Colorado Crash Analysis

By Dikshya Upreti

About the Dataset

- The data was extracted from Colorado Governmental Website which initially consisted of 81 columns and 86445 rows about traffic incidents from 2020.

- The dataset consists of informations like drivers age,weather,speed,location(county,city of crash),lightning, dui status,severity of the incidents and more.

- We will be dropping the columns with missing and null values and other columns that will not add value.

- EDA is done in Jupyter notebook using several libraries like pandas,numpy,matplotlib.

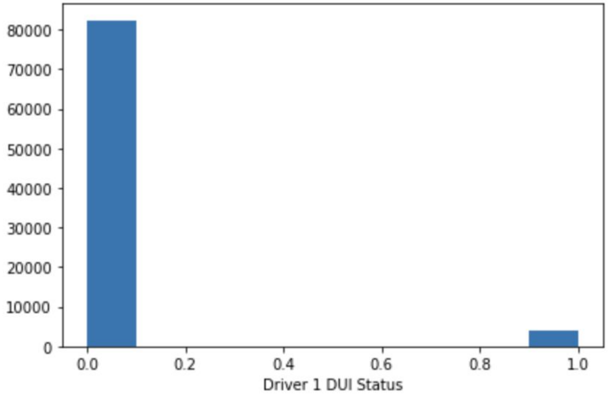
- Some basic visualizations are done in jupyter notebook.

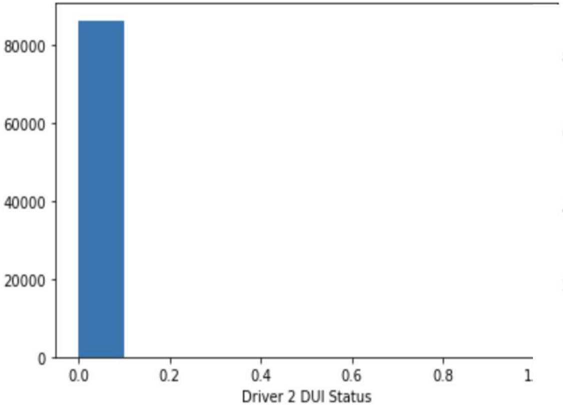
- Interactive visualization and dashboard are made in Tableau.

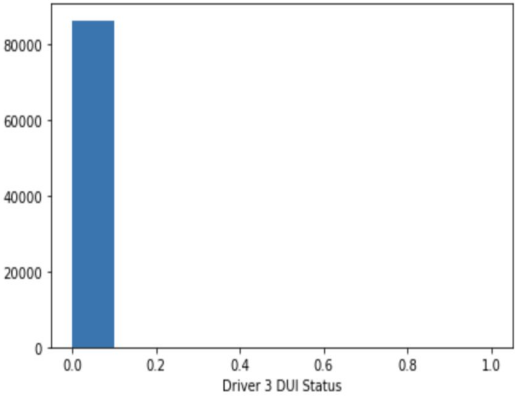
Research Questions

- Is Driver Impairment a major factor for the traffic incidents?
- What month in year does most accidents happen overall?
- What city system,weather and lightning condition has most incidents?
- What is the severity of the accidents overall?
- What ages are involved in most crashes?
- Did most accidents happen due to speeding?

Is Driver Impairment a major factor for the traffic incidents?

