# D-211 ADVANCED DATA ACQUISITION

PAOLA WILLIAMS

COLLEGE OF IT, WESTERN GOVERNOR’S UNIVERSITY

APRIL 6th, 2023

## Table of Contents

## Part I: Interactive Data Dashboard

### A. Interactive data dashboard

*A1. Datasets*

The datasets chosen are the churn dataset provided by WGU and an additional dataset “Per capita Income by County (2021) vs. Education” (Kaggle, 2021). Both were merged into a single dataset using PostgreSQL, then the tables in the new dataset were used to build a dashboard on Tableau.

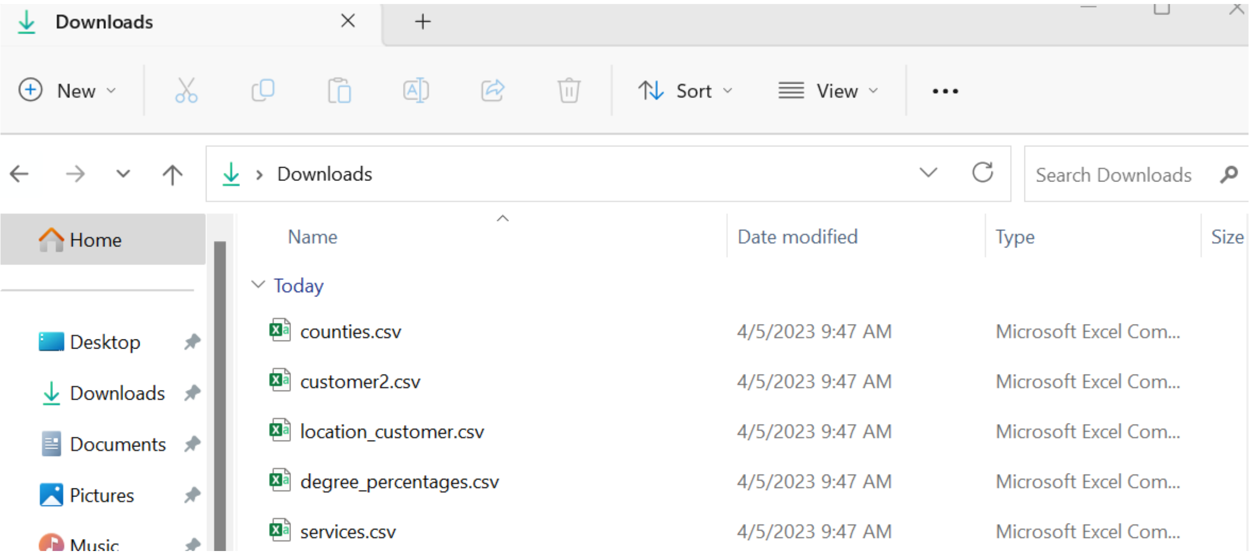
*A2. Installation instructions*

Use the following instructions to install the dashboard in Tableau:

1. Download the csv files to the local computer.

**Figure 1**

*CSV files in the Downloads folder*

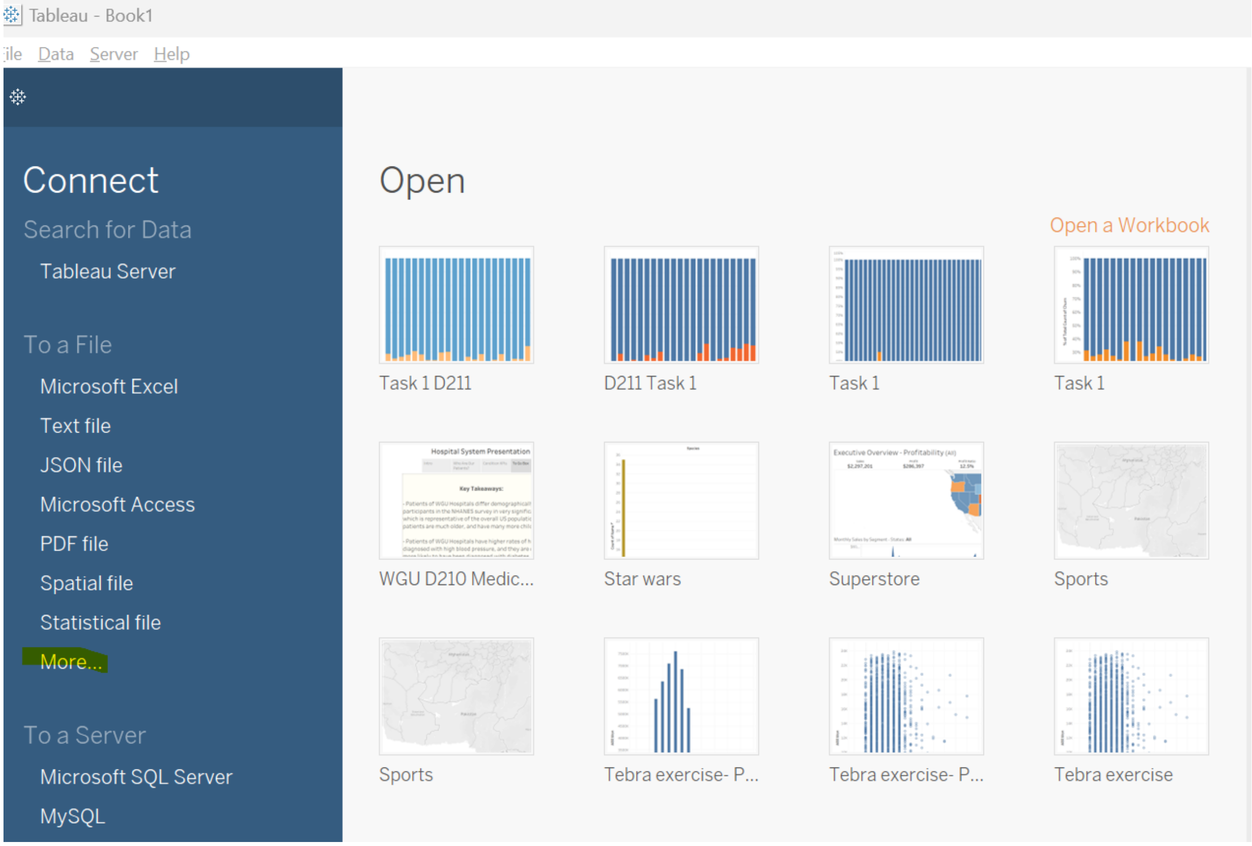


| Name | Date modified | Type | Size |
|---|---|---|---|
| counties.csv | 4/5/2023 9:47 AM | Microsoft Excel Com... | |
| customer2.csv | 4/5/2023 9:47 AM | Microsoft Excel Com... | |
| location_customer.csv | 4/5/2023 9:47 AM | Microsoft Excel Com... | |
| degree_percentages.csv | 4/5/2023 9:47 AM | Microsoft Excel Com... | |
| services.csv | 4/5/2023 9:47 AM | Microsoft Excel Com... | |

2. Double click the Tableau file Task 1 D211 to open.
3. Click on “Data” on the header.
4. Select “New Data Source” and you will get a Connect window. Click on “More”.

