ACCIDENT SEVERITY PREDICTION REPORT

Nasser

1. INTRODUCTION

Traffic accidents are one of the leading causes of death and injuries, According to International Transport Forum, 37 133 persons lost their lives in traffic crashes in the United States in 2017.

Many factors affects the traffic accidents, some related to the road, the car, or the person.

The scope of this project is to predict the severity of an accident through supervised machine learning algorithms using historical dataset, prediction of severity of the road based on certain features or condition such as "Road Condition, Weather Condition, Light Conditions, others".

The questions that are answered by this project

- What is the type of severity of a collision?
- What is the possibility of a person getting into a car accident and how severe it would be?

The importance of such project:

1- You can notify people to avoid the factors that have high probability of accident.
2- Emergency support can be prepared in case of an accident occurs in specific roads and can send the prober support based on the expected severity.
3- Special preparation can take place from government (Ex. Bad weather, Holidays,).
4- While continues development and enhancement in this model will help in Enhancing roads, notify people, advice…

Audience:

- Government entities: Traffic Department
- Emergency department

2. - Data

The Dataset is provided by SPD and recorded by Traffic Records (suggested by the course material); it includes all types of collisions within this Timeframe: 2004 to Present. Collisions will display at the intersection or mid-block of a segment.

The dataset can be downloaded from the below link
https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv.

The metadata for the dataset can be downloaded from the below link
https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

The following is initial information about the dataset.

- It contains of 194,673 rows and 38 attributes.
- Challenges
- Some missing values for specific columns, some limited rows of specific categories.
- For the complete description of the attributes and metadata, kindly follow the above link for the metadata.

- *SEVERITYCODE* is the label; it is biased label for code 1; it contains of 136,485 for sever ity 1 and 58188 for severity 2.

Features that can be extracted from the data:

- *ADDRTYPE*: Address Type
  Block 126926
  Intersection 65070
  Alley 751 few number for this category.

- *COLLISIONTYPE* is category of 10 types
- INCDATE the date of the incident.
- INCDTTM The date and time of the incident.
- JUNCTIONTYPE Category of junction at which collision took place

- WEATHER a description of the weather conditions during the time of the collision.

- ROADCOND The condition of the road during the collision.
- LIGHTCOND the light conditions during the collision.
- SPEEDING Whether or not speeding was a factor in the collision.

Attributes will be investigated more to check the impact and might be involved in feature engineering.

- PERSONCOUNT The total number of people involved in the collision
- PEDCOUNT the number of pedestrians involved in the collision.
- PEDCYLCOUNT the number of bicycles involved in the collision.
- VEHCOUNT the number of vehicles involved in the collision.
- INJURIES the number of total injuries in the collision.
- SERIOUSINJURIES The number of serious injuries in the collision.
- FATALITIES the number of fatalities in the collision.

Attributes have some issues:

- *INATTENTIONIND* (Y/N) it has only Y with 29805 but the rest are null, no values for N.
- *UNDERINFL* no enough information, there is no description for the values [1, 0].

For the complete list of attributes and description, check the URL
https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf.

3. Analysis and Methodology

I started by deleting the attributes that will not help in the analysis, below is summary for the deleted attributes mentioning the reason for deleting each one.

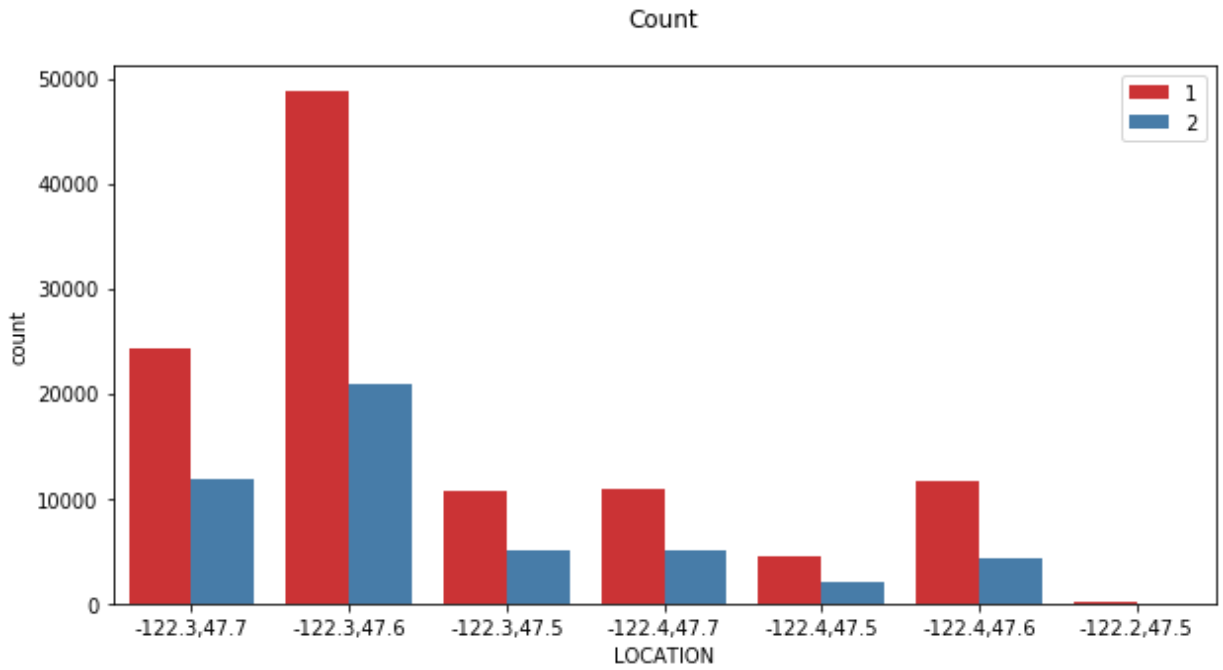
Ignore the following attributes:

Attribute Name	Reason to ignore
OBJECTID	Key
INCKEY	Key
COLDETKEY	Key
REPORTNO	Key
STATUS	No Enough information
INTKEY	Key
LOCATION	Not category data or unique
EXCEPTRSNCODE	No Enough information
EXCEPTRSNDESC	No Enough information
SEVERITYCODE.1	Duplicate attribute
SEVERITYDESC	Text Description
SDOTCOLNUM	Number not doesn't have impact
SDOT_COLDESC	Text Description
SDOT_COLCODE	No enough information, will use State Collision Code instead

- CROSSWALKKEY is deleted because 98 % of the CROSSWALKKEY attribute has value 0 and it represents only key.
- SEGLANEKEY is deleted because 99 % of the SEGLANEKEY attribute has value 0 and it represents only key.
- INCDATE is deleted as it is considered as duplicate information, the date and time can be extracted from INCDTTM.

