US Road Safety: Analyzing Car Accidents in LA, California for Insights

May 5, 2024

Rachel Chong

INTRODUCTION AND OVERVIEW

- **Problem Statement:** Los Angeles faces challenges with high traffic volume and car accidents, impacting safety and traffic flow. This project leverages data science and machine learning to predict the severity levels of car accidents, aiding in the identification of accident hotspots and contributing factors.
- **Goal:** By leveraging data science and machine learning, we aim to substantially reduce road accidents in Los Angeles, contributing to a safer and more economically efficient society.

$340 Billion in 2019[3]

That's a lot of money for traffic accidents

~19,937 crashes[2]

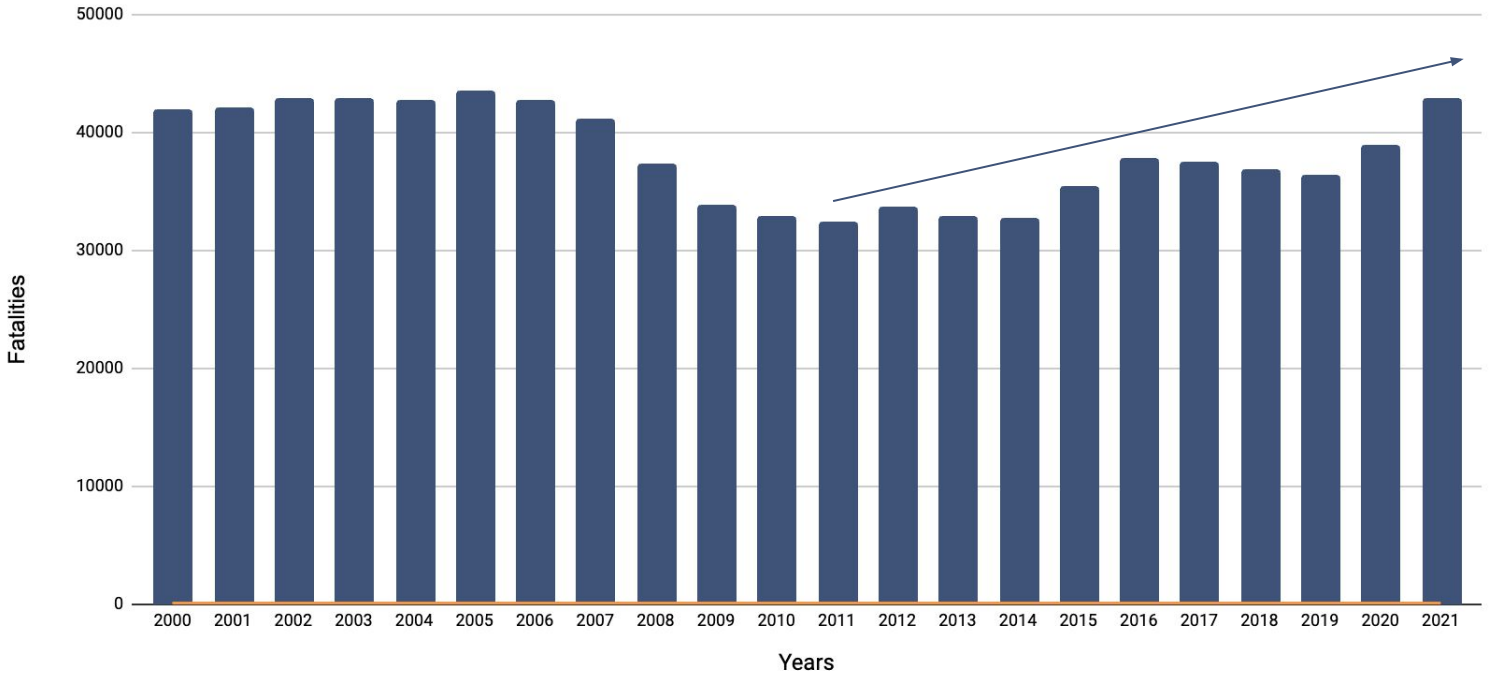
Each day

~118 deaths[2]

Each day

MOTOR VEHICLE SAFETY DATA

BUREAU OF TRANSPORTATION STATISTICS



BACKGROUND

Who is impacted?

