

CAR ACCIDENT SEVERITY PREDICTION

-GANAPATHI BALAJI



CONTENT

- Problem Definition
- Exploratory Data Analysis
- Feature Engineering
- Model Process
- Model Report
- Summary Conclusion

EXPLORATORY DATA ANALYSIS

	Unique_count
PEDCYLCOUNT	3
PEDCOUNT	7
VEHCOUNT	13
SDOT_COLCODE	39
PERSONCOUNT	47
SEGLANEKEY	1955
CROSSWALKKEY	2198
INTKEY	7614
X	23563
Y	23839
SDOTCOLNUM	114932
OBJECTID	194673
INCKEY	194673
COLDKETKEY	194673

Remove keys as
it won't add value

	count	unique	top	freq
SPEEDING	9333	1	Y	9333
EXCEPTRSNDESC	5638	1	Not Enough Information, or Insufficient Locati...	5638
PEDROWNOTGRNT	4667	1	Y	4667
INATTENTIONIND	29805	1	Y	29805
HITPARKEDCAR	194673	2	N	187457
STATUS	194673	2	Matched	189788
EXCEPTRSNCODE	84811	2		79173
SEVERITYDESC	194673	2	Property Damage Only Collision	136485
ADDRTYPE	192747	3	Block	128928
UNDERINFL	189789	4	N	100274
JUNCTIONTYPE	188344	7	Mid-Block (not related to intersection)	89800
ROADCOND	189661	9	Dry	124510
LIGHTCOND	189503	9	Daylight	116137
COLLISIONTYPE	189769	10	Parked Car	47987
WEATHER	189592	11	Clear	111135
SDOT_COLDESC	194673	39	MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END ...	85209
ST_COLDESC	189769	62	One parked-one moving	44421
ST_COLCODE	194655	115	32	27612
INCDATE	194673	5985	2006/11/02 00:00:00+00	98
LOCATION	191996	24102	BATTERY ST TUNNEL NB BETWEEN ALASKAN WY VI NB ...	276
INCDDTM	194673	182058	11/2/2006	98
REPORTNO	194673	194670	1776526	2

Removing these columns as these
are constant through out the data
and and unknown for most of the
observations

Removing description
columns as we also have
codes

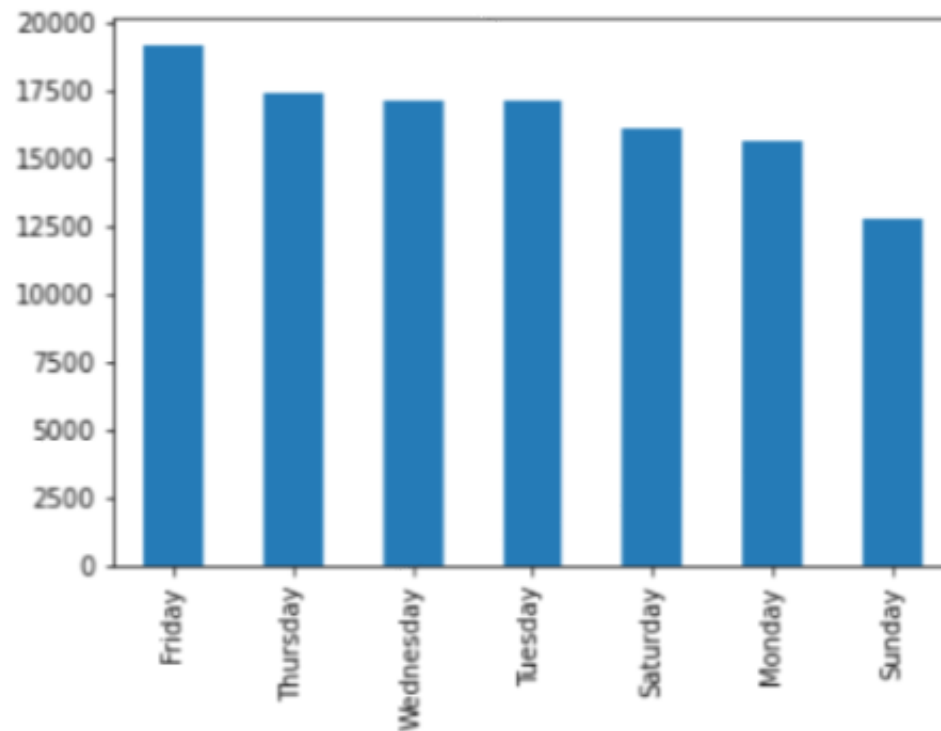
Removing time column
as we already have date
and LIGHTCOND fields

