1. **Data Acquisition and Cleaning**

For this paper, the data utilized was gathered by two main sources, IBM and Foursquare. From IBM, it was possible to obtain price and information about the houses sold in King County, USA between May of 2014 and May of 2015. This data also included latitudinal and longitudinal coordinates of these houses, enabling us to access the Foursquare API in order to pinpoint those positions. To get the amount of venues, Foursquare API retrieve a list of venues in a 500m radius from the geographical coordinates in 2020, that will be used as approximation for venues in that time period.

After having these data, some changes were required. For the house sales, most of it was already organised in a file with 21 thousand rows with the amount of bedrooms, of bathrooms, of floors, of square foot in the lot, of square foot in the actual residence, latitudinal and longitudinal coordinates, if it is in front of water, an overall grade, the condition of the house, the year built, the year renovated, the zip code and the date of the transaction. The data was already organised, so only the column with a defective index was dropped.

Nevertheless, the Foursquare data required some refinement. After retrieving the coordinates for some rows and applying to the API, a new data frame was formed with each of the venues connected to a certain house ID. Therefore, it was needed to create a new data frame with the amount of times that each ID appeared, that is, the amount of venues designed to each ID and merge it to the IBM data frame, adding a column named Venues Count to a smaller set of data.