Behavioral Risk Factor Surveillance

MOUNIKA VEMPALLI
Master's in Data Analytics
San Jose State University
San Jose, USA
mounika.vempalli@sjsu.edu

JAYA LAKSHMI GUNJI
Master's in Data Analytics
San Jose State University
San Jose, USA
jayalakshmi.gunji@sjsu.edu

VENU ANUPATI
Master's in Data Analytics
San Jose State University
San Jose, USA
venu.anupati@sjsu.edu

SAI SARAN PRASA
Master's in Data Analytics
San Jose State University
San Jose, USA
saisaran.parasa@sjsu.edu

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.**

I. 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.

II. Literature Review

1.The study, ”Trends in Health Behaviors and Health Outcomes Among US Adults: The Behavioral Risk Factor Surveillance System 1984-2017,” was conducted by Michael R. Kramer and colleagues (American Journal of Preventive Medicine, 2020). Using BRFSS data, this study looked at long-term changes in American adult health habits and outcomes. Data on physical activity, cigarette use, alcohol use, diet, and health outcomes like obesity, diabetes, and cardiovascular disease were all examined by the authors. They discovered that while some health behaviors (such as a decline in smoking rates) have been better over time, others have gotten worse (such as a decrease in physical activity). The study emphasizes the necessity of continuing to monitor health habits and results in order to pinpoint areas that require intervention.

2.”State-Level Variations in the Prevalence of Chronic Obstructive Pulmonary Disease (COPD) in the United States: Findings from the Behavioral Risk Factor Surveillance System (BRFSS)” by Kavita Singh et al. (International Journal of Environmental Research and Public Health, 2018).The prevalence of COPD in the US varied at the state level, according to this study's analysis of BRFSS data. Data on self-reported diagnoses of COPD as well as risk variables like smoking and air pollution were examined by the authors. They discovered that there were significant regional variations in the prevalence of COPD and that risk variables including smoking and air pollution had a significant impact on this prevalence. The study emphasizes the requirement for focused initiatives to lower the prevalence of COPD, particularly in states with high-risk factor prevalence.

3.”Prevalence of Sexual Orientation Identity and Health Disparities in Adults: The Behavioral Risk Factor Surveillance System, 2013” by Brian R. Hollenbeck et al. (American Journal of Public Health, 2018). The prevalence of sexual orientation identity and related health inequalities among US adults were investigated in this study using BRFSS data. Data on self-reported sexual orientation identity as well as health outcomes like mental health and substance use were examined by the authors. They discovered that compared to heterosexual adults, sexual minorities (including homosexual, lesbian, and bisexual people) had higher rates of mental health and drug use issues. This emphasizes the need for focused treatments to lessen health disparities in this community.

4.Bingrui Xie's cross-sectional study, "Association Between Sleep Duration and Cognitive Impairment: A Cross-Sectional Study Using the BRFSS Database," investigates the connection between sleep duration and cognitive impairment. The Behavioral Risk Factor Surveillance System (BRFSS) database, a sizable survey of the adult population in the United States, serves as the study's data source. For this investigation, a sample of almost 400,000 persons aged 45 and above was used. The study's key finding was that sleep duration and cognitive impairment had a U-shaped connection, meaning that both short and long sleep durations were linked to an increased risk of cognitive impairment. The study indicated that those who slept 7-8 hours per night were less likely to develop cognitive impairment than those who slept less than 5 hours or more than 9 hours per night. Overall, the research indicates that preserving cognitive function in middle-aged and older persons may depend on getting the recommended amount of sleep (7-8 hours each night).

III. Methodology

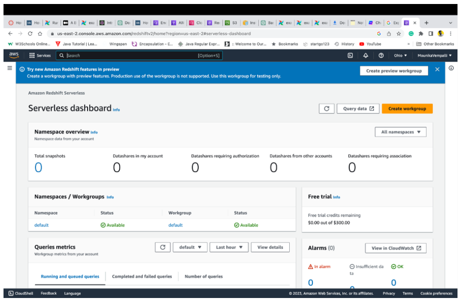
The dataset that we propose to use is called "Behavioral Risk Factor Surveillance System(BRFSS)" and it may be found on the "kaggle.com" website. Cleaning and preprocessing the data, so that it involves resolving inaccurate data and addressing outliers, and replacing duplicate entries with null values 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.

IV. 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.


us-east-2.console.aws.amazon.com/redshiftv2/home?region=us-east-2#serverless-dashboard

Try new Amazon Redshift features in preview
Create a workgroup with preview features. Production use of the workgroup is not supported. Use this workgroup for testing only.

