**D211 Performance Assessment**

**ADVANCED DATA ACQUISITION FOR MEDICAL DATA**

Fahim A. Akbar Student ID 001434895 Masters Data Analytics (January 1, 2021) Program Mentor: Mandy Rasmuson (801) 891-6865x5957 [fakbar3@wgu.edu](mailto:fakbar3@wgu.edu)

**Part 1: Data Dashboards**

Link to the dashboard: D211 Dashboard -

**A1. Datasets**

The data sets that serve as the data source for my dashboard are “medical\_data” and “healthcare\_dataset\_stroke\_data.” **DataSet 1** is represented by the medical\_data data set and **DataSet 2** is represented by the diabetic dataset.

Source for “medical\_data”

<https://tasks.wgu.edu/student/001434895/course/23540008/task/2804/overview>

Source for “healthcare\_dataset\_stroke\_data”

<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

**A2. Installation Guide:**

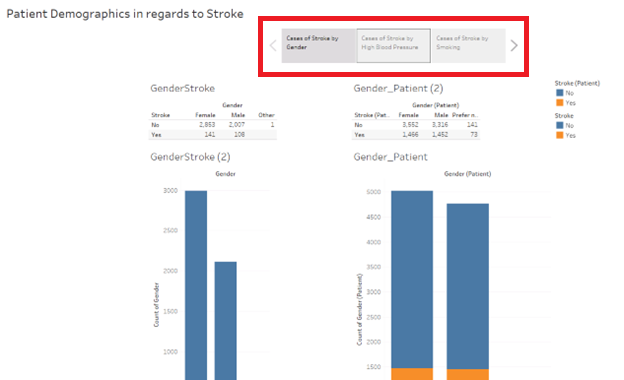
D211Task1.twbx file is attached with this submission, and a link to the D211Task1.twbx file that can be downloaded via Microsoft OneDrive to the virtual machine. The dashboard is provided with the task submission as a .twbx file, and can be opened with Tableau reader.

Here is how to install the dashboard on the Labs on Demand Virtual Machine:

* Open the virtual machine. Download the “D211Task1.twbx” file provided in this assessment submission, and open File Explorer in the virtual machine. Navigate to ‘C:\Users\Public\Downloads’ (directory) and paste the provided “D211Task1.twbx” file to this folder.
* Alternative way to open the file is to download directly through OneDrive with the provided link. Once downloaded, navigate to directory and open the file.
* Download the healthcare-dateaset-stroke-data.csv file from OneDrive. The file is also provided as an attachment with this submission.
* Navigate to pgAdmin in the virtual machine. Expand Servers (1), PostgresSQL 13, Databases (3). medical\_data, Schemas (1), public, Tables (7), and finally patient. Right click ‘patient’, select Query Tool and navigate to SQL tab
* Enter the following SQL code before proceeding.:
  + ALTER TABLE patient
  + RENAME COLUMN hignblood TO highblood;
* Exit pgAdmin
* Download the “healthcare\_dataset\_stroke\_data.csv” file provided in this assessment submission, and open File Explorer in the virtual machine. Navigate to ‘C:\Users\Public\Downloads’ (directory) and paste the provided “healthcare\_dataset\_stroke\_data.csv” file to this folder.
* Alternative way to open the file is to download directly through OneDrive with the provided link. Once downloaded, navigate to directory and open the file
* Double click on Tableau 2021.4 to launch
* At the top left, click File > Open and navigate to ‘C:\Users\Public\Downloads’. Select the Book1.twb file and open the file