- **Checking the X and Y against SEVERITYCODE:**

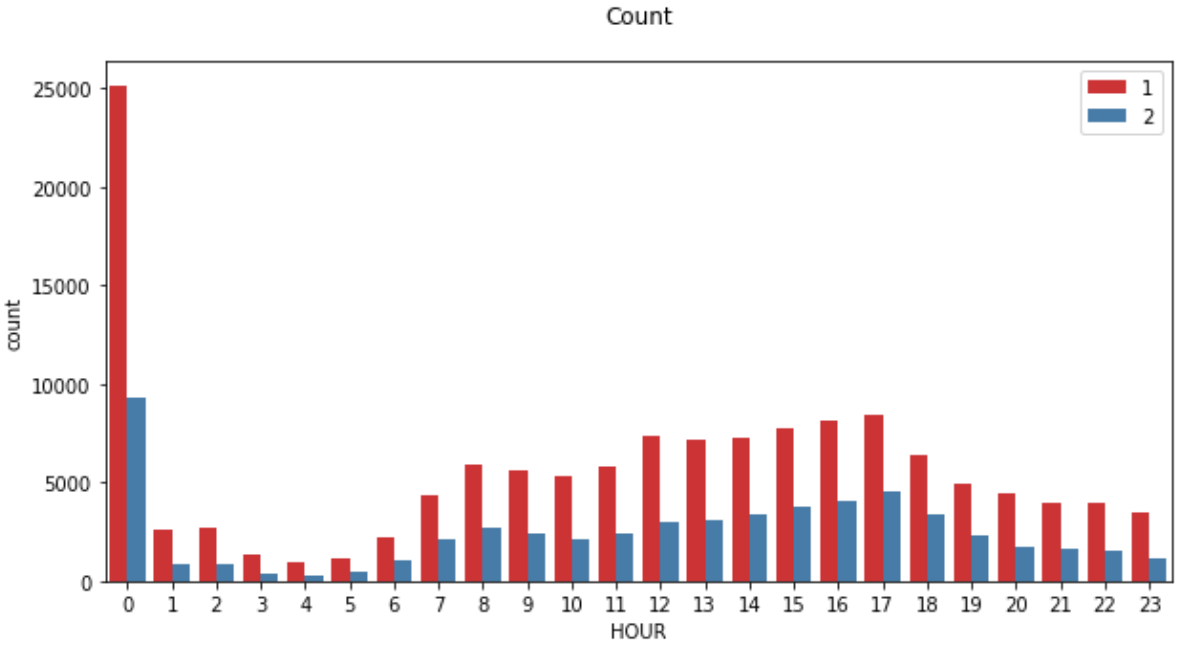
  - I deleted the LOCATION attribute because it is not categorical and just text but we cannot ignore that the location can be a factor for accident severity, so I tried to use the X and Y attribute to construct location with larger scale of Aria.
  - I did approximation to .1 which means to 110 Km, which gave me seven categories of location that can be used initially for understanding the location impact on the severity.



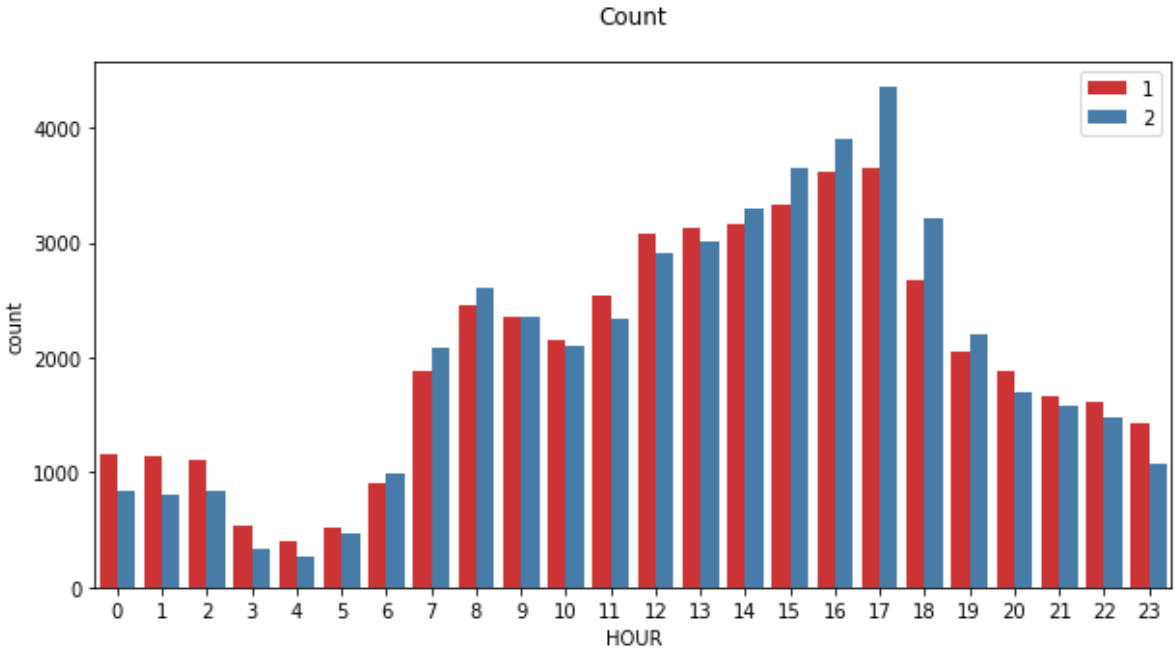
We see that almost similar severity code ratio for each Address that looks normal.

- **Checking the Hours against SEVERITYCODE:**

  - I extracted the Hour from the timestamp attribute, while I'm visualizing it, I found the following problem

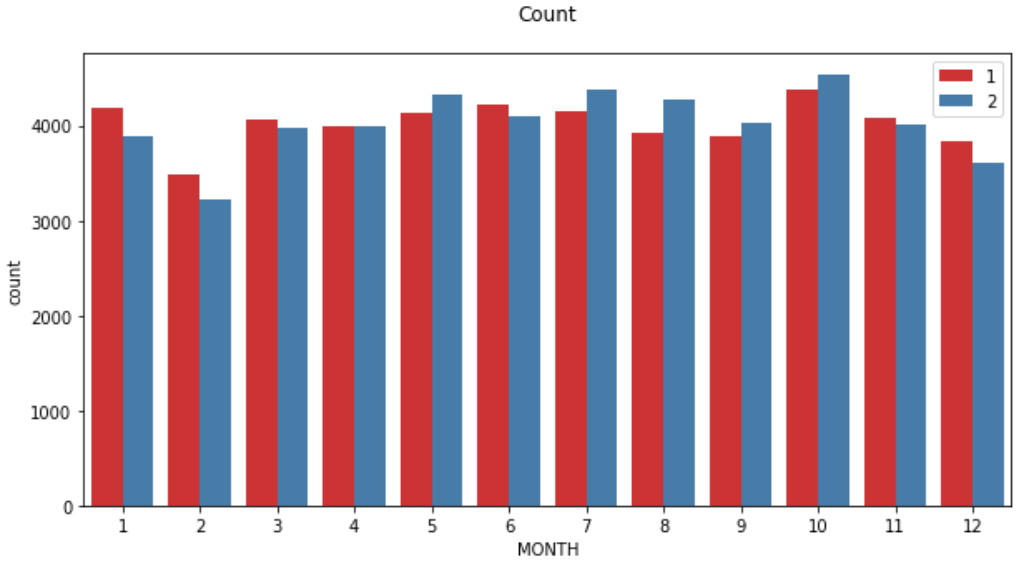


We can see that there is a problem with the data, at Hour zero the chart doesn't look normal, it has a peak not justified, after investigation, I found that some rows contains only date and don't have time information, so it is interpreted as 0 time, so I cleaned the data by removing these rows, below is the new graph after cleaning.