Commuters, law enforcement agencies, emergency services, city planners, and insurance companies located in LA, California

Anticipated impact:

- Enhance commuter safety
- Optimized emergency responses
- City planning for safer infrastructure
- Insurance risk assessment improvement
- Economic impact mitigation

DATASET OVERVIEW

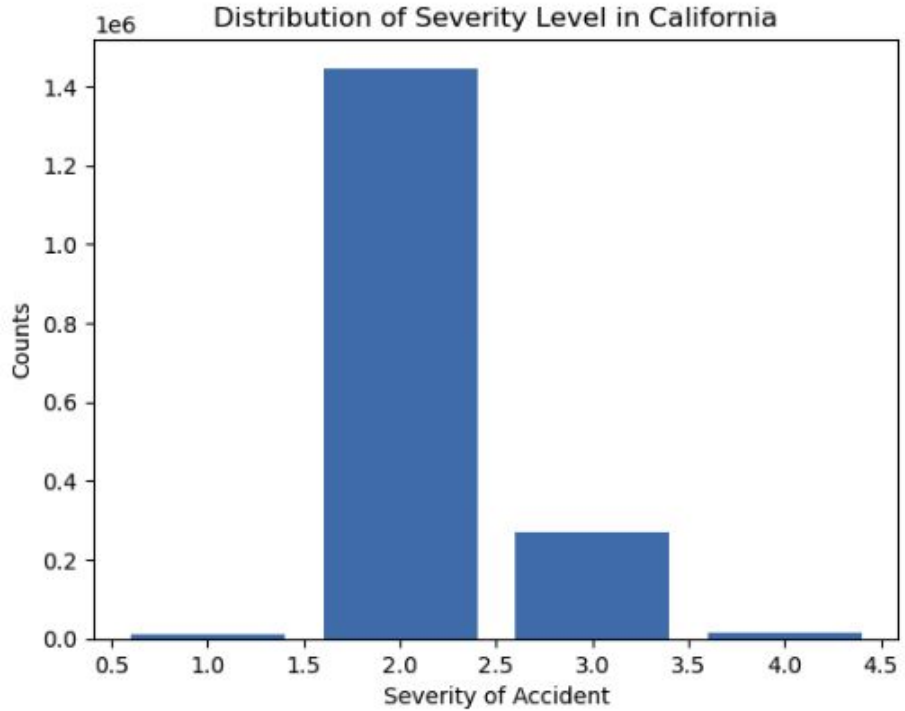
- Kaggle Dataset: comprehensive dataset of car accidents across the USA spanning from February 2016 to March 2023
- 7.7 million rows x 46 columns
  - Accident severity, timestamps, geographical coordinates, weather conditions, and various Points of Interest annotations

EDA AND DATA PREPROCESSING

- Out of the 45 columns, 22 columns had null values
  - End latitude, end longitude, precipitation, wind chill, and wind speed are the top 5 columns with most null values
- No duplicate rows
- Data types include: bool(13), float64(12), int64(1), object(20)

