**"Comprehensive Diabetes and Health Determinants Analysis: A Multifaceted Study"**

**Introduction:-** Diabetes stands as the prevailing ailment in the United States, impacting a significant portion of the population. Through our current research, we aim to investigate the intricate interplay between diabetes and a range of associated factors.

The research encompasses datasets with crucial variables including age, income, levels of physical activity, obesity, gender, insurance coverage, prevalence of diabetes, race and ethnicity, life expectancy, and educational attainment.

Our objective is to discern the influence of these diverse factors on the occurrence of diabetes, and to determine how diabetes prevalence varies across states and age brackets. Additionally, we seek to understand the determinants of diabetes within the United States.

To construct this research dataset, we have undertaken a comprehensive process of merging various datasets using the Big Query environment. This involved the integration of diverse variables and the implementation of various types of JOIN operations to unify them. Our approach centered on selecting a common variable as a pivot point for joining all the tables.

**Why did we choose those specific datasets?**

**Age:-** Age is a significant risk factor for developing type 2 diabetes, with prevalence increasing substantially in older adults. Research indicates that individuals aged 45-64 have a prevalence nearly four times higher than those aged 18-44, while those 65 and older have a prevalence over five times higher than the youngest age group.Hence it is important to consider age as a factor while studying type-2 diabetes.

**DiabetesPrev:-** Comparing prevalence across geographic regions, demographic groups, and risk factor strata highlight disparities in diabetes distribution, it helps in understanding the disease prevalence in specific areas which might be because of different geographic and climatic conditions which can affect the disease outcome.Hence it is important to consider Diabetes prevalence as a factor in the dataset.

**Education:-** Higher educational attainment is associated with lower risk of developing type 2 diabetes. Those with a college degree or more have around a 25-35% lower risk compared to those with less than a high school education, as people who are educated have a better understanding of the disease and the prevention of lifestyle factors which can cause type-2 diabetes.

**Gender:-** Men and women have inherent biological differences, like hormones and body composition, that influence their risk of developing type 2 diabetes. Analyzing gender differences can uncover sex-specific aspects of diabetes pathogenesis.

**Health\_insurance:-** Percentage of the population that is insured vs. uninsured. Lack of health insurance reduces access to preventive care and treatment in diabetes, which can negatively impact life expectancy and health outcomes.Hence it is important in considering diabetes as a factor in diabetes.

**Income:-** Income ranges or averages across the population. Lower income correlates strongly with reduced access to healthcare, food insecurity, worse health behaviors, and lower life expectancy. Income provides insight into socioeconomic status and its effect in type-2 diabetes.

**Life Expectancy:-** Life expectancy provides a high-level overview of the general health and mortality patterns within a population. Lower life expectancy may indicate poor health conditions related to uncontrolled diabetes.Hence it is important to consider life expectancy as a factor when considering type-2 diabetes and also people with type-2 diabetes have a higher rate of mortality in type-2 diabetes.

**ObesityPrev:-** Obesity is a strong risk factor for many leading causes of death and generally correlates with lower life expectancy and research says that people who are obese are highly susceptible to type-2 diabetes. Hence analyzing obesity rates can be helpful in studying type-2 diabetes.

**PhysicalInactivity:-** Lack of regular physical activity is strongly associated with higher rates of chronic diseases like heart disease and diabetes, which reduce life expectancy. Hence considering Physical inactivity is also an important factor when studying type-2 diabetes.

**Race/ethnicity:-** Genetics plays a major role in all the diseases. Some of the races may be highly susceptible for diabetes as they have homozygous alleles which might play an important role whether a person might have type-2 diabetes or not.

**Creating a Denormalized Table by Merging Individual Tables in BigQuery:-**

The process of joining multiple datasets into one large denormalized table started with 10 source datasets containing information related to demographics, health, physical inactivity, education, and other metrics. Joining the datasets together was done incrementally using the fips code column, a unique identifier present in all the datasets, as the key for linking the data. First, dataset 1 and 2 were joined on a fips code column into a common table. Then this common table was joined with dataset 3, also using the fips code column. This process continued by joining additional datasets one at a time, always using the fips code column as the join key. At each step of the joining process, only the necessary columns needed for later analysis were selected in order to minimize the width of the table. In the last two join steps, DISTINCT was used again after the SELECT to eliminate any duplicate rows and ensure only unique rows were included in the final denormalized table. The end result was a large denormalized table containing 25 columns and 590 billion rows of data, enabling powerful multidimensional analysis across demographic, health, economic, education, and other metrics.

**Types of Research you think this dataset could help inform:-**The current dataset which we have created using BigQuery has all the major factors that directly and indirectly cause type-2 diabetes in individuals. Studying these factors plays a major role in understanding how these factors affect the disease cause. Hence by taking these factors into consideration we can use this in the research scenarios given below.

* By using the above dataset local health departments can find out geographical hotspots of high diabetes prevalence.
* This dataset can be used in finding trends of diabetes in different ages,poverty,education and other socio economic factors.
* By implementing regression analysis to the current dataset we can find out impactful social determinants that cause type-2 diabetes.
* Current dataset can be combined with other datasets and can be used in analyzing utilization or costs in system wide burdens of diabetes.

**Description of the dataset:-** This dataset containing population-level information on demographics, socioeconomic factors, and health metrics related to diabetes provides a valuable real-world foundation for research studies. Analyzing this data enables epidemiologists and health services researchers to uncover insights and patterns that can drive the creation of evidence-based interventions, policies, and best practices. The knowledge gained from investigating the relationships within this data can then be implemented in healthcare and community settings to improve diabetes prevention, management, and outcomes. Ultimately, the availability of this comprehensive dataset facilitates research that helps reduce the incidence of diabetes and associated medical complications.

**Conclusion:-** This dataset offers a rich resource to study diabetes dynamics in the U.S. It contains many variables about age, income, exercise, obesity, gender, insurance, prevalence, race, education, and life expectancy.

These factors are included because research shows they influence diabetes risk, rates, and outcomes. Factors like age, income, insurance coverage, and education play key roles in diabetes patterns. Also, variables about gender, obesity, exercise, and race/ethnicity provide important insights into how health factors interact with diabetes.

The dataset has many research uses, like finding diabetes hotspots, seeing trends across groups, and statistical analyses of social factors. It could also be combined with other data to study diabetes healthcare use and costs.

By fully utilizing this dataset, researchers, epidemiologists, and health practitioners can tailor evidence-based interventions and policies. The goal is to improve diabetes prevention, care, and health for affected groups. Overall, the dataset offers a strong foundation to better understand and address diabetes through data-driven insights.

**References:-**

1. [Krishna S, Gillespie KN, McBride TM. Diabetes burden and access to preventive care in the rural United States. J Rural Health. 2010;26(1):3-11.](https://pubmed.ncbi.nlm.nih.gov/20105262/) doi:10.1111/j.1748-0361.2009.00259.x
2. [Walker RJ, Strom Williams J, Egede LE. Influence of Race, Ethnicity and Social Determinants of Health on Diabetes Outcomes. Am J Med Sci. 2016;351(4):366-373. doi:10.1016/j.amjms.2016.01.008](https://pubmed.ncbi.nlm.nih.gov/27079342/)
3. [Geiss LS, Wang J, Cheng YJ, et al. Prevalence and incidence trends for diagnosed diabetes among adults aged 20 to 79 years, United States, 1980-2012. JAMA. 2014;312(12):1218-1226. doi:10.1001/jama.2014.11494](https://pubmed.ncbi.nlm.nih.gov/25247518/)
4. [Khan MAB, Hashim MJ, King JK, Govender RD, Mustafa H, Al Kaabi J. Epidemiology of Type 2 Diabetes - Global Burden of Disease and Forecasted Trends. J Epidemiol Glob Health. 2020;10(1):107-111. doi:10.2991/jegh.k.191028.001](https://pubmed.ncbi.nlm.nih.gov/32175717/)

**"Sources of Dataset and Links"**

**Dataset Sources:-**

* [Small Area Income and Poverty Estimates (SAIPE) Program](https://www.census.gov/programs-surveys/saipe.html)
* [American Community Survey (ACS)](https://www.census.gov/programs-surveys/acs)
* [Small Area Health Insurance Estimates (SAHIE) Program](https://www.census.gov/programs-surveys/sahie.html)
* [USDA ERS](https://www.ers.usda.gov/)
* [HRSA Data Warehouse](https://data.hrsa.gov/)
* [CDC - NCHS - National Center for Health Statistics](https://www.cdc.gov/nchs/index.htm)
* [CDC WONDER](https://wonder.cdc.gov/)
* [NCHHSTP AtlasPlus](https://www.cdc.gov/nchhstp/atlas/index.htm)

**Links for datasets:-**

* Age - [Diabetes Data and Statistics | CDC](https://www.cdc.gov/diabetes/data/index.html)
* DiabetesPrev - [Diabetes Data and Statistics | CDC](https://www.cdc.gov/diabetes/data/index.html)
* Education - [American Community Survey (ACS)](https://www.census.gov/programs-surveys/acs)
* Gender - [American Community Survey (ACS)](https://www.census.gov/programs-surveys/acs)
* Health\_insurance - [Small Area Health Insurance Estimates (SAHIE) Program](https://www.census.gov/programs-surveys/sahie.html)
* Income - [Small Area Health Insurance Estimates (SAHIE) Program](https://www.census.gov/programs-surveys/sahie.html)
* Life\_Expectancy - [CDC - NCHS - National Center for Health Statistics](https://www.cdc.gov/nchs/index.htm)
* ObesityPrev - [CDC - NCHS - National Center for Health Statistics](https://www.cdc.gov/nchs/index.htm)
* PhysicalInactivity - [CDC - NCHS - National Center for Health Statistics](https://www.cdc.gov/nchs/index.htm)
* Race\_ethnicity - [American Community Survey (ACS)](https://www.census.gov/programs-surveys/acs)

**Google Drive link for the project :-** <https://drive.google.com/drive/folders/1pBK8nKlZ4ok5DccJlqIaBsBPVMo5cclx?usp=share_link>

**LLM usage:-**

**Chatgpt** : <https://chat.openai.com/>

**Prompt** : “Can i use the keyword distinct after select statement to remove the duplicate rows?”

**Output** :

Absolutely! If you're working with a relational database, you can use the DISTINCT keyword in a SELECT statement to retrieve unique rows. For example:

SELECT DISTINCT column1, column2, ...

