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

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.

Introduction

- **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.

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

Methodology

- **1.1. Python Libraries**

[1]:
```
# Import libraries
import numpy as np # Library to handle data in a vectorized manner

import pandas as pd # Library for data analysis
pd.set_option('display.max_columns', None) #used for setting the max number of columns in dataframe to display
pd.set_option('display.max_rows', None) #used for setting the max number of columns in dataframe to display

#!pip install bs4
#from bs4 import BeautifulSoup #to scrape information from web pages.

import json # Library to handle JSON files

import requests # Library to handle requests
from pandas.io.json import json_normalize # tranform JSON file into a pandas dataframe

#!conda install -c conda-forge geopandas
#!pip install geopandas
#import geopandas as gpd # to store geospatial data

import matplotlib.cm as cm # Matplotlib and associated plotting modules
import matplotlib.colors as colors
import matplotlib.pyplot as plt; plt.rcdefaults()
import plotly.express as px

#!conda install scikit-learn
#from sklearn.cluster import KMeans # import k-means from clustering stage

# To be installed in Mapping section
#!conda install -c conda-forge geopy --yes # geopy is a Python client for several popular geocoding web services
#!pip install geopy
#from geopy.geocoders import Nominatim # convert an address into latitude and Longitude values

#!conda install -c conda-forge folium=0.5.0 --yes #Folium is a Python library used for visualizing geospatial data.
#! pip install folium==0.5.0 #Folium is a Python library used for visualizing geospatial data, can select instead of conda
#import folium # map rendering library
#from folium import plugins # additional visualization

print('Libraries are imported!')
```

Methodology

- **1.2. Import Datasets**

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

#UK Census 2011 data.

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

Mean price paid for all house types

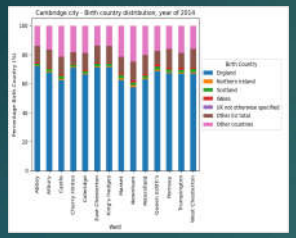
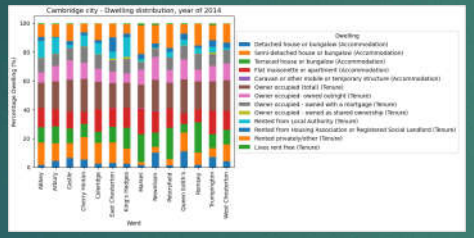
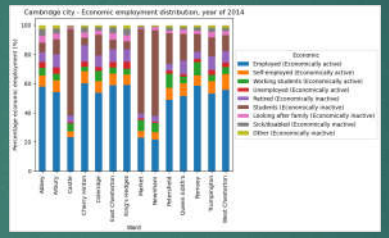
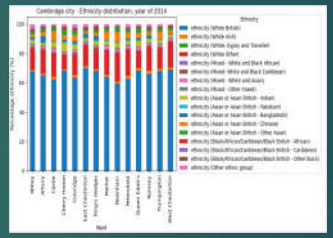
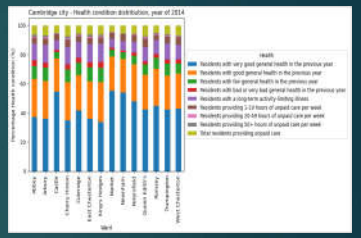
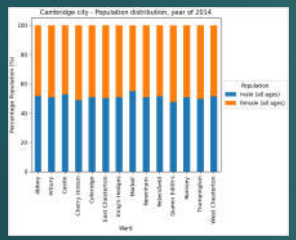
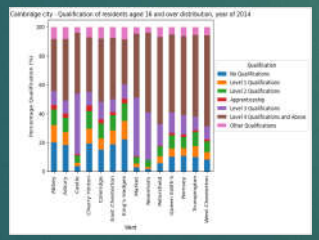
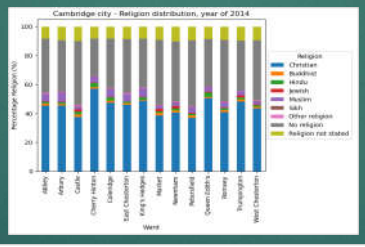
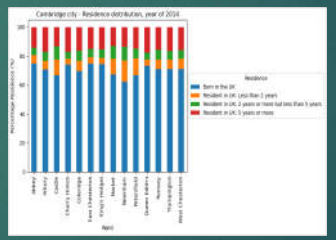
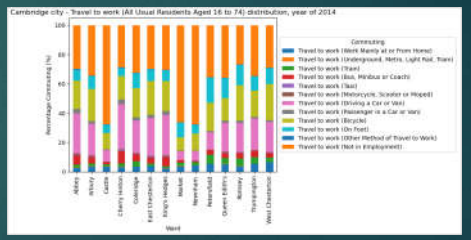
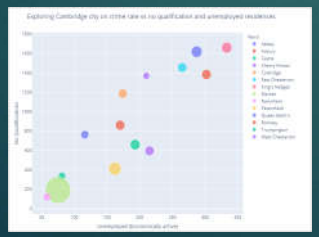
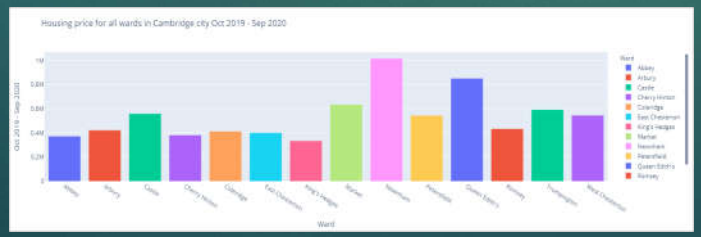
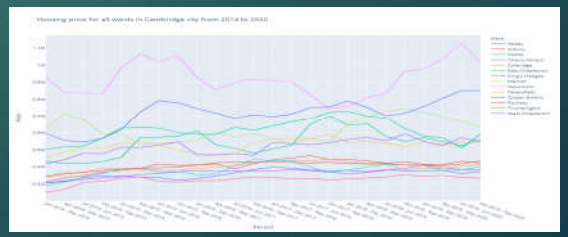
Methodology

