Final Project

We are trying to find out whether lighting conditions vs. road conditions, alone or in conjunction with any various small groupings of other factors, is a better predictor of crash severity with and without considering serious injury in the accidents. The data source used is a City of Chicago dataset focused on vehicles (not people, which is a distinct dataset): https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if

1) Data Pre-processing

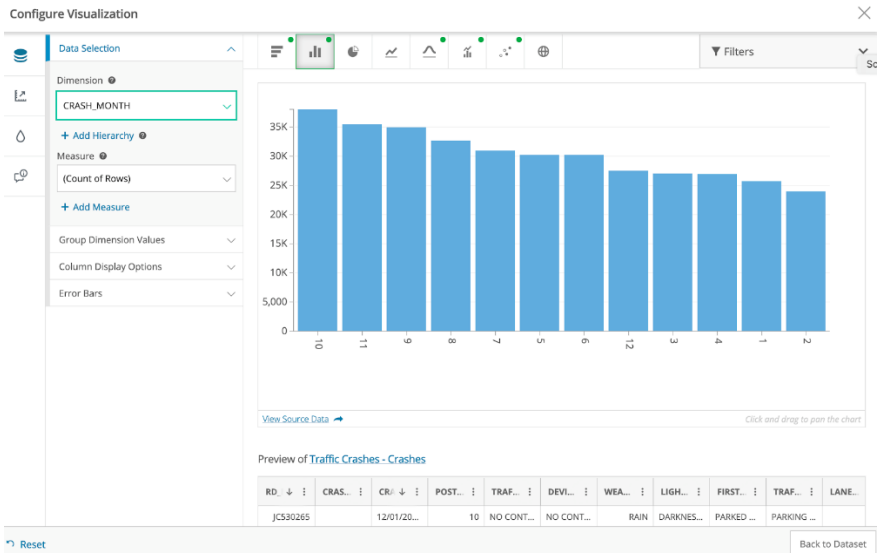
The original data source contains 48 variables mostly categorical and 363,249 observations. To facilitate the process of training models, we selected the recent 50,000 records by sorting on CRASH_DATE. Also, since our focus was mainly on two variables with combinations of other predictors hence, variables which has missing values were dropped. New featured were created by extracting or combining information from one or more than one variable. For example, by using the data in CRASH_DATE and CRASH_MONTH columns, a new feature CRASH_SEASON was generated.

Similarly, the data in CRASH_DAY_OF_WEEK was ordered numerical categorical data. This was changed to string categorical data one category is not greater or smaller than the other one. Hence, 1 was converted to Sunday and so on. Some of the tasks performed

