GEORGIA STATE UNIVERSITY

Department of Mathematics and Statistics

Investigating Causes of Student Performance

by

Jesse Annan

STAT 8670: SAS Programming and Data Analysis
Instructor: Li-Hsiang Lin PhD

2024

Contents

Introduction

As a pivotal stage in academic development, secondary education lays the foundation for future endeavors and significantly shapes students' future plans. It not only equips students with essential knowledge and skills but also fosters critical thinking, problem-solving abilities, and social-emotional growth. As such, the quality of secondary education profoundly influences individuals' academic trajectories, career opportunities, and overall life outcomes. Understanding the intricacies of student achievement is essential for educators, policymakers, and stakeholders alike to implement targeted interventions and support mechanisms. Therefore, this project seeks to construct a simple yet robust model that not only identifies key predictors of academic achievement but also offers actionable recommendations for enhancing educational outcomes.

1.1 Dataset

The dataset used in this project was sourced from UC Irvine Machine Learning Repository. Below is a list of all the features in the dataset, as well as the varaible type of each feature. See Table 3.2 for a complete description of each feature:

Table 1.1: Variable Datatypes

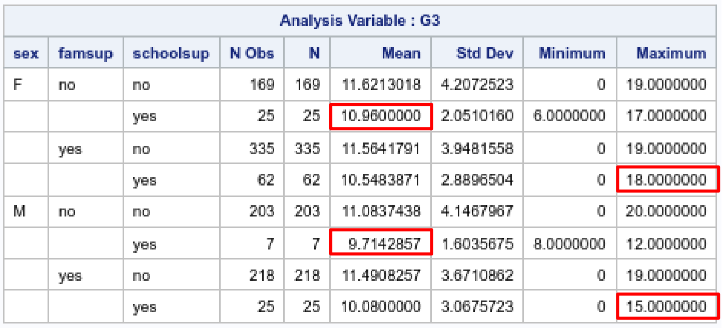
Variable	**Type**	**Variable**	**Type**	**Variable**	**Type**
school	binary	sex	binary	age	numeric
address	binary	famsize	binary	Pstatus	binary
Medu	numeric	Fedu	numeric	Mjob	nominal
Fjob	nominal	reason	nominal	guardian	nominal
traveltime	numeric	studytime	numeric	failures	numeric
schoolsup	binary	famsup	binary	paid	binary
activities	binary	nursery	binary	higher	binary
internet	binary	romantic	binary	famrel	numeric
freetime	numeric	goout	numeric	Dalc	numeric
Walc	numeric	health	numeric	absences	numeric
G1	numeric	G2	numeric	G3	numeric

1.2 Research Statement

This study utilizes datasets from two Portuguese schools, containing student grades, demographic, social, and school-related features. The data, collected through school reports and questionnaires, is made up of performance in two distinct subjects: Mathematics (math) and Portuguese language (port). This research aims to explore how variables such as absences, family and school support, and guardian type influence student performance, shedding light on strategies to enhance educational outcomes.

Exploratory Data Analysis (EDA)

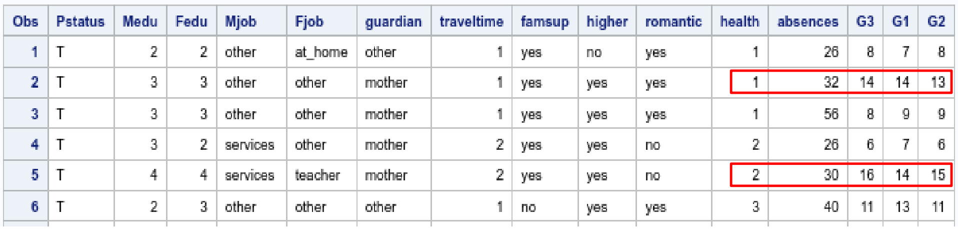
Exploratory Data Analysis (EDA) serves as a crucial preliminary step in understanding the underlying patterns, trends, and relationships within a dataset. In this chapter, we conduct a comprehensive exploration of the dataset, aiming to uncover key factors influencing student performance in secondary education. We begin our analysis by examining the demographic composition of the student population. The distribution of students by gender shows that there are more females than males. Additionally, we investigate the impact of family and school support on student performance, while considering gender variability. The frequency distribution and statistical summaries shed light on the distribution of final grades (G3) among the different subgroups. Notably, we observe differences in performance based on the presence of family and educational support, with females demonstrating higher performance levels, particularly in supportive environments.



Analysis Variable : G3

sex	famsup	schoolsup	N Obs	N	Mean	Std Dev	Minimum	Maximum
F	no	no	169	169	11.6213018	4.2072523	0	19.0000000
		yes	25	25	10.9600000	2.0510160	6.0000000	17.0000000
	yes	no	335	335	11.5641791	3.9481558	0	19.0000000
		yes	62	62	10.5483871	2.8896504	0	18.0000000
M	no	no	203	203	11.0837438	4.1467967	0	20.0000000
		yes	7	7	9.7142857	1.6035675	8.0000000	12.0000000
	yes	no	218	218	11.4908257	3.6710862	0	19.0000000
		yes	25	25	10.0800000	3.0675723	0	15.0000000

Next, we examine the variability of final grades across different schools, aiming to identify disparities and potential drivers of academic achievement. Through frequency analysis and graphical representations, we observe that Gabriel Pereira (GP) students perform better compared to Mousinho da Silveira's (MS) students. It can also be seen that relatively, GP has a larger number of supported students (107) compared to MS (12), which may indicate a greater emphasis on support services or resources at GP. Further investigation focuses on understanding the factors contributing to extreme absences among students and their implications for academic performance. By isolating students with excessive absences (25 or more), we analyze parental status, family support, and other socio-demographic attributes to observe potential causal factors. The observed patterns underscore the critical role of parental support and presence in mitigating absenteeism and fostering student success. Specifically, Among the students with the most absences, only five appear to have extreme health problems with health=1 or 2. Surprisingly, two students performed better than the average student.



Obs	Pstatus	Medu	Fedu	Mjob	Fjob	guardian	traveltime	famsup	higher	romantic	health	absences	G3	G1	G2
1	T	2	2	other	at_home	other	1	yes	no	yes	1	26	8	7	8
2	T	3	3	other	other	mother	1	yes	yes	yes	1	32	14	14	13
3	T	3	3	other	other	mother	1	yes	yes	yes	1	56	8	9	9
4	T	3	2	services	other	mother	2	yes	yes	no	2	26	6	7	6
5	T	4	4	services	teacher	mother	2	yes	yes	no	2	30	16	14	15
6	T	2	3	other	other	other	1	no	yes	yes	3	40	11	13	11

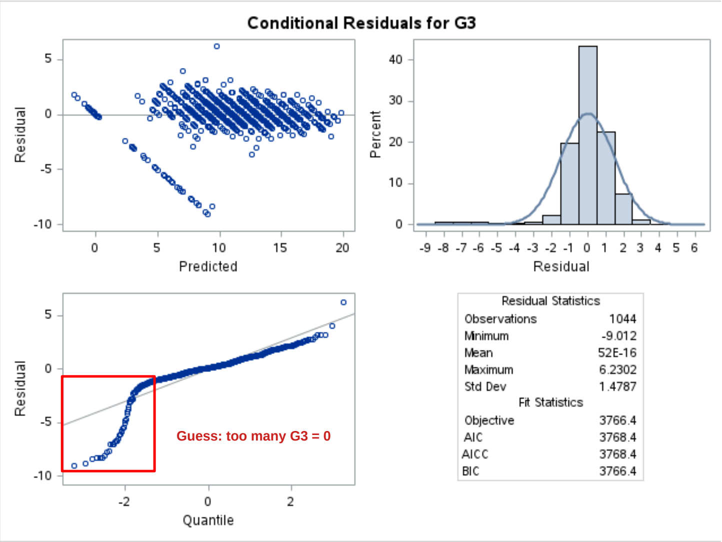
Another factor that seems to affect absences and subsequently performance is the parental status (pstatus), which indicates whether the guardian is staying together (T) with their child or apart (A), and family support. We can observe that two students with guardians away and without support performed worse than the average student, indicating that parental support and presence are necessary for students' success. Additionally, we investigate the interaction between alcohol consumption patterns and academic performance, with a focus on gender variability. The analysis reveals that alcohol consumption doesn't affect the performance of female students, but greatly affects the performance of male students. Finally, we examine the influence of guardian status (pstatus) on students' academic performance and their aspirations for higher education. Through graphical representations, it can be observed that regardless of guardian status, students who perform above average aspire to continue their education. For all detailed graphical representations, analysis, and comments, kindly see the appendix below.

Results

The GLMSELECT procedure was employed to select the most relevant (fixed-effect) features (Table 3.2) for predicting student academic performance (G3). The stepwise selection method was utilized to iteratively add or remove features based on their significance in predicting G3 scores in order to minimize the model's Akaike Information Criterion (AIC). The procedure identified the following features as significant fixed-effect predictors of student academic performance, with an R-squared value of approximately 0.84: Failures, Subject, Absences, G1, and G2. A random-effect model was constructed using the PROC MIXED procedure to investigate the relationship between the selected features and student performance while accounting for variability within schools.

$$\begin{aligned} G3 = \beta_0 + \beta_1 \times School + \beta_2 \times Failures + \beta_3 \times Subject + \beta_4 \times Absences \\ + \beta_5(Subject \times Failures) + \beta_6 \times G1 + \beta_7 \times G2 \end{aligned}$$

Judging by the residual plots, we can observe that the normal assumption isn't perfect. Based on further studies (see Appendix) of the problem, we believe the numerous zero final grade scores were the major reason for the bend in the QQ plot.

