Background

As we enter the new decade, major infrastructure projects are under way in the UK. HS2 (High Speed 2) rail and HS3 are aiming to provide greater capacity and faster connections between the UK's capital city London, other major cities such as Birmingham and Manchester, regional cities e.g. Leeds, Liverpool, Sheffield etc. A potential connection to Edinburgh and Glasgow at a future date has been mooted. There is evidence this will provide significant investment and gains in productivity to both the regional economies and London, albeit in two to three decades.

The regional economies are naturally rebalancing as the industrial heartlands of the Midlands and Northern England reinvent themselves with higher value industrial (automotive, aerospace pharmaceuticals) and services (media, game development, legal and financial services). There has been a significant increase in businesses 'north-shoring' or moving HQ's to other major cities for cost reasons.

On the other hand, London is facing issues of its own with costs outpacing income. Many of the city's residents have begun to question the feasibility of locating there, as have many economists. Despite higher median wages, net tax, the cost of living combined with higher house prices has eroded the disposable income and quality of life of Londoners*.

There is evidence that such costs and centralisation have crowded out innovation both in London and in the regions**. London as the capital is the largest economy, but it need not be the only location for higher value-added employment***

Rebalancing could lead to better outcomes for both the regions and London with greater investment and opportunities in the regions and a reduction of cost pressures that would benefit London's innovative service-oriented economy.

This has increased interest in analyses of the costs of living and business as well as the comparative attractiveness of neighbourhoods in London and alternative cities to private individuals looking to relocate in both directions.

The objective of this project is to investigate and compare median incomes, house prices and neighbourhoods for two metropolitan areas (London and Manchester) in the UK.

Only neighbourhoods at or below the London median income in the data (£31,500) will be considered.

Core Problem

Whilst government statisticians, decision makers and business have the means to use and understand data to draw inference most private individuals do not. Therefore, this project will seek to collate the required information, analyse, and present the data is a way that is simple to understand. It will also serve as base from which to draw further detailed analysis into the components.

Ultimately the objective of this project is about enabling private individuals to assess the attractiveness of two cities given their personal circumstances utilising a data centric approach. This is not about which one is 'better', appealing to emotions or regional loyalties.

Interest

The government, businesses and individuals alike require insights on how best to proceed with investment and employment decisions as both the regional economies and London transform with their various pros and cons. When deciding on a location private individual need to take account earning potential, taxes, housing costs, general cost of living and the surrounding neighbourhood.

Data Requirements

The two metropolitan areas that will be covered in this analysis will be Greater London and Greater Manchester, but this could equally apply to Birmingham, Leeds, Sheffield etc. The data for the neighbourhoods will be scraped of their respective Wikipedia pages:

https://en.wikipedia.org/wiki/List_of_areas_of_London and https://en.wikipedia.org/wiki/List_of_places_in_Greater_Manchester. GeoPy will be utilised to gather the Latitude and Longitude of the respective neighborhoods.

ForeSquare will be used to retrieve information on the types of venues present in each neighbourhood. A K-Means cluster algorithm from scikit will determine the clusters based on the frequency of types of venues in each neighbourhood. The clusters will be presented as basis of comparison to viewers.

We will use data from the Office of National Statistics (ONS) for data on median earnings and house prices whilst we will use websites such as https://listentotaxman.com/ and https://www.rightmove.co.uk/ will be used to calculate net tax income and housing costs based on the data provided by the ONS. Housing costs will be calculated as the mortgage cost of the median house price with a 20% down payment.

There will subsequently be manual calculations and adjustments to errors in data where it is more efficient to exercise in Excel. In this instance data frames will be saved, downloaded, edited and reuploaded. Such manual calculations and adjustment include coordinates, and formulas for calculating net tax and disposable income.

The data will be output in the form of a map in folium showing the geolocation of the neighbourhoods in Manchester and London as well a list of their respective cluster features. Additionally, bar charts displaying median, net tax and disposable income as well as a choropleth of house prices will be displayed to allow the viewer to quickly identify and compare neighbourhoods.

Extended Background

Reinvention is a chicken and egg problem****, geographical advantage (position relative to Europe) and three hundred years of centralisation has concentrated the majority of global businesses and national companies in London and the south east in a manner similar to France and South Korea but in opposition to the regionalism in Germany, Spain, Italy and the USA. Such reinvention will require government, businesses, and people to invest, upskill, relocate, and understand current and future opportunities and problems.

