Project Report

[A24] Data warehousing & ETL

1. Project Overview

The core of this project is to create a data warehouse. This warehouse will collect raw data from various sources, transform it into a usable format, and store it in a data warehouse for analysis and reporting. The process involves several key stages and database layers. The primary goal is to provide a reliable and consistent source of information for business users, enabling them to gain insights, identify trends, and make informed decisions.

The following diagram illustrates the key stages and database layers involved in this process:

2. Database Layers

The project includes the following database layers, each with a distinct purpose and data characteristics:

2.1 group_project_ADM (Administrative / Source DB)

This is the initial source of data. It gathers raw data directly from various business applications. The ADM database acts as the system of record, capturing data as it is generated by operational processes. It includes:

- Transactional data (e.g., orders, sales, invoices): Records of business events.
- Customer information: Details about customers, such as demographics, contact information, and purchase history.
- Product catalog: Information about the products or services offered by the business.
- Time-stamped operational logs: Records of system activities, providing a history of events.
- HR and financial records: Data related to human resources and financial transactions.
- Data is typically in a normalized format: This format reduces redundancy and improves data integrity for operational use.

2.2 group_project_ODS (Operational Data Store)

This layer consolidates data from multiple sources for near real-time operational reporting. It acts as an interim area with some data cleansing. The ODS provides a more integrated view of the data than the source systems, but it is still focused on supporting operational needs. It includes:

- Cleaned and harmonized data: Data from different sources is standardized and made consistent.
- De-duplicated records: Redundant data is removed to ensure accuracy.
- Potentially enriched data with business rules: Business logic may be applied to derive new data elements.
- Suitable for fast, near real-time analytics (not historical analysis): The ODS is optimized for current operational reporting, not long-term trend analysis.

2.3 group_project_STA (Staging Area)

This is a temporary storage area used before data is loaded into the Data Warehouse. It's used for raw data ingestion, transformation, and error checking. The Staging Area provides a space where data can be manipulated without affecting the source systems. It includes:

- Raw or lightly cleaned data: Data in its initial form, or with minimal modifications.
- Data in transit during the ETL/ELT process: Data being processed as it moves between systems.
- Audit logs of data loads: Records of data movement, used for tracking and troubleshooting.
- Temporary tables for validation or transformation: Intermediate data structures used during processing.

2.4 group_project_DWH (Data Warehouse)

This is the central repository for structured, historical data, optimized for Business Intelligence (BI), dashboards, and analytics. The Data Warehouse provides a stable and consistent foundation for decision support. It includes:

- Aggregated, historical data: Data is summarized and stored over time for trend analysis.
- Dimensional models (e.g., star or snowflake schemas): Data is organized into fact and dimension tables for efficient querying.
- Fact and dimension tables (e.g., sales, time, product, customer): Data is structured to represent business entities and their relationships.
- Denormalized and performance-optimized data: Data is structured for fast query retrieval, even with large datasets.
- Used for long-term analytics and reporting: The Data Warehouse supports in-depth analysis of historical trends and patterns.

3. ETL Process

The project implements an ETL process to move and transform data between these layers. The ETL process consists of three key stages, as illustrated in the diagram:

- **Extraction**: Data is extracted from the source systems (group_project_ADM). This involves reading data from various source databases, files, or other data sources. The extraction process should be designed to minimize the impact on the source systems' performance.
- **Transformation**: Data is cleaned, transformed, and prepared in the Staging Area (group_project_STA) and potentially in the ODS (group_project_ODS). This is the most complex stage, involving a series of operations to convert the data into a suitable format for the Data Warehouse. Common transformation operations include:
  - Cleansing (removing errors, inconsistencies): Identifying and correcting or removing inaccurate or incomplete data.
  - Harmonization (standardizing formats): Converting data into a consistent format (e.g., date formats, units of measure).
  - De-duplication: Eliminating redundant records to ensure data accuracy.
  - Enrichment (adding derived data): Calculating new data values based on existing data (e.g., calculating sales totals, profit margins).

- Filtering: Selecting only relevant data based on specific criteria.
- Aggregation: Summarizing data (e.g., calculating daily sales totals from individual transactions).
- Joining: Combining data from multiple sources.

- **Loading**: Transformed data is loaded into the Data Warehouse (group_project_DWH) for analysis. This involves writing the transformed data into the target database. The loading process should be optimized for performance to handle large volumes of data efficiently.

