SWINBURNE UNIVERSITY OF TECHNOLOGY

MASTER OF DATA SCIENCE

COS80025 – Data Visualisation

Semester 2, 2024

Deliverable 3: Final Report

Word count: 4828

Project facilitator: Mr Mohammad Abuhassan

Submitted by: Thi Ngan Ha Do

Student ID: 103128918

Table of Contents

1.0 Introduction

1.1. Context

For decades, vehicle accidents have been a serious global safety problem because of the injuries and fatalities they inflict (Stewart 2023). Beyond the economic burden, these incidents lead to profound losses for victims' families and affect overall public safety. (Cantillo, 2020). This project aims to analyse historical crash data to identify contributing factors, detect underlying trends, and evaluate the effectiveness of current safety measures, with the goal of raising awareness and improving road safety outcomes.

1.2. Target Audience

Given the widespread concern surrounding car accident analysis, this data visualization project aims to be accessible to individuals from diverse technical backgrounds, enabling them to extract insights without needing specialized domain knowledge. However, while the visualizations have been designed to be as user-friendly as possible, viewers with a basic understanding of statistics will likely find it easier to interpret the data more efficiently.

1.3. Project Objectives

The proposal outlined three primary research areas. However, to establish detailed requirements, it was essential to translate these research topics into clear, actionable requirements and break them down into sub-questions that could be addressed through data visualization. A comparison of these requirements with the project's objectives also involved discussing the various aspects of visualization analysis to ensure alignment with the overall project goals.

	Research topic	Requirement	Question	Reason
1	Identifying major factors contributing to traffic accidents, focusing on driver behaviours, infrastructure, and environmental conditions	Examine the role of driver actions, infrastructure setup, and environmental conditions on crash frequency and severity	1.1. How does driver behaviour contribute to crashes? 1.2. What infrastructure factors affect accident rates? 1.3. How do environmental conditions influence traffic accidents?	Pinpoint key factors and assess safety measures, with the aim of enhancing road safety and awareness
2	Spatio-temporal analysis of crash distribution and police response times.	Analyse the patterns of traffic crashes by location and time, and assess the police	2.1. How do traffic crashes vary over time? 2.2. What are the geographic patterns of crash occurrences?	Identify high-risk areas and time periods and evaluate the effectiveness of police response in

		response times to incidents		mitigating crash impacts
3	Crash Impact Analysis: Vehicle damage and injury outcomes	Examine the factors that most significantly affect the severity of vehicle damage and injury counts in traffic crashes	3.1. What influences the extent of vehicle damage in crashes? 3.2. What factors contribute to the severity of injuries in crashes?	Understand the key contributors to crash severity, providing insights for improving vehicle safety standards and minimizing injuries

Table 1. Revised Research Topics and proposed Sub-questions

2.0 Data

2.1 Data Source

In accordance with the project proposal, the "Traffic Crashes - Crashes" dataset was initially selected for visualization. However, during the exploratory phase, it became evident that the two additional datasets, "Traffic Crashes - People" and "Traffic Crashes - Vehicles," provided valuable information essential for a comprehensive understanding of the topic. Specifically, the vehicle data offered detailed background on the types of vehicles involved in crashes from 2015 to the present, which was crucial for analysing damage (research topic 3) and contributory factors (research topic 1). Likewise, the data on individuals involved in crashes was vital for injury analysis, particularly in examining demographic factors.

All three datasets, sourced from the Chicago Data Portal and last updated on October 13, 2024, were utilized, leading to an exploration of the entire Traffic Crashes database schema, with three main tables (crashes, people, and vehicles) linked by crash record ID, covering the period from 2015 to the present.

2.2 Data Processing

Upon reviewing the three datasets, several data pre-processing and feature selection challenges were detected:

Type	Issues	Relevant Columns	Solution
Data Pre-processing	Missing values	Entries missing over 10% values	Drop rows
		Nominal columns	Fill in missing values with False
	Data types	Date-time columns	Convert mm/dd/yyyy to dd/mm/yyyy format
	Duplicates	Referencing crash_record_id	Drop duplicates
	Files too heavy for Tableau to process (over	Referencing crash_record_id	Downsize files by filtering every 3 rows
			As crash record id is the primary key in the crashes dataset, and foreign keys for the remaining datasets, the data cleaning process made sure that the

	800,000KB in total)		three datasets preserve common ids after being cleaned.
Feature Selection	Unused columns (repetitive information and for underload the data size)	Crashes dataset: lane_cnt, location, intersection_related_i, not_right_of_way_i, street_name, beat_of_occurrence, photos_taken_i, dooring_i, work_zone_i, work_zone_type, workers_present_i	Those columns were manually deleted in Tableau
		Vehicles dataset: cmrc_veh_i, towed_i, fire_i, towed_by, towed_to, area_00_i, area_99_i, cmv_id, commercial_src, gvwr, carrier_name, hazmat_placards_i, un_no, mcs_report_i, idot_permit_no, wide_load_i, trailer1_width, axle_cnt, cargo_body_type	
Calculated Fields	Most common values in categorical columns	Columns: roadway_surface_cond, driver_action, alignment, lighting_condition, roadway_surface_cond, trafficway_type, vehicle_type, vehicle_use, weather_condition	Formula: { FIXED : MAX(IF { FIXED [Lighting Condition] : COUNT([Lighting Condition]) } = { FIXED : MAX({ FIXED [Lighting Condition] : COUNT([Lighting Condition]) }) } THEN [Lighting Condition] END) }
	Top N Filter	crash_record_id, injuries_total	Formula: • Accident Count Filter: RANK_UNIQUE(COUNT([Crash Record Id])) • Injuries Count Filter: RANK_UNIQUE(SUM([Injuries Total]))

	Time of Day Grouped	Crash_date	Formula:
			Time of Day Grouped
			`IF FLOOR(DATEPART('hour', [Crash Date]) / 3) = 0 THEN '00`
			`ELSEIF FLOOR(DATEPART('hour', [Crash Date]) / 3) = 1 THEN`
			`ELSEIF FLOOR(DATEPART('hour', [Crash Date]) / 3) = 2 THEN`
			`ELSEIF FLOOR(DATEPART('hour', [Crash Date]) / 3) = 3 THEN`
			`ELSEIF FLOOR(DATEPART('hour', [Crash Date]) / 3) = 4 THEN`
			`ELSEIF FLOOR(DATEPART('hour', [Crash Date]) / 3) = 5 THEN`
			`ELSEIF FLOOR(DATEPART('hour', [Crash Date]) / 3) = 6 THEN`
			`ELSEIF FLOOR(DATEPART('hour', [Crash Date]) / 3) = 7 THEN`
			`END`
			The calculation is valid. 3 Dependencies Apply OK

	Accident Severity	Damage, injuries_total	Formula:
			Accident Severity
			`IF [Injuries Total] = 0 AND [Damage] = "$500 OR LESS" THEN`
			`ELSEIF [Injuries Total] >= 1 AND [Injuries Total] <= 5 AND`
			`ELSEIF [Injuries Total] >= 10 OR [Damage] = "OVER $1,500"`
			`ELSE 'Moderate'`
			`END`
			The calculation is valid. 9 Dependencies Apply OK

	Count occurrence of categorical values (%)	injuries_total, injuries_fatal, injuries_incapacitating, injuries_non_incapacitating	Formula: COUNT(IF [Injuries Incapacitating] >= 1 AND [Injuries Fatal] = 0 THEN 1 END) / COUNT([Crash Record Id]) * 100
	Vehicle Age at Crash	Crash_date	Formula: YEAR([CRASH DATE]) - [Vehicle Year]

	Vehicle Age Bins	Vehicle Age at Crash	Vehicle Age Group
			`IF [Vehicle Age at Crash] <= 5 THEN "New"`
			`ELSEIF [Vehicle Age at Crash] <= 10 THEN "Middle-Aged"`
			`ELSEIF [Vehicle Age at Crash] <= 15 THEN "Aging"`
			`ELSEIF [Vehicle Age at Crash] <= 20 THEN "Old"`
			`ELSE NULL`
			`END`
			The calculation is valid. 3 Dependencies Apply OK

Table 2. List of Data Processing issues, Feature Selection and Feature Creation

The data cleaning process was performed using a Google Colab notebook. To run this file, Python must be installed, or alternatively, the script can be executed on local Python notebook platforms such as Anaconda. The input data consists of the original files from the City of Chicago website, and the output is a set of cleaned and downsized datasets. Calculated fields and feature selection were handled in Tableau, with the corresponding formulas and explanations provided in the table above.

3.0 Requirements

3.1 Must-Have Features

Question	Visualisation Name	Data set Type	Data source	Attribute Name	Data Type
Topic 1					
1.1 1.2 1.3	Overview of Contributory Factors in Traffic Incidents	Table	Crashes dataset, vehicles dataset	Weather_condition, lighting_condition, driver_vision, trafficway_type, alignment, roadway_surface_cond, vehicle_use, vehicle_type, vehicle_id, person_id, crash_record_id	Categorical, quantitative
1.2	Infrastructure Conditions by Trafficway Type	Table	Crashes dataset, people dataset	trafficway_type, traffic_control_device, device_condition, crash_record_id	Categorical
1.1	Driver Actions Contributing to Accidents by Gender	Table	Crashes dataset, people dataset	Sex, driver_action, person_id	Categorical
1.1	Top Behavioural Contributory Causes Leading to Crashes	Table	Crashes dataset	prim_contributory_cause, sec_contributory_cause, crash_record_id	Categorical
Topic 2					
2.1	Crash Frequency by Time and Day of the Week	Table	Crashes dataset	Crash_date, Time of Day Grouped (calculated field), crash_record_id	Time series, categorical
2.1	Emergency Response Times by Crash Severity and Report Status	Table	Crashes dataset	Crash_date, Accident Severity (calculated field), report_type	Time series, categorical
2.2	Spatial Distribution of Crashes in Illinois	Table	Crashes dataset, people dataset	Longitude, latitude, city, state, zipcode, crash_record_id	Geographical, categorical
2.2	Monthly Distribution of	Table	People dataset,	Crash_date, person_id, vehicle_id	Time series, quantitative

	People and Vehicles Involved in Crashes in Chicago		vehicles dataset		
Topic 3					
3.1 3.2	Damage Distribution by Cost and Injury Severity	Table	Crashes dataset	Damage, injuries_total, crash_record_id	Categorical, quantitative
3.2	Injury Distribution by Behavioural Factors	Table	Crashes dataset, people dataset	Driver_action, injuries_total, injury_classification, crash_record_id	Categorical, quantitative
3.2	Age Distribution and Injury Types	Table	Crashes dataset, people dataset	% of Injury Types (calculated fields), age, crash_record_id	Quantitative
3.1	Vehicle Age and Type Characteristics by Damage Cost and Crash Type	Table	Crashes dataset, vehicles dataset	Crash_type, vehicle age group (calculated field), damage, vehicle_type	Categorical, quantitative

