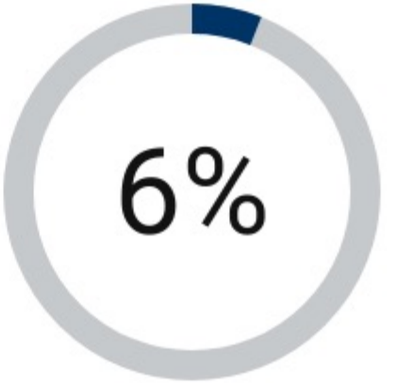
Car Accidents in the US

A presentation by Charlie Lee, Yini Zhong and Timothy Rivers

Background

1 in 55

Americans will get in a motor vehicle accident this year. That is a total average of approximately 6 million.



Deaths

Roughly, 6% of all accidents result in death.

Permanent Injury

An average of 30% of all accidents result in permanent injury

Dataset and Preparation

1

Kaggle Dataset

The dataset was collected from Kaggle, titled "US Accidents"

2

US Car Accidents

The dataset contains over 3 million records spanning from 2016-2019

3

Multiple Variables

The dataset includes over 40 variables, including: Time, State, Precipitation, Temperature, and Wind Speed.

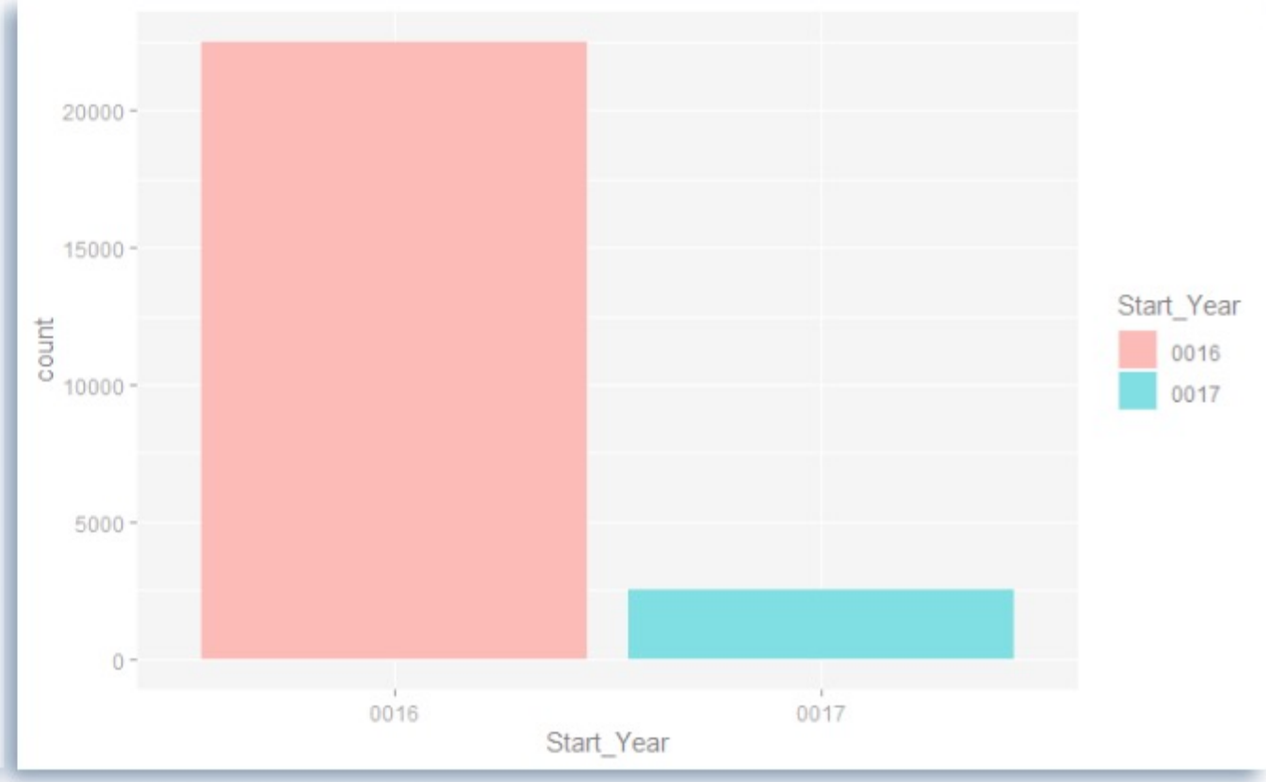
4

Severity

Each record is ranked from 1 to 4 based on a "Severity" score. This is calculated by impact on traffic.

