ROAD ACCIDENT DATA ANALYSIS

November 2022

Team Members

Aman Kumar Singh
20BCE1022

Japmann Kaur Banga
20BRS1048

ABSTRACT

The paper reviews, road traffic accident data analysis and visualization in R programming environment. One of the key objectives in accident data analysis to identify the main factors associated with a road and traffic accident. The aim is to show how to extract meaningful data from the raw database and visualize it. The results revealed the hour wise, day wise, month wise and year wise plots which allowed us to observe how road traffic accidents change in timescale. Visualization and data analysis of road traffic accidents led to make conclusions which would assist reduce the number of accidents. Graph based clustering algorithms are used for visulasing the data and Supervised learning model - Multilayer Perceptron Neural Network algorithm to predict and classify accidents based on weather conditions.

Scope/Objectives

The rapid and unplanned process of urbanization has caused an unprecedented revolution in global automotive growth. The alarming increase in morbidity and mortality from road traffic accidents (RTI) in recent decades has become a major concern worldwide. Motor vehicle crashes currently rank sixth in the list of burdens of illness, and by 2025 he is projected to be third. Across India in 2021, he will kill more than 1.55,000 people in road accidents, according to data from the National Crime Records Service. That's an average of 426 deaths per day and 18 per hour, the highest death toll ever recorded.

The growing number of road and traffic accidents poses a challenge to transportation systems. It is not only concerned with health issues, but also with the economic burden on society. As a result, it is critical for the safety analysis to conduct a comparative study of road accidents in order to identify the factors that cause an accident to occur, so that preventive measures can be implemented to reduce the accident rate and severity of accident consequences. To identify the various factors associated with road accidents, we will be coming up with a comparative study of road accidents by incorporsting the R language for data manipulation and graph based clustering algorithms in Python for data visualisation and identifying influential factors.

Problem Statement

Road accidents are one of the most important factors influencing untimely death and economic loss of public and private property. Road safety is a term that refers to the planning and implementation of specific strategies to prevent road and traffic accidents. Road accident data analysis is a critical tool for identifying various factors associated with traffic accidents and can aid in lowering the accident rate.

Accurate analysis is required to respond to the sheer number of traffic accidents in one place. This analysis is performed in more detail to determine the severity of traffic accidents using supervised learning techniques such as machine learning algorithms. Accident severity is categorized as fatality, serious injury, minor injury, and motor accident.

CONTRIBUTIONS

We are aiming to identify critical parameters such as weather conditions, light conditions, which are typically important factors in finding accident severity. We have analyzed several datasets collected from different resources and the performance of different ML algorithms. In this study, we have discussed the severity of accidents which will assist the researchers to work in road safety measures and minimizing casualties.

From past surveys, it is found that, due to urbanization, the standard of living of people has upgraded, and hence has boosted an increase in both population and vehicle. Every person owns one or more than one vehicles nowadays, which has increased. This has led to the increased commotion on the roads. Transportation is a fundamental part of our lives, as every person needs a vehicle at some point in the day. People going to schools, workplaces, recreational or shopping places need vehicles.

INTRODUCTION TO DOMAIN

Heterogeneity in road accident data is highly undesirable and unavoidable. The major disadvantage of heterogeneity of road accident data is that certain relationships may remain hidden such as certain accident factors associated with particular vehicle type may not be significant in entire data set the enormity of the effect of certain accident related factors may be different for various conditions severity levels for an accident contributing factors may be different for different accident types. In order to get more accurate results this heterogeneity of road accident data must be removed to deal with this heterogeneous nature of road accident data, divide the data into groups based on some exogenous attributes e.g. accident location, road condition, cause of accident and analyzed every group separately to identify several influential factors associated with road accidents in each group.

Data analysis is the process of examining, cleaning, transforming, and modeling data with the goal of discovering useful information, drawing conclusions, and supporting decision-making. There are multiple aspects and approaches to data analysis, including different techniques with different names and used in different fields of business, science and social sciences. In today's business world, data analytics play a role in helping organizations make more scientific decisions and operate more effectively.

ALGORITHMS

MULTILAYER PERCEPTRON NEURAL NETWORKS

we have used supervised learning algorithm **Multilayer perceptron neural networks**. The network is made of 3 fully connected layer with activation function “sigmoid”. For the first layer i.e. the input layer 25 neurons are used; in the second i.e. in the hidden layer 50 neurons are used and in the final layer i.e. the output layer is consisting of only 5 neurones.

Stochastic Gradient Descent

For optimizing we have used **SGD (Stochastic Gradient Descent)** Optimizer. The model is trained for 10 epochs with batch size of 32. The model is giving alsmost 89% accuracy.

Literature Review

A study [1] used Hadoop to analyze large amounts of data based on various criteria for predicting traffic accidents. Compared to other methods, Hadoop has proven to be the most efficient method for analyzing big data. The algorithms used in this document are CCMF and TCAMP, which effectively analyze datasets to predict traffic accident risk. The proposed algorithm attempts to predict the risk of traffic accidents using vast amounts of data on vehicle behavior and conditions favorable to traffic accidents.

This paper [2] focuses on finding and predicting traffic accident patterns based on severity, road type, accident type, climate, accident time, etc. The method of finding interesting and useful patterns from spatial databases is called spatial data mining. The spatio-temporal algorithm finds hidden patterns easier than traditional data mining techniques.

In [3], traffic accidents are inferred from heterogeneous data. A vast amount of heterogeneous data, including accident data and GPS recordings, was collected to investigate how vehicle mobility affects traffic accidents. This data is analyzed to build a Stack Denoise Auto-Encoder model that examines human locomotion characteristics to predict crash risk. The preparedness model can be used to simulate real-time accident risk and warn people of potential accidents to ensure safer routes and travel.

Anupama McCaret. In Al [4] he analyzed an accident data set of the last few years to predict traffic accidents. The approach proposed in this paper involves merging machine learning algorithms such as Bayes net, j48 graft and j48 decision tree in the data mining process, and studies the performance of algorithms in predicting accidents. Therefore, we have found that combining such algorithms yields better results than using a single algorithm. The results obtained help predict traffic accidents and contribute to their prevention and control.

This paper [5] used two predictive models to analyze historical and current accident data to predict the number of accidents that will occur this year. Multiple Linear regression and artificial neural networks were used for predictive analysis. After the analysis is performed, we conclude that the regression model had a larger error in the predicted values. On the other hand, predictions from artificial neural network analysis were more accurate and had fewer errors. Therefore, ANN has proven to be a better method for making his predictions of accidents.

EXISTING METHODOLOGIES AND LIMITATIONS SUMMARY

A hybrid approach of K-Means and Random Forest (RF) was developed to obtain the most important traffic accident variables. K-means extracts hidden information from traffic accident data and creates new functions in the training set. The distance between each cluster and the connecting line of k1 and k9 is calculated and chosen as the k-max value. k is the optimal value for splitting the training set. RF is used for severity prediction classification of accidents.

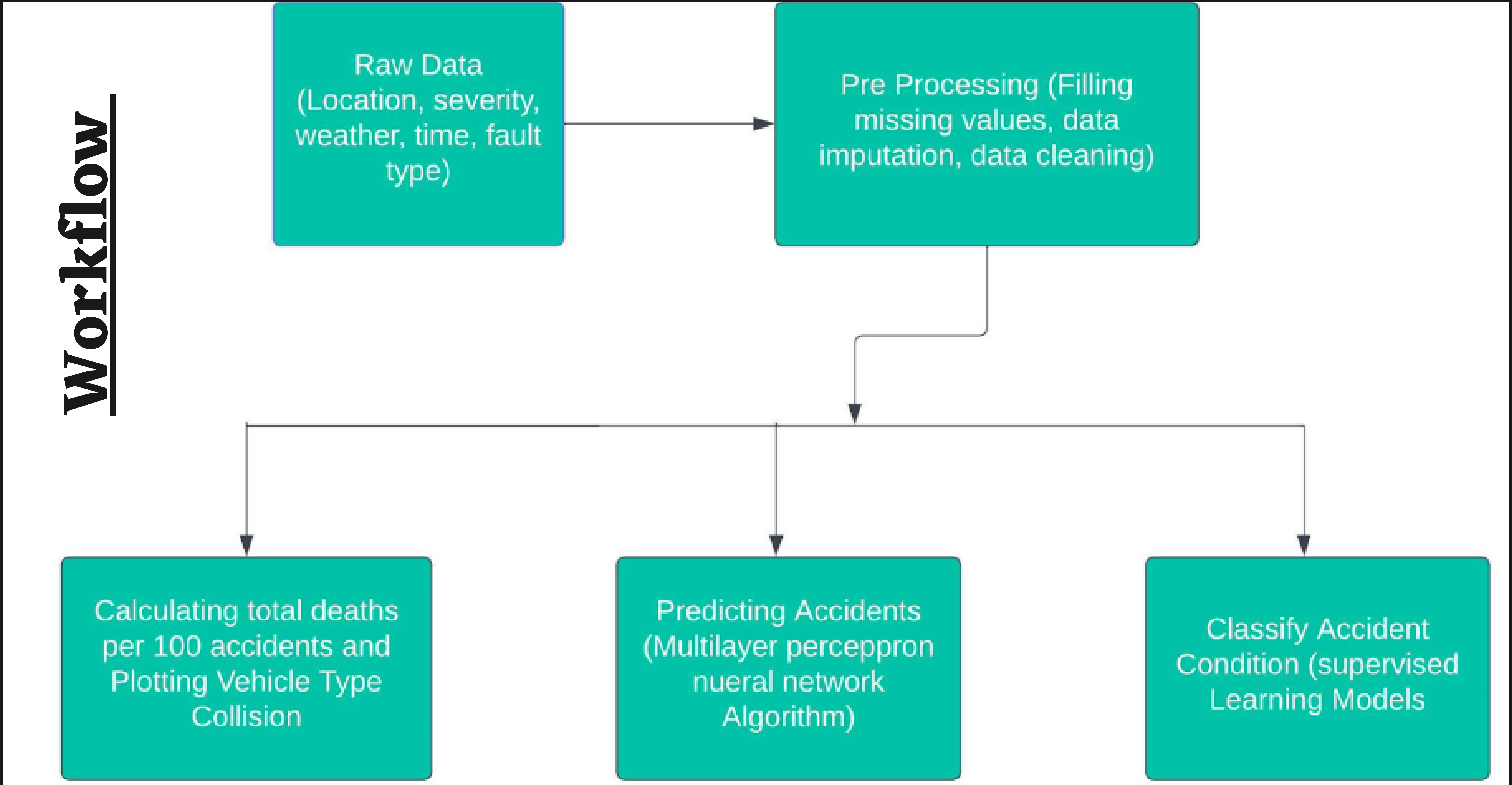
Other methodologies include multiple logistic regression and the pattern recognition type of artificial neural network (ANN) as a machine learning solution are used to recognize the most influential variables on the severity of accidents and the superior approach for accident prediction.

Limitations:

- The First limitation being the lack of availability of data-sets for Indian Road Accidents. There are no data-sets for Indian Road Accidents whereas we can find numerous data sets for countries like the USA, the UK, etc. If the data sets are found then one can easily train the data and can help in reducing fatalities in road accidents.
- Most current accident analysis techniques are event-based and do not adequately capture the dynamic complexity and non-linear interactions that characterize accidents in complex systems.
- Estimation methods vary from country to country. The composition of different vehicle fleets can lead to bias in mortality estimates and comparisons. Significantly higher than in Europe, which is misleading, as the risk of death while riding a motorcycle or moped is much higher than while driving a car.

