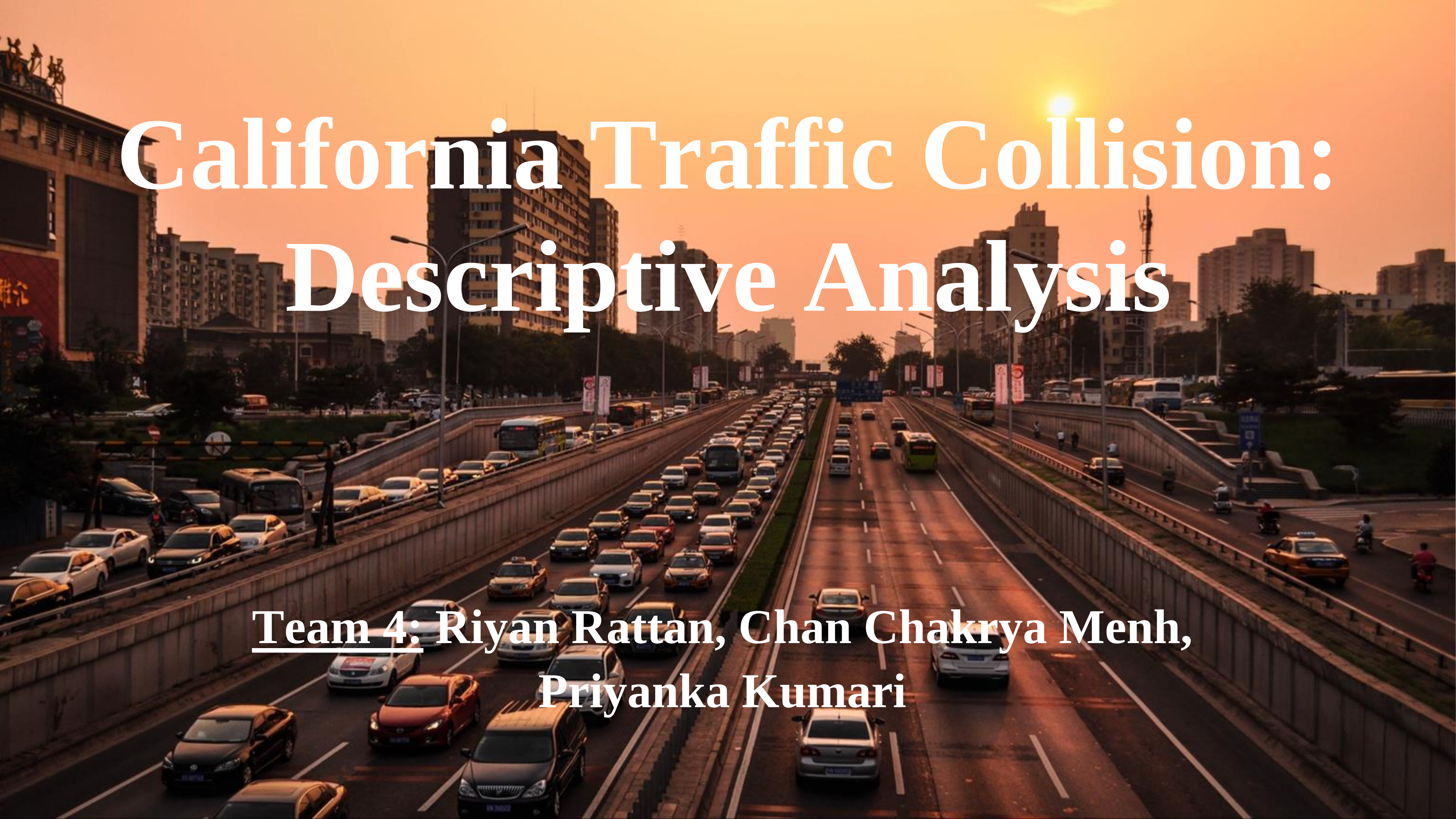


California Traffic Collision: Descriptive Analysis

Team 4: Riyan Rattan, Chan Chakrya Menh, Priyanka Kumari

Introduction

Background:

What is Traffic Collision?

A traffic collision or crash occurs when a vehicle collides with another vehicle, pedestrian, road barrier, or a stationary obstacle such as a tree or a utility pole. It may result in injury, death, vehicle damage, possession damage which causes death and disability, and financial burden.

The National Highway Traffic Safety Administration (NHTSA) disclosed its early estimation of traffic fatalities for 2021. NHTSA projects an estimated 42,915 individuals died in motor vehicle traffic crashes last year, a 10.5% expansion from the 38,824 fatalities in 2020. The projection is the highest fatalities since 2005 and the most significant annual percentage increase in the Fatality Analysis Reporting System history.

Motive/Goals:

- This project is to study California traffic collision by using dataset in 2019 to do descriptive analysis. This analysis will be achieved by two methods, one is utilizes Orange for machine learning to develop a model with training dataset and eventually with the test dataset, and second is Tableau for analysis insight

Research Questions

- How can we use machine learning to detect the type of collision?
- What type of collision that has more fatalities?
- Which month people was killed the most from traffic collision?
- What kind of weather has more fatalities?
- What genders are involved in the collision?
- What vehicle year has the most fatalities and injury?

Data Preparation

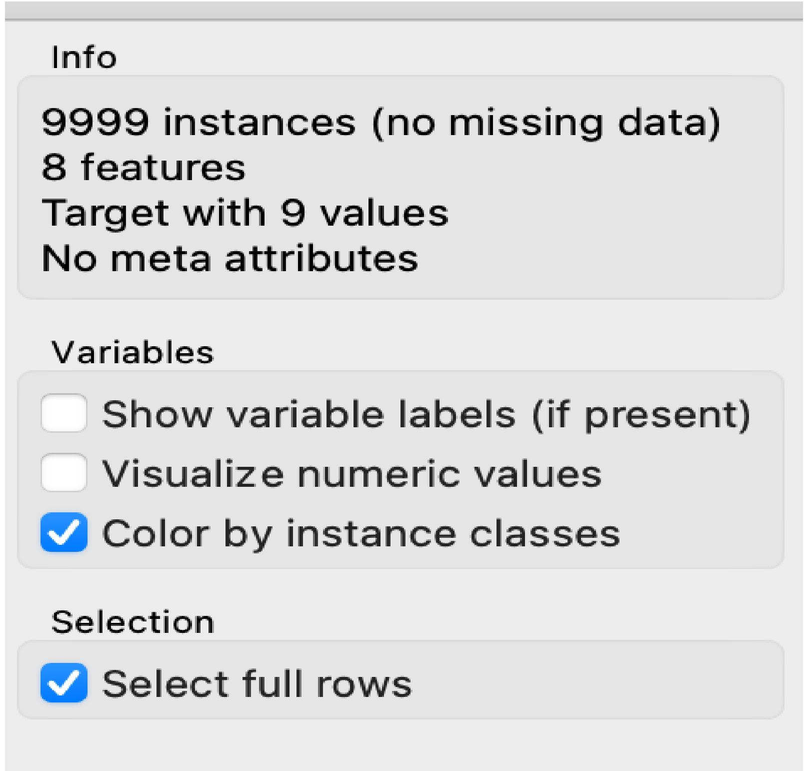
- Cleaning the dataset in csv format
  - Remove irrelevant columns, which eliminated missing values
  - Limit the sample size up to 10K to improve the process speed
- Import Data to Orange to check data attributes, statistics, and distribution
- Set Type of Collision and Weather as the target in the train data separately to compare
- Run test score, use cross validation, and run pipeline with Orange
- Use Tableau for data visualization

