6.1: Sourcing Open Data

Paul Maden.
Last Update: 11/01/2025

Project_Brief.pdf

Data Source:

A summary of your data source

The data chosen for this project was compiled by the open data platform *Kaggle*.

This dataset (‘***Super Market dataset***‘) was last updated 2 months ago and provides a detailed information of a superstore’s operations, encompassing transactional data, customer segmentation, and return trends. The temporal coverage is between 2021 – 2024, with geospatial coverage of the U.S.A and Canada.

Authors summary:

It has been structured into two sheets—Orders and Returns—offering a holistic view of retail dynamics. The inspiration behind this dataset lies in its potential to unlock actionable business insights, from sales analysis and customer behaviour trends to return rate predictions and product performance evaluation. Designed to spark creativity and enable practical applications, this dataset is perfect for analysts, students, and data enthusiasts looking to dive into retail analytics and beyond.

Data Variables Summary:

Sheet:
"Retail_Superstore"

Variable	Type	Description
Row ID	Integer	Unique identifier for each record.
Order ID	String	Unique identifier for each order.
Order Date & Ship Date	Date Time	Dates indicating when the order was placed/shipped.
Ship Mode	String	Specifies the shipping mode (e.g., Standard Class).
Customer ID	String	Unique identifier for each customer.
Customer Name	String	Name of the customer.
Segment	String	Customer segment (e.g., Consumer, Home Office).
Country/Region	String	Country or region of the order.
City	String	City where the order was placed.
State	String	State where the order was placed.
Postal Code	Integer	Postal code for the order location.
Region	String	Regional division (e.g., East, Central).
Product ID	String	Unique identifier for the product.
Category	String	Product category (e.g., Office Supplies).
Sub-Category	String	Sub-category of the product.
Product Name	String	Full name of the product.
Sales	Float	Transactional data on sales amount.
Quantity	Integer	Number of units purchased.
Discount	Float	Discount applied to the order.
Profit	Float	Profit margin per order.

Sheet:
"Retail_Superstore_Returns"

Variable	Type	Description
Returned	String	Indicates whether the product was returned ("Yes").
Order ID	String	Matches the order IDs in the other sheet for correlation.

An explanation for why you have chosen this data set.

I have chosen this dataset because my career spans 20 years in the retail industry, a sector in which I aim to continue working as a data analyst. This dataset allows me to demonstrate my analytical skills using industry-relevant data, showcasing my ability to generate insights and drive data-informed decision-making within the retail sector.

The data includes both continuous variables (Sales, Profit, Quantity, Discount) and categorical variables (Segment, Ship Mode, Region). It contains a geographical component with multiple regions, states, and cities. Also, the data is structured for sales performance analysis with the ability to link returns via the common Order ID. It also scored 100% for '*Usability*', with over 2000 downloads.

Clean the data. Conduct some basic data cleaning and consistency checks in Jupyter to ensure your data is ready for further analysis.

Excel – Initial basic checks completed (*before* loading into Jupyter):

a. Remove Duplicates using Data > Remove Duplicates
- No Duplicates found

b. Handle Blank Cells using Find & Select > Go to Special > Blanks.
- No cells found

c. Consistent Date formatting using Number > Short date.
- Completed on 'Order Date' & 'Ship Date.'

Jupyter – Checks to ensure the Dataset is cleaned: Retail_Superstore_Cleaned .ipynb

Missing Values:

- Using .isnull().sum()
- Check for any missing values in key columns such as Order ID, Customer ID, Sales, Quantity.
- If missing, either remove or impute based on the context.

Duplicated Records:

- Using .duplicated().sum()
- Identify and remove duplicate rows, especially checking columns like Order ID and Product ID.

Data Type Consistency:

- Checked for mixed data types: Using
```
for col in df_super.columns.tolist():
        weird = (df_super[col].map(type) !=type (df_super[col].iloc[0])).any()
```
- Ensure that numerical columns like Sales, Quantity, Profit are stored as numeric types.
- Verify date columns (Order Date, Ship Date) for correct datetime format.

