ANALYSIS OF ROAD ACCIDENTS IN CALIFORNIA

Hamsalakshmi Ramachandran, Soumya Challuru Sreenivas, Prathyusha Pingili, and Sugandha Chauhan

Master of Science in Data Analytics

San Jose State University

Spring 2024

Abstract

Millions of people are injured and lose their lives in traffic accidents every year, which raises serious health and safety concerns worldwide. The frequency and severity of these incidents in California pose difficult problems for the public, law enforcement, transportation authorities, and politicians. Understanding the mechanics of accidents is critical for designing effective methods to reduce their impact and increase road safety. This research underlines the importance of analyzing road accidents in California, highlighting the necessity for extensive data gathering to discover patterns, risk factors, and feasible actions to minimize accident severity. This study investigates how weather, time of day, and driving behavior might be used to improve road safety in California. The goal is to prevent accidents while improving the quality of life for inhabitants and visitors, resulting in a safer environment.

1 INTRODUCTION

In California, traffic accidents represent a significant and complex challenge in the state's transportation environment. California has an extensive network of highways, interstates, and country roads, and is home to a wide variety of accidents, from minor fender benders to catastrophic collisions. Factors such as population density, diverse demographics, changing road conditions, and ever-evolving traffic patterns compound the problem. Understanding California crash dynamics is critical to implementing effective safety measures, reducing injuries and fatalities, and improving overall traffic safety.

The Statewide Integrated Traffic Records System (SWITRS) in California is a complete database for traffic collision information across the state. SWITRS collects data from multiple law enforcement agencies to offer a range of information surrounding road accidents in California. It provides informed decision-making for policymakers, law enforcement agencies, researchers, and the public through crash reports and statistical analyses. SWITRS, with its vast collection of data covering many years, is a crucial resource for studying, and tackling traffic safety issues on California's varied road systems.

California has reported over 250,000 total traffic collisions per year, involving different types of vehicles and road users. A majority of this is speeding-related. The state has seen thousands of fatalities resulting from road accidents each year, with the number changing but usually exceeding 3000 deaths annually. In 2023, from January to June alone, showed 2,061 deaths.

2 SIGNIFICANCE TO THE REAL WORLD

Accidents in California present various challenges concerning public safety, facilities management, and policy development. These accidents vary in nature and severity, surrounding collisions involving vehicles, pedestrians, cyclists, and other road users. Factors such as traffic, weather, driver ac-

tions, and infrastructure deficiencies all play a crucial role in contributing to accidents.

Through the Statewide Integrated Traffic Records System (SWITRS), the California Highway Patrol (CHP) keeps extensive records of traffic collisions, offering useful information for analysis and prevention. The frequency, spread, and causes of accidents throughout the state and road types can be learned through analysis of this data. Hotspots for accidents frequently include highways, urban crossroads, and regions with significant pedestrian traffic.

Campaigns for public awareness that promote cautious driving, eliminate distractions, and boost the safety of bicyclists and pedestrians are essential to minimizing accidents. Law enforcement organizations play a major role in enforcing traffic regulations, preventing careless driving, and minimizing the prevalence of driving while inebriated and unfocused.

Furthermore, the goal of continuing investment in transportation infrastructure is to mitigate accident risks and enhance overall road safety. Examples of these investments include redesigning roads, modernizing intersections, and putting cutting-edge safety systems into place.

In general, tackling the problem of accidents in California demands a committed and data-driven strategy that makes use of accident data insights to determine focused actions and laws aimed to make roads safer for all users.

3 LITERATURE SURVEY

There is much helpful research in data analysis, accident prediction, and prevention. One of the journal reviews from "The New York Times" reveals a significant focus on leveraging various datasets and analytical tools to delve into accident patterns, contributing factors, and mitigation strategies. Researchers have explored the utilization of diverse datasets and real-time national databases to analyze accident rates, severity, and locations. One notable statement in most research is that analyzing accidents is vital in training self-driving vehicles. By studying accident data, we can develop algorithms and models that enable autonomous vehicles to navigate California's roads better. The effectiveness of past safety interventions and the potential of emerging technologies like autonomous vehicles have been subjects of investigation, paving the way for a safer transportation future. This research reflects a concerted effort to enhance traffic safety in California, leveraging accident analysis to train and improve the capabilities of self-driving vehicles and striving to reduce accidents and their impact on residents.

