SALES DATA ANALYTICS DOCUMENTATION

Prepared by: Ahmed Essam Eldin

SALES DATA ANALYTICS

1- Data Transformation:

1.1 Data Ingestion:

- Reading Data from 2 JSON us Pandas.
- Forecast table consists of 4 Columns {CountryRegion, Brand, Forecast, Year}.
- Sales table consists of 18 Columns.

```
raw_sales.info()
✓ 0.7s
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 298246 entries, 0 to 298245
Data columns (total 18 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   ProductKey     298246 non-null  int64
 1   Product Name   298246 non-null  object
 2   Brand          298246 non-null  object
 3   Color          298246 non-null  object
 4   Subcategory    298246 non-null  object
 5   Category       298246 non-null  object
 6   CustomerKey    298246 non-null  int64
 7   Customer Code  298246 non-null  object
 8   Name           29797 non-null   object
 9   Education      29797 non-null   object
 10  Occupation     29797 non-null   object
 11  Continent      298246 non-null  object
 12  City           298246 non-null  object
 13  State          298246 non-null  object
 14  CountryRegion  298246 non-null  object
 15  OrderDate      298246 non-null  object
 16  Quantity       298246 non-null  int64
 17  Net Price      298246 non-null  float64
dtypes: float64(1), int64(3), object(14)
memory usage: 41.0+ MB
```

```
raw_forecast.info()
✓ 0.0s
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 33 entries, 0 to 32
Data columns (total 4 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   CountryRegion  33 non-null     object
 1   Brand          33 non-null     object
 2   Forecast       33 non-null     int64
 3   Year           33 non-null     int64
dtypes: int64(2), object(2)
memory usage: 1.2+ KB
```

1.2 Cleaning:

- Removing white spaces and extra char using strip().
- only 3 columns in Sales table contain 268449 null values which are {Name, Education, Occupation}.
- As the columns are type of text then the most appreciate way is Replacing None values with “Unknown” value.
- Dropping duplicates from Sales table that contains 218008 duplicated value.
- Change “OrderDate” data type in sales table to DateTime.
- Computing SalesAmount = Quantity * NetPrice.

```
df_sales["Education"].value_counts()
✓ 0.0s
Education
Unknown             50441
Partial College      8071
Bachelors            7124
High School          6541
Graduate Degree      4942
Partial High School  3119
Name: count, dtype: int64
```

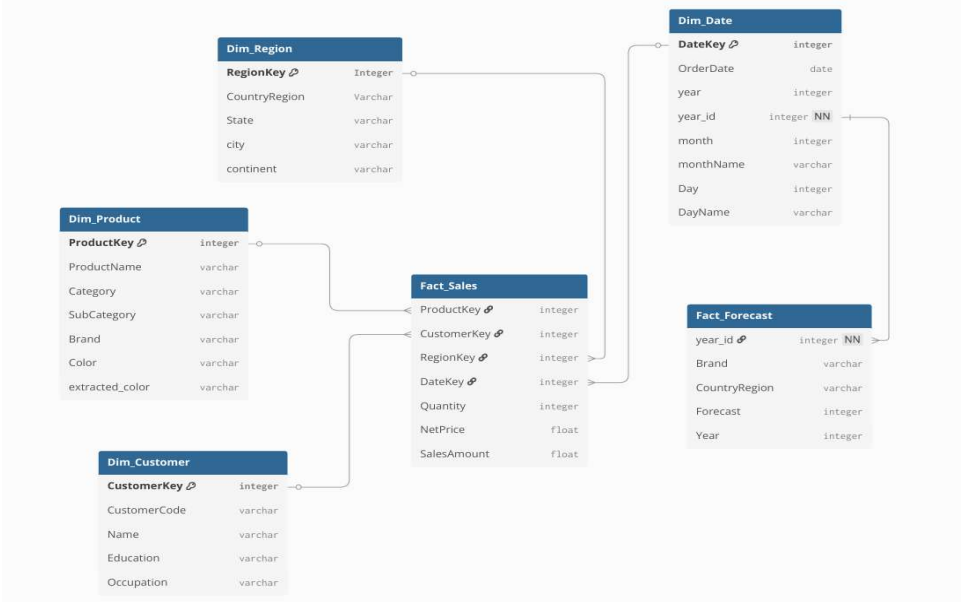
1.3 Loading the cleaned tables into csv files.