Removing Key value as it not
adding any value

EXPLORATORY DATA ANALYSIS

```
In [180]: df_under_sample['day_of_week'].value_counts().plot.bar()
```

```
Out[180]: <matplotlib.axes._subplots.AxesSubplot at 0x2983744c940>
```

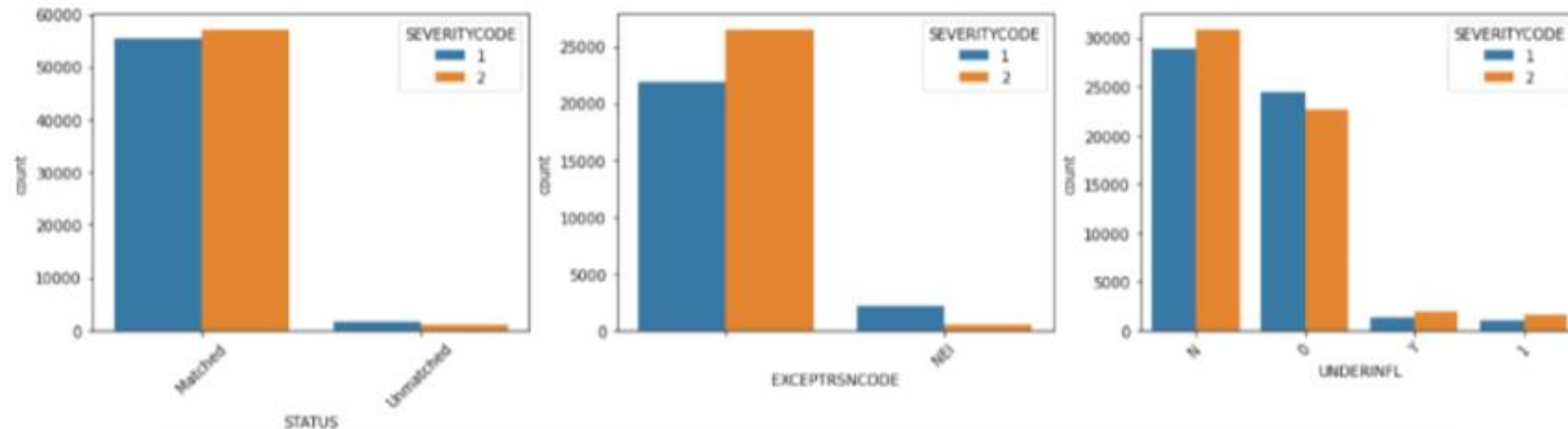


All most all the days are having similar frequency Friday&Sunday has little variation can be ignored as it is not having much impact on verall