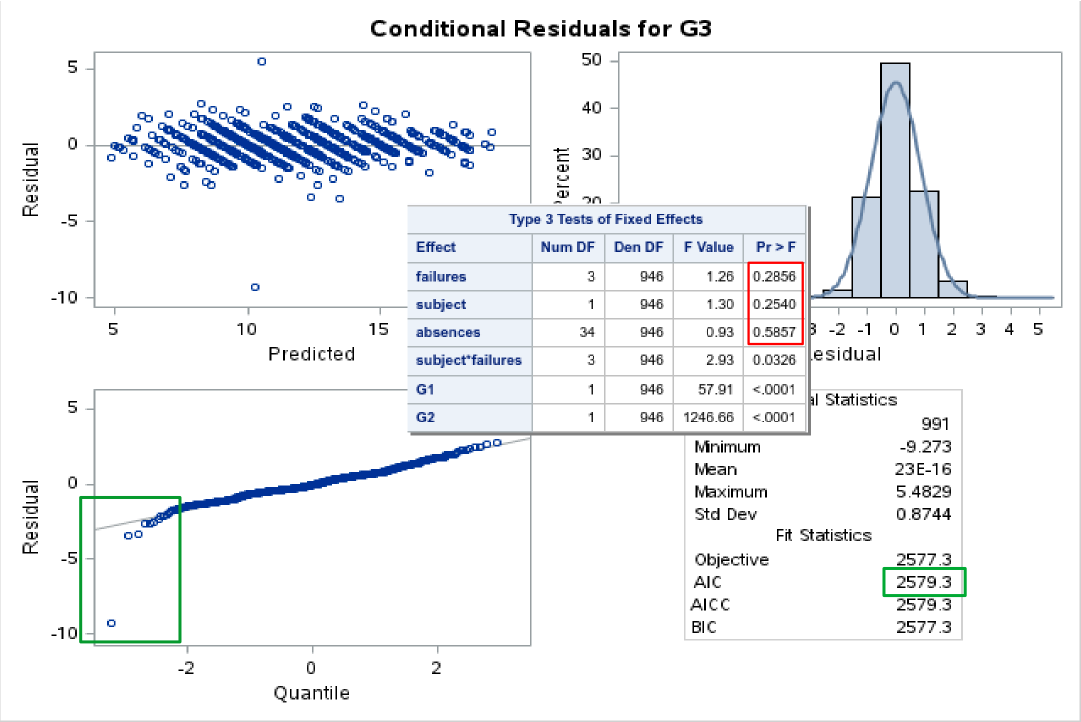


Residual Statistics	
Observations	1044
Minimum	-9.012
Mean	52E-16
Maximum	6.2302
Std Dev	1.4787
Fit Statistics	
Objective	3766.4
AIC	3768.4
AICC	3768.4
BIC	3766.4

Appendix A: Extra Results & Code

Ablation Study

In the ablation study, we conducted further analysis by removing all instances where the final grade (G3) was zero. This was done to examine the impact of excluding these data points on the model's performance and to assess whether they were influential in the overall analysis. First, the Akaike Information Criterion (AIC) value decreased significantly from 3768.4 to 2579.3. This reduction suggests that the modified model with zero G3 scores excluded provides a better fit to the data. Additionally, the residual plots, particularly the QQ plot, showed improvement in adherence to the normality assumption after removing the zero G3 scores. The QQ plot exhibited less deviation from the expected diagonal line, indicating a closer fit to the normal distribution. However, at significance level, $\alpha = 0.10$, the increase in the Pr $>$ F value suggests that, when the zero final grade scores were included in the analysis, the "failures", "subject", and "absences" features had a stronger and more significant relationship with student performance. However, upon removing these data points, the influence of the features may have diminished, leading to a higher p-value and reduced significance. This change implies that the presence of zero final grade scores may have disproportionately influenced the association between past class failures and academic performance in the original model. By removing these data points, the model's estimation of the effect of failures, subject, and absence on student performance may have become less precise or less reliable. This can be seen in the "Type 3 Test of Fixed Effects" table.



Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
failures	3	946	1.26	0.2856
subject	1	946	1.30	0.2540
absences	34	946	0.93	0.5857
subject*failures	3	946	2.93	0.0326
G1	1	946	57.91	<.0001
G2	1	946	1246.66	<.0001

l Statistics

	991
Minimum	-9.273
Mean	23E-16
Maximum	5.4829
Std Dev	0.8744

Fit Statistics

Objective	2577.3
AIC	2579.3
AICC	2579.3
BIC	2577.3

Table 3.2: Full Description of Features

Feature	**Description**
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, 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Fedu	father's education (numeric: 0 - none, 1 - primary education, 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Mjob	mother's job (nominal)
Fjob	father's job (nominal)
reason	reason to choose this school (nominal)
guardian	student's guardian (nominal)
traveltime	home to school travel time (numeric)
studytime	weekly study time (numeric)
failures	number of past class failures (numeric)
schoolsup	extra educational support (binary: yes or no)
famsup	family educational support (binary: yes or no)
paid	extra paid classes within the course subject (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)
freetime	free time after school (numeric)
goout	going out with friends (numeric)
Dalc	workday alcohol consumption (numeric)
Walc	weekend alcohol consumption (numeric)
health	current health status (numeric)
absences	number of school absences (numeric)
G1	first period grade (numeric)
G2	second period grade (numeric)
G3	final grade (numeric, output target)

```
*****************************
* IMPORTING BOTH DATA FILES *
*****************************;
proc import datafile='/home/u62534565/student_mat.csv'
    dbms='csv'
    out=mat
    replace;
    delimiter=';';
    getnames=yes;
    datarow=2;
run;

proc import datafile='/home/u62534565/student_por.csv'
    dbms='csv'
    out=por
    replace;
    delimiter=';';
    getnames=yes;
    datarow=2;
run;

data mat;
    set mat;
    subject="math";
    format subject $5.;
run;

data por;
    set por;
    subject="port";
    format subject $5.;
run;



**************************
* INSPECTING BOTH TABLES *
**************************;
title 'Student Performance on the Math Course';
proc print data=mat (obs=5);
run;
title;
```

```

title 'Student Performance on the Portuguese Language Course';
proc print data=por (obs=5);
run;
title;



************************************
* MERGING BOTH FILES INTO ONE DATA *
************************************;
proc sort data=mat;
    by _all_;
run;

proc sort data=por;
    by _all_;
run;

data students;
    merge mat por;
    by _all_;
run;

title 'Combined Student Performance';
proc print data=students (obs=5);
run;
title;



***************************
* OVERVIEW OF THE DATASET *
***************************;
proc contents data=students position;
run;

*converting G1 and G2 to numeric values;
data students;
    set students;
    G1_num = input(G1, best12.);
    G2_num = input(G2, best12.);
    drop G1 G2;
    rename G1_num=G1 G2_num=G2;
run;

%let score=g2; * observe: g1, g2, or g3;
proc sgplot data=students;
    title "Distribution of Subject Scores";
    histogram &score / group=subject transparency=0.5 nbins=30;
    xaxis label="Exam Scores (G1/G2/G3)";
    yaxis label="Distribution (in %)";
    keylegend / title="Subject";
run;
title;

proc sgplot data=students;
    vbox g3 / category=school group=subject;
    xaxis label="Subject";
    yaxis label="Final Grade (G3)";
    keylegend / title="Subject";
run;
* It seems it is easier to pass the Portuguese language course;



*****************************
* EXPLORATORY DATA ANALYSIS *
*****************************;
/* Question:
Does presence of family support and educational support affect
student performance while accounting for variability in gender? */
proc freq data=students;
    tables sex;
run;

proc means data=students;
    class sex famsup schoolsup;
    var g3;
run;

proc sgplot data=students;
    title "Distribution of Final Grade Given Family Support";
    histogram g3 / group=famsup transparency=0.5 nbins=30;
    xaxis label="Final Grade (G3)";
    yaxis label="Distribution (in %)";
    keylegend / title="Family Support";
run;
title;
/* Comments:
1. Although there are more females than males, performance
between genders isn't significantly different with roughtly
the same distribution.
2. However, with family support and school support, females
tend to perform better than males, with the best score for
males being 15 on Q3, whereas for females it's 18.
3. At the same time, with no support from either family or
school, we observe that both genders achieve their best
performances on average. */

/* Question:
Does the presence of educational support (schoolsup) affect
student performance (G3) while accounting for variability
within schools? */
proc freq data=students;
    tables school*schoolsup;
run;

proc means data=students mean stddev min max maxdec=3;
    class school schoolsup;
    var g3;
run;

proc sgplot data=students;
    title "Final grade variability between schools";
    vbox g3 / category=school;
    xaxis label="School";
    yaxis label="Final Grade (G3)";
run;
title;
/* Comments:
From the frequency and mean procedure, we observe that
Gabriel Pereira (GP) students perform better compared
to Mousinho da Silveira's (MS) students. It can also
be seen that relatively, GP has a larger number of
supported students (107) compared to MS (12), which
```

```
may indicate a greater emphasis on support services or
resources at GP */

/* Question:
What are the potential causes of extreme absenses? */
proc univariate data=students;
    var absences;
run;

ods graphics on;
proc freq data=students order=freq;
    tables absences / nocum plots=freqplot(orient=horizontal);
run;

%let absence=25; *absence threshold;
%let group=famsup; *observe: pstatus and famsup;
data ext_absence;
    set students;
    where absences > &absence;
    keep pstatus guardian medu fedu mjob fjob famsup higher traveltime
     ↪  romantic absences health g1 g2 g3;
run;

proc sort data=ext_absence;
    by health absences;
run;

proc print data=ext_absence;
run;

proc print data=ext_absence;
    where pstatus="A" and famsup="no";
run;

proc sgplot data=ext_absence;
    title "Relationship Between Absence and Final Grade";
    scatter x=absences y=g3 / group=&group;
    xaxis label="Absence";
    yaxis label="Final Grade (G3)";
    keylegend / title="Family Support";
run;
title;
```

```
/* Comment:
1. Among the students with the most absences, only five
appear to have extreme health problems with health=1 or 2.
Surprisingly, two students performed better than the average
student.
2. Another factor that seems to affect absences and
subsequently performance is the parental status (pstatus),
which indicates whether the guardian is staying together (T)
with their child or apart (A), and family support. We can
observe that two students with guardians away and without
support performed worse than the average student, indicating
that parental support and presence are necessary for students'
success. */

/* Question:
Does the number of past class failures predict final grades
differently across schools */
proc sgplot data=students;
    title "Variabily of Final Scores Across Schools Categorized by Failures";
    vbox g3 / category=failures group=school;
    xaxis label="Failures";
    yaxis label="Final Scores (G3)";
    keylegend / title="School";
run;
title;
/* Comment:
We observe that students in Gabriel Pereira (GP) who failed
more than once performed poorly in the final exam compared
to students at Mousinho da Silveira (MS). */

/* Question:
Is there an interaction between weekend alcohol consumption
(walc) and weekday alcohol consumption (dalc) in predicting
academic performance, and does this interaction vary between
gender (sex) */
%let category=walc; *takes: dalc or walc;
proc sgplot data=students;
    title "Variabily of Final Scores Across Alcohol Consumption Categorized by
     ↪  Gender";
    vbox g3 / category=&category group=sex;
    yaxis label="Final Scores (G3)";
    keylegend / title="Gender";
```

```
run;
title;

proc means data=students maxdec=3;
    class &category sex;
    var g3;
run;
/* Comment:
Based on the variability of the boxplots and averages, we can
observe that alcohol consumption doesn't affect the performance
of female students, but greatly affects the performance of
male students. This is evident with female students having a
minimum of 11 and 10 for daily and weekly consumption
respectively, while the minimum for male students is 5 and 0
for daily and weekly consumption respectively. */

/* Question:
Does the type of guardian (pstatus) influences students'
academic performance (g3) and their aspiration for higher
education (higher)
*/
proc sgplot data=students;
    title "Relationship between Guardian Status, Higher Education Aspiration, and Final Scores";
    vbox g3 / category=pstatus group=higher groupdisplay=cluster;
    xaxis label="Guardian Status (pstatus)";
    yaxis label="Final Scores (G3)";
    keylegend / title="Higher Education?";
run;
title;
/* Comment:
It can be observed that regardless of guardian status (i.e.,
whether the guardian is living together or apart from the
student), students who perform above average aspire to continue
their education. */



******************************************************
* FIXED EFFECT FEATURE SELECTION AND MODEL SELECTION *
******************************************************;
proc glmselect data=students;
```

```
    class sex address famsize pstatus medu fedu mjob fjob reason guardian schoolsup famsup paid activities nursery higher internet romantic subject;
    model g3 = sex age address famsize pstatus medu fedu mjob fjob reason guardian traveltime studytime failures schoolsup famsup paid activities nursery higher internet romantic famrel freetime goout dalc walc health absences subject g1 g2 / selection=stepwise(stop=none);
run;

data selected_features;
    set students;
    keep school failures subject absences g1 g2 g3; * selected features from glmselect;
run;

data correlation;
    format subj best12.;
    format g1 best12.;
    format g2 best12.;
    set selected_features;
    if subject = "math" then subj = 1;
    else if subject = "port" then subj = 0;
    keep failures subj absences g1 g2 g3;
run;

proc corr data=correlation;
    var failures subj absences g1 g2 g3;
run;

proc mixed data=selected_features method=reml covtest plots=(residualPanel) alpha=0.1;
    class school absences subject failures;
    model g3 = failures subject absences subject*failures g1 g2;
    random intercept school;
run;



******************
* ABLATION STUDY *
******************;
* Removing all G3 scores=0;
```

```
data nozeros;
    set students;
    where g3 > 0;
    keep school failures subject paid absences g1 g2 g3;
run;

proc mixed data=nozeros method=reml covtest plots=(residualPanel) alpha=0.1;
    class school absences subject failures;
    model g3 = failures subject absences subject*failures g1 g2;
    random intercept school;
run;
```