**A3. Dashboard Navigation:**

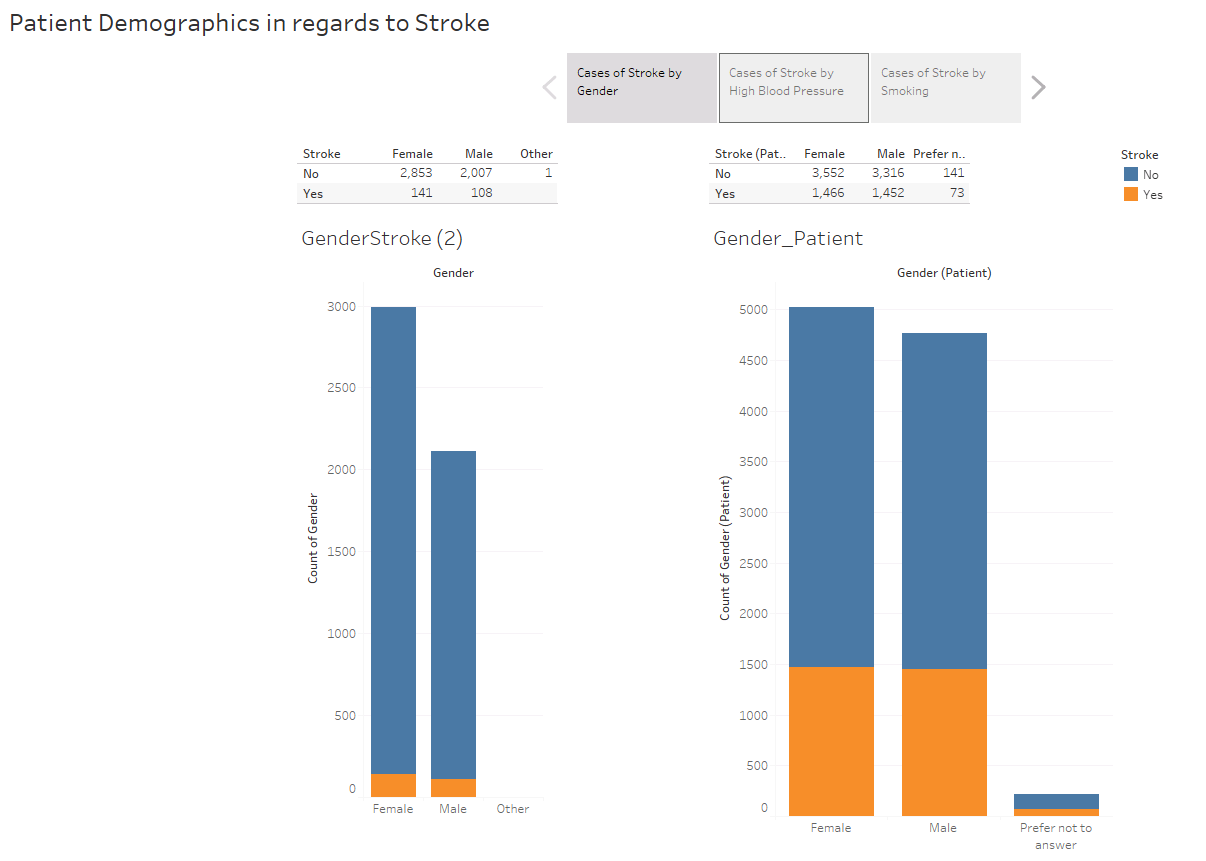
* The dashboard is presented as a story divided into three sections:
  + The first section “Cases of Stroke by Gender” focuses on the relationship between gender and likelihood of the stroke
  + The second section “Cases of Stroke by High Blood Pressure” focuses on the relationship between high blood pressure and the likelihood of stroke
  + The third section “Cases of Stroke by Smoking” represents the relationship between smoking and the likelihood of stroke



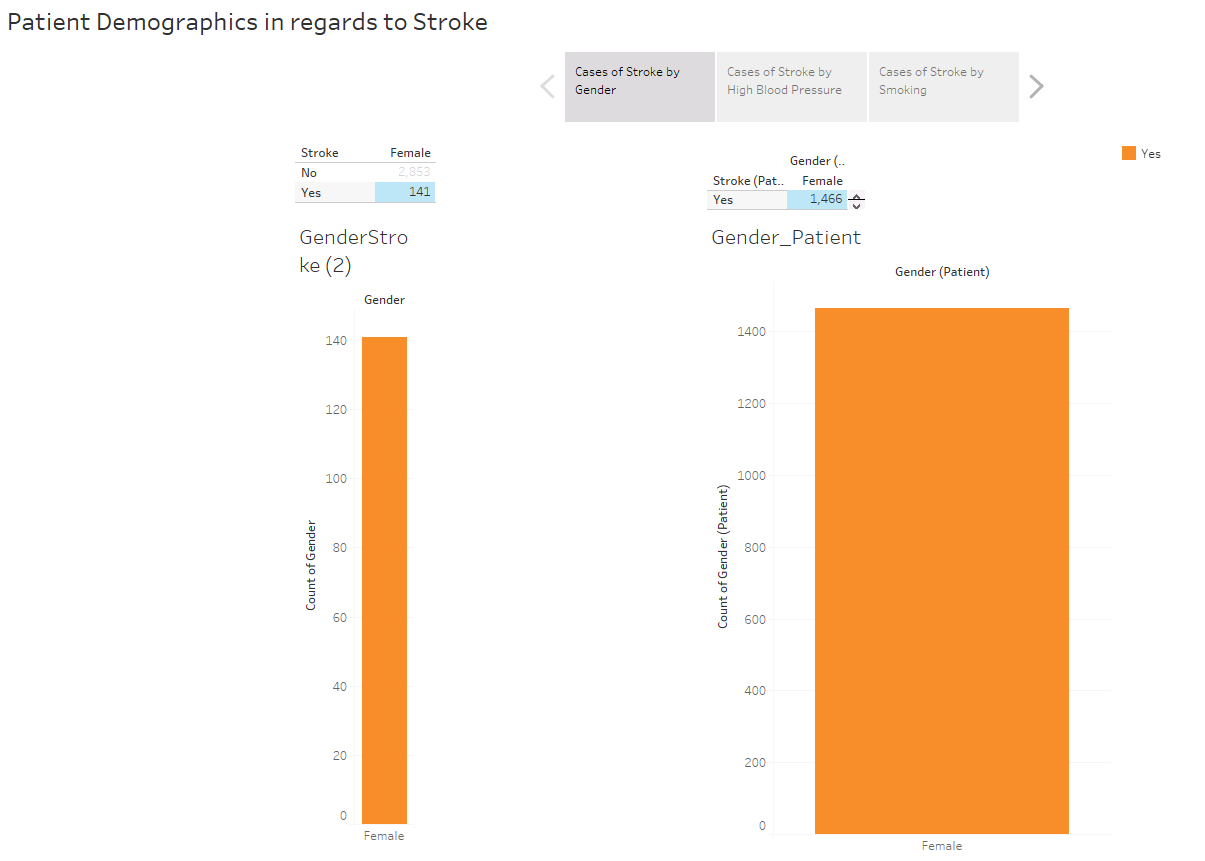
* To access a section, navigate to and select the respective tab at the top of the screen.
* To toggle between each section, use the arrow keys on the left and right ends of the tabs.



