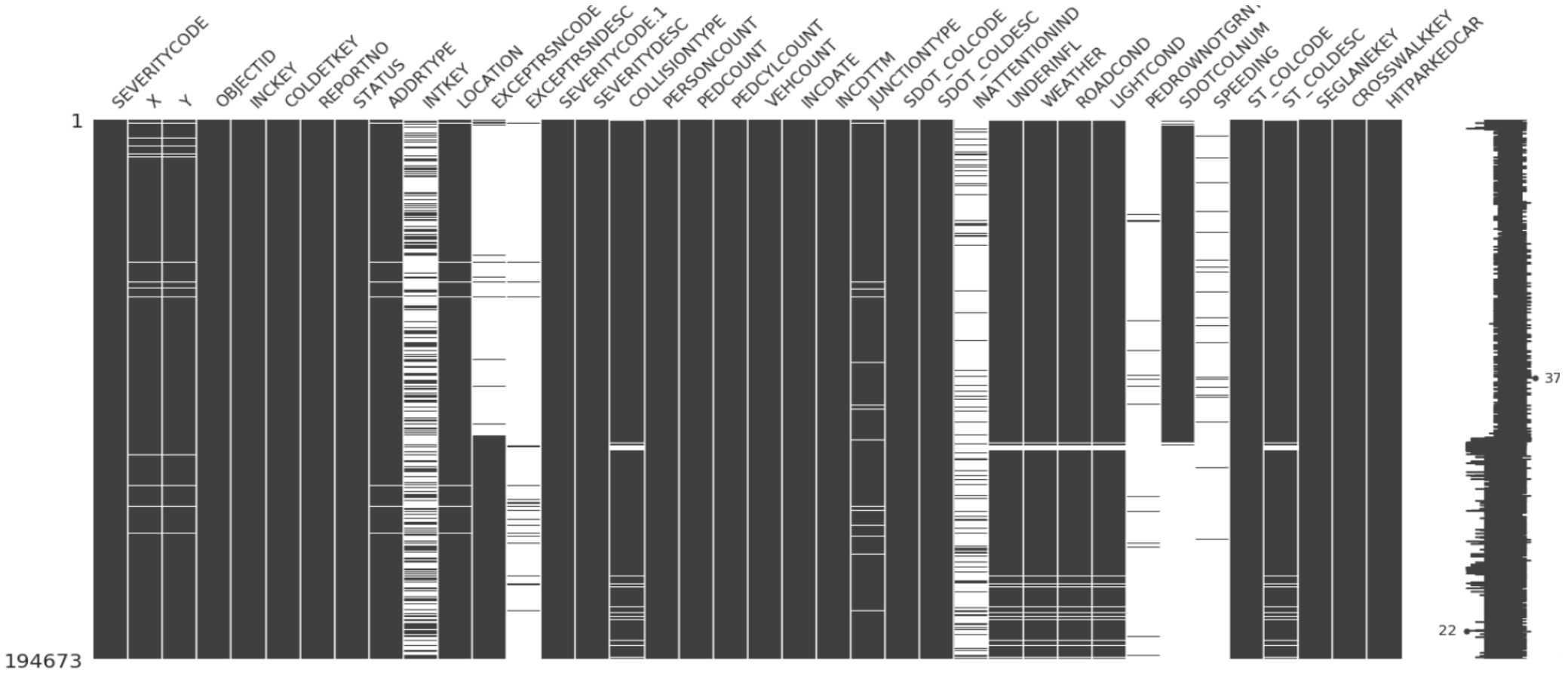