Improving communications technology has provided a potential opportunity for labour mobility on a virtual foundation by removing location as a primary factor. Unfortunately, this potential had not been explored prior to COVID-19.

*Despite a large concentration of global business, high income jobs and diversity London also has high levels of socioeconomic inequality and it is questionable whether for a given quartile Londoners are better off all things equal.

**This has raised the question of whether the UK, even London, may be more productive when regions are on a more even base. The theory is that increasing productivity in the regional economies does not have to come at the expense of London but can add to it; it may however require some priority initial investment likely paid for by London taxpayers.

***Bavaria, Germany has a larger economy, population, and pool of businesses than Hesse but the per capita (productivity and thus wages) is the same, due to a similar distribution of educated individuals and top quartile businesses by productivity. This suggests that there is nothing inherently un/productive about a location, with some exceptions, or an agglomeration of people or businesses per se. Instead it suggests productivity can be improved through education and capital investment in the development of top quartile businesses by productivity. In the United Kingdom the regions with the highest GVA growth per year perhaps unsurprisingly have the highest percentage of university graduates and subsequent capital investment.

****The origins of Silicon Valley owe to fortuitous events. William Shockley's family connections to California, his work at Bell Labs on radar and a subsequent requirement to replace fragile vacuum technology and the opportunities for the field present in California (after massive government investment due only to its location vis a visa Japan and subsequent future outlays for Korea, Vietnam etc.)

Methodology

Beautiful Soup scrape and data clean-up

After scraping the respective webpages for the neighbourhoods in question both sets of data require cleaning.

In the case of the data for Greater Manchester the Neighbourhoods are concatenated by comma values in a column known as 'Other components'. This requires a function to split all cell values separated by a comma into new rows all other data being equal. Then Columns 1 and 3 are dropped and any cells with an empty value are replaced with 'Not Assigned'. Lastly data cells in the column 'Other components' that have a value of 'Not Assigned' are replaced with the value in the 'Metropolitan borough' column. A new column 'address' is added concatenating the 'Other Components' and 'Metropolitan county'.

In the case of the Greater London data the neighbourhoods are already distributed by row. Instead the data requires the clean-up of footnotes and the dropping of the last four columns. Again, a new column is created by concatenating 'Location' with a manually inserted string 'Greater London'.

Using GeoPy to obtain altitude and longitude information.

GeoPy is imported and a function written to use GeoPy to generate the required geographical coordinates for each data set in a column called 'point'. Another useful column 'location' is created that we can use to cross reference erroneous data. The 'point' data which is a concatenation of the geodata required needs to be split into three new columns 'lat', 'long', 'dec'.

Manual editing of the files in Excel

At this stage, pandas has played its part and the remaining items are faster to execute in Excel. The data frames are saved and download to local storage and manually edited as follows:

- Erroneous latitude and longitude corrected. (If the location does not correspond to what is expected). A 2% error rate was identified (15 rows in total)
- Drop all columns except Metropolitan county, Metropolitan borough, Location (renamed as Neighborhood), lat (renamed as Latitude) and lon (renamed as Longitude) for each dataset. Neighborhood is intentionally misspelt in line with our transatlantic cousins for ease of coding.

The files are then reuploaded into pandas to serve as a basis for our K-Means clustering.

Exploratory analysis with ONS, listentotaxman and Rightmove data

Using the ONS data, as outlined in data requirements, along with estimated taxation and mortgage costs the net tax earnings, house price ratio and net income can be calculated as follows:

Field	Calculation
Net Tax Income	Median Earnings minus tax (income tax and national insurance)
Mortgage (80%) Dual Income	Mortgage cost with a 20% down payment based on the median house price and assuming dual income Median House Price minus 20% to Rightmove Mortgage Calculator. Divide by two (dual income, split cost)
Net Income	Net Tax Income minus mortgage cost
House Price Ratio	Median House Price/Median Earnings

After calculation, the new fields will be added into a new data frame with the following columns ‘Metropolitan county’, ‘Metropolitan borough’, ‘Code’, ‘Median Earnings’, ‘Median House Price’, ‘Ratio’, ‘Net Tax’, ‘Mortgage (80%) Dual Income’, ‘Net Income’. ‘Code’ is a geographical ID for use in the Choropleth map that will provide mapping to the JSON data.

