Behavioral Risk Factor Surveillance

Abstract

- The "Behavioral Risk Factor Surveillance System (BRFSS)" dataset is a compilation of survey information on people's health and well being in the United States. This dataset is a precious resource for people researching the topic, and public health officials, who are passionate about understanding the behavioral health and factors of risk of the United States population. The Centers for Disease Control and Prevention (CDC) conducted phone conversations with people from different US states to gather the data. A variety of subjects including health practices, chronic illnesses, sleep patterns, and the utilization of preventive services are covered in the survey. The dataset contains data on demographic traits, smoking patterns, levels of physical activity, sleep cycles, and patterns. Insights show that certain demographic groups, such as Black/African American and Hispanic/Latino adults, are more likely to report poor health outcomes and behaviors such as smoking and obesity.We will utilize SQL queries to analyze data, and visualize it with Python and Tableau. This mental health topic is often taboo and it is one of the main reasons why we wanted to dive deeper into the statistics.

Motivation

- This dataset concentrates on the prevalence of health behaviors and chronic conditions by providing insights into health disparities over time. By analyzing this dataset, we hope to identify trends in health behaviors such as tobacco use, physical activity, and alcohol consumption, as well as chronic health conditions such as diabetes, hypertension, and heart disease by highlighting the effects of sleep negligence. This topic of interest caught our attention in mental health as it is often the elephant in the room and people tend to neglect it. We have come across many stories of mental breakdowns. From Influential celebrities to the general public, many have fallen into mental breakdown trauma and committed suicides. Hence one can be able to infer how important it is to value not just the physical, but also a healthy mindset/environment.

- The BRFSS dataset provides a unique opportunity to gain insights into the health behaviors and risk factors of the US population and inform evidence-based public health interventions to promote healthier lives. Over time, people have disregarded how important sleep is and have incurred a number of health problems. We intend to identify compelling causes for the decline in people's health and mentality. These findings can then be used to inform public health policies and interventions aimed at reducing the prevalence of these risk factors and improving overall health outcomes.

Methodology

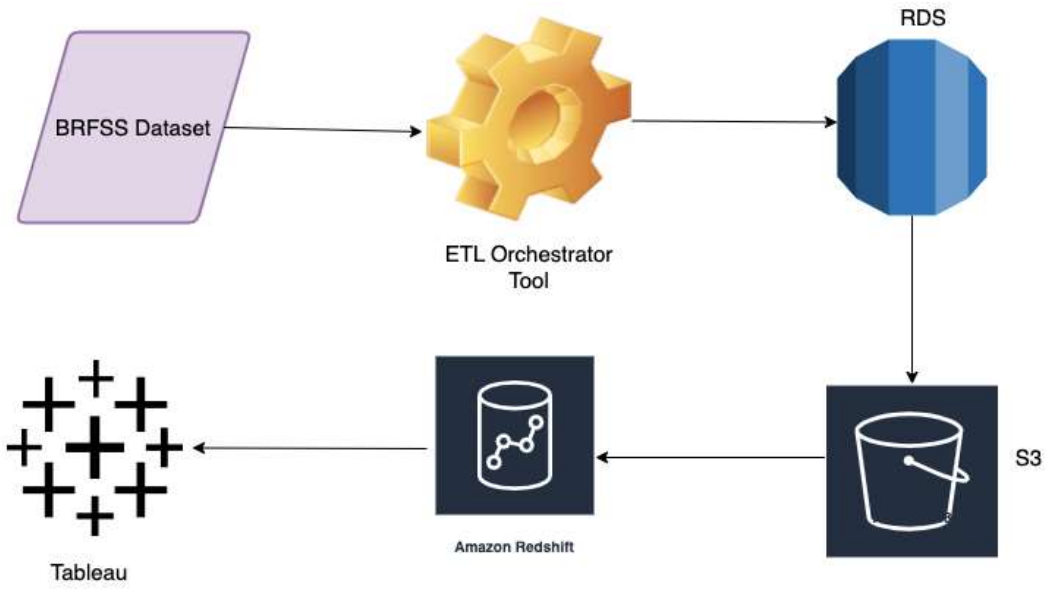
- The dataset that we propose to use is called "Behavioral Risk Factor Surveillance System(BRFSS)" and it may be found on the "kaggle.com" website. Cleaning and preprocessing the data, so that it involves resolving inaccurate data and addressing outliers, and replacing duplicate entries with null values using Python would be our first step in the data engineering process. By doing so, we would be able to create appropriate data models and move on to the next stage., which would be storing the data in AWS Cloud. The BRFSS dataset includes information from numerous surveys that were carried out over the years. With data integration, we would compile the results of numerous surveys into a single dataset using Python. Data warehousing and ETL methods will be applied to the dataset. Transforming the data, by aggregating it with a common group and calculating new variables based on existing data would be the next objective of our deliverables. Finally, we would use exploratory analysis to locate anomalies, connected patterns, and summarize them to understand the dataset. Upon doing so, we would be able to draw out newer insights and key factors for the root cause. Finally, to explore and present the insights that we have discovered through data analysis, we would utilize data visualization tools like Tableau and PowerBi and python data libraries like Matplotlib, Seaborn, etc wherever required. This would enable us to draw conclusions from our data that would be clear even to the average person in terms of the underlying reason. Finally in order to provide reports, we would use tools like Grammarly and Latex to improve the phrases and establish a standard format for language use.