The CONTENTS Procedure

Data Set Name	WORK.STUDENTS	**Observations**	1044
Member Type	DATA	**Variables**	34
Engine	V9	**Indexes**	0
Created	04/29/2024 15:23:08	**Observation Length**	224
Last Modified	04/29/2024 15:23:08	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	2
First Data Page	1
Max Obs per Page	584
Obs in First Data Page	556
Number of Data Set Repairs	0
Filename	/saswork/SAS_work4FC400009B58_odaws02-usw2.oda.sas.com/SAS_work389E00009B58_odaws02-usw2.oda.sas.com/students.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	1075166704
Access Permission	rw-r--r--
Owner Name	u62534565
File Size	384KB
File Size (bytes)	393216

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
27	Dalc	Num	8	BEST12.	BEST32.
8	Fedu	Num	8	BEST12.	BEST32.
10	Fjob	Char	10	$10.	$10.
31	G1	Char	4	$4.	$4.
32	G2	Char	4	$4.	$4.
33	G3	Num	8	BEST12.	BEST32.
7	Medu	Num	8	BEST12.	BEST32.
9	Mjob	Char	10	$10.	$10.
6	Pstatus	Char	3	$3.	$3.
28	Walc	Num	8	BEST12.	BEST32.
30	absences	Num	8	BEST12.	BEST32.
19	activities	Char	5	$5.	$5.
4	address	Char	3	$3.	$3.
3	age	Num	8	BEST12.	BEST32.
15	failures	Num	8	BEST12.	BEST32.
24	famrel	Num	8	BEST12.	BEST32.
5	famsize	Char	5	$5.	$5.

Alphabetic List of Variables and Attributes

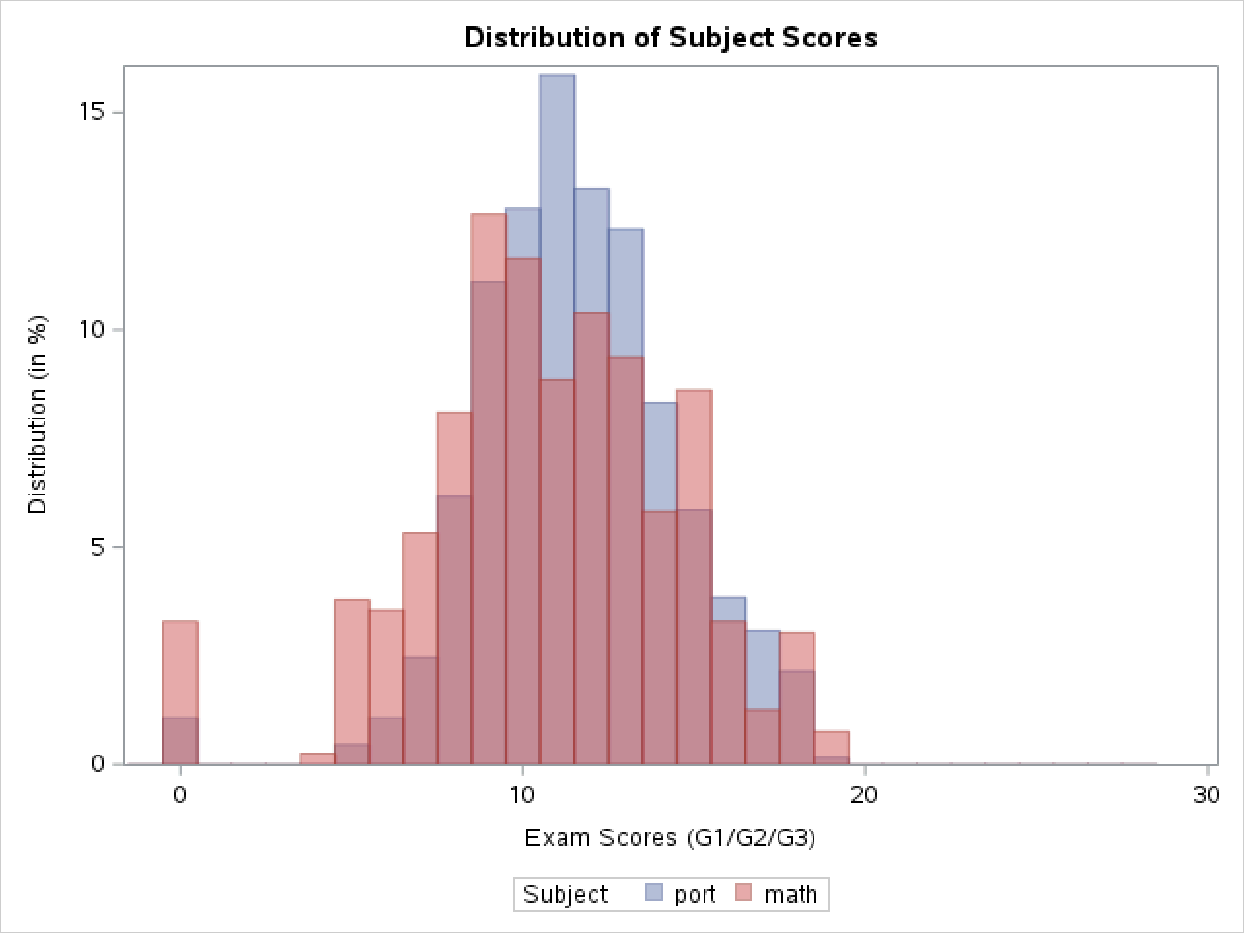
#	Variable	Type	Len	Format	Informat
17	famsup	Char	5	$5.	$5.
25	freetime	Num	8	BEST12.	BEST32.
26	goout	Num	8	BEST12.	BEST32.
12	guardian	Char	8	$8.	$8.
29	health	Num	8	BEST12.	BEST32.
21	higher	Char	5	$5.	$5.
22	internet	Char	5	$5.	$5.
20	nursery	Char	5	$5.	$5.
18	paid	Char	5	$5.	$5.
11	reason	Char	12	$12.	$12.
23	romantic	Char	5	$5.	$5.
1	school	Char	4	$4.	$4.
16	schoolsup	Char	5	$5.	$5.
2	sex	Char	3	$3.	$3.
14	studytime	Num	8	BEST12.	BEST32.
34	subject	Char	4	$5.	
13	traveltime	Num	8	BEST12.	BEST32.

