وزارة الاتصالات وتقنية المعلومات

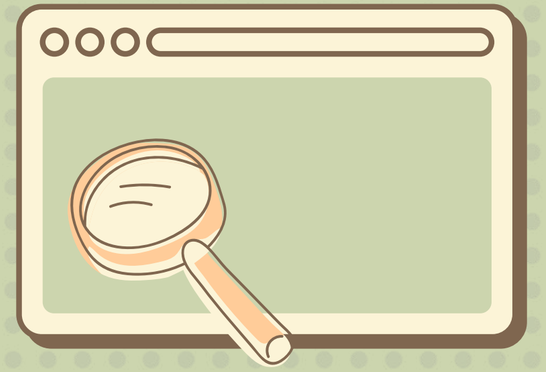
الأكاديمية السعودية الرقمية
SAUDI DIGITAL ACADEMY

CODING DOJO

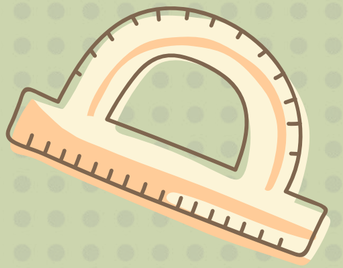
DESERT NINJAS

EDUCATIONAL LANDSCAPE IN THE MENA REGION

Capstone Two Project



- TEAM MEMBERS
- INTRODUCTION
- DATA PREPROCESSING
- EXPLORATORY DATA ANALYSIS
- DASHBOARD
- ML USING SPARK
- PIG
- FINDINGS



DESERT NINJAS

Reema Alaswad

Maha Alhazzani

Aljohara Alkanhal

Raghad Aleisa

Eman Aldosari

Reema Alaswad → Math

Maha Alhazzani → Physics

Aljohara Alkanhal → Physics

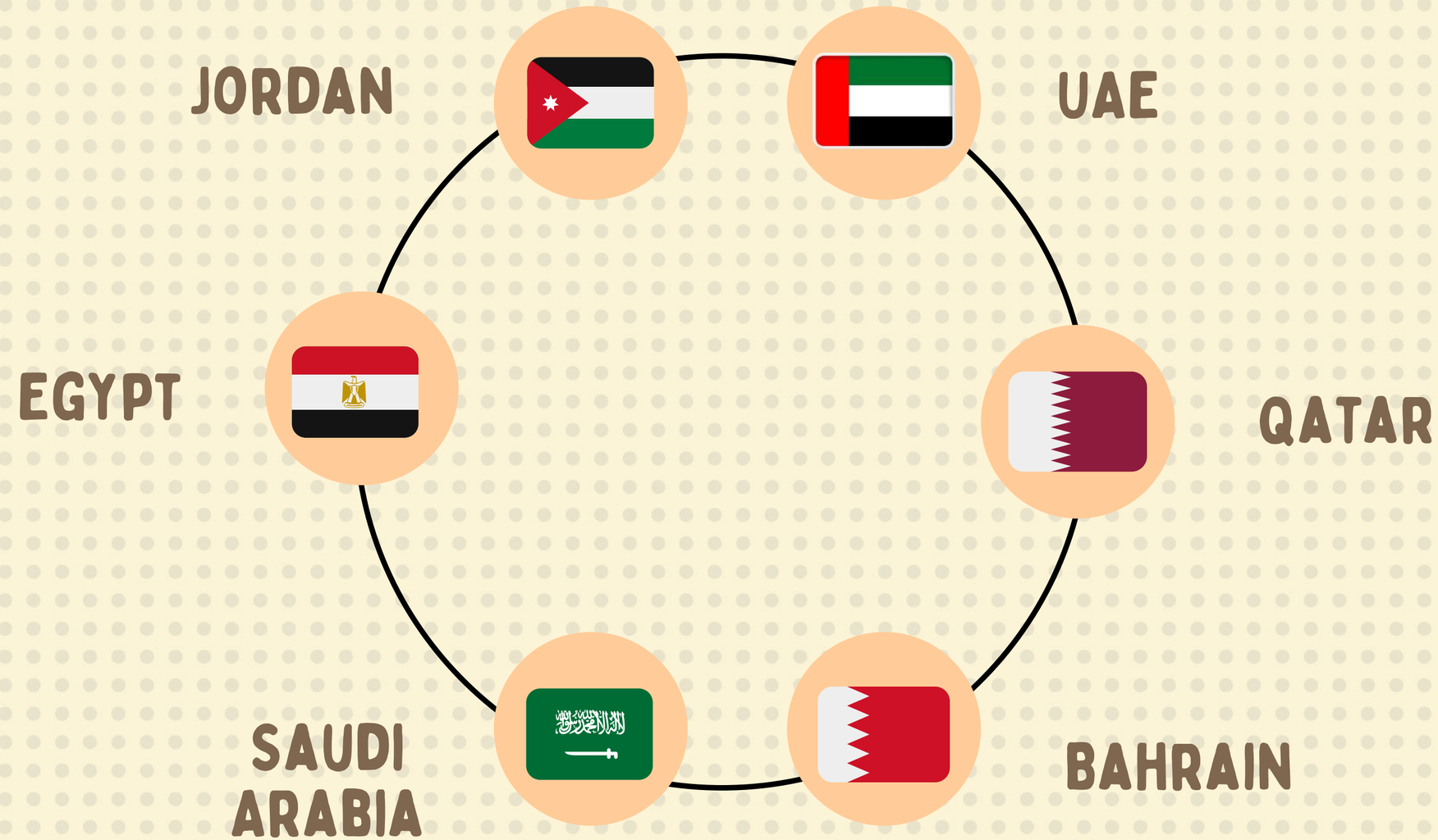
Raghad Aleisa → Physics

Eman Aldosari → Chemistry

What is your favorite subject?

INTRODUCTION

EDUCATION IN THE MENA REGION



EDUCATION IN THE MENA REGION

NUMBER OF STUDENTS

QUALITY OF EDUCATION

EDUCATION BUDGET

NUMBER OF STUDENTS IN MENA

1	2	3	4	5	6
EGYPT	SAUDI ARABIA	JORDAN	UAE	QATAR	BAHRAIN
24M	7M	2M	1M	390K	230K

BUDGET OF EDUCATION IN MENA

1	2	3	4	5	6
EGYPT	SAUDI ARABIA	QATAR	UAE	BAHRAIN	JORDAN
300B	135B	18B	10B	2B	600M

QUALITY OF EDUCATION IN MENA

1 QATAR

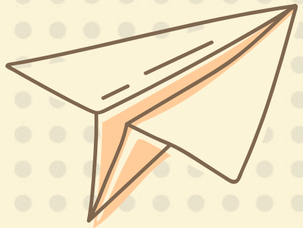
2 UAE

3 BAHRAIN

4 JORDAN

5 SAUDI ARABIA

6 EGYPT



SUSTAINABLE DEVELOPMENT GOALS

4 QUALITY EDUCATION

ENSURE INCLUSIVE AND EQUITABLE QUALITY EDUCATION AND PROMOTE LIFELONG LEARNING OPPORTUNITIES FOR ALL

5 GENDER EQUALITY

ACHIEVE GENDER EQUALITY AND EMPOWER ALL WOMEN AND GIRLS

SAUDI 2030 VISION TO ACHIEVE SUSTAINABLE DEVELOPMENT GOALS IN EDUCATION

A COMPREHENSIVE FRAMEWORK FOR THE PROFESSIONAL DEVELOPMENT OF TEACHERS AND EDUCATIONAL LEADERS.

SHIFTING TO DIGITAL EDUCATION TO SUPPORT TEACHER AND STUDENT PROGRESS.

ESTABLISH A PRACTICAL FRAMEWORK TO ALIGN UNIVERSITY GRADUATES WITH LABOR MARKET NEEDS.

SAUDI 2030 VISION TO ACHIEVE SUSTAINABLE DEVELOPMENT GOALS IN GENDER EQUALITY

WOMEN ASSUME MANY LEADERSHIP POSITIONS AND ARE ALLOCATED 20% OF SHURA COUNCIL SEATS.

LAUNCH OF SUPPORT PROGRAMS FOR WORKING WOMEN: WUSOOL AND QURRA.

LAUNCHING A TRAINING PROGRAM (DOROOB) WHICH AIMS TO PROVIDE PROFESSIONAL SUPPORT TO NEEDY GROUPS OF WOMEN

DATASET OVERVIEW

CASE STUDY: GRADE 12 HIGH SCHOOL STUDENTS PUBLIC RESULTS IN EGYPT

CASE STUDY: GRADE 12 HIGHSCHOOL STUDENTS PUBLIC RESULTS IN EGYPT

SCIENCE - HEALTH SCIENCES

SCIENCE - MATHEMATICS

LITERATURE

MAIN SUBJECTS:

ARABIC

FIRST FOREIGN LANGUAGE

SECOND FOREIGN LANGUAGE

RELIGION

NATIONAL EDUCATION

ECONOMICS STATISTICS

CASE STUDY: GRADE 12 HIGHSCHOOL STUDENTS PUBLIC RESULTS IN EGYPT

683K

45
COLUMNS

YEAR 2022
ONLY

OVERVIEW: GRADE 12 HIGHSCHOOL STUDENTS PUBLIC RESULTS IN EGYPT

683,287 STUDENTS

4334 SCHOOLS

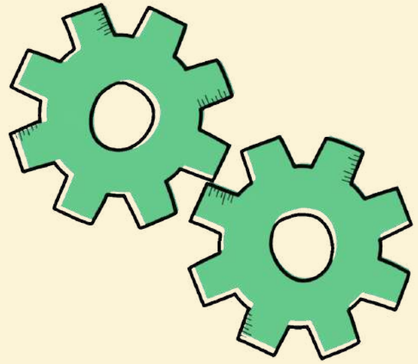
287 ADMINISTRATIONS

27 CITIES

3 BRANCHES

410 TOTAL DEGREES

DATA PREPROCESSING



DATASET BEFORE PREPROCESSING

desk_no	school_name	administration	city	branch	Percentage	status	arabic	first_foreign_lang
105410	الاورمان الرسمية لغات بنين	الدقى	الجيزة	أدبي	87.8	ناجح	61	27
105412	جمال عبد الناصرالرسمية لغات بنات	الدقى	الجيزة	علمي علوم	57.32	ناجح	47	25
105413	هضبة الاهرام ث التجريبية لغات بنين	الهرم	الجيزة	أدبي	83.41	ناجح	70	38
105415	التحرير الرسمية لغات بنين	أكتوبر	الجيزة	أدبي	53.17	ناجح	57	27
105416	التحرير الرسمية لغات بنين	أكتوبر	الجيزة	علمي رياضة	51.46	دور ثاني	56	25
105417	الحى العاشر الرسمية لغات بنات	أكتوبر	الجيزة	أدبي	93.66	ناجح	69	45
105418	الحى الخامس الرسمية المتميزة لغات بنين	أكتوبر	الجيزة	أدبي	79.02	ناجح	63	48
105419	الرؤية الثانوية الرسمية لغات بنين	أكتوبر	الجيزة	علمي علوم	65.37	ناجح	64	29
105420	الشيخ زايد الرسمية لغات بنات الحى الاول	الشيخ زايد	الجيزة	أدبي	71.46	ناجح	56	27
105421	القومية ث خ لغات بنات	العجوزة	الجيزة	أدبي	86.83	ناجح	67	40
105422	القومية ث خ لغات بنات	العجوزة	الجيزة	أدبي	88.05	ناجح	70	40
105430	نارمر ع خ لغات بنين	الدقى	الجيزة	أدبي	56.59	ناجح	40	27
105432	الحرية خ لغات بنين	الدقى	الجيزة	أدبي	78.29	ناجح	66	39
105433	الحرية خ لغات بنين	الدقى	الجيزة	أدبي	76.59	ناجح	65	41

683K RECORD

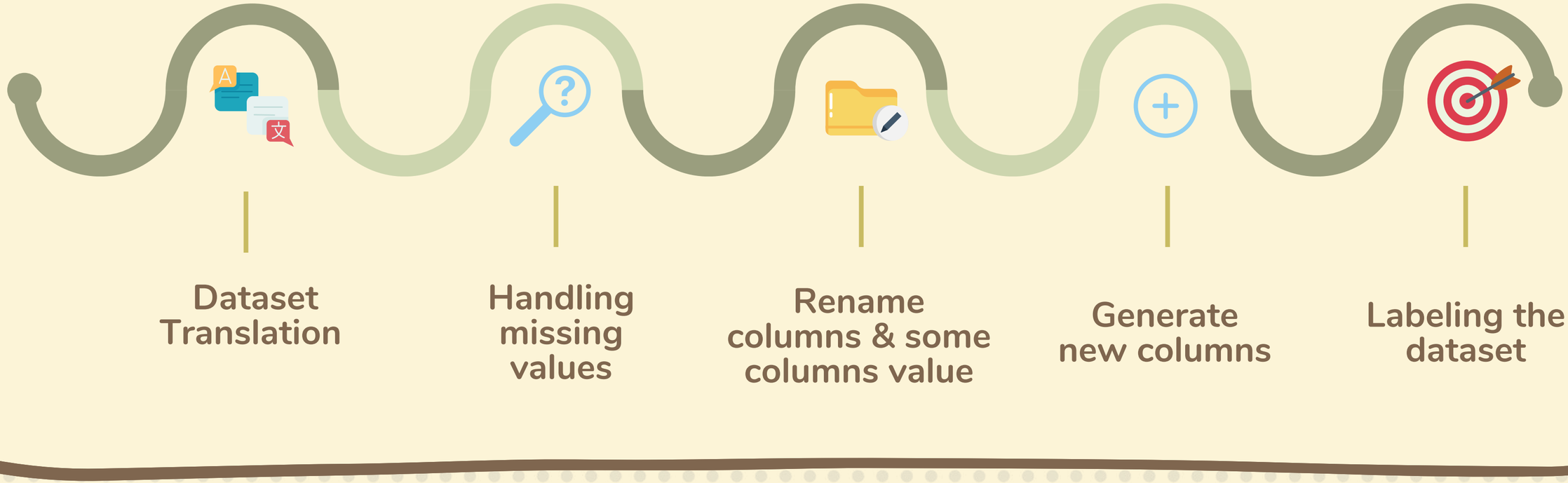
45 COLUMNS

8 CATEGORICAL COLS

37 NUMERIC COLS

1. THE DATASET WAS COMPLETELY IN ARABIC, EXCEPT FOR TWO COLUMNS.
2. THERE ARE **A LOT** OF MISSING VALUES.
3. MOST COLUMNS ARE NUMERIC → STUDENTS' GRADES.

DATA PREPROCESSING STEPS



1. DATASET TRANSLATION

1.1 Manual Translation

for branch, first and second attempt statuses

```
# branch column translation
df['branch'] = df['branch'].replace(['أدبي'], 'Literature')
df['branch'] = df['branch'].replace(['علمي علوم'], 'Science - Health sciences')
df['branch'] = df['branch'].replace(['علمي رياضة'], 'Science - Mathematics')
# first attempt status column translation
df['status'] = df['status'].replace(['ناجح'], 'Passed')
df['status'] = df['status'].replace(['راسب'], 'Failed')
df['status'] = df['status'].replace(['دور ثاني'], 'Second attempt')
# second attempt status column translation
df['status_2nd'] = df['status_2nd'].replace(['ناجح'], 'Passed')
df['status_2nd'] = df['status_2nd'].replace(['راسب'], 'Failed')
```

1. DATASET TRANSLATION

1.1 Manual Translation

for city column

```python
# city column translation
df['city'] = df['city'].replace(['القاهرة'], 'Cairo')
df['city'] = df['city'].replace(['الجيزة'], 'Giza')
df['city'] = df['city'].replace(['الشرقية'], 'Sharqia')
df['city'] = df['city'].replace(['الدقهلية'], 'Dakahlia')
df['city'] = df['city'].replace(['الاسكندرية'], 'Alexandria')
df['city'] = df['city'].replace(['القليوبية'], 'Qalyubiyya')
df['city'] = df['city'].replace(['المنيا'], 'Minya')
df['city'] = df['city'].replace(['الغربية'], 'Gharbia')
df['city'] = df['city'].replace(['المنوفية'], 'Monufia')
df['city'] = df['city'].replace(['البحيرة'], 'Beheira')
df['city'] = df['city'].replace(['اسيوط'], 'Assiut')
df['city'] = df['city'].replace(['كفر الشيخ'], 'Kafr El Sheikh')
df['city'] = df['city'].replace(['سوهاج'], 'Sohaj')
df['city'] = df['city'].replace(['بني سويف'], 'Beni Suef')
df['city'] = df['city'].replace(['الفيوم'], 'Faiyum')
df['city'] = df['city'].replace(['قنا'], 'Qena')
df['city'] = df['city'].replace(['دمياط'], 'Damietta')
df['city'] = df['city'].replace(['الإسماعيلية'], 'Ismailia')
df['city'] = df['city'].replace(['اسوان'], 'Aswan')
df['city'] = df['city'].replace(['الاقصر'], 'Luxor')
df['city'] = df['city'].replace(['بورسعيد'], 'Port Said')
df['city'] = df['city'].replace(['السويس'], 'Suez')
df['city'] = df['city'].replace(['شمال سيناء'], 'North Sinai')
df['city'] = df['city'].replace(['البحر الأحمر'], 'Red Sea')
df['city'] = df['city'].replace(['مطروح'], 'Matrouh')
df['city'] = df['city'].replace(['الوادي الجديد'], 'New Valley')
df['city'] = df['city'].replace(['جنوب سيناء'], 'South Sinai')
```

1. DATASET TRANSLATION

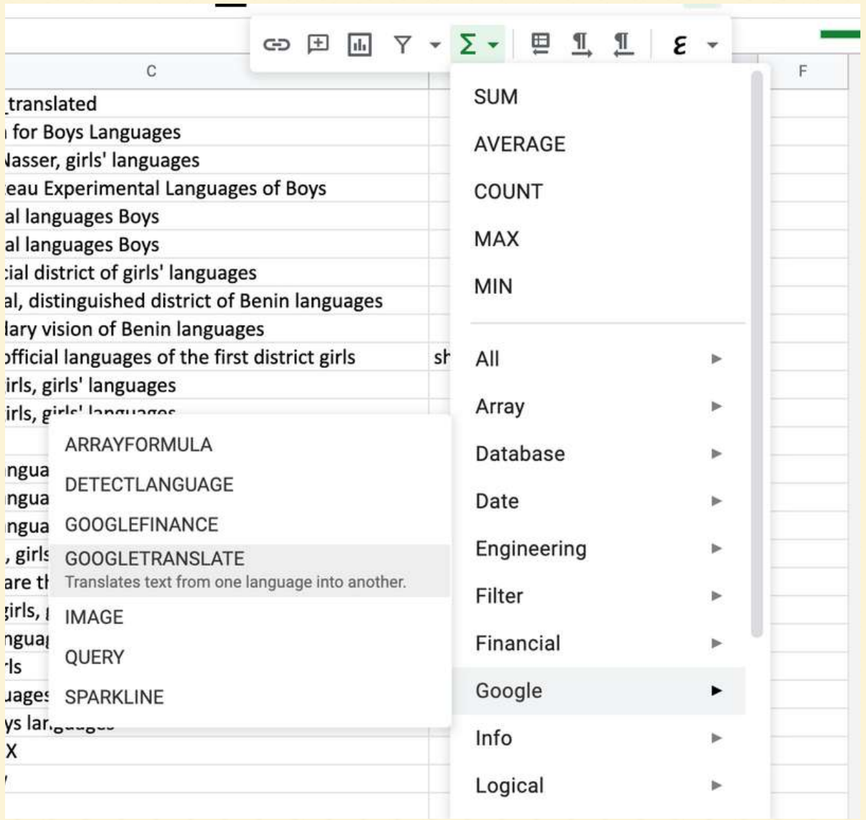
1.2 Automatic Translation using Google Translation Tool

