***Proof of Concept***

***On***

**Title: Analysis of Baby names**

**Submitted for the requirement of**

**Big Data Engineering course**

BACHELOR OF ENGINEERING

**COMPUTER SCIENCE & ENGINEERING**

**(Big Data and Analytics)**

**CST-357**

**Semester-6**

****

**Submitted to: Submitted by:**

Ms. Gurpreet Kaur

Project Supervisor (17BCS3943)

Raghav mandowara

# 

# ACKNOWLEDGEMENT

We would like to express our deepest appreciation to all those who provided us the possibility to complete this report.  A special gratitude we give to our 6th semester B.D.E project supervisor, Ms. Gurpreet Kaur,whose contribution in stimulating suggestions and encouragement, helped us to coordinate our project and especially in writing this report.Furthermore, we would also like to acknowledge with much appreciation her crucial role, in giving the permission to use all required equipment and the necessary materials to complete the task ‘Analysis of Baby names’ using Apache spark, python programming.

## OVERVIEW

AMCAT stands for Aspiring Minds Computer Adaptive Test. AMCAT is a test which helps the candidate to evaluate a range of their professional skills which includes aptitude skills, reasoning skills, technical skills and mental ability skills. AMCAT can be a really good way for the students of low tier colleges or the students which couldn’t get placed during the Campus Placement Drive to get a job. But AMCAT is not limited to that, this test is often used by reputed colleges like IIT’s, BIT and VIT for the recruitment process. Various companies which visit these colleges for recruitment use AMCAT score for scrutinizing the students for the first round. About 700+ companies including big companies like Accenture, HCL, and Amazon recognize AMCAT to recruit their candidates for the first round.

## RATIONALE

Educational institutions are interested in methods for providing effective learning to students, recognizing their achievements as well shortcomings. Given the recent trend of Big Data in the educational system for improving student results by using the answers given to questions in competitive exams, the present study aims to systematically analyze the individual performance of students using the concrete analytical approach of data mining with Cloudera QuickStart VM’s Hadoop Distribution.

## OBJECTIVES

1. To

2. To provide optimal guidance to the students,serving their educational development needs.

3. To upgrade education standards and ensure the quality education.

4. To know the weaker sections of students and work on it for better results.

**COLUMNS AND DATA TYPES:**

Sex Number

Name Chararray/String

Year Number

Frequency Number

**PROBLEM STATEMENTS:**

<https://perso.telecom-paristech.fr/eagan/class/igr204/datasets>

1. **Finding the number of students appeared for Amcat from each branch.**
2. **Summarizing the data using aggregated variables i.e. minimum marks, maximum marks, mean marks and standard deviation.**
3. **Finding students who scored more than 500+ in English,Quantitative Ability and Logical i.e. ELQ combined.**
4. **Finding topscorers from each branch.**
5. **Finding students who scored more than 50 in the subjectAutomata from the branches BE-CSE,CSE IBM (BD) and CSE IBM(IS).**
6. **Stratifying branch-wise, how many students appeared for Amcat exam.**
7. **Listing the students who score less than 400 in ELQ combined.**
8. **Checking branch-wise and gender-wise distribution of student marks.**
9. **Checking year-wise and branch-wise distribution of student marks.**

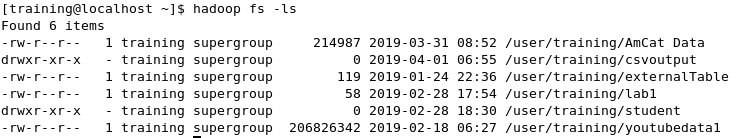
**SHELL SCRIPT:**

**PIG:**

**Loading data in HDFS:**

hadoop fs -copyFromLocal '/home/training/Desktop/Amcat Data.txt' '/user/training/Amcat Data'

Hadoop fs –ls

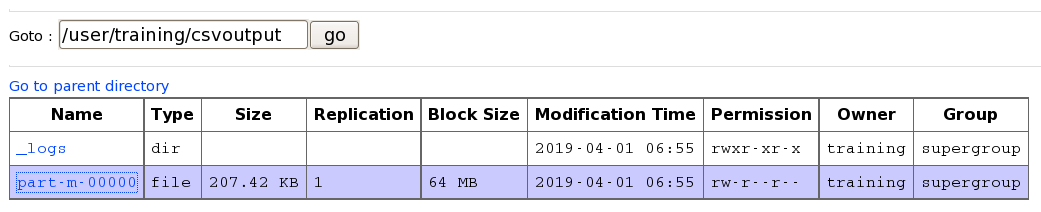


t1 = load '/user/training/AmCat Data' using PigStorage('\t') as

(universityRollNo:chararray, firstName:chararray, lastName:chararray, Branch:chararray, emailID:chararray, mobileNumber:long, English:int, Logical:int, Quantitative\_Ability:int, Sum\_of\_ELQ:int);

**STORE T1 AS CSV:**

store t1 into '/user/training/csvoutput' using PigStorage(',');





**PROBLEM STATEMENT 1: Number of students appeared for AmCat from each Branch**

**Code:**

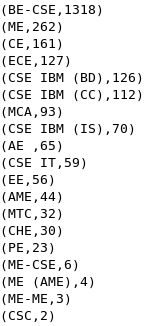
t2 = foreach t1 generate universityRollNo,Branch;

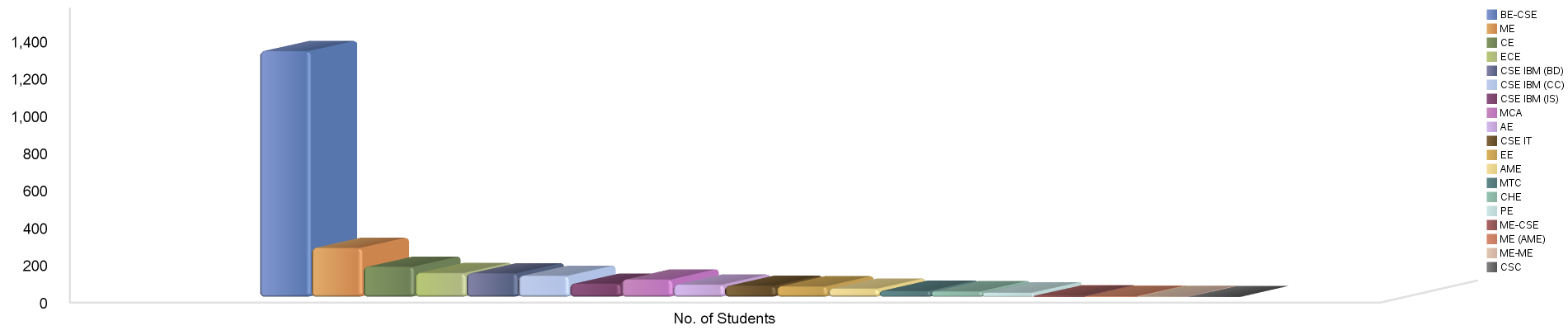
t3 = group t2 by Branch;

t4 = foreach t3 generate group,COUNT(t2.universityRollNo);

t5 = order t4 by $1 desc;

dump t5;





**PROBLEM STATEMENT 2: More than 500+ in English, Quantitative Ability and Logical and ELQ combined**

**ENGLISH:**

z1 = filter t1 by English>500;

z2 = foreach z1 generate 1 as one;

z3 = group z2 all;

z4 = foreach z3 generate SUM(z2.one) as count;

dump z4;

**OUTPUT:**

1380

**QUANTITATIVE ABILITY:**

q1 = filter t1 by Quantitative\_Ability>500;

q2 = foreach q1 generate 1 as one;

q3 = group q2 all;

q4 = foreach q3 generate SUM(q2.one) as count;

dump q4;

**OUTPUT:**

1074

**LOGICAL:**

l1 = filter t1 by Logical>500;

l2 = foreach l1 generate 1 as one;

l3 = group l2 all;

l4 = foreach l3 generate SUM(l2.one) as count;

gdump l4;

**OUTPUT:**

987

**ELQ COMBINED:**

a1 = filter t1 by Logical>500 and English>500 and Quantitative\_Ability>500;

a2 = foreach a1 generate 1 as one;

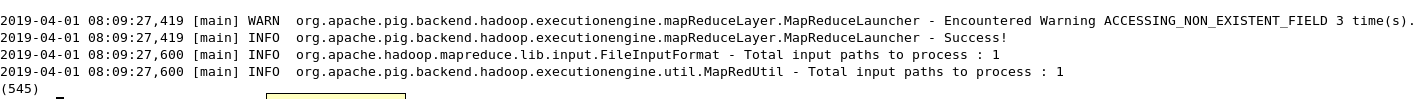
a3 = group a2 all;

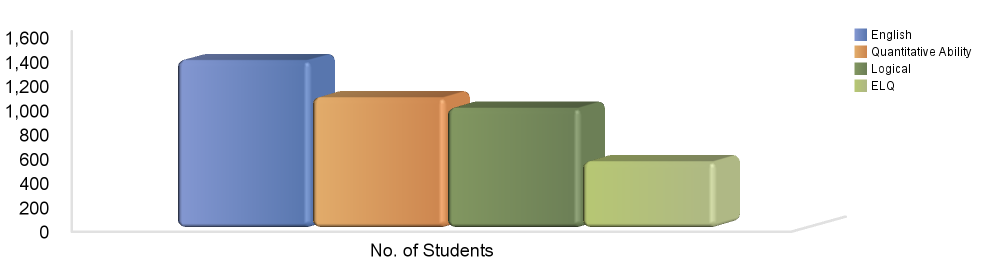
a4 = foreach a3 generate SUM(a2.one) as count;

dump a4;

