Qualitative Data Project

Carl Robinson 29/01/2018

UCI Students Performance Dataset

Objective

To use the SAS statistical software package to analyse a dataset of qualitative and quantitative variables, using descriptive statistics, multiple correspondence analysis, discriminant analysis and clustering techniques.

Data description, preprocessing and uni-dimensional analysis

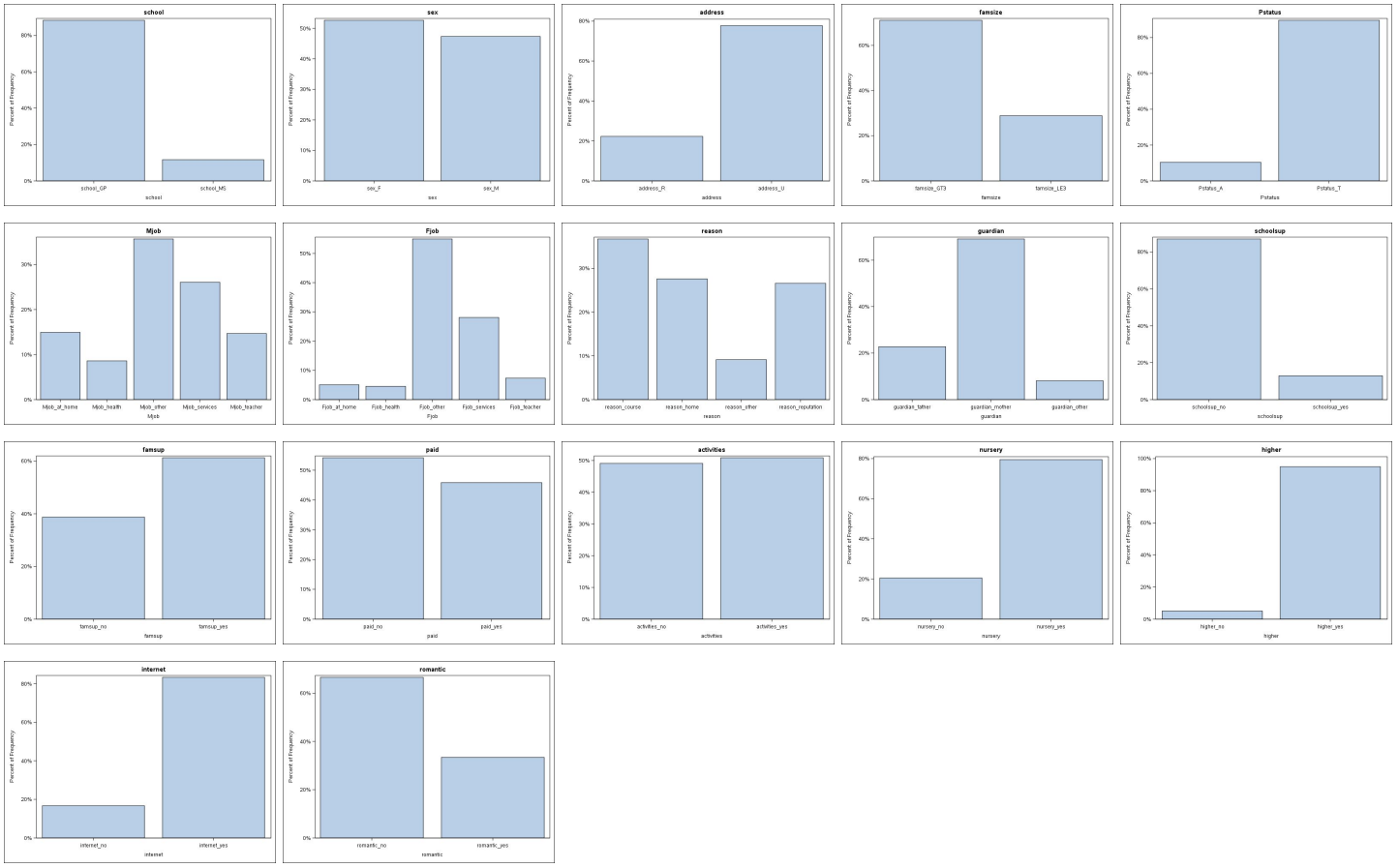
The data contain information on 395 students in two Portuguese secondary schools. Student grades, as well as demographic, social and school related features were collected using school reports and questionnaires. The data are split across two data files for the maths course and the Portuguese language course respectively; as these contain the same variables for the same students, we restrict our analysis to the maths course dataset only. The data is available for download here: https://archive.ics.uci.edu/ml/datasets/student+performance. There are 33 qualitative and quantitative variables present in the data:

Qualitative (categorical)

1. school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
2. sex - student's sex (binary: "F" - female or "M" - male)
3. address - student's home address type (binary: "U" - urban or "R" - rural)
4. famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
5. Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)
6. Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
7. Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
8. reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
9. guardian - student's guardian (nominal: "mother", "father" or "other")
10. schoolsup - extra educational support (binary: yes or no)

11. famsup - family educational support (binary: yes or no)
12. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
13. activities - extra-curricular activities (binary: yes or no)
14. nursery - attended nursery school (binary: yes or no)
15. higher - wants to take higher education (binary: yes or no)
16. internet - Internet access at home (binary: yes or no)
17. romantic - with a romantic relationship (binary: yes or no)

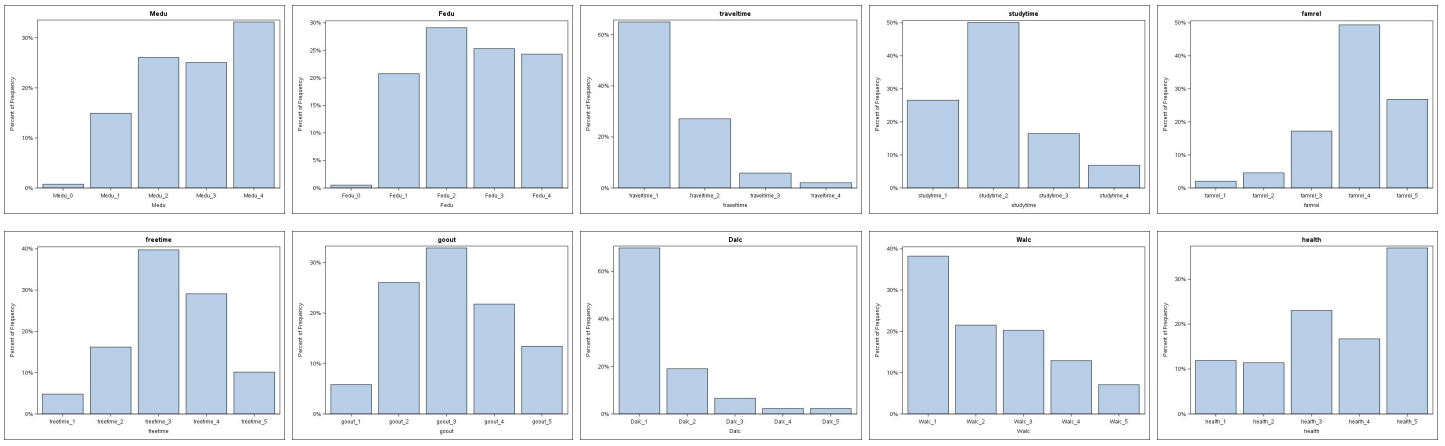
When using methods such as Multiple Correspondence Analysis (MCA), we must avoid confusion due to multiple modalities sharing the same name. Therefore the variable modalities were renamed to contain the variable name e.g. nursery_yes.



Qualitative (numeric)

18. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
19. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
20. 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)
21. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
22. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
23. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
24. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
25. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
26. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
27. health - current health status (numeric: from 1 - very bad to 5 - very good)

For the same reason as before, the variable modalities were renamed to contain the variable name e.g. health_3.