FROM your\_table;

This will return only distinct combinations of values in the specified columns, effectively removing duplicate rows from the result set. Keep in mind that the DISTINCT keyword applies to the entire row, so if there are duplicates in some columns but not others, those rows will still be considered distinct.

Is there a specific scenario you're dealing with?

**Prompt** : “How to write a report for a research dataset which is created by joining 10 different tables?”

**Output** :

Writing a report for a research dataset that involves joining 10 different tables can be a complex but rewarding task. Here's a general outline to help you structure your report:

### 1. Introduction

- Briefly introduce the purpose of the research and the significance of the dataset.

- Mention the scope and objectives of your analysis.

### 2. Data Sources

- Provide an overview of the 10 tables you used.

- Explain the rationale for selecting these tables and how they contribute to your research.

### 3. Data Schema

- Include a visual representation or description of the database schema.

- Highlight key relationships and connections between tables.

### 4. Research Questions

- Clearly state the research questions or hypotheses you aim to address with the dataset.

### 5. Data Preparation

- Discuss any data cleaning or preprocessing steps.

- Explain how missing values, outliers, or inconsistencies were handled.

### 6. Data Joining

- Describe the process of joining the 10 tables.

- Explain the criteria for joining and any challenges encountered.

### 7. Descriptive Statistics

- Present basic statistics and summary metrics for key variables.

- Use visualizations if appropriate to illustrate patterns or trends.

### 8. Analysis and Findings

- Address each research question or hypothesis.

- Provide detailed insights, supported by evidence from the dataset.

- Use visualizations, tables, and charts to enhance clarity.

### 9. Limitations

- Acknowledge any limitations of the dataset or your analytical approach.

- Discuss potential sources of bias or error.

### 10. Future Work

- Suggest areas for future research based on the limitations identified.

- Propose potential improvements or expansions to the dataset.

### 11. Conclusion

- Summarize key findings.

- Emphasize the relevance of your results to the research questions.

### 12. References

- Cite any sources, databases, or literature used in your analysis.

### 13. Appendices

- Include any additional information, code snippets, or supplementary materials.

### Tips:

- Use clear and concise language.

- Include visuals like charts, graphs, and tables to enhance understanding.

- Provide code snippets if the report is intended for a technical audience.

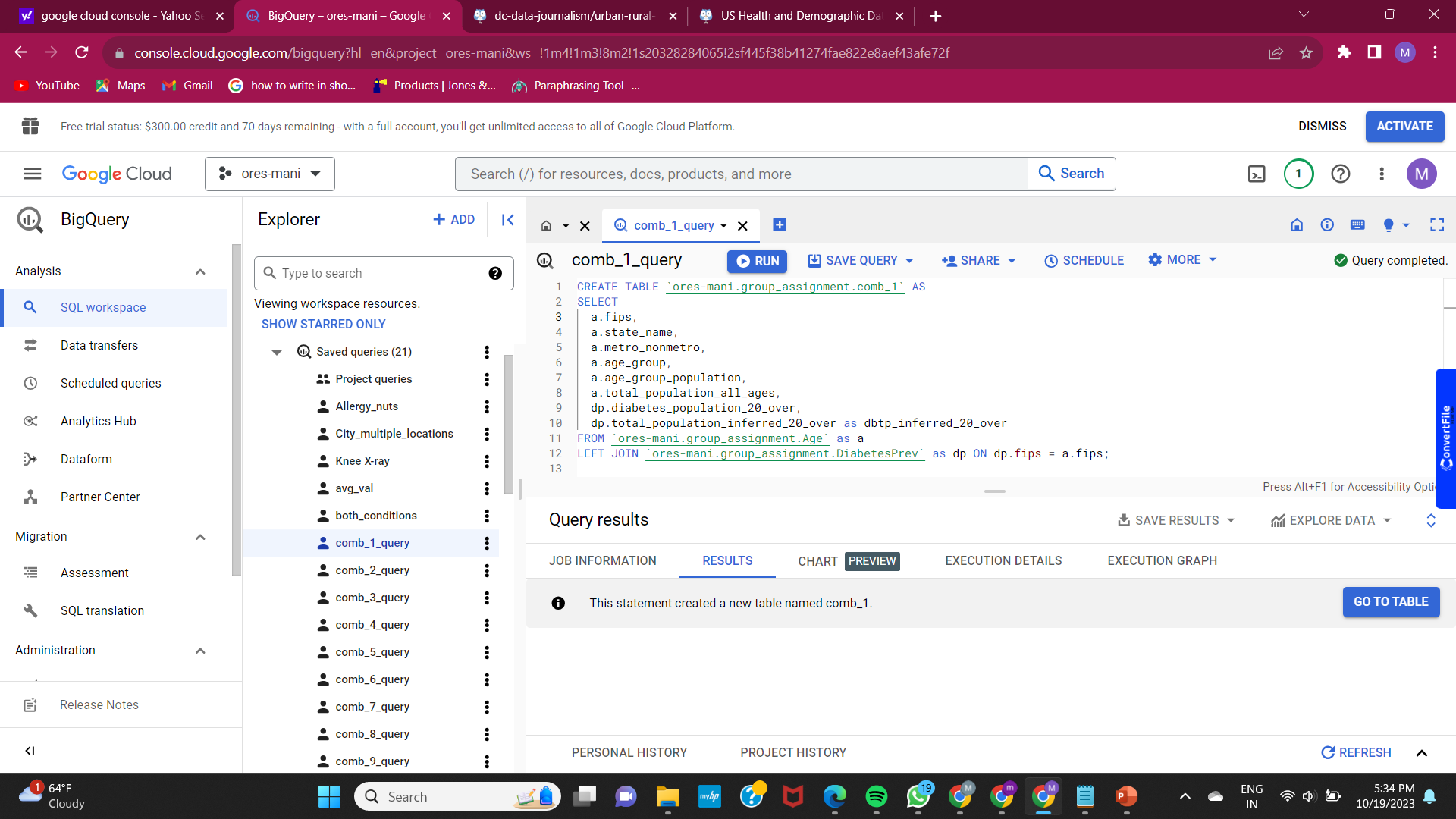
- Review and revise your report for clarity, coherence, and correctness.

Remember, the specific content and structure may vary depending on the nature of your research and your audience.