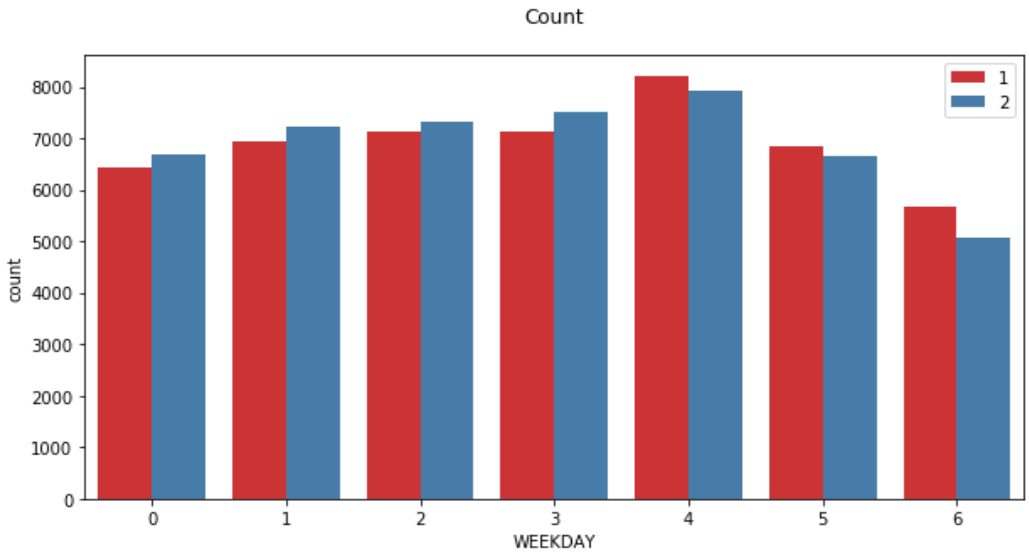


- New attributes are created Hour and Month and Year.
- We can get a conclusion from the above diagram that, some Hours has impact on number of collisions from 12 to 17.

- Month Histogram with SEVERITYCODE

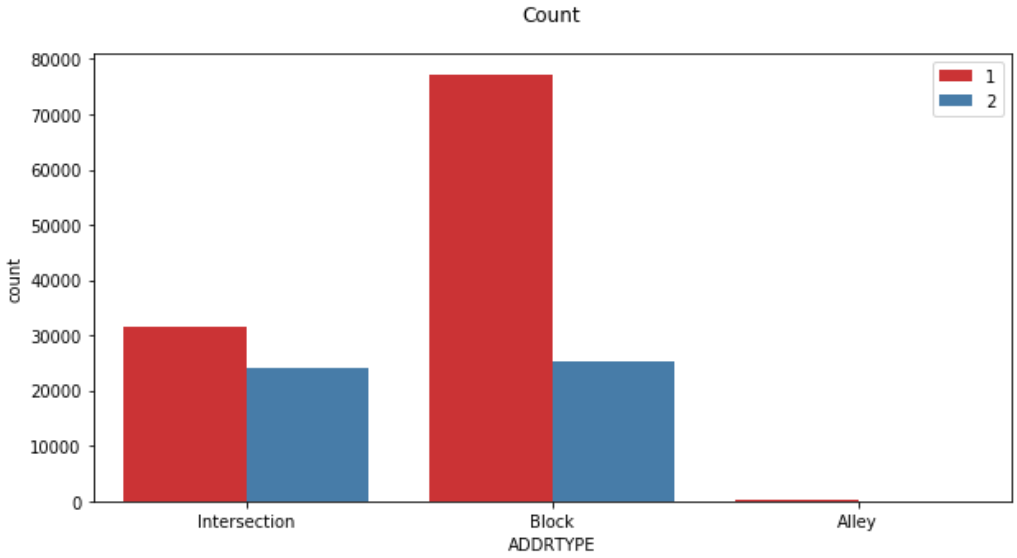


- We see that it is normal, no peaks, at month 2, there is less count but it can be because Feb sometimes come as 28 days only comparing to other months, so nothing special about the month.



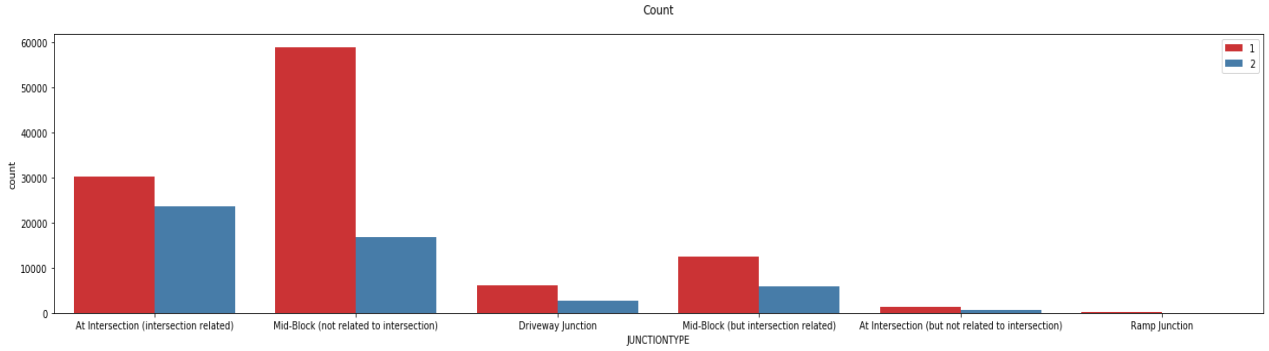
- Weekday histogram looks normal but number of accidents are lower in day 5 and 6.

- **Checking AddressType against our label.**



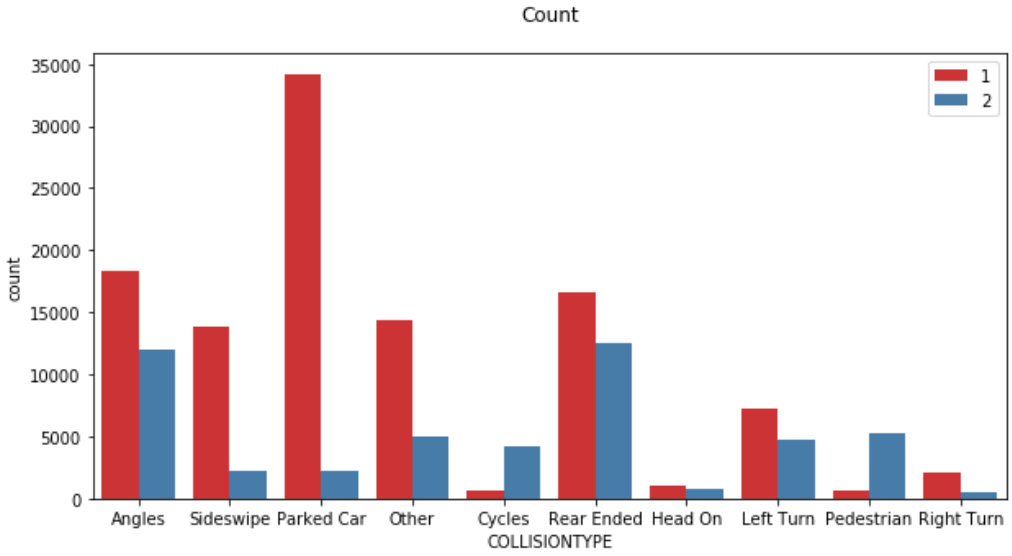
- Collisions at Address Type intersection has high severity Type ratio, which means that it is dangerous.