**OUTPUT:**

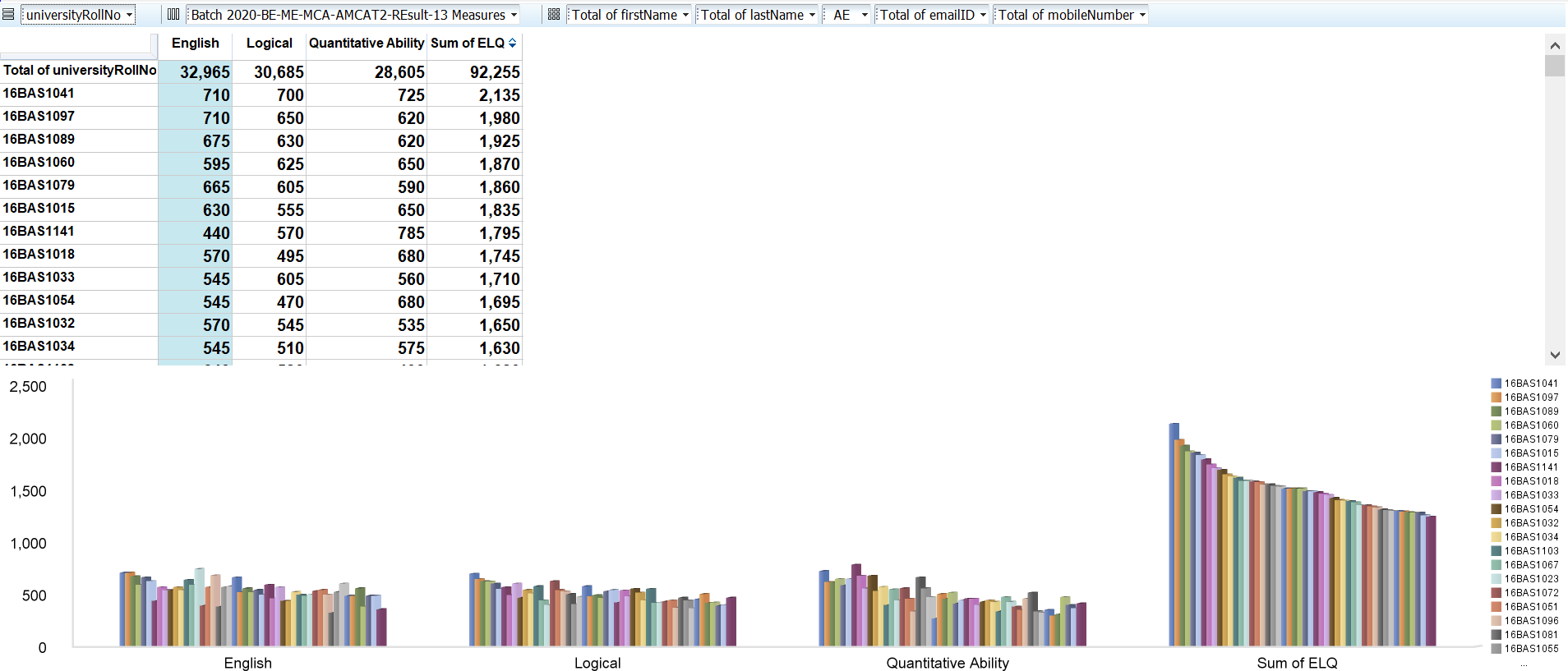
545



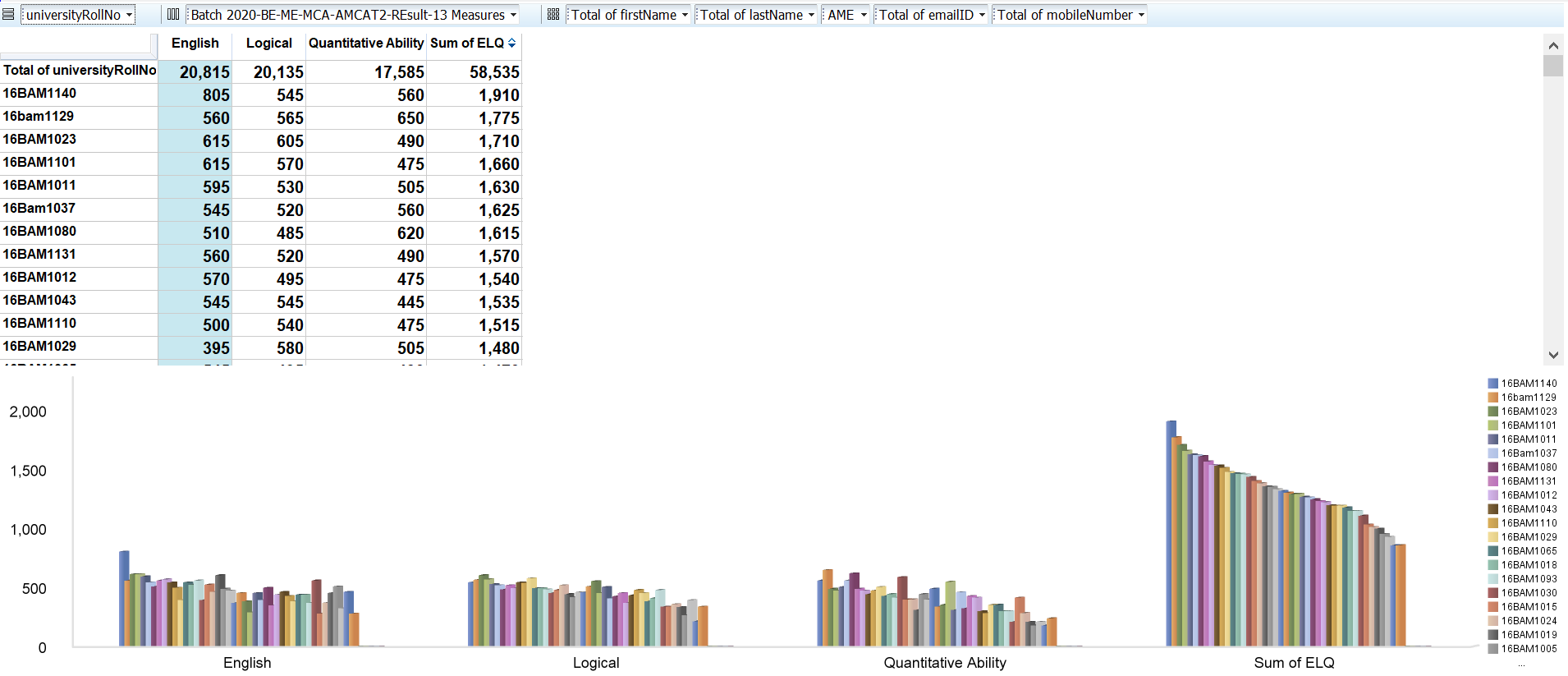


**PROBLEM STATEMENT 3: Top Scorers from each branch.**

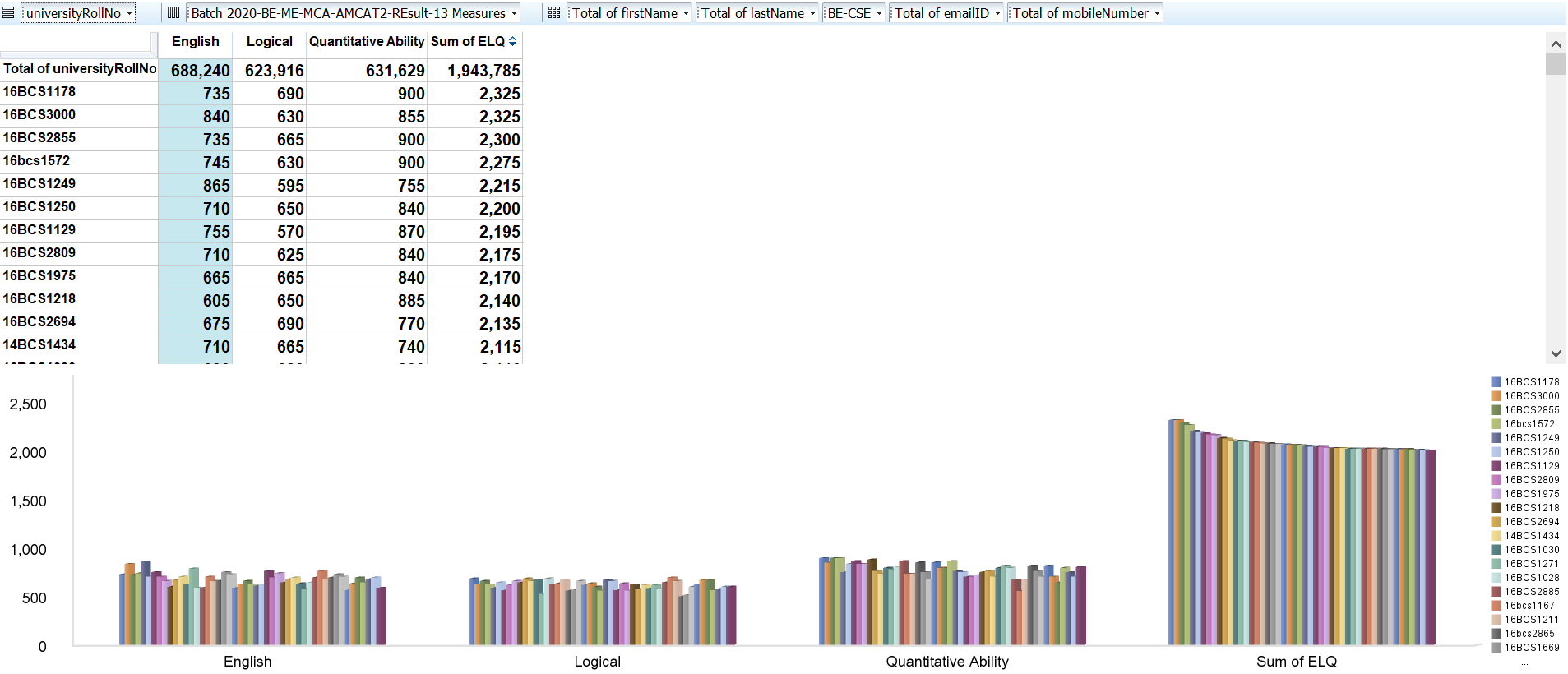
1. **AE:**



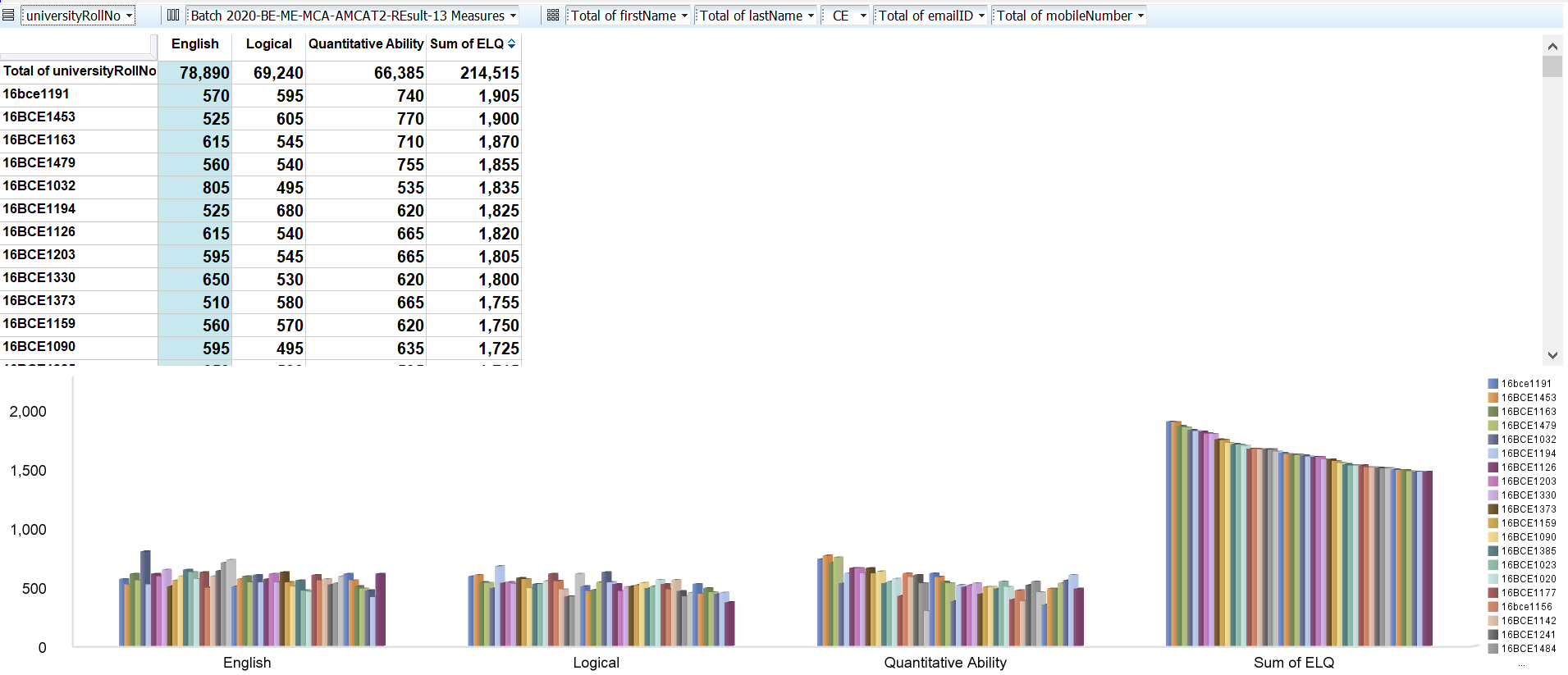
1. **AME:**



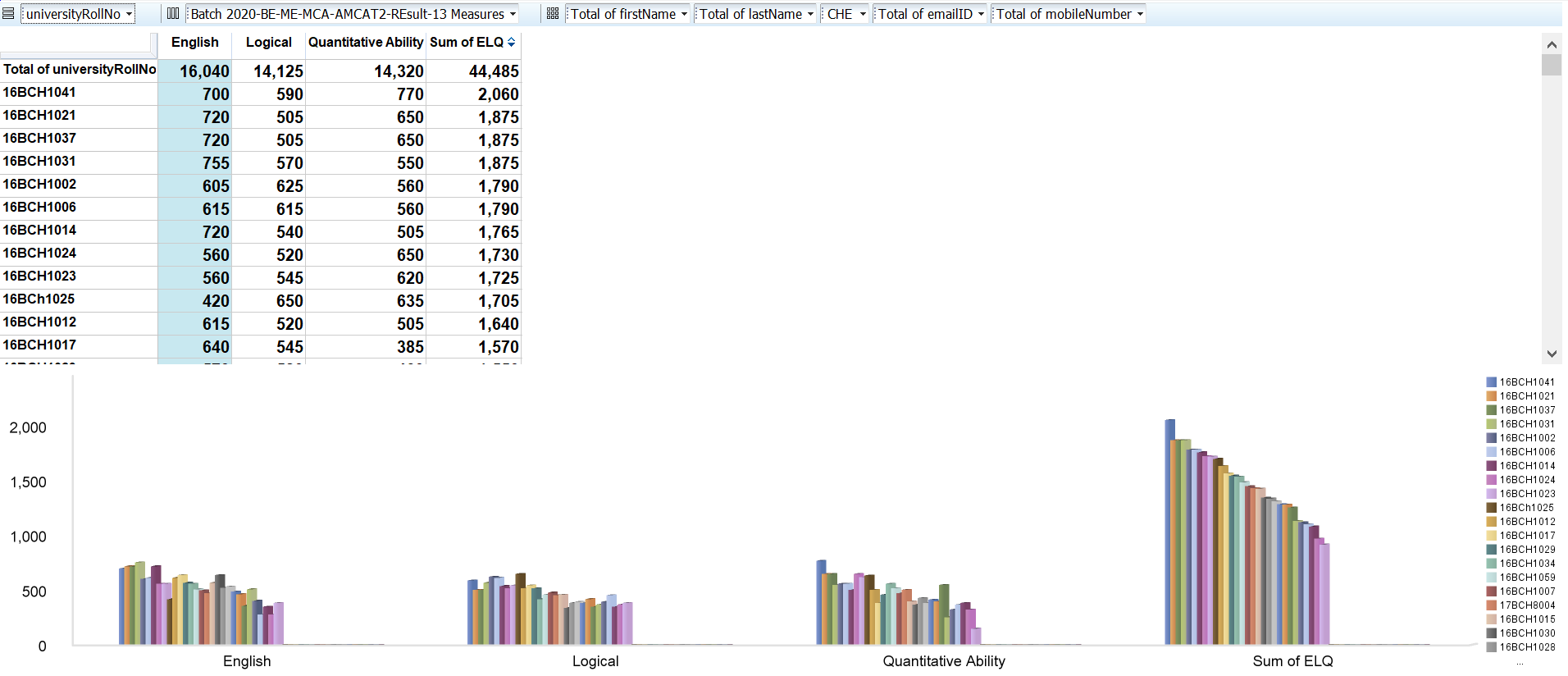
