Using data approach to identify the best place of living in Cambridge city, UK ¶

By Hiep Nguyen

Introduction

Background

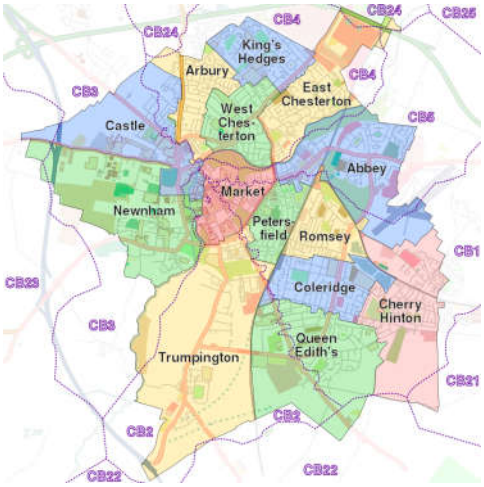
I recently received a job offer to move to Cambridge city, UK. After searching for some basic information on Google as such cost of living, schooling, housing, rate of crimes… in each district in Cambridge city and some nearby cities in Cambridgeshire, I am getting lost.
Especially when I read a news posted in a recent poll that indicated "Peterborough has retained its unwanted crown as England's worst place to live, topping an online poll for the third year running." https://www.cambridge-news.co.uk/news/local-news/peterborough-named-worst-place-live-19560796. While Peterborough is a cathedral city and unitary authority area in the north of Cambridgeshire. This will affect my selection where to live in Cambridge city or its nearby city/town/village.

Business problem

The main problem is there is no consolidated report based on data that combines all the information and shares insights about which areas are the best places to live in Cambridge city or Cambridgeshire. This should be based on variety and high dense of essential venues, affordable price of housing, lower rate of crimes…
For whoever like me, planning to move into Cambridge city or already in UK but looking to move into Cambridge city, can be beneficial to this insight to decide the best place to live, according to each own specific preference.

Facts

"Cambridgeshire is noted as the site of Flag Fen in Fengate, one of the earliest-known Neolithic permanent settlements in the United Kingdom, compared in importance to Balbridie in Aberdeen, Scotland." Wikipedia. Its in the Eastern part of England. Cambridge city is at the center of Cambridgeshire, homes to the prestigious University of Cambridge, the 2nd oldest English university in the world, dating from 1209. Cambridge city lays on the River Cam, around 55 miles (89km) north of London, heart of the high-technology Silicon Fen for software and bioscience (over 40% of workforce with higher education qualification). From the UK census 2011, its population is approximately 158,000 people.
There are 123 wards in Cambridgeshire in which 14 wards are in Cambridge city. Along with focusing on these 14 wards in Cambridge city, the more wards close to Cambridge are considered, the better information is available for end-user like me to understand the pros and cons to choose a suitable place to live.
Many available website as such UK local area, One Dome which provide quite interesting indicators for reference on the side.



Data

Data which is going to use in this project will include essential venues from Foursquare, price of housing over years, crimes record, traffic accidents, public wifi access… for Cambridge (and possible for other nearby villages if time and data allow).
Most of the data is collected from **data.cambridgeshireinsight** which is an open data portal for Cambridgeshire and Peterborough.

Other sources:

UK Postcode data
Cambridgeshire parishes database
UK Census 2011 data
Official Statistics of UK indices for deprivation in 2010 which is published 24 March 2011

Work and Output¶

1. Methodology

1.1. Python Libraries

All relevant python libraries for this project which not needed to install are imported in the first coding place. This is beneficial for me to quickly grasp what libraries were used in this project later. Other remaining libraries which needed to install will be kept as comments and to be installed in later section.

This will save time when running the code since the kernel will stop ("Dead kernel") after few hours when I am not active in this environment (later I found that using Skills Network Labs from Coursea is much better, no Monthly compute usage limit as Watson Studio lite-v1 plan).

1.2. Import Datasets

After exploring data and information for many different resources, I upload the most related data into GitHub repository. There will be different dataframes with different information which I will combine them using Python.

#Cambridgeshire_Names_Clean.csv: contains all the Output Area Codes for each wards in Cambridgeshire.

```
[2]: #Cambridgeshire_Names_Clean.csv: contains all the Output Area Codes for each wards in Cambridgeshire.
url = 'https://raw.githubusercontent.com/TNguyen50/Course10Capstone/master/Cambridgeshire_Names_Clean.csv'
df_cambridgeshire = pd.read_csv(url)
print ('Cambridgeshire has {} Wards and {} Output Area Codes.'.format(len(df_cambridgeshire['Ward name'].unique()),df_cambridgeshire.shape[0]))
#df_cambridgeshire.head(5)
df_cambridge = df_cambridgeshire[df_cambridgeshire['Local Authority Name'].str.match('Cambridge')].reset_index(drop=True)
print ('Cambridge city has {} Wards and {} Output Area Codes.'.format(len(df_cambridge['Ward name'].unique()),df_cambridge.shape[0]))
df_cambridge_ward = df_cambridge[['Ward name']]
df_cambridge_ward = df_cambridge_ward.groupby('Ward name').first().reset_index()
df_cambridge_ward.rename(columns={'Ward name': 'Ward'}, inplace = True)
df_cambridge_ward

Cambridgeshire has 123 Wards and 1937 Output Area Codes.
Cambridge city has 14 Wards and 372 Output Area Codes.
```

#UK Census 2011 data.

```
[3]: #UK Census 2011 data.
uk_census = ["birthcountry","dwelling","economic","ethnicity","health","population","qualifications","religion","residence","traveltowork"]
url_head = "https://raw.githubusercontent.com/TNguyen50/Course10Capstone/Census2011/Census-database_wards_V3_"
url_tail = ".csv"
df_census_list = [] # create empty list of dataframes for uk_census data
for census in uk_census:
    url_census = url_head + census + url_tail
    df_census_list.append(pd.read_csv(url_census))
```

Cambridgeshire crime rate data downloaded Jul-2021, data from 2007 to 2014.

Crime rate is the rate of crime per 1,000 residents in each ward (apart from Burglary in a dwelling which a rate per 1,000 dwellings).

```
[5]: # Cambridgeshire crime rate data downloaded Jul-2021, data from 2007 to 2014.
# Crime rate is the rate of crime per 1,000 residents in each ward (apart from Burglary in a dwelling which a rate per 1,000 dwellings).
url_crime = 'https://raw.githubusercontent.com/TNguyen50/Course10Capstone/master/RateperP1000peopleofcrimeinCambbyfinancialyrSD_0.csv'
df_cambridgeshire_crime_rate = pd.read_csv(url_crime)
df_cambridgeshire_crime_rate.rename(columns={'WardName': 'Ward'}, inplace=True)
df_cambridgeshire_crime_rate.head()
```

Then I keep only crime rate for Cambridge city:

[6]:
```
# Keep only crime rate for Cambridge city
df_cambridge_crime_rate = df_cambridgeshire_crime_rate.merge(df_cambridge_ward, how = 'inner', on = 'Ward')
df_cambridge_crime_rate.drop(columns = 'Ward Code', axis = 1, inplace=True)
df_cambridge_crime_rate.head()
```

[6]:

	Ward	Total Crime Rate 2007-2008	Total Crime Rate 2008-2009	Total Crime Rate 2009-2010	Total Crime Rate 2010-2011	Total Crime Rate 2011-2012	Total Crime Rate 2012-2013	Total Crime Rate 2013-2014	ASB Rate 2007-2008	ASB Rate 2008-2009	ASB Rate 2009-2010	ASB Rate 2010-2011	ASB Rate 2011-2012	ASB Rate 2012-2013	ASB Rate 2013-2014	Burglary Dwelling Rate 2007-2008	Bu Dw
0	Abbey	132.7	134.3	162.5	115.3	112.7	94.9	80.6	125.5	111.3	78.1	79.8	71.4	50.3	45.4	15.6	
1	Arbury	110.7	115.7	87.5	82.8	68.1	59.5	62.1	113.9	102.4	65.0	57.8	52.1	40.0	32.8	25.8	
2	Castle	66.8	62.1	53.2	57.3	42.5	33.7	36.1	25.0	26.2	18.1	17.2	15.9	10.5	11.9	14.0	
3	Cherry Hinton	62.9	56.7	50.6	45.7	44.7	40.4	29.4	66.4	57.6	34.6	42.1	39.8	24.4	29.0	9.7	
4	Coleridge	80.9	92.4	84.2	78.3	58.4	53.8	54.6	74.4	69.4	55.2	56.5	46.6	32.8	32.2	7.8	

And narrow it down to crime rate for latest year 2013-2014:

[7]:
```
# Keep only crime rate for latest year 2013-2014
filter_columns = ['Ward'] + [col for col in df_cambridge_crime_rate.columns if col.endswith('2014')]
df_cambridge_crime_rate_2014 = df_cambridge_crime_rate.loc[:, filter_columns]
df_cambridge_crime_rate_2014.head()
```

[7]:

	Ward	Total Crime Rate 2013-2014	ASB Rate 2013-2014	Burglary Dwelling Rate 2013-2014	Criminal Damage Rate 2013-2014	Domestic Abuse Rate 2013-2014	Thef
0	Abbey	80.6	45.4	11.2	8.0	25.4	
1	Arbury	62.1	32.8	12.3	8.3	27.0	
2	Castle	36.1	11.9	12.2	2.1	3.5	

Mean price paid for all house types

[9]:
```
# Mean price paid for all house types
url_house_price = 'https://raw.githubusercontent.com/TNguyen50/Course10Capstone/master/Meanpricepaidforallhousetypes Ward.csv'
df_cambridgeshire_house_price = pd.read_csv(url_house_price)
df_cambridgeshire_house_price.rename(columns={'NAME': 'Ward'}, inplace=True)
df_cambridgeshire_house_price.head()
```

[9]:

	Ward	Jan 2014 - Dec 2014	Apr 2014 - Mar 2015	Jul 2014 - Jun 2015	Oct 2014 - Sep 2015	Jan 2015 - Dec 2015	Apr 2015 - Mar 2016	Jul 2015 - Jun 2016	Oct 2015 - Sep 2016	Jan 2016 - Dec 2016	Apr 2016 - Mar 2017	Jul 2016 - Jun 2017	Oct 2016 - Sep 2017	Jan 2017 - Dec 2017	Apr 2017 - Mar 2018	Jul 2017 - Jun 2018
0	Abbey	307046	319164	335936	349681	346037	341013	338512	321922	332510	347865	365474	384914	399431	393395	381531

I then keep only housing price for Cambridge city:

```
[76]: # Keep only housing price for Cambridge city
df_cambridge_house_price = df_cambridgeshire_house_price.merge(df_cambridge_ward, how = 'inner', on = 'Ward')
df_cambridge_house_price.head()
```

[76]:

	Ward	Jan 2014 - Dec 2014	Apr 2014 - Mar 2015	Jul 2014 - Jun 2015	Oct 2014 - Sep 2015	Jan 2015 - Dec 2015	Apr 2015 - Mar 2016	Jul 2015 - Jun 2016	Oct 2015 - Sep 2016	Jan 2016 - Dec 2016	Apr 2016 - Mar 2017	Jul 2016 - Jun 2017	Oct 2016 - Sep 2017	Jan 2017 - Dec 2017
0	Abbey	307046	319164	335936	349681	346037	341013	338512	321922	332510	347865	365474	384914	399431
1	Arbury	311585	338043	348177	369381	381335	391521	417729	409080	411853	417625	388282	419680	441992

1.3. Datasets Cleaning

In fact, I have cleaned the data during converting excel file into csv due to some issue about the format of orginal files. Most of data cleaning work are normally during importing or datasets exploring. And other data cleaning processes for this project will go along with Foursquare venue section.

