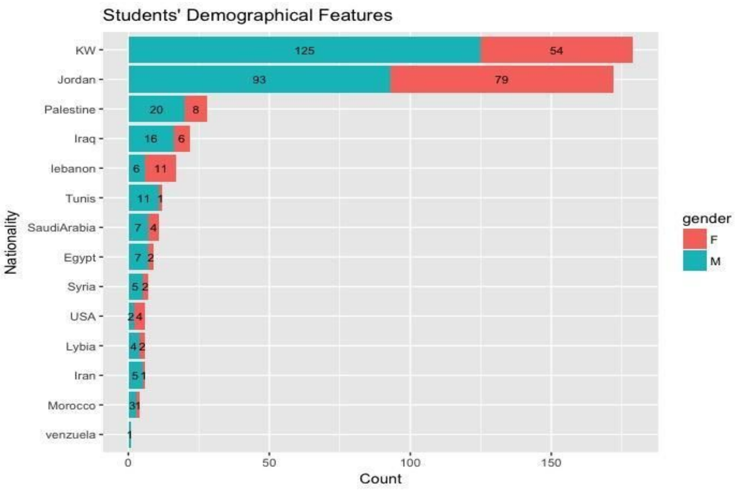
Problem 2: Students' Academic Performance Dataset (xAPI-Edu-Data)

In this work, a data analytics framework was proposed to predict students' grade including three parts ranging from data visualization to data preprocessing and to model building and evaluation. Please see below to find the detailed description of each part.

Part 1: Feature exploration and visualization to understand the dataset with R

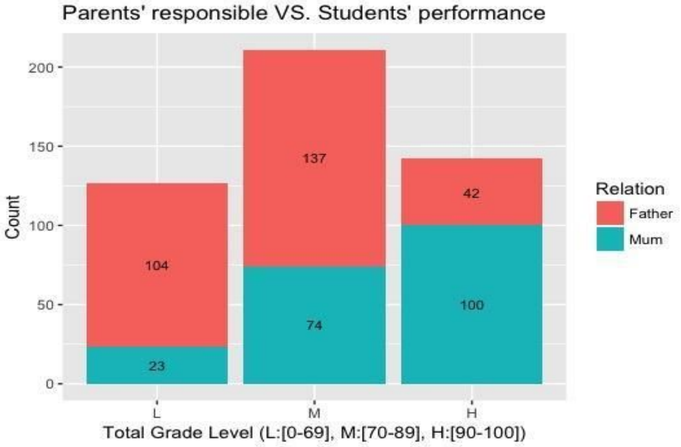
The data was collected through a learner tracker tool from 480 students records and 17 features including feature categories such as demographics, academy, parent participation, behavior and performance. For detailed feature list, please see Appendix Tab.1. In order to better understand the data, several data visualization was conducted, and some questions can be answered through the figures below. (Please check EDA_code.r for figure plot).

(1) Where are those students coming from? How many boys and girls?



