Restaurant Analysis

Introduction:

Entrepreneurship is inherently risky. It becomes even more risky if the venture is consumer facing in a geographic location with which you're not familiar. Restaurant businesses are one the most difficult businesses to run because the entrepreneur has to cater multiple factors some of which may not be in their control. Let's imagine we are an entrepreneur who wants to open a restaurant in Chicago but he is unfamiliar with the city and the type of people who visit the restaurant.

To find a favourable area to open the restaurant, we must focus on crime. Chicago has been in the news lately because of increasing street crime. Although, the violence may be predictable and specific to certain districts, as an entrepreneur who is new to Chicago one would want to know where the crimes are taking place. The restaurant business will be severely affected by street crime especially violent crime. Thus, as a risk averse entrepreneur we must explore Chicago's crime trends.

After, identifying safe districts I would also like to look at the famous venues in those districts to get a better idea of the type of restaurant visitors in that area. Customer segmentation is a very important part of any business.

Data:

For crime data, we will be using this dataset from data.cityofchicago.org. This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. order to protect the privacy of crime victims, addresses are shown at the block level only and specific locations are not identified. These crimes may be based upon preliminary information supplied to the Police Department by the reporting parties that have not been verified. The preliminary crime classifications may be changed at a later date based upon additional investigation and there is always the possibility of mechanical or human error. Therefore, the Chicago Police Department does not guarantee (either expressed or implied) the accuracy, completeness, timeliness, or correct sequencing of the information and the information should not be used for comparison purposes over time. The Chicago Police Department will not be responsible for any error or omission, or for the use of, or the results obtained from the use of this information. All data visualizations on maps should be considered approximate and attempts to derive specific addresses are strictly prohibited.

The dataset has more than six million data points and is greater than 1.8 GB in size. Since data from the past may not be relevant, only last five years data is used. As such a large dataset may not fit in a laptop, the dataset has been split, saved and uploaded from pickle format. For this paper, we have limited our analysis only to the type of crimes that are relevant to restaurant businesses. The data is content-rich since it has 30 columns. The geographical data has been used for geospatial analysis.

For more information about venues in the favorable areas, we will use Foursquare API. Data about 200 most popular venues was scraped. Using community area, districts, postal codes, zip

codes, etc data from the previous data, the venues only in favourable districts were analysed to see what type of visitors visit public spaces in those community areas and districts.

Methodology:

The crime dataset contained data from 2001 till the present date. As the data from year earlier is not relevant to the current crime situation of Chicago, the analysis has been limited to the last five years only. The dates in the data set were not datetime objects thus the date column in the dataset was converted to datetime object.

Changing Date Column in to datetime.datetime object

In [10]:
```
# Changing the data column in to datetime.datetime objects

crime_data['Date'] = crime_data['Date'].apply(lambda x: datetime.datetime.strptime(x,"%m/%d/%Y %I:%M:%S %p"))
crime_data['Date'].dtypes
```

Out[10]: dtype('<M8[ns]')

After the conversion, the dataset was filtered and all the observations before 2014 were removed.

Filtering the dataframe. Keeping only last five years data

In [11]:
```
# Since only last five years data is relevant to our problem
crime_data = crime_data.loc[(crime_data.Date.dt.year > 2014)]
crime_data.shape
```

Out[11]: (1193164, 30)

The filtered data was saved in pickle format so that laptops with low memory can access it too.

Saving the data in pickle so that it can be used on a low memory system

In [13]:
```
# save cleaned data to pickle file for easier loading from notebook start

crime_data.to_pickle('crime_data.pkl')
print('pickle size:', os.stat('crime_data.pkl').st_size)
```

```
pickle size: 283949523
```

In [3]:
```
crime_filtered = pd.read_pickle('crime_data.pkl')
crime_filtered.shape
```

Out[3]: (1193164, 30)

There are many crimes that happen in a particular city. Only some of them are relevant to the restaurant business. Thus, the dataset was filtered and limited only to relevant crimes.

