**ADTA5240 :** Harvesting Storing and Retrieving Data

**Final Project :** Analyzing Accidents Using NHTSA Data

**Member:** Dontireddy Manasa

**Analyzing Accidents Using NHTSA Data**

**Introduction**

Road safety is a subject of critical concern in many states and localities. Improving safety measures and enacting better vehicle regulations comes from being able to analyze and understand the factors in road accidents. In my project, I focus on a comprehensive analysis of accident data, utilizing two key datasets provided by the National Highway Traffic Safety Administration (NHTSA): the Fatality Analysis Reporting System (FARS) data and vehicle safety-specific data from the NHTSA’s Open Data Portal.

I aim to determine the correlation between various vehicle characteristics, such as make, model, and year, and accident occurrences to identify trends that could help establish better safety standards and procedures. For this analysis, I leverage the FARS dataset, which contains critical information on fatal accidents, such as accident circumstances and contributing factors.

By merging this data with vehicle safety-specific data, I examine how different vehicle types and their respective safety ratings affect accident severity. This analysis helps identify whether certain vehicle makes or models pose a higher risk of fatalities in the event of an accident.

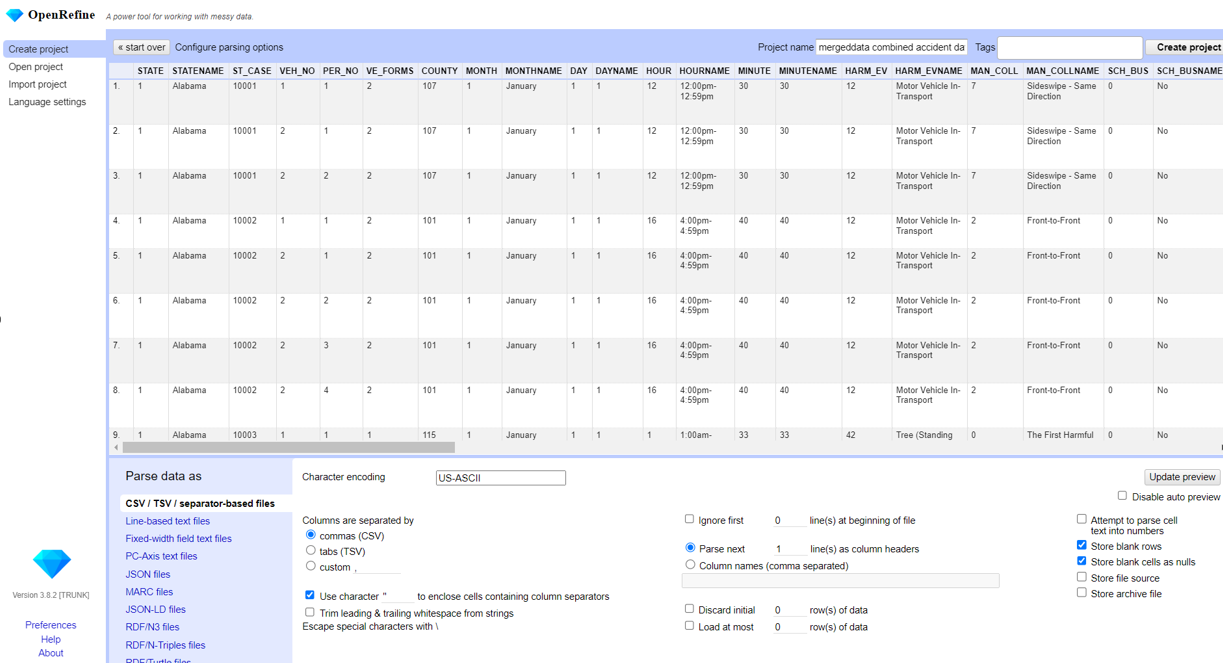
Ultimately, I strive to provide actionable insights to demonstrate the importance of high-safety vehicles in saving lives during accidents and highlight the risks associated with lower-safety vehicles. I hope my research offers valuable findings to inform regulatory decisions and increase public awareness about vehicle safety.

Additionally, I analyze the times of day when accidents are most frequent in the U.S. This information assists road safety teams in implementing enhanced security measures during peak accident hours. These insights enable authorities to identify critical periods when accidents occur, allowing them to introduce new regulations aimed at reducing fatalities, such as lowering speed limits, widening roads, or rerouting traffic.

