Unlocking Road Safety Insights

ANALYSIS OF CRASHSTATS DATA

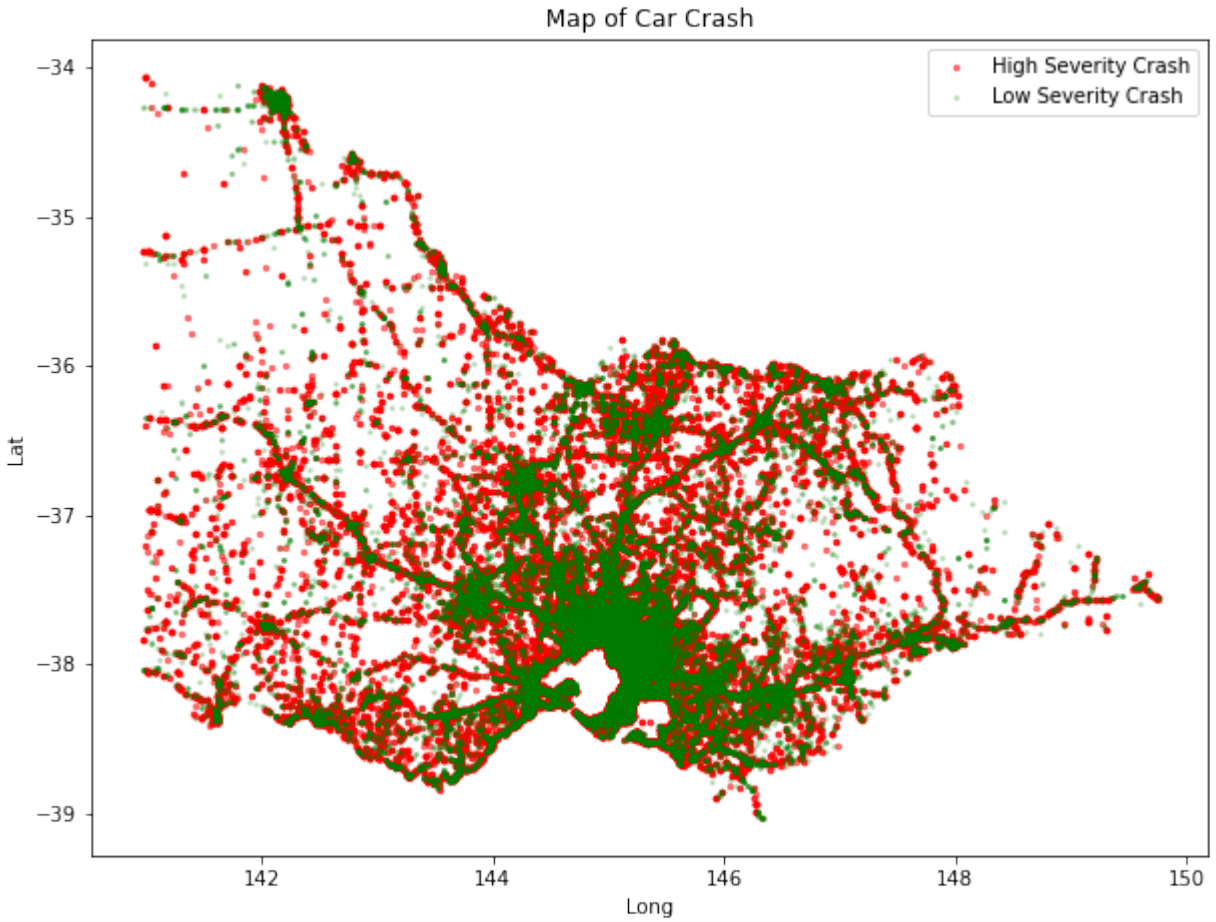
BRIAN YANG

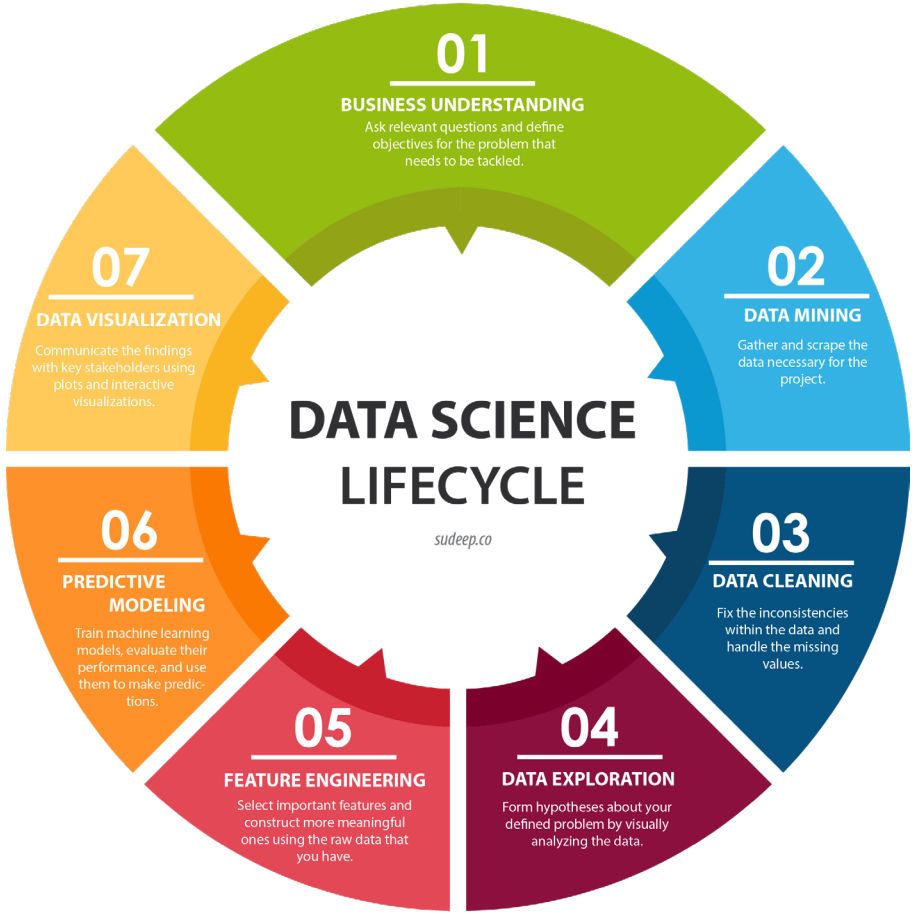
28 AUGUST 2023

Introduction

- Overview of the Project