1. **BE-CSE:**



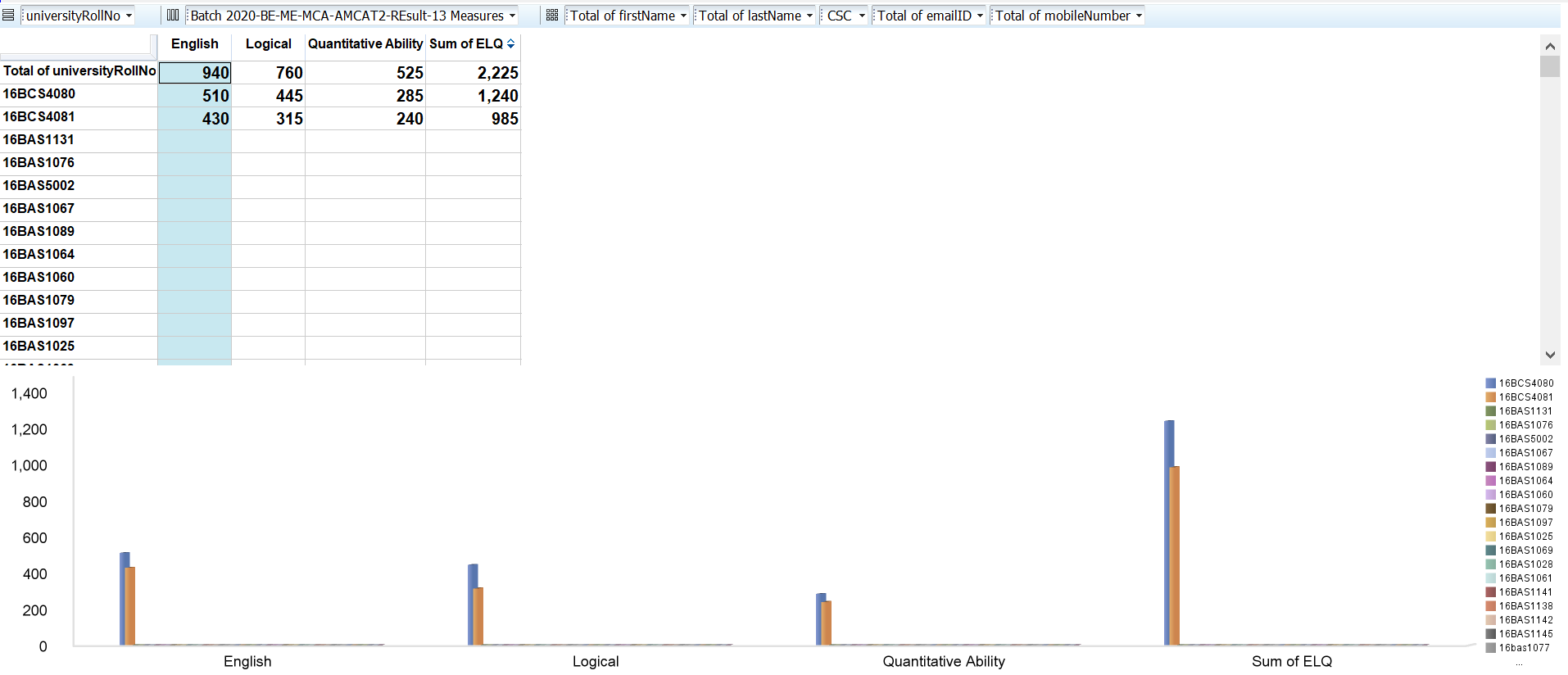
1. **CE:**



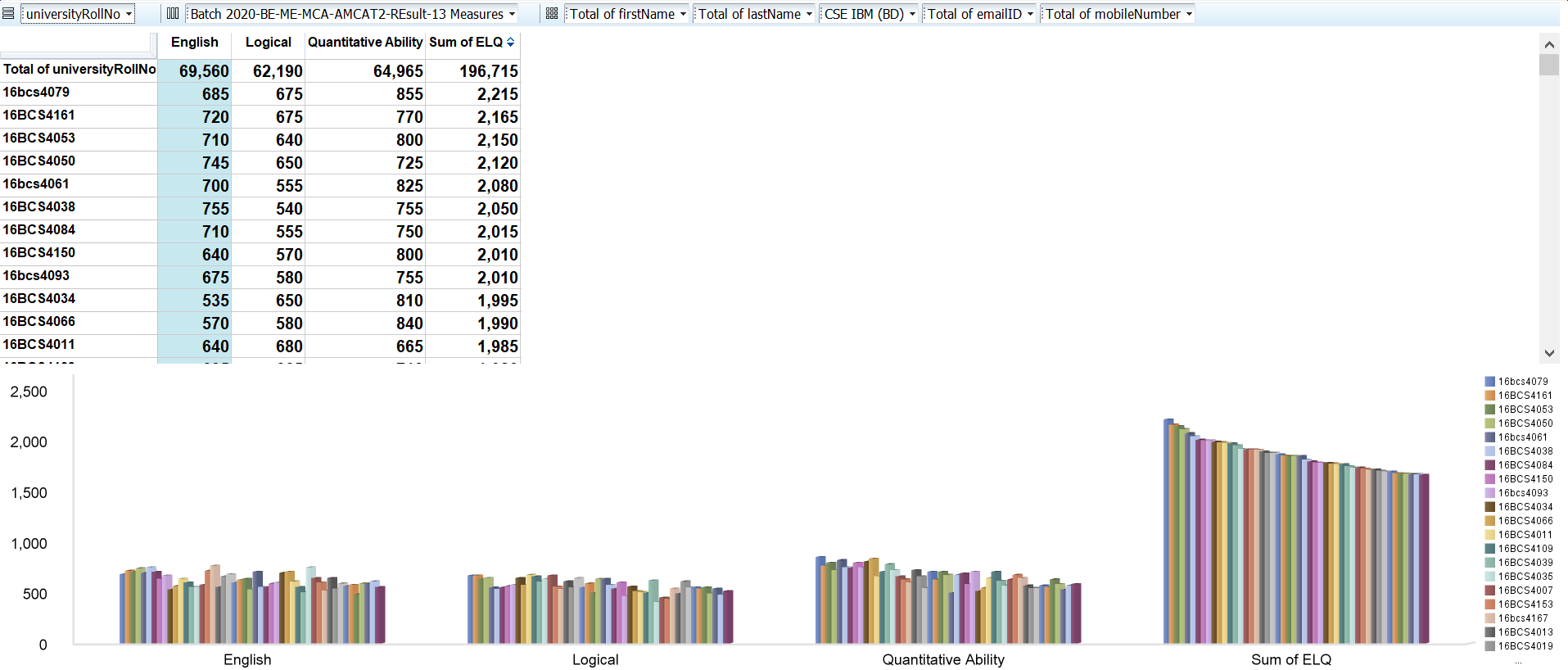
1. **CHE:**



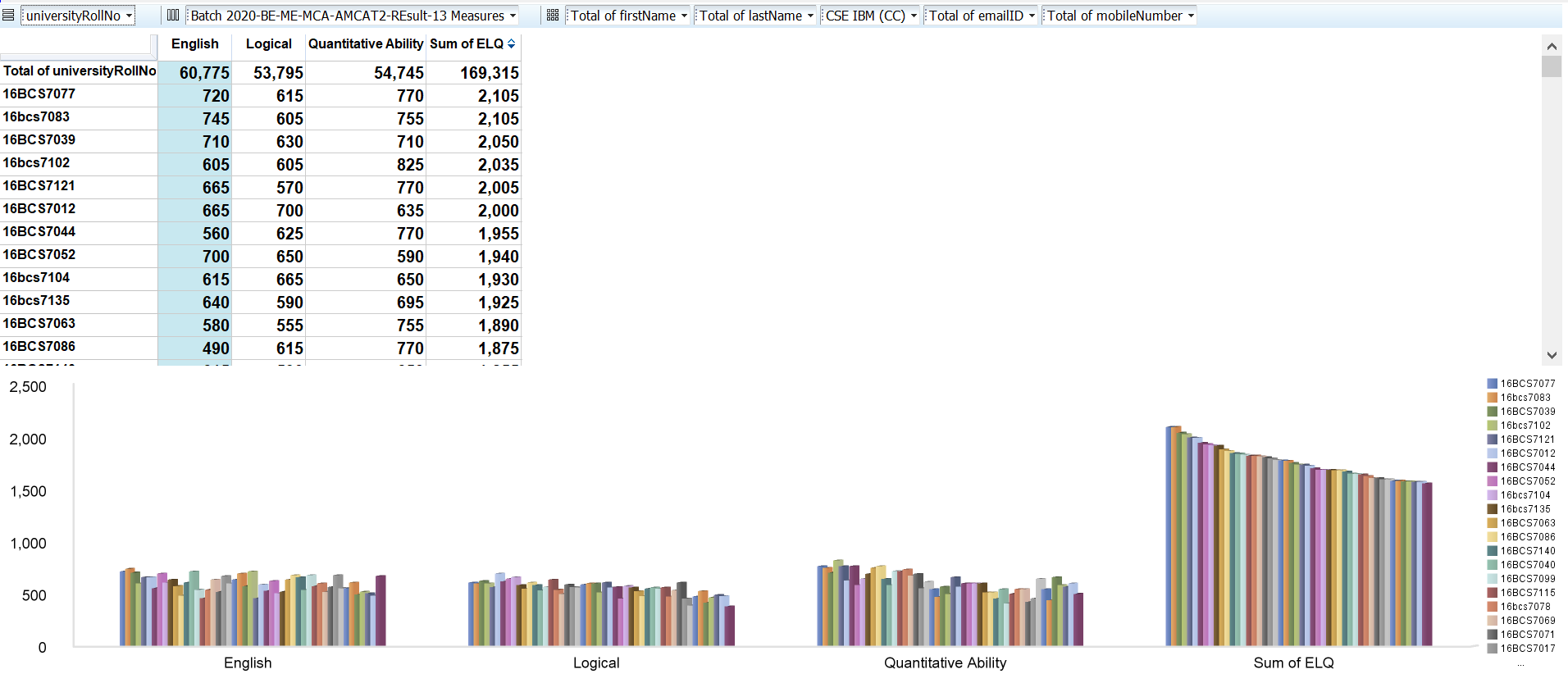
1. **CSC:**



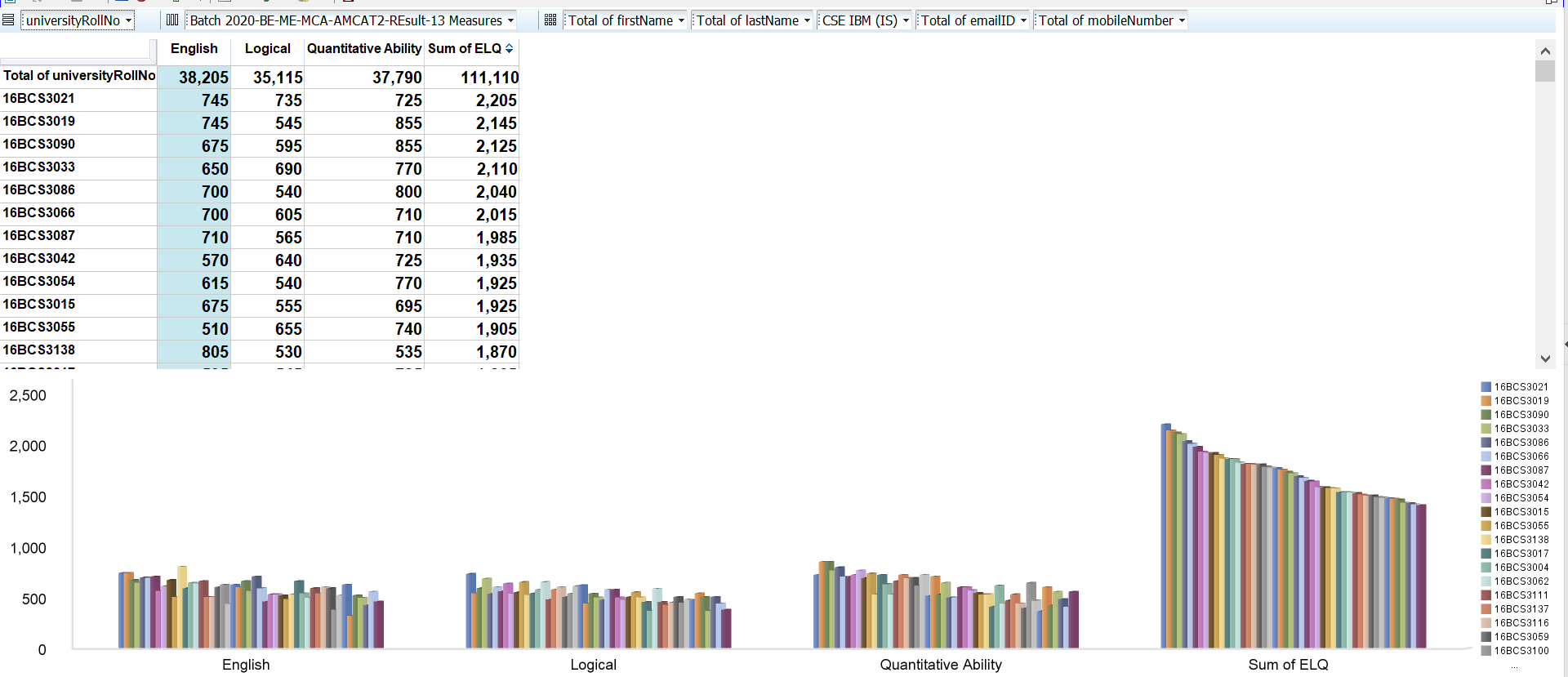
1. **CSE IBM (BD):**



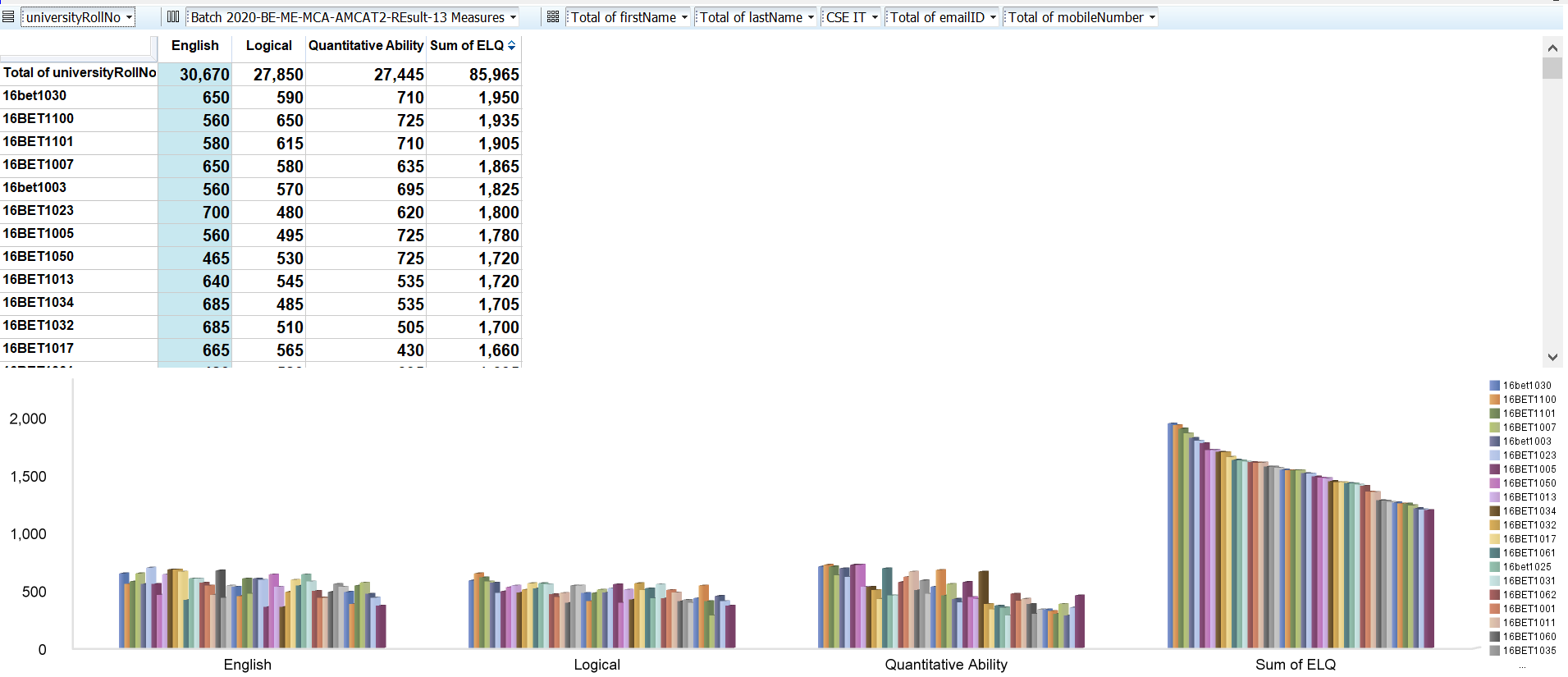
