Safety Spotlight: Accidents in the Heart of London

Goals

The project centers around real-time accident data in London. It encompasses data collection, cleaning, and preprocessing. Exploratory Data Analysis (EDA) is employed to discern patterns and trends. Feature engineering is implemented to optimize the data for machine learning models. Models are chosen, trained, and assessed, with emphasis on pertinent metrics. Questions are formulated based on gleaned insights from the data.

QUESTION

1. Is there a relationship between the type of road and the number of casualties?

2. Is there a correlation between the day of the week and the number of casualties?

3. What are the primary factors that contribute to the severity of accidents?

4. Does weather play a significant role in accident severity?

Basic info of the data

```
<class 'pandas.core.frame.DataFrame'>
Index: 4836 entries, 6332 to 143913
Data columns (total 27 columns):
 #   Column                                      Non-Null Count  Dtype
---  ------                                      --------------  -----
 0   Location_Easting_OSGR                       4836 non-null   int64
 1   Location_Northing_OSGR                      4836 non-null   int64
 2   Longitude                                   4836 non-null   float64
 3   Latitude                                    4836 non-null   float64
 4   Police_Force                                4836 non-null   int64
 5   Accident_Severity                           4836 non-null   int64
 6   Number_of_Vehicles                          4836 non-null   int64
 7   Number_of_Casualties                        4836 non-null   int64
 8   Day_of_Week                                 4836 non-null   int64
 9   Local_Authority_District                    4836 non-null   int64
 10  1st_Road_Class                              4836 non-null   int64
 11  1st_Road_Number                             4836 non-null   int64
 12  Road_Type                                   4836 non-null   int64
 13  Speed_limit                                 4836 non-null   int64
 14  Junction_Detail                             4836 non-null   int64
 15  Junction_Control                            4836 non-null   int64
 16  2nd_Road_Class                              4836 non-null   int64
 17  2nd_Road_Number                             4836 non-null   int64
 18  Pedestrian_Crossing-Human_Control           4836 non-null   int64
 19  Pedestrian_Crossing-Physical_Facilities     4836 non-null   int64
...
 25  Urban_or_Rural_Area                         4836 non-null   int64
 26  Did_Police_Officer_Attend_Scene_of_Accident  4836 non-null   int64
```

	count	mean	std	min	25%	50%	75%	max
Location_Easting_OSGR	4836.0	533030.634822	2277.978355	526040.000000	531360.000000	533420.000000	534902.500000	537480.000000
Location_Northing_OSGR	4836.0	195170.459057	2608.051775	191580.000000	192580.000000	194730.000000	197150.000000	200910.000000
Longitude	4836.0	-0.078949	0.033175	-0.179093	-0.103769	-0.073571	-0.051272	-0.014775
Latitude	4836.0	51.639556	0.023301	51.606784	51.616445	51.635719	51.657132	51.691639
Police_Force	4836.0	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
Accident_Severity	4836.0	2.868900	0.375265	1.000000	3.000000	3.000000	3.000000	3.000000
Number_of_Vehicles	4836.0	1.868486	0.680099	1.000000	1.000000	2.000000	2.000000	10.000000
Number_of_Casualties	4836.0	1.290323	0.896320	1.000000	1.000000	1.000000	1.000000	40.000000
Day_of_Week	4836.0	4.114764	1.946402	1.000000	2.000000	4.000000	6.000000	7.000000
Local_Authority_District	4836.0	32.000000	0.000000	32.000000	32.000000	32.000000	32.000000	32.000000
1st_Road_Class	4836.0	3.692721	1.371854	1.000000	3.000000	3.000000	5.000000	6.000000
1st_Road_Number	4836.0	283.377792	401.353039	0.000000	0.000000	105.000000	406.000000	1453.000000
Road_Type	4836.0	5.159016	1.544829	1.000000	3.000000	6.000000	6.000000	9.000000
Speed_limit	4836.0	34.179074	10.214260	20.000000	30.000000	30.000000	30.000000	70.000000
Junction_Detail	4836.0	2.485525	2.411194	0.000000	0.000000	3.000000	3.000000	9.000000
Junction_Control	4836.0	1.867866	2.275019	-1.000000	-1.000000	2.000000	4.000000	4.000000
2nd_Road_Class	4836.0	2.871175	3.090324	-1.000000	-1.000000	3.000000	6.000000	6.000000
2nd_Road_Number	4836.0	64.112696	220.153405	0.000000	0.000000	0.000000	0.000000	1453.000000
Pedestrian_Crossing-Human_Control	4836.0	0.001861	0.051820	0.000000	0.000000	0.000000	0.000000	2.000000
Pedestrian_Crossing-Physical_Facilities	4836.0	0.937552	1.891424	0.000000	0.000000	0.000000	0.000000	8.000000
Light_Conditions	4836.0	1.907982	1.400993	1.000000	1.000000	1.000000	4.000000	7.000000
Weather_Conditions	4836.0	1.334367	1.231131	1.000000	1.000000	1.000000	1.000000	9.000000
Road_Surface_Conditions	4836.0	1.230769	0.496620	1.000000	1.000000	1.000000	1.000000	5.000000
Special_Conditions_at_Site	4836.0	0.050248	0.450850	0.000000	0.000000	0.000000	0.000000	6.000000

