**3.0 DATA MANAGEMENT**

1. **Data Collection**

The data employed for this research work was collected from the Federal Road Safety Corps (FRSC), RS11.2, Ondo Sector. Each row of data contain information about every accident case recorded from the various unit commands in the state (RS11.21, RS11.22, RS11.23, RS11.24, and RS11.25) and the sector command: RS11.2 between 2016 and 2020. Each row of the Road Traffic Accidents (RTAs) table has the following attributes: ( 'accident\_id', 'unit\_command\_id', 'date', 'crash\_time', 'report\_time', 'arrival\_time', 'response\_time', 'route', 'location', 'vehicle\_no', 'vehicle\_type', 'vehicle\_category', 'vehicle\_make','vehicle\_model','fleet\_operator','name\_of\_driver','dl\_no','causes','number\_of\_injured\_male\_adult','number\_of\_injured\_female\_adult','number\_of\_injured\_male\_child','number\_of\_injured\_female\_child','total\_injured','number\_of\_killed\_male\_adult','number\_of\_killed\_female\_adult','no\_killed\_male\_child','number\_of\_killed\_female\_child','total\_killed','number\_of\_involved\_male\_adult','no\_involved\_female\_adult','no\_involved\_male\_child','no\_involved\_female\_child','total\_involved'. The collected data is not cleaned properly and contains a lot of missing fields.

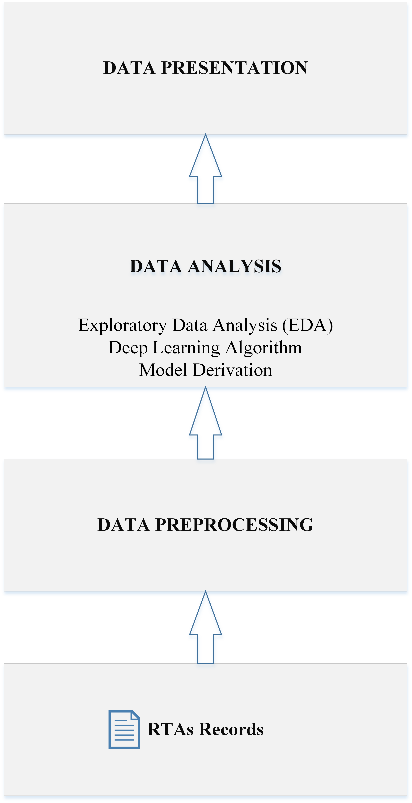


Figure 1: Data Management

1. **Data Preprocessing**

Data Preprocessing involves the preparation of the raw data collected for analysis. The data collected consists of 911 accident cases distributed for each year in this order: 2016 – 37 data points, 2017 – 222 data points, 2018 – 258 data points, 2019 – 215 data points, 2020 – 179 data points. The data collected for 2016 is small compared to others as it contains only information for two (2) months. Among the total set were some missing data, in order to mitigate for this, data points without information regarding the : crash time, report time and arrival time were removed completely. This reduced the data points to 874. Also, 4 data points have the ‘date’ feature missing, this is however a very sensitive feature. Rather than remove this row completely, the missing dates were backfilled using the date succeeding the missing row. In addition, the

