

**By**

**Sunil Raj Thota  
ALY6030 - Data Warehousing and SQL**

**Class Name: Fall 2021 CPS Quarter**

**Class Number: 202215**

**CRN: 70526**

**Week 4 Assignment 4**

**thota.su@northeastern.edu**

**Introduction:**

In this Assignment, I have proposed a study that focuses on the use of alcohol and alcohol-related problems among teenagers and college-aged individuals in Oklahoma. The data records were obtained from the Behavioral Risk Factor Surveillance System (BRFSS), an institution that conducts telephone interviews to assess wellness behaviors and serious health issues among Individuals in the united states. Around 20% of young adults fit the requirements for having an alcohol intake disorder. The information given was gathered in 2009 for the state of Oklahoma. The purpose of this study was to see how much alcohol was consumed on Oklahoma campuses. To learn more about how alcohol affects youth. To figure out which age groups are influenced by alcohol. To find out which gender is more influenced by drinking. We must create SQL queries as a phase of the analysis to estimate the number of teenagers who drink excessively as well as the region in which they live. For this report, I have SQL as the language to interprets the results by creating a database schema and defining the table attributes with their data types.

**Analysis:**

Let us use the “brfss” database schema throughout our analysis. This command enables the database to be used further.

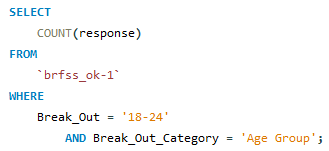
****

**Step 1: Examine the survey data ("Y") related to teenage alcohol intake.**

The analysis was focused on a BRFSS study of alcohol misuse among males in the state of Oklahoma. ‘Question', ‘Response', ‘Break out', ‘Break out the category', ‘Sample size', ‘Data-value', and ‘Zipcode' are the parameters in the information gathering. The characteristics of the people who were polled for this study are defined by the parameter 'Break out category.' 'Overall,' 'Race/Ethnicity,' 'Gender,' 'Household Income,' 'Education Attained,' and 'Age Group' are the metrics factors shown in the attribute. The property ‘Break out' gives the result associated with the variable's type.

The sampling size refers to the number of factors that were asked within every type of demographics survey. The Zipcode identifies the geographic location of the poll. The parameter 'data value' counts the set of surveyed people who completed yes to the question about alcohol misuse. The ‘YES' response variable indicates that the people were alcohol dependent. To determine the number of surveys done for the teenage age categories, I ran a simple SQL query that returned the number of respondents.

**Code:**



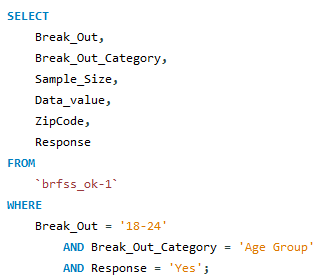
**Result:**

****

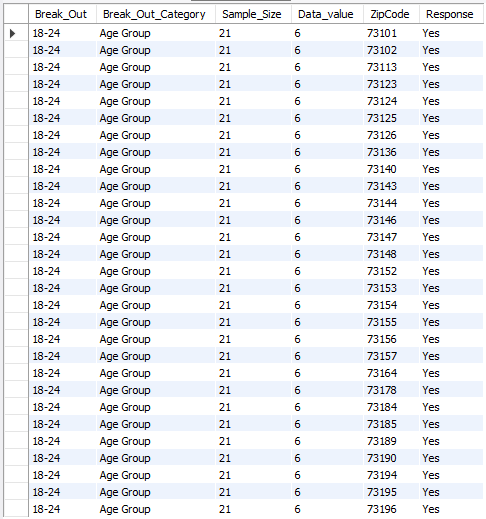
We can see from the results directly 29 evaluations for adolescent groups been undertaken in diverse geographic locations. Because we're seeking for the age group of teens, the 'break out category' parameter from the given dataset presented can be used, resulting in "Age Group" as the break out category.

