Sourcing Open Data

Sunita Lamichhane

3/4/024

Exercise 6.1

**Data Sourcing**

The Online Shopping Dataset is an external data source provided by Kaggle. <https://www.kaggle.com/datasets/jacksondivakarr/online-shopping-dataset>

It is a reliable source of data because Kaggle is the world's largest online data Sourcing community with powerful tools and resources to help people achieve their goals. Kaggle provides various and qualified data sets that are open for everyone.

**Data Collection**

Based on the Kaggle website the dataset has been scrapped from the internet. The author has not provided details on how the data was collected. The dataset is about product sales in the online platform.

**Data Ethics and Limitations**

The data collection method is not clear. So, we don’t know what stores or what websites they were collected from. This leads to the potential sampling and exclusion bias. Also, it is not clear if the ethical standards were followed collecting the data. The data doesn’t contain the PII information indicating no ethical issue in that case. But there might be other unknown issues like accuracy of the data. The data is from 2019 only and looks like the author hasn’t updated the data for recent years.

**Data Contents**

The data contains the online shopping data of United States different states years in 2019’s different months. The data sets have Customer ID, Gender, Location, Tenure\_Months, Transaction \_ID, Transaction\_Date, Product\_SKU, Product\_Description, Product\_Category, Quantity, Avg\_Price, Delivery\_Charges, Coupon\_Status, GST, Date, Offline\_Spend, Online\_Spend, Month, Coupon\_Code, Discount\_pct.

**Data Relevance**

The Online Shopping Dataset is relevant to our analysis since we can use this data to analyze different patterns of online shopping. From this dataset we can find which state’s people do more online shopping, people buy which product most? which gender do most of the online shopping etc.

**Data profile**

Data profile is attached in the Excel workbook.

**Why did I choose this data set?**

I chose online shopping dataset because it fulfills the project brief requirements, and which provides the online shopping pattern of different states and using those I can do different analysis of online shopping.

**Questions for Data Analysis**

1. Which Gender do most of the online shopping?
2. Which state spends most on online shopping?
3. What is the most popular product?
4. Which product has the highest and lowest average price?
5. What percentage of the people use coupon codes?
6. What is the relationship between coupon usage and total purchases?
7. Which month generates the highest revenue?