Road Accident Prediction and Severity Analysis

TEAM MEMBERS : PRAMIL PAUDEL, SUMIT BHATTARAI

Road Accidents…

Background and Facts :

- There were 33,654 fatal motor vehicle crashes in the USA in 2018
- In 2018, 36,560 deaths occurred in the M/V crashes in the USA
- 11.2 deaths/100K people and 1.13 deaths/100k Miles of travel

Source : National Highway Traffic Safety Administration

Road Accidents…

Data Source :

- This is a countrywide car accident dataset, which covers 49 states of the USA.
- The accident data are collected from February 2016 to June 2020.
- There were **3513617 (3 Million)** unique data.
- USA FIPS (Federal Information Processing Standards)data was used for county wise plotting.
- Link to data : https://www.kaggle.com/**sobhanmoosavi**/us-accidents

Note : Data is available for research/academic purpose.

Road Accidents…

Project's Purpose :

- Data Visualization
- Severity Analysis(EDA)
- Classification
- Time Series Forecasting

Road Accidents…

Project's Purpose :

VISUALIZATION

PRAMIL PAUDEL, SUMIT BHATTARAI

Road Accidents…

Data Visualization:

- Data are visualized using different properties like State, City, Day, Hours by aggregating their count.
- Data are visualized using day, month, and year.
- Accident count are plotted against City and Time too

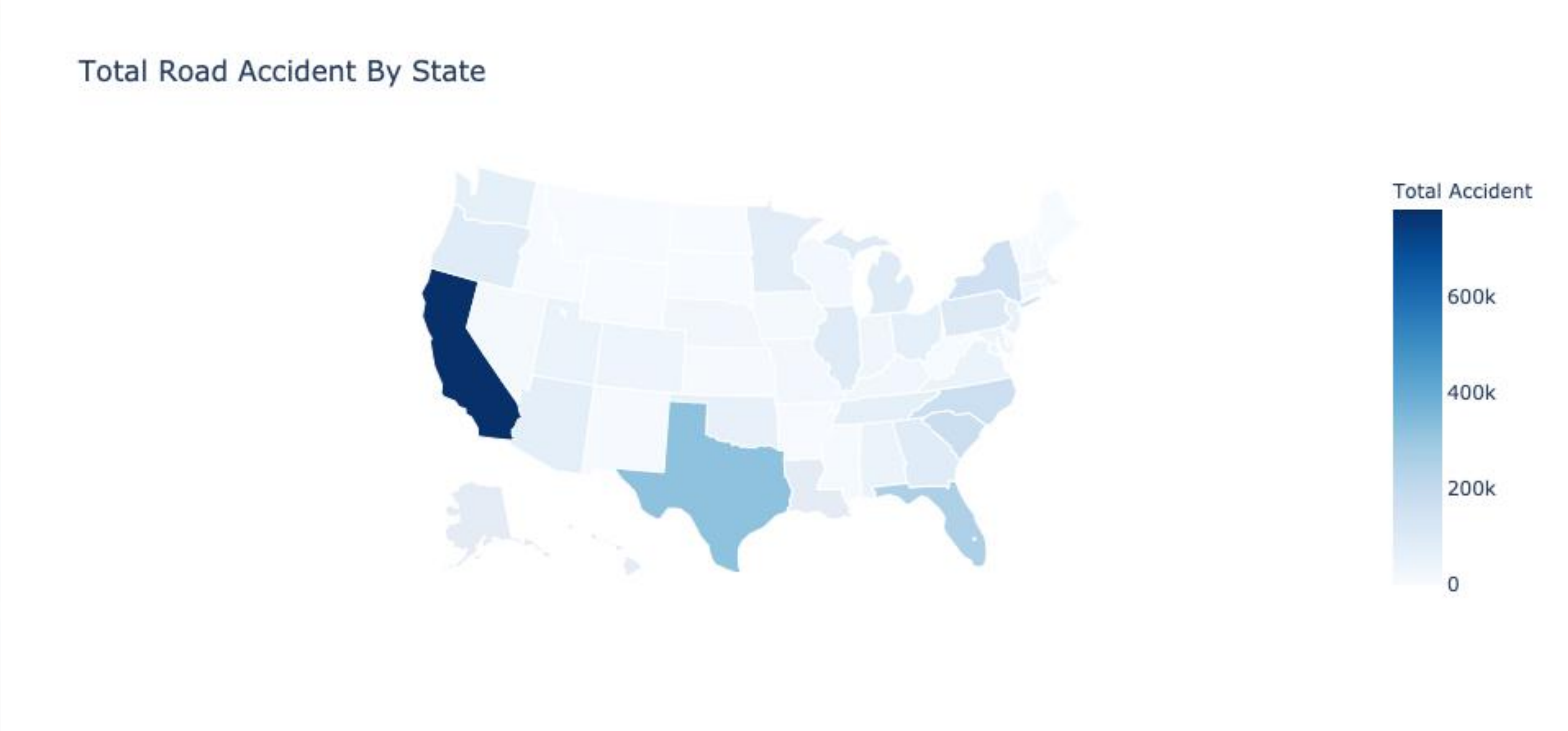
Road Accidents…

Data Visualization: Before Plotting in USA Map

- There were many cases having ***State*** and ***Zipcodes*** values N/A - dropped.
- For better visualization source data was merged with USA geo-graphical information to get FIPS (*Federal Information Processing Standards*)code
- Some attributed like ***Precipitation, Wind_Chill(F)*** containing many NA values were dropped to make data light weighted for easy processing
- These processed were applied only when they were applicable.

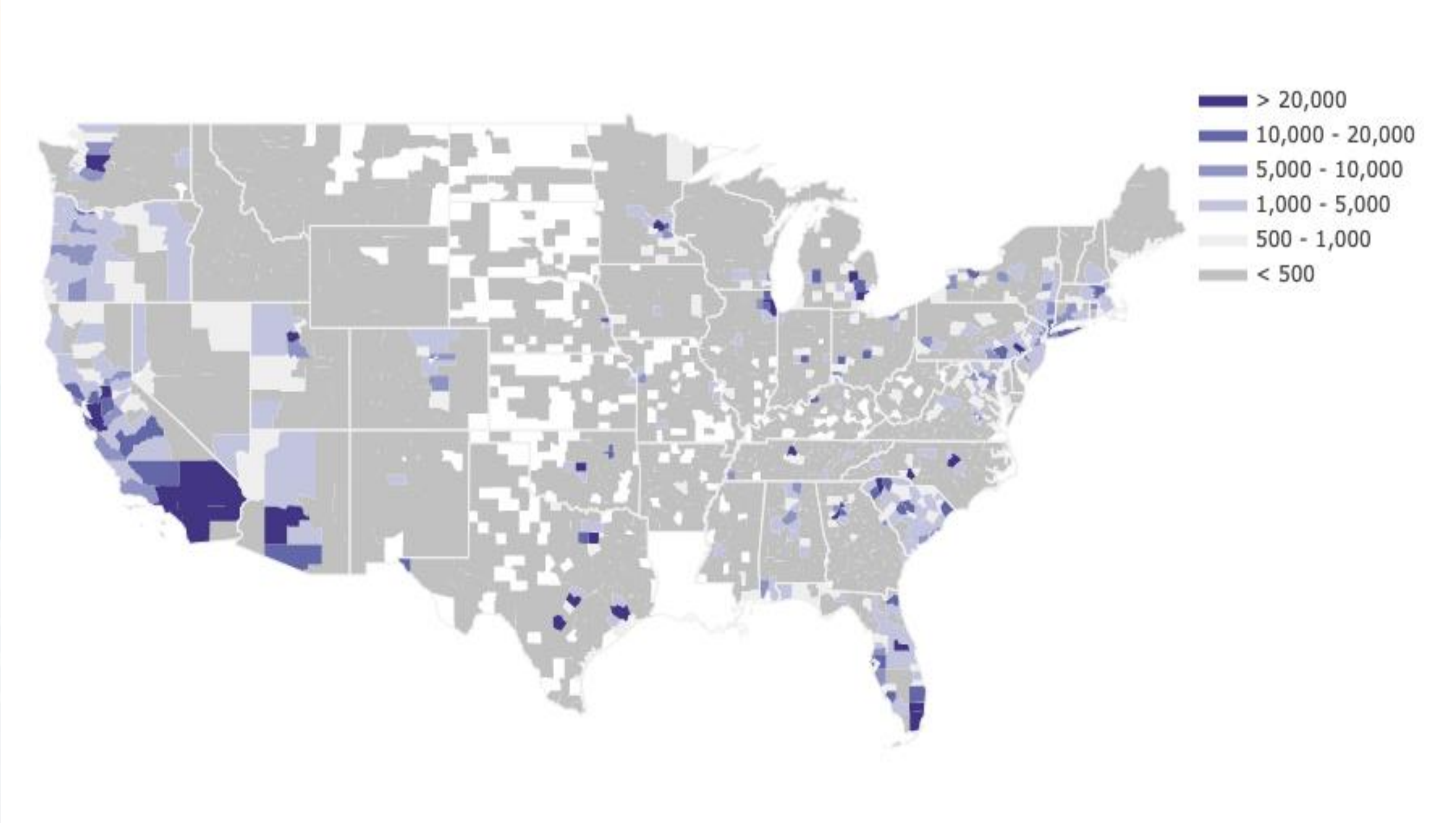
Road Accidents…

Data Visualization:



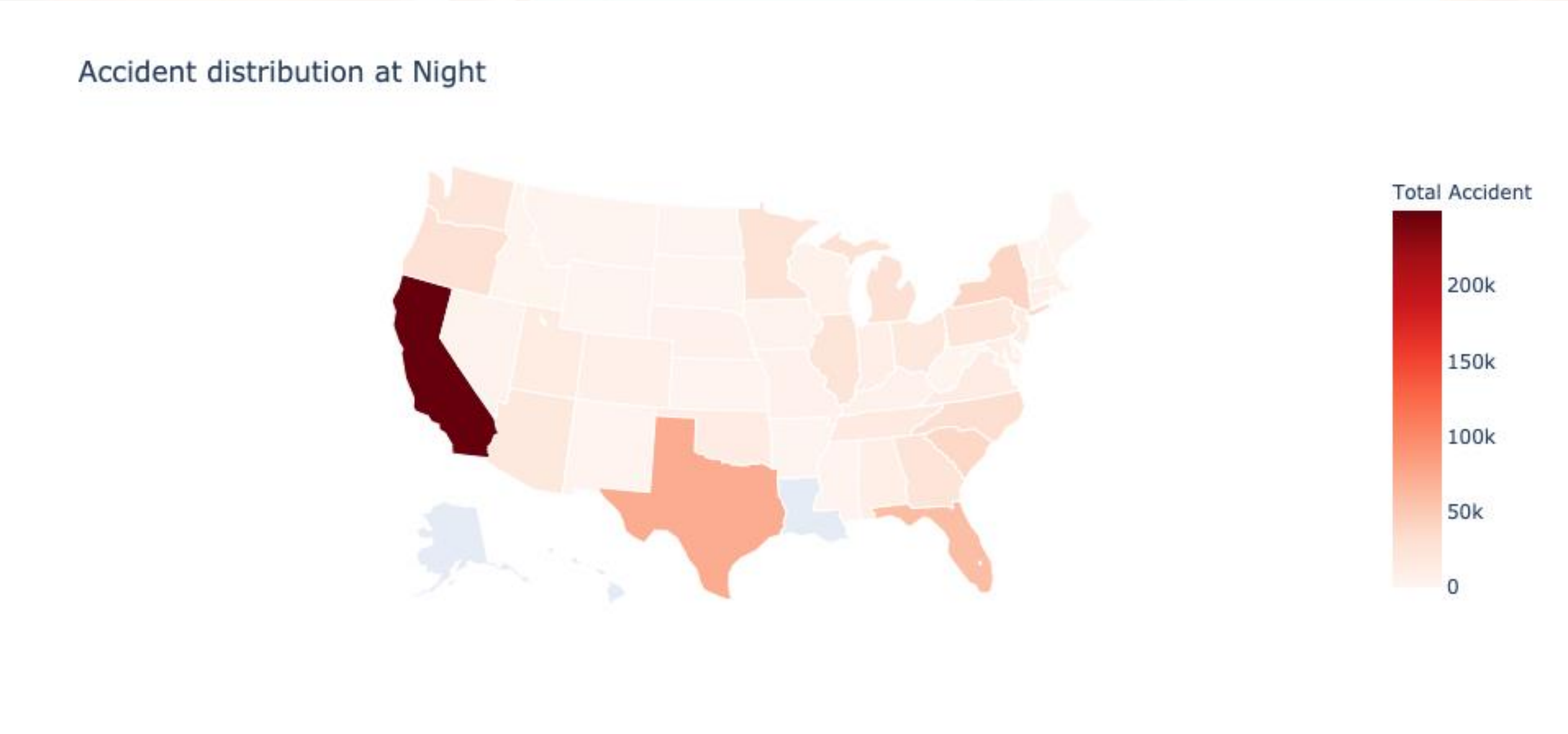
Road Accidents…

Data Visualization: Total Accidents in county level



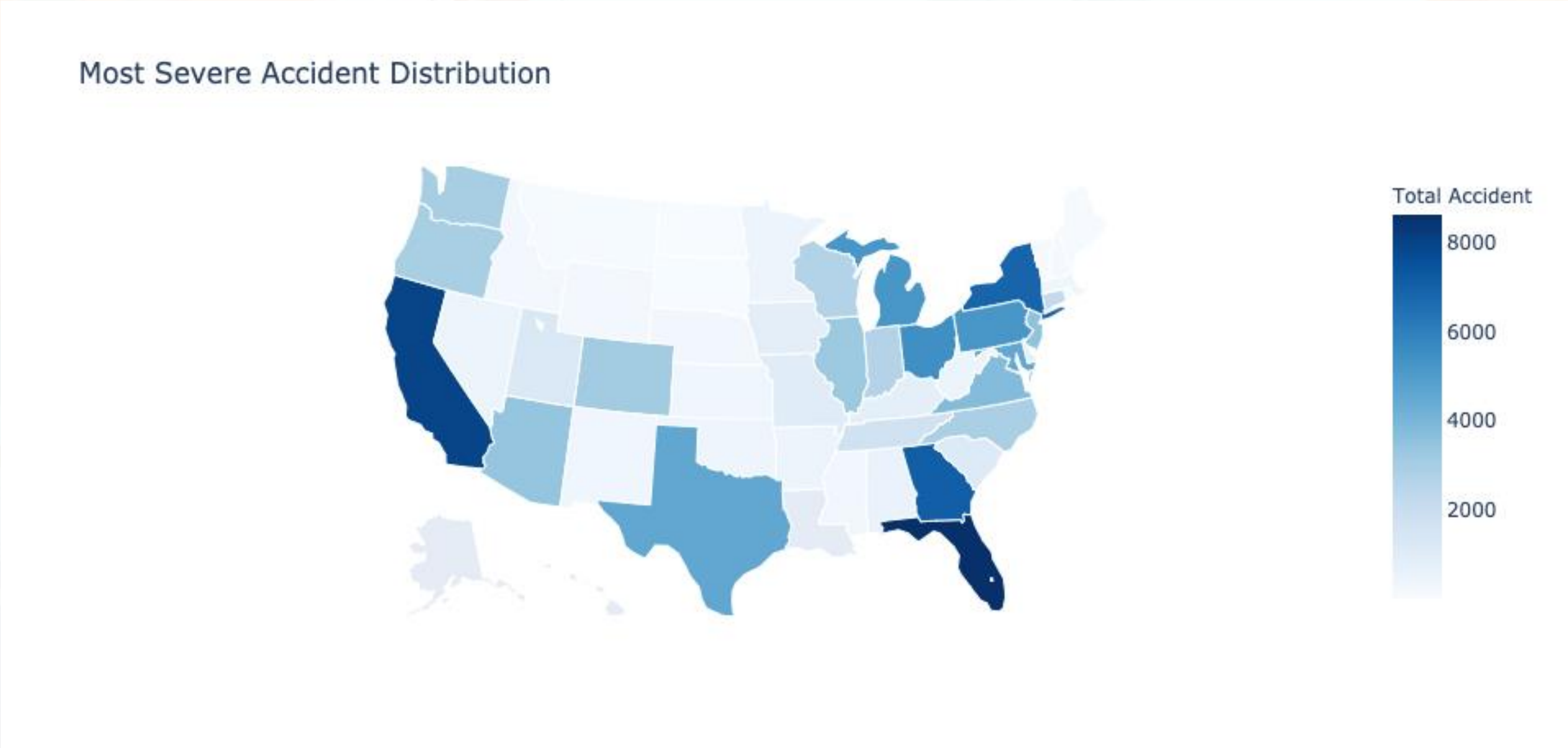
Road Accidents…

Data Visualization: Total Accidents in county level



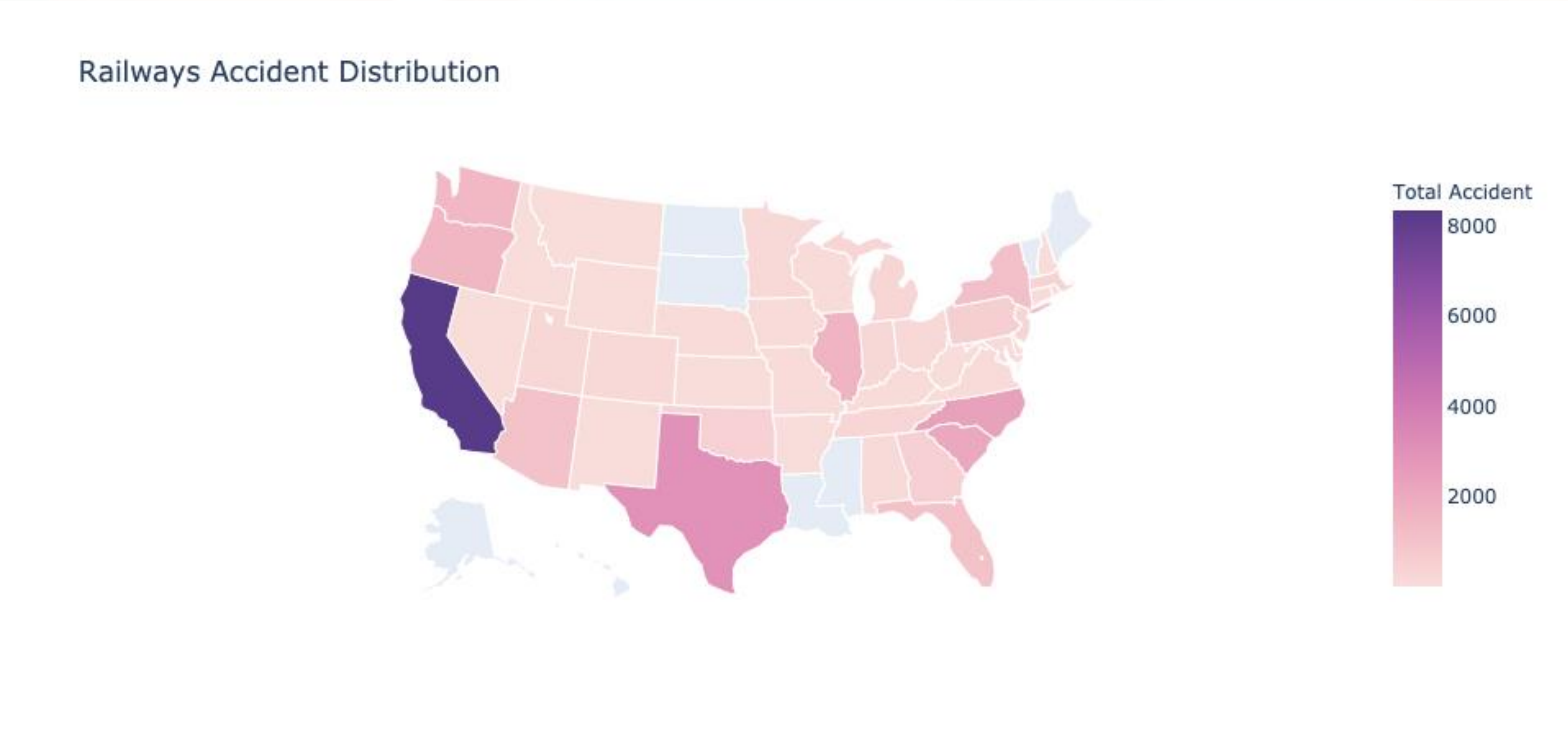
Road Accidents…

Data Visualization: Most severe accidents distribution



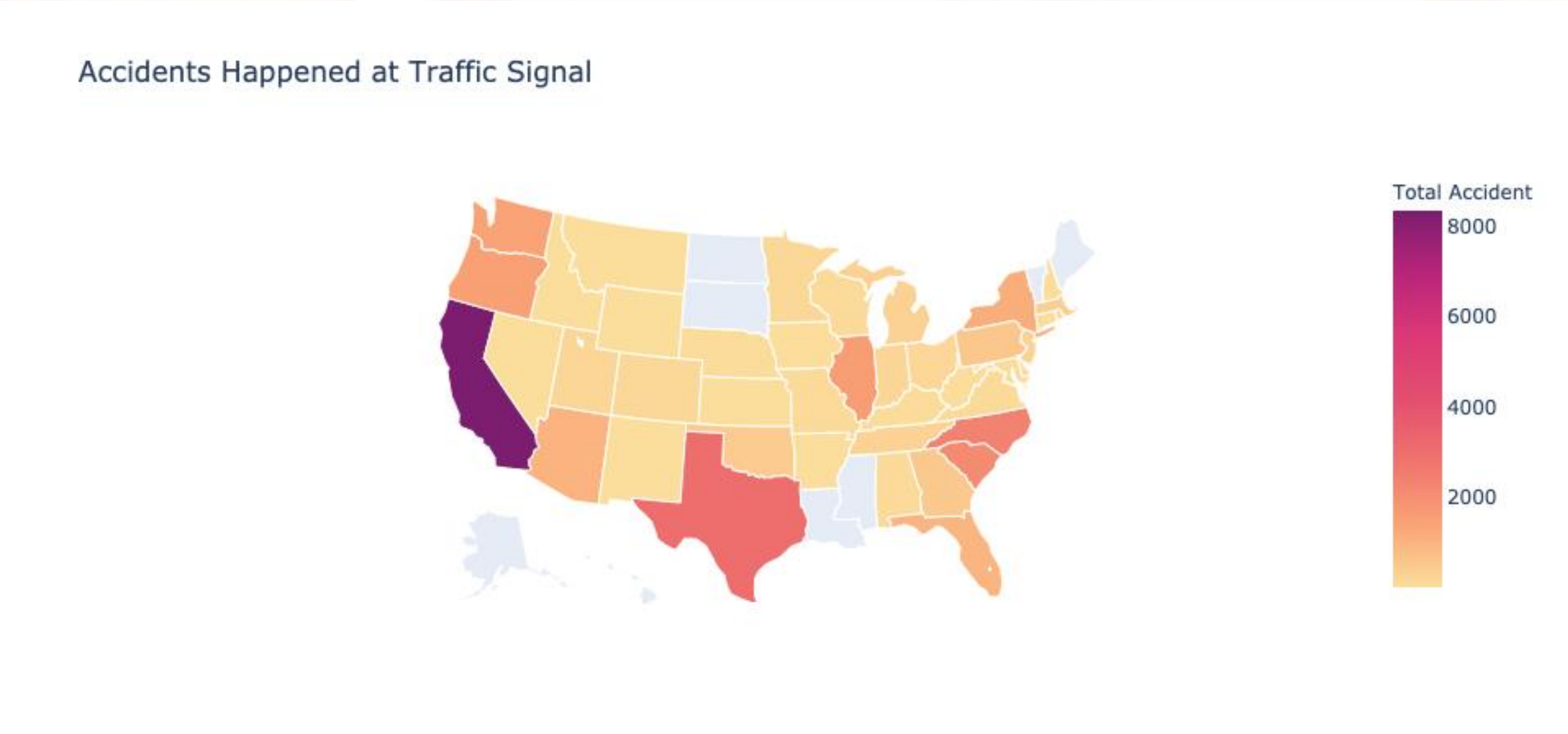
Road Accidents…

Data Visualization: Distribution of accidents due to railways



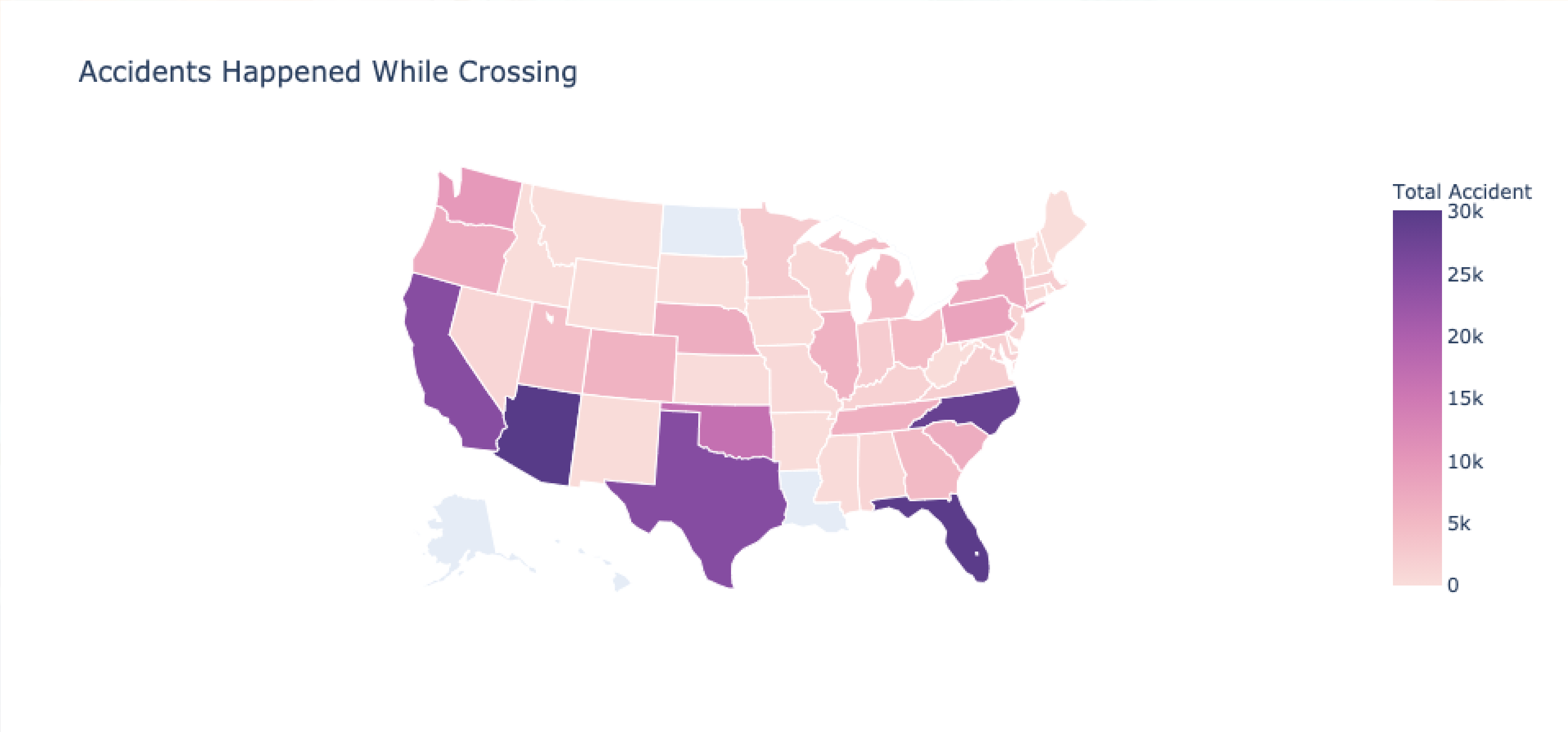
Road Accidents…

Data Visualization: Distribution of accidents -> Traffic Signal



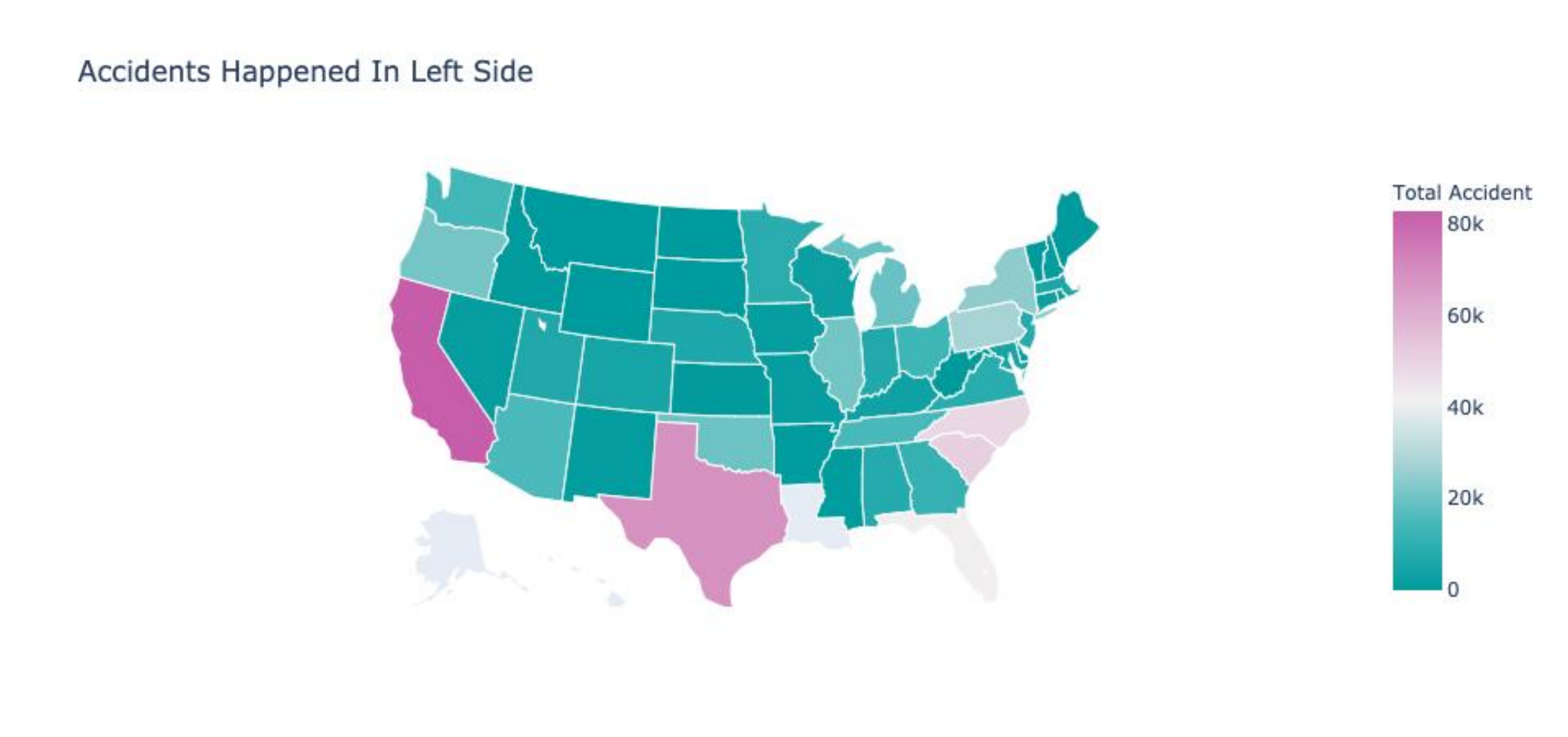
Road Accidents…

Data Visualization: Distribution of accidents -> While Crossing



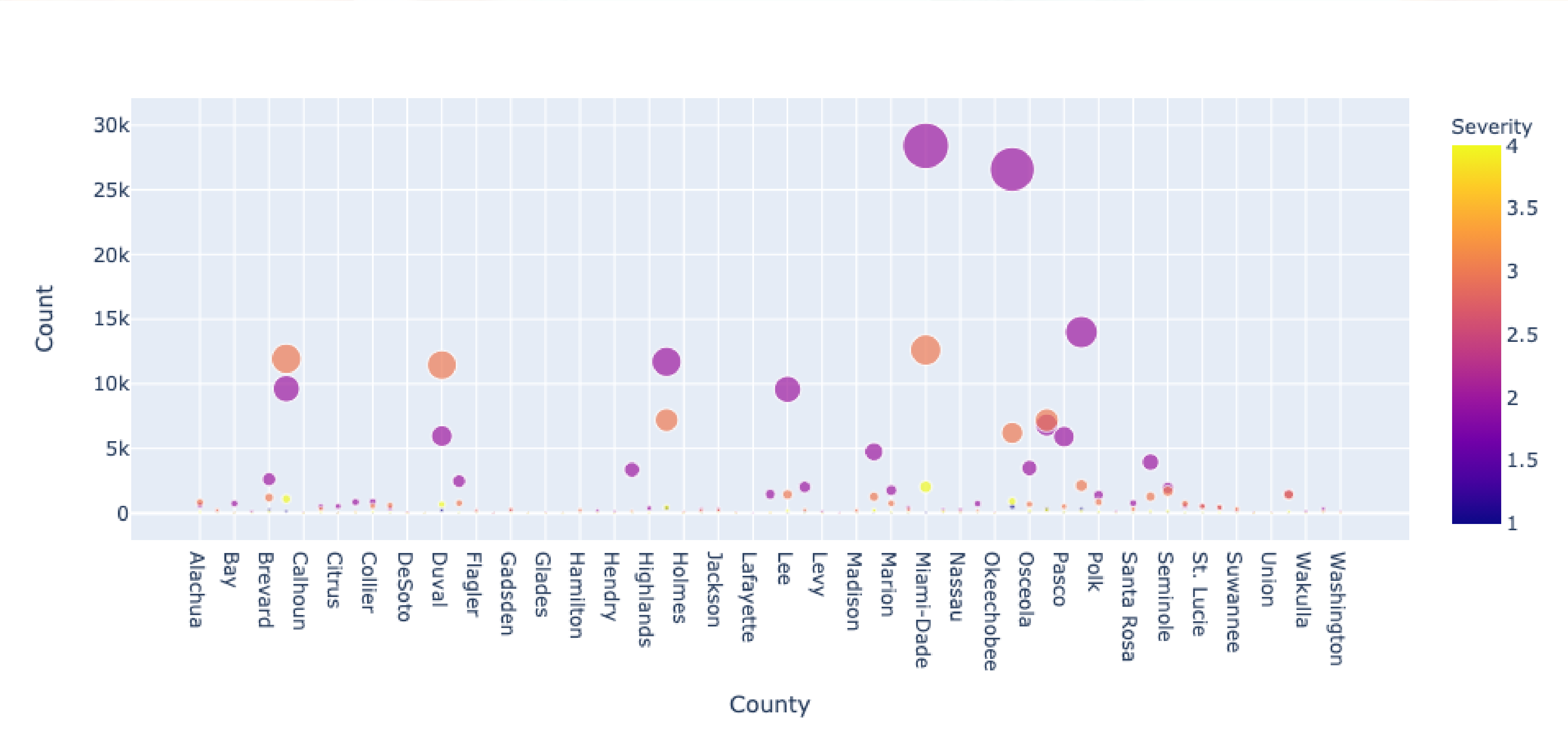
Road Accidents…

Data Visualization: Distribution of accidents -> Left Side



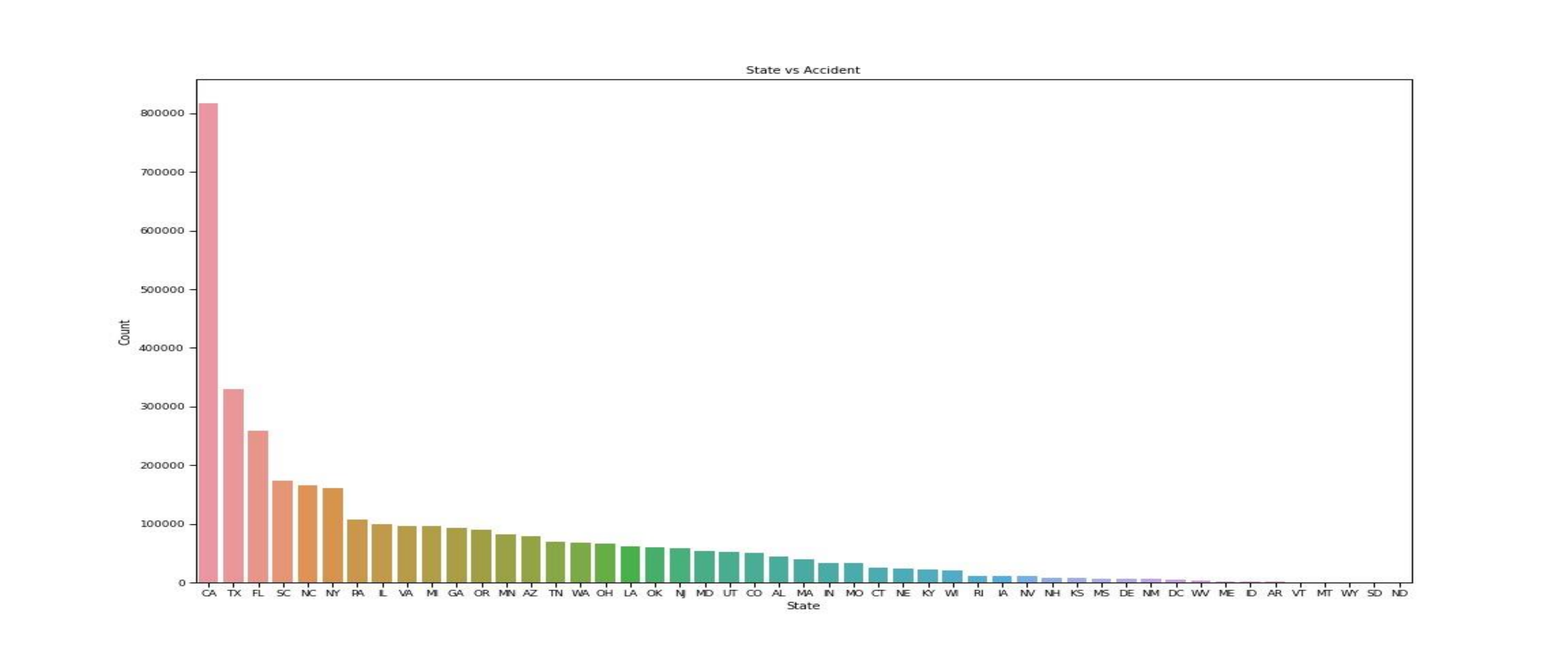
Road Accidents…

Data Visualization: Distribution of accidents -> Florida Counties



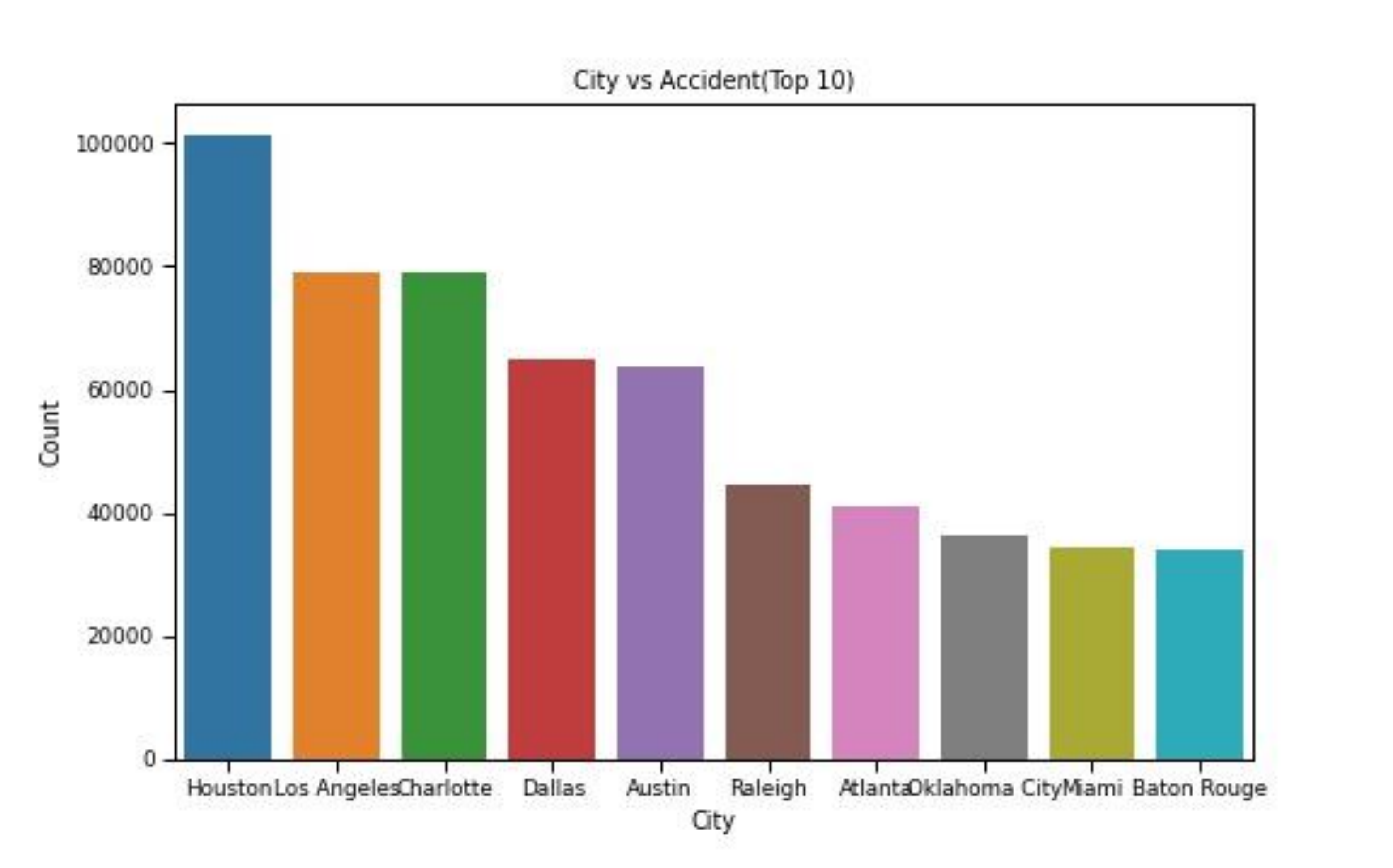
Road Accidents…

State vs Accident



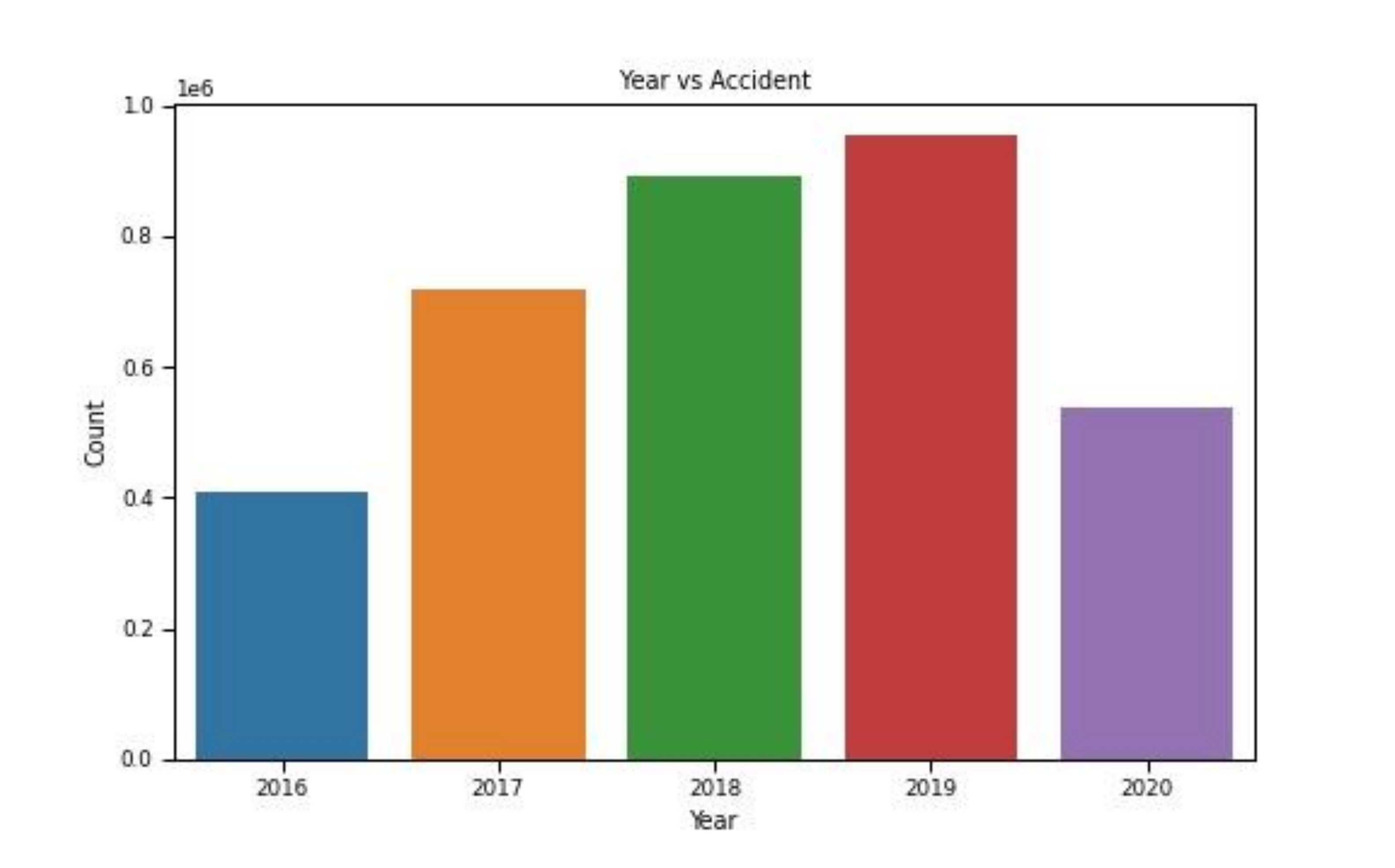
Road Accidents…

City vs Accident(Top 10):



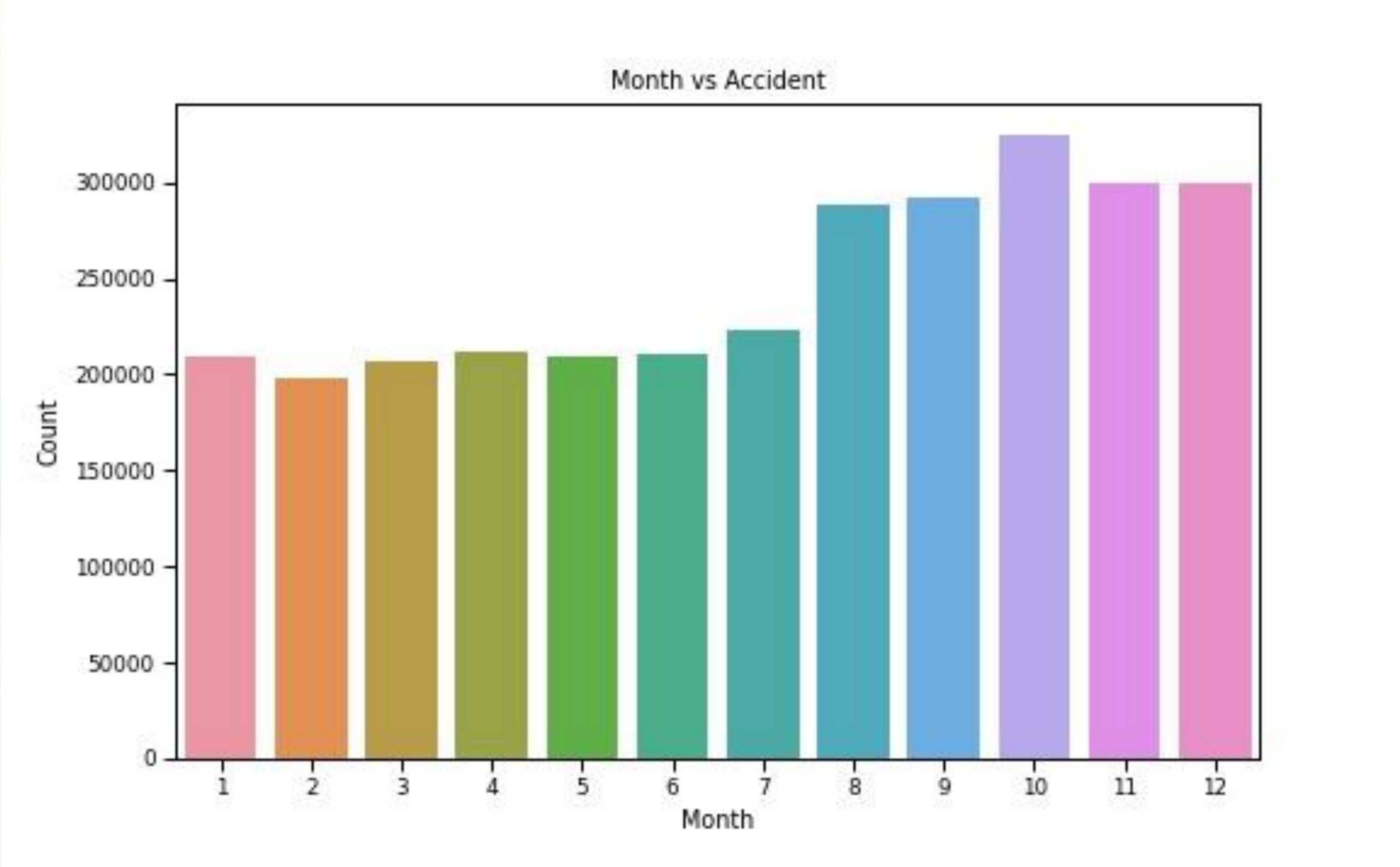
Road Accidents…

