CALIFORNIA TRAFFIC COLLISION ANALYSIS
PROJECT DOCUMENT

Abstract

The project focuses on studying the road traffic accidents specifically in the state of California from the year of 2019 –2020 to analyze the dataset as recorded by California Highway Patrol and gain insights

Team: Group 10

CALIFORNIA TRAFFIC COLLISIONS

Table of Contents

CALIFORNIA TRAFFIC COLLISIONS

Introduction

The increase of automobile demand has also increased the traffic rate across many countries. The increase in automobile demand has also increased the traffic and the traffic collision rate compared to the last decade. Over 1.2 million individuals die every year on the world's streets, and somewhere in the range of 20 and 50 million endure non-fatal injuries. In 2019, California's mileage death rate was 1.06 fatalities per 100 million miles traveled. That's according to the Global Traffic Scorecard released in March 2020 by INRIX, a data analytics company that studies how people move around the world – San Francisco, California was rated 7th among the most congested cities in the U.S. in 2019. The collisions are caused due to various reasons such as alcohol or drug consumption, cellphone in use, motorcycle, bicycle, pedestrian etc. During the pandemic in 2020 nationwide lockdown had restricted the movement of traffic. In this study, we are analyzing the dataset of Traffic collision rate of California from the Kaggle which was provided by California Highway portal records to the author. The data was collected, clean, manipulated, tabulated, and then analyzed. The analysis shows the accident prone regions of the country and the comparison of the fatal and non-fatal collisions in year 2019 and 2020.The significant findings from the analysis are: (a) most of the fatal and non-fatal accidents occurred in January 2019 (b) a more substantial number of deaths are from drivers in the 20-50 age group; (c) Most of the fatal collisions have occurred during the cloudy weather followed by rainy weather (d) about 13% accidents are attributed to drunk driving. (e) Passenger cars contribute to the highest number of collisions in 2019 and 2020. 22,464 collisions occurred in 2019 which dropped to 8320 in 2020 during COVID (f) Top 3 traffic violations leading to collisions were : Not following Traffic guidelines(11%), Unsafe speed (22% approx.) and Improper turns(17%).

CALIFORNIA TRAFFIC COLLISIONS

Dataset Size and Complexity

Data Source

The California Traffic collisions dataset that we have used is from Kaggle repository which is collected from the California Highway Patrol records by the author. It covers collisions from January 1st, 2001, until December 2020 and is available in form of SQLite Database. The dataset contains ~ 9 Million unique collision cases information

There are three main tables:

- **collisions:** Contains information about the collision, where it happened, what vehicles were involved
- **parties:** Contains information about the groups people involved in the collision including age, sex, and sobriety
- **victims:** Contains information about the injuries of specific people involved in the collision

Dataset Link

Dataset Size and Complexity

The dataset size of original dataset is 10 GB. Below are the statistics of columns in each table

1) *Collisions*: 75 columns, 9424334 rows
2) *Parties*: 32 columns, 18669166 rows
3) *Victims*: 12 columns, 9639334 rows

Data Wrangling

As part of the project scope, we selected traffic collisions that occurred in year 2019 and 2020. There are ~ 5k cases of traffic collisions.

The data set has 49958 unique cases and their related collisions, parties and victims data, it is not ready to use for analysis.

There are many anomalies in the dataset like:

- Null records
- Duplicate records
- Mismatched column

Below cleansing activities were performed to handle the anomalies and other discrepancies in dataset:

- To maintain data integrity , few duplicate records where deleted
- To handle data redundancy in source tables, the dataset was modelled into 3 NF form
- Datatypes of few columns were not recognized by tableau, therefore, datatype was changes
- To simplify the data complexity irrelevant columns were dropped
- Implemented relationship and constraints like primary key, foreign key, indexes, check constraints, assigned default values as part of data profiling and accuracy

CALIFORNIA TRAFFIC COLLISIONS

After data wrangling, the dataset size is 46.6 MB. It contains 9 entities as below:

Table Name	Columns	Rows
Collisions	35	49958
Parties	28	100674
Victims	12	45871
County	2	47
Collisions_location	7	41996
Road_Condition	3	46
Vehicle_Type	2	16
Violation	3	255
Weather_Effect	3	30

CALIFORNIA TRAFFIC COLLISIONS

Data Model

Normalization

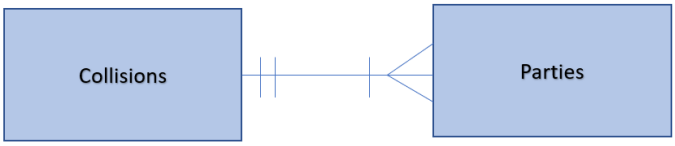
The initial California traffic collision dataset available contained three entities - Collisions, Parties and Victims in denormalised form. As part of this project, we have normalized the entities up to third Normal Form modelling it into a snowflake schema.

Entity Relationship Model

The entity-relationship (ER) model and its accompanying ER diagrams are widely used for database design and systems analysis.
California traffic collision entity relationship model is composed of entity type and specifies relationships that can exist between entities.

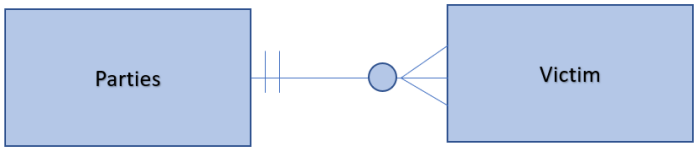
- **Collisions -Parties:** Collisions contain one entity instance (Case ID) for each traffic collision that occurs. For each collision there exists parties involved in the collision. For example: two cars collided with each other, therefore, there are two parties involved in the accident. Also, there must be at least on party involved in accident to cause collision

  Relationship: Collisions ➔ [Mandatory One -to-Many] ➔ Parties



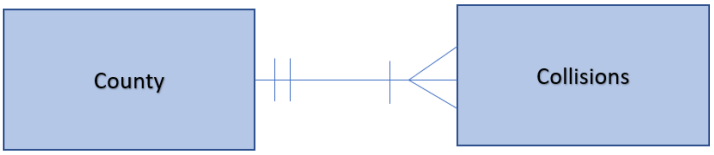
- **Parties – Victims:** For each party involved in the collision, victims may or may not exists, for example, no one was injured due to the collision. Also, for each party there can be one or more victims. For example, two cars A and B collided. There was driver plus two passengers in Car A and driver and one passenger of Car A gets injured. Therefore, there are two victims for party A during the collision

  Relationship: Parties ➔ [Optional One to Many] ➔Victims



- **County - Collisions:** Multiple collisions can occur in a particular county over a period. In a rare scenario, there might not be any collision occurring in a particular county

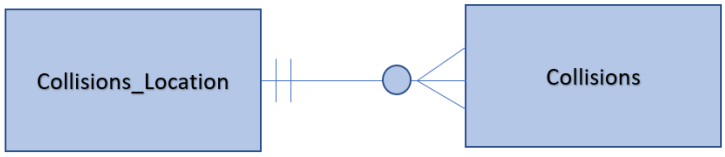
  Relationship: County ➔ [Optional One to Many] ➔ Collisions



CALIFORNIA TRAFFIC COLLISIONS

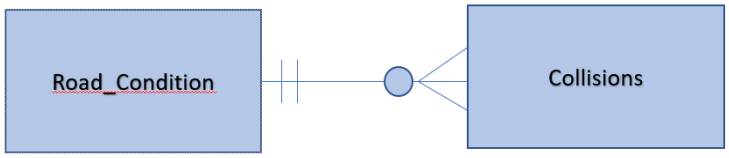
- **Collisions_Location - Collisions:** Multiple collisions can occur in a particular location (street, city etc.) over a period. In a rare scenario, there might not be any collision occurring in a particular location

  Relationship: Collisions_Location ➔ [Optional One to Many] ➔ Collisions



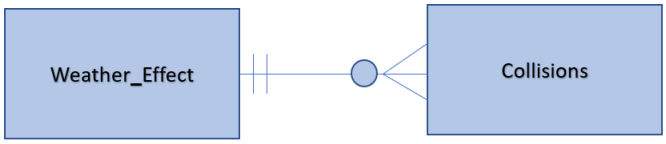
- **Road_Condition - Collisions:** Same type of road condition and lighting can cause may or may not cause collisions

  Relationship: Road_Condition ➔ [Optional One to Many] ➔ Collisions



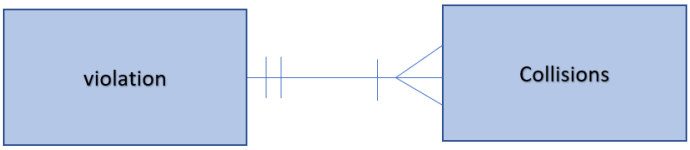
- **Weather Effect – Collisions:** Same type of weather conditions and road surface conditions may or may not cause traffic collisions

  Relationship: Weather_Effect ➔ [Optional One to Many] ➔ Collisions



- **Violation- Collisions:** From this relation we can determine in a collision which violation category was the cause

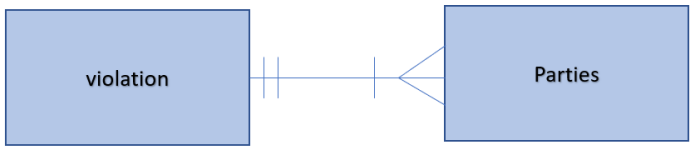
  Relationship: violation ➔ [Optional One to Many] ➔ Collisions



CALIFORNIA TRAFFIC COLLISIONS

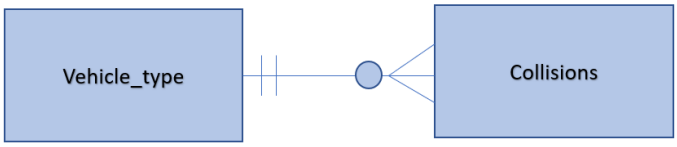
- **Violation -Parties:** From this relation we can determine which sort of violation was caused by the parties involved

  Relationship: violation ➔ [Optional One to Many] ➔ Collisions



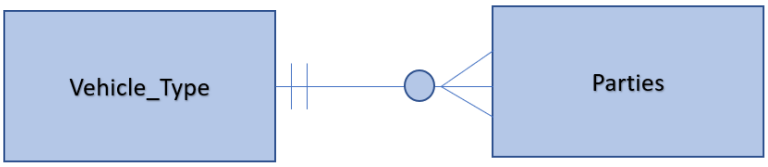
- **Vehicle_Type – Collisions:** From this relation we can infer what type of vehicles were involved in a collision

  Relationship: vehicle_type ➔ [Optional One to Many] ➔ Collisions



- **Vehicle_type - Parties:** From this relation we can infer what type of vehicles were used by the parties