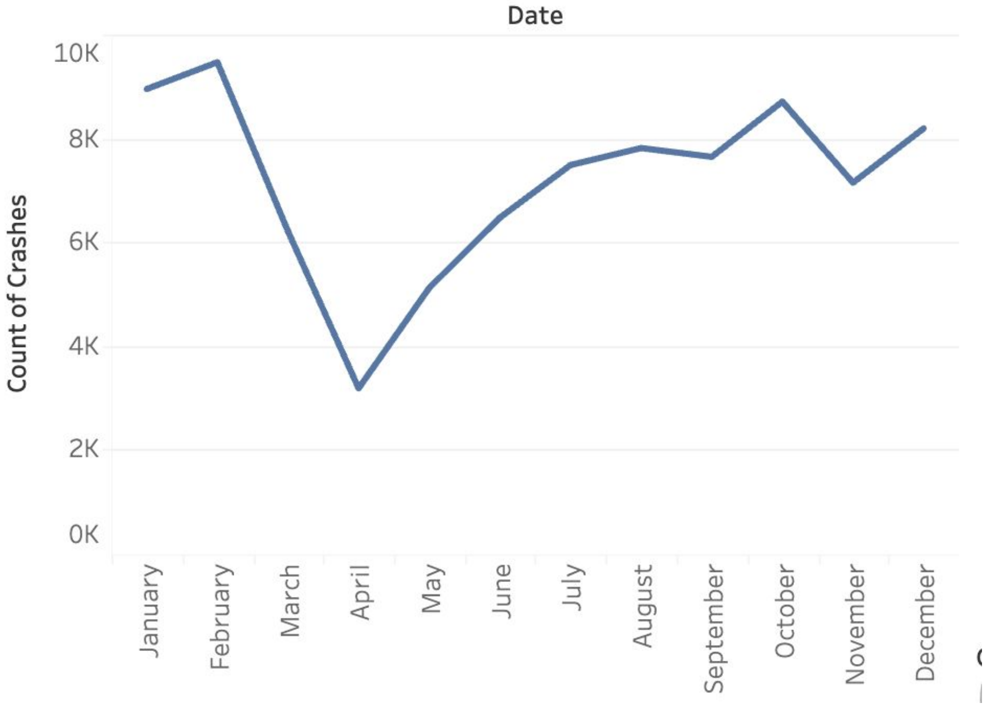






In the histograms above we are looking at 3 different drivers.On all 3 cases about 99% of accidents happen when drivers are not impaired in any means.This means driver impairment is not a major factor for traffic incidents.

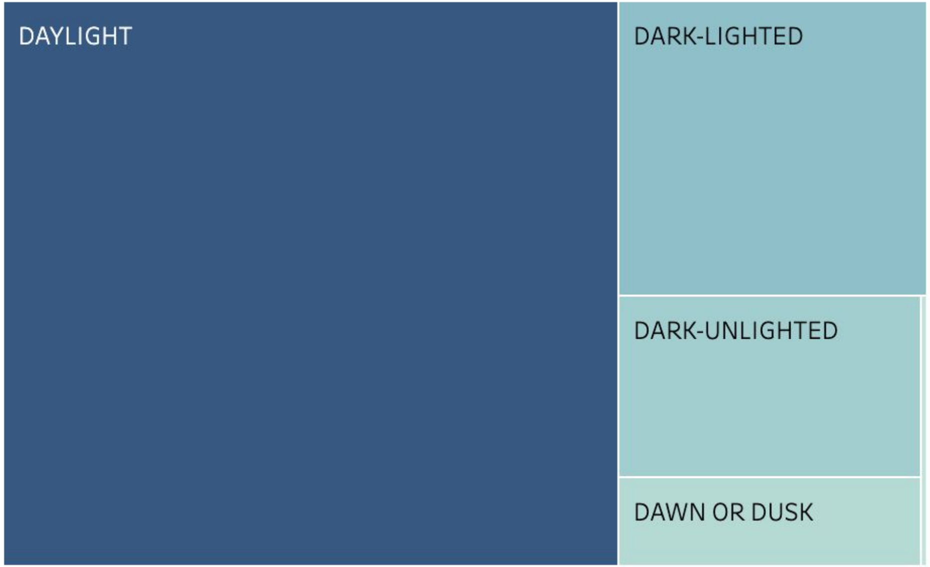
What month in year does most accidents happen overall?



Looking at the time series data we see that most accidents happen towards Jan,Feb and least during summer.This could possibly be because Jan,Feb have poor roads due to heavy snow.

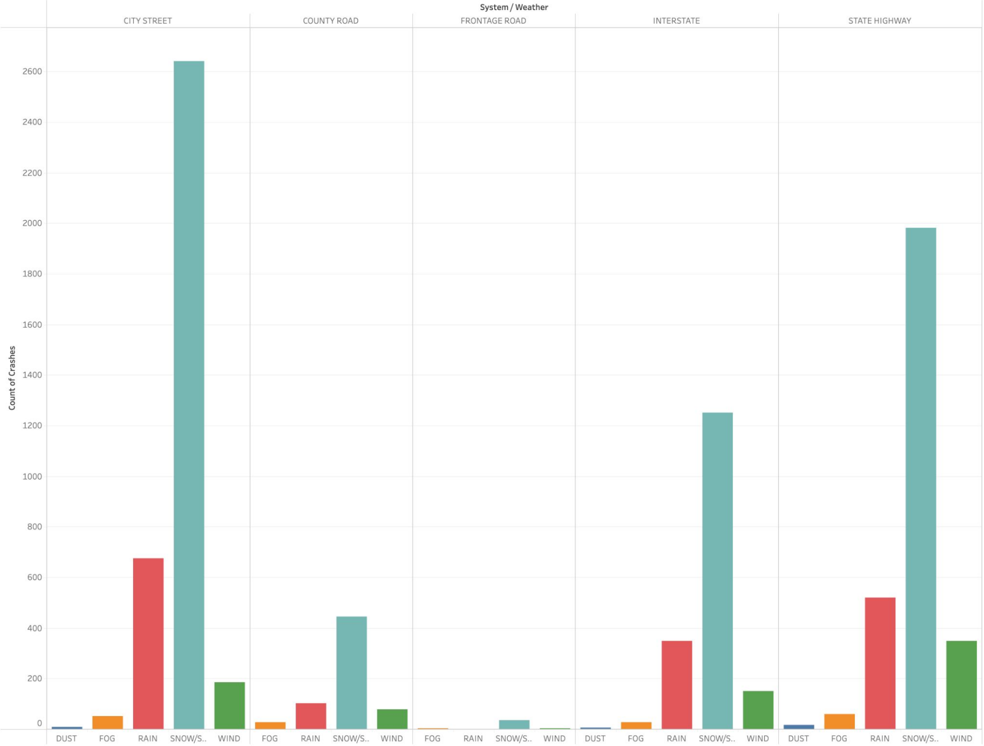
What city system,weather and lightning condition has most incidents?