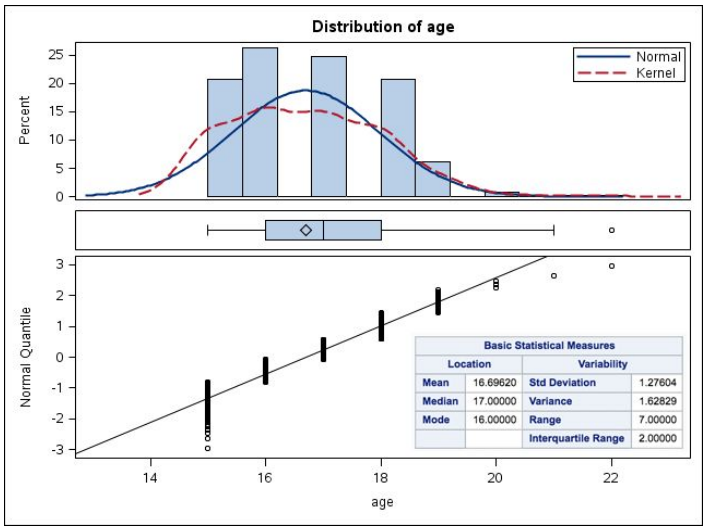


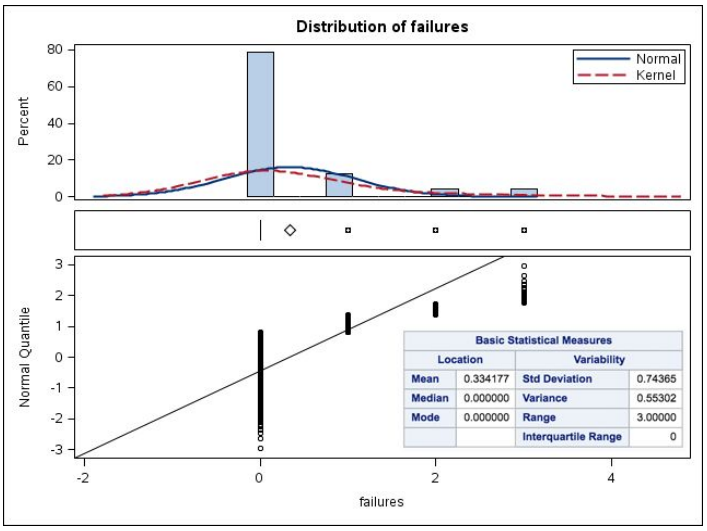
- By plotting the histograms we can quickly characterise each of the variables. We see that around 90% of students sampled come from one of the two schools, that almost all students want to go to university, and only around 10% are receiving extra educational support.
- A large majority of students live in an urban environment, with parents who remain together and are most likely to work in the service industries. Most students come from families of more than 3 people (meaning they have siblings), where their mother is their guardian. Additionally we see that most children went to nursery school when they were young, and have the internet at home.
- We conclude from this profile that the majority of students have a relatively high socio-economic status.
- We see that most students report a good or very good relationship with their family, are likely to have parents with an above average education level (especially the mother), and are generally in good/excellent health. However, there seems to be a significant number of students who drink at the weekends, and most report that their study time is average at best.

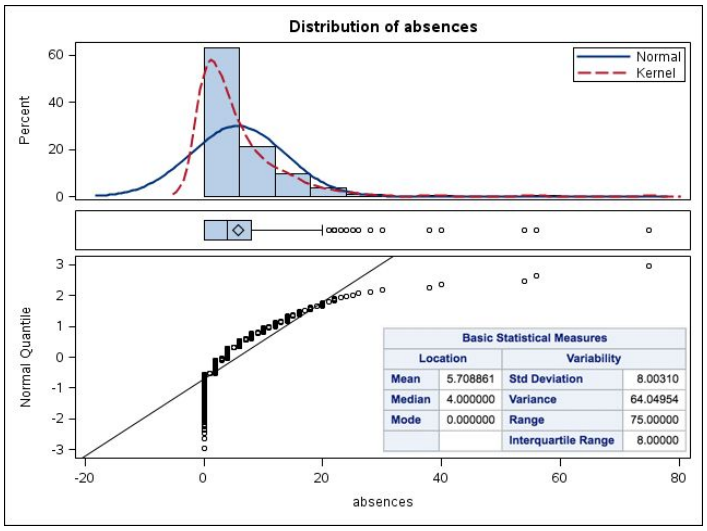
Quantitative (numeric)

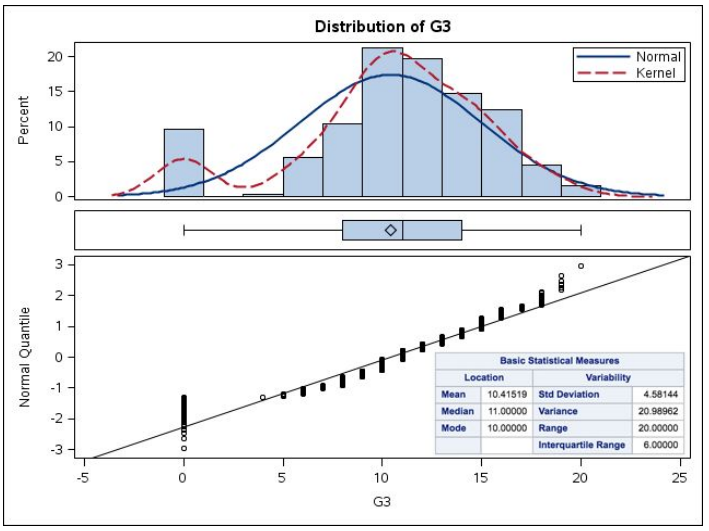
28. age - student's age (numeric: from 15 to 22)
29. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
30. absences - number of school absences (numeric: from 0 to 93)
31. G1 - first period grade for maths (numeric: from 0 to 20)
32. G2 - second period grade for maths (numeric: from 0 to 20)
33. G3 - final grade for maths (numeric: from 0 to 20, output target) - this is our **target variable**

The variables G1 and G2 were dropped as these are too highly correlated to the target variable G3, and therefore cannot provide any meaningful insights.









- G3 is our target variable, and follows a roughly Gaussian distribution that is skewed to the right. Students have a mean score of 10/20, which we would expect. There is a surprising number of students, 10%, who have received a score of 0/20. This could possibly be due to some administrative condition not being met, such as having failed to sit the exam, and will be explored in more depth.
- The boxplot shows us the interquartile range is 6, equally spread around the median of 11, meaning 50% of students achieved scores of 8-14.

The K-means clustering method was employed to convert the quantitative measurements to categorical variables, so that they could be analysed alongside the other variables using methods such as MCA. While quantisation will always lead to some loss of data, using K-means clustering helped us to preserve the underlying distribution of the data as much as possible.

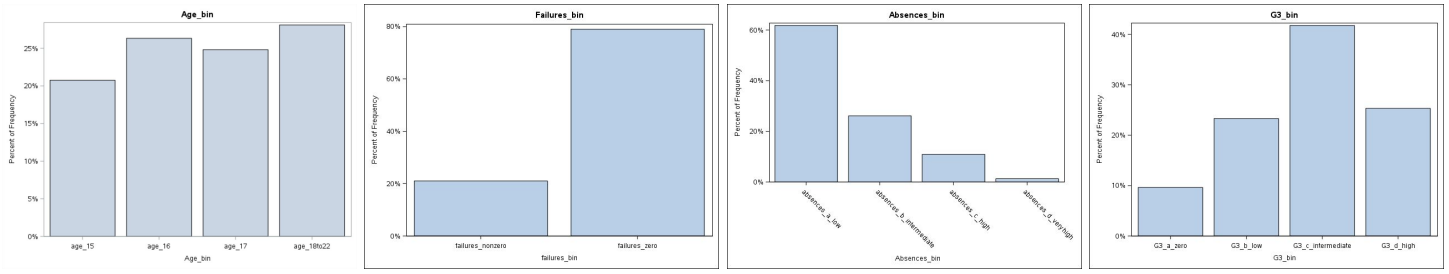
The following categorical variables were created using K-means, with K=4:

1. G3_bin - final grade for maths (nominal: “0 to 0”, “4 to 9”, “10 to 13”, “14 to 20”, output target)
2. Age_bin - student's age (nominal: “15 to 15”, “16 to 16”, “17 to 17”, “18 to 22”)
3. Absences_bin - number of school absences (nominal: “0 to 4”, “5 to 13”, “14 to 30”, “38 to 75”)

The following categorical variable was created by merging all non-zero failures into a second modality:

4. failures_bin - number of past class failures (nominal: zero, non-zero)

Here are the quantitative variables after quantisation into categorical variables:



- The quantisation process had the benefit of reducing the number of modalities for all variables. For instance, it created a roughly even distribution of ages across the 4 categories, which facilitated analysis of results.

Bi-dimensional analysis

Chi-square and Cramer's V

To discover the important relationships between pairs of variables, we used the chi-square and cramer's V analysis tools. Chi-square is a statistic that detects associations between categorical variables, to determine if they are related to each other in any way. Specifically, it compares the counts of categorical responses between two (or more) independent groups in a contingency table, and determines whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. The output, p, is a significance level that represents the likelihood of the observed frequencies occurring. A significance level close to zero means that the two variables are very unlikely to be completely unassociated in some population. However, this does not mean the variables are strongly associated; a weak association in a large sample size may also result in p = 0.000.

Cramer's V is a statistic that indicates how strongly two categorical variables are associated. It is the Pearson chi-square statistic rescaled to have values between 0 (no association) and 1 (perfect association). The interpretation of a Cramer's V can be made according to the following table:

Cramer's V	Verbal Description	COMMENTS
0.00	No Relationship	Knowing the independent variable does not help in predicting the dependent variable.
.00 to .15	Very Weak	Not generally acceptable
.15 to .20	Weak	Minimally acceptable
.20 to .25	Moderate	Acceptable
.25 to .30	Moderately Strong	Desirable
.30 to .35	Strong	Very Desirable
.35 to .40	Very Strong	Extremely Desirable
.40 to .50	Worrisomely Strong	Either an extremely good relationship or the two variables are measuring the same concept
.50 to .99	Redundant	The two variables are probably measuring the same concept.
1.00	Perfect Relationship.	If we the know the independent variable, we can perfectly predict the dependent variable.

As G3_bin (the categorical version of a student's final grade) was our target variable, chi-square and Cramer's V were calculated for all variables paired with G3_bin. When evaluated using the above mentioned criteria, the following variables were found to be correlated with G3_bin: Medu (mother's education); Failures (number of past failures); School Support (extra educational support); paid (extra paid classes); higher (wants to take higher education (yes/no)); Absences (number of school absences).

