**Data Engineering**

**Assignment 1: Healthcare Analytics**

Overview:

This assignment is designed to assess your practical skills and problem-solving abilities in the field

of data engineering. This assignment will test your understanding of data processing,

transformation, and integration. You will be required to ingest, transform and analyze healthcare data.

Data Source:

You will be using the "Health Insurance Marketplace Analysis" from Kaggle. The dataset

represents individual medical costs billed by health insurance and includes the following fields:

age, sex, bmi, children, smoker, region, and charges. You can download the dataset from this link.

<https://www.kaggle.com/code/bbhatt001/predictors-of-medical-expenses/input>

Tasks:

1. Ingestion and ETL: Load the dataset. Create an ETL job using the framework of your

preference, such as AWS Glue, Databricks, Apache Airflow, Talend, or any other tool you are

comfortable with. Perform necessary transformations, and store the transformed data.

Choose an appropriate data storage service.

2. Data Quality Checks: Implement data quality checks like checking for null values or verifying

the transformation logic output.

3. Data Analysis: Use SQL queries to analyze the data.

4. Output:

Understand the Market Dynamics for Health Insurance Plans & develop a Targeted

Marketing Strategy

Understand the distribution and characteristics of health insurance plans across different

states and age groups to identify potential markets to target or areas where the company

could improve its offerings. Understand the value proposition of different plans and

recommend the top 5 avenues where marketing effort should be spent by plan, age, and

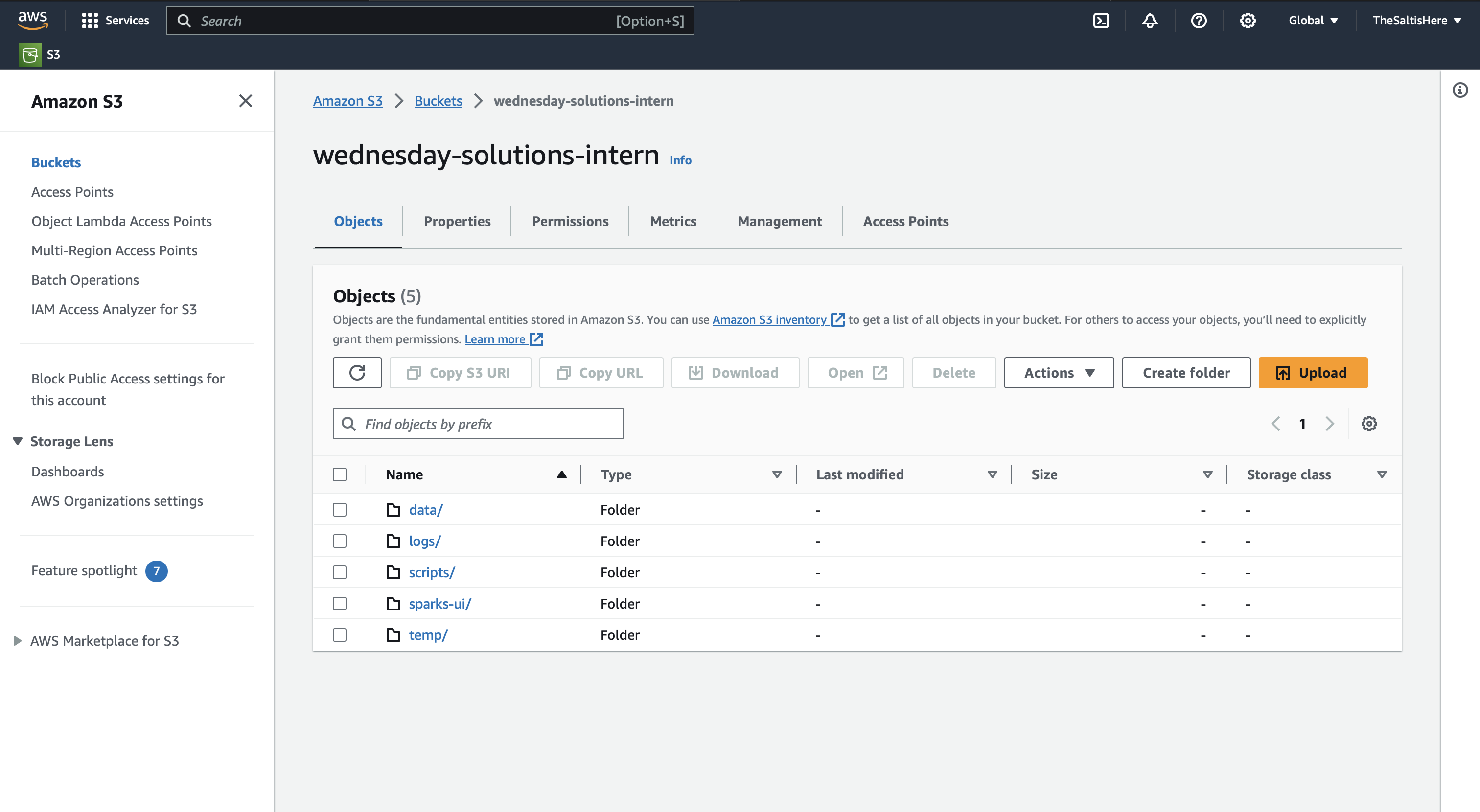
state.

For this assignment , I have used AWS SERVICES for doing everything including ETL ACTIVITY and DATA ANALYSIS Queries using SQL in AWS.

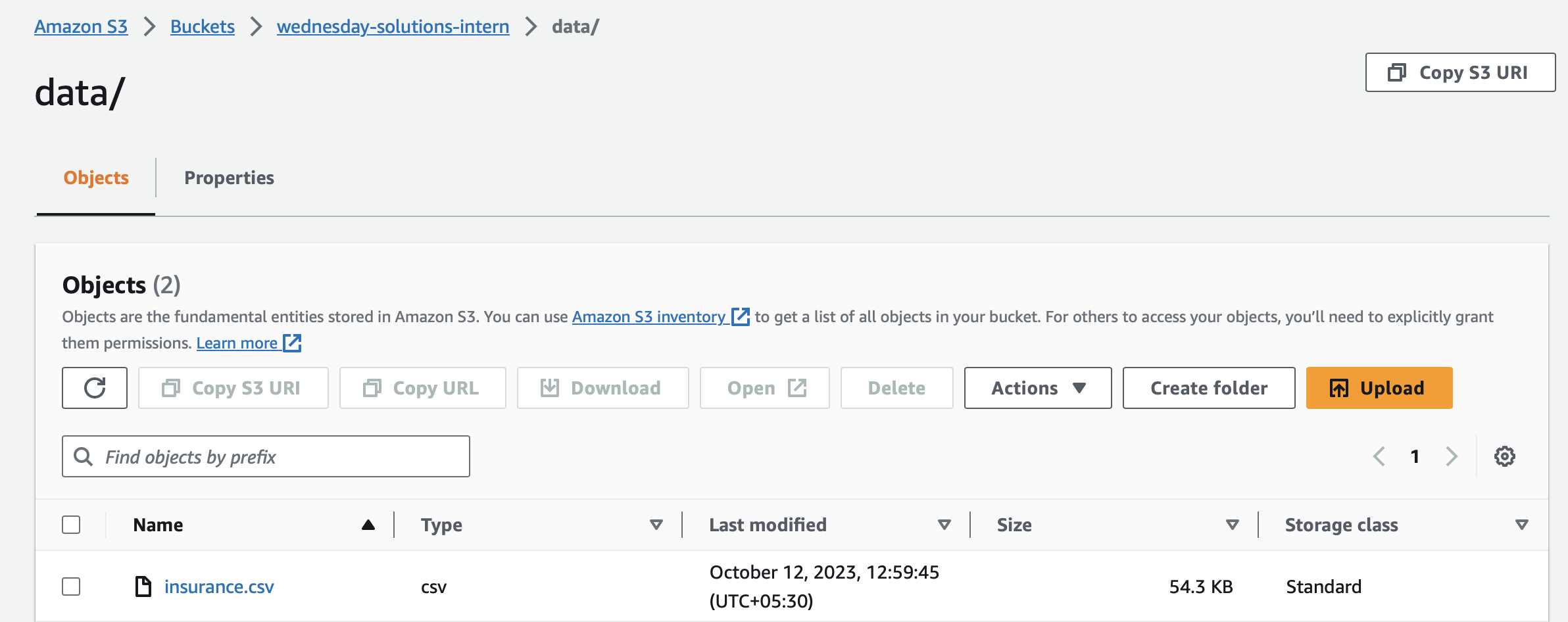
TASK 1

1.Creation of an AWS S3 Bucket USING AWS Console:

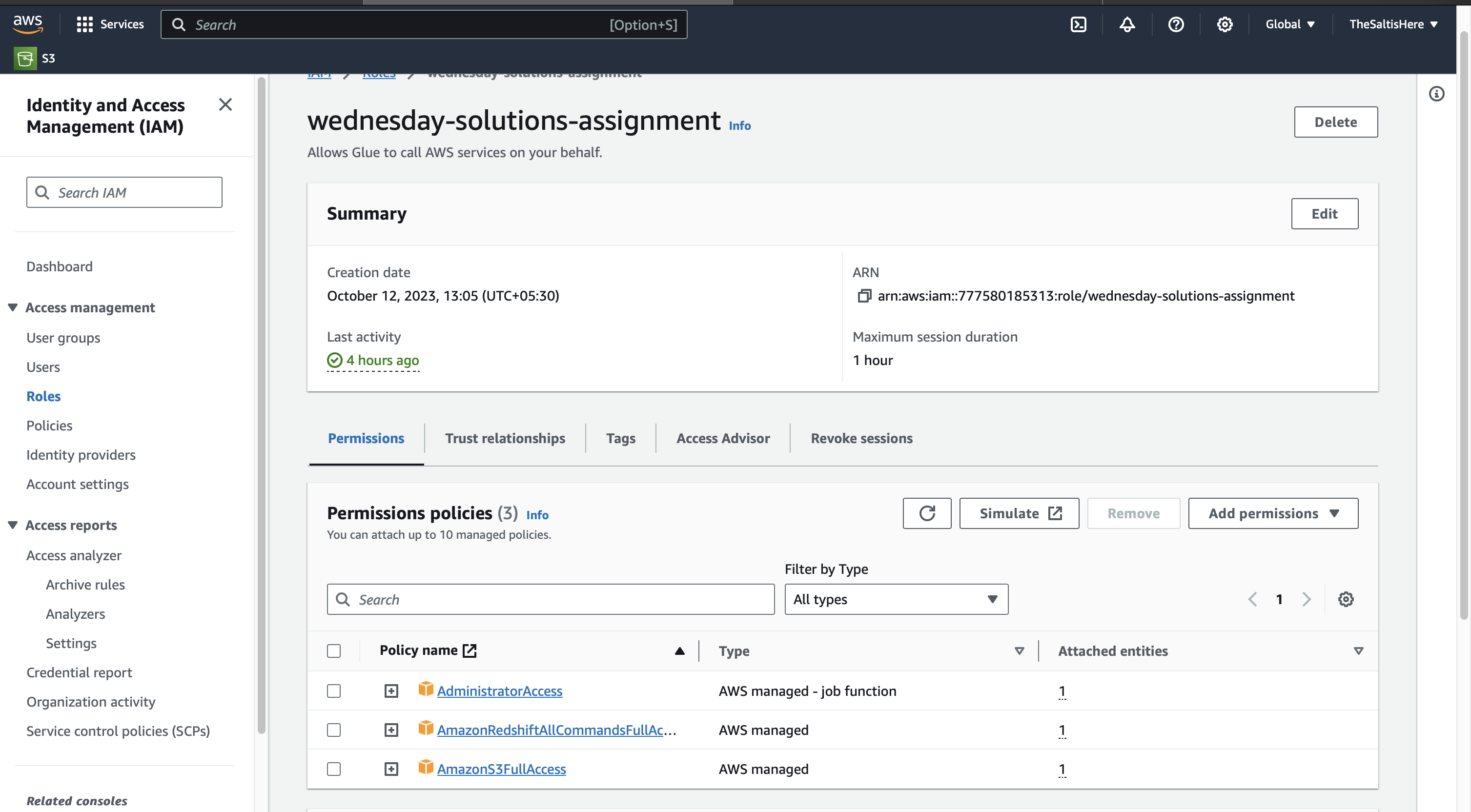
Firstly we create a AWS S3 bucket named Wednesday-solutions-intern. This bucket will store the output from the ETL Scripts that are written. The Data folder contains the Output files and the logs folder contain the scripts.



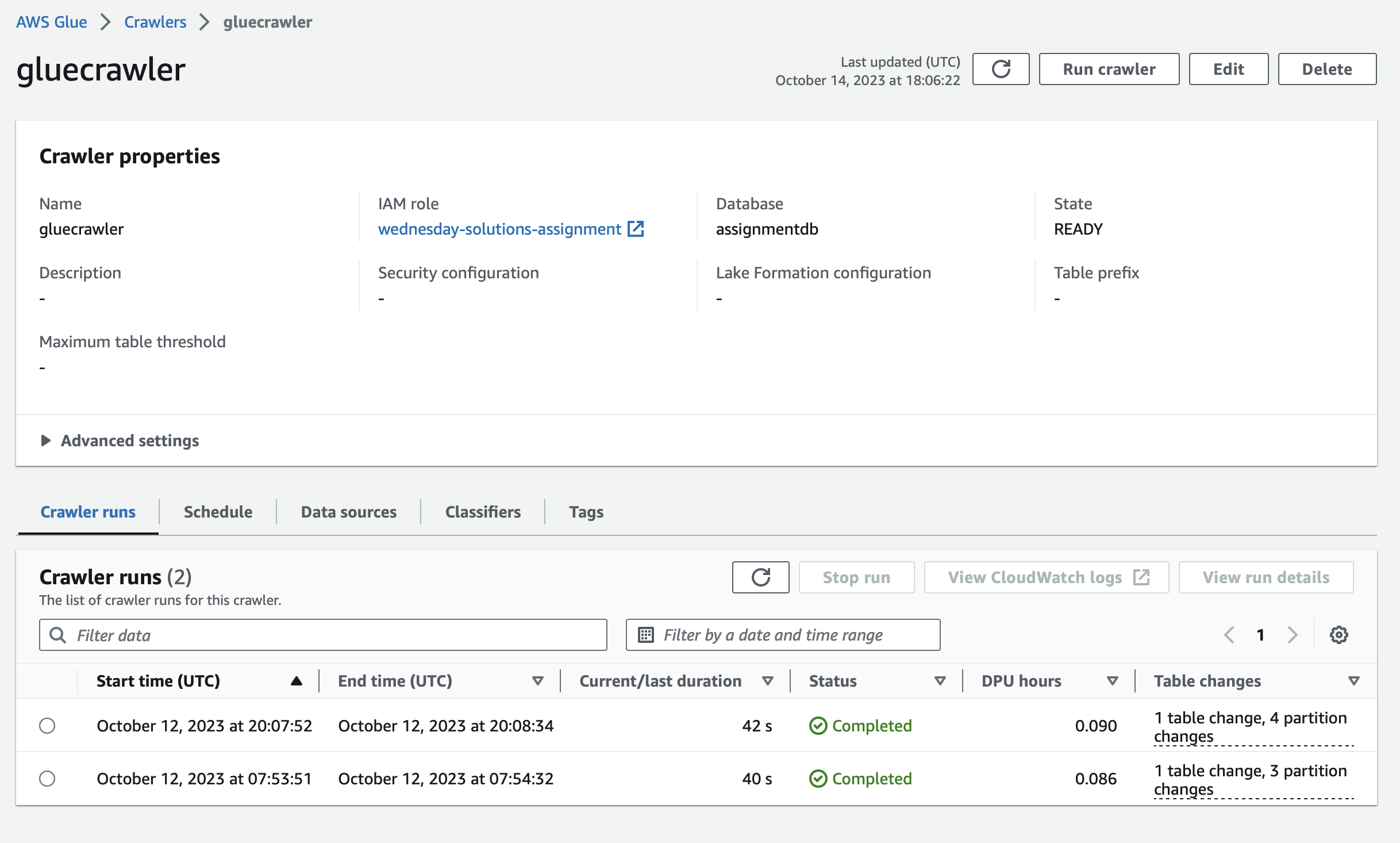
We Upload the Dataset insurance.csv into the data folder of this bucket.

