US Traffic Accident Analysis

Team :

Rachel Wesley, Shrey Patel, Kaustubh Vilas Wani, Kamal Patel

16:540:507: Data Analytics in Engineering Systems

December 12, 2022

Presentation Outline

- Introduction of the Dataset
- Cleaning the Data
- Goals of the Study and Main Conclusions
- Models
- Word Cloud
- Conclusion of Analysis
- Recommendations for Future Work

Introduce the Data Set

Data Source

The dataset was obtained from kaggle from members of department of Computer Science and Engineering from Ohio State University.

Collected by Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath

Its contains nationwide car accidents collected from 2016 to 2021 which covers 49 states.. Currently contains abound ~3 million accidents records in the dataset with 47 attributes.

Data Source

Why was this data collected?

This data was collected with intention for insight and use for accident prevention, studying accident hotspot locations, casualty analysis, predict accidents, or study the impact of environmental stimuli on accident occurrence.

Why did we choose this dataset?

We thought it would be interesting to see if we can use methods learned in this course to visualize the data as well as model the data of previous accidents. We wanted to see if we could identify any patterns in the data.

Dataset Variables- 47 Attributes

ID- This is a unique identifier of the accident record.

Severity- Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay because of the accident) and 4 indicates a significant impact on

Start_Time- Shows start time of the accident in the local time zone.

End_Time- Shows the end time of the accident in the local time zone.

Start_Lat- Shows latitude in the GPS coordinate of the start point.

Start_Lng- Shows longitude in the GPS coordinate of the start point.

End_Lat- Shows latitude in the GPS coordinate of the end point.

End_Lng- Shows longitude in GPS coordinate of the end point.

Distance(mi)- The length of the road extent affected by the accident.

Description- Shows natural language description of the accident.

Street- Shows the street name in the address field.

State- Shows the state in the address field.

Dataset Variables- 47 Attributes

Country- Shows the country in the address field.

Timezone- Shows timezone based on the location of the accident (eastern, central, etc.).

Airport Code- Denotes an airport-based weather station which is the closest one to location of the accident.

Weather_Timestamp- Shows the timestamp of a weather observation record (in local time).

Temperature(F)- Shows the temperature (in Fahrenheit).

Wind_Chill(F)- Shows the wind chill (in Fahrenheit).

Humidity(%)- Shows the humidity (in percentage).

Pressure(in)- Shows the air pressure (in inches).

Visibility(mi)- Shows visibility (in miles).

Wind_Direction- Shows wind direction.

Dataset Variables- 47 Attributes

Junction- A POI annotation which indicates presence of a junction in a nearby location.

No_Exit- A POI annotation which indicates presence of no_exit in a nearby location.

Railway- A POI annotation which indicates presence of railway in a nearby location.

Roundabout- A POI annotation which indicates the presence of a roundabout in a nearby location.

Station- A POI annotation which indicates presence of a station in a nearby location.

Stop- A POI annotation which indicates presence of stop in a nearby location.

Traffic_Signal- A POI annotation which indicates presence of traffic signal in a nearby location.

Turning_Loop- A POI annotation which indicates the presence of a turning loop in a nearby location.

Sunrise_Sunset- Shows the period of day (i.e., day or night) based on sunrise/sunset.

Astronomical_Twilight- Shows the period of day (i.e., day or night) based on astronomical twilight.

