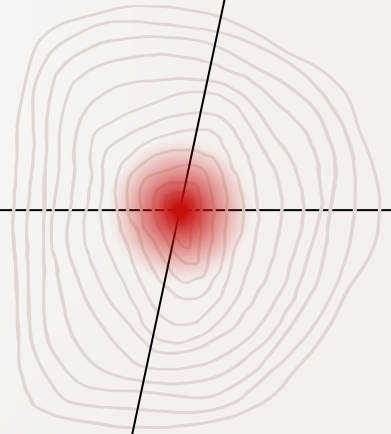
Correlation analysis of Taichung traffic accident database



Group 12

Yi-Fan Chen	M11215052
Ping-He Yeh	M11215032
Yi-Shan Cheng	M11215066
Po-Chun Hu	M11215110

Table of contents

Introduction

Background

In recent years, there has been a surge in traffic accidents, with law enforcement and government efforts to improve roads often criticized by the public for their lackluster results.

It is only through a scientific analysis of accident hotspots, including factors such as road conditions, and driver states in order to identify the underlying causes of accidents that we can genuinely reduce the frequency of accidents.

Dataset

Source

Using the traffic accident dataset from the **Taichung City Police Department** available on the government open data platform.

Period

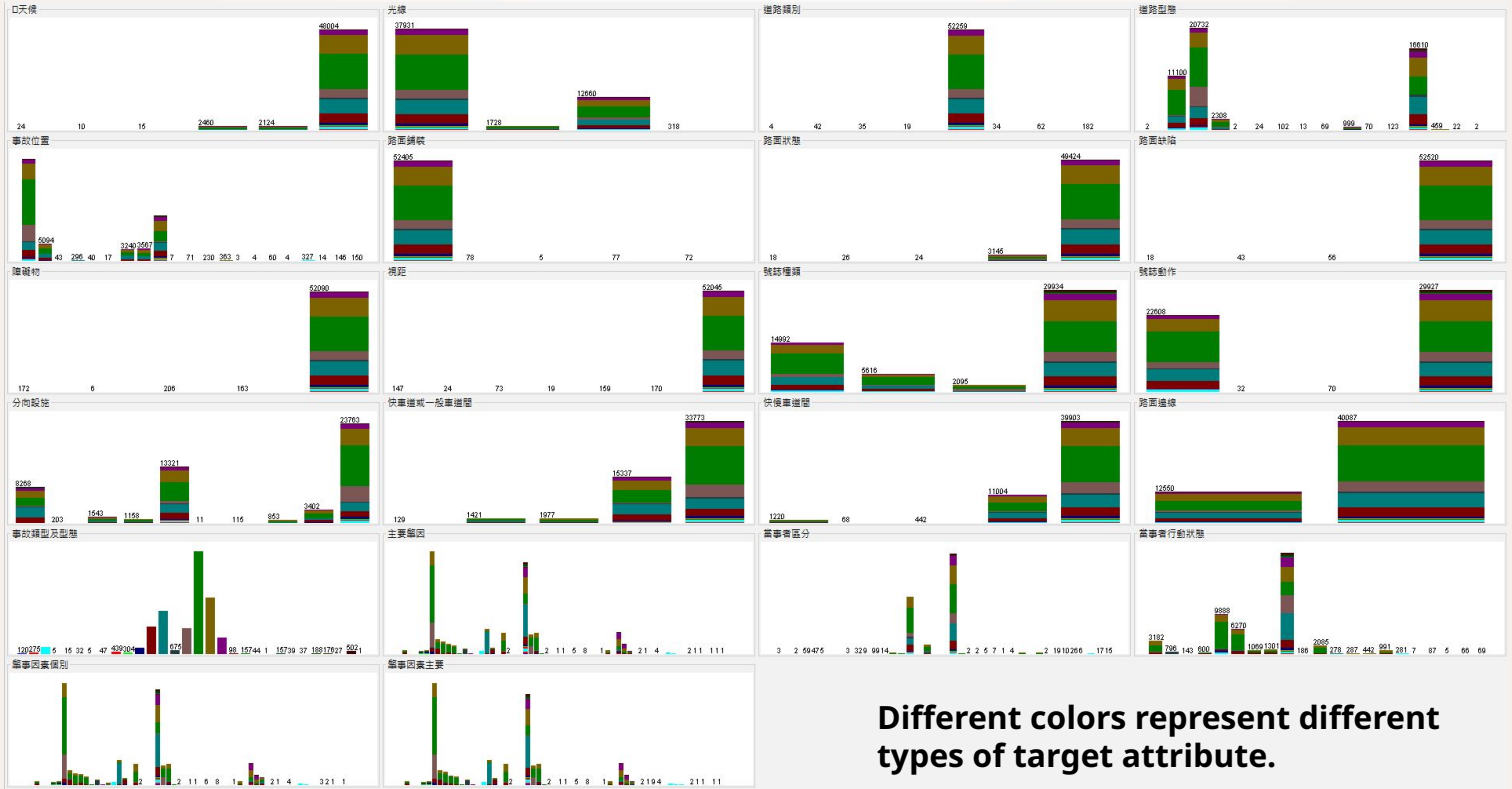
July 2022 to June 2023.

