# Ranking homes to maximize chances of long-term satisfaction

## Introduction

Many factors influence the decision to purchase a home. Home buyers typically focus on features of the home – square footage, number of bedrooms and bathrooms, and price. However, there are many other factors that go into determining whether a house can feel like home. For example, proximity to family, favorite shopping centers, and entertainment opportunities all influence satisfaction with one’s home.

My family will be looking to purchase a home soon, and I wanted to create a program that would take into consideration some of these factors to help rank houses so that my family can select the home most likely to provide long term satisfaction. I will be ranking homes based on four criteria:

1. Purchasing power – defined as the number of square feet per dollar spent and calculated as the square footage of the home divided by the list price.
2. Proximity to our favorite store – we are die-hard Costco enthusiasts, so we would like to live close to a Costco store if possible.
3. Proximity to family – we want to live close to our family (parents and siblings).
4. Entertainment – A proxy for entertainment potential will be created by determining the number of unique categories of nearby venues for each location, as this represents a range of dining and entertainment options.

As it is currently coded, this program is likely only of interest to myself, due to the personal value placed in proximity to Costco. However, the program could be generalized by removing this parameter and focusing on the other three, or by substituting a user’s favorite store in place of Costco.

## Data

As mentioned in the previous section, the rank of the home will be determined by four criteria. Below, we examine the data sources to make the ranking model possible.

1. Purchasing power – in order to calculate this variable, we will need data on what homes are for sale, as well as the sale price and square footage of the home. This information could be gathered from a residential realty aggregator like Zillow.com or Realtor.com. In this project, we will use Trulia.com.
2. Proximity to our favorite store – we will use the Foursquare API to find Costco locations nearby to the homes for sale, then calculate the distance to each of the homes.
3. Proximity to family – there are a variety of ways that this data could be obtained. We could supply addresses of family members to a geocoder through the geopy library and obtain latitude and longitude. Since there are only three points of interest, we could also go to google maps, right click the map where their address is located, and read the latitude and longitude. In this project, to preserve my family’s privacy, I have chosen arbitrary addresses in the city that they live, rather than use their home address.
4. Entertainment – we will use the Foursquare API to gather information on venues near the locations of each of the homes, then calculate the number of unique categories of venues for each address. This should give a good idea of variety of activity and dining opportunities available in an area.

## Methodology

This project required little in the way of exploratory data analysis, with most of the coding challenges related to scraping the initial data set and cleaning the data for later use. To that end, looking at the shape, data types, head, and tail of various dataframes were used to verify that data had been added, dropped, and modified as intended.

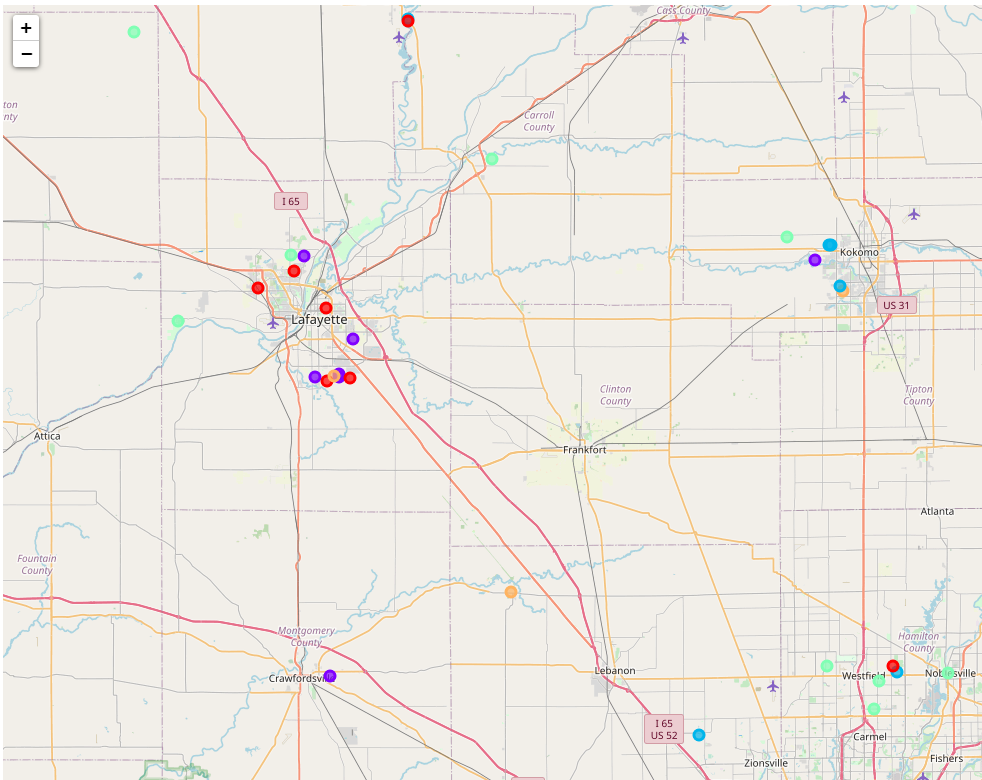
Once the dataframe had been cleaned as desired, variables of list price, square footage, number of bedrooms, number of bathrooms, latitude, longitude, and the ranking variables purchasing power, proximity to family, proximity to Costco, and venue diversity were scaled to allow for Kmeans clustering of homes. This clustering was performed so that if a house was taken off the market or a purchase offer was rejected, one could easily select another home from the same cluster, which should share some common features. In this case, it appears that the clusters were generated exclusively based on list price, with none of the other features showing consistent trends between groups.

## Results

The top ten homes are shown in Figure 1 while Figure 2 shows all homes visualized on a map.



**Figure 1**. Address, list price, size of home (square feet) and overall score for ranking (scale of 0-100) for the top ten homes.



**Figure 2.** Map of houses in final data set. Different colored markers signify different clusters which appear to be separated based on list price.

## Discussion

Prior to completion of the project, I expected Lafayette to have the highest ranked homes due to proximity to my parents, high venue diversity, and affordable homes. The results show that the top three homes (and the only homes with a ranking of 90 or above) are all located in the Westfield/Carmel area. This region is close to my older brother, but farthest from my parents, which gives the region a middling family distance score. However, venue diversity is high, and distance to Costco is low, both contributing to a high overall rank. While the Westfield/Carmel area is in general more expensive than the Lafayette area, this was not enough to offset the contribution of the other three factors to overall high scores for homes in this area.

Homes in the Kokomo area and the Northern most homes had the lowest ranking due to a long distance from Costco, high average distance to family, and low venue diversity. Furthermore, the purchasing power was low for several of these homes, likely because this variable only looks at the home size divided by its list price and does not account for lot size. Homes in these more rural areas likely have larger lots, contributing to higher price, which would decrease apparent purchasing power compared to the small lot homes of more suburban and urban areas, as lot size is not factored into this score.

## Conclusion

Home ranking could be improved by including other variables in the calculation of purchase power, such as age of home and lot size. Additionally, other variables that are important to the purchaser could be included, such as school rankings, to provide a more robust assessment of overall quality of the home and surrounding area. However, for the purposes of this course, the code works well to obtain information on homes listed for sale that fit the criteria important to me and utilizes data from Foursquare, as required by the capstone project.