PROPOSED PROJECT WORK IN DETAIL

Workflow



Raw Data (Location, severity, weather, time, fault type)

Pre Processing (Filling missing values, data imputation, data cleaning)

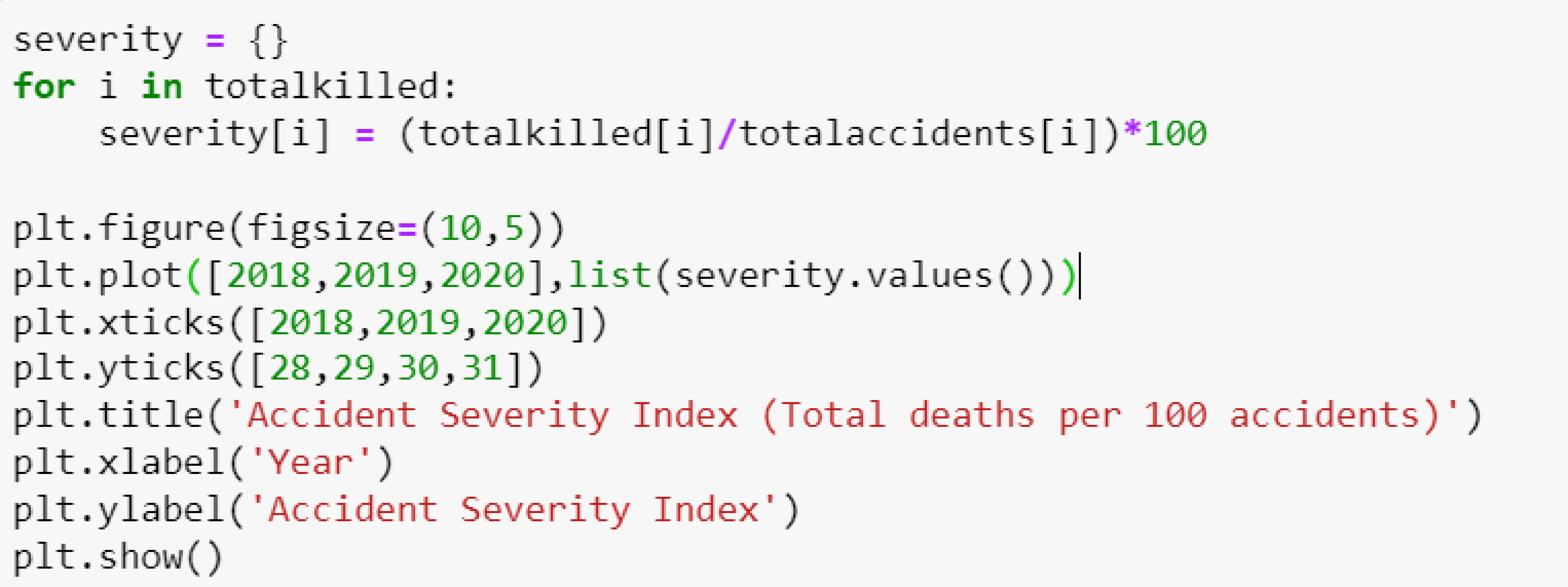
Calculating total deaths per 100 accidents and Plotting Vehicle Type Collision

Predicting Accidents (Multilayer perceppron nueral network Algorithm)

Classify Accident Condition (supervised Learning Models

Modules/ Algorithm Phases

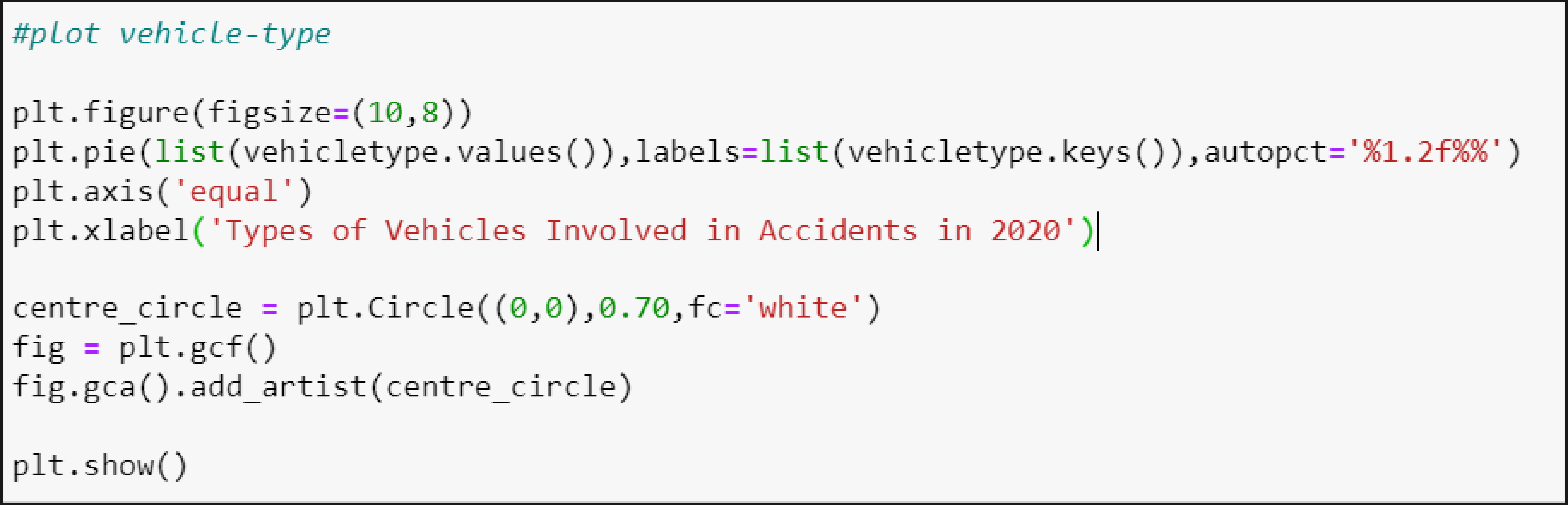
Accident Severity



```
severity = {}
for i in totalkilled:
    severity[i] = (totalkilled[i]/totalaccidents[i])*100

plt.figure(figsize=(10,5))
plt.plot([2018,2019,2020],list(severity.values()))
plt.xticks([2018,2019,2020])
plt.yticks([28,29,30,31])
plt.title('Accident Severity Index (Total deaths per 100 accidents)')
plt.xlabel('Year')
plt.ylabel('Accident Severity Index')
plt.show()
```

Modules/ Algorithm Phases

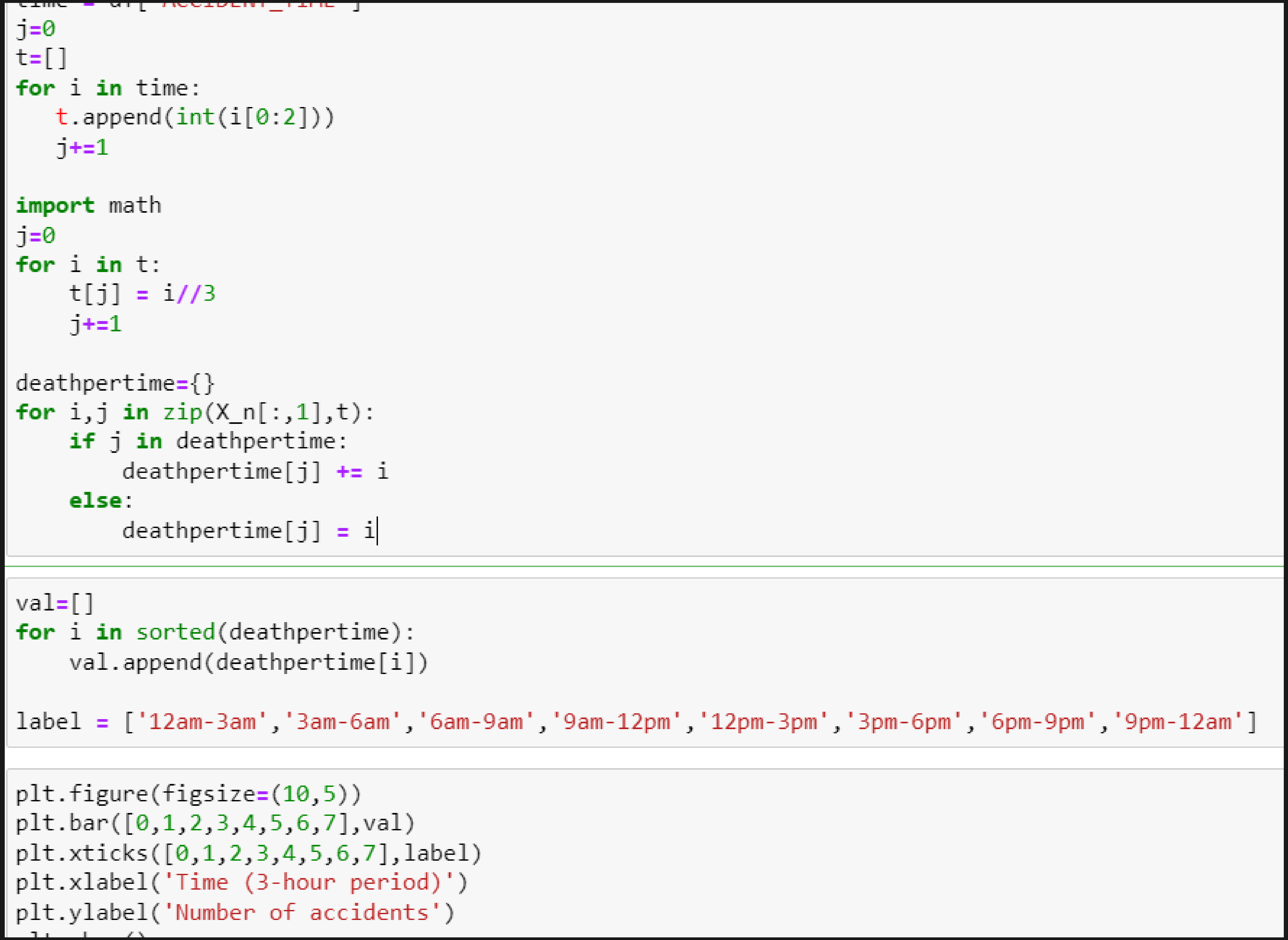
Vehicle type involved in accidents (2020)



```
#plot vehicle-type

plt.figure(figsize=(10,8))
plt.pie(list(vehicletype.values()),labels=list(vehicletype.keys()),autopct='%1.2f%%')
plt.axis('equal')
plt.xlabel('Types of Vehicles Involved in Accidents in 2020')

centre_circle = plt.Circle((0,0),0.70,fc='white')
fig = plt.gcf()
fig.gca().add_artist(centre_circle)

plt.show()
```