Attributes

There are a total of 55 attributes, with "**Accident Type and Pattern**" being the target attribute.

政府資料開放平台台中市政府警察局交通事故資料集https://reurl.cc/N4azM6

Data distribution map



Different colors represent different types of target attribute.

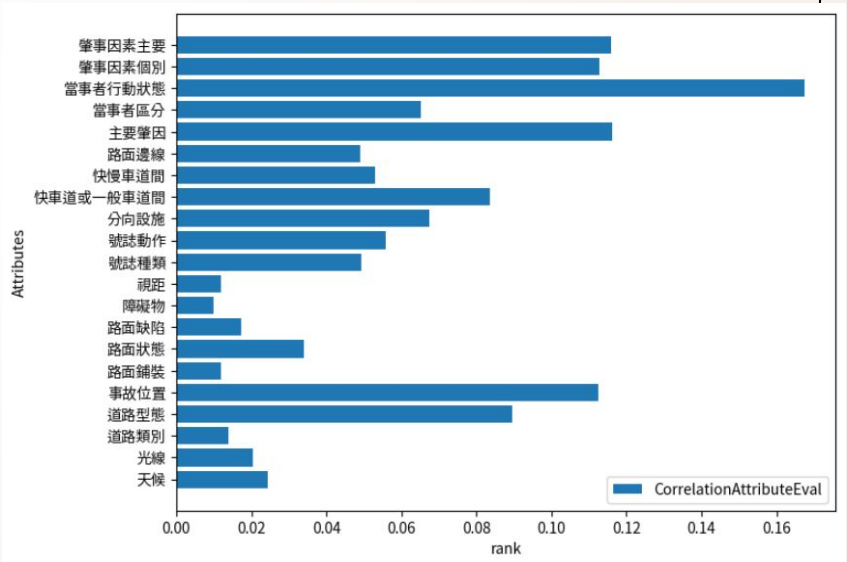
Select Attributes

Recursive feature elimination (RFE)

Only 5 attributes left, and the performance in subsequent Apriori association analysis and K-modes is not satisfactory.

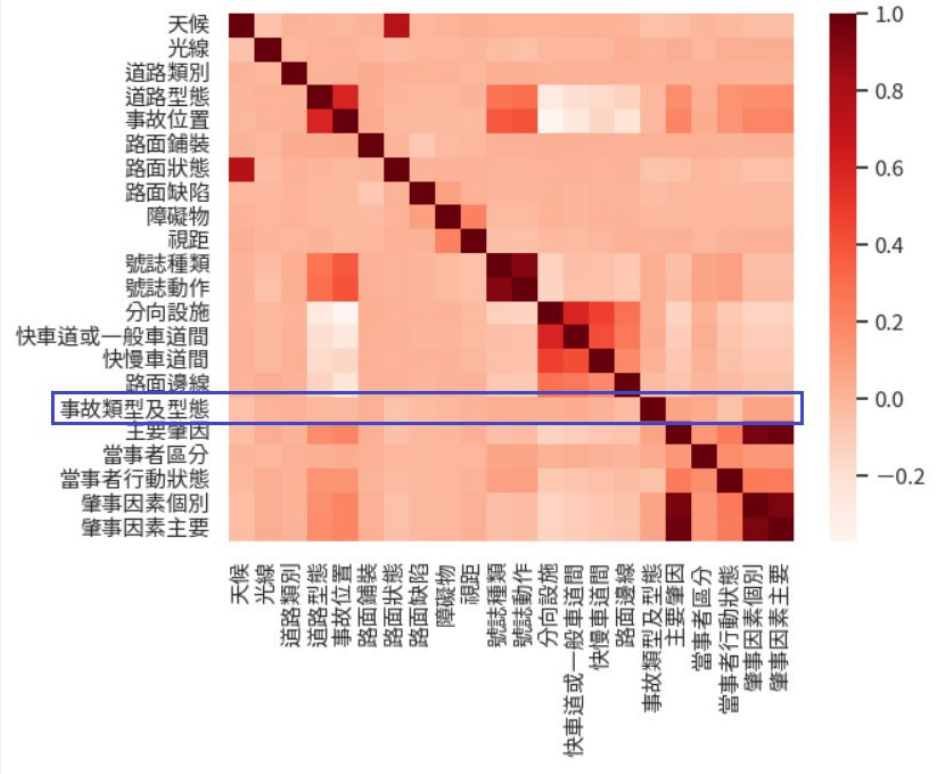
Attribute Evaluation in Weka

We retained the top-performing 9 attributes, but the performance in Apriori and K-modes is also not satisfactory.



