CS251 Final Project Arthur

Abstract:

After months of building and developing data analysis and visualization programs, it is only right for me to finally analyze my own dataset(s) extensively, while exploring and attempting to answer questions that intrigue me about the data. This is the essence of this project: the final project. In this project, I am going to dig deep into five data sets regarding the education system in Uganda. I chose to analyze education statistics because I feel that the level of education an individual attains plays a vital role in their standard of living, determining various aspects of the individual's life for example: the kind of jobs they will get, the circle of friends they will have, society's perception of them, to mention but a few. Therefore, discovering various trends in education data becomes a vital tool to better the standards of living of Uganda's population in the long run, because these trends will help inform Uganda's government on where improvements could be made in Uganda's education system.

The education journey in Uganda is divided into various segments and there are national exams to highlight transitions from one stage to the next. For my analysis, I will look at: the Primary Leaving Examinations (PLE), that take you from middle school to high school; the Uganda Certificate of Education (UCE) exams, for high school to advanced high school; and Uganda Advanced Certificate of Education (UACE) that see a student off to college. I intend to discover if these results from these exams are related at all, and if so, how they affect each other. I also intend to incorporate external factors like marital and child bearing status among female children, in addition to general trends centered around female children in Uganda.

Questions:

1. What relationships exist within the data for the three different education levels, that is: ple, uce, and uace results?

2. What relationships exist between the external factors: marital and childbearing status and education attained by Ugandans?

3. How does performance vary across genders in Uganda?

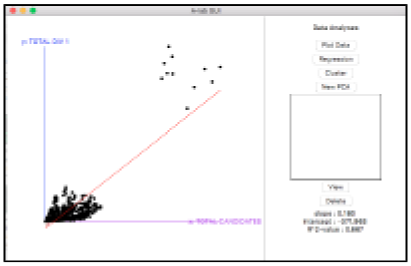
Methods, Exploration and Results:

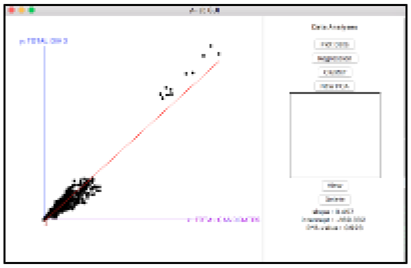
With the above questions in mind, I had to find the appropriate data sets to help me answer my questions. On getting my datasets, I had a huge task of cleaning them up. I had to curate my data, tweaking it to make it appropriate to my analysis arsenal. My first step in my data curation was dealing with missing data. Sometimes this involved calculating the missing values, while others I just simply disregarded columns with a lot of missing data. I then had to add a row specifying the data type in each column. I also ended up combining some data sets: either entirely or just adding columns across various data sets.

Below is the **User manual** for the A-Lab GUI:

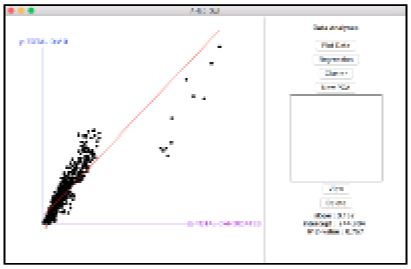
- Command q - quit
- command r - reset
- command o - open new data file
- command p - plot data
- command L - linear regression
- command s - save picture
- command a - pca analysis
- command c - cluster analysis

In my analysis, I mostly employed linear regression, pca and clustering analysis. Firstly, I look at the PLE results. Running linear regression analyses on the data, I get the following results:

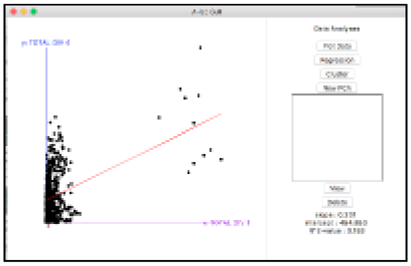




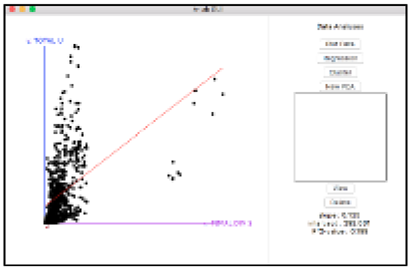
slope : [illegible]
intercept : [illegible]
R^2-value : [illegible]



