ĐỒ ÁN PHÂN TÍCH THỰC NGHIỆM

DS304.K21 GVHD: Đỗ Trọng Hợp

Võ Hoàng Thông
Đỗ Hùng Dũng
Nguyễn Xuân Vinh
Nguyễn Tấn Phong
Nguyễn Thị Thanh Kim

Data analysis

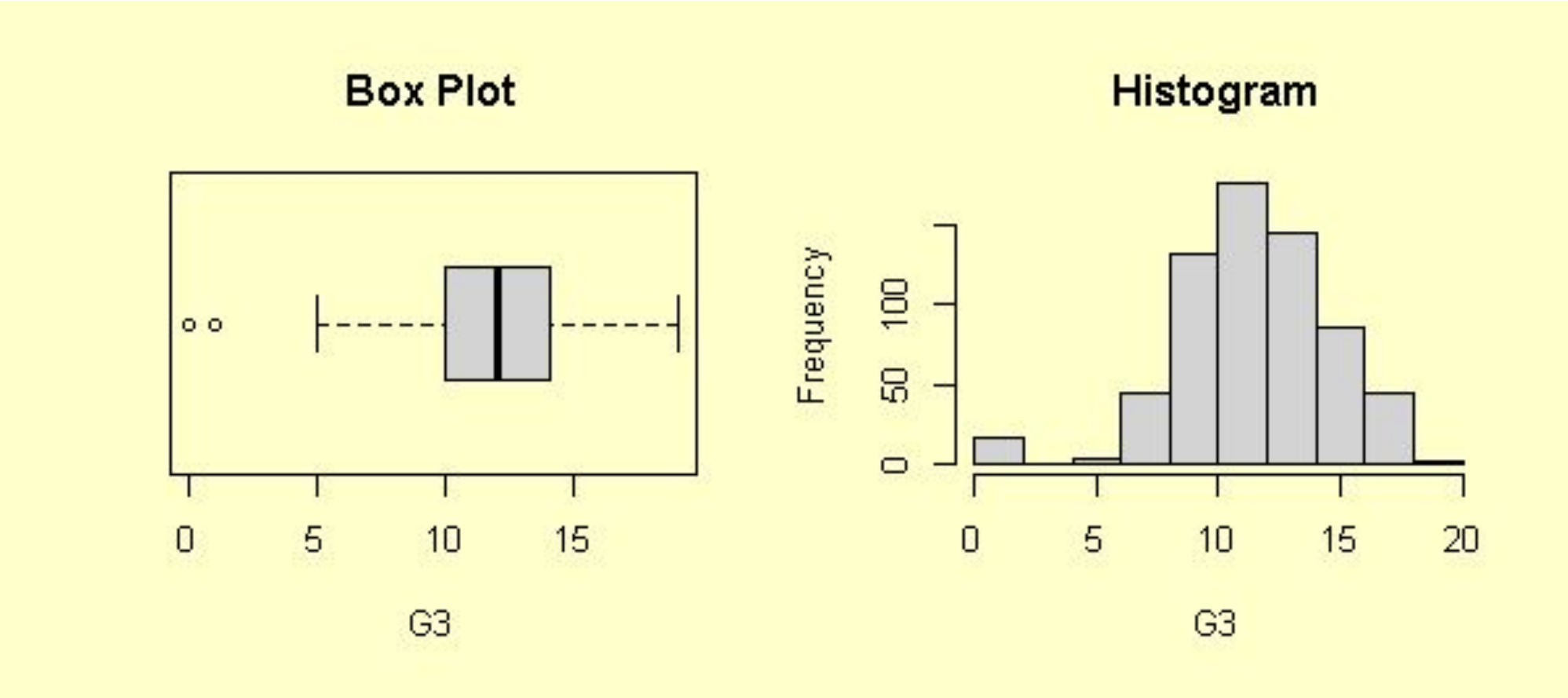
Bộ dữ liệu được sử dụng là Student Alcohol Consumption dựa trên dữ liệu được thu thập ở hai trường trung học ở Bồ Đào Nha. Các học sinh/sinh viên tham gia đến từ lớp Toán và tiếng Bồ Đào Nha. Số lượng học sinh lớp Toán tham gia khảo sát là 395, trong khi 649 học sinh/sinh viên tiếng Bồ Đào Nha tham gia khảo sát. Dữ liệu được thu thập ở hai địa điểm là Gabriel Pereira và Mousinho da Silveira bao gồm những thông tin liên quan như thông tin sơ bộ, nhân khẩu học và mức độ tiêu thụ rượu.

