Data Analysis

Matthew E. Heino

Advanced Data Acquisition

Introduction

The purpose of this paper is to discuss how one can go about creating a dashboard using a PostgreSQL database along with business intelligence software. The software that will be used in this assessment will be Tableau. The reason why this software was chosen will be discussed in a subsequent section of this document.

Background

This assessment seeks to combine the power of PostgreSQL and the associated database. It will use two separate data sources. One was provided by the university. It was included in the initial installation of the database. The file will have data about the patients. This data includes personal information and information about whether this patient has experienced certain types of conditions.

The database is composed of a few tables. These tables in a previous assignment were contained in one CSV file. For the assignment, the information was broken up into a series of separate tables. The tables are the following.

Table Name	Description
admission	Type of admission of the patient (emergency, elective, observation).
complication	The complication risk of the patient (unknown, high, medium, low)
job	The job of the patient.
location	The place of residence or location of the patient.
patient	Information about the patient e.g. age,

	income, etc.
all_medical_data	All the medical data was joined from the WGU dataset and the standalone dataset.
servicesaddon	The medical services that the patient underwent while the patient was in the hospital.
survey_responses_addon	Responses to the survey were presented to the patient.

Table 1. Tables of the medical_data Database

Note: Not all the tables will be used in the creation of the dashboards that are required for this assessment. More information about what tables and the information will be discussed in the following sections of this document.

Part 1 The Data Dashboard

In this section, there will be a discussion of the datasets that were used. There will be a step-by-step direction on how to install the dashboards that will be used in the presentation that will occur in a subsequent section. Instructions will also be provided on how to navigate the dashboards. There will be a section that will include all the code that was used to create the database as well as any ancillary code used to clean the data.

A1. The Datasets

The data that was used in this assessment are the following. The first dataset was provided by the university and was included in the VM that stored the database for the

assessment. The dataset is very similar to the dataset that was used in previous assessments that used the medical dataset. The only difference is that some additional information was added and the information was broken up over a few tables. These tables were discussed in the previous section of the paper. Please see the **Background** section for more information about the tables that are part of the dataset.

The second dataset was a cleaned data set that was composed of a series of files that were downloaded from the following website (*US Census Demographic Data*, n.d.):

- https://www.kaggle.com/datasets/muonneutrino/us-census-demographic-data?resource=download&select=acs2017_county_data.csv
- Data Set: **acs2017_county_data.csv**

There was one data file used from this site. The data file **acs2017_county_data.csv** provided the needed demographics that were needed to aid in the creation of the dashboard. The Kaggle CSV file was cleaned using PostgreSQL in pgadmin 4. More about how this file was cleaned and prepared will be discussed in a subsequent section of the paper

This file was read into a table called **us_census_data**. This table will hold all the columns that compose the ancillary CSV file. There were approximately 37 columns in the set along with 3220 rows of information. Columns that will be of interest to this assessment are the following:

- state
- county
- men – total male population by county
- women – total female population by county
- income – average income
- TotalPop – total population

- Unemployment – the unemployment rate as a percentage

Note: I am currently unaware of a way to selectively read only certain columns from the CSV file. I only know that is possible using Python and the pandas library. Most other manipulations of the data will occur in either the database or through the desktop interface of Tableau software.

A2. Dashboard Installation.

Installing the dashboard will be relatively straightforward. You will need the appropriate software that is available on the Tableau website. Based on the requirements of the assessment you can download and run the assessment packaged workbook. This file ends in the packaged workbook and will contain the database and all the required tables along with the data that is required to make the dashboards function. The files for the assessment are the following:

- **all_medical_data (medical_data).hyper** - This is the extract that will be used along with the packaged workbook. This will be needed to run the workbook without a direct connection to the database on Postgres.
- **Heino D211 Task 1 Final.twbx** – This is the packaged workbook. It will be "connected" to the extracted dataset that was discussed earlier.

