

## D205 Data Acquisition, PA -TGM1 - Task 1

Sean Simmons

**A. “Summarize a research question that can be answered using *both* the original database and the add-on CSV data. The question should require data from *both* these data sources.**

**1. Identify which data from the original data set and the add-on CSV file are needed to answer the research question.”**

The tables found within the churn database contain customer information and demographics. Included in the information were the gender and types of services purchased by the customers. Which led me to ask, Do customers who identify as male purchase device protection more than customers that identify as female? The data customer\_id and gender from the original data set and customer\_id and device protection data from the add-on CSV file are required to answer this research question because they provide the gender specifics and services purchased, which will answer my research question.

There is a 1.65% difference, whose relevance is to be determined by the business needs.

**B. “Create a logical data model for the add-on CSV file by evaluating the data contained in the file and emphasizing the relational constraints.**

**1. Write SQL code that creates a table that accommodates the extension of the logical data model to a physical data model by specifying the field types and relevant keys.**

**2. Write SQL code that loads the data from the add-on CSV file into the table created in part B1.”**

SQL Query 1:

```
1 /* This section creates a table for the services CSV file,
2 sets the key, and gives ownership */
3 CREATE TABLE services
4 (
5 customer_id character varying(30) NOT NULL,
6 inernetservice character varying(30),
7 phone character varying(3),
8 multiple character varying(3),
9 onlinesecurity character varying(3),
10 onlinebackup character varying(3),
11 deviceprotection character varying(3),
12 techsupport character varying(3),
13 CONSTRAINT customer_id
14 Primary KEY (customer_id)
15 )
16 TABLESPACE pg_default;
17
18 ALTER TABLE public.services
19 OWNER to postgres;
20 SELECT * FROM services;
```

	customer_id [PK] character varying (30)	inernetservice character varying (30)	phone character varying (3)	multiple character varying (3)	onlinesecurity character varying (3)	onlinebackup character varying (3)	deviceprotection character varying (3)	techsupport character varying (3)
1	A00088	Fiber Optic	Yes	Yes	Yes	No	No	No

### SQL Query 1 Code:

/\* This section creates a table for the services CSV file,  
sets the key, and gives ownership \*/

```
CREATE TABLE services  
(  
customer_id character varying(30) NOT NULL,  
internetservice character varying(30),  
phone character varying(3),  
multiple character varying(3),  
onlinesecurity character varying(3),  
onlinebackup character varying(3),  
deviceprotection character varying(3),  
techsupport character varying(3),  
CONSTRAINT customer_id  
Primary KEY (customer_id)  
)  
TABLESPACE pg_default;
```

```
ALTER TABLE public.services  
OWNER to postgres;  
SELECT * FROM services;
```

### SQL Query 2:

churn/postgres@PostgreSQL 13

Query EditorQuery History

```
1 --import with Header yes and delimiter using the import function or with the code below  
2 -- COPY services FROM 'c:\Labfiles\services.csv' DELIMITER ',' csv header;  
3 SELECT * FROM services
```