- Crashstat Dataset (VicRoads)
- Objective and Value

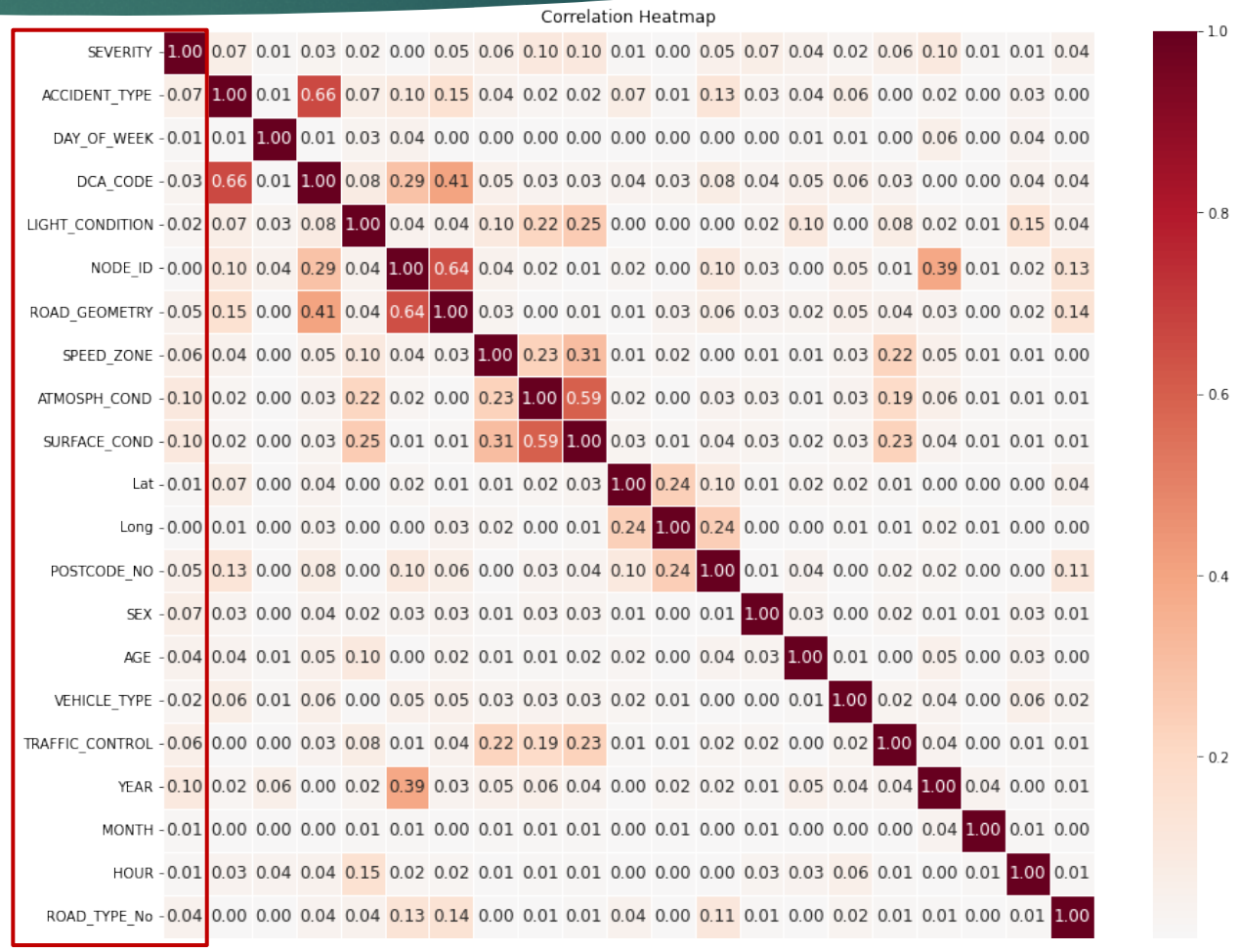
Overview of CrashStats Data Analysis



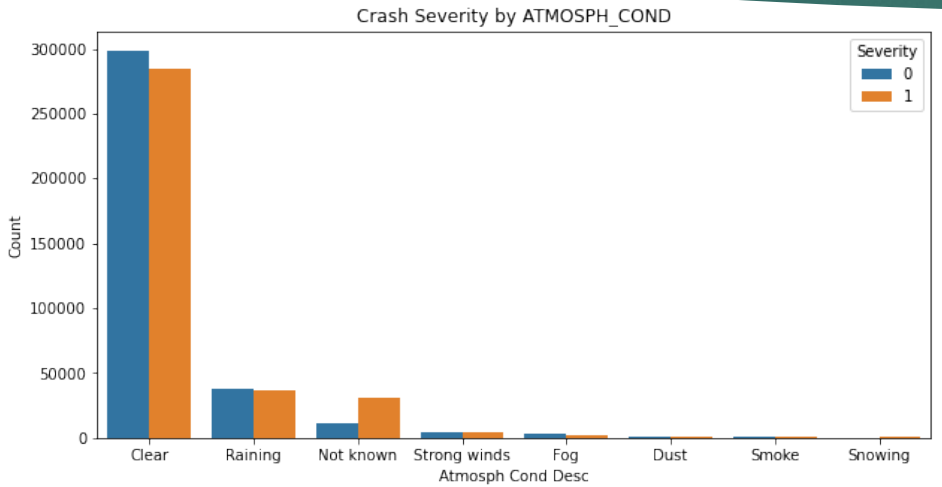


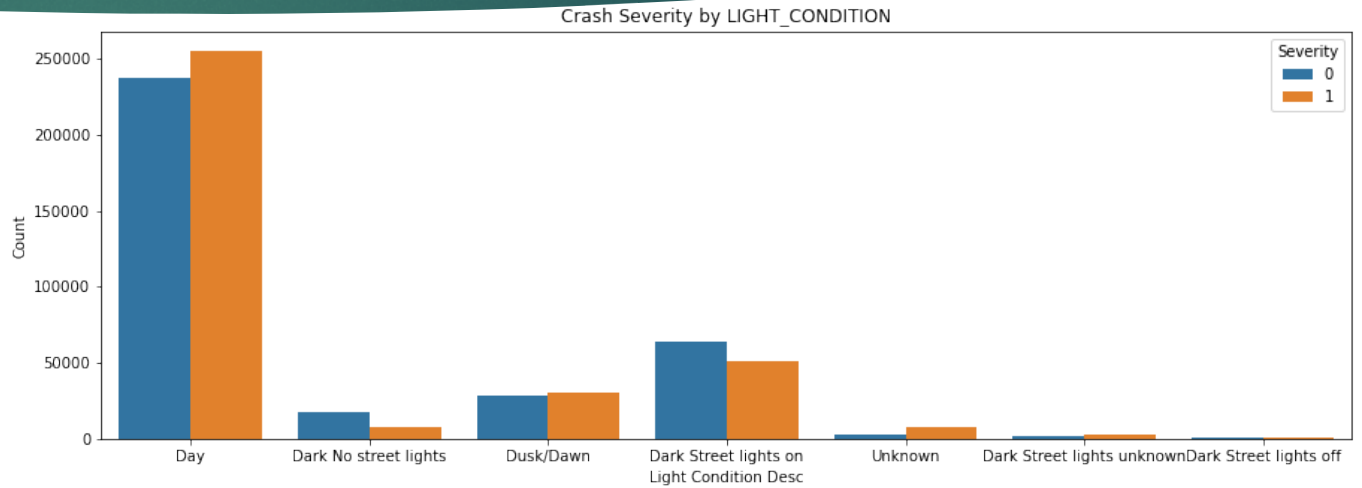