**Figure 2**

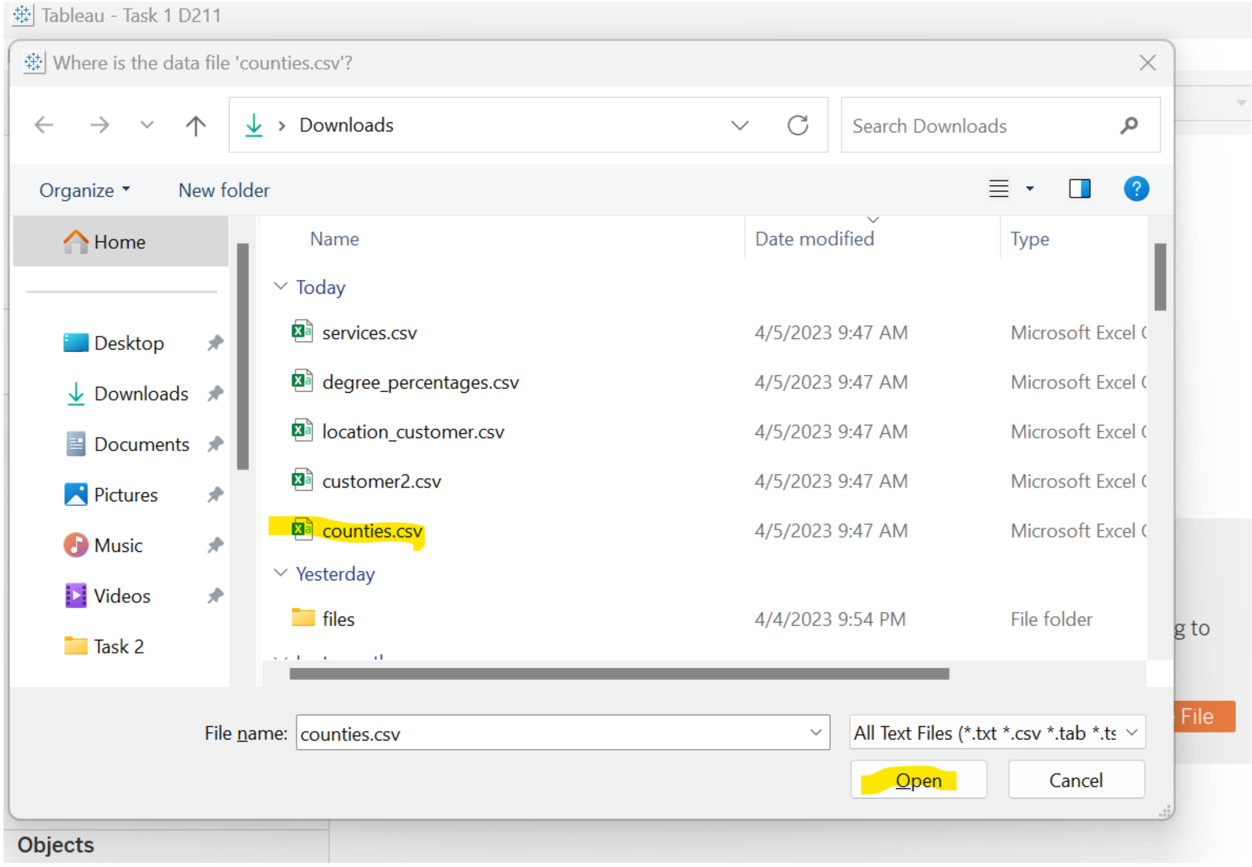
*Connect to data source (csv files)*



5. Choose the “counties.csv” file and click the “Open” button.

**Figure 3**

*Counties.csv file in the Downloads folder*



| Name | Date modified | Type |
|---|---|---|
| Today | | |
| services.csv | 4/5/2023 9:47 AM | Microsoft Excel ( |
| degree_percentages.csv | 4/5/2023 9:47 AM | Microsoft Excel ( |
| location_customer.csv | 4/5/2023 9:47 AM | Microsoft Excel ( |
| customer2.csv | 4/5/2023 9:47 AM | Microsoft Excel ( |
| counties.csv | 4/5/2023 9:47 AM | Microsoft Excel ( |
| Yesterday | | |
| files | 4/4/2023 9:54 PM | File folder |

