

<b>Module title:</b>	<b>Data mining and text analytics With application in SAS</b>
<b>Student number (id)</b>	6897234
<b>Assessment title:</b>	<b>Individual Assignment - Exploring Road Traffic Accident Data and Text Analytics Insights</b>

### Task 1 – Data Exploration and Cleaning [20 marks]

To show your skills in data exploration, visualization, summary statistics generation, and data cleaning.

The dataset contains detailed records of road accidents in Surrey, UK, during 2021, covering aspects like severity, road conditions, weather influences, and involved parties. It forms the basis for exploring accident patterns, predicting severity using machine learning, and analyzing related text data for actionable insights.

### Exploratory Data Analysis

Missing Data Frequencies

Legend: ., A, B, etc = Missing

Row	Frequency	Percent
Non-missing	2480	100.00

acci_ref	Frequency	Percent
Non-missing	2480	100.00

loc_east_osgr	Frequency	Percent
.	1	0.04
Non-missing	2479	99.96

loc_nor_osgr	Frequency	Percent
Non-missing	2480	100.00

longitude	Frequency	Percent
.	1	0.04
Non-missing	2479	99.96

latitude	Frequency	Percent
.	1	0.04
Non-missing	2479	99.96

police_force	Frequency	Percent
.	1	0.04
Non-missing	2479	99.96

acci_severity	Frequency	Percent
.	1	0.04
Non-missing	2479	99.96

num_of_vehi	Frequency	Percent
.	1	0.04
Non-missing	2479	99.96

num_of_casu	Frequency	Percent
.	1	0.04
Non-missing	2479	99.96

date	Frequency	Percent
.	1	0.04
Non-missing	2479	99.96

day_of_week	Frequency	Percent
.	1	0.04
Non-missing	2479	99.96

time	Frequency	Percent
Non-missing	2480	100.00

local_auth_distr	Frequency	Percent
Non-missing	2480	100.00

loc_auth_ons_distr	Frequency	Percent
Non-missing	2480	100.00

loc_auth_highw	Frequency	Percent
.	1	0.04
Non-missing	2479	99.96

first_road_class	Frequency	Percent
Non-missing	2480	100.00

first_road_num	Frequency	Percent
Non-missing	2480	100.00

road_type	Frequency	Percent
Non-missing	2480	100.00

speed_limit	Frequency	Percent
Non-missing	2480	100.00

junc_detail	Frequency	Percent
Non-missing	2480	100.00

junc_con	Frequency	Percent
Non-missing	2480	100.00

sec_road_class	Frequency	Percent
Non-missing	2480	100.00

sec_road_num	Frequency	Percent
Non-missing	2480	100.00

ped_cross_hum_con	Frequency	Percent
Non-missing	2480	100.00

ped_cross_phy_facil	Frequency	Percent
Non-missing	2480	100.00

light_con	Frequency	Percent
Non-missing	2480	100.00

weath_con	Frequency	Percent
Non-missing	2480	100.00

road_surf_con	Frequency	Percent
Non-missing	2480	100.00

spec_con_site	Frequency	Percent
Non-missing	2480	100.00

carri_haz	Frequency	Percent
Non-missing	2480	100.00

urb_or_rur_area	Frequency	Percent
Non-missing	2480	100.00

did_poli_offi_att	Frequency	Percent
Non-missing	2480	100.00

tru_road_flag	Frequency	Percent
Non-missing	2480	100.00

lsoa_of_acc_loc	Frequency	Percent
Non-missing	2480	100.00

Table 1

The exploratory data analysis revealed the following insights:

**Structure of the Dataset:**

The dataset comprises 2,480 rows and 35 columns, detailing road accidents in Surrey, UK, during 2021. It includes numerical, categorical, and temporal variables.

**Key Variables:**

- Longitude and latitude: Geographical location.
- Accident severity: Severity level.
- Number of vehicles involved.
- Number of casualties.
- Weather conditions: At the time of the accident.

**Missing Data:**

Variables such as longitude, latitude, accident severity, number of vehicles, and number of casualties each have one missing value. The **Describe Missing Values** feature in SAS Viya confirmed these patterns.

**Data Quality:**

The dataset is mostly complete, with minimal missing data. Temporal variables like time and date require format checks, and categorical variables may need cleaning to address inconsistencies.

## Summary Statistics

The dataset's summary statistics, generated in SAS Viya, included central tendencies (mean, median, mode) and dispersion (standard deviation, range). These statistics revealed insights into variables such as accident severity, number of casualties, and vehicles involved. Mean and standard deviation clarified data distribution while identifying potential outliers and inconsistencies.

Variable	Mean	Std Dev	Minimum	Maximum	N
time	50215.09	17892.30	60.0000000	86100.00	2480
Row	1240.50	716.0586568	1.0000000	2480.00	2480
acci_ref	451069167	33085.53	451011255	451160434	2480
loc_east_osgr	509570.09	14018.85	482163.00	543673.00	2479
loc_nor_osgr	157127.41	9087.04	132324.00	175208.00	2480
longitude	-0.4296579	0.2006301	-0.8317170	0.0570740	2479
latitude	51.3025884	0.0820440	51.0832120	51.4663730	2479
police_force	45.0000000	0	45.0000000	45.0000000	2479
acci_severity	2.7398144	0.4603663	1.0000000	3.0000000	2479
num_of_vehi	1.9096410	0.7675206	1.0000000	8.0000000	2479
num_of_casu	1.2803550	0.6953533	1.0000000	9.0000000	2479
date	22470.43	101.8460414	22281.00	22645.00	2479
day_of_week	4.0883421	1.9692607	1.0000000	7.0000000	2479
local_auth_distr	-1.0000000	0	-1.0000000	-1.0000000	2480
first_road_class	4.0766129	1.6026707	1.0000000	6.0000000	2480
first_road_num	355.2100806	785.1844174	0	3411.00	2480
road_type	5.1290323	1.6597056	1.0000000	9.0000000	2480
speed_limit	38.8548387	13.8880986	20.0000000	70.0000000	2480
junc_detail	2.0758065	2.9367233	0	9.0000000	2480
junc_con	1.1358871	2.3760271	-1.0000000	4.0000000	2480
sec_road_class	2.3786290	2.7298322	0	6.0000000	2480
sec_road_num	84.4302419	423.7698196	-1.0000000	3411.00	2480
ped_cross_hum_con	0.0237903	0.2102615	0	2.0000000	2480
ped_cross_phy_facil	0.6411290	1.8070059	0	8.0000000	2480
light_con	1.9665323	1.7018868	1.0000000	7.0000000	2480
weath_con	1.5403226	1.6205079	1.0000000	9.0000000	2480
road_surf_con	1.3133065	0.5700473	-1.0000000	5.0000000	2480
spec_con_site	0.1326613	0.7782217	-1.0000000	7.0000000	2480
carri_haz	0.0991935	0.6234617	-1.0000000	7.0000000	2480
urb_or_rur_area	1.4116935	0.4922394	1.0000000	2.0000000	2480
did_poli_offi_att	1.3834677	0.7281363	1.0000000	12.0000000	2480
tru_road_flag	1.8693548	0.3370798	1.0000000	2.0000000	2480
hour_of_day	13.5008065	4.9812071	0	23.0000000	2480

Table 2

## Data Visualization

The visualizations analyze factors influencing road accidents in Surrey in 2021, focusing on variables like accident severity, time, environmental conditions, and road characteristics. They uncover patterns, identify risk factors, and highlight areas for targeted interventions.

### Distribution of Accident Severity:

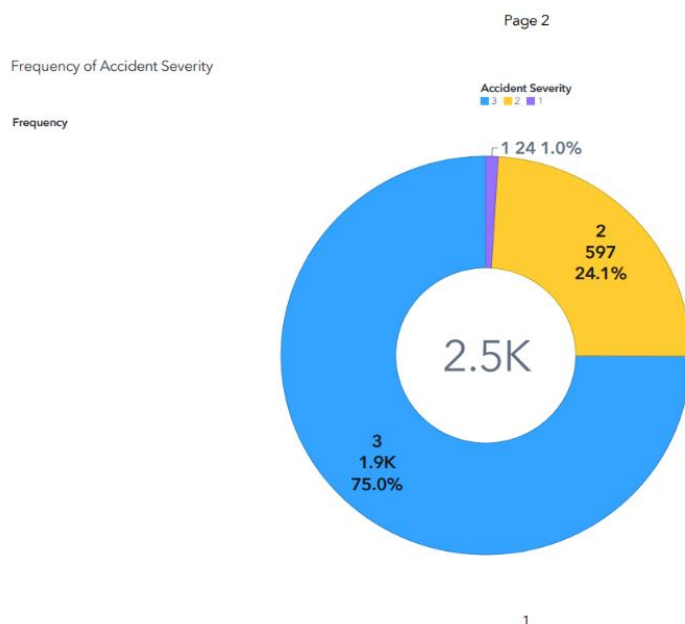


Image 1

The pie chart highlights the imbalance in the target variable, Accident Severity, which includes three levels:

1. **Severity Level 3 (Low Severity):** Majority class, accounting for 75% of the data.
2. **Severity Level 2 (Medium Severity):** Represents 24.1% of records.
3. **Severity Level 1 (High Severity):** Rarest class, making up only 1%.

This imbalance poses a challenge for predictive modelling, potentially biasing predictions toward the majority class. Techniques like oversampling or class-weighted algorithms may be necessary to improve performance on rare, high-severity cases.

### Frequency of Accidents by Hour of the Day:

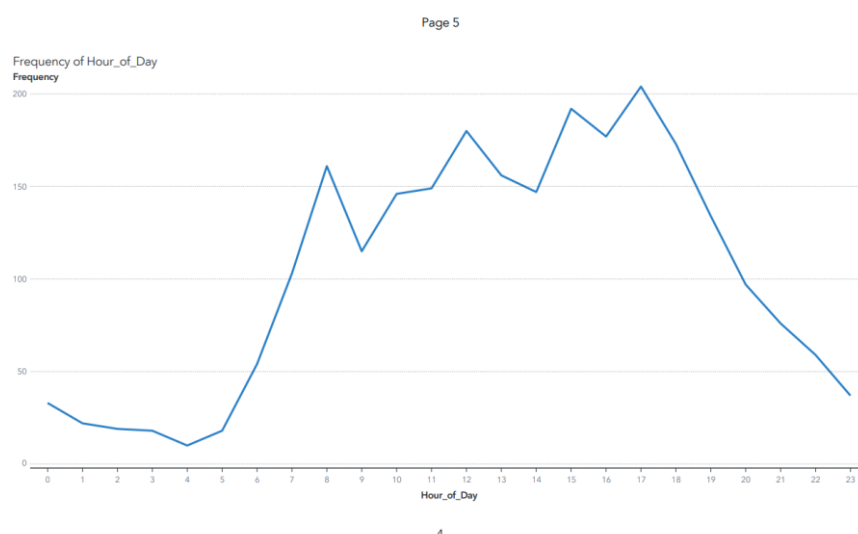


Image 2

The line graph shows the distribution of accidents across a 24-hour period:

- **Peak Hours:** Accidents spike during morning (8-9 AM) and evening (4-6 PM) rush hours due to increased traffic congestion.
- **Off-Peak Hours:** Accident frequency drops significantly between midnight and early morning (10 PM-6 AM) when traffic volume is lower.
- **Trend:** Higher accident frequency aligns with times of increased commuter activity.

This visualization highlights temporal patterns, emphasizing the need for targeted traffic management and safety measures during peak hours.

Accident Frequency by Junction Type:

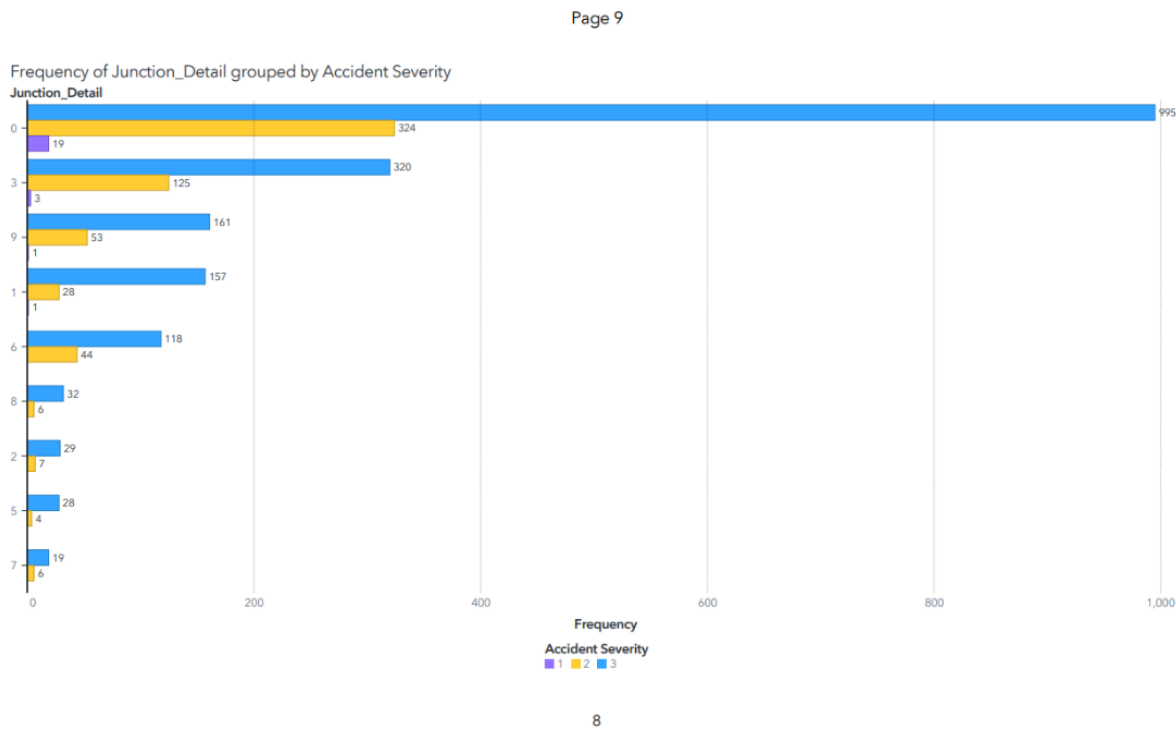


Image 4

The graph shows the distribution of accidents by severity (1: Fatal, 2: Serious, 3: Slight) across junction types. Non-junction areas (“0”) have the highest number of accidents, mainly slight, followed by serious and fatal. T or staggered junctions (“3”) are the second highest, primarily contributing to slight accidents. This highlights non-junction areas as major hotspots, with T or staggered junctions also playing a notable role in lower-severity accidents.

Frequency of Accidents by Road Type:

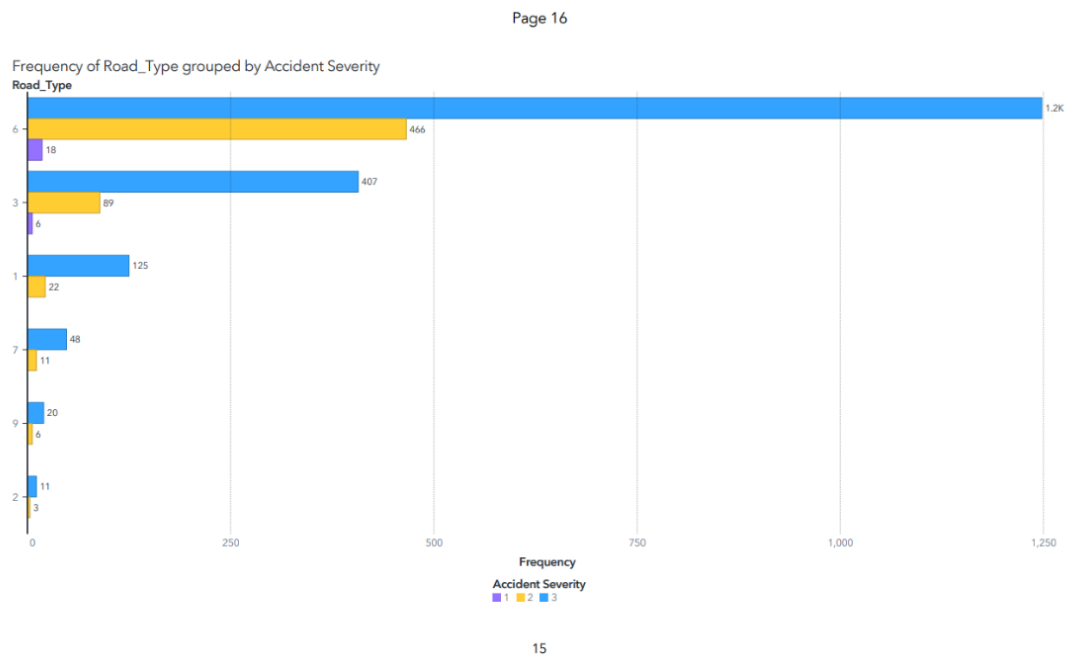


Image 3

The graph depicts accident frequency by severity (1: Fatal, 2: Serious, 3: Slight) across road types. Single carriageways have the highest number of accidents, primarily slight, followed by serious and fatal. Dual carriageways rank second, contributing significantly, while roundabouts have the lowest frequency and no fatal accidents, likely due to their safer design. This highlights single carriageways as major hotspots, with dual carriageways also notable and roundabouts relatively safer.

