

01 Descriptive

Colin Linke

2024-05-08

The data used by Meda et al. was gathered by a longitudinal study of first year students at to colleges in the North East of the United States using the semester System. The students in the study were given a an identification number and self reported their grades over 4 Semester, the amount of alcoholic drinks consumed and the amount of time they consumed marijuana. The gathered metrics for substance use did not come with further information on the metrical amount consumed, type or way of consumption, however there was an additional variable which tracked the amount of times a student on average consumed alcohol in the past month. Due to the fact that the authors of Meda et al. did not look further into this variable in the entirety of their paper, I dropped it in the replication attempt as well. Students were also provided a way to provide an interval for each consumed substance, in which case the middle point of the interval was used in the aggregation to the average amount consumed per semester.

In Mena et al. it is mentioned that the amount of times Marijuana was consumed was further classified into categories on a scale from 1 to 6, with 6 being the categories with the largest consumption. They've also stated that the substance use data was transformed using a logarithmic function, with the logarithmic being provided in the accessible data. By replicating the table on page 7 it becomes apparent that the base used in the logarithmic function is 10, however this leads to a contradiction in the supposed data generating process. The highest value in the provided MJ consumption data is 1.41, the required base to transform to the categorie 6 would be roughly 2.88. After visualizing the data (see below) it seems likely that solely the logarithmic value was used and a value of 25 was used as a cut of point for higher values.

Additionally, additional metrics for each participating student were gathered, such as their age, Their SAT scores, including the scores they scored in each part of it, their Parental Socio Economic situation (Parental_SES), the Beck Depression Index (BDI), the students Gender (Sex), their smoking status (Fager4_binary), their Family History of for Alcoholism (FH_binary, with positive implying that such a history exists) and their State Traite Anxiety Score (STAI), with the Parental SES, BDI, and STAI being questionnaires resulting in scores, with a higher score implying a higher socio economic parental standing, higher likelihood of developing a depression and higher Anxiety levels respectively, while Sex, Family History for Alcoholism and smoking status being categorical values. These variables were gathered once for each Student at the beginning of the study

A total of 1142 students participated over 4 semester with a 95 % possible participation rate. The authors followed this up grouping the substance usage into three clusters, with them being no to low alcohol usage / no to low marijuana usage, medium to high alcohol usage / no to low marijuana usage and medium to high alcohol usage / medium to high marijuana usage, which is explained further below. The following table serves as a direct replication of the table provided on page 7, aswell as an overview of the used data stratified by the clusters the students were placed into in Semester 1:

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
```

Both tables are generally identical, with the sole exception that for the Continuous Variables I used a Chi Squared Test, which discretizes the data and is therefore an approximation, unlike Mena et al. which uses an ANOVA test to directly compare the groups. Due to the large sample size there are only very minor differences and both have qualitative equal results, with all but Age, the SAT Sections on Verbal skills and Family History of Alcoholism having significant differences between the stratified clustered. (footnote, testing if the substance usage significantly differ between the clusters is somewhat trivial and self proofing, as the clusters were directly created by clustering on alcohol and marijuana usage.)

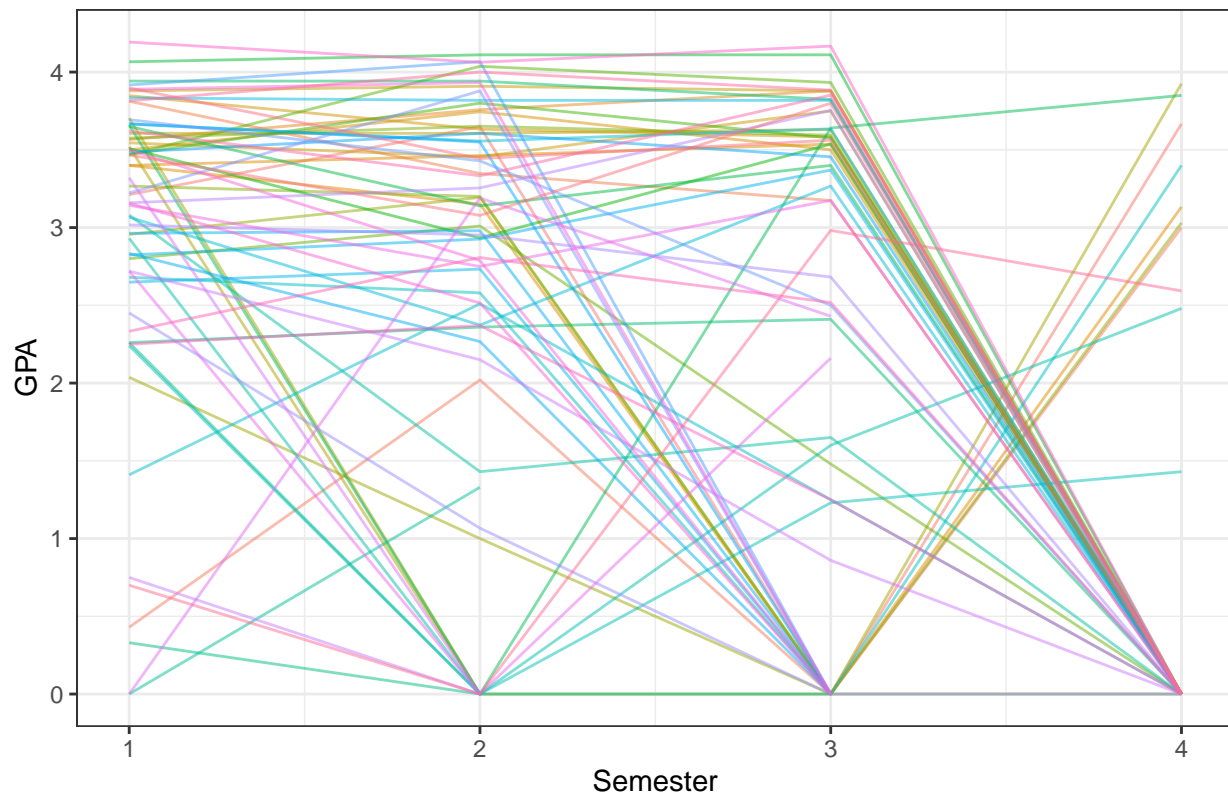
There are also minor differences in the Mean and standard deviation, this can mostly down to algorithmic alculatation or rounding of these values with the exceptions to the GPA and SAT Section. In the provided data there were multiple instances in which these variables have entries of 0. In the case of the GPA this could for example be due to the fact that students are dropping out, or that they these are missing values. These cases can be seen here

```
plot.gpa0 + theme(legend.position="none")
```

```
## Warning: Removed 24 rows containing missing values or values outside the scale range
```

```
## (`geom_line()`).
```

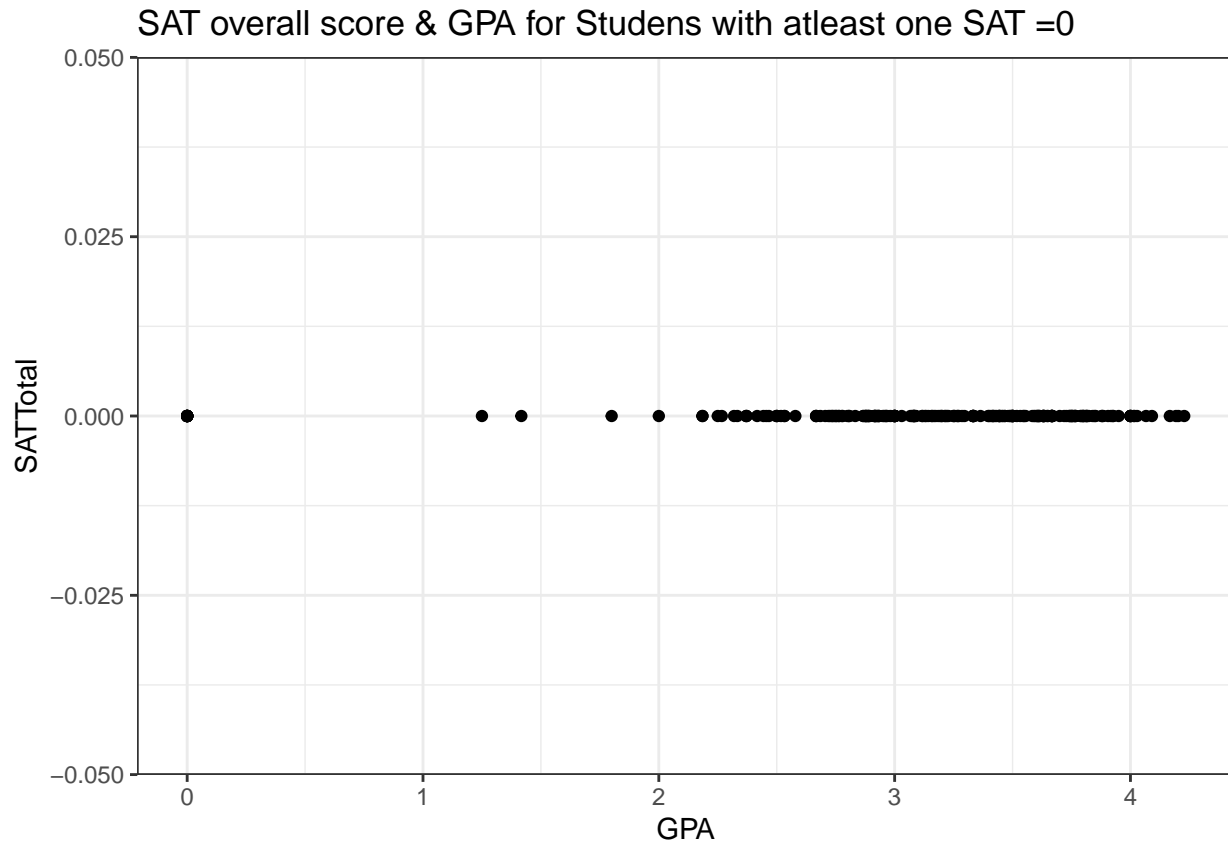
GPA over all Semester for Studens with atleast one GPA =0



