = 0.1) +
  xlab("Semester") + ylab("Grade Point Average") + labs(colour = "cluster") +
  ggtitle("GPA along original Cluster classification (1. Semester)") +
  facet_wrap(~Cluster_SEM1) + scale_colour_manual(values = cluster.colors, na.translate = FALSE) + theme
gpa.spaghetti
```

```
## Warning: Removed 759 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

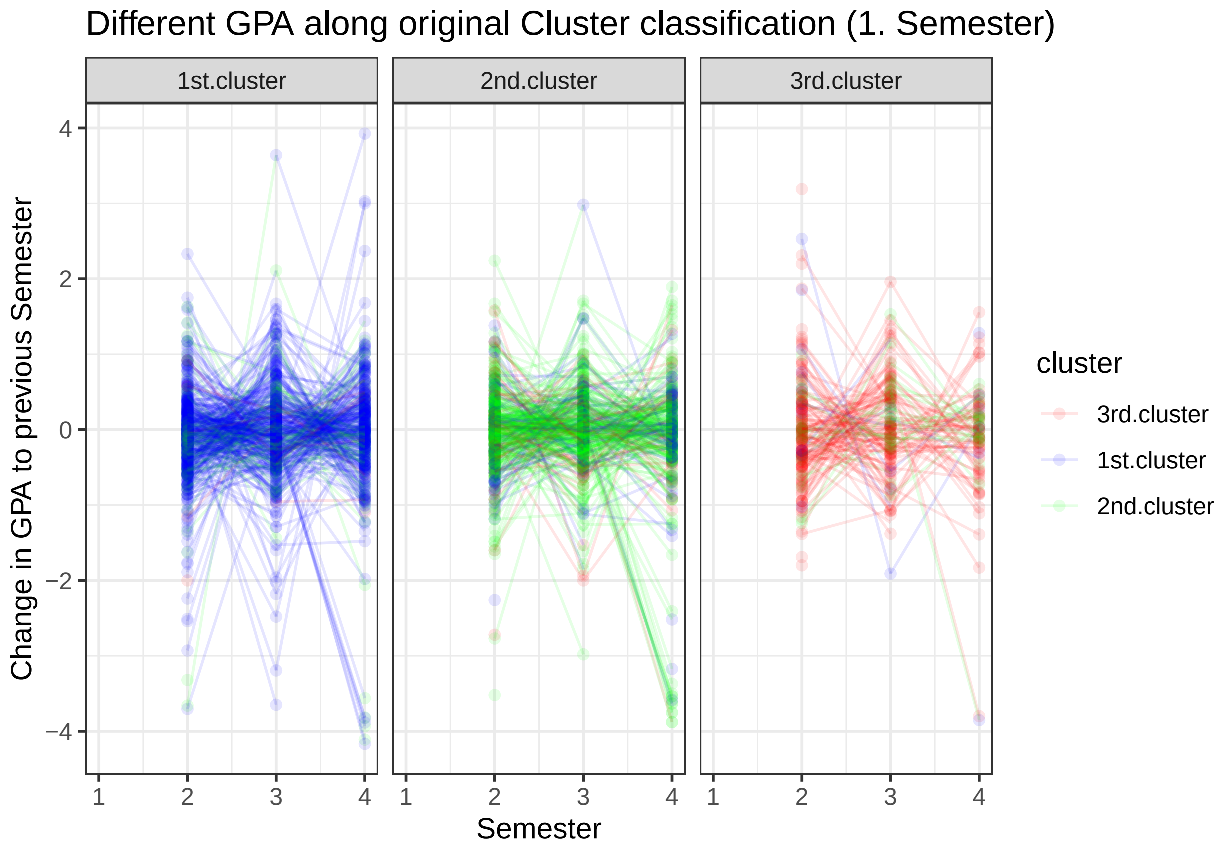
```
## Warning: Removed 711 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



```
gpa.spaghetti.diff <- ggplot(data.file.long %>% filter(!is.na(Cluster_SEM1)), aes(x = Semester, y = dif
  geom_point(alpha = 0.1) +
  geom_line(alpha = 0.1) +
  xlab("Semester") + ylab("Change in GPA to previous Semester") + labs(colour = "cluster") +
  ggtitle("Different GPA along original Cluster classification (1. Semester)") +
  facet_wrap(~Cluster_SEM1) + scale_colour_manual(values = cluster.colors, na.translate = FALSE) + them
gpa.spaghetti.diff
```

```
## Warning: Removed 1857 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

