D205 – Data Acquisition

WGU M.S. Data Analytics

Lyssa Kline

January 31, 2024

A, Research Question

My research question for this project is, “Are there more customers aged 60 and above who seek tech support compared to those under the age of 60?” This research question is crucial to the business because it identifies a significant customer demographic that could be beneficial in multiple contexts. By determining the number of customers who require tech support services and the bulk of the customer demographic that needs support, the company can better evaluate the need for these services.

In the future, this question could be expanded to explore additional areas of opportunity, such as recurring issues and requests across different demographic groups, including regular outages, equipment failures, and the frequency of support requests.

A1, Identify Data

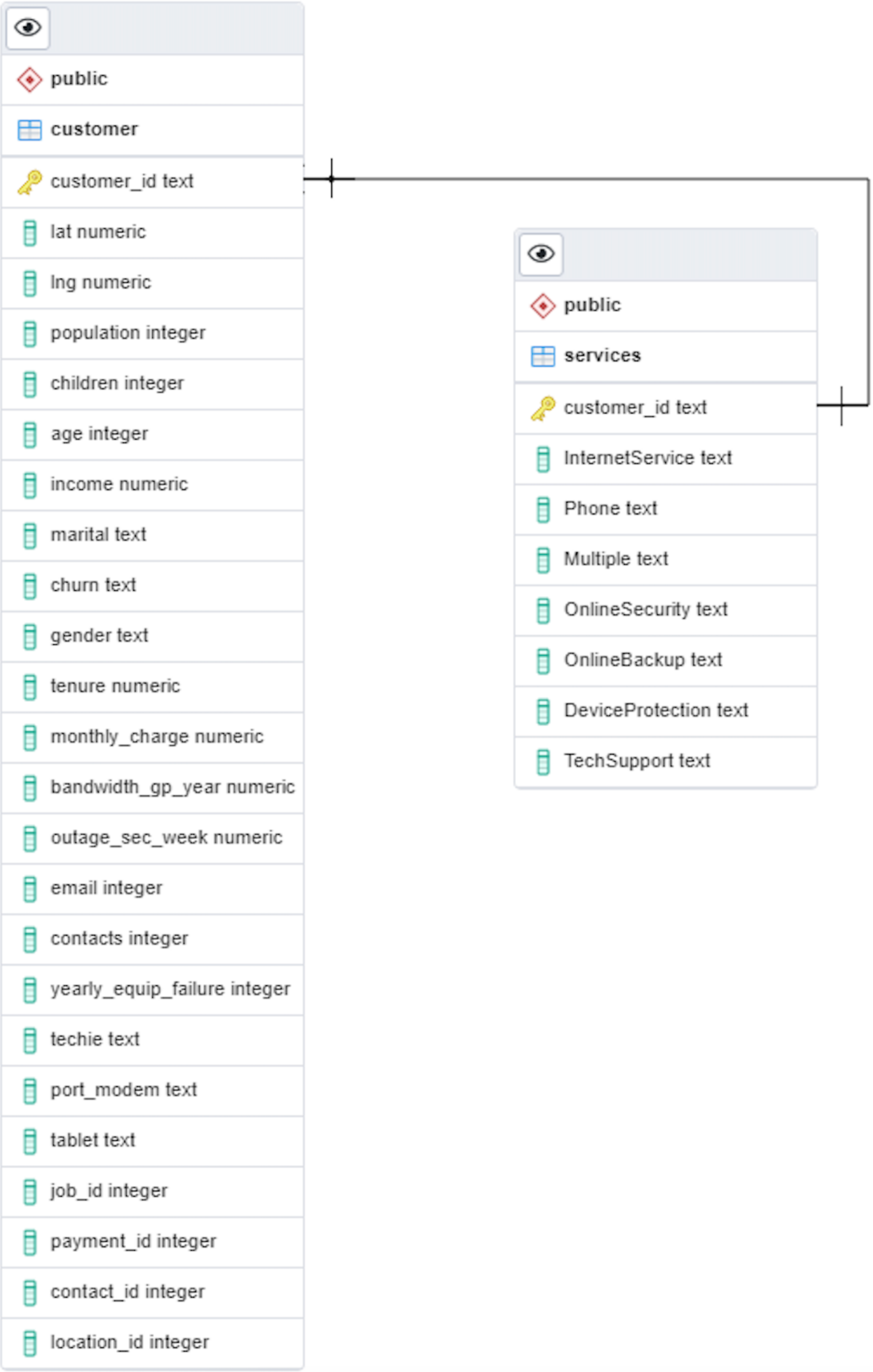
To adequately analyze the research question, three columns will need to be retrieved from the provided dataset along with an additional CSV file. Specifically, the 'customers' dataset within the ‘churn’ database will be used to extract 'customer_id' and 'age'. Additionally, 'customer_id' and 'techsupport' from the 'services.csv' file will be utilized to effectively answer the targeted research question. The ‘customer_id’ field between both datasets has a data type of text, ‘techsupport’ from the services table also has a data type of text. Text data types store any kind of text data (i.e. unique customer ID’s, or text indicators). The ‘age’ field from the customer table has a datatype of integer. Integers are whole numbers typically used to store counts and values (i.e. age numbers).