1.4. Datasets Exploring

In this section, I will review the datasets with different charts to find some insights.

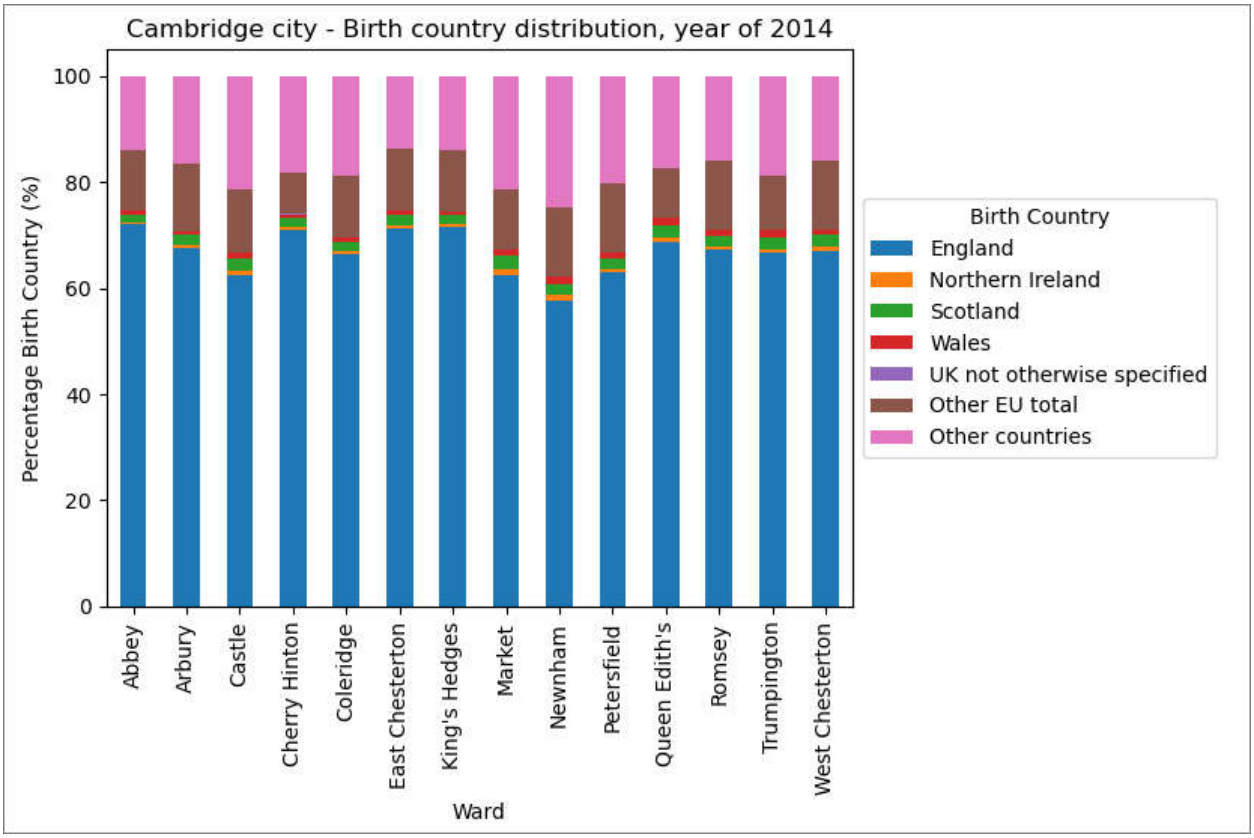
1.4.1. Birth country

Statistics:

```
[12]: df_census_cambridge_list[0].describe()
```

[12]:

	Total residents	England	Northern Ireland	Scotland	Wales	UK not otherwise specified	Other EU total	Other countries
count	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000
mean	8847.642857	5920.000000	63.500000	179.785714	84.571429	1.142857	1021.285714	1577.357143
std	774.690470	775.013945	19.657647	20.725972	22.193950	1.231456	160.713998	236.948098
min	7150.000000	4453.000000	41.000000	149.000000	55.000000	0.000000	686.000000	1278.000000
25%	8407.000000	5465.000000	51.500000	171.000000	66.000000	0.000000	892.500000	1418.750000
50%	9098.500000	6163.500000	57.500000	177.500000	80.500000	1.000000	1090.500000	1520.500000
75%	9352.500000	6265.750000	67.000000	189.000000	101.250000	1.750000	1123.500000	1660.750000
max	9907.000000	7127.000000	102.000000	220.000000	127.000000	4.000000	1214.000000	2103.000000



Observation: Newnham town has the most diversity of residences.

1.4.2. Dwelling

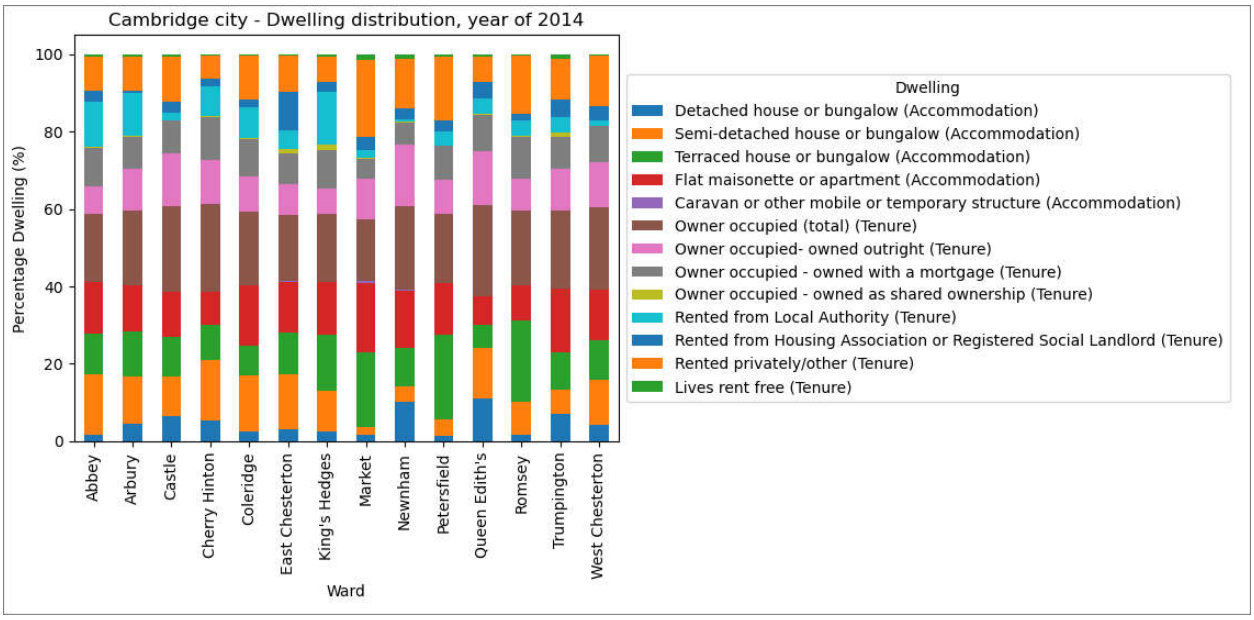
Statistics:

1.4.2. Dwelling

```
[14]: df_census_cambridge_list[1].describe()
```

[14]:

	Detached house or bungalow (Accommodation)	Semi-detached house or bungalow (Accommodation)	Terraced house or bungalow (Accommodation)	Flat maisonette or apartment (Accommodation)	Caravan or other mobile or temporary structure (Accommodation)	Owner occupied (total) (Tenure)	Owner occupied-owned outright (Tenure)	Owner occupied - owned with a mortgage (Tenure)	Owner occupied - owned as shared ownership (Tenure)	Rented from Local Authority (Tenure)	Rented from Housing Association or Registered Social Landlord (Tenure)	Rented privately/other (Tenure)	Lives rent free (Tenure)
count	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000
mean	349.285714	905.785714	1002.928571	1036.428571	5.714286	1621.214286	831.357143	752.285714	37.571429	507.785714	279.571429	875.571429	52.571429
std	235.357806	493.888056	459.981850	313.464687	10.858733	436.192649	217.736470	275.422034	39.853027	419.126964	226.541460	291.141709	14.949733
min	61.000000	84.000000	438.000000	616.000000	0.000000	619.000000	411.000000	193.000000	5.000000	28.000000	63.000000	548.000000	37.000000
25%	173.250000	538.000000	732.250000	722.000000	1.000000	1496.750000	718.250000	698.250000	14.250000	165.000000	155.750000	625.750000	44.000000
50%	314.500000	1037.500000	912.000000	1125.500000	1.000000	1728.000000	775.000000	811.000000	21.500000	353.500000	215.500000	853.500000	48.500000
75%	444.500000	1306.250000	1105.250000	1286.250000	2.500000	1860.250000	1016.500000	927.500000	33.000000	743.500000	335.750000	1013.250000	60.750000
max	952.000000	1546.000000	2001.000000	1462.000000	32.000000	2156.000000	1214.000000	1075.000000	135.000000	1293.000000	983.000000	1426.000000	93.000000



Observation: Market ward has the largest rented privately/other tenure.