Modules/ Algorithm Phases

Hourly Basis Accidents



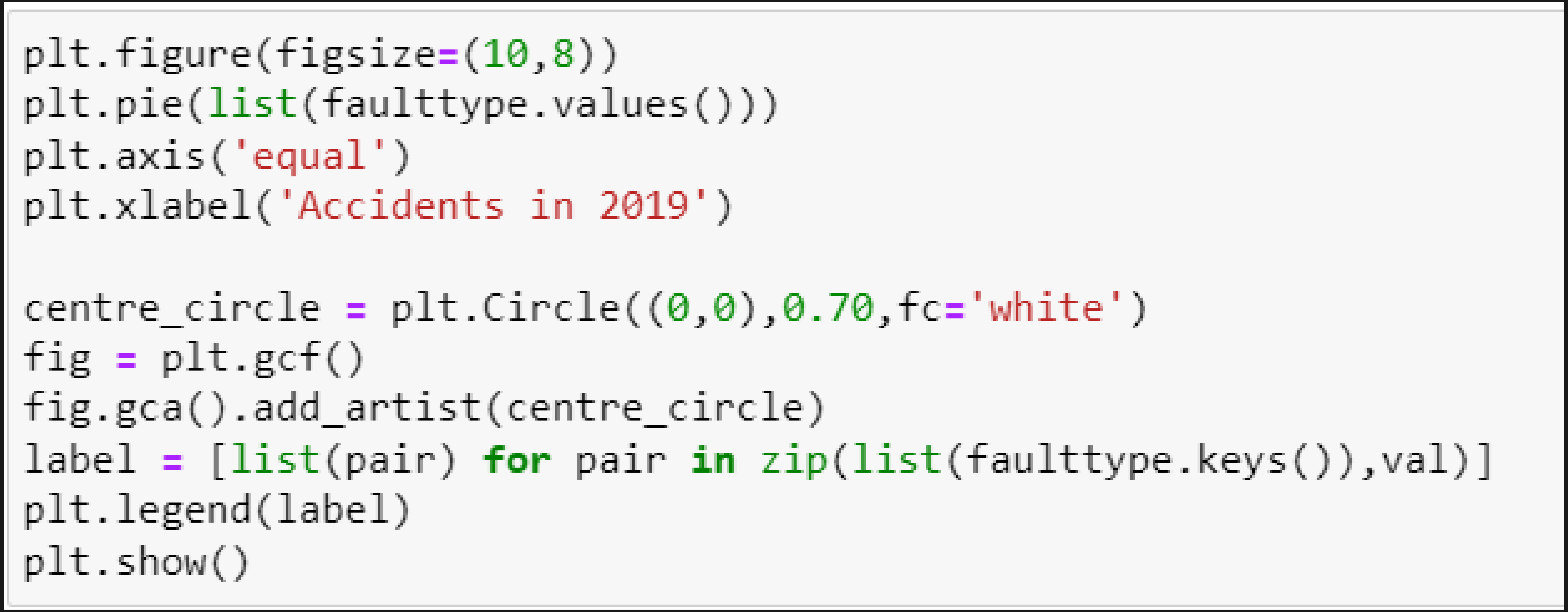
```
time = df['ACCIDENT_TIME']
j=0
t=[]
for i in time:
   t.append(int(i[0:2]))
   j+=1

import math
j=0
for i in t:
    t[j] = i//3
    j+=1

deathpertime={}
for i,j in zip(X_n[:,1],t):
    if j in deathpertime:
        deathpertime[j] += i
    else:
        deathpertime[j] = i
```

```
val=[]
for i in sorted(deathpertime):
    val.append(deathpertime[i])

label = ['12am-3am','3am-6am','6am-9am','9am-12pm','12pm-3pm','3pm-6pm','6pm-9pm','9pm-12am']
```

```
plt.figure(figsize=(10,5))
plt.bar([0,1,2,3,4,5,6,7],val)
plt.xticks([0,1,2,3,4,5,6,7],label)
plt.xlabel('Time (3-hour period)')
plt.ylabel('Number of accidents')
```

Modules/ Algorithm Phases

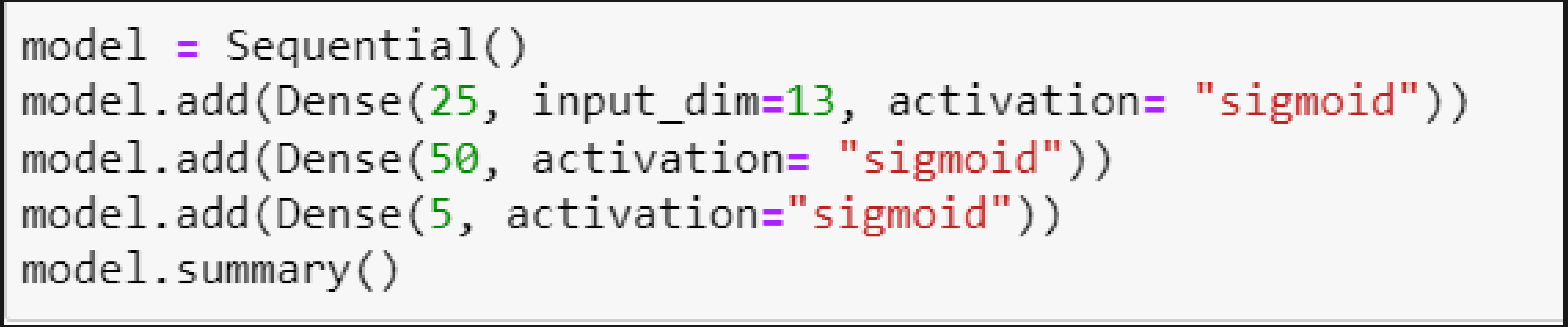
Accidents based on Fault Type



```
plt.figure(figsize=(10,8))
plt.pie(list(faulttype.values()))
plt.axis('equal')
plt.xlabel('Accidents in 2019')

centre_circle = plt.Circle((0,0),0.70,fc='white')
fig = plt.gcf()
fig.gca().add_artist(centre_circle)
label = [list(pair) for pair in zip(list(faulttype.keys()),val)]
plt.legend(label)
plt.show()
```

Modules/ Algorithm Phases

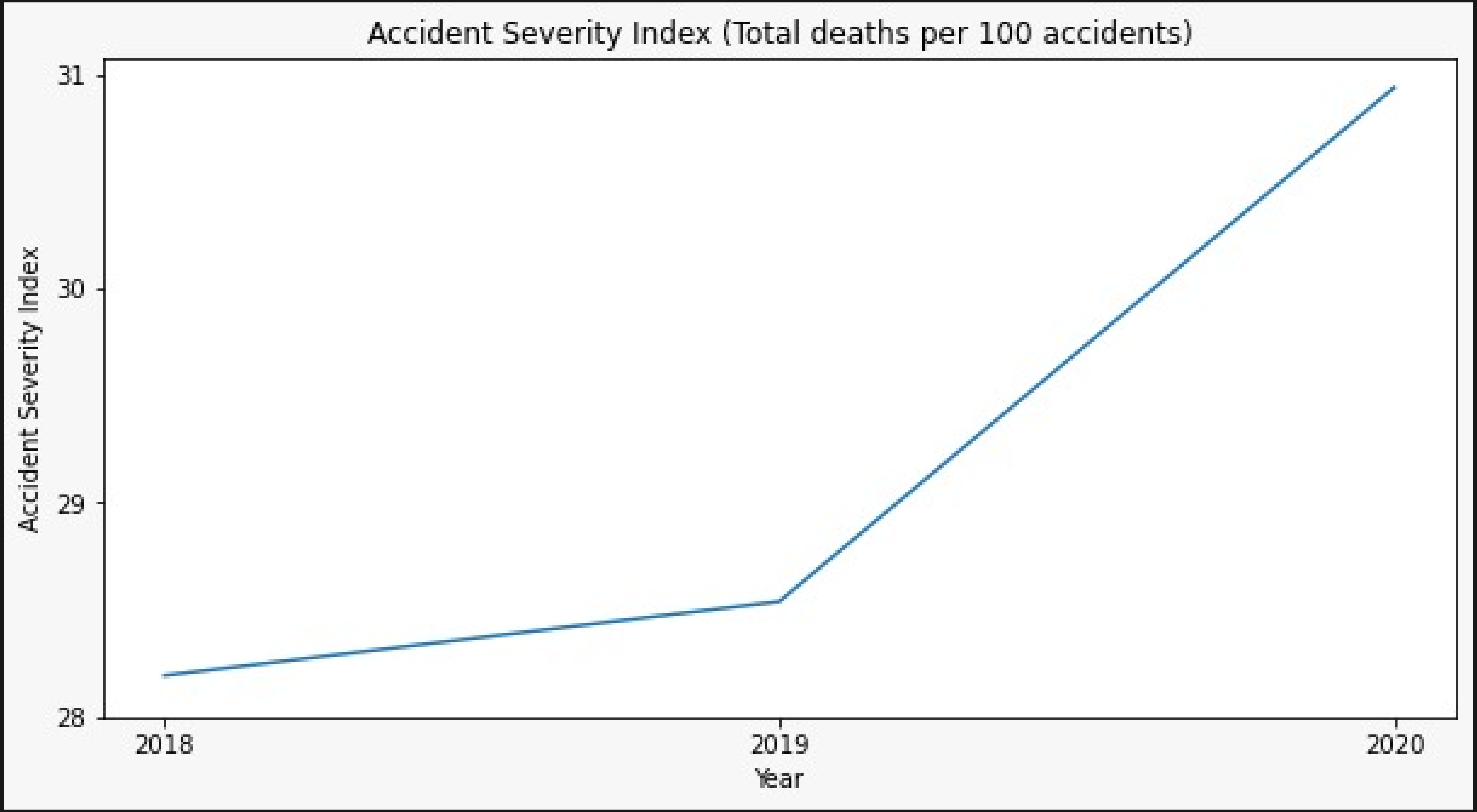
Multilayer Perceptron Neural Networks



```
model = Sequential()
model.add(Dense(25, input_dim=13, activation= "sigmoid"))
model.add(Dense(50, activation= "sigmoid"))
model.add(Dense(5, activation="sigmoid"))
model.summary()
```