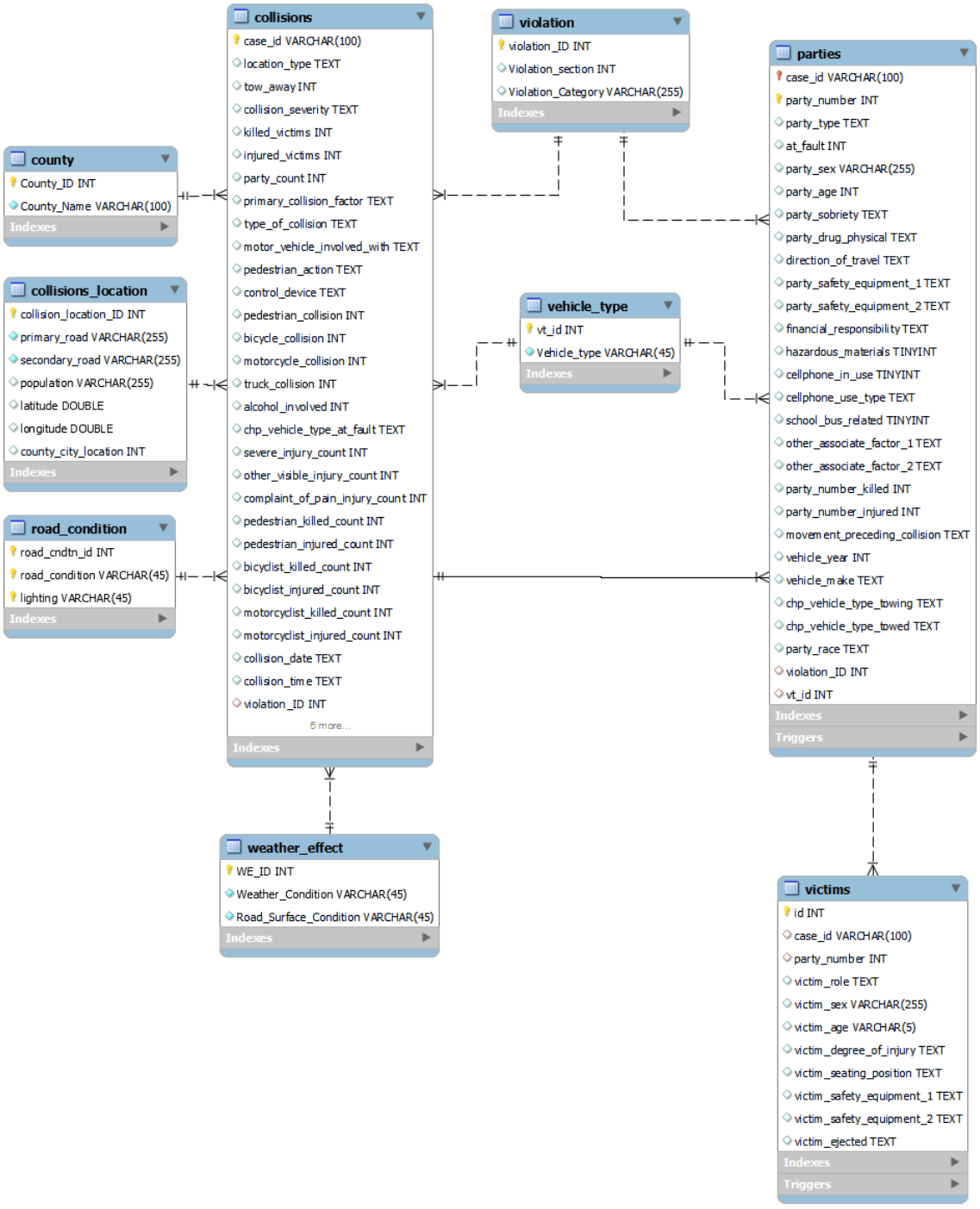
  Relationship: vehicle_type ➔ [Optional One to Many] ➔ Collisions



CALIFORNIA TRAFFIC COLLISIONS

Entity -Relationship Diagram

Below is the entity relationship diagram for the database `Cal_Road_Accident`



CALIFORNIA TRAFFIC COLLISIONS

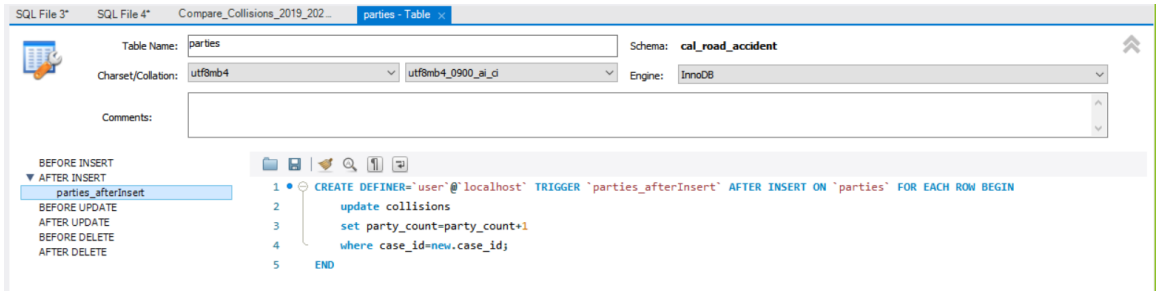
Usage of Triggers and Procedures

Triggers

Trigger Name: parties_afterInsert

When a record is inserted in victims table for a particular case update party_count in collisions table. This field gives the aggregated count of party involved in a particular collision briefly

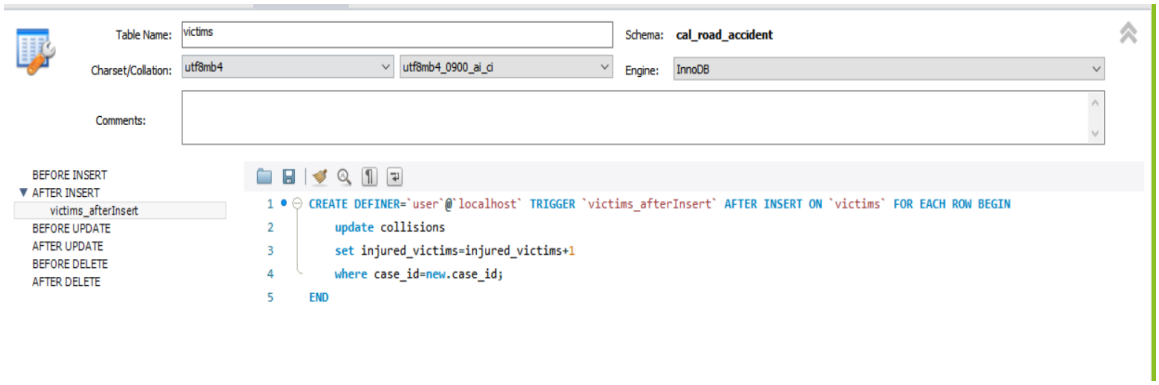
- Code Snippet:



Trigger Name: victims_afterInsert

When a record is inserted in victims table for a particular case update victim_count in collisions table. This field gives the aggregated count of victims in a particular collision briefly

- Code Snippet:



CALIFORNIA TRAFFIC COLLISIONS

Procedures

Procedure Name: Compare_Collisions_2019_2020

Stored procedures can reduce network traffic between clients and servers, because the commands are executed as a single batch of code. This means only the call to execute the procedure is sent over a network, instead of every single line of code being sent individually. We have created a stored procedure Compare_Collisions_2019_2020 and this procedure is further used for visualization.

- Code Snippet:

CALIFORNIA TRAFFIC COLLISIONS

Data Design

Table Structure and Business Rules

Table Name: Collisions

Field	Type	Null	Key	Default
case_id	varchar(100)	NO	PRI	NULL
location_type	text	YES		NULL
tow_away	int	YES		NULL
collision_severity	text	YES		NULL
killed_victims	int	YES		NULL
injured_victims	int	YES		NULL
party_count	int	YES		NULL
primary_collision_factor	text	YES		NULL
type_of_collision	text	YES		NULL
motor_vehicle_involved_with	text	YES		NULL
pedestrian_action	text	YES		NULL
control_device	text	YES		NULL
pedestrian_collision	int	YES		NULL
bicycle_collision	int	YES		NULL
motorcycle_collision	int	YES		NULL
truck_collision	int	YES		NULL
alcohol_involved	int	YES		NULL
chp_vehicle_type_at_fault	text	YES		NULL
severe_injury_count	int	YES		NULL
other_visible_injury_count	int	YES		NULL
complaint_of_pain_injury_count	int	YES		NULL
pedestrian_killed_count	int	YES		NULL
pedestrian_injured_count	int	YES		NULL
bicyclist_killed_count	int	YES		NULL
bicyclist_injured_count	int	YES		NULL
motorcyclist_killed_count	int	YES		NULL
motorcyclist_injured_count	int	YES		NULL
collision_date	text	YES		NULL
collision_time	text	YES		NULL
violation_ID	int	YES	MUL	NULL
county_ID	int	NO	MUL	NULL
location_ID	int	YES	MUL	NULL
WE_ID	int	YES	MUL	NULL
rc_ID	int	YES	MUL	NULL
vehicle_type_at_fault_id	int	YES	MUL	NULL

CALIFORNIA TRAFFIC COLLISIONS

Table Name: Parties

Field	Type	Null	Key	Default
case_id	varchar(100)	NO	PRI	NULL
party_number	int	NO	PRI	NULL
party_type	text	YES		NULL
at_fault	int	YES		NULL
party_sex	varchar(255)	YES		NA
party_age	int	YES		NULL
party_sobriety	text	YES		NULL
party_drug_physical	text	YES		NULL
direction_of_travel	text	YES		NULL
party_safety_equipment_1	text	YES		NULL
party_safety_equipment_2	text	YES		NULL
financial_responsibility	text	YES		NULL
hazardous_materials	tinyint	YES		NULL
cellphone_in_use	tinyint	YES		NULL
cellphone_use_type	text	YES		NULL
school_bus_related	tinyint	YES		NULL
other_associate_factor_1	text	YES		NULL
other_associate_factor_2	text	YES		NULL
party_number_killed	int	YES		NULL
party_number_injured	int	YES		NULL
movement_preceding_collision	text	YES		NULL
vehicle_year	int	YES		NULL
vehicle_make	text	YES		NULL
chp_vehicle_type_towing	text	YES		NULL
chp_vehicle_type_towed	text	YES		NULL
party_race	text	YES		NULL
violation_ID	int	YES	MUL	NULL
vt_id	int	YES	MUL	NULL

CALIFORNIA TRAFFIC COLLISIONS

Table Name: Victims

Field	Type	Null	Key	Default	Extra
id	int	NO	PRI	NULL	auto_increment
case_id	varchar(100)	YES	MUL	NULL	
party_number	int	YES	MUL	NULL	
victim_role	text	YES		NULL	
victim_sex	varchar(255)	YES		NA	
victim_age	varchar(5)	YES		NULL	
victim_degree_of_injury	text	YES		NULL	
victim_seating_position	text	YES		NULL	
victim_safety_equipment_1	text	YES		NULL	
victim_safety_equipment_2	text	YES		NULL	
victim_ejected	text	YES		NULL	
party_ID	mediumint	NO	MUL	NULL	

Table Name: Collisions_Location

Field	Type	Null	Key	Default	Extra
collision_location_ID	int	NO	PRI	NULL	auto_increment
primary_road	varchar(255)	NO	MUL	NULL	
secondary_road	varchar(255)	NO	MUL	NULL	
population	varchar(255)	YES	MUL	NULL	
latitude	double	YES	MUL	NULL	
longitude	double	YES	MUL	NULL	
county_city_location	int	YES	MUL	NULL	

Table Name: County

Field	Type	Null	Key	Default	Extra
County_ID	int	NO	PRI	NULL	auto_increment
County_Name	varchar(100)	NO	UNI	NULL	

Table Name: Road_Condition

Field	Type	Null	Key	Default	Extra
road_cndtn_id	int	NO	PRI	NULL	auto_increment
road_condition	varchar(45)	NO	PRI	NULL	
lighting	varchar(45)	NO	PRI	NULL	

CALIFORNIA TRAFFIC COLLISIONS

Table Name: Vehicle_Type

Field	Type	Null	Key	Default	Extra
vt_id	int	NO	PRI	NULL	auto_increment
Vehicle_type	varchar(45)	NO		NULL	

Table Name: Violation

Field	Type	Null	Key	Default
violation_ID	int	NO	PRI	NULL
Violation_section	int	YES		NULL
Violation_Category	varchar(255)	YES		NULL

Table Name: Weather_Effect

Field	Type	Null	Key	Default	Extra
WE_ID	int	NO	PRI	NULL	auto_increment
Weather_Condition	varchar(45)	NO		NULL	
Road_Surface_Condition	varchar(45)	NO		NULL	

CALIFORNIA TRAFFIC COLLISIONS

Representing Primary Keys, Foreign Key and Constraints

The data model has been designed to ensure integrity of data is maintained thought out the process. Below is the table that lists the constraints applied on table sin schema `Cal_Road_Accident`