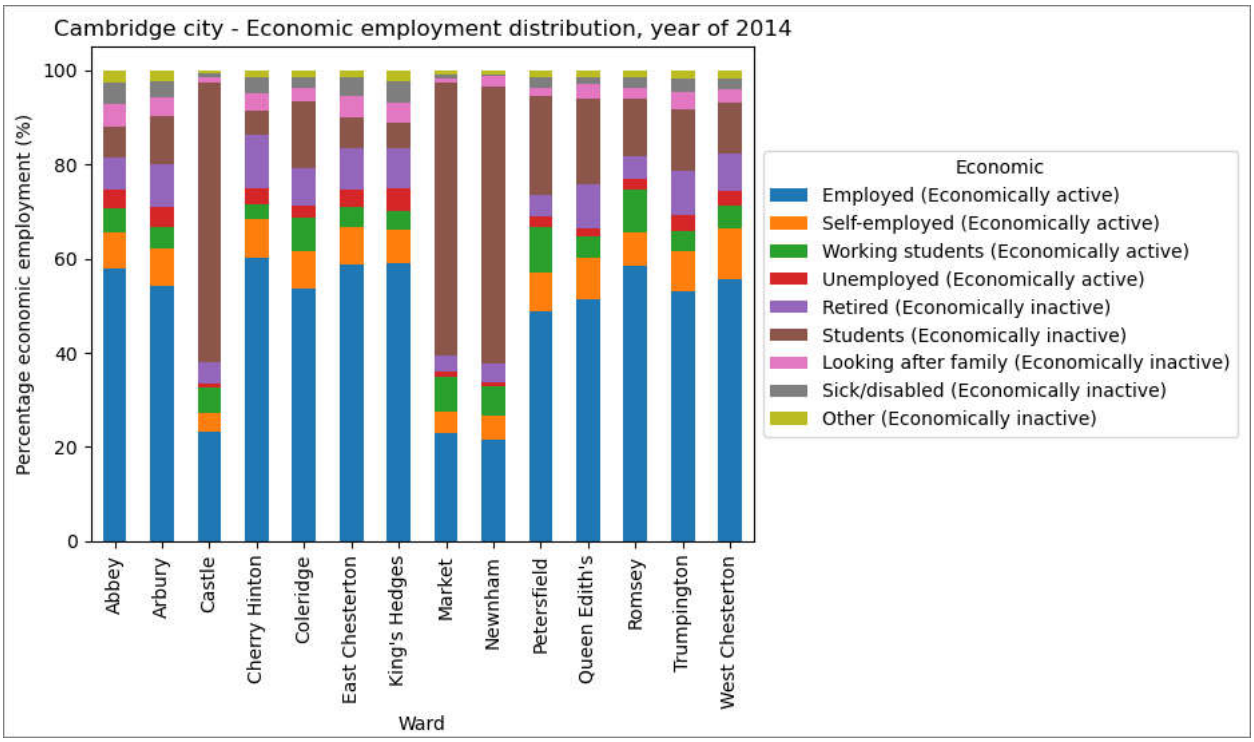
1.4.3. Economic

Statistics:

```
[16]: df_census_cambridge_list[2].describe()
```

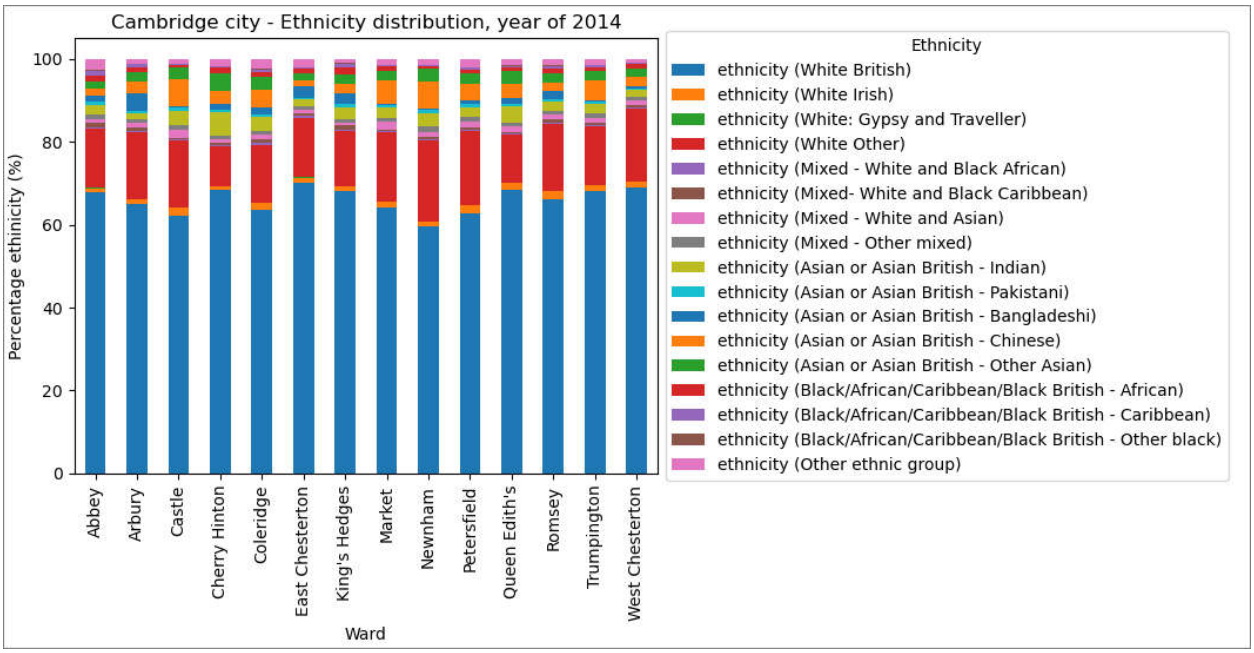
[16]:

	Employed (Economically active)	Self-employed (Economically active)	Working students (Economically active)	Unemployed (Economically active)	Retired (Economically inactive)	Students (Economically inactive)	Looking after family (Economically inactive)	Sick/disabled (Economically inactive)	Other (Economically inactive)	total population aged 16-74
count	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000
mean	3382.214286	520.785714	400.214286	188.642857	500.071429	1533.071429	210.642857	174.071429	110.500000	7020.214286
std	1007.438349	117.056317	150.260890	87.037222	155.955351	1561.606050	83.073283	95.886523	40.099396	649.798688
min	1439.000000	306.000000	211.000000	58.000000	213.000000	333.000000	62.000000	27.000000	53.000000	6025.000000
25%	3246.250000	496.000000	290.500000	127.500000	381.250000	554.250000	155.750000	104.250000	98.750000	6711.500000
50%	3757.000000	553.500000	355.500000	183.500000	550.000000	853.500000	209.000000	172.000000	104.000000	6911.500000
75%	4032.000000	583.000000	477.250000	252.500000	613.250000	1406.250000	259.000000	228.750000	116.250000	7207.250000
max	4454.000000	739.000000	686.000000	333.000000	738.000000	5190.000000	357.000000	346.000000	188.000000	8749.000000



Observations: There is huge number of students in Castle, arket and Newnham wards. The highest unemployed percentage is at King's Hedges ward.

1.4.4. Ethnicity



Observations: More other Asian ethnicity population in Cherry Hinton ward, where it may be better for other Asian communities.

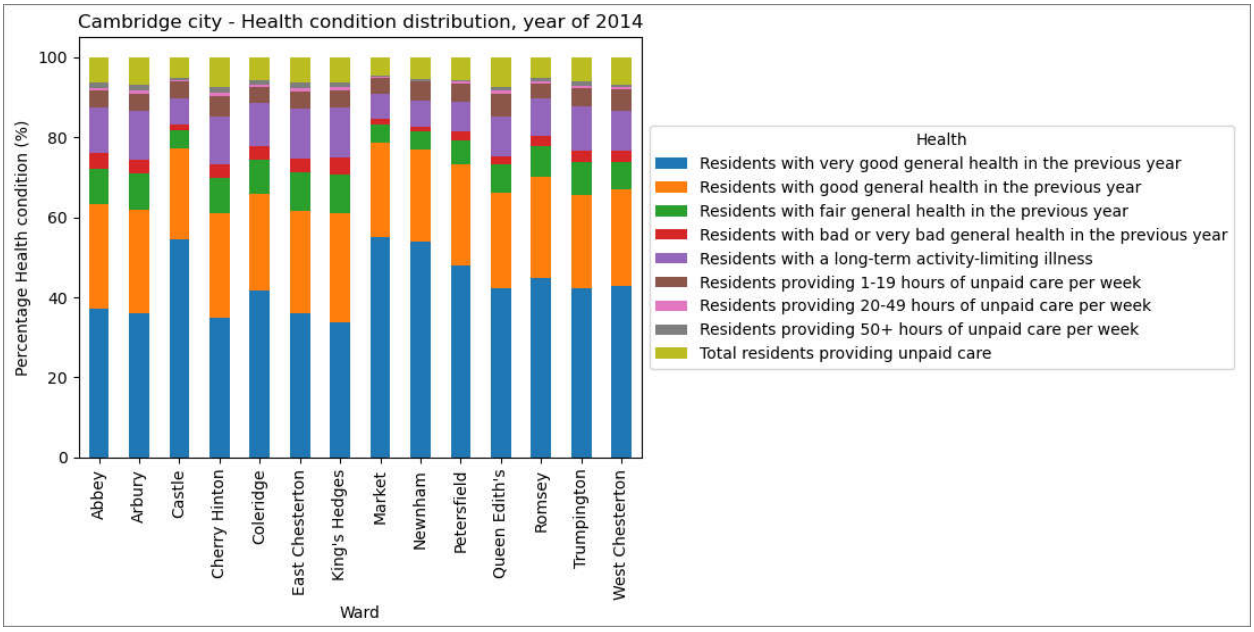
1.4.5. Health

Statistics:

```
[20]: df_census_cambridge_list[4].describe()
```

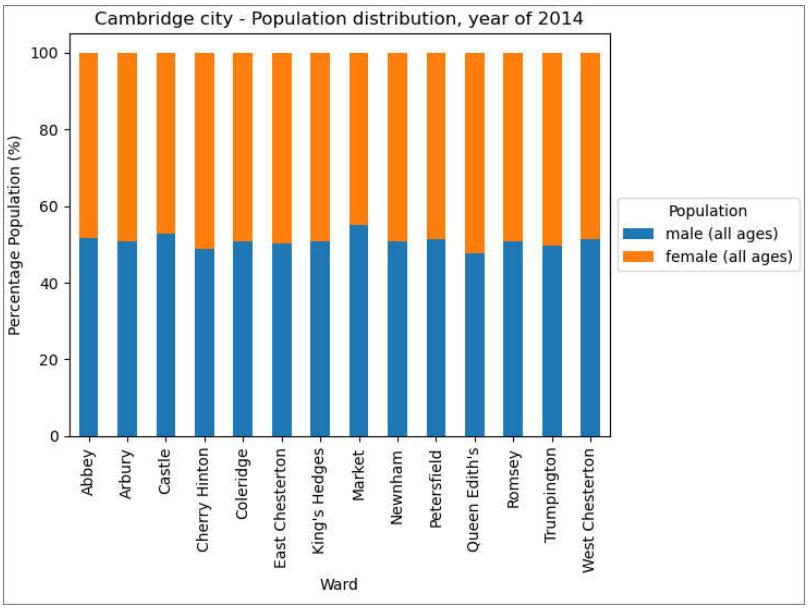
[20]:

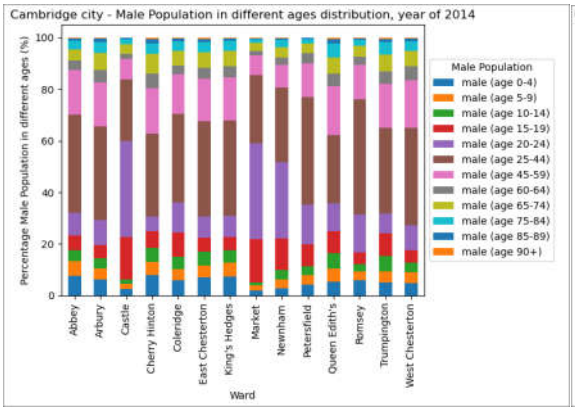
	Residents with very good general health in the previous year	Residents with good general health in the previous year	Residents with fair general health in the previous year	Residents with bad or very bad general health in the previous year	Residents with a long-term activity-limiting illness	Residents providing 1-19 hours of unpaid care per week	Residents providing 20-49 hours of unpaid care per week	Residents providing 50+ hours of unpaid care per week	Total residents providing unpaid care
count	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000
mean	4838.785714	2828.571429	858.642857	321.642857	1147.428571	510.428571	74.428571	113.500000	698.357143
std	570.004073	423.256190	280.265700	132.027158	356.110336	91.621894	31.653689	54.430973	151.107505
min	4125.000000	1993.000000	374.000000	114.000000	530.000000	326.000000	19.000000	39.000000	384.000000
25%	4451.250000	2612.250000	656.000000	231.500000	852.500000	466.000000	52.750000	62.500000	595.250000
50%	4840.000000	2906.500000	857.500000	310.000000	1166.500000	506.000000	79.000000	105.500000	724.500000
75%	5104.750000	3148.750000	1102.500000	421.500000	1464.750000	538.000000	100.000000	164.500000	811.250000
max	6402.000000	3409.000000	1207.000000	513.000000	1571.000000	709.000000	116.000000	191.000000	907.000000

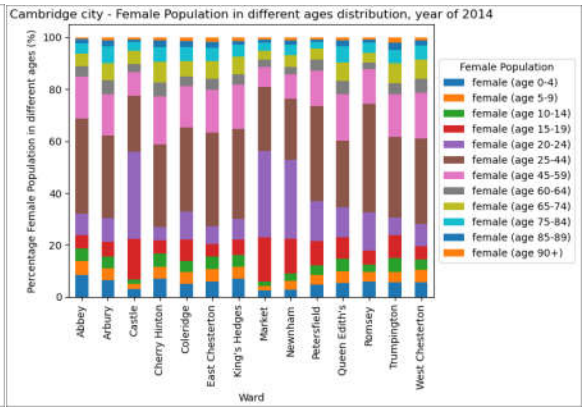


Observation: Residences in Market ward have the best health condition.

1.4.6. Population







Observation: Depending on the age of each child in the family, it may be beneficial to select the ward that has more similar age population. So that, the child or even adult can have more friends at similar age. Lets say for a child at 10 years old, its better to select Trumpington or Cherry Hinton ward to live.

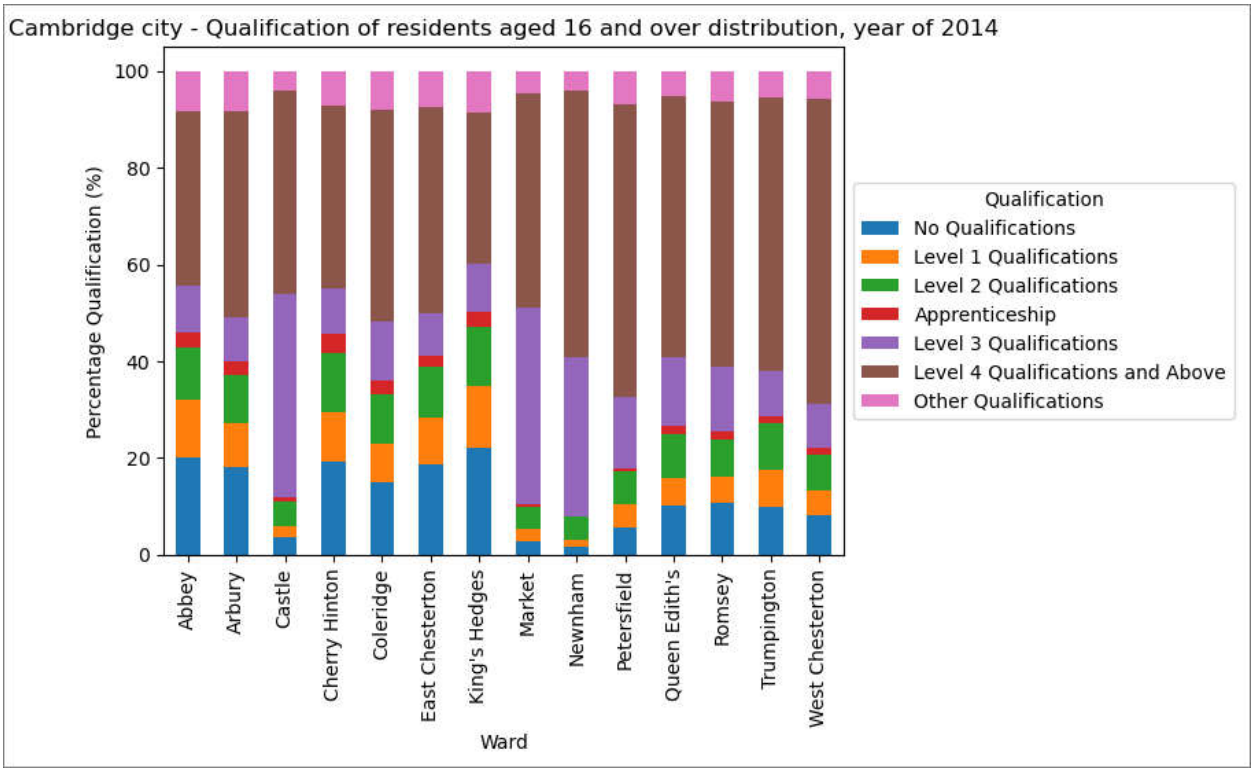
1.4.7. Qualifications

Statistics:

```
[27]: df_census_cambridge_list[6].describe()
```

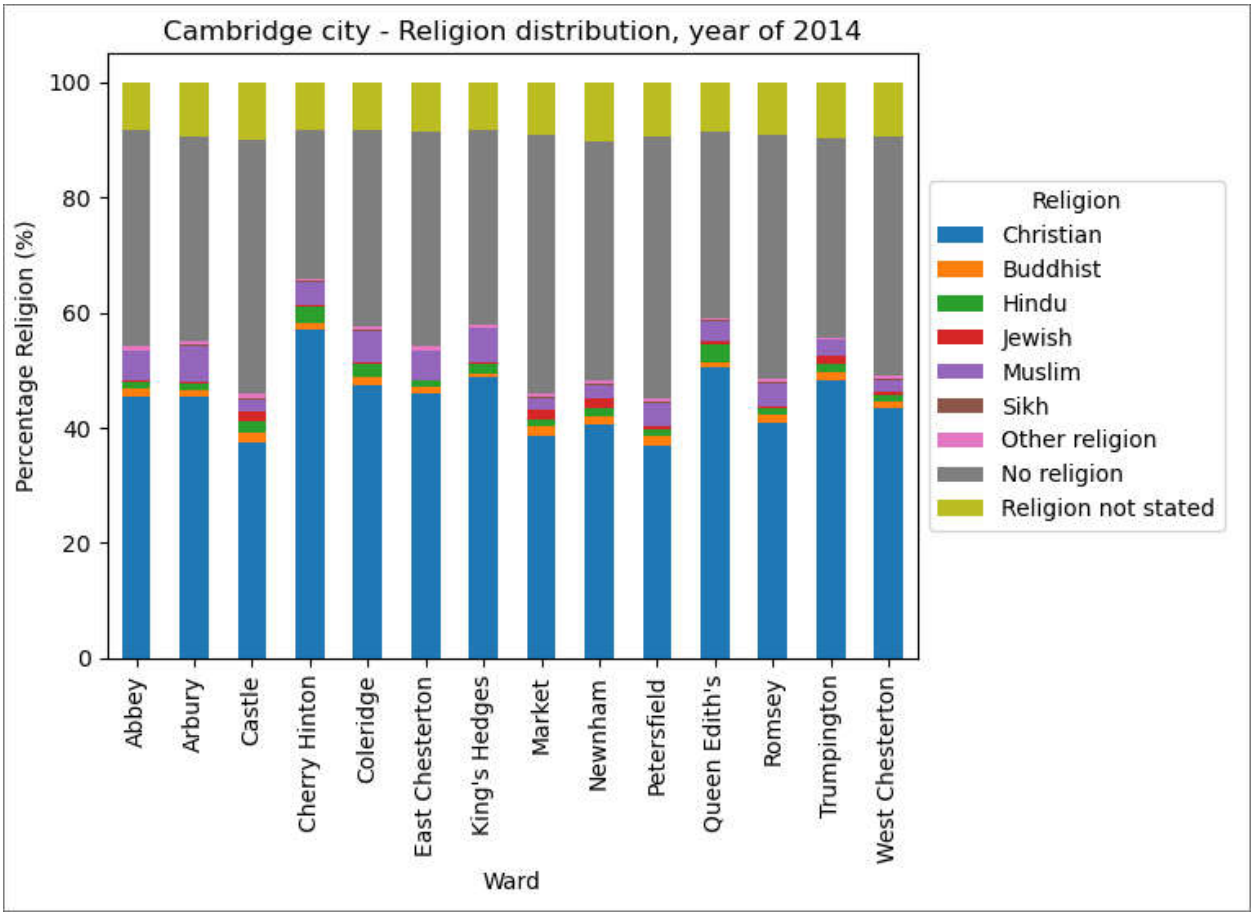
[27]:

	All Usual Residents Aged 16 and Over	No Qualifications	Level 1 Qualifications	Level 2 Qualifications	Apprenticeship	Level 3 Qualifications	Level 4 Qualifications and Above	Other Qualifications
count	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000
mean	7571.928571	900.428571	521.000000	654.214286	144.571429	1281.642857	3583.500000	486.571429
std	622.352048	538.918534	273.659811	200.421164	89.428712	978.740368	699.481869	129.475697
min	6688.000000	120.000000	104.000000	323.000000	13.000000	614.000000	2362.000000	278.000000
25%	7176.250000	457.250000	354.250000	506.000000	82.250000	690.250000	3054.500000	382.500000
50%	7555.500000	811.000000	477.000000	669.500000	142.500000	869.000000	3631.500000	501.000000
75%	7886.000000	1379.500000	718.750000	817.250000	221.000000	1080.750000	4060.500000	615.750000
max	9129.000000	1660.000000	956.000000	925.000000	290.000000	3836.000000	4663.000000	663.000000



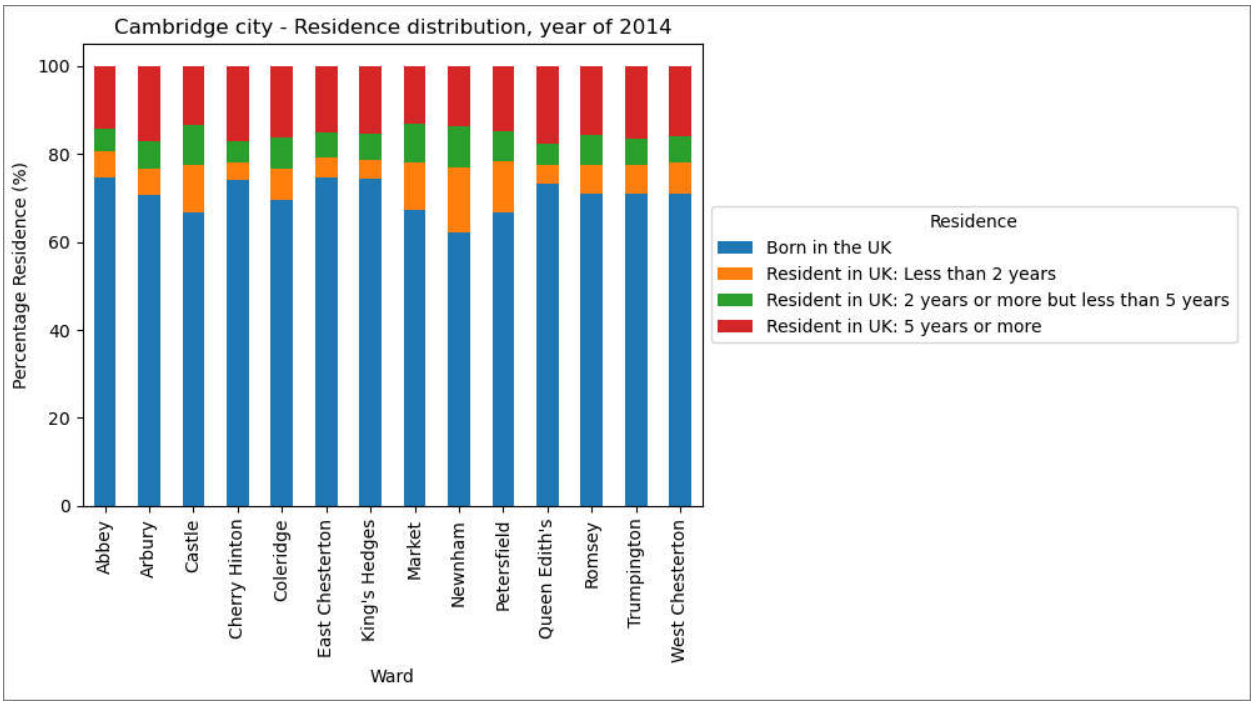
Observation: King's Hedges ward has the highest no qualification residences which may be better to avoid if you are looking for high academic place to live.

1.4.8. Religion



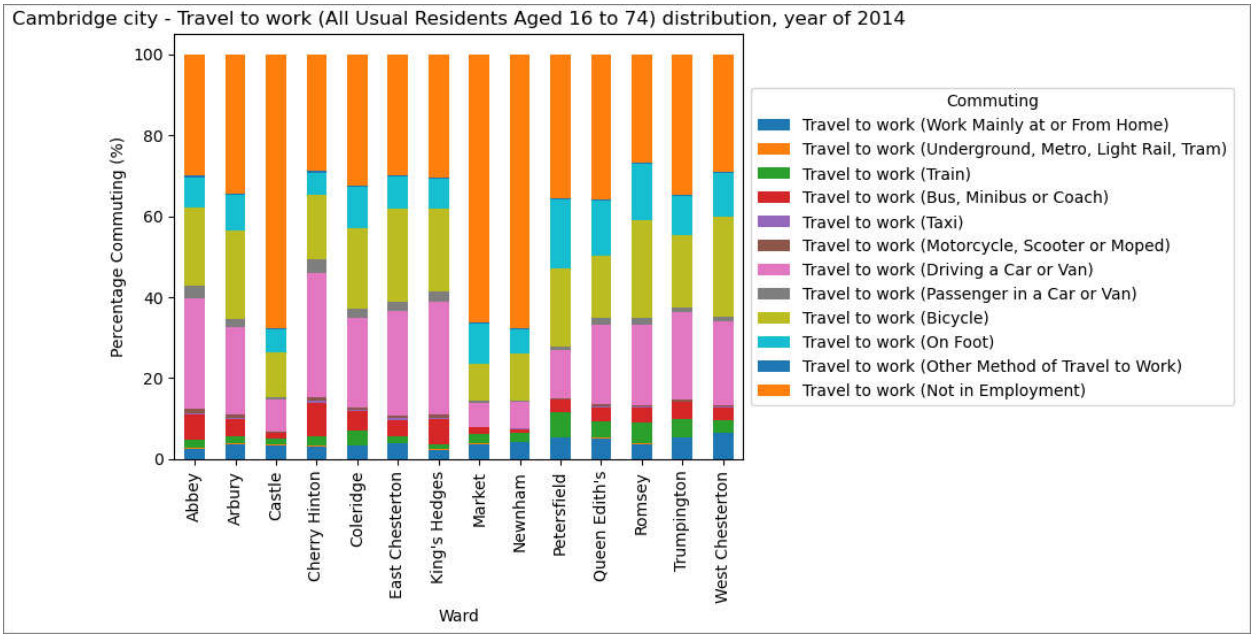
Observation: Cherry Hinton ward has the highest percentage of Christian residences.

1.4.9. Residence



Observation: Newnham ward has the highest number of residences that stays less than 2 years.