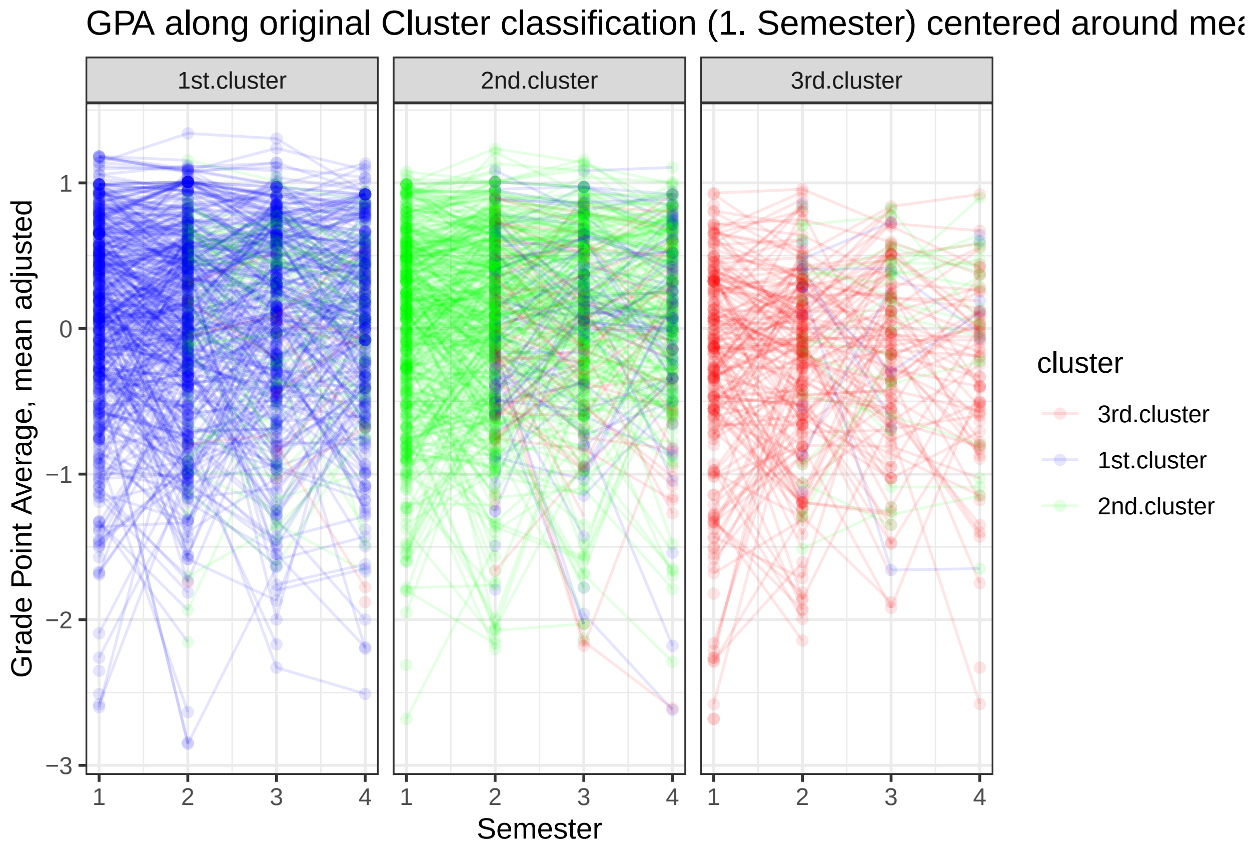
```
## Warning: Removed 1829 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



