**Module title:** **Data mining and text analytics with application in SAS**

**Student number (id):** **6793260**

**Assessment title:** **Individual Assignment - Exploring Road Traffic Accident Data and Text Analytics Insights**

**Table of Contents**



**List of Tables**



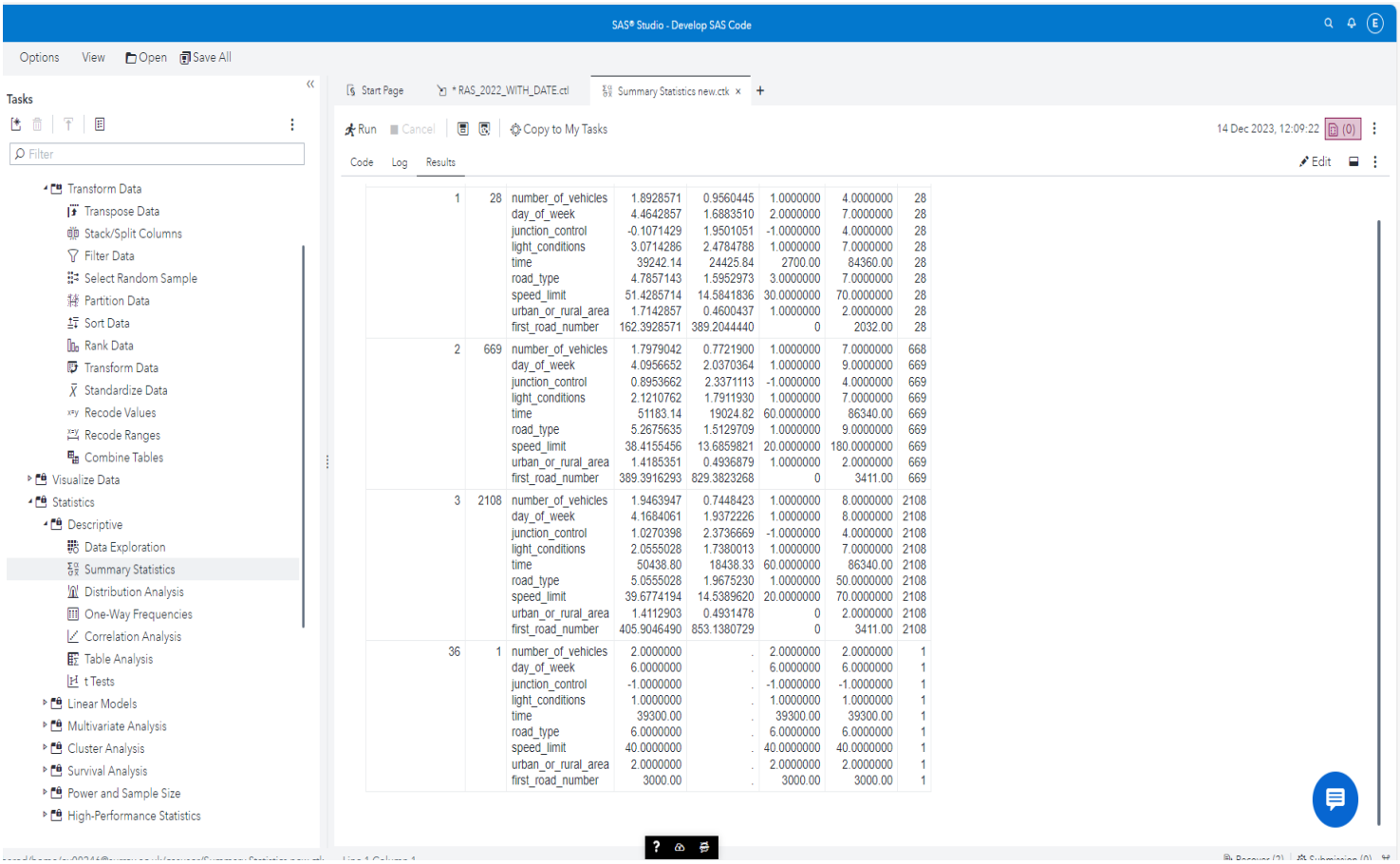
**List of Figures**

## **Task 1** – Data Exploration and Cleaning

### 1.1 Overview of the Dataset

The dataset comprises a range of variables related to road accidents. Each row represents an individual accident, and the dataset includes 35 variables, encompassing various aspects such as geographic coordinates, accident severity, number of vehicles and casualties, road conditions, and more. The dataset structure contains 2807 rows and 35 columns with the target variable being "accident_severity".

### 1.2 Summary Statistics




| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 28 | number_of_vehicles | 1.8928571 | 0.9560445 | 1.0000000 | 4.0000000 | 28 |
| | | day_of_week | 4.4642857 | 1.6883510 | 2.0000000 | 7.0000000 | 28 |
| | | junction_control | -0.1071429 | 1.9501051 | -1.0000000 | 4.0000000 | 28 |
| | | light_conditions | 3.0714286 | 2.4784788 | 1.0000000 | 7.0000000 | 28 |
| | | time | 39242.14 | 24425.84 | 2700.00 | 84360.00 | 28 |
| | | road_type | 4.7857143 | 1.5952973 | 3.0000000 | 7.0000000 | 28 |
| | | speed_limit | 51.4285714 | 14.5841836 | 30.0000000 | 70.0000000 | 28 |
| | | urban_or_rural_area | 1.7142857 | 0.4600437 | 1.0000000 | 2.0000000 | 28 |
| | | first_road_number | 162.3928571 | 389.2044440 | 0 | 2032.00 | 28 |
| 2 | 669 | number_of_vehicles | 1.7979042 | 0.7721900 | 1.0000000 | 7.0000000 | 668 |
| | | day_of_week | 4.0956652 | 2.0370364 | 1.0000000 | 9.0000000 | 669 |
| | | junction_control | 0.8953662 | 2.3371113 | -1.0000000 | 4.0000000 | 669 |
| | | light_conditions | 2.1210762 | 1.7911930 | 1.0000000 | 7.0000000 | 669 |
| | | time | 51183.14 | 19024.82 | 60.0000000 | 86340.00 | 669 |
| | | road_type | 5.2675635 | 1.5129709 | 1.0000000 | 9.0000000 | 669 |
| | | speed_limit | 38.4155456 | 13.6859821 | 20.0000000 | 180.0000000 | 669 |
| | | urban_or_rural_area | 1.4185351 | 0.4936879 | 1.0000000 | 2.0000000 | 669 |
| | | first_road_number | 389.3916293 | 829.3823268 | 0 | 3411.00 | 669 |
| 3 | 2108 | number_of_vehicles | 1.9463947 | 0.7448423 | 1.0000000 | 8.0000000 | 2108 |
| | | day_of_week | 4.1684061 | 1.9372226 | 1.0000000 | 8.0000000 | 2108 |
| | | junction_control | 1.0270398 | 2.3736669 | -1.0000000 | 4.0000000 | 2108 |
| | | light_conditions | 2.0555028 | 1.7380013 | 1.0000000 | 7.0000000 | 2108 |
| | | time | 50438.80 | 18438.33 | 60.0000000 | 86340.00 | 2108 |
| | | road_type | 5.0555028 | 1.9675230 | 1.0000000 | 50.0000000 | 2108 |
| | | speed_limit | 39.6774194 | 14.5389620 | 20.0000000 | 70.0000000 | 2108 |
| | | urban_or_rural_area | 1.4112903 | 0.4931478 | 0 | 2.0000000 | 2108 |
| | | first_road_number | 405.9046490 | 853.1380729 | 0 | 3411.00 | 2108 |
| 36 | 1 | number_of_vehicles | 2.0000000 | . | 2.0000000 | 2.0000000 | 1 |
| | | day_of_week | 6.0000000 | . | 6.0000000 | 6.0000000 | 1 |
| | | junction_control | -1.0000000 | . | -1.0000000 | -1.0000000 | 1 |
| | | light_conditions | 1.0000000 | . | 1.0000000 | 1.0000000 | 1 |
| | | time | 39300.00 | . | 39300.00 | 39300.00 | 1 |
| | | road_type | 6.0000000 | . | 6.0000000 | 6.0000000 | 1 |
| | | speed_limit | 40.0000000 | . | 40.0000000 | 40.0000000 | 1 |
| | | urban_or_rural_area | 2.0000000 | . | 2.0000000 | 2.0000000 | 1 |
| | | first_road_number | 3000.00 | . | 3000.00 | 3000.00 | 1 |

**Fig 1 – Summary of key Variables**

## 1.3 Data Visualization

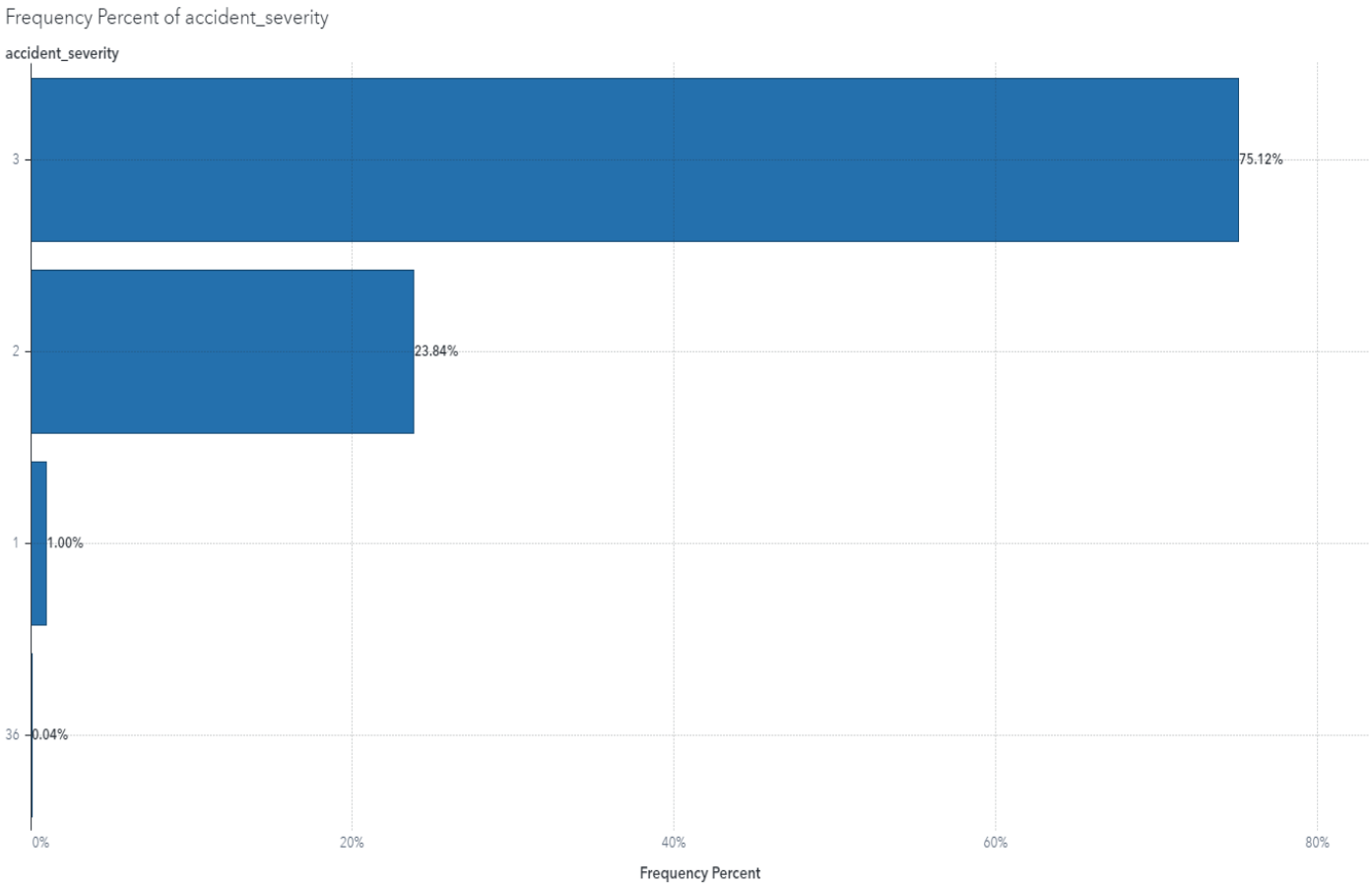



**Fig 2 - Percentage Distribution of Road Accidents by Severity**

The chart indicates a common trend seen in traffic accident data where most accidents result in minor injuries (75%), a smaller proportion result in serious injuries (24%), and a very small percentage are fatal (1%). This is consistent with general traffic accident statistics where minor accidents are far more common than those resulting in serious harm or fatalities.

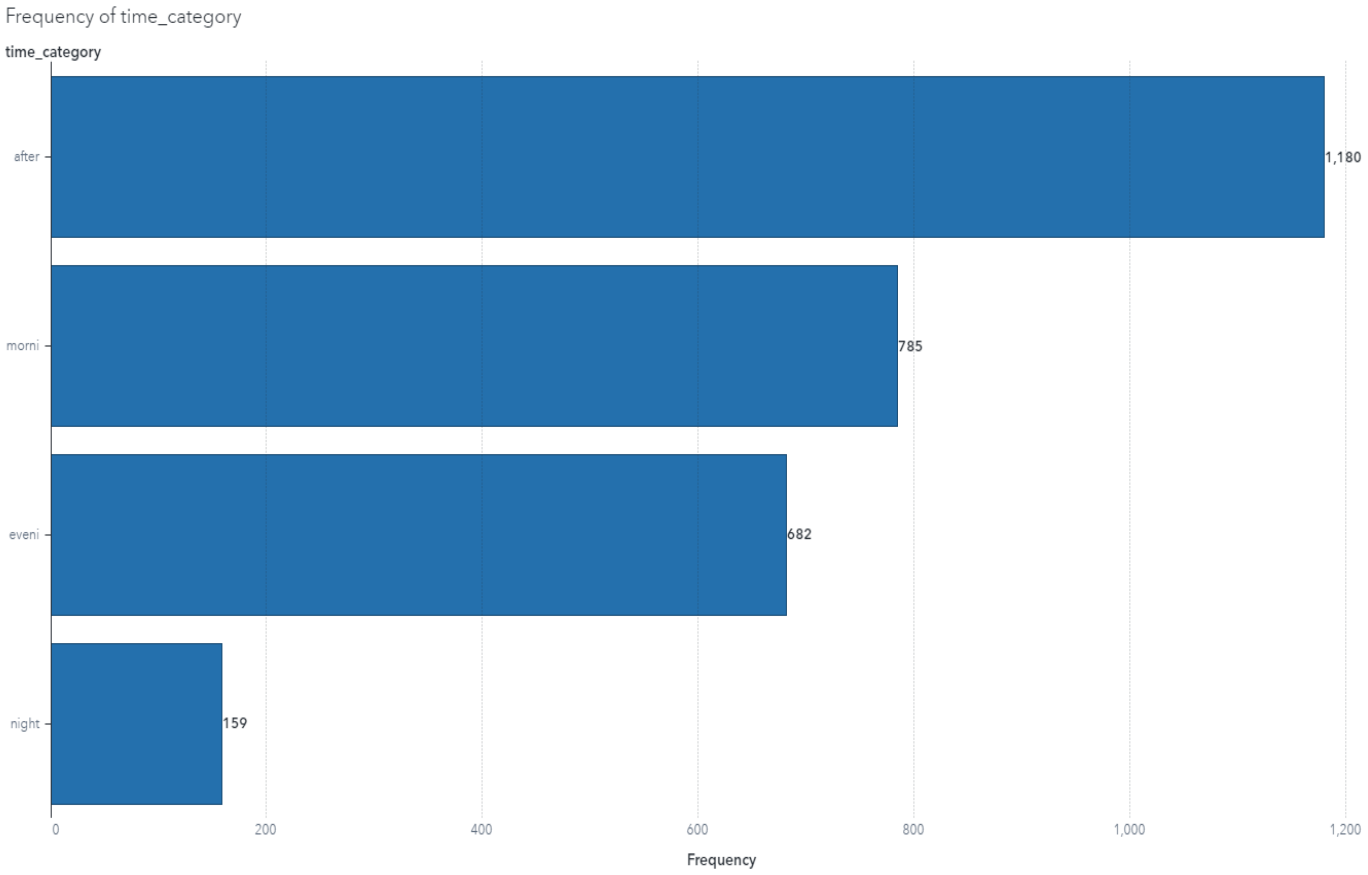



**Fig 3 – Accident Frequency by Time Category**

The majority of incidents occur throughout the afternoon, with mornings being the second most accident-prone time of day.