Lightning Conditions



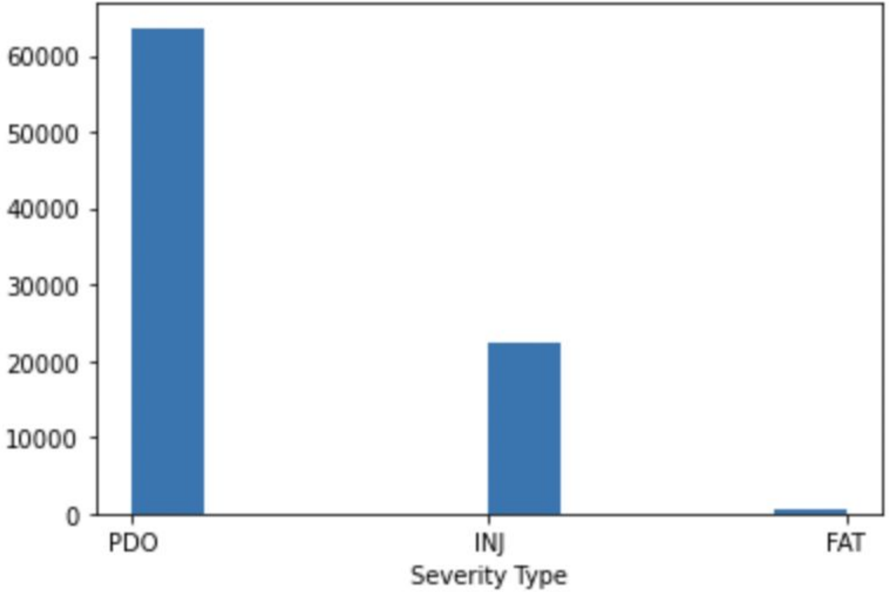
Looking at Treemap we can see that Daylight is when most accidents happen.This can be potentially because roads are most busy due to commute.

City System and weather



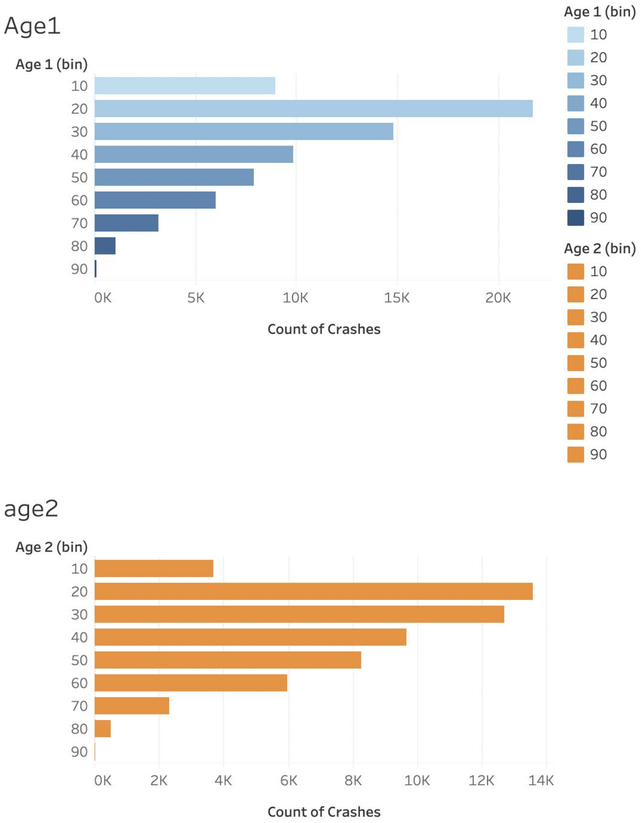
Its evident that most incidents happen in city roads during snow as assumed by our previous findings.

