This submission will eventually become your **Introduction/Business Problem** section in your final report. So I recommend that you push the report (having your Introduction/Business Problem section only for now) to your Github repository and submit a link to it.

1. A description of the data and how it will be used to solve the problem. (**15 marks)**
2. A description of the data and how it will be used to solve the problem. (**15 marks)**
3. 
4. Describe the data that you will be using to solve the problem or execute your idea. Remember that you will need to use the Foursquare location data to solve the problem or execute your idea. You can absolutely use other datasets in combination with the Foursquare location data. So make sure that you provide adequate explanation and discussion, with examples, of the data that you will be using, even if it is only Foursquare location data.
5. This submission will eventually become your **Data** section in your final report. So I recommend that you push the report (having your **Data** section) to your Github repository and submit a link to it.

**Battle of Zip codes – Charlotte Edition**

**Introduction/Business Problem**:

The goal of this capstone project is to develop and deliver recommendation for the best location (zip code) to open a new restaurant in the city of Charlotte NC.

Our client is planning to open a marquee fine dining establishment that serves Indian food in the city of Charlotte. Although there are a few Indian restaurants, our client sees a need for a date-worthy establishment that caters not just to people of Indian origin, but to a wide variety of patrons. Our client, though has an idea of a good location for the restaurant, she would like a recommendation based on data. She’d like to know the best zip code to open the restaurant. She’d also like to understand the parameters used for coming up with the recommendation and how each factor affected the recommendation. She wants the model to be flexible enough to get a new recommendation without starting from scratch, if she were to change her mind about the concept of the restaurant.

About Charlotte, NC:

According to Wikipedia, Charlotte, NC is the most populous city in the state of North Carolina and 17th most populous city in the US. It is also the third most populous city in the US with approximately eight hundred thousand residents. Bank of America and east coast operations of WellsFargo are based out of Charlotte among several other companies. It is also home to a diverse population.

**Approach**

Solving this problem starts with identifying and collecting data that can help us. We need to collect information along the following four dimensions.

* **Crime Rate/Safety:** Safety is of prime concern to the patrons. A restaurant won’t be successful if it is not in a safe location. For the kind of restaurant that our client is considering, we need to treat this factor very important
* **Demographics:** Understanding demographics helps us understand if the location is the right location. A high-end restaurant will not do well in a low-income neighborhood, for example. Among all the data that is being considered, we’d expect this to be of highest importance after safety.
* **Competition:** Number of competing restaurants vs complimentary restaurants are in a location (when combined with demographic data) will give us valuable information on future success of the project. I’d consider this to be the third in the list in the order of importance.
* **Economic activity in the area:** Commerce and economic vibrancy in an area would indicate its viability for new business. To measure economic vibrancy, we will look at all the shopping centers/areas in a particular location. Compare one location relative to the other to determine what is are the preferable locations. Though I would not consider this to be the most important, it nevertheless is a valuable data point in triangulating best location for the restaurant.

Once we collect the data, we will need to profile each zip code along the respective dimensions. For each dimension, we will group all the zip codes that are alike. K-means clustering algorithm will come in handy, in this case.

For the sake of simplicity, we will try to group them into three clusters. We will analyze each cluster and assign a somewhat subjective weight/rating on their desirability as a location for the restaurant.

In some cases, for example – Crime/Safety, we might conclude that a high-crime zip code is categorically undesirable. Where as in other cases, as in the case of economic activity in the area, even if a cluster/zip code rates low, we may not entirely rule it out depending on how the zip code measures on other dimensions.

Once we cluster zip codes along the four dimensions, we’ll evaluate each zip code on all four dimensions and pick the zip code that comes out at the top.

Specific details of how the methodology is used will be elaborated in Methodology section of this report. We will focus on specific data that we collected and used in the next section.

**Data:**

The data that was collected/used for this project is from three sources.

* US Zipcodes database from uszipcodes package in Python
* Data from Charlotte NC Open Data Portal at data.Charlottenc.gov
* Foursquare
* **Crime Rate/Safety:**

The City of Charlotte does not have all the crime data openly available to public. However, they do have three APIs/files for the three categories below

* + CMPD Officer-Involved Shootings – Individuals at <http://data.charlottenc.gov/datasets/cmpd-officer-involved-shootings-individuals-1/geoservice>
  + CMPD Officer Involved Shootings – Officers at <http://data.charlottenc.gov/datasets/cmpd-officer-involved-shootings-officers-1/geoservice>
  + CMPD Officer-Involved Shootings – Incidents at <http://data.charlottenc.gov/datasets/cmpd-officer-involved-shootings-incidents-1/geoservice>

As you can see this data only accounts for crime/safety incidents where a fire-arm was discharged. However, we can safely assume that this data is a good proxy for Safety/Crime in general.

Since I would consider all the incidents above are of same magnitude, I will only interested in latitude and longitude at which the incident occurred and total number of incidents. This will help us in identifying the zip code in which the incident occurred and how many incidents occurred. Since the data set is pretty small, I’ll be using all of the data regardless of when the crime occurred. Please follow the links if you are interested in all the data provided by the API.

Note that I will be using # incidents / zip code population as input to standard scaler while clustering the data (to make comparison apples to apples)

* **Demographics:**

I have used built-in US Zipcodes module in Python to gather demographic information for each zip code. I have taken the following into account while clustering the data

* + Zipcode
  + Total population
  + Population density
  + # of house holds
  + # of occupied homes
  + Median household income
* **Economic activity in the area:**

City of Charlotte provides an API for Existing Shopping Centers at <http://data.charlottenc.gov/datasets/existing-shopping-centers/geoservice>

This API provides information about all shopping centers in and around Charlotte with shape objects for each of the shopping centers. Name, shopping center type (convenience, community, regional, super regional etc.), area of the shopping center and anchor stores along with other details.

Among this data, I will be using the shape (or the boundaries) of the shopping centers to extract latitude and longitude of the shopping center to get the zip code as to where it is located.

Shopping center type and area are also of significance. They are proxies for economic vibrancy in that area and this are crucial for our model.

* **Competition**

I’ll be using FourSquare to gather information about Restaurants in each zip code. Number of restaurants in relation to population density give us a good idea on if a zip code is saturated vs there is an opportunity waiting in plain sight. Category of each of these restaurants will help us understand if the zip is ripe for a specific type of restaurant (or replete with it). I will be using the following data for each zip code:

* + Category
  + Longitude, latitude

I will again be using longitude and latitude to get the zip code for each restaurant. I would be interested in looking at Number of restaurants in each category by zipcode (in relation to population density). This information will be used to cluster the competition to identify opportunities/threats.