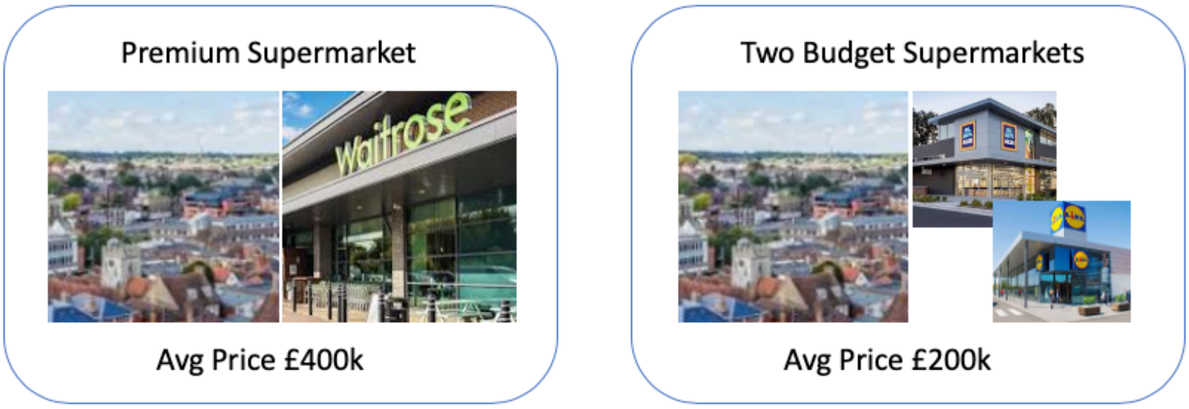
Inferring Disposable Income from Property Values and Supermarket Proximity

Gerry Baird

May 2020

Introduction

If you'd ever been shopping for groceries in the UK you wouldn't need a statistical genius to tell you which of the two towns illustrated below probably had the highest disposable income. We don't need to know anything about the household earnings to form an opinion based on the average property price and the presence of certain supermarkets.



Beyond this intuitive understanding is it possible to quantify this somehow by applying machine learning? Can we leverage the detailed analysis performed by the supermarket chains themselves to infer anything about disposable income?

Grocery shopping is dominated by large supermarket chains, but they target different customer segments. Premium supermarkets are operated by Waitrose, mainstream supermarkets are operated by Tesco, Sainsburys, Asda and Morrisons. Budget supermarkets are operated by Lidl and Aldi. Can we leverage the detailed analysis performed by these supermarket chains that led to them to open stores in some areas and not others to infer anything about these areas in terms of disposable income?

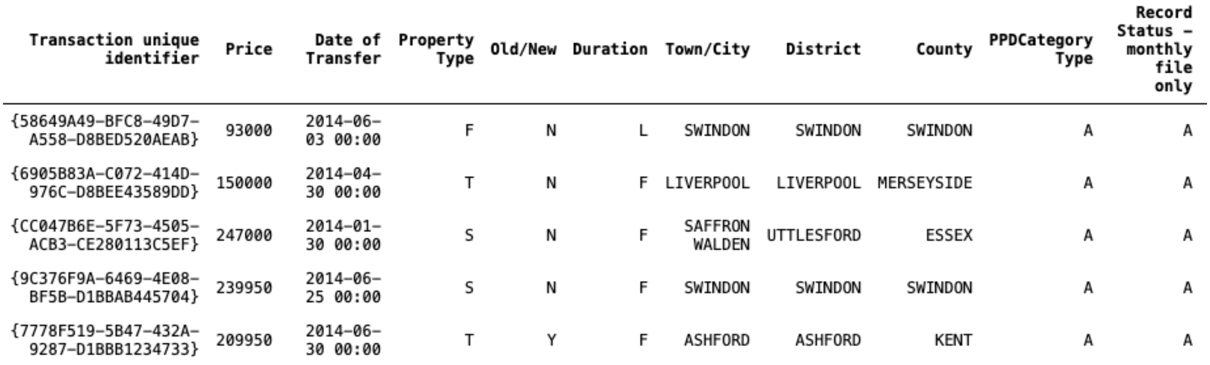
This project will analyse towns and cities in England and Wales to see if any insights into disposable income can be derived from presence of certain supermarkets combined with property values. The presence of a supermarket can raise property prices, so the two features being analysed aren't truly independent. A study by Lloyds bank in 2017 suggested that a Waitrose can add as much as £36k to local property prices. £22k for a mainstream supermarket and £6k for a budget supermarket [1].