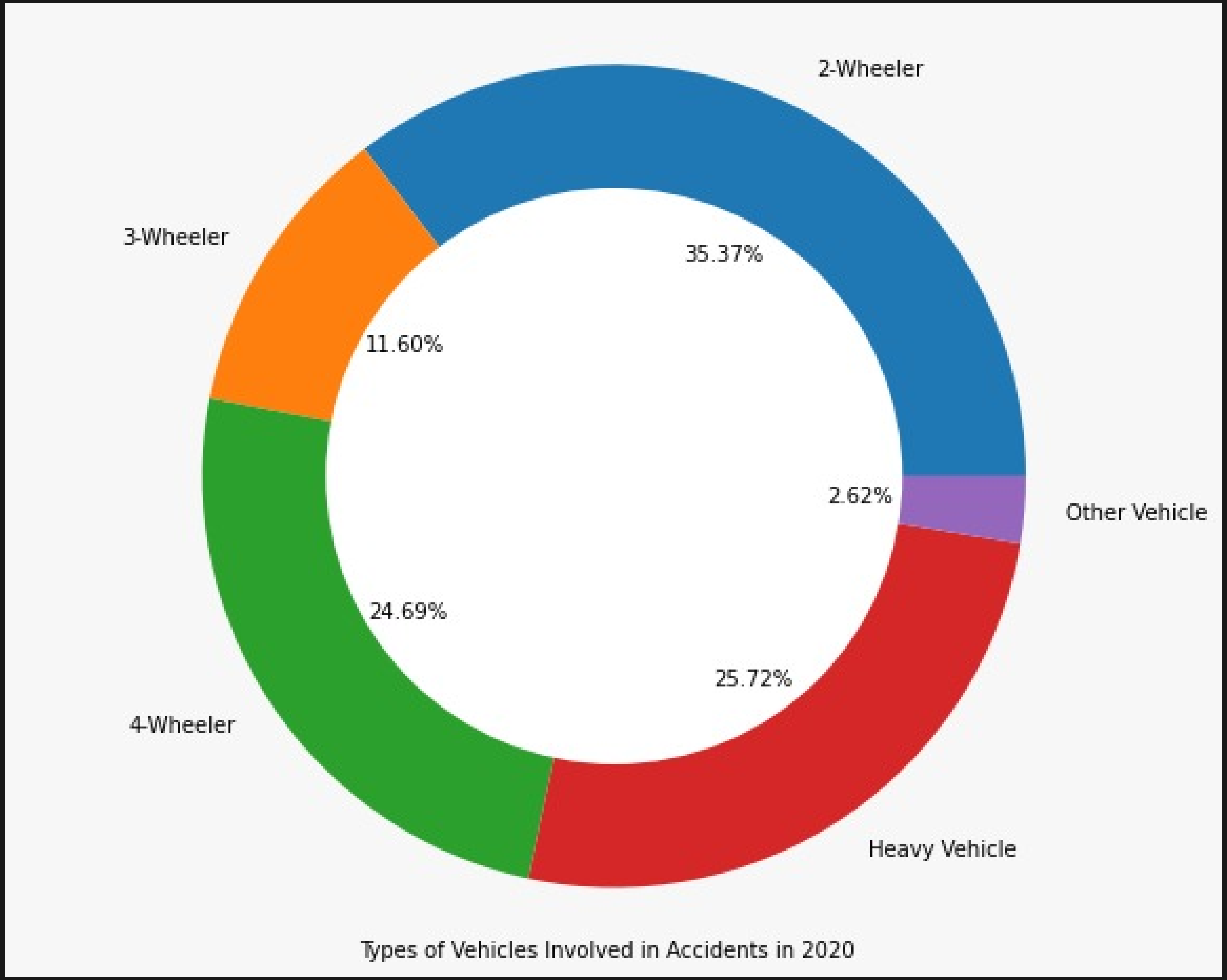
RESULTS

Accident Severity



RESULTS

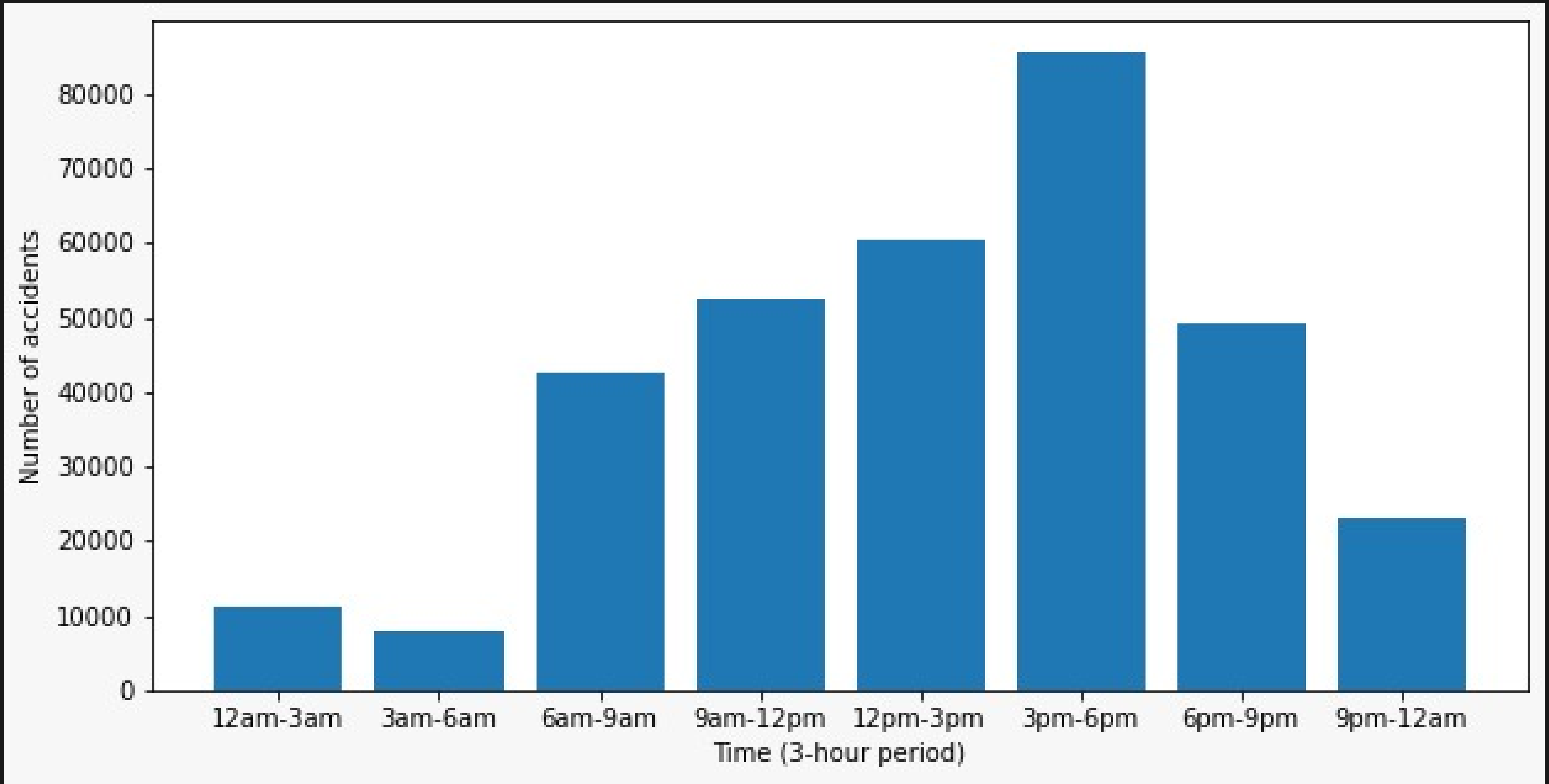
Vehicle types involved in Accidents



Types of Vehicles Involved in Accidents in 2020

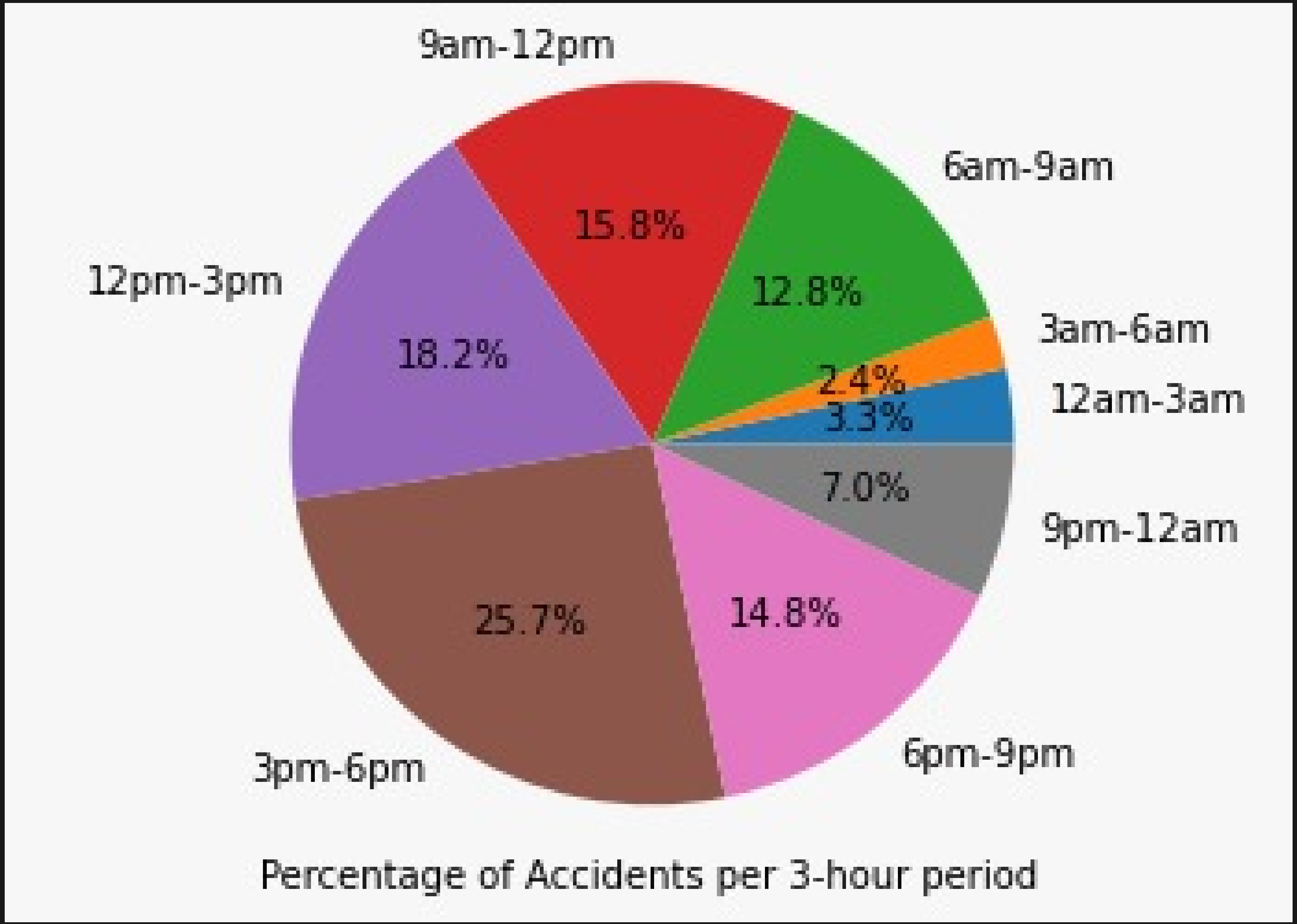
RESULTS

Rush hours for accidents



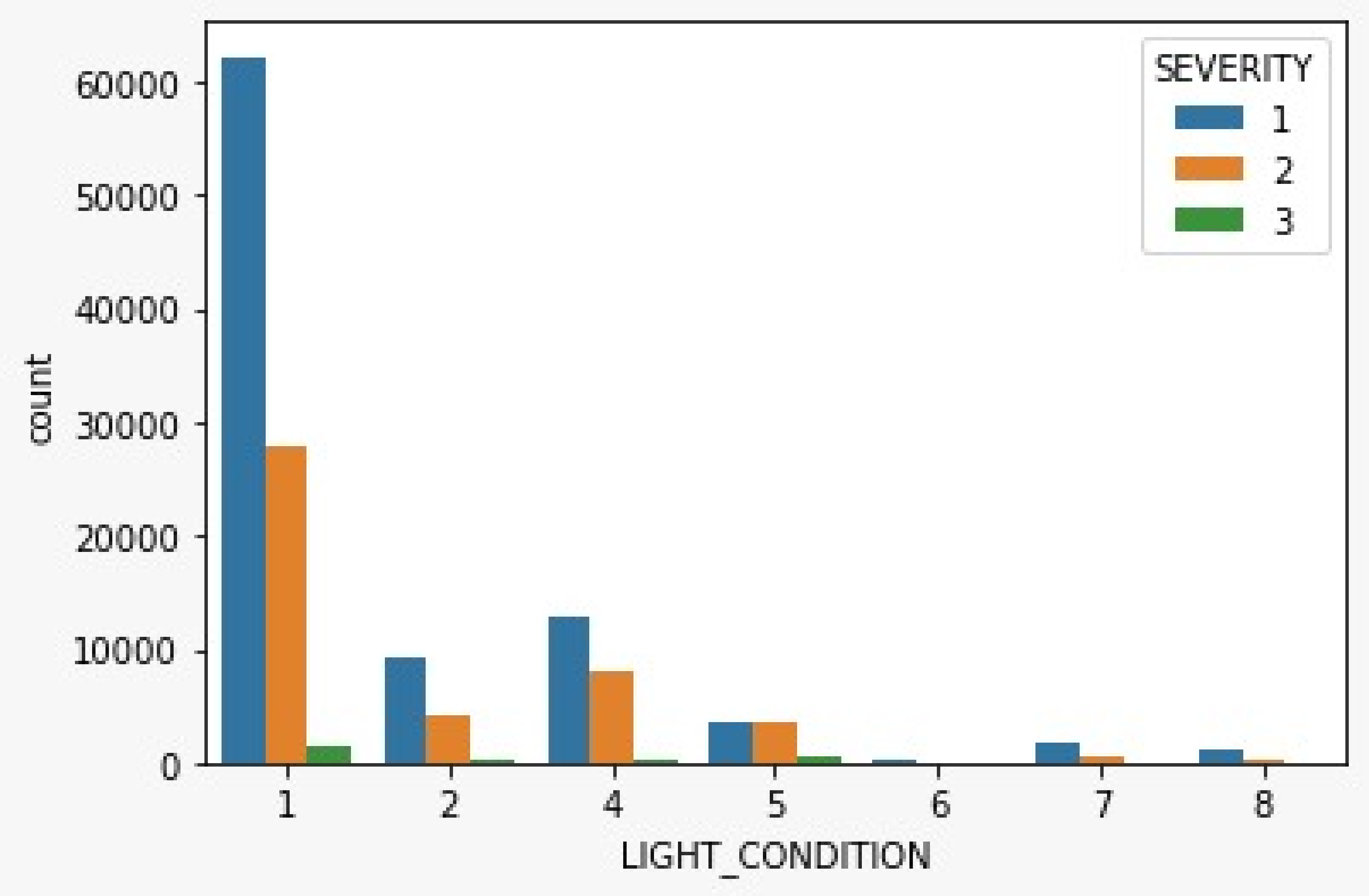