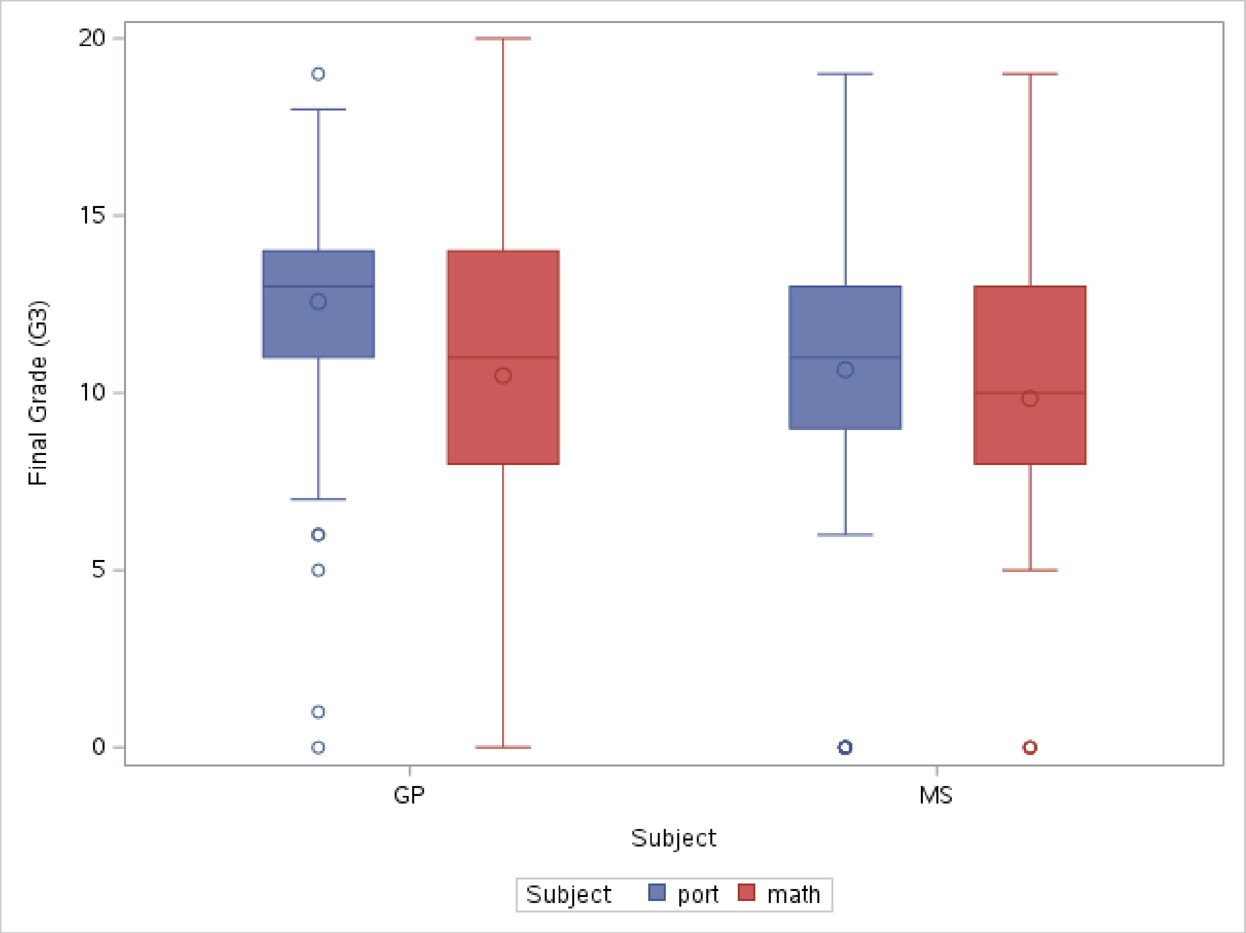
Variables in Creation Order

#	Variable	Type	Len	Format	Informat
1	school	Char	4	$4.	$4.
2	sex	Char	3	$3.	$3.
3	age	Num	8	BEST12.	BEST32.
4	address	Char	3	$3.	$3.
5	famsize	Char	5	$5.	$5.
6	Pstatus	Char	3	$3.	$3.
7	Medu	Num	8	BEST12.	BEST32.
8	Fedu	Num	8	BEST12.	BEST32.
9	Mjob	Char	10	$10.	$10.
10	Fjob	Char	10	$10.	$10.
11	reason	Char	12	$12.	$12.
12	guardian	Char	8	$8.	$8.
13	traveltime	Num	8	BEST12.	BEST32.
14	studytime	Num	8	BEST12.	BEST32.
15	failures	Num	8	BEST12.	BEST32.
16	schoolsup	Char	5	$5.	$5.
17	famsup	Char	5	$5.	$5.
18	paid	Char	5	$5.	$5.
19	activities	Char	5	$5.	$5.
20	nursery	Char	5	$5.	$5.
21	higher	Char	5	$5.	$5.
22	internet	Char	5	$5.	$5.
23	romantic	Char	5	$5.	$5.
24	famrel	Num	8	BEST12.	BEST32.
25	freetime	Num	8	BEST12.	BEST32.
26	goout	Num	8	BEST12.	BEST32.
27	Dalc	Num	8	BEST12.	BEST32.

Variables in Creation Order

#	Variable	Type	Len	Format	Informat
28	Walc	Num	8	BEST12.	BEST32.
29	health	Num	8	BEST12.	BEST32.
30	absences	Num	8	BEST12.	BEST32.
31	G1	Char	4	$4.	$4.
32	G2	Char	4	$4.	$4.
33	G3	Num	8	BEST12.	BEST32.
34	subject	Char	4	$5.	



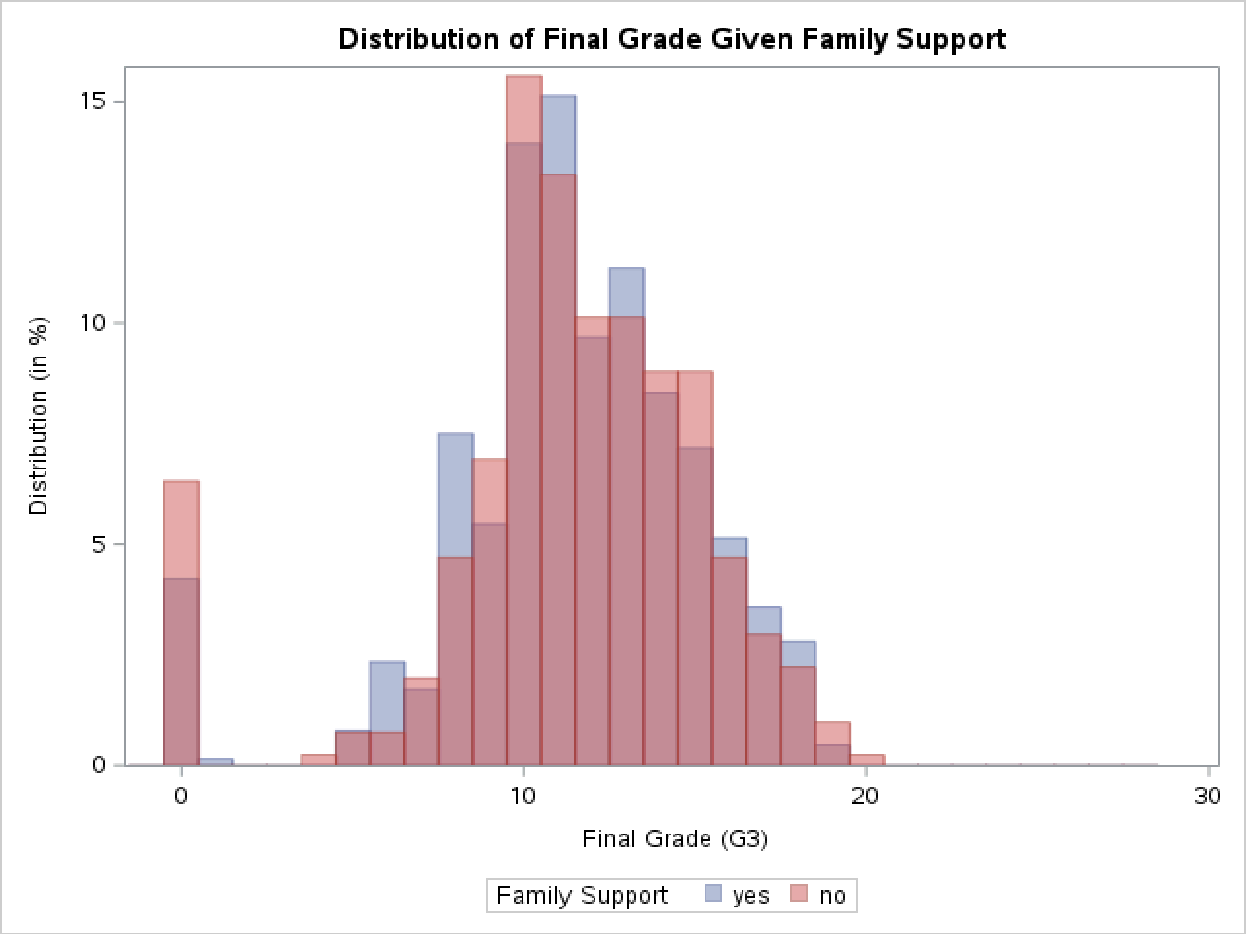


The FREQ Procedure

sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	591	56.61	591	56.61
M	453	43.39	1044	100.00

The MEANS Procedure

Analysis Variable : G3								
sex	famsup	schoolsup	N Obs	N	Mean	Std Dev	Minimum	Maximum
F	no	no	169	169	11.6213018	4.2072523	0	19.0000000
		yes	25	25	10.9600000	2.0510160	6.0000000	17.0000000
	yes	no	335	335	11.5641791	3.9481558	0	19.0000000
		yes	62	62	10.5483871	2.8896504	0	18.0000000
M	no	no	203	203	11.0837438	4.1467967	0	20.0000000
		yes	7	7	9.7142857	1.6035675	8.0000000	12.0000000
	yes	no	218	218	11.4908257	3.6710862	0	19.0000000
		yes	25	25	10.0800000	3.0675723	0	15.0000000



