# Data Quality Assessment, Anomaly Detection, and Exploratory Data Analysis in E-Commerce Customer Data: Test Project

# Data Dictionary

|  |  |
| --- | --- |
| Column Name | Description |
| customer\_id | A unique identifier for each customer. |
| age | The age of the customer, typically between 18 and 80 years. |
| income | Annual income of the customer, rounded to the nearest 100. Ranges between 20,000 and 120,000. |
| purchase\_amount | The most recent purchase amount (currency) made by the customer , typically between 20 and 2,000. |
| gender | The gender of the customer, either 'Male' or 'Female' |
| region | The geographic region where the customer is located. One of 'North', 'South', 'East', 'West'. |
| total\_purchases | The total number of purchases made by the customer, typically between 1 and 50. |
| preferred\_product\_category | The customer's favorite product category. One of 'Electronics', 'Clothing', 'Home & Garden', 'Toys', 'Health'. |
| customer\_loyalty\_score | A loyalty score assigned to each customer based on their behavior, ranging from 0 to 100. |
| signup\_date | The date when the customer signed up for the service, typically between 2021-01-01 and 2022-12-31. |
| last\_purchase\_date | The date when the customer made their last purchase, typically after signup\_date. |

# Project Deliverables

* Data Quality Report:
* Data quality checks (missing values, duplicates, outliers, anomalies etc)
* Methods used to clean the data.
* Exploratory Data Analysis (EDA) Report:
* Descriptive Statistics: Provide basic descriptive statistics for numerical as well as categorical features and derive insights
* Univariate Analysis: Visualizations (histograms, box plots, count plots etc) for each variable.
* Bivariate Analysis: Analyze relationships between two variables, such as age vs gender, income vs. purchase\_amount, or customer\_loyalty\_score vs. total\_purchases etc and derive insights.
* Multivariate Analysis: Analyze relationship between two or more variables and insights from the same
* Correlation Analysis: Analyze the correlation between numerical variables using correlation matrices and heatmaps.
* Time-Based Analysis: Visualize trends over time, like signups per month or last purchase dates.
* Python Jupyter Notebook: A well-documented Jupyter notebook for data cleaning, anomaly detection, and EDA.
* A brief summary on key findings along with insights.
* Presentation (Optional): A brief presentation summarizing key findings, visualizations, and insights.