What is the severity of the accidents overall?



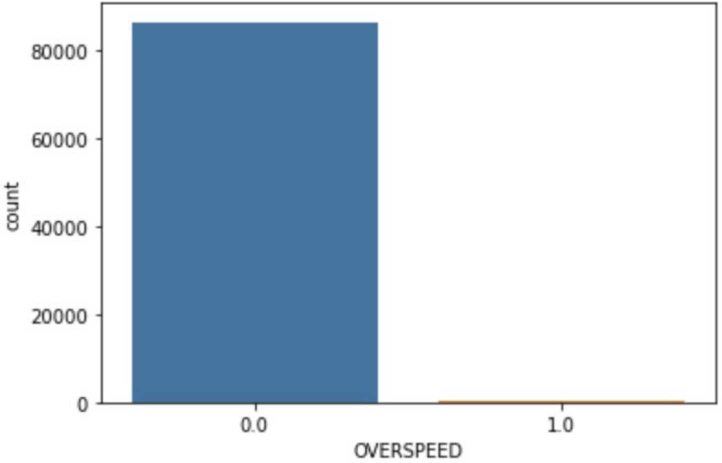
From the above visualizations we can see that most incidents were properly damage only.

What ages are involved in most crashes?



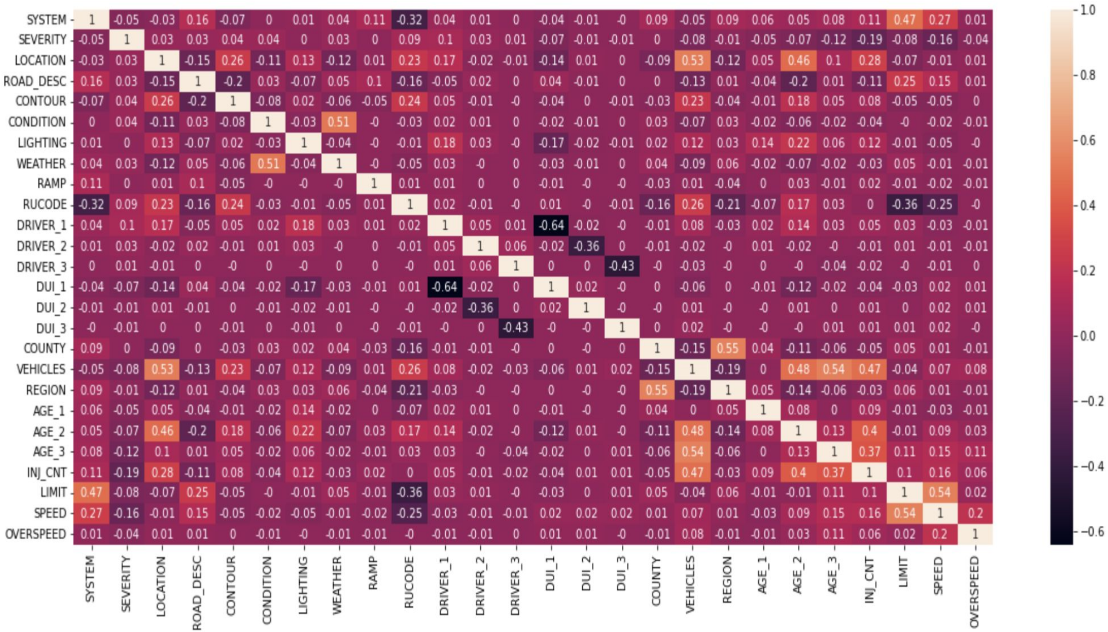
Looking at ages for both drivers involved in accidents,its evident that most drivers are in 20s and 30s who are presumably new drivers.

Did most accidents happen due to speeding?



After calculating the posted limit and the speed of vehicles we divided speed into two categories.1 represents overspeed and 0 represents driving in limit.

Heatmap Findings



Observations:

- Speed is highly positively correlated with
  Limit, System and negatively correlated with RUCode.
- Limit is highly Positively correlated with System and Road_Desc and negatively correlated with RUCode.
- Injury count is highly positively correlated with no.of vehicles involved in accident and age of the drivers.
- RUCode is highly -vely correlated with system, limit & speed.

