

**D205: Data Acquisition Task 1**

Justin Huynh

Student ID: 012229514

M.S. Data Analytics

May 12, 2024

## **A. Research Question**

My research question for this project is “How many people who use Fiber Optic service make at least six figures?”

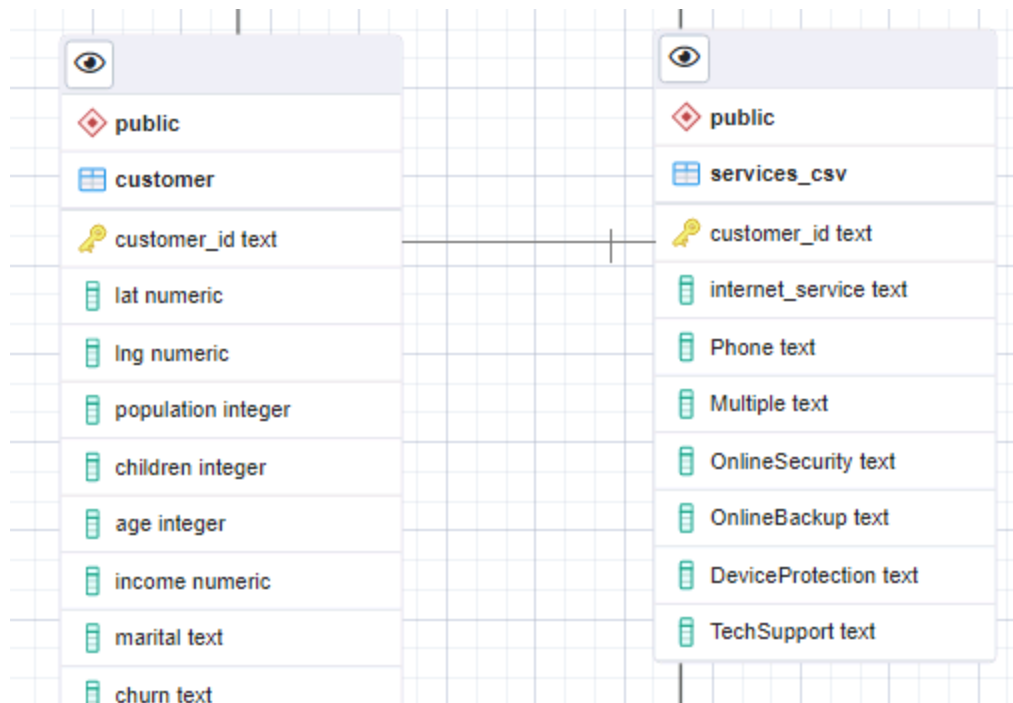
### **A1. Question Justification**

This question will be answered through combining the original customer table from the churn database with the add-on CSV file Services.csv in order to pull the necessary columns to identify exactly how many people who make at least six figures use this particular internet service.

### **A2. Identifying Data**

The research question requires two columns from the customer table in the original churn database and two columns from the Service.csv file. From the customer table, we'll be using the “customer\_id” and “income” fields while the Service.csv file will use the “customer\_id” and “InternetService” fields. Since we'll be joining the “customer\_id” field from both tables in order to conduct our research on the other two fields, we'll have only three columns in our created table to look at the relationship between the customers' income and the internet service they've chosen.

## **B: Entity Relationship Diagram (ERD)**



This ERD shows a 1:1 relationship between the customer table and the services\_csv table. pgAdmin 4 doesn't support a 1:1 relationship in their GUI tool. ("ERD Tool", pgAdmin Documentation)

[https://www.pgadmin.org/docs/pgadmin4/development/erd\\_tool.html](https://www.pgadmin.org/docs/pgadmin4/development/erd_tool.html)

### B1: Relationship Discussion

In this 1:1 relationship, the customer\_id fields from both tables are primary keys since they both have unique records that ties them to only one service record and there cannot be two customers with the same customer\_id. One issue we may come across is schema evolution. If the structure of one table changes via new data/fields imported, the same changes would need to be added to the corresponding table, which may cause more challenges during data migration such as increased time or data quality.

### B2: Statement For The ERD

```
CREATE TABLE public.services_csv (
    customer_id text NOT NULL,
```

```

    "internet_service" text NOT NULL,
    "Phone" text NOT NULL,
    "Multiple" text NOT NULL,
    "OnlineSecurity" text NOT NULL,
    "OnlineBackup" text NOT NULL,
    "DeviceProtection" text NOT NULL,
    "TechSupport" text NOT NULL,
    PRIMARY KEY (customer_id)
);

ALTER TABLE public.services_csv
    OWNER to postgres;

```

This script was generated in the GUI tool as I was importing the data from the add on Services.csv file. Only the "customer\_id" and "internet\_service" fields are relevant for my research question.

### **B3: Loading CSV Data**

```

COPY services_csv
FROM 'C:\LabFiles\Services.csv'
DELIMITER ','
CSV HEADER;

```

This will be used to add data from the Services.csv file to the services\_csv table.

### **C: SQL Query**

```

SELECT
    COUNT(*) AS six_figure_customers,

```




```

        s.internet_service
FROM customer AS c
INNER JOIN services_csv
        ON c.customer_id = s.customer_id
WHERE income > 100000
        AND internet_service = 'Fiber Optic'
GROUP BY s.internet_service;

```

This is the query that will achieve the desired results to answer our research question.

### C1: CSV Files

Data Output				Explain	Messages	Notifications
		six_figure_customers bigint			internet_service text	
1			174		Fiber Optic	

The result of the query answers the research question in this picture and was submitted with the report.

### D: Add-On File Time Period

Since our research question is based on understanding our six figure customers' internet service provider, I believe that a reasonable time period for refreshing the add on file to the database and keep the data relevant business activities relevant would be roughly around once a month but it could be subject to change to a longer timeframe such as every two or three months depending on the goal of the business.

**D1: Explanation Of Time Period**

A time period of one to three months for a refresh would seem adequate to identify trends of our six figure customers and what type of internet service they use. Since customers don't change their internet services often, we wouldn't want to set a precedent that they'll be contacted about switching providers too often. That is why I believe this time period would reflect enough time to see how many of the customers with a higher income bracket of six figures use fiber optic and potentially upsell them with a faster but more expensive service such as dark fiber or dedicated leased line.

**E: Panopto Video**

The URL link for the Panopto video can be found [here](#). It will also be submitted in the Performance Assessment task submission and has been uploaded.

**F: Web Sources**

1. "ERD Tool." Retrieved from  
[https://www.pgadmin.org/docs/pgadmin4/development/erd\\_tool.html](https://www.pgadmin.org/docs/pgadmin4/development/erd_tool.html)
2. "Import CSV File Into PostgreSQL Table." Retrieved from  
<https://www.postgresqltutorial.com/postgresql-tutorial/import-csv-file-into-posgresql-table/>