Amazon Redshift Serverless

Serverless dashboard Info

Namespace overview Info
Namespace data from your account

Total snapshots	Datashares in my account	Datashares requiring authorization	Datashares from other accounts	Datashares requiring association
0	0	0	0	0

Namespaces / Workgroups Info

Namespace	Status	Workgroup	Status
default	Available	default	Available

Free trial Info
Free trial credits remaining
$0.00 out of $300.00

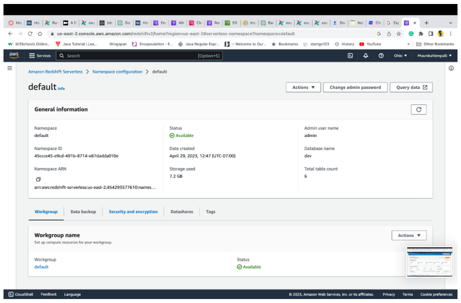
Queries metrics
Workgroup metrics from your account

Alarms (0)

In alarm	Insufficient data	OK
0	0	0

© 2023, Amazon Web Services, Inc. or its affiliates.

Fig. 1. Amazon Redshift Serverless Cluster


us-east-2.console.aws.amazon.com/redshiftv2/home?region=us-east-2#serverless-namespace?namespace=default

Amazon Redshift Serverless > Namespace configuration > default

default Info

General information

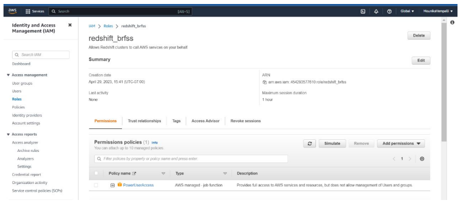
Namespace default	Status Available	Admin user name admin
Namespace ID 45ccce45-e9cd-491b-8714-e87dadda010e	Date created April 29, 2023, 12:47 (UTC-07:00)	Database name dev
Namespace ARN arn:aws:redshift-serverless:us-east-2:454293577610:names...	Storage used 7.2 GB	Total table count 6

Workgroup name
Set up compute resources for your workgroup.

Workgroup	Status
default	Available

© 2023, Amazon Web Services, Inc. or its affiliates.

Fig. 2. Amazon Redshift Serverless Cluster


IAM > Roles > redshift_brfss

redshift_brfss
Allows Redshift clusters to call AWS services on your behalf

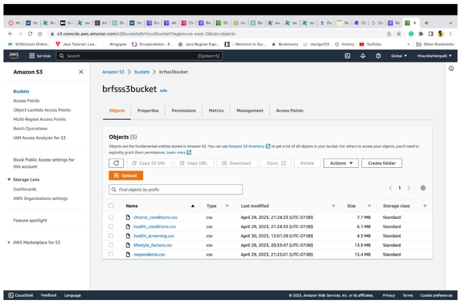
Summary

Creation date April 29, 2023, 15:41 (UTC-07:00)	ARN arn:aws:iam::454293577610:role/redshift_brfss
Last activity None	Maximum session duration 1 hour

Permissions policies (1) Info
You can attach up to 10 managed policies.

Policy name	Type	Description
PowerUserAccess	AWS managed - job function	Provides full access to AWS services and resources, but does not allow management of Users and groups.

Fig. 3. Creating roles


s3.console.aws.amazon.com/s3/buckets/brfsss3bucket?region=us-east-2&tab=objects

Amazon S3 > Buckets > brfsss3bucket

brfsss3bucket Info

Objects (5)
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Name	Type	Last modified	Size	Storage class
chronic_conditions.csv	csv	April 29, 2023, 21:24:23 (UTC-07:00)	7.7 MB	Standard
health_conditions.csv	csv	April 29, 2023, 21:24:33 (UTC-07:00)	6.1 MB	Standard
health_screening.csv	csv	April 30, 2023, 12:01:28 (UTC-07:00)	4.3 MB	Standard
lifestyle_factors.csv	csv	April 29, 2023, 20:33:47 (UTC-07:00)	13.9 MB	Standard
respondents.csv	csv	April 29, 2023, 21:25:01 (UTC-07:00)	15.4 MB	Standard

© 2023, Amazon Web Services, Inc. or its affiliates.

Fig. 4. Uploaded Extarcted to S3 Bucket

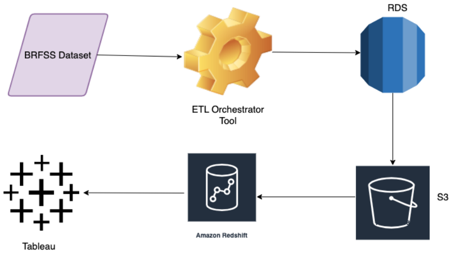
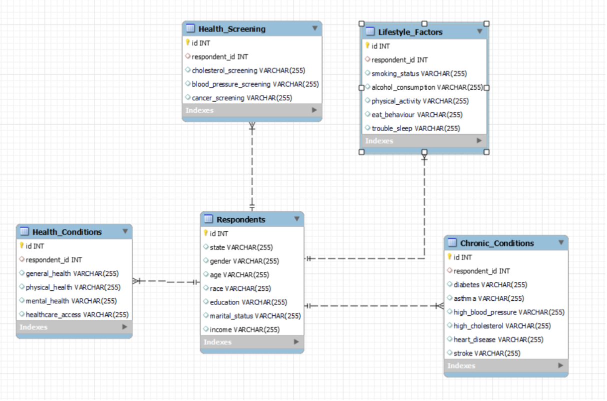


Fig. 5. ARCHITECTURE DIAGRAM

V. ER DIAGRAM

We created the following ER diagram using the reverse engineering method in MySQL Workbench.