2- Data Model Architecture:

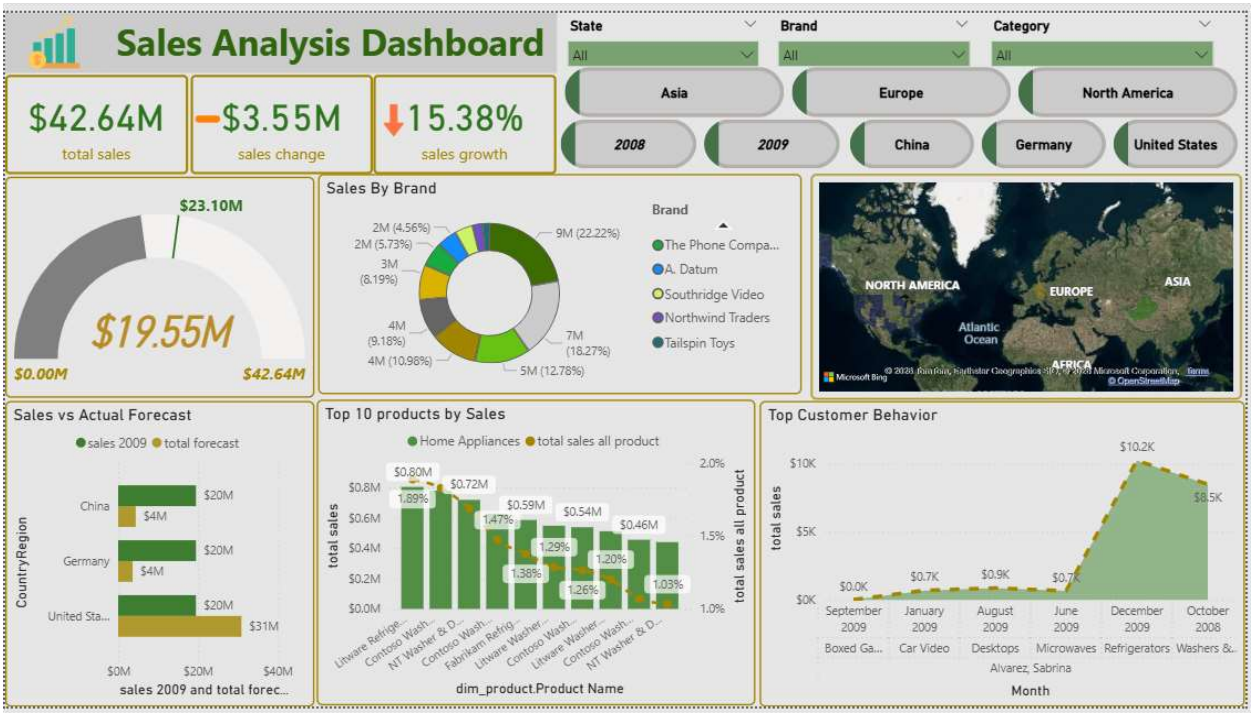
The following tables have been created and saved as CSV files:

Table Type	Table Name	Columns	Description
Fact	**fact_sales**	**DateKey**, **ProductKey**, **CustomerKey**, **RegionKey**, Quantity, NetPrice, SalesAmount	Contains transactional quantities and sales amounts.
Fact	**fact_forecast**	CountryRegion, Brand, Forecast, Year	Contains the 2009 targets for Country and Brand.
Dimension	**dim_product**	**ProductKey**, Category, SubCategory, extracted_color, Color, Brand, ProductName.	Product details, including Category and Brand.
Dimension	**dim_customer**	**CustomerKey,** CustCode, Name, Education, Occupation.	Customer demographics and attributes.
Dimension	**dim_region**	**RegionKey,** CountryRegion, State, City, Continent.	Geographic hierarchy: Continent, Country, State, and City.
Dimension	**dim_date**	**DateKey**, OrderDate, Month, Year.	The central bridge connecting both facts.

- Now The Star Schema Contains 2 Fact tables and 4 Dimensions tables.
- Extracting Year, Month, Day from OrderDate column in Dim_Date.
- Extracting Color from ProductName in Dim_Product table using split with space delimiter and choosing the last element as extracted_color.
- The Fact Forecast table links to Dim_Date via year_id and to Dim_Region via CountryRegion.



3- Dashboard:



3.1 Executive Summary (KPIs):

- Total Sales: $42.64M, representing the sum of SalesAmount from the sales fact table.

- Sales Change: A negative variance of -$3.55M.

- Sales Growth: A decline of 15.38%.

- Gauge Chart: Shows a current progress of $19.55M against a target of $42.64M.

3.2 Sales by Brand & Product:

- Sales by Brand: A donut chart breaks down performance by brands such as Contoso (23.13%), Tailspin Toys, and Litware.

- Top 10 Products: A bar chart ranks individual products (e.g., Litware Refrigerators, Contoso Washers) by total sales.

3.3 Geographic Distribution:

Using the Dim_Region table, the dashboard filters and maps data by location:

- Interactive Map: Visualizes global sales across North America, Europe, and Asia.

- Region Slicers: Users can filter the entire dashboard by State, Country/Region (e.g., China, Germany, United States), or specific continents.

3.4 Comparative & Trend Analysis:

- Sales vs. Actual Forecast: A horizontal bar chart compares 2009 sales against the total forecast for China, Germany, and the United States, utilizing the relationship between Fact Sales and Fact Forecast.

- Top Customer Behavior: A line chart tracks total sales over time (Months), derived from the Dim_Date table. It highlights peak sales activity in December 2009 ($10.2K).