Detailed Class Feedback on PE1

We design these programming exercises to put your programming skills into practice, but as with most data exploration and visualisation tasks there are multiple possible answers. Therefore, with this document, we try to provide a full description of what possible responses we were looking for and to help you to better understand what you might have missed or any individual feedback you received. As always, talk to your marker if you need further guidance.

Data Cleaning:
Your data cleaning section should have identified a number of irregularities and explained how you wrangled them. These were not complex, but they are the sort of errors you will commonly encounter in data. You had to use visualisations from Tableau to demonstrate them (and may have even found them that way). This often was no more complex than a bar chart that showed how the related values were odd, compared to the rest of the data.

There were actually more than three types of irregularities in the data. To get full marks, you had to correctly describe and visualise at least three different types/fields, not just three instances of an irregularity. You also had to explain and justify how you handled them. Let's talk about the main irregularities you could have found and visualised:

Numbers for day of the week not consistent: There was an irregularity in the numbers associated with days of the week, e.g., 1 for Sunday, 2 for Monday. The data didn't have this recorded consistently. Some days of the week used the numbers 0-6, starting from Sunday. Others had 1-7. To fix this, you either had to assign a specific number to each day, or you just used the actual date to determine the day of the week and then assign it a number (either 1-7 or 0-6).

An accident in the bay: There was one record that had longitude and latitude values corresponding to a location in the middle of the water of Port Phillip Bay. This was caused by adding a '9' in the coordinates. All you needed to do to fix it was to correct the coordinates or make the latitude or longitude null for this accident. If you found any other location irregularity, you had to argue well why it was an irregularity.

Duplicated accident: There was also a row that seemed to be repeated. This could be discovered in various ways by visualisation. The crucial thing is that it had to show that the data claimed that the same accident occurred at the same location at the same time. The best way to handle this is to delete the duplicates.

The other errors were a little harder to understand.

Column type string converted to integer: Many identified an irregularity in the SPEED_ZONE data because of the values “777”, “888” and “999”, but most students thought these were outliers beyond the speed limits in the State of Victoria. They weren't, because the values in the data are speed zone categories, not speed limit numbers. According to the metadata published by the government for this data, the following categories are valid:

> "040 40 km/hr 050 50 km/hr 060 60 km/hr 075 75 km/hr 080 80 km/hr 090 90 km/hr 100 100 km/hr 110 110 km/hr 777 Other speed limit 888 Camping grounds, off road 999 Not known"
> https://vicroadsopendatastorehouse.vicroads.vic.gov.au/opendata/Road_Safety/RCIS%20Documents/Metadata%20-%20Victoria%20Road%20Crash%20data.pdf

If you wanted to ignore these last three values during your analysis, the best way would have been to either explicitly filter them out in Tableau when visualising this column (which is not a permanent way of cleaning your data), or set them to NULL and then filter that out. It was incorrect to delete the entire record as 1) that is a lot of accident records that would been deleted, and 2) the records contained other data that should be considered for other analysis.

Similar LGA Name: The LGA_NAME field had some values that were strings that started and finished with curved brackets, e.g., "(MOUNT BULLER)". This is not an error. They are areas of Victoria that aren't within a local government area (LGA) due to their special nature, but are managed by the state government. However, there was one record that had the value "(MT BULLER ALPINE RESORT)". This was an error that could have been easily fixed by changing it to "(MOUNT BULLER)" or filtering it out of related analysis.

Some LGAs are regarded as belonging to multiple categories for DEG_URBAN_NAME, e.g., some accidents in the LGA of Ballarat were RURAL_VICTORIA and others were LARGE_PROVINCIAL_CITIES. This is fine, as some LGAs are large and the accident could be anywhere in the area, and the categories are not exclusive. If you claimed this needed to be cleaned, you would have had to justify how it impacted your analysis and have implemented a sensible response to minimise the effect.

Light condition irregularities: Some accidents occurred in the middle of the day or night, but had values describing lighting conditions that were inconceivable, like "Dawn/Dusk" or "Day" when it clearly wasn't that time. It is hard to set threshold times for some of these, but there certainly were some clear errors. The crucial thing for PE1 is that you weren't expected to use the lighting conditions in your analysis, so it didn't need to be fixed as it would not have any impact on your data exploration. However, if you did then the safest cleaning would be to reset any values that are clearly errors to NULL.

