Capstone Project – Detailed Report

Coursera – IBM Certificate in Data Science Project Submission v

July 2020

1. Introduction

The IBM Certificate in Data Science course on Coursera culminates in a Capstone Project, which has a defined outline to the requirements for qualifying and passing the final course assessment.

The parameters of the Capstone project guide that users should apply the knowledge obtained during the course to a real life business opportunity that would benefit from the techniques found in the course. Specifically for this one, the requirement is to apply the K-Clustering approach to the defined business opportunity.

The other key parameter is that the submission must use the FourSquare API to obtain venue data, which is to be used in the analysis.

This presentation and its accompanying Jupyter notebook (which can be found at the link below) outline such a business opportunity and the associated analysis to address this need.

2. The Business Opportunity

This business issue is a hypothetical one as I don't have a real case that fulfils the requirements for the Capstone Project i.e. use FourSquare API and data. This case involves a business with investments in the Food & Beverage / Hospitality sector looking to expand into the Richmond-upon-Thames administration area. The Richmond-upon-Thames administration area lies within the boundaries of the Greater London are and consists of 19 main neighbourhoods. This includes famous landmarks such as the Twickenham Rugby Stadium, which is considered the home of rubgy

The business is looking to enter this area by opening another outlet. The business runs multiple arms in the business, which includes coffee shops, fast food joints, restaurants and wine bars. As a first step in opening an outlet, the business would like to get an initial understanding of the neighbourhoods and the clustering of food related companies operating in these regions. This will then feed into further analysis necessary for decision making including socio-economic distributions, footfall patterns, food venues opening/closing data, market gaps.

Assumptions, business logic: The assumption behind the analysis is that we can use unsupervised machine learning to create clusters of neighbourhoods that will provide us a list of areas for consideration for the restaurant. The intent is the opening will be ideally located in a potentially underserved neighbourhood or one with a gap in offerings (e.g. lack of a high-end burger place).

This notebook addresses this first step in the analysis.

3. The Data, Scope, Approach, Assumptions and Limitations

The following are the key data sources to be used.

- List of the neighbourhoods in the Richmond admin area -> This will be scraped from the website www.doogal.co.uk .. This provides some key location information. I have used the Postcode, Ward (which I have renamed to Neighborhood), Latitude and Longitude

- Top venues of neighbourhoods -> Foursquare API is used to collect the venue data such as Name, Category, geographic location details such as Latitude and Longitude

- Co-ordinate data is inherent in the data coming from www.doogal.co.uk

Scope, Approach, Assumptions and Limitations

- Scope- we will work through all 19 neighbourhoods of the admin area. In doing this, though there is postcode data that provides greater granularity, for brevity and focus, we will look at the neighbourhood names and group all the subsequent processing on this attribute. Therefore analysis is performed at the neighbourhood level.

- Approach: After tidying up the data, we will apply K-means machine learning technique for creating clusters of the neighbourhoods. We will use silhouette score for choosing the optimal number of clusters.

- Assumption: The data obtained from both data sources is considered complete and accurate. No separate verification process has been performed.

- Limitations: Due to multiple considerations, the data gathering from the FourSquare API is limited to 500m around the co-ordinates being used for each of the neighbourhoods. An initial limitation on the venues retrieved was set to 50. However, the results returned were limited so it was decided to increase this number to 100.

4. The Analysis

1.1. Data Cleansing

As a first step, the data was scraped from the website (www.doogal.co.uk) and examined. This highlighted duplications of the neighbourhoods based primarily on the postcodes. For the purposes of this assessment, this would introduce a very large set of datasets to assess. So I made a decision to limit the number of neighbourhoods on the names rather than unique postcodes. This produced the following results:-

```
(7847, 4)
```

Out[3]:

	PCode	Area	Lat	Long
0	KT1	Hampton Wick	51.412355	-0.31185
1	KT1	Hampton Wick	51.409578	-0.310033
2	KT1	Hampton Wick	51.404226	-0.309936
3	KT1	Hampton Wick	51.399714	-0.314472
4	KT1	Hampton Wick	51.410868	-0.312193

```
(48, 4)
```