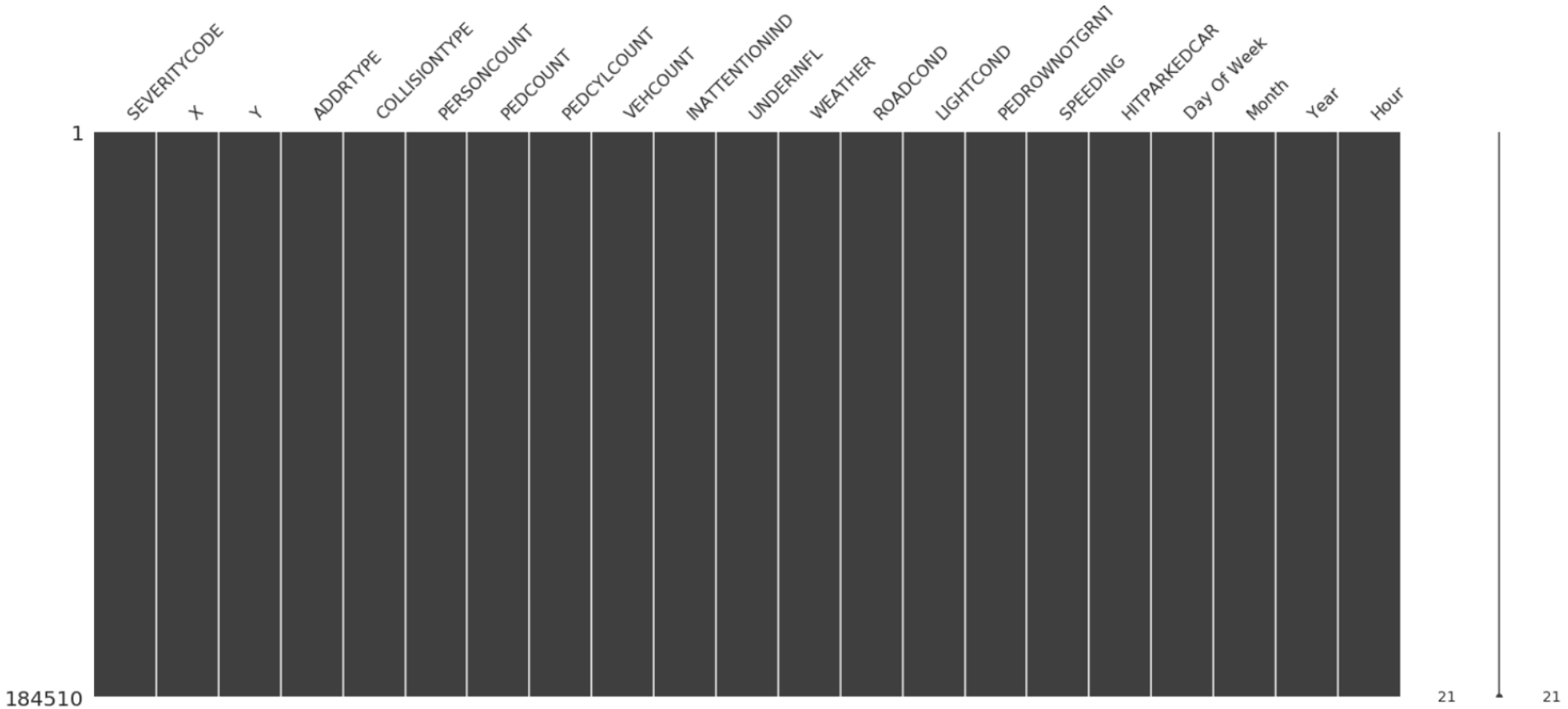
Data Feature and missing value Fixed

