D205: Data Acquisition

Lucas Benish

Western Governor’s University

# Summarize a research question that can be answered with *both* the original database and add-on csv data.

## A1. Identify which data from the original data set and add-on csv file are necessary to answer the question.

The research question that I am interested in answering is: Which customers in each state may respond to an ad campaign to upgrade their internet service, assuming it is likely they may upgrade if they already utilize more than one service (internet and phone) but do not have fiber optic internet, and have both monthly charges below the average and incomes above the average?

Answering this question will require me to determine which clients have both phone and internet services, have internet service that is not fiber, have incomes above the group average and monthly payments below the group average, and determine what state each customer lives in. Therefore, I’ll require the following columns from the customer and services tables:

customer.customer\_id, customer.monthly\_charge, customer.income, customer.tenure, services2.customer\_id, services2.phone, services2.internet, location.customer\_id, and location.state.

# Create a logical data model for the add-on csv file by evaluating the data contained in the file and emphasizing relational constraints.

Please see the attached D205 ERD.pdf for this. This file demonstrates the relationships between the entities of customer, location, job, contract, payment, and services2.

Each relationship is a one-to-many relationship, with the “many” relating to the customer entity. The reason for this is that each of the attributes in the outlying entities (job, contract, location, services2, payment) can correlate to multiple entries from the customer table, but a specific attribute from the customer table can only relate to a single entry from the outer entities.

For example, each customer can have a single “job\_id,” but the same “job\_id” can be utilized for multiple customers. This same can be said for the attributes of contract, payment, location, and services2. Each service can be given to multiple customers, but each customer only has a single list of services utilized (RelationalDBDesign, n.d.).

## B1. 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 type and relevant keys.

The code that I used to create my table is as follows:

CREATE TABLE services (

customer\_id text PRIMARY KEY,

Internet text,

Phone text,

Multiple text,

OnlineSecurity text,

OnlineBackup text,

DeviceProtection text,

TechSupport text);

## B2. Write SQL code that loads the data from the add-on csv file into the table created in B1.

The code that I used to copy the table data from the CSV file is as follows:

COPY services FROM ‘C:\Users\LabUser\Desktop\Services.csv’ DELIMITER ‘,’ CSV HEADER;

I found this information in an email from Dr. Gagner that was sent out in previous weeks. The specific portion of the email is below:

“**You can write a script like this:**

**COPY table\_name from ‘c:\users\labuser\desktop\filename.csv’ DELIMITER ‘,’ CSV HEADER;”** (D. Gagner, personal communication, 12/31/21)

# Write SQL statements for a query that inform the research question asked in part A.

The query that I ran to investigate my research question was as follows:

SELECT c.customer\_id, l.state AS state

FROM customer AS c

INNER JOIN location AS l

USING (location\_id)

--selecting customers with below avg charges, above avg income

--been with company for 2 years, and have both phone/internet but NOT fiber

WHERE customer\_id IN

(

SELECT customer\_id

FROM services2

INNER JOIN customer

USING (customer\_id)

WHERE “Internet” = ‘DSL’

AND “Phone” = ‘Yes’

AND tenure > 2

AND monthly\_charge < (SELECT AVG(monthly\_charge) FROM customer)

AND income > (SELECT AVG(income) FROM customer)

)

ORDER BY state;

In order to perform this query appropriately, I needed to determine why I was consistently getting error messages. I learned from StackOverflow that PostgreSQL requires quotations around columns that have “CamelCase” in their headers (StackOverflow, 2018).

## C1. Provide a csv file or files that capture the results of the query or queries.

Please see the attached CustWithDSL\_Phone.CSV file for the results of the above query. This query gives results of every customer\_ID and their associated state. This query could be investigated further and each customer\_id could be counted to show which states have the greatest number of potential respondents. I did not do that currently in order to show each individual customer\_id.

In order to create a CSV file, I first created a new folder under C:\users\public and then utilized a COPY TO script on C:\users\public\publicfolder\CustWithDSL\_Phone.CSV with commas as the delimiter, and then emailed myself the file (Sainsbury, n.d., and DataCamp, 2021).

# 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 research question.

I think that at a minimum that this information should be refreshed monthly, and may benefit from being refreshed every two weeks as well. This will capture new results for individuals whose tenure has only recently exceeded 2 years, or whose income may have changed relative to the average.

# Create a SQL script that performs the process of loading the add-on data.

The code that I used to copy the table data from the CSV file is as follows:

COPY services FROM ‘C:\Users\LabUser\Desktop\Services.csv’ DELIMITER ‘,’ CSV HEADER; (D. Gagner, personal communication, 12/31/2021)

# 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.

# Record the web sources used to acquire data or segments of third-party code to support the application.

# Acknowledge sources, using in-text citations for reference.

Please see the above in-text citations and source page below.

# Demonstrate professionalism in content and presentation of performance assessment.

Sources

1. *Three Types of Relationships in an ERD Diagram.* RelationalDBDesign. (n.d.). Retrieved January 6, 2022, from <https://www.relationaldbdesign.com/database-design/module6/three-relationship-types.php>.
2. D. Gagner, personal communication, 12/31/2021
3. *Postgresql Column Doesn’t Exist [duplicate].* StackOverflow. (2018, August 24). Retrieved January 6, 2022, from <https://stackoverflow.com/questions/52007364/postgresql-column-doesnt-exist>.
4. Niel Sainsbury (n.d.) *Permission Denied error when using PostgreSQL’s COPY FROM/TO command.* Retrieved January 6, 2022, from <https://www.neilwithdata.com/copy-permission-denied>.
5. *How to export postgresql data to a CSV or Excel File*. The Data School. (2021, August 29). Retrieved January 6, 2022, from <https://dataschool.com/learn-sql/how-to-export-data-to-csv-or-excel/>.