These are the only files that will be required to use the database. Please note that some of the range functions on some of the dashboards may not be active when using this method to access the data and run the story/dashboards that are included with this submission. This will be further explained in the next section.

A3. Dashboard Navigation

Data Analysis

This assessment will be composed of a few dashboards. This is to keep the display clean and easy to see. The dashboards were included in a story. The story is **WGU Patient Demographics and Census Data.** This was to allow for easier viewing of the data and to help differentiate between the different data sources that were used in the presentation. These dashboards are best viewed in **Full Screen** mode of the virtual machine.

To navigate the dashboard it is best accessed using the story (**WGU Patient Demographics and Census Data**) that was mentioned above. There is one information dashboard and three informational dashboards that provide information about the patients and the citizens of the United States. The dashboards are:

- **Introduction** – just for the assessment can be skipped
- **Census Demographics** – Presents US Census demographics that are compatible with the data provided by the university. **Note**: some of the data for the counties will not appear.
- **WGU Patient Demographics** – Presents the patient data as found in the university-provided dataset.
- **Patient Income Map** – This shows a density map of the income for the patients in the university dataset.

The information about the citizens comes from the US Census. Please note that if the screen goes blank there may be no information for that particular combination of search features. Please press the green reset button located on the right side bottom of the dashboard. This should reset the filters.

Pressing the reset pin located on the left-hand side of the map will center the map. **Note:** This can be used at any time to recenter the map on any of the dashboards that are included within this story.

Census Demographics Dashboard Navigation:

The navigation of this dashboard is accomplished by using the various filters located on the right-hand side of the dashboard. You can filter by various components these filters are the following:

- State
- Gender
- Readmis – Readmission (Yes or No)
- Male population (a range value) – based on county
- Female population (a range value) -based on county

** Please note that the range sliders may or may not work if you are using the extracted dataset that will accompany this file. If there is a blank screen as a result of a selection of the filters you have two options:

1. Select the green reset button located at the bottom right of the dashboard and this will reset the whole view and reset it to the full United States view of the census data.
2. You can select a different combination of filters until the view returns. This is useful if you would like to stay within the same state, but want to try other filter combinations.

WGU Patient Demographics Dashboard Navigation:

This dashboard will give you additional information about the individual patients. You will be able to see some individual information about the patient. The gender of the patient is a color-coded point on the map.

To zoom in on the map click on a location that you are interested in, through the filters on the right side of the dashboard and the map should zoom in on that point. Hoovering over the colored point will give you information about the patient as well as a calculated field that

displays the average income for the patients that are in the university's data. Please note that not all points will have all the data. Some data points will have some data missing. This is because it was not available for the patient and it was not corrected. This will not affect the overall display of the patient data points on the map. Filter options on this page are the following:

- State
- City
- Gender
- Arthritis
- Asthma
- Diabetes
- High Blood Pressure
- Overweight
- Stroke

If you encounter a "blank" page please follow the steps that were given in the previous dashboard (Census Demographics Dashboard Navigation). You will be able to look at a combination of these filters to see where the patients are located within the country.

Patient Income Map:

The Patient Income Map will show where the income of the patient as a point on the map relative to the income of the patient. To zoom in on the map click on a location that you are interested in, through the filters on the right side of the dashboard and the map should zoom in on that point. The map will not always zoom or pan to the location of choice. You may need to click in the general vicinity of the state or location you are interested in. This is after choosing the state from the State filter that is located on the right side of the page.

There are two filtering options for this page. The filtering options for this page are:

- State
- Age categories – to view the income of individual patients by age categories.

If you encounter a "blank" page please follow the steps that were given in the previous dashboard (Census Demographics Dashboard Navigation). You will be able to look at a combination of these filters to see where the patients are located within the country.

A4. Code for the Dashboard.

The code for supporting the dashboard is shown below. The code that was used to prepare the datasets from Kaggle and the included tables is shown below.

The first SQL query that was created was the one to extract all the information from the university supplied is shown in the following screenshot[1].