Data

This project will use property transaction data published by the UK Govt Land Registry on Kaggle, combined with geolocation data from Google, augmented with location insights provided by FourSquare.

Property Transaction Data

The transaction data is hosted on Kaggle [2]. There is 2.2 GB of data in the complete dataset, containing property transactions from 1995 until 2017. An example of the data is shown below.

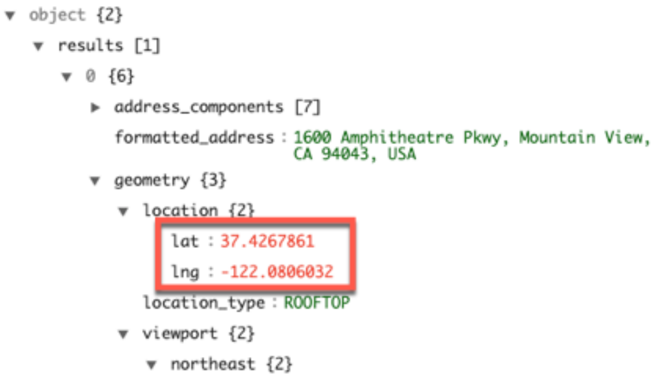


Transaction unique identifier	Price	Date of Transfer	Property Type	Old/New	Duration	Town/City	District	County	PPDCategory Type	Record Status - monthly file only
{58649A49-BFC8-49D7-A558-D8BED520AEAB}	93000	2014-06-03 00:00	F	N	L	SWINDON	SWINDON	SWINDON	A	A
{6905B83A-C072-414D-976C-D8BEE43589DD}	150000	2014-04-30 00:00	T	N	F	LIVERPOOL	LIVERPOOL	MERSEYSIDE	A	A
{CC047B6E-5F73-4505-ACB3-CE280113C5EF}	247000	2014-01-30 00:00	S	N	F	SAFFRON WALDEN	UTTLESFORD	ESSEX	A	A
{9C376F9A-6469-4E08-BF5B-D1BBAB445704}	239950	2014-06-25 00:00	S	N	F	SWINDON	SWINDON	SWINDON	A	A
{7778F519-5B47-432A-9287-D1BBB1234733}	209950	2014-06-30 00:00	T	Y	F	ASHFORD	ASHFORD	KENT	A	A

For this project we are interested in the price, the transaction date, the town, district and county. One significant limitation of the data is that precise location data isn't provided, we have to infer the location from town, district in county. This is rather imprecise for large cities resulting in a lack of detail for any large city such as Birmingham, London or Liverpool as shown above.

Geolocation Data

Google provides a geolocation API as part of the Google Maps platform [3]. Given an address, the geolocation API will try to provide a precise latitude and longitude even if the address is ambiguous or incomplete. The example response below shows the location information returned by the API for Google's head office.

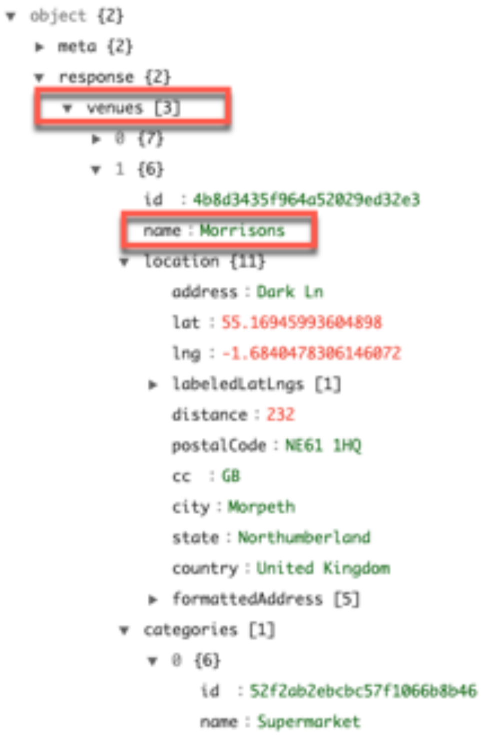


```
▼ object {2}
  ▼ results [1]
    ▼ 0 {6}
      ▶ address_components [7]
        formatted_address : 1600 Amphitheatre Pkwy, Mountain View,
                            CA 94043, USA
      ▼ geometry {3}
        ▼ location {2}
            lat : 37.4267861
            lng : -122.0806032
          location_type : ROOFTOP
        ▼ viewport {2}
          ▼ northeast {2}
```

