DATA ANALYTICS-1

ASSIGNMENT-2

PART-1

Team members:

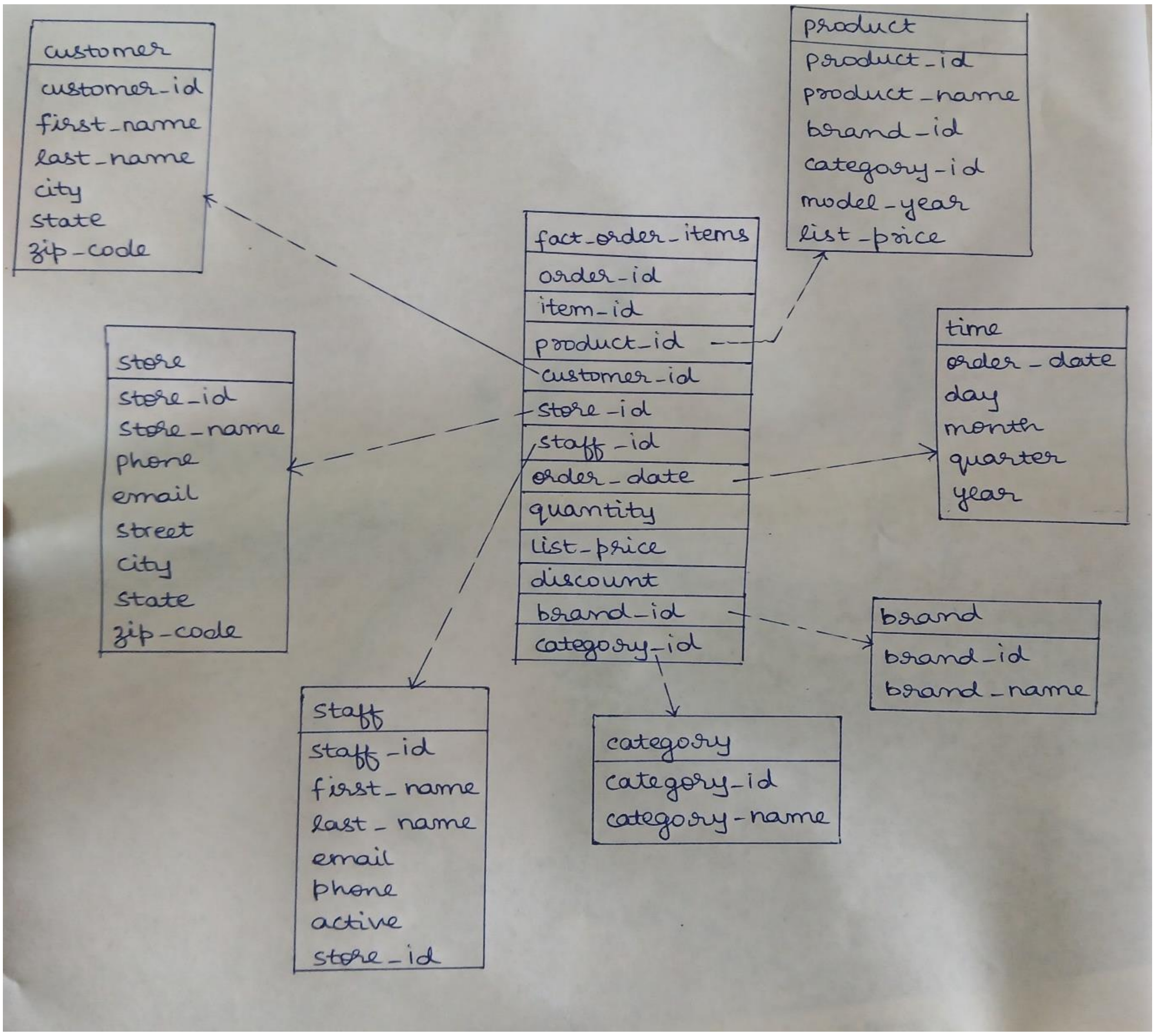
Saloni Goyal (2023115001) - Part 2 (From BUC)

Sachi Thonse Rao (2023101120) (Part1+Part2 AOI)

PART 1

Task1

ER Diagram



A star schema is a widely used data warehousing model that organizes data into a central fact table connected to multiple dimension tables. This structure simplifies analytical queries and enhances performance for aggregation and slicing operations.

For the question, the schema is as follows:

- Fact Table: fact_order_items
- Dimension Tables: dim_product, dim_customer, dim_store, dim_staff, dim_time, dim_brand, dim_category

Task2

Description of the ETL Process

The ETL (Extract, Transform, Load) process is designed to populate a star schema–based data warehouse from raw CSV files. The dataset comprises transaction-level sales data and associated dimensional attributes like products, customers, stores, etc.

