Applied Data Science Capstone Project - The Battle of Neighborhoods

Introduction/Business Problem

As I am from London UK it seemed logical to focus this project on the London area. In this project I will attempt to develop a model to predict the median property value for a neighborhood in Inner London based on factors such as distance to central London, number of schools, doctors offices, parks, transport links such as train stations. The purpose of this is to understand the factors which influence property value in London so that the effect on property values of adding or removing infrastructure or amenities can be estimated and the cost/benefit of a change can influence decisions about public or private sector spending. It will also be interesting to see how changes in working practises following the Covid-19 pandemic may influence house prices in London.

Data

First, I will determine how to divide London into neighborhoods so that there are enough that statistically significant conclusions can be drawn but yet not so many that analysis is impossible. I will also need to use definitions of neighborhoods for which data is available concerning median property values.

There are approximately 322,000 detailed postcodes in London which can be rolled up into 276 ‘outcodes’. There are 32 Boroughs in London of which 12 are considered ‘Inner London’. There are 627 Wards in London of which 221 are in ‘Inner London’.

The ONS (Office of National Statistics) website is a good source for a dataset which includes longitude and latitude for electoral wards in the UK so I will use this to identify the geographical coordinates for the wards in London. I will take the median property value data from data.london.gov.uk which is published by the office of the Mayor of London. The distance to central London will be based on co-ordinate data and the number of schools, doctor’s offices and other amenities will be obtained from Foursquare.