Code

Project

Sergio Abbate

20/7/2021

1 Data Preparation

1.1 Dataset Overview

The Dataset that was selected for this project is a Student Performance Dataset. It includes 649 observations of students of two Portuguese Schools, and it contains both numeric and categorical attributes, describing information like demographics, academic features and family/life related. The outcome variable is the final year grade 'G3' (measure of performance) and is a numeric attribute. The dataset was found in Kaggle.

1.2 Project Objective

Determine which are the most important variables that best describe a student grade. Which attributes play a bigger role in explaining/predicting a student grade? Are the academic factors, demographics, family related variables or the Life Balance variables? Which specifically? And how do they affect a student performance in general?

1.3 Data Type Conversion

Code

```
## Rows: 649 Columns: 33
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (17): school, sex, address, famsize, Pstatus, Mjob, Fjob, reason, guardia
n, ...
## dbl (16): age, Medu, Fedu, traveltime, studytime, failures, famrel, freetime,
go...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this messa
ge.
```

Code

```
##  [1] "school"     "sex"        "age"        "address"    "famsize"    "Pstatu
s"
##  [7] "Medu"       "Fedu"       "Mjob"       "Fjob"       "reason"     "guardia
n"
## [13] "traveltime" "studytime"  "failures"   "schoolsup"  "famsup"     "paid"
## [19] "activities" "nursery"    "higher"     "internet"   "romantic"   "famrel"
## [25] "freetime"   "goout"      "Dalc"       "Walc"       "health"     "absence
s"
## [31] "G1"         "G2"         "G3"
```

Code

```
##  [1] "age"        "Medu"       "Fedu"       "traveltime" "studytime"  "failure
s"
##  [7] "famrel"     "freetime"   "goout"      "Dalc"       "Walc"       "health"
## [13] "absences"   "G1"         "G2"         "G3"
```

Code

```
##  [1] "school"     "sex"        "address"    "famsize"    "Pstatus"    "Mjob"
##  [7] "Fjob"       "reason"     "guardian"   "schoolsup"  "famsup"     "paid"
## [13] "activities" "nursery"    "higher"     "internet"   "romantic"
```

Code

The following 'numeric' columns are actually categories, since each level has it's own meaning, reason why I decided it is more appropriate to treat them as categories.

Code

The following variables are numeric scales from 1 to 5. Since each number does not have a specific meaning (e.g. 2 = bad, 3 = normal, 4 = good), I decided to keep them as numeric. To some extent, it make sense to talk about averages of these variables and since every level does not have it's own meaning, I treat these variables as numeric:

- famrel, freetime, goout, Dalc, Walc, health.

Code

```
##  [1] "age"      "failures" "famrel"   "freetime" "goout"    "Dalc"     "Walc"
##  [8] "health"   "absences" "G1"       "G2"       "G3"
```

Code

```
##  [1] "school"     "sex"        "address"    "famsize"    "Pstatus"    "Medu"
##  [7] "Fedu"       "Mjob"       "Fjob"       "reason"     "guardian"   "travelt
ime"
## [13] "studytime"  "schoolsup"  "famsup"     "paid"       "activities" "nurser
y"
## [19] "higher"     "internet"   "romantic"
```

1.4 Summary of variables

Code

```
##  school   sex          age        address famsize   Pstatus Medu    Fedu
##  GP:423   F:383   Min.   :15.00   R:197   GT3:457   A: 80   0:  6   0:  7
##  MS:226   M:266   1st Qu.:16.00   U:452   LE3:192   T:569   1:143   1:174
##                   Median :17.00                             2:186   2:209
##                   Mean   :16.74                             3:139   3:131
##                   3rd Qu.:18.00                             4:175   4:128
##                   Max.   :22.00
##        Mjob           Fjob            reason      guardian   traveltime studyt
ime
##  at_home :135   at_home : 42   course    :285   father:153   1:366      1:212
##  health  : 48   health  : 23   home      :149   mother:455   2:213      2:305
##  other   :258   other   :367   other     : 72   other : 41   3: 54      3: 97
##  services:136   services:181   reputation:143                4: 16      4: 35
##  teacher : 72   teacher : 36
##
##     failures      schoolsup famsup     paid     activities nursery   higher
##  Min.   :0.0000   no :581   no :251   no :610   no :334    no :128   no : 69
##  1st Qu.:0.0000   yes: 68   yes:398   yes: 39   yes:315    yes:521   yes:580
##  Median :0.0000
##  Mean   :0.2219
##  3rd Qu.:0.0000
##  Max.   :3.0000
##  internet  romantic      famrel         freetime        goout            Dalc
##  no :151   no :410   Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.
000
##  yes:498   yes:239   1st Qu.:4.000   1st Qu.:3.00   1st Qu.:2.000   1st Qu.:1.
000
##                      Median :4.000   Median :3.00   Median :3.000   Median :1.
000
##                      Mean   :3.931   Mean   :3.18   Mean   :3.185   Mean   :1.
502
##                      3rd Qu.:5.000   3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:2.
000
##                      Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :5.
000
##       Walc           health         absences            G1              G2
##  Min.   :1.00   Min.   :1.000   Min.   : 0.000   Min.   : 0.0   Min.   : 0.00
##  1st Qu.:1.00   1st Qu.:2.000   1st Qu.: 0.000   1st Qu.:10.0   1st Qu.:10.00
##  Median :2.00   Median :4.000   Median : 2.000   Median :11.0   Median :11.00
##  Mean   :2.28   Mean   :3.536   Mean   : 3.659   Mean   :11.4   Mean   :11.57
##  3rd Qu.:3.00   3rd Qu.:5.000   3rd Qu.: 6.000   3rd Qu.:13.0   3rd Qu.:13.00
##  Max.   :5.00   Max.   :5.000   Max.   :32.000   Max.   :19.0   Max.   :19.00
##        G3
##  Min.   : 0.00
##  1st Qu.:10.00
##  Median :12.00
##  Mean   :11.91
##  3rd Qu.:14.00
##  Max.   :19.00
```

Code

Code

Attribute	Missing Values	Unique Values	Mean	Min	Max	SD
age	0	8	16.7442219	15	22	1.2181376
failures	0	4	0.2218798	0	3	0.5932351
famrel	0	5	3.9306626	1	5	0.9557169
freetime	0	5	3.1802773	1	5	1.0510926
goout	0	5	3.1848998	1	5	1.1757661
Dalc	0	5	1.5023112	1	5	0.9248344
Walc	0	5	2.2804314	1	5	1.2843800
health	0	5	3.5362096	1	5	1.4462591
absences	0	24	3.6594761	0	32	4.6407588
G1	0	17	11.3990755	0	19	2.7452651
G2	0	16	11.5701079	0	19	2.9136387
G3	0	17	11.9060092	0	19	3.2306562

Code

Attribute	Missing Values	Unique Values
school	0	2
sex	0	2
address	0	2
famsize	0	2
Pstatus	0	2
Medu	0	5
Fedu	0	5
Mjob	0	5
Fjob	0	5
reason	0	4
guardian	0	3
traveltime	0	4
studytime	0	4
schoolsup	0	2
famsup	0	2

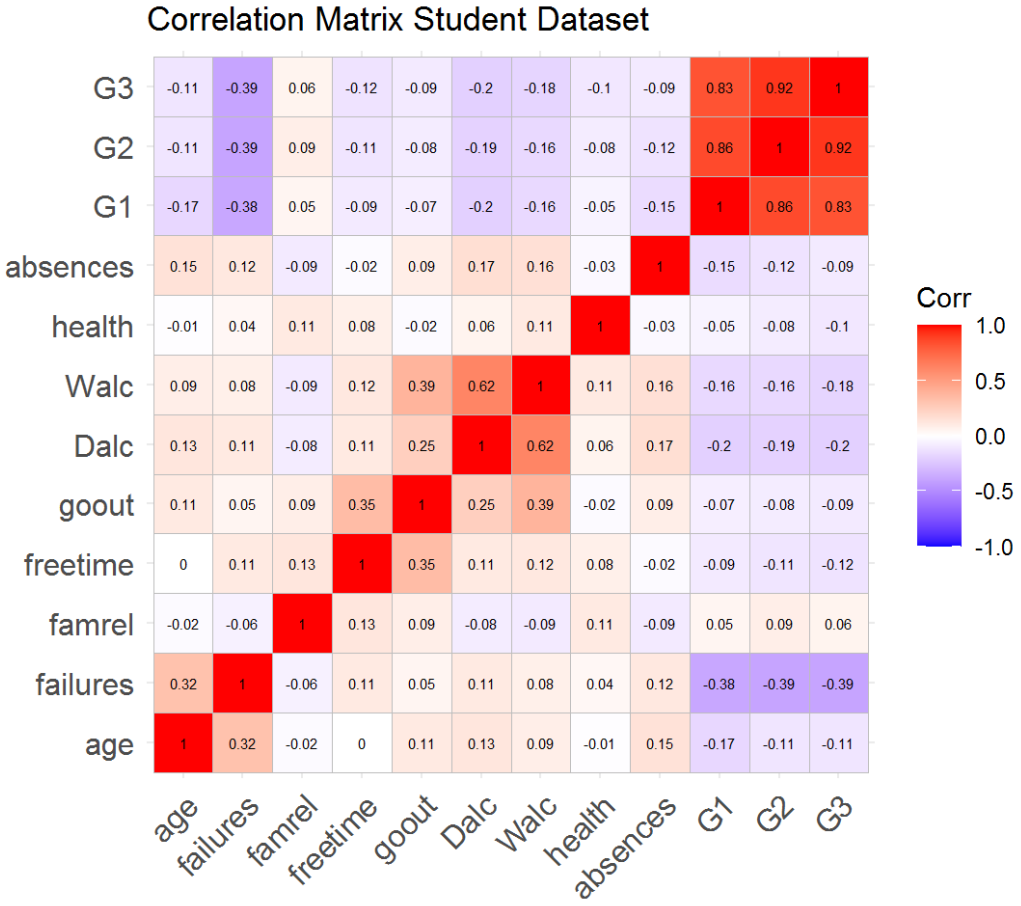
Attribute	Missing Values	Unique Values
paid	0	2
activities	0	2
nursery	0	2
higher	0	2
internet	0	2
romantic	0	2

Observations: Most students live in urban areas, with a family size greater than 3 members, Mother's and Father's education is almost evenly distributed (not too many high education levels). Most of the student's have a relatively low traveltime, however their reason of attendance to school is not necessarily 'close to home'. Most of them have no extra school support or paid classes. Most of them wish to take higher education. Most of them have a good family relationship, normal freetime and go-out time. Low alcohol consumption in students, high health status and low absences as well.

Observations: Overall, the dataset is pretty clean, there aren't any unusual or illogical values in any of the columns, each column's meaning is clear and understandable and there aren't any NA values in the dataset.

1.5 Un-used and Correlated variables

Code



Code

```
## [1] 0.6474766
```

Observations:

- High Correlations between G1, G2 and G3 as expected since G3 is somehow composed by G2 and G1. Since G3 is the real output variable and G1 and G2 are just the intermediate grades and not the final grade, I droped G1 and G2.
- Relatively high correlation between Dalc and Walc (Daily vs Weekend Alcohol Consumption), so I decided to keep only one of them (Walc).
- Another important and relatively high correlation found is between Mother's education and Father's education. To some extent is this relationship reasonable but I believe that these 2 variables do not neccesarily tell the same information so I kept both.

Code

school	avg_grade
GP	12.57683
MS	10.65044

Code

```
##
##  Welch Two Sample t-test
##
## data:  G3 by school
## t = 6.7545, df = 340.49, p-value = 6.212e-11
## alternative hypothesis: true difference in means between group GP and group MS
is not equal to 0
## 95 percent confidence interval:
##  1.365411 2.487368
## sample estimates:
## mean in group GP mean in group MS
##         12.57683         10.65044
```

Code

The dataset contains students from 2 schools. There is a statistically significant difference in the grades of both school. At this point I can consider 2 different subpopulations when performing several Data Mining Techniques to see if both schools behave differently. For the time being, I will ignore the school when performing the EDA phase and I will get back to it further on in the Project.

2 Logical Groupings

For ease of understanding I decided to group the variables that are closely related in 4 different groups:

- Demographics: sex, age, address.

- Family: famsize, Pstatus, Medu, Fedu, Mjob, Fjob, guardian, famsup, famrel
- Academic: reason, studytime, failures, schoolsup, paid, nursery, higher, absences
- Life Balance: traveltime, activities, internet, romantic, freetime, goout, Walc, health