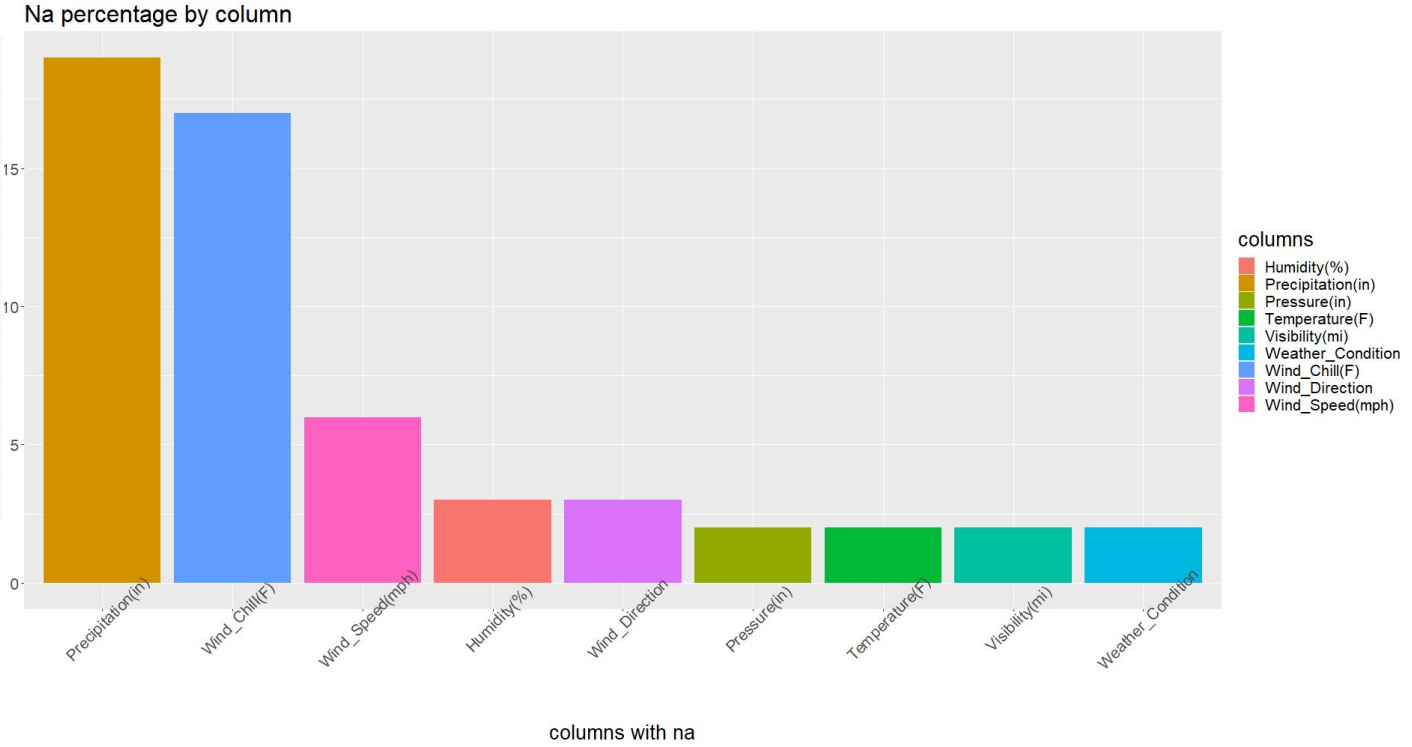
Cleaning Dataset

Data Cleaning

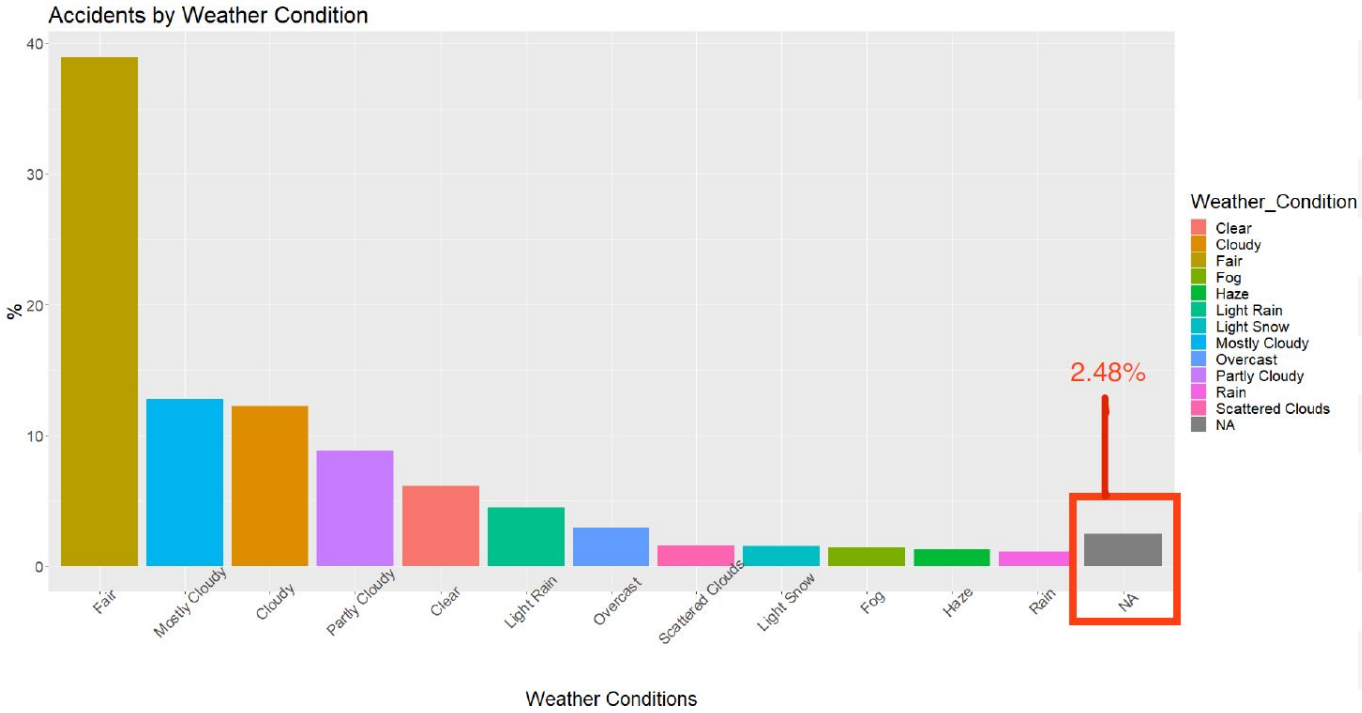
- Handle Missing Values
- Drop Weather condition NA Levels
- Location related variables
- Handle NA Values in Continuous Variables
- Handle NA values in Categorical Variables

Handle Missing Values

	columns	na
1	Precipitation(in)	19
2	Wind_Chill(F)	17
3	Wind_Speed(mph)	6
4	Humidity(%)	3
5	Wind_Direction	3
6	Temperature(F)	2
7	Pressure(in)	2
8	Visibility(mi)	2
9	Weather_Condition	2



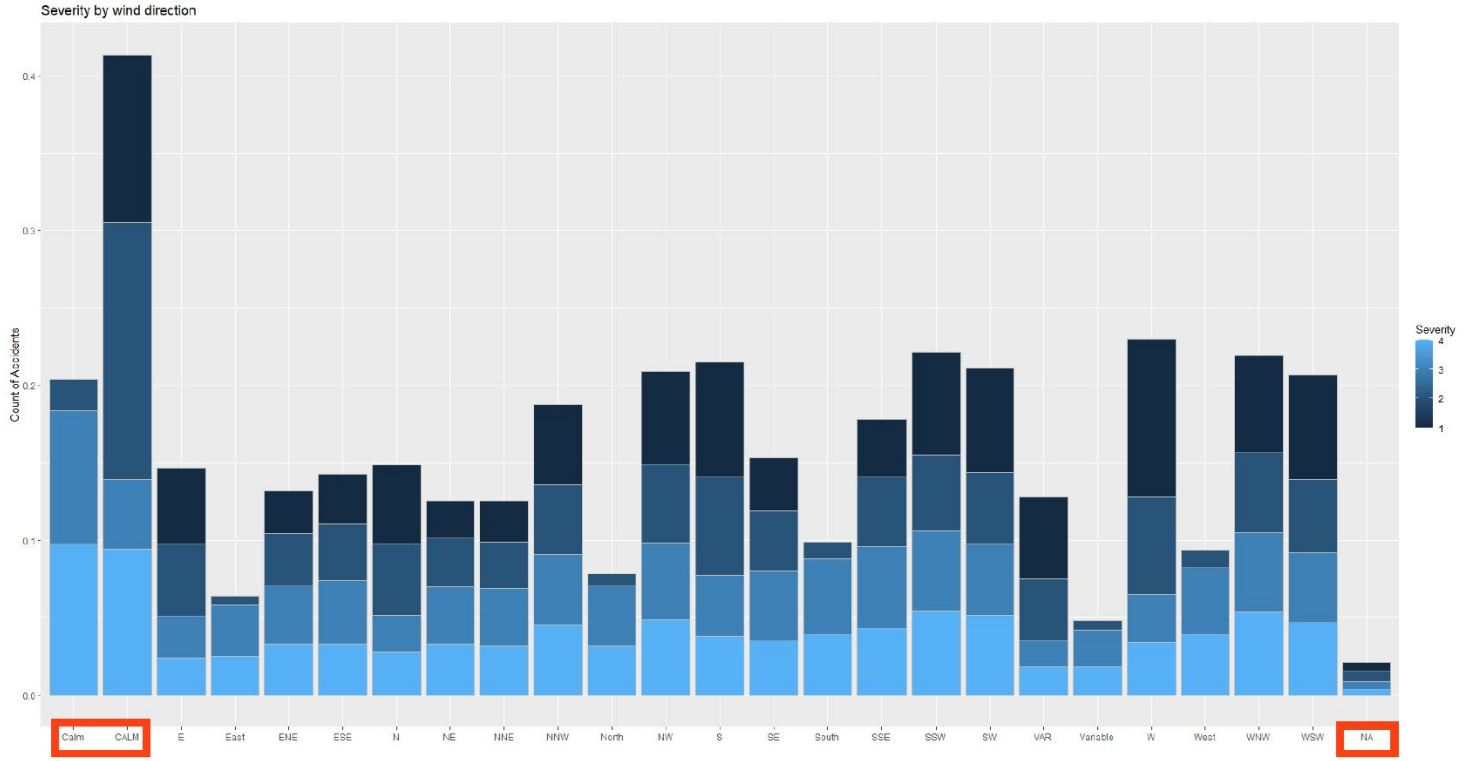
Weather NA Values



A tibble: 13 × 2

weather_condition	percentage
<chr>	<dbl>
Fair	38.912510
Mostly Cloudy	12.791397
Cloudy	12.257472
Partly Cloudy	8.784146
Clear	6.109037
Light Rain	4.512744
Overcast	2.983191
NA	2.482514
Scattered Clouds	1.586171
Light Snow	1.537671
Fog	1.448894
Haze	1.277667
Rain	1.091046

Severity by Wind Direction



Wind_Direction	n
CALM	431846
S	169216
W	167330
WNW	144519
NW	140891
SSW	136829
WSW	130361
SW	126537
SSE	125066
NNW	124109
E	123614
N	123159
SE	109147
VAR	103824
ESE	102357
ENE	94648
NE	88482
NNE	84701
Calm	75989
South	39027
West	39026
North	29397
East	23061
Variable	22065
NA	17505

Missing Categorical Values

We handle missing categorical values using two methods:

1. Remove all records containing the variables NA value.
2. Use NA as new level so other feature information is not lost.