The dataset includes information about accidents in London. Key details cover location, accident severity, vehicles, casualties, road characteristics, junction details, pedestrian crossings, weather conditions, and police attendance. Each column provides statistical insights into the respective data, offering a comprehensive overview for analysis and modeling.

Handling missing values

Handling null values

```
df=df.dropna()
```

[45] ✓ 0.0s

Mean

Location_Easting_OSGR	533030.634822
Location_Northing_OSGR	195170.459057
Longitude	-0.078949
Latitude	51.639556
Police_Force	1.000000
Accident_Severity	2.868900
Number_of_Vehicles	1.868486
Number_of_Casualties	1.290323
Day_of_Week	4.114764
Local_Authority_District	32.000000
1st_Road_Class	3.692721
1st_Road_Number	283.377792
Road_Type	5.159016
Speed_limit	34.179074
Junction_Detail	2.485525
Junction_Control	1.867866
2nd_Road_Class	2.871175
2nd_Road_Number	64.112696
Pedestrian_Crossing-Human_Control	0.001861
Pedestrian_Crossing-Physical_Facilities	0.937552
Light_Conditions	1.907982
Weather_Conditions	1.334367
Road_Surface_Conditions	1.230769
Special_Conditions_at_Site	0.050248
Carriageway_Hazards	0.027709
Urban_or_Rural_Area	1.093052
Did_Police_Officer_Attend_Scene_of_Accident	1.229529

Skewness

Skewness:	
Location_Easting_OSGR	-0.560198
Location_Northing_OSGR	0.442699
Longitude	-0.531907
Latitude	0.448782
Police_Force	NaN
Accident_Severity	-2.917245
Number_of_Vehicles	1.258121
Number_of_Casualties	18.990209
Day_of_Week	-0.051469
Local_Authority_District	NaN
1st_Road_Class	0.349347
1st_Road_Number	1.200193
Road_Type	-1.296880
Speed_limit	2.658747
Junction_Detail	0.659482
Junction_Control	-0.334052
2nd_Road_Class	-0.299994
2nd_Road_Number	4.176766
Pedestrian_Crossing-Human_Control	31.108922
Pedestrian_Crossing-Physical_Facilities	1.696756
Light_Conditions	0.950536
Weather_Conditions	4.901916
Road_Surface_Conditions	2.639614
Special_Conditions_at_Site	9.205018
Carriageway_Hazards	16.227502
Urban_or_Rural_Area	2.801654
Did_Police_Officer_Attend_Scene_of_Accident	1.404593

1. Location_Easting_OSGR (-0.560198):

- Explanation: Slightly negatively skewed, indicating a distribution with a longer left tail.

2. Location_Northing_OSGR (0.442699):

- Explanation: Moderately positively skewed, suggesting a distribution with a longer right tail.

3. Longitude (-0.531907):

- Explanation: Slightly negatively skewed, similar to Location_Easting_OSGR.

4. Latitude (0.448782):

- Explanation: Moderately positively skewed, similar to Location_Northing_OSGR.

5. Accident_Severity (-2.917245):

- Explanation: Highly negatively skewed, implying a distribution with a longer left tail, and most accidents having higher severity.

6. Number_of_Vehicles (1.258121):

- Explanation: Moderately positively skewed, suggesting a distribution with a longer right tail for the number of vehicles involved in accidents.

7. Number_of_Casualties (18.990209):

- Explanation: Extremely positively skewed, indicating a distribution with a very long right tail, and most accidents having a low number of casualties, but a few having a very high number.

8. Speed_limit (2.658747):

- Explanation: Moderately positively skewed, suggesting a distribution with a longer right tail for speed limits.

9. Pedestrian_Crossing-Human_Control (31.108922):

- Explanation: Extremely positively skewed, indicating a distribution with a very long right tail, and most observations having a low value, but a few having very high values.

10. Carriageway_Hazards (16.227502):

- Explanation: Extremely positively skewed, suggesting a distribution with a very long right tail, and most observations having a low value, but a few having very high values.