- There is 4334 unique school names.
- There is 287 unique administration.
- From google sheet → Functions → GOOGLETRANSLATE Function
- Syntax:

column "ar"

GOOGLETRANSLATE(text, [source_language, target_language])

"en"



1. DATASET TRANSLATION

1.2 Automatic Translation using Google Translation Tool

SCHOOL_NAME	SCHOOL_NAME_TRANSLATED	ADMINISTRATION	ADMINISTRATION_TRANLATED
الاورمان الرسمية لغات بنين	OFFICIAL ORMAN FOR BOYS LANGUAGES	الدقى	DOKKI
هضبة الاهرام ث التجريبية لغات بنين	AL -AHRAM PLATEAU EXPERIMENTAL LANGUAGES OF BOYS	الهرم	PYRAMID
الحى العاشر الرسمية لغات بنات	THE TENTH OFFICIAL DISTRICT OF GIRLS' LANGUAGES	أكتوبر	OCTOBER
الحى الخامس الرسمية المتميزة لغات بنين	THE FIFTH OFFICIAL, DISTINGUISHED DISTRICT OF BENIN LANGUAGES	أكتوبر	OCTOBER
الرؤية الثانوية الرسمية لغات بنين	OFFICIAL SECONDARY VISION OF BENIN LANGUAGES	أكتوبر	OCTOBER
الشيخ زايد الرسمية لغات بنات الحى الاول	SHEIKH ZAYED OFFICIAL LANGUAGES OF THE FIRST DISTRICT GIRLS	الشيخ زايد	SHEIKH ZAYED
الحرية خ لغات بنات	FREEDOM, GIRLS, GIRLS	الدقى	DOKKI

WE NOTICE THAT THE AUTOMATIC TRANSLATION WAS A LITERAL TRANSLATION AND THE VALUES NEED TO BE CLEANED UP

1. DATASET TRANSLATION

1.3 Translation mistakes cleaning

```
# All school names are converted to lowercase letters to faster the cleaning process with fewer
possibilities.
df.school_name_translated=df.school_name_translated.str.lower()
# Convert the mistranslation of "benin" to boys with whitespace consideration.
df.school_name_translated=df.school_name_translated.str.replace(" benin ", " boys ", regex=True)
df.school_name_translated=df.school_name_translated.str.replace(" benin", " boys", regex=True)
# Convert the mistranslation of "languages" to international school with whitespace consideration.
df.school_name_translated=df.school_name_translated.str.replace(" languages ",
                                                                 " international school ", regex=True)
df.school_name_translated=df.school_name_translated.str.replace(" languages",
                                                                 " international school", regex=True)
df.school_name_translated=df.school_name_translated.str.replace(" girls, ",
                                                                 " international school ", regex=True)

# Remove useless words.
df.school_name_translated=df.school_name_translated.str.replace(" are ", " ", regex=True)
df.school_name_translated=df.school_name_translated.str.replace(" is ", " ", regex=True)
df.school_name_translated=df.school_name_translated.str.replace(" w ", " ", regex=True)
df.school_name_translated=df.school_name_translated.str.replace(" kh ", " ", regex=True)
df.school_name_translated=df.school_name_translated.str.replace(",", " ", regex=True)
```

1. DATASET TRANSLATION

1.3 Translation mistakes cleaning

- The result after cleaning:

SCHOOL_NAME	SCHOOL_NAME_TRANSLATED	ADMINISTRATION	ADMINISTRATION_TRANLATED
الاورمان الرسمية لغات بنين	OFFICIAL ORMAN FOR BOYS INTERNATIONAL SCHOOL	الدقى	DOKKI
هضبة الاهرام ث التجريبية لغات بنين	AL -AHRAM PLATEAU EXPERIMENTAL INTERNATIONAL SCHOOL OF BOYS	الهرم	PYRAMID
الحى العاشر الرسمية لغات بنات	THE TENTH OFFICIAL DISTRICT OF GIRLS' INTERNATIONAL SCHOOL	أكتوبر	OCTOBER
الحى الخامس الرسمية المتميزة لغات بنين	THE FIFTH OFFICIAL DISTINGUISHED DISTRICT OF BOYS INTERNATIONAL SCHOOL	أكتوبر	OCTOBER
الرؤية الثانوية الرسمية لغات بنين	OFFICIAL SECONDARY VISION OF BOYS INTERNATIONAL SCHOOL	أكتوبر	OCTOBER
الشيخ زايد الرسمية لغات بنات الحى الاول	SHEIKH ZAYED OFFICIAL INTERNATIONAL SCHOOL OF THE FIRST DISTRICT GIRLS	الشيخ زايد	SHEIKH ZAYED
الحرية خ لغات بنات	FREEDOM INTERNATIONAL SCHOOL GIRLS	الدقى	DOKKI

2. HANDLING MISSING VALUES

2.1 The number of missing values per column

- There is a large number of missing values for all students' marks, whether on the first or second attempt.
- The dataset is for all students from all three majors and for both attempts. Therefore, It is normal for values to be missing in these frequent cases:

1. The grades of literary subjects are missing values for students of the Science - health sciences major.
2. All marks for all subjects in the second attempt will be missing for students who only took the first attempt and succeeded.

- As for other columns that contain missing values. We can conclude that there are 939 students who did not attend the first attempt. And there are 529k students who did not attend the second attempt (mostly they passed the first time).

```
desk_no                      0
school_name                  0
school_name_translated       0
administration               0
administration_translated    0
city                         0
branch                       0
Percentage                 939
status                     939
arabic                    4731
first_foreign_lang        6048
second_foreign_lang       6629
pure_mathematics        586558
history                 425132
geography               424860
philosophy              424953
psychology              424990
chemistry               264934
biology                 360852
geology                 360511
applied_math            586482
physics                 265399
total                      939
religion                 72739
```

```
national_education          72547
economics_statistics        73158
gender                        787
Percentage_2nd              529454
status_2nd                  529454
arabic_2nd                  529726
first_foreign_lang_2nd      529806
second_foreign_lang_2nd     530239
pure_mathematics_2nd        666236
history_2nd                 610500
geography_2nd               610491
philosophy_2nd              610446
psychology_2nd              610451
chemistry_2nd               602777
biology_2nd                 619702
geology_2nd                 619659
applied_math_2nd            666180
physics_2nd                 602857
total_2nd                   529454
religion_2nd                562349
national_education_2nd      562321
economics_statistics_2nd    562656
_merge                           0
dtype: int64
```

2. HANDLING MISSING VALUES

2.2 Filling the missing values of all students' marks

- How to fill in the missing values in the numeric columns (the students' grades) ?

1. If the student did not attend all the first attempt exams → the missing values will be filled in with -1
2. If the student did not attend all the second attempt exams → the missing values will be filled in with -2
3. Whether on the first, second, or both attempts; If the subjects are not related to the student's major → the missing values will be filled in with -3
4. Whether on the first, second, or both attempts; If a student missed one of the 10 required courses → the missing values will be filled in with -4

2. HANDLING MISSING VALUES

2.2 Filling the missing values of all students' marks

- How to fill in the missing values in the numeric columns (the students' grades) ?

1. If the student did not attend all the first attempt exams → the missing values will be filled in with -1

```
# Mark the students who did not enter the first attempt exams and whose total and
percentage were missing with the number "-1"
df.loc[(df['Percentage'].isnull()) & (df['total'].isnull()), ['Percentage', 'total']] = -1
df.loc[(df['Percentage']==-1) & (df['total']==-1),
['arabic','first_foreign_lang','second_foreign_lang','pure_mathematics','history','geograp
hy','philosophy','psychology','chemistry','biology','geology','applied_math','physics','re
ligion','national_education','economics_statistics']] = -1
```

2. HANDLING MISSING VALUES

2.2 Filling the missing values of all students' marks

- How to fill in the missing values in the numeric columns (the students' grades) ?

2. If the student did not attend all the second attempt exams → the missing values will be filled in with -2

```
#Mark the students who did not enter the second attempt exams and whose total and
percentage were missing with the number "-2"
df.loc[(df['Percentage_2nd'].isnull()) & (df['total_2nd'].isnull()), ['Percentage_2nd',
'total_2nd']] = -2
df.loc[(df['Percentage_2nd']==-2) & (df['total_2nd']==-2), ['arabic_2nd',
'first_foreign_lang_2nd','second_foreign_lang_2nd','pure_mathematics_2nd','history_2nd','g
eography_2nd','philosophy_2nd','psychology_2nd','chemistry_2nd','biology_2nd','geology_2nd
','applied_math_2nd','physics_2nd','religion_2nd','national_education_2nd','economics_stat
istics_2nd']] = -2
```

2. HANDLING MISSING VALUES

2.2 Filling the missing values of all students' marks

- How to fill in the missing values in the numeric columns (the students' grades) ?

3. Whether on the first, second, or both attempts; If the subjects are not related to the student's major → the missing values will be filled in with -3

MAJOR	THE SUBJECTS ARE NOT RELATED TO THE MAJOR
LITERATURE	CHEMISTRY, PHYSICS, BIOLOGY, GEOLOGY, PURE_MATHEMATICS, APPLIED_MATH
SCIENCE - HEALTH SCIENCES	HISTORY, PHILOSOPHY, PSYCHOLOGY, GEOGRAPHY, PURE_MATHEMATICS, APPLIED_MATH
SCIENCE - MATHEMATICS	HISTORY, PHILOSOPHY, PSYCHOLOGY, GEOGRAPHY, BIOLOGY, GEOLOGY

2. HANDLING MISSING VALUES

2.2 Filling the missing values of all students' marks

3. Whether on the first, second, or both attempts; If the subjects are not related to the student's major → the missing values will be filled in with -3

- An example of literature major students for each first or second attempt or both.

```
#For students majoring Literature and who entered the first attempt only, we will mark the specialized subject from other majors with "-3"
df.loc[(df['branch']=='Literature') & (df['_merge']=='left_only') & (df['Percentage']≠-1) & (df['total']≠-1), ['chemistry', 'physics','biology','geology','pure_mathematics', 'applied_math']] = -3
#For students majoring Literature and who entered the second attempt only, we will mark the specialized subjects from other majors with "-3"
df.loc[(df['branch']=='Literature') & (df['_merge']=='right_only') & (df['Percentage_2nd']≠-2) & (df['total_2nd']≠-2),['chemistry_2nd','physics_2nd', 'biology_2nd','geology_2nd','pure_mathematics_2nd','applied_math_2nd']] = -3
#For students majoring Literature and who entered the both attempt, we will mark the specialized subjects from other majors with "-3"
df.loc[(df['branch']=='Literature') & (df['_merge']=='both') & (df['Percentage']≠-1) & (df['total']≠-1) & (df['Percentage_2nd']≠-2) & (df['total_2nd']≠-2), ['chemistry', 'physics','biology','geology','pure_mathematics','applied_math','chemistry_2nd', 'physics_2nd','biology_2nd','geology_2nd','pure_mathematics_2nd','applied_math_2nd']] = -3
```

2. HANDLING MISSING VALUES

2.2 Filling the missing values of all students' marks

4. Whether on the first, second, or both attempts; If a student missed one of the 10 required courses → the missing values will be filled in with -4

- An example of looping over all majors students and their subjects to mark the absentees in first attempt exams with -4.

```
# On the first attempt and for all majors, if the student was absent in one of the ten compulsory subjects exams (which vary depending on the major), we will fill in the subjects in which he/she was absent with -4
for index, row in df.iterrows():
    if row['_merge'] == "left_only":
        if row['branch'] == "Literature":
            for col in ['arabic', 'first_foreign_lang', 'second_foreign_lang', 'history', 'geography', 'philosophy', 'psychology', 'religion', 'national_education', 'economics_statistics']:
                if pd.isnull(row[col]):
                    df.at[index, col] = -4
        if row['branch'] == "Science - Health sciences":
            for col in ['arabic', 'first_foreign_lang', 'second_foreign_lang', 'chemistry', 'geology', 'physics', 'biology', 'religion', 'national_education', 'economics_statistics']:
                if pd.isnull(row[col]):
                    df.at[index, col] = -4
        if row['branch'] == "Science - Mathematics":
            for col in ['arabic', 'first_foreign_lang', 'second_foreign_lang', 'chemistry', 'physics', 'applied_math', 'pure_mathematics', 'religion', 'national_education', 'economics_statistics']:
                if pd.isnull(row[col]):
                    df.at[index, col] = -4
```

2. HANDLING MISSING VALUES

2.2 Filling the missing values of the statuses

- **The status of the student in the first attempt:**

Since the student did not pass, fail, and did not deserve a second attempt.

Therefore, they **didn't attend** the first attempt exams.

- **The status of the student in the second attempt:**

Since the student did not pass or fail on the second attempt, then they did not attend the exams for the second attempt ➜ they **actually passed** from the first attempt.

```
Passed                  450945
Second attempt          154940
Failed                   76463
Did not attend             939
Name: status,     dtype: int64
```

```
Passed from the first attempt    529454
Passed                           136385
Failed                            17448
Name: status_2nd, dtype: int64
```

```
df['status'] =  df['status'].fillna('Did not attend')
df['status_2nd'] =  df['status_2nd'].fillna('Passed from the first attempt')
```

2. HANDLING MISSING VALUES

2.2 Filling the missing values of the gender

- Now, there are only missing values in the Gender column, and they are very few.
- We will fill them based on the name school.
- If the school name contains "girls" the gender will be Female.
- If the school name contains "boys" the gender will be Male.
- Otherwise, we will drop them (They belong to a mixed school).

```
desk_no                    0
school_name                0
school_name_translated     0
administration             0
administration_translated  0
city                       0
branch                     0
Percentage                 0
status                     0
arabic                     0
first_foreign_lang         0
second_foreign_lang        0
pure_mathematics           0
history                    0
geography                  0
philosophy                 0
psychology                 0
chemistry                  0
biology                    0
geology                    0
applied_math               0
physics                    0
total                      0
religion                   0
```

```
national_education         0
economics_statistics       0
gender                   787
Percentage_2nd              0
status_2nd                  0
arabic_2nd                  0
first_foreign_lang_2nd      0
second_foreign_lang_2nd     0
pure_mathematics_2nd        0
history_2nd                 0
geography_2nd               0
philosophy_2nd              0
psychology_2nd              0
chemistry_2nd               0
biology_2nd                 0
geology_2nd                 0
applied_math_2nd            0
physics_2nd                 0
total_2nd                   0
religion_2nd                0
national_education_2nd      0
economics_statistics_2nd    0
_merge                      0
dtype: int64
```

2. HANDLING MISSING VALUES

2.2 Filling the missing values of the gender

```python
# This function will fill the missing values in the gender based on the school name,
# if the school name contains the word "girls" then the gender will be female,
# and if it contains the word "boys" then the gender will be male.
# Otherwise, the school will be mixed school, and it will be difficult to determine the
gender, so we will drop them later.
def gender_fillna(school_name, gender):
    if gender == 'null':
        if "بنات" in school_name:
            gender = "Female"
        if "للبنات" in school_name:
            gender = "Female"
        if "بنين" in school_name:
            gender = "Male"
        if "للبنين" in school_name:
            gender = "Male"
        return gender
    return gender
```

3. RENAME COLUMNS & SOME COLUMNS' VALUE

3.1 Column names & values manipulation

1

```
df['_merge'] = df['_merge'].replace(['left_only'], 'First attempt only')
df['_merge'] = df['_merge'].replace(['right_only'], 'Second attempt only')
df['_merge'] = df['_merge'].replace(['both'], 'Both attempts')
```

2

```
df['gender'] = df['gender'].replace(['F'], 'Female')
df['gender'] = df['gender'].replace(['M'], 'Male')
```

3. RENAME COLUMNS & SOME COLUMNS' VALUE

3.1 Column names & values manipulation

3

```
#Rename some of the columns to more meaningful names.
df= df.rename(columns =
{'_merge':'no_of_attempts','Percentage':'percentage',
'Percentage_2nd':'percentage_2nd'}, inplace = True)
```

4. GENERATE NEW COLUMN

4.1 Generate school type column

- The most common words in the school name

```
In [53]: from collections import Counter
         Counter(" ".join(df["school_name"]).split()).most_common(200)
```

```
Out[53]: ,(340806 ,'الثانوية')]
         ,(188928 ,'بنات')
         ,(142521 ,'ث')
         ,(137049 ,'بنين')
         ,(100621 ,'المشتركة')
         ,(77636 ,'الشهيد')
         ,(65857 ,'الثانويه')
         ,(48433 ,'منازل')
         ,(45608 ,'محمد')
         ,(41767 ,'لغات')
         ,(41114 ,'خدمات')
         ,(39363 ,'م')
         ,(38199 ,'إدارة')
         ,(35873 ,'احمد')
         ,(35702 ,'عبد')
         ,(29228 ,'العسكرية')
         ,(26097 ,'الرسمية')
         ,(21922 ,'.')
         ,(21864 ,'الخاصة')
```

4. GENERATE NEW COLUMN

4.1 Generate school type column

PRIVATE SCHOOL

PUBLIC SCHOOL

```
# Private school type
df.loc[df['school_name'].str.contains("الخاصة"), 'school_type'] = "Private School"
df.loc[df['school_name'].str.contains("خاصة"), 'school_type'] = "Private School"
df.loc[df['school_name'].str.contains("الخاصه"), 'school_type'] = "Private School"
df.loc[df['school_name'].str.contains("خاصه"), 'school_type'] = "Private School"
```

```
# Public school type
df.loc[df['school_name'].str.contains("العامة"), 'school_type'] = "Public School"
df.loc[df['school_name'].str.contains("العامه"), 'school_type'] = "Public School"
df.loc[df['school_name'].str.contains("عامة"), 'school_type'] = "Public School"
df.loc[df['school_name'].str.contains("عامه"), 'school_type'] = "Public School"
```

4. GENERATE NEW COLUMN

4.1 Generate school type column

INTERNATIONAL SCHOOL

SCHOOL FOR THE BLINS

```python
# International school type
df.loc[df['school_name'].str.contains("لغات"), 'school_type'] = "International School"
df.loc[df['school_name'].str.contains("للغات"), 'school_type'] = "International School"
df.loc[df['school_name'].str.contains("اللغات"), 'school_type'] = "International School"
```

```python
# School for the blind type
df.loc[df['school_name'].str.contains("للمكفوفين"), 'school_type'] = "School for the Blind"
df.loc[df['school_name'].str.contains("مكفوفين"), 'school_type'] = "School for the Blind"
df.loc[df['school_name'].str.contains("الكفيفات"), 'school_type'] = "School for the Blind"
df.loc[df['school_name'].str.contains("وضعاف البصر"), 'school_type'] = "School for the Blind"
df.loc[df['school_name'].str.contains("ضعاف البصر"), 'school_type'] = "School for the Blind"
df.loc[df['school_name'].str.contains("كفيفات"), 'school_type'] = "School for the Blind"
df.loc[df['school_name'].str.contains("للكفيفات"), 'school_type'] = "School for the Blind"
```