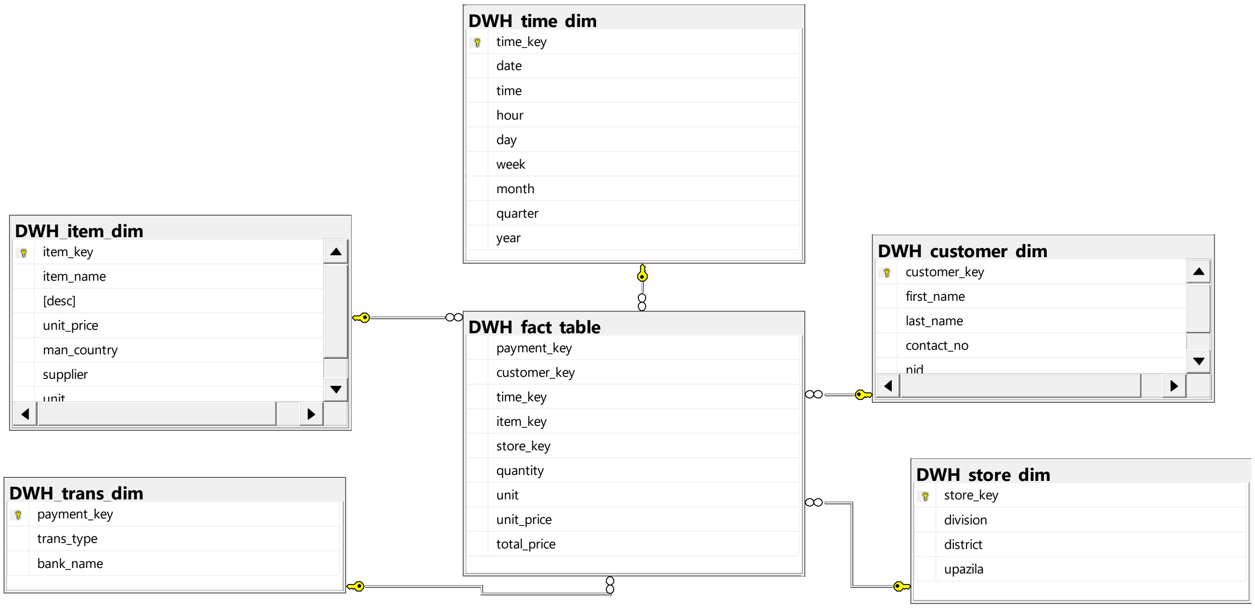
4. Data Flow and ER Diagrams

In summary, the data flows as follows:

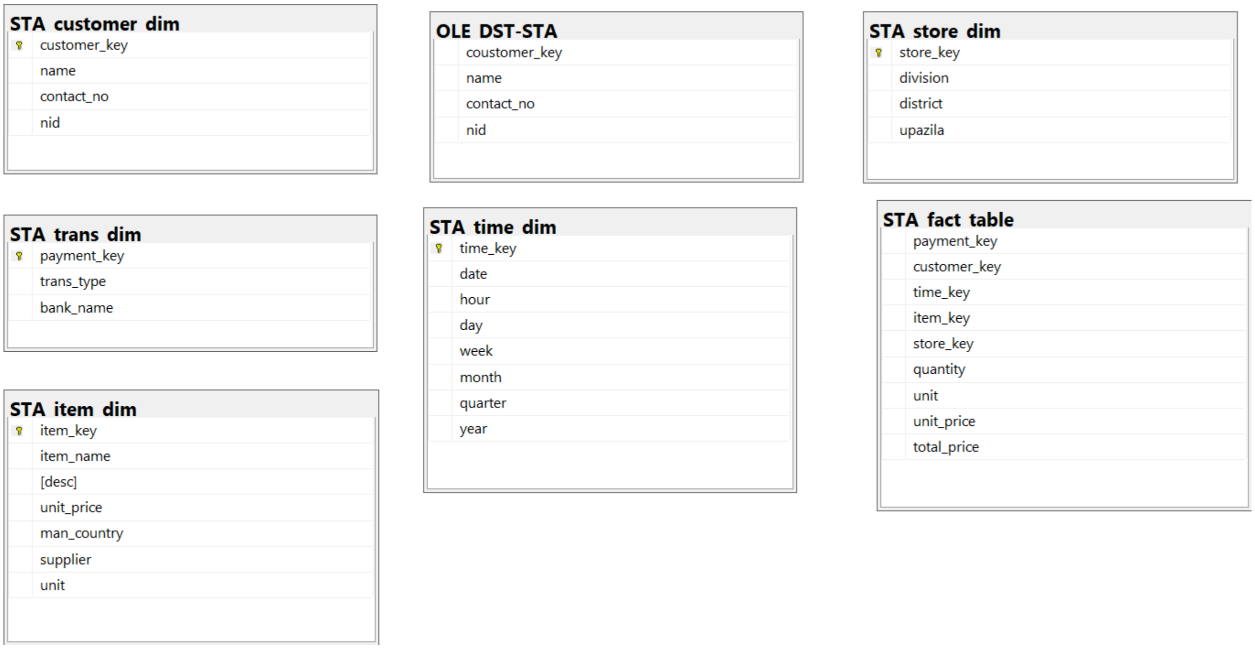
- Raw data is collected in the group_project_ADM.
- Data is consolidated and cleaned in the group_project_ODS for operational reporting.
- Data is staged and transformed in the group_project_STA.
- Cleaned and transformed data is loaded into the group_project_DWH for long-term analysis and business intelligence.