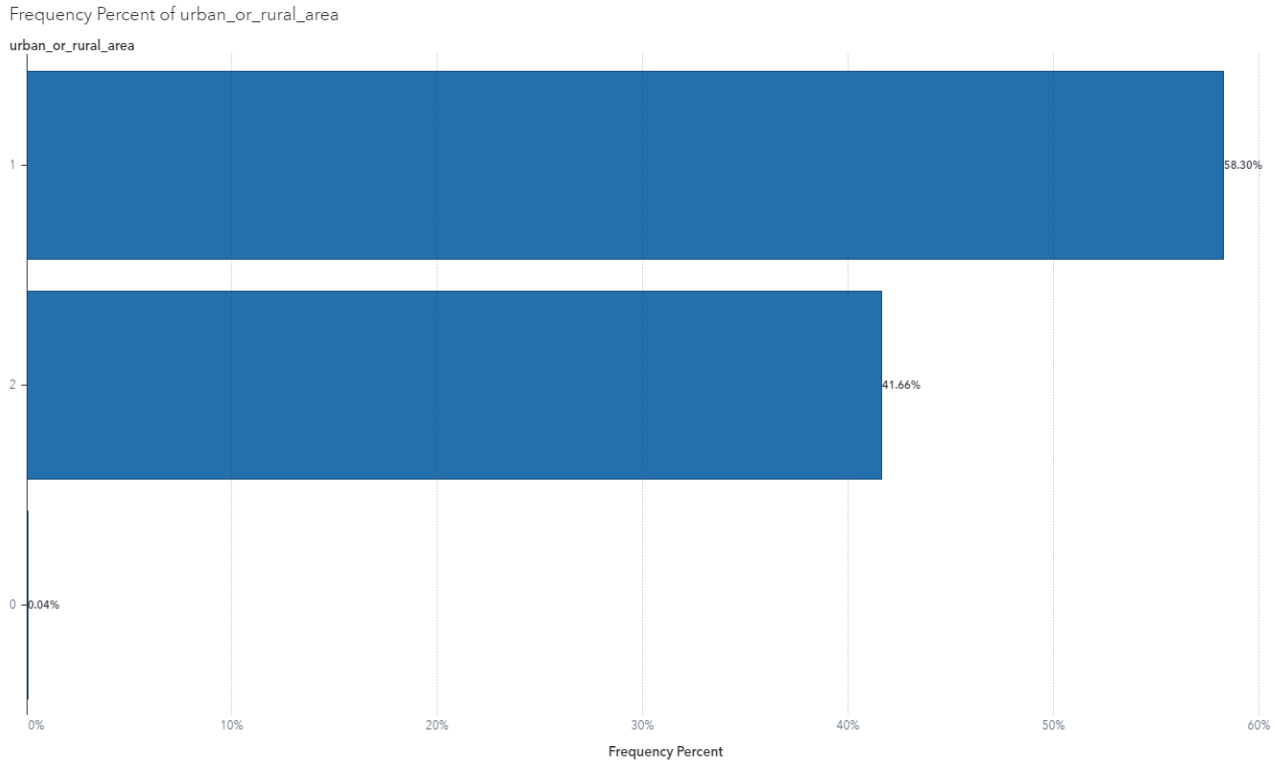



**Fig 4 – Percentage Distribution of Road Accidents in Urban vs. Rural Areas**

The chart indicates that a higher proportion of road accidents occur in urban areas compared to rural areas. This could be due to various factors such as higher traffic density, more complex road systems, and a greater number of vehicles in urban areas.

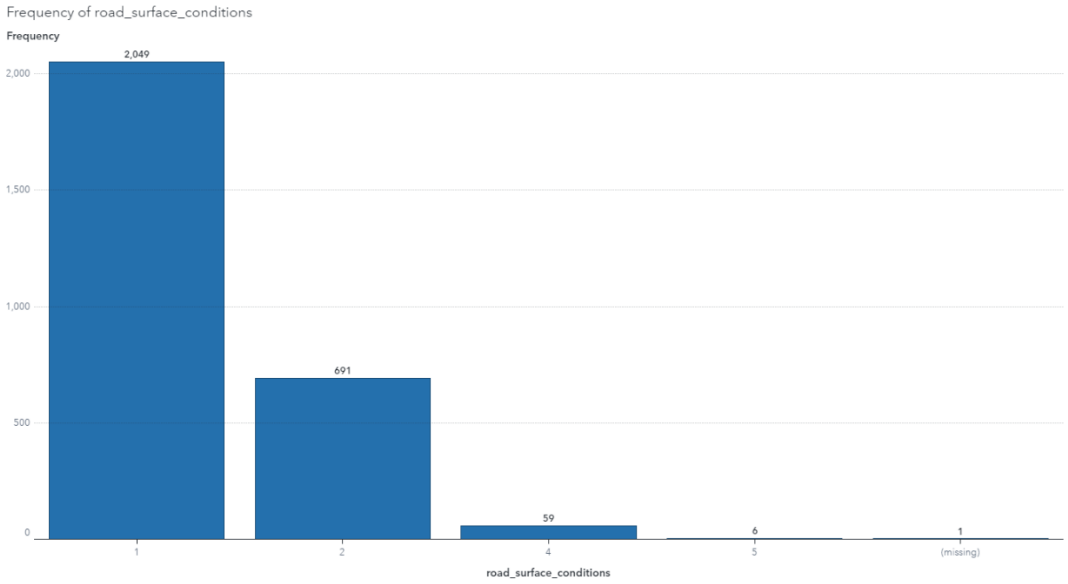



**Fig 5 – Accident Frequency by Road Surface Condition**

Accidents primarily occur in dry weather, with wet conditions being the second most common.

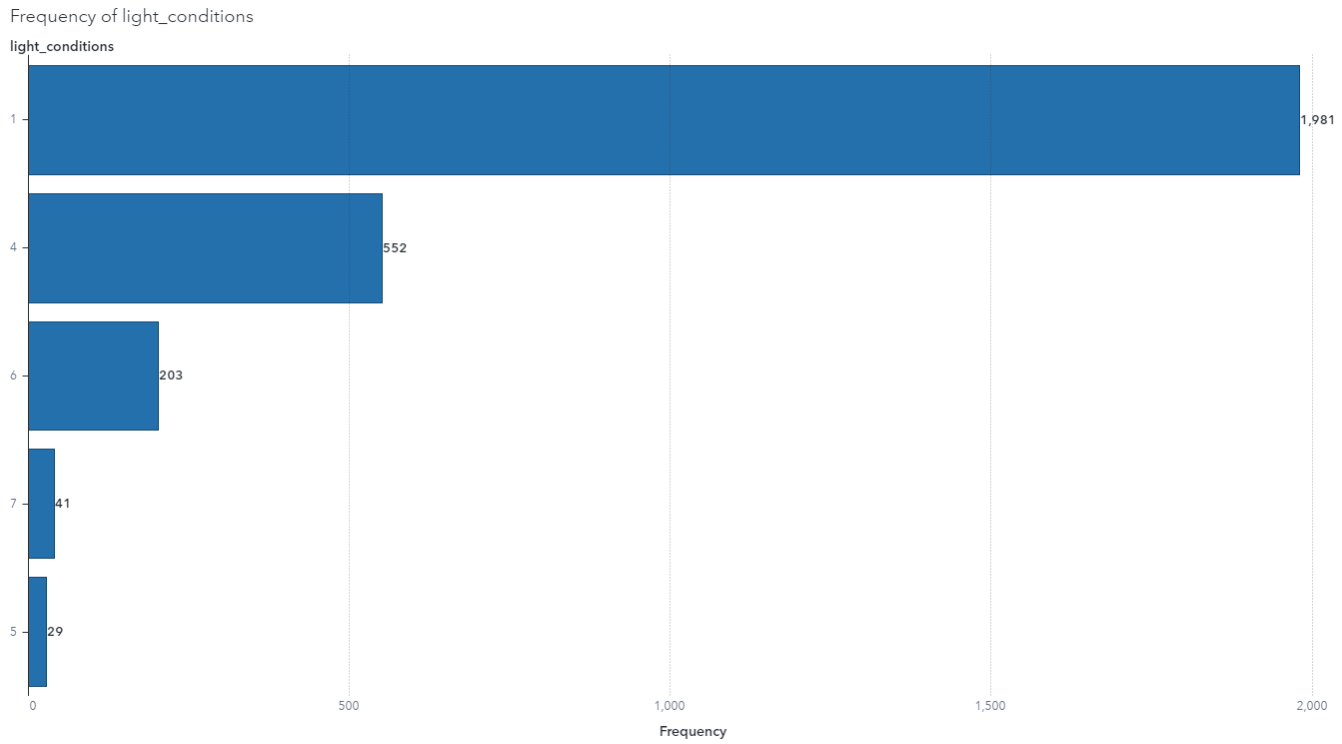



**Fig 6 – Accident Frequency by Light Conditions**

The majority of accidents tend to occur during daylight hours or in well-illuminated conditions at night.

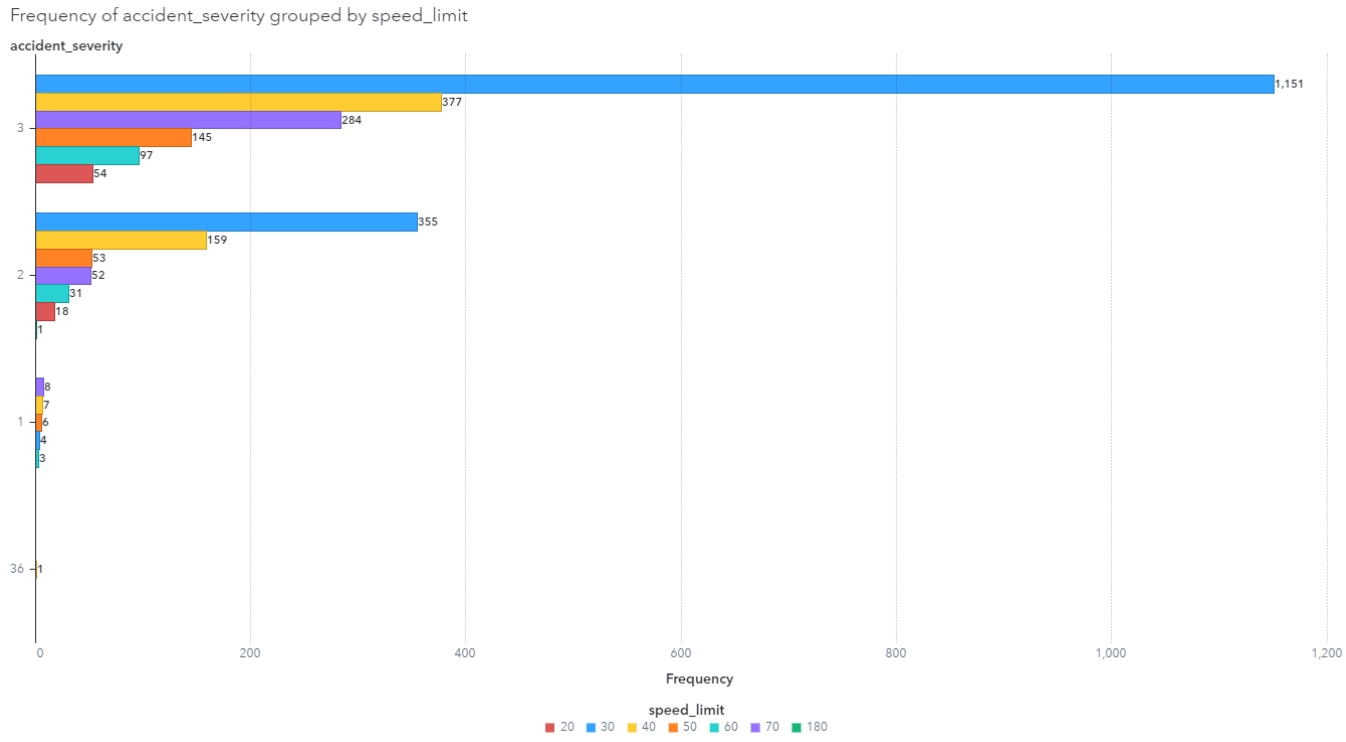



**Fig 7 – Frequency of Accident Severity Grouped by Speed Limit**

Highways that enforce a speed limit of 30mph experience the most catastrophic incidents.

