Socioeconomic and Academic Distribution of University Students in Colombia

Anjan Majhi, Kanak Sangvi, Rishit Kulhari, Rishu Anand
Project for Introduction to Statistics and Computation with Data
Academic Year: 2024–2025

Introduction

Higher education serves as a critical determinant of professional advancement and economic mobility. However, access to such education is significantly shaped by socioeconomic conditions, tuition costs, and financial aid availability. This study investigates how socioeconomic status (SES) influences career choice, tuition fees, and access to financial assistance at a public university in Colombia—Universidad Nacional de Colombia.

Using data from the National Directorate of Academic Information, Registration, and Enrollment for the first semester of 2021, the research provides an in-depth look into the interplay between students' economic backgrounds and their academic trajectories.

Objective

The primary aim of this project is to explore the relationship between socioeconomic factors and academic decisions. Specifically, the project seeks to:

- Analyze the link between SES, tuition fees, and career choices.
- Compare tuition fees based on students' secondary school background (public vs. private).
- Investigate how SES affects scholarship allocation.

By examining these aspects, we aim to highlight the structural patterns that influence higher education access and affordability.

Background

In Colombia, socioeconomic status is categorized on a scale from 0 to 7—where 0 represents the most economically vulnerable and 7 the most affluent. This classification is based on various household factors, including income, housing quality, and access to public services.

The dataset used in this study captures these elements, providing insights into:

- Parental income
- Residence type and location
- Secondary school type
- Tuition and scholarship details
- PBM (Basic Tuition Score), a composite index reflecting economic capacity

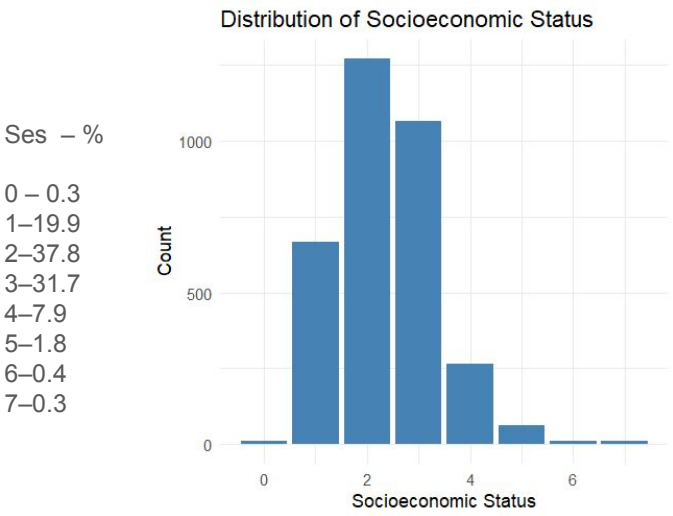
PBM plays a central role in determining tuition fees—higher PBM scores result in higher tuition, ensuring that fee structures are aligned with students' financial capacity.

Dataset Overview

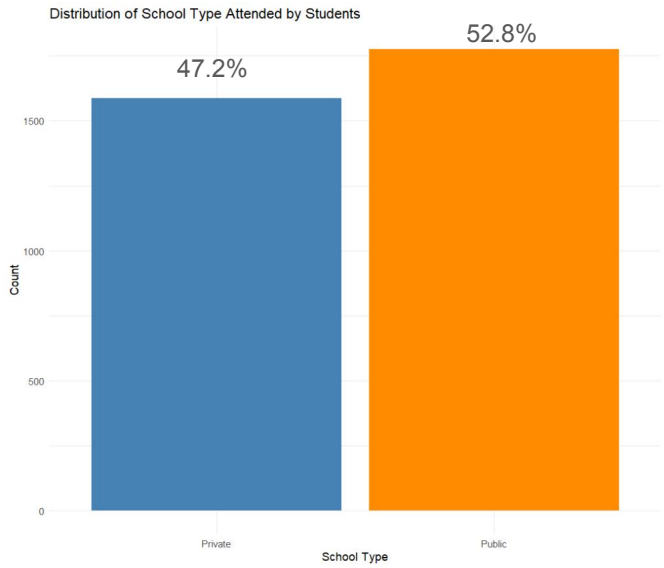
The dataset, titled **ColombiaTuitionSET**, comprises **3,361 anonymized student records** from the first semester of 2021. Key variables include:

- Tuition fee (COP)
- Family income (father's and mother's income)
- School type (public/private)
- SES level (0 to 7)
- Scholarship status
- Career/program affiliation
- Residence type (urban/rural/special)
- Sibling count

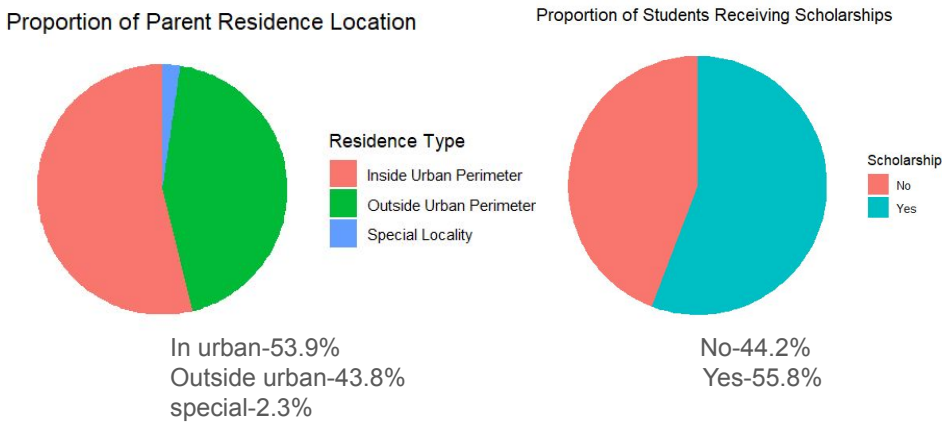
The structure of this dataset supports a wide range of statistical analyses to identify trends and relationships between economic status and academic outcomes.



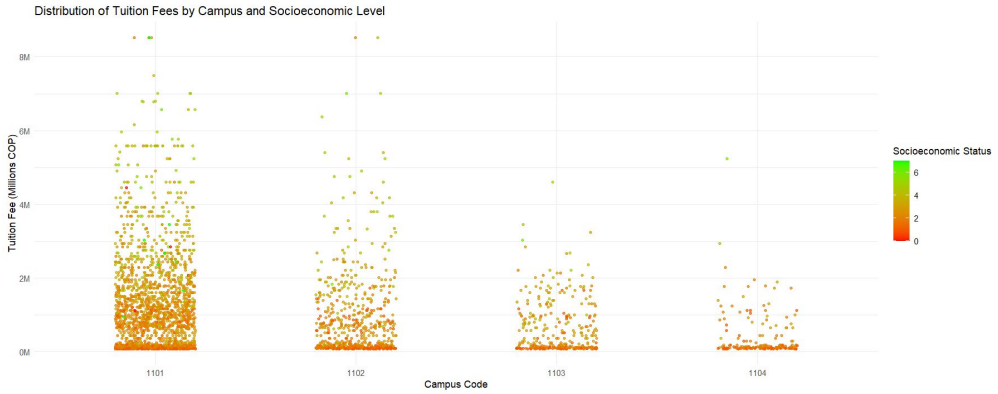
Observation: The Graph highlights that the majority of students come from lower SES groups SES2:37.8% ;SEs3:31.7% ;SES1:19.9% SES 5, 6, 7 combined account for 2.5% of the population. SES 0 and SES 4 are also minimal, each under 8%.