Table 3. List of must-have fields for visualisations

4.0 Visualisation Design

4.1 Topic 1: Contributory Factors Analysis: Driver Behaviour, Infrastructure, and Environmental Conditions

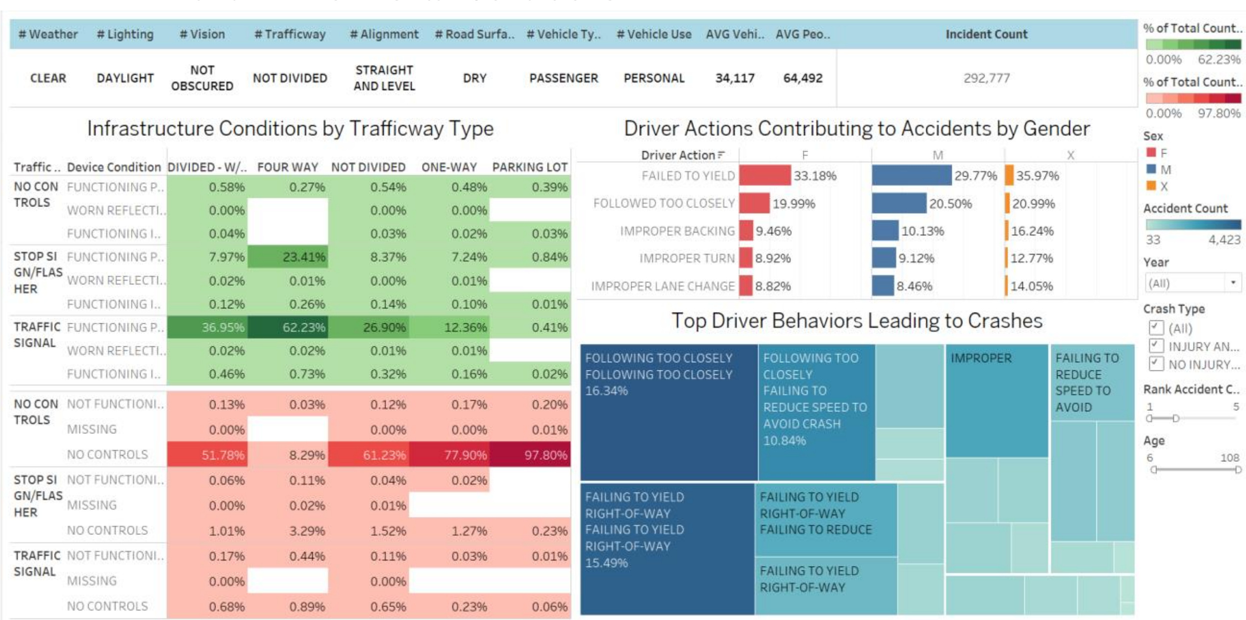


Figure 1. Dashboard 1

4.1.1. Chart Analysis

Cards - Overview of Contributory Factors in Traffic Incidents

# Weather	# Lighting	# Vision	# Trafficway	# Alignment	# Road Surfa..	# Vehicle Ty..	# Vehicle Use	AVG Vehi..	AVG Peo..	Incident Count
CLEAR	DAYLIGHT	NOT OBSCURED	NOT DIVIDED	STRAIGHT AND LEVEL	DRY	PASSENGER	PERSONAL	34,117	64,492	292,777

Figure 2 provides an overview of environmental and infrastructure conditions, vehicle usage, and average count of people involved in crashes.

Attributes	Data type	Mark	Channel	Encoding
Weather, Lighting, Vision, Trafficway, Alignment, Road Surface, Vehicle Type, Vehicle Use	Categorical	Text Labels	Aligned horizontally to represent each specific field, presented in a tabular format for comparison	The categorical fields are displayed as text, representing the most frequently occurring values
Average Vehicle Count, Average People Count, Total Incident Count	Quantitative			The average counts of vehicles and people and the total incident count are shown in numerical form

Table 4. Data encoding Chart 1

Heatmap - Infrastructure Conditions by Trafficway Type

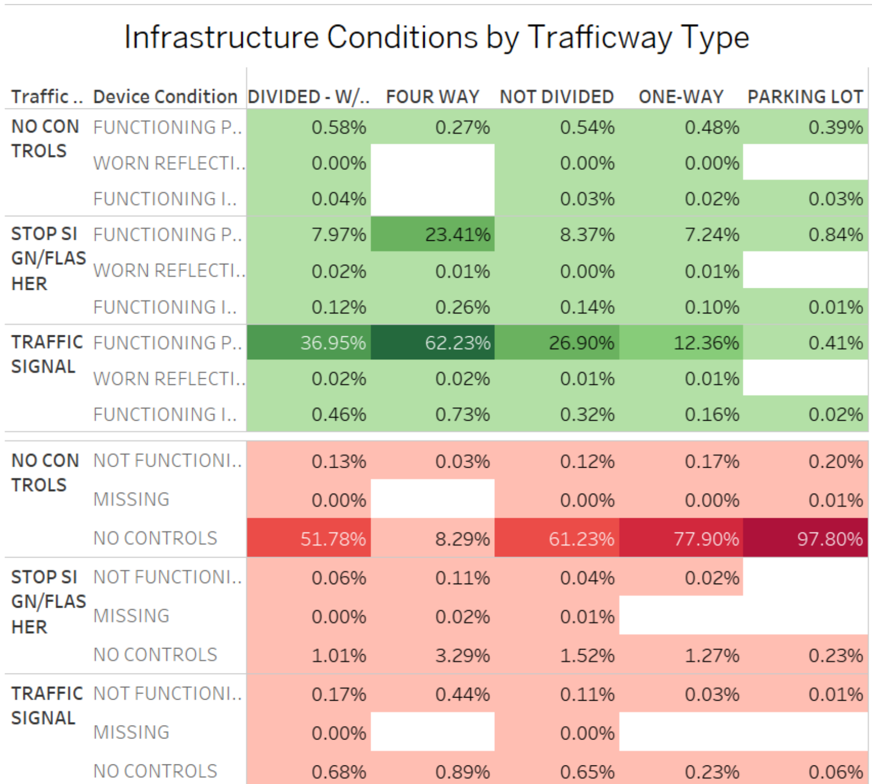


Figure 3 visualises the performance of various traffic control devices under different trafficway configurations

Attributes	Data type	Mark	Channel	Encoding
Trafficway Type	Categorical	Rectangles	Positions along the x-axis	Encoded using positions on the x-axis, the order is shown in the original data source. Filtered by top 5 trafficway types with highest crash counts
Traffic Control Device	Categorical		Positions along the y-axis	Encoded using positions along the y-axis to show different control devices
Device Condition	Categorical		Secondary categorization on the y-axis. Colour	ncoded using different colors to distinguish between functioning (green) and non-functioning conditions (red)
Percentage of Incidents/Conditions	Quantitative	Text labels	Colour intensity	Encoded using different shades of colour, where darker shades indicate higher percentages

Table 5. Data encoding Chart 2

Bar Chart - Driver Actions Contributing to Accidents by Gender

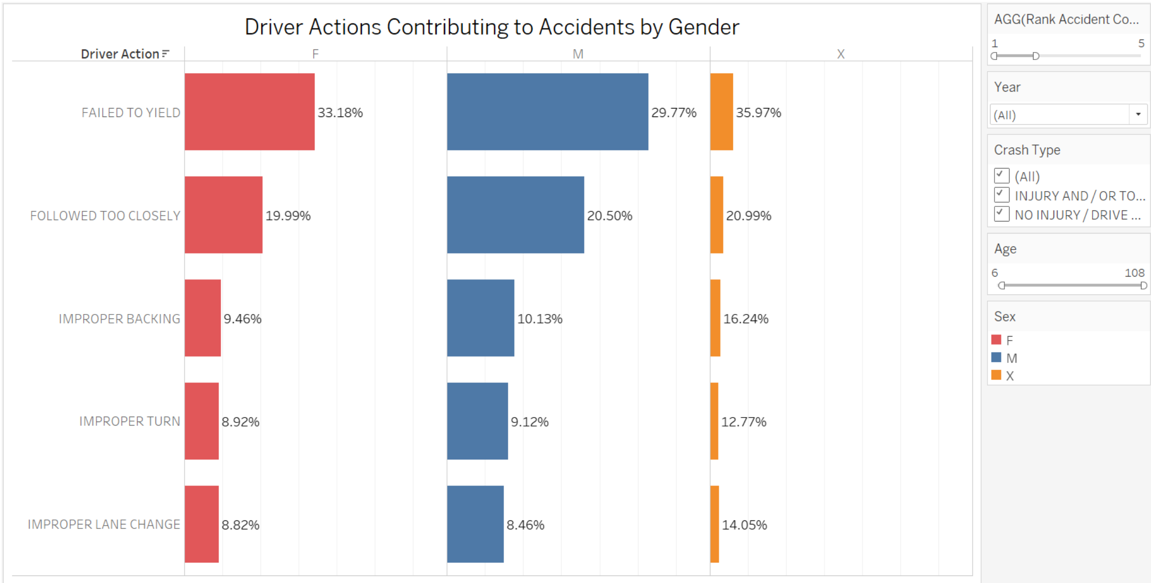


Figure 4 highlights the most common driver behaviours leading to accidents, categorised by gender

Attributes	Data type	Mark	Channel	Encoding
Driver Action	Categorical	Bars	Positions along the y-axis	Encoded using positions along the y-axis to differentiate types of driver actions. Additional filter to display top N driver actions with highest accident count
Sex	Categorical	Bars	Positions along the x-axis. Colours coded	Encoded using positions along the x-axis and colour to differentiate between gender groups
Percentage of Accidents	Quantitative	Bar length	Length of bars. Text labels	Encoded using the length of the bars and text labels to represent exact accident percentages