Bộ dữ liệu mà đồ án này sẽ sử dụng chỉ bao gồm dữ liệu chứa thông tin các học sinh/sinh viên lớp tiếng Bồ Đào Nha

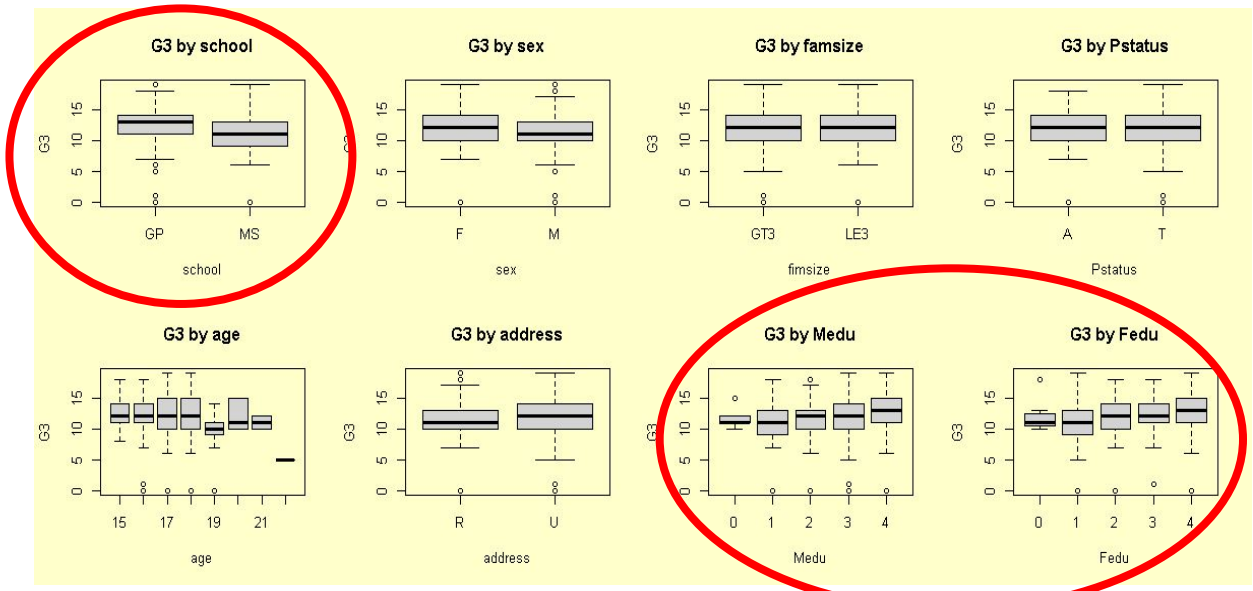
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
1	school	sex	age	addres	famsiz	Pstatu	Medu	Fedu	Mjob	Fjob	reasor	guardi	travelt	studyt	failure	school	famsu	paid	activit	nurser	higher	intern	roman	famrel	freetin	goout	Dalc	Walc	health	absen	G1	G2	G3
2	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mothe	2	2	0	yes	no	no	no	yes	yes	no	no	4	3	4	1	1	3	4	0	11	11
3	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes	no	no	no	yes	yes	no	5	3	3	1	1	3	2	9	11	11
4	GP	F	15	U	LE3	T	1	1	at_home	other	other	mothe	1	2	0	yes	no	no	no	yes	yes	yes	no	4	3	2	2	3	3	6	12	13	12
5	GP	F	15	U	GT3	T	4	2	health	services	home	mothe	1	3	0	no	yes	no	yes	yes	yes	yes	yes	3	2	2	1	1	5	0	14	14	14
6	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes	no	no	yes	yes	no	no	4	3	2	1	2	5	0	11	13	13
7	GP	M	16	U	LE3	T	4	3	services	other	reputa	mothe	1	2	0	no	yes	no	yes	yes	yes	yes	no	5	4	2	1	2	5	6	12	12	13
8	GP	M	16	U	LE3	T	2	2	other	other	home	mothe	1	2	0	no	no	no	no	yes	yes	yes	no	4	4	4	1	1	3	0	13	12	13
9	GP	F	17	U	GT3	A	4	4	other	teacher	home	mothe	2	2	0	yes	yes	no	no	yes	yes	no	no	4	1	4	1	1	1	2	10	13	13
10	GP	M	15	U	LE3	A	3	2	services	other	home	mothe	1	2	0	no	yes	no	no	yes	yes	yes	no	4	2	2	1	1	1	0	15	16	17
11	GP	M	15	U	GT3	T	3	4	other	other	home	mothe	1	2	0	no	yes	no	yes	yes	yes	yes	no	5	5	1	1	1	5	0	12	12	13
12	GP	F	15	U	GT3	T	4	4	teacher	health	reputa	mothe	1	2	0	no	yes	no	no	yes	yes	yes	no	3	3	3	1	2	2	2	14	14	14
13	GP	F	15	U	GT3	T	2	1	services	other	reputa	father	3	3	0	no	yes	no	yes	yes	yes	yes	no	5	2	2	1	1	4	0	10	12	13
14	GP	M	15	U	LE3	T	4	4	health	services	course	father	1	1	0	no	yes	no	yes	yes	yes	yes	no	4	3	3	1	3	5	0	12	13	12
15	GP	M	15	U	GT3	T	4	3	teacher	other	course	mothe	2	2	0	no	yes	no	no	yes	yes	yes	no	5	4	3	1	2	3	0	12	12	13