Data Analysis

- Create training and test data set
- Set target for the model to get traffic collision type and weather by using quantitative method
- Better data models for training the data set:
    - kNN
    - Decision Tree
    - Random Forest

Orange:

Create two pipeline for train and test data



Info

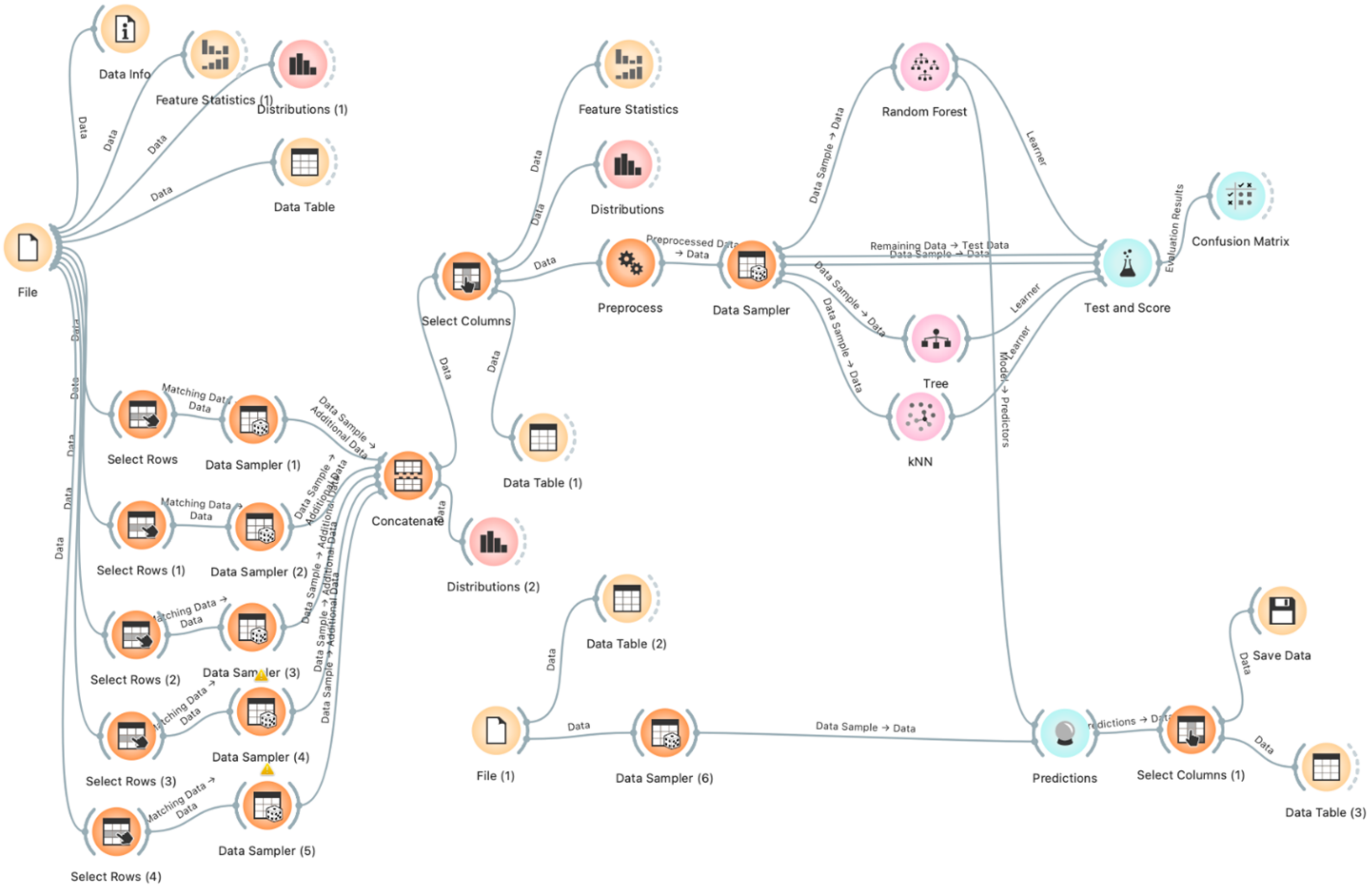
9999 instances (no missing data)
8 features
Target with 9 values
No meta attributes