Year vs Accident:



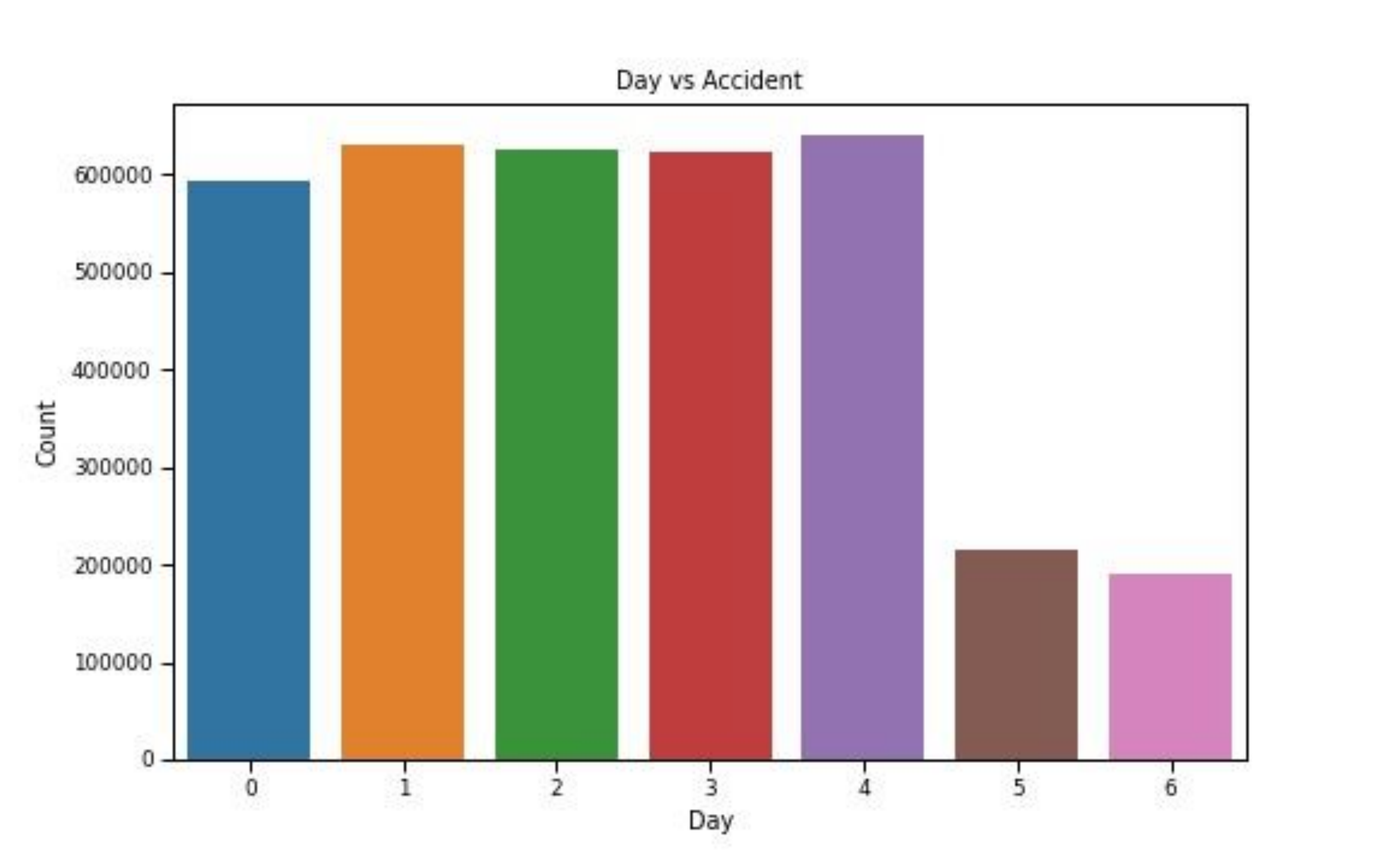
Road Accidents…

Month vs Accident:



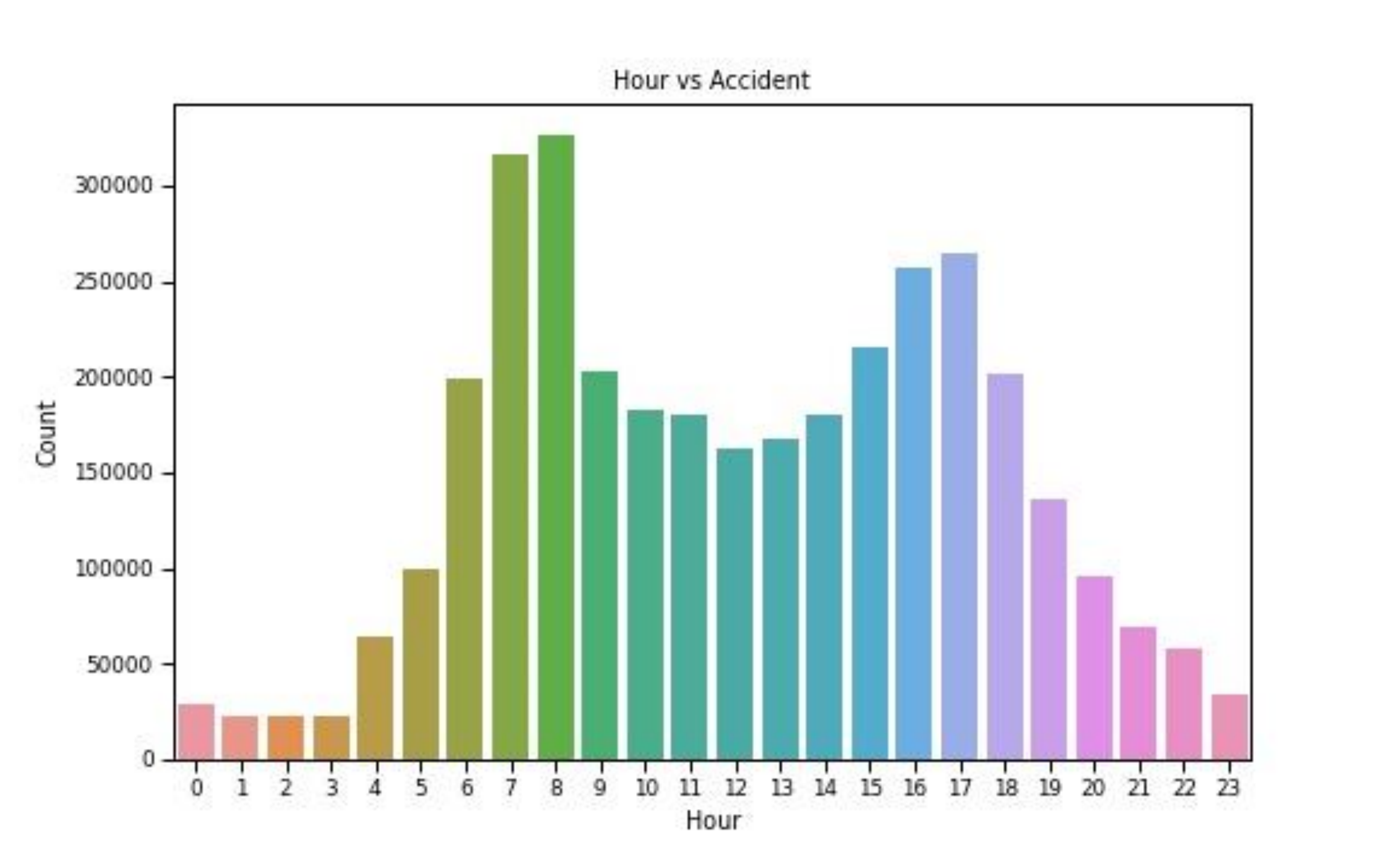
Road Accidents…

Day vs Accident:



Road Accidents…

Hour vs Accident:



Road Accidents…

Result -> Visualization:

- Most Populous state and Cities has highest number of accident.
- The accident number is increasing every year.
- For year 2020, data is available until June and it is already greater than half of 2019.
- The accident was highest in October and overall it occurred more in last five month of the year. (Plot excludes 2020 data)
- October is designated as 'National Pedestrian Safety Month'.
- As expected, accident are higher during weekdays.
- Most of the accident occurred during commute hour/rush hours.

Road Accidents…

Data Visualization: Distribution of accidents due to railways

Severity Analysis

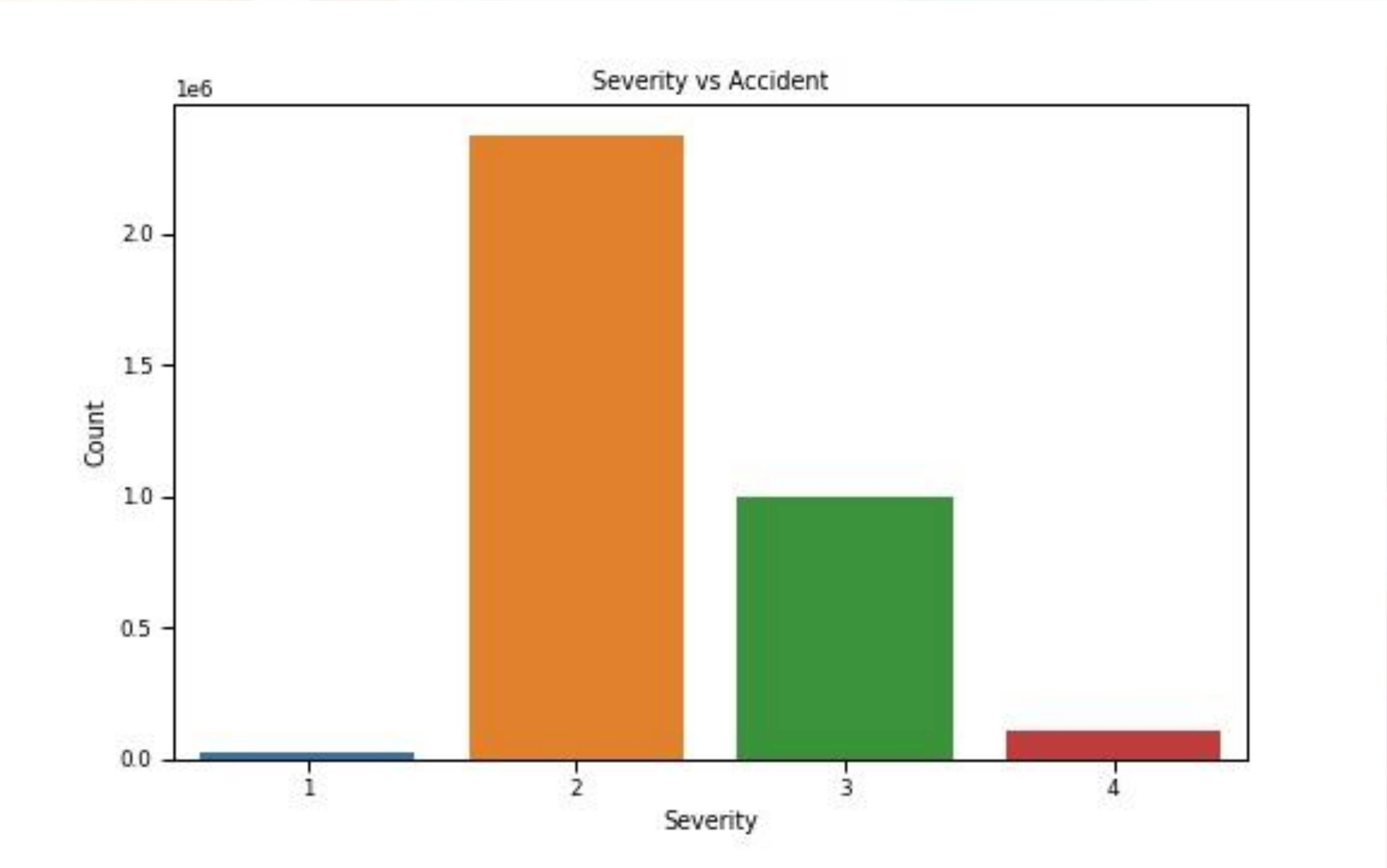
Road Accidents…

Severity Analysis :

- The data is aggregated with respect to severity for different properties.
- Stacked bar plot is used to analyze the data.
- The data is later converted to percentage/rate of severity and are analyzed with respect to different properties.