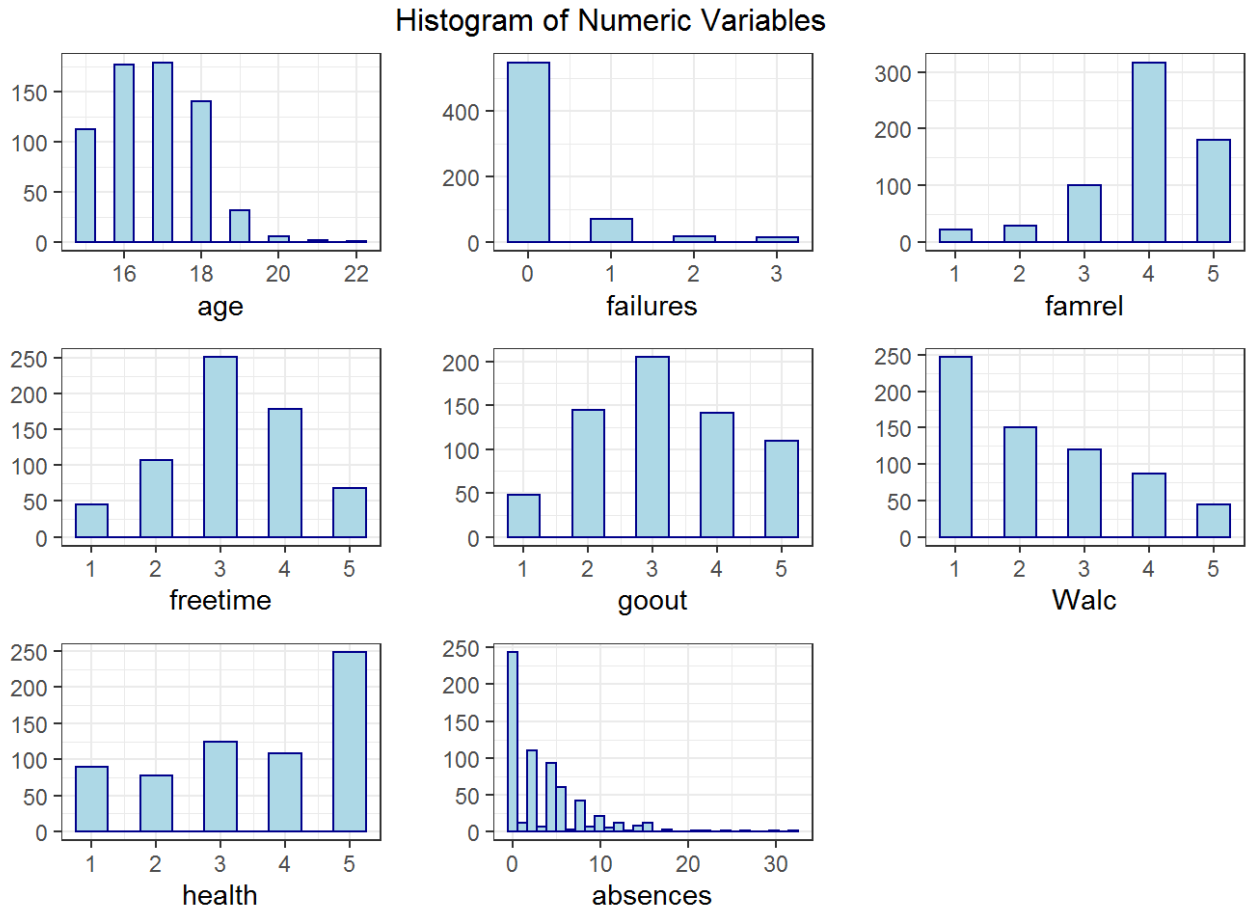
Code

3 Univariate Analysis

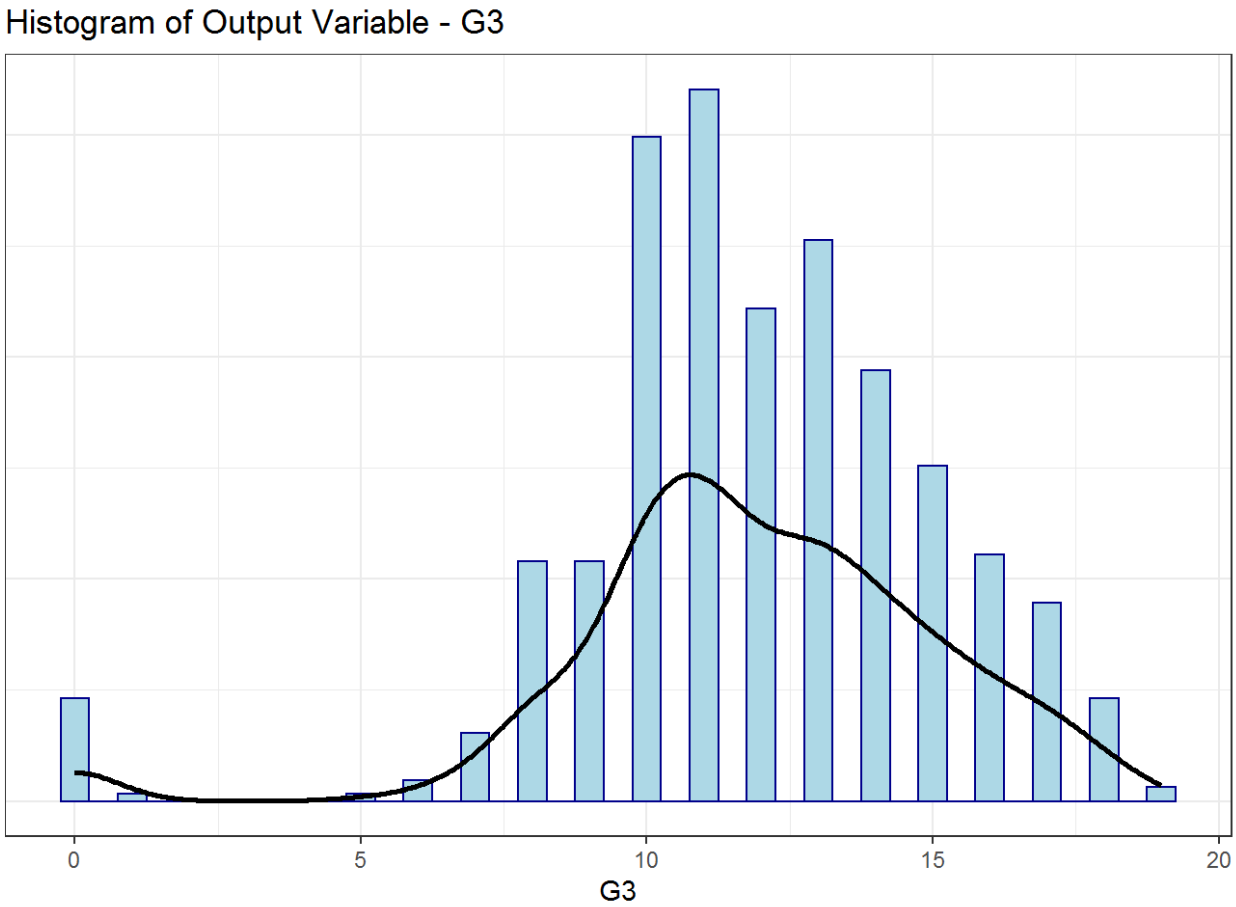
3.1 Numeric Attributes

Code

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



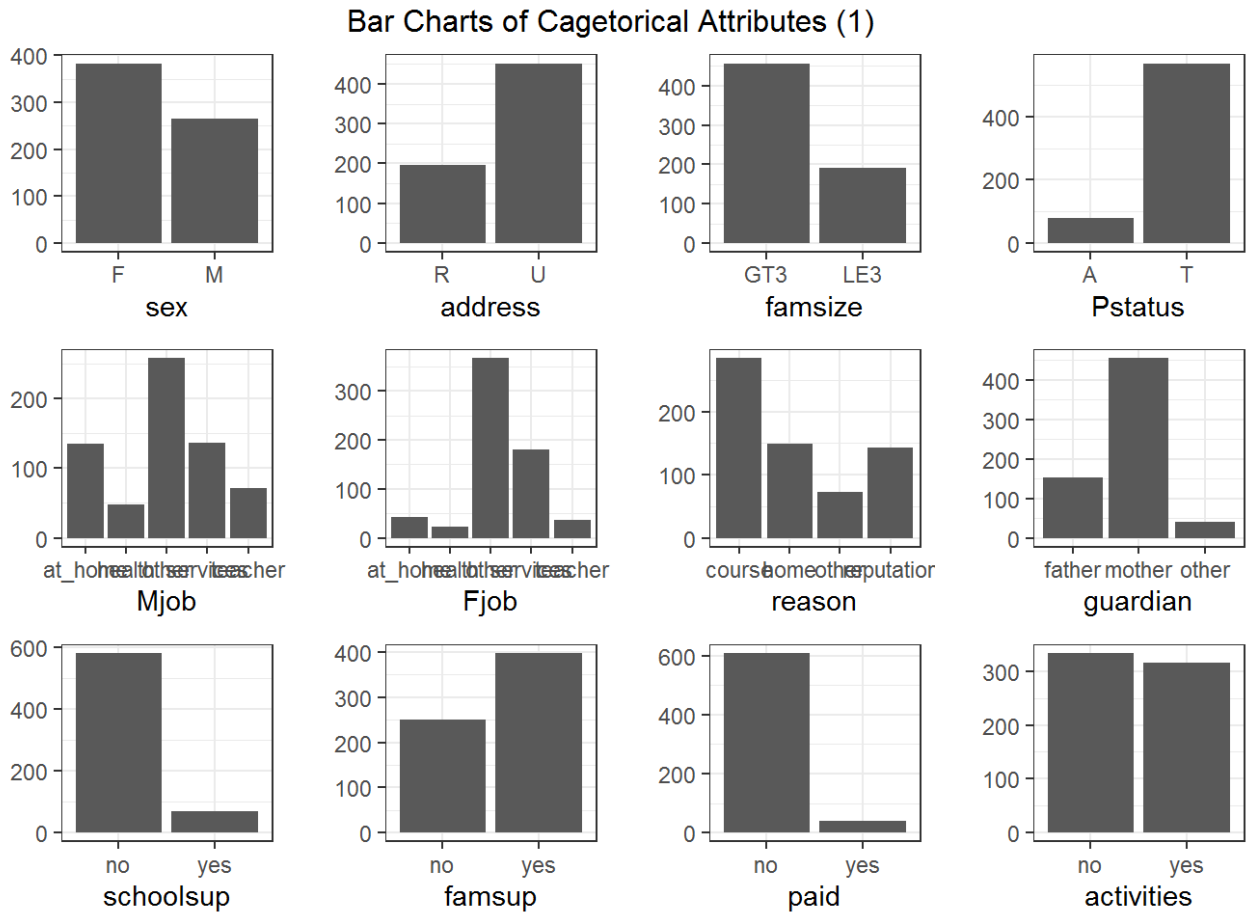
Code



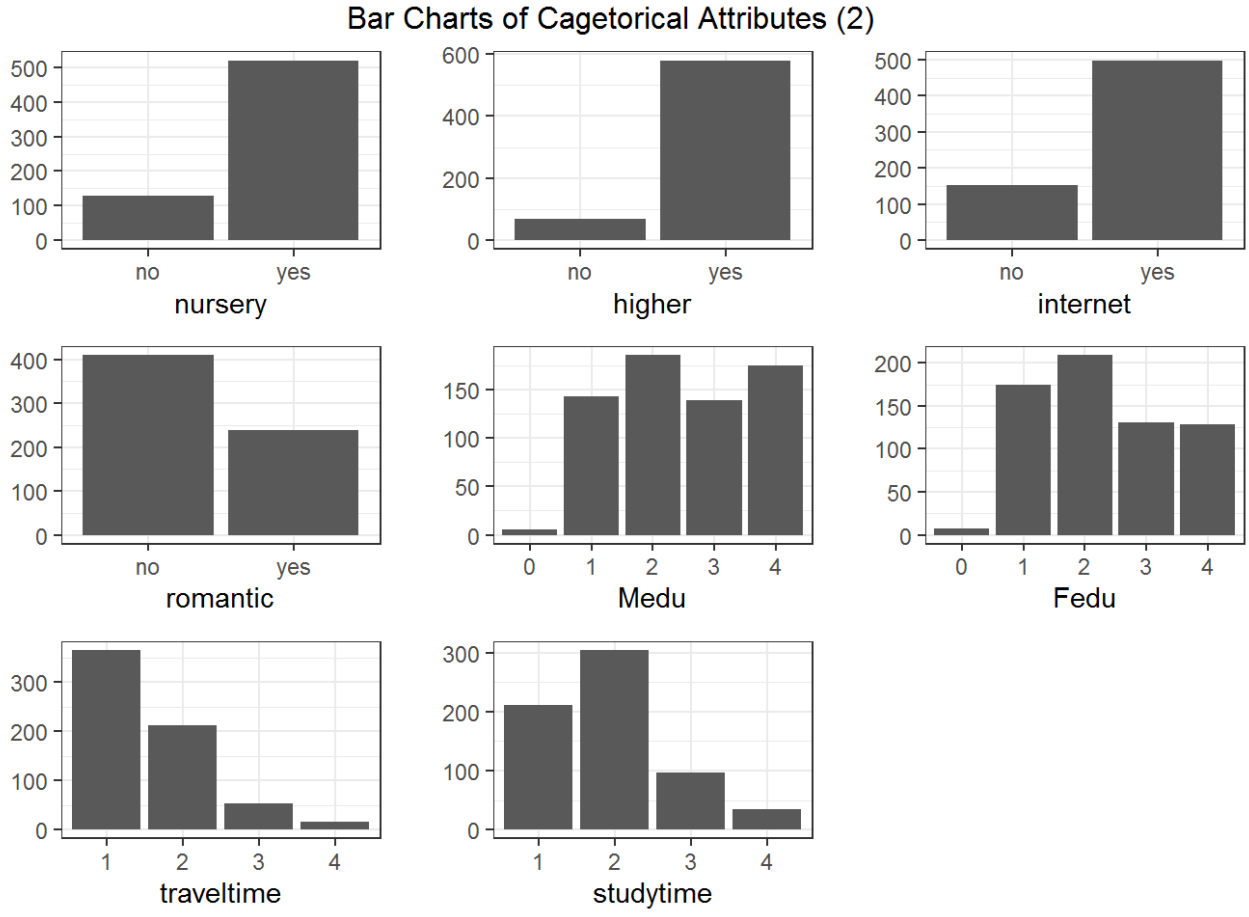
Observations: Most students between an age of 15 and 18, very few older. Low number of failures in general, pretty good family relationship, normal freetime and go-out time, relatively low alcohol consumption and good health. Not too many absences but a fat tail starting at 10 absences. Grades are aparently normally distributed with a peak around 10-12 and some very low grades (0).

3.2 Categorical Attributes

Code



Code



Observations: Most students female, urban, with a larger family size. Parents are mostly together, work as 'other'. Student's reason of attendance is mostly course preference, most of them with no extra school educational support but family educational support, mostly not paid.

50 % students have extra curricular activities and 50 % don't, most of them attended nursery school, want to pursue higher education, have internet access and no romantic relationship. Mother's and Father's education are evenly distributed, some high, some low, traveltime is in general low and so is studytime.

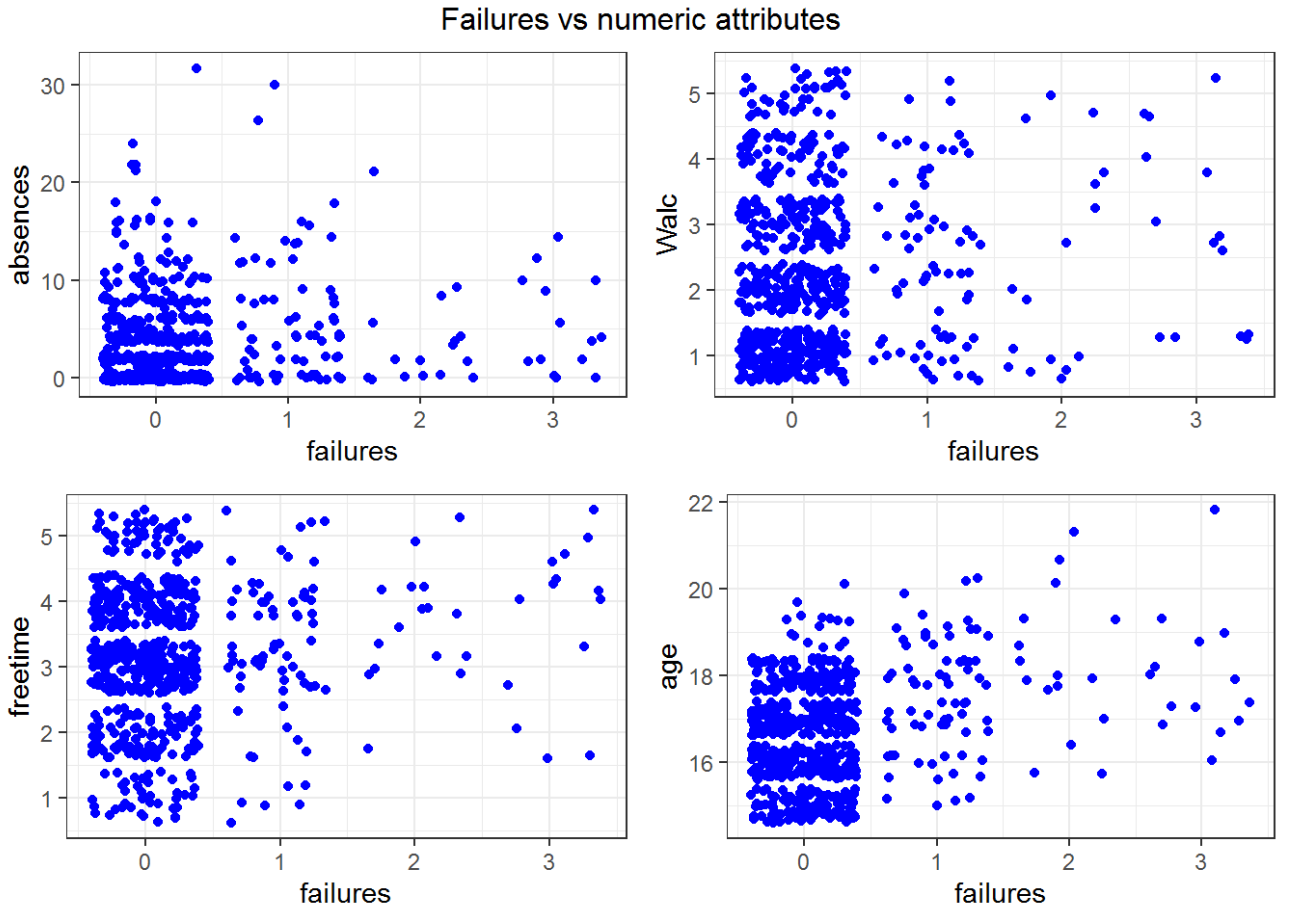
4 Bivariate Analysis

Since we have over 30 columns, I will investigate some (not all) of the most important bivariate relations. However, further in this Project I will investigate deeper the relationship between the output variable G3 and the other variables.

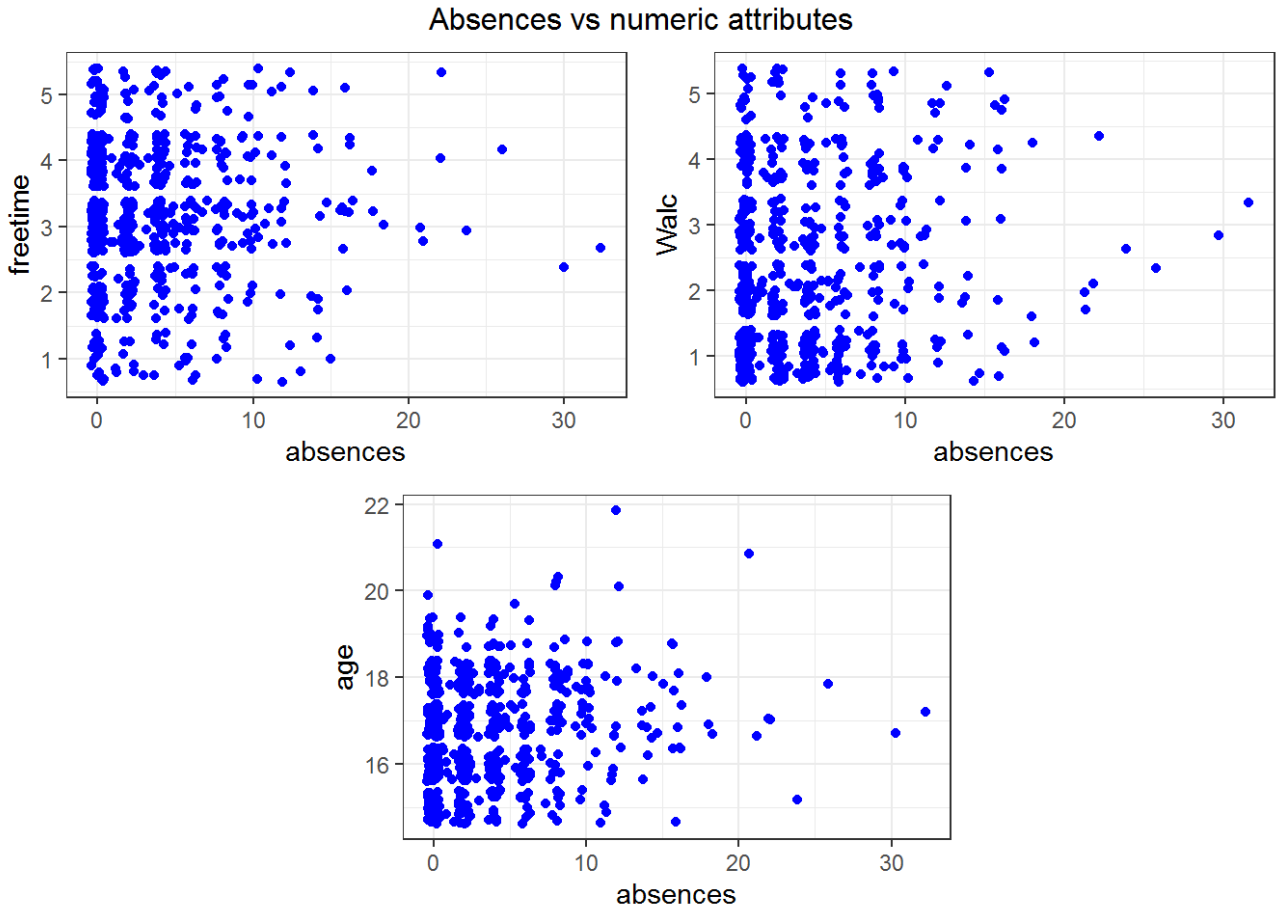
I subselected a set of 10 variables to investigate their interaction. I considered the academic variables to be highly important. I investigated variables like failures, absences, studytime, higher and schoolsup. In Life Balance I considered important freetime and Walc. One of the most important family related variables is Fedu and Medu (just picked Fedu) and regarding demographics I chose age and sex.