4. GENERATE NEW COLUMN

4.1 Generate school type column

- The final result

```
Public School                 607811
International School           54211
Private School                 20268
School for the Blind             210
Name: school_type, dtype: int64
```

4. GENERATE NEW COLUMN

4.2 Generate mixed school column

```
# Schools for girls only
df.loc[df['school_name'].str.contains("بنات"), 'mixed_school'] = "Girls only"
df.loc[df['school_name'].str.contains("للبنات"), 'mixed_school'] = "Girls only"
# Schools for boys only
df.loc[df['school_name'].str.contains("بنين"), 'mixed_school'] = "Boys only"
df.loc[df['school_name'].str.contains("للبنين"), 'mixed_school'] = "Boys only"
# Schools for girls and boys
df.loc[df['school_name'].str.contains("المشتركة"), 'mixed_school'] = "Mixed"
df.loc[df['school_name'].str.contains("المشتركه"), 'mixed_school'] = "Mixed"
```

4. GENERATE NEW COLUMN

4.2 Generate mixed school column

- There are 206K missing values, this function will fill them based on gender.

```
def get_mixed_school(gender):
    if gender == 'Female':
        mixed_school='Girls only'
    if gender == 'Male':
        mixed_school='Boys only'
    return mixed_school
```

4. GENERATE NEW COLUMN

4.2 Generate mixed school column

- The final result

```
Girls only                 307703
Boys only                  252067
Mixed                      122730
Name: mixed_school, dtype: int64
```

4. GENERATE NEW COLUMN

4.3 Generate homeschooling column

- According to the Egyptian Ministry of Education, the students who repeated the third year of high school twice, or the students whose days of absence are more than their attendance, are transferred to homeschooling school.

```
# If the school name contains homeschooling then we'll mark them as "Yes"
df.loc[df['school_name'].str.contains("منازل"), 'homeschooling'] = "Yes"
# Else, they'll be marked as "No"
df['homeschooling'].fillna("No", inplace = True)
```

4. GENERATE NEW COLUMN

4.3 Generate homeschooling column

- The final result

```
No                  634067
Yes                  48433
Name: homeschooling, dtype: int64
```

4. GENERATE NEW COLUMN

4.4 Number of failed courses for each student

- There are 10 subjects that all students from all majors must pass in order to graduate. Their total and percentage will be calculated based on their performance in the first seven subjects, but they need to pass the remaining three general subjects regardless of their final grade.
- A student is considered to have failed the course when: ⚠

1. Gets a score less than half of the full score.
2. Was absent on the day of the exam.

- For example:
- Arabic subject is out of 80. when the student gets 39 out of 80 → Failed in the subject.

SUBJECTS IN EACH MAJOR

HEALTH SCIENCES

SUBJECTS:

- ARABIC 80
- FIRST FOREIGN LANGUAGE (ENGLISH) 50
- SECOND FOREIGN LANGUAGE 40
- CHEMISTRY 60
- BIOLOGY 60
- GEOLOGY 60
- PHYSICS 60

- RELIGION 25
- NATIONAL EDUCATION 25
- ECONOMICS STATISTICS 50

LITERATURE

SUBJECTS:

- ARABIC 80
- FIRST FOREIGN LANGUAGE (ENGLISH) 50
- SECOND FOREIGN LANGUAGE 40
- HISTORY 60
- GEOGRAPHY 60
- PHILOSOPHY 60
- PSYCHOLOGY 60

- RELIGION 25
- NATIONAL EDUCATION 25
- ECONOMICS STATISTICS 50

MATHEMATICS

SUBJECTS:

- ARABIC 80
- FIRST FOREIGN LANGUAGE (ENGLISH) 50
- SECOND FOREIGN LANGUAGE 40
- PURE MATHEMATICS 60
- CHEMISTRY 60
- APPLIED MATH 60
- PHYSICS 60

- RELIGION 25
- NATIONAL EDUCATION 25
- ECONOMICS STATISTICS 50

SUBJECTS IN EACH MAJOR

- Three functions will calculate the number of subjects in which the student failed based on each major. All subjects will be included starting from the specialized subjects for each major, basic language subjects. (Arabic, foreign 1st language, and foreign 2nd language), as well as the three general subjects.

Literature 📚	Science "Health Sciences" 🫀	Science "Mathematics" 🔢	Full Mark ✅	Failure Mark ⚠️
Arabic	Arabic	Arabic	80	39 and less
First foreign language	First foreign language	First foreign language	50	24 and less
Second foreign language	Second foreign language	Second foreign language	40	19 and less
History	Chemistry	Chemistry	60	29 and less
Philosophy	Physics	Physics	60	29 and less
Psychology	Biology	Applied Mathematics	60	29 and less
Geography	Geology	Pure Mathematics	60	29 and less
Religion	Religion	Religion	25	12.4 and less
National education	National education	National education	25	12.4 and less
Economics statistics	Economics statistics	Economics statistics	50	24 and less

4. GENERATE NEW COLUMN

4.4 Number of failed courses for each student

- The final result

```
0.0     429971
1.0      82261
3.0      43914
2.0      39561
4.0      35841
5.0      23137
6.0      15242
7.0       7710
8.0       2991
9.0       1097
10.0       775
Name: no_of_failed_courses, dtype: int64
```

4. GENERATE NEW COLUMN

4.5 Generate the final grade column

GRADING SYSTEM IN EGYPT

PERCENTAGE	QUALIFICATION
85–100	EXCELLENT
75–84	VERY GOOD
65–74	GOOD
50–64	ACCEPTABLE OR PASS
30–49	WEAK
0–29	VERY WEAK

4. GENERATE NEW COLUMN

4.5 Generate the final grade column

- The final result

```
Acceptable          291129
Good                151728
Very good            99376
Weak                 77396
Excellent            55786
Very weak             7085
Name: final_grade, dtype: int64
```

5. LABEL THE DATASET

5.1 Generate can join university column

- We label the dataset based on the lowest university entry total by major, We can summarize these totals in this table:

MAJOR	THE LOWEST TOTAL
SCIENCE - HEALTH SCIENCES	369 OUT OF 410
SCIENCE - MATHEMATICS	342 OUT OF 410
LITERATURE	265 OUT OF 410

5. LABEL THE DATASET

```
def get_can_join_major_university(no_of_attempts, branch, total, total_2nd):
    can_join_major_university="No"
    if no_of_attempts == "First attempt only":
        if branch == "Literature":
            if total >= 265:
                can_join_major_university="Yes"
        if branch == "Science - Health sciences":
            if total >= 369:
                can_join_major_university="Yes"
        if branch == "Science - Mathematics":
            if total >= 342:
                can_join_major_university="Yes"
    if no_of_attempts == "Second attempt only" or no_of_attempts == "Both attempts":
        if branch == "Literature":
            if total_2nd >= 265:
                can_join_major_university="Yes"
        if branch == "Science - Health sciences":
            if total_2nd >= 369:
                can_join_major_university="Yes"
        if branch == "Science - Mathematics":
            if total_2nd >= 342:
                can_join_major_university="Yes"
    return can_join_major_university
```

5. LABEL THE DATASET

5.1 Generate can join university column

- The final result

```
No     565749
Yes    116751
Name: can_join_uni, dtype: int64
```

DATASET AFTER PREPROCESSING

desk_no	gender	branch	city	dministratic	chool_nam	school_type	omeschooli	ixed_schoo	_of_attem	f_failed_co	an_join_un	final_grad
105410	Male	Literature	Giza	Dokki	official orm	Internatior	No	Boys only	First attem	0	Yes	Excellent
105412	Female	Science - H	Giza	Dokki	gamal abde	Internatior	No	Girls only	First attem	0	No	Acceptab
105413	Male	Literature	Giza	Pyramid	al -ahram p	Internatior	No	Boys only	First attem	1	Yes	Very gooc
105415	Male	Literature	Giza	October	official edit	Internatior	No	Boys only	First attem	1	No	Acceptab
105416	Male	Science - M	Giza	October	official edit	Internatior	No	Boys only	Both attem	1	No	Acceptab
105417	Female	Literature	Giza	October	the tenth o	Internatior	No	Girls only	First attem	0	Yes	Excellent
105418	Male	Literature	Giza	October	the fifth off	Internatior	No	Boys only	First attem	1	Yes	Very gooc
105419	Male	Science - H	Giza	October	official secc	Internatior	No	Boys only	First attem	0	No	Good
105420	Female	Literature	Giza	sheikh Zaye	sheikh zaye	Internatior	No	Girls only	First attem	1	Yes	Good
105421	Female	Literature	Giza	Agouza	nationalisn	Internatior	No	Girls only	First attem	0	Yes	Excellent
105422	Female	Literature	Giza	Agouza	nationalisn	Internatior	No	Girls only	First attem	0	Yes	Excellent
105430	Male	Literature	Giza	Dokki	narmer p	Internatior	No	Boys only	First attem	0	No	Acceptab
105432	Male	Literature	Giza	Dokki	freedom in	Internatior	No	Boys only	First attem	0	Yes	Very gooc
105433	Male	Literature	Giza	Dokki	freedom in	Internatior	No	Boys only	First attem	0	Yes	Very gooc
105434	Male	Science - M	Giza	Dokki	freedom in	Internatior	No	Boys only	First attem	1	No	Acceptab

682K RECORD

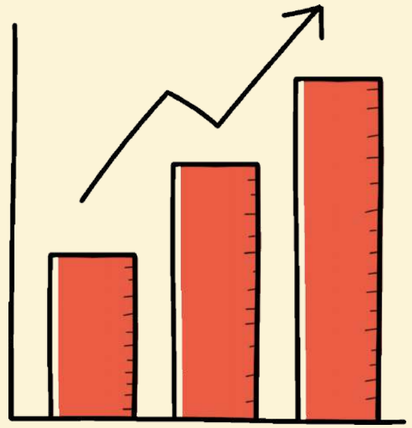
51 COLUMNS

13 CATEGORICAL COLS

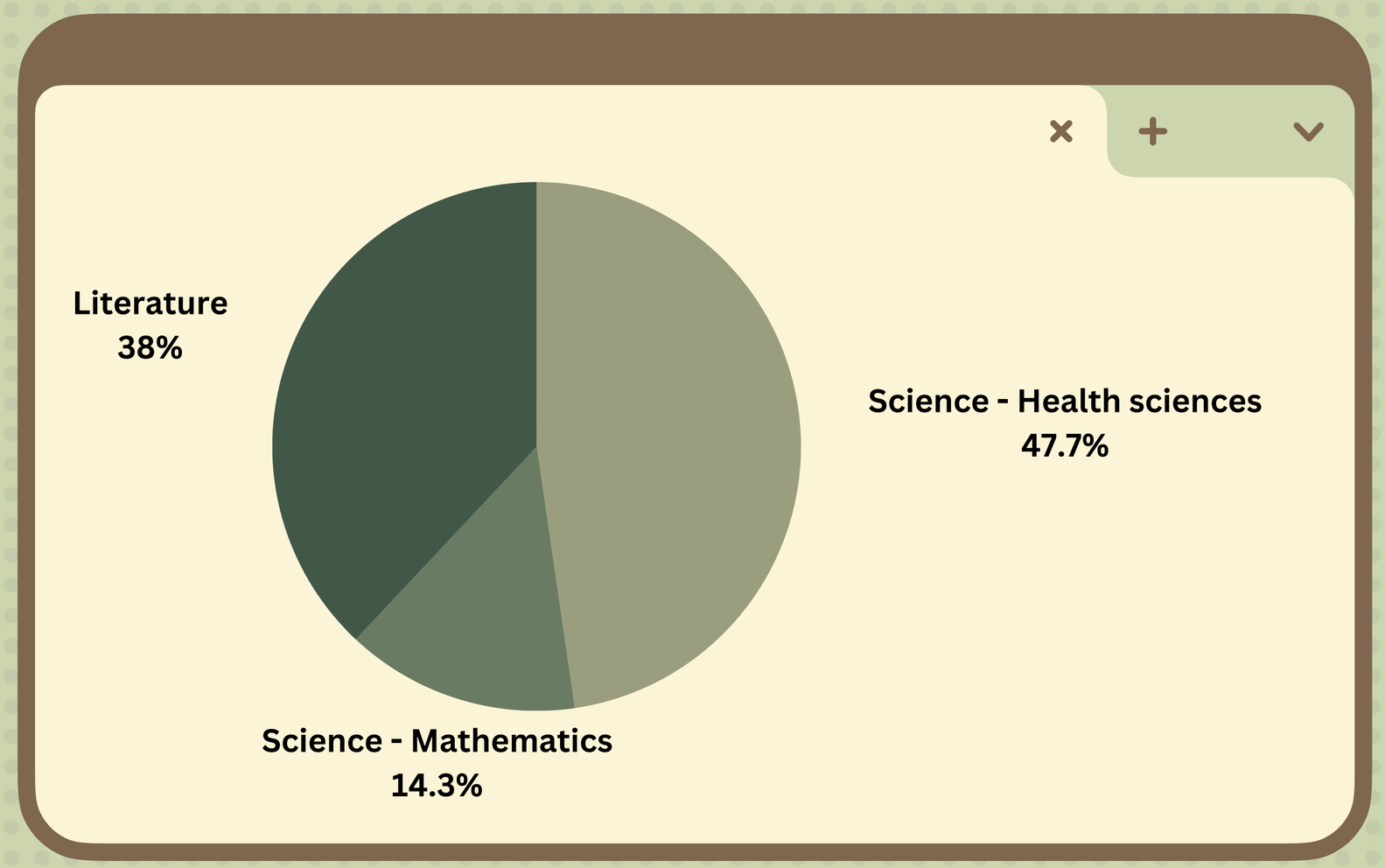
38 NUMERIC COLS

- Do grades differ based on the branch?
- Has **Egypt** achieved the perfect **normal distribution** for the grading curve?
- Do we have **gender equality** in our schools?
- Were there any unusual cases that happened to students during their exams?
- What exactly happens if a student fails or misses their exam? Are they given *another chance*? And does their **score improve** once they get a second chance?
- How do students with special needs perform in their exams?
- For **Egypt** and **Saudi**, Do we have the **same schooling system**? How do schools handle students with special needs?

EXPLOTARY DATA ANALYSIS



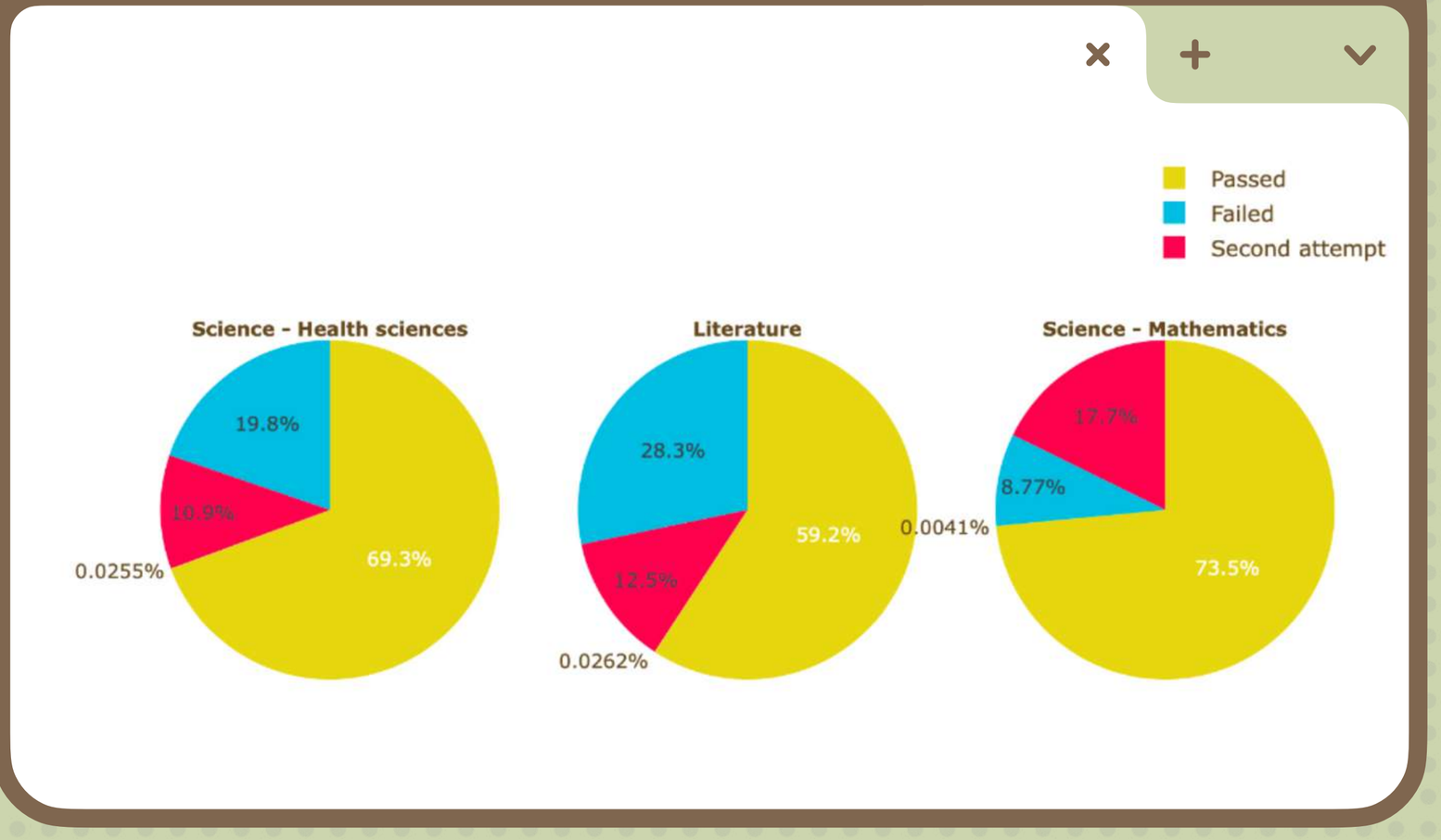
HIGH SCHOOL BRANCH DISTRIBUTION



KEY INSIGHTS:

- The **majority of students are studying Health sciences where they represent 48%**
- **38%** represent students are studying **Literature**
- While **14% represents students are studying the Mathematics** branch