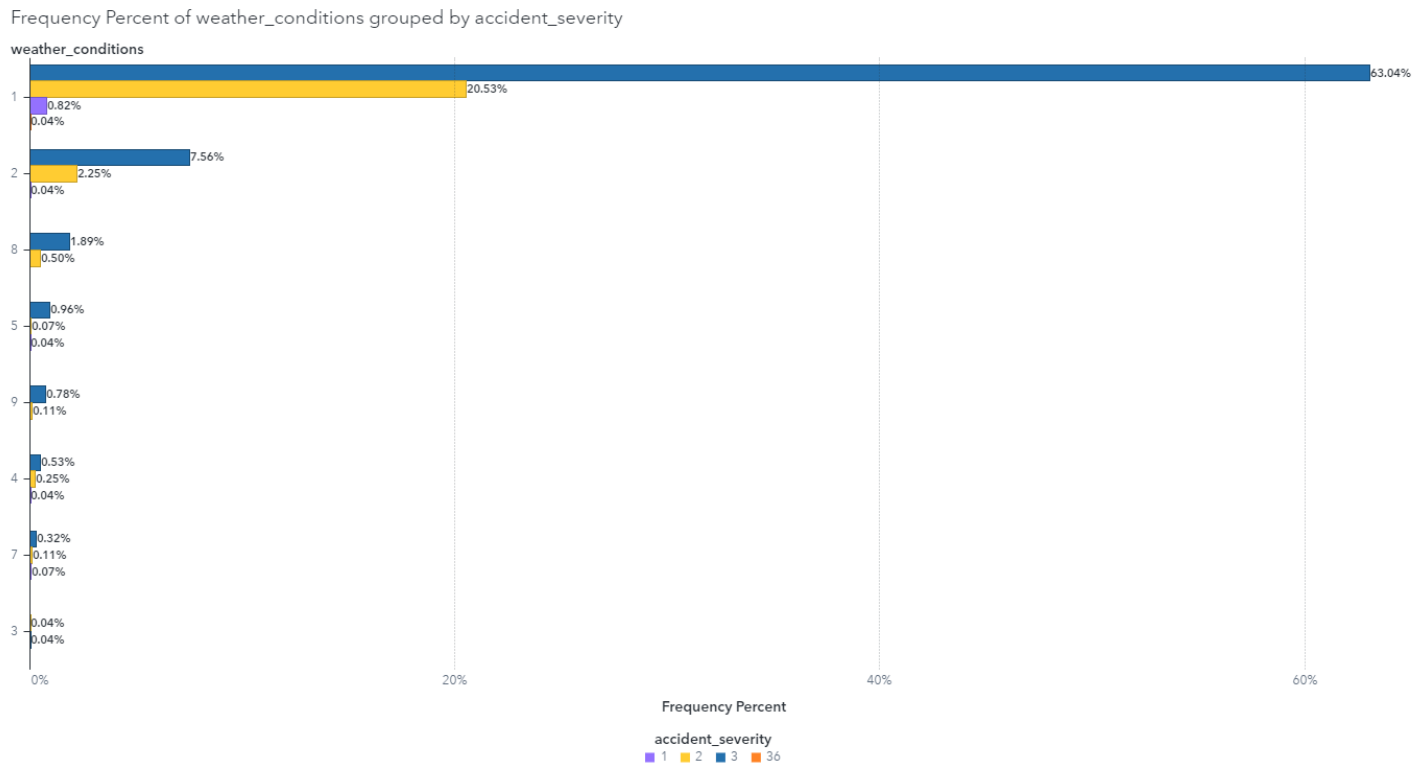



**Fig 8 – Frequency Percent of Accidents in different Weather Conditions Grouped by Severity**

A significant percentage of accidents, comprising both severe (20%) and minor (63%) incidents, occurred within fair weather conditions free of strong winds.

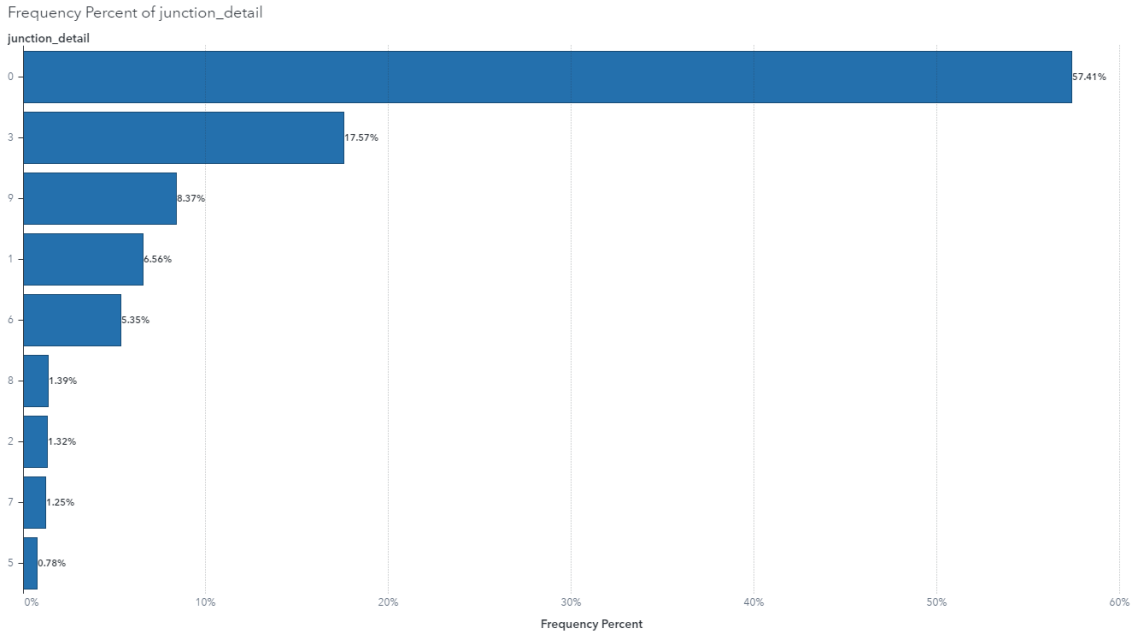



**Fig 9 – Frequency Percent of Accidents by Junction Type**

T-junctions account for 17% of all accidents.

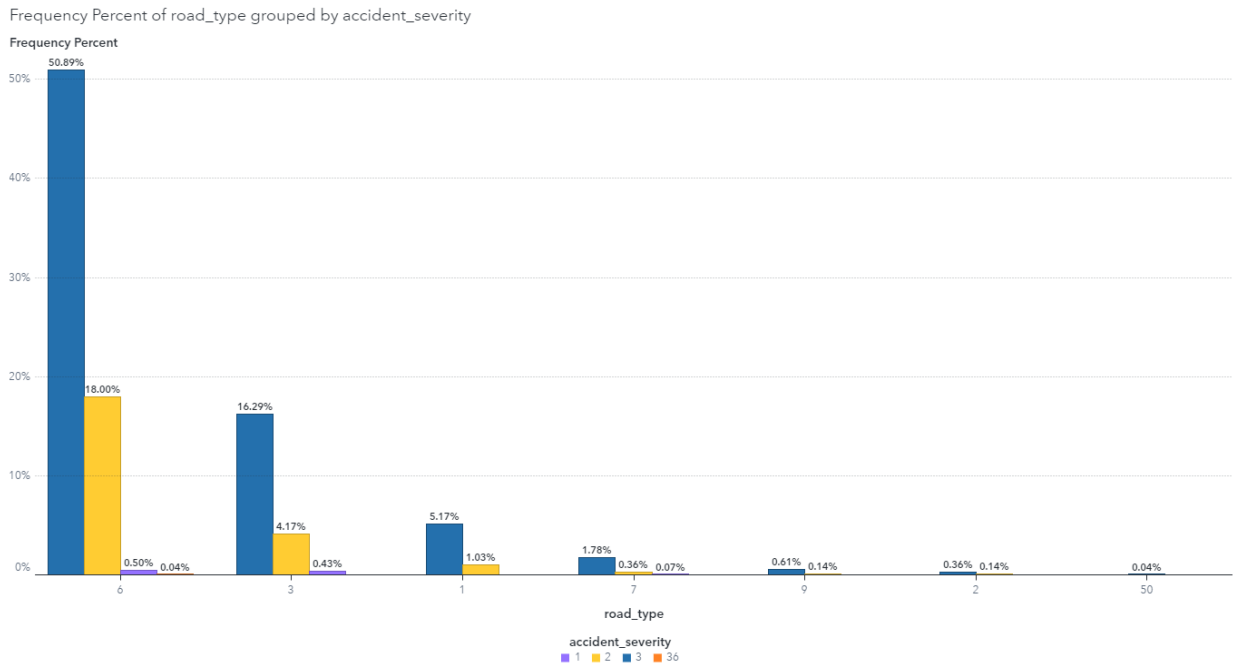



**Fig 10 – Frequency Percent of Accidents by Road Type Grouped by Severity**

Single carriageways are the primary site for serious accidents, with dual carriageways ranking second in frequency.

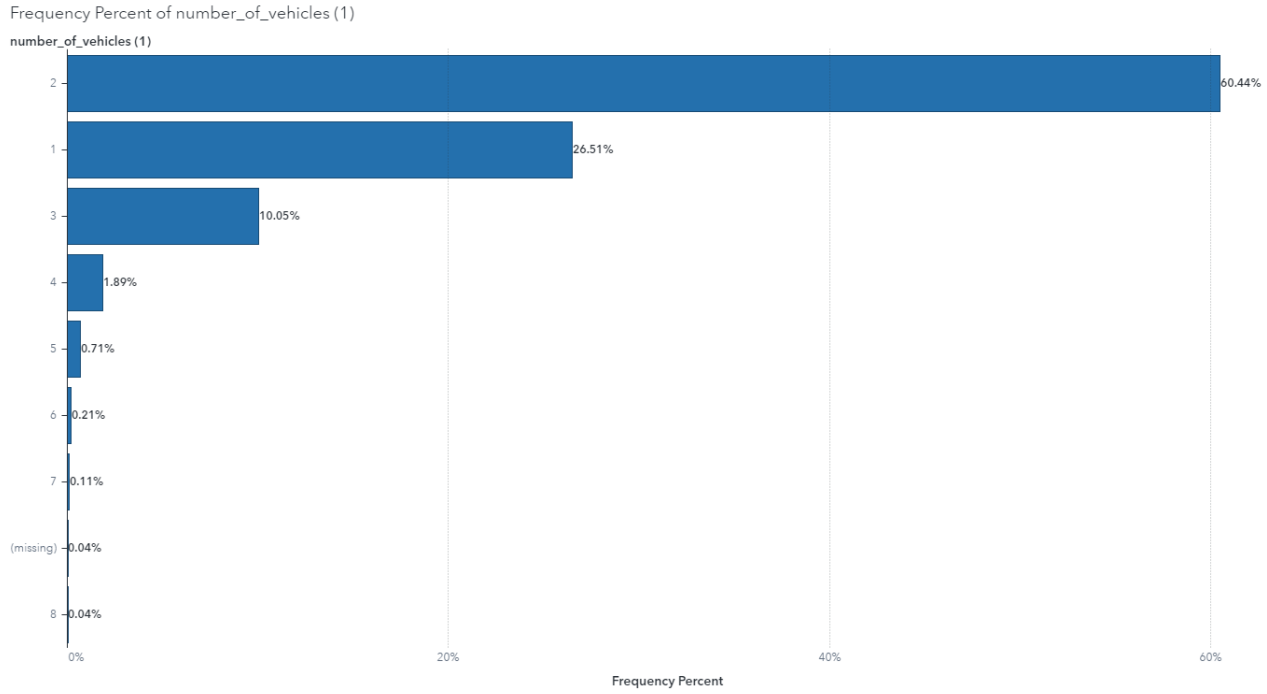



**Fig 11 – Distribution of Road Accidents by Number of Vehicles Involved**

About sixty percent of accidents are identified by the involvement of two vehicles, whereas 26% of accidents involve only a single vehicle

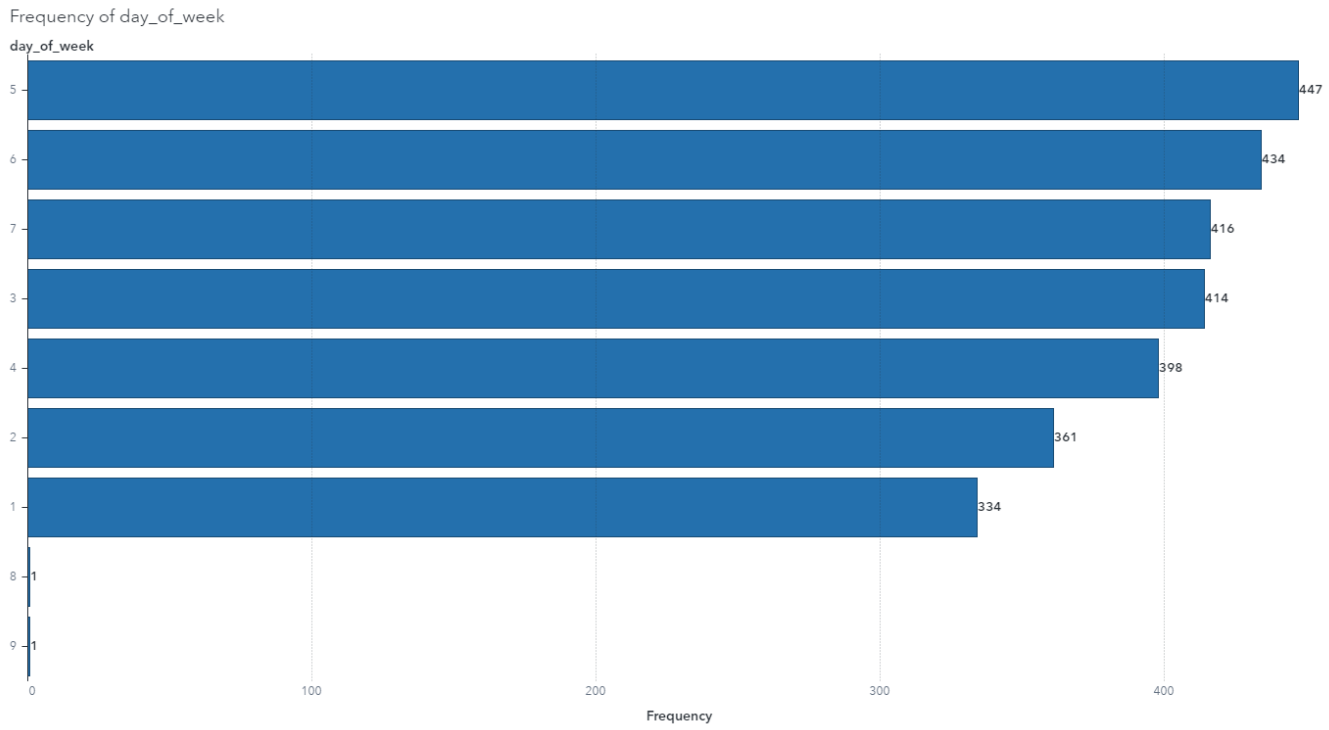



**Fig 12 – Accident Frequency by Weekday**

The days of the week with the highest frequency of accidents are Thursdays and Fridays.