4.1 Measure to Measure

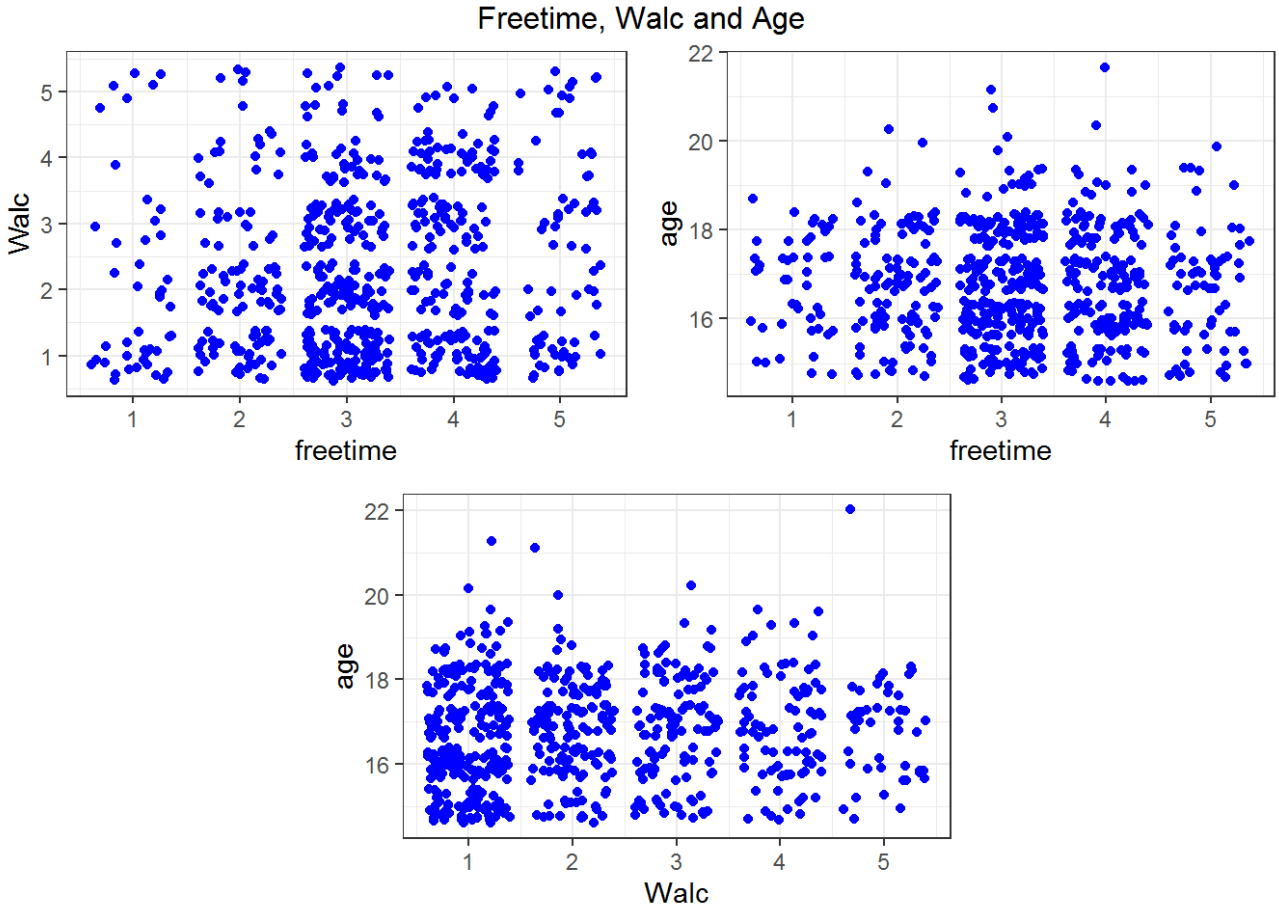
Code



Code



Code



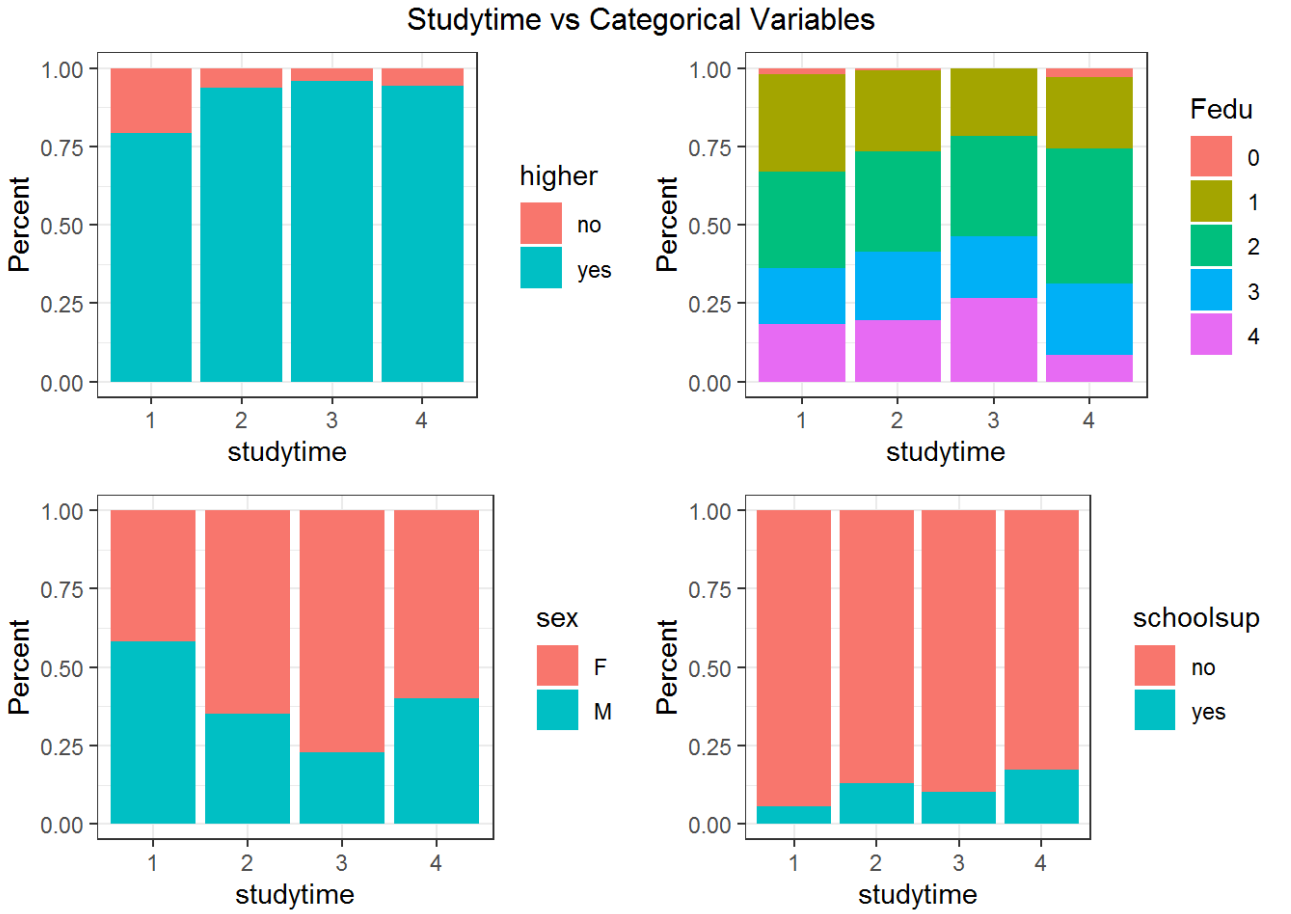
Observations: Not too many clear relationships and insights are drawn from these graphs, but some of them are:

- Most of students with failures > 1 also have freetime > 3.

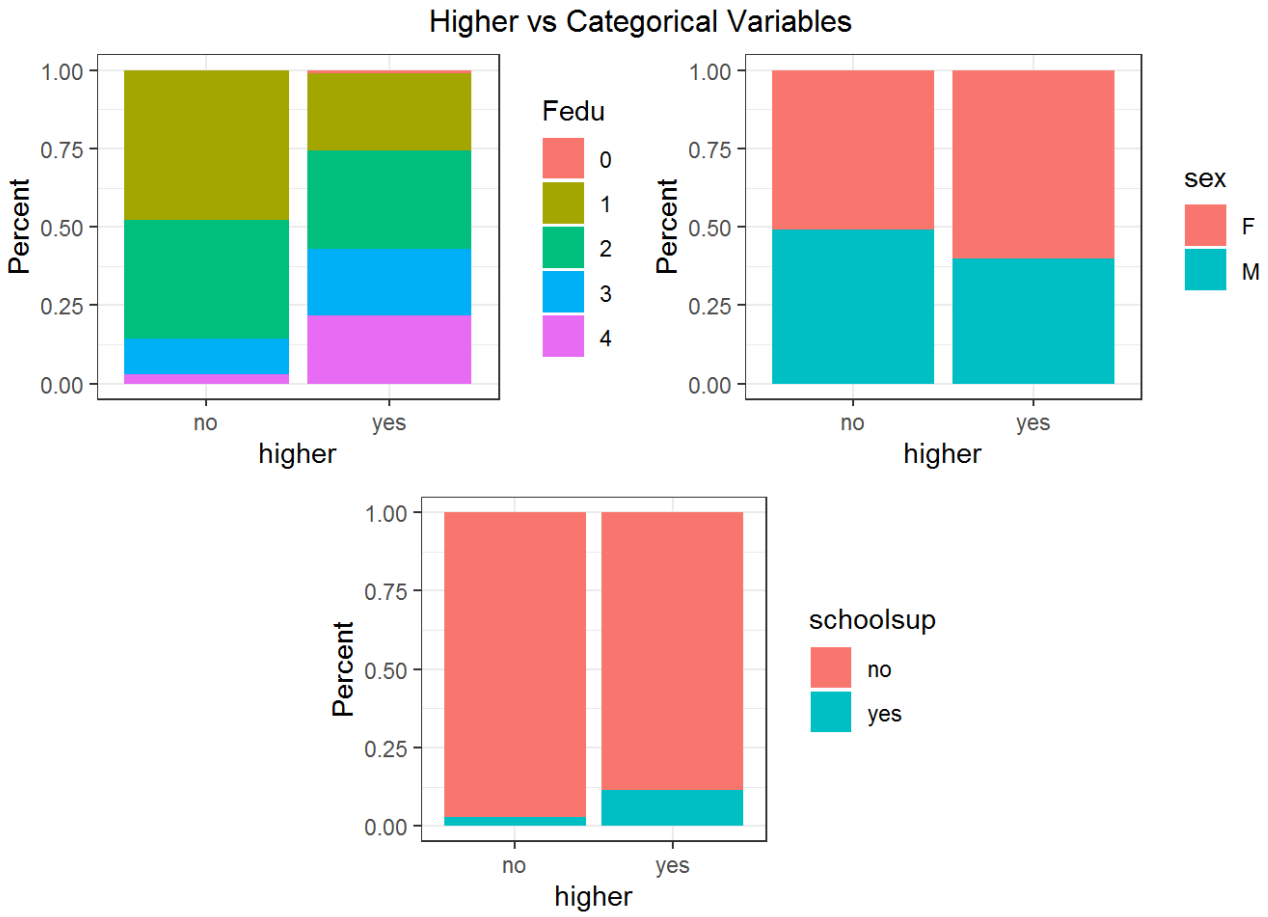
- Students with failures are in general older (reasonable).
- Students with higher absences are in general younger.

4.2 Category vs Category

Code



Code



Code

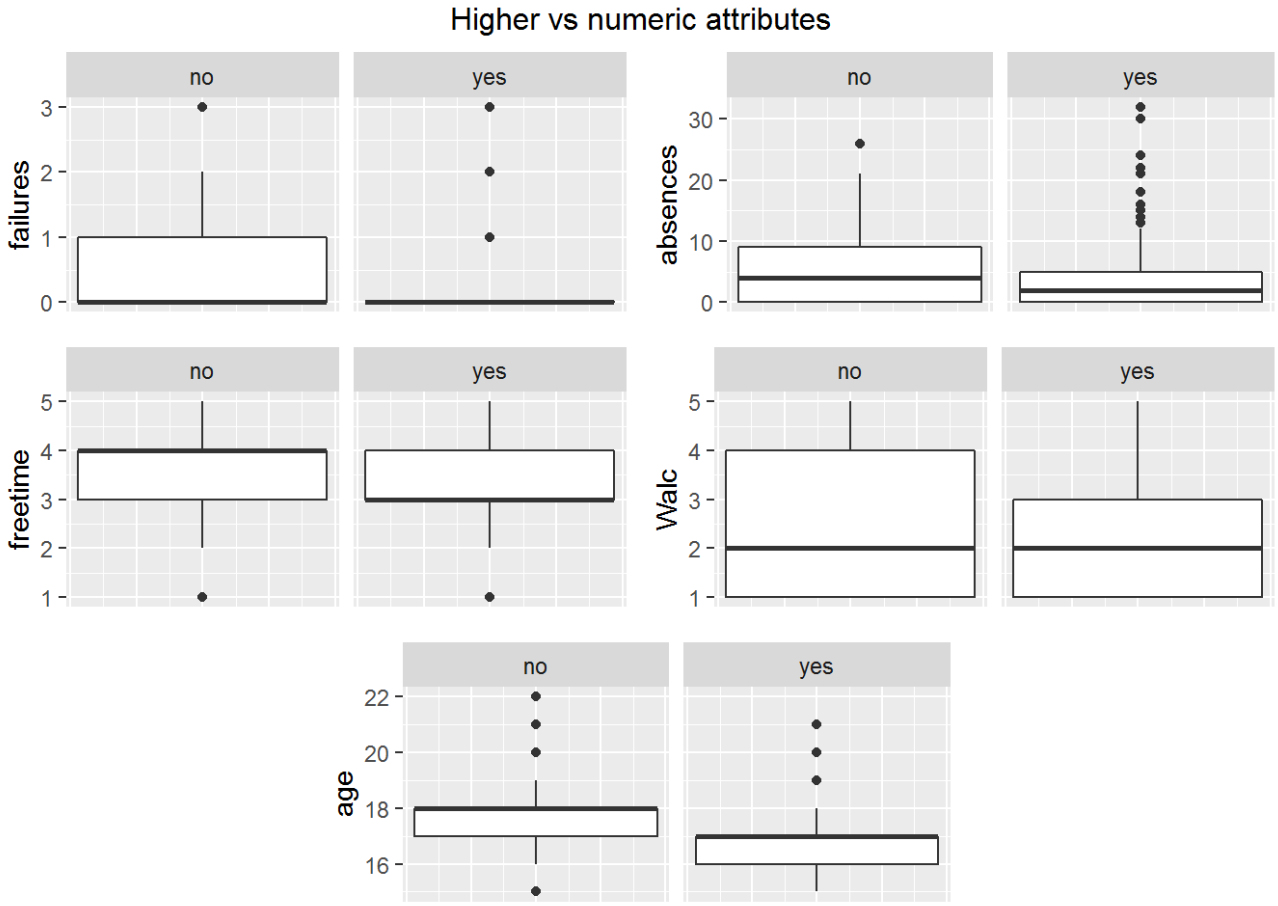
Observations: Some of the relationships and insights that are drawn from these graphs are:

- Students that don't wish to get higher education also have less studytime (reasonable).