Frequency of Accidents by Speed Limit:

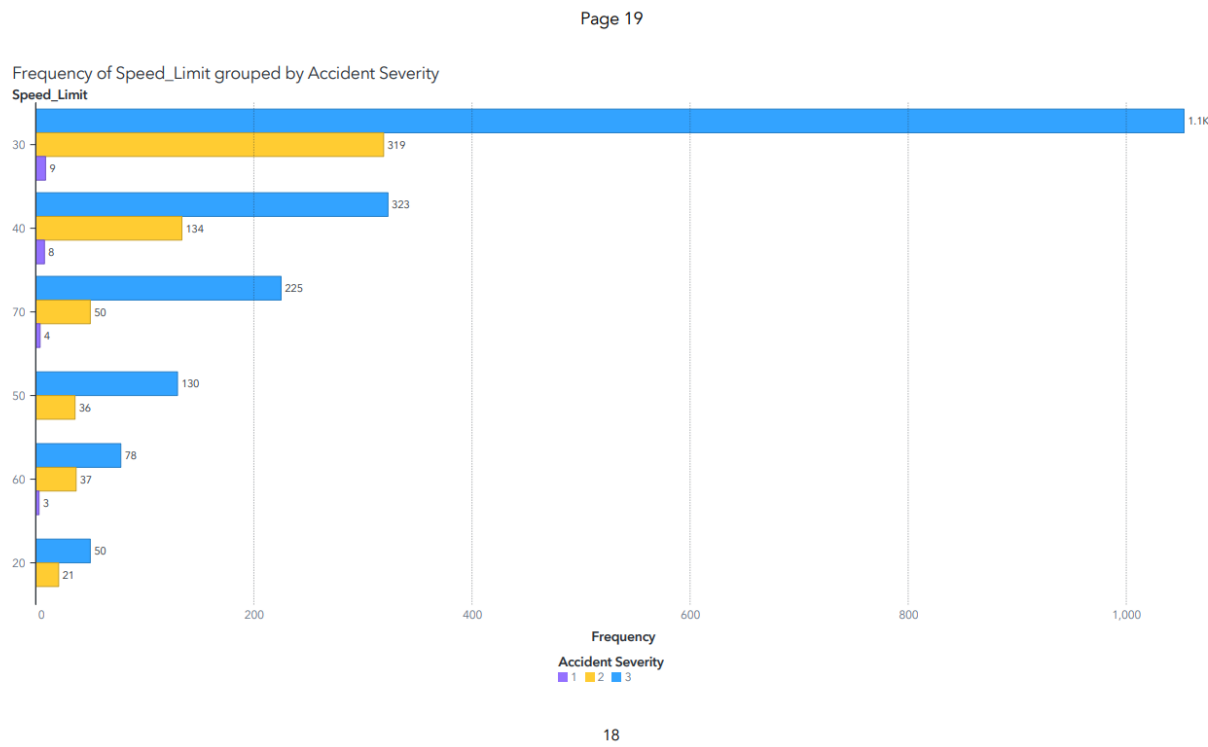


Image 5

The graph displays accident frequency by severity (1: Fatal, 2: Serious, 3: Slight) across speed limits. Most accidents occur in 30 mph zones, mainly slight, followed by serious and fatal, reflecting urban traffic patterns. While 40 mph and 70 mph zones see fewer accidents, they show significant serious and slight proportions. Higher-speed zones generally have fewer accidents but a greater risk of severe outcomes, emphasizing the dangers of high-speed travel.

### Frequency by Number of Casualties:

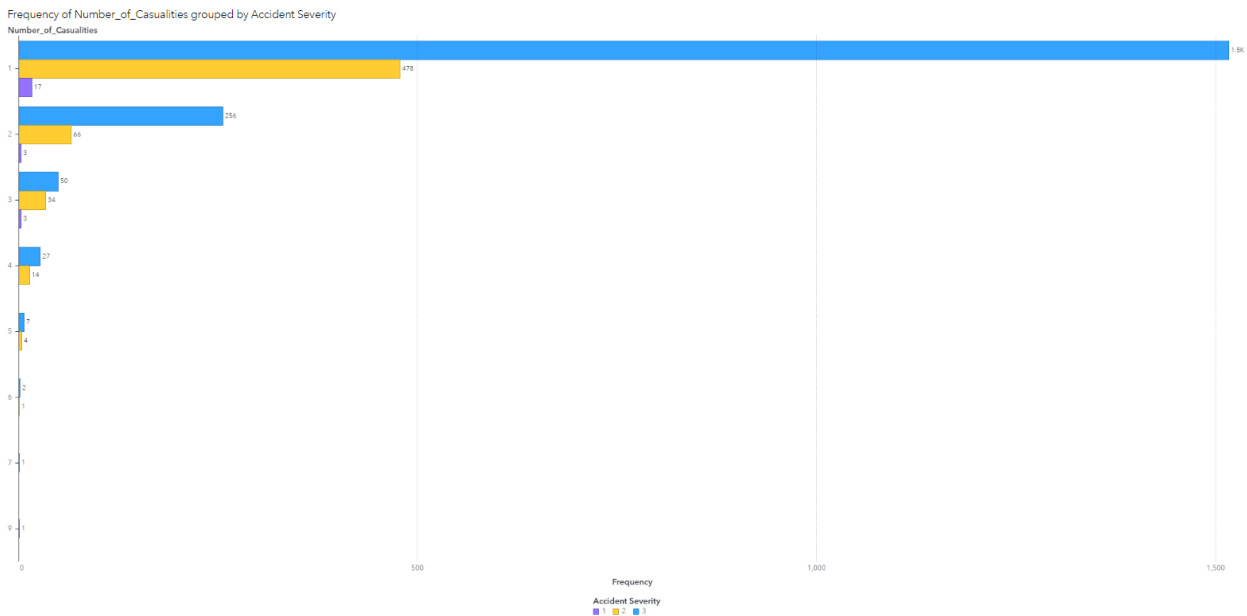


Image 6

The graph shows that accidents with a single casualty are most frequent, predominantly slight in severity. Accidents with multiple casualties are less common, with higher severity levels observed as the number of casualties increases. This highlights the link between casualty count and accident severity.

### Frequency by Urban/Rural:

Page 21

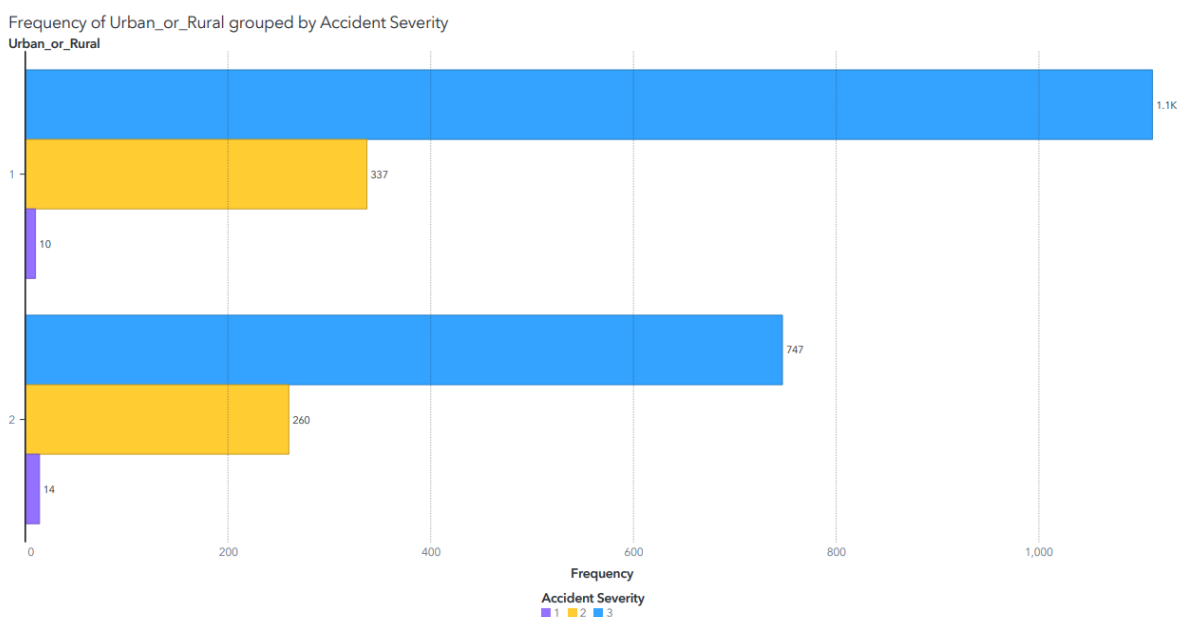
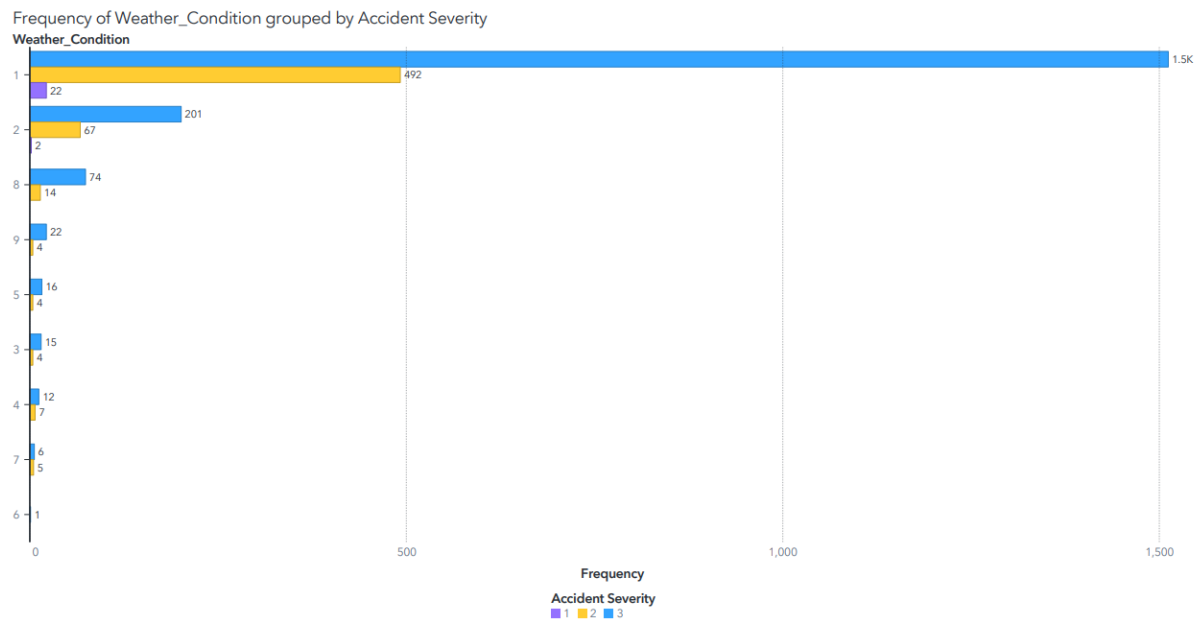


Image 7

The graph illustrates the frequency of road accidents by severity (1: Fatal, 2: Serious, 3: Slight) in urban and rural areas. Urban areas have a higher number of accidents, predominantly slight, while rural areas, despite fewer accidents, show a higher proportion of fatal incidents. This indicates that rural accidents are generally more severe compared to urban ones.

Accident Frequency by Weather Condition:

Page 22



21

Image 8

The graph displays the frequency of accidents by severity (1: Fatal, 2: Serious, 3: Slight) across weather conditions. Most accidents occur in fine weather, dominated by slight accidents, followed by serious and fatal. Light rain also contributes significantly, primarily to slight and serious accidents. While fine weather sees the highest accident frequency due to greater traffic volume, weather conditions do not strongly influence accident severity.



## Accident Severity Distribution across Surrey:

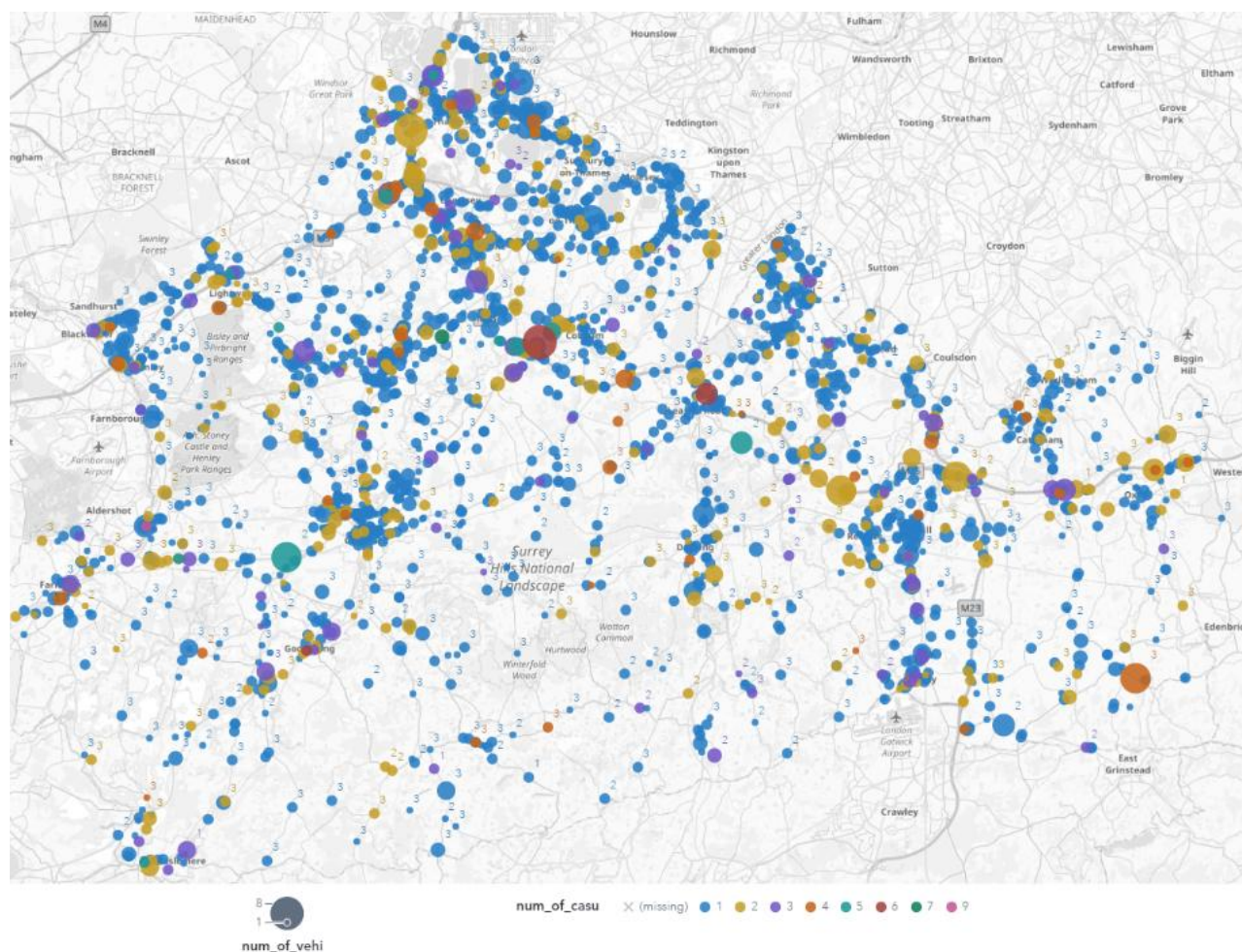


Image 9

The map visualizes the geographical distribution of accidents across Surrey, with severity levels (1: Fatal, 2: Serious, 3: Slight) labeled for each point. The color of the points indicates the number of casualties involved, while the size of the circles represents the number of vehicles involved. This provides a clear spatial understanding of accident hotspots and their severity, along with the impact in terms of vehicles and casualties involved.

## Data Cleaning:

The data cleaning process resolved missing values and inconsistencies, ensuring readiness for analysis. Missing values in Accident Severity and other columns (e.g., longitude, latitude, Number of Vehicles, and Number of Casualties) were addressed using SAS Viya's "Replace Missing Values" feature via code and flows. Numerical variables were imputed with the median, and categorical variables with the mode. This process produced a complete and consistent dataset, ready for predictive modelling and analysis.