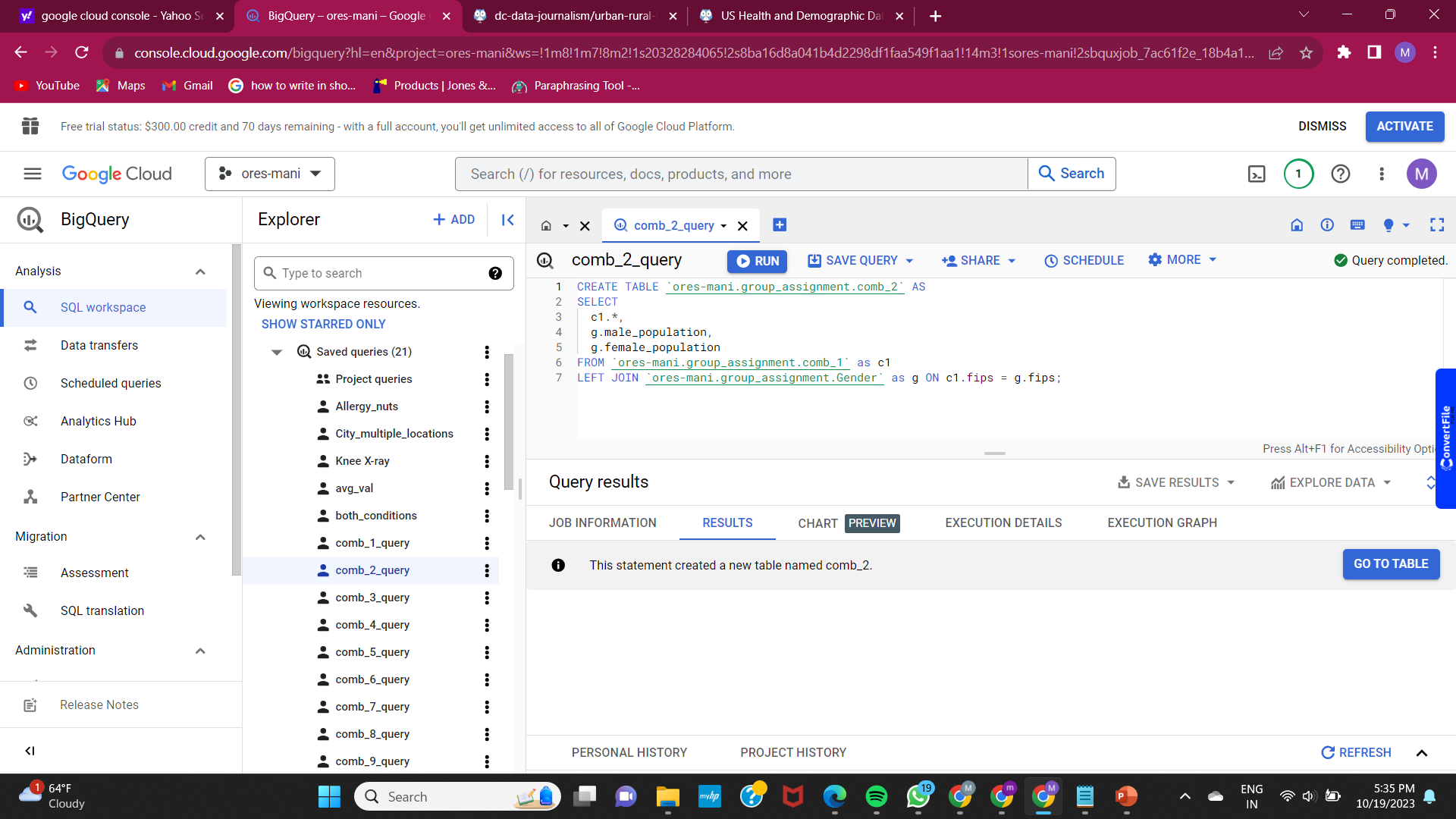
**Appendix:-**

**Process of Creating a Denormalized table:-**

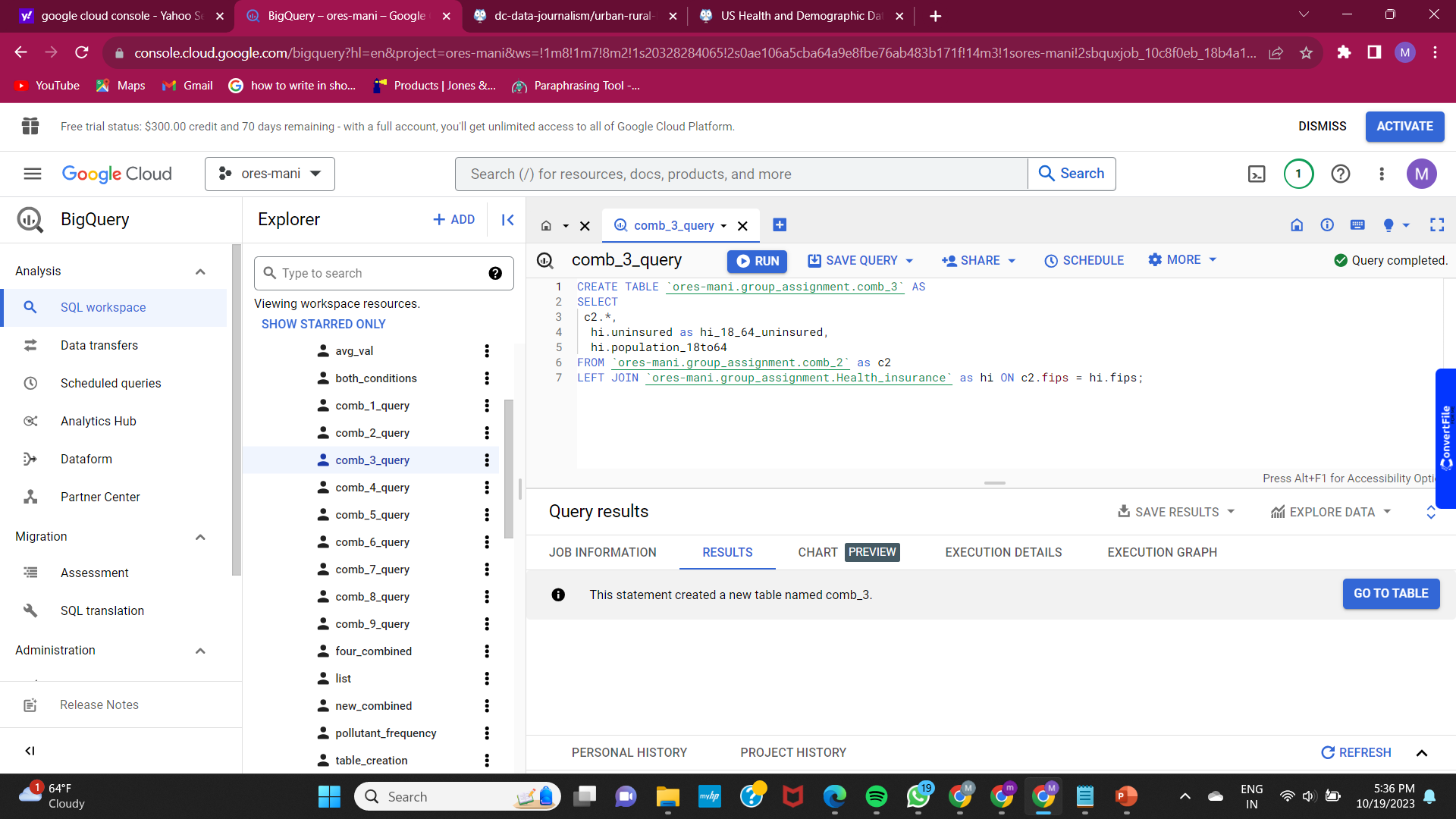
**Comb\_1\_query:-** for joining tables Age and DiabetesPrev



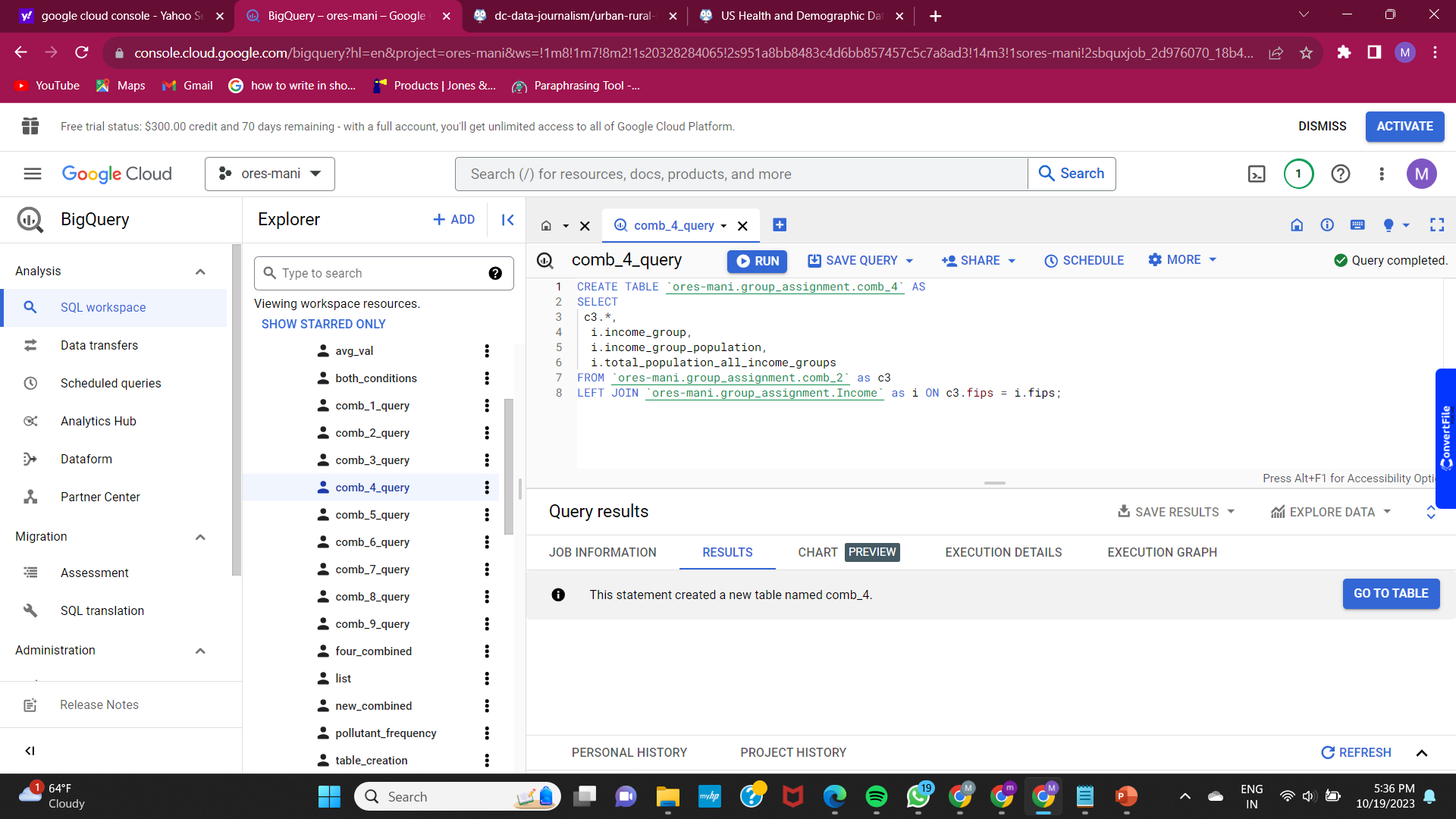
**Comb\_2\_query:-** The below query is used to join “comb\_1 table” and “Gender”.