Venue Data

The FourSquare Places API provides details of venues within a radius of a given location [4]. This can be a general search or a search within a specific category of venue such as a supermarket. For the purposes of this exercise I was searching for supermarket venues within a 3000 meter radius of a given longitude and latitude.

The example below shows the search result for Morpeth in Northumberland.



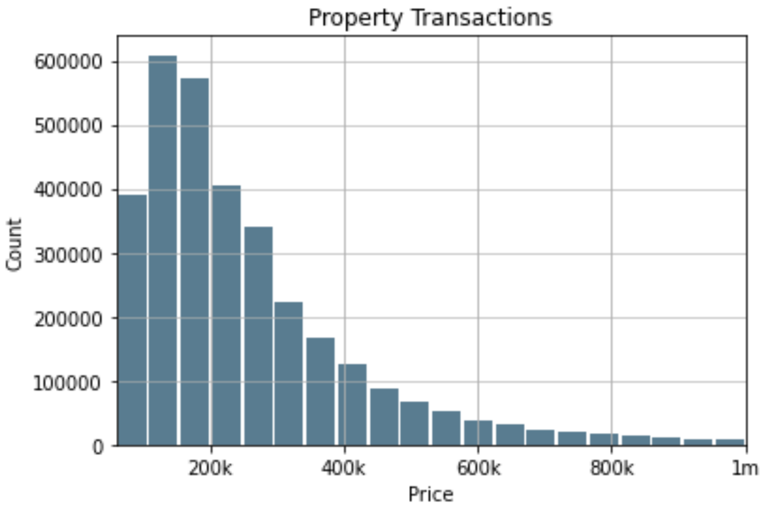
```
▼ object {2}
  ► meta {2}
  ▼ response {2}
    ▼ venues [3]
      ► 0 {7}
      ▼ 1 {6}
          id  : 4b8d3435f964a52029ed32e3
          name : Morrisons
        ▼ location {11}
            address : Dark Ln
            lat : 55.16945993604898
            lng : -1.6840478306146072
          ► labeledLatLngs [1]
            distance : 232
            postalCode : NE61 1HQ
            cc  : GB
            city : Morpeth
            state : Northumberland
            country : United Kingdom
          ► formattedAddress [5]
        ▼ categories [1]
          ▼ 0 {6}
              id  : 52f2ab2ebcbc57f1066b8b46
              name : Supermarket
```

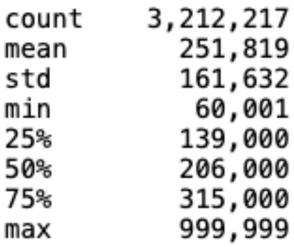
For this search there were 3 venues found within the search radius where the category was supermarket. I've expanded the second result which is for a Morrisons supermarket.

Methodology

Exploratory Data Analysis

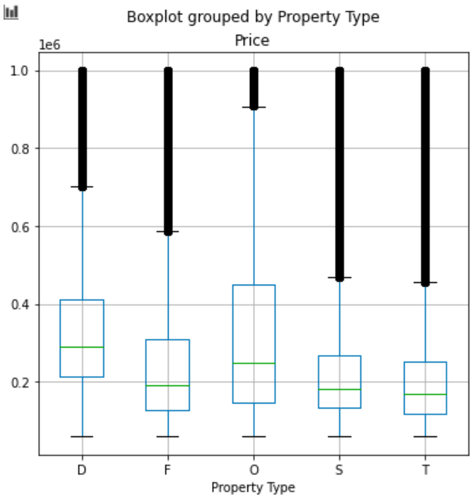
There is 2.2 GB of data in the complete dataset, containing all property transactions from 1995 until 2017. We are only interested in recent data as it isn't easy to accurately estimate a property's value today from a transaction that took place 20 years ago, we will only include sales after 2013. We'll also remove any outliers as these could skew the average price in a given tow, we only include transactions between £60,000 and £1,000,000 but there are still 3.2m transactions.





```
count    3,212,217
mean       251,819
std        161,632
min         60,001
25%        139,000
50%        206,000
75%        315,000
max        999,999
```

For simplicity we will calculate the average property price for any area regardless of property type, but it should be noted that there is a difference between different types of property even if there is a lot of overlap.