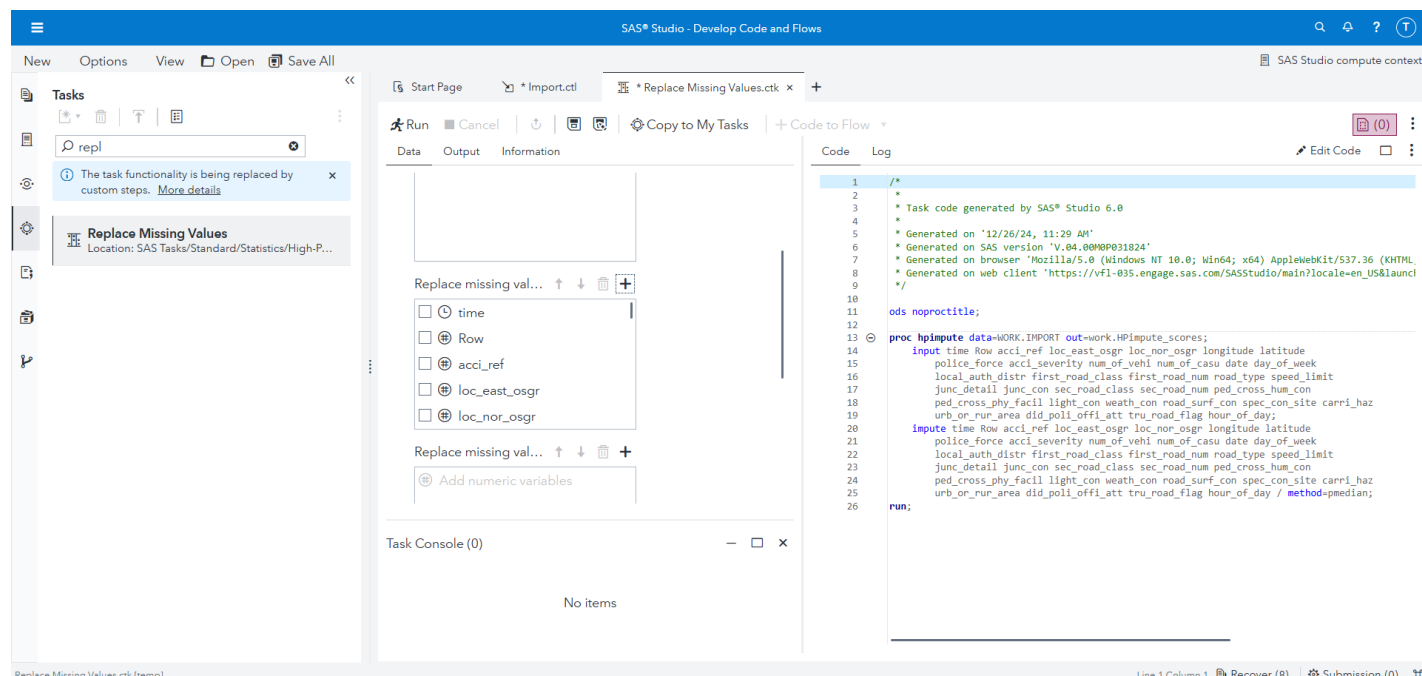


Image 10

The successful execution of the "Replace Missing Values" task in SAS Viya, as shown in the accompanying screenshot, was followed by a detailed output table summarizing the imputed values and their respective imputation methods. This thorough cleaning process ensures the dataset is complete and retains the reliability essential for accurate predictive modeling.

Performance Information	
Execution Mode	Single-Machine
Number of Threads	4

Data Access Information			
Data	Engine	Role	Path
WORK.IMPORT	V9	Input	On Client
WORK.HPIMPUTE_SCORES0001	V9	Output	On Client

Imputation Results					
Variable	Imputation Indicator	Imputed Variable	N Missing	Type of Imputation	Imputation Value (Seed)
time	M_time	IM_time	0	Pseudo Median	51680
Row	M_Row	IM_Row	0	Pseudo Median	1240
acci_ref	M_acci_ref	IM_acci_ref	0	Pseudo Median	451068264
loc_east_osgr	M_loc_east_osgr	IM_loc_east_osgr	1	Pseudo Median	507080
loc_nor_osgr	M_loc_nor_osgr	IM_loc_nor_osgr	0	Pseudo Median	158193
longitude	M_longitude	IM_longitude	1	Pseudo Median	-0.46401
latitude	M_latitude	IM_latitude	1	Pseudo Median	51.31257
police_force	M_police_force	IM_police_force	1	Pseudo Median	45.00000
acci_severity	M_acci_severity	IM_acci_severity	1	Pseudo Median	3.00000
num_of_vehi	M_num_of_vehi	IM_num_of_vehi	1	Pseudo Median	2.00000
num_of_casu	M_num_of_casu	IM_num_of_casu	1	Pseudo Median	1.00000
date	M_date	IM_date	1	Pseudo Median	22474
day_of_week	M_day_of_week	IM_day_of_week	1	Pseudo Median	4.00000
local_auth_distr	M_local_auth_distr	IM_local_auth_distr	0	Pseudo Median	-1.00000
first_road_class	M_first_road_class	IM_first_road_class	0	Pseudo Median	4.00000
first_road_num	M_first_road_num	IM_first_road_num	0	Pseudo Median	24.00000
road_type	M_road_type	IM_road_type	0	Pseudo Median	6.00000
speed_limit	M_speed_limit	IM_speed_limit	0	Pseudo Median	30.00000
junc_detail	M_junc_detail	IM_junc_detail	0	Pseudo Median	0
junc_con	M_junc_con	IM_junc_con	0	Pseudo Median	-1.00000
sec_road_class	M_sec_road_class	IM_sec_road_class	0	Pseudo Median	0
sec_road_num	M_sec_road_num	IM_sec_road_num	0	Pseudo Median	-1.00000
ped_cross_hum_con	M_ped_cross_hum_con	IM_ped_cross_hum_con	0	Pseudo Median	0
ped_cross_phy_facil	M_ped_cross_phy_facil	IM_ped_cross_phy_facil	0	Pseudo Median	0
light_con	M_light_con	IM_light_con	0	Pseudo Median	1.00000
weath_con	M_weath_con	IM_weath_con	0	Pseudo Median	1.00000
road_surf_con	M_road_surf_con	IM_road_surf_con	0	Pseudo Median	1.00000
spec_con_site	M_spec_con_site	IM_spec_con_site	0	Pseudo Median	0
carri_haz	M_carri_haz	IM_carri_haz	0	Pseudo Median	0
urb_or_rur_area	M_urb_or_rur_area	IM_urb_or_rur_area	0	Pseudo Median	1.00000
did_poli_offi_att	M_did_poli_offi_att	IM_did_poli_offi_att	0	Pseudo Median	1.00000
tru_road_flag	M_tru_road_flag	IM_tru_road_flag	0	Pseudo Median	2.00000
hour_of_day	M_hour_of_day	IM_hour_of_day	0	Pseudo Median	14.00000

Table 3

## Task 2 – Predicting Accident Severity [30 marks]

You will apply machine learning techniques to predict accident severity using the dataset

The goal of this task is to develop machine learning models to predict accident severity based on the provided dataset.

### Data Balancing

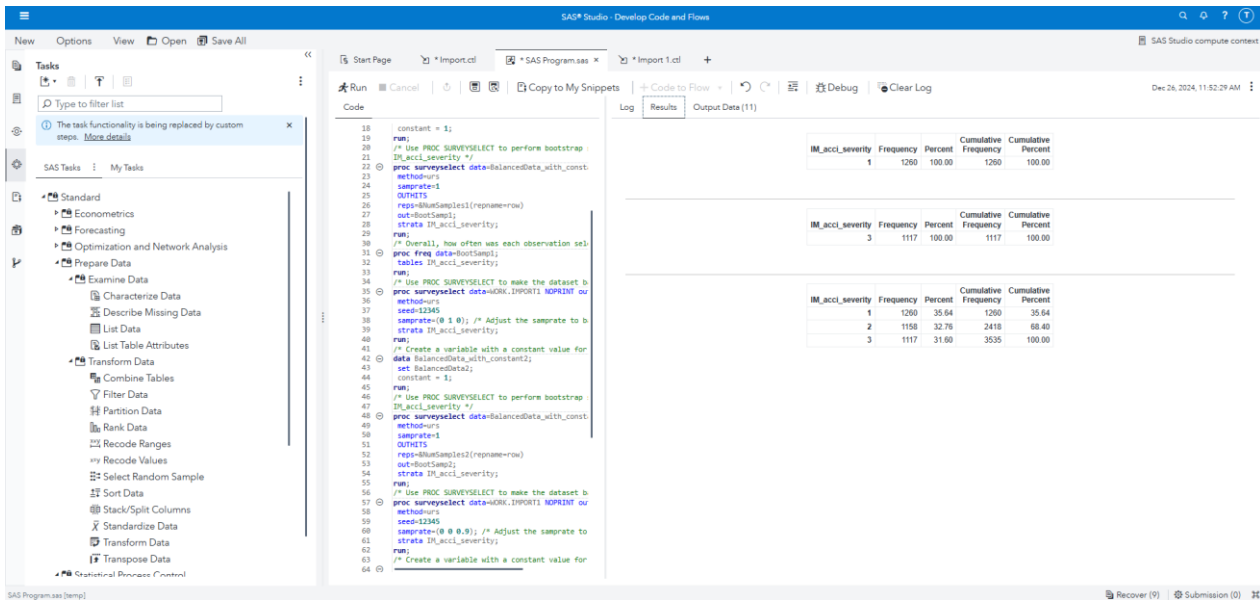


Image 11

Given the significant imbalance in the target variable, with fatal accidents comprising only 1% of the data, the professor-provided SAS code was used to balance the dataset after imputing missing values. This ensured a fair representation of all classes in the target variable. Using the balanced dataset, multiple predictive models were built and evaluated to identify the most effective approach for classifying accident severity and understanding the key factors influencing severe outcomes.

### Machine Learning Pipeline

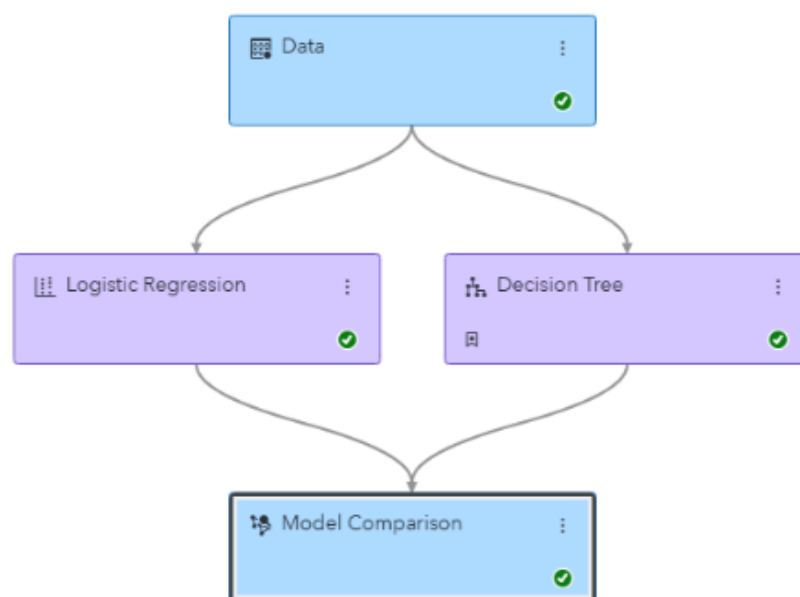


Image12

This SAS pipeline outlines the machine learning workflow for predicting accident severity using the balanced dataset. The process starts with input data containing pre-processed and balanced data, followed by training three supervised learning models: Logistic Regression and Decision Tree.

## Feature Selection

Feature selection was performed using variable importance scores from the Decision Tree model. This method measures each variable's contribution to splitting nodes and improving model performance. High-priority features like the hour of the day and Local Authority District were identified as the most strongly associated with accident severity.

This approach ensures the inclusion of only the most relevant features, reducing complexity, enhancing interpretability, and minimizing overfitting. As a result, the model is better equipped to generalize to unseen data while maintaining accuracy and efficiency.

The variable importance chart ranks features based on their impact on predicting accident severity. The most influential variables include hour of the day, Local Authority District, junction details, and first road number. Other features, such as longitude, number of casualties, and weather conditions, contribute significantly but with relatively lower importance. These insights underline the relationship between these variables and accident severity, aiding in model refinement.

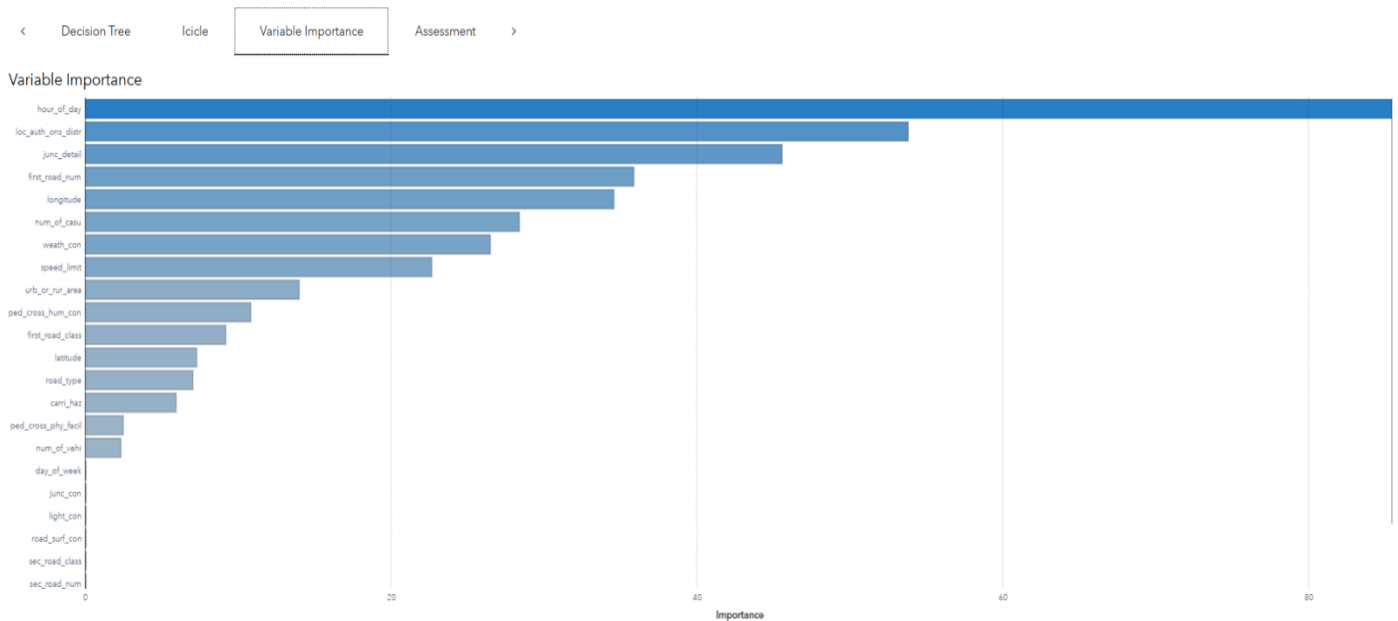


Image 13



Logistic Regression:

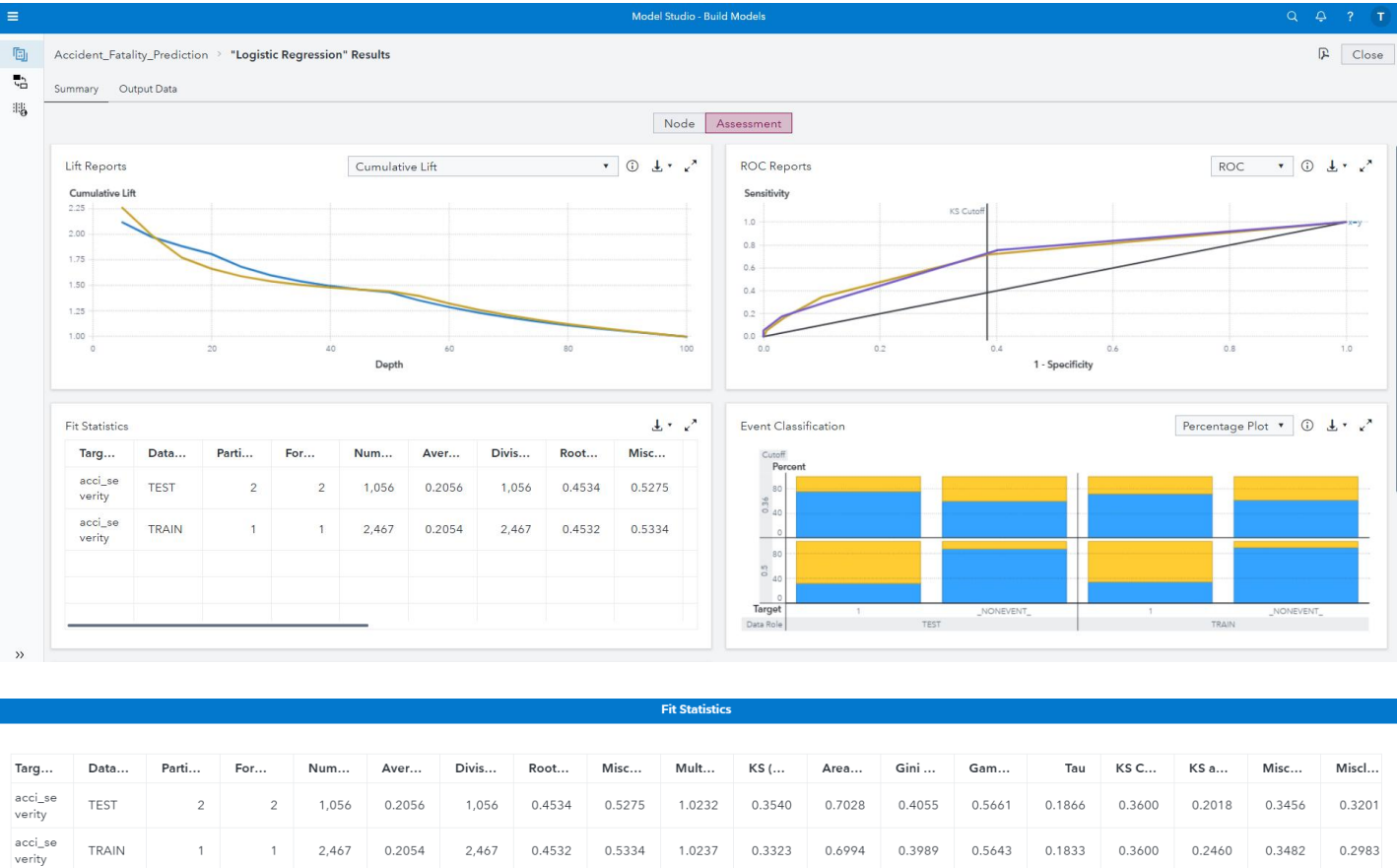


Image 14

The logistic regression model was evaluated on training and testing datasets, yielding the following observations:

- Model Performance Metrics:**
  - The AUC was 0.6994 (training) and 0.7028 (testing), indicating moderate discriminatory power.
  - Misclassification rates of 53.34% (training) and 52.75% (testing) highlight challenges in predicting severity, particularly for minority classes like fatal accidents.
  - KS statistic values of 0.3323 (training) and 0.3540 (testing) show reasonable separation between predicted probabilities and actual classes.
  - A lift value of 1.023 demonstrates limited ability to prioritize severe cases effectively.
- Confusion Matrix Insights:**



The model performs well for the dominant class (slight accidents) but struggles with less frequent classes (serious and fatal accidents), reflected in high false negative rates for minority classes.
- Strengths:**
  - Consistency:** Similar metrics across training and testing datasets indicate good generalization.
  - Interpretability:** Provides clear insights into predictors of accident severity.
  - Baseline Performance:** Serves as a foundation for comparing advanced models.
- Weaknesses:**

- **Class Imbalance:** Struggles with minority classes, reducing recall and precision for severe accidents.
- **High Misclassification:** Over 50% misclassification highlights difficulty in accurate predictions.
- **Limited Predictive Power:** Moderate KS and lift values indicate room for improvement.

This analysis highlights the model's strengths as a baseline and its limitations, particularly in addressing class imbalance and predictive power for severe cases.

5.Variable Importance:

Selection Summary



Step	Effect Entered	Effect Removed	Number of Effects	SBC	Optimal SBC
0	Intercept		1	5,422.2064	0
1	num_of_vehi		2	5,143.4869	0
2	loc_auth_ons_distr		3	4,957.5827	0

Table 4

The stepwise logistic regression analysis identified two key predictors of accident severity:

1. **Number of Vehicles Involved:**  
This was the most influential feature, strongly associated with greater severity due to the increased likelihood of multiple casualties or more severe outcomes.
2. **Local Authority District:**  
This feature highlights geographical differences in severity, driven by variations in road conditions, traffic density, and enforcement of traffic regulations across districts.

These variables significantly improve the model's ability to distinguish severity levels, emphasizing their importance in predictive analysis.

Conclusion:

The logistic regression model serves as a baseline for predicting accident severity but is limited by class imbalance and its reliance on linear assumptions. To improve performance, advanced models like Random Forests or Gradient Boosting should be considered. Rebalancing techniques, such as oversampling or under sampling, are also recommended to enhance recall and precision for severe accident classes.

## Decision Trees:

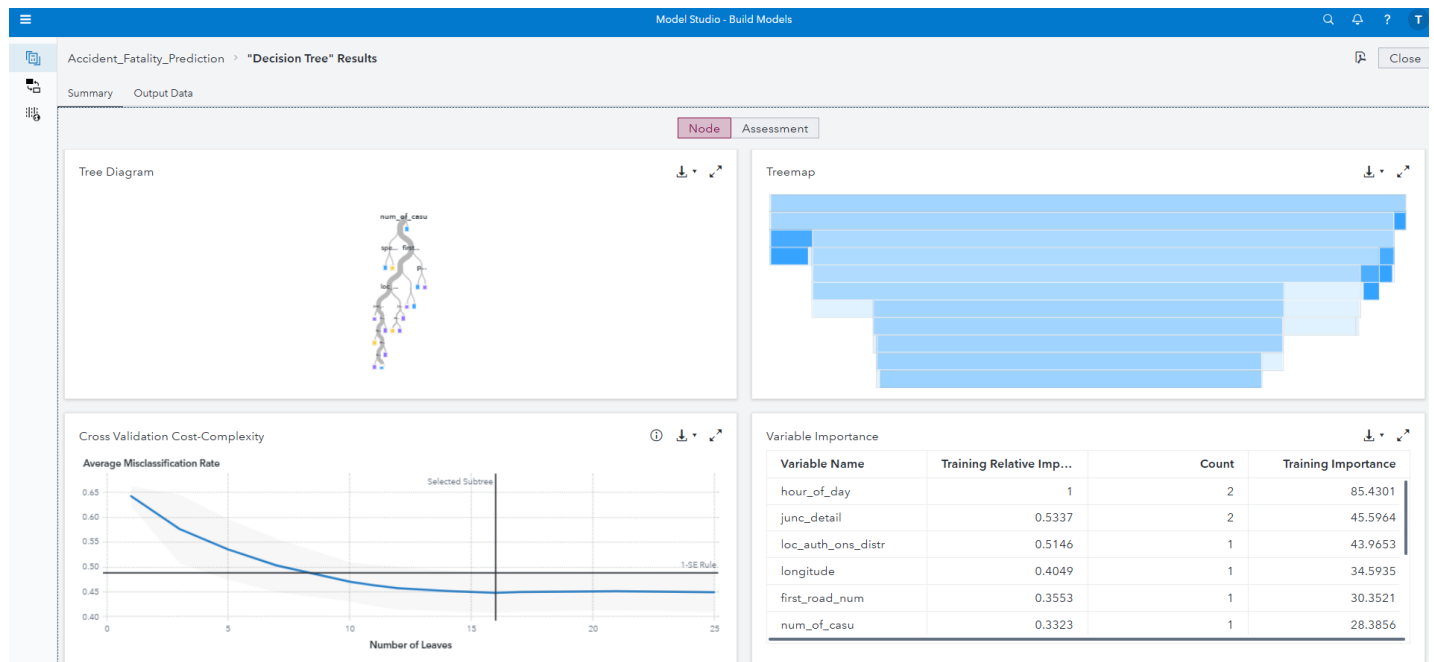


Image 15

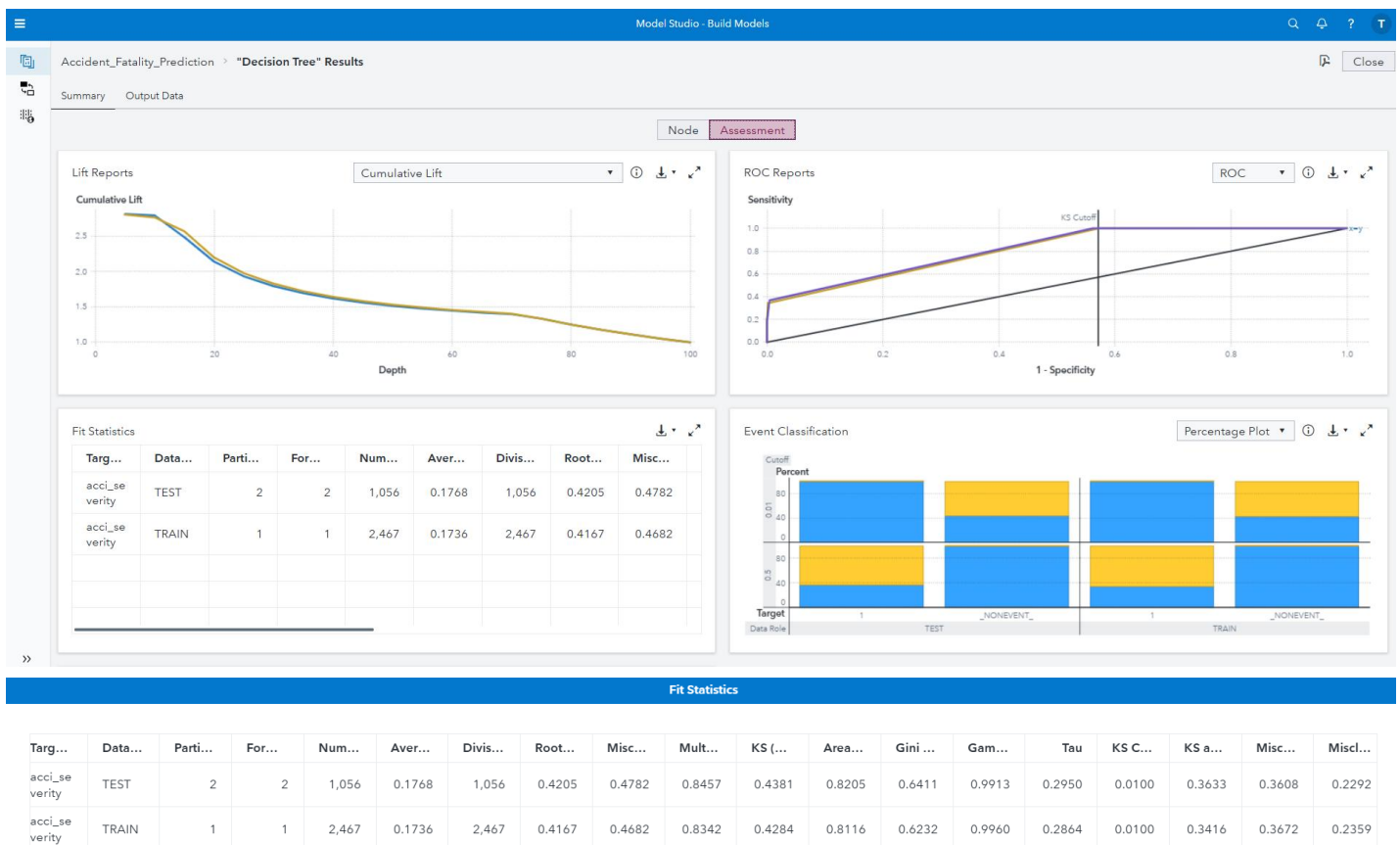


Image 16

The decision tree model was evaluated using training and testing datasets, with the following observations:

### 1. Model Performance Metrics:



- Achieved an AUC of 0.8116 (training) and 0.8205 (testing), showing better discriminatory power than logistic regression.
- Misclassification rates of 46.82% (training) and 47.82% (testing) indicate improved accuracy but challenges in predicting minority classes persist.
- KS statistic values (0.4284 training, 0.4381 testing) highlight stronger separation of predicted probabilities.
- Lift value of 0.8457 demonstrates the model's ability to prioritize severe cases, though improvements are needed.

## 2. **Confusion Matrix Insights:**

The model predicts slight accidents well but struggles with serious and fatal accidents. Improved handling of minority classes is evident compared to logistic regression due to a lower misclassification rate.

## 3. **Strengths:**

- **Non-Linear Relationships:** Effectively captures interactions and improves accuracy.
- **Feature Prioritization:** Identifies critical predictors for targeted interventions.
- **Interpretability:** Tree visualization aids understanding of decision-making.
- **Improved Accuracy:** Outperforms logistic regression in misclassification rates and AUC.

## 4. **Weaknesses:**

- **Class Imbalance:** Struggles with accurate predictions for minority classes.
- **Overfitting Risk:** Prone to overfitting despite good generalization.
- **Complexity:** Large trees may hinder interpretation and implementation.

## 5. **Variable Importance:**

Key predictors include:

- **Number of Casualties:** Strong correlation with accident severity.
- **Hour of the Day:** Higher severity at nighttime or early morning.
- **Speed Limit:** Severe accidents increase with higher speed limits.
- **Local Authority District:** Reflects geographic variations in road conditions.
- **Junction Details and Road Type:** Influence severity based on accident location.

## **Conclusion:**

The decision tree model outperforms logistic regression, effectively capturing non-linear relationships and prioritizing key features. While class imbalance and tree complexity remain challenges, ensemble methods like Random Forests or Gradient Boosting and rebalancing techniques can further improve performance.

### **Recommendations for Improving Road Safety:**

Based on the analysis of accident severity predictors and model results, the following recommendations are proposed:

#### **1. Implement Time-Sensitive Interventions**

- **Peak Hour Monitoring:** Increase traffic enforcement during critical peak hours to reduce accident frequency.
- **Dynamic Traffic Management:** Leverage real-time traffic data to optimize signal timings and manage congestion effectively during high-risk periods.

#### **2. Improve Road Design and Infrastructure**

- **High-Risk Junctions:** Enhance junction safety with better signage, traffic signals, and roundabouts to reduce accidents at prone locations.
- **Spatial Hotspots:** Install safety measures like speed cameras and reflective markers in accident-prone areas identified through geographic analysis.

#### **3. Enhance Speed Management**

- **Speed Limit Enforcement:** Strengthen speed monitoring in high-risk zones to minimize speed-related accidents.
- **Variable Speed Limits:** Introduce adaptive speed limits during adverse weather or high traffic volumes to lower accident risks.

#### **4. Address Environmental Risks**

- **Weather-Specific Measures:** Deploy warning systems for adverse weather and ensure proper road maintenance to prevent skidding.
- **Improved Drainage:** Invest in drainage systems to address waterlogging, reducing rain-related accidents.

#### **5. Leverage Data for Proactive Safety**

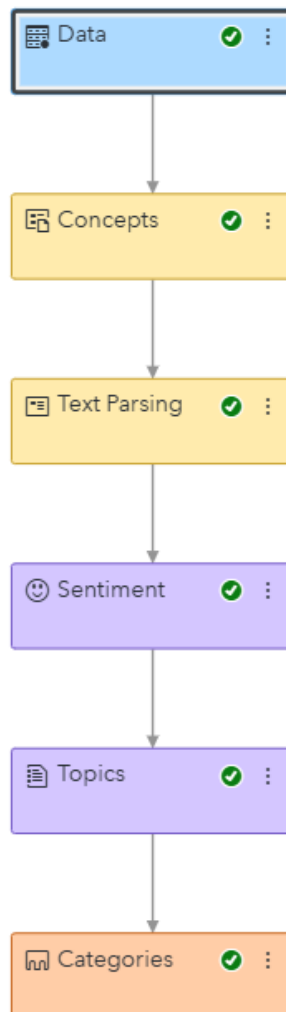
- **Real-Time Monitoring:** Use predictive models to identify emerging risks and deploy resources pre-emptively.
- **Collaborate with Local Authorities:** Partner with governments to address high-risk areas through targeted infrastructure improvements.

These targeted actions can address key factors contributing to accidents, enhance safety measures, and reduce accident severity effectively.

**Task 3 – Text Analysis of Tweets [20 marks]**

You will need to summarizing key insights and findings from your text analysis

The dataset, containing tweets about road accidents in Surrey, was loaded into SAS Viya for text analysis. The exploration focused on understanding its composition by reviewing variations in tone, language, and content, including descriptive accident reports, opinions, and general commentary..



