**Enhancing Traffic Safety Through Data-Driven Analysis**

Madhulika Gangarajula

**Introduction:**

**Project Overview:**

In our project aimed to make roads safer by using data to understand traffic accidents better. We started by gathering lots of information through a set of data about accidents and to examine the traffic collision incidents in the City of Los Angeles from 2010 to present. This data helped us learn about past accidents. Then, we added real-time updates from systems that report traffic incidents and cameras that watch the roads. We used different tools like Google Cloud Platform to store data, Hadoop to process it, and BigQuery, Hive, and Spark to analyze it. By looking at both old and new data, we figured out where accidents happen most, why they happen, and how bad they can be. This helped us create models to predict accidents and manage traffic better. Next, we plan to improve these models using fancy techniques, set up systems to monitor traffic in real-time, and work with transportation agencies to make roads safer for everyone.

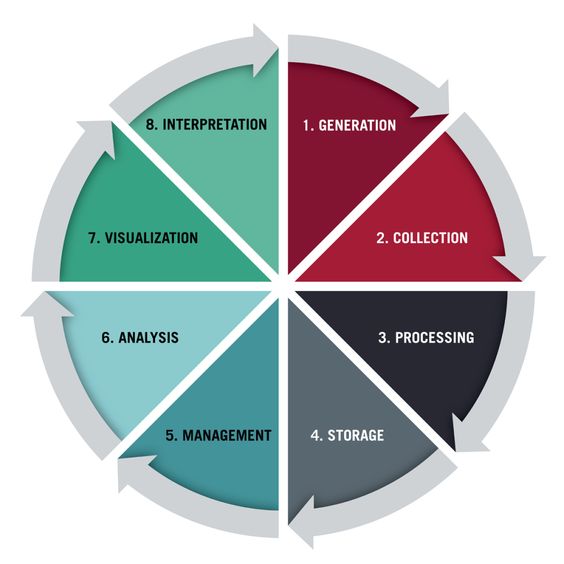
**Objectives:**

This study seeks to delve into inquiries aimed at uncovering insights through data-driven analysis to pinpoint crucial factors contributing to traffic accidents. Also, it aims to formulate focused interventions geared towards diminishing accident rates and enhancing road safety.

1. How to Evaluate the Top crime locations with high number of accidents in the city of Los Angeles?
2. How to Analyze the crimes by day of week?
3. What are the age groups of the victims?
4. How to analyze the crime rates by gender and area?

**Data lifecycle:**

Understanding the data lifecycle is fundamental as it delineates the phases traversed by data from inception to storage to processing, ensuring its sustained accuracy and reliability. Comprising eight sequential steps, this lifecycle mandates adept handling to effectively manage data. Throughout each stage, a plethora of techniques and resources are available for data processing, facilitating efficient management. Upon completing the data lifecycle, valuable insights can be gleaned from the dataset, underscoring its significance in informing decision-making processes. In this project, adherence to the eight steps of the data lifecycle will enable us to clean, store, and extract insights from the dataset methodically.



**Data Generation:**

Data originates from diverse origins, including sensors, human inputs, online transactions, or data transfers, marking the inception of the data lifecycle. At this initial stage, the data is raw and unrefined, rendering it unsuitable for analysis to drive data-driven decisions. This raw data typically contains null values and unwanted elements that necessitate preprocessing and cleaning before it can be utilized effectively for analysis purposes.

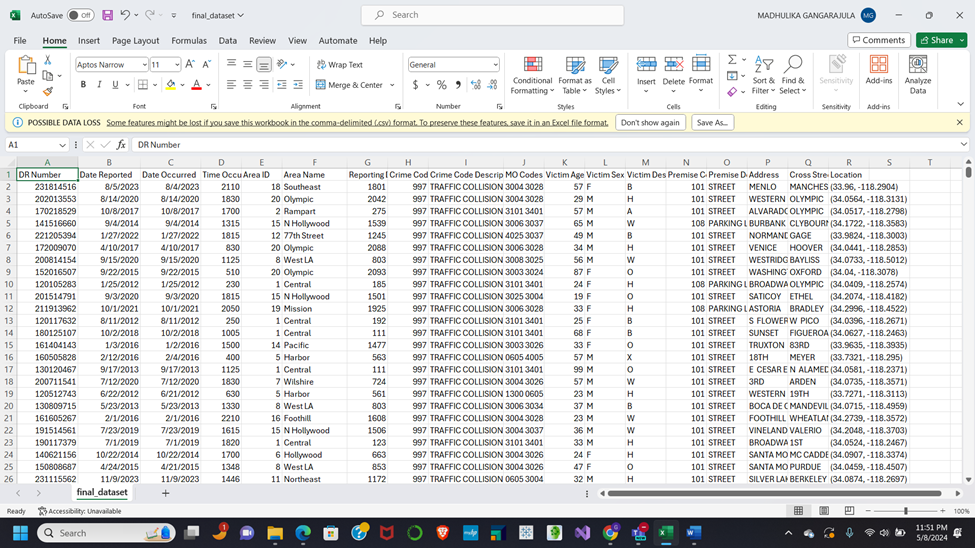
This data set we will use in this project is taken from "data.gov". This dataset shows the traffic collisions that have occurred in Los Angeles from 2010 to present with Victim's details along with the location where it occurred. This dataset contains the fields such as: DR Number, Date occurred, Time occurred, Area ID, Area Name, Reporting District, Crime Code, Crime Code Description, MO Codes, Victim Age, Victim Sex, Victim Descent, Premise Code, Premise Description, Address, Cross Street, Location.

