**ST1501 CA2 Group Tasks**

**Class:** DAAA/FT/2A/01

**Group No:** 4

**Group Members:**

|  |  |  |
| --- | --- | --- |
| **Student No** | **Name** | **Team Lead (Y/N)** |
| 2309024 | Jaysoorya Rajendran | Y |
| 2309389 | Tay Yikoon | N |
| 2209520 | Don Goh | N |
| 2309376 | Yong Bong Hong Bin | N |

**Group Tasks Solution**

1. Submit OLTP\_insert.sql including SQL queries that insert data into the OLTP tables.

Identify data quality issues in dataset provided. List down details in the table below.

|  |  |  |
| --- | --- | --- |
| Issue No. | Description of issue (which file, which line/part, what is the problem) | Solution (how do you resolve the issue) |
| 1 | order.csv, line 857 CustomerID is missing | Dropped the row with missing value |
| 2 | order.csv, line 2722 CustomerID has typo (c0068xxxxxxxx) | Replaced value with ‘c0068’ |
| 3 | employee.docx and modelling.xlsx are not in .csv format | Converted employee.docx to .csv and separated all the worksheets in modelling.xlsx into .csv files |
| 4 | order.csv, Required Model where model is ‘logR’ (e.g. line 15) is not capitalized like in modeling.xlsx (LogR) | Replaced ‘logR’ with ‘LogR’ |
| 5 | customer.csv & employee.docx, Contact Number have duplicated values (e.g. customer line 2 and 13, Employee ID e00001 and e00008) | Did nothing as the error did not affect our queries |
| 6 | There is an extra worksheet in modelling.xlsx named ‘great ideas’ which does not give information about the company. | Excluded it from the dataset. |
| 7 | employee.docx has a Gender column with values ‘Male’ and ‘Female’. However, the OLTP database create table for Employee only allows char(1) data type, meaning values are ‘M’ and ‘F’. | Converted Male and Female in employee.docx to M and F. |
| 8 | order.csv, Order Date and Completion Date, the format for the dates within these columns are not in ‘yyyy-mm-dd’ format, which is the format SQL Server requires for inserting date data. | Changed the format for Order Date and Completion Date columns to ‘yyyy-mm-dd’. |

Explain how you check if your table creation and data insertion are correct.

* We ensured that the table creation and data insertion is correct by selecting all the columns from the fact table and each dimension and seeing if any columns are missing or mislabeled. We checked if the data insertion is correct by selecting the count of rows for the table and each dimension and checking against the dataset files to ensure that all values are the same and inserted and also in the correct formats.

1. Paste your data warehouse design here (database diagram). Write a short description to explain your design such as choice of measurement and levels of details.

|  |  |
| --- | --- |
| DW design | Show all details of the DB diagram. Format your layout and size of tables to ensure nothing is hidden. |
| Explain your design | This data warehouse has a ‘snowflake’ schema where a sub-dimension table ‘ModelType’ is connected to the ‘Model’ dimension table. It has the central OrderFacts table where it has the foreign keys of all the other dimension tables’ surrogate keys and its own surrogate key as a primary key. The facts table also has a Price attribute because it is an important part of analysis for the company. We made the dataset table a dimension table instead of a sub-dimension, hence the Model dimension does not have an DatasetID attribute anymore. We also created an ‘Order’ dimension to store additional order details like ‘RequiredAcc’ and ‘OrderDate’. We created a Time dimension to store each order’s completion date, and it stores attributes ‘Date’, ‘Day’, ‘Month’, ‘Quarter’, ‘Year’ and ‘DayOfWeek’. |

1. Implement the data warehouse you designed in b) using MS SQL server.

Submit DW\_create.sql with all the SQL statements that creates DW tables, and DW\_insert.sql with all the SQL statements that query data from OLTP tables and insert into DW tables.

Briefly explain how you verify that your data warehouse is setup correctly, and that the data is inserted correctly.

To verify that the data warehouse is setup correctly, we used select statements for the facts table and each dimension to check if the attributes are all inside. Afterwards, to check if the data is inserted correctly, we selected the count of the fact table and each dimension and checked against the number of lines in each data file to see if the counts do not match each other.

1. Implement the queries and explain your findings to the 3 questions. You can list no more than 2 findings/queries for each question below. Modify the template accordingly.

* Q1:

What do you want to find out?

We want to find out about what the total profit trend is from 2021 to 2023 to see if the company is thriving and getting business. We also wanted to see the growth rate of the company.

Insert your query here:

A screenshot of a computer code

Description automatically generated

Insert your results here:

A screenshot of a data table

Description automatically generated

Explain what you find based on the results:

\

for the months:

- From this, we can observe certain things. First, we can see that the most profit is always during December with a value of 55,113 in 2021, 112,816 in 2022 and 223,587 in 2023. We can also see that January does the worst in terms of profit gained, with 1,103 in 2021, 1,158 in 2022 and 3,031 in 2023. As compared to December, the total amount of profit earned in all 3 years in January is 5292 while that of December is 391516, which is which a whopping 7300% or 73 times more.

- Moreover, we can also observe a general trend where the profit seems to be a lot better during the end of the year (September, October, November, December) as compared to the start of the year (January, February, March, April). This can be because of the seasonal variations where companies might purposely invest more in AI solutions before the holiday season to optimize sales and customer experience.

What do you want to find out?