The process is implemented entirely in DuckDB, leveraging the read_csv_auto function for efficient ingestion and transformation.

1. Extract

The data is extracted from raw CSV files located in the directory M25_DA_A2_Part1. Each file corresponds to a specific entity in the star schema. The tool used is read_csv_auto ('file_path'). This DuckDB function automatically infers data types and parses the structure.

2. Transform

Several transformations are performed during data ingestion:

Dimension Tables

- Selected only required columns using SELECT column1, column2, ...
- Ensured unique keys (e.g., customer_id, store_id, etc.)
- Used proper typing for numerical, date, and boolean fields

dim_time Table

- Derived date components (day, month, quarter, year) using DuckDB's EXTRACT() function from order_date
- Applied DISTINCT to prevent duplicate entries

Fact Table

- Merged three sources: order_items.csv, orders.csv, and products.csv
- Performed JOINs:
    - orders joined on order_id to add customer, staff, store, and order_date
    - products joined on product_id to include brand_id and category_id (for denormalized storage)
- Calculated derived metrics like quantity * list_price * (1 - discount) during analysis

3. Load

The transformed data is **loaded into the schema tables** using INSERT INTO ... SELECT FROM read_csv_auto(...).

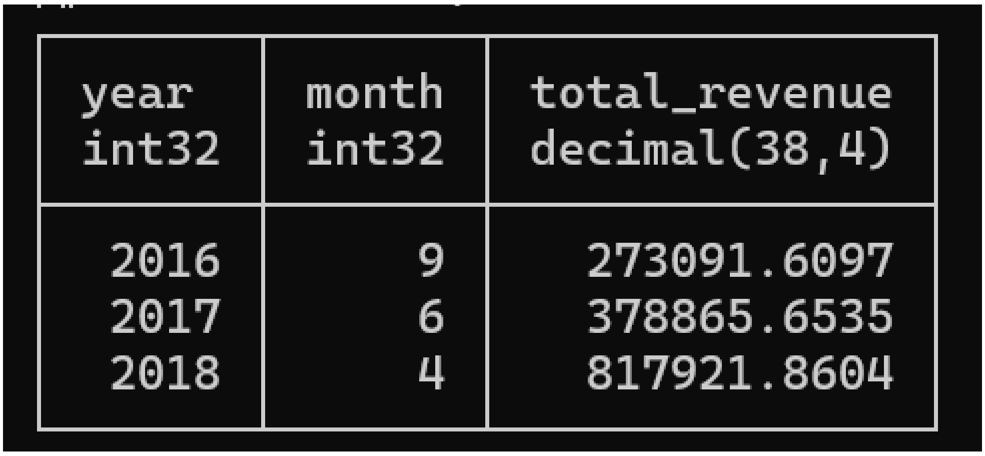
After the ETL process completes:

- The database contains fully populated dimension and fact tables.
- Data is ready for OLAP-style analytical queries, such as roll-up, drill-down, and cube aggregations.

<u>Task3</u>

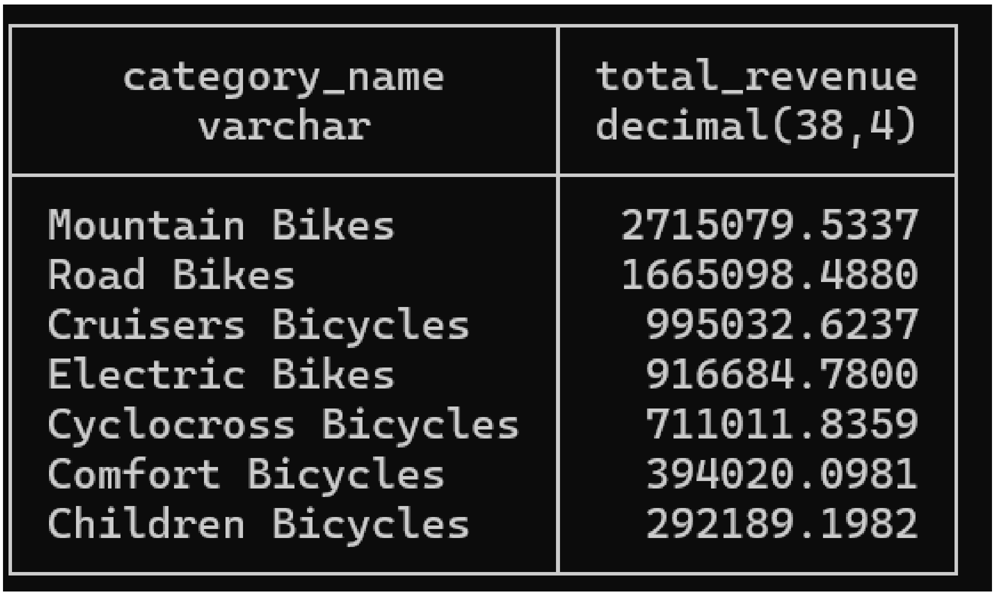
Query Outputs (For the ones where number of rows is large refer to the CSV files provided with the complete output):

Q2.