A tibble: 10 × 8

Temperature	Wind_Chill	Humidity	Pressure	Visibility	Wind_Speed	Precipitation	Weather_Condition
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
17.6	10.7	93	29.83	2.0	4.6	NA	NA
12.0	-4.0	77	30.17	9.0	15.0	0	NA
12.0	-4.0	77	30.17	9.0	15.0	0	NA
NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	29.70	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA
30.2	18.1	93	29.50	2.5	18.4	NA	NA
30.2	18.5	93	29.49	1.0	17.3	NA	NA
64.4	NA	49	30.28	10.0	10.4	NA	NA

- Before we decide, there is one interesting thing: when weather related values is missing, there is a good chance that other weather related variables will be missing too.

Handle NA Values in Continuous Variables

- There are continuous variables with NA values, which is not a big issue.
- We can replace these NA values with the **mean** of the corresponding variable.

A tibble: 10 × 13

Distance	Temperature	Wind_Chill	Humidity	Pressure	Visibility	Wind_Speed	Precipitation
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
3.230	42.1	36.1	58	29.76	10.0	10.4	0.00
0.747	36.9	NA	91	29.68	10.0	NA	0.02
0.055	36.0	NA	97	29.70	10.0	NA	0.02
0.123	39.0	NA	55	29.65	10.0	NA	NA
0.500	37.0	29.8	93	29.69	10.0	10.4	0.01
1.427	35.6	29.2	100	29.66	10.0	8.1	NA
0.227	33.8	NA	100	29.63	3.0	2.3	NA
0.521	33.1	30.0	92	29.63	0.5	3.5	0.08
0.491	39.0	31.8	70	29.59	10.0	11.5	NA
0.826	32.0	28.7	100	29.59	0.5	3.5	0.05

Transform Time Variable

- Feature Engineering

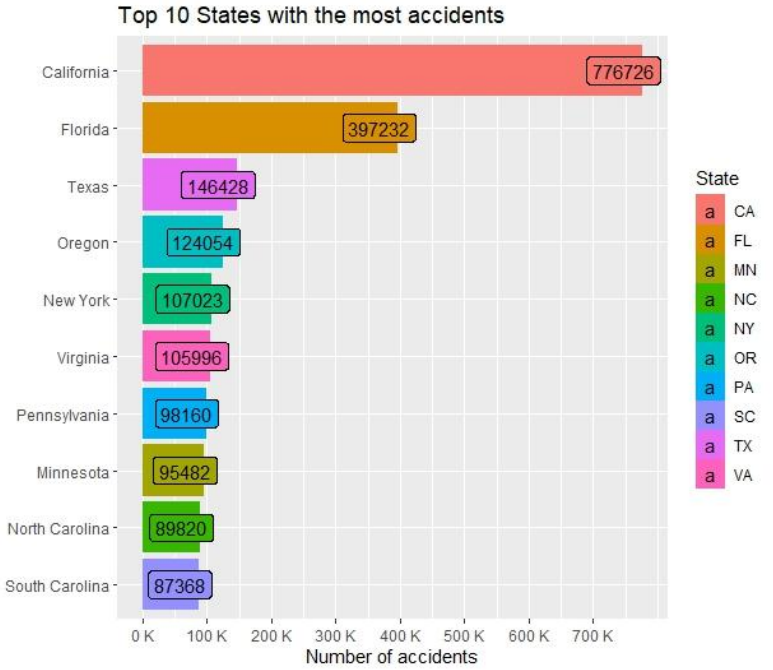
A tibble: 5 × 2

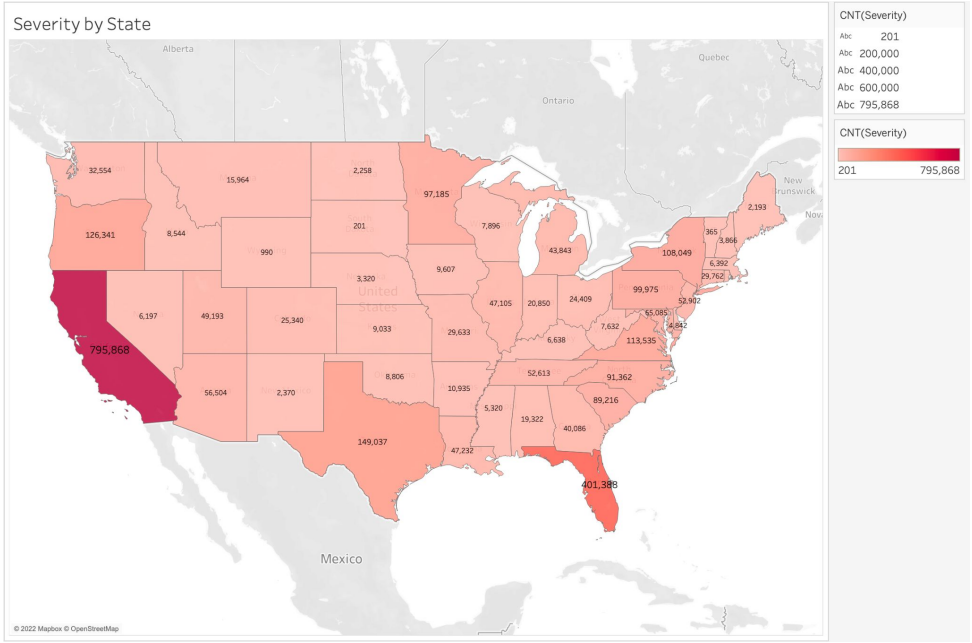
Start_Time	End_Time
<dttm>	<dttm>
2016-02-08 00:37:08	2016-02-08 06:37:08
2016-02-08 05:56:20	2016-02-08 11:56:20
2016-02-08 06:15:39	2016-02-08 12:15:39
2016-02-08 06:51:45	2016-02-08 12:51:45
2016-02-08 07:53:43	2016-02-08 13:53:43

A tibble: 5 × 6

Year	Month	Day	Hour	Wday	Duration
<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>
2016	02	08	00	2	360
2016	02	08	05	2	360
2016	02	08	06	2	360
2016	02	08	06	2	360
2016	02	08	07	2	360

Selecting Data to Analyze





- California had the most data points after cleaning the NA values

Goals & Conclusions

Project Goals

- Perform exploratory data analysis and try generate insights about traffic accidents in the US
- Predict Severity levels based on traffic accidents using various models.
- Use techniques learned during the course to visualize and model the data to answer questions

5 Questions To Answer

1. Is there a discernible pattern when accidents occur (day of the week, time)?
2. Can we use the time of day, weather, and or environmental stimuli to predict traffic accidents severity?
3. Which attributes contribute the most to accident occurrence?
4. Which major interstates in CA are more accident prone?
5. Can we group together the states based on Severity and the Number of accidents?