The FREQ Procedure

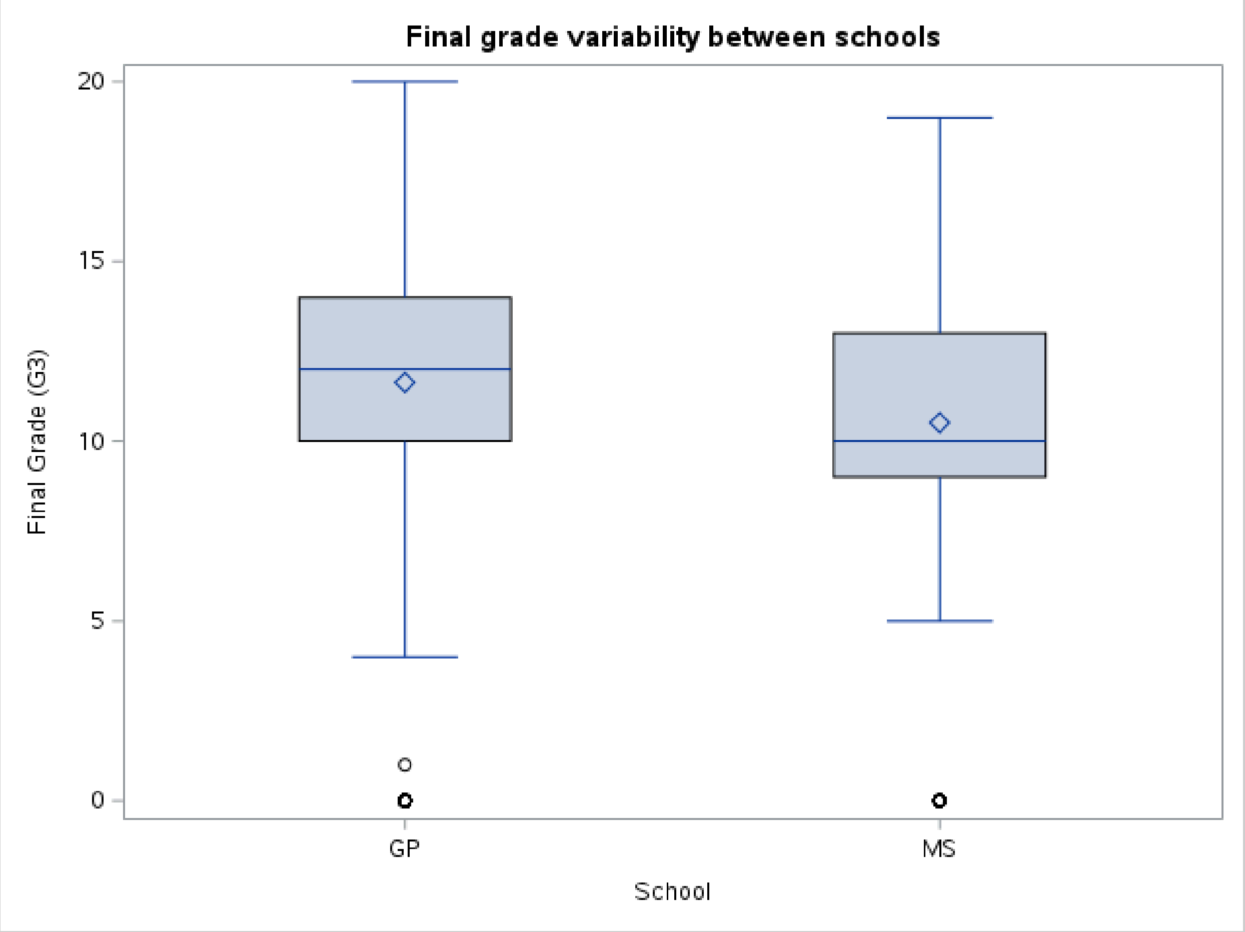
Frequency Percent Row Pct Col Pct

Table of school by schoolsup			
	schoolsup		
school	no	yes	Total
GP	665 63.70 86.14 71.89	107 10.25 13.86 89.92	772 73.95
MS	260 24.90 95.59 28.11	12 1.15 4.41 10.08	272 26.05
Total	925 88.60	119 11.40	1044 100.00

The MEANS Procedure

Analysis Variable : G3						
school	schoolsup	N Obs	Mean	Std Dev	Minimum	Maximum
GP	no	665	11.823	3.949	0.000	20.000
	yes	107	10.458	2.500	0.000	17.000
MS	no	260	10.504	3.899	0.000	19.000

Analysis Variable : G3						
school	schoolsup	N Obs	Mean	Std Dev	Minimum	Maximum
	yes	12	10.750	4.288	0.000	18.000



The UNIVARIATE Procedure
Variable: absences

Moments			
N	1044	Sum Weights	1044
Mean	4.4348659	Sum Observations	4630
Std Deviation	6.21001656	Variance	38.5643057
Skewness	3.7413466	Kurtosis	26.5962003
Uncorrected SS	60756	Corrected SS	40222.5709
Coeff Variation	140.027155	Std Error Mean	0.19219519

Basic Statistical Measures			
Location		Variability	
Mean	4.434866	Std Deviation	6.21002
Median	2.000000	Variance	38.56431
Mode	0.000000	Range	75.00000
		Interquartile Range	6.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	23.0748	Pr > |t|	<.0001
Sign	M	342.5	Pr >= |M|	<.0001
Signed Rank	S	117477.5	Pr >= |S|	<.0001

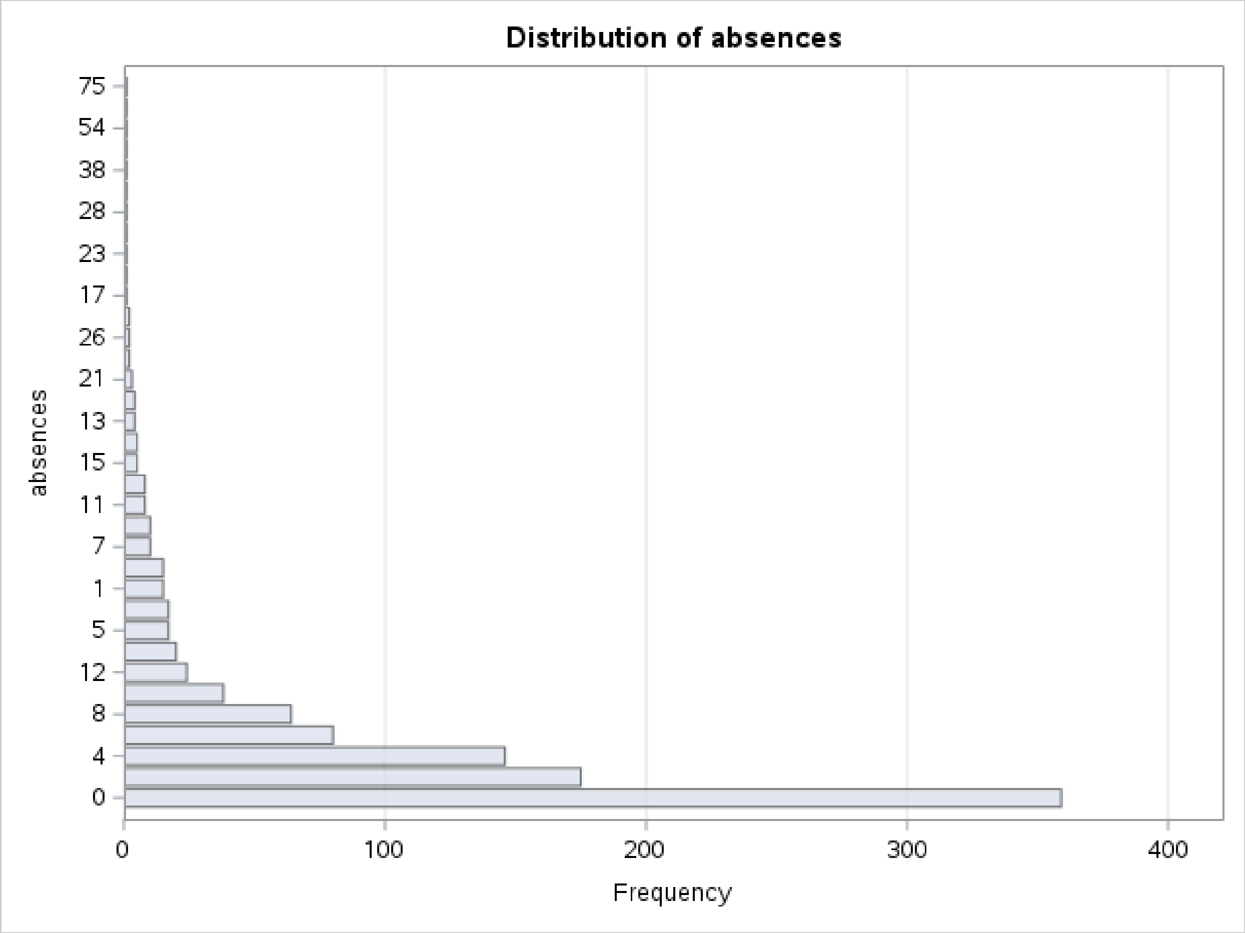
Quantiles (Definition 5)	
Level	Quantile
100% Max	75
99%	26
95%	16
90%	12
75% Q3	6
50% Median	2
25% Q1	0
10%	0
5%	0
1%	0
0% Min	0