STUDENT BRANCHES BY THE SCORE STATUS

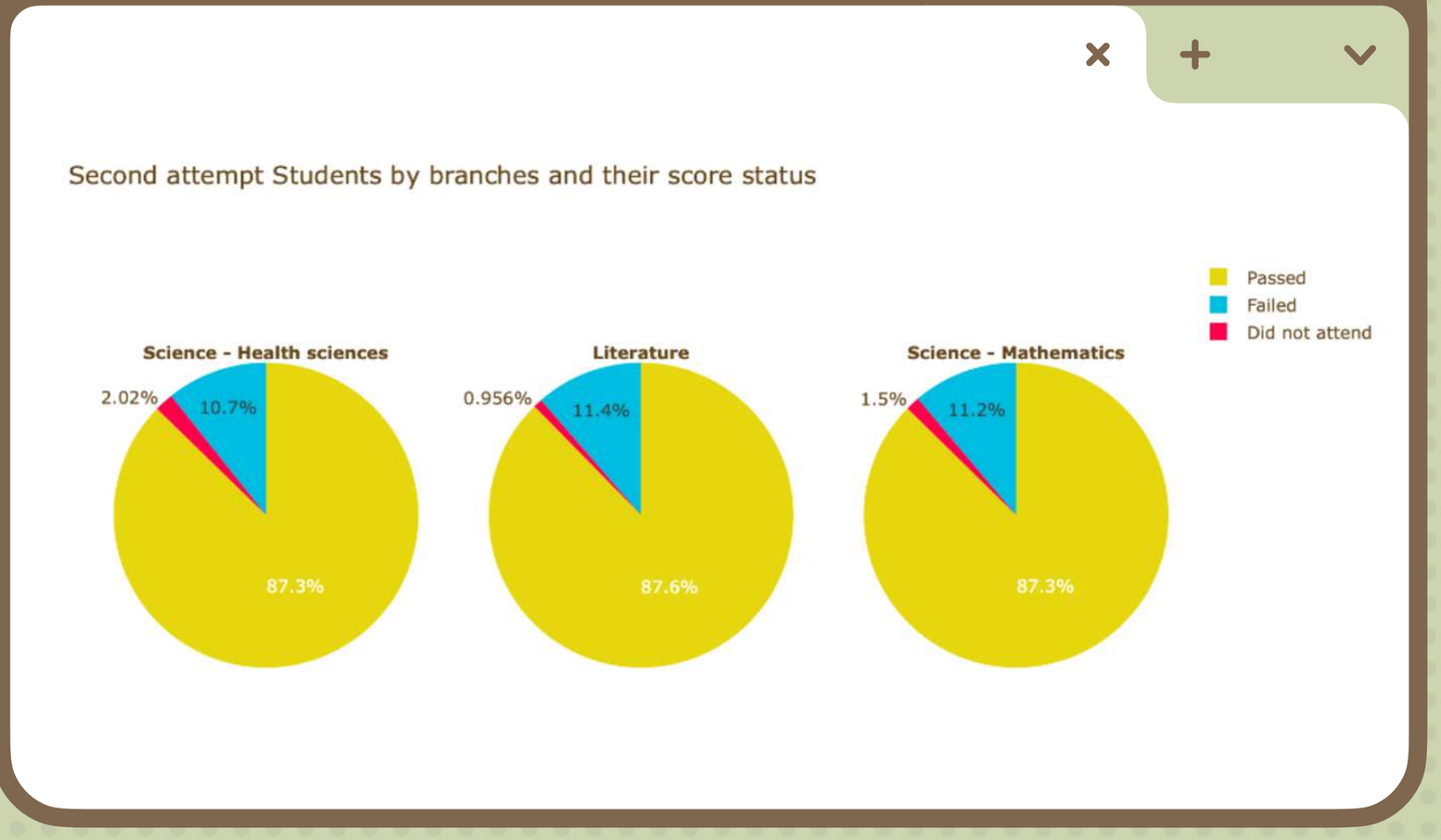


KEY INSIGHTS:

- **Health sciences and Literature students have similar distribution of scores**, with the majority passing their exams, and both have the **highest failure rates among the branches**
- While **73% of Mathematics students are passing their exams**, however, **~18% are moving towart their second attempts** (i.e, This is for students who fail in one or two subjects only)

WE WANT TO TAKE A CLOSER LOOK INTO THE STUDENT BRANCHES BY STUDENTS WHO WENT FOR THEIR SECOND ATTEMPT

DID STUDENT'S SCORE STATUS CHANGE AFTER THEIR SECOND ATTEMPT?



KEY INSIGHTS:

- **All branches have similar scores distribution**
- **~87 of the Students that failed in one or two subjects, have passed the exams in their second attempt**
- ~11 of the Students failed their exams after their second attempt too
- A minority of Students didn't care attend their second attempt

WE WANT TO INVESTIGATE THE SECOND ATTEMPT STUDENTS WHAT ARE THE SUBJECTS THAT THEY ARE FAILING MOSTLY AT?

F WHAT ARE THE SUBJECTS THAT STUDENTS ARE FAILING MOSTLY AT?

USED CRITERIA:

- FOR EACH BRANCH FILTER BY SECONED ATTEMPT STUDENTS
- GET THE MEAN COMPARED WITH THE TOTAL SCORE

IF THE MEAN IS LESS THAN HALF

THIS IS A COMMON SUBJECT FOR STUDENTS TO FAIL AT

EXAMPLE:

MATH TOTAL SCORE IS OUT OF 60

30 IS THE PASS RATE

HEALTH SCIENCES

SUBJECTS:

- ARABIC
- FIRST FOREIGN LANGUAGE (ENGLISH)
- SECOND FOREIGN LANGUAGE
- CHEMISTRY
- BIOLOGY
- GEOLOGY
- PHYSICS

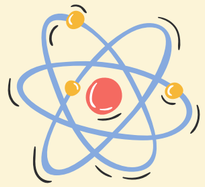
SUBJECTS THAT STUDENTS ARE FAILING MOSTLY AT?

HEALTH SCIENCES

SUBJECTS:

- ARABIC
- FIRST FOREIGN LANGUAGE (ENGLISH)
- SECOND FOREIGN LANGUAGE
- CHEMISTRY
- BIOLOGY
- GEOLOGY
- PHYSICS

SUBJECTS THAT STUDENTS ARE FAILING MOSTLY AT?



PHYSICS

25.1/60

CHEMISTRY

28.7/60

LITERATURE

SUBJECTS:

- ARABIC
- FIRST FOREIGN LANGUAGE (ENGLISH)
- SECOND FOREIGN LANGUAGE
- HISTORY
- GEOGRAPHY
- PHILOSOPHY
- PSYCHOLOGY

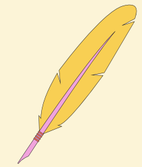
SUBJECTS THAT STUDENTS ARE FAILING MOSTLY AT?

LITERATURE

SUBJECTS:

- ARABIC
- FIRST FOREIGN LANGUAGE (ENGLISH)
- SECOND FOREIGN LANGUAGE
- HISTORY
- GEOGRAPHY
- PHILOSOPHY
- PSYCHOLOGY

SUBJECTS THAT STUDENTS ARE FAILING MOSTLY AT?



ENGLISH — 23.7/60

HISTORY — 28.6/60

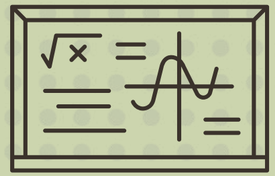
GEOGRAPHY — 29.8/60

MATHEMATICS

SUBJECTS:

- ARABIC
- FIRST FOREIGN LANGUAGE (ENGLISH)
- SECOND FOREIGN LANGUAGE
- PURE MATHEMATICS
- CHEMISTRY
- APPLIED MATH
- PHYSICS

SUBJECTS THAT STUDENTS ARE FAILING MOSTLY AT?



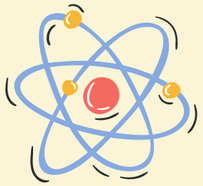
MATHEMATICS

SUBJECTS:

- ARABIC
- FIRST FOREIGN LANGUAGE (ENGLISH)
- SECOND FOREIGN LANGUAGE
- PURE MATHEMATICS
- CHEMISTRY
- APPLIED MATH
- PHYSICS

SUBJECTS THAT STUDENTS ARE FAILING MOSTLY AT?

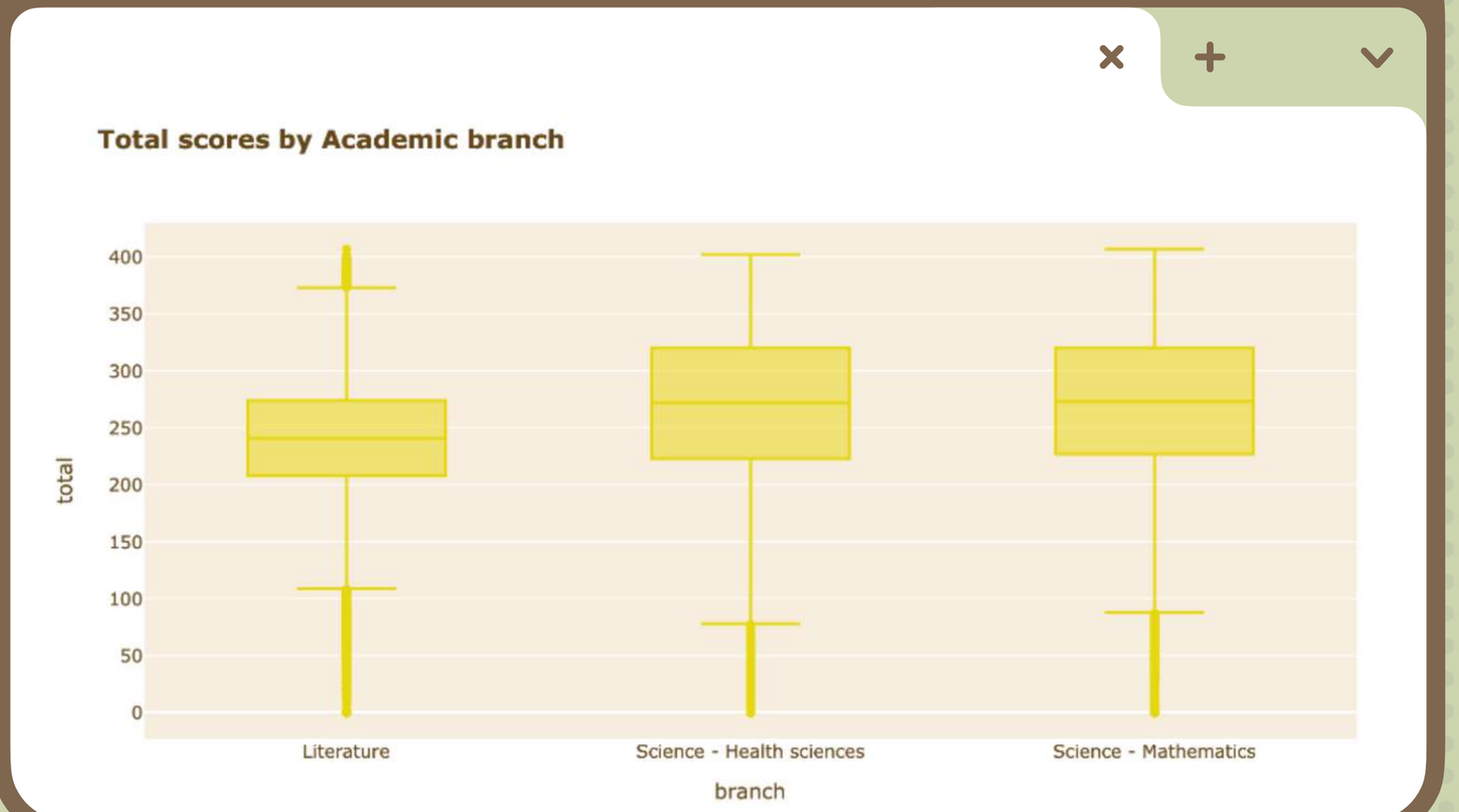
PURE MATHEMATICS — 27.5/60



PHYSICS — 28.8/60

CHEMISTRY — 29.2/60

NOW LET'S TAKE A LOOK AT THEIR TOTAL SCORES

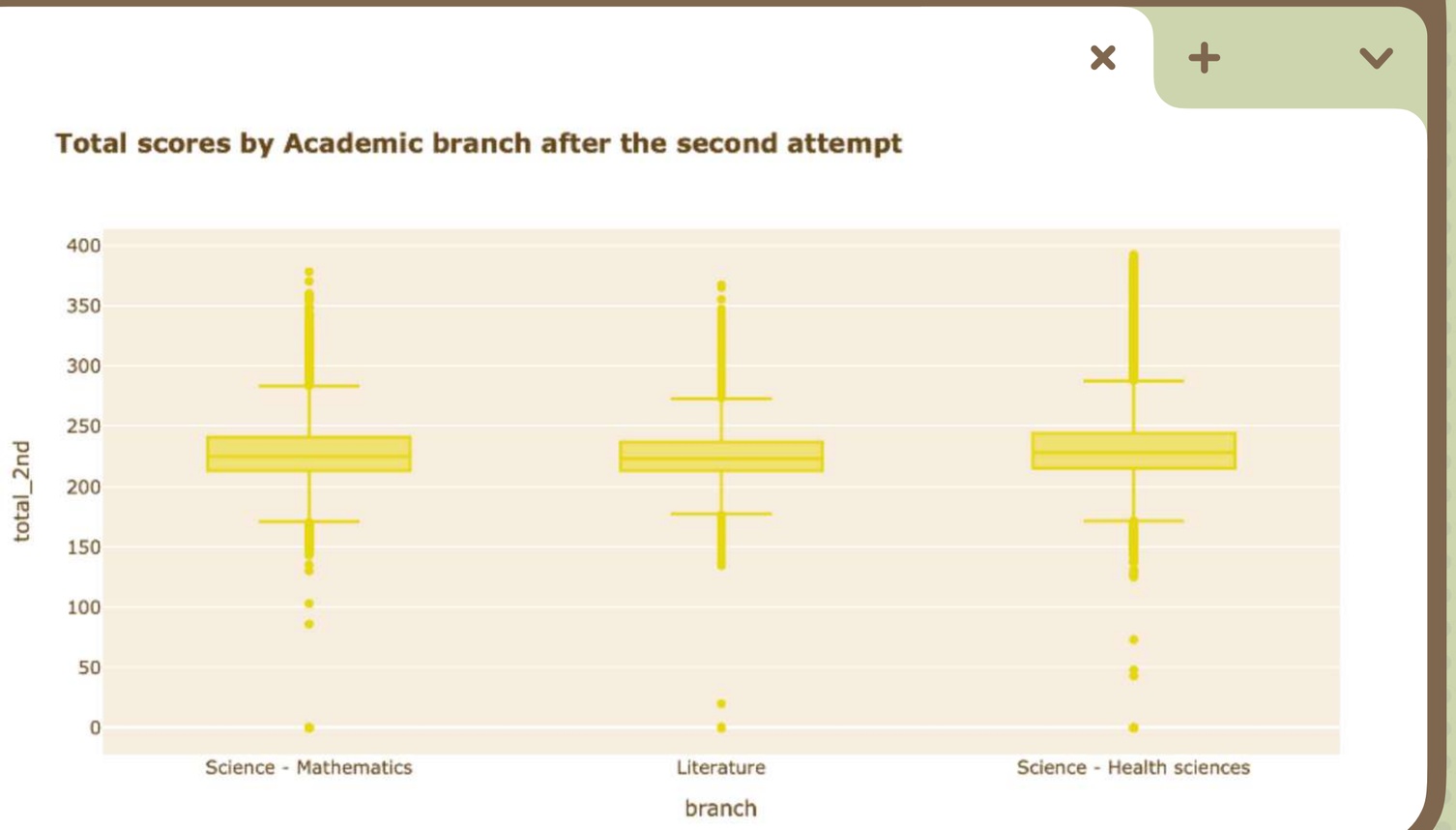


KEY INSIGHTS:

- None of the +600k students obtained a full mark 410/410 the **highest score recorded was at 407**, which is just 3 points away from a full mark
- **Health sciences and Mathematics** students have similar distribution of scores, **on average students score around (225 - 320)/ 410**
- While **on average Literature students score on a lower scale around (208 - 274)/ 410**

WE WANT TO INVESTIGATE THE TOTAL SCORES AFTER HEADING TO THEIR SECOND ATTEMPT ?

NOW LET'S TAKE A LOOK AT THEIR SECOND ATTEMPT'S TOTAL SCORES



KEY INSIGHTS:

- None of the +600k students obtained a full mark 410/410 the **highest score recorded of the second attempt was at 392,** which is **lower than the first attempt students**
- **Fewer outliers on the lower scores scale**, students pull on more efforts this time to pass their exams

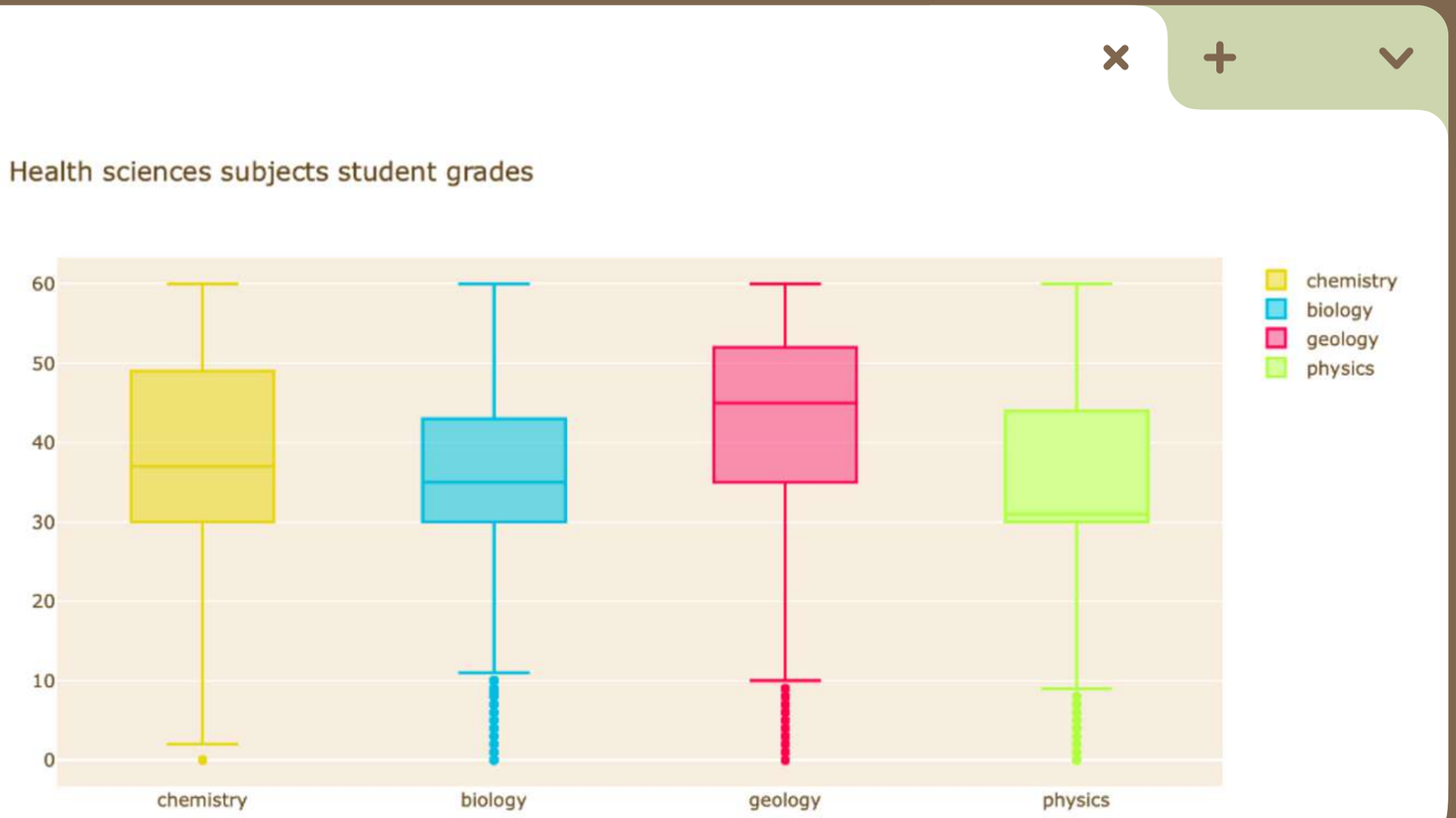
COMMON SUBJECTS BETWEEN ALL BRANCHES GRADES DISTRIBUTION

KEY INSIGHTS:

- **Religion, national education and economics subjects are not included within the cumulative final grade**, and they contain 100 grades
- Arabic, English, and second foreign languages have different grading, in total, they contain 160 out of the total grade

WE WANT TO INVESTIGATE THE SUBJECTS BY EACH BRANCH

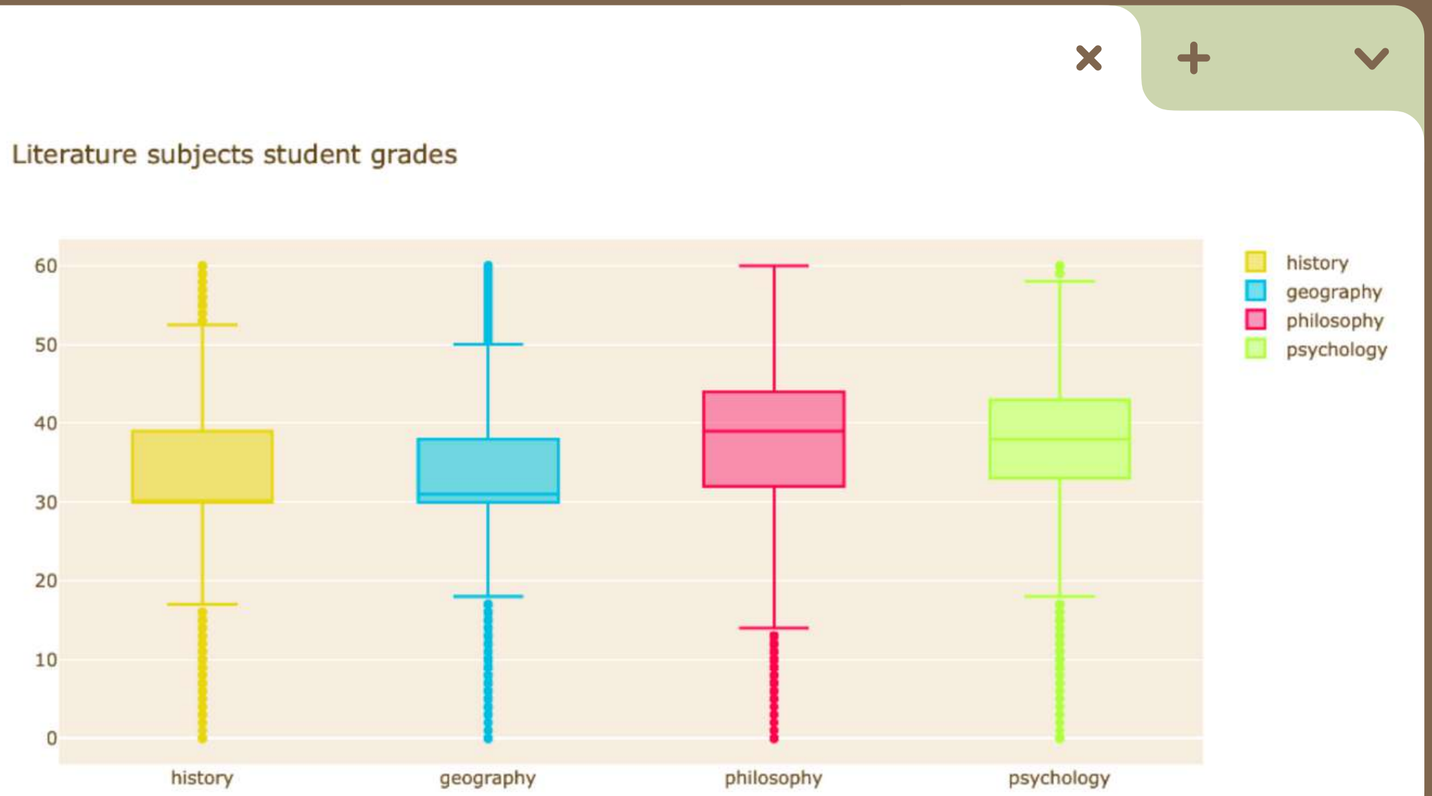
HEALTH SCIENCES SUBJECTS STUDENT GRADES



KEY INSIGHTS:

- Compared to the common subjects, here we can clearly see that the grades are out of 60, and **the averages are slightly above the minimum pass score**
- **Chemistry has the lowest scores among all subjects**

LITERATURE BRANCH SUBJECTS STUDENT GRADES

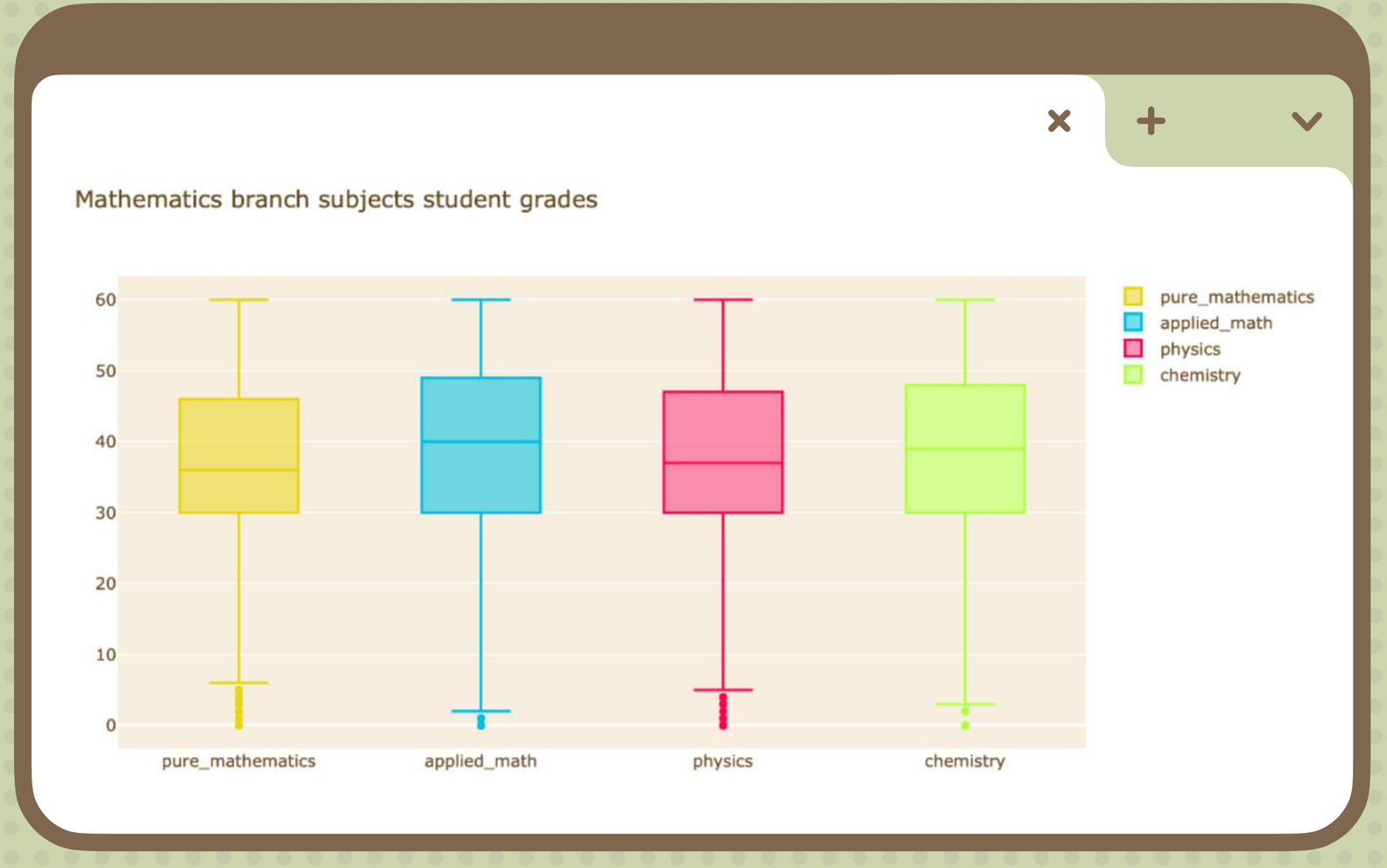


KEY INSIGHTS:

- We can clearly spot that there are **a lot of outliers at the higher and lower scale scores**
- Philosophy has the lowest scores among all subjects, however, history has the lowest on average

MATHEMATICAL BRANCH SUBJECTS STUDENT GRADES

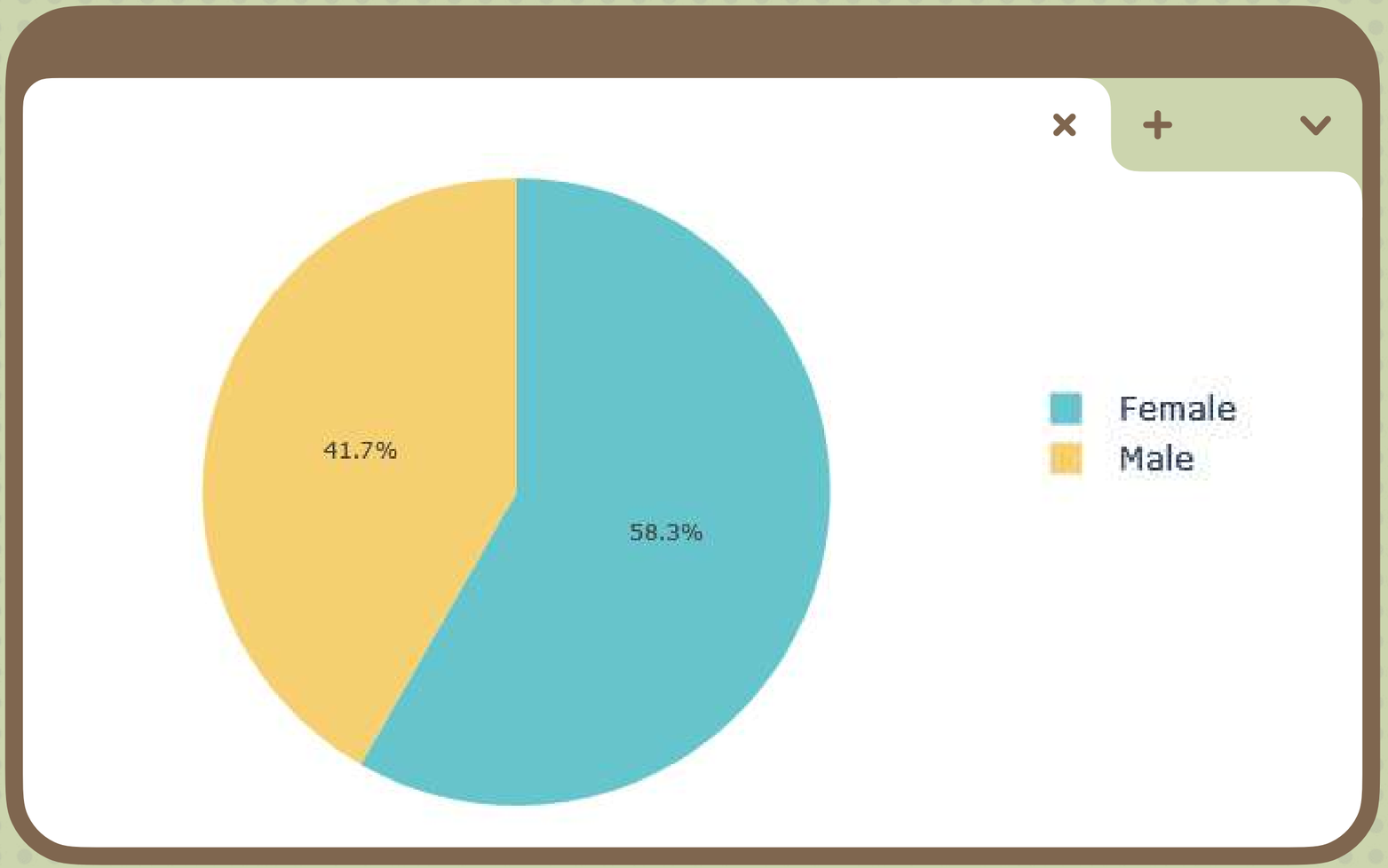
$$V = \frac{4}{3}\pi r^3$$



KEY INSIGHTS:

- Compared to Literature students, **mathematics branch students tend to achieve full marks**
- A **higher passing rate،due to scores reaching above average**

SELF SCAMMED STUDENTS BY GENDER



Scam Alert

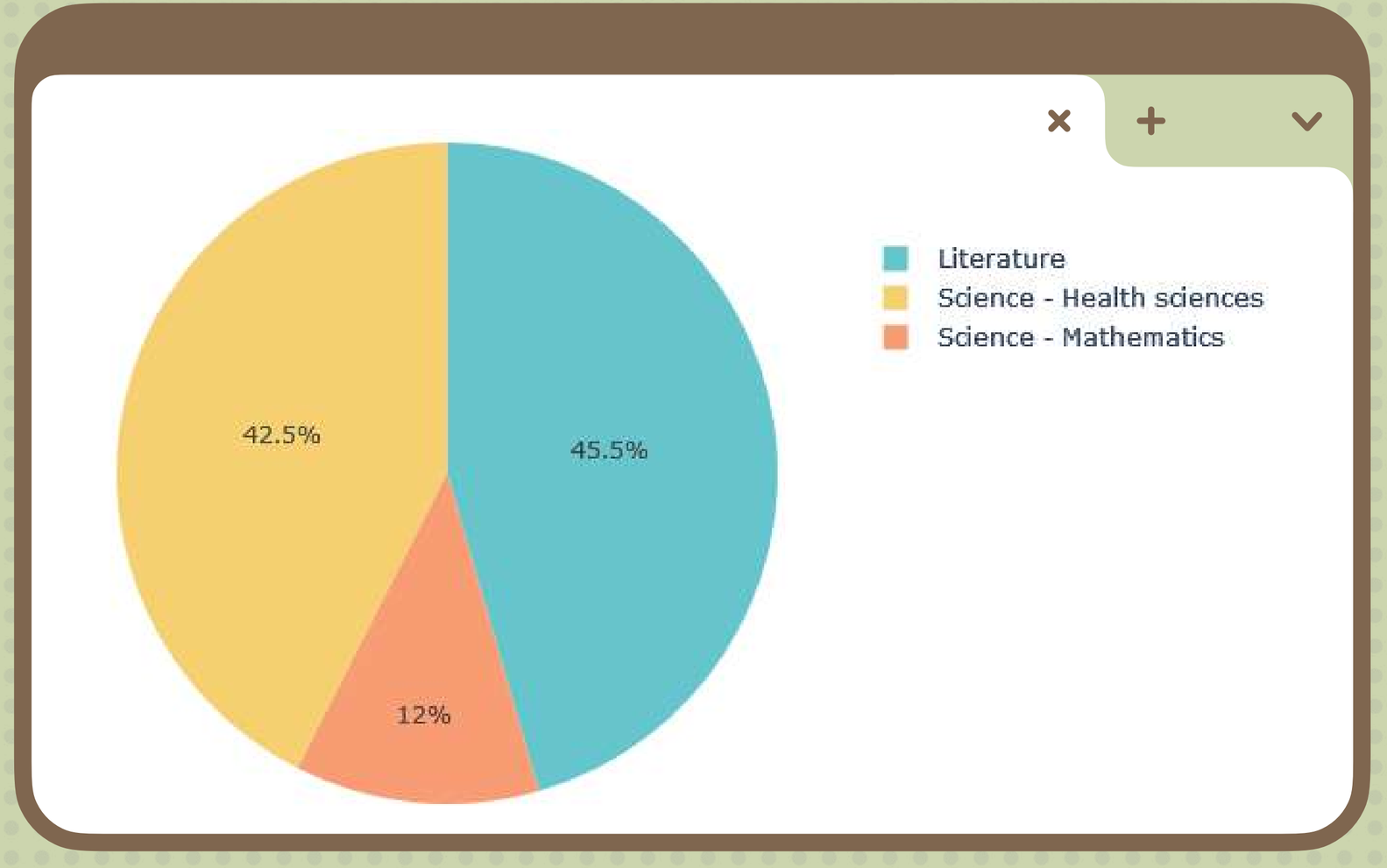
KEY INSIGHTS:

Females have tried *both attempts* and received **lower** grade on the second attempt