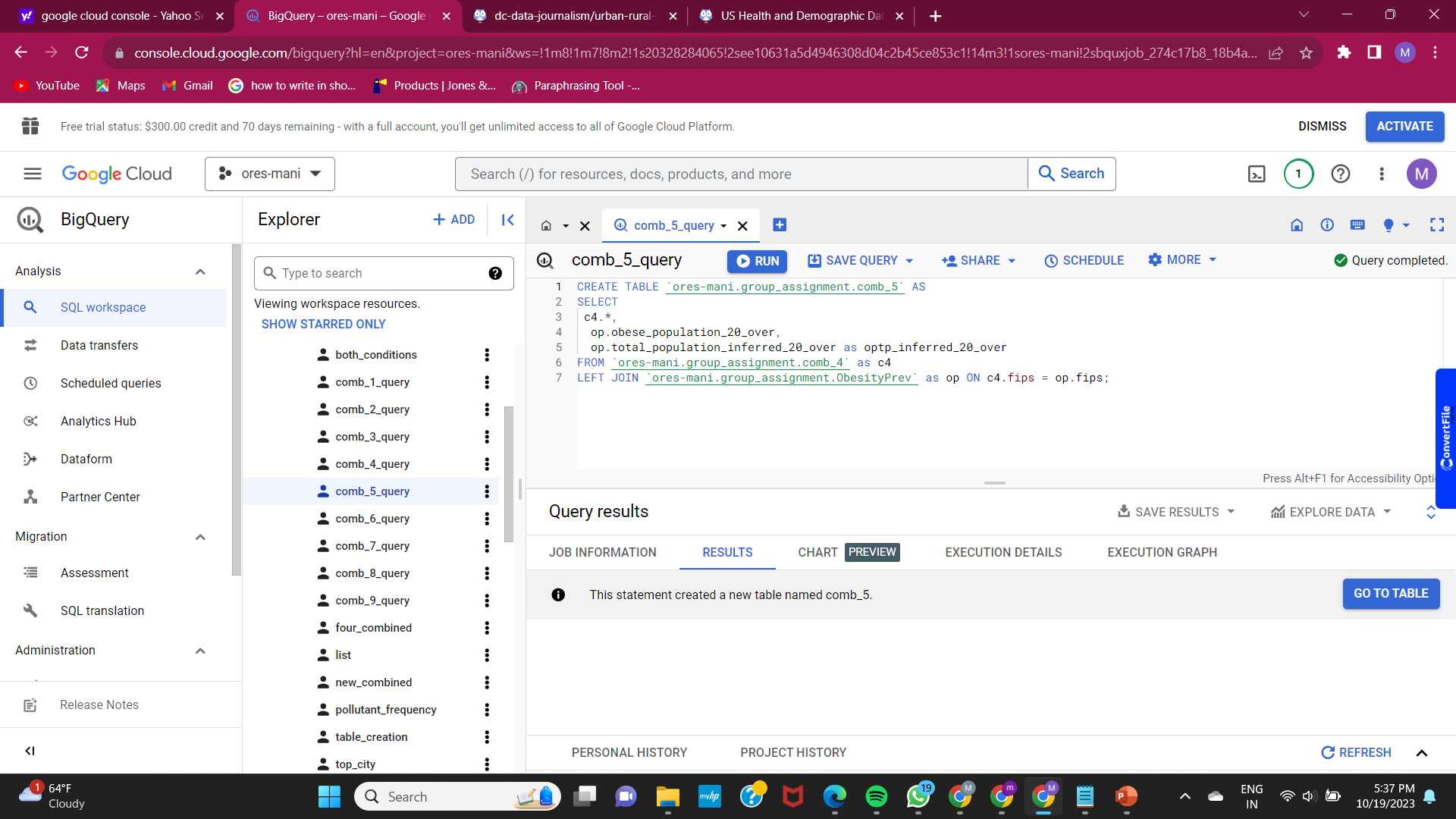
**Comb\_3\_query:-** The below query is used to join “comb\_2 table” and “Health\_insurance”



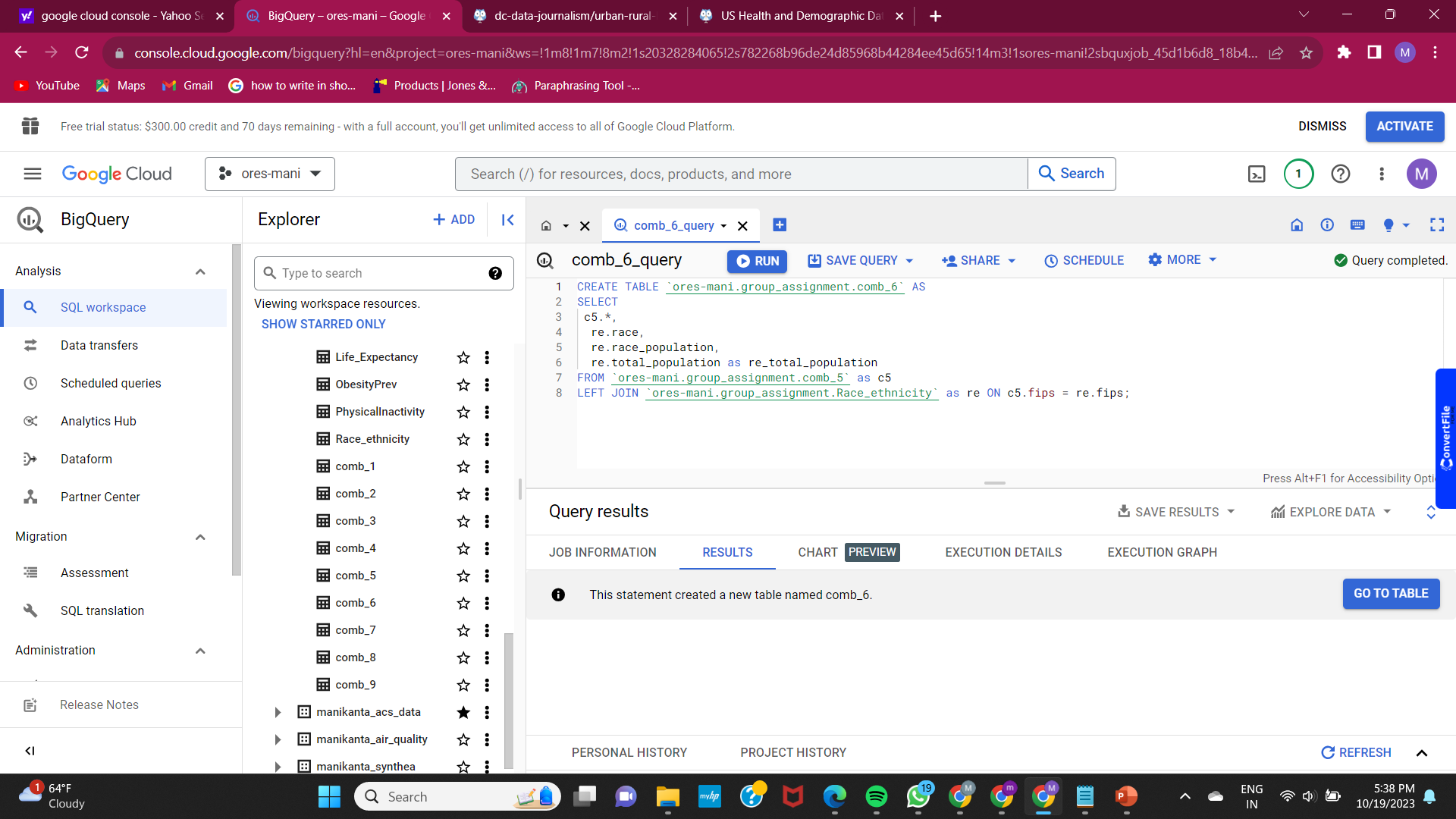
**Comb\_4\_query:-** The below query is used to join “comb\_3 table” and “Income”