Main Conclusions

1. Is there a discernible pattern when accidents occur (day of the week, time)?
   → *Yes, weekdays around 6-9am and 3-6pm*

2. Can we use the time of day, weather, and or environmental stimuli to predict traffic accidents severity?
   → *Yes, Decision Trees, Naïve Bayes, Random Forest*

3. Which attributes contribute the most to accident occurrence?
   → *Duration and Precipitation are top two attributes*

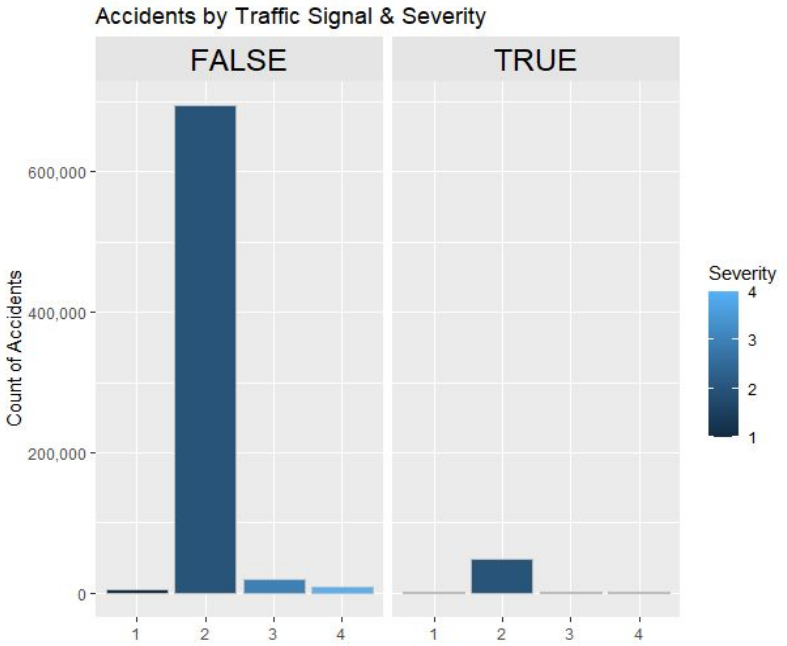
4. Which major interstates in CA are more accident prone?
   → *I-5N, I-10E, I-5S are top 3*

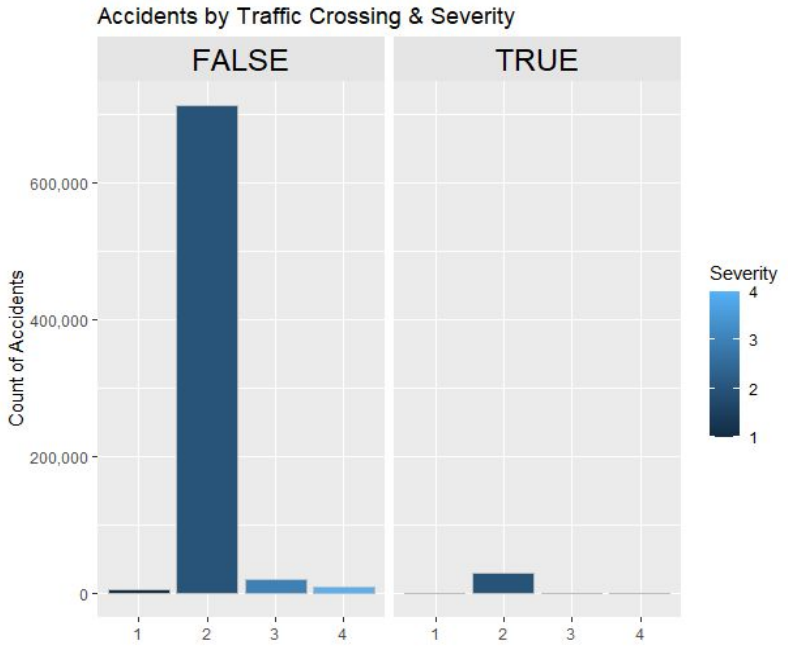
5. Can we group together the states based on Severity and the Number of accidents?