2) Exploratory Data analysis

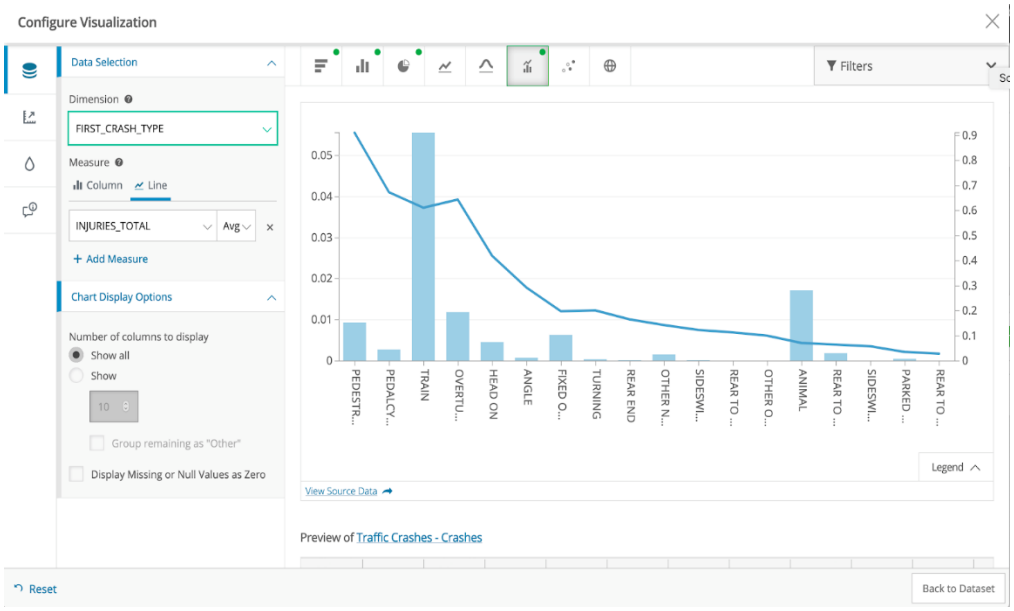
- General Exploratory Findings

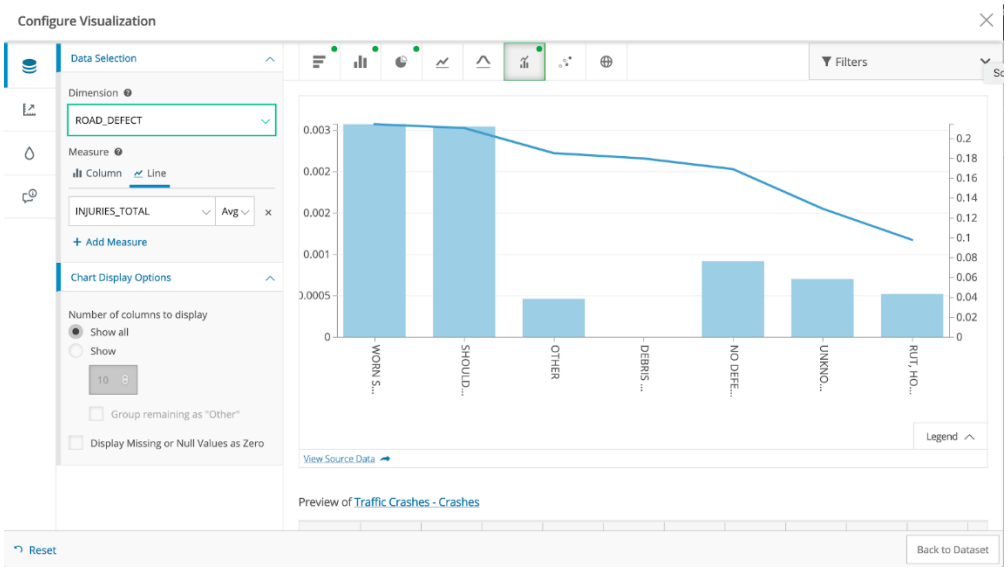
  Below are some of the general exploratory analysis done without any data pre-processing:

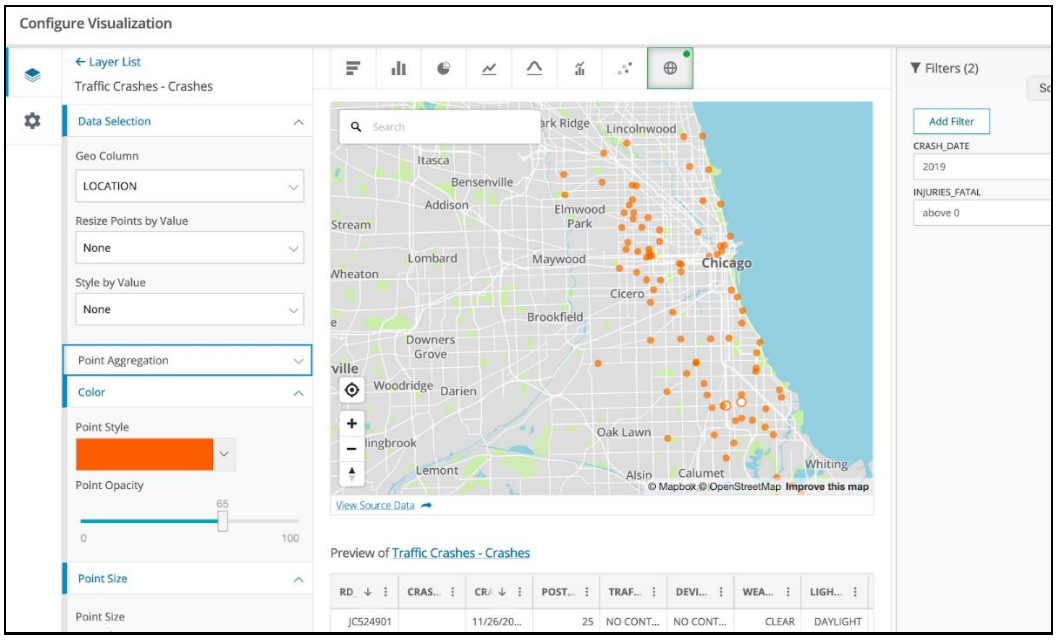


More crashes are observed later in the year.