- **Checking JunctionType against our label.**



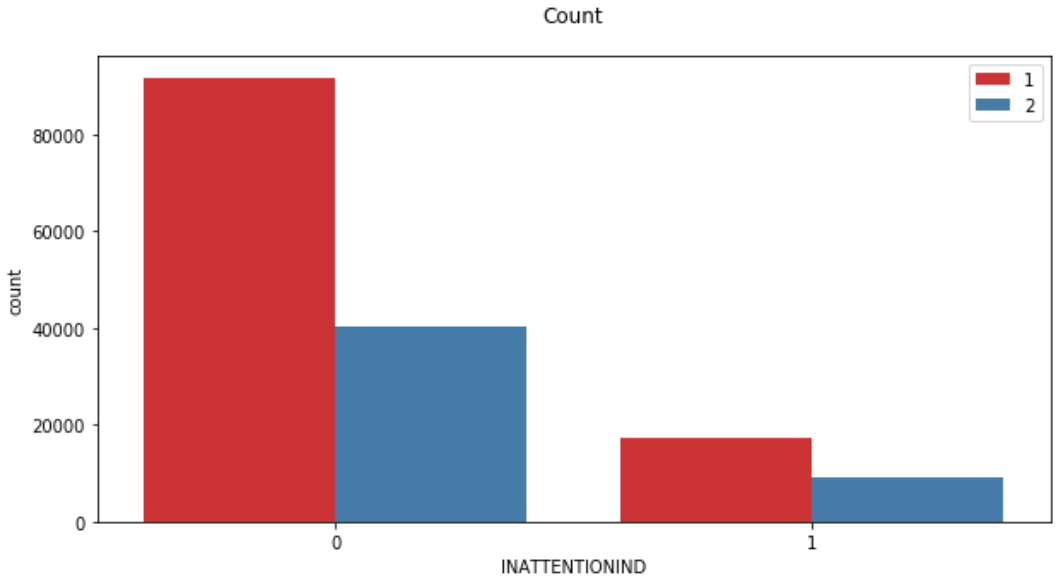
- Collisions at Junction Type intersectionRelated has high severity Type ratio.

- **Checking Collision Type against our label.**



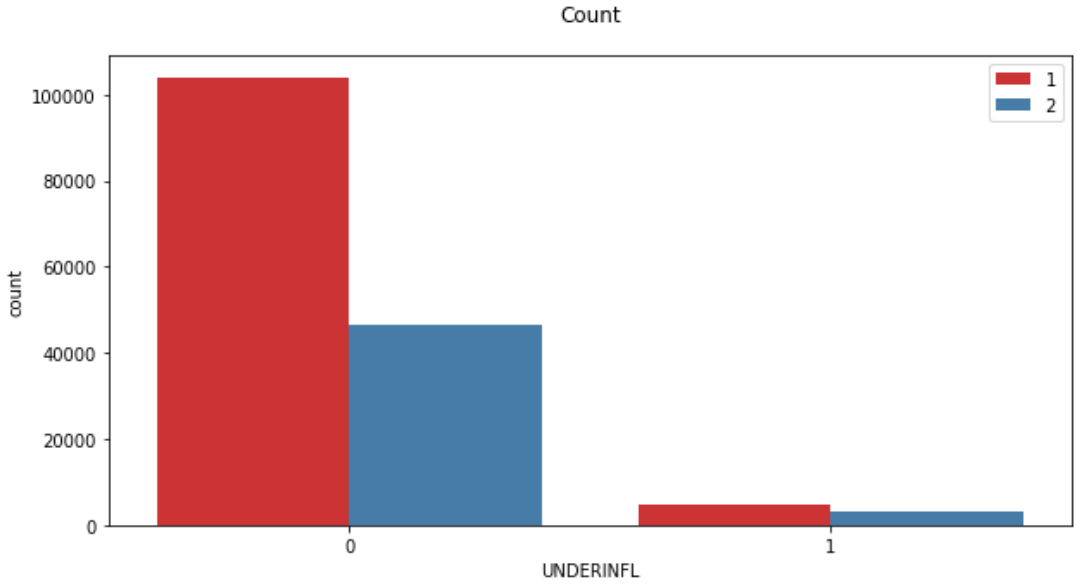
- Collisions at **Collision Type** Pedestrian and Cycles have high severity.
- *4703 rows have null values, since there is a type "Other" so the null values can fit in this value, so I changed the null values to other.*

- **Checking INATTENTIONIND against our label.**



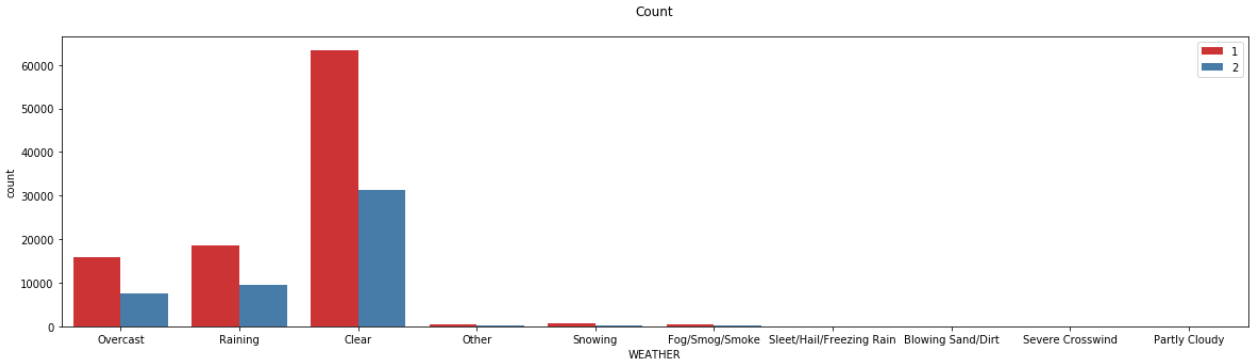
- It has little impact on severity.
- *Since this is Y, N column, I updated the Null values with N, then use numeric values instead, 0 for N, 1 for Y*

- **Checking UNDERINFL against our label.**



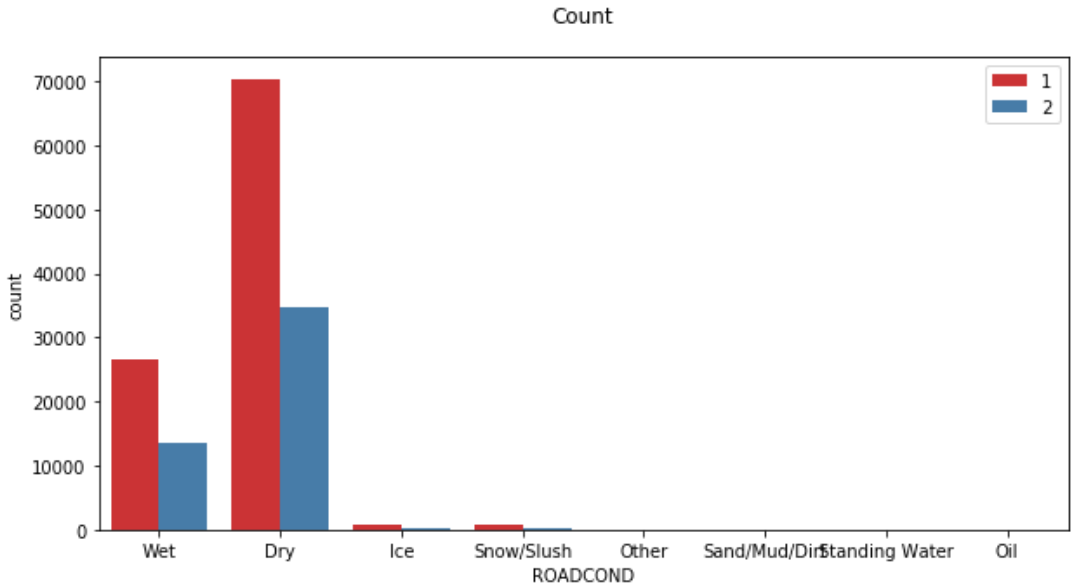
- It has high impact on severity 2.
- *I fixed inconsistency in the data (Some rows has 1, 0 values and others have N, Y values), i updated the N value with 0 and Y value with 1.*