   → *Yes, using clustering we can group the states together*

Data Visualization

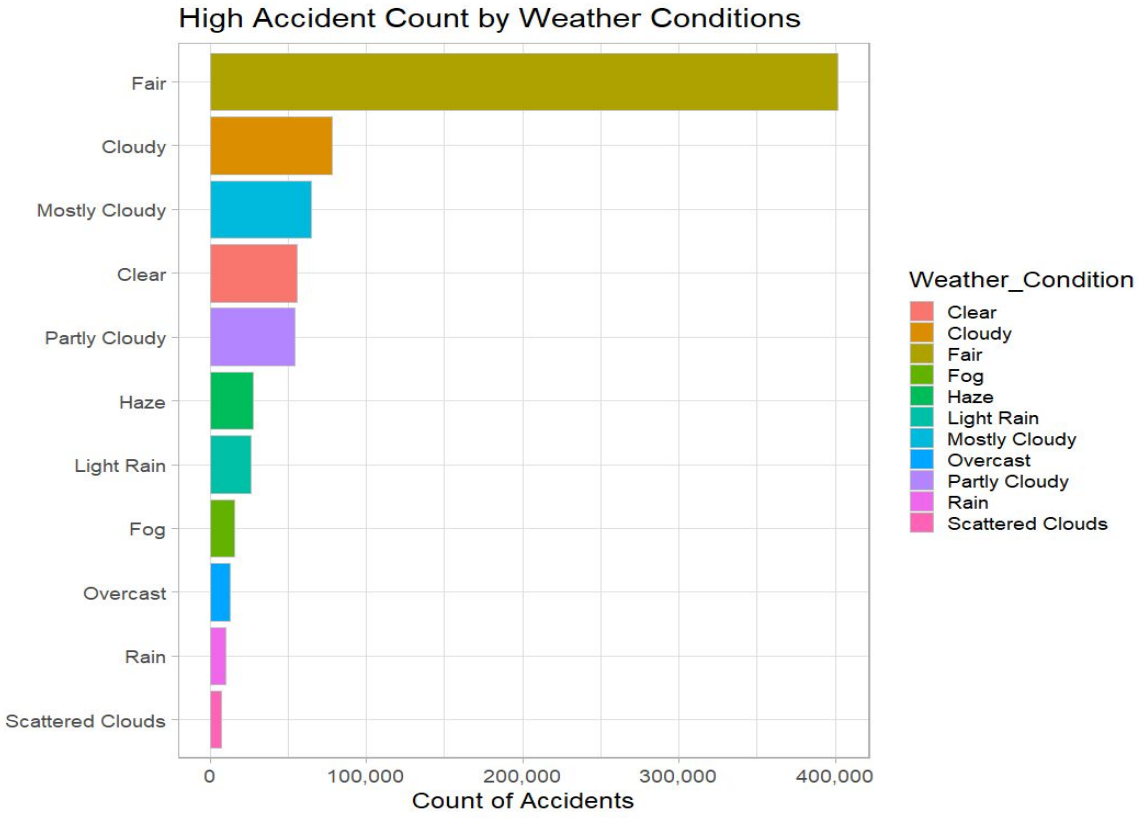
Accident by Traffic and Crossing Signal





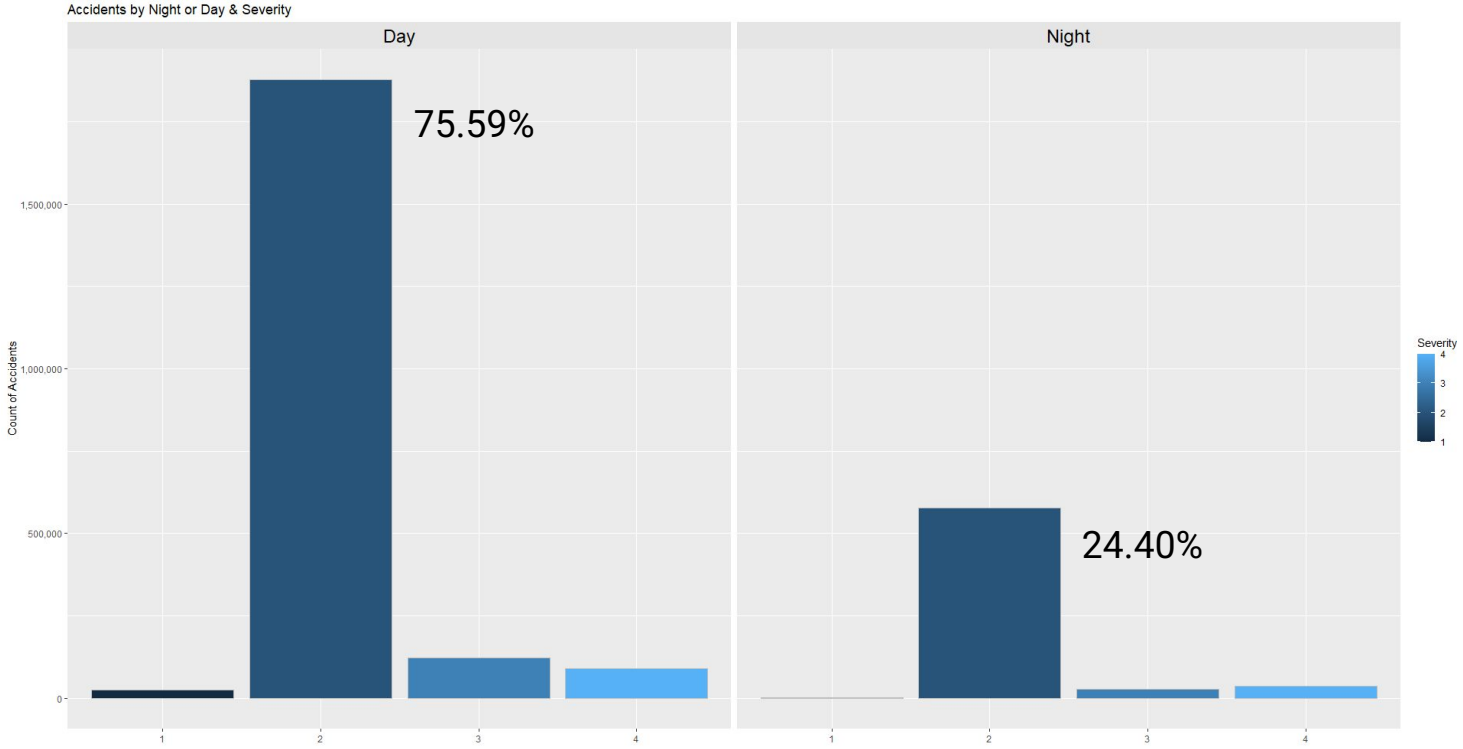
- Most accidents occur when no traffic signal or crossing was present

Accident by Weather Condition



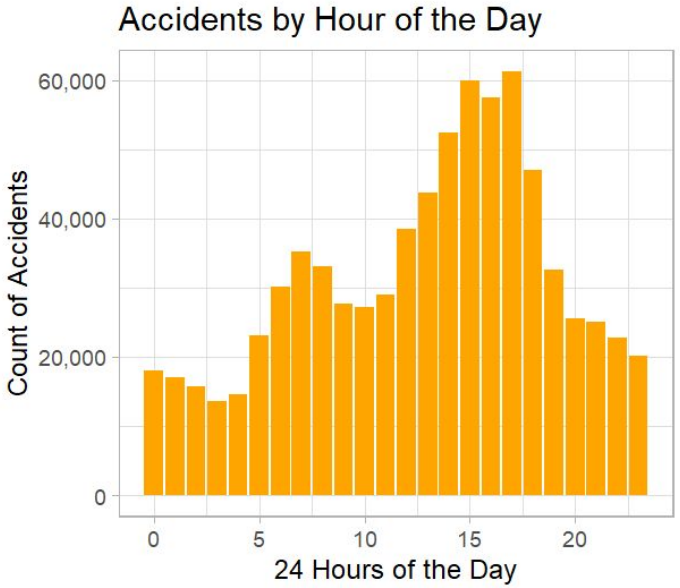
- Most accidents occur in fair weather, not bad weather conditions