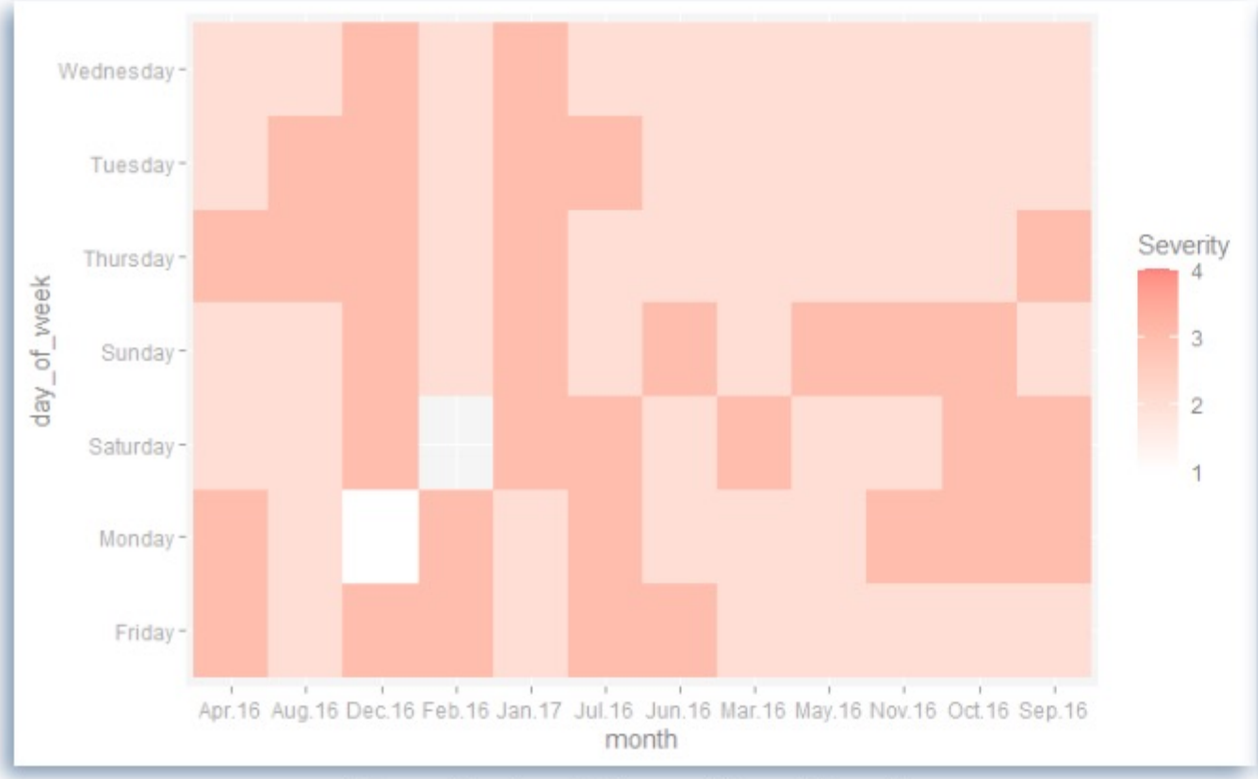
Number of Accidents

Number of accidents by State

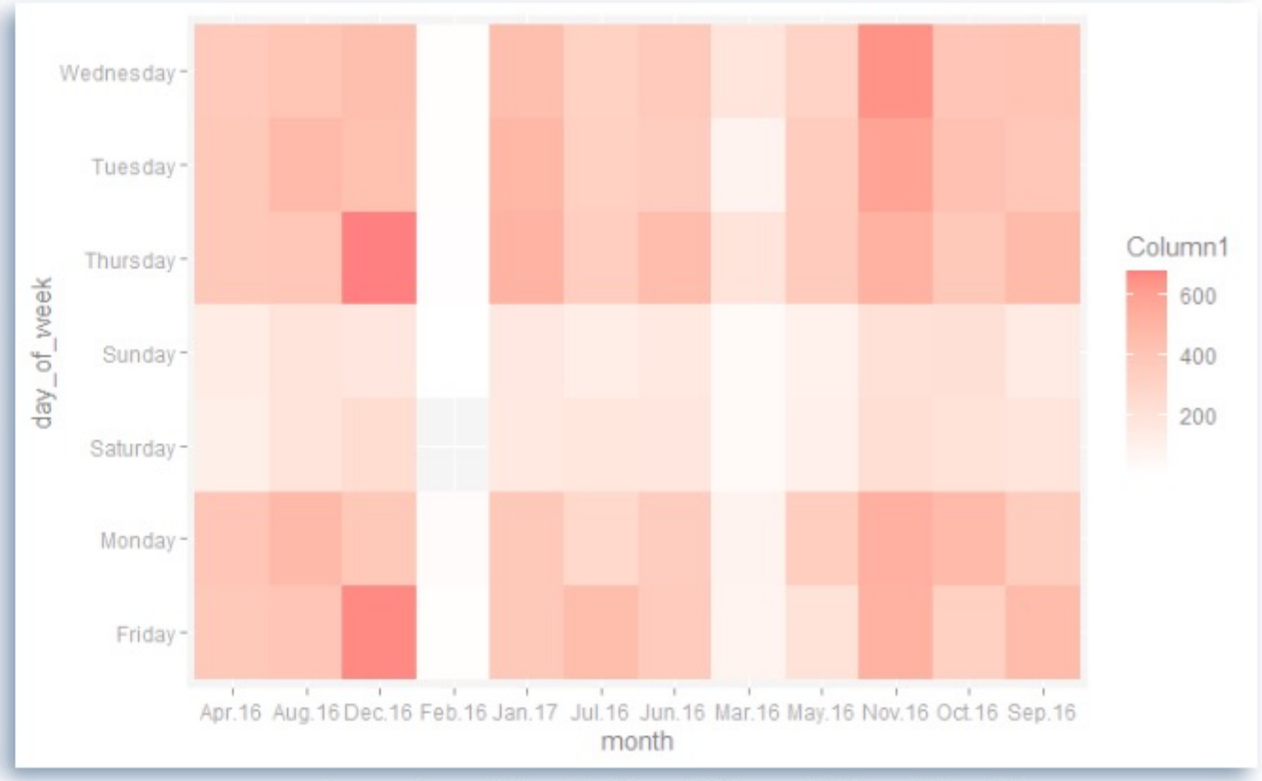


Number of accidents by year

Visualize the differences in Severity and number of accidents by Day/Month

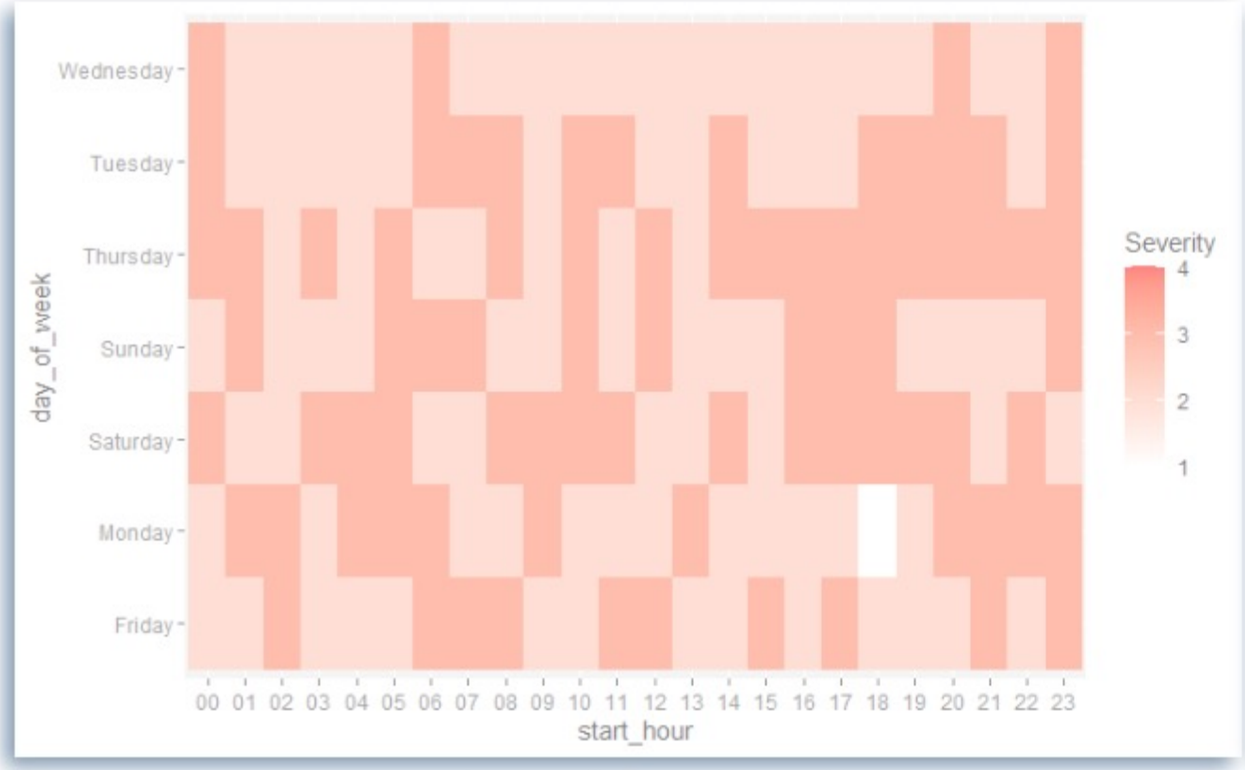


Severity for different Day/Month

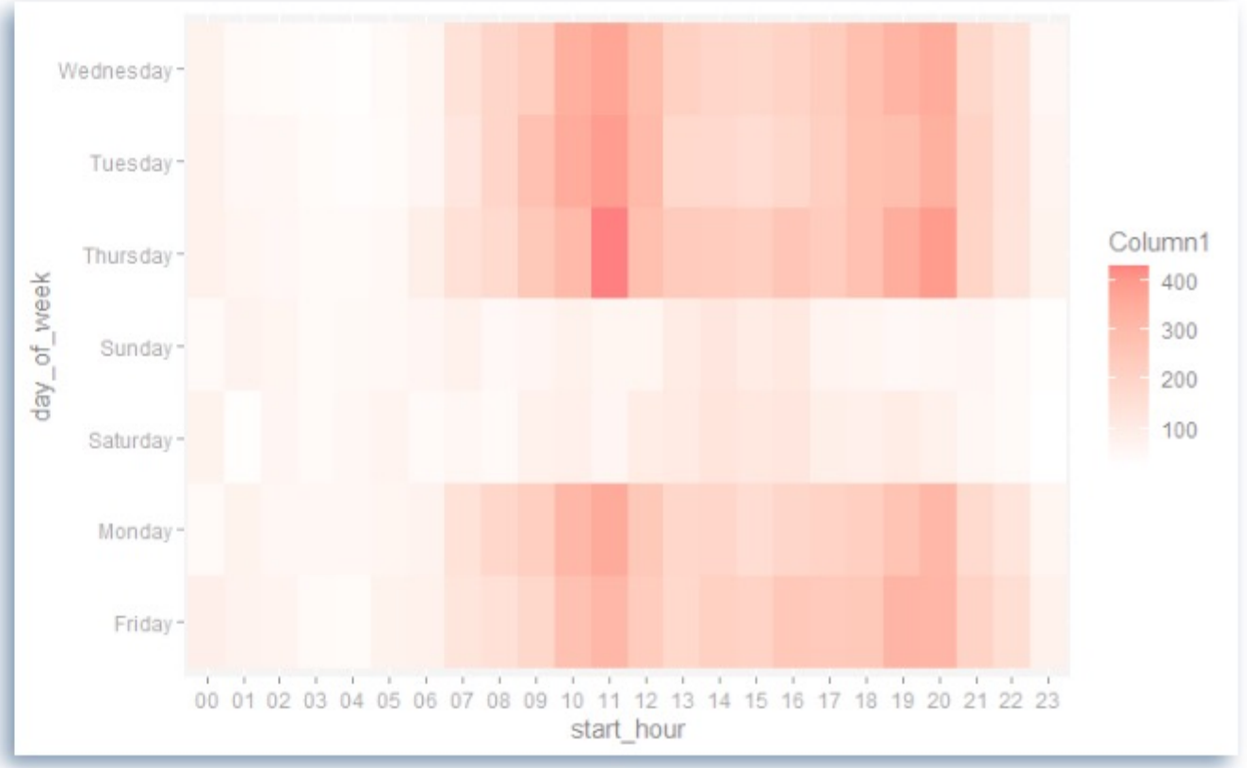


number of accidents for different Day/Month

Visualize the differences in Severity and number of accidents by Hour/Day