(This is done with the entire dataset on the website visualizer)


Configure Visualization
Data Selection
Dimension
FIRST_CRASH_TYPE
Measure
Column
Line
INJURIES_TOTAL
Avg
+ Add Measure
Chart Display Options
Number of columns to display
Show all
Show
10
Group remaining as "Other"
Display Missing or Null Values as Zero
Filters
0.05
0.04
0.03
0.02
0.01
0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
PEDESTR...
PEDALCY...
TRAIN
OVERTU...
HEAD ON
ANGLE
FIXED O...
TURNING
REAR END
OTHER N...
SIDESWI...
REAR TO ...
OTHER O...
ANIMAL
REAR TO ...
SIDESWI...
PARKED ...
REAR TO ...
Legend
View Source Data
Preview of Traffic Crashes - Crashes
Reset
Back to Dataset


Configure Visualization
Data Selection
Dimension
ROAD_DEFECT
Measure
Column
Line
INJURIES_TOTAL
Avg
+ Add Measure
Chart Display Options
Number of columns to display
Show all
Show
10
Group remaining as "Other"
Display Missing or Null Values as Zero
Filters
0.003
0.002
0.002
0.001
.0005
0
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
WORN S...
SHOULD...
OTHER
DEBRIS ...
NO DEFE...
UNKNO...
RUT, HO...
Legend
View Source Data
Preview of Traffic Crashes - Crashes
Reset
Back to Dataset



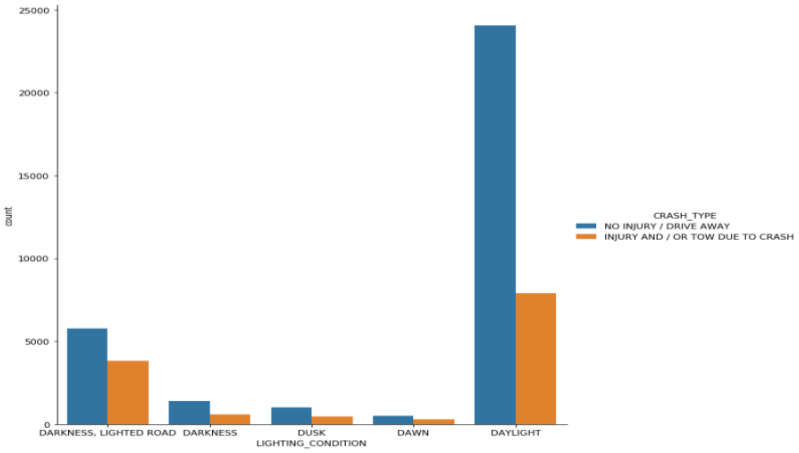
2019 fatalities – there are no definite patterns discernible, other than they don't seem to be on the highways, so it could be that the data are gathered by state police or allotted to state/federal metrics, being interstate roads. This would also explain why 30/35 mph speed zones seemed to be the average for both fatalities and injuries in general, no real uptick in fatality for high speed zone.

Analysis done from Bar/Line graphs:

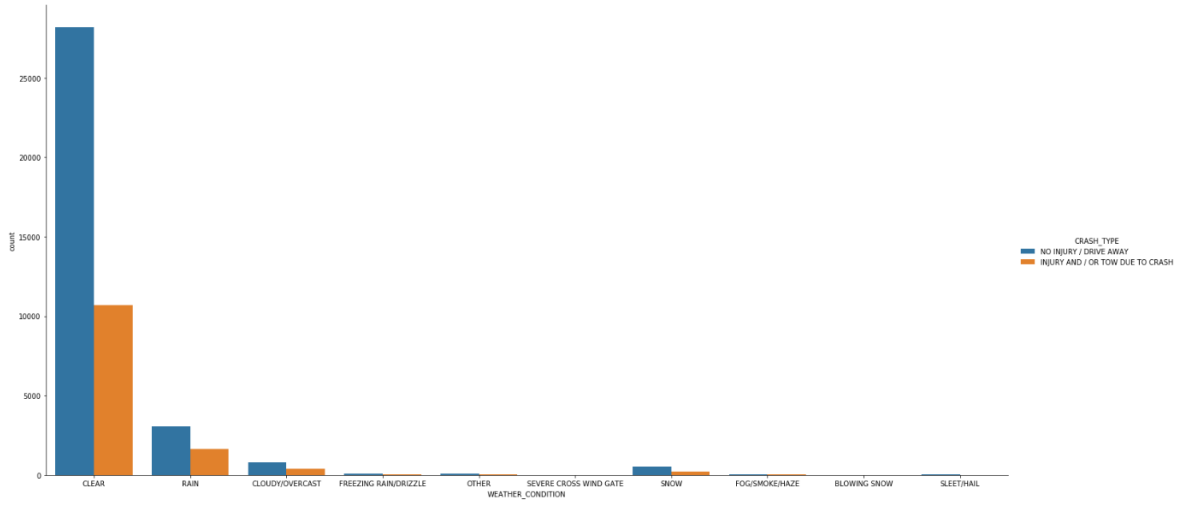
X axis qty	X axis value	Line graph	Remarks
Average fatalities	Weather conditions	Average total injuries	Rain associated with small spike in fatality average
Average fatalities	Alignment	Average total injuries	Curve on hillcrest spike fatalities (probably meaning-less, small n)
Average fatalities	Beat_of_occurence	Average total injuries	5 of highest fatality rates beats are in Englewood. Poor response/medical access?
Average fatalities	crash_Month	Average total injuries	NOT as much an effect on injuries as on number crashes (apparent)
Average fatalities	first_crash_type	Average total injuries	Train and animal spike in fatality
Average fatalities	not_right_of_way	Average total injuries	Crash in the public right of way twice as likely to be a fatality
Average fatalities	num_units	Average total injuries	More units, more fatality, especially 7!
Average fatalities	posted_speed_limit	Average total injuries	This makes more sense, faster means more injury. Es-pecially from 25 -> 45 increase 200%
Average fatalities	prim_contribu-tory_cause	Average total injuries	Driver's condition and exceeding speed limit results to fatalities.

Average fatalities	road_defect	Average total injuries	Worn surface and shoulder defect stand out ~ fatalities
Average fatalities	sec_contributing_cause	Average total injuries	Seven of these are relevant, but they're just descriptions of what happened. Don't plan to use.
Average fatalities	workers_present_i	Average total injuries	Present doubles rate of fatality, over no entry though, so murky
Average fatalities	work_zone	Average total injuries	And no work zone at all is worse. So i will ignore this one.
Average fatalities	work_zone_type	Average total injuries	similar to "present", construction type zone triples fatality over no response

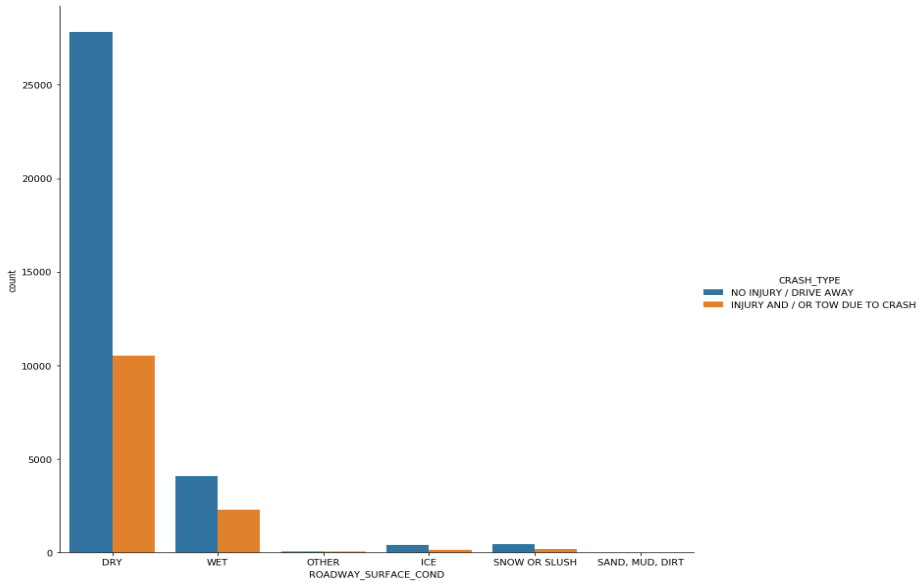
- Exploratory Findings After data pre-processing



