**D205: Data Acquisition**

**Performance Assessment – Task 1**

Leng Yang

WGU MSDA

Student ID: 012298452

July 11, 2024

**A. Question**

This paper’s focus question is: What is the top internet service subscription amongst male customers?

**A1. Question Justification**

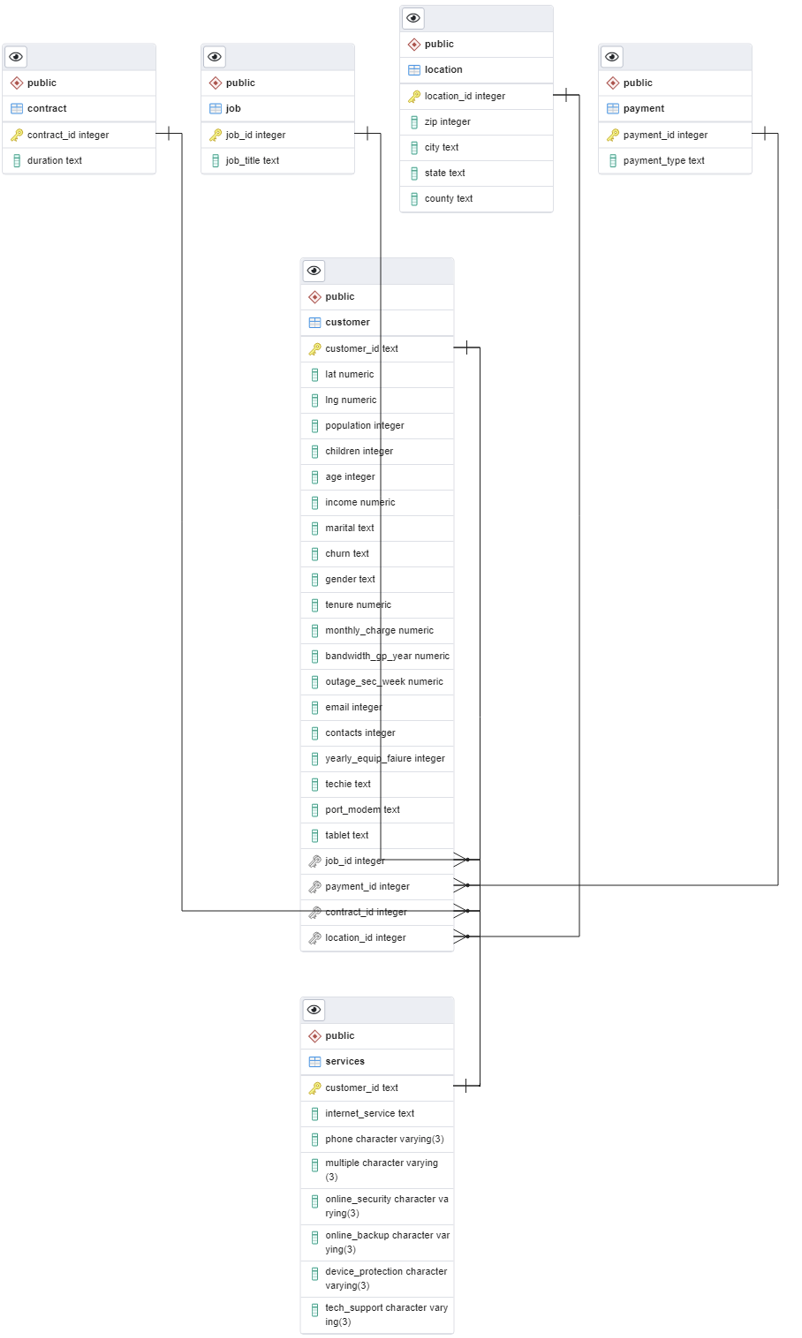
The question will be answered using pgAdmin4 and pgSQL, the relational database management system (RDBMS) that hosts the churn dataset containing one of the tables needed, specifically, the customer table. Additionally, the services.csv file, an external file, will be loaded into pgAdmin4. This external file is required because it contains the type of internet service each customer has.

**A2. Identifying Data**

The customer table will be used from the churn dataset. The two columns of interest from this table are “customer\_id” and “gender,” which are of text datatype. The customer\_id field will be used to match records in the services table during the join, and the gender field will be used to filter results only on customers who identify as male.

