Western Governors University
School of Technology, College of IT
Master of Science, Data Analytics

D211 Advanced Data Acquisition

Performance Assessment

Tyson Biegler
Student ID: 012170282
10-20-2024

A1. Data sets

I used the provided churn data set and census data on the US population. I will include both data sets in the uploaded files and a cited link to the external CSV file in section D.

2. Installation steps: I've also included the .sql file that can be run directly in the pgAdmin.

1. Run the following SQL code to create the table for the external csv file.

```
CREATE TABLE us_pop_by_state (
    STATE INT PRIMARY KEY,
    NAME VARCHAR(255),
    Name_ID VARCHAR(2),
    ESTIMATESBASE2020 INT,
    POPESTIMATE2020 INT,
    POPESTIMATE2021 INT,
    POPESTIMATE2022 INT,
    POPESTIMATE2023 INT
);
```

2. Update the security settings of the CSV file or simply use the import/export feature to add the CSV data to the table. Updating the file's security settings might be too complicated for the average user. In this case, I'd recommend using the built-in import/export feature to add the CSV data.
    a. **right click the CSV file** > **Properties** > **Security** > **Edit** > **Add** > then type "**Everyone**" in the text box.
3. Run the 'COPY FROM' query to add the data to the table.

```
-- Adding CSV data to the table
COPY us_pop_by_state (STATE, NAME, Name_ID, ESTIMATESBASE2020, POPESTIMATE2020,
POPESTIMATE2021, POPESTIMATE2022, POPESTIMATE2023)
FROM 'C:\Users\LabUser\Downloads\NST-EST2023-POPCHG2020_2023_CLEANED.csv'
DELIMITER ','
CSV HEADER;
```

4. Lastly, double-click on **Dashboard_D211V2.twbx,** and the dashboard will be ready. There is no need to run the join queries. Tableau will handle the joins automatically.

3. Navigation instructions:

General instructions:

To interact with the dashboard, use the various charts and filters to modify what is displayed on the dashboard. Click on elements within the charts, such as states on the map or segments in the donut

chart, to filter the data accordingly. The filter panel at the top right is another way to filter the data just as if you were to click on individual elements.
Below are the specific instructions on how to interact with each of the charts found in the dashboard.

Churn by Region:

Hover over each donut to see the churn and retention rates for the individual regions. The regions were determined by referencing census regions and divisions map **(U.S. Census Bureau, n.d.).** If you click on one of the donut charts, the dashboard will update and only include information pertaining to the selected region. Selecting multiple regions can be done in two ways. First, click on a region and then hold down the CTRL key while selecting another region. The alternate way to do this is to use the 'Region' drop-down menu in the dashboard's top right. By checking the box next to a desired region, the dashboard will update to include the information that pertains to the selected regions.

Churn by Age:

Each bar in the churn by age chart represents 5-year increments. If you hover your mouse pointer over a bar, you will see the number of customers in that specific age group. Likewise, clicking on the bar, or 'CTRL + clicking' on multiple bars, will filter the dashboard's data to only include what is relevant to the selected age bars.

Churn by Gender:

By hovering over a bar in the churn by gender chart, you can see the churn rate, the number of customers, and the percentage of the total customers represented by each gender. Just like in the other charts, clicking on a gender or CTRL-clicking on multiple genders will filter the dashboard to only include information that pertains to the selected genders.

Churn by Income:

Each bar in the churn by income chart represents $20,000 increments. By hovering over any of the bars in the chart, the churn rate, number of customers, and the percentage of customers in the income group all become visible. Once again, by clicking on a bar or CTRL-clicking on multiple bars, the user can filter the data in the dashboard to include only the data relevant to the selected income groups.

Churn by State:

The churn-by-state chart is where most of the information is displayed. By hovering over a state, the user can see various statistics. These include the average tenure, churn rate, market penetration, population, and population change in the past year, as well as the total number of current customers alongside the number of churned customers. These statistics are accompanied by a description that helps explain what the statistics mean if this dashboard user is not technically savvy.

Like the other charts, clicking on or CTRL-clicking multiple states will filter the dashboard data to include only the relevant information. However, in the top right of the dashboard, the user can also

select individual or multiple states and regions rather than clicking on each state or region on the map chart.

4. SQL code:

```
-- Creating the table for the external CSV file
CREATE TABLE us_pop_by_state (
    STATE INT PRIMARY KEY,
    NAME VARCHAR(255),
    Name_ID VARCHAR(2),
    ESTIMATESBASE2020 INT,
    POPESTIMATE2020 INT,
    POPESTIMATE2021 INT,
    POPESTIMATE2022 INT,
    POPESTIMATE2023 INT
);

-- Adding csv data to table
COPY us_pop_by_state (STATE, NAME, Name_ID, ESTIMATESBASE2020, POPESTIMATE2020, POPESTI-
MATE2021, POPESTIMATE2022, POPESTIMATE2023)
FROM 'C:\Users\LabUser\Downloads\NST-EST2023-POPCHG2020_2023_CLEANED.csv'
DELIMITER ','
CSV HEADER;
```

First, I created the table for the external CSV file in the PostgreSQL database. Before adding the csv file data, I had to change the permissions on the csv file so It could be accessed. This may be too complicated for the average user, so I would encourage the user to use the import/export feature in PostgreSQL. However, this is the method I used. Then, I used the 'COPY FROM' query to add the CSV data to the newly created table.

Secondly, I joined the customer and location tables on the 'location_id' fields. Customer.location_id and location.location_id share an identification number that I can use to match with a state abbreviation in the next step. Then I joined the location.State and 'us_pop_by_state.name_id' together to connect the customer table location id and the external census data's state population data.

This is important because the customer table gives me data about the customers, such as gender, age, income, churn status, etc., and the external census data gives me population data. Combining the churn data set and the census data file lets me gather insights about population changes or market penetration.

Below are the previously mentioned join statements Tableau generated in the background on my behalf. **(Sewell, 2024, slide 22).**

```sql
-- this query joins the customer table to the location table
SELECT "customer"."age" AS "age",
  "customer"."bandwidth_gp_year" AS "bandwidth_gp_year",
  "customer"."children" AS "children",
  CAST("customer"."churn" AS TEXT) AS "churn",
  CAST("location1"."city" AS TEXT) AS "city (location1)",
  "customer"."contacts" AS "contacts",
  "customer"."contract_id" AS "contract_id",
  CAST("location1"."county" AS TEXT) AS "county (location1)",
  CAST("customer"."customer_id" AS TEXT) AS "customer_id",
  "customer"."email" AS "email",
  CAST("customer"."gender" AS TEXT) AS "gender",
  "customer"."income" AS "income",
  "customer"."job_id" AS "job_id",
  "customer"."lat" AS "lat",
  "customer"."lng" AS "lng",
  "location1"."location_id" AS "location_id (location1)",
  "customer"."location_id" AS "location_id",
  CAST("customer"."marital" AS TEXT) AS "marital",
  "customer"."monthly_charge" AS "monthly_charge",
  "customer"."outage_sec_week" AS "outage_sec_week",
  "customer"."payment_id" AS "payment_id",
  "customer"."population" AS "population",
  CAST("customer"."port_modem" AS TEXT) AS "port_modem",
  CAST("location1"."state" AS TEXT) AS "state (location1)",
  CAST("customer"."tablet" AS TEXT) AS "tablet",
  CAST("customer"."techie" AS TEXT) AS "techie",
  "customer"."tenure" AS "tenure",
  "customer"."yearly_equip_faiure" AS "yearly_equip_faiure",
  "location1"."zip" AS "zip (location1)"
FROM "public"."customer" "customer"
```

```
-- this query joins the location table to the aditional csv census data table named
"us_pop_by_state"
SELECT CAST("location"."city" AS TEXT) AS "city",
  CAST("location"."county" AS TEXT) AS "county",
  "us_pop_by_state1"."estimatesbase2020" AS "estimatesbase2020 (us_pop_by_state1)",
  "location"."location_id" AS "location_id (location)",
  "us_pop_by_state1"."name" AS "name (us_pop_by_state1)",
  "us_pop_by_state1"."name_id" AS "name_id (us_pop_by_state1)",
  "us_pop_by_state1"."popestimate2020" AS "popestimate2020 (us_pop_by_state1)",
  "us_pop_by_state1"."popestimate2021" AS "popestimate2021 (us_pop_by_state1)",
  "us_pop_by_state1"."popestimate2022" AS "popestimate2022 (us_pop_by_state1)",
  "us_pop_by_state1"."popestimate2023" AS "popestimate2023 (us_pop_by_state1)",
  "us_pop_by_state1"."state" AS "state (us_pop_by_state1)",
  CAST("location"."state" AS TEXT) AS "state",
  "location"."zip" AS "zip"
FROM "public"."location" "location"
  INNER JOIN "public"."us_pop_by_state" "us_pop_by_state1" ON (CAST("location"."state" AS TEXT)
= "us_pop_by_state1"."name_id")
```