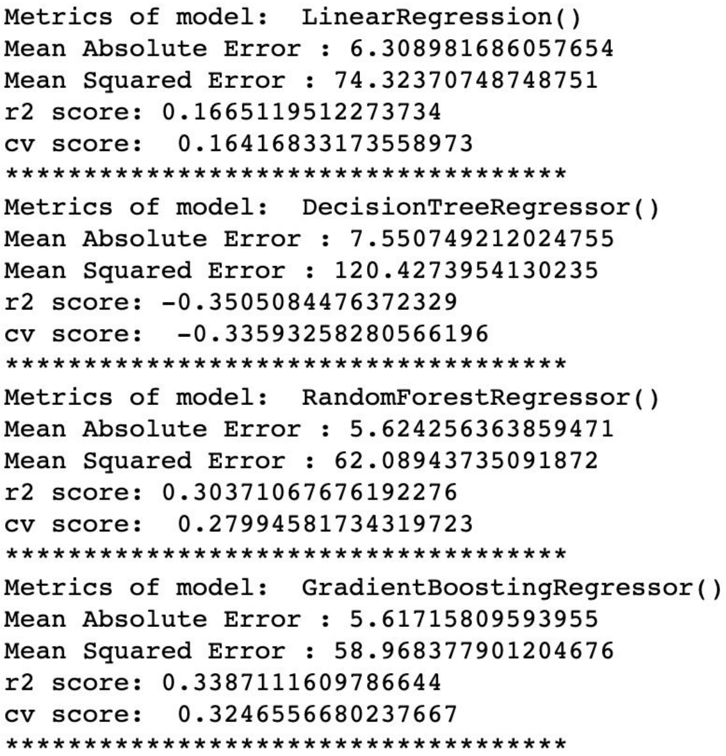
(Note: These are the major useful observations remaining have no much impact or related to each other.)

Target Variable : Speed

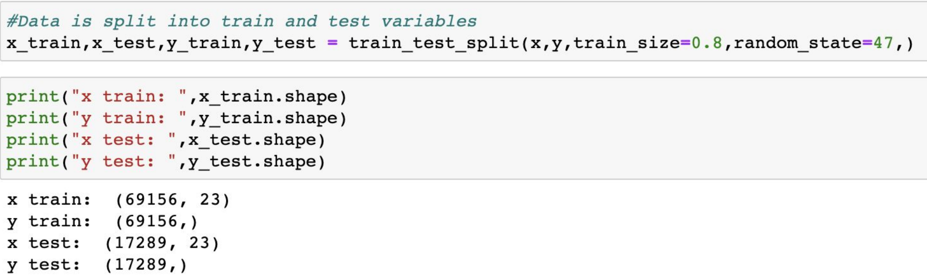
Machine Learning Models tested

- LinearRegression
- DecisionTreeRegressor
- RandomForestRegressor
- GradientBoostingRegressor

ML Model 1



```
Metrics of model:  LinearRegression()
Mean Absolute Error : 6.308981686057654
Mean Squared Error : 74.32370748748751
r2 score: 0.1665119512273734
cv score:  0.16416833173558973
************************************
Metrics of model:  DecisionTreeRegressor()
Mean Absolute Error : 7.550749212024755
Mean Squared Error : 120.4273954130235
r2 score: -0.3505084476372329
cv score:  -0.33593258280566196
************************************
Metrics of model:  RandomForestRegressor()
Mean Absolute Error : 5.624256363859471
Mean Squared Error : 62.08943735091872
r2 score: 0.30371067676192276
cv score:  0.27994581734319723
************************************
Metrics of model:  GradientBoostingRegressor()
Mean Absolute Error : 5.61715809593955
Mean Squared Error : 58.968377901204676
r2 score: 0.3387111609786644
cv score:  0.3246556680237667
************************************
```



```
#Data is split into train and test variables
x_train,x_test,y_train,y_test = train_test_split(x,y,train_size=0.8,random_state=47,)
```

```
print("x train: ",x_train.shape)
print("y train: ",y_train.shape)
print("x test: ",x_test.shape)
print("y test: ",y_test.shape)
```

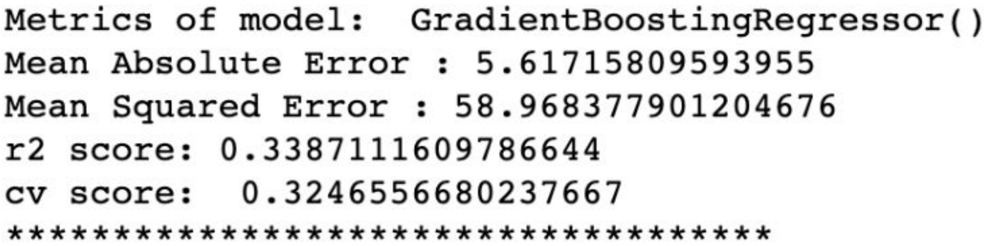
```
x train:  (69156, 23)
y train:  (69156,)
x test:  (17289, 23)
y test:  (17289,)
```

80% of the data is used for train model and remaining 20% of testing model.

