ISM 6208
Data Warehousing

Data Warehousing Final Project

Topic: Analysis of Student Performance Factors in Education Data Mart

By- Group 7

Advith Reddy Kunda

Neshma Kasamneni

Purna Chandra Shekar Reddy Gosala

Raajitha Sai Bondada

Sai Anirudh Mantha

Shyam Rishi Kashyap Sai Kalaparthy

Executive Summary

The document explores the creation and analysis of a data mart focused on student performance, integrating data from mathematics and Portuguese courses. By establishing a structured data warehouse environment, we've enabled detailed analytical queries across various dimensions such as schools, students, and courses.

Through SQL queries, we investigated key factors affecting student performance, including attendance, family support, and educational outcomes across different schools and courses. This analysis provides insights into how demographics and educational environments influence academic success.

The findings from this data mart offer valuable perspectives for educational institutions to tailor interventions and support mechanisms, aiming to enhance student achievements and address identified challenges effectively.

Problem Statement

The educational sector often grapples with understanding the multifaceted factors that influence student performance. Schools and educational policymakers face challenges in pinpointing the specific variables that significantly impact learning outcomes, such as family background, teaching methods, and student engagement. The lack of a consolidated data framework makes it difficult to analyze these variables comprehensively, hindering the development of targeted strategies to improve student achievement and overall educational quality.

To address this issue, there is a need for a robust data mart that integrates diverse student data sets, enabling detailed analysis of performance across various subjects, like mathematics and Portuguese. The absence of a unified analytical platform limits the ability to perform deep dive analyses into the correlations between student demographics, support systems, and academic results. Establishing a comprehensive data mart could provide the necessary insights to craft effective educational policies and interventions, ultimately fostering an environment conducive to student success and academic excellence.

Literature Review

In the realm of educational research, the analysis of student performance data has been a focal point, aiming to uncover the dynamics affecting academic success. Studies have extensively explored variables such as socio-economic status, parental involvement, and educational resources, highlighting their significant impact on student outcomes. The integration of data warehousing and business intelligence tools in education, as seen in works by Marjanovic, O. (2010) and Romero, C., & Ventura, S. (2013), has paved the way for more sophisticated data-driven decision-making processes. These methodologies facilitate a holistic view of the educational landscape, enabling stakeholders to derive actionable insights from complex data sets.

The literature also emphasizes the importance of multidimensional data analysis in understanding educational phenomena. For instance, Baker, R. S., & Inventado, P. S. (2014) discuss the application of learning analytics to predict student performance, suggesting that data from various educational contexts can be synthesized to forecast and improve learning outcomes. However, challenges remain in terms of data integration and interpretation, as noted by Siemens, G., & Baker, R. S. (2012), who argue for the development of more nuanced analytical frameworks that can handle the complexities of educational data. The evolution of data mart solutions in education underscores a growing recognition of the need for robust, scalable systems that can support comprehensive analysis and foster educational improvement.

Data Collection and Preparation:

Data Set: https://www.kaggle.com/datasets/larsen0966/student-performance-data-set

The data, sourced from Kaggle, underwent a series of preparation steps including data cleaning, preprocessing, and transformation to make it suitable for analysis. Initially in CSV format, the raw data was imported into a database where the ETL (Extraction, Transformation, and Loading) process was executed. This led to the creation of tables within the database, enabling the execution of various queries to extract meaningful insights.

Data Set information:
The raw data comprised numerous columns, offering a wealth of potentially valuable information. However, to maintain the clarity and efficiency of our data mart, we selectively utilized only those data elements that were deemed most relevant and informative, avoiding the inclusion of data that could introduce complexity and hinder the analysis process.

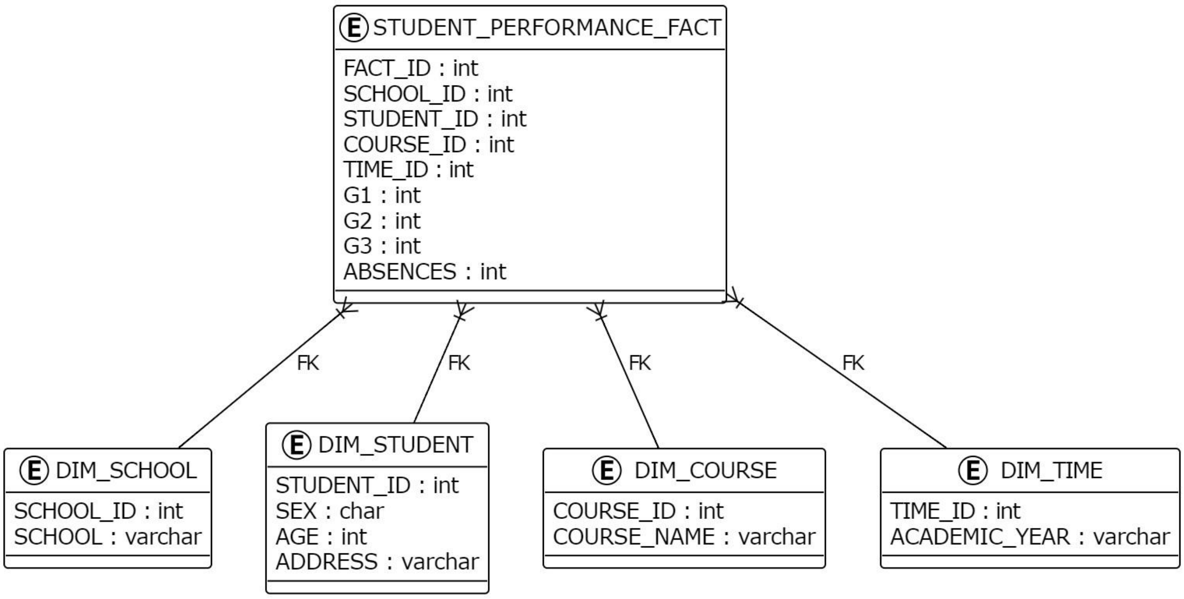
# Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

1 school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
2 sex - student's sex (binary: "F" - female or "M" - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: "U" - urban or "R" - rural)
5 famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
6 absences - number of school absences (numeric: from 0 to 93)

# these grades are related with the course subject, Math or Portuguese:
7 G1 - first period grade (numeric: from 0 to 20)
8 G2 - second period grade (numeric: from 0 to 20)
9 G3 - final grade (numeric: from 0 to 20, output target)
#As they were two different files for each subject we have created a dimension with course in it

10 Course_Name (Math or Portuguese)

Entity Relationship Diagram:



Part B: Dimensional Modelling

Fact Table: STUDENT_PERFORMANCE_FACT

The `STUDENT_PERFORMANCE_FACT` table is designed to store key performance indicators for students within an educational data mart. Here is a brief overview of each column in the table:

- `FACT_ID`: Serves as a primary key for the fact table.
- `SCHOOL_ID`: A numerical identifier for the school where the student is enrolled. This serves as a foreign key linking to a `DIM_SCHOOL` table that contains more detailed information about each school.
- `STUDENT_ID`: A numerical identifier unique to each student, acting as a foreign key that connects to a `DIM_STUDENT` table, providing more detailed demographic and other relevant information about the student.
- `COURSE_ID`: A character or number identifier representing different courses, like Math or Portuguese, linked to a `DIM_COURSE` table which would detail each course.
- `G1`, `G2`, `G3`: These are numerical fields representing grades or scores for the student at three different time points or assessment stages, enabling analysis of academic performance over time.
- `ABSENCES`: A numerical field tracking the number of times a student has been absent.
- `TIME_ID`: This number ties each record to a specific time, academic year, and would relate to a `DIM_TIME` table containing details about it.