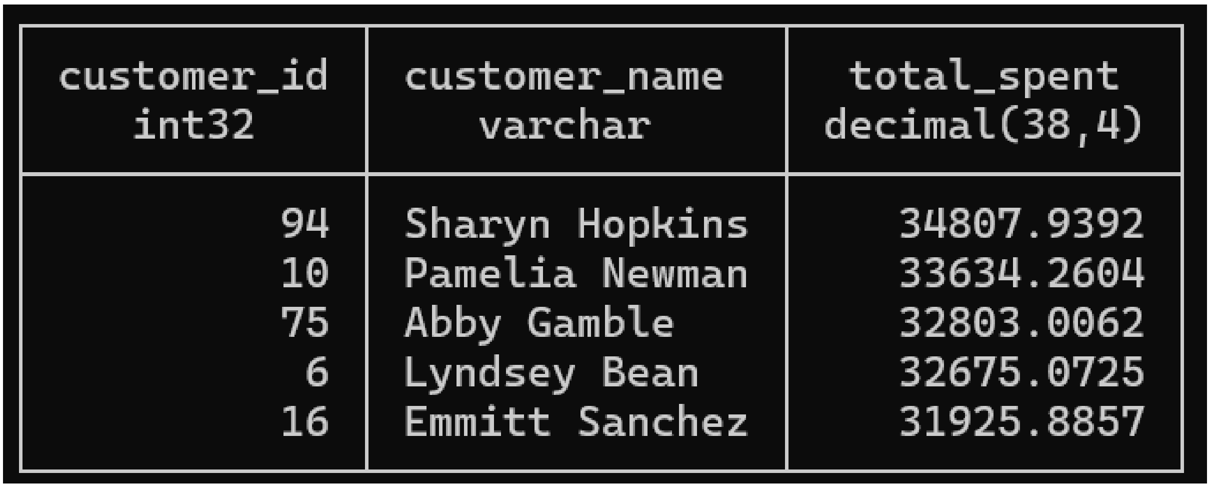
```
┌───────┬───────┬────────────────┐
│ year  │ month │ total_revenue  │
│ int32 │ int32 │ decimal(38,4)  │
├───────┼───────┼────────────────┤
│  2016 │     9 │   273091.6097  │
│  2017 │     6 │   378865.6535  │
│  2018 │     4 │   817921.8604  │
└───────┴───────┴────────────────┘
```

Q3.



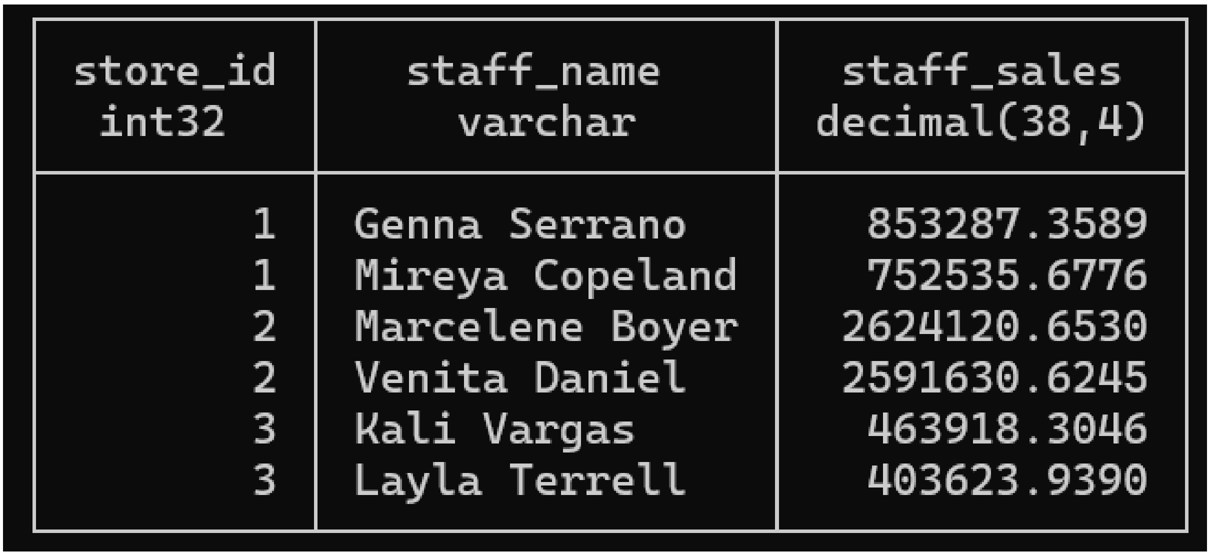
```
┌─────────────────────┬────────────────┐
│    category_name    │ total_revenue  │
│       varchar       │ decimal(38,4)  │
├─────────────────────┼────────────────┤
│ Mountain Bikes      │  2715079.5337  │
│ Road Bikes          │  1665098.4880  │
│ Cruisers Bicycles   │   995032.6237  │
│ Electric Bikes      │   916684.7800  │
│ Cyclocross Bicycles │   711011.8359  │
│ Comfort Bicycles    │   394020.0981  │
│ Children Bicycles   │   292189.1982  │
└─────────────────────┴────────────────┘
```

