Original Research Article

An Example of Using MATLAB for Data Analysis—The Correlation between College Entrance Exam and Students' Performance at Universities

Zhihao Ren*

California Crosspoint High School, CA 94545, USA. E-mail: calebren21@gmail.com

Abstract: Many universities all over the world use entrance exams as a tool for filtering and selecting applicants to their courses. While entrance exams provide a standardized testing mechanism, it is not clear whether they are a predictor of the student's future performance at the university. As an initial investigation, the author found and analyzed a raw dataset of students' entrance exam scores and their performance in the university during the first three semesters. The author carried out statistical analysis for the entire cohort and also according to gender. The analysis was carried out using Matlab. The analysis carried out shows that for the given dataset, there is no correlation between entrance exam scores and university scores. Also, there is no significant difference in the performance at the entrance exam and university scores, between male and female students.

Keywords: Histogram; Statistics; Scatterplot; Correlation

1. Objective: analyze the datasets with the programing tool

The idea for this article arose out of the author's curiosity to understand whether entrance examination grades are a predictor of a student's performance at the university, especially in the initial phase and his motivation to apply the data analysis tools he has learnt to large datasets.

2. Background: identify MATLAB as the tool to analyze the chosen dataset

Matlab is a really convenient language for technical computing and was ranked among the top 10 programming languages by IEEE for 2020[1]. It has a good computing, calculating, and programming environment. Compared to the traditional language such as Java or Python, Matlab has powerful built-in routines that can perform very complex and tedious calculations. It also has an easy-to-master chart design that visualizes the data, which makes it clear and helps us to master the data.

The dataset used for analysis is taken from the entrance examination scores of Grande do Sul Federal University, South Rio DE Janeiro, and average scores of students in the first three semesters of university[2]. Each row in the dataset contains scores from nine exams an applicant took during the college application process, as well as anonymous information about their corresponding GPA during the first three semesters of college. The dataset has 43,303 lines, one student for each line.

doi: 10.18282/ii.v2i2.370

The matrix derived from the original dataset has 11 columns:

(1) Column 1 - Gender: where 0 denotes female and 1 denotes male

(2) Column 2 - Score on the physics exam

(3) Column 3 - Score on the biology exam

(4) Column 4 - Score on the history exam

(5) Column 5 - Score on the second language exam

(6) Column 6 - Score on the geography exam

(7) Column 7 - Score on the literature exam

(8) Column 8 - Score on Portuguese essay exam

(9) Column 9 - Score on the math exam

(10) Column 10 - Score on chemistry exam

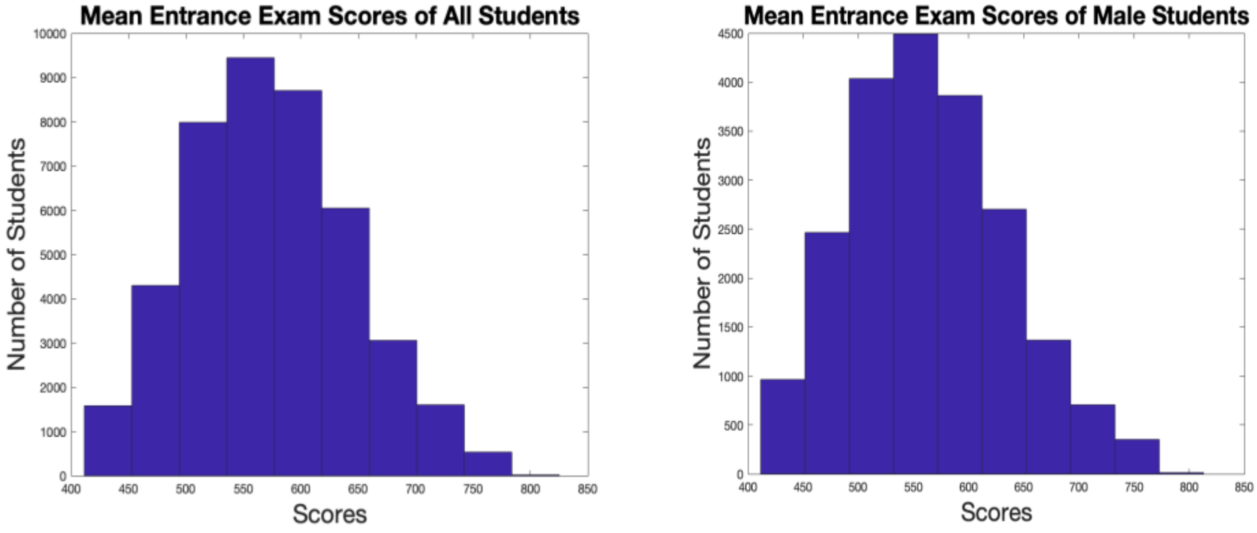
(11) Column 11 - Score of the mean GPA during first three semesters at university, on a 4.0 scale