Data OutputMessagesNotificationsExplain

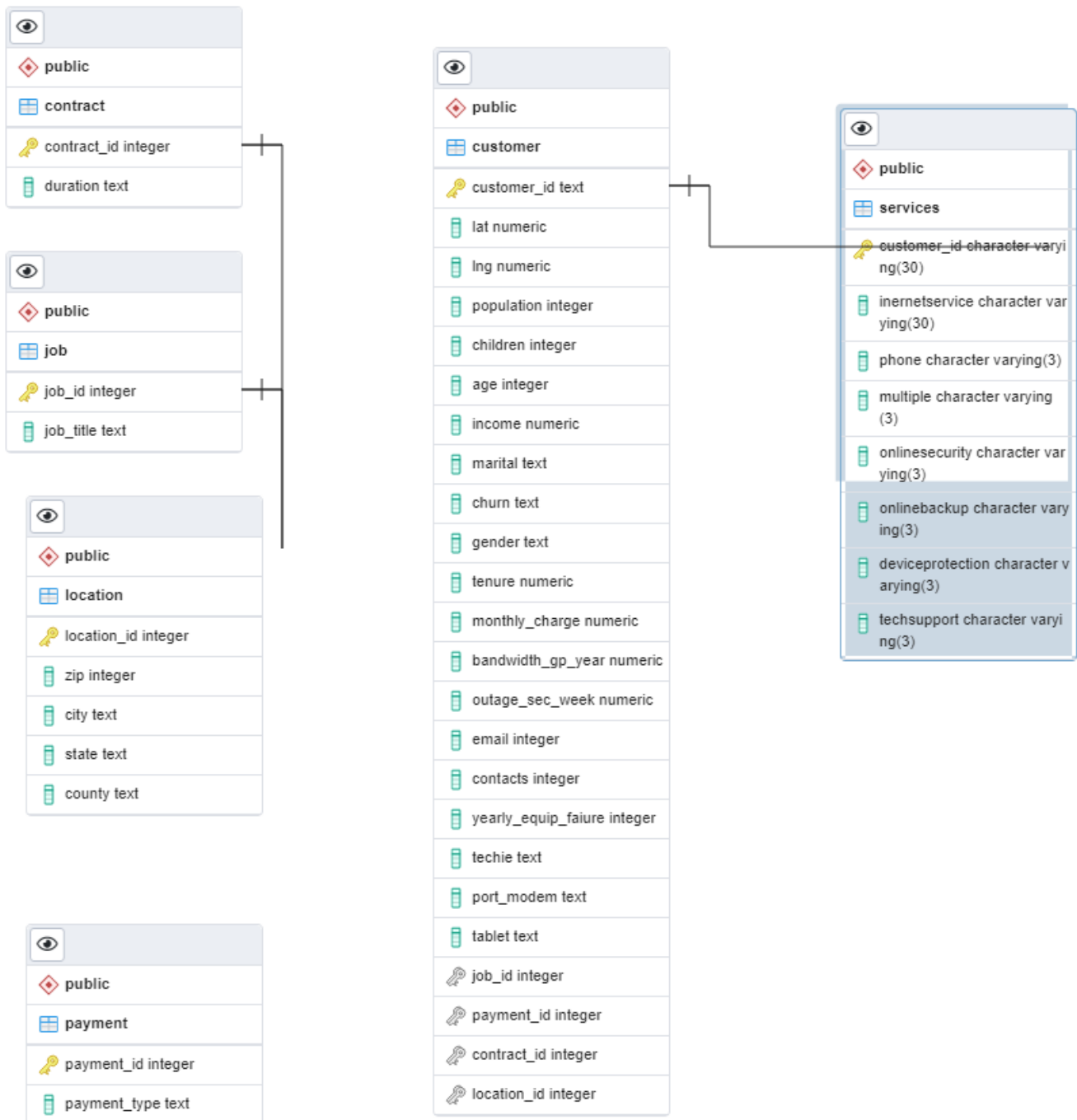
	customer_id [PK] character varying (30)	internetservice character varying (30)	phone character varying (3)	multiple character varying (3)	onlinesecurity character varying (3)	onlinebackup character varying (3)	deviceprotection character varying (3)	techsupport character varying (3)
1	A00088	Fiber Optic	Yes	Yes	Yes	No	No	No
2	A04204	DSL	Yes	Yes	Yes	Yes	Yes	Yes
3	A04378	None	Yes	Yes	No	No	No	Yes
4	A04830	DSL	Yes	Yes	Yes	No	Yes	No
5	A05946	Fiber Optic	No	No	No	No	No	No
6	A06667	DSL	Yes	Yes	No	Yes	Yes	Yes
7	A08755	Fiber Optic	Yes	No	No	Yes	Yes	Yes
8	A100934	DSL	Yes	No	Yes	No	No	No
9	A103291	DSL	No	Yes	No	No	No	Yes
10	A105398	DSL	Yes	No	Yes	No	Yes	Yes
11	A106835	Fiber Optic	Yes	Yes	No	No	Yes	No
12	A118391	DSL	Yes	No	No	No	No	No
13	A121664	Fiber Optic	Yes	Yes	Yes	No	No	Yes
14	A124923	None	Yes	No	No	Yes	No	Yes
15	A125688	DSL	Yes	Yes	Yes	Yes	Yes	Yes

### SQL Query 2 Code:

```
--import with Header yes and delimiter using the import function or with the code below
```

```
COPY services FROM 'c:\Labfiles\services.csv' DELIMITER ',' csv header;
SELECT * FROM services
```

