Western Governors University
School of Technology, College of IT
Master of Science, Data Analytics

D205 Performance Assessment

Do customers who subscribe to the tech support service have a lower churn?

Tyson Biegler
Student ID: 012170282
D205 - Data Acquisition
June 3, 2024

A, Question:

I will seek to answer the question: “Do customers who subscribe to the tech support service have a lower churn?” Although it might be wise to investigate this same question for each available service, I will only be examining the tech support service at this time.

Understanding this question is important for many reasons. It helps shed light on the customer's subscription preferences. It will also allow me to determine if there appears to be, at first glance, a correlation between receiving tech support or other services and churn status. However, deciding whether or not there is a statistically significant correlation between tech support and churn status is outside this project's scope. So, I will determine if there is a difference between those who churn and those who don't.

If it is determined that customers who subscribe to tech support do have a lower churn rate, then following up with statistical tests is necessary. The company might consider adding tech support services to future marketing campaigns or sales efforts if there is a correlation.

A1 Question Justification:

I will answer the research question in a few steps. First, I will join the customer table with the services table and then count the total customers in the services table because I need to know how many customers are, or are not, subscribed to ‘techsuport.’ Because I am only interested in discovering the relationship between tech support subscriptions and churn, I will use a WHERE clause to only return the customers from the ‘customer’ table who have churned. Lastly, I will group and order the results by the service ‘techsupport.’ The results will display the number of customers who have churned and differentiate them by whether or not they were subscribed to tech support.

A2, Identifying Data:

To answer the research question, I must utilize four columns of data from two tables. Two columns are from the original ‘customer’ table, and the other two are from the addon CSV file.

Although I only used two columns in this services table to answer my research question, I loaded all the CSV data into the table I named ‘services.’

The ‘services’ table includes a column named ‘customer_id.’ This serves as the primary key in the ‘services’ table while simultaneously acting as a foreign key that references the ‘customer_id’ in the ‘customer’ table. To differentiate the two customer_id fields, I will refer to the customer_id in the services table as ‘services.customer_id’ and the customer_id in the customer table as ‘customer.customer_id’.