Statistics for Table of G3_bin by Medu

Statistic	DF	Value	Prob
Chi-Square	12	38.4026	0.0001
Likelihood Ratio Chi-Square	12	41.4299	<.0001
Mantel-Haenszel Chi-Square	1	18.5533	<.0001
Phi Coefficient		0.3118	
Contingency Coefficient		0.2977	
Cramer's V		0.1800	

Statistics for Table of G3_bin by failures_bin

Statistic	DF	Value	Prob
Chi-Square	3	54.5736	<.0001
Likelihood Ratio Chi-Square	3	57.4715	<.0001
Mantel-Haenszel Chi-Square	1	54.1370	<.0001
Phi Coefficient		0.3717	
Contingency Coefficient		0.3484	
Cramer's V		0.3717	

Statistics for Table of G3_bin by schoolsup

Statistic	DF	Value	Prob
Chi-Square	3	22.9471	<.0001
Likelihood Ratio Chi-Square	3	26.1428	<.0001
Mantel-Haenszel Chi-Square	1	3.9866	0.0459
Phi Coefficient		0.2410	
Contingency Coefficient		0.2343	
Cramer's V		0.2410	

Statistics for Table of G3_bin by paid

Statistic	DF	Value	Prob
Chi-Square	3	11.6037	0.0089
Likelihood Ratio Chi-Square	3	12.3502	0.0063
Mantel-Haenszel Chi-Square	1	3.5603	0.0592
Phi Coefficient		0.1714	
Contingency Coefficient		0.1689	
Cramer's V		0.1714	

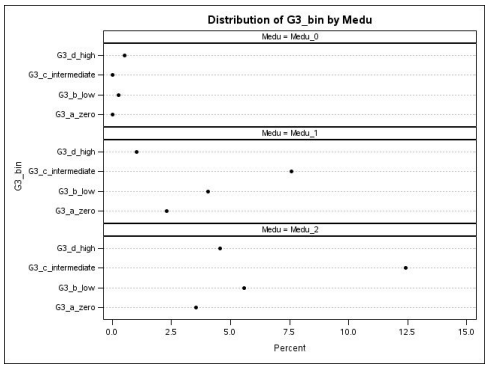
Statistics for Table of G3_bin by higher

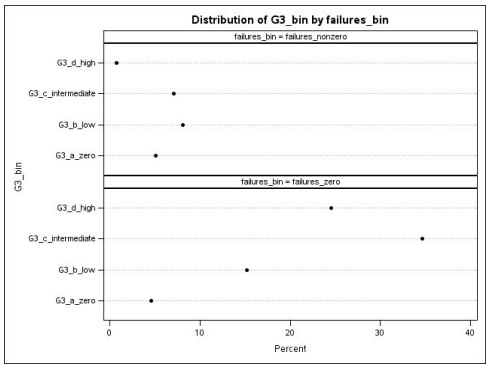
Statistic	DF	Value	Prob
Chi-Square	3	15.8998	0.0012
Likelihood Ratio Chi-Square	3	17.6927	0.0005
Mantel-Haenszel Chi-Square	1	15.1133	0.0001
Phi Coefficient		0.2006	
Contingency Coefficient		0.1967	
Cramer's V		0.2006	
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

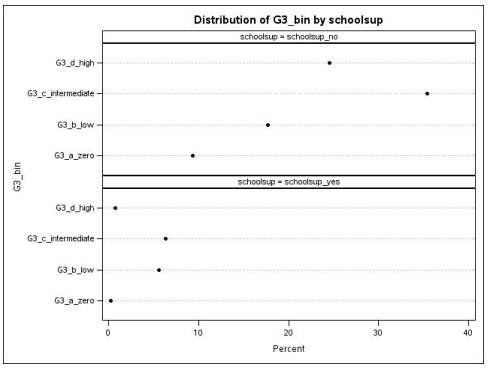
Statistics for Table of G3_bin by Absences_bin

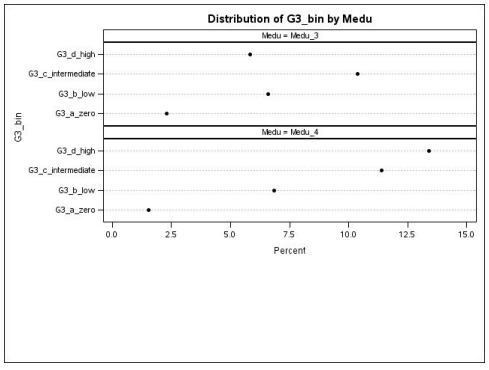
Statistic	DF	Value	Prob
Chi-Square	9	55.5360	<.0001
Likelihood Ratio Chi-Square	9	67.4335	<.0001
Mantel-Haenszel Chi-Square	1	0.1141	0.7355
Phi Coefficient		0.3750	
Contingency Coefficient		0.3511	
Cramer's V		0.2165	
WARNING: 31% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

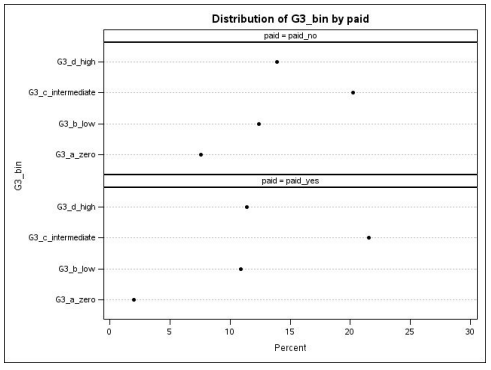
To determine the direction of the correlations, we plotted the categorical values on dot-plots:

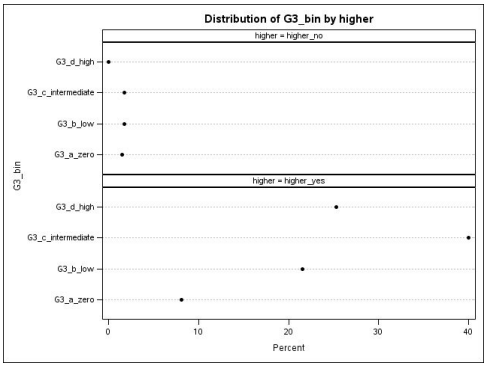


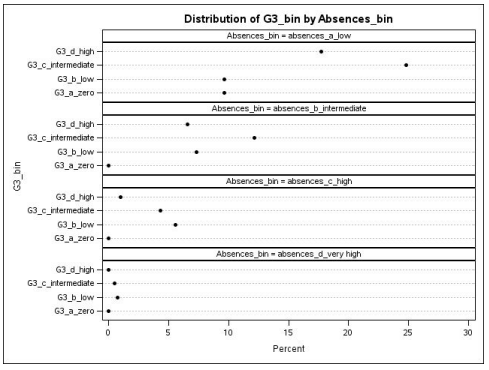












A summary of the strength (based on Cramer's V) and direction (based on the dot-plots) of the correlations:

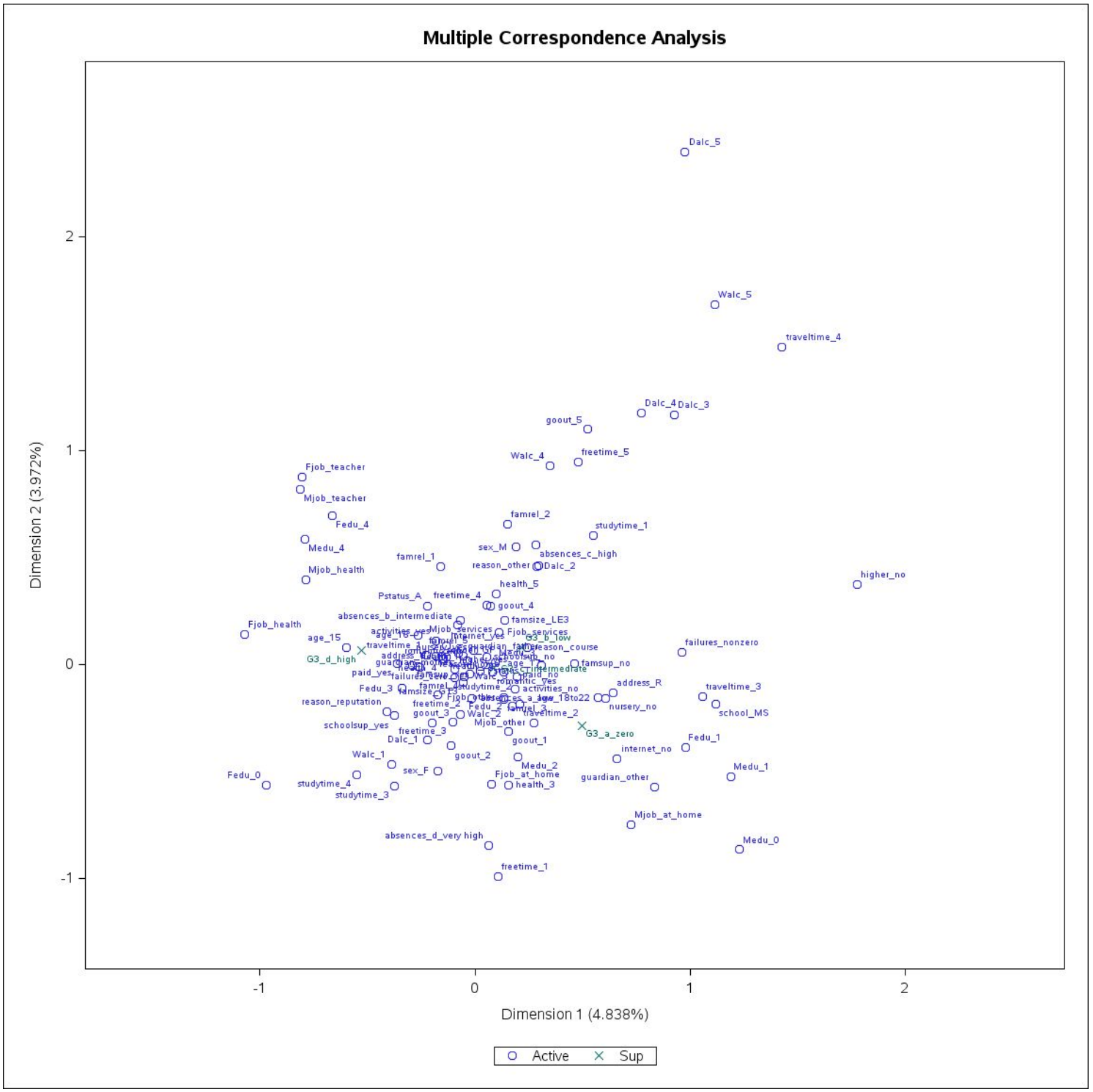
- Medu: mother's education - **weak positive** correlation (the higher the education level, the higher the grade)
- Failures: number of past failures - **very strong negative** correlation (non-zero failure suggests a lower grade)
- School Support: extra educational support - **moderate negative** correlation (receiving school support suggests a lower grade)
- paid: extra paid classes - it's **unclear**, as the dot-plots are almost identical
- higher: wants to take higher education (yes/no) - **moderate positive** correlation (wanting to take higher education suggests a higher grade)
- Absences: number of school absences - **moderate negative** correlation (the more absences, the lower the grade)

Multi-dimensional Analysis

Multiple Correspondence Analysis (MCA)

MCA allowed us to study both the individuals and the variables in the dataset. The method exposed the differences (variability) between the instances by finding dimensions (principal axes) that separate very different individuals from average individuals, based on categories of variables they belong to. In terms of the variables, two qualitative variables are linked if the categories of one have a connection with categories of the other. MCA allowed us to visualise these relations by creating quantitative variables based on qualitative variables (in a similar way to PCA).

The MCA plot was first generated using all of the variables available, specifying our target variable, G3_bin, as the supplementary variable (shown in green):



PROC CORRESP creates a Burt table from categorical data and performs a multiple correspondence analysis. The centre of gravity of any variable is the origin, and so the centre of gravity of the whole point cloud is also the origin. This is why the categories are mostly clustered around the origin.

The variance of category k is the square of the distance between category k and the origin is $(1/p_k) - 1$, where p_k is the sum of all values in that variable's category column (in the one-hot representation). This means rare categories are further from the origin, as the distance from the origin increases greatly as the category becomes less common. We see in the plot that very rare categories such as Dalc_5 (very high workday alcohol consumption) and Fjob_teacher (father's job is a teacher) are positioned at the far edges of the plot.

The inertia of category k is calculated as the squared distance from category k to the origin * weight of category = $(1-p_k)/J$ where J is the total number of variables. When a category is rare, the distance to the origin increases but its

weight decreases. Both of these factors work against each other, but overall, rare categories have high inertia. Having said this, MCA doesn't over-exaggerate influence of extremely rare categories either. The variables with the top 10 and bottom 10 levels of inertia were:

Obs	_NAME_	Inertia
1		2.36667
2	Fedu_0	0.01401
3	Medu_0	0.01398
4	absences_d_very high	0.01391
5	famrel_1	0.01380
6	traveltime_4	0.01380
7	Dalc_4	0.01376
8	Dalc_5	0.01376
9	Fjob_health	0.01344
10	famrel_2	0.01344
92	guardian_mother	0.00435
93	Dalc_1	0.00424
94	famsize_GT3	0.00406
95	address_U	0.00314
96	failures_zero	0.00296
97	nursery_yes	0.00289
98	internet_yes	0.00235
99	schoolsup_no	0.00182
100	school_GP	0.00164
101	Pstatus_T	0.00146
102	higher_yes	0.00071

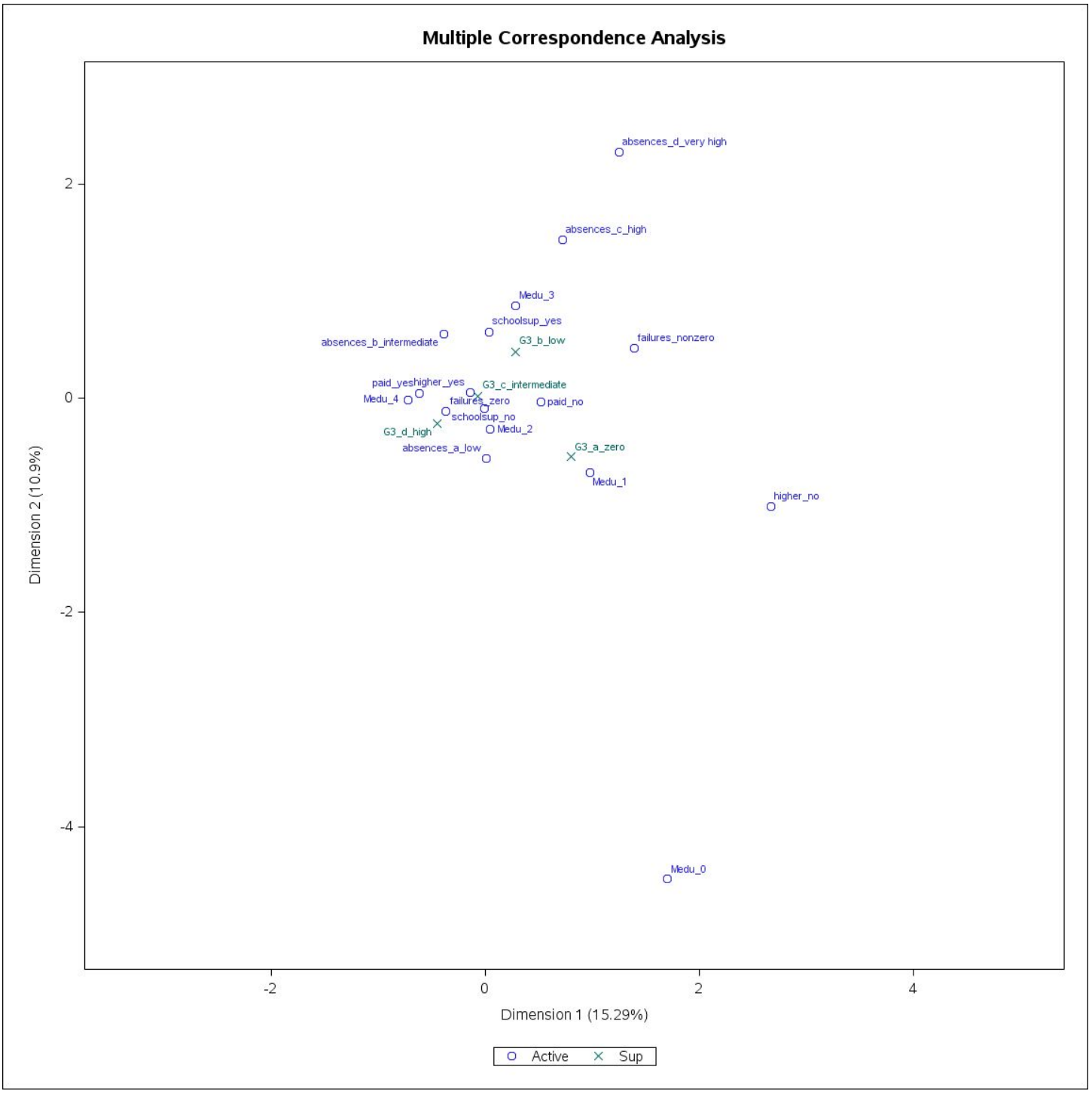
We saw, along with the variables that appeared at the edges of the plot, that the extreme categories such as absences_d_very_high and traveltime_4 had the highest inertia. By referencing the histograms plotted earlier, we confirmed that these are rare categories with few students in. Conversely, the categories with the lowest inertia were located around the origin, and are common amongst students.