Observation: Almost equal number of students come from both public and private schooling(the difference is not that high)



Observation: Urban: 53.9%
Outside Urban (Rural): 43.8%
Special Locality: 2.3%
Scholarship :44.2% doesn't get scholarship and 55.8 % does get it.



Observation: Majority of students live in Hostel code 1101 and 1102 with lower SES student lying in bottom of graph

summery of data provided

- **SES Distribution:** Most students fall in SES levels 1–3, with 37.8% in SES 2 and 31.7% in SES 3.
- **School Type:** Slight majority attended public schools (52.8%), while 47.2% came from private schools.
- **Parental Residence:** 53.9% of students' parents live in urban areas, 43.8% in rural, and 2.3% in special localities.
- **PBM and Tuition Correlation:** A strong positive correlation ($r = 0.921$) exists between PBM and tuition fees, confirming PBM as a key driver of tuition costs.
- **Campus Variation:** Some campuses (e.g., codes 1101 and 1102) show a broader range of tuition fees, with certain cases exceeding 6 million COP.

Hypothesis 1 & 2

Objective

The aim of this project is to explore and compare statistical methods for analyzing relationships between variables, with a focus on non-parametric tests and correlation analysis.

Hypothesis 1: Students who went to Private schools pay more tuition fee than those who went to Public schools.

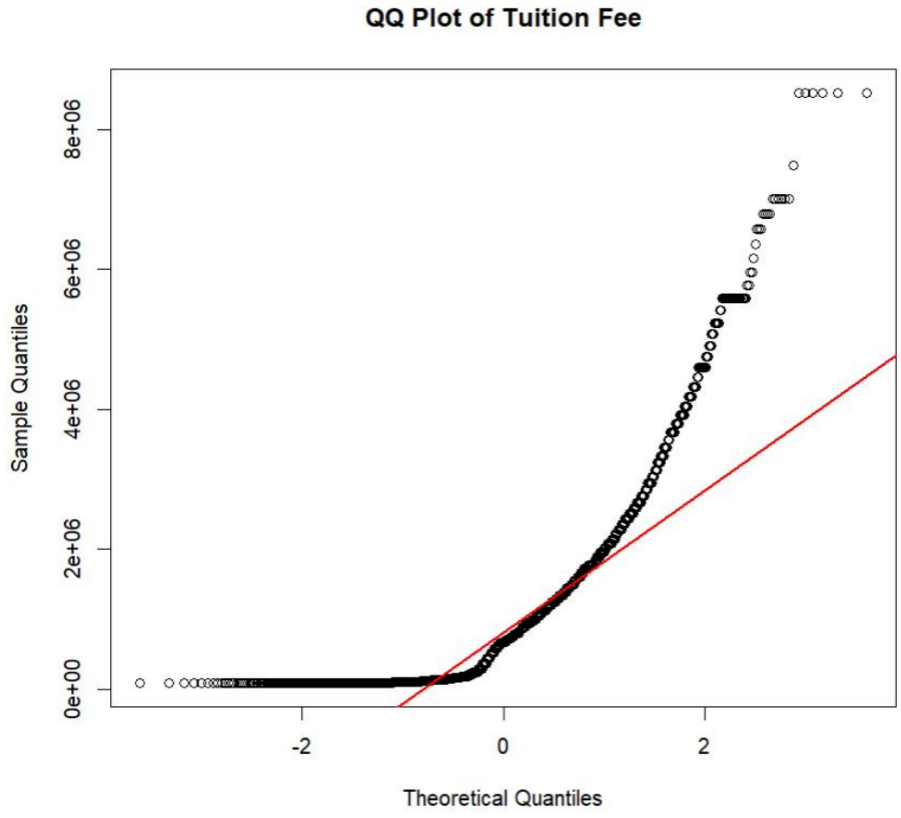
Defining Null & Alternative Hypothesis:

H_0 (Null Hypothesis): The average tuition fees of the students coming from private schools is same as that of students coming from public schools.

H_a (Alternative Hypothesis): The tuition fee of the students coming from private schools is significant higher than the students that are coming from public schools.

We need to check if we can use t-test to reject (or fail to reject) H_0.

Checking normality:



Conclusion: Normality doesn't hold due to negative skewness.

Check homoscedasticity:

Levene's test checks whether different groups have the same variance. Equal Variances (homoscedasticity) are a key assumption for tests like ANOVA and t-test.
We define Z_{ij} as

$$Z_{ij} = |X_{ij} - \tilde{X}_i|$$

Where

- $\tilde{X}_i$ is the mean of the i-th group. (In case of Brown-Forsythe test is the median)
- X_{ij} is the i-th group's j-th subject.

For Levene's test the Null and Alternative Hypothesis are

H_0 (Null Hypothesis): Variance of every group is equal.

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

H_a (Alternative Hypothesis): At least one group has a significantly different variance.

$$\sigma_i^2 \neq \sigma_j^2 \quad (\text{for some } i \neq j)$$

Now we perform the ANOVA test on Z_{ij} and find the F-test statistic as

$$F = \frac{\text{variance between groups of } Z_{ij}}{\text{variance within groups of } Z_{ij}}$$

where

- variance between groups of Z_{ij} is

$$\sum_{i=1}^{k} \frac{n_i(\bar{Z}_i - \bar{Z})^2}{k-1}$$

  with degree of freedom $k-1$. (k =no. of groups)

- variance within groups of Z_{ij} is

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} \frac{(Z_{ij} - \bar{Z}_i)^2}{N-k}$$

  with degree of freedom $N-k$. (N = total observations)

And we take our $\alpha = 0.05$.

For our data:

```
Levenes Test for Homogeneity of Variance (center = median)
        Df F value Pr(>F)
group    1   611.71 < 1.1e-16 ***
      3359
```

Low enough p-value. Therefore we reject H_0 of the Levene's test, that is our data is not homoscadastic.