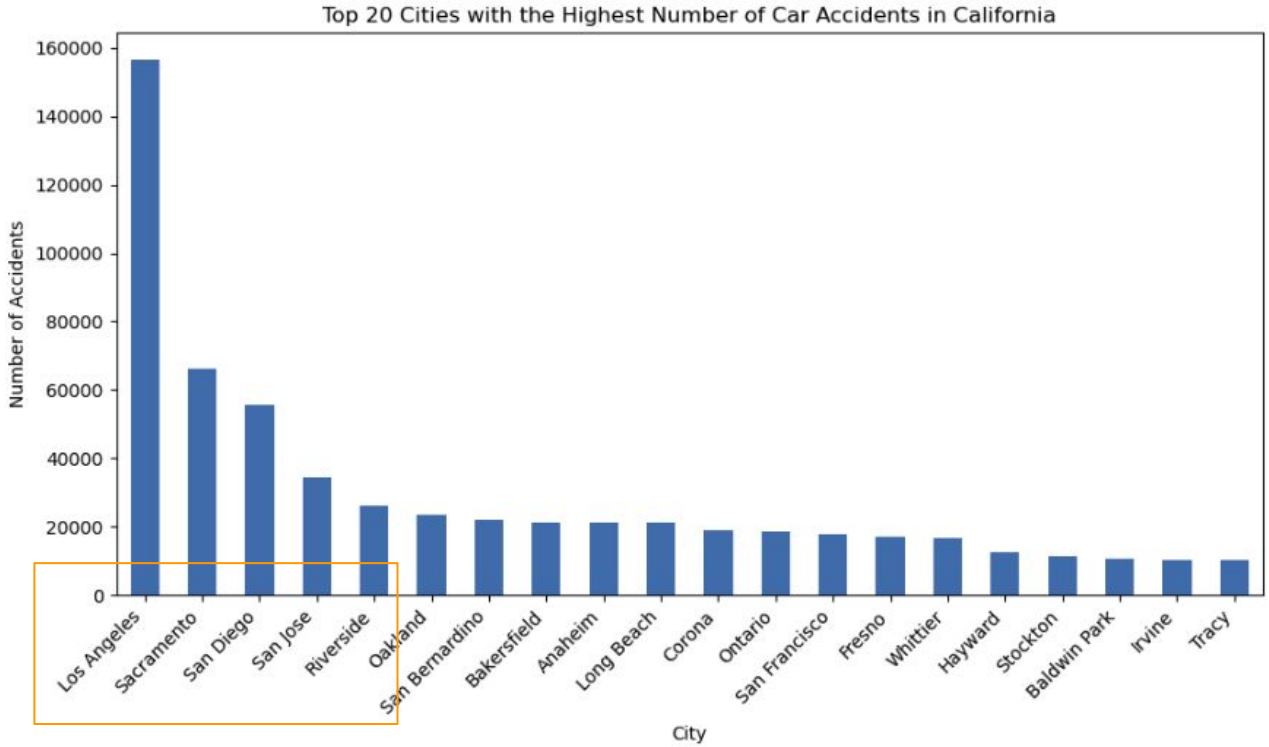
Columns	Null Values
End_Lat	3402762
End_Lng	3402762
Precipitation(in)	2203586
Wind_Chill(F)	1999019
Wind_Speed(mph)	571233
Visibility(mi)	177098
Wind_Direction	175206
Humidity(%)	174144
Weather_Condition	173459
Temperature(F)	163853
Pressure(in)	140679
Weather_Timestamp	120228
Sunrise_Sunset	23246
Civil_Twilight	23246
Nautical_Twilight	23246
Astronomical_Twilight	23246
Airport_Code	22635
Street	10869
Timezone	7808
Zipcode	1915
City	253
Description	5

