Alcohol and Crime in Neighbourhoods in the Australian Capital Territory

# Part 1 – Background and Problem

## Discussion of Background

Over the last several years in Australia, there has been strong concerns about alcohol and violent crime. The Australian state of New South Wales introduced laws restricting trading hours for businesses supplying alcohol to the public, such as bottle shops and bars.

Since the introduction of the laws, there has been a rigorous debate whether there is an association between the supply of alcohol and violent crime. Many businesses have complained there is not and want the laws repealed.

The Australian Capital Territory currently does not have New South Wales-style alcohol control laws. The study aims to inform policy and policing decision-making in the Australian Capital Territory to ensure policy and policing decisions have a data-driven basis. The target audience is the Australian Capital Territory Government.

## Description of Problem

There has been no research (the author is aware of) examining alcohol and violent crime in the Australian Capital Territory. To address this gap, the present study examines whether neighbourhoods with more venues that supply alcohol have higher levels of violent crime than neighbourhoods with fewer of these venues. This research could support informing policing decisions to prevent and address violent crime. The problem is important to solve to ensure there is objectively informed decisions for alcohol and crime control.

# Part 2 – Dataset and Solution

## Overview of Dataset and Solution

The study will use crime data from the Australian Federal Police, neighbourhood population estimates from the Australian Capital Territory Government and venue data from Foursquare. An OPTICS variable-density DBSCAN-based cluster analysis will be used on both crime per capita figures and venue data to determine whether violent crime and alcohol venues have similar feature clustering. (OPTICS clustering searches for clusters based on local density within a dataset and is like DBSCAN by having a concept of outliers). A brute force model will also be used to test whether suburbs in the same clusters also appear in the same cluster in each other dataset (that is code will check one-by-one whether a suburb that is in the same cluster as another suburb in one set of clusters is also in the same cluster as that suburb in the other set of clusters).

## Description of Crime Data

The Australian Federal Police publish crime data for the Australian Capital Territory at <https://www.data.act.gov.au/Justice-Safety-and-Emergency/ACT-Crime-Statistics/2egm-dieb>, including features such as:

* counts of types of crime (such as assault, homicide and sexual assault)
* a breakdown of each quarter (such as Q2 Apr-June 2019)
* a breakdown by neighbourhood (Amaroo, Lyneham etc.).

The Python3 geopy.geocoders package will be used to add longitude and latitude values to the above dataset. Only violent crime will be examined. Canberra neighbourhood population estimates will also be used to transform the data to per capita crime rates.

This data has been manually manipulated in Excel to normalise its layout and uploaded again to <https://raw.githubusercontent.com/caseyj2/IBM-Data-Science-Capstone/master/Crime%20in%20the%20Australian%20Capital%20Territory%2C%201%20July%202018-30%20June%202019.csv>.

The investigation uses crime statistics for the Australian fiscal year that ended 30 June 2019.

## Description of Foursquare Data

Foursquare contains venue data for the Australia Capital Territory. The study will use this data to extract the details of venues that supply alcohol (bars, bottleshops etc.). Examples of the features used will be:

* venue name
* longitude
* latitude.

## Description of Neighbourhood Population Estimates

Population estimates are available from the Australian Capital Territory Government via the API endpoint http://www.data.act.gov.au/kci6-ugxa. The dataset's features are population estimates by:

* neighbourhood
* sex
* age.

For this project, the estimates will be aggregated and only pivoted by neighbourhood, because the study is not examining variances by sex and age.

# Part 3 - Methodology

## Inferential Statistics Used

This study did not use any inferential statistics, because the data available was population-level or estimates (not from random samples). Inferential statistics (t-tests, ANOVAs etc.) generally have assumptions, such as the central limit theorem. The data sources were not able to be validated against these assumptions.

## Machine Learning Techniques Used and Why

The study used the unsupervised machine learning technique called OPTICS clustering. This is similar to dbscan, but examines the local density of data to be able to accomodate datasets with clusters of varying density. Reasons why for this approach included:

* upon manual inspection, the data sources showed features with clusters of varying densities
* the venue dataset from Foursquare was dimension and using OPTICS eliminated the need for parameter hypertuning (such as determining the number of clusters etc. in k-means, search radius in DBSCAN etc.)
* OPTICS is in sklearn and a well developed and reliable algorithm.

## Exploratory Data Analysis

### Overview

OPTICS clustering was used on per capita crime data and venue data from Foursquare to extract clustering from the datasets. Initially, K-means was used; however, the results were not meaningful and so the analysis changed to use OPTICS because of being able to deal with variable density data with outliers. Python code was developed to check if there were common patterns of clustering in the two datasets. No common patterns were found.

### Packages Used

The analysis used the six nonstandard packages below. Standard Python3.7 objects were also used (such as sleep etc.). Versions were set to avoid compatibility issues on future re-runs. A description of each package is below.