SInce normality and equal variance assumptions are violated, We use Mann-Whitney U/Wilcoxon rank sum test. (Since we have independent samples)

The Mann-Whitney U test is based on ranks of values, not raw data.
Now let,

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \quad \& \quad U_2 = n_1 n_2 - U_1$$

Now the test statistic U is defined as

$$U = \min\{U_1, U_2\}$$

Where

- $n_1,\ n_2$ are the sample sizes of two groups
- $R_1,\ R_2$ are sum of ranks of each groups
- $U_1,\ U_2$ are the Mann-Whitney U-statistic for each group

Now The Null & Alternative Hypothesis are
H_0 (Null Hypothesis): The Rank sums of two groups do not differ

$$R_1 = R_2$$

H_a (Alternative Hypothesis): The Rank sums differ significantly

$$R_1 \neq R_2$$

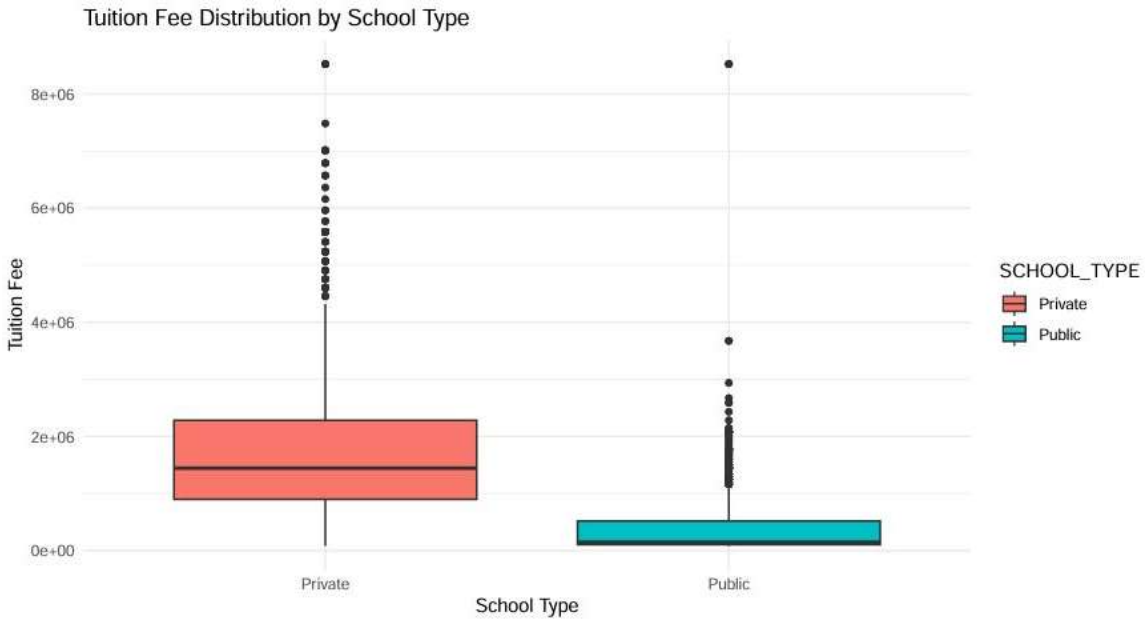
Again we take $\alpha = 0.05$ and by the data we find the p-value as

```
        Wilcoxon rank sum test with continuity correction

data:  TUITION_FEE by SCHOOL_TYPE
W = 2491853, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

⟹ We reject null hypothesis

Conclusion: Tuition fees paid by students who come from private schools is on average different from those who come from public schools.


Tuition Fee Distribution by School Type
Tuition Fee
8e+06
6e+06
4e+06
2e+06
0e+00
Private
Public
School Type
SCHOOL_TYPE
Private
Public

Hypothesis 2: Lowe SES students are more likely to recieve scholarships.

Definining Null & Alternative Hypothesis:

H_0 (Null Hypothesis): Scholarship distribution is independent of SES.

H_a (Alternative Hypothesis): Scholarship distribution is dependent on SES.

We check if χ^2 test is valid:

```
        Pearson's Chi-squared test

data:  table_ses_scholarship
X-squared = 603.96, df = 7, p-value < 2.2e-16

Warning message:
In chisq.test(table_ses_scholarship) :
  Chi-squared approximation may be incorrect

