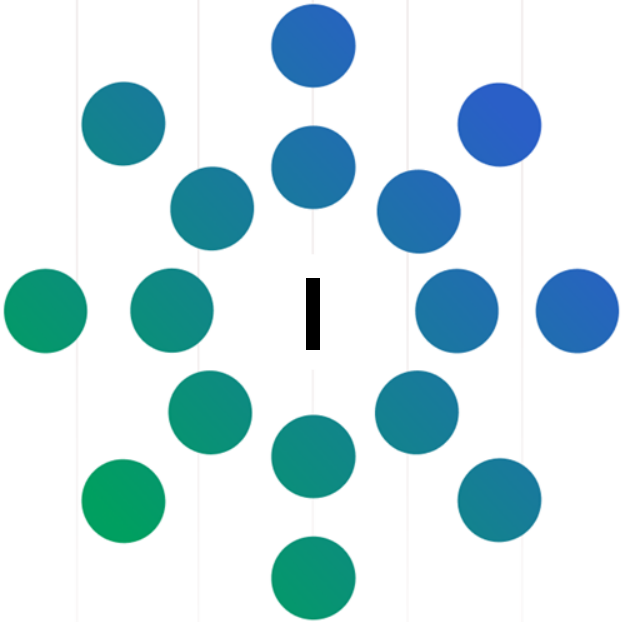
Mini Project Data Science

Analysis of Traffic Conditions in Bandung City using Waze Data

by Adrian Maulana Muhammad

https://github.com/adrn-mm/DSLS_Mini-Project-DS_Adrian-Maulana-Muhammad

Table of Contents



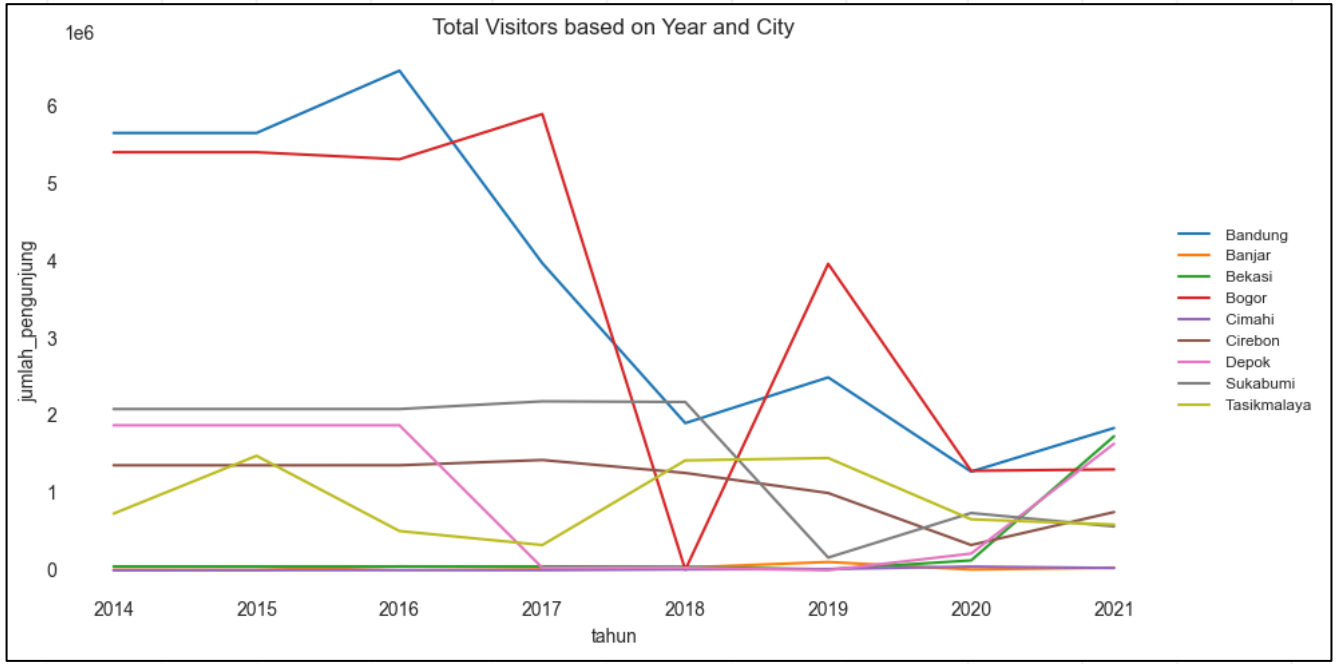
Business Understanding

Visitors in Bandung City

Based on an analysis of visitor data for cities in West Java province provided by Open Data Jabar, Bandung has the highest number of visitors and a positive trend in visitor growth over the last two years. This phenomenon may impact traffic conditions on all roads in Bandung. By utilizing traffic data provided by Waze for Cities, we can analyze the traffic conditions in Bandung, including information on traffic jams, irregularities, and alerts.

Total Visitors based on Year and City

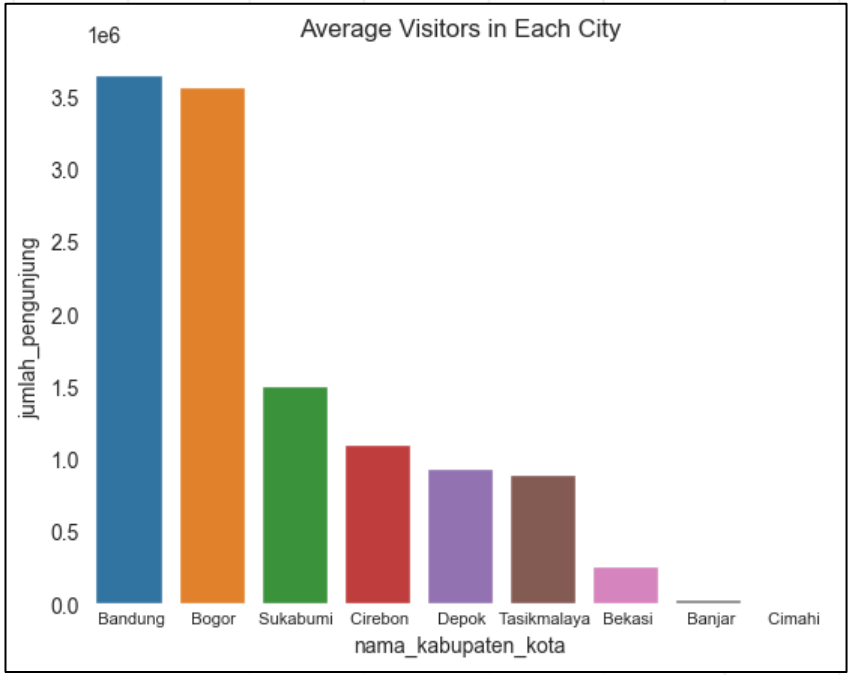
The chart below shows that Bandung has a positive trend in visitor growth over the last two years.



Source: Open Data Jabar

Average Total Visitors in Each City

The chart below shows that Bandung has the highest number of visitors.



