Battle of Neighborhoods

Finding Relationships Between Similar Clusters of Neighborhoods and Housing Prices

Data

- **kc_house_data**: The dataset has been taken from Kaggle and is widely used for a hands-on experience on training regression models. We will leverage this dataset to add another feature to it and understand if we can find correlations between housing prices and the clusters of neighborhoods that we will be creating further in this project.
  In this dataset, the most important column for our project is the column 'zipcode'. We will find the unique number of zipcodes that we have and later group our entire dataset according to zipcode.

  https://www.kaggle.com/harlfoxem/housesalesprediction

  Below is the list of columns that the dataset has:

id:	a notation for a house
date:	Date house was sold
price:	Price is prediction target
bedrooms:	Number of Bedrooms/House
bathrooms:	Number of bathrooms/House
sqft_living:	square footage of the home
sqft_lot:	square footage of the lot
floors:	Total floors (levels) in house
waterfront:	House which has a view to a waterfront
view:	Has been viewed
condition:	How good the condition is (overall)
grade:	overall grade given to the housing unit, based on King County grading system
sqft_above:	square footage of house apart from basement
sqft_basement:	square footage of the basement
yr_built:	Built Year
yr_renovated:	Year when house was renovated
zipcode:	zip
lat:	Latitude coordinate
long:	Longitude coordinate
sqft_living15:	Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area
sqft_lot15:	lotSize area in 2015(implies-- some renovations)

- **Coordinates using geolocator library**: Once we have the zipcodes, we will try to find the coordinates for each of them using geolocator library and put it together into a dataframe. We will further be using these coordinates to explore nearby venues using Foursquare APIs.
- **Foursquare data**: Once we have the coordinates, we will analyze the neighborhoods for each of the zipcodes using Foursquare data. The data that we will receive will be in a JSON format. We will be extracting the necessary components from the data (categories) and then convert the extracted data into a pandas dataframe using **json_normalize.** We will then cluster the areas defined by zipcodes and later try to find if neighborhood has any impact on prices or not.