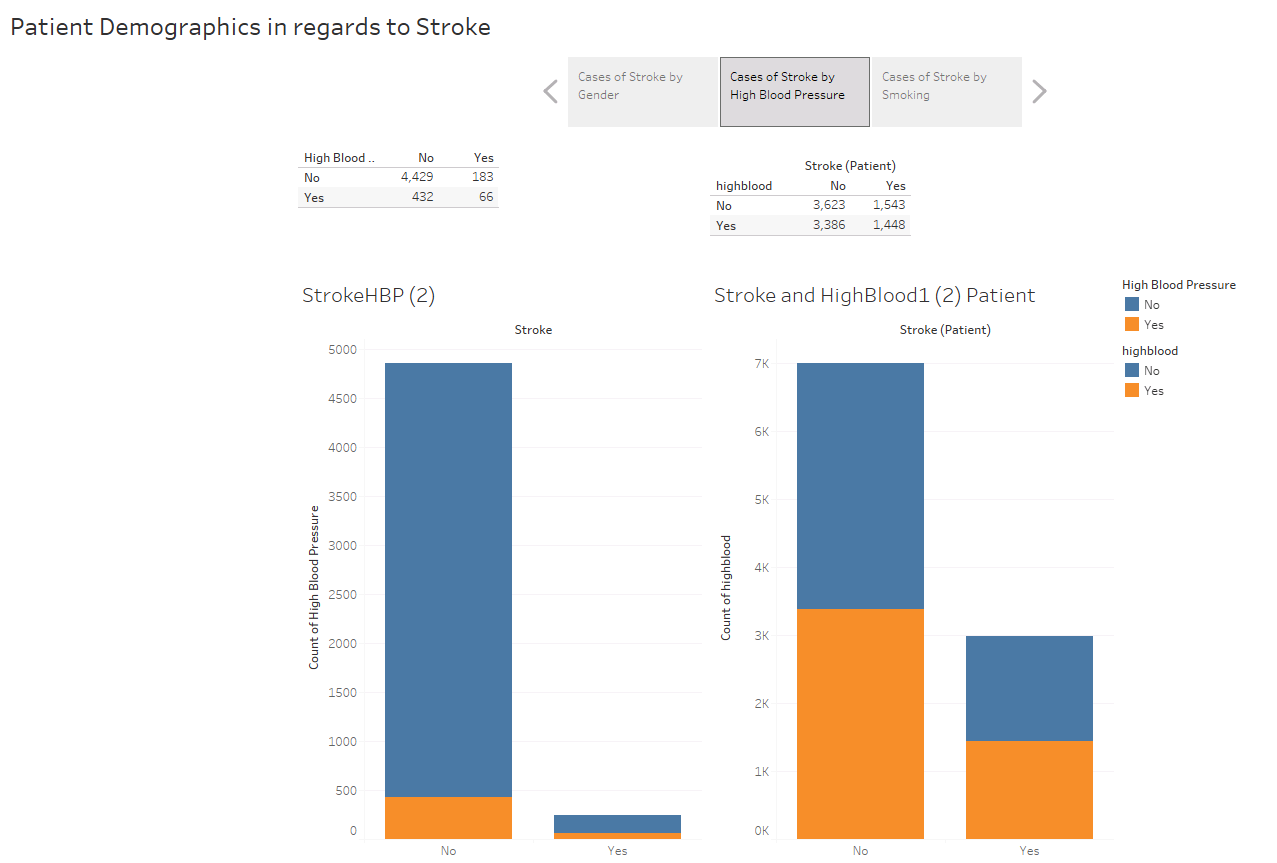
* The bar graphs for gender and smoking in this dashboard follow a same color-coded legend for Stroke: Blue indicates “No” for stroke and Orange indicates “Yes” for stroke. These colors were chosen to account for potential viewers being color blind.
* The bar graphs for High blood pressure section in this dashboard follow a similar color-coded legend for High blood pressure: Blue indicates “No” for High blood pressure and Orange indicates “Yes” for High blood pressure. These colors were chosen to account for potential viewers being color blind.
* Beginning with section one, to explore the section select the tile titled “Cases of Stroke by Gender”