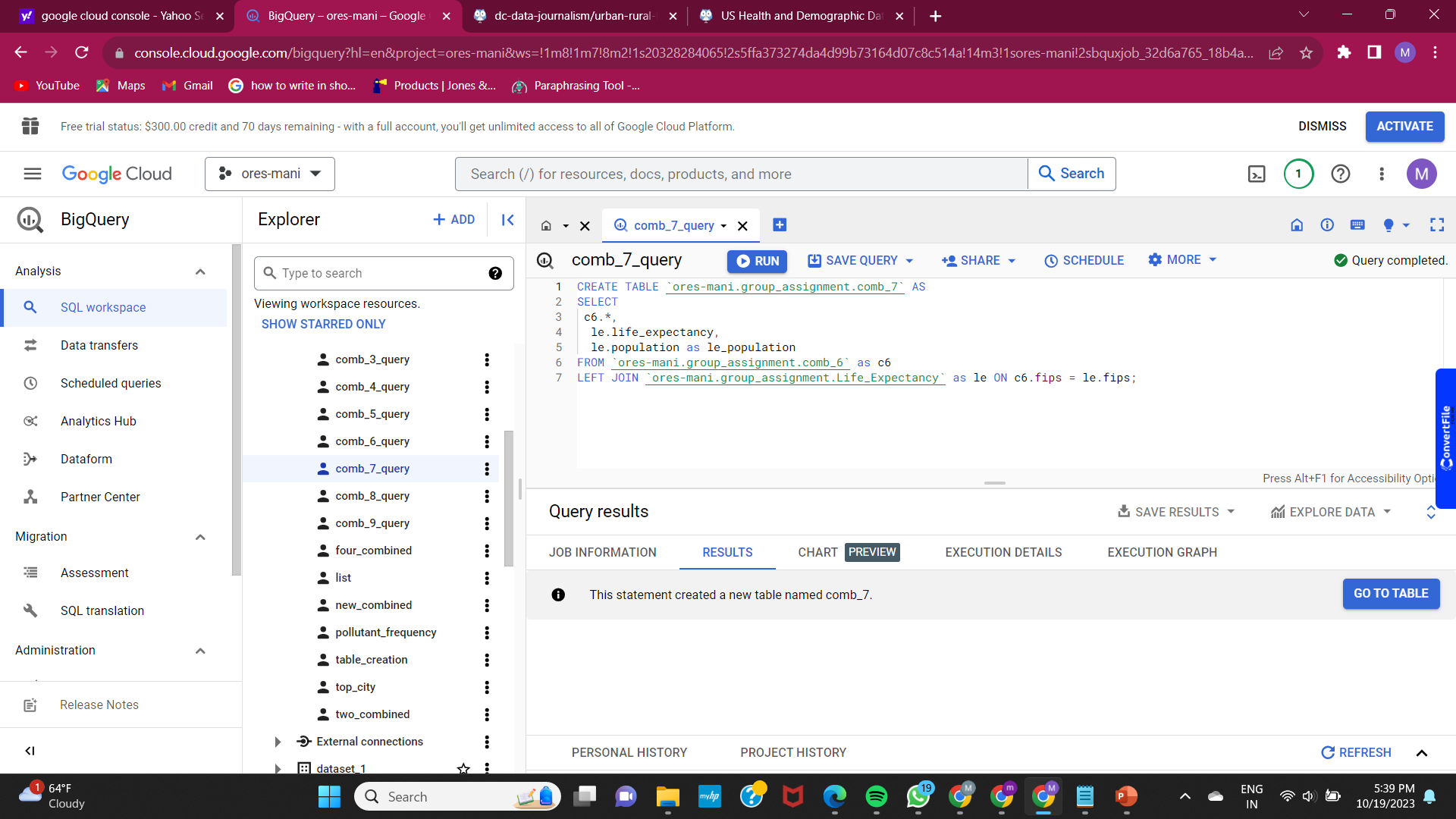
**Comb\_5\_query:-** The below query is used to join “comb\_4 table” and “ObestityPrev”



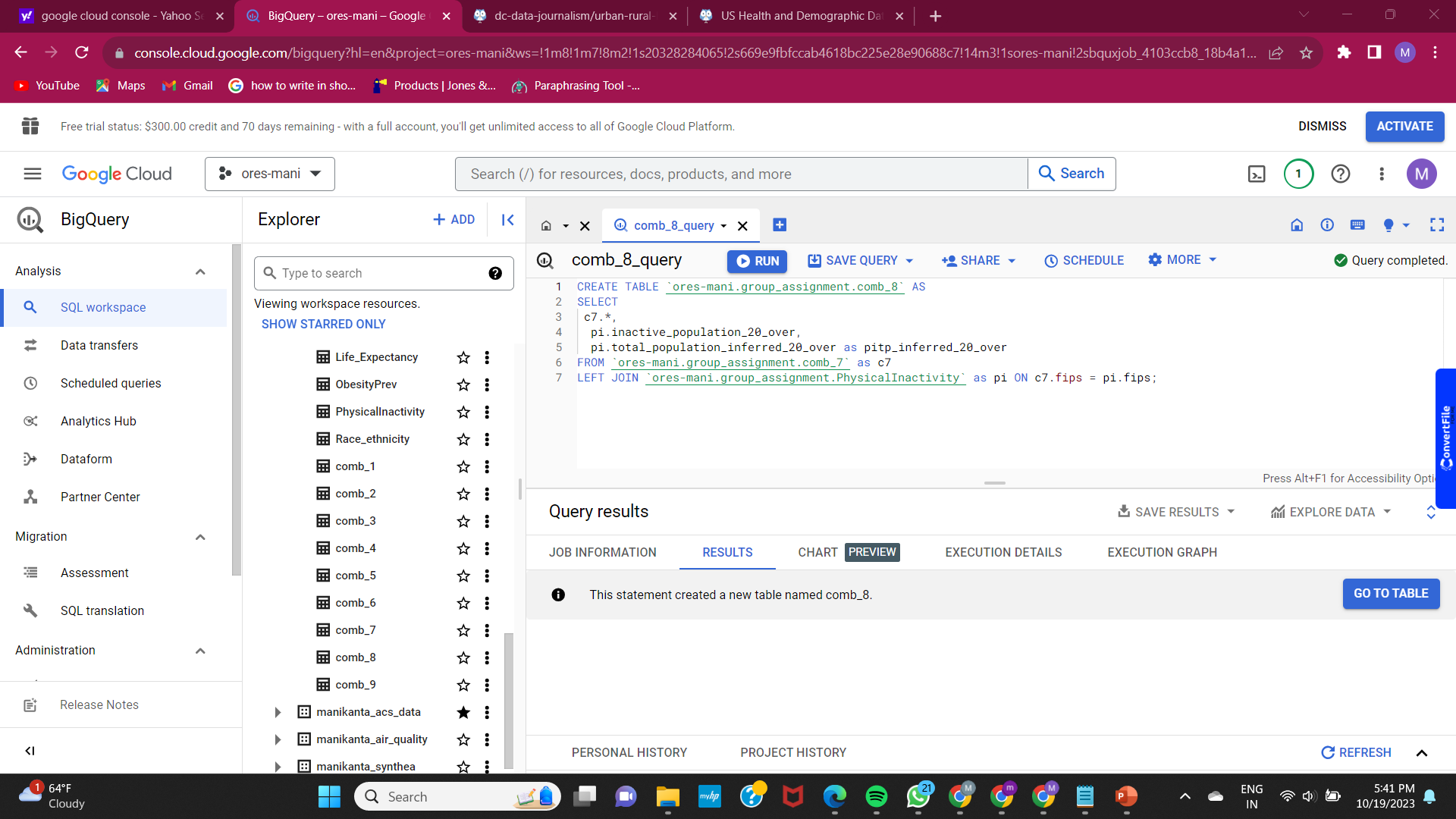
**Comb\_6\_query:-** The below query is used to join “comb\_5 table” and “Race\_ethnicity”