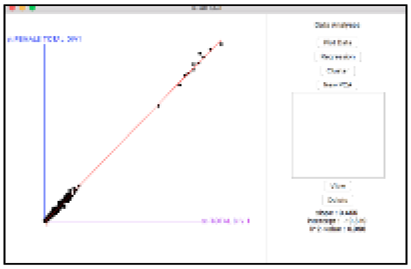
slope : [illegible]
intercept : [illegible]
R^2-value : [illegible]



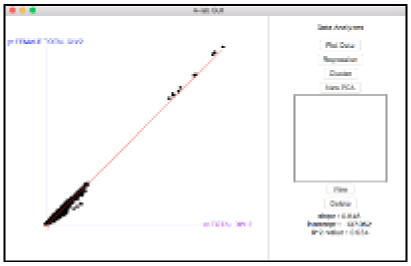
slope : [illegible]
intercept : [illegible]
R^2-value : [illegible]



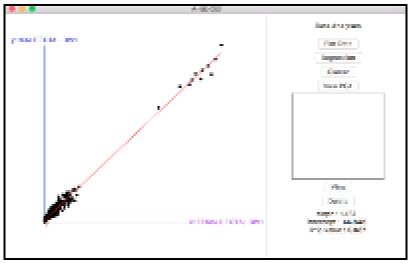
slope : [illegible]
intercept : [illegible]
R^2-value : [illegible]

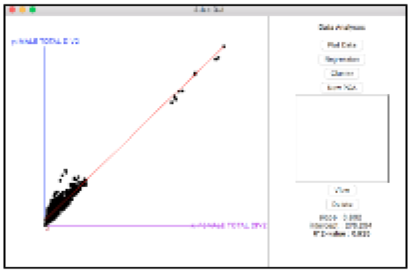


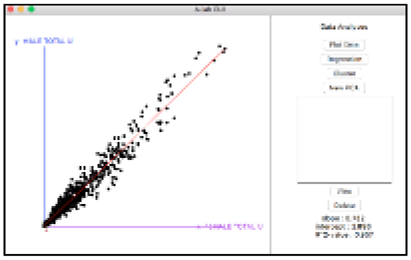
slope : [illegible]
intercept : [illegible]
R^2-value : [illegible]



slope : [illegible]
intercept : [illegible]
R^2-value : [illegible]

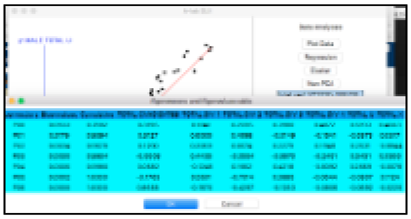




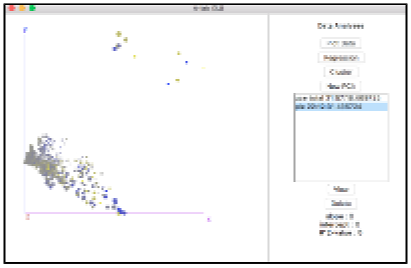


From these analyses, as expected, there is a relatively strong correlation between the total number of students and each respective grade, for example there is a correlation coefficient of 0.923 between the total number of candidates and the number of candidates in division two. However, there is a weak correlation between the different grades, for example the correlation coefficient between division four and division one is 0.153. There is a strong correlation between the total number of candidates and the performance of each sex. Additionally here is a strong correlation between the performance of females and that of males for example if many females got division ones, so did males (R2 coefficient of 0.925).

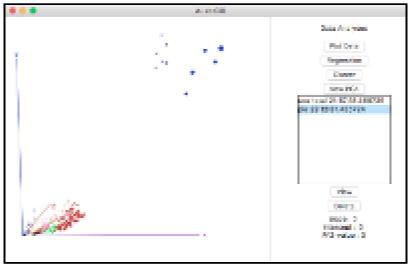
This makes sense because this data is two dimensional, as shown in the figure below: a pca analysis on the total number of candidates, total division one, two, three, ungraded. Two new features are required to represent 93% of the variation in the data.