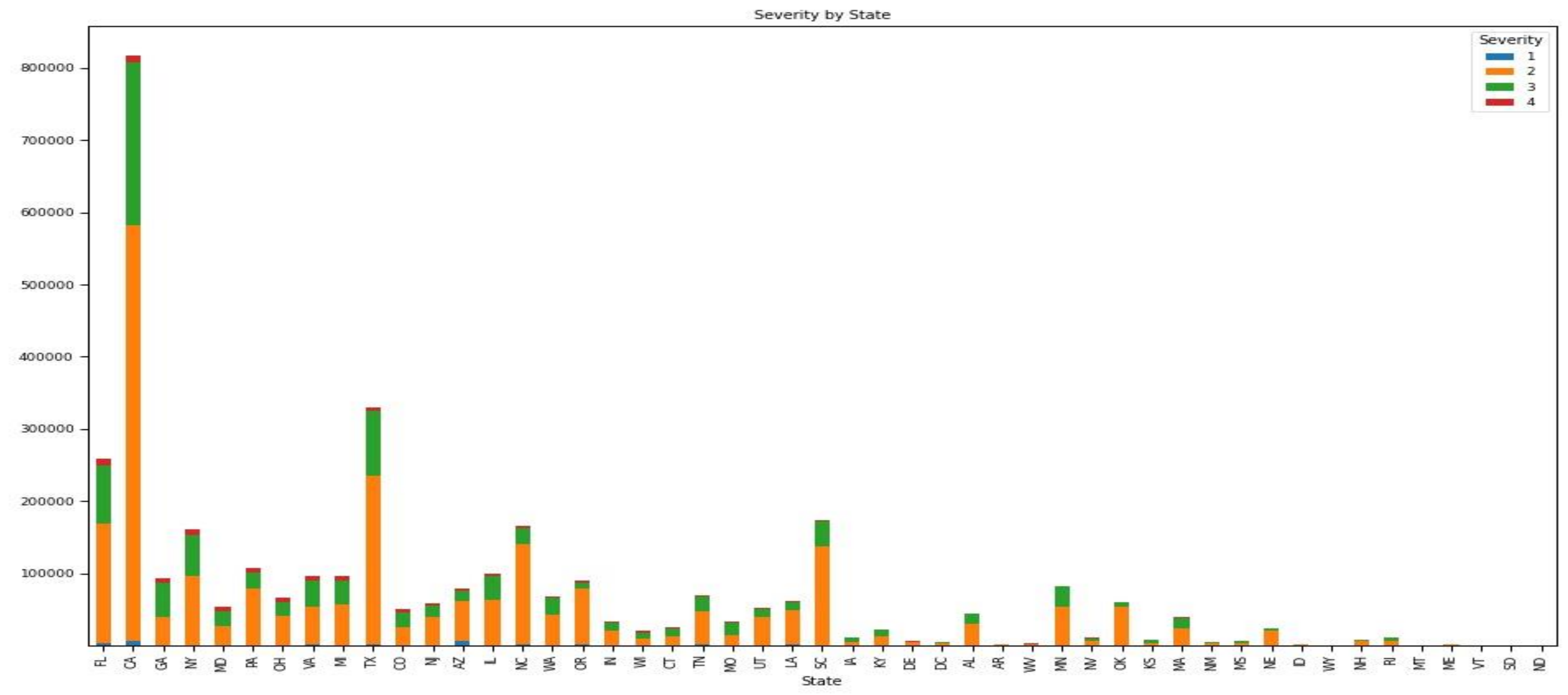
Road Accidents…

Severity Analysis :



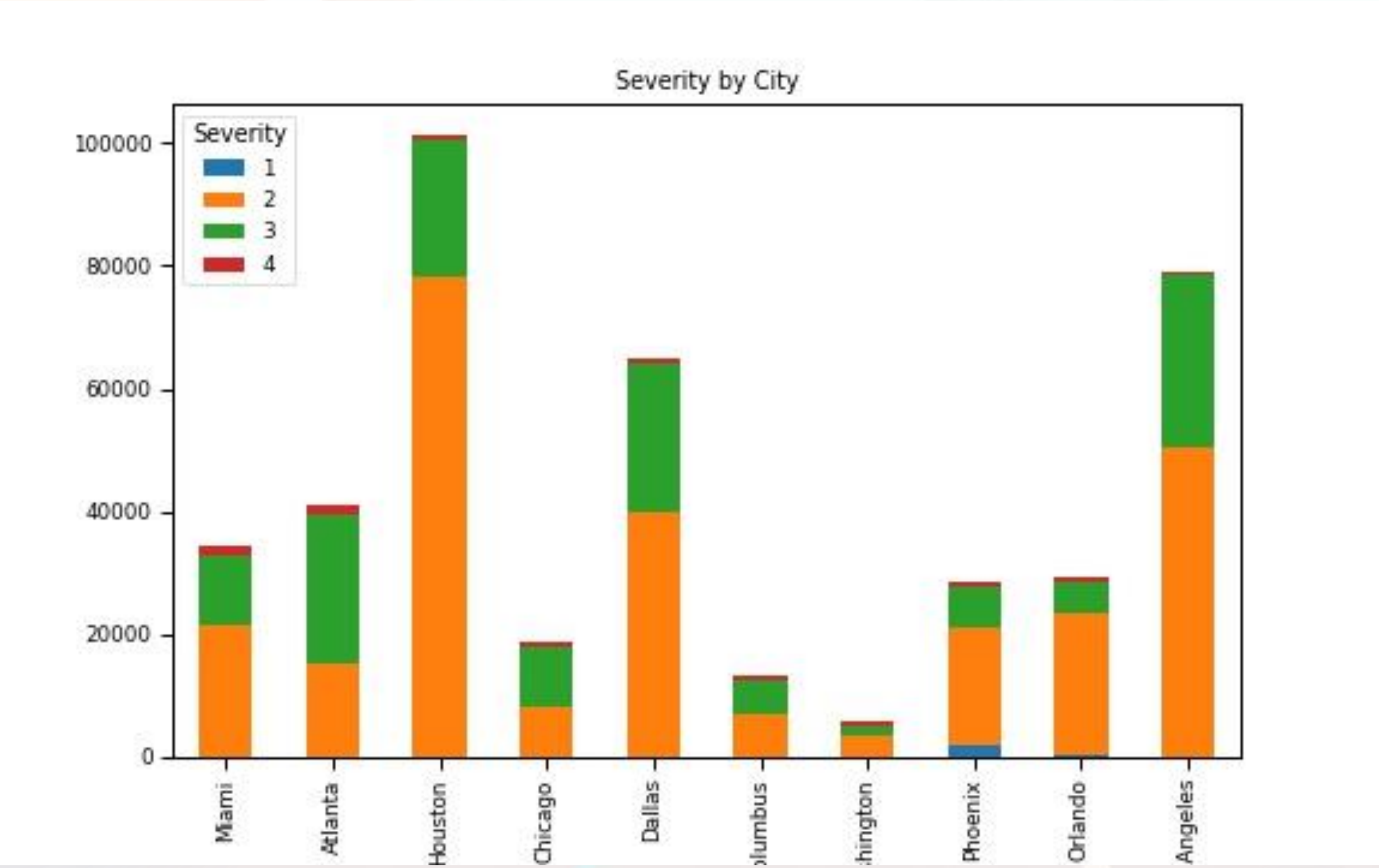
Road Accidents…

Severity By State:



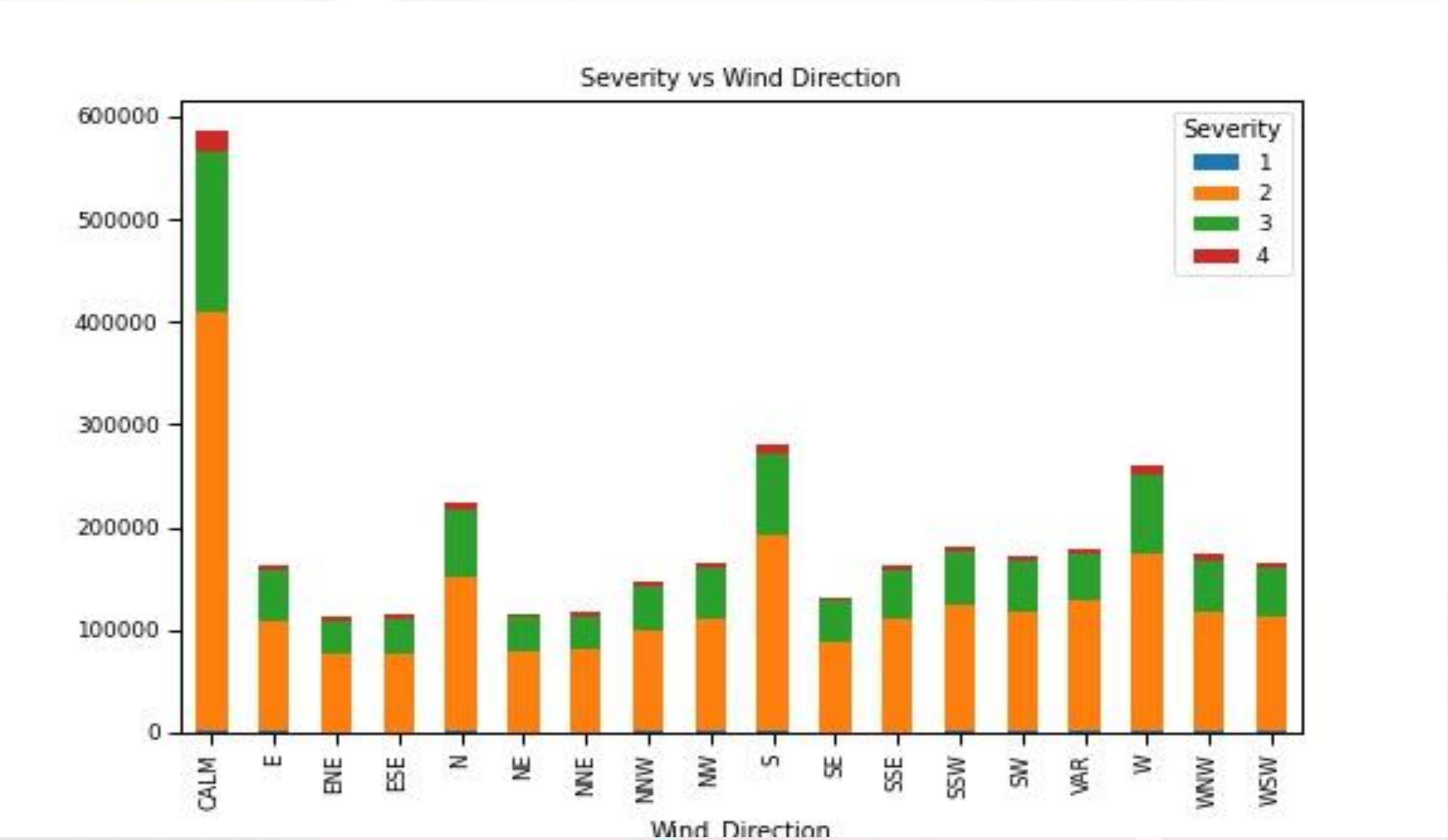
Road Accidents…

Severity By City:



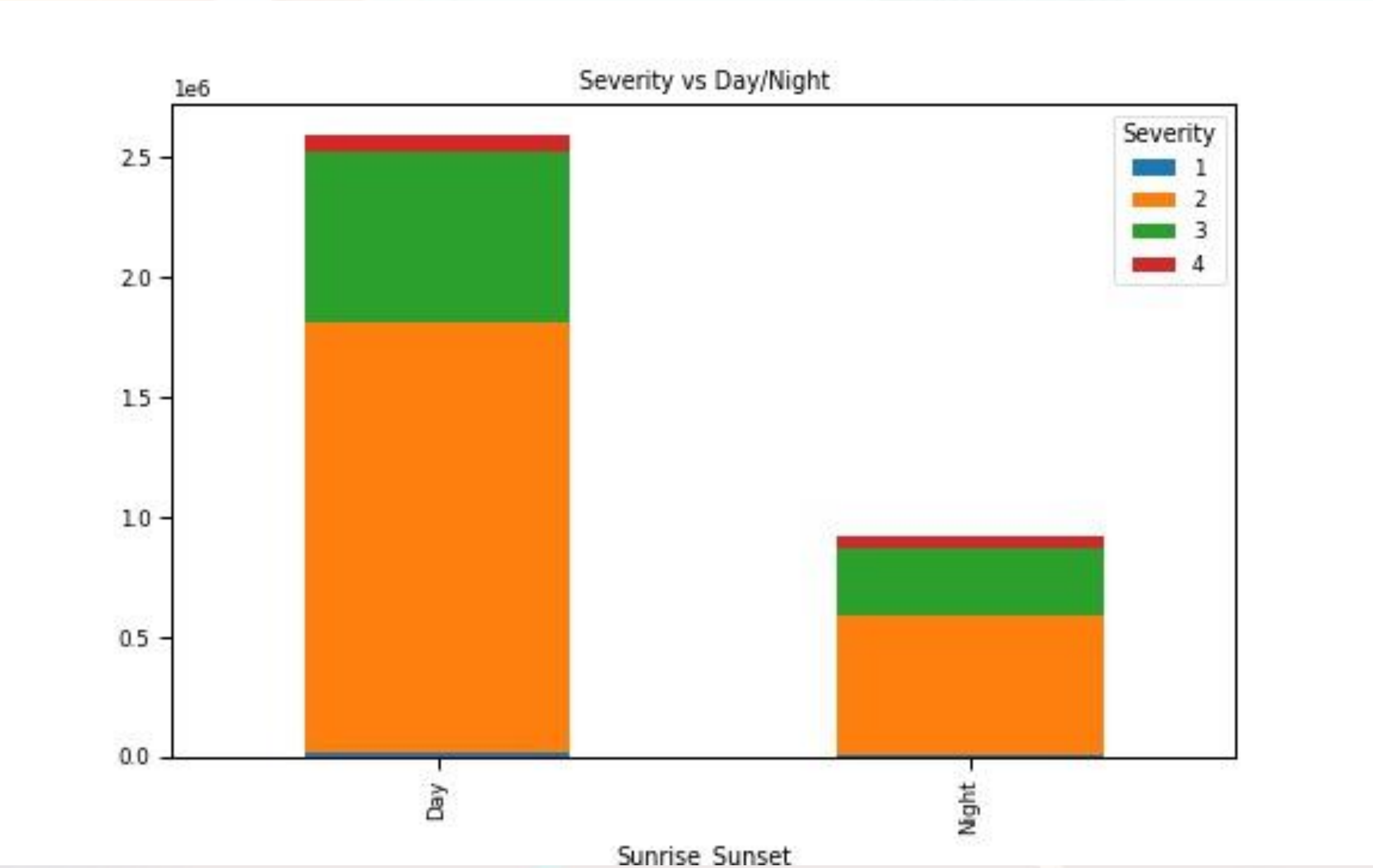
Road Accidents…

Severity Vs Wind Direction:



Road Accidents…

Severity vs Day/Night:



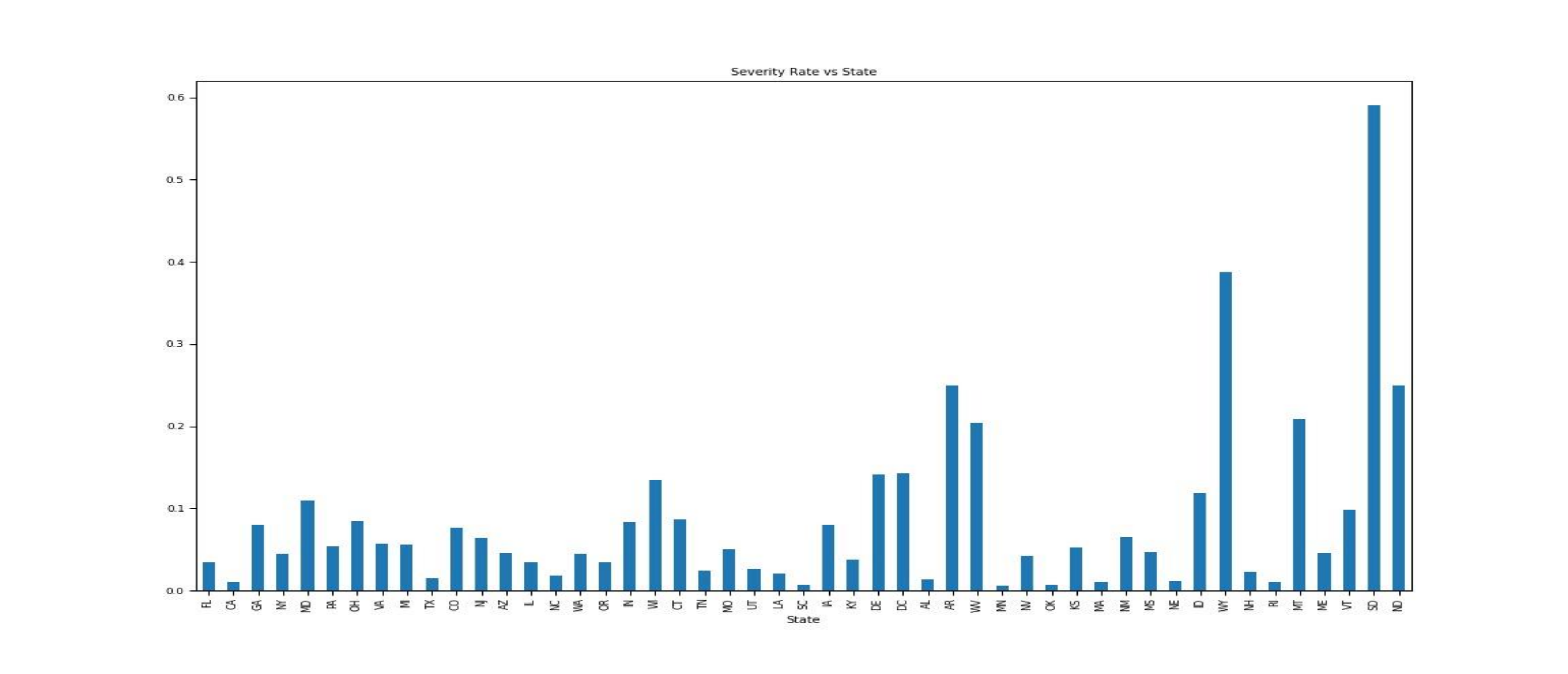
Road Accidents…

Conclusion:

- Most of the accident reported were of level 2.
- Most severe accident occurred in biggest state and cities.
- It was high during day when most of the people commute.
- It was high when the weather direction was normal.
- All the result using count/numbers are expected result. So, let us analyze these along with other properties using the rate/percentage of severity of level 4.

