Midterm Test WQD7005

Student id: s2141070

Name: Loh Cin Ceat

Our Assignment Dataset

In this project student performance data is obtained from the UCI Machine Learning Repository website: https://archive.ics.uci.edu/ml/datasets/Student+Performance. This data consists of student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. We had sampled the data which consists of 396 rows of data and 16 columns of variables regarding the performance in Mathematics subject. The table below shows the features present in the dataset and their descriptions.

No	Feature	Data type	Description
1	sex	Categorical	student's sex • F = female • M = male
2	age	Numeric	student's age • 15 to 22
3	address	Categorical	student's home address type • U = urban • R = rural
4	Medu	Numeric	mother's education • 0 = none • 1 = primary education • 2 = 5th to 9th grade • 3 = secondary education • 4 = higher education
5	Fedu	Numeric	father's education • 0 = none

			• 1 = primary education • 2 = 5th to 9th grade • 3 = secondary education • 4 = higher education
6	Mjob	Categorical	mother's job • teacher, health care related • civil services • at home • other
7	Fjob	Categorical	Father's job • teacher, health care related • civil services • at home • other
8	studytime	Numeric	weekly study time • 1 = <2 hours • 2 = 2 to 5 hours • 3 = 5 to 10 hours • 4 = >10 hours
9	failures	Numeric	number of past class failures • n if $1 <= n < 3$ • else 4
10	famsup	Categorical	family educational support • yes • no
11	internet	Categorical	Internet access at home • yes • no
12	freetime	Numeric	free time after school

			• 1 = very low to 5 = very high
13	goout	Numeric	going out with friends • 1 = very low to 5 = very high
14	health	Numeric	current health status • 1 = very bad to 5 = very good
15	absences	Numeric	number of school absences 0 to 93
16	G3	Numeric	final grade 0 to 20, output target

1. Use your group assignment dataset and draw a schema diagram for the data warehouse using Star schema.

To draw a star schema few attributes are added in to our dataset which includes

17	Student_id	Categorical	Student id for the students records
18	Subject_id	Categorical	Subject id for the subjects records
19	School_id	Categorical	School id for the schools records
20	Exam_id	Categorical	Exam id for the exams records
21	Subject_names	Categorical	• portuguese • mathematics
22	School_names	Categorical	• "AP" = abriel Pereira • "MS" = Mousinho da Silveira
23	School_address	Categorical	School's address type • U = urban • R = rural
24	First_exam_score	Numeric	• First exam grade 0 to 20
24	Mid_exam_score	Numeric	• Mid exam grade 0 to 20
24	Final_exam_score	Numeric	• Final exam grade 0 to 20

The star schema is then sketched as diagram 1.0 below.