The data will be uploaded as a data frame into pandas to serve as a basis for our folium maps and seaborn bar charts that will provide visual comparison between the neighbourhoods and the metropolitan county.

Using ForeSquare to pull nearby venues.

Each dataset will be run against ForeSquare API to generate information on venues.

After connecting to ForeSquare a function is written to pull venues close to the given neighbourhood into a new data frame. The number of venues for each neighbourhood are counted before one hot encoding is used to reshape the data frame with each neighbourhood a separate row and the venues separate columns. Finally, the mean occurrence of each venue is calculated.

The above mean values are used to write a function that returns the most common venues in descending order. A new data frame is composed with each neighbourhood again a new row but with only ten columns representing the first to tenth most commons venue for each neighbourhood.

Merging of datasets

After each dataset is completed it is time to merge the data frames into one single data frame. The data frames we created to represent each neighbourhood and the ten most common venues for Greater Manchester and Greater London will be merged; this will be our main data frame. Additionally, the data frame composed of neighbourhoods as rows and the venues as columns with the mean occurrence of each venue will be merged and will need the extra step of cleaning. Not every type of venue exists in each neighbourhood, so the NaN values need to be converted to 0. This data frame will be used to generate the cluster.

Lastly the two original data frames we used in ForeSquare will be merged as these have important information such as the ‘Metropolitan county’, ‘Metropolitan borough’, ‘Latitude’ and ‘Longitude’. The new data frame will subsequently be merged with the first merged data frame we created to generate a new data frame with the important information outlined above and the top ten most common venues.

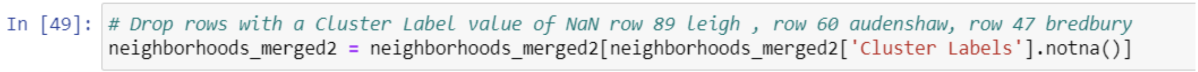
This will give us two final new data frames.

K-Means Clustering

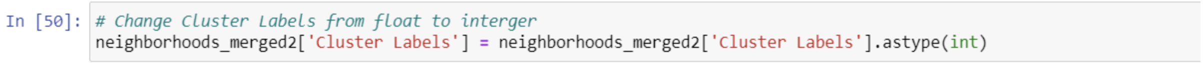
The required library is imported from scikit. Using the data frame with the mean occurrence of venues a K-Means clustering algorithm is applied to the neighbourhoods. This will generate five clusters as written based on the mean occurrence of venues in each neighbourhood. The function written will create a new column with a mapping of the cluster (a label).

After K-Means clustering has been performed we can merge the column created (cluster label) with the data frame we created at the end of the last stage. Now we should have a new data frame with information on the neighbourhood (e.g. 'Metropolitan country', 'Latitude'), the top ten most common venues and now the respective cluster it belongs to.

Finally, before we start plotting data, we need to do a sanitary check on the data. Any rows with a cluster value of NaN will be dropped and all cluster values will be converted to integer (float type does not work nicely with folium).



```
In [49]: # Drop rows with a Cluster Label value of NaN row 89 leigh , row 60 audenshaw, row 47 bredbury
         neighborhoods_merged2 = neighborhoods_merged2[neighborhoods_merged2['Cluster Labels'].notna()]
```



```
In [50]: # Change Cluster Labels from float to interger
         neighborhoods_merged2['Cluster Labels'] = neighborhoods_merged2['Cluster Labels'].astype(int)
```

