COURSERA CAPSTONE PROJECT

BY MAX NETHERCOTT

INTRODUCTION

- This project will be aimed at identifying suitable areas in Bristol (UK) for home hunters to purchase a property.
- By utilising geospatial data and the fourSquared API this project will be able to identify and cluster neighbourhoods based on their surrounding amenities/venues.
- Kmeans machine learning algorithm

DATA ACQUISITION AND CLEANING

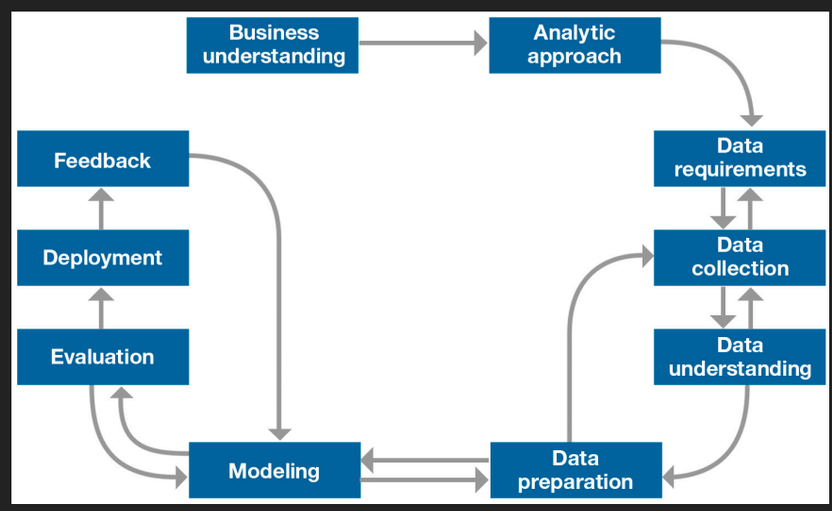
- Data will be collected from wikipedia to create a list of post codes and areas. (https://en.wikipedia.org/wiki/BS_postcode_area)
- FourSquared API will be used to capture data relating to the surrounding venues in the area.

	Neighborhood	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	Almondsbury, Bradley Stoke	Construction & Landscaping	Food Truck	Lake	Golf Course	Garden Center	Women's Store	Discount Store	Falafel Restaurant	English Restaurant	Electronics Store
1	Alveston, Rudgeway, Severn Beach, Pilning, Tho...	Fast Food Restaurant	Pub	Mobile Phone Shop	Asian Restaurant	Diner	Falafel Restaurant	English Restaurant	Electronics Store	Discount Store	Women's Store
2	Avonmouth, Shirehampton, Lawrence Weston	Historic Site	Pub	Gun Range	Convenience Store	Dessert Shop	Falafel Restaurant	English Restaurant	Electronics Store	Discount Store	Diner
3	Bedminster Down, Bishopsworth, Hartcliffe, Wit...	Pharmacy	Supermarket	Women's Store	Fish & Chips Shop	Construction & Landscaping	Convenience Store	Cosmetics Shop	Deli / Bodega	Department Store	Dessert Shop
4	Bedminster, Southville, Bower Ashton, part of ...	Coffee Shop	Pizza Place	Pub	Bar	Fast Food Restaurant	Pharmacy	Harbor / Marina	Café	Restaurant	BBQ Joint

Example of final enriched data set.

METHODOLOGY

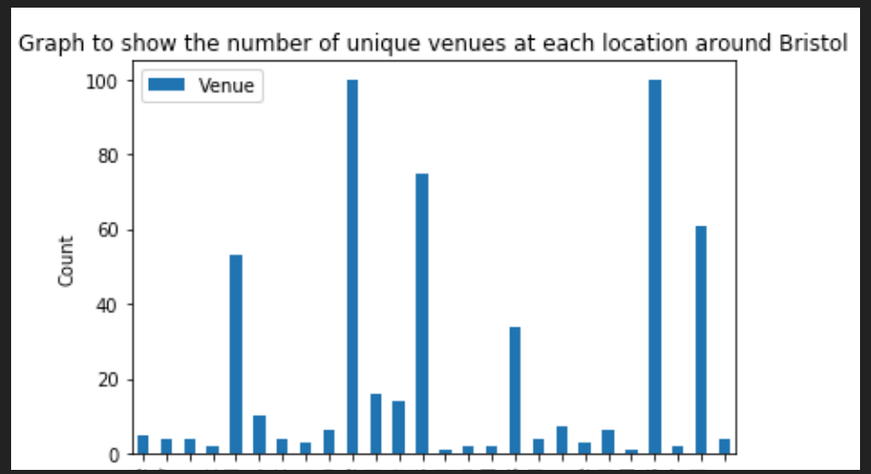
- Enrich raw data from Wikipedia with geolocation coordinates
- Utilise FourSquared API to get Venues and activities to do around coordinates.
- One hot encoding was used to convert categorical variables into a form that can be provided to machine learning algorithms.
- In this case Kmeans clustering was used to place neighbourhoods with similar venues/things to do in the same clusters.
- Visualise using:
  - Folium - Python tool for mapping
  - Seaborn - Python visualisation