CorrelationAttributeEval

Select Attributes



Considering the results of **dataframe.corr()**, all attributes have some impact on the target attribute. Therefore, we are considering using all attributes for association analysis.

Preprocess


Data integration
Remove missing values
Data type conversion
Mapping target attribute

Result & Analytics

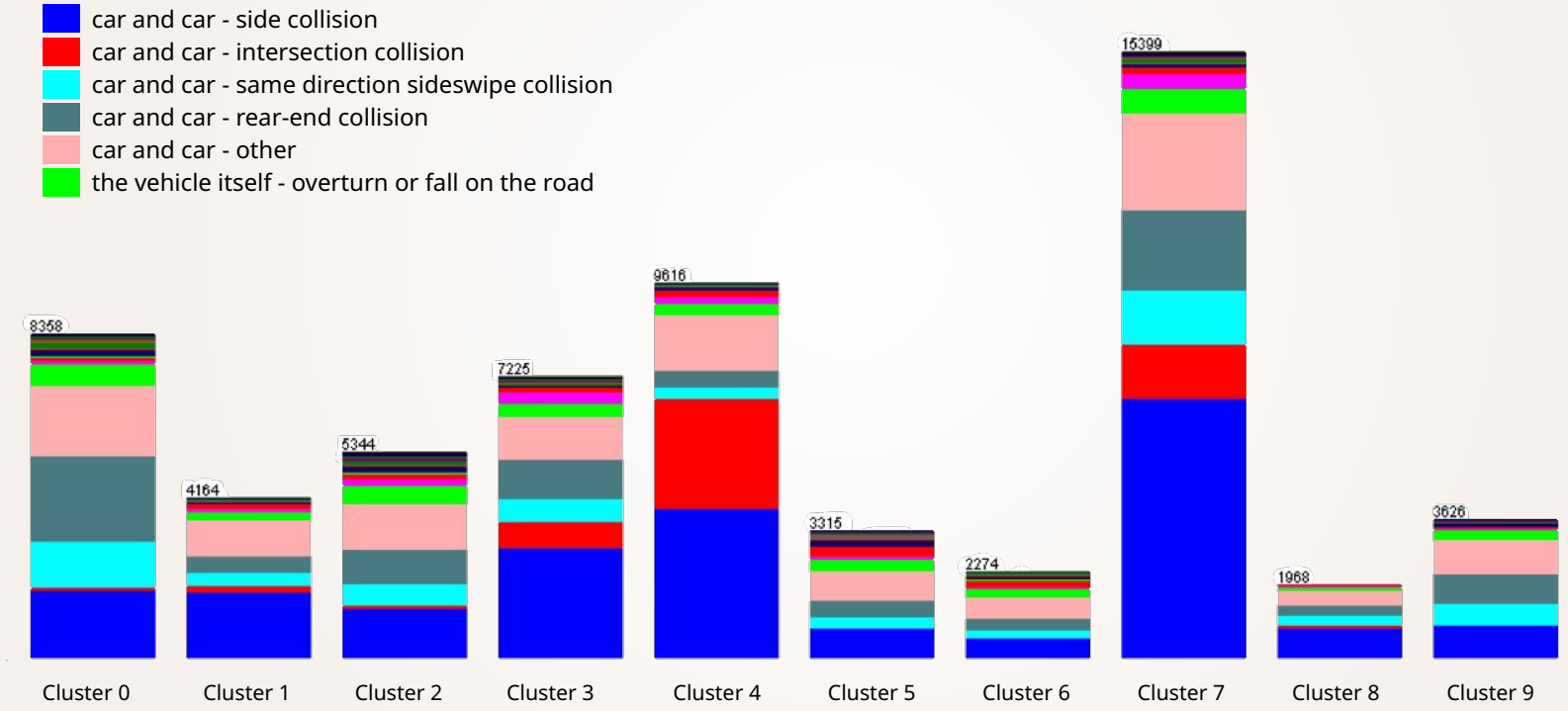
K-modes

K-Modes is a clustering algorithm for categorical data, like text or category labels. It groups data into clusters with similar features, applied in market research and user segmentation.

Key Steps:
1. Initialization: Randomly select K initial centroids.
2. Assignment: Assign each data point to its nearest centroid's cluster.
3. Update: Update centroids using the mode (most frequent category) of data points in each cluster.
4. Repeat: Iterate steps 2 and 3 until convergence criteria are met.

K-Modes efficiently handles discrete features and finds use in market research and user segmentation.