Dataset Link: <https://catalog.data.gov/dataset/traffic-collision-data-from-2010-to-present>

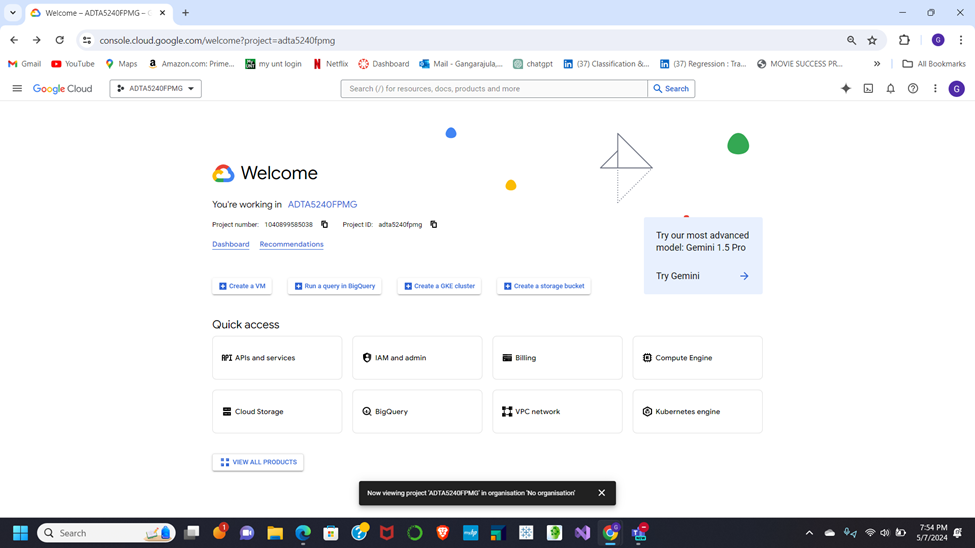
**Data Collection/ Processing:**

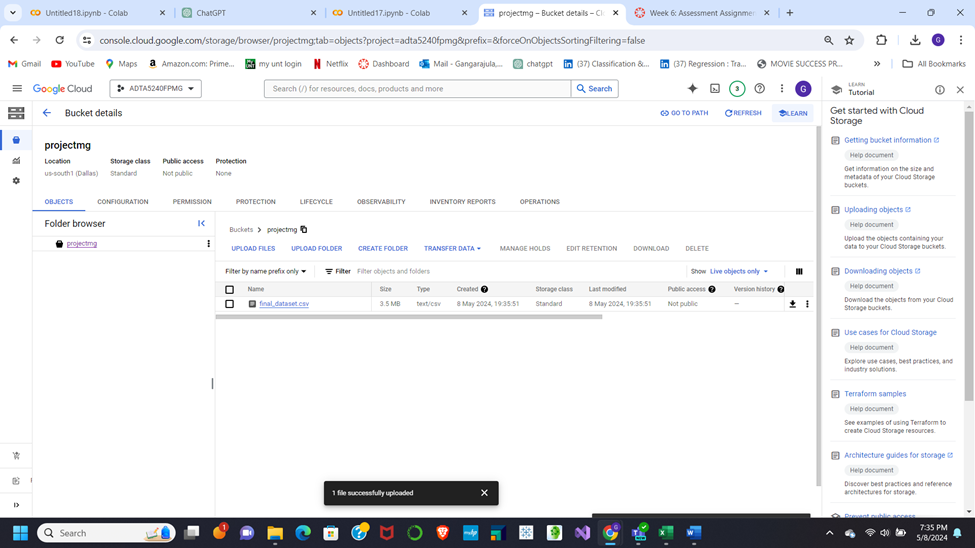
The collection and processing of data holds significant importance in the management and analysis of data. Gathering data means that data is gathered in different sources such as surveys, sensors, or databases to fulfil specific goals. This is an important point in which we need to make careful plans so that the data would be accurate and relevant. Subsequently, data processing occurs, where raw data undergoes refinement and transformation to rectify errors, standardize formats, and consolidate information from multiple sources. This phase is vital for preparing the data for analysis, guaranteeing its reliability, and structuring it appropriately for extracting insights or integrating it into decision-making models. Together, these stages form the cornerstone for conducting data analysis.

