MINI GROUP ASSIGNMENT

DBMS-CASE STUDY

GROUP 4

INDEX

1. Team Members
2. Introduction to the Problem Statement
3. Aim
4. Technologies Used & Skills Used
5. Data Description
6. Data Collection and Cleaning
7. Problem Solving Steps
8. What could have been done better?
9. Takeaways and Conclusions
10. Future Steps

TEAM MEMBERS

- Hithashree J.
- Aditya Aryan
- Shrawani Hemant Deshmukh
- Adarsh Kumar

INTRODUCTION TO THE PROBLEM STATEMENT

- In the first dataset provided, we have the test cricket details as well as links to profiles of numerous cricketers who played in test matches from the 19th century up until a few years ago. The primary task to accomplish from the data was analysing it by separating data from columns which were filled with more than one determining values and looking at various factors in order to create best possible groups of batsmen.

- For the second set of problems, we have the database of a food supply chain in business with many companies associated with food supply and customers worldwide. The entire data is divided into multiple tables to maintain a systemized organization. The main work was to get insights of the business by finding out the most profitable orders, companies and products, and getting a better analysis of supplier-customer relations.

AIM

- **ICC Test Cricket:**

The primary task to accomplish from the data is analysing it by separating data from columns which were filled with more than one determining values and looking at various factors in order to create best possible groups of batsmen.

- **Supply chain database :**

The main work is to get insights of the business by finding out the most profitable orders, companies and products, and getting a better analysis of supplier-customer relations.

- **TECNOLOGIES USED**

✓ Structured Query Language (SQL)

✓ MySQL Workbench

- **SKILLS USED**

✓ Programming skills

✓ Data analysis skills

DATA DESCRIPTION

- **ICC Test Cricket**

❑ In the dataset, there are 12 columns and 3001 rows.

❑ There are no null values or duplicate records but missing values are denoted by hyphens (-).

❑ There are 9 categorical attributes in the dataset.

❑ The maximum total runs scored by any cricketer is 15921. The mean total runs of the data is 745.01.

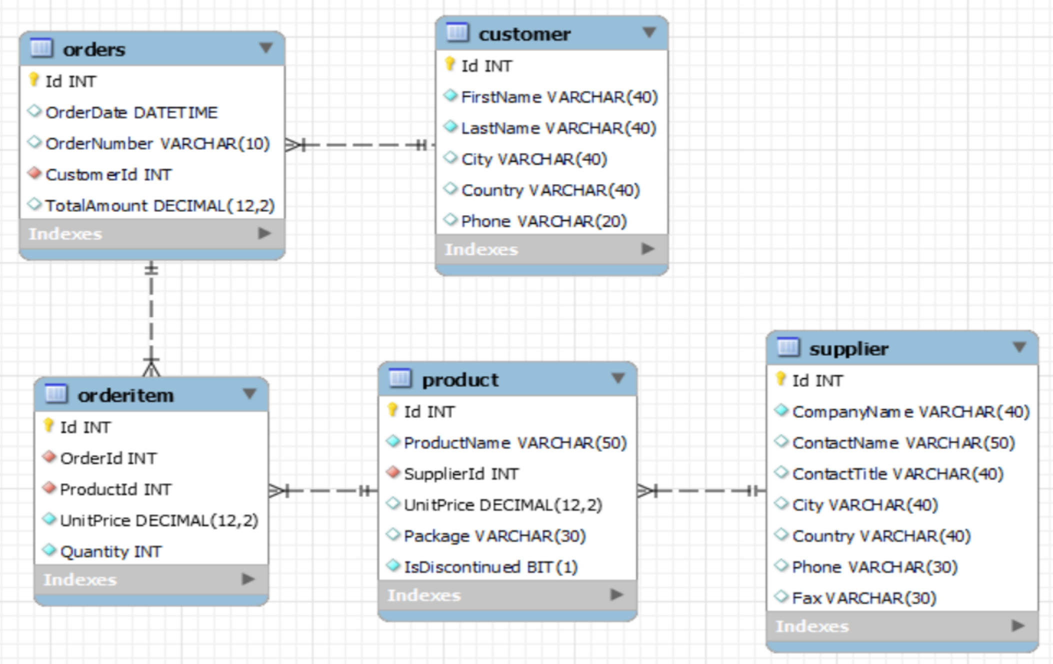
❑ The maximum average runs per match of any cricketer is 160.5. The mean is 20.54.

