**D211 - Advanced Data Acquisition**

Aaron Balke

December 22nd, 2023

Dr. William Sewell

Western Governors University - College of Information Technology

**Part 1: Data Dashboards**

**A. Dashboards**

The Tableau Dashboards are provided as “Advanced Data Acquisition.twb” with the submission.

**A1. Datasets**

WGU provides the WGU Medical Dataset through the virtual machine.

The other dataset is provided by the National Institute of Diabetes and Digestive and Kidney Diseases through the following link:

https://www.kaggle.com/datasets/nanditapore/healthcare-diabetes/

A copy of the processed datasets as csv files is provided with the submission.

**A2. Installation Instructions**

You will need the “queries.sql” and “Advanced Data Acquisition.twb” files provided with this submission.

Step One: Prepare the Database

Open pgAdmin on your virtual machine and navigate to the medical_data database. Once the medical_data database has been selected, right-click on the database and open the Query Tool. At this time, you may open the ‘queries.sql’ file from the query tool, or paste the file's contents into the blank query space. Then choose Execute File from the top options (or press F5). This will run all the sql queries necessary for this assessment. Please note, that this will execute a curl command, downloading the additional dataset from a remote location.

Step Two: Connect Tableau Dashboard

At this time open the “Advanced Data Acquisition.twb” by double-clicking it. When prompted to connect to the database, please use the credentials provided by the virtual machine to connect - at this time these credentials are username: “postgres” and password: “Passw0rd!”.

**A3. Navigation Instructions**

Please Verify you are on the Advanced Data Acquisition Story Page at the bottom of the Tableau Window. Then choose Presentation Mode (press F7). You may change the dashboard being viewed by using the arrows at the top, or clicking the specific dashboard you are looking for. When you are done press the “Esc” key to leave presentation mode. You may exit the program at this time.

The first dashboard provides data comparing the demographics of our hospital system to our competitors. This includes gender, high blood pressure, and overweight features.

The second dashboard provides KPIs comparing our extreme indicators, the patients with all healthy indicators, and the patients with all unhealthy indicators.

**A4. SQL Code**

```
-- 1 Create View of just Wgu Important Data, Also create AllHealthy and AllUnHealthy Features
CREATE VIEW wgu_view AS(
	SELECT
		patient.age,
		patient.gender,
		patient.hignblood AS highBlood,
		servicesaddon.overweight,
		servicesaddon.diabetes,
		CASE
			WHEN (servicesaddon.overweight = 'No')
			AND (servicesaddon.diabetes = 'No')
			AND (patient.hignblood = 'No')
			THEN 'Yes'
			ELSE 'No'
		END AS allHealthy,
		CASE
			WHEN (servicesaddon.overweight = 'Yes')
			AND (servicesaddon.diabetes = 'Yes')
			AND (patient.hignblood = 'Yes')
			THEN 'Yes'
			ELSE 'No'
		END AS allUnHealthy
	FROM patient
	NATURAL JOIN servicesaddon
);

-- 2 Create Table for DIA Data
CREATE TABLE IF NOT EXISTS dia (
	Id INTEGER PRIMARY KEY,
```

```
	Pregnancies INTEGER NOT NULL,
	Glucose INTEGER NOT NULL,
	BloodPressure INTEGER NOT NULL,
	SkinThickness INTEGER NOT NULL,
	Insulin INTEGER NOT NULL,
	BMI NUMERIC NOT NULL,
	DiabetesPedigreeFunction NUMERIC NOT NULL,
	Age INTEGER NOT NULL,
	Outcome INTEGER NOT NULL
);

-- 3 If the table exists set make us the owner
ALTER TABLE IF EXISTS dia OWNER to postgres;

-- 4 Import DIA Data from remote location
COPY dia(
	Id,
	Pregnancies,
	Glucose,
	BloodPressure,
	SkinThickness,
	Insulin,
	BMI,
	DiabetesPedigreeFunction,
	Age,
	Outcome
) FROM PROGRAM
'curl
"https://docs.google.com/spreadsheets/d/1eoXsxpAGHLGDPXmbH5xhs1jmufM8XDfFqj4hxfh3c
OU/gviz/tq?tqx=out:csv&sheet=Healthcare-Diabetes"'
WITH (
 FORMAT csv,
 HEADER true,
 ENCODING utf8
 );

-- 5 Create View of just DIA Important Data, Also create AllHealthy and AllUnHealthy Features
CREATE VIEW dia_view AS (
	SELECT
		age,
		gender,
		diabetes,
		overweight,
		highBlood,
		CASE
			WHEN (overweight = 'No')
			AND (diabetes = 'No')
			AND (highBlood = 'No')
			THEN 'Yes'
			ELSE 'No'
		END AS allHealthy,
```

```
        CASE
            WHEN (overweight = 'Yes')
            AND (diabetes = 'Yes')
            AND (highBlood = 'Yes')
            THEN 'Yes'
            ELSE 'No'
        END AS allUnHealthy
    FROM
        (
        -- Subquery converts features to match WGU features using industry standards
        SELECT
            age,
            'Female' AS gender,
            case when outcome = 1 then 'Yes' else 'No' end as diabetes,
            case when bmi > 24.9 then 'Yes' else 'No' end as overweight,
            case when bloodpressure > 80 then 'Yes' else 'No' end as highBlood
        FROM dia
    ) AS subquery
);
```