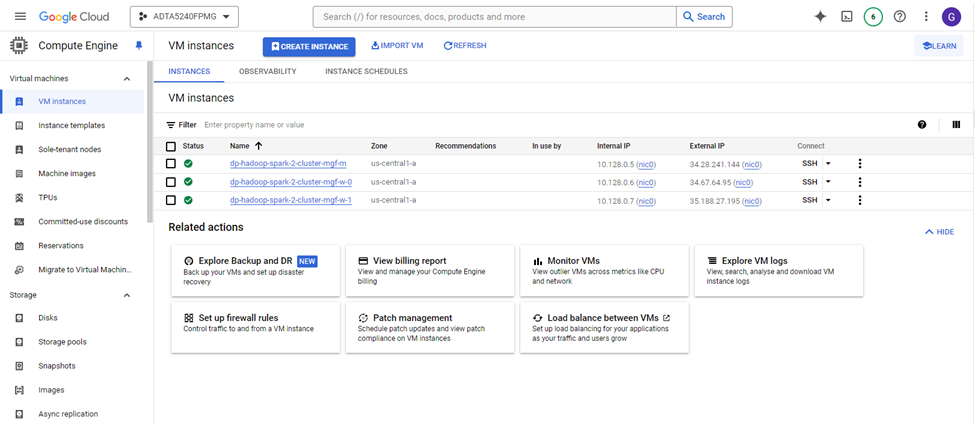
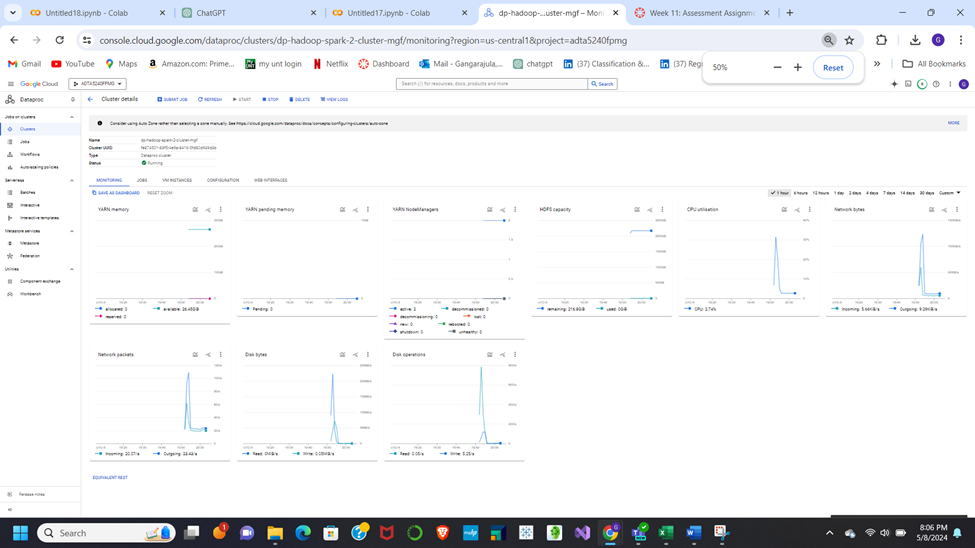
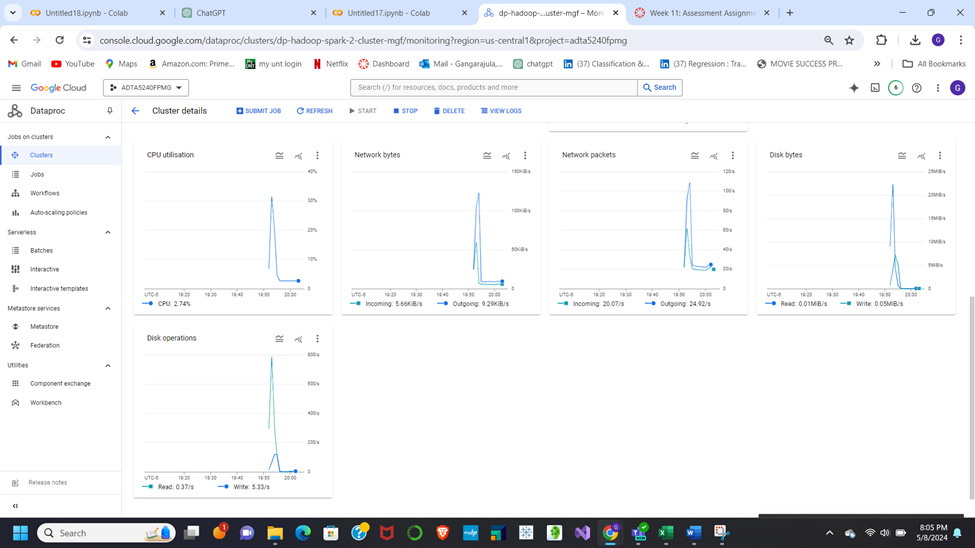