The ER Diagram is showed below:

- ER Diagram of group_project_DWH



- ER Diagram of group_project_STA



STA customer dim
- customer_key
- name
- contact_no
- nid

OLE DST-STA
- coustomer_key
- name
- contact_no
- nid

STA store dim
- store_key
- division
- district
- upazila

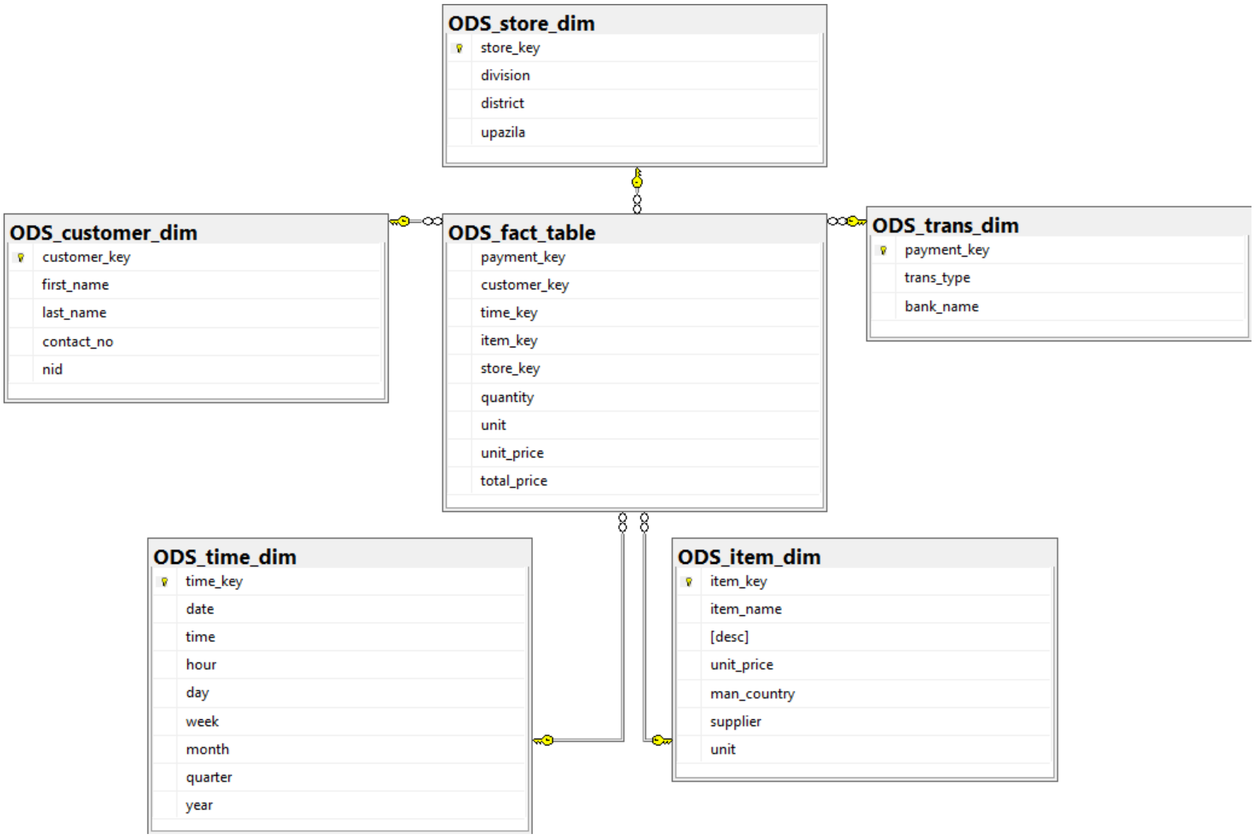
STA trans dim
- payment_key
- trans_type
- bank_name

STA time dim
- time_key
- date
- hour
- day
- week
- month
- quarter
- year

STA fact table
- payment_key
- customer_key
- time_key
- item_key
- store_key
- quantity
- unit
- unit_price
- total_price

STA item dim
- item_key
- item_name
- [desc]
- unit_price
- man_country
- supplier
- unit