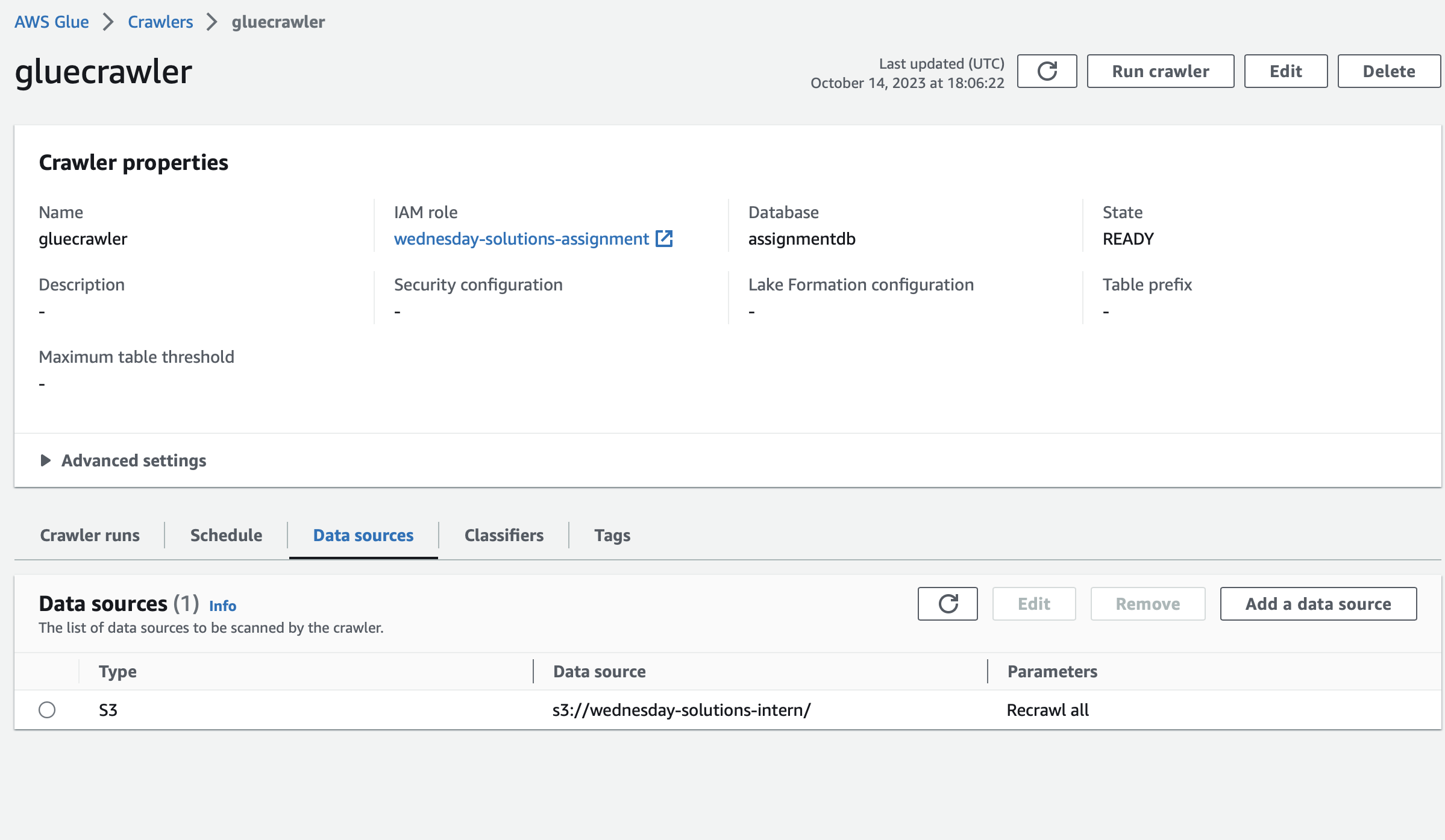


2. After this we create an IAM role for this project giving access to all the files and different AWS services like S3, Redshift, Admin Access….

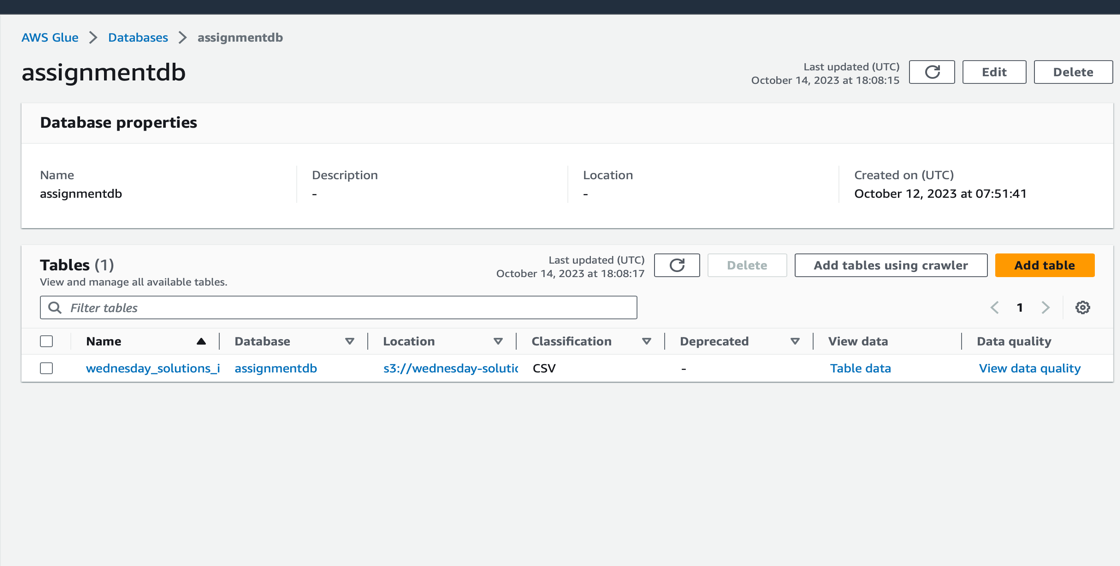


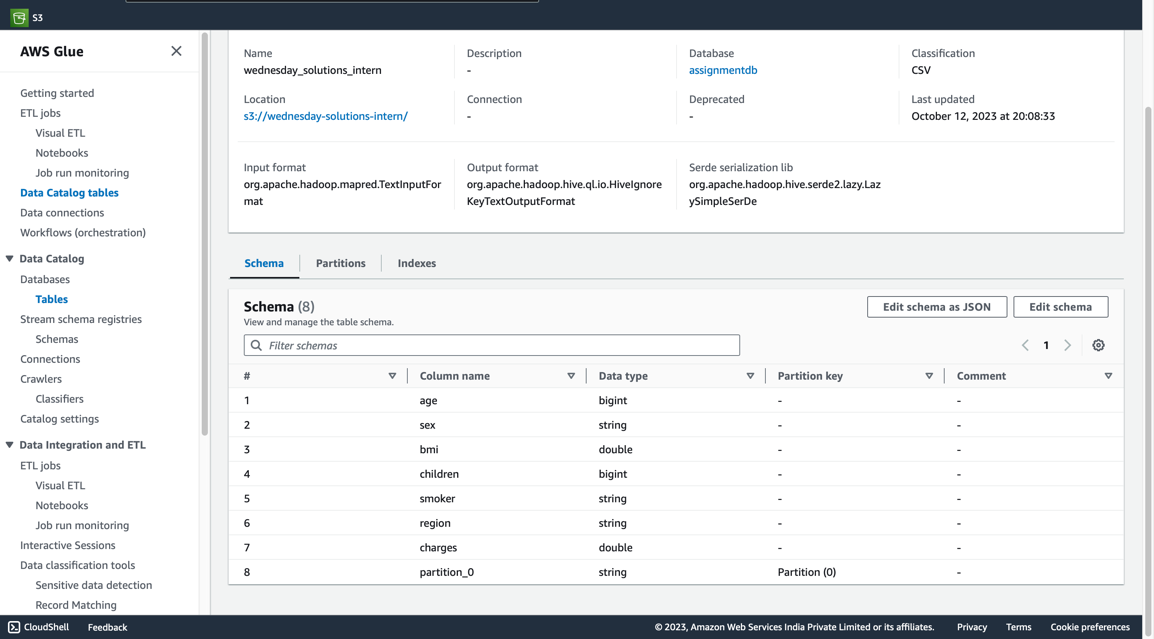
3. After this we move on to make a Crawler to go through the S3 Bucket and the Crawlers job is to make a Database and the tables in it will be extracted from the Source





4. After the crawler does its Work , Database assignmentdb is made containing the tables.

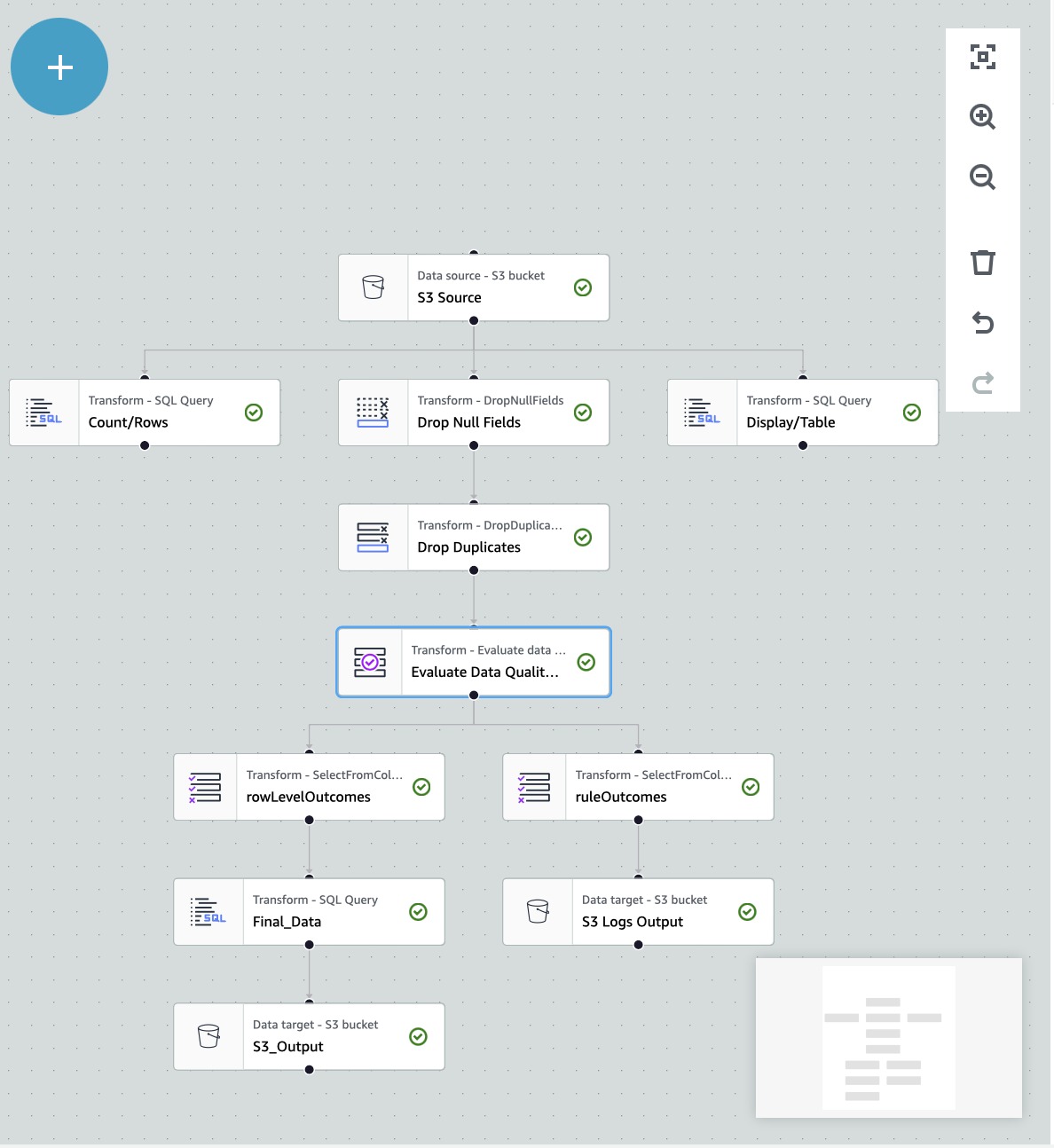




5. Making an AWS Glue Job using these Tables. So we create a Glue job with ETL and Python Scripts using Pyspark Framework.

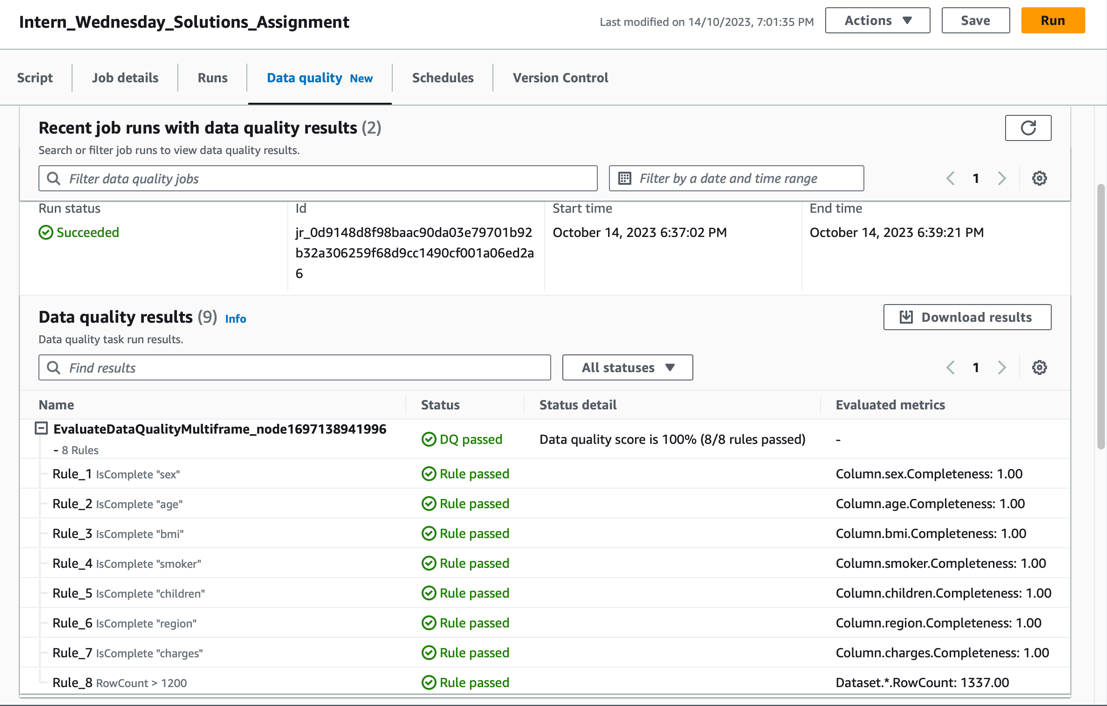
For the Entire ETL Process we use a pipeline that looks something like this

