CAPSTONE PROJECT:

FIND THE KEY AREAS TO DO BUSINESS IN BANGALORE, INDIA.

Introduction / Business Problem:

- **In a highly competitive environment where everyone wants to be in the prime areas, it is always a problem to find the "Best Areas" in a city to start any form of business.**
- **Depending on the type of business the "Best Areas" can be filtered. If the business is B2B then the "Best Areas" can be filtered as per other required businesses present in the vicinity. If it is a B2C model, then malls and street food chains can be used to filter the appropriate areas.**
- **There is also a need to identify the least popular areas to start a business as the expenditure on the real estate in these areas would be a lot cheaper.**
- **An algorithm which throws the "Best Areas" in Bangalore based a set criterion can come in handy.**
- **In this Capstone I am trying to list out the top 5 places in popularity in Bangalore, INDIA. Popularity is defined by:**
    - **clusters of number of venues present within vicinity**
    - **number of houses present in that area – Which gives a rough estimate of the working people present in that area.**

DATA:

- **I will be using the data from:**
  https://raw.githubusercontent.com/suvajit/opendata/master/BBMP/data/CSV/BBMPwards.csv

- **A snip of the Dataframe:**

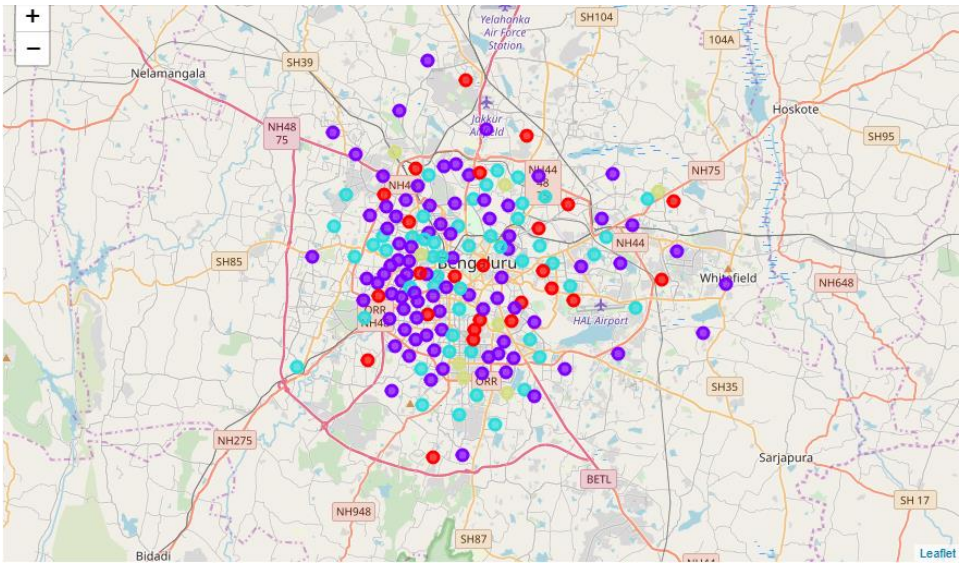
OBJECTID	ASS_CONST_	ASS_CONST1	WARD_NO	WARD_NAME	TOT_HH	POP_TOTAL	POP_M	POP_F	POP_SC	POP_ST	POP_LIT	POP_WORK
186	150	Yelahanka	1	Kempegowda Ward	8647	34783	18197	16586	2816	1097	27748	14794
1	150	Yelahanka	2	Chowdeswari Ward	9506	36602	19060	17542	3941	810	27160	16865
2	150	Yelahanka	3	Atturu	14605	58129	30799	27330	6480	1859	46738	23818
3	150	Yelahanka	4	Yelahanka Satellite Town	10583	41986	21799	20187	6319	1065	33599	17722
13	152	Byatarayanapura	5	Jakkuru	12387	52025	27269	24756	6423	973	37879	20445

- **Description of the column names are as follows:**
  - **ASS_CONST_ – Constituency Number**
  - **ASS_CONST1 – Constituency Name**
  - **WARD_NO – Ward #. (Wards exist within Constituency, like an area)**
  - **WARD_NAME– Name of the WARD**
  - **TOT_HH– Total houses present in that area**
  - **POP_TOTAL– Total population in that area**
  - **POP_M– Total Population male**
  - **POP_F– Total Population Female**
  - **POP_SC– Total population of Scheduled Caste community**
  - **POP_ST- Total population of Scheduled Tribe community**
  - **POP_WORK– number of people who are working**
  - **AREA_SQ_KM- Area of the WARD/AREA in sq KM.**
  - **LAT – Latitude of the WARD/AREA**
  - **LON – Longitude of the WARD/AREA**
  - **RESERVATIO – majority of the people present here fall under which reservation category. “General” means that they do not fall under any reservation.**
- **NOTE: The data provided belongs to a CENSUS done in 2000.**
- **I am using the GeoJSON file from: https://raw.githubusercontent.com/openbangalore/bangalore/master/bangalore/GIS/bangalore_pincode.json**
- **For finding the neighbouring venues I am using the api provided by Foursquare.**

METHODOLOGY:

- **Data Preparation –WARD_NO, WARD_NAME, TOT_HH, LAT and LON are required from the obtained Dataframe. We need to convert the TOT_HH to a normalised value so that it makes it easier to compare and group to a cluster.**
- **Using the Foursquare API to get venues and their related info.**
- **Store the Venue Names, categories, Latitude and Longitude into a dataframe.**
- **Get the nearby areas belonging to the same neighbourhood from the above dataframe.**
- **Analyse each neighbourhood and check for which venues are present in that neighbourhood. (Using get_dummies). Use the frequency of the venues being repeated in that area and provide a list of neighbourhoods with their top 5 venues.**
- **Use the normalised TOT_HH and the neighbourhoods with their top 5 venues to train a K-means clustering model.**
- **I am considering 5 clusters/5 areas to show as "BEST AREA".**
- **Use the labels got from training the model and group all labels.**
- **Plot them on a map using FOLIUM.**

RESULTS:



DISCUSSION:

- **Since I am using the houses present during 2000 and the most current venues from Foursquare to compare, there is a high chance of the map showing me an erroneous neighbourhood.**
- **Foursquare in INDIA is not as thorough as in other parts of the world. The only venues which were thrown by the api were restaurants, pubs, etc (entertainment related). Using other APIs one could find Hospitals, Schools etc.**
- **Depending on the business, the criteria for filtering the dataset can be adjusted before modelling.**
- **I am only picking top 5 categories and their venues for modelling. This again can be increased to give you a more precise location.**

CONCLUSION

Clusters contain frequently visited venues and with similar number of houses. I pick the biggest cluster and pull out the top 5 neighbourhoods to show the "Best Areas". I Plot them individually on a map.