```
gpa.spaghetti.mean <- ggplot(data.file.long %>% filter(!is.na(Cluster_SEM1)), aes(x = Semester, y = mear
  geom_point(alpha = 0.1) +
  geom_line(alpha = 0.1) +
  xlab("Semester") + ylab("Grade Point Average, mean adjusted") + labs(colour = "cluster") +
  ggtitle("GPA along original Cluster classification (1. Semester) centered around mean") +
  facet_wrap(~Cluster_SEM1) + scale_colour_manual(values = cluster.colors, na.translate = FALSE) + theme
gpa.spaghetti.mean
```

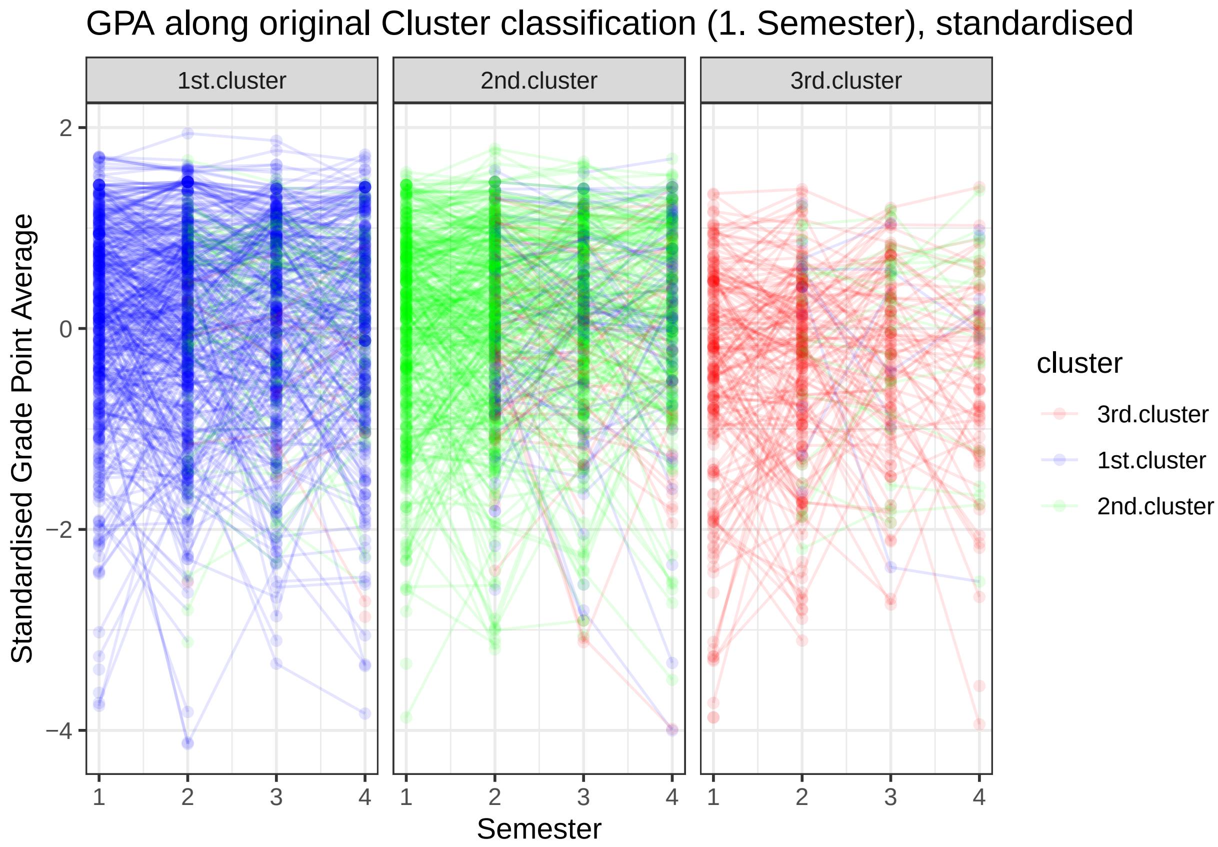
```
## Warning: Removed 759 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning: Removed 711 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



```
gpa.spaghetti.std <- ggplot(data.file.long %>% filter(!is.na(Cluster_SEM1)), aes(x = Semester, y = std_(
  geom_point(alpha = 0.1) +
  geom_line(alpha = 0.1) +
  xlab("Semester") + ylab("Standardised Grade Point Average") + labs(colour = "cluster") +
  ggtitle("GPA along original Cluster classification (1. Semester), standardised") +
  facet_wrap(~Cluster_SEM1) + scale_colour_manual(values = cluster.colors, na.translate = FALSE) + theme(
gpa.spaghetti.std
```