The code that was used to alter the database will be discussed in the following paragraphs. The first code segment is the creation of the table that will hold the data for the data from the cleaned dataset that was discussed earlier in this part of the assessment. The code is shown in the screenshot below.

```
DROP TABLE IF EXISTS wgu_medical_data;
CREATE TABLE wgu_medical_data AS(

        SELECT pt.lat AS latitude, pt.lng AS longitude,
               pt.age, pt.readmis, pt.gender, pt.income AS "Patient Income",
               loc.state, loc.county, loc.zip, loc.city,
               serv.arthritis, serv.overweight, serv.asthma, serv.diabetes,
               pt.hignblood AS high_blood_pressure, pt.stroke
        FROM patient AS pt
        INNER JOIN location AS loc ON pt.location_id = loc.location_id
        INNER JOIN servicesaddon AS serv ON pt.patient_id = serv.patient_id

);
```

Image 1. Code for the creation of the wgu_medical_data table.

This script was created through the use of the pgadmin application. It allowed us to easily create a table that would be assimilated into the existing database. This script created a table that contained all the information that was required from the tables that were supplied by

[1] All code will be screenshots. This is to make it easier to keep the formatting in its state. If required you can view the .zip file that has the original SQL files.

the university. Please note that three tables are being used for the creation of this table. The tables are the following

- patient
- location
- servicesaddon

These provided all the informational demographics that are required to create the dashboard utilizing this information. This table will later be joined in a subsequent SQL script.

The next script will be a simple script to change the "Prefer not to answer" to the "Nonbinary." This is to be in keeping with how the data dictionaries in past assessments refer to this category. It was not required but it was performed in keeping with the previous assignment requests for preparation. The script is shown below.

```
-- Change the gender.
UPDATE wgu_medical_data
SET gender = REPLACE (gender,
                                        'Prefer not to answer',
                                        'Nonbinary');
```

Image 2. Code for the changing of gender.

The third script is the one that will create the table for the ancillary dataset. The census data from Kaggle. The script will read in all the columns that are in the dataset. None of the columns of the source CSV will be omitted. There are 37 columns in the source CSV. The script is shown in the following image.

```
DROP TABLE IF EXISTS us_census_data;

CREATE TABLE us_census_data(
        countyId integer NOT NULL,
        state text,
        county text,
        total_pop integer,
        men Integer,
        women Integer,
        hispanic float,
        white float,
        black float,
        native float,
        asian float,
        pacific float,
        voting_age_citizen integer,
        income integer,
        income_error integer,
        income_per_cap integer,
        income_per_cap_err integer,
        poverty float,
        child_poverty float,
        professional float,
        service float,
        office float,
        construction float,
        production float,
        drive float,
        carpool float,
        transit float,
        walk float,
        other_trans float,
        workAtHome float,
        mean_Commute float,
        employed float,
        private_Work float,
        public_work float,
        self_employed float,
        family_work float,
        unemployment float

    CONSTRAINT countyId_pkey PRIMARY KEY (countyId)

)
```

Image 3. Code for the creation of the us_census_data table.

Note: After completing this step you will need to import the file named that was given earlier in section **acs2017_county_data.csv**. This will provide the data for the table. This will need to be completed before moving on to the next set of SQL scripts.

The next script is to modify how the county is displayed in the ancillary file. The add-on file uses the word "County" in all references to the state's county. This is not the case in the dataset that is supplied by the university. This discrepancy will be changed. This will make the column in the census data into just the name of the county and nothing. This will make it easier to join as it will be one of the fields that will be required to have a successful join. The code can be seen in the following image.

```
-- Change county to a single word
UPDATE us_census_data
SET county = TRIM (TRAILING 'County' FROM county);
```

Image 4. Code for removing "County" from the us_census_data table.