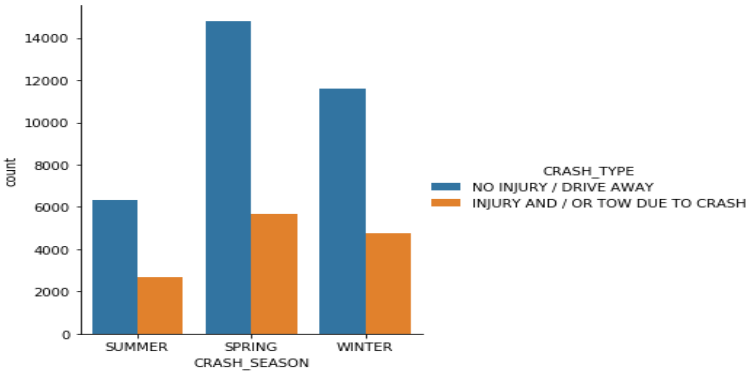
Maximum accidents seem to be happening during daylight which is quite understandable since it's the office hours. Next category is during the late evening, night and early morning i.e. office closing time, people are tired, and traffic.



Most people prefer to drive when the weather condition is conducive, hence the maximum accidents. Rain and snow make the roads slippery thus making more accidental prone.

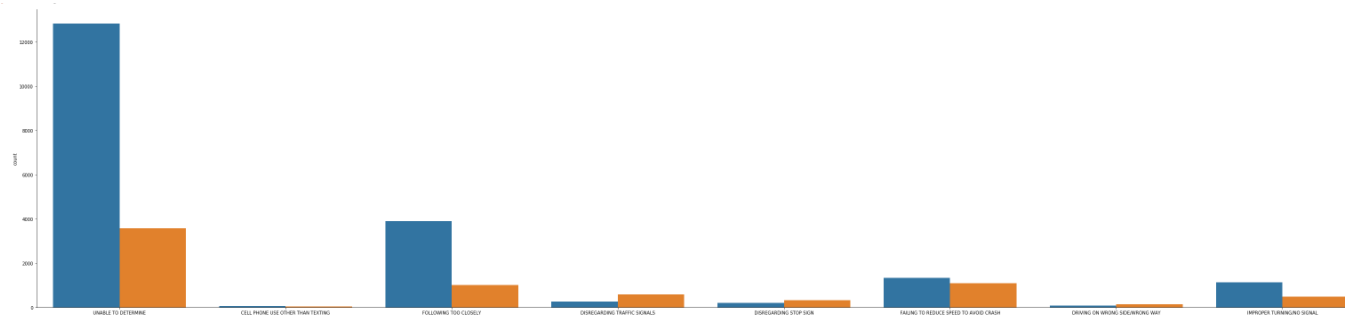


Same as WEATHER_CONDITION variable