4]:

	PCode	Area	Lat	Long
0	KT1	Hampton Wick	51.412355	-0.31185
18	KT1	Teddington	51.412046	-0.31724
118	KT2	Ham, Petersham and Richmond Riverside	51.427081	-0.293606
132	KT8	Hampton	51.406791	-0.348605
152	SW13	Mortlake and Barnes Common	51.471836	-0.248994

The box on the left shows the Shape of the data before cleaning, whilst the box on the right shows the resultant cleaned data frame.

NB. In a real life business case, I would have chosen a different approach in streamlining this data without losing the granularity that is offered by the initial dataset. However, this approach would be governed by the business needs.

4. Analysis (Cont'd)

1.2 Obtaining Venue Information from FourSquare

The next step after the data cleansing is to obtain venue information from the FourSquare API. In this step, we use the selected neighbourhoods and use their geographic information (latitude and longitude) to base as the centre of the query whilst limiting the search to 500m around this point. We capture 100 venues at the maximum.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	\
0	Hampton Wick	51.412355	-0.31185	
1	Hampton Wick	51.412355	-0.31185	
2	Hampton Wick	51.412355	-0.31185	
3	Hampton Wick	51.412355	-0.31185	
4	Hampton Wick	51.412355	-0.31185	

	Venue	Venue Latitude	Venue Longitude	Venue Category
0	John Lewis & Partners	51.411538	-0.306803	Department Store
1	Pomegranate Bistro	51.412219	-0.310923	Italian Restaurant
2	Côte Brasserie	51.410510	-0.308183	French Restaurant
3	The Foresters Arms	51.413021	-0.311352	Pub
4	Local Hero	51.410753	-0.306694	Café

The above table shows the first 5 rows of the resultant data frame that contains the neighbourhood details and the venue details (name, category, latitude, longitude). This will be used for the next stages of the analysis.

4. Analysis (Cont'd)

1.3 Prepare the Data for K-Clustering.

The first stage in this process is to perform the One-Hot encoding. The picture below shows samples of the output from this stage.

```
   Neighborhood  American Restaurant  Argentinian Restaurant  Art Gallery  \
0  Hampton Wick                    0                       0            0
1  Hampton Wick                    0                       0            0
2  Hampton Wick                    0                       0            0
3  Hampton Wick                    0                       0            0
4  Hampton Wick                    0                       0            0

   Arts & Crafts Store  Asian Restaurant  Athletics & Sports  Auto Garage  \
0                    0                 0                   0            0
1                    0                 0                   0            0
2                    0                 0                   0            0
3                    0                 0                   0            0
4                    0                 0                   0            0

   BBQ Joint  Bakery  ...  Tapas Restaurant  Tea Room  Tennis Court  \
0          0       0  ...                 0         0             0
1          0       0  ...                 0         0             0
2          0       0  ...                 0         0             0
3          0       0  ...                 0         0             0
4          0       0  ...                 0         0             0
```

	Neighborhood	American Restaurant	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Garage	BBQ Joint	Bakery	...	Tapas Restaurant	Tea Room	Tennis Court	Thai Restaurant	Theater	Track	Trail	Train Station
0	Barnes	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.038462	...	0.000000	0.000000	0.000000	0.038462	0.000000	0.038462	0.000000	0.000000
1	East Sheen	0.000000	0.000000	0.000000	0.045455	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.045455	0.000000	0.000000	0.000000	0.000000	0.000000
2	Fulwell and Hampton Hill	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.043478	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	Ham, Petersham and Richmond Riverside	0.000000	0.000000	0.016393	0.000000	0.000000	0.016393	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.016393	0.000000	0.016393	0.000000
4	Hampton	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.071429
5	Hampton North	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