- Query Used:

```
SELECT
    table_name,
    COLUMN_NAME,
    CONSTRAINT_NAME,
    IFNULL(REFERENCED_COLUMN_NAME, 'NA') AS REFERENCED_COLUMN_NAME,
    IFNULL(REFERENCED_TABLE_NAME, 'NA') AS REFERENCED_TABLE_NAME
FROM
    information_schema.KEY_COLUMN_USAGE
WHERE
    constraint_schema = 'cal_road_accident'
ORDER BY table_name;
```

TABLE_NAME	COLUMN_NAME	CONSTRAINT_NAME	REFERENCED_COLUMN_NAME	REFERENCED_TABLE_NAME
collisions	case_id	PRIMARY	NA	NA
collisions	location_ID	collisions_ibfk_1	collision_location_ID	collisions_location
collisions	county_ID	fk_collision_countyID	County_ID	county
collisions	rc_ID	fk_collision_rc	road_cndtn_id	road_condition
collisions	violation_ID	fk_collision_violation	violation_ID	violation
collisions	vehicle_type_at_fault_id	fk_collision_vt	vt_id	vehicle_type
collisions	WE_ID	fk_collision_we	WE_ID	weather_effect
collisions_location	collision_location_ID	PRIMARY	NA	NA
county	County_Name	County_Name_UNIQUE	NA	NA
county	County_ID	PRIMARY	NA	NA
parties	case_id	PRIMARY	NA	NA
parties	party_number	PRIMARY	NA	NA
parties	case_id	fk_parties_case_id	case_id	collisions
parties	violation_ID	fk_parties_violation	violation_ID	violation
parties	vt_id	fk_parties_vt	vt_id	vehicle_type
road_condition	road_cndtn_id	PRIMARY	NA	NA
road_condition	road_condition	PRIMARY	NA	NA
road_condition	lighting	PRIMARY	NA	NA
vehicle_type	vt_id	PRIMARY	NA	NA
victims	id	PRIMARY	NA	NA
victims	case_id	fk_victim_party_case	case_id	parties
victims	party_number	fk_victim_party_case	party_number	parties
violation	violation_ID	PRIMARY	NA	NA
weather_effect	WE_ID	PRIMARY	NA	NA

CALIFORNIA TRAFFIC COLLISIONS

SQL Code/Queries

Data Definition Language Queries

Attached is the .sql file containing below queries:

Cal_Road_Accident_Database_OnlyQueries.sql

- Create Database
- Create Tables
- Triggers and Procedures

Data Manipulation Language Queries

Attached is the .sql file containing below queries:

Queries of normalization -updating FK and drop columns in Collisions and Parties table

DataWranglingQueries.sql

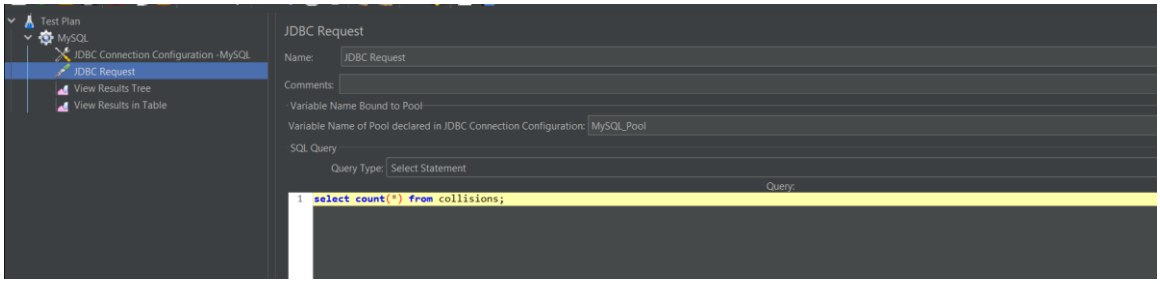
Database Dump – Cal_Road_Accident

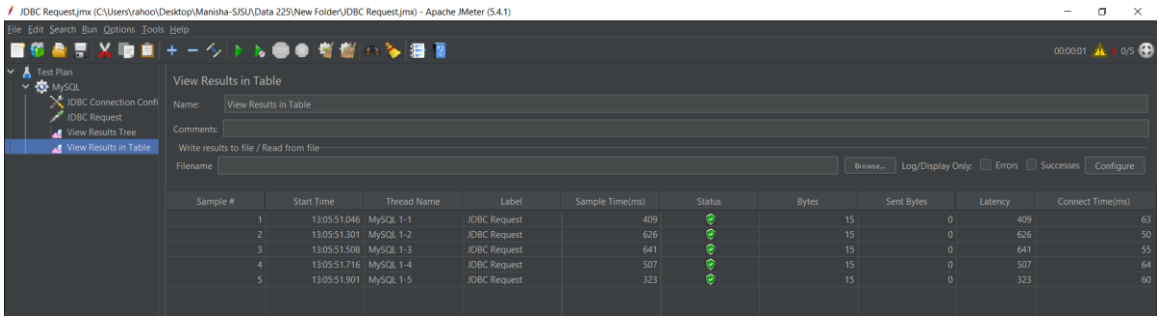
California_Traffic_Collision_Database_Export_WithData.sql

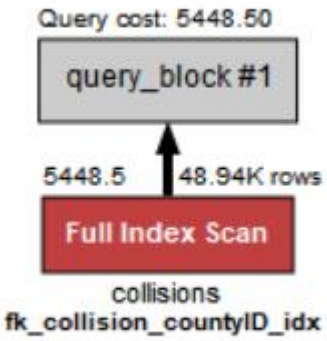
CALIFORNIA TRAFFIC COLLISIONS

SQL Performance Measurement (Select)

Query to select count of records in Collisions

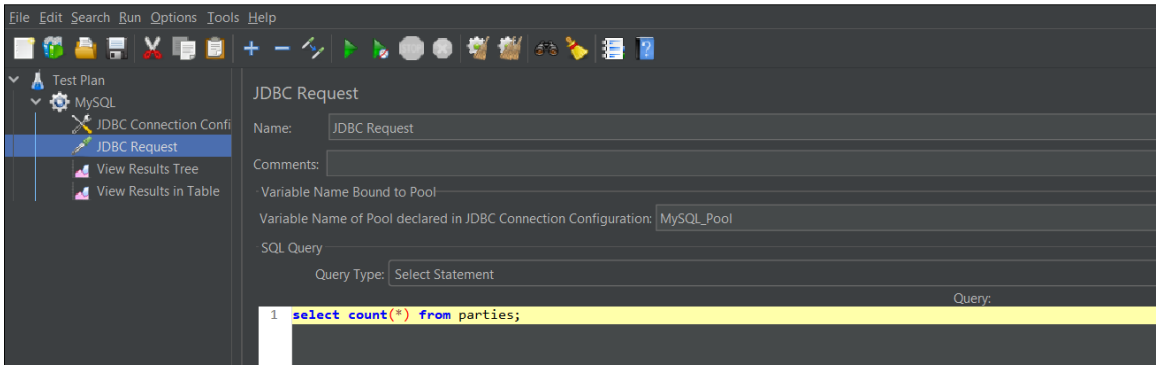


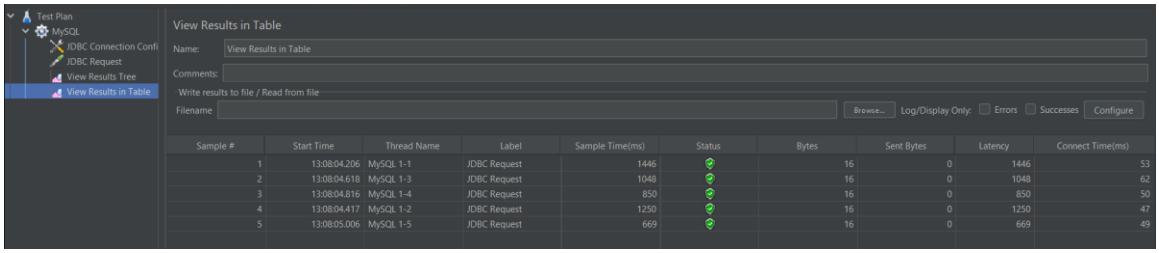


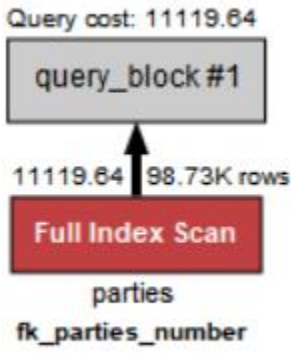


CALIFORNIA TRAFFIC COLLISIONS

Query to select count of records in Parties

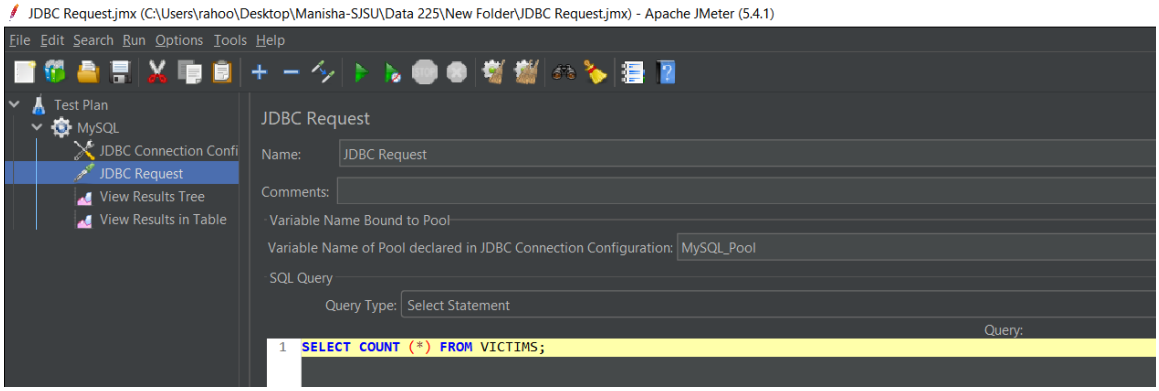


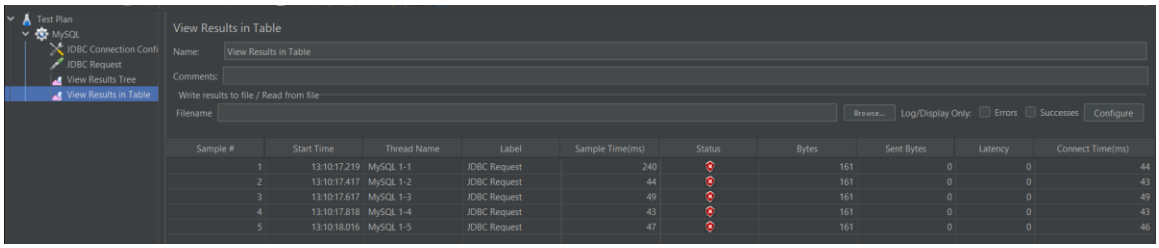


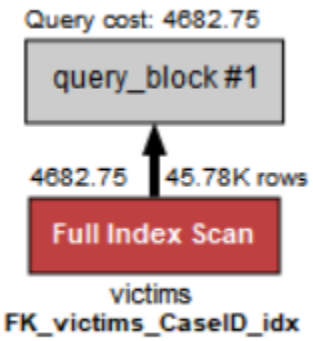


CALIFORNIA TRAFFIC COLLISIONS

Query to select count of records in Victims

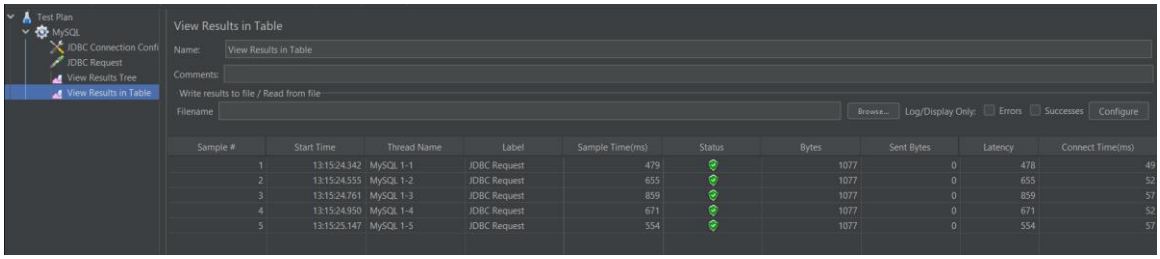


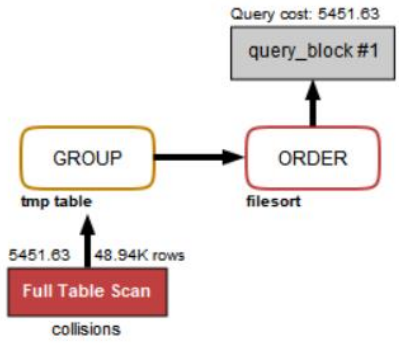




CALIFORNIA TRAFFIC COLLISIONS

Query for Fatal -Non-Fatal Collisions from January 22019 -December 2020





CALIFORNIA TRAFFIC COLLISIONS

Query for visualization of Count of Victims vs County Based on Location Type

JDBC Request

Name: JDBC Request

Comments:

Variable Name Bound to Pool

Variable Name of Pool declared in JDBC Connection Configuration: MySQL_Pool

SQL Query

Query Type: Select Statement

Query:

```
SELECT `county`.`County_Name` AS `County_Name`,
SUBSTRING(`collisions`.`location_type`, 1, 1024) AS `location_type`
FROM `collisions`
INNER JOIN `county` ON (`collisions`.`county_ID` = `county`.`County_ID`)
WHERE ((NOT (`county`.`County_Name` IN ('monterey', 'solano', 'sonoma'))) AND
 (CASE WHEN (((`county`.`County_Name` = 'humboldt') AND (SUBSTRING(`collisions`.`location_type`, 1, 1024) = 'highway'))
 OR ((`county`.`County_Name` = 'san joaquin') AND (SUBSTRING(`collisions`.`location_type`, 1, 1024) IN ('highway', 'ramp')))) THEN 0 ELSE 1 END)
AND (CASE WHEN (SUBSTRING(`collisions`.`location_type`, 1, 1024) IN ('intersection', 'Other')) THEN 0 ELSE 1 END)
AND (NOT (`collisions`.`county_ID` NOT IN (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15))))
GROUP BY 1,  2
```

View Results in Table

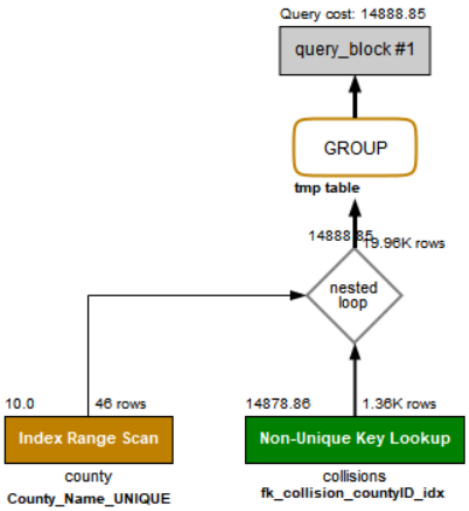
Name: View Results in Table

Comments:

Write results to file / Read from file

Filename: Browse... Log/Display Only: Errors Successes Configure

Sample #	Start Time	Thread Name	Label	Sample Time(ms)	Status	Bytes	Sent Bytes	Latency	Connect Time(ms)
1	13:32:14.561	MySQL 1-1	JDBC Request	2520	✓	651	0	2520	49
2	13:32:14.762	MySQL 1-2	JDBC Request	2319	✓	651	0	2319	52
3	13:32:14.963	MySQL 1-3	JDBC Request	2417	✓	651	0	2417	54
4	13:32:15.366	MySQL 1-5	JDBC Request	2034	✓	651	0	2034	60
5	13:32:15.163	MySQL 1-4	JDBC Request	2237	✓	651	0	2237	54



CALIFORNIA TRAFFIC COLLISIONS

Query for visualization of Weather Condition vs Collision Severity

JDBC Request

Name: JDBC Request

Comments:

Variable Name Bound to Pool

Variable Name of Pool declared in JDBC Connection Configuration: MySQL_Pool

SQL Query

Query Type: Select Statement

Query:

```
SELECT `county`.`County_Name` AS `County_Name`,
  `weather_effect`.`Weather_Condition` AS `Weather_Condition`,
  SUBSTRING(`collisions`.`collision_severity`, 1, 1024) AS `collision_severity`
FROM `collisions`
  INNER JOIN `county` ON (`collisions`.`county_ID` = `county`.`County_ID`)
  INNER JOIN `weather_effect` ON (`collisions`.`WE_ID` = `weather_effect`.`WE_ID`)
WHERE ((NOT (`weather_effect`.`Weather_Condition` IN ('clear', 'NA', 'other', 'snowing', 'wind'))))
AND (CASE WHEN (SUBSTRING(`collisions`.`collision_severity`, 1, 1024) = 'other injury') THEN 0 ELSE 1 END)
AND (CASE WHEN (SUBSTRING(`collisions`.`location_type`, 1, 1024) = 'Other') THEN 0 ELSE 1 END))
GROUP BY 1,2,  3
```

View Results in Table

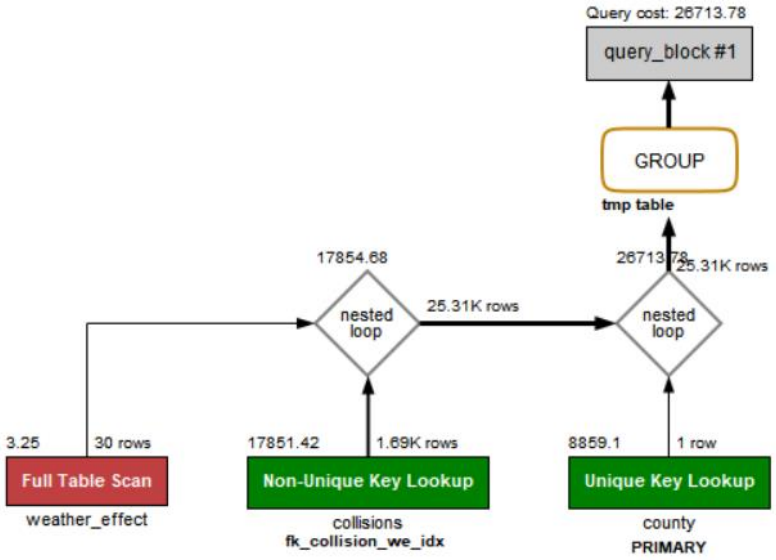
Name: View Results in Table

Comments:

Write results to file / Read from file

Filename: Browse... Log/Display Only: Errors Successes Configure

Sample #	Start Time	Thread Name	Label	Sample Time(ms)	Status	Bytes	Sent Bytes	Latency	Connect Time(ms)
1	13:41:10.989	MySQL 1-1	JDBC Request	1077		6951	0	1076	50
2	13:41:11.196	MySQL 1-2	JDBC Request	1073		6951	0	1072	44
3	13:41:11.402	MySQL 1-3	JDBC Request	1063		6951	0	1062	53
4	13:41:11.791	MySQL 1-5	JDBC Request	800		6951	0	799	50
5	13:41:11.591	MySQL 1-4	JDBC Request	1000		6951	0	999	49



CALIFORNIA TRAFFIC COLLISIONS

Query for visualization of Alcohol vs time and type of collision

JDBC Request

Name: JDBC Request

Comments:

Variable Name Bound to Pool

Variable Name of Pool declared in JDBC Connection Configuration: MySQL_Pool

SQL Query

Query Type: Select Statement

Query:

```
SELECT collisions.collision_time AS "collision_time",
  SUM(collisions.alcohol_involved) AS "sum:alcohol_involved",
  collisions.type_of_collision AS "type_of_collision"
FROM collisions
WHERE (CASE WHEN ((collisions.type_of_collision IN ('other')) OR (collisions.type_of_collision IS NULL)) THEN FALSE ELSE TRUE END) GROUP BY 1,  3
```

View Results in Table

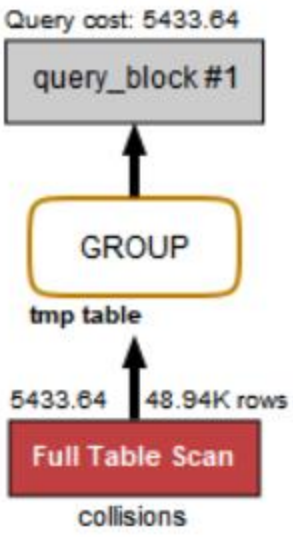
Name: View Results in Table

Comments:

Write results to file / Read from file

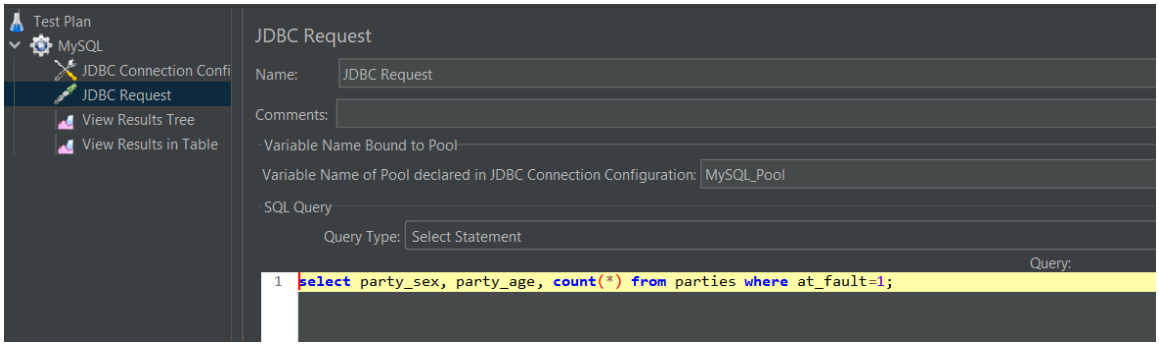
Filename Browse... Log/Display Only: Errors Successes Configure

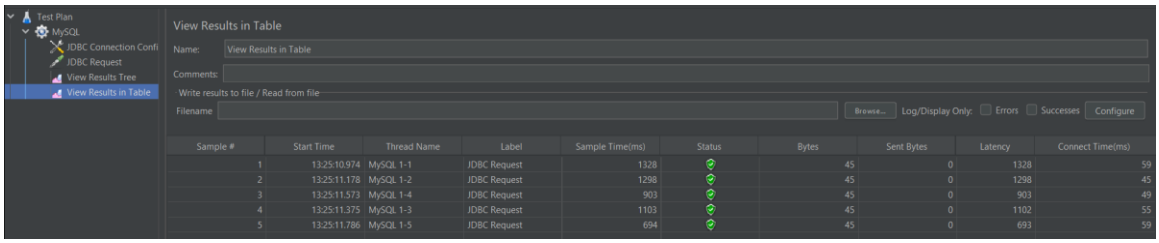
Sample #	Start Time	Thread Name	Label	Sample Time(ms)	Status	Bytes	Sent Bytes	Latency	Connect Time(ms)
1	13:51:21.228	MySQL 1-1	JDBC Request	424		179746	0	398	50
2	13:51:21.642	MySQL 1-3	JDBC Request	567		179746	0	550	66
3	13:51:21.440	MySQL 1-2	JDBC Request	793		179746	0	778	50
4	13:51:21.841	MySQL 1-4	JDBC Request	622		179746	0	612	57
5	13:51:22.029	MySQL 1-5	JDBC Request	616		179746	0	607	54