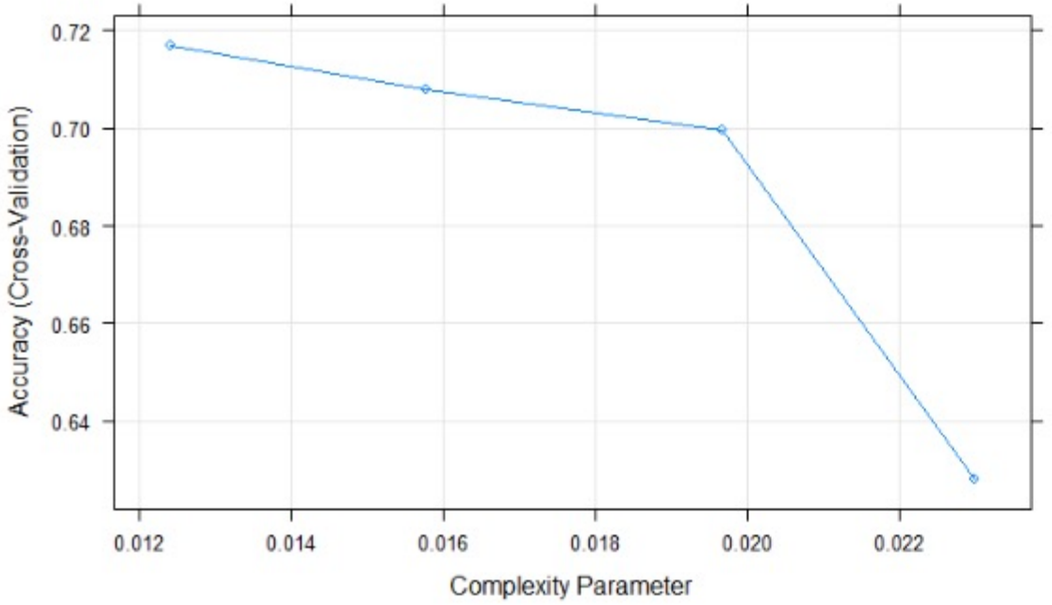


Severity for different Hour/Day

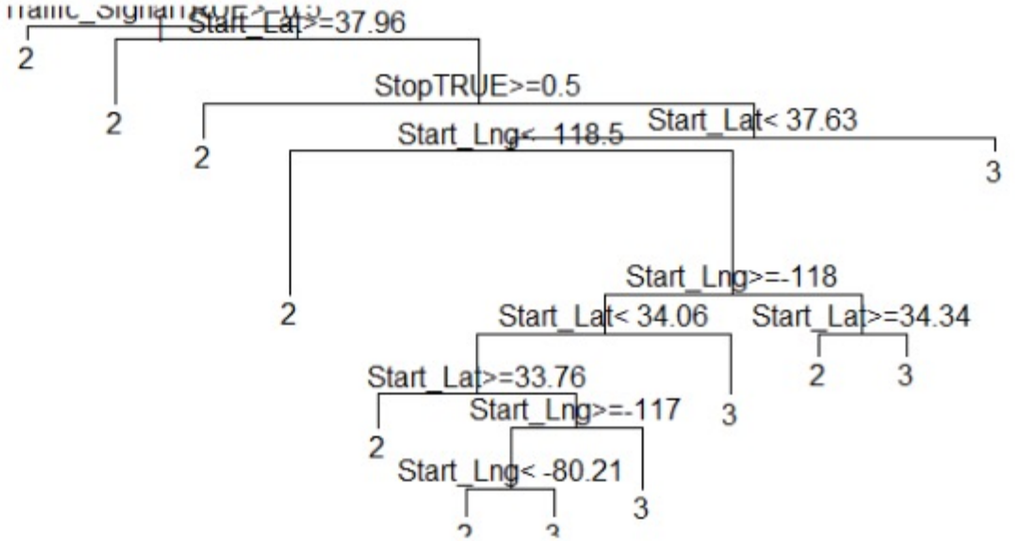


Number of accidents for different Hour/Day

Decision Tree Method



Trend of CP Value



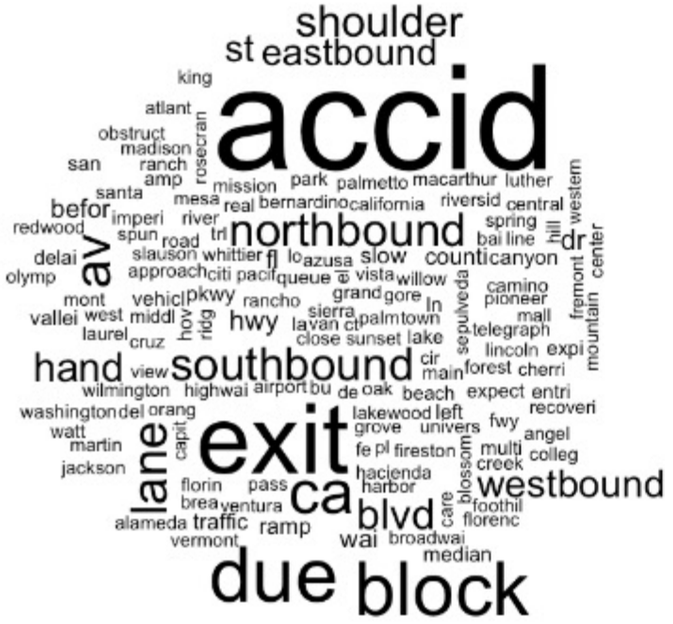
Decision Tree Method

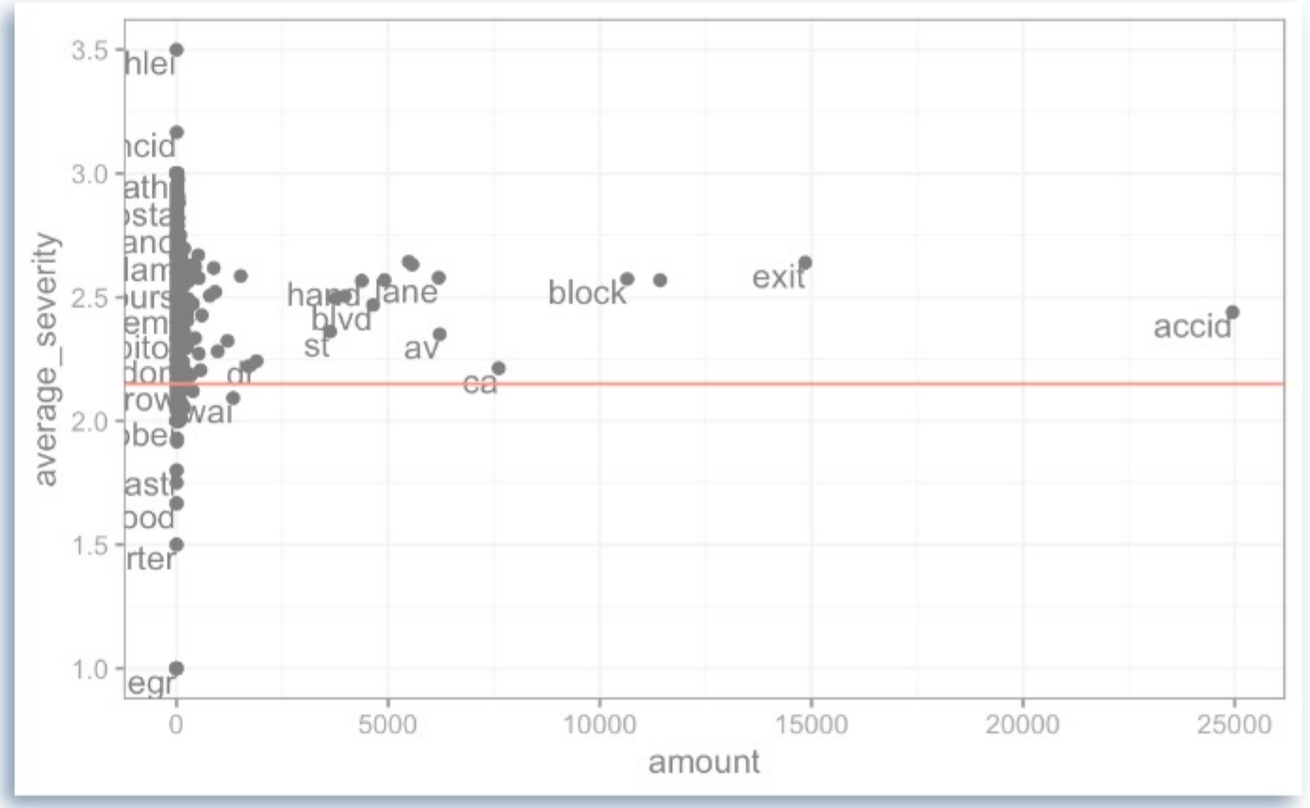
```
   user  system elapsed
  26.39    0.26   26.87
CART

25000 samples
   28 predictor
    4 classes: '1', '2', '3', '4'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 22500, 22500, 22500, 22500, 22499, 22500, ...