Table 6. Data encoding Chart 3

Tree map - Top Behavioural Contributory Causes Leading to Crashes

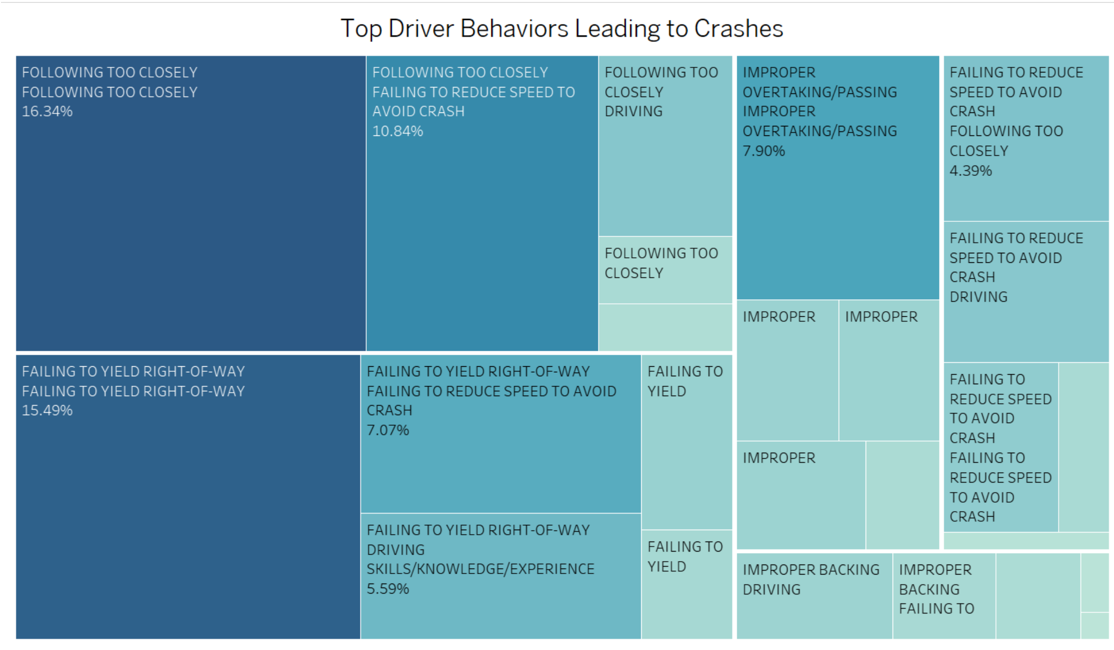


Figure 5 visualises the most frequent driver behaviours contributing to accidents

Attributes	Data type	Mark	Channel	Encoding
Primary Contributory Cause	Categorical	Rectangles	Positions within the tree map	Encoded using the text labels inside the rectangles to differentiate contributory causes (driver behaviours)
Secondary Contributory Cause	Categorical	Rectangles	Positions within the treemap	
Percentage of Accidents	Quantitative	Bar length	Size of the rectangles. Colour coded	Encoded using the size of the rectangles and text labels to represent the percentage of crashes caused by each behaviour

Table 7. Data encoding Chart 4

Chart Analysis

Visualisation	Chart Type	Insight
Figure 2	The cards displaying most occurred categorical values were chosen to display a wide range of data fields at once. Despite its limitations on exploring further details, it provides a grand picture of categorical fields	Driver behaviour and infrastructure design play a larger role than adverse environmental conditions Environment: Most crashes occurred under clear weather conditions during daylight, with no vision obstructions and on dry, straight, and level roads. Infrastructure: The crashes primarily took place on undivided roads.

		Driver behaviour: Passenger vehicles used for personal purposes were most frequently involved.
Figure 3	Heatmap was chosen to compare two categorical fields (trafficway type and device condition) for identification of patterns and outliers	Positive traffic control devices, especially at critical junctions like four-way intersections and divided roads, play a crucial role in reducing accidents, while areas with no or malfunctioning controls see significantly higher accident rates
Figure 4	Side-by-side bar chart was selected to visualise the relationship between driver actions and accident counts, sliced by genders for comparison between categories	Males are involved in most accidents. Failing to yield and following too closely are leading driver actions to accidents across all genders, with slight variations in the proportion of crashes caused by each behaviour
Figure 5	Tree map was selected to compare primary and secondary contributory causes (hierarchical categorical data) that lead to crashes	Most common contributory cause is following too closely. Secondary cause, failing to reduce speed to avoid a crash, exacerbates the impact of behaviour, contributing to an additional 10.84% of incidents

Table 8. Dashboard 1 Analysis

4.1.2. Visualisation Design

Gestalt's data visualization design principles focus on how individuals interpret visual elements, emphasizing concepts like proximity, similarity, and continuity to help users identify patterns and relationships in the data (Ye, Xue & Lin 2021). While this data visualization project was built with these principles in mind, it's essential to evaluate the design elements to assess how well the dashboards communicate information and achieve aesthetic appeal.

	Principle	**Critique**
1	Proximity	Grouping of related visualisations: The row of cards placed on top provides an overall understanding to common factors in crashes. Then, the dashboard is split into two sections, with the left exploring infrastructure conditions, and the right diving into driver behaviours (actions and contributory causes)
2	Similarity	Intuitive colours are used across the dashboard (shades of blue, red, and orange were used to represent gender categories; shades of green and red were used to display the positive/negative conditions of traffic devices)
3	Enclosure	Unclear enclosure, missing bounding boxes around groups of related visualizations, which makes it harder for viewers to understand how the visualizations are related
4	Continuity	Unclear continuity between the charts. Viewers need to read context to follow the contents
5	Figure-Ground	The dashboard's background is light and does not distract from the figures

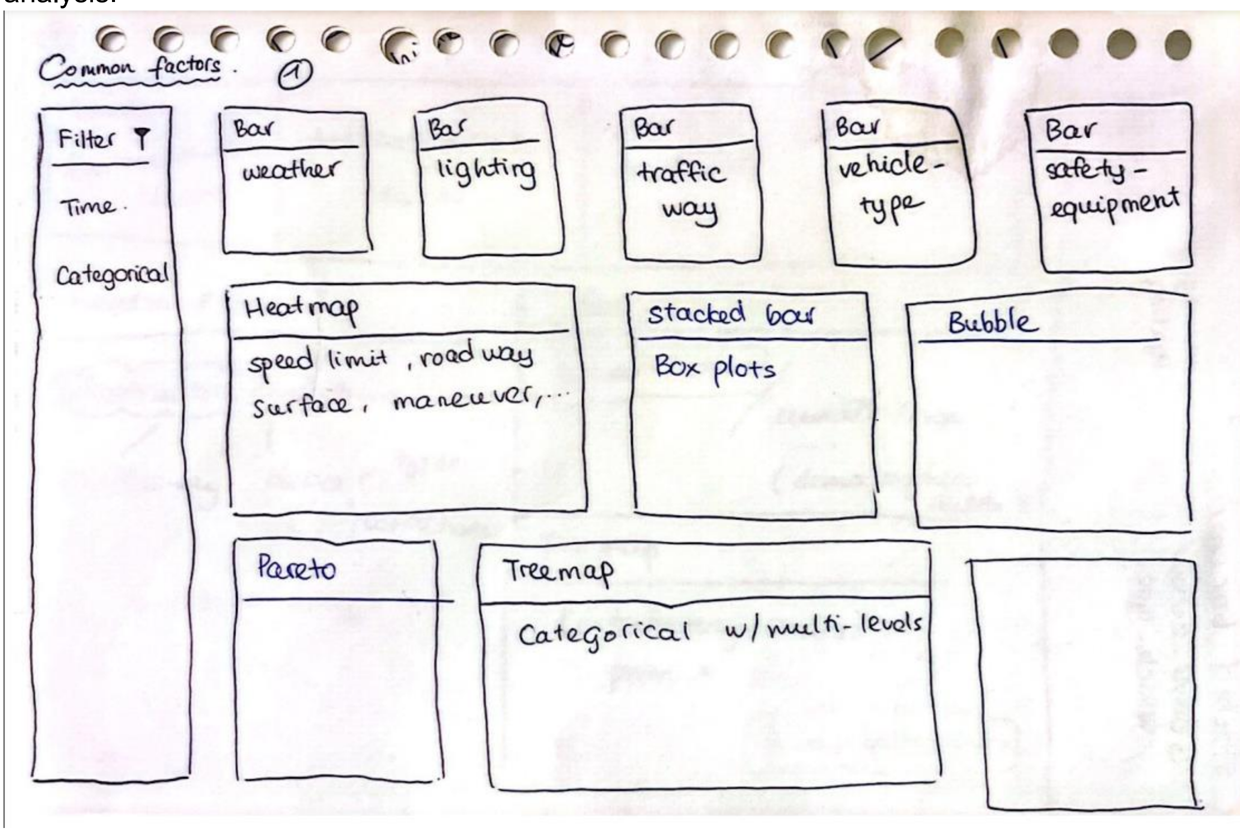
6	Closure	Overwhelming with too much detailed data in the heatmap. Should replace with on-hover details and categories grouping
7	Symmetry and Order	The dashboard displays balance in the layout, the heatmap is cramped because of excessive information

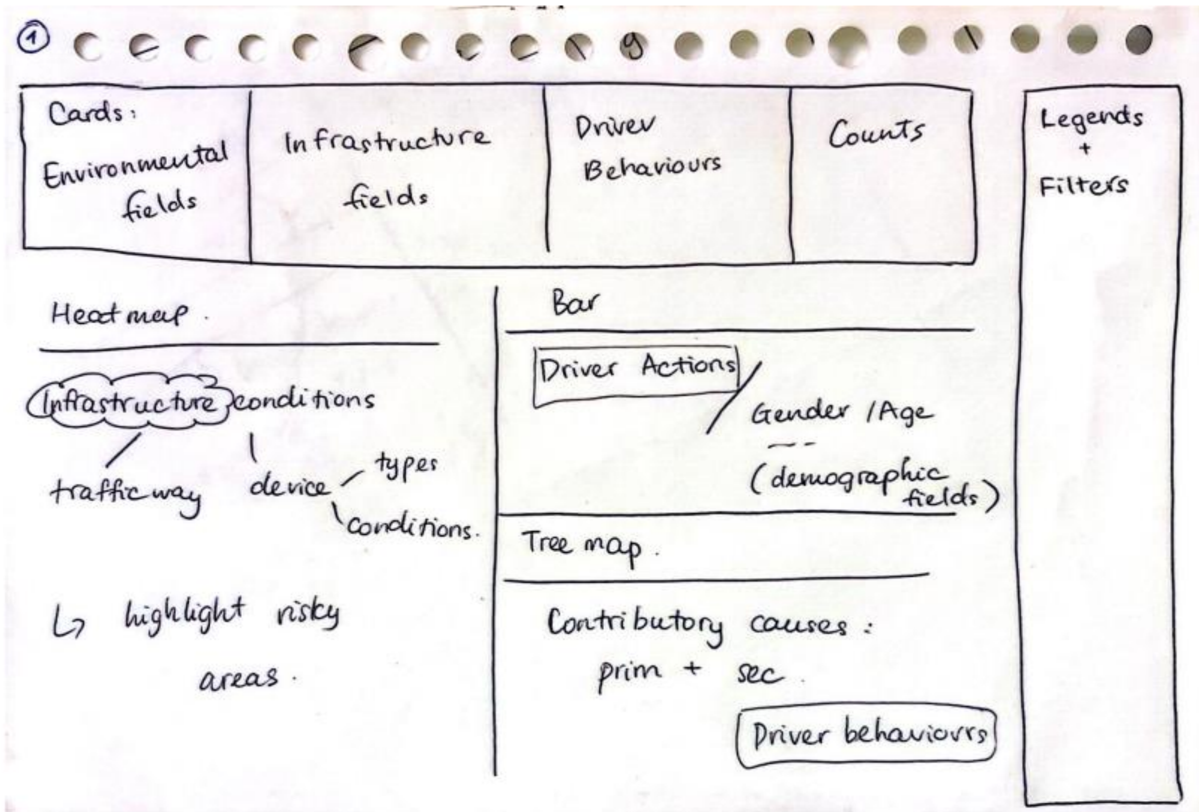
Table 9. Design principles Evaluation of Dashboard 1

The dashboard incorporates interactive features such as filters for crash types, age, and year, enabling users to explore the data more thoroughly. A filter for selecting the top N driver actions with the highest accident counts is particularly added to address the issue of limited space for displaying multiple data fields. Additionally, users can hover over elements to view detailed tooltips. However, the organization of the filters is unclear due to the absence of proper grouping.

