1. A short summary:

No matter how complex the data or challenging business needs, my passion for data science will never disappear. The enthusiasm for machine learning, neural networks and their applications in the modern world began a long time ago. Even as a child, I found numbers fascinating and had a unique interest in both statistics and computer science. Moreover, the ability of a data scientist to combine math’s and computer science to find valuable insights from this huge amount of data and give back to world has always fascinated me.

1. Dataset:

For this project, the following data is needed:

• List of neighborhoods in Los Angeles

• Latitude and longitude coordinates of neighborhoods

• Crime data in Los Angeles

• Venues Details

1. List of neighborhoods in Los Angeles data is from Wikipedia website, there is an article

called “List of districts and neighborhoods in Los Angeles” (1). This article is about a list of notable districts and neighborhoods (residents, commercial, and industrial) within the city of Los Angeles, California, present and past. The following screenshot is a part of data from this article.

Graphical user interface, application

Description automatically generated

In order to get the latitude and longitude coordinates of neighborhoods, I will use the Google’s Geocoding API (2) to get the geolocation data for each neighborhood. The Geocoding API is a service that provides geocoding, which is the process of converting addresses into geographic coordinates like latitude and longitude, which I can use to place markers on a map or position the map. When I access the Geocoding API through an HTTP interface, the response is returned in JSON format with the latitude and longitude. For the crime data in LA, this dataset reflects incidents of crime in the City of Los Angeles dating back to 2020. This data is transcribed from original crime reports that are typed on paper.

1. Those datasets are fully satisfying all three required criteria. First of all, those datasets are

fully accessible. Wikipedia is free content that anyone can edit, use, modify, and distribute. And the Google’s Geocoding API and Foursquare API. Secondly, I believe those datasets are released with an open license, because when I go to the privacy policy document on Wikipedia website, I found there is one sentence: “Text is available under the Creative Commons Attribution-ShareAlike License”. Finally, this dataset is machine processible, because the format of this dataset is digital, so it supports to be machine processible.

1. In order to access those datasets, I will use some Python library to parse the Wikipedia

page to get the list of neighborhoods and districts in LA, and then I will first save it as a csv file and then upload this csv file into relational database, it could be MySQL, or might be Microsoft SQL database. After that, I will create some queries to access those data from database.

For the Latitude and longitude coordinates of neighborhoods, I will use the Google’s Geocoding API (2) to get the geolocation data for each neighborhood.

To analyze criminal activity for each neighborhood I will use Los Angeles Crime & Arrest Data: from Beginning 2020 to Present dataset (3) from Kaggle website. It contains information about location, time, category and other miscellaneous data from the LA Police Department.

1. Sketch out a simple data science project:

Los Angeles is a very vibrant city with a lot of neighborhoods, each with unique character. Some neighborhoods are quiet and cozy, has convenient store locations, while others offer a lot of fun and nightlife activities. Choosing a neighborhood to live in or open a business can be a complicated task to do, but with the help of location data from Foursquare and crime data, we can make it a little bit easier. The objective of this project is to analyze and select the best locations in the city of Los Angeles, California to choose a neighborhood to live in or open a new business. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Los Angeles, California, what would be a better place to live in or start a business?

Reference:

(1): <https://en.wikipedia.org/wiki/List_of_districts_and_neighborhoods_in_Los_Angeles>

(2): <https://developers.google.com/maps/documentation/geocoding/start?utm_source=google&utm_medium=cpc&utm_campaign=FY18-Q2-global-demandgen-paidsearchonnetworkhouseads-cs-maps_contactsal_saf&utm_content=text-ad-none-none-DEV_c-CRE_315916118159-ADGP_Hybrid%20%7C%20AW%20SEM%20%7C%20SKWS%20~%20Geocoding%20API-KWID_43700039136946657-kwd-335278985932-userloc_1003571&utm_term=KW_%2Bgeocoder%20%2Bapi-ST_%2Bgeocoder%20%2Bapi&gclid=CjwKCAiA9bmABhBbEiwASb35VxYDeNjxxNqYvcHe-VqP5urIvUHWqmyIJkVOX_ZzAr-CGFQlYyx2dxoC53cQAvD_BwE>

(3): <https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8>

(4): <https://developer.foursquare.com/docs/places-api/>