Variables

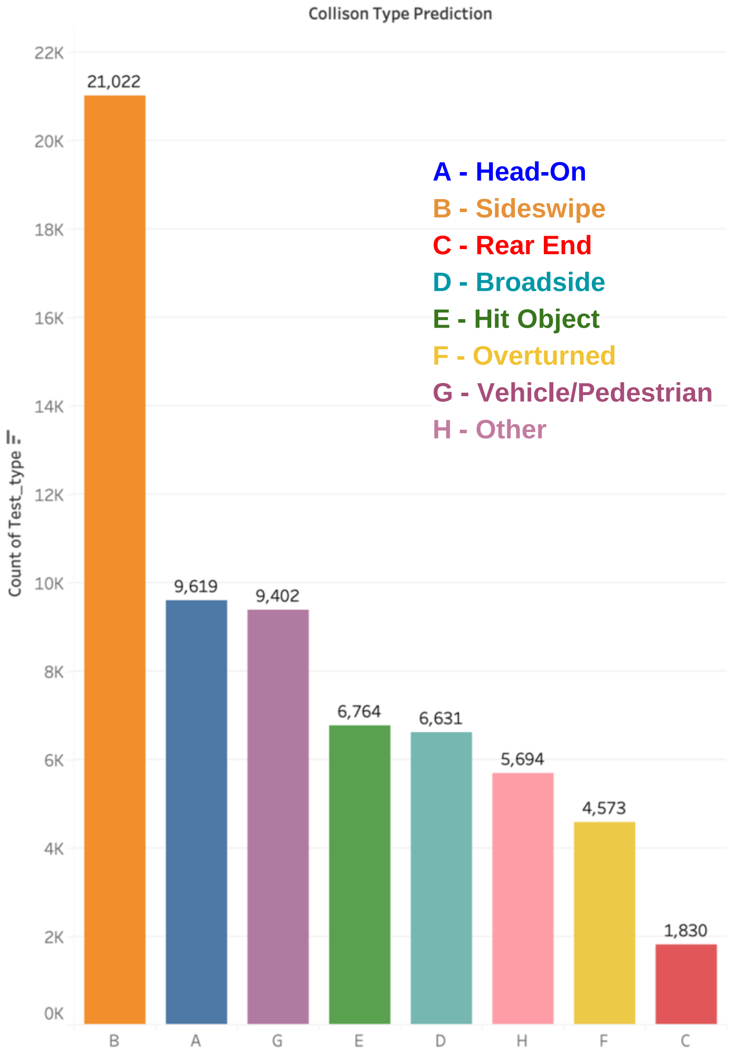
□ Show variable labels (if present)
□ Visualize numeric values
☑ Color by instance classes

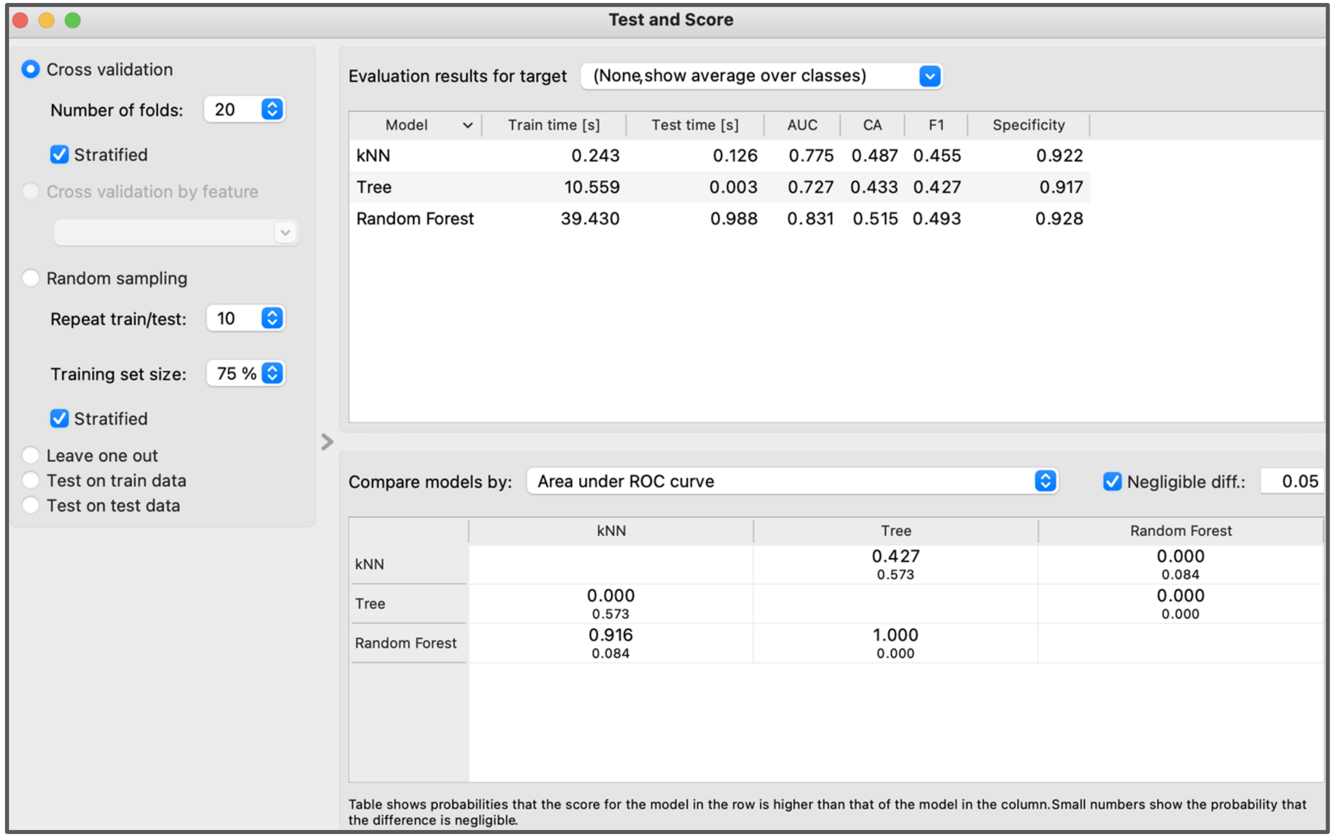
Selection

☑ Select full rows



Collision Type





Test and Score

● Cross validation

Number of folds: 20

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 75 %

☑ Stratified

○ Leave one out

○ Test on train data

○ Test on test data

Evaluation results for target (None,show average over classes)

Model	Train time [s]	Test time [s]	AUC	CA	F1	Specificity
kNN	0.243	0.126	0.775	0.487	0.455	0.922
Tree	10.559	0.003	0.727	0.433	0.427	0.917
Random Forest	39.430	0.988	0.831	0.515	0.493	0.928

Compare models by: Area under ROC curve ☑ Negligible diff.: 0.05

	kNN	Tree	Random Forest
kNN		0.427 0.573	0.000 0.084
Tree	0.000 0.573		0.000 0.000
Random Forest	0.916 0.084	1.000 0.000	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column.Small numbers show the probability that the difference is negligible.