4.1.3. Evolution of Design

Initially, the approach for analysing common factors was to explore as many fields as possible, given the nature of the selected datasets. Since most of the data was categorical, bar charts were chosen to display the most frequent values. However, during the exploration, it became apparent that this layout was too space-consuming and repetitive. As a result, the bar charts were replaced with a card row, though this limited the ability to perform drilldown analysis.





4.2. Topic 2: Spatio-temporal Analysis: Traffic crash density and Police response times

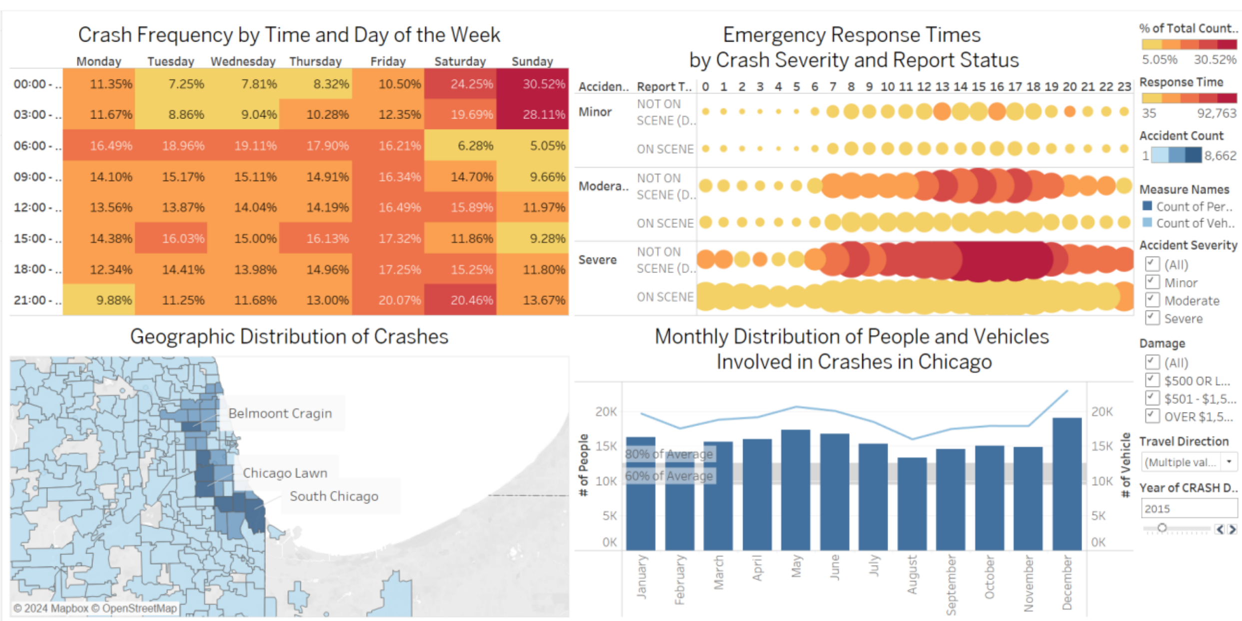


Figure 6. Dashboard 2

4.2.1. Chart Analysis

Heatmap - Crash Frequency by Time and Day of the Week

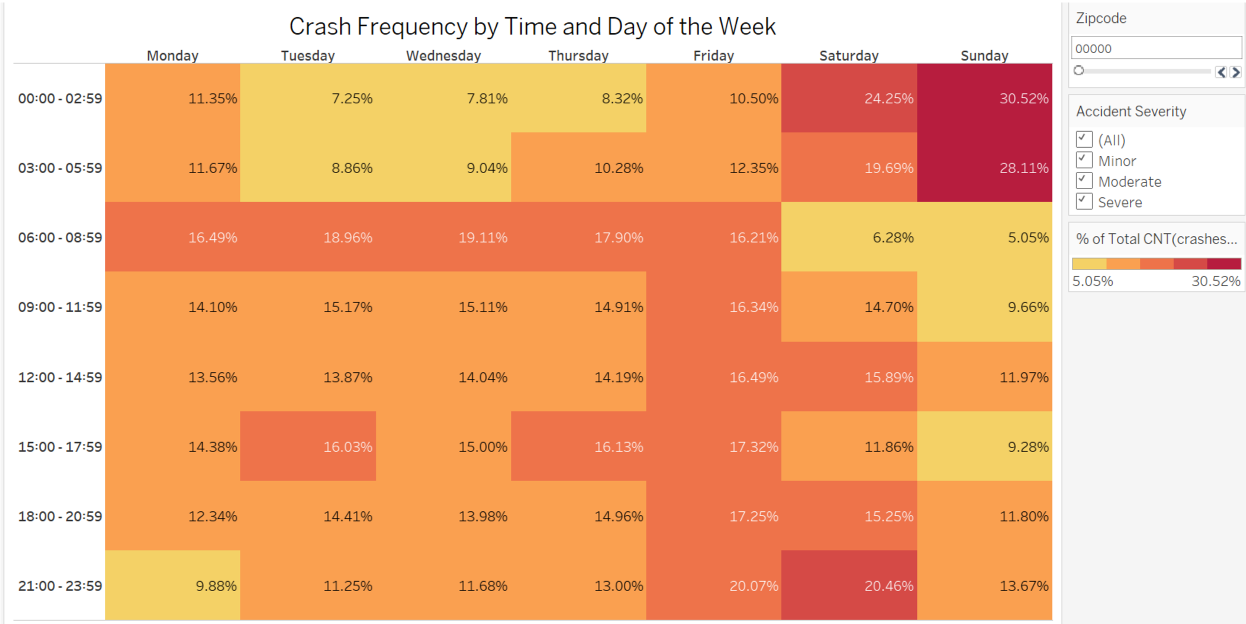


Figure 7 visualises the frequency of traffic crashes by the time of day and day of the week

Attributes	Data type	Mark	Channel	Encoding
Time of Day	Categorical	Rectangles	Positions along the y-axis	Encoded using positions along the y-axis to show different time intervals of the day
Day of the Week	Categorical	Rectangles	Positions along the x-axis	Encoded using positions along the x-axis to show different days of the week
Percentage of Accidents	Quantitative	Colour and text labels	Colour intensity	Encoded using a gradient of color intensity from yellow (low frequency) to red (high frequency) and text labels to show the exact percentages

Table 10. Data encoding Chart 5

Bubble Chart - Emergency Response Times by Crash Severity and Report Status

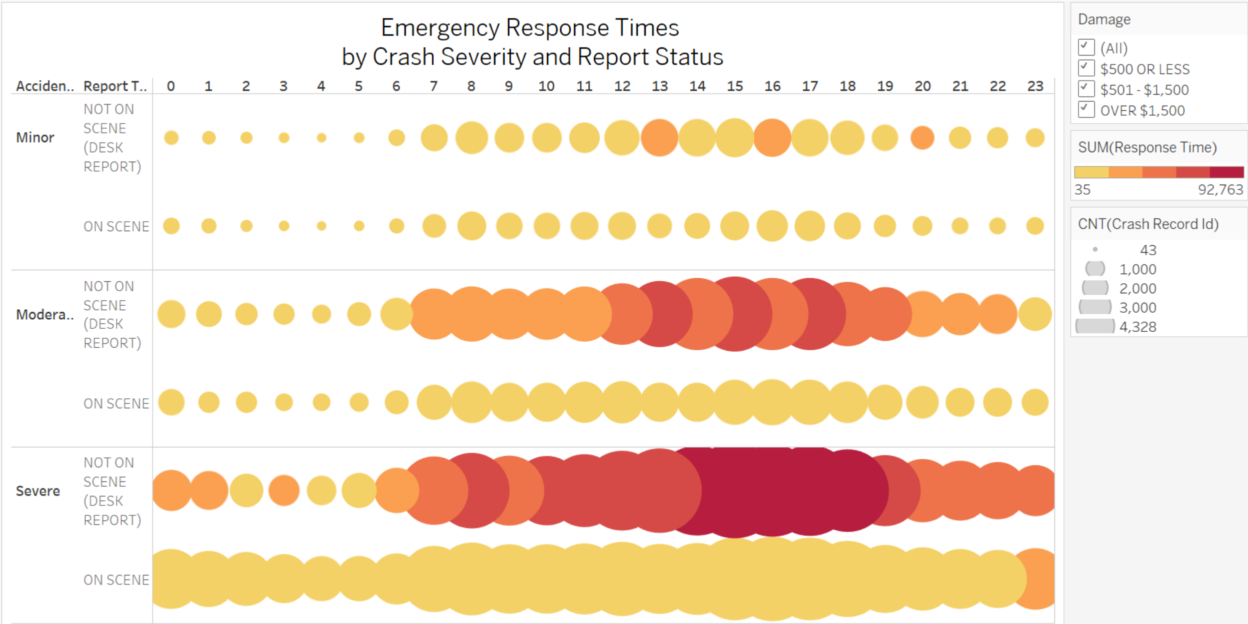


Figure 8 compares the response times based on crash severity and report type

Attributes	Data type	Mark	Channel	Encoding
Crash Severity	Categorical	Circles	Positions along the y-axis	Encoded using positions along the y-axis to show different crash severity levels
Report Type	Categorical	Circles	Positions along the y-axis (secondary level)	Encoded using positions along the y-axis to differentiate report type
Crash Hour	Categorical	Circles	Positions along the x-axis	Encoded using positions along the x-axis to represent each hour of the day
Number of Crashes	Quantitative	Circle size	Size of circles	Encoded using the size of circles, where larger circles represent a higher count of crashes
Response Time	Quantitative	Colour intensity	Color of circles	Encoded using colour intensity, where darker red represents longer response times and lighter yellow represents shorter response times