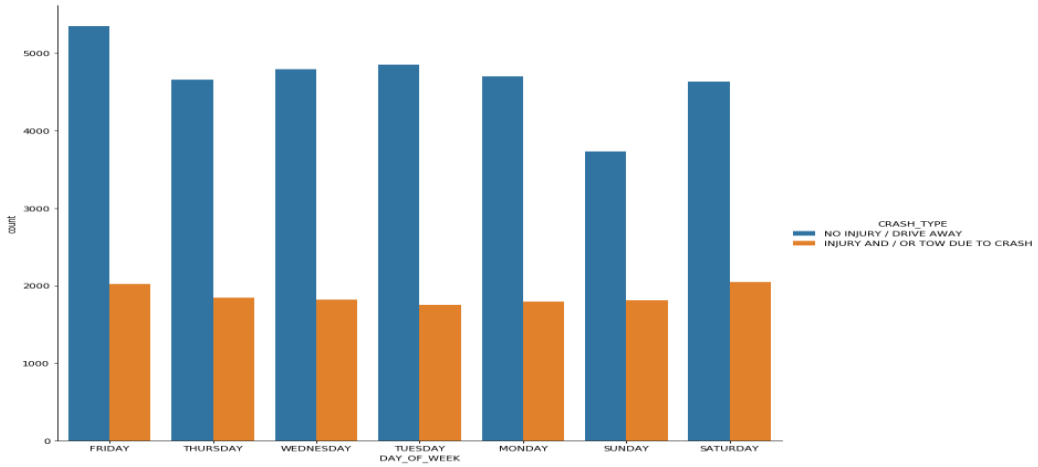


Spring and Winter seasons observe maximum accidents. Cars skid in slippery roads hence, more accidents in winters. Roads are more crowded in Spring thus increasing accidents.

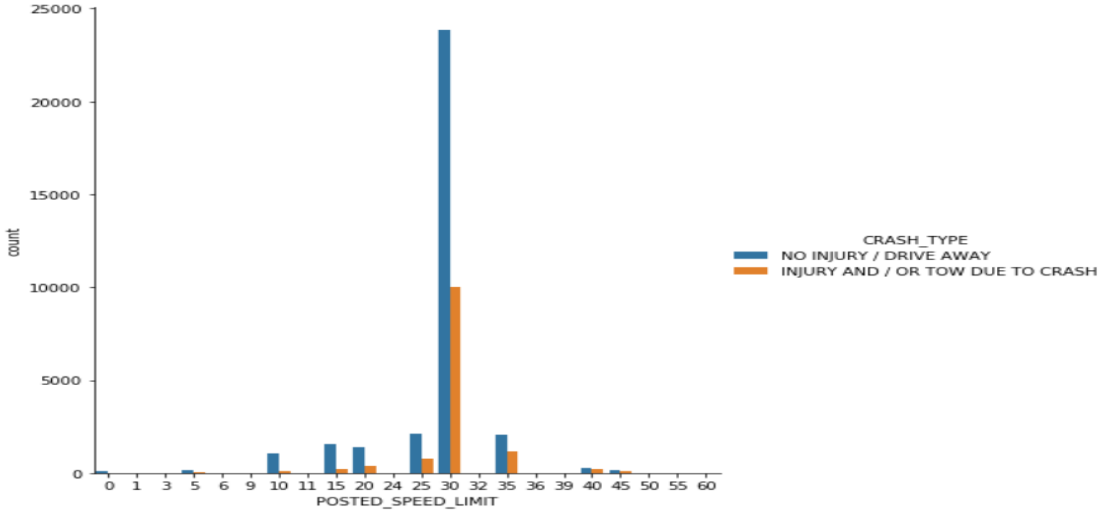
Plotted PRIM_CONTRIBUTORY_CAUSE as x-axis parameter, but plot is not visible in the report properly due to large number of categories.



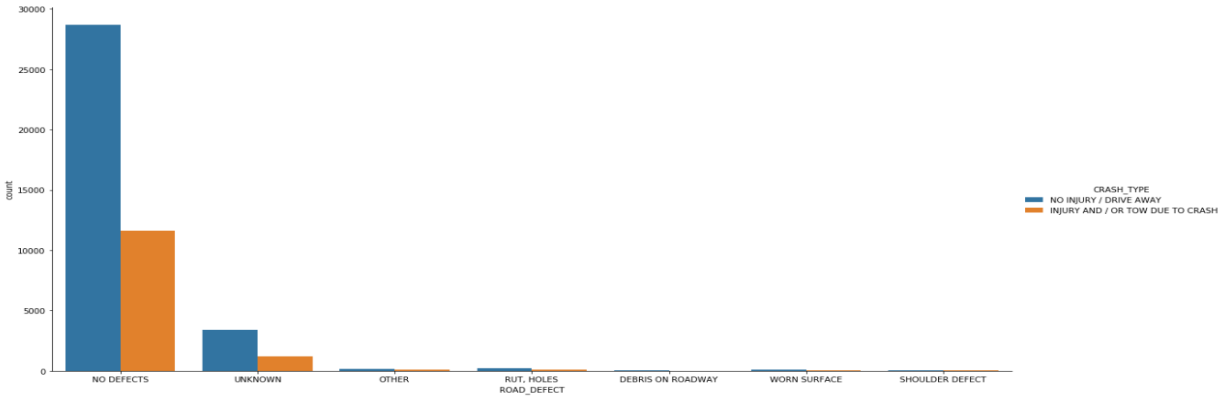
The primary reason for crash seems to be "Following too closely" something very common in heavy traffic.