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)

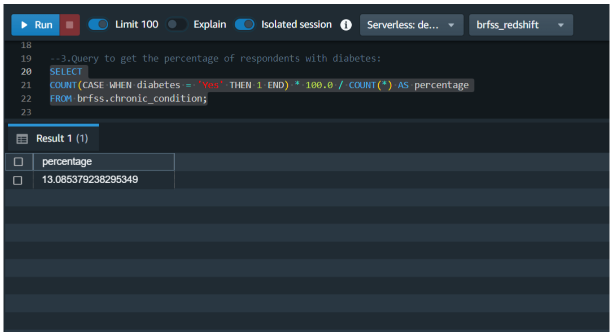
Health_Conditions
- id INT
- respondent_id INT
- general_health VARCHAR(255)
- physical_health VARCHAR(255)
- mental_health VARCHAR(255)
- healthcare_access VARCHAR(255)

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)

Fig. 6. ER Diagram

VI. INSIGHTS



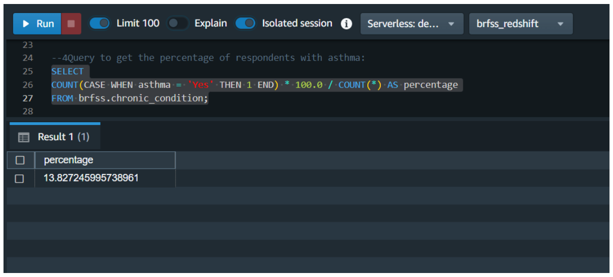
```
--3.Query to get the percentage of respondents with diabetes:
SELECT
COUNT(CASE WHEN diabetes = 'Yes' THEN 1 END) * 100.0 / COUNT(*) AS percentage
FROM brfss.chronic_condition;
```

```
percentage
13.085379238295349
```

Fig. 7. Query to get the percentage of respondents with diabetes

We have done as many queries as possible to get more insights and to understand the data better .From all the querying and analysis that we have done, we found out the following observations.

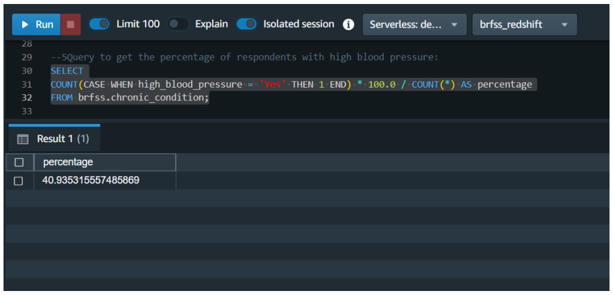
• California, Colorado, and Kansas have the highest number of respondents, while Alaska, Arkansas, and the District of Columbia have the lowest number of respondents among the states.



```
--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;
```

```
percentage
13.827245995738961
```

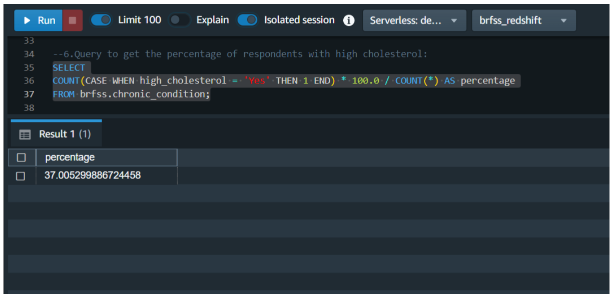
Fig. 8. Query to get the percentage of respondents with asthma



```
--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;
```

```
percentage
40.935315557485869
```

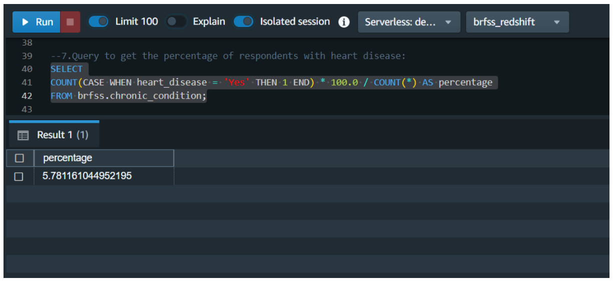
Fig. 9. Query to get the percentage of respondents with high blood pressure



```
--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;
```

```
percentage
37.005299886724458
```

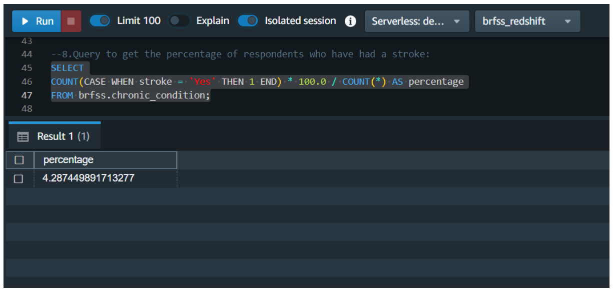
Fig. 10. Query to get the percentage of respondents with high cholesterol



```
--7.Query to get the percentage of respondents with heart disease:
SELECT
COUNT(CASE WHEN heart_disease = 'Yes' THEN 1 END) * 100.0 / COUNT(*) AS percentage
FROM brfss.chronic_condition;
```

```
percentage
5.781161044952195
```

