Mixed Model on Polimi Students Data

Riccardo Bertoglio - Paola Mussida

07/07/2020

Target

Apply Multilevel Model on PoliMI data to maximise performance on dropout prediction.

Students are rarely indipendent, they are clustered or nested in a way that makes the observations not truly independent.

Agenda

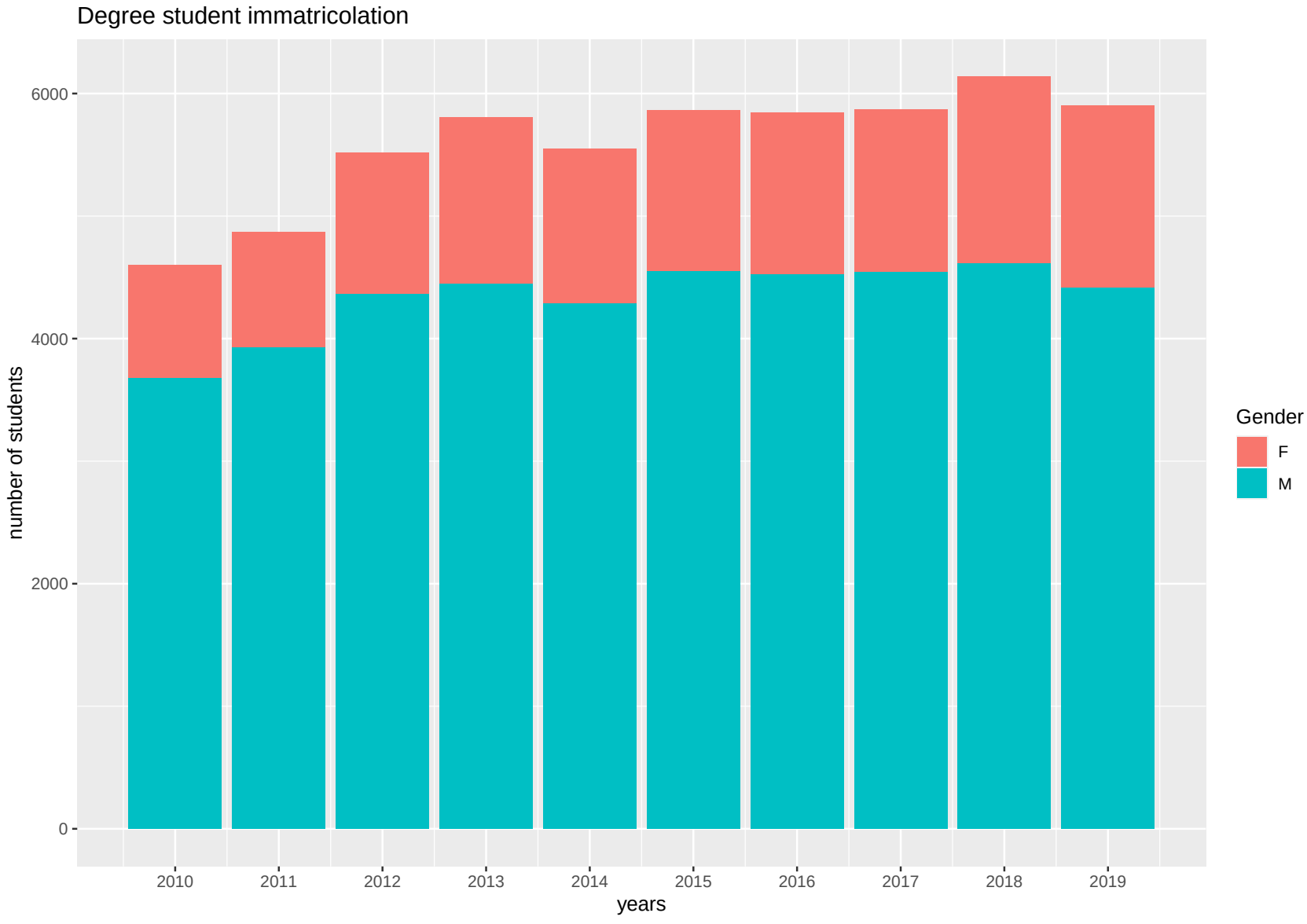
1. Data Preparation

- Data Exploration
- Data Cleaning
- Data Aggregation
- Feature Selection
- Missing Values