Within the datasets selected, the ‘customer_id’ field will be used as both the primary key and foreign key of the tables and will be used to join the ‘customers’ table with the ‘services’ table. By joining the tables on ‘customer_id’, it allows us to use the data from the CSV file. This will allow us to start to evaluate the potential correlation between the number of customers, the age of those customers, and the amount of tech support that is required.

B, Data Diagram

The ERD Diagram shows an overview of the 'customer' dataset, highlighting the key fields necessary to address the research question in scope. Specifically, ERD shows the 'customer_id' and 'age' fields. Both the Primary Key and Foreign Key of the customer dataset is the ‘customer_id’, this field is a unique identifier that can allow the linking of additional tables. The diagram also features a 'services' table, which has been loaded to PGAdmin for analysis. This table contains the 'customer_id' and 'techsupport' fields, with the former serving as both the Primary and Foreign Key. To answer our research question, we will employ a left join between the two tables, bringing in the 'techsupport' field.

As seen below there is a one-to-one relationship between customers and services. I believe that over time, a single customer in the customer table could potentially have many services however this table does not contain a date or timestamp field indicating that these services are recorded on one single row in the services table; therefore, a one-to-one relationship occurs.



public

customer

customer_id text

lat numeric

lng numeric

population integer

children integer

age integer

income numeric

marital text

churn text

gender text

tenure numeric

monthly_charge numeric

bandwidth_gp_year numeric

outage_sec_week numeric

email integer

contacts integer

yearly_equip_failure integer

techie text

port_modem text

tablet text

job_id integer

payment_id integer

contact_id integer

location_id integer

public

services

customer_id text

InternetService text

Phone text

Multiple text

OnlineSecurity text

OnlineBackup text

DeviceProtection text

TechSupport text

B1, Create Table for CSV Data

The following SQL code successfully creates the 'services' table within the churn database, utilizing the fields contained within the CSV file. The code designates 'customer_id' as both the Primary Key and Foreign Key of the table. This code effectively generates a table equipped with the necessary fields to facilitate the research question in scope.

```
---CREATION OF TABLE FOR ADD ON CSV FILE 'SERVICES'---
CREATE TABLE public.services
(
	customer_id text,
	InternetService text,
	Phone text,
	Multiple text,
	OnlineSecurity text,
	OnlineBackup text,
	DeviceProtection text,
	TechSupport text,
	PRIMARY KEY (customer_id),
	CONSTRAINT customer_id_fkey FOREIGN KEY (customer_id)
		REFERENCES public.customer (customer_id)
);

---ALTER CREATED TABLE -- OWNER TO POSTGRES---
ALTER TABLE public.services
	OWNER to postgres;
```

B2, Load CSV Data

The SQL code below loads data from the ‘services.csv’ file to the ‘services’ table created previously. The code worked successfully to load data from the CSV file into the created table. This query works by telling PGAdmin to copy the full CSV file (saved inside of the C drive) and saves it to the designated table location. The query uses ‘WITH CSV HEADER’ at the end to indicate that the file has headers, and to ignore them.

```
---LOAD ALL DATA FROM CSV FILE TO CREATED TABLE---
COPY public.services FROM 'C:\LabFiles\Services.csv' WITH CSV HEADER;
```