We want to find out the amount of profit each model type brings, and see which model earns the most profit for the company.

Insert your query here:

A screenshot of a computer code

Description automatically generated

Insert your results here:

A screenshot of a data

Description automatically generated

Explain what you find based on the results:

- Neural Network has the highest total profit, at 633,131. This is almost 5x the profit or 466.6% of the model with the second highest total profit, Support Vector Machine, at 111,736. This means that Neural Network has historically been the model that has earned the most profit for the company, since its start. As for the rest of the models, we can see that their total profits are around the same, with Random Forest being the lowest at 94,341. This shows that these models make around the same amount of sales

- As for the average profit we can see that Neural Networks clearly has the highest profit, at 465, which is 114 more than the second best, Support Vector Machine at 351. As for the rest of the models, we can also observe that make around the same amount of sales, with most of their average profits being around the 350 mark.

- INSIGHT: It is recommended that the company focus on providing promotions and incentives for the model types other than Neural Network, especially Support Vector Machine, to ensure that all model types are profitable.

* Q2:

What do you want to find out?

We want to find out what accuracy customers usually require from the company’s models and see if the company’s models are over or under performing.

Insert your query here:

A screenshot of a computer program

Description automatically generated

Insert your results here:

A screenshot of a data

Description automatically generated

Explain what you find based on the results:

- From this, we can see the Accuracies required by the customers as compared to the accuracies of the models given to the companies. We can observe that the average of the average accuracy seems to be around 85% while the average of the average customer required accuracy seems to be around 64.5%. From this, we can see that this company ensures that its customer's requirements are met and in fact does more to ensure that they exceed customer expectations. Instead of giving the customer an above average model with around 65% accuracy, this company ensures that all its customers receive a good model with around 80% accuracy. Moreover, this proves that the Company provides a higher quality of service, which could have been a reason as to why the company grew rapidly from 2021 to 2023. This is as they were able to attract more customers and retain current ones due to the quality of service they provide, which helped them distinguish themselves in the market.

What do you want to find out?

We wanted to find out the amount spent by each company, and the amount of accuracy required by these companies.

Insert your query here:

A screenshot of a computer program

Description automatically generated

Insert your results here:

A screenshot of a data

Description automatically generated

Explain what you find based on the results:

- The company who has spent the most money is Nu Company, with the total amount spent being 326,511, which is 145.6 % more spent as compared to the second company, Gamma Enterprises, which only spent 132,948. The company which spent the least amount of money is ABC Company. Which spent only 11,022. As for the total number of models ordered, we can see that Nu Company ordered 742 models which is more than 2x of how many models Gamma Enterprises ordered. Thus we can conclude that Nu Company is the most valuable customer.

- As for Accuracy, we can see that all of the companies have around the same requirements. As for average, we can see that all the companies are in the ballpark of about 64% accuracy. As for highest accuracy, we can see that all the companies are in the ballpark of 77.8%. From this we can say that all the companies only required a model that performs better than average, rather than models close to perfection.

* Q3:

What do you want to find out?

We want to find out how much profit each employee has generated for the company and how many orders have they fulfilled.

Insert your query here:

A screenshot of a computer

Description automatically generated

Insert your results here:

A screenshot of a table

Description automatically generated

Explain what you find based on the results:

- Mia Lewis has the highest amount of total orders fulfilled at 651 orders, and the highest total profit, at 295,219. This shows that Mia fulfilled more than 2.5x the number of orders as compared to Emily Scott and generated a profit which is 238.5% more what Emily generated, which is 87,203. As for the average profit per model, we can see the all the employees are in the range of 340 – 460, with Mia being the highest at 453, and Liam being the lowest at 340.

- Moreover, we can also notice that for some cases, an employee may have sold more models but made lesser profit. For example, Isabella Collins sold 141 models, while Matthew Adams sold 139. Yet, Isabella’s total profit was 53,082, while Matthew’s was 54,726, which is more than 1,5k higher. This can be due to the type of models they sold as some models cost more than others, which can be seen by our queries from Question 1.

- INSIGHT: It is important for the company to provide incentives for employees to earn higher profits for the company, and they should start doing this.

What do you want to find out?

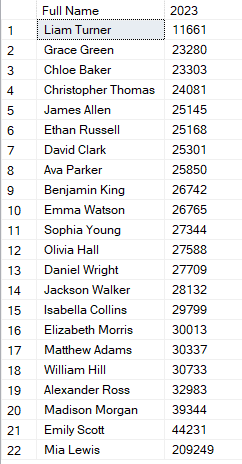
We want to find out which employee is performing poorly in 2023, the company’s most profitable year, by querying each employee’s average profit.

Insert your query here:

A screen shot of a computer

Description automatically generated

Insert your results here:



Explain what you find based on the results:

- This query shows us the performance of each employee during 2023. We decided to use this query as the company thrived the most in 2023, with the highest amount of profits. Thus, this allowed us to see which employee contributed most or contributed little to the success of this company in 2023. Thus, from this, we can see that Mia Lewis definitely contributed the most. However, Liam Turner contributed extremely poorly, with only 11,661 contributed, which is 50% lesser than what the 2nd least contribution was, which is 23,280. Moreover, Liam was not anywhere close the average profit contributed by each employee, which seems to be around 27k. It is suggested that the company consults Liam on his poor performance and maybe provides him help as he may be struggling, to be able to see better results from him in the future.