- **Checking WEATHER against our label.**



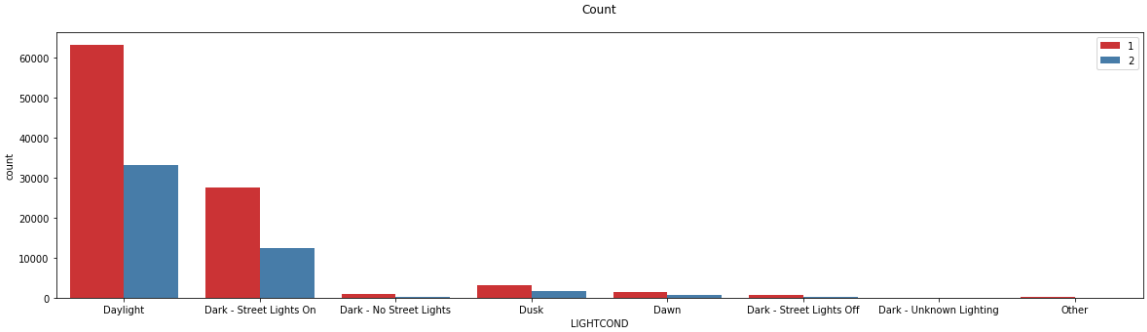
- *There are 11673 records with value UNKNOWN, as UNKOWN means that it could be any other category, so i do not consider it as a valuable data, 93 % of the UNKNOWN Values have severity code 1, and I deleted the UNKNOWN Values.*

