**Data 422 Group Assignment**

A mock Report to the Christchurch City Council advising the locations and conditions regarding serious and fatal traffic collisions.

**Contributors**:

Kenneth Rosal (kro109)

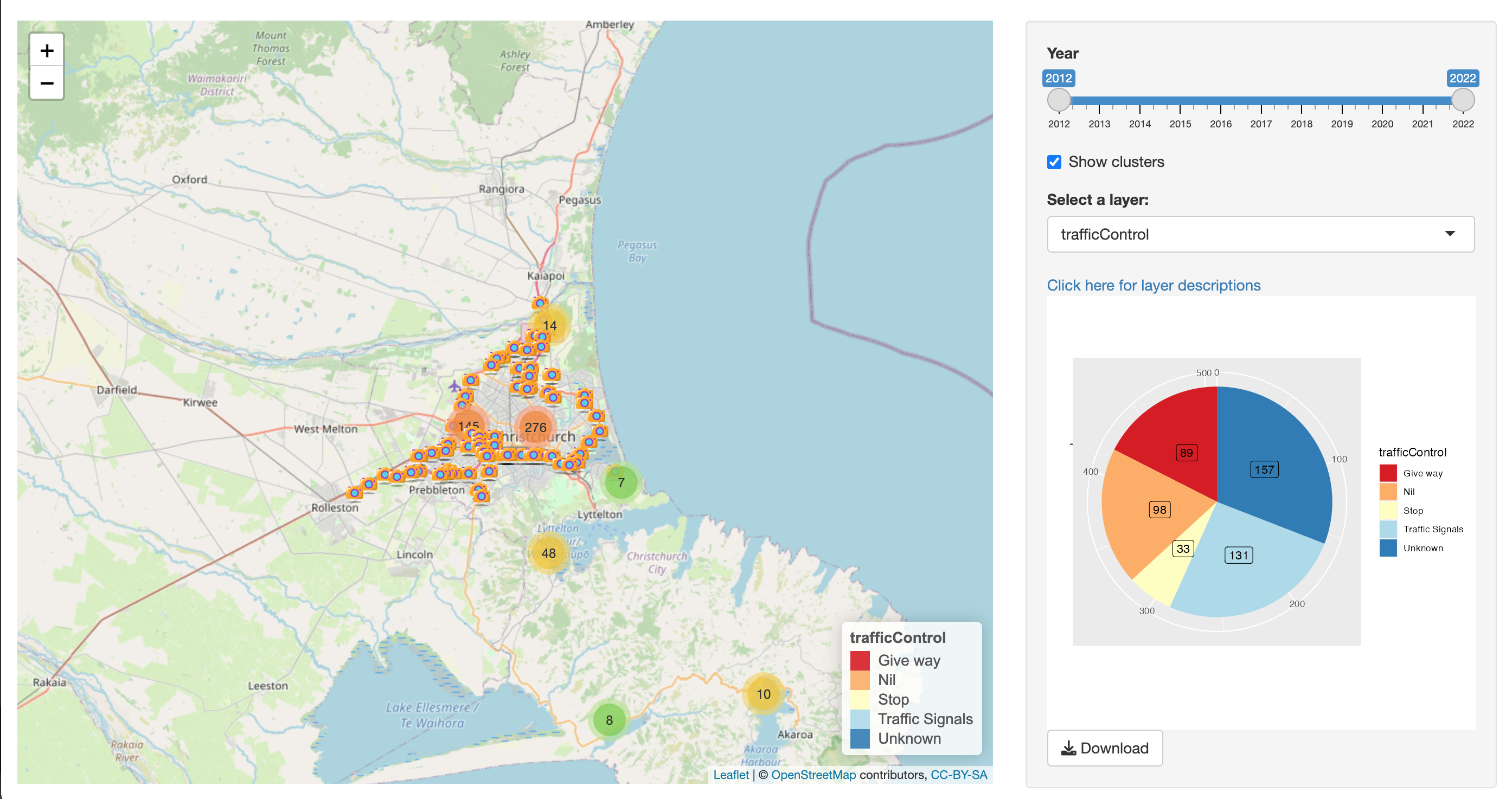
Kyle Hargraves (kha166)

Miguel Montalla (mmo191)

Xiaoshi Xie (xxi53)

**Purpose**

The purpose of the project is to develop a mock report to the Christchurch City Council which identifies the location of serious and fatal accidents defined by the Crash Analysis System (CAS) at the NZ Transport Agency within Christchurch. The report may motivate some additional research by the Council with the objective of developing initiatives to reduce the incidence of collisions at the locations identified.

**Intended Audience**

The project ought to be of interest to anyone who is in an administrative local government position within Canterbury or anyone associated with the transport industry or indeed any member of the public concerned with reducing the incidence of fatal or serious traffic collisions.

**Methodology**

The question that confronts any data scientist is how best to present the data to obtain

1. the maximum information from the dataset and
2. to present the data (tables, graphs, descriptions) in a comprehensive and comprehensible manner.

The question does not lend itself to an uniform answer but is contingent upon the nature or completeness or integrity of the data and, to a large extent, the intended audience. The technical specifics are often omitted for a general audience.

This project is obliged to address the forgoing question also. One approach may well be to consider the location with the least number of missing values and from that location collate the instances by street name as a group and count the instances. A second approach, which is the adopted approach for the project, is to consider the two locations provided [referred to in the data file] as 'location1' and 'location2' and undertake a frequency of collisions upon that basis. The former approach is likely to provide a greater number of incidents of collision but it is deemed by the group that the serious collusions identified will be representative of serious collusions in general.