Source: Open Data Jabar

Problem Description

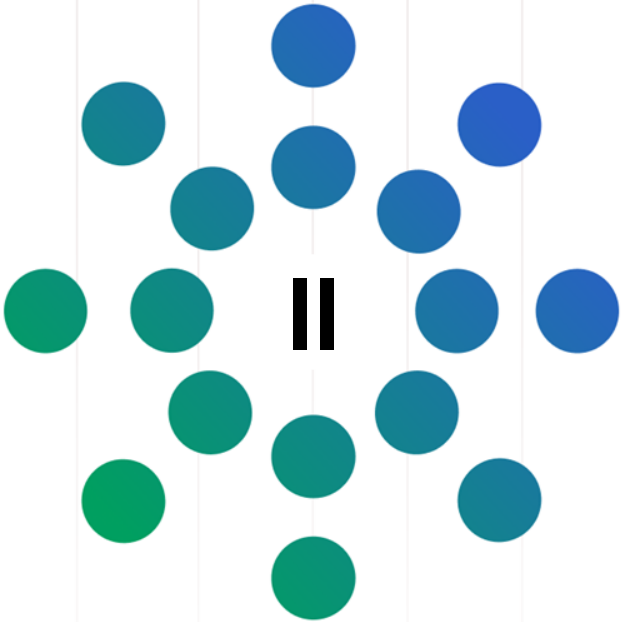
I have identified several summary questions to address the problem from the background information provided:

1. What are the traffic alert patterns in Bandung City?
2. How is the pattern and distribution of traffic congestion in Bandung City?
3. How can we group Bandung City areas with similar traffic characteristics?

Project Goals

To answer this question, I set the goals and methods that can answer the problem.

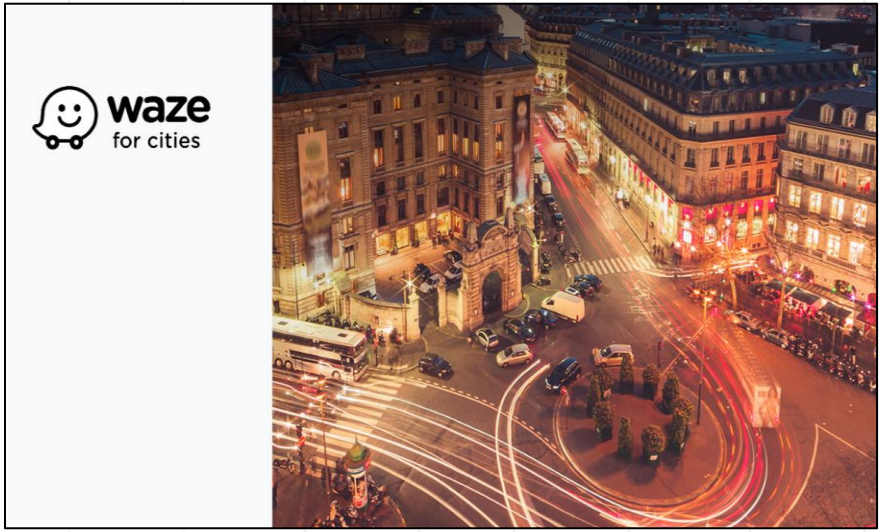
1. Analyze the traffic alert pattern in the city of Bandung using Waze for Cities data analysis to see the most dominant type of alert.
2. Analyzing traffic congestion patterns in Bandung by analyzing Waze for Cities data for Bandung in various circumstances.
3. Conducting a Machine Learning algorithm, which is clustering, to get a pattern of group division for roads in Bandung with similar traffic characteristics.



Data Understanding

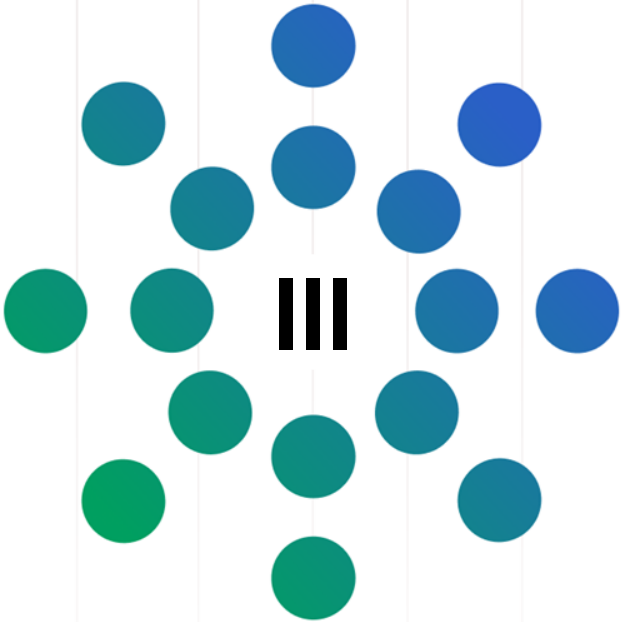
Waze for Cities Data

Through the Waze for Cities initiative, which involves a data-sharing partnership between Waze and public sector partners worldwide, we can utilize the data to analyze traffic safety in Bandung City for visitors.



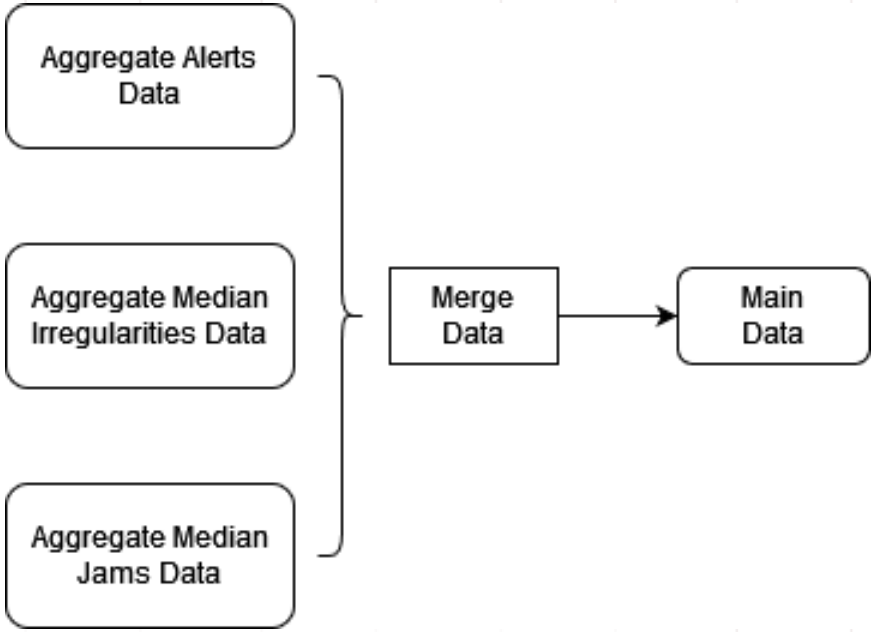
Data Summaries

Data	Summary
Aggregate Alerts	This data provides information on the number and types of incident reports sent by Waze users within a specific area during a certain period of time. This data can be helpful for road users and public service providers identify incidents affecting traffic and mobility in a particular area.
Aggregate Median Irregularities	This data provides information on the condition of roads, particularly in terms of irregularities on the road surface. This data can assist public service providers in monitoring road conditions and planning for repairs.
Aggregate Median Jams	This data provides information on traffic congestion on a particular road during a certain period of time. This data can assist road users in planning more efficient travel routes and avoiding traffic congestion.



