STUDENT DATA ANALYSIS

Dataset description

This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details).

Attribute Information:

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

- school – student's school (binary: 'GP' – Gabriel Pereira or 'MS' – Mousinho da Silveira)
- sex – student's sex (binary: 'F' – female or 'M' – male)
- age – student's age (numeric: from 15 to 22)
- address – student's home address type (binary: 'U' – urban or 'R' – rural)
- famsize – family size (binary: 'LE3' – less or equal to 3 or 'GT3' – greater than 3)
- Pstatus – parent's cohabitation status (binary: 'T' – living together or 'A' – apart)
- Medu – mother's education (numeric: 0 – none, 1 – primary education (4th grade), 2 5th to 9th grade, 3 secondary education or 4 higher education)
- Fedu – father's education (numeric: 0 – none, 1 – primary education (4th grade), 2 5th to 9th grade, 3 secondary education or 4 higher education)
- Mjob – mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- Fjob – father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- reason – reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
- guardian – student's guardian (nominal: 'mother', 'father' or 'other')
- traveltime – home to school travel time (numeric: 1 – <15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour, or 4 – >1 hour)
- studytime – weekly study time (numeric: 1 – <2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours, or 4 – >10 hours)
- failures – number of past class failures (numeric: n if 1<=n<3, else 4)
- schoolsup – extra educational support (binary: yes or no)
- famsup – family educational support (binary: yes or no)
- paid – extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
- activities – extra-curricular activities (binary: yes or no)
- nursery – attended nursery school (binary: yes or no)
- higher – wants to take higher education (binary: yes or no)
- internet – Internet access at home (binary: yes or no)
- romantic – with a romantic relationship (binary: yes or no)
- famrel – quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
- freetime – free time after school (numeric: from 1 – very low to 5 – very high)
- goout – going out with friends (numeric: from 1 – very low to 5 – very high)
- Dalc – workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
- Walc – weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
- health – current health status (numeric: from 1 – very bad to 5 – very good)
- absences – number of school absences (numeric: from 0 to 93)

The following grades are related with the course subject, Math or Portuguese:

- G1 – first period grade (numeric: from 0 to 20)
- G2 – second period grade (numeric: from 0 to 20)
- G3 – final grade (numeric: from 0 to 20, output target)

Data Import

student-mat.csv

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	6	5	6	6
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	4	5	5	6
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	10	7	8	10
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	2	15	14	15
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	4	6	10	10
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
390	MS	M	20	U	LE3	A	2	2	services	services	...	5	5	4	4	5	4	11	9	9	9
391	MS	M	17	U	LE3	T	3	1	services	services	...	2	4	5	3	4	2	3	14	16	16
392	MS	M	21	R	GT3	T	1	1	other	other	...	5	5	3	3	3	3	3	10	8	7
393	MS	M	18	R	LE3	T	3	2	services	other	...	4	4	1	3	4	5	0	11	12	10
394	MS	M	19	U	LE3	T	1	1	other	at_home	...	3	2	3	3	3	5	5	8	9	9

395 rows × 33 columns

student-por.csv

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	4	0	11	11
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	2	9	11	11
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	6	12	13	12
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	0	14	14	14
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	0	11	13	13
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
644	MS	F	19	R	GT3	T	2	3	services	other	...	5	4	2	1	2	5	4	10	11	10
645	MS	F	18	U	LE3	T	3	1	teacher	services	...	4	3	4	1	1	1	4	15	15	16
646	MS	F	18	U	GT3	T	1	1	other	other	...	1	1	1	1	1	5	6	11	12	9
647	MS	M	17	U	LE3	T	3	1	services	services	...	2	4	5	3	4	2	6	10	10	10
648	MS	M	18	R	LE3	T	3	2	services	other	...	4	4	1	3	4	5	4	10	11	11