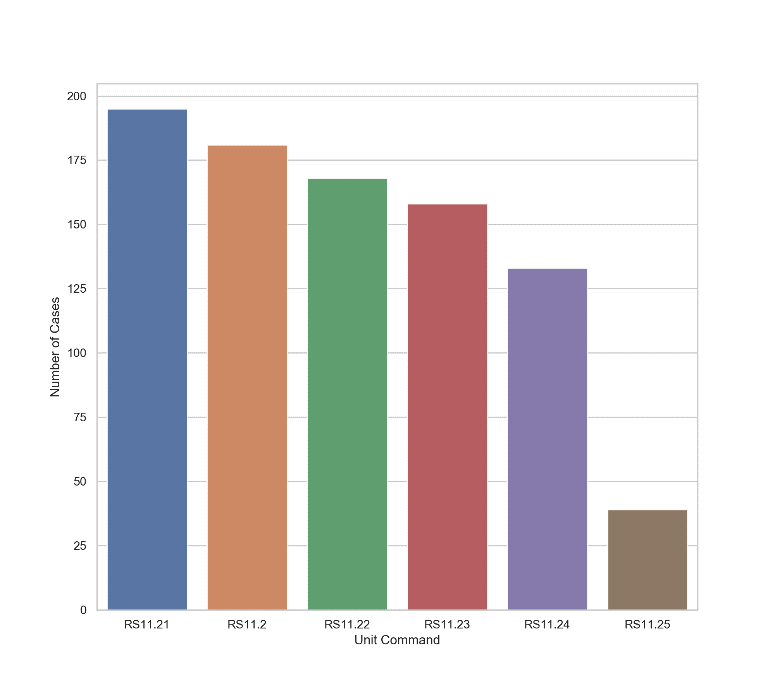
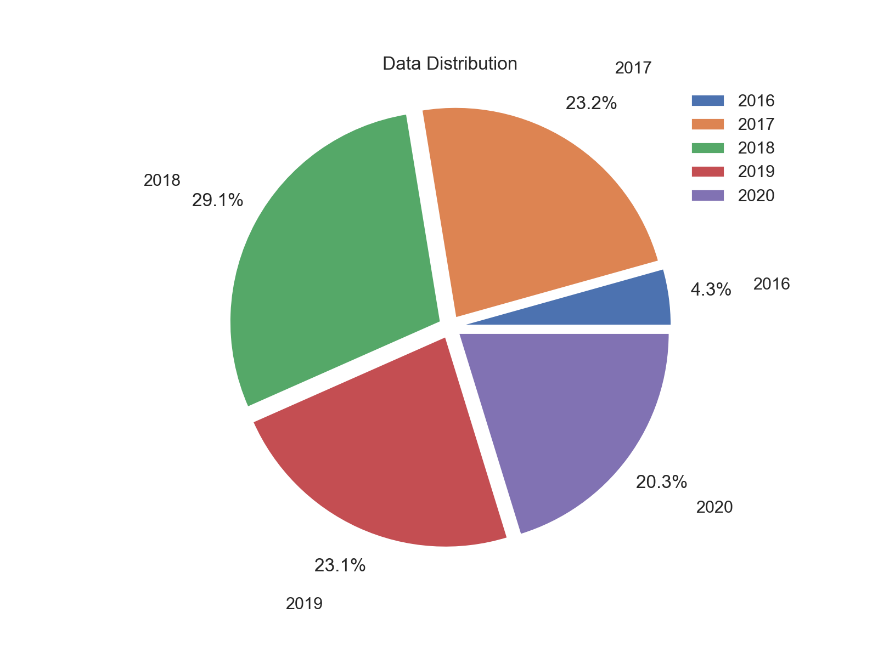


Figure 3.1: Data Distribution on a yearly basis Figure 3.2: Data Distribution based on Unit Command