B: Panopto video:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b53e9162-b6db-4e71-b8f3-b1cf012f5f81

C1. Explanation of functionality:

The performance assessment scenario section suggests that telecom companies can experience churn rates of up to 25%. (WGU, n.d). Amy Gallo of Harvard Business Review: *"...acquiring a new customer is anywhere from five to 25 times more expensive than retaining an existing one" (**Gallo, 2014**).* So, identifying the customers and the factors that lead to churn is essential for a business to continue operating. This dashboard will allow the executives to filter demographic information based on churn, targeting specific individuals that match a profile or target a particular area that may be contributing to churn.

2. Justification of the tools:

This dashboard was created using a combination of PostgreSQL for access and manipulation of the churn dataset and Tableau for data visualization.

PostgreSQL, like any other database management systems I've used, can easily handle large datasets. PostgreSQL was the most intuitive and user-friendly, making it an easy choice.

Tableau was selected primarily because I am building on the tableau dashboard that I created in D210. By connecting Tableau directly to the PostgreSQL database, I was able to get real-time updates to the data if I were to make changes.

3. Explanation of the steps taken to prepare the data:

The external CSV file I downloaded from the U.S. Census Bureau contains population data over several years. It also differentiates the population by state. Because of this, the data was ready to use once it was loaded into the table.

4. Explanation of the steps taken to create the dashboard:

This dashboard consists of 5 charts and 3 KPIs. The charts were created using various measures, dimensions, and calculated fields.

Calculated fields:

Market penetration represents the relationship between the total population of a state retrieved from the external CSV file and the count of customers. Simply put, the market penetration number indicates the percentage of the population that are customers. This calculation is helpful because if the market penetration and tenure is low, this could give rise to a referral program to recruit new customers.

KPIs:

1. The churn rate was calculated by first getting the sum of customers whose churn status is "Yes," meaning that they have churned in the last month, and then dividing that number by the count of 'customer id.' This results in a churn rate of 26.50% for the dataset.
    a. **SUM(IF [Churn] = "Yes" THEN 1 ELSE 0 END) / COUNT([Customer Id])**
2. Population change was calculated by getting the sum of the 2023 population minus the 2022 population and then dividing that number by the 2022 population number.
    a. **(SUM([Popestimate2023]) - SUM([Popestimate2022])) / SUM([Popestimate2022]) * 100**
3. Market penetration was calculated by calculating the number of customers who had not churned and then dividing that number by the population. If a user selects a state or region, the calculation would be divided by the population of that area. The dataset has a market penetration of 0.0022%, meaning that 0.0022% of the population are customers.