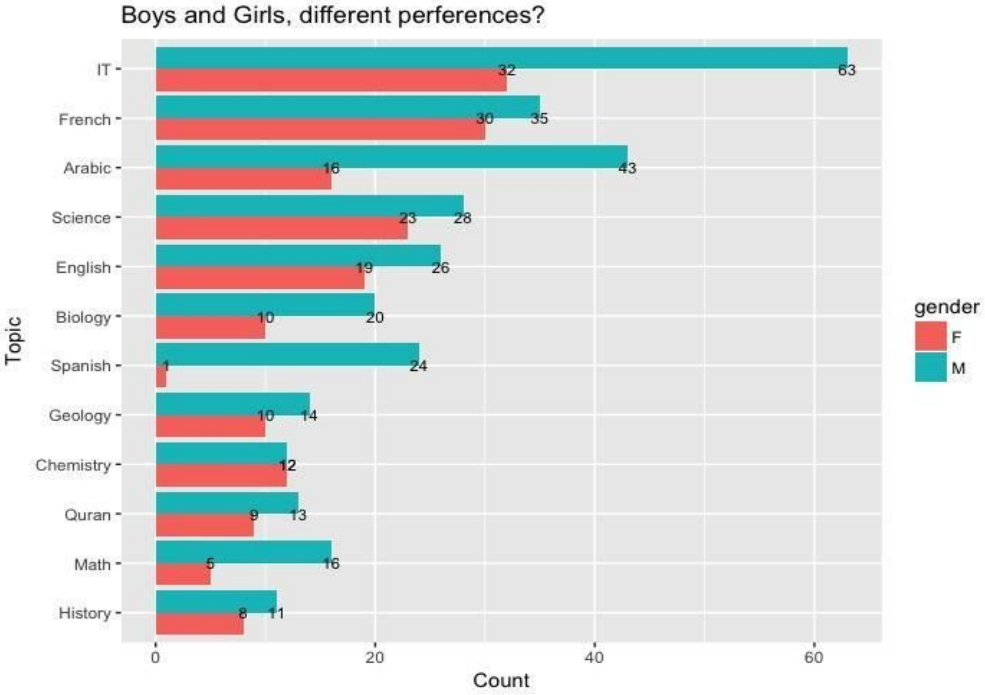
In this bar plot, the x-axis shows the number of students from a country, the y-axis shows the country's name. In each bar, the red color part represents the female student with the number marked on it, the green color part represents the male student with the number marked on it. Among those 305 males and 175 females students, students from Kuwait(KW) and Jordan occupies the majority (179 students are from Kuwait, 172 students are from Jordan), more boy students than girl students.

(2) Does mom and dad affect differently on students?



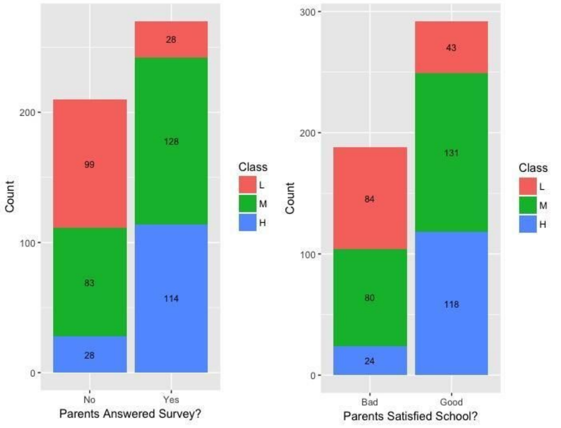
In this plot, x-axis shows the total grade level of students, L represents grade value between 0 and 69, M represents grade value between 70-89, H represents grade value between 90 and 100. Y-axis shows the number of students in each level. For example, in the L bar, the green part means 23 students are responsible by mother, 104 students are responsible by father. It is found that students with low level (L) grade are more likely to be responsible by father.

(3) Does gender unbalance exist under different topics?

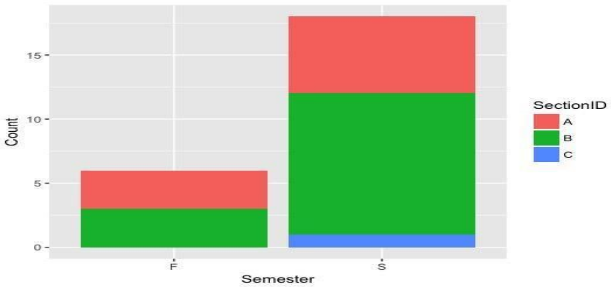


Yes, more boys in IT, Spanish, Arabic, Math, Biology.

(4) Does parents' interaction with student will affect students grade? How about the satisfaction of school?