**Image 17**

The **Concepts Node** in SAS Viya was utilized to extract and classify both predefined and custom concepts from the tweets dataset, enabling structured analysis of unstructured text.

**Key Highlights:****1. Predefined Concepts:**

- Extracted entities like dates (nlpDate), locations (nlpPlace), times (nlpTime), and organizations (nlpOrganization).
- These categories provide critical details about accident locations, times, and affected organizations.

**2. Custom Concepts:**

- Concepts such as *Cities*, *Time*, *Congestion*, *Direction*, *Highway*, *Junctions*, *Road*, and *Lanes* were created to capture terms specific to cities (e.g., "Woking," "Guildford") and road features (e.g., "lane 1").
- These concepts enhance domain-specific insights relevant to road accidents.

**3. Matched Data:**

- Of the 598 tweets analyzed, 274 matched the *Cities* concept, frequently mentioning cities involved in accidents.
- Highlighted matches provide context on the usage and relevance of specific concepts in the dataset.

This process was essential for identifying structured patterns and extracting meaningful insights from the text data.

The screenshot shows the 'Model Studio - Build Models' interface. On the left, under 'Concepts', there are 'Predefined Concepts (9)' including nlpDate, nlpMeasure, nlpMoney, nlpNounGroup, nlpOrganization, nlpPercent, nlpPerson, nlpPlace, and nlpTime. There are also 'Custom Concepts (2)' including Road\_lane and \_Cities\_. The main area is titled 'Edit Concept' and 'Sandbox'. It shows a list of documents with a search bar and a table of matches. The table has columns for 'Text' and 'Fact ...'. The 'Text' column contains snippets of tweets mentioning locations like Woking, Farnham, and Guildford. The 'Fact ...' column shows a count of matches for each concept.

Image 18

## Text Preprocessing

The screenshot shows the 'Model Studio - Build Models' interface. On the left, under 'Text Parsing - Manage Terms', there are 'Kept Terms (536)' and 'Dropped Terms (2795)'. The main area shows a table of terms with columns for 'Term', 'Role', 'Documents', and 'Frequency'. The 'Kept Terms' table lists terms like collision, closed, lane, m25, road, due, and surrey. The 'Dropped Terms' table lists terms like the, be, a, and, to, in, and on. The bottom section shows a list of documents with a search bar and a table of matches. The table has columns for 'Text' and 'Fact ...'. The 'Text' column contains snippets of tweets mentioning locations like M25, Surrey, and Guildford.

Image 19

Text preprocessing was conducted using the **Text Parsing** node in SAS Viya to improve dataset quality and consistency. Key steps included:

1. **Removing Special Characters and Punctuation:** Eliminated symbols and noise, retaining meaningful text content.
2. **Tokenization:** Split text into individual words or phrases, enabling word-level analysis and facilitating frequency and sentiment analysis.
3. **Handling Start and Stop Words:** Removed common words (e.g., "and," "the") to reduce redundancy and focus on significant terms.

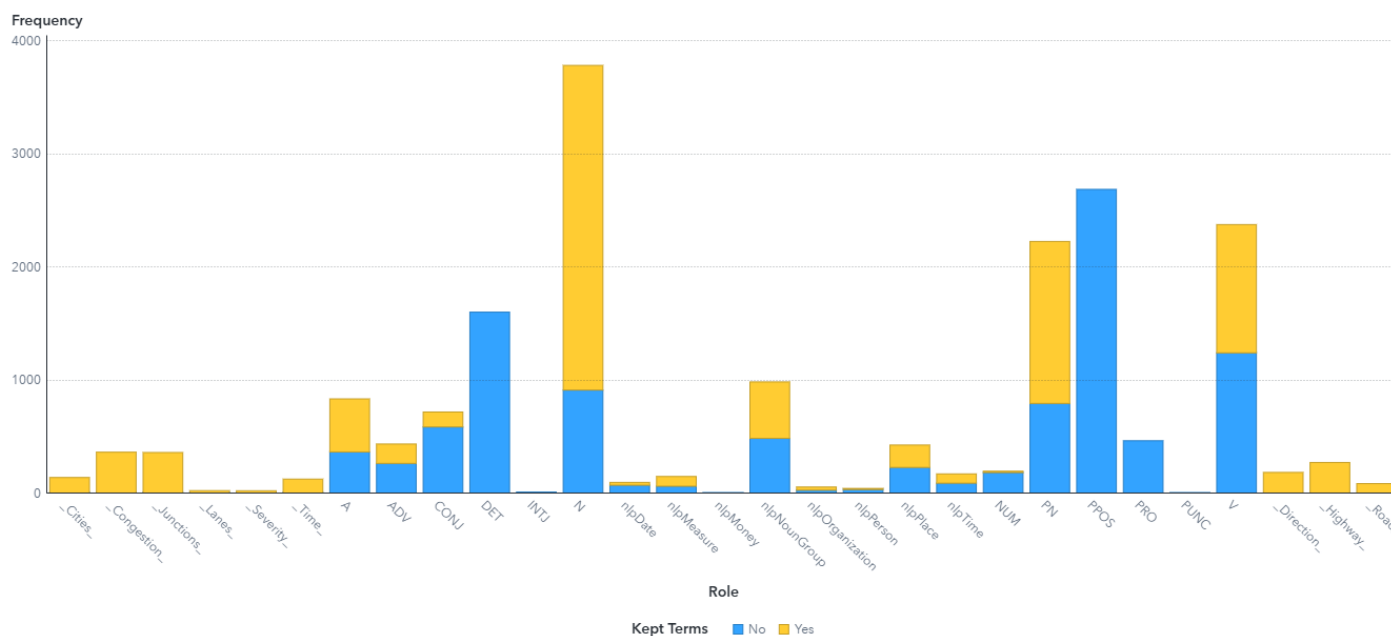


Image 20

The Text Parsing node streamlined preprocessing, creating a clean and standardized dataset for analysis. This automated approach ensured efficiency and consistency, providing a solid foundation for advanced text analytics.

## Exploratory Analysis

Exploratory text analysis uncovered patterns and key themes in the dataset using word clouds for keywords, concepts, and text frequency. Key findings include:



Image 21

Model Studio - Build Models

tweets3 > Text Parsing - Manage Terms

Run node

Close

Kept Terms (536)

Term	Role	Documents	Frequency
<input type="checkbox"/> collision	N	486	507
<input type="checkbox"/> closed	_Congestion_	180	224
<input type="checkbox"/> ▶ lane	N	167	212
<input type="checkbox"/> m25	_Highway_	163	175
<input type="checkbox"/> ▶ road	N	120	135
<input type="checkbox"/> due	CONJ	131	132
<input type="checkbox"/> surrey	PN	112	126
<input type="checkbox"/> ▶ car	N	113	120
<input type="checkbox"/> road	PN	79	107

Dropped Terms (2795)

Term	Role	Documents	Frequency
<input type="checkbox"/> the	DET	359	685
<input type="checkbox"/> ▶ be	V	338	631
<input type="checkbox"/> a	DET	400	607
<input type="checkbox"/> and	CONJ	314	428
<input type="checkbox"/> to	PPOS	312	423
<input type="checkbox"/> ▶ in	PPOS	310	414
<input type="checkbox"/> ▶ on	PPOS	245	331
<input type="checkbox"/> of	PPOS	227	308
<input type="checkbox"/> between	PPOS	235	241

Documents

All (598)

Matched

Text

Anticlockwise #M25 is slow in patches from J11 Chertsey to J9 Leatherhead after a breakdown and collision earlier - all lanes are back open

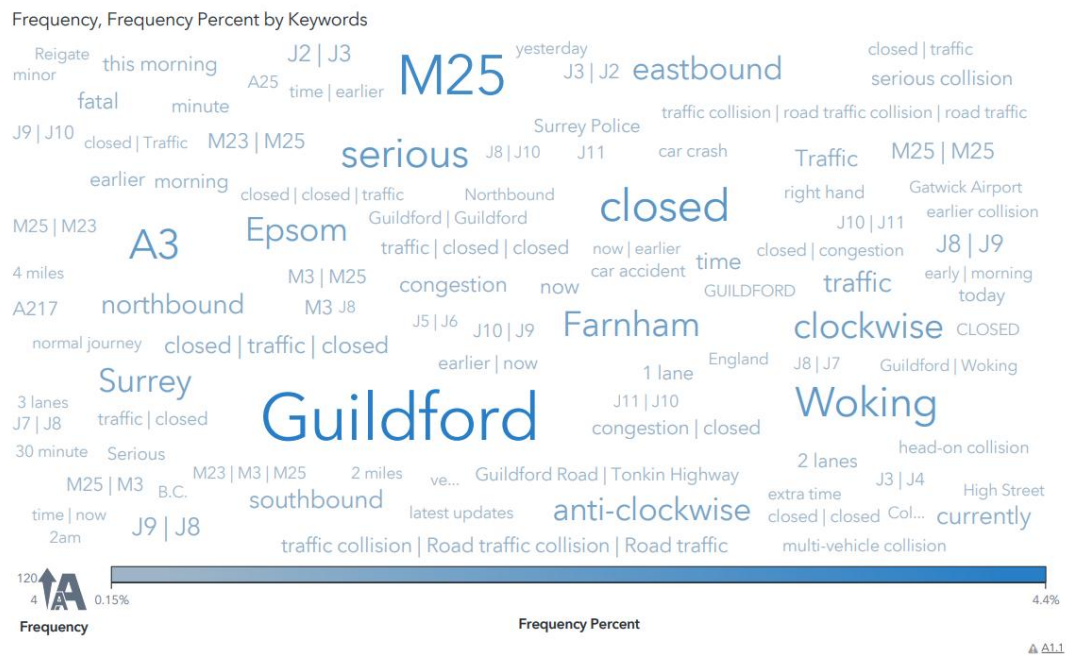
1 lane (of 4) remains closed on the #M23 northbound in #Surrey between J9 (@Gatwick\_Airport) and J8 (#M25) following a collision. Recovery is now taking place. Delays are now less than 10 minutes. Thanks for your patience.

2 lanes (of 4) are closed on the #M25 clockwise in #Surrey between J10 (Cobham) and J11 (Chertsey) due to a collision. Emergency services are in attendance. There's a 45 minute delay on approach with 6.5 miles of congestion.

Document 1 of 598

### Image 22

- **Keyword Analysis:** Frequent terms like "Guildford," "closed," "M25," "Woking," and "serious" emphasize road closures, traffic disruptions, and accident severity in Surrey.



### Image 23

- **Concept Analysis:** Common concepts such as "Congestion," "Direction," "Highway," and "Severity" highlight traffic impact and accident outcomes.

Frequency, Frequency Percent by Concept Name

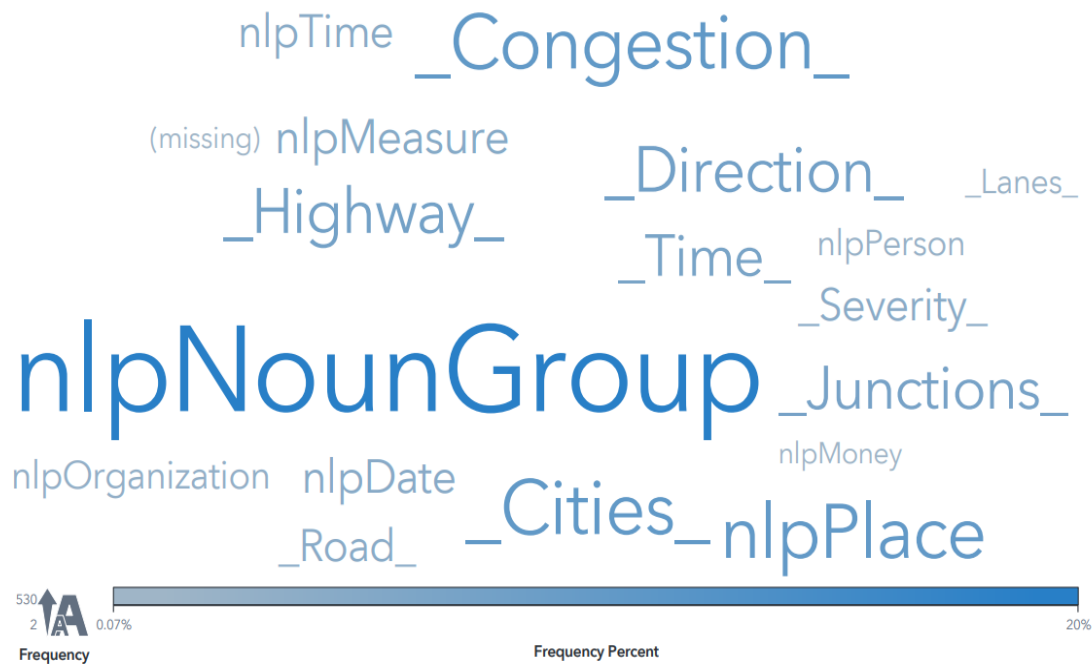


Image 24

- **Text Patterns:** Recurring phrases focus on road closures, serious collisions, and traffic updates, particularly on major highways like the M25 and A3.

Frequency, Frequency Percent by Text

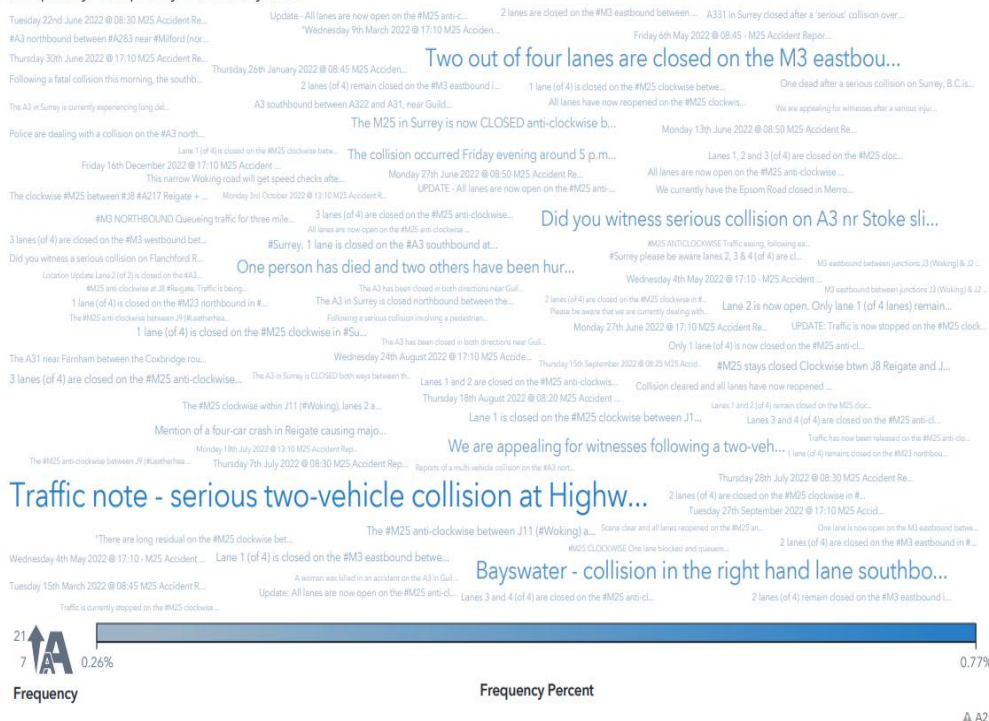


Image 25

Insights

The analysis reveals a focus on high-traffic areas, accident severity, and traffic disruptions, providing a foundation for further sentiment analysis and topic modeling.

Sentiment Analysis

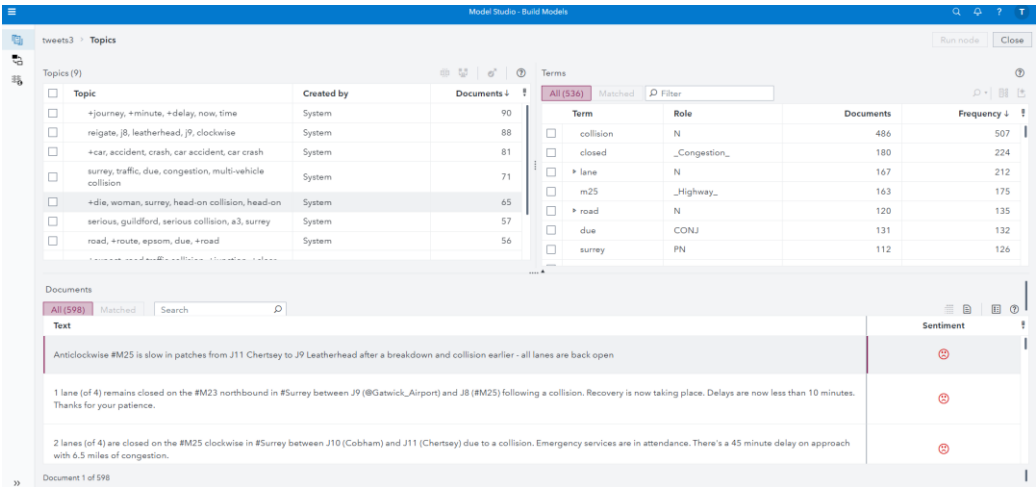


Image 26

Sentiment analysis was performed on the tweets to understand the overall sentiment (positive, negative, or neutral) regarding road accidents and traffic in Surrey. The analysis results are summarized below:

1. Sentiment Distribution

- The majority of tweets exhibit **negative sentiment**, reflecting public concern, frustration, and the serious nature of traffic accidents and road closures.
- A smaller proportion of tweets convey **neutral sentiment**, often factual updates on traffic conditions and closures.
- Few tweets show **positive sentiment**, typically related to safety initiatives or expressions of gratitude.