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.01242067  0.7125194  0.4230913
  0.01577516  0.7061587  0.4102005
  0.01967362  0.6973582  0.3920254
  0.02298277  0.6400839  0.2311121

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.01242067.
```

```
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3    4
         1    0    0    0    0
         2   13 9988 3172    6
         3    6 3982 7826    7
         4    0    0    0    0

Overall Statistics

               Accuracy : 0.7126
                 95% CI : (0.7069, 0.7182)
    No Information Rate : 0.5588
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4221
```

Text Mining

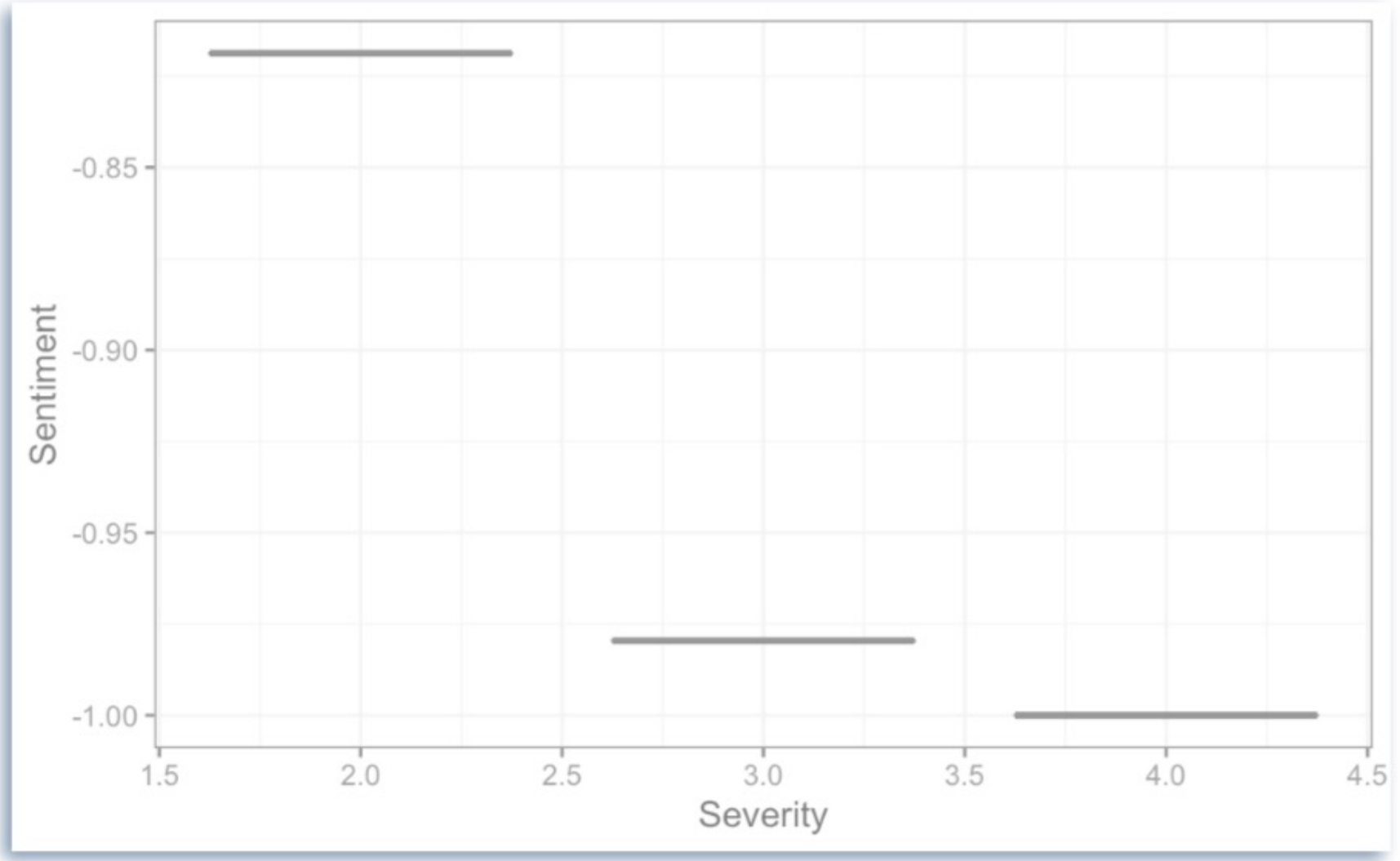
After applying tokenization, stemming and stopwords, we finally found the frequency of words. As you can see, the 3 most frequent words are "accident", "exit", "block". which means most accidents happened at the exits and blocked the traffic