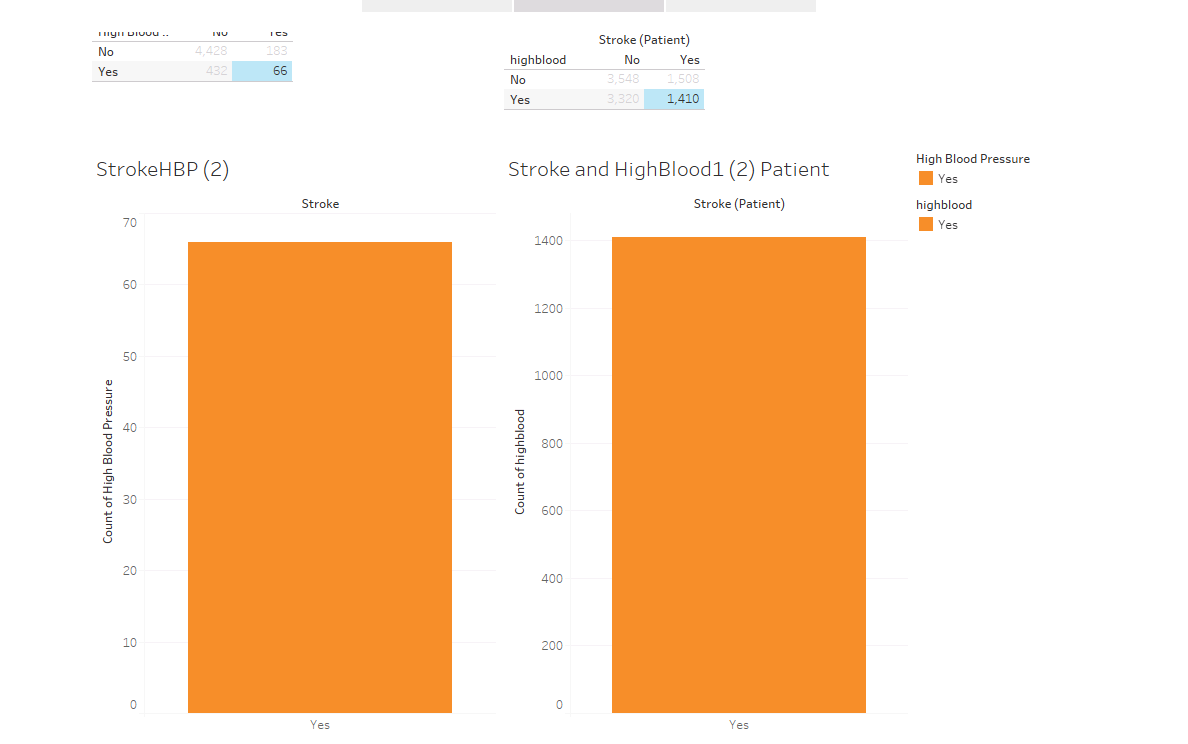
* The figures on the left represent data from healthcare\_dataset\_stroke\_data and the figures on the right represent data from medical\_data.
* The two tables on the top of the dashboard represent filters that can be used to show only instances of a specific demographic.
* For example, if you only want to see the number of female patients with stroke from both datasets, you would click on the field that represents the number of patients that are female with “Yes” for stroke, as shown below:



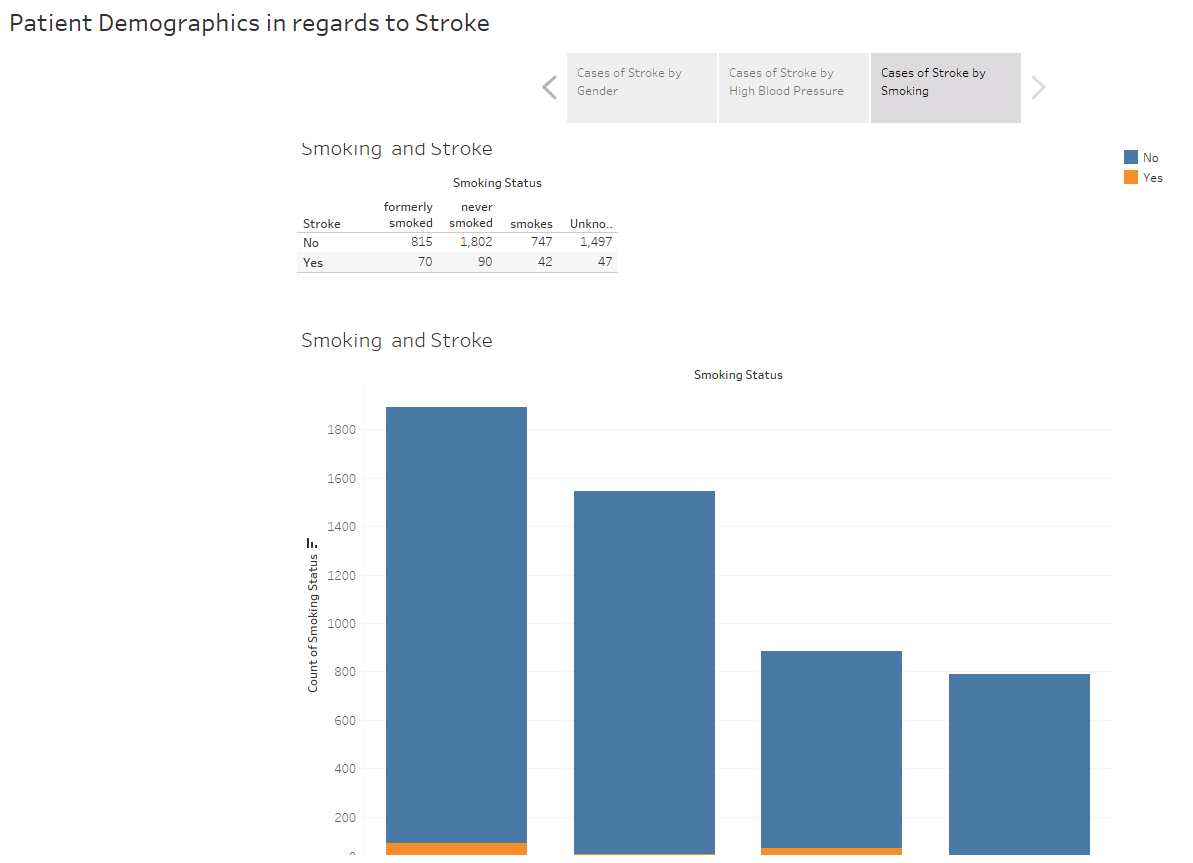
* To explore section 2, select the tile titled “Cases of Stroke by High Blood Pressure”



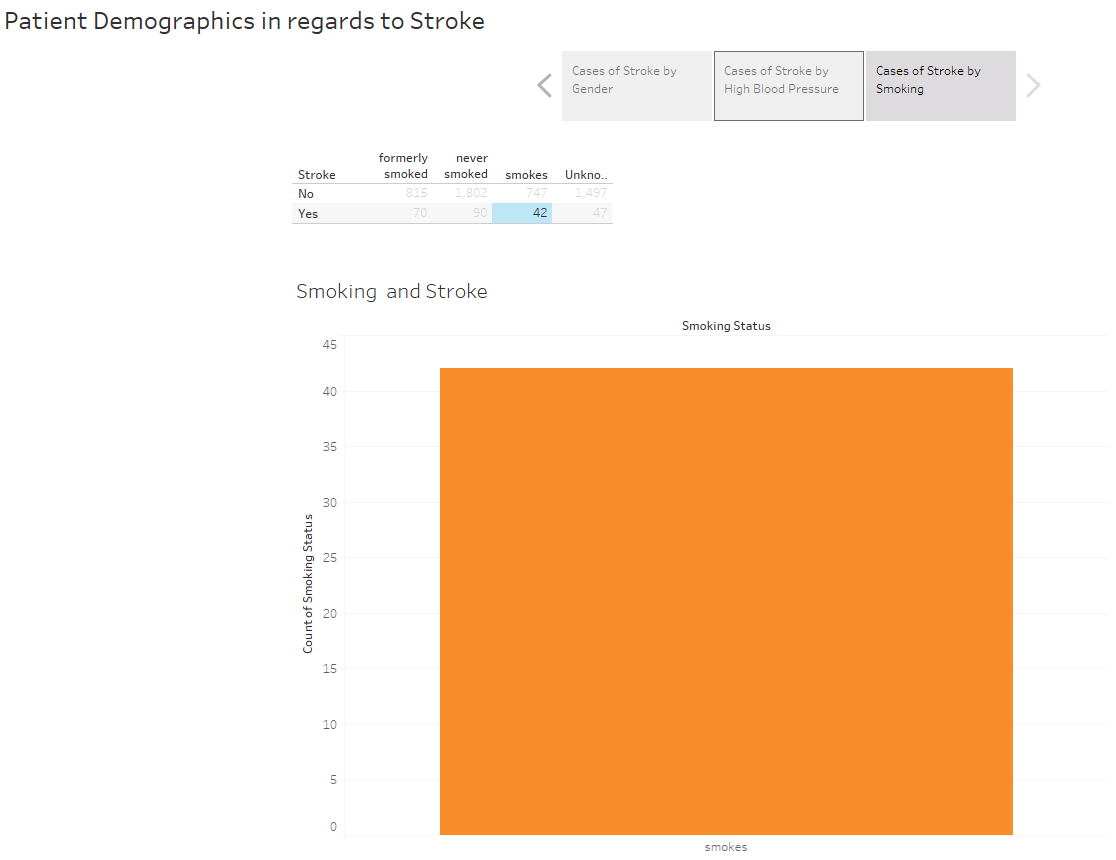
* The figures on the left represent data from healthcare\_dataset\_stroke\_data and the figures on the right represent data from medical\_data.
* The two tables on the top of the dashboard represent filters that can be used to show only instances of a specific demographic.
* For example, if you only want to see the amount of patients with both stroke and high blood pressure from both datasets, you would click on the field that represents the number of patients that are female with “Yes” for stroke, as shown below:



* Beginning with section one, to explore the section select the tile titled “Cases of Stroke by Smoking”



* The figures represents data from healthcare\_dataset\_stroke\_data
* The table on the top of the dashboard represent filters that can be used to show only instances of a specific demographic.
* For example, if you only want to see the amount of patients who smoke and have had stroke, you would click on the field that represents the number of patients under smoking with “Yes” for stroke, as shown below:



**A4: SQL code and supporting code:**

SQL Code for cleaning the data:

ALTER TABLE patient

RENAME COLUMN hignblood TO highblood;

SQL Code for the medical\_clean patient table:

-- Table: public.patient

-- DROP TABLE public.patient;

CREATE TABLE public.patient

(

patient\_id text COLLATE pg\_catalog."default" NOT NULL,

lat numeric,

lng numeric,

population integer,

children integer,

age integer,

income numeric,

marital text COLLATE pg\_catalog."default",

readmis text COLLATE pg\_catalog."default",

gender text COLLATE pg\_catalog."default",

initial\_days numeric,

totalcharge numeric,

additional\_charges numeric,

vitd\_levels numeric,

doc\_visits integer,

full\_meals integer,

vitd\_supp integer,

soft\_drink text COLLATE pg\_catalog."default",

highblood text COLLATE pg\_catalog."default",

stroke text COLLATE pg\_catalog."default",

job\_id integer,

compl\_id integer,

admis\_id integer,

location\_id integer,

CONSTRAINT patient\_pkey PRIMARY KEY (patient\_id),

CONSTRAINT patient\_admis\_id\_fkey FOREIGN KEY (admis\_id)

REFERENCES public.admission (admins\_id) MATCH SIMPLE

ON UPDATE NO ACTION

ON DELETE NO ACTION,

CONSTRAINT patient\_compl\_id\_fkey FOREIGN KEY (compl\_id)

REFERENCES public.complication (complication\_id) MATCH SIMPLE

ON UPDATE NO ACTION

ON DELETE NO ACTION,

CONSTRAINT patient\_job\_id\_fkey FOREIGN KEY (job\_id)

REFERENCES public.job (job\_id) MATCH SIMPLE

ON UPDATE NO ACTION

ON DELETE NO ACTION,

CONSTRAINT patient\_location\_id\_fkey FOREIGN KEY (location\_id)

REFERENCES public.location (location\_id) MATCH SIMPLE

ON UPDATE NO ACTION

ON DELETE NO ACTION

)

TABLESPACE pg\_default;

ALTER TABLE public.patient

OWNER to postgres;

Convert to custom sql:

SELECT "patient"."additional\_charges" AS "additional\_charges",

"patient"."admis\_id" AS "admis\_id",

"patient"."age" AS "age (patient)",

"patient"."children" AS "children",

"patient"."compl\_id" AS "compl\_id",

"patient"."doc\_visits" AS "doc\_visits",

"patient"."full\_meals" AS "full\_meals",

CAST("patient"."gender" AS TEXT) AS "gender (patient)",

CAST("patient"."highblood" AS TEXT) AS "highblood",

"patient"."income" AS "income",

"patient"."initial\_days" AS "initial\_days",

"patient"."job\_id" AS "job\_id",

"patient"."lat" AS "lat",

"patient"."lng" AS "lng",

"patient"."location\_id" AS "location\_id",

CAST("patient"."marital" AS TEXT) AS "marital",

CAST("patient"."patient\_id" AS TEXT) AS "patient\_id",

"patient"."population" AS "population",

CAST("patient"."readmis" AS TEXT) AS "readmis",

CAST("patient"."soft\_drink" AS TEXT) AS "soft\_drink",

CAST("patient"."stroke" AS TEXT) AS "stroke (patient)",

"patient"."totalcharge" AS "totalcharge",

"patient"."vitd\_levels" AS "vitd\_levels",

"patient"."vitd\_supp" AS "vitd\_supp"