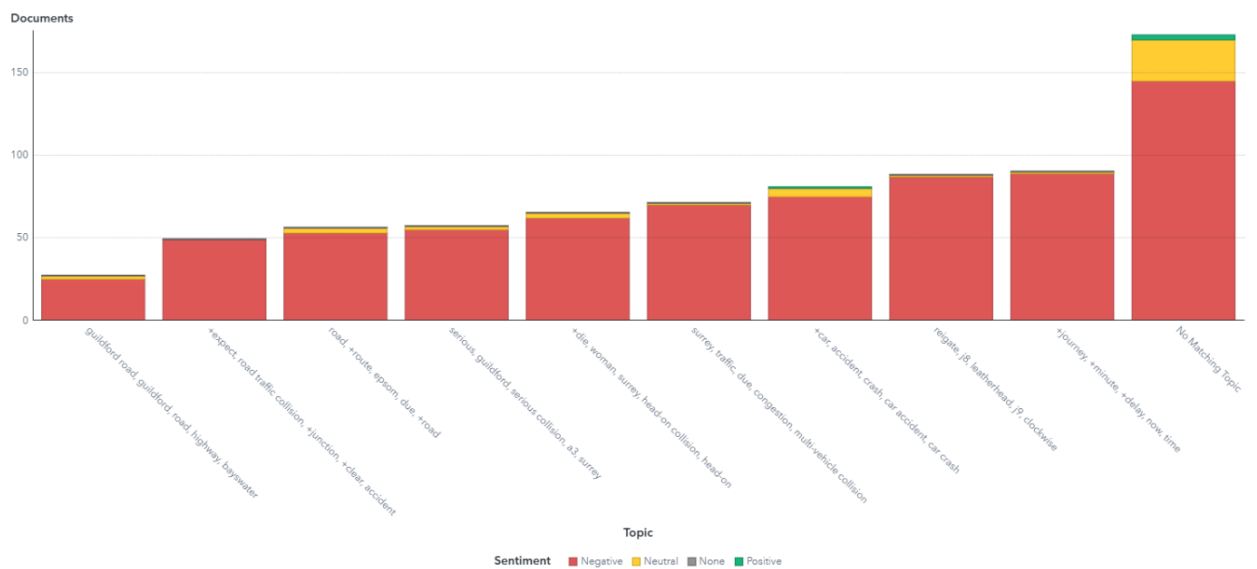
2. Topic-Specific Sentiment

The sentiment was analyzed across key topics identified in the dataset:

- **Accidents and Collisions:** Predominantly negative due to the serious and often fatal nature of incidents.
- **Traffic and Congestion:** Negative sentiment driven by delays and disruptions.

Visual Representation





**Image 27**

The bar chart categorizes sentiment for various topics, illustrating that negative sentiment dominates across most topics. The chart emphasizes areas of public concern, such as severe accidents and significant traffic disruptions.

### Insights

- The overwhelming negativity highlights public frustration and concern surrounding accidents and traffic conditions.
- Positive sentiment, while rare, provides insight into appreciation for safety measures or community support.

### Conclusion

The sentiment analysis reflects public perception of road conditions and traffic incidents in Surrey, with a clear emphasis on negative sentiment. These insights can inform authorities about areas of public dissatisfaction and help shape strategies to improve communication and road safety measures.

## Text Analysis Report

### Key Insights and Findings

1. **Public Concern:** Negative sentiment dominates the dataset, reflecting public dissatisfaction with road conditions, accidents, and congestion in Surrey.
2. **Focus on High-Traffic Areas:** Locations like Guildford and the M25 are frequently mentioned, highlighting their importance in traffic and accident discussions.
3. **Accident Severity:** Discussions emphasize serious and fatal accidents, underlining the need for enhanced road safety measures.
4. **Traffic and Congestion:** Key themes include traffic delays, lane closures, and multi-vehicle collisions, demonstrating the widespread impact of these issues.
5. **Community Engagement:** Positive sentiment, though rare, highlights appreciation for safety initiatives and public responses to accidents.

### Conclusion

The analysis underscores public frustration with accidents and traffic disruptions while emphasizing the need for better communication, safety measures, and infrastructure improvements in high-traffic areas. These findings provide actionable insights for addressing public concerns and improving road safety.

**Task 4 – Decision-Maker's Summary and Recommendations [20 marks]**

Maximum 2 pages including tables, figures (do not use appendix for this task)

This report analyses road accidents in Surrey, UK, during 2021, providing key insights and actionable recommendations to improve road safety and reduce accident severity. The findings are based on a detailed dataset and data-driven analysis, with clear strategies outlined for decision-makers.

**Key Findings****1. Accident Patterns**

- **Severity:** Most accidents are slight, with serious and fatal accidents representing a smaller share. Fatal accidents, while rare, require targeted focus due to their severity.
- **Timing:** Accidents are most frequent during rush hours, correlating with high traffic volumes. Off-peak hours see significantly lower accident frequencies.

**2. High-Risk Locations**

- **Non-Junction Areas:** Account for the highest number of accidents, predominantly slight.
- **Single Carriageways:** Major contributors to accidents, highlighting the need for safety interventions.
- **Hotspot Areas:** Certain local authority districts experience a higher density of accidents, necessitating focused action.

**3. Contributing Factors**

- **Speed:** Most accidents occur in 30 mph zones, but higher-speed areas are associated with greater severity.
- **Weather Conditions:** Poor weather and road surface conditions increase accident risks, emphasizing the importance of infrastructure maintenance.

**Recommendations****1. Time-Based Interventions**

- Deploy traffic enforcement during rush hours to manage congestion and reduce accidents.
- Implement dynamic traffic management systems for real-time signal adjustments and congestion mitigation.

**2. Targeted Infrastructure Improvements**

- Enhance safety at high-risk junctions with improved signage, traffic lights, and roundabouts.
- Address accident hotspots with speed cameras, reflective markers, and better lighting.

**3. Speed Control**

- Strengthen speed enforcement in high-risk zones, especially urban areas.
- Introduce variable speed limits during adverse weather or peak traffic periods.

**4. Public Awareness Campaigns**

- Educate drivers about accident risks during peak hours and in specific road conditions.
- Promote safety campaigns focusing on speeding, distracted driving, and adverse weather.

**5. Data-Driven Safety Measures**

- Use predictive models to monitor risk patterns and deploy resources pre-emptively.
- Collaborate with local authorities to invest in accident-prone areas and improve road conditions.

**Conclusion**

By addressing the identified risk factors and implementing the recommended strategies, Surrey can significantly enhance road safety and reduce accident severity. The insights derived from this analysis provide a roadmap for targeted interventions, effective resource allocation, and long-term safety improvements.

Visualizations and data summaries are included to support these findings, ensuring clarity and actionable insights for decision-makers.

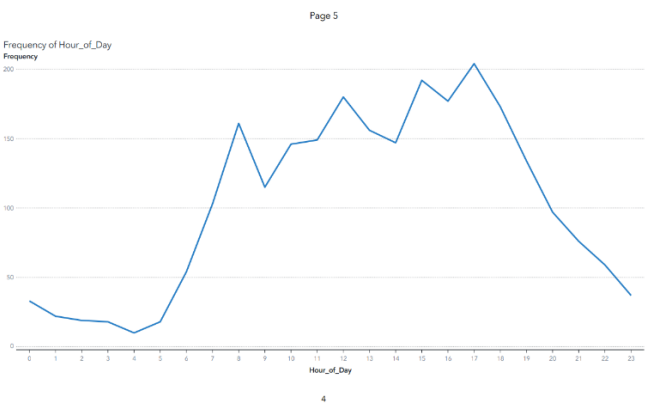
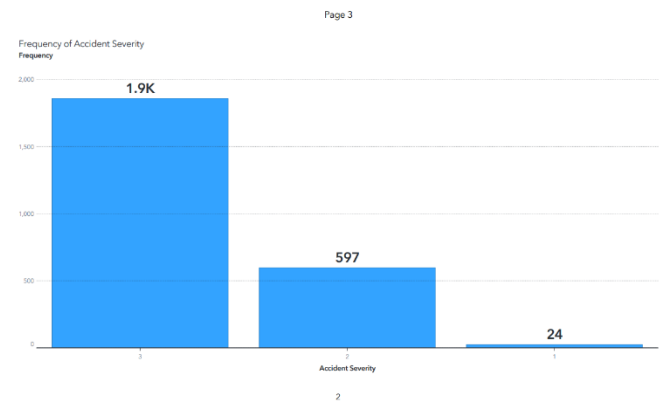


Image 28

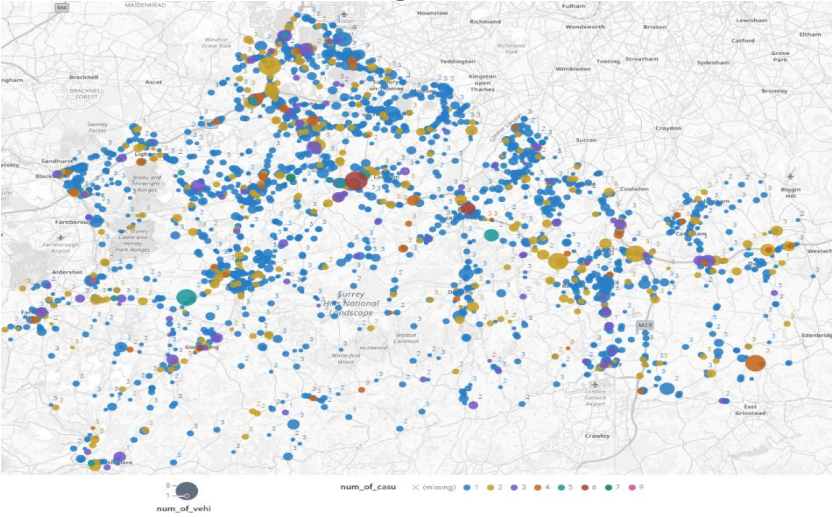


Image 29

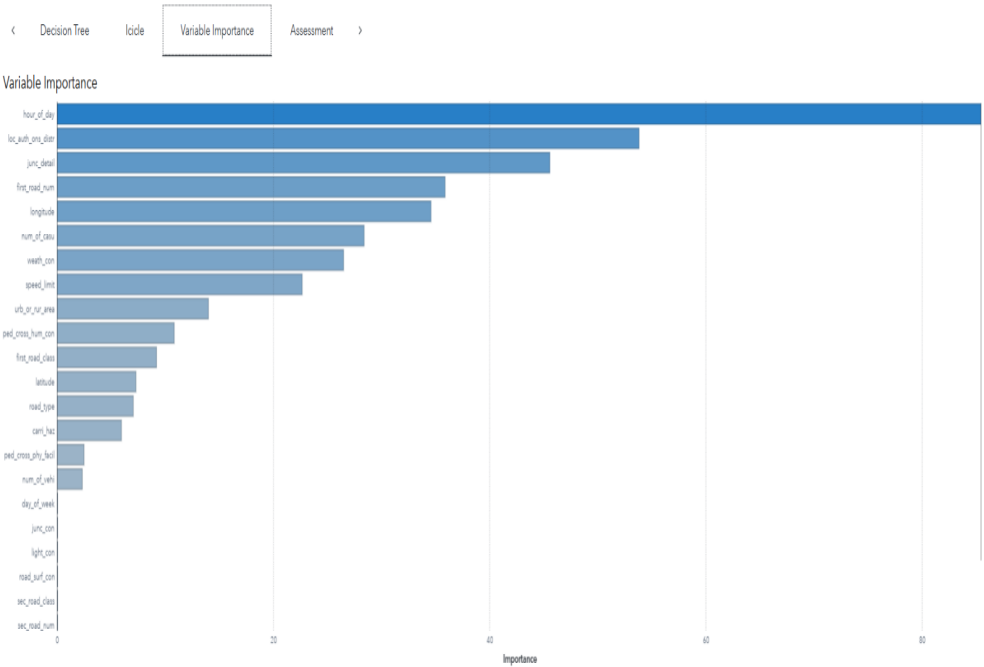
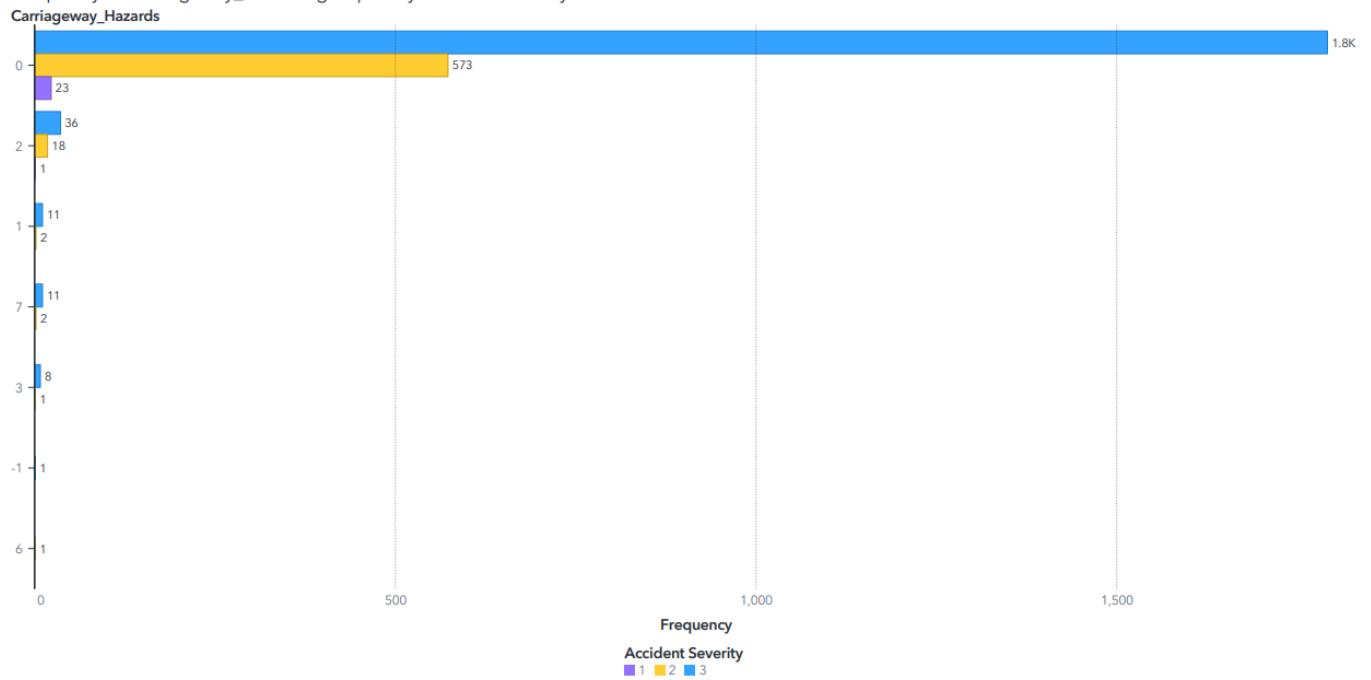