Before



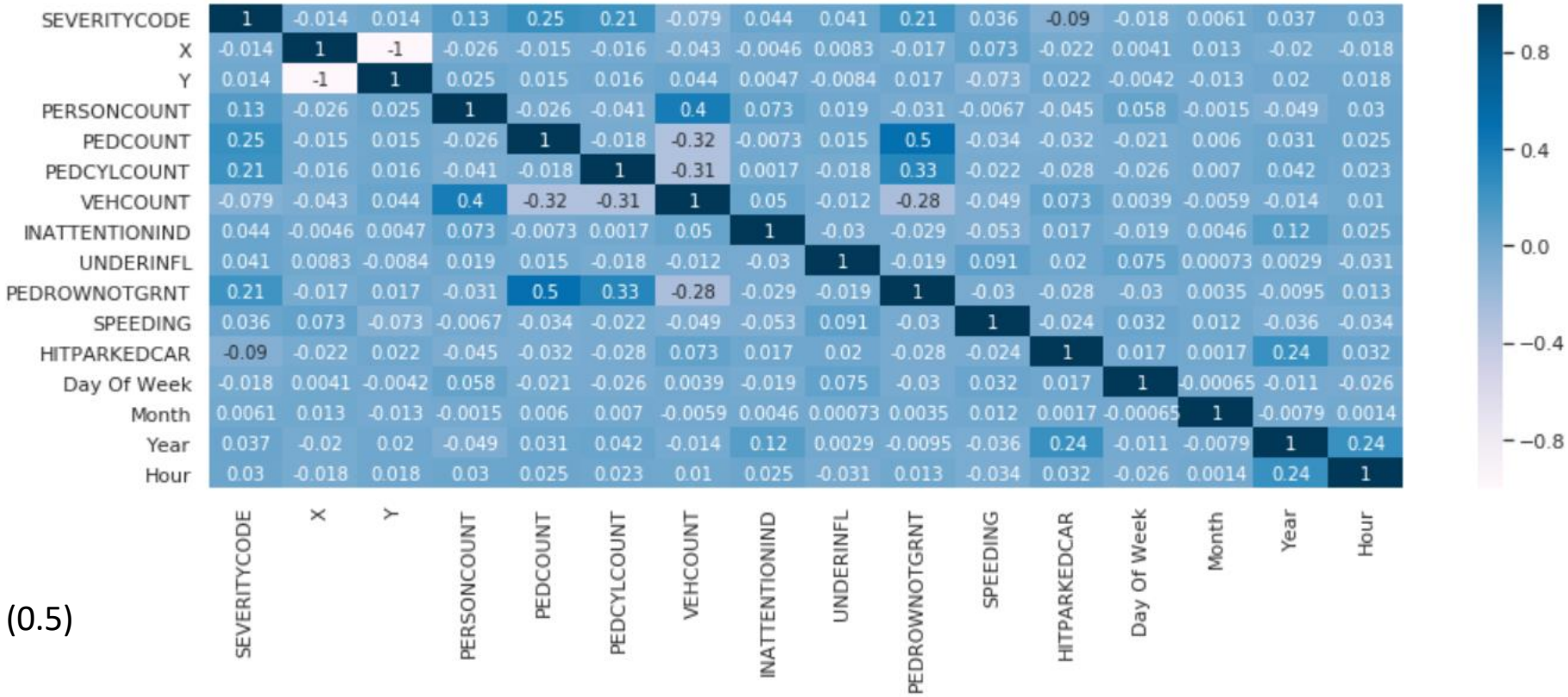
Data Feature and missing value Fixed

After



Exploratory data analysis

Correlation Matrix



The most correlation we have are:

Independent variable:

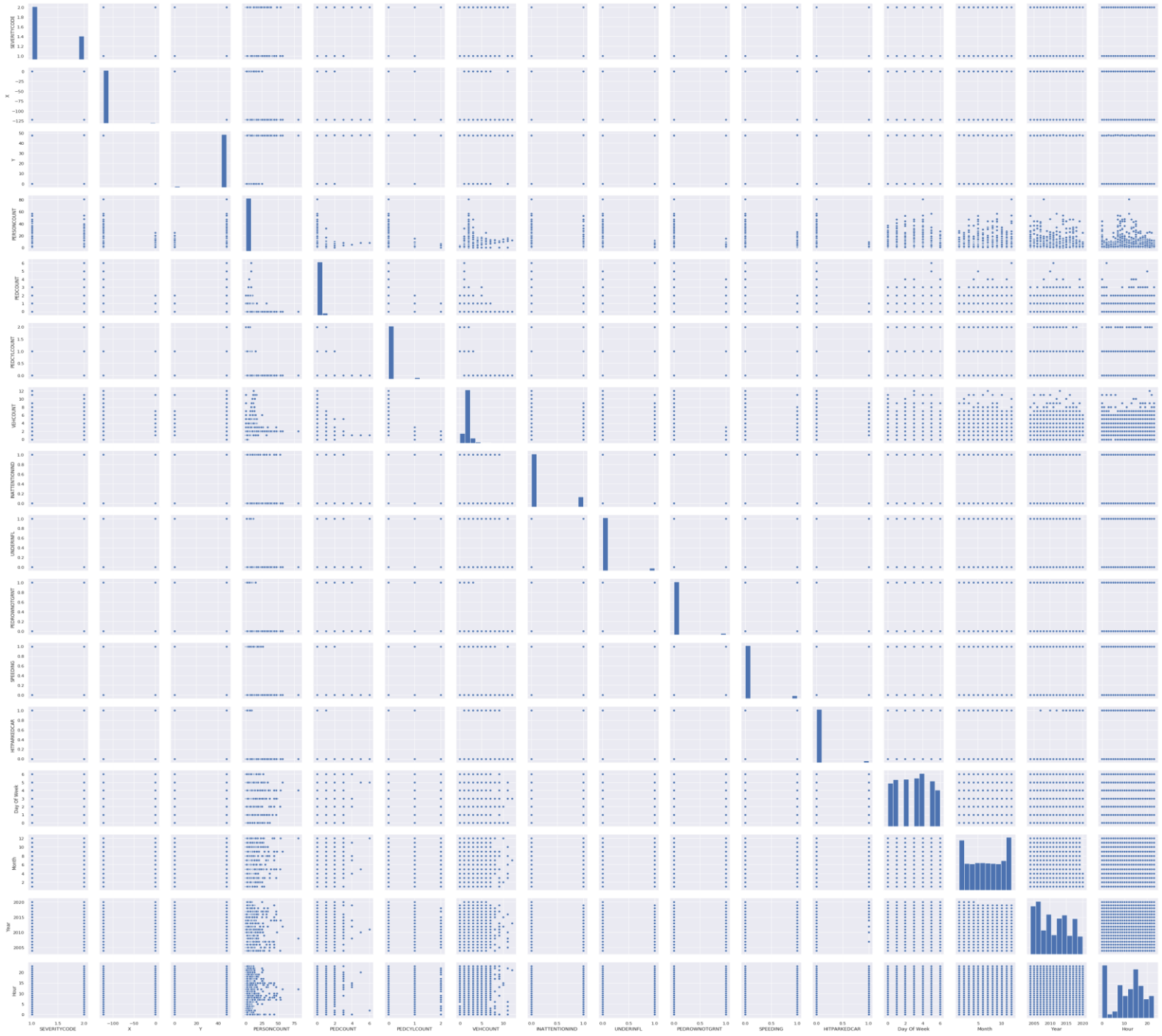
- PEDROWNOTGRANT – PEDCYLCOUNT (0.5)
- VEHCOUNT – PERSONCOUNT (0.4)
- SPEEDING- PEDCYLCOUNT(0.33)