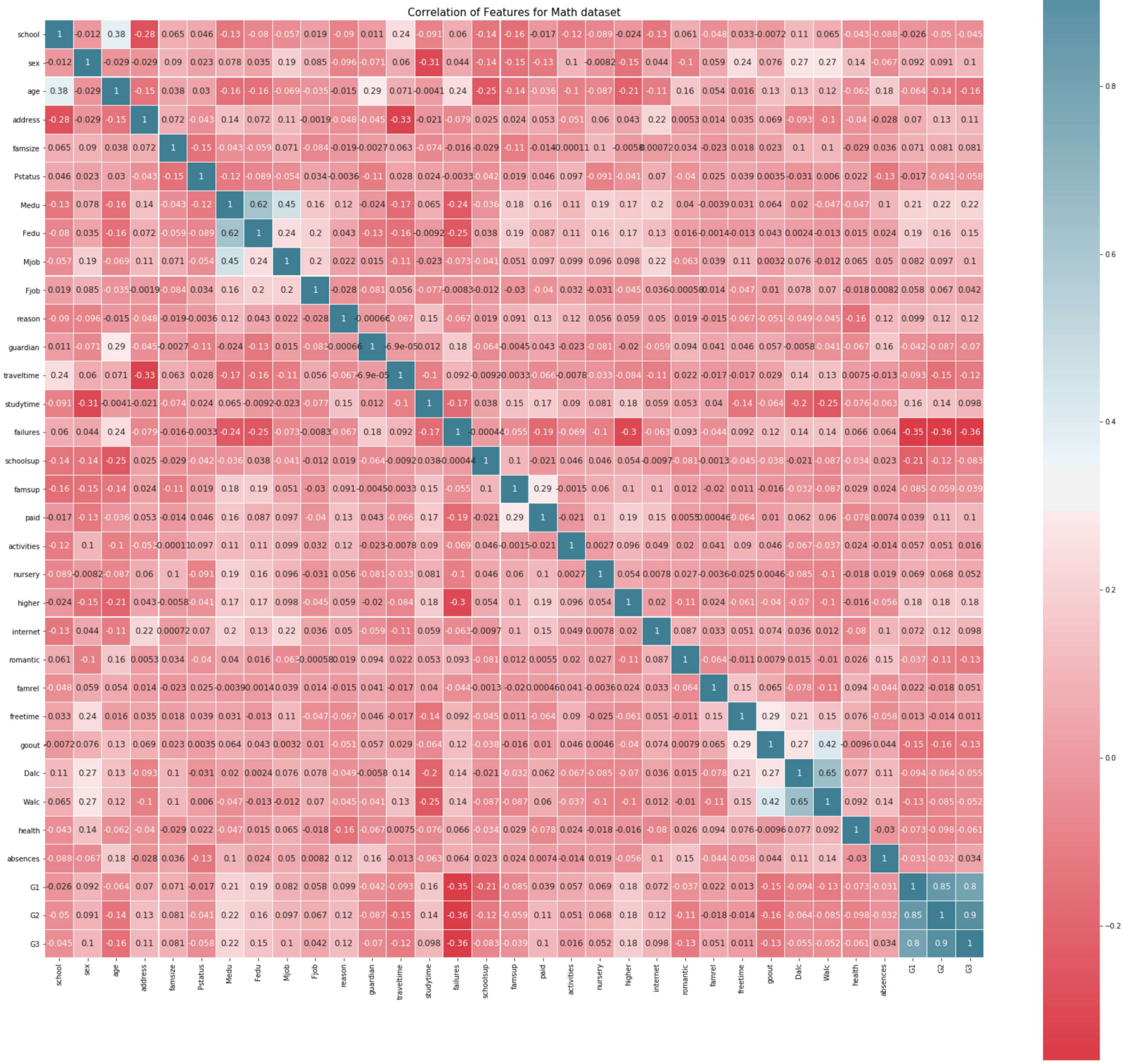
649 rows × 33 columns

As we can see, there is no features with continuous type. Except G1/G2 score, others are all binary/five-level data

Exploratory data analysis

Correlation analysis

Examine the correlation between ecah features and G3

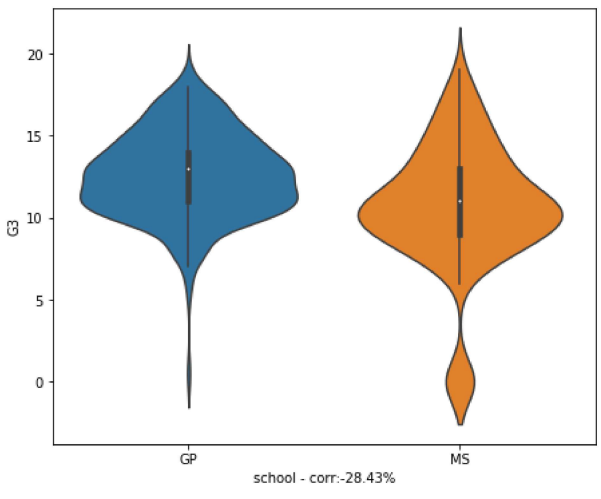


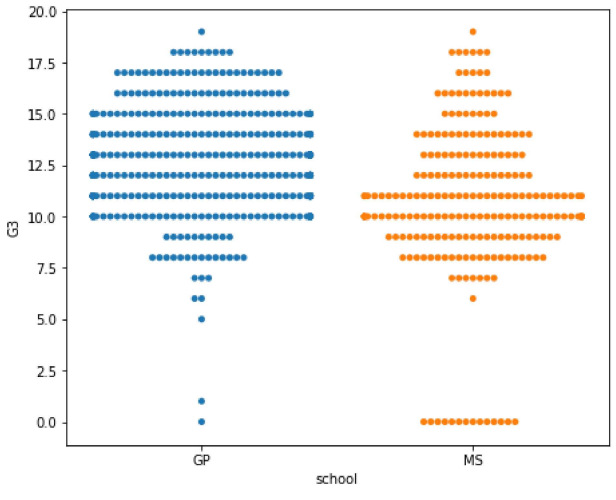
Some observation on it:

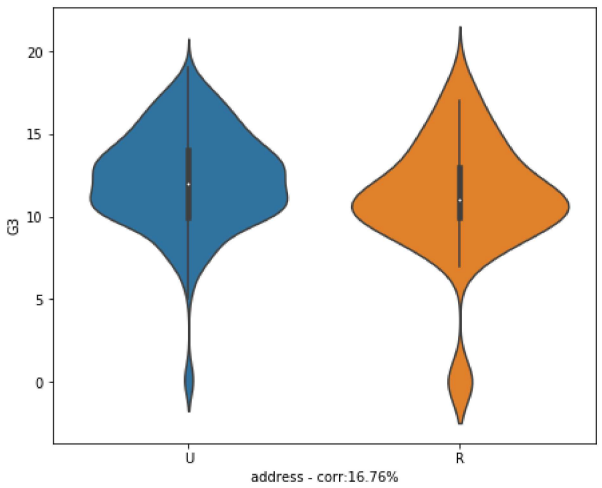
1. Mother and Father's education and jobs are related.
2. Alcohol consumption and gout are related.
3. G1, G2, G3 scores are related