Slide Type

Crimes Affecting Open Restaurants

In [5]:

```
# create and preview dataframe containing crimes that may affect a restaurant business
#Ten crime types were selected out of thirty three types

colm = ['Date','Primary Type','Arrest','Domestic','District','Community Area',
        'Zip Codes','X Coordinate','Y Coordinate','Latitude','Longitude']
res_crimes = crime_filtered[colm]
res_crimes = res_crimes[res_crimes['Primary Type']\
                .isin(['ASSAULT','PUBLIC INDECENCY','ROBBERY','INTIMIDATION','CRIMINAL TRESPASS','PUBLIC PEACE VIOLATION',
                       'CRIM SEXUAL ASSUALT','HOMICIDE','NARCOTICS','WEAPONS VIOLATION','NARCOTICS','PROSTITUTION','THEFT'])]

# clean some rouge (0,0) coordinates
#res_crimes = res_crimes[res_crimes['X Coordinate']!=0]

res_crimes.head()
```

Now, that we had data for all the relevant crimes and relevant years, the cluster analysis was performed to identify districts for which the crimes was high. For this purpose, one-hot encoding was performed to bring the dataset into the required format.

In [13]:

```
cr_m = pd.merge(cr_grouped, cr_loc, on ='District')
cr_m.head()
```

Out[13]:

	District	ASSAULT	CRIMINAL TRESPASS	HOMICIDE	INTIMIDATION	NARCOTICS	PROSTITUTION	PUBLIC INDECENCY	PUBLIC PEACE VIOLATION	ROBBERY	THEFT	WEAPONS VIOLATION	Latitu
0	1.0	0.077350	0.058369	0.000725	0.000925	0.019031	0.001500	0.000175	0.009578	0.049791	0.778778	0.003776	41.8759
1	2.0	0.183064	0.053856	0.005543	0.001487	0.067646	0.002884	0.000090	0.010636	0.112759	0.534544	0.027491	41.8101
2	3.0	0.228812	0.061919	0.007752	0.001634	0.098155	0.000794	0.000000	0.016530	0.127761	0.406864	0.049778	41.7711
3	4.0	0.237039	0.061397	0.006664	0.001040	0.122833	0.002196	0.000077	0.018566	0.091672	0.400008	0.058509	41.7348
4	5.0	0.230083	0.064984	0.009455	0.001200	0.120033	0.013198	0.000096	0.016078	0.085621	0.380255	0.078998	41.6885

Slide Type

Crime and location data for each of the districts was used to cluster the districts into low and high crime clusters. As we were interested in two clusters, the value of K was chosen as 2.

Slide Type

K-Means Clustering

In [14]:

```
from sklearn.cluster import KMeans
# set number of clusters
kclusters = 2

cr_clustering = cr_grouped.drop('District', 1)


# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(cr_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_
```

Out[14]:

```
array([1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0,
       1])
```

After identification of safe districts, venue information for the city of Chicago was extracted using the Foursquare API. Using zip codes from the previous dataset, we focused on the venues of only safe districts. After getting information about venues, it enabled used to analyse the type of

visited places in different areas. This analysis improved our understanding about the consumer segments present in those areas. Based on this information, we can make informed decisions about our own restaurant.

Results:
After clustering, two clusters were identified. For the districts in **Cluster 0**, the crime rate is **high**. We can see the mean values for Asssault, Narcotics, Prostitution, Robbery, and Weapons Violations are considerably high for districts in Cluster 0. Thus, as per crime data, entrepreneurs should avoid the districts in Cluster 0. For the districts in **Cluster 1**, the crimes rates are relatively **very low**. These districts are very good choice for opening a restaurant as the negative externalities are reduced. Thus, the areas in cluster 2 were our areas of interest.
To add more granularity to the data, the community areas and zip codes were added to the analysis. As a district contains many community areas, the favourable community areas increased our understanding of safe areas in Chicago city.
Using the Foursquare API, we identified 1700+ venues in our areas of interest which gave us more information about particular areas.

Discussion:

The following Chicago districts were identified as low crime districts: 1, 2, 12, 14, 16, 17, 18, 19, 20, 24, and 31. The number of safe community areas in these districts are 45.

Conclusion:

In conclusion, we have identified eleven districts that are very safe for consumer facing business. Moreover, we have also identified the famous venues and their location in this particular districts which gives us rich information regarding the consumer psychology of potential customers.