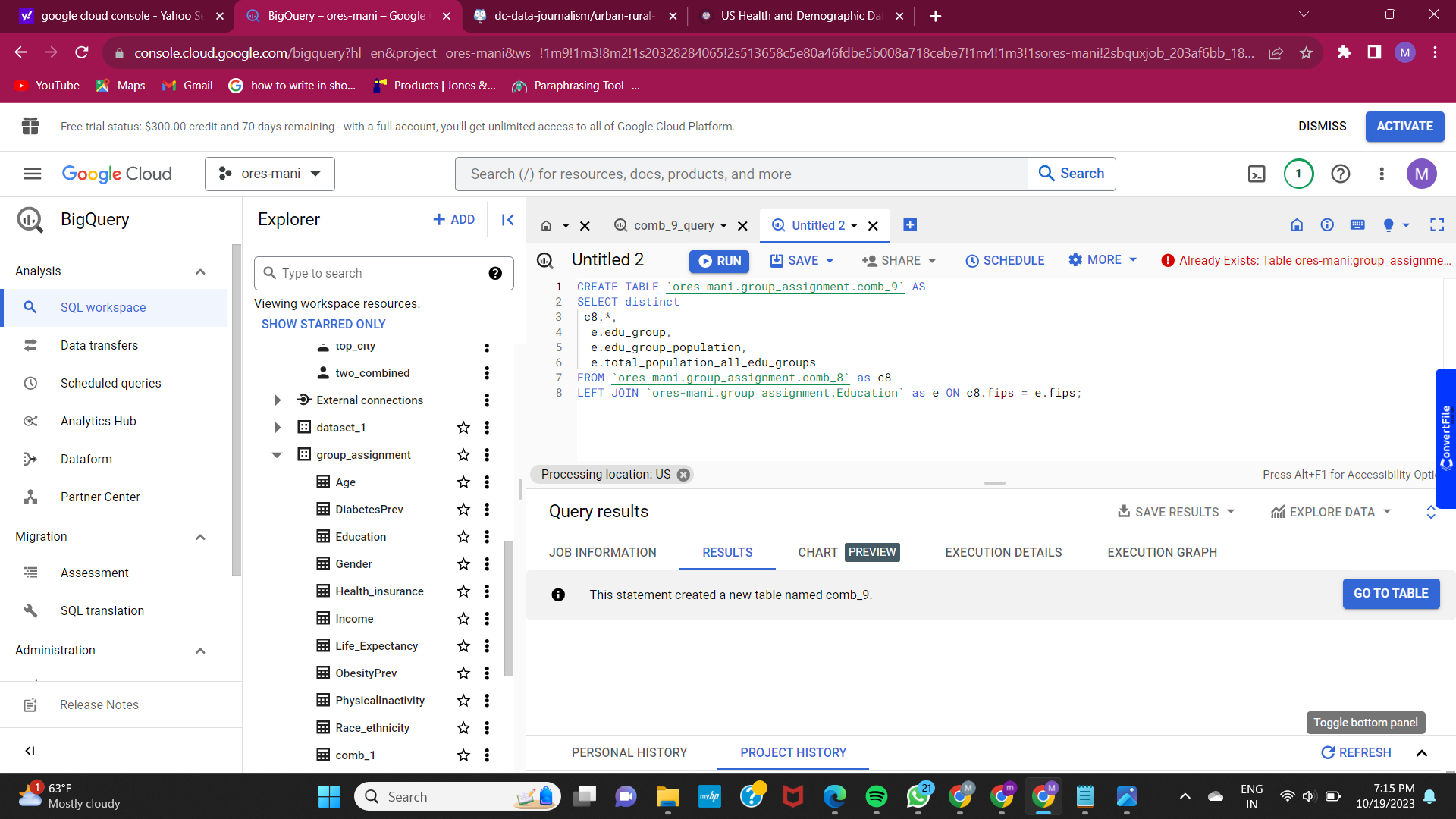
**Comb\_7\_query:-** The below query is used to join “comb\_6 table” and “Life\_Expectancy”



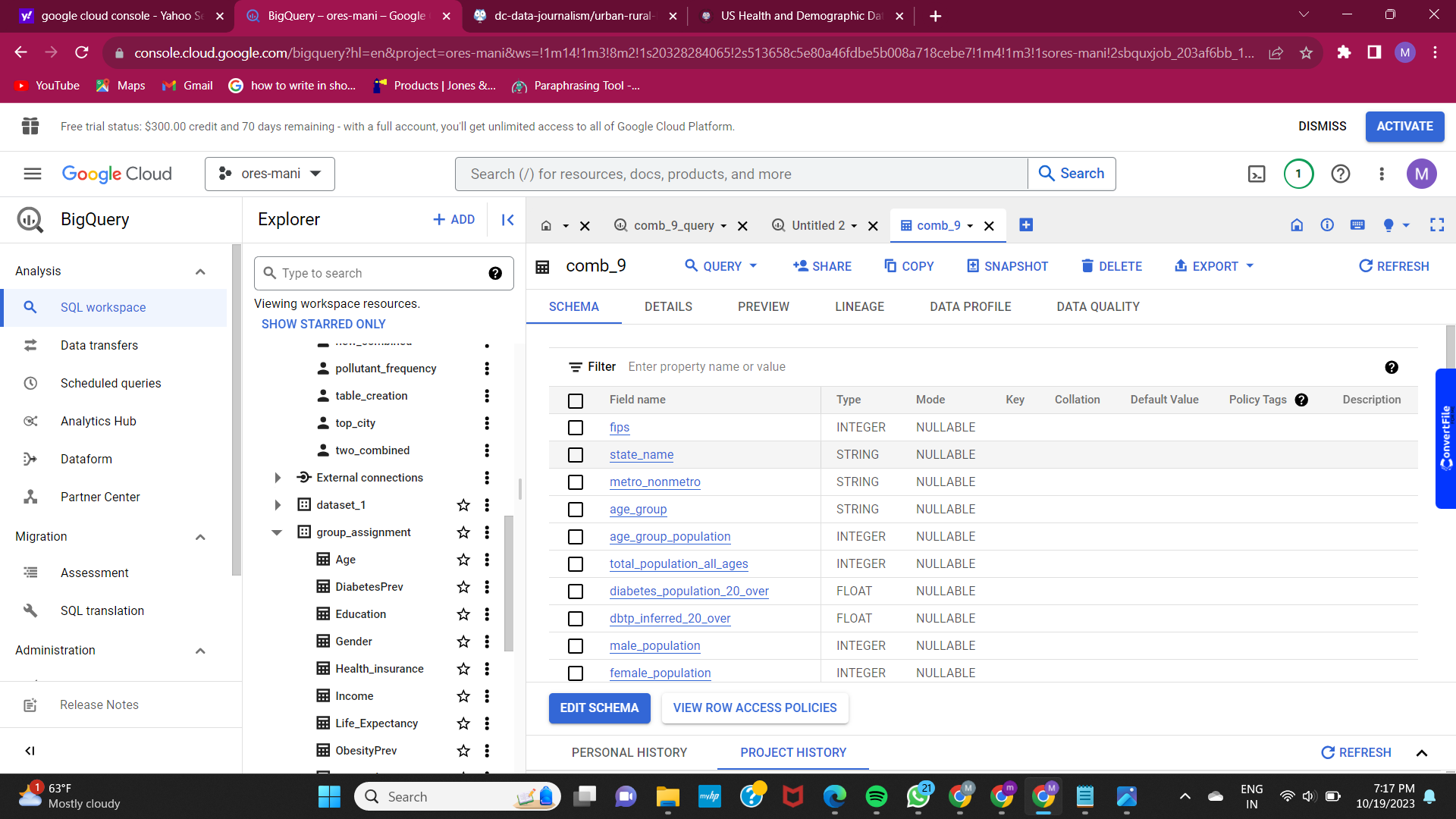
**Comb\_8\_query:-** The below query is used to join “comb\_7 table” and “PhysicalInactivity”



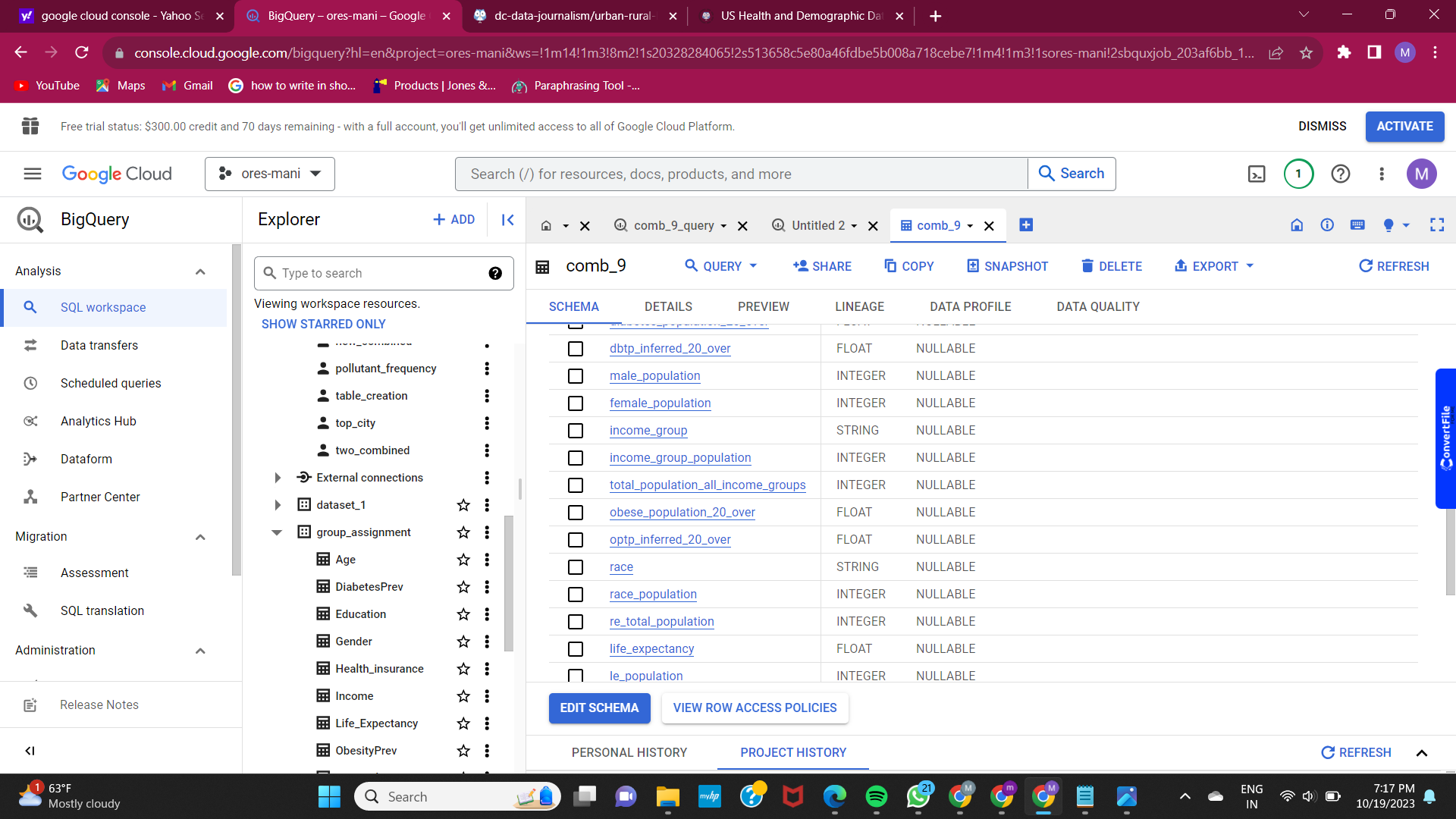
**Comb\_9\_query:-** The below query is used to join “comb\_8 table” and “Education”



