Tyler Boda

The code snippets in module 4’s dataset show how code can use weak customer information and turn it into useful data for decision-making. This process is the DataFrame, a table format that stores rows of customer info such as age, gender, income, and purchase behavior. Functions like df.head(), df.shape, and df.info() gives a look at the data’s content, size, and quality, by highlighting missing values in satisfaction scores or duplicate records that need to be fixed.

df.head() – shows top rows of data stored … df.tail() shows bottom rows of data stored

df.shape() – returns rows/columns of data… df.info() shows you information like how many values are present (non-null counts), data types, and memory usage. (User can use this information to determine anomalies and missing entries)

Statistical summaries with df.describe() is used to calculate averages, ranges, and variability of numbers like income and spending. This helps spot anything unusual, like extremely high income results. Histograms are used to display numerical features which are continuous and used as ‘bins’. Bar charts are used to display categorical features which compare totals/counts between categories.

The correlation matrix shows relationships between variables. For example, customers who spend more also tend to have higher average order values. This can help decide which features to focus on or combine when building predictive models. Using text or categories, classes like OrdinalEncoder turn ordinal labels like education levels ("High School", "Bachelor", "Master", "PhD") into numbers for machine learning algorithms to understand. This numerical translation is key because most models only work with numbers.

Missing data is common, and simple fixes include dropping incomplete rows or filling gaps with an average (‘imputed’). This allows the dataset to continue to be usable for training models. Sometimes groups say, male and female customers are uneven in number. Techniques like resampling help balance these groups to avoid biased predictions. Outlier detection looks for unusual data points that could mislead models. Statistical rules like the interquartile range (IQR) flag customers with incomes far above or below normal, letting users to decide whether to exclude or adjust them.

These steps prepare data to build models that can predict customer behavior, segment markets, or improve services. They clean, organize, and convert data into a numeric format. This workflow can be applied across industries where complex data must be turned into useable information.