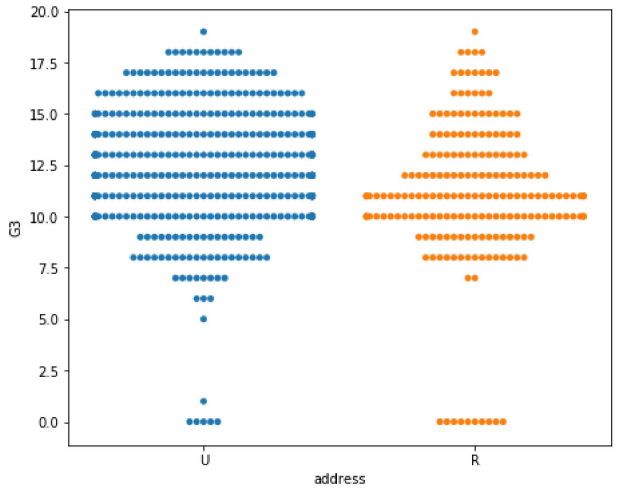
Correlation analysis for G3 in Portuguese language performance and other features

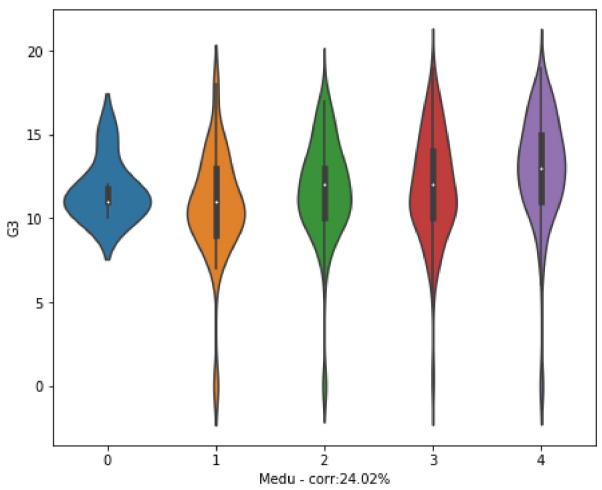
List the correlation value greater thean 15%