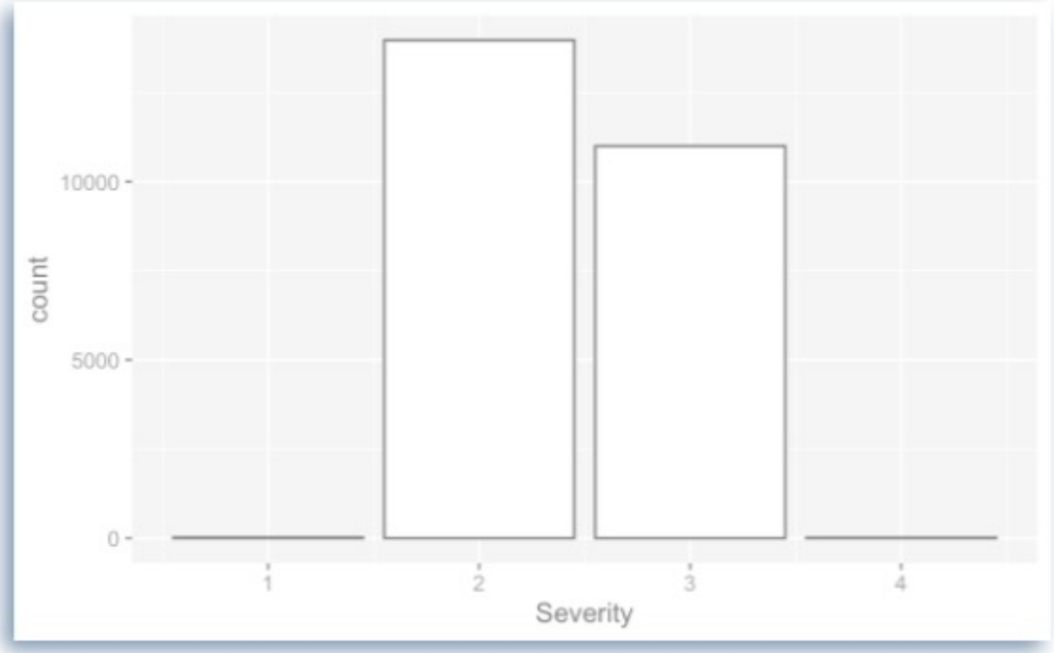


Text Mining

We can see the sentiment scores drops while the Severity increases



Text Mining



Groups 1 and 4 had too few data.

multinomial_naive_bayes

```
Confusion Matrix and Statistics

          Reference
Prediction Minor Server
    Minor      0      0
    Server  1478   1869

               Accuracy : 0.5584
                 95% CI : (0.5414, 0.5753)
    No Information Rate : 0.5584
    P-Value [Acc > NIR] : 0.5072

                  Kappa : 0
```

SVM Linear

```
Confusion Matrix and Statistics

          Reference
Prediction Minor Server
    Minor     89     41
    Server  1389   1828

               Accuracy : 0.5728
                 95% CI : (0.5558, 0.5896)
    No Information Rate : 0.5584
    P-Value [Acc > NIR] : 0.049

                  Kappa : 0.0423
```

ML Prediction

Location: Start_Lng, Start_Lat, distance.mi

Weather: Temperature.F, Humidity..., Pressure.in, Visibility.mi,
Weather_Condition, Amenity

Road Condition: Bump, Crossing, Give_Way, Junction, No_Exit, Railway, Roundabout,
Station, Stop, Traffic_Calming, Traffic_Signal, Turning_Loop

Time: Sunrise_Sunset

```
feature_lst <- c('Severity','Start_Lng','Start_Lat','Distance.mi.','State','Temperature.F.','Humidity...','Pressure.in.',
'Visibility.mi.','Weather_Condition','Amenity','Bump','Crossing','Give_Way','Junction','No_Exit','Railway','Roundabout','Station','Stop','Traffic_Calming','Traffic_Signal','Turning_Loop','Sunrise_Sunset'
)
```