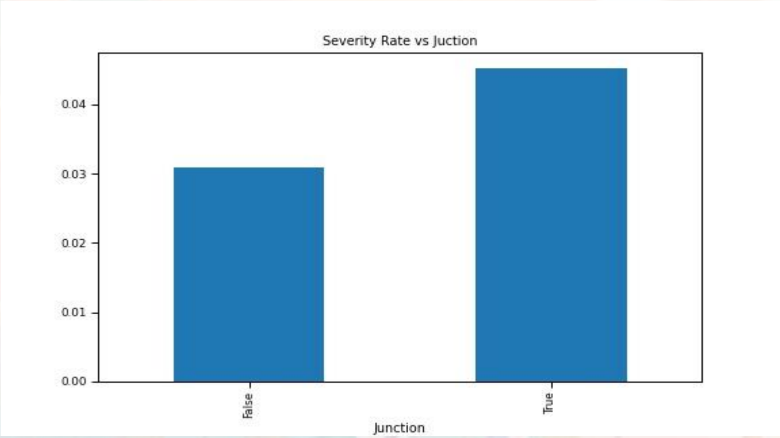
Road Accidents…

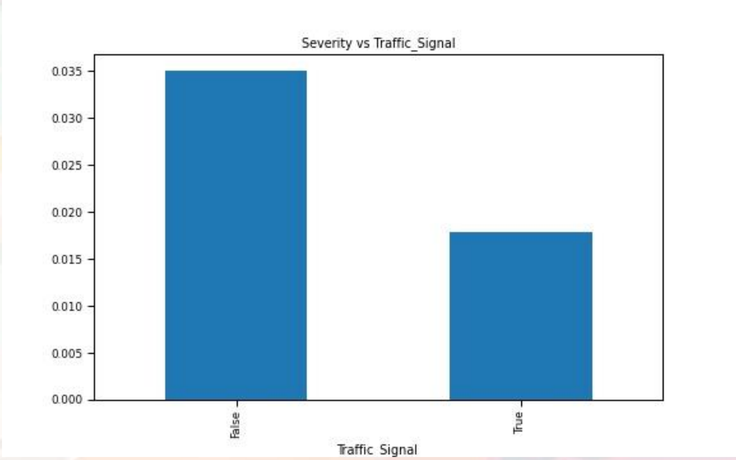
Severity Rate:

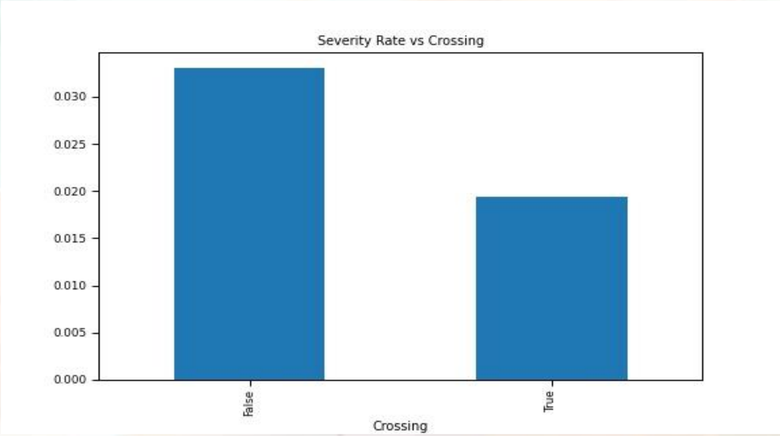


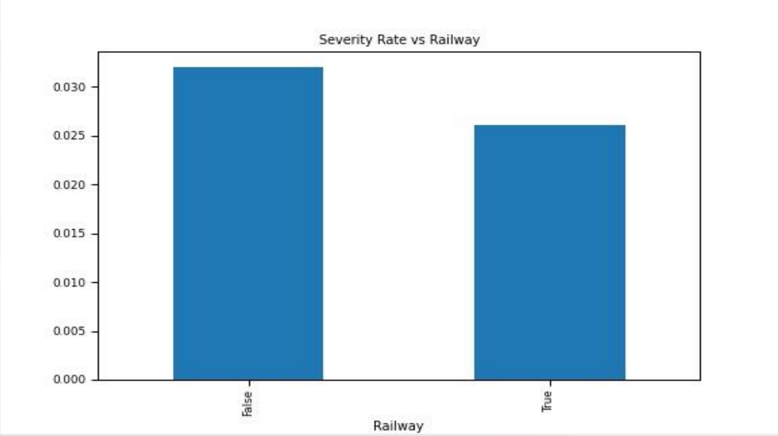
Road Accidents…

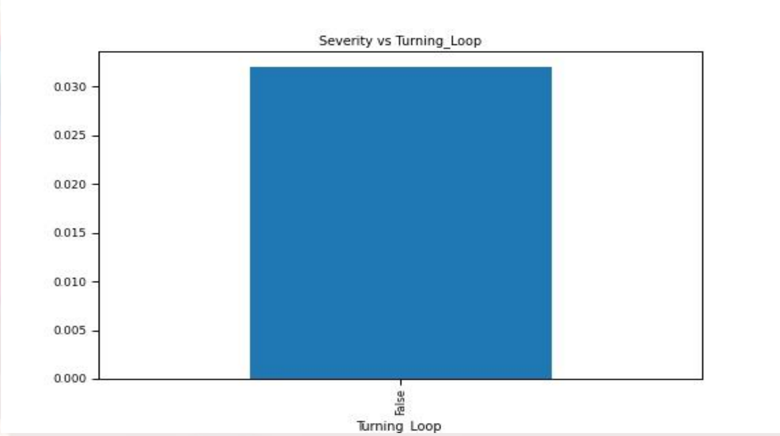
Severity Rate:





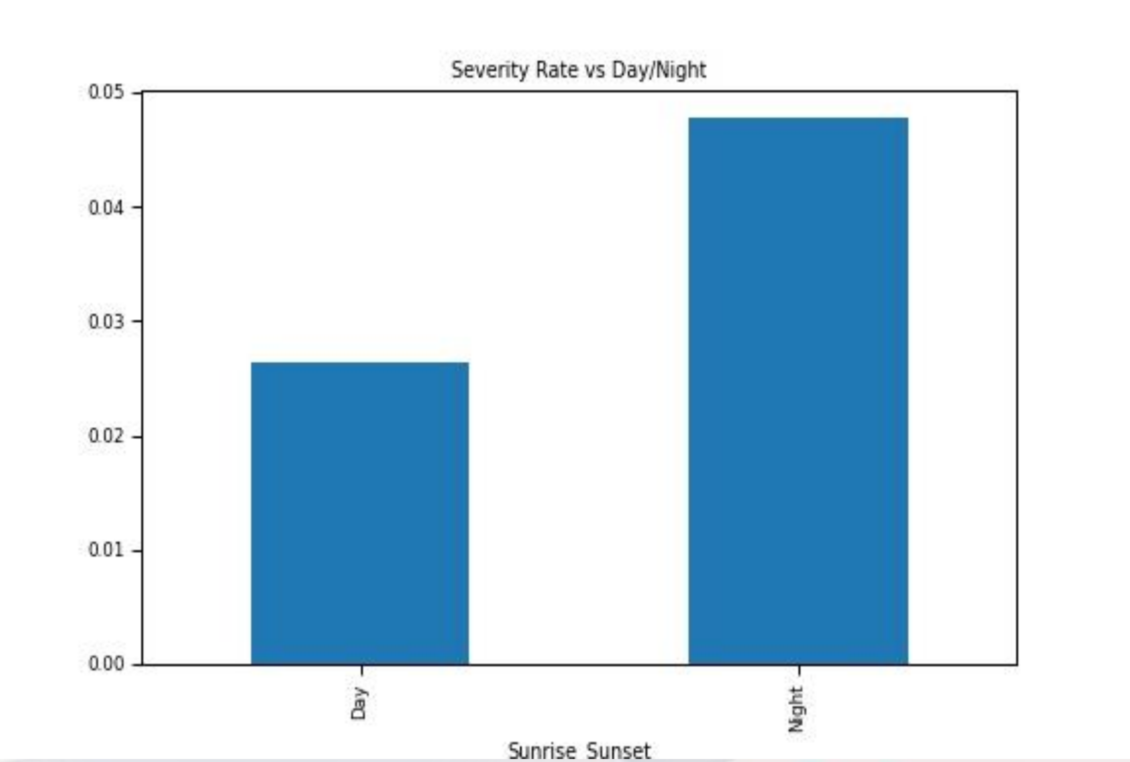


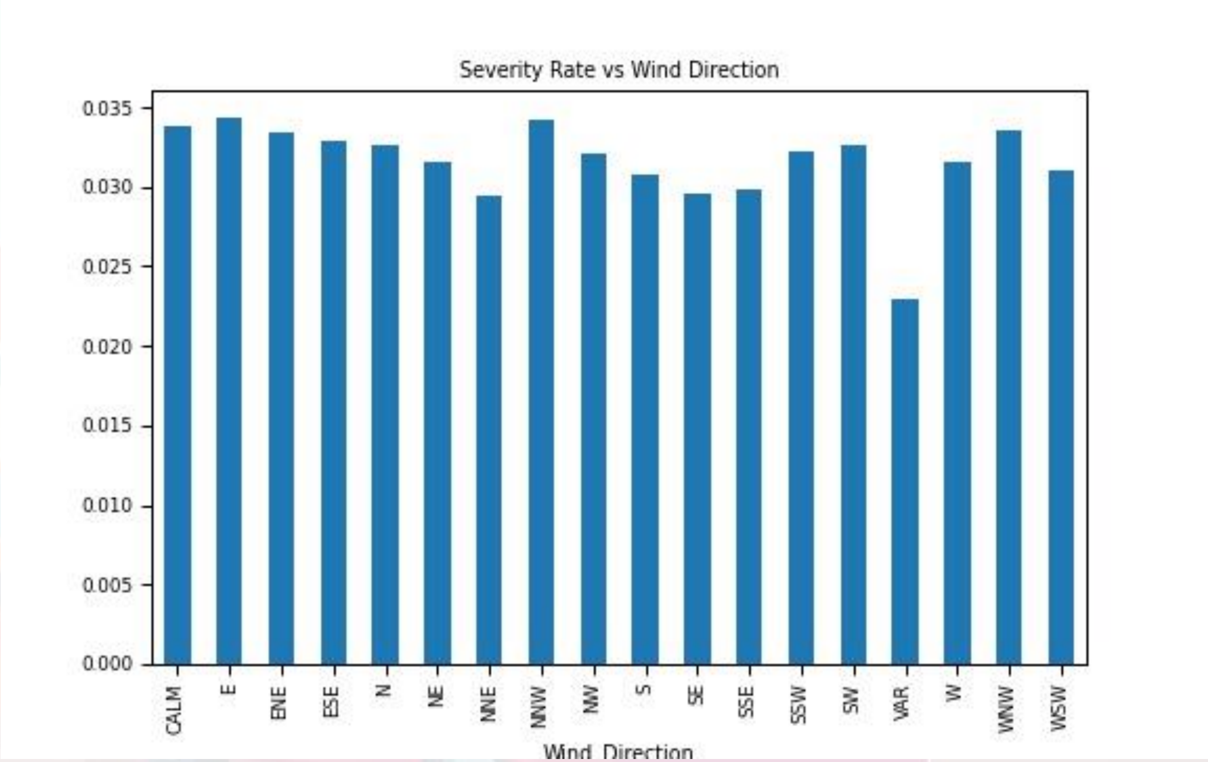




Road Accidents…

Severity Rate:





Road Accidents…

Result -> Severity Analysis :

- State with severe climate condition like South Dakota, North Dakota has highest severity rate.
- Biggest states like California, Florida, etc., has lowest severity rate.
- It seems most of the people are cautious in areas with crossing, traffic signal, turning loops, etc.
- No severe accident were reported in turning loops.
- Severity rate is high for accident reported at Night.
- Almost every wind direction has similar severity rate.

Road Accidents…

CLASSIFICATION

PRAMIL PAUDEL, SUMIT BHATTARAI

Road Accidents…

Classification :

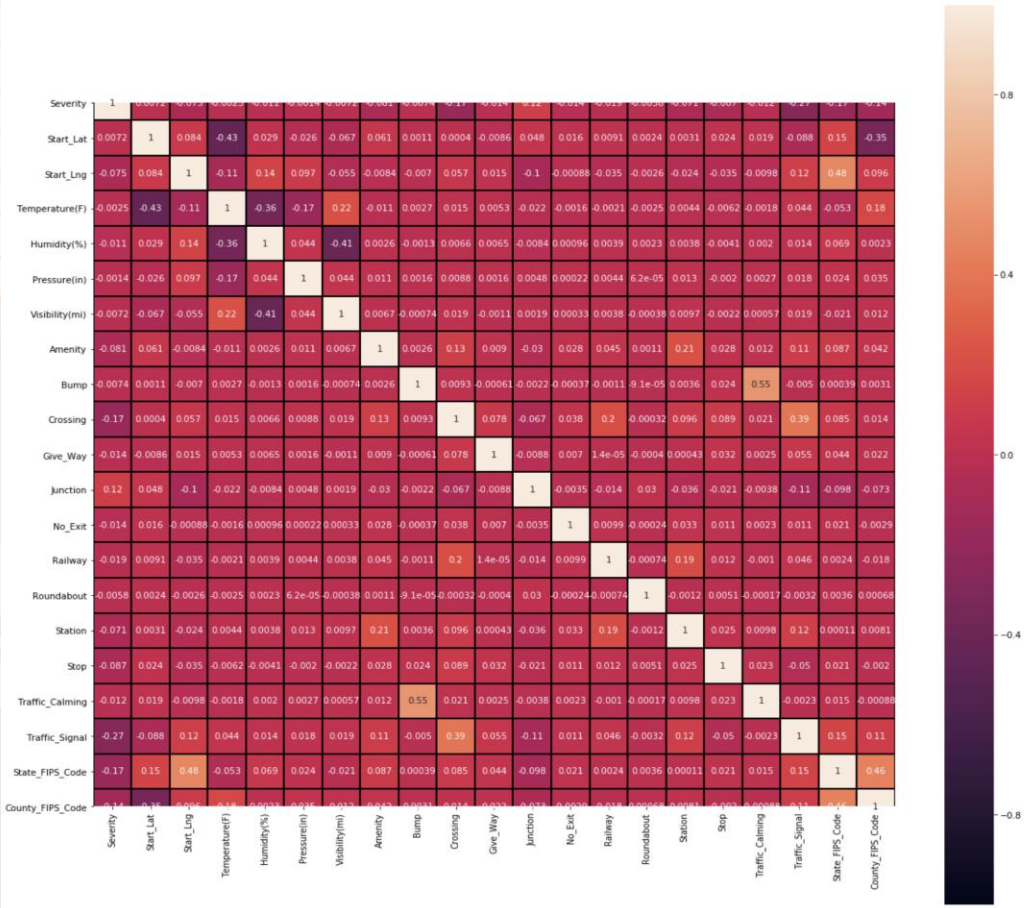
Following are the assumption made while doing classification

- Sample data of size : 499999 , was taken using terminal split over pre-processed data.
- Only 38 columns were selected initially containing non-NA values
- Features sharing only high correlation > 0.05 with targets values are selected.
- Two classifications are used 1. Decision Tree 2. KNN

Road Accidents…

Classification : feature selection

At first correlation heat map was produced for many features

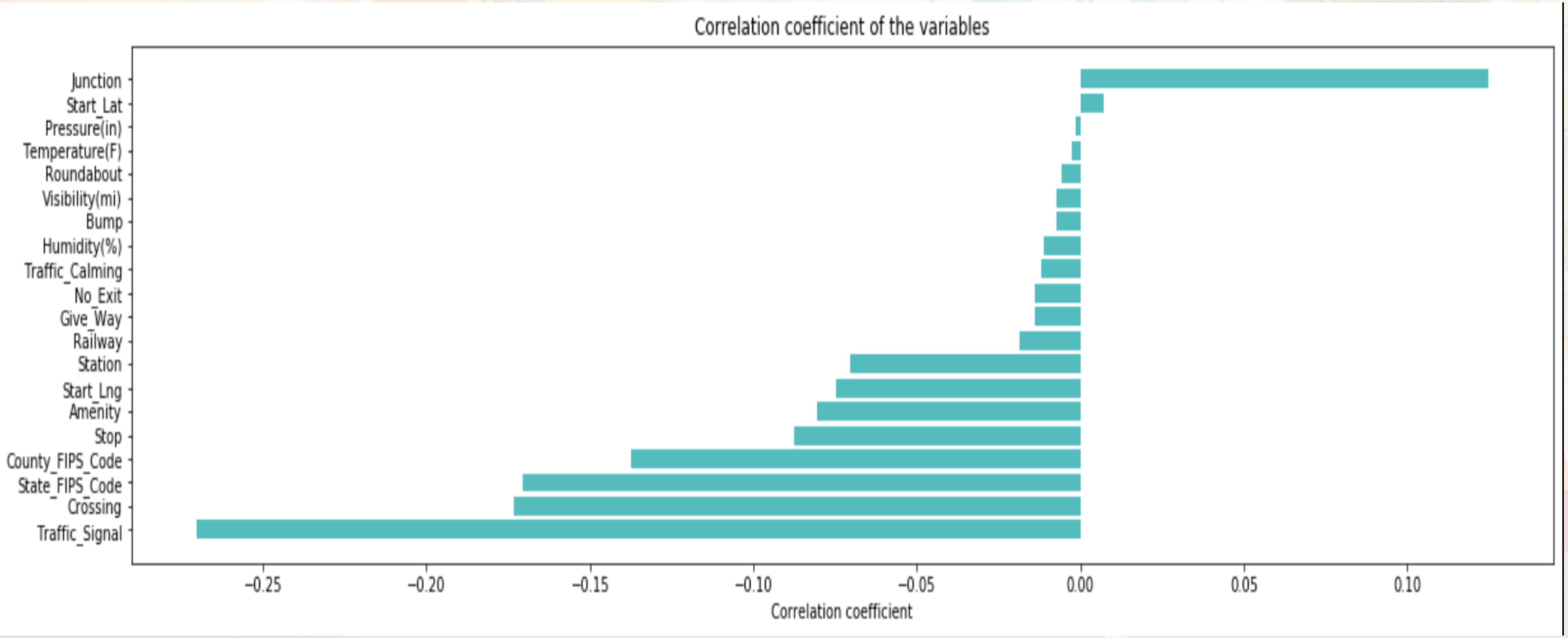


PRAMIL PAUDEL, SUMIT BHATTARAI

Road Accidents…

Classification : feature selection

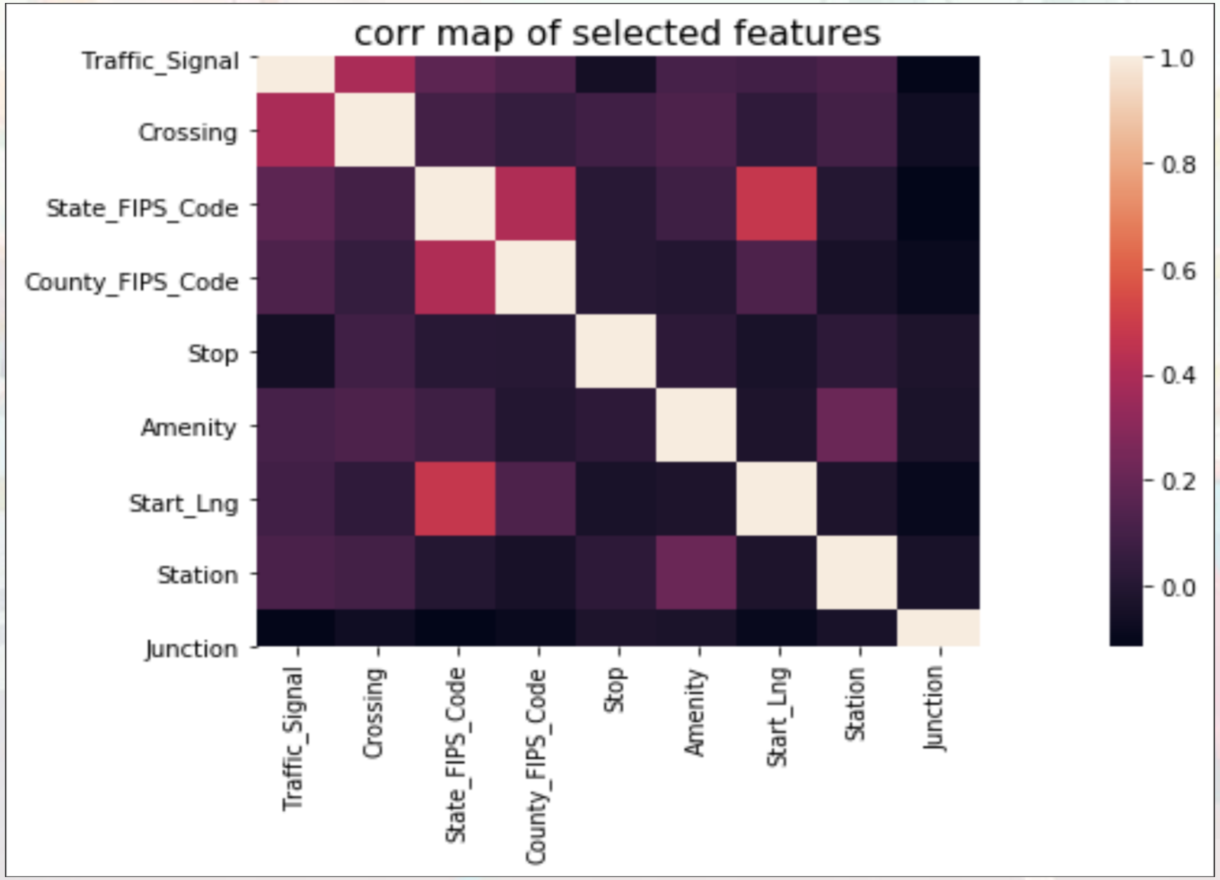
- Correlation values are calculated and plotted as below
- Traffic Signal, Crossing, State, County, etc. have highest correlation factor



Road Accidents…

Classification : feature selection

- Only features containing more than 0.05 correlation factor are selected.