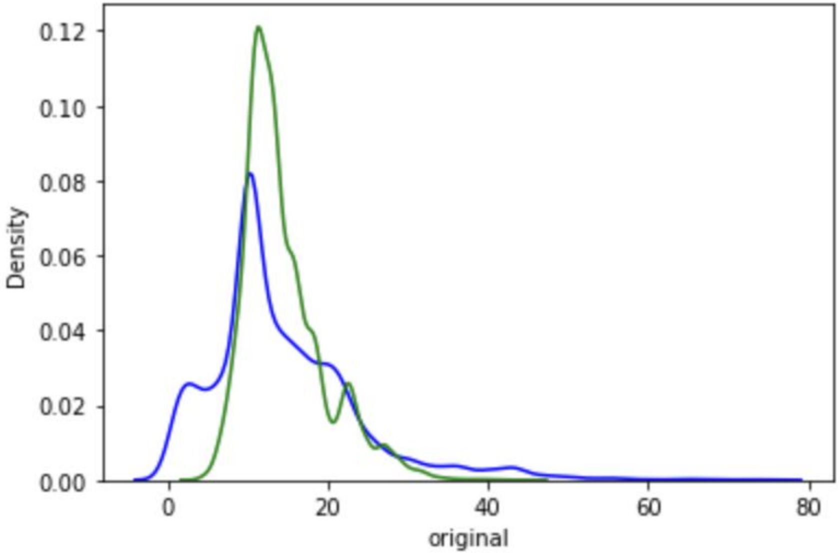
There are 23 independent columns.

Considering several factors like mean absolute error,and difference between r2 score and cv score we can see that Gradient Boosting Regressor Model Is Performing best.

Model 1



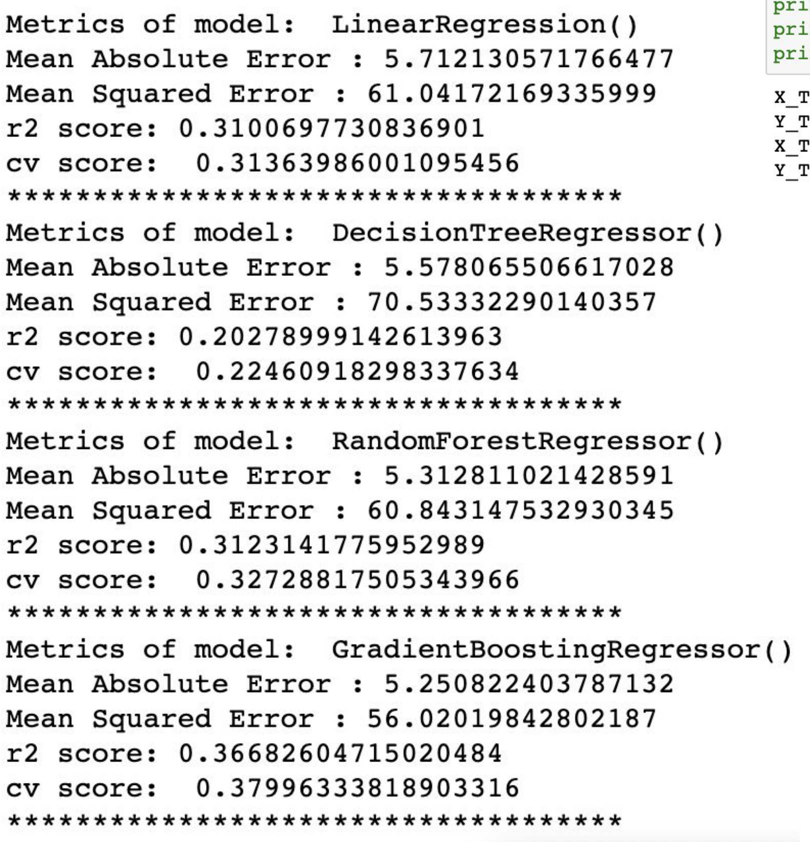
```
Metrics of model:  GradientBoostingRegressor()
Mean Absolute Error : 5.61715809593955
Mean Squared Error : 58.968377901204676
r2 score: 0.338711609786644
cv score:  0.3246556680237667
************************************
```