C, SQL Queries

The SQL statement written below shows the answer to the research question “Are there more customers aged 60 and above who seek tech support compared to those under the age of 60?”.

```
---PULL COUNT OF CUSTOMERS REQUIRING TECHSUPPORT---
WITH age_groups AS (
SELECT
	a.customer_id
	, b.techsupport
```

```
        , CASE WHEN a.age >= 60 THEN '60 and over'
               WHEN a.age < 60 THEN 'under 60'
         END AS age_group

        FROM public.customer AS a
        LEFT JOIN public.services AS b
                ON a.customer_id = b.customer_id
        )
SELECT
age_group
, SUM(CASE WHEN techsupport = 'Yes' THEN 1 ELSE 0 END) AS techsupport_count
, COUNT(customer_id) AS customer_count
FROM age_groups
GROUP BY age_group
ORDER BY age_group
```

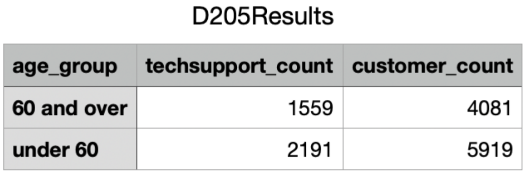
To obtain the necessary information for review, this query efficiently combines the 'customer' and 'services' tables through a left join. The left join allows us to pull in all customers from the ‘customer’ table with the matching records from the ‘services’ table. A Common Table Expression (CTE) is created, featuring key fields and a CASE statement to determine demographic groups. The SELECT statement then pulls together the age group and customer data retrieved from the CTE to accurately count the number of customers and quantity of tech support services.

The output of the query informs us about a few key demographic areas of the business that we can now start to conclude. Our analysis reveals that the total customer count is 10,000, with 41% of these customers aged 60 and over, and 59% under the age of 60. Within these groups 38% of the customers aged 60 and over required tech support, while 37% of customers under the age of 60 required tech support.

Based on the data analysis, two conclusions can be drawn. Firstly, when analyzing individual demographic groups, it was observed that the 'over 60' group required slightly more technical support than the 'under 60' group, with a difference of 1%. Secondly, when considering the total number of customers, the 'under 60' group had a higher count compared to the 'over 60' group. This informs us that 18% more customers were under the age of 60, and of those under 60, the overall count of tech support was higher.

While it appears that there are more customers under the age of 60 and more customers in this group required tech support; we can see that both groups had a very similar % demand for tech services. Thus, in conclusion, based on our observation of the customer groups and their tech service needs, the most important finding is that there is a consistent 37% demand for tech services across all customers and demographics. By having an idea of the demand for tech services, this knowledge can be leveraged to ensure that the company has enough resources and support staff available to meet customer needs.

C1, Query Results



D205Results

age_group	techsupport_count	customer_count
60 and over	1559	4081
under 60	2191	5919

A picture of the query's results is provided here, see the CSV file submitted alongside this project for the full report. The results CSV shows the results of the research question in scope. This data output helps to answer the research question and assists in the conclusion drawn under section C. As seen in the results chart, the query provided an output that shows the two customer age groups in question, '60 and over' and 'under 60'; a count of tech support requested; as well as a count of the total number of customers in each group.

D, CSV File Update/Refresh Time Period

The primary use case for the question in scope would be for the company to ensure they have adequate resources on hand to ensure customers can get the tech support help that they need. Due to this, to my knowledge, the data contained in these sources should be refreshed daily. There are two tables within the scope of this research question, both of which should be updated daily.

The customer table should be updated daily to ensure all new and existing customer transactions are captured within the company data. The services table should also be refreshed daily to ensure all company services are being captured on a timely basis for new and existing customer transactions. This will allow the company to capture any additional customers as well as receive tech support.

D1, Time Period Explanation

Company resources needed for tech support services would require a frequent review. Due to this, a daily refresh period has been deemed adequate. A daily refresh of the datasets used is relevant for the business to be able to conclude the needs of tech support. Frequent and up-to-date insights ensure that a company can review on a timely basis all allocated resources. This will help ensure the company has adequate availability of employees who can provide tech support, in return will ensure that customers will receive the assistance they require.

E, Panopto Video Recording Executions

The video recording for this assignment includes a vocalized demonstration of all the code, the code being executed, and the results of the code being represented inside this report.

E1, Panopto Video Recording Explanation

The video recording for this project can be found inside the Panopto drop box titled "Master of Science, Data Analytics TGM2 | D205 (Student Creators) [assignments]" as a URL link.

Panopto video link: https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b4c6d2b6-ff4e-40ea-a1be-b108013ada20

F, Web Sources

No sources, or segments of third-party code, were used to acquire data or to support the report.

G, Acknowledge Sources

I acknowledge that no sources, or segments of third-party code, were used to acquire data or to support the report.