16	GP	M	15	U	GT3	A	2	2	other	other	home	other	1	3	0	no	yes	no	no	yes	yes	yes	yes	4	5	2	1	1	3	0	14	14	15
17	GP	F	16	U	GT3	T	4	4	health	other	home	mothe	1	1	0	no	yes	no	no	yes	yes	yes	no	4	4	4	1	2	2	6	17	17	17
18	GP	F	16	U	GT3	T	4	4	services	services	reputa	mothe	1	3	0	no	yes	no	yes	yes	yes	yes	no	3	2	3	1	2	2	10	13	13	14
19	GP	F	16	U	GT3	T	3	3	other	other	reputa	mothe	3	2	0	yes	yes	no	yes	yes	yes	no	no	5	3	2	1	1	4	2	13	14	14
20	GP	M	17	U	GT3	T	3	2	services	services	course	mothe	1	1	3	no	yes	yes	yes	yes	yes	yes	no	5	5	5	2	4	5	2	8	8	7
21	GP	M	16	U	LE3	T	4	3	health	other	home	father	1	1	0	no	no	no	yes	yes	yes	yes	no	3	1	3	1	3	5	6	12	12	12
22	GP	M	15	U	GT3	T	4	3	teacher	other	reputa	mothe	1	2	0	no	no	no	no	yes	yes	yes	no	4	4	1	1	1	1	0	12	13	14
23	GP	M	15	U	GT3	T	4	4	health	health	other	father	1	1	0	no	yes	yes	no	yes	yes	yes	no	5	4	2	1	1	5	0	11	12	12
24	GP	M	16	U	LE3	T	4	2	teacher	other	course	mothe	1	2	0	no	no	no	yes	yes	yes	yes	no	4	5	1	1	3	5	0	12	13	14
25	GP	M	16	U	LE3	T	2	2	other	other	reputa	mothe	2	2	0	no	yes	no	yes	yes	yes	yes	no	5	4	4	2	4	5	2	10	10	10
26	GP	F	15	R	GT3	T	2	4	services	health	course	mothe	1	3	0	yes	yes	no	yes	yes	yes	yes	no	4	3	2	1	1	5	2	10	11	10
27	GP	F	16	U	GT3	T	2	2	services	services	home	mothe	1	1	0	no	yes	no	no	no	yes	yes	no	1	2	2	1	3	5	6	10	11	12
28	GP	M	15	U	GT3	T	2	2	other	other	home	mothe	1	1	0	no	yes	no	no	yes	yes	yes	no	4	2	2	1	2	5	8	11	12	12
29	GP	M	15	U	GT3	T	4	2	health	services	other	mothe	1	1	0	no	no	no	no	yes	yes	yes	no	2	2	4	2	4	1	0	11	11	11
30	GP	M	16	U	LE3	A	3	4	services	other	home	mothe	1	2	0	yes	yes	yes	yes	yes	yes	yes	no	5	3	3	1	1	5	2	12	12	13
31	GP	M	16	U	GT3	T	4	4	teacher	teacher	home	mothe	1	2	0	no	yes	yes	yes	yes	yes	yes	yes	4	4	5	5	5	5	4	12	11	12
32	GP	M	15	U	GT3	T	4	4	health	services	home	mothe	1	2	0	no	yes	yes	no	no	yes	yes	no	5	4	2	3	4	5	0	10	11	11
33	GP	M	15	U	GT3	T	4	4	services	services	reputa	mothe	2	2	0	no	yes	no	yes	yes	yes	yes	no	4	3	1	1	1	5	2	15	15	15