Data Visualisation - Folium Map, Choropleth, Seaborn

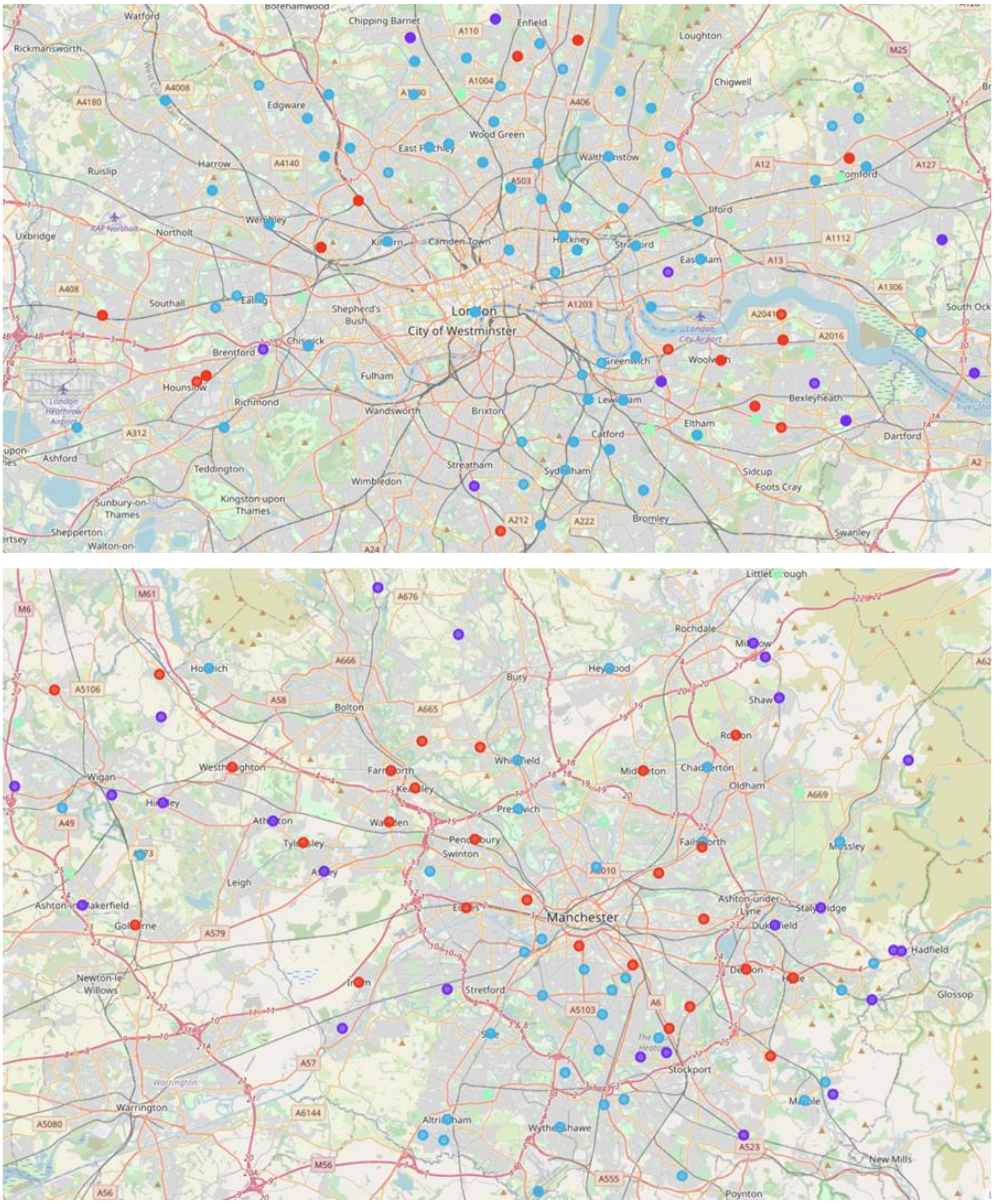
After importing the required json, matplotlib, seaborn and folium libraries we can begin to visualise the data. The design of the visualisations is critical to conveying the story and, in this case, guiding the audience. In the context of the project the visual data must be easy to interpret and draw inference from the government statisticians to the layman.

The first visualisations are folium maps centred on Greater London and Grater Manchester, for comparative purposes, of the neighbourhood clusters. Straight away the colour coded clusters make it easy to identify similar neighbourhoods. A table for each respective cluster will be generated for further analysis.

Using the final data frame, we uploaded initially, a Choropleth map is used to display the median house prices of their respective Metropolitan boroughs in Greater London and Greater Manchester. This data frame is split at a higher level – metropolitan borough and includes the data on median earnings, housing prices, net tax etc. The more expensive areas are immediately obvious with their darker red hue. This can be compared straight away with the neighbourhoods in the previous folium map and the seaborn charts.