FROM "public"."patient" "patient"

SELECT CAST("patient"."gender" AS TEXT) AS "gender (patient)",

CAST("patient"."stroke" AS TEXT) AS "stroke (patient)"

FROM "public"."patient" "patient"

GROUP BY 1,

2

SELECT CAST(""patient"".""gender"" AS TEXT) AS ""gender (patient)"",

CAST(""patient"".""stroke"" AS TEXT) AS ""stroke (patient)""

FROM ""public"".""patient"" ""patient""

WHERE ((CAST(""patient"".""gender"" AS TEXT) >= 'Female') AND (CAST(""patient"".""gender"" AS TEXT) <= 'Prefer not to answer'))

GROUP BY 1,

2

SQL behind innerjoin of 'gender' between each data set:

SELECT "healthcare-dataset-stroke-data.csv"."gender" AS "gender",

"healthcare-dataset-stroke-data.csv"."stroke" AS "stroke"

FROM "db2002"."TableauTemp"."healthcare-dataset-stroke-data#csv" "healthcare-dataset-stroke-data.csv"

INNER JOIN "public"."Tableau\_14\_73AE7894-9F48-4570-BC1E-8CC9B5E2615F\_4\_FQ\_Temp\_1" "FQ\_Temp\_1" ON ("healthcare-dataset-stroke-data.csv"."gender" = "FQ\_Temp\_1"."gender")

WHERE ("healthcare-dataset-stroke-data.csv"."gender" = 'Male') AND ("healthcare-dataset-stroke-data.csv"."gender" = 'Female') <= 'unknown'))

GROUP BY 1,

2

SQL behind visualizations for the Patient Table:

SELECT CAST("patient"."gender" AS TEXT) AS "gender (patient)",

CAST("patient"."stroke" AS TEXT) AS "stroke (patient)"

FROM "public"."patient" "patient"

GROUP BY 1,

2

"SELECT CAST(""patient"".""highblood"" AS TEXT) AS ""highblood"",

CAST(""patient"".""stroke"" AS TEXT) AS ""stroke (patient)""

FROM ""public"".""patient"" ""patient""

GROUP BY 1,

2"

**Part 2: Demonstration**

**B. Panopto Demonstration**

Link to the Panopto Recording:

**Transcript of the presentation:**

today I would like to give you all an overview of the demographics of patients with stroke.

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

The dashboards for this assessment were all created in Tableau, and were designed to showcase the likelihood of a patient having stroke based on three factors: gender, high blood pressure, and smoking.

2. Demonstrate the functionality of the dashboards.

>

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

The SQL code used to support the dashboard creation selects the gender and stroke fields from the patient table, and casts the gender to a text type to be displayed in a Tableau visualization, such as a bar graph or text table. Similar SQL code was used for generating visualizations regarding high blood pressure and stroke from the patient table.

The Select is pulling the gender and stroke fields from the patients table of the public medical\_clean database. Cast makes these fields into a datatype that can then be presented

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

Medical\_data was prepared by editing the patient table column “hignblood” to “highblood” in postgress before adding a connection to Tableau.

Healthcare\_dataset\_stroke\_data was prepared by removing all null values

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

6. Demonstrate how the databases were created.

>show the sql code

ALTER TABLE patient

RENAME COLUMN hignblood TO highblood;

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

>join between data sets via ‘gender’

**D211 Performance Assessment**

**ADVANCED DATA ACQUISITION FOR MEDICAL DATA**

Fahim A. Akbar Student ID 001434895 Masters Data Analytics (January 1, 2021) Program Mentor: Mandy Rasmuson (801) 891-6865x5957 [fakbar3@wgu.edu](mailto:fakbar3@wgu.edu)

**Part 3: Report**

**C1. Purpose and function of dashboard**

The goal of this project was to construct a data dashboard aimed at allowing executive leaders to delve into medical data, recognize patterns, and analyze essential metrics. The purpose is to assist a hospital chain in evaluating patient data for insights into risks and contributing factors of stroke. The created dashboard incorporates 10,000 patient records from the hospital chain and 4190 patient records from another hospital with the focus study being on patients likelihood of stroke.

**C2. Business intelligence tool selection**

The dashboard was created in Tableau Desktop. Tableau was selected as the business intelligence tool for this assessment because of its ability to create intricate and interactive data visualizations effortlessly. Tableau makes the process of communicating analysis findings easier and facilitates the exploration of intricate data, enabling the discovery of new relationships and intriguing observations. (Loth, 2019).

**C3. Data cleaning and preparation steps**

The medical\_data did not require much cleaning, as there were no nulls, duplicates, or inappropriate entries. One column name had a misspelling. The external data was loaded into a new table in the medical\_data database.