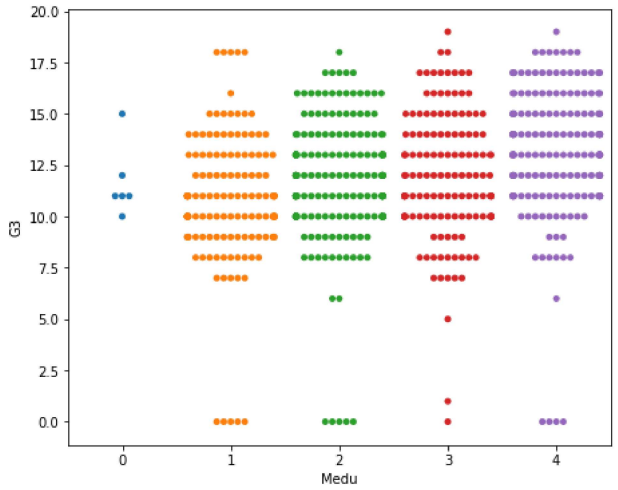


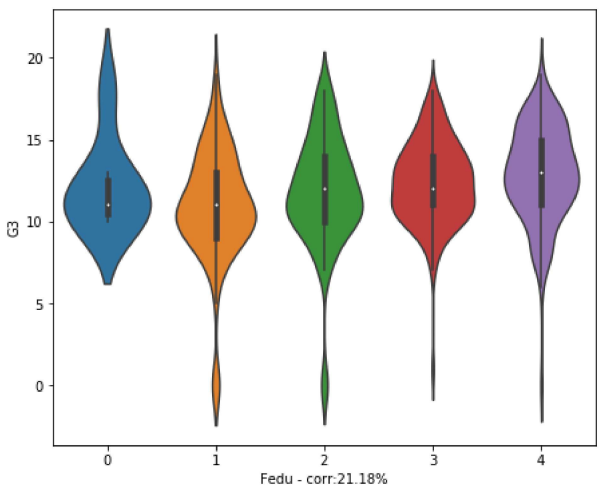


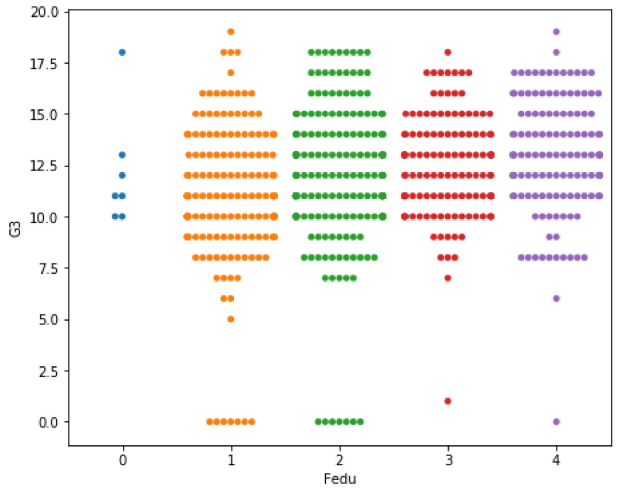

G3
U
R
address - corr:16.76%

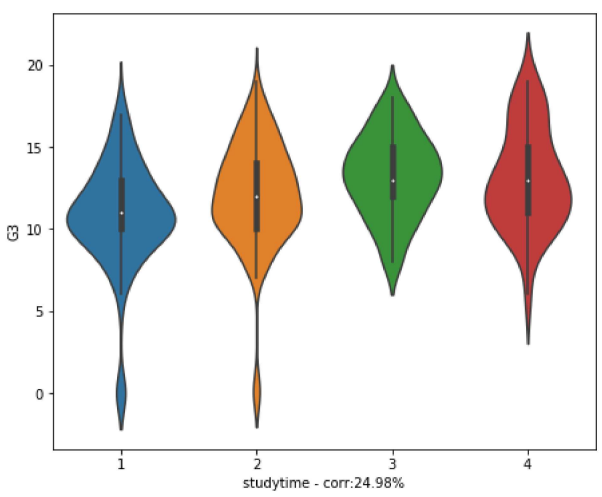

G3
U
R
address

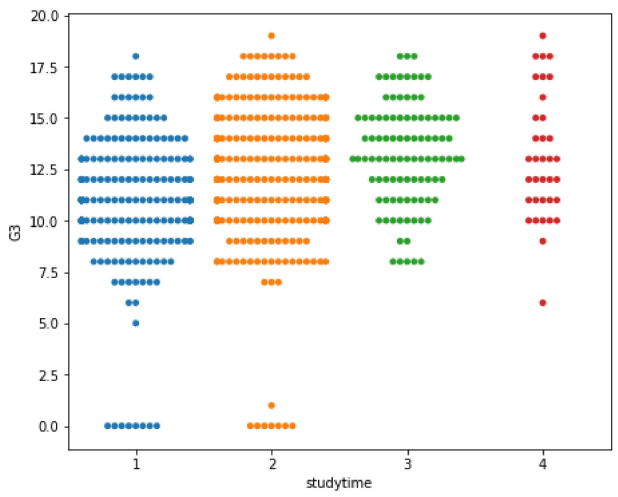

G3
Medu - corr:24.02%

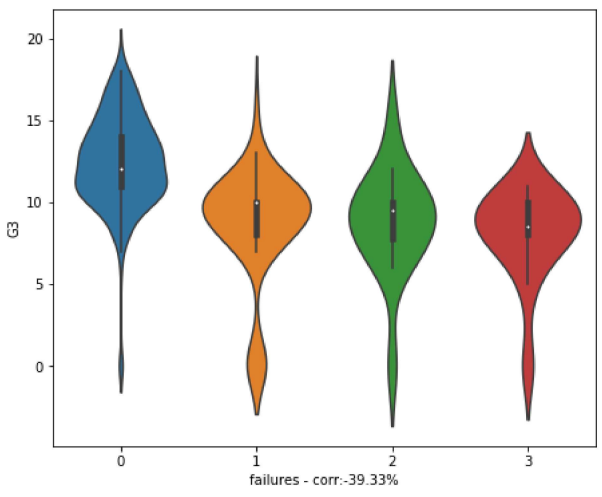

G3
Medu

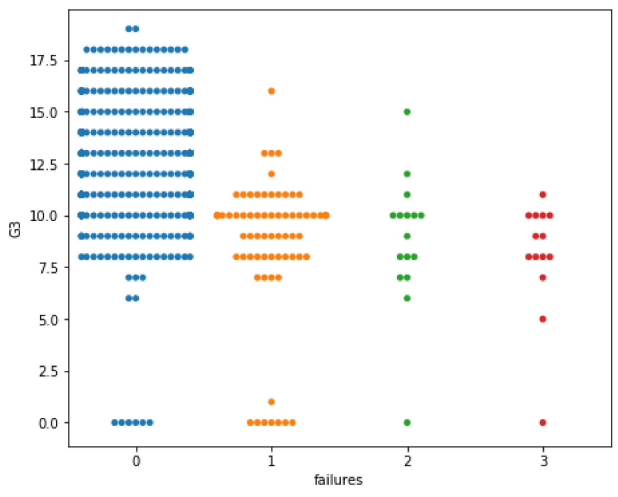

G3
Fedu - corr:21.18%

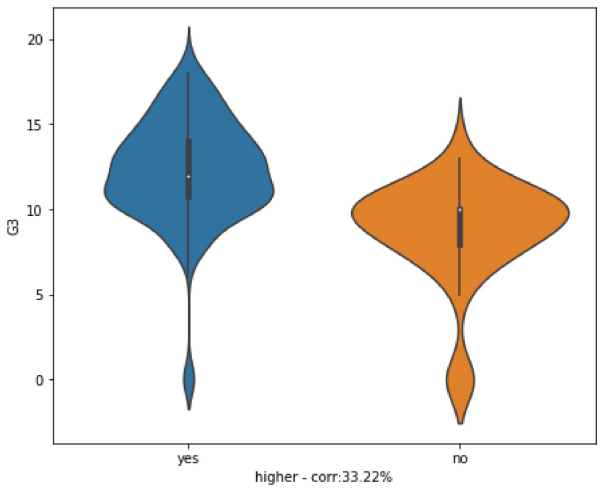

G3
Fedu

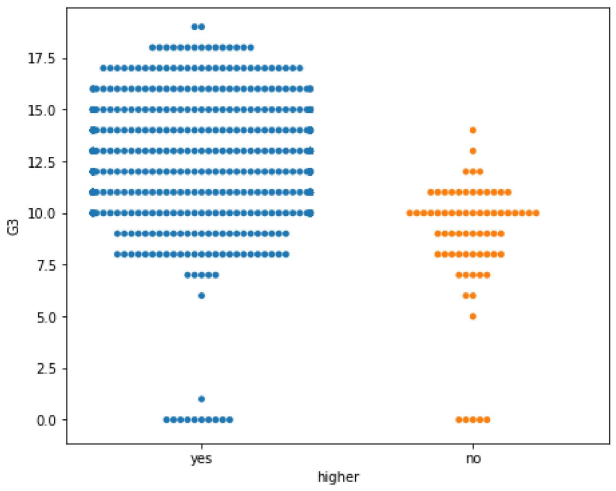

G3
studytime - corr:24.98%

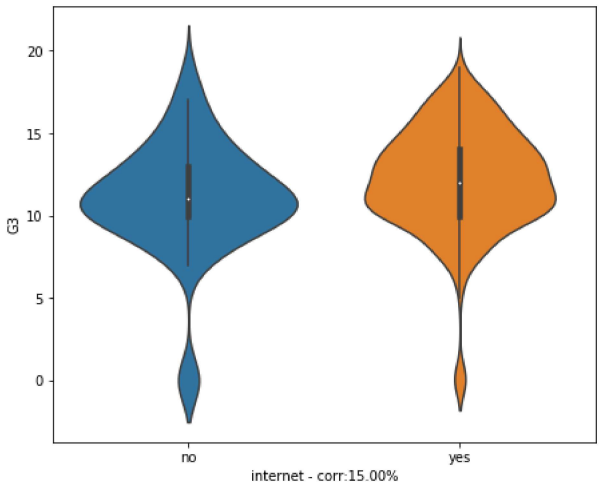

G3
studytime

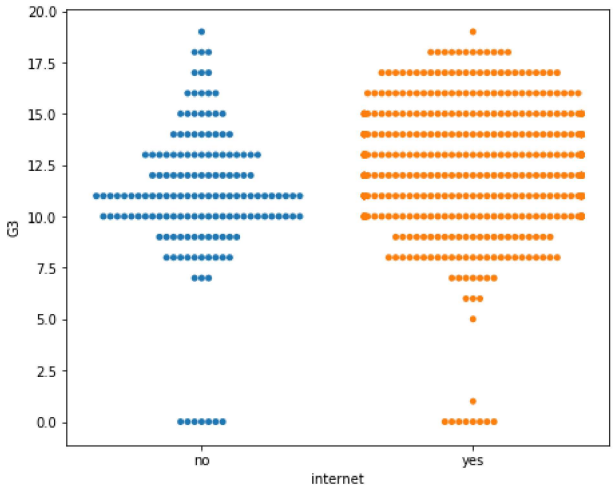

G3
failures - corr:-39.33%

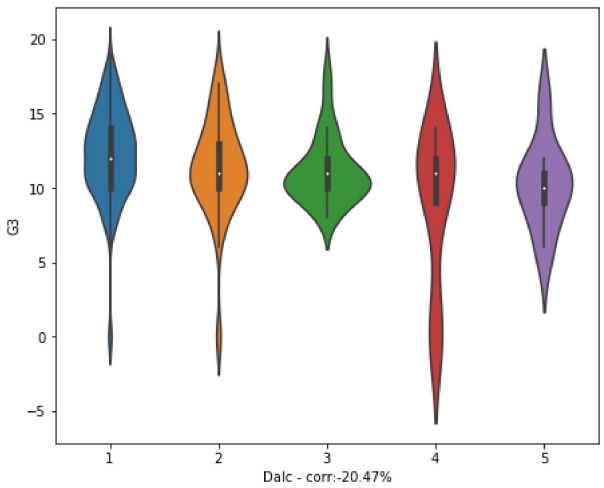

G3
failures

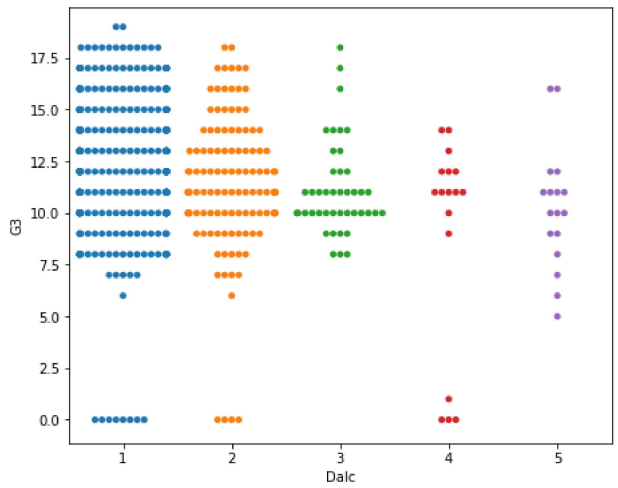

G3
yes
no
higher - corr:33.22%

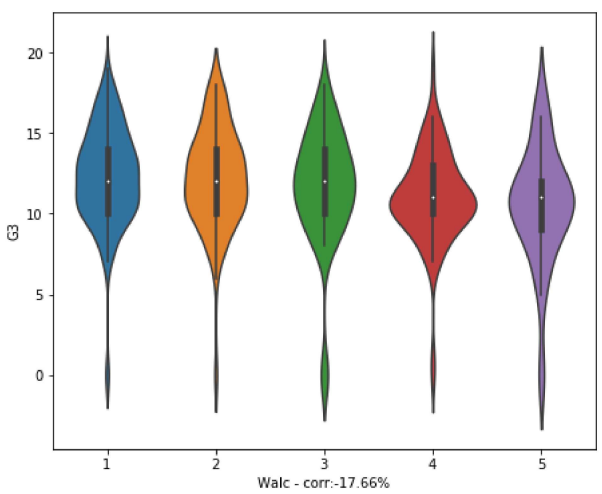

G3
yes
no
higher

