North-East England Housing Sales Prices and Venues Data Analysis

Fern Vass

June 2020

1 Introduction

1.1 North-East England Background

North East England is one of nine official regions of England at the first level of NUTS. It is known as being the cheapest region in England based on average house prices taken in January 2020 with a monthly price drop of 2.6% since December 2019.

The most populous city in the North East is Newcastle, which is the eighth most populous urban area in the UK. It hosts many major corporate headquarters and has strengths in learning, digital technology, retail, tourism and cultural centres. Investment property comprises a large proportion of terraced housing. Towards the city centre, there are many large, impressive Georgian structures however, large family homes are available towards the outer areas and many new-build developments have emerged around the city. Newcastle beings in many students as it houses Newcastle University and Northumbria University, so there many many student lets especially towards the centre of the city.

From the perspective of a property investor/landlord, we will want to be able to able to clearly visualise the average house sale prices in different areas so we can narrow down the areas we are looking to buy into. We want to find the ideal place to purchase our next home and have a rough idea of the social/cultural environment we are buying into. As we are looking at homes for long-term tenants, we will be looking at Newcastle, Durham, Sunderland, South Tyneside and North Tyneside.

1.2 Data

1. We will be looking at the Lower Layer Super Output Areas (LSOA) of the UK and assign the median house prices to these areas. We can download this data (last updated September 2019) from the Office for National Statistics[4]. This

will be cleaned so we gather data for Newcastle Upon Tyne, County Durham, Sunderland, South Tyneside and North Tyneside.

2. To be able to visualise this on a map, we will need the boundaries of each LSOA area which we can pull from a GeoJSON online project[3]. These are given as TopoJSON files so we will convert them to GeoJSON using an online converter and merging them using an online GeoJSON merger. We can use this to add a Choropleth map.

3. To be able to cluster the data, we will add markers onto the map at the centroid location of each area. We can gather the centroid coordinates from another CSV file provided by the Office for National Statistics[5]. Again, we will clean this CSV file to give us our North East areas.

4. Using these centroid locations for each area, we can pull up the 10 most common venues around these areas use FourSquare API[2] and we can use k-means clustering to cluster this data by venues.

2 Methodology

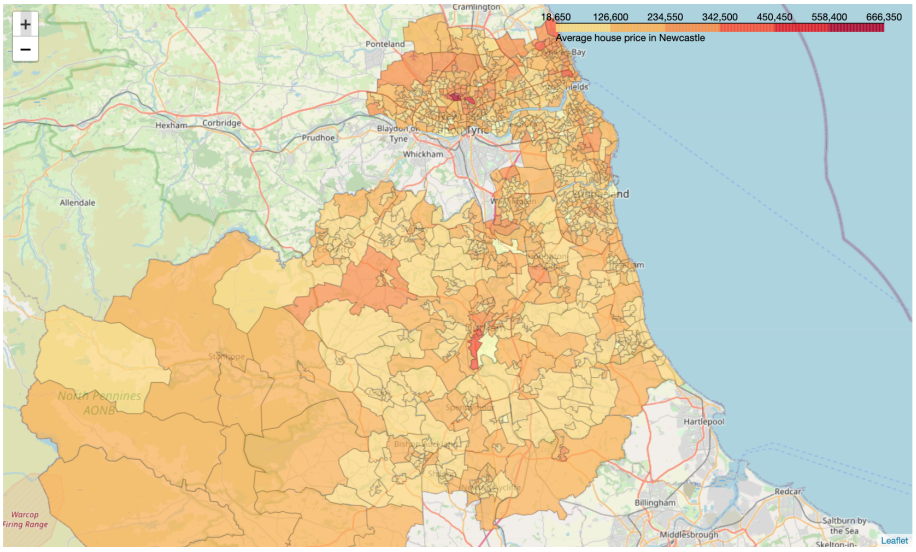
2.1 Cleaning the Data

We will be looking at the Lower Layer Super Output Areas (LSOA) of the UK and assign the median house prices to these areas. We can download this data last updated September 2019 from the Office for National Statistics[4]. After gathering relevant data for Newcastle Upon Tyne, County Durham, Sunderland, South Tyneside and North Tyneside[1]. We will clean it getting rid of any null values and converting the string prices to integers. The first 5 rows of our dataframe are now displayed below.