> min(expected_vals)
[1] 3.979173
```

So, we can't use χ^2 since the expected frequency is less than 1. instead we use the Fisher test.

Fisher's Exact test is a statistical significance test used to determine whether there are nonrandom associations between two categorical variables. Originally designed for 2×2 tables, it can be generalized to an $m \times n$ contingency table. (For continuous distribution of data points we sub-divide the spectrum to get discreet values.)

First we make a table of the data

	C_1	C_2	$\cdots$	C_n	**Row Total**
R_1	a_{11}	a_{12}	$\cdots$	a_{1n}	r_1
R_2	a_{21}	a_{22}	$\cdots$	a_{2n}	r_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
R_m	a_{m1}	a_{m2}	$\cdots$	a_{mn}	r_m
Column Total	c_1	c_2	$\cdots$	c_n	N

Where

- a_{ij}: count in cell at row i and column j
- $r_i = \sum_{i=1}^{n} a_{ij}$ sum of i-th row.
- $c_j = \sum_{i=1}^{m} a_{ij}$ sum of j-th column.
- $N = \sum_{i=1}^{m} \sum_{i=1}^{n} a_{ij}$ total number of observations.

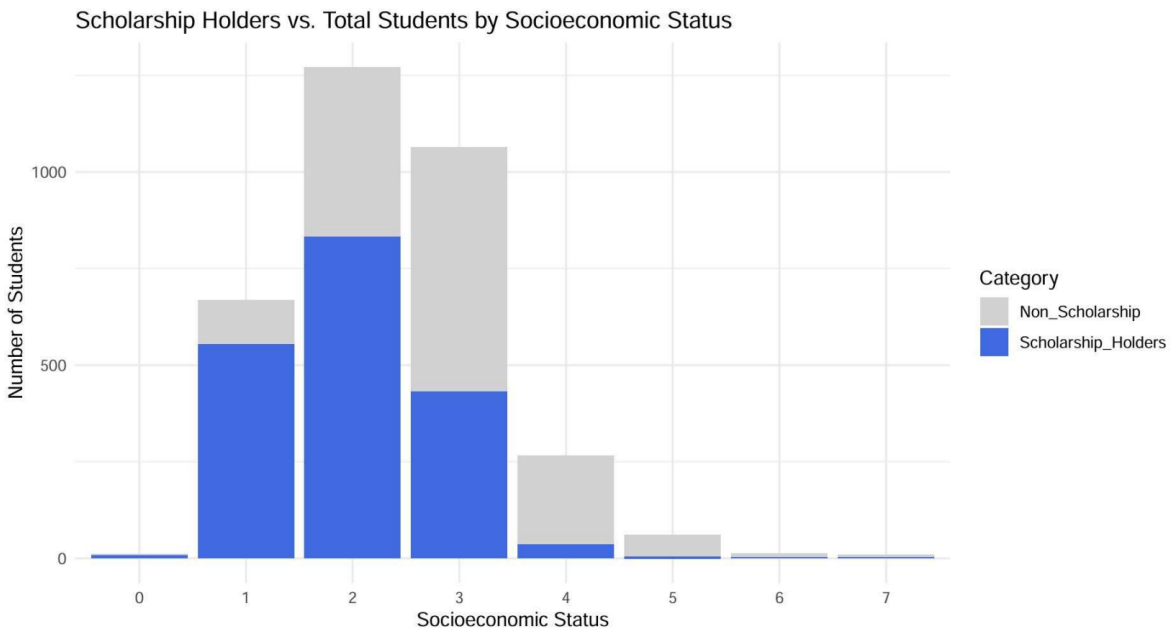
Now Fisher's Exact test has the probability of obtaining observed table is

$$p = \frac{\left(\prod_{i=1}^{m} r_i!\right) \cdot \left(\prod_{j=1}^{n} c_j!\right)}{N! \cdot \left(\prod_{i=1}^{m} \prod_{j=1}^{n} a_{ij}!\right)}$$

And after taking $\alpha = 0.05$ and running the R code of our data, we get $p = 9.999e^{-5}$. As the p-value is extremely small. we reject H_0 and conclude that scholarship distribution depends on SES.

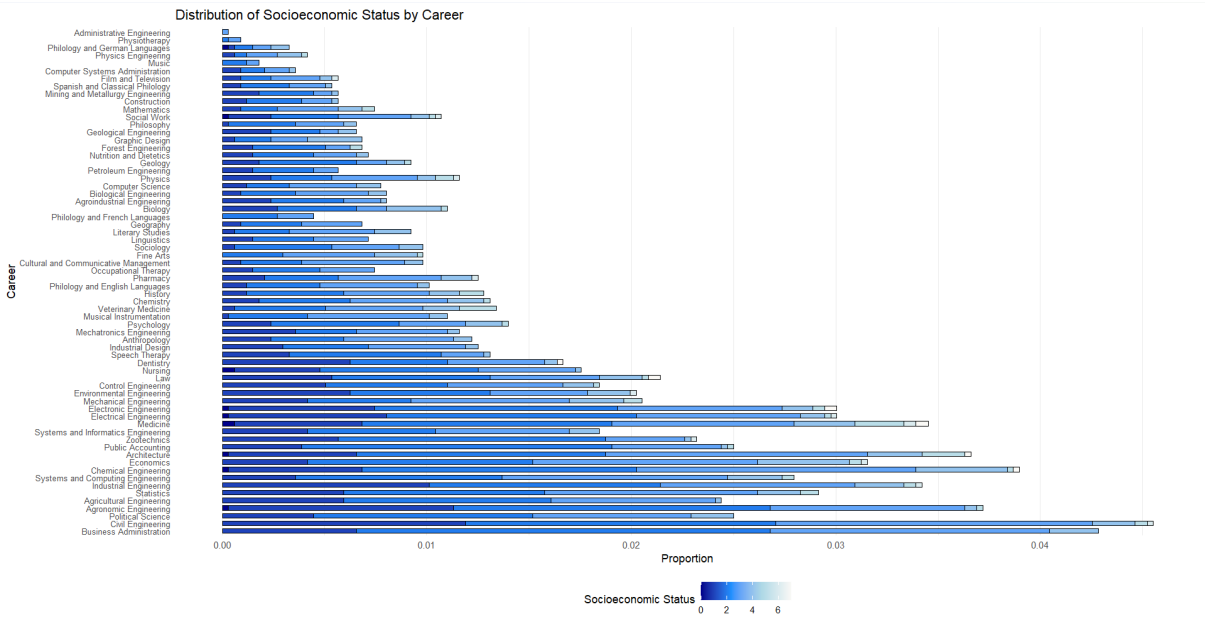
Causes that might favor this behaviour:

- There might be some govt scheme/Finantial aid programs which target lowe SES students.
- Higher SES students might have other funding sources like loans, family support, private scholarships.


Scholarship Holders vs. Total Students by Socioeconomic Status
Number of Students
1000
500
0
0
1
2
3
4
5
6
7
Socioeconomic Status
Category
Non_Scholarship
Scholarship_Holders

Objective

The aim of this project is to explore how a student's socioeconomic status (SES) relates to their choice of career. Specifically, we investigate whether tuition fees—an outcome that might be influenced by SES—are also linked to different career paths. Since tuition fees are officially determined by a metric called PBM (a socio-economic score), we explore whether other factors, like career choice, also play a role.



Observations from the Data

1. **Enrollment Variation:** The number of students varies significantly across different courses. For example, courses such as Civil Engineering, Business Administration, and Political Science have high enrollment, while courses like Administrative Engineering, Physiotherapy, and Philosophy & German have low enrollment.

2. **SES Distribution by Career:** The SES composition of students varies across careers.

   - Careers like *Graphic Design* show a higher proportion of high-SES students.
   - Careers like *Zootechnics* and *Agricultural Engineering* show a higher proportion of low-SES students.

This suggests that socioeconomic status may influence career choice.

Why Correlation Doesn't Work Here

Career choice is qualitative (categorical) data, while SES is quantitative. Therefore, computing a direct correlation is not valid.

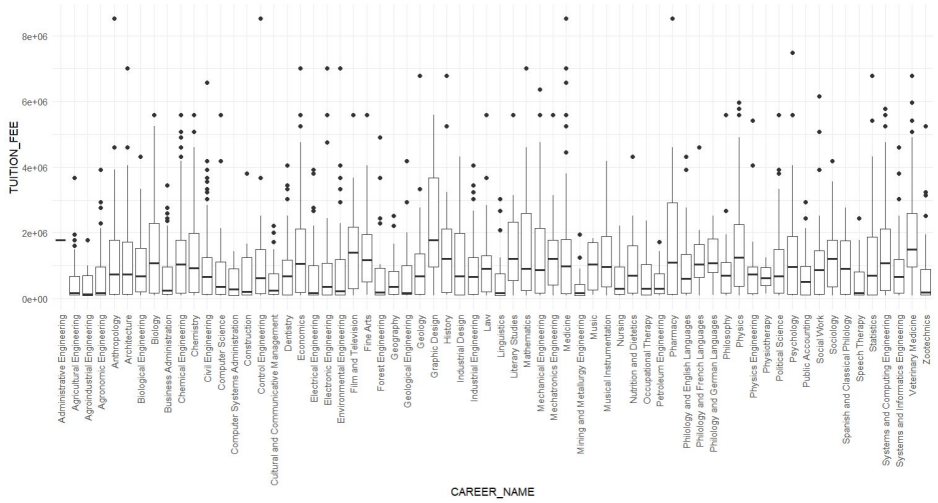