This fact table is central to the data mart, providing a comprehensive view of student performance metrics across various dimensions such as time, student demographics, and academic courses.

Dimension Tables:
DIM_SCHOOL:
This table categorizes the schools involved in the study, identified by `SCHOOL_ID`, a numerical value assigned conditionally (`1` for 'GP' and `2' for 'MS'). The `SCHOOL` column contains the names of the schools. This dimension table allows for analysis based on specific schools.

DIM_STUDENT:
In the `DIM_STUDENT` table, each student is uniquely identified by a `STUDENT_ID`, generated as a sequential number. Additional attributes include `SEX`, `AGE`, and `ADDRESS`, `FAMSIZE`, providing demographic details of the students. This table facilitates demographic-based analysis and segmentation of student performance data.

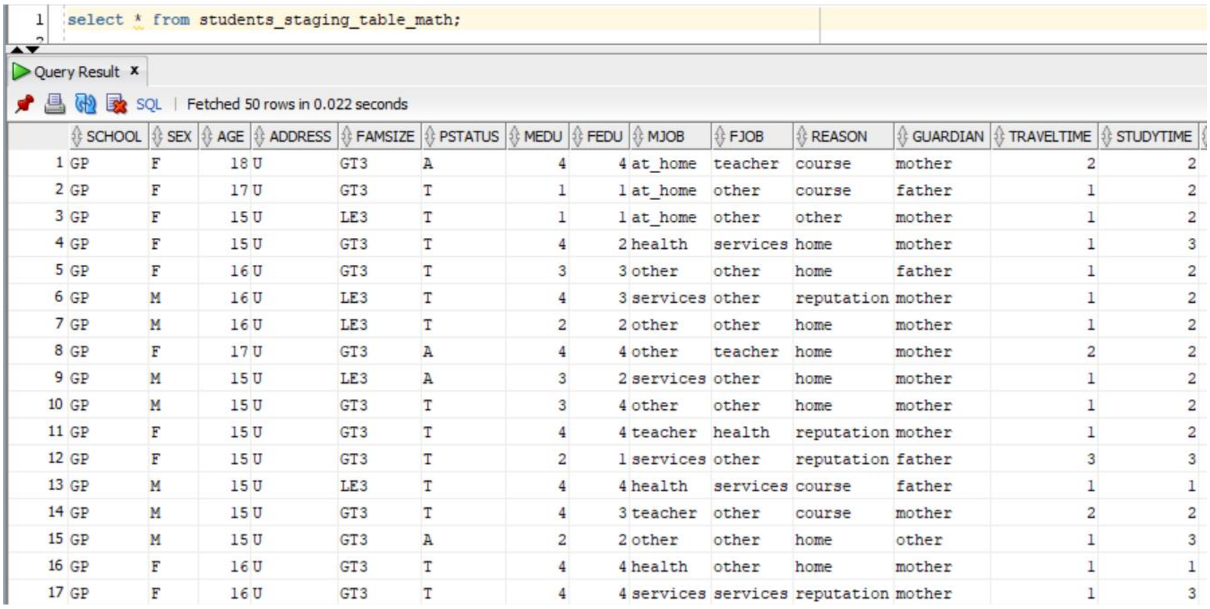
DIM_COURSE:
This table defines the courses studied, with `COURSE_ID` as a unique identifier for each course (e.g., `1` for Math, `2` for Portuguese) and `COURSE_NAME` specifying the course's name. It's essential for analyzing performance across different academic courses.

DIM_TIME:
The `DIM_TIME` table includes `TIME_ID` as a unique identifier for different time periods (like academic years) and `ACADEMIC_YEAR` detailing the specific year or term. This dimension allows for the temporal analysis of student performance across different time frames.

PART-C: Create Table Statements

-Creating the staging tables for both the csv files

```
CREATE TABLE "STUDENTS_STAGING_TABLE_MATH"
  (	"SCHOOL" VARCHAR2(50 BYTE),
	"SEX" CHAR(1 BYTE),
	"AGE" NUMBER,
	"ADDRESS" CHAR(1 BYTE),
	"FAMSIZE" VARCHAR2(50 BYTE),
	"PSTATUS" CHAR(1 BYTE),
	"MEDU" NUMBER,
	"FEDU" NUMBER,
	"MJOB" VARCHAR2(100 BYTE),
	"FJOB" VARCHAR2(100 BYTE),
	"REASON" VARCHAR2(100 BYTE),
	"GUARDIAN" VARCHAR2(100 BYTE),
	"TRAVELTIME" NUMBER,
	"STUDYTIME" NUMBER,
	"FAILURES" NUMBER,
	"SCHOOLSUP" VARCHAR2(50 BYTE),
	"FAMSUP" VARCHAR2(50 BYTE),
	"PAID" VARCHAR2(50 BYTE),
	"ACTIVITIES" VARCHAR2(50 BYTE),
	"NURSERY" VARCHAR2(50 BYTE),
	"HIGHER" VARCHAR2(50 BYTE),
        "INTERNET" VARCHAR2(50 BYTE),
        "ROMANTIC" VARCHAR2(50 BYTE),
        "FAMREL" NUMBER,
        "FREETIME" NUMBER,
        "GOOUT" NUMBER,
        "DALC" NUMBER,
        "WALC" NUMBER,
        "HEALTH" NUMBER,
        "ABSENCES" NUMBER,
        "G1" NUMBER,
        "G2" NUMBER,
        "G3" NUMBER
);
```

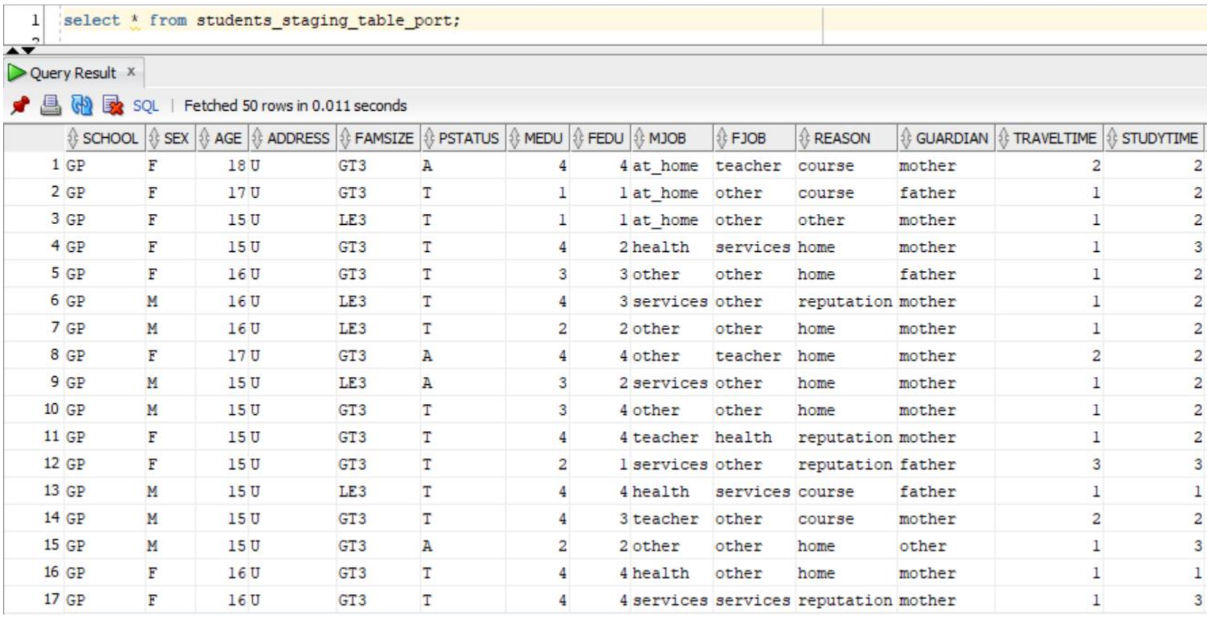


```
select * from students_staging_table_math;
```