Data Description & Pre-processing

Denormalized Data



Data Shape & Data Types

- The main data has 15 columns and 19.809 rows
- The list of data types is as follows, I will convert the Datetime column to the DateTime type.

```
Data columns (total 15 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Street                              19801 non-null  object
 1   Datetime                            19809 non-null  object
 2   Jam Level                           19809 non-null  int64
 3   Jam Length (meters)                 19809 non-null  float64
 4   Jam Time Spent (seconds)            19809 non-null  float64
 5   Jam Speed (Km/h)                    19809 non-null  float64
 6   Total Jam Records                   19809 non-null  int64
 7   Jam Geometry                        19809 non-null  object
 8   Irregularities Length (meters)      19809 non-null  float64
 9   Irregularities Time Spent (seconds) 19809 non-null  float64
 10  Irregularities Speed (Km/h)         19809 non-null  float64
 11  Total Irregularities Records        19809 non-null  int64
 12  Irregularities Geometry             19809 non-null  object
 13  Alert Type                          19809 non-null  object
 14  Total Alert Records                 19809 non-null  int64
dtypes: float64(6), int64(4), object(5)
```

Checking Missing Values

Since there are only eight rows with missing values out of 19.809 rows, I will drop those rows with missing values.

```
Street                                  8
Datetime                                0
Jam Level                               0
Jam Length (meters)                     0
Jam Time Spent (seconds)                0
Jam Speed (Km/h)                        0
Total Jam Records                       0
Jam Geometry                            0
Irregularities Length (meters)          0
Irregularities Time Spent (seconds)     0
Irregularities Speed (Km/h)             0
Total Irregularities Records            0
Irregularities Geometry                 0
Alert Type                              0
Total Alert Records                     0
dtype: int64
```

Checking Duplicate Rows

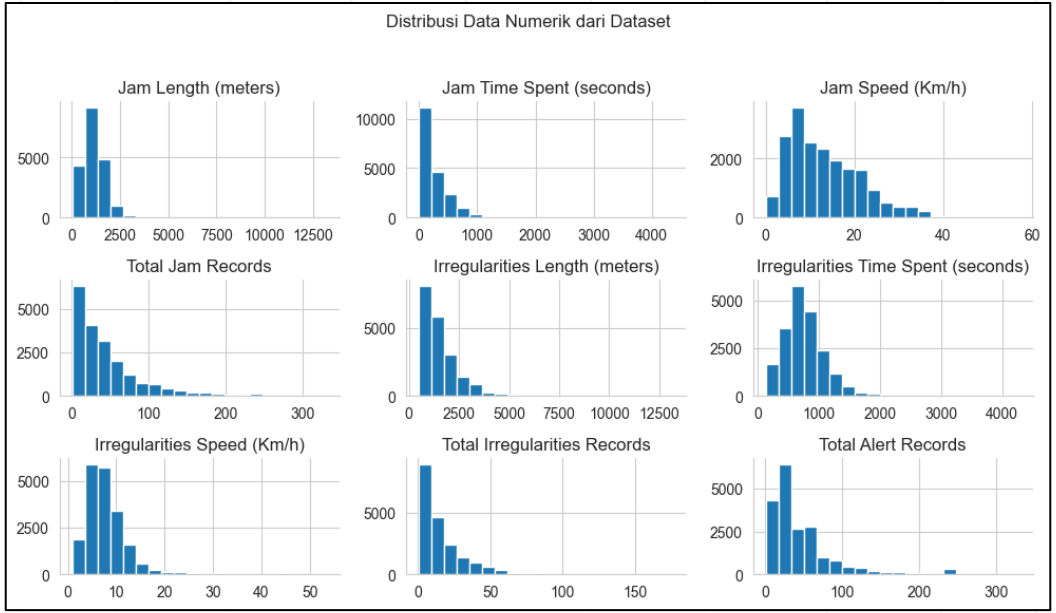
There are no duplicate rows in the main data

```
if True in df_merge.duplicated().tolist():
    print("There are duplicate rows in the data.")
else:
    print("There are no duplicate rows in the data")
```

```
There are no duplicate rows in the data
```

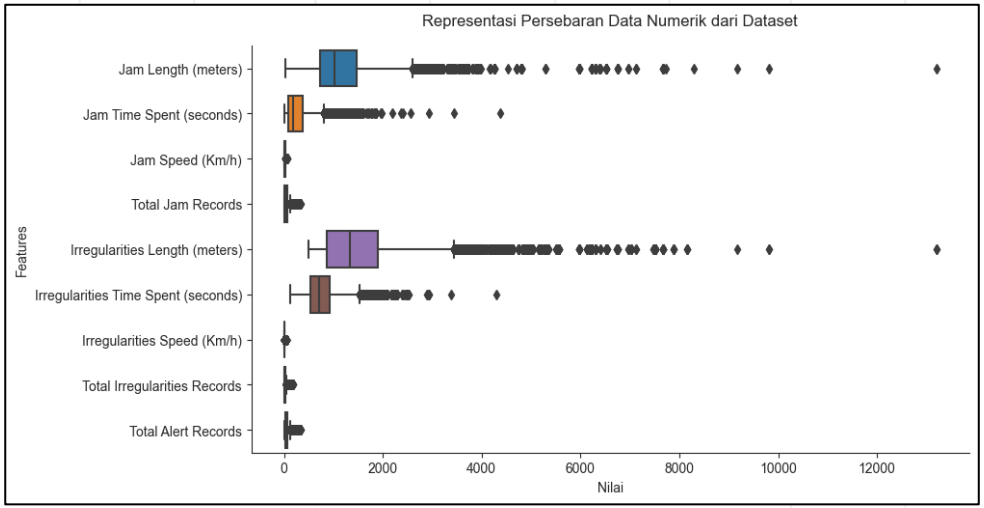
Checking Data Distribution for Numerical Data

Most of the numerical data on the dataset has a skewed distribution.



Checking Outliers in Numerical Data

Most of the dataset has outliers, but the outliers pattern in the Jams and Irregularities columns is almost the same. So there is a possibility that these outliers are not an unusual error or condition, so I will not delete the outliers for now because the outliers can be aware of the conditions in the dataset. But, I will remove the outliers before the Machine Learning Stage.



