D205 Data Acquisition, Performance Assessment

Alexa R. Fisher

Western Governors University

Degree: M.S. Data Analytics

Table of Contents

A. RESEARCH QUESTION

The research question selected for this thesis is, "**How many customers selected the internet service 'Fiber Optic' for each economic income class per state?"** This is an analysis of the number of customers per income classification based on the Pew Research Center (Bennett, 2021). These classifications are within income brackets named upper-income class, middle-income class, and lower-income class. The class system is based on a minimum family size of three members. The groupings are a breakdown based on Pew Research Center's data and inflation, excluding the location cost of living (Snider, 2021). The U.S. News & World Report has supplied a visual representation chart below (Snider, 2021)

INCOME GROUP	INCOME
Low income	Less than $52,200
Middle income	$52,200 - $156,600
Upper income	More than $156,600

Table 1: Economic Class System based on Income (Snider,2021).

The total count of these customers within the income groups is separated based on acquiring the internet service, "Fiber Optic." These results are further filtered to the state location. To solve this research question, a variety of means are used such as current income systems and the various data sets provided within the PostgreSQL churn public database as well as an addon CSV file. The original data sets include the customer and location data sets along with the add-on services CSV file. This research question can provide insight into the affordability and access

of fiber optic internet service across the United States. Affordability is determined if the count of customers in the lower-class bracket is apparent. Availability is visualized based on any amount found in each state location.

I. IDENTIFYING DATA

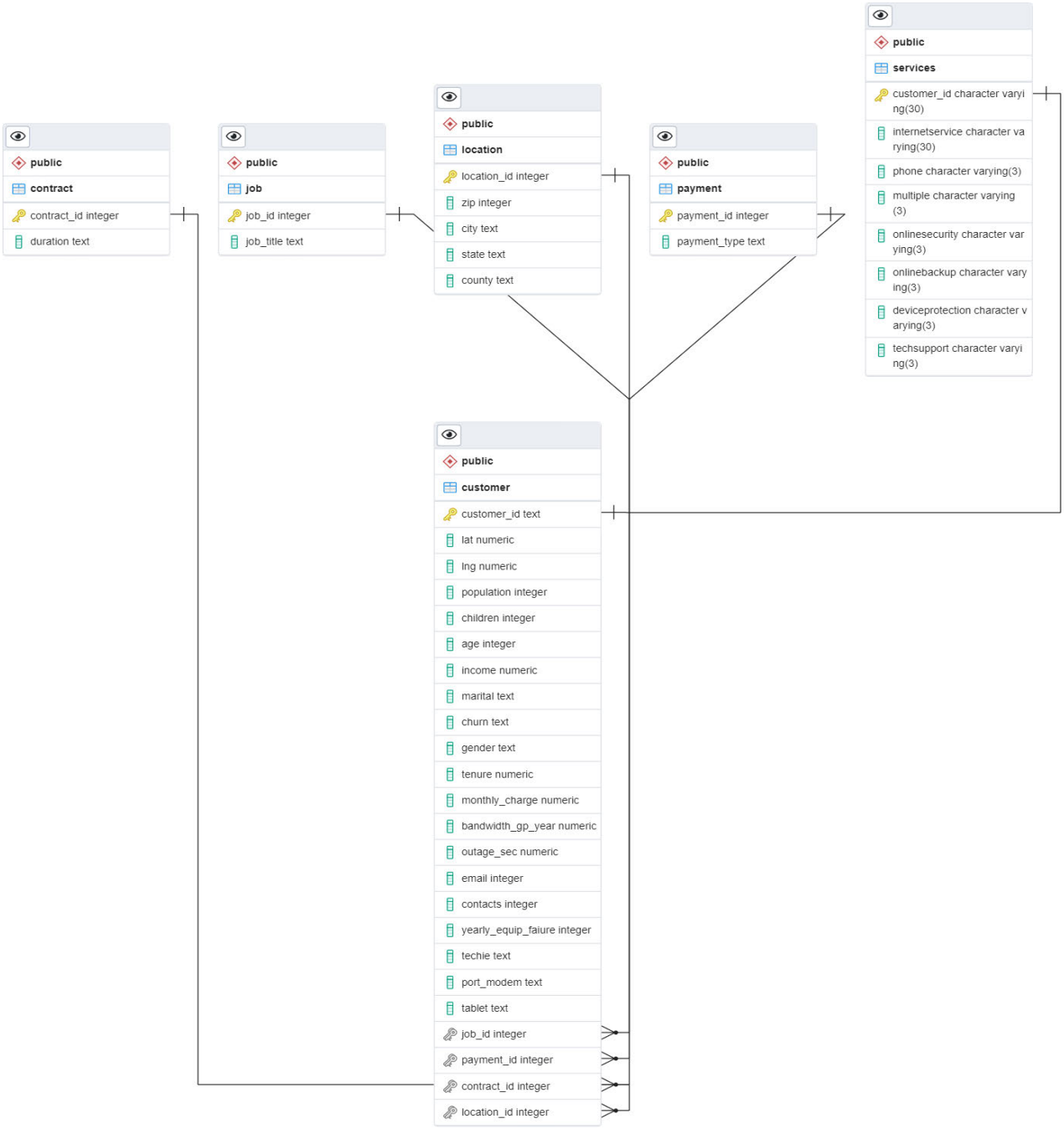
To answer the research question, there are various data sets needed to be used. The following will provide the sampling of data required within the Postgres database system. The *Customer* data set, the *Location* data set, as well the addon CSV file called services. The CSV file will be imported into the database as a newly created *Services* data set. Within these tables, the following attributes or columns can be queried to provide information to answer the research question: customer_id, income, children, marital status, location_id, state, and internet service. Join the Services data set and Customer data set on the customer_id attribute using the join feature. The *Location* data set will be joined with the *Customer* data set on the location_id attribute.