WE WANT TO INVESTIGATE THE BRANCH WERE THEY GOT SCAMMED THE HIGHEST

SELF SCAMMED STUDENTS BY BRANCH

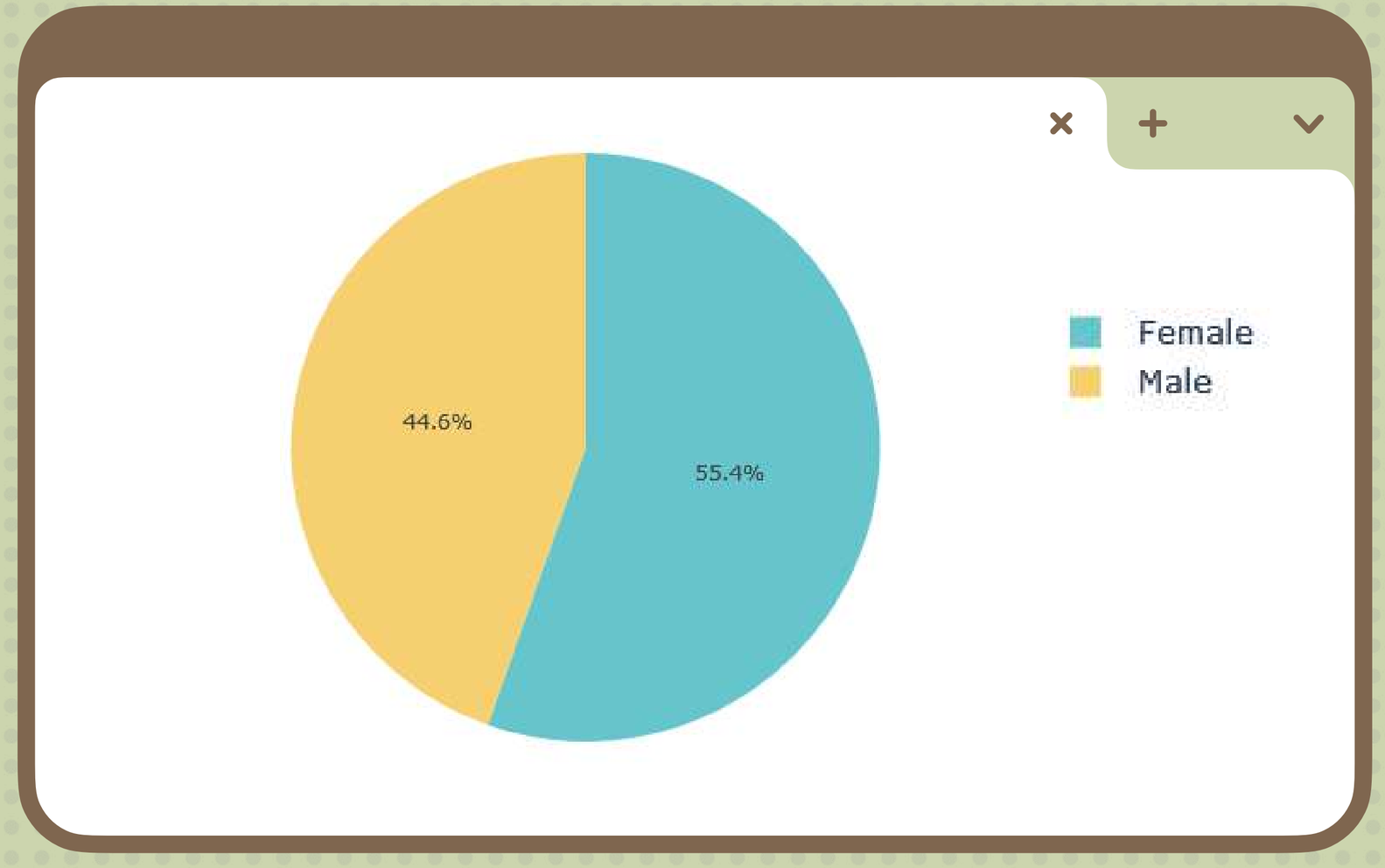
Scam Alert



KEY INSIGHTS:

Literature and **Health sciences** have the highest proportion of getting **lower** grades on the *second attempt.*

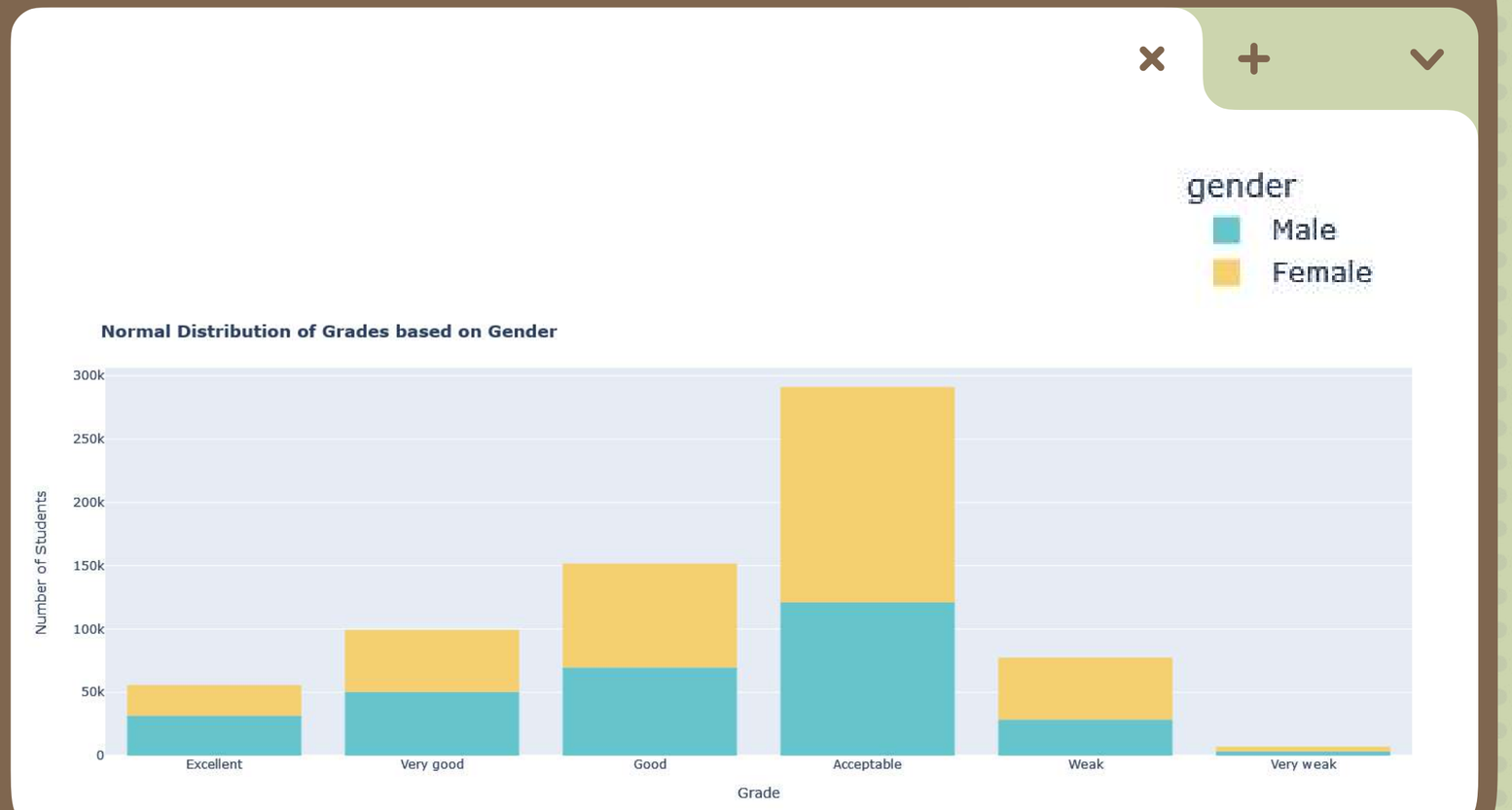
GENDER DISTRIBUTION IN EGYPT 2022



KEY INSIGHTS:

The proportion of males to females is very close. We can say that Egypt has **gender equality** in their schools

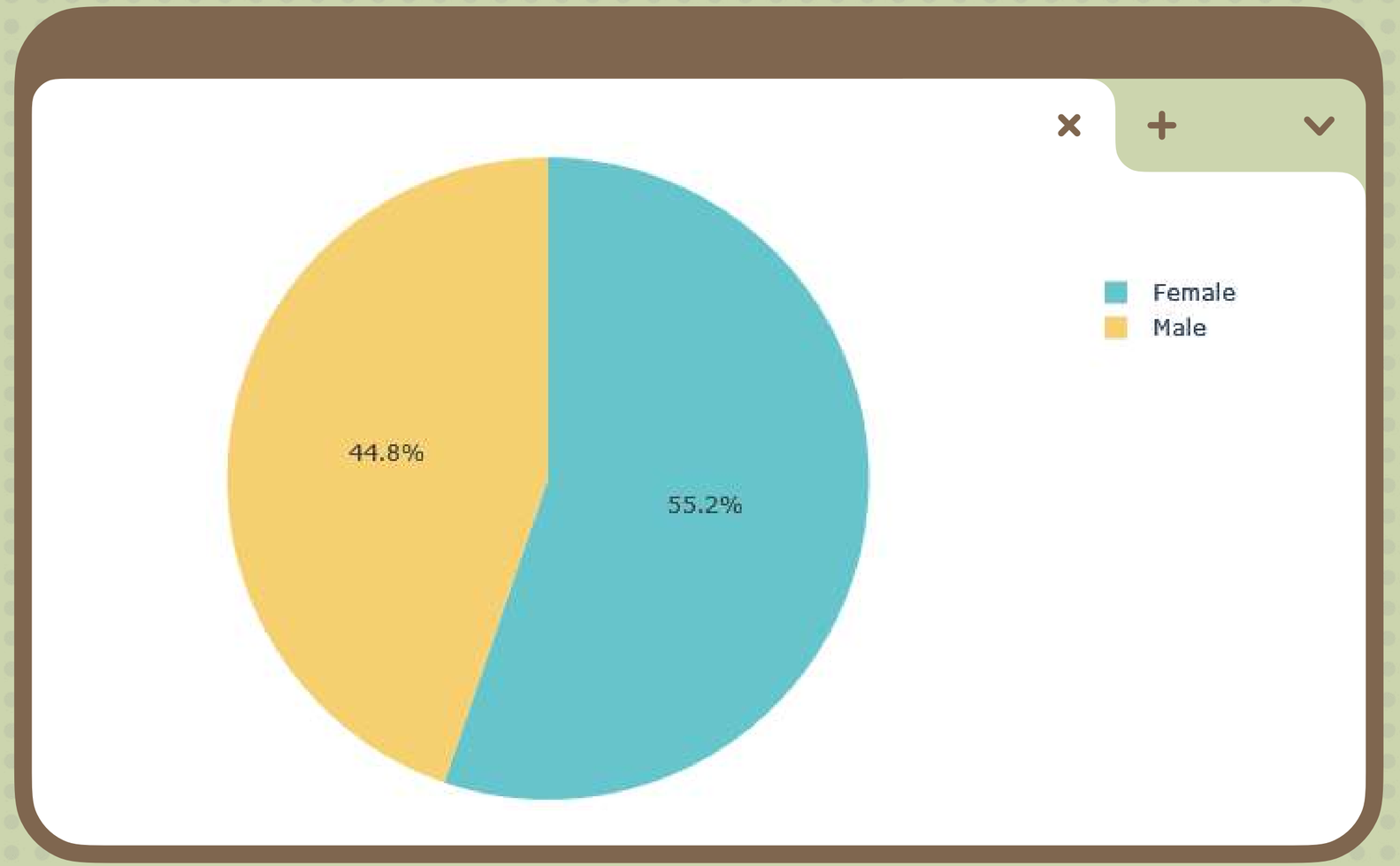
GRADES DISTRIBUTION BY GENDER



KEY INSIGHTS:

- We have a normal distribution for grades. "Schools dream"
- Only excellent, acceptable, and very weak have different proportions between males and females.
- Females have more acceptable and weak grades.
- Males have more excellent grades.

SPECIAL NEEDS GENDER DISTRIBUTION

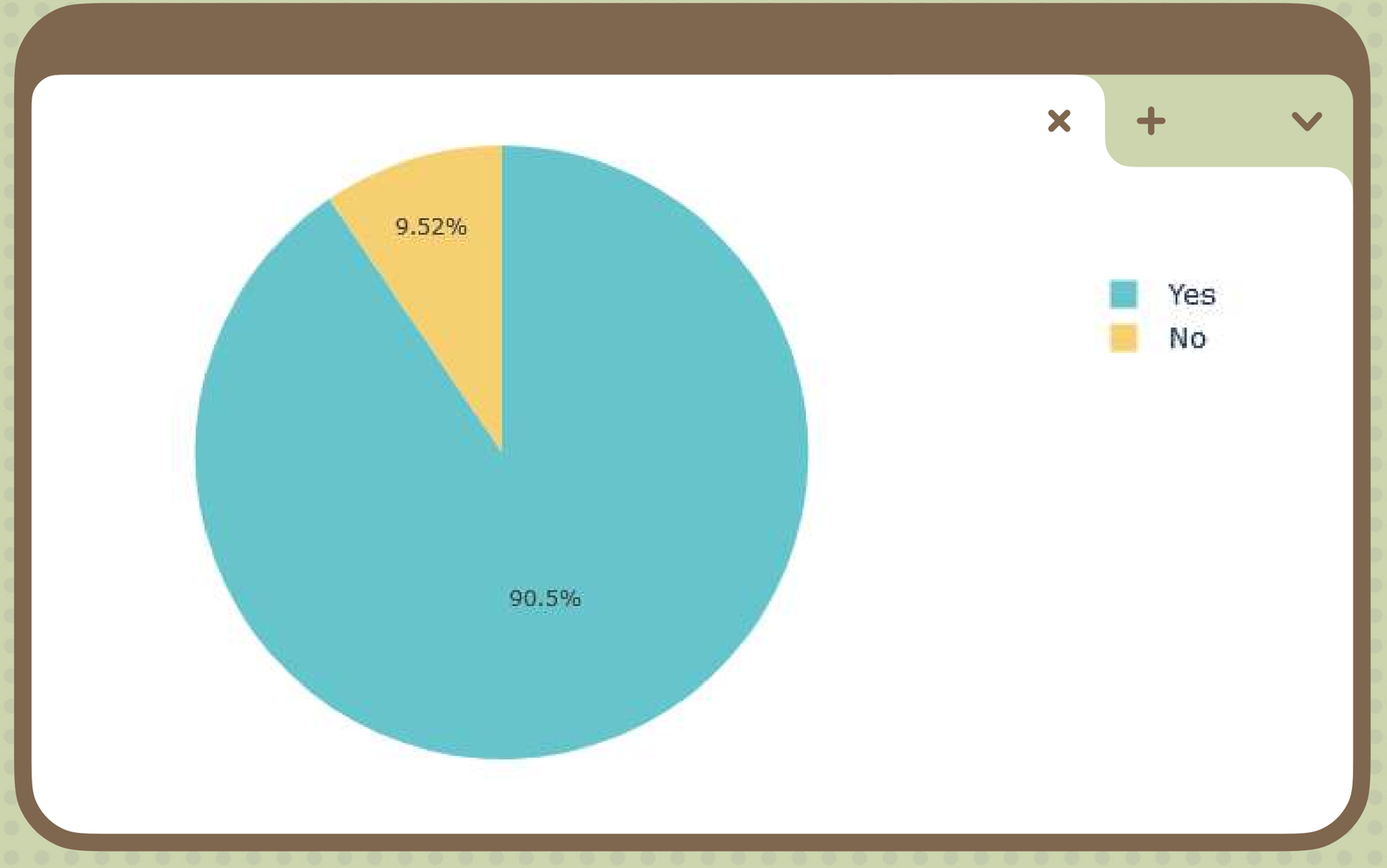


KEY INSIGHTS:

Majority of students with **disabiliites** are *females*.

WE WANT TO INVESTIGATE THE CHANCE OF THEM GETTING ACCEPTED INTO UNIVERSITIES

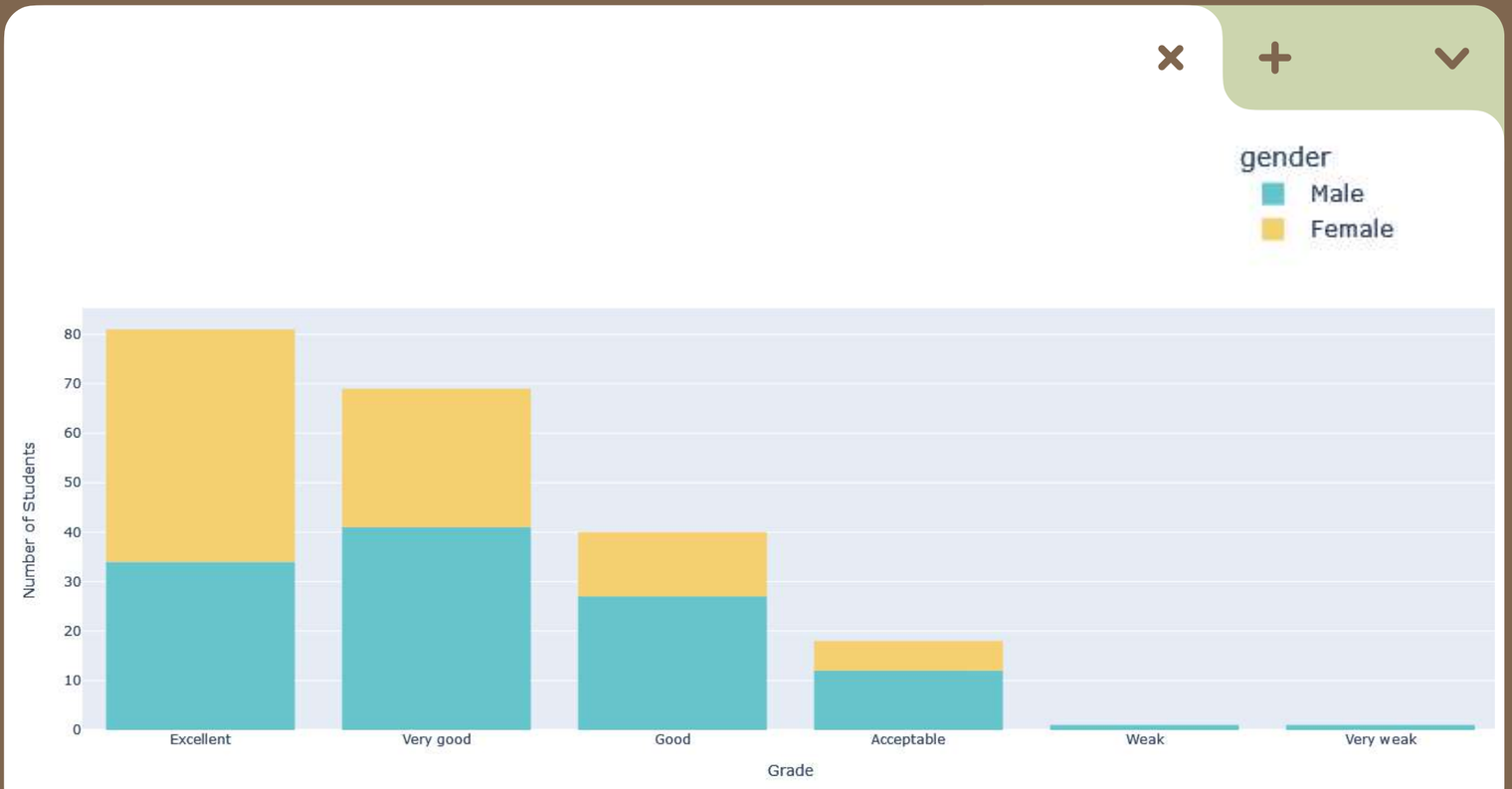
HOW MANY STUDENTS WITH SPECIAL NEEDS CAN JOIN THE UNIVERSITY?



KEY INSIGHTS:

The majority of students **can join** university.

