# Case study evaluation

| Week 1 case study – Adelaide road crash data | Week 2 case study – Modelling viral spread |
|---|---|
| **Introduction** | |
| **Purpose** – To identify areas with high risk of road fatalities in Greater Adelaide in order to provide information relevant to budgeting and public health to stakeholders.<br><br>**Stakeholders** – South Australian government, local government, city planners, emergency services, and road users.<br><br>**Data** – Provided by *Data.SA*. Key variables included: Lambert coordinates (ACCLOC_X and ACCLOC_Y), area speed, total casualties (all quantitative variables).<br><br>**Modelling types used** – k-means clustering. | **Purpose** – (Initially) To classify locations based on the severity of COVID-19 (as indicated by the number of infections and deaths in each location) and compare policy/response effectiveness.<br><br>**Stakeholders** – Organizations such as WHO, various governments/ departments of health, clinicians, genetic researchers and immunologists.<br><br>**Data source** – Provided by the *ECDC*. Key variables included: Country, confirmed cases, and confirmed deaths. Preview available here<br><br>**Modelling types** – k-means clustering, k-nearest neighbour, |
| **Summary of methods** | |
| The position was plotted as a scatter plot using Lambert coordinates, coloured by the number of casualties. This revealed a road map of Greater Adelaide but didn't show any relationships between position and number of crashes. K-means clustering (K of three, no justification provided) was then performed to isolate groups by speed (1401 in 60km/hr or less, 1898 in 70-100km/hr, and 10300 in 100km/hr or more). Little variation was observed between speed and the mean number of casualties. From here, only instances of three casualties or more were considered in order to reduce the amount of data. When this subset of the data was grouped with k-means clustering (k of 3), clear grouping was observable geographically with cluster one in the north, cluster two in the south, and cluster three (5, 8, and 210 instances respectively) in the centre/metropolitan area. K-means clustering was used again to determine if there was any variation between car speed and mean number of casualties for instances with three or more casualties, all groups had between 3-4 casualties. | K-means clustering (k of five, no justification provided) was used to categorise 195 locations into five clusters. The cluster number was then used as the target variable in a k-nearest neighbour algorithm. Using random sampling, 90% of the data was used as training data and 10% was reserved to test the model. It is not clear what value was selected for k. The accuracy of the model was tested 1000 times, of those it appears around 800 showed an accuracy of more than 0.90. This may be misleading as will be discussed in the next section. This process creates a model that can predict which countries would have similar conditions to a country with a given prediction set. |

## Evaluation

**Method suitability –**The k-means clustering model requires that the user sets the number of clusters, k, but is otherwise unsupervised. K-means clustering requires quantitative data, it is unclear if any pre-processing was done to salvage qualitative variables (eg, one-hot encoding) such as those relating to drug or alcohol involvement, weather, or road condition. Additionally, k-means clustering isn't suitable if the cluster in question is non-convex, or if there are outliers present, one of the following may be better suited: agglomerative clustering, mean-shift clustering, DBSCAN, or other density-based clustering techniques (Anderson 2009; Gnjatović et al. 2022; Ket et al. 2020; Kumar & Toshniwal 2016).

**Support for conclusion –** No comment is made on how many repetitions were made to ensure centroids were optimally placed, nor any variance values provided (eg, through an elbow curve) nor a justification for the selection of 3 for k as is common practice (Anderson 2009; Holmgren et al. 2020; Klyavin et al. 2021; Kumar & Toshniwal 2016; Sinclair & Das 2021). While these finding may be accurate, they are limited and poorly validated.

**Depth/quality of conclusion –** Simply measuring the variation of mean number of casualties between the speed groups makes no comment on the frequency or severity of the occurrence which will be of interest to stakeholders.

**Table 1.2:** K-means clustering with three clusters of sizes 39, 57 and 127. Cluster means

| | Total Cars | Area Speed |
|---|---|---|
| 1 | 3.49 | 79.74 |
| 2 | 3.60 | 104.04 |
| 3 | 3.46 | 57.64 |

**Method suitability –** limitations of K-means clustering have already been discussed. It has been used in this setting previously, however the impact of outliers and the multidimensional shape of the data should be investigated (Gohari et al. 2022; González-Collazo et al. 2022). The arbitrary choice of k (as well as data cleaning choices) may have resulted in too few instances within clusters one, four, and five.

The k-nearest neighbour algorithm is a supervised method used to predict a classification. The purpose of the study has shifted from comparing responses of the past to guide future policy, to predicting what country 'Country X' (with a given set of predictors) is most similar to.

**Support for conclusion –** It is unwise to make comparisons with the assumption that all strains of the virus behave the same way (El-Shabasy et al. 2022). Additionally, the accuracy measures are misleading because some clusters are unlikely to appear in the testing data at all with 13 of 195 countries making up three of the clusters, sampling bias associated with the target variable leading to an availability bias (Schwartz et al. 2022). It may appear accurate and useful, but appearances aren't everything.