Outlier Detection:

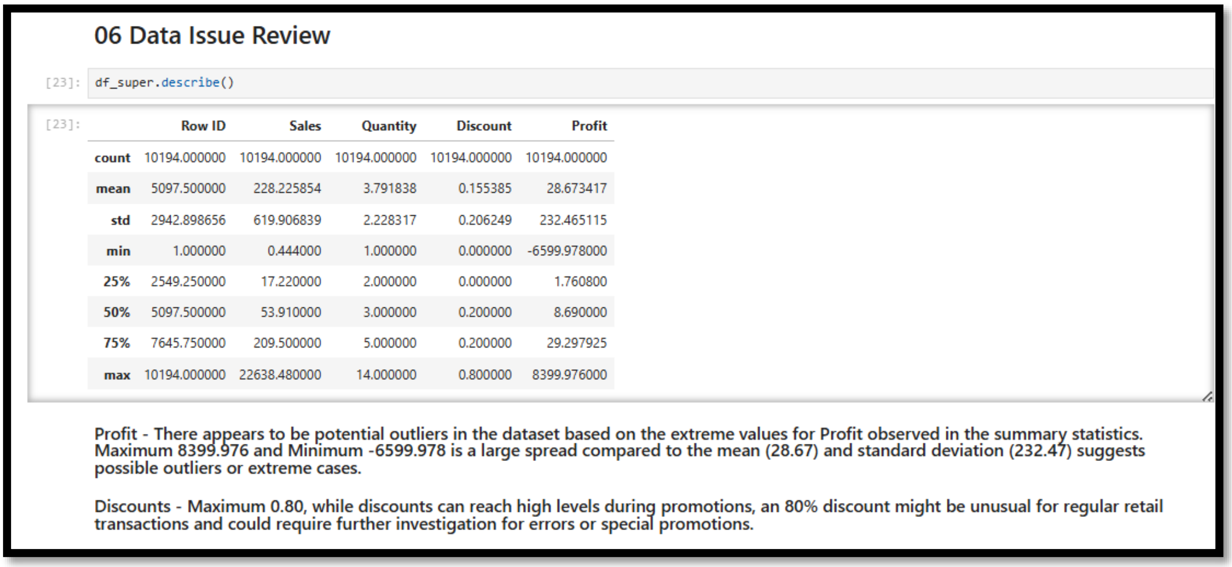
- Check for outliers in the Profit column as .describe() illustrated potential outliers.
- Identify outliers using statistical method - the Interquartile Range (IQR). 1913 rows.

Merge data:

- Join the Sample - Superstore dataset with Retail_Superstore_Returns on Order ID and create a new column indicating whether an order was returned.
- Verify merge success using .value_counts() & .duplicated.sum()
- Three new columns added to dataset:
    - 'Profit_Outlier' ('Normal' or 'Outlier')
    - 'Returned' ('Yes' or 'No')
    - 'Duplicate_Flag' ('Unique' or 'Duplicate') – simply highlighting four rows in the dataset.

Understand your data. Develop a basic understanding of your data set by reviewing the variables and performing basic descriptive statistical analysis.

As per the Jupyter file, the variables reviewed, and BDA illustrated below:



06 Data Issue Review

```
[23]: df_super.describe()
```

[23]:

	Row ID	Sales	Quantity	Discount	Profit
count	10194.000000	10194.000000	10194.000000	10194.000000	10194.000000
mean	5097.500000	228.225854	3.791838	0.155385	28.673417
std	2942.898656	619.906839	2.228317	0.206249	232.465115
min	1.000000	0.444000	1.000000	0.000000	-6599.978000
25%	2549.250000	17.220000	2.000000	0.000000	1.760800
50%	5097.500000	53.910000	3.000000	0.200000	8.690000
75%	7645.750000	209.500000	5.000000	0.200000	29.297925
max	10194.000000	22638.480000	14.000000	0.800000	8399.976000

Profit - There appears to be potential outliers in the dataset based on the extreme values for Profit observed in the summary statistics. Maximum 8399.976 and Minimum -6599.978 is a large spread compared to the mean (28.67) and standard deviation (232.47) suggests possible outliers or extreme cases.

Discounts - Maximum 0.80, while discounts can reach high levels during promotions, an 80% discount might be unusual for regular retail transactions and could require further investigation for errors or special promotions.

