**Final Report for the Capstone Project in Data Science**

1. Introduction:

Starting a new business venture by opening an establishment requires a lot of effort and careful consideration. It might be useful to use precedence in this matter to ascertain the success of your new business.

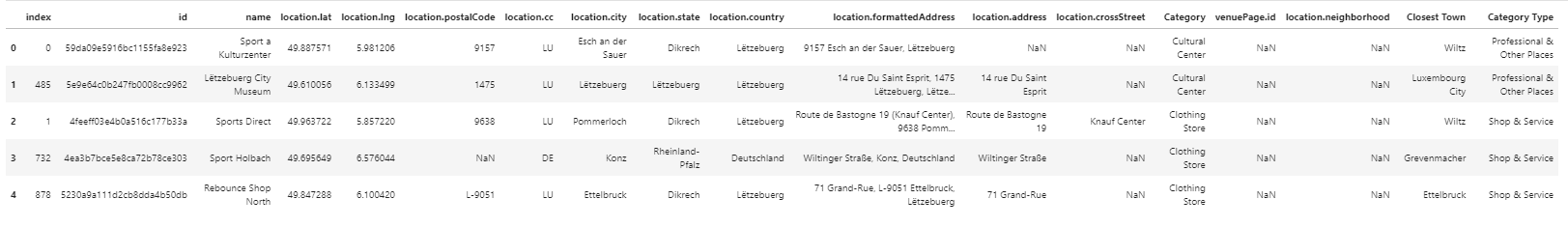
The problem I’m trying to describe and solve in this project is: if you had to start a venture in the city of Luxembourg what location would be most appropriate for you to do so, based on the success of previous businesses of the same category. I believe this maybe of interest to any party who wishes to start a business in the country.

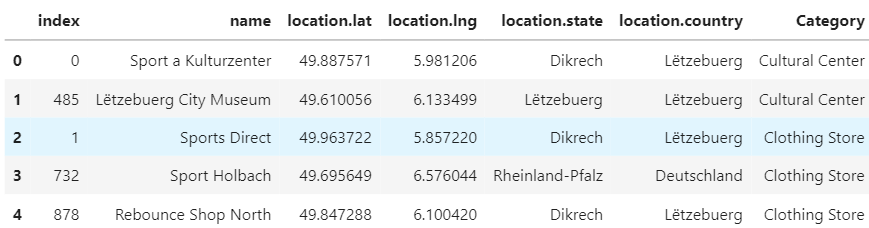
1. Data Acquisition and Cleaning:

The Dataset used in this project has been sourced via Kaggle as a csv file named *luxembourg\_places* (a copy exists in the repository). It has been read into a pandas dataframe (*df\_places)* and the columns are as follows: *index, id, name, location.lat* , *location.lng, location.postalCode, location.cc, location.city, location.state, location.country, location.address, location.formattedAddress, location.crossStreet, Category, Nearest City, Category type, location.neighbourhood and VenuePage.*

The data was cleaned to more precisely fit exclusively to Luxembourg and not surrounding countries. Further, all venues where state or category wasn’t available were dropped from the dataframe. There was also a redundancy in features including country code, cross street, country, etc. which were also dropped considering the area in question was entirely within Luxembourg. Finally, to replace the *nan* values in the *location.state* column, I created a function to find the states for those indices using the foursquare search query with the help of the latitude and longitude information of the venue.

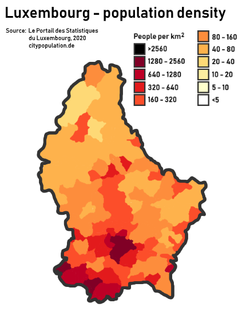
*First five rows of the dataset before and after cleaning:*





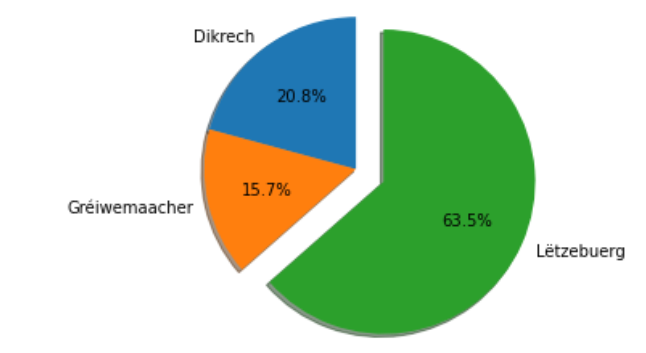
1. Data Analysis:

*Population density of Luxembourg:*

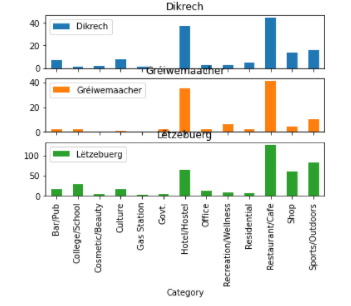


Most of the population of Luxembourg is dense around the southern area of the country, which is the state of Luxembourg, with the state Diekirch being the next most populated region and Grevenmecher is the least populated area. So it is safe to assume that most types of venues would likely be successful in Luxembourg or Diekirch.

*Percentage division of venues in Luxembourg:*



*Division of venues by Category and State*



Then I created a one-hot encoded dataframe by Category which I passed through a function to get the top 10 venues in the State.

Then I used the k-means clustering algorithm, on this dataframe to get the labels for each venue.

1. **Results:**

On passing a vector of length 13 (symbolizing the number of categories; it will be all 0s except in the position which is the category of this instance, which will be 1), we obtain the cluster label of the new entry. We can match that with the Dataframe to get the most appropriate location for the new venue

1. **Conclusion:**

I would hope that this effort helps someone looking to start a new venue anywhere in the world. Machine Learning is a powerful tool and with the volume of data that we have today, should be used extensively to better our situation in any way possible.

*fin*