Data Analysis of USA Countrywide Accidents Report

***Motivation & Goal* —In many parts of the world, accident cases are ignored and are not taken seriously. Statistics show that there are between 20 and 50 million accidents per year. According to the National Highway Traffic Safety Administration, there are 64 million road crashes in the United States each year. Each day, 90 people die in traffic accidents. This is the leading cause of death in the USA. There is something that needs to be done about this. These statistics piqued our curiosity and motivated us to find out what causes accidents.**

The reduction of road accidents is an important public safety policy that requires extensive analysis. Since this is one of the major issues in the current world, we decided to analyze the accident dataset available from Kaggle. We analyzed traffic accident data thoroughly and performed analytics on the top 20 USA cities to identify contributing causes based on the trends and propose remedies.

Keywords—Google Big Query, Informatica cloud, oracle cloud MYSQL

I. Introduction

Data analysis involves gathering, transforming, cleaning, and modeling data to find the information needed. In today's world, data analytics techniques are used in every field, including business, finance, criminal justice, science, medicine, and government. You can take raw data and uncover patterns to extract valuable insights.

In terms of our project, we chose to work on a car accident database. The reason is those car accidents are not taken seriously; data shows that 19937 crashes occur every day in the United States. Road transport is the most unsafe circumstance individuals face every day, but casualty figures from such occurrences pull in less media attention than other, less visited types of catastrophes.

Before we analyzed the dataset, we made the preprocessing with the help of python such as removing the null values, removing the unnecessary columns, replacing the NULL values, etc. We analyzed the data from car accident databases and obtained results to provide some meaningful insights into the data. We also worked with the data to determine the major causes of the accidents, as well as the locations where the accidents occurred. We used different types of technologies such as the oracle cloud instance for deploying the on-premises MYSQL server and using google big query to perform analytics.

II. About the Data

First, Our dataset has approximately 619912 million recordings that are sourced from a variety of APIs. These include records from transportation organizations, police agencies, traffic cameras, and traffic sensors integrated into road networks. The data, which has been analyzed, contains information on occurrences that have taken place in 20 cities, with crucial details dispersed among 47 columns. The collection includes latitude and longitude information, which allows us to determine the exact location of a site and display the data using maps.

We decided to trim down this dataset by removing about 4 columns that contained extraneous information from the 47 columns. The civil twilight, nautical light, and astronomical twilight columns have been removed because all the data in these columns are contained in the column Sunrise Sunset. This column indicates when sunrise or sunset occurs. Several columns contained missing data, so we calculated the percentage of missing values in each column to identify any inconsistencies. Column 'number' has 61% of the null values of any column, so it was removed. So, all the columns with fewer missing values have been linearly interpolated.

Since the data set contains numerous abnormalities, we normalized it to satisfy 1NF, 2NF, and 3NF. Data was divided into 5 CSV files as entities after normalization, and relationships between these entities were discovered. Those 5 CSV files are accident details, location details, road details, weather details, and state city. So, the details related to accidents are stored in the accident details CSV, and other details related to location, weather, and road are stored in their respective CSV with ID numbers for all the data to make the data unique. You can find the details of the features with their descriptions.

	A	B
1	**Column Nmae**	**Description**
2	ID	This is a unique identifier of the accident record.
3	Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic
4	Start_Time	Shows start time of the accident in local time zone.
5	End_Time	Shows end time of the accident in local time zone.
6	Start_Lat	Shows latitude in GPS coordinate of the start point.
7	Start_Lng	Shows longitude in GPS coordinate of the start point.
8	End_Lat	Shows latitude in GPS coordinate of the end point.
9	End_Lng	Shows longitude in GPS coordinate of the end point.
10	Distance(mi)	he length of the road extent affected by the accident.
11	Description	Shows natural language description of the accident.
12	Number	Shows the street number in address record.
13	Street	Shows the street name in address record.
14	Side	Shows the relative side of the street (Right/Left) in address record.
15	City	Shows the city in address record.
16	County	Shows the county in address record.
17	State	Shows the state in address record.
18	Zipcode	Shows the zipcode in address record.
19	Country	Shows the country in address record.
20	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).
21	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.
22	Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).
23	Temperature(F)	Shows the temperature (in Fahrenheit).
24	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).
25	Humidity(%)	Shows the humidity (in percentage).
26	Pressure(in)	Shows the air pressure (in inches).
27	Visibility(mi)	Shows visibility (in miles).
28	Wind_Direction	Shows wind direction.
29	Wind_Speed(mph)	Shows wind speed (in miles per hour).
30	Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset
31		
32		
33		