Understanding these insights early in the analysis ensures data accuracy and informs appropriate preprocessing steps for more reliable results.

Understand Consider limitations and ethics. Outline any limitations and ethical considerations presented by the content of your data, its source, and/or how it was collected.

Considerations for using this dataset:

Data Source and Collection Methods:

- **Limitation:** The Kaggle dataset lacks clear information on *how* the data was collected (e.g., direct POS data, surveys, third-party aggregation).
- **Ethical Concern:** If the source does not clearly indicate if the data were collected with proper consent, using it for analysis could violate data privacy norms.

Completeness and Representativeness:

- **Limitation:** The dataset does exclude certain product categories, regions, or customer segments, providing incomplete analysis.
- **Ethical Concern:** Drawing general conclusions from incomplete data can be misleading and unfairly represent certain groups or markets.

Potential Biases in the Dataset:

- **Limitation:** If the data was collected from a single store, chain, or geographic area, it might introduce a sampling bias.
- **Ethical Concern:** Bias can lead to inaccurate generalizations, such as over-representing a particular consumer group while under-representing others.

Data Privacy and Confidentiality:

- **Limitation:** If the dataset includes personal identifiers like customer names or contact details, there could be a risk of privacy violations.
- **Ethical Concern:** Managing such data without anonymization could breach data protection laws (e.g., GDPR).

Data Profile:

Define questions to explore. In a third section of your project document, define a list of questions to explore with your analysis.

Key questions to explore with my analysis based on the project brief:

Descriptive Analytics:

- What are the sales trends over time? (e.g., monthly, quarterly, yearly)
- Which product categories and subcategories contribute the most to total sales and profit?
- What is the average discount offered across different regions and how does it impact profit margins?
- Which regions, cities, and states/provinces have the highest and lowest sales performance?
- How does the distribution of sales and profit vary across customer segments?

Geospatial Analysis:

- Which regions or cities generate the highest revenue?
- Are there any geographic patterns in sales performance and returns?

Customer Behaviour & Segmentation:

- What are the most common buying patterns among customers?
- Is there a relationship between customer segments and sales/profitability?
- Are there specific products frequently returned by certain customer segments?

Advanced Analytical Techniques: Regression Analysis:

- What factors have the strongest influence on sales and profit? (e.g., discounts, region, quantity sold)
- Is there a relationship between discount levels and profit margins?

Clustering Analysis:

- Can customers be grouped into clusters based on their purchase behaviour?
- Can products be grouped based on sales performance and profitability?

Time Series Analysis:

- What are the seasonal trends in sales and profits?
- Are there recurring patterns in monthly or yearly sales?

Ethical Considerations & Data Limitations:

- Are there any biases in the dataset due to the data source, region, or time frame?
- Is the dataset representative of the entire market or specific to a subset?
- Is the data current and relevant for decision-making?

Exploring Hypothesis:

Impact of Discounts on Profitability

- **Hypothesis:** Higher discount rates negatively impact profit margins.
  (H_0 = There is no statistically significant relationship between discount rates and profit margin).
- **Rationale:** Offering higher discounts may reduce profit margins despite potentially increasing sales volume. This hypothesis can be assessed using regression analysis between the Discount and Profit columns.
- **Potential Tests:** Correlation heatmap, scatter plot with a trendline, linear regression analysis, using a significance level 0.05.

Regional Influence on Sales Performance

- **Hypothesis:** Sales and profitability vary significantly by region, with certain regions consistently outperforming others.
  (H_0 = There is no statistically significant difference in sales and profitability across regions).
- **Rationale:** Geographic factors, such as local demand, competition, and economic conditions, can affect sales performance.
- **Potential Tests:** Geospatial heatmaps, box plots comparing sales across regions, descriptive statistics for each region.

Customer Segmentation and Purchase Behaviour

- **Hypothesis:** Customers in the "Corporate" segment generate higher average order values compared to the "Consumer" segment.
  (H_0 = There is no statistically significant difference in the average order values between the 'Corporate' and 'Consumer' segments).
- **Rationale:** Corporate clients often place bulk orders, leading to higher sales per transaction compared to individual consumers.
- **Potential Tests:** Box plots comparing sales by customer segment, group-by analysis for average sales, cluster analysis for segmentation.