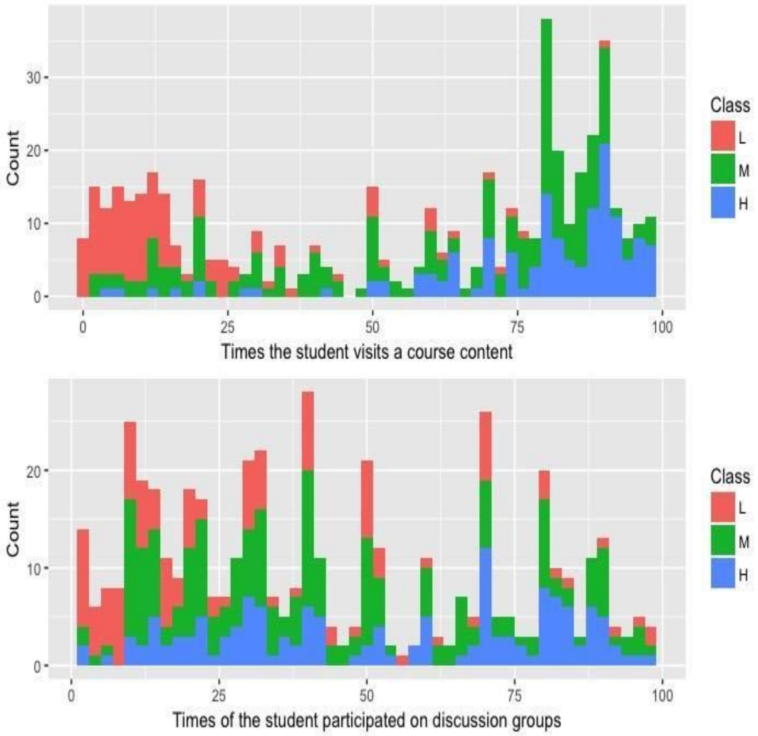
#1



#2

If parents could involve more with school's activity, students could have better performance. But still a lot of parents are not satisfied with school, especially students with low grades. Why parents are still not happy with school when students with high grades? The plot #2 shows high grade students with parents unsatisfied about school. Most students are within second semester of school from classroom A and B. Those parents might not be satisfied about school's education quality with the time. (Please check "bad" dataframe in R code)

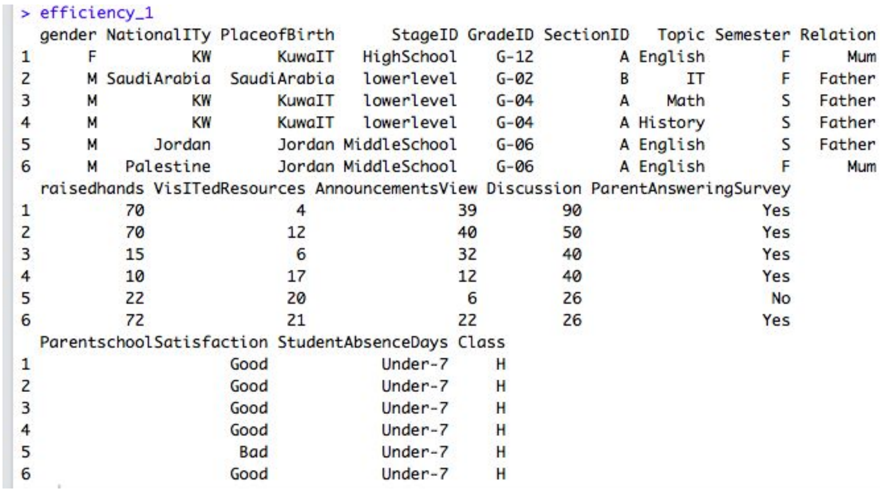
(5) Is there any outlier in student, less work with high degree? Yes.



The plot shows students' learning behavior with grades. In the upper part of the plot, the x-axis represents the number of times the student visited the course content, y-axis shows the number of students. For example, about 15 students visited a course content 50 times among the total 480 students. For the students who visited 50 times of course content, they have different grades results which means they have different learning efficiencies or being affected by other factors so they did not get a same grade. Similarly, the lower part of the plot, which shows the statistics analysis on the times of the student participated on discussion groups related to their grades. For example, about 20 students attended 50 times discussion groups activities. Usually a student will probably have high grades if he/she visits the course content more frequently, attends discussion groups more frequently.