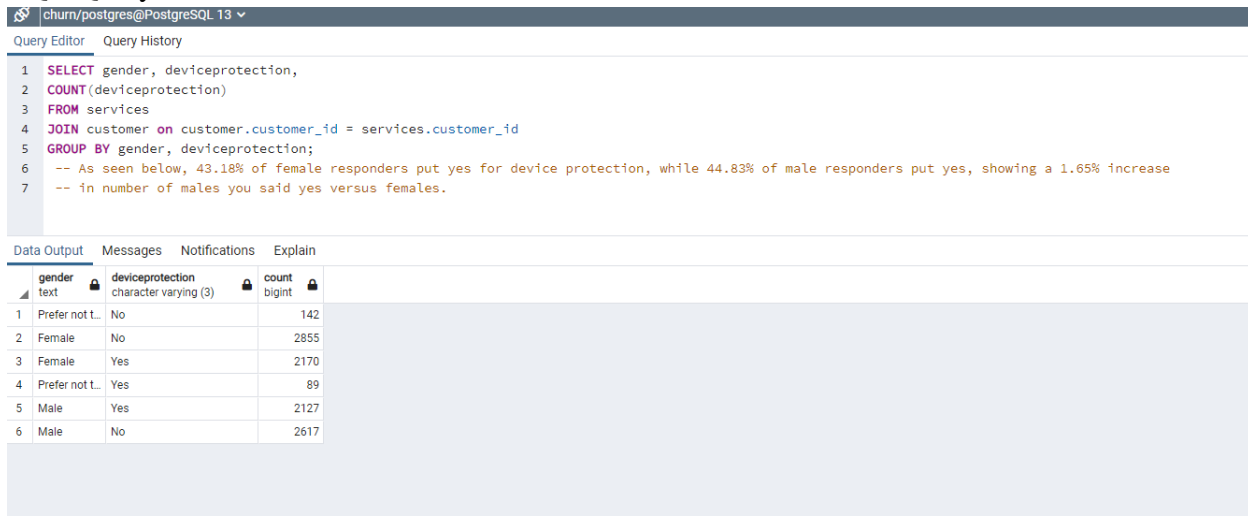
ERD: The new table is on the right side of the primary customer table



**C. “Write SQL statement(s) for a query or queries that inform the research question summarized in part A.**

**1. Provide a CSV file or files that capture the results from the query or queries.”**

SQL Query:

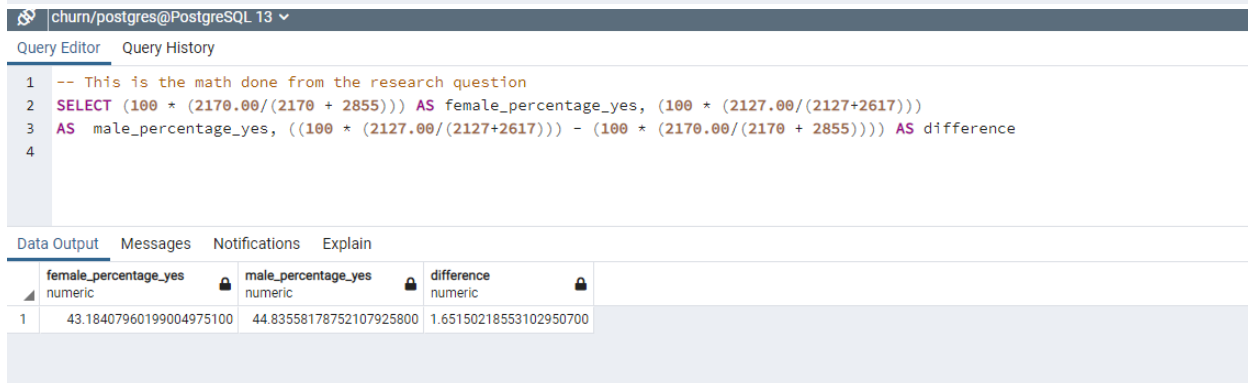


The screenshot shows a PostgreSQL Query Editor interface. The query editor has tabs for "Query Editor" and "Query History". The query is as follows:

```
1 SELECT gender, deviceprotection,
2 COUNT(deviceprotection)
3 FROM services
4 JOIN customer on customer.customer_id = services.customer_id
5 GROUP BY gender, deviceprotection;
6 -- As seen below, 43.18% of female responders put yes for device protection, while 44.83% of male responders put yes, showing a 1.65% increase
7 -- in number of males you said yes versus females.
```

Below the query editor, there are tabs for "Data Output", "Messages", "Notifications", and "Explain". The "Data Output" tab is selected, showing a table with 3 columns: "gender", "deviceprotection", and "count".

gender	deviceprotection	count
1 Prefer not to answer	No	142
2 Female	No	2855
3 Female	Yes	2170
4 Prefer not to answer	Yes	89
5 Male	Yes	2127
6 Male	No	2617



The screenshot shows a PostgreSQL Query Editor interface. The query editor has tabs for "Query Editor" and "Query History". The query is as follows:

```
1 -- This is the math done from the research question
2 SELECT (100 * (2170.00/(2170 + 2855))) AS female_percentage_yes, (100 * (2127.00/(2127+2617)))
3 AS male_percentage_yes, ((100 * (2127.00/(2127+2617))) - (100 * (2170.00/(2170 + 2855)))) AS difference
4
```