From the services table, the two columns of interest are “customer\_id” and “internet\_service.” Both columns are of text datatype. Again, customer\_id is used to match records during the join. For the internet\_service column, the records will be counted and then grouped by each internet service type.

**B. Entity Relationship Diagram**



The entity relationship diagram (ERD) was generated using pgAdmin’s generate ERD tool. Due to the limitations of the auto-generate feature, only 1:M and M:N relationships could be featured. To circumvent this limitation, the diagram was edited to correctly show the 1:1 relationship between the customer and services table.

**B1. Relationship Discussion**

The relationship between the customer table and the services table is a 1:1 relationship. This relationship is held because each record in one table is only related to one record in the other table and vice versa (Dan-Chuku, 2018). In this case, customer IDs are unique for every customer, so each customer\_id record in the services table uniquely corresponds to each customer\_id record in the customer table.

One issue that could arise with the relationship between the customer table and the services table is due to the relational constraint set during the services table creation. Since the services table also sets the customer\_id field as a foreign key pointing towards the customer table's customer\_id field, adding records with a new customer ID in the services table is only possible if one already exists in the customer table.

**B2. Statement for the ERD**

CREATE TABLE public.services (

customer\_id text NOT NULL,

internet\_service text,

phone VARCHAR(3),

multiple VARCHAR(3),

online\_security VARCHAR(3),

online\_backup VARCHAR(3),

device\_protection VARCHAR(3),

tech\_support VARCHAR(3),

PRIMARY KEY (customer\_id),

FOREIGN KEY (customer\_id)

REFERENCES public.customer (customer\_id) MATCH SIMPLE

ON UPDATE NO ACTION

ON DELETE NO ACTION

NOT VALID

);

ALTER TABLE public.services

OWNER to postgres;

The table was created using pgAdmin’s “Create Table” function. After designing the features of the table, the generated SQL code was copied (Table Dialog).

**B3. Loading CSV Data**

COPY public.services

FROM 'C:/LabFiles/Services.csv'

DELIMITER ','

CSV HEADER;

Data was loaded from the services.csv file into pgAdmin utilizing the framework of the code from www.postgressqltutorial.com (Import CSV file into PosgresSQL Table).

**C. SQL Query**

SELECT internet\_service,

COUNT(\*) AS total\_subscribed

FROM services AS s

INNER JOIN customer AS c

USING (customer\_id)

WHERE gender = 'Male'

GROUP BY internet\_service

ORDER BY total\_subscribed DESC;

The SQL query joins the customer and services table on customer\_id and filters for those whose gender identifies as male. The records are totaled and grouped using this filter by internet service type. Lastly, the result set gets sorted by the number of subscribers from most to least.

**C1. CSV Files**

The result from the SQL query is attached to the submission labeled as result.csv.

**D. Add-on File Time Period**

For the add-on file to remain relevant to which internet service is the most popular among males, it should be updated monthly.

**D1. Explanation of the Time Period**

The timeframe for why the add-on file should be updated monthly is mainly due to how internet service subscriptions usually work. Most internet service providers bill for subscription services every month; thus, customers are locked into this cycle. Even if a customer were to end their internet service early, they would still have access to it until the end of their monthly cycle. And so, the underlying data regarding internet service would change monthly. Hence, it would make the most sense to update the file monthly as that is the best opportunity to track updates, whether that be the addition of new customers and services or a change in service type.

**E. Panopto Video**

Below is the link for the video submission, which was also added to the link section during submission.

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=a41f9168-c7e2-4adc-876d-b1ab001da264>

**F. Web Sources**

Dan-Chuku, C. (2018, October 10). *How to implement one to one, one to many and many* *to many relationships when designing a database.* Medium. <https://medium.com/@emekadc/how-to-implement-one-to-one-one-to-many-and-many-to-many-relationships-when-designing-a-database-9da2de684710>

*Import CSV file into PosgreSQL Table*. PostgreSQL Tutorial. (2024, February 1). <https://www.postgresqltutorial.com/postgresql-tutorial/import-csv-file-into-posgresql-table/>

*Table Dialog¶*. Table Dialog - pgAdmin 4 8.9 documentation. (n.d.). <https://www.pgadmin.org/docs/pgadmin4/8.9/table_dialog.html>