DATA PREPARATION AND EXPLORATION

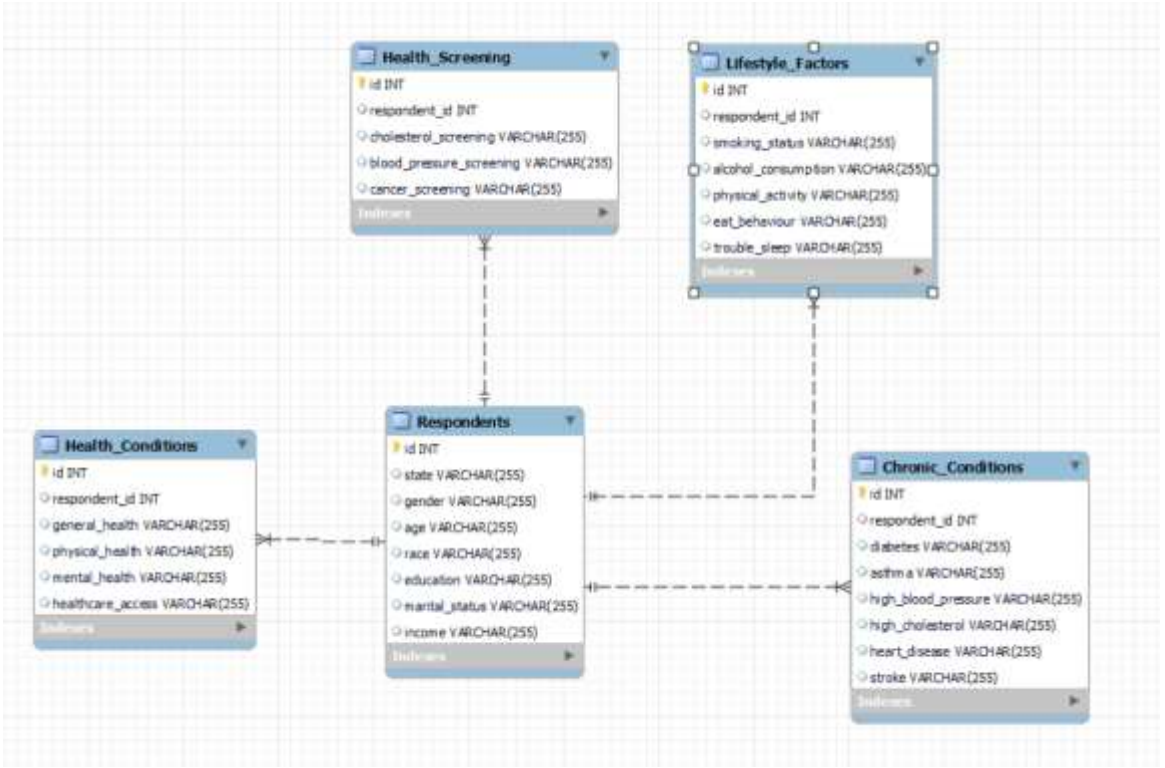
- The crash data is provided as a comma-separated document (.csv). We took the 5 csv and loaded it into Amazon S3 cloud storage, then put it through AWS Glue to create a data warehouse, where we divided the 5 csv files as a Data Model.

- For preliminary exploration we used python to understand the data structure and for removal of noisy / null data. Based on the column the null values were replaced by the measures of central tendencies such as mean, mode.
- We connected Amazon Redshift to tableau for data visualization and performed operations for meaningful insights.