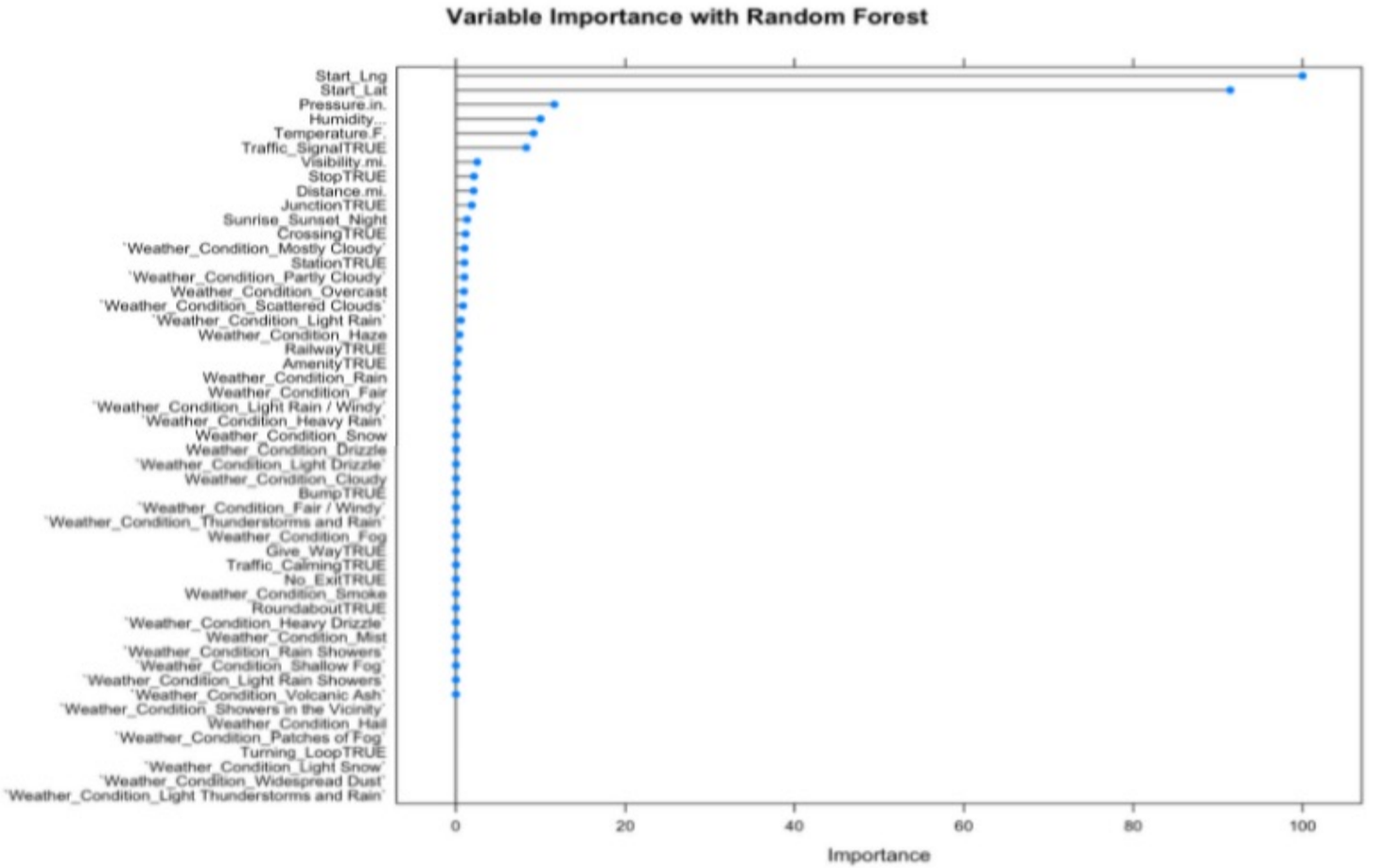
ML Predition

```
Random Forest

15182 samples
   51 predictor
    4 classes: '1', '2', '3', '4'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 13666, 13664, 13664, 13663, 13664, 13663, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.5536384  0.03120531
  26    0.8873645  0.77383588
  51    0.9012628  0.80185090

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 51.
```

