None

HW3 Solutions

Note:

1. Only one possible right answer is shown. All possible right answers will be given full credit.
2. Only the final solution is shown, and the details of actual code is not shown.
3. Some of the plots in Problem-B have extra customization that is not specifically mentioned in the HW problem statement. Those customizations are for the purpose of illustration only. No credits will be deducted for extra customization.
4. You may come to the office hours or the help sessions to discuss the HW solutions.
5. If you find any typos or issues, kindly contact your section instructor, or send a text @ **smujahid** on MS teams.

Table of Contents

Problem-A

```
--------------------------------------------------Problem #A--------------------------------------------------
A-1:

To display the first 10 rows, we use df.head(10), where 10 is the argument that represents number of rows.
```

	school	gender	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	schoolsup	famsup	paid	activities	internet	health	absei
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	yes	no	no	no	no	3	
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	no	yes	no	no	yes	3	
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	yes	no	yes	no	yes	3	
3	GP	F	15	U	GT3	T	4	2	health	services	...	no	yes	yes	yes	yes	5	
4	GP	F	16	U	GT3	T	3	3	other	other	...	no	yes	yes	no	no	5	
5	GP	M	16	U	LE3	T	4	3	services	other	...	no	yes	yes	yes	yes	5	
6	GP	M	16	U	LE3	T	2	2	other	other	...	no	no	no	no	yes	3	
7	GP	F	17	U	GT3	A	4	4	other	teacher	...	yes	yes	no	no	no	1	
8	GP	M	15	U	LE3	A	3	2	services	other	...	no	yes	yes	no	yes	1	
9	GP	M	15	U	GT3	T	3	4	other	other	...	no	yes	yes	yes	yes	5	

10 rows × 25 columns

```
----------------------------------------------------------------------------------------------------
A-2:

Note: df.count(), gives total NON-null values for each column.
The number of rows in data is 395, and the number of columns in the data is 25.
Column school contains 395 non-null rows.
Column gender contains 395 non-null rows.
Column age contains 395 non-null rows.
Column address contains 395 non-null rows.
Column famsize contains 395 non-null rows.
Column Pstatus contains 395 non-null rows.
Column Medu contains 395 non-null rows.
Column Fedu contains 395 non-null rows.
Column Mjob contains 395 non-null rows.
Column Fjob contains 395 non-null rows.
Column reason contains 395 non-null rows.
Column guardian contains 395 non-null rows.
Column traveltime contains 395 non-null rows.
Column studytime contains 395 non-null rows.
Column failures contains 395 non-null rows.
Column schoolsup contains 395 non-null rows.
Column famsup contains 395 non-null rows.
Column paid contains 395 non-null rows.
Column activities contains 395 non-null rows.
Column internet contains 395 non-null rows.
Column health contains 395 non-null rows.
Column absences contains 395 non-null rows.
Column G1 contains 395 non-null rows.
Column G2 contains 395 non-null rows.
Column G3 contains 395 non-null rows.
----------------------------------------------------------------------------------------------------
A-3:

Note: df.select_dtypes(), can be used for selecting columns based on data type.
The numeric columns are: ['age', 'Medu', 'Fedu', 'traveltime', 'studytime', 'failures', 'health', 'absences',
'G1', 'G2', 'G3'].
The non-numeric columns are: ['school', 'gender', 'address', 'famsize', 'Pstatus', 'Mjob', 'Fjob', 'reason', '
guardian', 'schoolsup', 'famsup', 'paid', 'activities', 'internet'].

Statistical summaries for numeric columns:
```

	age	Medu	Fedu	traveltime	studytime	failures	health	absences	G1	G2	G3
count	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000
mean	16.696203	2.749367	2.521519	1.448101	2.035443	0.334177	3.554430	5.708861	11.664557	11.467089	11.154430
std	1.276043	1.094735	1.088201	0.697505	0.839240	0.743651	1.390303	8.003096	3.336388	3.781659	4.586196
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000	1.000000	0.000000	4.000000	0.500000	0.500000
25%	16.000000	2.000000	2.000000	1.000000	1.000000	0.000000	3.000000	0.000000	9.000000	9.500000	9.000000
50%	17.000000	3.000000	2.000000	1.000000	2.000000	0.000000	4.000000	4.000000	11.500000	11.500000	11.500000
75%	18.000000	4.000000	3.000000	2.000000	2.000000	0.000000	5.000000	8.000000	14.000000	14.000000	14.500000
max	22.000000	4.000000	4.000000	4.000000	4.000000	3.000000	5.000000	75.000000	20.000000	20.000000	21.000000

```
Statistical summaries for non-numeric columns:
```

	school	gender	address	famsize	Pstatus	Mjob	Fjob	reason	guardian	schoolsup	famsup	paid	activities	internet
count	395	395	395	395	395	395	395	395	395	395	395	395	395	395
unique	2	2	2	2	2	5	5	4	3	2	2	2	2	2
top	GP	F	U	GT3	T	other	other	course	mother	no	yes	no	yes	yes
freq	349	208	307	281	354	141	217	145	273	344	242	214	201	329

```
----------------------------------------------------------------------------------------------------
A-4:

Note: masking is an easy way to select the rows.
Following table presents the statistical summaries of all the numerical columns, whose rows are related to mal
e students.
```

	age	Medu	Fedu	traveltime	studytime	failures	health	absences	G1	G2	G3
count	187.000000	187.000000	187.000000	187.000000	187.000000	187.000000	187.000000	187.000000	187.000000	187.000000	187.000000
mean	16.657754	2.839572	2.561497	1.491979	1.764706	0.368984	3.764706	5.144385	11.954545	11.852941	11.689840
std	1.356181	1.100311	1.087670	0.750405	0.808713	0.788152	1.343337	5.980749	3.402532	3.868612	4.527699
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000	1.000000	0.000000	4.000000	0.500000	0.500000
25%	16.000000	2.000000	2.000000	1.000000	1.000000	0.000000	3.000000	0.000000	9.500000	9.500000	9.500000
50%	16.000000	3.000000	3.000000	1.000000	2.000000	0.000000	4.000000	4.000000	11.500000	12.000000	12.000000
75%	18.000000	4.000000	3.500000	2.000000	2.000000	0.000000	5.000000	8.000000	14.500000	14.500000	14.500000
max	22.000000	4.000000	4.000000	4.000000	4.000000	3.000000	5.000000	38.000000	19.500000	20.000000	21.000000

```
----------------------------------------------------------------------------------------------------
A-5:

Note: masking is an easy way to select the rows.
Following table presents the statistical summaries of all the non-numerical columns, where students age is eit
her 15, 17 or 19.
```

	school	gender	address	famsize	Pstatus	Mjob	Fjob	reason	guardian	schoolsup	famsup	paid	activities	internet
count	204	204	204	204	204	204	204	204	204	204	204	204	204	204
unique	2	2	2	2	2	5	5	4	3	2	2	2	2	2
top	GP	F	U	GT3	T	other	other	course	mother	no	yes	no	no	yes
freq	186	110	155	148	180	69	104	77	131	175	130	113	104	170

```
----------------------------------------------------------------------------------------------------
A-6:

Note: masking is an easy way to select the rows.
The following table presents the summary statistics of column "famsize" for those students who score above ave
rage in
columns G1, G2 and G3, respectively.
count     163
unique      2
top       GT3
freq      113
Name: famsize, dtype: object
----------------------------------------------------------------------------------------------------
A-7:

The condition is same as mother's job is teachers, health or services; or father's job is teachers, health or
services (or both).
The number of students whose one(or both) of the parents job is categorized as teacher, health, or services is
246.
```

```
----------------------------------------------------------------------------------------------------
A-8:
The required first 10 students in ascending order of G3:
```