4 DEVELOPMENT METHODOLOGY

- Pair programming, an agile approach, was used in the creation of this project. It is a technique for developing software.
- We used Jira Kanban to capture the progress of our tasks. Link: JIRA KANBAN
- The project's coding and testing phase's time management has improved because of pair programming.
- The team was also able to collaborate effectively. We can conclude that, thus far, integrating a pair programming approach alongside the development has had a very beneficial impact on the project output.
- Zoom meetings were set up to discuss the project. Google Docs and Google Sheets were used to collaborate to adhere to pair programming. All team members had their chance to be observers and drivers.

5 DATABASE STRUCTURE

The database consists of 3 tables namely accidents, parties, and victims. Accidents is the main table. The other tables are joined to this.

5.1 accidents(12 columns):

- **case_id**: unique identifier of the accident case
- **weather**: weather condition at the time of the accident (clear, raining, snowing, fog, wind, other)

- **collision_severity**: worst injury suffered by any victim in the accident (fatal, pain, property damage only, severe injury, other)
- **killed_victims**: number of killed victims in each accident
- **injured_victims**: number of injured victims in each accident
- **party_count**: number of parties involved in the accident (a vehicle counts as one party regardless of the number of occupants)
- **pcf_violation_category**: a value computed from the law section that was given as the primary cause of the accident
- **road_surface**: roadway surface condition at the time of the accident (dry, wet, slippery, snowy)
- **road_condition**: roadway condition at the time of the accident (like flood, construction, holes etc)
- **lighting**: lighting conditions at the accident location and the time of the accident (dark with street lights, dark with no street lights, dark with street lights not functioning, daylight, dusk, or down)
- **collision_date**: the date when the accident occurred
- **collision_time**: the time when the accident occurred(24 hour time)

5.2 victims(8 columns):

- **id**: unique identifier of the victim case
- **case_id**: unique identifier of the accident case
- **party_number**: the unique identifier of the party in the accident that this victim belongs to
- **victim_role**: role of the victim (1 - Driver, 2 - Passenger (includes non-operator on a bicycle or any victim on/in a parked vehicle or multiple victims on/in a non-motor vehicle), 3 - Pedestrian, 4 - Bicyclist, 5 - Other (single victim on/in the non-motor vehicle), 6 - Non-Injured Party
- **victim_sex**: gender of the victim (male, female, non-binary)
- **victim_age**: age of the victim at the time of the accident
- **victim_degree_of_injury**: severity of the injury to the victim (1 - Killed, 2 - Severe Injury, 3 - Other Visible Injury, 4 - Complaint of Pain, 5 - Suspected Serious Injury, 6 - Suspected Minor Injury, 7 - Possible Injury, 0 - No Injury)
- **victim_seating_position**: seating position of the victim (1 - Driver, 2 thru 6 - Passengers, 7 - Station Wagon Rear, 8 - Rear Occupant of Truck or Van, 9 - Position Unknown, 0 - Other Occupants, A thru Z - Bus Occupants)

5.3 parties(11 columns):

- **id**: unique identifier of the victim case
- **case_id**: unique identifier of the accident case
- **party_number**: number that together with the case_id uniquely identifies a party in a accident
- **party_type**: involved party type (driver, pedestrian, bicyclist, parked vehicle, operator, other)
- **party_sex**: gender of the party (male, female, non-binary)
- **party_age**: age of the party at the time of the accident
- **cellphone_use**: usage of cellphone at the time of accident(1 - cell phone handheld in use, 2 - cell phone hands-free in use, 3 - cell phone not in use, 4 - cell phone use unknown)
- **party_number_killed**: number of killed victims in the party
- **party_number_injured**: number of injured victims in the party
- **movement_preceding_collision**: the action of the vehicle before the accident and before evasive action(like stopped, u-turn, merging etc)

- **party_vehicle**: company of the vehicle involved in the accident

6 ENTITY RELATIONSHIP DIAGRAM

The relationships between entity sets in a database are shown in an entity-relationship diagram (ERD). An individual data unit is an entity in this sense. An "entity package" is a collection of similar entities. These entities' qualities are specified by their attributes. We can see from the ERD that the collisions table is relationships with parties and victims tables along with a foreign key for the same.