THIS IS THE VISUAL REPRESENTATON OF THE ETL



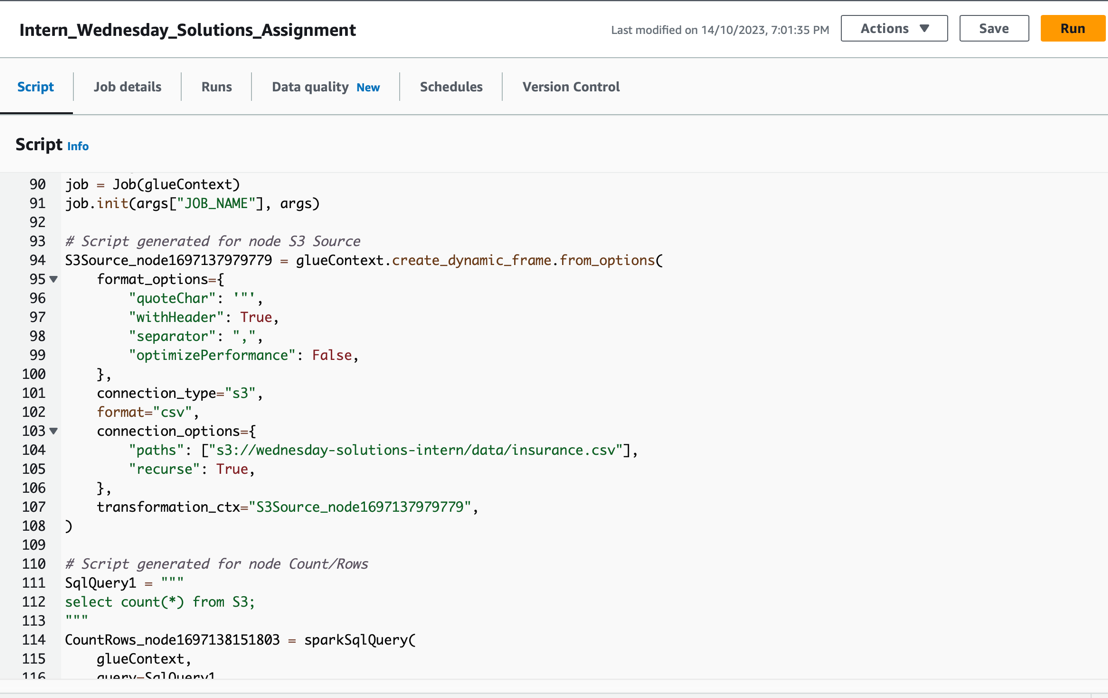
THE ETL (Extraction, Transformation, Loading) job consists of :

1. EXTRACTION (Getting the Data from S3): The First part of the Process is extracting the Data which has been done here.
2. TRANSFORMATION (Transforming the Data from the Input Source):
3. Dropping Null Fields: Fields having “”, null as column name.
4. Dropping Duplicates in the Data: Removing multiple records having the same field name
5. Replacing Null Values: In this dataset there are no null values, it is cleaned but in General for numeric data we use **Mean** to replace empty values and for Categorical values we can have **Mode** of that Column or have some **Custom Rules set for that.**
6. Evaluating the Data Quality on the Data using some rules.
7. See if there are Null entries in the Data for all the columns. In this Data there are no null entries. For that we have used isComplete “column” for all the Columns.
8. Also see if the final created Data after all the pre-processing stage has number of rows > 90 % of the initial dataset ie Number of rows > 0.9\*1337 … that is ~1200

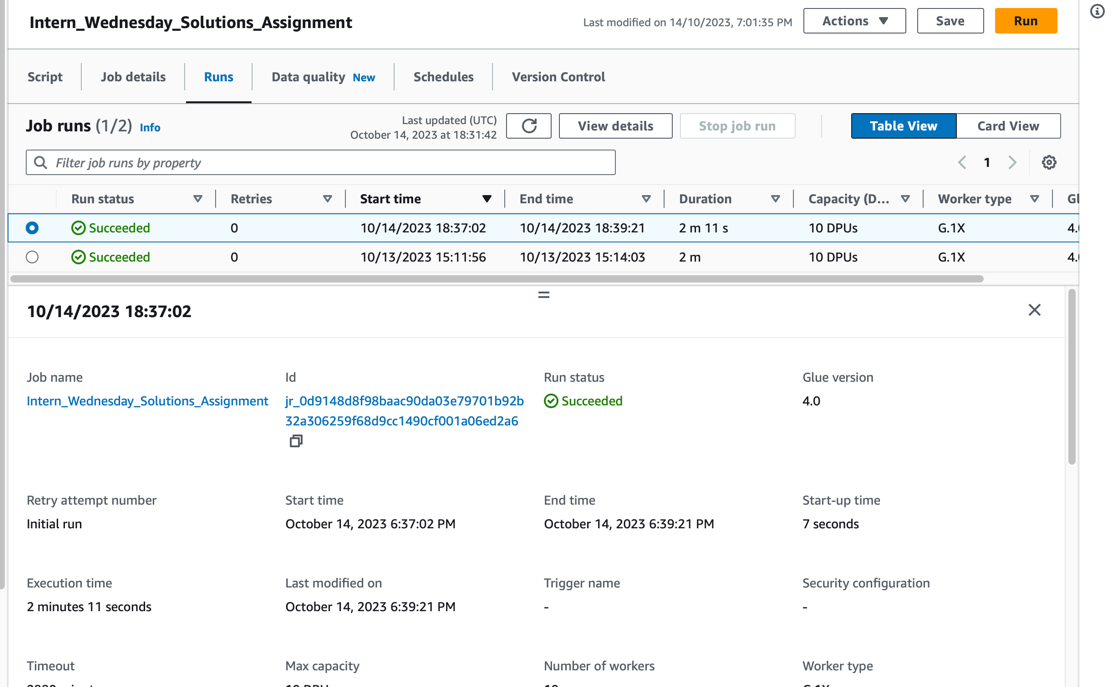


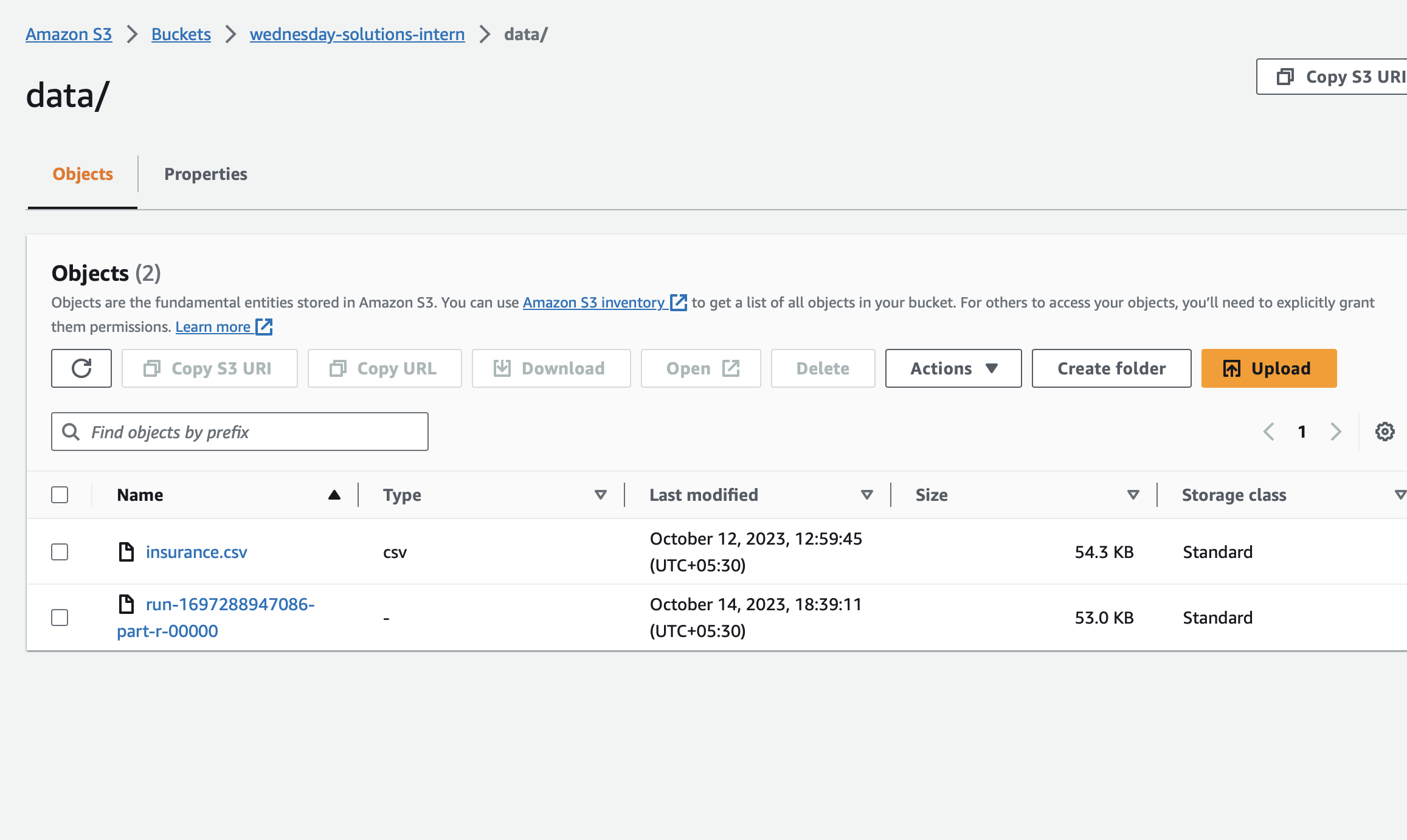
1. The Logs of Data Quality are stored in the logs folder in the S3 Bucket

The Entire Python Script Handling this is uploaded onto the Github Repo and the Code is available in Intern\_Wednesday\_Solutions\_Assignment.py



1. Finally the Data is Output in the S3 Bucket data folder after the JOB gets fully executed.

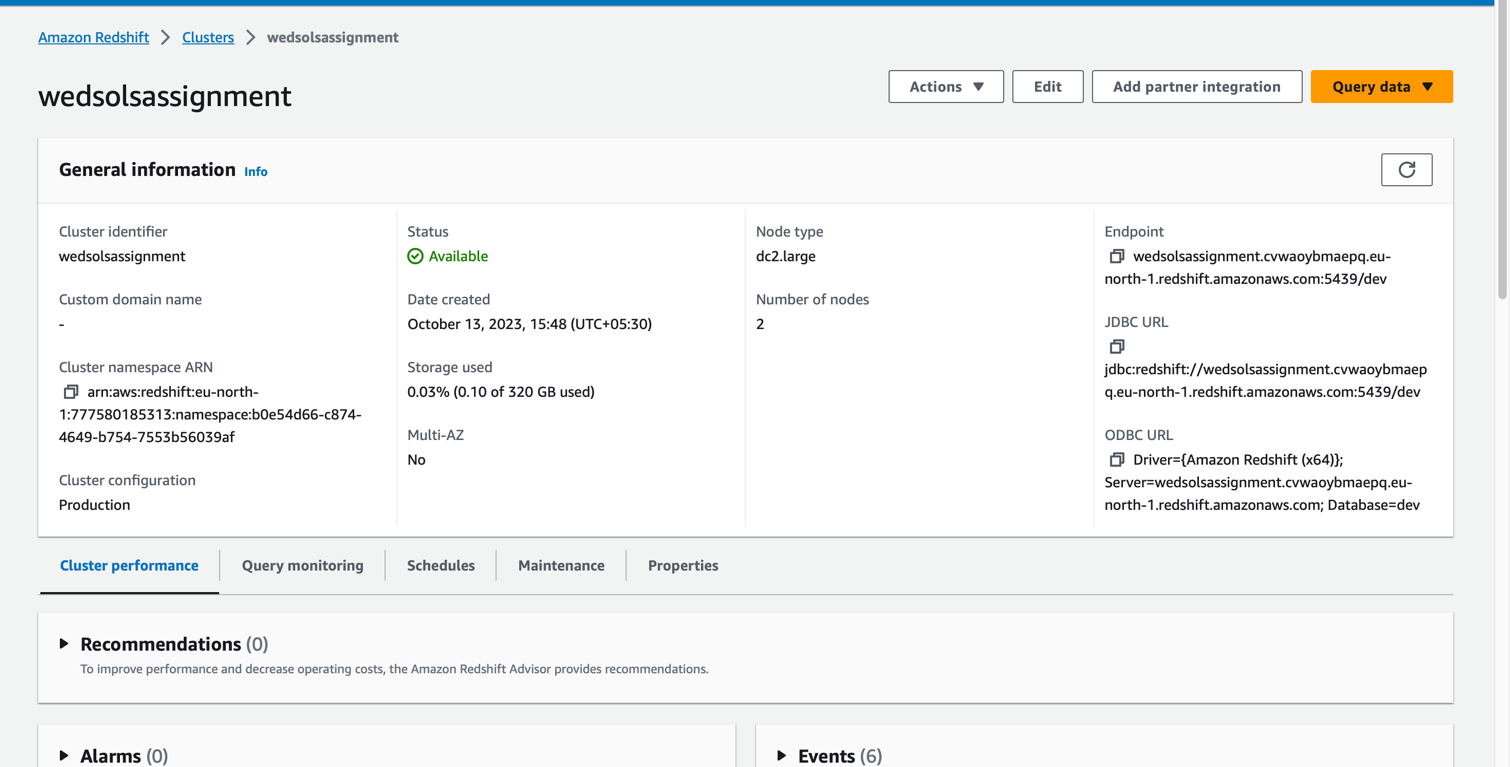




Run-1697288947086-part-r-00000 is created and it’s the output file

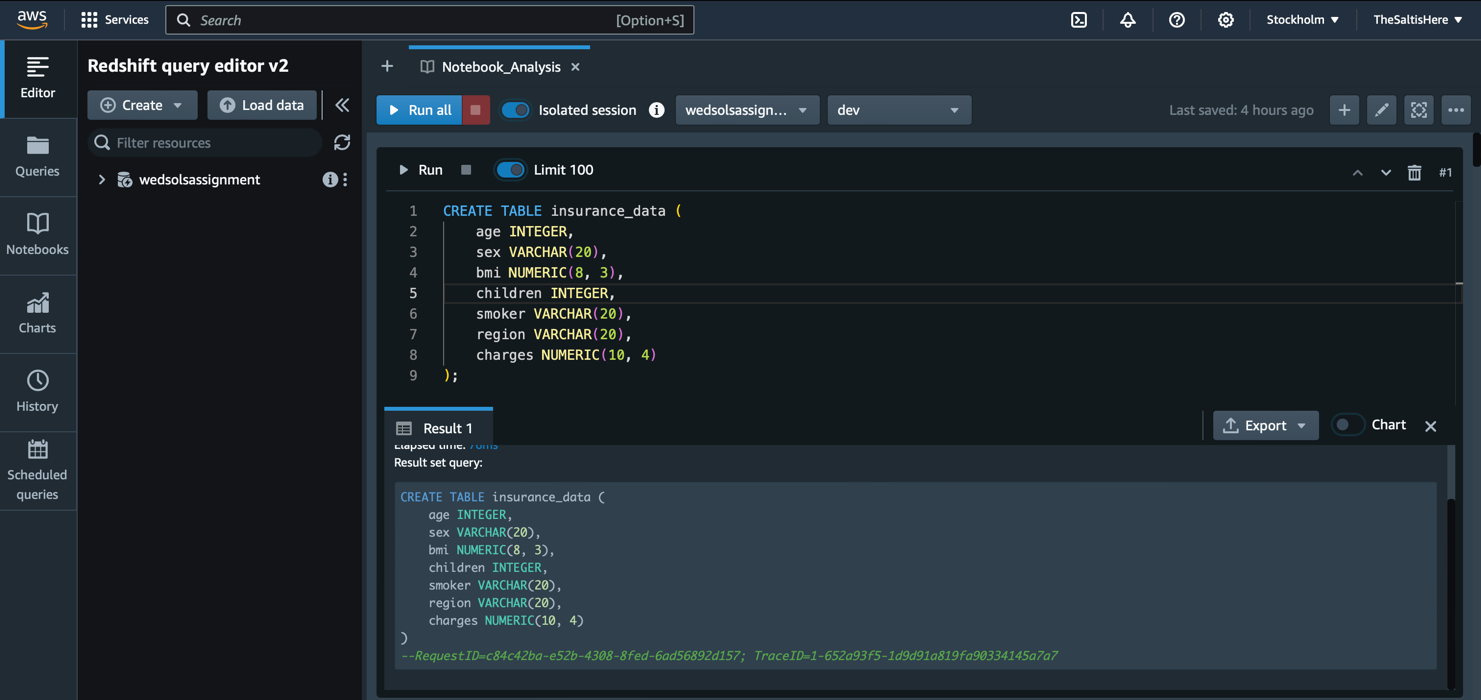
This File will be Used Further on.