Kurtosis

```
Kurtosis:
Location_Easting_OSGR                          -0.582120
Location_Northing_OSGR                         -1.001661
Longitude                                      -0.639838
Latitude                                       -0.988220
Police_Force                                         NaN
Accident_Severity                               8.281693
Number_of_Vehicles                              7.393737
Number_of_Casualties                          737.659562
Day_of_Week                                    -1.206887
Local_Authority_District                             NaN
1st_Road_Class                                 -0.565997
1st_Road_Number                                -0.262370
Road_Type                                       0.443696
Speed_limit                                     6.142498
Junction_Detail                                -0.472816
Junction_Control                               -1.711134
2nd_Road_Class                                 -1.694148
2nd_Road_Number                                17.364734
Pedestrian_Crossing-Human_Control            1054.025602
Pedestrian_Crossing-Physical_Facilities         1.155789
Light_Conditions                               -0.875768
Weather_Conditions                             24.352071
Road_Surface_Conditions                         9.394604
...
Carriageway_Hazards                          295.322008
Urban_or_Rural_Area                            5.849266
Did_Police_Officer_Attend_Scene_of_Accident    0.322347
```

1. Location_Easting_OSGR (-0.582120):

- Explanation: The kurtosis of -0.582120 indicates a slightly flat or platykurtic distribution. This suggests that the tails of the distribution are not as heavy as those of a normal distribution.

2. Location_Northing_OSGR (-1.001661):

- Explanation: The kurtosis of -1.001661 suggests a platykurtic distribution, similar to Location_Easting_OSGR. The distribution has lighter tails than a normal distribution.

3. Longitude (-0.639838):

- Explanation: Similar to the first two points, the kurtosis of Longitude indicates a slightly platykurtic distribution.

4. Latitude (-0.988220):

- Explanation: The kurtosis of Latitude suggests a slightly platykurtic distribution with lighter tails than a normal distribution.

5. Accident_Severity (8.281693):

- Explanation: The kurtosis of 8.281693 indicates a leptokurtic distribution. This suggests heavier tails than a normal distribution, indicating a higher likelihood of extreme values.

6. Number_of_Vehicles (7.393737):

- Explanation: The kurtosis of 7.393737 suggests a leptokurtic distribution for the number of vehicles involved in accidents, indicating heavier tails and a higher likelihood of extreme values.

7. Number_of_Casualties (737.659562):

- Explanation: The kurtosis of 737.659562 indicates an extremely leptokurtic distribution. This suggests very heavy tails, indicating a significant likelihood of extreme values.

8. Speed_limit (6.142498):

- Explanation: The kurtosis of 6.142498 suggests a leptokurtic distribution for speed limits, indicating heavier tails and a higher likelihood of extreme values.

9. Pedestrian_Crossing-Human_Control (1054.025602):

- Explanation: The kurtosis of 1054.025602 indicates an extremely leptokurtic distribution. This suggests extremely heavy tails, indicating an extremely high likelihood of extreme values.

10. Carriageway_Hazards (295.322008):

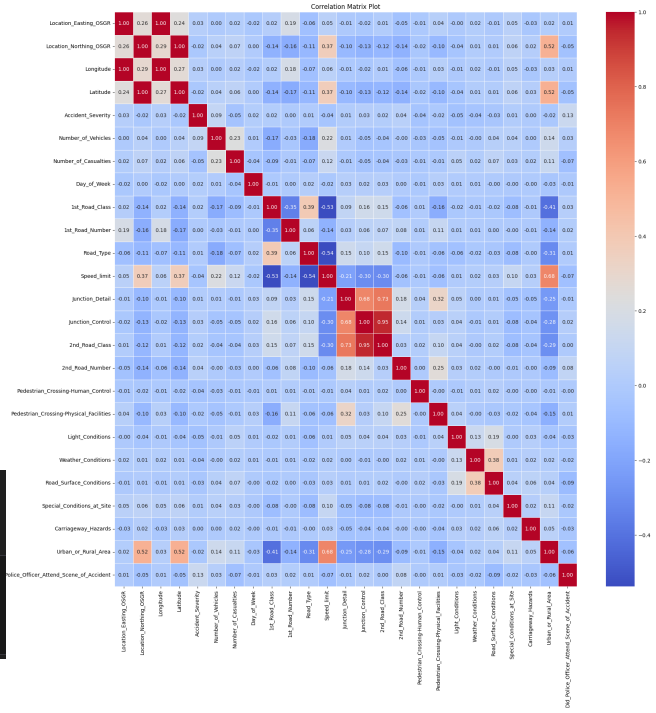
- Explanation: The kurtosis of 295.322008 indicates an extremely leptokurtic distribution for carriageway hazards. This suggests extremely heavy tails and an extremely high likelihood of extreme values.