Fetched 50 rows in 0.022 seconds

	SCHOOL	SEX	AGE	ADDRESS	FAMSIZE	PSTATUS	MEDU	FEDU	MJOB	FJOB	REASON	GUARDIAN	TRAVELTIME	STUDYTIME
1	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2
2	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2
3	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2
4	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3
5	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2
6	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1	2
7	GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2
8	GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2
9	GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2
10	GP	M	15	U	GT3	T	3	4	other	other	home	mother	1	2
11	GP	F	15	U	GT3	T	4	4	teacher	health	reputation	mother	1	2
12	GP	F	15	U	GT3	T	2	1	services	other	reputation	father	3	3
13	GP	M	15	U	LE3	T	4	4	health	services	course	father	1	1
14	GP	M	15	U	GT3	T	4	3	teacher	other	course	mother	2	2
15	GP	M	15	U	GT3	A	2	2	other	other	home	other	1	3
16	GP	F	16	U	GT3	T	4	4	health	other	home	mother	1	1
17	GP	F	16	U	GT3	T	4	4	services	services	reputation	mother	1	3

```
CREATE TABLE "STUDENTS_STAGING_TABLE_PORT"
(       "SCHOOL" VARCHAR2(50 BYTE),
        "SEX" CHAR(1 BYTE),
        "AGE" NUMBER,
        "ADDRESS" CHAR(1 BYTE),
        "FAMSIZE" VARCHAR2(50 BYTE),
        "PSTATUS" CHAR(1 BYTE),
        "MEDU" NUMBER,
        "FEDU" NUMBER,
        "MJOB" VARCHAR2(100 BYTE),
        "FJOB" VARCHAR2(100 BYTE),
        "REASON" VARCHAR2(100 BYTE),
        "GUARDIAN" VARCHAR2(100 BYTE),
        "TRAVELTIME" NUMBER,
        "STUDYTIME" NUMBER,
        "FAILURES" NUMBER,
        "SCHOOLSUP" VARCHAR2(50 BYTE),
        "FAMSUP" VARCHAR2(50 BYTE),
        "PAID" VARCHAR2(50 BYTE),
        "ACTIVITIES" VARCHAR2(50 BYTE),
        "NURSERY" VARCHAR2(50 BYTE),
        "HIGHER" VARCHAR2(50 BYTE),
        "INTERNET" VARCHAR2(50 BYTE),
        "ROMANTIC" VARCHAR2(50 BYTE),
        "FAMREL" NUMBER,
        "FREETIME" NUMBER,
        "GOOUT" NUMBER,
        "DALC" NUMBER,
        "WALC" NUMBER,
        "HEALTH" NUMBER,
        "ABSENCES" NUMBER,
        "G1" NUMBER,
        "G2" NUMBER,
        "G3" NUMBER
  );
```

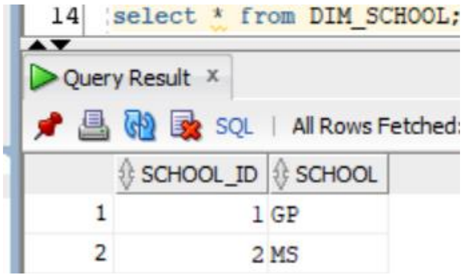


```
select * from students_staging_table_port;
```

	SCHOOL	SEX	AGE	ADDRESS	FAMSIZE	PSTATUS	MEDU	FEDU	MJOB	FJOB	REASON	GUARDIAN	TRAVELTIME	STUDYTIME
1	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2
2	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2
3	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2
4	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3
5	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2
6	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1	2
7	GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2
8	GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2
9	GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2
10	GP	M	15	U	GT3	T	3	4	other	other	home	mother	1	2
11	GP	F	15	U	GT3	T	4	4	teacher	health	reputation	mother	1	2
12	GP	F	15	U	GT3	T	2	1	services	other	reputation	father	3	3
13	GP	M	15	U	LE3	T	4	4	health	services	course	father	1	1
14	GP	M	15	U	GT3	T	4	3	teacher	other	course	mother	2	2
15	GP	M	15	U	GT3	A	2	2	other	other	home	other	1	3
16	GP	F	16	U	GT3	T	4	4	health	other	home	mother	1	1
17	GP	F	16	U	GT3	T	4	4	services	services	reputation	mother	1	3

-Creating dimension tables:

```
CREATE TABLE DIM_SCHOOL AS
SELECT DISTINCT
   CASE SCHOOL
      WHEN 'GP' THEN 1
      WHEN 'MS' THEN 2
      ELSE NULL
   END AS SCHOOL_ID,
   SCHOOL
FROM
   (SELECT SCHOOL FROM DW169.STUDENTS_STAGING_TABLE_MATH
    UNION
    SELECT SCHOOL FROM DW169.STUDENTS_STAGING_TABLE_PORT);
```

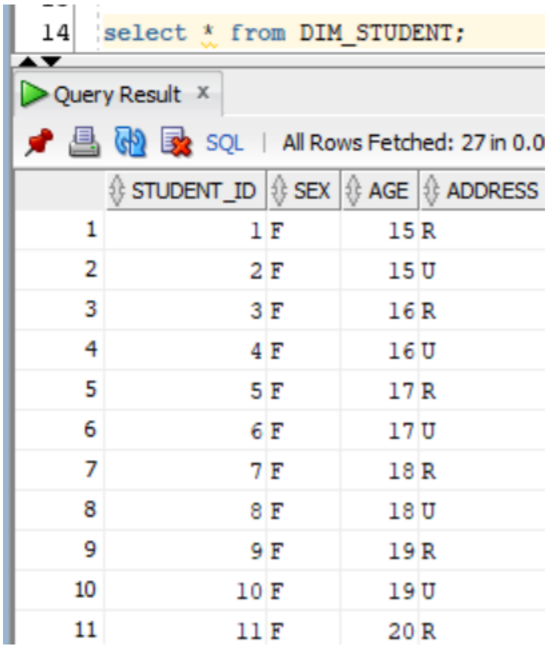


```
select * from DIM_SCHOOL;
```

All Rows Fetched:

	SCHOOL_ID	SCHOOL
1	1	GP
2	2	MS

```
CREATE TABLE DW169.DIM_STUDENT AS
SELECT
   ROW_NUMBER() OVER (ORDER BY SEX, AGE, ADDRESS) AS STUDENT_ID,
   SEX,
   AGE,
   ADDRESS
FROM
   (SELECT DISTINCT SEX, AGE, ADDRESS FROM
DW169.STUDENTS_STAGING_TABLE_MATH
    UNION
    SELECT DISTINCT SEX, AGE, ADDRESS FROM
DW169.STUDENTS_STAGING_TABLE_PORT);
```