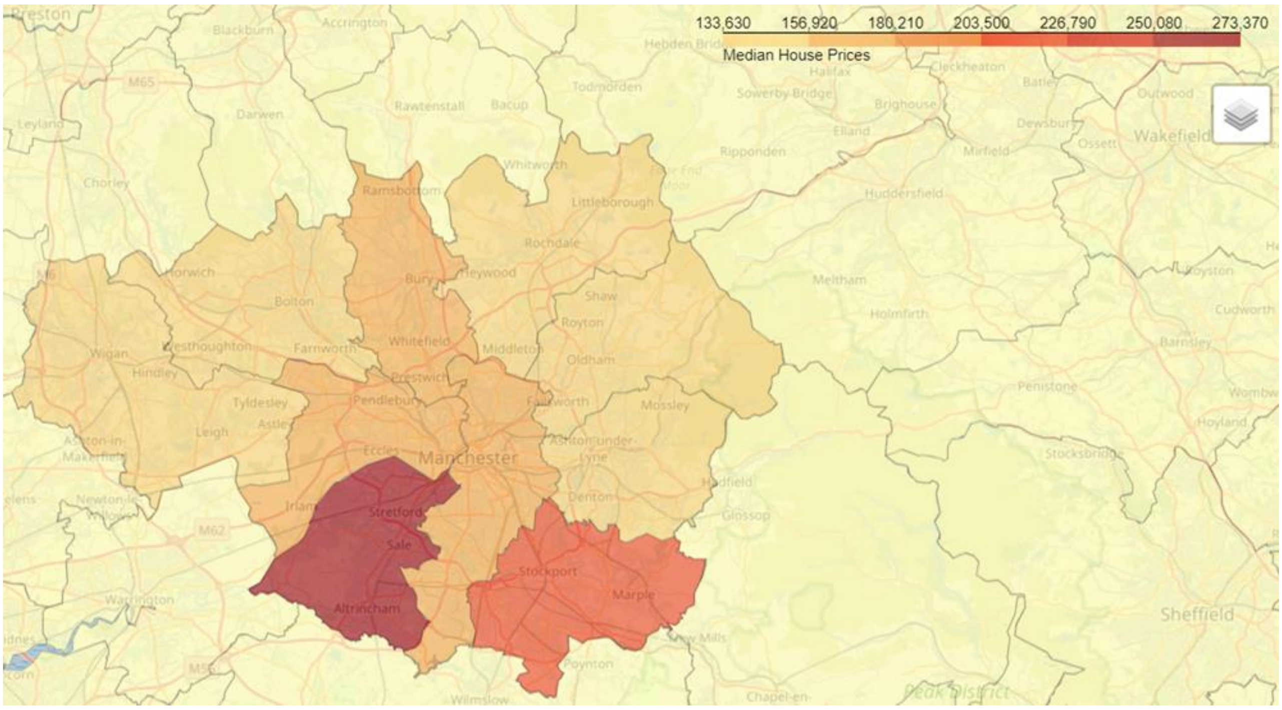
Finally, several Seaborn bar charts are generated are generated based on median earnings, the house price to median earnings ratio, net tax earnings, estimated mortgage costs and net income. The charts will be colour coded by Metropolitan county e.g. Greater Manchester and Greater London as the primary objective of this analysis is to compare the neighbourhoods and their higher-level boroughs for the respective county.