III. Project Overview and Architecture

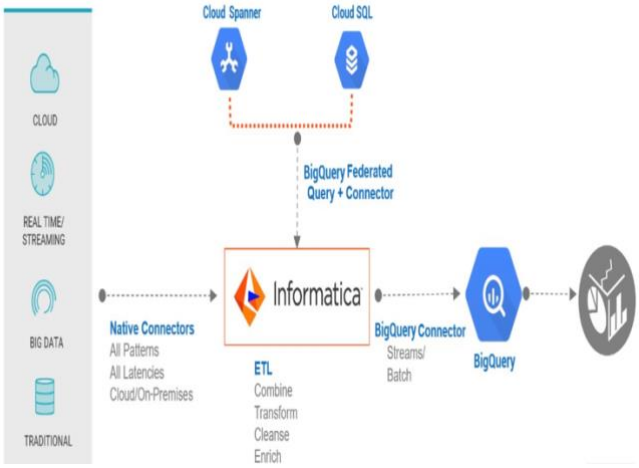
We are building a robust cloud model for our project. Considering efficiency, usability, and cost, we chose Google Cloud Platform, which offers the fastest and most reliable network architecture and latency. In terms of data, we got our dataset from Kaggle which is unnormalized and includes US accidents that occurred in 49 states.

However, in our model, we chose the top 20 cities with the highest accident rate. In total, 619912 data points were collected from 20 cities. We normalized the data into 5 CSV files which are accident details, location details, road details, weather details, and state city.

We have developed python scripts in such a way that the datasets have been deployed directly to the MySQL cloud

database, configured in an oracle cloud instance. The Cloud instance is configured with a private VPC network that only authorized individuals can access post we have used Informatica cloud as an ETL tool to deploy our data from the oracle cloud MYSQL instance to the google Big query instance, also known as GBQ.

One successful connection has been established in Informatica ETL between oracle MYSQL and cloud BIG query. The data has been successfully transformed into google big query. Finally, using the BIQ SQL queries, analytics is performed and saved in separate new tables in the same dataset that is used for data reporting.



IV. WORKFLOW

The project flow is mainly divided into seven steps and each step has its own characteristics that cannot be skipped in the entire flow.

1. Set up an Oracle cloud account OCI and create a virtual instance.
2. Install the MYSQL server by using the docker.
3. Create an account with Informatica cloud and run the Informatica cloud agent service.
4. Install the Google big query and MYSQL server extension on the Informatica cloud ETL.
5. Configure both instances in the Informatica cloud ETL.
6. Successful connection at both ends will allow us to transfer the data.
7. Create the mapping between the two on-premises objects and run the job.

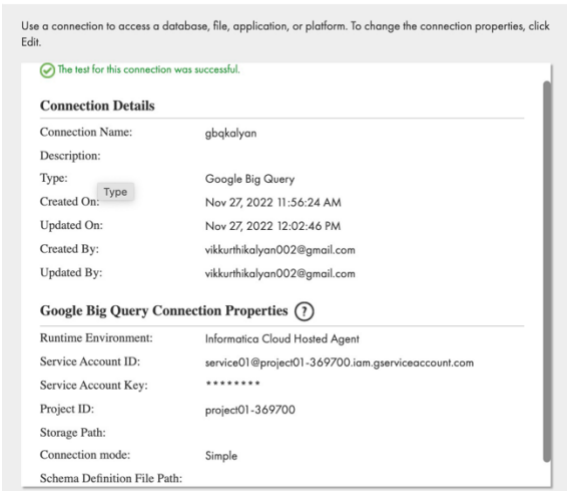
V. ETL(EXTRACT TRANSFORM LOAD) PROCESS

We have chosen the Informatica cloud as an ETL process. The Extract, Transform, and Load Process was performed by writing Python scripts and coordinating with them as this is an information integration device based on ETL architecture. It offers data integration software and services to a wide range of businesses, industries, and government organizations, including those in telecom, health care, finance, and insurance. Informatica PowerCenter is a high-end data integration tool. It is one the most effective data transformation tools and converts data from one application to another's application. The major reliability of using Informatica is it reads the data, row by row, from a table (or group of related tables) in a database, or from a file.