The categories of the supplementary variable, G3_bin, are arranged in a gentle diagonal bottom-right to top-left pattern across the plot. It can be said that the closer an active category is to one of of these supplementary categories, the closer the relationship in the data. We see that at the top-left, categories relating to a high level of parent education, parents who work in the health industry, and good family relationships are related to G3_d_high (a final grade of between 14 and 20). In the bottom-right we see the opposite; low levels of parental education, guardians other than the parent, no internet at home, and no nursery education. Therefore we can say this first diagonal appears to relate to the **quality of the home environment.**

There is a second diagonal, perpendicular to the first, running from the bottom-left to the top-right. In the bottom-left we find high levels of study time, low levels of weekend alcohol consumption, extra school support received, but interestingly, also low levels of father's education. In the top-right we see the opposite; high workday and weekend alcohol consumption, high traveltime, high freetime after school and low study time. Therefore we can say this second diagonal appears to relate to the **student's personal habits.**

The distance between two categories k and k' is a function of p_k and p_k', and $p_{kk'}$ (the proportion of individuals that are in both categories k and k'). The more individuals there are that in only one of the two categories, the greater the distance between the categories. The more individuals two categories share, the closer they are to each other. We used this property to infer the following from the plot:

- Students who drink a lot of alcohol, go out to see their friends a lot, have a high number of absences or spend a long time travelling to/from school, generally do not spend much time studying.
- Male students are more likely to be involved in these distracting activities than female students.
- Students who live in a higher quality home environment and have more educated parents tend to score more highly in their exams.

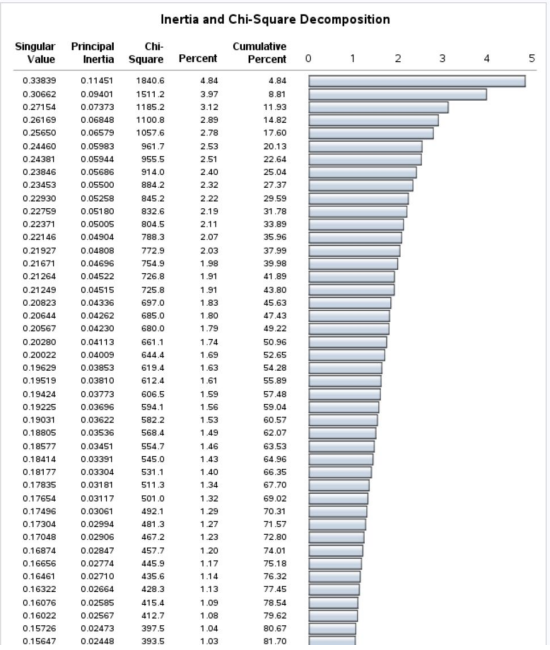


- A second MCA was performed using only the variables that were found to be correlated using the bidimensional analysis.
- Here we clearly see that the G3_bin grades are arranged in a bottom-left to top-right diagonal (ignoring G3_a_zero for the moment). Along the same diagonal the absences categories are arranged, with low absences corresponding to a high final grade, and vice-versa. Close to G3_d_high are categories for very high mother's education level, extra paid classes within the course subject, and a desire to attend higher education. Conversely, categories close to G3_b_low include non-zero exam failures, high mother's education, and extra support received in school.
- Along the perpendicular diagonal (bottom-right to top-left) are categories for low and very low mother's education levels, and to not attend higher education. Interestingly the zero G3 grades do not seem to be closely related to the level of absences, suggesting that another factor is at play here, e.g., perhaps students who do not want to attend university often decide not to sit the exams, because they have no need for the grades. More data would be required to discover the relationship between these variables.

Discriminant Factor Analysis

STEPDISC

The STEPDISC procedure performs a stepwise discriminant analysis to select a subset of the quantitative variables for use in discriminating among the classes. In our case, the classification variable is G3_bin, and the quantitative variables are the MCA dimension coefficients corresponding to our variable categories. The STEPDISC procedure is a useful prelude to further analyses with the CANDISC procedure, as it reduces the number of inputs.



Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent
0.33839	0.11451	1840.6	4.84	4.84
0.30662	0.09401	1511.2	3.97	8.81
0.27154	0.07373	1185.2	3.12	11.93
0.26169	0.06848	1100.8	2.89	14.82
0.25650	0.06579	1057.6	2.78	17.60
0.24460	0.05983	961.7	2.53	20.13
0.24381	0.05944	955.5	2.51	22.64
0.23846	0.05686	914.0	2.40	25.04
0.23453	0.05500	884.2	2.32	27.37
0.22930	0.05258	845.2	2.22	29.59
0.22759	0.05180	832.6	2.19	31.78
0.22371	0.05005	804.5	2.11	33.89
0.22146	0.04904	788.3	2.07	35.96
0.21927	0.04808	772.9	2.03	37.99
0.21671	0.04696	754.9	1.98	39.98
0.21264	0.04522	726.8	1.91	41.89
0.21249	0.04515	725.8	1.91	43.80
0.20823	0.04336	697.0	1.83	45.63
0.20644	0.04262	685.0	1.80	47.43
0.20567	0.04230	680.0	1.79	49.22
0.20280	0.04113	661.1	1.74	50.96
0.20022	0.04009	644.4	1.69	52.65
0.19629	0.03853	619.4	1.63	54.28
0.19519	0.03810	612.4	1.61	55.89
0.19424	0.03773	606.5	1.59	57.48
0.19225	0.03696	594.1	1.56	59.04
0.19031	0.03622	582.2	1.53	60.57
0.18805	0.03536	568.4	1.49	62.07
0.18577	0.03451	554.7	1.46	63.53
0.18414	0.03391	545.0	1.43	64.96
0.18177	0.03304	531.1	1.40	66.35
0.17835	0.03181	511.3	1.34	67.70
0.17654	0.03117	501.0	1.32	69.02
0.17496	0.03061	492.1	1.29	70.31
0.17304	0.02994	481.3	1.27	71.57
0.17048	0.02906	467.2	1.23	72.80
0.16874	0.02847	457.7	1.20	74.01
0.16656	0.02774	445.9	1.17	75.18
0.16461	0.02710	435.6	1.14	76.32
0.16322	0.02664	428.3	1.13	77.45
0.16076	0.02585	415.4	1.09	78.54
0.16022	0.02567	412.7	1.08	79.62
0.15726	0.02473	397.5	1.04	80.67
0.15647	0.02448	393.5	1.03	81.70

Using the inertia and chi-square output from the MCA procedure, we see that 80% of the inertia in the system is represented by the first 43 dimensions. We will work only with these dimensions from here on.

MCA outputs the quantitative coordinate vectors for each individual across these 43 dimensions, which were used as input to STEPDISC. A stepwise discriminant analysis was performed to select a subset of the quantitative variables for use in discriminating among the classes.

The STEPDISC Procedure

The Method for Selecting Variables is FORWARD			
Total Sample Size	395	Variable(s) in the Analysis	43
Class Levels	4	Variable(s) Will Be Included	0
		Significance Level to Enter	0.15

Number of Observations Read	395
Number of Observations Used	395

Class Level Information

G3_bin	Variable Name	Frequency	Weight	Proportion
G3_a_zero	G3_a_zero	38	38.0000	0.096203
G3_b_low	G3_b_low	92	92.0000	0.232911
G3_c_intermediate	G3_c_intermediate	165	165.0000	0.417722
G3_d_high	G3_d_high	100	100.0000	0.253165

The STEPDISC Procedure

Forward Selection Summary

Step	Number In	Entered	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	Dim1	0.1080	15.79	<.0001	0.89195619	<.0001	0.03601460	<.0001
2	2	Dim4	0.0493	6.74	0.0002	0.84796192	<.0001	0.05179267	<.0001
3	3	Dim43	0.0436	5.92	0.0006	0.81096594	<.0001	0.06506235	<.0001
4	4	Dim7	0.0395	5.32	0.0013	0.77891452	<.0001	0.07796075	<.0001
5	5	Dim33	0.0301	4.00	0.0080	0.75548932	<.0001	0.08726038	<.0001
6	6	Dim21	0.0293	3.89	0.0093	0.73333989	<.0001	0.09548747	<.0001
7	7	Dim35	0.0299	3.95	0.0085	0.71144361	<.0001	0.10380996	<.0001
8	8	Dim24	0.0267	3.51	0.0155	0.69247928	<.0001	0.11170576	<.0001
9	9	Dim40	0.0269	3.53	0.0151	0.67385127	<.0001	0.11960452	<.0001
10	10	Dim9	0.0261	3.42	0.0175	0.65624759	<.0001	0.12680639	<.0001
11	11	Dim42	0.0214	2.78	0.0407	0.64217160	<.0001	0.13244729	<.0001
12	12	Dim8	0.0216	2.80	0.0400	0.62830094	<.0001	0.13819721	<.0001
13	13	Dim32	0.0184	2.37	0.0702	0.61673235	<.0001	0.14379105	<.0001
14	14	Dim38	0.0148	1.89	0.1312	0.60762919	<.0001	0.14817071	<.0001

The 14 dimensions (variable categories) found with STEPDISC have potential discriminatory power, and can be used as input to the CANDISC procedure.

CANDISC

Canonical discriminant analysis is a dimension-reduction technique that finds linear combinations of quantitative variables that provide maximal separation between classes or groups. The CANDISC procedure derives canonical variables, which are linear combinations of the quantitative variables that summarize inter-class variation. The quantitative variables in our case are the coefficients of the individuals in the dimensions discovered with the CORRESP procedure. We added to this our classification variable, G3_bin, to enable visualisation of classes. CANDISC outputs canonical coefficients both and scored canonical variables.