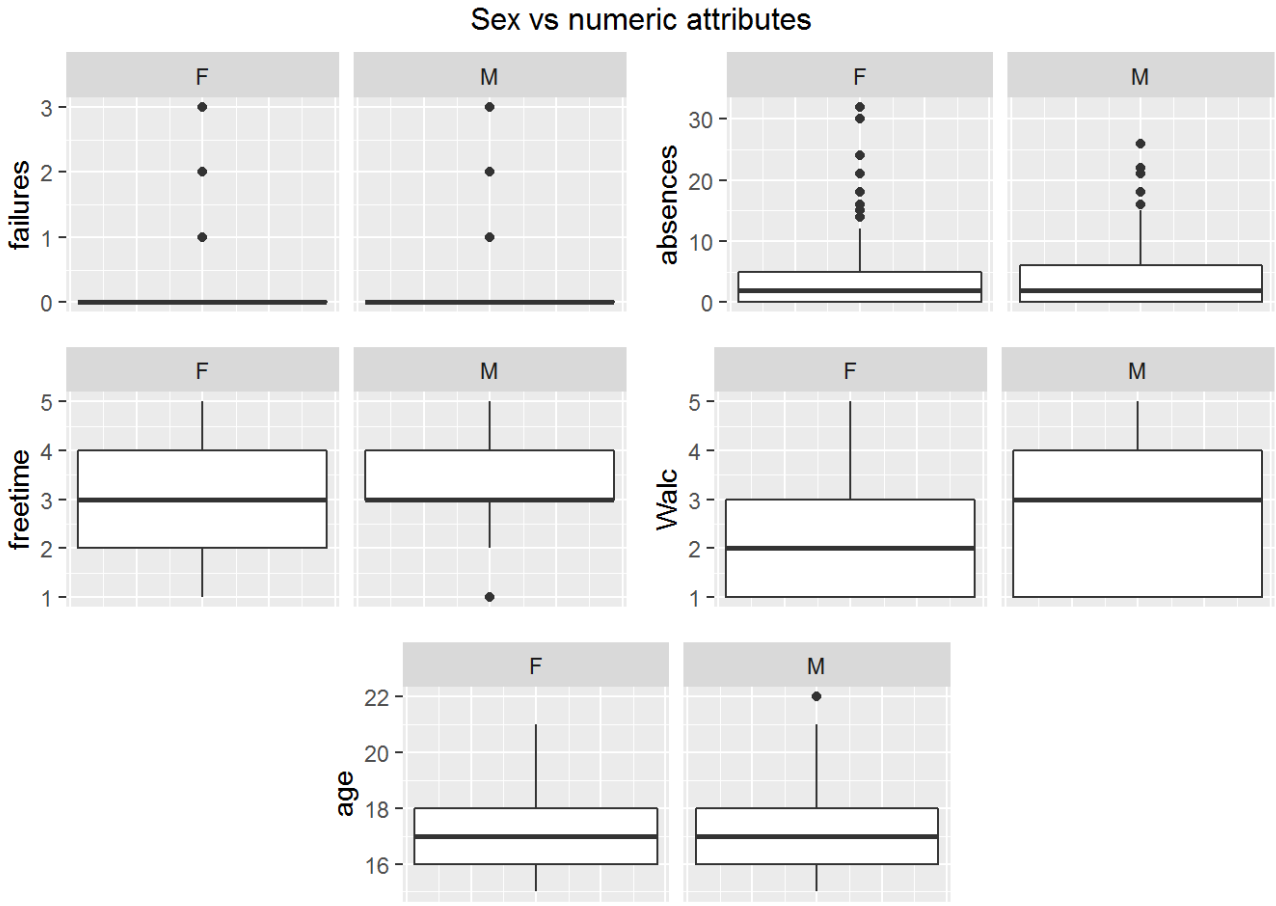
- Students with higher studytime are mostly females.
- Students with Fathers with higher levels of education are also the ones that wish to get higher education.
- Student's who's fathers had low levels of education also need extra school educational support.

4.3 Measure vs Category

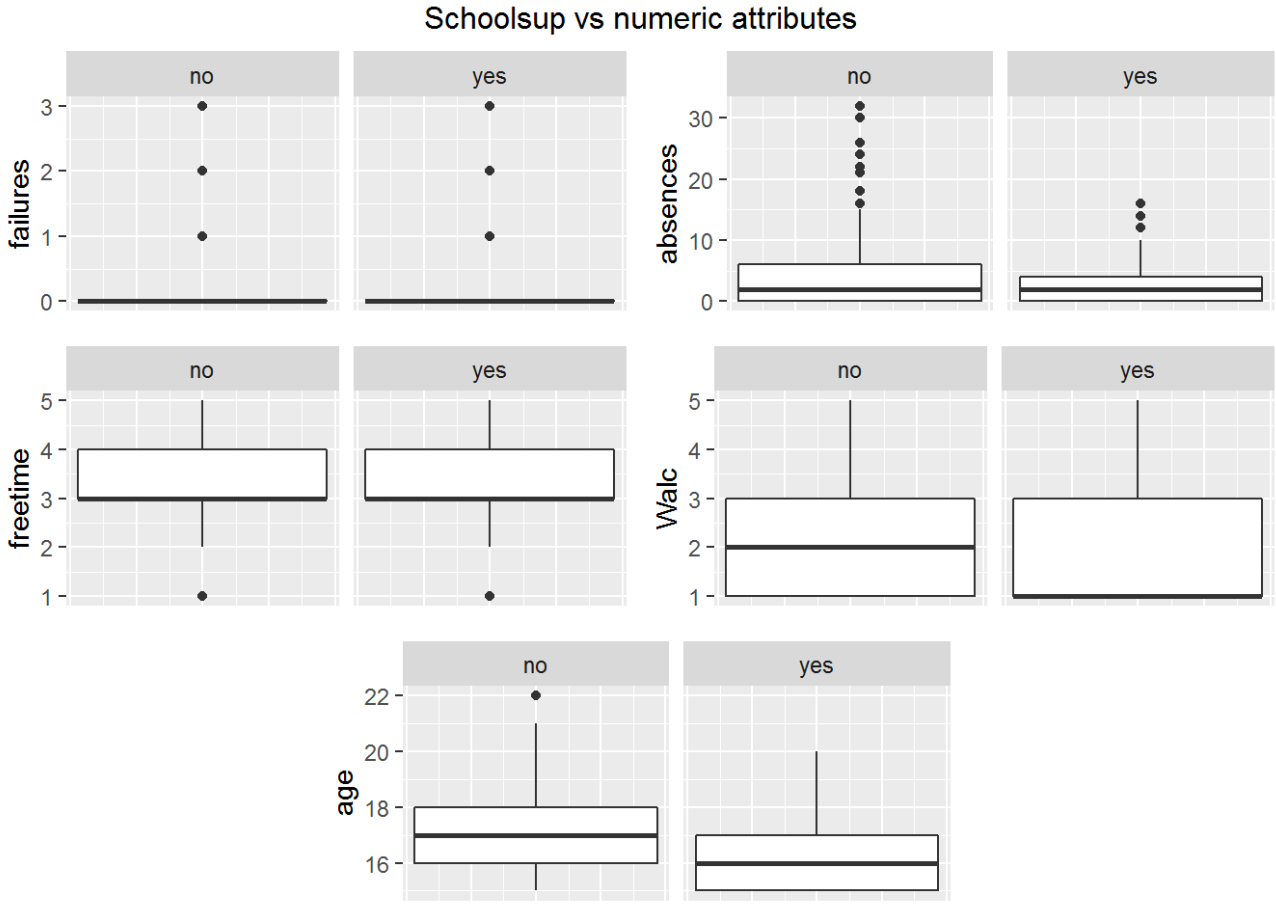
Code



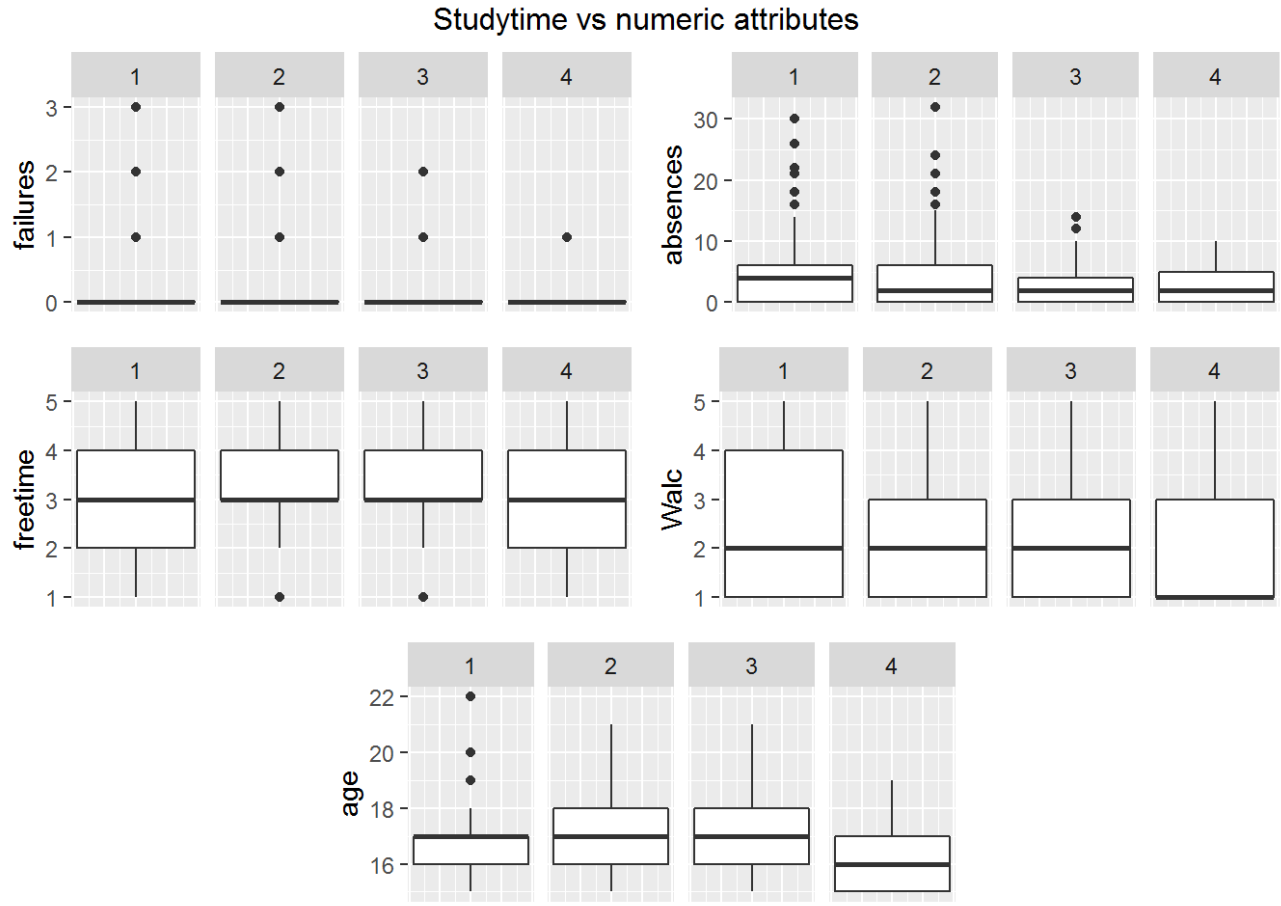
Code


Sex vs numeric attributes
F
M
failures
absences
freetime
Walc
age

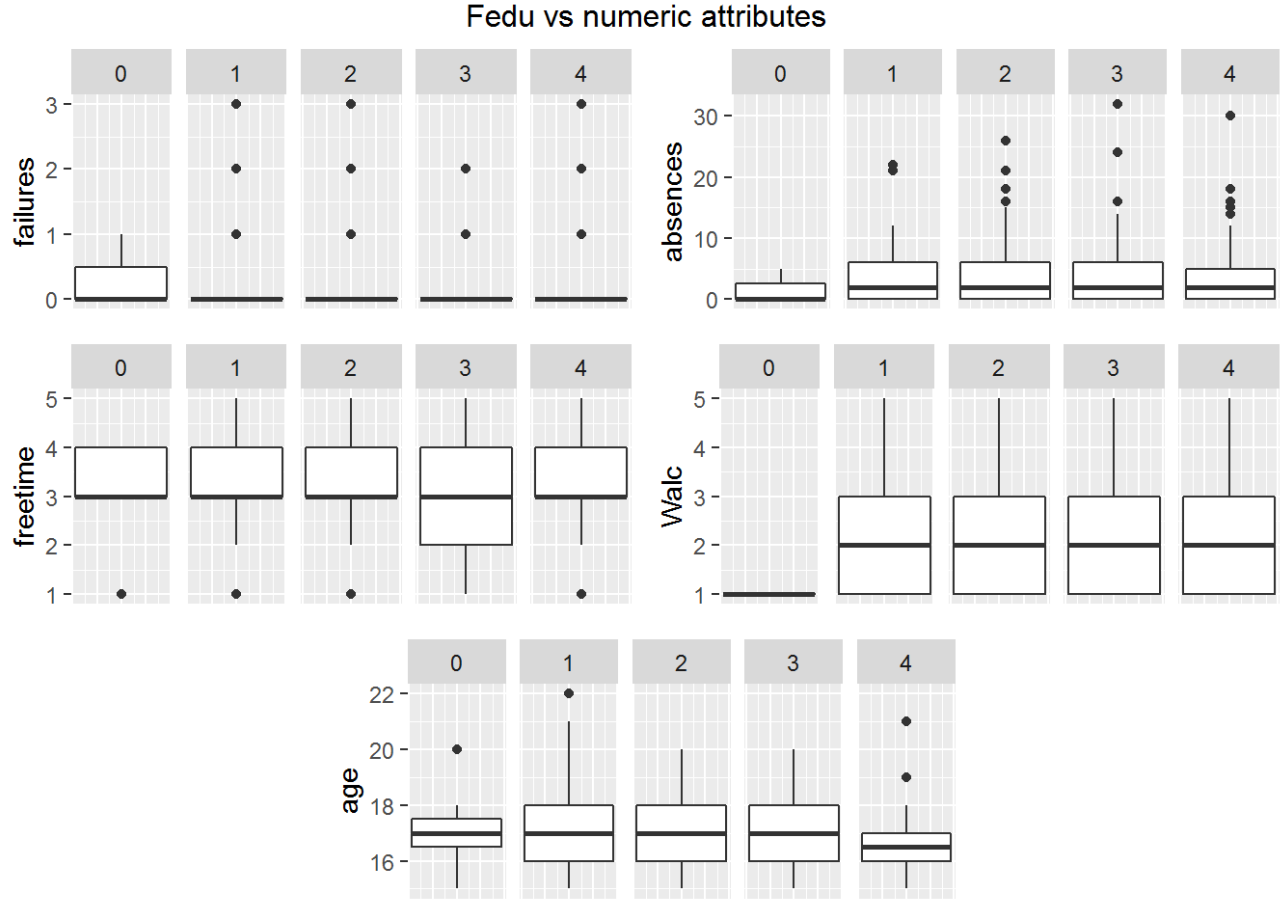
Code


Schoolsup vs numeric attributes
no
yes
failures
absences
freetime
Walc
age

Code



Code



Observations: Some of the relationships and insights that are drawn from these graphs are:

- Most students with absences do not want to take higher education.

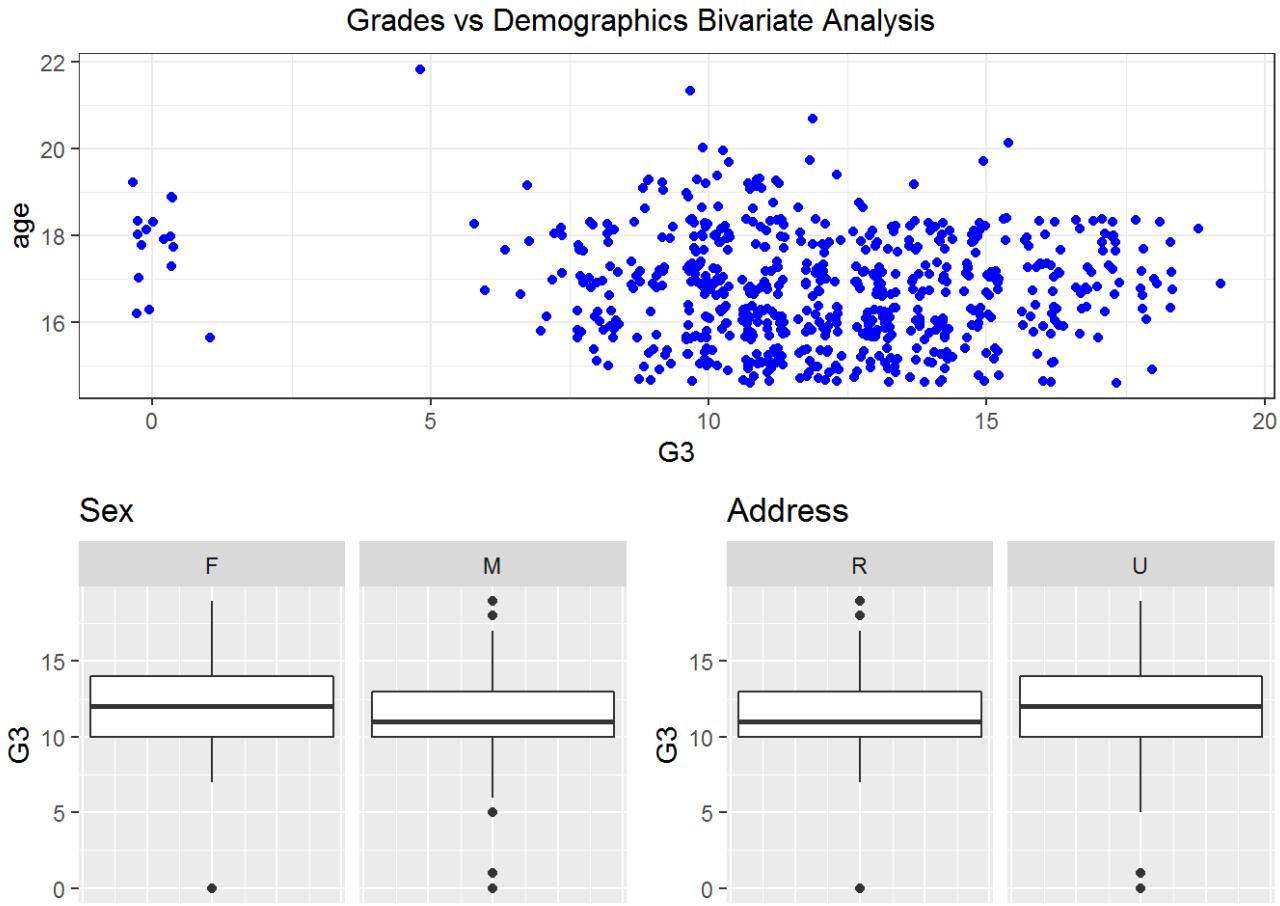
- Higher levels of alcohol consumption seems to be related to not wanting to take higher education, to males more than females, to no extra educational support and to lower studytimes.
- Students with higher studytimes are in general younger.