Road Accidents…

Classification : Decision Tree Classifier

- There are two Decision Tree Classification are used based on Criteria
  - ❑ Entropy – where entropies of each features is calculated, and tree is generated.
  - ❑ Gini - Tree is generated used Gini Coefficient
- Depth Limitation : Decision tree pruning is applied using predefined tree depth.
  - ❑ Classification is tested using different depth.
- 70% Train and 30% Test data is used in classification

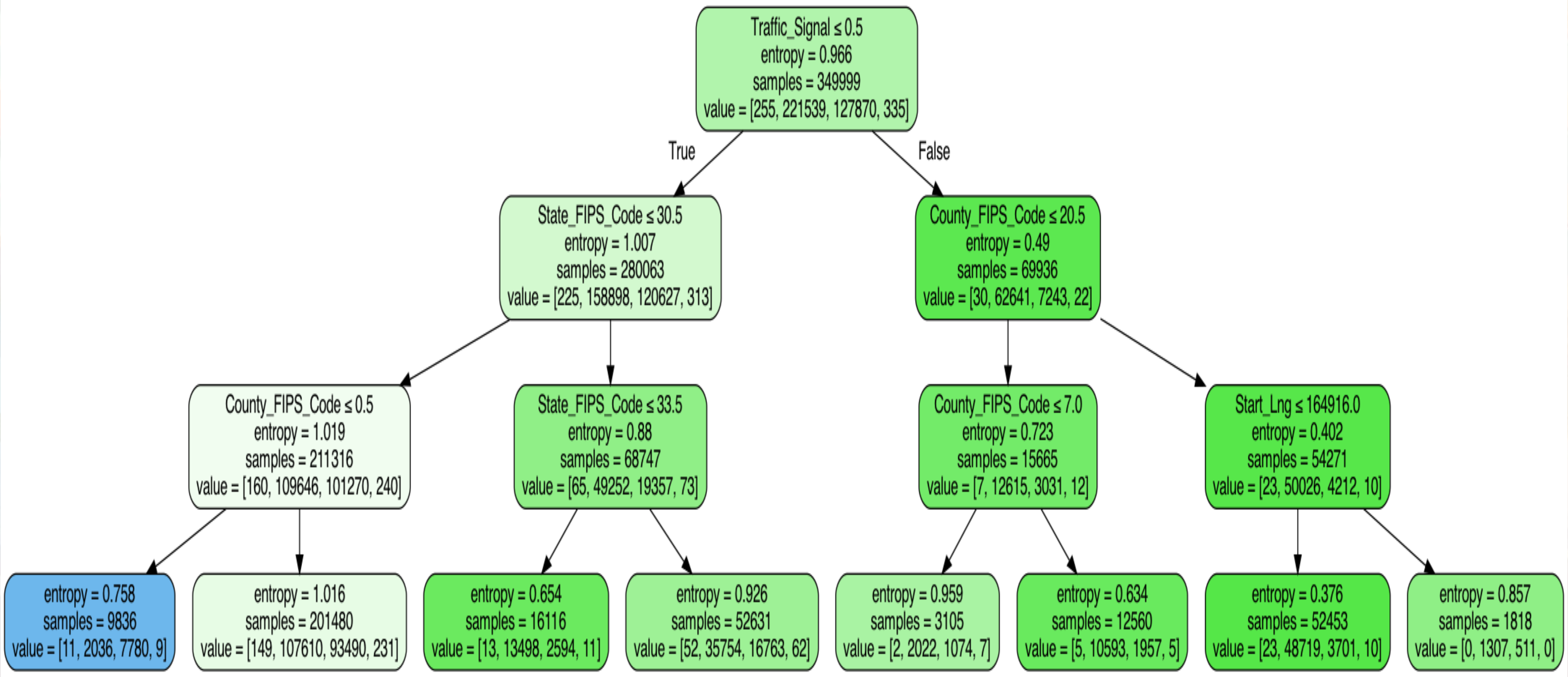
Road Accidents…

Classification : Decision Tree Classifier -> Entropy

Accuracy of Entropy Classifier using 70.0 % training and depth 1: : : : 0.6360133333333333
Accuracy of Entropy Classifier using 70.0 % training and depth 2: : : : 0.6360133333333333
Accuracy of Entropy Classifier using 70.0 % training and depth 3: : : : 0.6519
Accuracy of Entropy Classifier using 70.0 % training and depth 4: : : : 0.6543533333333333
Accuracy of Entropy Classifier using 70.0 % training and depth 5: : : : 0.66366
Accuracy of Entropy Classifier using 70.0 % training and depth 6: : : : 0.6882533333333334
Accuracy of Entropy Classifier using 70.0 % training and depth 7: : : : 0.7021666666666667
Accuracy of Entropy Classifier using 70.0 % training and depth 8: : : : 0.7262066666666667
Accuracy of Entropy Classifier using 70.0 % training and depth 9: : : : 0.7318066666666667
Accuracy of Entropy Classifier using 70.0 % training and depth 10: : : : 0.74606
Accuracy of Entropy Classifier using 70.0 % training and depth 11: : : : 0.7554733333333333
Accuracy of Entropy Classifier using 70.0 % training and depth 12: : : : 0.7715
Accuracy of Entropy Classifier using 70.0 % training and depth 13: : : : 0.7841733333333333
Accuracy of Entropy Classifier using 70.0 % training and depth 14: : : : 0.7954333333333333
Accuracy of Entropy Classifier using 70.0 % training and depth 15: : : : 0.8052333333333334
Accuracy of Entropy Classifier using 70.0 % training and depth 16: : : : 0.8128733333333333
Accuracy of Entropy Classifier using 70.0 % training and depth 17: : : : 0.82224
Accuracy of Entropy Classifier using 70.0 % training and depth 18: : : : 0.8303133333333333
Accuracy of Entropy Classifier using 70.0 % training and depth 19: : : : 0.8402933333333333
Accuracy of Entropy Classifier using 70.0 % training and depth 20: : : : 0.8497533333333334

Road Accidents…

Classification : Decision Tree Classifier -> Entropy



Road Accidents…

Classification : Decision Tree Classifier -> Entropy -> Observation

Confusion Matrix

1	80	36	0
33	83861	11496	12
6	10691	43600	35
0	36	112	1

Increased Accuracy with Depth ?

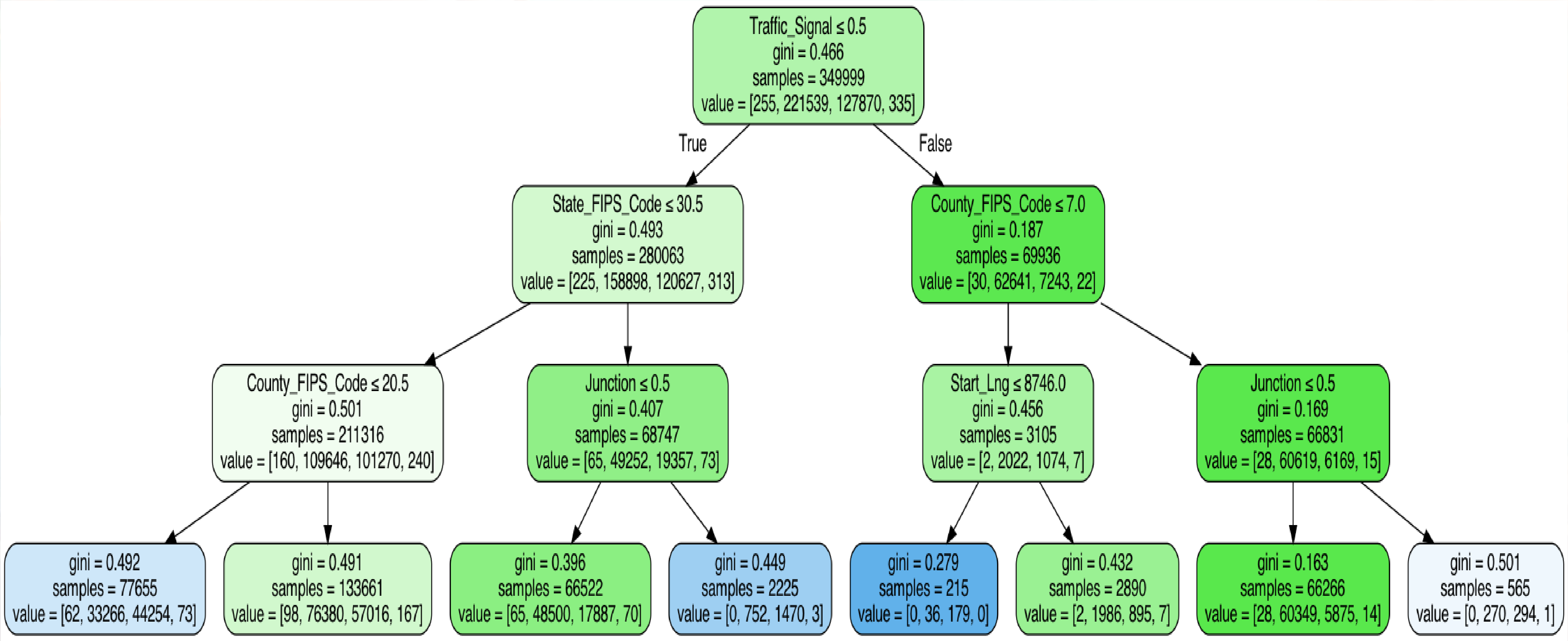
Road Accidents…

Classification : Decision Tree Classifier -> Gini

Accuracy of Gini Classifier using 70.0 % training and depth 1: : : : 0.6360133333333333
Accuracy of Gini Classifier using 70.0 % training and depth 2: : : : 0.6360133333333333
Accuracy of Gini Classifier using 70.0 % training and depth 3: : : : 0.6663866666666667
Accuracy of Gini Classifier using 70.0 % training and depth 4: : : : 0.6819066666666667
Accuracy of Gini Classifier using 70.0 % training and depth 5: : : : 0.6943
Accuracy of Gini Classifier using 70.0 % training and depth 6: : : : 0.7009333333333333
Accuracy of Gini Classifier using 70.0 % training and depth 7: : : : 0.7124733333333333
Accuracy of Gini Classifier using 70.0 % training and depth 8: : : : 0.7339066666666667
Accuracy of Gini Classifier using 70.0 % training and depth 9: : : : 0.7475333333333334
Accuracy of Gini Classifier using 70.0 % training and depth 10: : : : 0.7635866666666666
Accuracy of Gini Classifier using 70.0 % training and depth 11: : : : 0.7734
Accuracy of Gini Classifier using 70.0 % training and depth 12: : : : 0.7911733333333333
Accuracy of Gini Classifier using 70.0 % training and depth 13: : : : 0.8015533333333333
Accuracy of Gini Classifier using 70.0 % training and depth 14: : : : 0.8101066666666666
Accuracy of Gini Classifier using 70.0 % training and depth 15: : : : 0.81742
Accuracy of Gini Classifier using 70.0 % training and depth 16: : : : 0.8262866666666666
Accuracy of Gini Classifier using 70.0 % training and depth 17: : : : 0.833
Accuracy of Gini Classifier using 70.0 % training and depth 18: : : : 0.8433533333333333
Accuracy of Gini Classifier using 70.0 % training and depth 19: : : : 0.8517666666666667
Accuracy of Gini Classifier using 70.0 % training and depth 20: : : : 0.86004

Road Accidents…

Classification : Decision Tree Classifier -> Gini



Road Accidents…

Classification : Decision Tree Classifier -> Gini-> Observation

Confusion Matrix

Actual

1	**79**	**37**	**0**
23	84921	10449	9
3	10217	44083	29
0	37	111	1

Predicted

Recall = TP/(TP+FN),
Precision = TP/(TP+FP)

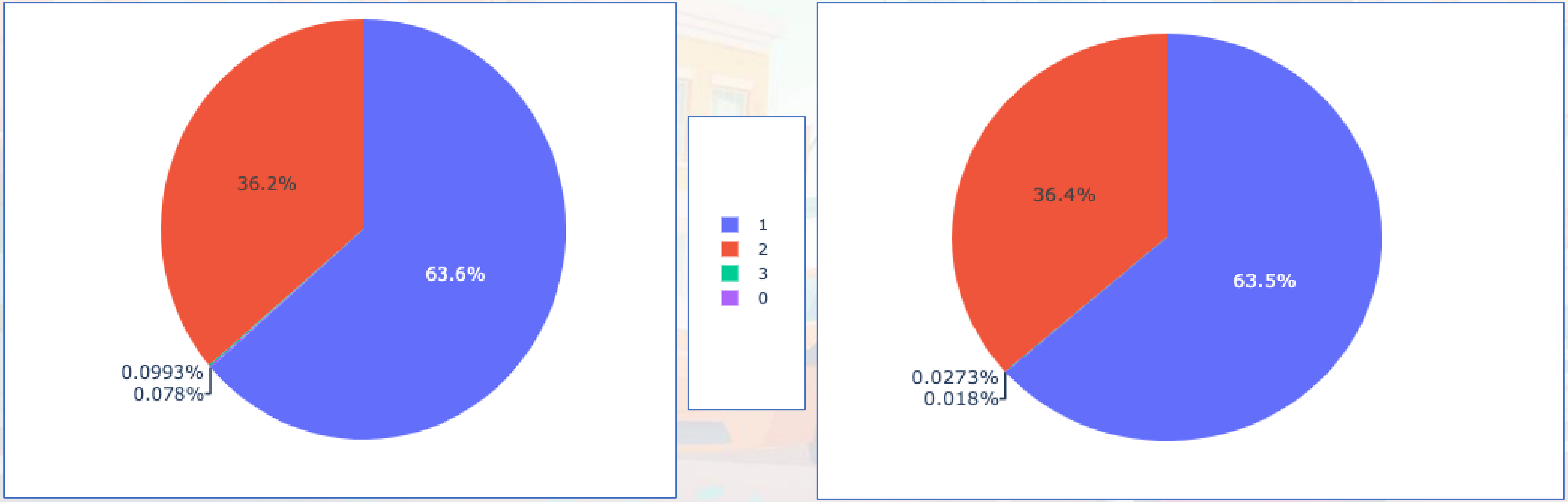
Increased Accuracy with Depth ?

Road Accidents…

Classification : Decision Tree Classifier -> Gini-> Observation

Test Set Severity Distribution

Classified Severity Distribution (Gini)



Road Accidents…

Classification : Decision Tree Classification

- The main reason behind increment of the depth is increasing accuracy is due to presence of continuous variables (having numbers not quantified properly) like State Code, County Code and Lattitude.
- Continuous variable are treated based on ***variance***
- So let's remove these variables from feature list and check..

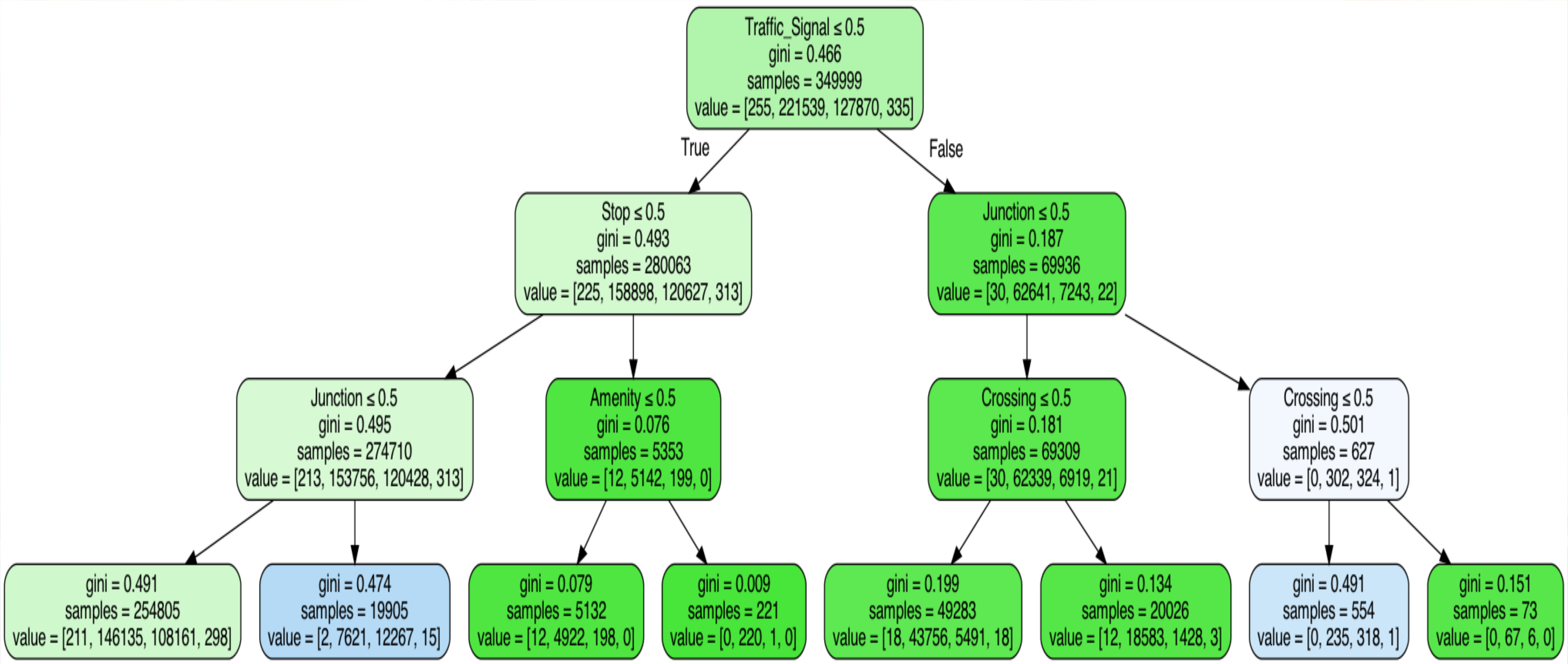
Road Accidents…

Classification : Decision Tree Classification -> Modified Gini

Accuracy using Gini Classifier using 70.0 % training and depth 1: : : : 0.6360133333333333
Accuracy using Gini Classifier using 70.0 % training and depth 2: : : : 0.6359333333333334
Accuracy using Gini Classifier using 70.0 % training and depth 3: : : : 0.64788
Accuracy using Gini Classifier using 70.0 % training and depth 4: : : : 0.6478933333333333
Accuracy using Gini Classifier using 70.0 % training and depth 5: : : : 0.6479
Accuracy using Gini Classifier using 70.0 % training and depth 6: : : : 0.6479
Accuracy using Gini Classifier using 70.0 % training and depth 7: : : : 0.6479
Accuracy using Gini Classifier using 70.0 % training and depth 8: : : : 0.6479
Accuracy using Gini Classifier using 70.0 % training and depth 9: : : : 0.6479
Accuracy using Gini Classifier using 70.0 % training and depth 10: : : : 0.6479
Accuracy using Gini Classifier using 70.0 % training and depth 11: : : : 0.6479
Accuracy using Gini Classifier using 70.0 % training and depth 12: : : : 0.6479
Accuracy using Gini Classifier using 70.0 % training and depth 13: : : : 0.6479
Accuracy using Gini Classifier using 70.0 % training and depth 14: : : : 0.6479
Accuracy using Gini Classifier using 70.0 % training and depth 15: : : : 0.6479
Accuracy using Gini Classifier using 70.0 % training and depth 16: : : : 0.6479
Accuracy using Gini Classifier using 70.0 % training and depth 17: : : : 0.6479
Accuracy using Gini Classifier using 70.0 % training and depth 18: : : : 0.6479
Accuracy using Gini Classifier using 70.0 % training and depth 19: : : : 0.6479
Accuracy using Gini Classifier using 70.0 % training and depth 20: : : : 0.6479