EXPLORATORY DATA ANALYSIS

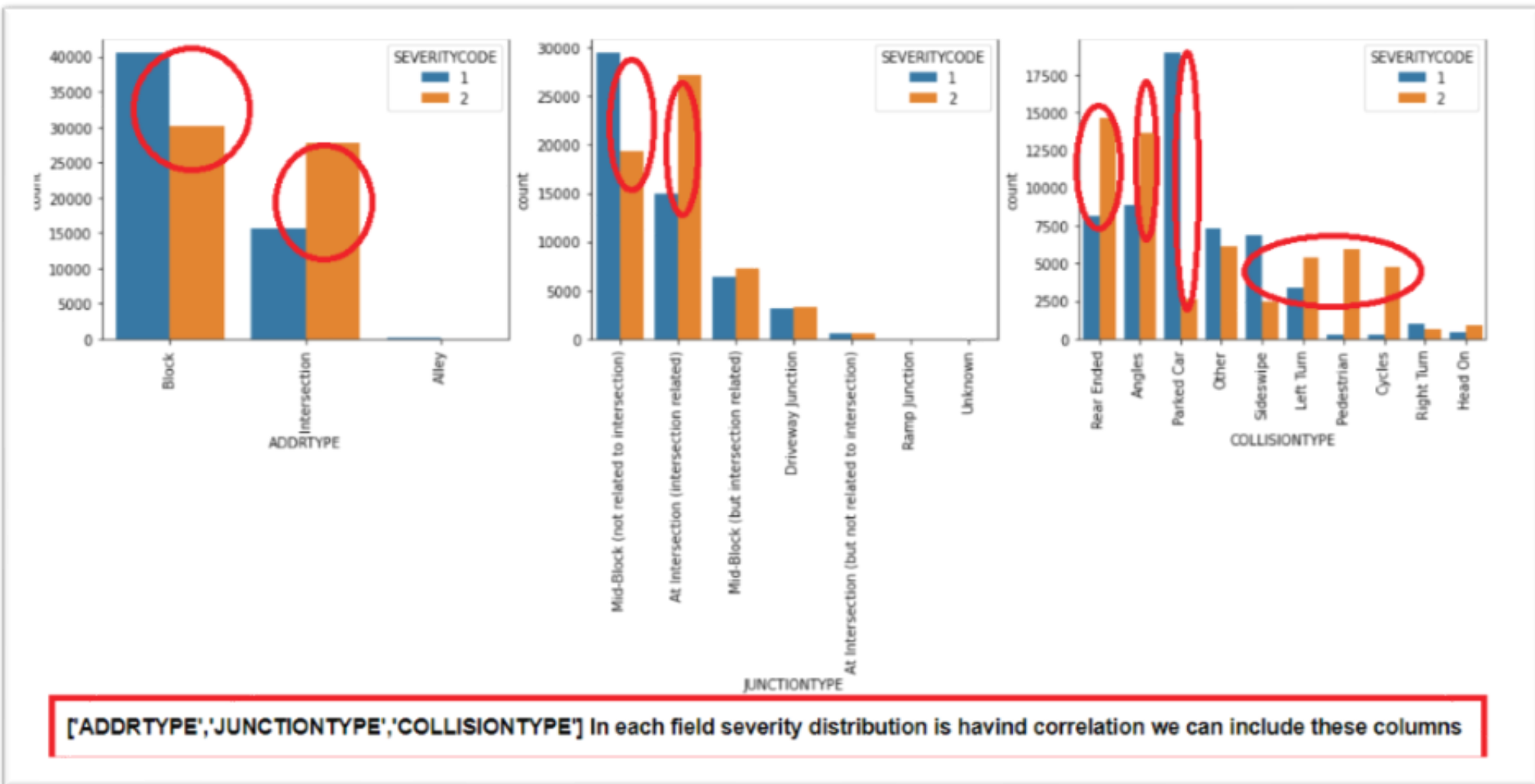
`['STATUS','EXCEPTSNCODE','UNDERINFL']`

```
In [ ]: f, ax = plt.subplots(1,3,figsize=(18, 4))
sns.countplot(x='STATUS', hue='SEVERITYCODE',data=df_under_sample,order=df_under_sample['STATUS'].value_counts().index,ax=ax[0])
sns.countplot(x='EXCEPTSNCODE', hue='SEVERITYCODE',data=df_under_sample,order=df_under_sample['EXCEPTSNCODE'].value_counts().index,ax=ax[1])
sns.countplot(x='UNDERINFL', hue='SEVERITYCODE',data=df_under_sample,order=df_under_sample['UNDERINFL'].value_counts().index,ax=ax[2])
ax[0].set_xticklabels(ax[0].get_xticklabels(), rotation=45)
ax[1].set_xticklabels(ax[1].get_xticklabels(), rotation=45)
ax[2].set_xticklabels(ax[2].get_xticklabels(), rotation=45)
f.show()
```