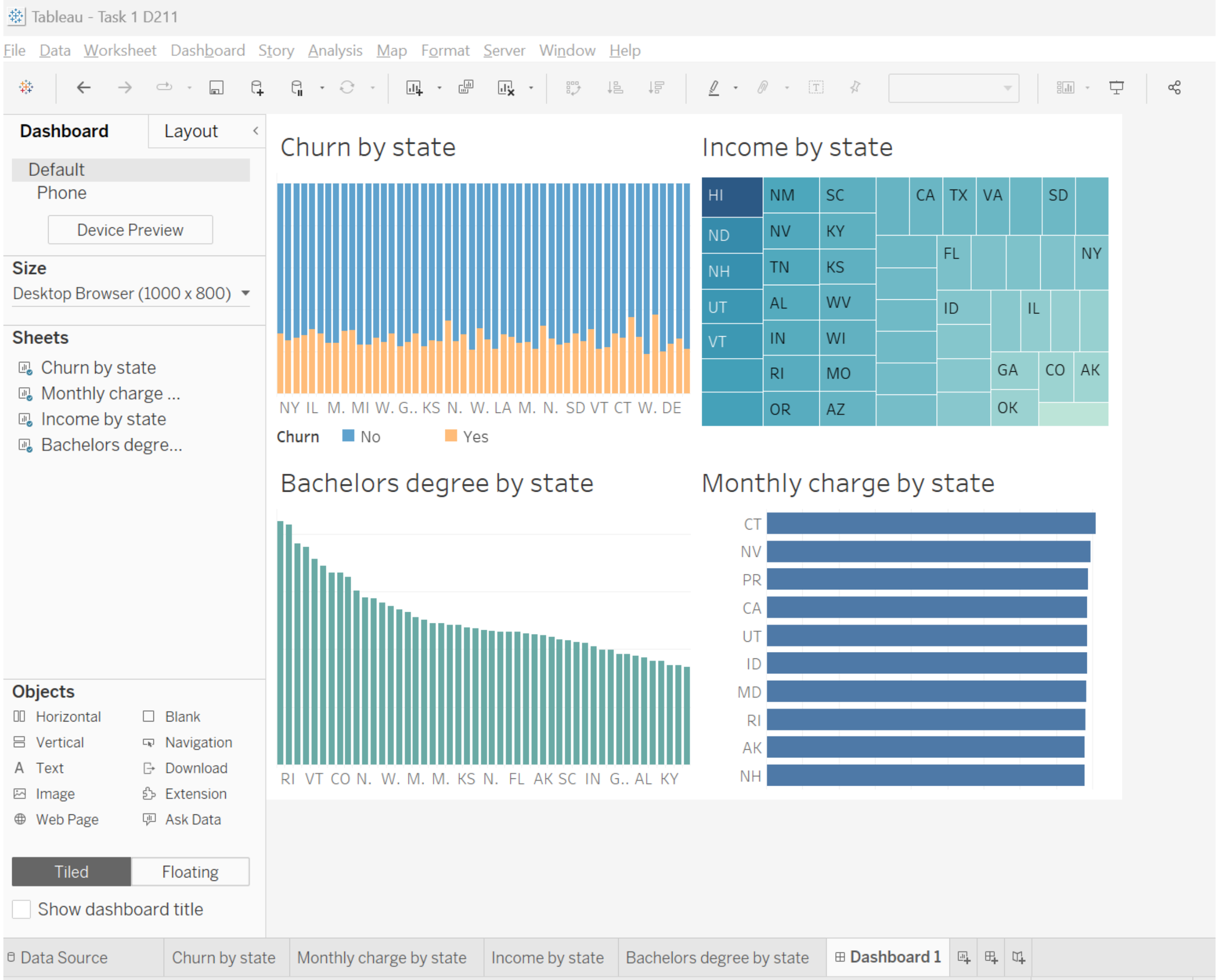
File name: counties.csv

All Text Files (*.txt *.csv *.tab *.ts

6. The dashboard will appear when the connection to the csv files is established.

**Figure 4**

*Dashboard in Tableau*



*A3. Navigation instructions*

The dashboard includes 4 different charts based on the two joined datasets mentioned before.