There were some records that already had missing data or NULL values. The assignment said to not treat these as errors.

Questions 1 and 2:
These had many possible solutions, depending on how you approached the visualisation and wrangling. The issues often for both were 1) were you showing the right attributes and 2) did you visually allow the user to compare and contrast the values for multiple attributes.

Question 1 asked about when specific accident types occurred, per year, month, day of the week and hour. This was commonly visualised with multiple line graphs. If you left out a time measure or didn't include accident types in your visualisations, you might have lost marks. You were expected to recognise various trends in these visualisations, like the peaks on Fridays, during the morning and afternoon peak hours and the drop in 2020 due to the pandemic lockdowns. You could have also commented on the different behaviours of

different accident types and the fact that 2023 did not have data for the entire year. (Note this is not a irregularity or needing cleaning; it just needs to be understood and taken into account when exploring the data)

Question 2 was harder. It asked you to consider the spatial data related to the accidents and how this related to the road geometry, speed zones and the urban/rural aspect of the location. This required multiple visualisations, at least one of which was expected to be a map. A number of students included three maps, one for each of the non-spatial attributes, but did not provide any visualisations (map or otherwise) that visually showed whether any of the attribute values occurred in the same locations as each other. As such, the visualisations did not completely explore the data. The visualisations also had to be clear in what they showed. If the plots on a map overlapped each other a lot, it was hard to determine what plots you couldn't see. It was also hard to determine on a symbol map which attribute values were more common. That was more suitable for graphs or a density map.

A lot of students had issues with the rural/urban aspect. It had to be clear what data was considered to be about rural accidents and what was urban accidents. Often the DEG_URBAN_NAME values "MELB_URBAN" and "RURAL_VICTORIA" were used in the commentary and/or visualisation without any consideration of the remainder of the data, e.g., accidents in SMALL_CITIES, MELB_CDB, TOWNS. As a result, it was hard to accurately interpret patterns in the data relating to the rural/urban distinction. For instance, you couldn't say which road geometry was more common in one than the other.

Combining Q1 and Q2: Q2 also asked you to consider whether what you learnt from the data and visualisations in Q1 and Q2 supported or challenged each other. This was not well answered. The most obvious response is that together they and other visualisations can be used to explore whether particular accidents in certain locations (e.g., rural vs urban) occurred at certain times. A lot of students did not address this using visualisations and just speculated about possible relationships. This is part of the role of data visualisation. Visual analytics is not just used to help answer specific research questions; it is also to help explore and see how elements of the data inter-relate to each other. You were expected to consider this in PE1 (and hopefully in your DEP).

Report Writing and Clear Visualisations: Finally, you needed to write a clear and well written report. As we are particularly using visualisations you need to make sure your visualisations are understandable. This means that your visualisation is clear, that the font sizes and colours are suitable for the page. You should have colour legends and axes to identify the unit of measurement. Each figure should have had a clear caption and reference. Some have titles too. There should be at least one of these, but a title and repeated title as a caption is not a useful caption. Be careful of your precious page limit and use the space wisely. Furthermore, you should mention each figure or table in your text and clearly identify it, e.g., "as seen in Figure 1 and Table 3". Make things clear for the reader/user.

Try not to use black outlines around your figures or legends. Think about what you want to stand out to the reader and ensure that the visualisation and what is included is readable (and does not have missing parts of the text, or underscores etc). Many provided screenshots of the entire Tableau page rather than the visualisation itself.

Hopefully this assignment has made you more familiar with Tableau as well as introduced you to using visual analytics for data exploration. What is most crucial here is that in order to understand the visualisations for Q1 and Q2, you needed to consider them together, not just in isolation. You will need to consider how you can use multiple displays when designing your visualisation for your DVP. The visual elements really need to be designed so they can work together to give the audience a better understanding of the data as a whole.

Thanks for all your submissions and insights,

Michael, and Sarah, on behalf of the entire FIT5147 teaching team