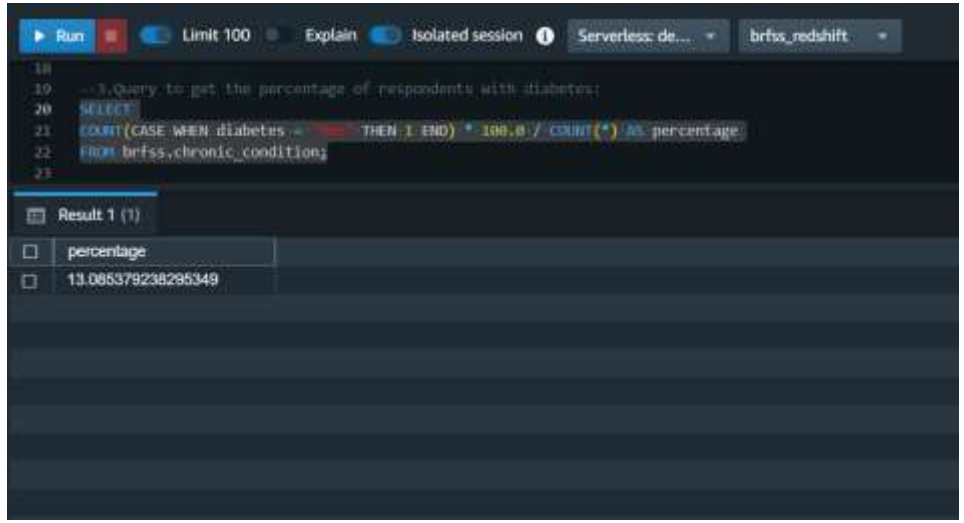
Architecture Diagram

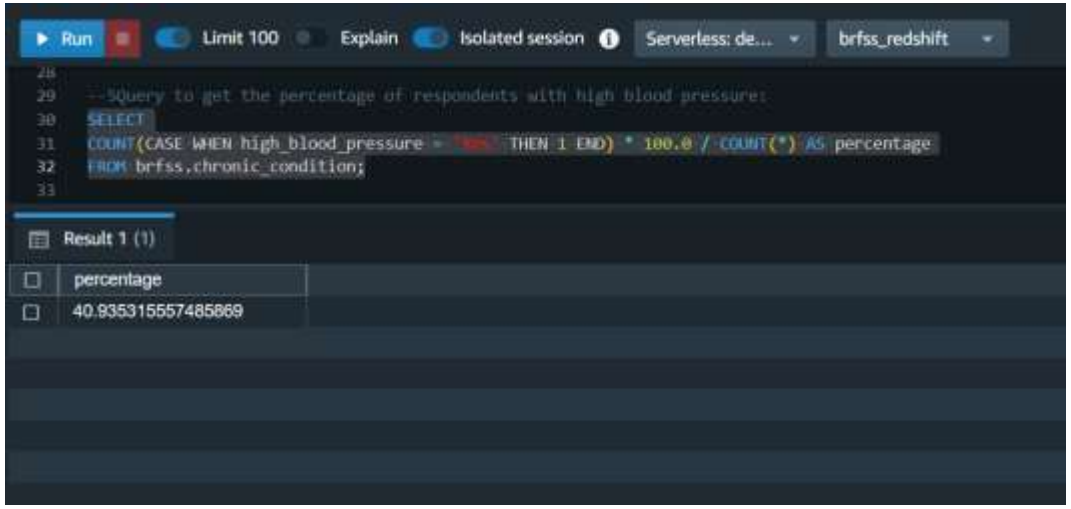


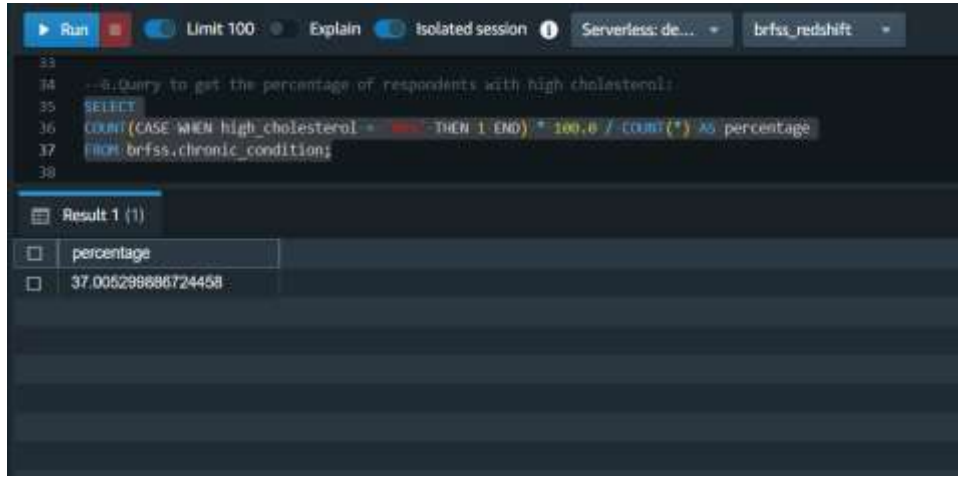
ER

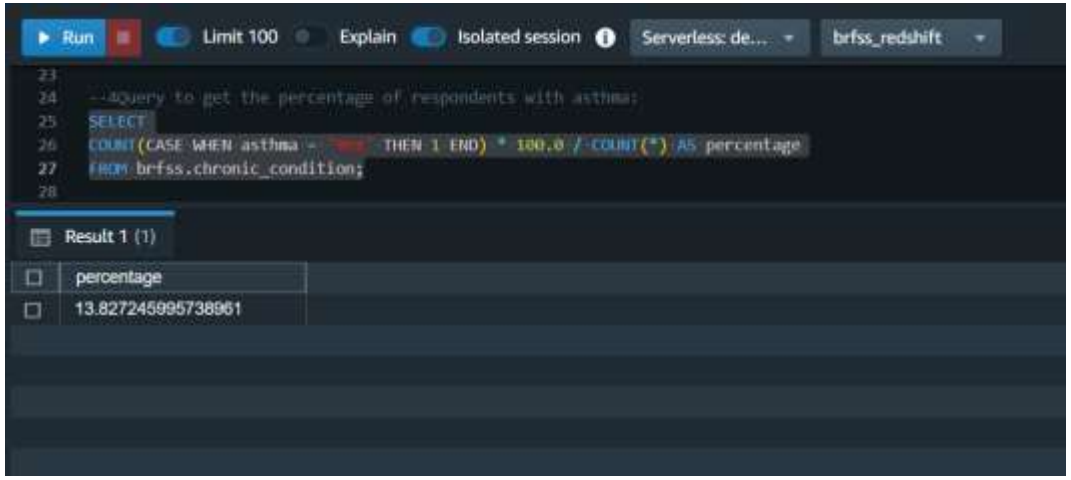

Health_Screening
id INT
respondent_id INT
cholesterol_screening VARCHAR(255)
blood_pressure_screening VARCHAR(255)
cancer_screening VARCHAR(255)
Lifestyle_Factors
id INT
respondent_id INT
smoking_status VARCHAR(255)
alcohol_consumption VARCHAR(255)
physical_activity VARCHAR(255)
eat_behaviour VARCHAR(255)
trouble_sleep VARCHAR(255)
Health_Conditions
id INT
respondent_id INT
general_health VARCHAR(255)
physical_health VARCHAR(255)
mental_health VARCHAR(255)
healthcare_access VARCHAR(255)
Respondents
id INT
state VARCHAR(255)