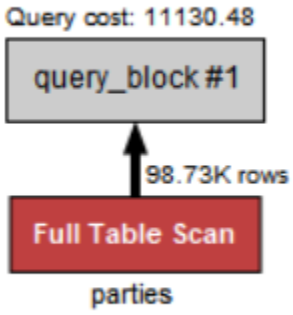


CALIFORNIA TRAFFIC COLLISIONS

Query for visualization of Age vs Gender of Parties at Fault

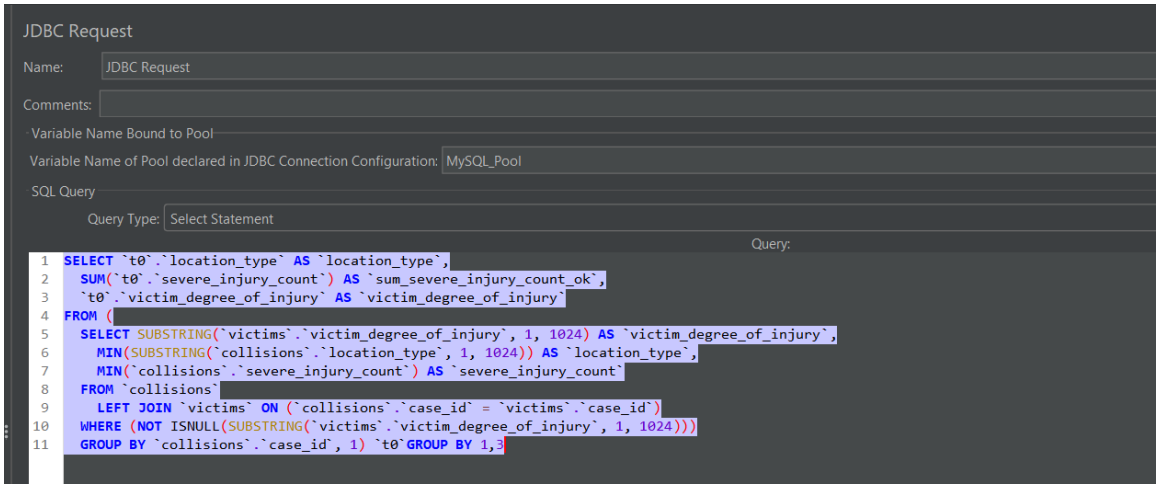


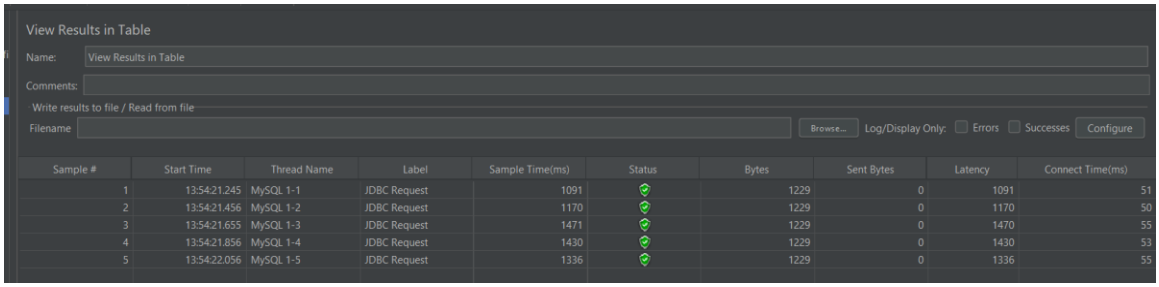


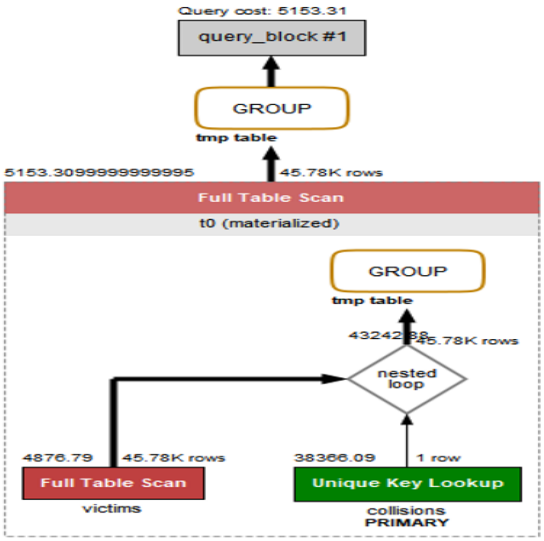


CALIFORNIA TRAFFIC COLLISIONS

Query for Victims Degree of Injury vs Injury Count







CALIFORNIA TRAFFIC COLLISIONS

Queries in .sql file used for Performance Measurement

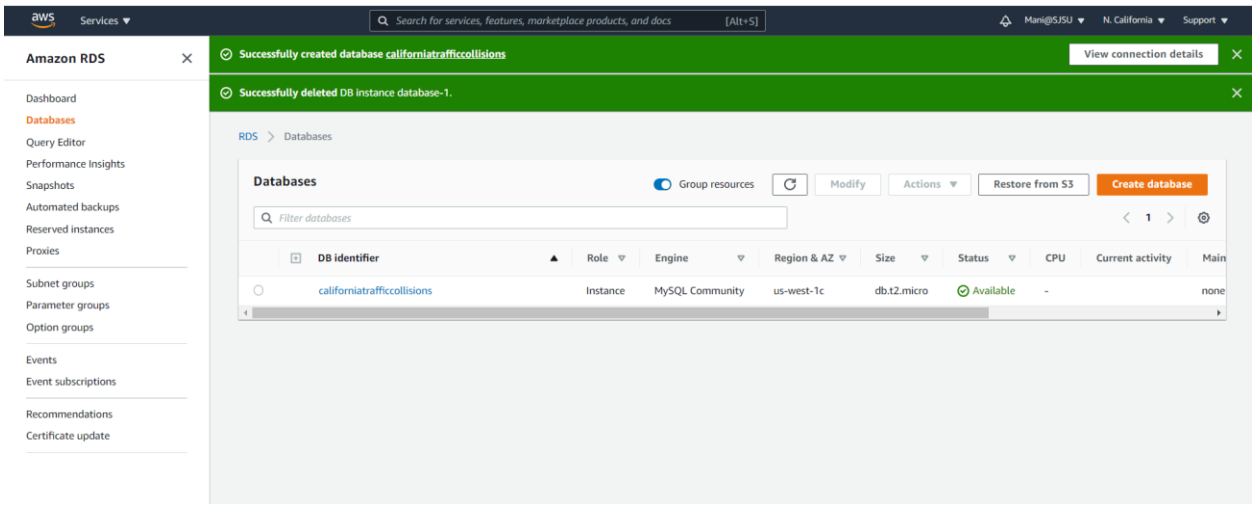
SQl_Performance_Measurement_Queries.sql

CALIFORNIA TRAFFIC COLLISIONS

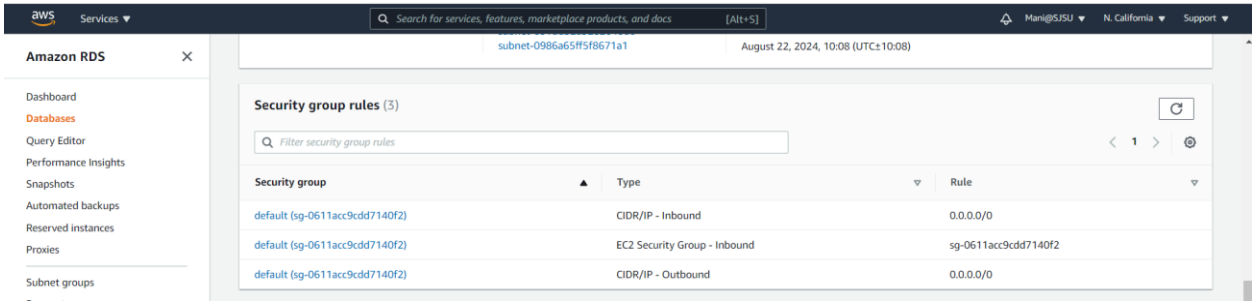
Connectivity to AWS Python

Upload MySQL Project Database into RDS

Step 1: Logged in to AWS and created RDS database

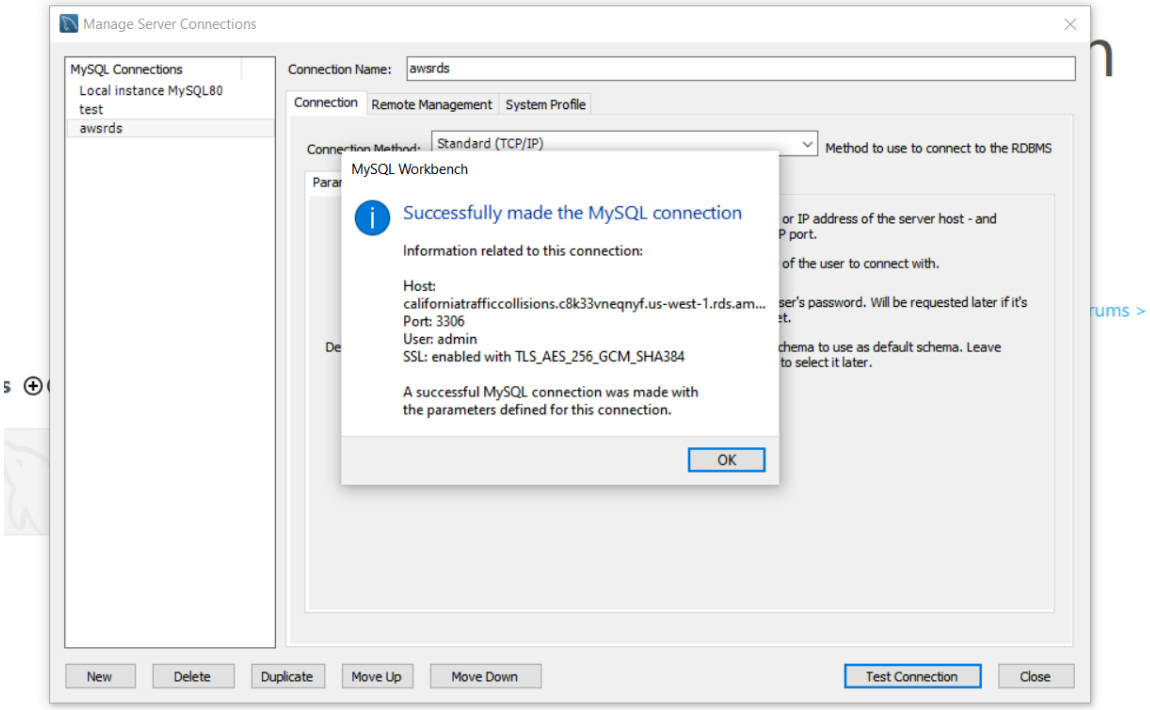


Step 2: Creating security group in EC2

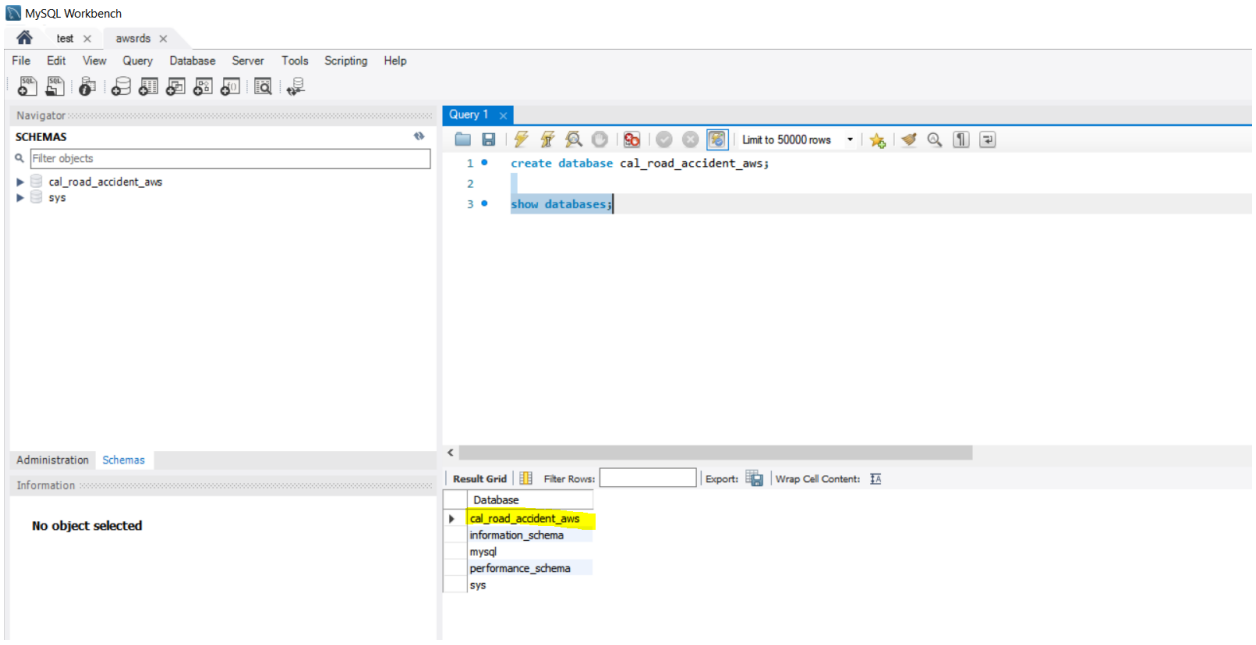


CALIFORNIA TRAFFIC COLLISIONS

Step 3: Connection established in MYSQL workbench

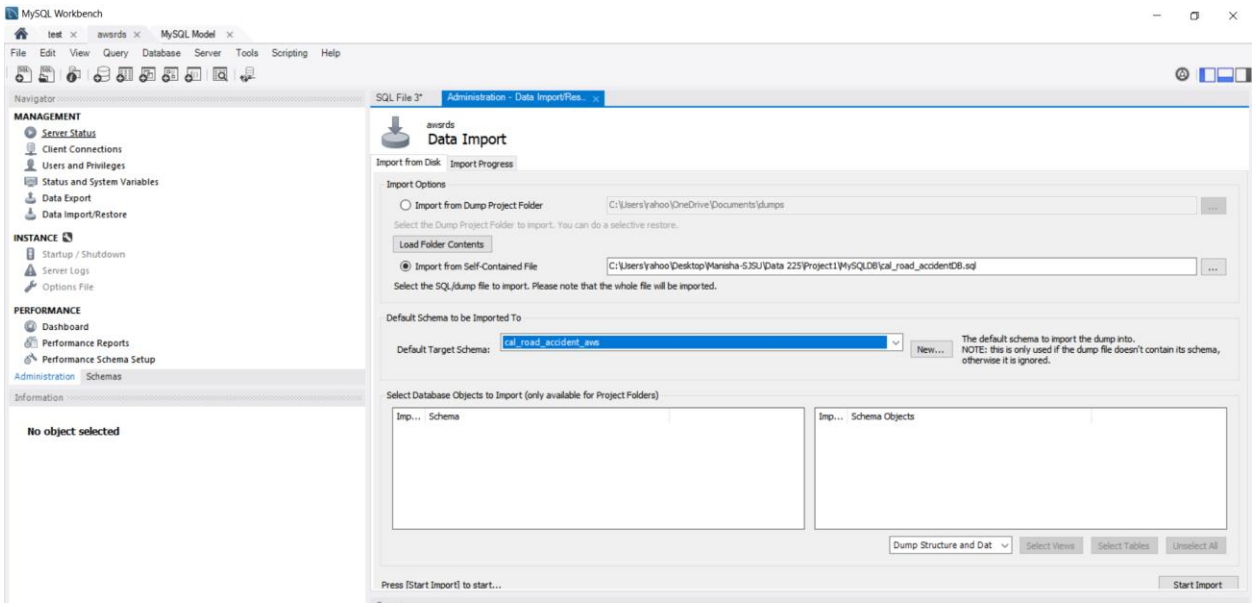