5 Bivariate (Output-variables) Analysis

To make it more organized I divided this section of the bivariate analysis (output vs other variables) by the 4 logical groupings: Demographics, Family, Academic, Life Balance

5.1 Demographics

Code



Observations:

- Best grades are not found in older students

5.1.1 Significance Testing

Code

```
##
##  Welch Two Sample t-test
##
## data:  G3 by sex
## t = 3.2747, df = 547.44, p-value = 0.001125
## alternative hypothesis: true difference in means between group F and group M i
s not equal to 0
## 95 percent confidence interval:
##  0.3390334 1.3554639
## sample estimates:
## mean in group F mean in group M
##        12.25326        11.40602
```

Code

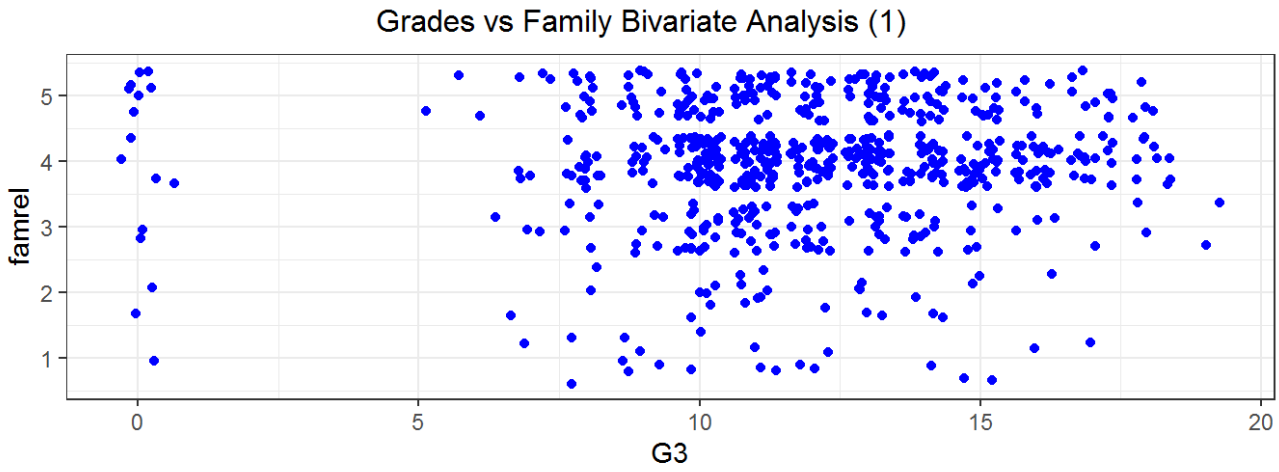
```
##
##  Welch Two Sample t-test
##
## data:  G3 by address
## t = -4.0199, df = 318.49, p-value = 7.274e-05
## alternative hypothesis: true difference in means between group R and group U i
s not equal to 0
## 95 percent confidence interval:
##  -1.7530260 -0.6009338
## sample estimates:
## mean in group R mean in group U
##        11.08629        12.26327
```

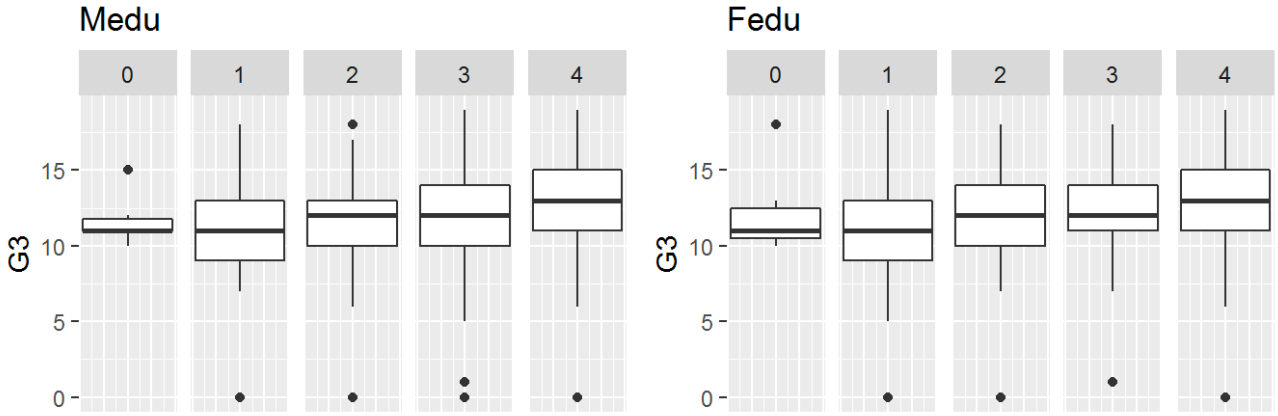
Observations:

- Statistically significant differences in grades across sex and address. Difference in mean is not really high though.

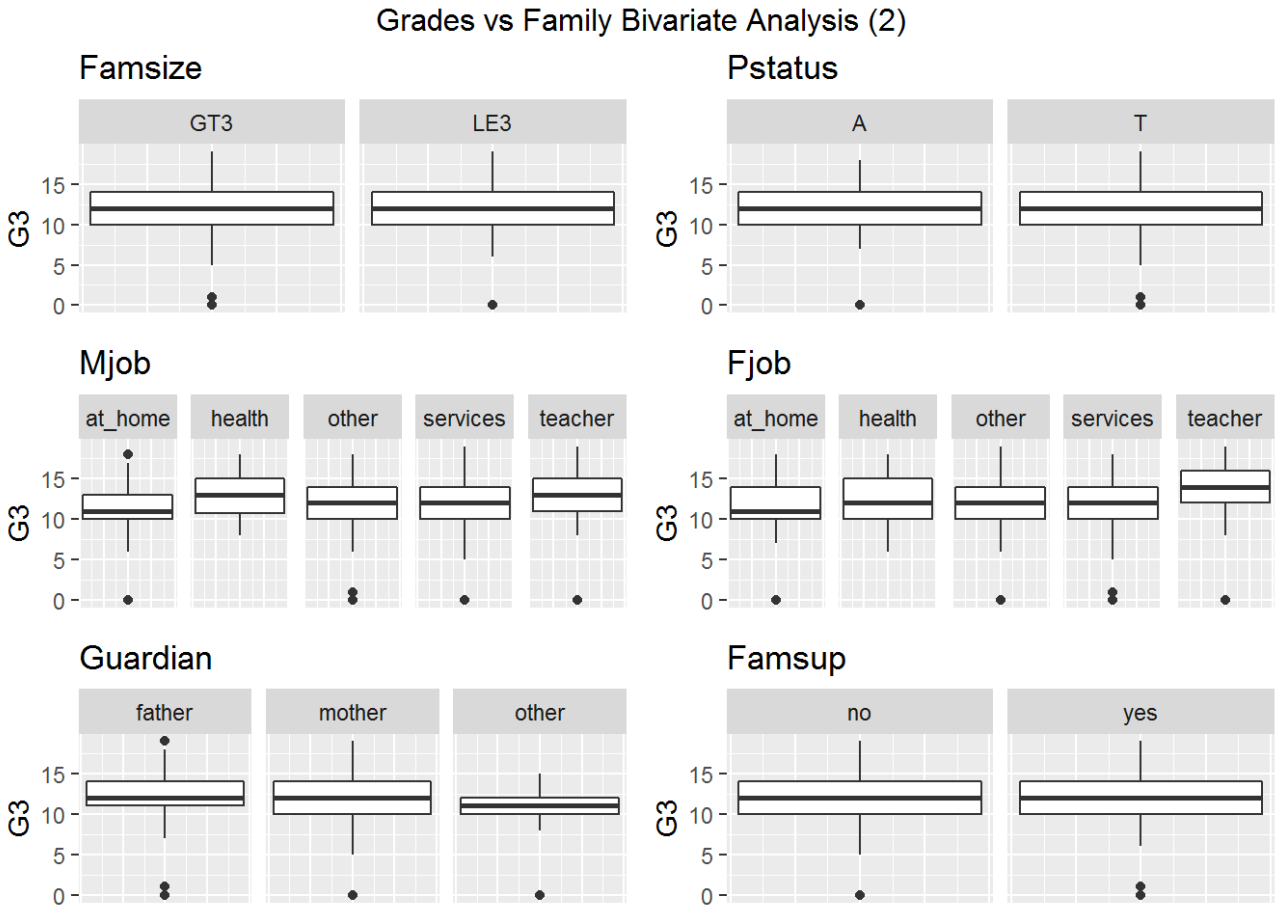
5.2 Family

Code





Code



Observations:

- In general, grades increase slightly as Mother and Father's Education increase.

- Grades increase slightly when Mothers and/or Fathers are teachers

5.2.1 Significance Testing

Code

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Medu          4    424  105.96   10.76 1.89e-08 ***
## Residuals   644   6339    9.84
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Code

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = G3 ~ Medu, data = student)
##
## $Medu
##             diff         lwr      upr     p adj
## 1-0 -0.869463869 -4.44607335 2.707146 0.9637645
## 2-0 -0.005376344 -3.56529890 3.554546 1.0000000
## 3-0  0.254196643 -3.32448453 3.832878 0.9996817
## 4-0  1.401904762 -2.16151243 4.965322 0.8187853
## 2-1  0.864087525 -0.09045546 1.818631 0.0973063
## 3-1  1.123660512  0.10137744 2.145944 0.0229212
## 4-1  2.271368631  1.30387403 3.238863 0.0000000
## 3-2  0.259572987 -0.70270343 1.221849 0.9475471
## 4-2  1.407281106  0.50342325 2.311139 0.0002281
## 4-3  1.147708119  0.17258280 2.122833 0.0117372
```

Code

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## Fedu          4    329   82.16   8.223 1.8e-06 ***
## Residuals   644   6435    9.99
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Code

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = G3 ~ Fedu, data = student)
##
## $Fedu
##           diff         lwr      upr     p adj
## 1-0 -1.2060755 -4.53937199 2.127221 0.8599528
## 2-0 -0.3581681 -3.68065300 2.964317 0.9983499
## 3-0  0.2388222 -3.11556480 3.593209 0.9996788
## 4-0  0.7790179 -2.57736254 4.135398 0.9693494
## 2-1  0.8479074 -0.03947319 1.735288 0.0690240
## 3-1  1.4448978  0.44467223 2.445123 0.0008181
## 4-1  1.9850934  0.97820309 2.991984 0.0000010
## 3-2  0.5969904 -0.36659224 1.560573 0.4378555
## 4-2  1.1371860  0.16668694 2.107685 0.0122963
## 4-3  0.5401956 -0.53445563 1.614847 0.6439212
```

Code

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Mjob          4    296   74.01    7.37 8.31e-06 ***
## Residuals   644   6467   10.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Code

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = G3 ~ Mjob, data = student)
##
## $Mjob
##                          diff         lwr         upr     p adj
## health-at_home     2.01805556  0.56127595  3.47483516 0.0015465
## other-at_home      0.62609819 -0.29472309  1.54691947 0.3402133
## services-at_home   1.10261438  0.04943071  2.15579805 0.0349271
## teacher-at_home    2.09444444  0.82939398  3.35949491 0.0000691
## other-health      -1.39195736 -2.75461280 -0.02930193 0.0425209
## services-health   -0.91544118 -2.37081530  0.53993295 0.4218975
## teacher-health     0.07638889 -1.53893482  1.69171260 0.9999369
## services-other     0.47651619 -0.44207995  1.39511232 0.6155742
## teacher-other      1.46834625  0.31293477  2.62375774 0.0049086
## teacher-services   0.99183007 -0.27160165  2.25526178 0.2014363
```

Code

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Fjob          4    135   33.68   3.273 0.0114 *
## Residuals   644   6629   10.29
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Code

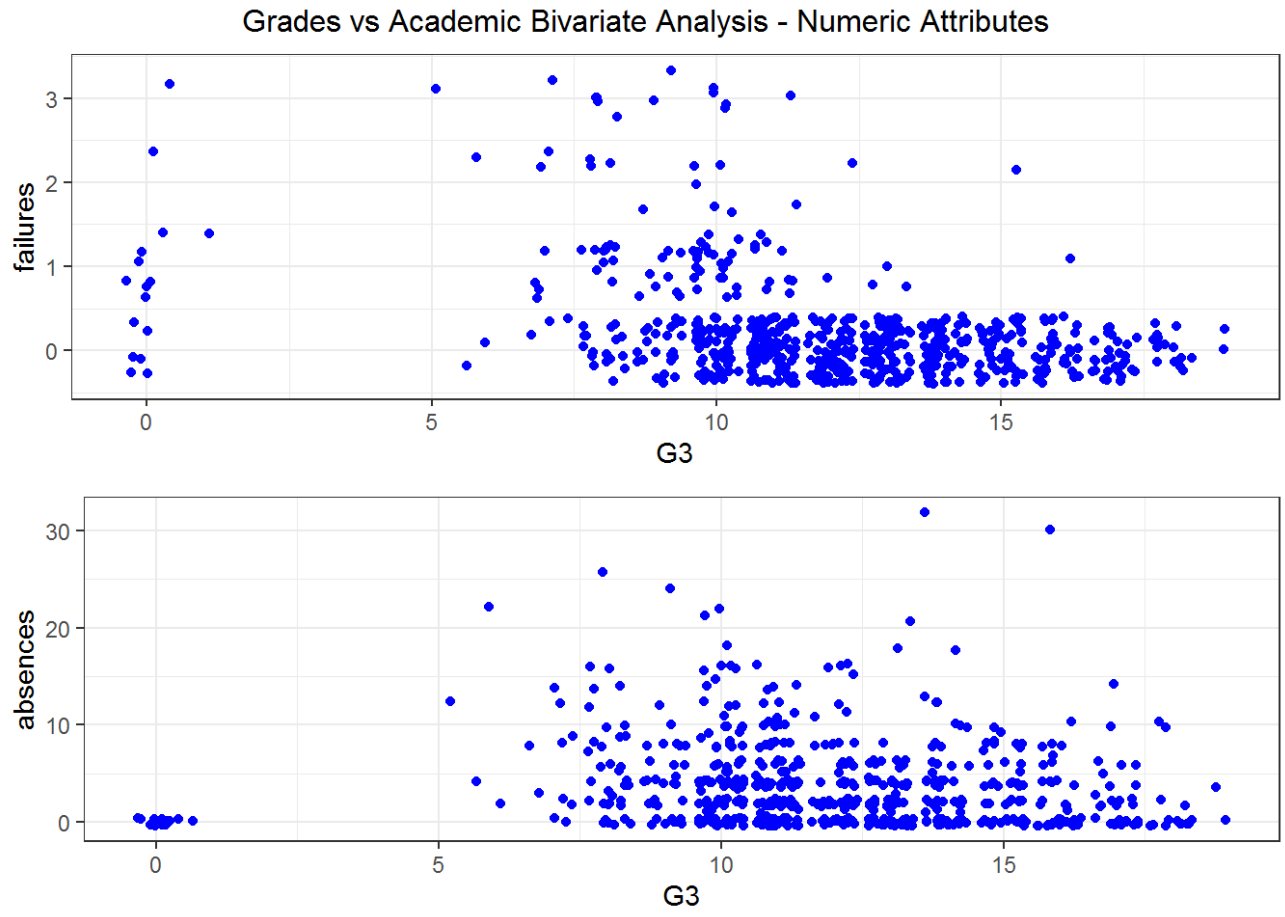
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = G3 ~ Fjob, data = student)
##
## $Fjob
##                         diff        lwr       upr     p adj
## health-at_home    1.1366460 -1.1398868 3.4131787 0.6498410
## other-at_home     0.4624367 -0.9671475 1.8920210 0.9024803
## services-at_home  0.2012628 -1.3018594 1.7043850 0.9961581
## teacher-at_home   2.1547619  0.1614411 4.1480827 0.0266084
## other-health     -0.6742092 -2.5606411 1.2122227 0.8652923
## services-health  -0.9353831 -2.8781365 1.0073702 0.6806761
## teacher-health    1.0181159 -1.3245838 3.3608156 0.7578446
## services-other   -0.2611739 -1.0582940 0.5359461 0.8982699
## teacher-other     1.6923252  0.1595656 3.2250847 0.0220009
## teacher-services  1.9534991  0.3519321 3.5550661 0.0079584
```