A. *Extract*

Using an Informatica cloud agent service, relational transactional data was extracted. The connection was made with a connection string that included the project name, the region where the Cloud SQL instance is hosted, and the name of the Cloud SQL Instance. The authorization is carried out with the help of appropriate IAM roles and permissions to access the Cloud SQL instance, as well as user and secured password parameters. After successfully establishing the connection, the data were extracted by selecting the appropriate fields from the source object MYSQL. Please find the below picture for your reference about the connection properties that we used for the oracle cloud MYSQL.

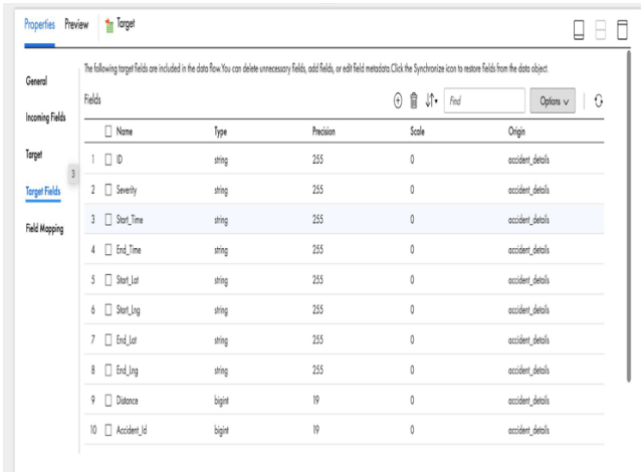
Configuration files:



B. *Transformation*

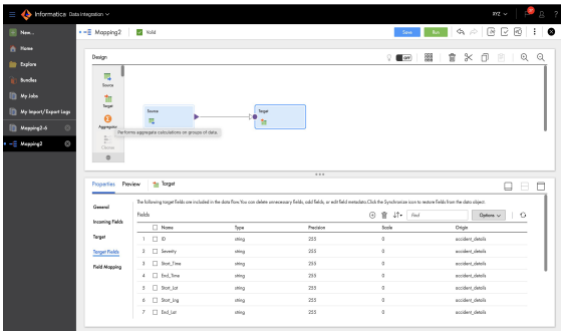
Data preprocessing was performed using pandas after the data was extracted and saved in a panda's data frame object.

Several data transformation steps are being performed, including renaming columns according to big query standards, handling null values, and imputing them with appropriate measures such as last known value, mean, and so on, depending on the use case. We also added extra columns that are more suitable for analysis. At the time of transformation, the Informatica ETL tool provides us the flexibility to select what columns we need for the target destination and can also perform data type manipulations and can perform some mathematical functions to get a new column. Please find the below screenshot for your reference with the details of the target fields selection and successful job properties.

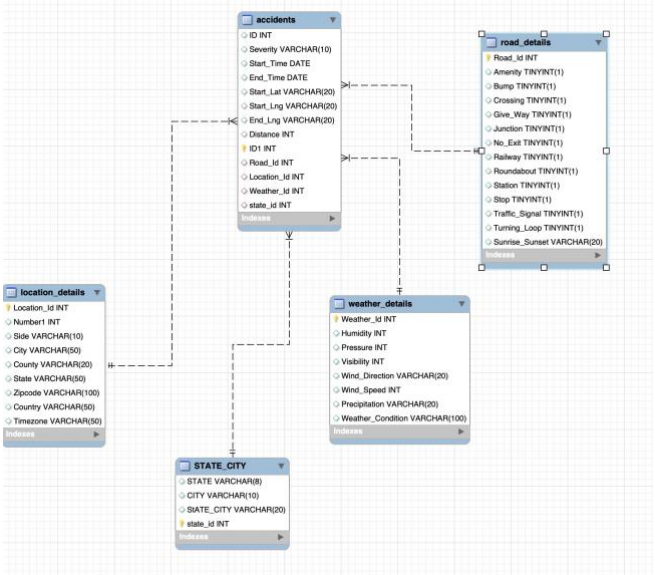


C. Transformation or Deploy

At the time of deploying to the target source, the Informatica cloud agent will take care of all the necessary environments that are required for the source and target databases. Once we establish the connection from the oracle cloud MYSQL database and Google big query we can save the mapping and run the jobs. Informatica will then ingest the data into a Cloud Big Query Instance as a destination object. Cloud logging has been implemented at each stage to track the data flow throughout the pipeline. It will be helpful when if we face any errors at the time of running the jobs to capture Exception Handling. Please find the below screenshot for detailed information about how the pipeline is developed and respective jobs run successfully without any errors.