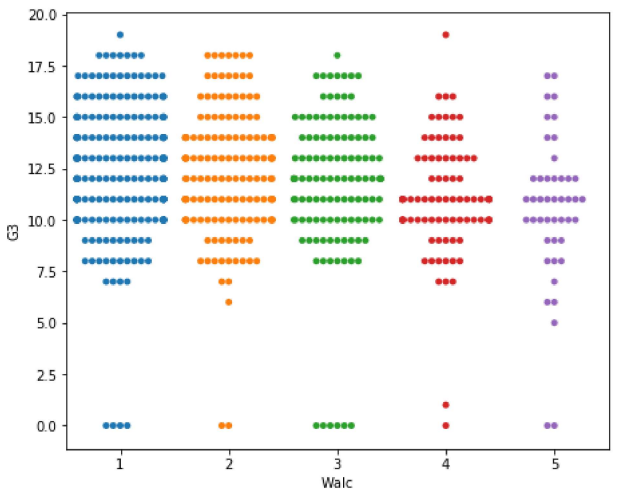

G3
no
yes
internet - corr:15.00%

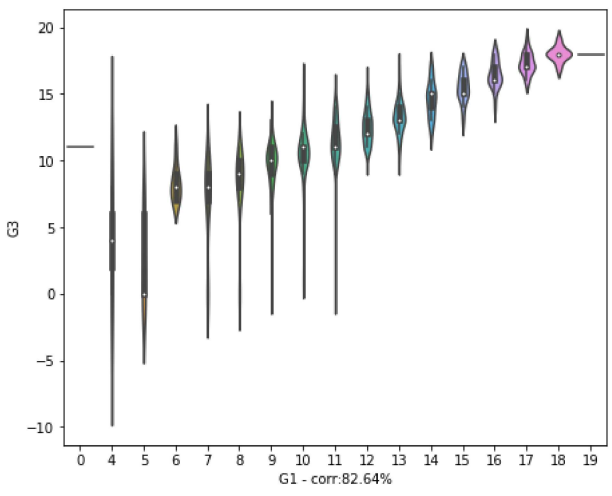

G3
no
yes
internet

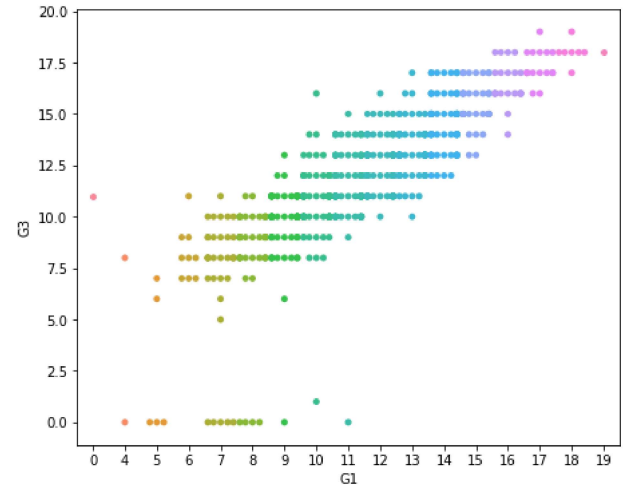

G3
Dalc - corr:-20.47%

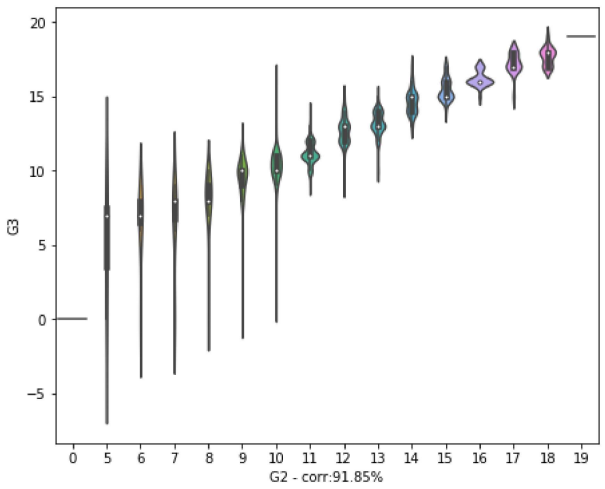

G3
Dalc

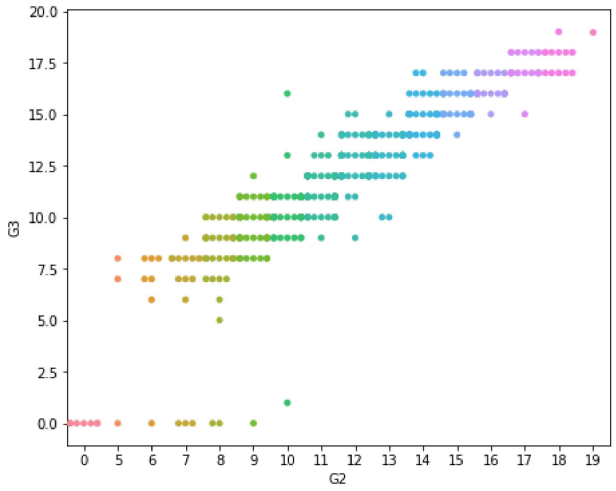

G3
Walc - corr:-17.66%


G3
Walc


G3
G1 - corr:82.64%


G3
G1





correlation table for student_mat.csv

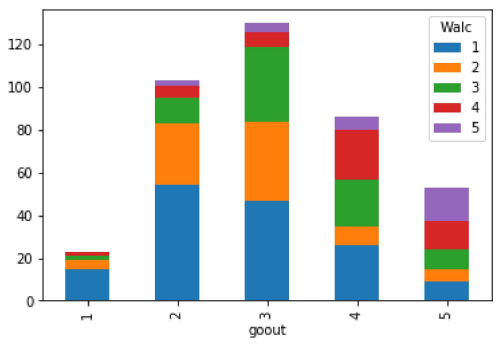
	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absenc
school	1.000000	-0.083050	0.087170	-0.354520	0.022252	0.028120	-0.254787	-0.209806	-0.206829	-0.081872	...	-0.031597	0.034666	0.044632	0.047169	0.014169	-0.058599	-0.1639
address	-0.354520	0.025503	-0.025848	1.000000	0.046113	-0.094635	0.190320	0.141493	0.159761	-0.006535	...	-0.033897	-0.036647	0.015475	-0.047304	-0.012416	0.003787	0.07365