The next script that will be used to create a suitable dataset is the changing of the state column in the census data. The values stored in the column for the census use the full name of the state or territory. This is not how it was stored in the university-provided dataset. This dataset stores only the two-letter abbreviation for the state. The code below shows how these values were changed. It only affected 23 values indicating that not all states are reflected in the dataset.

```
-- Change the state to two letter abbreviation

-- A states
UPDATE us_census_data SET state = 'AL' WHERE state = 'Alabama';
UPDATE us_census_data SET state = 'AK' WHERE state = 'Alaska';
UPDATE us_census_data SET state = 'AZ' WHERE state = 'Arizona';
UPDATE us_census_data SET state = 'AR' WHERE state = 'Arkansas';

-- C states
UPDATE us_census_data SET state = 'CA' WHERE state = 'California';
UPDATE us_census_data SET state = 'CO' WHERE state = 'Colorado';
UPDATE us_census_data SET state = 'CT' WHERE state = 'Connecticut';

-- D states.
UPDATE us_census_data SET state = 'DE' WHERE state = 'Delaware';
UPDATE us_census_data SET state = 'DC' WHERE state = 'District of Columbia';

-- F states
UPDATE us_census_data SET state = 'FL' WHERE state = 'Florida';

-- G states.
UPDATE us_census_data SET state = 'GA' WHERE state = 'Georgia';

-- H states.
UPDATE us_census_data SET state = 'HI' WHERE state = 'Hawaii';

-- I states.
UPDATE us_census_data SET state = 'ID' WHERE state = 'Idaho';
UPDATE us_census_data SET state = 'IL' WHERE state = 'Illinois';
UPDATE us_census_data SET state = 'IN' WHERE state = 'Indiana';
UPDATE us_census_data SET state = 'IA' WHERE state = 'Iowa';

--K states
UPDATE us_census_data SET state = 'KA' WHERE state = 'Kansas';
UPDATE us_census_data SET state = 'KY' WHERE state = 'Kentucky';

-- L states.
UPDATE us_census_data SET state = 'LA' WHERE state = 'Louisiana';
```

```
-- M states.
UPDATE us_census_data SET state = 'ME' WHERE state = 'Maine';
UPDATE us_census_data SET state = 'MD' WHERE state = 'Maryland';
UPDATE us_census_data SET state = 'MA' WHERE state = 'Massachusetts';
UPDATE us_census_data SET state = 'MI' WHERE state = 'Michigan';
UPDATE us_census_data SET state = 'MN' WHERE state = 'Minnesota';
UPDATE us_census_data SET state = 'MS' WHERE state = 'Mississippi';
UPDATE us_census_data SET state = 'MO' WHERE state = 'Missouri';
UPDATE us_census_data SET state = 'MT' WHERE state = 'Montana';

-- N states.
UPDATE us_census_data SET state = 'NE' WHERE state = 'Nebraska';
UPDATE us_census_data SET state = 'NV' WHERE state = 'Nevada';
UPDATE us_census_data SET state = 'NH' WHERE state = 'New Hampshire';
UPDATE us_census_data SET state = 'NJ' WHERE state = 'New Jersey';
UPDATE us_census_data SET state = 'NM' WHERE state = 'New Mexico';
UPDATE us_census_data SET state = 'NY' WHERE state = 'New York';
UPDATE us_census_data SET state = 'NE' WHERE state = 'Nebraska';
UPDATE us_census_data SET state = 'NC' WHERE state = 'North Carolina';
UPDATE us_census_data SET state = 'ND' WHERE state = 'North Dakota';

-- O states
UPDATE us_census_data SET state = 'OH' WHERE state = 'Ohio';
UPDATE us_census_data SET state = 'OR' WHERE state = 'Oregon';
UPDATE us_census_data SET state = 'OK' WHERE state = 'Oklahoma';

-- P states.
UPDATE us_census_data SET state = 'PA' WHERE state = 'Pennsylvania';
UPDATE us_census_data SET state = 'PR' WHERE state = 'Puerto Rico';

-- R states.
UPDATE us_census_data SET state = 'RI' WHERE state = 'Rhode Island';

-- S states.
UPDATE us_census_data SET state = 'SC' WHERE state = 'South Carolina';
UPDATE us_census_data SET state = 'SD' WHERE state = 'South Dakota';
```

```
-- T states.
UPDATE us_census_data SET state = 'TN' WHERE state = 'Tennessee';
UPDATE us_census_data SET state = 'TX' WHERE state = 'Texas';

-- U states
UPDATE us_census_data SET state = 'UT' WHERE state = 'Utah';

-- V states
UPDATE us_census_data SET state = 'VT' WHERE state = 'Vermont';
UPDATE us_census_data SET state = 'VA' WHERE state = 'Virginia';

-- W states.
UPDATE us_census_data SET state = 'WA' WHERE state = 'Washington';
UPDATE us_census_data SET state = 'WV' WHERE state = 'West Virginia';
UPDATE us_census_data SET state = 'WI' WHERE state = 'Wisconsin';
UPDATE us_census_data SET state = 'WY' WHERE state = 'Wyoming';
```