1. LOADING : The Output file is loaded onto AWS Redshift. For this I have created a Cluster which is connected to the DB and I have chosen AWS Redshift as it can give SQL Query as well as Visuals for understanding the Output table . It will help us in understanding the Analysis Phase.

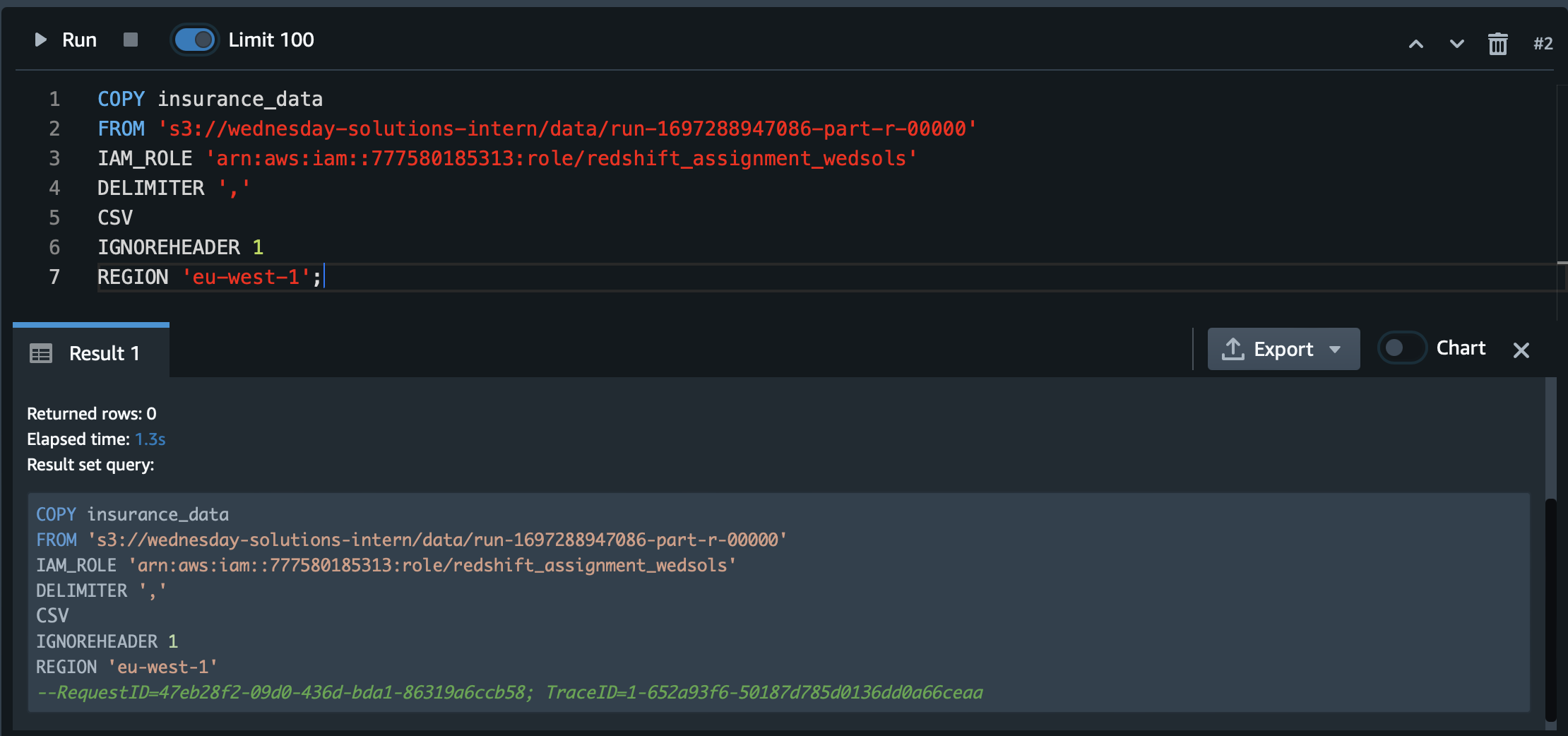


DATA ANALYSIS QUERIES:

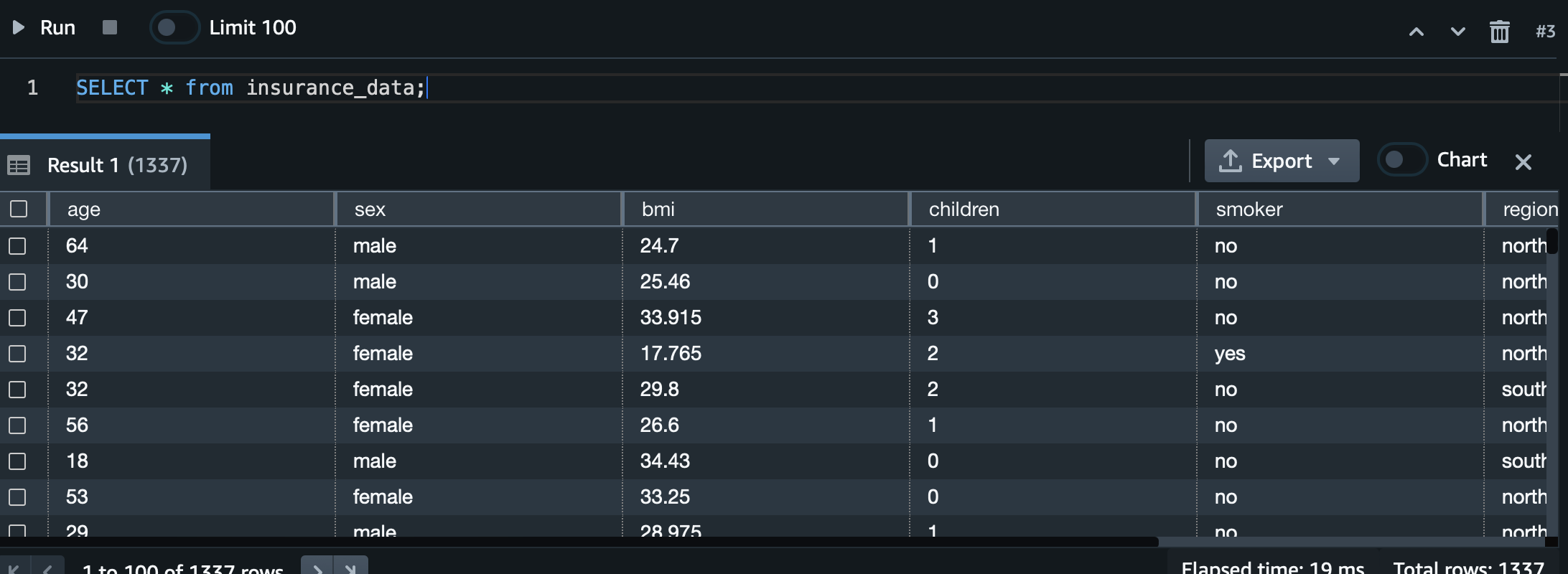
1.Now I have gone to Query Editor V2 in Redshift and created a table insurance\_data and successfully created it.



2. We now connect to S3 Bucket to get the Output Data.



3.This is a look at the Data and we see that table is displayed.