**Step 2: 29 Samples who responded as ‘Yes’**

**Code:**

****

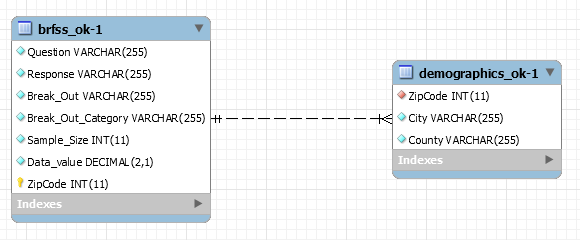
**Result:**

****

The samples of the acquired result are shown in the picture on top. We see this when a query for alcohol dependence in the adolescent population was run in Oklahoma, 29 entries were given as result. The 21 people who were evaluated for this study were picked at random via 21 distinct zip codes in 21 diverse regional places. The 'data value' was consistent at 6 in all situations, implying that 6 out of 21 people were victims of alcohol consumption in general. It also implies that now in Oklahoma, one in every three teenagers aged 18 to 24 was drinking excessively.

The next step is to determine the subgroup of teenagers who are much more inclined to produce an alcohol-related issue. I have proposed a few ways for you to consider, this might aid mostly in quick recognition of these people as well as raising people's awareness well about the issue. - The addict's family's annual salary must be documented and assessed because a larger income allows for increased spending. - For certain age categories, investigators must create a model depending on city-wise classification. They'd be ready to pinpoint the exact site of the issue if they did it. - We could also categorize them by racial and ethnic groups to see if there are any patterns in teenage alcohol issues basis on race.

**Step 3: ERD Diagram**

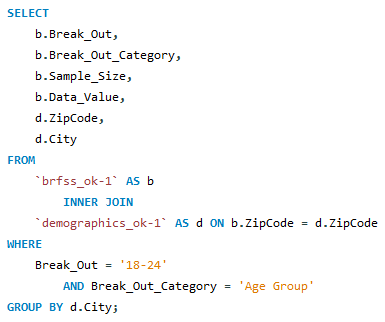


The 'ZipCode' is the primary key in the 'brfss table,' which also works as a foreign key in the 'demographics' table, as seen in the ERD diagram above. The identities of the parameters and the types of data relating to every parameter are also displayed in the tables. The 'INT' type is used to hold elements with only numeric data. The ‘VARCHAR' form is used to hold variables that include both character and numeric values. To connect two tables, one-to-many links are employed.

According to the results, out of the 21 people surveyed, every one of the regions in Oklahoma exhibited 6 positive replies. As a result, neither of the places can be deemed to have the highest or lowest majority of individuals who have an alcohol problem. The SQL queries that identify the city and county of teenagers who've been abusing alcohol are listed here. I utilized joins to have the result because we needed to combine the two tables for such inquiries.

**Step 4: A SQL Query to fetch the areas by the City of Oklahoma State which have the highest and lowest number of respondents for adolescent alcohol abuse.**

**Code:**

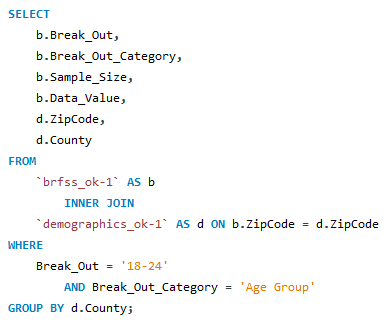


**Result:**

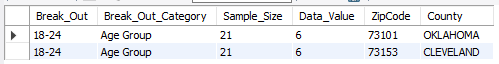


**Step 5: A SQL Query to fetch the areas by County of Oklahoma State which have the highest and lowest number of respondents for adolescent alcohol abuse.**

**Code:**



**Result:**

****

**Conclusion:**

I'm glad I chose this data since it enabled me to try out several techniques for generating tables and loading entries, like joins and setting Primary Keys, Foreign Keys, and ERDs. The lowest and highest cities and counties were found as an outcome of the abovementioned research. Also, Integrating sampling techniques besides surveys. The study might've been funded with sufficient funds. To encompass a wide expanse, the study may have employed a substantial number of people. The inner joins were used to define these elements in the tables. I'd like to experiment with more complex datasets and data warehousing solutions employing sophisticated star schema designs.