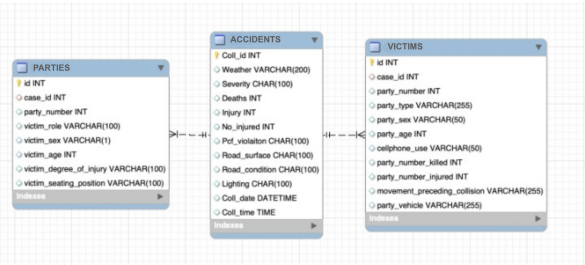

Figure 1: Entity Relationship Diagram

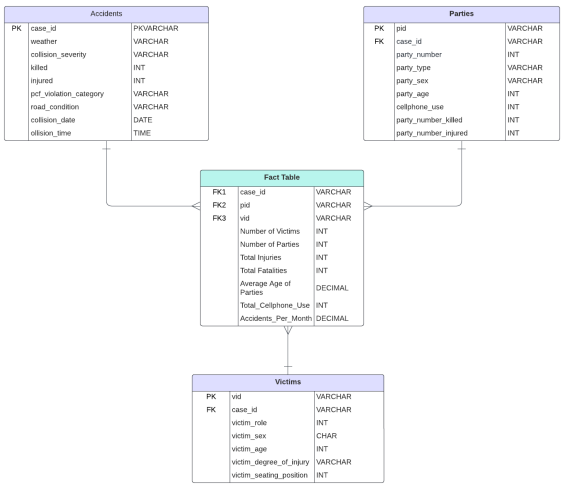

Figure 2: Star Schema for Data Warehouse

7 DESIGN STEPS

	Task Performed	Tool Used
1	Choosing a topic & its scope	Team members brainstormed ideas in person and on Zoom meetings and decided the topic
2	Proposal	Google Docs, Zoom meeting
3	Dataset	California Highway Patrol website, Kaggle
4	Data model	Mysql Workbench, Lucid Chart
5	Schema	PostgreSQL
6	Data storage	CSV, S3 Bucket
7	Data cleaning & ETL	Excel, Amazon Web Services (AWS Glue, Crawler), Python Pandas
8	Loading data into tables	pgAdmin, AWS Glue, MongoDB
9	Visualization	MongoDB Charts, Redshift Charts
11	Data Analysis	Postgre, Amazon Redshift data warehouse, MongoDB
12	Version Control	Github repository
13	Slides	Google Slides, Canva, Microsoft Powerpoint
14	Report	Google Docs, LaTex, Google Sheets

8 TECHNICAL DIFFICULTIES

- In Redshift, the default datatype for time was timestamp but the requirement was to have time without timezone. So we had to default setting of the Redshift datatype.
- While making connections from Glue to Redshift, we faced issues with endpoints, when making connections from S3 to AWS Glue to Redshift. We reconfigured the entire Glue pipeline from scratch.
- Learning to use AWS was time consuming as its terminologies were difficult to understand.

9 DATA VISUALIZATION

Data visualization is the graphical representation of data or information.Visualization can be in the form of charts, maps or graphs. Highly complex

qualitative and quantitative data can be effectively represented visually thus making it easier to process and analyze. In our project, we have used Redshift charts and MongoDB charts for visualization.