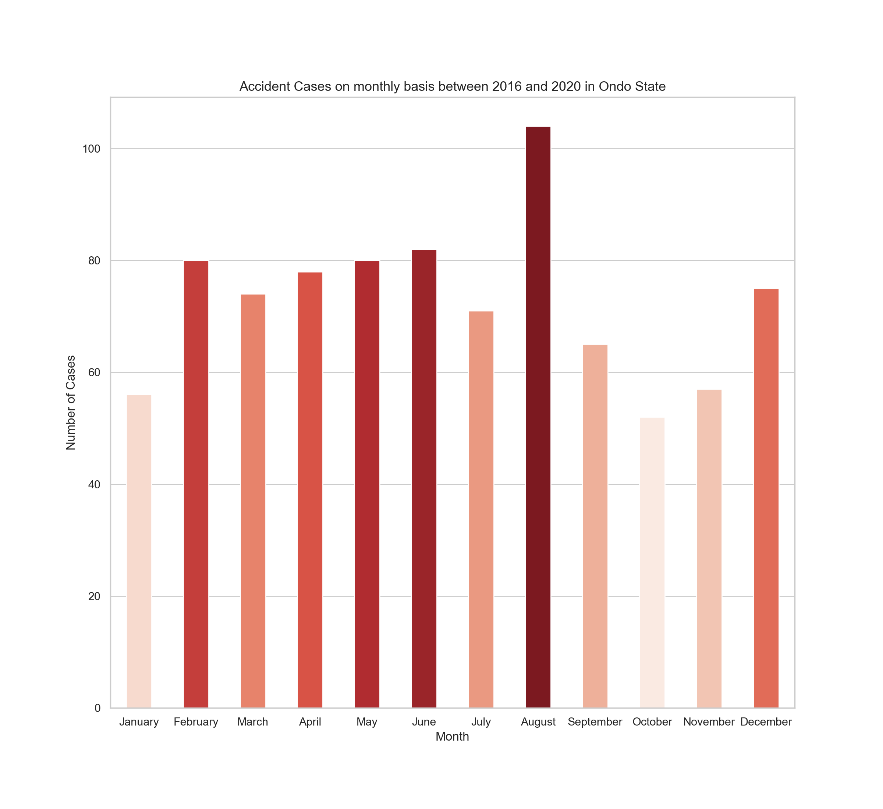
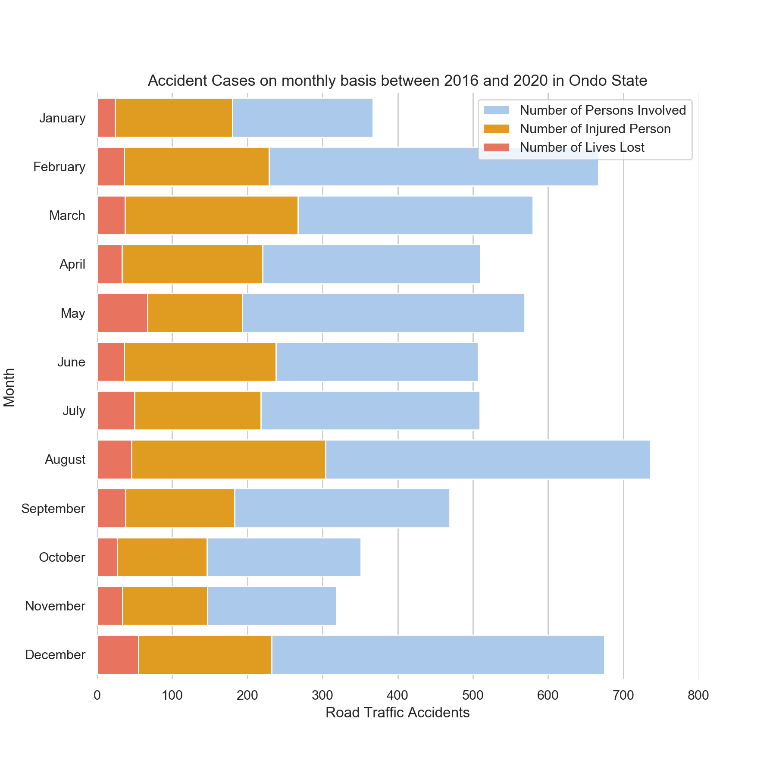
There was inconsistencies in the format of the 33 features found in the data, some features are expected to maintain a standard format such as the number of person involved (numeric) , the number of fatal cases(numeric) , vehicle type (string). Rather, the order was random. However, for the purpose of smooth analysis, each feature needs to maintain consistent representation across the dataset. Hence, the data was cleaned properly.

1. **Data Transformation**

Data transformation involves the transformation of the raw data into some set of new attributes. Highly significant features such as the total number of automobile involved, the month, the day of week were not included in the 33 features provided. To mitigate for this, some features were transformed from the 33 features intelligently.

**3. 1 Exploratory Data Analysis**

Exploratory Data Analysis involves the analysis of a dataset in the bid to summarize the characteristics using statistical tools and visualization techniques. Figure 3.2 shows the number of RTA total cases recorded each month between 2016 and 2020 in Ondo state. The visualization indicates August to be the month with the highest number of cases, followed by June, February, May and December. The figure 3.3 also indicates the number of persons involved and the number of lives lost on a monthly basis, the figure shows that several fatal cases were recorded most in May, December and August. December appearing on both figures can infer a lot of automobiles on the road as such period.