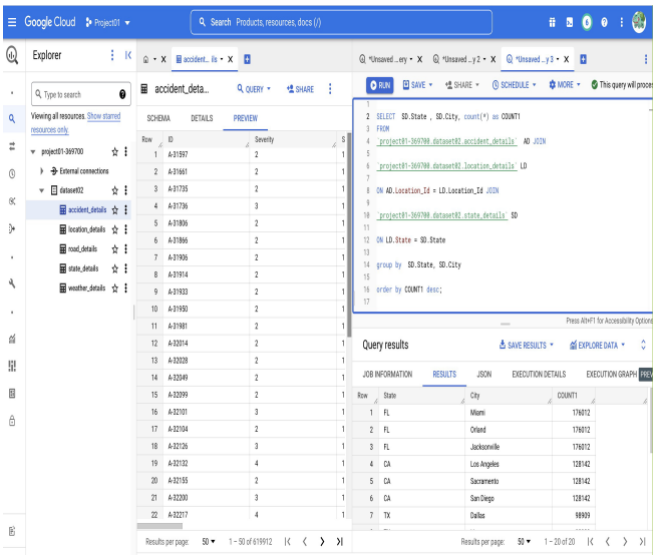
D. MYSQL Model: This is an MYSQL database model that has been deployed into the oracle cloud instance with the help of docker.



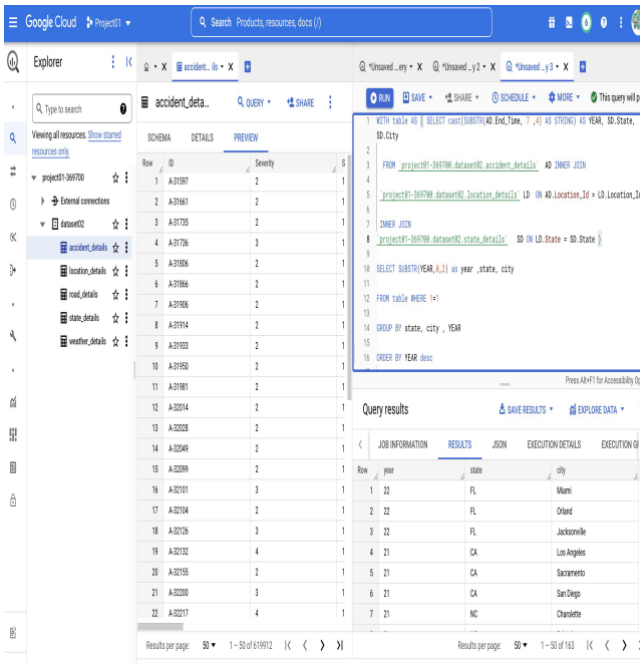
VI. Data Analytics for Execution Queries

After successful data migration to the google big query, we have performed various analytics from the tables that are imported. Google big query provides a lot of flexibility that supports the structured query language as same as a relational database.

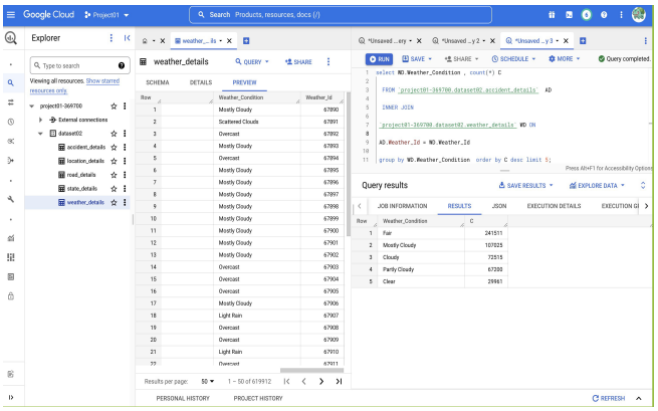
A. Number of accidents occurred across each state and city combination