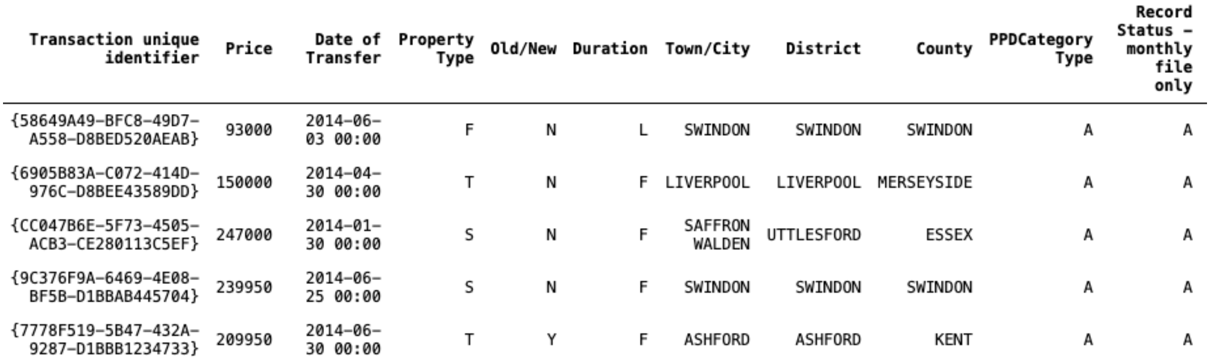


D : Detached
F : Flats/Apartments
O : Other
S : Semi-Detached
T : Terraced

On average flats and apartments are cheaper than terraced houses, terraced houses are cheaper than semi-detached houses, which in turn are cheaper than detached houses.

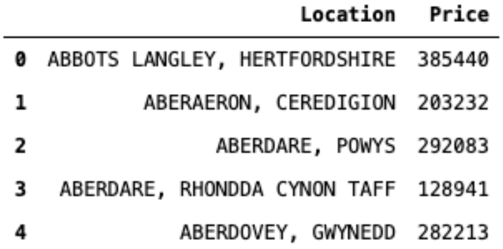
Preparing the Data

To prepare the data we need to remove any redundant data and calculate an average price for each town. The base data in the dataset had the structure below:



Transaction unique identifier	Price	Date of Transfer	Property Type	Old/New	Duration	Town/City	District	County	PPDCategory Type	Record Status - monthly file only
{58649A49-BFC8-49D7-A558-D8BED520AEAB}	93000	2014-06-03 00:00	F	N	L	SWINDON	SWINDON	SWINDON	A	A
{6905B83A-C072-414D-976C-D8BEE43589DD}	150000	2014-04-30 00:00	T	N	F	LIVERPOOL	LIVERPOOL	MERSEYSIDE	A	A
{CC047B6E-5F73-4505-ACB3-CE280113C5EF}	247000	2014-01-30 00:00	S	N	F	SAFFRON WALDEN	UTTLESFORD	ESSEX	A	A
{9C376F9A-6469-4E08-BF5B-D1BBAB445704}	239950	2014-06-25 00:00	S	N	F	SWINDON	SWINDON	SWINDON	A	A
{7778F519-5B47-432A-9287-D1BBB1234733}	209950	2014-06-30 00:00	T	Y	F	ASHFORD	ASHFORD	KENT	A	A

The data was loaded into a Pandas dataframe which makes it very easy to manipulate the data. I dropped columns that weren't relevant to the analysis then I created a new location field in the dataset that combined the existing town/city with county as this would be used in the subsequent geolocation. I then grouped the data using the new location field and calculated the average price.



