T.LaFrance – Capstone Final Project

**2.0 Data**

**Data Availability and Sources**

The data needed fall into two regimes; one dataset related to location of the businesses of interest and one dataset related to the crimes and the locations of the crimes. For this pilot study, an urban area is likely to provide a good set of data. At first glance a city like New Orleans, New York, LA or Chicago might be ideal. Choosing the entire United States or all of those cities is not ideal as many states have jurisdiction over their crime records and they keep them in certain locations and formats. Some data seems to be available in reports and files for the US as a whole, but it is largely high-level summary data and does not contain the level of detail needed.

The retail locations appear to be accessible using Foursquare for venue search to obtain type and location. Foursquare data has the advantage of being easily plotted on maps. One concern about Foursquare might be that some venues might be classified in a way that is not helpful for our analysis. For example, a country club that serves substantial amounts of alcohol might not be captured as a retail alcohol location.

The crime incidents and locations appear to be accessible in datasets that are available as court “Open Records” in the county in which the crime was committed. A quick survey of a few of these indicates these datasets are retrievable as .json or other formats and a description of what fields are in each dataset is usually available. Previewing some of these gives some insight; some datasets list the incident, the date, time, arresting officer, and a plethora of other information, but not details about the crime. Other data sets give the crime in a description field, but only give a general geographic location or an X, Y co-ordinate which is not helpful. Some had a street address, but the majority of the address fields had NaN values. These present a problem for the information sought, but so many datasets are out there that one will be found to contain functional information.

Focusing on a certain year or a recent few years will likely seem appropriate to control that amount of data. It will have to be explored as to whether all crimes should be looked at or if they could be filtered to those that are most related to alcohol. This is not completely straightforward either as, in recent years, the terms for intoxicated drivers are a multitude; DUI, DWI, OVI, OWI and sometimes other terms are used.

**Data Utilization**

The expectation is to retrieve data for alcohol related establishments for the city/urban area in question from Foursquare and plot these locations on a Folium map. The second part of the analysis is to pull crime data and locations from a relevant county crime dataset in the selected city/urban area and plot that on a Folium map. By combining information from these two sources, a visualization is expected that will allow the user to look for trends and investigate specific regions and make some suggestions for areas that are at greater risk for criminal activity, possibly induced by alcohol.

**Sources of Relevant Information**

<https://www.cdc.gov/motorvehiclesafety/impaired_driving/impaired-drv_factsheet.html>

<https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/tables/table-18>

<https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812450>