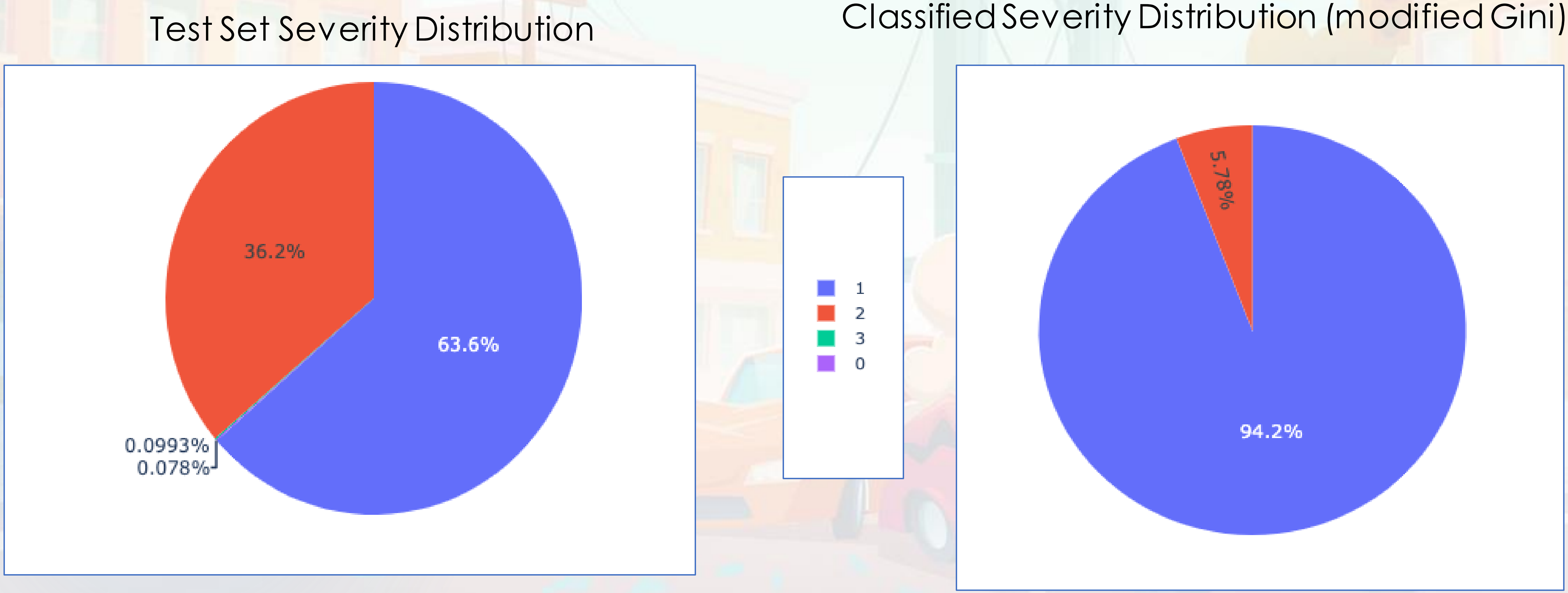
Road Accidents…

Classification : Decision Tree Classification -> Modified Gini



Road Accidents…

Classification : Decision Tree Classifier -> M. Gini-> Observation



Road Accidents…

Classification : Decision Tree Classification -> Result

- Geographical Information are important for classification (to be included in features)
- Around 86% of accuracy is achieved using decision tree classification

Road Accidents…

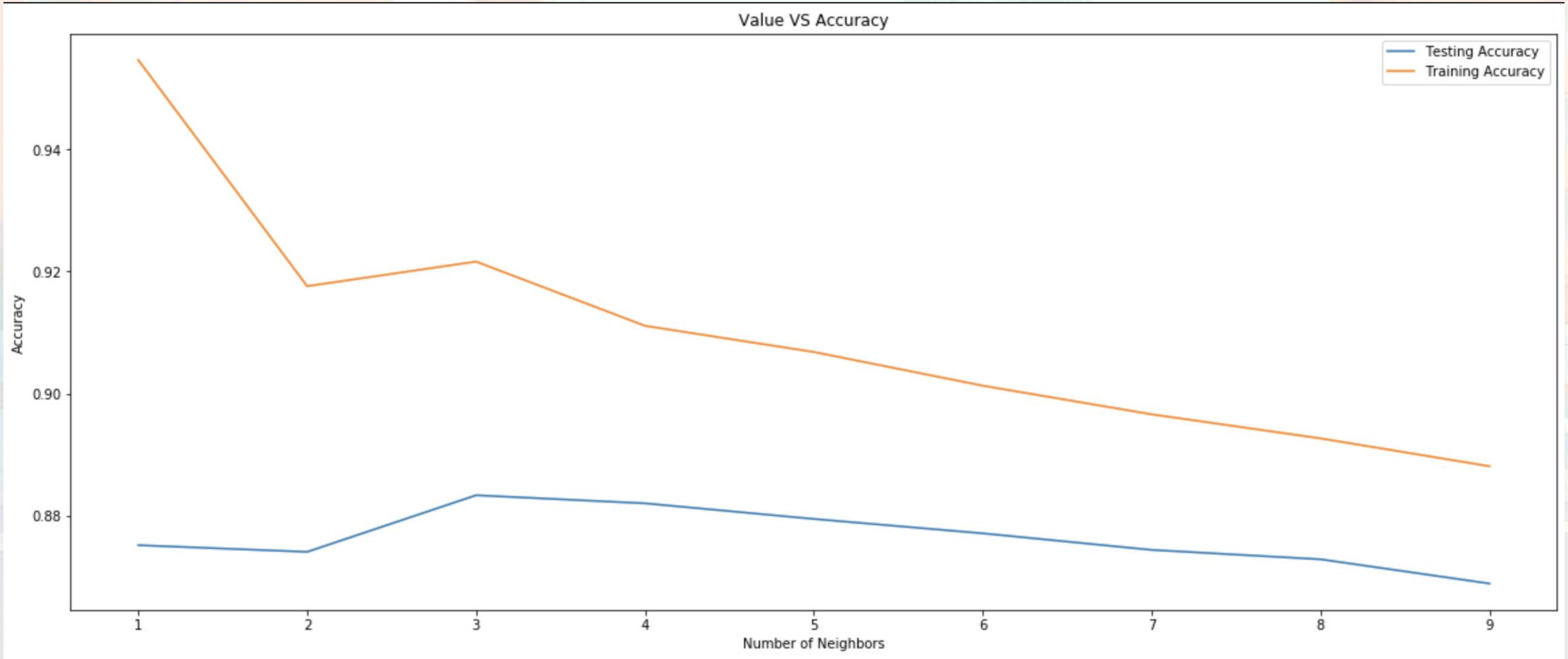
Classification : KNN Classification

Up to 14 neighbors KNN fitting was tested with following result

Accuracy Using KNN-Classifier with 70.0 % training and 1 neighbors is 0.87514:
Accuracy Using KNN-Classifier with 70.0 % training and 2 neighbors is 0.8740466666666666:
Accuracy Using KNN-Classifier with 70.0 % training and 3 neighbors is 0.8833266666666667:
Accuracy Using KNN-Classifier with 70.0 % training and 4 neighbors is 0.882:
Accuracy Using KNN-Classifier with 70.0 % training and 5 neighbors is 0.87944:
Accuracy Using KNN-Classifier with 70.0 % training and 6 neighbors is 0.87708:
Accuracy Using KNN-Classifier with 70.0 % training and 7 neighbors is 0.8743666666666666:
Accuracy Using KNN-Classifier with 70.0 % training and 8 neighbors is 0.8728266666666666:
Accuracy Using KNN-Classifier with 70.0 % training and 9 neighbors is 0.8688533333333334:
Accuracy Using KNN-Classifier with 70.0 % training and 10 neighbors is 0.86782:
Accuracy Using KNN-Classifier with 70.0 % training and 11 neighbors is 0.86432:
Accuracy Using KNN-Classifier with 70.0 % training and 12 neighbors is 0.8632066666666667:
Accuracy Using KNN-Classifier with 70.0 % training and 13 neighbors is 0.8596533333333334:
Accuracy Using KNN-Classifier with 70.0 % training and 14 neighbors is 0.85814:

Road Accidents…

Classification : KNN Classification



Road Accidents…

Classification : Decision Tree Classifier -> Gini-> Observation

Confusion Matrix N = 3

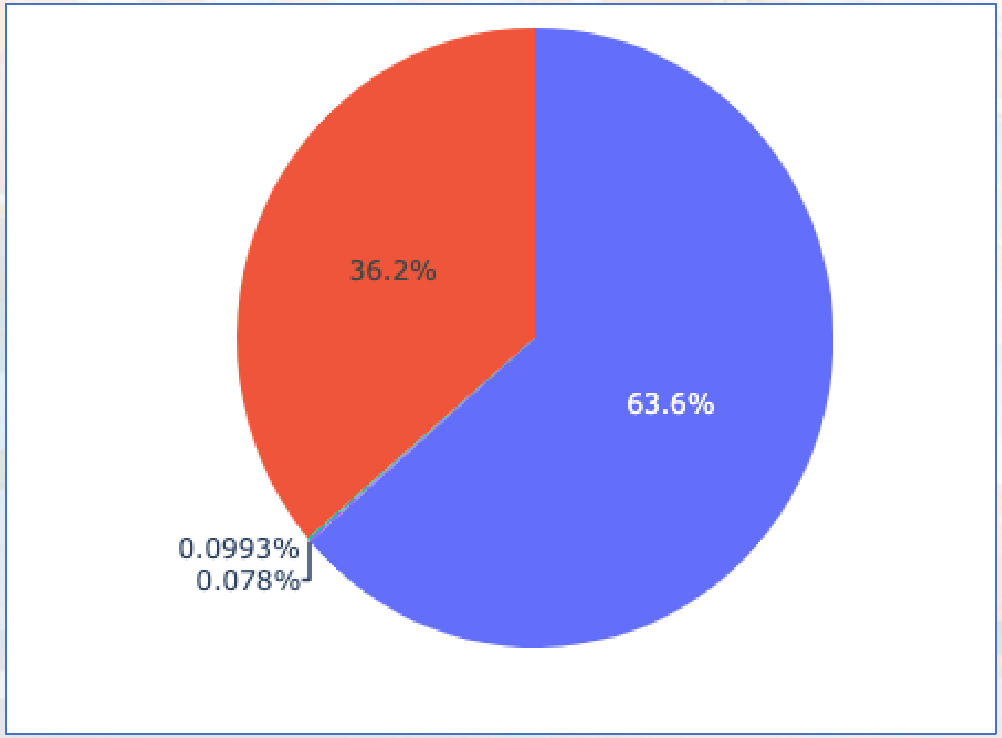
0	90	27	0
47	86538	8816	1
30	8284	45956	62
0	35	109	5

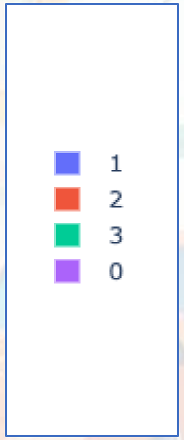
Maximum Accuracy with N = 3
(Criteria = Uniform)

Road Accidents…

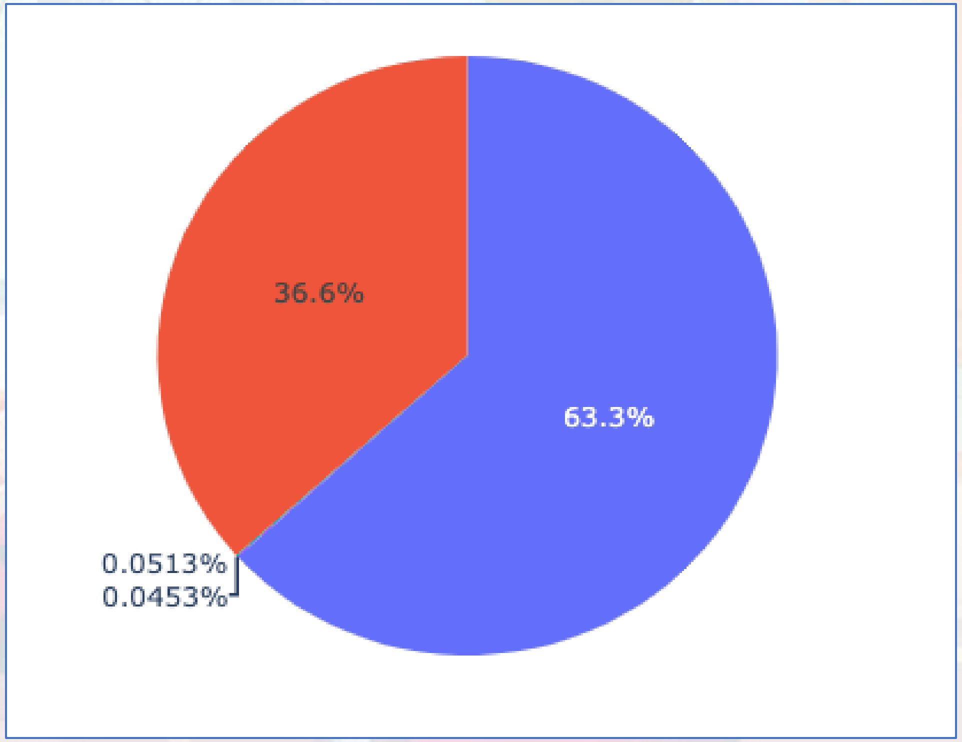
Classification : KNN Classifier -> Observation

Test Set Severity Distribution





Classified Severity Distribution (KNN)



Road Accidents…

Classification : Conclusion

- KNN and DT both are working great for classification.
- KNN better handled continuous variables like latt. and counties code.

Assumption

- Training set could be done either under/over sampling for better result

Road Accidents…

Time Series Forecasting

PRAMIL PAUDEL, SUMIT BHATTARAI

Road Accidents…

Data Preparation :

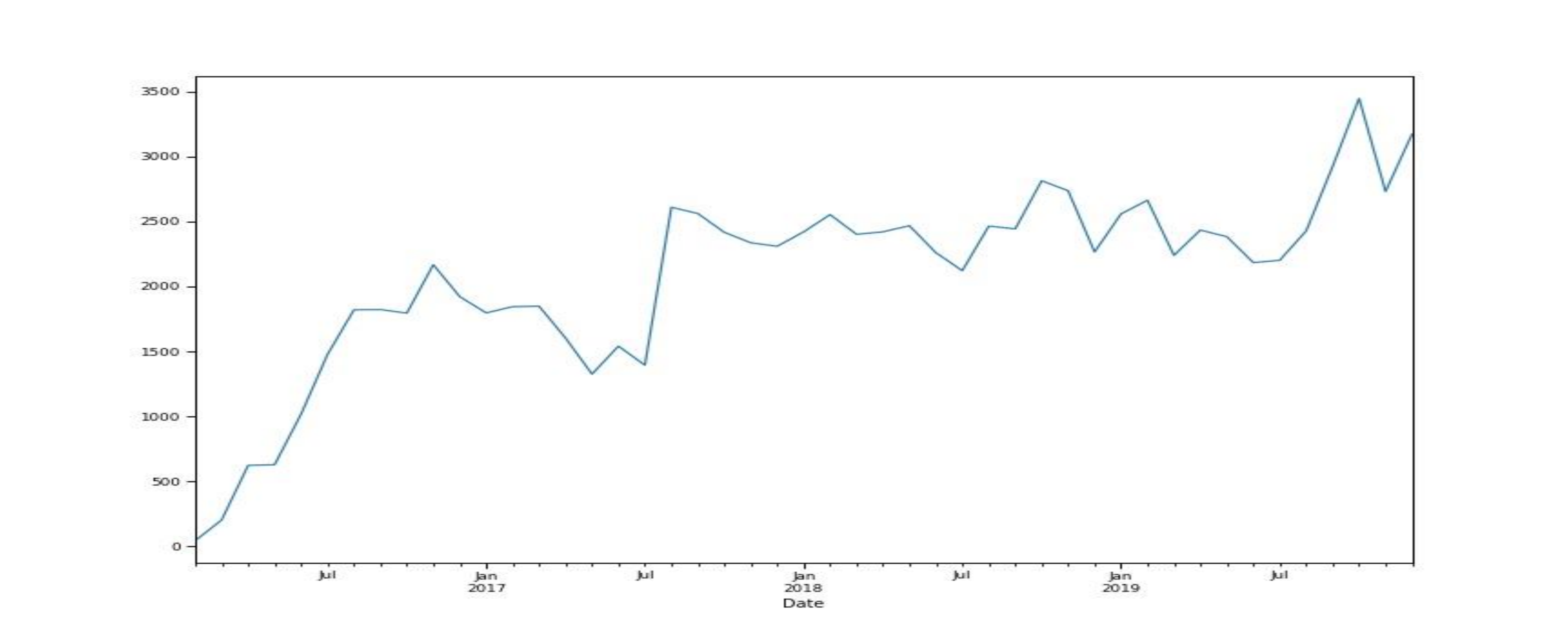
- Data is prepared by taking the aggregation of count based on the Accident Start Date property for year 2016 to 2019.
- There were significant number of accident in each day.
- Data is split into 70% train and 30% test dataset.

Procedure :

- Decomposition is used to detect the trend and seasonality of the data.
- ARIMA,SARIMAX and fbprophet are used to model the time series forecasting model.
- Grid Search is used to find the order (p,d,q) of the ARIMA model.
- Mean squared error is calculated for each p,d,q in range of (0,3) to find the order with least mean square error.
- Grid Search is used to find the parameters of SARIMAX with lowest AIC value.

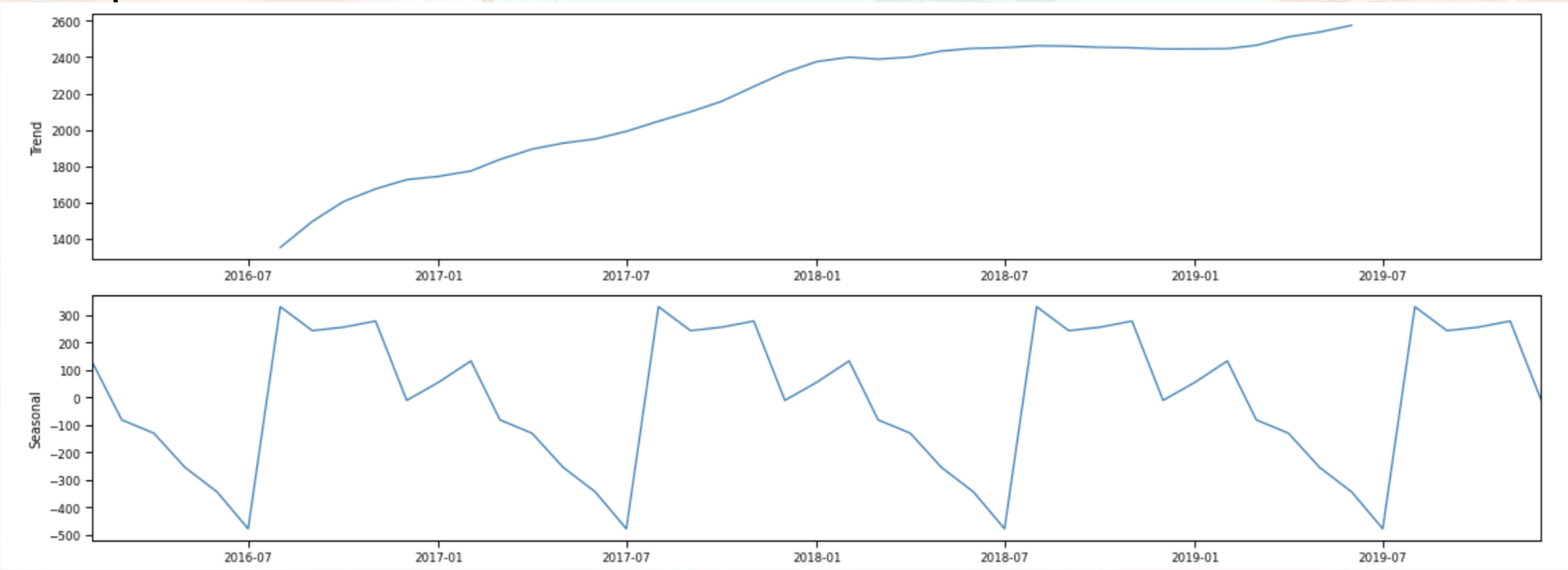
Road Accidents…

Moving Average Plot of the data:



Road Accidents…

Decomposition:



- The trend is increasing.
- There is a seasonality pattern in the data.