Also, we cannot assign arbitrary numbers to career choices because there's no meaningful scale or order among them. Instead, we need a way to *numerically classify* career choices based on an associated measurable quantity—such as tuition fees.

Career Choice and Tuition Fees

We noticed a pattern:

- Careers with a higher proportion of high-SES students often have **higher tuition fees**.
- Careers with a higher proportion of low-SES students often have **lower tuition fees**.

This leads us to the following hypothesis:



Hypothesis 3

Career choices influence tuition fees.

Statistical Hypothesis Testing

Null and Alternative Hypotheses

- H_0: Career choice does not affect tuition fee.
- H_1: Career choice does affect tuition fee.

Checking Assumptions for ANOVA

We tested for:

- Normality
- Homoscedasticity (equal variance)

Both assumptions were **violated**, so we used a **non-parametric test** instead.

Kruskal-Wallis Test

We applied the Kruskal-Wallis test to examine whether tuition fees significantly differ across career choices.

Result: We rejected the null hypothesis. **Conclusion:** Career choices **do** affect average tuition fees.

Categorizing Career Choices

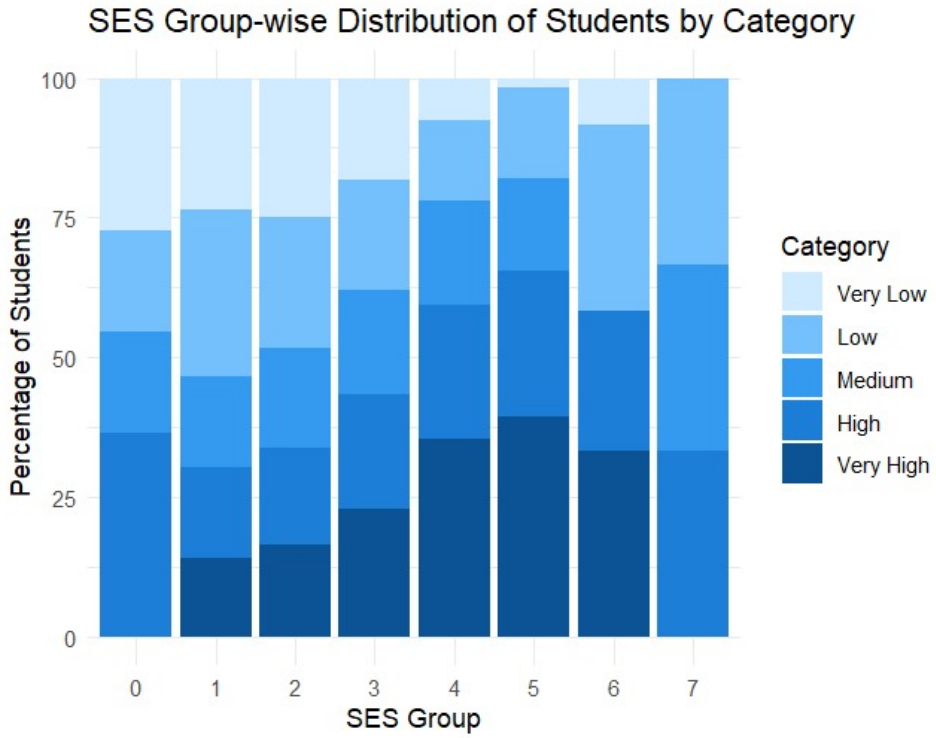
Based on tuition fees, we divided career choices into 5 groups:

- **Very High Fee:** Top 20 percentile (80–100%)
- **High Fee:** 60–80%
- **Medium Fee:** 40–60%
- **Low Fee:** 20–40%
- **Very Low Fee:** Bottom 20 percentile (0–20%)

This classification allows us to analyze whether SES distribution varies across these tuition fee tiers.

	A	B	C
1	Career N	Average	Tuition Cat
2	Graphic I	2E+06	Very High
3	Veterinar	2E+06	Very High
4	Pharmac	2E+06	Very High
5	Administ	2E+06	Very High
6	Physics	2E+06	Very High
7	Literary S	2E+06	Very High
8	Mathema	2E+06	Very High
9	Film and T	2E+06	Very High
10	Fine Arts	2E+06	Very High
11	History	2E+06	Very High
12	Biology	1E+06	Very High
13	Systems	1E+06	Very High
14	Economic	1E+06	Very High
15	Mechanic	1E+06	Very High
16	Medicine	1E+06	High
17	Sociology	1E+06	High
18	Mechatro	1E+06	High
19	Philology	1E+06	High
20	Psycholo	1E+06	High
21	Chemistr	1E+06	High
22	Chemical	1E+06	High
23	Anthropo	1E+06	High
24	Philology	1E+06	High
25	Musical I	1E+06	High
26	Physics I	1E+06	High
27	Social W	1E+06	High
28	Statistics	1E+06	High
29	Architect	1E+06	Medium
30	Spanish	1E+06	Medium
31	Control E	1E+06	Medium
32	Industrial	1E+06	Medium
33	Compute	1E+06	Medium
34	Geology	1E+06	Medium
35	Biologica	1E+06	Medium
36	Political S	1E+06	Medium
37	Philology	1E+06	Medium
38	Nutrition	1E+06	Medium
39	Law	998738	Medium
40	Philosoph	992094	Medium
41	Music	991702	Medium
42	Systems	902064	Low
43	Dentistry	890133	Low
44	Civil Engi	887413	Low
45	Environm	871007	Low
46	Industrial	852968	Low
47	Forest Er	829964	Low
48	Geologic	806057	Low
49	Electroni	798427	Low
50	Construc	747444	Low
51	Physioth	676095	Low
52	Geograp	656143	Low
53	Public Ac	647284	Low
54	Electrical	639788	Low
55	Zootechr	632475	Very Low
56	Business	632394	Very Low
57	Agronom	616419	Very Low
58	Nursing	611821	Very Low
59	Occupati	600925	Very Low
60	Cultural a	600090	Very Low
61	Linguistic	587519	Very Low
62	Petroleur	530988	Very Low
63	Compute	522895	Very Low
64	Speech T	494026	Very Low
65	Agricultu	483410	Very Low
66	Mining ar	414061	Very Low