Checking Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
Jam Length (meters)	19801.0	1145.147291	672.797284	24.00	728.00	1032.000000	1476.00	13201.00
Jam Time Spent (seconds)	19801.0	273.201202	250.995594	-1.00	95.00	179.000000	378.00	4368.50
Jam Speed (Km/h)	19801.0	13.376370	8.367126	0.00	6.69	11.639999	18.67	57.28
Total Jam Records	19801.0	45.598000	46.864158	1.00	13.00	31.000000	60.00	331.00
Irregularities Length (meters)	19801.0	1546.701101	943.741351	500.00	874.00	1339.000000	1900.00	13201.00
Irregularities Time Spent (seconds)	19801.0	759.225191	330.727332	132.00	537.50	720.000000	934.50	4295.50
Irregularities Speed (Km/h)	19801.0	8.001935	4.923337	0.88	5.04	7.060000	9.73	53.49
Total Irregularities Records	19801.0	16.485228	17.235238	1.00	4.00	11.000000	23.00	176.00
Total Alert Records	19801.0	46.739660	44.933235	1.00	20.00	31.000000	60.00	331.00

Checking Negative Values

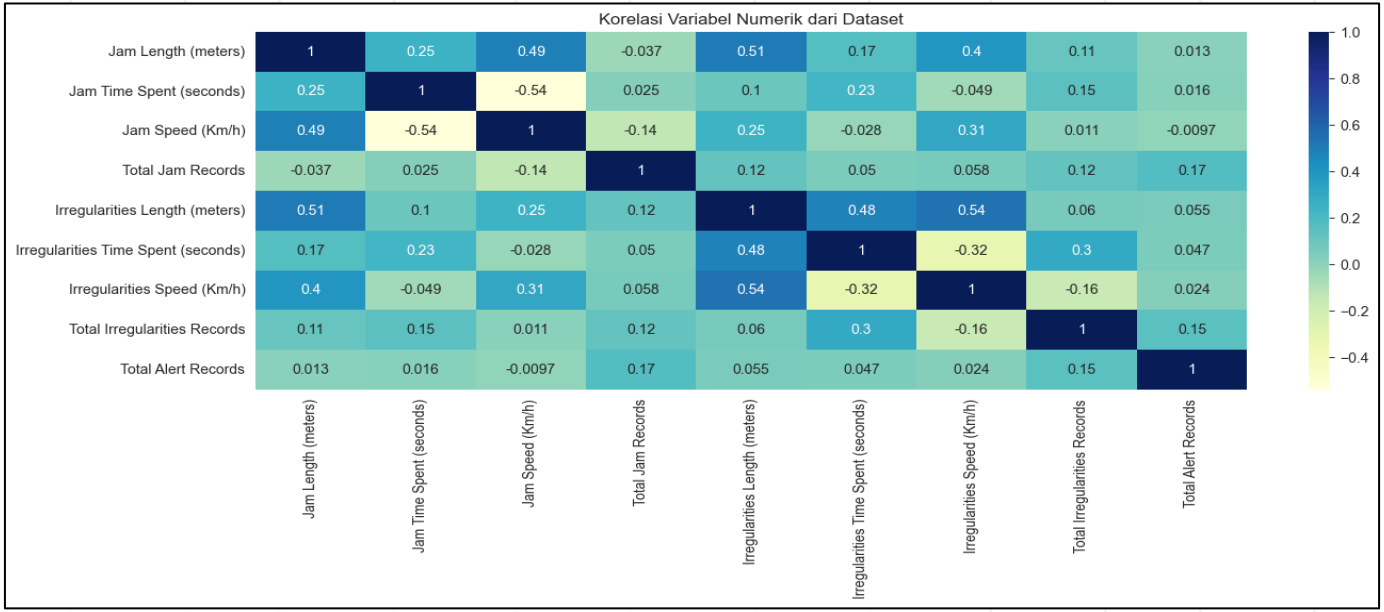
There are many negative values in the Jam Time Spent (seconds) column. This negative value means that traffic in the area does not experience delays or even faster than normal speed. In the context of data delay of hours, negative values indicate the remaining time, which is the difference between the time needed to pass the road and the actual time needed. So that this negative value is not an error or an unusual thing, I do not delete this negative value.

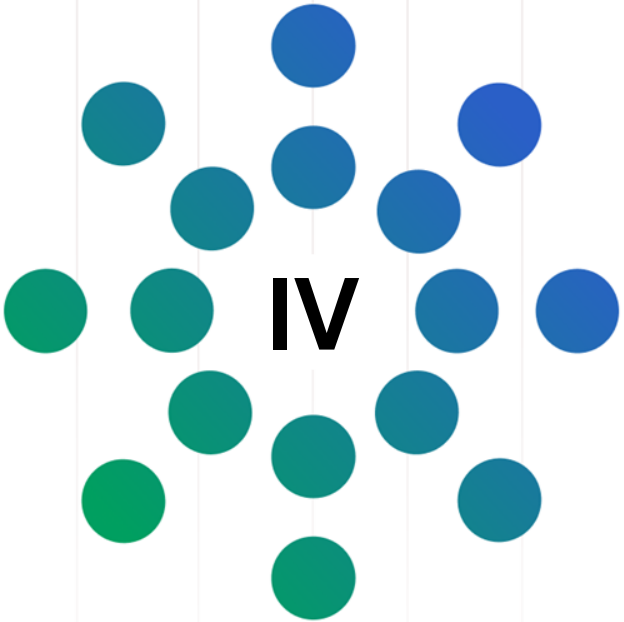
```
negative_values = []
for idx, val in enumerate(df_merge["Jam Time Spent (seconds)"].tolist()):
    if val < 0:
        negative_values.append(val)
print("The size of negative values is {}".format(len(negative_values)))
```

```
The size of negative values is 329
```

Checking Correlation in Numerical Data

The time, distance, and speed column in Jams and Irregularities data appears to have a moderate positive correlation.