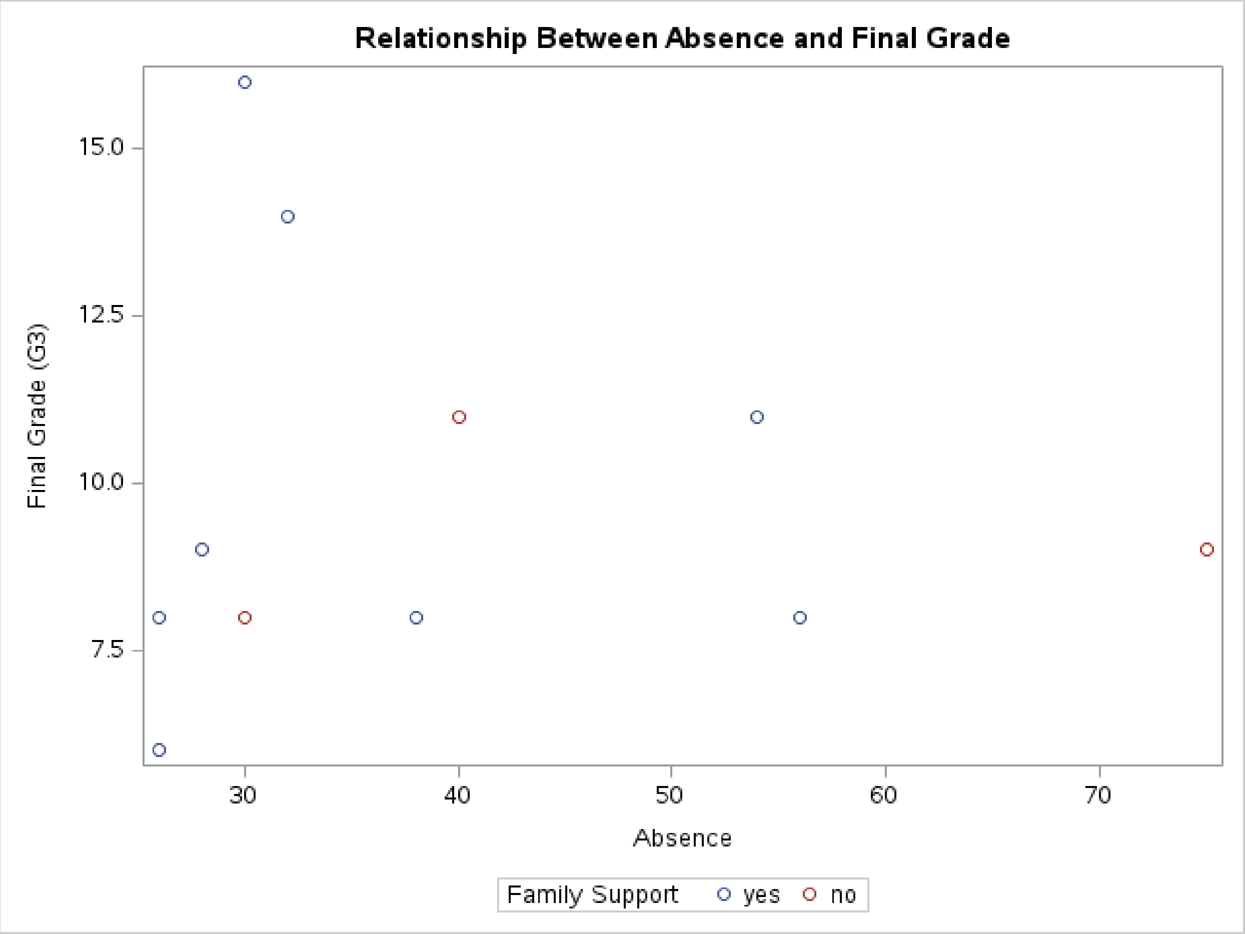
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	1037	38	765
0	1036	40	396
0	1031	54	141
0	1030	56	313
0	1026	75	319

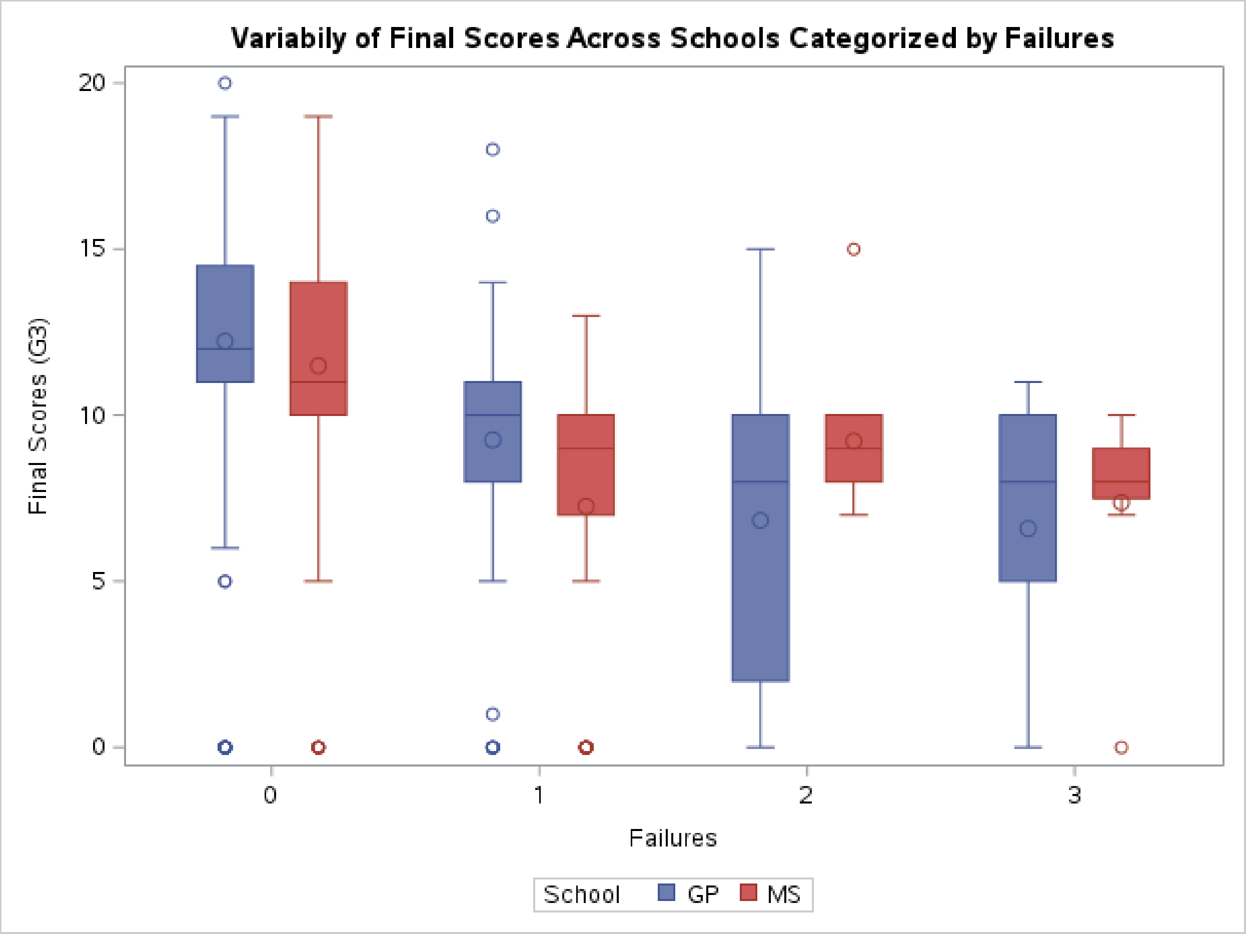
The FREQ Procedure

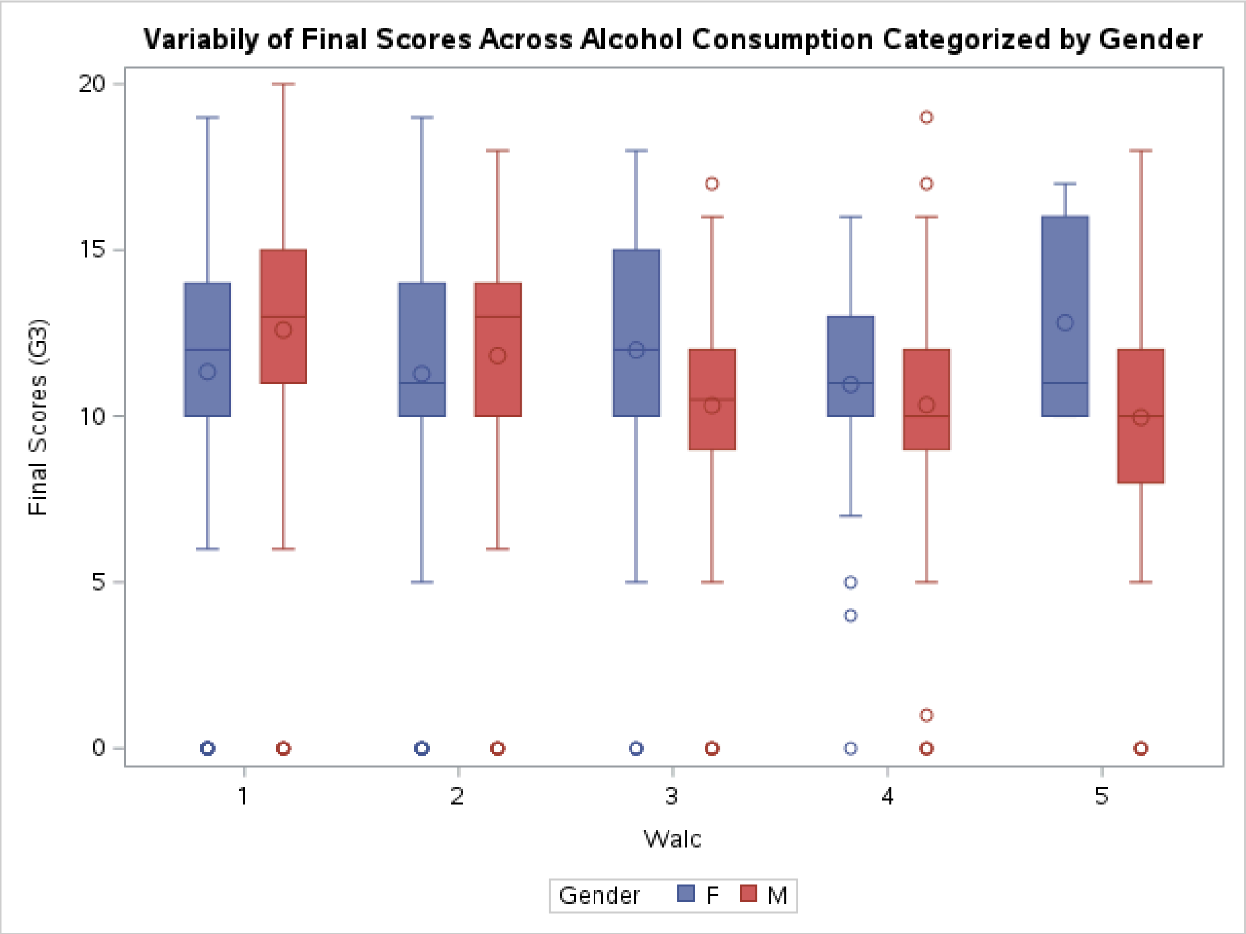
absences	Frequency	Percent
0	359	34.39
2	175	16.76
4	146	13.98
6	80	7.66
8	64	6.13
10	38	3.64
12	24	2.30
14	20	1.92
5	17	1.63
16	17	1.63
1	15	1.44
3	15	1.44
7	10	0.96
9	10	0.96
11	8	0.77
18	8	0.77

absences	Frequency	Percent
15	5	0.48
22	5	0.48
13	4	0.38
20	4	0.38
21	3	0.29
24	2	0.19
26	2	0.19
30	2	0.19
17	1	0.10
19	1	0.10
23	1	0.10
25	1	0.10
28	1	0.10
32	1	0.10
38	1	0.10
40	1	0.10
54	1	0.10
56	1	0.10
75	1	0.10