Image 5. Code for updating the states to their abbreviation.

To make sure the columns that will be used to join the two tables together are properly formatted. There was a need to clean up the whitespace that was found in the columns. If this was not done there would be an inability to join the two tables together. The script that was used is in the image below.

```
-- TRIM whitespace
UPDATE us_census_data
SET county = TRIM (BOTH FROM county);

UPDATE wgu_medical_data
SET county = TRIM (BOTH FROM county);
```

Image 6. Code for trimming the whitespace.

The script while not required to make the tables "joinable" was used to make the table easier to read. The script drops all unnecessary columns that will not be included in the final table which will include the data from both the university-supplied tables as well as the census

data that is included in this table. The image shows the script that was used to drop the unneeded columns from the table.

```
ALTER TABLE us_census_data
DROP COLUMN IF EXISTS hispanic,
DROP COLUMN IF EXISTS white,
DROP COLUMN IF EXISTS black,
DROP COLUMN IF EXISTS asian,
DROP COLUMN IF EXISTS pacific,
DROP COLUMN IF EXISTS native,
DROP COLUMN IF EXISTS voting_age_citizen,
DROP COLUMN IF EXISTS income_error,
DROP COLUMN IF EXISTS income_per_cap,
DROP COLUMN IF EXISTS income_per_cap_err,
DROP COLUMN IF EXISTS poverty,
DROP COLUMN IF EXISTS child_poverty,
DROP COLUMN IF EXISTS professional,
DROP COLUMN IF EXISTS service,
DROP COLUMN IF EXISTS office,
DROP COLUMN IF EXISTS construction,
DROP COLUMN IF EXISTS production,
DROP COLUMN IF EXISTS drive,
DROP COLUMN IF EXISTS carpool,
DROP COLUMN IF EXISTS transit,
DROP COLUMN IF EXISTS walk,
DROP COLUMN IF EXISTS other_trans,
DROP COLUMN IF EXISTS workathome,
DROP COLUMN IF EXISTS mean_commute,
DROP COLUMN IF EXISTS employed,
DROP COLUMN IF EXISTS private_work,
DROP COLUMN IF EXISTS public_work,
DROP COLUMN IF EXISTS self_employed,
DROP COLUMN IF EXISTS family_work,
DROP COLUMN IF EXISTS unemployed
;
```

Image 7. Code for dropping the columns.

The second to final script will be the creation of a single table that will hold the information that is joined from the two tables of data. This script will join the two tables on two

columns. The columns are county and state. This join will use a full join and will result in 10,904 rows. The script is shown below.

```
-- Create the joined table.
DROP TABLE IF EXISTS all_medical_data;

CREATE TABLE all_medical_data AS(
        SELECT latitude,longitude, age,
        readmis, gender, wmd.state,
        wmd.county, zip, city,wmd."Patient Income", arthritis, overweight, asthma, diabetes,
        high_blood_pressure, stroke,
        usc.state AS "Census State",
        usc.county AS "Census County", usc.total_pop as "Total Population",
        usc.income AS "Income", usc.men AS "Men Population",
        usc.women AS "Women Population", usc.unemployment AS "Census Unemployment"
        FROM wgu_medical_data AS wmd
        Full JOIN US_census_data as usc ON wmd.state = usc.state
        AND wmd.county = usc.county
);
```

Image 8. Code for creating the all_medical_data table.