	school	gender	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	schoolsup	famsup	paid	activities	internet	health	al
341	GP	M	18	U	GT3	T	4	4	teacher	services	...	no	yes	no	yes	yes	2	
140	GP	M	15	U	GT3	T	4	3	teacher	services	...	yes	yes	no	no	yes	3	
242	GP	M	16	U	LE3	T	4	3	teacher	other	...	no	no	no	yes	yes	3	
148	GP	M	16	U	GT3	T	4	4	teacher	teacher	...	no	yes	no	no	yes	5	
134	GP	M	15	R	GT3	T	3	4	at_home	teacher	...	no	yes	no	no	no	5	
130	GP	F	15	R	GT3	T	3	4	services	teacher	...	no	yes	no	no	yes	5	
386	MS	F	18	R	GT3	T	4	4	teacher	at_home	...	no	yes	yes	yes	yes	5	
209	GP	F	17	R	GT3	T	4	3	teacher	other	...	no	yes	yes	yes	yes	4	
49	GP	F	15	U	GT3	T	4	4	services	teacher	...	yes	yes	no	yes	yes	3	
180	GP	M	16	U	GT3	T	4	3	teacher	other	...	no	yes	yes	yes	yes	3	

10 rows × 25 columns

```
Note: Since it is not mentioned, the sorting can be ascending or descending order.
----------------------------------------------------------------------------------------------------
A-9:
```

For applying lambda or custom function on every element of a column use df[column].apply() method.

For applying lambda or custom function on every element of more than one column use df[columns].applymap() method.

```
The required summary statistics of columns G1, G2 and G3 are:
```

	G1	G2	G3
count	395.000000	395.000000	395.000000
mean	11.915190	11.720253	11.388608
std	3.368226	3.818150	4.600016
min	4.500000	0.500000	0.500000
25%	9.500000	9.500000	9.500000
50%	11.500000	11.500000	11.500000
75%	14.500000	14.500000	14.500000
max	20.500000	20.500000	21.500000

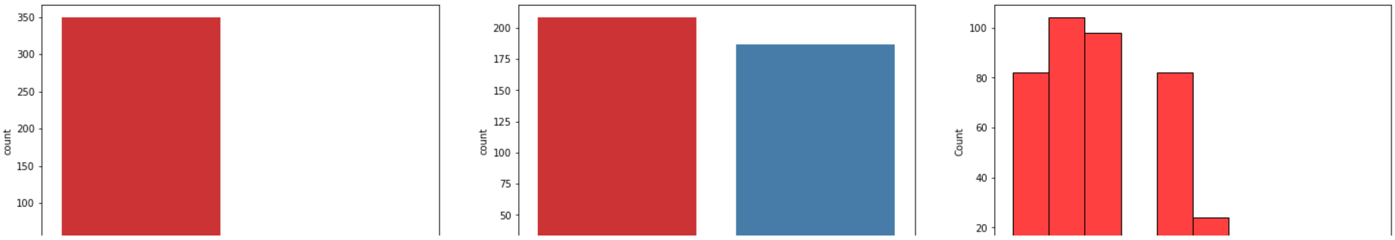
```
----------------------------------------------------------------------------------------------------
A-10:
```