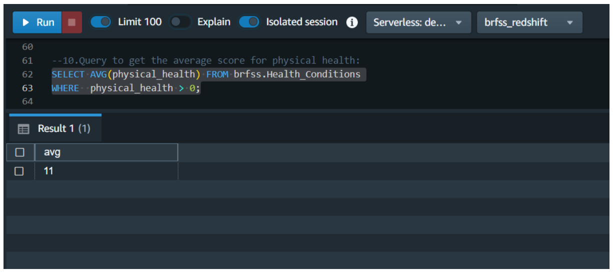
Fig. 11. Query to get the percentage of respondents with heart disease



```
--8.Query to get the percentage of respondents who have had a stroke:
SELECT
COUNT(CASE WHEN stroke = 'Yes' THEN 1 END) * 100.0 / COUNT(*) AS percentage
FROM brfss.chronic_condition;
```

□	percentage
□	4.287449891713277

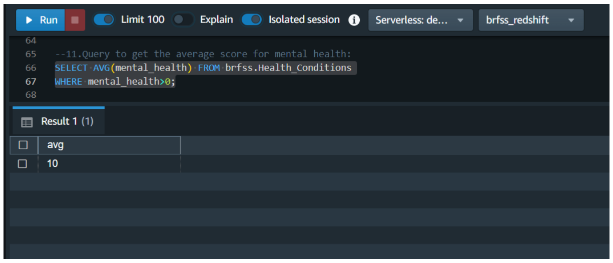
Fig. 12. Query to get the percentage of respondents who have had a stroke



```
--10.Query to get the average score for physical health:
SELECT AVG(physical_health) FROM brfss.Health_Conditions
WHERE  physical_health > 0;
```

□	avg
□	11

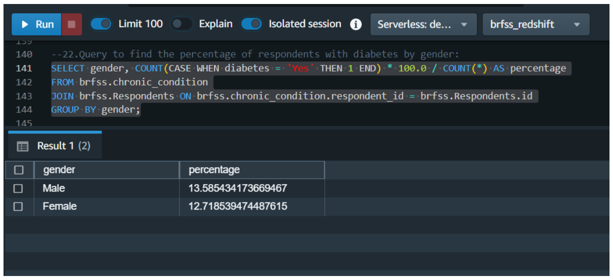
Fig. 13. Query to get the average score for physical health



```
--11.Query to get the average score for mental health:
SELECT AVG(mental_health) FROM brfss.Health_Conditions
WHERE mental_health>0;
```

□	avg
□	10

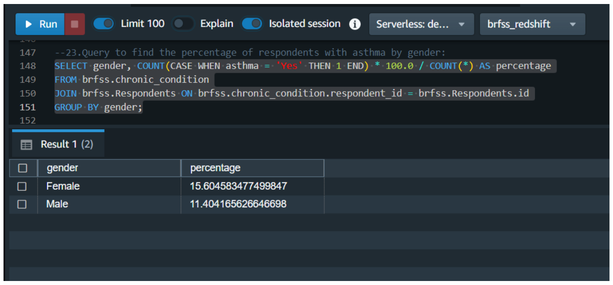
Fig. 14. Query to get the average score for mental health



```
--22.Query to find the percentage of respondents with diabetes by gender:
SELECT gender, COUNT(CASE WHEN diabetes = 'Yes' THEN 1 END) * 100.0 / COUNT(*) AS percentage
FROM brfss.chronic_condition
JOIN brfss.Respondents ON brfss.chronic_condition.respondent_id = brfss.Respondents.id
GROUP BY gender;
```

□	gender	percentage
□	Male	13.585434173669467
□	Female	12.718539474487615

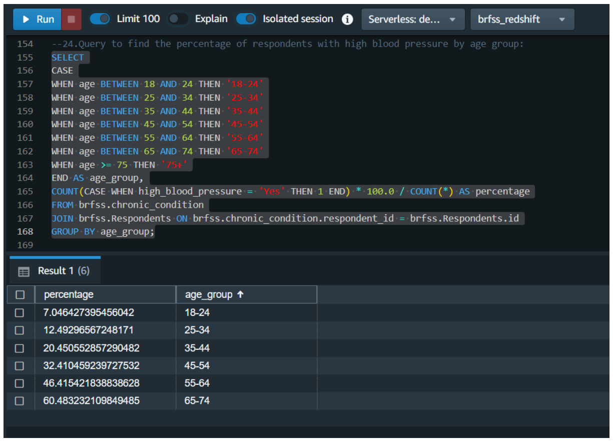
Fig. 15. Query to find the percentage of respondents with diabetes by gender



```
--23.Query to find the percentage of respondents with asthma by gender:
SELECT gender, COUNT(CASE WHEN asthma = 'Yes' THEN 1 END) * 100.0 / COUNT(*) AS percentage
FROM brfss.chronic_condition
JOIN brfss.Respondents ON brfss.chronic_condition.respondent_id = brfss.Respondents.id
GROUP BY gender;
```

