# Debre Berhan University

## Computing College

## Department of Software Engineering

## Fundamentals of Big Data Analytics and BI

# E-commerce Behavior Data Transformation and Visualization Report

**By:**

**Yewoynhareg Mulugeta**

**0035/13**

**Submitted to: Derbew Felasman(MSc)**

**Submission Date: 06/06/2017 E.C**

# Table of Contents

# E-commerce Behavior Data Transformation and Visualization Report

## 1. Introduction

This report presents an end-to-end data pipeline designed to process e-commerce behavioral data from a multi-category store and challenges that occurred during the process. The pipeline involves data extraction, transformation, loading (ETL), and visualization to support business decision-making. This data was extracted from the Kaggle Data Provider Website then the dataset was transformed using Pandas, and then it is loaded to PostgreSQL, and key insights are visualized using Microsoft Power BI.

## 2. Data Pipeline Overview

### 2.1 Data Extraction

The e-commerce dataset was extracted from Kaggle Data Provider. The dataset was extracted using Python and Pandas.

The dataset originally consisted of rows but due to the large volume of the dataset, my personal computer kept crushing. The RAM required to transform this data was a minimum of 16GB while my computer is of 8GB. Although I have tried to use the batch processing method, it still was too large for my computer's capacity. This forced me to minimize the size of the dataset from over 13,776,051 to 1,010,000 rows and I also minimized the columns from 7 to 5 columns. It includes attributes such as `event_time`, `event_type`, `product_id`, `brand`, and `price`.

### 2.2 Data Transformation

The data was cleaned to remove duplicates, handle missing values, and convert data types where necessary to match the PostgreSQL schema. The dataset had 1,010,000 records before transformation and become a dataset of 990,351 records after cleaning.

### 2.3 Data Loading

A connection to PostgreSQL was established using the command prompt:

-A database named ‘ecommercek_db’ was created

- A structured table named `ecommerce_data` was created in ‘ecommercek_db’ database

- 990,351 records were successfully loaded into PostgreSQL.

SQL Table Schema:

```
CREATE TABLE ecommerce_data (
   id SERIAL PRIMARY KEY,
   event_time TIMESTAMP,
   event_type VARCHAR(50),
   product_id VARCHAR(50),
   brand VARCHAR(100),
   price NUMERIC(10,2));
```

## 3. Database Connection and Data Storage

### 3.1 Connection Verification

- Verified the PostgreSQL connection using:

```
psql -U postgres -d ecommercek_db
```

- Checked if the table was successfully created using

```
\d ecommerce_data
```

- Queried the total number of records:

```
SELECT COUNT(*) FROM ecommerce_data;
```

Result: 990,351 rows

### 3.2 Data Validation

Ensured data integrity using:

```
SELECT COUNT(*) FROM ecommerce_data WHERE event_time IS NULL OR event_type IS NULL;
```

Result: No null values found.

Verified distinct event types:

```
SELECT event_type, COUNT(*) FROM ecommerce_data GROUP BY event_type;
```

## 4. Business Intelligence Visualization

### 4.1 BI Tool Used

Power BI was used for visualization. I connected Power BI to my PostgreSQL and imported the cleaned ecommerce dataset.

### 4.2 Key Insights

**1. Sales Trends Over Time** - A time-series analysis was performed to identify peak sales periods. The dataset was of a single month so the sales trends over a month were analyzed.

This was visualized using Line Chart because they are best for showing trends over time. On the X-Axis, event time was represented; on the Y-Axis, price was represented. This was filtered by purchase event type in order to exclude views and cart additions.

**2. Top-selling Products** – Identified top-selling products and brands.

This was visualized using Bar Chart because they are ideal for categorical comparisons. On the X-Axis, product id was represented; on the Y-Axis, price was represented. This was also filtered by purchase event type.