67	Agroindu	388235	Very Low

Now we will plot a stacked bar graph of SES Group-wise Distribution of Students by Catagory



Observation: Students from higher socioeconomic status (SES) backgrounds tend to pursue career choices associated with higher tuition fees

Discussion: Does SES Determine Career Choice?

The data suggests that low-SES students tend to choose low-fee careers, possibly due to financial constraints. However, we must be cautious:

- PBM affects tuition, and SES affects PBM.
- So, SES may affect tuition fees *indirectly* via PBM.
- Additionally, **other unobserved factors** (e.g., job prospects, interests) may also influence career choices.

Therefore, while the trend is clear, we cannot definitively conclude causality without more data on motivation, return on investment, or other variables.

Conclusion

There is statistical evidence that career choice influences tuition fee levels, and there is a visible pattern where SES seems to affect the kind of careers students choose. However, we must interpret this relationship cautiously due to confounding factors.

Regression Analysis Report

1 Objective

The goal of this analysis is to understand how socioeconomic and academic factors affect the tuition fees paid by students at a Colombian public university. Specifically, we aim to build a regression model that accurately captures this relationship and can be used for future prediction and interpretation. Our independent variable is Tuition fee.

2 Variable Selection and Motivation

We began by identifying potentially influential variables including PBM (Basic Tuition Score), SES (Socioeconomic Status), parental income, school type, and tuition-related metrics. To avoid multicollinearity—a condition where independent variables are highly correlated, potentially distorting the estimates—we computed the Variance Inflation Factor (VIF).

Threshold: Variables with VIF > 5 were excluded. **Exception:** PBM was retained despite its high VIF due to its policy relevance and known strong influence on tuition fees. We decided on keeping the following:

Predictor	**VIF**
SOCIOECONOMIC STATUS	*1.782*
FATHER INCOME	*1.380*
MOTHER INCOME	*1.311*
MONTHLY SCHOOL FEE	*2.487*
PBM	*5.980*

Table 1: Model Metrics After Outlier Removal

3 Initial Exploration via Scatterplots

We visualized the relationship between tuition fees and various predictors using scatterplots.

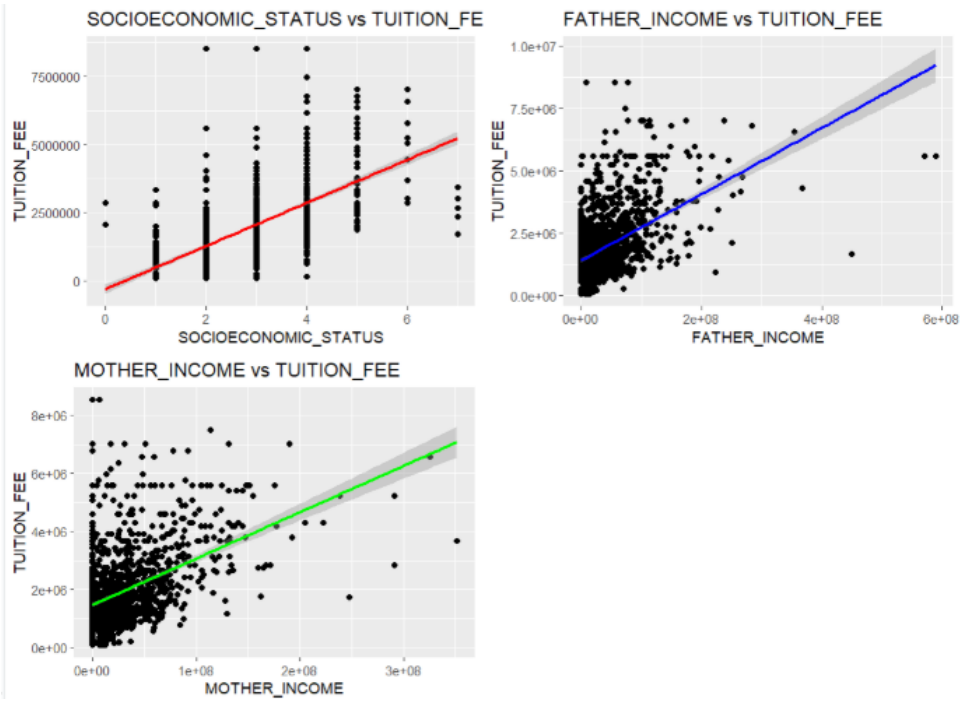


Figure 1: Scatterplots showing relationships between Tuition Fee and predictors such as PBM, SES, and Parental Income.

Inference: The predictors demonstrate positive correlation with tuition fees. However, the trend does not look exactly linear. We still go ahead with a linear model.

4 Linear Regression

We fit an initial multiple linear regression model using the selected variables using R. We get the following model:

$$\begin{aligned}\text{Tuition Fee} = \; & -984805 + 46414 \cdot \text{SES} + 0.0125 \cdot \text{Father Income} \\ & + 0.0523 \cdot \text{Mother Income} + 151.7 \cdot \text{Monthly School Fee} + 4568 \cdot \text{PBM}\end{aligned}$$

But while checking the residual plots we ran into problems.

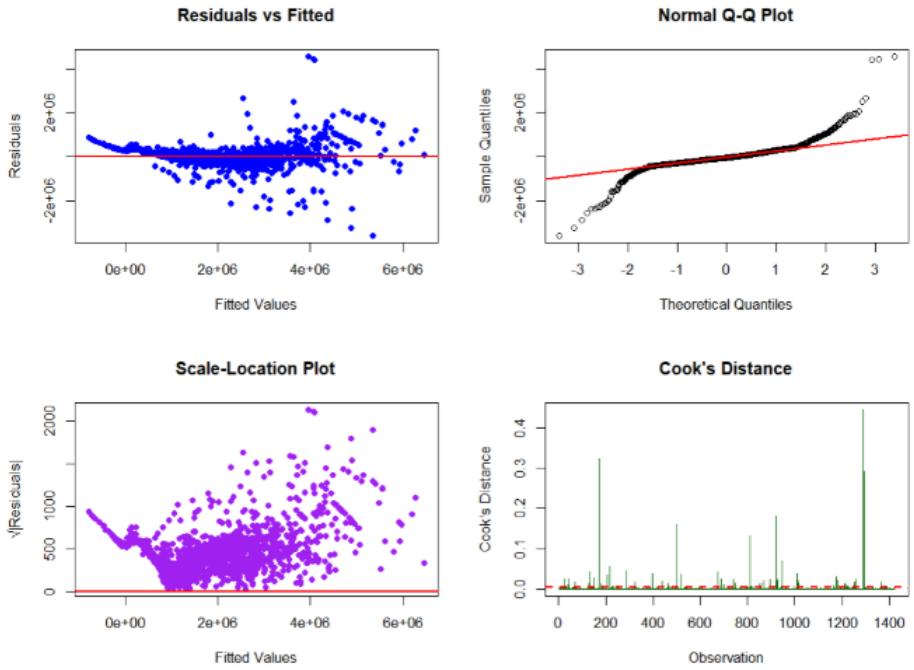


Figure 2: Residuals vs Fitted values for linear model.