Q6.



```
┌─────────────┬─────────────────┬────────────────┐
│ customer_id │  customer_name  │  total_spent   │
│    int32    │     varchar     │ decimal(38,4)  │
├─────────────┼─────────────────┼────────────────┤
│          94 │ Sharyn Hopkins  │     34807.9392 │
│          10 │ Pamelia Newman  │     33634.2604 │
│          75 │ Abby Gamble     │     32803.0062 │
│           6 │ Lyndsey Bean    │     32675.0725 │
│          16 │ Emmitt Sanchez  │     31925.8857 │
└─────────────┴─────────────────┴────────────────┘
```

Q7.



```
┌──────────┬──────────────────┬────────────────┐
│ store_id │    staff_name    │  staff_sales   │
│  int32   │     varchar      │ decimal(38,4)  │
├──────────┼──────────────────┼────────────────┤
│        1 │ Genna Serrano    │    853287.3589 │
│        1 │ Mireya Copeland  │    752535.6776 │
│        2 │ Marcelene Boyer  │   2624120.6530 │
│        2 │ Venita Daniel    │   2591630.6245 │
│        3 │ Kali Vargas      │    463918.3046 │
│        3 │ Layla Terrell    │    403623.9390 │
└──────────┴──────────────────┴────────────────┘
```

Analysis of the queries and the data gathered from the operations:

Q1: Total Sales Revenue Drill-Down (Year → Quarter → Month)

- Revenue consistently increased across the years.
- Q3 of most years saw relatively high revenue.
- Example: In 2016, March (Q1) alone saw revenue over 180K, with Q2 and Q3 showing stronger momentum.

Q2: Month with the Highest Sales per Year

- September 2016, June 2017, and April 2018 were peak sales months.
- April 2018 recorded the highest monthly revenue at over ₹817K, suggesting a potential seasonal trend (possibly new product launches or campaigns in Q2).

Q3: Highest Revenue-Generating Product Categories

- The Mountain Bikes category dominates sales, with total revenue exceeding ₹2.7M.

- Followed by:
    - Road Bikes: ₹1.66M
    - Cruisers Bicycles and Electric Bikes also perform well.

Investments and promotions should prioritize Mountain Bikes, the clear revenue leader.

Q4: Drill-Down: Category → Product

- Within Children Bicycles, the top-selling product is:
    - Electra Girl's Hawaii 1 (20-inch) generating over ₹41K.
- Each product contributes incrementally to category totals, but many products still show moderate sales.

There is a clear product-level hierarchy under each category, which helps in micro-targeting promotions.

Q5: CUBE Aggregation (Brand, Category, Year)

- Brand Electra dominates the Children Bicycles and Comfort Bicycles categories.
- The highest revenue for Electra was in 2016, with over ₹154K in Comfort Bicycles alone.

Electra is a consistently high-performing brand across multiple years and categories.

Q6: Top 5 Customers by Total Purchases

- Sharyn Hopkins is the highest spender with over ₹34.8K.
- The top customers are consistent across different cities and likely form the loyal customer base.

Consider rewarding high-value customers with exclusive offers or loyalty programs.

Q7: Staff Performance by Store

- Marcelene Boyer and Venita Daniel from Store 2 dominate with ₹2.6M+ in combined sales.
- Genna Serrano from Store 1 is the top performer there.

Staff training and incentives for high performers like Marcelene and Venita should be scaled.

Q8: CUBE Aggregation (Category, Store, Year)

- Baldwin Bikes (Store) consistently performs well in Children Bicycles, with revenue peaking in 2017.
- Rowlett Bikes also sees significant sales in that category.

Store-level specialization in certain product categories can be exploited further.

1. Mountain Bikes are the most lucrative product category.
2. Brand Electra shows dominance across multiple segments.
3. Store 2 and its staff contribute the largest share of sales.
4. Targeted promotions should focus on high-performing months like April and September.