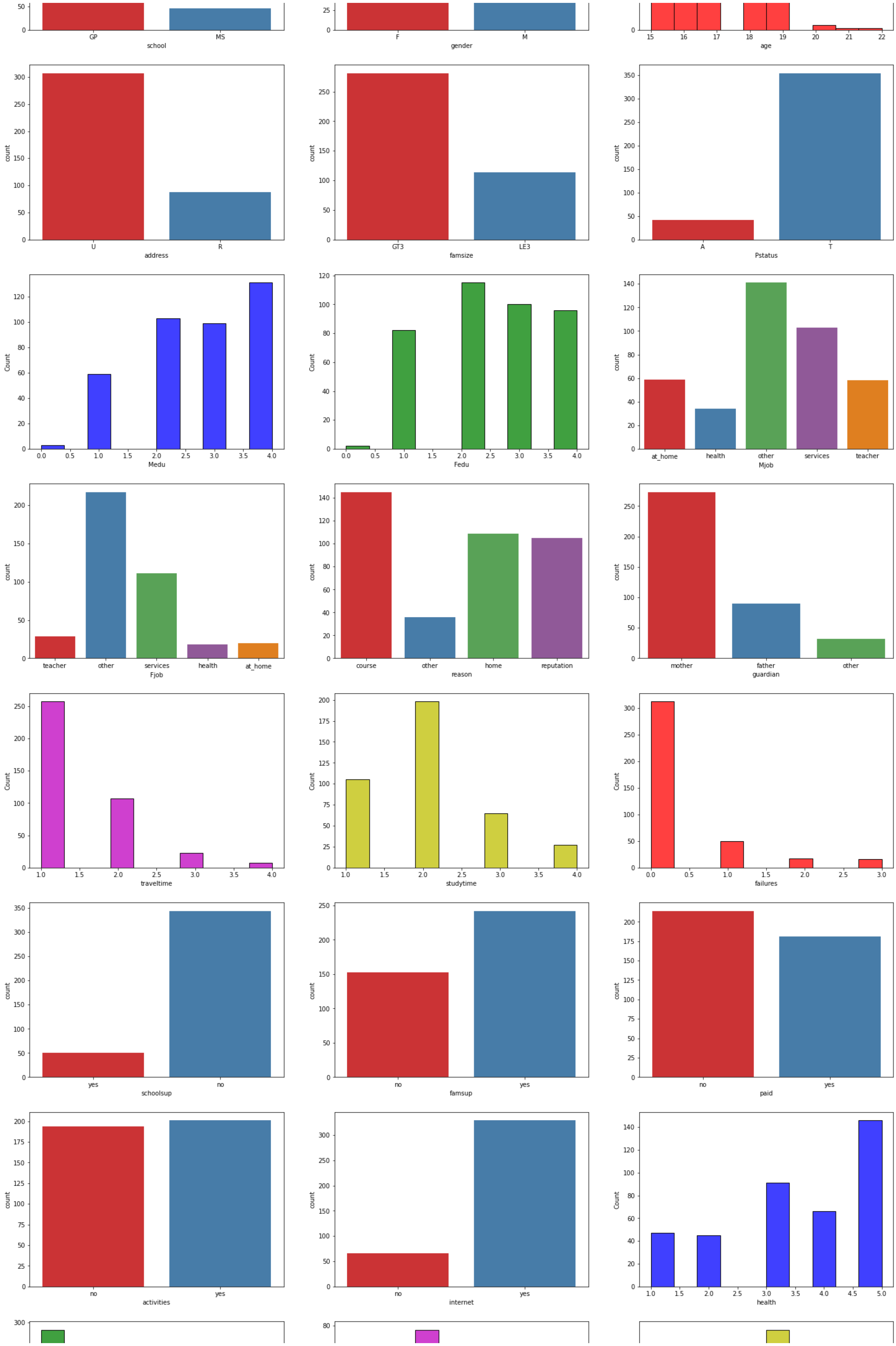
For applying lambda or custom function that take the entire row as input, use df.apply() method, with parameter axis=1.

```
The required value counts for the advising column are:
normal       329
follow-up     58
concern        6
medical        2
Name: advising, dtype: int64
~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~
```

