Statistical Analysis of Baccalaureate Exam Results

Created by: Póra Boglárka, Rafain Emőke, Simon Andrea

Our project focuses on conducting a Statistical Analysis of Baccalaureate Exam Results in Romania, exploring data from 2020 to 2023. In the introduction, we emphasize the significance of the baccalaureate exam, with a historical overview, including its introduction in 1925 and changes over time, particularly during the communist era and contemporary measures to maintain credibility.

We set a comprehensive research question: "To what extent do various factors influence the academic performance and promotion outcomes of high school students?" Factors include gender, educational form, language, Romanian and native grades, mandatory and optional grades, candidate environment, and the appeal process. We aimed to use inferential statistical methods, including correlation analysis, T-tests, ANOVAs, Chi-Square tests, and regression analysis, to answer the research question.

Before embarking on the project, we explored existing studies and found a webpage with visualizations. Our objectives include testing hypotheses and concluding the relationships between variables, anticipating correlations between grades.

Data for the study was collected from data.gov.ro, with four separate datasets representing the years 2020 to 2023. We used Python and Google Colaboratory for analysis. The datasets contain both categorical (e.g., candidate code, gender) and numerical data (e.g., grades). The data cleaning process involved transforming, adding, and dropping variables, resulting in a standardized set of final variables for analysis.

We were anticipating conclusions that demonstrate a high level of dependence between variables, considering factors such as subject performance correlations and the impact of appeal grades on averages. Overall, the study aims to provide insights into the complex relationships influencing high school students' academic performance in Romania.

Descriptive Statistics

Most Frequent Categorical Variables between 2020 and 2023

Feature	Most Frequent Value
Gender	Slightly male-dominated, roughly the same
Specialization	Mathematics and Informatics
Profile	STEM (‘Real’)
Educational Path	theory-based (‘Teoretic’)
Form of Study	attendance by day
Candidate Environment	urban
Native Paper (non-Romanian candidates)	Hungarian
Mandatory Paper	Mathematics
Chosen Paper	Biology
Foreign Language	English

Figure: Table of most frequent values

Important Measures between 2020 and 2023

Measure/Year	2020	2021	2022	2023
Passing Rate of Candidates	61.34%	66.28%	72.23%	72.02%
Count of Candidates	155650	133664	126453	130522
Average Romanian Grade (Passing)	7.97	7.95	7.86	7.83
Average Native Grade (Passing)	8.0	8.17	8.07	8.04
Average Mandatory Grade (Passing)	8.19	8.06	8.36	8.2
Average Chosen Grade (Passing)	8.41	8.44	8.43	8.29
Average Final Grade (Passing)	8.10	8.03	8.09	7.98

Figure: Table of important measures between 2020 and 2023

Inferences

Correlation:

H0: There is no linear correlation between the written exam grades and their appeals.

H1: There is a strong linear correlation between the written exam grades and their appeals.

T-Test:

H0: There is no significant difference in final averages between genders.

H1: The difference in final averages between genders is statistically significant.

H0: There is no significant difference in mandatory grades between profiles.

H1: The difference in mandatory grades between profiles is statistically significant.

ANOVA:

H0: There are no differences between average grades across specializations.

H1: There are differences between average grades across specializations.

H0: There are no differences between average grades across education forms.

H1: There are differences between average grades across education forms.

Regression:

H0: There is no linear relationship between final score and independent grades.

H1: There is linear relationship between final score and independent grades.

Chi-square test:

H0: There is no association between Status and Gender.

H1: There is an association between Status and Gender.

H0: There is no association between Status and Profile.

H1: There is an association between Status and Profile.

Contribution

Part Written	Student(s)
Intoduction	Póra Boglárka
Project Objective	Póra Boglárka
Data Collection and Modeling	Póra Boglárka
Descriptive Statistics	Rafain Emőke
Inferences	Póra Boglárka, Rafain Emőke & Simon Andrea
Conclusions	Póra Boglárka, Rafain Emőke & Simon Andrea

Figure: Table of Contribution

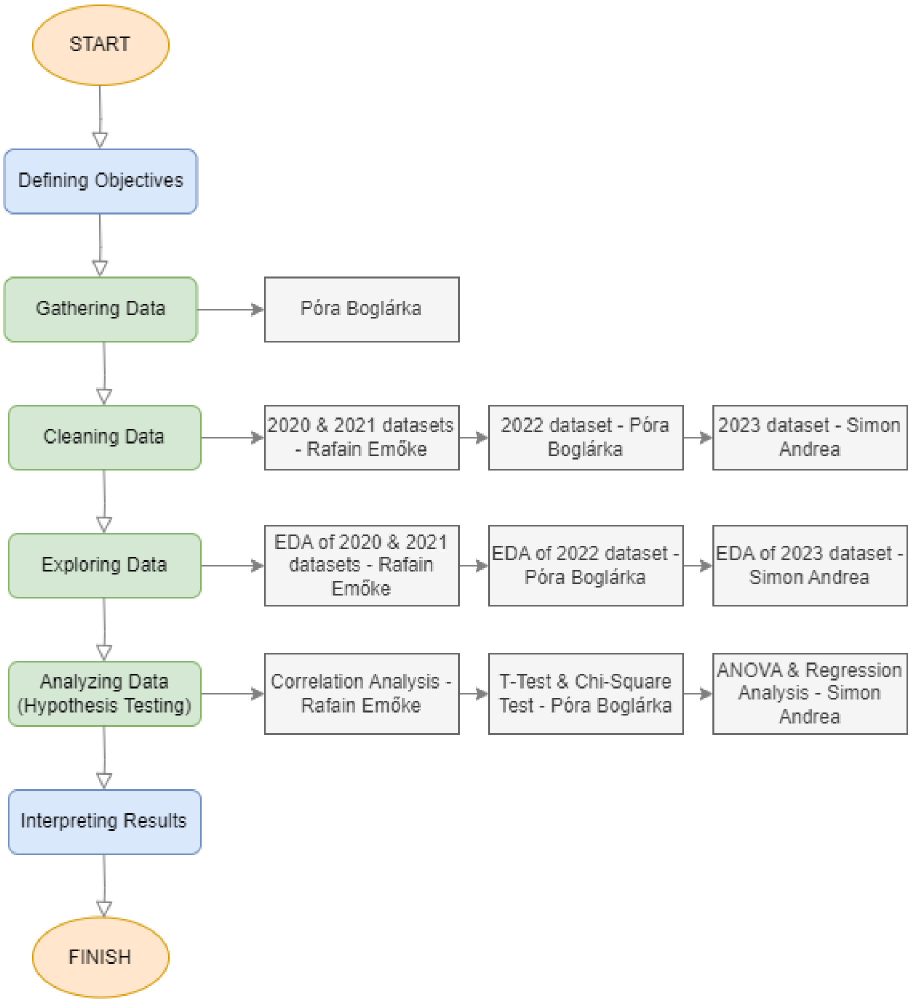
START
Defining Objectives
Gathering Data
Póra Boglárka
Cleaning Data
2020 & 2021 datasets - Rafain Emőke
2022 dataset - Póra Boglárka
2023 dataset - Simon Andrea
Exploring Data
EDA of 2020 & 2021 datasets - Rafain Emőke
EDA of 2022 dataset - Póra Boglárka
EDA of 2023 dataset - Simon Andrea
Analyzing Data (Hypothesis Testing)
Correlation Analysis - Rafain Emőke
T-Test & Chi-Square Test - Póra Boglárka
ANOVA & Regression Analysis - Simon Andrea
Interpreting Results
FINISH

Figure: Flow Chart of Analysis and Contribution