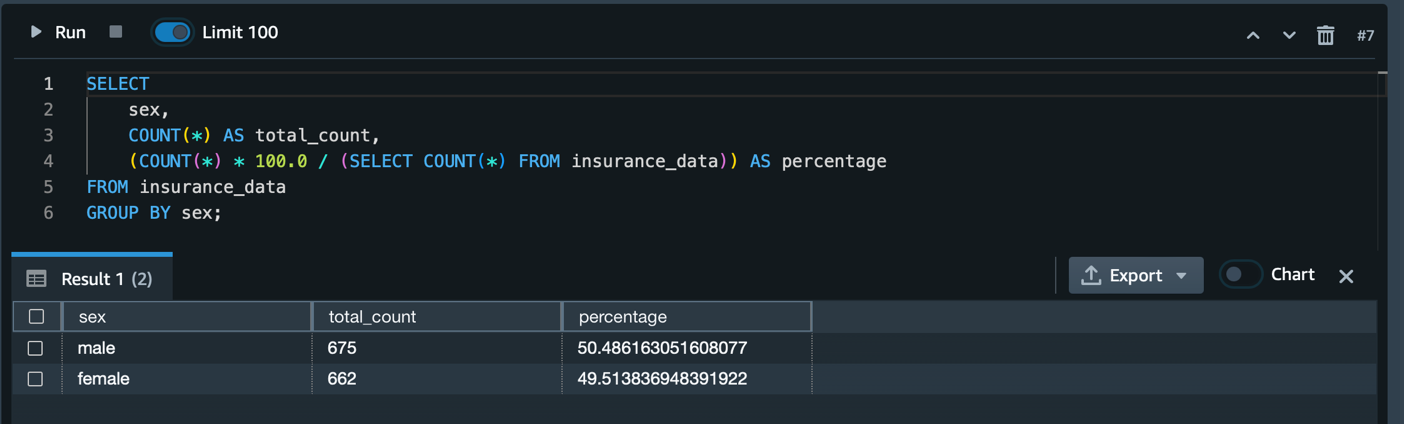


1. Taking a look at Statistics for all the Columns: Age,Sex,BMI,Children , Smoker,Region,Charges

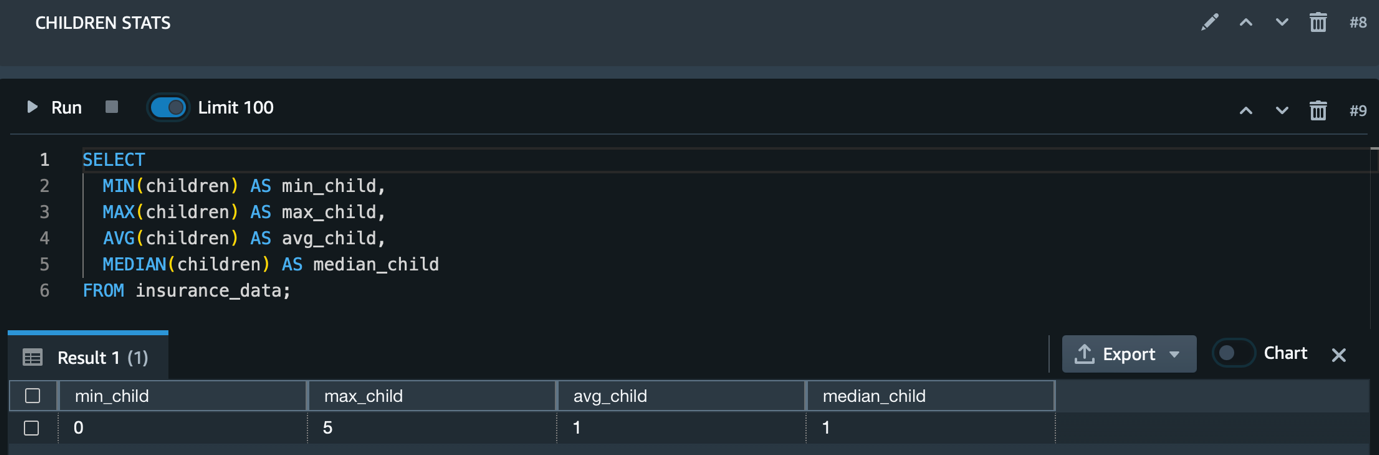
AGE

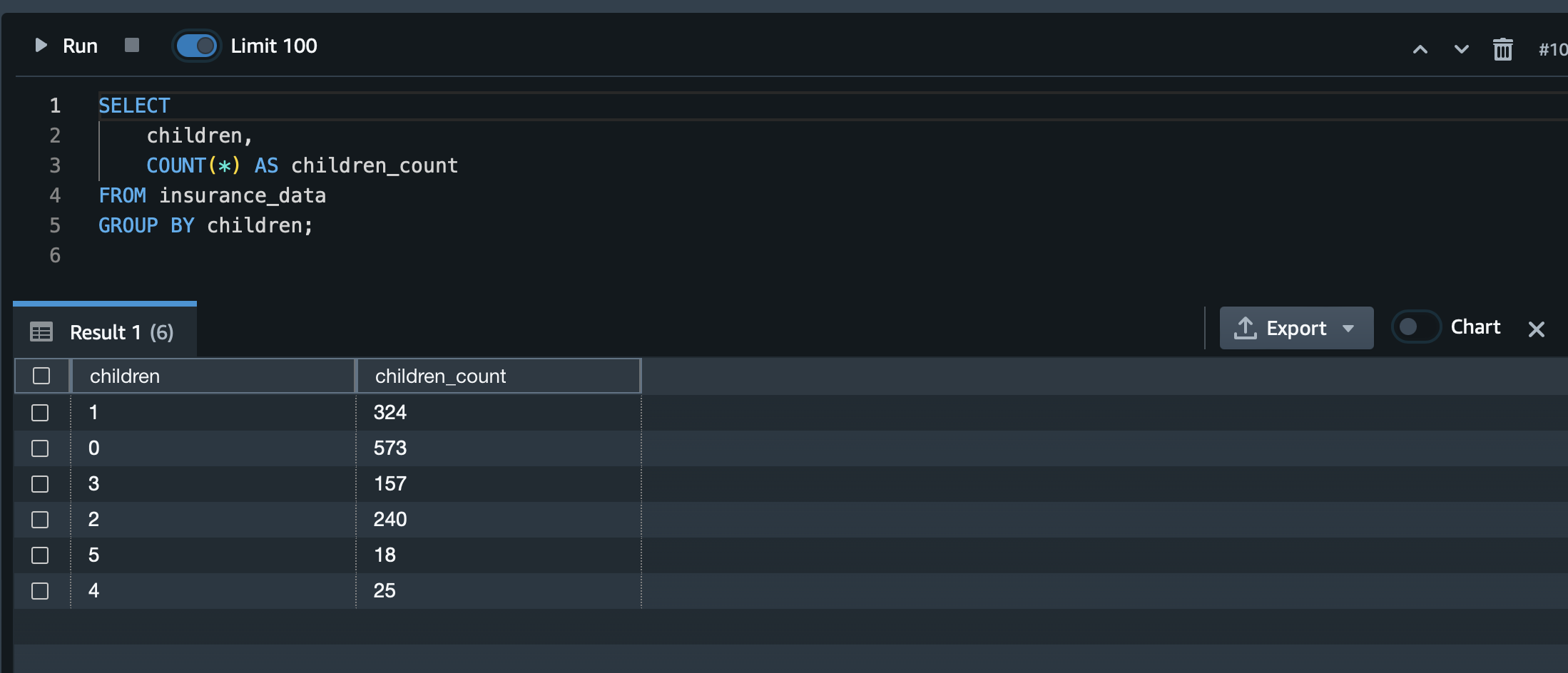


SEX

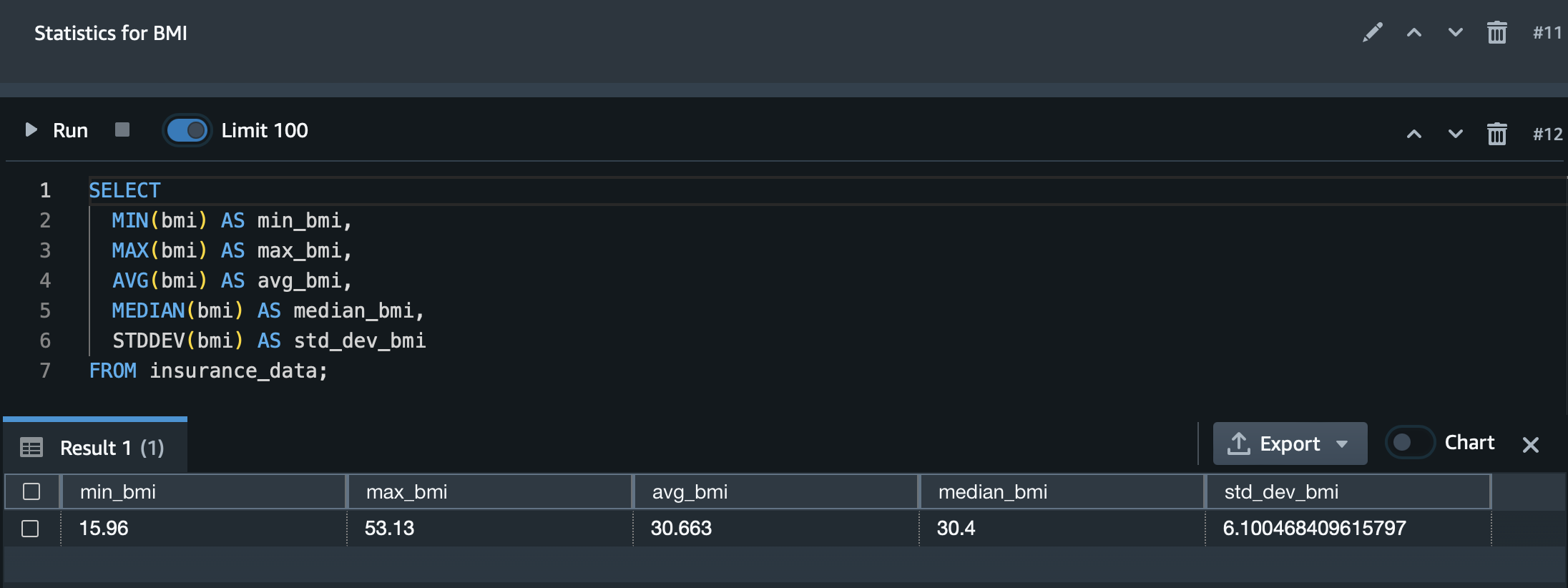


CHILDREN

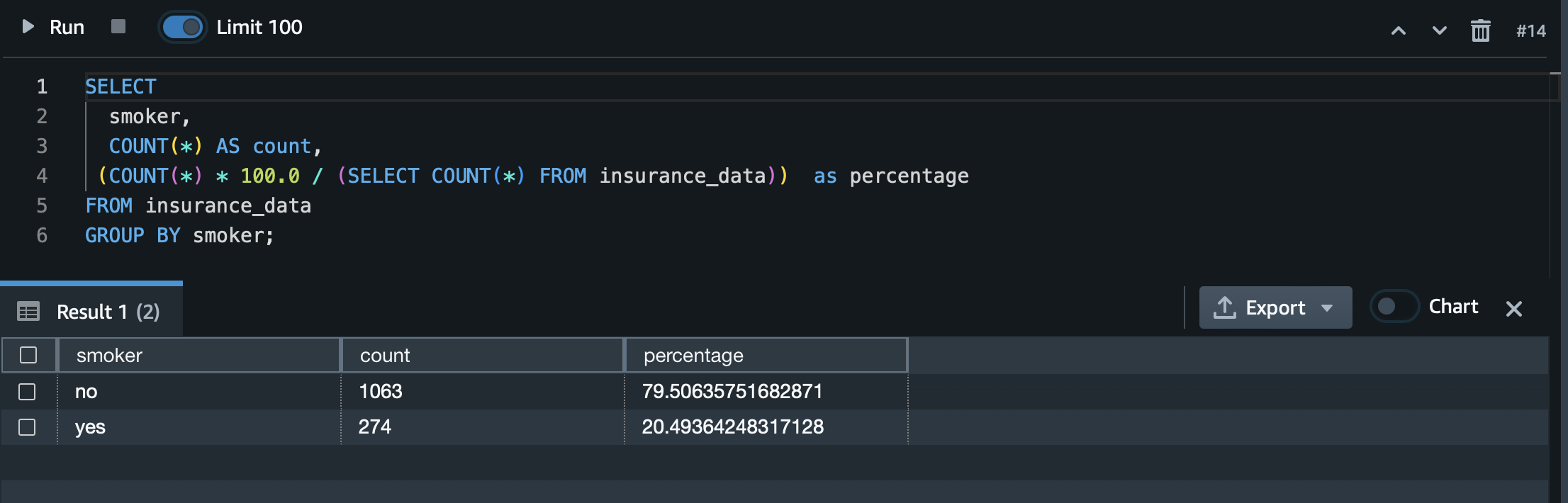




BMI

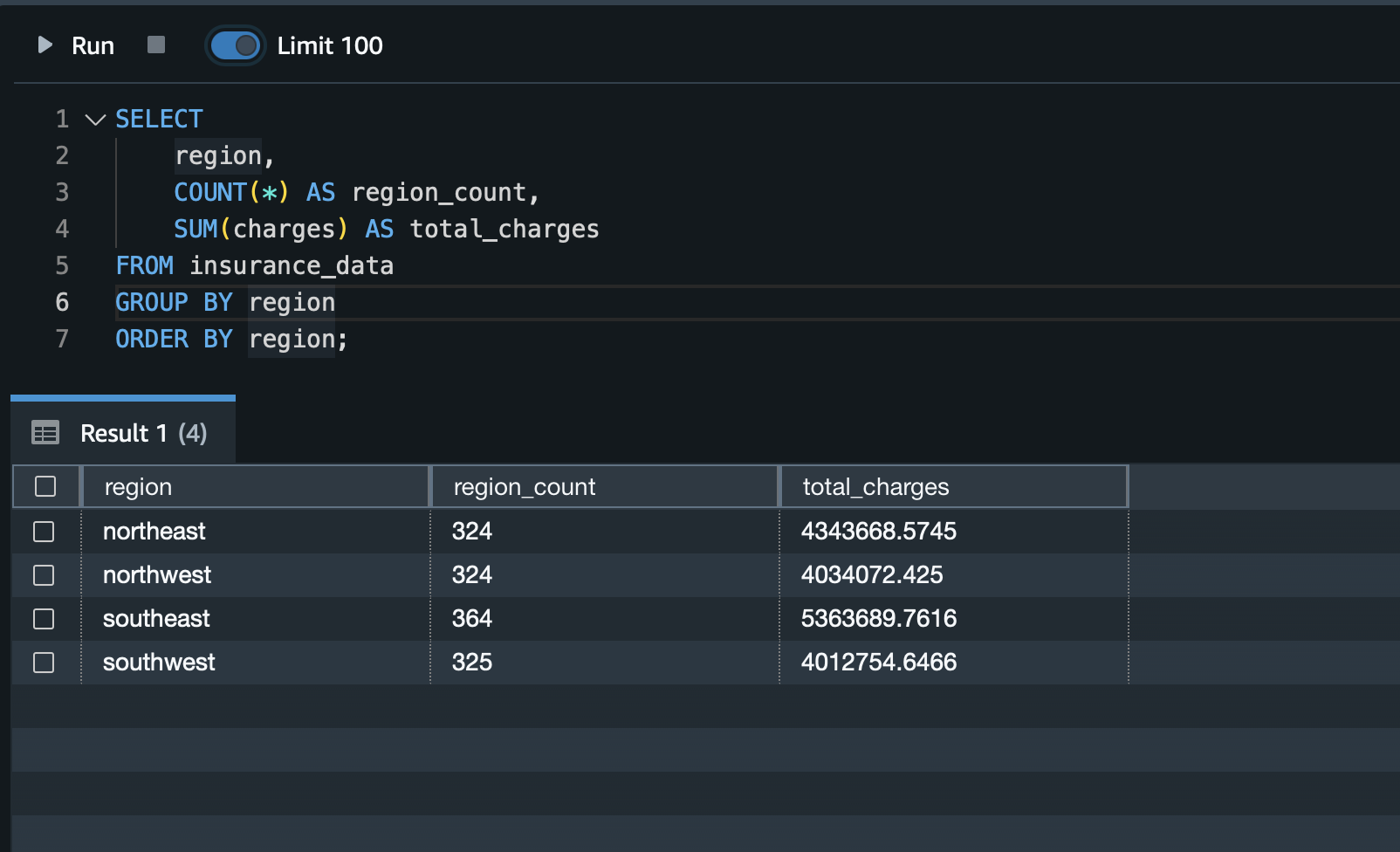


SMOKER

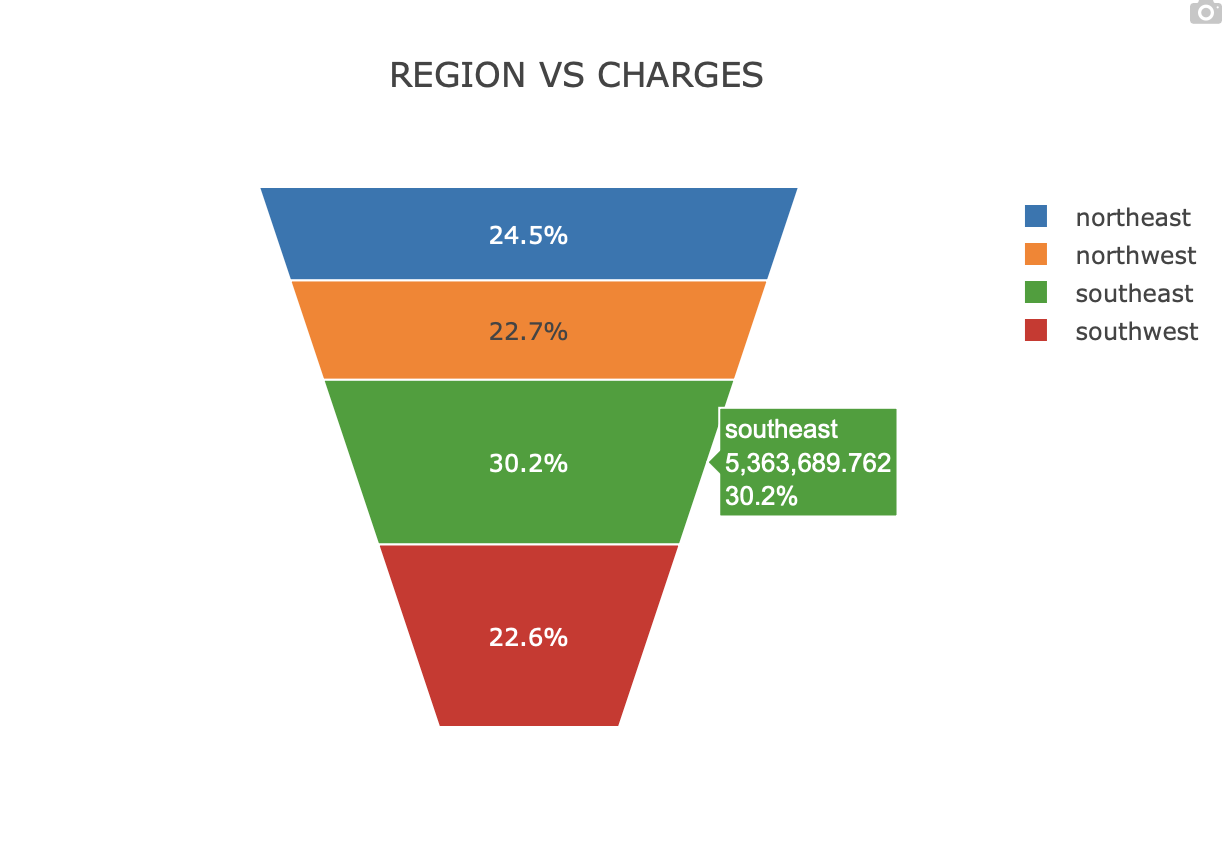


REGION

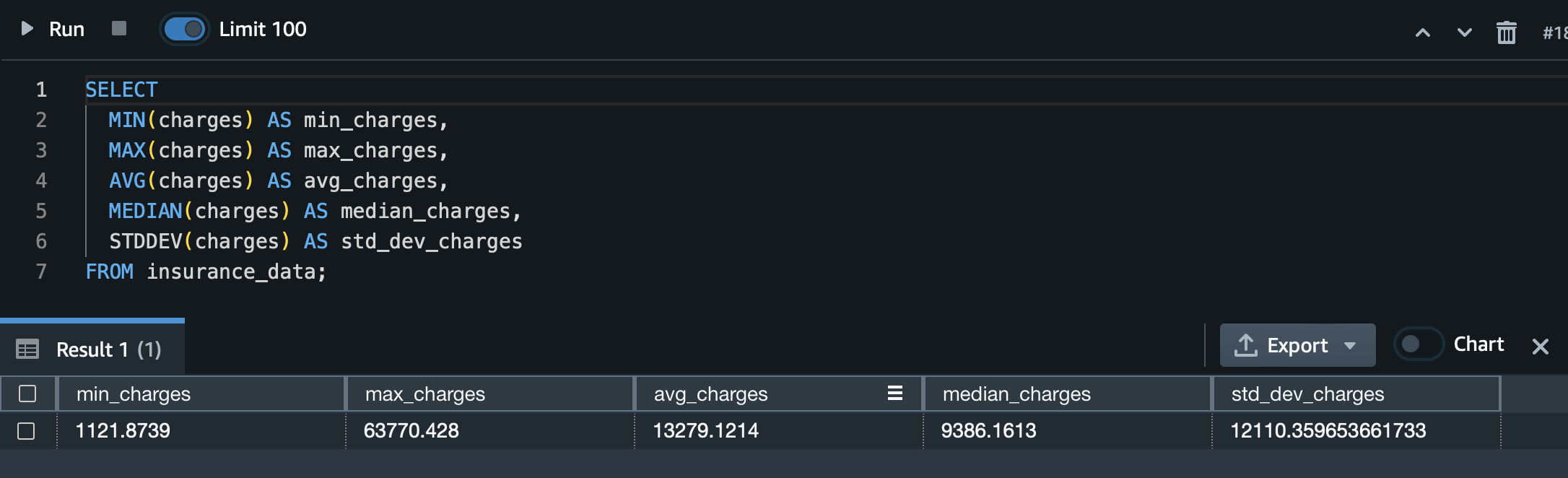
A table showing the region, region\_count and the total charges as a sum aggregate in the region.