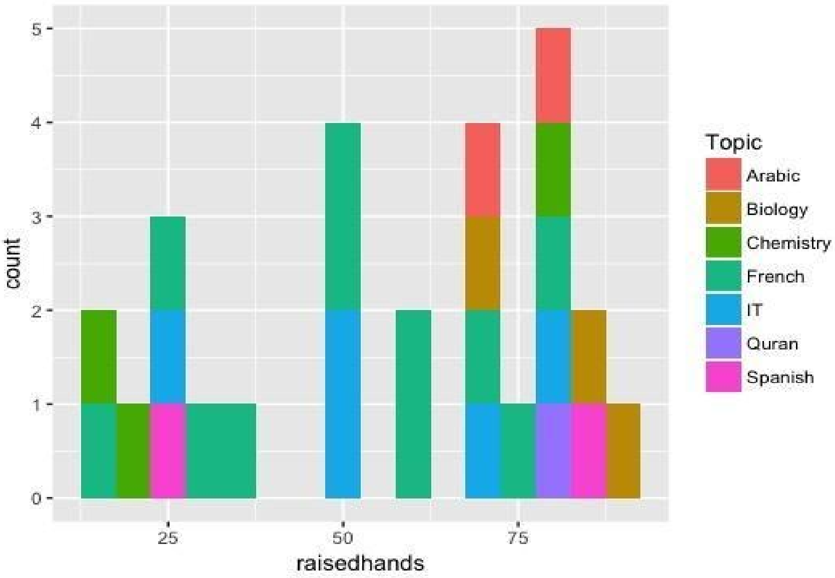
But outliers exist, how about those students less time visit a course content (eg.less than 25) but with high grades? (Please check"efficiency1" dataframe in R code)

What does students less often visit course resources with high grades look like? From the efficiency1 dataframe shown in the R code, only 1 female student exists and she likes discussion groups very much! The rest male students do not show distinctive features.

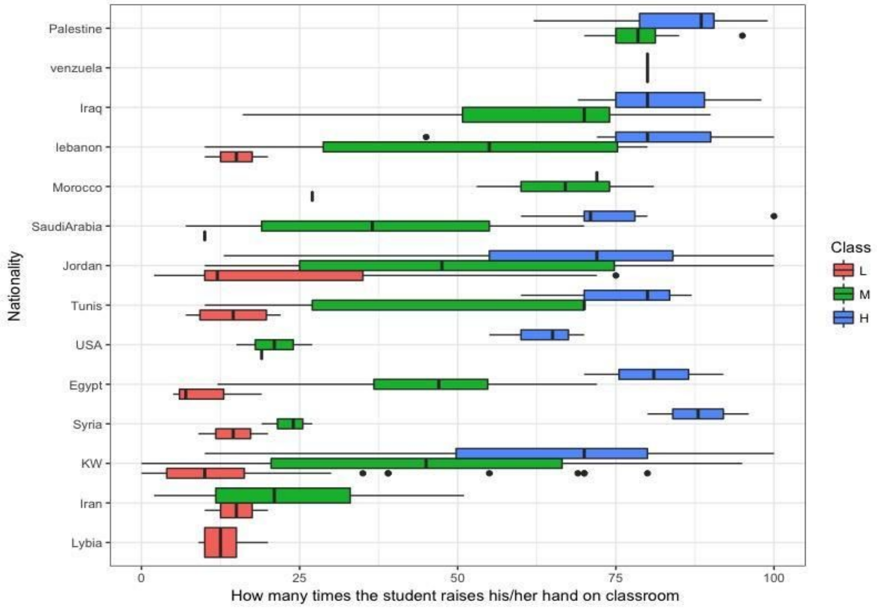


```
> efficiency_1
  gender NationalITy PlaceofBirth      StageID GradeID SectionID   Topic Semester Relation
1      F          KW       KuwaIT   HighSchool    G-12         A English        F      Mum
2      M SaudiArabia  SaudiArabia   lowerlevel    G-02         B      IT        F   Father
3      M          KW       KuwaIT   lowerlevel    G-04         A    Math        S   Father
4      M          KW       KuwaIT   lowerlevel    G-04         A History        S   Father
5      M      Jordan       Jordan MiddleSchool    G-06         A English        S   Father
6      M   Palestine       Jordan MiddleSchool    G-06         A English        F      Mum
  raisedhands VisITedResources AnnouncementsView Discussion ParentAnsweringSurvey
1          70                4                39         90                   Yes
2          70               12                40         50                   Yes
3          15                6                32         40                   Yes
4          10               17                12         40                   Yes
5          22               20                 6         26                    No
6          72               21                22         26                   Yes
  ParentschoolSatisfaction StudentAbsenceDays Class
1                     Good            Under-7     H
2                     Good            Under-7     H
3                     Good            Under-7     H
4                     Good            Under-7     H
5                      Bad            Under-7     H
6                     Good            Under-7     H
```