3. Methods: convert the datasets into a matrix for efficient analysis

Using Matlab, the comma separated file was converted into a matrix for efficient analysis. An additional column which was the mean of the scores in the entrance examination was added for each student. The dataset was also sorted and split into two smaller datasets based on the gender data provided. The data was then saved as a .mat file to avoid recalculation and for easier processing by Matlab.

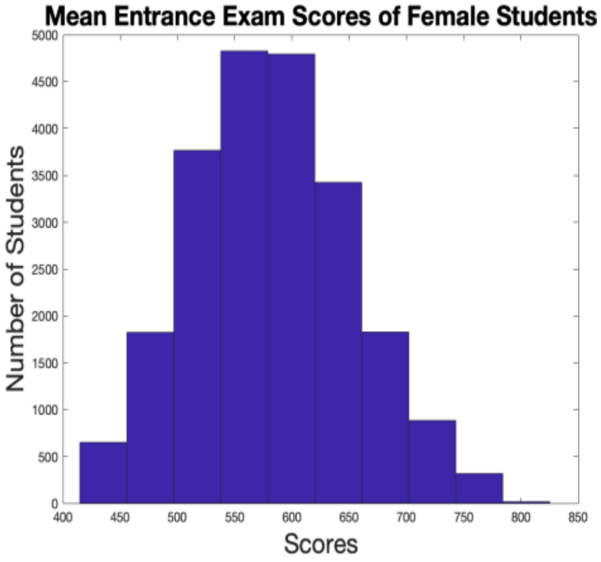
Thereafter, the mean, median and mode of the data was calculated for all students and for male and female students separately. Histogram analysis of the entrance scores and CGPA was also carried out. Finally, a scatterplot was performed to check if there was any correlation between mean entrance exam scores and the average GPA during the first three semesters. In addition to the statistical analysis, different software approaches were used to understand the performance of Matlab.

4. Results and observations



a

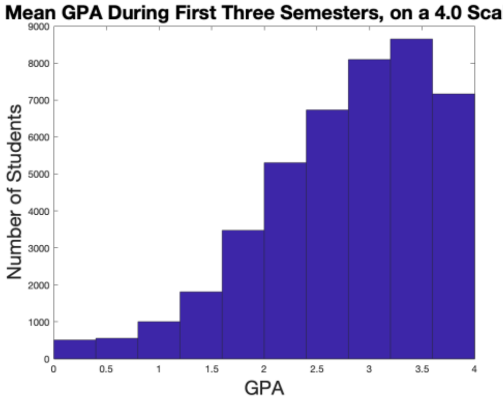
b



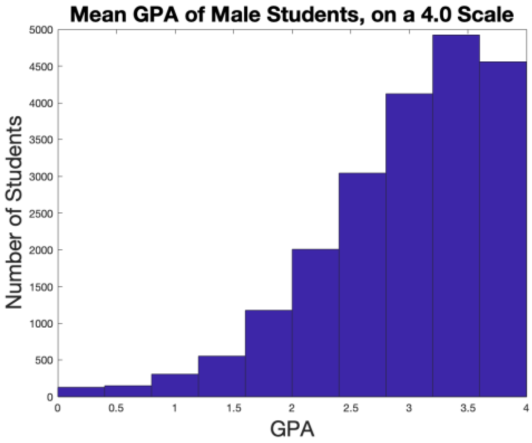
c

Figure 1. Mean entrance exam scores for a) all, b) male and c) female students.