Table 11. Data encoding Chart 6

Map - Spatial Distribution of Crashes in Illinois

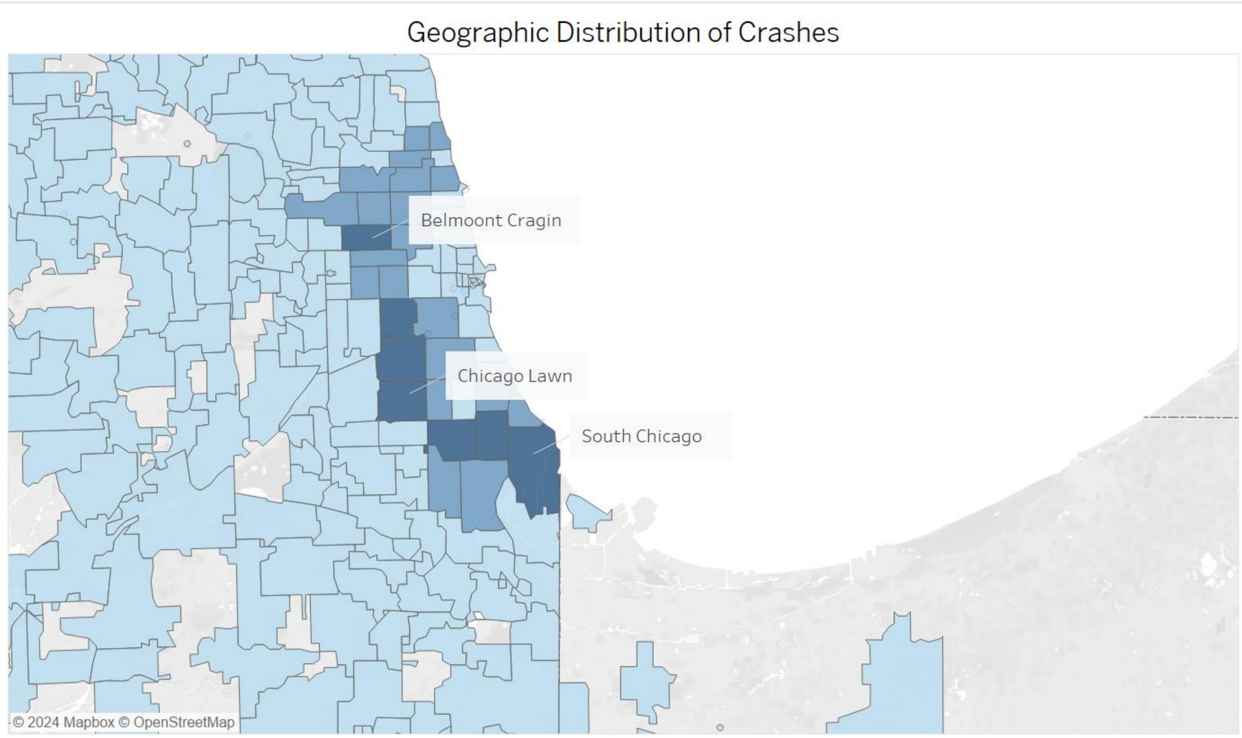


Figure 9 illustrates the spatial distribution of crashes across Chicago's neighbourhoods highlighted for higher crash densities

Attributes	Data type	Mark	Channel	Encoding
Geographic Location	Categorical	Polygons	Positions along latitude and longitude	Encoded using latitude and longitude to position geographic regions within Illinois
Accident Count	Quantitative	Colour intensity and text	Colour of polygons	Encoded using colour intensity, where darker blues represent higher accident counts. Details: Encoded using tooltips to show the exact accident count when hovering over each region.
Zipcode	Categorical	Polygons	Region boundaries	Encoded using distinct geographic regions defined by zipcode boundaries

Table 12. Data encoding Chart 7

Mixed Chart - Monthly Distribution of People and Vehicles Involved in Crashes in Chicago

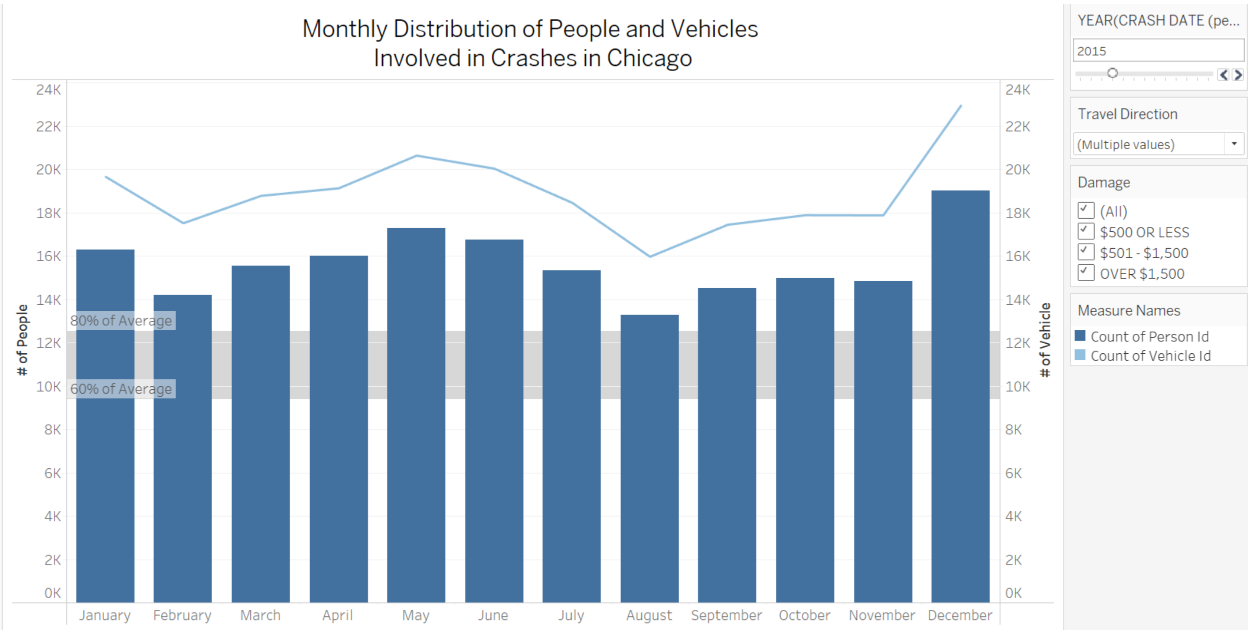


Figure 10 provides an overview of the temporal crash patterns in Chicago, showing the number of people and vehicles involved in crashes each month

Attributes	Data type	Mark	Channel	Encoding
Month	Categorical	Bars	Positions along the x-axis	Encoded using positions along the x-axis to represent each month
Number of People	Quantitative	Bars length	Height of bars	Encoded using the height of bars, where taller bars represent a higher number of people involved in crashes
Number of Vehicles	Quantitative	Line	Position of line	Encoded using the line position, where higher points represent more vehicles involved in crashes

Table 13. Data encoding Chart 8

Chart Analysis

Visualisation	Chart Type	Insight
Figure 7	Heatmap was selected to represent two categorical variables (time and day of the week) to identify patterns and hotspots at specific times and days	Weekends and late-night hours (especially between 00:00 and 05:59) have the highest crash frequency
Figure 8	Bubble chart was selected to display the correlation between two quantitative variables: response time and accident count	Longer response times for more serious accidents, particularly when the crash is on-scene. Minor crashes generally have quicker response times, and the number of crashes increases significantly for severe injuries

Figure 9	The map was used to show geographical distribution and visualise the relationship between location (categorical) and crash frequency (quantitative)	Crashes are highly concentrated in Belmont Cragin, Chicago Lawn, and South Chicago
Figure 10	Combined bar and line chart was selected to visualise two quantitative variables (number of people and vehicles involved in crashes) over time (monthly distribution)	The number of people involved in crashes remains relatively stable throughout the year, but there is a notable spike in vehicle involvement in December, suggesting seasonal factors like weather or increased holiday travel as potential contributors

Table 14. Dashboard 2 Analysis

4.2.2. Visualisation Design

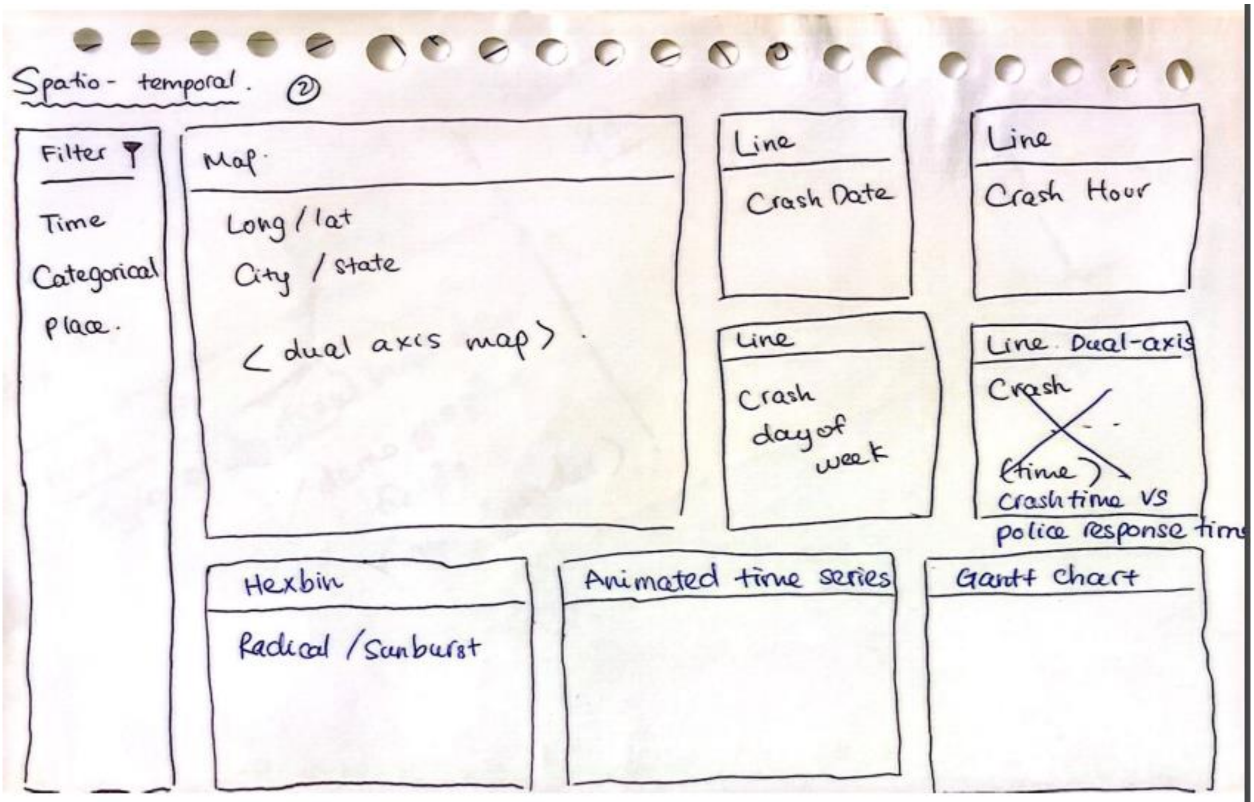
	Principle	Critique
1	Proximity	Grouping of related visualisations: The dashboard is split into two sections, with the upper exploring time analysis, and the lower exploring location analysis. This layout is different from the first dashboard, which may cause viewers to be confused.
2	Similarity	Consistent use of colours helps distinguish two types of analysis (warm colours for time analysis, and cold colours for location analysis). However, there is inconsistency between the geographic map and the bar chart where the colour scales differ significantly.
3	Enclosure	Unclear enclosure, missing bounding boxes around groups of related visualizations
4	Continuity	The top two charts are well-grouped together, but the bottom two are more isolated, disrupting the natural flow for viewers to follow the data narrative
5	Figure-Ground	The dashboard's background is light and does not distract from the figures
6	Closure	Overwhelming with too much detailed data in the heatmap. Should replace with on-hover details and categories grouping
7	Symmetry and Order	The dashboard displays balance in the layout, but the map is slightly out of balance due to its smaller size

