Western Governors University

D205 – Data Acquisition Performance Assessment

By Krista Moik

Table of Contents

A. Research Question:

My research question using the customer and services data is: Do older customers require more technical support than younger customers? This is an important business question to have information on as profitable customers are one of the business goals, so it is necessary to know which customers may result in higher expenses due to staffing requirements to handle tech support requests and inquiries. It would be helpful if the files provided also included the number of times tech support was requested per customer, but it is out of the scope of the provided materials. For this question, we will consider older customers as those 65 years of age and older. We will consider middle-aged customers as those who are 35 to 64 years of age. For everyone under the age of 35, we will consider young customers.

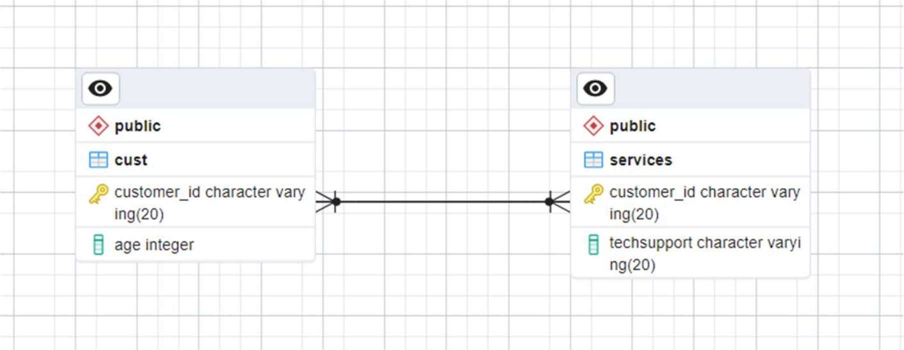
A1. Identifying Data

To answer this research question, I will use the following columns on the churn_clean CSV: customer_id and age. For the add-on files, I will use the services.csv. On this file, I will use the customer_id to join the tables, as well as the TechSupport column that states whether a customer used tech support or not. Using these columns will allow me to query the data to see the total number of these customers based on their age group, how many in each group of the total used tech support, and then calculate the percentage of each age group that used tech support. Once I have that calculation, I will be able to determine if older customers required tech support more than young customers.

B. Create an entity relationship diagram (ERD) for the add-on CSV file and any other tables and columns used to answer the question from part A by evaluating the data contained in the file and identifying the m:n relationships and relational constraints.

Below are two types of entity relationship diagrams. From the churn_clean.csv, I created a table called cust, and from the services.csv, I created a table called services. The primary key for both tables is customer_id. That primary key was used to join the 2 tables to run my queries. The ERDs shown below was created using the insert table function in Microsoft word and the second one was created using the create ERD tool in pgadmin4.

Table	Keys
Cust	Primary key: customer_id
Cust	Age
Services	Primary key: customer_id
Services	Tech support


public
cust
customer_id character varying(20)
age integer
public
services
customer_id character varying(20)
techsupport character varying(20)

B1. Code for the ERD: Write SQL code, in text format, that creates a table based on the ERD and specifies the columns and relevant keys.

I used the CREATE TABLE data definition language to create two tables: cust and services. The cust table included the customer_id and age columns from the churn_clean spreadsheet. The services table included the customer_id and techsupport columns. The primary key used to link these tables was the customer_id column that was present and provided unique identification in both tables. Age was the only column that was specified to be an integer. Below is the SQL query used to create the tables with the aforementioned columns:

```
CREATE TABLE cust (
customer_id varchar(20),
age int);

CREATE TABLE services (
customer_id varchar(20),
techsupport varchar(3));
```

B2. Loading CSV Data: Write SQL code, in text format, that loads the data from one of the add-on CSV files into the table created in part B1.

Note: Do not include SQL code as a screenshot.

After I created the above tables with the desired number of columns, I then used the import feature in pgadmin4 to import the specified columns into the table from the files. I imported the churn data into the cust table. The services.csv was imported into the services table. Below is the code that was generated in pgadmin4 after the successful importation of both files into their respective tables:

```
--command " "\\copy public.cust (customer_id, age, techie) FROM
'C:/Users/KMOIKW~1/Desktop/D205/cust.csv' DELIMITER ',' CSV HEADER QUOTE '\"' ESCAPE '''';""

--command " "\\copy public.services (customer_id, techsupport) FROM
'C:/Users/KMOIKW~1/Desktop/D205/Services.csv' DELIMITER ',' CSV HEADER QUOTE '\"' ESCAPE '''';""
```

C. SQL Query: Write a SQL statement or statements in text format for a query or queries that answer the question from part A.

Note: Do not include SQL statements as a screenshot.