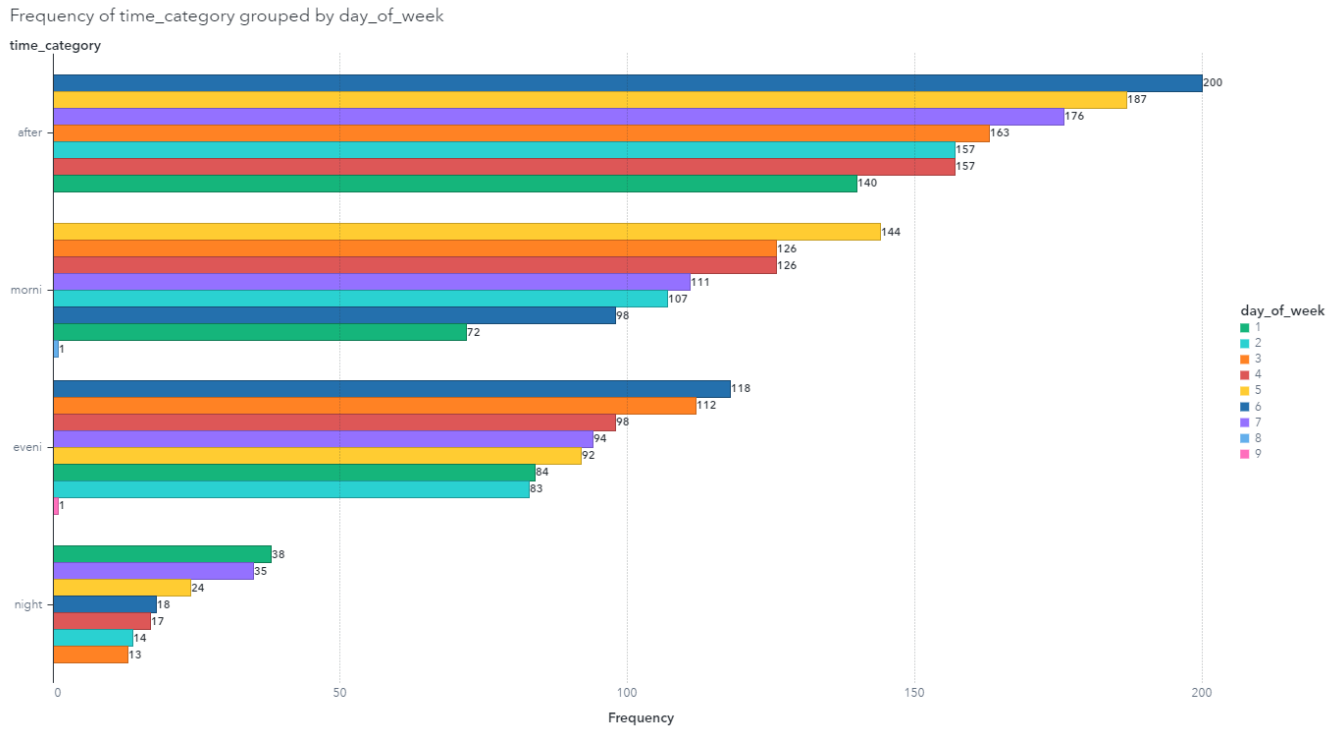



**Fig 13 – Accident Frequency by Time Category Grouped by Day of Week**

Afternoon accidents are more likely to occur on Fridays, whereas morning accidents are more common on Thursdays, and night accidents are more frequent on Sundays.

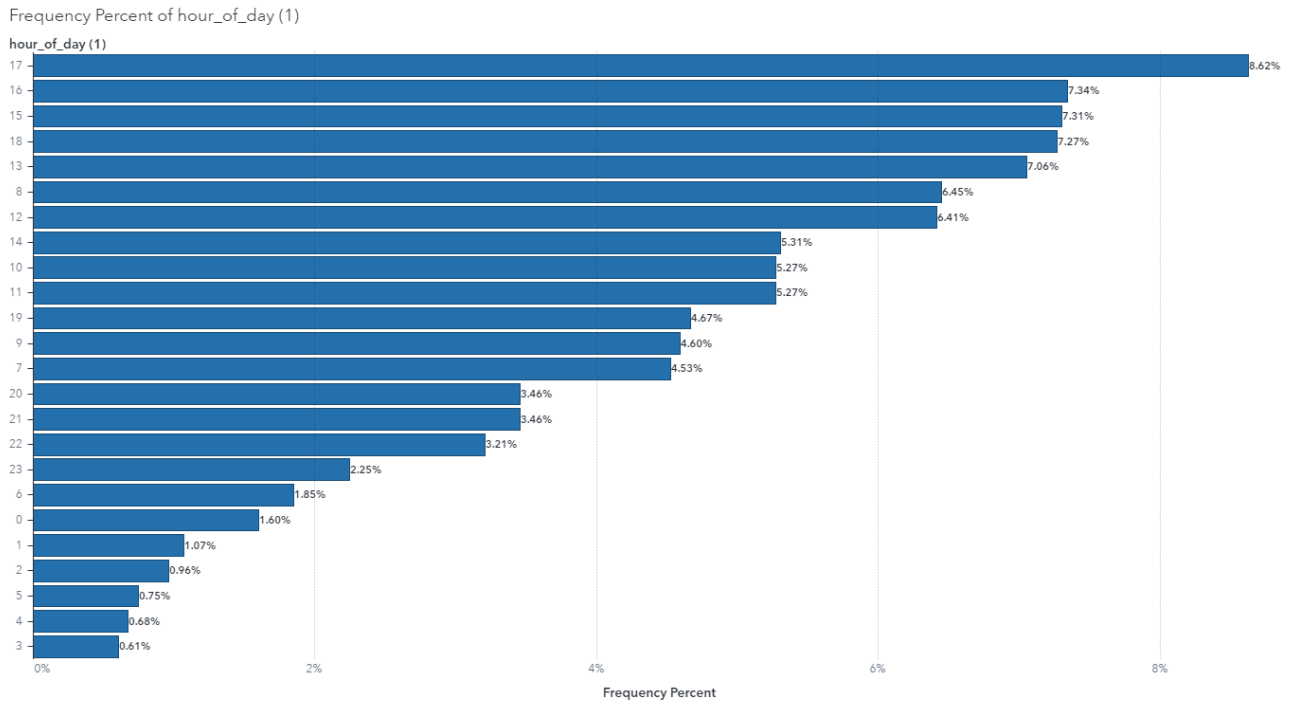



**Fig 14 – Frequency Percent of Accidents by Hour of Day**

The hour with the highest number of accidents, accounting for 8.62% of the total is determined to be 5 pm.

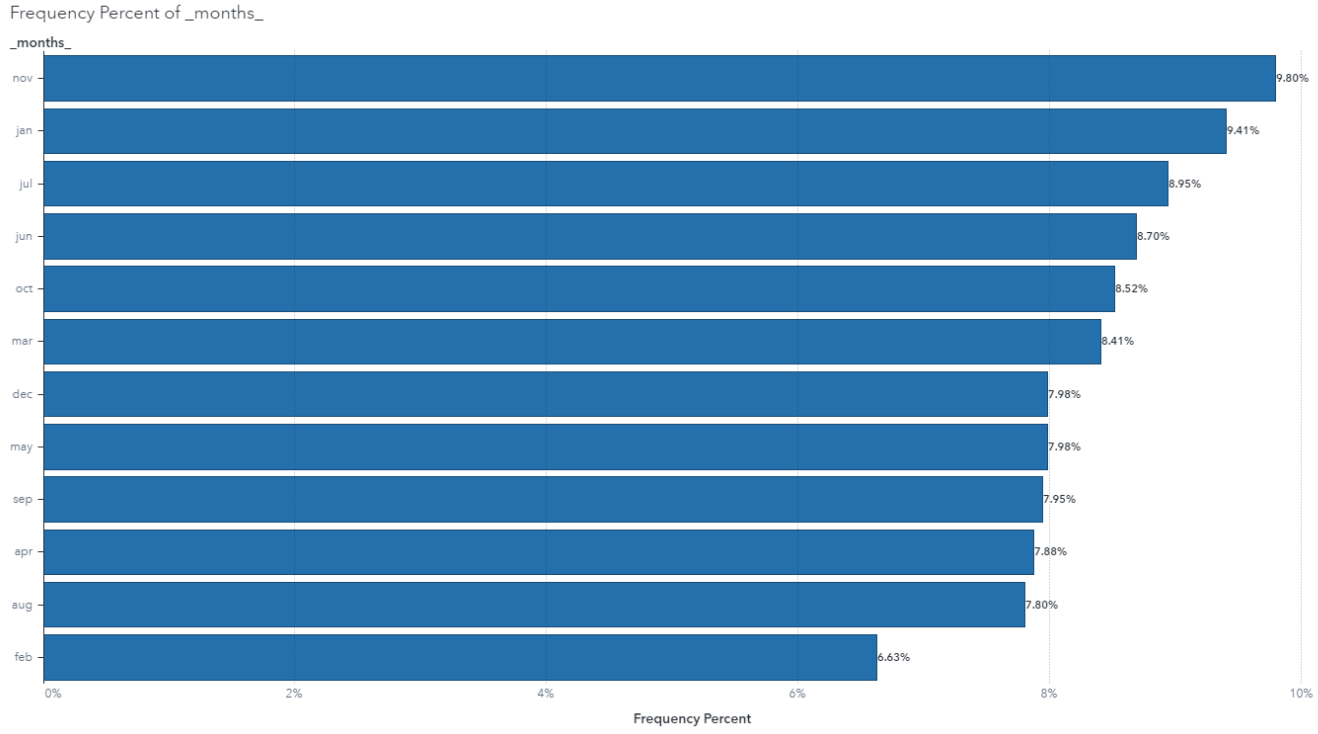



**Fig 15 – Frequency Percent of Accidents by Month**

The accident rates in November and January are the highest, representing 9.8% and 9.4% respectively.

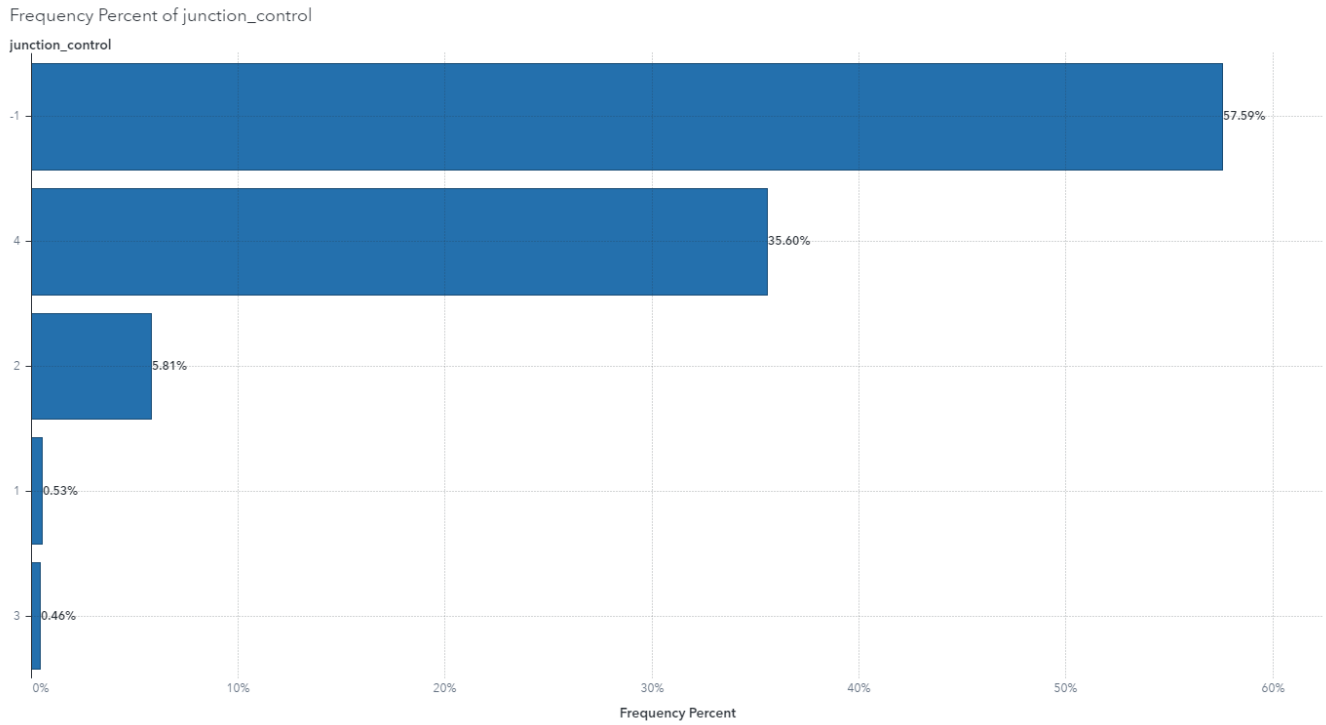



**Fig 16 – Frequency Percent of Accidents by Junction Control**

Unregulated intersections are involved in 35% of accidents.

### 1.4 Data Quality Issues

There is an anomalous data point in the speed_limit column, marked as 180, which is unusually high and not consistent with typical public highway speed limits. In the accident severity column, there is a value of 36, which is inconsistent with the established categories of 1 for fatal, 2 for serious, and 3 for slight, indicating a likely input error. Additionally, there are single incorrect entries in the "police_force" and "road_type" columns, as well as two inaccuracies in the "day_of_week" column. Furthermore, the "road_surface_conditions" and "number_of_vehicles" columns each have one missing entry.

### 1.5 Strategies Implemented for Data Cleansing

1. Outliers – Rows with outliers were removed.
2. Erroneous data – Rows with incorrect data entries were deleted.
3. Missing entries – Mode values were used to fill in missing data points.
4. Long column names – Columns with names exceeding 32 characters were renamed for conciseness.
5. Date adjustment – The date column was transformed to display only the month.
6. Time refinement – The time column was split into two new columns: one indicating the time category and the other the hour of the day.

### 1.6 Data Imbalance and Preparation for Modeling

The dataset exhibited a pronounced imbalance, with a majority of the records (75%) being slight accidents, compared to a smaller proportion of fatal (1%) and serious accidents (24%). To create a balanced dataset for effective modeling, appropriate data-balancing techniques were applied.

## **Task 2** – Predicting Accident Severity

### 2.1 Scenario

In the city one morning, under clear daylight, a serious accident happened at an uncontrolled junction where a slip road merged. Two vehicles were involved: one was crossing the junction while the other was entering from the slip road. The driver on the slip road, perhaps misjudging the timing or speed, entered the junction simultaneously with the other vehicle. This resulted in a collision at the uncontrolled junction, despite the visibility provided by daylight. The impact was severe. One of the drivers sustained serious injuries and required immediate medical attention. Emergency services, including police and an ambulance, responded quickly. The injured driver was taken to the hospital for treatment.