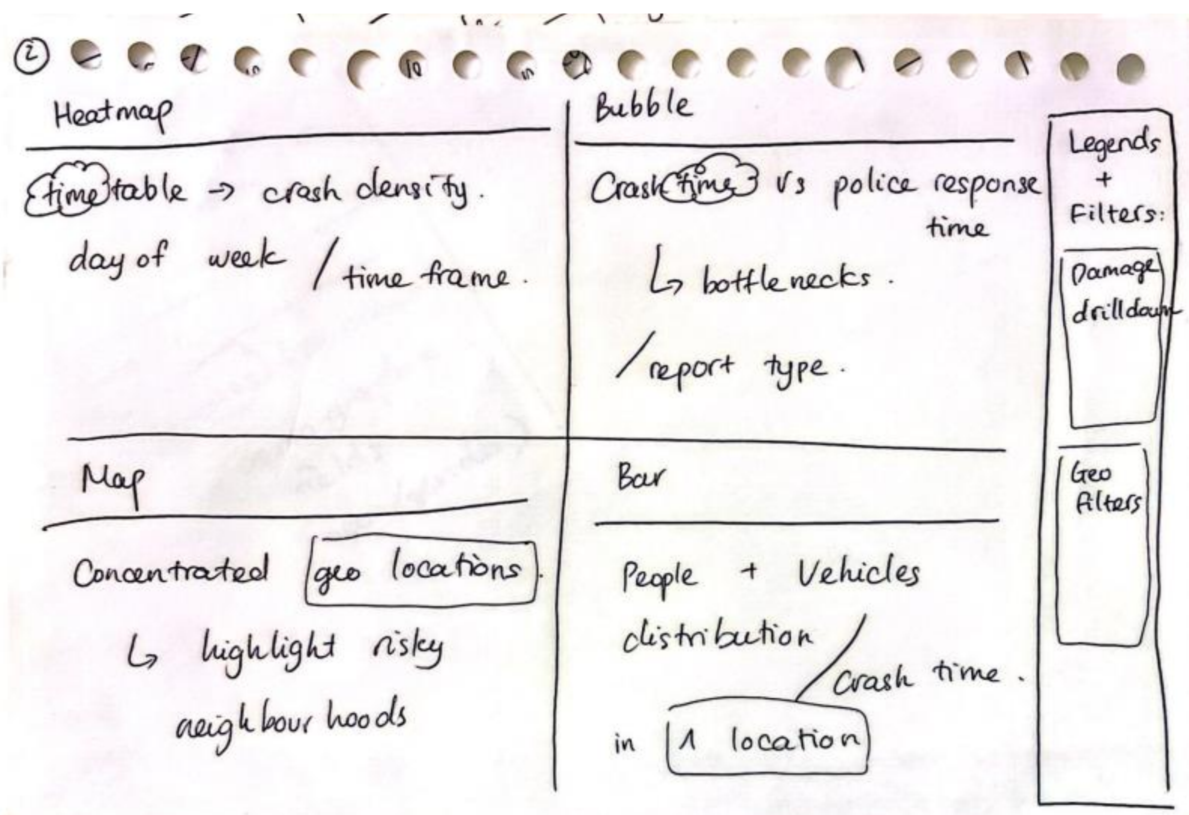
Table 15. Design principles Evaluation of Dashboard 2

The dashboard incorporates interactive features such as filters for accident severity and damage (vehicles damage filters), zip code and travel directions (geographical filters), and year (chronological filter), as this grouping is an improvement from the first dashboard. Additionally, users can hover over elements to view detailed tooltips.

4.2.3. Evolution of Design

Similar to the initial design of Dashboard 1, Dashboard 2 was originally designed to display all time series data using line charts, as they effectively show category distributions over time. However, this approach proved unnecessary since traffic crash data requires hourly updates. Consequently, the time analysis was replaced with a timetable-style breakdown, highlighting accident density by the hour.


Spatio-temporal. ②
Filter
Time
Categorical
place.
Map.
Long / lat
City / state
< dual axis map >.
Line
Crash Date
Line
Crash Hour
Line
Crash day of week
Line. Dual-axis
Crash (time)
Crash time VS police response time
Hexbin
Radical / Sunburst
Animated time series
Gantt chart


②
Heatmap
(time) table → crash density.
day of week / time frame.
Bubble
Crash (time) Vs police response time
↳ bottlenecks.
/ report type.
Legends + Filters:
Damage drilldown
Geo filters
Map
Concentrated geo locations.
↳ highlight risky neighbourhoods
Bar
People + Vehicles distribution / crash time.
in 1 location

4.3. Topic 3: Crash Impact Analysis: Vehicle damage and injury outcomes

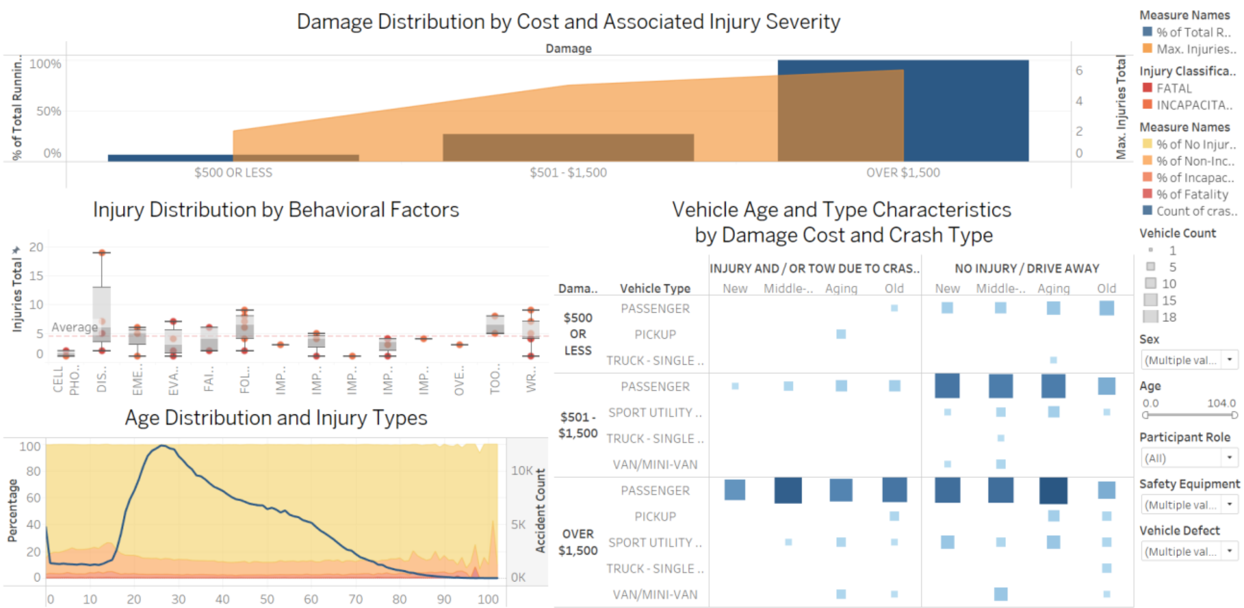


Figure 11. Dashboard 3

4.3.1. Chart Analysis

Pareto Chart - Damage Distribution by Cost and Injury Severity

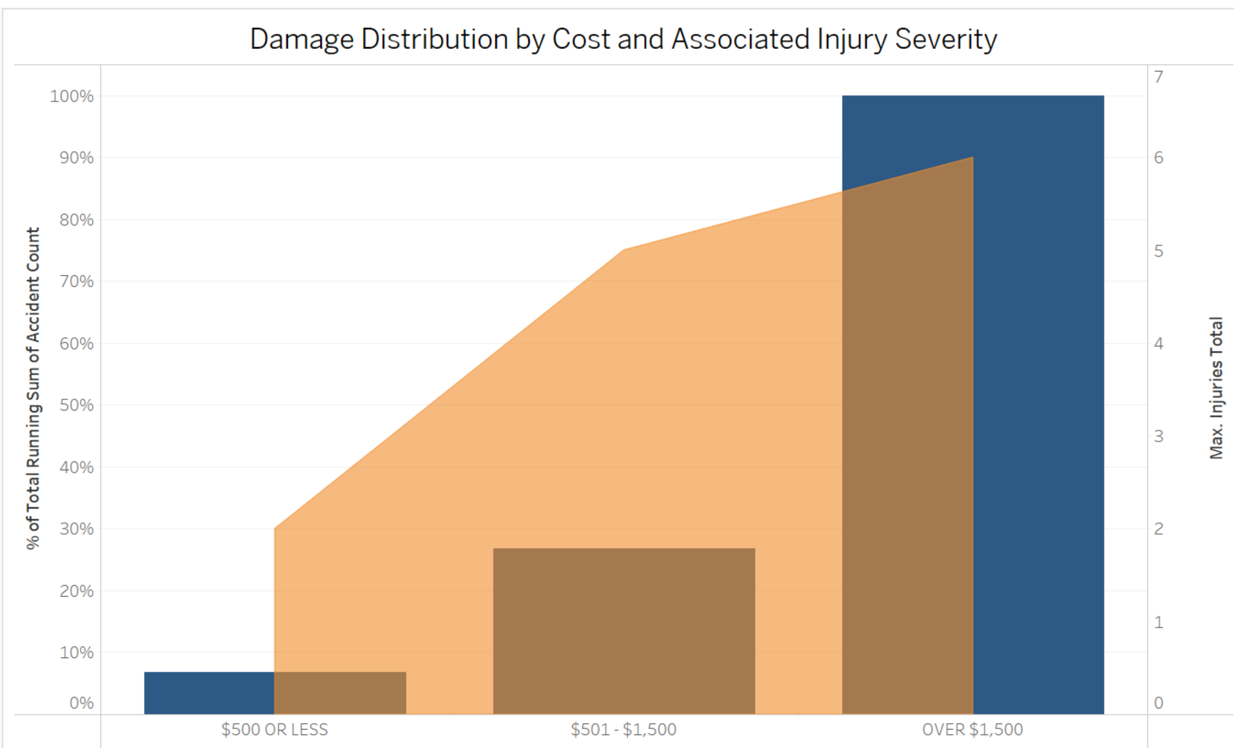


Figure 12 illustrates the distribution of vehicle damage costs and the corresponding injury severity