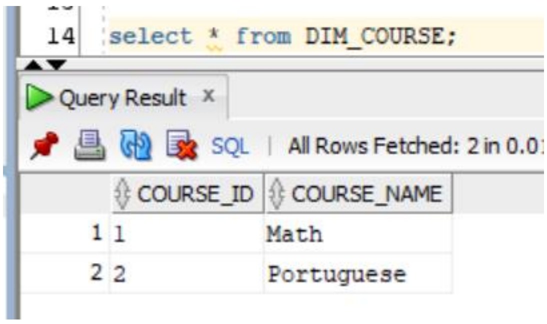


```
select * from DIM_STUDENT;
```

All Rows Fetched: 27 in 0.0

	STUDENT_ID	SEX	AGE	ADDRESS
1	1	F	15	R
2	2	F	15	U
3	3	F	16	R
4	4	F	16	U
5	5	F	17	R
6	6	F	17	U
7	7	F	18	R
8	8	F	18	U
9	9	F	19	R
10	10	F	19	U
11	11	F	20	R

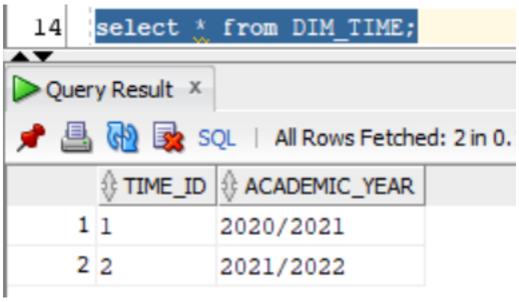
```
CREATE TABLE DW169.DIM_COURSE AS
SELECT DISTINCT
   COURSE_ID,
   COURSE_NAME
FROM
   (SELECT '1' AS COURSE_ID, 'Math' AS COURSE_NAME FROM DUAL
    UNION
    SELECT '2' AS COURSE_ID, 'Portuguese' AS COURSE_NAME FROM DUAL);
```



```
select * from DIM_COURSE;
```

	COURSE_ID	COURSE_NAME
1	1	Math
2	2	Portuguese

```
CREATE TABLE DW169.DIM_TIME AS
SELECT DISTINCT
    TIME_ID,
    ACADEMIC_YEAR
FROM
    (SELECT '1' AS TIME_ID, '2020/2021' AS ACADEMIC_YEAR FROM DUAL
     UNION
     SELECT '2' AS TIME_ID, '2021/2022' AS ACADEMIC_YEAR FROM DUAL);
```

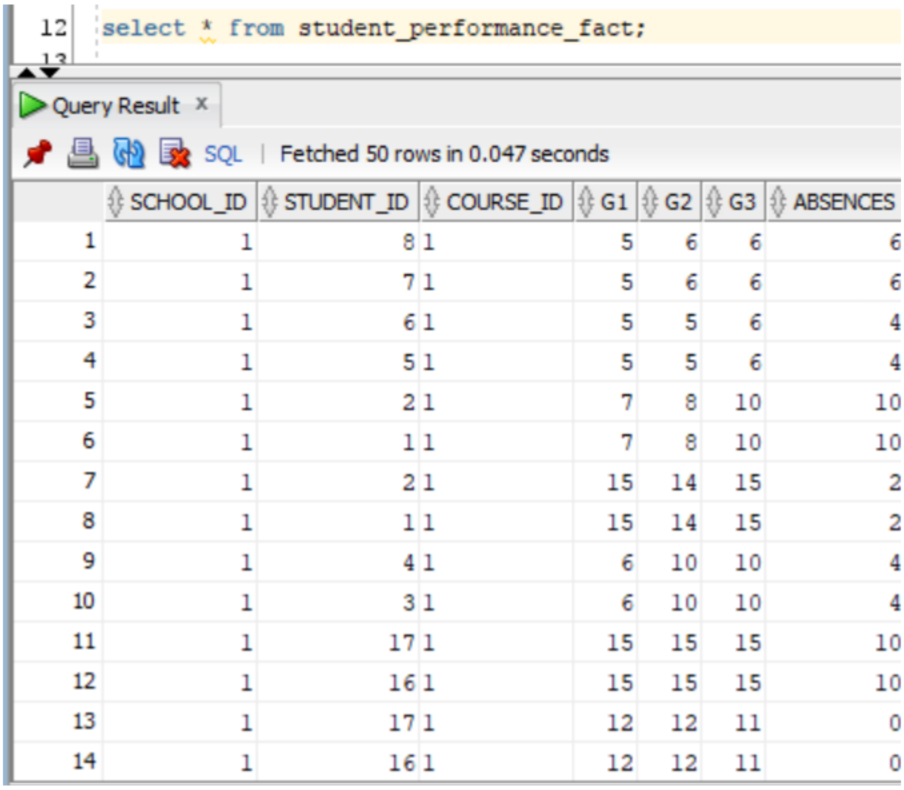


```
select * from DIM_TIME;
```

	TIME_ID	ACADEMIC_YEAR
1	1	2020/2021
2	2	2021/2022

```
CREATE TABLE STUDENT_PERFORMANCE_FACT AS
SELECT
    S.SCHOOL_ID,
    ST.STUDENT_ID,
    C.COURSE_ID,
    M.G1,
    M.G2,
    M.G3,
    M.ABSENCES
FROM
    DW169.STUDENTS_STAGING_TABLE_MATH M
JOIN DW169.DIM_SCHOOL S ON M.SCHOOL = S.SCHOOL
JOIN DW169.DIM_STUDENT ST ON M.SEX = ST.SEX AND M.AGE = ST.AGE
JOIN DW169.DIM_COURSE C ON C.COURSE_NAME = 'Math'
UNION ALL
SELECT
    S.SCHOOL_ID,
    ST.STUDENT_ID,
    C.COURSE_ID,
    P.G1,
    P.G2,
    P.G3,
    P.ABSENCES
```

```
FROM
    DW169.STUDENTS_STAGING_TABLE_PORT P
JOIN DW169.DIM_SCHOOL S ON P.SCHOOL = S.SCHOOL
JOIN DW169.DIM_STUDENT ST ON P.SEX = ST.SEX AND P.AGE = ST.AGE
JOIN DW169.DIM_COURSE C ON C.COURSE_NAME = 'Portuguese';
```



```
select * from student_performance_fact;
```

Fetched 50 rows in 0.047 seconds

	SCHOOL_ID	STUDENT_ID	COURSE_ID	G1	G2	G3	ABSENCES
1	1	8	1	5	6	6	6
2	1	7	1	5	6	6	6
3	1	6	1	5	5	6	4
4	1	5	1	5	5	6	4
5	1	2	1	7	8	10	10
6	1	1	1	7	8	10	10
7	1	2	1	15	14	15	2
8	1	1	1	15	14	15	2
9	1	4	1	6	10	10	4
10	1	3	1	6	10	10	4
11	1	17	1	15	15	15	10
12	1	16	1	15	15	15	10
13	1	17	1	12	12	11	0
14	1	16	1	12	12	11	0

PART-D: ETL Process

ETL Process Breakdown (Oracle SQL Developer)

In the above design, the ETL (Extraction, Transformation, and Loading) process is central to consolidating and organizing data from the staging tables (`STUDENTS_STAGING_TABLE_MATH` and `STUDENTS_STAGING_TABLE_PORT`) into a structured data mart. Here’s a brief overview of each ETL phase in this context:

Extraction
Data is extracted from the source staging tables, where raw data is collected. In this case, the sources are the staging tables for math and Portuguese courses. These tables contain comprehensive student data, including demographics, academic performance, and other relevant attributes.