- **Richard's Supply**

□ There are 5 tables in the database with 41 columns altogether and multiple rows.

□ There are null values as well as duplicate records in more than one tables.

□ There are 23 text columns, 17 numerical columns and 1 datetime column.



orders
- Id INT
- OrderDate DATETIME
- OrderNumber VARCHAR(10)
- Custom erId INT
- TotalAmount DECIMAL(12,2)
- Indexes

customer
- Id INT
- FirstName VARCHAR(40)
- LastName VARCHAR(40)
- City VARCHAR(40)
- Country VARCHAR(40)
- Phone VARCHAR(20)
- Indexes

orderitem
- Id INT
- OrderId INT
- ProductId INT
- UnitPrice DECIMAL(12,2)
- Quantity INT
- Indexes

product
- Id INT
- ProductName VARCHAR(50)
- SupplierId INT
- UnitPrice DECIMAL(12,2)
- Package VARCHAR(30)
- IsDiscontinued BIT(1)
- Indexes

supplier
- Id INT
- CompanyNam e VARCHAR(40)
- ContactName VARCHAR(50)
- ContactTitle VARCHAR(40)
- City VARCHAR(40)
- Country VARCHAR(40)
- Phone VARCHAR(30)
- Fax VARCHAR(30)
- Indexes

DATA COLLECTION AND CLEANING

- For the first set of tasks, the table was imported from the ‘ICC Test Batting Figures’ CSV file.
- Since the file had missing values represented by hyphens (-), we first imported the missing value columns as text so as to not leave any records from the file. Then we updated the table replacing all the hyphens with NULL values. Finally, we altered the table to convert the columns concerned into suitable datatypes.
- The database and tables for the Richard’s Supply data were created by executing the following SQL scripts in this order:

i. 1_ddl_case study

ii. 2_data

iii. 3_data constraints

- Data cleaning was not required for these tables.

PROBLEM SOLVING STEPS

- **ICC Test Cricket**

- We created a new database and made it the current database to work. Thereafter, we imported the table from the provided CSV file. After cleaning the data, we gave the columns their suitable datatypes.
- ‘Player Profile’ column was dropped using the ALTER TABLE statement.
- To extract countries, we took out the characters after the last occurrence of ‘(’ in values of ‘Player Profile’ column and trimmed the final ‘)’. We extracted names by taking out all the characters on the left of ‘(’. Data from both of these was placed in two new separate columns.
- Start and end years were extracted by taking out first four and last four characters from the ‘Span’ column respectively.
- If there was an ‘*’ present in ‘HS’ column, every character except the last was extracted, otherwise the data was entirely taken into the new column. Likewise, if ‘*’ was present, we inserted 1, else 0 for not-outs.
- We wrote queries to get names of cricketers with 2019 between their start and end years for the respective countries and limited them to 6.
- Count of cricketers in each country was found out by applying COUNT() function on ‘name’ column and grouping by ‘country’.
- We again put COUNT() on ‘name’ column with condition as ‘IND’, ‘PAK’, ’AFG’, ’BDESH’ or ’SL’ to be present in ‘country’ column to find no. of Asian cricketers, and the opposite for no. of non-Asian cricketers.

- **<u>Richard's Supply</u>**

- Saving per order was found out by subtracting the product of actual price and quantity from the product of selling price and quantity and grouping by order ID. Result was then ordered by descending savings.
- We found out the highest demand products by ordering products in descending order of their count of order IDs and limiting to 5. This was then used in a join to find out suppliers for those products.
- All the tables were joined to find out customers and suppliers which belonged to the same country. We then queried out customers with countries not present in the list of supplier countries and vice versa.
- A view was created with companies, countries and sales by them. Using the same, companies were ranked by their countries and top two were displayed for each country.
- Products and supplier countries were shown for which consumers were from UK but suppliers were not.
- We created the two tables as mentioned and also the desired trigger with AFTER DELETE to insert values in the new table after deleting records.