- **Checking ROADCOND against our label.**

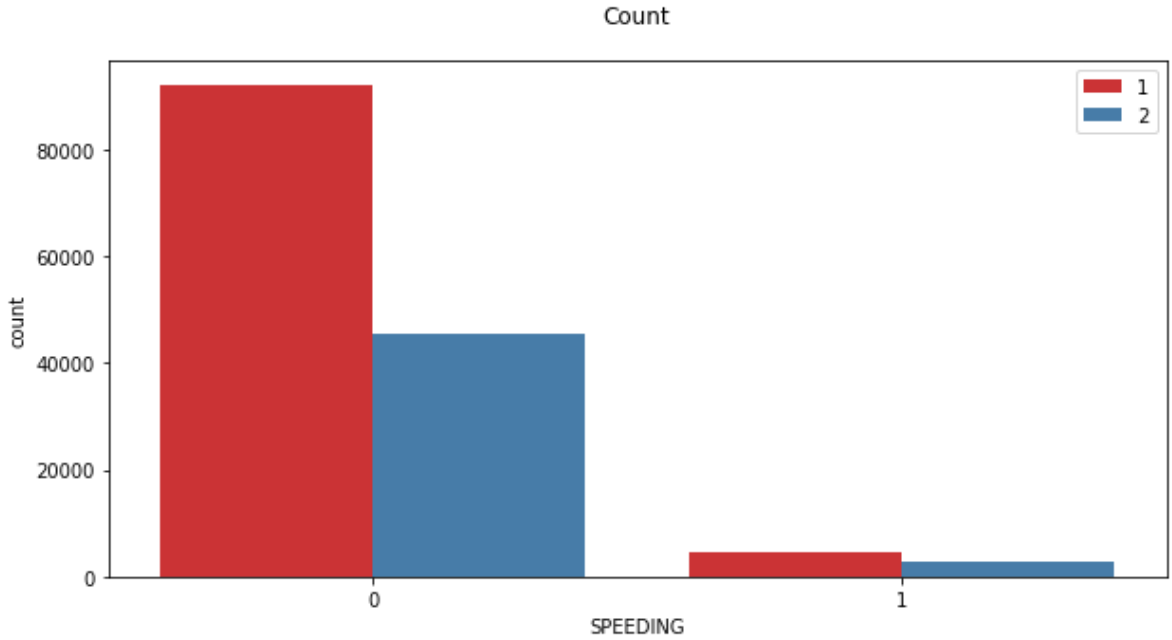


- *I deleted the Unknown values as it is misleading information.*

- **Checking LIGHTCOND against our label.**

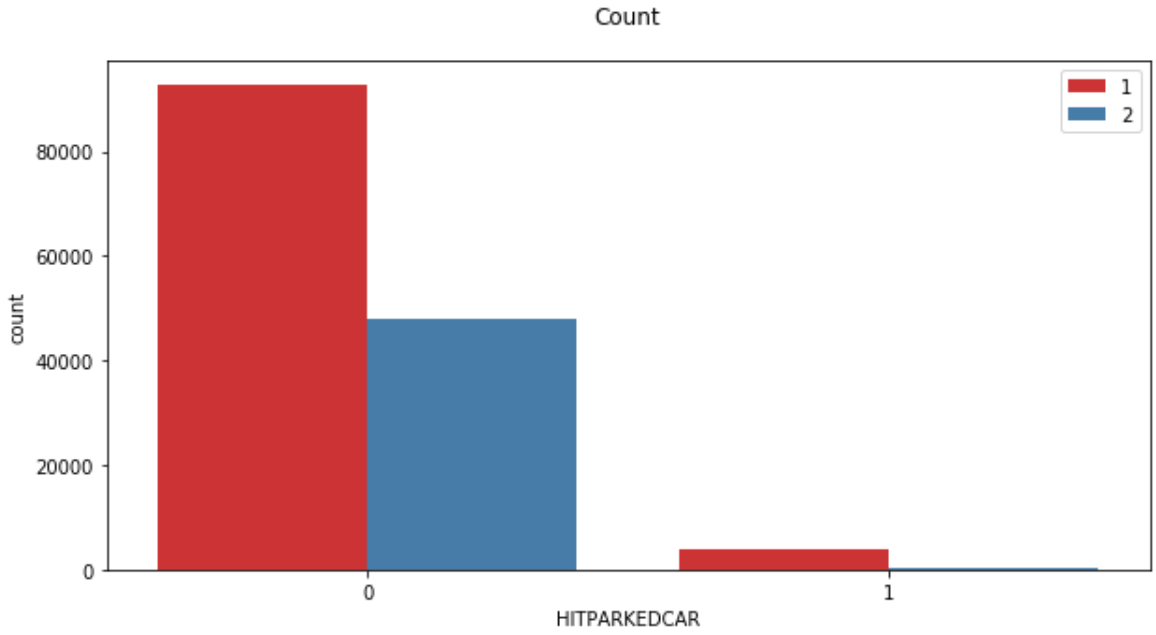


- **Checking SPEEDING against our label.**



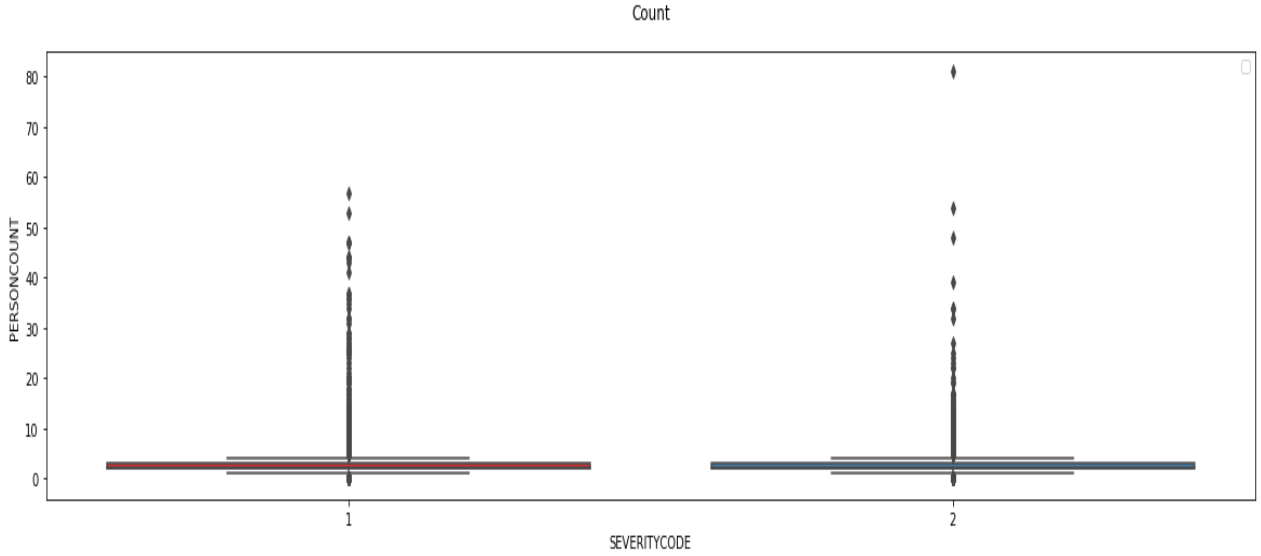
- *As we see, speed affect the High/Low severity ratio.*

- **Checking HITPARKEDCAR against our label.**



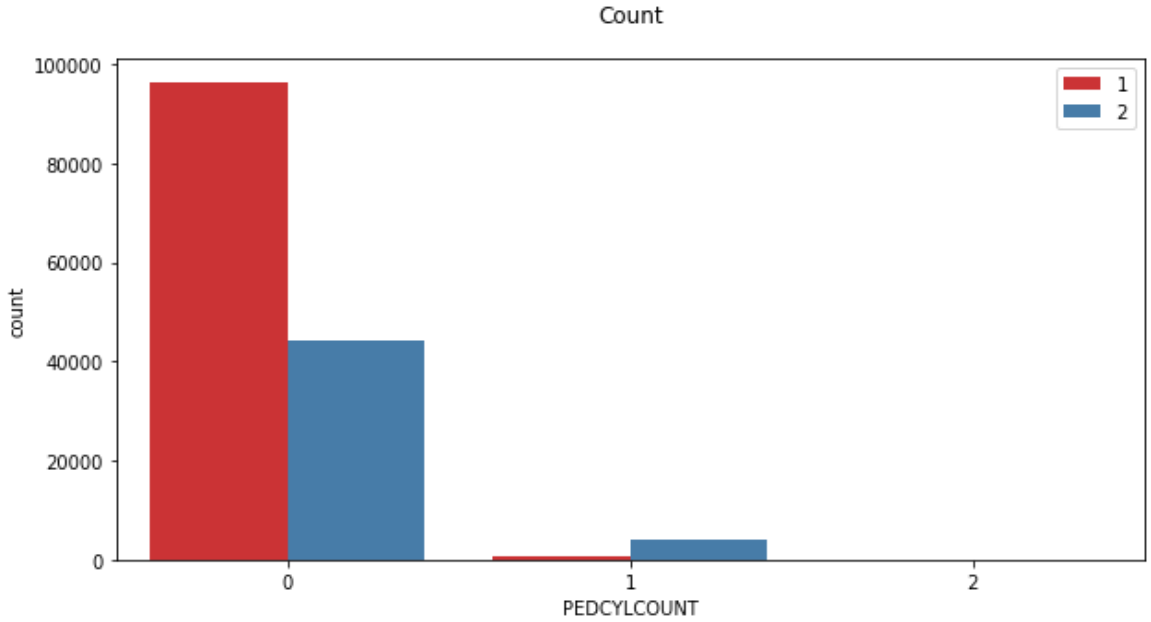
- *It does not have big impact on the severity.*