**Part 2: Demonstration**

**B. Presentation**

https://youtu.be/RB0j06XxjT8

**Part 3: Report**

**C1. Dashboard Purpose and Function**

The purpose of our dashboards is to compare health metrics between our hospital system and our competitor's, the DIA Hospital System. The comparison of these datasets will allow us to see, where we are overachieving, and where we are falling short. We will be able to use this data to first, play to our strengths and advertise our strengths, and secondly, apply more resources to the places we are falling short. The main strength of our WGU Hospital System is the amount of older patients we have. This implies are much longer and sustained patient population when compared to our competitor. Who mostly has twenty-year-old patients. A Major strength of our system is the ability to keep patients living longer with more of them having high-quality lives - a larger population of elderly patients without high blood pressure, or diabetes and who aren't overweight.

**C2. Justification of Business Intelligence Tool**

Tableau was chosen as the business Intelligence Tool since Tableau provides easy integration with SQL Databases and easy-to-use developer interfaces. The ability to connect to a Postgres database was paramount, and the Enterprise version of Tableau has this functionality. Additionally, the flexible developer experience makes building and deploying analysis visualizations effective and efficient. The only other major Business Intelligence tool is Microsoft Power BI, which has limited data connectivity and database connection options while having tight integration with other Microsoft products, which was not required for this project (Biswal).

**C3. Data Preparation**

All data preparation took place in the SQL Database queries.

First, a view is created for the WGU data. A view is used to filter the data being sent to Tableau to only be the necessary data, previously in D210, I had all data being imported into Tableau which made development messy. This view filters the data to the Age, Gender, High Blood Pressure, Overweight, and Diabetes Features. High Blood Pressure had to be labeled differently using AS. Additionally, two new Features were made, All Healthy and All Unhealthy. This took the overweight, high blood pressure, and diabetes features, and if all values were yes or no returned yes or no for the corresponding feature. It is important to note that some of these features were on the patient table, and others on the servicesaddon table. These were joined naturally along the patient_id key, creating a larger table of the two tables combined.

Second, A table was created for the additional dataset. This was called, dia, and all values were set as either integer if discrete, or numeric if continuous. All values were set to be not null, and we were set to be the owner.

Third, we import the data from a remote location. I stored the data in a Google sheet accessed by a curl command to pull the data. This makes it easier for users and the evaluation. It is then imported into the dia table we made previously.

Finally, a view is created using the dia dataset, to be similar to the wgu view. Age, Gender, Diabetes, Overweight, and High Blood Pressure features are added to the view, and then identical All Healthy and All UnHealthy features are created. It is important to note that in the original additional dataset, features that are boolean in the wgu dataset, are provided as continuous variables in the dia dataset. For example, BMI is provided instead of an overweight boolean. To fix this, the continuous variables are converted into booleans using industry-standard conversions.