** **

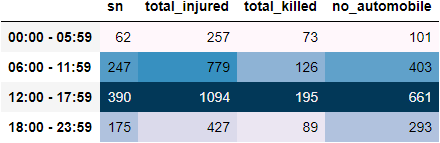
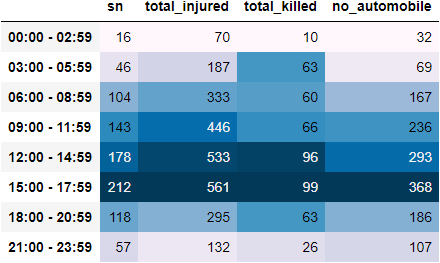
**Figure 3.2:** Accident Cases recorded each month **Figure 3.3**: Accident Cases showing the number of persons involved and lives lost

In relation to this work, the further analysis was carried out on the data acquired:

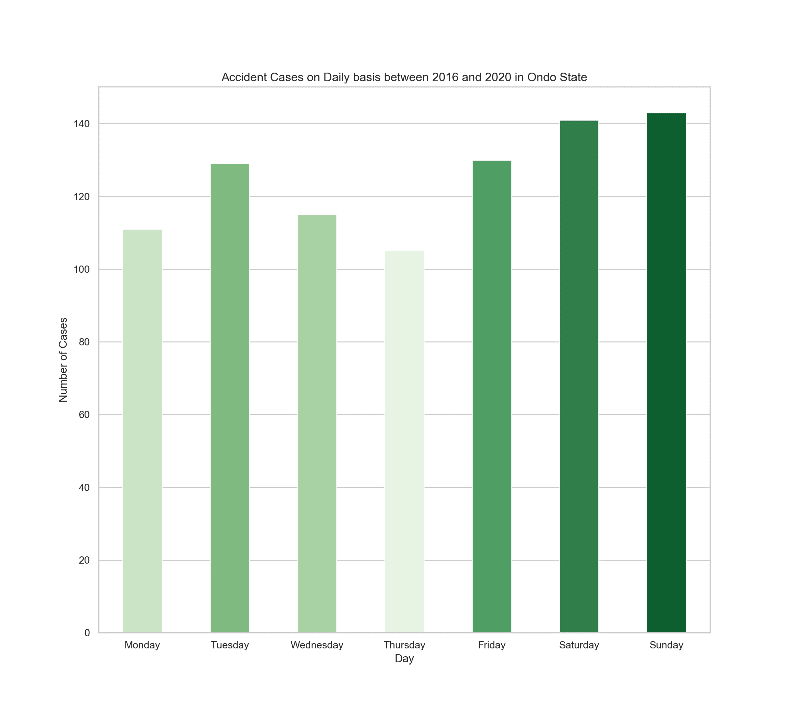
1. Analysis based on time
2. Analysis based on location

**3.1.1 Analysis based on time**

The figure 3.4 and figure 3.5 show the accident cases every 3-hour and 6-hour respectively. The ‘sn’ column represents that number of cases recorded. The tables shows that the most RTAs are recorded in daytime, between the 6th hour and 18th hour. Also, the table indicates that most fatal cases are recorded during the mentioned period. However, cases recorded at midnight are very minimal when compared with other periods of each day. In addition, from the data collected one can infer from figure 3.4 that RTAs occur mostly on weekends, with the least cases recorded on Thursdays and the most on Sundays.