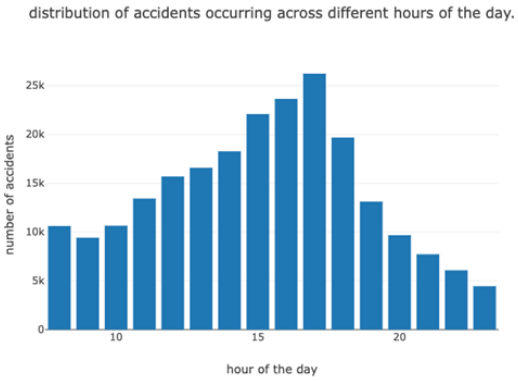

Figure 3: Age distribution of drivers involved in fatal accidents

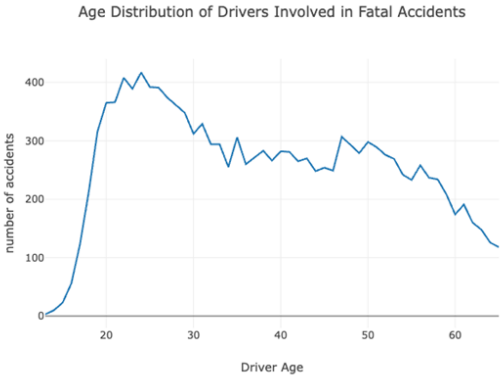

Figure 4: Distribution of accidents occurring across different hours of the day

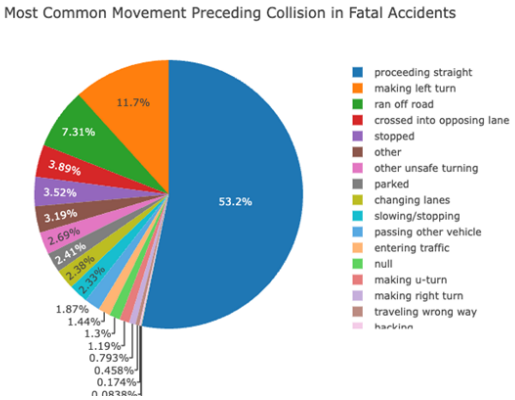

Figure 5: Most common movement preceding collision in fatal accidents

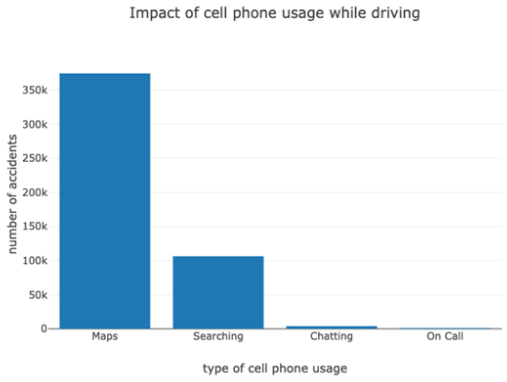

Figure 6: Impact of cell phone usage while driving

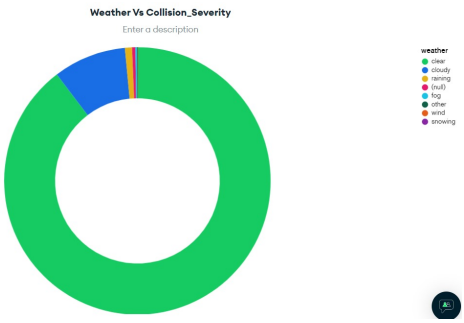

Figure 7: Weather vs Collision Severity

10 ANALYSIS

- It appears clear weather is predominant, suggesting that factors other than weather play a significant role in collision severity.
- Peaking in the mid to late afternoon especially at 5-6pm, there is a higher risk of accidents during these hours, which could correlate with office to home return traffic.
- Over half of the fatal accidents occurred while vehicles were proceeding straight. The next significant movements are making a left turn and running off the road, highlighting potential areas for targeted safety interventions.
- Impact of cell phone usage like using maps (on cell phone) has been associated with a substantial number of accidents, including other usages like searching or chatting.
- This emphasizes the importance of hands-free operation and the potential dangers of distractions while navigating.

11 AWS IMPLEMENTATION

In AWS, we created a VPC endpoint first, then created Redshift clusters and S3 bucket. We used Glue to integrate a local host PostgreSQL to AWS RDS. Then we used Crawler to load the data of all three tables from the relational database into S3 bucket and then into Redshift.

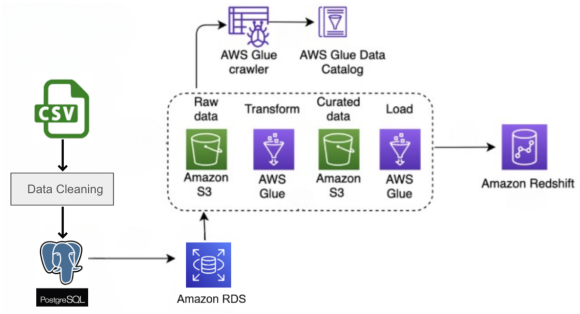

Figure 8: AWS Redshift Implementation

12 LESSONS LEARNED

- We learned how to use AWS Solution Stack to build end-to-end Data Engineering/ Data Warehousing Solutions.
- We uploaded the data to the PostgreSQL database from csv files. Before uploading data, we made sure to clean the files according to the PostgreSQL datatype format.
- Data cleaning with an ETL tool like AWS Glue and Excel is much easier than python pandas.