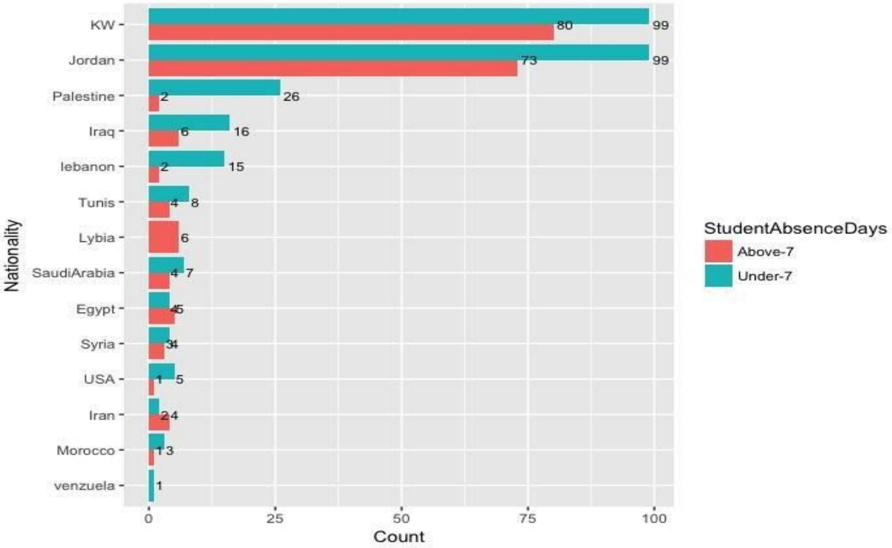
Moreover, how about those students less discussion group (eg.less than 25) time but with high grades? (Please check”efficiency2” dataframe in R code) Those students are more likely to raise hands.



(6) Students from which country are more active during class? Palestine, Venezuela (only 1), Iraq.



(7) Who does not like to go to class? Students from Lybia don’t like go to class, all of the student from Lybia absent from classroom for more than 7 days.



Part 2: Feature Engineering

Data Preprocessing (data_preprocessing.py): The dataset contains 12 categorical data columns and 4 numerical data columns. The categorical data columns were dummied and converted to numerical data type resulting in a 480*72 feature table where 480 is the number of records, and 72 for number of variables. The target variable is “Class” with content “L”, “M”, ”H”. Please see code “data_preprocessing.py” for this process. More work could be done in this process such as PCA, feature value normalization (standardization) as well as correlation-based variable selection

Part 3: Student Grade Prediction with Variable Selection

Lasso Regression and Random Forest are utilized for variable selection. These models can also find most meaningful features and provide nature of variables.

(1) Lasso and Logistic Regression: (variable_selection_lasso.py, model_logitr.py)

Lasso regression with the L1 sparsity norm can perform feature selection in order to improve the prediction accuracy, and provide faster and more cost-effective predictors by setting different alpha values to its regularization term in the objective function. In this process, the alpha is set to be 1e-3 resulting in the removal of 27 features. Of course, alpha value could be tuned with a classifier in a cross-valuation cycle, where the optimal value should be the one corresponding to the best classification performance.

Confusion matrix and accuracy rate have been adopted to evaluate the performance of the model. The performance of predictive model logistic regression in “model_logitr.py” is shown as below with accuracy rate of 0.6875.

