



BI Homework 2023

Le Belle e la Bestia

Altieri Mariarosaria

Della Libera Davide

Longoni Letizia

Pre-processing - Part 1



Goal: understand the data deeply and obtain a more effective analysis

- ★ **Column formatting and editing** (exams dates, immatriculation age...)

- ★ **Removing didactic activities that do not provide a grade**

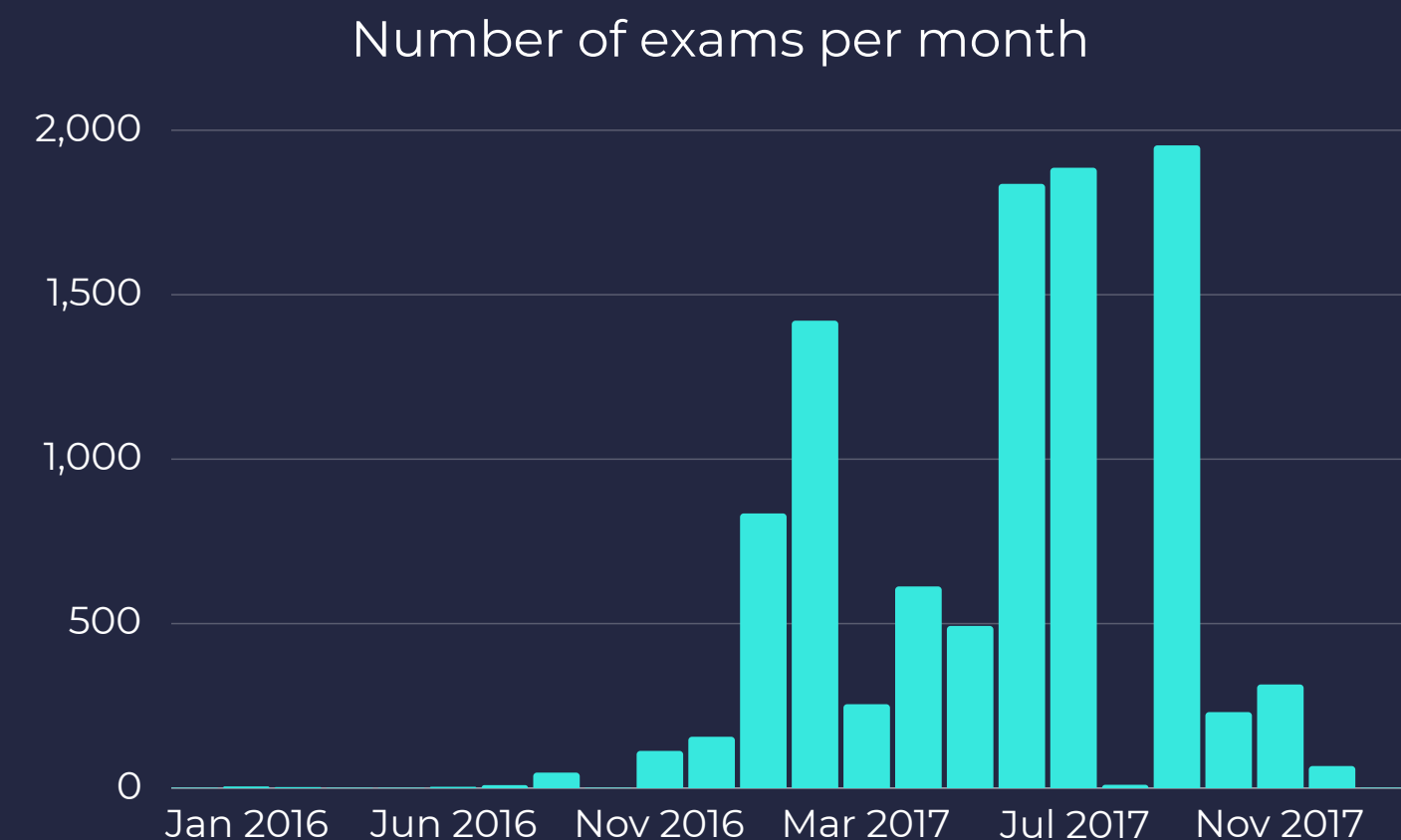
- *Situation:* exam passed but *NULL* grade
- *Examples:* seminars and internships

- ★ **Updating course of study codes**

CDSCOD	CDS Name	Type	Exams
524	Scienze dell'Educazione	L2	1
E1901R	Scienze dell'Educazione	L2	325

- ★ **Analyzed period**

From *November 2016* to *December 2017*



Pre-processing - Part 2



Open and Closed Exams

Pending registration

All the columns related to registration have value equal to 0 and *NULL* grade

An exam is said to be **open**

- ✳ if all enrollments of the same have the pending registration
- ✳ if there are no subsequent exam dates for that didactic activity

Otherwise is said to be **closed** and students with the pending registration are treated as absent



Exams Stats Table



- ✳ *Pre-computed statistics* of exams
- ✳ *Optimize the efficiency* of queries on the normalized database