	Location	Price
0	ABBOTS LANGLEY, HERTFORDSHIRE	385440
1	ABERAERON, CEREDIGION	203232
2	ABERDARE, POWYS	292083
3	ABERDARE, RHONDDA CYNON TAFF	128941
4	ABERDOVEY, GWYNEDD	282213

Enriching the Data and Feature Engineering

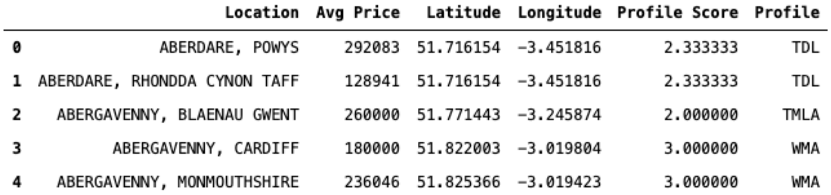
Enrichment was comprised of three stages for each town in the dataset.

1. Determine the precise longitude and latitude of the town.
2. Use the precise location the search for supermarkets within a 3km radius
3. Aggregate the supermarkets together and derive a score for the town.

Stages 1 & 2 used public API's from Google [3] and FourSquare [4] respectively, and I won't describe them in detail here.

Stage 3 involved feature engineering, that is, creating a new feature by applying domain knowledge to existing features. The goal was to create a score that reflected the makeup of supermarkets in a given town, a higher score would be given for premium and mainstream supermarkets, a score would be lowered if there were budget supermarkets nearby.

This required counting the occurrences of each type of supermarket within a town and producing an aggregate score. Any premium supermarket (Waitrose) was given a score of 5, mainstream supermarkets were given a score of 3 and the budget supermarkets were given a score of 1. An average score called Profile Score was calculated and then added to the dataset.



	Location	Avg Price	Latitude	Longitude	Profile Score	Profile
0	ABERDARE, POWYS	292083	51.716154	-3.451816	2.333333	TDL
1	ABERDARE, RHONDDA CYNON TAFF	128941	51.716154	-3.451816	2.333333	TDL
2	ABERGAVENNY, BLAENAU GWENT	260000	51.771443	-3.245874	2.000000	TMLA
3	ABERGAVENNY, CARDIFF	180000	51.822003	-3.019804	3.000000	WMA
4	ABERGAVENNY, MONMOUTHSHIRE	236046	51.825366	-3.019423	3.000000	WMA

The profile column shows which supermarkets were discovered. Waitrose(W), Sainsburys(S), Tesco(T), Morrisons(M), Asda (D), Aldi(A) and Lidl(L). The method for calculating the profile score Abergavenny, shown in row 4, is illustrated below. Abergavenny has Waitrose(W), Morrisons(M) and Aldi(A) nearby, hence WMA.

Score: 5

Score: 3

Score: 1

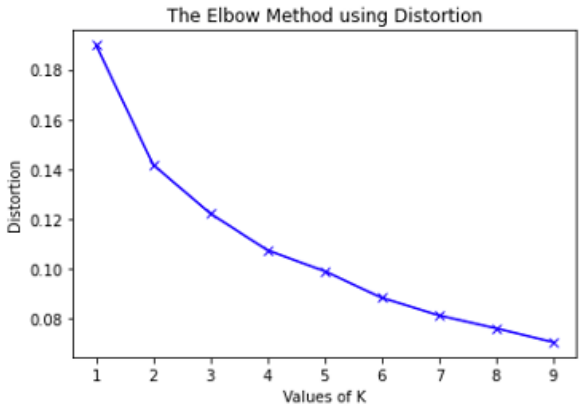
$$\text{Profile Score} = \frac{5 + 3 + 1}{3} = 3$$

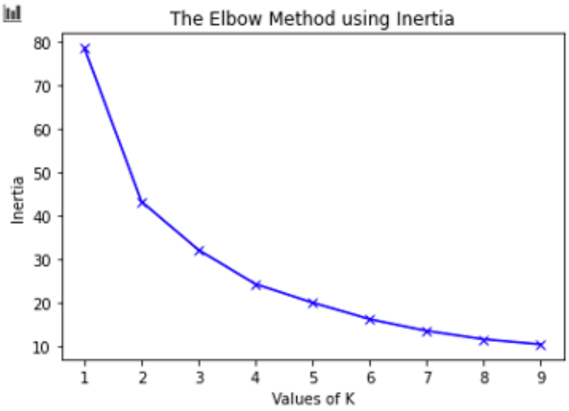
The profile score is a new feature within the data and is intended to represent the presence of certain supermarkets within a town. The intuition behind the score is best explained with some further examples. Consider a small town where the only supermarket is a Waitrose, this would have a profile score of 5. Consider another town that only had a Lidl and an Aldi, this town would have a score of 1.