Attributes	Data type	Mark	Channel	Encoding
Damage	Categorical	Bars, Area	Positions along the x-axis	Encoded using positions along the x-axis to represent damage categories
Percentage of Accident Count	Quantitative	Bars	Height of bars. Colour	Encoded using the height of bars (blue), where taller bars represent a higher percentage of accidents
Max Injuries Total	Quantitative	Area	Height of area. Colour	Encoded using the height of the area (orange), where higher areas represent more injuries

Table 16. Data encoding Chart 9

Box Plot - Injury Distribution by Behavioural Factors

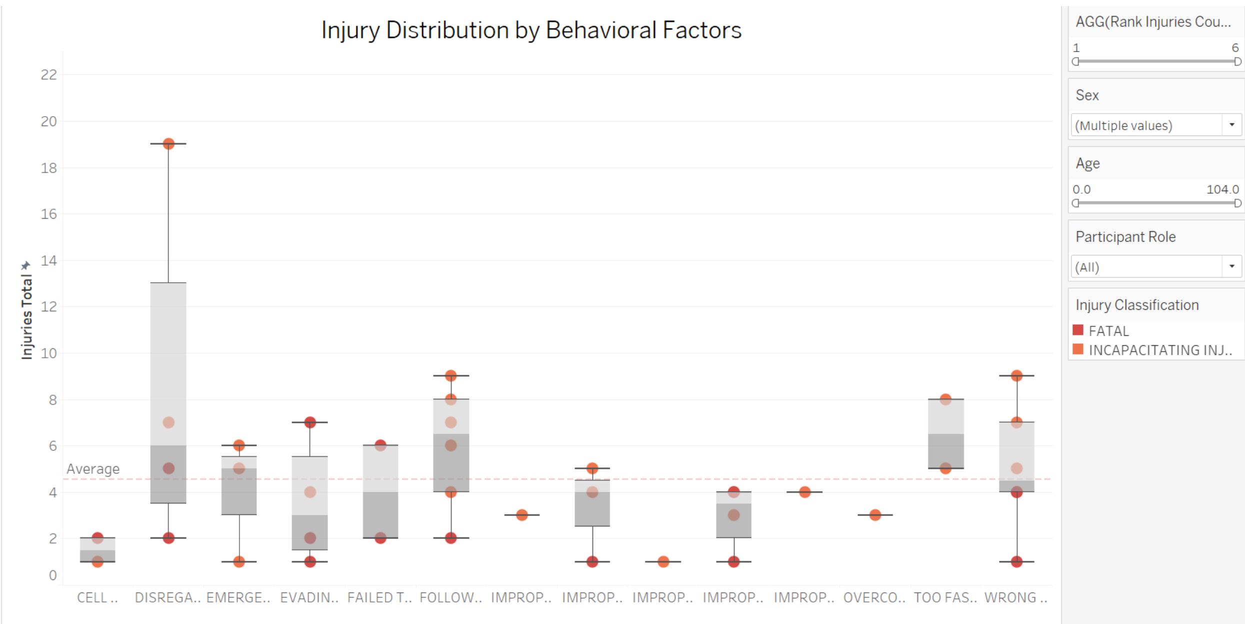


Figure 13 visualises the distribution of injuries caused by different driver behaviours and highlights the high variability in injuries related to specific actions

Attributes	Data type	Mark	Channel	Encoding
Driver Action	Categorical	Box Plot	Positions along the x-axis	Encoded using positions along the x-axis to represent different driving behaviours
Injuries Total	Quantitative	Box Plot	Height of boxes, length of whiskers	Encoded using box plots to show the distribution of total injuries, with the whiskers capturing the spread
Injury Classification	Categorical	Circles	Colour	Encoded using colour to differentiate between injury types (red for fatal, orange for incapacitating)

Table 17. Data encoding Chart 10

Area Chart - Age Distribution and Injury Types

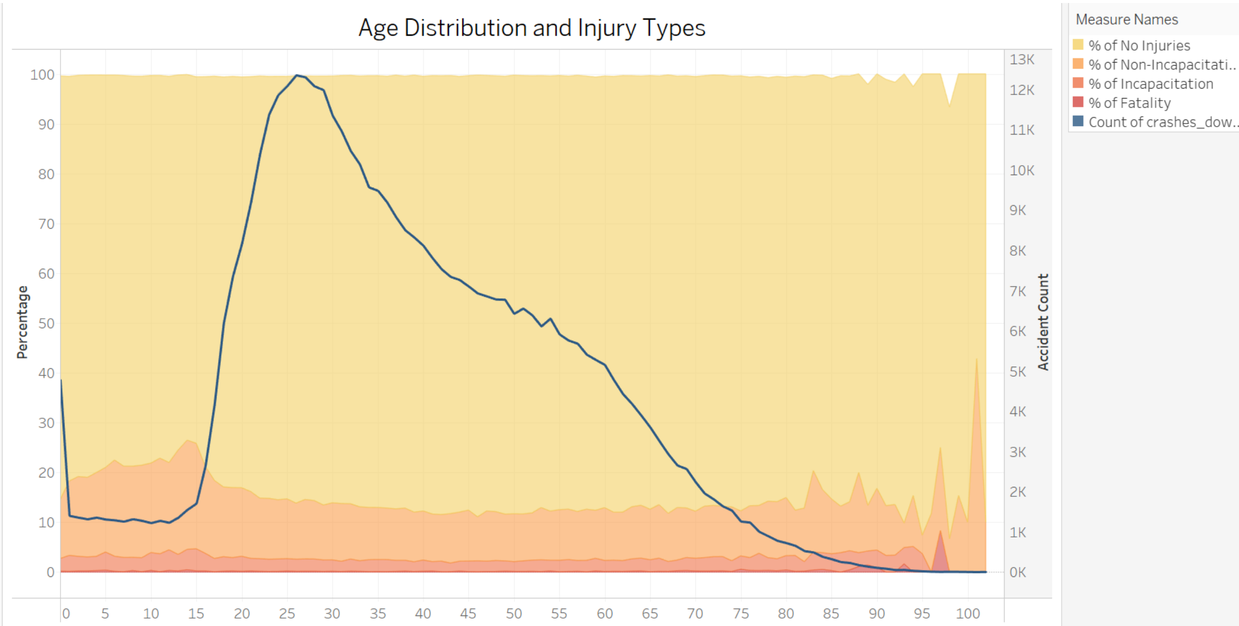


Figure 14. shows the distribution of different injury types across various age groups

Attributes	Data type	Mark	Channel	Encoding
Age	Quantitative	Area, Line	Positions along the x-axis	Encoded using positions along the x-axis to represent different age groups
Injury Types	Categorical	Area and colour	Height of stacked areas and colour intensity	Encoded using the height of the stacked areas to represent the percentage of each injury type per age group Encoded using different colours for each injury type: yellow for no injuries, orange for non-incapacitating, dark orange for incapacitating, and red for fatality
Accident Count	Quantitative	Line	Height of line	Encoded using the height of the line chart to show the total number of crashes for each age group

Table 18. Data encoding Chart 11

Heatmap - Vehicle Age and Type Characteristics by Damage Cost and Crash Type

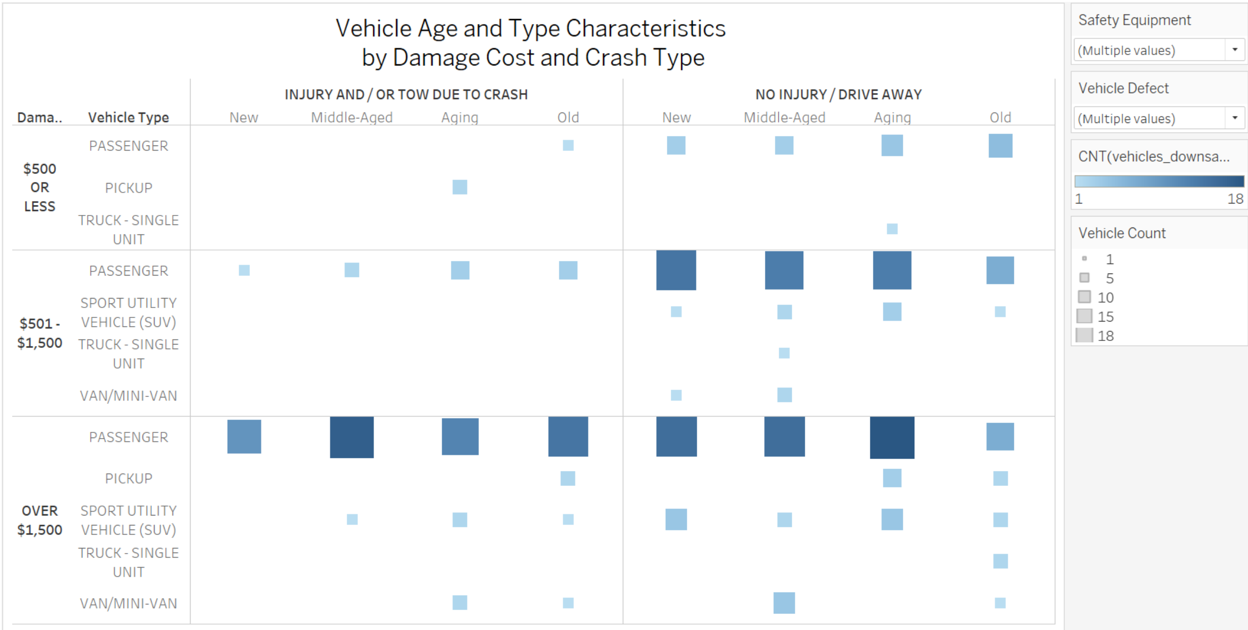


Figure 15 compares different vehicle types and ages against damage costs and crash outcomes

Attributes	Data type	Mark	Channel	Encoding
Damage	Categorical	Squares	Positions along the y-axis	Encoded using positions along the y-axis to represent different damage cost categories
Vehicle Type	Categorical		Positions along the y-axis	Encoded within each damage category to represent different vehicle types
Vehicle Age	Categorical		Positions along the x-axis	Encoded using positions along the x-axis to represent different vehicle age groups
Crash Type	Categorical		Secondary categorization on the x-axis	Encoded as a secondary variable on the x-axis to represent crash type (injury/no injury)
Vehicle Count	Quantitative	Size and colour intensity	Size and colour of squares	Encoded using size and colour intensity of the squares, where larger and darker squares represent a higher count of vehicles