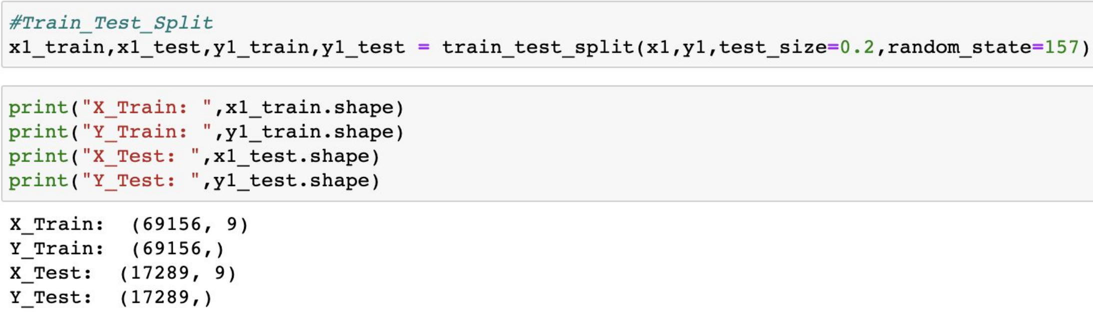
- The model is predicting well only when the original values are in range of 5-20.

- We can see there is lot of noise in our target data, which is causing this issue of low performance of model in terms of percentage by r2 score.

Model 2



```
Metrics of model:  LinearRegression()
Mean Absolute Error : 5.712130571766477
Mean Squared Error : 61.04172169335999
r2 score: 0.3100697730836901
cv score:  0.31363986001095456
************************************
Metrics of model:  DecisionTreeRegressor()
Mean Absolute Error : 5.578065506617028
Mean Squared Error : 70.53332290140357
r2 score: 0.20278999142613963
cv score:  0.22460918298337634
************************************
Metrics of model:  RandomForestRegressor()
Mean Absolute Error : 5.312811021428591
Mean Squared Error : 60.843147532930345
r2 score: 0.3123141775952989
cv score:  0.32728817505343966
************************************
Metrics of model:  GradientBoostingRegressor()
Mean Absolute Error : 5.250822403787132
Mean Squared Error : 56.02019842802187
r2 score: 0.36682604715020484
cv score:  0.37996333818903316
************************************
```



```
#Train_Test_Split
x1_train,x1_test,y1_train,y1_test = train_test_split(x1,y1,test_size=0.2,random_state=157)
```

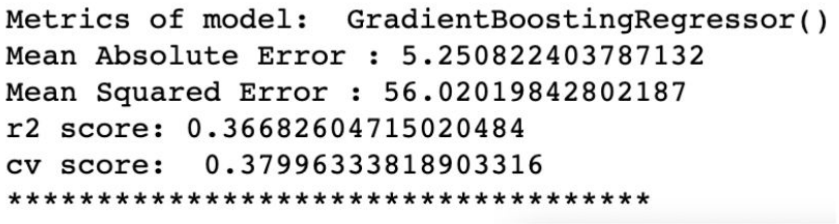
```
print("X_Train: ",x1_train.shape)
print("Y_Train: ",y1_train.shape)
print("X_Test: ",x1_test.shape)
print("Y_Test: ",y1_test.shape)
```

```
X_Train:  (69156, 9)
Y_Train:  (69156,)
X_Test:  (17289, 9)
Y_Test:  (17289,)
```