gender VARCHAR(255)
age VARCHAR(255)
race VARCHAR(255)
education VARCHAR(255)
marital_status VARCHAR(255)
income VARCHAR(255)
Chronic_Conditions
id INT
respondent_id INT
diabetes VARCHAR(255)
asthma VARCHAR(255)
high_blood_pressure VARCHAR(255)
high_cholesterol VARCHAR(255)
heart_disease VARCHAR(255)
stroke VARCHAR(255)

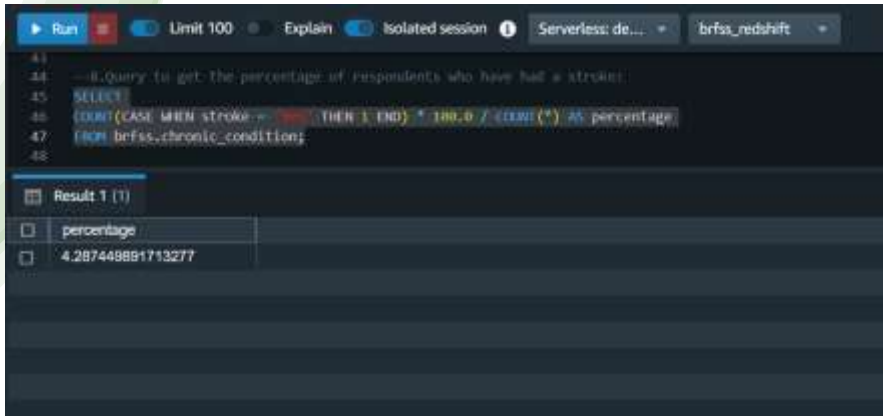
Insights

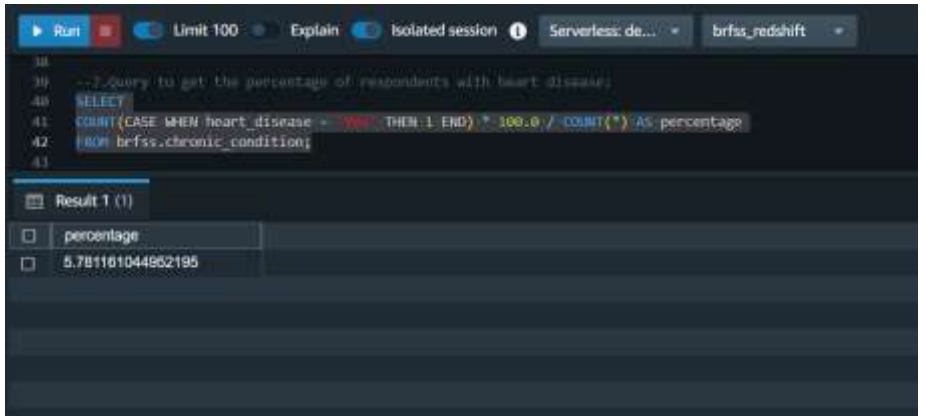

Run
Limit 100
Explain
Isolated session
Serverless: de...
brfss_redshift
--3.Query to get the percentage of respondents with diabetes:
SELECT
COUNT(CASE WHEN diabetes = 'Yes' THEN 1 END) * 100.0 / COUNT(*) AS percentage
FROM brfss.chronic_condition;
Result 1 (1)
percentage
13.085379238295349