Charts:

1. Bar chart of average churn based on income (bins). I created income bins that increment by $20,000 and then put the count of customers on the y-axis to show the total customers in each income bin.
2. Bar chart of the total customers in each age group. Instead of creating an age bin, I created an age group. The reason for the change is because bins suggested that there were customers under 18, whereas bins do not. I also overlaid a bar chart of the churn rate in each age group using dual axis.
3. Bar chart displaying the churn rate differentiated by gender.
4. Map chart displaying each state's churn rate, market penetration, total customers, average tenure, and the population. This chart will allow an executive to see several statistics based on states or regions.
5. A donut chart that shows the churn and retention rates by region. I used a dual axis chart here as well to get the donut effect and then used the measure values for churn rate and retention rate.
    a. Churn Rate: **SUM(IF [Churn] = "Yes" THEN 1 ELSE 0 END) / COUNT([Customer Id])**
        i. The formula checks if the churn status is 'Yes' and then adds 1. After all the churn values are added up then it will divide that number by the count of customers.
    b. Retention Rate: **SUM(IF [Churn] = "No" THEN 1 ELSE 0 END) / COUNT([Customer Id])**

i. Retention rate is calculated the same way churn rate is calculated except for it checks if the churn status is “No”, which would indicate that the customer has been retained.

5. Discuss the results of your data analysis and how it supported the purpose and function of your dashboard.

One benefit of this dashboard is that it can create customer profiles. For example, by selecting various items in the charts, I found that people whose gender is “prefer not to answer” and are between 18 and 22 years old and living in the Midwest have a 50% churn rate. Similarly, males between 18 and 22 in the Midwest who made less than $20,000 per year have a 45.83% churn rate. This is alarmingly high compared to the 26.5% national average.

Using the churn-by-state map, I found that some areas have an average tenure of around 3 years. Take Idaho, for example. In Idaho, the average tenure of a customer is 3.186 years. That seems to be on the higher end as far as tenure goes. However, Idaho has gained 1.33% or 25730 new residents over the past year, while the company lost approximately 24.69% or 20 of its 81 customers In the state. So the available customer pool is increasing, but the company is losing customers. For some reason that cant be determined by this dashboard, the new residents are not being attracted to the company's services. Combining this knowledge with the average tenure of over 3 years, the company might benefit from offering a referral bonus to current customers. By offering current customers a reward for recruiting new residents would increase market penetration.

6. Explanation of limitations:

One major limitation, and perhaps the most detrimental regarding analysis, is the lack of time data, specifically yearly or quarterly dates. Because of this, the dashboard can not accurately display revenue changes over time or the impact of specific services on revenue.

The database is more of a list of facts about the customers rather than financial data. However, the dictionary describes a relationship between a company's revenue and customer churn. Retaining customers safeguards against revenue loss far better than earning new customers. However, without knowing the data date ranges and creating new tables for calculations across time, any monthly revenue calculations would have to make arbitrary assumptions about the year or quarter in which the data was gathered.

Adding in the US census population data allowed me to gather population changes over time, and assuming the churn data is current, I was able to put this data together to calculate market penetration and make recommendations on incentive programs. However, as I mentioned in the previous paragraph, the time frame of the churn data set is unclear, so these previously mentioned calculations and recommendations are based on the assumption that the churn data is current.

D. Web sources

U.S. Census Bureau. (n.d.). NST-EST2023-POPCHG2020_2023 [CSV file]. https://www2.census.gov/programs-surveys/popest/datasets/2020-2023/state/totals/NST-EST2023-POPCHG2020_2023.csv

E. Additional Sources

1. **Gallo, A.** (2014, October 29). The Value of Keeping the Right Customers. Harvard Business Review. Retrieved August 10, 2024, from https://hbr.org/2014/10/the-value-of-keeping-the-right-customers
2. **U.S. Census Bureau.** (n.d.). Census regions and divisions of the United States. U.S. Department of Commerce. https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf
3. **Western Governors University.** (n.d.). *Advanced Data Acquisition - D211*. WGU. Retrieved August 8, 2024, from https://tasks.wgu.edu/student/012170282/course/34320018/task/4303/overview
4. **Sewell, W.** (2024, August 18). *SQL Sunday Postgres D211 PowerPoints* [PowerPoint slides]. Western Governors University. https://westerngovernorsuniversity-my.sharepoint.com/:p:/g/personal/william_sewell_wgu_edu/EQtfWHAWXOxOtRGnIjWRVhMB78w3O31MSh5exT41fofaIQ?e=etWBhA

F. Panopto video:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=84c3cdef-e061-4410-9b31-b20f011ec852