Key Insights from the Analysis

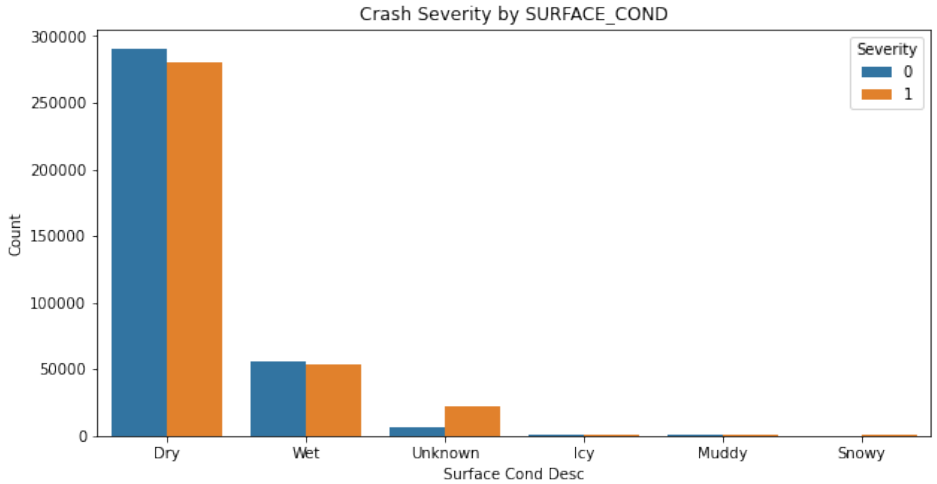
- Crash Severity Patterns
- Contributing factors
  - Location
  - Road
  - Human
  - Vehicle
- Temporal Trends
- Road Type Significance