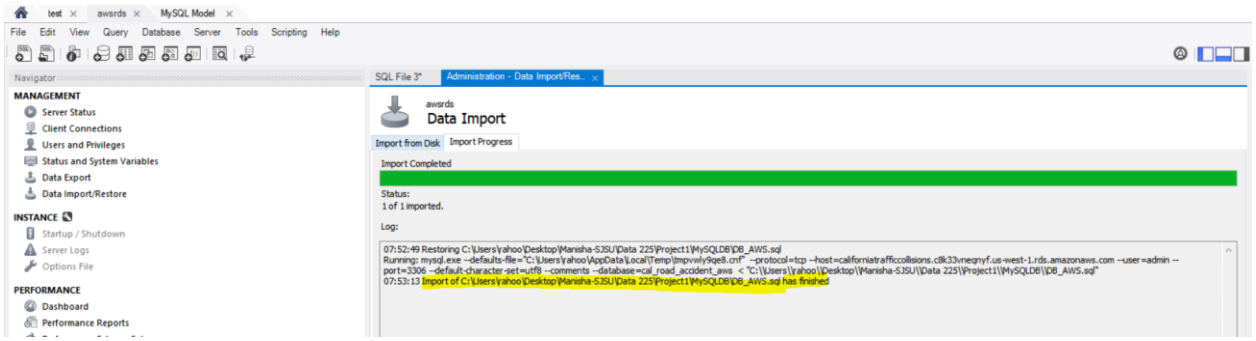
Step 4: Uploading cal_road_accident database from MYSQL to AWS RDS
First created database in MYSQL using AWS database connection



CALIFORNIA TRAFFIC COLLISIONS

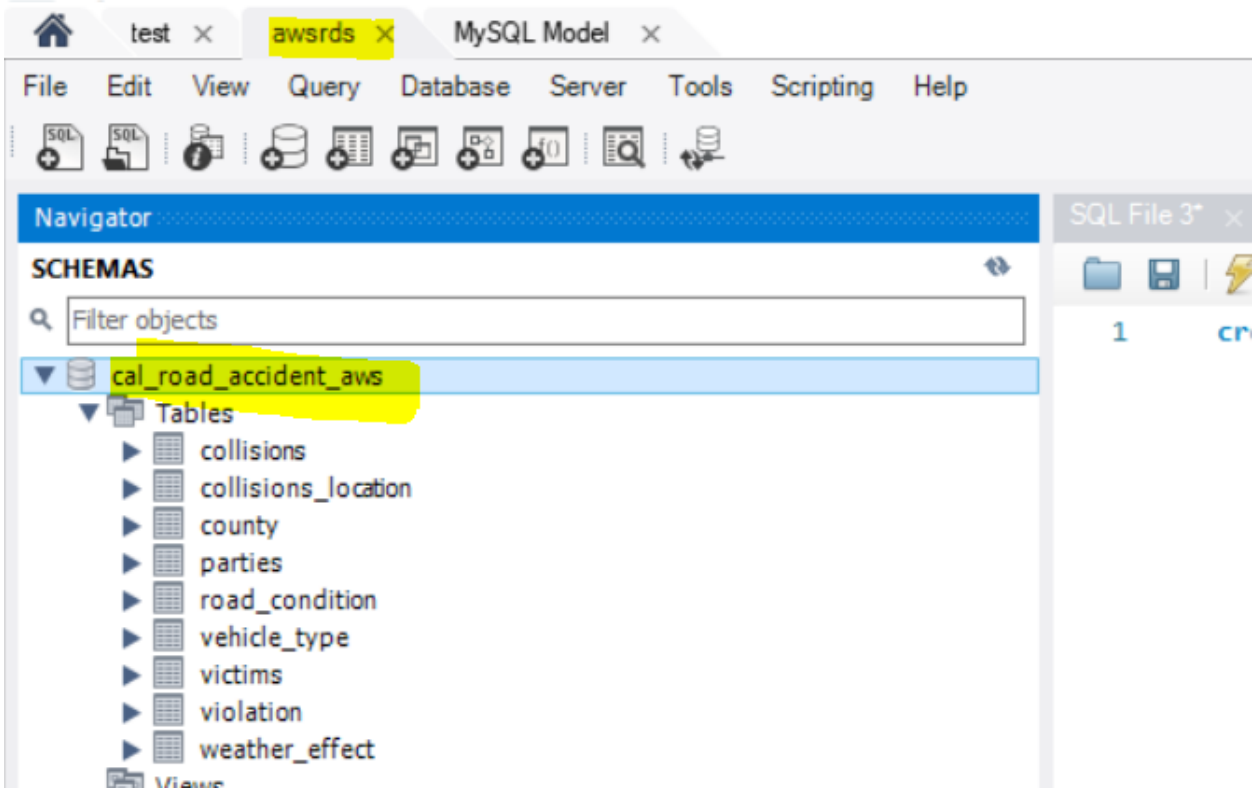
Step 5: Importing Database Cal_Road_Accident database SQL file into Cal_Road_Accident_aws database using connection awsrds





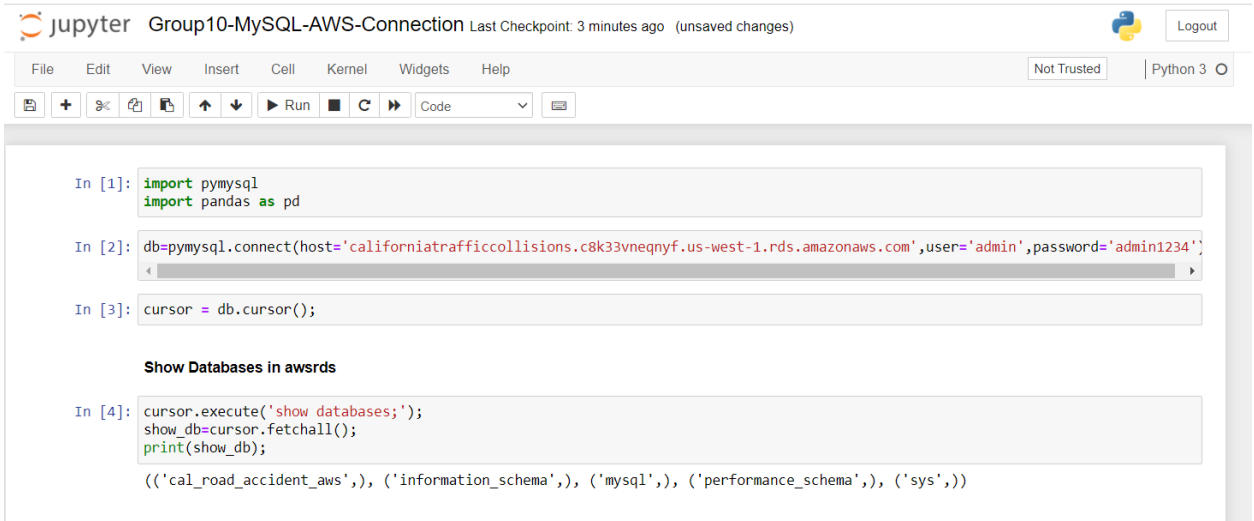
CALIFORNIA TRAFFIC COLLISIONS

Step 6: Database imported successfully



Connecting AWS RDS in python

Step 7: Querying on cal_road_accident_aws database



CALIFORNIA TRAFFIC COLLISIONS

Select query on collision table

In [5]:
```
cursor.execute('select case_id from cal_road_accident_aws.collisions limit 5');

select_collision=cursor.fetchall();
dataframe_collision= pd.DataFrame(select_collision, columns=['case_id']);
print(dataframe_collision);
```

```
    case_id
0   8008500
1   8008532
2   8008550
3  80972854
4  80976438
```