	LSOAcode	LSOAname	houseprice	houseprice1
0	E01020591	County Durham 015B	168000	168,000
1	E01020592	County Durham 011A	87000	87,000
2	E01020593	County Durham 011B	84000	84,000
3	E01020594	County Durham 013A	88000	88,000
4	E01020595	County Durham 013B	196250	196,250

2.2 Adding Choropleth map based on house prices

To be able to visualise this on a map, we will need the boundaries of each LSOA area which we can pull from this UK-GeoJSON project[3]. These are given as TopoJSON files so after converting to GeoJSON and merging all the boundaries, we have all the boundaries of all the LSOA areas we want[6]. We can now plot the Choropleth data onto the map.



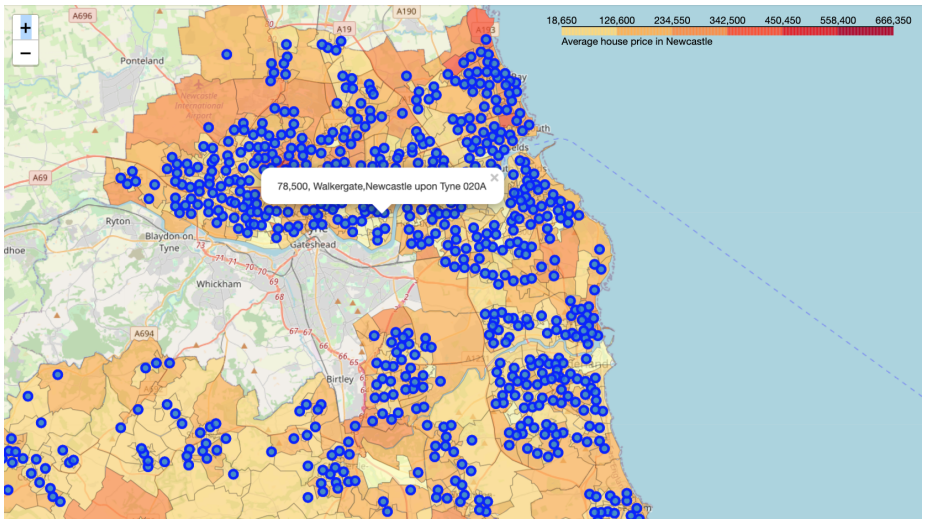
2.3 Adding centroid markers onto map

We can obtain the latitude and longitude for each lower layer super output area centroid using data from the Office of National Statistics[5].

We download the CSV file for the UK and clean it again so we get our specific areas[7]. Our data gives use the X,Y Northing-Easting coordinates so we'll convert them into latitude and longitude values for each centroid to our dataframe. WD15NM is our ward name which we will add to our label. Below are the first 5 rows of our new dataframe.

	LSOAcode	LSOAname	houseprice	Latitude	Longitude	WD15NM
0	E01020591	County Durham 015B	168000	54.854128	-1.514874	Lumley
1	E01020592	County Durham 011A	87000	54.856330	-1.583208	Chester-le-Street West Central
2	E01020593	County Durham 011B	84000	54.854644	-1.576591	Chester-le-Street West Central
3	E01020594	County Durham 013A	88000	54.849857	-1.572715	Chester-le-Street East
4	E01020595	County Durham 013B	196250	54.850199	-1.566768	Chester-le-Street East

Now we want to plot the centroid onto our map and have the labels display house price, ward and LASO name.



2.4 Pulling up the top 10 Venues for each area

Using the FourSquare API, we can use the latitude and longitude for each area and get the nearby venues around them using the limit of 100 venues and radius of 500m for each point. Now for each area name (ward), we can list the top 10 most common venues, as shown below.