- ER Diagram of group_project_ODS



ODS_store_dim
- store_key
- division
- district
- upazila

ODS_customer_dim
- customer_key
- first_name
- last_name
- contact_no
- nid

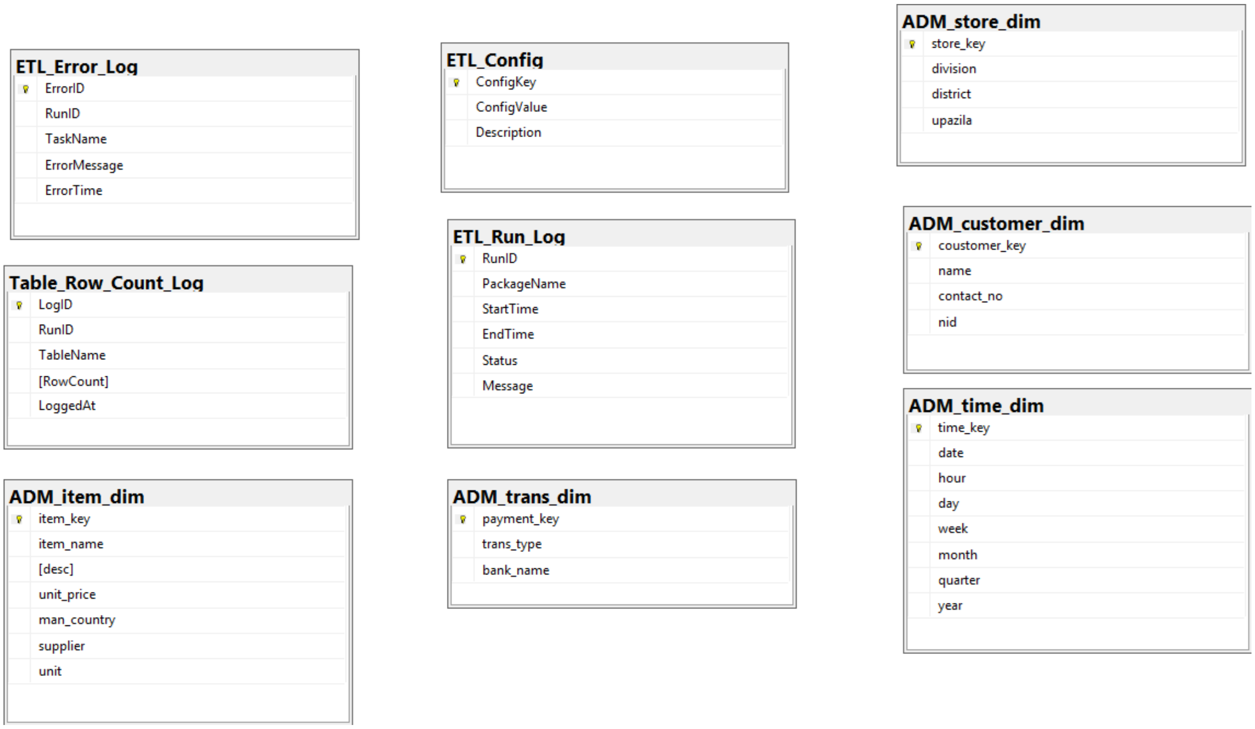
ODS_fact_table
- payment_key
- customer_key
- time_key
- item_key
- store_key
- quantity
- unit
- unit_price
- total_price

ODS_trans_dim
- payment_key
- trans_type
- bank_name

ODS_time_dim
- time_key
- date
- time
- hour
- day
- week
- month
- quarter
- year

ODS_item_dim
- item_key
- item_name
- [desc]
- unit_price
- man_country
- supplier
- unit

- ER Diagram of group_project_ADM



ETL_Error_Log
ErrorID
RunID
TaskName
ErrorMessage
ErrorTime

ETL_Config
ConfigKey
ConfigValue
Description

ADM_store_dim
store_key
division
district
upazila

Table_Row_Count_Log
LogID
RunID
TableName
[RowCount]
LoggedAt

ETL_Run_Log
RunID
PackageName
StartTime
EndTime
Status
Message

ADM_customer_dim
coustomer_key
name
contact_no
nid

ADM_item_dim
item_key
item_name
[desc]
unit_price
man_country
supplier
unit

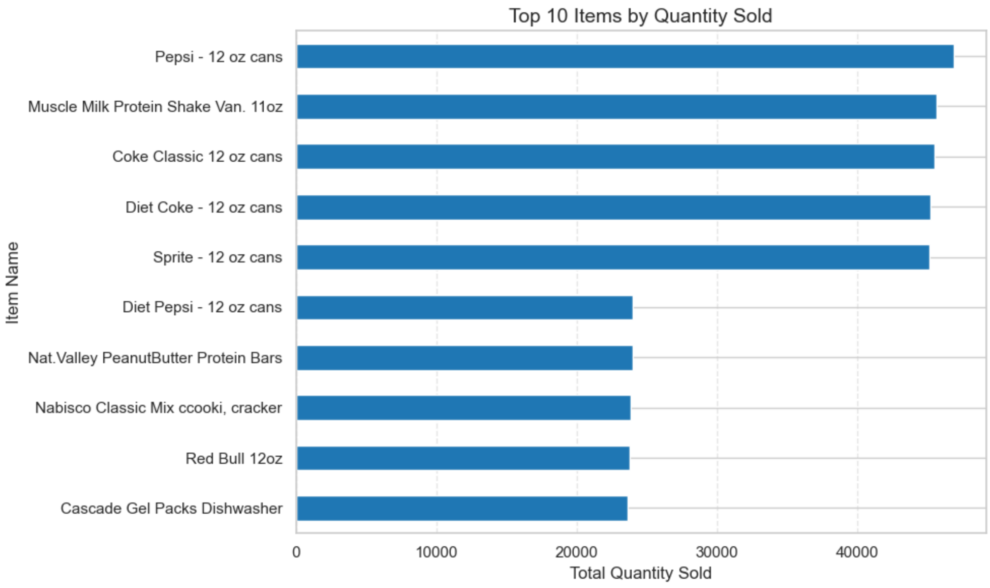
ADM_trans_dim
payment_key
trans_type
bank_name