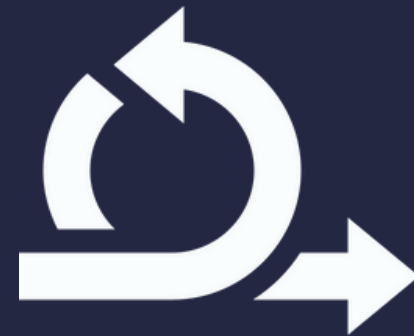
- ✳ Open-Closed exam
- ✳ Number of enrolled, passed, failed, insufficient, withdrawn and absent students
- ✳ Passing rate
- ✳ Average and median grade

Query 1 - Enrollments Distribution



Granularity

- Didactic activity
- Course of study
- Exams



Output Data

- Number of enrollments
- Number of exams
- Average number of enrollments per exam



If two or more exams have the same *course of study*, *didactic activity* and *date* but different teachers, they will be treated as a single exam instance

Query 2 - Hardest Didactic Activities

Difficulty of exam

1

Passing Rate

Ratio between number of passed and enrolled students

2

Aggregated Grade

Mean between average and median grades



Aggregated Grade Algorithm

Issue: median or average grades equal *NULL*

- ★ Passing Rate ≥ 0.5
Average and Median grades are numbers
- ★ Passing Rate = 0
Average and Median grades are NULL and considered as 0
- ★ $0 < \text{Passing Rate} < 0.5$
Median grade is NULL



A

Passing Rate range

[0.001, 0.499] with width equal to 0.498

B

Insufficient grade range

[1, 17] with width equal to 16

C

Ranges association

Width of grade subranges (WGS): $0.498 / 16$

$$\text{MedianGrade} = \frac{\text{PassingRate}}{\text{WGS}} + 1$$

Query 2 - Hardest Didactic Activities



Difficulty of Didactic Activity

1

Average Number of Trials
(students)

2

Median of Passing Rate
(exams)

3

Median of Aggregated Grade
(exams)



Final Ranking Algorithm

For each difficulty parameter within the Course of Study, compute the rank:



Median of Passing Rate

Median of Aggregated Grade



Average Number of Trials

Difficulty Index

$$DifficultyIndex = 0.4 * AvgTrials + 0.3 * MedianAggGrade + 0.3 * MedianPassRate$$



- As parameters of this last formula, we use the previously computed **ranks**
- For final ranking, we consider just Didactic Activities with two or more exams

Query 3 - Commitment Rate

For each **course of study**, the **commitment rate** is computed as:

$$CR = 0.7 * \frac{CommitmentDates * AvgCommitment}{DistinctDates} + 0.3 * \frac{NumAd}{NumExams}$$

where:

1

CommitmentDate

number of dates with overlap (2 or more exams on the same day, of different DA)

2

AvgCommitment

average number of different exams on dates with overlaps

3

DistinctDates

number of distinct dates on which exams are held

4

NumAd

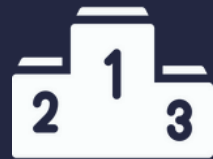
number of Didactic Activities of the course

5

NumExams

number of exams, considering all dates and all exams on those dates

Query 4 - Difficulty Score



Individuate the *most difficult and easiest didactic activities* for each course of study, in terms of **average grade**, assigning an **ad-hoc score**

1

Average Grade

Average grade taken by students who passed the exam

2

Passing Rate

Ratio between number of passed and enrolled students

$$\text{Score} = 0.7 * \text{AvgGrade} + 0.3 * \text{PassingRate}$$



- ✦ Before computing the Score, Average Grade has been properly normalized (**min-max**)
- ✦ Between two exams with the same score, the top-ranked will be the one with more exams



- ✦ Didactic Activities with at least three exams are considered
- ✦ On average, at least 2 enrollments for each exams per Didactic Activity

Query 5: Fast and Furious Index

For **each student** we considered two parameters:

Fast

- 1 **Period of activity:** computed as the number of months between the first and last exam taken

Furious

- 1 **Average grade:** average of grades of exams passed by each student
- 2 **Number of passed exams**
- 3 **Success ratio:** ratio between the number of passed exams and the number of enrollments

Query 5: Fast and Furious Index

Steps:

1

Parameters Normalization

To avoid parameters with a wider scale to have a greater impact on the final F&F index, each one is normalized in the range [0,1] using *min-max normalizaiton*

2

Parameters Weighting and final formula

For each student, the F&F index is computed as a linear combination of the parameters described before, properly weighted:

$$F\&F = 0.25 * (1 - Activity) + 0.25 * AvgGrade + 0.25 * ExamPassed + 0.25 * SuccessRatio$$



Why (1 - Activity)?

Because we want that the shortest period of activity gives the maximum contribution to F&F Index

Query 6: Trial and error rate

Step 1

Count the number of **failures** (Insufficiencies **I**, withdrawals **W** and absences **A**) for each **student** within each **DA**

Compute T&E for each **student** in each **DA**:

$$T\&E = 1.5I + 1W + 0.5A$$

Step 2

Compute the **average T&E** for each DA and rank them to identify the most problematic

Step 3

T&E index for each **Course**, computing the average T&E of all its DA, and rank them



Only didactic activities with more than 2 students are considered

Query 7: Gold Mortarboard

Goal

Identify the best students of the year for each Course of Study, deserving of a scholarship, obtaining the award **“Gold Mortarboard”**.

Fairness

The number of students selected per Course will be **proportional** to the number of "active students" in that specific one.

At least one student has to be selected for each Course of Study.

Normalized parameters

The following three parameters are based on the results of queries **6**, **2** and **5** respectively, properly *normalized* with respect to the belonging course.

- ★ **Average Trial&Error Index** (AT&E)
For all the DA that the student has passed
- ★ **Average Difficulty Index** (ADI)
For all the DA that the student has passed. In this case, since “lower the score, harder the DA”, we consider 1-DI
- ★ **Fast&Furious Index** (FFI)

Gold Mortarboard Score

Obtained as the weighted **sum** of the three parameters described before

$$Score = 0.5 * FFI + 0.25 * AT\&E + 0.25 * ADI$$

Consider only the students that passed at least **six exams**

Performance comparison - Pro vs Cons

Normalized

PROS

- ✓ Better for doing complex operations
- ✓ The usage of the exams stats table made it possible to optimize the operations performed
- ✓ The separation of the entities, primary and foreign keys makes the process more understandable (even if there are JOIN operations to perform)

CONS

- ✗ The creation of the views led to a decrease of performances

Denormalized

PROS








- ✓ Better for reading the data

CONS

- ✗ We don't have a primary key to identify entities
- ✗ Always scrolling through all lines
- ✗ Fragmentation of queries to get better code-readability; creation of sub queries and then joining them

Performance comparison - Time

The following times are obtained as a **average of 5 executions** of each query, to obtain more realistic and reliable measures:

	Query 1	Query 2	Query 3	Query 4	Query 5	Query 6	Query 7
Normalized	0,020 	0,430 	0,017 	0,025 	0,544 	0,358 	67,20
Denormalized	0,528	4,582	0,424	1,056	0,693	0,430	59,21 



Time of *Exams Stats Table* used only for the Normalized DB is **0,372**

Dashboard - Guide



Click here to see the dashboard

1

Here you can select the specific course to focus on

2

“**Students**” section: you can see general info about the active students enrolled to the course and their F&F score

3

“**Exam sessions**” section: you can see the number of exams enrollments per month, as well as for each exam date of a specific DA. It is also possible to see details about the CR of the course

4

“**DA**” section: info about the difficulty of each DA of the course and the T&R associated to each of them

5

“**Overview**” section: general information about all the courses, including the number of exams held each day, courses with the highest CR and the “Gold Mortarboard” assignation

COURSE

E2702Q - SCIENZE E TECNOLOGIE CHIMICHE

1

NAVIGATE



2



3



4

OVERVIEW



5

E2702Q - SCIENZE E TECNOLOGIE CHIMICHE

Type of course: L2 Duration: 3 years

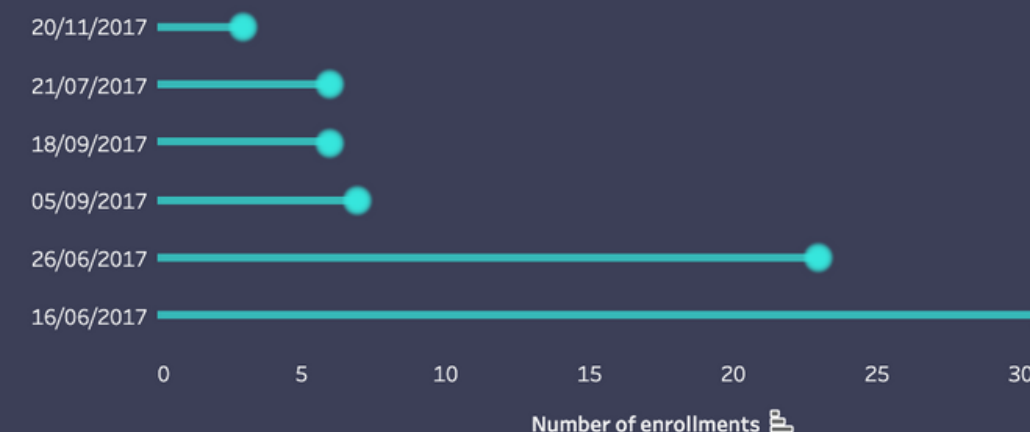
Number of students enrolled to exams



Number of students enrolled to exams of a didactic activity (DA):

Select the DA: E2702Q075 - MATEMATICA II

Exam date



Proportion of Courses with lower CR

Commitment rate:
1,009

% of days with at least one overlap:
48,10%

Average number of overlaps:
2,816

Average number of sessions for each DA:
4,933



Thank you!

for your attention