Medu	-0.254787	0.119127	-0.107832	0.190320	-0.014325	-0.057174	1.000000	0.647477	0.459337	0.152582	...	0.024421	-0.019686	0.009536	-0.007018	-0.019766	0.004614	-0.0085
Fedu	-0.209806	0.083913	-0.121050	0.141493	-0.039538	-0.031856	0.647477	1.000000	0.290703	0.211604	...	0.020256	0.006841	0.027690	0.000061	0.038445	0.044910	0.02985
studytime	-0.137857	-0.206214	-0.008415	0.062023	-0.010945	-0.008748	0.097006	0.050400	0.057176	-0.019125	...	-0.004127	-0.068829	-0.075442	-0.137585	-0.214925	-0.056433	-0.1183
failures	0.113788	0.073888	0.319968	-0.063824	-0.066068	-0.009881	-0.172210	-0.165915	-0.117882	-0.055415	...	-0.062645	0.108995	0.045078	0.105949	0.082266	0.035588	0.12277
higher	-0.136112	-0.058134	-0.265497	0.076706	0.004523	0.022726	0.213896	0.191735	0.148116	0.089929	...	0.048239	-0.102618	-0.069105	-0.131663	-0.084327	0.017290	-0.1298
internet	-0.240486	0.065911	0.013115	0.175794	0.013357	0.059754	0.266052	0.183483	0.260658	0.088625	...	0.082214	0.063268	0.092869	0.042811	0.060651	-0.022792	0.06730
Dalc	0.047169	0.282696	0.134768	-0.047304	0.060482	0.041513	-0.007018	0.000061	0.049576	0.055389	...	-0.075767	0.109904	0.245126	1.000000	0.616561	0.059067	0.17295
Walc	0.014169	0.320785	0.086357	-0.012416	0.081958	0.070976	-0.019766	0.038445	0.025657	0.044607	...	-0.093511	0.120244	0.388680	0.616561	1.000000	0.114988	0.15637
G1	-0.292626	-0.104109	-0.174322	0.157127	0.047230	0.015251	0.260472	0.217501	0.181551	0.109847	...	0.048795	-0.094497	-0.074053	-0.195171	-0.155649	-0.051647	-0.1471
G2	-0.269776	-0.104005	-0.107119	0.154600	0.038891	0.018689	0.264035	0.225139	0.153875	0.086343	...	0.089588	-0.106678	-0.079469	-0.189480	-0.164852	-0.082179	-0.1247