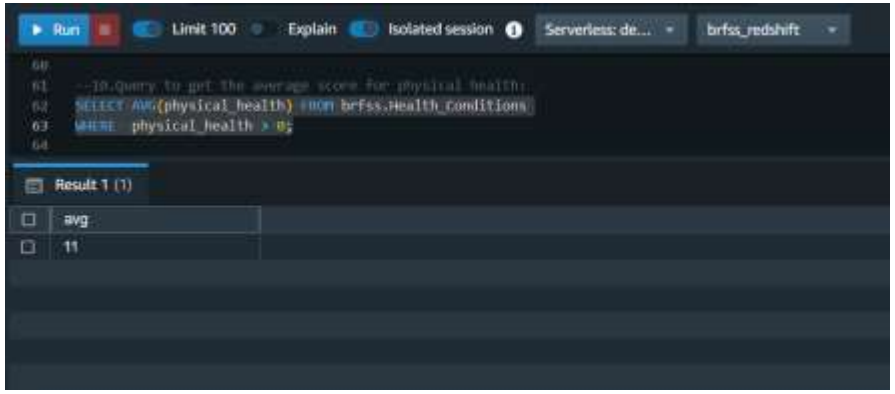

Run
Limit 100
Explain
Isolated session
Serverless: de...
brfss_redshift
--5Query to get the percentage of respondents with high blood pressure:
SELECT
COUNT(CASE WHEN high_blood_pressure = 'Yes' THEN 1 END) * 100.0 / COUNT(*) AS percentage
FROM brfss.chronic_condition;
Result 1 (1)
percentage
40.935315557485869


Run
Limit 100
Explain
Isolated session
Serverless: de...
brfss_redshift
--6.Query to get the percentage of respondents with high cholesterol:
SELECT
COUNT(CASE WHEN high_cholesterol = 'Yes' THEN 1 END) * 100.0 / COUNT(*) AS percentage
FROM brfss.chronic_condition;
Result 1 (1)
percentage
37.005298686724458


Run
Limit 100
Explain
Isolated session
Serverless: de...
brfss_redshift
--4Query to get the percentage of respondents with asthma:
SELECT
COUNT(CASE WHEN asthma = 'Yes' THEN 1 END) * 100.0 / COUNT(*) AS percentage
FROM brfss.chronic_condition;
Result 1 (1)
percentage
13.827245995738961


Run
Limit 100
Explain
Isolated session
Serverless: de...
brfss_redshift
--8.Query to get the percentage of respondents who have had a stroke:
SELECT
(COUNT(CASE WHEN stroke = '1' THEN 1 END) * 100.0 / COUNT(*)) AS percentage
FROM brfss.chronic_condition;
Result 1 (1)
percentage
4.287449881713277

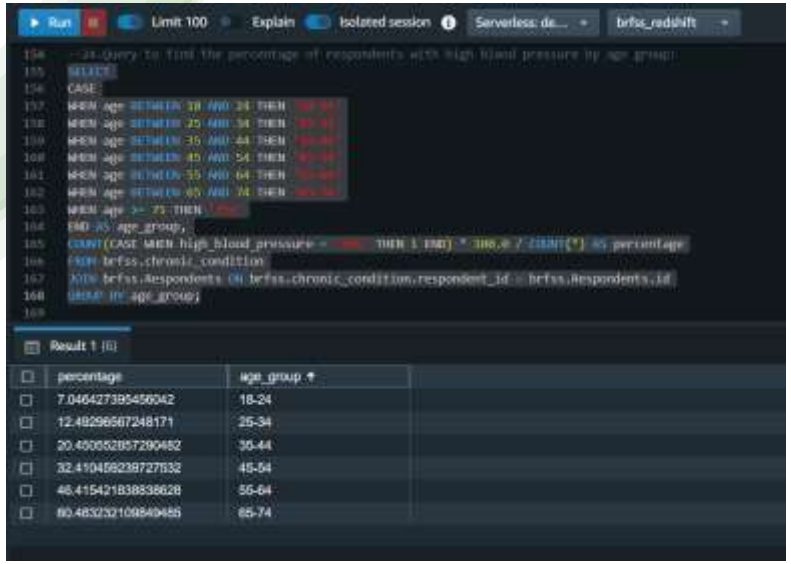

Run
Limit 100
Explain
Isolated session
Serverless: de...
brfss_redshift
--7.Query to get the percentage of respondents with heart disease:
SELECT
COUNT(CASE WHEN heart_disease = '1' THEN 1 END) * 100.0 / COUNT(*) AS percentage
FROM brfss.chronic_condition;
Result 1 (1)
percentage
5.781161044962195