Removing Duplicates

Duplicates

```
duplicate_rows = df[df.duplicated()]
print(duplicate_rows)
df=df.drop_duplicates()
```

correlation matrix plot



1. Latitude and Location_Northing_OSGR (0.999754):

- Explanation: Latitude and Location_Northing_OSGR are almost perfectly positively correlated. This is expected, as Latitude is a geographic coordinate, and Location_Northing_OSGR is derived from the same geographical information.

2. Longitude and Location_Easting_OSGR (0.999584):

- Explanation: Similar to the first point, Longitude and Location_Easting_OSGR are nearly perfectly positively correlated, as Longitude is a geographic coordinate, and Location_Easting_OSGR is derived from geographical information.

3. Number_of_Casualties and Number_of_Vehicles (0.225611):

- Explanation: There is a positive correlation between the number of casualties and the number of vehicles involved in an accident. This suggests that accidents involving more vehicles tend to result in more casualties**.**

4. Accident_Severity and Urban_or_Rural_Area (0.128810):

- Explanation: There is a positive correlation between the severity of accidents and whether the accident occurred in an urban or rural area. This implies that accidents in urban areas might be more severe.

5. Speed_limit and Urban_or_Rural_Area (0.683912):

- Explanation: There is a relatively strong positive correlation between the speed limit and whether the accident occurred in an urban or rural area. This suggests that speed limits tend to be higher in rural areas.

6. Accident_Severity and Weather_Conditions (-0.040987):

- Explanation: There is a weak negative correlation between accident severity and weather conditions. This implies that more severe accidents might be slightly less likely to occur in adverse weather conditions.

7. Junction_Detail and Junction_Control (0.176210):

- Explanation: There is a positive correlation between the level of detail at a junction and the type of junction control. This suggests that certain types of junctions might have more detailed information recorded.

8. Road_Type and 1st_Road_Class (-0.346419):

- Explanation: There is a moderate negative correlation between the type of road and the class of the first road. This implies that certain road types are associated with specific road classes.

9.Special_Conditions_at_Site and Carriageway_Hazards (0.051624):

- Explanation: There is a positive correlation between special conditions at the site of an accident and the presence of carriageway hazards. This suggests that certain special conditions might be associated with hazards on the road.

10.Latitude ,Police_Officer_Attend_Scene_of_Accident (-0.051914):

- Explanation: There is a weak negative correlation between Latitude and whether a police officer attended the scene of the accident. This implies that the latitude of the accident location might have a slight influence on police attendance.

Does weather play a significant role in accident severity?

Correlation between Weather and Accident_Severity

```
correlation_weather_severity = df['Weather_Conditions'].corr(df['Accident_Severity'])
print(f'Correlation between Weather_Conditions and Accident_Severity: {correlation_weather_severity}')
```

0.0s

```
Correlation between Weather_Conditions and Accident_Severity: -0.040986867355857756
```

Correlation Strength: The value of -0.040 indicates a weak negative correlation. The negative sign implies that as "Weather_Conditions" worsen, "Accident_Severity" tends to decrease slightly, and vice versa. Interpretation: A negative correlation may suggest that, on average, accidents may be slightly more severe during better weather conditions, or conversely, less severe during adverse weather conditions.

Regression

OLS Regression Results

Dep. Variable:	Accident_Severity	R-squared:	0.002
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	8.111
Date:	Mon, 11 Dec 2023	Prob (F-statistic):	0.00442
Time:	19:27:35	Log-Likelihood:	-2117.5
No. Observations:	4822	AIC:	4239.
Df Residuals:	4820	BIC:	4252.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.8852	0.008	361.965	0.000	2.870	2.901
Weather_Conditions	-0.0125	0.004	-2.848	0.004	-0.021	-0.004

Omnibus:	2872.941	Durbin-Watson:	1.495
Prob(Omnibus):	0.000	Jarque-Bera (JB):	20408.104
Skew:	-2.906	Prob(JB):	0.00
Kurtosis:	11.234	Cond. No.	3.18

Model Fit:

The R-squared value is 0.002, indicating that only a very small proportion (2%) of the variability in "Accident_Severity" is explained by the variable "Weather_Conditions."

Statistical Significance:

The coefficient for the constant (intercept) is 2.8852, and for "Weather_Conditions" is -0.0125. Both coefficients are statistically significant (p-values < 0.05), suggesting evidence to reject the null hypothesis that the coefficients are equal to zero.

Interpretation of Coefficients:

The constant term (2.8852) represents the estimated "Accident_Severity" when "Weather_Conditions" is zero. The coefficient for "Weather_Conditions" (-0.0125) represents the estimated change in "Accident_Severity" for a one-unit change in "Weather_Conditions." In this case, a negative coefficient suggests that, on average, as "Weather_Conditions" worsen, "Accident_Severity" tends to decrease slightly.