Please note that some of the columns were given aliases to help differentiate them. For example, income is a common column but refers to two different types of data being recorded. One is the patient's data while the other is from the census. The columns do not have the same data and should be titled differently.

Looking at the data there was a realization that there were numerous rows that were appended to the bottom of the table that were no use. They were empty or null. These will need to be removed as they serve no purpose when it comes to being analyzed. The SQL script is shown in the following image.

```
-- Drop empty rows from the table.
-- The empty rows will not be needed
-- in the final part of the assessment.
-- DROP the reows that do not have latitude.

DELETE FROM all_medical_data
WHERE latitude IS NULL;
```

Image 9. Code for dropping null rows.

After all these scripts are run there will be a single table that can be used for the database. All the other tables will be left in place and will not be dropped as they may be needed for future analysis projects. Using one data stream (all_medical_data) will make it easier to connect to and see the relevant information for the desired dashboards. You will not need to look through all the tables to find the needed information.

Part 2. Demonstration.

This section contains the link to the Panopto presentation. Topics that will be covered are the following:

- Description of the technical environment.
    - The technical environment is a virtual machine that currently has a Postgres database and a copy of Tableau version 2021.4. The virtual environment is provided by Labs on Demand.
- Functionality of the dashboards.
    - The functionality of the dashboard is to provide some idea of who the patients are what are some of the conditions that they exhibit and in what age range these are most prevalent.
    - The dashboard will allow the user to look at information about the patient as well as get a feel as to how the patients are distributed across the United States.
- Explanation of the SQL scripts used to support the dashboards.

  - The scripts were used to create the ancillary table and to clean some of the data to make it more like the previous assignment. Some SQL scripts were used to extract the data from the database for use in the Tableau dashboards. These will be the scripts that were discussed in section **A4 Coding the Dashboard**.
- How the streams were prepared to support analysis. To streamline the process. There will be essentially one stream to the Postgres database. The connection will be to the main table which is the joining of the data provided by the university and the census data. The name of the table in the **medical_data** database is **all_medical_data**.
- Description of how the data were aligned with the other data points.
  - The data that was included gave additional stats about where the patients live and what the demographics are as a whole for that particular county.
  - It is to give better insight into the types of people that reside there. And to get a feel for those who may exhibit demographics that are not in line with other demographics that may need more or additional services.
- Demonstrate how the databases were created.
  - There is a single database that was used to create the dashboards. While there were many tables they were eventually consolidated into one table that had all the required information. This was to make it easier to gather the correct information for the desired components of the dashboard.
- Explanation of referential integrity was enforced in the database.
  - To enforce the concept of referential integrity there will be a need to uniquely identify the rows within the table this was accomplished by assigning a primary key to the tables.

The link for the Panopto video is given below.

-

Part 3. The Report

This section will include a discussion of the purpose of the dashboard. There will be a justification for the business intelligence tool used. There will be an explanation of how the data was prepared and readied to be used in analysis and the creation of the dashboard.

C1. Purpose of the Dashboards.

The purpose of the dashboards is to show the demographics of the patients and how they relate to information that is available from an open and trusted source like the US Census. The US Census provided the resources of for looking at the patient versus the country as a whole. While some information was not available for all the states and counties that were found in the university-provided dataset. There was enough of an association to make some interesting observations.

The patient's data is more than just the tests that were performed. A patient is an amalgam of all their information. To better help the patient we need to understand who they are and where they come from. This includes their location as well as the types of conditions that they may have.

Using this dashboard will help see who are patients are and where they are located and can look at the patients and a granular level by looking at them by state, city, etc. This can be seen in the layout of the dashboards and the types of information that they present to the user.