Clustering

K-Means clustering was applied to the average property price and profile score. As the profile score is already biased to reflect the higher score of premium supermarkets, I decided to use a min/max scaler rather than a standard scaler as the standard scaler would try to represent the significance of the score in the scaled values, essentially double counting the bias I had already engineered into the profile score.

K-Means is an intuitive unsupervised clustering method that tries to group the observations into a predefined number of clusters (K). To determine the optimal number of clusters I ran the algorithm several times and measured the distortion and inertia. Ideally, we are looking for a clear elbow in the graph that indicates where the distortion levels off and further clusters add little or no value.

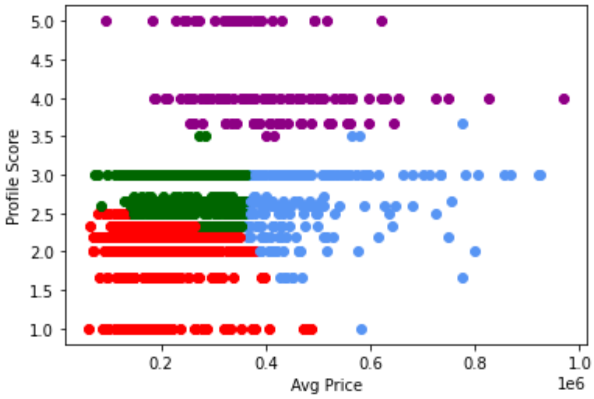




Distortion is a measure of how spread out the cluster centres are. Inertia is a measure of how closely clumped the clusters are. This analysis didn’t reveal a clear elbow, just a flattening from 4-6 on the inertia plot. As such I decided that 4 clusters would provide some new information without making the interpretation of the results overly complex.

Results

With a K of 4 the clustering algorithm gave each observation a label of 0, 1,2 or 3. When plotted on a scatter plot the following clusters were revealed:



Understanding the results of clustering requires some interpretation and domain knowledge. Based on my understanding I drew the following conclusions from the clustering.

- Cluster 0: Blue. High property prices, mix of all supermarkets
- Cluster 1: Red. Lower property prices, budget and mainstream supermarkets.
- Cluster 2: Purple. Mid-high property prices, bias towards premium supermarkets.
- Cluster 3: Green. Lower property prices, mix of all supermarkets

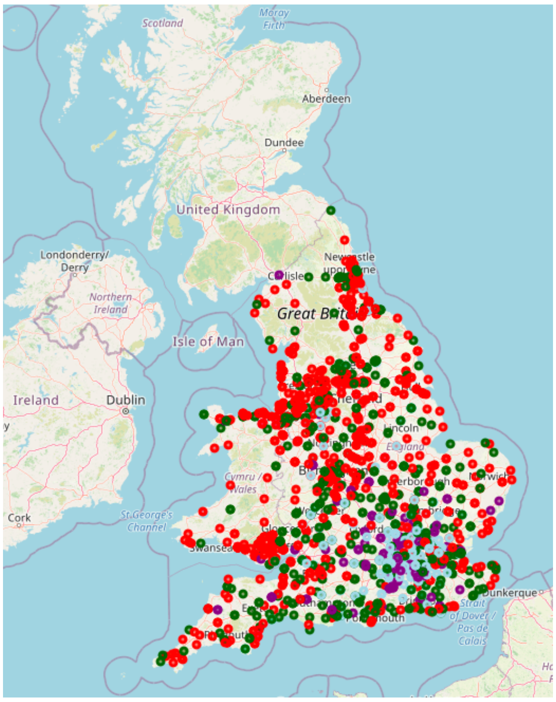
What can we infer about disposable income from this analysis? The red cluster contains towns where property prices are lower than average, this alone doesn't reveal much about income but the bias towards budget and mainstream supermarkets suggests that wages are also lower on average in these areas.

The average house price in the green cluster is no different from that in red, yet we start to see the presence of premium supermarkets which suggests wages are higher, housing costs are similar and therefore disposable income should be higher too.