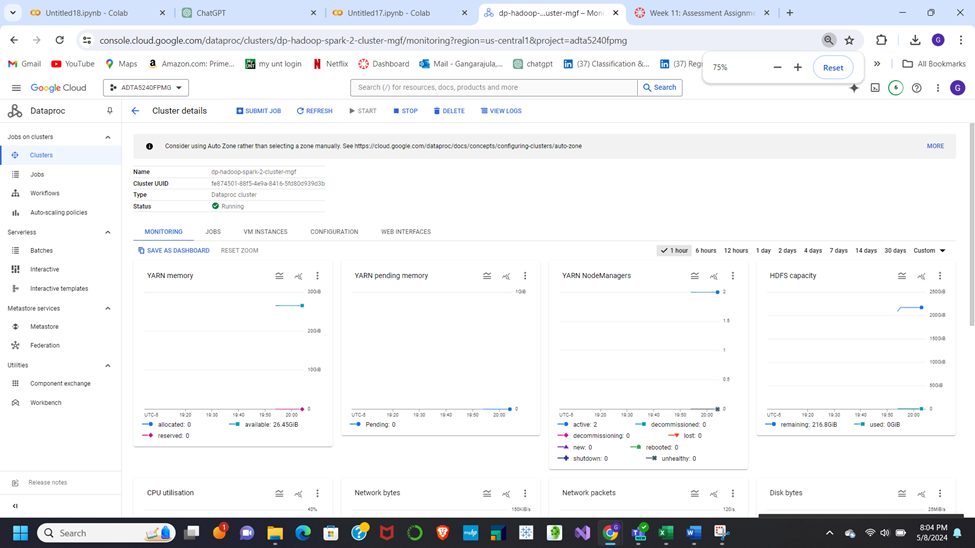
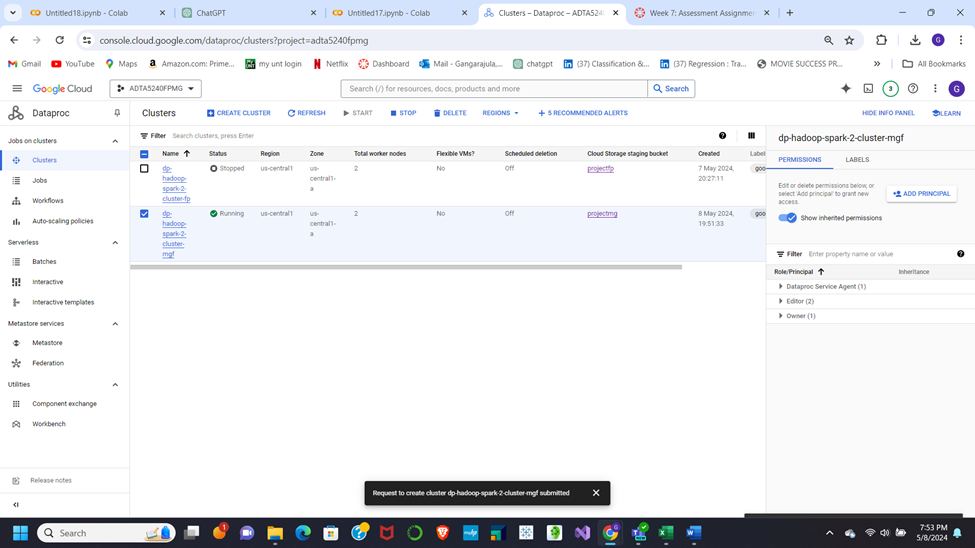
Dataset after cleaning:



**Data Storage/Management:**

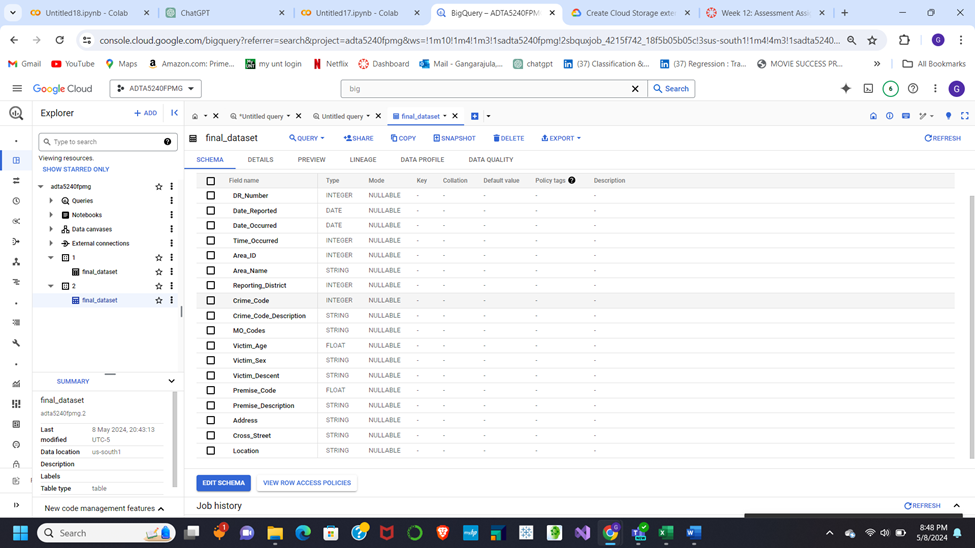
After cleaning the data, we need to store the data. We will be using GCP to store the data have created a new project with the name “ADTA5240FPMG”.Next we have created a bucket in the project that we have created with the name “projectmg” and uploaded the dataset that we have cleaned.



After enabling the Compute Engine API, we have successfully established clusters and nodes. These clusters and nodes are operational and ready for utilization in any subsequent analyses as needed. 

**Data Analysis:**

This stage of the life cycle is where the analysis will start to analyze the data, we will be using the BigQuery on the Google Cloud Platform.