TAKEAWAYS AND CONCLUSIONS

- **ICC Test Cricket**

1. MA Agarwal, V Kohli, CA Pujara, RG Sharma, RR Pant and AM Rahane are the top six Indian batsmen by average runs per match in 2019.
2. Likewise, V Kohli, CA Pujara, AM Rahane, RG Sharma, KL Rahul and R Ashwin are the top Indian batsmen by total number of centuries.
3. V Kohli, MA Agarwal, RG Sharma, CA Pujara, KL Rahul and AM Rahane are the best by total number of half-centuries.
4. Six South African players of 2019 who hold the highest average runs per match are found to be S Muthusamy, HM Amla, F du Plessis, Q de Kock, D Elgar and AK Markram.
5. Similarly, HM Amla, D Elgar, F du Plessis, Q de Kock, AK Markram and T Bavuma hit the greatest number of centuries.
6. England has the highest number of players while Ireland the lowest.
7. Out of 3001, 797 players belong to countries in Asia and 2204 belong to the rest of the world.

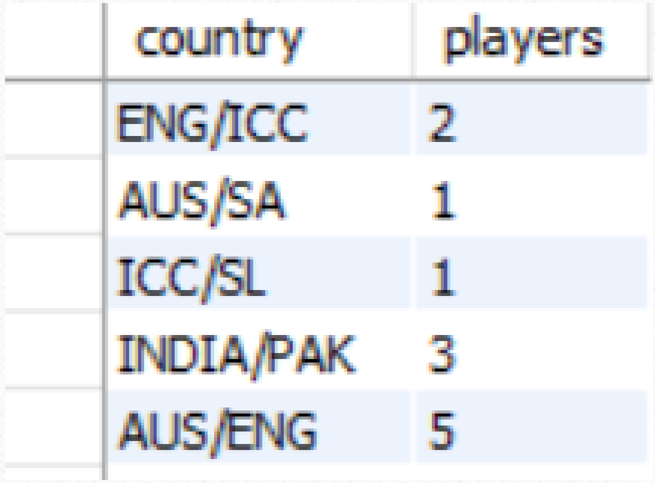
- **Richard’s Supply**

- Products of highest demand based on total number of orders are Raclette Courdavault, Gorgonzola Telino, Camembert Pierrot, Guaraná Fantástica and Gnocchi di nonna Alice.
- In Japan, Australia, Singapore and Netherlands, the suppliers only sell to foreign customers.

WHAT COULD HAVE BEEN DONE BETTER?

Countries from the ‘Player’ column of the Test Cricket data could have been extracted individually instead of as groups of countries with another country or ‘ICC’.

The above-mentioned would have made the presentation of count of players in different countries better. Also, counting the number of Asian and non-Asian players would have been easier.



country	players
ENG/ICC	2
AUS/SA	1
ICC/SL	1
INDIA/PAK	3
AUS/ENG	5

FUTURE STEPS

- The start and end years in the ICC Test Cricket dataset can be used to calculate the career duration of all the cricketers in years.
- We can find out the most experienced players by ordering them in descending order of total number of matches played.
- Using the ‘0’ column, the batsmen who have never been duck out can be found out.
- From the Richard’s Supply database, we can find out the companies with the largest variety of products.
- The expansion of Richard’s Supply can be determined by counting the number of countries it has suppliers in and the number of countries the chain supplies to as well as the total number of suppliers and customers.
- We can find out the most frequently ordered products by using the ‘OrderDate’ column in the ‘Order’ table.
- Customers who have placed the greatest number of orders can be found out and special discounts can be offered to them.
- We can calculate the country-wise count of customers to find out countries with the largest consumer bases.
- To get an insight into the progress of the supply chain, we can get the year-wise total sale amounts.
- Customers with no orders placed can be spotted and special bonuses can be offered by contacting them to encourage them into buying products.

THANK YOU