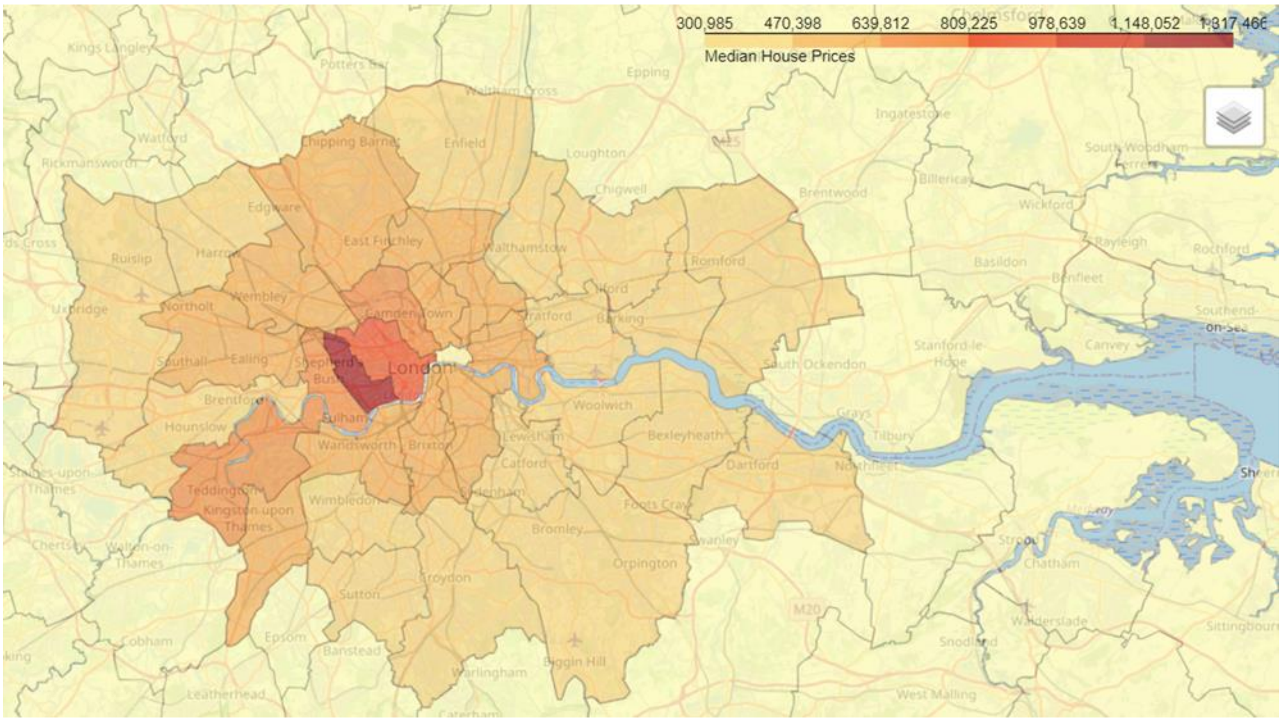
Results



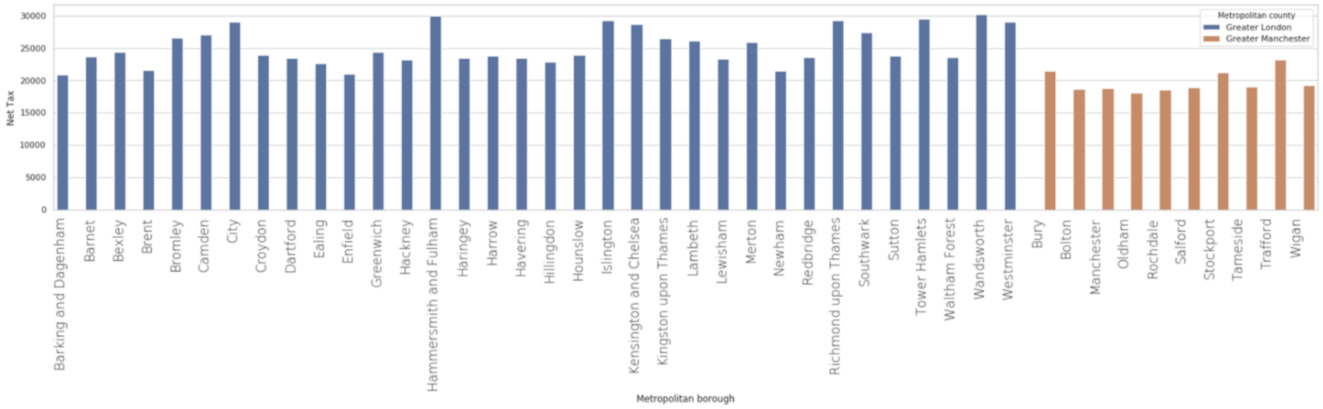
(The colour of the clusters have been changed from the Python workbook for viewing purposes)

One thing that is immediately obvious when looking at the folium map displaying the clustered neighbourhoods is that both Greater London and Greater Manchester have a diversity of neighbourhoods according to the clusters. Interestingly Greater Manchester has the greater diversity of clusters, which may come as a surprise. Whether this is a ‘good’ thing is debatable. Furthermore, each the placement of the neighbourhood has some visual correlation with median house prices.