Friday and Saturday observe the maximum number of accidents. This could be due to the rush observed on roads during office closing time.



As the data collected could be from the interstate roads hence, more accidents are observed in the speed zone of 30 mph.



Vehicles tends to speed up in smooth roads which increases the probability of car crashes.

Finally, Correlation matrix is used to check correlations between numerical predictors and response variable. For categorical variables, the predictors were first encoded using OrdinalEncoder and the response variable with LabelEncoder. Then SelectKBest with Chi-squared is used for categorical feature selection. Finally dropped observations with 'UNKNOWN' as the category under several variables.

For Classification, the response variable is CRASH_TYPE. We have not considered injuries involved for simplification. The variables used are:

- WEATHER_CONDITION
- LIGHTING_CONDITION
- FIRST_CRASH_TYPE
- TRAFFICWAY_TYPE
- ALIGNMENT
- ROADWAY_SURFACE_COND
- ROAD_DEFECT
- PRIM_CONTRIBUTORY_CAUSE
- SEC_CONTRIBUTORY_CAUSE
- DAY_OF_WEEK
- CRASH_SEASON

- For Clustering, new features are generated based on the exploratory analysis. For example, **Speed35** and **anInjury** are newly generated features to name a few**.** The value of Speed35 variable is 1 if POSTED_SPEED_LIMIT is greater than or equal to 35. Similarly, if INJURIES_FATAL is greater than or equal to 1 then, aFatal = 1.

Following are the of newly generated features (based on the analysis from bar/line graphs):

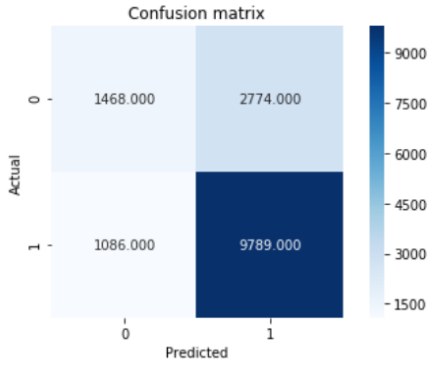
- Speed35
- AnimalTrain
- Defect
- RightOfWay
- Construction
- Workers
- 7Units
- aFatal
- anInjury

3) Data Modeling

- **Classification**

  The two classification techniques used in this project are Decision Tree and K-Nearest Neighbor.
  The best model is obtained by K-Nearest Neighbor with n_neighbors = 7. F1-score for majority class is 0.835 and minority class is 0.432. Since, this dataset is imbalanced and hence, macro and weighted F1-score is also calculated. The weighted F-1 score for this model is 0.72. Since this is highly imbalanced dataset, hence best model is selected considering both F1-score and weighted F-1 score.
  n_neighbors = 3,4,5,6,8 were also evaluated.
  Similarly, Decision Trees with random_state = 0,1,2 were experimented.

```
[0.7446583316795661, 0.2553416683204339, array([0.57478465, 0.77919287]), array([0.34606318, 0.90013793]), array([0.43201883, 0.83
531018])]
F1 score (average: macro) = 0.6336645073290428
F1 score (average: weighted) = 0.7221420985941576
```



- **Clustering**

  The two classification techniques are used in this project are DBSCAN, K-Means.

  Using the newly engineered features with different parametric values DBSCAN performed the best. The response variable in this model is **anInjury**. The model performed best with the following parameters: eps = 2, min_samples = 5, metric = "euclidean".