C2. Justification of the Business Intelligence Tool.

The chosen tool was Tableau and it relied on the Postgres database to be the data store from which it will draw in the needed data. Tableau was chosen because it provides a way to access data in a variety of different ways. Tableau is a platform that provides multiple different ways to visualize the data. This will make it easier for interested parties or stakeholders to see relationships between various data elements. Tableau has the ability for the user to interact or create dashboards that can model the data. It can present data in an easy-to-understand manner through the use of the appropriate visualization (Biswal, n.d.).

The audience for this dashboard was those who were considered executives. Tableau allows the dissemination of information in a manner that can provide a larger amount of information in an easy-to-grasp manner. This can be accomplished through visualization the use of appropriate colors and the inclusion of tooltips with the appropriate information. These tooltips can provide more granular information.

C3. Steps Used to Clean the Data.

The data provided by the university seemed to be clean. There were a few anomalies that were found in the dataset when this data was mapped. If you look at some of the points of the provided maps they do not match with the filters. There was no easy way to remedy these errors since this would require the changing of the latitude and longitude. I am not aware of a resource that is available that I can use to remedy these errant points on the map.

Regarding the US Census data, there were no empty or null values in the columns that were summarized by looking at the appropriate informational data on the Kaggle site for the CSV file.

The only time that the data needed to be prepared was in getting it ready to be joined. This was mostly done for the US Census data. The columns for state and county needed to be in a format that would allow the two tables to be joined on these fields.

This was done by removing the "County" from the name of the county. The word "County" does not appear in the university-supplied data. So to join this field they will need to contain the same information.

In the same manner, the state in the US Census data needed to be changed to the two-letter abbreviations that were used in the university data. It also makes it easier to create the appropriate drop-down in the dashboard.

After the joining of the university data and the US Census data, there were null rows located at the bottom of the table. These were rows that did not match with the university data but were included. These were dropped since they were of no use to the creation of the dashboard.

Certain items were not checked for in this assessment. Looking for outliers was not conducted as the types of visualization and calculations were not going to be influenced by these.

As stated earlier there were errant longitude and latitude values in the dataset for the university data. I am not sure if this was on purpose, a coding error on my part, or just uncorrectable errors in the data.

C4. Summarize the Steps Used to Create the Dashboards.

The dashboard was constructed of three informational dashboards that give various demographic information about the patients. The three dashboards are:

- Census Demographics
- WGU Patient Demographics
- Patient Income Map

The steps in creating these were very similar to each other. It is the information that is displayed on them where they differ.

Census Demographics Dashboard

The first dashboard was the Census Demographics dashboard. This dashboard was used to familiarize the user with some general information about the citizenry of the United States. The data displayed on this page made use of a map. This required the longitude and latitude that were supplied by the university data set. The longitude was placed on the Columns shelf and the latitude was placed on the Rows shelf. Next, the Marks card was populated with Income and taking the average. This was used to provide color and as a tooltip.

The county was added as a label for the counties on the map. The total population was added to provide the tooltip with the population of the county in question. The readmission count was added. This count may not be reflective of all the patients as some of the counties may not have data associated with them in the US Census dataset. The unemployment rate was added as tool-tip to provide some employment information about the county

This dashboard has five filters associated with it. If you look at the Filter card you will see more. Some of the additional are "inherited" from other dashboards that are included. They were left as they could prove useful in other analyses.

The filters that are active for this dashboard are:

- State – allows the used to filter by the state
- Gender – to filter by gender

- Readmis – to filter by readmission status
- Men Population – to filter by the population of men
- Women Population – to filter by the population of women

Patient Demographics Dashboard

This dashboard will provide individual information on each of the patients. It will be composed of nine filters to help look at various combinations of attributes that are exhibited by the patients.

The data displayed on this page made use of a map. This required the longitude and latitude that were supplied by the university dataset. The longitude was placed on the Columns shelf and the latitude was placed on the Rows shelf.