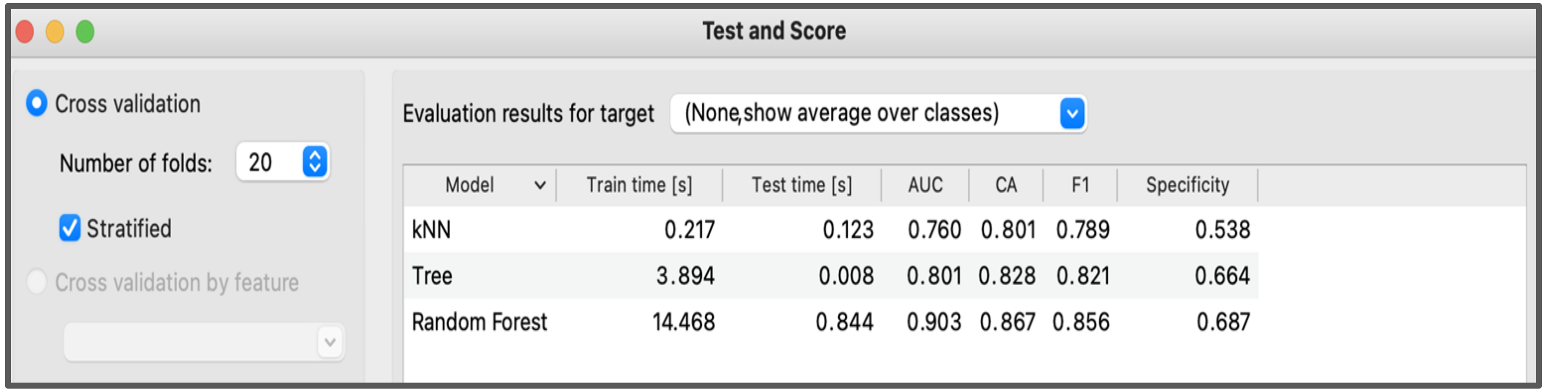
Model Evaluation (Train Dataset)

Weather Type

Target type is weather and the best result is Random Forest

Weather 1:

A - Clear
B - Cloudy
C - Raining
D - Snowing
E - Fog
F - Other
G - Wind



Test and Score

● Cross validation

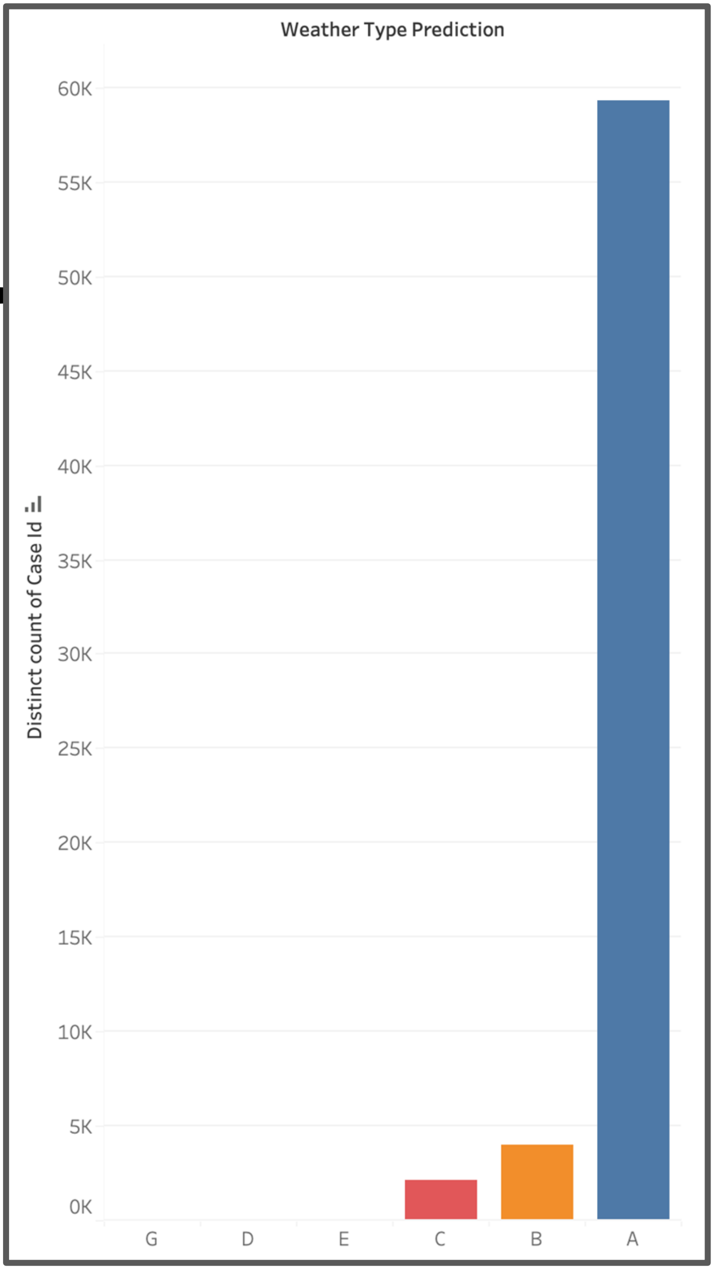
Number of folds: 20

☑ Stratified

○ Cross validation by feature

Evaluation results for target (None, show average over classes)

Model	Train time [s]	Test time [s]	AUC	CA	F1	Specificity
kNN	0.217	0.123	0.760	0.801	0.789	0.538
Tree	3.894	0.008	0.801	0.828	0.821	0.664
Random Forest	14.468	0.844	0.903	0.867	0.856	0.687