12 rows × 33 columns

Below is some observations base on the Correlation analysis results

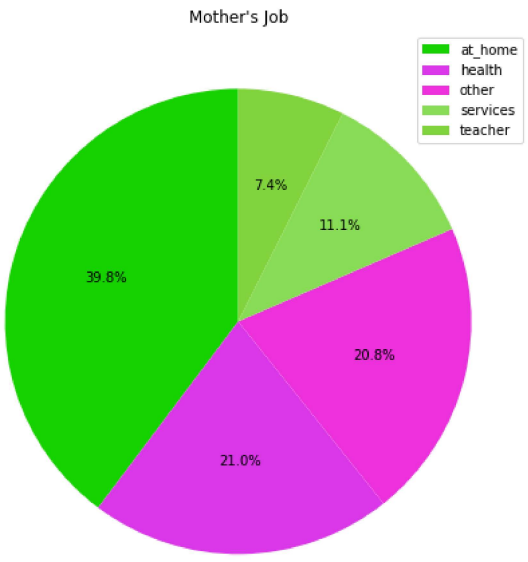
1. G3 has strong correlation with G1 or G2
2. G3 has weak correlation with failures, Medu, higher
3. G3 has no obvious correlation with sex, school or other single feature

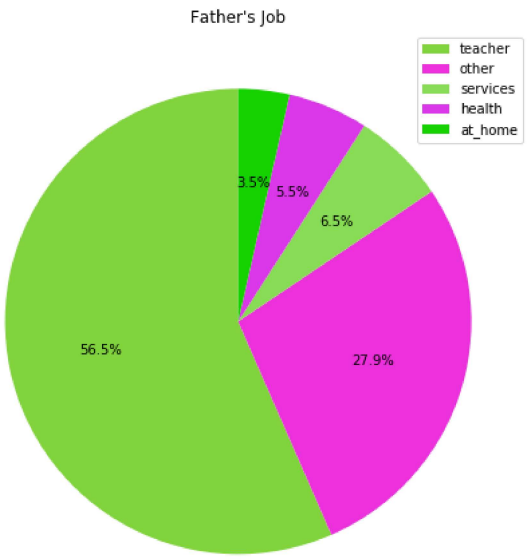
Weekend alcohol consumption and goout

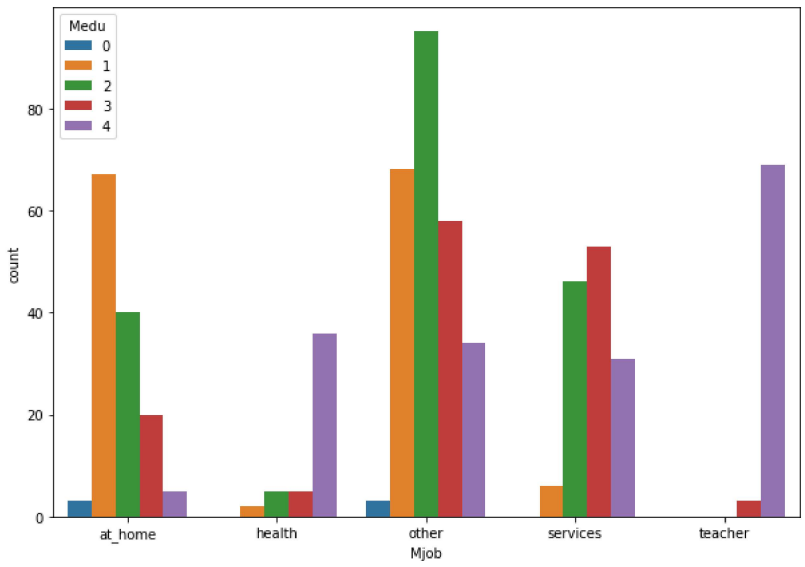


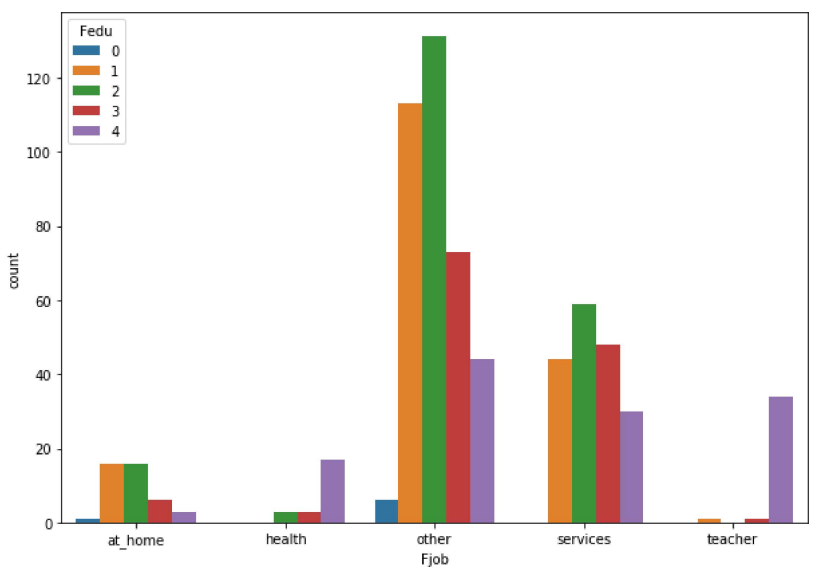
- Students who go out more will consume more alcohol.
- Most students will go out sometimes at weekend and consume alcohols

Parents' jobs and education









- Teacher jobs normally need highest education
- Mothers are more likely to be at home