DATA EXPLORATION
DISTRIBUTION OF SEVERITY LEVEL IN CALIFORNIA

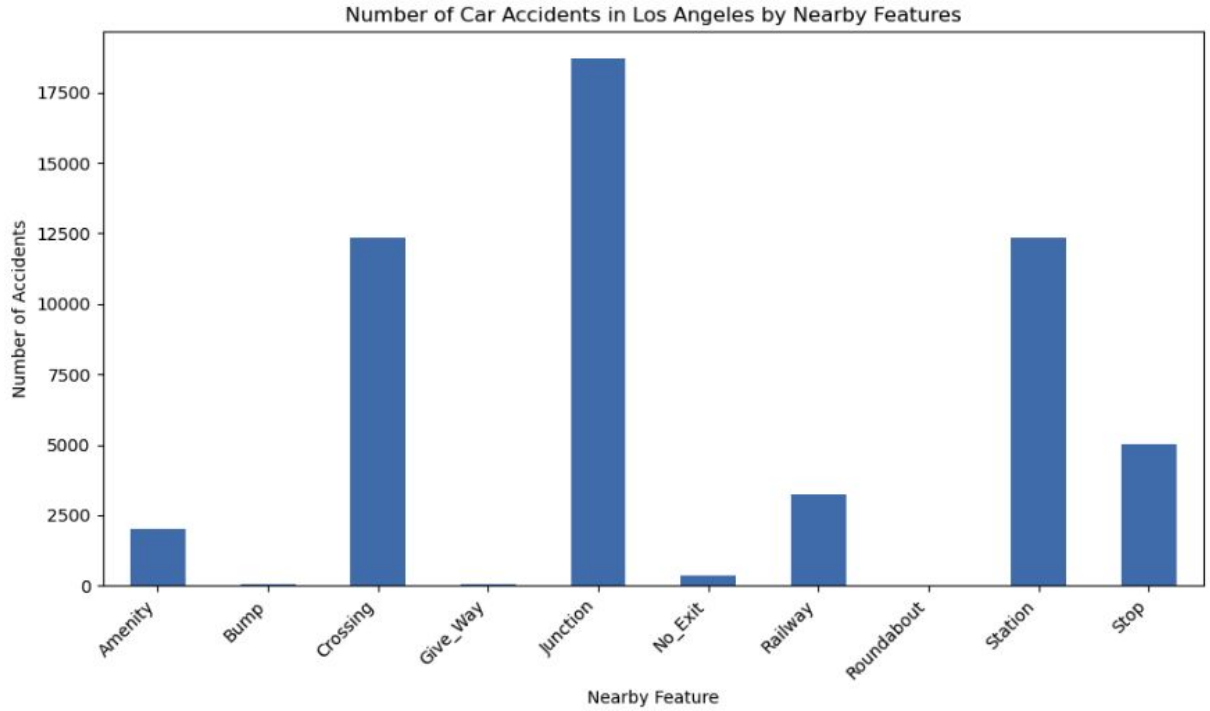


- Severity of the accident on a scale of 1 to 4 where 1 indicates minimal impact and 4 significant