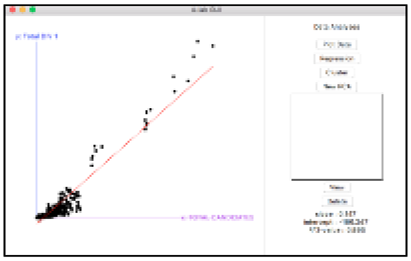
Below is a plot on the first five dimensions:

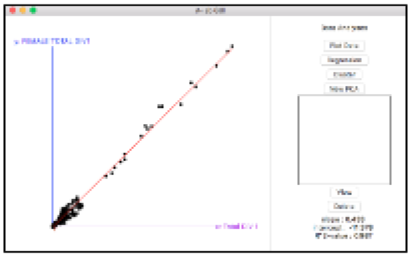


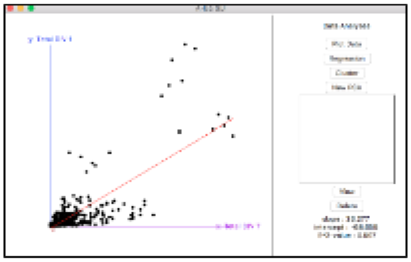
Due to curiosity, I go ahead to discover the groups in this data set. I run a clustering analysis using the l1norm distance metric. Below are the five groups in the data:

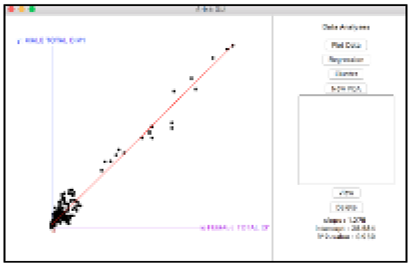


I then move on to the UCE results. On running linear regression on this data, similar trends as in PLE persist. For example there is a high positive correlation (0.895) between the total number of candidates and the number in division one. Further evidence is attached below:

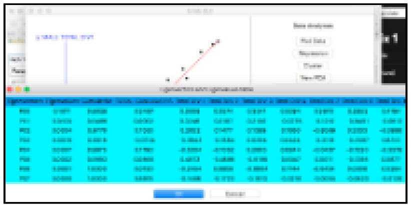




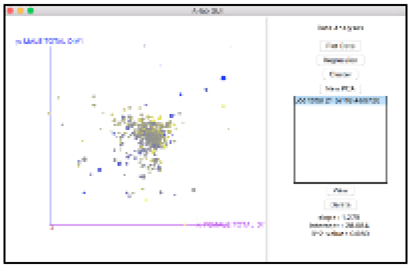




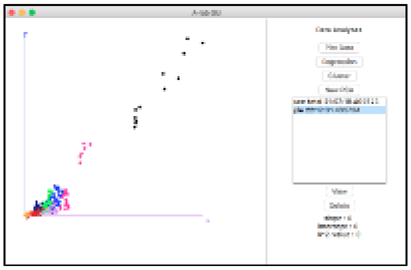
Similarly, this data is also two dimensional as only two new features are required to represent 0.95% of the variation in the data as shown below:



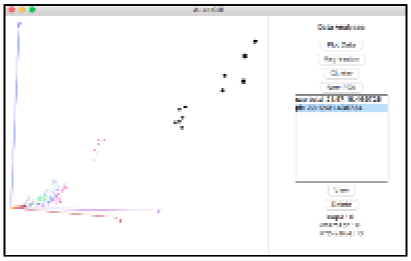
And below is a plot of the data projected on the first five Eigenvectors:



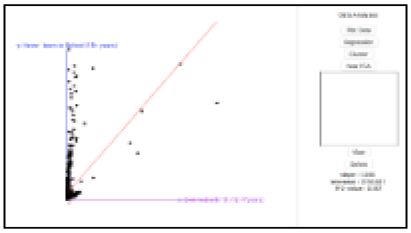
I go ahead to discover the groups in this data set. I run a clustering analysis using the l1norm distance metric. Below are the ten groups in the data:



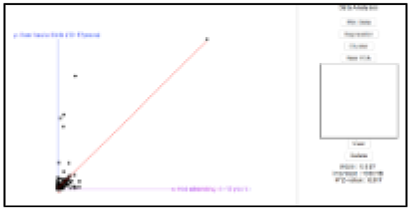
On rotating a little bit:



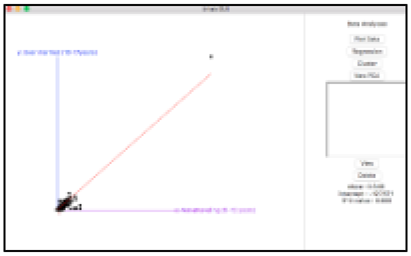
I would go ahead and analyze the UACE data set but due to there being so many zeros in the data, it is insufficient. It is not all around encompassing since many districts do not have data for students going to eleventh and twelfth grade. I therefore move on to do inter-dataset analyses. To do this, I start by combining the education characteristics and marital and childbearing status data sets into one file. I then run linear regression analyses on the data. Below are my findings:



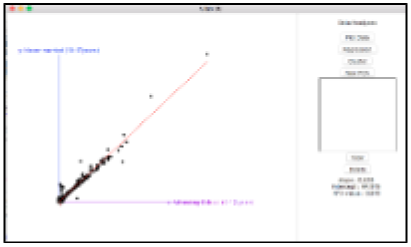
Interestingly, as shown above, there is barely any correlation between the number of Ugandans have not been to school and those that have had a birth between the age of 12 and 17 (R^2 = 0.187). This could because early child births are not the primary reasons as to why Ugandans do not attend school. Factors like the high prices of attending school might be playing a bigger role.



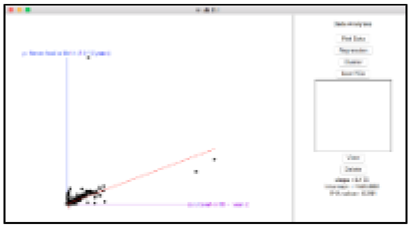
However, there is an moderate relationship between children between 6 and 12 not attending school and having an early birth (between 12 and 17 years), as shown above. However, the early birth parameter includes births after age 13.

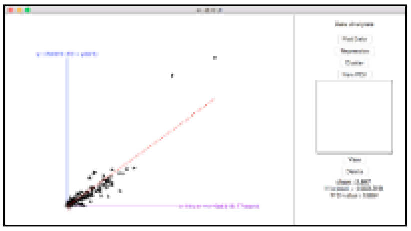


As shown above, there is a high correlation between early marriages (10 -17 years) and children not attending school from age 6 to 12 (R^2 = 0.9). This could be because of the fact that when the children are married off, they leave school to become wives, which is sad.



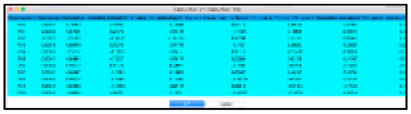
The results of this linear regression reinforce the findings in the previous analysis. There exists a strong correlation between attending school and never being married(R^2 value of 0.97).



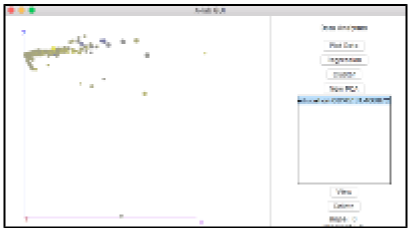


From the last two analyses, I get interesting results. Despite there being a strong relationship (R^2 =0.854) between never being married and the literacy rate, there is barely any relationship (R^2 = 0.291) between having an early birth and the literacy rate. This suggests that sometimes having an early birth does not necessarily stop someone from continuing school, however, if they go ahead to get married, then they might quit school.

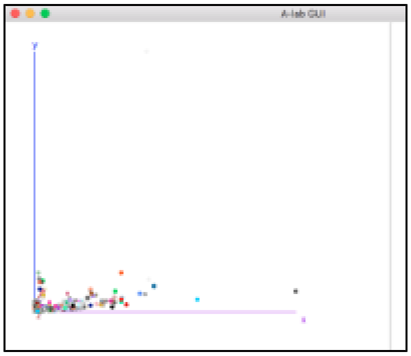
I then go ahead to execute a pca analysis in the data to discover the dimensionality of the data set. I get the following results:



The data is three dimensional with three features needed to represent 93% of the variation of the data.



I go ahead to discover the groups in this data set. I run a clustering analysis using the l1norm distance metric. Below are ten groups in the data:



Conclusion:

From my analyses, it is evident that indeed for females in school, they have equal ability as males. Their performance is strongly related. This is very crucial to take into account especially in Uganda where some parents think it is risky to send their daughters to school because they are scared they might not perform as well as boys. However it is evident that various factors play a big role in the girl-child-education. An example of this is having early births and early marriages. Early marriages tend to make children drop out of school. I contend that reducing the rate at which girls get early into marriages, or give birth early would increase the literacy rates in Uganda. This would consequently improve the standard of living across the country.

References:

Professor Stephanie Taylor

The source of my data files is Data.ug.

http://catalog.data.ug/dataset?q=&sort=views_recent+desc