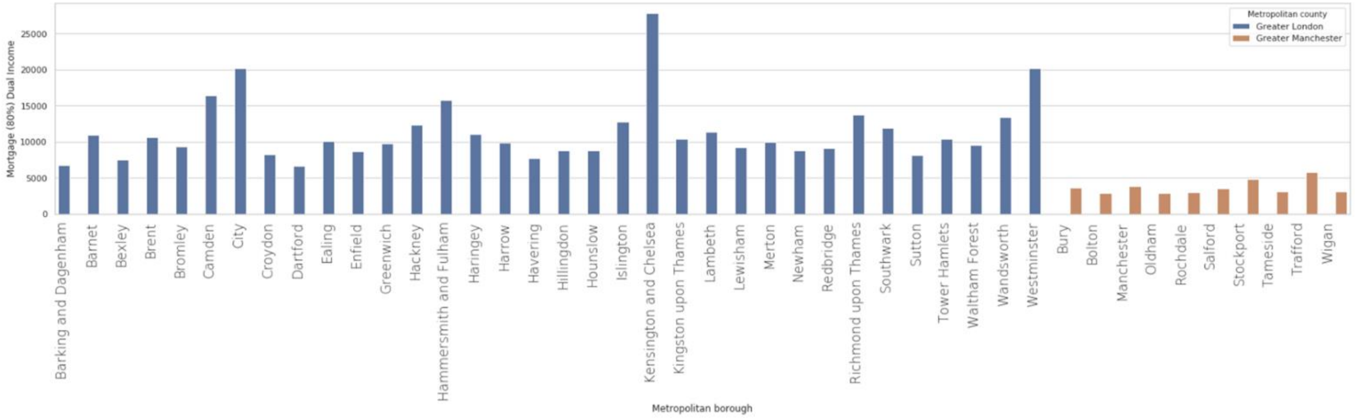


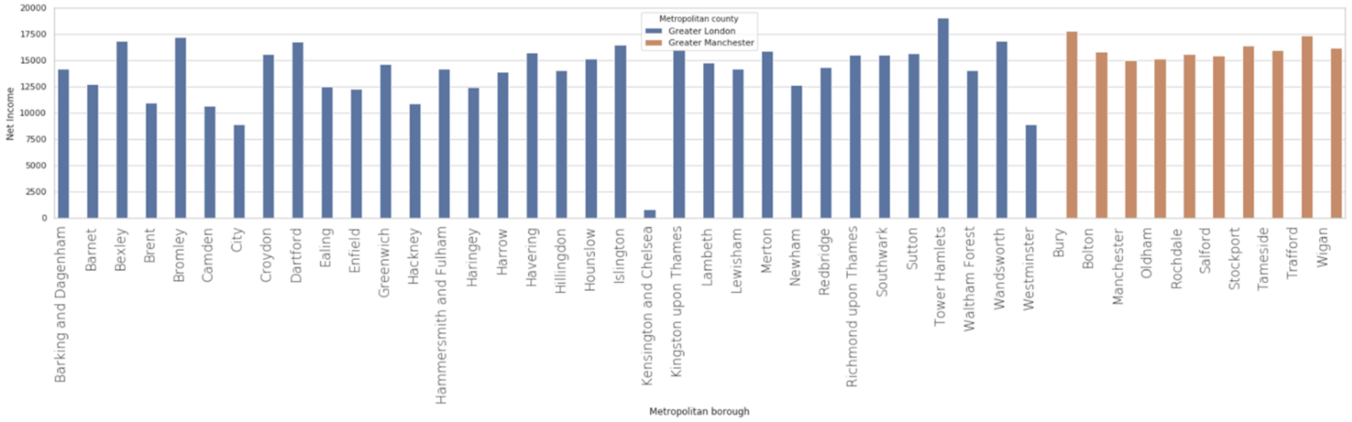


When we look at the folium map of median house prices it is obvious that both Greater London and Greater Manchester have a range of median house prices according to the respective metropolitan boroughs and that there is generally a concentration. In Greater London, this concentration is around the centre-west and inner south west. In Greater Manchester, this concentration is in the centre-south and south. Additionally, the spread of price both as an absolute and percentage is greater in Greater London.



The most interesting and surprising results come from the Seaborn charts. Greater London boroughs enjoy higher median earnings, even after tax, than their counterparts in Greater Manchester. Despite this there are some metropolitan boroughs of Greater London comparable to Greater Manchester, even net tax.