ADM_time_dim
time_key
date
hour
day
week
month
quarter
year

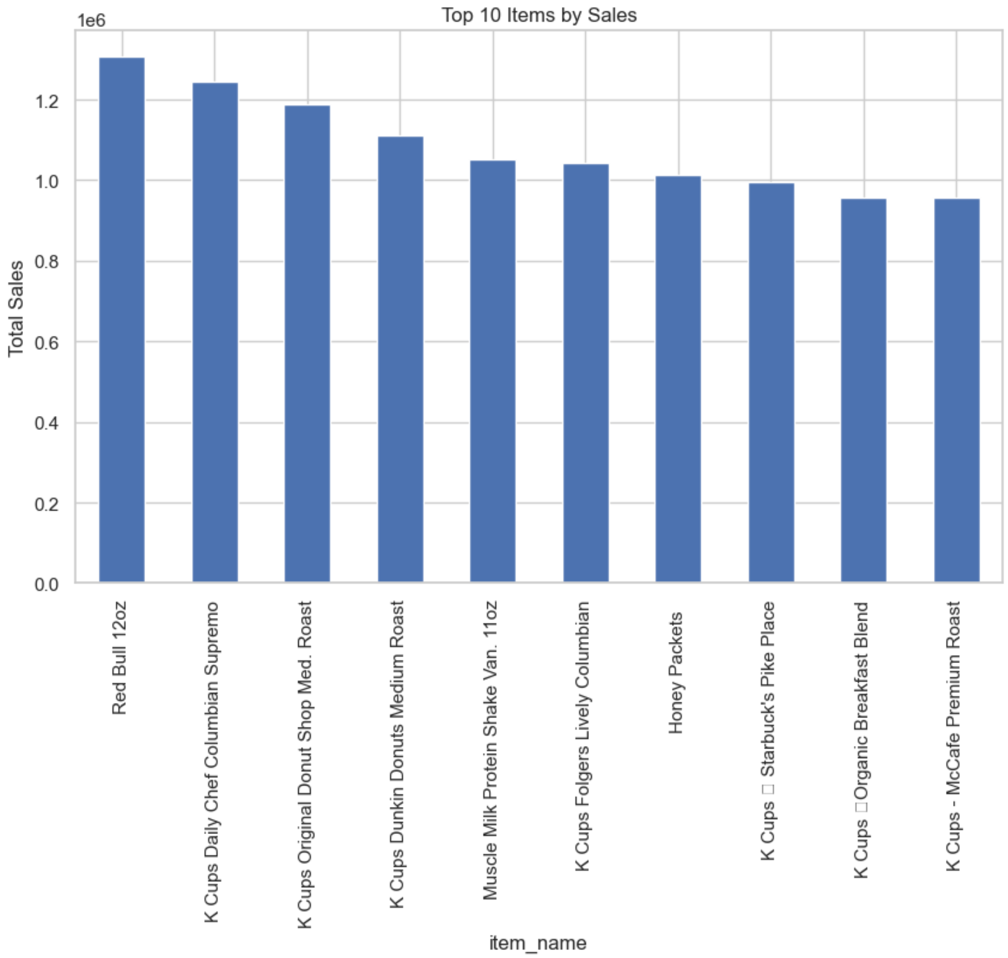
5. Data Analytics

Some practical answers which are discovered from the data warehouse are showed below.