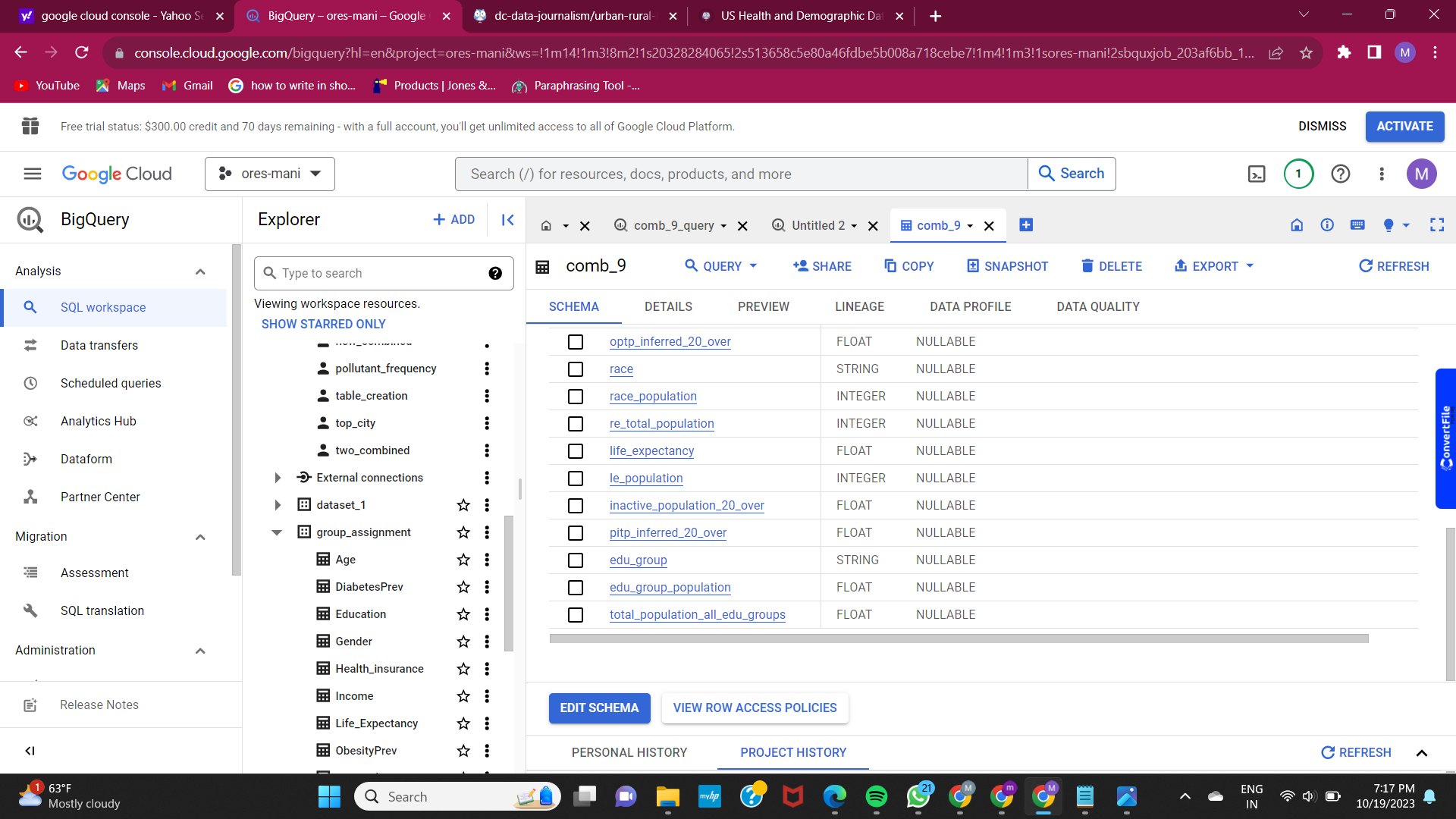
**Schema of Dataset-1:-** The below three screen-shots show all the columns in the denormalized table.



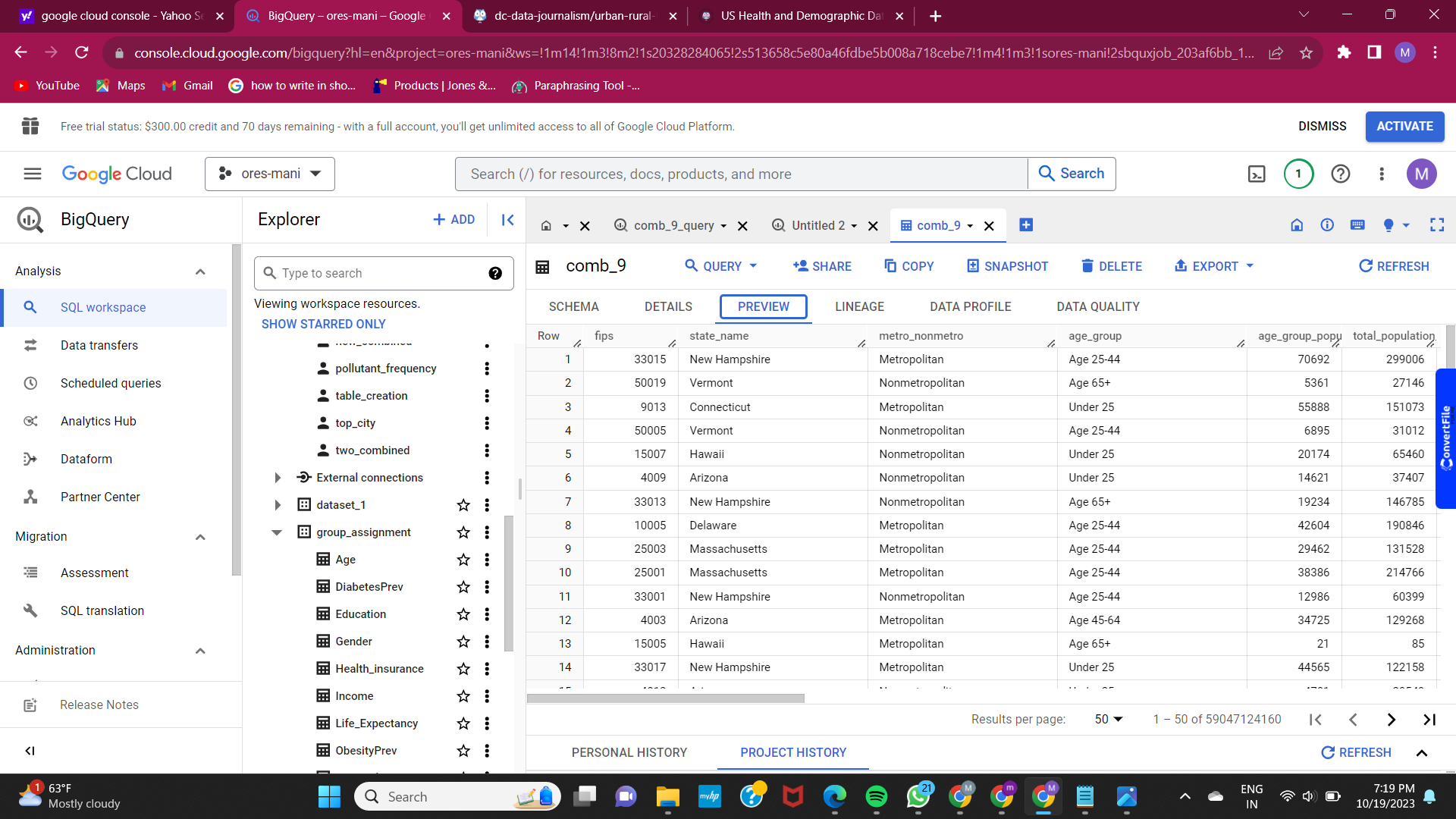
**Schema of Dataset-2:-** The below three screen-shots show all the columns in the denormalized table.