Run
Limit 100
Explain
Isolated session
Serverless: de...
brfss_redshift
--10.Query to get the average score for physical health:
SELECT AVG(physical_health) FROM brfss.Health_Conditions
WHERE physical_health > 0;
Result 1 (1)
avg
11

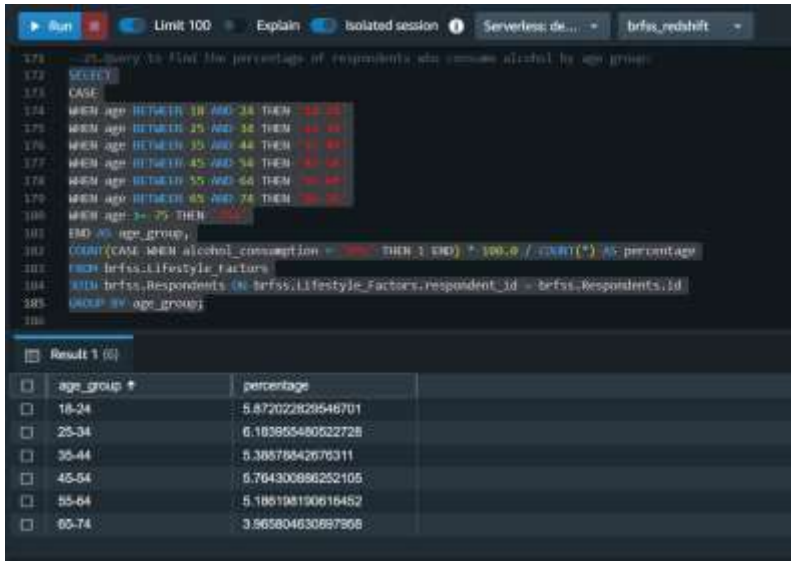

Run
Limit 100
Explain
Isolated session
Serverless: de...
brfss_redshift
--11.Query to get the average score for mental health:
SELECT AVG(mental_health) FROM brfss.Health_Conditions
WHERE mental_health>0;
Result 1 (1)
avg
10

Insights

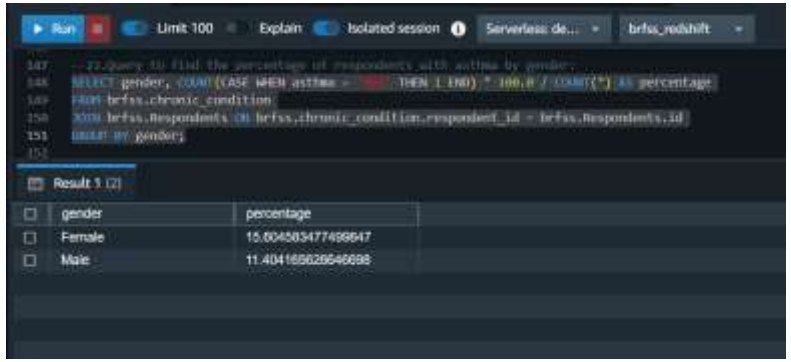
insights



percentage	age_group
7.046427385456042	18-24
12.49296567248171	25-34
20.490052857290482	35-44
32.41045923872732	45-54
48.41542183838628	55-64
60.48323210984985	65-74



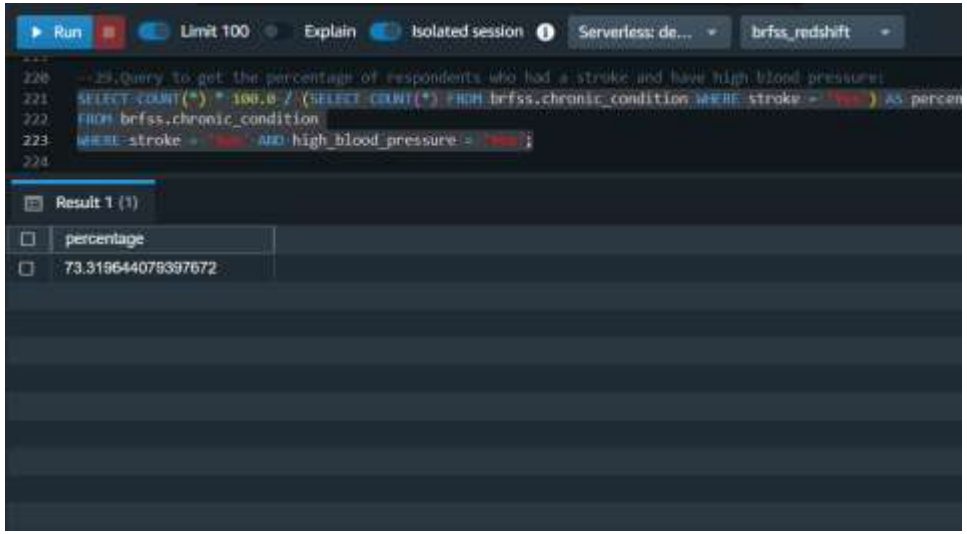
age_group	percentage
18-24	5.872022829546701
25-34	6.183955480522728
35-44	5.38578842676311
45-54	5.764300896252105
55-64	5.186198190616452
65-74	3.965804630697958