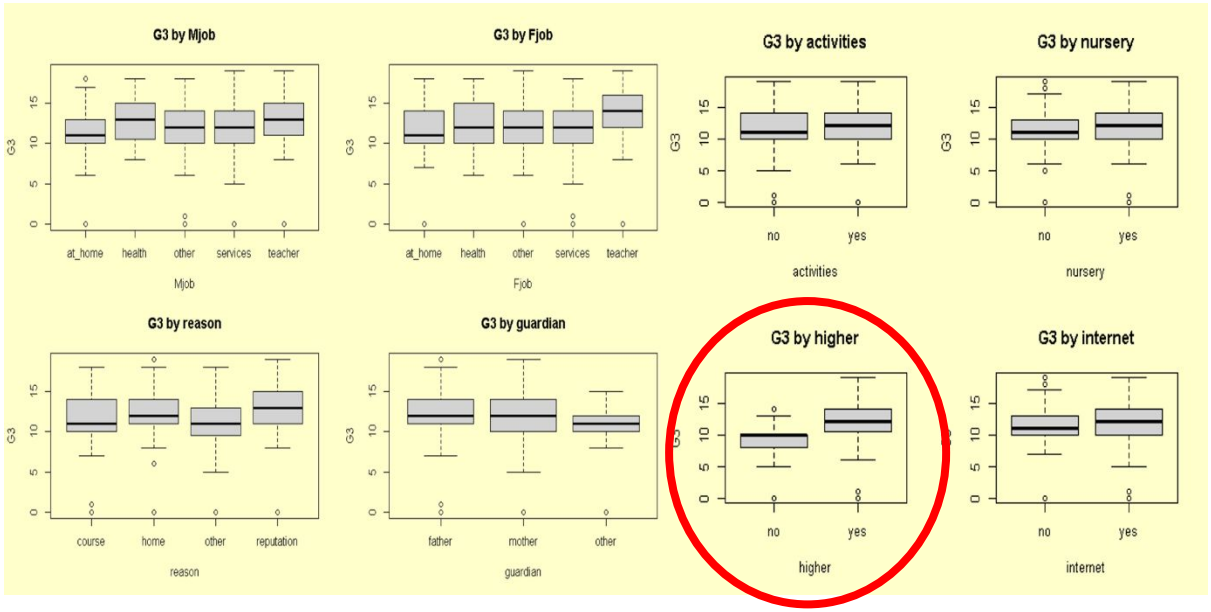
33 thuộc tính

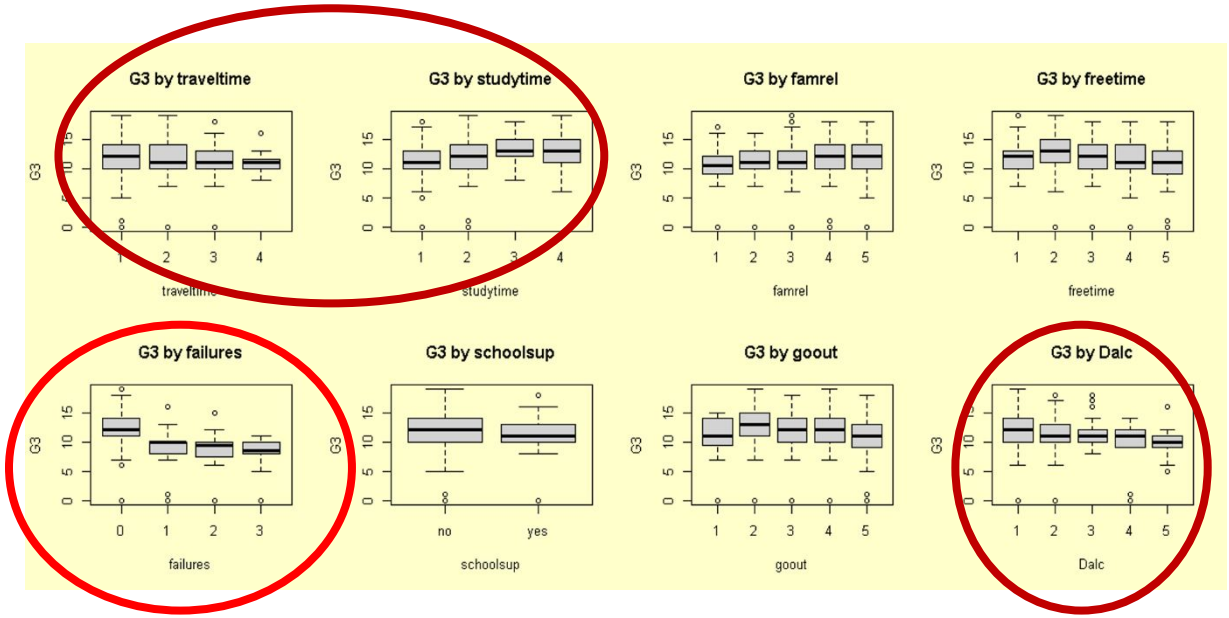
Sex Age School Address famsize Pstatus Medu Fedu Mjob Fjob reason guardian traveltime studytime failures schoolsup famsup activities paid Internet nursery higher romantic famrel freetime goout Dalc Walc health absences G1 G2 G3