2. Models

- Data Partition
- Model Creation
- Results Evaluation

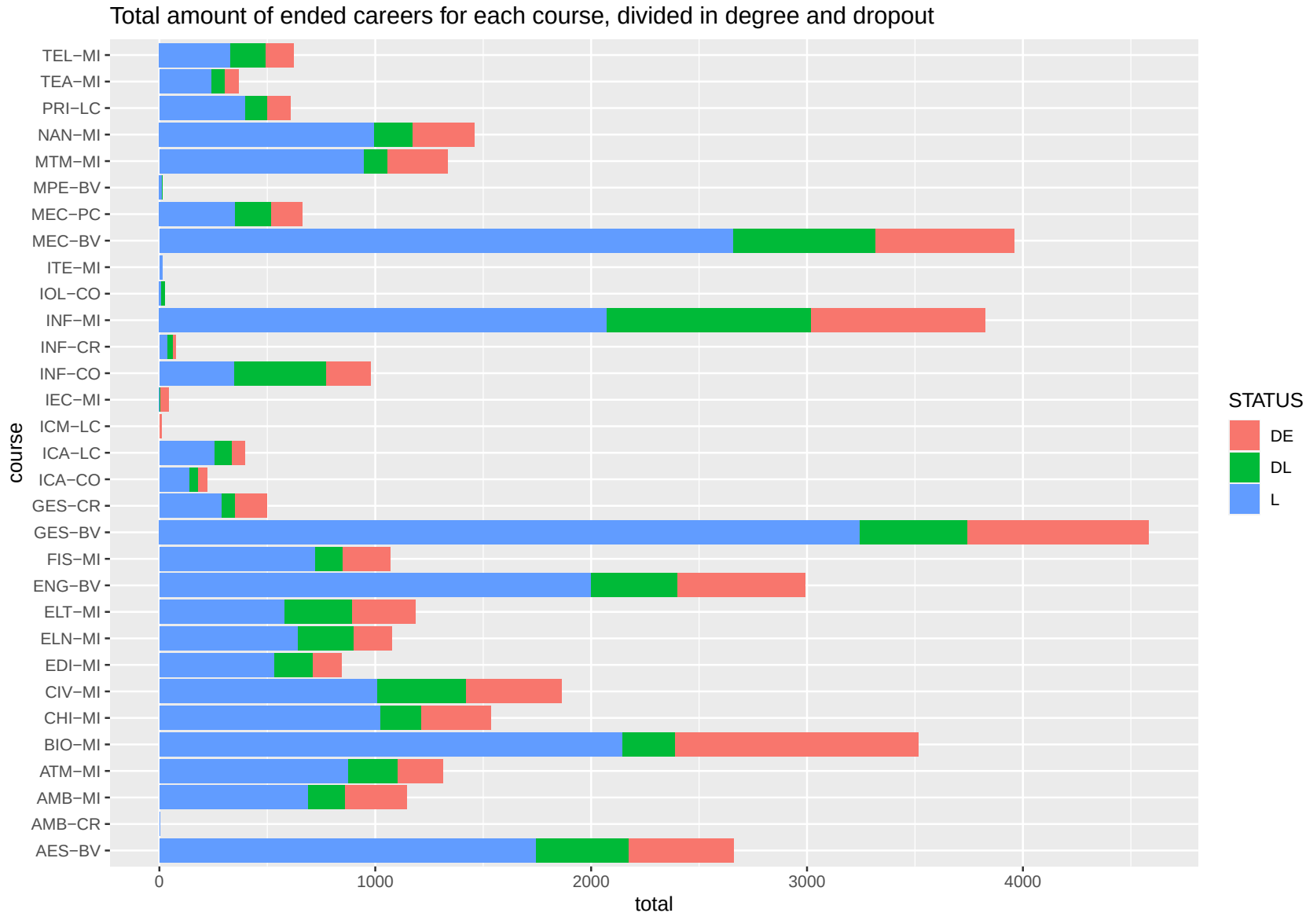
Careers


Degree student immatricolation
number of students
6000
4000
2000
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
years
Gender
F
M