Observations:

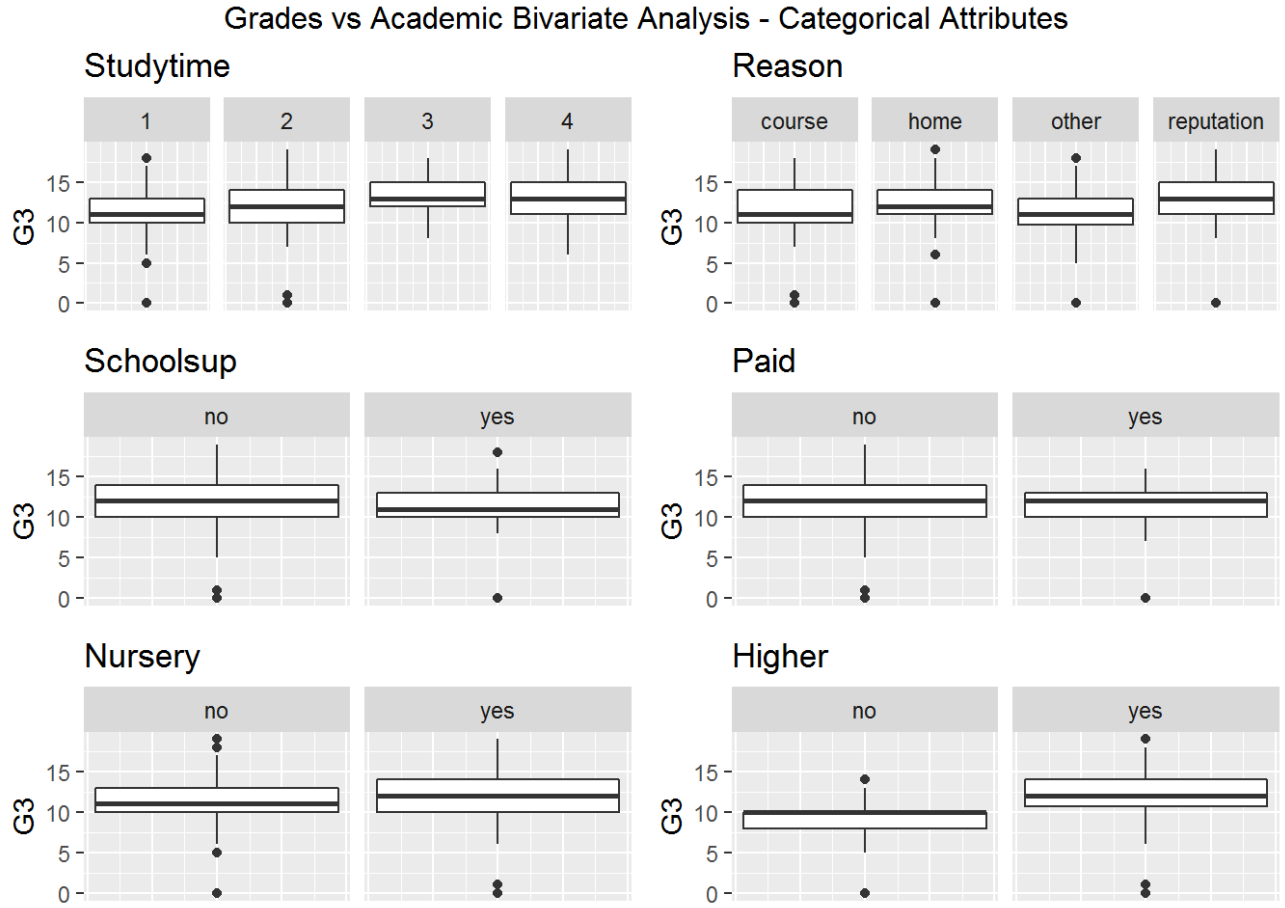
- Statistically significant differences in grades between Medu 4-2 and Medu 4-1, as observed in the visualizations. All other combinations are not statistically significant.
- Statistically significant differences in grades between Fedu 4-1 and Fedu 3-1. All other combinations are not statistically significant.
- Statistically significant differences in grades between Mjob health-at_home, teacher-at_home and teacher-other. All other combinations are not statistically significant.
- Statistically significant differences in grades between Fjob teacher-services. All other combinations are not statistically significant.

5.3 Academic

Code



Code



Observations:

- Best grades are within students that do not have any failure (reasonable).

- In general, best grades are found in students with less than 10 absences.
- Grades increase slightly as studytime increases.
- Grades are considerably higher in students that wish to pursue higher education than those who don't.

5.3.1 Significance Testing

Code

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## studytime     3    465  155.03   15.88 5.71e-10 ***
## Residuals   645   6298    9.76
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Code

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = G3 ~ studytime, data = student)
##
## $studytime
##           diff        lwr      upr     p adj
## 2-1  1.2474637  0.5277536 1.967174 0.0000559
## 3-1  2.3824645  1.3958316 3.369097 0.0000000
## 4-1  2.2128032  0.7442940 3.681312 0.0006612
## 3-2  1.1350008  0.1967750 2.073227 0.0103107
## 4-2  0.9653396 -0.4710944 2.401774 0.3083885
## 4-3 -0.1696613 -1.7567354 1.417413 0.9927036
```

Code

```
##
##  Welch Two Sample t-test
##
## data:  G3 by higher
## t = -9.1593, df = 86.036, p-value = 2.323e-14
## alternative hypothesis: true difference in means between group no and group ye
s is not equal to 0
## 95 percent confidence interval:
##  -4.233783 -2.723738
## sample estimates:
##  mean in group no mean in group yes
##          8.797101         12.275862
```

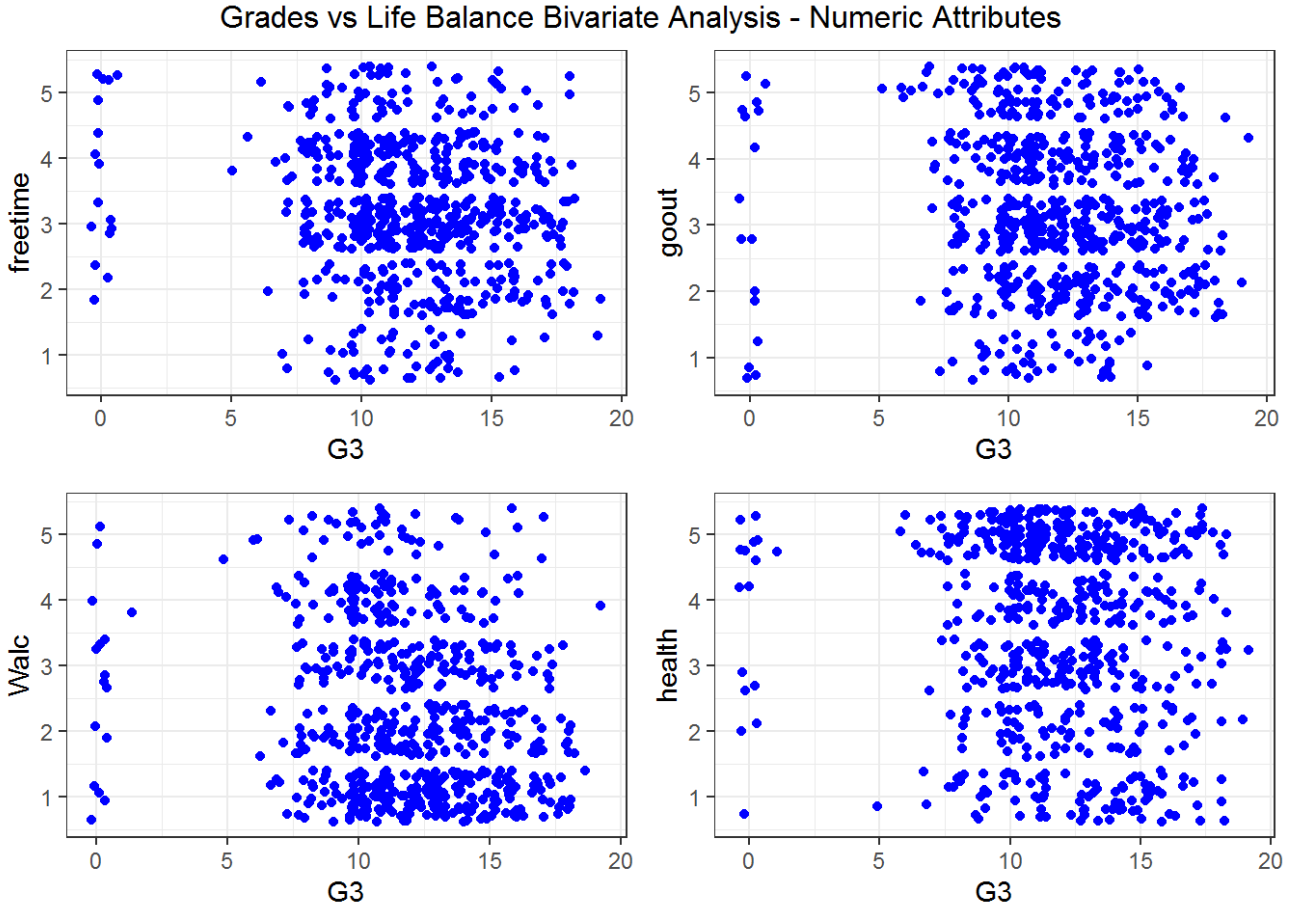
Observations:

- Statistically significant differences in grades between studytime 1 and 2,3,4. All other combinations are not statistically significant.

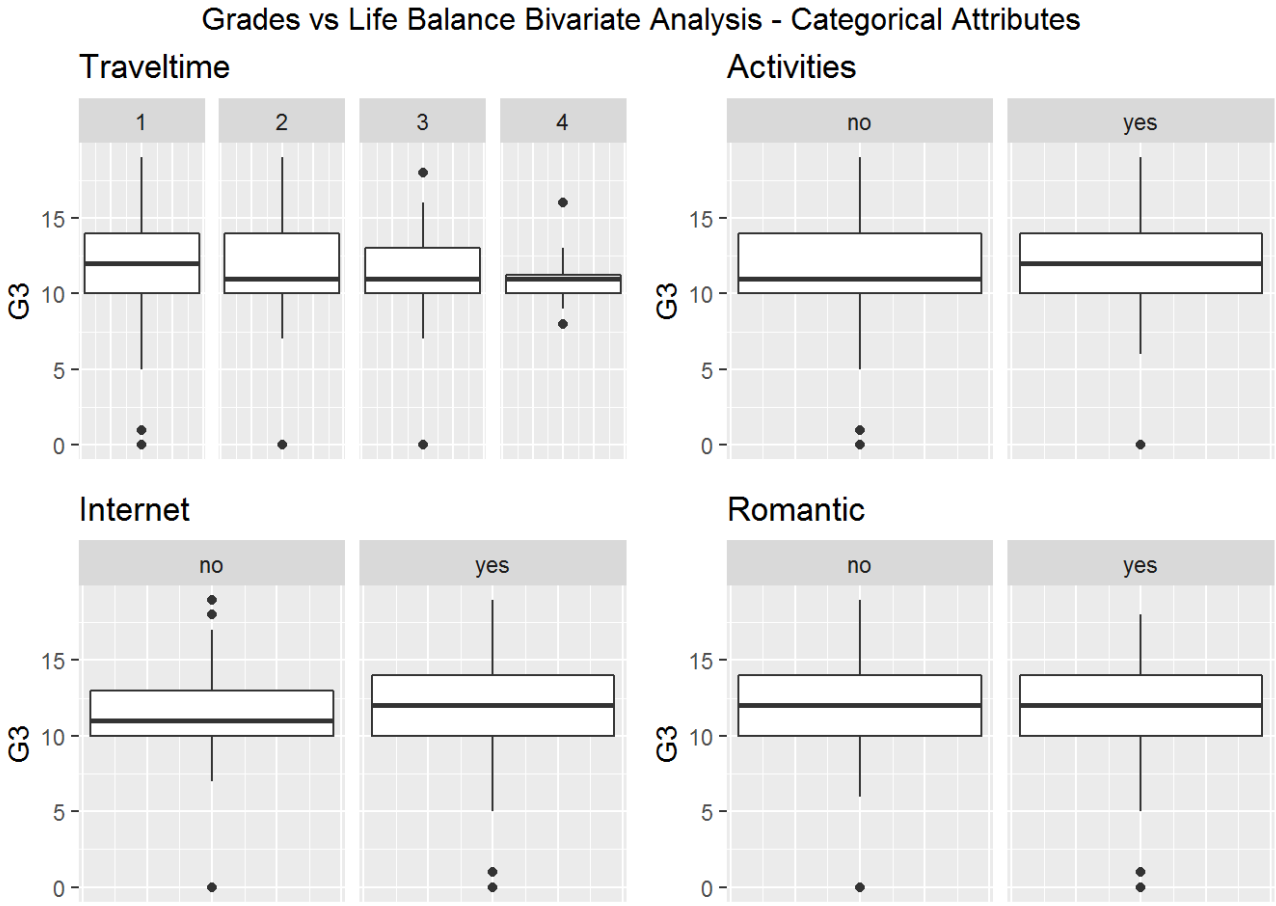
- Statistically significant difference in grades between higher = yes and higher = no. Group means difference is high.

5.4 Life Balance

Code



Code



Observations:

- Grades are to consistently lower for students with a high traveltime.
- No clear difference in grades for the other Life related attributes.

5.4.1 Significance Testing

Code

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## traveltime    3    113   37.72   3.659 0.0123 *
## Residuals   645   6650   10.31
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Code

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = G3 ~ traveltime, data = student)
##
## $traveltime
##           diff       lwr        upr     p adj
## 2-1 -0.6739013 -1.386664 0.03886177 0.0715848
## 3-1 -1.0846995 -2.290356 0.12095678 0.0951406
## 4-1 -1.3763661 -3.488724 0.73599138 0.3360080
## 3-2 -0.4107981 -1.670899 0.84930249 0.8355013
## 4-2 -0.7024648 -2.846363 1.44143372 0.8334100
## 4-3 -0.2916667 -2.645786 2.06245239 0.9887514
```

Code

```
##
##  Welch Two Sample t-test
##
## data:  G3 by internet
## t = -3.658, df = 229.37, p-value = 0.0003153
## alternative hypothesis: true difference in means between group no and group ye
s is not equal to 0
## 95 percent confidence interval:
##  -1.7635937 -0.5288077
## sample estimates:
##  mean in group no mean in group yes
##          11.02649          12.17269
```

Observations:

- No statistically significant differences in grades between traveltimes.
- Statistically significant difference in grades between internet = yes and higher = no. Group means difference is not high though.

I did not investigate the relationship between all categories vs categories since our outcome variable is not categorical.

6 Data Mining Techniques - Analytical Ready Data

Right now, I got a better understanding of the data and the variables. There weren't any major Data Quality Issues and I droped the variables that were not useful for the analysis. The data is ready for modelling. Depending on the data mining method, I performed additional changes to the dataset.