* pandas (Version 0.25.3) - pandas provided data analysis and wrangling features.
* sodapy (Version 2.0.0) - Sodapy allowed interfacing more simply with the Australian Capital Territories Government's data endpoints.
* scikit-learn (Version 0.22) - scikit learn provided OPTICS clustering and supporting functions.
* foursquare (Version 1!2019.9.11) - foursquare provided venue data
* geopy (Version 1.20.0) - geopy allowed attributing neighbourhoods to geocodes and vice-versa.

### Wrangle Crime Data

For the crime data, a multi-step wrangling process was used. From the source, the data was not suitable for use in pandas, due to an unorthodox format. To fix this, I manually re-formatted the data in Excel to be in a useable format. Then, I used pandas to import the data, translate the neighbourhoods to the same set of identifiers as in the other dataset. Neighbourhood population estimates were used to transfer the data to a per capita basis.

### OPTICS Clustering

After transforming the data to a per capita basis, OPTICS clustering was used to group the data into clusters. OPTICS clustering was also used to group the venue data into clusters.

# Part 4 - Results

## Clustering of Crime Data

### Outliers

ACT East, Ainslie, Amaroo, Aranda, Banks, Bonner, Braddon, Calwell, Campbell, Chapman, Charnwood, Chifley, Chisholm, Conder, Cook, Crace, Curtin, Deakin, Dickson, Downer, Duffy, Dunlop, Evatt, Fadden, Farrer, Fisher, Florey, Flynn, Forde, Forrest, Franklin, Fraser, Gilmore, Gordon, Gowrie, Greenway, Griffith, Gungahlin, Hackett, Hall, Harrison, Hawker, Higgins, Holder, Holt, Hughes, Isaacs, Kambah, Kingston-Barton, Kowen, Latham, Lawson, Lyneham, Lyons, Macarthur, Macquarie, Majura, Mawson, Mckellar, Melba, Mitchell, Monash, Narrabundah, Ngunnawal, O'Connor, O'Malley, Oxley, Page, Palmerston, Pearce, Reid, Richardson, Rivett, Scullin, Spence, Stirling, Theodore, Torrens, Wanniassa, Weetangera are outliers.

### Cluster 0

Cluster 0 includes City, Garran, Hume, Molonglo and Phillip. Cluster 0 has the highest rates of all types of violent crime and the highest rate overall.

### Cluster 1

Cluster 1 has Belconnen, Bruce, Nicholls, Waramanga, Weston and Yarralumla. Cluster 1 has the second highest rates of all types of violent crime, except family violence which is second lowest. Cluster 1 also has the second highest rate overall.

### Cluster 2

Cluster 2's neighbourhoods are ACT South West, Bonython, Casey, Giralang, Isabella Plains, Macgregor. Cluster 2 has the lowest rates of all violent crime, except family violence which is second highest and other offences against a person which is second lowest. Cluster 2 also has the lowest rate overall.

### Cluster 3

Cluster 3 has Acton, Kaleen, Red Hill, Turner and Watson. Cluster 3 has the second lowest rates of all violent crime, except family violence and other offences against a person that were lowest. Cluster 3 also has the second lowest rate overall.

## Clustering of Venue Data

### Outliers

The neighbourhoods Acton, Ainslie, Aranda, Barton, Belconnen, Braddon, Bruce, Calwell, Chapman, Charnwood, City, Crace, Curtin, Deakin, Downer, Evatt, Flynn, Forrest, Franklin, Fyshwick, Garran, Greenway, Griffith, Gungahlin, Hall, Holder, Holt, Hume, Kaleen, Kambah, Kingston, Lyneham, Mawson, Mckellar, Mitchell, Narrabundah, Ngunnawal, Nicholls, O'Connor, Page, Phillip, Stirling, Turner, Wanniassa, Waramanga, Watson, Weston and Yarralumla were outliers.

### Cluster 0

Cluster 0 contains Amaroo, Bonner, Conder, Dickson, Hawker, Oaks Estate and Pearce. Cluster 0 has the lowest number of alcohol-related venues in the Foursquare data.

## Cluster 1

Cluster 1 has the neighbourhoods of Chisholm, Duntroon, Hughes, Macquarie, Melba and Reid. Cluster 1 has the highest number of venues.

## Overlap of Clustering

The analysis shows no overlap between the clusters for the crime data and those of the venue data.

# Part 5 - Discussion

The analysis suggests there is no association between the availability of alcohol-related venues and patterns of violent crime, for clustering from an OPTICS clustering analysis. There were many outliers within the analysis; however, and so there is a need for subsequent research to validate the results. The discovery of clusters for patterns of violent crime also merits further research to determine the drivers of the patterns.