SWINBURNE UNIVERSITY OF TECHNOLOGY

MASTER OF DATA SCIENCE

COS80025 – Data Visualisation

Semester 2, 2024

Deliverable 1: Project Proposal

Due date: Sunday 24th August 2024

Project facilitator: Mr Mohammad Abuhassan

Submitted by: Thi Ngan Ha Do

Student ID: 103128918

1.0 Introduction

1.1 Background and Motivation

For decades, road accidents have been one of the main causes of unnatural death, posing a substantial global safety concern due to the associated injuries and fatalities (World Health, 2023). According to the (Berg et al., 2023), from 1913 to 2022, the number of motor-vehicle fatalities in the United States increased by 996%, including passenger cars, trucks, buses, and motorcycles.

Despite the alarming increase in traffic crashes, we believe this issue can be addressed by raising road user awareness and implementing a more effective traffic control system. Therefore, it is imperative to examine past data to identify contributing reasons and assess the efficiency of current safety measures. Understanding the underlying causes of road accidents is critical to designing effective safety measures.

There are several compelling reasons why traffic crashes have been selected as the focus of this project. From a social perspective, traffic collisions are a longstanding issue within society. In 2023, the economic impact of car crashes accounted for over 1.4% of the U.S. GDP (Tandrayen-Ragoobur, 2024), highlighting the substantial financial burden these incidents impose on the economy, including significant productivity losses.

On a personal level, traffic is an unavoidable part of daily life, yet it poses serious risks, with participants facing potentially life-threatening dangers. The repercussions of traffic crashes extend beyond property damage and physical injury; they also include the profound losses experienced by the families of victims. Around 75% of low-income households who lost a member in a traffic crash reported a decline in their standard of living (Abdulhafedh, 2017). Given these considerations, it is crucial—both personally and collectively—for individuals to understand the factors contributing to accidents and to explore ways to reduce or eliminate the likelihood of such events.

1.2 Project Objectives

The purpose of this project is to examine the impact of many critical elements to uncover common trends in traffic crash severity and evaluate the effectiveness of current safety measures.

	Research topic	**Objective**	**Reason**
1	Identifying common contributory factors in traffic accidents	Understanding the impact of driver behaviour, infrastructure and environmental conditions on the frequency and severity of crashes.	• Individuals: Avoid risk factors for safer driving practices • For policy makers: Guide the development of targeted interventions and safety regulations

2	Spatio-temporal analysis of traffic crash density and police response times	Map out the distribution of traffic crashes across various locations and time periods, and analyse the response times of police to these incidents	• Individuals: take precautionary measures (avoid high-risk areas and time periods) • For policy makers: resource allocation, traffic management, and law enforcement deployment
3	Predictive modelling of vehicle damage and influential factors	Identify and analyse the attributes that most significantly influence the extent of vehicle damage	• Individuals: practice safe driving and enhance awareness of high-risk situations. • For policy makers: prioritize interventions that reduce the severity of vehicle damage

1.3 Project Schedule

In accordance with the syllabus and assignment timeline, a project schedule has been established to ensure consistent weekly progress and the completion of the project within the designated timeframe.

Week	Task	Description	Start Date	End Date
Week 1-4	Data collection and project proposal	(1) Data Collection: • Identify and gather relevant datasets to the project's objectives • Document the data sources and attributes for future reference (2) Project Proposal: • Define the research questions • Develop a project plan (objectives and scope) • Obtain feedback and approval on the project proposal	29/07/2024	25/08/2024
Week 5	Data processing	(1) Data Cleaning: • Check for errors and correct any inconsistencies	26/08/2024	01/09/2024

		• Standardize date-time formats (2) Feature Engineering: • Standardise features for modelling • Create new columns • Extract temporal features		
Week 6-9	Data exploration and building dashboards	(1) Data Exploration: • Perform exploratory data analysis and identify potential patterns • Develop predictive models (2) Building Dashboards: • Design dashboards • Integrate predictive models into the dashboards	02/09/2024	06/10/2024
Week 10	Finalize dashboards	(1) Dashboard Refinement: • Review the dashboard • Collect feedback (2) Documentation: • Document the methodology, data sources, and key insights	07/10/2024	13/10/2024
Week 11	Final Report	• Summarize the entire project • Discuss the implications of the findings and provide actionable recommendations	14/10/2024	20/10/2024

The most time-consuming task involves data exploration, which will be conducted by building dashboards aligned with the proposed research questions. Additionally, regarding vehicle damage prediction (research topic 3), data modelling will initially be performed using a Python notebook, with the analytical results subsequently visualised on a Tableau dashboard. This will be complemented by a thorough descriptive and predictive analysis of the key factors significantly influencing property damage.