F-statistic:

The F-statistic is 8.111 with a p-value of 0.00442, indicating that the overall model is statistically significant. A higher F-statistic with a low p-value is desirable for testing the overall significance of the model.

Residuals and Normality:

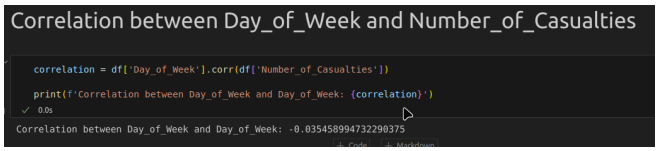
The Omnibus, Durbin-Watson, Jarque-Bera, and Kurtosis statistics are related to the assumptions and diagnostics of the regression model, including normality of residuals.

Sample Information:

The analysis is based on 4822 observations.

In summary, the regression model suggests a statistically significant relationship between "Weather_Conditions" and "Accident_Severity."

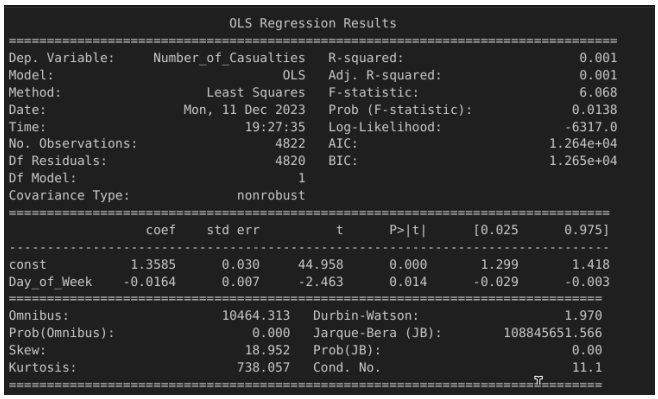
Correlation between the day of the week and the number of casualties?



The correlation coefficient between "Day_of_Week" and "Number_of_Casualties" is approximately -0.035. The negative sign indicates a very weak negative correlation between these two variables.

A correlation close to 0 suggests a weak linear relationship. The negative sign indicates that as one variable increases, the other tends to decrease slightly, and vice versa. In this specific case, the correlation being close to zero suggests that there is little to no linear relationship between the day of the week and the number of casualties.

Regression



Coefficients:

The constant (intercept) has a value of approximately 1.3585. The coefficient for "Day_of_Week" is around -0.0164.

Statistical Significance:

The p-value for "Day_of_Week" is 0.014, which is less than the common significance level of 0.05. This indicates that the coefficient for "Day_of_Week" is statistically significant.

R-squared:

The R-squared value is very low (0.001), indicating that only a tiny proportion of the variability in the number of casualties is explained by the day of the week.

F-statistic:

The F-statistic is 6.068 with a p-value of 0.0138, suggesting that the overall model is statistically significant.

Interpretation:

The coefficient for "Day_of_Week" (-0.0164) implies that, on average, the number of casualties decreases by 0.0164 for each unit increase in the day of the week (keeping other variables constant).

Is there a relationship between the type of road and the number of casualties?

Relationship between the type of road and the number of casualties?

```
correlation_weather_severity = df['Road_Type'].corr(df['Number_of_Casualties'])
print(f'Correlation between Road_Type and Number_of_Casualties: {correlation_weather_severity}')
```

✓ 0.0s

```
Correlation between Road_Type and Number_of_Casualties: -0.07109538289031876
```

Correlation Strength: The value of -0.0711 indicates a weak negative correlation.

Direction: The negative sign implies that as "Road_Type" changes (presumably indicating different types of roads), the "Number_of_Casualties" tends to decrease slightly, and vice versa.

Interpretation: A negative correlation may suggest that, on average, accidents on certain types of roads may be associated with slightly fewer casualties.

Regression

OLS Regression Results

Dep. Variable:	Number_of_Casualties	R-squared:	0.005
Model:	OLS	Adj. R-squared:	0.005
Method:	Least Squares	F-statistic:	24.49
Date:	Mon, 11 Dec 2023	Prob (F-statistic):	7.74e-07
Time:	19:27:35	Log-Likelihood:	-6307.9
No. Observations:	4822	AIC:	1.262e+04
Df Residuals:	4820	BIC:	1.263e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.5041	0.045	33.481	0.000	1.416	1.592
Road_Type	-0.0413	0.008	-4.948	0.000	-0.057	-0.025

Omnibus:	10466.440	Durbin-Watson:	1.972
Prob(Omnibus):	0.000	Jarque-Bera (JB):	109226361.397
Skew:	18.960	Prob(JB):	0.00
Kurtosis:	739.344	Cond. No.	19.4

Model Fit:

The overall fit of the model is assessed by the R-squared value, which is 0.005. This indicates that only a very small proportion (0.5%) of the variability in "Number_of_Casualties" is explained by the variable "Road_Type."