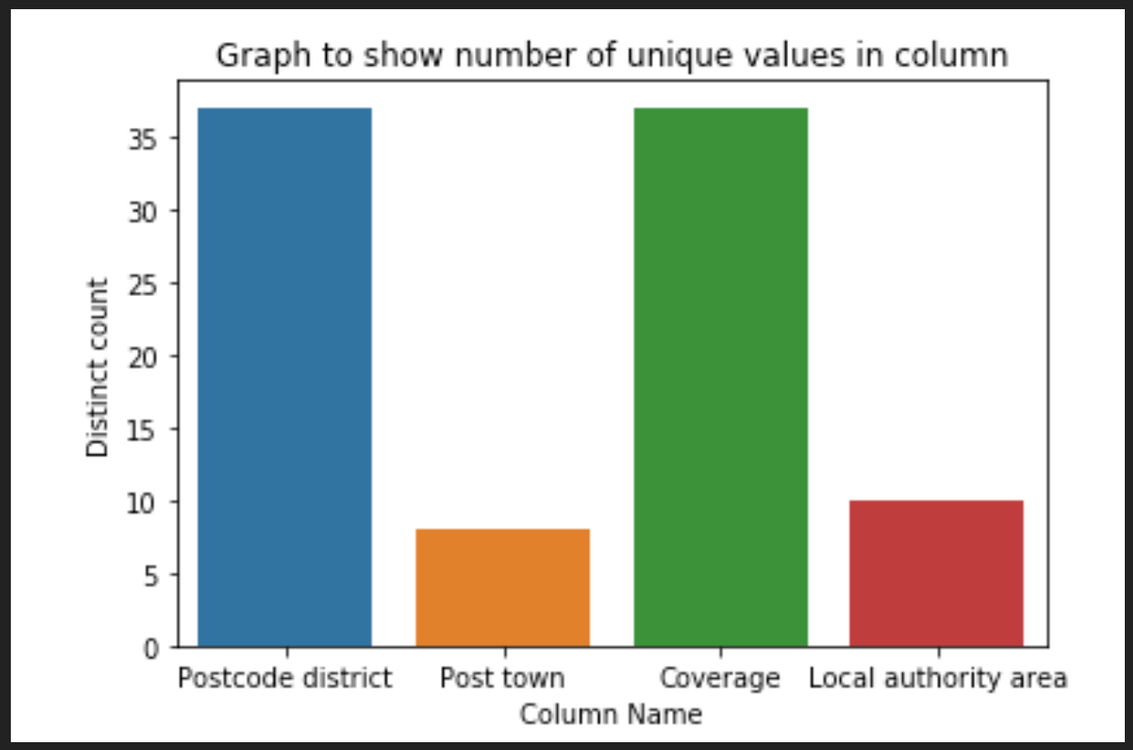


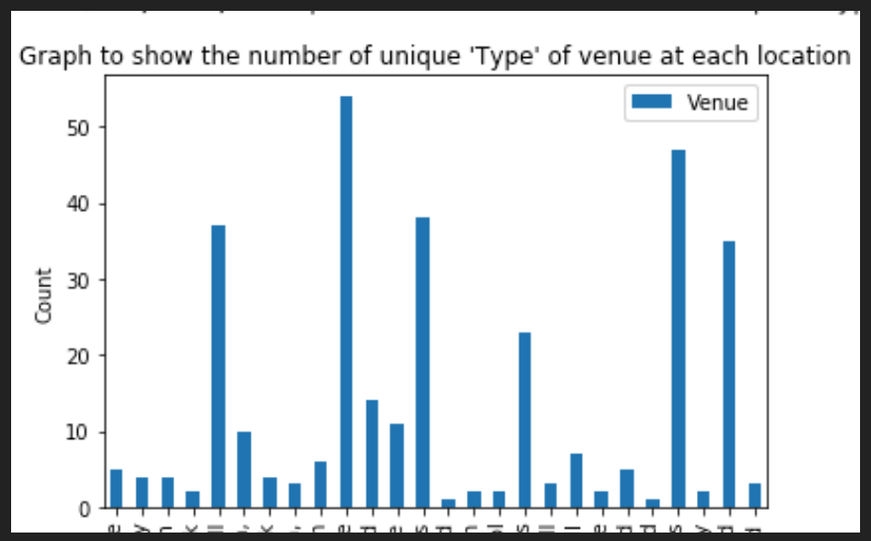
Following IBM Data Science lifecycle

ANALYSIS

- Analysis to show the number of unique values per column.
- Distribution of unique venues per neighbourhood and unique types of venues per neighbourhood.