I performed 3 Data Mining techniques to better understand the data and the relationships between the variables: * Clustering

- Principal Components Analysis

- Regression (Trees)

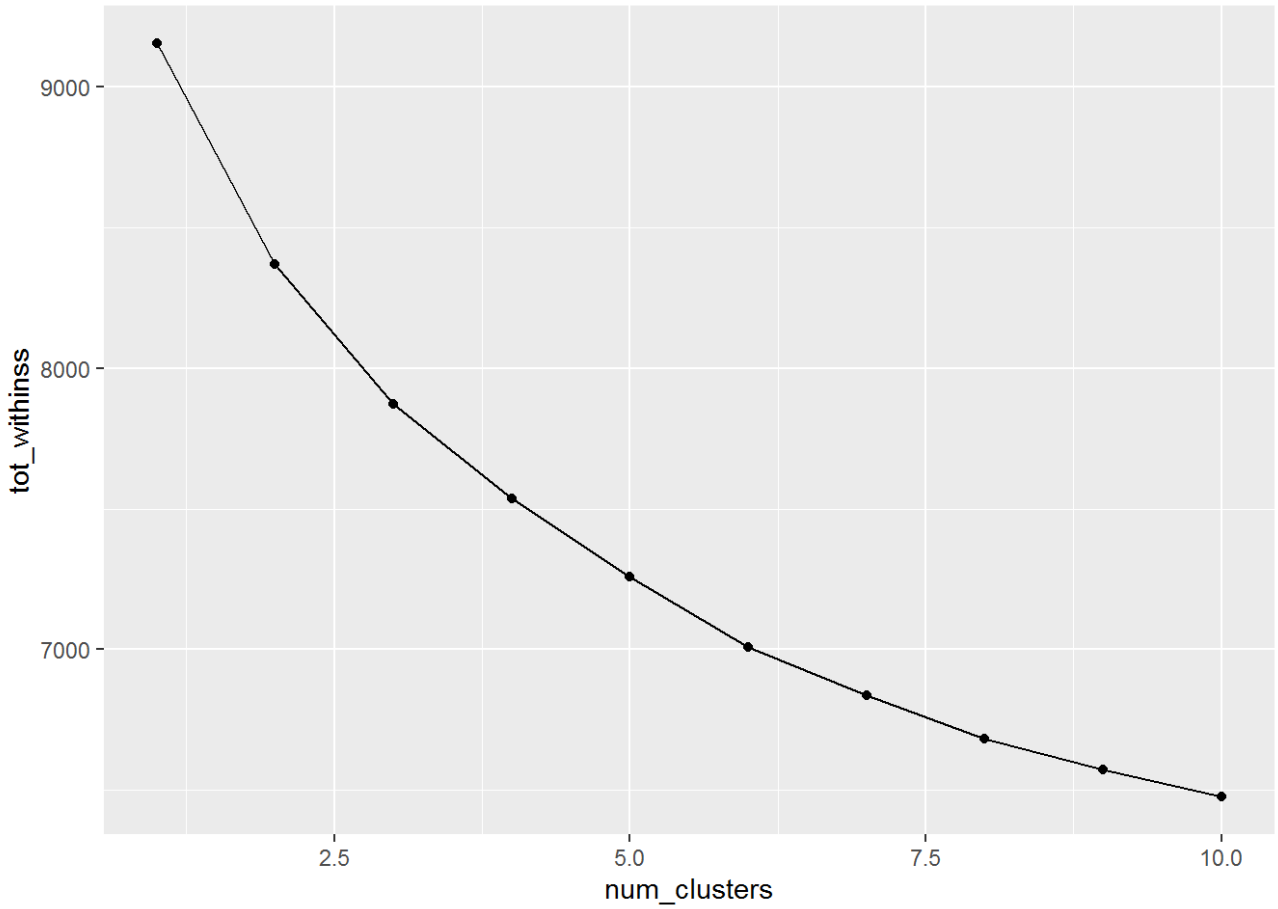
7 Clustering

Code

```
## $H
## [1] 0.280409
```

There are natural clusters in the data: hopkins statistic below 0.5

Code



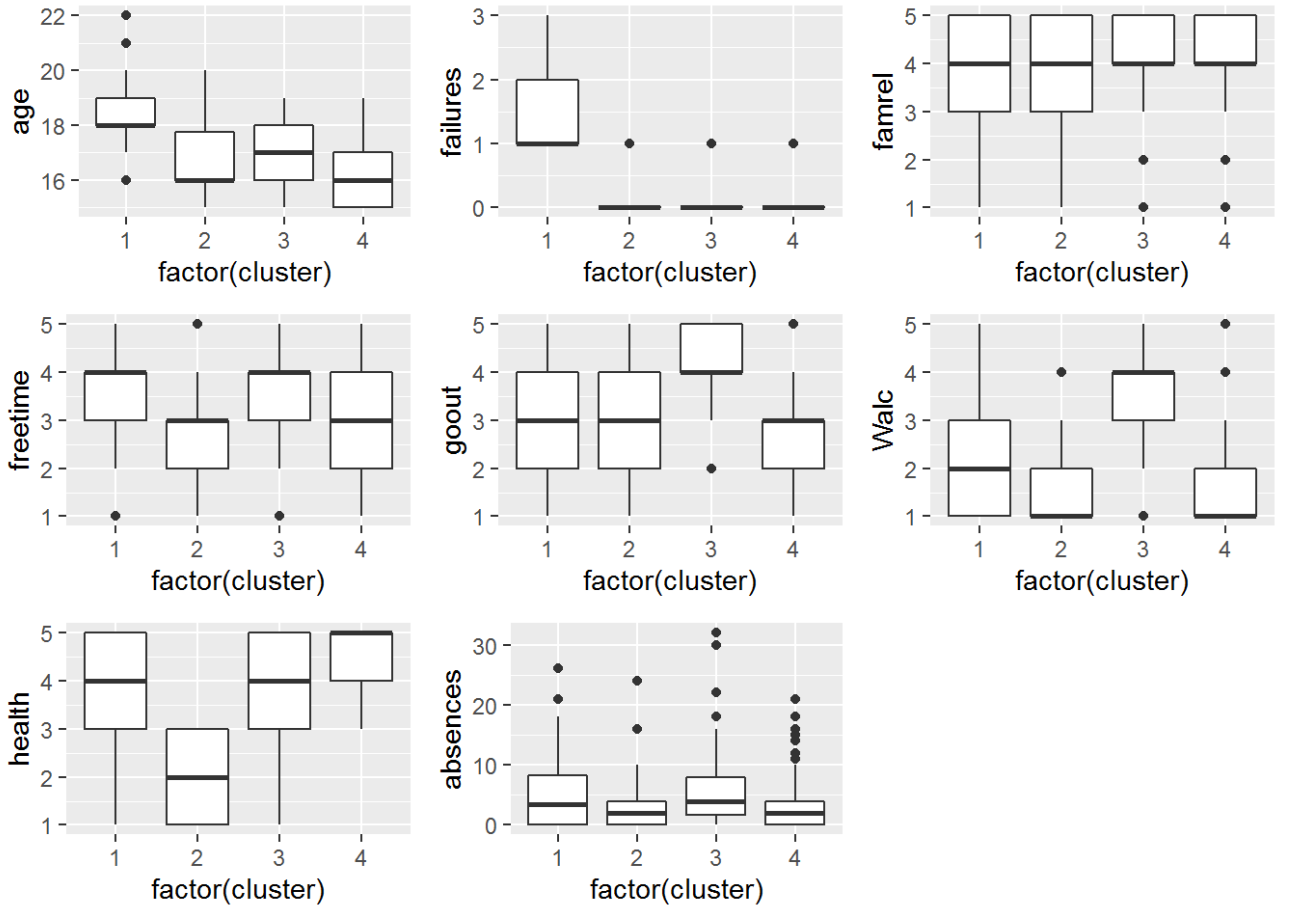
It seems that either 4 or 5 could be a good choice for the number of clusters.

Code

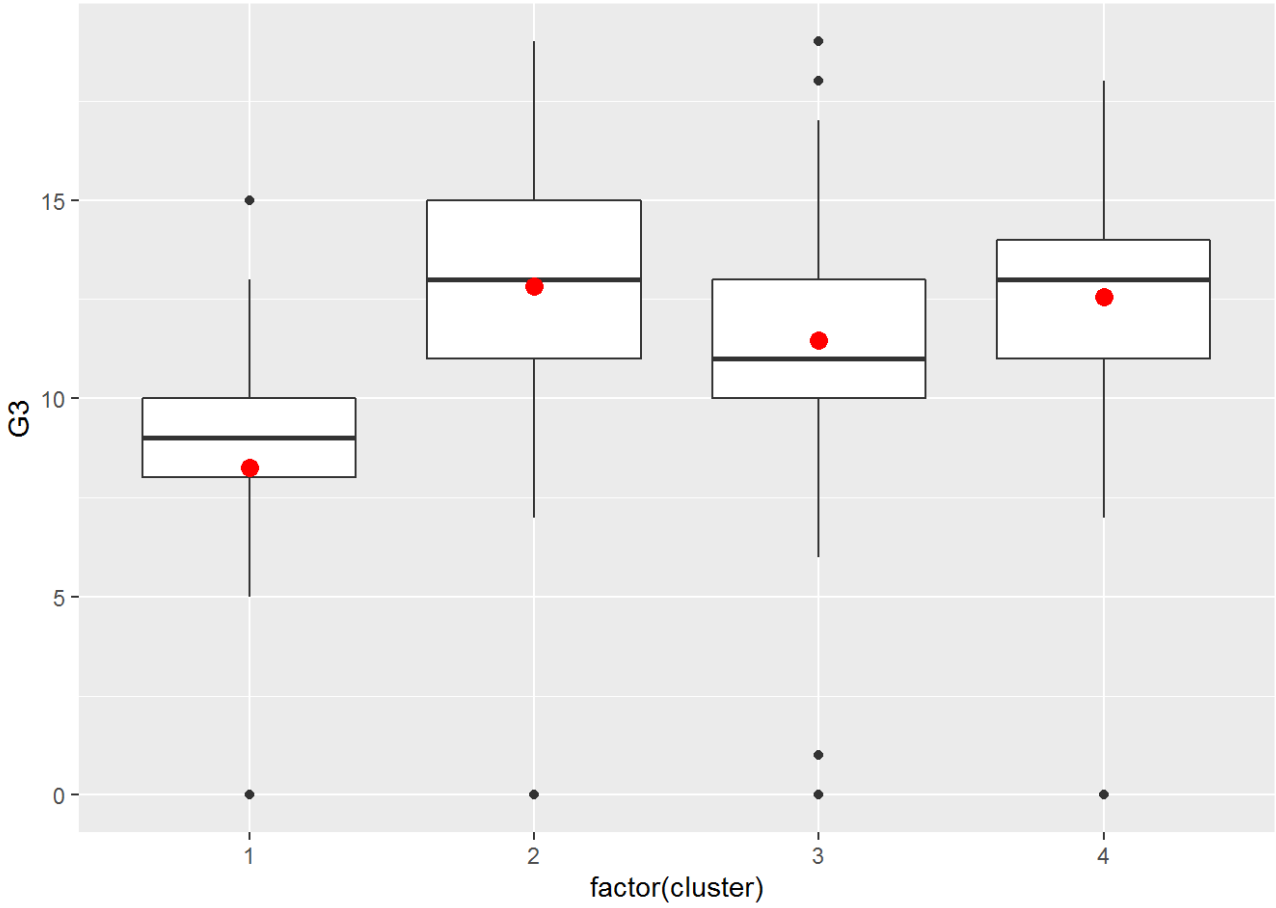
nms	V2	V3	V4	V5
cluster	1	2	3	4
Size	68	190	168	223
G3	8.26	12.84	11.45	12.57
age	18.25	16.6	16.87	16.31
failures	1.65	0.05	0.1	0.03
famrel	3.79	3.81	3.93	4.08
freetime	3.5	2.79	3.71	3.01

nms	V2	V3	V4	V5
goout	3.21	2.92	4.28	2.58
Walc	2.32	1.71	3.63	1.74
health	3.81	1.92	3.91	4.55
absences	5.1	2.59	5.38	2.83

Code



Code



Code

nms	V2	V3	V4	V5
cluster	1	2	3	4
Size	68	190	168	223
sex	F	F	M	F
address	U	U	U	U
famsize	GT3	GT3	GT3	GT3
Pstatus	T	T	T	T
Mjob	other	other	other	other
Fjob	other	other	other	other
reason	course	course	course	course
guardian	mother	mother	mother	mother
schoolsup	no	no	no	no
famsup	yes	yes	yes	yes
paid	no	no	no	no
Medu	1	2	2	4
Fedu	1	2	2	2
traveltime	2	1	1	1

nms	V2	V3	V4	V5
studytime	1	2	2	2
activities	no	no	yes	no
nursery	yes	yes	yes	yes
higher	yes	yes	yes	yes
internet	yes	yes	yes	yes
romantic	no	no	no	no

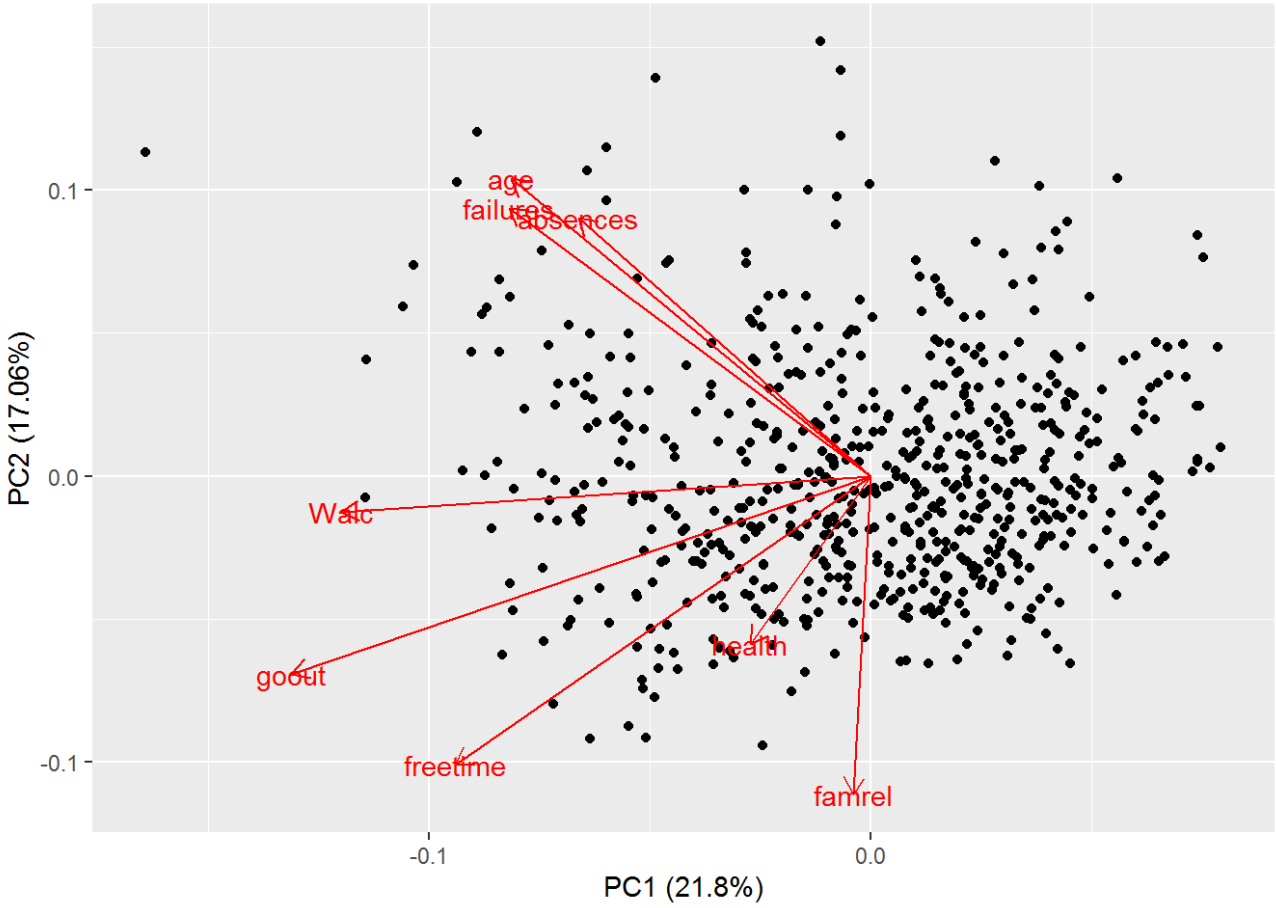
Clusters Description:

- Cluster 1: Majority of students, have higher grades, youngest students, almost no failures, absences or Alcohol consumption, highest Mother's education and females. Good family relationship and health but not differentiable from the other cluster in all the remaining variables. In general, really good, responsible, young students with good family relation and education.
- Cluster 2: Second highest number of students, have the highest grades in average. Similar characteristics as cluster 1 (young, females, no failures, absences or alcohol consumption). However they go out more frequently, have the worst health among the 4 clusters, and their Mother's education is not that high as cluster 1. In general, best students, responsible, young but maybe a little more loose, with worse health and more go out time and worst family education. with good family relation and education
- Cluster 3: Relatively good grades, high go-out times and Alcohol consumption, several absences, mostly males. All the other variables are to some extent the same as the other clusters, no differentiation.
- Cluster 4: Minority, oldest students with the worst grades (pretty low). High number of failures and absences. However no high alcohol consumption or go-out times. Probably the students that repeated courses one or many times. Mostly females, with the lowest mother and father's education and study time among the 4 clusters.

Principal variables used to differentiate: age, failures, absences, Walc, goout, Medu, Fedu, sex.