Dependent variable:

- SEVERIYCODE- PEDCOUNT(0.25)
- SEVERIYCODE- PEDCYLCOUNT(0.21)
- SEVERIYCODE- PEDROWNOTGRANT (0.21)

Pair plot

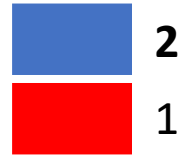
There is no probability distribution function
That we recognize for a feature.

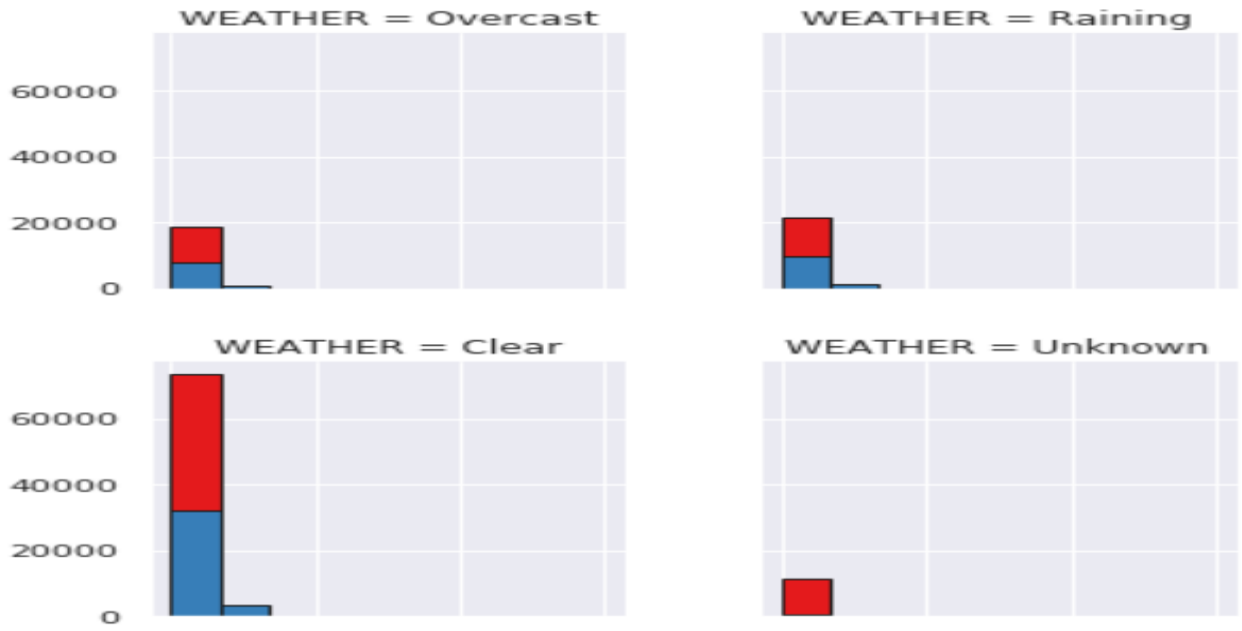


Categorical Data impact on collision and collision severity

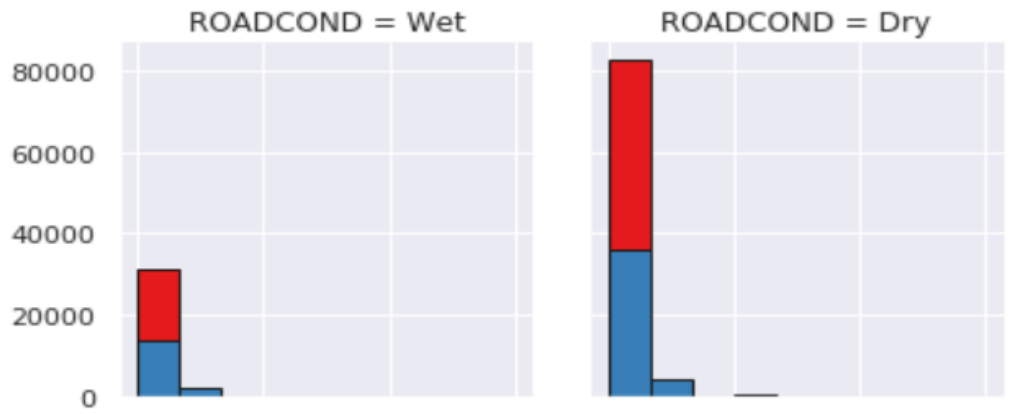
Collison address Type

Weather


2
1


WEATHER = Overcast
60000
40000
20000
0
WEATHER = Raining
WEATHER = Clear
60000
40000
20000
0
WEATHER = Unknown