The blue cluster contains more expensive housing than green, yet the supermarket mix is very similar suggesting similar wages to those in green. If incomes are similar between blue and green then the higher housing costs incurred in blue would suggest that disposable incomes in blue are lower than those in green.

The purple cluster has the highest presence of premium supermarkets yet a wide range of property prices too. This suggests that wages are higher in this cluster compared to other to other clusters where property prices are similar. As such, disposable incomes are highest in this cluster.

Towns from every cluster are spread around the UK but it is clear that most towns in the blue cluster are grouped around London.



Broadly speaking the results confirm the intuition around disposable income as house prices around London are very high yet there a mix of supermarkets catering for all budgets.

Discussion

The results produced from this analysis confirmed the intuition around disposable income, house prices and the presence of premium supermarkets. However, a number of compromises had to be accepted.

Early work with the property transaction data was done with a small subset of counties I know well. When the data was grouped to provide an average price for a town it became clear that some transactions are recorded in the wrong county. Furthermore, the same town can be represented twice if different town and district combinations are used just because of the way the data is grouped. For example, Whychavon, Wyre Forest, Worcestershire might also be reported as Whychavon, Worcestershire, Worcestershire. As such the district feature was discarded and I relied upon Google geolocation to locate the towns.

Property sales in large cities are reported at the city level so precise locations aren’t available. This also places the centre of the search radius a long way from potential supermarkets. It should be possible to improve the accuracy of the score by basing the search radius on town population, the larger the town the greater the search radius from its centre.

Some extra bias could be introduced for very small towns that can support a Waitrose. For example, Marlborough in Wiltshire is very small yet has a Tesco and a Waitrose.

Some premium food retailers are hard to find. I limited my search to Waitrose, Asda, Tesco, Sainsburys, Morrisons, Aldi and Lidl. I tried to include Marks and Spencer, but they are categorised as Supermarkets where they just sell food or department stores where a food supermarket is embedded into a traditional M&S department store. I was constrained by FourSquare API limits but if I was doing this for a commercial project I would augment the supermarket category search with a broader venue search in every town. I'd also widen the search to include other supermarkets such as Co-Op and Iceland.

Some food outlets, particularly from Waitrose and M&S couldn't be described as regular supermarkets as they are found on petrol station forecourts and motorway service areas, as such these should be excluded.

I've tried to simplify the supermarket landscape in the UK to keep this project manageable but there are further complexities around the feature engineering of the profile score that could be explored. The scores themselves are rather arbitrary but they could be based on some derived average price for certain common foods, such as, raw chicken, gin, sliced bread. Finally, some towns have multiple Tesco's or Sainsbury whereas my model only counts the first occurrence.

Conclusion

Property prices and the mix of supermarkets in a given area can be an indicator of disposable income but the clustering results require domain knowledge and are open to some interpretation. Whilst I believe my interpretation is reasonable, others are possible, especially if the number of clusters is increased.

Opportunities exist to improve the descriptive power of the model by incorporating additional features such as population data and a more exhaustive search for supermarket venues based on a wider range of supermarket chains and a more flexible search radius. The model could be extended to include the presence of certain car dealerships in an area to. Rather than looking for a Waitrose we'd also be looking for a Porsche or BMW dealership.

It should be noted that the presence of a certain supermarket isn't an indicator for individual shopping behaviour. Many people, myself included, will shop at a wide range of stores depending on need. It is however a reflection of the detailed analysis a supermarket chain made, perhaps many years ago, that led to an outlet opening in a particular location.

Finally, supermarkets that were once thriving may close, new entrants such as Aldi and Lidl bring disruption. A town that could easily sustain a Waitrose or M&S may not have one because it is already saturated or has limited viable sites. The analysis that led to a store opening five years ago may give a very different result if repeated today.

References

[1] The Waitrose Effect. https://www.independent.co.uk/property/house-prices-latest-waitrose-effect-sainsburys-marks-and-spencer-uk-property-a7760926.html

[2] UK Property Data, free login required. https://www.kaggle.com/hm-land-registry/uk-housing-prices-paid

[3] Google Geocoding API. https://developers.google.com/maps/documentation/geocoding/start

[4] FourSquare API. https://enterprise.foursquare.com/products/places