Variables for predicting accident severity

1. junction_control
2. light_conditions
3. number_of_vehicles
4. _months_
5. speed_limit
6. time_category
7. urban_or_rural_area
8. day_of_week
9. road_type
10. first_road_number

### 2.2 Predictive Models

The following predictive models are used:

1) Neural Networks
2) Decision Trees
3) Logistic Regression

### 2.2.1 Neural Networks

**Table 1 – Neural Network Results**

| | |
|---|---|
| KS (youden) | 0.7047 |
| Average square error (ASE) | 0.0749 |
| AUC | 0.9057 |
| Accuracy | 0.8556 |
| Cumulative Lift | 2.5984 |

A KS score of 0.7047 indicates that the model exhibits a high level of discrimination between the three classes. An ASE of 0.0749 signifies that the model's predictions have, on average, an error squared value of 0.0749, indicating relatively minor errors. Achieving an AUC value of 0.9057 indicates a high level of performance for the model, implying that it possesses a strong ability to differentiate between the three classes with great effectiveness. An accuracy score of 85.56% shows that the model makes valid predictions for the result in 85.56% of cases. A cumulative lift of 2.5984 indicates that the model is successfully distinguishing situations that have a higher probability of achieving the intended outcome, in contrast to a random selection.

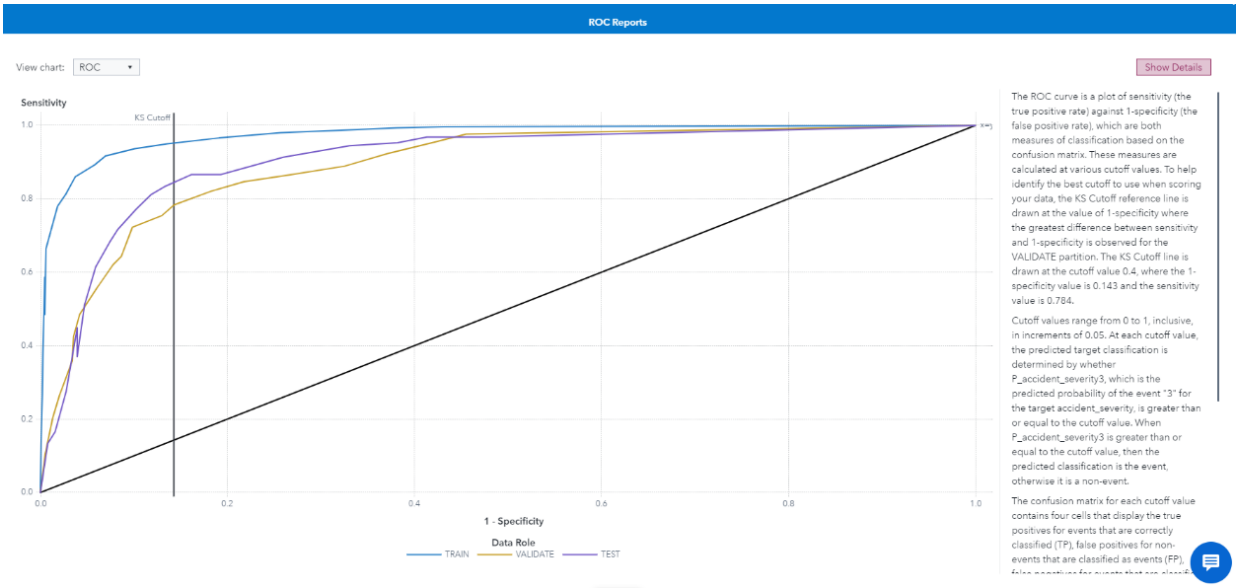



**Fig 17 – ROC - Neural Networks**

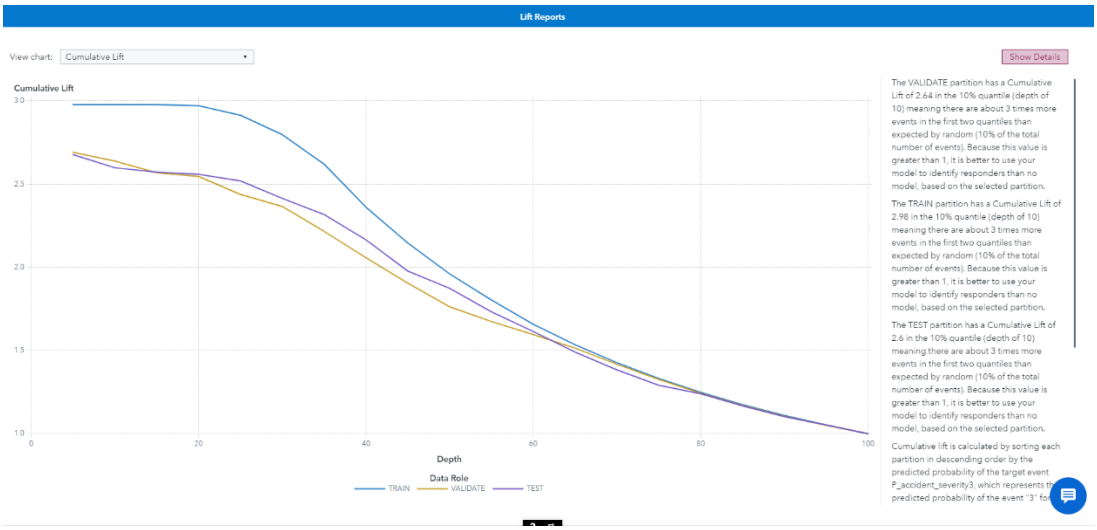



**Fig 18 – Cumulative Lift - Neural Networks**

### 2.2.2 Decision Trees

**Table 2 – Decision Tree Results**

| | |
|---|---|
| KS (youden) | 0.3976 |
| Average square error (ASE) | 0.1527 |
| AUC | 0.7316 |
| Accuracy | 0.7585 |
| Cumulative Lift | 2.7756 |

A KS value of 0.3976 indicates that the model possesses a moderate capability to distinguish between the classes. An ASE of 0.1527 signifies that, on average, the model's predictions deviate from the actual data by a squared difference of 0.1527. The AUC value of 0.7316 indicates that the model possesses a strong ability to make accurate predictions. An accuracy of 75.85% suggests the model accurately predicts the outcome for about 75.85% of the instances.

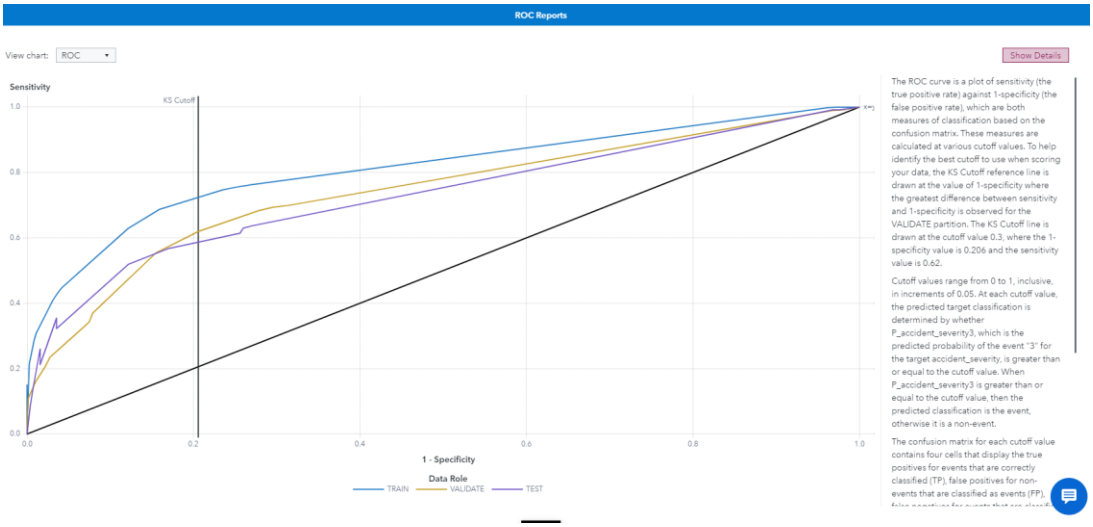



**Fig 19 – ROC - Decision Trees**

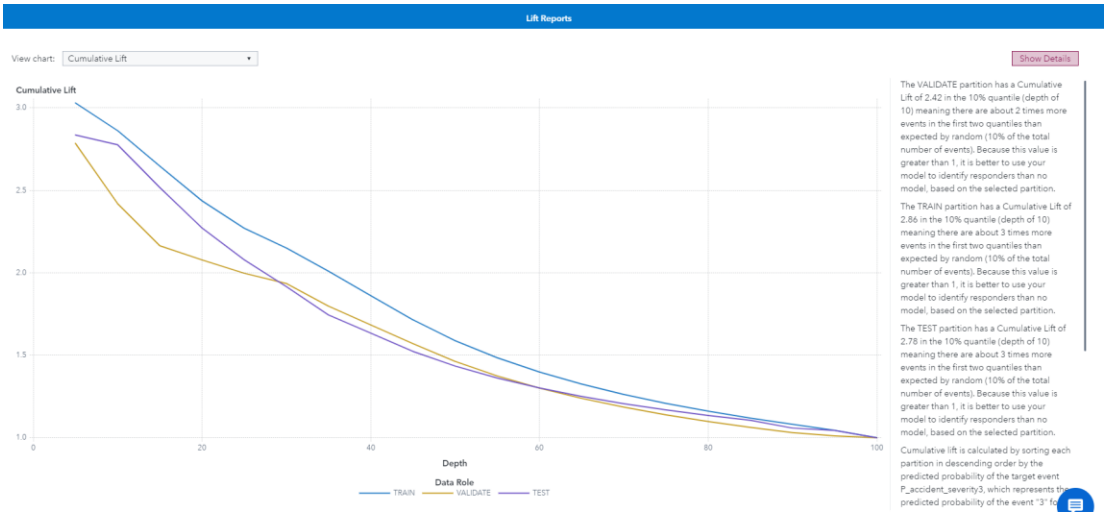



**Fig 20 – Cumulative Lift - Decision Trees**

### 2.2.3 Logistic Regression

**Table 3 – Logistic Regression Results**

| | |
|---|---|
| KS (youden) | 0.1929 |
| Average square error (ASE) | 0.1976 |
| AUC | 0.6136 |
| Accuracy | 0.6667 |
| Cumulative Lift | 1.4488 |

The KS value of 0.1929 indicates a somewhat weak discriminatory ability of the model. A value of 0.1976 for the ASE suggests that the model's predictions deviate somewhat significantly from the true values, on average. A value of 0.6136 for the AUC suggests a moderate level of predictive capability.

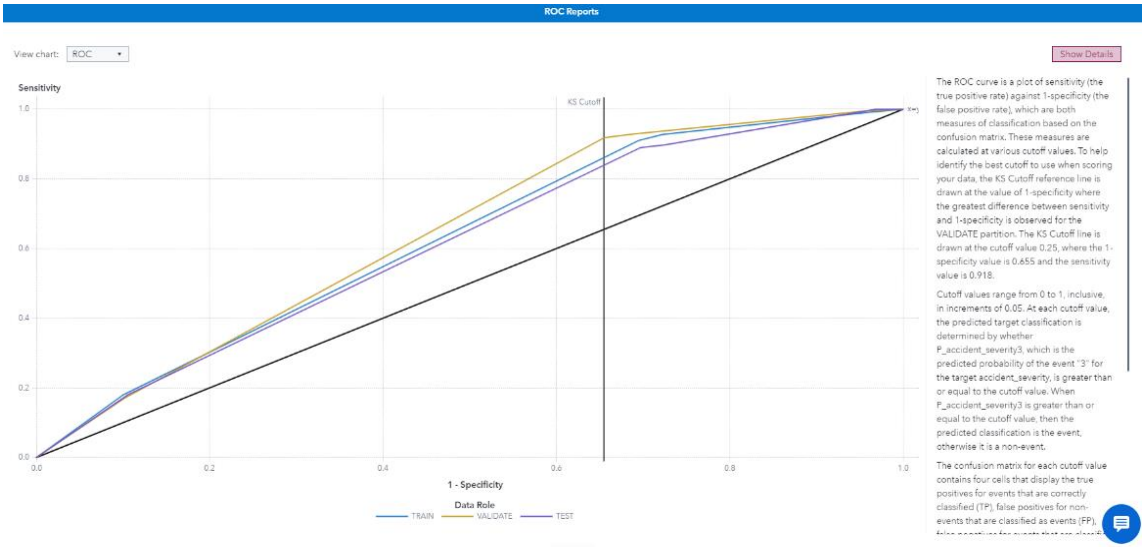



**Fig 21 – ROC - Logistic Regression**

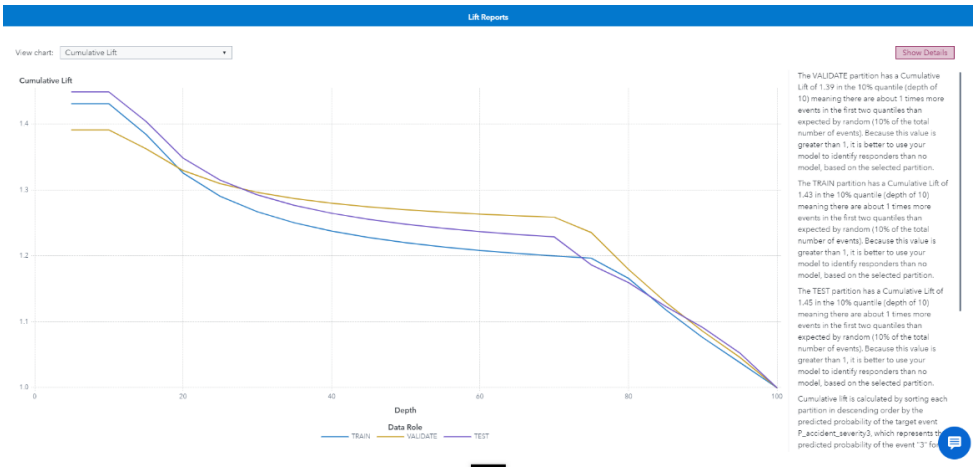



**Fig 22 – Cumulative Lift - Logistic Regression**

### 2.2.4 Comparative Analysis of the performance of models

**Table 4 – Model Comparison**

| | KS (youden) | Average square error (ASE) | AUC | Accuracy | Cumulative Lift |
|---|---|---|---|---|---|
| **Neural Network** | **0.7047** | **0.0749** | **0.9057** | **0.8556** | **2.5984** |
| Decision Tree | 0.3976 | 0.1527 | 0.7316 | 0.7585 | 2.7756 |
| Logistic Regression | 0.1929 | 0.1976 | 0.6136 | 0.6667 | 1.4488 |

From Table 4, it is clear that the Neural Network model demonstrates superior performance compared to the Decision Tree and Logistic Regression models in most criteria. It possesses the greatest KS (Kolmogorov-Smirnov) statistic, which suggests exceptional proficiency in discriminating between several classes. The Average Square Error (ASE) is the smallest, indicating a superior match to the data. The Neural Network exhibits the largest AUC (Area Under the ROC Curve), indicating its superior performance in classification.

The Decision Tree has intermediate KS and AUC values, surpassing the Neural Network in ASE but still maintaining comparatively high accuracy and the highest Cumulative Lift. This reflects its robust performance in accurately selecting the positive class compared to a random guess.