1. **CSE IBM (CC):**



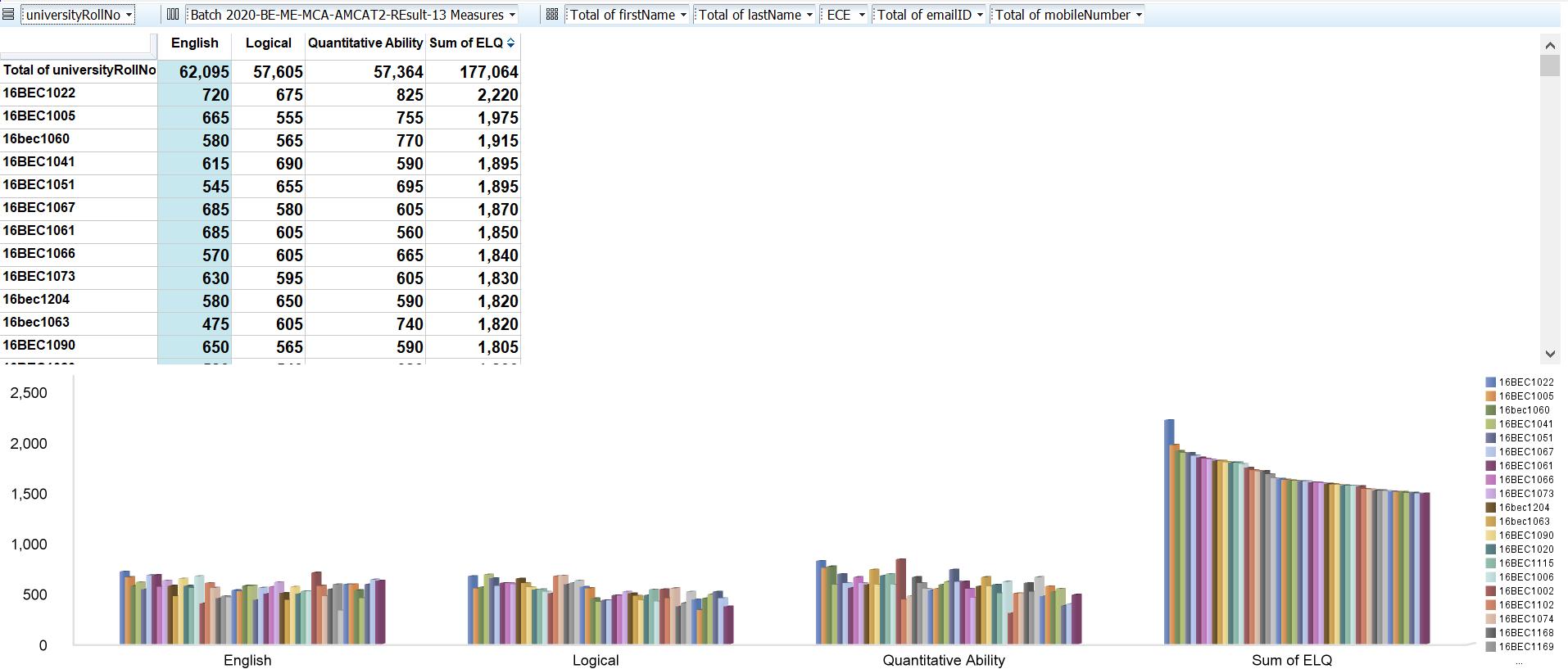
1. **CSE IBM (IS):**



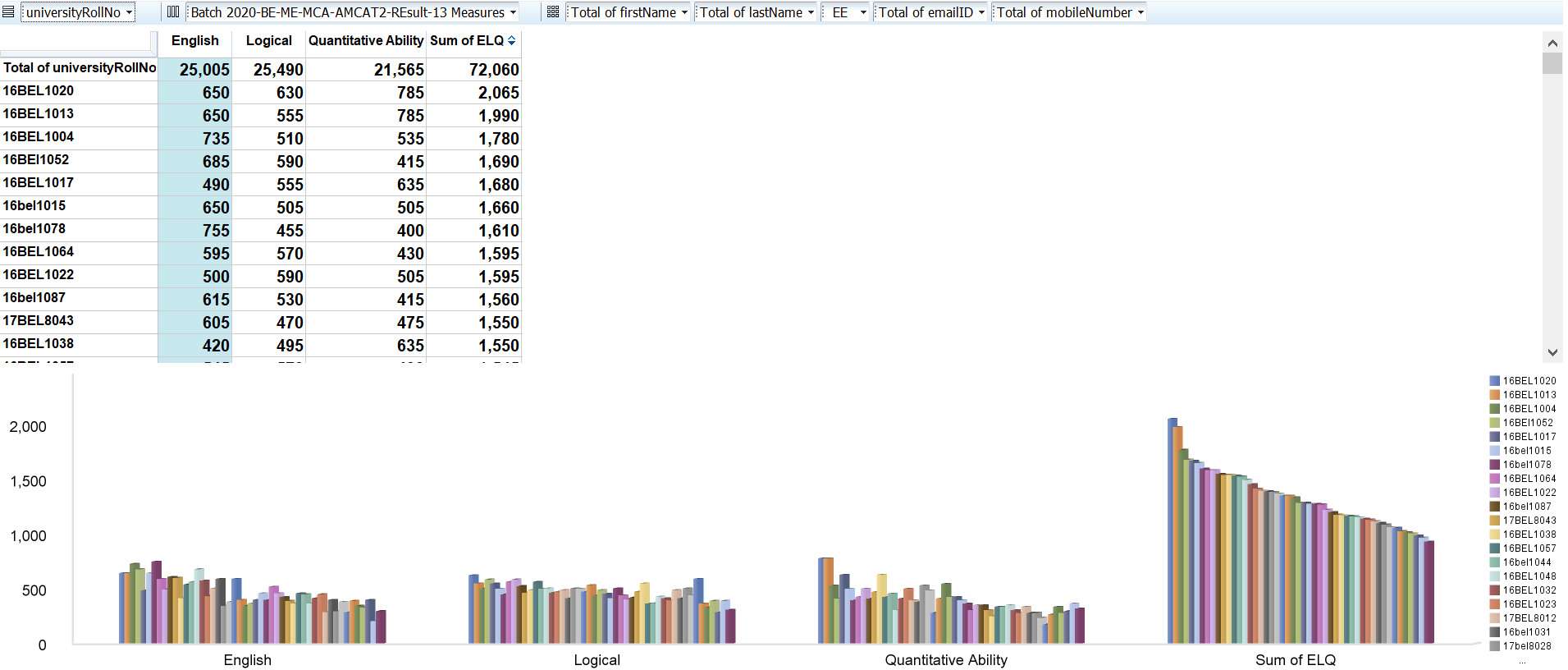
1. **CSE IT:**



1. **ECE:**



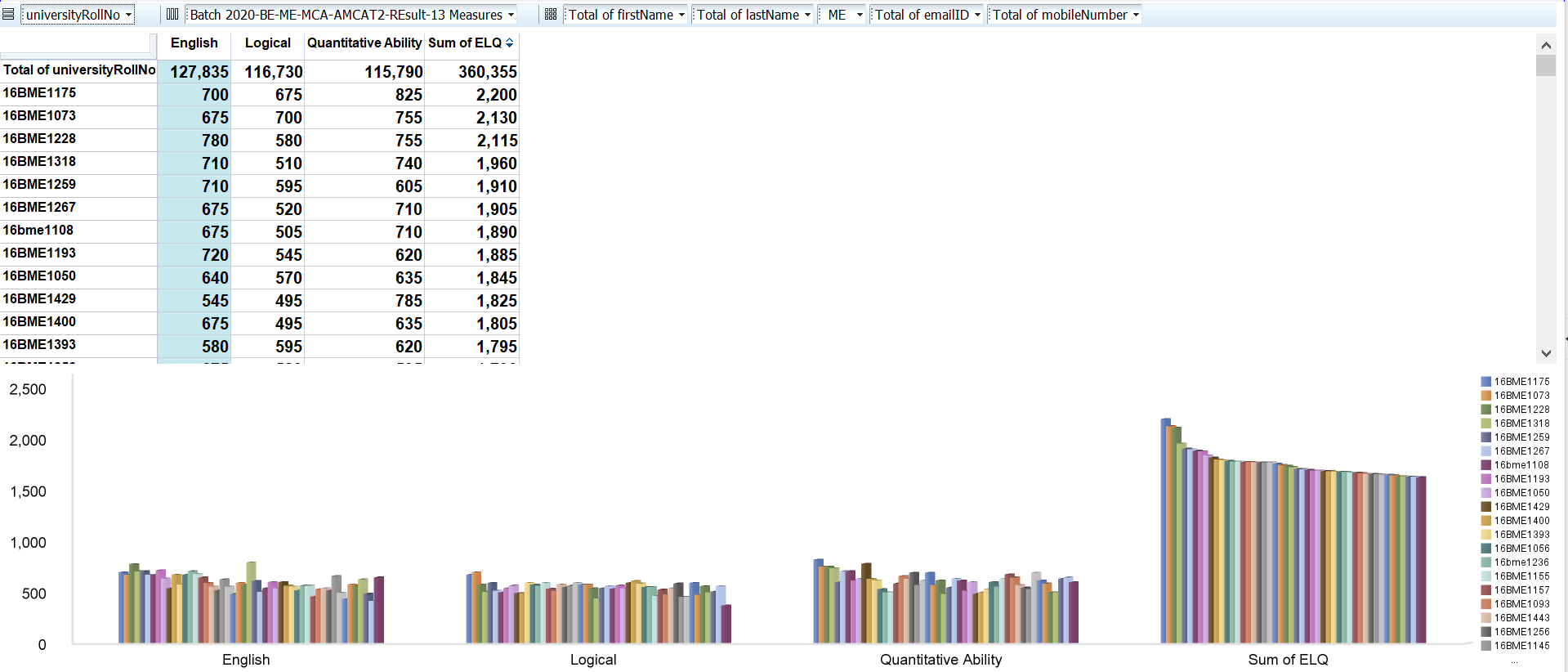
1. **EE:**



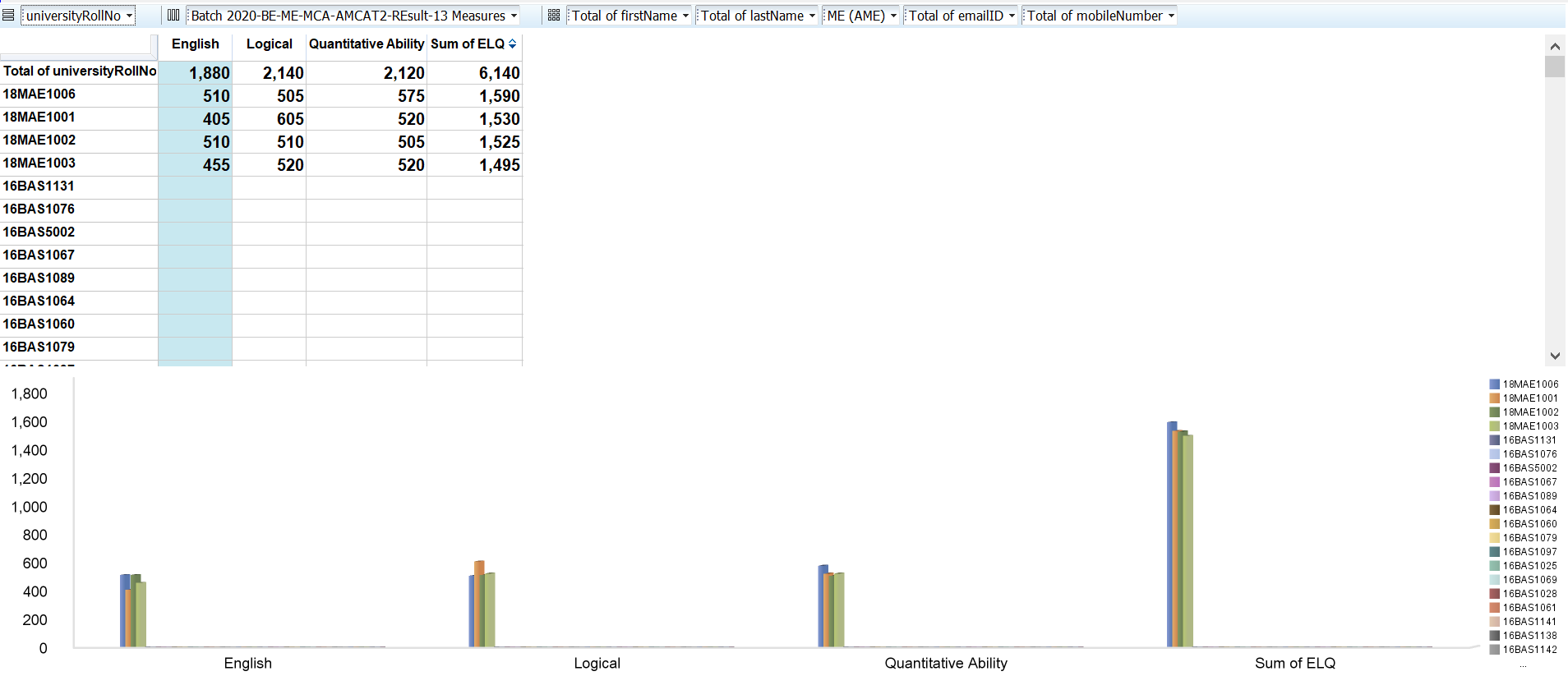
1. **MCA:**