□	gender	percentage
□	Female	15.604583477499847
□	Male	11.404165626646698

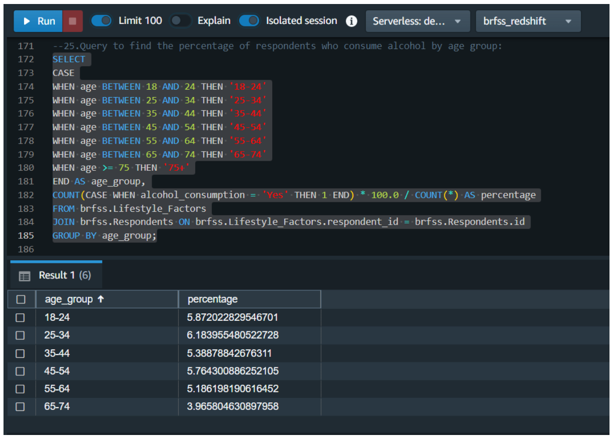
Fig. 16. Query to find the percentage of respondents with asthma by gender



```
--24.Query to find the percentage of respondents with high blood pressure by age group:
SELECT
CASE
WHEN age BETWEEN 18 AND 24 THEN '18-24'
WHEN age BETWEEN 25 AND 34 THEN '25-34'
WHEN age BETWEEN 35 AND 44 THEN '35-44'
WHEN age BETWEEN 45 AND 54 THEN '45-54'
WHEN age BETWEEN 55 AND 64 THEN '55-64'
WHEN age BETWEEN 65 AND 74 THEN '65-74'
WHEN age >= 75 THEN '75+'
END AS age_group,
COUNT(CASE WHEN high_blood_pressure = 'Yes' THEN 1 END) * 100.0 / COUNT(*) AS percentage
FROM brfss.chronic_condition
JOIN brfss.Respondents ON brfss.chronic_condition.respondent_id = brfss.Respondents.id
GROUP BY age_group;
```

□	percentage	age_group ↑
□	7.046427395456042	18-24
□	12.49296567248171	25-34
□	20.450552857290482	35-44
□	32.410459239727532	45-54
□	46.415421838838628	55-64
□	60.483232109849485	65-74

Fig. 17. Query to find the percentage of respondents with high blood pressure by age group



```
--25.Query to find the percentage of respondents who consume alcohol by age group:
SELECT
CASE
WHEN age BETWEEN 18 AND 24 THEN '18-24'
WHEN age BETWEEN 25 AND 34 THEN '25-34'
WHEN age BETWEEN 35 AND 44 THEN '35-44'
WHEN age BETWEEN 45 AND 54 THEN '45-54'
WHEN age BETWEEN 55 AND 64 THEN '55-64'
WHEN age BETWEEN 65 AND 74 THEN '65-74'
WHEN age >= 75 THEN '75+'
END AS age_group,
COUNT(CASE WHEN alcohol_consumption = 'Yes' THEN 1 END) * 100.0 / COUNT(*) AS percentage
FROM brfss.Lifestyle_Factors
JOIN brfss.Respondents ON brfss.Lifestyle_Factors.respondent_id = brfss.Respondents.id
GROUP BY age_group;
```

□	age_group ↑	percentage
□	18-24	5.872022829546701
□	25-34	6.183955480522728
□	35-44	5.38878842676311
□	45-54	5.764300886252105
□	55-64	5.186198190616452
□	65-74	3.965804630897958

Fig. 18. Query to find the percentage of respondents who consume alcohol by age group

- The majority of respondents fall into the 65-74 age group, followed by the 45-54 age group. The 18-24 age group has the lowest representation.
- Diabetes prevalence is slightly higher in males (13.59percentage) than in females (12.72percentage).
- Asthma prevalence is significantly higher in females (15.60percentage) than in males (11.40percentage).
- High blood pressure and high cholesterol are prevalent among respondents, affecting 40.94percentage and 37.00 percentage of them, respectively.
- Heart disease and stroke have a lower prevalence, affecting 5.78percentage and 4.29percentage of respondents, respectively.
- General health scores are mostly rated as ”Good” (3 out of 5), with average physical health and mental health scores at 11 and 10 (out of 30), respectively.
- A large percentage of respondents (86.75) have had cholesterol screenings, while 10.11percentage have had blood pressure screenings, and only 5.13percentage have had cancer screenings.