Logistic Regression exhibits the poorest performance in this particular scenario, as indicated by the lowest values in KS, AUC, and accuracy, and a high ASE. The model's Cumulative Lift is notably inferior to that of the other models, suggesting it has a lesser efficacy in detecting the positive class.

In summary, the Neural Network is the most effective model for predicting accident severity in this scenario.

### 2.2.5 Most Important Features for Champion Model (Top 4)

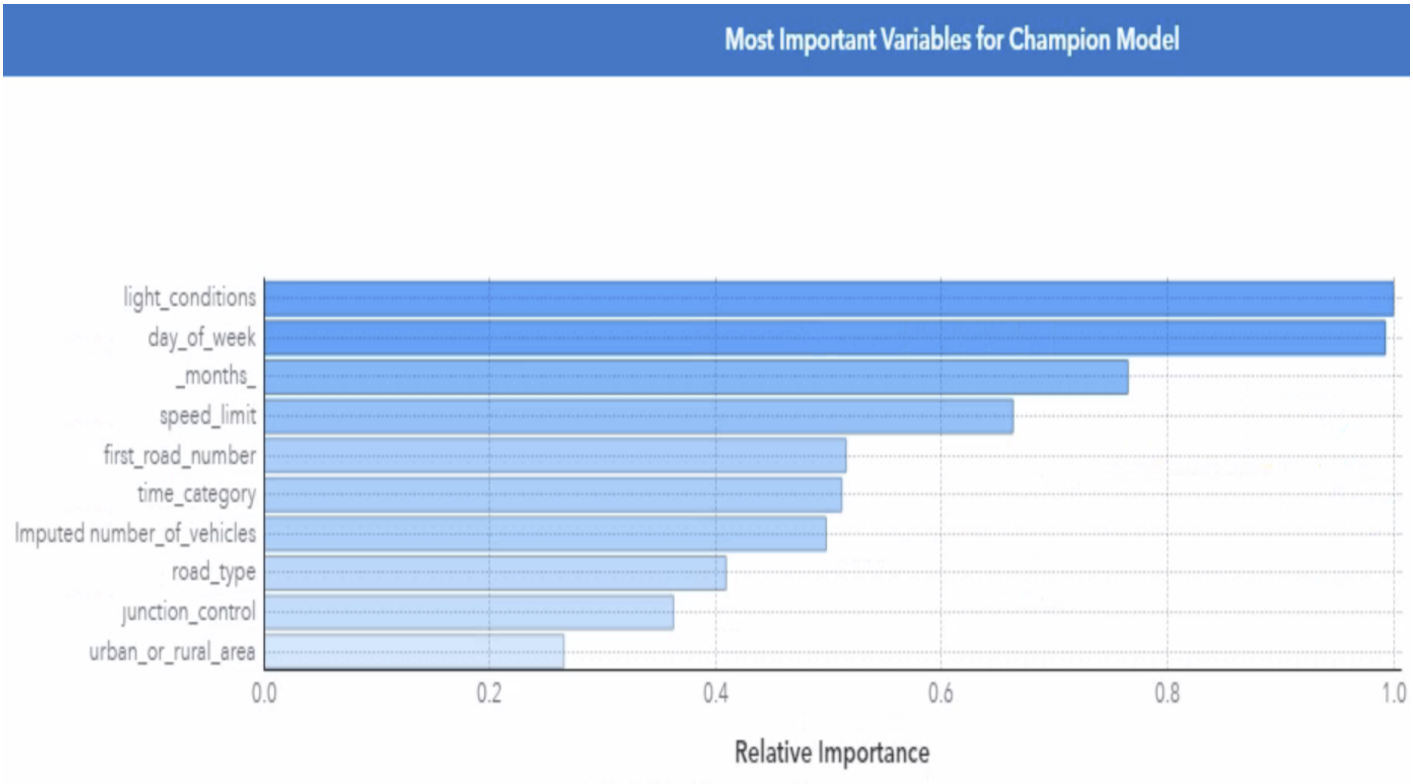



**Fig 23 – Relative Importance of Variables for Champion Model**

The primary feature for the Champion Model is 'light_conditions', having the highest importance score of 1. This is closely followed by 'day_of_week' with a score of 0.99. The third significant feature is 'months', with an importance score of 0.76, and finally, 'speed_limit' ranks fourth with a score of 0.66.

## **Task 3** – Text Analysis of Tweets

### 3.1 Data

The dataset consists of textual data collected from tweets regarding road traffic accidents in the Surrey region.

### 3.2 Text Analysis Pipeline

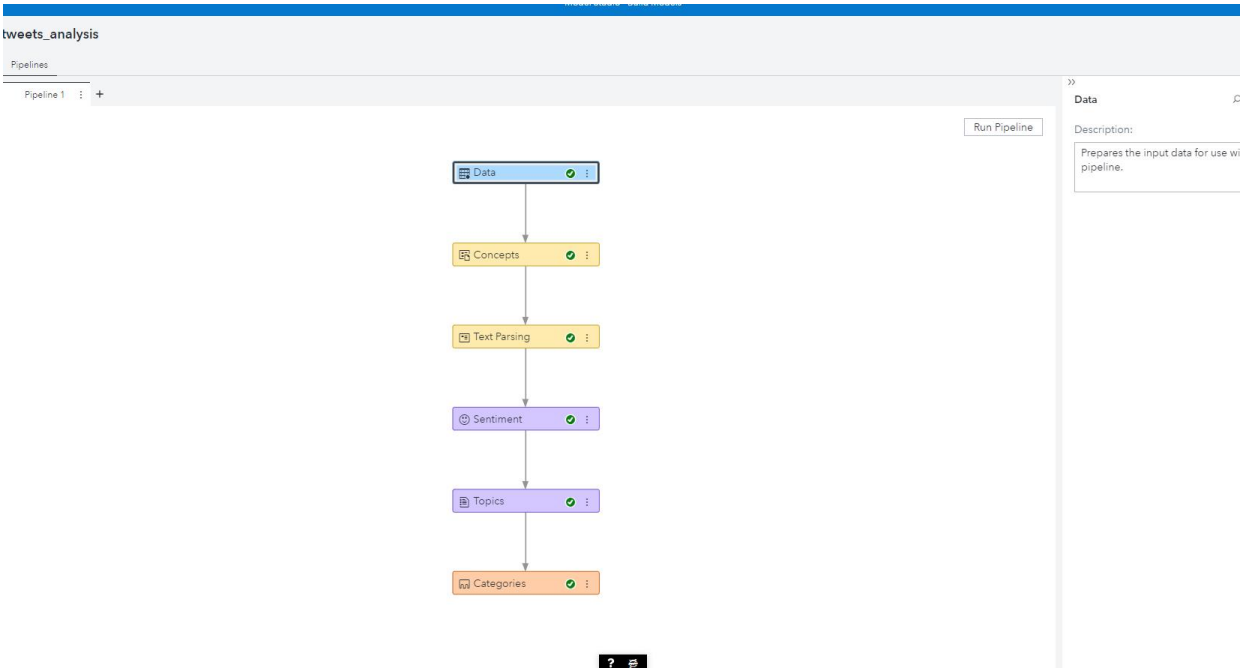



**Fig 24 – Text Analysis Pipeline**

Figure 24 depicts the text analysis pipeline with each node representing a step in processing and analyzing the text data. Each node in the pipeline will be explained below.

#### 3.2.1 Data Node

This is the starting point of the pipeline and represents the input data that will be used in the analysis.

#### 3.2.2 Concepts Node

The concepts node is used for identifying and extracting key concepts from the text data. This could involve identifying significant phrases that represent the main ideas or topics within the text.

*Custom Concepts*

The following user-defined concepts were created using appropriate classifiers and expressions to identify the main ideas within the text.

*_collision_*

CLASSIFIER:collisions

CLASSIFIER:collision

CLASSIFIER:Clash

CLASSIFIER:collided

CLASSIFIER:collide

CLASSIFIER:Head-On

CLASSIFIER:crash

*_rollover_*

CLASSIFIER:Rollover

CLASSIFIER:Overturned

CLASSIFIER:instability

*_Pedestrian_involved_*

CLASSIFIER:Pedestrian

CLASSIFIER:Crosswalk

*_weekend_*

CLASSIFIER:Saturday

CLASSIFIER:Sunday

CLASSIFIER:Weekend

*_weekday_*

CLASSIFIER:Monday

CLASSIFIER:Tuesday

CLASSIFIER:Wednesday

CLASSIFIER:Thursday

CLASSIFIER:Friday

CLASSIFIER:Weekday

*_Road_M25_*

CLASSIFIER:M25

*_Road_A3_*

CLASSIFIER:A3

*_Road_M23_*
CLASSIFIER:M23

*_Road_M3_*
CLASSIFIER:M3

*_Road_A31_*
CLASSIFIER:A31

*_Road_A322_*
CLASSIFIER:A322

*_Road_A217_*
CLASSIFIER:A217

*_Road_A25_*
CLASSIFIER:A25

*_Road_A283_*
CLASSIFIER:A283

*_Road_A243_*
CLASSIFIER:A243

*_Road_A331_*
CLASSIFIER:A331

*_Junction_J8_*
CLASSIFIER:J8

*_Junction_J9_*
CLASSIFIER:J9

*_Junction_J10_*

CLASSIFIER:J10

*_Junction_J3_*

CLASSIFIER:J3

*_Junction_J2_*

CLASSIFIER:J2

*_Junction_J11_*

CLASSIFIER:J11

*_Junction_J7_*

CLASSIFIER:J7

*_Junction_J6_*

CLASSIFIER:J6

*_Junction_J4_*

CLASSIFIER:J4

*_Minor_Accident*

CLASSIFIER:Minor

CLASSIFIER:Low

CLASSIFIER:Slight

CLASSIFIER:Dent

*_Serious_Accident*

CLASSIFIER:Serious

CLASSIFIER:Fatal

CLASSIFIER:Emergency

CLASSIFIER:Fire

CLASSIFIER:Major

CLASSIFIER:Severe

CLASSIFIER:Injury

CLASSIFIER:Hospitalized

CLASSIFIER:Explosion

CLASSIFIER:Life-Threatening

CLASSIFIER:Killed

CLASSIFIER:Death

CLASSIFIER:Died

*_car_*

CLASSIFIER:Car

*_bicycle_*

CLASSIFIER:Cyclist

CLASSIFIER:Bicycle

CLASSIFIER:Cycling

*_motorcycle_*

CLASSIFIER:Motorcycle

CLASSIFIER:Biker

CLASSIFIER:Motorbike

CLASSIFIER:Motorcyclist

*_Truck_*

CLASSIFIER:18-Wheeler

CLASSIFIER:Tanker

CLASSIFIER:Truck

*_Morning_Accident_*

CLASSIFIER:Morning

*_Afternoon_Accident*

CLASSIFIER:Afternoon

CLASSIFIER:noon

*_Night_Accident_*

CLASSIFIER:Night

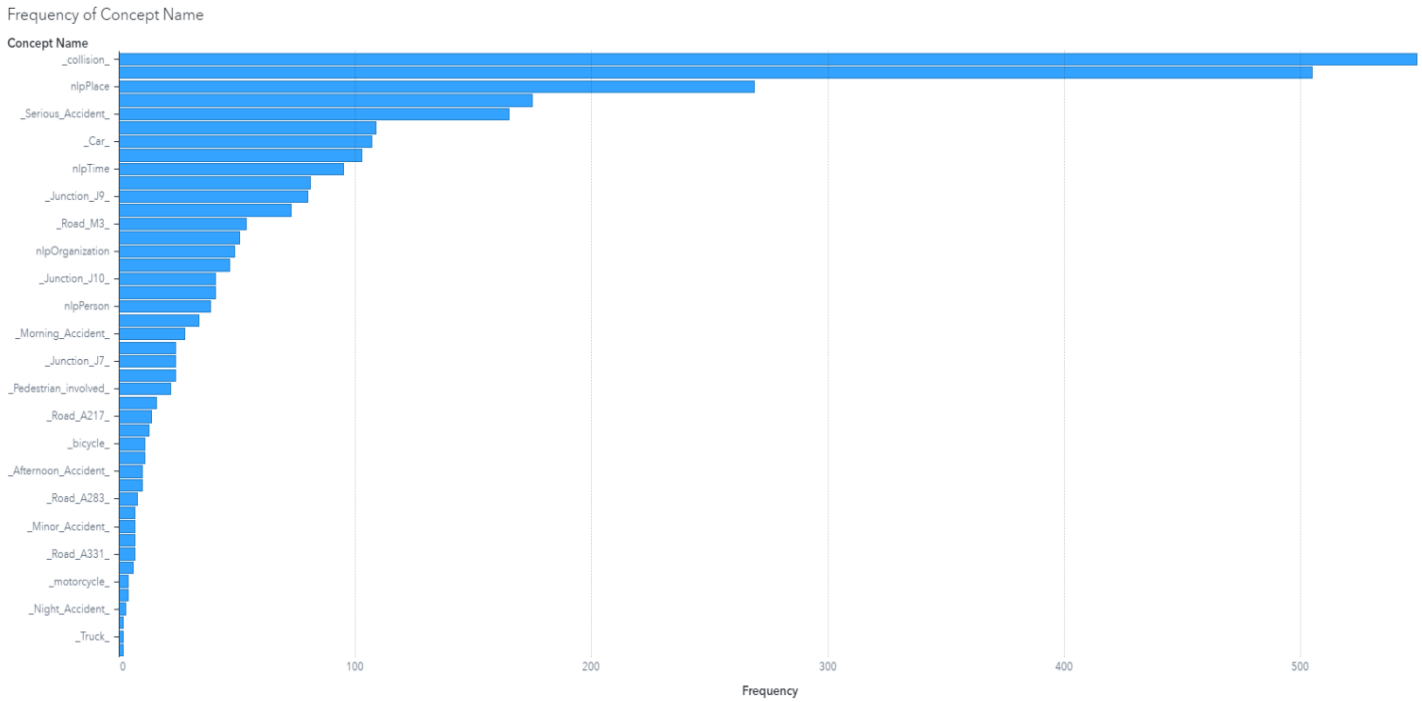



**Fig 25 - Frequency Distribution of Concepts**

The concept "_collision_" has the highest frequency, suggesting that the term "collision" is very common in the dataset. It is likely one of the primary subjects of the tweets related to road accidents. A significant number of tweets discuss serious accidents, indicating the severity of the incidents being reported or discussed. The high frequency of the concept "Road_M25" suggests that Road M25 is more likely to be a location with a higher incidence of accidents. The "_Car_" concept has a significantly high frequency which indicates that most of the time cars are involved in road accidents compared to other vehicles. The frequent occurrence of the concepts "_Junction_J9_" and "_Junction_J10_" suggests that junctions 'J9' and 'J10' are particularly prone to accidents. The frequent occurrence of the concept "Road_M3" implies that Road M3 is likely to be a site with a greater prevalence of accidents. By comparing the frequencies of the concepts "_Morning_accidents_", "_Afternoon_Accidents_", and "_Night_Accidents_", it can be inferred that a greater number of accidents occur in the morning as compared to other times of the day. The occurrence rate of the concepts "_Pedestrian_involved_" and "_bicycle_" suggests that pedestrians and cyclists are also involved in the accidents mentioned in the tweets.