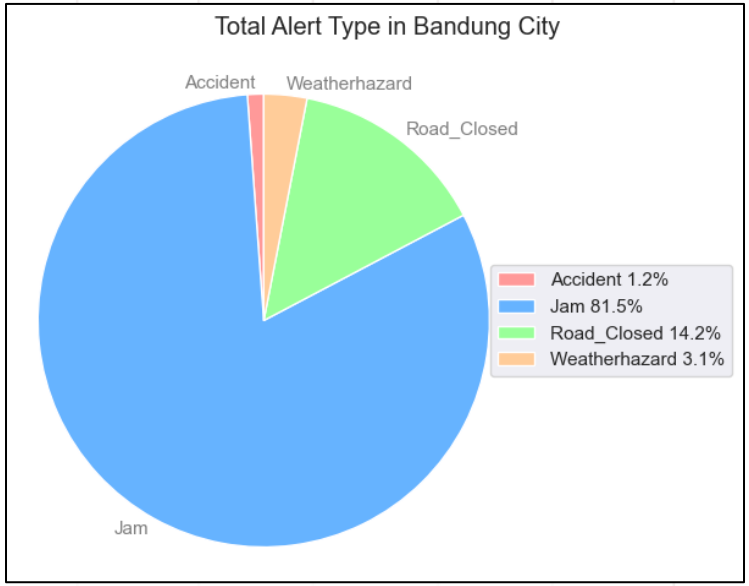




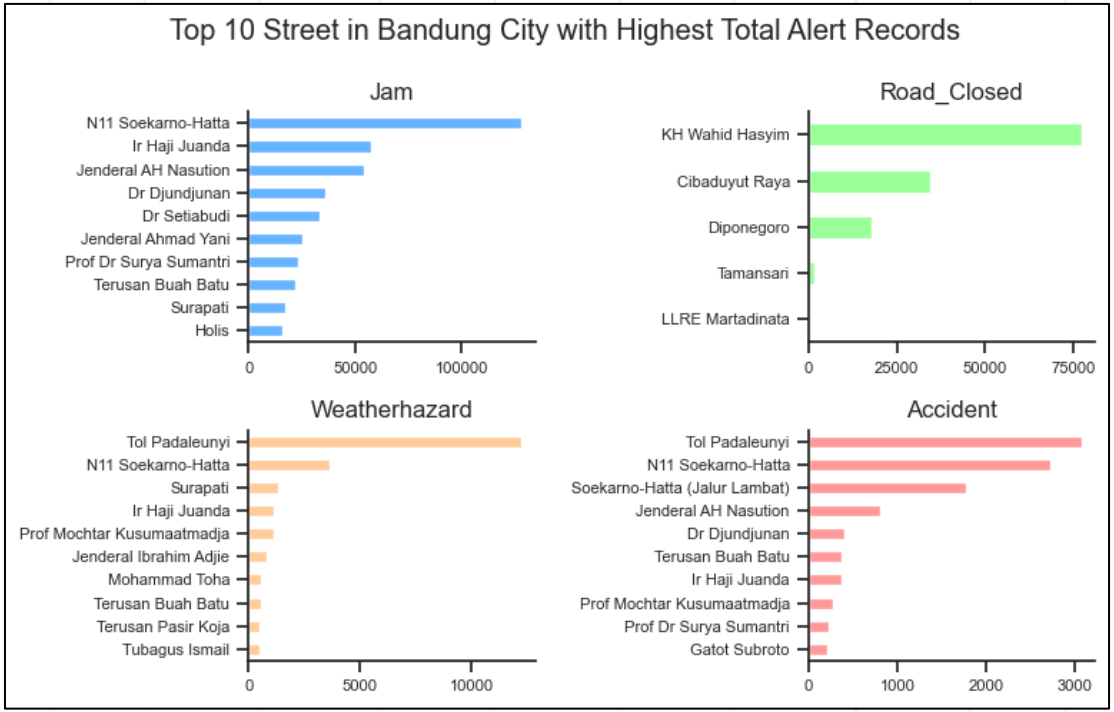
Exploratory Data Analysis (EDA)

Alert Type Proportion in Bandung City

As we can see from the chart below, the most frequently reported Alert Type by drivers in Bandung City is traffic jam.

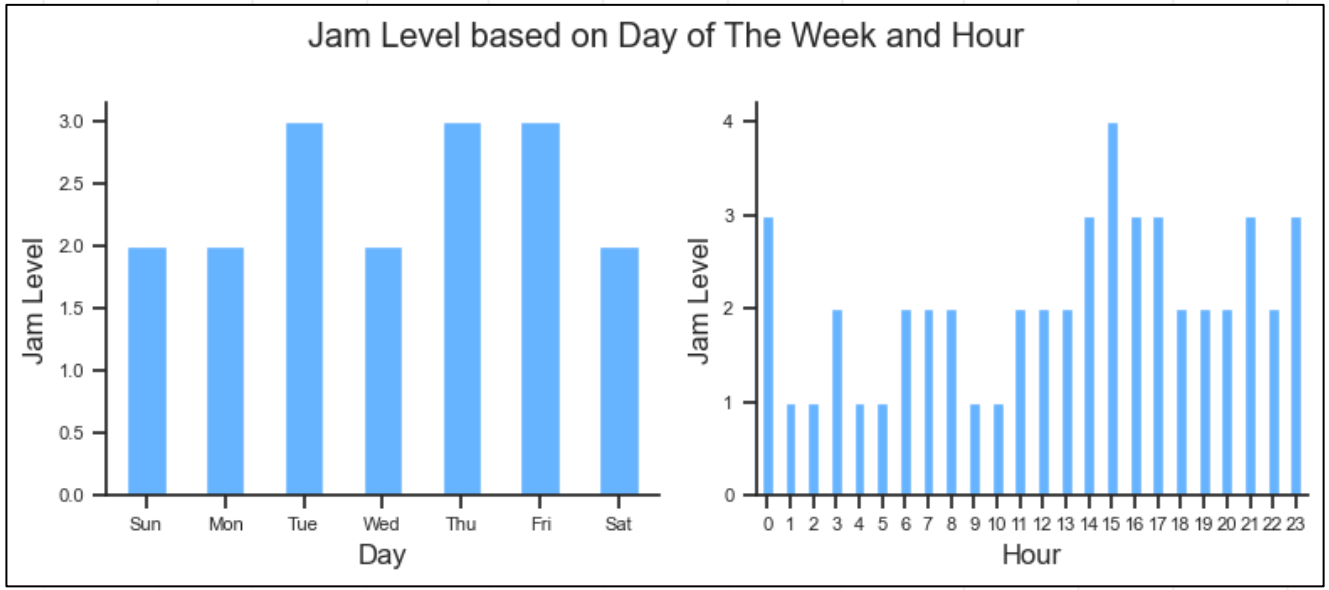


Street in Bandung City with Highest Alert Reports



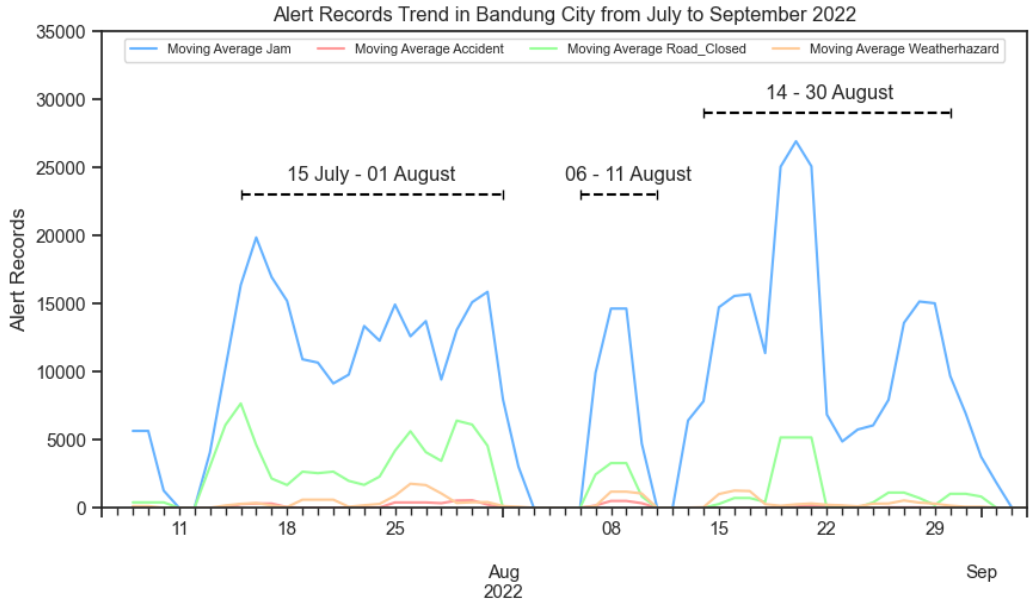
Jam Level Based On Day of The Week & Hour

It is noticeable that the average traffic congestion is quite severe on Tuesdays, Thursdays, and Fridays. Meanwhile, traffic congestion tends to increase from 9 a.m. to 3 p.m.