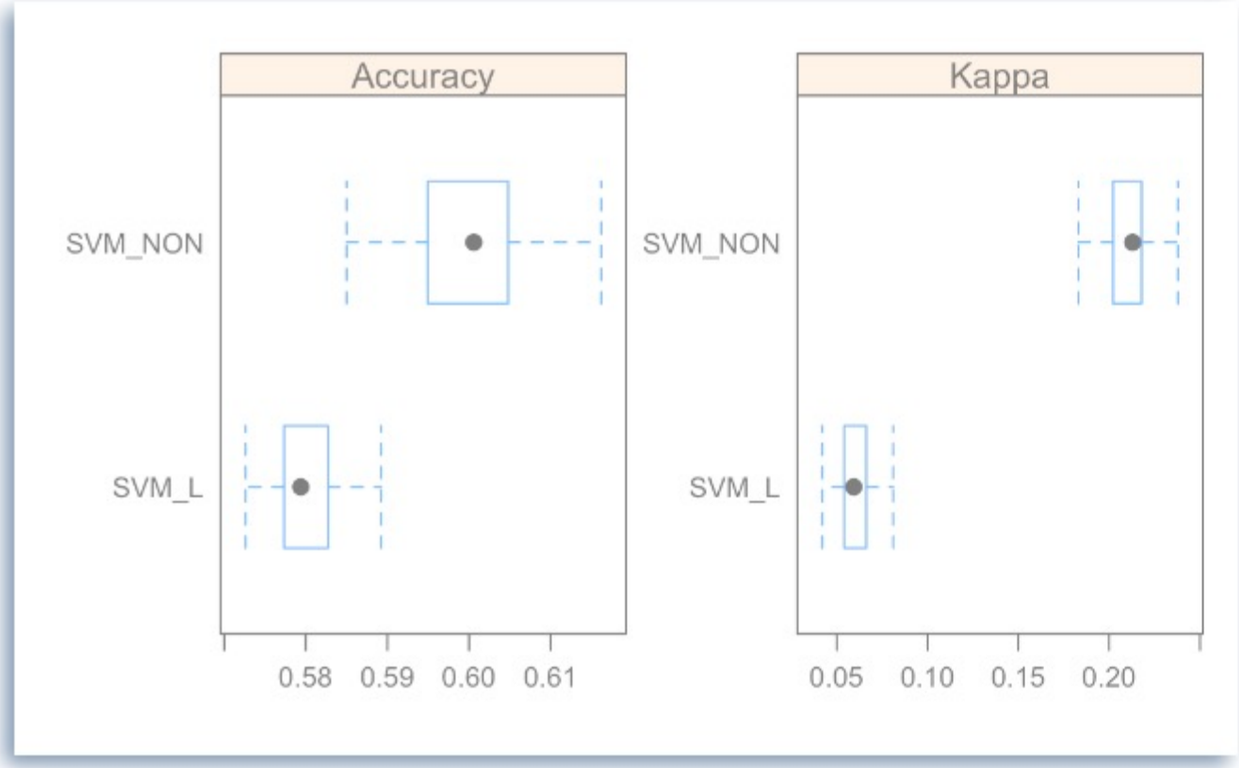
```
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3    4
         1    0    0    0    0
         2    4 3168  260    0
         3    1  361 2709    1
         4    0    0    0    0

Overall Statistics

                Accuracy : 0.9036
                  95% CI : (0.8962, 0.9107)
    No Information Rate : 0.5426
    P-Value [Acc > NIR] : < 2.2e-16

                   Kappa : 0.8065
```

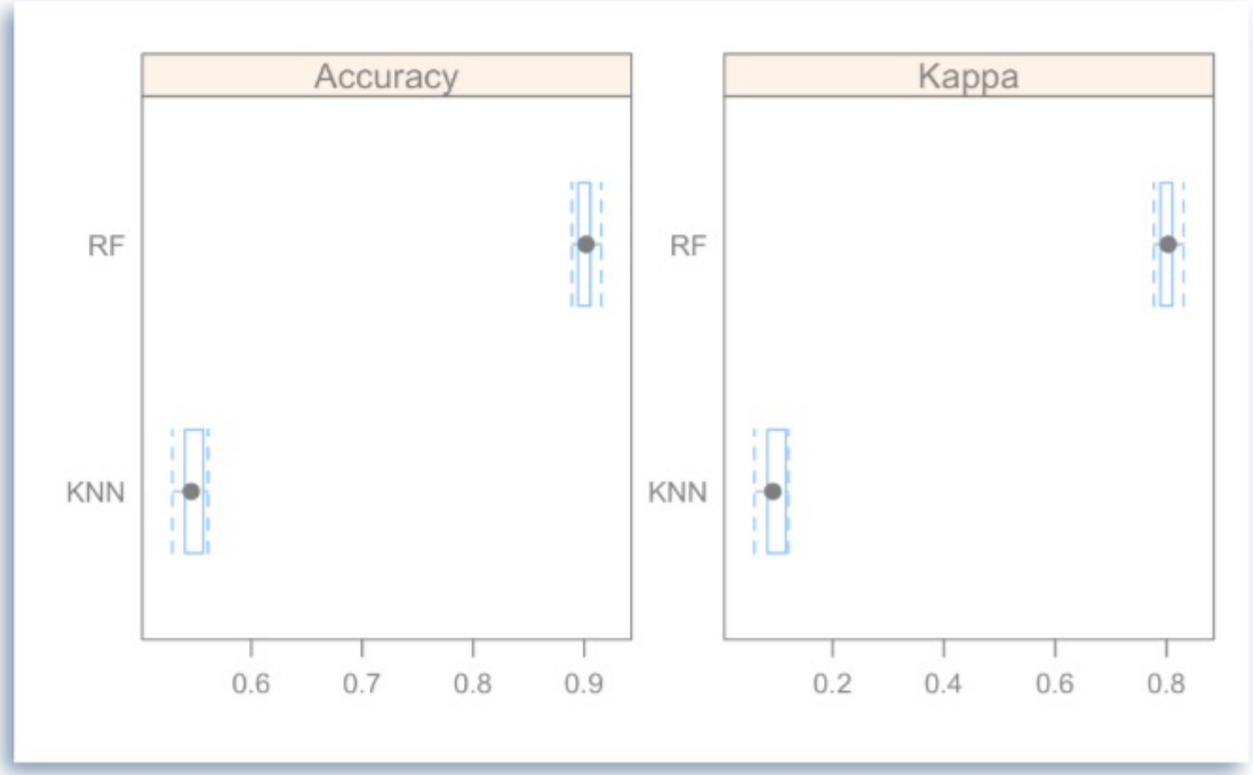
ML Predition

	Overall <dbl>
Start_Lng	100.0000000
Start_Lat	91.4666615
Pressure.in.	11.6220914
Humidity...	9.9948449
Temperature.F.	9.1810136
Traffic_SignalTRUE	8.3172216
Visibility.mi.	2.5071577
StopTRUE	2.1309127
Distance.mi.	2.0847302
JunctionTRUE	1.8732345



ML Prediction





Only Random Forest
got the best result.

Key Take Aways

1

Initial Observations

- Saturday and Sundays had the lowest number of accidents
- Right before noon and around 8:00 PM had the worst
- Severity based on traffic may not be the most appropriate indicator

2

Predictions

- Precipitation was not a strong indicator that severity would increase
- Accuracy is low, though random forests had the highest accuracy

3

Next steps

- Find ways to decrease the number of variables
- Retrain the models
- Predictive models for other variables, such as state or time of day

Questions

THANK YOU

Are there any questions?