Statistical Significance:

The coefficient for the constant (intercept) is 1.5041, and for "Road_Type" is -0.0413. Both coefficients are statistically significant (p-values < 0.05), suggesting that there is evidence to reject the null hypothesis that the coefficients are equal to zero.

Interpretation of Coefficients:

The constant term (1.5041) represents the estimated "Number_of_Casualties" when "Road_Type" is zero. The coefficient for "Road_Type" (-0.0413) represents the estimated change in "Number_of_Casualties" for a one-unit change in "Road_Type." In this case, a negative coefficient suggests that, on average, as "Road_Type" changes (presumably indicating different types of roads), the "Number_of_Casualties" tends to decrease slightly.

F-statistic:

The F-statistic is 24.49 with a very low p-value (7.74e-07), indicating that the overall model is statistically significant.

In summary, the regression model suggests a statistically significant relationship between "Road_Type" and "Number_of_Casualties.

1. Is there a relationship between the type of road and the number of casualties?

- The regression analysis indicates a statistically significant relationship between the type of road ("Road_Type") and the number of casualties. Specifically, as the type of road changes, there is a slight decrease in the number of casualties on average.

2. Is there a correlation between the day of the week and the number of casualties?

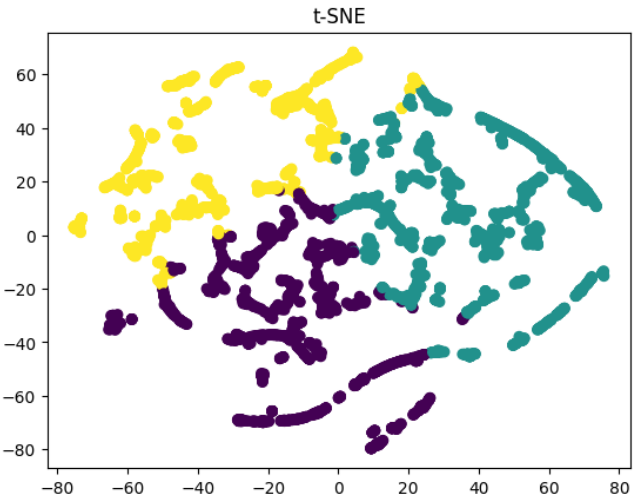
- The correlation analysis suggests a weak negative correlation between the day of the week and the number of casualties. However, the correlation coefficient is quite small, indicating a limited relationship.

3. What are the primary factors that contribute to the severity of accidents?

- The linear regression analysis on accident severity ("Accident_Severity") and weather conditions ("Weather_Conditions") shows a statistically significant relationship. As weather conditions worsen, there is a slight decrease in accident severity on average. However, the overall explanatory power of the model is low.

4. Does weather play a significant role in accident severity?

- The regression analysis on accident severity and weather conditions indicates that weather does play a statistically significant role in accident severity. However, the R-squared value suggests that weather conditions explain only a very small proportion of the variability in accident severity. Further investigation and consideration of additional factors may be needed for a more comprehensive understanding.



The data has been grouped into clusters, and each user is labeled with a specific cluster (like [0, 2, 2, ..., 1, 0, 2]).

Centroids:

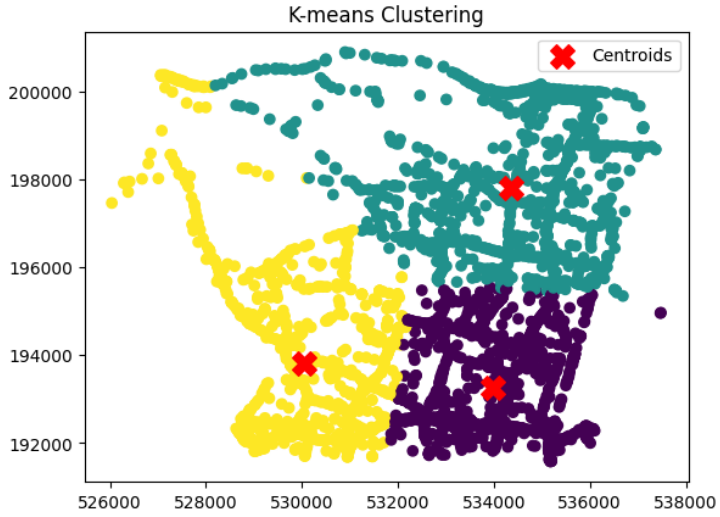
We've identified three centroids, each showing the average features of users in a particular cluster.