Below the query editor, there are tabs for "Data Output", "Messages", "Notifications", and "Explain". The "Data Output" tab is selected, showing a table with 4 columns: "female\_percentage\_yes", "male\_percentage\_yes", and "difference".

female_percentage_yes	male_percentage_yes	difference
43.18407960199004975100	44.83558178752107925800	1.65150218553102950700

SQL Query Code:

First Query

SELECT gender, deviceprotection,

COUNT(deviceprotection)

FROM services

JOIN customer on customer.customer\_id = services.customer\_id

GROUP BY gender, deviceprotection;

-- As seen below, 43.18% of female responders put yes for device protection, while 44.83% of male responders put yes, showing a 1.65% increase

-- in number of males you said yes versus females.

Second Query

```
-- This is the math done from the research question
SELECT (100 * (2170.00/(2170 + 2855))) AS female_percentage_yes, (100 *
(2127.00/(2127+2617)))
AS male_percentage_yes, ((100 * (2127.00/(2127+2617))) - (100 * (2170.00/(2170 + 2855))))
AS difference
```

CSV file: That is the CSV file that was submitted with this document.

**D. “Determine 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 research question.”**

My research question asks about the device protection purchase and the gender of the customers. As this is a general question, the add-on file can be acquired and refreshed monthly to continue to monitor the gender norms of the business to better understand their customers and behaviors. Any longer and it does not provide the business with enough information to accurately predict how gender influences their sales. Any shorter and it provides an excess of data that may not accurately reflect the causation we are looking for since any number of variables could affect the weekly service sales.

**E. “Create an SQL script that performs the process of loading the add-on data.”**

SQL Query:

```
--import with Header yes and delimiter using the import function or with the code below,
COPY services FROM 'c:\Labfiles\services.csv' DELIMITER ',' csv header;
SELECT * FROM services
```

churn/postgres@PostgreSQL 13									
Query Editor   Query History									
<pre> 1  --import with Header yes and delimiter using the import function or with the code below 2  -- COPY services FROM 'c:\Labfiles\services.csv' DELIMITER ',' CSV HEADER; 3  SELECT * FROM services </pre>									
Data Output   Messages   Notifications   Explain									
	customer_id [PK] character varying (30)	internetservice character varying (30)	phone character varying (3)	multiple character varying (3)	onlinesecurity character varying (3)	onlinebackup character varying (3)	deviceprotection character varying (3)	techsupport character varying (3)	
1	A00088	Fiber Optic	Yes	Yes	Yes	No	No	No	
2	A04204	DSL	Yes	Yes	Yes	Yes	Yes	Yes	
3	A04378	None	Yes	Yes	No	No	No	Yes	
4	A04830	DSL	Yes	Yes	Yes	No	Yes	No	
5	A05946	Fiber Optic	No	No	No	No	No	No	
6	A06667	DSL	Yes	Yes	No	Yes	Yes	Yes	
7	A08755	Fiber Optic	Yes	No	No	Yes	Yes	Yes	
8	A100934	DSL	Yes	No	Yes	No	No	No	
9	A103291	DSL	No	Yes	No	No	No	Yes	
10	A105398	DSL	Yes	No	Yes	No	Yes	Yes	
11	A106835	Fiber Optic	Yes	Yes	No	No	Yes	No	
12	A118391	DSL	Yes	No	No	No	No	No	
13	A121664	Fiber Optic	Yes	Yes	Yes	No	No	Yes	
14	A124923	None	Yes	No	No	Yes	No	Yes	
15	A125688	DSL	Yes	Yes	Yes	Yes	Yes	Yes	

**F. “Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the programming environment.”**

Link to the video that can also be found with my submission:

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=5f784e36-f81d-4179-a260-af85015e26c9>

**G. “Record the web sources used to acquire data or segments of third-party code to support the application. Be sure the web sources are reliable.”**

I did not use web sources as described above. All lines of code were informed from prior knowledge and learning in this course. However, the churn database, tables, and add-on files were provided to me by WGU as a part of the digital lab environment to complete this assignment.

**H. “Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.”**

I did not use sources as described above. However, the text in bold is pulled directly from my performance assessment task that was assigned to me through the D205 course. I have decided to use it to make this assignment clear and concise.

**I. “Demonstrate professional communication in the content and presentation of your submission.”**

This aspect of the performance assessment carries throughout this document and cannot be summarized.