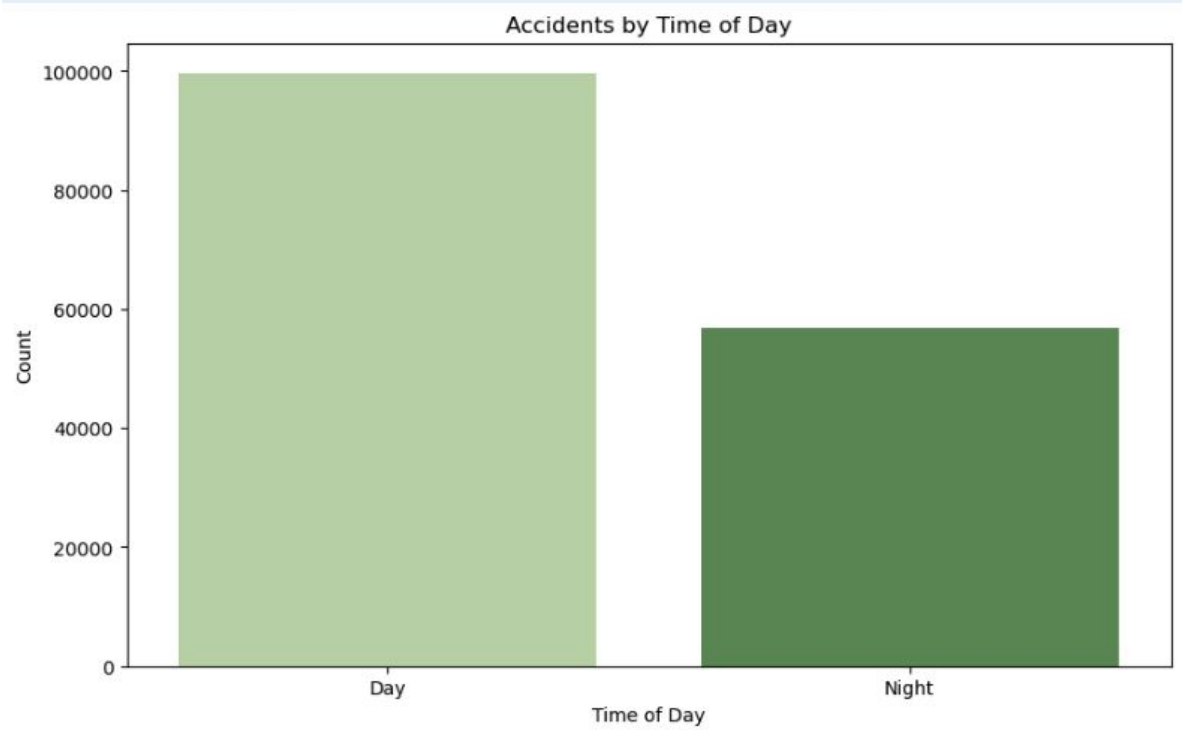
DATA EXPLORATION
NUMBER OF ACCIDENTS BY CITY IN LA



DATA EXPLORATION
NUMBER OF ACCIDENTS BY NEARBY FEATURES



DATA EXPLORATION
ACCIDENTS BY TIME OF DAY



MODEL COMPARISON

- **Model Evaluations:**
  - Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Gradient Boosting.
  - Compare performance metrics: accuracy, precision, recall, F1-score.
- **Strengths & Weaknesses:**
  - **Logistic Regression:** Simple, interpretable, may struggle with complex patterns.
  - **KNN:** Good performance, sensitive to outliers, and data scaling.
  - **Decision Tree:** Interpretability, risk of overfitting, sensitive to parameter tuning.
  - **Random Forest:** Strong performance, feature importance, potential for high variance.
  - **Gradient Boosting:** High accuracy, robust to imbalanced data, complex model.

MODEL COMPARISON CONTINUED

- **Challenges & Considerations:**
  - Addressing class imbalance.
  - Hyperparameter tuning for optimal performance.
  - Evaluating generalization through cross-validation.

MODEL INTERPRETATION

- Gradient Boosting Model.
- Reasons for Selection:
  - Performance: High accuracy, precision, recall, F1-score.
  - Interpretability: Feature importance analysis aids understanding.
  - Generalization: Consistent performance across data splits.
- Feature Importance:
  - Highlights key factors influencing accident severity.
  - Guides targeted road safety interventions and policy decisions.
- Model Robustness:
  - Handles complex relationships in data.
  - Good performance even with unbalanced classes.

NEXT STEPS

- Further address class imbalance (e.g., re-sampling techniques such as SMOTE)
- Feature engineering: investigate new features to enhance performance and interpretability
- Hyperparameter Tuning: Optimize learning rate, tree depth, and other parameters
- Collaborations: Partner with government agencies or transportation organizations to validate the model and impact

THANK YOU!

Any questions?

GENERAL REFERENCES

1. CDC. (2020, December 14). Road Traffic Injuries and Deaths—A Global Problem. Centers for Disease Control and Prevention. https://www.cdc.gov/injury/features/global-road-safety/index.html#:~:text=Road%20traffic%20crashes%20are%20a

2. How Many Car Accidents Occur Each Hour, Day & Year? (n.d.). Https://Amarolawfirm.com/. Retrieved January 11, 2024, from https://amarolawfirm.com/how-many-car-accidents-occur-each-hour-day-year-in-the-u-s/#

3. NHTSA: Traffic Crashes Cost America $340 Billion in 2019 | NHTSA. (2023, January 10). Www.nhtsa.gov. https://www.nhtsa.gov/press-releases/traffic-crashes-cost-america-billions-2019

DATASET REFERENCES

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.
- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.