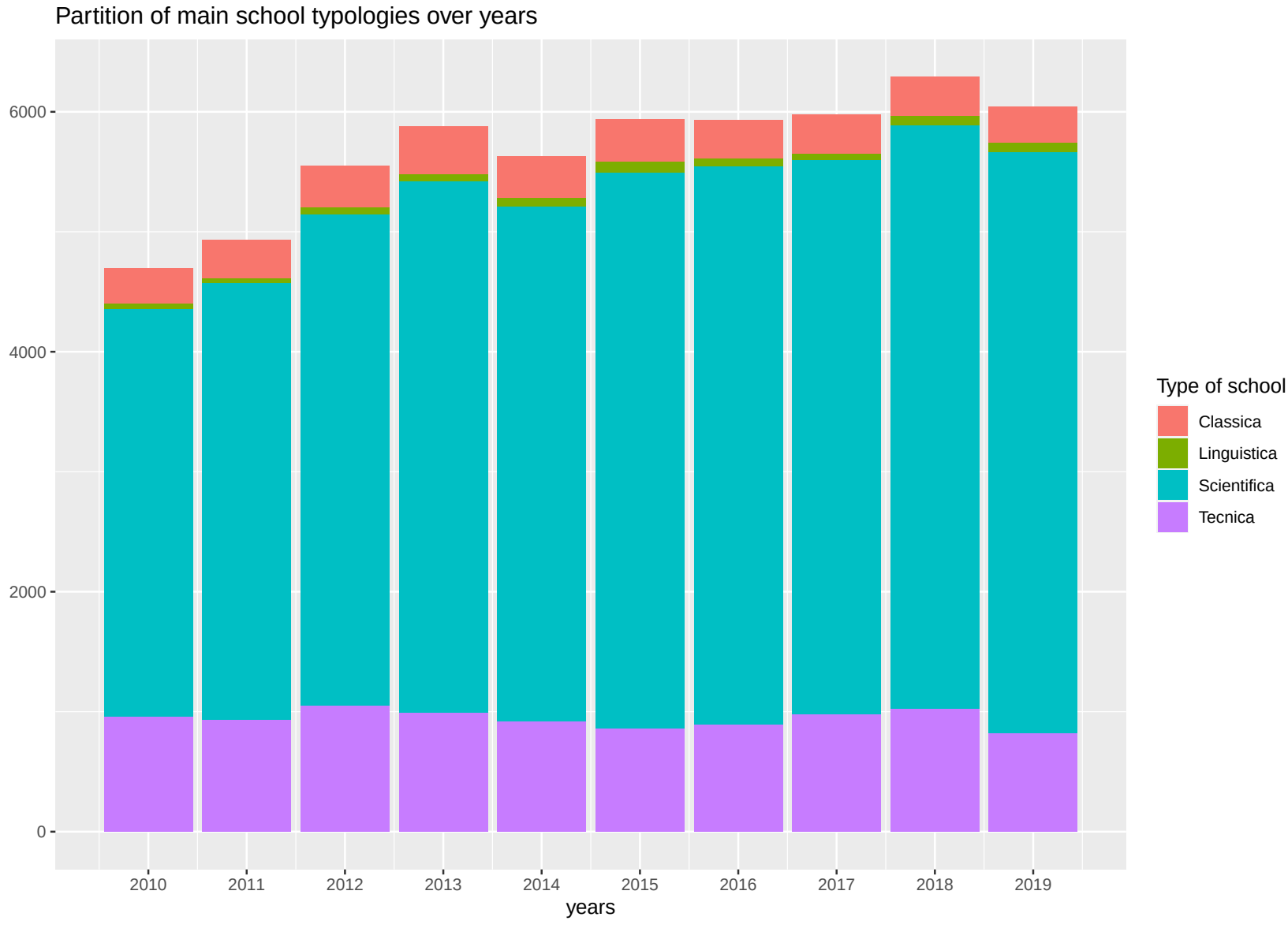
Career Description

- personal data
- previous studies
- admission score
- degree course
- exams
- degree / dropout