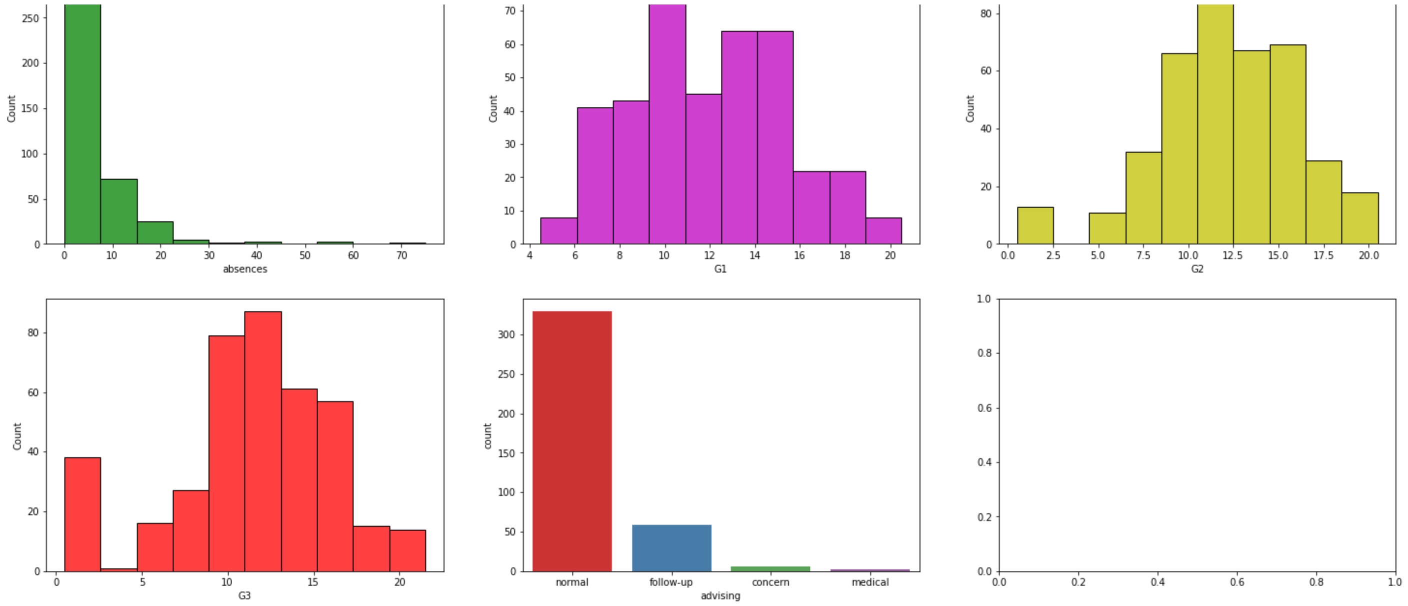
Problem-B

```
--------------------------------------------------Problem #B--------------------------------------------------
B-1: The histograms for all numeric and nonnumeric columns are as follows (for numeric columns 10 bins are taken).
```