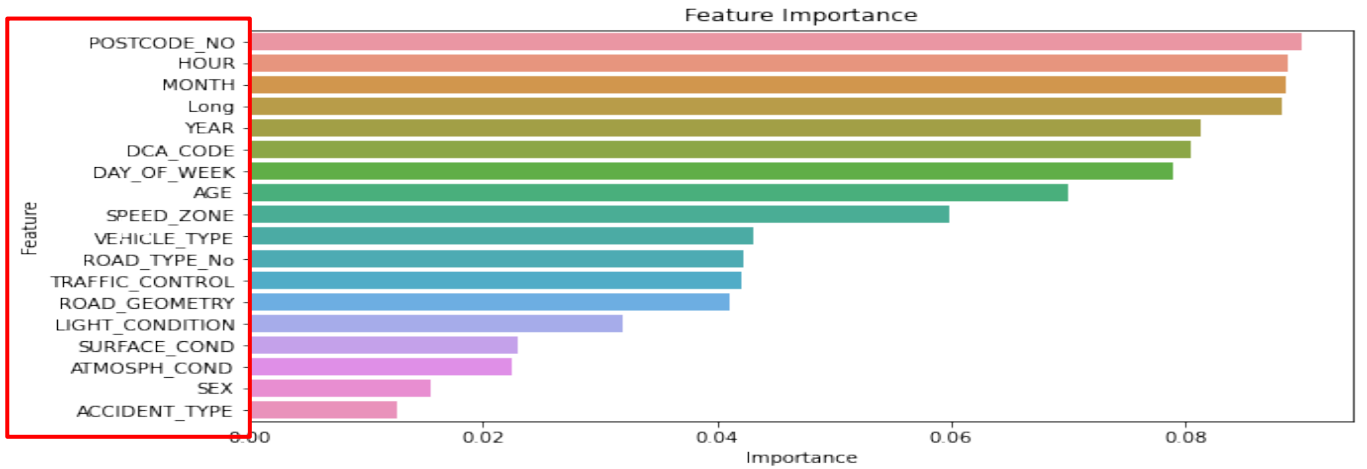


Contributing Factors and Risk Identification





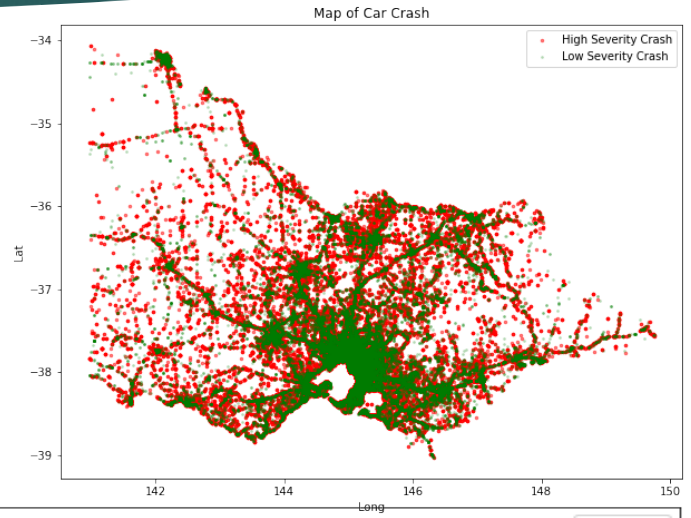


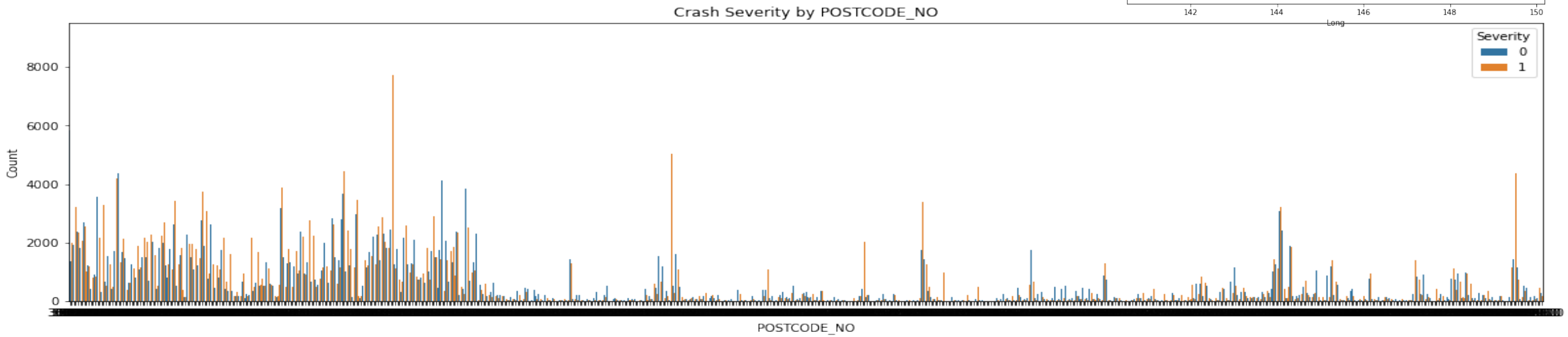


Geographical Crash Hotspots