Ended careers by course

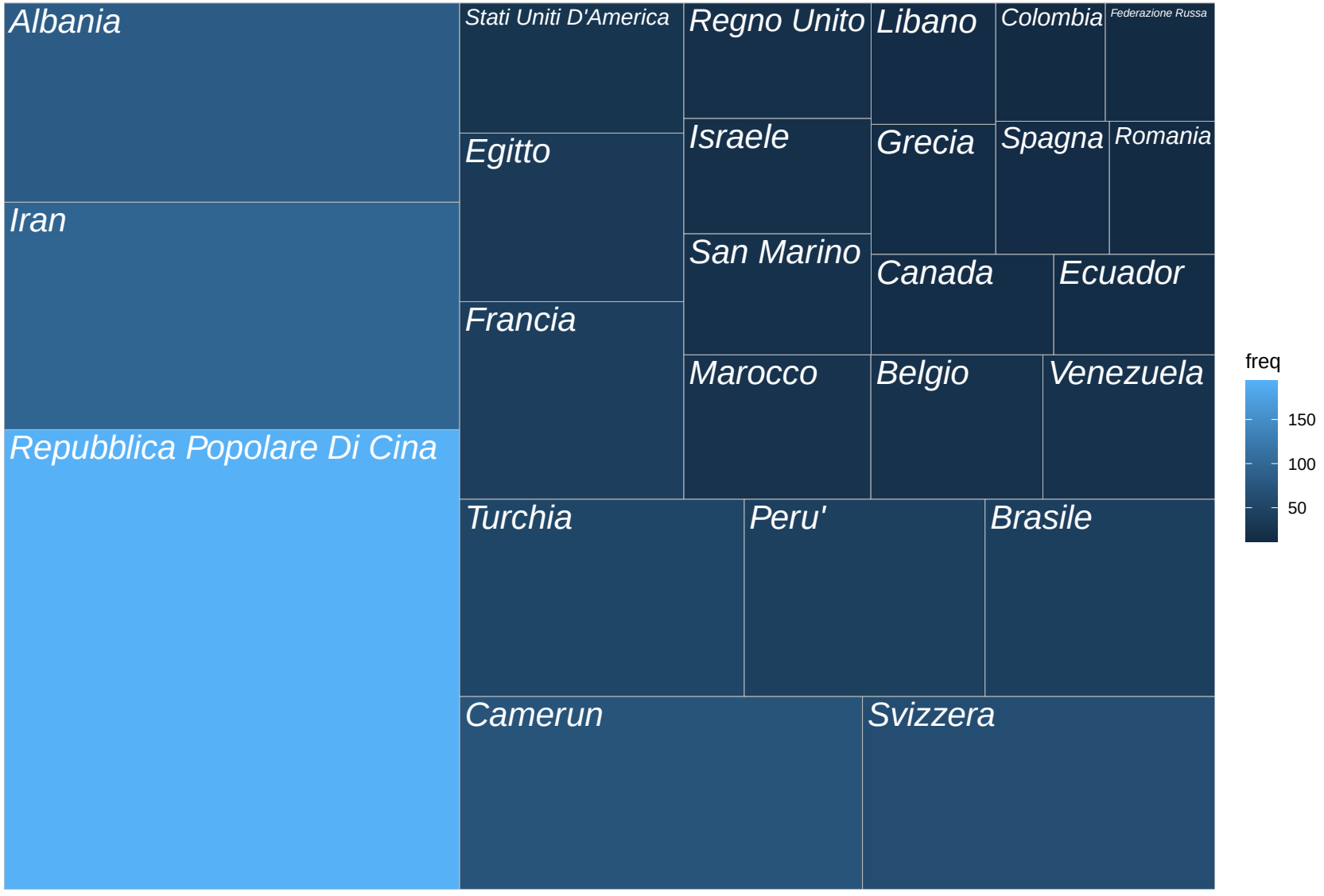


Previous Studies

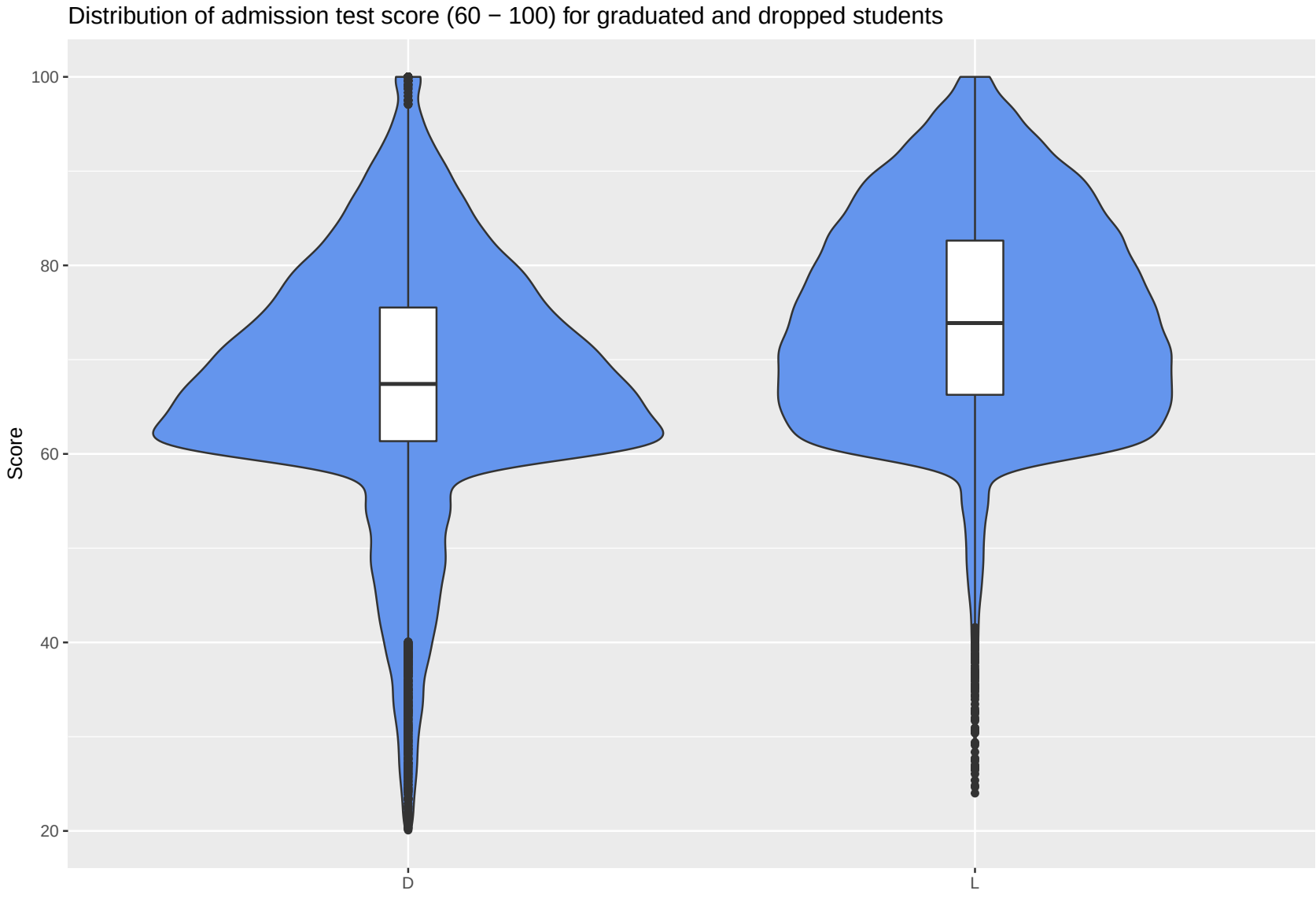


Foreign Previous Studies

Principal countries of foreign previuos studies



Admission Test - score distribution



Data Cleaning

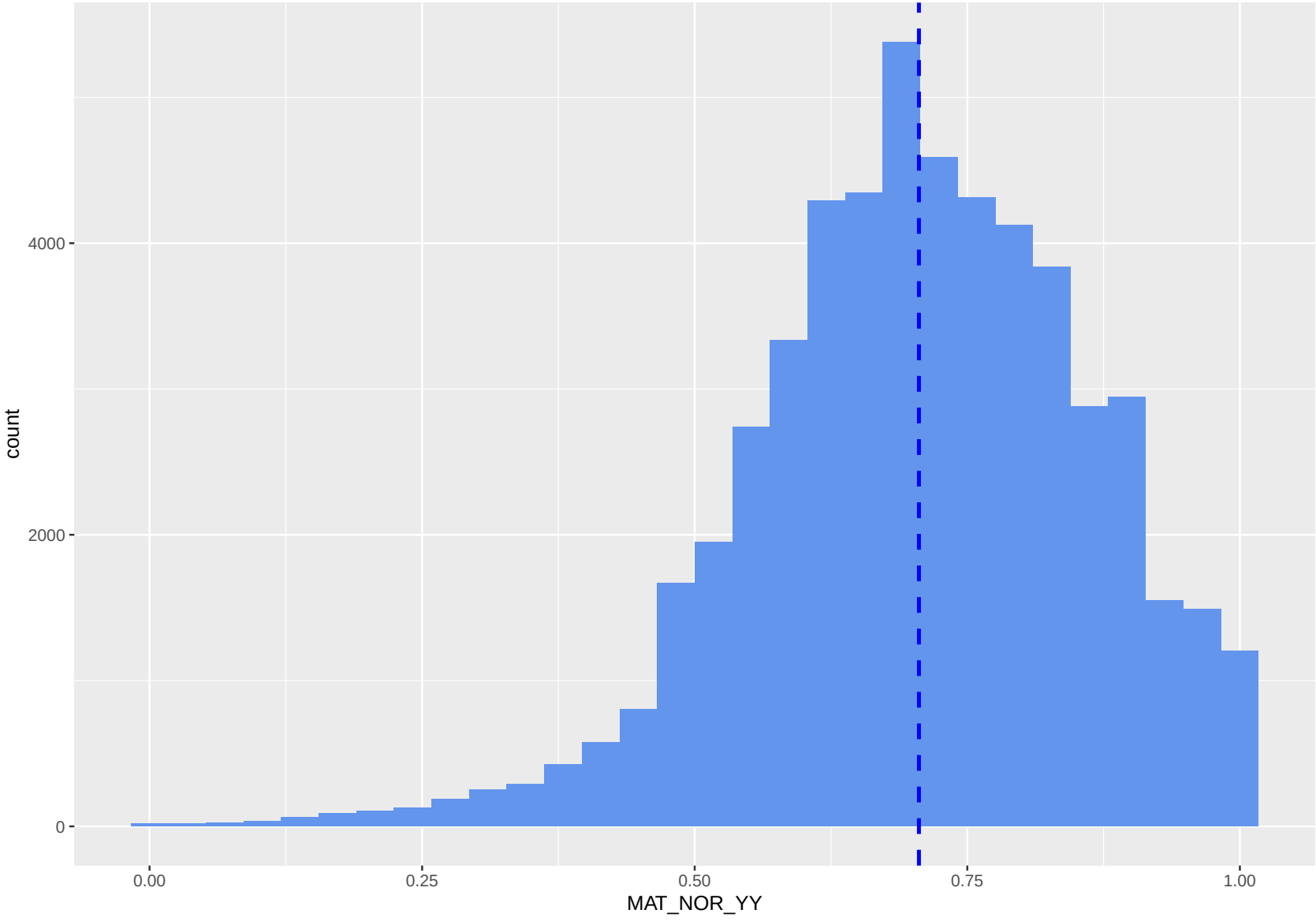
- We focus only on young students (less than 30 years old) because of data distribution. Older people have different paces and different ways to handle the studies, so they cannot be used to predict majority.
- We remove also maximum value for school grade different than 100 means that those people are outlier (for example, 60 as maximum grade was used in the past)