Select query on parties table

In [6]:
```
cursor.execute('select case_id,party_number,party_sobriety,cellphone_in_use,hazardous_materials,cellphone_use_type from cal_road_
select_parties=cursor.fetchall();
dataframe_parties=pd.DataFrame(select_parties, columns=['case_id','party_number','party_sobriety','cellphone_in_use','hazardous_m
print(dataframe_parties);
```

```
   case_id  party_number                    party_sobriety  \
0  0081715             1                    not applicable
1  0081715             2                    not applicable
2  0726202             1                impairment unknown
3  8008483             1  had been drinking, under influence
4  8008483             2                    not applicable

   cellphone_in_use hazardous_materials    cellphone_use_type
0               0.0                None  cellphone not in use
1               0.0                None  cellphone not in use
2               NaN                None                  None
3               0.0                None  cellphone not in use
4               NaN                None                  None
```

Select query on victims

In [7]:
```
cursor.execute('SELECT id, case_id,party_number, victim_role, victim_sex, victim_age FROM cal_road_accident_aws.victims limit 5;'
select_victims=cursor.fetchall();
dataframe_victims=pd.DataFrame(select_victims, columns=['id', 'case_id','party_number', 'victim_role', 'victim_sex', 'victim_age'
print(dataframe_victims);
```

```
        id  case_id  party_number victim_role victim_sex victim_age
0  3078083  8008484             2      driver       male         33
1  3078084  8008484             2   passenger       male       None
2  3078087  8008488             2      driver       male         26
3  3078088  8008488             2   passenger     female         26
4  3078090  8008491             1      driver     female         34
```

Analysis on sobriety of parties consuming alcohol

In [8]:
```
cursor.execute('select c.alcohol_involved ,p.party_sobriety , count(*) from cal_road_accident_aws.collisions c join cal_road_acci
select_output=cursor.fetchall();
dataframe_analysis_result=pd.DataFrame(select_output, columns=['alcohol_involved','party_sobriety','no. of collisions']);
print(dataframe_analysis_result);
```

```
   alcohol_involved                    party_sobriety  no. of collisions
0                 1  had been drinking, under influence               5069
1                 1                not applicable               2311
2                 1           had not been drinking               3314
3                 1                            None                324
4                 1   had been drinking, impairment unknown           1253
5                 1  had been drinking, not under influence            874
6                 1                impairment unknown                225
```

CALIFORNIA TRAFFIC COLLISIONS

MYSQL-AWS-CONNECTION NOTEBOOK

Group10-MySQL-AWS-Connection.ipynb

CALIFORNIA TRAFFIC COLLISIONS

Visualization

Count of Victims vs County Based on Location Type

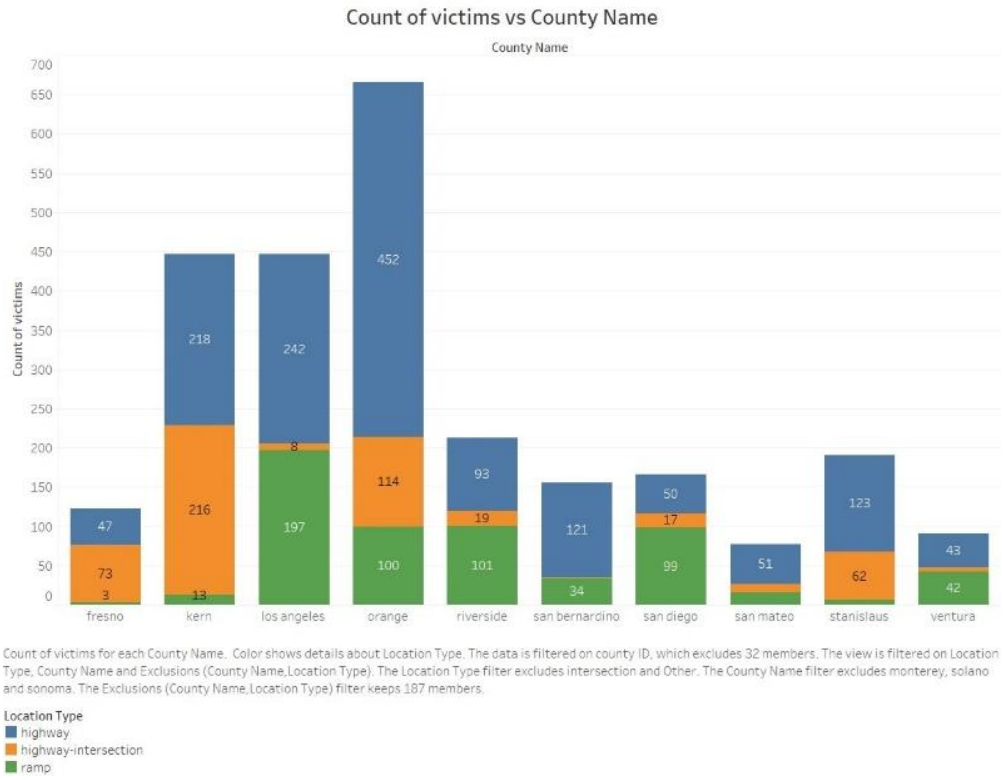


FIGURE 1: GRAPH REPRESENTING COUNT OF VICTIMS VS COUNTY NAME BASED ON LOCATION TYPE

The stacked bar graph gives a pictorial description on the total number of victims who have been in a collision in different counties in the State of California for the years 2019-2020 based on location where collision occurred. We can infer from the graph that most of the accidents occur in large counties such as Los Angeles, Kern and Orange on Highway and Highway Road Intersections.

CALIFORNIA TRAFFIC COLLISIONS

Injured Victims vs Hour of Collision Time

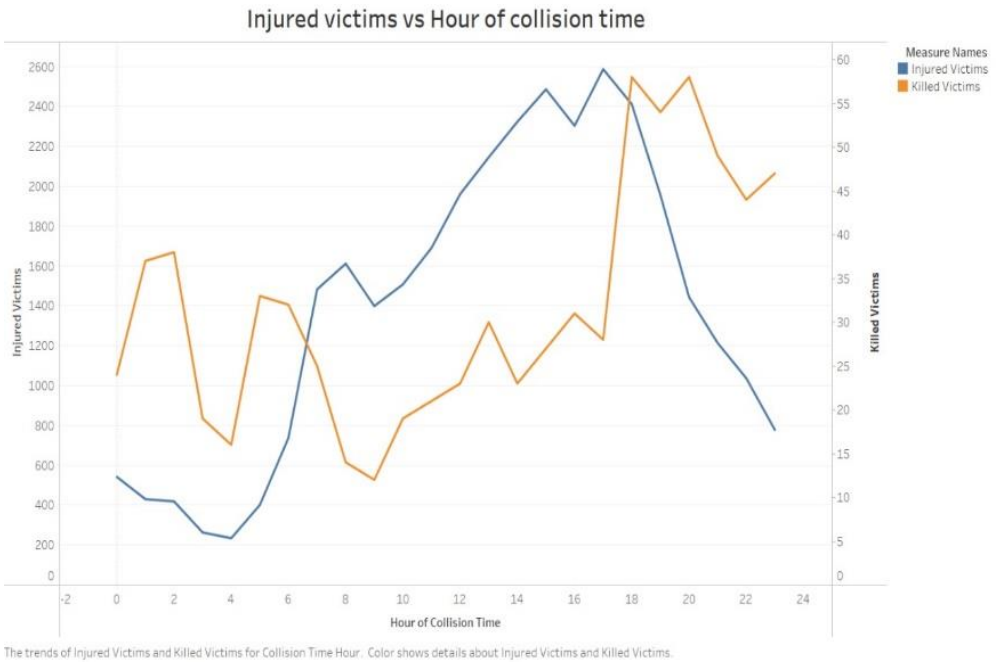


FIGURE 2: GRAPH REPRESENTING THE INJURED VICTIMS VS HOUR OF COLLISION TIME

Above line graph illustrates the correlation between number of injured and killed victims' and time of the day. Units measured are in hours (0-24-hour format)
The number of killed victims is at the peak from evening 6pm till midnight. It can conclude that collisions are more severe during nights compared to daytime.

CALIFORNIA TRAFFIC COLLISIONS

Weather Condition vs Collision Severity

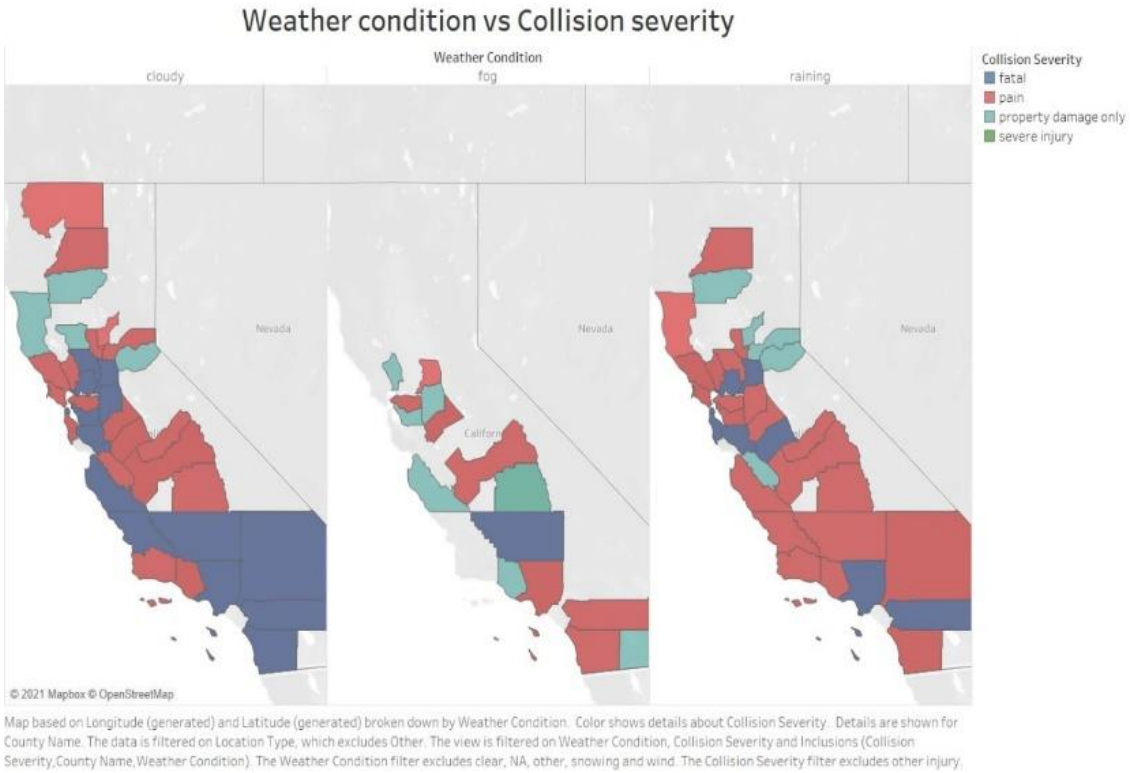


FIGURE 2: GRAPH REPRESENTING WEATHER CONDITION VS COLLISION SEVERITY

Open Street Graph illustrates the collision severity with respect to changes in weather.
The major contributors for weather related accidents are cloudy, fog and rain. Highest fatality rate can be witnessed when it's cloudy. Whereas it's the least when the weather is foggy. Highest pain is witnessed when it's raining and it's the least when the weather is foggy. This is a clear indication that foggy weather is comparatively better than cloudy and rainy weather condition.