- Current smokers make up 9.63percent of respondents, and 4.99percent consume alcohol.
- Regular physical activity is reported by 46.25 percent of respondents, and 84.07 percent have good eating behavior.
- Higher education levels correlate with higher average income, and married respondents have a higher average income compared to other marital statuses.
- The percentage of respondents with healthcare access varies by state, with Alabama having the highest percentage (14.03) and the District of Columbia having the lowest (5.86).
- The prevalence of high blood pressure increases with age, and alcohol consumption is slightly higher among the 25-34 and 45-54 age groups.
- Respondents with multiple chronic conditions are relatively rare.
- A strong correlation exists between having a stroke and having high blood pressure, with 73.32 percent of respondents who had a stroke also having high blood pressure.
- A smaller percentage of respondents (9.42) who smoke also have heart disease.
- Cancer screening rates vary by age group and gender, with the highest percentages found in the 65-74 and 55-64 age groups for both males and females.

VII. 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.

VIII. Visualizations

Data visualization is the act of transforming raw data into meaningful graphical representations such as charts, graphs, pictures, and films. It is advantageous to gain insights from it in this manner since it simplifies the explanation of the digits and numbers. The relevant images are generated using the previously stated energy consumption algorithm.

IX. Technical Difficulty

In order to do analysis and get insights on risk factors producing various health conditions, we have used a variety of approaches throughout this project. This project has be implemented using AWS cloud infrastructure and services like AWS Redshift. These technologies has improved the project's

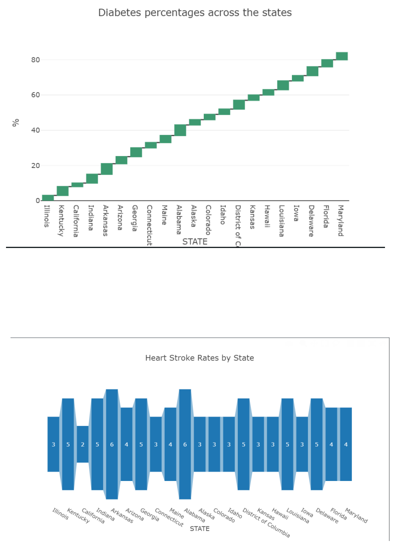


Fig. 19. Query to get the percentage of respondents who smoke and have heart disease

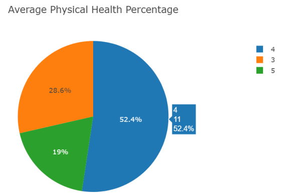


Fig. 20. Query to get the percentage of respondents who smoke and have heart disease

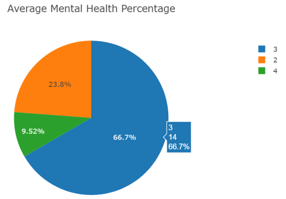


Fig. 21. Query to get the percentage of respondents who smoke and have heart disease

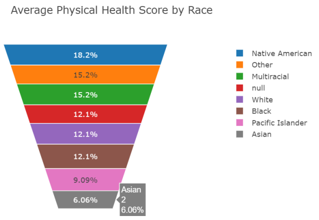


Fig. 22. Query to get the percentage of respondents who smoke and have heart disease

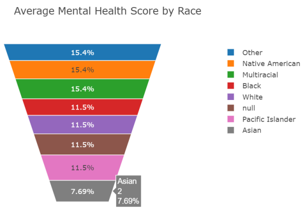


Fig. 23. Query to get the percentage of respondents who smoke and have heart disease

productivity while also given us a chance to practice utilizing these services and gain knowledge of ideas like data modeling, exploratory data analysis, and warehousing.

X. Impact

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.

XI. Relevance to the Course

This project provided us with an opportunity to work with more complex queries and analyses. We have covered the concepts such as normalization and indexing which are important considerations when designing and managing a database. It has also helped us to cover the concepts such as designing a relational schema, defining the relationships between them, querying the database to retrieve specific subsets of data, joining tables together to combine related data or aggregating data to calculate statistics, maintaining data integrity by performing ETL operations and visualizing the data.

XII. Uniqueness

The dataset that we took for this project is vast with over 120,000 entries from individuals across different states in the US, providing a large and diverse sample from both national and regional levels. Data related to the health status of the people along with their exact geographical location of each data entry, their low confidence and high confidence limits enables us to examine and identify potential discrepancies in terms of quality of life between gender, age, geographic areas, or other groups making it a comprehensive resource for understanding health outcomes.

XIII. MoM(Pair Programming)

Agenda:

- Review progress on Data model and ETL tool
- Discuss potential issues or concerns
- Make decisions and assign action items

Discussion:

- Venu presented their proposed solution to the data modeling and suggested using a dataset which is available on Kaggle.
- Mounika challenged the proposed solution, expressing concerns about its feasibility and meaningful entities, and further suggesting to use the dataset from the original BRFSS archives.
- A lively discussion ensued, with both parties presenting their perspectives and providing evidence to support their arguments.
- After careful consideration of all viewpoints, the group decided to complete the data model first and ensure that the dataset has enough scope to perform data modeling to draw meaningful entities.