8 Principal Components Analysis

Code


age
failures
absences
Walc
goout
freetime
health
famrel
0.1
0.0
-0.1
-0.1
0.0
PC2 (17.06%)
PC1 (21.8%)

Code

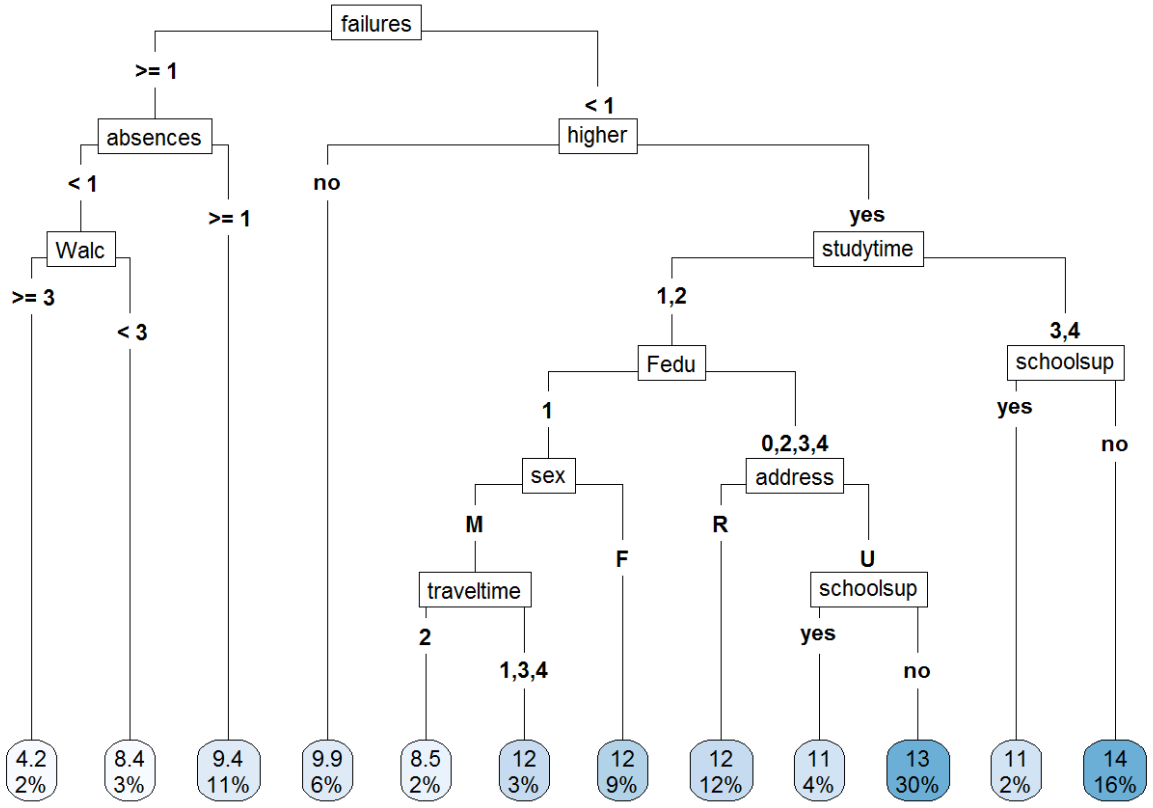
```
##                 PC1         PC2        PC3         PC4         PC5         PC
6
## age      -0.33507846  0.42946436 -0.3868196  0.14269682  0.07337588  0.4536463
2
## failures -0.33730866  0.38644669 -0.4658193  0.06497017 -0.30747743 -0.2195446
9
## famrel   -0.01649139 -0.45967936 -0.4752590  0.17703537  0.60319437  0.1954076
3
## freetime -0.38756734 -0.41653955 -0.1218389  0.26381461 -0.25163446 -0.5451468
1
## goout    -0.54069839 -0.28529965  0.2441009  0.22242655  0.02838899  0.2049119
3
## Walc     -0.49399503 -0.05133313  0.3930659 -0.33230608 -0.06113315  0.3323443
3
## health   -0.11248664 -0.24213904 -0.3673919 -0.83428085 -0.04256201 -0.0659151
5
## absences -0.27281094  0.37248720  0.2088162 -0.13620345  0.68304870 -0.5038348
0
##                 PC7        PC8
## age      -0.53483126 -0.1899146
## failures  0.58346987  0.1802601
## famrel    0.33068574 -0.1408176
## freetime -0.29524484 -0.3796960
## goout    -0.07303492  0.6848807
## Walc      0.34758406 -0.5033895
## health   -0.21734000  0.2105258
## absences -0.06008672  0.0256210
```

Observations:

- Component 1 seems to capture and represent primarily the variables Go out and Weekend Alcohol Consumption roughly evenly (they have fairly the same magnitude in coefficient). Other variables that are considered in the Principal Component 1 are age, failures and free time but with lower magnitude. Health and absences are considered but in a pretty lower magnitude. Seems that Component 1 captures the variance primarily of goout and Walc. Family relationship seems not to be in component 1 with a coefficient close to 0.

9 Regression Trees

Code



Code

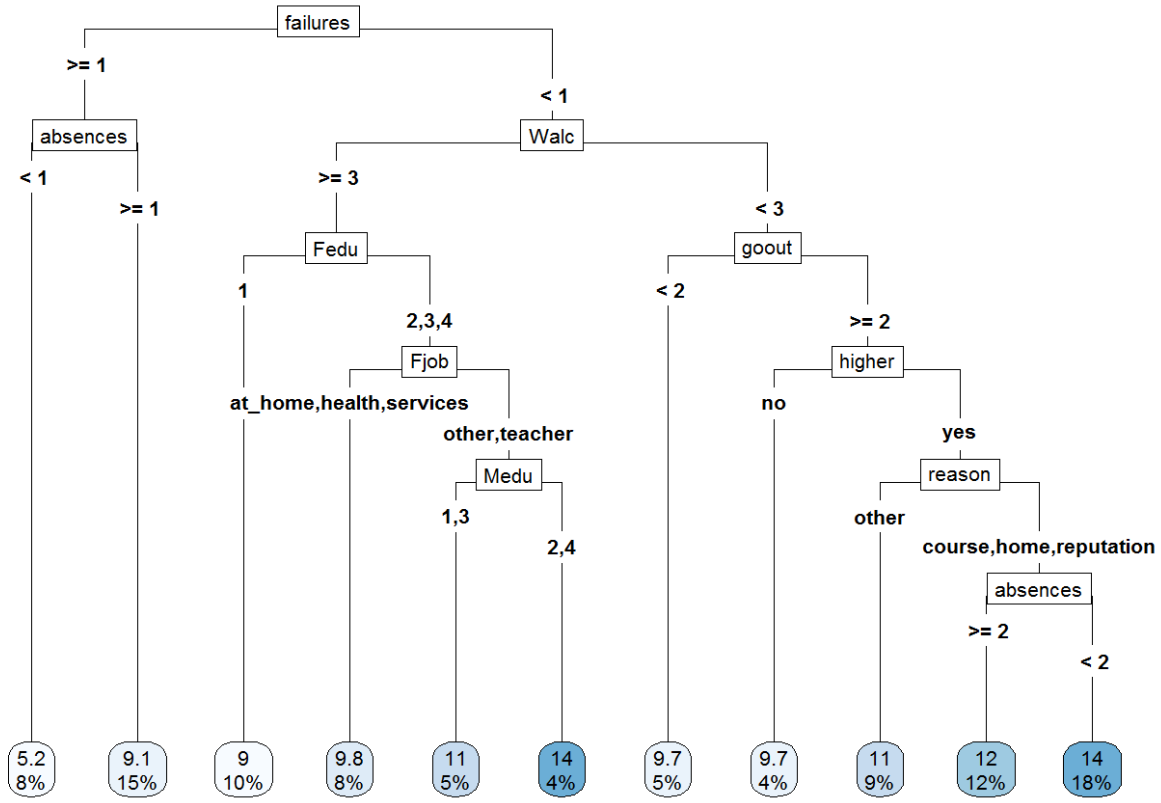
```
##    failures      higher    schoolsup        Walc    absences         age   stud
ytime
## 1299.881664  264.688077  181.189238  168.961472  166.728973  153.140254  144.1
85284
##  traveltime        Fedu         sex     address        Fjob        Medu      f
amrel
##  127.507127  125.808058   95.017653   72.539099   48.973992   44.509982   39.8
22222
##       goout     nursery    guardian    freetime      famsup        Mjob
##   30.459812   17.066667   12.998817    6.853961    6.225926    6.225926
```

Observations:

- Some of the most important variables that are used to split the tree are: failures, higher, absences, Walc and schoolsup.
- By looking at the variable importance for the tree, we also see that failures, higher, schoolsup, Walc and absences are the most important variables that are used to make predictions (in that order).
- We see that a number of failures greater or equal to 1 considerably reduces the prediction of a student grade. The same with higher = no.
- High alcohol consumption reduces the prediction of a high grade as well.
- However, something I found somehow odd is that branches with schoolsup = no have considerably higher grades than branches with schoolsup = yes. However the best 2 students in the dataset, both have schoolsup = no.

Since I found statistically significant difference in G3 in both schools, I analyzed those 2 subpopulations deeper.

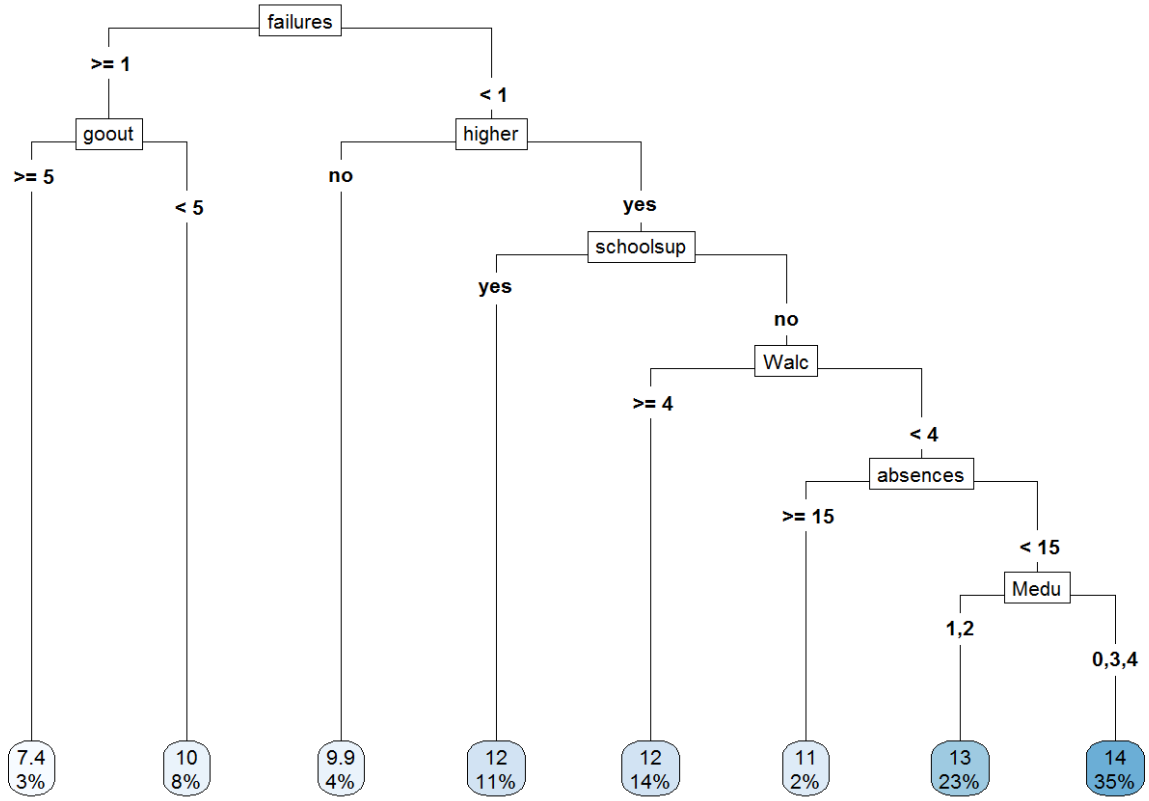
Code



Code

```
##   failures   absences        age       Walc   guardian       Fedu      goout
## 580.567855 267.469179 142.403436 111.618767 111.429301 108.726354 104.875164
##     reason       Fjob     higher       Medu       Mjob    address  studytime
##  93.363142  90.608390  76.716808  53.620022  52.712434  38.356641  31.855186
##  schoolsup        sex traveltime   freetime     famsup       paid     famrel
##  28.734168  28.347623  20.500541  18.130541  16.182323  12.413780   6.864286
##    nursery
##   6.864286
```

Code



Code

```
##   failures     higher  schoolsup       Walc      goout   absences        age
## 490.231678 152.406649 126.003846  91.187284  72.099747  52.317859  31.291384
##       Medu       Mjob     reason       Fedu traveltime   internet    nursery
##  29.205455  13.234404  12.016624  10.839138   8.718097   1.505436   1.505436
```

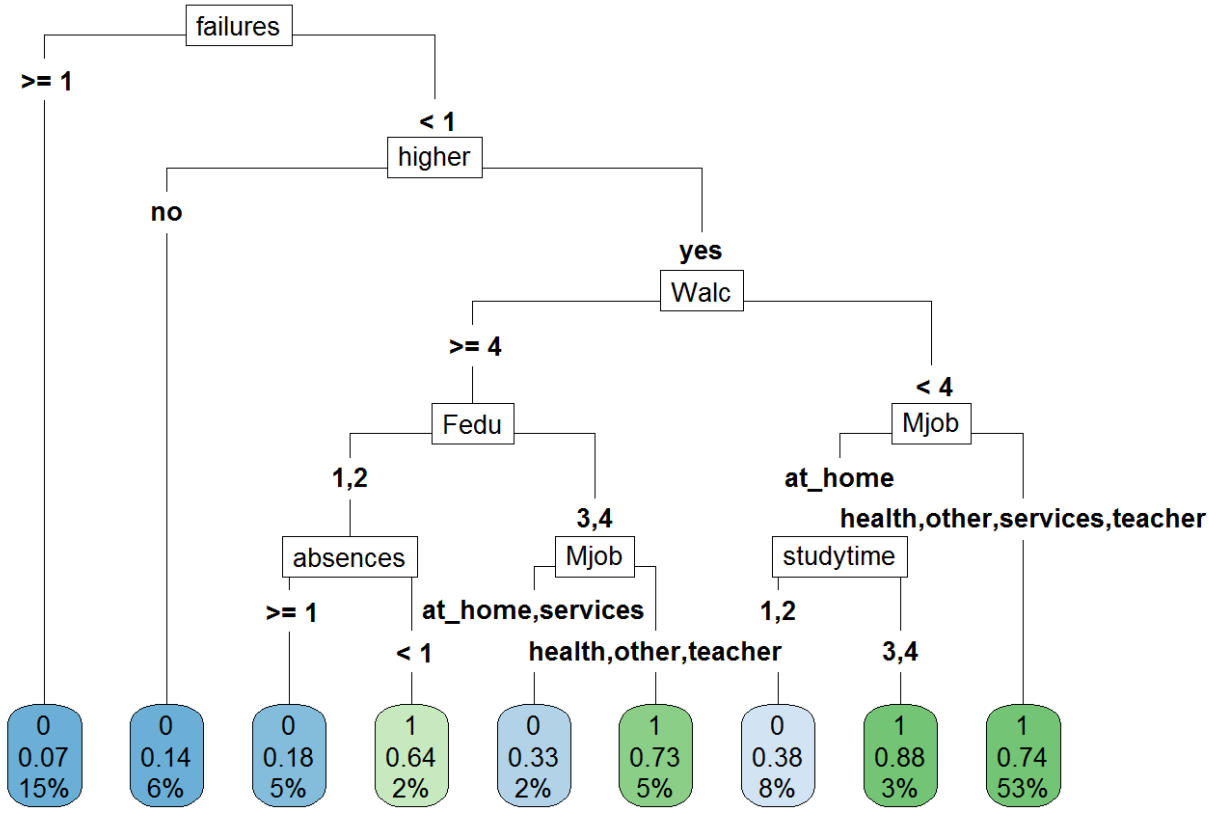
Observations:

- Same variables more or less are the most important in the 2 subpopulation trees, however I found much more complicated trees in the subpopulations.
- Variables like failure, absences, higher, age, Walc and schoolsup are also found among the top important variables.
- Some other variables are also used but in a smaller magnitude, like guardian, Fedu and goout.
- In general, variables used for the 2 subpopulations trees are almost the same as the general tree

9.1 Binning - Classification Trees

The Histogram of the Grades (G3) does not show reasonable separations in order to do binning. However, I tried converting the final grade G3 to a pass/no pass grade and performing classifications methods on this variable. To convert G3 to a pass/no pass variable, I set up a threshold of above 60% of the total scale (20) is a pass, and below a no pass.

Code



Code

```
##   failures     higher       Mjob       Walc  studytime        age   absences
## 51.3883514 17.9190394 11.3705224  9.1330304  7.1232484  5.6527187  5.2547268
##       Fedu       Medu traveltime   freetime     famsup   guardian   internet
##  4.0833333  3.2459176  1.6832809  1.1909722  1.0208333  0.5138835  0.4267677
```

Observations:

- Important variables for classifying a student grade into pass/no pass are also failures, higher and Walc. However, Mjob and studytime have gained more importance when using trees to classify and not to predict the actual grade.

10 Final Conclusions and Recommendations

By analyzing all the bivariate relationships, the output to variables relationships, as well as the main features used for the Principal Components Analysis, Clustering and Regression and Classification Trees, we can summarize the most important variables and relationships as follows:

- Clear relationships between variables and G3:
- Younger age have better grades
- High Mothers and fathers education (Medu, Fedu) are related to higher grades
- Decrease in failure is related to an increase in grades
- Decrease in absences is related to an increase in grades
- A desire to pursue higher education is related to higher grades

- High traveltime -> low grades

*Top features used for clustering: age, failures, absences, Walc, goout, Medu, Fedu, sex.

*Top features found in PCA: gout, Walc, age, failures, freetime.

*Top features used in Regression trees: failures, higher, absences, Walc and schoolsup.

Summarizing these findings we can state to some extent that the following variables are among the best predictores since they are more frequent among all the techniques:

- Failures, absences and higher

Other really important variables are:

- Age, Walc, Goout, Sex, Medu, Fedu and Schoolsup.

From initally having over 30 attributes, the dataset can be reduced to principally 10 variables or columns, that capture most of the variation of the grades and act as the best predictors.

Focusing on the logical groupings, we can state that the most powerfull predictors can be found in the Academic group (Failures, absences, higher), which is completely reasonable since this focuses mainly on the academic side of the student. Regarding family related attributes, the Parents Education is the most important. When it comes to Life Balance, we see that Alcohol Consumption and Goout times play an important role in explaining a students performance, and regarding demographics, mainly sex and age are the best predictors.

Most of these variables are really personal attributes that students manage themselves. There are really no recommendations to be made regarding higher, failures, age, goout or Medu. However, some recommendations for the schools, if they want to increase the student's grades, would be:

- Try to lower the amount of absences in the students
- Disencourage the alcohol consumption
- Maybe focus the educational efforts more on males, than on females since males usally have lower grades in the dataset