****

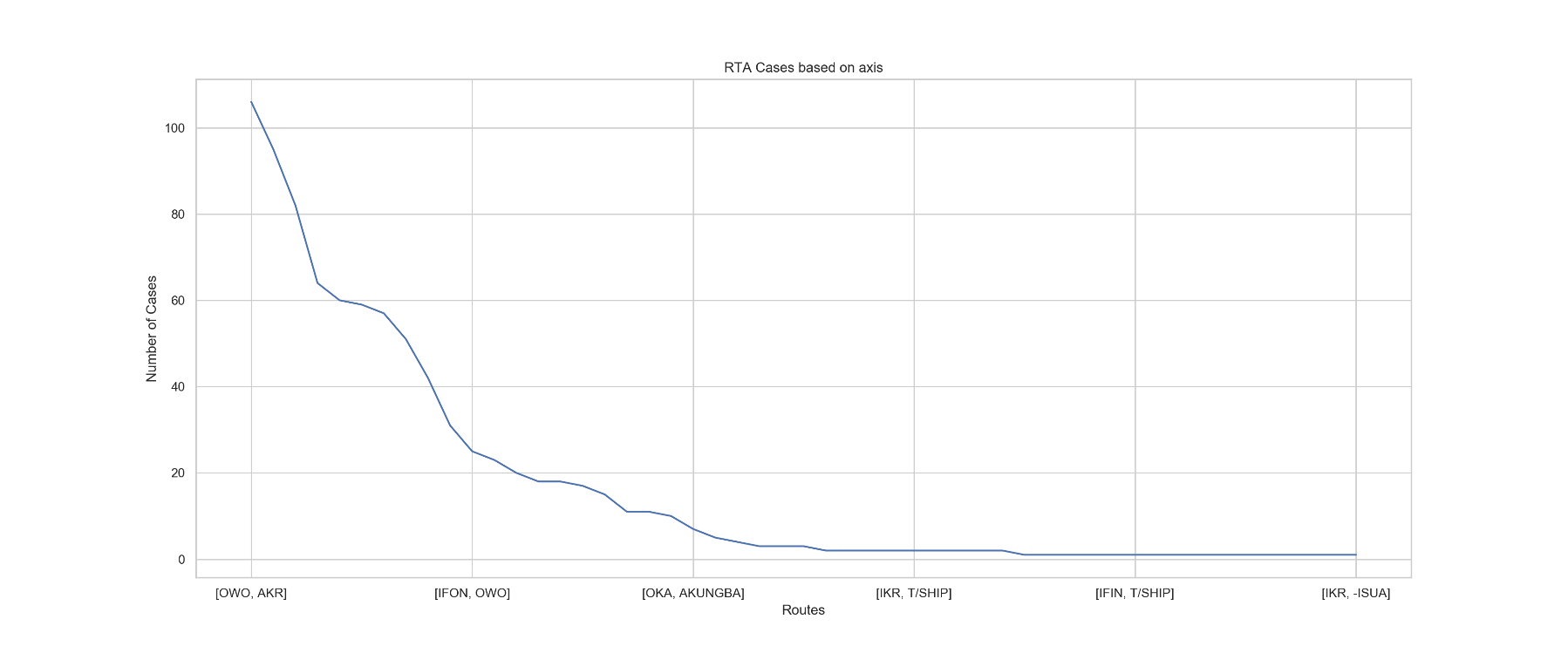
**Table 3.1:** Accident Cases report on 3-Hour basis **Table 3.2:** Accident Cases report on 6-Hour basis



**Figure 3.4:** Daily Accident Cases recorded on a daily basis

**3.1.2 Analysis based on Routes**

For location based analysis, data is sorted based on the route indicates that most RTAs cases recorded occurs along the Owo-Akure and Ifon-Owo axes compared with other axes in the state.



**3.4 Feature Engineering**

Only relevant features from the 33 attributes in the total data record will be useful in the training of the model. Hence, it is important to select the relevant features from the attributes and also extract new features from the relevant ones.

**3.4.1 Feature Selection**

However, before embarking on, some attributes were removed completely. The 'fleet\_operator','name\_of\_driver','dl\_no’ which represent the name of fleet operator, the name of driver, the vehicle registration number have over 98% of the total records missing. Hence, their removal from the total set.

**3.4.2 Feature Extraction**

In addition, the following features ‘report\_minus\_crash’, ‘arrival\_minus\_crash’, ‘arrival\_minus\_report’ were extracted from the ‘crash\_time’,’ report\_time’, ‘arrival\_time’, and ‘response\_time’ attributes via arithmetic operations.

1. ‘report\_minus\_crash’ - Number of seconds between the crash time and report time
2. ‘arrival\_minus\_crash’ - Number of seconds between the crash time and arrival time
3. ‘arrival\_minus\_report’ - Number of seconds between the arrival time and crash time

Having done this, the ‘report\_time’, ‘arrival\_time’, and ‘response\_time’ were dropped from the features. Also, the ‘crash\_time’ recorded in ‘datetime’ format was replaced with the hour the crash occurs. These were done in the bid to further make the attributes usable.

Machine Learning models are very sensitive to numeric values. Hence, it is imperative to scale continuous values in every column to a specific range (i.e. normalization). Furthermore, computation cannot texts which can be found in some significant attributes such as 'vehicle\_type' and 'vehicle\_category'. It is therefore important that features extracted intelligently from these attributes.

|  |  |  |  |
| --- | --- | --- | --- |