The information obtained from this approach is presented in summary as a table and also graphically as a heat

map {or interactive map}

**Data Sources**

The principal source for the project was the (so called) Crash dataset containing nationwide collisions that are available from the NZTA website[[1]](#footnote-1) as a csv file. The file was fitered for the Christchurch region in Microsoft Excel and the result was exported to a new file called "Crash\_data\_chch.csv" which contained almost 52,000 rows and 72 columns recording official traffic accidents over the last 22 years (from 2000). The data for 2022, at the time of writing of the report is acknowledged as incomplete. For this type of undertaking there is no significant alternative source.

**General Wrangling**

From the initial dataframe the content was filtered for events post 2010 on the basis that serious collisions over the preceding decade would be of interest to the Council.

Former issues may well have been addressed but, moreover, attention needed to be given to the the more current road-grid. To this end the time interval applied was from 2012 to 2022. When filtered for serious or fatal accidents over the interval there were approximately 1,600 records.

A number of options presented themselves for collating frequency counts by location regarding serious or fatal collisions. Moreover, there are a range of vehicles to consider along with speed restrictions etc. Ultimately, an accident count for each location per year was obtained with the results ranked the the count. We chose the combination of crashLocation1 and crashLocation2[formal field names] instead of NZGD2000 coordinates [field names X and Y respectively] as geographical reference for grouping. Location by street name provided the desired resolution for "hot spots"

For a bird's-eye appreciation of serious collisions a location count was aggregated over the interval of interest (2012-2022). This approach provided a bit over 200 locations where there had been one or more serious collusion but when appealing to coordinates (more than one set of coordinates for a given location) the number of records increased to a bit of 500

For the purpose of making the presentation interactive by user choice or selection the quantity of columns was reduced to apparent attributes that reflected environmental or pertinent conditions such as vehicle type, speed zone and weather.

The dataframe for interactive use is reduced to the following (user selected) attributers:

lat, lon, crashLocation1, crashLocation2, crashYear, trafficControl, light, advisorySpeed, bicycle,

cliffBank, flatHill, guardRail, motorcycle, NumberOfLanes, roadLane, speedLimit, streetLight,

trafficIsland, weatherA, weatherB

which will be utilised by the leaflet package and R shiny to build an interactive map.

Julia has been utilised to create a .csv file that has been webscrapped from NZTA traffic cam API that provides information obtained by the camera markers.

**Ethical Considerations**

For some decades, indeed from the inception of computer usage by 'industry' in the largest sense of the word, there has been a tacit distinction between behaviour in regard to computers and that of filing cabinets or office draws. In general the law prevents intrusions of filing cabinets etc via trespass but specific legalisation has been enacted in all countries with with regard to 'computer hacking'. It would appear, from a legal point of view, that the latter activity is deemed more detrimental than the former.

Nevertheless, Garczarek and Steuer (2019, p. 2) observe that "There is currently no generally agreed definition

of data science"[[2]](#footnote-2)

but the authors do cite an reference noting that

1. data science should become a profession in the sense of a "grounding" in science, and a responsibility to society in the context that
2. exploratory statistics is a historical predecessor of data science.

The principal ethical dilemma for data science is concerned with the balance between individual privacy and the right of access to reliable information about the country and the location in which one resides. The Royal Statistical Society advocates five principles but the two main principles are to:

1. seek to enhance the value of data science for society : i.e. recognise a professional responsibility and
2. avoid harm : (of any form; financial or reputation) to others and, by implication, self

**Results (and mini user manual)**

**Conclusion**

**Project Organisation and the use of GitHub** ([www.github.com](http://www.github.com/))

A ‘repo’ was created on github that also included a ‘wiki’ page for general correspondence. A dedicated folder for the minutes of meeting became part of the repo that is identified in footnotes 3 & 4. The wiki and the minutes serve as the project diary.

This "Crash\_data\_chch.csv" was put in a public repository on Github[[3]](#footnote-3). The source code for this project can be found in a repository called DATA422-Group-Project[[4]](#footnote-4). The notebook should run in situ although to install the necessary libraries the removal of a ‘#” (or un-commenting) will be required.

On running the code the interactive map on the user’s localhost becomes available with the port specified in the output. A deployed version that is also hosted on the Rshiny server by visiting https://mmo191.shinyapps.io/deploy422

**References**

Garczarek, U. Steuer, D. 2019 Approaching Ethical Guidelines for Data Scientists Research gate

https://www.researchgate.net/publication/330410013

Floridi1, L. Taddeo1, M. 2016 What is data ethics? Oxford Institute, University of Oxford,

http://dx.doi.org/10.1098/rsta.2016.0360

Royal Statistical Society (RSS) 2019 A Guide for Ethical Data Science

1. https://opendata-nzta.opendata.arcgis.com/datasets/NZTA::crash-analysis-system-cas-data-1/explore?location=12.136558%2C0.000000%2C1.97 [↑](#footnote-ref-1)
2. concurring with the Royal Statistical Society (RSS) p. 2 [↑](#footnote-ref-2)
3. https://github.com/zelta1990/crash\_data\_nz [↑](#footnote-ref-3)
4. https://github.com/zelta1990/DATA422-Group-Project [↑](#footnote-ref-4)