Data visualization: Tableau

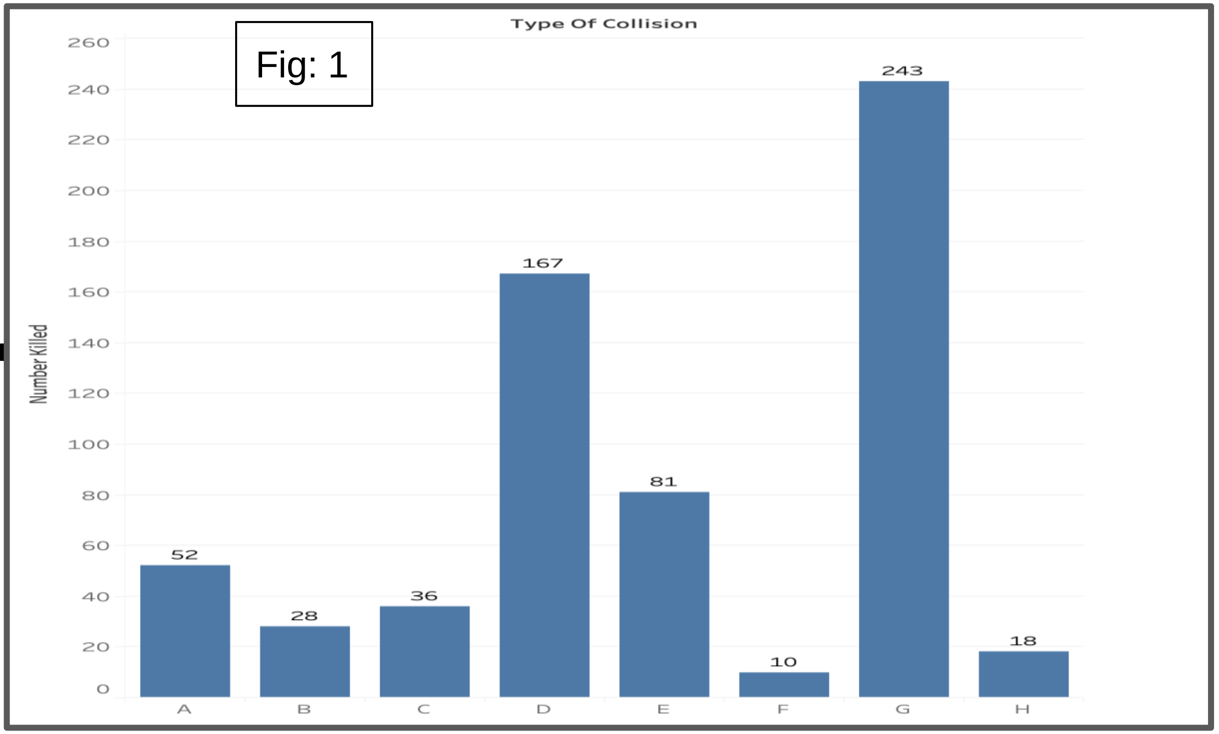


Fig: 1

Fig1: Number Killed Vs Type of Collision.

G type has most fatalities, and second is D type.

Fig2: Number Killed Vs Number of Injured,

D type has about 167 fatalities and about 27K injured.

Type of Collision:

A - Head-On

B - Sideswipe

C - Rear End

D - Broadside

E - Hit Object

F - Overturned

G - Vehicle/Pedestrian

H - Other

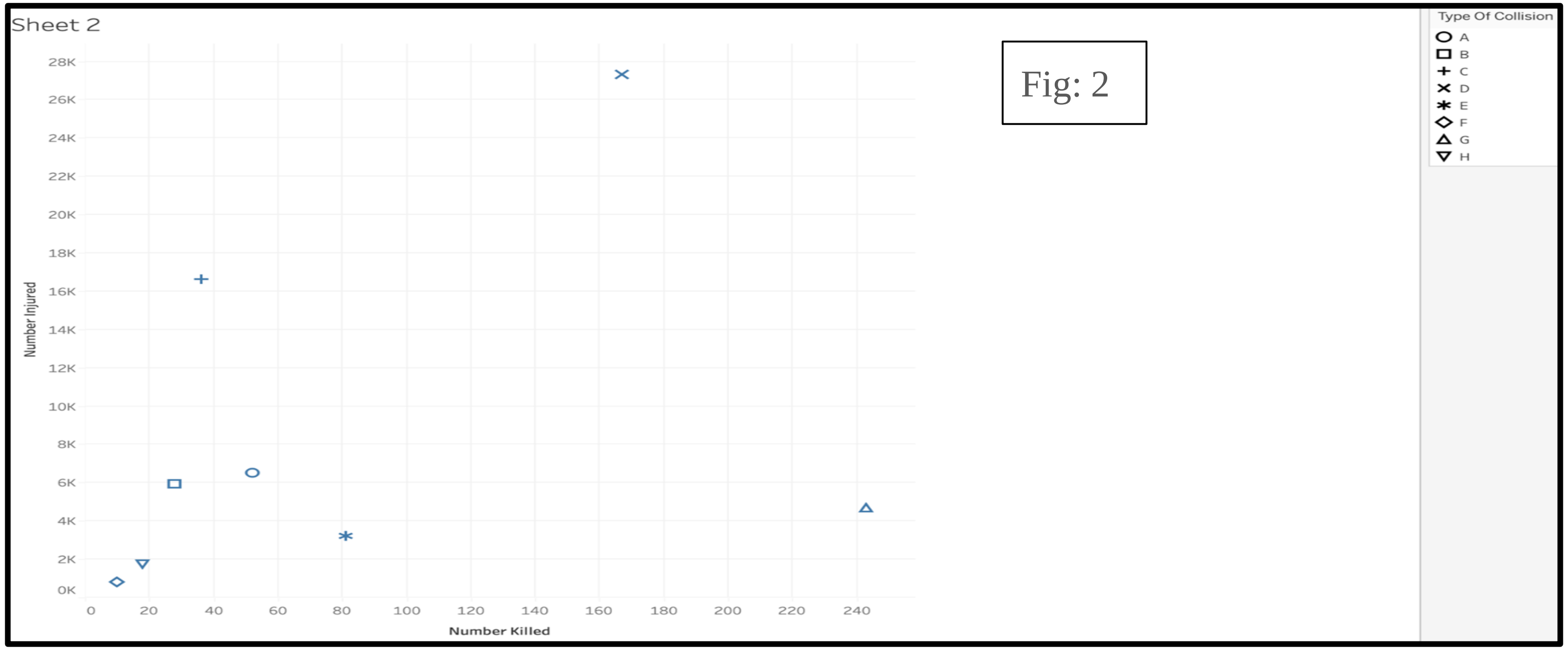


Fig: 2

Weather Vs Number Killed

This graph shows that most people died when the Weather is clear.

