**Data Acquisition (TGM2) – D205**

**Performance Assessment**

**Western Governors University**

**By: Christian LeBlanc**

**Task A- Providing a question**

The question that I will be answering in this project is “What customers are using less than the average bandwidth and does not have Tech Support?” This question can help identify customers that could be needing of Tech Support. The reason I believe that these customers could be needing Tech Support is those that use less bandwidth might be less tech savvy because of lack of learning through using. I have previous experience in businesses that offered Tech Support and it is one of the higher margin products so this should be a service that should always be looking at ways to increase the number of customers with it.

**Task A1- Identifying data**

The question I proposed will require data from the customer table that is located in the churn database and the Services CSV file data. I will be using the bandwidth\_gp\_year column from the customer table to get the average. From there I will be using the customer\_id and bandwidth\_gp\_year columns from customer table and TechSupport column from the Services CSV data that will be uploaded as a table named Services. I will be using a join on customer\_id columns to connect the data in the two tables to see the customers that satisfy my question.

**Task B-Creating an ERD**

I used the services.csv file for the add-on to answer the question from Task A. I named the table I created with this data services. This table has customer\_id as primary key, as well as a foreign key. The entity relationship diagram created shows the relationships between all the preloaded tables in the churn database, but as I only needed one of these, customer, I will not go over the tables not used. The customer table uses customer\_id as a primary key. It also has 4 foreign keys each one relates it to a different preloaded table. I used three columns to get the answer for the question in task A. I used customer\_id which was used to join the customer table with the service table, the bandwidth\_pg\_year from the customer table, and the tech\_support column from the services table. Below is a picture of the ERD created from pgAdmin.

A screen shot of a graph

Description automatically generated

**Task B1- SQL code for table based on the ERD**

The services table that was created with the data from the services.csv file was created by using the create table feature in pgAdmin and produced the following code:

CREATE TABLE public.services

(

customer\_id text NOT NULL,

internet\_service text,

phone text,

multiple text,

online\_security text,

online\_backup text,

device\_protection text,

tech\_support text,

PRIMARY KEY (customer\_id),

CONSTRAINT 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;

This created a table with eight columns. The columns are named customer\_id, internet\_services, phone, multiple, online\_security, online\_backup, device\_protection, and tech\_support. They are all text data type, which worked for this assessment but other data types could be used to save space in more complex assignments. I made customer\_id a primary key. It needed to be a primary key to ensure that every customer has a unique ID and that no two customers can have the same one and to make sure that when a customer changes something in the services they receive that it updates their data and doesn’t create a second record. The customer\_id is also a foreign key with the column customer\_id in the customer table. This is done to ensure the referential integrity of the data within the customer\_id columns, and not allowing any customers to be listed in the services table that is not already in the customer table.

**Task B2- SQL code that loads data from CSV file into table**

The code created from pgAdmin when using the import feature to fill the services table with the data from the Services CSV file is below.

COPY services

FROM 'C:\LabFiles\Services.csv'

DELIMITER ','

CSV HEADER;

**Task C- SQL statement to answer question from Task A**

I needed to run two different statements to achieve the answer from the question in task A. I am sure this could have been done in one more efficient statement, but the two I used worked to give the answer. I ran the following to give me the average bandwidth used from all the customers.

SELECT AVG(bandwidth\_gp\_year) AS average\_bandwidth

From customer

This gave me the average bandwidth for the customers, which came out to be 3392.34154973529. I dropped the decimal places and used 3392 as the average as the question being answered does not need to be exact because it will just be used as a guide to help salesmen identify upselling opportunities.

SELECT customer.customer\_id, customer.bandwidth\_gp\_year, services.tech\_support

FROM services

JOIN customer ON services.customer\_id = customer.customer\_id

WHERE customer.bandwidth\_gp\_year < 3392 AND services.tech\_support = 'No';

**Task C1- Results from query**

I will have two files uploaded to show the results from my queries from Task C along with this document. I will also include below a picture of the first few rows of the results that answer my question when I opened them in Excel. I will only be showing some in the picture below because the results have 3,144 rows returned.

A screenshot of a table

Description automatically generated

**Task D- Time period for refreshing data**

I would recommend this data to be refreshed on a quarterly basis. This will reduce the number of customer that need to be reached out to about Tech Support if they have already decided to purchase it during the quarter as this should be a quarterly drive to increase Tech Support. More information about the bandwidth\_gp\_year would be useful, but for the purpose of this assignment I will assume that this number is on a rolling 12 month period and will be different each quarterly pull.

**Task D1- Why time period is relevant to business needs**

My reasoning for choosing the quarterly refreshing is customer happiness driven. The updated data pulls would coincide with salesmen reaching out to pitch Tech Support. I worry that anything sooner than quarterly would be off-putting to the customer. These calls can be a way to show the customers they are valuable by asking them about the current services then translate into offering Tech Support and other services.

**Task E- Panopto Video**

My Panopto video has me opening pgAdmin in a new lab and I lead a presentation of a recreation of how I accomplished all the steps that has been included in this submission.

**Task E1- Panopto Video describing**

I spoke aloud on my Panopto video describing the processes being completed as I did them. My video is saved in the “Master of Science, Data Analytics TGM1 | D205 (Student Creators) [assignments]” drop box. I emailed [assessmentservices@wgu.edu](mailto:assessmentservices@wgu.edu) because I could not find “Master of Science, Data Analytics TGM2 | D205 (Student Creators) [assignments]." I got a reply saying that it will not be an issue.

**Task F- Web data sources**

I did not use any web sources to acquire data or segments of third-party code in this Performance Assessment.

**Task G- Sources/ In-text citations/ References**

I did not use any sources, in-text citations, or references in this Performance Assessment.

**Task H- Professional Communication**

The content in this Performance Assessment is set up and presented with the highest professional standards.