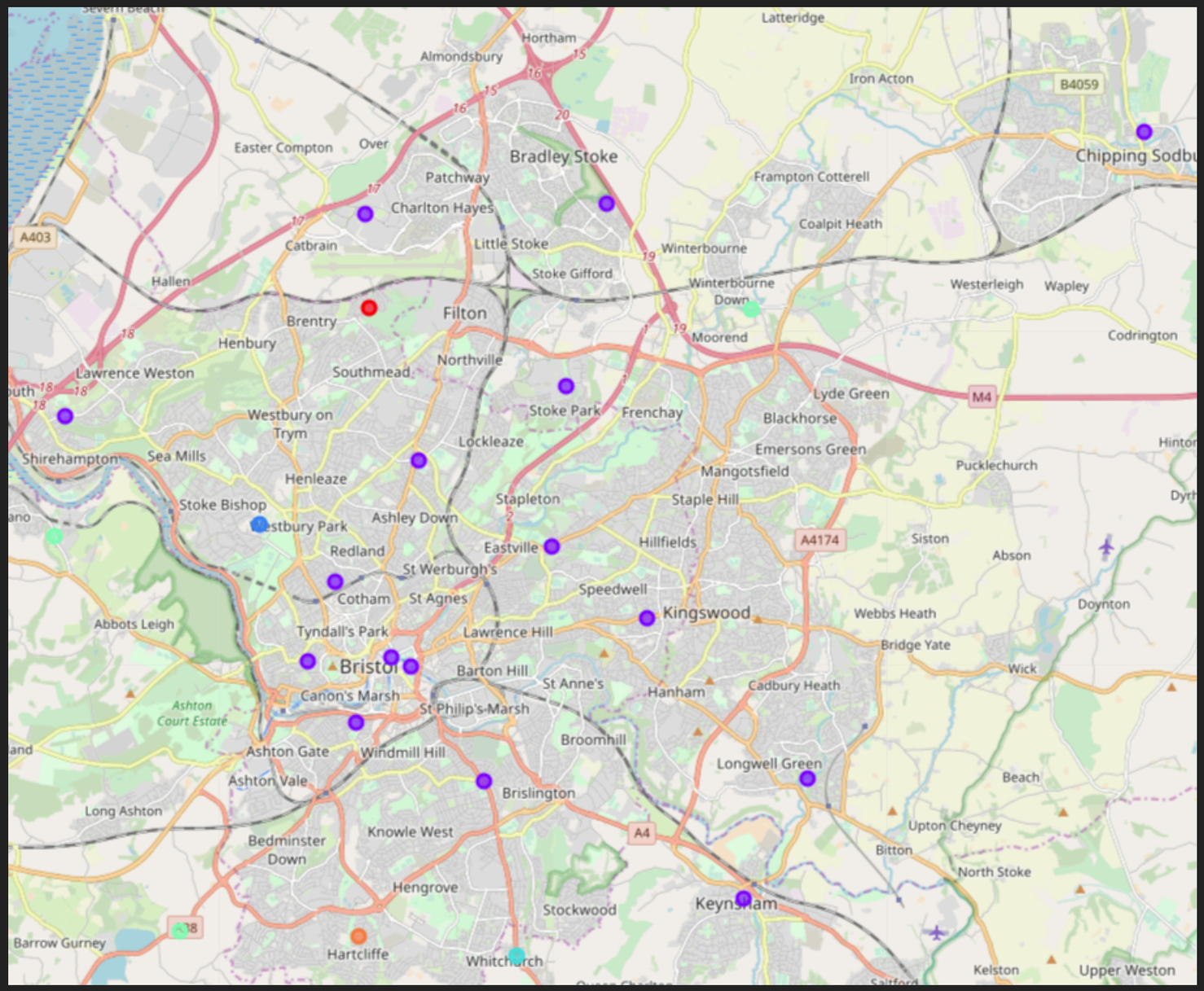






RESULTS

- Neighbourhoods around Bristol have similar venues/things to do
- Cluster 2 (Shown in purple on the map) - has a good mixture of pubs, bars and restaurants.
- Cluster 3 (Shown in Blue) has a nice park and Botenical Gardens in the vicinity.
- Maybe this area would be more suited to the elderly or someone looking for a more tranquil area to live.
- To improve the results more data would need to be incorporated.
- Data relating to transport, work places and crime rates could be included to deepen the understanding of each neighbourhood.



CONCLUSION

- It is possible to use geospatial data, relating to bristol and its neighbourhoods, combined with the use of fourSquared API to cluster potential living areas based on surrounding amenities/venues.
- The results could be used to advise home hunters on which neighbourhood would be best suited to them.
- The process could be improved by incorporating a more varied dataset but initial results show good potential.