Data Acquisition – D205

# Bailey Psuik

## Western Governors University

### December 10, 2023

1. **Question**

For this project, data was analyzed from a popular hospital chain that serves patients across the country. The question being investigated in this project is as follows: What are the top five states that have the highest count of overweight patients? To answer this question, data from the original database and a file containing additional medical services data were explored.

**A1. Identifying Data**

The original database contains data on patient demographics and includes a table of location data. From the *location* table, the *state* column was used, which contains data of the text datatype. Additional data was provided for this analysis in the form of a file containing data on services received by patients and details on their health such as whether they are overweight, have high blood pressure, are habitual soft drink consumers, etc. From this *services* table, the *overweight* column was used, which contains data of the variable character (varchar) datatype.

A screenshot of a computer

Description automatically generated

1. **Entity Relationship Diagram (ERD)**

The figure to the left shows an entity relationship diagram (ERD) for the tables being used to answer the original question – What are the top five states that have the highest count of overweight patients? As previously stated, the *state* column from the *location* table is needed, as is the *overweight* column from the *services* table. A common column does not exist between these two tables, so it is necessary to include an intermediate table, *patient*, which allows for joining the *location* and *services* tables. To uphold referential integrity, a foreign key must be added to the services table that references the location\_id in the location table. This ensures that all records in the services table point to existing records in the location table.

**B1. Code for the ERD**

For this project, data from the additional file containing medical service data needed to be uploaded into the database with the original data. To create this new table, the SQL code below was executed:

|  |
| --- |
| CREATE TABLE services (  patient\_id text,  services varchar(20),  overweight varchar(3),  arthritis varchar(3),  diabetes varchar(3),  hyperlipidemia varchar(3),  back\_pain varchar(3),  anxiety varchar(3),  allergic\_rhinitis varchar(3),  reflux\_esophagitis varchar(3),  asthma varchar(3)  ); |

Once this new table, *services*, was created and had its columns’ datatypes defined, it then needed to have both a primary and a foreign key defined to uphold referential integrity between the tables. To add these constraints, the code below was used:

|  |
| --- |
| ALTER TABLE services ADD PRIMARY KEY (patient\_id);  ALTER TABLE services  ADD COLUMN location\_id integer;  ALTER TABLE services  ADD CONSTRAINT services\_fkey FOREIGN KEY (location\_id)  REFERENCES location(location\_id); |

**B2. Loading .csv Data**

The next step was to import the file containing the medical service data. To import this data into the PostgreSQL database through the use of pgAdmin, the built-in functionality on this platform was used. The following code shows the file path used to execute this:

|  |
| --- |
| --command " "\\copy public.services (patient\_id, services, overweight, arthritis, diabetes, hyperlipidemia, back\_pain, anxiety, allergic\_rhinitis, reflux\_esophagitis, asthma) FROM 'C:/Users/LabUser/Desktop/MSERVI~1.CSV' DELIMITER ',' CSV HEADER QUOTE '\"' ESCAPE '''';"" |

1. **SQL Query**

To answer the original question – What are the top five states that have the highest count of overweight patients? —an additional SQL query was needed. The SELECT statement calls the *overweight* column and the *state* column. Because the *overweight* column is of the varchar datatype, and a COUNT is needed, use of two CASE statements was necessary – assigning a value of 1 to all ‘Yes’ rows and aliasing them as *overweight* and *not\_overweight*, respectively. An intermediate table was needed to join the *location* and *services* table. The ERD shows that *patient\_id* is a common column between *services* and *patient*, and *location\_id* is a common column between *patient* and *location*. Two INNER JOINs were used in this query to join the three tables. The query results were grouped by state and sorted in descending order based on the count of overweight patients in each state.

|  |
| --- |
| SELECT  COUNT(CASE WHEN s.overweight = 'Yes' THEN 1 END) AS overweight,  COUNT(CASE WHEN s.overweight = 'No' THEN 1 END) AS not\_overweight,  l.state  FROM services AS s  INNER JOIN patient AS p        ON s.patient\_id = p.patient\_id  INNER JOIN location AS l        ON p.location\_id = l.location\_id  GROUP BY l.state  ORDER BY overweight DESC  ; |

**C1. .csv Files**

Upon executing the query above, the full results covering the *overweight* and *not\_overweight* counts were returned, with states with the highest *overweight* counts returned first. The full results of the query above are included in a separate file. Below are the results from the top five states, which provide an answer to the original question: Pennsylvania (PA), California (CA), Texas (TX), New York (NY), and Illinois (IL) are the top five states with the highest count of overweight patients.

**A screenshot of a computer

Description automatically generated**

1. **Time Period**

The data being used in this analysis needs to remain relevant to the hospitals using it to mitigate readmission penalties. To accomplish this, the medical services file that was added to the original database should be acquired and refreshed quarterly. This will provide the decision-makers at the hospitals with updated data four times per year to help them identify potential trends and estimates for their facilities in the future.

**D1. Explanation of Time Period**

Quarterly updates to this data will be sufficient for keeping medical services data relevant. Acquiring and refreshing this data more frequently, such as at a monthly or weekly cadence, would be costly and time-consuming. Acquiring and refreshing this data less frequently, such as annually or biannually, would not provide sufficient insight into emerging trends in patient data.

1. **Panopto Video of Code**

A full video explaining the code used is available as a separate file.

**E1. Panopto Video of Programs**

A full video explaining the code used is available as a separate file.

1. **Web Sources**

To complete this project, data from the provided database and the add-on file on medical data were used. No additional web sources were used to acquire data or segments of third-party code.

1. **Sources**

Data from the provided database and add-on file on medical data were used to answer the original question this project was asking. No additional sources were used to acquire data or segments of third-party code.