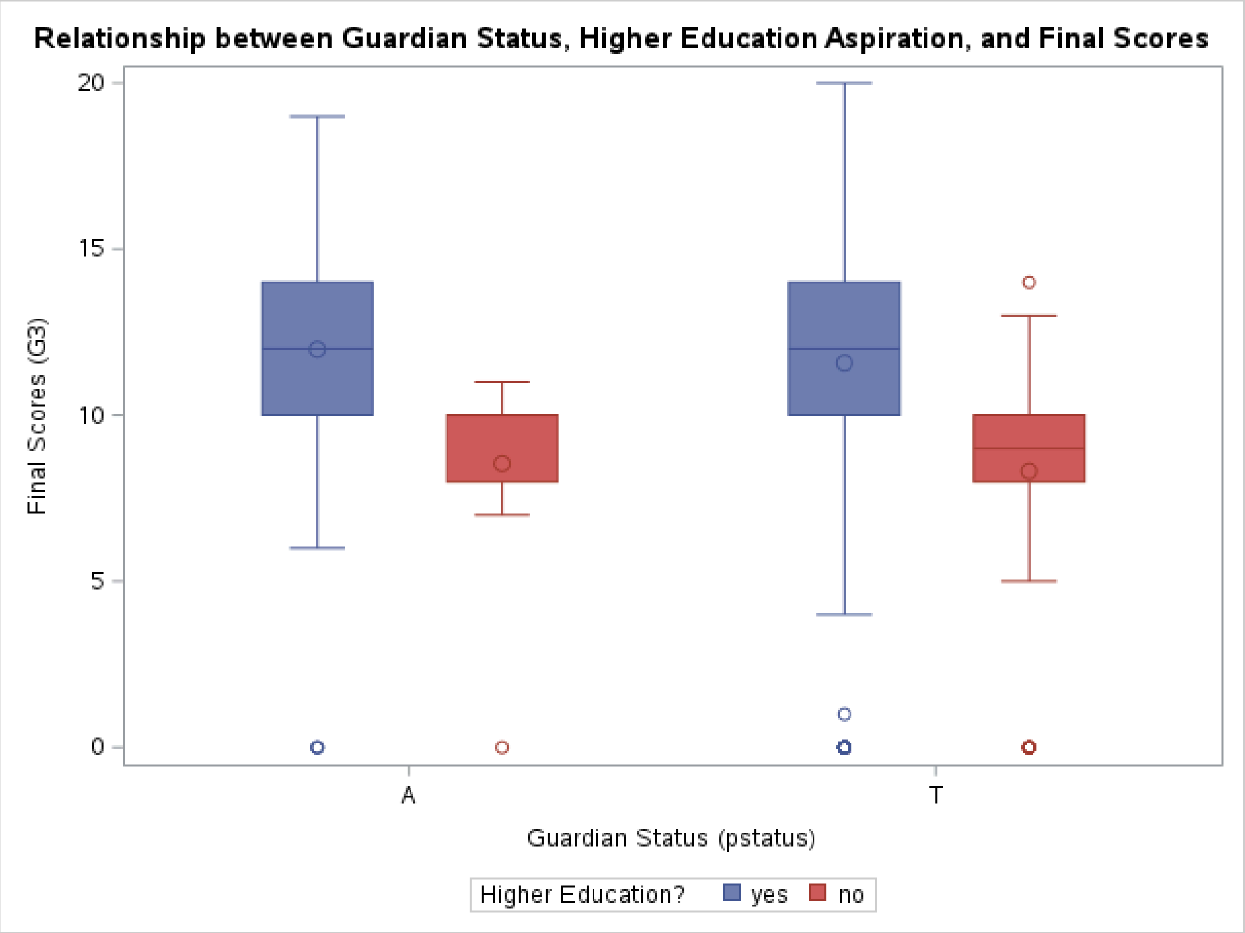

Relationship Between Absence and Final Grade
Final Grade (G3)
15.0
12.5
10.0
7.5
30
40
50
60
70
Absence
Family Support
yes
no


Variabily of Final Scores Across Schools Categorized by Failures
Final Scores (G3)
20
15
10
5
0
0
1
2
3
Failures
School
🟦 GP
🟥 MS



The MEANS Procedure

Analysis Variable : G3							
Walc	sex	N Obs	N	Mean	Std Dev	Minimum	Maximum
1	F	270	270	11.337	3.996	0.000	19.000
	M	128	128	12.602	4.032	0.000	20.000
2	F	150	150	11.273	4.056	0.000	19.000
	M	85	85	11.824	3.883	0.000	18.000
3	F	116	116	11.991	3.689	0.000	18.000
	M	84	84	10.321	3.475	0.000	17.000
4	F	44	44	10.955	3.019	0.000	16.000
	M	94	94	10.340	3.258	0.000	19.000
5	F	11	11	12.818	2.786	10.000	17.000
	M	62	62	9.968	3.785	0.000	18.000



The GLMSELECT Procedure

Data Set	WORK.STUDENTS
Dependent Variable	G3
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	None
Effect Hierarchy Enforced	None

Number of Observations Read	1044
Number of Observations Used	1044

Class Level Information		
Class	**Levels**	**Values**
sex	2	F M
address	2	R U
famsize	2	GT3 LE3
Pstatus	2	A T
Medu	5	0 1 2 3 4
Fedu	5	0 1 2 3 4
Mjob	5	at_home health other services teacher
Fjob	5	at_home health other services teacher
reason	4	course home other reputation

Class Level Information		
Class	**Levels**	**Values**
guardian	3	father mother other
schoolsup	2	no yes
famsup	2	no yes
paid	2	no yes
activities	2	no yes
nursery	2	no yes
higher	2	no yes
internet	2	no yes
romantic	2	no yes
subject	2	math port

Dimensions	
Number of Effects	33
Number of Parameters	67

The GLMSELECT Procedure

Stepwise Selection Summary					
Step	**Effect Entered**	**Effect Removed**	**Number Effects In**	**Number Parms In**	**SBC**
0	**Intercept**		1	1	2828.7360
1	**G2**		2	2	989.1170
2	**subject**		3	3	963.3152
3	**G1**		4	4	949.5348
4	**absences**		5	5	940.5029
5	**failures**		6	6	938.3318*
*** Optimal Value of Criterion**					

Selection stopped as adding or dropping any effect does not improve the selection criterion.

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 5).

Effects:	Intercept failures absences subject G1 G2

Analysis of Variance				
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**
Model	5	13115	2622.92952	1104.83
Error	1038	2464.27481	2.37406	
Corrected Total	1043	15579		

Root MSE	1.54080
Dependent Mean	11.34195
R-Square	0.8418
Adj R-Sq	0.8411

AIC	1954.62696
AICC	1954.73506
SBC	938.33185

Parameter Estimates				
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**
Intercept	1	-0.602424	0.218769	-2.75
failures	1	-0.239012	0.079192	-3.02
absences	1	0.032872	0.007824	4.20
subject math	1	-0.659918	0.100523	-6.56
subject port	0	0	.	.
G1	1	0.139292	0.031461	4.43
G2	1	0.938053	0.028708	32.68

The CORR Procedure

6 Variables:	failures subj absences g1 g2 G3

Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**
failures	1044	0.26437	0.65614	276.00000	0	3.00000
subj	1044	0.37835	0.48521	395.00000	0	1.00000
absences	1044	4.43487	6.21002	4630	0	75.00000
g1	1044	11.21360	2.98339	11707	0	19.00000
g2	1044	11.24617	3.28507	11741	0	19.00000
G3	1044	11.34195	3.86480	11841	0	20.00000

Pearson Correlation Coefficients, N = 1044 Prob > |r| under H0: Rho=0						
	failures	**subj**	**absences**	**g1**	**g2**	**G3**
failures	1.00000	0.08304 0.0073	0.10000 0.0012	-0.37417 <.0001	-0.37717 <.0001	-0.38315 <.0001
subj	0.08304 0.0073	1.00000	0.16013 <.0001	-0.07973 0.0100	-0.12646 <.0001	-0.18717 <.0001
absences	0.10000 0.0012	0.16013 <.0001	1.00000	-0.09242 0.0028	-0.08933 0.0039	-0.04567 0.1403
g1	-0.37417 <.0001	-0.07973 0.0100	-0.09242 0.0028	1.00000	0.85874 <.0001	0.80914 <.0001
g2	-0.37717 <.0001	-0.12646 <.0001	-0.08933 0.0039	0.85874 <.0001	1.00000	0.91074 <.0001
G3	-0.38315 <.0001	-0.18717 <.0001	-0.04567 0.1403	0.80914 <.0001	0.91074 <.0001	1.00000

The Mixed Procedure

Model Information	
Data Set	WORK.SELECTED_FEATURES
Dependent Variable	G3
Covariance Structure	Variance Components

Model Information	
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Class Level Information

Class	Levels	Values
school	2	GP MS
absences	35	0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 28 30 32 38 40 54 56 75
subject	2	math port
failures	4	0 1 2 3

Dimensions	
Covariance Parameters	3
Columns in X	52
Columns in Z	3
Subjects	1
Max Obs per Subject	1044

Number of Observations	
Number of Observations Read	1044
Number of Observations Used	1044
Number of Observations Not Used	0

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	3766.43285802	
1	1	3766.43285802	0.00000000

Convergence criteria met but final Hessian is not positive definite.

Estimated G matrix is not positive definite.