1.4.10. Travel to work

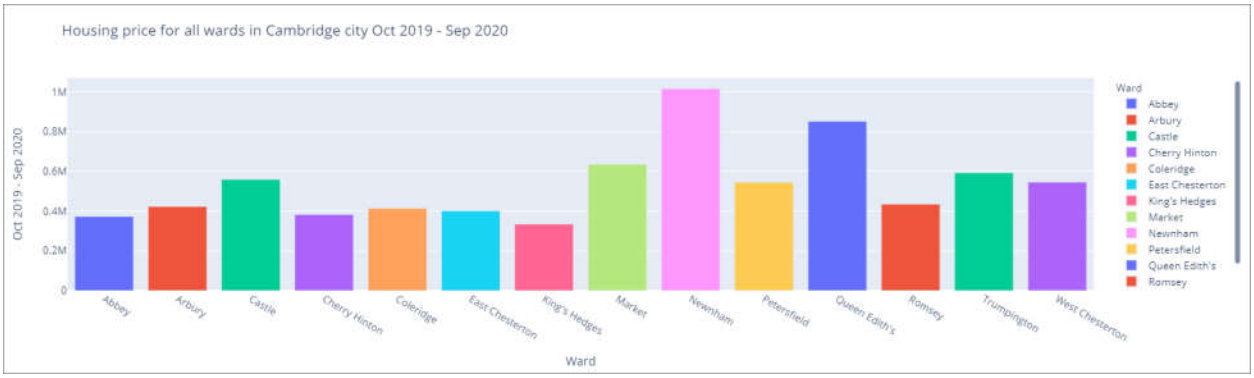


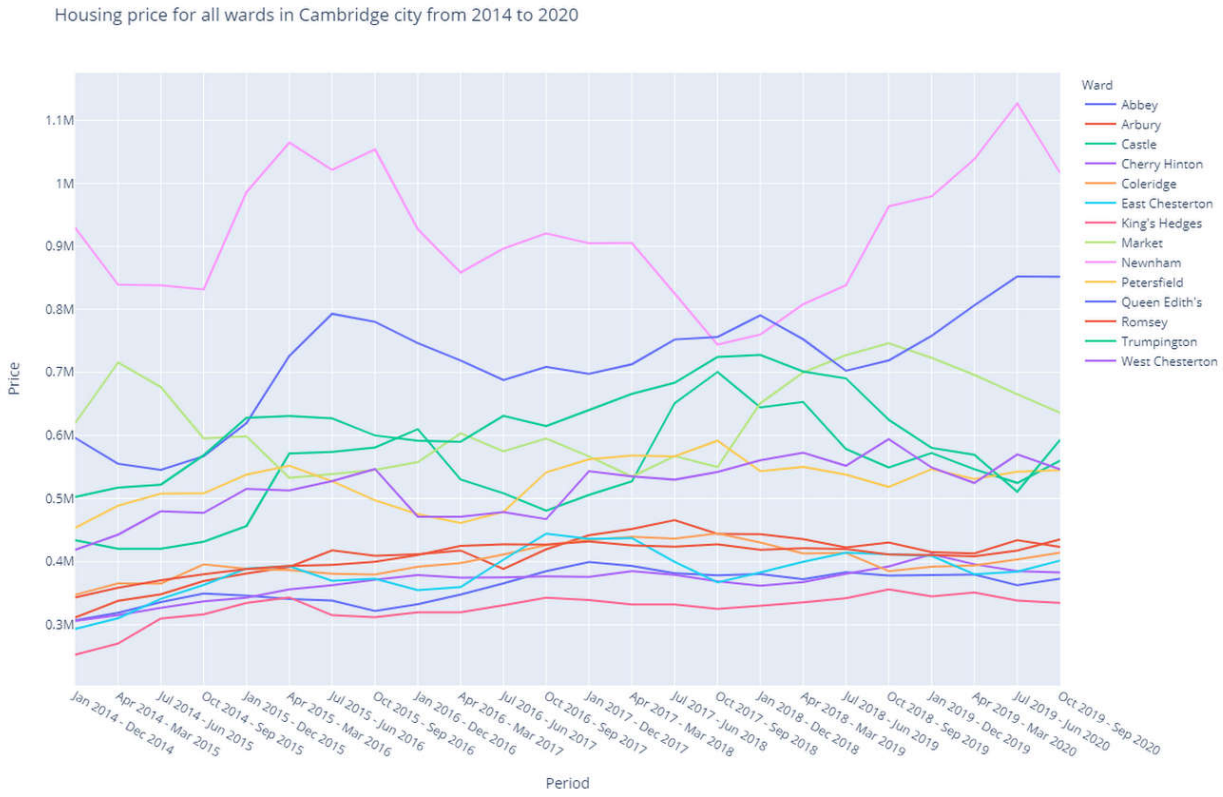
Observation: In Castle, Market and Newnham ward, large percentages are traveling to work but not in employment. Travel to work by bicycle is quite common in most of wards.

1.4.11 Crime rate

Observation: Its quite clear that the higher no qualification number will lead to higher unemployed residences, however the total crime rate is not following this trend.

1.4.12 Housing price



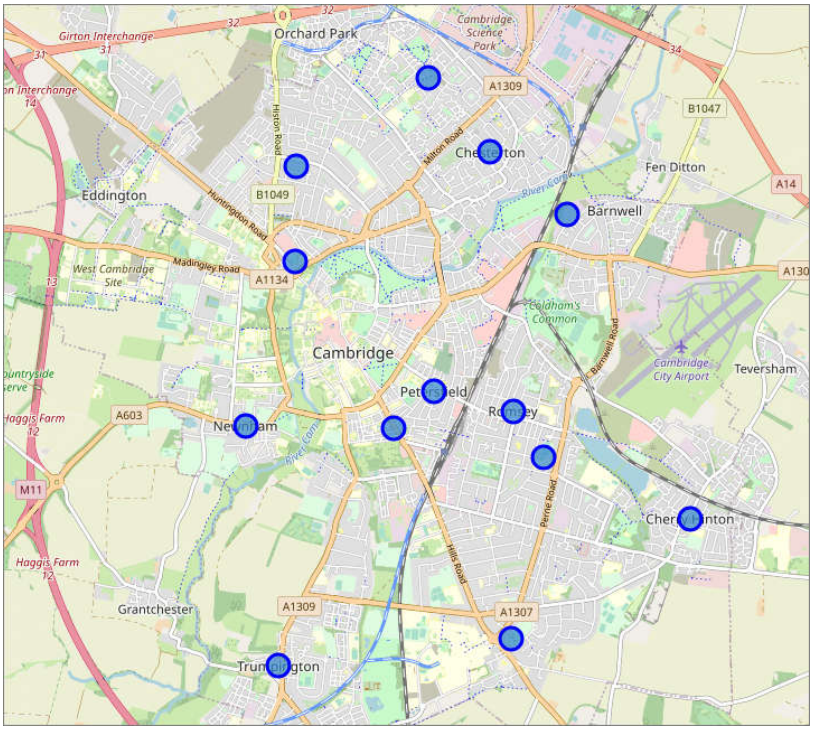


Observation: Newnham ward is the most expensive place to buy houses.

1.5. Mapping

In this section, I will use geopy to collect all coordinates (Latitude and Longitude) of each wards and show them in the map. Map will be plotted with folium library.

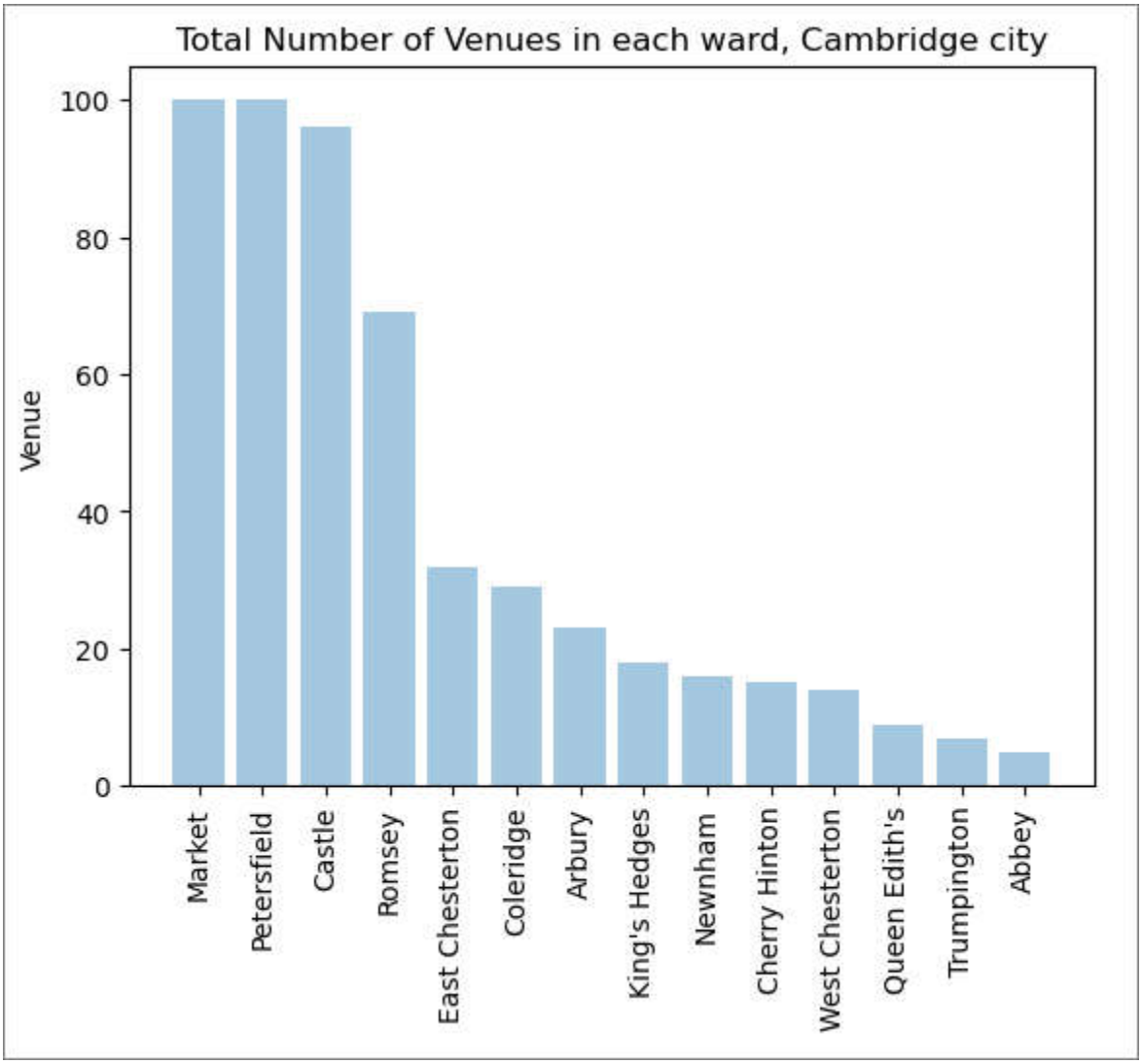
	Ward	Ward Latitude	Ward Longitude	Address
0	Abbey	52.3138	0.0502717	(The Abbey, Over Mereway, Fenlow Farm, Willing...
1	Arbury	52.221	0.114808	(Arbury Ward, Chesterton, Cambridge, Cambridge...
2	Castle	52.212	0.114699	(Castle Mound, Cambridge, Cambridgeshire, East...
3	Cherry Hinton	52.1878	0.175241	(Cherry Hinton, Cambridge, Cambridgeshire, Eas...
4	Coleridge	52.1937	0.152713	(Coleridge Community College, Radegund Road, R...
5	East Chesterton	52.2165	0.156338	(E, Beadle Industrial Estate, Barnwell, Cambri...
6	King's Hedges	52.2292	0.135074	(King's Hedges Ward, Chesterton, Cambridge, Ca...
7	Market	52.1964	0.129776	(HSBC UK, 62, Hills Road, Petersfield, Cambrid...
8	Newnham	52.1965	0.107044	(Newnham, Cambridge, Cambridgeshire, East of E...
9	Petersfield	52.1998	0.135933	(Petersfield, Cambridge, Cambridgeshire, East ...
10	Queen Edith's	52.1766	0.147808	(Queen Edith's Ward, Cambridge, Cambridgeshire...
11	Romsey	52.198	0.148062	(Romsey, Cambridge, Cambridgeshire, East of En...
12	Trumpington	52.174	0.112116	(Trumpington, Cambridge, Cambridgeshire, East ...
13	West Chesterton	52.2223	0.144592	(Chesterton, Cambridge, Cambridgeshire, East o...



1.6. Collect FourSquare Venues

In this section, I will collect top 100 venues in each Cambridge ward with radius of 1000m using the Foursquare API.