Tab.2 Confusion Matrix of Lasso

	L(Predicted)	M(Predicted)	H(Predicted)
L	24	5	0
M	2	27	12
H	1	10	15

(2) Random Forest (RF): (variable_selection_rf.py, model_rf.py)

Random Forest ensembles results from a number of decision trees and is popular for its simplicity and avoidance of overfitting. Every node in the decision tree represents a single feature with a condition and splits the dataset into two based on the impurity of the tree to group similar response, which can help figure out duplicated features. In this work, python’s RandomForestClassifier in sklearn package was applied. Number of trees and depth of trees were set to be 50 and 3 respectively. Again, these values could be tuned via k-fold cross-validation. Features with importance values above 10% were selected resulting in only 16 features. Please check plot “RF_FeatureSelection_Results.png” for features’ importance values.

The confusion matrix out of “model_rf.py” is shown as below with accuracy rate of 0.8646. It is obviously that random forest outperformed logistic regression, or we could say the dataset is more suitable for tree models.

However, it is just an initial trial of the random forest, different parameters' (number of trees, depth of trees, etc.) tuning could be deployed by using sklean's gridsearch, control loop.

Tab. 3 Confusion Matrix of Random Forest

	L(Predicted)	M(Predicted)	H(Predicted)
L	29	0	0
M	1	32	8
H	0	4	22

(3) Study of SVM: (model_svm.py)

This is a quick study of SVM. As we can see, without variable selection, SVM did not perform well compared to the above two models. The confusion matrix is shown as below with accuracy rate of 0.6354.

Tab. 4 Confusion Matrix of SVM

	L(Predicted)	M(Predicted)	H(Predicted)
L	15	14	0
M	1	37	3
H	0	17	9

(4) Comparison of three models

Tab. 5 Accuracy Comparison

	Accuracy Score
Logistic Regression	0.6875
Random Forest	0.8646
Support Vector Machine	0.6354

RF performs the best while Lasso + LogitR is the fastest. SVM without a variable selection process before classification is with the lowest accuracy and highest computational time.

Appendix

Tab. 1 Feature Description

Feature Category	Feature	Description
Demographics	Nationality	’ Kuwait’,’ Lebanon’,’ Egypt’,’ SaudiArabia’,’ USA’,’ Jordan’,’ Venezuela’,’ Iran’,’ Tunis’,’ Morocco’,’ Syria’,’ Palestine’,’ Iraq’,’ Lybia’
	Gender	'Male','Female’
	Place of Birth	’ Kuwait’,’ Lebanon’,’ Egypt’,’ SaudiArabia’,’ USA’,’ Jordan’,’ Venezuela’,’ Iran’,’ Tunis’,’ Morocco’,’ Syria’,’ Palestine’,’ Iraq’,’ Lybia’
	Parent Responsibility	’mom’,’father’
Academy	Education Stages	‘Lowerlevel’,’MiddleSchool’,’HighSchool’
	Grade Levels	‘G-01’, ‘G-02’, ‘G-03’, ‘G-04’, ‘G-05’, ‘G-06’, ‘G-07’, ‘G-08’, ‘G-09’, ‘G-10’, ‘G-11’, ‘G-12 ‘
	Section ID	’A’,’B’,’C’
	Semester	’ First’,’ Second’
	Topic	’ English’,’ Spanish’, ‘French’,’ Arabic’,’ IT’,’ Math’,’ Chemistry’, ‘Biology’, ‘Science’,’ History’,’ Quran’,’ Geology’
	Absence Days	above-7, under-7
Parent Participation	Parent Answered Survey	parent answered the surveys provided from school or not (’Yes’,’No’)
	Parent Satisfaction	the Degree of parent satisfaction from school(nominal:’Yes’,’No’)
Behavior	Discussion Groups	how many times the student participate on discussion groups (numeric:0-100)
	Visited Resources	how many times the student visits a course content(numeric:0-100)
	Raised Hands on Class	how many times the student raises his/her hand on classroom (numeric:0-100)
	Viewing Announcements	how many times the student checks the new announcements(numeric:0-100)
Performance	Class	The students are classified into three numerical intervals

		based on their total grade/mark: Low-Level(L): interval includes values from 0 to 69, Middle-Level(M): interval includes values from 70 to 89, High-Level(H): interval includes values from 90-100.