Image 30

## Appendix

Page 4

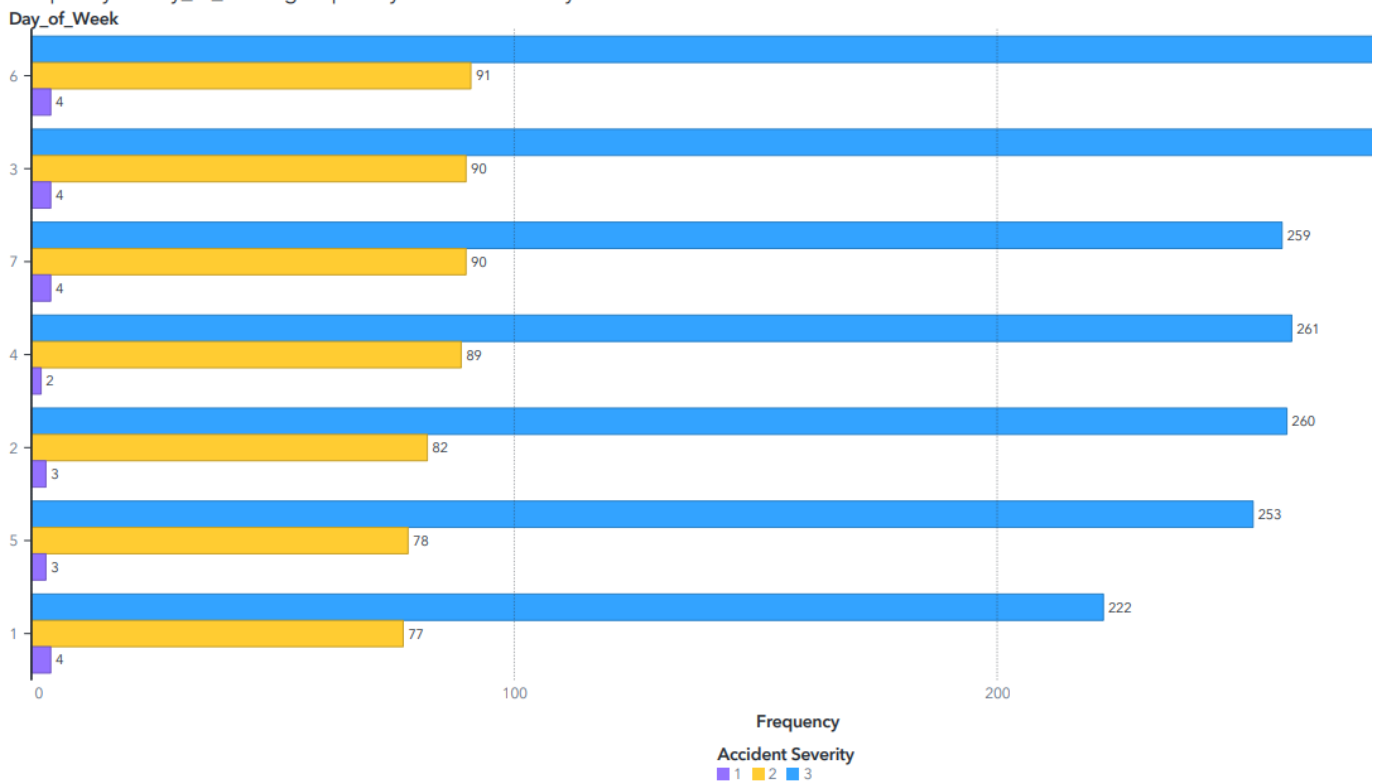
Frequency of Carriageway\_Hazards grouped by Accident Severity



3

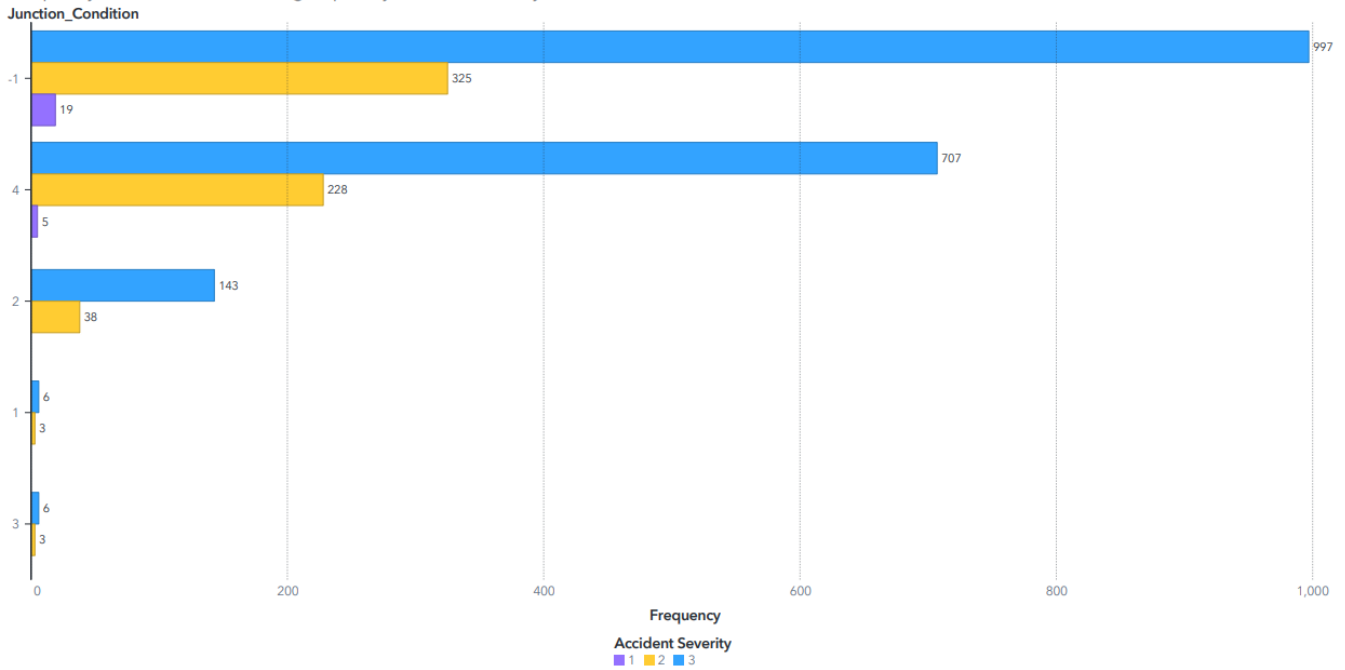
Page 6

Frequency of Day\_of\_Week grouped by Accident Severity



Page 8

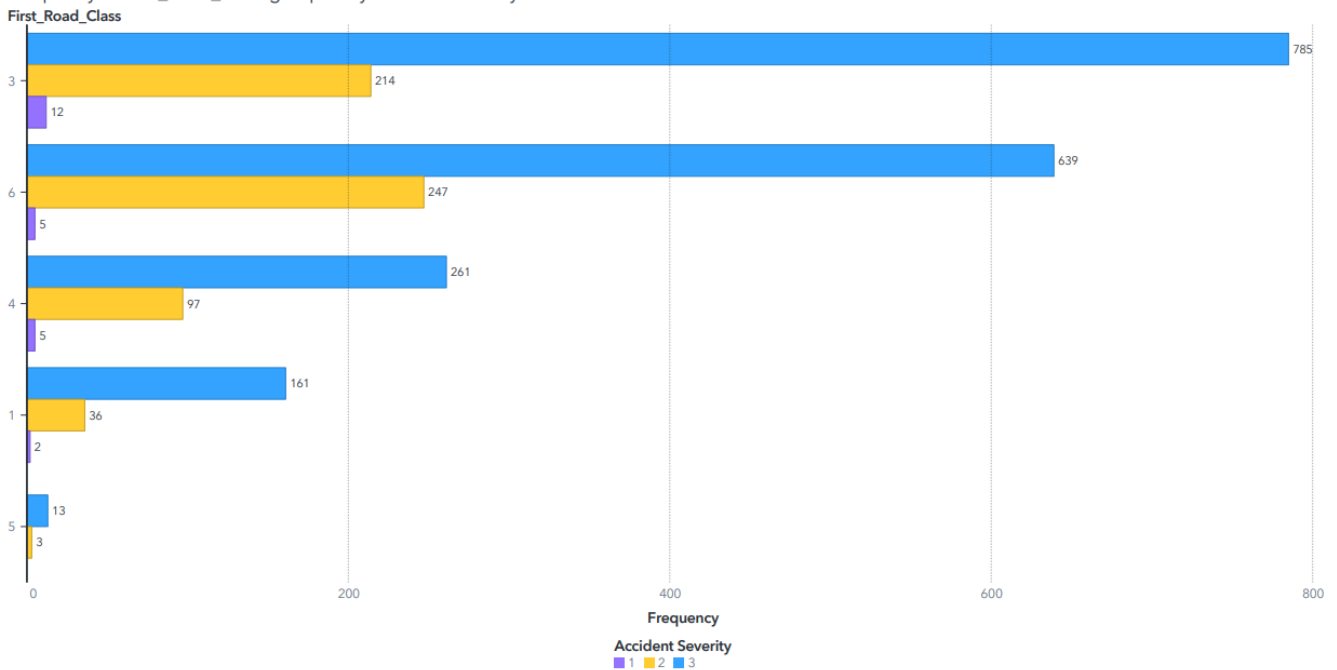
Frequency of Junction\_Condition grouped by Accident Severity



7

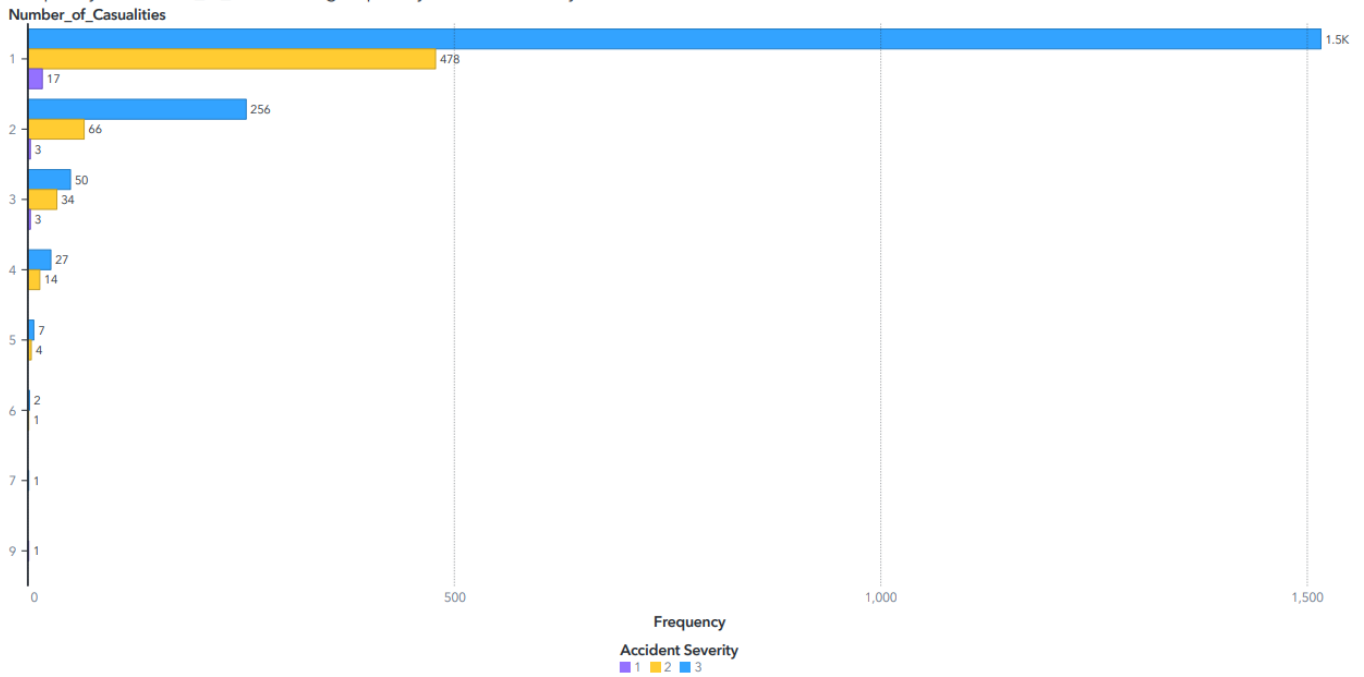
Page 7

Frequency of First\_Road\_Class grouped by Accident Severity

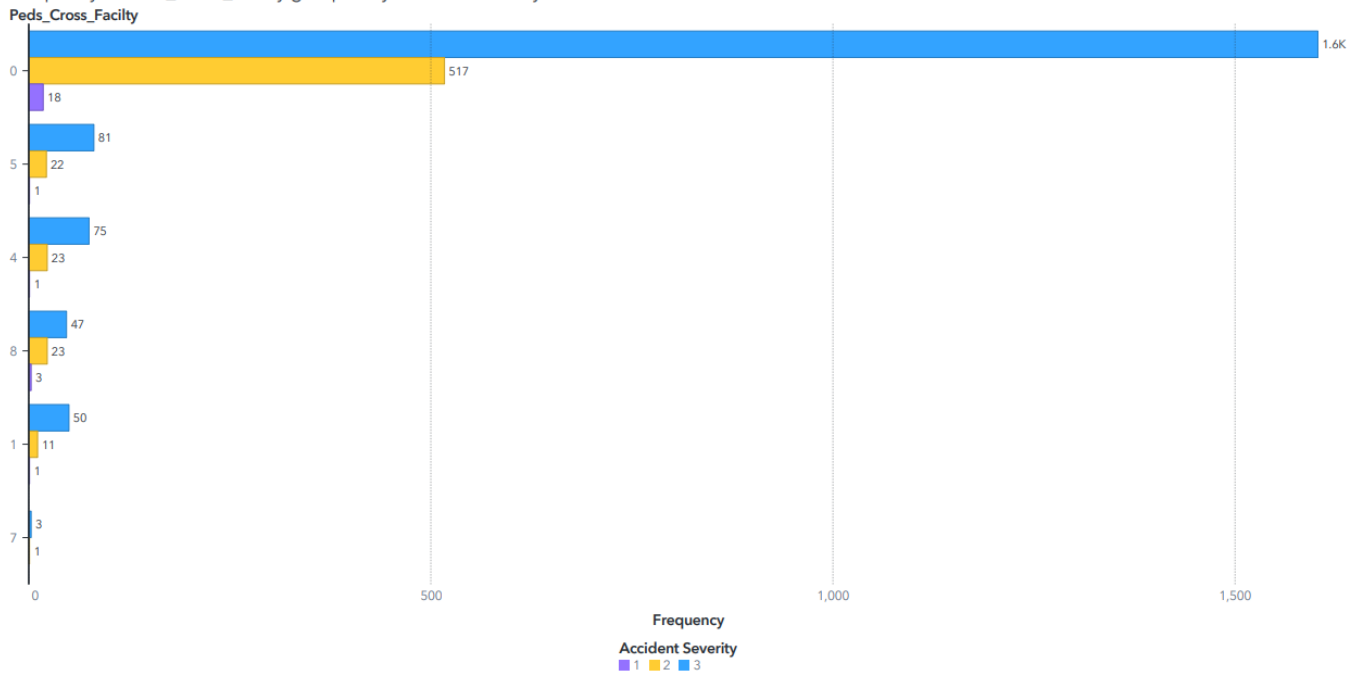


6

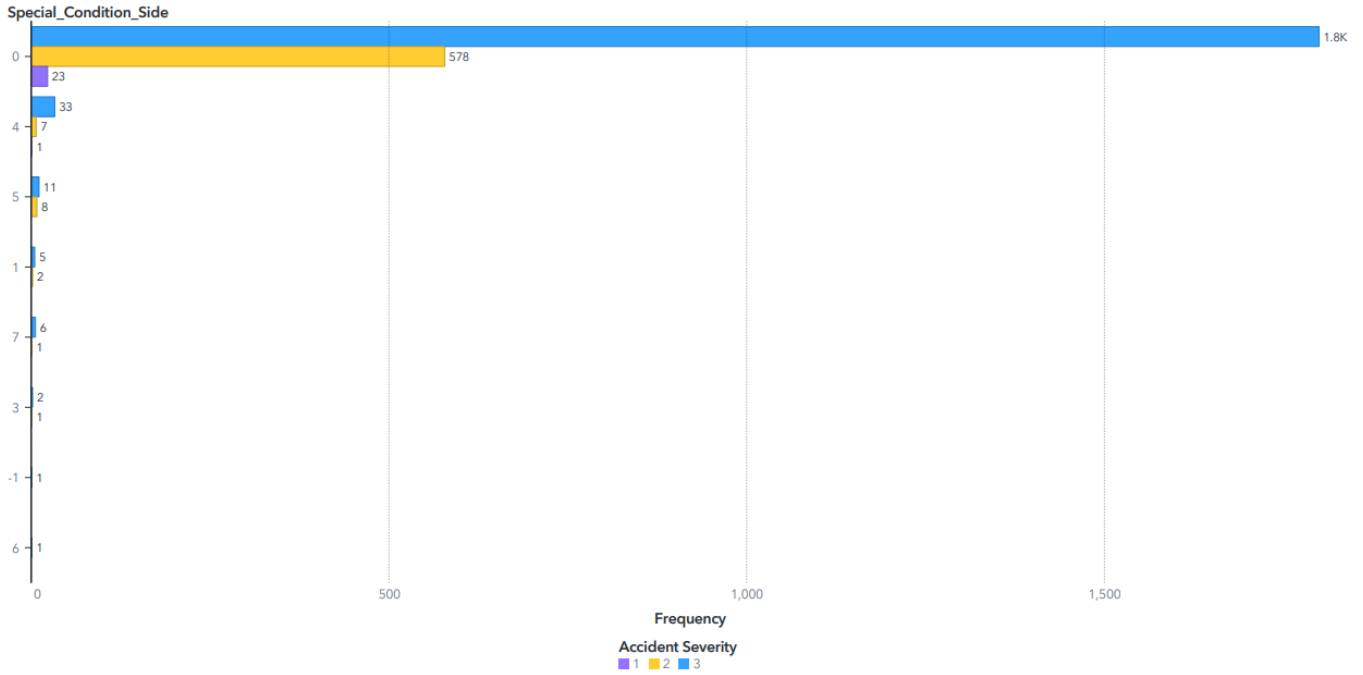
Frequency of Number\_of\_Casualties grouped by Accident Severity



