6.1: Sourcing Open Data

Shirin Younesian

Data Source

Dataset Name

Global_Superstore2

Dataset Link

https://www.kaggle.com/datasets/apoorvaappz/global-super-store-dataset?select=Global_Superstore2.csv

Data Sourcing

This dataset is an external and open data source

Data Contents

Data contains details of the order done online by people across the globe in the time frame 1-jan-2011 to 31-dec-2014.

Data Volume

Initial dataset consists of 24 columns and 51290 rows

Data Cleaning (Data Integrity & Consistency)

Dropping Columns

Excluding the ‘customer name’ & ‘postal code’ columns from the dataset, to protect the privacy of customers` information.

Dropping these irrelevant columns: ‘order priority’, ‘region’, ‘ship date’ & ‘product name’ (there is ‘product id’ column in dataset).

Missing Data

There is not any missing data in the dataset.

Renaming Columns

The names of ‘Row ID’, ‘Order ID’ & ‘Order Date’ columns, were changed to ‘id’, ‘order_id’ & ‘order_date’, according to the naming convention.

Duplicate Data

There is not any duplicate data in the dataset.

Transforming Data

In the order date column, dates were stored in different formats, which were changed and transformed to the ‘dd/mm/yyyy’ format.

Data Profile

Column Name	Description	Data Type	Time Variant
id	Identity number for each row in the dataset	Quantitative	Invariant
Order_id	Identity number for each order in the dataset	Quantitative	Invariant
Order_date	Date of order	Quantitative	Variant
Ship_mode	Kind of shipment	Qualitative	Invariant
Customer_id	Identity number for each customer in the dataset	Quantitative	Invariant
segment	Kind of customer	Qualitative	Invariant
city	City name	Qualitative	Invariant
state	State name	Qualitative	Invariant
country	Country name	Qualitative	Invariant
market	Market regional classification	Qualitative	Invariant
Product_id	Identity number for each product in the dataset	Quantitative	Invariant
category	Product classification	Qualitative	Invariant

Sub_category	Product sub classification	Qualitative	Invariant
sales	Amount of sales	Quantitative	Invariant
quantity	Amount of product order	Quantitative	Invariant
discount	Percentage discount for each row of the order	Quantitative	Invariant
profit	Amount of profit or loss for each row of the order	Quantitative	Invariant
Shiping_cost	Cost of order shipping	Quantitative	Invariant
year	Year of order date	Quantitative	Variant
month	Month of order date	Quantitative	Variant

Data Limitation and Ethics

The time frame of the dataset is from 2011 to 2014 and it is not possible to update the dataset.

The ‘customer name’ & ‘postal code’ columns from the dataset were excluded to protect the privacy of customers` information.

Questions to explore

1- Which customer segment is most profitable in each year?
2- Which customer segment registers the most orders each year?
3- Is there a relationship between the customer segment and the amount of profit?
4- How are the customers distributed across the market?
5- Which product category is the most profitable in each year?
6- What categories of products are the most ordered in each market?
7- What months are the most profitable in each market?
8- What is the trend of sales and sales profit between 2011 to 2014?