Alcohol Influence vs Time and Type of Collision

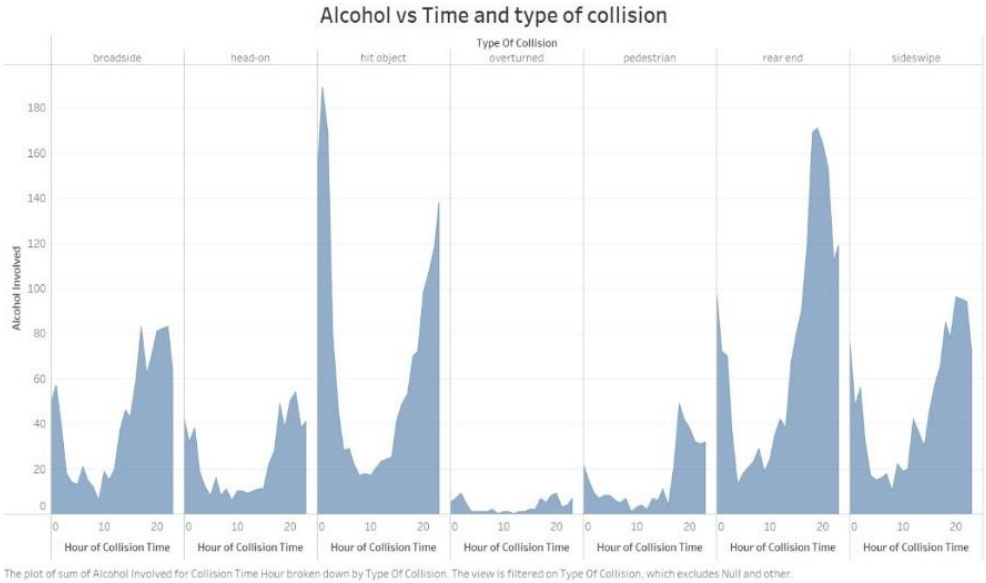


FIGURE 4: GRAPH REPRESENTING ALCOHOL INFLUENCE VS TIME AND TYPE OF COLLISION

The above graph gives a pictorial description on the type of collision and the time of occurrence when the person is under the influence of alcohol. We can infer from the graph that the most common type of collision is 'Hit Object' collision and the most common time for an accident to occur while the person is under the influence of alcohol is usually during the nighttime.

Diversity of Party Gender Involved in Collision

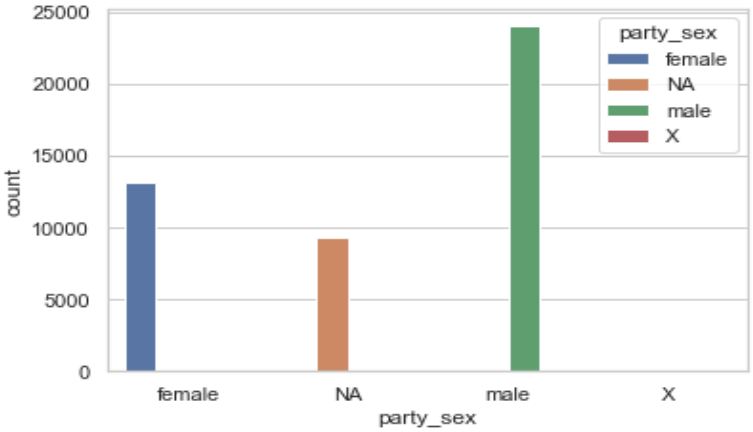


Figure 5: Graph showing gender vs count of the parties

The above count plot graph [Python visualization] gives us insights on how many people were involved in the collision and their gender either Male or Female or Transgender. NA is the unspecified data in the Dataset which means their gender was not specified. It is clear that Males have a larger count with a number of almost 25000 and females have a count of 12500

Age vs Gender of Parties at Fault

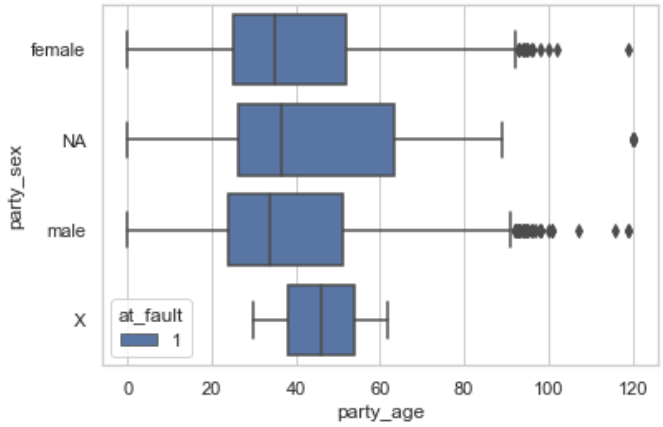


FIGURE 6: GRAPH SHOWING THE AGE GROUP VS GENDER OF PEOPLE CAUSING A COLLISION

The above boxplot gives us insights on the party age, party sex based on Female or Male or Transgender and NA. From the graph we can say that the mean age for male victim at fault for causing an accident is 35 and the mean age for the female victim causing an accident is 37.
From the graph we can say that the age group that are at fault is mainly between early 20's and 40's for both the genders. NA is the not specified gender which tells us that there are many people between the age group of 30 and early 60's who are at fault for the accident and their gender has not been specified while collecting data. We can infer from the above two graphs that the total number of men at fault are the highest when compared to the female at fault.

CALIFORNIA TRAFFIC COLLISIONS

Number of Fatal Collisions [January 2019 – December 2020]

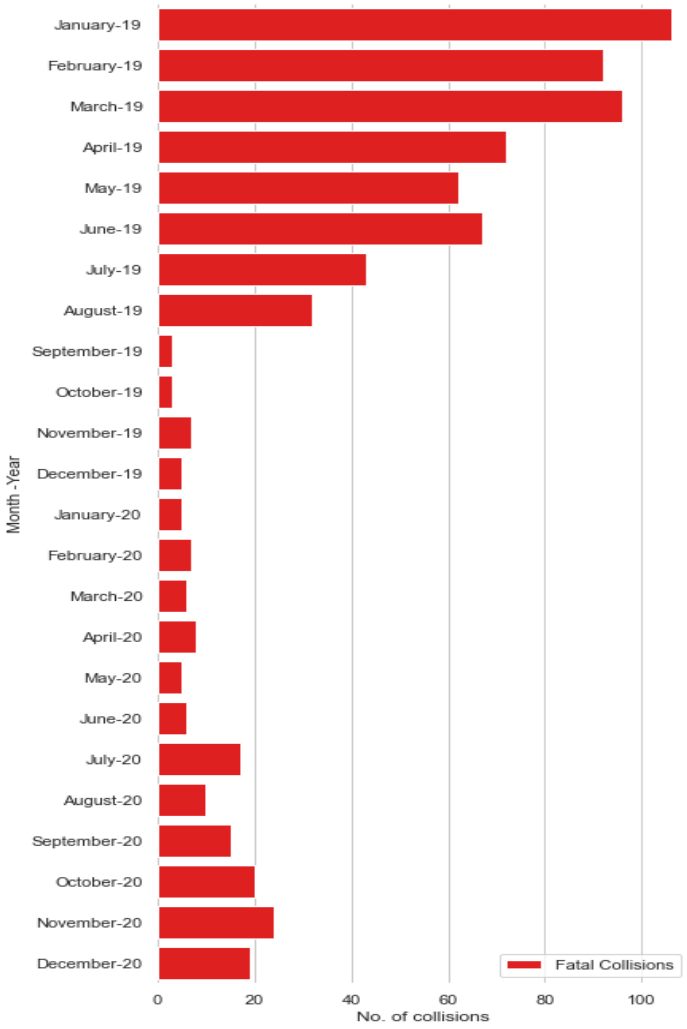


FIGURE 7: GRAPH REPRESENTS NUMBER OF FATAL COLLISIONS FROM JANUARY 2019 – DECEMBER 2020

The above bar plot provides insights on how many fatal collisions occurred from January 2019 to December 2020. The number of fatal collisions pre-covid was more as there was more movement amongst the people of California.

January-2019 has the most with a count of almost greater than 100 fatal collisions. But as the pandemic started to get worse the movement within the state reduced so did the total number of collisions in the State as well. The months from September 2019 till July 2020 (considering there was s stay at home order imposed in April 2020), from this we can infer that the movement of people reduced and gradually so did the collisions. But soon as they started to ease the lockdown rules, the rate of collisions increased as people started to travel.

CALIFORNIA TRAFFIC COLLISIONS

Number of Non-Fatal Collisions [January 2019 – December 2020]

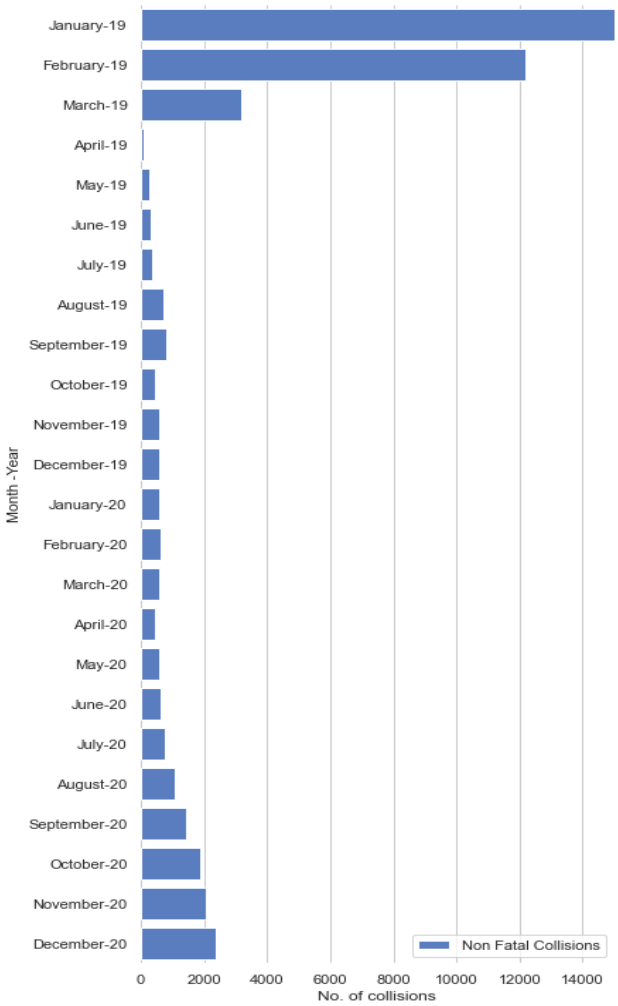


FIGURE 8: GRAPH REPRESENTS NUMBER OF NON-FATAL COLLISIONS FROM JANUARY 2019 – DECEMBER 2020

The above bar plot provides insights on how many non-fatal collisions takes place from January 2019 to December 2020. The number of non-fatal collisions pre-covid was more as there was more movement amongst the people of California.

January-2019 has the most with a count of almost greater than 15000 non-fatal collisions. We can clearly see from the graph above that the number of non-fatal collisions has reduced drastically as the pandemic got worse. The months from April 2019 till July 2020 had the least number of non-fatal accidents in the State of California. The number of non-fatal collisions started increasing gradually once people started to make movement across California.

CALIFORNIA TRAFFIC COLLISIONS

Visualization Code Documents

Tableau Dashboard

LINK

Visualization - Python Jupyter Notebook

CaliforniaTrafficCollisionPythonVisualization.ipynb

Visualization -Tableau Workbook

Tableau_Wookbook_Folder

Queries in .sql file used for Visualization

Python_Tableau_Visualization_Queries.sql

CALIFORNIA TRAFFIC COLLISIONS

Conclusion

Despite small number of vehicles operating in the year 2020, the level of crash accident recorded in California, made the state one of the top in the United States for traffic collisions. Through this data analysis a variety of insights concerning the location, time, weather, and points-of-interest of an accident are found. The analysis helps us understand the best month, day, and hour of the day to commute. Also, it can help us to predict what are the accident prone areas in the state such as Los Angeles, Kern and Orange with Highway and Highway road Intersections. It also shows that the highest death is happening between the 20- 50 age group and most of the accidents have occurred during a cloudy weather. The top 3 violations causing maximum collisions were: not following Traffic guidelines (11%), Unsafe speed (22% approx.) and Improper turns (17%).