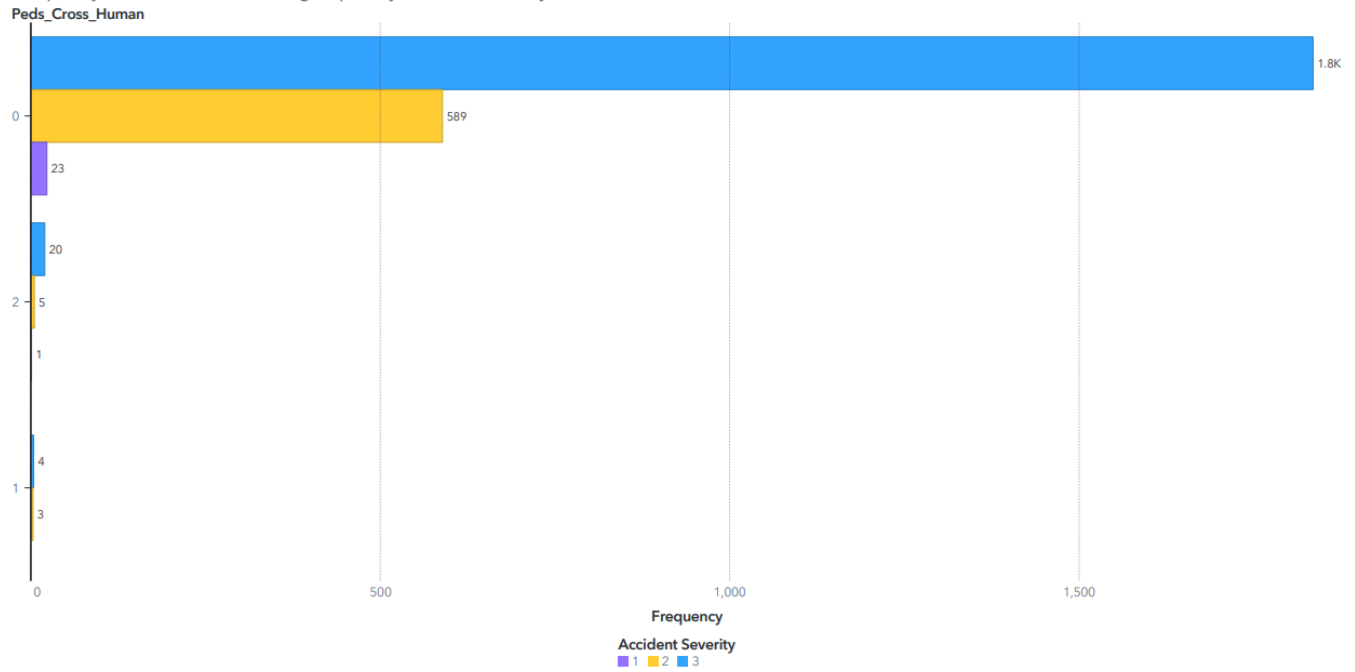
Frequency of Peds\_Cross\_Facility grouped by Accident Severity



Frequency of Special\_Condition\_Side grouped by Accident Severity



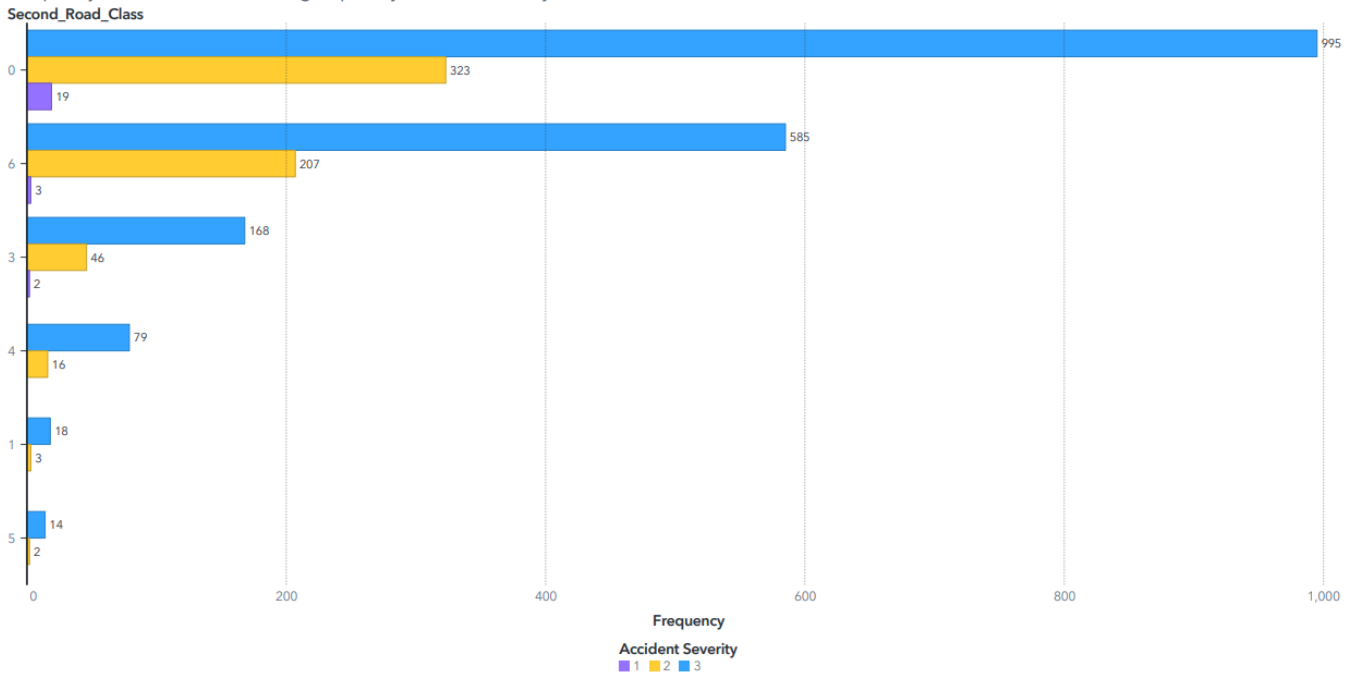
Frequency of Peds\_Cross\_Human grouped by Accident Severity





Page 17

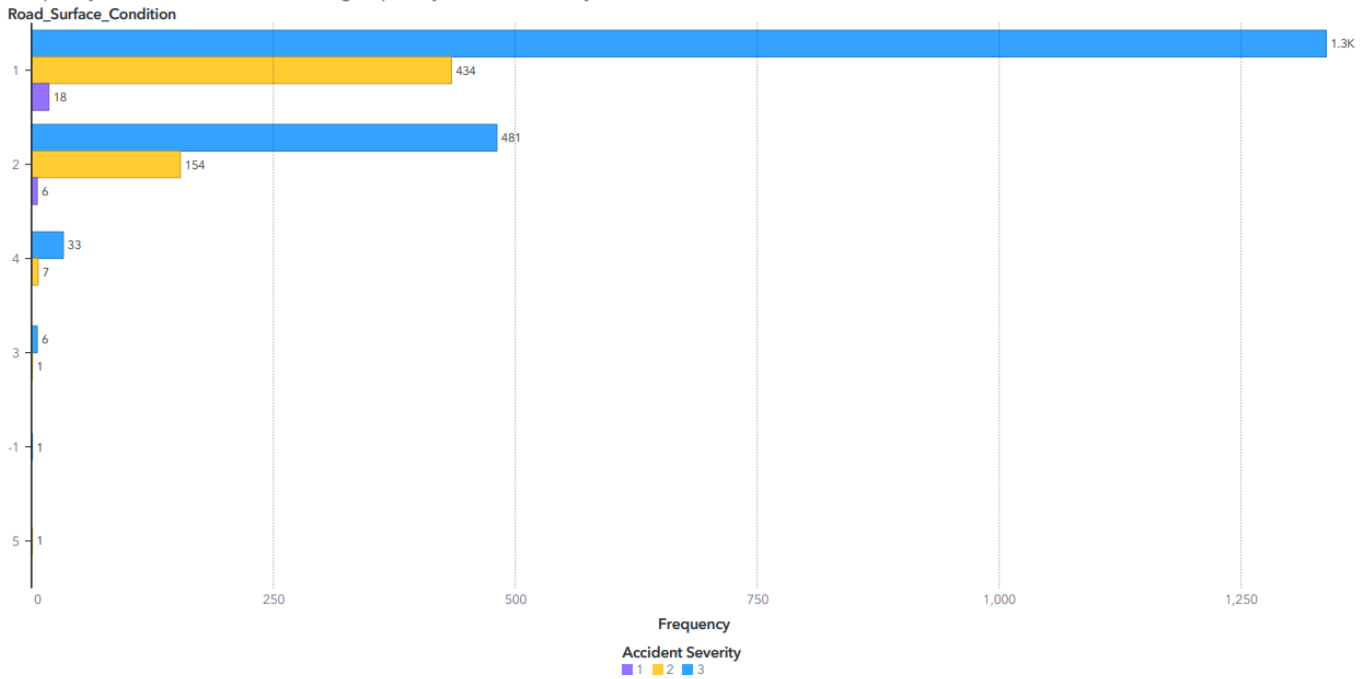
Frequency of Second\_Road\_Class grouped by Accident Severity



16

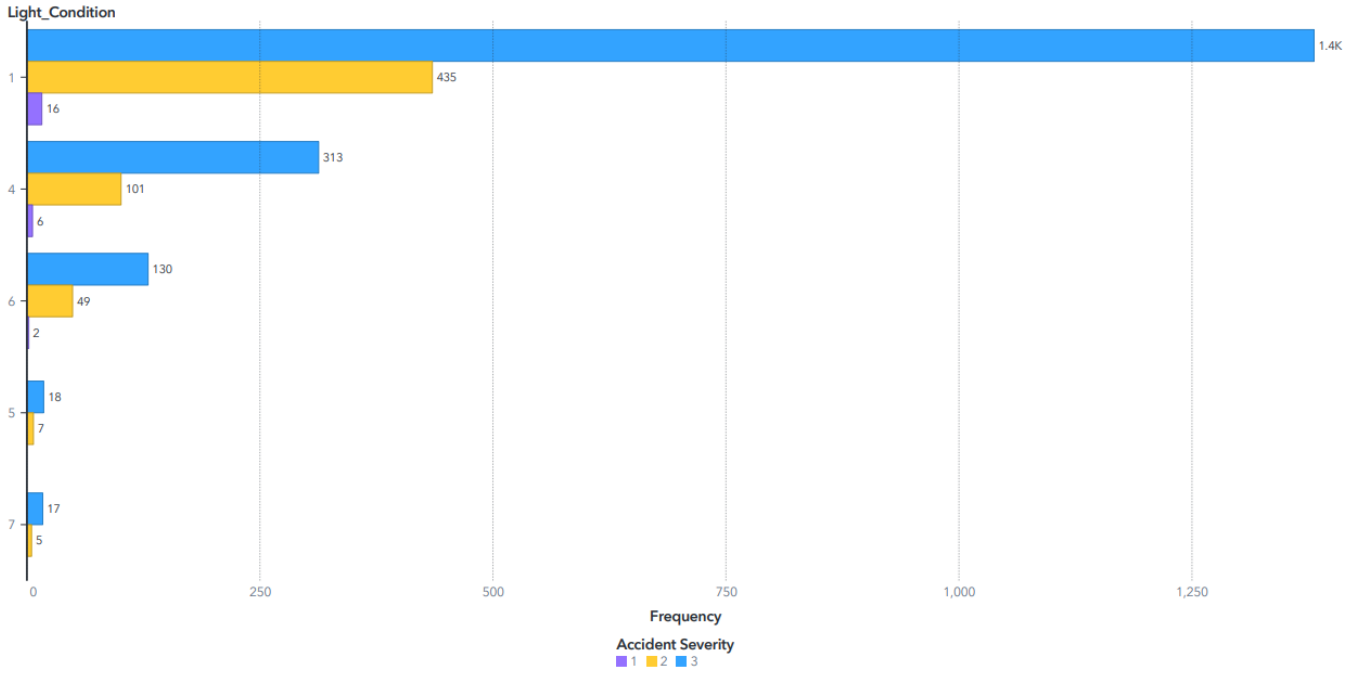
Page 15

Frequency of Road\_Surface\_Condition grouped by Accident Severity

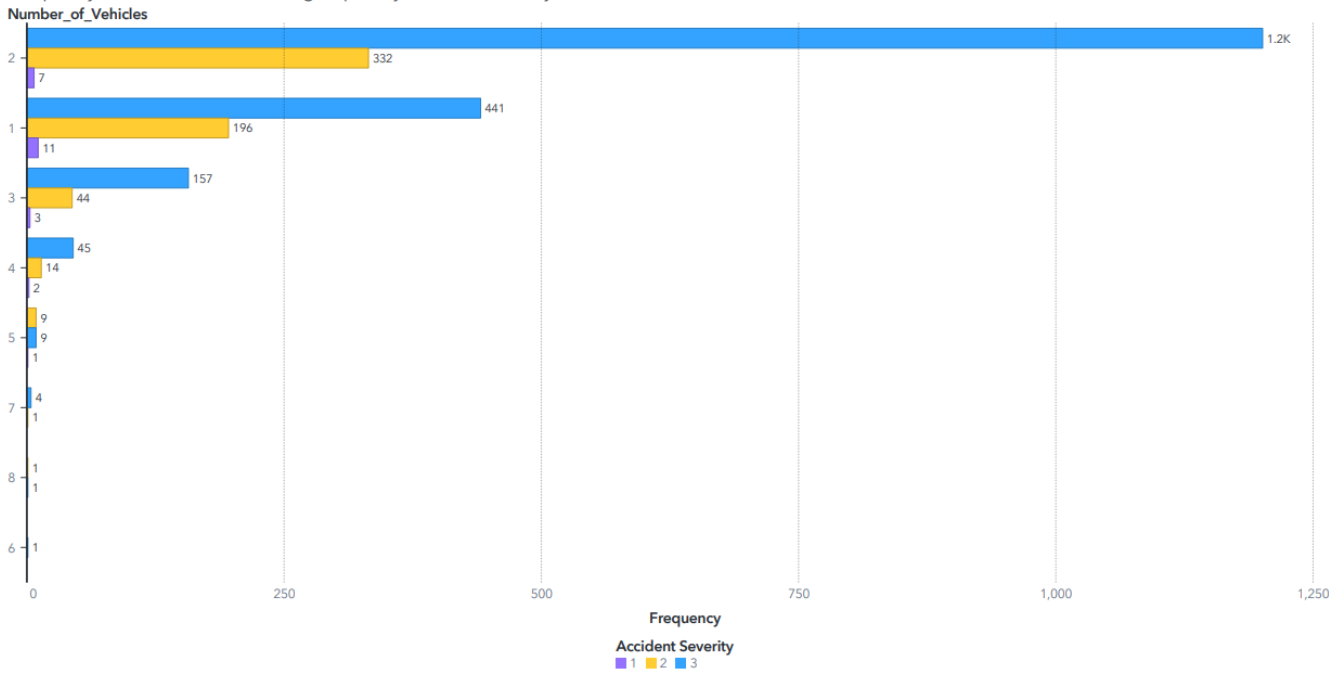


14

Frequency of Light\_Condition grouped by Accident Severity



Frequency of Number\_of\_Vehicles grouped by Accident Severity



Task 1, 2 and 3 (Technical report): Maximum 3000 words excluding tables, figures and appendices  
Task 4 (managerial report): Maximum 2 pages including tables, figures (no appendix for this task)

1. One file (Word or PDF) containing two parts: a technical report for tasks 1, 2, and 3 (maximum of 3000 words, excluding the title page, tables, figures, and appendix) and a managerial report for Task 4 (maximum of 2 pages, including tables, figures, no appendix for this task)
2. A PDF file of your saved SAS project.