**Fig 26 – Word Cloud of Concepts**


**Fig 27 – Word Cloud of keywords**

### 3.2.3 Text Parsing Node

*3.2.3.1 Text Parsing Process*

Text parsing is a series of operations that converts unstructured text into a structured format to comprehend and analyze the data. The process commences with tokenization when the text is divided into discrete pieces such as words or sentences. Next, the process of normalization ensues, encompassing tasks such as unifying the case of all characters and eliminating punctuation to minimize inconsistencies in the text. The following step is filtering, in which commonly used terms, referred to as stopwords, are eliminated to prioritize the more significant text. Stemming and lemmatization enhance the process by reducing words to their fundamental or canonical form.

Subsequently, part-of-speech tagging allocates grammatical functions to individual words, facilitating comprehension of sentence structure. Named Entity Recognition (NER) is a process that detects and classifies proper nouns and other important phrases. Syntactic parsing, on the other hand, breaks down the grammatical structure of a sentence, aiding in the understanding of the links between words. The ultimate phase, semantic analysis, aims to grasp the significance and context, enabling sophisticated comprehension and future utilization such as sentiment analysis or topic modeling.

*3.2.3.2 Kept Terms and Dropped Terms*

The terms that were included as a result of parsing are known as kept terms and the terms that were excluded are known as dropped terms. The relevant terms from the dropped terms are transferred to kept terms, while the unwanted terms from the kept terms are moved to dropped terms.

*3.2.3.3 Term Map*

A term map is a graphical depiction of the interconnections among various terms in a collection of texts. The visualization displays the frequency of terms and their interconnections based on their co-occurrence in the text. Nodes of greater size indicate terms that occur more frequently, while the lines signify the co-occurrence of two terms in the same context. This facilitates the identification of patterns, trends, and critical topics within the data.

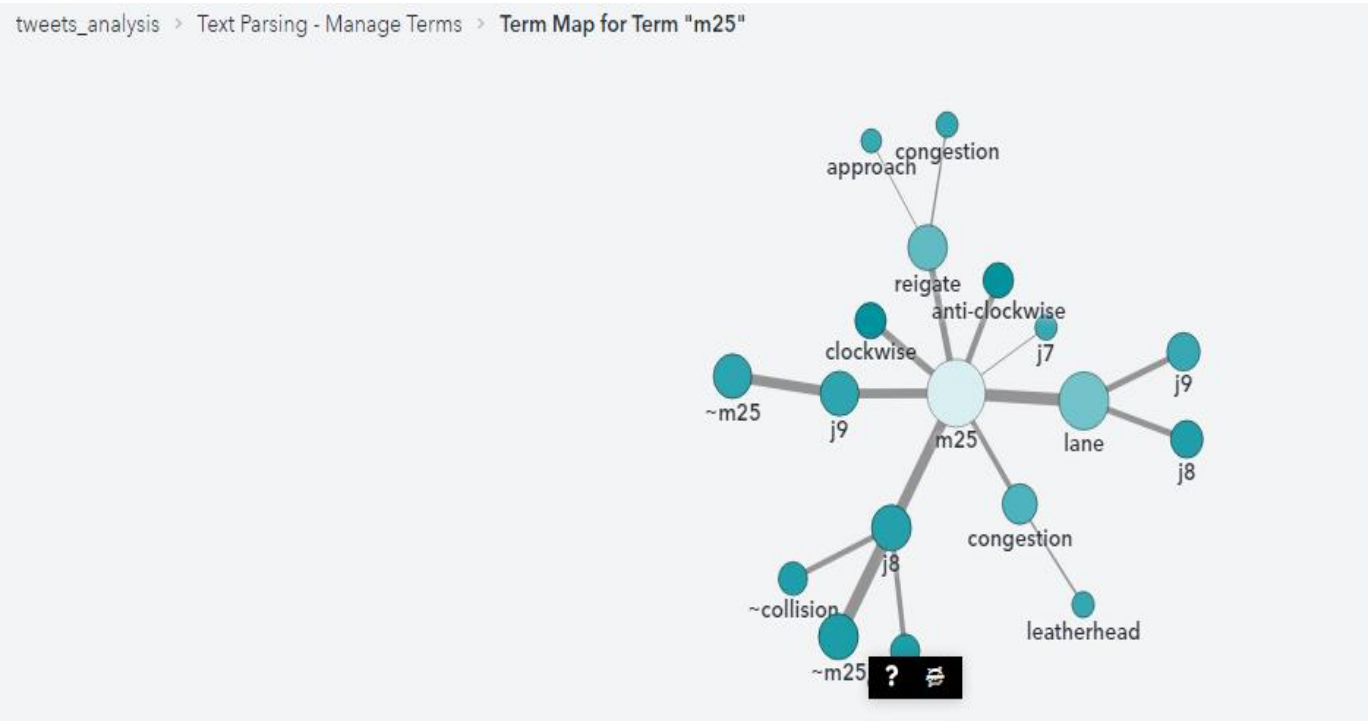



**Fig 28 – Term map**

For example, the term map in Figure 28 visualizes keywords from data concerning road accidents on the M25 motorway. Terms like "collision" and "congestion" suggest common themes in the data. Junctions "j7," "j8," and "j9" are prominent, possibly pinpointing frequent accident locations or traffic bottlenecks. "Leatherhead" and "Reigate" appear as sites of note for incidents or heavy traffic.

### 3.2.4 Sentiment Node

The sentiment analysis node is used to determine the sentiment expressed in the text. It typically categorizes the sentiment as positive, negative, or neutral.

### 3.2.5 Topics Node

In the text analysis pipeline, the "topics" node generally denotes the component that is accountable for discerning and extracting the primary themes or subjects from a set of texts. This is commonly accomplished through a technique referred to as "topic modeling." Topic modeling algorithms detect groups of terms that commonly appear together in the texts. Every cluster represents a "topic", which is an intangible notion that categorizes associated phrases.

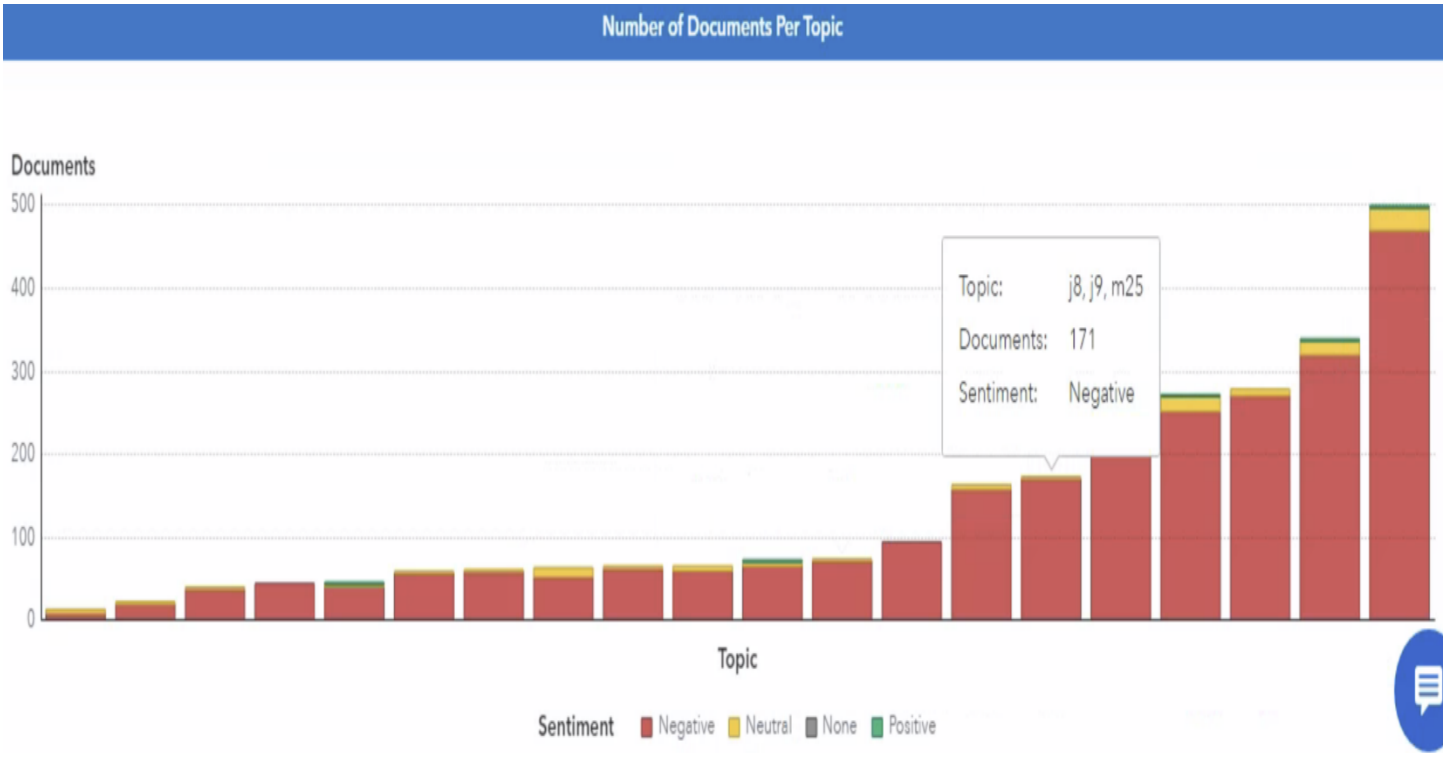



**Fig 29 – Number of Documents Per Topic**

For instance, within the realm of road traffic accidents, terms such as "j8", "j9", and "m25" may constitute a topic known as "Accident hotspots". Likewise, different topics such as "Vehicle", "accident severity", "months", "Weekday", "Locations", "Roads", "accident type" and "Time Category" has been identified in the document.

### 3.2.6 Categories Node

The "Categories" node in a text analysis pipeline is tasked with categorizing or grouping the documents into predetermined or identified groups depending on their content. The topics based on their terms are added as categories. The following are the discovered categories.

*Major_Accident*

(OR,(AND,"collision"),(AND,"serious"),(AND,"crash"),(AND,"multi-vehicle collision"),(AND,"killed"),(AND,"serious"),(AND,"fatal"),(AND,"died"),(AND,"head-on collision"),(AND,(OR,"injuries","injury")),(AND,"emergency"))

*Emergency_Response*

(OR,(AND,"police"),(AND,"emergency"),(AND,"hospital"),(AND,(OR,"investigation","investigations")),(AND,"recovery"),(AND,(OR,"called","call","calling")),(AND,(OR,"reports","report","reported","reporting")),(AND,(OR,"ambulances","ambulance")))

*Car_Accident*

(OR,(AND,(NOT,"collision"),(OR,"cars","car")))

*Accidents_in_Trunk_Roads*

(OR,(AND,"a3"),(AND,"a31"),(AND,(OR,"late","latest")),(AND,"a217"),(AND,"a25"))

*Motorway_Accidents*

(OR,(AND,"m25"),(AND,"m3"))

### 3.3 Summary

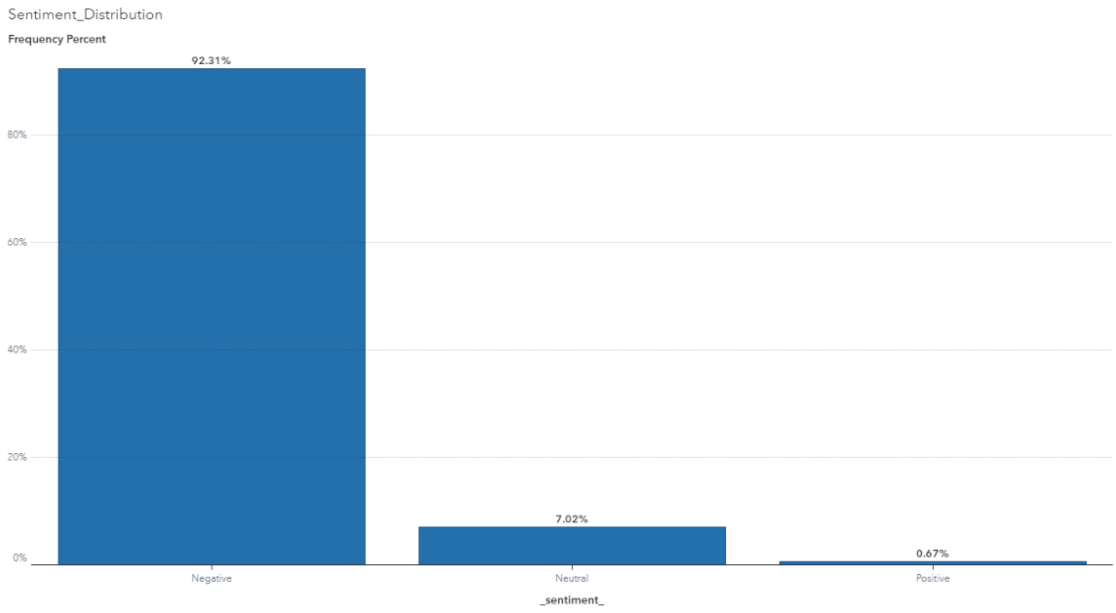


**Fig 30 – Sentiment Distribution**

The substantial proportion (92.31%) suggests that the prevailing sentiment of tweets concerning road traffic accidents in Surrey is negative. This could be attributed to the fact that traffic accidents are commonly linked to adverse occurrences such as delays, property damage, injuries, or stress. A relatively small portion (7.02%) of the tweets are classified as neutral. These could be tweets that report accidents without expressing any emotion or opinion, such as factual updates from traffic news sources. A very small fraction (0.67%) of the tweets are positive. These statements could express thanks towards emergency services, relief that no significant injuries happened, or favorable remarks about the effective handling of traffic despite an accident.