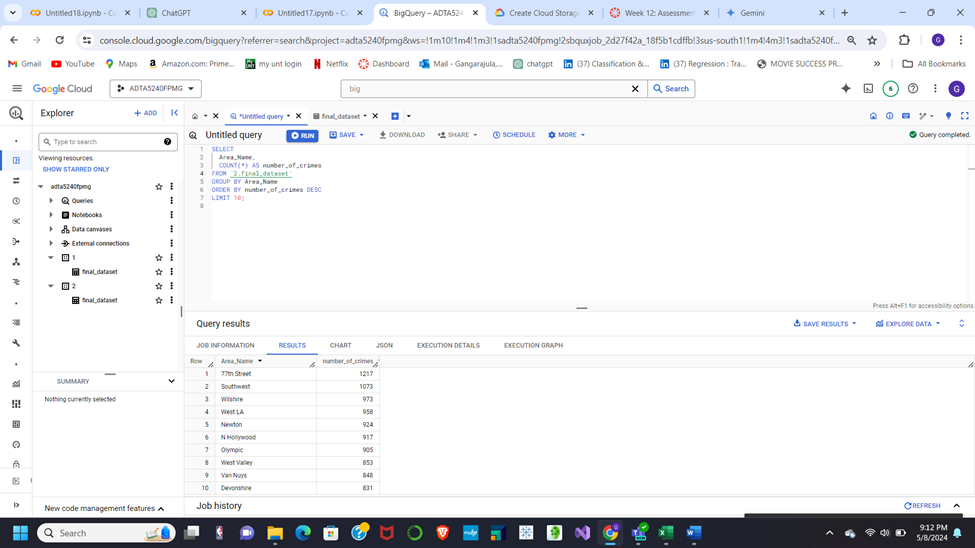


**Problem Statement:**

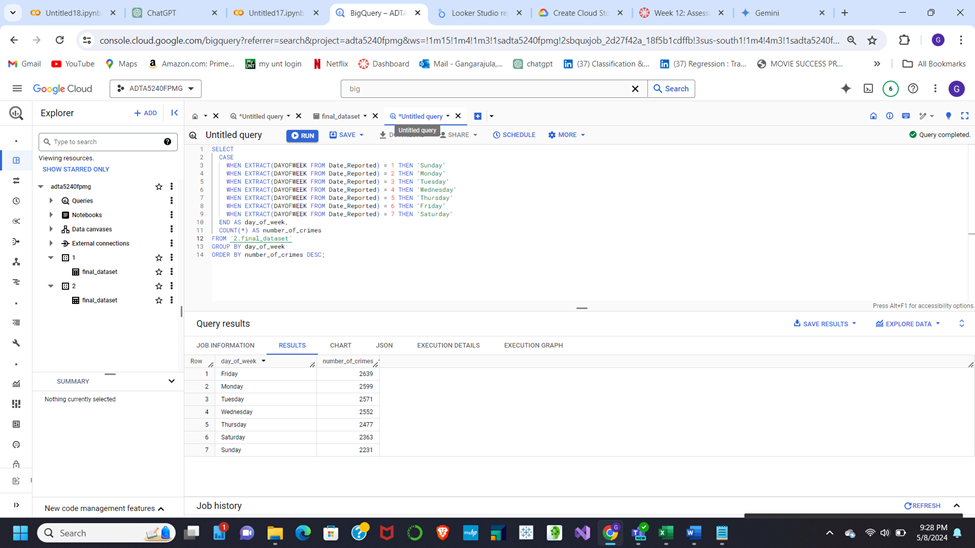
All throughout the world, we consider and deliberate over how to drive safely every day. It is impossible to comprehend how many people die in car accidents each year. There is still a pressing need for us to develop new strategies for improving traffic safety even though infrastructure and technology have rapidly changed. We suggest using empirically based analysis to address this issue since it will enable us to identify the primary causes of traffic accidents and develop targeted safety measures that will both lower the frequency of accidents and improve overall traffic safety.

1. Top Crime Locations:

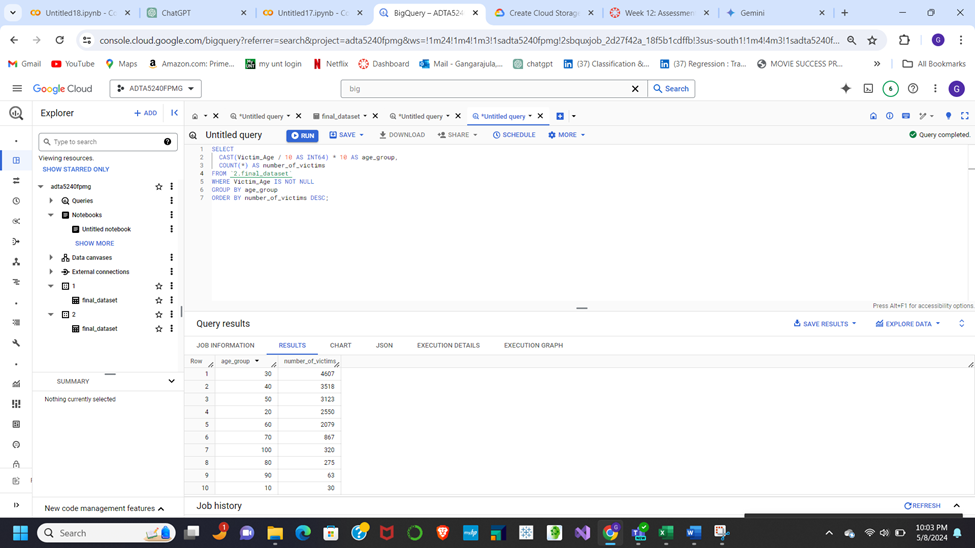
This query identifies areas with the highest number of reported crimes.