**Depth/quality of conclusion –** Comparing countries based on the number of deaths and infections without considering population densities is non-sensical. You can't make comparisons when the social contexts are so varied. If the data had accounted for population size (eg, by dividing by population size and area, using %infected and %death instead of counts) then perhaps the clusters would be better represented in both the training and testing sets; leading to a more meaningful accuracy score (Shao & Xiong 2022).

| | |
|---|---|
| **Additional suggested improvements** | |
| - The data is a bit dated.<br>- The grouping should have read “50-70km/hr”, not “50-60km/hr”.<br>- Three or more casualties was an arbitrary choice, reducing by another feature may have been more meaningful (eg, Rear-end crashes only, fatalities, etc.)<br>- To add to the above point, it is unclear if ‘no of casualties’ for each instance refers to the passengers in that car, or in that incident – in which case duplicate casualty values may exist in the data.<br>- Time variables in conjunction with direction of travel and road slope, could be used to engineer a visibility variable which may be informative (González-Collazo et al. 2022).<br>- Unclear how missing values, outliers, and other cleaning was performed (reproducibility issues).<br>- The number of passengers is less meaningful than the condition of the driver as mentioned above. Colouring instances by some of these qualitative variables on another scatter plot would provide stakeholders with essential information (Ket et al. 2020). | - The number of neighbours was not outlined, nor was any weighting method indicated (reproducibility issues).<br>- The 90/10 split may not be the optimal split.<br>- There are many assumptions made in the premise. Not everyone that’s infected will get tested, there has also been cases of false negatives and intentional incorrect reporting (Burki 2022). Additionally, not all strains have the same severity/behaviour. In reality the fight against covid is multifaceted. To compare responses as suggested is senseless, because all infections are being treated as though they are of the same strain, all regions as though their innate immunity is equivalent, and all places as though the weather and social norms are identical (El-Shabasy et al. 2022; Riley 2010) (Atchadé & Sokadjo 2022)..<br>- With the above in mind, we can reduce computational costs by selecting countries with similar population densities and social norms as these will hold the most meaning and have similar genetic phenotypes relating to COVID-19 immunity. |

## *References*

Anderson, TK 2009, 'Kernel density estimation and K-means clustering to profile road accident hotspots', *Accident Analysis & Prevention*, vol. 41, no. 3, pp. 359-364.

Atchadé, MN & Sokadjo, YM 2022, 'Overview and cross-validation of COVID-19 forecasting univariate models', *Alexandria Engineering Journal*, vol. 61, no. 4, pp. 3021-3036.

Burki, T 2022, 'COVID-19 in North Korea', *The Lancet*, vol. 399, no. 10344, p. 2339.

El-Shabasy, RM, Nayel, MA, Taher, MM, Abdelmonem, R & Shoueir, KR 2022, 'Three wave changes, new variant strains, and vaccination effect against COVID-19 pandemic', *International Journal of Biological Macromolecules*.

Gnjatović, M, Košanin, I, Maček, N & Joksimović, D 2022, 'Clustering of Road Traffic Accidents as a Gestalt Problem', *Applied Sciences*, vol. 12, no. 9, p. 4543.

Gohari, K, Kazemnejad, A, Sheidaei, A & Hajari, S 2022, 'Clustering of countries according to the COVID-19 incidence and mortality rates', *BMC Public Health*, vol. 22, no. 1, pp. 1-12.

González-Collazo, SM, del Río-Barral, P, Balado, J & González, E 2022, 'Detection of direct sun glare on drivers from point clouds', *Remote Sensing*, vol. 14, no. 6, p. 1456.

Holmgren, J, Knapen, L, Olsson, V & Masud, AP 2020, 'On the use of clustering analysis for identification of unsafe places in an urban traffic network', *Procedia Computer Science*, vol. 170, pp. 187-194.

Ket, P, Dhembare, A, Bhide, S & Kolhe, S 2020, 'Accident Black Spot Detection on Greater Mumbai Region', in *Proceedings of the International Conference on Recent Advances in Computational Techniques (IC-RACT)*.

Klyavin, V, Sysoev, A, Drurechenskaya, A & Mamedov, A 2021, 'Approaches to Traffic Accidents Clustering to Form Effective Marketing Campaign', in *2021 3rd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA),* IEEE, pp. 978-980.

Kumar, S & Toshniwal, D 2016, 'Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC)', *Journal of Big Data*, vol. 3, no. 1, pp. 1-11.

Riley, JC 2010, 'Smallpox and American Indians revisited', *Journal of the history of medicine and allied sciences*, vol. 65, no. 4, pp. 445-477.

Schwartz, R, Vassilev, A, Greene, K, Perine, L, Burt, A & Hall, P 2022, 'Towards a Standard for Identifying and Managing Bias in Artificial Intelligence'.

Shao, D & Xiong, W 2022, 'Does High Spatial Density Imply High Population Density? Spatial Mechanism of Population Density Distribution Based on Population–Space Imbalance', *Sustainability*, vol. 14, no. 10, p. 5776.

Sinclair, C & Das, S 2021, 'Traffic accidents analytics in uk urban areas using k-means clustering for geospatial mapping', in *2021 International Conference on Sustainable Energy and Future Electric Transportation (SEFET),* IEEE, pp. 1-7.