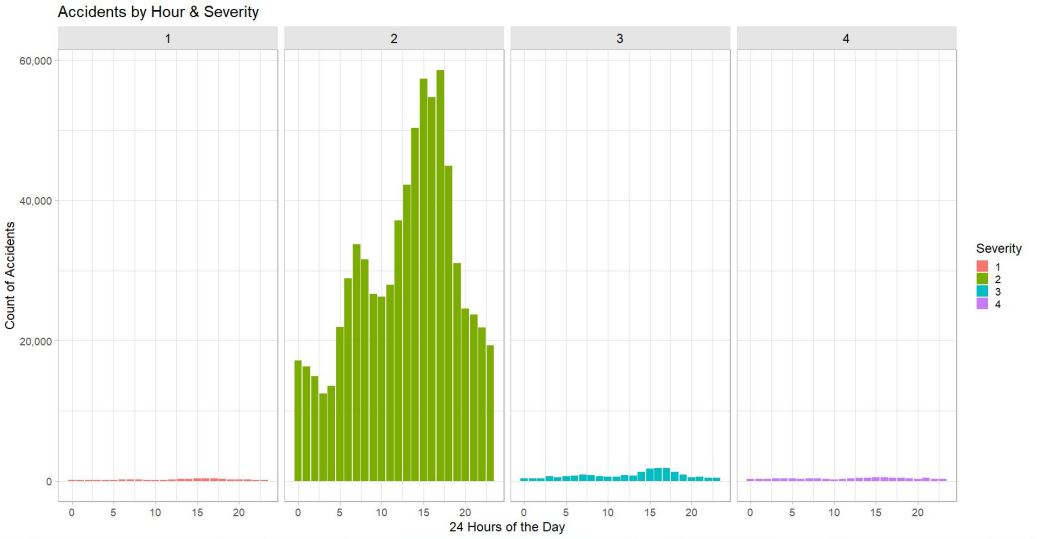
Accidents by Astronomical Light



- Most accidents occur in daytime conditions

Accident by Time of Day



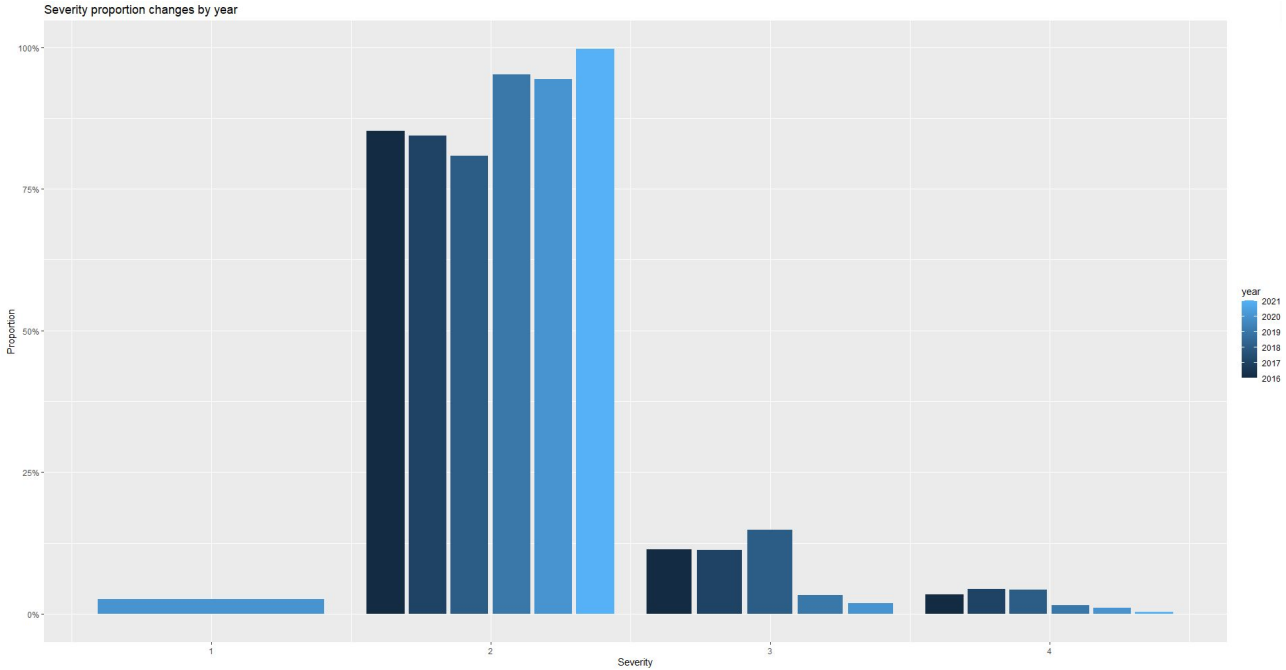


- Majority of Accidents occur between 6-9am and 3-6pm
- And the majority of accidents falls under medium severity

Accident Severity by Year

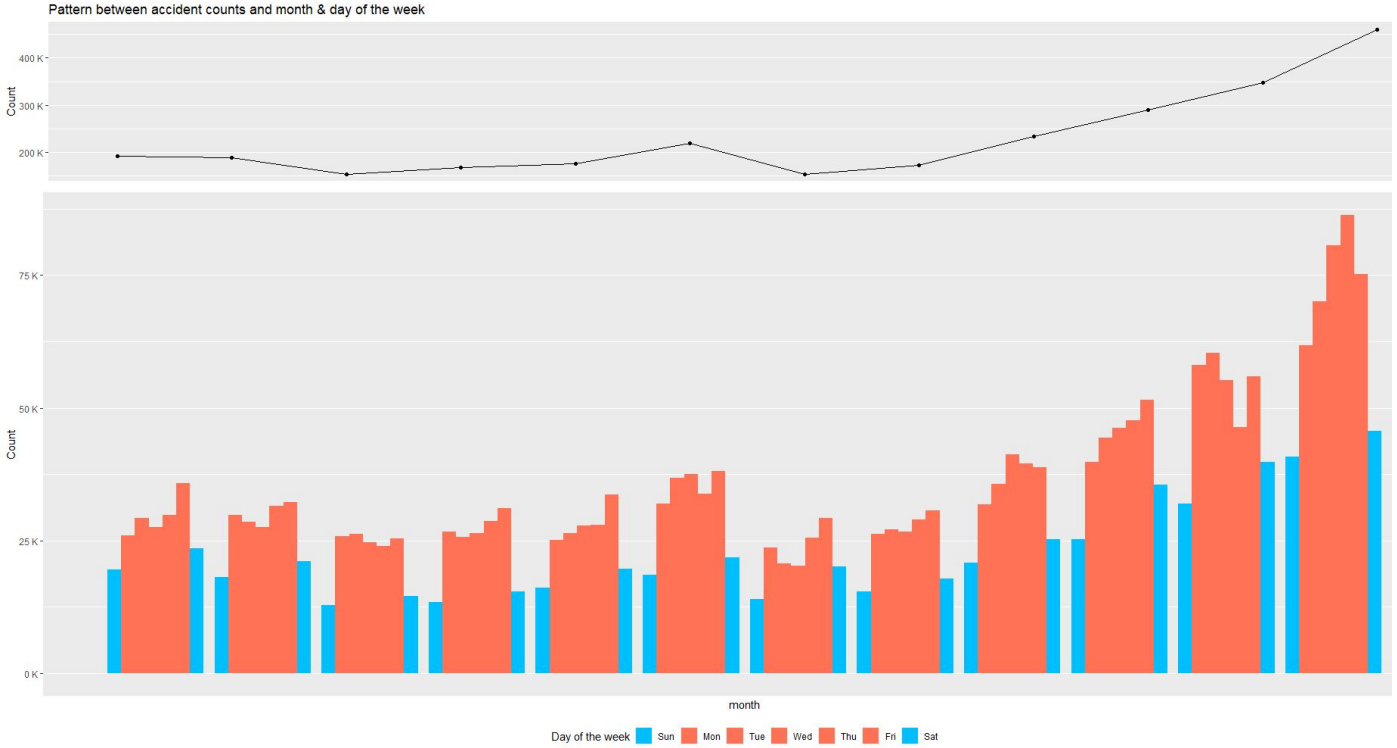
Total Number of Accidents

year	n
<dbl>	<int>
2016	33111
2017	34588
2018	33911
2019	103916
2020	188582
2021	376695



- More severity level 2 accidents from 2019 - 2021 and decrease in level 3 and 4 of severity

Accident Pattern by Month



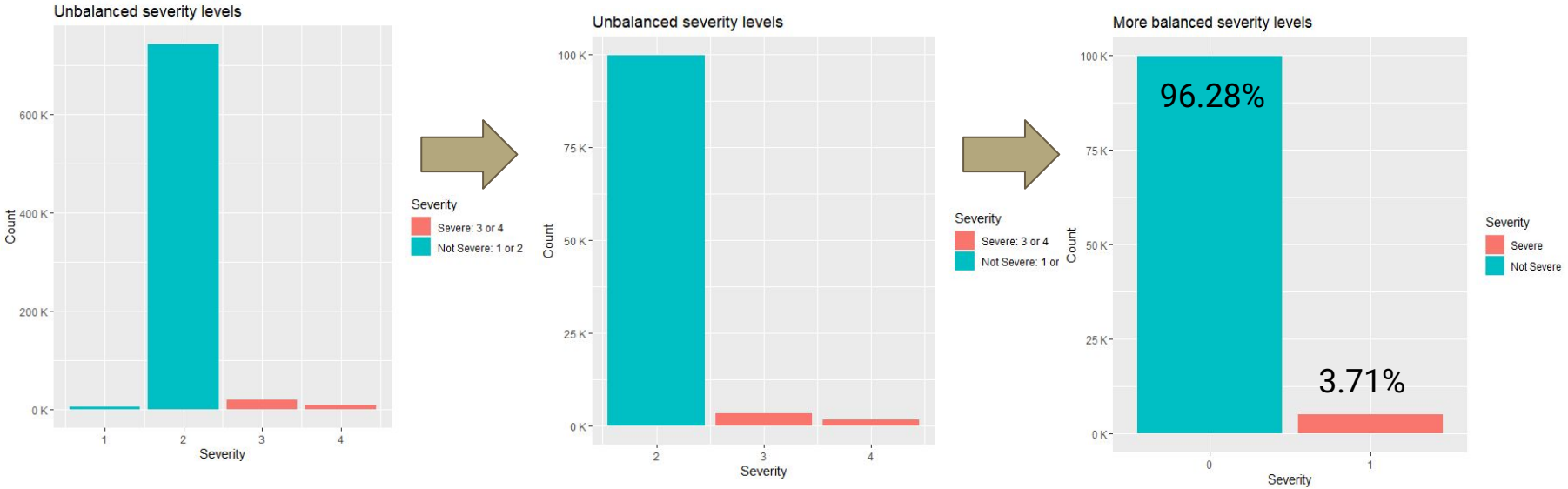
- More accidents on weekdays in every month. More accidents occured in the month of December