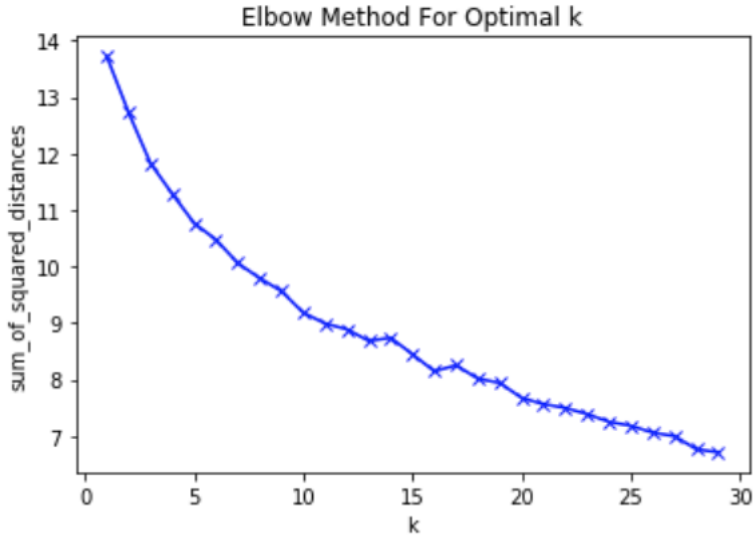
	Area_Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Annfield Plain	IT Services	Supermarket	Park	Construction & Landscaping	Soccer Field	Cupcake Shop	Cricket Ground	Laser Tag	Funeral Home	Chinese Restaurant
1	Aycliffe East	Supermarket	Soccer Field	Cricket Ground	Convenience Store	Warehouse Store	Gym / Fitness Center	Fast Food Restaurant	Grocery Store	Bakery	Bus Stop
2	Aycliffe North and Middridge	Bus Stop	Bar	Supermarket	Food	Pharmacy	Pizza Place	Food & Drink Shop	Indian Restaurant	Pub	Construction & Landscaping
3	Aycliffe West	Pub	Bar	Bus Stop	Convenience Store	Playground	Pizza Place	Photography Studio	Pharmacy	Fast Food Restaurant	Supermarket
4	Barnard Castle East	Café	Gym / Fitness Center	Pub	Supermarket	Museum	Coffee Shop	Zoo	Food Service	Flower Shop	Food

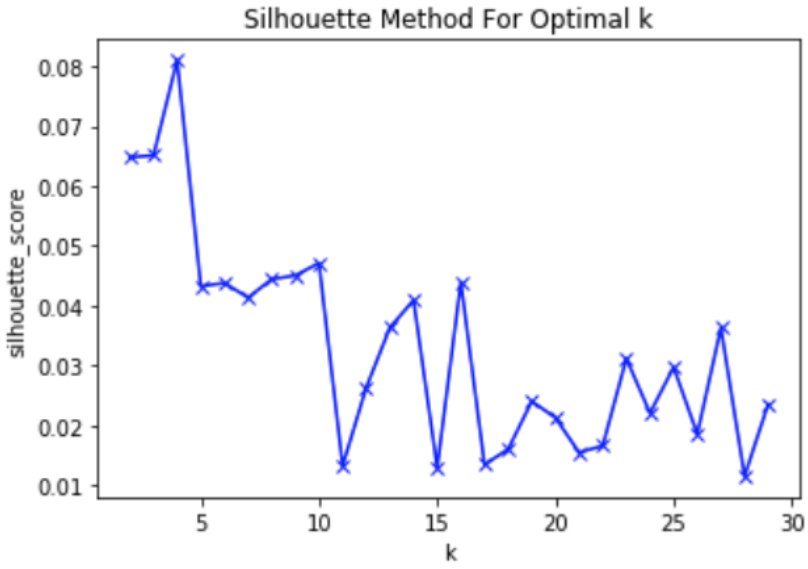
2.5 K-Means Clustering

K-means clustering is an unsupervised method which we will use to cluster the wards. Firstly, we want to find our optimum value of k, which is the number of clusters we should have to give us the best grouping.

Below we can see using the elbow method to determine the optimal k is not giving us a very clear 'knee' or bend. So we will look to perhaps using other methods to find the optimal number of clusters such as the Silhouette method.

The Silhouette method figure below suggests 4 clusters maximises the average silhouette values, 10 clusters coming in as second optimal number of clusters. So using this we will use K-means algorithm with k = 4.





3 Results

3.1 Analysing clusters

Now that we have our optimal k, we can cluster the areas easily. Below we can see our merged table containing the area names, their corresponding cluster

labels and their top 10 most common venues.