A FUNNEL AREA PLOT FOR THE SAME TOO…

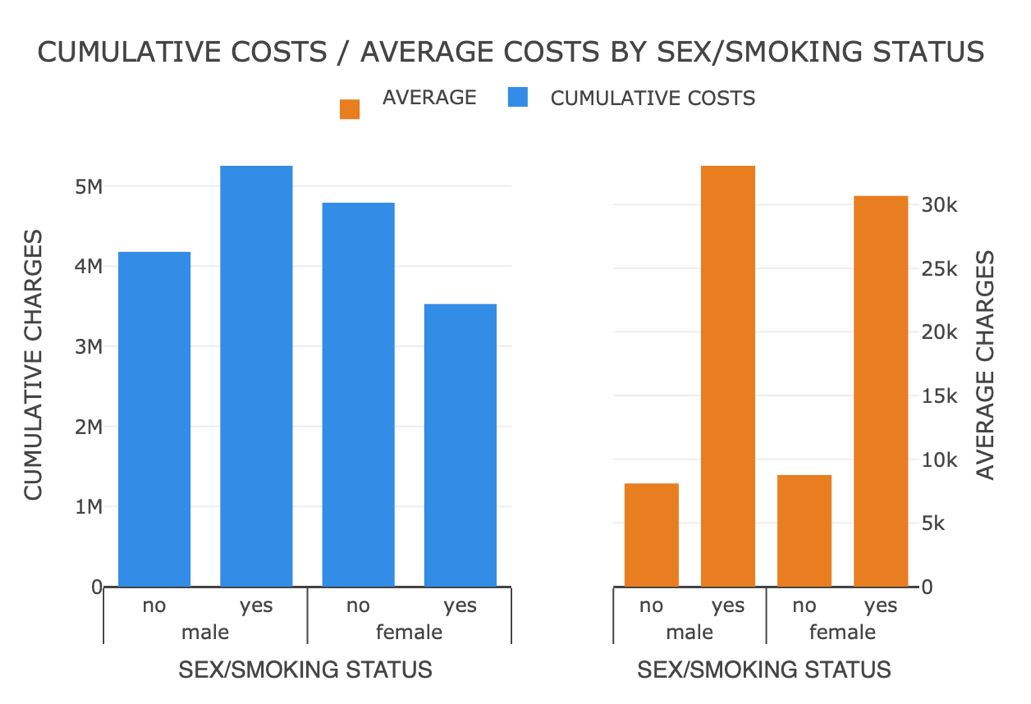


CHARGES



1. Having a look at Smokers vs Non Smokers Mean and Cumulative Costs

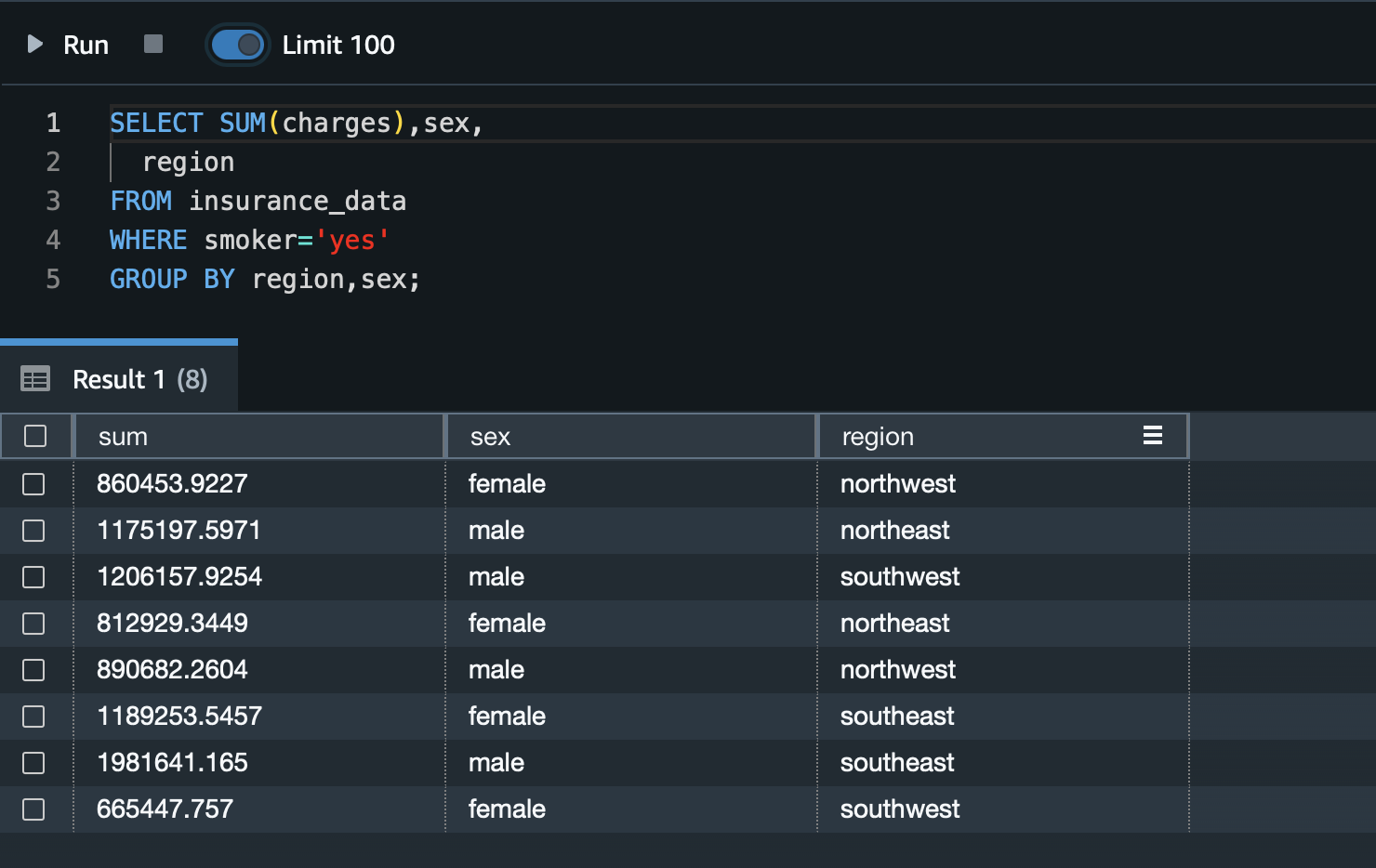


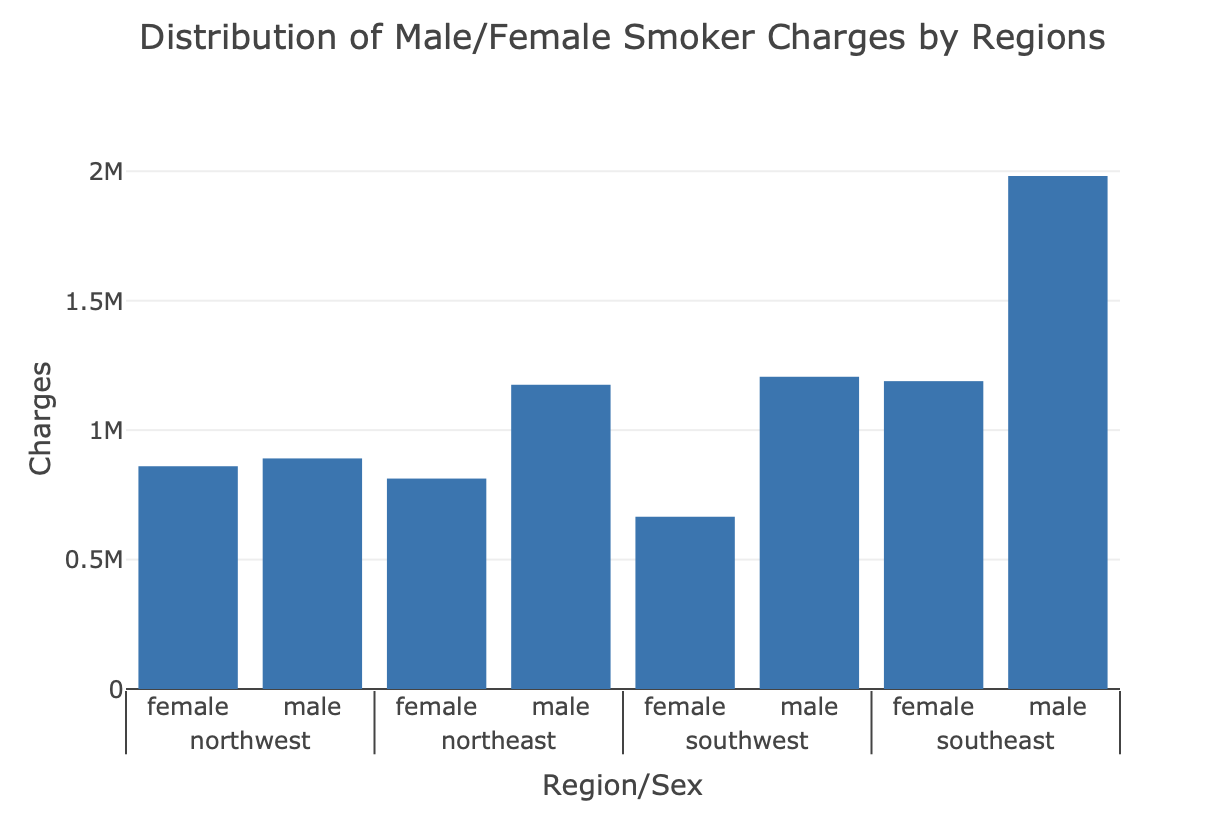


So we can see that on average, males have accumulated higher charges cumulatively. On average, too, among smokers, on the other hand, in non-smokers, females have more higher average/cumulative charges than men.

So we can understand that non-smoking-related issues are more in females, while for men, smoking-related issues are more.

6.Lets look at the region wise Distribution of those who are Smokers…





We understand by this plot that the regions where males/females can be targeted based on smoking activity, especially in the Southeast. Male smokers are more, so insurance premium plans should be higher here for males..

7.Lets have An Analysis on BMI

So Certain Rules for BMI :

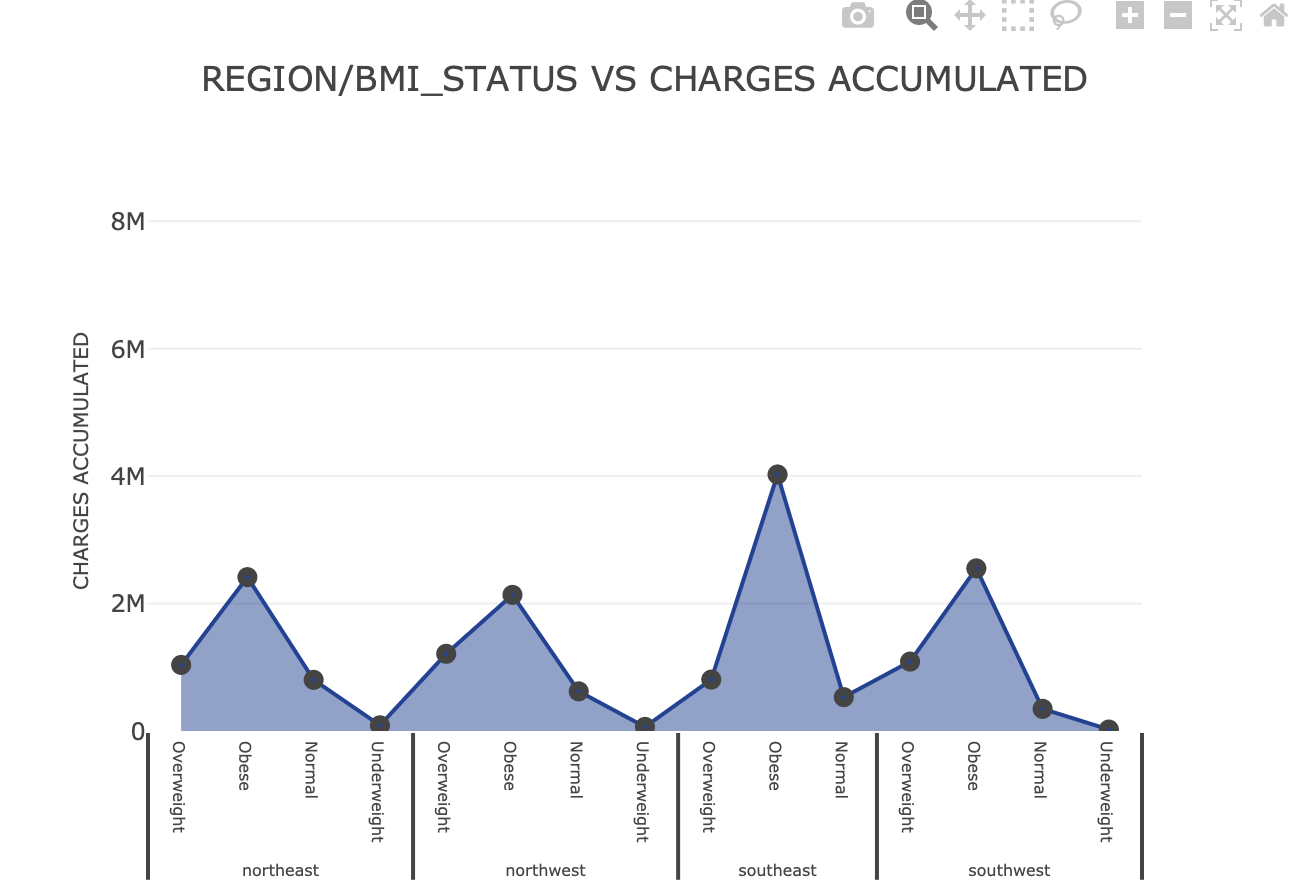
UNDERWEIGHT < 18,

HEALTHY BMI BW 18-24.9

OVERWEIGHT BW 25-29.9

OBESE AFTER THAT

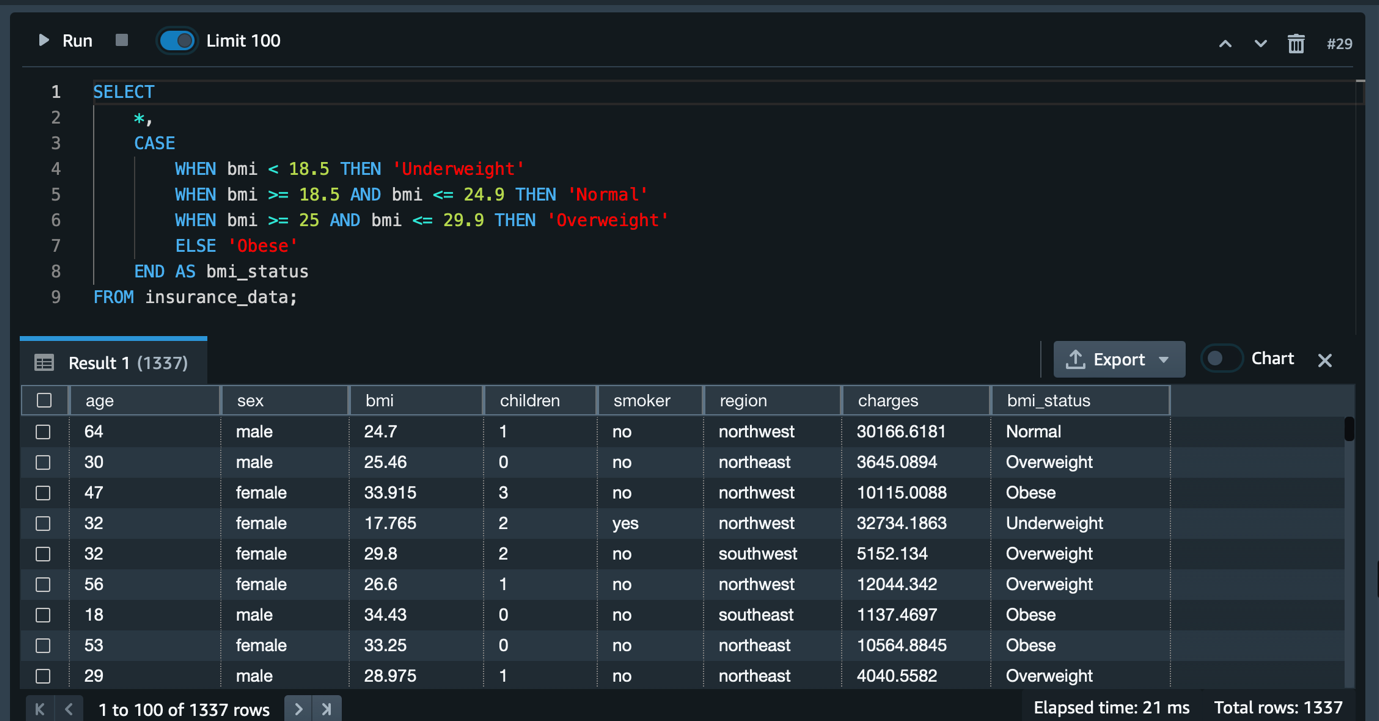
So we will try to build a column "BMI\_STATUS" filling values based on the above conditions and create some visuals.