Weather 1:

A - Clear

B - Cloudy

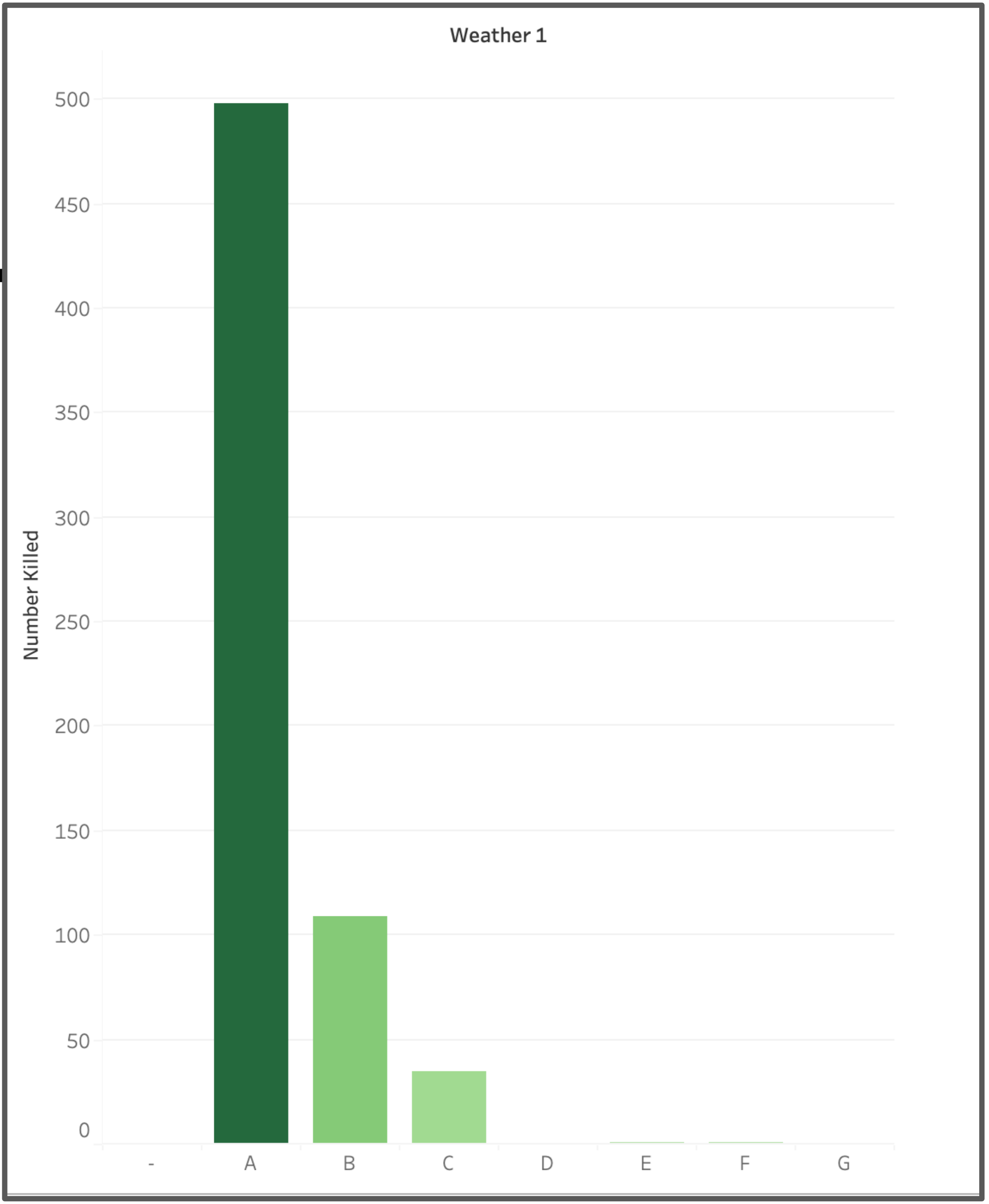
C - Raining

D - Snowing

E - Fog

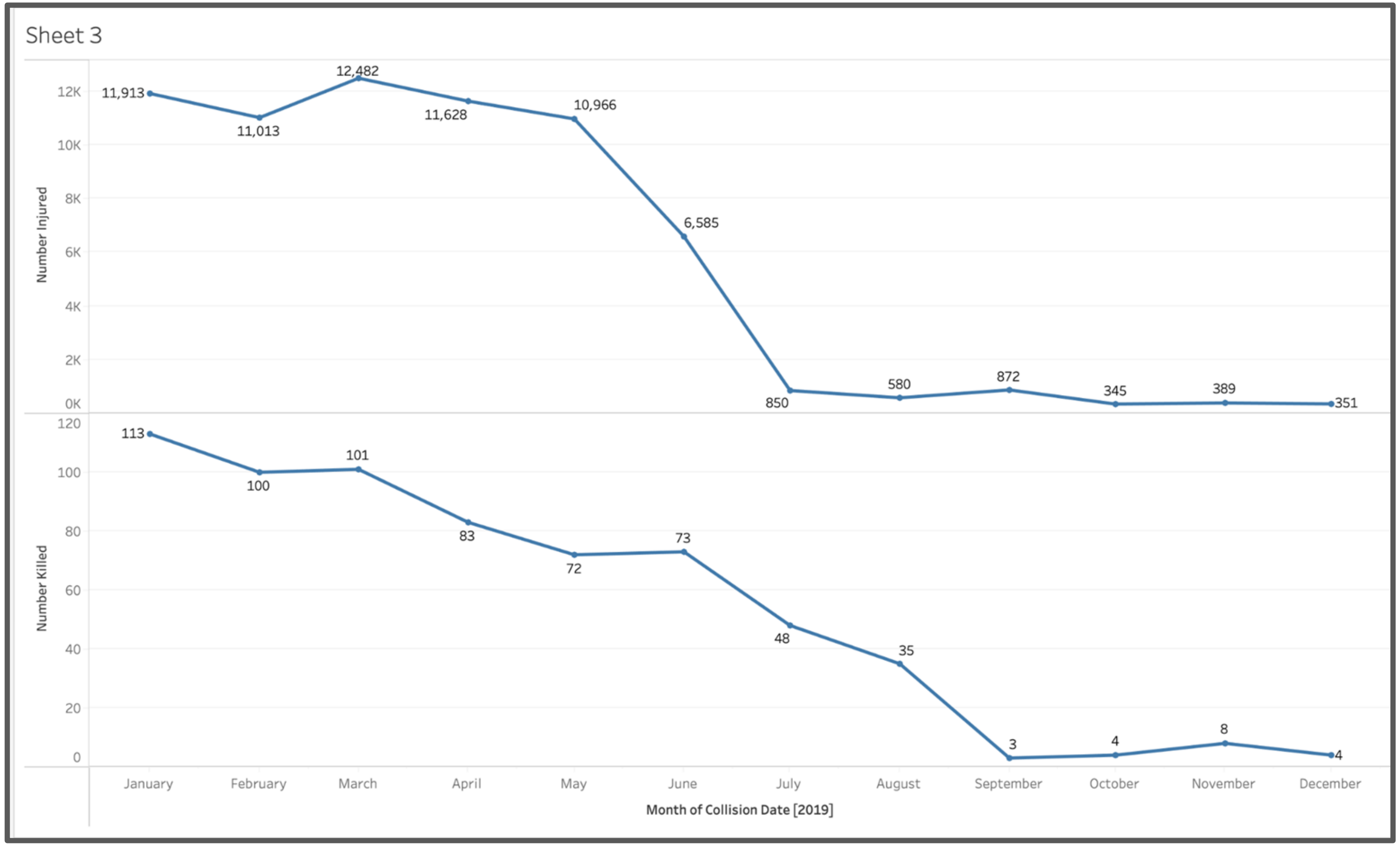
F - Other

G - Wind



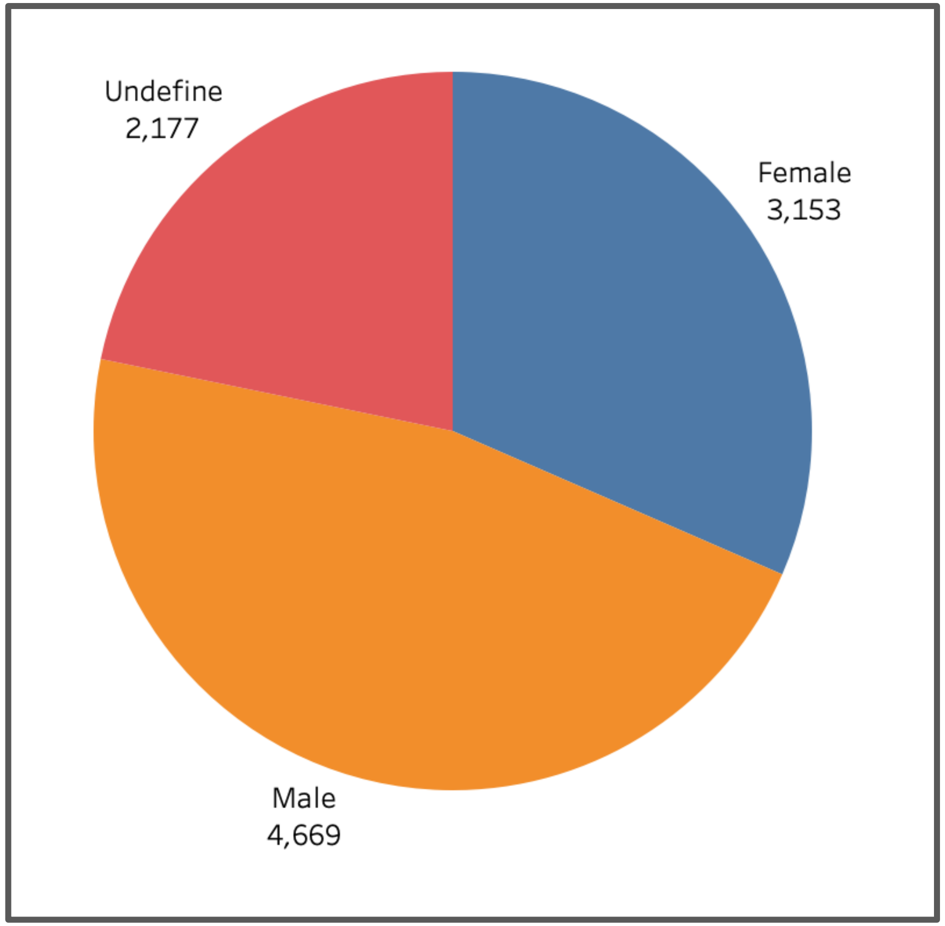
Number Killed Vs Number Injured by Month