Gender was added to the color card and this was used to color code the patients based on their gender. Most of the information will be found in the tool tip as this seemed like the easiest way to display this information. This tooltip is composed of the following information:

- Gender
- Age
- ZIP code
- Income
- City
- Readmitted

The filters that are active for this dashboard are:

- State
- Asthma
- Overweight
- City
- Arthritis
- Stroke
- Gender
- Diabetes
- Arthritis
- High Blood Pressure

These filters will allow the user to filter based on any combination of the filters. This assumes that there is data available for this particular patient.

Patient Income Map

Data Analysis

The patient income map shows the income of the patients within the database. The data displayed on this page made use of a map. This required the longitude and latitude that were supplied by the university dataset. The longitude was placed on the Columns shelf and the latitude was placed on the Rows shelf.

This dashboard utilized a calculated field to display the average income of the patients in the database. This was then affixed to the tooltip. This was to show how the patient's income compared to the average of others in the database.

Most of the composition of this dashboard was adding the appropriate information to the tooltip. Information that will be included on the tooltip is the following:

- Patient Income
- State
- Gender
- City
- Age
- Average Patient Income – the calculated field

The filters that are active for this dashboard are:

- State
- Age categories – to allow the user to look at the patients based on age bins.
    - These bins are in ten-year increments.

Each of these dashboards included a reset button to reset the filter. This was created by including another worksheet in the dashboard. The reset worksheet made use of actions that allowed the filters to be reset to their original state.

First, create a new worksheet and name it. To accomplish this you need to go to the **Worksheet → Actions → Add Action → Filter**. Then give the **Filter Action** a name. Select the worksheet name for the **Source Sheets** section. Choose **Select** in the **Run action on** section. Then on the **Target Sheets** section choose the appropriate sheet and select **Show all values** for the **Clearing the selection will** section. Then you will be able to add this to the dashboard.

C5. Results of the Data Analysis.

The results of the data show that our patients are found mostly in the eastern half of the United States. The data shows that patients are found mostly east of the Mississippi River. The number of patients dwindles as we move further west. This does not seem that surprising as the middle of the country is more sparsely populated

If we look at the gender of the we can see that it seems that there are slightly more women in the sample that was provided in the university sample. Looking at the census data we can see that as we choose a county some of the counties have more women or some have more women.

Looking at the results it may be worth looking to fill the holes in the data. Some data points have missing data and this could be handled in a few ways. The data could be left as is. The data could be imputed, but this approach may yield erroneous observations. We could gather further data that would fill in the gaps.

Drilling down in the data using the various filters we can see that there is a lot of diversity in the patient population. We can see that the patients come from various economic and geographic backgrounds.

In conclusion, using the maps to look at the distribution of patients we can see what constitutes the patient base of the university dataset. We can see that we do not have many patients in the interior of the country. This could be an area that could be a place to expand the services that are offered. Before that route is pursued it might be beneficial to see what are the causes of this lack of patients in this area of the country. The data that was used will not be able to provide answers to this observation.

C6. Limitations of the Analysis. Looking at the analysis we can see that we have gaps in both data sets. Not all the counties of the United States have data in the Census data that was used in the creation of the dashboards. We do not have data for those under the age of 18 as this was always been commented upon in past analysis.

There is a need to procure a secondary dataset that more closely matches the types of data that are found in the university-provided data. This could come from looking into datasets that are from the US census itself as opposed to the ones that are offered on sites like Kaggle.

References

D. Web sources for the data or other sources used in the document.

US Census Demographic Data. (n.d.). Kaggle. Retrieved December 27, 2023, from https://www.kaggle.com/datasets/muonneutrino/us-census-demographic-data?resource=download&select=acs2017_county_data.csv

E. In-text Citation and References.

Biswal, A. (2023, November 30). *Power BI vs Tableau: Which Is Better Data Visualization Tool*. Simplilearn.com. Retrieved December 23, 2023, from https://www.simplilearn.com/tutorials/power-bi-tutorial/power-bi-vs-tableau