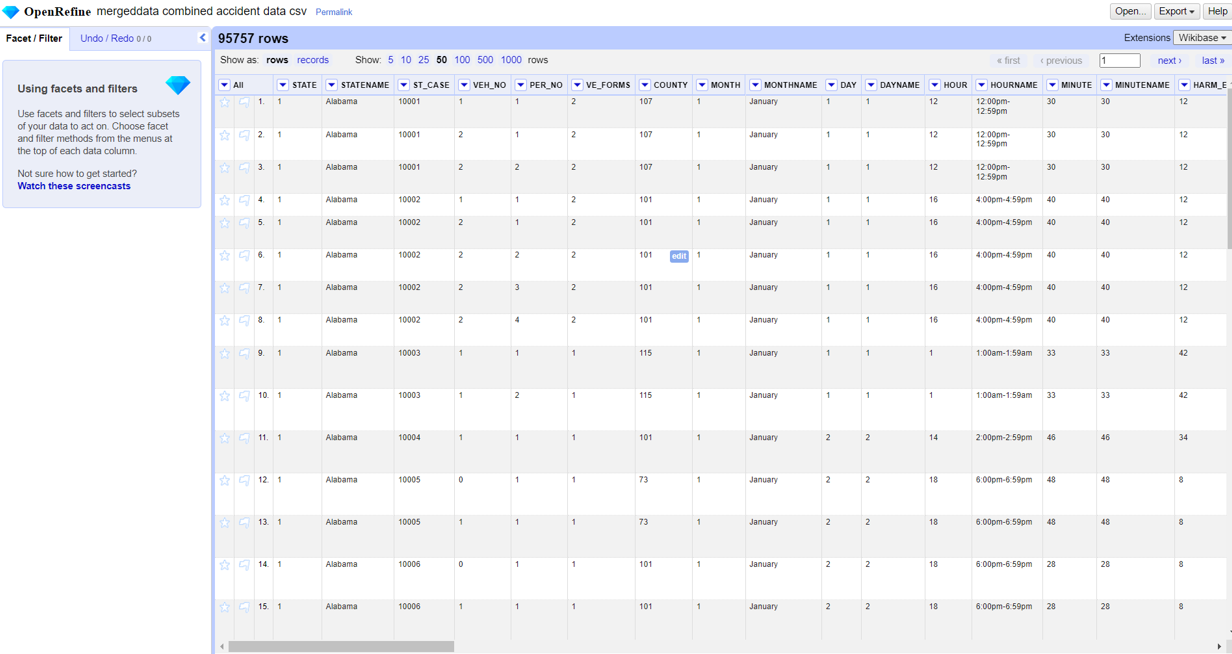
**Phase 1 - Data Collection**

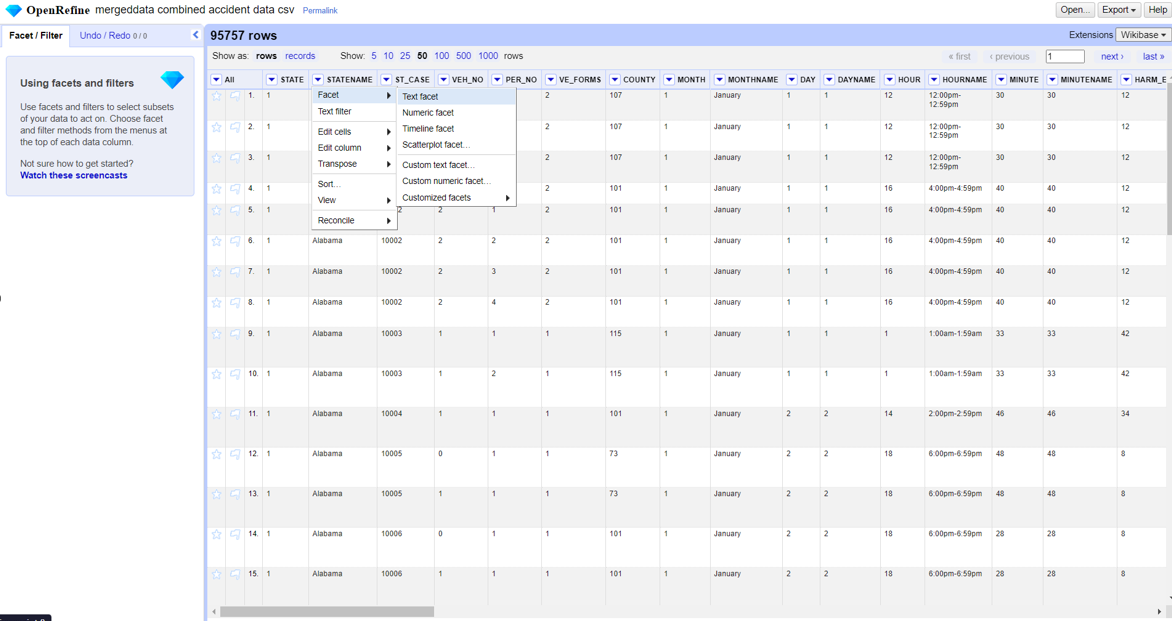
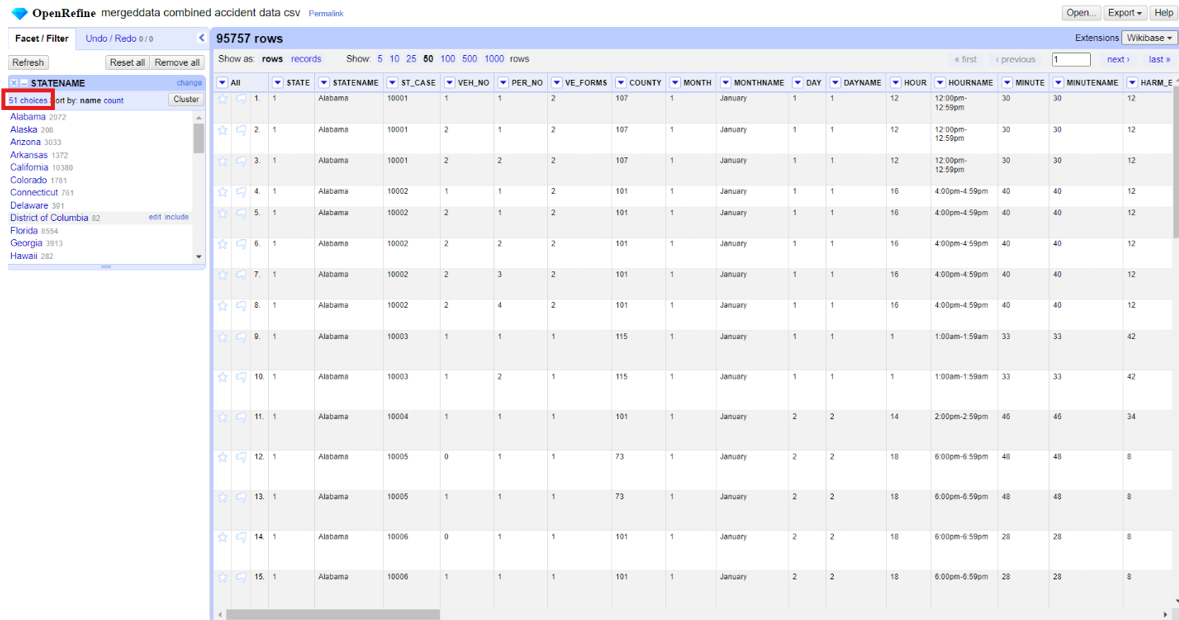
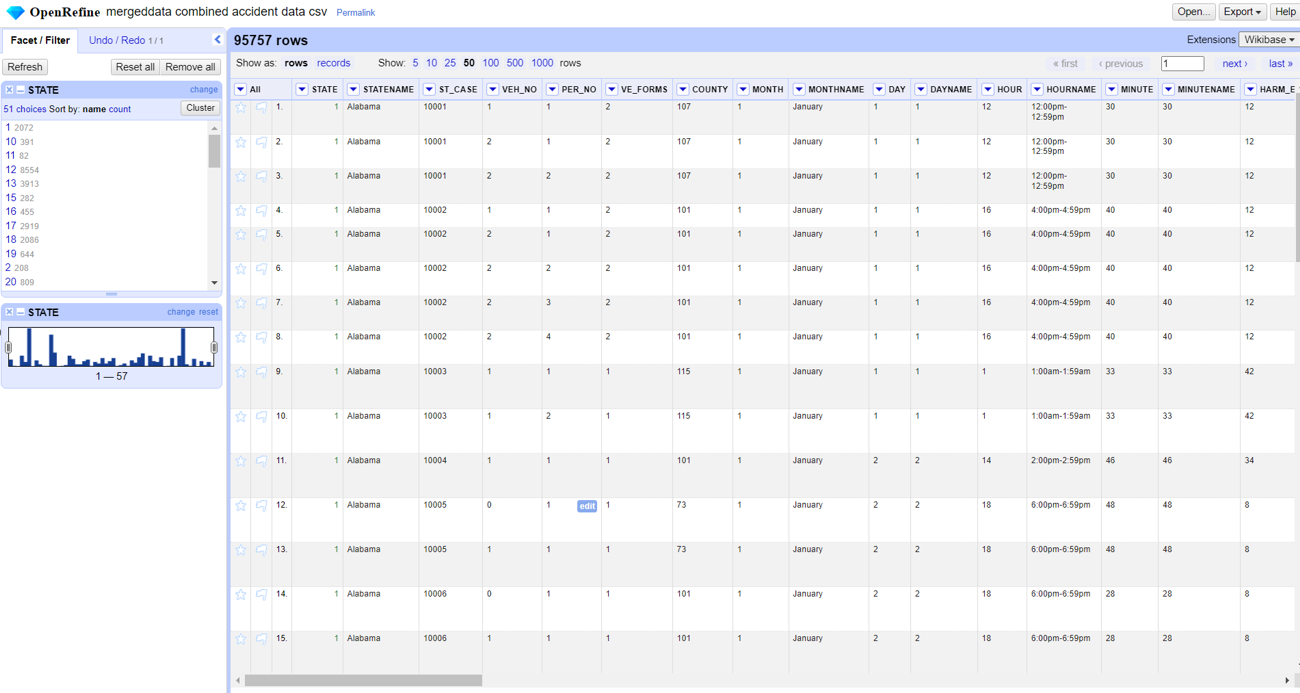
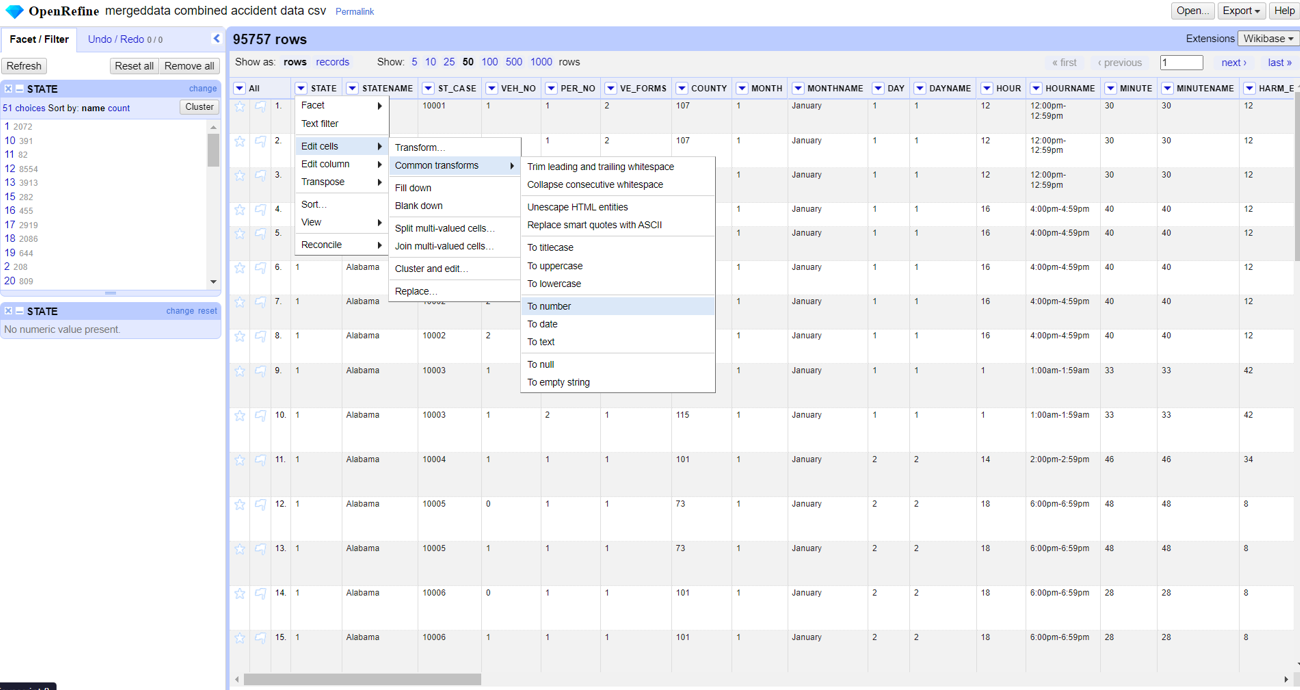
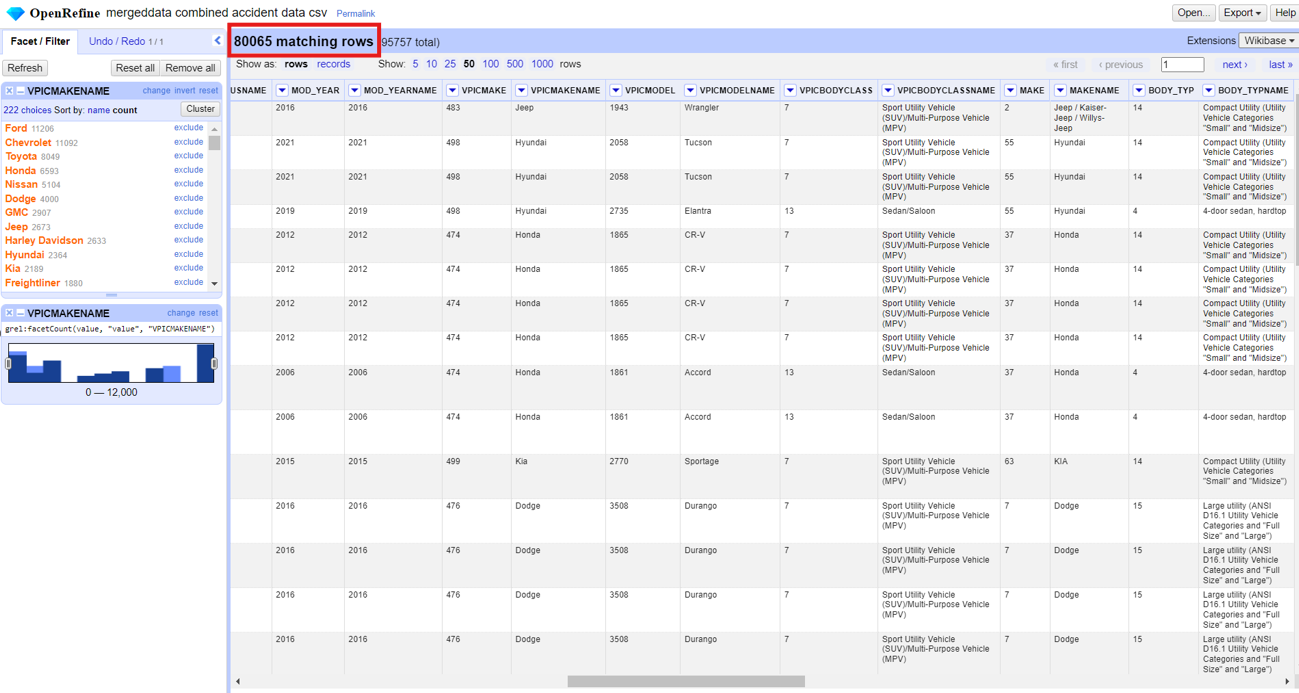
**Static Dataset:** The accident data from the Fatality Analysis Reporting System (FARS), downloaded from NHTSA, is used for analyzing vehicle safety during accidents.

Once we import the file, we will have an option to customize the file upload. Generally, we leave the options by default.



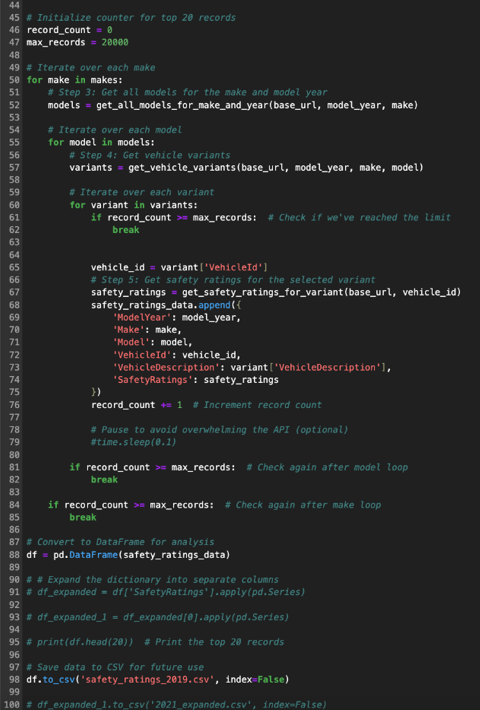
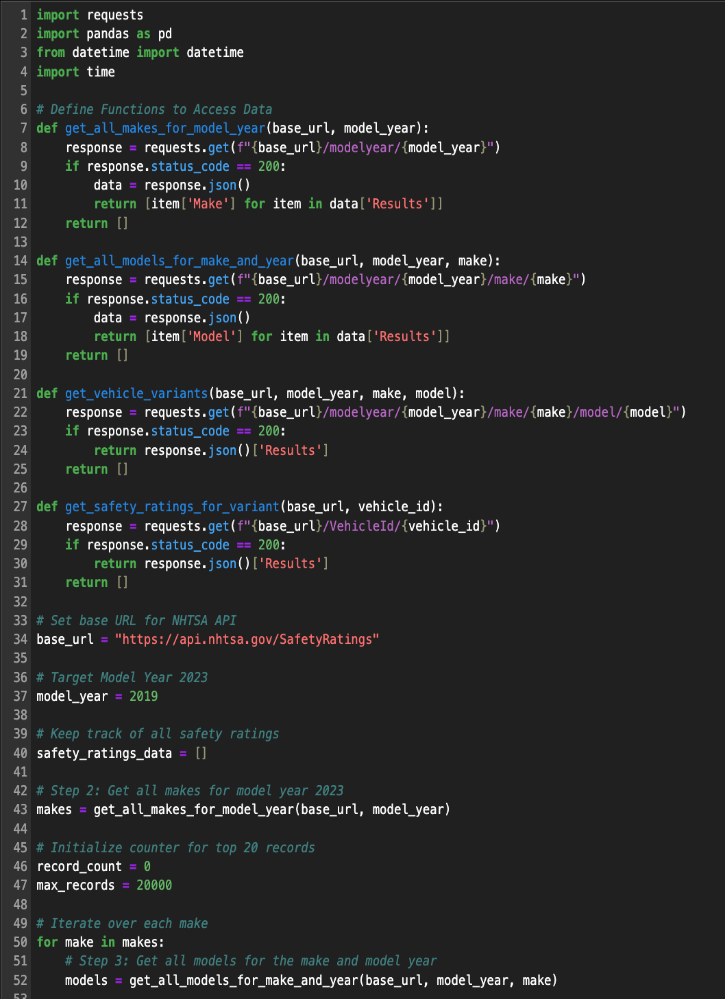
Screenshot of the imported file after creating the project with a total of 95757 records:



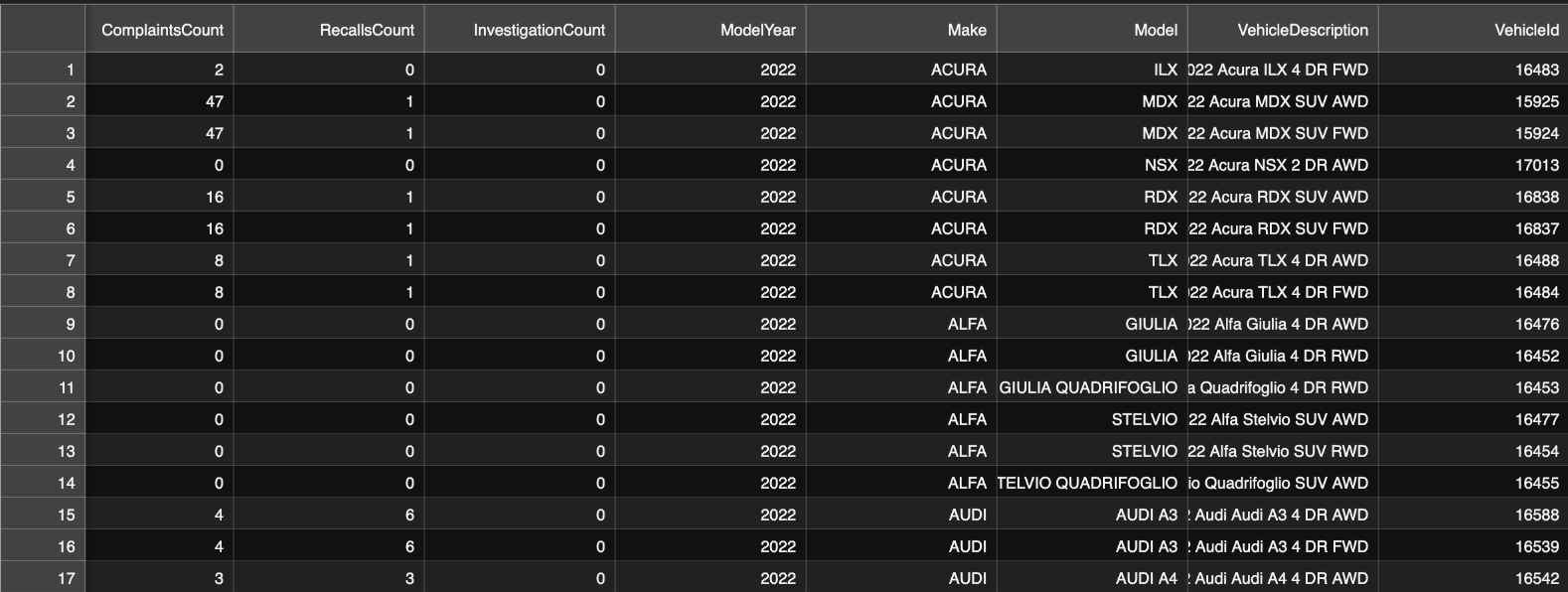
1. I have checked for any spelling mistakes or blanks in the State Name column and confirmed that everything is correct. There are a total of 51 entries, including the District of Columbia.**** 
2. Changed the State code from text to integers:
3. cleaned the data by eliminating blanks from the Vehicle Name column, resulting in a total of 80,065 records remaining. 

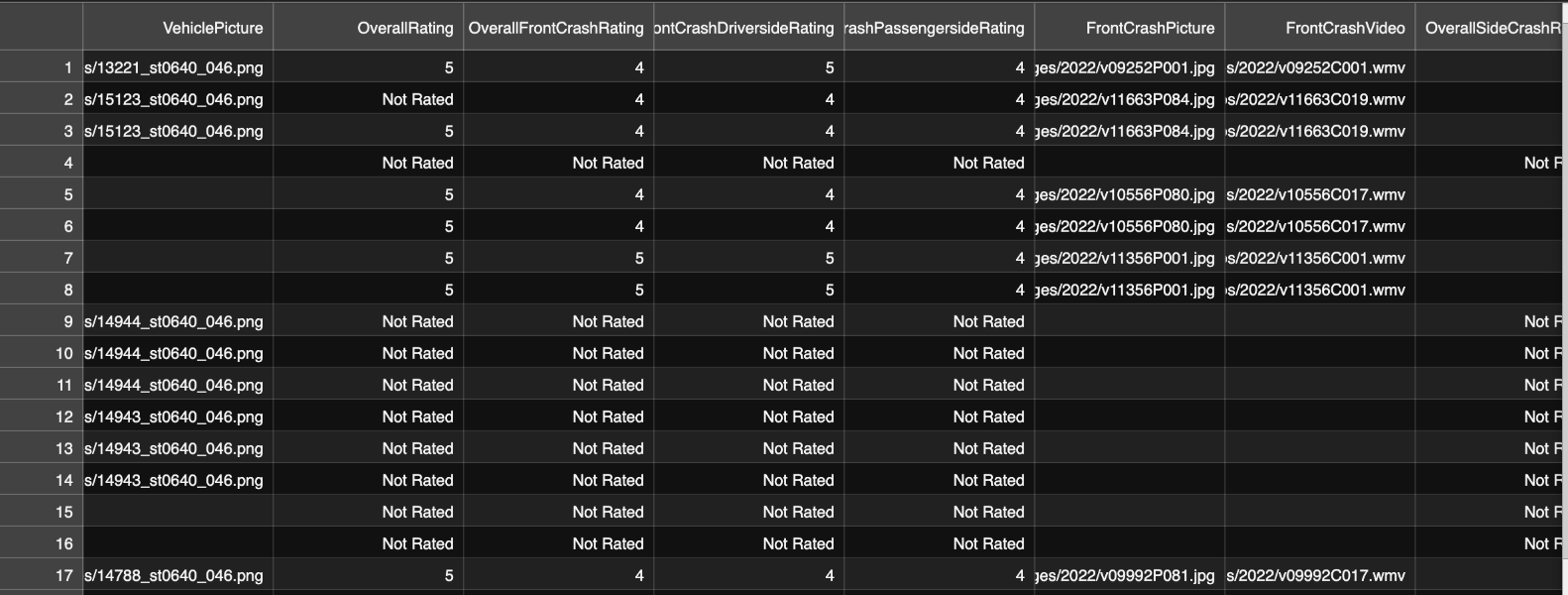
**Streaming Dataset:** NHTSA Datasets; Vehicle make, model, overall rating and year.

I have used Jupyter to pull streaming data of Vehicle Safety Specific information using API. Below are the screenshots of imported data and attaching the Jupyter notebook file:



Results:

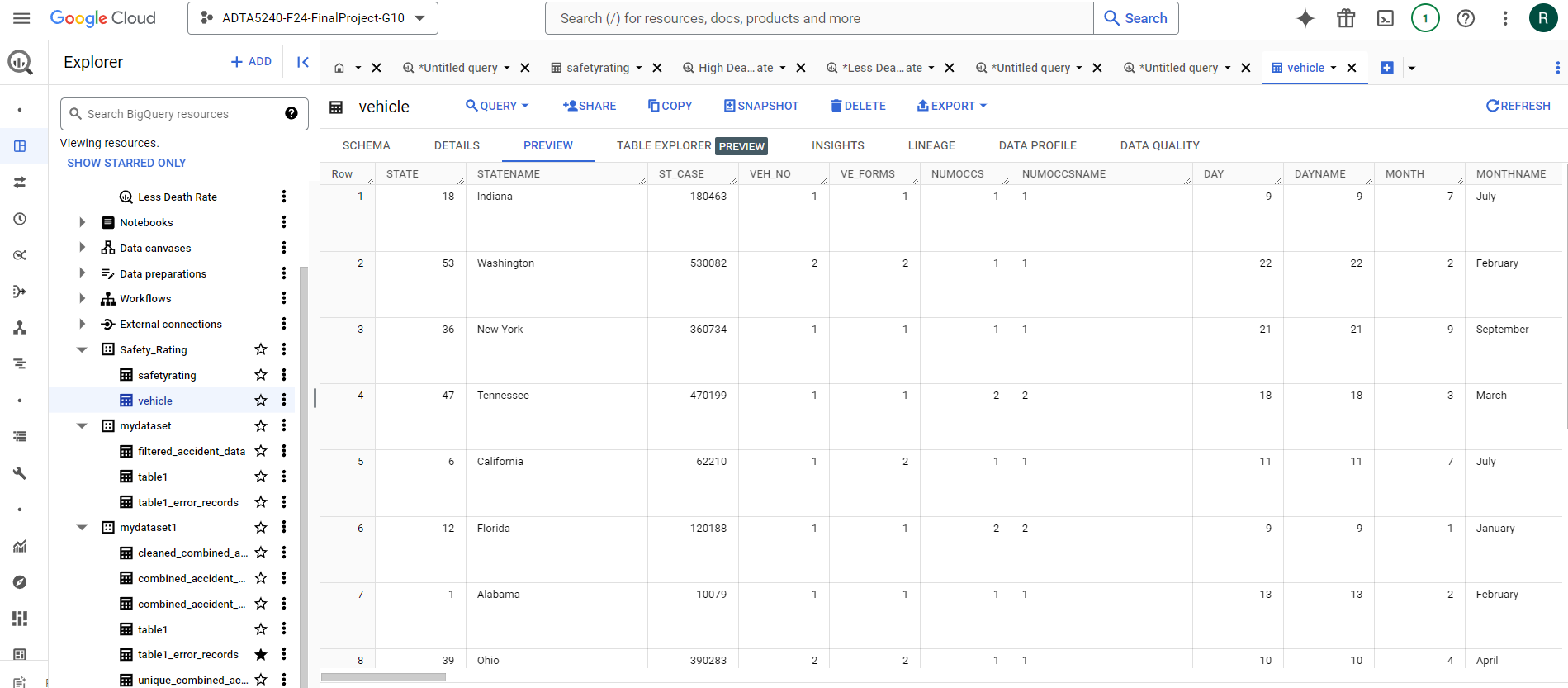






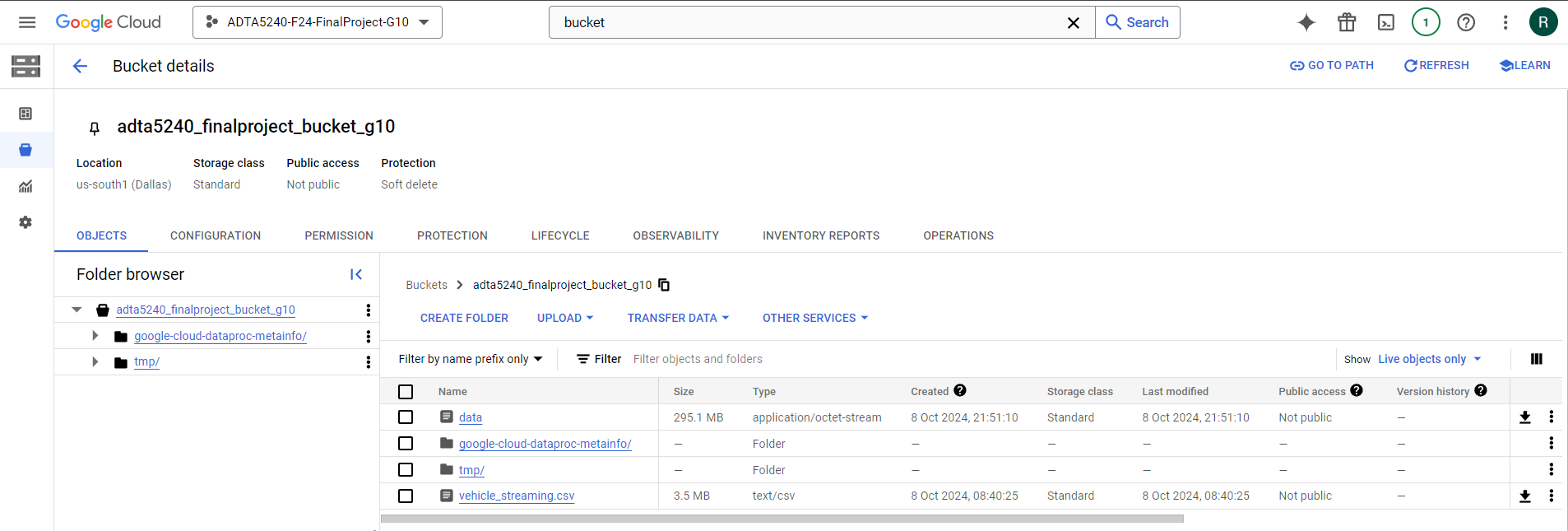
**Phase2 -data storage and pipelines**

In this scenario both static and streaming data were stored. Below is a preview of the static data loaded into the Big query ready for further preprocessing:



**Data Storage - GCP**

Once the FARS dataset and API is processed , the data is stored in Google Cloud Platform platform (GCP) . A new project has been created under name “ADTA-5240-F24-FinalProject-G10” also respective bucket “adta5240\_finalproject\_bucket\_g10” have been initiated . Also, the vehicle rating , vehicle accidents files and streaming data has been added to the bucket.