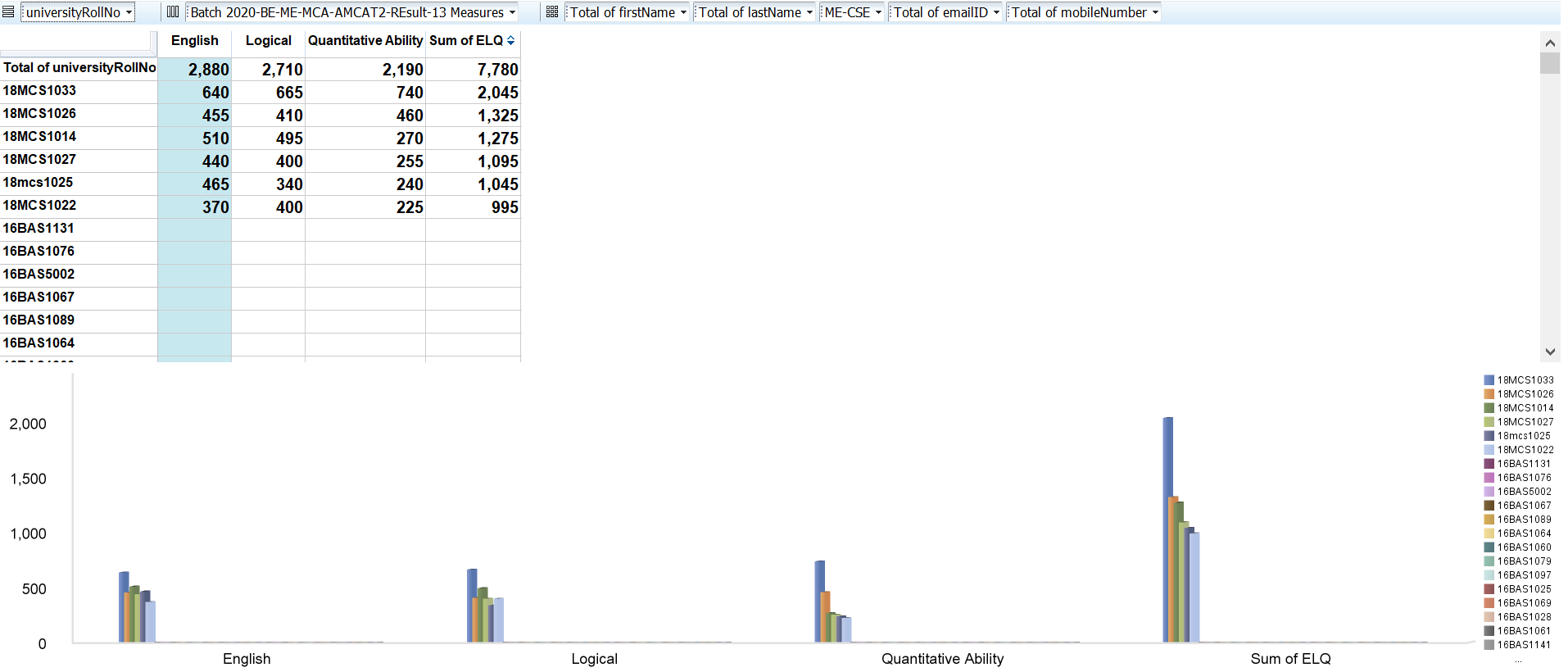
1. **ME:**



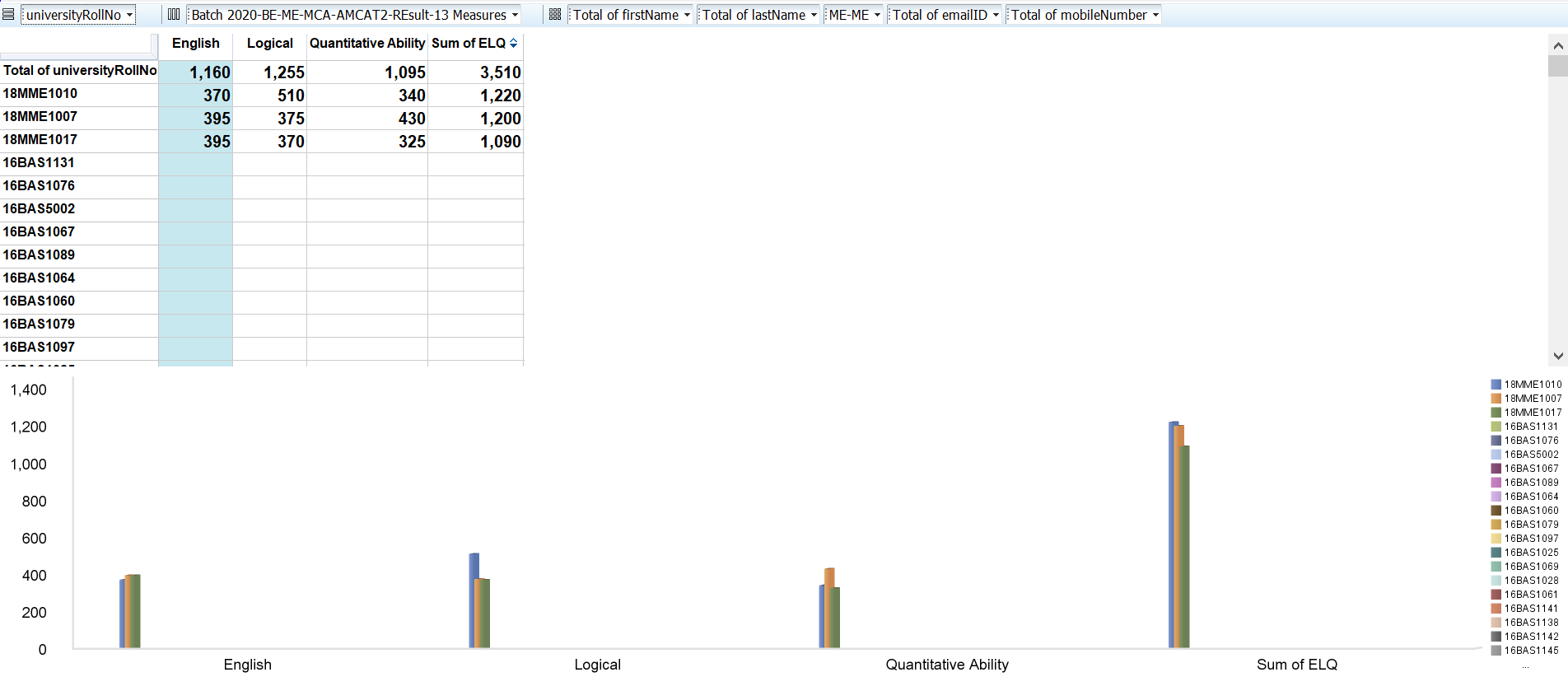
1. **ME (AME):**



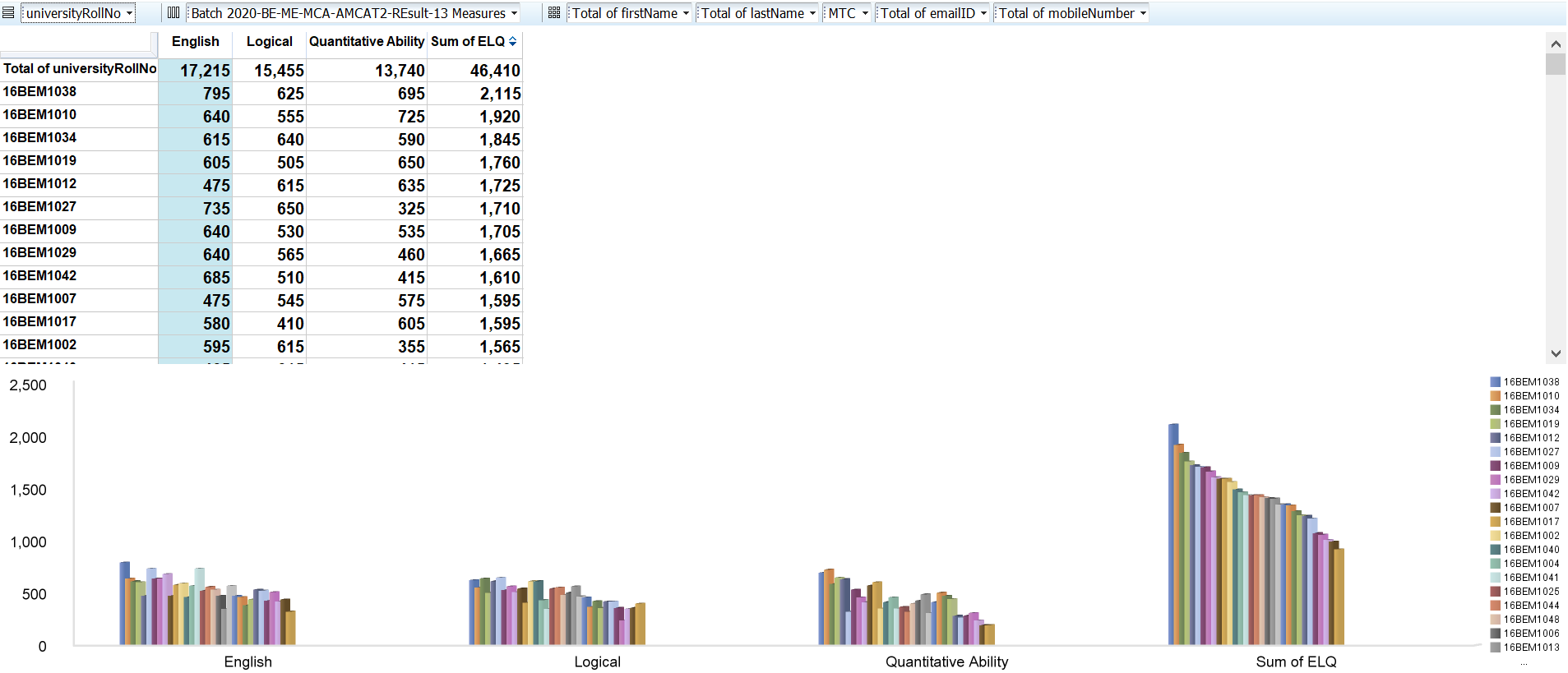
1. **ME-CSE:**



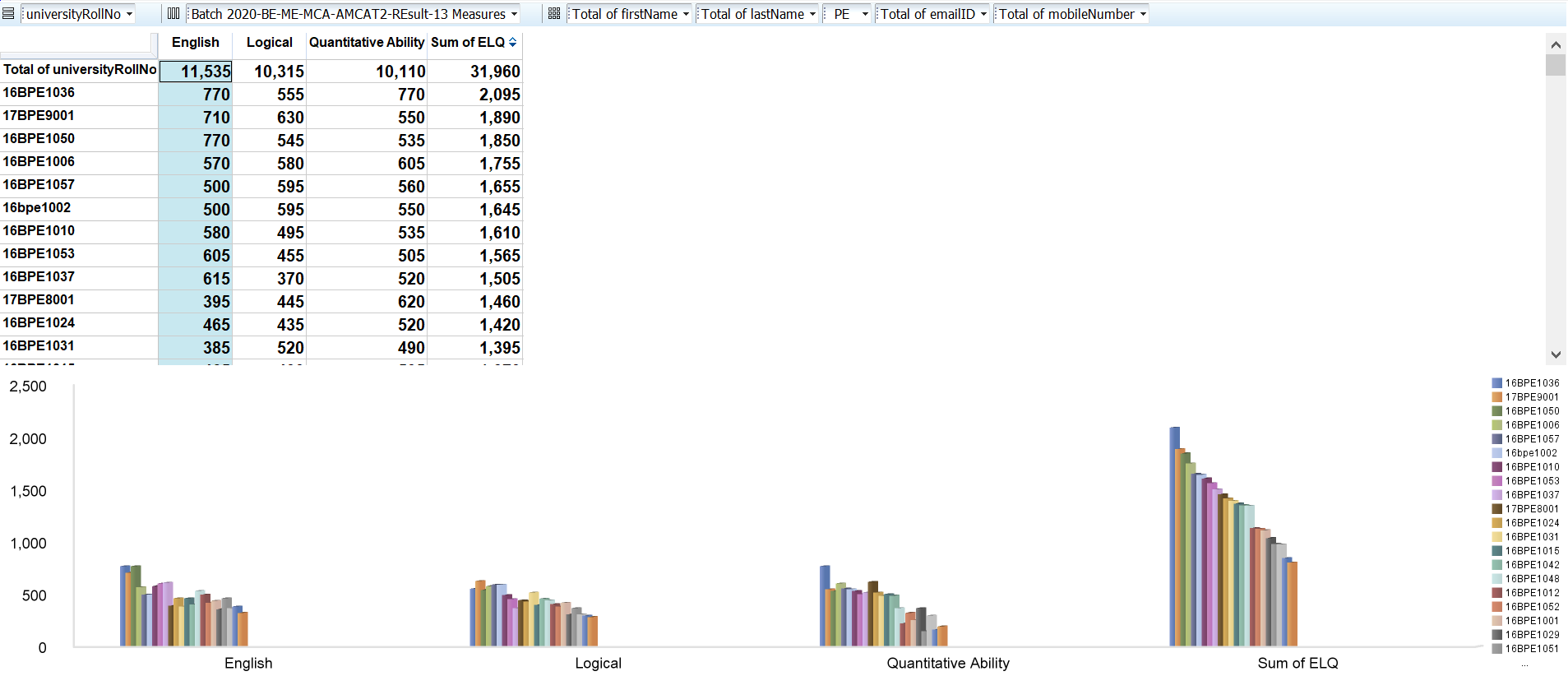
1. **ME-ME:**



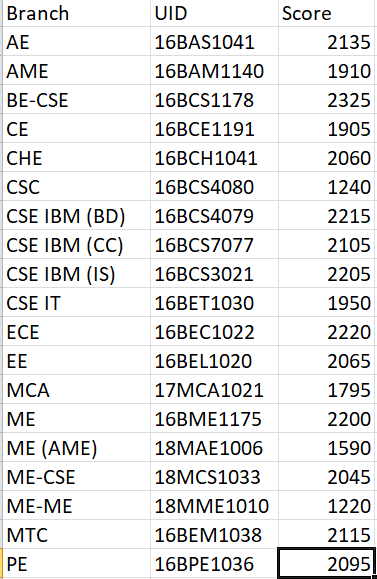
1. **MTC:**