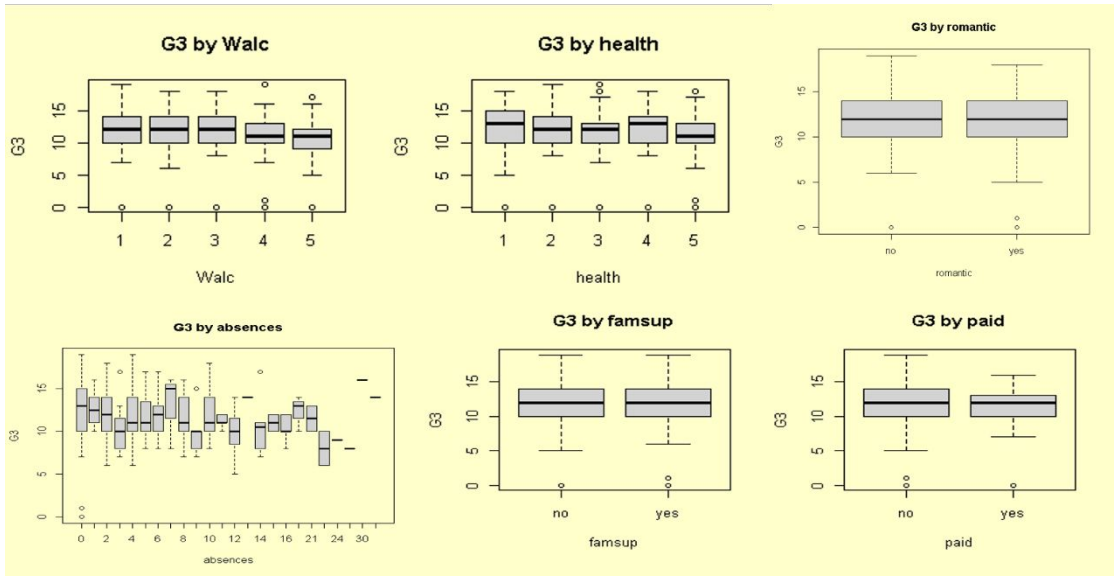


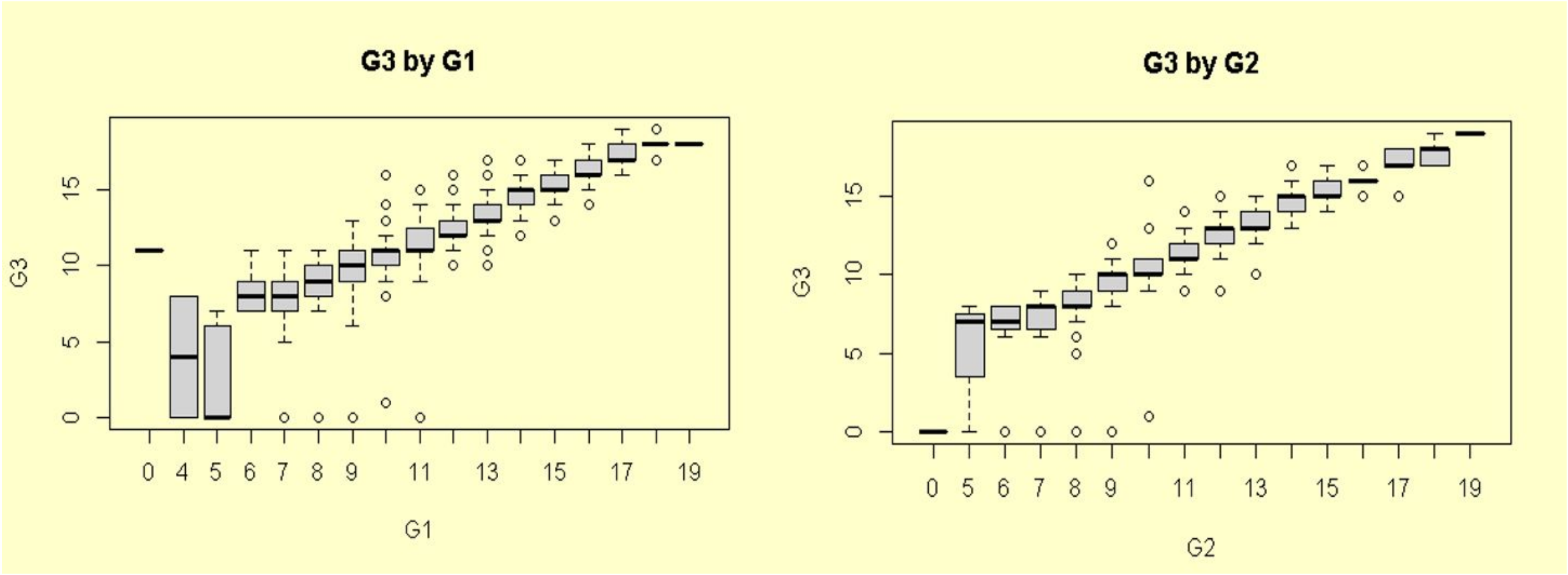
Trực quan hóa dữ liệu G3


G3 by school
G3
0 5 10 15
GP MS
school
G3 by sex
F M
sex
G3 by famsize
GT3 LE3
famsize
G3 by Pstatus
A T
Pstatus
G3 by age
15 17 19 21
age
G3 by address
R U
address
G3 by Medu
0 1 2 3 4
Medu
G3 by Fedu
0 1 2 3 4
Fedu


G3 by Mjob
at_home health other services teacher
Mjob
G3 by Fjob
at_home health other services teacher
Fjob
G3 by activities
no yes
activities
G3 by nursery
no yes
nursery
G3 by reason
course home other reputation
reason
G3 by guardian
father mother other
guardian
G3 by higher
no yes
higher
G3 by internet
no yes
internet


G3 by traveltime
1 2 3 4
traveltime
G3 by studytime
1 2 3 4
studytime
G3 by famrel
1 2 3 4 5
famrel
G3 by freetime
1 2 3 4 5
freetime
G3 by failures
0 1 2 3
failures
G3 by schoolsup
no yes
schoolsup
G3 by goout
1 2 3 4 5
goout
G3 by Dalc
1 2 3 4 5
Dalc


G3 by Walc
1 2 3 4 5
Walc
G3 by health
1 2 3 4 5
health
G3 by romantic
no yes
romantic
G3 by absences
0 2 4 6 8 10 12 14 16 21 24 30
absences
G3 by famsup
no yes
famsup
G3 by paid
no yes
paid