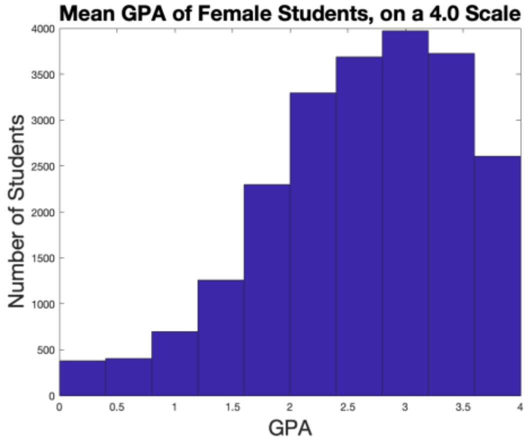
As can be seen in **Figure 1**, the entrance exam scores follow a bell curve with most of the students scoring around the mean score whereas the mean GPA during the first three semesters is highly skewed towards higher GPA, as shown in **Figure 2.** This difference is observed for all students as well as for male and female students separately as can be seen from Figures 1 and 2. One can infer that the difficulty of entrance examinations was set such that the scores followed the normal distribution curve. However, for the university examinations, we observe that scores are biased towards the maximum GPA. This either means that there was improvement in the students' understanding capacity due to better guidance at the university or it could also mean that while the entrance exams tested the students across a wide spectrum of subjects, selective choosing of courses that the students are more interested in or have an aptitude for, resulted in better CGPA during the first three semesters.



a



b



c

Figure 2. Mean university CGPA during first three semesters for a) all, b) male and c) female students.

The error bars shown in **Figure 3** demonstrate that there is no significant difference in the performance of male and female students, for both the entrance examinations and their university CGPA. Especially of interest is the fact that female students perform better than male students when it comes to their performance in STEM subjects in the entrance exam, dispeling notions of male advantage in STEM subjects.

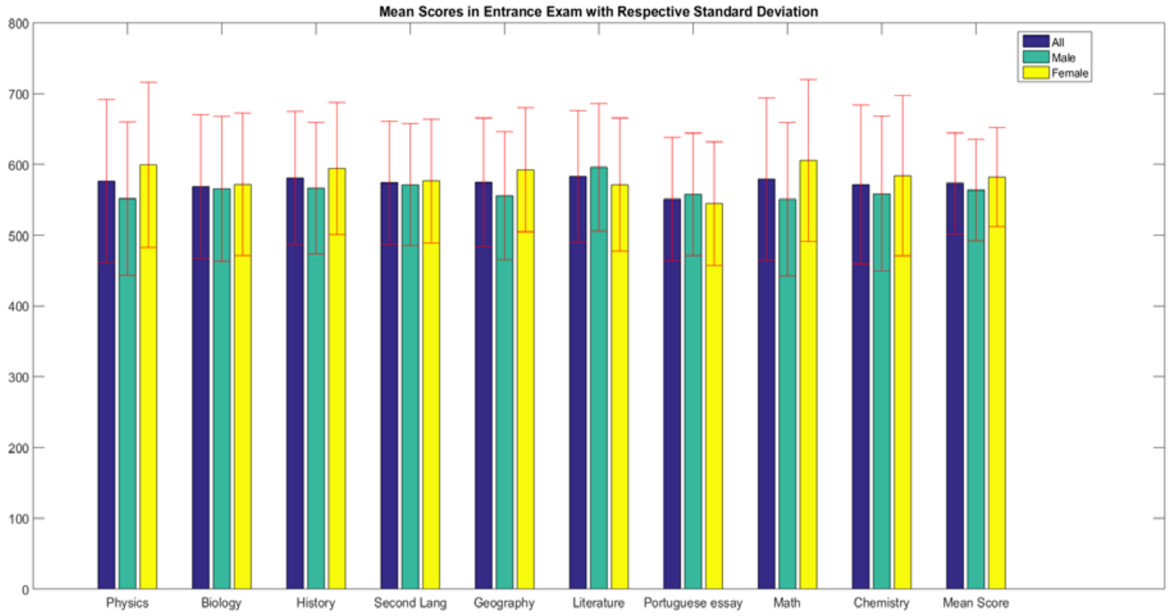


Figure 3a. Mean scores in entrance exam with respective standard deviation.