```
ggplot(ind.sat0, aes(y = SATTotal, x = GPA)) + geom_point() + theme_bw() + ggtitle("SAT overall score & GPA")
```

```
## Warning: Removed 4 rows containing missing values or values outside the scale range
```

```
## (`geom_point()`).
```



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

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

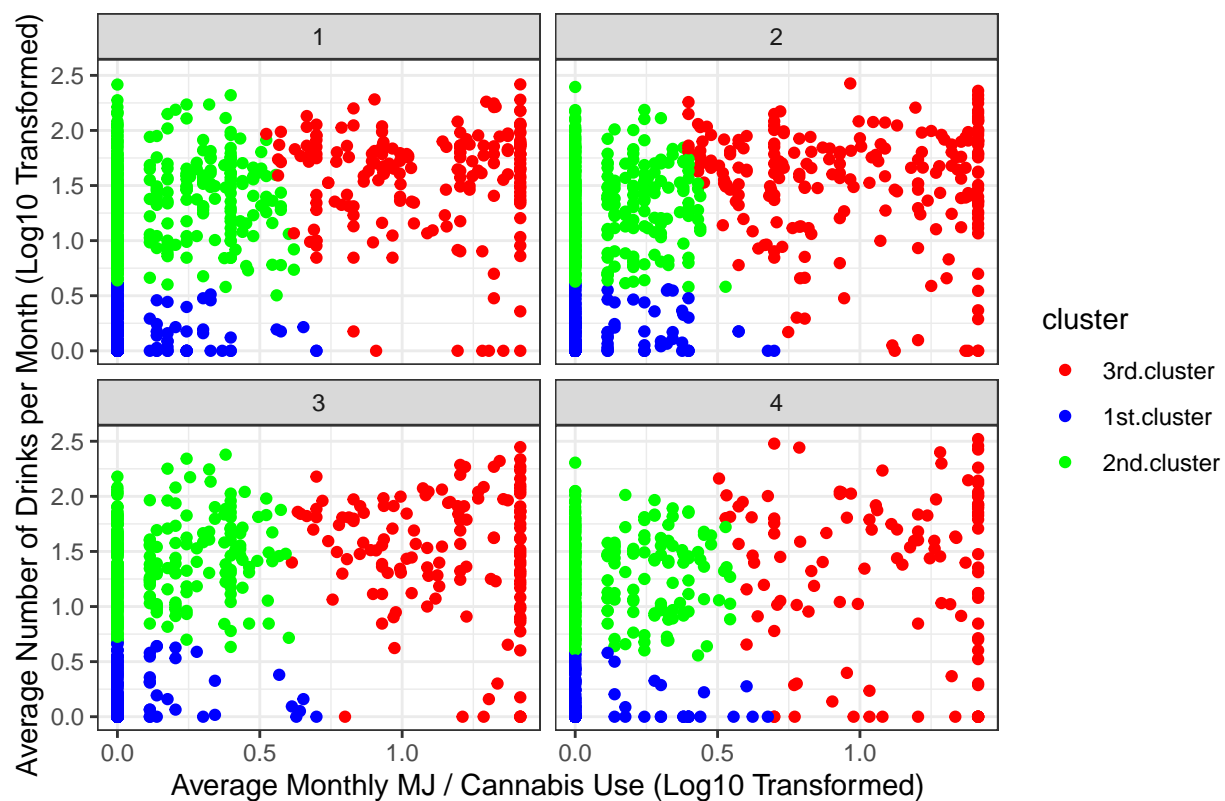
Considering that students usually have non zero GPA data points before and after they are assigned a value of '0', it seems likely that a lot of students were wrongly assigned '0' for NAs here. Additionally, a completed SAT can not have a point total of zero. It is also extremely likely '0' entries here are missing data, as most students also have a nonzero GPA, therefore implying that these are actually attendances. Zero entries were subsequently imputed as NAs to avoid introducing bias into later estimates.

All other provided data was taken as it was given and assumed to be true. It is generally not observable if the data is wrong or not and further adjusting it may introduce bias through wrong assumptions.

The following plots show the composition of the three different clusters 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 represents the cluster allocation given the untransformed alcohol and marijuana consumptions. In the third plot, the untransformed variables are shown, however only values up to on average 12 monthly consumed alcoholic beverages & 10 times average MJ consumptions are being displayed in order for the separation between cluster 1 and 2 to become more visible apparent.

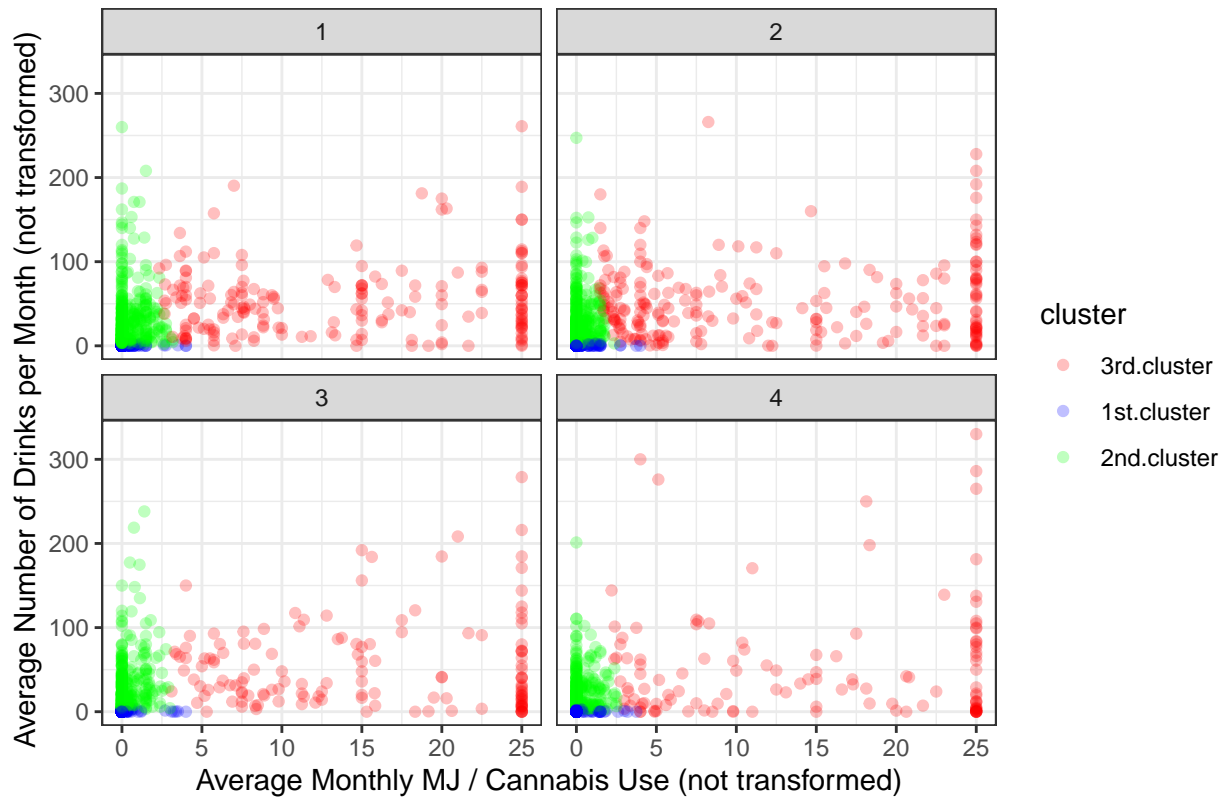
```
suppressWarnings(print(plot.page5))
```

Cluster allocation given Alcohol & MJ consumption (log10 transformed)



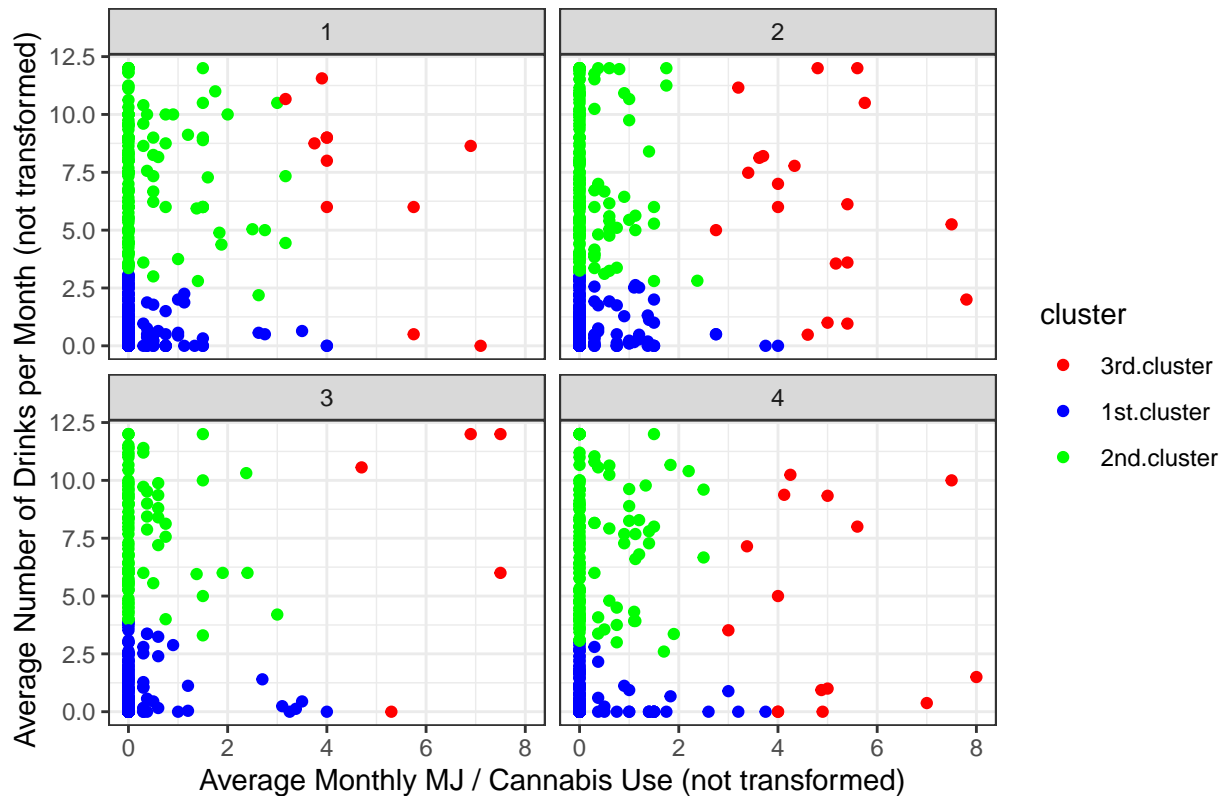
```
suppressWarnings(print(plot.page5.nottransformed))
```

Cluster allocation given Alcohol & MJ consumption



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

Showing the Separation



The first plot is a direct replication of the shown graphic on page 5 of Mena et al. Here they show the relationship between the consumption use and the assigned Cluster in each of the 4 semester for each student. Mena et al. used a two step clustering algorithm, an initially hierarchically ordering the substance usage, which was then used to create various cluster, from which the cluster structure with the smallest AIC score was chosen. The cluster are time varying and based on the AIC there wasn't an additional 4th cluster constructed to differentiate between No-low MJ usage and medium - high MJ usage for medium - high alcoholic usage. The reported kappa statistic for the interreliaby cluster coherence between subsequent semester to the first are