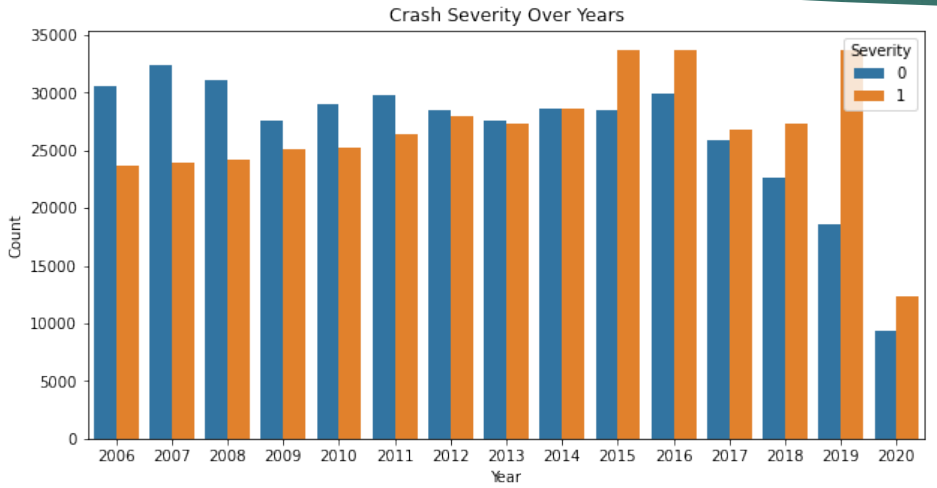
Trends and Patterns

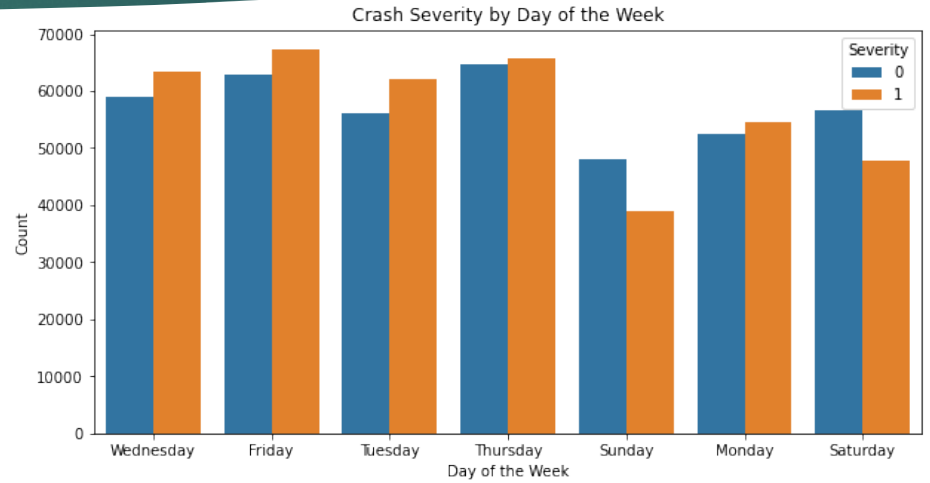
- Certain area(postcode) have more crash
- Metro – crash hotspots but less serious
- Regional – more server crashes on HW