Road Accidents…

ARIMA model :

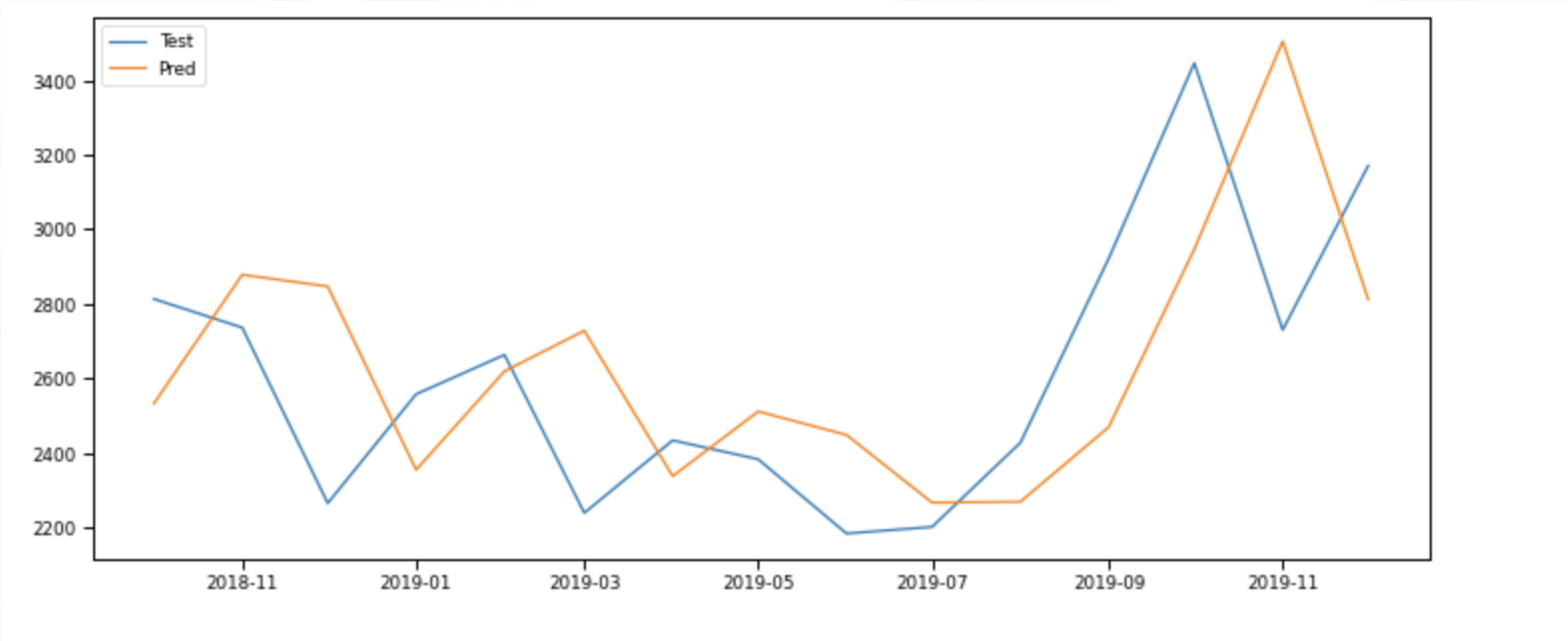
Best (p,d,q) = (2,1,1)

```
                              ARIMA Model Results
==============================================================================
Dep. Variable:                    D.y   No. Observations:                   46
Model:                 ARIMA(2, 1, 1)   Log Likelihood                -328.773
Method:                       css-mle   S.D. of innovations            301.797
Date:                Mon, 23 Nov 2020   AIC                            667.546
Time:                        14:53:05   BIC                            676.689
Sample:                             1   HQIC                           670.971

==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         52.3009     15.002      3.486      0.000      22.897      81.705
ar.L1.D.y      0.7624      0.151      5.046      0.000       0.466       1.059
ar.L2.D.y      0.0526      0.161      0.325      0.745      -0.264       0.369
ma.L1.D.y     -1.0000      0.068    -14.781      0.000      -1.133      -0.867
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.2106           +0.0000j            1.2106            0.0000
AR.2          -15.7182           +0.0000j           15.7182            0.5000
MA.1            1.0000           +0.0000j            1.0000            0.0000
-----------------------------------------------------------------------------
```

Road Accidents…

ARIMA model :



The predicted value shifts by a month.

Road Accidents…

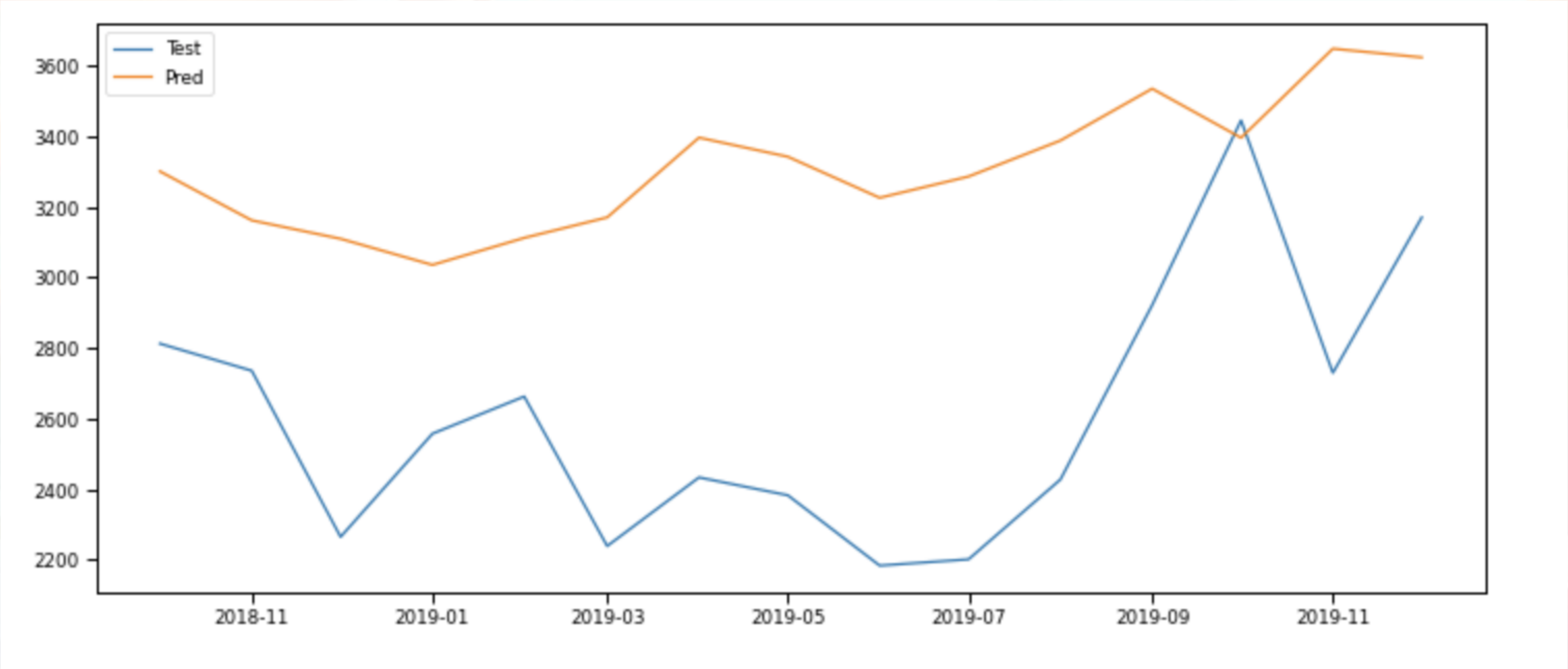
SARIMAX model :

Best params:
P,d,q = (0,1,1), seasonal param= (0,1,2,7)

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  47
Model:             SARIMAX(0, 1, 1)x(0, 1, [1, 2], 7)   Log Likelihood              -284.174
Date:                             Mon, 23 Nov 2020   AIC                           576.349
Time:                                     15:14:30   BIC                           583.003
Sample:                                          0   HQIC                          578.736
                                              - 47
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.0531      0.184     -0.289      0.773      -0.414       0.308
ma.S.L7       -0.9604      0.235     -4.082      0.000      -1.421      -0.499
ma.S.L14       0.2800      0.286      0.981      0.327      -0.280       0.840
sigma2      1.089e+05   2.81e+04      3.875      0.000    5.38e+04    1.64e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.10   Jarque-Bera (JB):                 6.59
Prob(Q):                              0.75   Prob(JB):                         0.04
Heteroskedasticity (H):               1.46   Skew:                             0.80
Prob(H) (two-sided):                  0.50   Kurtosis:                         4.23
===================================================================================
```

Road Accidents…

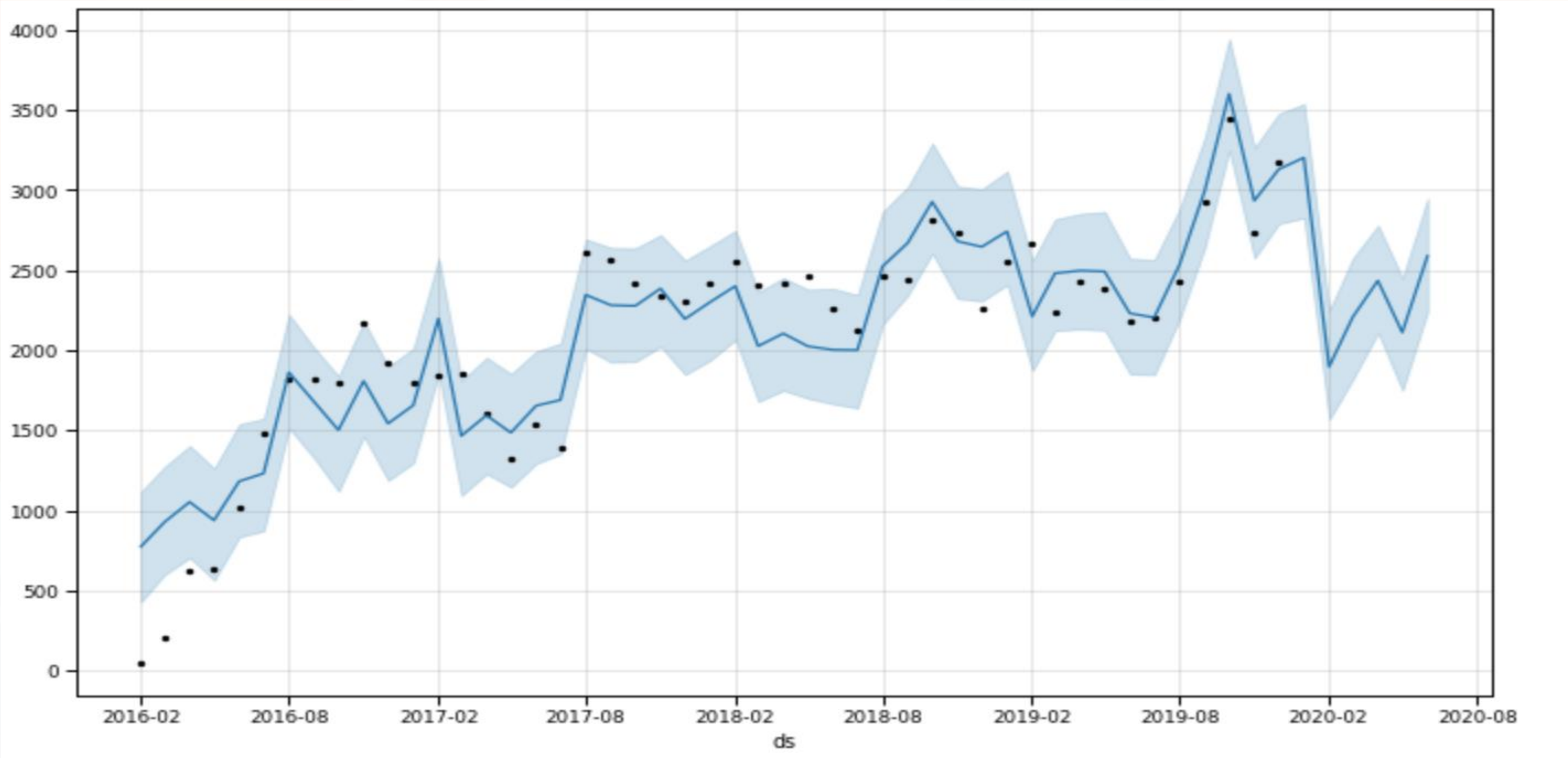
SARIMAX model :



The predicted values are much higher than the expected values.

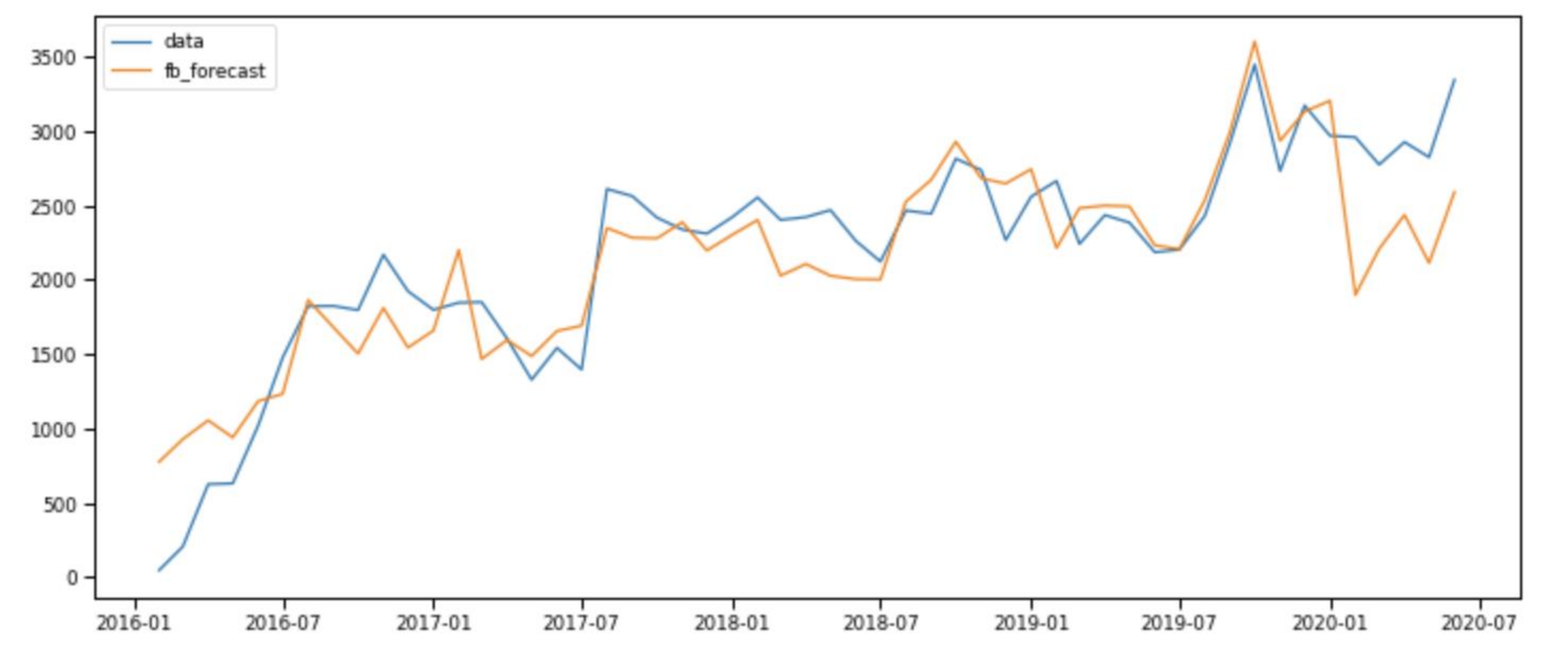
Road Accidents…

Facebook Prophet :



Road Accidents…

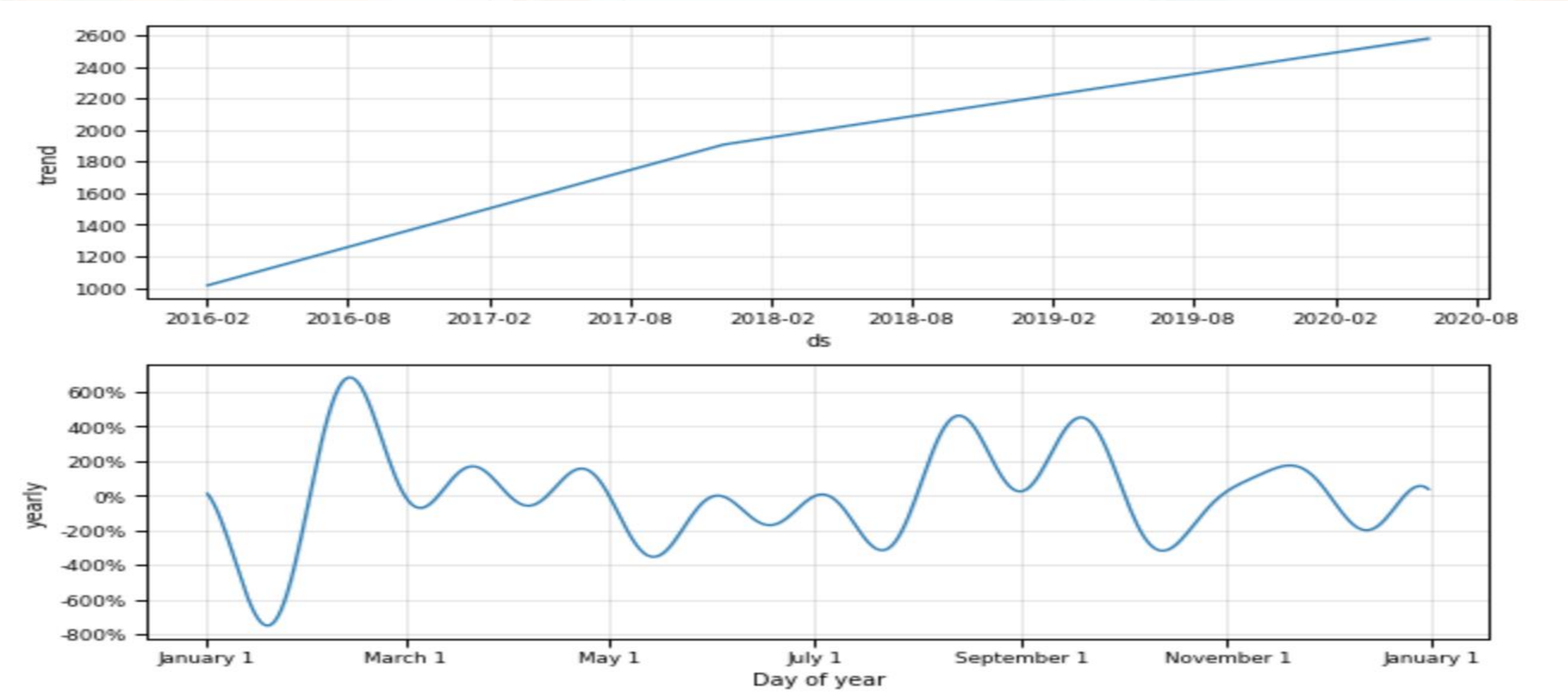
Facebook Prophet :



- The predicted result for next six month is less than the expected value.
- The pattern of the test dataset is preserved by the model.

Road Accidents…

Facebook Prophet :



Road Accidents…

Conclusion :

- ARIMA model has predicted value that is one step/month further. Some performance tuning might help.
- SARIMAX is predicting value much higher than the test value preserving some pattern.
- Facebook Prophet predicts value which is less than the actual test value and follows the test pattern. So, it seems Facebook Prophet is currently working better for this dataset.

Road Accidents…

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.*

Road Accidents…

Thank you !!

PRAMIL PAUDEL, SUMIT BHATTARAI