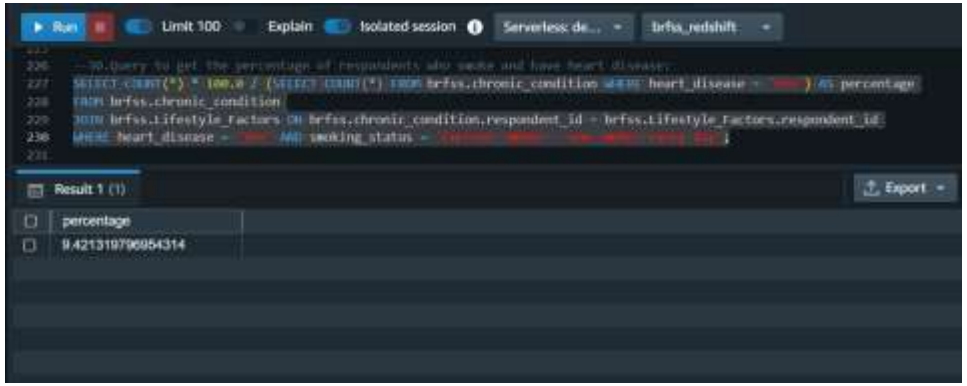


gender	percentage
Female	15.604583477498647
Male	11.404166629646698

gender	percentage
Male	13.585434173669467
Female	12.718538474487615

Insights





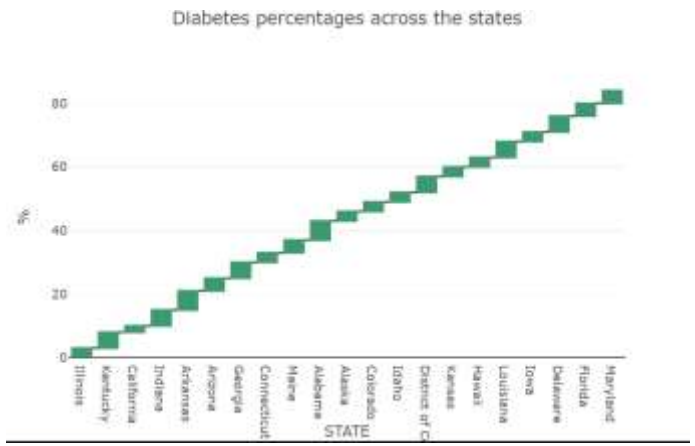
Result 1 (12)

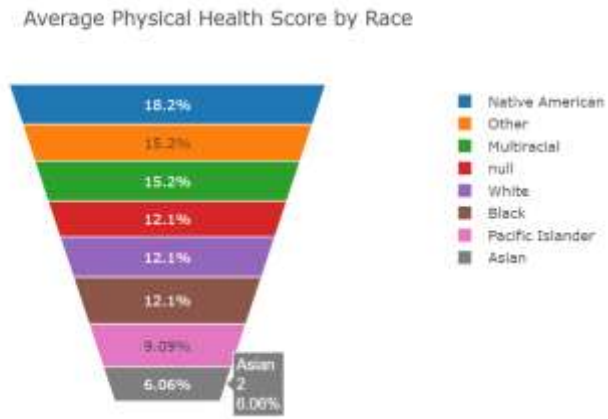
age_group	gender	percentage
18-24	Male	0
18-24	Female	0
25-34	Male	0
25-34	Female	0.011827321111768
35-44	Male	0.02348796241926
35-44	Female	0.036903773410831
45-54	Female	3.101534443356186