XIV. Scrum Meeting MoM

Minutes of Meeting

Project Name: BRFSS Dataset Project

Date: April 17 2023

Attendees: Venu, Saran, Mounika, Jaya

Agenda:

- Review progress on data warehousing, ETL orchestration, data cleaning and data preprocessing, and visualizations
- Identify any potential roadblocks and challenges
- Assign action items

Discussion:

- Saran and Jaya shared updates on data cleaning and data preprocessing, highlighting that the team had made significant progress in cleaning and preprocessing the data, and that they would soon start visualizing the data.
- Mounika discussed her progress in developing the ETL orchestration and mentioned that the team would soon have a functional ETL process.
- Venu shared an update on data warehousing using Redshift and mentioned that progress was on track and that the team could soon start loading data into the warehouse.
- The team discussed potential roadblocks and identified that the integration of the various components of the project could pose a challenge.
- Venu proposed scheduling a meeting to discuss integration and identify any potential issues that may arise.
- The team agreed to schedule a meeting to discuss integration and identified the need for frequent communication among the team members.

Action Items:

- Saran and Jaya will continue working on data cleaning and preprocessing and start visualizing the data
- Mounika will continue developing the ETL orchestration and work on integrating it with the data warehousing.
- Venu will continue working on data warehousing using Redshift.
- The team will schedule a meeting to discuss integration and identify any potential issues that may arise.

XV. VERSION CONTROL

We used GitHub to publish and make our project public. https://github.com/monikavempalli/brfss-db

XVI. CONCLUSION

• This project analyzed a diverse sample of respondents across various states, age groups, and backgrounds. The data reveals several key trends and insights that can inform public health policies and interventions.

• Firstly, chronic conditions such as high blood pressure, high cholesterol, and diabetes are highly prevalent among the respondents. Specifically, high blood pressure is the most common condition, affecting over 40percent of respondents. This underscores the need for effective public health measures to prevent and manage these conditions, including promoting lifestyle changes and ensuring access to appropriate healthcare services.

• Gender differences are also evident in the prevalence of certain conditions. For instance, females have a higher prevalence of asthma, while males have a higher prevalence of diabetes. These differences should be taken into account when designing targeted interventions and healthcare services.

• Age is another significant factor influencing the prevalence of chronic conditions. Older respondents have higher rates of high blood pressure, high cholesterol, and stroke. This highlights the importance of age-specific interventions and healthcare services, as well as the need to address the unique health challenges faced by older populations.

• Education and marital status both have an impact on respondents' average income. College graduates earn the highest average income, while those with less than a high school education earn the lowest. Similarly, married respondents earn the highest average income, while separated respondents earn the lowest. These findings underscore the importance of addressing socioeconomic disparities in health outcomes and access to healthcare services.

• The wide variation in the distribution of respondents across states and the differences in healthcare access by state indicate that state-specific policies and interventions may be necessary to address the unique health challenges faced by each state's population.

• Lifestyle factors such as smoking, alcohol consumption, and physical activity also play a significant role in respondents' health. Smoking is linked to heart disease, and alcohol consumption varies by age group, with the highest consumption observed in the 25-34 age group. Regular physical activity is reported by just over 46percent of respondents, suggesting that efforts to promote physical activity could have a significant impact on overall health outcomes.

• In conclusion, the findings from this project reveal a complex interplay of factors influencing the health of respondents, including chronic conditions, age, gender, socioeconomic status, and lifestyle choices. Public health policies and interventions should be designed to address these diverse factors in a comprehensive and targeted manner, with the goal of improving overall health outcomes and reducing disparities in access to healthcare services.

REFERENCES

[1] "Behavioral Risk Factor Surveillance System" 1993-2010 Health-Related Quality of Life Data Dataset: https://www.kaggle.com/datasets/thedevastator/behavioral-risk-factor-surveillance

[2] "Behavioral Risk Factor Surveillance System-Centers of Disease Control and Prevention" Basic Information about Dataset: https://www.cdc.gov/brfss/index.html

[3] B. Xie, "Association Between Sleep Duration and Cognitive Impairment: A Cross-Sectional Study Using the BRFSS Database," 2020 International Conference on Public Health and Data Science (ICPHDS), Guangzhou, China, 2020, pp. 108-115, doi: 10.1109/ICPHDS51617.2020.00030.

[4] "Trends in Health Behaviors and Health Outcomes among US Adults: The Behavioral Risk Factor Surveillance System 1984-2017" by Michael R. Kramer et al. (American Journal of Preventive Medicine, 2020)

[5] "State-Level Variations in the Prevalence of Chronic Obstructive Pulmonary Disease (COPD) in the United States: Findings from the Behavioral Risk Factor Surveillance System (BRFSS)" by Kavita Singh et al. (International Journal of Environmental Research and Public Health, 2018)

[6] "Prevalence of Sexual Orientation Identity and Health Disparities in Adults: The Behavioral Risk Factor Surveillance System, 2013" by Brian R. Hollenbeck et al. (American Journal of BRFSS Public Health, 2018)

[7] Remington PL, Smith MY, Williamson DF, Anda RF, Gentry EM, Hogelin GC. Design, characteristics, and usefulness of state-based behavioral risk factor surveillance: 1981-87. Public Health Rep. 1988 Jul-Aug;103(4):366-75. PMID: 2841712; PMCID: PMC1478092.