Once the data is stored and loaded into Big Query, we now can begin analyzing the accidents and their fatality rates. The screenshot below shows that the data has been successfully uploaded to Big Query for further analysis:

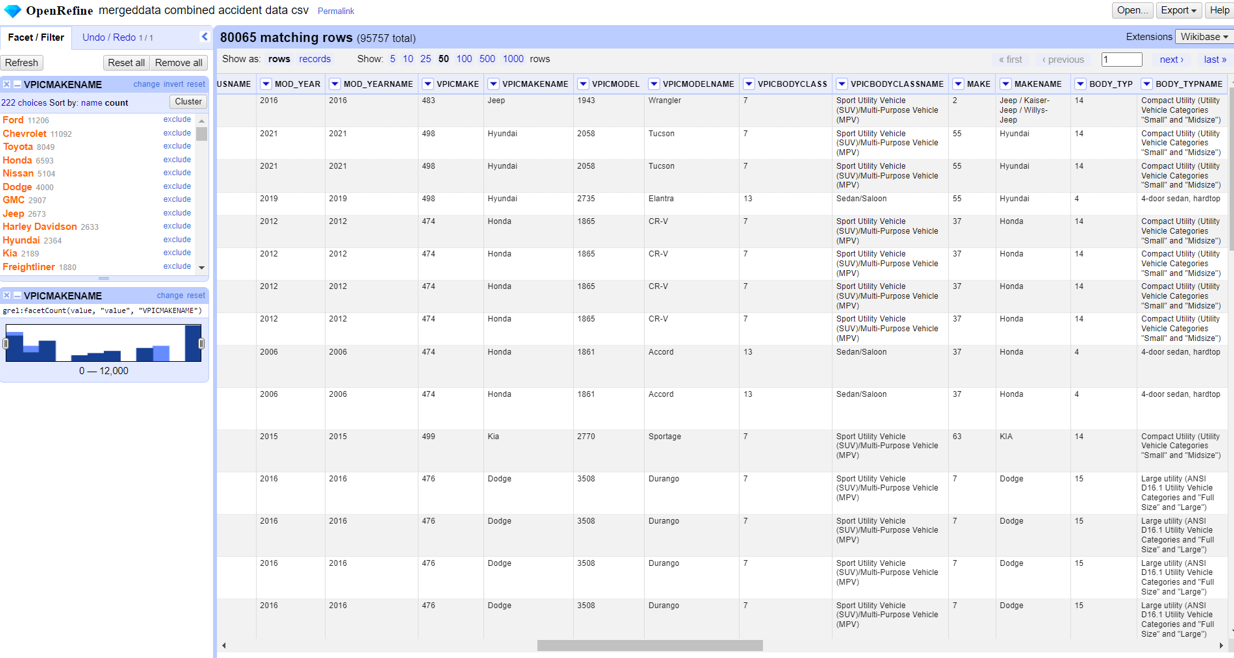
**Data Management – Hadoop**

* Once the project and bucket has been created in GCP, now Hadoop Data Proc clusters needs to be implemented to retrieve/query the vehicle safety ratings information from the accident data. Make sure to enable Cloud Data Proc API and Computer Engine in GCP. It helps to manage large amount of data.

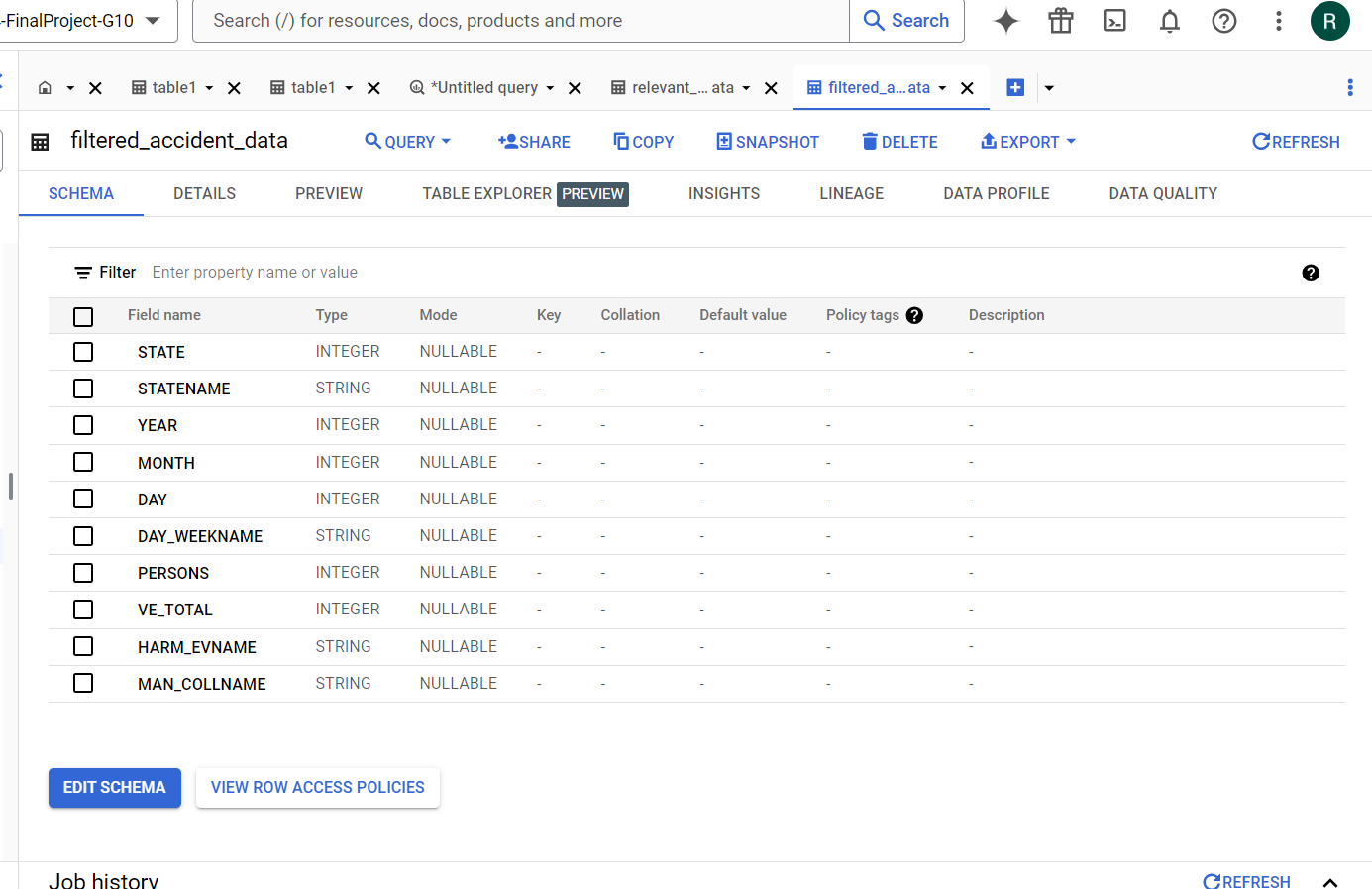


**Phase 3-Data cleaning**

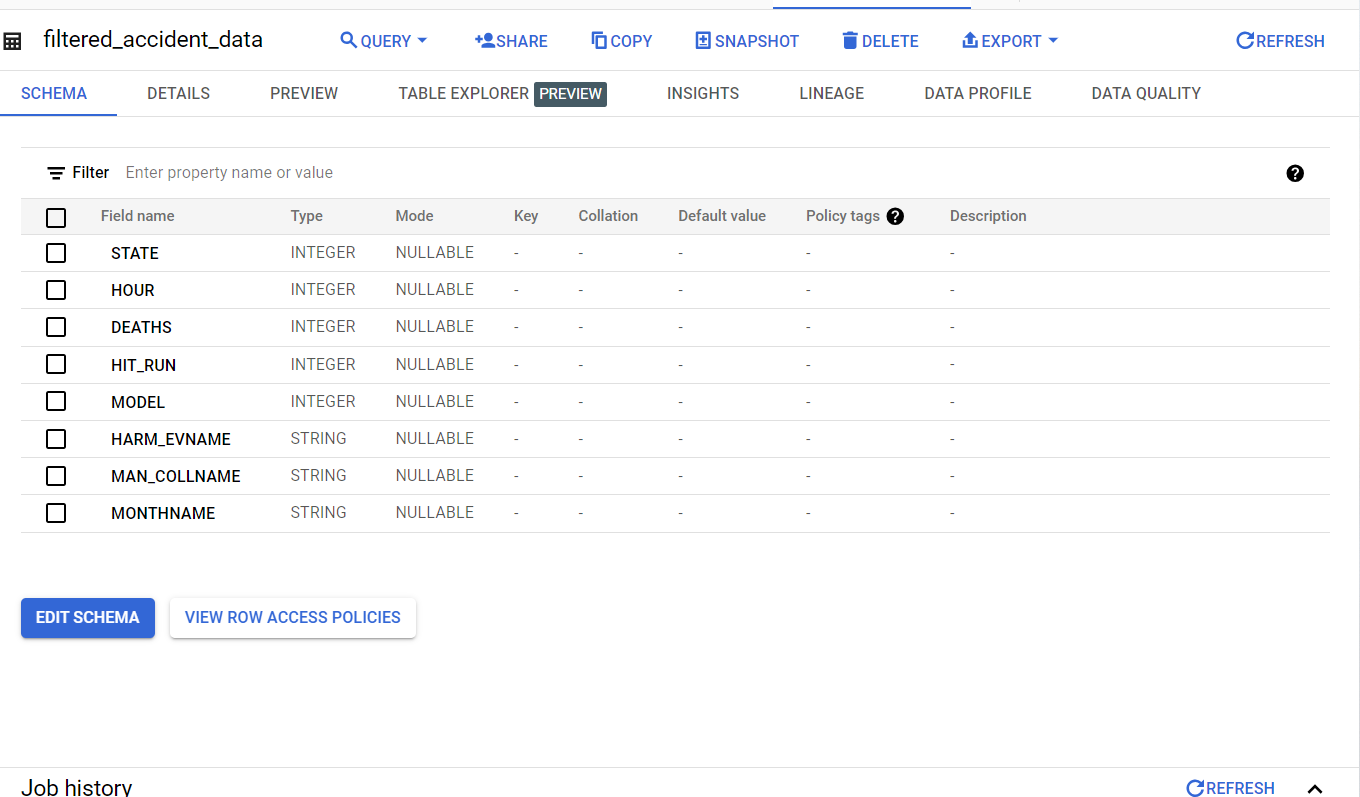
Both datasets contained numerous unnecessary columns and rows that needed preprocessing before any further actions could be taken. One dataset has already been cleaned in Open Refine, and now we plan to clean the other dataset in GCP. Since the current dataset is nearly clean, we will focus on removing the irrelevant columns and retaining only those that are essential for our further analysis.



The new filtered accident dataset with the most relevant columns for further analysis