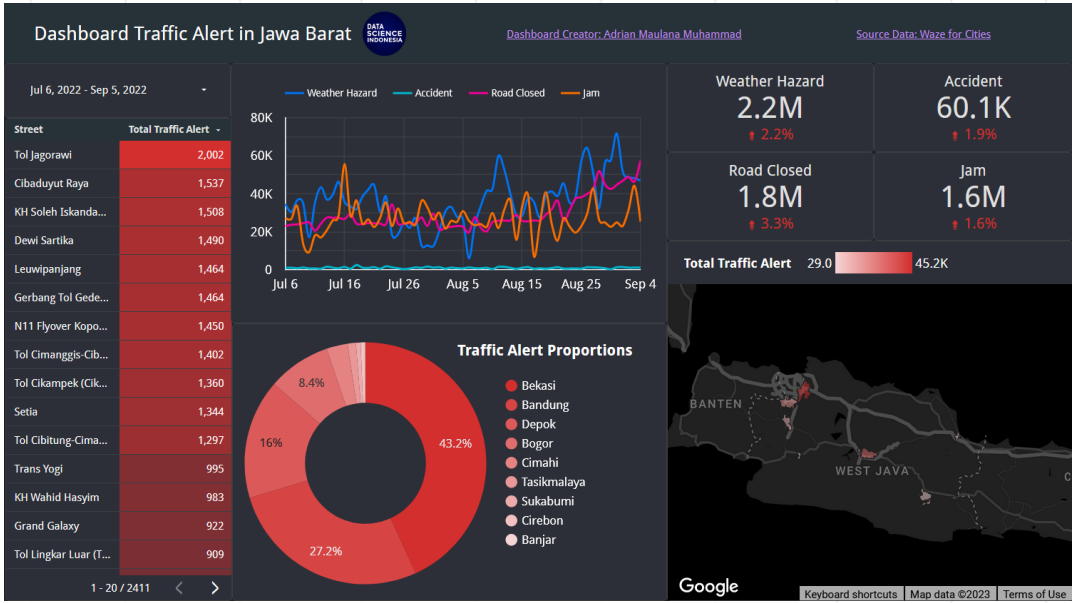
Alert Records Trend

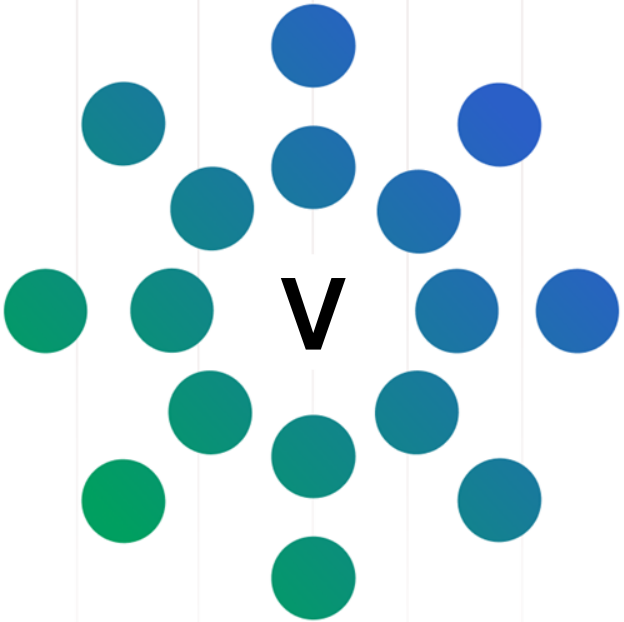
There are three time periods from July to September 2022, where the trend for each alert report has significantly increased. A further diagnostic analysis is needed to determine the causes of these trends during these periods.



Analytics Dashboard

This analytics dashboard shows the proportion of traffic alerts in the city and street of West Java. The local government can use the dashboard to monitor and get insight into their area's traffic condition. The dashboard can be accessed at this link.

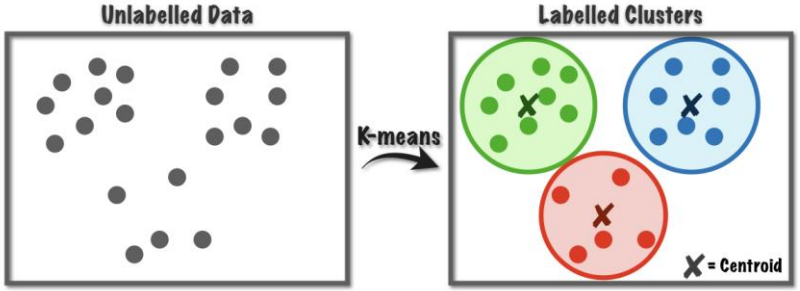




Machine Learning Model – Clustering

K-Means Algorithm

At this stage, we will group or clustering street in Bandung City based on data length, time spent, and speed for irregularities in traffic. The clustering algorithm that I chose is K-Means. K-means clustering is a type of unsupervised learning used with unlabeled datasets. The goal of this algorithm is to find K groups in the data. The algorithm works iteratively to assign each data point to one of K groups based on the provided features. Data points are clustered based on feature similarity.



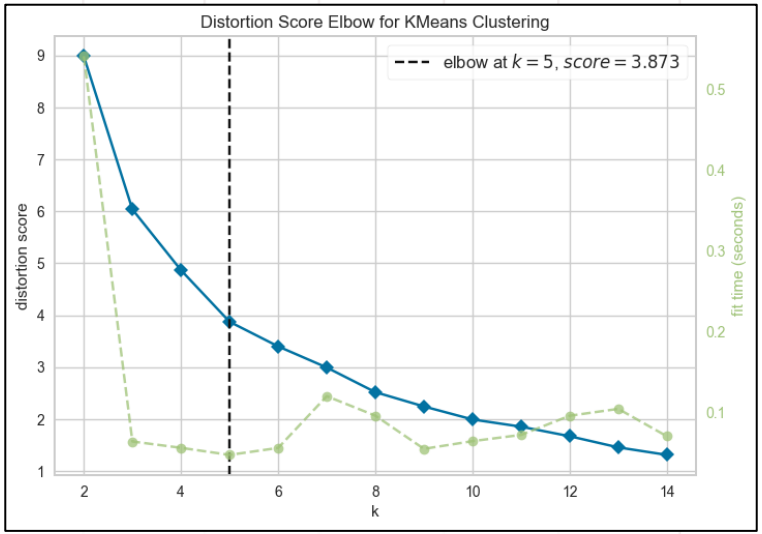
Feature Scaling

Feature scaling is a common preprocessing step in machine learning that is often performed before applying the k-means clustering algorithm. The reason for performing feature scaling is to ensure that all features or variables have equal importance in the clustering process, regardless of their original scale or units of measurement.

```
minMaxScaler = MinMaxScaler()
normalized_df = filtered_df.copy()
normalized_df[normalized_df.columns.tolist()] = minMaxScaler.fit_transform(
    normalized_df[normalized_df.columns.tolist()]
)
```