RESULTS

Rush hours for accidents



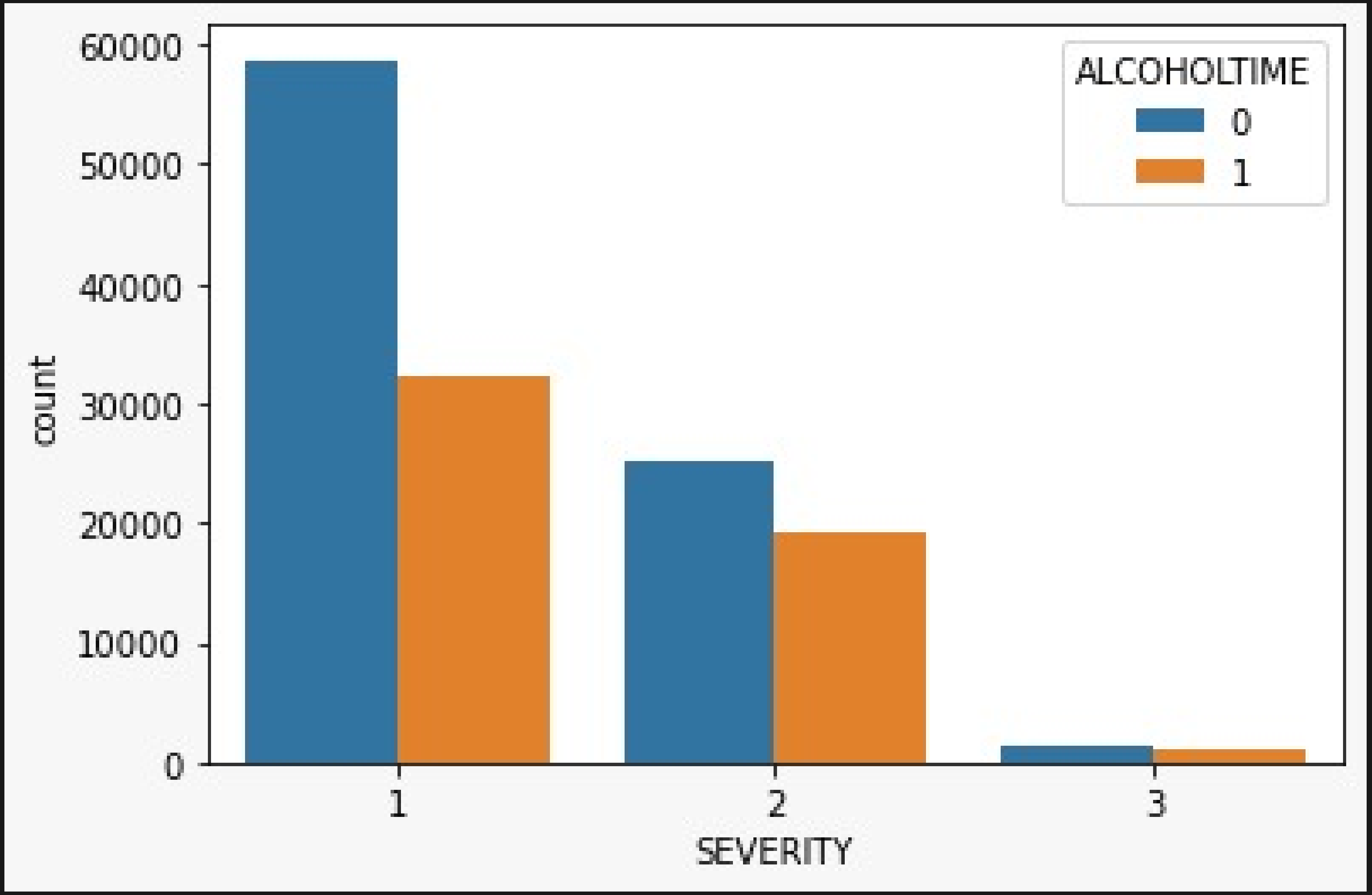
RESULTS

Light Conditions



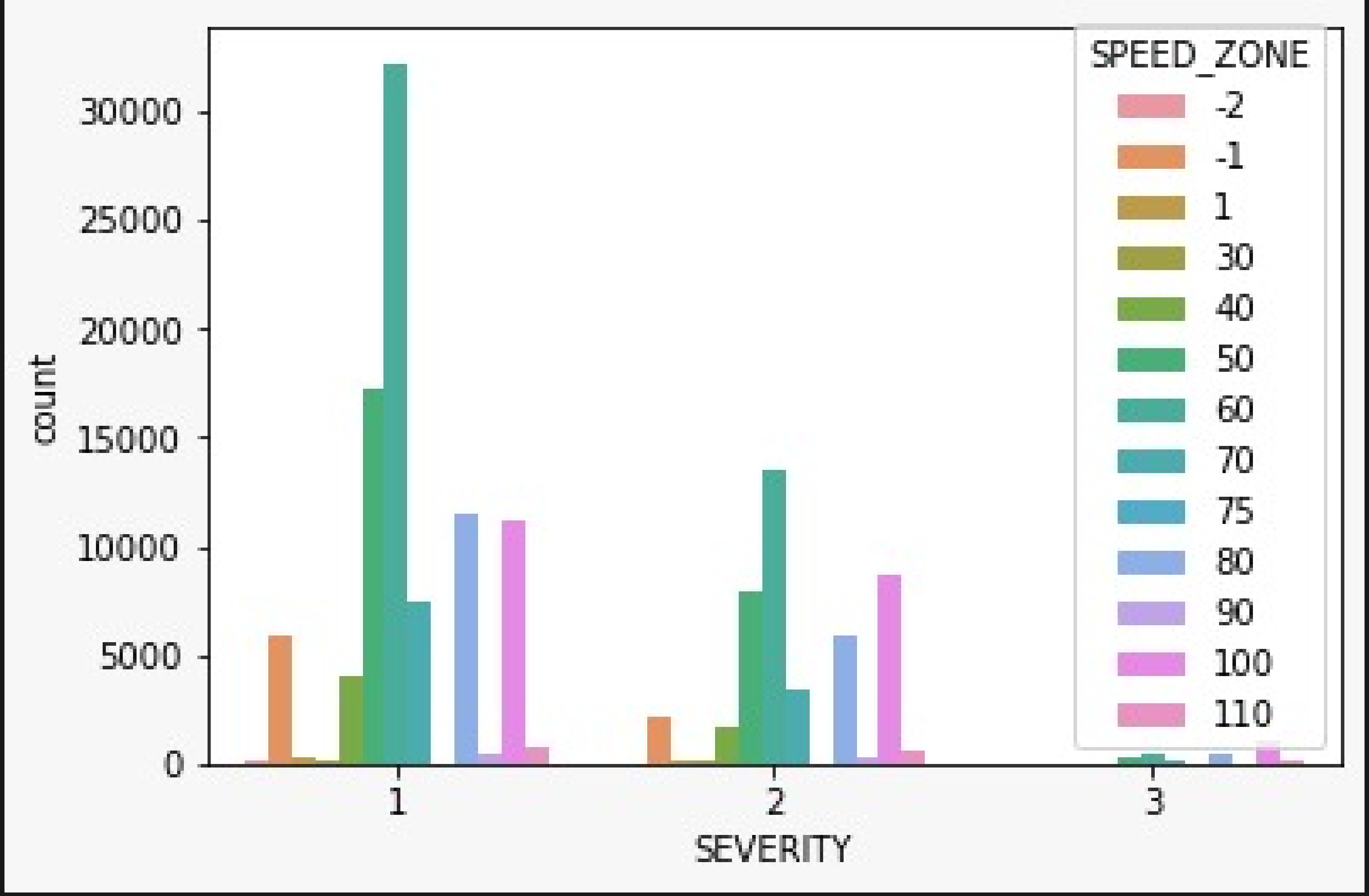
RESULTS

Severity and Alcohol Time



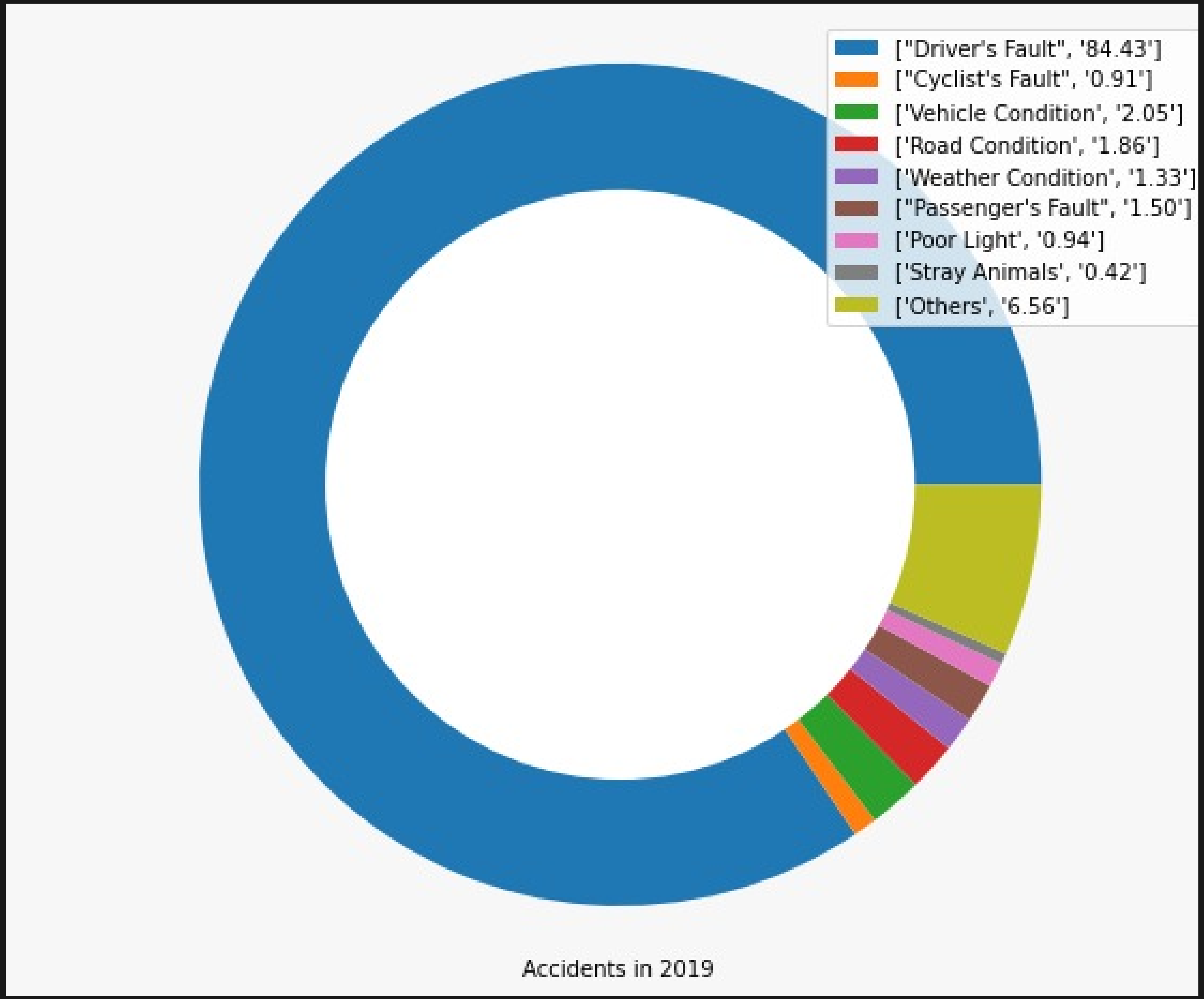
RESULTS

Severity and Speed Time



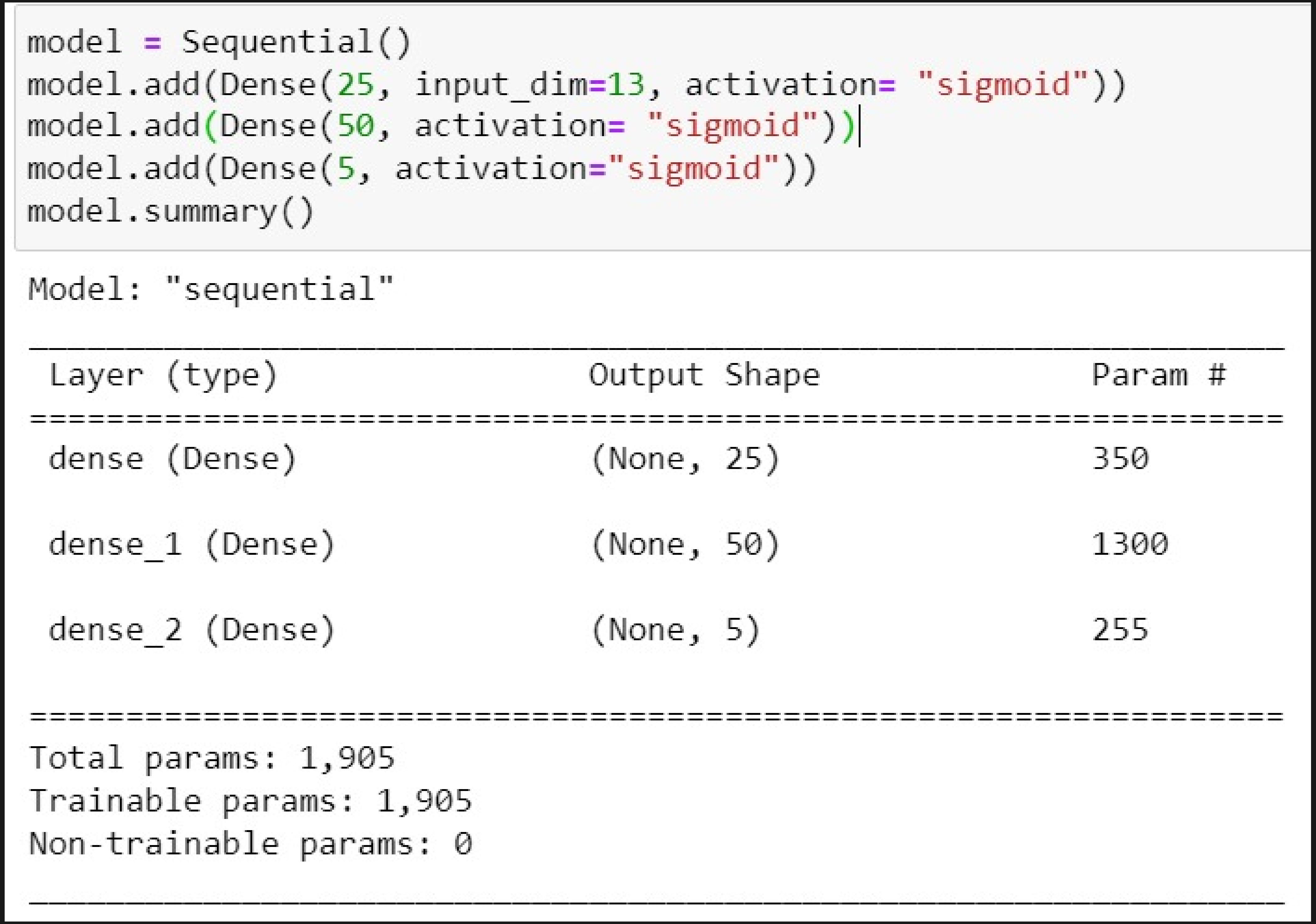
RESULTS