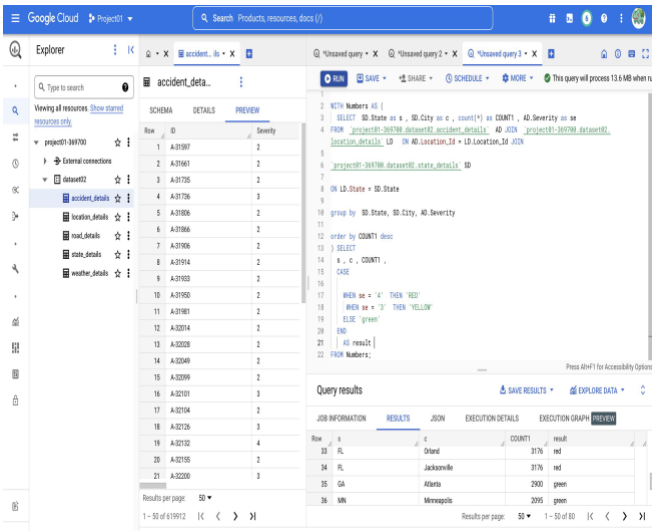
B. *The number of accidents that occurred each year*



C. *The most affected weather conditions for most of the accidents.*

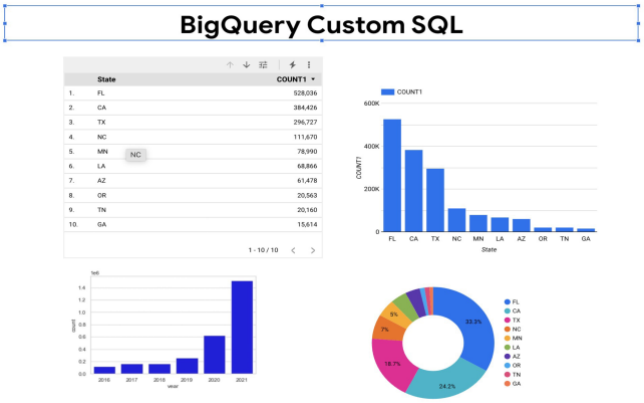


D. *The severity of the accidents that occurred in each state and city.*

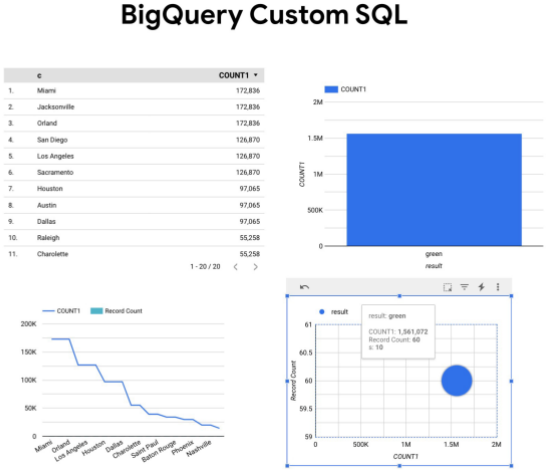


Big Query Looker Studio above results:

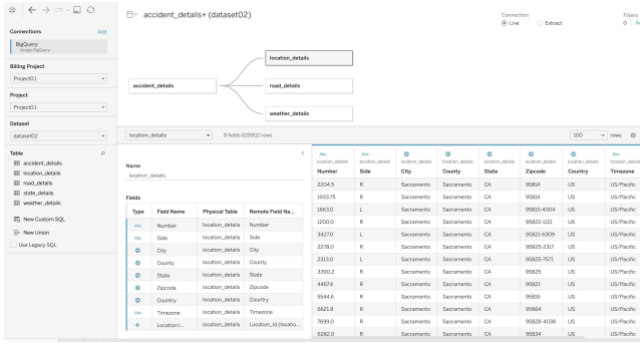
Top 20 states with most number of accidents occurred.



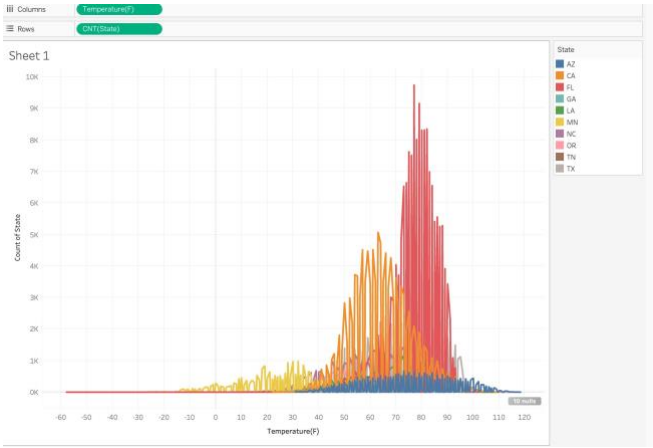
The most number of accidents occurred in terms of severity and could see that most of the accidents are having low severity.