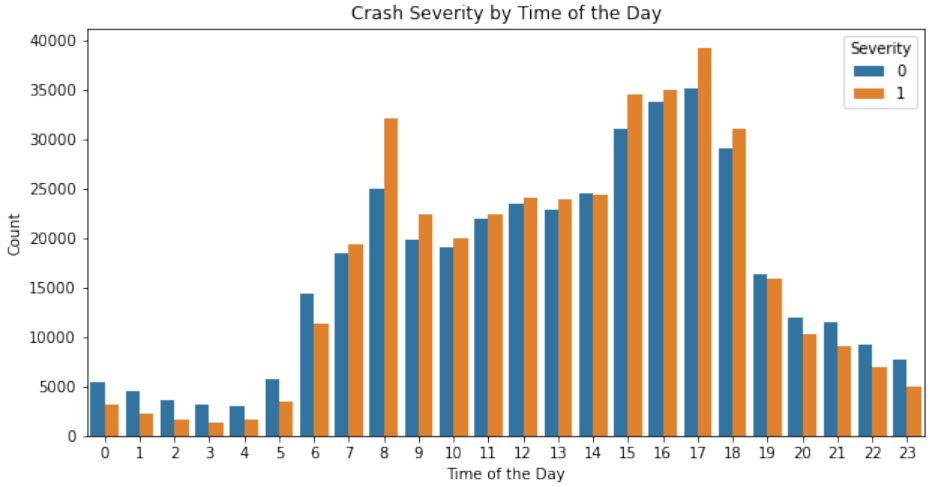




Time Trends and Seasonal Patterns





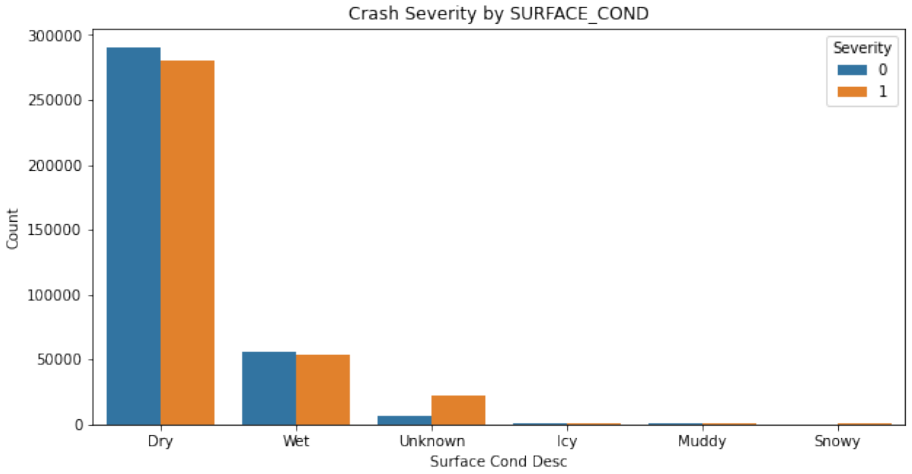


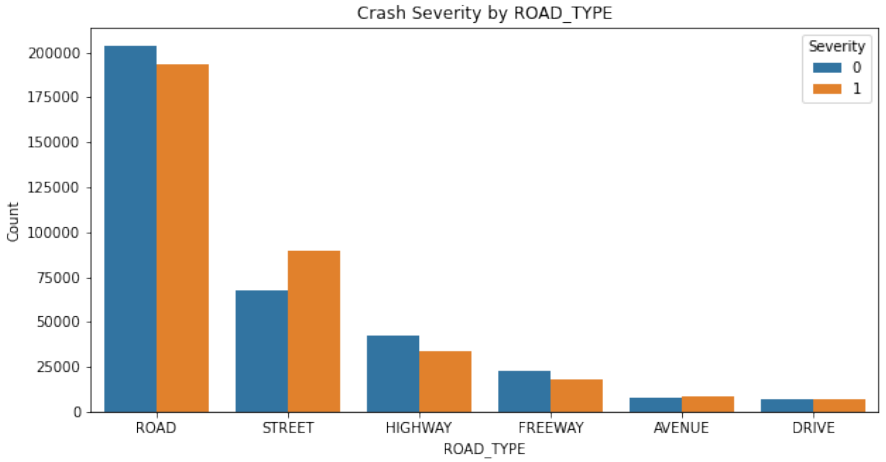
Trends and Patterns

- Peak accident times (**8am, 3-6pm**)
- Peak days (**Tue-Fri**)
- Total Incidents ▼
- High to Low Severity Ratio ▼

Road Type and Speed Zone Analysis

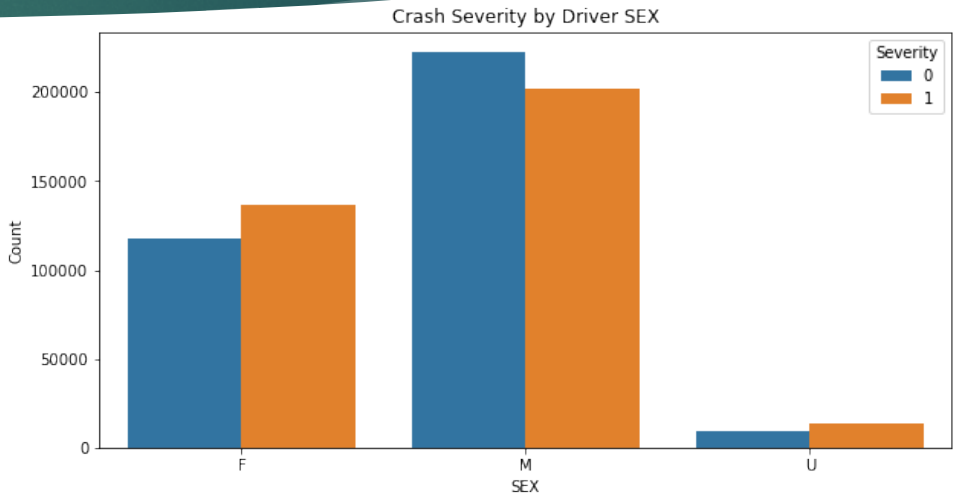
- Road Type
  - More car crashes on **RD**, **ST**, **HW and FW** happening in **Dry** surface condition
- Speed Zone
  - High speed crashes tend to more serious (e.g. 80 and 100)
  - Most of crashes are in **60 speed zone**