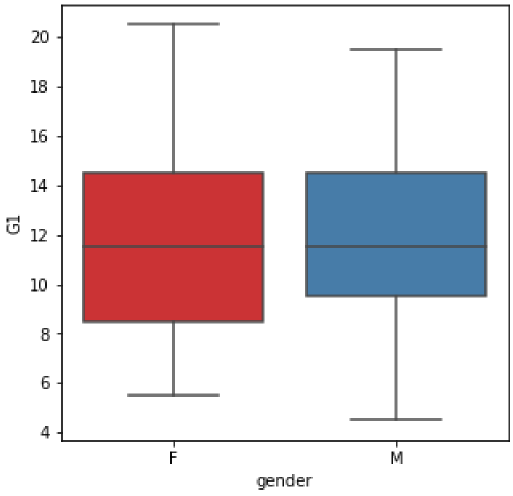




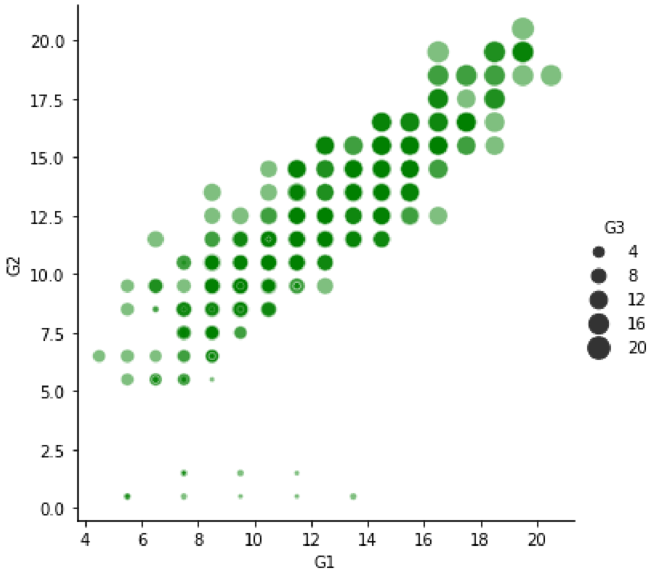
0
50
GP
MS
school
0
25
F
M
gender
0
15
16
17
18
19
20
21
22
age
count
0
50
100
150
200
250
300
U
R
address
count
0
50
100
150
200
250
GT3
LE3
famsize
count
0
50
100
150
200
250
300
350
A
T
Pstatus
Count
0
20
40
60
80
100
120
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Medu
Count
0
20
40
60
80
100
120
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Fedu
count
0
20
40
60
80
100
120
140
at_home
health
other
services
teacher
Mjob
count
0
50
100
150
200
teacher
other
services
health
at_home
Fjob
count
0
20
40
60
80
100
120
140
course
other
home
reputation
reason
count
0
50
100
150
200
250
mother
father
other
guardian
Count
0
50
100
150
200
250
1.0
1.5
2.0
2.5
3.0
3.5
4.0
traveltime
Count
0
25
50
75
100
125
150
175
200
1.0
1.5
2.0
2.5
3.0
3.5
4.0
studytime
Count
0
50
100
150
200
250
300
0.0
0.5
1.0
1.5
2.0
2.5
3.0
failures
count
0
50
100
150
200
250
300
350
yes
no
schoolsup
count
0
50
100
150
200
250
no
yes
famsup
count
0
25
50
75
100
125
150
175
200
no
yes
paid
count
0
25
50
75
100
125
150
175
200
no
yes
activities
count
0
50
100
150
200
250
300
no
yes
internet
Count
0
20
40
60
80
100
120
140
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
health
300
80



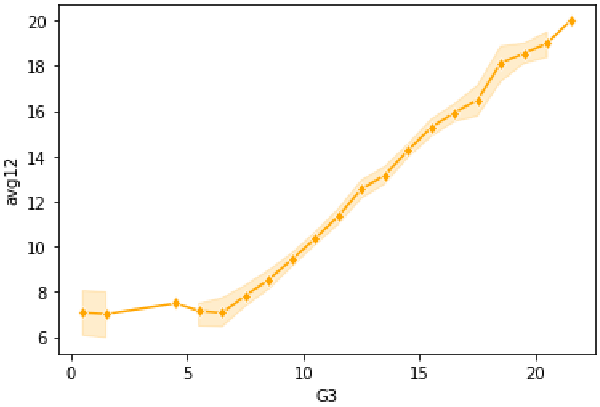
```
----------------------------------------------------------------------------------------------------
B-2:
The box plots for G1 differentiated by gender.
```



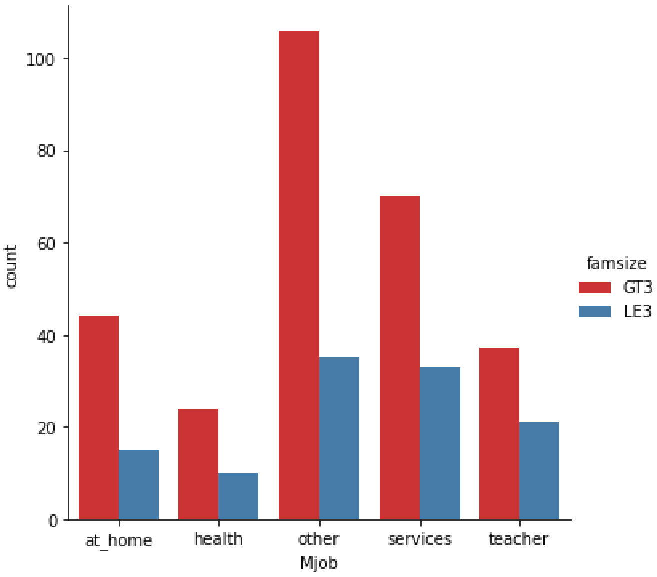
```
----------------------------------------------------------------------------------------------------
B-3:
The scatter plot between columns G1 and G2, where the size of the marker is based on column G3.
```