Models

- Supervised
  - ➔ Decision Tree
  - ➔ Random Forest
  - ➔ Naïve-Bayes
- Unsupervised
  - ➔ Clustering

Data Processing for Modeling

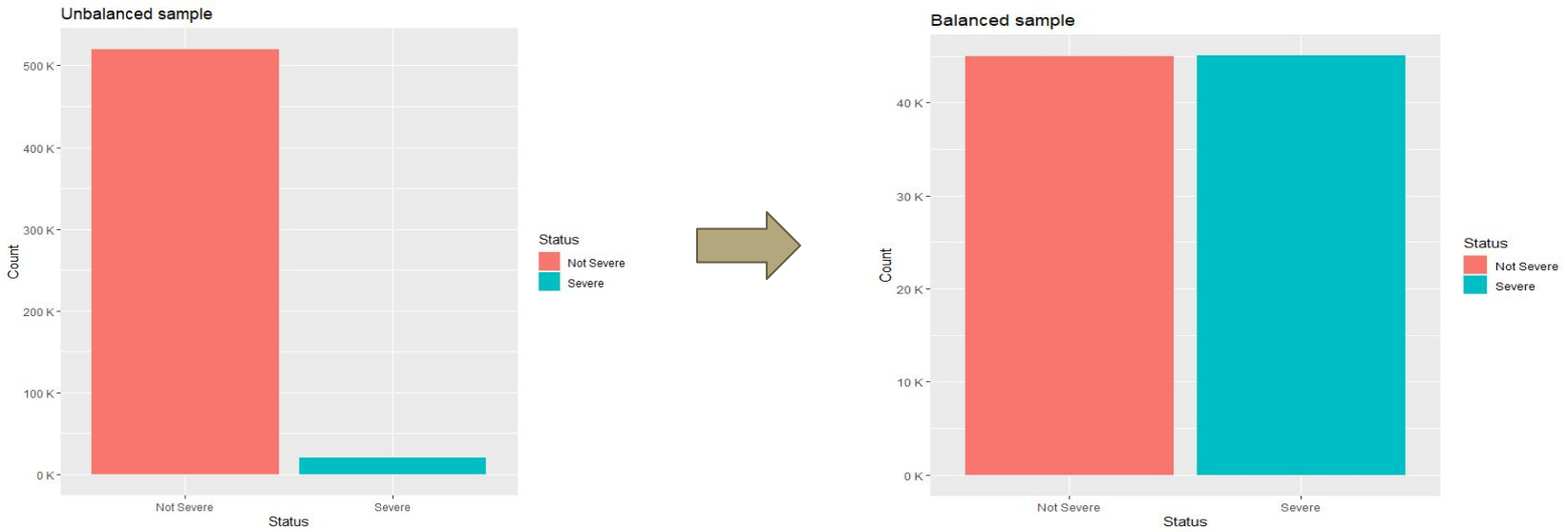
- Data set was unbalanced, needed to adjust the groupings for analysis



- Remove severity level 1
- Group severity levels 3 and 4 together

Data Processing for Modeling

- Balancing the data by random undersampling the majority data and over-sampling the minority data.



Decision Tree

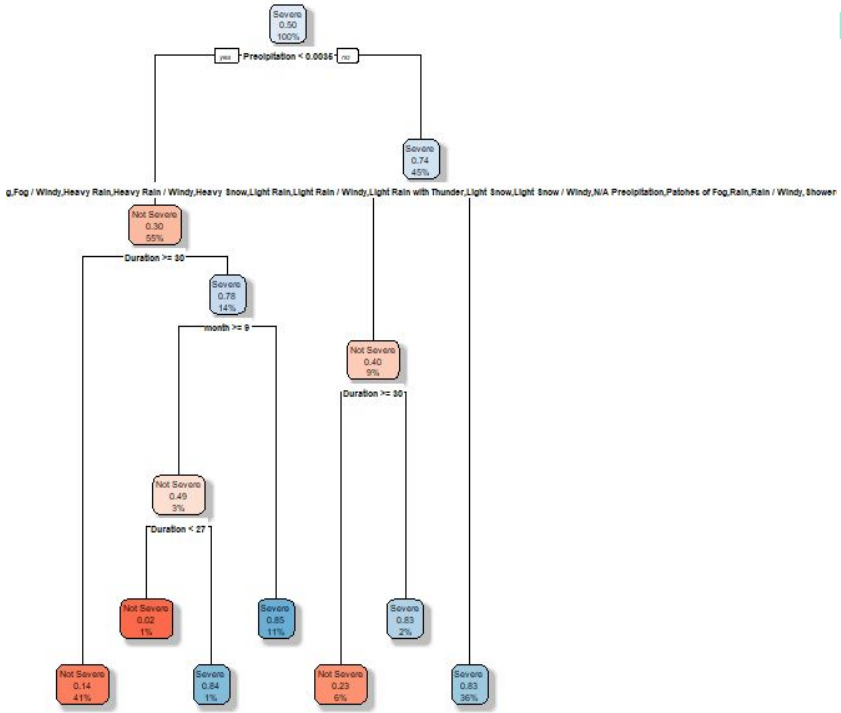
For Unbalanced sample

```
Confusion Matrix and Statistics


          0     1
  0 29701   827
  1   246   712

               Accuracy : 0.9659
                 95% CI : (0.9639, 0.9679)
    No Information Rate : 0.9511
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5535

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9918
            Specificity : 0.4626
         Pos Pred Value : 0.9729
         Neg Pred Value : 0.7432
             Prevalence : 0.9511
         Detection Rate : 0.9433
   Detection Prevalence : 0.9696
      Balanced Accuracy : 0.7272

       'Positive' Class : 0
```

For Balanced sample

```
Confusion Matrix and Statistics


              Not Severe Severe
  Not Severe     183862   1251
  Severe          38787   7341

               Accuracy : 0.8269
                 95% CI : (0.8253, 0.8284)
    No Information Rate : 0.9628
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2194

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8258
            Specificity : 0.8544
         Pos Pred Value : 0.9932
         Neg Pred Value : 0.1591
             Prevalence : 0.9628
         Detection Rate : 0.7951
   Detection Prevalence : 0.8005
      Balanced Accuracy : 0.8401

       'Positive' Class : Not Severe
```

Decision Tree



Important Features

	imp
Duration	14439.267864
Precipitation	9586.802080
Weather_Condition	8091.909952
Wind_Chill	4890.790573
Pressure	3404.656521
Wind_Speed	1902.468289
month	492.475751
Distance	113.694163
Humidity	15.424257
Temperature	2.059972

Random Forest

For Unbalanced data

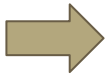
```
Confusion Matrix and Statistics

   predict_rf
       0     1
  0 29765   182
  1   684   855

               Accuracy : 0.9725
                 95% CI : (0.9706, 0.9743)
    No Information Rate : 0.9671
    P-Value [Acc > NIR] : 1.58e-08

                  Kappa : 0.65

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9775
            Specificity : 0.8245
         Pos Pred Value : 0.9939
         Neg Pred Value : 0.5556
             Prevalence : 0.9671
         Detection Rate : 0.9453
   Detection Prevalence : 0.9511
      Balanced Accuracy : 0.9010

       'Positive' Class : 0
```