Input all 43 dimensions

To begin with, we used all 43 dimensions as input to PROC CANDISC. To find the three canonical variables that best separate the students in terms of final grade, we set NCAN=3 so only the first three canonical variables are calculated.

A canonical correlation is performed on the discriminating predictor variables (the MCA coefficients) and the set of dummy variables generated from the class variable, G3_bin. The result is a set of canonical correlation values, where the first canonical correlation is the greatest possible multiple correlation with the classes that can be achieved by using a linear combination of the quantitative variables. In the student data this was found to be 0.570304. We see from the P-values that only the first two canonical correlations are statistically significant.

Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				Test of H0: The canonical correlations in the current row and all that follow are zero				
				Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
0.570304	0.500417	0.033994	0.325247	0.4820	0.2164	0.5802	0.5802	0.49218298	2.17	129	1046.7	<.0001
0.458102	0.370003	0.039807	0.209857	0.2656	0.1824	0.3197	0.8998	0.72942658	1.42	84	700	0.0105
0.277204	0.114102	0.046508	0.076842	0.0832		0.1002	1.0000	0.92315815	0.71	41	351	0.9072

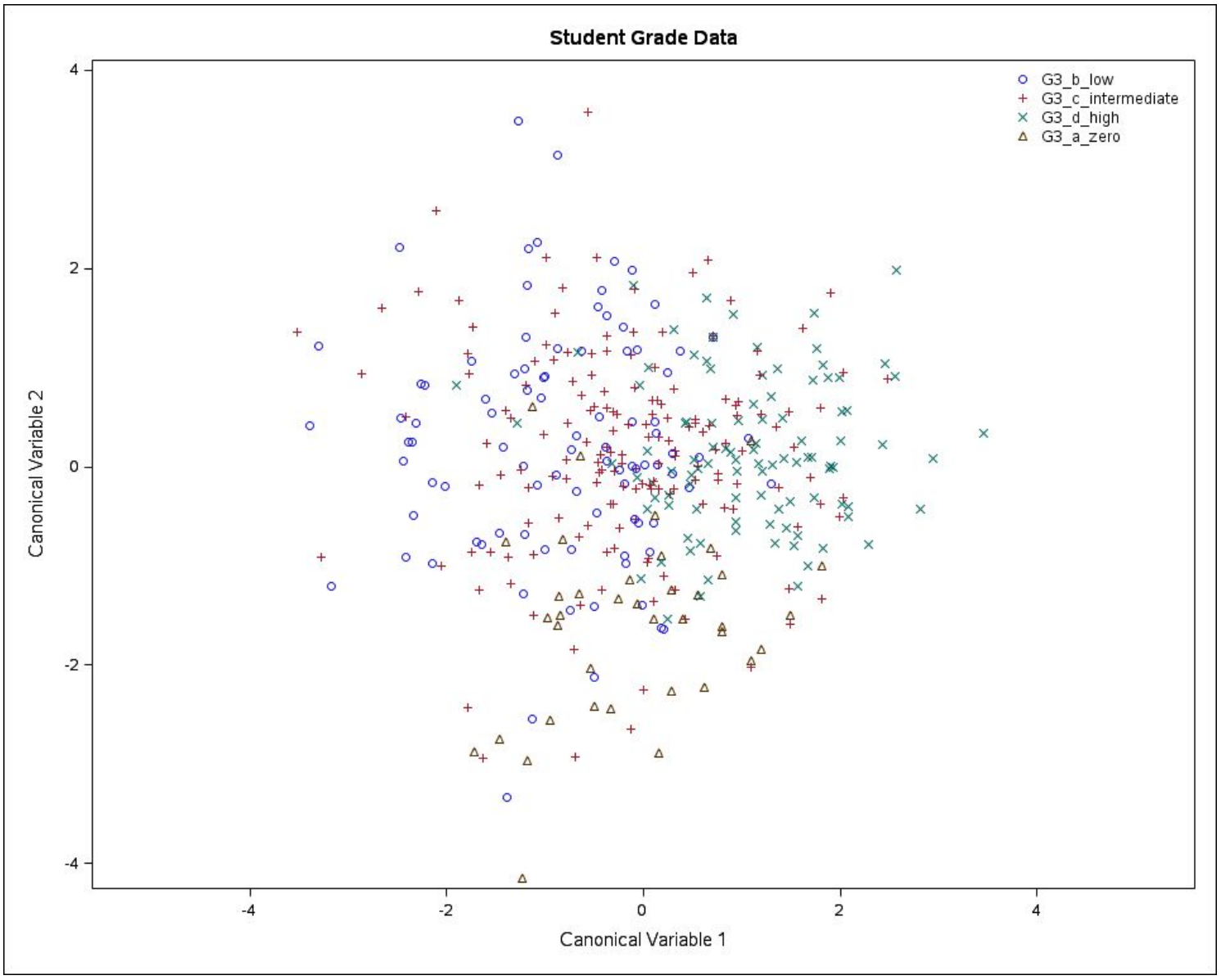
The squares of the canonical correlations can be interpreted similarly to R^2 coefficient of determination, in that they are the proportion of the variance in the canonical variate of one set of variables explained by the canonical variate of the other set of variables. Here we see that for the first canonical correlation, only 32.5% of the variance of the class variable G3_bin is explained by the discriminating predictor variables (the MCA coefficients). This is not as high as we would hope, and shows that much of the variance of the student's grades is not represented by the variables in the dataset. For the second canonical correlation, only 21.0% of the variance is explained. Nevertheless, we will visualise the result to identify any patterns that exist.

Raw Canonical Coefficients			
Variable	Can1	Can2	Can3
Dim1	-1.800717469	-1.083190689	-0.626243020
Dim2	0.063037027	0.752296655	0.490651818
Dim3	-0.052737814	0.233383838	0.779599042
Dim4	1.593863593	-0.675777346	-0.847502249
Dim5	-0.696463922	0.138505042	-0.048363403
Dim6	-0.701303755	0.548814955	0.372595129
Dim7	-0.058099287	-1.762631426	1.345708078
Dim8	1.001348680	0.307502324	-0.953974188
Dim9	-1.053609712	0.719106323	-0.915490619
Dim10	-0.492421334	0.250485435	0.446033477
Dim11	0.363264132	0.449951725	-0.834839034
Dim12	-0.293739074	-0.462868132	-0.340424671
Dim13	0.066206090	0.277406153	0.338656282
Dim14	-0.786320433	0.459177979	-0.685646363
Dim15	0.086281706	0.896120869	-0.291241228
Dim16	0.365303210	0.336830685	0.234972269
Dim17	0.795650921	-0.512925199	0.697470231

The first canonical variable, Can1, is the linear combination of the centered variables with the raw canonical coefficients that separates the students most effectively. From the table on the right, this is calculated as:

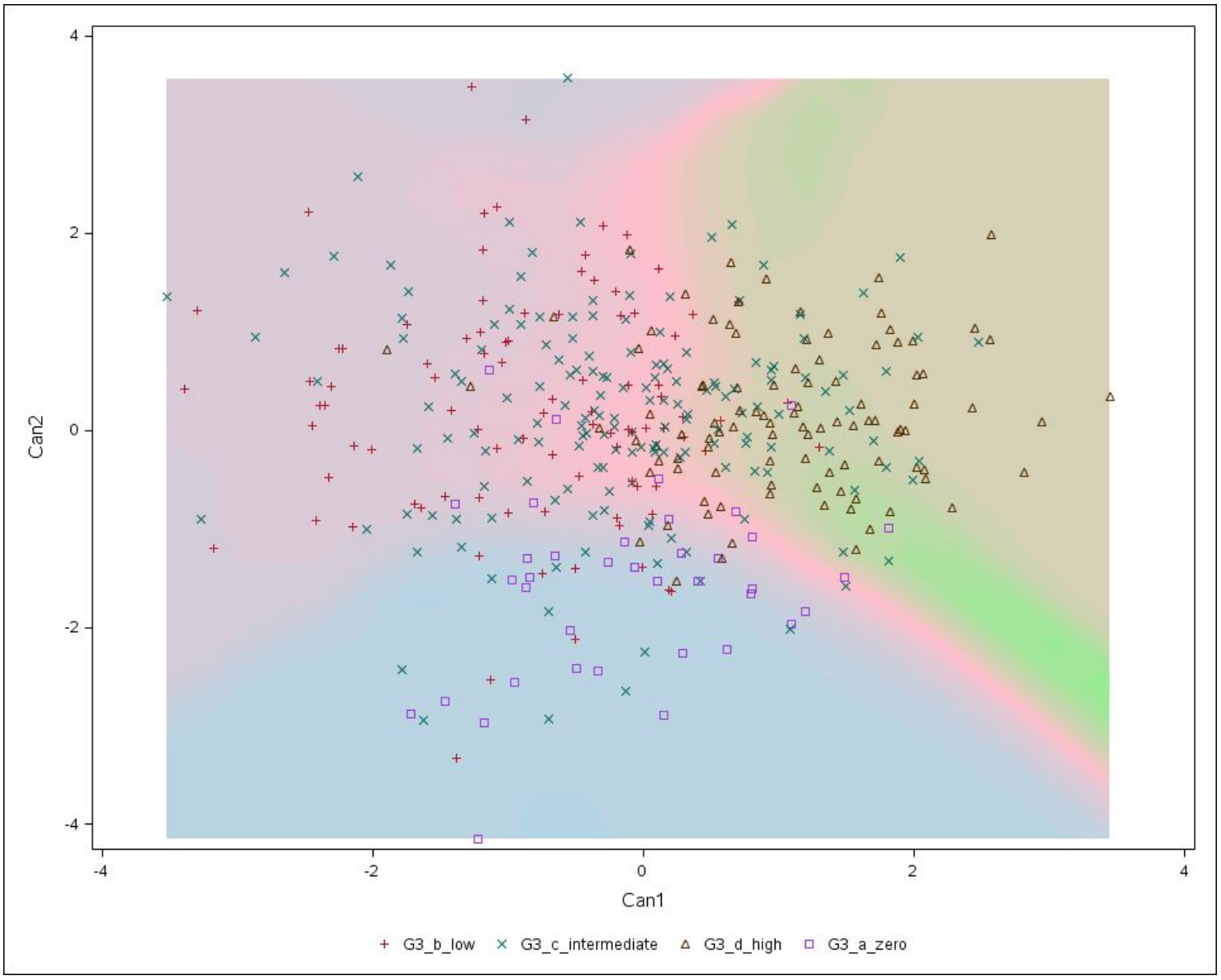
```
Can1 = -1.800*dim1 + 0.063*dim2 + ... + 2.46*dim43.
```