Admission test - partial values

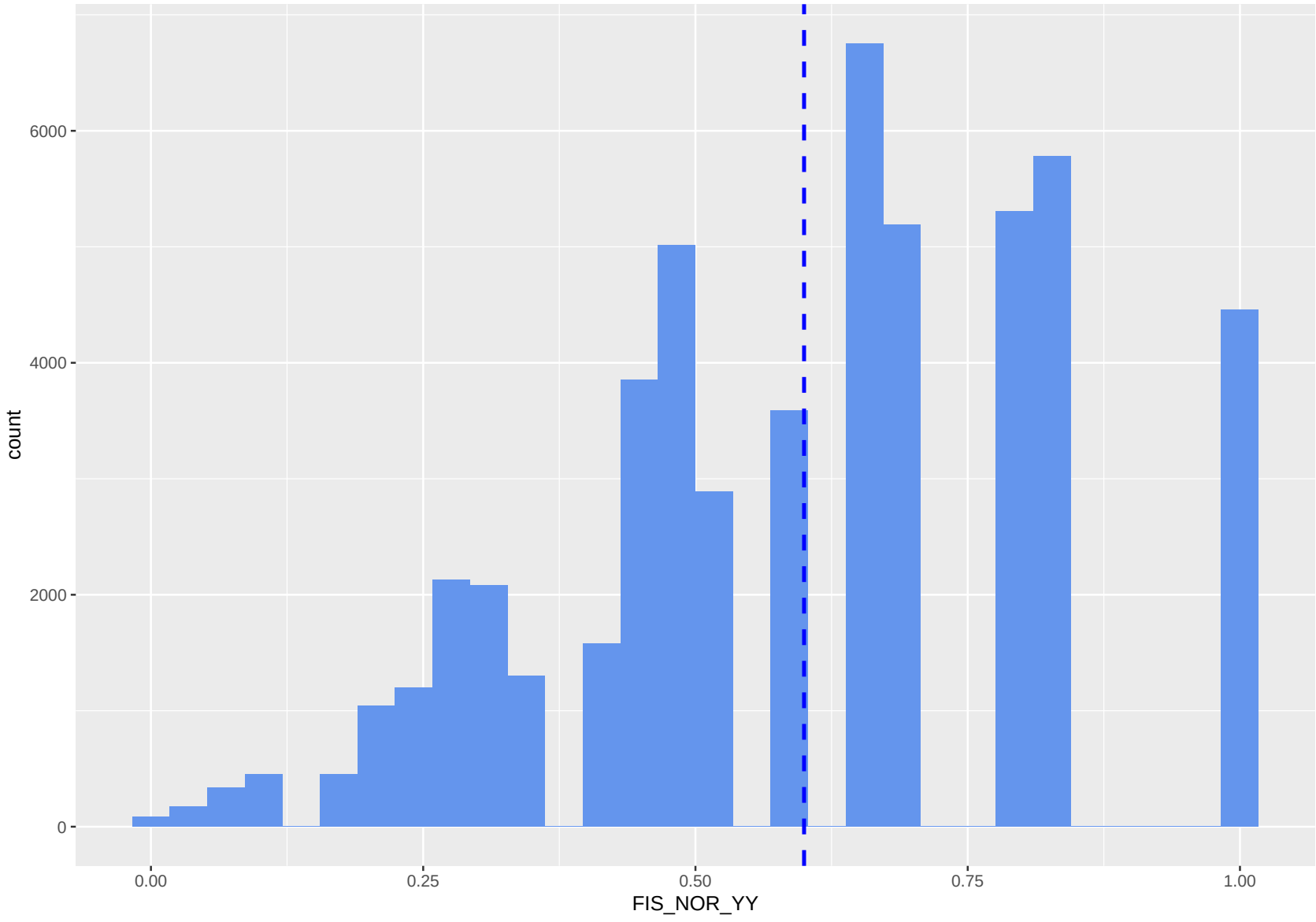
- The admission test if composed on 4 sections:
  - math
  - physics
  - reading comprehension
  - english
- each session has a different score and different importance, we normalized the value by mean

```
##  CV_NOR_YY ENG_NOR_YY FIS_NOR_YY MAT_NOR_YY
##  0.7829635  0.8331675  0.5997384  0.7056361
```