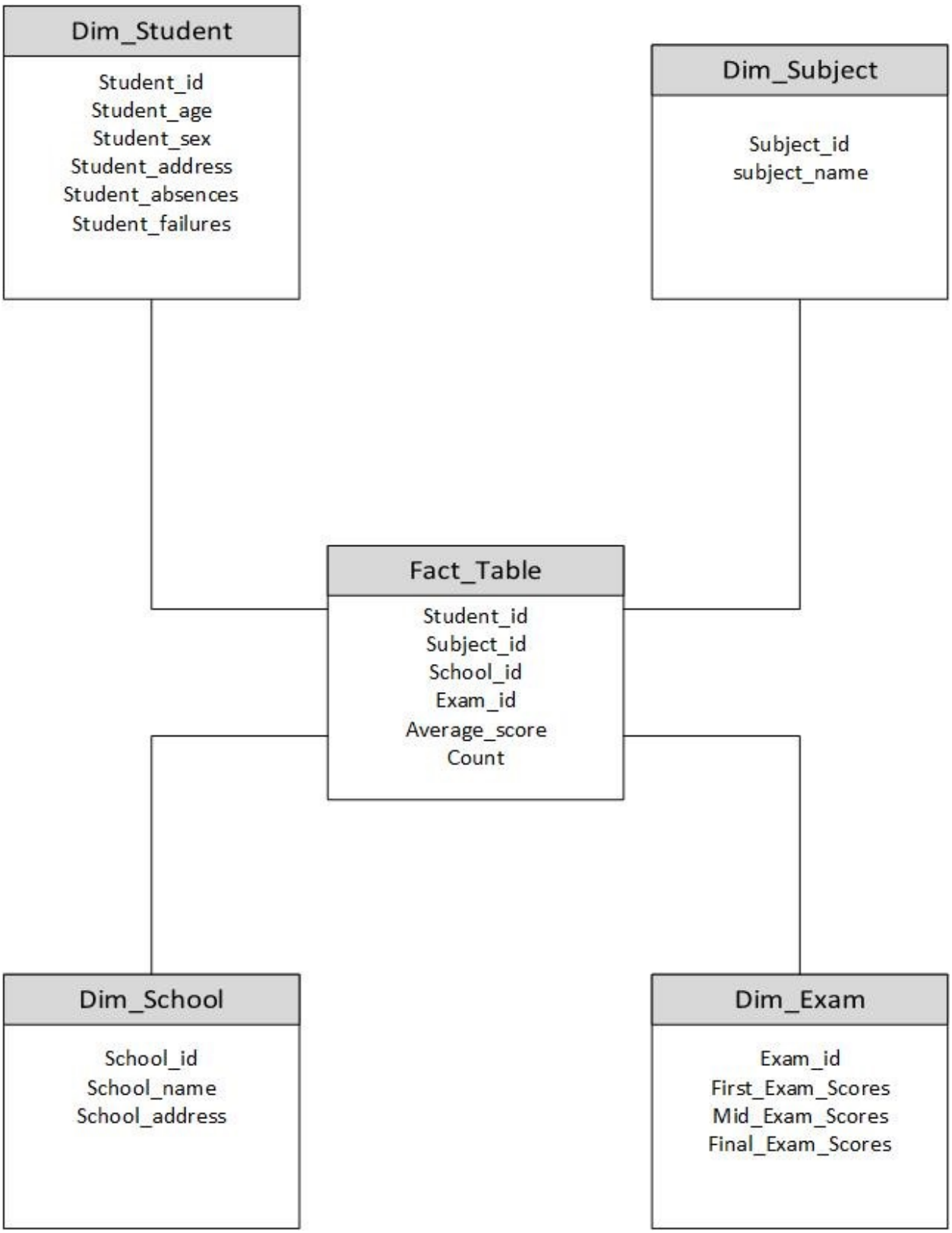


Diagram 1.0: Star Schema diagram for Student Performance Dataset

2. Define your words of Data Mining.

Data mining is a process used to extract usable data from a larger batch of any raw data by extracting and discovering patterns in the large data sets using one or more software where machine learning, statistics, and database systems often are used.

3. Discuss the differences between database and data warehouse with own three examples which are based on group assignment dataset.

A database is any collection of data that has been structured for storage, access, and retrieval. A data warehouse is a sort of database that consolidates copies of transaction data from several source systems and makes them available for analysis. For example, our assignment dataset is a small database which contains all the data information for the student performance records. However, we can see the example of data warehouse form the star schema above in diagram 1.0 where the student performance data is a stored in several sources like school information, exam information, student information and subject information.

Moreover, data in databases has been normalised. The purpose of normalisation is to decrease, if not eliminate, data redundancy, which is the practise of storing the same piece of data more than once. This decrease in duplicate data results in enhanced consistency and, as a result, more accurate data because the database saves it just once. Normalizing data divides it into several tables. Each table represents a distinct data object. For example, our assignment dataset is a database documenting Student performance. However, after normalization, the results datawarehouse will consists of **school** information, **exam** information, **student** information and **subject** information datasets.

A database is designed to update (add, alter, or remove) data as quickly and efficiently as possible. Database response times must be exceedingly fast for efficient transaction processing. Data warehouses are designed to conduct a small number of complicated queries on big multidimensional information quickly. For example our student performance dataset can be easily be processed by adding, altering, or remove any attributes or row of observations when there is a change in any of the information of the student performance dataset. However, for data warehouse, the data is stored in several sources with the split information of the database which makes any changes to the data to be difficult since all the sources are interconnected where while altering one source others source needed to be amended simultaneously. On the other hand, the data warehouse with split information will make the retrieval of information of students only more quickly.

3. Data quality can be assessed in terms of several issues, including incomplete, noisy, inconsistent and intentional. Show any two issues in your group assignment dataset.