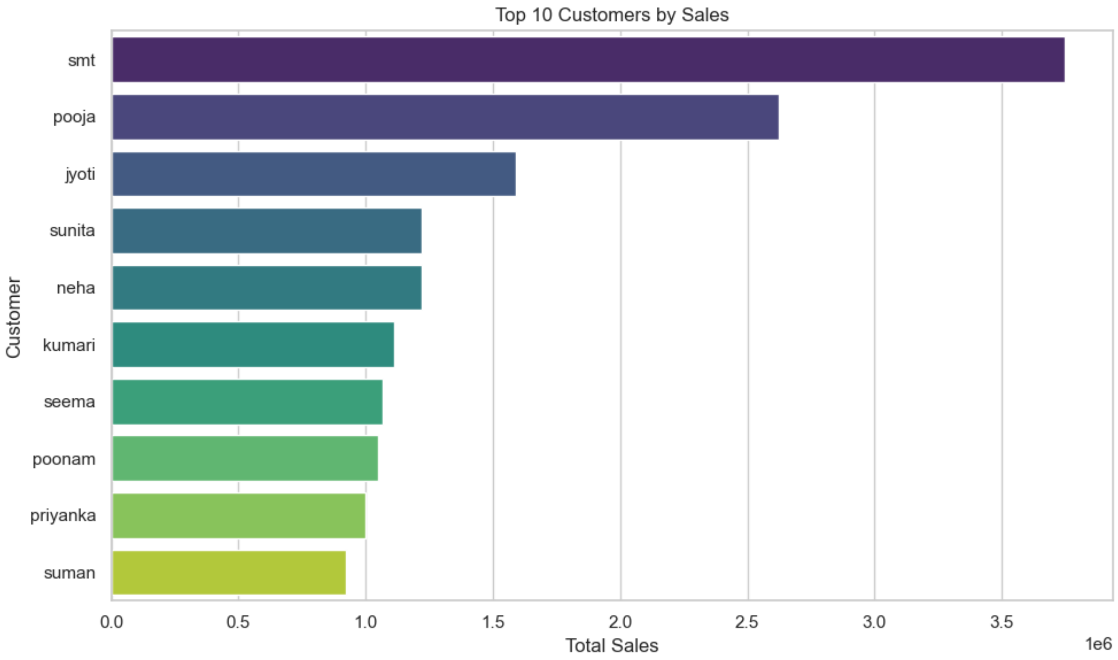
- Question 1: Key Performance Metrics (KPI): 105024137.75
- Question 2: Top 10 product based moving (measure in quantity)



