**Student Number:** C15311966

**Name:** Maksymilian Drzezdzon

**Programme Code:** TU060/1

# Section 1 - Research Question(s)

**Does a student’s/parent’s level of Sisben (Government aid) influence their level of academic achievement?**

My hypothesis is based on the assumption that in poorer households there is more stress/pressure on said students when it comes to higher education, compensation and doing well, this is why id like to investigate if there is relationship between students grades and their family’s socioeconomic status through their exam results and level of Sieben students/parents receive. This will be attempted using **dimension reduction** as it’s something I didn’t have time to properly cover in my CA portfolio. This is a super important concept as its useful for feature selection/extraction.

# Section 2 - Dataset

The dataset contains the results in national assessments for secondary and university education in engineering students and contains academic, social, economic information for 12,411 students. The data was collected as part of the Master's Degree in Engineering project of the Technological University of Bolívar (UTB) titled Academic Efficiency Analysis in Engineering students.

**Representativeness**

The gender distribution of students is 40.63% for women and 59.37% for men. This distribution of 40/60 should not cause any major issues and can be left as is without preforming any variable weightings.

**The variables of interest to your research questions and their statistical measurement type**

Internet, TV, Computer, WASHING\_MCH, MIC\_OVEN, CAR, DVD, FRESH, PHONE, MOBILE are all binary.

Sisben is a discrete ordinal variable.

Saber 11 exams: MAT\_S11, CR\_S11, CC\_S11, BIO\_S11, ENG\_S11 are all discrete variables.

Saber Pro exams: CR\_PRO, QR\_PRO, CC\_PRO, WC\_PRO, ENG\_PRO are all discrete variables.

**Provide relevant descriptive statistics and visuals/Assess normality where relevant**

Both Completed In r code for the sake of space.

**Missingness**

There is no missing data apart from a column called ‘…10’ this column was probably created when two datasets were joined as its completely empty and its in-between data from the two tables shown in the dataset descriptor.

## Option B

## Section 3.1 Dimension Reduction

Reported in R notebook

* Correlation matrix
* Bartlett’s Test of Sphericity
* KMO
* Loadings
* Communalities
* Variance Explained
* Factors/Components extracted

## Section 3.2 – Model

### Section 3.2.1 Statistical Evidence

In R notebook, evaluating variables and preforming the above tasks in section 3.1

### Section 3.2.2 Model

Reported in R notebook

* Report the model as discussed during lectures
* Make sure you comment on the fit and usefulness of the model in your report
* Make sure you comment on the predictors and their significance
* Make sure you assess the model against assumptions
* Make sure you illustrate how the model works for the dataset e.g. state the regression equation and illustrate how it works by using mean/median values for continuous data and the differences by looking at the equation for all values of the differential effects.

The equation used in the PC analysis



# Section 4 – Discussion/Conclusion

In this section you should reflect on your results from the perspective of your research question(s). You should also suggest some changes/additional research that could be conducted to try to better answer this research question.

# Section 5 – Appendix/Bibliography

A full descriptor is available at: <https://www.sciencedirect.com/science/article/pii/S2352340920304315#utbl0001>

The dataset is available for download at <https://data.mendeley.com/datasets/83tcx8psxv/1>

Delahoz-Dominguez, Enrique, Rohemi Zuluaga, and Tomas Fontalvo-Herrera. "Dataset of academic performance evolution for engineering students." *Data in Brief* (2020): 105537.