In comparison, number of injured on monthly bases is more than number of killed.



Collision Case by Gender

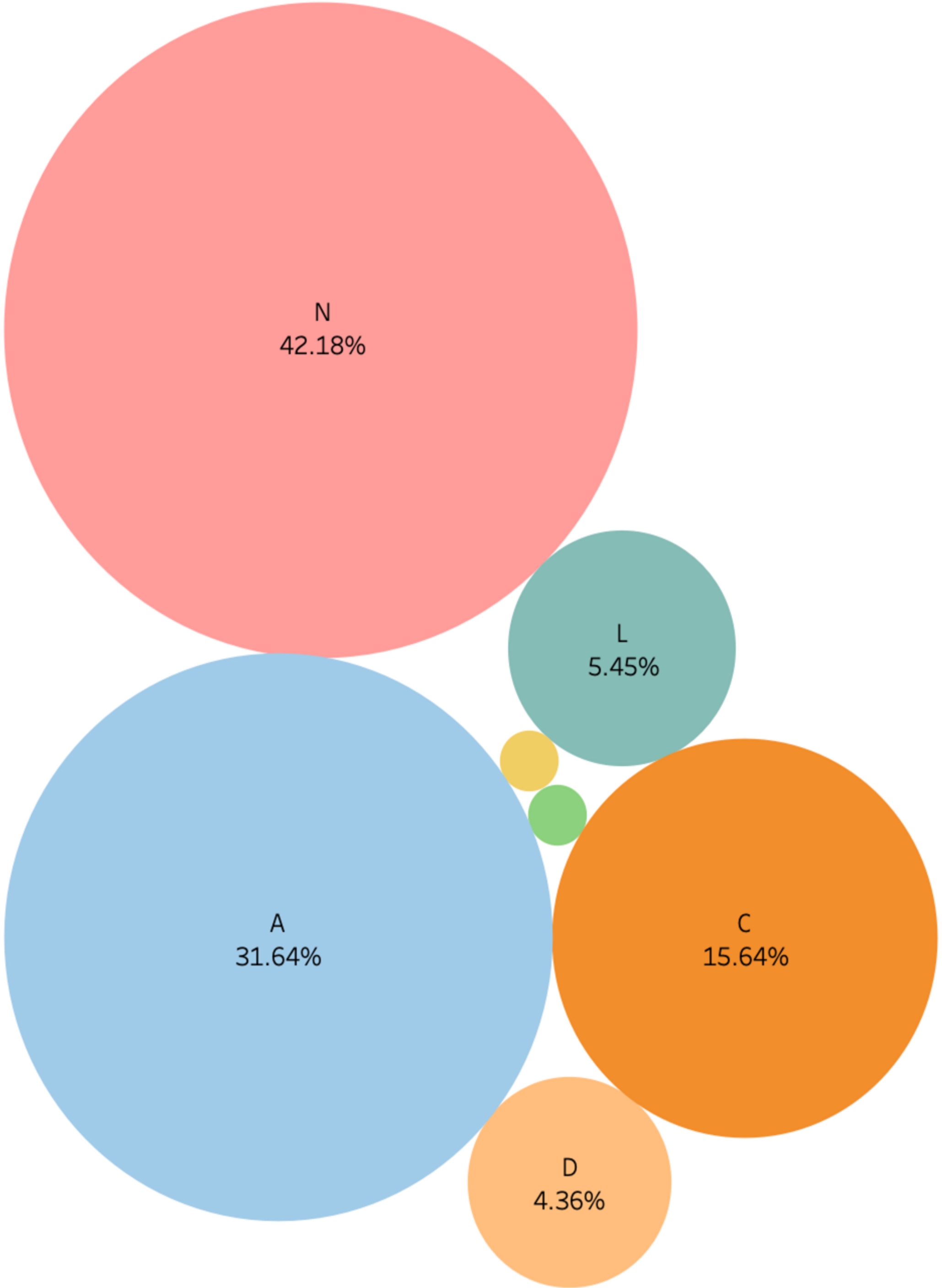
Proportion from the Dataset (10K) for female, male and undefine/unknown.