13 VERSION CONTROL

We uploaded our work in github with public viewing. Anyone can check out our project. Link: https://github.com/hamsaram14/DATA225-ANALYSIS-OF-ROAD-ACCIDENTS-IN-CALIFORNIA/tree/main

14 CONCLUSION

In conclusion, the analysis of road accidents data in California reveals challenges that require extensive solutions. The data emphasizes the urgency of addressing key risk factors such as distracted driving behavior. By making use of evidence-based strategies, and raising awareness for responsible driving, we can strive towards a future where road accidents are minimized, and the safety of all road users is safeguarded.

15 REFERENCES

- https://www.chp.ca.gov/programs-services/services-information/switrs-internet-statewide-integrated-traffic-records-system
- https://www.kaggle.com/datasets/alexgude/california-traffic-collision-data-from-switrs/data
- https://www.ksbw.com/article/several-crashes-california-two-weeks-opening-san-benito-county-turbo-roundabout/60034442
- Our blog with pitch video: https://crunchthatdata.blogspot.com/

16 RUBRICS

Criteria	Points	Comments
Presentation Skills Includes time management	5 pts	Evaluation by professor
Code Walkthrough	3 pts	Code is included in github, walkthrough will be done during presentation
Discussion / Q&A	4 pts	During presentation
Demo	3 pts	During presentation
Version Control Use of Git / GitHub or equivalent; must be publicly accessible	3 pts	https://github.com/hamsaram14/DATA225-ANALYSIS-OF-ROAD-ACCIDENTS-IN-CALIFORNIA/tree/main
Significance to the real world	5 pts	Included in the report
Lessons learned Included in the report and presentation? How substantial and unique are they?	5 pts	Included in the report
Innovation	5 pts	We have conducted an analysis of driver behavior and pre-collision movements, a perspective not explored previously in other analyses.
Teamwork	5 pts	Practiced pair-programming, agile, JIRA
Technical difficulty	4 pts	Included in the report
Practiced pair programming?	2 pts	Yes, Team Viewer and Zoom meets
Practiced agile / scrum (1-week sprints)? Submit evidence on Canvas - meeting minutes, other artifacts	3 pts	JIRA Kanban https://sugandhachauhan1995.atlassian.net/jira/software/projects/DP/boards/1?assignee=unassigned
Used Grammarly / other tools for language? Grammarly free version is sufficient; can use other tools as well. Submit report screenshot on Canvas.	2 pts	Used Grammarly for plagiarism check and grammatical errors, screenshot added in github
Slides	5 pts	During presentation & added in github
Report	7 pts	Relevant topics are covered in the report,
Format, completeness, language, plagiarism, whether turnitIn could process it (no unnecessary screenshots), etc		plagiarism check was done
Used unique tools E.g.: LaTeX for writing report (submit .tex that is not generated from another format such as .docx; generating from .lyx and similar LaTeX editor outputs is fine. Also checkout https://www.overleaf.com/LinksLinksLinks Links to an external site. Unique features of Prezi or powerpoint, etc	5 pts	Used Overleaf for LaTex for writing report(screenshot in github) , used Google Slides to make the presentation slides, used Canva for slides' designs
Performed substantial analysis using database techniques Project must include an analytics component	3 pts	Used Redshift charts for visualization, Postgre DB, used pandas for data cleaning
Used a new database or data warehouse tool not covered in the HW or class	3 pts	Used Amazon Redshift and PostgreSQL
Used appropriate data modeling techniques	5 pts	Used MySQL Workbench to generate ER diagram and Draw.IO for Star Schema
Used ETL tool	1 pts	Used AWS Glue
Demonstrated how Analytics support business decisions	3 pts	Included in report
Used RDBMS Idea is to exercise as many topics from the course as possible	1 pts	Used PostgreSQL with pgAdmin in relational database
Used Datawarehouse Idea is to exercise as many topics from the course as possible	1 pts	Used Amazon Redshift Data Warehouse
Includes DB Connectivity / API calls Possibly using Python	1 pts	Connected MongoDB with python using API
Used NOSQL	1 pts	Used MongoDB NoSQL database