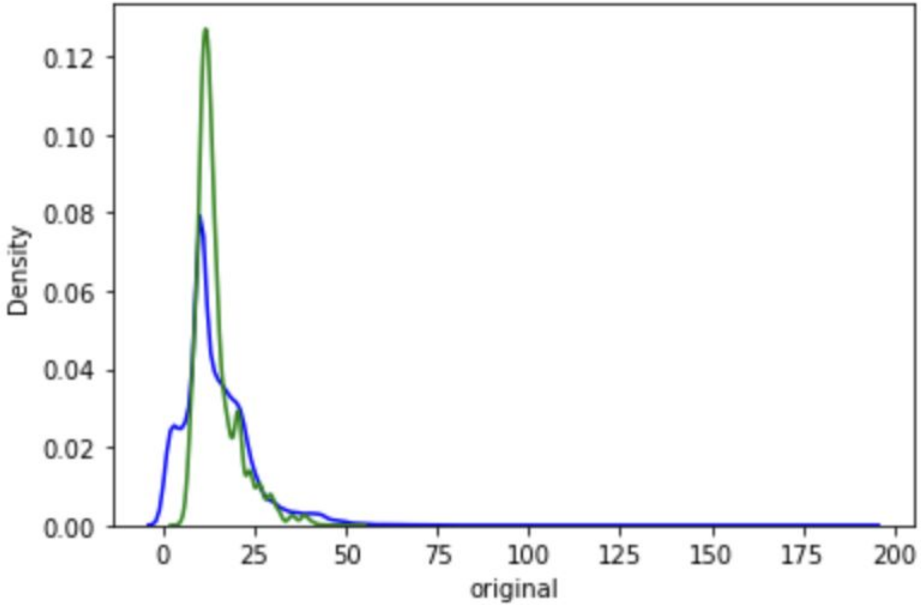
- 80% of the data is used for train model and remaining 20% of testing model.
- There are 9 independent columns.
- Considering several factors like mean absolute error,and difference between r2 score and cv score we can see that Gradient Boosting Regressor Model Is Performing best.

(Note: Dropped columns that are not correlated from analyzing the heatmap)

Model 2



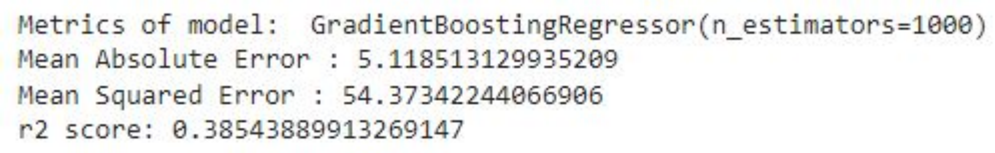
```
Metrics of model:  GradientBoostingRegressor()
Mean Absolute Error : 5.250822403787132
Mean Squared Error : 56.02019842802187
r2 score: 0.36682604715020484
cv score:  0.37996333818903316
************************************
```



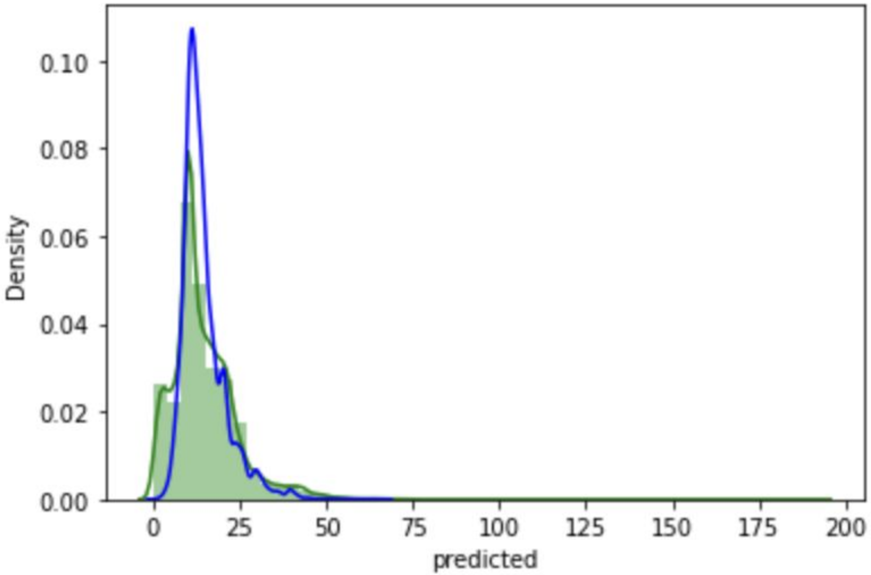
- The model is predicting well only when the original values are in range of 5-50.

- This is an improved model because it has better performance compared to Model 1.

Hyperparameter tuning



```
Metrics of model:  GradientBoostingRegressor(n_estimators=1000)
Mean Absolute Error : 5.118513129935209
Mean Squared Error : 54.37342244066906
r2 score: 0.38543889913269147
```



After tuning the model the model performance further improved.

Parameters changed:

'criterion': 'friedman_mse',
'loss': 'squared_error',
'n_estimators': 1000

Accuracy improved from 36% to 39%

Conclusions and Findings

- Driver Impairment is not the major factor for most of the crashes
- More incident take place during beginning of the year which may possibly be due to snow and poor road conditions and start to increase around June when I believe roads are busiest.
- Most accidents happen in daylight because of busier roads.
- Beginner drivers aged 20-30 are involved in most crashes.
- Speed is not the main reason for crashes as 99% of drivers are in limit when the accidents happened.

Thank You!