Data

The data used in this project will be sourced from Wikipedia, and government data collection websites. I have used data from a range of different sources so that I may piece together a picture of different factors affecting the viability of new property developments. The data sources can be found in the appendix, along with being referenced in the notebook, and consist of two types of data accumulation; csv tables and API requests. I have used the csv tables as the basis for my datasets, and the API requests to retrieve location data for the universities and buildings that I have used.

API

The API references are as follows:

- GeoPy was able to return latitude and longitude coordinates given the names of universities, which was essential to helping me to place the universities on the map in bulk. It was also used to calculate the distances between locations given their coordinates.
- Foursquare was used to return nearby student accommodation companies around each university.
- Finally, Postcodes.io was used to return the borough of each location to me in a format that would be reliably read each time it was run.

CSV

In terms of the csv files, the wiki tables were scraped of the websites using the very useful website 'https://wikitable2csv.ggor.de/'. The data from data.london.gov.uk was easily accessible in the csv format. The files were uploaded to the IBM database and accessed from there into my notebook. The csv file sources were used as follows:

- Wikipedia was used to obtain a list of universities within London. This was an essential first step as it gave me only the universities within the city I was studying. Wikipedia also gave me a list of student numbers for every university within the UK. This table was merged with the list of universities within London.
- The gov.uk website was used to give me housing price data for boroughs within the UK. This dataset was over 12000 rows initially, however after processing this dropped to around 40 rows. I selected the median property prices for London boroughs.
- LatLong was used to give me an accurate value for the centre of London from which I could calculate the distance of universities.

Of the universities given to me by Wikipedia, about 4 of them did not return a value for latitude or longitude by the Foursquare API. I made the decision to remove these universities from my data as they did not represent a large proportion of my data, nor would their absence impact my results enough to bring a different outcome.