Transformation
The transformation step involves cleaning, standardizing, and restructuring the data to fit the data mart schema. This includes:

- Assigning unique identifiers to entities (like schools and students) to create dimension tables (`DIM_SCHOOL`, `DIM_STUDENT`, `DIM_COURSE`, `DIM_TIME`).

- Converting the school names into numerical identifiers (`SCHOOL_ID`) in the `DIM_SCHOOL` table.
- Generating sequential `STUDENT_ID`s for the `DIM_STUDENT` table to uniquely identify students.
- Aggregating course data into the `DIM_COURSE` table with distinct identifiers.
- Creating time dimensions in the `DIM_TIME` table to represent different academic periods.

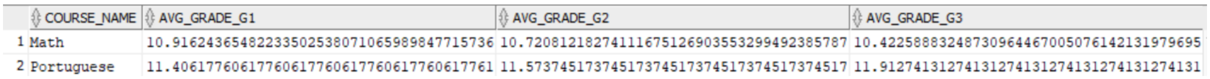
Loading
The transformed data is then loaded into the respective dimension tables and the fact table (`STUDENT_PERFORMANCE_FACT`). This table combines key performance metrics and foreign keys to the dimension tables, linking data points like student performance (grades and absences) with dimensions like school, student demographics, course, and time.

The fact table serves as the central repository for analysis, supporting queries that can span across multiple dimensions to provide insights into student performance and trends over time. This ETL process enables the structured analysis of educational data, facilitating a comprehensive understanding of the factors affecting student outcomes.

Exploratory Data Analysis

1. Average Grades Analysis by Course

```
SELECT
   C.COURSE_NAME,
   AVG(F.G1) AS AVG_GRADE_G1,
   AVG(F.G2) AS AVG_GRADE_G2,
   AVG(F.G3) AS AVG_GRADE_G3
FROM
   STUDENT_PERFORMANCE_FACT F
JOIN
   DW169.DIM_COURSE C ON F.COURSE_ID = C.COURSE_ID
GROUP BY
   C.COURSE_NAME;
```



	COURSE_NAME	AVG_GRADE_G1	AVG_GRADE_G2	AVG_GRADE_G3
1	Math	10.91624365482233502538071065989847715736	10.72081218274111675126903553299492385787	10.42258883248730964467005076142131979695
2	Portuguese	11.40617760617760617760617760617760617761	11.57374517374517374517374517374517374517	11.91274131274131274131274131274131274131

Output Data Explanation:

The output of the query provides the following average grades for each course:

- Math:

  - Average Grade G1: 10.92

  - Average Grade G2: 10.72

  - Average Grade G3: 10.42

- Portuguese:

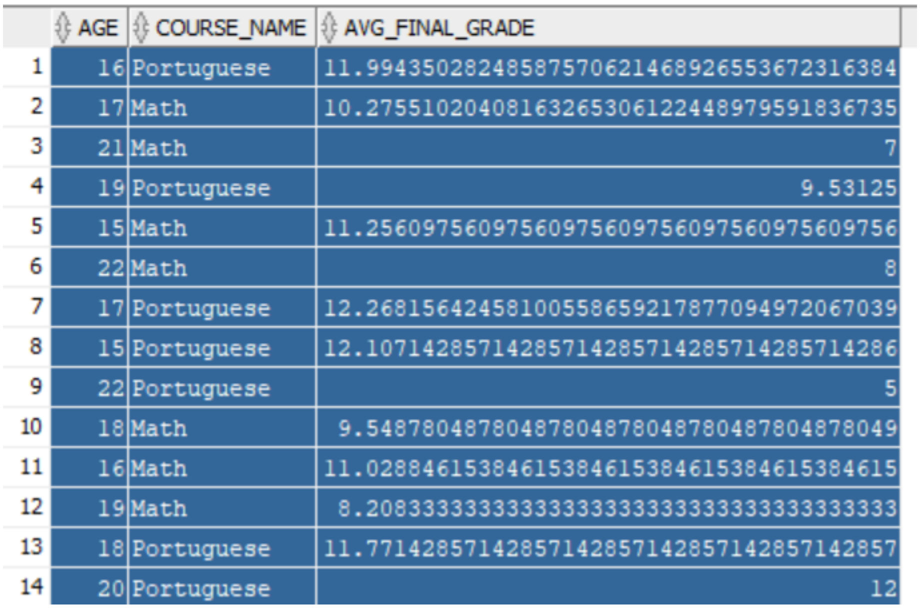
  - Average Grade G1: 11.41

- Average Grade G2: 11.57

- Average Grade G3: 11.91

2. Analysis of Average Final Grades by Age and Course

```
SELECT
   ST.AGE,
   C.COURSE_NAME,
   AVG(F.G3) AS AVG_FINAL_GRADE
FROM
   STUDENT_PERFORMANCE_FACT F
JOIN
   DW169.DIM_STUDENT ST ON F.STUDENT_ID = ST.STUDENT_ID
JOIN
   DW169.DIM_COURSE C ON F.COURSE_ID = C.COURSE_ID
GROUP BY
   ST.AGE,
   C.COURSE_NAME;
```



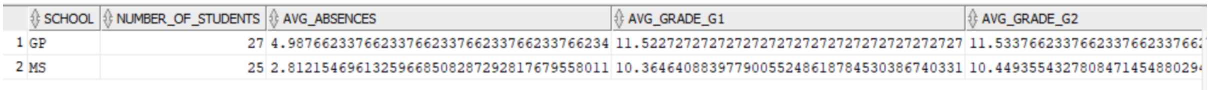
	AGE	COURSE_NAME	AVG_FINAL_GRADE
1	16	Portuguese	11.99435028248587570621468926553672316384
2	17	Math	10.27551020408163265306122448979591836735
3	21	Math	7
4	19	Portuguese	9.53125
5	15	Math	11.25609756097560975609756097560975609756
6	22	Math	8
7	17	Portuguese	12.26815642458100558659217877094972067039
8	15	Portuguese	12.10714285714285714285714285714285714286
9	22	Portuguese	5
10	18	Math	9.548780487804878048780487804878048780049
11	16	Math	11.02884615384615384615384615384615384615
12	19	Math	8.208333333333333333333333333333333333333
13	18	Portuguese	11.77142857142857142857142857142857142857
14	20	Portuguese	12

Output Data Explanation:
The output presents the average final grade (G3) for students of different ages in each course. Some notable data points include:

- 16-year-olds in Portuguese have an average final grade of approximately 12.00.
- 17-year-olds in Math have an average final grade of around 10.28, while in Portuguese, it's higher at about 12.27.
- Students aged 21 in Math have a significantly lower average final grade of 7.
- The data shows variability in academic performance based on age and course. For instance, 20-year-olds have an average final grade of 14 in Math, which is a notable peak.
- Conversely, 22-year-olds in Portuguese show a considerable dip, with an average final grade of 5.