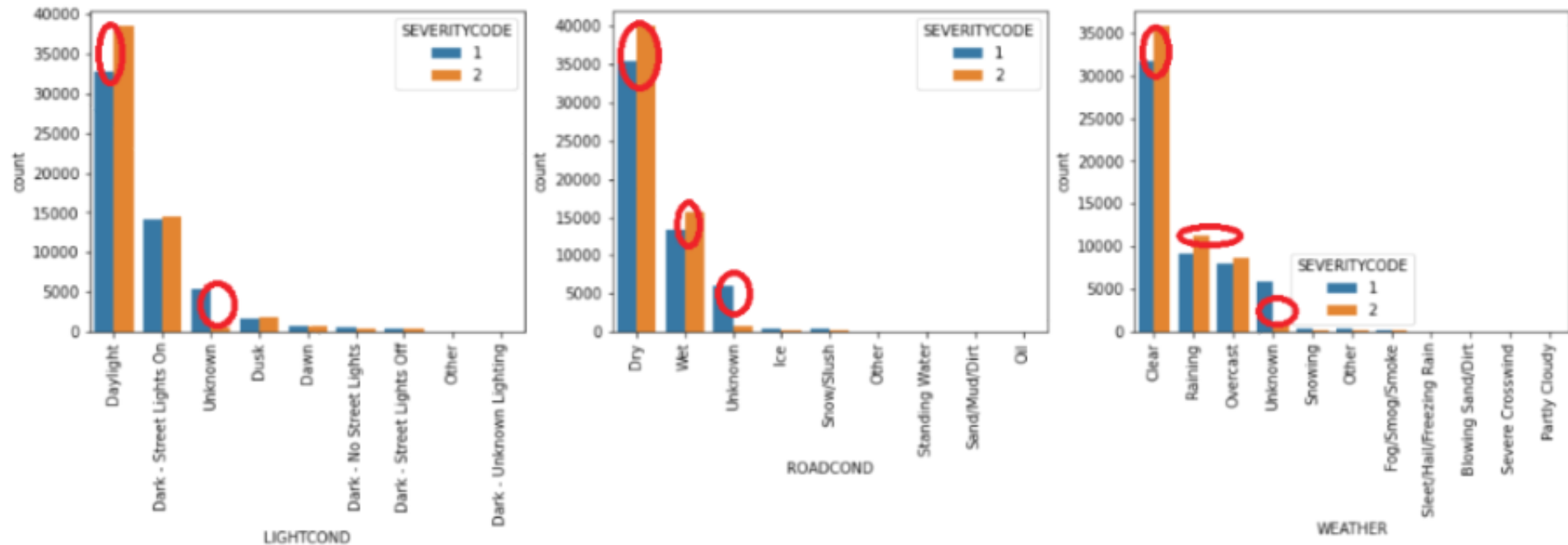


`['STATUS','EXCEPTSNCODE','UNDERINFL']` In each field severity distribution is similar we can ignore these columns