Once I successfully imported the data files into their respective tables, I needed to join the tables to view all the data. To do this, I selected all the columns that I imported from the data files into the tables and used the WHERE clause to join the tables using the primary key: customer_id. Here is the SQL code that joined the tables and allowed me to view the data included in both:

```
SELECT *
FROM cust, services
WHERE cust.customer_id = services.customer_id
```

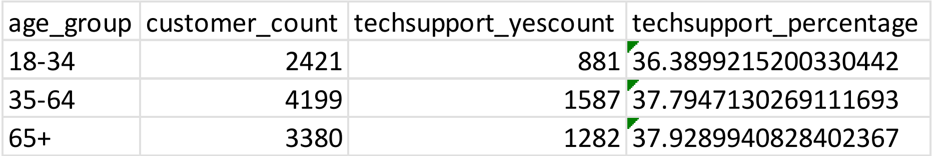
Once I confirmed the tables were joining properly, I then broke down what I needed to do to determine if older customers required more tech support. As previously stated, I decided to group young customers as those between the ages of 18 and 34, middle-aged customers as 35-64 years old, and older customers as 65 years or older. Not only did I need to group by age, but I also needed to count how many customers were in each age group as well as how many in each age group required tech support. Finally, I needed to add a column that would provide the percentage of each age group that required tech support by dividing the number in each age group that required tech support by the total number in that age group and multiplying that by 100 to obtain the percentage.
Here is the updated code I used to process that query:

```
SELECT
        CASE
                WHEN cust.age BETWEEN 18 AND 34 THEN '18-34'
                WHEN cust.age BETWEEN 35 AND 64 THEN '35-64'
                ELSE '65+'
                END AS age_group,
                COUNT(*) AS customer_count,
                COUNT(CASE WHEN services.techsupport = 'Yes' THEN 1 END) AS techsupport_yescount,
                (COUNT(CASE WHEN services.techsupport = 'Yes' THEN 1 END) * 100.0 / COUNT(*)) AS
techsupport_percentage
FROM cust, services
WHERE cust.customer_id = services.customer_id
GROUP BY age_group
```

By completing this query, I obtained a new view of the data that showed 4 new columns from my above query: age_group, customer_count, techsupport_yescount, and techsupport_percentage. The results showed that there are 2,421 customers in the 18-34 year age group and of that, 881 required tech support. This results in approximately 36.39% of young customers requiring tech support. The 35-64-year age group has a total of 4,199 customers. Of that, 1,587 of them, or 37.79%, required tech support. Finally, our old age group of 65+ had a total of 3,380 customers. Of those customers, 1,282, or 37.93%, required tech support.

While the data seems to confirm my hypothesis that older customers do require more tech support, the percentages are so close and with the lack of additional data provided to act as additional constraints, more information is probably needed before the business should determine that older customers are more costly to attract and retain due to their need for increased tech support.

C1. CSV Files: Provide a data file or files that capture the results from the query or queries.



age_group	customer_count	techsupport_yescount	techsupport_percentage
18-34	2421	881	36.3899215200330442
35-64	4199	1587	37.7947130269111693
65+	3380	1282	37.9289940828402367

I have also attached these output results in the PDF file named: Krista Moik D205 output.

D. Add-On File: Identify the specific time period for how often the add-on file should be acquired and refreshed in the database for the data to remain relevant to the business and the question from part A.

As the data constantly changes including ages as birthdays come and go, new tech support requests that come in, and normal customer churn that occurs, quarterly may be the best time to refresh the data. To do it daily would be unnecessary and not show too much change overall. Refreshing the data yearly, while not unreasonable, does not give the business opportunities to make adjustments in staffing, marketing, etc. throughout the year in response to what the data shows.

D1. Explanation of Time Period: Explain why the time period identified in part D is relevant to the business needs.

Quarterly is the most reasonable and best-suited time frame to refresh the data. As previously stated, daily is too often to get any real value out of the refreshed data, but to refresh yearly means missed opportunities to adjust business strategy in response to the data. While refreshing the data monthly may provide some insights, refreshing the data quarterly provides the most valuable insights into trends that are occurring and allows the business the opportunity to capitalize on positive trends and minimize potentially negative trends where possible.

E. Panopto Video of Code: Provide a Panopto video recording that includes the presenter and a vocalized demonstration showing all code used, the code being executed, and the results of all code used in the task.

E1: Panopto Video of Programs: Include a vocalized demonstration within the Panopto video recording provided in part E that describes the programs used to complete the task.

My Panopto recording was uploaded to the above Panopto drop box and the link to that video is below:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=0de2d6ce-3f1c-4f71-bee7-b0df013efd14

F. Web Sources: Acknowledge web sources used to acquire data or segments of third-party code to support the application. Be sure the web sources are reliable.

Note: Submit web sources for part F separate from the sources in part G, or state none were used.

I did not use any outside web sources.

G. Sources: Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

I did not use any outside sources.

H. Professional Communication: Demonstrate professional communication in the content and presentation of your submission.