Elbow Method

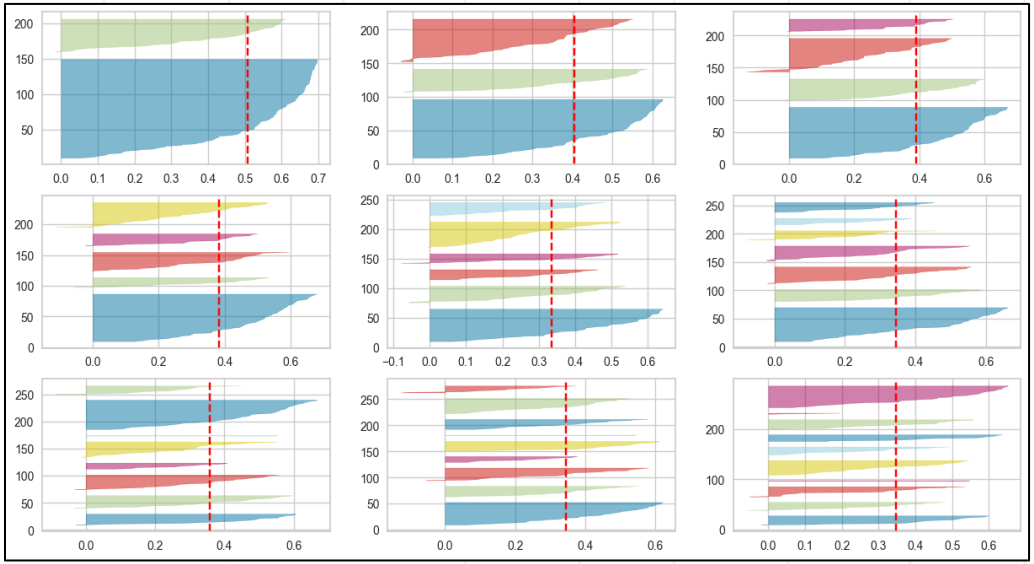
The elbow method is commonly used to determine the optimal number of clusters in K-Means clustering. The optimal number of groups is where the "elbow" of the curve is located in the plot. The below graph selects an Elbow point at K=5. Let's validate the value of K using the Silhouette plot.



Silhouette Score

Silhouette score is a widely used metric for evaluating the quality of clusters in K-Means clustering. The Silhouette score ranges from -1 to 1, where a score of 1 indicates good clustering, and a score of -1 indicates poor clustering.

	n_clusters	average_scores
0	2	0.506845
1	3	0.403974
2	4	0.391234
3	5	0.382560
4	8	0.356966
5	10	0.348226
6	9	0.344923
7	7	0.343975
8	6	0.336116



Choosing K=4

The Silhouette score is maximum (0.50) for K = 2, but that's not sufficient to select the optimal K.

The following conditions should be checked to pick the right 'K' using the Silhouette plots:

- For a particular K, all the clusters should have a Silhouette score more than the average score of the dataset (represented by a red dotted line). The x-axis represents the Silhouette score.
- There should not be wide fluctuations in the size of the clusters. The width of the groups represents the number of data points

The value of 4 for clusters looks optimal. The silhouette score for each group is above average silhouette scores. Also, the thickness is more uniform than the plot with a bunch of 2 or 3, with one cluster thickness much more than the other. **Thus, one can select the optimal number of groups as 4.**

Fitting K-Means with K=4

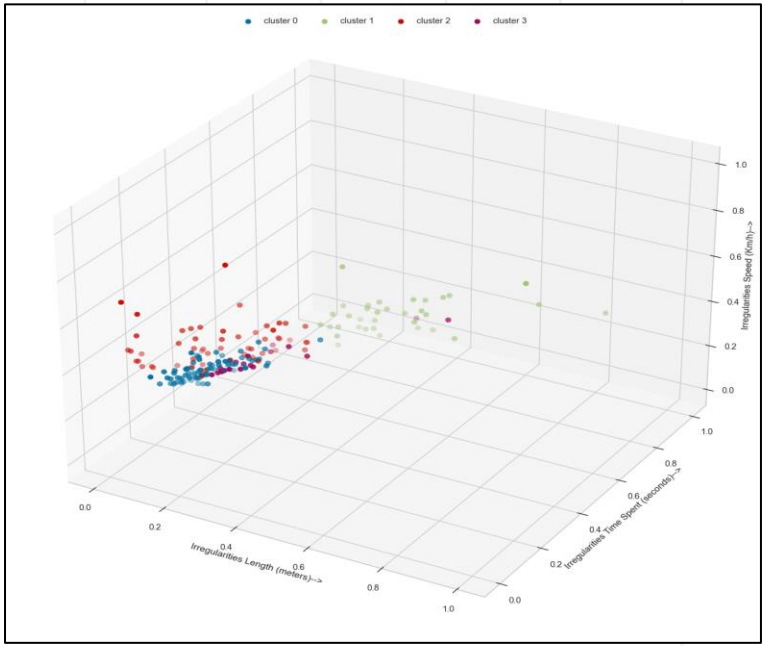
After determining the optimal value of K, the next step is to perform the process of fitting the K-Means model to the features of the data. In this step, the K-Means algorithm is applied to the data, and the model is trained to cluster the data points into K clusters. Once the model is fitted to the data, it can be used to predict the cluster labels of new data points based on their proximity to the centroids.

```
cluster_model = KMeans(n_clusters=4, random_state=42)
preds = cluster_model.fit_predict(X)
filtered_df["K-Means"] = preds

# save the model
pickle.dump(preds, open('../../models/clustering_kmeans.pkl', 'wb'))
```

Clusters Visualization

The previous K-Means cluster prediction results can be visualized in a three-dimensional chart shown below. This visualization allows for a deeper understanding of the clustering structure and pattern of the data, and can serve as a useful tool for exploratory data analysis and feature selection.



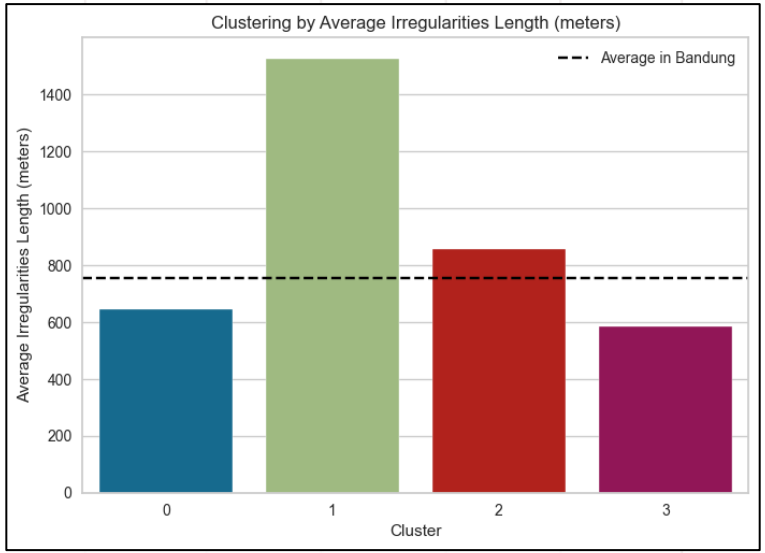
Cluster's Characteristics

The characteristics of each cluster can be observed and analyzed through the table presented below

	Cluster	Total Member	Average Irregularities Length (meters)	Average Irregularities Time Spent (seconds)	Average Irregularities Speed (Km/h)
0	1	34	1523.75	752.50	7.62
1	2	53	856.00	513.00	5.92
2	0	80	644.00	644.00	3.82
3	3	20	583.50	878.25	2.39