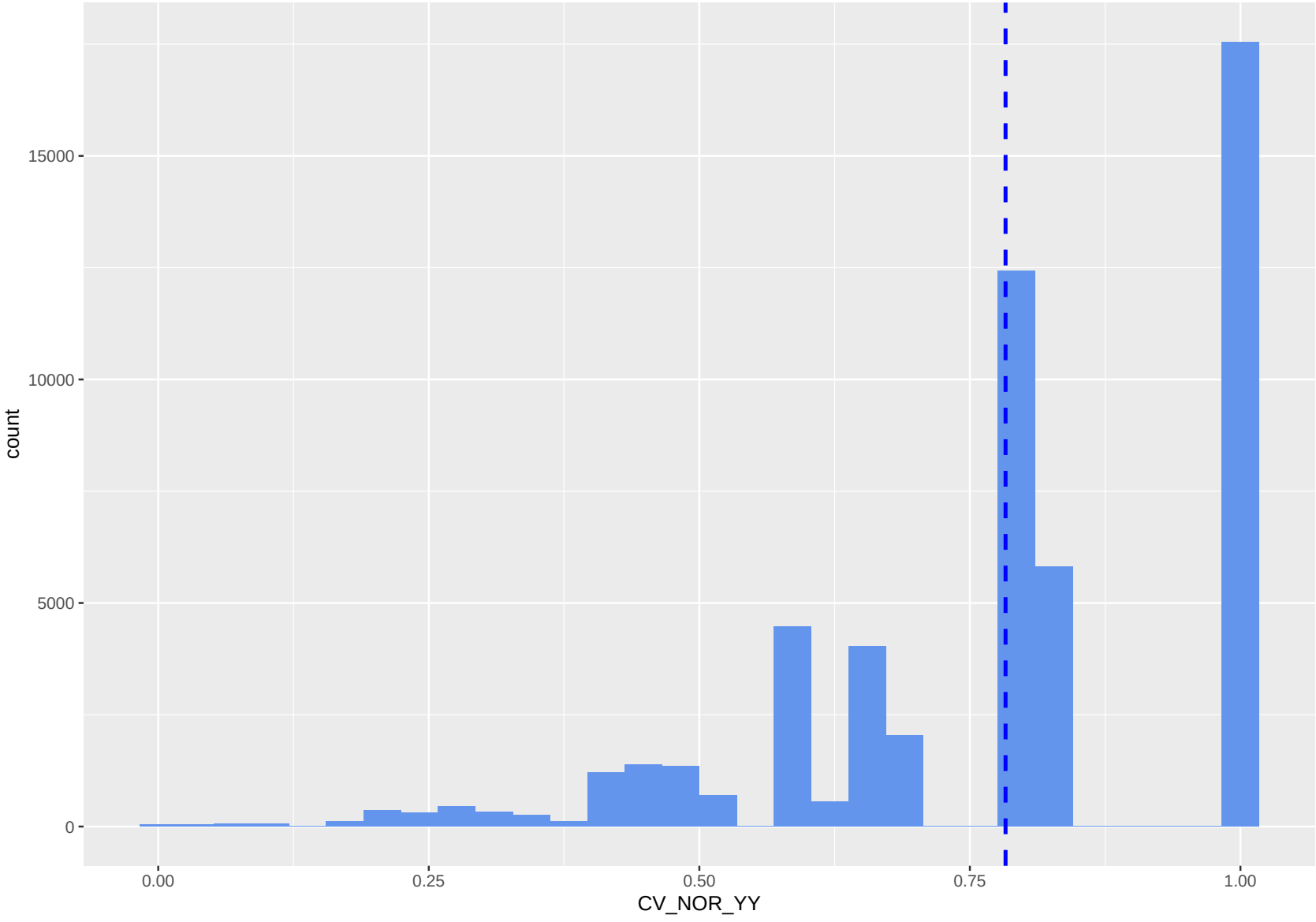
Admission test - Math score



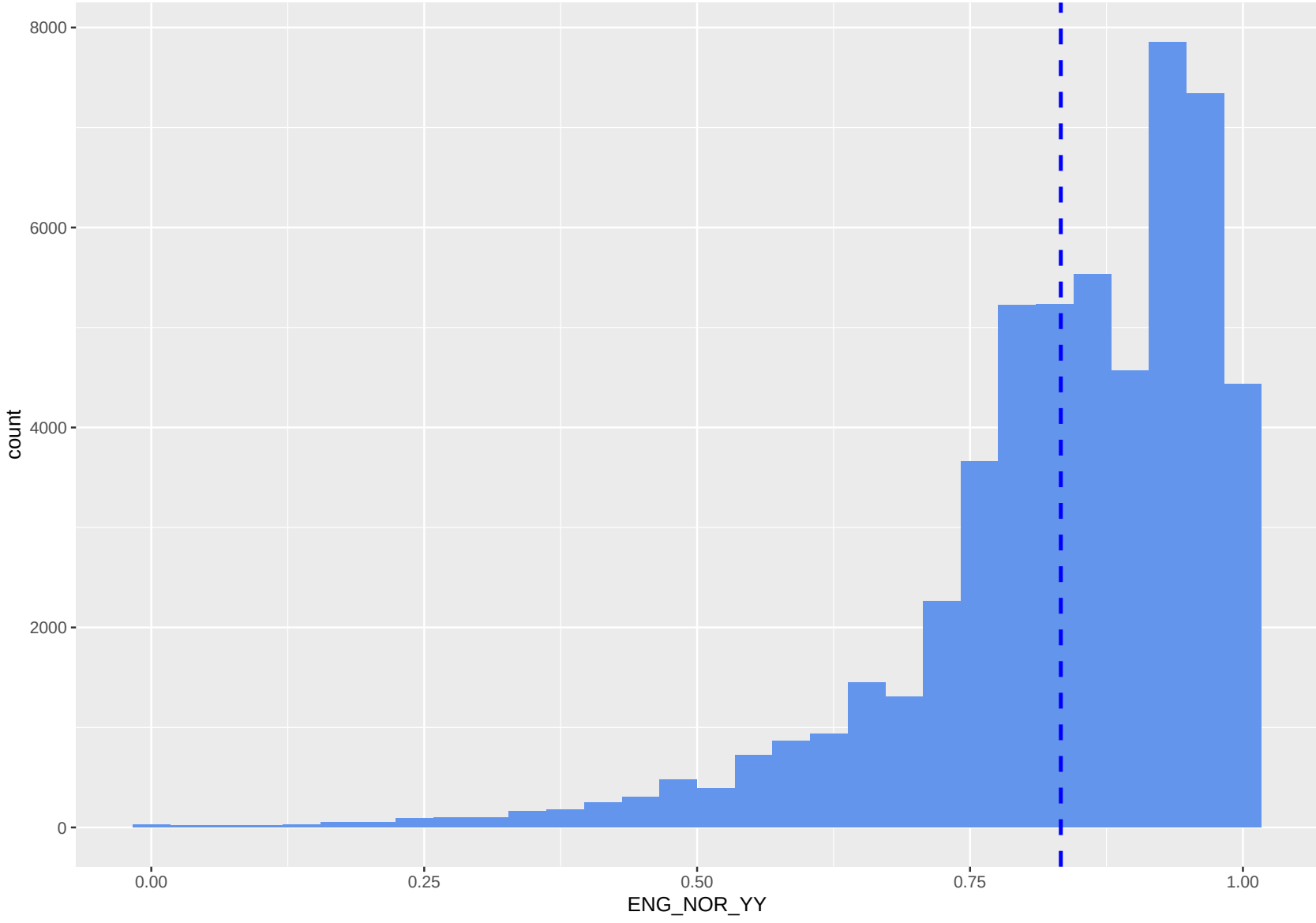
Admission test - Physics score



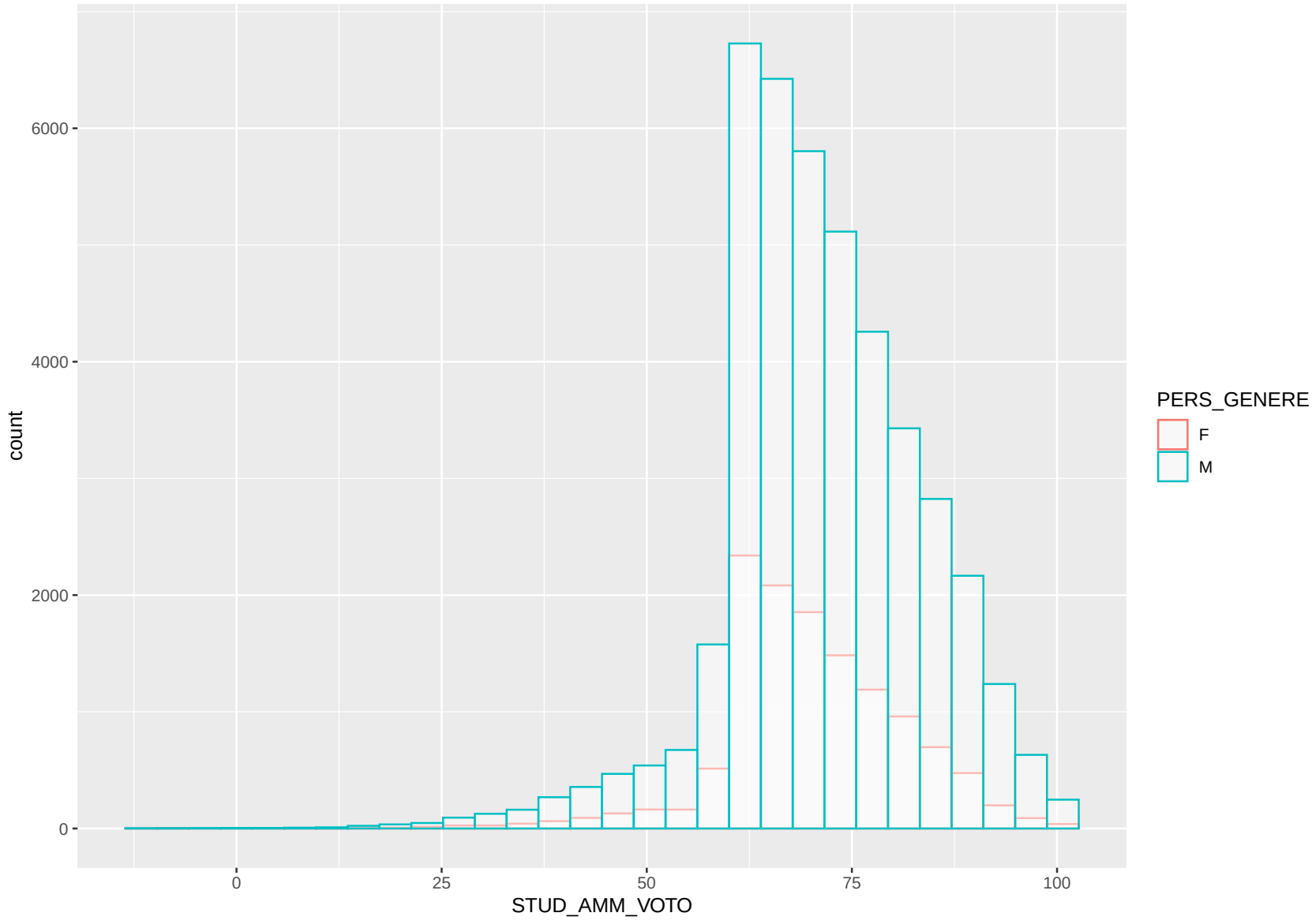
Admission test - Reading comprehension score