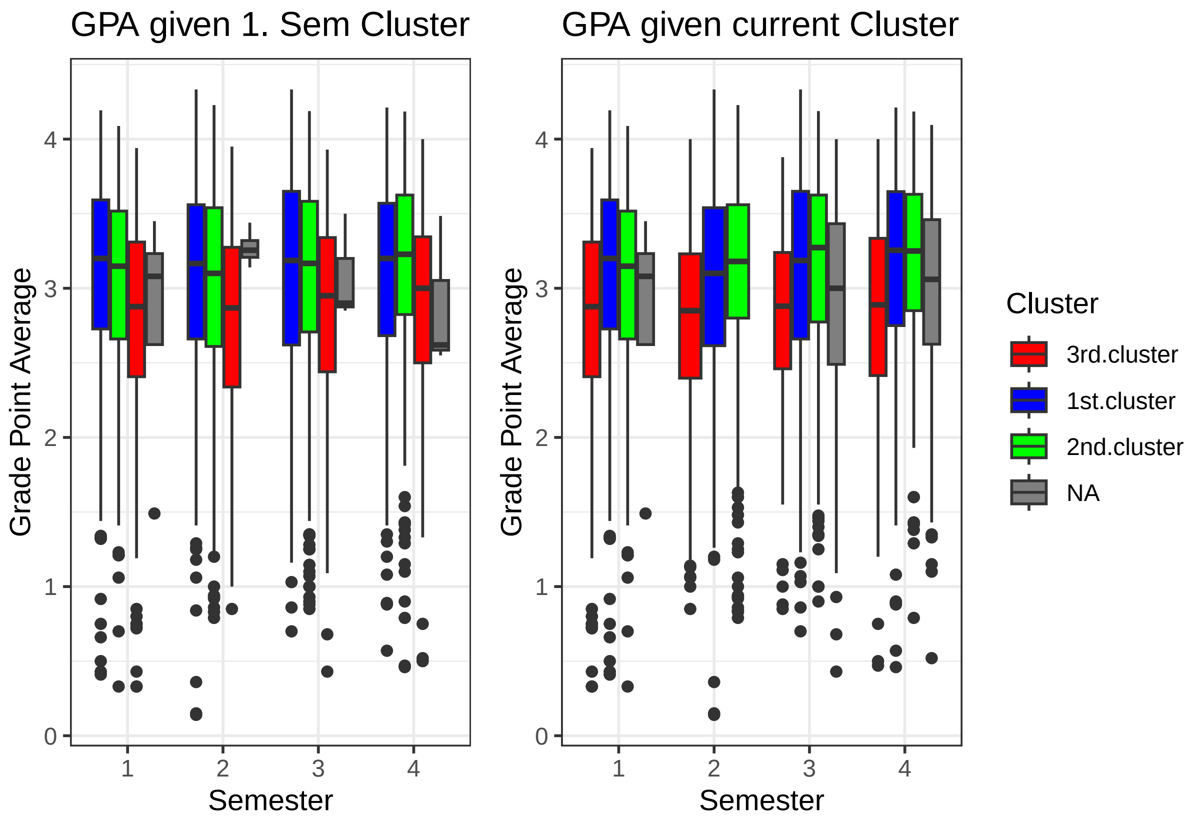
```
## Warning: Removed 759 rows containing missing values or values outside the scale range
## (`geom_point()`).
## Removed 711 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



```
plot.cluster.sem1.GPA <- ggplot(data = data.file.long, aes(x=as.factor(Semester), y=GPA)) +
  geom_boxplot(aes(fill=Cluster_SEM1)) +
  xlab("Semester") + ylab("Grade Point Average") +
  ggtitle("GPA given 1. Sem Cluster") +
   scale_fill_manual(values = cluster.colors, na.translate = TRUE) + theme_bw() + guides(fill="none")
#plot.cluster.sem1.GPA

plot.cluster.current.GPA <- ggplot(data = data.file.long, aes(x = as.factor(Semester), y=GPA)) +
  geom_boxplot(aes(fill=Cluster_current)) +
  xlab("Semester") + ylab("Grade Point Average") + guides(fill=guide_legend(title="Cluster")) +
  ggtitle("GPA given current Cluster") +
  scale_fill_manual(values = cluster.colors, na.translate = TRUE) + theme_bw()
#plot.cluster.current.GPA

