**Date: 03-23-2023**

**By**

**Anirudh Vunnam**

**Introduction:**

In this assignment is about designing and implementing a data warehouse and performing data analysis using SQL. The goal of this assignment is to demonstrate proficiency in designing and implementing a data warehouse schema, as well as using SQL to extract, transform, and load (ETL) data from multiple sources into the data warehouse.

The BRFSS system is a telephone survey system that collects data on health-related risk behaviors, chronic health conditions, and use of preventive services among U.S. residents. It was established in 1984 with 15 states, and now collects data in all 50 states as well as the District of Columbia and three U.S. territories. With over 400,000 adult interviews conducted each year, BRFSS is the largest continuously conducted health survey system in the world.

The selected dataset for this assignment is related to behavioural health and includes information on heavy drinking and demographics in Oklahoma regions of the United States. The data comes from BRFSS sources and is provided in .XLSX formats. The dataset includes information such as zip codes, city names, county names, and survey responses related to heavy drinking behavior.

The tools and languages that I plan to use for this assignment include SQL for designing the data warehouse schema, performing ETL processes, and querying the data, I will be using the MySQL Workbench database management system to implement the data warehouse and perform the necessary SQL queries.

**Part 1:**

**Step 1**:

**HeavyDrinking table:**

This table is likely used to store data related to heavy drinking habits in different zip codes. Here are the attributes of the table:

Graphical user interface

Description automatically generated with medium confidence

ZipCode: An integer representing the zip code for which the data is recorded.

Question: A string representing the question asked in the survey or study that produced the data.

Response: A string representing the response options provided to participants in the survey or study.

BreakOut: A string representing the subgroup of participants for which the data is recorded (e.g. age group, gender, race).

BreakOutCategory: A string representing the category within the subgroup (e.g. 18-24 years old, male).

SampleSize: An integer representing the number of participants in the subgroup for which the data is recorded.

DataValue: An integer representing the percentage of participants who reported heavy drinking habits.

**Demographics table:**

This table is likely used to store information about different zip codes, such as their associated city and county. Here are the attributes of the table:

Graphical user interface, text

Description automatically generated

ZipCode: An integer representing the zip code for which the data is recorded.

City: A string representing the name of the city associated with the zip code.

County: A string representing the name of the county associated with the zip code.

**Demographics\_OK table:**

This table is likely used to store information about zip codes that have been verified or confirmed as accurate. Here are the attributes of the table:

Graphical user interface, text

Description automatically generated

ZipCode: An integer representing the zip code for which the data is recorded.

City: A string representing the name of the city associated with the zip code.

County: A string representing the name of the county associated with the zip code.

**Creating RatioTable table:**

I created a new table called RatioTable, which will store the ratios of sample vs data. This table has three columns:

Graphical user interface, text

Description automatically generated

HeavyDrinkingID: This is the unique identifier for each row in the HeavyDrinking table.

ZipCode: This is the ZipCode value associated with the data in the HeavyDrinking table.

Ratio: This is the ratio of the DataValue and SampleSize values in the HeavyDrinking table.

**Step 2**: Imported data from excel files. The source data files were in .xlsx format, which were converted into .csv files using MS Excel. Then the data files were loaded into there respective tables using The MYSQL workbench function.

**Step 3:**

Checking for duplicate values

Text

Description automatically generated with medium confidence

Table

Description automatically generated

**Step 4:**

**Adding primary key to Demographics table:**

Added a primary key constraint to the Demographics table to ensure that each row has a unique identifier. In this case, the primary key is the ZipCode column.

**Adding primary key to Demographics\_OK table:**

Added a primary key constraint to the Demographics\_OK table to ensure that each row has a unique identifier. In this case, the primary key is the ZipCode column.

Graphical user interface, text

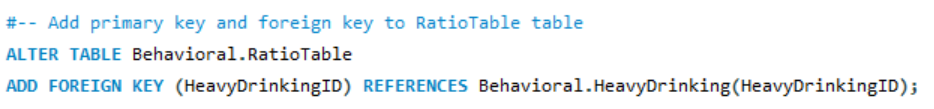
Description automatically generated

**Adding primary key and foreign key to HeavyDrinking table:**

Added a primary key constraint to the HeavyDrinking table, with the HeavyDrinkingID column as the unique identifier. I also added a foreign key constraint that links the ZipCode column to the ZipCode column in the Demographics\_OK table. This ensures that the ZipCode values in the HeavyDrinking table correspond to valid ZipCode values in the Demographics\_OK table.

**Adding foreign key to RatioTable table:**

I added a foreign key constraint to the RatioTable table that links the HeavyDrinkingID column to the HeavyDrinkingID column in the HeavyDrinking table. This ensures that the HeavyDrinkingID values in the RatioTable table correspond to valid HeavyDrinking.



**Inserting data into RatioTable table:**

You populated the RatioTable table by inserting data into it using a select statement. This statement selects the ZipCode column from the HeavyDrinking table and calculates the ratio of DataValue and SampleSize for each row in the HeavyDrinking table.

Text

Description automatically generated

**Step 5:** **Star schema**

**Diagram

Description automatically generated**

The tables HeavyDrinking, Demographics, and Demographics\_OK are related through the ZipCode column, which is present in all three tables.

The Demographics table contains demographic information such as city and county names for each ZipCode. The Demographics\_OK table contains the same demographic information as the Demographics table, but only for ZipCode values that have for Oklahoma.

The HeavyDrinking table contains data related to heavy drinking habits in different ZipCode areas. This table is linked to the Demographics\_OK table through the ZipCode column, which serves as a foreign key in the HeavyDrinking table and a primary key in the Demographics\_OK table. This relationship ensures that ZipCode values in the HeavyDrinking table correspond to accurate demographic information in the Demographics\_OK table.