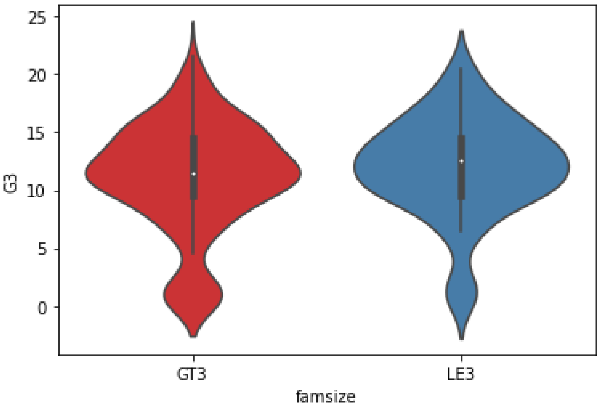
B-4:
The plot of G3 in ascending order on the x-axis, and the corresponding average of G1 and G2 on the y-axis.



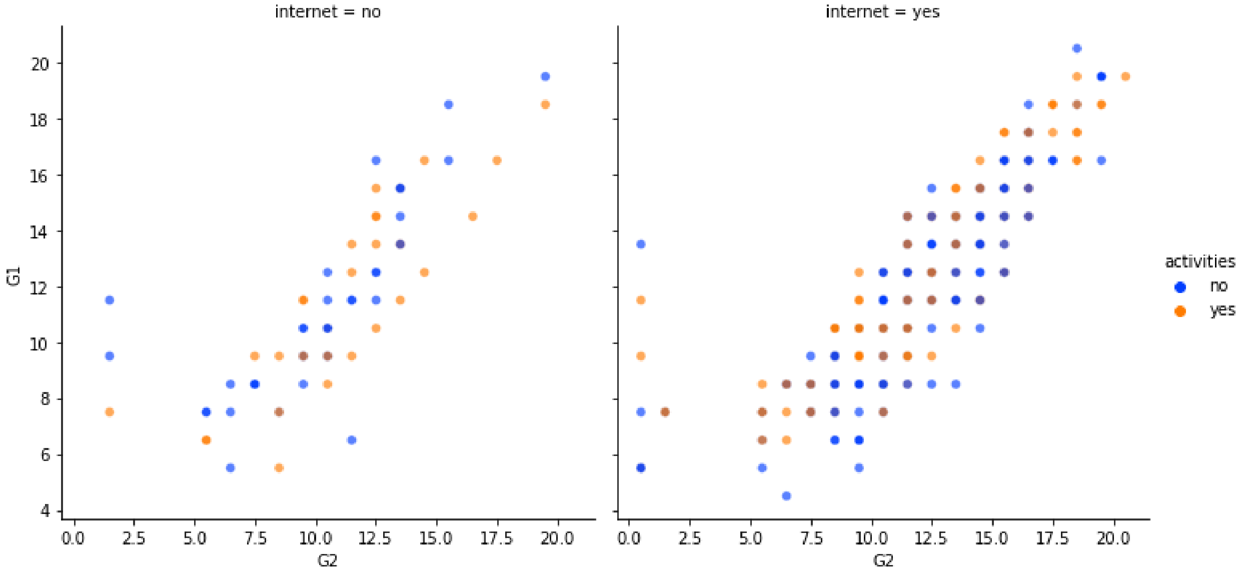
B-5:
The count plots of Mjob, differentiated by famsize.



