**CAPSTONE PROJECT: THE BATTLE OF NEIGHBORHOODS**

**Expats Location Engine**

1. **Introduction:** Every year a great number of professionals change their home for mid or long term periods, accompanied most of the time by their families. For a newcomer in the town it is very difficult to make the decision on which side of the city to set foot and find an appropriate venue. Target Users: Relocation services, Expats, Exchange Program Services
2. **Data:** For this Location Engine data available on the Foursquare's API will be used. It will be used to check data terms of their neighborhoods. The collected data will be prepared by applying normalization techniques: get rid of null values, duplicate rows, data wrangling and formatting the data i.e. standardization. After the data is standardized the qualified data is ready to be processed. Population, Average Income per Neighborhood, Demographics could be integrated too.
3. **Methodology**

Geo locations will be queried using geocoder library and venues from FourSquare API. The locations for which foursquare doesn't provide data will be droped because not useful for the analysis. Using the amenities provided by foursquare in the neighbourhood, the nearby venues will be clustered to find out the top ten amenities available for each location and extract those features. Based on the user preferences the clusters are analyzed and recommended for living.

The choices are targeted to the different profiles that the interested might have:

* Single
* Family

This will be crossed on a matrix base with other basic dimensions to create the full profile:

* Age
* w/wo children
* w/wo pets
* City center/suburb preference

The list of dimensions might increase with the time in order to refine the profile by using ML algorithms.

User preferences are taken into consideration such as Banks, Shopping Malls, Bus Station, Grocery, Resturants, Coffee Shops, Deli Foods/ Bakeries etc ... and will be find out which ones are best for living.

One of the technique which will be used is KMeans to group neighborhoods with similar venues.

1. **Results**

Based on the amenities available in each area, the city-county combinations are divided into 4 clusters. Each cluster has a unique combination.

For example:

* **Cluster 1** is a group of amenities found in any key junction of neighborhood with parks, shopping, grocery, bank.
* **Cluster 2** is the group which contains needs for tourist ppl such as restaurants and motels for accomodation
* **Cluster 3** is mostly about food joints and super markets
* **Cluster 4** is the grouping of most happening places consisting of restaurants, nightlife, pubs, malls, banks & entertainment centers

1. **Discussion**

Without the initial data exploration and methodology phase, it could not be possible to figure out what are top amenities in the neighborhood that helps in making the decision for living in that area.

Though the data has gone through exploratory analysis, some of the issues can be found during actual run of data. For ex: few locations didn't return geo location and few others didn't return any result for FourSquare API and couldn't find nearby venues.

1. **Conclusion**

Data Science is a highly iterative process which needs going back and forth to tune the data as needed.