Moreover, the extremity of the median house prices compared to median earnings (the ratio) significantly degrades the net income after estimated mortgage/housing costs. The estimated monthly mortgage payment, on a dual income basis, is at least double in Greater London compared to Greater Manchester. The net tax median earnings cannot make up for this and thus the net income, all things equal, for most metropolitan boroughs in Greater Manchester is at least comparable to Greater London.

Discussion

As with any analysis the results only show what is captured by data and greater context is needed when discussing that they show. Whilst the results may bust the myth of a ‘grim, northern’ Greater Manchester and a ‘prosperous, southern’ Greater London there are several other factors to consider.

Firstly, whilst the neighbourhoods are clustered on their similarity and thus the neighbourhood clusters are diverse compared to each other this does not indicate that the neighbourhoods themselves are diverse. To demonstrate this logic one cluster of neighbourhoods could have a diversity of venues e.g. ethnic restaurants, gardens, museums, whereas another cluster may have a high frequency of homogenous venues e.g. pub or only Italian restaurants and another cluster may have a high frequency of other homogenous venues e.g. Supermarkets and fish and chip shops.

Secondly there are a wealth of statistics including but not limited to crime, employment (industry), education, ethnicity, innovation, creative arts etc. that could influence how neighbourhoods are clustered.

Lastly, whilst it is indisputable that the median earnings in Greater London are insufficient to cover higher house prices this does not tell the whole picture. The ONS data on median earnings also contained data for deciles and quartiles. One must consider that the upper two deciles of earnings is significantly higher in Greater London than Manchester. This may indicate that people are willing to endure temporary pain for potentially greater economic opportunities later*. Greater London hosts the majority of global and domestic business HQ’s.

*The the sacrifices made by individuals living in Greater London were noted in the ONS data with a significantly higher rate of leasehold apartments and flats (less living space) and larger household sizes. It was also noted that the population is generally younger, more educated with more full-time employment and higher instances of dual income. Lastly there was a noted tendency for emigration in the 30-40 age group, likely due to the affordability of a family sized housing.

Conclusion

Notwithstanding these considerations what is interesting about the data is that is appears to confirm two theories that have increasingly been quoted as anecdotal evidence.

Firstly, that neighbourhood diversity exists in minor and major cities as well as those cities defined as megacities, although the exact number and diversity of neighbourhood may differ, Greater London has a higher number of diverse neighbourhoods. Nevertheless, we can see from the results that there is sufficient diversity within the neighbourhoods of Greater London and Greater Manchester that such that migration should not cause loss of access to diverse venues.

Secondly, the commonly held belief that living standards are higher as a result of higher median earnings is at least if not questionable demonstrably false on a like or like basis. The higher median earnings in Greater London are insufficient to cover higher tax and median house prices. The results indicate that there is at least rough parity, all things equal, between most boroughs in Greater Manchester and Greater London in terms of net income. It is likely the bottom half of earners in Greater London are significantly worse off than those in Greater Manchester but the upper two deciles are likely better off.

The human tendency towards survivorship bias has perhaps distorted the full picture of earnings particularly those at the median and lower deciles. The reality of a pyramid shaped earnings profile may come as a shock to graduates, young families and even businesses and government officials.

Further avenues of research may include:

- Whether broadly similar neighbourhoods are feature across other countries' major cities and whether there are identifiably unique variables in global cities.

- Given that disposable income does not consider the size of the dwelling (only the amount of earnings spent on rent/mortgage) how can this be normalised for comparison?

- Is there a pattern in cities with higher median earnings where the top two deciles are significantly richer and whether this has negative connotations on the bottom 50%? New York, Los Angeles, Chicago, Paris, Tokyo etc. would make good studies.

- Do perceptions of higher economic opportunity influence migration between cities, how and whether these are factually based?

- How much economic mobility exists in each decile of earnings and the statistical mobility particularly in the context of migration to larger urban areas for this purpose?

- To what extent is GDP and per capita data is distorted by such factors as the reporting location of the HQ, percentage of people in part time work, retirees etc?

- To what extent does is productivity inherent to location. What agglomeration economies e.g. population size, number of business, education, capital investment cause productivity improvements and what is their respective weightage? Are there other discrete or not logically consistent factors e.g. social, political?

- Chicken and Egg. How can government influence productivity improvements?

- How advances in communication technology could affect location as factor in agglomeration economics?

- How much does productivity vary across cities and in what sectors. What activities are carried out more efficiently?