Fault Type based accidents



RESULTS

Model Training - Multilayer Perceptron Neural Network

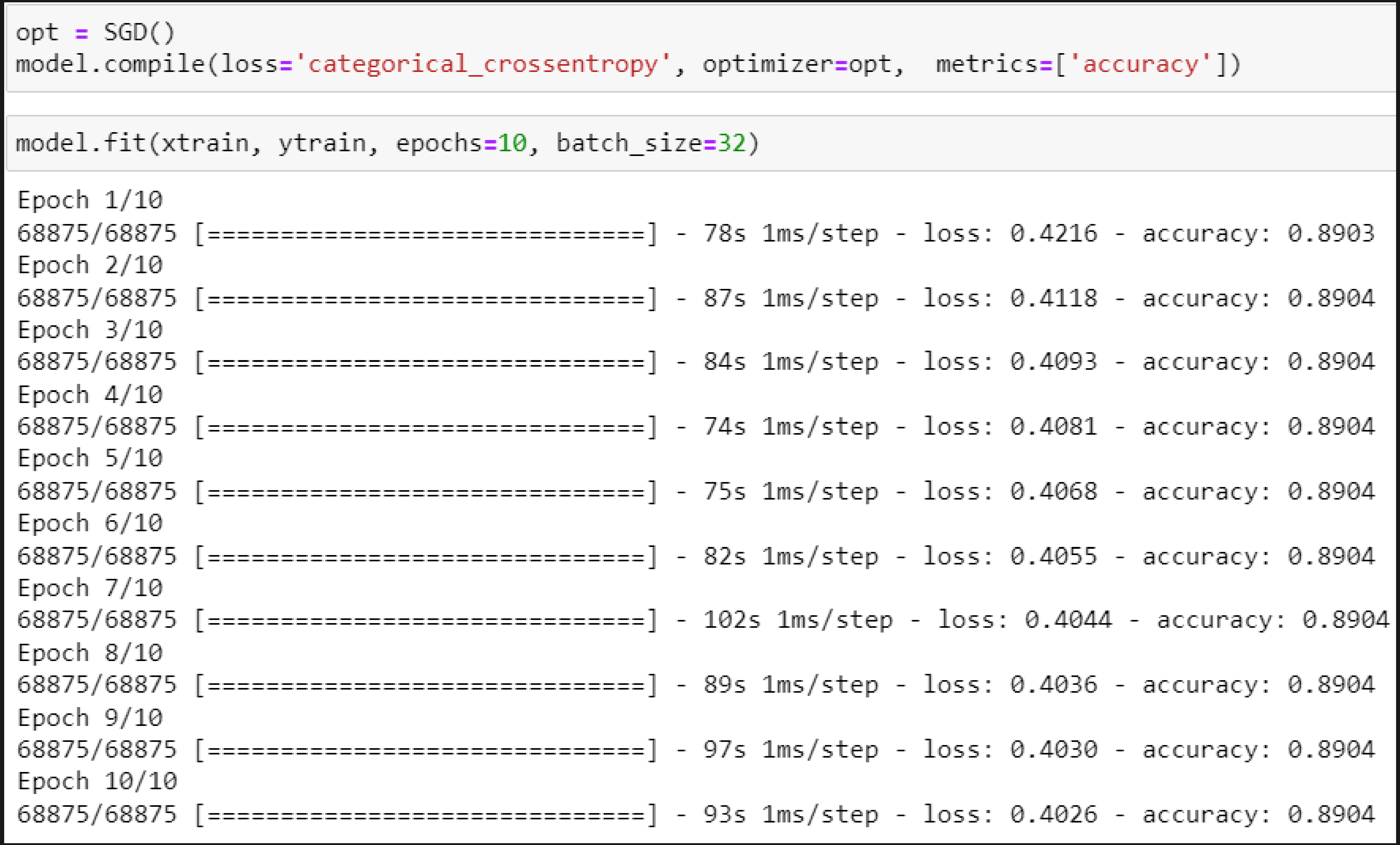


```
model = Sequential()
model.add(Dense(25, input_dim=13, activation= "sigmoid"))
model.add(Dense(50, activation= "sigmoid"))
model.add(Dense(5, activation="sigmoid"))
model.summary()
```