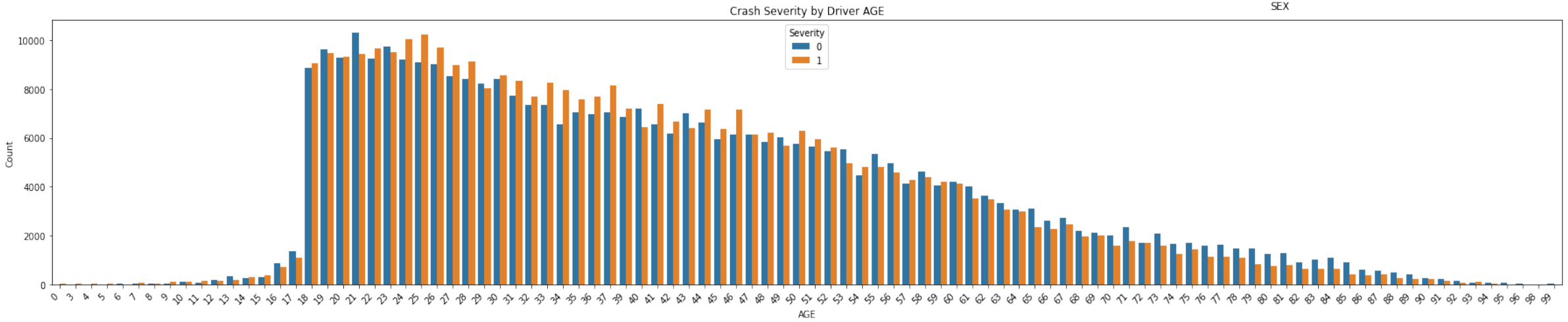




Crashing fact about drivers

- Age
  - **Younger** drivers cause more car crashes.
- Sex
  - Male drivers are responsible for more crashes, while crashes caused by females tend to be more serious.





Data-Driven Risk Management Strategies

- Utilising insights act as a guide for developing targeted risk management strategies
- Leveraging insights for resource allocation, safety interventions, and infrastructure improvements
- Collaboration for safer roads by leveraging data-driven risk management

Q&A Session

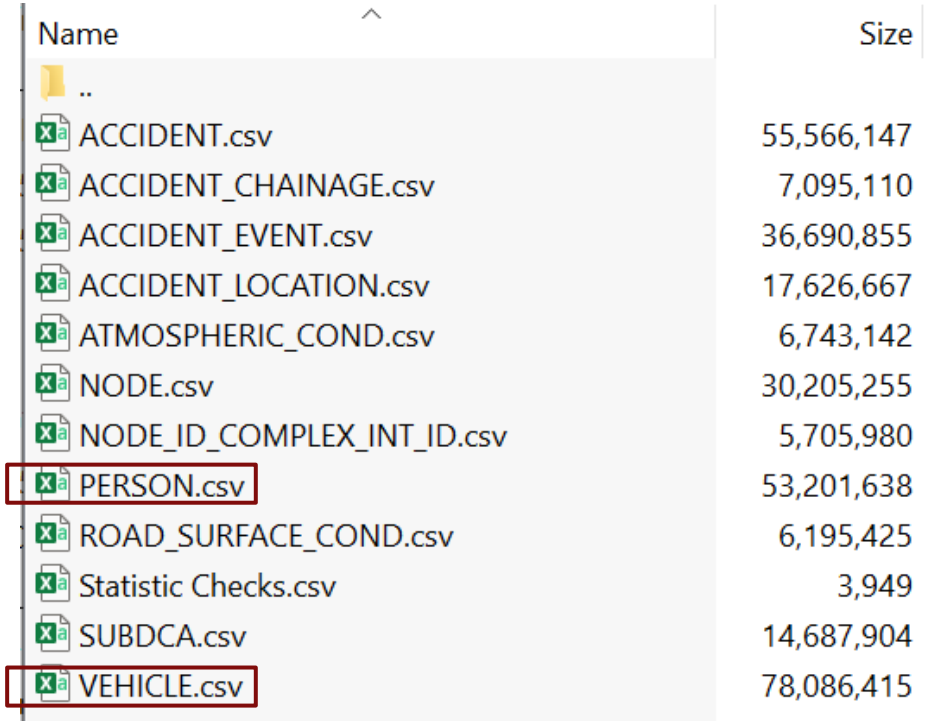
Thank You!