3. School-wise Student Attendance and Performance Analysis

```
SELECT
   S.SCHOOL,
   COUNT(DISTINCT F.STUDENT_ID) AS NUMBER_OF_STUDENTS,
   AVG(F.ABSENCES) AS AVG_ABSENCES,
   AVG(F.G1) AS AVG_GRADE_G1,
   AVG(F.G2) AS AVG_GRADE_G2,
   AVG(F.G3) AS AVG_GRADE_G3,
   SUM(CASE WHEN C.COURSE_NAME = 'Math' THEN F.ABSENCES ELSE 0 END)
AS TOTAL_MATH_ABSENCES,
   SUM(CASE WHEN C.COURSE_NAME = 'Portuguese' THEN F.ABSENCES ELSE 0
END) AS TOTAL_PORTUGUESE_ABSENCES
FROM
   STUDENT_PERFORMANCE_FACT F
JOIN
   DW169.DIM_SCHOOL S ON F.SCHOOL_ID = S.SCHOOL_ID
JOIN
   DW169.DIM_COURSE C ON F.COURSE_ID = C.COURSE_ID
GROUP BY
   S.SCHOOL;
```



	SCHOOL	NUMBER_OF_STUDENTS	AVG_ABSENCES	AVG_GRADE_G1	AVG_GRADE_G2
1	GP	27	4.9876623376623376623376623376623376623 4	11.5227272727272727272727272727272727272727	11.533766233766233766233766
2	MS	25	2.8121546961325966850828729281767955801 1	10.3646408839779005524861878453038674033 1	10.449355432780847145488029

Output Data Explanation:
The output shows aggregated data for each school, represented here by 'GP' and 'MS'. Key findings include:

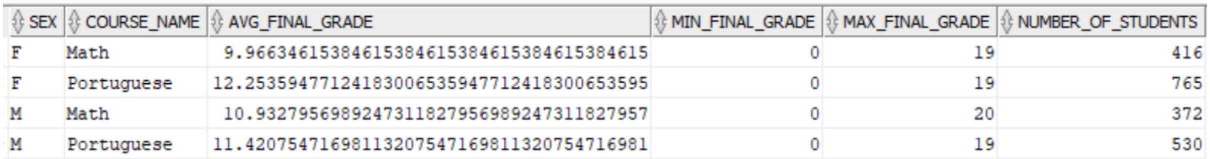
- School GP:
  - Total Students: 27
  - Average Absences: ~4.99
  - Average Grades: G1 - 11.52, G2 - 11.53, G3 - 11.64
  - Total Absences in Math: 4148
  - Total Absences in Portuguese: 3533

- School MS:
  - Total Students: 25
  - Average Absences: ~2.81
  - Average Grades: G1 - 10.36, G2 - 10.45, G3 - 10.52
  - Total Absences in Math: 343
  - Total Absences in Portuguese: 1184

4. Performance by Gender and Course

```
SELECT
   ST.SEX,
   C.COURSE_NAME,
   AVG(F.G3) AS AVG_FINAL_GRADE,
   MIN(F.G3) AS MIN_FINAL_GRADE,
   MAX(F.G3) AS MAX_FINAL_GRADE,
   COUNT(F.STUDENT_ID) AS NUMBER_OF_STUDENTS
```

```
FROM
   STUDENT_PERFORMANCE_FACT F
JOIN
   DW169.DIM_STUDENT ST ON F.STUDENT_ID = ST.STUDENT_ID
JOIN
   DW169.DIM_COURSE C ON F.COURSE_ID = C.COURSE_ID
GROUP BY
   ST.SEX,
   C.COURSE_NAME;
```



SEX	COURSE_NAME	AVG_FINAL_GRADE	MIN_FINAL_GRADE	MAX_FINAL_GRADE	NUMBER_OF_STUDENTS
F	Math	9.966346153846153846153846153846153846 15	0	19	416
F	Portuguese	12.2535947712418300653594771241830065359 5	0	19	765
M	Math	10.932795698924731182795698924731182795 7	0	20	372
M	Portuguese	11.420754716981132075471698113207547169 81	0	19	530

Output Data Explanation:
The results break down as follows for each gender-course pairing:

- Female Students in Math:
  - Average Final Grade: ~9.97
  - Minimum Final Grade: 0
  - Maximum Final Grade: 19
  - Number of Students: 416

- Female Students in Portuguese:
  - Average Final Grade: ~12.25
  - Minimum Final Grade: 0
  - Maximum Final Grade: 19
  - Number of Students: 765

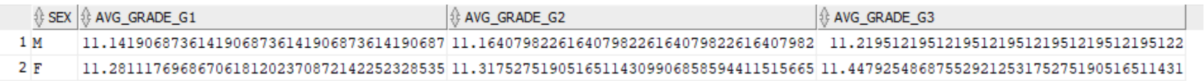
- Male Students in Math:
  - Average Final Grade: ~10.93
  - Minimum Final Grade: 0
  - Maximum Final Grade: 20
  - Number of Students: 372

- Male Students in Portuguese:
  - Average Final Grade: ~11.42
  - Minimum Final Grade: 0
  - Maximum Final Grade: 19
  - Number of Students: 530

5. Gender-based Analysis of Student Academic Performance

```
SELECT
   ST.SEX,
   AVG(F.G1) AS AVG_GRADE_G1,
   AVG(F.G2) AS AVG_GRADE_G2,
   AVG(F.G3) AS AVG_GRADE_G3
FROM
   STUDENT_PERFORMANCE_FACT F
JOIN
```

```
    DW169.DIM_STUDENT ST ON F.STUDENT_ID = ST.STUDENT_ID
GROUP BY
    ST.SEX;
```



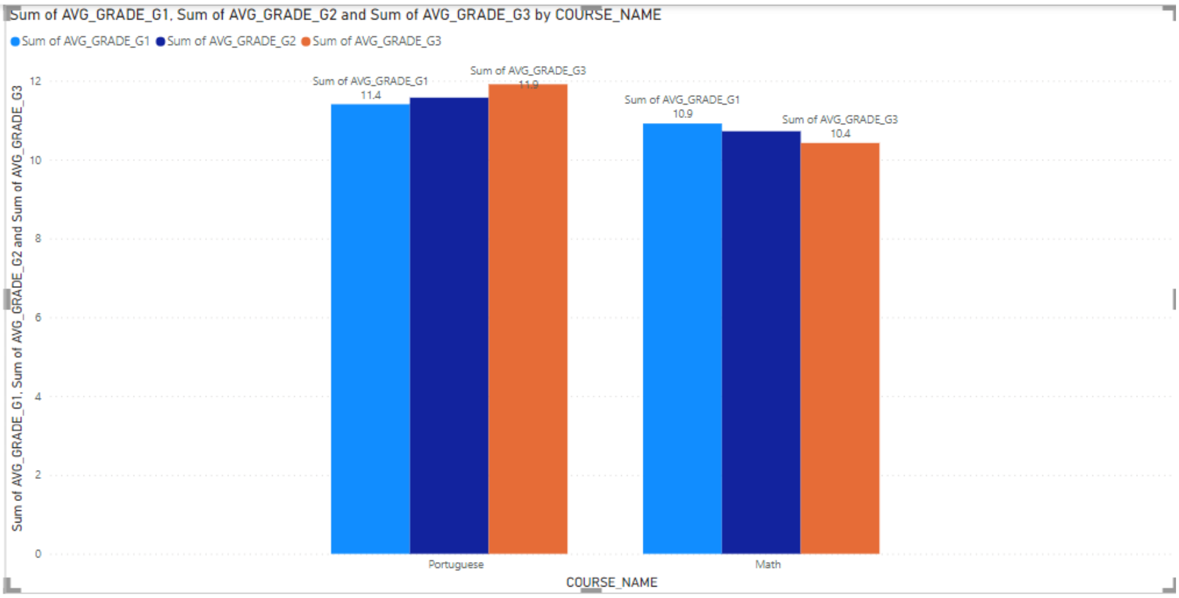
	SEX	AVG_GRADE_G1	AVG_GRADE_G2	AVG_GRADE_G3
1	M	11.141906873614190687361419068736141906 87	11.164079822616407982261640798226164079 82	11.219512195121951219512195121951219512 2
2	F	11.281117696867061812023708721422523285 35	11.317527519051651143099068585944115156 65	11.447925486875529212531752751905165114 31