By applying the linear combination of these 43 coefficients to the dimensions of each of the individuals, and plotting the resulting Can1 and Can2 for each, we could observe the ability of the coefficients to discriminate between classes of student grades.



There is a clear left-to-right horizontal separation of students with low grades (blue), intermediate grades (red) and high grades (green). This shows that canonical variable 1 (Can1) does discriminate between students based on their final grade.

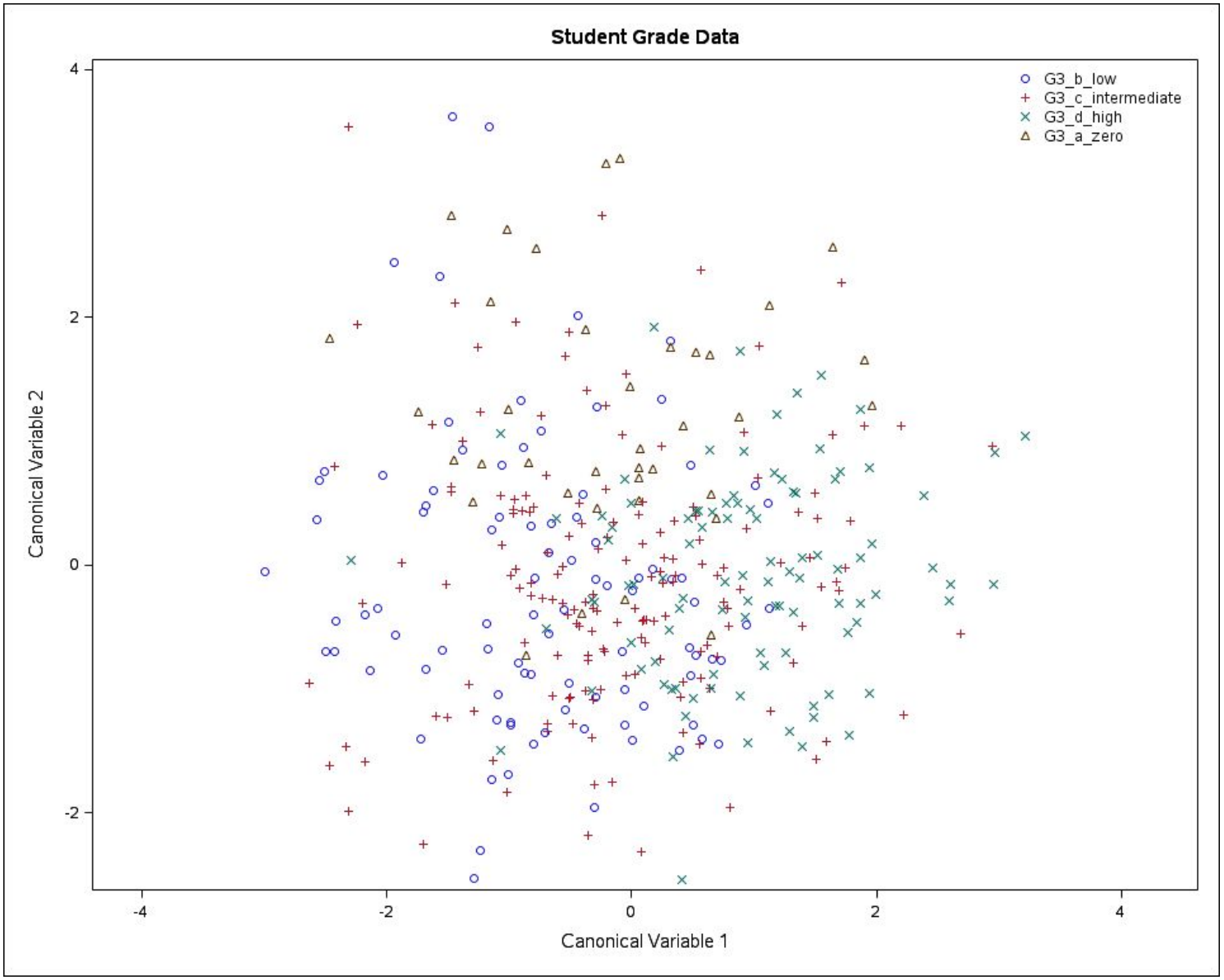
The students with zero grade (black) are clustered in the lower half of the plot, with negative values for canonical variable 2 (Can2). This shows that Can2 discriminates well between students who received a zero grade and all other students who received some nonzero grade. The plot also shows that these students are relatively central in the horizontal, meaning that Can1 cannot discriminate this class. This also corroborates our earlier findings that the factors causing students to receive a zero grade are quite different from those causing students to receive a low grade (e.g. absences).



A contour plot highlights the separation between classes. The horizontal left-to-right progression from low to intermediate to high grade is seen from pink to green. The students with a zero grade are clearly separated in the blue region below.

Input 14 selected dimensions

The same process was then performed using only the 14 dimensions selected by the STEPDISC procedure:



The resulting plot still displays the horizontal separation between low, intermediate and high grade students. This shows that canonical variable 1 retains its ability to discriminate even when calculated using 14 dimensions as opposed to 43. However, canonical variable 2 has lost its ability to discriminate the zero graded students, as we now see the black triangle markers spread over the entire vertical range of the plot. This shows that we have lost some important variance in the data by selecting a subset of the dimensions.

Classification using DISCRIM

The DISCRIM function performs a similar operation to the CANDISC function, but also outputs a classification table based on the linear discriminant function. Values in the diagonal of the classification table reflect the correct classification of individuals into groups based on their scores on the discriminant dimensions.

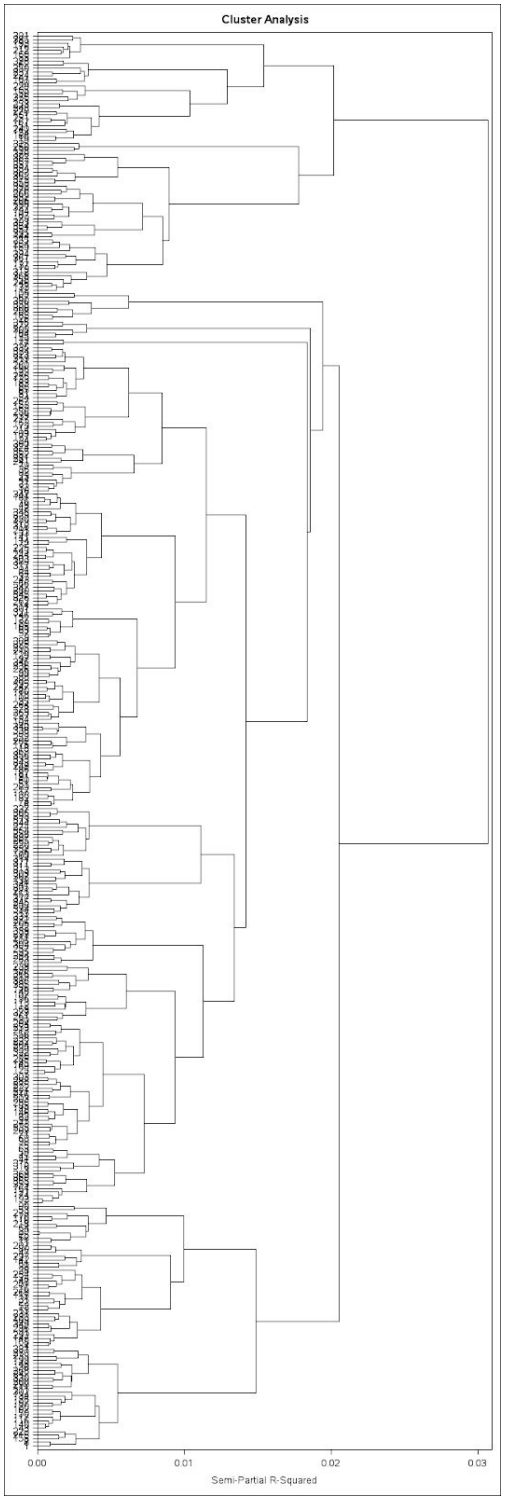
The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.DISCRIM_INPUT
Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into G3_bin					
From G3_bin	G3_a_zero	G3_b_low	G3_c_intermediate	G3_d_high	Total
G3_a_zero	31 81.58	3 7.89	2 5.26	2 5.26	38 100.00
G3_b_low	12 13.04	51 55.43	22 23.91	7 7.61	92 100.00
G3_c_intermediate	21 12.73	48 29.09	63 38.18	33 20.00	165 100.00
G3_d_high	6 6.00	5 5.00	18 18.00	71 71.00	100 100.00
Total	70 17.72	107 27.09	105 26.58	113 28.61	395 100.00
Priors	0.25	0.25	0.25	0.25	