For Balanced data

```
Confusion Matrix and Statistics

              predict_rf
               Not Severe Severe
  Not Severe       196644  26005
  Severe             1027   7565

               Accuracy : 0.8831
                 95% CI : (0.8818, 0.8844)
    No Information Rate : 0.8548
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3185

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9948
            Specificity : 0.2254
         Pos Pred Value : 0.8832
         Neg Pred Value : 0.8805
             Prevalence : 0.8548
         Detection Rate : 0.8504
   Detection Prevalence : 0.9628
      Balanced Accuracy : 0.6101

       'Positive' Class : Not Severe
```

Random Forest

Important features



	MeanDecreaseGini
Duration	15374.7693
Precipitation	5096.9097
Weather_Condition	4825.4374
Distance	3457.6909
month	2832.7630
Pressure	2657.8357
Wind_Chill	2436.0153
Humidity	2176.5820
Temperature	1906.3347
hr	1735.1334
Wind_Speed	1667.7490
Junction	353.6893
Astronomical_Twilight	280.1761
Traffic_Signal	151.1067

Naïve Bayes

For Unbalanced data

```
Confusion Matrix and Statistics

   predict_nb
       0     1
0  29947     0
1   1539     0

               Accuracy : 0.9511
                 95% CI : (0.9487, 0.9535)
    No Information Rate : 1
    P-Value [Acc > NIR] : 1

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

              Precision : 1.0000
                 Recall : 0.9511
                     F1 : 0.9749
             Prevalence : 1.0000
         Detection Rate : 0.9511
   Detection Prevalence : 0.9511
      Balanced Accuracy :     NA

       'Positive' Class : 0
```

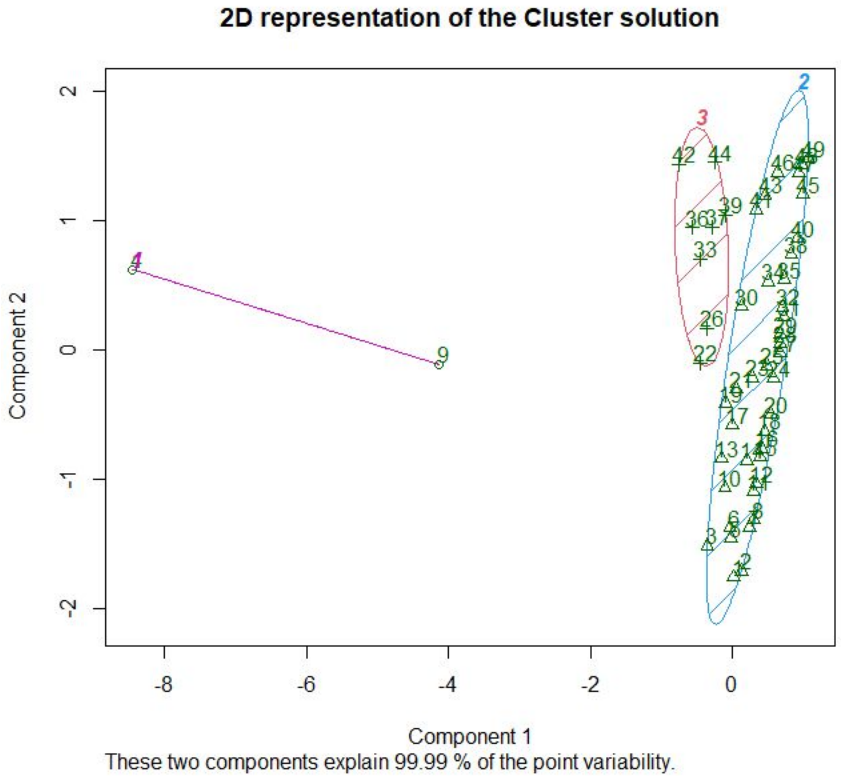
Comparison of Models

- Duration and Precipitation are the two most important attributes for the Decision Tree and Random Forest

Model	Accuracy	Sensitivity	Specificity	Precision	*F*-Measure	Recall
Decision Tree (Unbalanced)	96.59%	99.18%	46.26%	97.29%	98.22%	99.18%
Decision Tree (Balanced)	82.69%	82.58%	85.44%	99.32%	90.18%	82.58%
Random Forest (Unbalanced)	97.25%	97.75%	82.45%	97.75%	98.57%	99.39%
Random Forest (Balanced)	88.31%	99.48%	22.54%	88.31%	93.57%	99.48%
Naïve Bayes (Unbalanced)	95.11%	100.00%	0.00%	100.00%	97.49%	95.11%

Clustering

- K-means clustering
- Clusters based on Severity Index and Total number of Accidents
- CA(4) and FL(9) highest accidents same cluster



Clustering

- Normalized severity Index and total number of Accidents (0 to 1)
- Maximum CA(4) therefore 1
- Followed by FL(9)

	resstate	sevindex	restotacc
1	AL	0.0253437425	0.0243652768
2	AR	0.0147363593	0.0134112815
3	AZ	0.0711824284	0.0704024163
4	CA	1.0000000000	1.0000000000
5	CO	0.0413177145	0.0317423249
6	CT	0.0421811209	0.0380529056
7	DC	0.0117484266	0.0110422189
8	DE	0.0073735799	0.0060108800
9	FL	0.5194631759	0.5127411715
10	GA	0.0624543360	0.0506234683
11	IA	0.0135273361	0.0117583868
12	ID	0.0104917541	0.0101366298
13	IL	0.0753450767	0.0604007167
14	IN	0.0316061393	0.0262530019
15	KS	0.0101391488	0.0094866296
16	KY	0.0093805710	0.0081918188
17	LA	0.0597796185	0.0594821535
18	MA	0.0089320317	0.0077169684
19	MD	0.0781506711	0.0689259687
20	ME	0.0025825954	0.0025468071

Word Cloud

- Text Mining analysis on Description attribute to transform unstructured data in to suitable structured data for deriving insights.

	Var1	Freq
1	I-5 N	20748
2	I-10 E	16420
3	I-5 S	16200
4	I-10 W	15735
5	I-405 N	11773
6	I-80 W	9187
7	I-80 E	8662
8	I-405 S	7687
9	I-880 N	5847
10	I-880 S	5702
11	I-580 W	5470
12	I-580 E	5382
13	I-605 S	5061

Conclusions of Analysis

Analysis Conclusions

- Even after various safety measures implemented by the government, the number of accidents has increased over the years.
- The popular belief is that the most accidents happens during bad weather like rain and fog, but through the analysis it appears most accidents happened during fair weather
- Most accidents happen on the weekdays an hour before and after the typical workday hours. The government could adopt various measure to reduce the traffic flow at that time by promoting public transportation
- I-5 N freeway in California has the highest frequency of the accidents and California is the state with highest number of accidents

Data-Driven Solutions: Proposed Safety Policies

- Utilize data on crash patterns and road design to adjust speed limits dynamically based on real-time traffic conditions and safety considerations.
- Replacing high-collision intersections with roundabouts, which have been shown to reduce the frequency and severity of crashes.
- Allocate resources for targeted infrastructure improvements like better signage, guardrails, and wider shoulders at high-risk locations.

Recommendations for Future Work

Further Improvements

- One-hot encoding can be tried for some of the features - which could be used in clustering e.g determining which state fall in severity levels.
- Weighted XGBoost or other similar models can be used instead of resampling the dataset since the dataset is very imbalanced.
- Having better text descriptions could give better analysis and understandings
- Implement findings with companies like uber or lyft to to make suggestions for better routing

References

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, arXiv preprint arXiv:1906.05409 (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.