A: The Question

"Does a customer's income impact the likelihood they will purchase an Online Security package?"

Exploring the relationship between income and optional extras could alert the business to affordability concerns with the offered Online Security product. Should such a relationship emerge, the business could look into more competitive pricing models, discounts, and income-based vouchers, driving an increase in usage of the service.

Within the existing database, I will be looking at the customer table. We will be using the numeric field customer.income. From the Services.csv file, we will be using the OnlineBackup column, represented as services.online_backup using a text value.

Both tables will be joined on the customer_id field, which is a text field shared both by the existing table and the table generated from Services.csv. In addition to these existing columns, the report will generate three new columns. The first will be income_bracket, a text field that will be generated by separating records by customer.income according to three pre-defined thresholds. The second column will be total_customers, a numeric field. This column will be a count of each record falling into the income_bracket groups from before. The third column is also numeric, percentage_with_online_security. This will be the count of records where services.online_backup = true divided by total_customers and multiplied by 100 for a result expressed as a percentage.

B: Entity Relationship Diagram

Above is the ERD for the two tables in question. They are joined by a shared customer_id. This column is unique in each table and serves as the primary key. Each record from `customer` corresponds to exactly one record from `services` establishing a one-to-one relationship between the tables.

Adding the services table to the database:

CREATE TABLE services (

   customer_id TEXT PRIMARY KEY,

   internet_service TEXT,

   phone TEXT,

   multiple TEXT,

online_security TEXT,

online_backup TEXT,

device_protection TEXT,

tech_support TEXT,

FOREIGN KEY (customer_id) REFERENCES customer(customer_id)

);

Populating the services table with data from the csv:

COPY services(customer_id, internet_service, phone, multiple, online_security, online_backup, device_protection, tech_support) FROM 'C:\LabFiles\Services.csv' DELIMITER ',' CSV HEADER;

C: SQL

The intention is to use aggregate functions to determine the percentage of customers have Online Security packages within three income brackets: Below 20,000, between 20,000 and 60,000, and over 60,000.

This is the statement:

SELECT income_bracket, COUNT(*) AS total_customers, COUNT(CASE WHEN s.online_security = 'Yes' THEN 1 END) * 100.0 / COUNT(*) AS percentage_with_online_security FROM (SELECT c.customer_id, c.income, CASE WHEN c.income < 20000 THEN 'Below 20,000' WHEN c.income BETWEEN 20000 AND 60000 THEN 'Between 20,000 and 60,000' ELSE 'Above 60,000' END AS income_bracket FROM customer c) AS income_groups JOIN services s ON income_groups.customer_id = s.customer_id GROUP BY income_bracket;

Output:

The results of the query are included in the D205_Results.csv file.

D: Refresh Timing

In order to keep the report timely, the data should be refreshed annually. While income tends to increase for the average person over time, general shifts in income are not that dramatic within shorter intervals.

After times of larger change, it would help for the business to refresh the data and take another look at the numbers to see if changes in income over time makes a dramatic difference, allowing them a further insight into subscription levels as they apply to customer income.

E: Panopto Video

Below is a link to the Panopto video for the task.

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=a818fd2c-10d0-4028-86f3-b1a1013e9776