plot.cluster.sem1.GPA + plot.cluster.current.GPA
```

```
## Warning: Removed 362 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

```
## Warning: Removed 362 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

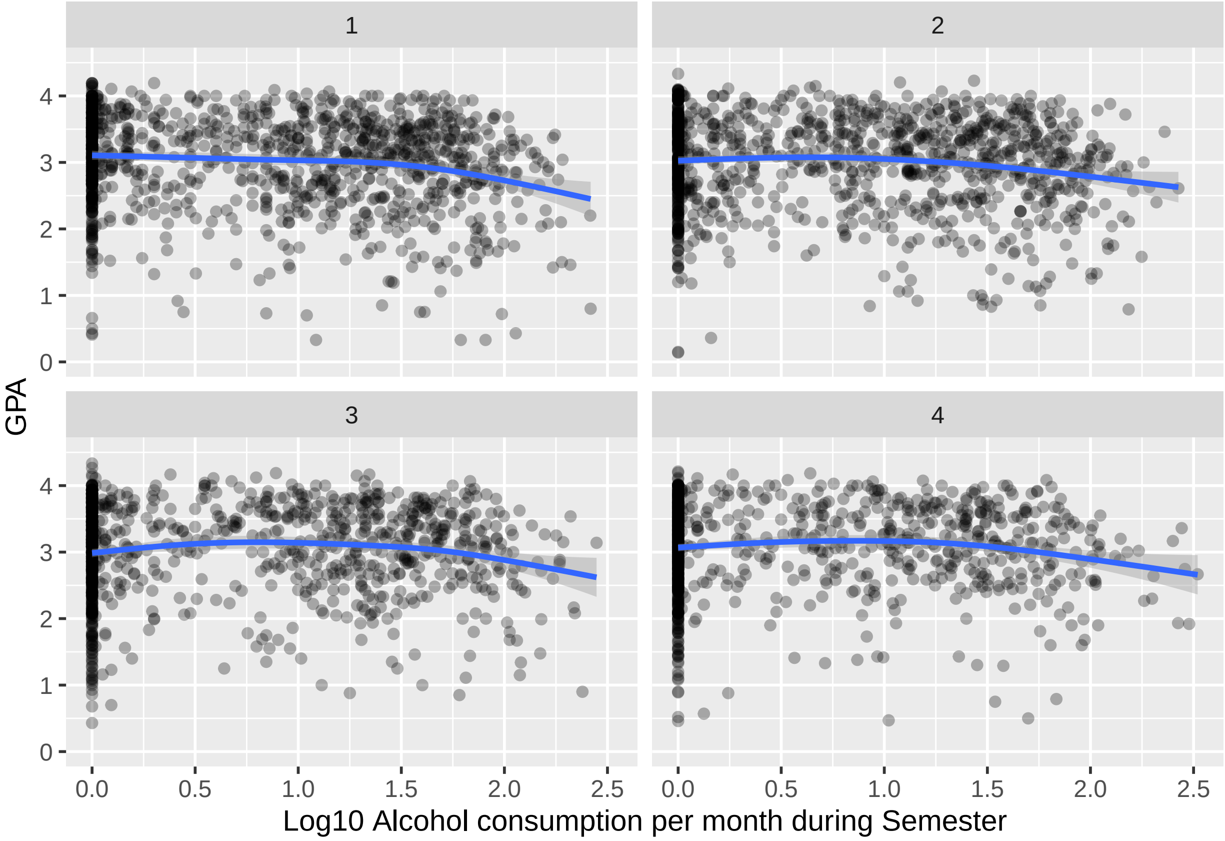


```r
plot.log.alcoholGPA <- ggplot(data = data.file.long, aes(x=LOG_Avg_Drinks_current, y=GPA)) +
  geom_point(alpha = 0.3) + geom_smooth() + ylim(0, 4.5) + facet_wrap(~Semester) +
  labs(x="Log10 Alcohol consumption per month during Semester" , y="GPA")
plot.log.alcoholGPA
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 362 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

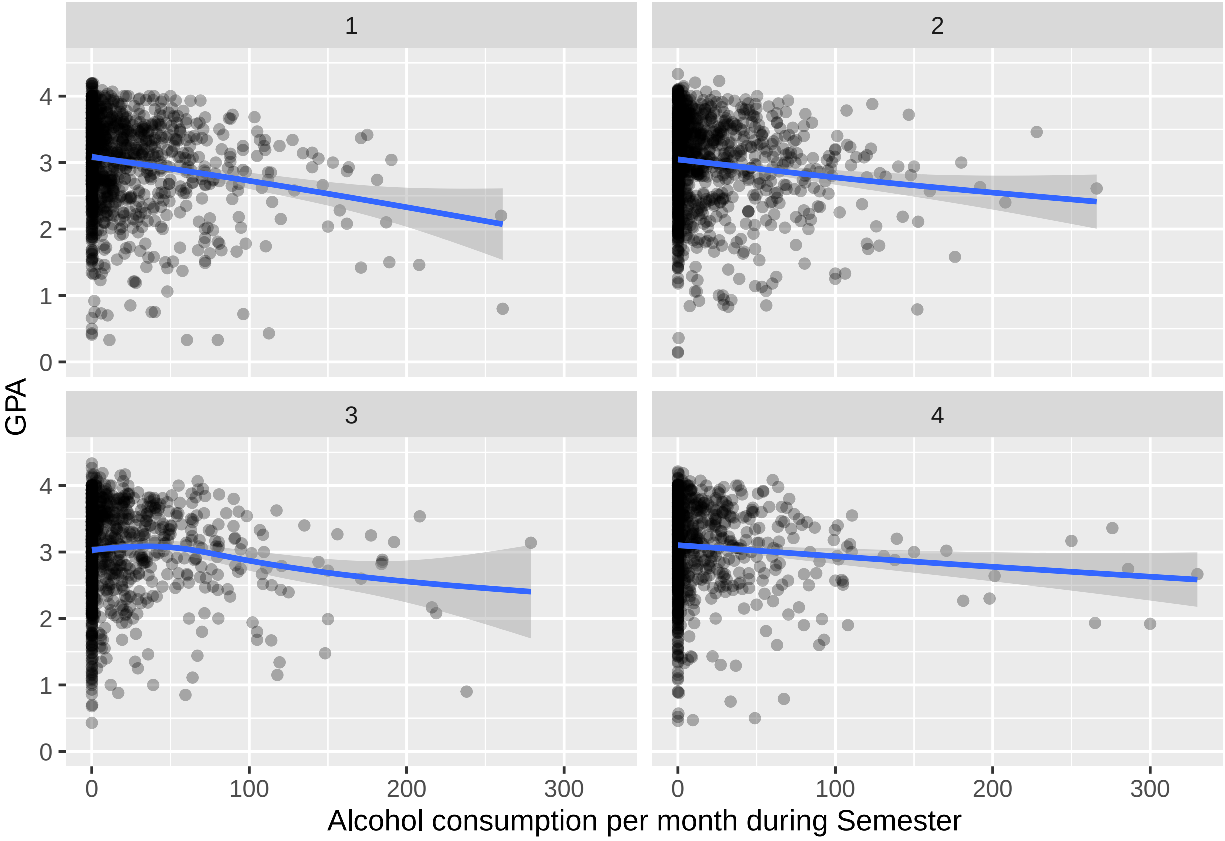
```
## Warning: Removed 362 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



```
plot.alcoholGPA <- ggplot(data = data.file.long, aes(x=Avg_Drinks_current, y=GPA)) +
  geom_point(alpha = 0.3) + geom_smooth() + ylim(0, 4.5) + facet_wrap(~Semester) +
  labs(x="Alcohol consumption per month during Semester" , y="GPA")
plot.alcoholGPA
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

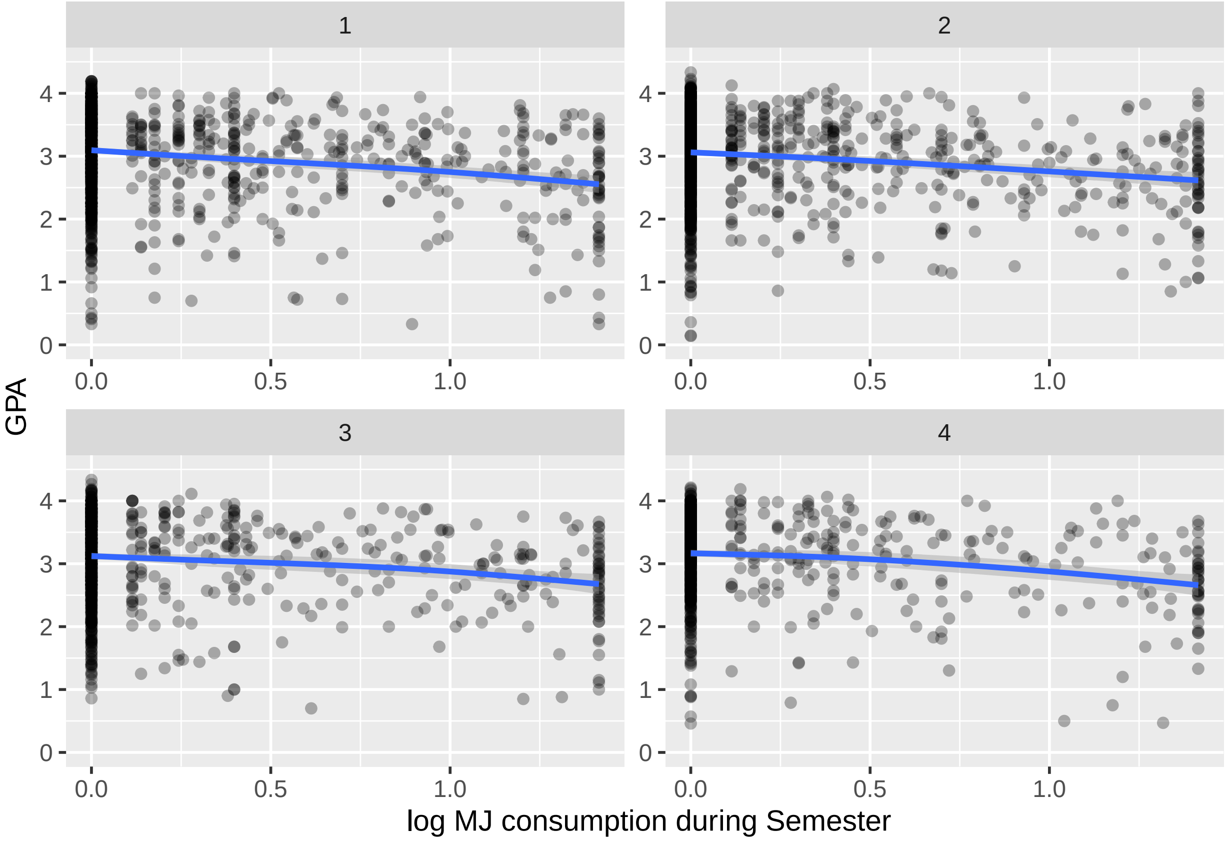
```
## Warning: Removed 362 rows containing non-finite outside the scale range
## (`stat_smooth()`).
## Removed 362 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Relationship appears to be nonlinear for log transformed Alcohol variable, and somewhat non linear for the transformed variable

```
plot.log.MJGPA<-  ggplot(data = data.file.long, aes(x=LOG_Avg_MJ_current, y=GPA)) +
  geom_point(alpha = 0.3) + geom_smooth() + ylim(0, 4.5) + facet_wrap(~Semester, scales = "free") +
  labs(x=" log MJ consumption during Semester" , y="GPA")
plot.MJGPA<-  ggplot(data = data.file.long, aes(x=Avg_MJ_current, y=GPA)) +
  geom_point(alpha = 0.3) + geom_smooth() + ylim(0, 4.5) + facet_wrap(~Semester, scales = "free") +
  labs(x="MJ consumption during Semester" , y="GPA")
plot.log.MJGPA
```

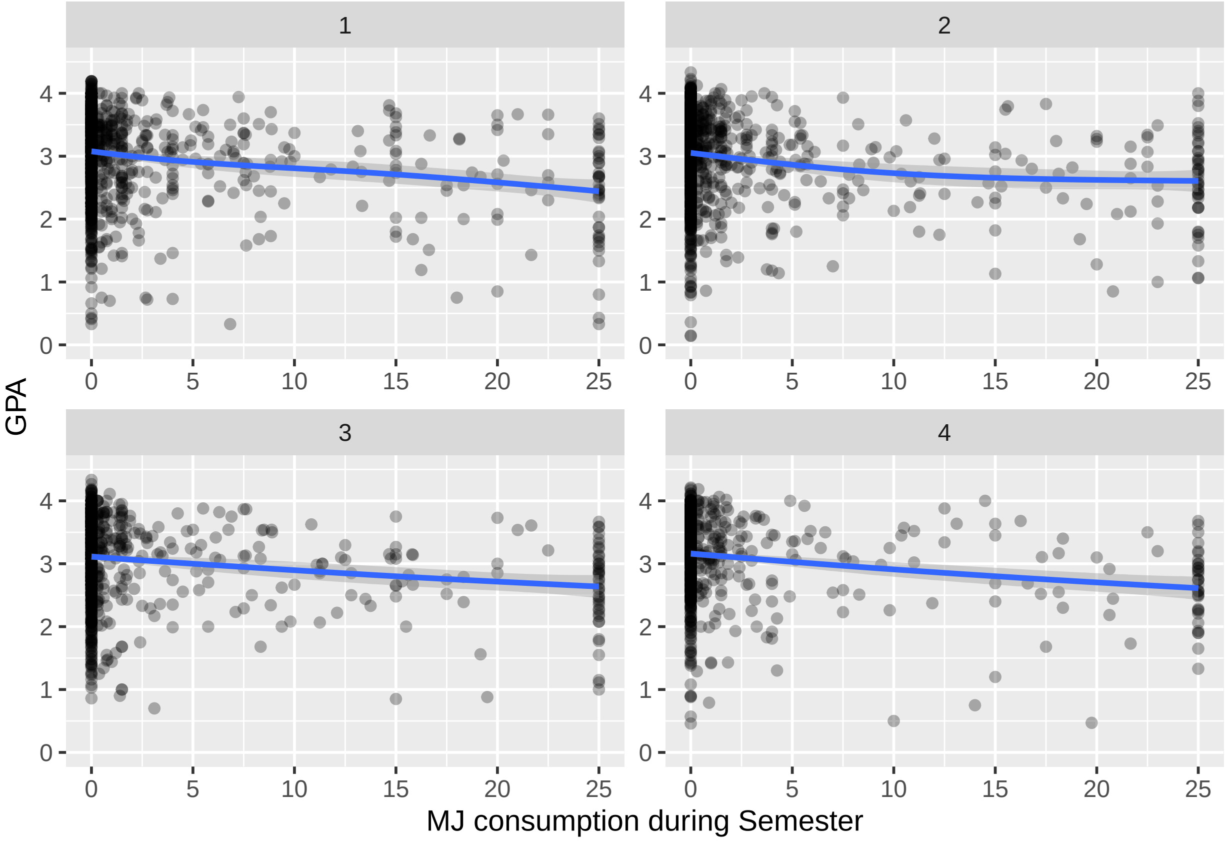
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

## Warning: Removed 766 rows containing non-finite outside the scale range
## (`stat_smooth()`).

## Warning: Removed 766 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



```
plot.MJGPA
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 766 rows containing non-finite outside the scale range
## (`stat_smooth()`).
## Removed 766 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



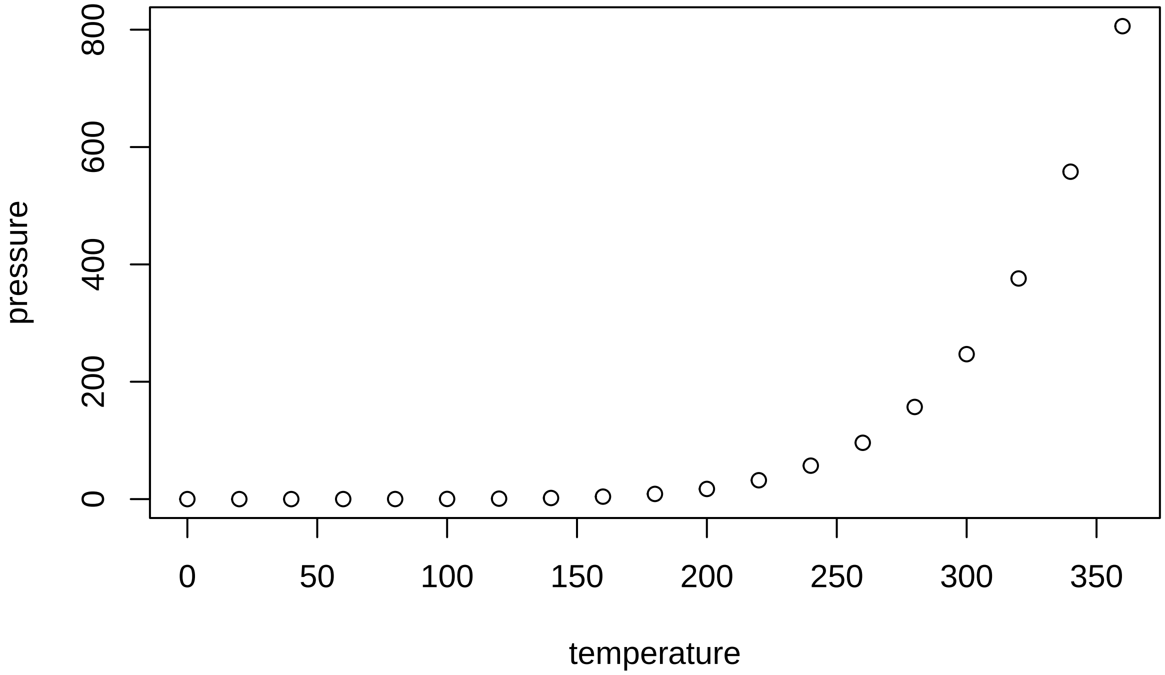
relationship seems to be mostly linear for both transformed and untransformed MJ usage

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:


pressure
0
200
400
600
800
temperature
0
50
100
150
200
250
300
350