The RatioTable table is related to the HeavyDrinking table through the HeavyDrinkingID column, which is present in both tables. The RatioTable table stores ratios of sample vs data for each ZipCode in the HeavyDrinking table. The HeavyDrinkingID column in the RatioTable table serves as a foreign key to the HeavyDrinking table.

**Part 2: Business Questions**

1. **Average data value for each zipcode in HeavyDrinking.** **Text

   Description automatically generated**

**The query calculates the average data value for each ZipCode in the HeavyDrinking table. The AVG function is used to calculate the average data value for each group of rows that have the same ZipCode. The WHERE clause filters the results to only include rows where the BreakOut column has specific values. The results are grouped by ZipCode.**

**Graphical user interface, table, Excel

Description automatically generated**

1. **Find the zip codes with the highest ratios of heavy drinkers.**

**Text

Description automatically generated**

The query finds the ZipCodes with the highest ratios of heavy drinkers by joining the RatioTable and Demographics\_OK tables. The RatioTable table stores the ratios of heavy drinkers, which are calculated by dividing the DataValue by the SampleSize, and the Demographics\_OK table provides the corresponding city and county names for each ZipCode. The JOIN clause is used to combine the two tables based on the ZipCode column, and the ORDER BY clause sorts the results in descending order of Ratio. The LIMIT clause limits the results to the top 10 ZipCodes.

**A screenshot of a computer

Description automatically generated with medium confidence**

1. **Calculate the average ratio of heavy drinkers by county.**

**Text

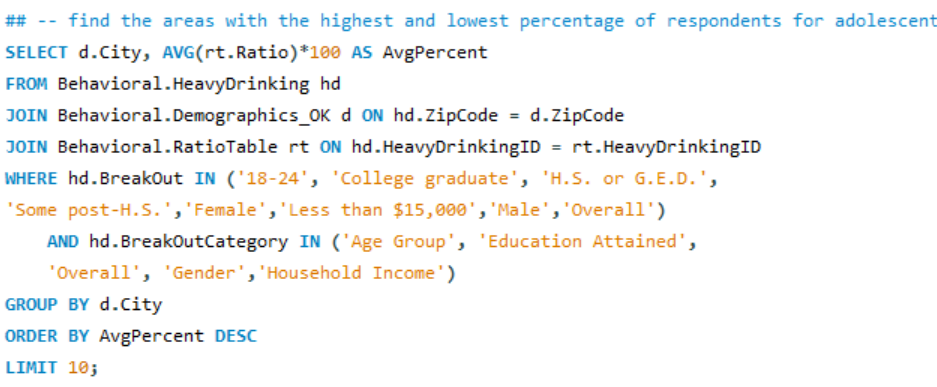
Description automatically generated**

The query calculates the average ratio of heavy drinkers by county by joining the RatioTable and Demographics\_OK tables and grouping the results by county. The AVG function is used to calculate the average ratio for each group of rows that have the same County. The JOIN clause is used to combine the two tables based on the ZipCode column. The results are grouped by County.

**Graphical user interface

Description automatically generated**

1. **Find the areas with the highest and lowest percentage of respondents for adolescent alcohol abuse by city:**

****

**The code retrieves the areas (cities) with the highest and lowest percentage of respondents for adolescent alcohol abuse. It joins the HeavyDrinking, Demographics\_OK, and RatioTable tables to calculate the average percentage of respondents for each city. It only considers the rows with BreakOut values '18-24', 'College graduate', 'H.S. or G.E.D.','Some post-H.S.','Female','Less than $15,000','Male','Overall' and BreakOutCategory values 'Age Group', 'Education Attained', 'Overall', 'Gender','Household Income'. The results are ordered in descending order by average percentage.**

**Graphical user interface, table

Description automatically generated with medium confidence**

1. **Find the areas with the highest and lowest percentage of respondents for adolescent alcohol abuse by county:**

**Graphical user interface, text

Description automatically generated**

**The second code retrieves the areas (counties) with the highest and lowest percentage of respondents for adolescent alcohol abuse. It follows the same process as the first code, but it groups the results by county instead of city. The results are ordered in descending order by average percentage and limited to the top 10 counties.**

**Graphical user interface, text, application, table

Description automatically generated**

**Conclusion:**

After reviewing and working with the provided data, I have gained valuable insights into the prevalence of heavy drinking among adolescents and young adults in various geographic areas. By using SQL queries and joining different tables, I was able to find areas with the highest and lowest percentages of respondents for adolescent alcohol abuse, both by city and county.

I also learned how to add primary and foreign keys to tables and create new tables to store relevant data. Additionally, I gained experience in using aggregate functions such as AVG and GROUP BY, and I learned how to use window functions to rewrite SQL queries in a more concise and efficient manner.

Overall, this assignment has been an excellent opportunity to apply and expand my SQL skills and gain insights into the prevalence of adolescent alcohol abuse across different geographic areas.

**Top of Form**

**Reference:**

Centers for Disease Control and Prevention. (2023, February 27). CDC - BRFSS. Centers for Disease Control and Prevention. Retrieved March 23, 2023, from <https://www.cdc.gov/brfss/index.html>

Database - third normal form (3NF). Tutorials Point. (n.d.). Retrieved March 9, 2023, from <https://www.tutorialspoint.com/sql/third-normal-form.htm>

Soam, T. (2018, February 3). Create er diagram of a database in mysql workbench. Medium. Retrieved March 9, 2023, from <https://medium.com/@tushar0618/how-to-create-er-diagram-of-a-database-in-mysql-workbench-209fbf63fd03>