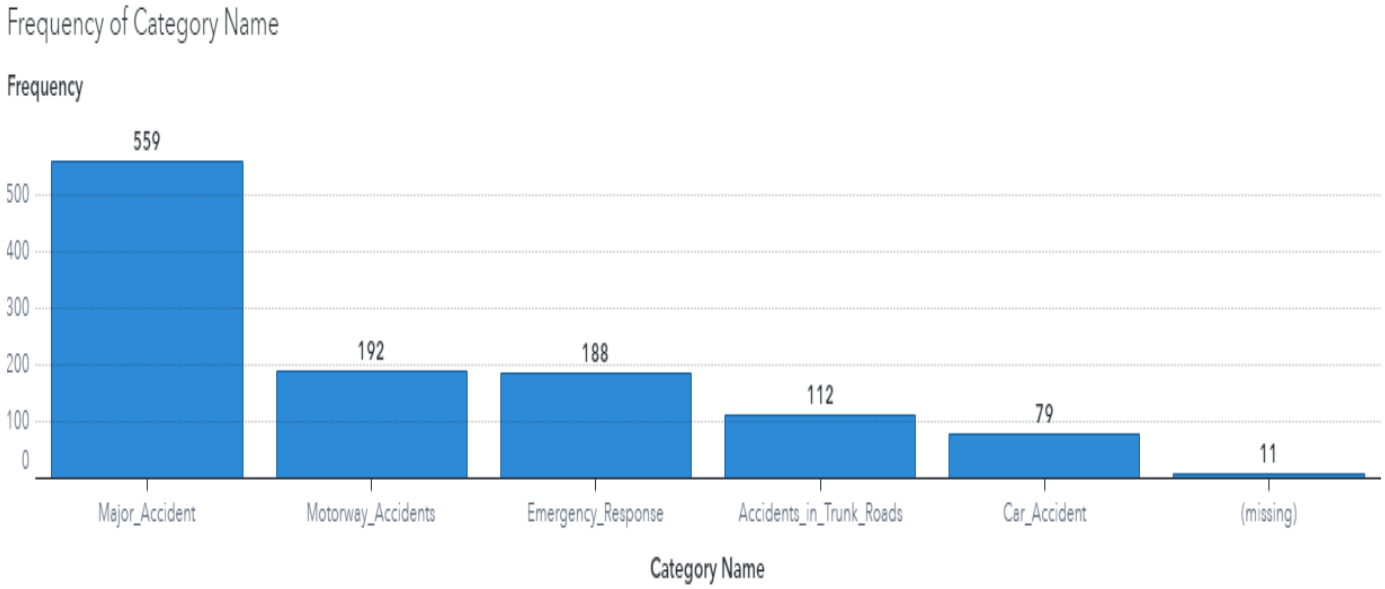



**Fig 31 – Number of Tweets Per Category**

The category with the highest number of tweets is "Major_Accident," with 559 occurrences, suggesting that this topic is discussed most frequently in the dataset and possibly reflects the most significant concern among the public or authorities. The "Motorway_Accidents" category, with 192 tweets, likely covers incidents occurring on high-speed motorways, which are significant traffic arteries. The "Emergency_Response" category is just slightly less frequent, with 188 tweets, which would include discussions around the actions taken by emergency services in response to road incidents. "Accidents_in_Trunk_Roads" has 112 tweets and refers to incidents on major roads that are not classified as motorways. The "Car_Accident" category, with 79 tweets, focuses specifically on accidents involving cars.

## **Task 4** – Decision-Maker's Summary and Recommendations

### 4.1 Accident Severity Prediction

This dataset includes data regarding vehicular accidents that occurred in the Surrey region throughout the year 2022. Each row in the dataset corresponds to a single accident, and it contains 35 columns that cover different characteristics such as geographic coordinates, accident severity, number of vehicles and casualties, road conditions, and other relevant information.

#### 4.1.2 Insights from Data Analysis

*Weather Conditions*: The majority of accidents, both serious (20%) and minor (63%), occur in fine weather without strong winds.

*Speed Limits and Accident Severity*: Highways with a 30mph speed limit report the most severe accidents.

*Road Types*: Single carriageways are the most common accident locations, with dual carriageways following.

*Road Surface Conditions*: Dry conditions are predominant in accident occurrences, closely followed by wet conditions.

*Lighting Conditions*: Accidents frequently happen in daylight or well-lit conditions at night.

*Day of the Week*: Thursdays and Fridays witness the highest number of accidents.

*Vehicle Involvement*: Around 60% of accidents involve two vehicles, while 26% involve only one.

*Casualties*: The majority of accidents (79%) result in one fatality, and 14% result in two fatalities.

*Junction Types*: T-junctions are involved in 17% of the accidents.

*Junction Control*: Uncontrolled junctions feature in 35% of accidents.

*Time of Day*: Afternoons, followed by mornings, see the most accidents.

*Weekly Trends*: Fridays are prone to afternoon accidents, Thursdays to morning accidents, and Sundays to night accidents.

*Monthly Trends*: November and January have the highest accident rates

*Peak Hour*: 5 pm is identified as the hour with the most accidents.

#### 4.1.3 Recommendations for Enhancing Road Safety

1) *Improved Road Markings and Signage*: It is important to take measures to improve the road markings and signage around uncontrolled intersections, particularly in areas where slip roads merge. To alert drivers, this would include visible yield or stop signs, as well as warnings that are painted on the surface of the road.

2) *Speed Control Measures*: To urge drivers to reduce their speed, it is recommended that speed control measures be implemented at the intersection. These measures may include rumble strips or speed bumps.

3) *Improved Lighting and Visibility*: Increase lights in places that are considered to be high-risk. It is extremely important to encourage the use of reflective materials and lighting on vehicles, as well as on pedestrians and bicycles, especially during times when visibility is low.

4) The fourth recommendation is to increase the amount of *traffic monitoring and enforcement* on Thursdays and Fridays, which are the days of the week when the number of accidents is at its highest.

5) *Maintenance of the road surface* should be performed regularly to ensure it is kept in good shape. This maintenance should focus on both dry and wet circumstances with equal importance.

### 4.2 Text Analysis

The dataset comprises textual information gathered from tweets on road traffic accidents in the Surrey region.

#### 4.2.1 Insights

1. The high share (92.31%) implies Surrey road traffic accident tweets are negative. Only 7.02 percent of tweets are neutral and only 0.67 percent of tweets are positive.
2. The dataset has 559 tweets on "Major_Accident," showing that this topic is addressed most often and may be the most concerning to the public or authorities. The "Motorway_Accidents" category, with 192 tweets, certainly covers motorway accidents, which are major traffic arteries. The "Emergency_Response" category, which covers emergency services' road incident responses, has 188 tweets. "Accidents_in_Trunk_Roads" features 112 tweets about serious non-motorway road accidents. The "Car_Accident" category has 79 tweets on car accidents.

#### 4.2.3 Recommendations

1. Given the high volume of negative sentiment, there is an opportunity to launch public awareness campaigns focused on road safety. These campaigns can educate drivers, pedestrians, and cyclists about safe practices, the importance of being attentive, and the consequences of risky behaviors such as speeding or distracted driving.
2. A small proportion of tweets with favorable sentiment may indicate discontentment with emergency services or traffic control. Efficient and proactive management of traffic events is necessary to enhance the general public's sentiment. This objective can be accomplished by increasing the allocation of financial resources and personnel to the emergency services and traffic control division.
3. With "Motorway Accidents" a major issue, road safety must be improved. This might include better lighting, signs, more frequent road safety patrols, and advanced weather and traffic warning systems.
4. To specifically address car accidents, efforts can be directed towards boosting seat belt usage, preventing distracted driving (such as using mobile phones while driving), and implementing anti-drunk driving programs.

## References


1. SAS Institute Inc. 2019. Exploring SAS® Viya®: Visual Analytics, Statistics, and Investigations. Cary, NC: SAS Institute Inc.
2. SAS Institute Inc. 2019. Exploring SAS® Viya®: Data Mining and Machine Learning. Cary, NC: SAS Institute Inc.
3. Larose, DT 2015, Data Mining and Predictive Analytics, John Wiley & Sons, Incorporated, Newark. Available from: ProQuest Ebook Central. [5 January 2024].


## Appendix

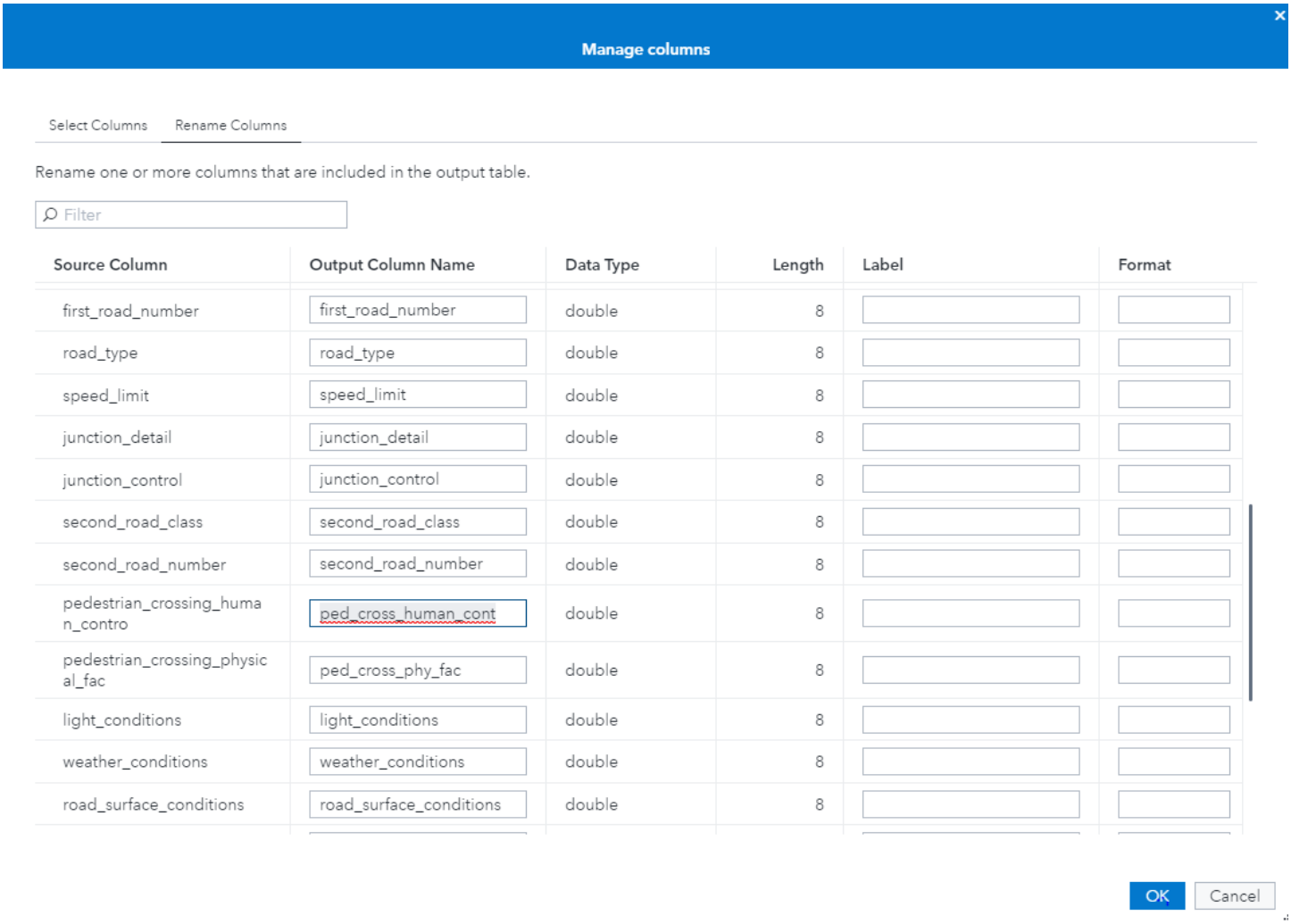



**Fig 32 – Column Renaming**

Renamed all the Columns having more than 32 characters

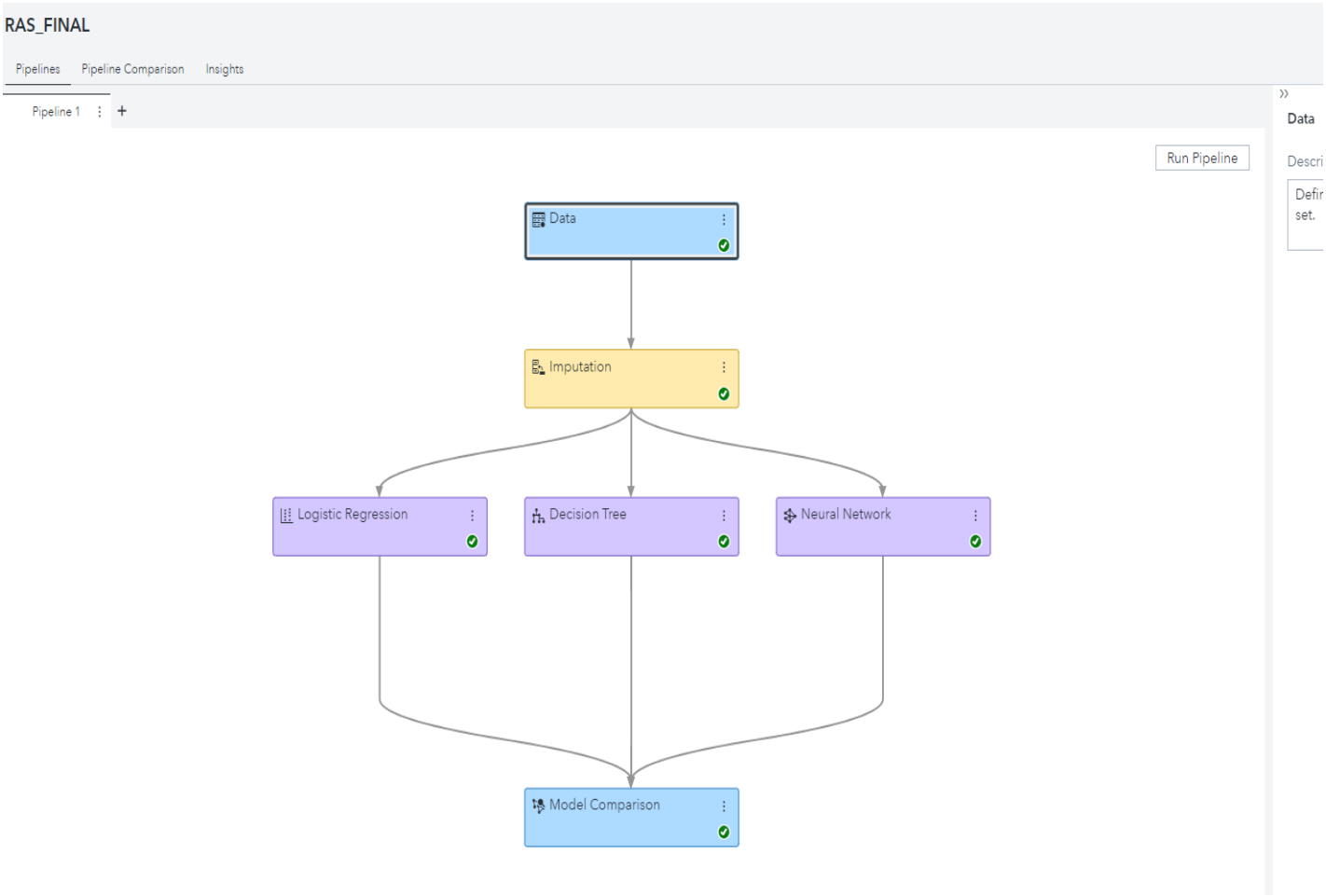



**Fig 33 – Machine Learning Pipeline**