```
Sem1_to <- c("Sem 2", "Sem 3", "Sem 4")
kappa_value <- c(0.64, 0.70, 0.67)
t(data.frame(Sem1_to, kappa_value))
```

```
##           [,1]    [,2]    [,3]
## Sem1_to    "Sem 2" "Sem 3" "Sem 4"
## kappa_value "0.64"  "0.70"  "0.67"
```

with reported p value of < 0.001 , which suggest substantial agreement of cluster assignment over time.

These cluster assignments are taken as they were given by the authors of the paper, though this approach has a few short comings. Firstly, the clustering methods applied are not robust, different starting points can lead to different cluster assignments. Due to scarce information in Mena et al. it is unclear if additional steps were taking for a more robust cluster assignment. Secondly, the fact that the cluster assignment is time varying can make sense from the statistical perspective, however this means that students with identical substance usage across all semesters can be assigned to different clusters in the different semesters. Later interpretation of cluster parameters in regressions should therefor only be between Cluster assignment within a semester. In addition Mena et al. investigate effect the effect group transitions have on GPA, with group transition being used as a stand in for a significant change in substance usage. This approach can lead skewed results as minor consumption changes can lead to a new cluster assignment for some students, while for others much major

significant change is required, yet both would be counted the same. Thirdly, transforming two variables into a single factor offers the benefit of mitigating the impact of outliers, which is later shown to be warranted, and simplifies the analysis and interpretation, this leads to a loss in information. There is also not a risk of multicollinearity, as the correlation between the two variables is only moderately high at 0.41 (footnote 0.47 for the log transformed data).

```
## [1] 0.4090057
```

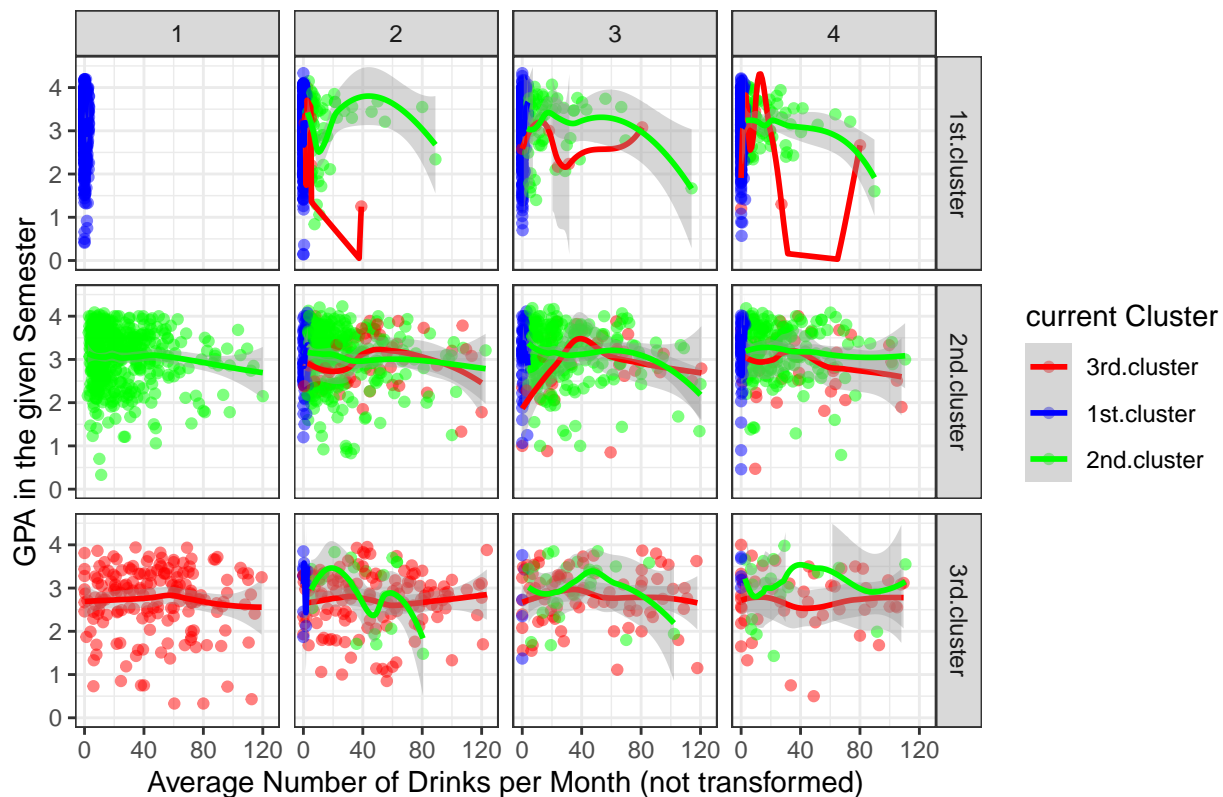
```
## [1] 0.467915
```

Lastly, the field of statistics is an interdisciplinary field and using only statistical methods can come at the cost of interpreting the results and / or potentially lack a real world context. It can be seen in the two untransformed plots above that the cutoff between Cluster 1 and Cluster 2 at roughly 3 alcoholic drinks per months, varying through the semester. The labeling of these clusters must be seen as a relative description between the clusters, but labeling the consumption of 5 alcoholic drinks per month as moderate alcohol usage seems subjective. The cluster assignments could have also been based on, or partially based on previously established guidelines or research. The National Institute on Alcohol Abuse and Alcoholism for example defines heavy drinking as “consuming five or more drinks on any day or 15 or more per week [for men, or] consuming four or more on any day or 8 or more drinks per week [for women]”, which would give an inherent interpretation of the cluster assignment. The same institute also funded the research of Mena et al. as is stated on page 1.

```
suppressWarnings(print(plot.alcgpa))
```

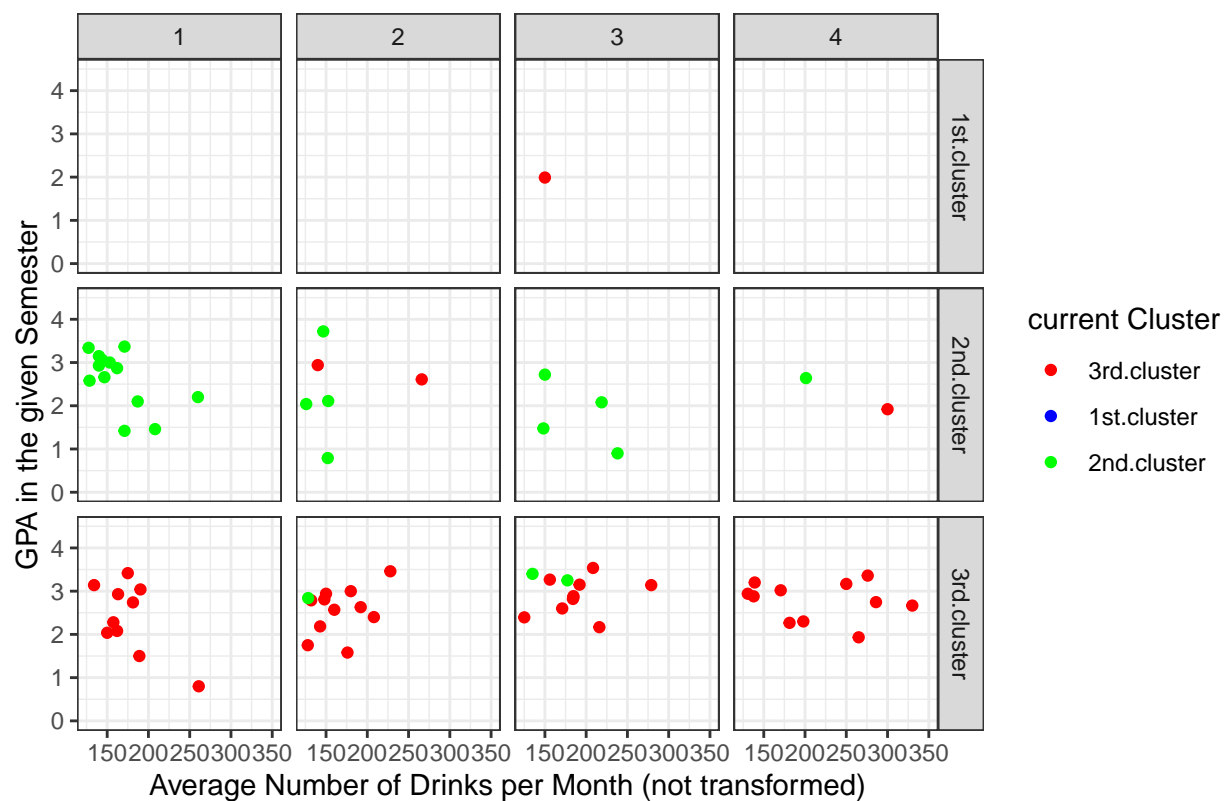
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

Alcohol consumption & GPA given original Cluster specification



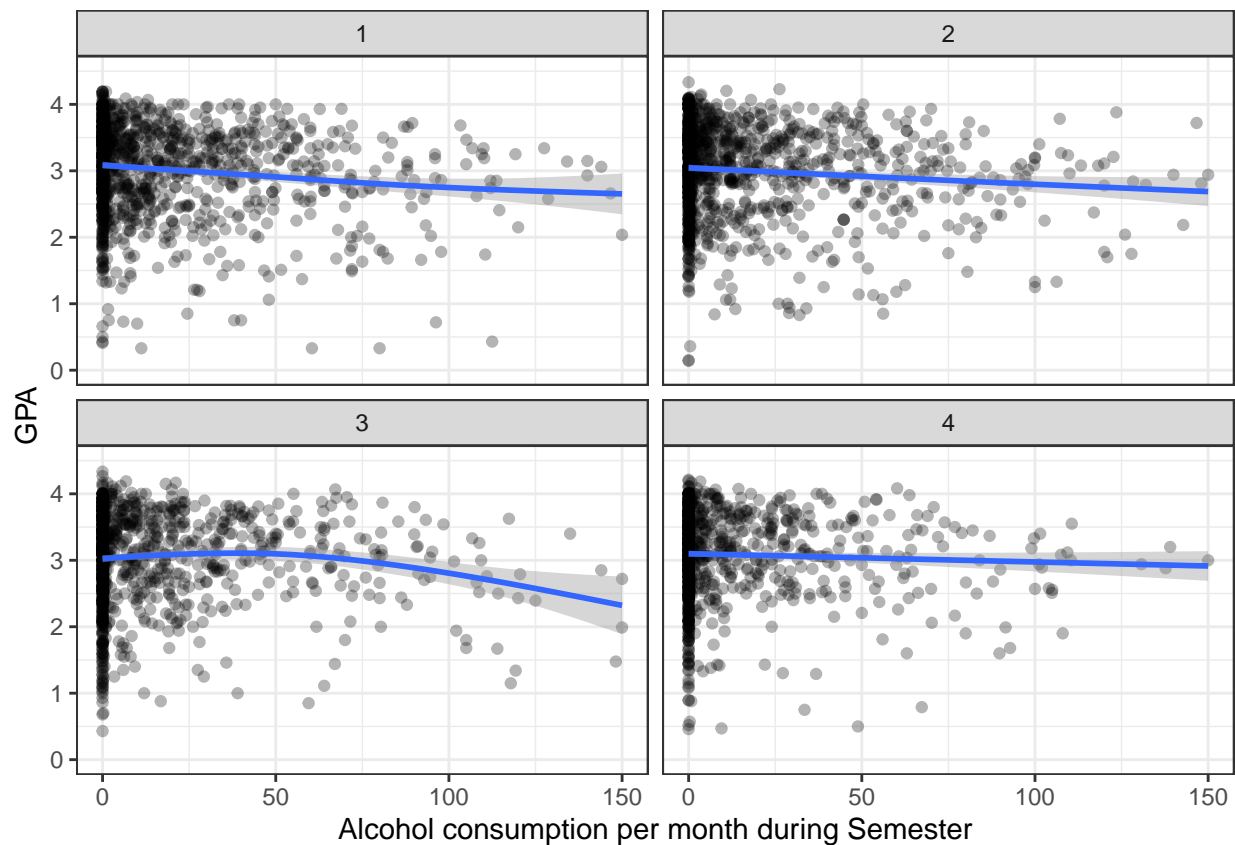
```
suppressWarnings(print(plot.alcgpa.outliers))
```


Alcohol consumption & GPA given original Cluster specification



```
suppressWarnings(print(plot.alcoholGPA))
```

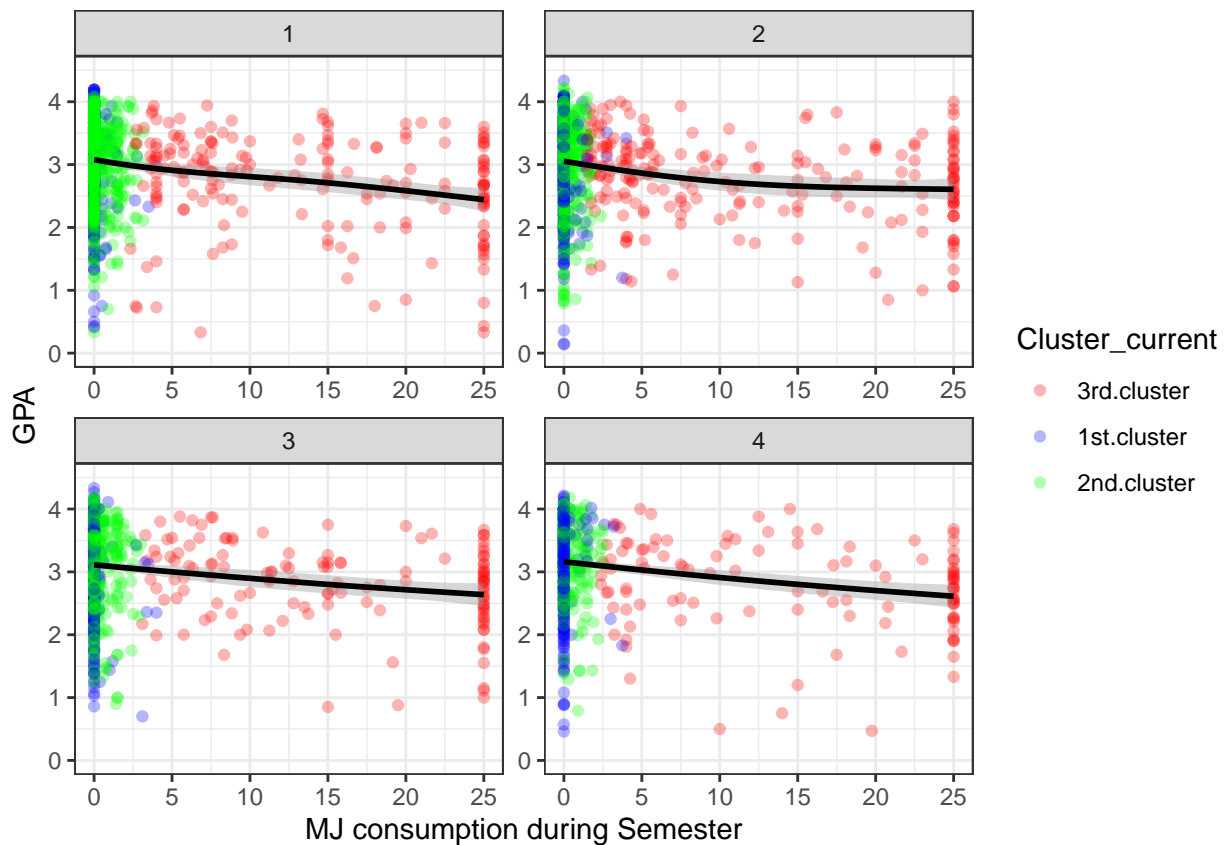
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



```
plot.MJGPA<- ggplot(data = data.file.long, aes(x=Avg_MJ_current, y=GPA, colour = Cluster_current)) +
  geom_point(alpha = 0.3) + geom_smooth(aes(group = 1), color = "black") + ylim(0, 4.5) + facet_wrap(~S
  labs(x="MJ consumption during Semester" , y="GPA") + scale_colour_manual(values = cluster.colors, na.
```

```
suppressWarnings(print(plot.MJGPA))
```

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



Seems like a linear trend with higher MJ consumption having a negative effect on GPA

Missing Data

```
table(data.file.long$sum.GPAna)
```

```
##
```

```
##    0    1    2    3    4
```

```
## 3692 412 364  92   8
```

```
data.file.long %>% group_by(Semester) %>% summarise(MeanGPA = mean(GPA, na.rm = TRUE)) %>%
  pivot_wider(names_from = Semester, values_from = MeanGPA, names_prefix = "Semester ")
```

```
## # A tibble: 1 x 4
```

```
##   `Semester 1` `Semester 2` `Semester 3` `Semester 4`
```

```
##           <dbl>           <dbl>           <dbl>           <dbl>
```

```
## 1           3.01           2.99           3.03           3.08
```

```
data.file.long %>% group_by(Semester, sum.GPAna) %>%
```

```
  summarise(MeanGPA = mean(GPA, na.rm = TRUE), .groups = 'drop') #>%
```

```
## # A tibble: 20 x 3
```

```
##   Semester sum.GPAna MeanGPA
```

```
##   <fct>      <int>   <dbl>
```

```
## 1 1         0      3.06
```

```
## 2 1         1      2.87
```

```
## 3 1         2      2.89
```

```
## 4 1         3      2.23
```

```
## 5 1         4     NaN
```

```
## 6 2         0      3.04
```

```
## 7 2      1      2.85
## 8 2      2      2.65
## 9 2      3      1.33
## 10 2     4      NaN
## 11 3     0      3.06
## 12 3     1      2.74
## 13 3     2      1.83
## 14 3     3      NaN
## 15 3     4      NaN
## 16 4     0      3.09
## 17 4     1      2.77
## 18 4     2      3.08
## 19 4     3      NaN
## 20 4     4      NaN
```

```
# pivot_wider(names_from = c(Semester, sum.GPAna), values_from = MeanGPA, names_prefix = "Semester ",
data.file.long %>% group_by(Semester, Cluster_SEM1) %>%
  summarise(missing_GPA = sum(is.na(GPA)), .groups = 'drop') #>%
```

```
## # A tibble: 16 x 3
##   Semester Cluster_SEM1 missing_GPA
##   <fct>      <chr>          <int>
## 1 1      1st.cluster          3
## 2 1      2nd.cluster          1
## 3 1      3rd.cluster          3
## 4 1      <NA>                0
## 5 2      1st.cluster        14
## 6 2      2nd.cluster        14
## 7 2      3rd.cluster          8
## 8 2      <NA>                0
## 9 3      1st.cluster        49
## 10 3     2nd.cluster        48
## 11 3     3rd.cluster        27
## 12 3     <NA>                1
## 13 4     1st.cluster        69
## 14 4     2nd.cluster        80
## 15 4     3rd.cluster        44
## 16 4     <NA>                1
```

```
# pivot_wider(names_from = c(Semester, Cluster_SEM1), values_from = missing_GPA, names_prefix = "Seme
```

censoring through dropout

APPENDIX.

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, group = Cluster_SEM1)) +
  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_minimal()
gpa.spaghetti
```

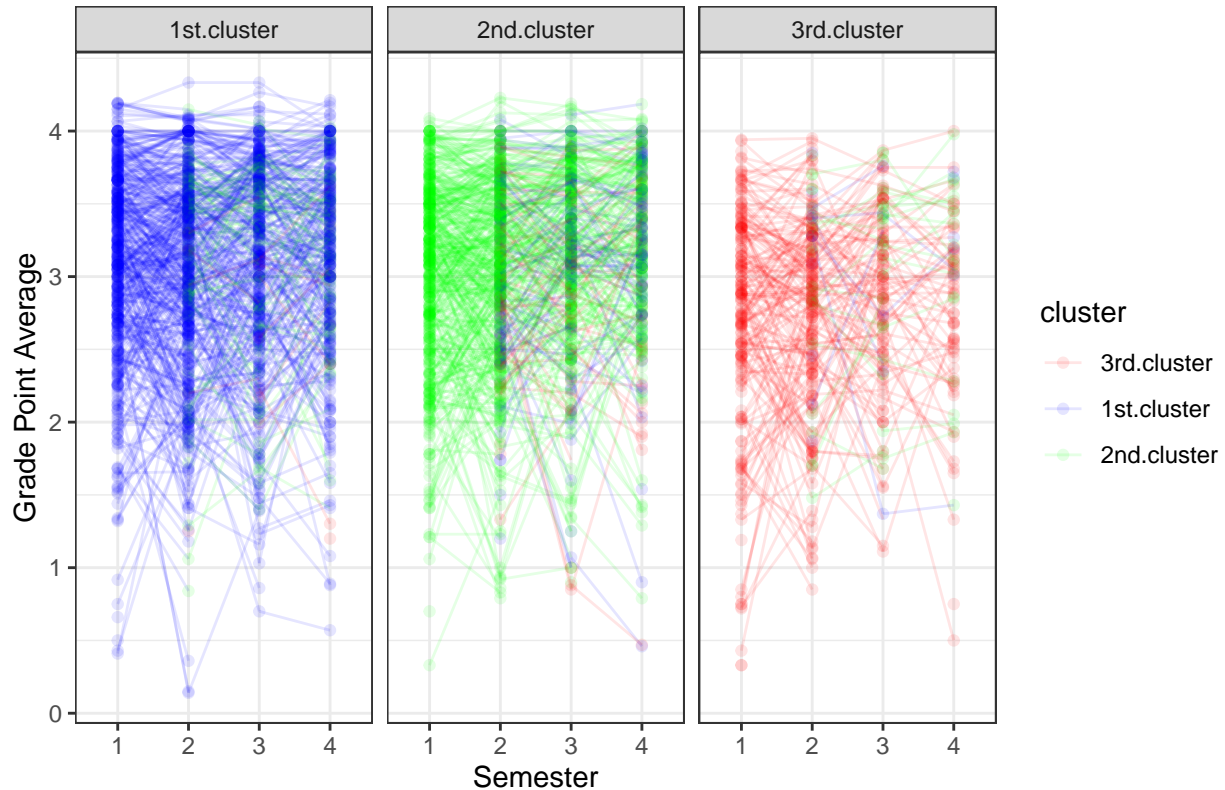
```
## Warning: Removed 759 rows containing missing values or values outside the scale range
```

```
## (`geom_point()`).
```

```
## Warning: Removed 711 rows containing missing values or values outside the scale range
```

```
## (`geom_line()`).
```

GPA along original Cluster classification (1. Semester)



```
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) + theme
gpa.spaghetti.diff
```

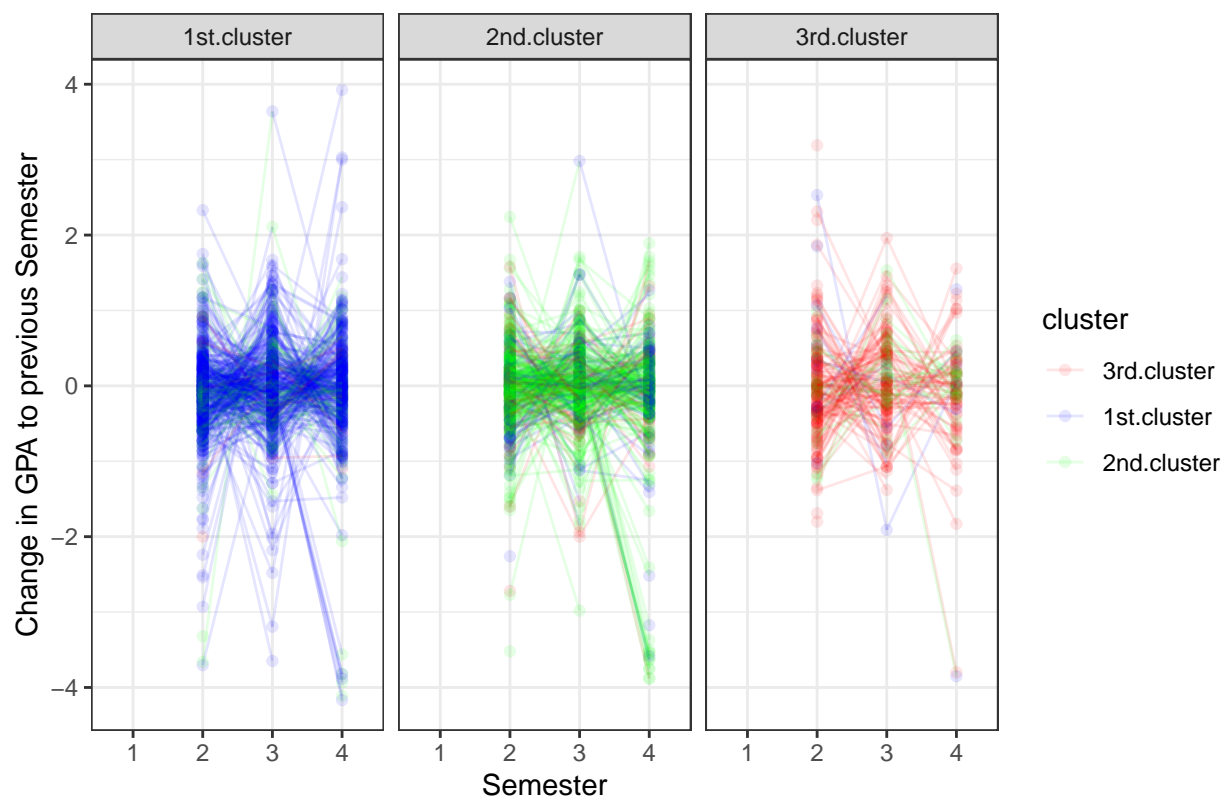
```
## Warning: Removed 1857 rows containing missing values or values outside the scale range
```

```
## (`geom_point()`).
```

```
## Warning: Removed 1829 rows containing missing values or values outside the scale range
```

```
## (`geom_line()`).
```

Different GPA along original Cluster classification (1. Semester)

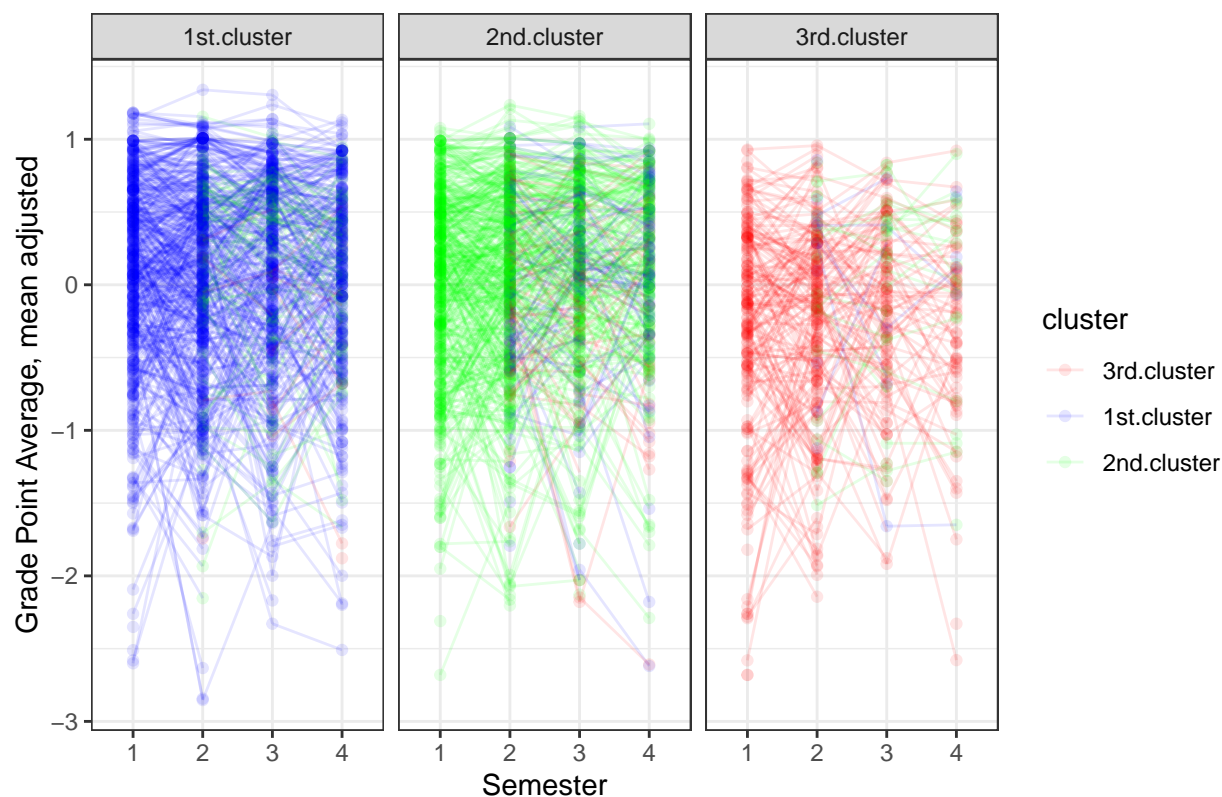


```
gpa.spaghetti.mean <- ggplot(data.file.long %>% filter(!is.na(Cluster_SEM1)), aes(x = Semester, y = mean_gpa)) +
  geom_point(alpha = 0.1) +
  geom_line(alpha = 0.1) +
  xlab("Semester") + ylab("Grade Point Average, mean adjusted") + labs(colour = "cluster") +
  ggtitle("GPA along original Cluster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

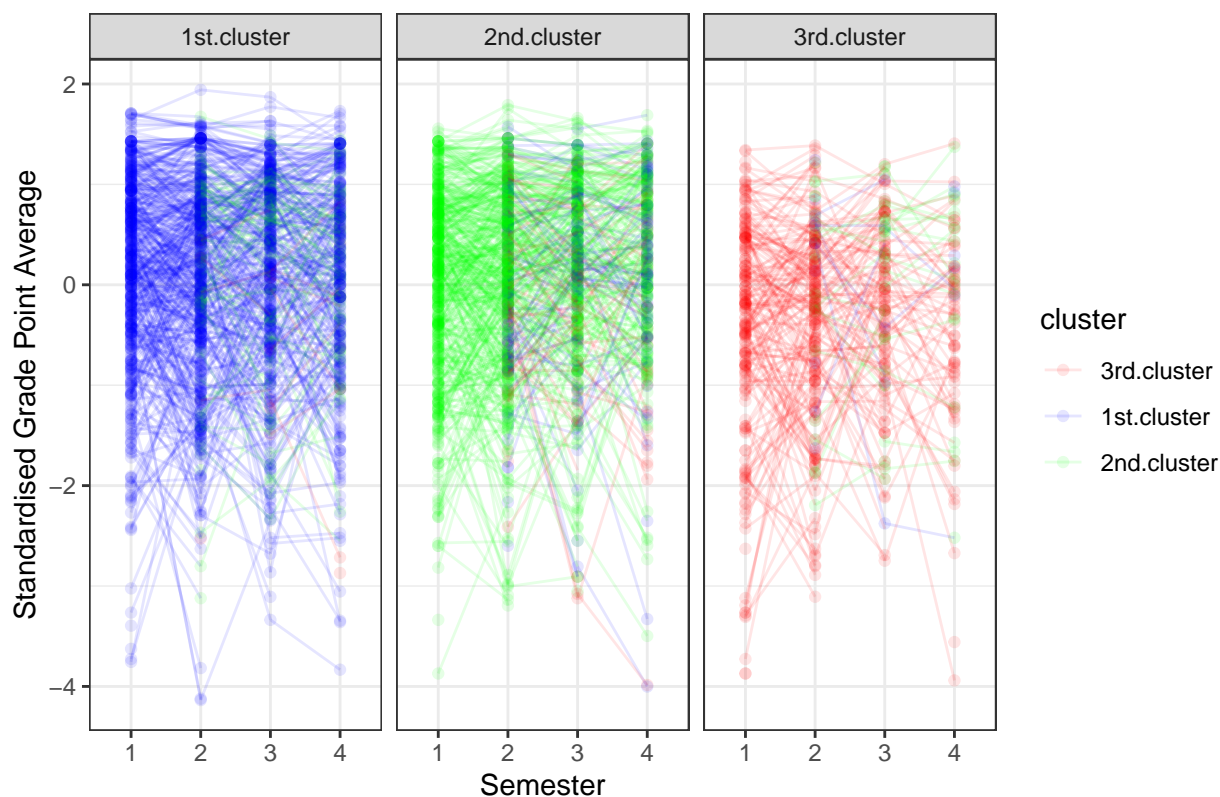
GPA along original Cluster classification (1. Semester) centered around mean



```
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()`).
```

GPA along original Cluster classification (1. Semester), standardised



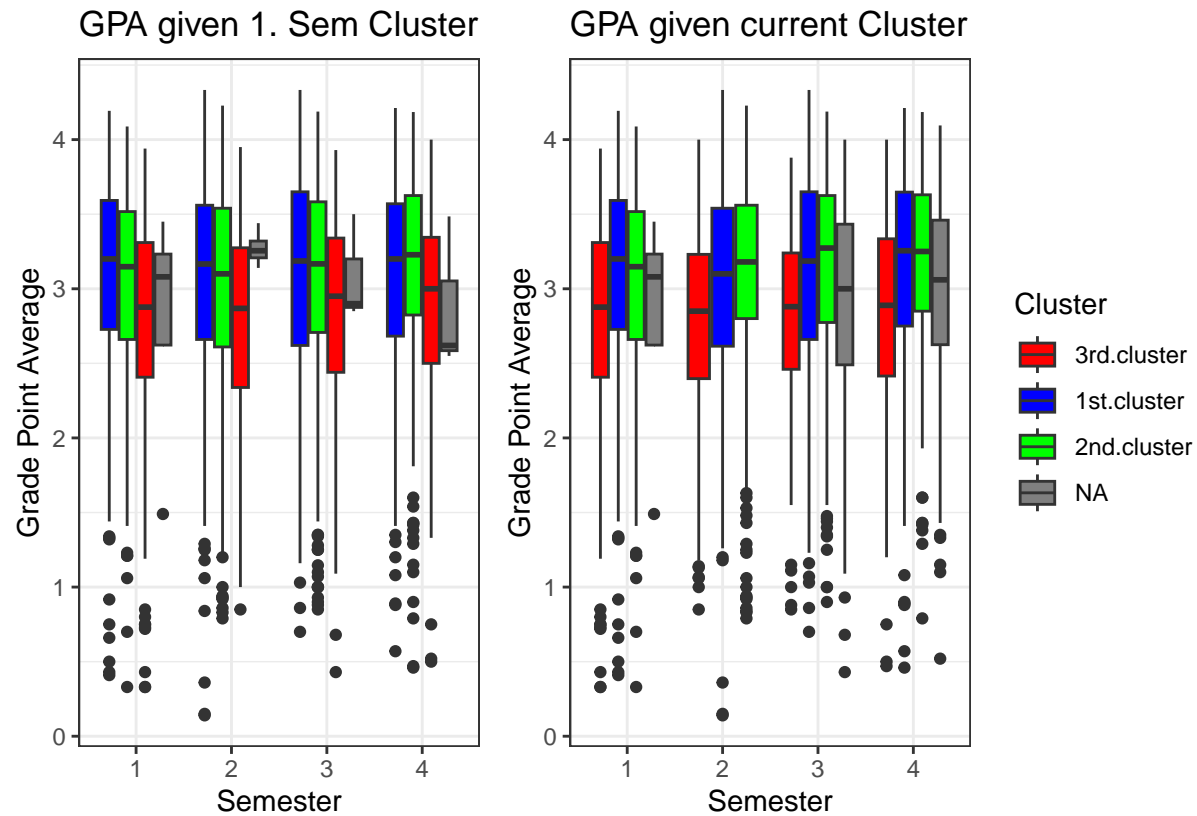
```
plot.cluster.sem1.GPA <- ggplot(data = data.file.long, aes(x=as.factor(Semester), y=GPA)) +
  geom_boxplot(aes(fill=Cluster_SEM1)) +
  xlab("Semester") + ylab("Grade Point Average") +
  ggtitle("GPA given 1. Sem Cluster") +
  scale_fill_manual(values = cluster.colors, na.translate = TRUE) + theme_bw() + 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()`).
```

```
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
```

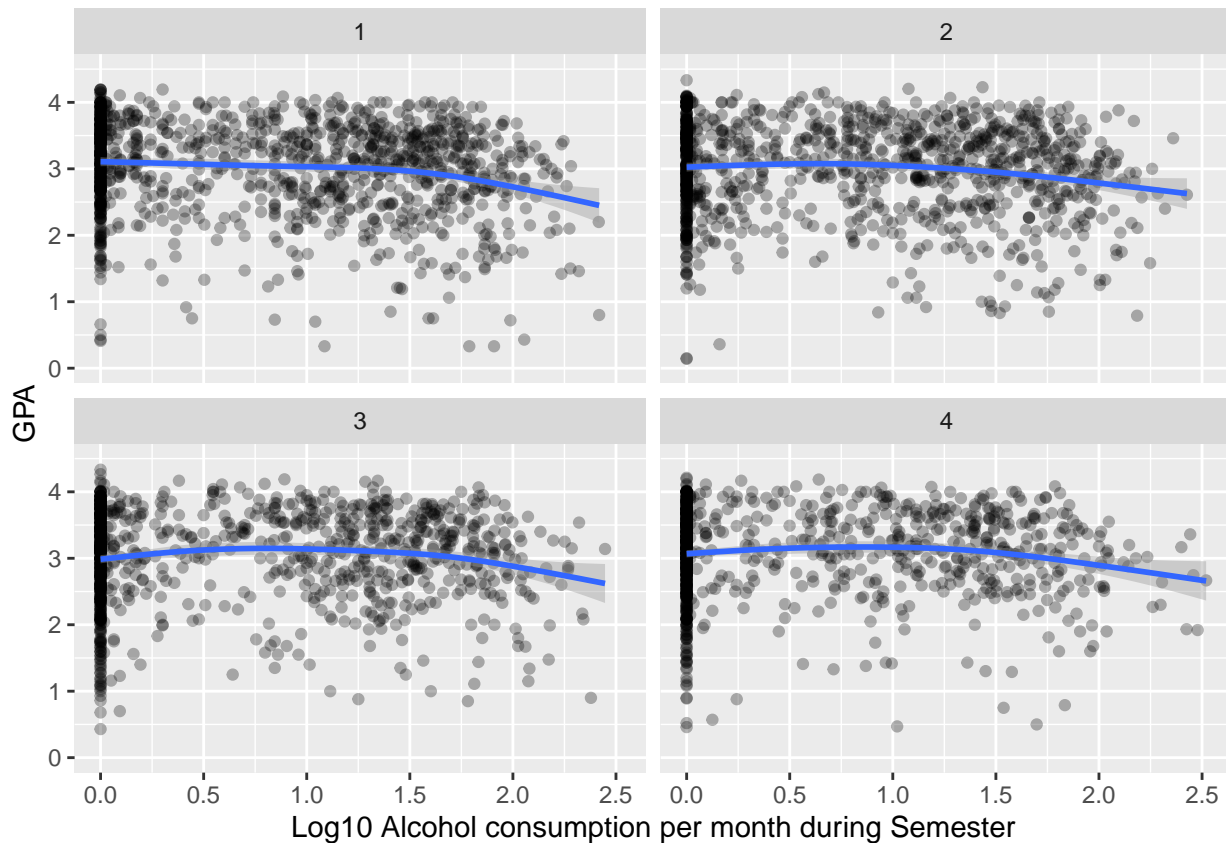
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 362 rows containing non-finite outside the scale range
```

```
## (`stat_smooth()`).
```

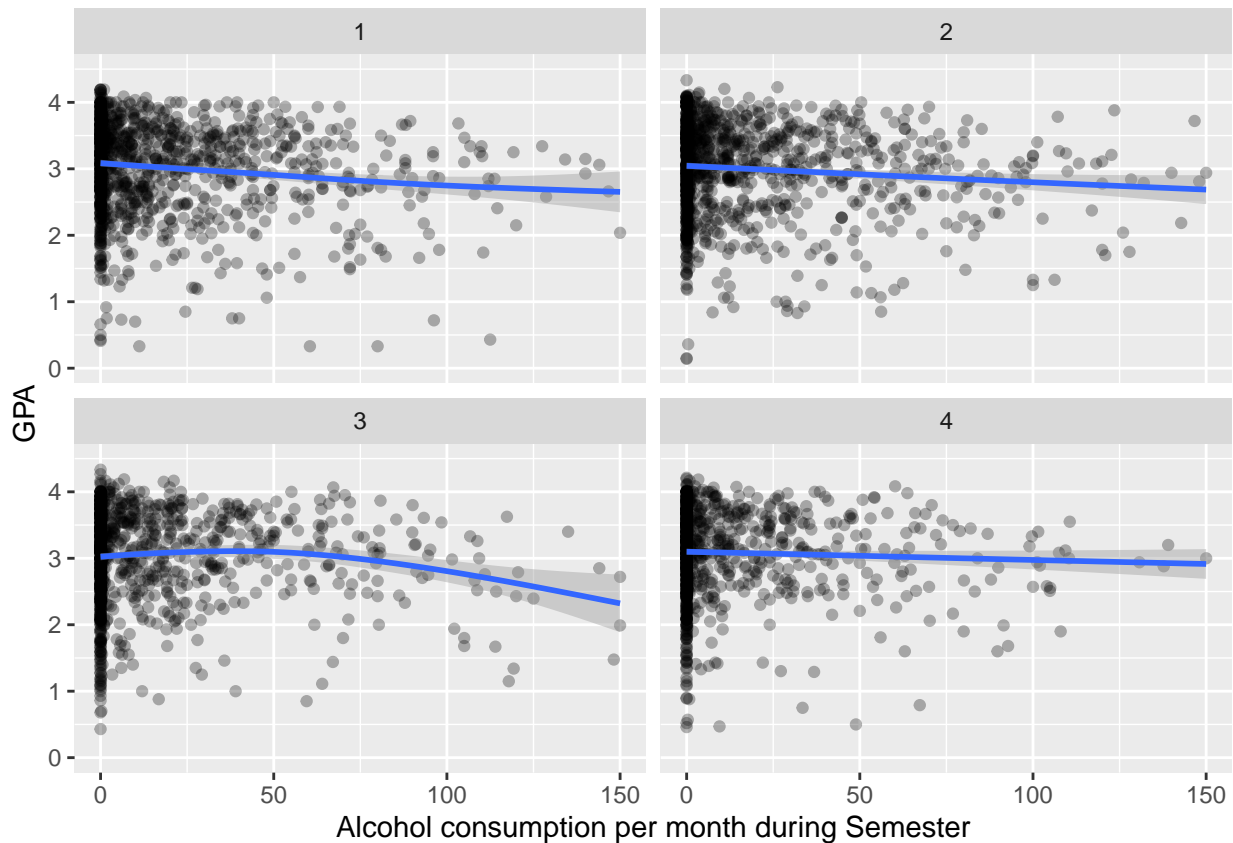
```
## Warning: Removed 362 rows containing missing values or values outside the scale range
```

```
## (`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") + xlim(0, 150)
plot.alcoholGPA

## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
## Warning: Removed 408 rows containing non-finite outside the scale range
## (`stat_smooth()`).
## Warning: Removed 408 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
```

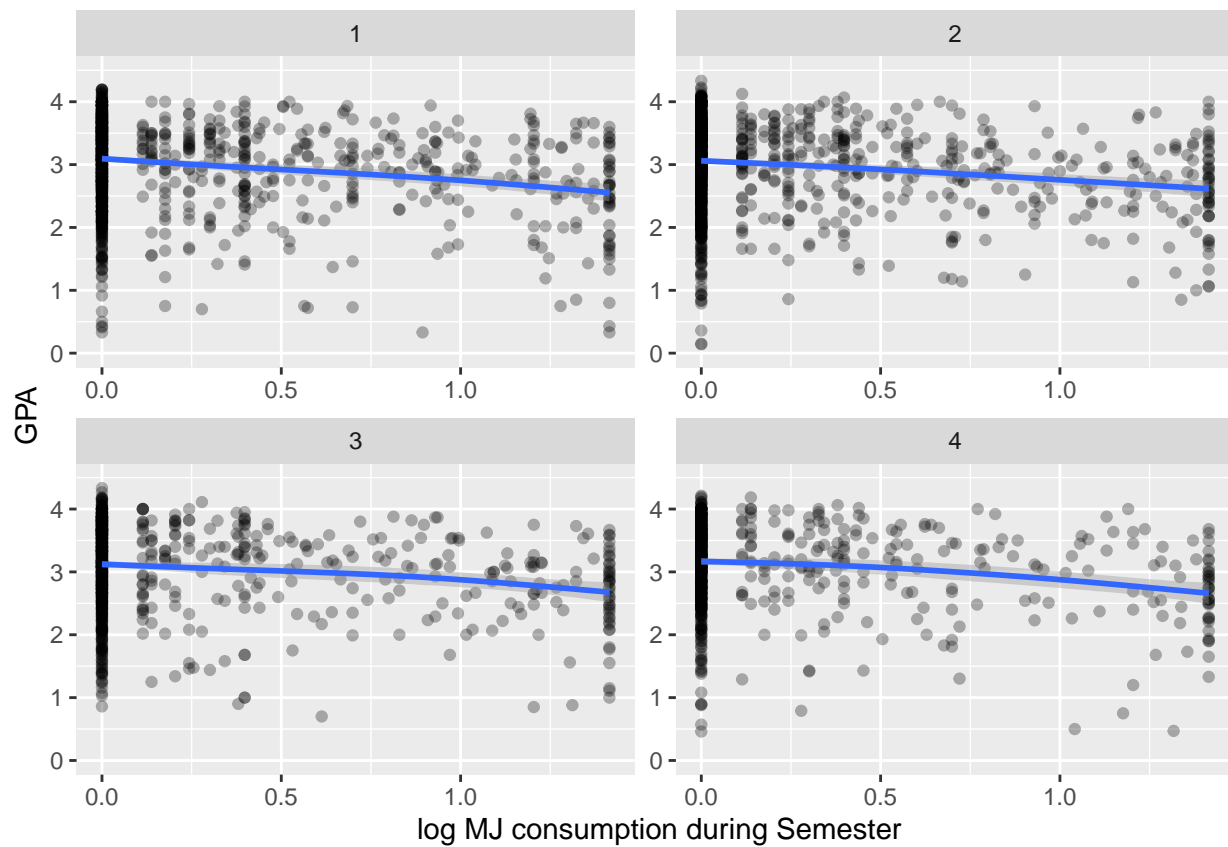
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 766 rows containing non-finite outside the scale range
```

```
## (`stat_smooth()`).
```

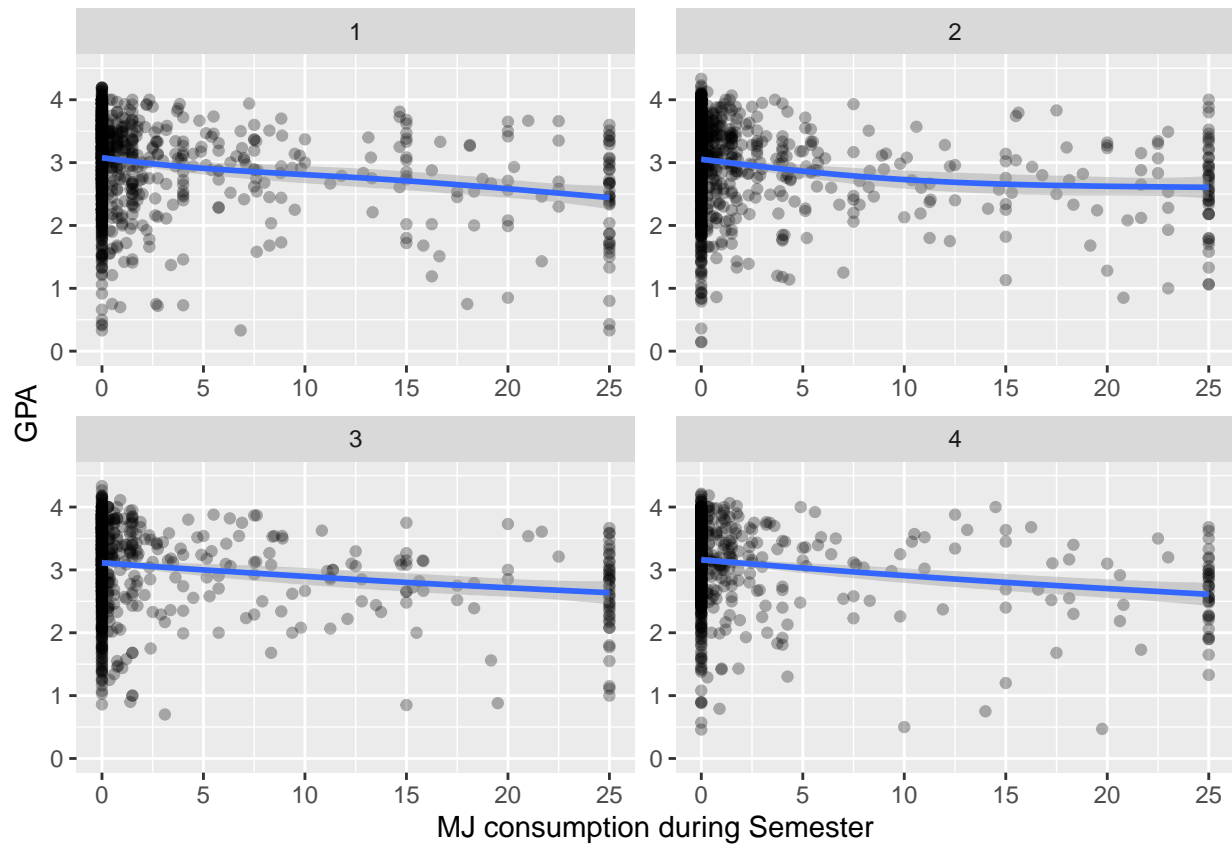
```
## Warning: Removed 766 rows containing missing values or values outside the scale range
```

```
## (`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