- **Checking PERSONCOUNT against our label.**



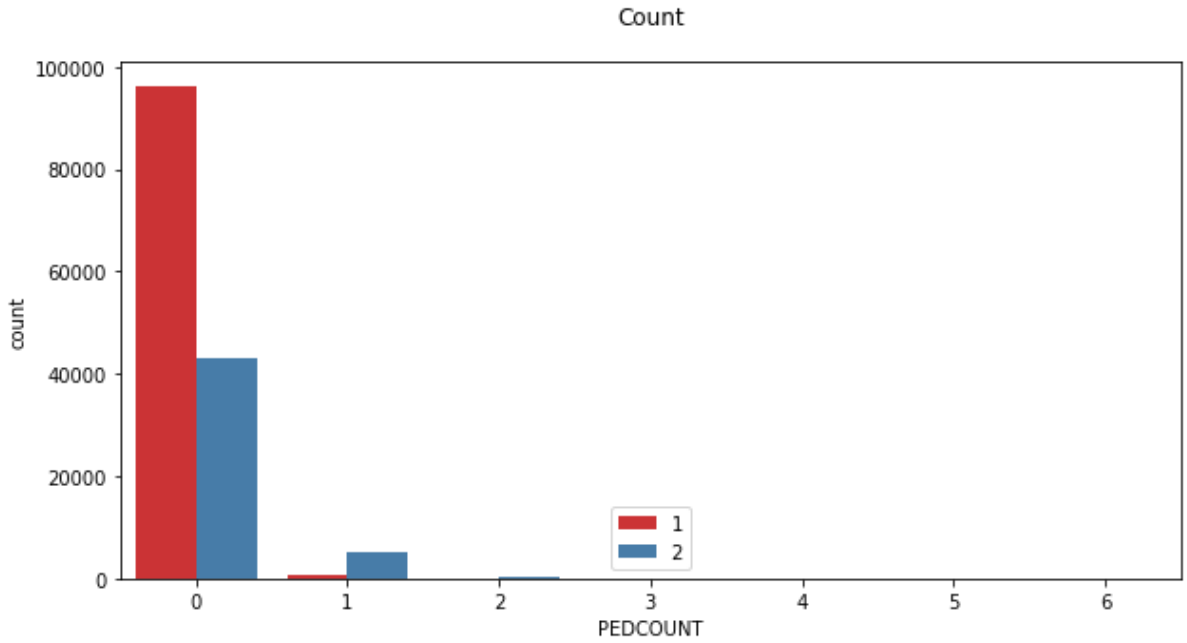
- *It has many outliers.*

- **Checking PEDCYLCOUNT against our label.**



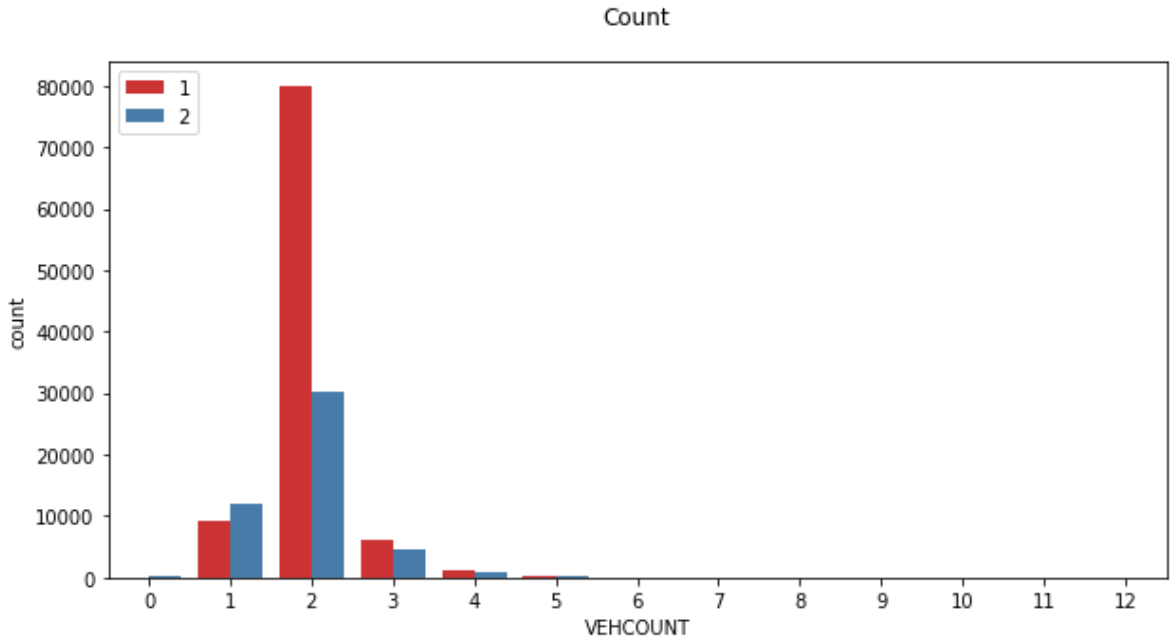
- *It clear that collision with PEDCYLCOUNT have high severity.*

- **Checking PEDCOUNT against our label.**



It clear that collision with PEDCOUNT bigger than 0 have high severity.

- **Checking VEHCOUNT against our label.**

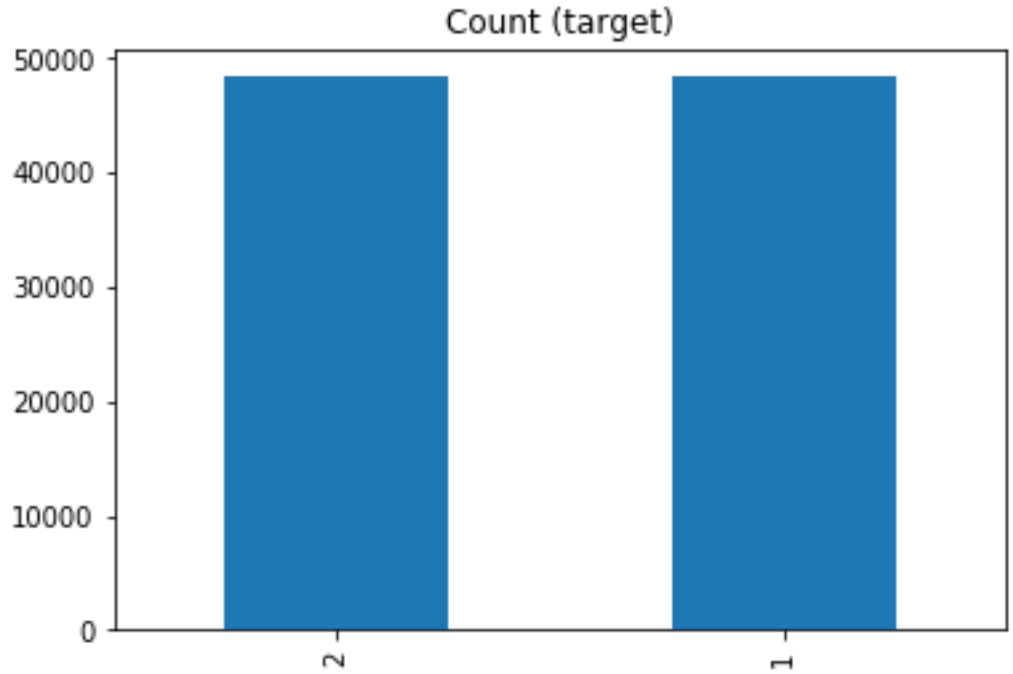


- **Checking the label SEVERITYTYPE.**

1 96859
2 48370
Name: SEVERITYCODE, dtype: int64

We see that we have unbalanced label.

As we have many records, so I decided to use under sampling to balance the dataset.

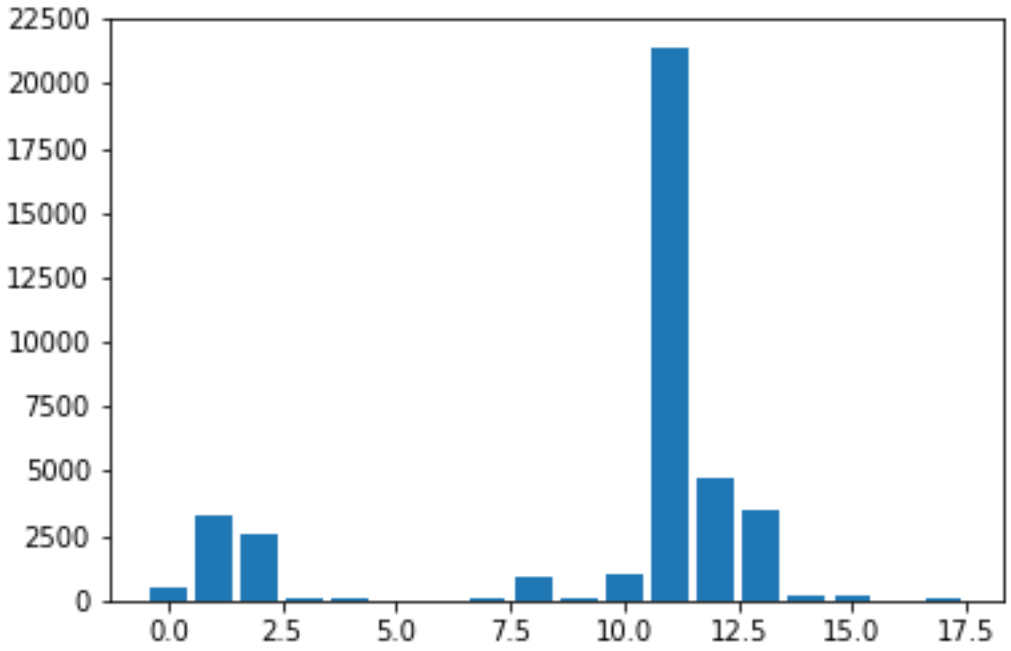


Dataset after balancing.

Features Selection

As we saw from the above analysis, that some attributes have impact on the label and we will confirm that by using some feature selection technique.

I used Chi-Squared Feature Selection and Mutual Information Feature Selection to highlight the important features and confirm my observations from the previous data analysis.




```
Feature 0 : ADDRTYPE:        544.758354
Feature 1 : COLLISIONTYPE    3292.224995
Feature 2 : JUNCTIONTYPE     2548.979959
Feature 3 : INATTENTIONIND   56.333333
Feature 4 : UNDERINFL        86.744498
Feature 5 : WEATHER          0.105541
Feature 6 : ROADCOND         15.359304
Feature 7 : LIGHTCOND        62.882922
Feature 8 : PERSONCOUNT      961.004300
Feature 9 : SPEEDING         59.770011
Feature 10: HITPARKEDCAR     1061.335013
Feature 11: ST_COLCODE       21408.436815
Feature 12: PEDCOUNT         4738.513894
Feature 13: PEDCYLCOUNT      3463.410336
Feature 14: VEHCOUNT         136.969809
Feature 15: HOUR             155.768906
Feature 16: MONTH            14.621536
Feature 17: WEEKDAY          57.432404
```

Some features have high impact but we will take into consideration what is the main requirements and the question that we are trying to solve.