By applying boxplot graph in SAS enterprise miner, we found that the **absences** attributes have noisy data looking at the diagram 2.0. From the box plot below we can see that there are many outlier with student absences greater than 20 days until almost 80 days which may be abnormal or noisy data since student with such amount of absences. It may cause by error in data entry in surveying.

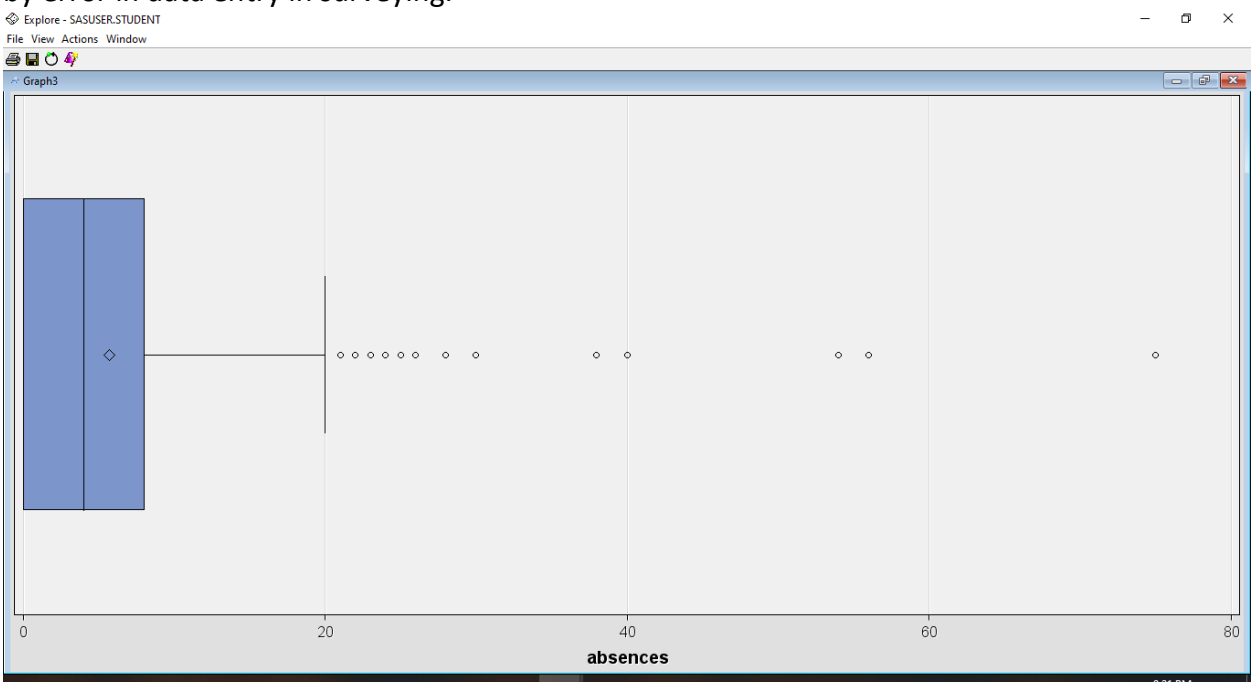
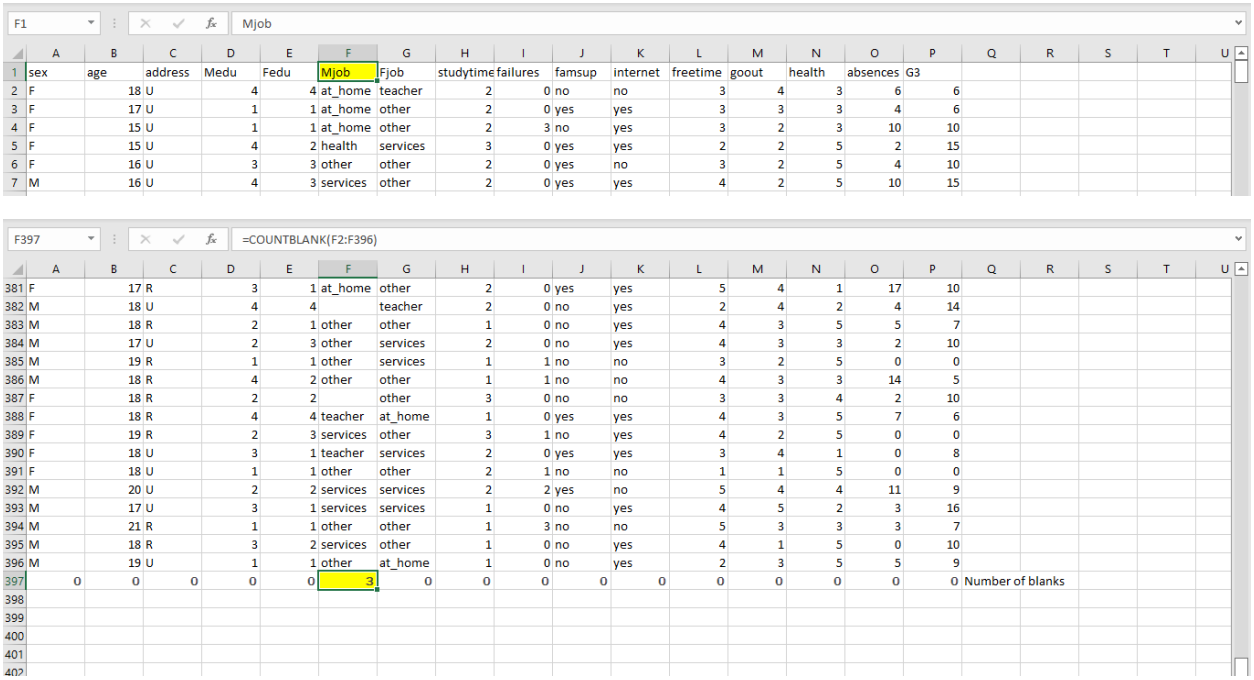