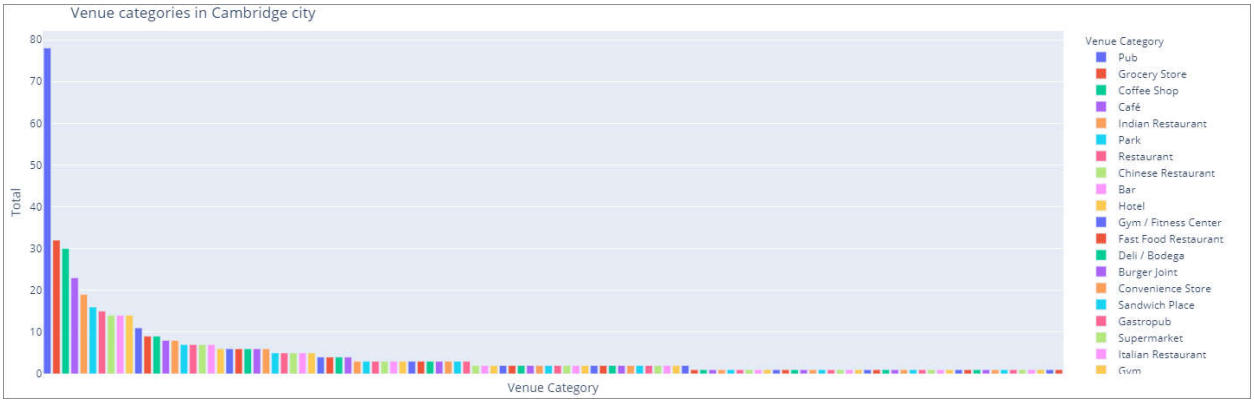
Summary of data mining for venues



Observation: We can observe that Market, Petersfield, Castle and Romsey has most number of venues. Abbey presents the least venue ward. The result reflects only within 1km radius of each selected address which presents for each ward. So we may need to clarify more specific address for further comparison.

1.7. Data Analyzing

In this section, I will analyze the venue categories using one hot encoding to convert venue categories to numerical formats for each ward.



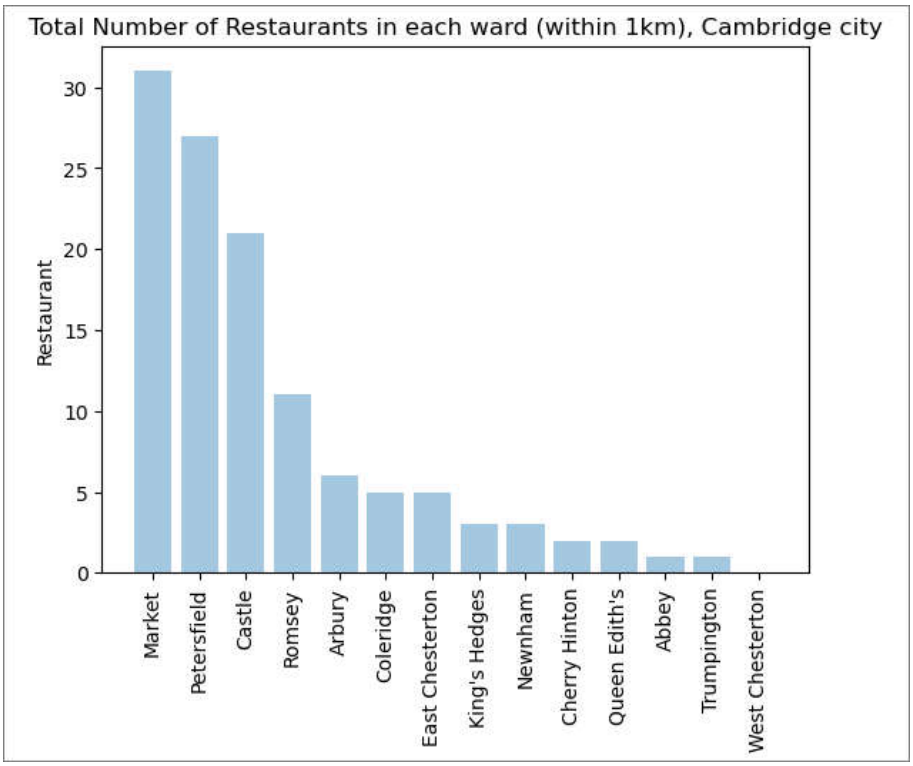
Observation: Its quite obvious that pub is the most popular venue in Cambridge, more than double the grocery store.

Now, lets look at the number of restaurants in each ward

```
[141]: #Select only restaurant
df_cambridge_restaurant = df_cambridge_venues_grouped_sum[df_cambridge_venues_grouped_sum["Venue Category"].str.contains("Restaurant")].reset_index(drop=True)
df_cambridge_restaurant
```

[141]:

	Venue Category	Total	Abbey	Arbury	Castle	Cherry Hinton	Coleridge	East Chesterton	King's Hedges	Market	Newnham	Petersfield	Queen Edith's	Romsey	Trumpington	West Chesterton
0	Indian Restaurant	19	1	3	3	1	0	1	0	3	0	4	0	3	0	0
1	Restaurant	15	0	0	4	1	1	0	1	4	1	3	0	0	0	0
2	Chinese Restaurant	14	0	1	1	0	0	0	1	4	0	6	0	1	0	0
3	Fast Food Restaurant	9	0	0	0	0	1	3	0	0	0	1	2	1	1	0
4	Italian Restaurant	7	0	0	2	0	1	0	0	2	0	1	0	1	0	0
5	Seafood Restaurant	6	0	0	0	0	1	0	0	2	1	1	0	1	0	0
6	Sushi Restaurant	6	0	0	3	0	1	0	0	1	0	0	0	1	0	0
7	African Restaurant	6	0	0	0	0	0	0	0	2	0	2	0	2	0	0
8	French Restaurant	5	0	0	1	0	0	0	0	2	0	2	0	0	0	0
9	Asian Restaurant	5	0	0	0	0	0	0	1	3	0	1	0	0	0	0
10	Thai Restaurant	5	0	0	0	0	0	1	0	2	1	1	0	0	0	0



Observation: Similarly with the trend of total venues, Market ward has the most number of restaurants, which will be quite convenient for anyone who is interested in eating outside regularly.

Create a new dataframe with top 10 venues for each ward

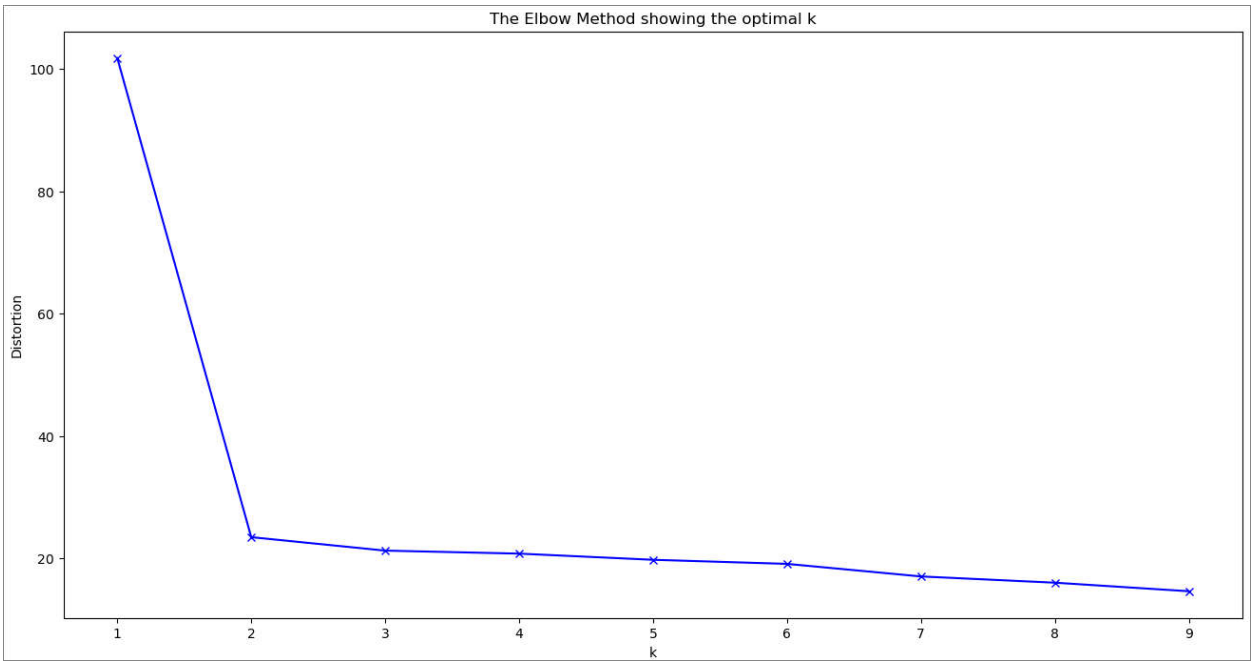
	Ward	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Abbey	Pub	Grocery Store	Indian Restaurant	Market	Pet Store	Field	Convenience Store	Creperie	Deli / Bodega	Department Store
1	Arbury	Pub	Indian Restaurant	Grocery Store	Convenience Store	Coffee Shop	Boat or Ferry	Field	Noodle House	Park	Pizza Place
2	Castle	Pub	Café	Park	Burger Joint	Coffee Shop	Restaurant	Grocery Store	Indian Restaurant	Sushi Restaurant	Wine Shop
3	Cherry Hinton	Park	Pharmacy	Plaza	Grocery Store	Gym / Fitness Center	Warehouse Store	Convenience Store	Indian Restaurant	Gastropub	Construction & Landscaping
4	Coleridge	Pub	Deli / Bodega	Café	Grocery Store	Hotel	Fish & Chips Shop	Italian Restaurant	Fast Food Restaurant	Seafood Restaurant	Park
5	East Chesterton	Fast Food Restaurant	Convenience Store	Coffee Shop	Bus Stop	Furniture / Home Store	Sporting Goods Shop	Soccer Stadium	Pharmacy	Pizza Place	Platform
6	King's Hedges	Grocery Store	Bed & Breakfast	Pub	Bus Station	Office	Convenience Store	Park	Restaurant	Chinese Restaurant	Coffee Shop
7	Market	Pub	Coffee Shop	Hotel	Bar	Café	Chinese Restaurant	Grocery Store	Restaurant	Asian Restaurant	Indian Restaurant
8	Newnham	Pub	Park	Grocery Store	Tennis Court	Restaurant	Rugby Stadium	Seafood Restaurant	Bakery	Soccer Field	Thai Restaurant
9	Petersfield	Pub	Café	Coffee Shop	Chinese Restaurant	Grocery Store	Hotel	Indian Restaurant	Bar	Bakery	Deli / Bodega
10	Queen Edith's	Coffee Shop	Fast Food Restaurant	Food Court	Bus Station	Soccer Field	Pub	Greek Restaurant	Gift Shop	Construction & Landscaping	Convenience Store
11	Romsey	Pub	Grocery Store	Coffee Shop	Café	Indian Restaurant	Bar	Sandwich Place	Deli / Bodega	African Restaurant	Hotel
12	Trumpington	Pub	Supermarket	Department Store	Bus Station	Fast Food Restaurant	Food & Drink Shop	Greek Restaurant	English Restaurant	Convenience Store	Creperie
13	West Chesterton	Bed & Breakfast	Grocery Store	Pub	Post Office	Coffee Shop	Playground	Platform	Park	Beer Garden	Pop-Up Shop

1.8. Apply Machine Learning

I use k-Means, an Unsupervised Machine Learning algorithm to group the data into k number of clusters.