EXPLORATORY DATA ANALYSIS

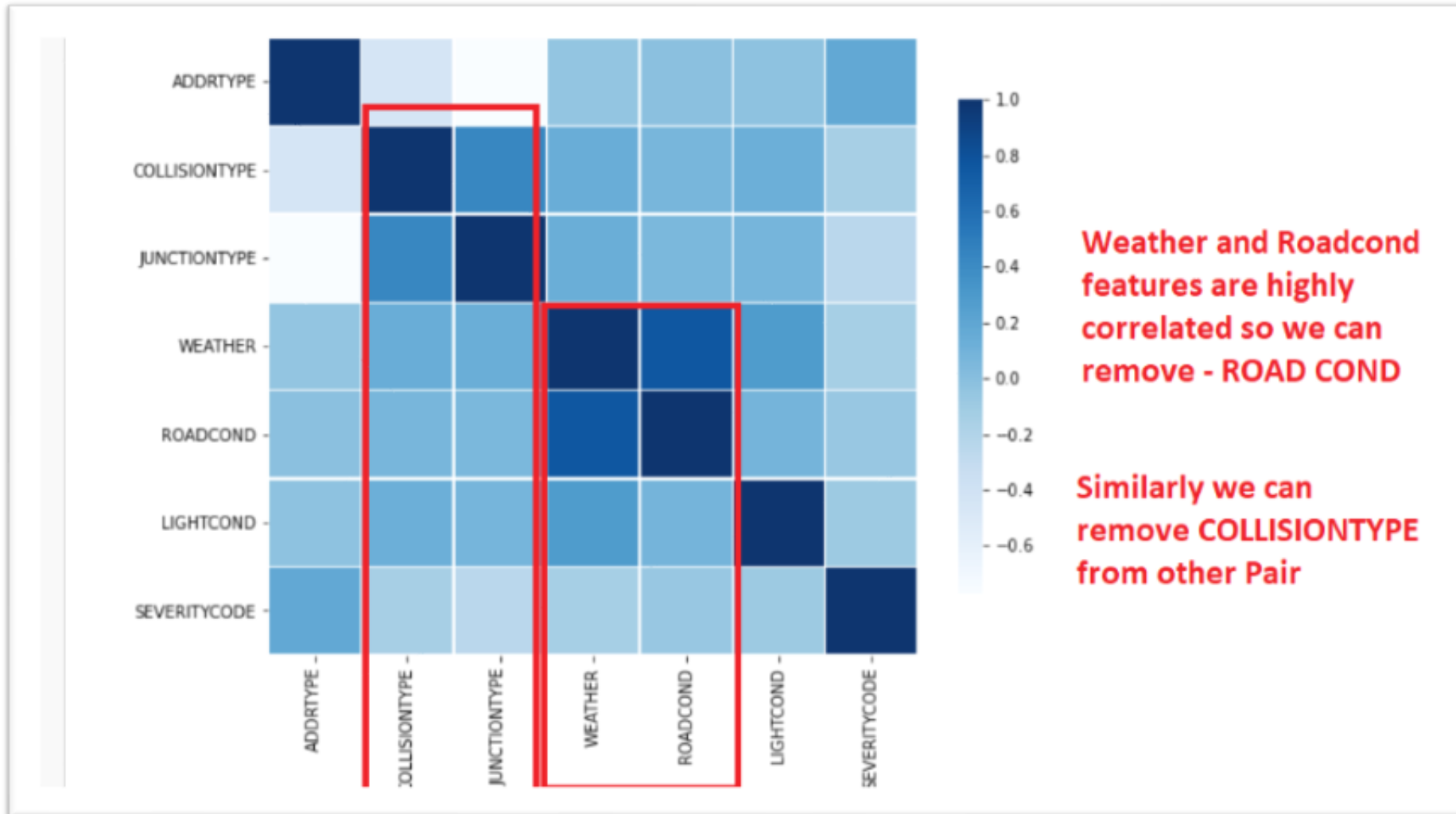


EXPLORATORY DATA ANALYSIS

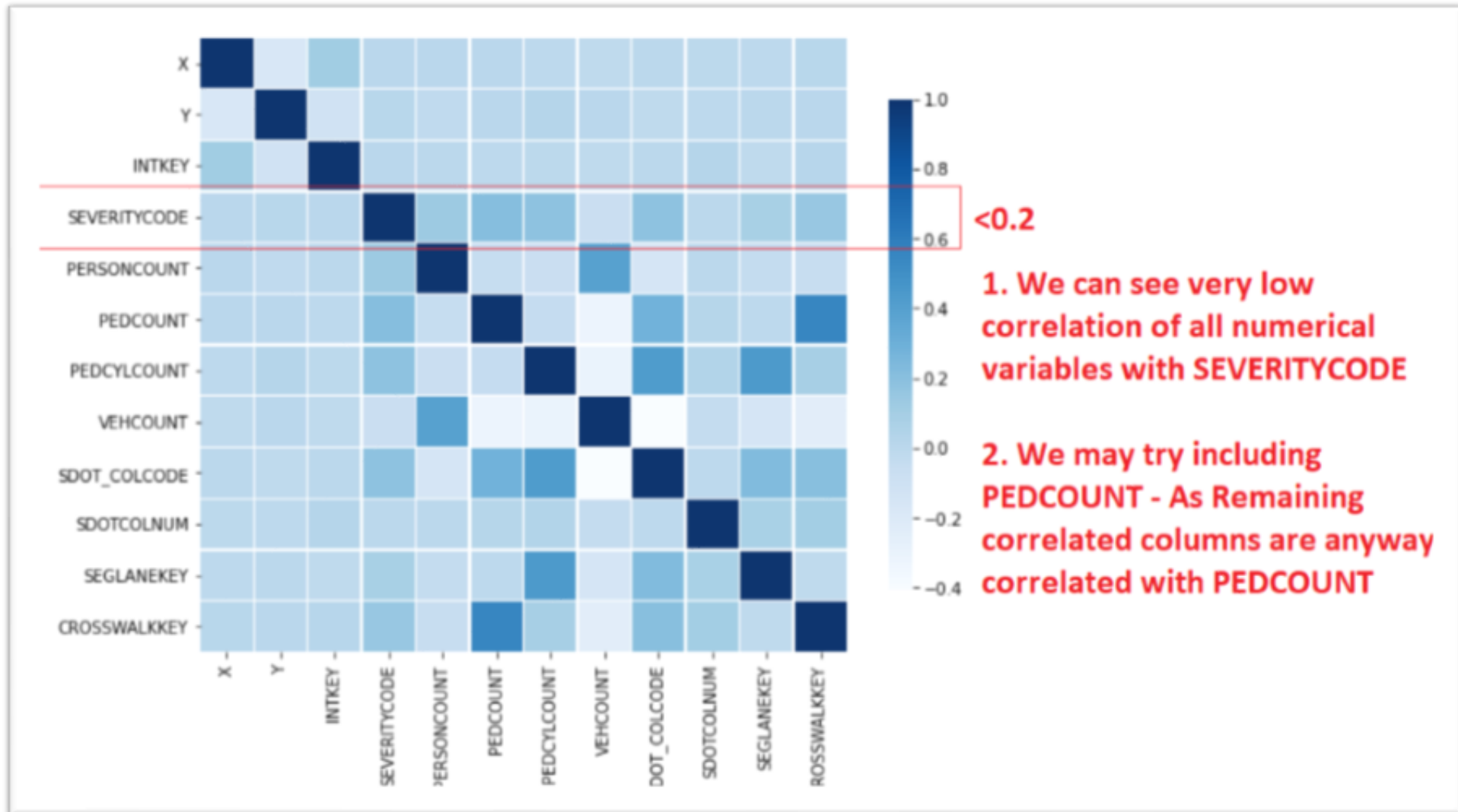


['LIGHTCOND','ROADCOND','WEATHER'] Severity distribution is having some correlation with fields so we can include these columns and test the performance

EXPLORATORY DATA ANALYSIS



EXPLORATORY DATA ANALYSIS



FEATURE ENGINEERING

- After all the EDA I can say the final list of features are as below
 - From Numerical features (#PEDCOUNT)
 - From Categorical features (ADDRTYPE, COLLISIONTYPE, JUNCTIONTYPE, WEATHER, ROADCOND, LIGHTCOND)

```
print('label:',label)
print('features:')
final_features

label: SEVERITYCODE
features: ['ADDRTYPE',
          'COLLISIONTYPE',
          'JUNCTIONTYPE',
          'WEATHER',
          'ROADCOND',
          'LIGHTCOND',
          'PEDCOUNT']
```



Thank You