2.0 Data

2.1 Data Source

2.1.1. Data Acquisition

The “Traffic Crashes - Crashes” dataset was sourced from the Chicago Data Portal and was last updated on August 23, 2024, as per the last access date. This dataset, provided by the City of Chicago (2024), is part of a broader Traffic Crashes database schema, which also includes datasets on people and vehicles. However, the analysis and visualization for this

project focused solely on the crashes dataset, excluding the other two. The data covers a collection period from 2015 to the present.

2.1.2. Metadata

The dataset comprises 48 attributes and contains over 800,000 records, with each row corresponding to a distinct traffic accident incident. The metadata is comprehensively detailed on the official city government website:

	Column	Description	Data Type	Category
1	crash_record_id	Unique identifier	Text	Categorical
2	crash_date_est_i	Crash date estimated	Text	Categorical
3	crash_date	Date and time of crash	Datetime	Temporal
4	posted_speed_limit	Posted speed limit	Integer	Numerical
5	traffic_control_device	Traffic control device present	Text	Categorical
6	device_condition	Control device condition	Text	Categorical
7	weather_condition	Weather condition	Text	Categorical
8	lighting_condition	Light condition	Text	Categorical
9	first_crash_type	Type of first collision in crash	Text	Categorical
10	trafficway_type	Trafficway type	Text	Categorical
11	lane_cnt	Total number of through lanes in either direction, excluding turn lanes (0 = intersection)	Integer	Numerical
12	alignment	Street alignment	Text	Categorical
13	roadway_surface_cond	Road surface condition	Text	Categorical
14	road_defect	Road defects	Text	Categorical
15	report_type	Administrative report type (at scene, at desk, amended)	Text	Categorical
16	crash_type	Severity classification: Injury and/or Tow Due to Crash or No Injury / Drive Away	Text	Categorical
17	intersection_related_i	A field observation whether an intersection played a role	Text	Categorical
18	not_right_of_way_i	Whether the crash began, or first contact was made outside of the public right-of-way	Text	Categorical
19	hit_and_run_i	Crash did/did not involve a driver who caused the crash and fled the scene	Text	Categorical
20	damage	A field observation of estimated damage	Text	Categorical
21	date_police_notified	Calendar date on which police were notified of the crash	Datetime	Temporal

22	prim_contributory_cause	The most significant causing factor	Text	Categorical
23	sec_contributory_cause	The second most significant causing factor	Text	Categorical
24	street_no	Street address number	Text	Categorical
25	street_direction	Street address direction	Text	Categorical
26	street_name	Street address name	Text	Categorical
27	beat_of_occurrence	Chicago Police Department Beat ID	Integer	Numerical
28	photos_taken_i	Whether the Chicago Police Department took photos at the location of the crash	Text	Categorical
29	statements_taken_i	Whether statements were taken from unit(s) involved	Text	Categorical
30	dooring_i	Whether crash involved a vehicle occupant opening a door into the travel path of a bicyclist	Text	Categorical
31	work_zone_i	Whether the crash occurred in an active work zone	Text	Categorical
32	work_zone_type	The type of work zone	Text	Categorical
33	workers_present_i	Whether construction workers were present in an active work zone	Text	Categorical
34	num_units	Number of units involved. Each unit represents a mode of traffic	Integer	Numerical
35	most_severe_injury	Most severe injury sustained by any person involved	Text	Categorical
36	injuries_total	Total persons sustaining fatal, incapacitating, non-incapacitating, and possible injuries	Integer	Numerical
37	injuries_fatal	Total persons sustaining fatal injuries	Integer	Numerical
38	injuries_incapacitating	Total persons sustaining incapacitating injuries	Integer	Numerical
39	injuries_non_incapacitating	Total persons sustaining non-incapacitating injuries	Integer	Numerical
40	injuries_reported_not_evident	Total persons sustaining possible injuries	Integer	Numerical
41	injuries_no_indication	Total persons sustaining no injuries in the crash	Integer	Numerical