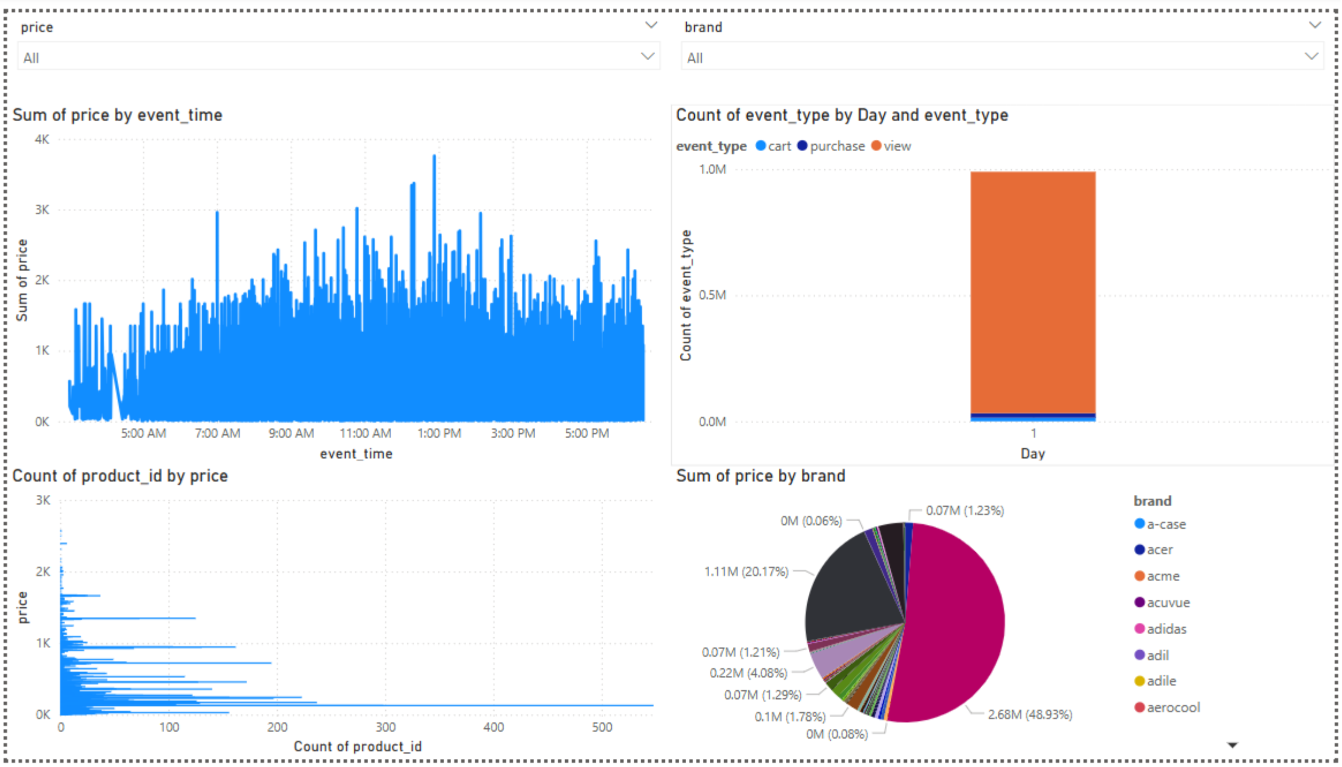
**3. Brand Performance** – Helps analyze which brands generate the most revenue.

This was represented using Pie Chart because they show proportions. On the Category section, brand was represented; on the values section, price was represented. This was also filtered by purchase event type.

**4. Customer Interaction Trends** – Helps track customer behavior trends.

This was represented using Stacked Column Chart because it helps to because different event types over time. On the X-Axis, event time was represented; on the Y-Axis count of event type was represented and on the Legend section event type was represented.

## 4.3 Sample Visualization



To enhance the interactivity of the dashboard, slicers and filters were used. The price slicer allows users to filter data by product price range. The Brand slicer filters all visuals to show data related to a specific brand. For the sake of easy selection and better user experience, the slicers were set dropdown format.

## 5. Conclusion

This project implemented a data pipeline that performs ETL (Extraction, Transformation and Loading) of e-commerce behavioral data.

The dataset was cleaned and reduced to a manageable size for analysis using pandas. PostgreSQL was used for structured data storage or loading. Power BI dashboards provided insights on sales trends, top-selling products, brand performance, and customer interactions. It also included price and brand slicers so that the user can filter the visualized data be price and brand respectively.