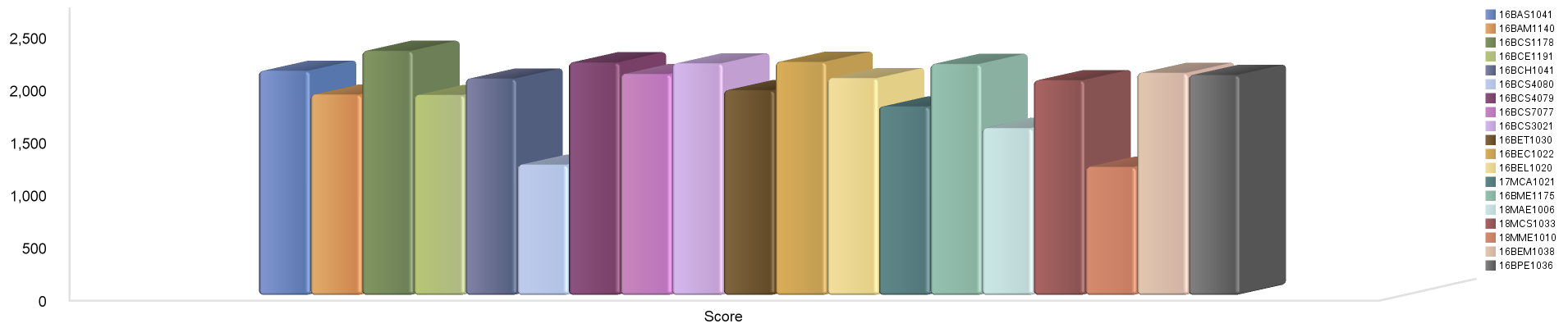


1. **PE:**



**OUTPUT:**

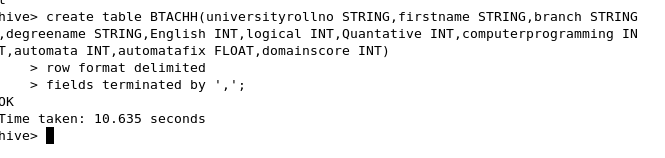




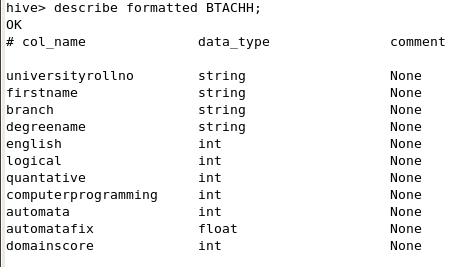
**HIVE:**

**PROBLEM STATEMENT 4: Finding students who scored more than 50 in the subject Automata from the branches BE-CSE, CSE IBM (BD) and CSE IBM(IS).**