Table	Attribute
customer	customer_id
customer	children
customer	income
customer	location_id
customer	marital
location	location_id
location	state
services	customer_id
services	internetservice

B. LOGICAL DATA MODEL

The logical data model below is created using the original tables within the churn public database. The PostgreSQL pgadmin4 environment allows an entity relational diagram (ERD) model to be visualized. The ERD model consists of the *Customer* table with a primary key of

customer_id and various columns to include the foreign keys of contract_id, job_id, location_id, and payment_id. The foreign keys have a one-to-many relationship with the customer table. For example, one location is assigned to many customers. The only one-to-one relationship found in this database would be the newly added services table. There is only one service to one customer.



I. CODE FOR PHYSICAL DATA MODEL

The SQL code to create the Services table that accommodates the extension of the logical data model to the physical data model is the following:

```sql
-- Table: public.services

-- DROP TABLE IF EXISTS public.services;

CREATE TABLE IF NOT EXISTS public.services
(
    customer_id character varying(30) COLLATE pg_catalog."default" NOT NULL,
    internetservice character varying(30) COLLATE pg_catalog."default",
    phone character varying(3) COLLATE pg_catalog."default",
    multiple character varying(3) COLLATE pg_catalog."default",
    onlinesecurity character varying(3) COLLATE pg_catalog."default",
    onlinebackup character varying(3) COLLATE pg_catalog."default",
    deviceprotection character varying(3) COLLATE pg_catalog."default",
    techsupport character varying(3) COLLATE pg_catalog."default",
    CONSTRAINT services_pkey PRIMARY KEY (customer_id)
)

TABLESPACE pg_default;

ALTER TABLE IF EXISTS public.services
    OWNER to postgres;
```

The fields needed with this data set are customer id, internet service, phone, multiple, online security, online backup, device protection, and tech support. All the fields are VARCHAR datatypes, which means character varying and it allows both numbers and letters (Carchedi, n.d.). The customer id and internet service have a maximum character allotment of thirty characters. All the other attributes required a three-character limit to allow “yes” or “no” designations. The customer id is set as the primary key constraint that cannot be null or empty.