Data Extraction and Sources

- Primary Data Source: ACCIDENT.zip (crash data from 2006 to 2020) and Data Loading

Tables

- accident (basic accident details, time, severity, location)
- person (person based details, age, sex etc)
- vehicle (vehicle based data, vehicle type, make etc)
- accident_event (sequence of events e.g. left road, rollover, caught fire)
- road_surface_cond (whether road was wet, dry, icy etc)
- atmospheric_cond (rain, winds etc)
- sub_dca (detailed codes describing accident)
- accident_node (master location table - NB subset of accident table)
- Node Table with Lat/Long references



Name	Size
..	
ACCIDENT.csv	55,566,147
ACCIDENT_CHAINAGE.csv	7,095,110
ACCIDENT_EVENT.csv	36,690,855
ACCIDENT_LOCATION.csv	17,626,667
ATMOSPHERIC_COND.csv	6,743,142
NODE.csv	30,205,255
NODE_ID_COMPLEX_INT_ID.csv	5,705,980
PERSON.csv	53,201,638
ROAD_SURFACE_COND.csv	6,195,425
Statistic Checks.csv	3,949
SUBDCA.csv	14,687,904
VEHICLE.csv	78,086,415

Data Cleaning and Preparation

- Dropping unused data
  a) Remove empty space
  b) Remove unrelated features
  c) Remove certain human factors that highly correlated with the target
- Handling missing values
  a) Numerical features – mean
  b) Categorical features – mode (the most common value)
- Data formatting
  a) Datetime
  b) Road type

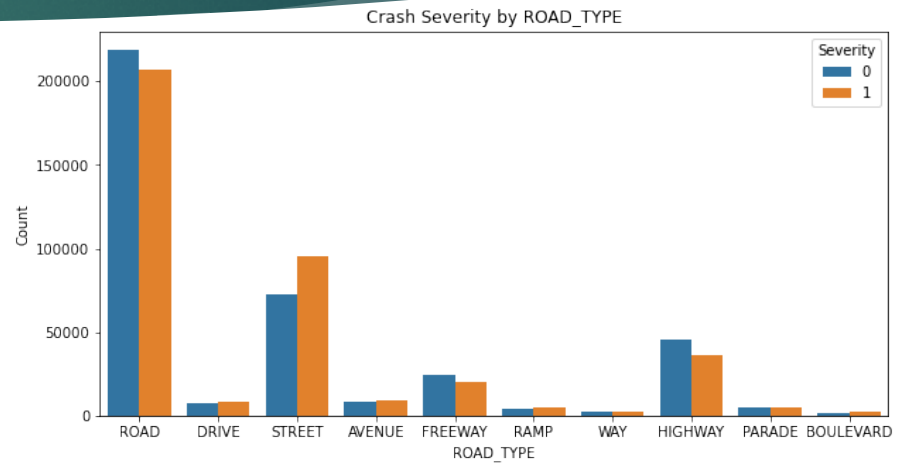
Resampling and Balancing

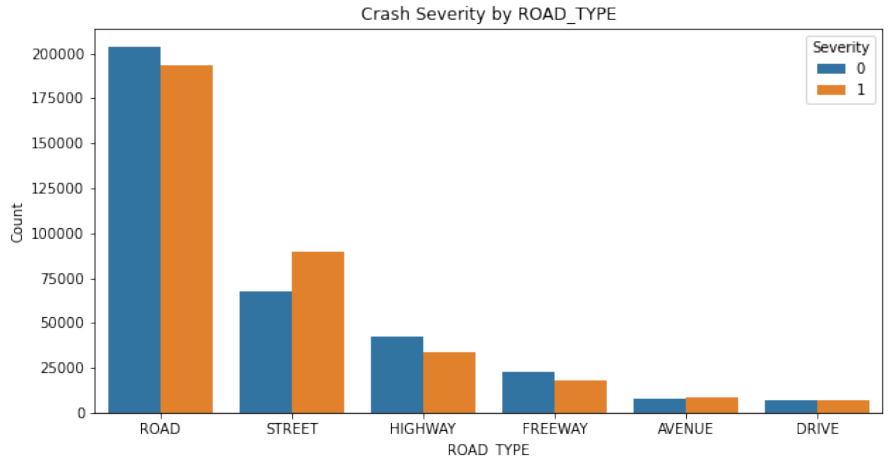
- Why Resampling and Balancing Matter
  1. Tackling Imbalance
  2. Strengthening Predictions
- The Impact of Imbalance
  1. Underperform in predicting rare outcomes
  2. Overestimate frequent outcomes
- Resampling Techniques
  - Oversampling
  - Undersampling

```
'SEVERITY' feature vaule count before resampling:
3    559666
2    252780
1     11796
4         8
Name: SEVERITY, dtype: int64
'SEVERITY' feature vaule count after resampling:
1    559674
0    264576
Name: SEVERITY, dtype: int64
Balanced resampled data: 0    400000
1    400000
Name: SEVERITY, dtype: int64
```

Data Exploration and Engineering

- Remove uncommon features
  a. Select the top 6 most frequent road types
  b. Select the top 40 most frequent accident types
- Remove bias data
  a. Drop rows with specified values in SPEED_ZONE
  b. Drop rows with specified values in ATMOSPH_COND
  c. Drop rows with specified values in SURFACE_COND
  d. Drop value [38] in AGE column





Modelling and Classification

- Classification models
  - **Random Forest:** An ensemble of decision trees that collaboratively make predictions. Their collective wisdom delivers robust and accurate results.
  - **Extra Trees**: Like Random Forest, but with a unique twist—randomising the process further. This introduces diversity and strengthens predictions.
- Training Test Split (test_size=0.2)
- Model Accuracy
  - Random Forest Accuracy: 0.9813
  - Extra Trees Accuracy: 0.9814