- **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.

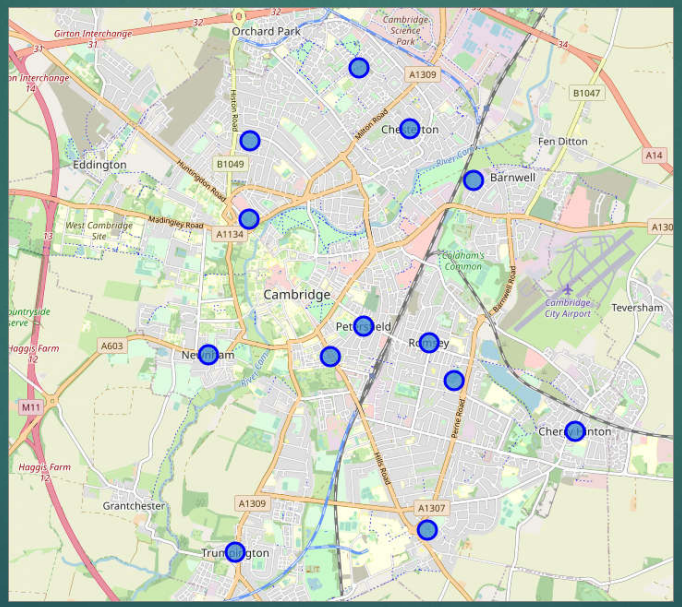
Methodology

- **1.4. Datasets Exploring**

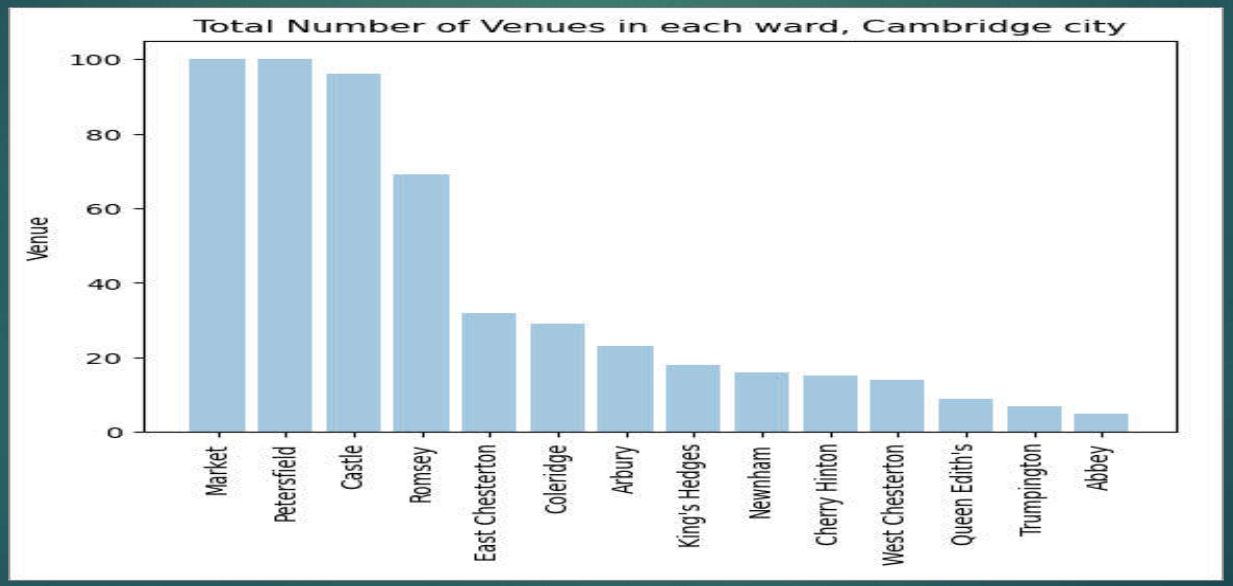
Methodology

- **1.5. Mapping**



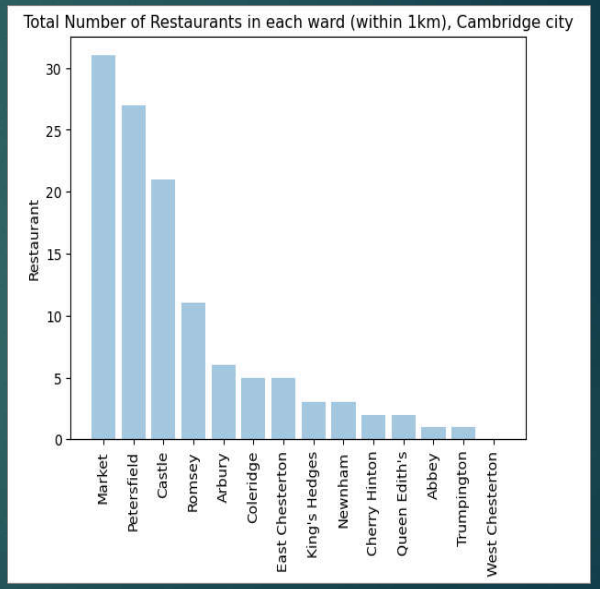
Methodology

- **1.6. Collect FourSquare Venues**



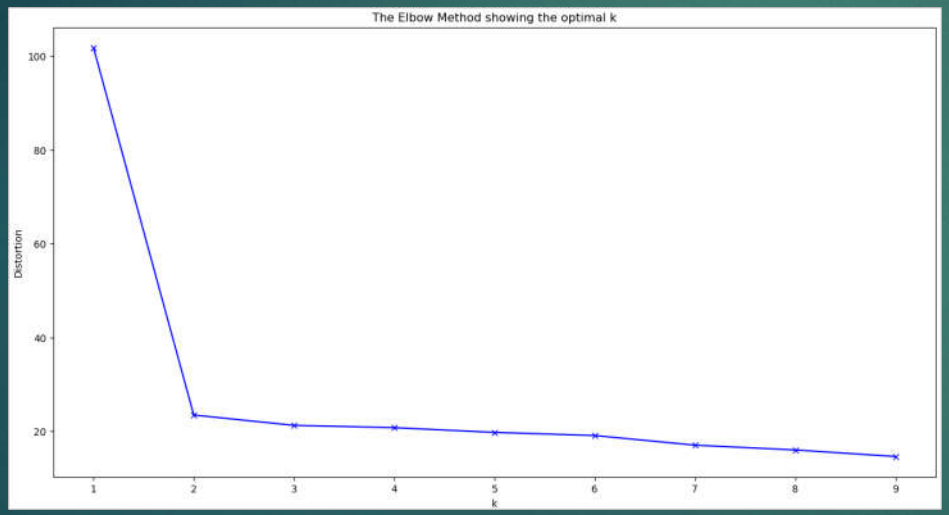
Methodology

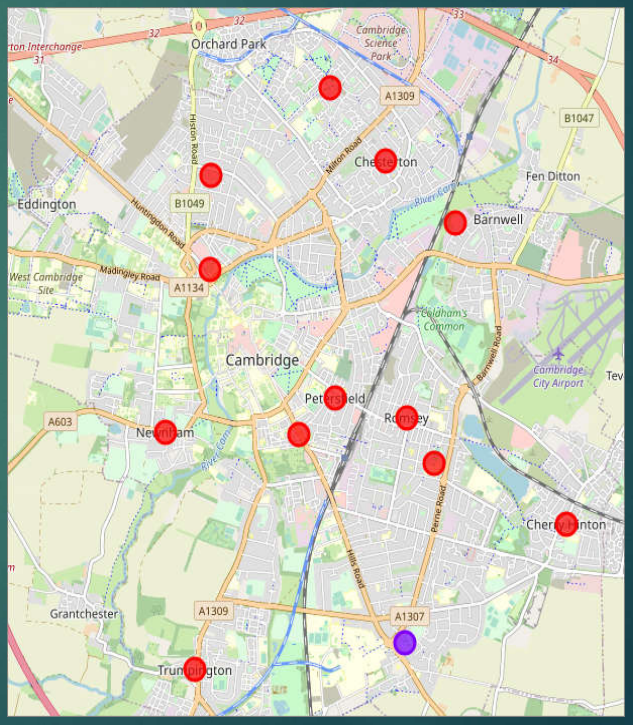
- **1.7. Data Analyzing**



Methodology

- **1.8. Apply Machine Learning**





2. Results

- Newnham town has the most diversity of residences.
- Market ward has the largest rented privately/other tenure.
- There is huge number of students in Castle, arket and Newnham wards. The highest unemployed percentage is at King's Hedges ward.
- More other Asian ethnicity population in Cherry Hinton ward, where it may be better for other Asian communities.
- Residences in Market ward have the best health condition.
- 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.
- 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.
- Cherry Hinton ward has the highest percentage of Christian residences.
- Newnham ward has the highest number of residences that stays less than 2 years.
- 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.
- 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.
- 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.
- Newnham ward is the most expensive place to buy houses.
- 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 these data are 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