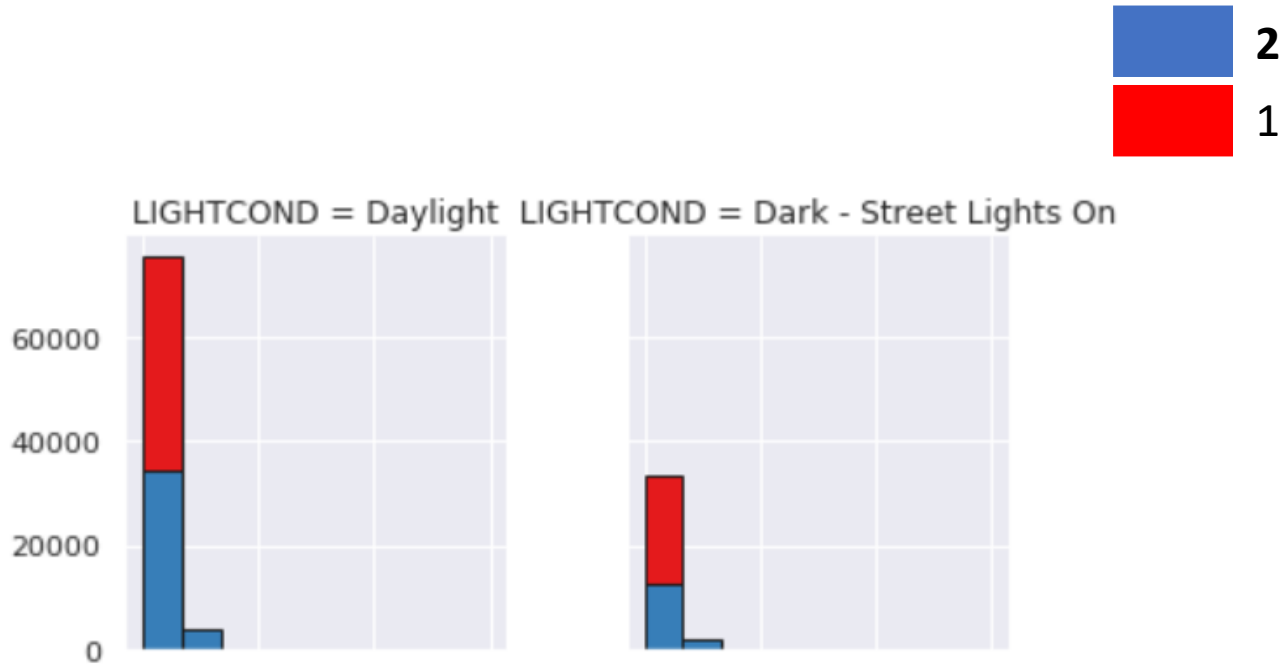
Road Condition



2

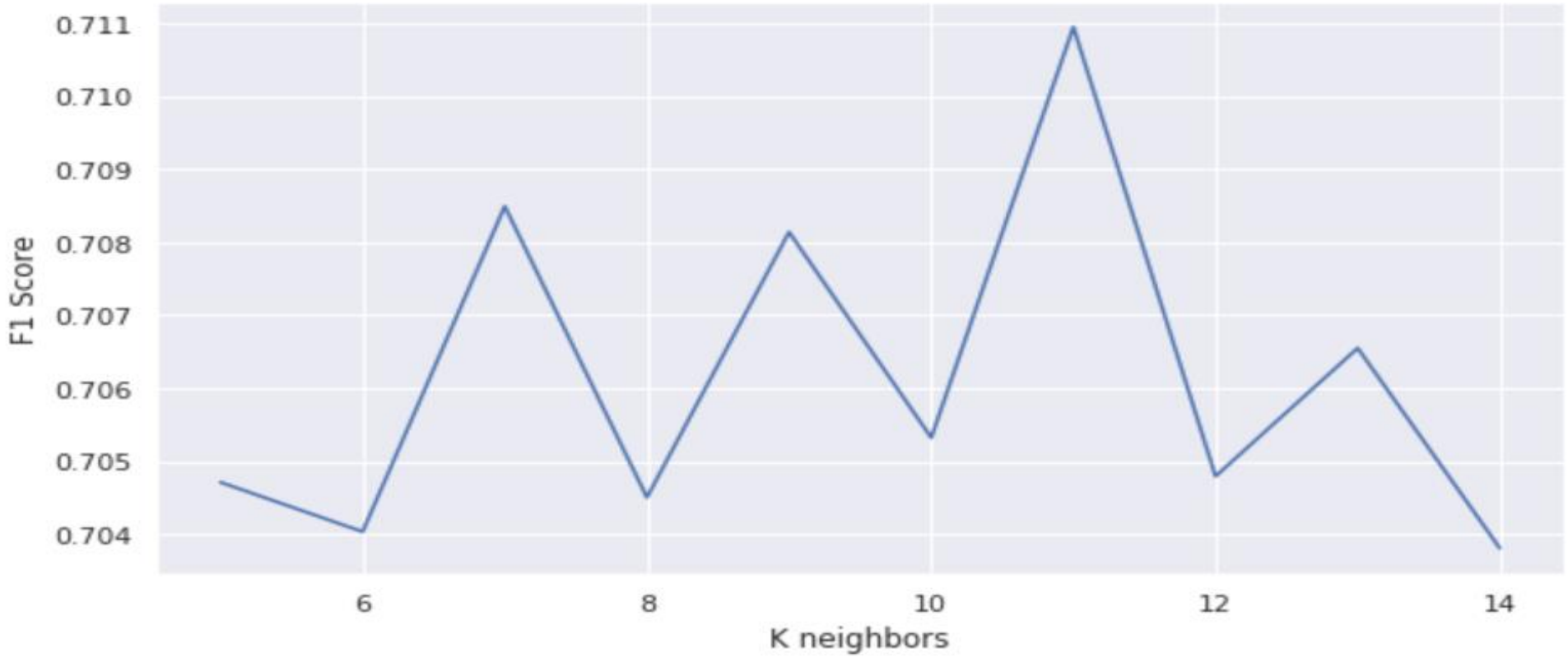
1

Light Condition



Predictive Model

K choice



We choose k=11

	precision	recall	f1-score	support
1	0.76	0.90	0.82	6300
2	0.63	0.38	0.48	2926
micro avg	0.73	0.73	0.73	9226
macro avg	0.69	0.64	0.65	9226
weighted avg	0.72	0.73	0.71	9226

- **The model has a very interesting precision for both classes.**
- **The recall for class1 is nearly perfect while it's lower for the 2 class.**
- **The f1 score is pretty much good in average**

- Model of choice:
  - KNN Classifier with 11 neighbors as parameter