Covariance Parameter Estimates

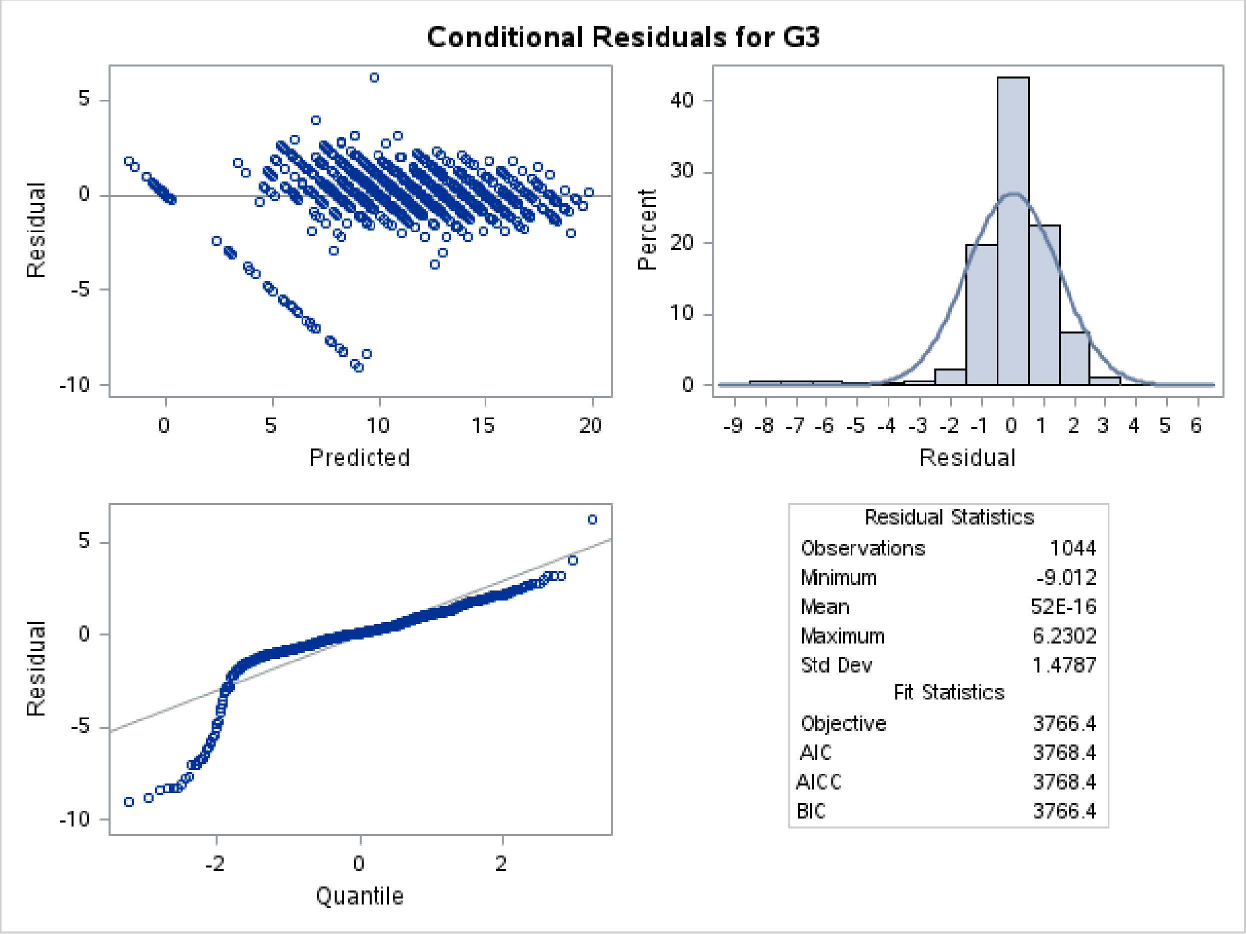
Cov Parm	Estimate	Standard Error	Z Value	Pr > Z	Alpha	Lower	Upper
Intercept	0	.	.	.	.	.	.
school	0	.	.	.	.	.	.
Residual	2.2805	0.1020	22.36	<.0001	0.1	2.1220	2.4585

Fit Statistics	
-2 Res Log Likelihood	3766.4
AIC (Smaller is Better)	3768.4
AICC (Smaller is Better)	3768.4
BIC (Smaller is Better)	3766.4

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
failures	3	999	7.45	<.0001

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
subject	1	999	18.88	<.0001
absences	34	999	2.53	<.0001
subject*failures	3	999	2.21	0.0850
G1	1	999	27.70	<.0001
G2	1	999	1001.97	<.0001


Conditional Residuals for G3

Residual Statistics	
Observations	1044
Minimum	-9.012
Mean	52E-16
Maximum	6.2302
Std Dev	1.4787
Fit Statistics	
Objective	3766.4
AIC	3768.4
AICC	3768.4
BIC	3766.4

The Mixed Procedure

Model Information	
Data Set	WORK.NOZEROS
Dependent Variable	G3
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Class Level Information		
Class	**Levels**	**Values**
school	2	GP MS
absences	35	0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 28 30 32 38 40 54 56 75

Class Level Information		
Class	**Levels**	**Values**
subject	2	math port
failures	4	0 1 2 3

Dimensions	
Covariance Parameters	3
Columns in X	52
Columns in Z	3
Subjects	1
Max Obs per Subject	991

Number of Observations	
Number of Observations Read	991
Number of Observations Used	991
Number of Observations Not Used	0

Iteration History			
Iteration	**Evaluations**	**-2 Res Log Like**	**Criterion**
0	1	2577.26721649	
1	1	2577.26721649	0.00000000

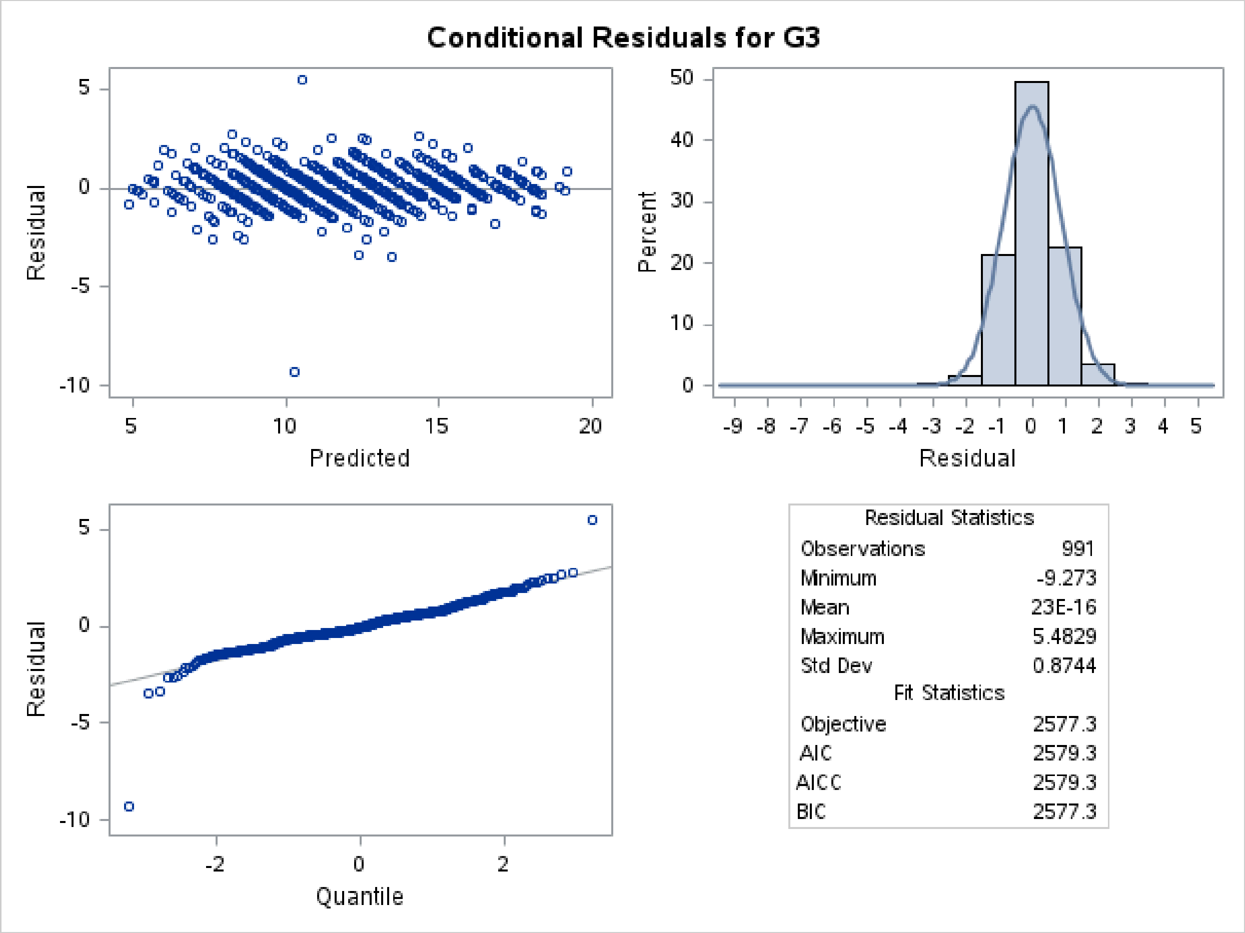
Convergence criteria met but final Hessian is not positive definite.

Estimated G matrix is not positive definite.

Covariance Parameter Estimates							
Cov Parm	**Estimate**	**Standard Error**	**Z Value**	**Pr > Z**	**Alpha**	**Lower**	**Upper**
Intercept	0	.	.	.	.	.	.
school	0	.	.	.	.	.	.
Residual	0.7993	0.03673	21.76	<.0001	0.1	0.7423	0.8635

Fit Statistics	
-2 Res Log Likelihood	2577.3
AIC (Smaller is Better)	2579.3
AICC (Smaller is Better)	2579.3
BIC (Smaller is Better)	2577.3

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
failures	3	946	1.26	0.2856
subject	1	946	1.30	0.2540
absences	34	946	0.93	0.5857
subject*failures	3	946	2.93	0.0326
G1	1	946	57.91	<.0001
G2	1	946	1246.66	<.0001



Conditional Residuals for G3

Residual Statistics	
Observations	991
Minimum	-9.273
Mean	23E-16
Maximum	5.4829
Std Dev	0.8744
Fit Statistics	
Objective	2577.3
AIC	2579.3
AICC	2579.3
BIC	2577.3