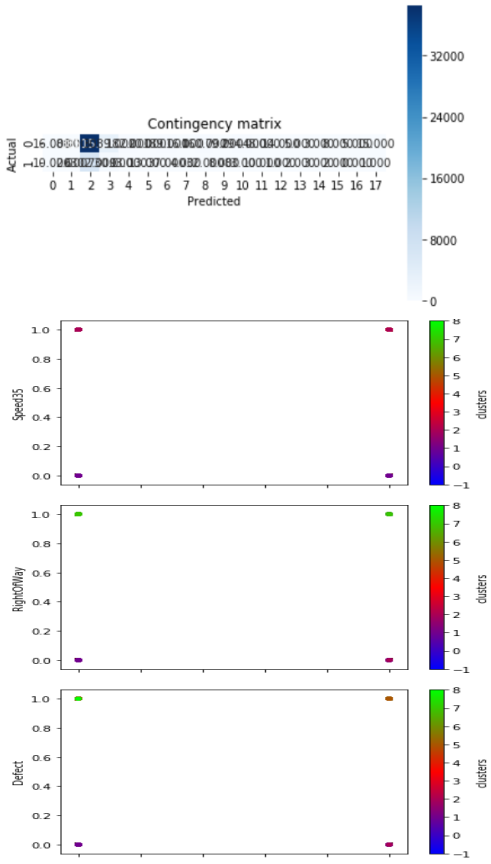
  The Adjusted Rand index came out to be 0.038 and Silhouette coefficient is 0.99.

  Keeping in mind both the supervised and unsupervised metric, DBSCAN with the parameters performed the best.

  The variables used are: 'Defect','RightOfWay','SpeedingOrDriverCond','7Units','Speed35'.

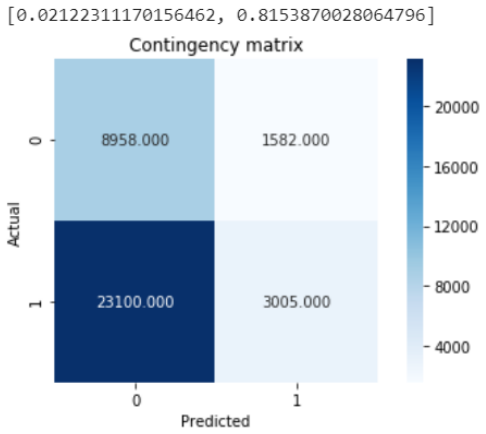
  Using these variables, we are getting cohesive and well-separated clusters.

```
[0.038024676054597674, 0.9990875535575344]
```



The above result was obtained using completely new features.

Separate experimentation was done by using existing categorical and numeric variables. K-mean clustering technique was u sed. The first step was to encode the categorical variables. Then combination of only encoded categorical variables with nu meric variables was fed into the models. The best result was obtained by using the two variables: 'WEATHER_CONDITION', ' LIGHTING_CONDITION' and parameters: n_clusters = 2, init = 'k-means++', n_init = 10, random_state=1. Adj Rand Index = 0. 02 and Silhouette coefficient= 0.815. The Silhouette coefficient's value signifies that the clusters are cohesive and well-sepa rated. Tried with various combinations of variables and parameters but the same results were same.



Also, tried DBSCAN but it kept giving an error on the cluster labels: Number of labels is 1. Valid values are 2 to n_samples - 1 (inclusive)
. It might be since there are no distinct cluster formation using DBSCAN.

4) Conclusion

Classification:

When response variable is CRASH_TYPE, KNN algorithm is the best. Data might be non-linearly separated and hence KNN produces decision boundaries of any shape. Also, the variables that were chosen (as mentioned above) has a ma- jor impact on road accidents.

Clustering:

For newly engineered features DBSCAN gave the best result. DBSCAN works well for non-spherical dense data. This could be another reason why KNN worked well.

Whereas K-MEANS provided good Silhouette coefficient when original variables were used. Thus, we can conclude that Lightening and weather conditions does play a major role when it comes to road accidents.

5) Challenges faced

Most of the variables are categorical and each variable contains at least 10 categories. Creating dummy variables in- creased the dimension of the dataset exponentially. This led to increase in time to train models. Hence, experimenta- tion was done using only a limited number of combinations of variables and parameters.
When calculating Silhouette coefficient during clustering, MemoryError was being thrown. The experiment was re- run by decreasing the observations and variables but in vain.
Note: MemoryError was due to some internal system issue. Restarted the laptop and the issue got resolved.