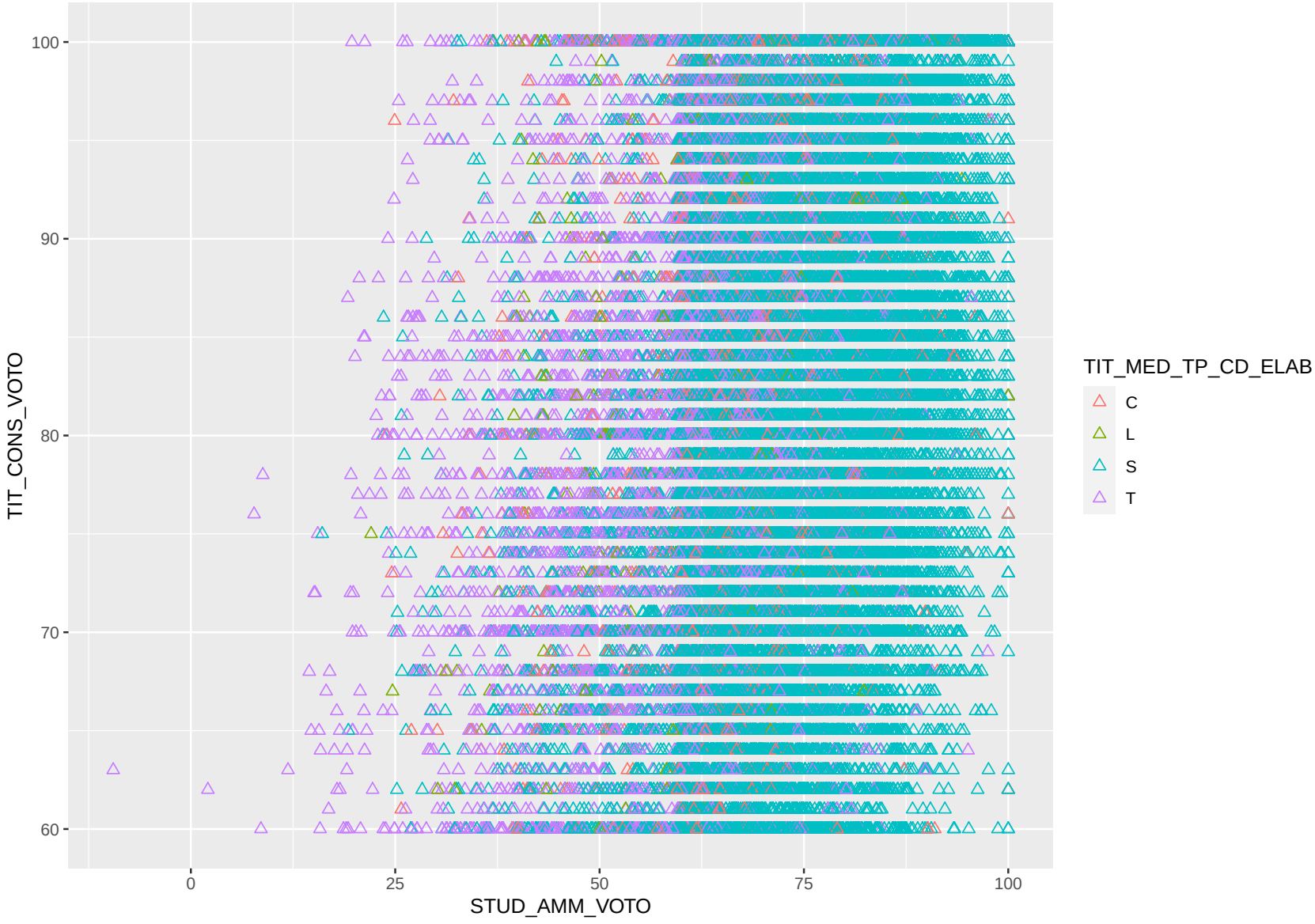
Admission test - English



Admission test by gender



Admission test and Previous studies



Exams

- The information are aggregate by year and only few features remain in the dataset:
  - number of passed exams (CFU)
  - number of failed exams (CFU)
  - avarage exams

Data Cleaning

Data cleaning - admission score

The admission test is multiple choice with zero average (min: 60, max: 100), we remove 7 students with a negative score

Reduce Columns

Only few column were indipendent and significant, so we reduce the dataset for a smaller model