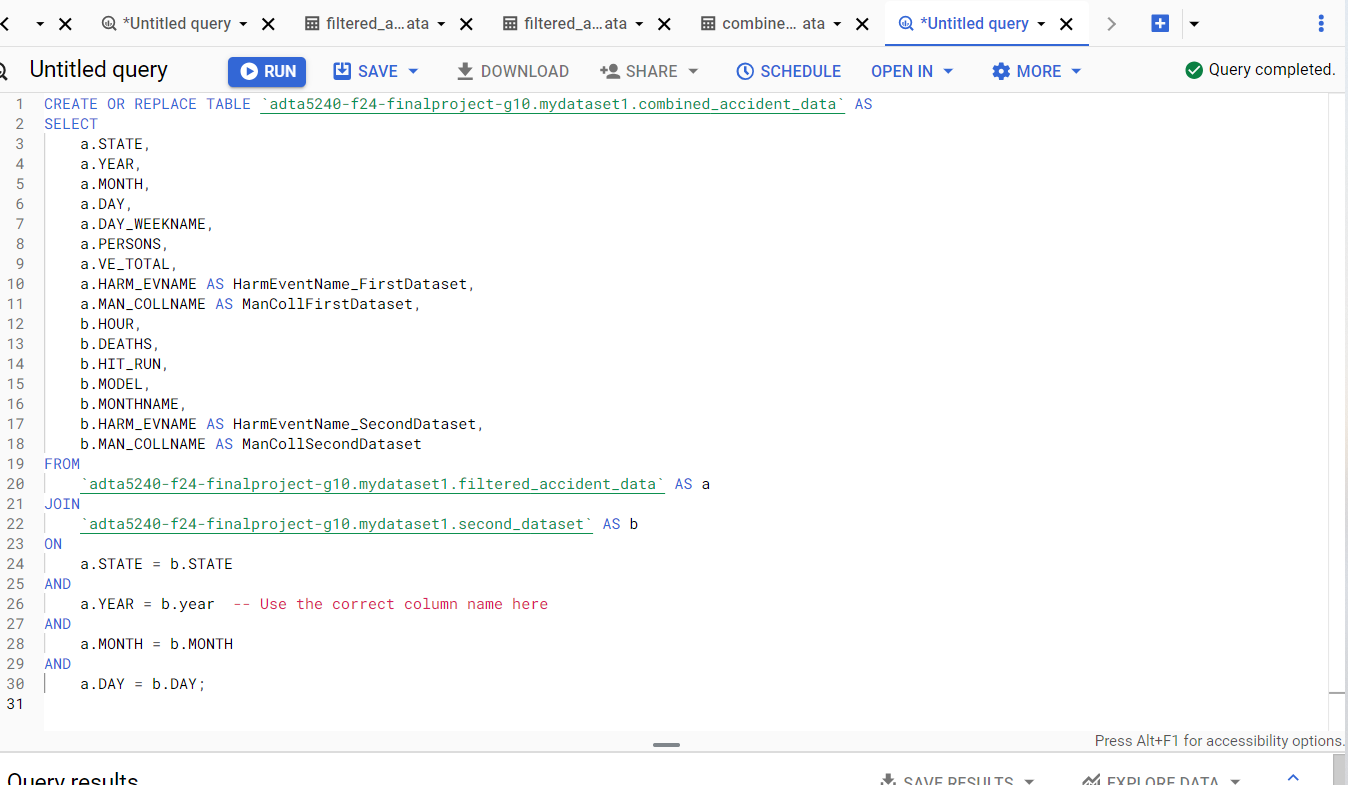


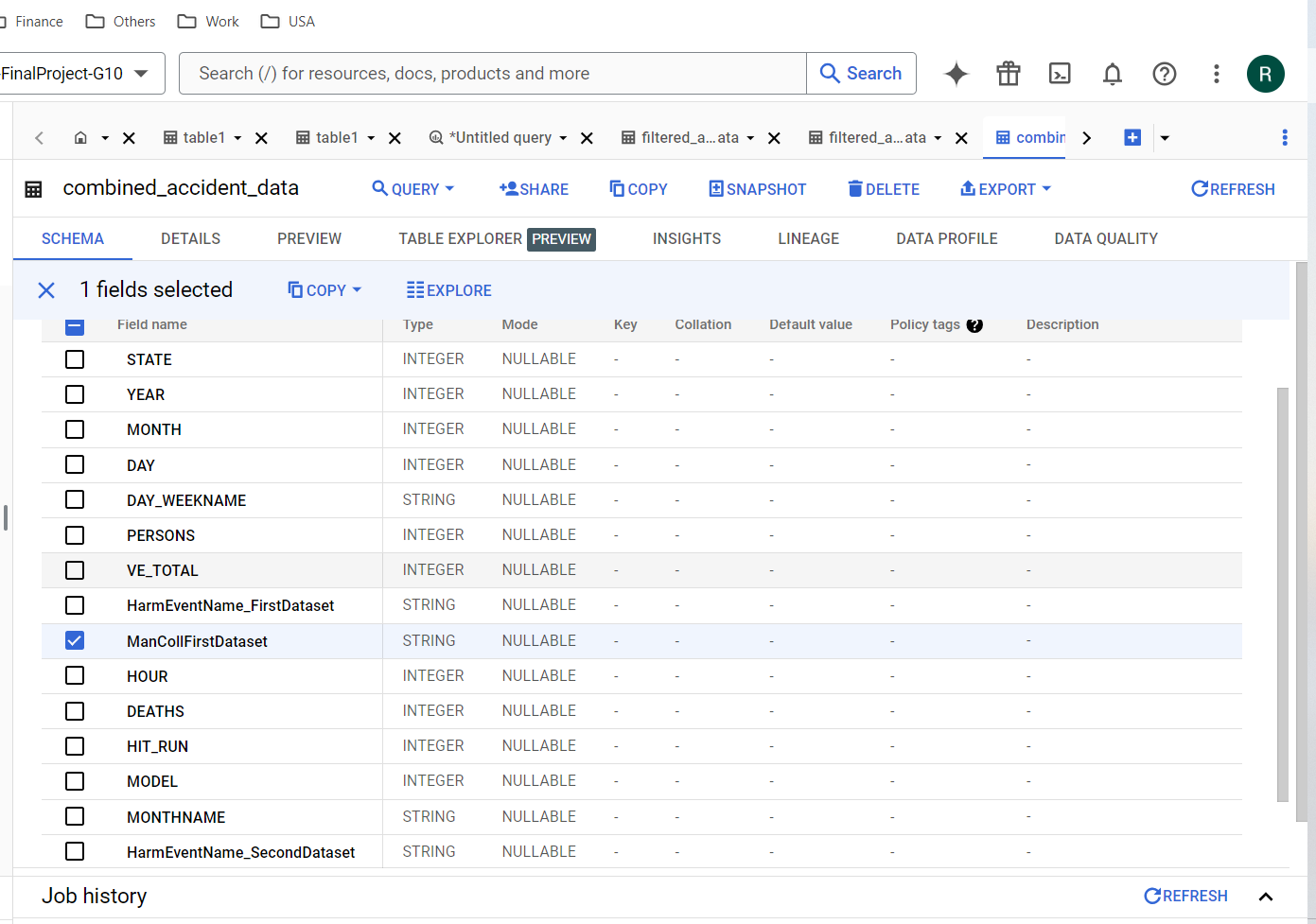
For the static dataset as well, it was filtered with the most relevant columns.



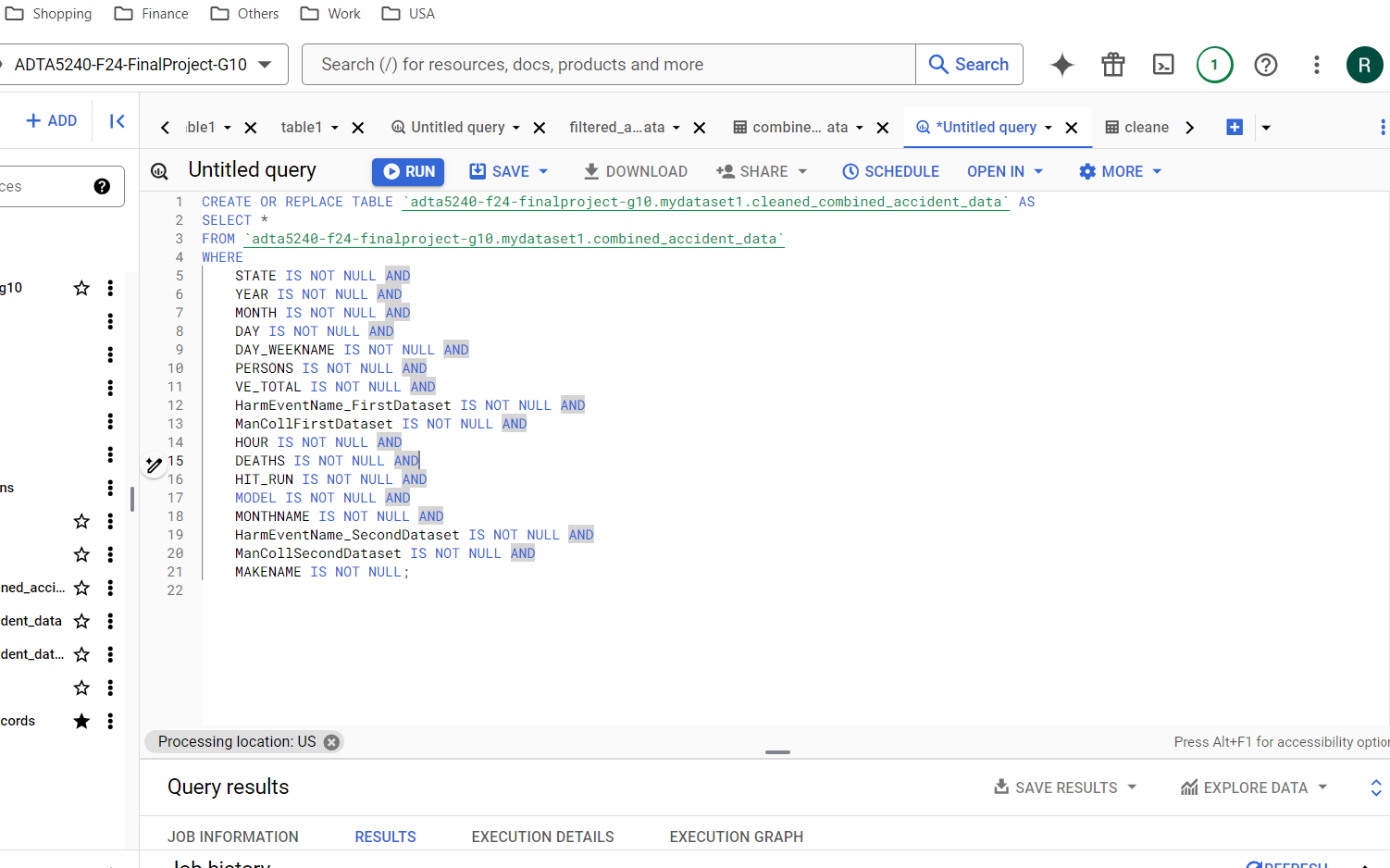
**Data Analysis:**

After filtering the necessary columns and combining the two datasets, the following screenshots display the schema of the final dataset.

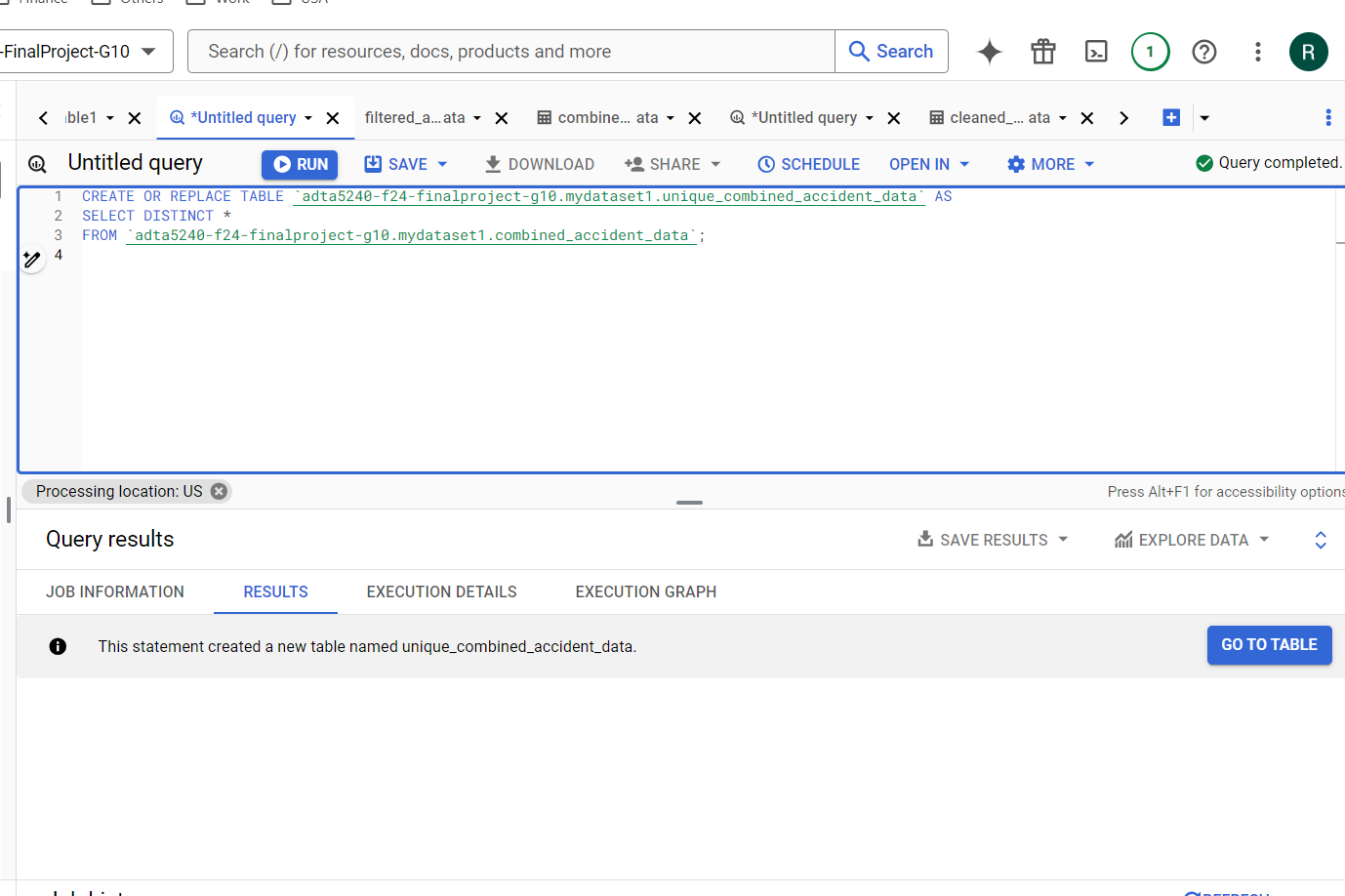




I filtered out all null values to ensure that the analysis is precise and provides accurate data without any discrepancies.