**Schema of Dataset-3:-** The below three screen-shots show all the columns in the denormalized table.

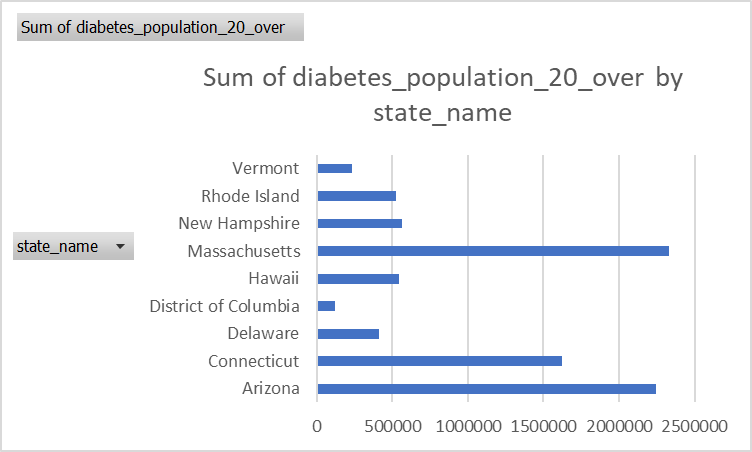


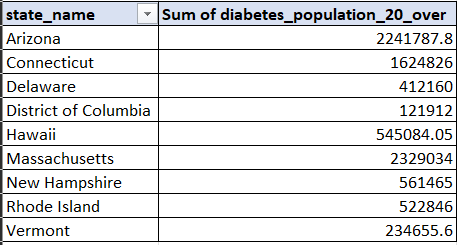
**Preview of dataset:-** The below screenshot shows table structure.



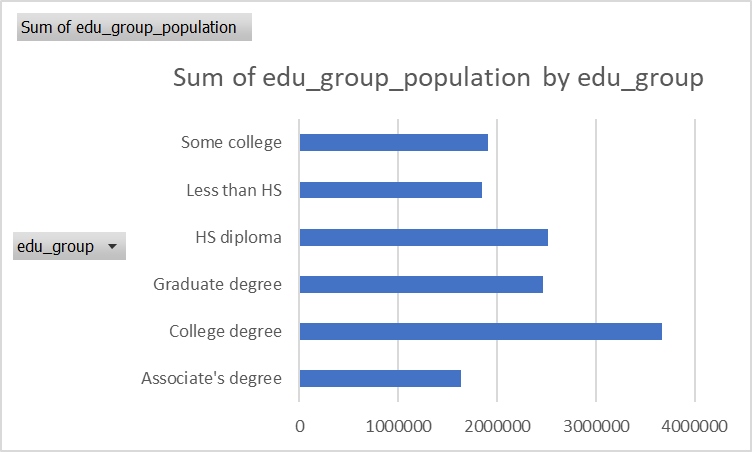
Examples showing how we can use this dataset to find out basic research questions.

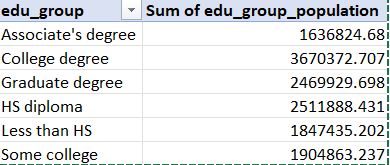
1. Graph showing Diabetes population of people over the age of 20 in X-axis and different states in the Y-axis. By looking at the graph we can see that Massachusetts has the highest rate of diabetes among the population who are aged above 20 years.



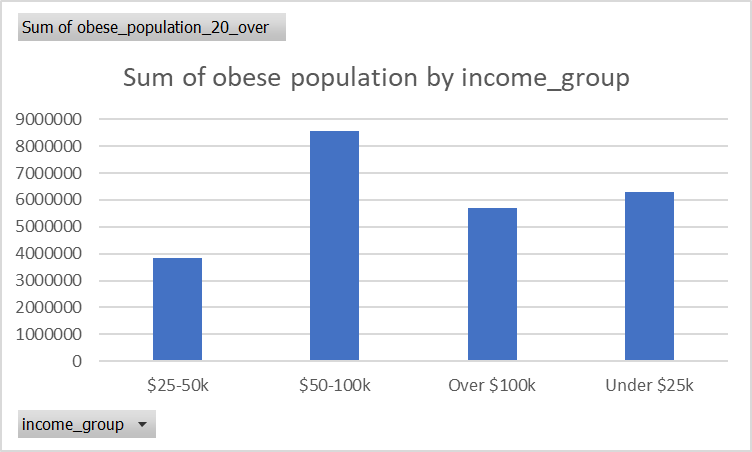


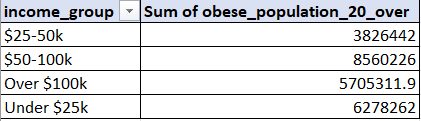
1. Graph showing the sum of education group population on the X-axis and their highest degree of education on the Y-axis. By looking at the graph we can see that Most of the people have a College degree as highest level of education.





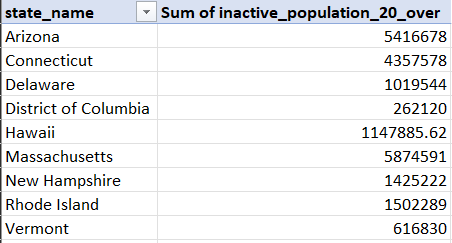
1. Graph showing Sum of obese population on the Y-axis and income group on the X-axis.By looking at the graph we can say that who earn between $50-100k are more obese when compared to people of other income groups.





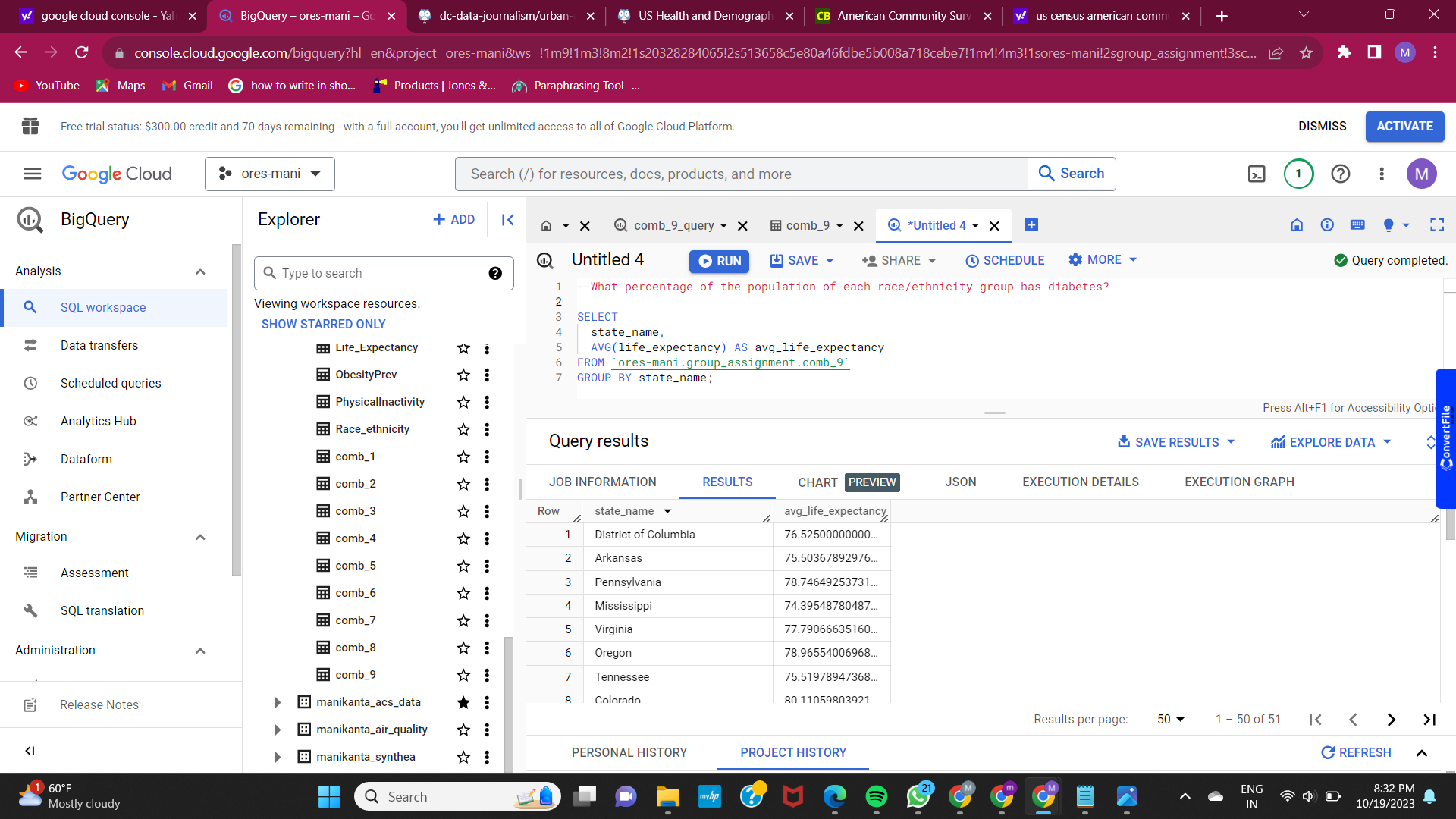
1. Graph showing Sum of inactive population on X-axis and states on the Y-axis. We can say that Massachusetts has a more inactive population who are aged above 20 years.





**Analyzing Population Health Trends with the Denormalized Dataset:**

**Question 1**: What percentage of the population of each race/ethnicity group has diabetes?



**Question 2**: What is the obesity rate for each income group?