Diagram 2.0: Boxplot for absences attributes for the Student performance dataset of our assignment.

Moreover, from diagram 3.0 by using excel **Count blanks function,** 3 empty cells are found in the Mjob attributes.



F1 | Mjob

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	sex	age	address	Medu	Fedu	Mjob	Fjob	studytime	failures	famsup	internet	freetime	goout	health	absences	G3					
2	F	18	U	4	4	at_home	teacher	2	0	no	no	3	4	3	6	6					
3	F	17	U	1	1	at_home	other	2	0	yes	yes	3	3	3	4	6					
4	F	15	U	1	1	at_home	other	2	3	no	yes	3	2	3	10	10					
5	F	15	U	4	2	health	services	3	0	yes	yes	2	2	5	2	15					
6	F	16	U	3	3	other	other	2	0	yes	no	3	2	5	4	10					
7	M	16	U	4	3	services	other	2	0	yes	yes	4	2	5	10	15					

F397 | =COUNTBLANK(F2:F396)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
381	F	17	R	3	1	at_home	other	2	0	yes	yes	5	4	1	17	10					
382	M	18	U	4	4		teacher	2	0	no	yes	2	4	2	4	14					
383	M	18	R	2	1	other	other	1	0	no	yes	4	3	5	5	7					
384	M	17	U	2	3	other	services	2	0	no	yes	4	3	3	2	10					
385	M	19	R	1	1	other	services	1	1	no	no	3	2	5	0	0					
386	M	18	R	4	2	other	other	1	1	no	no	4	3	3	14	5					
387	F	18	R	2	2		other	3	0	no	no	3	3	4	2	10					
388	F	18	R	4	4	teacher	at_home	1	0	yes	yes	4	3	5	7	6					
389	F	19	R	2	3	services	other	3	1	no	yes	4	2	5	0	0					
390	F	18	U	3	1	teacher	services	2	0	yes	yes	3	4	1	0	8					
391	F	18	U	1	1	other	other	2	1	no	no	1	1	5	0	0					
392	M	20	U	2	2	services	services	2	2	yes	no	5	4	4	11	9					
393	M	17	U	3	1	services	services	1	0	no	yes	4	5	2	3	16					
394	M	21	R	1	1	other	other	1	3	no	no	5	3	3	3	7					
395	M	18	R	3	2	services	other	1	0	no	yes	4	1	5	0	10					
396	M	19	U	1	1	other	at_home	1	0	no	yes	2	3	5	5	9					
397	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	Number of blanks				
398																					
399																					
400																					
401																					
402																					

Diagram 3.0 Excel spreadsheet showing the number of blanks observation for each attributes.