D211 Advanced Data Acquisition Performance Task 1

Data Analysis

Sean Simmons

WDU Data Analytics

MSDA D211

February 2023

Part 1: Data Dashboards

A. Provide a copy of your dashboards that support executive decision-making.

The dashboard has been included as "Dashboard.twb" and an image is included in the instructions. I have included a pdf of the dashboard as a back-up, appropriately named so.

1. Provide *both* data sets that serve as the data source for the dashboards.

The original data set provided by WGU is the attached "churn_clean.csv" file. In this assignment, the data set was already loaded into the churn database on the virtual lab machine so I did not have to do anything with churn_clean, although I will still include it. The external dataset is attached as "organised_gen"

2. Provide step-by-step instructions to guide users through the dashboard installation.

Instructions have been provided in the file "Instructions PA: Advanced Data Acquisition (SLM1)."

3. Provide clear instructions to help users navigate the dashboards.

Instructions have been provided in the file "Instructions PA: Advanced Data Acquisition (SLM1)."

4. Provide a copy of *all* SQL code and other code supporting the dashboards.

A copy of the queries referred to and used in the installation can be found in the file "queries.txt." and the full sql file has been included as "creating_table"

The join and connections Tableau made have been downloaded and can be performed in your SQL query after loading the data into the new electricity 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("location"."city" AS TEXT) AS "city",
  "customer"."contacts" AS "contacts",
  "customer"."contract_id" AS "contract_id",
  CAST("location"."county" AS TEXT) AS "county",
  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",
  "location"."location_id" AS "location_id (location)",
  "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("location"."state" AS TEXT) AS "state",
  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",
  "location"."zip" AS "zip"
FROM "public"."customer" "customer"
  LEFT JOIN "public"."location" "location" ON ("customer"."location_id" =
"location"."location_id")
```

Part 2: Demonstration

B. Provide a link to a Panopto multimedia presentation in which you present the dashboards to an audience of data analytics peers. You should do *all* of the following in your presentation:

Key points and takeaways from the actual video that answers these points are also summarized/included below.

1. Describe the technical environment used to create the dashboards.

Pgadmin Postgresql 4 version 5.2, Tableau desktop version 2021.4.

2. Demonstrate the functionality of the dashboards.

Shown in video and can't be summarized.

3. Explain the SQL scripts used to support the creation of the dashboards.

Provided as a text file, sql file, explained in the instructions guide, and summarized in the section of the video where I show the actual database and query.

4. Explain how the data streams were prepared to support the analysis.

The churn set and tables are already prepared and cleaned in the database provided. For organised_gen the data is already cleaned and just needs to be put into the churn database.

5. Describe how data were aligned with other data points.

States from organised_gen to the state abbreviations in churn were used as keys to align the data. The customer and location tables were joined using the location ids. The new data was electricity generated by state, adding to the geographic identification information of the customers and providing another way to identify trends.

6. Demonstrate how the databases were created.

Using the postgresql churn database provided, and then upload the organized gen, download the queries that create a table for it. We will connect the tables using the state names in tableau.

7. Explain how referential integrity was enforced in the database.

I used the location id initially, and later the state abbreviations from both data sets as a foreign and primary key to enforce the referential integrity of the database through tableau (Ian 2016).

The link to my panopto video presentation:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=fd651e7f-c888-46b0-80af-afad00205e4e

Part 3: Report

C. Write a report to outline the data exploration, use of advanced SQL operations, and the analysis of the data. Do the following as part of your report:

1. Explain how the purpose and function of your dashboard aligns with the needs outlined in the data dictionary associated with your chosen data set.

The purpose of this dashboard aligns with the needs of the organization because the telecommunications company is searching for actionable insight from their customer dataset and

the external dataset. My chosen dataset, orgaised_gen is electricity generation information throughout all 50 states we can use to gain further insights into the churn data. The purpose of my dashboard is to find actionable insight into a major performance metric, churning, through other variables and to find actionable insights that link the organised_gen file to our churn dataset. By actionable insight, I mean I am looking for trends or connections through bivariate and univariate analysis with the representations in the dashboard that I can use to recommend future actions to the organization.

My dashboard analyzes the relationship between electricity generated and monthly charge, perhaps informing the organizations key metric of customer turnover with their main profit force, monthly charge. The other representation shows the relationship between the 50 United States and electricity generation with outage seconds per week, a key performance metric that can inform the organization of which states need more infrastructure support. The function of showing these data trends and using electricity generation to inform the churn dataset allows for actionable insights to be made that align with the needs in the data dictionary; retaining customers and predicting which customers are at a high rate of churn by understanding what areas need more support and resources.

2. Justify the selection of the business intelligence tool you used.