Note that this is a resubstitution error (training error), not a cross-validation error, so the values are artificially high. We can summarise the results as follows:

- Zero grade: 81.58% correctly classified
- Low grade: 55.43% correctly classified
- Intermediate grade: 38.18% correctly classified
- High grade: 71.00% correctly classified

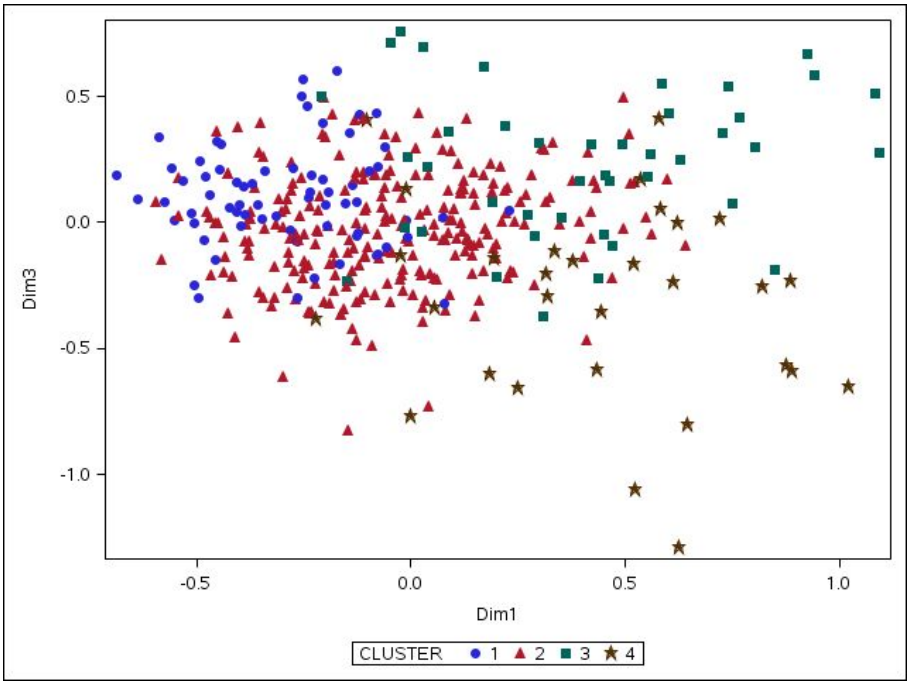
The intermediate graded students have the lowest classification percentage as these individuals lie in the centre of the point cloud, and therefore in centre of the multidimensional space. This means that they are surrounded by and mixed with individuals from other classes to a much greater extent than the other classes, with many more individuals being located on or near a decision boundary.



Classification using Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering (AHC) is a bottom-up clustering (classification) method which works by measuring the dissimilarity between the objects to decide which to group together. The resulting dendrogram showed the progressive grouping of the data, so it was possible to gain an idea of a suitable number of classes into which the data can be grouped. Here we used the euclidean distance as a measure of dissimilarity, and the Ward clustering criterion.

From the dendrogram we can see there are approximately 4 main clusters of data, separated by large vertical increases in distance before being clustered together. We used K=4 to cut the dendrogram and assign individuals to the clusters, which are plotted below against dimensions 1 and 3:



We can clearly see that the ACH method has created four clusters whose variance is mostly represented by dimension 1, as indicated by the horizontal separation of clusters. There is a certain amount of vertical separation across dimension 3 too, most noticeable with clusters 3 and 4.

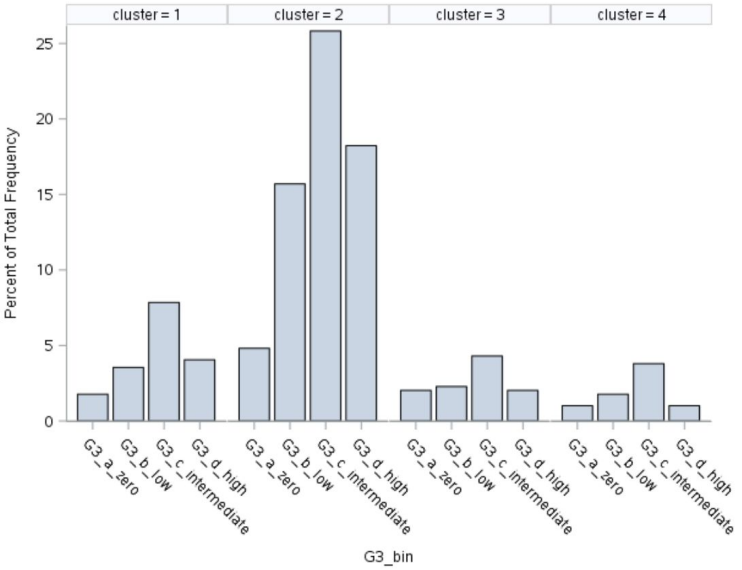
This pattern confirms that our canonical discriminant coefficients are effective at distinguishing between clusters of students. We can explore the membership of each cluster to determine its characteristics:

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of G3_bin by cluster

G3_bin	cluster 1	2	3	4	Total
G3_a_zero	7 1.77 18.42 10.29	19 4.81 50.00 7.45	8 2.03 21.05 19.05	4 1.01 10.53 13.33	38 9.62
G3_b_low	14 3.54 15.22 20.59	62 15.70 67.39 24.31	9 2.28 9.78 21.43	7 1.77 7.61 23.33	92 23.29
G3_c_intermediate	31 7.85 18.79 45.59	102 25.82 61.82 40.00	17 4.30 10.30 40.48	15 3.80 9.09 50.00	165 41.77
G3_d_high	16 4.05 16.00 23.53	72 18.23 72.00 28.24	8 2.03 8.00 19.05	4 1.01 4.00 13.33	100 25.32
Total	68 17.22	255 64.56	42 10.63	30 7.59	395 100.00



- By examining the column percentages, we see that there is a significant different in the composition of the clusters. Cluster 2 is made up of 7.45% of zero graded students, whereas 19.05% of cluster 3 students received a zero grade. In contrast, 28.24% of Cluster 2 are high grade students, whereas only 13.33% of cluster 4 received a high grade.
- We also see that every cluster has a roughly similar percentage of low and intermediate students.
- These findings show that the CAH clustering has some success at discriminating the students at the extremes (zero and high) from those in the centre (low and intermediate), but does poorly at discriminating those in the centre from each other. It is therefore the worst performing of the three classification methods.

Classification using K-Nearest Neighbours

K-Nearest Neighbours (KNN) makes predictions using the training dataset directly; there is no training step. Predictions are made for an individual by searching through the entire training set for the K nearest individuals based on a distance measure, and classifying the individual using a majority vote of the classes of the neighbours.

For the students data, we chose to use cross validation to discover the true classification accuracy. SAS cross validation option splits the dataset into 10 folds, chooses 1 fold as the "test set", and searches for neighbors in the remaining 9. It then repeats this for each of the other folds, and combines the results.

To tune the parameter K, the classification process was run using a variety of values for K. The highest cross validated classification accuracy was found with K=3:

- Zero grade: 21.05% correctly classified
- Low grade: 30.43% correctly classified
- Intermediate grade: 32.73% correctly classified
- High grade: 38.00% correctly classified

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.DISCRIM_INPUT
Cross-validation Summary using 3 Nearest Neighbors

Number of Observations and Percent Classified into G3_bin

From G3_bin	G3_a_zero	G3_b_low	G3_c_intermediate	G3_d_high	Total
G3_a_zero	8 21.05	8 21.05	14 36.84	8 21.05	38 100.00
G3_b_low	20 21.74	28 30.43	31 33.70	13 14.13	92 100.00
G3_c_intermediate	27 16.36	51 30.91	54 32.73	33 20.00	165 100.00
G3_d_high	18 18.00	18 18.00	26 26.00	38 38.00	100 100.00
Total	73 18.48	105 26.58	125 31.65	92 23.29	395 100.00
Priors	0.25	0.25	0.25	0.25	

Error Count Estimates for G3_bin

	G3_a_zero	G3_b_low	G3_c_intermediate	G3_d_high	Total
Rate	0.7895	0.6957	0.6727	0.6200	0.6945
Priors	0.2500	0.2500	0.2500	0.2500	

We see that the classification accuracy for the high grade students is the best, likely due to the relative homogeneity of the cluster seen in the MCA plot of individuals earlier.