Issues Identified:

- **Heteroscedasticity:** Residuals had increasing spread.
- **Non-normality:** Q-Q plot showed heavy tails.

These observations indicated the inadequacy of a linear model. We next tried polynomial regression.

5 Polynomial Regression

We fitted polynomial regression models of degrees 1 through 5 and compared them using R tools. Comparison metrics used:

- Adjusted R^2
- AIC (Akaike Information Criterion)

Best Fit: Degree 5 polynomial initially performed best. However, several terms were statistically insignificant. We simplified the model by retaining only significant terms.

$$\begin{aligned}\text{Tuition Fee} \sim\ & \text{poly(SES, 4)} + \text{poly(Father Income, 1)} + \text{poly(Mother Income, 2)} \\ & + \text{poly(Monthly School Fee, 4)} + \text{PBM}\end{aligned}$$

6 Comparing Linear and Polynomial models

Metric	Linear	Polynomial
Adjusted R^2	*0.849*	*0.8734*
Max Error	*4574625*	*3954657*

Table 2: Comparison of Linear and Polynomial Regression Models

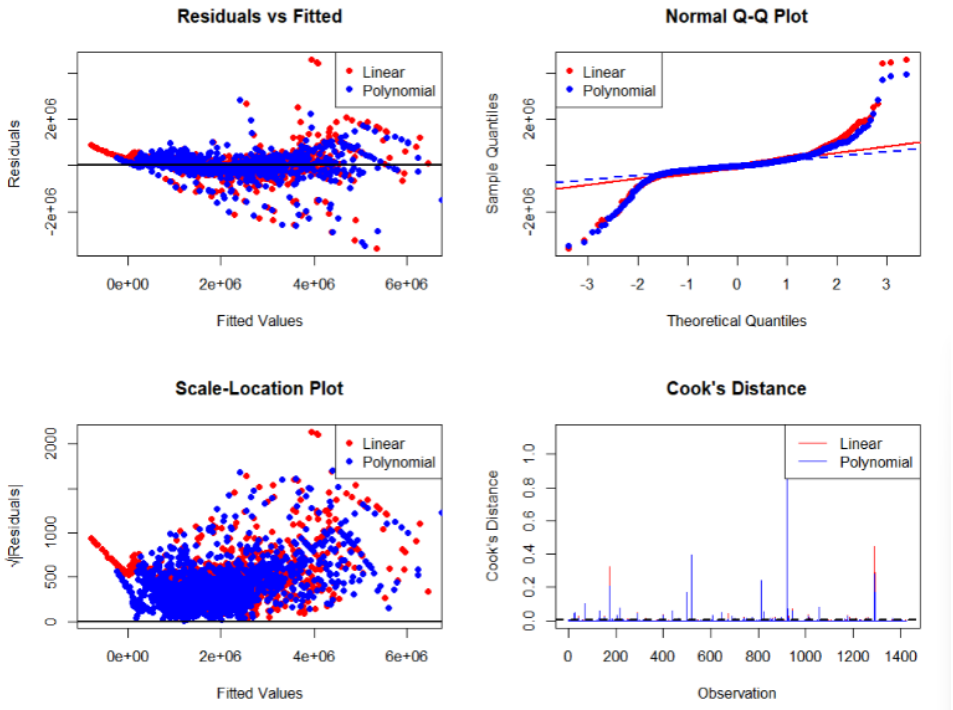


Figure 3: Comparison of residuals from Linear and Polynomial models.

Inference: Polynomial regression performed better than linear regression, but it is still not perfect.

Improvements with Polynomial Regression:

- Captures non-linearity better
- Reduces bias
- Handles heteroscedasticity better

Remaining Issues:

- Residual patterns persist
- Outliers still pose a problem. Infact, polynomial regression is more sensitive to outliers (as can be seen from the cook's distance plot)

7 Outlier Detection and Cleaning

Outliers were detected using **Cook's Distance**, with a threshold of $\frac{4}{n}$. Roughly 80 outliears were detected. We found that most of these students had extremely high parental incomes ($>$ 100M COP), hence they do not represent the whole student body, justifying the removal of most of these outliers. These outliers were removed and the models were re-evaluated.

8 Re-Fitting Models Without Outliers

Both linear and polynomial models were refitted using the cleaned dataset.

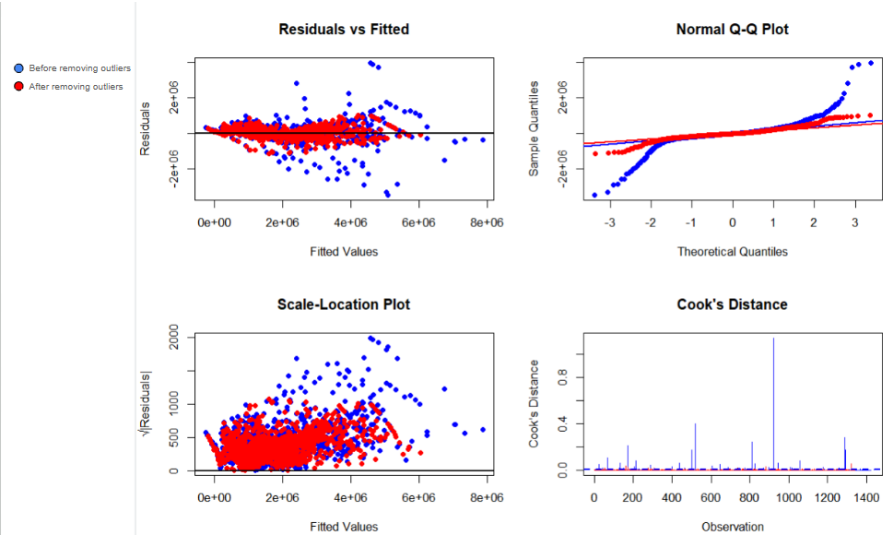


Figure 4: Residuals of the polynomial model after removing outliers.

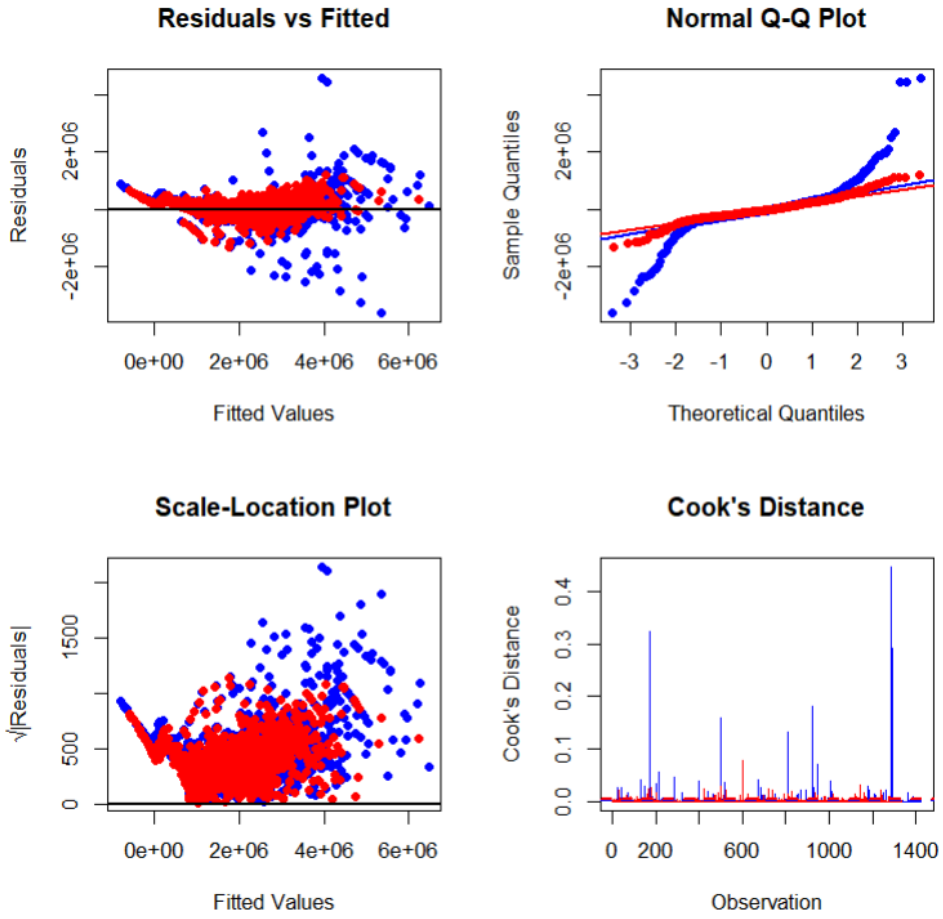


Figure 5: Residuals of the linear model after removing outliers.

Now we recompare the linear and polynomial models.

Model	**Adjusted R^2**	**AIC**
Linear (cleaned)	*0.9296*	*3.7146*
Polynomial (cleaned)	*0.9495*	*3.7081*

Table 3: Model Metrics After Outlier Removal

We can clearly see that the linear model is almost as good as the polynomial model, and since polynomial models are more complex, we land on the linear model as the final model.

9 Final Model

The final chosen model is:

$$\begin{aligned}\text{Tuition Fee} = &-802300 + 21430 \cdot \text{SES} + 17760 \cdot \text{Father Income} + 13940 \cdot \text{Mother Income} \\ &+ 14770 \cdot \text{Monthly School Fee} + 42360 \cdot \text{PBM}\end{aligned}$$

Key Findings:

- PBM remains the most influential predictor.
- Parental income and historical school fee data add significant explanatory value.
- SES has a non-linear but meaningful impact.

Correlation of two variables:

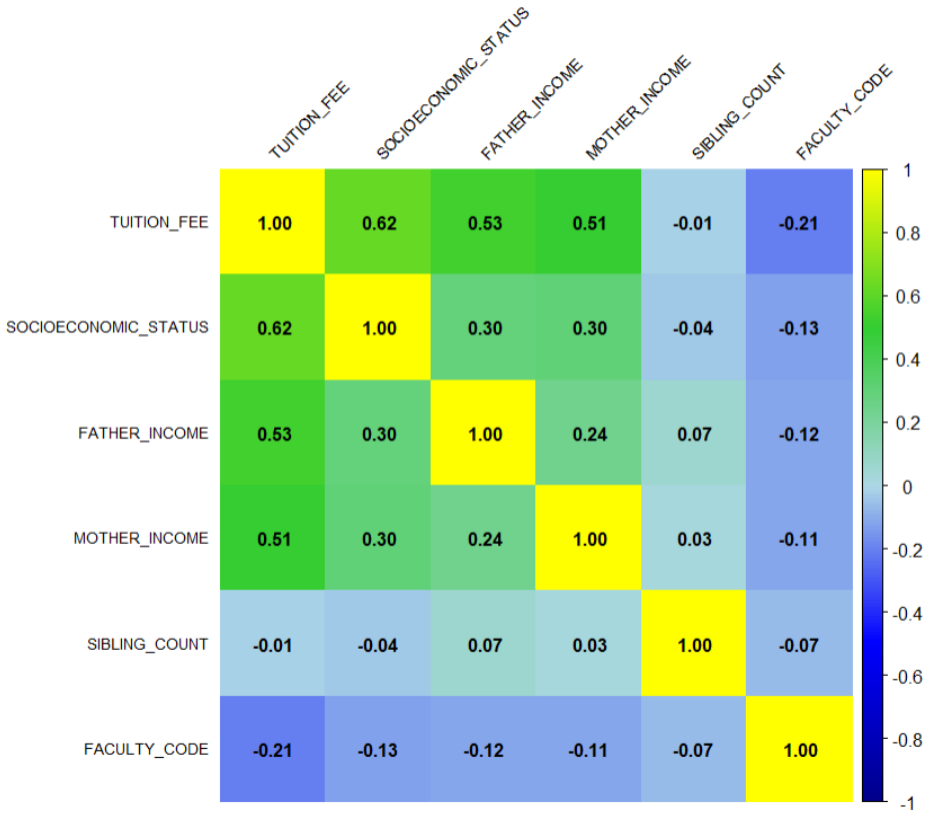
The correlation of two variables X, Y is given as