Churn numbers by state shows a bar chart with the percentage of churn (Yes or No) by state. The tooltip allows the user to check the percentages and the chart itself helps visualize the distribution of churn in a single state.

The user can also see which states have the highest churn and which states have the lowest churn.

Monthly charged by state shows how much in average is charged to customers. The user can hover over the bars to check the average amount charged by state.

Average income by state is a heatmap which highlights which states have the highest income.

In the case of bachelor's degree by state, we are using a sorted bar chart with the bachelor's degree percentages by state. The user may want to check if there is a relationship in bachelor's degree percentage and the average income by state.

*A3. SQL Code*

1. The following code was used in PG Admin:

```
--------------------------------------------------------------------------
-- PART 1 Create the tables in Churn
--------------------------------------------------------------------------
DROP TABLE IF EXISTS all_fields;
CREATE TABLE all_fields (
   customer_id text,
   lat numeric,
   lng numeric,
   population integer,
   children integer,
   age integer,
   income numeric,
   marital text,
   churn text,
   gender text,
   tenure numeric,
   monthly_charge numeric,
   bandwidth_gp_year numeric,
   outage_sec_week numeric,
   email integer,
   contacts integer,
   yearly_equip_faiure integer,
   techie text,
   port_modem text,
   tablet text,
   duration text,
   job_title text,
   zip integer,
   city text,
   state text,
   county text,
   payment_type text
);
```

```
INSERT INTO all_fields (
   customer_id, lat, lng, population, children, age, income, marital, churn, gender, tenure,
monthly_charge, bandwidth_gp_year, outage_sec_week, email, contacts, yearly_equip_faiure,
techie, port_modem, tablet,
   duration,
   job_title,
   zip, city, state, county,
   payment_type
)
SELECT
   customer.customer_id, customer.lat, customer.lng, customer.population, customer.children,
customer.age, customer.income, customer.marital, customer.churn, customer.gender,
customer.tenure, customer.monthly_charge, customer.bandwidth_gp_year,
customer.outage_sec_week, customer.email, customer.contacts,
customer.yearly_equip_faiure, customer.techie, customer.port_modem, customer.tablet,
   contract.duration,
   job.job_title,
   location.zip, location.city, location.state, location.county,
   payment.payment_type
FROM customer
LEFT OUTER JOIN job ON customer.job_id = job.job_id
LEFT OUTER JOIN location ON customer.location_id = location.location_id
LEFT OUTER JOIN contract ON customer.contract_id = contract.contract_id
LEFT OUTER JOIN payment ON customer.payment_id = payment.payment_id;

DROP TABLE IF EXISTS services;
DROP TABLE IF EXISTS location_customer;
DROP TABLE IF EXISTS customer2;

CREATE TABLE customer2 (
   customer_id text COLLATE pg_catalog."default" NOT NULL PRIMARY KEY,
   children integer,
   age integer,
   income numeric,
   marital text COLLATE pg_catalog."default",
   gender text COLLATE pg_catalog."default",
   techie text COLLATE pg_catalog."default",
   tablet text COLLATE pg_catalog."default",
   job_title text COLLATE pg_catalog."default"
);

INSERT INTO customer2 SELECT customer_id, children, age, income, marital, gender, techie,
tablet, job_title FROM all_fields;
```

```
CREATE TABLE services (
    id serial,
    churn text COLLATE pg_catalog."default",
    tenure numeric,
    monthly_charge numeric,
    bandwidth_gp_year numeric,
    outage_sec_week numeric,
    email integer,
    contacts integer,
    yearly_equip_faiure integer,
    port_modem text COLLATE pg_catalog."default",
    duration text COLLATE pg_catalog."default",
    payment_type text COLLATE pg_catalog."default",
    customer_id text COLLATE pg_catalog."default" UNIQUE,
    CONSTRAINT services_pkey PRIMARY KEY (id),
    CONSTRAINT services_customer_id_fkey FOREIGN KEY (customer_id)
        REFERENCES public.customer2 (customer_id) MATCH SIMPLE
        ON UPDATE NO ACTION
        ON DELETE NO ACTION
        NOT VALID
);

INSERT INTO services(churn, tenure, monthly_charge, bandwidth_gp_year, outage_sec_week,
email, contacts, yearly_equip_faiure, port_modem, duration, payment_type, customer_id)
SELECT churn, tenure, monthly_charge, bandwidth_gp_year, outage_sec_week, email,
contacts, yearly_equip_faiure, port_modem, duration, payment_type, customer_id FROM
all_fields;

CREATE TABLE location_customer (
    id serial PRIMARY KEY,
    lat numeric,