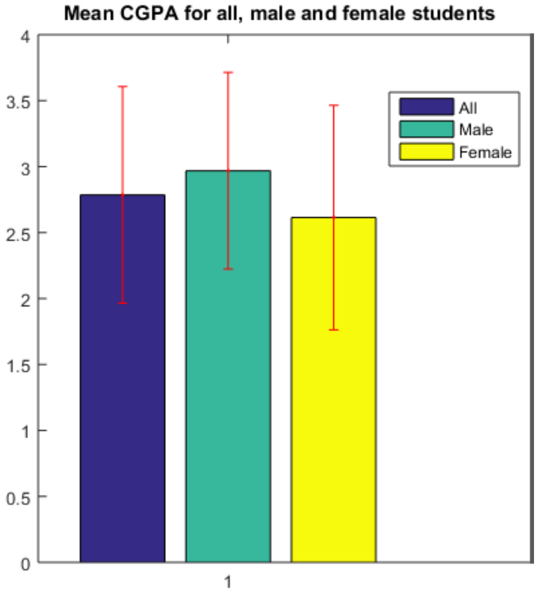


Figure 3b. Mean university CGPA for first three semesters with respective standard deviation.

Scatter plot drawn in **Figure 4** between the mean entrance exam score and university CGPA further confirms that no prediction can be made regarding a student's future performance at the university based on their entrance exam scores.

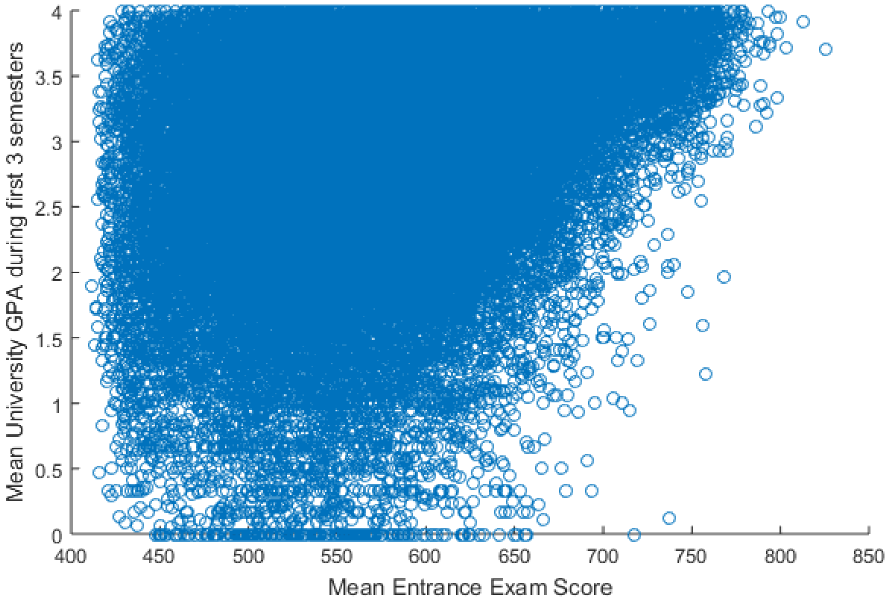


Figure 4. Scatter plot of mean CGPA versus mean entrance exam score.

The following table shows the goodness of the curve fitting carried out using the fit function in Matlab. A linear curve of the form f(x) = 0.003396*x + p2 was the best fit curve found by Matlab. As can be seen from the table, the correlation coefficient is very low indicating almost no correlation between the mean entrance exam scores and the mean university GPA.

Sum of squared errors (sse)	2.6621e+04
Correlation coefficient (R^2)	0.0874
Root mean squared error (rmse)	0.7841

5. Conclusion

The dataset, while not huge in size, was large enough to enable the author to understand some of the challenges faced while dealing with big data analysis. Initially, the author tried to calculate various statistical parameters using nested for loops which was suboptimal with respect to execution time on Matlab. Thereafter, he took advantage of Matlab's in-built functions and matrix operations to perform the same calculations, which significantly increased the speed. The author also sorted the data according to gender and split the dataset into two based on them. The author not only calculated various statistical parameters but also tried to demonstrate them in meaningful ways using error bars and histograms, which can be drawn using Matlab's plotting functions. In addition to the technical task, the author also tried to infer the results obtained from the above operations.

References

1. Cass S. Top programming languages. IEEE Spectrum 2020; 57(8): 22. Available from: https://spectrum.ieee.org/at-work/tech-careers/top-programming-language-2020.
2. Castro da Silva B. UFRGS entrance exam and GPA data [Internet]. Harvard Dataverse; 2019. Available from: https://doi.org/10.7910/DVN/O35FW8.