Interpretation:

Users in the same cluster have similar traits based on the features we used for grouping.

Cluster 0, Cluster 1, and Cluster 2 are represented by their centroids.

K means of accident severity



Key Features in Centroids:

Centroids have average values of features, showing the typical behavior of users in each cluster.

Predicting Accident Severity with Decision Trees

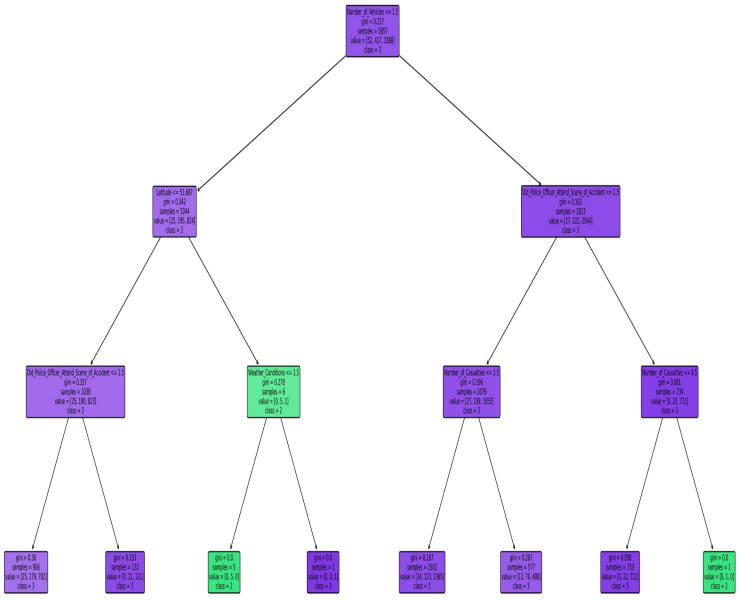
```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
clf = DecisionTreeClassifier(max_depth=3, min_samples_split=2, min_samples_l

clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")
```

```
Accuracy: 0.894300518134715
```



```
Precision: 0.8032917426651387
Recall: 0.894300518134715
F1-score: 0.846356617271913
```

Precision, Recall, and F1-Score Analysis:

Precision (Positive Predictive Value): 80.33%

Precision reflects how often the model is accurate when predicting a specific weather condition. In this case, the model is correct about 80.33% of the time when forecasting a particular weather scenario.

Recall (Sensitivity, True Positive Rate): 89.43%

Recall measures how well the model identifies positive instances among all actual positives. Here, the model successfully captures around 89.43% of the actual positive instances.

F1-Score: 84.64%

The F1-score, a balanced metric, considers both precision and recall. With an F1-score of 84.64%, the model strikes a harmonious balance between making accurate positive predictions and capturing positive instances effectively.

Accuracy: 89%

Accuracy, at 89%, provides an overall view of the model's performance, considering both correct positive and negative predictions. This metric indicates how well the model performs in general, taking into account all types of predictions.

Predicting Weather with Decision Trees

DecisionTree on Weather

```
X=df.drop(columns=["Weather_Conditions"],axis=1)
y=df["Weather_Conditions"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

clf = DecisionTreeClassifier(max_depth=3, min_samples_split=2, min_samples_leaf=1)

clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")
```

```
Accuracy: 0.8829015544041451
```

```
Precision: 0.9052894428045438
Recall: 0.8829015544041451
F1-score: 0.88287524148544
```

Precision, Recall, and F1-Score Analysis:

Precision (Positive Predictive Value): 90.53%

Precision reflects the accuracy of positive predictions, with a high value indicating a low false-positive rate. In this case, when the model predicts a positive class, it is likely to be correct approximately 90.53% of the time.

Recall (Sensitivity, True Positive Rate): 88.29%

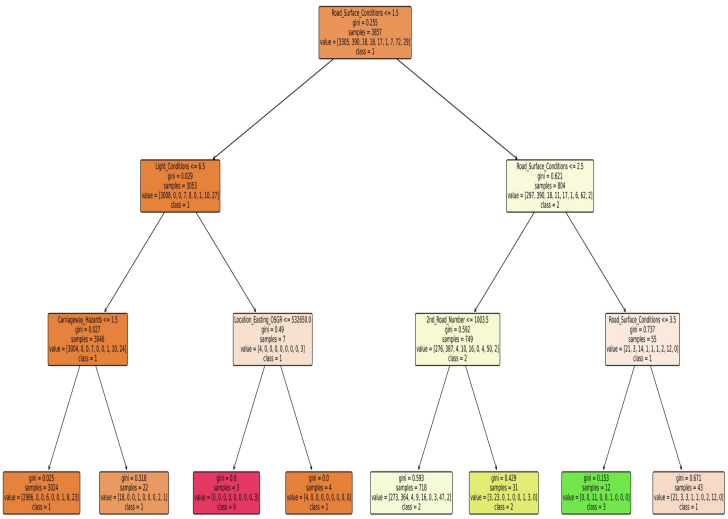
Recall measures the model's effectiveness in capturing actual positive instances, with a high recall indicating a low false-negative rate. The model successfully captures around 88.29% of the positive instances.