    lng numeric,
    zip integer,
    city text COLLATE pg_catalog."default",
    state text COLLATE pg_catalog."default",
    county text COLLATE pg_catalog."default"
);

INSERT INTO location_customer(lat,lng,zip,city,state,county)
SELECT DISTINCT lat, lng, zip, city, state, county FROM all_fields;
```

```
--------------------------------------------------------------------------
-- PART 2 Create tables for additional dataet
--------------------------------------------------------------------------

DROP TABLE IF EXISTS education_vs_per_capita_income;
CREATE TABLE education_vs_per_capita_income (
   county_fips INT NOT NULL,
   state VARCHAR(2) NOT NULL,
   county VARCHAR(50) NOT NULL,
   per_capita_personal_income_2019 INT NOT NULL,
   per_capita_personal_income_2020 INT NOT NULL,
   per_capita_personal_income_2021 INT NOT NULL,
   associate_degree_numbers_2016_2020 INT NOT NULL,
   bachelor_degree_numbers_2016_2020 INT NOT NULL,
   associate_degree_percentage_2016_2020 NUMERIC(4, 2) NOT NULL,
   bachelor_degree_percentage_2015_2019 NUMERIC(4, 2) NOT NULL
);

COPY education_vs_per_capita_income FROM 'C:\Users\LabUser\Downloads\Education vs per capita income.csv' DELIMITER ',' CSV HEADER;

DROP TABLE IF EXISTS per_capita_personal_income;
DROP TABLE IF EXISTS degree_numbers;
DROP TABLE IF EXISTS degree_percentages;
DROP TABLE IF EXISTS counties;

CREATE TABLE counties (
   county_fips INT PRIMARY KEY,
   state VARCHAR(2) NOT NULL,
   county VARCHAR(50) NOT NULL,
   UNIQUE(state, county)
);

INSERT INTO counties (county_fips, state, county) SELECT county_fips, state, county FROM education_vs_per_capita_income;


CREATE TABLE per_capita_personal_income (
   id SERIAL PRIMARY KEY,
   year SMALLINT NOT NULL,
   county_fips INT NOT NULL,
   income INT NOT NULL,
   UNIQUE(year, county_fips)
```

```
);

INSERT INTO per_capita_personal_income (year, county_fips, income)
SELECT 2019, county_fips, per_capita_personal_income_2019 FROM
education_vs_per_capita_income UNION
SELECT 2020, county_fips, per_capita_personal_income_2020 FROM
education_vs_per_capita_income UNION
SELECT 2021, county_fips, per_capita_personal_income_2021 FROM
education_vs_per_capita_income;


CREATE TABLE degree_numbers (
    id SERIAL PRIMARY KEY,
    year_range TEXT NOT NULL,
    degree VARCHAR(20) NOT NULL,
    county_fips INT NOT NULL,
    numbers INT NOT NULL,
    UNIQUE(year_range, degree, county_fips),
    CONSTRAINT degree_numbers_county_fips_fkey FOREIGN KEY (county_fips)
        REFERENCES public.counties (county_fips) MATCH SIMPLE
        ON UPDATE NO ACTION
        ON DELETE NO ACTION
        NOT VALID
);

INSERT INTO degree_numbers (year_range, degree, county_fips, numbers)
SELECT '2016-2020', 'Associate', county_fips, associate_degree_numbers_2016_2020 FROM
education_vs_per_capita_income UNION
SELECT '2016-2020', 'Bachelor', county_fips, bachelor_degree_numbers_2016_2020 FROM
education_vs_per_capita_income;


CREATE TABLE degree_percentages (
    id SERIAL PRIMARY KEY,
    year_range TEXT NOT NULL,
    degree VARCHAR(20) NOT NULL,
    county_fips INT NOT NULL,
    percentage NUMERIC(4, 2) NOT NULL,
    UNIQUE(year_range, degree, county_fips),
    CONSTRAINT degree_percentages_county_fips_fkey FOREIGN KEY (county_fips)
        REFERENCES public.counties (county_fips) MATCH SIMPLE
        ON UPDATE NO ACTION
        ON DELETE NO ACTION
        NOT VALID
```

```
);

INSERT INTO degree_percentages (year_range, degree, county_fips, percentage)
SELECT '2016-2020', 'Associate', county_fips, associate_degree_percentage_2016_2020 FROM
education_vs_per_capita_income UNION
SELECT '2015-2019', 'Bachelor', county_fips, bachelor_degree_percentage_2015_2019 FROM
education_vs_per_capita_income;


COPY counties TO 'C:\Users\LabUser\Downloads\counties.csv' WITH (FORMAT CSV, HEADER);
COPY degree_numbers TO 'C:\Users\LabUser\Downloads\degree_numbers.csv' WITH (FORMAT
CSV, HEADER);
COPY degree_percentages TO 'C:\Users\LabUser\Downloads\degree_percentages.csv' WITH
(FORMAT CSV, HEADER);
COPY per_capita_personal_income TO
'C:\Users\LabUser\Downloads\per_capita_personal_income.csv' WITH (FORMAT CSV,
HEADER);
COPY customer2 TO 'C:\Users\LabUser\Downloads\customer2.csv' WITH (FORMAT CSV,
HEADER);
COPY services TO 'C:\Users\LabUser\Downloads\services.csv' WITH (FORMAT CSV, HEADER);
COPY location_customer TO 'C:\Users\LabUser\Downloads\location_customer.csv' WITH
(FORMAT CSV, HEADER);


DROP TABLE IF EXISTS education_vs_per_capita_income;
---DROP TABLE IF EXISTS location_customer;
DROP TABLE IF EXISTS per_capita_personal_income;
```