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X, \sigma_Y} = \frac{\sum XY - \sum X \sum Y}{\sqrt{(\sum X^2 - (\sum X)^2) \cdot (\sum Y^2 - (\sum Y)^2)}}$$

Correlation values interprets as

- $\rho_{X,Y} = +1$: Perfect positive correlation ($X \propto Y$)
- $\rho_{X,Y} = -1$: Perfect negative correlation ($X \propto \frac{1}{Y}$)
- $\rho_{X,Y} = 0$: No Linear correlation (we can't say anything about the relationship)
- $\rho_{X,Y} \in (0, 1)$: Positive correlation (X, Y are not completely proportional, but as $X \uparrow$, Ytends to $\uparrow$)
- $\rho_{X,Y} \in (-1, 0)$: Negaitive correlation (X, Y are not completely inversely proportional, but as $X \uparrow$, Ytends to $\downarrow$)

Correlation of Numerical attributes



Key observations:

- SIBLING_COUNT does not correlate any other variable significantly.
- There are some "seemingly numerical attributes" which does not really correlate properly with anything.
- SOCIOECONOMIC_STATUS assignments has some meaningful insights.

Correlation of Numerical & Transformed Non-numeric attributes:

Key observations:

- SCHOOL_TYPE is assigned as (B1)
  - Private = 0.95
  - Public = 1

  but differnt assignments does not change the correlation if the order i.e. Private ¡ Public is not changed. That is because It is Invariant under Linear Transformation, Suppose X is our original variable, and we change it into $X' = aX + b$ where $a \neq 0$ then,

$$\rho_{X',Y} = \frac{\text{Cov}(X', Y)}{\sigma_{X'}\sigma_Y} = \frac{\text{Cov}(aX + b, Y)}{(|a| \cdot \sigma_X)\sigma_Y} = \frac{a \cdot \text{Cov}(X, Y)}{|a| \cdot \sigma_X\sigma_Y} = \pm\rho_{X,Y}$$

  We get that

$$\rho_{X',Y} = \begin{cases} \rho_{X,Y} & \text{if } a > 0 \\ -\rho_{X,Y} & \text{if } a < 0 \end{cases}$$

- CAREER_CODE doesn't tell us Career's attribute, rather it is just enumeration. So, it's not worth a hassle to find correlation.

Limitations:

The dataset covers the first semester of 2021, after Colombia's COVID-19 lockdown (March–August 2020), reflecting a post-lockdown reality where families and institutions had adapted to pandemic-related challenges. As it includes data from only one semester, it cannot capture long-term trends but serves as an exploratory foundation for studying socioeconomic factors in education. Like other cross-sectional studies, it provides insights into educational inequalities, addressing a gap in Latin American data on educational equity. Future work could integrate data from multiple semesters to identify trends and support more robust policy development.

References:

- All the abstract hypothesis test theories came from Wikipedia
  - https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
  - https://en.wikipedia.org/wiki/Levene%27s_test
  - https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_test
- The SES Score came from
  - http://www.dane.gov.co
- All the R-Codes made during this project can be found here
  - https://github.com/kanaksanghvi/Stat1-project
- Tests for model fitting
  - https://datascienceplus.com/identify-describe-plot-and-removing-the-outliers

Acknowledgement:

Thanks to our beloved professor **Rituparna Sen**, we had the opportunity to deep-dive into a dataset like this, learn various statistical techniques and real life applications.