Find optimal number of Clusters (Elbow Method)

For this method, the dataset is fit with the k-means model for a range of values (1-10). The distortions for each value of k is stored and then plotted on a line chart. The point of inflection is a good indication that the model fits best at that point.

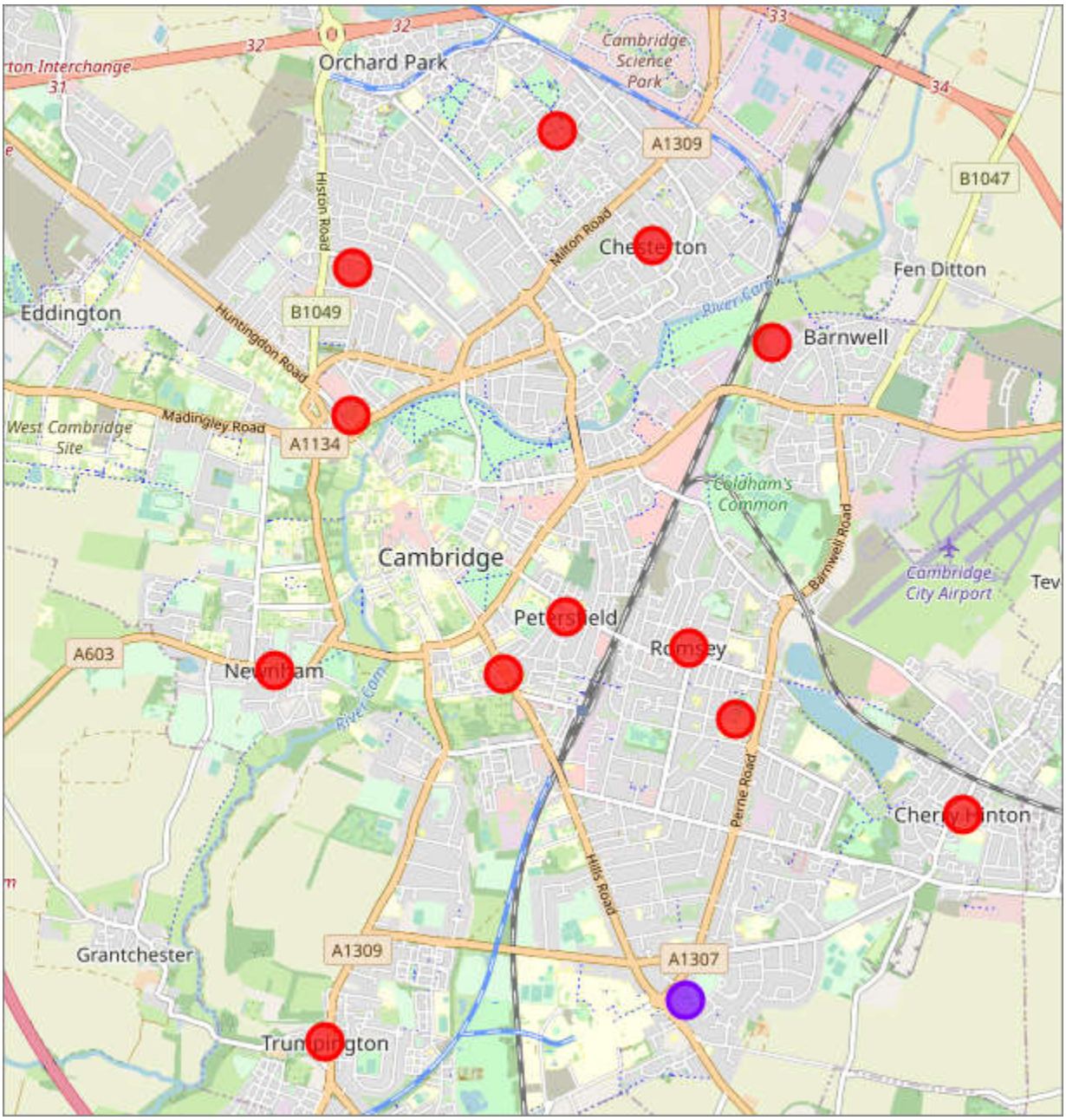


KneeLocator is used to compute the point of inflection especially when it is difficult to locate the point of inflection from the curve

```
[152]: from kneed import KneeLocator
       kl = KneeLocator(range(1, 10),
                        distortions,
                        curve="convex",
                        direction="decreasing")
       print('The optimum number of clusters is: ' + str(kl.elbow))

       The optimum number of clusters is: 2
```

Visualize the resulting clusters



Examine Clusters

```
[156]: # Cluster 0
df_cambridge_merged.loc[df_cambridge_merged['Cluster Label'] == 0, df_cambridge_merged.columns[[0] + list(range(4, df_cambridge_merged.shape[1]))]]
```

[156]:

	Ward	Cluster Label	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	Abbey	0	Pub	Grocery Store	Indian Restaurant	Market	Pet Store	Field	Convenience Store	Creperie	Deli / Bodega	Department Store
1	Arbury	0	Pub	Indian Restaurant	Grocery Store	Convenience Store	Coffee Shop	Boat or Ferry	Field	Noodle House	Park	Pizza Place
2	Castle	0	Pub	Café	Park	Burger Joint	Coffee Shop	Restaurant	Grocery Store	Indian Restaurant	Sushi Restaurant	Wine Shop
3	Cherry Hinton	0	Park	Pharmacy	Plaza	Grocery Store	Gym / Fitness Center	Warehouse Store	Convenience Store	Indian Restaurant	Gastropub	Construction & Landscaping
4	Coleridge	0	Pub	Deli / Bodega	Café	Grocery Store	Hotel	Fish & Chips Shop	Italian Restaurant	Fast Food Restaurant	Seafood Restaurant	Park
5	East Chesterton	0	Fast Food Restaurant	Convenience Store	Coffee Shop	Bus Stop	Furniture / Home Store	Sporting Goods Shop	Soccer Stadium	Pharmacy	Pizza Place	Platform
6	King's Hedges	0	Grocery Store	Bed & Breakfast	Pub	Bus Station	Office	Convenience Store	Park	Restaurant	Chinese Restaurant	Coffee Shop
7	Market	0	Pub	Coffee Shop	Hotel	Bar	Café	Chinese Restaurant	Grocery Store	Restaurant	Asian Restaurant	Indian Restaurant
8	Newnham	0	Pub	Park	Grocery Store	Tennis Court	Restaurant	Rugby Stadium	Seafood Restaurant	Bakery	Soccer Field	Thai Restaurant
9	Petersfield	0	Pub	Café	Coffee Shop	Chinese Restaurant	Grocery Store	Hotel	Indian Restaurant	Bar	Bakery	Deli / Bodega
11	Romsey	0	Pub	Grocery Store	Coffee Shop	Café	Indian Restaurant	Bar	Sandwich Place	Deli / Bodega	African Restaurant	Hotel
12	Trumpington	0	Pub	Supermarket	Department Store	Bus Station	Fast Food Restaurant	Food & Drink Shop	Greek Restaurant	English Restaurant	Convenience Store	Creperie
13	West Chesterton	0	Bed & Breakfast	Grocery Store	Pub	Post Office	Coffee Shop	Playground	Platform	Park	Beer Garden	Pop-Up Shop

```
[157]: # Cluster 1
df_cambridge_merged.loc[df_cambridge_merged['Cluster Label'] == 1, df_cambridge_merged.columns[[0] + list(range(4, df_cambridge_merged.shape[1]))]]
```

[157]:

	Ward	Cluster Label	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
10	Queen Edith's	1	Coffee Shop	Fast Food Restaurant	Food Court	Bus Station	Soccer Field	Pub	Greek Restaurant	Gift Shop	Construction & Landscaping	Convenience Store

Observation: For the clustering and ranking for the most common venue, most of the wards are equipped with pub. This is probably a traditional aspect in Cambridge as well as in UK.

2. Results

From all the data collected, below are few of insights I found:

1. Newnham town has the most diversity of residences.
2. Market ward has the largest rented privately/other tenure.
3. There is huge number of students in Castle, arket and Newnham wards. The highest unemployed percentage is at King's Hedges ward.
4. More other Asian ethnicity population in Cherry Hinton ward, where it may be better for other Asian communities.
5. Residences in Market ward have the best health condition.
6. Depending on the age of each child in the family, it may be beneficial to select the ward that has more similar age population. So that, the child or even adult can have more friends at similar age. Lets say for a child at 10 years old, its better to select Trumpington or Cherry Hinton ward to live.
7. King's Hedges ward has the highest no qualification residences which may be better to avoid if you are looking for high academic place to live.
8. Cherry Hinton ward has the highest percentage of Christian residences.
9. Newnham ward has the highest number of residences that stays less than 2 years.
10. In Castle, Market and Newnham ward, large percentages are traveling to work but not in employment. Travel to work by bicycle is quite common in most of wards.
11. Based on the descriptive statistics, the highest crime rate is for Anti Social Behaviour (ASB) and the 2nd highest is theft of pedal cycles in 2013-2014 fiscal year.
12. Its quite clear that the higher no qualification number will lead to higher unemployed residences, however the total crime rate is not following this trend.
13. Newnham ward is the most expensive place to buy houses.
14. Pub is the most popular venue in Cambridge, more than double the grocery store.

3. Discussion

The most consuming time for me in this project is to find the data I need and learn different ways to wrangle the data. It is really challenging at some points where there data is not in the required format. This process eventually allows me to practise python and googling faster.

4. Conclusion

Even though, I myself found quite many interesting insights from the data I collected, I still see that many gaps can be improved, which will require more time. Those are:

- More accurate ward coordinates (I did not want to manually input in my table) from **data.cambridgeshireinsight**, in geojson format. I was not able to extract the geojson data yet from full geojson data of UK (> 500Mb).
- Incorporate more data (wait on the soon release census survey in UK on 2021)
- Explore more with currently available census data