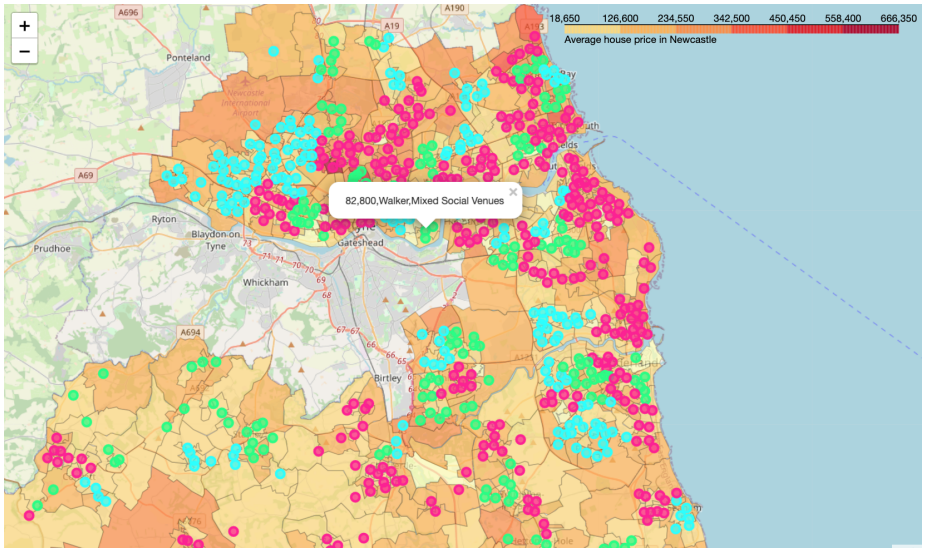
	LSOAcode	LSOAname	houseprice	Latitude	Longitude	WD15NM	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	E01020591	County Durham 015B	168,000	54.854128	-1.514874	Lumley	1	Pub	Grocery Store	Bus Stop	Cosmetics Shop	Electronics Store	Gaming Cafe	Furniture / Home Store
1	E01020592	County Durham 011A	87,000	54.856330	-1.583208	Chester-le-Street West Central	1	Pub	Coffee Shop	Supermarket	Discount Store	Food & Drink Shop	Doctor's Office	Indian Restaurant
2	E01020593	County Durham 011B	84,000	54.854644	-1.576591	Chester-le-Street West Central	1	Pub	Coffee Shop	Supermarket	Discount Store	Food & Drink Shop	Doctor's Office	Indian Restaurant
3	E01020594	County Durham 013A	88,000	54.849857	-1.572715	Chester-le-Street East	1	Pub	Playground	Bar	Entertainment Service	Gym	Auto Workshop	Locksmith
4	E01020595	County Durham 013B	196,250	54.850199	-1.566768	Chester-le-Street East	1	Pub	Playground	Bar	Entertainment Service	Gym	Auto Workshop	Locksmith

Now, we want to analyse the venues in the different clusters and assign a clear label that can tell us information about the type of area of each ward.

3.2 Assigning Cluster Labels

Analysing above, we may label the clusters as such:

1. Cluster 0: High-street areas
2. Cluster 1: Pub, Bars Restaurants
3. Cluster 2: Mixed Social Venues
4. Cluster 3: Accommodation



4 Discussion

Now we have the basic level of information about the areas and house prices in the North East. There are many more things that we could explore that would be of use to property/investors landlords. For example, we could also add to the areas the average rental price in that particular area which would give more insight into the returns on the investment. When analysing the market before we invest, we also want to take a look at factors such as:

1. Average household income income growth
2. Average crime rate in the area
3. Population growth
4. Job growth

If we wanted to expand beyond the North East and look at other possible areas in the UK to invest in, we would want to add in this information for better analysis of the market.

For areas with Universities, we may also want to make clear the areas where students tend to be living.This way, landlords can get a good idea on where students are situated (they may want to invest here for the purpose of student lets or even avoid these areas altogether).

5 Conclusion

Real estate is a big market which many people are trying to enter. The key to being successful is to be able to analyse the market and do enough research on your area before jumping straight in. By performing this analysis, we can pick the right areas to buy based on price and social environment. For landlords looking to target mainly families, they may choose to buy in 'Mixed social venues' areas with range £50,000-£80,000 sale price.

With access to location data using platforms such as FourSquare, we can use data analysis to give us quick answers to find areas we want to invest in based on our own desires for the property.