Output Data Explanation:
The dataset yields the following insights into the average grades of male (M) and female (F) students:
- Male Students:
  - Average Grade G1: 11.14
  - Average Grade G2: 11.16
  - Average Grade G3: 11.22

- Female Students:
  - Average Grade G1: 11.28
  - Average Grade G2: 11.32
  - Average Grade G3: 11.45

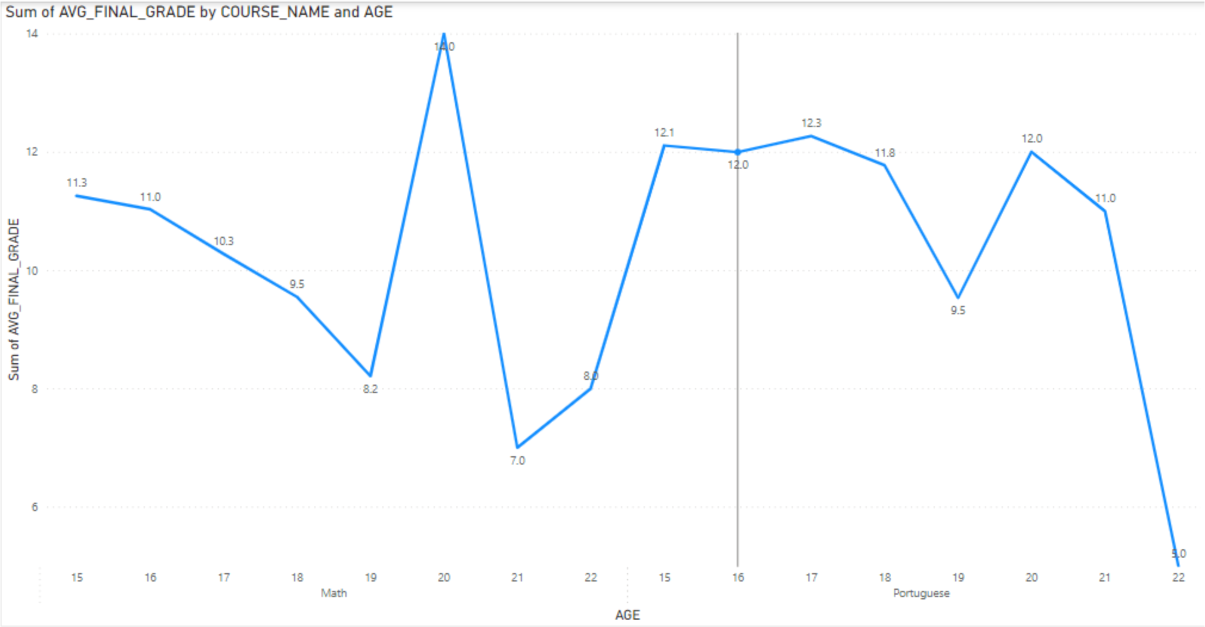
Reporting, Modelling and Storytelling

1. What is the Average Grade by course?



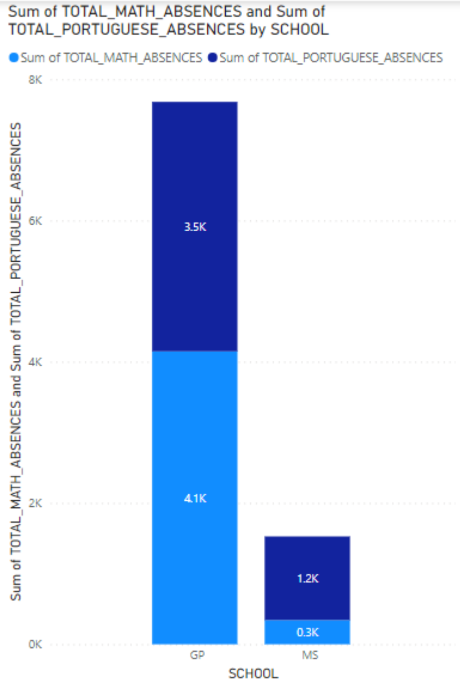
The data indicates that, on average, students scored higher in Portuguese than in Math across all three grading periods. Moreover, there is an upward trend in the average grades from G1 to G3 for Portuguese, suggesting improvement over time. In contrast, the average grades for Math show a slight decline from G1 to G3.

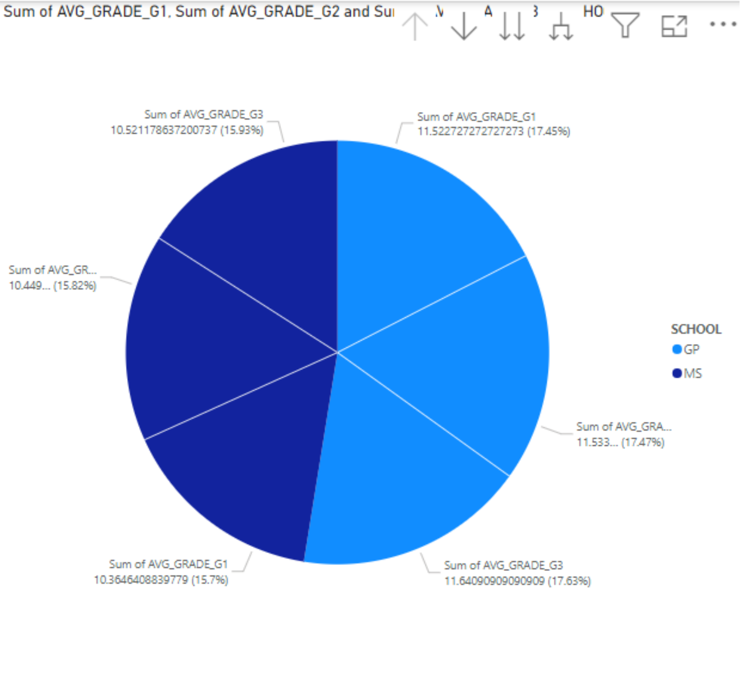
2. How does the student age affect the performance in each course?



This analysis suggests that academic performance in terms of final grades is influenced by the age of the students and the specific courses they are enrolled in, with some age groups performing better in certain subjects than others.

