D205 Performance Assessment

D205 DATA ACQUISITION

Medina, Luis
WESTERN GOVERNORS UNIVERSITY
MARCH 30, 2024

A: Research Question

My research question will be, “Do customer who have higher income are more likely to have better internet package?” Asking this question can be useful for the business as this will give them insights into other things they can include. For example, companies will have a better understanding of who to offer promotions for the internet or who can be a candidate for faster internet speed packages. If wanted, we can also use other columns from the database to further expand on the idea of who to offer internet packages too.

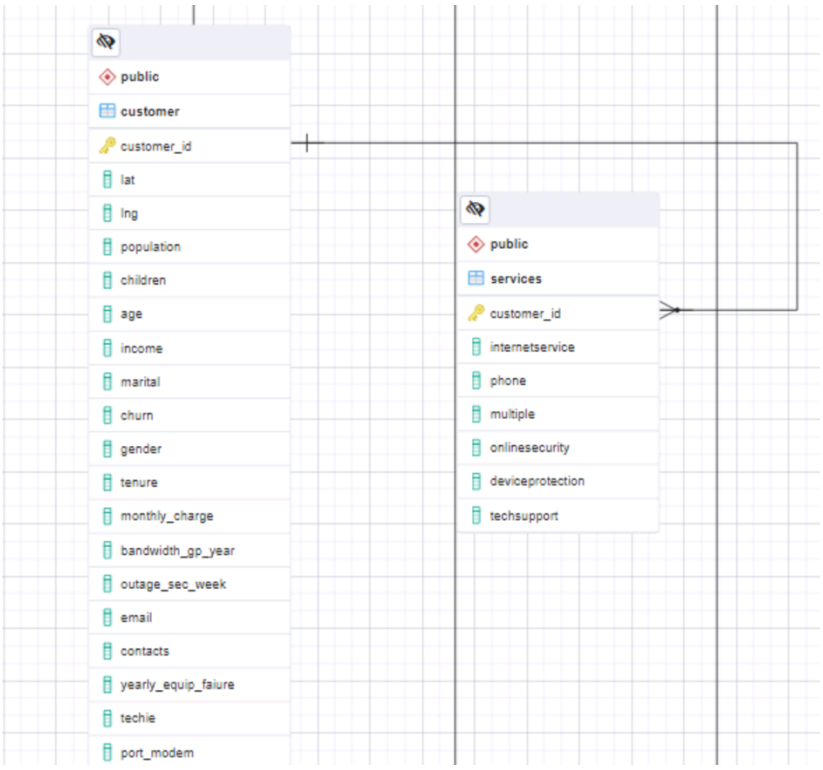
A1: Question Justification

To answer this question, we would need to investigate two columns from the database and two columns from the CSV file. These 4 columns once combined into one table will give us an understanding of how people from different income classes buy internet packages, meaning we will be able to make decisions from the data. If needed, we can always add more columns if the decision is not a clear decision, adding more factors can also help with the decision making.

A2: Identifying Data

For this question it will require the use of the churn database with the services.csv file to help answer the question. First, from the churn database we require the columns “customer_id” and “income.” Then, for the “services.csv” file we will require “customer_id” to join the two tables and “InternetService” column, which is one of the columns that will help answer our research questions. After joining the two tables we will be able to look at the different internet packages people have and see how income can play a role in all this.

B. Entity Relationship Diagram



After creating the ERD diagram this is a 1:M relationship as the customers table will have multiple records in the service table. The relational constraints of the new table are the Primary Key customer_id for the services table. Also, this is the Foreign Key that connects it to the customer's table.

B1: Relationship Discussion

The relationship between the newly created table and the table that already existed is a one-to-many relationship as one customer can have multiple records in the services table. But it's not the same the other way around, meaning that each record in the services table can only correspond to one customer. One issue that can arise is data integrity, as this does not ensure that the column "customer_id" in the services table exists in the customer's table.

B2: Statements for the ERD

Here is the SQL statement used to create the new table in the database.

```
CREATE TABLE public.services (
        customer_id varchar(64) NOT NULL,
        InternetService varchar(64),
        Phone  varchar(3),
        Multiple varchar(3),
        OnlineSecurity varchar(3),
        OnlineBackup varchar(3),
        DeviceProtection varchar(3),
        techSupport varchar(3),
        PRIMARY KEY (customer_id),
        CONSTRAINT customer_id_fkey FOREIGN KEY (customer_id)
        REFERENCES public.customer ( customer_id)
        );
```

The script above is the one I generated to use to create the new table for the data from the CSV file. There were no errors with this code, and it successfully generated a new table, now I will add the data from the CSV file.

B3: Loading CSV Data

Here is the script used to load the CSV data into the table created earlier.

--command " "\\copy public.services(customer_id, internetservice, phone, multiple, onlinesercurity, onlinebackup, deviceprotection, techsupport) FROM 'C:/LabFiles/Services.csv' DELIMITER ',' CSV HEADER QUOTE '\"ESCAPE"";""

C: SQL Query

```
WITH customer_info AS (
	SELECT c.customer_id, c.income, s.internetservice
	FROM customer AS c
	INNER JOIN services AS a
	ON c.customer_id = s.customer_id
)

SELECT internetservice,
	COUNT(CASE
			WHEN income < 30000 THEN 1
			ELSE NULL END) AS lower_class,
	COUNT(CASE
			WHEN income >= 30000 AND income < 58000 THEN 1
			ELSE NULL END) AS lower_middle,
	COUNT(CASE
			WHEN income >= 58000 AND income < 94000 THEN 1
			ELSE NULL END) AS middle,
	COUNT(CASE
			WHEN income >= 94000 AND income < 153000 THEN 1
			ELSE NULL END) AS upper_middle,
	COUNT(CASE
			WHEN income >= 153000 THEN 1
			ELSE NULL END) AS upper_class
FROM customer_info
GROUP BY internetservice;
```

Here I created a CTE first so it's easier to just look at the data we need to answer the research question. Then, I created a CASE statement so that way I can break the incomes into a separate group so we can see the different tax brackets to easily look at what internet package they have.

C1: CSV Files

Data Output | Explain | Messages | Notifications

	internetservice character varying (64)	lower_class bigint	lower_middle bigint	middle bigint	upper_middle bigint	upper_class bigint
1	DSL	1528	1163	584	175	13
2	Fiber Optic	2018	1514	660	198	18
3	None	950	727	347	98	7

Here is a picture of the results gathered from the query above, these will also be saved on a csv file that will be submitted.

D: Add-On File Time Period

The data gathered from the CSV file should be updated and combined again at least once a year for better results. Monthly updates can also work too if we start to look at monthly income instead of yearly income.

D1: Explanation of Time Period

The reason for the time period mentioned above is because income is something that usually will change monthly and yearly. For example, people's yearly salary could change whether they changed jobs or got a raise at their current job. So, updating this monthly or yearly will be accurate with the change in their income. Also, for updating internet packages monthly and yearly, make sense because when getting an internet package, you will have to get a contract for a yearly special. Meaning that once the year is up people might switch what kind of internet service they have. Then, there are people who after their contract is up, they just pay month to month, meaning that it can change any month they want to upgrade or downgrade their services. This is why the CSV file should be updated monthly and yearly along with the data from the database.

E: Panopto Video

The Panopto video will be pasted in the Links option.

F: Web Sources

No web sources were used for the application.