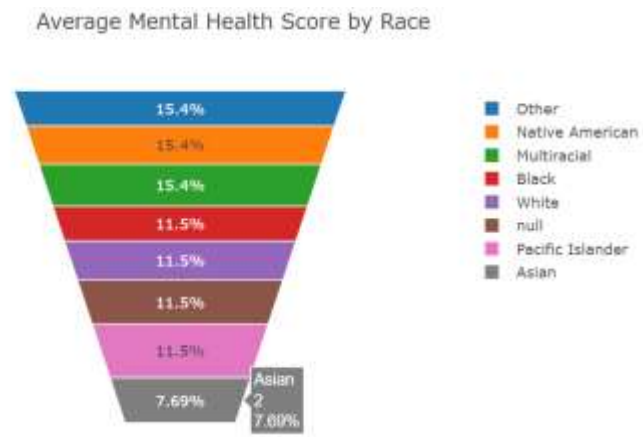
Trends based on Insights

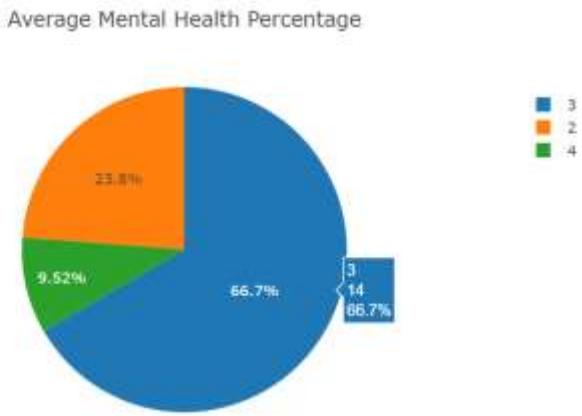
- Chronic conditions such as high blood pressure, high cholesterol, and diabetes are prevalent among respondents, with high blood pressure being the most common.
- Females have a higher prevalence of asthma, while males have a higher prevalence of diabetes.
- Age is a significant factor in the prevalence of chronic conditions, with older respondents having higher rates of high blood pressure, high cholesterol, and stroke.
- Education and marital status both influence respondents' average income.
- There is a wide variation in the distribution of respondents across states, as well as in healthcare access by state.
- Lifestyle factors such as smoking, alcohol consumption, and physical activity vary among respondents, with smoking being linked to heart disease and alcohol consumption varying by age group.

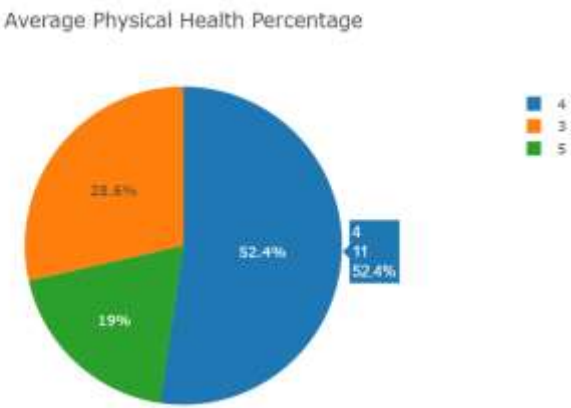
Visualizations

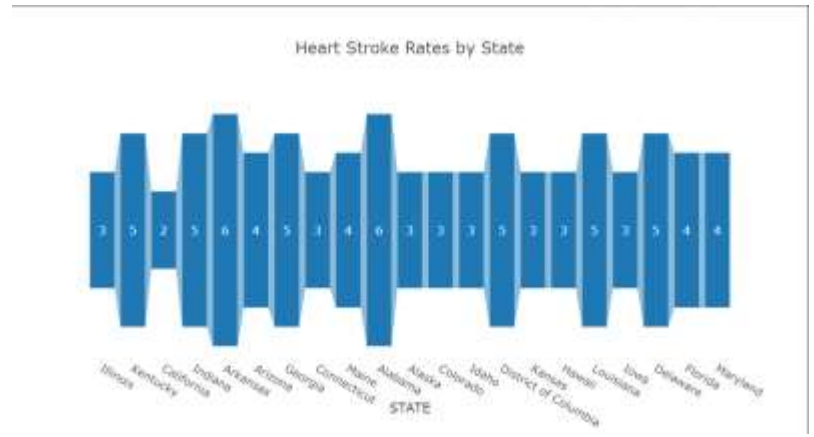












Impact on Society

- By enhancing our knowledge of behavioral health, factors of risk, and outcomes of the US population, this project has the potential to have a positive impact on society. We can spot trends and patterns in health-related behaviors and results, which can enable public health policies and programs to create focused interventions that can help people and communities live healthier lives by preventing chronic diseases.

Project Development Methodology

CONCLUSION

1. The project analyzed a diverse sample and found key trends and insights that can inform public health policies and interventions.
2. Chronic conditions such as high blood pressure, high cholesterol, and diabetes are highly prevalent among respondents, underscoring the need for effective public health measures.
3. Gender differences exist in the prevalence of certain conditions, and these differences should be taken into account in designing targeted interventions and healthcare services.
4. Age is a significant factor influencing the prevalence of chronic conditions, highlighting the need for age-specific interventions and healthcare services.
5. Education and marital status have an impact on respondents' average income, emphasizing the importance of addressing socioeconomic disparities in health outcomes and access to healthcare services.
6. The distribution of respondents across states and the differences in healthcare access by state indicate that state-specific policies and interventions may be necessary.
7. Lifestyle factors such as smoking, alcohol consumption, and physical activity play a significant role in respondents' health.
8. The findings reveal a complex interplay of factors influencing health, and public health policies and interventions should be designed to address these factors comprehensively and in a targeted manner.

THANK YOU