There was one misspelled column name, "hignblood" which was changed to "highblood" using the following sql code:

ALTER TABLE patient

RENAME COLUMN hignblood TO highblood;

The external data file “healthcare\_dataset\_stroke\_data” was cleaned in excel before being imported into Tableau by removing all null values. The file was then imported into Tableau via “Add connection.” In the Add panel, look under “To a file” and select the “more” option. Once selected, “healthcare\_dataset\_stroke\_data” was opened and loaded into Tableau.

**C4. Dashboard creation steps**

Data source connection: To connect Tableau to the medical\_data database in Postgres, go to Add connection.

* In the Add panel, open the dropdown for “To a Server” and select the “more” option.
* Under “Installed Connectors” select “PostgreSQL.” Once selected you will be prompted to enter the appropriate information for the fields.
  + For server, enter “localhost” and for port enter “5432.” For the database, enter “medical\_data.” For authentication, select “Username and Password.” Once selected, enter your user credentials and then select “sign in.”
* To connect Tableau to the healthcare\_dataset\_stroke\_data.csv, go to Add connection.
* In the Add panel, look under “To a file” and select the “more” option. Once selected, you will be redirected to the file manager. Here, select the dataset “healthcare\_dataset\_stroke\_data” and click ‘open.’

Once the data sources are connected, we can proceed with creating the dashboards. Each dashboard in the presentation utilizes two types of visualizations: table

To create a text table:

* Create a new worksheet
* Select the option for text table under “Show me”
* For columns, enter the desired fields.
* Add “stroke” to the rows section
* Under marks, add the selected column. Right click and change measure to ‘count’.
* In “healthcare\_dataset\_stroke\_data” stroke cases are represented by 0 (No) and 1(Yes). For these visualizations, you will need to right click and select “Edit Alias” and convert 0/1 to No/Yes respectively.

To create a bar graph:

* Create a new worksheet
* Select the option for bar graphs under “Show me”
* For columns, enter the desired fields.
* Add “stroke” to the rows section
* Under marks, add the selected column. Right click and change measure to ‘count’.

To create each dashboard section:

* On the left hand panel, select the worksheets for the text table and bar graphs.
* On the right side of each text table, select the and enable the filter key.
* For each bar graph, generate and color legend. Use orange and blue to account for color blindness.

To create the final story:

* Once all the dashboards are done, go to the bottom and click the New Story tab
* Double-click a dashboard on the left to add it to a story point
* Click Add a caption to summarize the story point
* To add another story point, click duplicate to use the initial story point as the basis for a new one
* Once all the story points are set up, add the respective dashboards into the story point

**C5: Results of Data Analysis**

Three primary insights were derived during this analysis. First, both data sets showed that the majority of stroke cases were women. However, this may not necessarily mean that women are more susceptible to stroke then men are. For the patient data from “healthcare-dateaset-stroke-data.csv” 141 women had stroke in contrast to 108 men. For the patient data from “medical\_data” 1,466 women had stroke, in contrast to 1,452 men. These values show that likelihood of stroke is very close for both genders, and the disparity is due to more women in both data sets. Next, the data from both sheets didn’t appear to show a significant correlation between high blood pressure and stroke. For the patient data from “healthcare-dateaset-stroke-data.csv,” 66 patients had both high blood pressure and stroke. For the patient data from “medical\_data,” 1,448 patients had high blood pressure and stroke. This does not necessarily mean that high blood pressure isn’t a strong contributor to likelihood of stroke, but rather it is not a strong contributor for the patient populations of both datasets. Lastly, the relationship between smoking and stroke from the patient data in “healthcare-dateaset-stroke-data.csv” revealed surprising results. The number of stroke patients who formerly smoked exceeded the number of patients who currently smoke, by 70 to 42 respectively. This may suggest that smoking long-term could increase the odds of having a stroke, and even quitting the habit might not be enough to mitigate damage.

**C6: Limitations of Data Analysis**

One limitation of the data analysis is how the nature of the two datasets differ. While their fields and features compliment each other, “healthcare-dateaset-stroke-data.csv” was a case study specific to stroke amongst the patient population of 4,910 patients. “Medical\_data” was a broader study encompassing a variety of features across a patient population of 10,000. Another limitation is that neither dataset specifies if these patients had experienced stroke in the past or multiple times, so these visualizations and conclusions were made under the assumption that “Yes” values represented at least 1 instance of stroke. Lastly, the analysis is limited by the scope of the assessment. While the features tested provided valuable insight, additional features such as age, weight, or other pre-existing conditions could be tested in future analyses to see what other factors contribute to the likelihood of stroke.

**D. Sources**

**Works Cited:**

Loth, A., Vogel, N. and Sparkes, S. (2019) *Visual analytics with tableau*. Indianapolis, IN: John Wiley & Sons, Inc.

**Source for external data:**

<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>