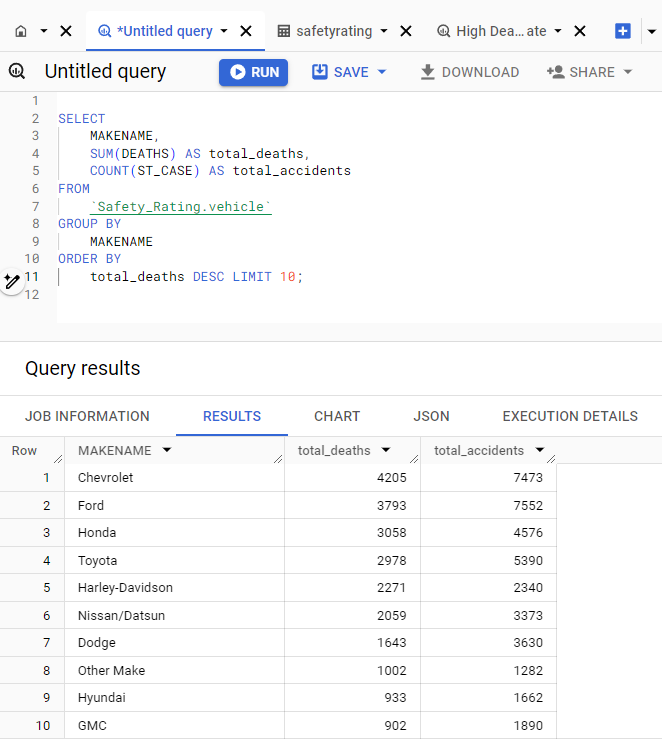
If the dataset contains duplicates, it can result in inaccurate analysis and hinder our ability to determine the exact number of accidents and fatalities. Although I removed all duplicates during the data cleaning process, I ran a query again to ensure that no duplicates remain. This will enable us to accurately count the accidents and fatalities that occurred.

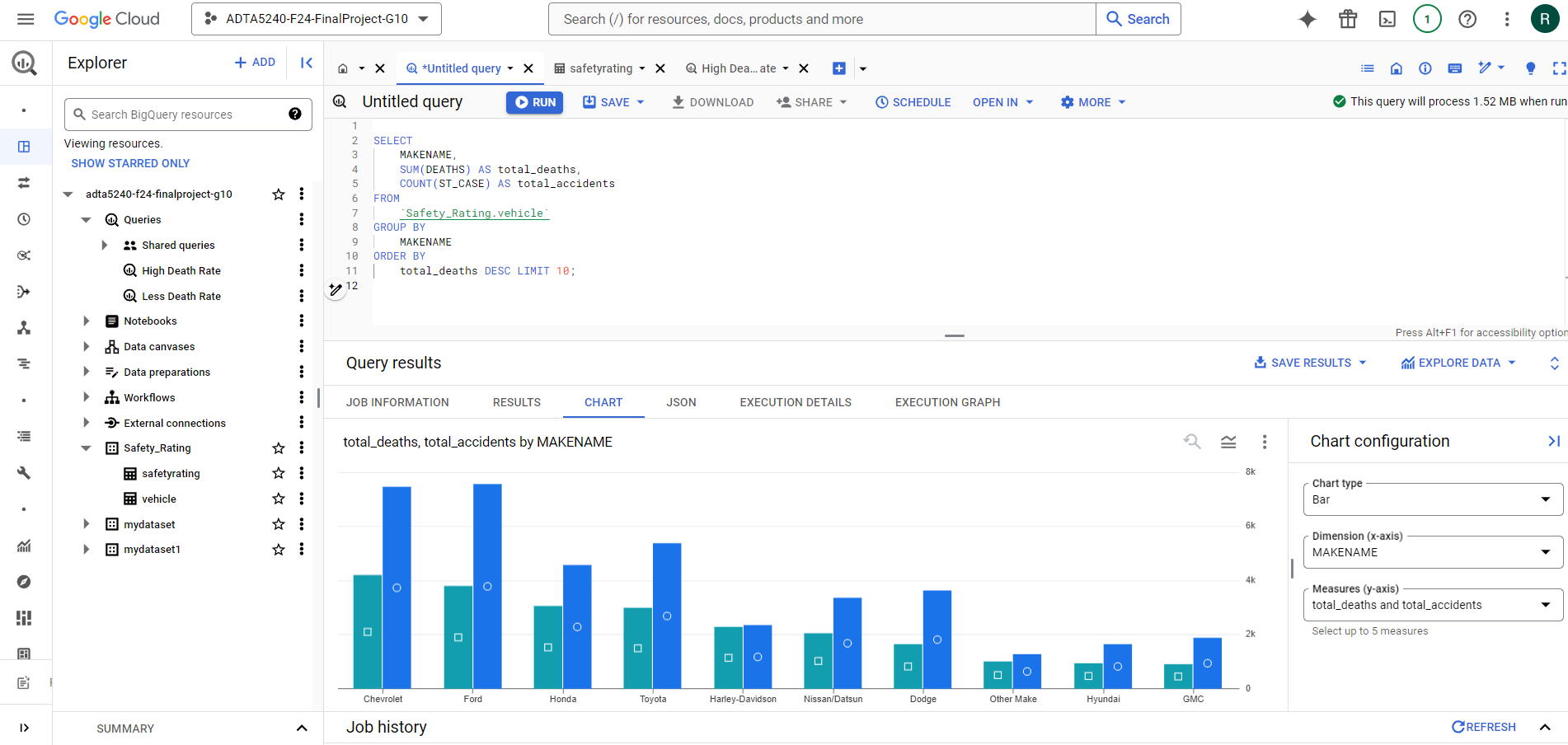


**Phase 4 -Data analysis and Visualizations**

Several valuable insights have emerged from the final dataset, which will assist society in choosing vehicles based on overall safety ratings, ultimately helping to reduce the fatality rate. Additionally, these insights will inform authorities about the times of day when accidents are most frequent, enabling them to implement new rules and road projects in high-incident areas, such as widening roads or increasing lighting at those locations.

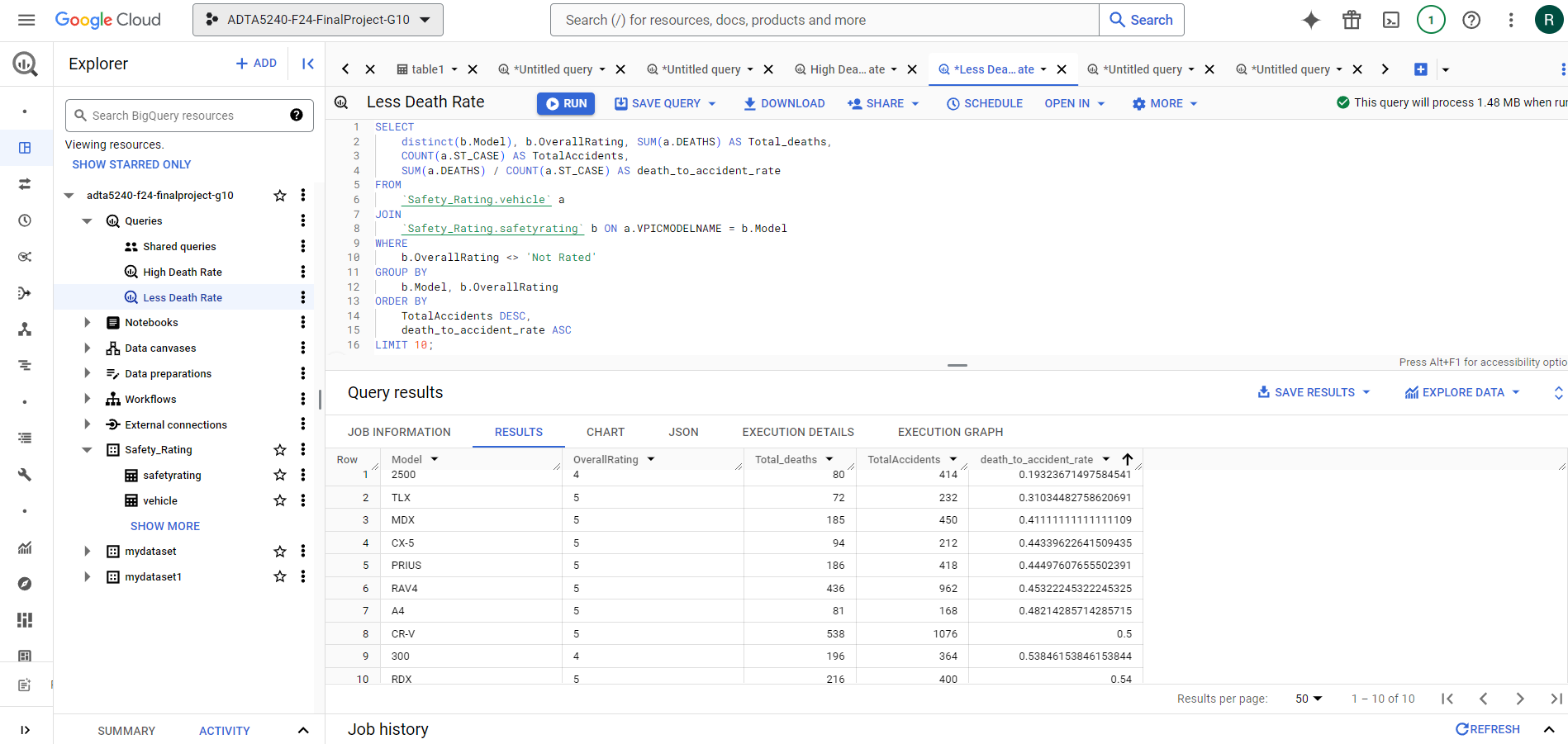
* Top ten vehicle makes with the highest number of accidents are shown in the below analysis:

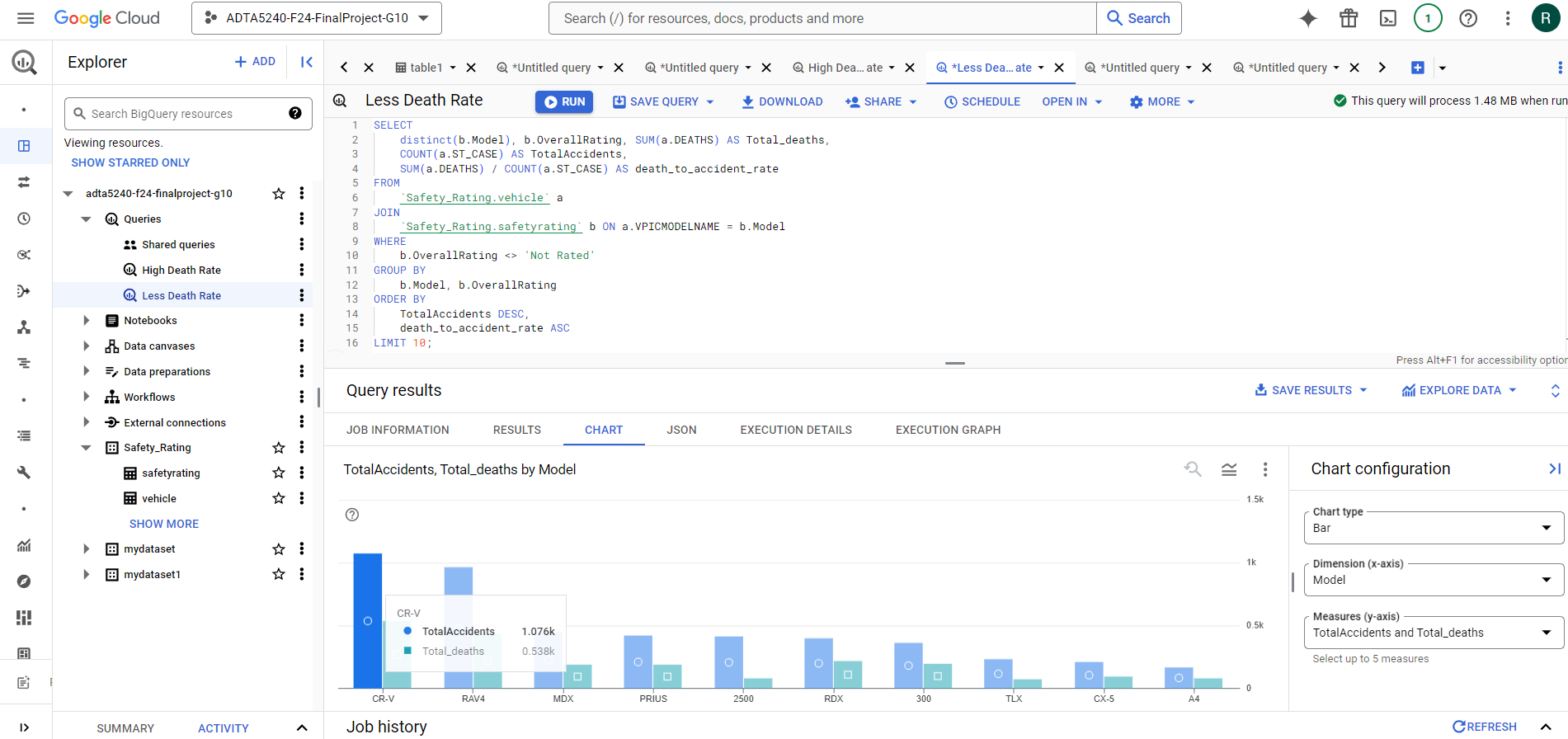




Ford and Chevrolet are involved in the most vehicles, with 7552 and 7473 total accidents, respectively. Both makes also have a significant number of fatalities, with 3793 for Ford and 4205 for Chevrolet. Third on the list is Toyota, with 5390 accidents and 2978 fatalities. Notably, Harley-Davidson has almost 1:1 fatal accident ratio, indicating that accidents involving this make are often fatal. Hyundai and GMC reflect this trend as well, showing a low number of accidents but a high number of fatalities. Based on these insights, there appear to be differences in vehicle severity that warrant further study to identify vehicle-specific safety strategies.