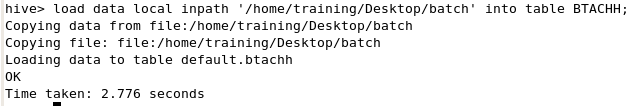
1. **Create table BTACHH:**



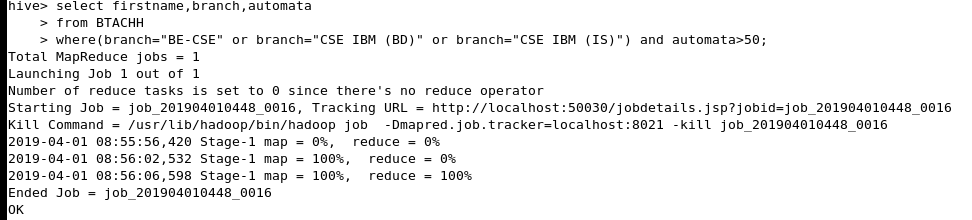
**Data type:**



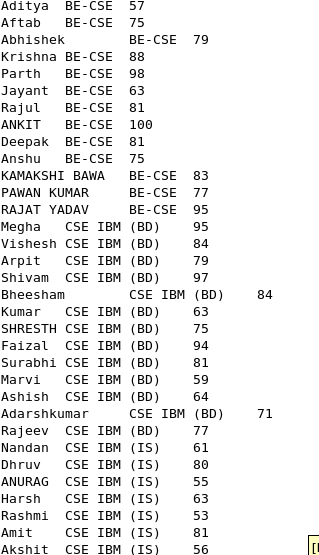
**Load data:**



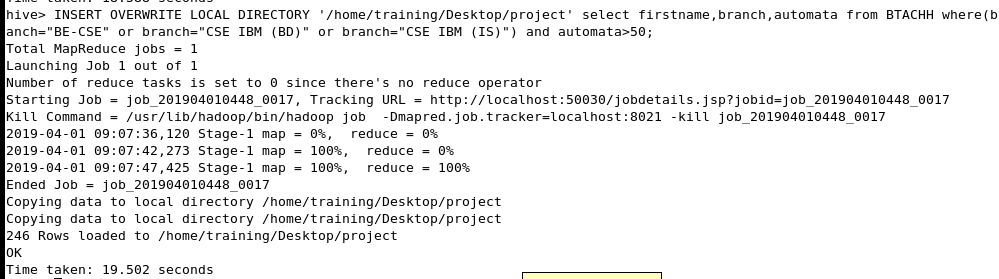
**Query:**

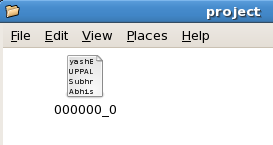


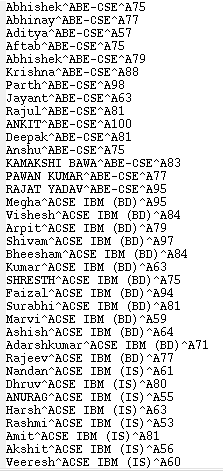
**Output:**



**Csv file:**

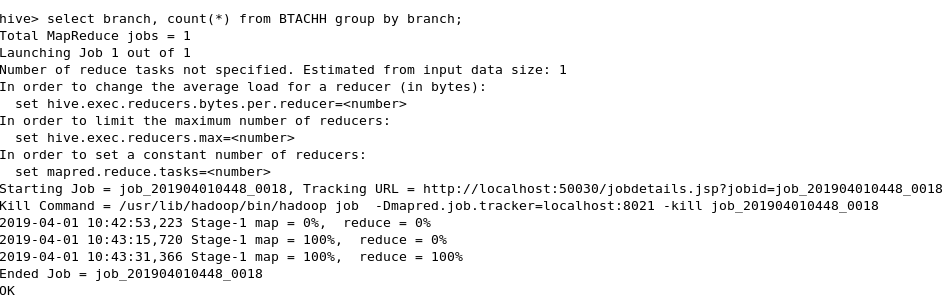




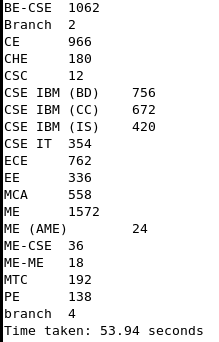


**PROBLEM STATEMENT 5: Stratifying branch-wise, how many students appeared for Amcat exam.**

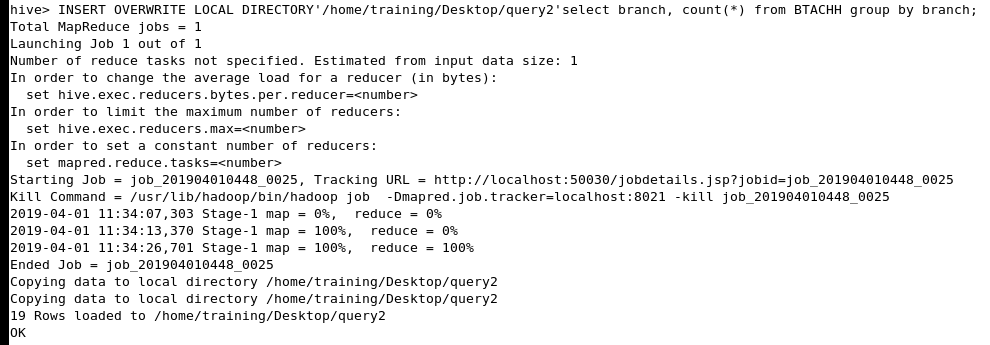
**Query:**

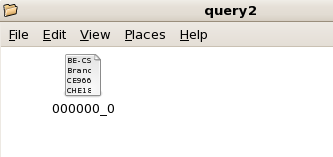


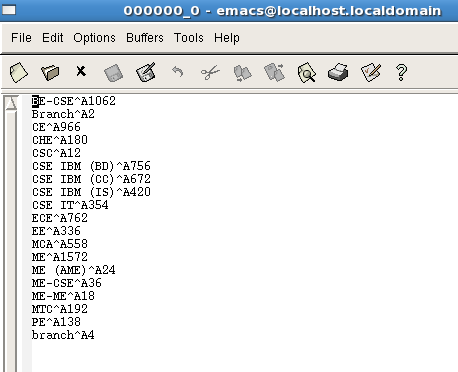
**Output:**



**Csv:**



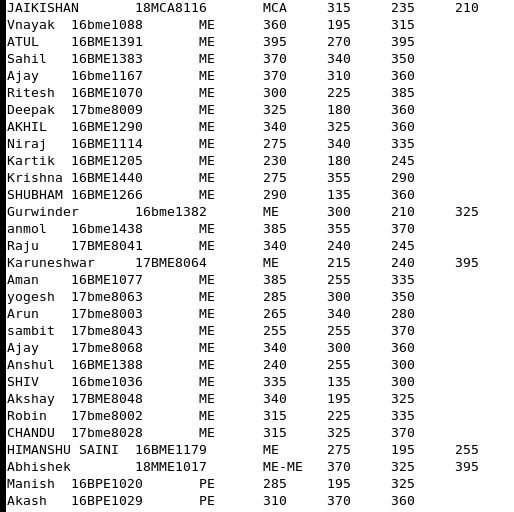
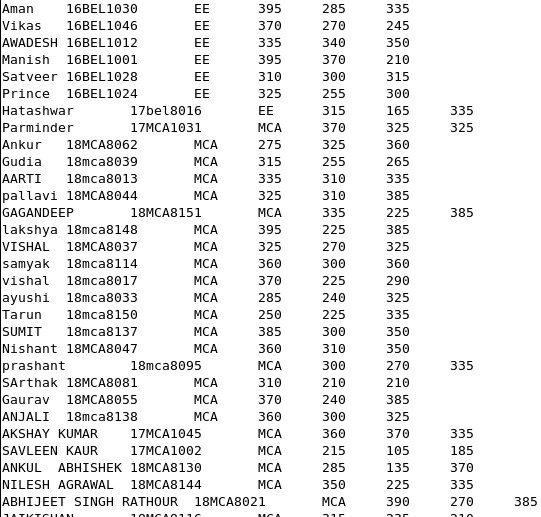


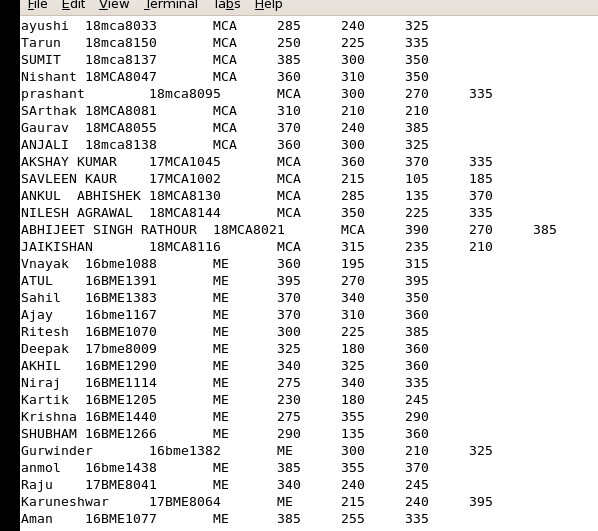
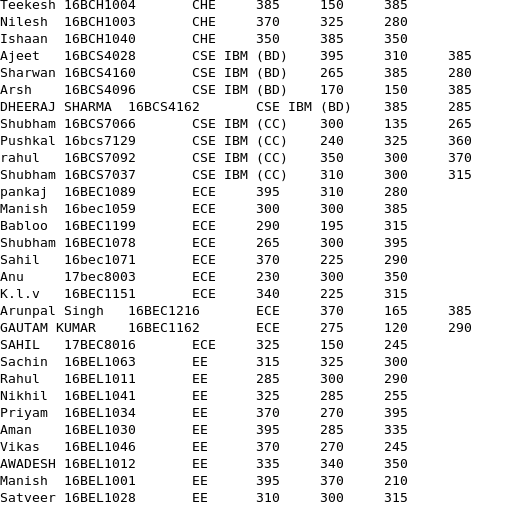


**PROBLEM STATEMENT 6: Listing the students who score less than 400 in ELQ combined.**

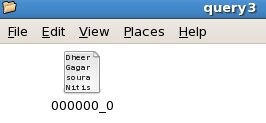
**Query:**Select firstname,universityrollno,branch,logical,quantative,English, from BTACHH where English<400 and logical<400 quantative<400;