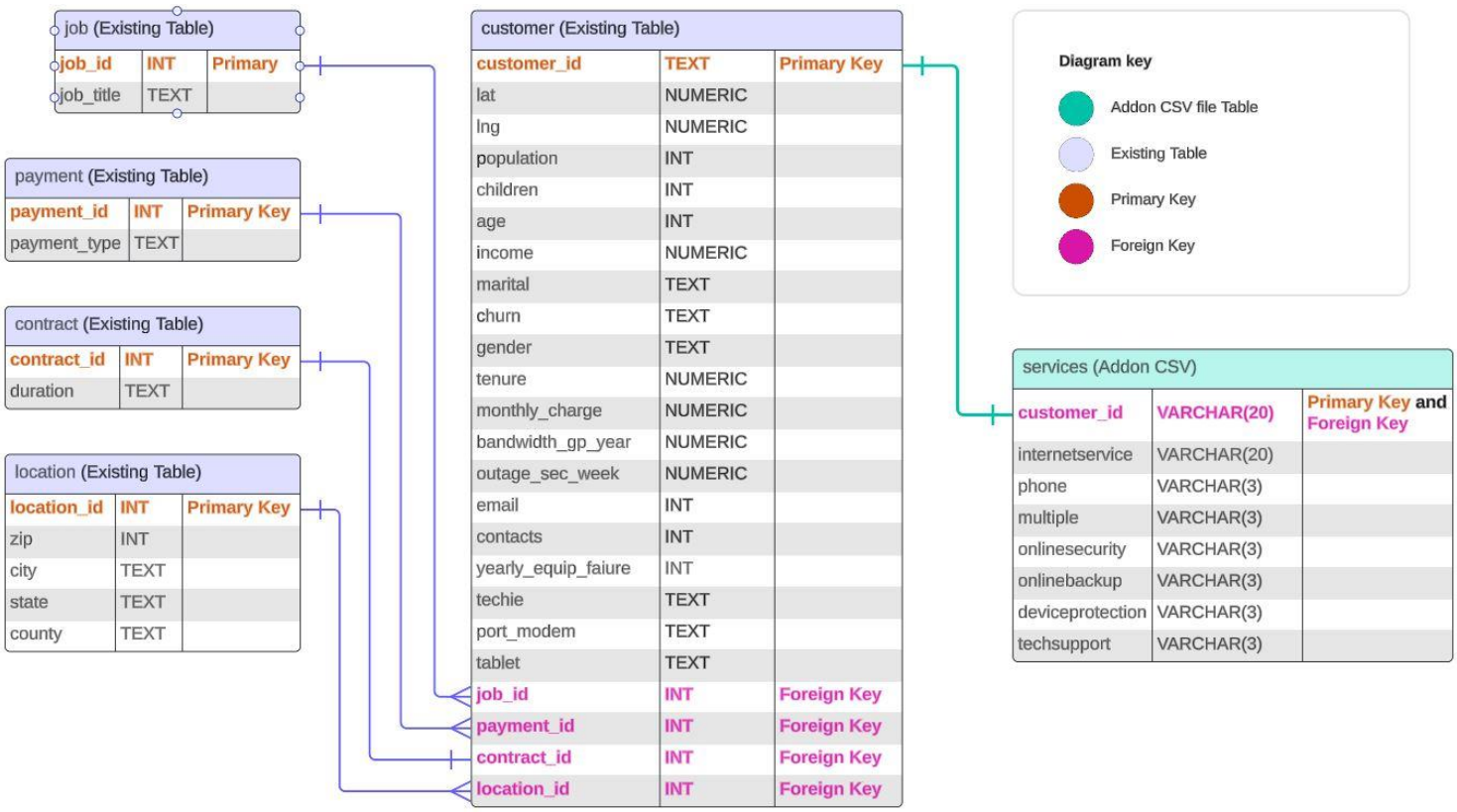
The following are the descriptions of the data types for each column in the services table. The data type for the services.customer_id column is a VARCHAR(20) to allow for plenty of characters without using up any unnecessary space (**Sewell, 2023, 13:00**). Even though the current maximum character length for any service is only 11 characters,

'internetservice,' being the longest character length is a VARCHAR(20). Having the character limit set to 20 will allow for additional internet services in the future that might require more characters without wasting storage space. Additionally, all other services offered, 'phone,' 'multiple,' 'onlineSecurity,' 'onlineBackup,' 'deviceProtection,' and 'techSupport,' are 'Yes' or 'No' values. These only require a VARCHAR(3) because the values can never exceed three characters. Using a VARCHAR(3) also helps clarify these columns' intent by informing future database users that these columns are meant for storing minimal-length data.

The customer.customer_id and the 'churn' columns will be used from the existing customer table. The customer.customer_id column is used to join with the services table, while the churn column identifies which customers have churned.

From the services table, I will need the customer_id. The services.customer_id column will be used to join with the customer table, as mentioned previously, and will also be used to get the customer count. I will also utilize the 'techsupport' column in the services table as I only investigate the relationship between tech support status and churn status.

B, Entity Relationship Diagram:



B1, Relationship Discussion:

The existing 'customer' table and the 'services' table, created from the addon CSV file, share a one-to-one relationship. This is established through the 'customer_id' column found in both tables. Each customer.customer_id in the *customer* table corresponds to one and only one record in the *service* table, and vice versa. There can never be multiple rows for any customer ID, thus indicating a one-to-one relationship. As a result, both tables contain demographic and account records or subscription status records for all 10,000 customers. This is true regardless of the number of services any particular customer subscribes to.

Problems might occur with the ERD if this company decides to, for example, add a unique or exclusive service that is only offered to specific customers; the relationship between the customer table and the services table would need to be changed.

Some of the current columns in these tables are lacking in specificity. For example, very complex queries would be required if the company wanted to gather specific data such as activation date or expiration of contracts since the tenure and contract duration fields do not specify this information.