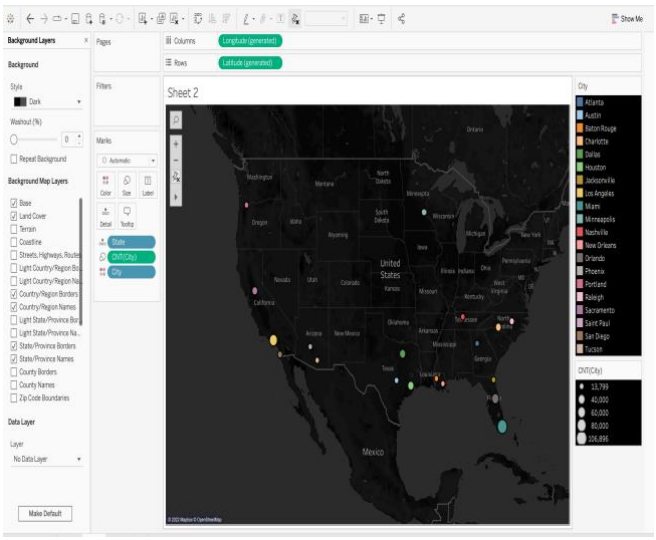
*E. The **Visualization with Tableau***



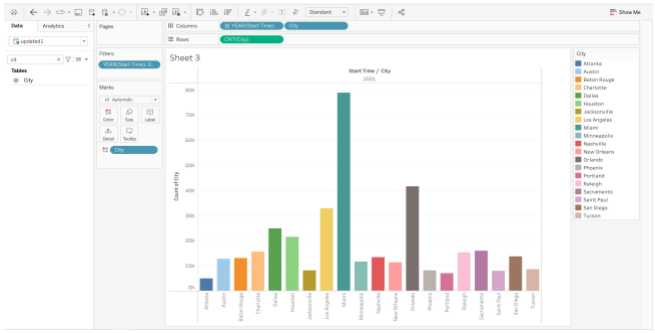
F. It is seen that temperature plays a major role in deciding accidents and could see that Florida recorded the highest across the states.



G. It is the column s latitude and longitude, we have developed a Map chart where state and city are shown with their own positions



H. We have checked the city-wise accidents that occurred in the year 2021 and could see that Miami took the place highest.



VII. LIMITATIONS

- In the data, there are outliers where a small number of cities have very few data points, resulting in abnormal values. This abnormality causes all analyses to be altered, resulting in incorrect results. The anomalies in the data took time to discover, and we removed them all. Weather analysis for accidents includes precipitation as one of the most significant features. However, the dataset has many missing values, so we used the top 20 cities where all precipitation values were accurate.

VIII. KEY LEARNINGS

- We have explored the various configured as part of the ETL tool such as Oracle cloud MTSQL database and Google Big Query.
- We understand that studying the original documentation plays a key role when we are working with new technologies.
- We get hands-on experience in the fields of oracle cloud MYSQL databases, Google big query, and Google Cloud, Informatica tool.
- Moreover, worked on a data integration tool in Informatica cloud agent service which is a part of ETL architecture.
- We were able to analyze how road accidents affect current road transport with the dataset

IX. TECHNICAL DIFFICULTY

- The datasets we have collected on Kaggle contain several contradictory and missing variables. The clarification of this data took a long time.
- We have filled in the missing data and found the relation between all the attributes to normalize into the 3NF form.

- “compliment”, “discreet” and “discrete”, “principal” and “principle”.
- Initially, we were unable to configure the Informatica ETL for the source oracle cloud MYSQL and google big query. Took lots of time to figure it out and to make a successful connection.
- We also encountered problems in loading the data into the oracle cloud MYSQL database and we have developed a python script to deploy the data into the database.

X. Conclusions

Analyzed how road accidents affects the current road transport and factors for increase in accident rate.

We analyzed the data through google big query looker studio and also connected google big query server to tableau and analyzed in it too.

So, in our analysis we successfully found that even during the COVID period, accidents continue to rise even though the number peaked in 2021.
And also there are few analysis we made which tells the weather conditions effecting the accident rate and severity of it.

XI. Future Work

For this project, we worked on static data but in the future, we wanted to work with real-time data. With real-time data, we will have many challenges, but we want to take up the challenge and analyze the real-time data. For the present analysis, we used the top 20 cities but, in the future, we want to work on all the cities in the USA because there might be a few more trigger points we can get from such big data.

GitHubLink:

https://github.com/MylarusettyNikhil/DB_GROUP5