[8] Macera CA, Ham SA, Yore MM, Jones DA, Ainsworth BE, Kimsey CD, Kohl HW 3rd. Prevalence of physical activity in the United States: Behavioral Risk Factor Surveillance System, 2001. Prev Chronic Dis. 2005 Apr;2(2):A17. Epub 2005 Mar 15. PMID: 15888228; PMCID: PMC1327711.

[9] Centers for Disease Control and Prevention (CDC). Methodologic changes in the Behavioral Risk Factor Surveillance System in 2011 and potential effects on prevalence estimates. MMWR Morb Mortal Wkly Rep. 2012 Jun 8;61(22):410-3. PMID: 22672976.

[10] Siegel, Paul Z., et al. "Behavioral Risk Factor Surveillance, 1986–1990." Morbidity and Mortality Weekly Report: Surveillance Summaries, vol. 40, no. SS-4, 1991, pp. 1–23. JSTOR, http://www.jstor.org/stable/24675437. Accessed 9 Mar. 2023.

Appendix

Criteria	Pts	Comments
Presentation Skills Includes time management	5 pts	
Code Walkthrough	3 pts	The Code walkthrough during the presentation would cover the architecture and orchestration of the project and other related tools. All the code and scripts used in the project are stored on GitHub, and the project report includes screenshots of each outcome.
Discussion / Q&A	4 pts	
Demo	5 pts	
Version Control Use of Git / GitHub or equivalent; must be publicly accessible	3 pts	https://github.com/monikavempalli/brfss-db Used git for our project.
Significance to the real world	5 pts	The dataset from the Behavioral Risk Factor Surveillance System (BRFSS) is significant in the real world as it provides crucial information on the health behaviors and risk factors of adults in the United States. This information can be used to develop and implement public health programs and policies aimed at improving the overall health of the population.
Lessons learned Included in the report and presentation? How substantial and unique are they?	5 pts	Yes, we have covered all the relevant sections.
Innovation	5 pts	Using the BRFSS dataset, we were able to draw insights of health patterns across several states in the United States. The innovation in this project is the key challenge to data model a dataset of huge attributes and clean it.
Teamwork	5 pts	As a team, everyone has actively taken part in each and every phase of the project and we practiced agile methodology during the course of the project.
Technical difficulty	4 pts	We have learnt so many technologies like AWS Redshift, RDS, S3 and techniques like warehousing.

This criterion is linked to a learning outcome Practiced pair programming? See: https://en.wikipedia.org/wiki/Pair_programming Li nksLinksLinksLinksLinksLinksLinksLinksLinksL in ksLinksLinks Links to an external site. to an external site. to an external site. to an external site. to an external site. to an external site. to an external site. to an external site. to an external site. to an external site. to an external site. to an external site. to an external site	2 pts	Yes, we have practiced pair programming and attached the minutes of meeting as a proof.
Practiced agile / scrum (1-week sprints)? Submit evidence on Canvas - meeting minutes, other artifacts	3 pts	Attached MOM of one of our scrum meetings as a proof.
Used Grammarly / other tools for language? Grammarly free version is sufficient; can use other tools as well. Submit report screenshot on Canvas.	2 pts	Grammarly and Google docs were used
Slides	5 pts	Have attached it in Canvas.
Report Format, completeness, language, plagiarism, whether turnItIn could process it (no unnecessary screenshots), etc	5 pts	We utilized the IEEE format for the document created in overleaf and selectively incorporated features in Microsoft PowerPoint to design the slides for our presentations.
Used unique tools E.g.: LaTeX for writing reports (submit .tex that is not generated from another format such as .docx; generating from .lyx and similar LaTeX editor outputs is fine. Also checkout https://www.overleaf.com/LinksLinksLinks Links to an external site. to an external site.) Unique features of Prezi or powerpoint, etc	5 pts	We have used LaTex using overleaf website to generate the project report
Performed substantial analysis using database techniques Project must include an analytics component	3 pts	We used Tableau and Redshift to visualize our findings.
Used a new database or data warehouse tool not covered in the HW or class	3 pts	AWS Redshift and AWS RDS
Used appropriate data modeling techniques	5 pts	Out of 330 columns in the dataset, selected (27) columns and normalized the dataset into 3NF separating entities. For visual reference, we used draw.io to capture relationships.
Used ETL tool	1 pts	Developed a custom ETL orchestration in python to perform data loading.
Demonstrated how Analytics support business	3 pts	We were able to draw conclusions of

decisions		health patterns which may help any business that is involved across health sector
Used RDBMS Idea is to exercise as many topics from the course as possible	1 pts	We have used RDBMS to run various queries on MySQL Workbench
Used Datawarehouse Idea is to exercise as many topics from the course as possible	1 pts	Used AWS redshift for querying
Includes DB Connectivity / API calls Possibly using Python	1 pts	Used MySQL Connector to connect to an RDS instance.
Used NOSQL	1 pts	