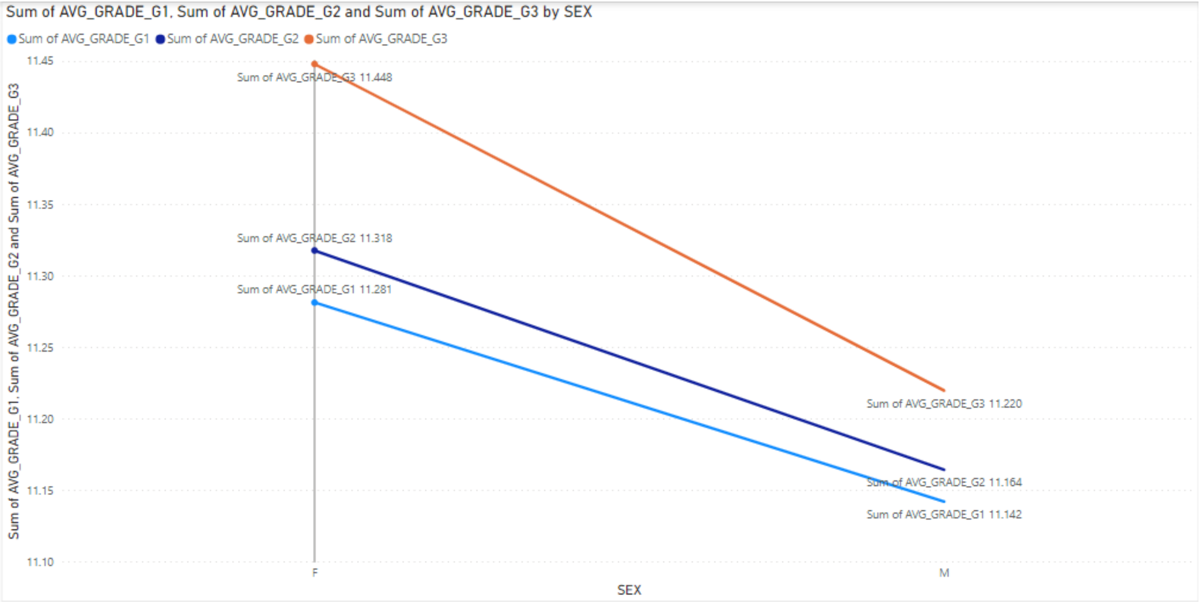
3. What is School-wise Student Attendance and Performance Analysis?





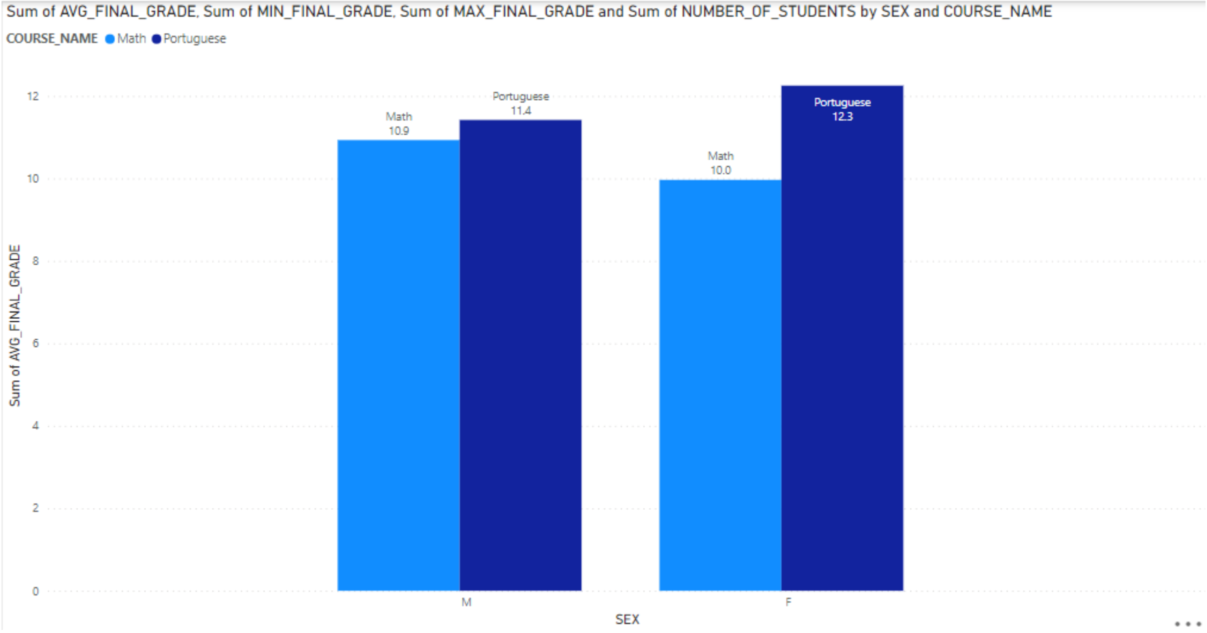
From the data, we can infer that School GP has a higher average in terms of both absences and grades compared to School MS. Additionally, both schools show higher absences in Math than in Portuguese, with a notably higher figure in School GP. This analysis can help identify patterns and areas for improvement in student attendance and academic performance across different schools and subjects.

4. What is the Average Grade by Gender?



From this data, we observe that female students have a higher average final grade in Portuguese compared to Math, whereas male students show a relatively balanced performance in both subjects. Notably, the minimum final grades are zero for all groups, which may indicate failing scores or lack of participation. The higher number of female students in Portuguese might suggest a greater female interest or enrolment in this course. This analysis can guide educational strategies and resource allocation to address gender disparities and academic performance in different subjects.

5. What is Gender-based Analysis of Student Academic Performance?



The data indicates that female students, on average, score slightly higher than male students in all three grading periods. The incremental increase in the average grades from G1 to G3 for both genders suggest an overall improvement in academic performance as the course progresses. This analysis highlights the importance of gender-specific educational strategies and support mechanisms to foster academic growth and address any disparities.

Conclusion

The series of SQL queries and analyses conducted on the educational data mart, focusing on student performance across various dimensions, reveals significant insights:

1. Course-based Performance Analysis: The average grades across different courses show distinct patterns, with students generally performing better in Portuguese than in Math. This suggests that the curriculum or teaching methodologies might need adjustments to enhance understanding and performance in Math.

2. Age and Course Performance Correlation: The data indicates that student performance in courses varies with age, with older students showing varying results in Math and Portuguese. This variability could be levered to tailor educational content and support services to different age groups.

3. School-wise Attendance and Performance: The analysis points to differences in student attendance and performance between schools, highlighting the need for targeted interventions in schools with higher absences and lower grades to improve academic outcomes.

4. Gender-based Performance Metrics: Gender differences in performance were observed, with female students generally achieving higher average grades than their male counterparts. This calls for a gender-sensitive approach in educational planning and resource allocation to ensure equitable opportunities for success.

5. General Academic Trends: Across the analyses, while some patterns of improvement were noted over time within courses, there were also areas where performance dipped, indicating potential areas for targeted support and intervention.

In conclusion, the comprehensive analysis of the educational data mart underscores the critical role of data-driven decision-making in the education sector. The insights obtained from these queries can guide stakeholders in refining educational strategies, ensuring resources are optimally allocated, and tailoring interventions to meet the diverse needs of students, ultimately aiming to enhance the educational experience and outcomes for all students.

References

1. Data Set: https://www.kaggle.com/datasets/larsen0966/student-performance-data-set
2. https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/books/data-warehouse-dw-toolkit/