The referential integrity of the relationships between the tables is consistent because the primary key of the customer table (customer.customer_id) is referenced by the foreign key of the services table (services.customer_id), which is also the primary key of the services table. This query uses a cascade clause in the services table to maintain referential integrity by preventing orphaned values if a customer is removed from the customer table (***PostgreSQL Documentation: 16: 5.4. Constraints*, n.d.**). In addition, transactions can prevent unintended data loss by allowing the cascade to be rolled back. (**PostgreSQL, n.d.**).

B2, Statement for the ERD:

I created the services table from the data in the services.csv. As shown below, the primary key 'customer_id' in the services table is also the foreign key to the 'customer_id' in the customer table.

```
CREATE TABLE public.services (
-- The customer_id column in the services table is the primary key of the table
	customer_id VARCHAR(20) NOT NULL PRIMARY KEY, internetService VARCHAR(20) NOT NULL,
	phone VARCHAR(3) NOT NULL,
	multiple VARCHAR(3) NOT NULL,
	onlineSecurity VARCHAR(3) NOT NULL,
	onlineBackup VARCHAR(3) NOT NULL,
	deviceProtection VARCHAR(3) NOT NULL,
	techSupport VARCHAR(3) NOT NULL,
-- The 'customer_id' column in the 'services' table is also the foreign key that references the
primary key 'customer_id' in the 'customer' table.
	CONSTRAINT services_customer_id FOREIGN KEY (customer_id)
		REFERENCES public.customer (customer_id) ON DELETE CASCADE
);
ALTER TABLE public.services
	OWNER to postgres;
```

B3 Loading CSV Data:

I used the COPY FROM method to load the data from services.csv.

```
COPY services (
	customer_id,
	internetService,
	phone,
	multiple,
	onlineSecurity,
	onlineBackup,
	deviceProtection,
	techSupport
)
FROM 'C:\LabFiles\Services.csv'
DELIMITER ','
CSV HEADER;
```

C, SQL Query:

```
SELECT
CASE
	WHEN s.techsupport = 'No' THEN 'Not subscribed'
	WHEN s.techsupport = 'Yes' THEN 'Subscribed'
END AS tech_support,
COUNT(*) AS churned_customers
FROM
	services s
JOIN
	customer c ON s.customer_id = c.customer_id
WHERE
	c.churn = 'Yes'
GROUP BY
	tech_support
ORDER BY
	tech_support;
```

C1, CSV Files:

Below is a screenshot of the table results. The Query_results.csv file will be submitted with this PA.

The results indicate that customers subscribing to the tech support service have a lower churn. However, correlation does not always imply causation. As mentioned earlier, these results do not demonstrate if there is a statistical correlation between the tech support service and the customers who have churned. These results suggest that a correlation could exist but should be investigated further.

	tech_support text	churned_customers bigint
1	Not subscribed	1616
2	Subscribed	1034

D, Add-On File Time Period:

This query should be run or refreshed monthly because this will match the frequency at which customers can discontinue services.

D1, Explanation of Time Period:

Letting a customer go rather than retaining them can be a significant cut into the profits of a company, as noted by Amy Gallo of Harvard Business Review: *"...acquiring a new customer is anywhere from five to 25 times more expensive than retaining an existing one" (**Gallo, 2014**).* Running this query monthly will ensure that the churn data remains current and will allow the company to monitor the effectiveness of its services, such as tech support service, in this case.

E, Panopto Video:

The pre-production code review will be submitted as directed in the PA requirements.

F, Web Sources: (code and other)

PostgreSQL. (n.d.). Documentation: 16: 5.4. Constraints. Retrieved May 20, 2024, from https://www.postgresql.org/docs/16/ddl-constraints.html#DDL-CONSTRAINTS-FK

PostgreSQL. (n.d.). Documentation: 16: BEGIN. Retrieved June 2, 2024, from https://www.postgresql.org/docs/16/sql-begin.html

Sewell, W. (2023). D205 SQL Sunday Presentation_default. Retrieved May 27, 2024, from https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=8f778f3a-0e6b-41d7-a052-b07d00ecceb5

Gallo, A. (2014, October 29). The Value of Keeping the Right Customers. Harvard Business Review. Retrieved May 27, 2024, from https://hbr.org/2014/10/the-value-of-keeping-the-right-customers