Number of Fatality by Vehicle Type

Statewide Vehicle Type:

A - Passenger Car/Station Wagon
B - Passenger Car with Trailer
C - Motorcycle/Scooter
D - Pickup or Panel Truck
E - Pickup or Panel Truck with Trailer
F - Truck or Truck Tractor
G - Truck or Truck Tractor with Trailer
H - Schoolbus
I - Other Bus
J - Emergency Vehicle
K - Highway Construction Equipment
L - Bicycle
M - Other Vehicle
N - Pedestrian
O - Moped
Blank - Not Stated



Number of Fatalities Vs Injured by Vehicle Year

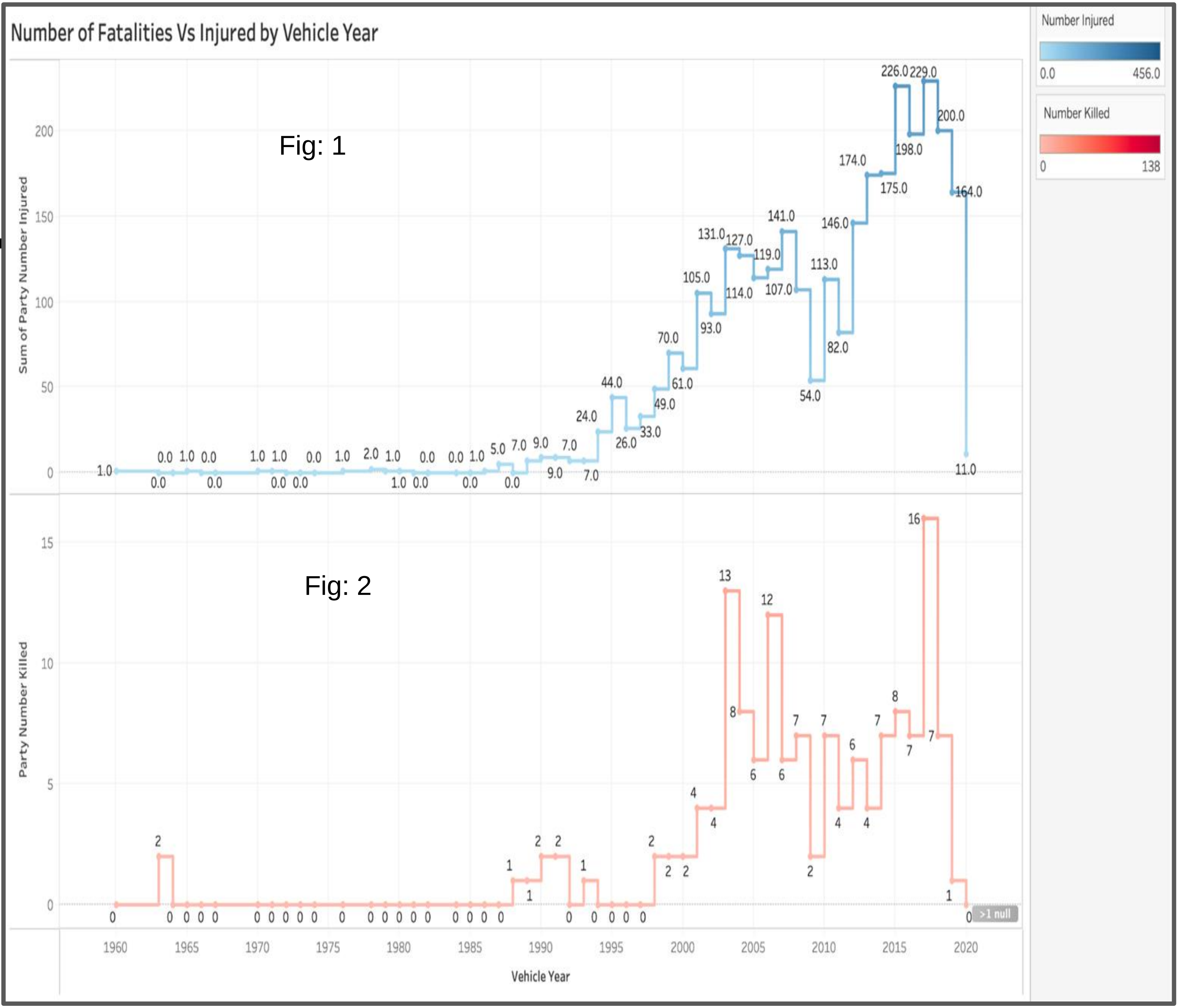


Fig 1: Describe about the total number of killed/fatalities in 1960 to 2020.

Fig 2: Describe about the total number of injuries in 1960 to 2020.

Project Action Plan

No.	Description	Start Date	End Date	Status
1.	Data Cleaning	July 17	July 24	15%
2.	Data Analyzing	July 25	July 31	45%
3.	Data Modeling	Aug 1	Aug 7	65%
4.	Data Visualization	Aug 8	Aug 14	75%
5.	Data Report Submission	Aug 15	Aug 24	100%

Future Plan

- Working on Modeling, improve accuracy and prediction analysis.
- Explore more on data

Sources:https://www.kaggle.com/datasets/sonicpsionic/california-switrs-collision-reports

Thank you!