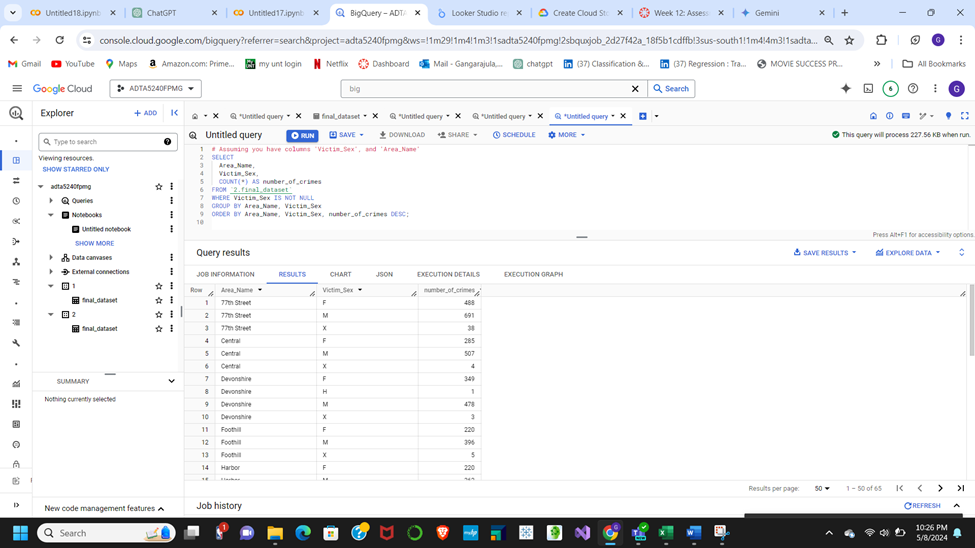
1. Crimes by Day of Week:



1. Victim Age Groups:



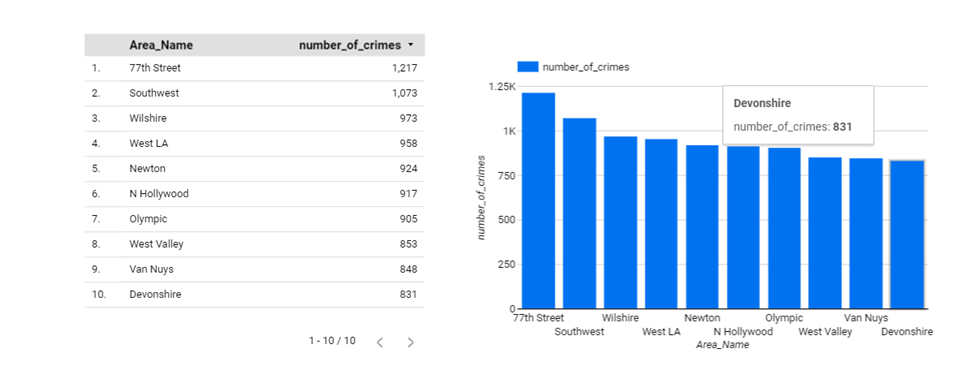
1. Analyze Crime Rates by Gender and Area:



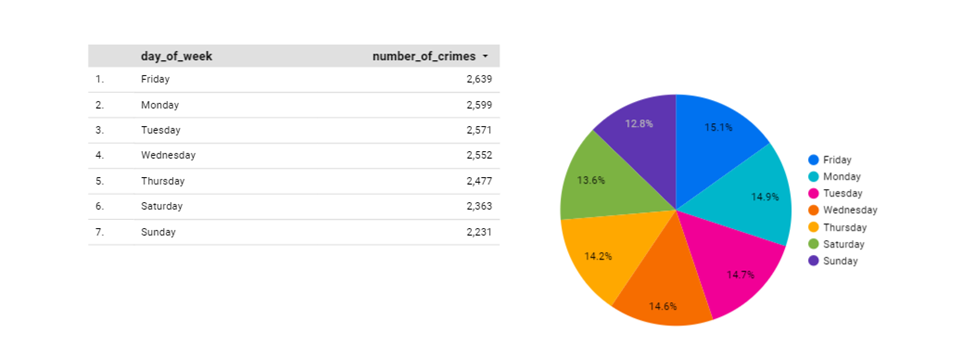
**Data Visualization:**

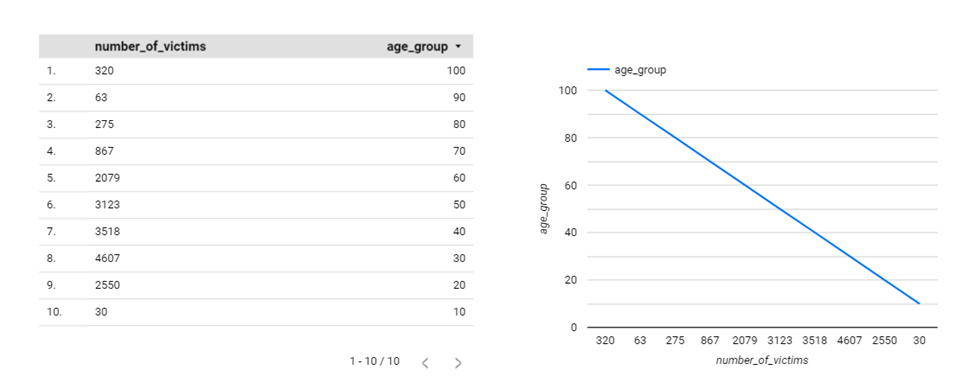
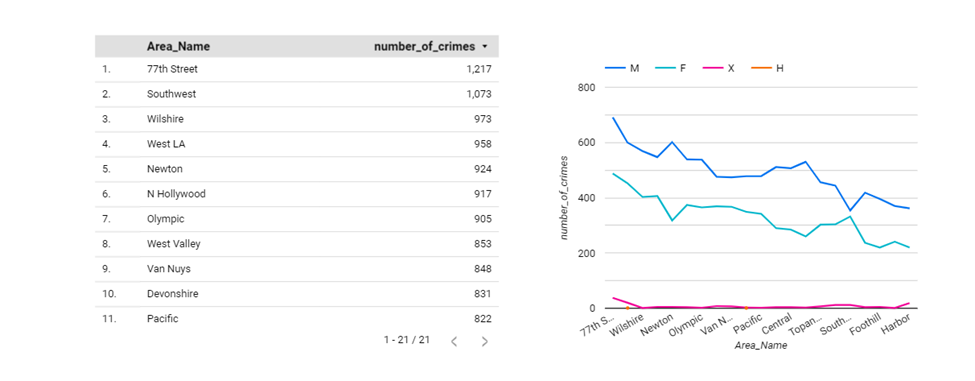
To successfully convey insights, data must be transformed into visual representations like charts, graphs, and maps. This process is known as data visualization. This methodology enables the visualization of patterns, trends, and anomalies present in datasets. Stakeholders can rapidly assimilate information and arrive at well-informed decisions through the display of facts. Through aesthetically pleasing representations, visualization improves the accessibility and comprehensibility of information, which is a critical function of data analysis. We will use Locker Studio for this project to complete this step.

The graph below is a bar graph for Problem statement 1 that depicts ascending to descending number of crimes in the city of Los Angeles with the area name and the count of the number of crimes. This is a graphical presentation that shows us the data visualization on our topic vividly.



The other kind of graphical representation for Problem statement 2 given below is a data visualization in pie chart format showing a percentage of the number of crimes with respect today of the week.

The data visualization (line chart) below is depicted on problem statement 3 with respect to number of victims versus the age groups.

 The line chart depicts the visualization of data correlating area names with the frequency of crimes, illustrating the frequency of incidents occurring in specific locations.

**Data Interpretation:**

Data interpretation, the last step in the data lifecycle, goes beyond visualization alone. While visualization presents the data in a clear format, interpretation adds meaning by explaining what the visualizations reveal. This understanding allows us to draw conclusions and recognize patterns, aiding decision-making, and policy formulation.

According to the data visualization analysis, in problem statement 1, we can conclude that there were more crimes in 77th street and Devonshire had the least. According to the data visualization analysis in problem statement 2, we can conclude that the number of crimes were observed to take place on Friday and the least number of accidents would occur on Sunday. In problem statement 3, the age groups 20-60 had more accidents than other age groups. In the 4th case, 77th street has the highest number of accidents.

**Conclusion:**

The insights provided by this data-driven study can be very helpful in enhancing traffic safety. Major areas for improvement were identified by means of data visualization and interpretation. First, according to the data, Devonshire has the least incidences while 77th Street is identified as a high-crime area. This shows that in high-risk regions like 77th street, focused measures are required to address safety problems. Second, there is the highest crime on Fridays and the lowest on Sundays. This demonstrates the significance of traffic control and enforcement, particularly on days with heavy traffic. Third, drivers between the ages of 20 and 60 are more likely to be involved in accidents, according to the research. This highlights how important it is to create awareness campaigns and specialized instructional programs for this group of people in order to encourage safe driving practices. Lastly, the data shows that the most accidents also occur on 77th street. This emphasizes how urgently this area needs traffic safety measures and infrastructural upgrades.

These results show the effectiveness of data-driven strategies overall. Targeted interventions can be developed to lower accident rates and enhance overall road safety by identifying critical elements that lead to accidents. All parties involved may develop more sustainable and safe transportation networks by utilizing data and evidence-based decision-making.