The business intelligence tool I used was Tableau Desktop. This business analysis tool is one of the leading tools to assess, visualize, and analyze data. Personally, I am familiar with Tableau and enjoy the user interface. As a tool, it allowed me to integrate data from postgres and then manipulate the variables to easily create all of my desired visualizations. Tableau was also able to easily handle the joining of the data that was easily brought in from an SQL database.

The steps to do so and the range of abilities tableau offers made it the business intelligence tool I chose to use.

3. Explain the steps used to clean and prepare the data for the analysis.

Please see the provided instructions file for detailed and comprehensive instructions.

1. The first step in cleaning and preparing the data was downloading the organised_gen file onto the virtual lab machine provided to me from WGU. This virtual lab machine already included a database filled with the churn tables and connections. The data file provided, churn_clean, needed no cleaning or preparation as it was already cleaned.
2. Import organised_gen into the sql coding environment
3. Install the sql database into tableau
4. Connect the customer and location table through location_id
5. Use tableau to create visualizations through the state name connections

4. Summarize the steps used to create the dashboards.

The full steps used to create the dashboard can be found in the file "Instructions PA: Advanced Data Acquisition (SLM1)." However, I will copy the main steps for a summary here. To begin creating the dashboard I used, you must follow the preparation steps above and have a fully connected database. Once there, upload the database to tableau. Connect the customer and location tables through the location id's. Next, create a relationship using the state abbreviations from electricity and customer. Finally, we can begin going to sheet 1 and dragging variables into the UI to create the dashboard.

1. Exclude null values for states in all processes and filters throughout the dashboard. We have now set our data source and can create the four visualizations. Start by navigating to sheet 1.
2. 1st representation: Create a heatmap of the United States with energy per state. Drag states into the column section. Drag the variables generation (megawatthours) and bandwidth into the rows section, this will take a few minutes for tableau to execute the query. Click “Show me” on the top right of the screen and find the option for the geographic visualization. Filter out the unknown. Choose the gray color from the color-blind accessible color palette. We will add the ranking calculation and tool tip described later in this instruction guide.
3. 2nd representation. Exclude US-total and nulls. Create sheet 2. Drag the state variable into the column, drag generation (megawatthours) and outage seconds per week into the columns. Go to show me and choose the shape and size graph. Select a color from the color-blind accessible color palette.
4. For color blind accessibility, bring up the color-blind color palette and choose colors from the list with high contrast, I chose the gray option.
5. Drag the corresponding legends from the list of variables to the tooltip option to create the legends.
6. Take the target columns, average monthly charge, generation (megawatthours), and outage seconds per week for our various representations and drag them to the tool tip section to create legends.
7. Click create a dashboard and then double click all of the sheets we have created to automatically bring them into the dashboard.

8. Finally, run the optimizer to resolve any issues, save and share the dashboard.

5. Discuss the results of your data analysis and how it supports executive decision-making.

Our dashboard informs us of monthly charge and outage seconds per week trends with the 50 states and electricity generated in them. The electricity generation and average monthly charge can be used to inform customer retention methods and demographics. By further analyzing monthly charges with another data trend, we can begin to understand our customer demographics and identify what support we can give them. I recommend the organization to identify their states of interest and put more support into lowering the average monthly charge. The representation using the size map and variables of both datasets can be used to gather data to inform the organization's retention efforts, seeing how electricity generation and outage seconds per week can be affected to increase retention by improving customer support. I recommend the organization to put resources and support into decreasing the outage seconds per week for the customers in the states with the highest averages and to continue to analyze how this trend is associated with electricity generated.

6. Discuss the limitation(s) of your data analysis.

The limitations of my data analysis are that electricity generated by the state does not have a breakdown of city, county, or area like the churn database so we are limited in how far we can drill down the data. Also, there is no direct link between the specific churn customers and the electricity generated. Thirdly, electricity generated does not dive into electricity consumed by specific demographics of the state (consumer, industrial, corporate, etc.) therefore limiting the

connection we can make with churn. Finally, due to the nature of the data, we can draw attention to trends and perform basic analysis, but direct causality cannot be reached with our dashboard.

D. Record the web sources used to acquire data or segments of third-party code used to support the application. Ensure the web sources cited are reliable.

References

How to JOIN Tables in SQL. (2021, March 3). LearnSQL.com. https://learnsql.com/blog/how-to-join-tables-sql/#:~:text=To%20join%20two%20tables%20in%20SQL%2C%20you%20need

Mester, T. (2022, July 10). *How to Import Data into SQL Tables*. Data36. https://data36.com/how-to-import-data-into-sql-tables/#:~:text=How%20to%20Import%20Data%20into%20SQL%20Tables%201

E. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

References

Ian. (2016, May 28). *What is referential integrity? | database.guide*. Database.guide. https://database.guide/what-is-referential-integrity/

F. Demonstrate professional communication in the content and presentation of your submission.

This aspect of the rubric is evaluated through the entirety of this report and I hope professionalism has shown continuously.