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 25)                350

 dense_1 (Dense)             (None, 50)                1300

 dense_2 (Dense)             (None, 5)                 255

=================================================================
Total params: 1,905
Trainable params: 1,905
Non-trainable params: 0
_________________________________________________________________
```

RESULTS

Stochastic Gradient Descent



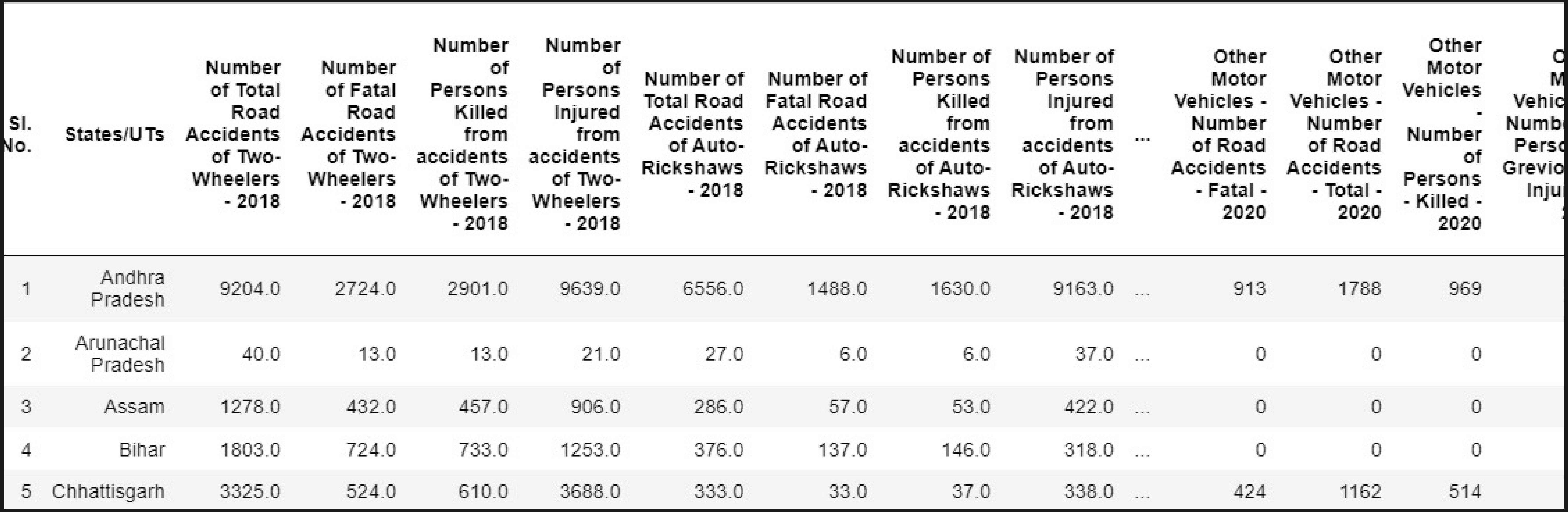
```
opt = SGD()
model.compile(loss='categorical_crossentropy', optimizer=opt,  metrics=['accuracy'])
```