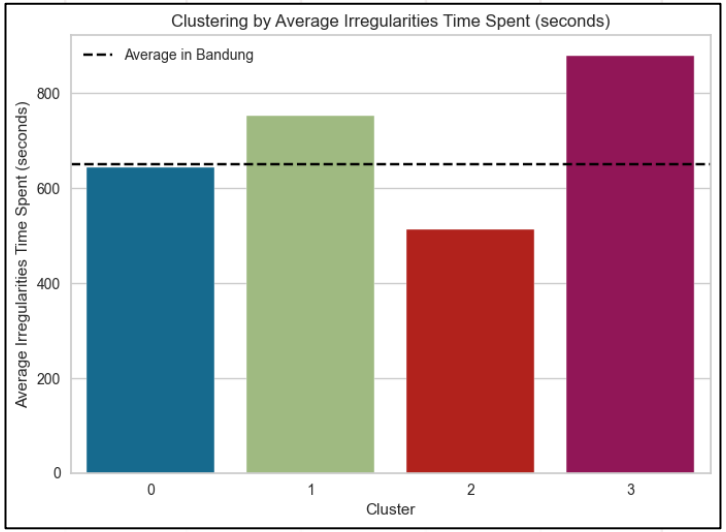
Cluster’s Average Irregularities Length

Based on the average Irregularities Length, only Cluster 1 and Cluster 2 can exceed the overall average in Bandung. This finding suggests that these two clusters may have distinct characteristics and represent areas with higher road irregularities or poor road conditions.



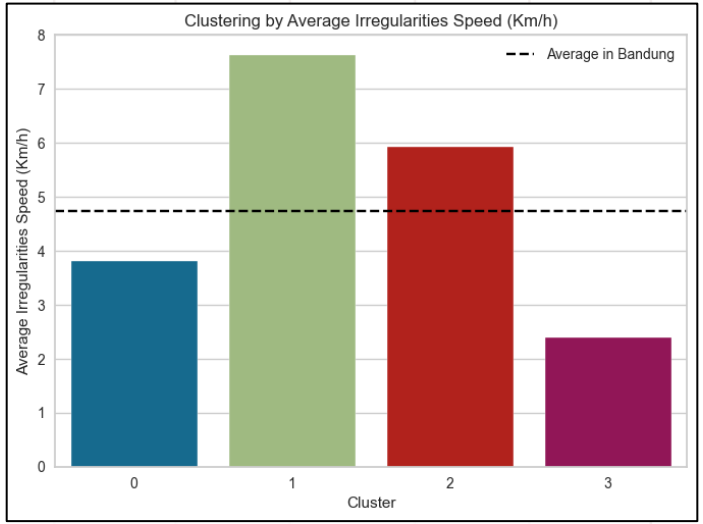
Cluster's Average Irregularities Time Spent

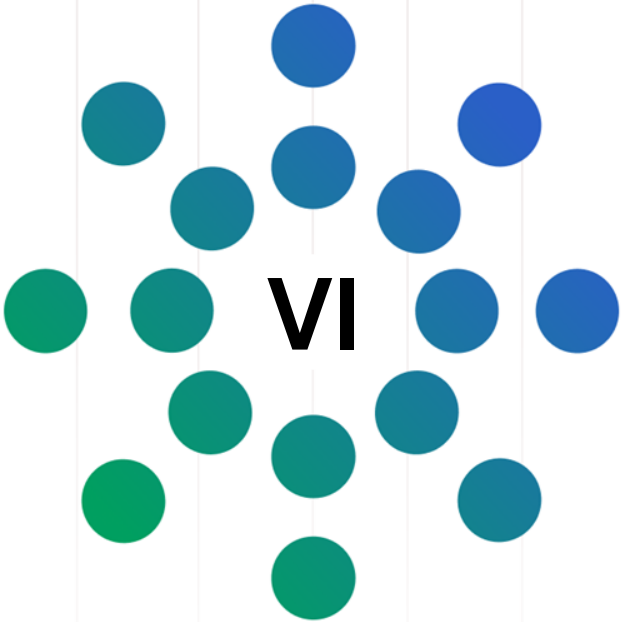
Based on the average Irregularities Time Spent, only Cluster 1 and Cluster 3 can exceed the overall average in Bandung. This finding suggests that these two clusters may have distinctive travel patterns and preferences compared to the other clusters and may represent areas with longer travel times or congestion during peak hours.



Cluster's Average Irregularities Speed

Based on the average Irregularities Speed, only Cluster 1 and Cluster 2 can exceed the overall average in Bandung. This finding suggests that these two clusters may have distinct traffic characteristics and flow patterns compared to the other clusters and may represent areas with higher traffic congestion, bottlenecks, or accidents.





Summaries & Recommendations

Summaries

- Traffic jam is the most frequently reported Alert Type by drivers in Bandung City.
- Traffic congestion is severe on Tuesdays, Thursdays, and Fridays and tends to increase between 9 a.m. to 3 p.m.
- There are three time periods (July to September 2022) where the trend for each alert report has significantly increased, requiring further diagnostic analysis to determine the causes of the trends.
- Only Streets in Cluster 1 and 2 have an average Irregularities Length that exceeds the overall average in Bandung, indicating these clusters may represent areas with higher road irregularities or poor road conditions.
- Only Streets in Cluster 1 and 3 have an average Irregularities Time Spent that exceeds the overall average in Bandung, suggesting these two clusters may represent areas with longer travel times or congestion during peak hours and have distinctive travel patterns and preferences compared to the other clusters.
- Only Streets in Cluster 1 and 2 have an average Irregularities Speed that exceeds the overall average in Bandung. These clusters may represent areas with higher traffic congestion, bottlenecks, or accidents and have distinct traffic characteristics and flow patterns compared to the other groups.

Recommendations

- To manage severe traffic congestion on Tuesdays, Thursdays, and Fridays, stakeholders should consider implementing road network optimization, rerouting, and information dissemination to provide alternative travel options to motorists.
- The three time periods where the trend for each alert report has significantly increased require further diagnostic analysis to determine the causes of the trends. This analysis can provide insight into identifying and addressing potential problems, allowing stakeholders to mitigate future effects of similar issues.
- Streets in Cluster 1 and 2 may represent areas with higher road irregularities or poor road conditions, necessitating an evaluation of the roads' maintenance and repair needs by local authorities and transport infrastructure providers.
- Streets in Cluster 1 and 3 may represent areas with longer travel times or congestion during peak hours. Stakeholders should consider implementing traffic management and control systems to alleviate congestion and improve traffic flow.
- Streets in Cluster 1 and 2 may represent areas with higher traffic congestion, bottlenecks, or accidents, necessitating traffic calming measures, road widening, and accident prevention strategies.

VII References

- https://www.kaggle.com/code/prashant111/k-means-clustering-with-python/notebook
- https://towardsdatascience.com/elbow-method-is-not-sufficient-to-find-best-k-in-k-means-clustering-fc820da0631d
- https://www.kaggle.com/code/shrutimechlearn/step-by-step-kmeans-explained-in-detail
- https://pennmusa.github.io/MUSA_801.io/project_8/index.html
- https://www.arcgis.com/apps/dashboards/d1ee0d06645b49d5a01e3bad3add4c5a

Thank You!