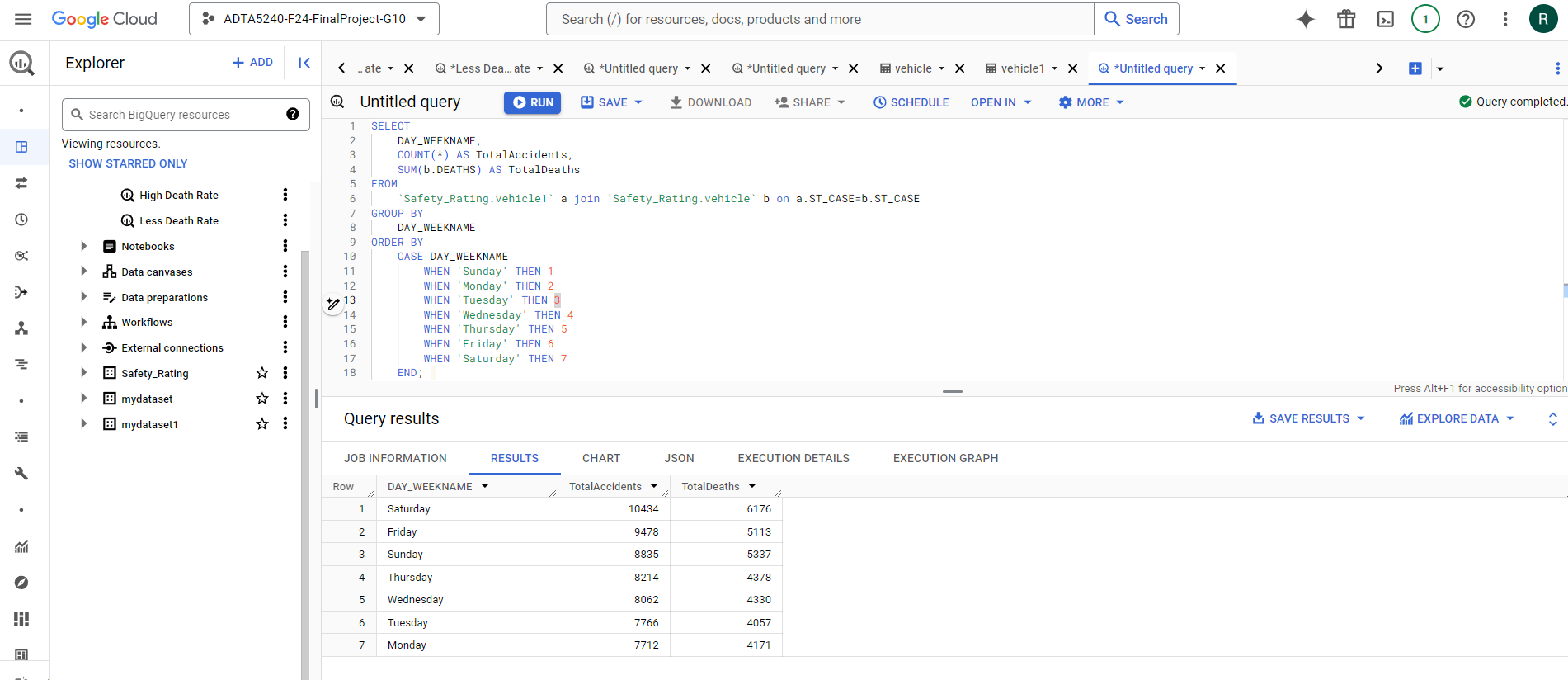
* I conducted an analysis of vehicles that experienced the highest number of accidents but had a lower likelihood of fatalities. Below, you can see the results highlighting vehicles with 4 and 5-star ratings and their corresponding low death rates.

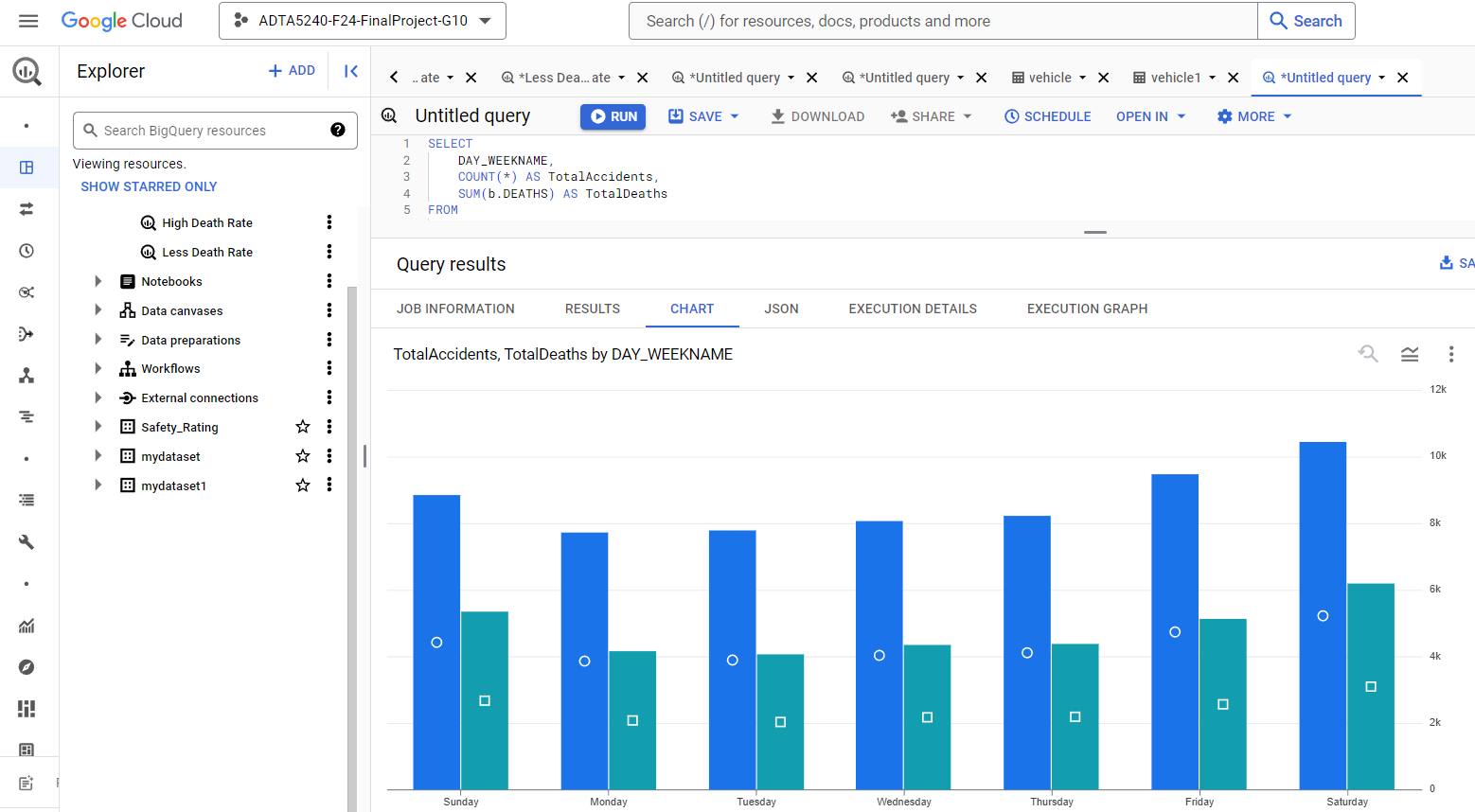




Notably, the RAM 2500 model exhibits the lowest fatality rate at 0.19.

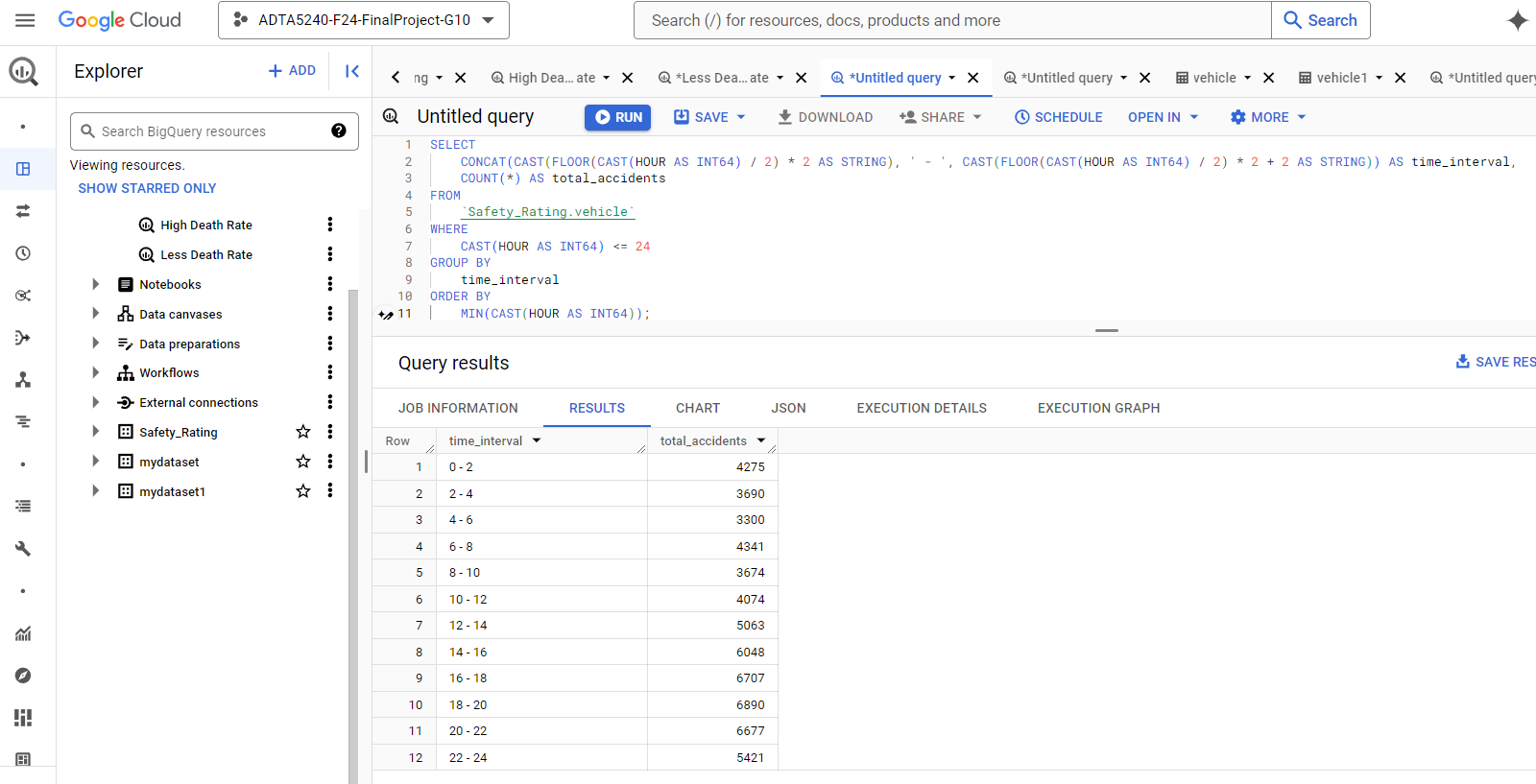
* I have compiled data on accidents that occurred on all days of the week, allowing authorities to be more vigilant on days with higher accident rates. By implementing additional rules on these specific days, it is anticipated that the number of accidents and fatalities will decrease.