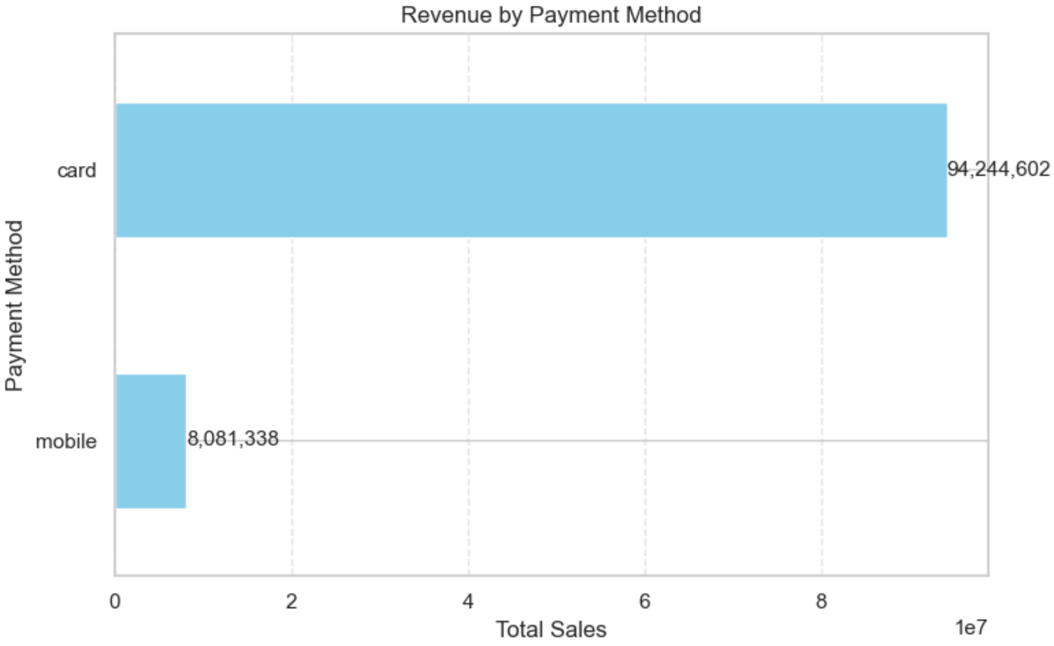
- Question 3: Top 10 product by sales



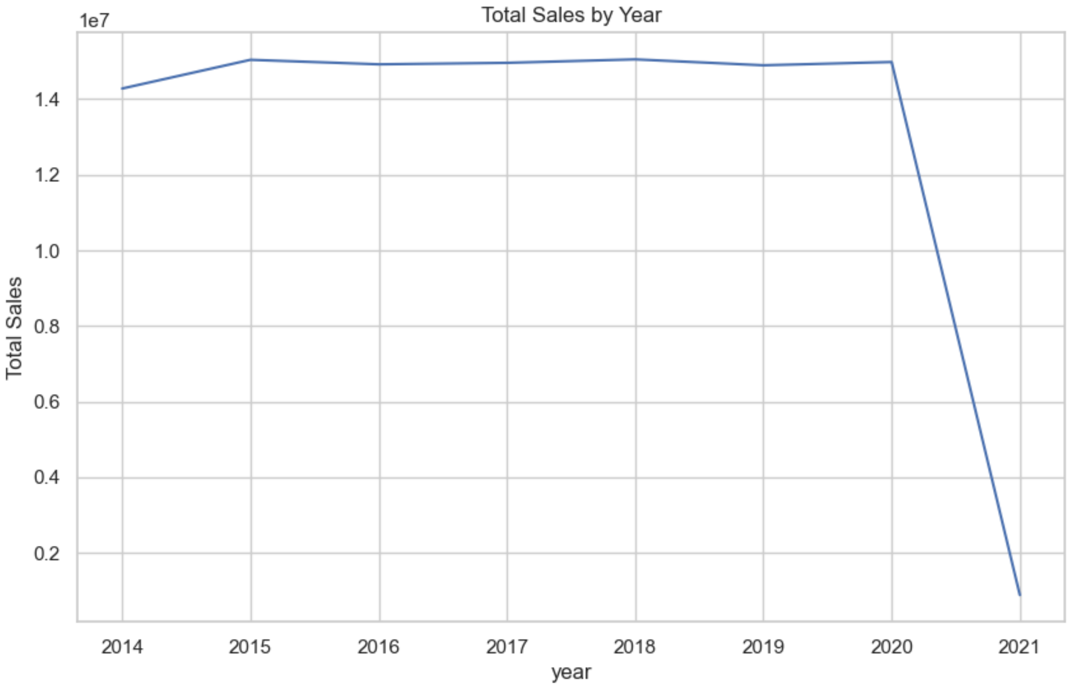
- Question 4: Top 10 customers



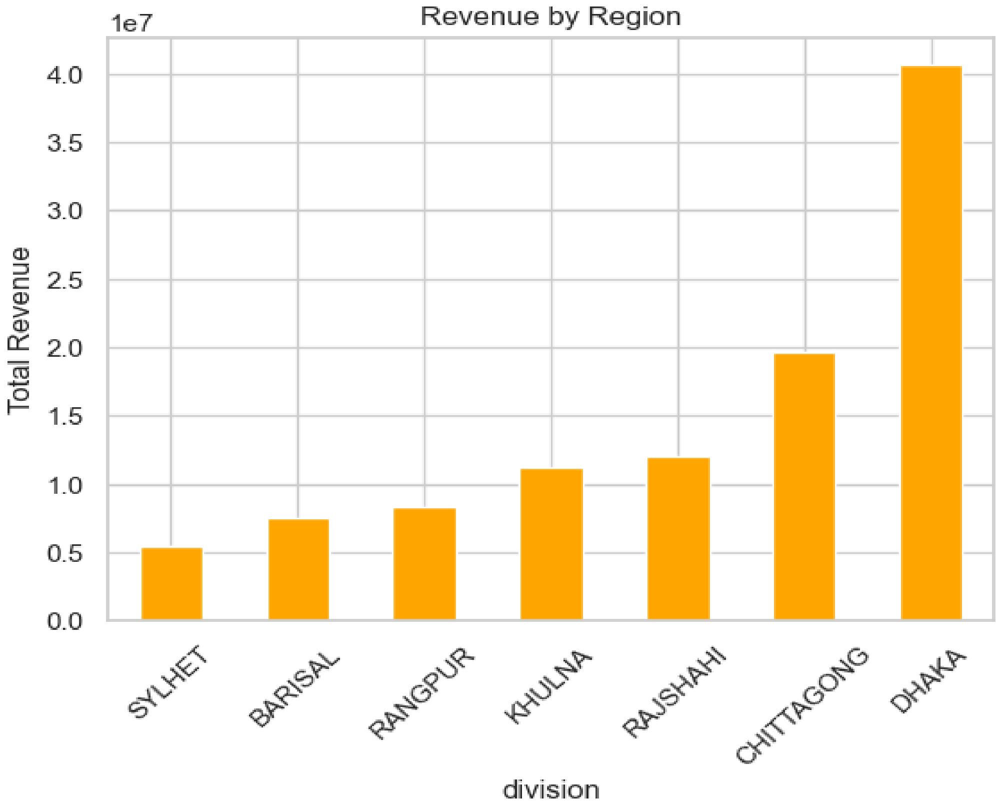
- Question 5: Sales by payment method



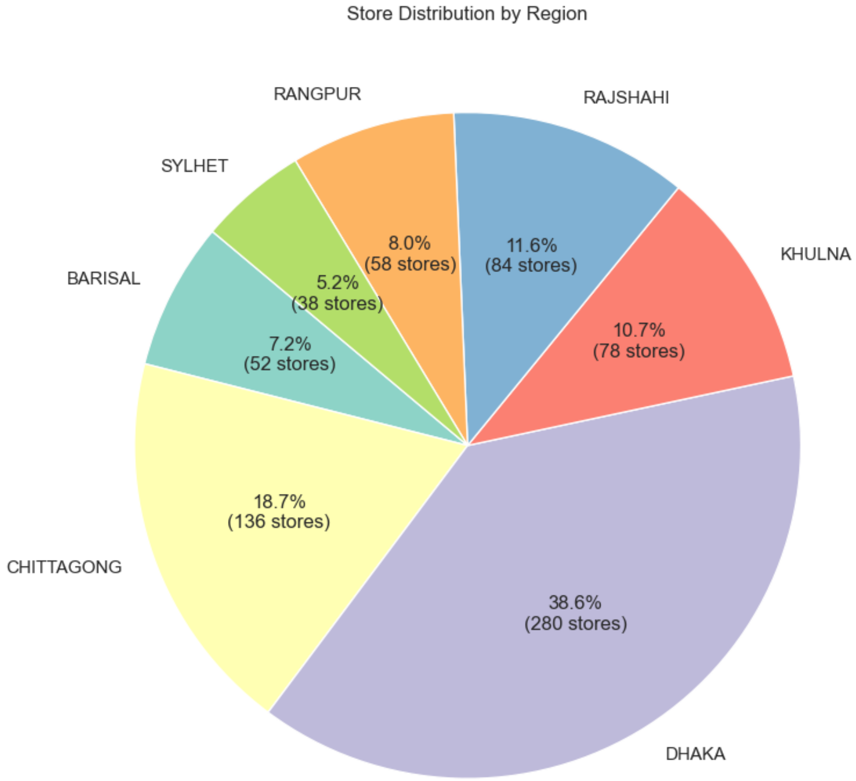
- Question 6: Total sales by year



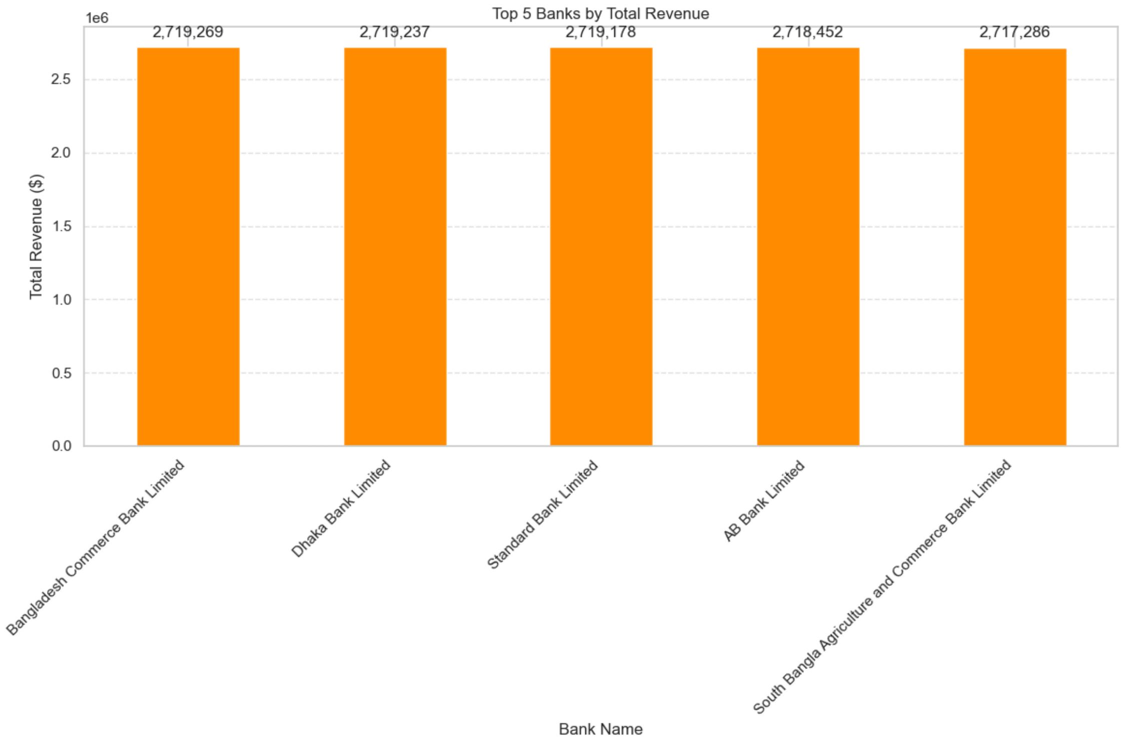
- Question 7: Revenue by regions



- Question 8: Store distribution by region (count and percentage)



Question 9: Banks handle the most transactions



5. Comparing raw and cleaned data

File Name	Original Rows	Cleaned Rows	Original Nulls	Cleaned Nulls	Rows Difference	Nulls Difference
customer_dim.csv	9191	8524	27	0	-667	-27
time_dim.csv	99999	99999	0	0	0	0
trans_dim.csv	39	38	1	0	-1	-1
item_dim.csv	264	263	1	0	-1	-1
store_dim.csv	726	726	0	0	0	0
fact_table.csv	1000000	900188	3723	0	-99812	-3723