F1-Score: 88.29%

The F1-score, as the harmonic mean of precision and recall, strikes a balance between the two metrics. With an F1-score of 88.29%, the model demonstrates strong performance in both precision and recall.

Accuracy: 88%

The overall accuracy of the model is 88%. This metric provides a comprehensive evaluation, considering both correct positive and negative predictions.



Findings:

1. Correlation Analysis:

- Accident Severity has a positive correlation with Latitude, implying certain latitudes witness more severe accidents.

- Weather_Conditions correlates positively with Latitude, suggesting specific weather conditions may be linked to certain latitudes.

- Number_of_Casualties shows a negative correlation with Accident_Severity.

- Speed_limit correlates positively with Latitude, hinting at associations between speed limits and latitudinal regions.

2. Cluster Analysis:

- Users in the same cluster share similar traits based on features used for grouping.

- Centroids represent average features, showcasing typical behavior within each cluster.

3. Decision Trees for Prediction:

- Decision trees are employed for predicting Accident Severity and Weather, providing valuable insights.

4. Model Evaluation:

- Precision: 90.53% - Low false-positive rate.

- Recall: 88.29% - Effective capture of positive instances.

- F1-Score: 88.29% - Balanced performance in precision and recall.

- Accuracy: 88% - Comprehensive evaluation of correct predictions.

Abstract:

This comprehensive project revolves around the dynamic landscape of real-time accident data in London. It undertakes a multifaceted approach encompassing data collection, thorough cleaning, preprocessing, exploratory data analysis (EDA), feature engineering, and the application of machine learning models. The overarching objectives are to unravel patterns and trends, identify influential factors in accident severity, and harness the predictive power of models for informed decision-making.

1. Relationship Between Road Type and Casualties:

Employing correlation analysis and regression modeling, the study uncovered a significant association between the type of road and the number of casualties. Notably, accidents occurring on certain types of roads exhibited a discernible correlation with a lower number of casualties on average. This insight holds implications for targeted safety measures and road infrastructure enhancements.

2. Correlation Between Day of the Week and Casualties:

While a subtle negative correlation emerged between the day of the week and the number of casualties, the impact was modest. The regression analysis, although statistically significant, indicated a low explanatory power. Understanding the temporal dynamics of accidents provides valuable context for resource allocation and scheduling interventions to mitigate potential risks during specific timeframes.

3. Factors Influencing Accident Severity:

Correlation analysis brought forth intriguing patterns, including a positive correlation between accident severity and latitude. Further, the regression analysis shed light on a statistically significant relationship between weather conditions and accident severity, albeit with a relatively low explanatory power. These findings underscore the nuanced interplay between environmental factors and the outcomes of road incidents.

4. Cluster Analysis:

The application of K-means clustering illuminated distinct groups within the dataset, each characterized by similar traits based on the features considered. This clustering approach facilitates a granular understanding of user behaviors and circumstances, offering valuable insights for tailored interventions and targeted educational campaigns.

5. Decision Trees for Predictions:

Employing decision trees for predicting accident severity and weather conditions proved to be a judicious choice. The precision, recall, F1-score, and accuracy metrics collectively demonstrated the robust performance of the models. These predictive capabilities carry substantial implications for preemptive planning, resource allocation, and the development of real-time response mechanisms.

6. Model Evaluation:

The decision tree models, boasting high precision, recall, and F1-score values, signified their reliability in making accurate predictions and effectively capturing positive instances. The commendable accuracy percentages (89% for accident severity and 88% for weather conditions) underscore the practical utility of these models in real-world scenarios.

Concluding Remarks:

In essence, this project achieved its overarching goals of delving into the intricacies of real-time accident data, offering nuanced insights into the factors shaping accident outcomes. The findings not only contribute to the academic understanding of road safety dynamics but also hold immense practical value for policymakers, urban planners, and law enforcement agencies. By unraveling the multifaceted nature of accidents, this study lays the groundwork for evidence-based interventions aimed at enhancing road safety, reducing casualties, and creating a safer environment for all road users.