Phân phối dữ liệu G3 bởi G1 và G2

One-way ANOVA

Kết quả kiểm tra bằng bảng ANOVA cho thấy hầu hết các yếu tố có ý nghĩa thống kê với mức ý nghĩa 5%. T

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
school	1	547	546.6	56.89	1.57e-13	***
sex	1	113	112.68	10.96	0.000982	***
age	7	262	37.40	3.687	0.000644	***
address	1	190	190.06	18.71	1.76e-05	***
famsize	1	14	13.71	1.314	0.252	
Pstatus	1	0	0.004	0	0.985	
Medu	1	390	390.1	39.6	5.75e-10	***
Fedu	4	329	82.16	8.223	1.8e-06	***
Mjob	4	296	74.01	7.37	8.31e-06	***
Fjob	4	135	33.68	3.273	8.31e-06	*
reason	3	308	102.57	10.25	1.34e-06	***
guardian	2	55	27.40	2.638	0.0723	.
traveltime	3	113	37.72	3.659	0.0123	*
studytime	3	465	155.03	15.88	5.71e-10	***
failures	3	1305	434.9	51.39	<2e-16	***
famsup	1	24	23.71	2.276	0.132	
paid	1	20	20.38	1.956	0.162	
activities	1	24	24.18	2.321	0.128	
nursery	1	6	5.591	0.535	0.465	
higher	1	746	746.2	80.24	<2e-16	***
internet	1	152	152.22	14.9	0.000125	***
romantic	1	55	55.49	5.353	0.021	*
famrel	4	151	37.76	3.678	0.00568	**
freetime	4	183	45.86	4.489	0.00139	**
goout	4	281	70.32	6.986	1.65e-05	***
Dalc	4	328	81.88	8.193	1.9e-06	***
Walc	4	225	56.20	5.535	0.000219	***
health	4	101	25.25	2.44	0.0457	*
absences	23	248	10.76	1.032	0.421	
G1	16	4821	301.30	98.03	<2e-16	***
G2	15	5762	384.1	242.7	<2e-16	***

Two-way ANOVA

Từ bảng ANOVA trên, ta có thể thấy ngoài các yếu tố chính ra, còn có những interaction 2 yếu tố ảnh hưởng tới G3.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
sex	1	112.7	112.7	114.421	< 2e-16	***
Medu	1	448.0	448.0	454.870	< 2e-16	***
studytime	3	286.8	95.6	97.081	< 2e-16	***
failures	3	895.7	298.6	303.188	< 2e-16	***
higher	1	165.2	165.2	167.799	< 2e-16	***
G1	16	3008.6	188.0	190.938	< 2e-16	***
G2	15	925.0	61.7	62.619	< 2e-16	***
sex:Medu	1	0.3	0.3	0.299	0.585058	
sex:studytime	3	3.2	1.1	1.072	0.360711	
sex:failures	3	14.4	4.8	4.887	0.002405	**
sex:higher	1	3.0	3.0	3.030	0.082553	.
sex:G1	13	40.9	3.1	3.198	0.000138	***
sex:G2	13	80.8	6.2	6.309	8.47e-11	***
Medu:studytime	3	1.0	0.3	0.330	0.803428	
Medu:failures	3	9.9	3.3	3.339	0.019437	*
Medu:higher	1	2.6	2.6	2.659	0.103777	
Medu:G1	13	22.7	1.7	1.776	0.044932	*
Medu:G2	14	47.6	3.4	3.451	2.63e-05	***
studytime:failures	6	14.9	2.5	2.527	0.020703	*
studytime:higher	3	11.3	3.8	3.839	0.009938	**
studytime:G1	33	19.0	0.6	0.585	0.968968	
studytime:G2	31	28.1	0.9	0.921	0.591589	
failures:higher	3	6.3	2.1	2.134	0.095439	.
failures:G1	19	88.4	4.7	4.722	7.56e-10	***
failures:G2	11	52.6	4.8	4.858	4.95e-07	***
higher:G1	6	7.3	1.2	1.242	0.283808	
higher:G2	6	27.1	4.5	4.585	0.000165	***
G1:G2	45	68.4	1.5	1.544	0.017145	*
Residuals	377	371.3	1.0			