For this processing, I would also like to mention that the boolean features are left as, or converted to “Yes” and “No” as opposed to 0, 1. The reasoning behind this choice was that analysis would be limited to creating dashboards that would output “Yes” and “No” at the end. Additionally, the WGU dataset is provided as string values, if future data is added it would most likely be in the same format, so the use of “Yes” and “No” will future-proof the analysis.

**C4. Dashboard Creation**

The first step to creating the dashboards is to import the WGU and DIA Views from the database. This is completed by accessing the Data Sources tab, selecting Add Connection, and choosing Postgres. At this point, you will input credentials and then connect the database files. Each line graph is created identically. For the chosen dataset, place the record count into the “Rows” location, and the Age feature into the “Columns” Location, and set it to be a dimension to distribute the data properly. Then drop the chosen Feature into the Colors Mark to give the “Yes” and “No” lines different colors. These line graphs are then pasted into the dashboard to create a complete visualization. Pie charts are the only other visualization in these dashboards, which require “Pie” to be chosen in the Marks tool tab, the record counts as the label, with a quick calculation for percentage total, and finally the chosen Feature as the color. These will create all the provided dashboard visualizations. Text is available from the dashboard panel.

**C5. Results**

Our data analysis and visualizations help paint the main point for executive decision-making, we have a patient population that is living longer higher-quality lives compared to our competitor. Our analysis showed first off, our population age demographics are very even, we have the same amount of twenty-year-olds as eighty-year-olds. This is in stark contrast to our competitor, which has a massive amount of twenty-year-olds, and then a sharp decline in thirty-plus patients. This in of itself is a massive strength of our system and provides a metric to make resource and marketing decisions based on, however, additionally, from our All Healthy Indicators, we can that our elderly patient population also has an equal quantity of all healthy patients as younger patients. Our entire patient population has an equal amount of patients without high blood pressure or diabetes, without being overweight. This means our patients are living longer, healthier lives when compared to our competitors. This means we do not have to allocate resources as intensely to our elderly population, but rather can spread them evenly across all age groups.

**C6. Limitations**

With our analysis, the largest limitation was the differences between the WGU Hospital System dataset and the additional “DIA Hospital System” dataset. The biggest problem would be that the DIA dataset was gender-limited, only providing female records. I confirmed that the removal of male and nonbinary patients from our dataset did not affect the generalized ratio of each boolean feature for our analysis, this asserted the story and conclusions of the analysis. As an example, the ratio of High Blood Pressure Patients to Non-High Blood Pressure patients was roughly the same between the All-gender group and only the Females.

**D. Web Sources**

Van den Broeck, M. (n.d.). Intro to Tableau. Datacamp. from https://app.datacamp.com/learn/courses/introduction-to-tableau

Lacroix, H. (n.d.). Analyzing Data in Tableau. Datacamp. from https://app.datacamp.com/learn/courses/analyzing-data-in-tableau

Lacroix, H. (n.d.). Creating Dashboards in Tableau. Datacamp. from https://app.datacamp.com/learn/courses/creating-dashboards-in-tableau

Van den Broeck, M. (n.d.). Statistical Techniques in Tableau. Datacamp. from https://app.datacamp.com/learn/courses/statistical-techniques-in-tableau

Valkenburg, M. (n.d.). Visualizing Geospatial Data in Python. Datacamp. from https://app.datacamp.com/learn/courses/visualizing-geospatial-data-in-python

Holy Sheet! Remote Access CSV Files from Postgres. (2023, June 12). Crunchy Data. https://www.crunchydata.com/blog/holy-sheet-remote-access-csv-files-from-postgres

**E. Other Sources**

Biswal, A. (2023, January 30). Power BI Vs Tableau: Difference and Comparison. Simplilearn.com. https://www.simplilearn.com/tutorials/power-bi-tutorial/power-bi-vs-tableau