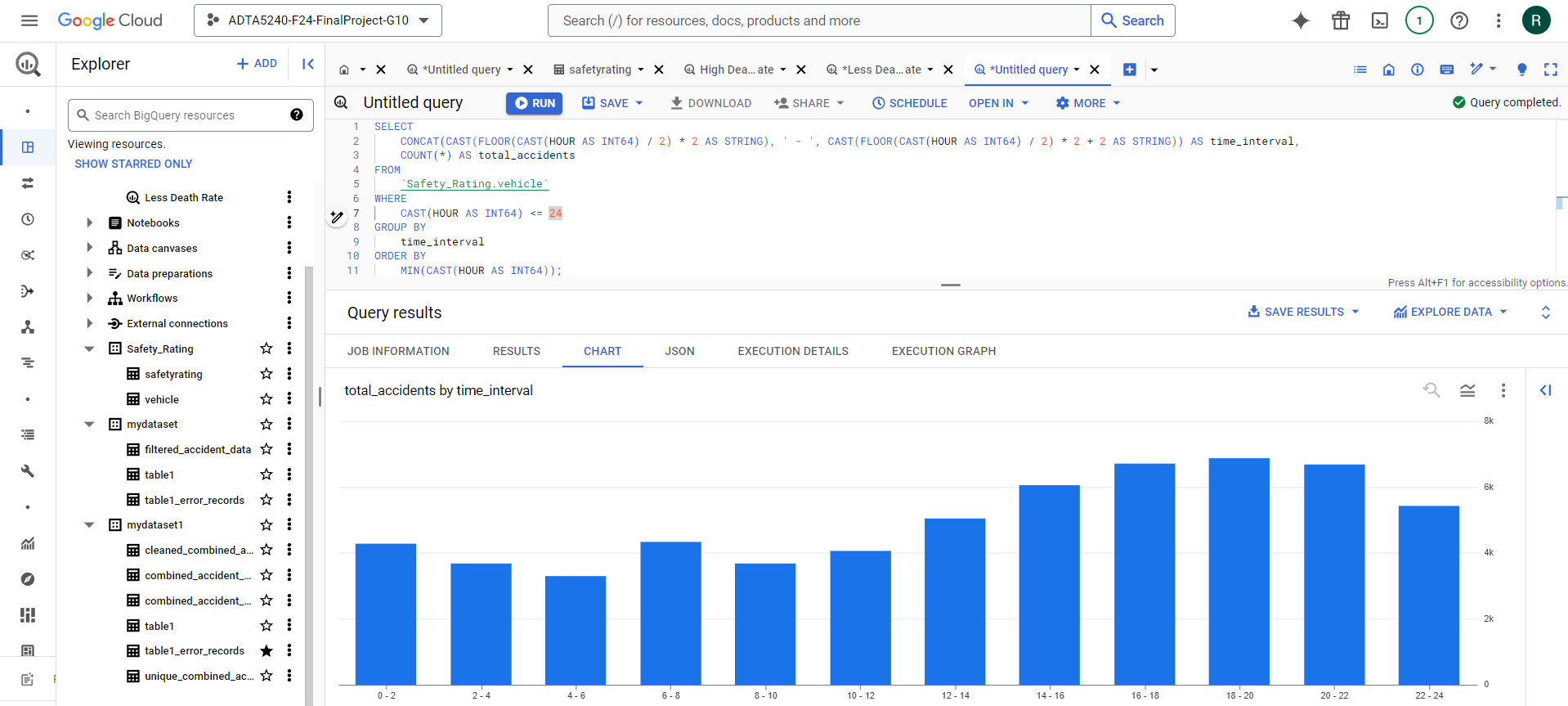




The data indicates that weekends, particularly Saturdays, have the highest rates of accidents and fatalities, suggesting an increased risk during leisure time. Generally, weekdays see fewer incidents. Therefore, there is a need for targeted safety measures and awareness campaigns during the weekends to minimize risks associated with increased road activity and social behavior during this time.

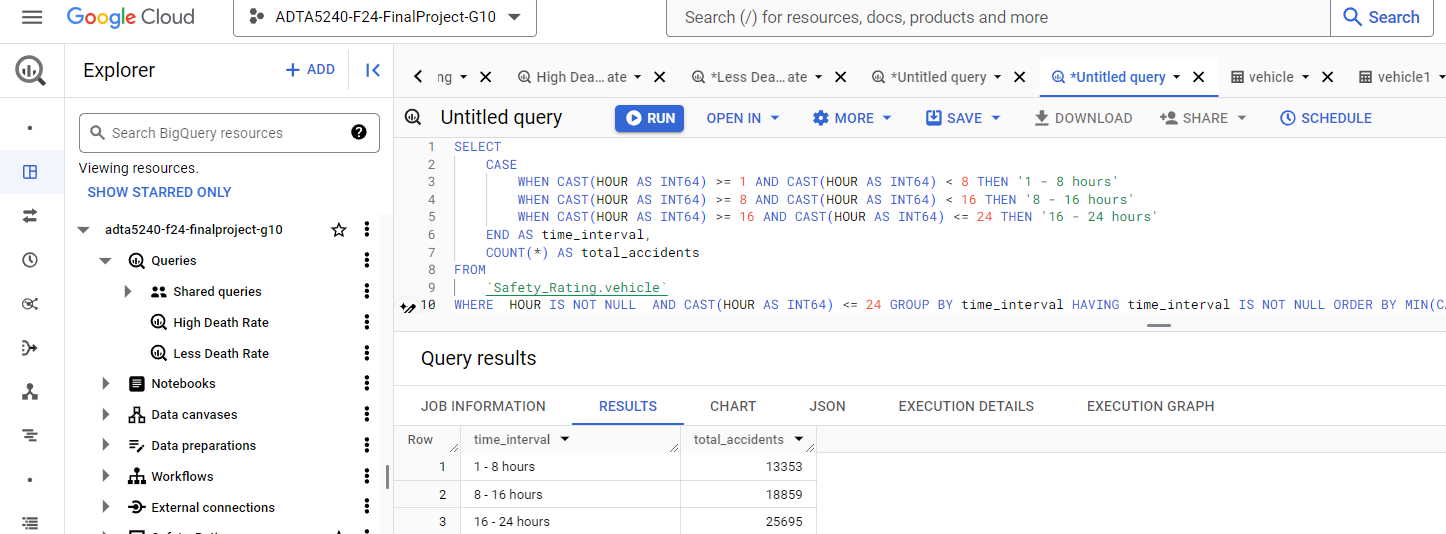
* I have also analyzed the hours of the day when the most accidents occur. This analysis will help individuals identify high-risk times, allowing them to plan their activities accordingly to avoid accidents. Additionally, it enables officials to implement enhanced safety measures during these peak hours.

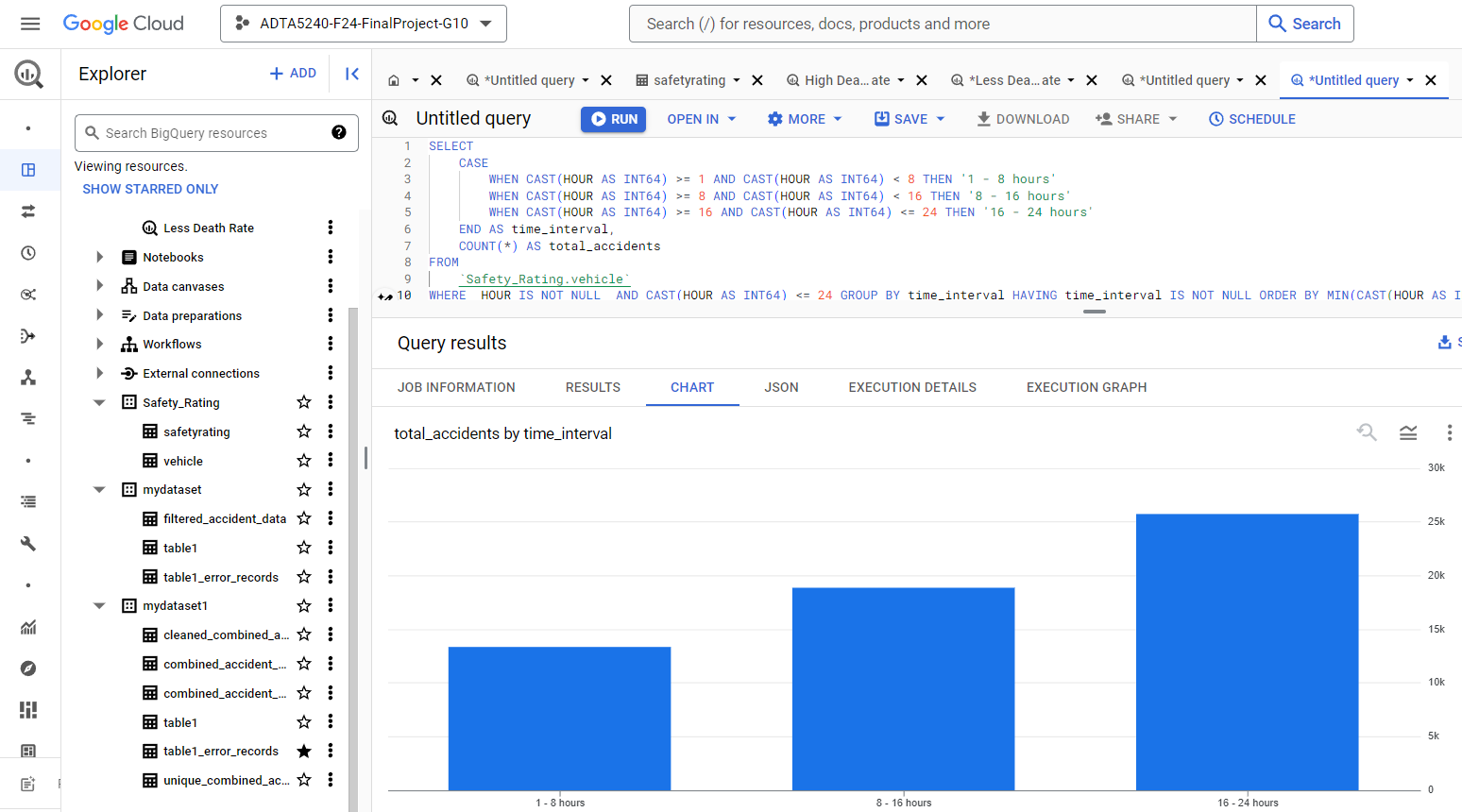




The data indicates that the highest number of accidents occurs in the evening, especially between 4 PM and 10 PM, with a sum of 28,322 accidents recorded at between 4-10PM. This trend aligns with peak commuting times, highlighting the need for targeted safety measures to reduce incidents during these hours.

* The query below will help us retrieve data similar to the previous analysis, but it will also clarify which times of day (morning, afternoon, evening) are more likely to have accidents with fatalities. This understanding will assist officials in implementing additional safety measures, and the report will provide insights that are easily comprehensible for anyone interested in daily accident trends.





Accidents occurring during 04-12PM traffic hours are notably higher than remaining hours of the day. Off-peak hours are linked to a greater number of injuries and fatalities compared to accidents that take place during rush hours.

**Data Interpretation:**

* From the analysis and visualizations, it is clear that the fatality rate can be significantly reduced by using vehicles with high safety ratings (4 or 5 stars) for transportation. We have also extracted data on the vehicle makes that are involved in the most accidents and their associated fatality rates. This analysis highlights the importance of considering safety ratings when the public plans to purchase a new vehicle.
* In addition to the above, I have analyzed busy traffic days and hours, identifying weekends and the period from 4 PM to 12 AM as the times with the highest fatality rates. This insight enables authorities to enhance safety measures during these peak periods and informs the implementation of additional infrastructure projects to improve road safety.

**Conclusion**

I have conducted a thorough analysis of the NHTSA datasets, uncovering various insights related to vehicle types, accident timing, and severity. Our findings indicate that certain vehicles and types are more frequently linked to severe accidents, with fatalities and incidents occurring more often on weekends, particularly on Saturdays. Additionally, the highest traffic volume is observed between 4 PM and 12 AM, highlighting a strong correlation between accidents and fatality rates. This analysis emphasizes the need for targeted safety measures during high-risk periods such as weekends and rush hours, as well as the necessity to improve vehicle safety standards, which have significantly contributed to the reduction in fatalities. Understanding these trends can inform future safety regulations and enhance public awareness regarding accident prevention and vehicle safety.

**References**

Peura, C., Kilch, J. A., & Clark, D. E. (2015). Evaluating adverse rural crash outcomes using the NHTSA State Data System. Accident Analysis & Prevention, 82, 257-262.

Wang, X., Liu, Q., Guo, F., Xu, X., & Chen, X. (2022). Causation analysis of crashes and near crashes using naturalistic driving data. Accident Analysis & Prevention, 177, 106821.

Admassu, K., Linge, M., & Shirazi, A. (2018). Dataset Analysis and Trend Learning of the NHTSA 2015 Traffic Fatalities Data.

ChatGpt- For Importing data from streaming data using Jupyter.