COLLISIONTYPE, PERSONCOUNT, HITPARKEDCAR, ST_COLCODE, PEDCOUNT, PEDCYLCOUNT ,
These features have high impact but we will ignore them as they are known only after the collision happen, our objective is to notify people to avoid the factors that makes collision happen with high severity, so people can change the time or direction if possible.

The final select features

```
SEVERITYCODE','ADDRTYPE','JUNCTIONTYPE','WEATHER','ROADCOND','LIGHTCOND'
'INATTENTIONIND','UNDERINFL','MONTH','WEEKDAY','HOUR'
```

I used one hot encoder to encode the Categorical text attributes.

```
'ADDRTYPE','JUNCTIONTYPE','WEATHER','ROADCOND','LIGHTCOND'
```

Methods and Models

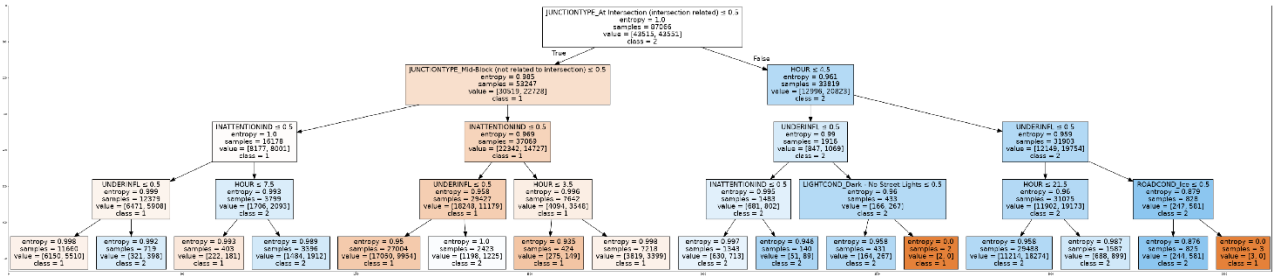
- *I created training and testing data from the dataset, with 20% testing, so I can use it in evaluation.*
- *I used four algorithm to train the model and compare the results to select the best one that fit with the cases.*

The result was better with including collision Type and other features that only known after the accidents, it was giving better accuracy when training with them but it will not fit the requirements.
Below is the result for the training set:

1- K Nearest Neighbor (KNN)

```
Accuracy:  0.6052244342020455
```

2- Decision Tree



```
Accuracy:  0.5948752873073976
```

3- Support Vector Machine

```
f1 score                  : 0.5991658031738754
jaccard_similarity_score : 0.5992411436354571
```

4- Logistic Regression

```
jaccard_similarity_score : 0.5957630519646353
log_loss                  : 0.669996197420805
```

4. Result

1- K Nearest Neighbor (KNN)

```
jaccard Accuracy          : 0.5460684997588037
f1 score                  : 0.5460579063238686
```

2- Decision Tree (Best Result)

```
jaccard                   : 0.5952036386189787
f1 score                  : 0.5943432661034356
```

3- Support Vector Machine

```
jaccard_similarity_score : 0.5941699400454827
f1 score                  : 0.594084365659453
```

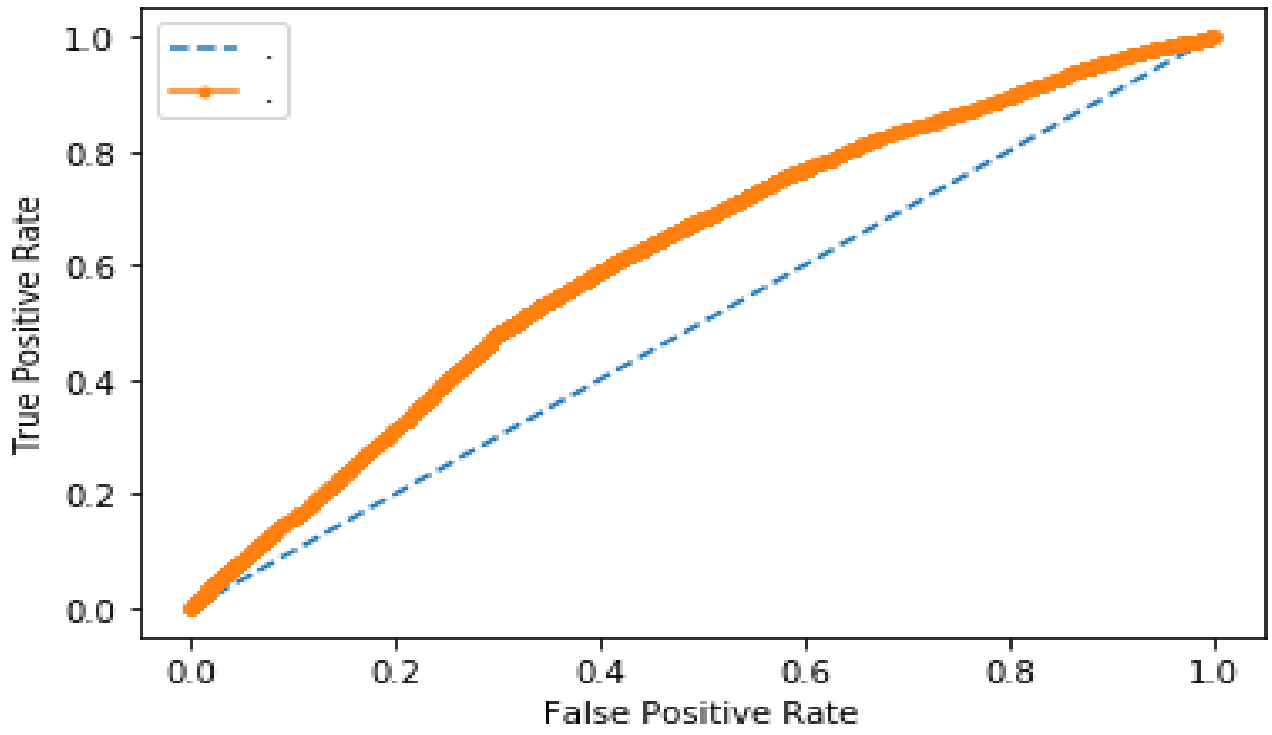
4- Logistic Regression

```
jaccard_similarity_score : 0.5941010268072496
f1 score                  : 0.5934465075124995
log_loss                  : 0.6707890936622034
```

The result after adding other features like collision type, it increase the accuracy of the model to 0.65 +

Calculating ROC curve with Logistic Regression

```
ROC AUC=0.619
```



5. Conclusion

As we have seen in the analysis part, Some features affects the severity of the accidents and all are highlighted in the data analysis section, such as Pedestrians, Collision type with Cycle and Pedestrians, Address Type ,Junction Type that involve Intersections, Some Weather Conditions, Some Road Conditions, also we have seen that some hours during the day, have bigger number of accidents, some weekdays also same, So we have to be careful and notify drivers about the Pedestrians impact and traffic departments enhance these pedestrians safety by many ways.

Regarding the prediction model, I have ignored the features that only known after the collision to focus on helping people to predict the days and the road/weather conditions that may result in high severity accident if it occurred.

6. Future Work

Expand the analysis to include the gender that might affect the result, try to identify the location of the perdetraints that have huge number of accidents and trying to avoid that.

7. References

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv.