Upon my uncle's request to help him find the location for his new dental clinic in the Waltham area, I tried to predict the location that would give the highest profit by using machine learning classifiers from other competitor dental clinics in the area.

First, I collected the data from Google Maps to get the data of all the competitor dental clinics in Waltham. Data collected were address, latitude, longitude, ratings, and reviews. I then collected census tract data of Waltham to get the population and median income data for each census tract where these dental clinics are located. I separated the Waltham district by census tract, which will give more detailed data about the region and put those variables into machine learning classifiers. After collecting the competitor dental clinic data, I collected population center data from the Waltham district, which consisted of the same features from the competitor dataset, and saved the data as a CSV file for later use.

There were a lot of data preprocessing steps to be done after scrapping the data because the data had been collected from the web and it was raw. Therefore, I preprocessed the data by converting string values of ratings, and reviews count to float and removed words such as "#" and "Unit" in addresses to use it for the geopandas library to get latitude and longitude for each clinic.
Then I calculated latitude and longitude data with competitors' coordinates to get the average distance to other dental clinics. Moreover, some dental clinic data from the web was incorrect, and their distance was abnormally higher than other data, so I made a boxplot to see the outliers and removed them for better results.

Now, I have X variables ready, and it's time to label the data since I will use supervised machine learning classifiers to predict the outcome. After talking to my uncle who had run a dental clinic before, I found that there is no way to find out their annual profit for the year since it is a private business but labeling them based on reviews would work because it shows how many loyal patients that the clinic has and enables how many patients are satisfied with the practice. The threshold for label 1 data was that the rating should be more than 3.7, and the review count should be more than 10. I built a classifier that labeled the data, and my data was ready to be explored.

The first step that I took for data exploration was to build a correlation heatmap and see which variables had the highest correlation with the label. The heatmap showed that average distance, median income, and the population had the highest correlation to the label, and I picked these three variables to be X variables to predict the labels for the population center dataset.
Secondly, I plotted bar plots of averages of label 1 and label 0 data to see if there was a difference between these variables. It showed that the average median income, average distance, and average rating were higher on label 1 but lower on label 0, and the Population variable had the opposite result. Lastly, I made a scatter plot with average distance and median income variables grouped by label 1 and 0 data points. The result showed that label 0 data is more centered toward lower average distance and median income, but label 1 data was more centered toward higher median income and average distance.

I was able to find some relationships between the variables and the labels then I went to see which machine-learning classifier worked the best. Firstly, I started by using logistic regression and set the testing data to 30 percent of the whole dataset. After training the data, the logistic regression model showed an accuracy of 0.81. The ROC/AUC curve also gave me good results, with the curve being top left corner and the area below 0.71.
Then I tested the SVM model with linear, polynomial, and Gaussian kernels. The accuracies of these three models were 0.68, 0.66, and 0.64, which were lower than the logistic regression model. Lastly, I tested the Naïve Bayes model, which gave me an accuracy of 0.74, which was also lower than the logistic regression model. Therefore, I decided to use a logistic regression model to predict the labels based on population center data.
X variables were set with population, average distance, and median income, and after putting these population center variables in the logistic regression model, location with index 0, 1, 4 got prediction of label 1. So the best location to open in the Waltham district was 441 Lincoln Street, Waltham, MA 02451, 320 Prospect Hill Road, Waltham, MA 02451, and College Drive, Waltham, MA 02452.
After getting the result, I made an interactive map with the folium library and showed how competitor dental clinics are distributed in population centers and locations that were predicted in label 1.