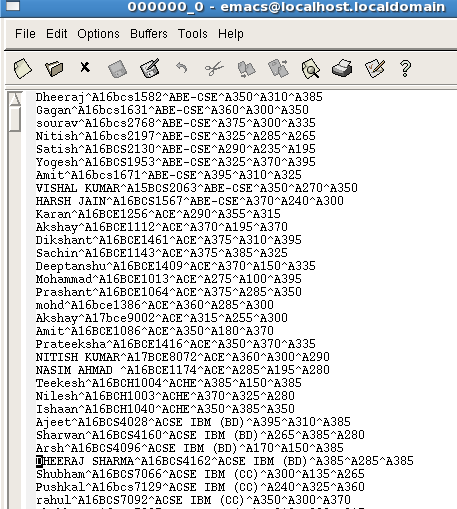
**Output:**





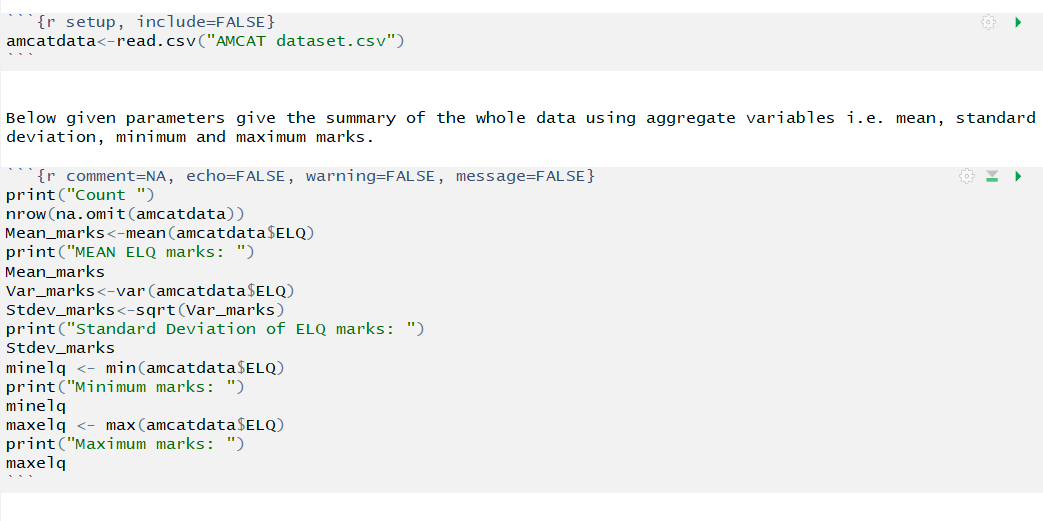
**Csv:**



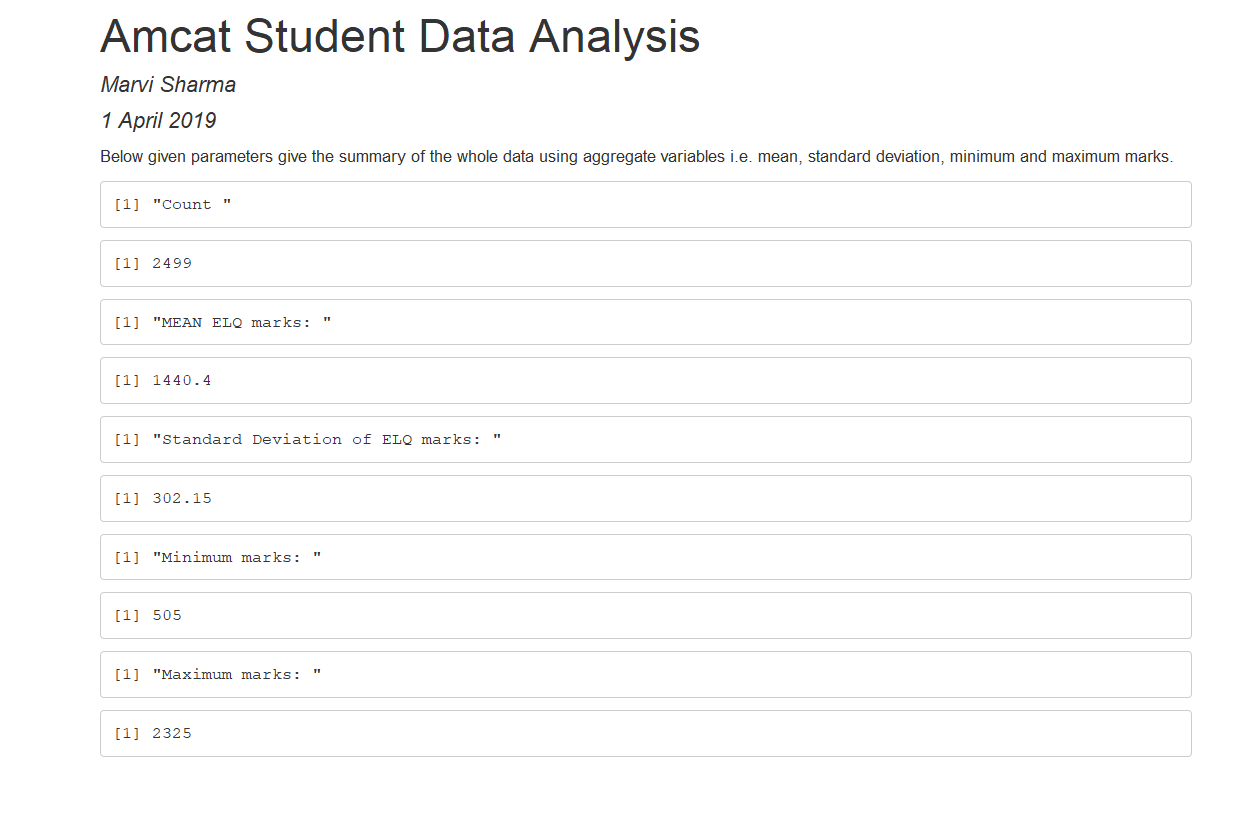


**PROBLEM STATEMENT 7: Listing the students who score less than 400 in ELQ combined.**

**R Markdown code:**

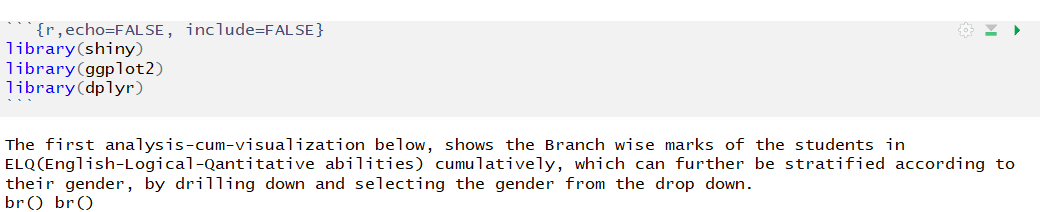
****

**Output graph:**

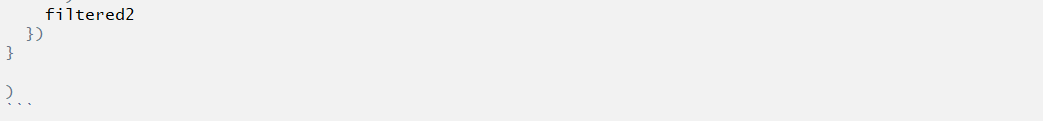
****

**PROBLEM STATEMENT 8: Checking branch-wise and gender-wise distribution of student marks.**

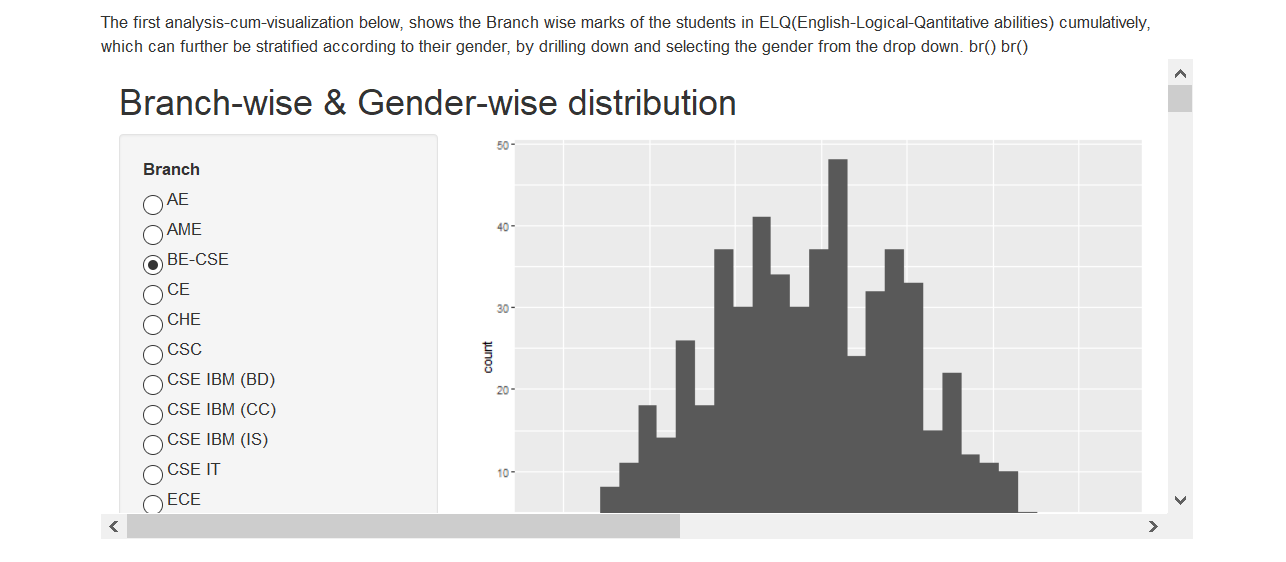
**R Markdown code:**

****

****

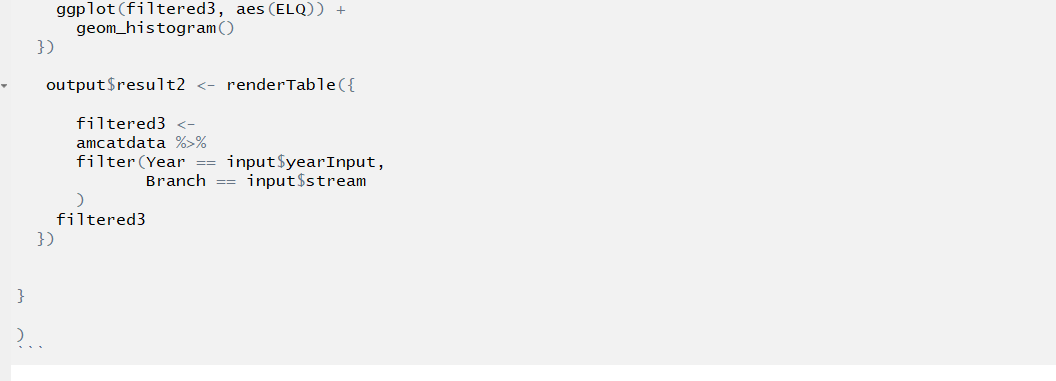
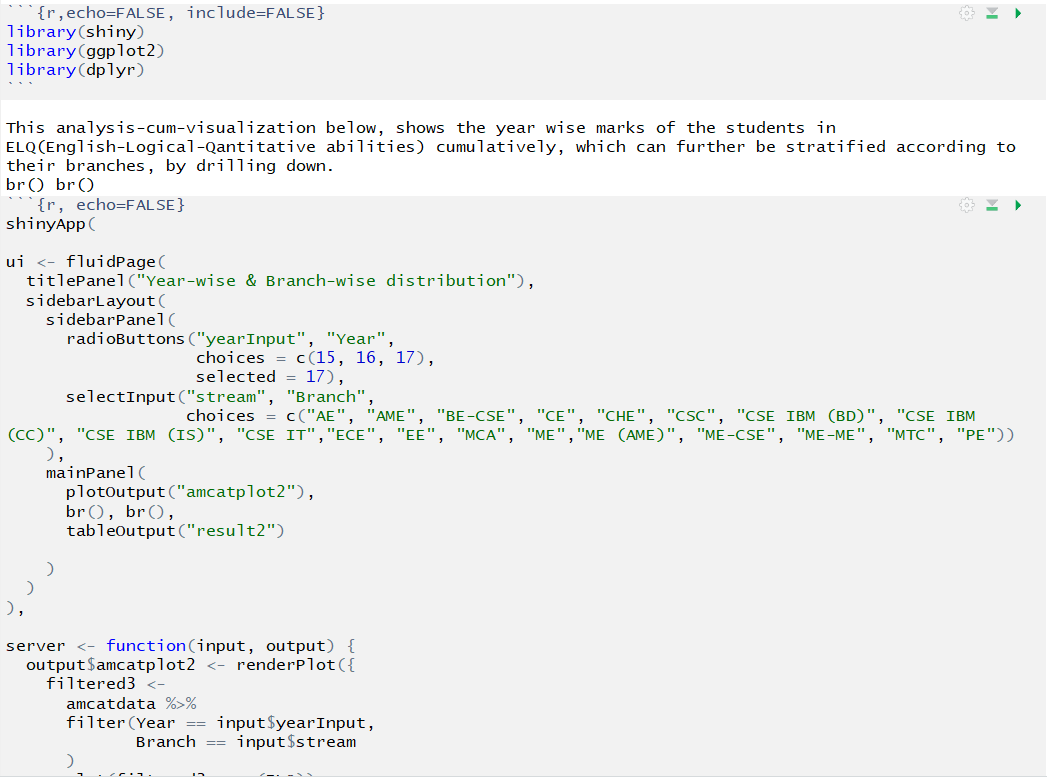
****

**Output graph:**

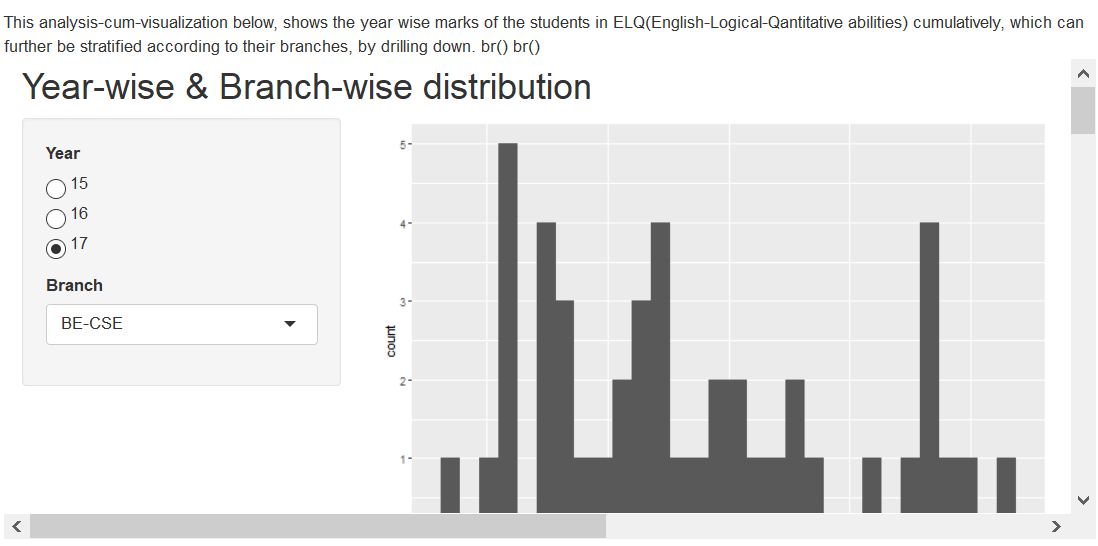


**PROBLEM STATEMENT 9: Checking year-wise and branch-wise distribution of student marks.**

**R Markdown code:**

****

**Output graph:**

****

**https://perso.telecom-paristech.fr/eagan/class/igr204/datasets**