## Part 2: Demonstration

### B. Panopto video

## Part 3: Report

### C. Written report

#### *C1. Dashboard alignment*

The dashboard created helps the user visualize not only the percentage of customers that have churned by state, but also the average income. It also allows the user to find if there is a relationship between income and education.

Another feature would be to verify if the states with the highest amount charged to customers are also the ones with highest income.

### *C2. Business intelligence tool*

We are using Tableau as the tool since it allows us to connect to our serve in PostgreSQL. In addition, we can join the tables within the schema and prepare our dataset for our visualizations.

Moreover, we can create the visualizations needed for the analysis based on the datasets we have previously prepared on PG Admin.

### *C3. Data Cleaning*

We are using PostgreSQL to change the structure of our Churn dataset since it is not ideal for our analysis. We will create a view and within this view we will create the tables needed for the analysis, some coming from the Churn dataset and others from the additional dataset Education vs per capita income.

1. Restructuring of the existing tables in Churn

We will create 3 different tables in Churn that will contemplate all the columns we currently have. For the table “customer”, we will take those columns that describe a customer, e.g., age, children, income, etc. The table “services” will store those services available to the customer or service-related fields, e.g., tenure, monthly charge, payment_type, etc. Lastly, the “location” table will contain those fields related, e.g., county, state, lat, etc.

2. Creating new tables for the additional dataset in Churn

For the additional dataset, we will create 4 tables. “Counties” will contain county_fips (a county identifier), county, and state. “Degree numbers” will consist of the number of bachelor’s and associate degrees by state while “degree percentage” will include the percentages of the degrees mentioned also by state. “Per_capita_personal_income” will have the average income by state.