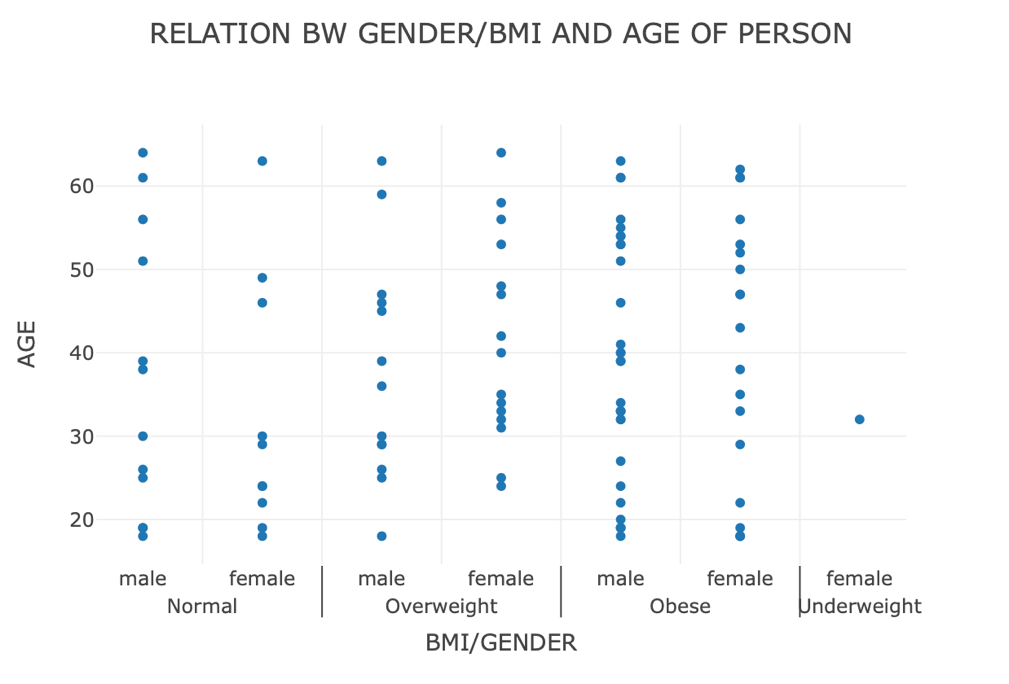


We observe here that the peaks are observed in obese people in all the regions. The order goes like this for the charges accumulated: obese > overweight > normal > underweight.

Hence, the premium plans should be made more expensive for obese individuals and then decrease gradually for overweight, normal, and underweight categories.

8. Creating an Analysis on BMI/ Gender vs Age of the Person





We can see here by this plot that the obese category is in males mostly in the region of the 45-55 age bracket. In females, by the scatter plot, we see that obese individuals mostly lie in the range of 45-50.

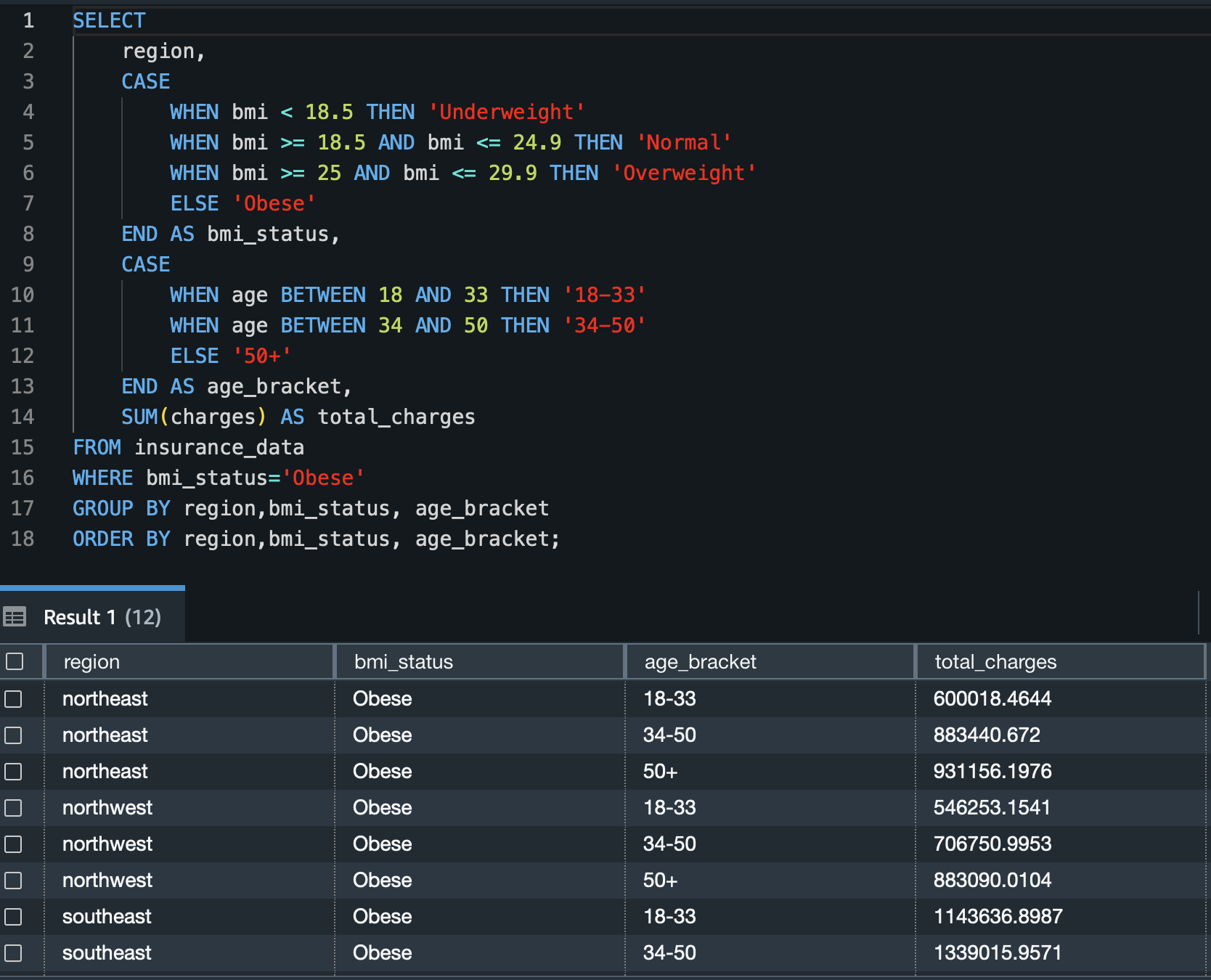
9. Creating an Analysis on Age

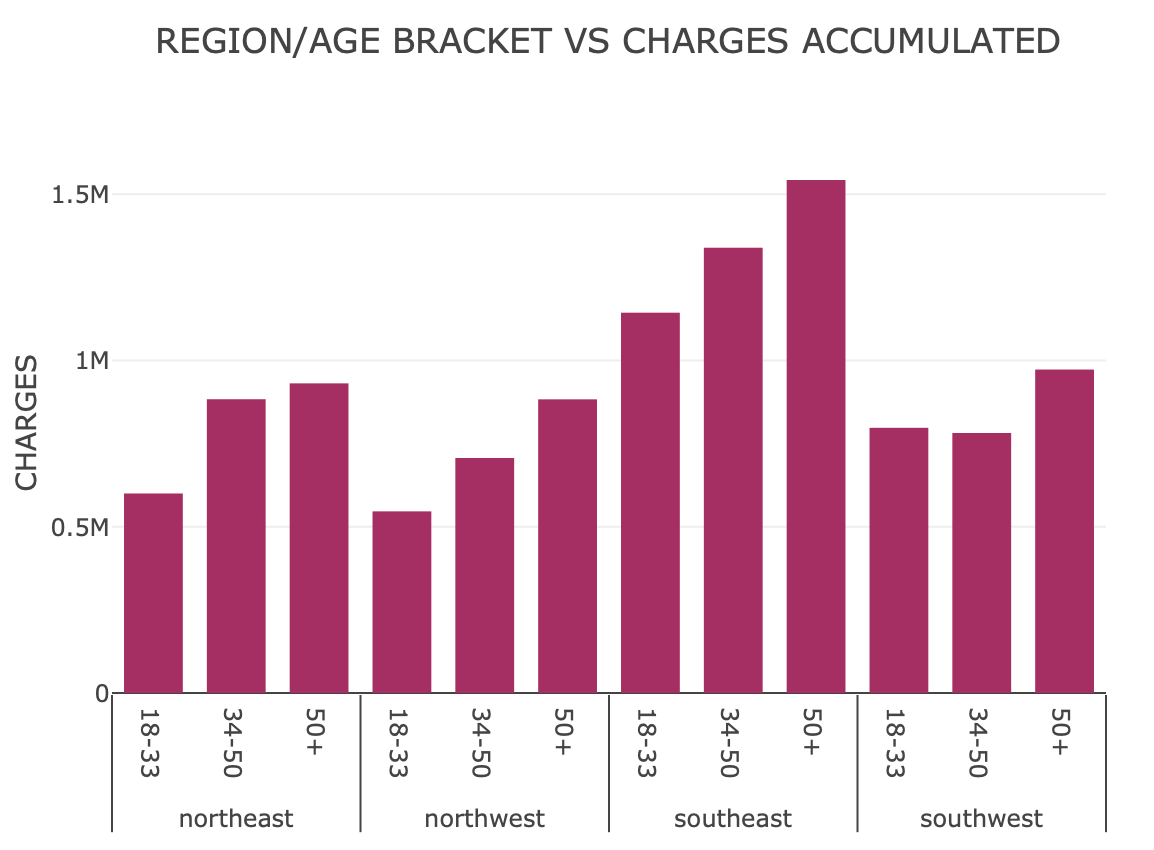
Let's make age brackets as follows:

18-33

34-50

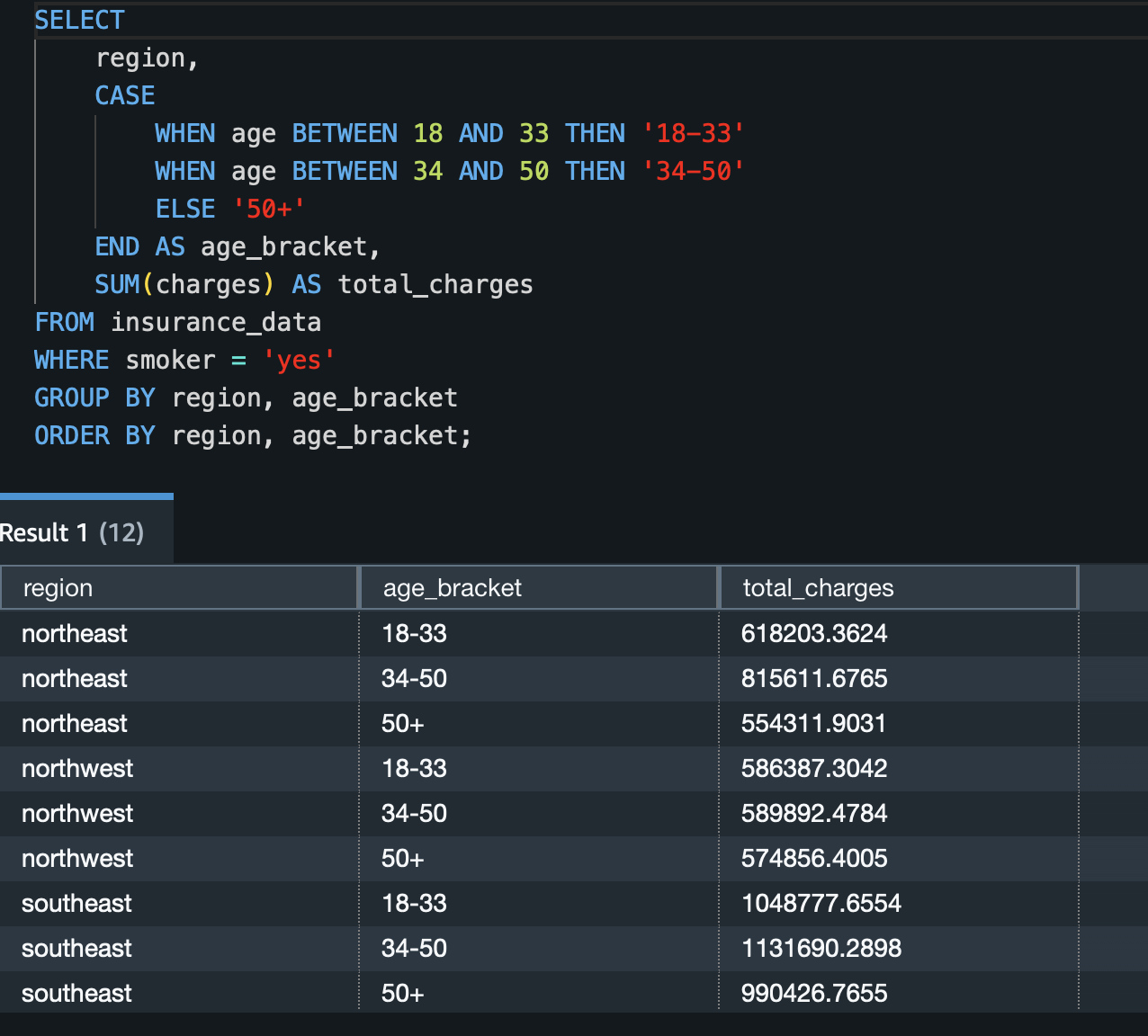
50+

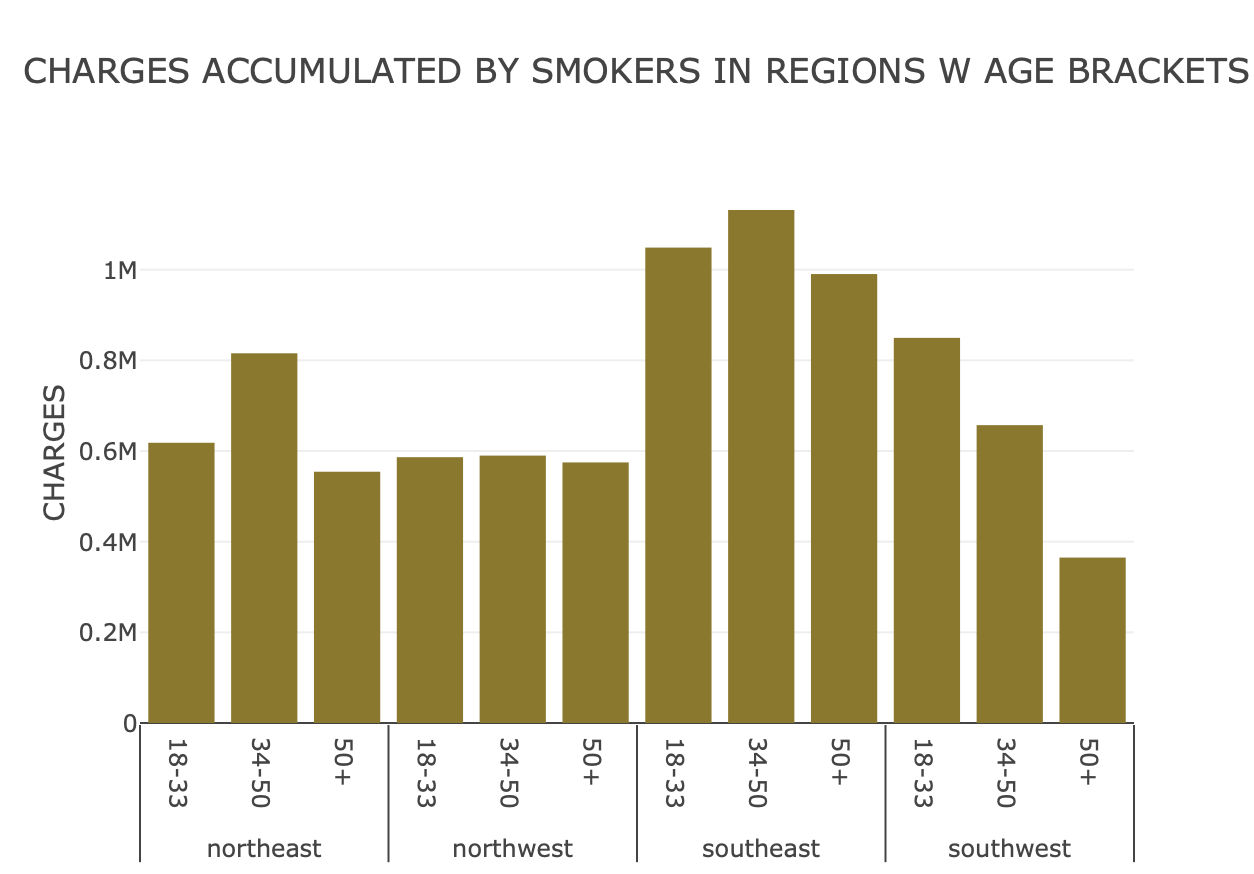
This segregation has been made by min-max age calculation.



This plot shows us the relationship between region, age bracket, and charges accumulated for **obese** people. For all regions, the 50+ age bracket needs to have a higher premium on health plans. This plot can be automatically changed for different BMI\_STATUS categories.

10. ANALYSIS OF SMOKERS WITH REGION /AGE BRACKET VS CHARGES

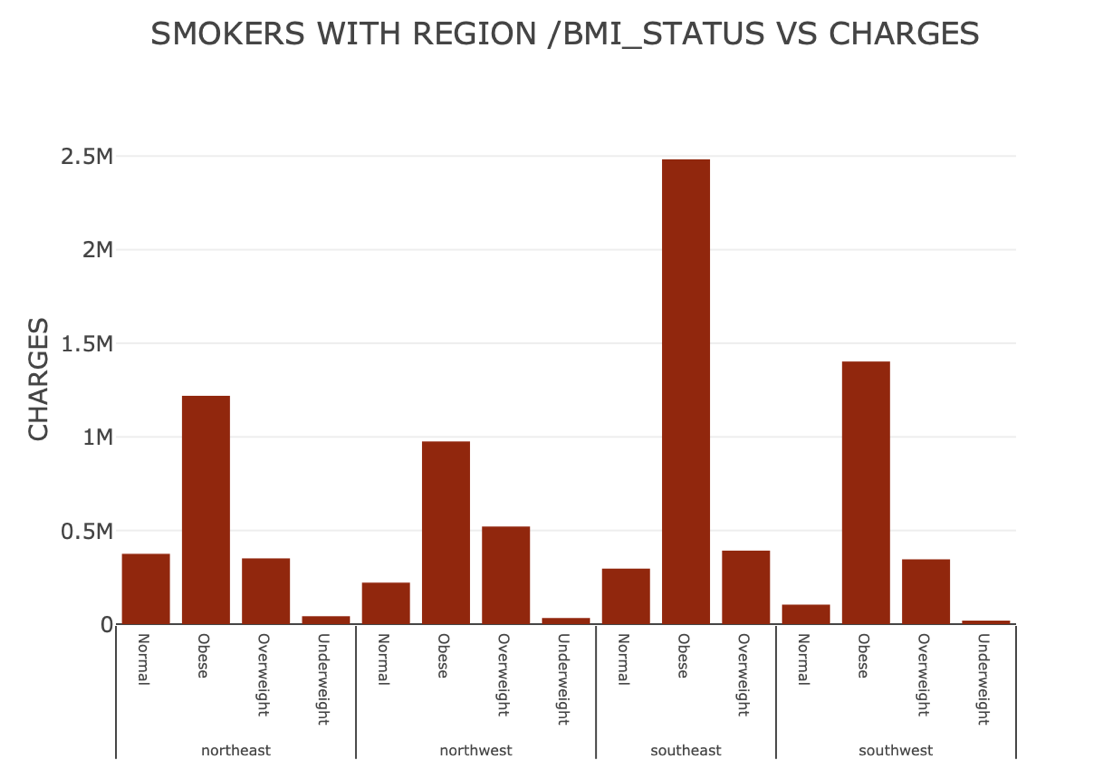




Analysing this plot, we can see different age -brackets in different regions having higher charges than the rest. According to needs, plans can be made out for smokers in different regions based on age bracket too.

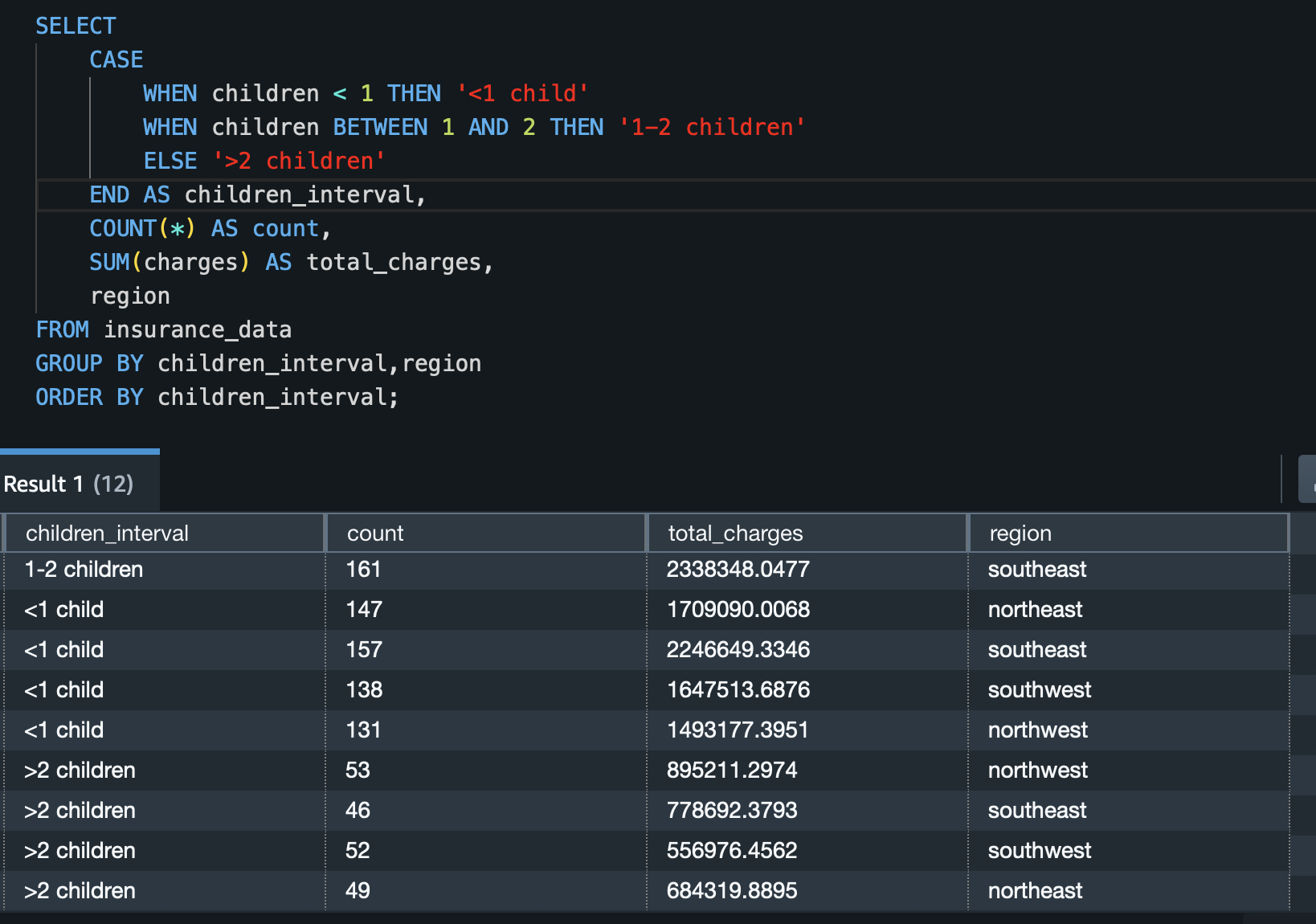
11. ANALYSIS OF SMOKERS WITH REGION /BMI\_STATUS VS CHARGES

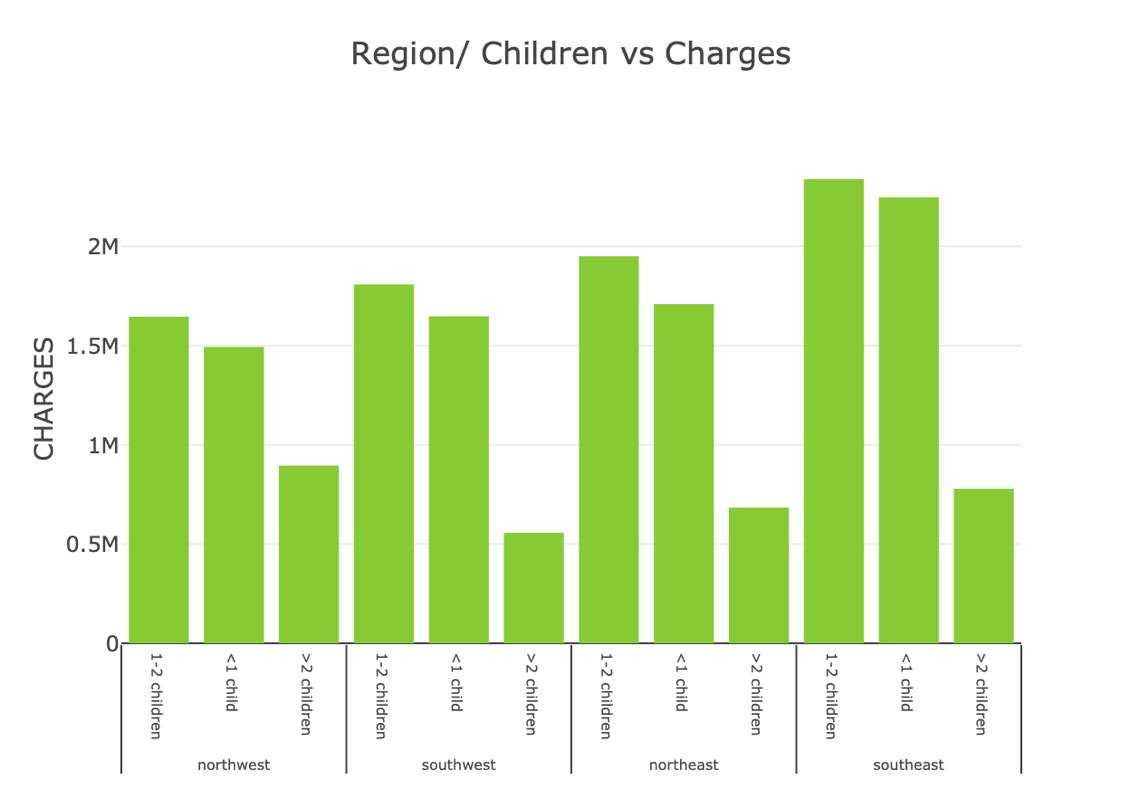




Analysing this plot, we can see that Obese people who are smokers have relatively more charges than the remaining groups in all the states and are at higher

12. ANALYSIS OF SMOKERS WITH REGION /BMI\_STATUS VS CHARGES





Observing this we can see that People having 1-2 Children are accumulating more Charges. Hence , more premium should be paid for plans having them and Vice versa.

All the above SQL queries are written in a script called

SQL\_Queries\_Analysis.ipynb

This Script can be accessed by opening in a .ipynb file. This file is directly output from the Redshift query editor.

OUTPUT :

Our marketing efforts should be strategically directed, primarily based on the top-5 regional analyses:

1.In various regions, it is evident that males who smoke should be a key focus for our efforts in promoting better insurance premium plans. This demographic tends to incur higher hospital bills. Additionally, we should consider running targeted social media campaigns to reduce smoking rates, thereby enhancing life expectancy.

2.Across all regions, it's clear that individuals with obesity tend to accumulate higher medical expenses. Therefore, the insurance company should offer tailored insurance plans with increased coverage for this group.

3.Across all regions, we've observed a consistent trend where healthcare costs rise with age. To align with this trend, our marketing efforts should be more attentive to elderly individuals, emphasizing premium plans designed to meet their specific needs.

4.We can tailor our marketing efforts further by considering the Body Mass Index (BMI) status of individuals, such as obese, normal, underweight, etc. Different age groups can be targeted with insurance plans customized to their BMI status.

5.An interesting observation reveals that families with 1-2 children tend to have higher accumulated charges, with the order of preference being 1-2 children, followed by <1 child, and then >2 children. To capitalize on this trend, marketing efforts should be structured accordingly to target these demographics.

Learnings and Challenges faced

Learnings:

End-to-End Data Engineering: This assignment taught the complete data engineering process, including data ingestion, transformation, and loading into a storage service.

AWS Services: Utilizing AWS services like S3, Glue, and Redshift for ETL and analysis showcased the power of cloud services in handling data tasks.

Data Quality Checks: Checking for null values and verifying transformation output is vital for data reliability.

SQL Analysis: Using SQL in Redshift enabled in-depth data exploration and insights.

Targeted Marketing Strategy: Data analysis helped create a strategy for tailored insurance plans and marketing based on demographics.

Challenges Faced:

Data Cleaning: Dealing with messy data, especially missing or inconsistent values.

AWS Learning Curve: Navigating AWS services and their configurations can be challenging.

ETL Complexity: Designing efficient and robust ETL processes can be complex.

Query Optimization: Optimizing SQL queries for large datasets in Redshift requires expertise.