Result & Analytics

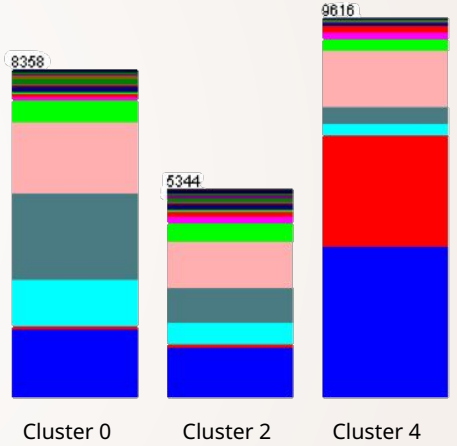


Result & Analytics

	weather	light	road condition	road type	accident location	road defect	obstacle
0	sunny	daytime	dry	urban area	**straight road (without fast or slow lanes)**	none	none
2	sunny	**daytime**	dry	urban area	**three-way intersection**	none	none
4	sunny	**lighted at night**	dry	urban area	**four-way intersection**	none	none

Result & Analytics

	Most	Second	Third
0	car and car rear-end collision	car and car side collision	car and car other
2	car and car side collision	car and car other	car and car rear-end collision
4	car and car side collision	car and car intersection collision	car and car other

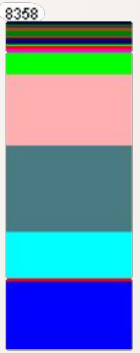


Result & Analytics - cluster 0

	weather	light	road condition	road type	accident location	road defect	obstacle
0	sunny	daytime	dry	urban area	**straight road (without fast or slow lanes)**	none	none

Possible Reasons Analysis:

Due to clear weather, daytime visibility, dry road conditions, and an urban straight road, the accident might be attributed to factors such as driver inattention, speeding, or other individual behavioral factors.



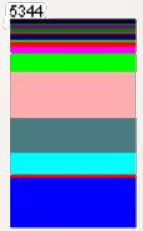
Cluster 0

Result & Analytics - cluster 2

	weather	light	road condition	road type	accident location	road defect	obstacle
2	sunny	**daytime**	dry	urban area	**three-way intersection**	none	none

Possible Reasons Analysis:

Similar to Case 0, with good weather and road conditions, the accident might be due to factors related to a three-way intersection, driver decision-making, or other elements.



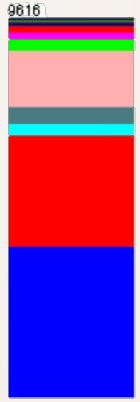
Cluster 2

Result & Analytics - cluster 4

	weather	light	road condition	road type	accident location	road defect	obstacle
4	sunny	**lighted at night**	dry	urban area	**four-way intersection**	none	none

Possible Reasons Analysis:

Unlike the first two cases, this incident occurred at night. Even with street lighting, visibility is still relatively poor compared to daytime, which could lead to drivers noticing pedestrians too late to avoid and subsequently side colliding other vehicles.



Cluster 4

Result & Analytics

- Common possible reasons include:
    - Driver's lack of concentration.
    - Complexity and traffic flow at intersections.
    - Individual driver behaviors such as speeding or non-compliance with traffic rules.

Result & Analytics

Accidents involving overturning and falling off the road

Total number of incidents: 3327

urban roads: 3298 (99.1%), asphalt road: 3277 (98.5%), no obstacles:3275 (98.4%)

Cluster A: 672(22%)

clear weather, nighttime(or in tunnels, underground), well-lit, three-way intersection, near crossroads, no traffic signals, no lane dividers

Cluster B: 541(16%)

rainy weather, daytime, wet road, roundabouts, within the lane

Result & Analytics

Pedestrian struck while crossing the road

Total number of incidents: 1400

urban roads: 3298 (99.1%)、asphalt road: 3277 (98.5%)、
no obstacles:3275 (98.4%)

Cluster A: 251(18%)

daytime, Four-way intersection, within an intersection,
no traffic signals, no lane dividers

Cluster B: 225(16%)

Nighttime (or in tunnels, underground passages, culverts),
within an intersection, traffic light, no lane dividers

Result & Analytics

Collisions in the same direction

Total number of incidents: 6051

weather (clear): 5579 (92.2%), urban roads: 5997 (99.1%), asphalt roads: 6032 (99.7%)

Cluster A: 1052 (17%)

daytime,three-way intersection, within an intersection, no lane dividers, no traffic signals

Cluster B: 527 (9%)

nighttime (or in tunnels, underground passages), four-way intersection, near intersections, separation between fast and slow lanes, traffic control signals (with pedestrian signals).

Conclusion

In our report, we explored various methods for attribute selection, including RFE and attribute evaluation in Weka. Additionally, we utilized K-modes to categorize the data into 10 clusters. We conducted an analysis of the results after clustering to identify attributes correlated with accidents. Furthermore, we delved into potential underlying causes and proposed several improvement measures.

Thanks