While restructuring and creating the new tables in Churn, we will have to make sure we are using primary and foreign keys to ensure referential integrity, e.g, one single customer can have multiple services but only one location.

*C4. Dashboard creation*

1. Joining the tables

We will be creating 2 different tables from joining our tables already created. For the first table, we will join customer, location and services on the customer_id. These tables are originally coming from the Churn dataset.

From “Data Source”, we drag the “customer” table and click on it. Then we drag “location” and “services” to join them with “customer”.

Consequently, we drag the “counties” table to the panel and proceed to do the same as the “customer” table above, but using the “degree_percentages” table.

To create the custom queries, click on the table, then on “Data” and “Convert to custom SQL” (Tableau). We then remove those columns that are not relevant to the analysis.

2. Create the worksheets

The dashboards consist of worksheets placed together, so we will have to create a chart per worksheet “Churn by State”, “Income by State”, and “bachelor’s degree by State”.

To create a worksheet, click on the “New Worksheet” at the bottom of the panel.

a. Churn by State

Drag the “State” field to the Columns and the “Churn” field to the Rows.

To create the stacked bar chart, click on the “CNT(Churn)” downward arrow and select Create “Table Calculation”.

Under “Calculation Type” select “Percent of Total” and under “Compute Using” choose Table (down).

b. Monthly Charge by State

Drag the “State” field to the Rows and the “Monthly Charge” field to the Columns.

Change the default measure sum to average to see the average amount charged to customers in a specific state.

c. Income by State

To create a tree map with the average income by State, drag the field “Per capita income” to the “Marks” section over “Color”, “Size”, and “Label”.

Also, add the “State” field to the “Label” mark to show the state name on the tree map as well.

d. Bachelor’s degree by State

Drag and drop the “State” field to the Rows and the “Percentage” to the Columns.

Since the default measure is sum, click on the downward arrow to change the measure to average.

### *C5. Data analysis results*

The states that have highest churn rates are CT, DE, RI, PR, and HI. On the other hand, the states less likely to churn are MS, LA, NH, UT, and SD. This can help executive decision-making since we can identify which states we need to create strategies, so customers decide to stay at the company.

In the case of the monthly amount charged, the states that experience the highest amounts are SC, NM, AK, CT, and IN. We can see that CT is one of the states with most churned customers and it is also one of the ones with the highest amount charged. We can imply that for this state, the amount charged could be one of the reasons customers are leaving. Lowering the amount charged could be a strategy to retain customers in those states where this amount is not proportional to the average income.

The states with the highest customer’s average income are AR, NM, AZ, NE, and ID. NM is the only state that its average monthly charge is one of the highest as well as its average income. The customers in the rest of the states are probably being charged more compared to what their income is. This could be an issue since customers may want to leave when the cost is perceived as high.

The highest percentage of bachelor’s degrees correspond to MA, RI, CO, UT, and HI. We can see that in this case, these states are do not perceive the highest income. This might mean that most of our customers do not have a bachelor’s degree and because of this reason they are not part of the highest income within their state.

### *C6. Analysis limitations*

One of the limitations is that we are only studying geography as a variable related to churn, but there are more variables in the Churn dataset that can be studied as well and would make the analysis more robust.

Another limitation is that the additional dataset “Education vs per capita income” is two years old. The economy has changed in the past two years, and this might change income data which will not be representative of what is really happening.

**D. Sources**

Siegel, A. F. (2012). *Practical Business Statistics(Sixth Edition).* Academic Press.

*SQL Commands.* (n.d.). Retrieved from Scaler topics: https://www.scaler.com/topics/dbms/sql-commands/

*Tableau.* (n.d.). Retrieved from Connect to a Custom SQL Query: https://help.tableau.com/current/pro/desktop/en-us/customsql.htm

**E. Professional communication**