Table 19. Data encoding Chart 12

Chart Analysis

Visualisation	Chart Type	Insight
Figure 12	Stacked bar and area chart was used to compare two quantitative variables (injury severity and total running sum of accidents) across a	Damage cost increases as the proportion of severe injuries increases, with the highest costs (over $1,500) associated with the most significant

	categorical variable (damage cost). The area chart displays the range of injuries, while the bar chart visualises the total accident counts	injuries and the highest running sum of accident counts
Figure 13	Box plot was used to display the distribution of injuries for each driver behaviour (categorical data), with the addition of circles for specific injury classifications (fatal and incapacitating)	Improper turns and following too closely show a wide range of injury totals, with several fatal or incapacitating injuries represented by outliers. Cell phone usage and disregarding traffic signals have fewer injury counts overall
Figure 14	Stacked area chart was used to show age distribution and injury types (both quantitative) over a continuous variable (age), while the line chart represents the total accident count per age group	Younger drivers (ages 16-30) are associated with the highest accident counts, but they tend to have fewer severe injuries. Older drivers (ages 60+) show fewer accidents, but they have a higher proportion of incapacitating injuries
Figure 15	Heatmap was used to compares two categorical variables (vehicle age and vehicle type) with damage cost and crash type (injury or no injury)	Older vehicles and SUVs involved in crashes with higher damage costs (over $1,500) tend to be associated with injury-related crashes. In contrast, newer vehicles and passenger cars are more commonly involved in non-injury crashes

Table 20. Dashboard 3 Analysis

4.3.2. Visualisation Design

	Principle	Critique
1	Proximity	Grouping of related visualisations: The top bar chart provides an overall understanding of how vehicles damage and injuries severity are associated. Then, the dashboard is split into two sections, with the left exploring injury distribution, and the right exploring damage analysis. This layout is similar from the first dashboard
2	Similarity	Consistent use of colours helps distinguish two types of analysis (warm colours for injury analysis, and cold colours for damage analysis). However, this colour scheme is similar to dashboard 2, which may be confusing for viewers as the two dashboards are of different topics
3	Enclosure	Unclear enclosure, missing bounding boxes around groups of related visualizations
4	Continuity	The display follows a logical left-to-right order, with the bottom left two charts are well-grouped together, but the bottom right chart is more isolated, showing an imbalance in information
5	Figure-Ground	The dashboard's background is light and does not distract from the figures
6	Closure	Overwhelming with too much detailed data in the heatmap. Should replace with on-hover details and categories grouping

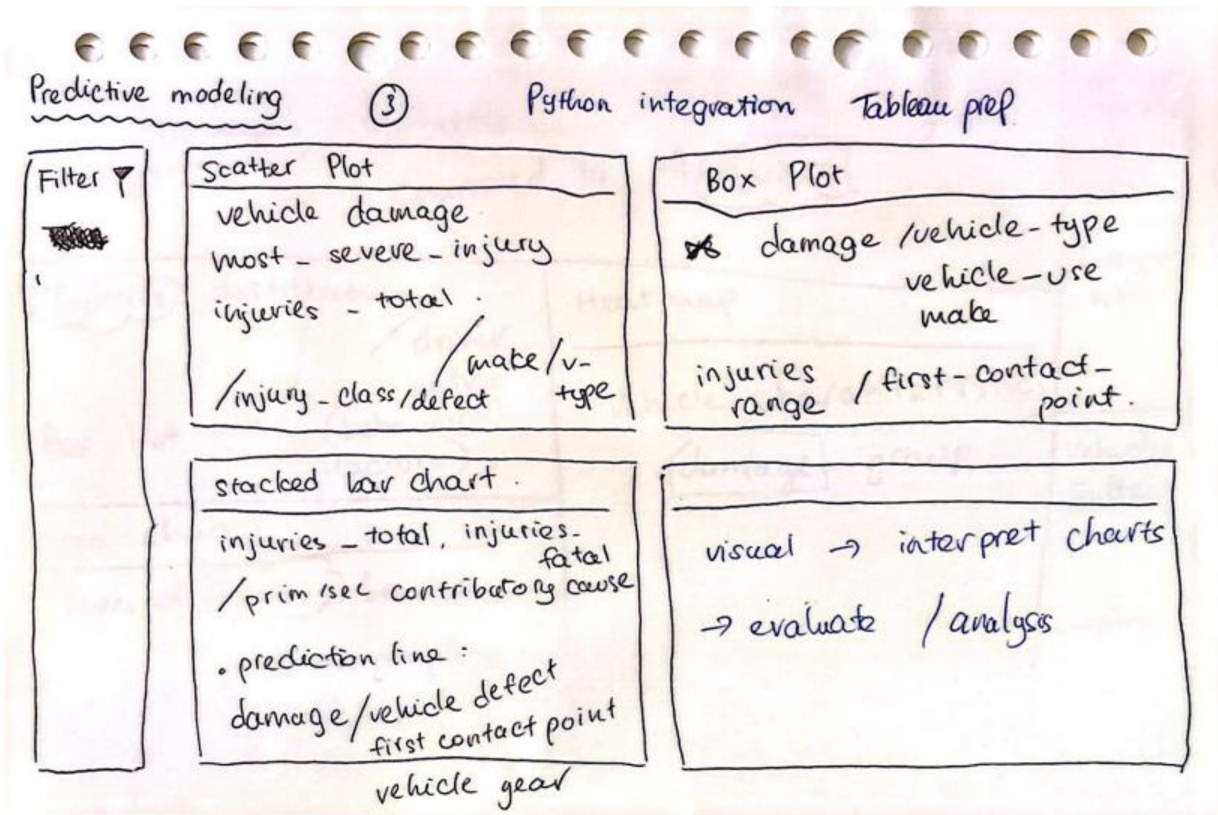
7	Symmetry and Order	The dashboard displays balance in the layout, but the area chart is slightly out of balance due to its smaller size

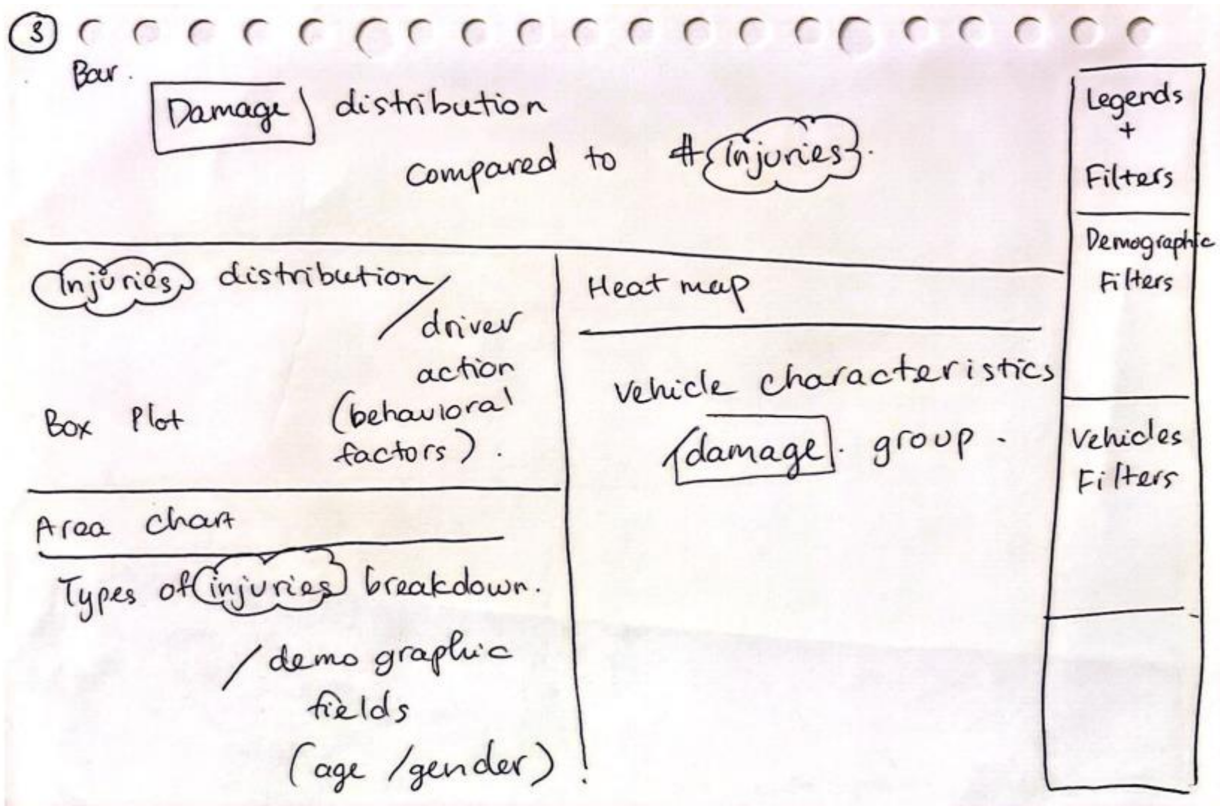
Table 21. Design principles Evaluation of Dashboard 3

The dashboard incorporates interactive features such as filters for gender, age, and participant role (driver characteristics filters), safety equipment and vehicle defect (vehicle characteristics filters), as this grouping is an improvement from the first dashboard. Additionally, users can hover over elements to view detailed tooltips.

4.3.3. Evolution of Design

Topic 3 was the only topic that could not be addressed, as the original plan involved building predictive models, regression for the numerical field "injuries total" and classification for categorical fields like "damage groups." However, this modelling required Python integration, which was beyond the scope of the project. Instead, the focus shifted to exploring two remaining aspects of the datasets: damage analysis and injury analysis. The design was adjusted to first display the relationship between these two aspects, followed by a detailed breakdown of each.





5.0 Conclusion

This data visualization project focused on analysing car crash data to uncover patterns and insights into driver behaviours, infrastructure, and environmental factors that contribute to traffic incidents. As a Data Science student, I learned valuable skills in data preparation, from handling missing values to feature engineering and creating calculated fields. I also gained a deeper understanding of the iterative design process, using interactive features and design principles to create intuitive dashboards that communicate insights clearly to audiences.

6.0 Reference List

Click or tap here to enter text.

Cantillo, Víctor, Luis Márquez, and Carmelo J. Díaz. "An exploratory analysis of factors associated with traffic crashes severity in Cartagena, Colombia." Accident Analysis & Prevention 146 (2020): 105749.

City of Chicago. (2024). Traffic crashes - Crashes [Data set]. Chicago Data Portal.

https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if/about_data

City of Chicago. (2024). Traffic crashes - People [Data set]. Chicago Data Portal.

https://data.cityofchicago.org/Transportation/Traffic-Crashes-People/u6pd-qa9d/about_data

City of Chicago. (2024). Traffic crashes - Vehicles [Data set]. Chicago Data Portal.

https://data.cityofchicago.org/Transportation/Traffic-Crashes-Vehicles/68nd-jvt3/about_data

Stewart, T., 2023. Overview of motor vehicle traffic crashes in 2021 (No. DOT HS 813 435).

Ye, Z, Xue, C & Lin, Y 2021, "Visual perception based on Gestalt theory," in Advances in intelligent systems and computing, pp. 792–797, https://doi.org/10.1007/978-3-030-68017-6_118.