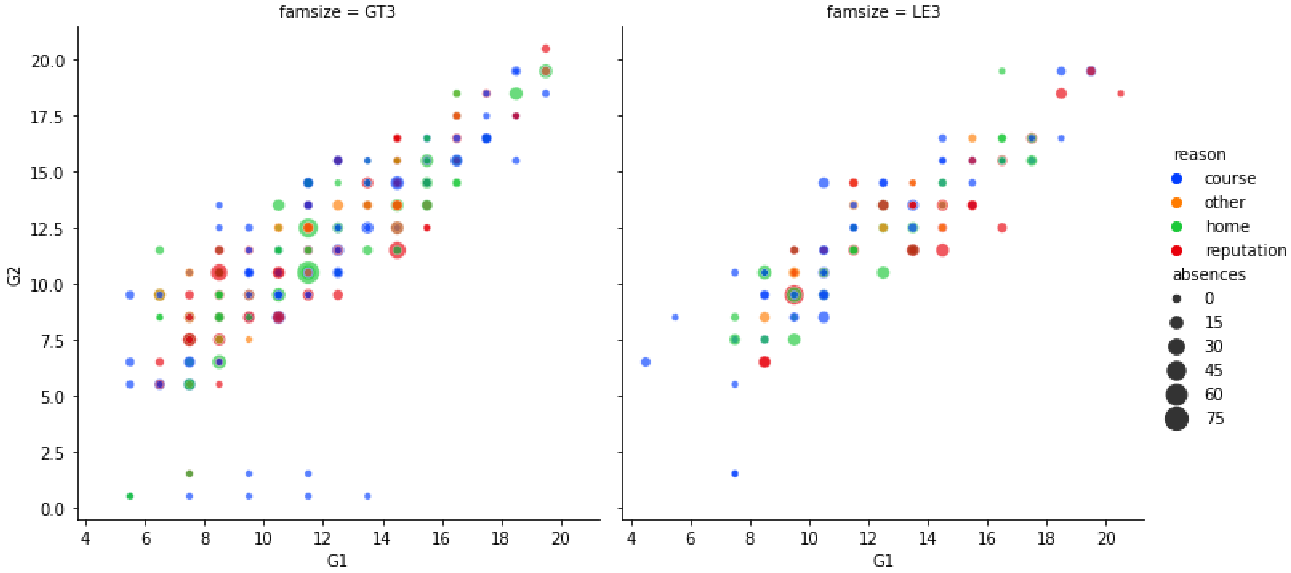
B-6:
The violin-plots of G3, differentiated by famsize.



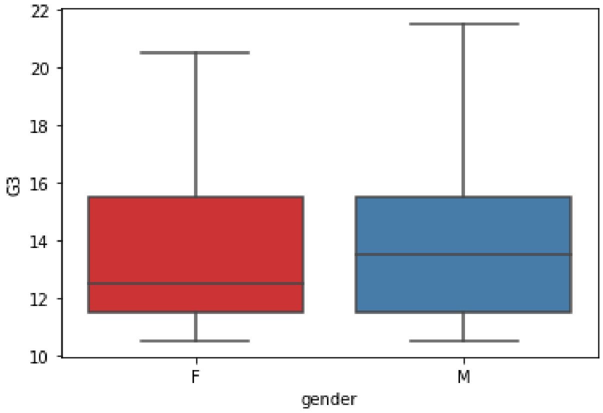
```
------------------------------------------------------------------------------------------------------------------------
B-7:
The plot of G1 vs G2 differentiated by activities and internet.
```



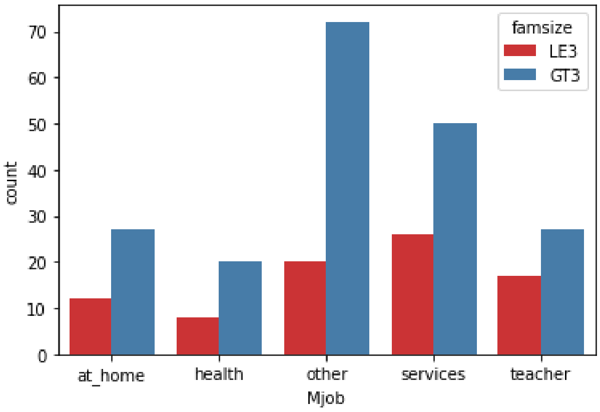
```
------------------------------------------------------------------------------------------------------------------------
B-8:
The required plot of G1 and G2.
```



```
------------------------------------------------------------------------------------------------------------------------
B-9:
The required box plot for G3.
```



```
----------------------------------------------------------------------------------------------------
B-10:
The required count plots of Mjob.
```



```
-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~~
```