```
model.fit(xtrain, ytrain, epochs=10, batch_size=32)
```

```
Epoch 1/10
68875/68875 [==============================] - 78s 1ms/step - loss: 0.4216 - accuracy: 0.8903
Epoch 2/10
68875/68875 [==============================] - 87s 1ms/step - loss: 0.4118 - accuracy: 0.8904
Epoch 3/10
68875/68875 [==============================] - 84s 1ms/step - loss: 0.4093 - accuracy: 0.8904
Epoch 4/10
68875/68875 [==============================] - 74s 1ms/step - loss: 0.4081 - accuracy: 0.8904
Epoch 5/10
68875/68875 [==============================] - 75s 1ms/step - loss: 0.4068 - accuracy: 0.8904
Epoch 6/10
68875/68875 [==============================] - 82s 1ms/step - loss: 0.4055 - accuracy: 0.8904
Epoch 7/10
68875/68875 [==============================] - 102s 1ms/step - loss: 0.4044 - accuracy: 0.8904
Epoch 8/10
68875/68875 [==============================] - 89s 1ms/step - loss: 0.4036 - accuracy: 0.8904
Epoch 9/10
68875/68875 [==============================] - 97s 1ms/step - loss: 0.4030 - accuracy: 0.8904
Epoch 10/10
68875/68875 [==============================] - 93s 1ms/step - loss: 0.4026 - accuracy: 0.8904
```

RESULTS - TABLES

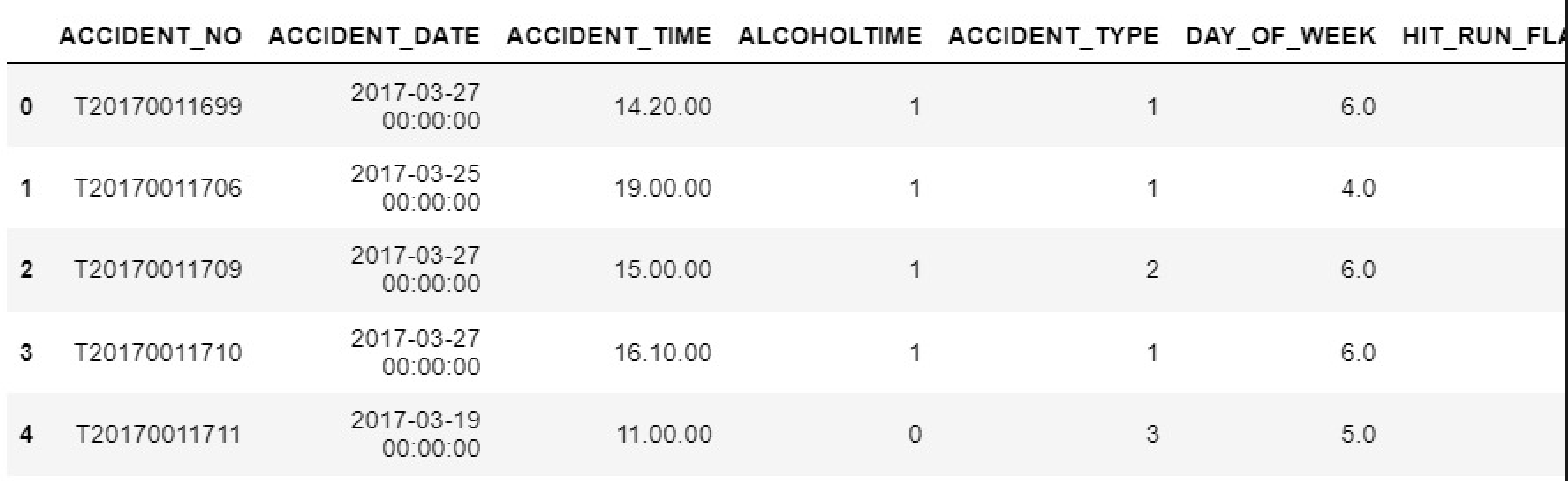
Accidents



Sl. No.	States/UTs	Number of Total Road Accidents of Two-Wheelers - 2018	Number of Fatal Road Accidents of Two-Wheelers - 2018	Number of Persons Killed from accidents of Two-Wheelers - 2018	Number of Persons Injured from accidents of Two-Wheelers - 2018	Number of Total Road Accidents of Auto-Rickshaws - 2018	Number of Fatal Road Accidents of Auto-Rickshaws - 2018	Number of Persons Killed from accidents of Auto-Rickshaws - 2018	Number of Persons Injured from accidents of Auto-Rickshaws - 2018	...	Other Motor Vehicles - Number of Road Accidents - Fatal - 2020	Other Motor Vehicles - Number of Road Accidents - Total - 2020	Other Motor Vehicles - Number of Persons - Killed - 2020	O M Vehic Numb Perso Grevio Inju
1	Andhra Pradesh	9204.0	2724.0	2901.0	9639.0	6556.0	1488.0	1630.0	9163.0	...	913	1788	969	
2	Arunachal Pradesh	40.0	13.0	13.0	21.0	27.0	6.0	6.0	37.0	...	0	0	0	
3	Assam	1278.0	432.0	457.0	906.0	286.0	57.0	53.0	422.0	...	0	0	0	
4	Bihar	1803.0	724.0	733.0	1253.0	376.0	137.0	146.0	318.0	...	0	0	0	
5	Chhattisgarh	3325.0	524.0	610.0	3688.0	333.0	33.0	37.0	338.0	...	424	1162	514	

RESULTS - TABLES

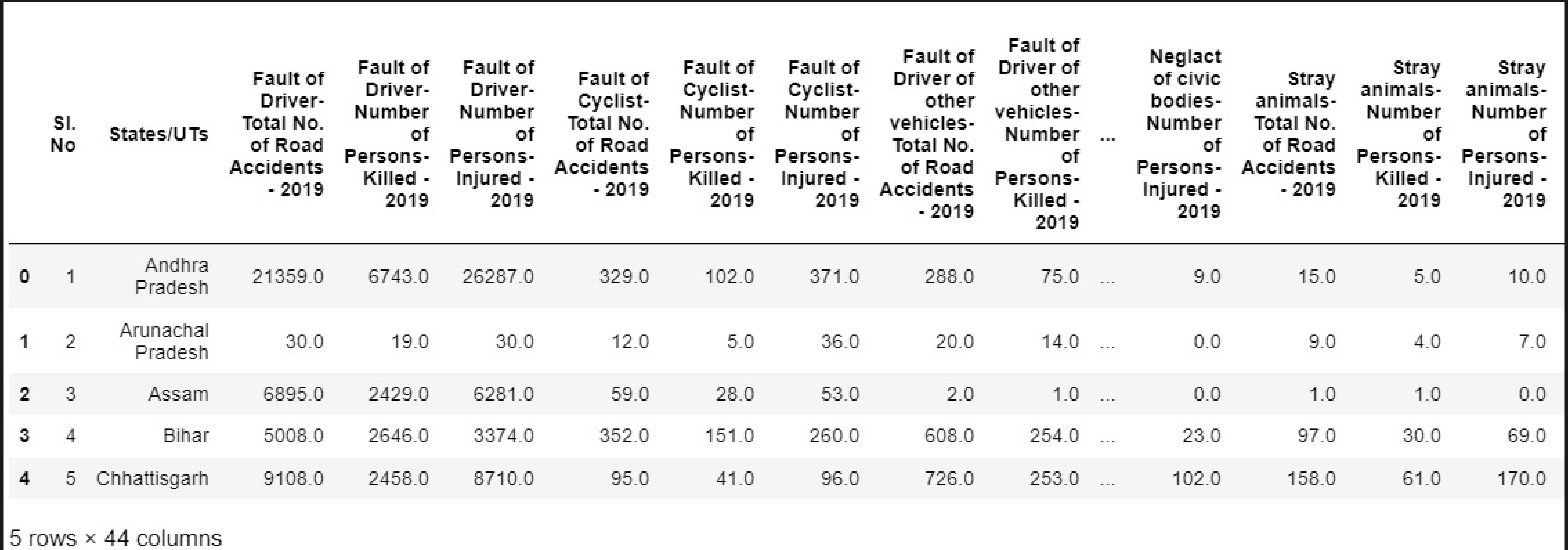
Time Based



	ACCIDENT_NO	ACCIDENT_DATE	ACCIDENT_TIME	ALCOHOLTIME	ACCIDENT_TYPE	DAY_OF_WEEK	HIT_RUN_FLA
0	T20170011699	2017-03-27 00:00:00	14.20.00	1	1	6.0	
1	T20170011706	2017-03-25 00:00:00	19.00.00	1	1	4.0	
2	T20170011709	2017-03-27 00:00:00	15.00.00	1	2	6.0	
3	T20170011710	2017-03-27 00:00:00	16.10.00	1	1	6.0	
4	T20170011711	2017-03-19 00:00:00	11.00.00	0	3	5.0	

RESULTS - TABLES

Fault Type

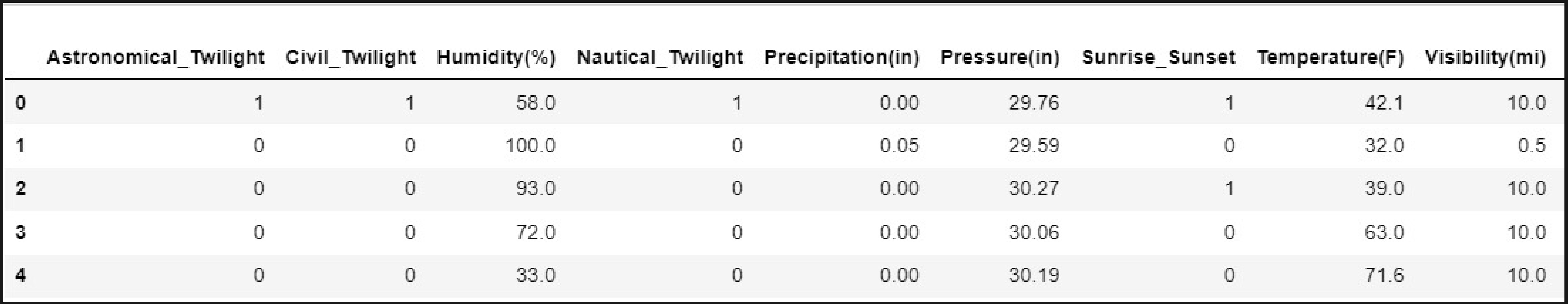


	Sl. No	States/UTs	Fault of Driver-Total No. of Road Accidents - 2019	Fault of Driver-Number of Persons-Killed - 2019	Fault of Driver-Number of Persons-Injured - 2019	Fault of Cyclist-Total No. of Road Accidents - 2019	Fault of Cyclist-Number of Persons-Killed - 2019	Fault of Cyclist-Number of Persons-Injured - 2019	Fault of Driver of other vehicles-Total No. of Road Accidents - 2019	Fault of Driver of other vehicles-Number of Persons-Killed - 2019	...	Neglact of civic bodies-Number of Persons-Injured - 2019	Stray animals-Total No. of Road Accidents - 2019	Stray animals-Number of Persons-Killed - 2019	Stray animals-Number of Persons-Injured - 2019
0	1	Andhra Pradesh	21359.0	6743.0	26287.0	329.0	102.0	371.0	288.0	75.0	...	9.0	15.0	5.0	10.0
1	2	Arunachal Pradesh	30.0	19.0	30.0	12.0	5.0	36.0	20.0	14.0	...	0.0	9.0	4.0	7.0
2	3	Assam	6895.0	2429.0	6281.0	59.0	28.0	53.0	2.0	1.0	...	0.0	1.0	1.0	0.0
3	4	Bihar	5008.0	2646.0	3374.0	352.0	151.0	260.0	608.0	254.0	...	23.0	97.0	30.0	69.0
4	5	Chhattisgarh	9108.0	2458.0	8710.0	95.0	41.0	96.0	726.0	253.0	...	102.0	158.0	61.0	170.0

5 rows × 44 columns

RESULTS - TABLES

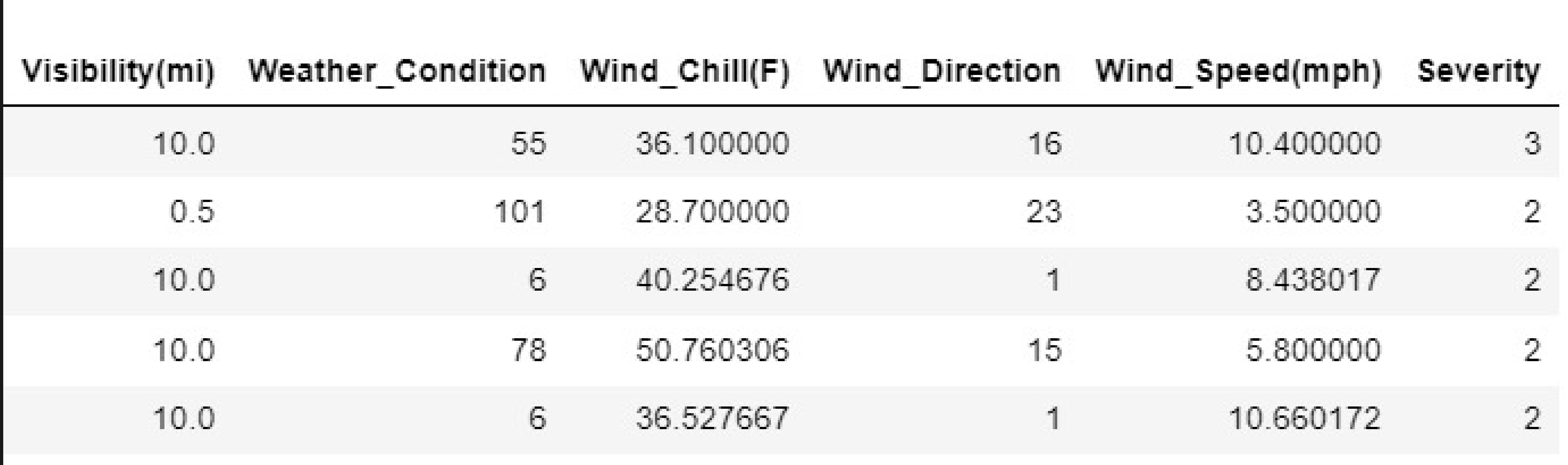
Weather Conditions and Predicted accident Severity



	Astronomical_Twilight	Civil_Twilight	Humidity(%)	Nautical_Twilight	Precipitation(in)	Pressure(in)	Sunrise_Sunset	Temperature(F)	Visibility(mi)
0	1	1	58.0	1	0.00	29.76	1	42.1	10.0
1	0	0	100.0	0	0.05	29.59	0	32.0	0.5
2	0	0	93.0	0	0.00	30.27	1	39.0	10.0
3	0	0	72.0	0	0.00	30.06	0	63.0	10.0
4	0	0	33.0	0	0.00	30.19	0	71.6	10.0

RESULTS - TABLES

Weather Conditions and Predicted accident Severity



Visibility(mi)	Weather_Condition	Wind_Chill(F)	Wind_Direction	Wind_Speed(mph)	Severity
10.0	55	36.100000	16	10.400000	3
0.5	101	28.700000	23	3.500000	2
10.0	6	40.254676	1	8.438017	2
10.0	78	50.760306	15	5.800000	2
10.0	6	36.527667	1	10.660172	2

Conclusion

SO, WE CAN CONCLUDE THAT THE RUSH HOURS ARE THE TIMES WHEN PEOPLE COME BACK FROM THE OFFICES AND THE KIDS, CHILDREN COME BACK FROM THE STUDENTS AND THE TRAFFIC ON THE ROAD IS MAXIMUM. MOST OF THE TIME THE HEAVY VEHICLES LIKE BUSES, TRUCKS AND THE MOST USEFUL VEHICLES IN DAILY LIFE LIKE 2 WHEELERS ARE THE MAIN CAUSES OF THE ROAD ACCIDENTS.

FROM OUR DATASET WE ALSO CONCLUDED THAT MOST OF THE TIME (85% TIMES IN 2018) IT'S THE DRIVER'S FAULT IN A ROAD ACCIDENT. ALSO, THE SEVERITY CLASSIFICATION DEEP LEARNING MODEL IS WORKING WELL WITH A GOOD ACCURACY. SO, BY OUR WORK WE CAN PREDICT WHETHER THERE WILL BE AN ACCIDENT OR NOT AND IF THE ACCIDENT OCCURS HOW SEVERE IT MAY BE. SO NECESSARY MEASURES SHOULD BE TAKEN TO PREVENT THESE ACCIDENTS.

Future Scope

THE FUTURE PLAN FOR THIS PROJECT IS TO CONNECT AN IOT DEVICE TO A DEEP LEARNING MODEL USING A RASPBERRY PI THAT TAKES WEATHER CONDITION INPUT FROM THE ENVIRONMENT AND FEEDS IT TO THE DEEP LEARNING MODEL. YOU CAN TAKE STEPS TO AVOID ACCIDENTS.

THEREFORE, WE DEVELOPED AN APP THAT CAN EFFICIENTLY PREDICT TRAFFIC ACCIDENTS BASED ON THE ABOVE FACTORS. THERE ARE MANY REASONS TO CHOOSE NEURAL NETWORKS, INCLUDING: ABILITY TO EXTRACT INFORMATION FROM INCOMPLETE AND NOISY DATA. GAIN EXPERIENCE AND KNOWLEDGE THROUGH SELF-EDUCATION AND ORGANIZATION. POSSIBILITY OF VERY FAST OPTIMIZATION. YOUR APTITUDE FOR PROBLEMS FOR WHICH ALGORITHMIC SOLUTIONS ARE DIFFICULT OR NON-EXISTENT.

REFERENCES

[1] ATHANASIOS THEOFILATOS AND GEORGE, YANNIS NATIONAL “A REVIEW OF THE EFFECT OF TRAFFIC
AND WEATHER CHARACTERISTICS ON ROAD SAFETY”, ACCIDENT ANALYSIS AND PREVENTION 72 (2014) 24425. (2014)

[2] K MESHRAM AND H.S. GOLIYA “ACCIDENT ANALYSIS ON NATIONAL HIGHWAY-3 BETWEEN INDORE TO DHAMNOD” INTERNATIONAL JOURNAL OF APPLICATION OR INNOVATION IN ENGINEERING & MANAGEMENT (IJAIEM) VOLUME2, ISSUE 7, JULY 2013.

[3] R.R. DINU, A. VEERARAGAVAN “RANDOM PARAMETER MODELS FOR ACCIDENT PREDICTION ON TWO LANE UNDIVIDED HIGHWAYS IN INDIA” , JOURNAL OF SAFETYRESEARCH 42(2011) 39-42, 2011

[4] S. NAGENDRA BABU, J. JEBAMALAR TAMILSELVI,” A DATA MINING FRAMEWORK TO ANALYZE ROAD ACCIDENT DATA USING MAP REDUCE METHODS CCMF AND TCAMP ALGORITHMS” , IJSSST, 2013

REFERENCES

[5] SERAFÍN MORAL-GARCÍA , JAVIER G. CASTELLANO , CARLOS J. MANTAS , ALFONSO MONTELLA AND JOAQUÍN ABELLÁN , “DECISION TREE ENSEMBLE METHOD FOR ANALYZING TRAFFIC ACCIDENTS OF NOVICE DRIVERS IN URBAN AREAS”, 2019

[6] GABRIELA V. ANGELES PEREZ, JOSE CASTILLEJOS LOPEZ, ARACELI L. REYES CABELLO, EMILIO BRAVO GRAJALES, ADRIANA PEREZ ESPINOSA, JOSE L. QUIROZ FABIAN, “ROAD TRAFFIC ACCIDENTS ANALYSIS IN MEXICO CITY THROUGH CROWDSOURCING DATA AND DATA MINING TECHNIQUES”, 2018

Thank you!