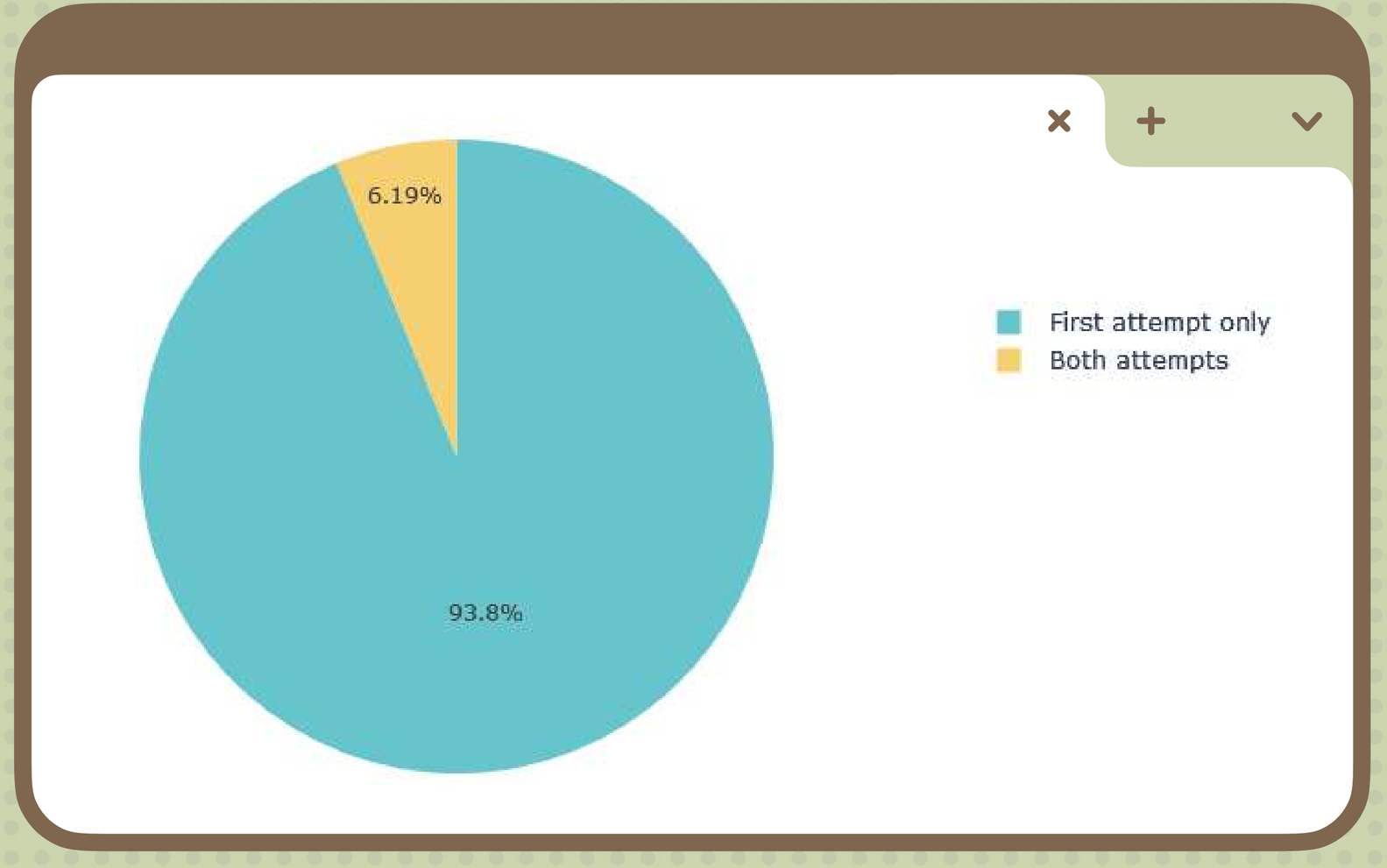
SPECIAL NEEDS GRADING DISTRIBUTION



KEY INSIGHTS:

The graph is **right-skewed** which *doesn't achieve* the normal grading distribution.

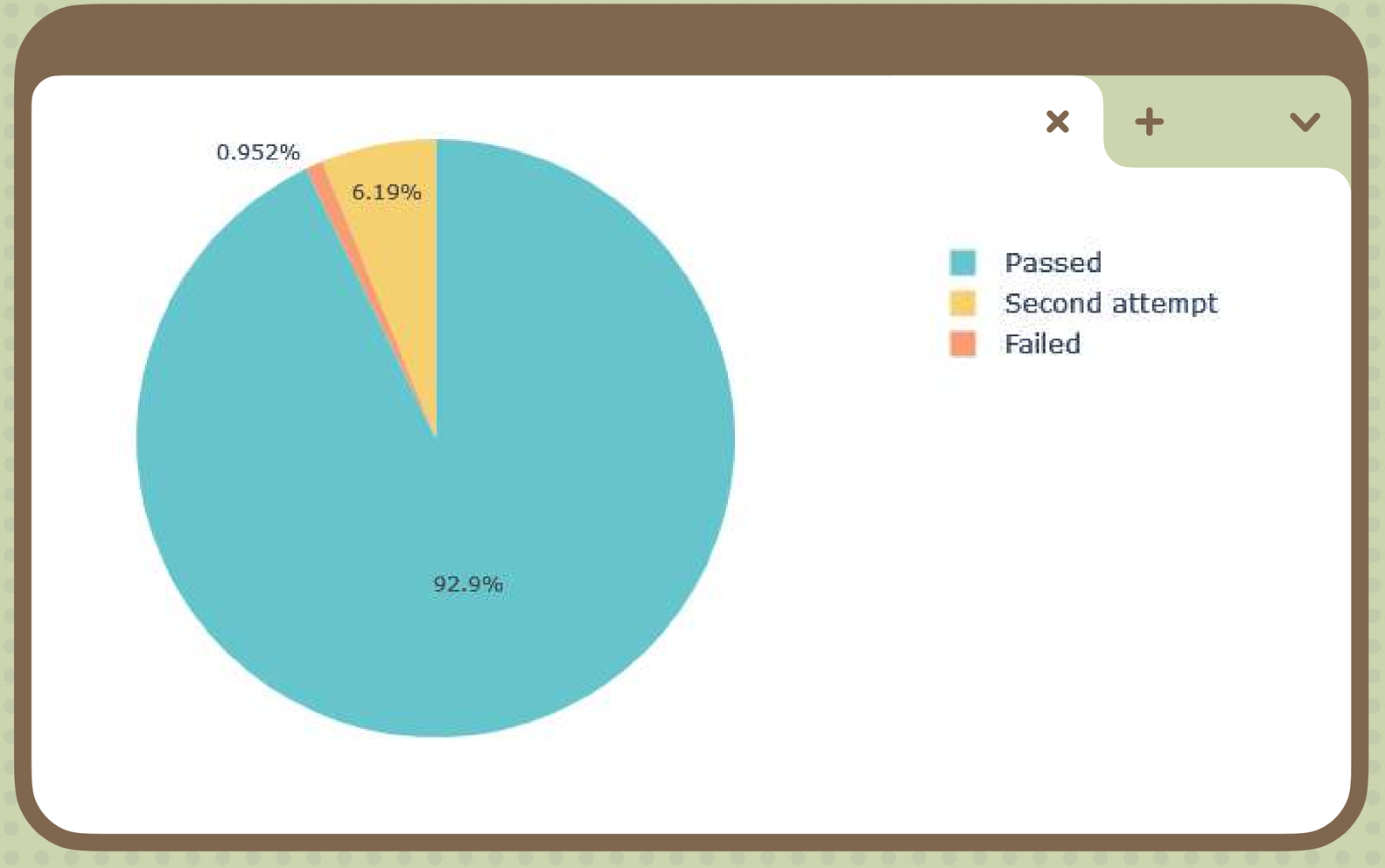
SPECIAL NEEDS EXAM'S ATTEMPTS



KEY INSIGHTS:

None of the students had only the second attempt, as a conclusion, **no unusual cases** have happened.

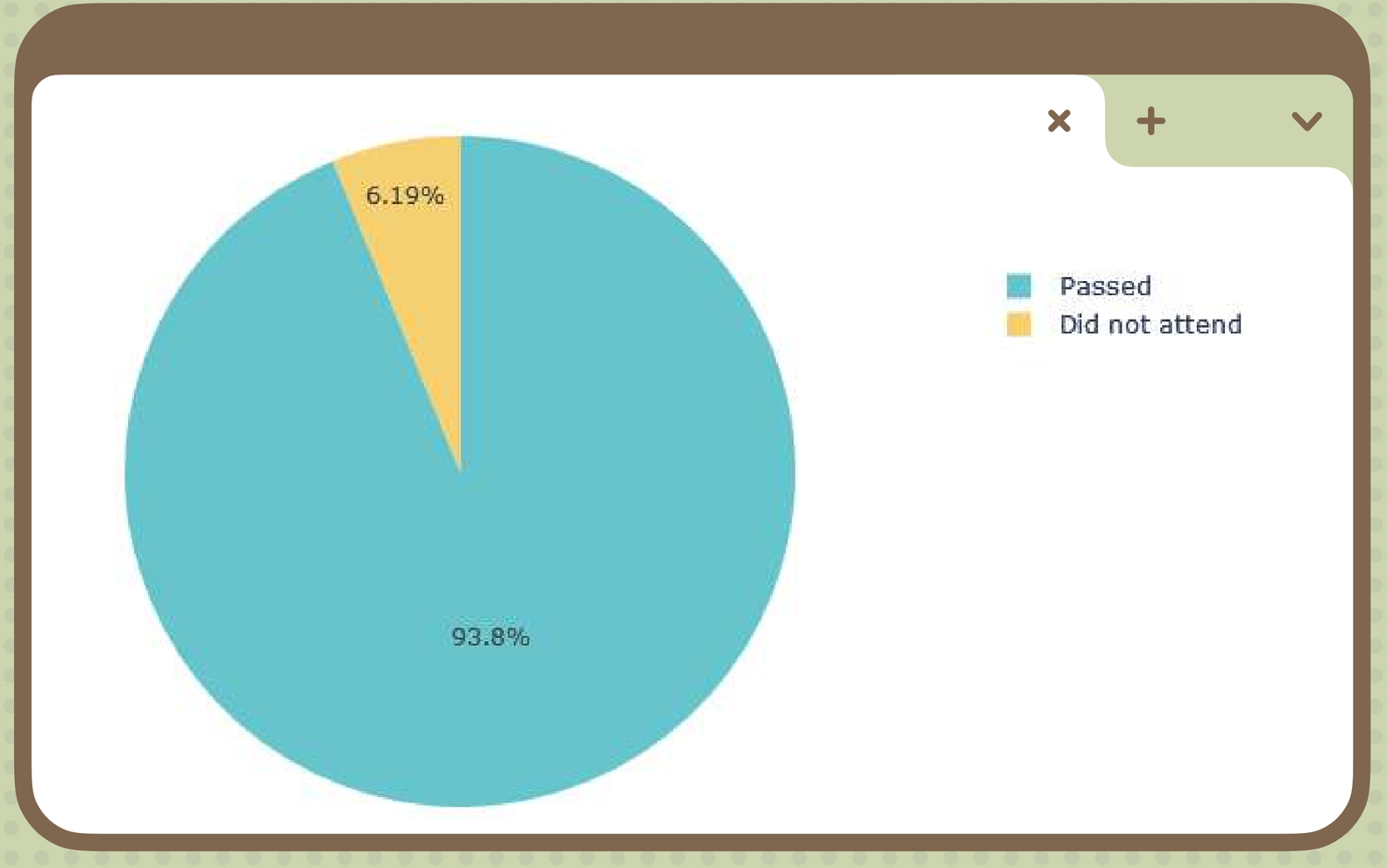
SPECIAL NEEDS EXAM'S STATUS



KEY INSIGHTS:

Most students have **passed**. from the *first attempt*, and a small minority had a **second chance**. While a *very little* have **failed** their exams.

SPECIAL NEEDS EXAM'S SECOND STATUS

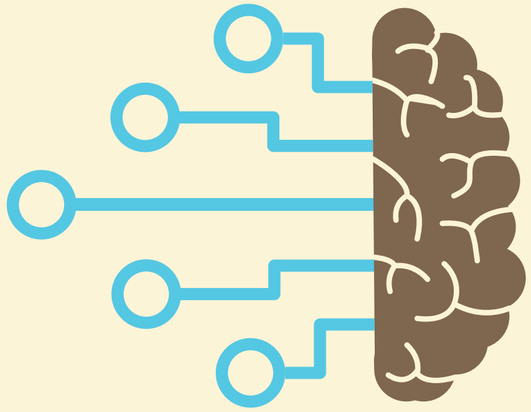


KEY INSIGHTS:

Most students have *passed* their exams in their second chance, while a small minority *did not attend* their second attempt maybe due to **unusual cases**.

DASHBOARDS

ML USING SPARK



FEATURE ENGINEERING STEPS

1

REMOVING COLUMNS

2

OVERSAMPLING

3

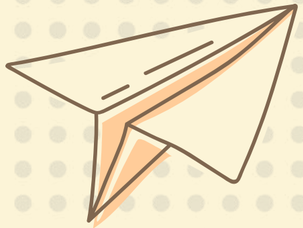
LABEL ENCODING

4

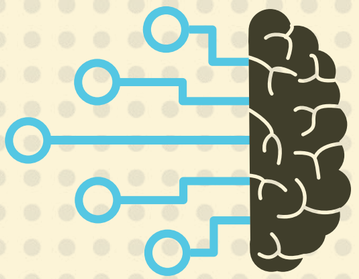
VECTOR ASSEMBLER

5

STANDARD SCALING



ML MODELS

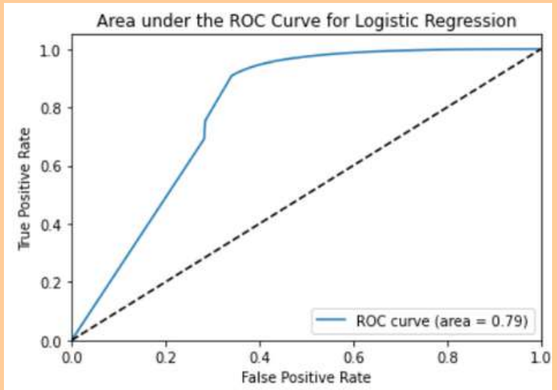


1. DECISION TREE
2. RANDOM FOREST
3. LOGISTIC REGRESSION
4. GRADIENT BOOST

ML MODELS EVALUATION

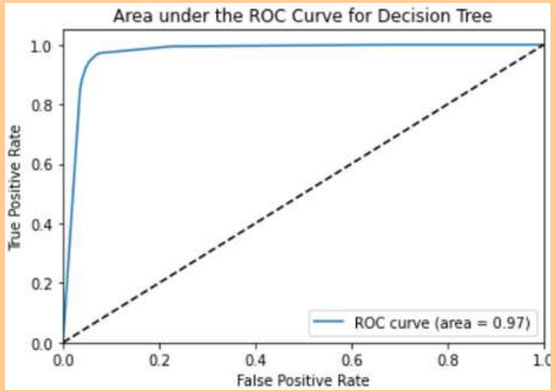
55%

LOGISTIC REGRESSION



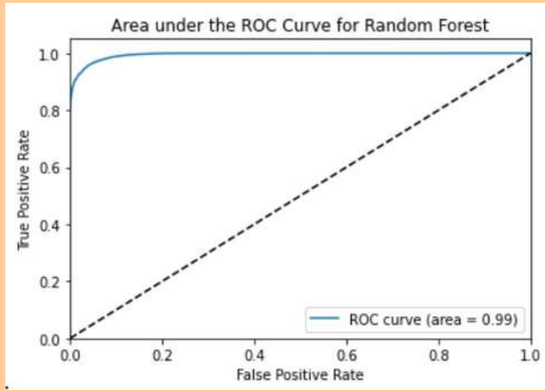
95%

DECISION TREE



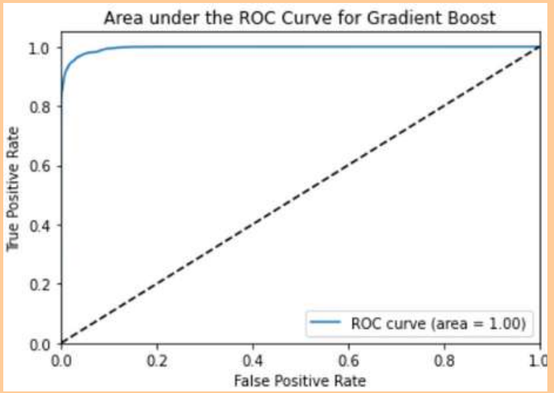
95.5%

RANDOM FOREST



96%

GRADIENT BOOST



MODEL SELECTION & OPTIMIZATION

GRADIENT BOOST:

96%

97%

PIG QUERIES

HOW MANY 12 GRADE STUDENTS HAVE ENTERED THE PUBLIC EXAMS IN THE YEAR 2022 FROM EGYPT'S CAPITAL CITY CAIRO?

```
Filter_Cairo = FILTER Egyption_school BY (city == 'Cairo');

--Count the number of students using desk no:
student_group_all = Group Filter_Cairo All;
student_count = foreach student_group_all  Generate COUNT(Filter_Cairo.desk_no);
Dump student_count;
```

Cairo: al-Qāhirah, is **the capital of Egypt and its largest city, home to 10 million people**

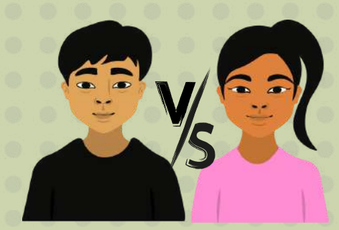
There are 102,446 grade 12 students at the capital city of Egypt (Cairo), which is around **1% of it's total population**

HOW MANY ARE THE GRADE 12 STUDENTS BY THE TYPE OF SCHOOL?

```
--Count the number of students using desk no:
student_group_school_type = Group Egyption_school by school_type;
student_count = foreach student_group_school_type GENERATE group AS School_type,
COUNT(Egyption_school.desk_no)AS count_of_student;
Dump student_count;
```

TYPE	#OFSTUDENTS
 PUBLIC SCHOOL	607,660
 PRIVATE SCHOOL	20,268
 INTERNATIONAL SCHOOL	54,207
 SCHOOL FOR THE BLIND	210

WHAT IS THE AVERAGE SCORE PERCENTAGE FOR BOYS COMPARED TO GIRLS?




```
--filter the dataset where status is Passed:
Filter_Passed = FILTER Egyption_school BY (status == 'Passed');

--The average percentage of scores:
student_group_gender = Group Filter_Passed by gender;
student_average = foreach student_group_gender GENERATE group AS Gender,
AVG(Filter_Passed.percentage)AS average_persentage_by_gender;
Dump student_average;
```

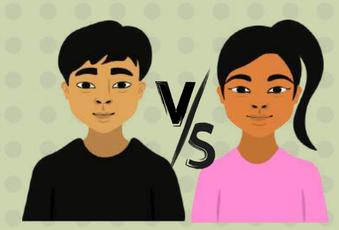
Girls typically outperform boys in humanities, languages and reading tests, while boys do better in maths

But when grades are awarded by teachers, girls do better in all subjects, Sky news

71.86%

69.93%

WHAT WAS THE HIGHEST AND LOWEST TOTAL PERCENTAGE FOR GIRLS COMPARED TO BOYS?




```
--Group by the gender:
group_gender = Group Egyption_school by gender;

-- Count the minimum and maximum percentage
min_max_percentage = FOREACH group_gender GENERATE MIN(Egyption_school.percentage) AS Min_percentage,
MAX(Egyption_school.percentage) AS Max_percentage;

--Describe the results.
DESCRIBE min_max_percentage;

DUMP min_max_percentage;
```

Girls typically outperform boys in humanities, languages and reading tests, while boys do better in maths

But when grades are awarded by teachers, girls do better in all subjects, Sky news

	MAXIMUM	MINIMUM
	99.27%	0%
	99.02%	0%

WHAT IS THE AVERAGE SCORE PERCENTAGE BY BRANCH?

```
--filter the dataset where status is Passed:
Filter_Passed = FILTER Egyption_school BY (status == 'Passed');

--The average percentage of scores:
student_group_branch = Group Filter_Passed by branch;
student_average = foreach student_group_branch GENERATE group AS branch,
AVG(Filter_Passed.percentage)AS average_persentage_by_branch;
Dump student_average;
```

BRANCH	AVERAGE %
 HEALTH SCIENCES	73.59
 MATHEMATICS	72.86
 LITERATURE	65.86

WHAT ARE THE HIGHEST AND LOWEST AVERAGE PERCENTAGE RATE BY CITY?

```
--filter the dataset where city is Ciro:
Filter_Passed = FILTER Egyption_school BY (status == 'Passed');

--The average percentage of scores:
student_group_city = Group Filter_Passed by city;
student_average = foreach student_group_city GENERATE group AS branch, AVG(Filter_Passed.percentage)AS
average_persentage_by_city;
order_student_average = ORDER student_average BY average_persentage_by_city DESC;
Dump order_student_average;
```

LOWEST 3

CITY	MINYA	ASWAN	GIZA
AVERAGE %	67.96	67.90	67.52

TOP 3

	CITY	AVERAGE %
	NORTH SINAI	77.82
	DAKAHLIA	75.30
	PORT SAID	74.35

WHAT IS THE NUMBER OF HOMESCHOOLED STUDENTS?

```
--filter the dataset where homeschooling is yes:
Filter_homeschooling = FILTER Egyption_school BY (homeschooling == 'Yes');

--Count the number of students using desk no:
student_group_all = Group Filter_homeschooling All;
student_count = foreach student_group_all  Generate COUNT(Filter_homeschooling.desk_no);
Dump student_count;
```

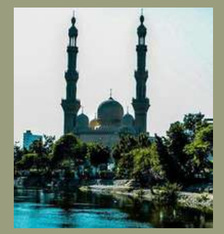
According to the Egyptian Ministry of Education, the **students who repeated the third year of high school twice, or the students whose days of absence are more than their attendance**, are **transferred to homeschooling schools**

There are 48,428 grade 12 homeschooled students at Egypt

WHAT IS THE GENDER, SCHOOL NAME, AND ADMINISTRATION OF THE PEOPLE WHO GOT THE HIGHEST SCORE IN 2022?

```
--2. filter the data based on the highest total
filtered = FILTER Egyption_school BY (total>=407);

--3. group the data and get the maximum total
grouped_data = GROUP filtered BY (gender, school_name_translated, administration_translated);

max_grade = FOREACH grouped_data GENERATE FLATTEN(group) AS (gender, school_name_translated,
administration_translated), MAX(filtered.total) AS Highest_grade;

DUMP max_grade;
```

INTERNATIONAL SCHOOL



EAST ZAGAZIG

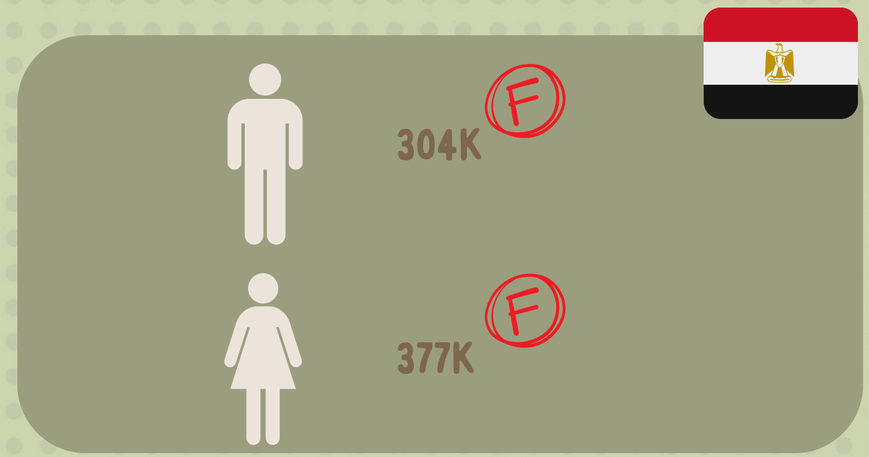
×2

PUBLIC SCHOOL

KAFR SAQR

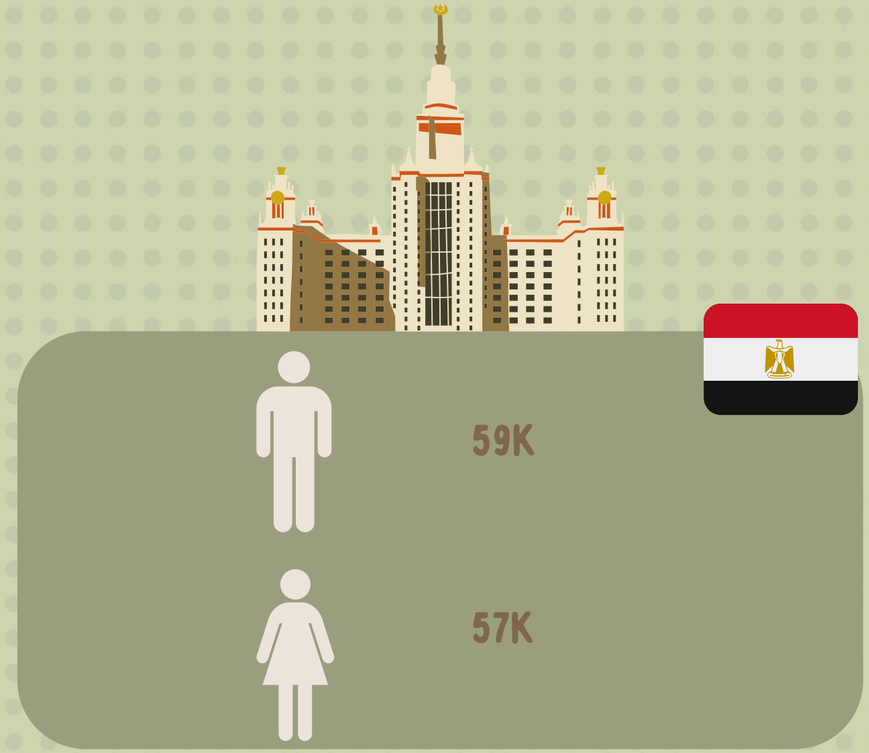
WHAT IS THE NUMBER OF FAILED COURSES PER GENDER?

```
--2. group the data and get the maximum failed courses
grouped_data = GROUP Egyption_school BY gender;

gender_failed_courses = FOREACH grouped_data GENERATE group AS gender,
COUNT(Egyption_school.no_of_failed_courses) AS Number_of_failed_courses;

DUMP gender_failed_courses;
```



WHICH GENDER HAS A HIGHER NUMBER OF JOINING THE UNIVERSITY BASED ON THEIR CURRENT MAJOR?

```
--2. filter the data based on joining condition
filtered = FILTER Egyption_school BY (can_join_uni == 'Yes');

--3. group the data and get the students count
grouped_data = GROUP filtered BY gender;

join_uni = FOREACH grouped_data GENERATE group AS gender, COUNT(filtered.desk_no) AS gender_join_uni;

DUMP join_uni;
```



FOR EACH CITY, WHICH BRANCH HAS THE HIGHEST STUDENT NUMBER?

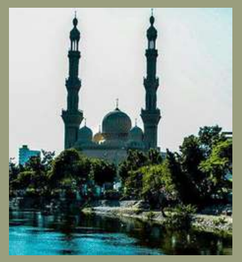
```
--2. group the data and Get the count of students for each city per branch
grouped_data = GROUP Egyption_school BY (city, branch);

branch_count_data = FOREACH grouped_data GENERATE FLATTEN(group) AS (city, branch),
COUNT(Egyption_school.desk_no) AS branch_count;

--3. Order the result in descending order
order_by_data = ORDER branch_count_data BY branch_count DESC;

DUMP order_by_data;
```

	BRANCH	STUDENTS
	LITERATURE	47K
	HEALTH SCIENCES	34K
	MATHEMATICS	21K

WHAT IS THE NUMBER OF MIXED SCHOOL STUDENTS IN EACH CITY?

```
--2. Filter only the mixed schools
filtered = FILTER Egyption_school BY (mixed_school=='Mixed');

--3. group the data we need to show and Get the count of students for each city per mixed school
grouped_data = GROUP filtered BY (city, mixed_school);

mixed_count_data = FOREACH grouped_data GENERATE FLATTEN(group) AS (city, mixed_school),
COUNT(filtered.desk_no) AS mixed_count;

--5. Order the result in descending order
order_by_data = ORDER mixed_count_data BY mixed_count DESC;
DUMP order_by_data;
```

TOP 3

	CITY	STUDENTS
	DAKAHLIA	17K
	SHARQIA	13K
	MINYA	11K

FOR PEOPLE WHO DID ONLY THE FIRST ATTEMPT, HOW MANY OF THEM FAILED?


```
--2. Filter only the first attempt and failed status
filtered = FILTER Egyption_school BY (no_of_attempts=='First attempt only') AND (status=='Failed');

--3. group the data we need to show and Get the count of students based on filter
grouped_data = GROUP filtered BY (gender, no_of_attempts, status);

failed_count_data = FOREACH grouped_data GENERATE FLATTEN(group) AS (gender, no_of_attempts, status),
COUNT(filtered.desk_no) AS failed_count;

--5. Order the result in descending order
order_by_data = ORDER failed_count_data BY failed_count DESC;

DUMP order_by_data;
```

WHAT IS THE LOWEST AND HIGHEST SCORE IN ARABIC SUBJECT?

```
--filter the dataset where status is Passed:
Filter_Passed = FILTER Egyption_school BY (status == 'Passed');

max_arabic = FOREACH (GROUP Filter_Passed ALL ) GENERATE MAX(Egyption_school.arabic) as
max_arabic;
-- Find the minimum score in the Arabic subject
min_arabic = FOREACH (GROUP Filter_Passed ALL ) GENERATE MIN(Egyption_school.arabic) as
min_arabic;
```

العربية

★★★★★ 80

★★★★★ 0

WHAT IS THE AVERAGE SCORE PERCENTAGE FOR BOYS COMPARED TO GIRLS WITH SPECIAL NEEDS?

```
--filter the dataset where status is Passed and the school is for the blind:
Filter_Passed = FILTER Egyption_school BY (status == 'Passed') and (school_type == 'School for the
Blind');

--The average percentage of scores:
student_group_gender = Group Filter_Passed by gender;
student_average = foreach student_group_gender GENERATE group AS Gender,
AVG(Filter_Passed.percentage)AS average_persentage_by_gender;
Dump student_average;
```

Girls typically outperform boys in humanities, languages and reading tests, while boys do better in maths

But when grades are awarded by teachers, girls do better in all subjects, Sky news

79.7%

84.2%

HOW MANY STUDENTS FAILED IN ALL THEIR COURSES?


```
-- Select only rows where no_of_failed_courses is equal to 10
filtered_data = FILTER Egyption_school BY no_of_failed_courses == 10;
-- Count the number of rows in the filtered dataset
failed_10_courses = FOREACH (GROUP filtered_data ALL) GENERATE
COUNT(filtered_data);
dump failed_10_courses;
```

620

HOW MANY SCHOOL FOR SPECIAL NEEDS PER CITY?

```
-- Filter only the mixed schools
filtered = FILTER Egyption_school BY (school_type == 'School for the Blind');;
-- group the data we need to show
grouped_data = GROUP filtered BY (city, school_type);
-- Get the count of students for each city per school_type
blind_count_data = FOREACH grouped_data GENERATE FLATTEN(group) AS (city, school_type),
COUNT(filtered.desk_no) AS blind_count;
-- Order the result in descending order
order_by_data = ORDER blind_count_data BY blind_count DESC;
-- Describe and display the results
DESCRIBE order_by_data;
DUMP order_by_data;
```

TOP 3

	CITY	SCHOOL
	GIZA	24
	MINYA	19
	QALYUBIYYA	16

FOR PEOPLE WHO DID ONLY THE FIRST ATTEMPT, HOW MANY OF THEM FAILED AND WERE STUDENTS WITH SPECIAL NEEDS?

```
-- Filter only the first attempt, failed status and blind school
filtered = FILTER Egyption_school BY (no_of_attempts=='First attempt only') AND (status=='Failed') AND
(school_type == 'School for the Blind');
-- group the data we need to show
grouped_data = GROUP filtered BY (gender, no_of_attempts, status,school_type);
-- Get the count of students for each city per mixed school
failed_count_data = FOREACH grouped_data GENERATE FLATTEN(group) AS (gender, no_of_attempts,
status,school_type), COUNT(filtered.desk_no) AS failed_count;
-- Order the result in descending order
order_by_data = ORDER failed_count_data BY failed_count DESC;
-- Describe and display the results
DUMP order_by_data;
```

2

0

WHAT IS THE AVERAGE RATE PER BRANCH FOR MIXED SCHOOL STUDENTS?

```
--filter the dataset where status is Passed and mixed schools:
Filter_Passed = FILTER Egyption_school BY (status == 'Passed') and (mixed_school=='Mixed');

--The average percentage of scores:
student_group_branch = Group Filter_Passed by branch;
student_average = foreach student_group_branch GENERATE group AS branch,
AVG(Filter_Passed.percentage)AS average_persentage_by_branch;
Dump student_average;
```

	BRANCH	STUDENTS
	LITERATURE	66.20%
	HEALTH SCIENCES	73.80%
	MATHEMATICS	74.19%

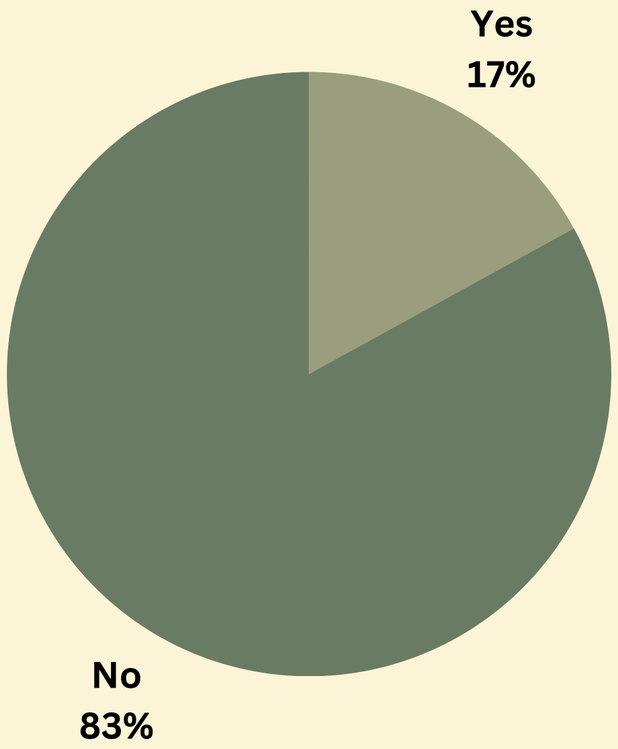
HOW MANY STUDENTS FAILED IN ALL THEIR COURSES IN MIXED SCHOOLS?

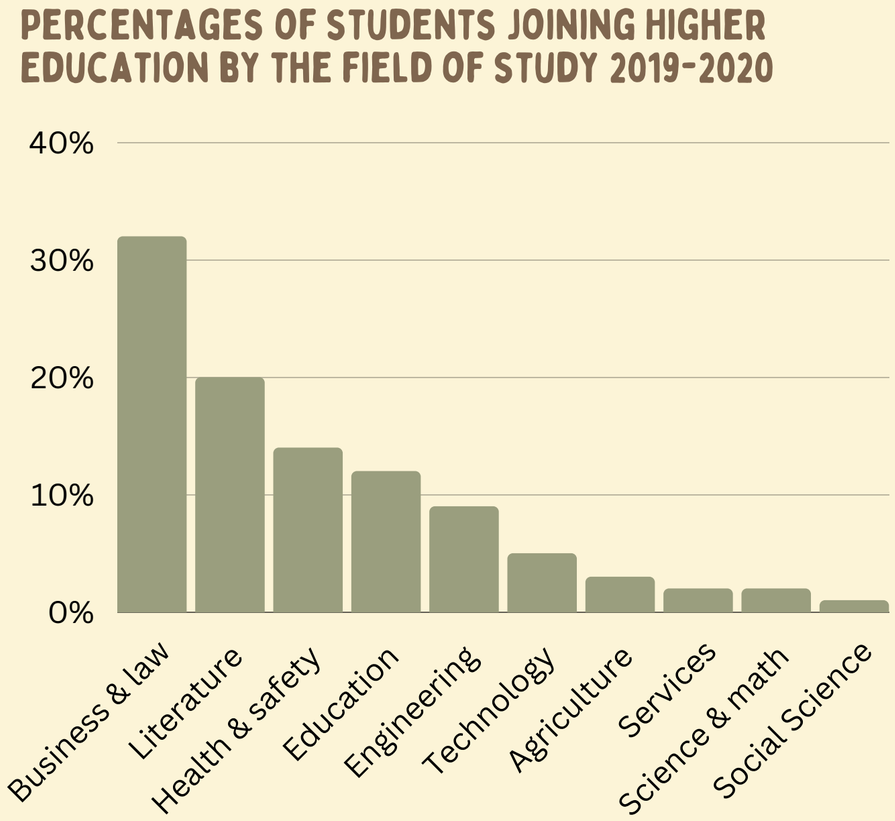

```
--filter the dataset where no_of_failed_courses is 10 :
Filter_failed = FILTER Egyption_school BY (mixed_school=='Mixed') and (no_of_failed_courses == 10);
;
-- Count the number of students
failed_10_courses_mixed = FOREACH (GROUP Filter_failed ALL) GENERATE COUNT(Filter_failed);

-- show the result
dump failed_10_courses_mixed;
```

84

FINDINGS

CAN THE STUDENT JOIN THE UNIVERSITY, MATCHING HIS CURRENT FIELD OF STUDY?





WHY DO YOU THINK THIS IS AN ISSUE AND HOW CAN WE GET OVER IT?

RECOMMENDATIONS

CHANGING THE APPROACH OF LEARNING FROM MEMORIZATION TO ONE THAT EMPHASIZES THE DEVELOPMENT OF SELF-RELIANCE, IMAGINATION, AND CRITICAL THINKING

CUMULATIVE EVALUATION OF ASSESSING SKILLS RATHER THAN MEMORIZATION, THE PRACTICE OF EVALUATION BASED ON RISKY FINAL EXAMS IS MINIMIZED

CHANGING THE CENTRAL COORDINATION SYSTEMS, GIVING UNIVERSITIES THE CHOICE OF ADMISSION ON THE CAPABILITIES OF EACH UNIVERSITY COLLEGE, AS AN ALTERNATIVE FOR STUDENTS

PROVIDING SCHOLARSHIPS BASED ON THE SKILLS OR ACADEMIC ACCOMPLISHMENTS STUDENT MAINTAINS, RATHER THAN TIEING IT WITH A SPECIFICSCORE

THANK YOU

Do you have any questions ?

HOPEFULLY, WE WILL CROSS PATHS AT THE FUTURE

BEST REGARDS,
ALL CLIMATES NINJAS FOR THE LAST TIME