42	injuries_unknown	Total persons for whom injuries sustained are unknown	Integer	Numerical
43	crash_hour	The hour of the day component of CRASH_DATE	Integer	Numerical
44	crash_day_of_week	The day of the week component of CRASH_DATE. Sunday=1	Integer	Numerical
45	crash_month	The month component of CRASH_DATE	Integer	Numerical
46	latitude	The latitude of the crash location	Float	Numerical
47	longitude	The longitude of the crash location	Float	Numerical
48	location	The crash location as derived from the reported address		Point

2.2 Data Processing

This dataset, collected and pre-processed by an official provider, is considered a reliable source with minimal errors. However, in preparation for the data exploration phase, it is crucial to ensure data consistency and to derive additional insights by creating new columns from the existing data to address the proposed research questions.

Issues	Relevant Columns	Solution
Missing values	Entries missing over 30% values	Drop rows
	Nominal columns	Fill in missing values with False/No
Data types	Date-time columns	Convert mm/dd/yyyy to dd/mm/yyyy format
New columns creation	Crash_date and Date_police_notified	Calculate time differences in crash date and police notified date
	Crash_date	Extract the hour, day of the week, month, and year components to establish a date hierarchy

Several rules have been established to handle missing values and duplicates. Given the large volume of entries, it is acceptable to remove rows that do not meet the criteria. Throughout the data exploration process, maintaining data integrity is prioritized to ensure the accessibility and accuracy of the final report.

3.0 Requirements

3.1 Must-Have Features

Column	Research question	Reason
posted_speed_limit, device_cond, weather_cond, lighting_cond, trafficway_type,	Identifying common contributory factors in traffic accidents	Critical environmental, infrastructural, and behavioral factors that

roadway_surface_cond, crash_type, prim_contributory_cause, sec_contributory_cause, dooring_i,		contribute to traffic accidents
crash_date, date_police_notified, street_no, street_name, latitude, longtitude	Spatio-temporal analysis of traffic crash density and police response times	Temporal and spatial data to analyze the distribution of crashes over time and across locations, as well as the efficiency of police response times
num_units, injuries_total, damage, crash_hour, crash_day_of_week, first_crash_type, trafficway_type	Predictive modelling of vehicle damage and influential factors	Key variables (both numerical and categorical) that influence the severity of vehicle damage, enabling the development of accurate predictive models

3.2 Optional Features

Column	Reason
lane_cnt, alignment, report_type, location	Additional information on the roadway conditions, may not be necessary for the scope of this analysis. Regarding location, the data format is incompatible with visualisation tools like Tableau, replace with Longtitude and Latitude columns
intersection_related_i, not_right_of_way_i, beat_of_occurence, photo_taken_i, statements_taken_i, work_zone_i, work_zone_type, workers_present_i	Columns with over 30% missing values are suboptimal for visualization purposes. The nominal data in these columns offers limited utility for data modeling.

4.0 Reference List

Abdulhafedh, A. (2017). Road traffic crash data: an overview on sources, problems, and collection methods. *Journal of transportation technologies*, *7*(2), 206-219.

Berg, C., Evans, S., Luby, E., Medrano, I., Reyes, H., Shepard, E.,…Cody, P. (2023). National Safety Council: Our Driving Concern Program Evaluation.

City of Chicago. (2024). Traffic crashes - Crashes [Data set]. *Chicago Data Portal*. https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if/about_data

Tandrayen-Ragoobur, V. (2024). The economic burden of road traffic accidents and injuries: A small island perspective. *International Journal of Transportation Science and Technology*.

World Health, O. (2023). *Global status report on road safety 2023: summary*. World Health Organization.