**CECS 456 — Data Description Report**

**Study Design and Data Source**

* Dataset Name: **Amazon Sales 2025**
* Source: **Zahid Feroze from Kaggle**
* Number of Observations (Rows):**250**
* Number of Variables (Columns): **10**
* **Description of Dataset:** The dataset contains 250 transactions records of amazon sales. Each has information represents an individual sale completed on the Amazon platform, capturing detailed information about the product purchased, its category, price, quantity sold, total sales value, customer details

(Write a short paragraph explaining what the dataset is about and what each observation represents.)

**Outcome Variable**

* Outcome Variable Name: **product association rules**
* Descriptio**n:** The outcome would be a relationship between products. The association would not be making a outcome category but finding a pattern between categories and products. The algorithm would find the relation of the product.

|  |  |  |
| --- | --- | --- |
| **Outcome Category** | **Count** | **Percentage** |
| **Frequent product associations** | **N/a** | **N/A** |
|  |  |  |
|  |  |  |

**Descriptive Statistics for All Variables**

**Categorical Variables**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Category** | **Count** | **Percentage** |
| **Product category** | **Footwear**  **Electronics**  **Clothing**  **Books**  **Home Appliances** | **27**  **118**  **40**  **25**  **40** | **10.8**  **47.2**  **16**  **10**  **16** |
| **Payment methods** | **Credit card**  **Debit card**  **PayPal**  **Gift card**  **Amazon pay** | **54**  **53**  **60**  **42**  **41** | **21.6**  **21.2**  **24**  **1638**  **16.4** |
|  |  |  |  |
|  |  |  |  |

**Continuous Variables**

| **Variable Name** | **Min** | **Max** | **Mean** | **Median** | **Standard Deviation** |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

(Add a table like this for each continuous variable)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable Name | Min | Max | Mean | Median | Standard Deviation |
| Price | 15 | 1200 | 343.58 | 150 | 380.63 |
| Quantity | 1 | 5 | 2.856 | 3 | 1.43 |

**Feature Decision and Engineering Section**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Action (Keep / Bin / Drop / Transform)** | **Explanation** |
| **Order ID** | **Kept** | **This is would be used to keep track of grouping products purchased together in the same transaction** |
| **Date** | drop | we are not looking for season trends but to see what items are purchased together |
| Product | keep | This is important and the key into finding the pattern of items because this variable is the item itself and would need to be kept. |
| category | keep | This would be used to group similar products. |
| Price | bin | Useful for analyzing rules by product type and could also group similar products. |
| Quantity | bin | Continuous values should be binned into price ranges Low, Medium, High |
| Total Sales | bin | bin into "Low Sale," "Medium Sale," "High Sale" |
| Customer name | Drop | We won’t need to use the customer’s name to find a pattern that relates to the teams |
| Customer Location | keep | Used to discover location-based purchasing patterns and the location could be key to find relation patterns between items |
| Payment Method | Keep | find associations between certain products and preferred payment methods. |
| status | Drop | We won’t need to consider the status because the grouping of items have already been made and won’t need to refer if it was complete, returned, or canceled |

* For **each variable**, clearly explain:
  + Why did you decide to keep it, bin it, drop it, or transform it?
  + If binning, describe the bin ranges and reasoning.
  + If transforming, describe the method used.
  + If dropping, justify why the variable is not useful.

**Optional Visualizations**

(Insert simple plots here if you create any — bar charts, histograms, pie charts, etc.)  
A graph showing different colored squares

AI-generated content may be incorrect.

A graph of sales distribution

AI-generated content may be incorrect.

**End of Report**