Ryan James Calabio
D211: Advanced Data Acquisition
Prof. Sewell

A1:

The external dataset that was used in this assignment was included in the submission. The other churn dataset is stored with pgAdmin 4 on the labs on demand.

A2:

The datasets are included alongside this document in the submission. It uses the churn dataset, and a dataset taken from Kaggle on census information for adults in the United States.

A3:

The dashboard can also be accessed by downloading and installing Tableau, then opening the attached .twbx file.

Once the file is opened, there is a presentation mode that can be entered by pressing the button on the bottom right of Tableau.

A4:

The SQL that was used in this project is:

```
CREATE TABLE census (
        "Age" int NOT NULL,
        "Workclass" varchar(255) NOT NULL,
        "Final Weight" int NOT NULL,
        "Education" varchar(255) NOT NULL,
        "EducationNum" int NOT NULL,
        "Marital Status" varchar(255) NOT NULL,
        "Occupation" varchar(255) NOT NULL,
        "Relationship" varchar(255) NOT NULL,
        "Race" varchar(255) NOT NULL,
        "Gender" varchar(255) NOT NULL,
        "Capital Gain" int NOT NULL,
        "capital loss" int NOT NULL,
        "Hours per Week" int NOT NULL,
        "Native Country" varchar(255) NOT NULL,
        "Income" varchar(255) NOT NULL
);

DROP TABLE census;

SELECT * FROM census;

SELECT *
FROM census
WHERE
        "Age" IS NULL OR
        "Workclass" IS NULL OR
        "Final Weight" IS NULL OR
        "Education" IS NULL OR
        "EducationNum" IS NULL OR
        "Marital Status" IS NULL OR
        "Occupation" IS NULL OR
        "Relationship" IS NULL OR
        "Race" IS NULL OR
        "Gender" IS NULL OR
        "Capital Gain" IS NULL OR
        "capital loss" IS NULL OR
        "Hours per Week" IS NULL OR
        "Native Country" IS NULL OR
        "Income" IS NULL;
```

B:

Panopto Link is included in the submission.

C1:

1. Explain how the purpose and function of your dashboard aligns with the needs of the stakeholders for your chosen dataset.

For this dashboard, I wanted to focus on the difference between the average American by age with the average customer of the company. This can give insights to the business like if there is a certain age segment that is missing from their customer base, or one that the product is especially popular for a certain age base. The included visuals are:

- Distribution of Census Gender vs Customer Gender by Age
- Number of People Making Over and Under 50k by Age
- Average Hours Worked per Week Working by Age

C2:

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

In the creation of this dashboard, there were two main tools that were used. First, pgAdmin 4 was used to set up the churn database, import the external dataset, and then connect to the second business intelligence tool that was used. The second tool was Tableau. Tableau imported the database using PostgreSQL and then visualized the data in a way that was useful for the stakeholder.

C3:

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

In order to import the data into pgAdmin 4, I needed to delete the header from the CSV, as it would automatically import the data in based upon the set types set in the SQL table. If I kept the headers in, the data it pulled would not match the types that were initially set in the table creation. I started this by using excel to delete the first row of the .csv from the file and resaving it. Then importing it into pgAdmin 4.

I also checked for nulls using this query:

```
SELECT *
FROM census
WHERE
        "Age" IS NULL OR
        "Workclass" IS NULL OR
        "Final Weight" IS NULL OR
        "Education" IS NULL OR
        "EducationNum" IS NULL OR
        "Marital Status" IS NULL OR
        "Occupation" IS NULL OR
        "Relationship" IS NULL OR
        "Race" IS NULL OR
        "Gender" IS NULL OR
```

```
"Capital Gain" IS NULL OR
"capital loss" IS NULL OR
"Hours per Week" IS NULL OR
"Native Country" IS NULL OR
"Income" IS NULL;
```

C4:

4. Summarize the steps used to create the dashboard.

The external data was added to the database in pgAdmin 4 using the following steps:

- Open pgAdmin 4
- Open Databases, churn, rightclick on Tables, and press Query Tool
- Use the following SQL:

```
CREATE TABLE census (
        "Age" int NOT NULL,
        "Workclass" varchar(255) NOT NULL,
        "Final Weight" int NOT NULL,
        "Education" varchar(255) NOT NULL,
        "EducationNum" int NOT NULL,
        "Marital Status" varchar(255) NOT NULL,
        "Occupation" varchar(255) NOT NULL,
        "Relationship" varchar(255) NOT NULL,
        "Race" varchar(255) NOT NULL,
        "Gender" varchar(255) NOT NULL,
        "Capital Gain" int NOT NULL,
        "capital loss" int NOT NULL,
        "Hours per Week" int NOT NULL,
        "Native Country" varchar(255) NOT NULL,
        "Income" varchar(255) NOT NULL
);
```

- Refresh the Tables in the right click menu, press import/export on the right click menu from census, and then select the included CSV.

Once the data was added to Tableau, I started by designing each of the visual in an individual sheet. I started with “Average Hours Worked per Week Working by Age“ by using Age and Hours Worked per Week from the census data. Next, I made “Number of People Making Over and Under 50k by Age”. This was made using Age and Census Income. Then finally, I created the visual with two graphs: “Average Hours Worked per Week Working by Age”. This visual used the Age, Count of Census Gender, and Count of Customer Gender. I created a dashboard, created a title, and then brought in all of the aforementioned visuals.

C5:

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

The results of my analysis were that according to census data, that the age range with the most people making greater than 50,000 dollars a year are around age 46, and as they get older and younger from there, it starts to go down. It looks like a normal distribution. It may be useful for the business to target customers that are in that age range as they may have more disposable income then age demographics with higher levels of those who make under 50,000. This matches up with the working hours per week by age graph, which has around that age range as the ones working the most hours. If the customer is working that many hours, they may have a greater need for telecommunication products to keep in contact with family and friends. Lastly, the graph for Distribution of Census Gender vs Customer Gender by Age can give us a few different insights. First of all, it can show us the distribution of number of customers compared to the census age. The census shows that there are many more young people than there are older. However, the customer base is very uniform. This means that maybe the business needs to focus more on attaining younger customers, as that group in general is much larger than the older one.

C6:

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

The limitation of my data analysis was that because I could only connect the two datasets by either age or gender, I had to keep it to analysis that was related to these two columns of data. If not for this limitation, the visuals and analysis that could be done would be much more extensive. Another limitation is that the census data was split into test and training datasets, and only one of these datasets were used. Therefore, the census data may be lacking a small portion of data that could affect insights from the analysis.

D: Sources

https://app.datacamp.com/learn/custom-tracks/custom-d211-advanced-data-acquisition

https://www.kaggle.com/datasets/tawfikelmetwally/census-income-dataset?resource=download