II. LOADING CSV DATA

The SQL statement needed to load the data from the add-on CSV file into the table is as such:

```
--command " "\\copy public.services (customer_id, internetservice, phone, multiple,
onlinesecurity, onlinebackup, deviceprotection, techsupport) FROM
'C:/Users/alexa/DOCUME~1/WGU/MSDA/D205/Services.csv' DELIMITER ',' CSV HEADER
QUOTE '\"' ESCAPE '";""
```

C. SQL QUERY

The SQL statements below provide the query result to the stated research question of "**How many customers selected the internet service 'Fiber Optic' for each economic income class per state?"** For this query, common table expressions or CTE were used. Common table expressions are temporary named result-sets that can be referred to within a SELECT statement (Carchedi, n.d.). Three CTEs were created for each economic class based on the monetary brackets provided by Pew Research Center (Bennett, 2021). In addition to the income brackets, the use of children and marital status was accounted for. Being married designated an adult pairing of two people, with all other marital statuses to be noted as one person. The children's field was considered to add the additional needed to make groupings of at least three members when combined with the adults derived from the marital column. All the needed data sets were combined using the JOIN statement and further filtered by the WHERE statement of internet service being fiber optic.

```sql
WITH
    upper_class AS
        (SELECT *
        FROM customer
        WHERE (income > 156600 AND children >= 1 AND marital = 'Married')
            OR (income > 156600 AND children >= 2 AND marital <> 'Married')
        ),
    middle_class AS
        (SELECT *
        FROM customer
        WHERE (income BETWEEN 52200 AND 156600 AND children >= 1 AND marital = 'Married')
            OR (income BETWEEN 52200 AND 156600 AND children >= 2 AND marital <> 'Married')
        ),
    lower_class AS
        (SELECT *
        FROM customer
        WHERE (income < 52200 AND children >= 1 AND marital = 'Married')
            OR (income < 52200 AND children >= 2 AND marital <> 'Married')
        )
SELECT
    loc.state AS state,
    COUNT(u.customer_id) AS upper_class_fiber,
    COUNT(m.customer_id) AS mid_class_fiber,
    COUNT(l.customer_id) AS lower_class_fiber
FROM customer AS c
LEFT JOIN upper_class AS u
ON c.customer_id = u.customer_id
LEFT JOIN middle_class AS m
ON c.customer_id = m.customer_id
LEFT JOIN lower_class AS l
ON c.customer_id = l.customer_id
LEFT JOIN location AS loc
ON c.location_id = loc.location_id
LEFT JOIN services AS s
ON c.customer_id = s.customer_id
WHERE s.internetservice = 'Fiber Optic'
GROUP BY loc.state
```

I. CSV FILES

Please see attached CSV data file: afisher_income_fiber.csv to view the results of the executed SQL query.

D. ADD-ON FILE

The add-on file should be updated at the time the customer data set is refreshed or annually at the onset of analysis by the organization. This will allow the most up-to-date data to be analyzed. It can provide marketing strategies geared to affordability and geographical locations.

E. SQL SCRIPT

The SQL script that performs the process of loading the add-on data, Services.csv.

```
\\copy public.services (customer_id, internetservice, phone, multiple, onlinesecurity,
onlinebackup, deviceprotection, techsupport) FROM
'C:/Users/alexa/DOCUME~1/WGU/MSDA/D205/Services.csv' DELIMITER ',' CSV HEADER
QUOTE '\"' ESCAPE '"';""
```

F. PANOPTO VIDEO

Please see attached Panopto video link. This is a video providing an overview of the PostgreSQL database system and its graphical user interface pgadmin4 environment (Group, n.d.). The recording will demonstrate the environment features such as Import/Export Data, Generate ERD, and the Query Tool. It will also show the execution of the SQL code to provide the research question results.

[redacted]

G. WEB SOURCES

No web sources were used to acquire data or segments of third-party code to support the application. All code and segments are original work.

H. RESOURCES

Bennett, J., Fry, R., & Kochhar, R. (2021, August 3). *Are you in the American middle class? find out with our income means*. Pew Research Center. Retrieved April 28, 2022, from https://www.pewresearch.org/fact-tank/2020/07/23/are-you-in-the-american-middle-class/

Carchedi, N., Grossenbacher, T., Sulmont, L., Ismay, C., Khalil, M., & Gorenshteyn, D. (n.d.). *D205 Data Acquisition Custom Track*. DataCamp. Retrieved April 28, 2022, from https://app.datacamp.com/learn/custom-tracks/custom-d205-data-acquisition

Group, P. S. Q. L. G. D. (2022, April 29). PostgreSQL. Retrieved April 28, 2022, from https://www.postgresql.org/

Snider, S., & Kerr, E. (2021, December 16). *Where do I fall in the American economic class system ...* US News. Retrieved April 29, 2022, from https://money.usnews.com/money/personal-finance/family-finance/articles/where-do-i-fall-in-the-american-economic-class-system