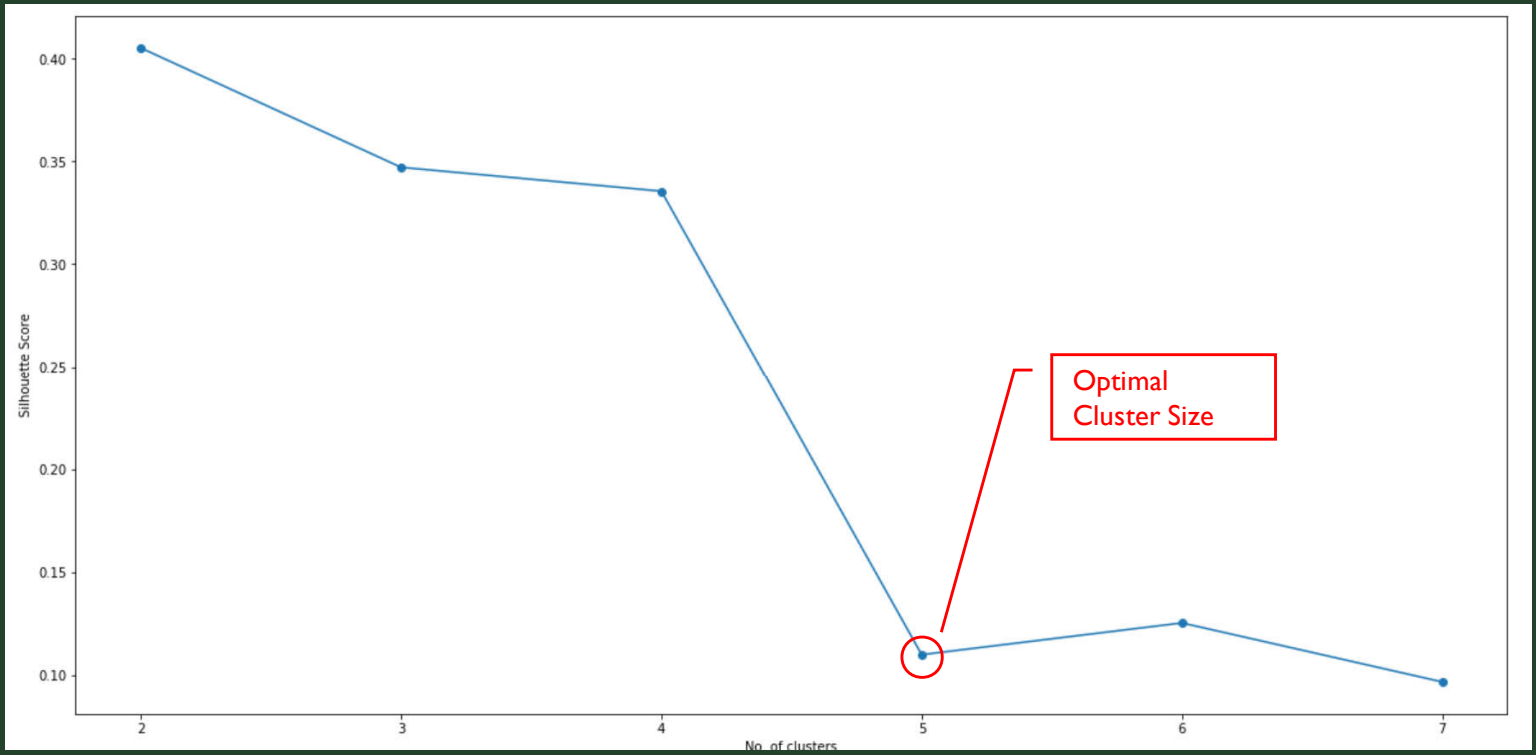
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Barnes	Food & Drink Shop	Café	Pub	Farmers Market	Park	Lake	Italian Restaurant	Community Center	Movie Theater	Coffee Shop
1	East Sheen	Coffee Shop	Pub	Pizza Place	Stationery Store	Pharmacy	Creperie	Plaza	Dance Studio	Chinese Restaurant	Middle Eastern Restaurant
2	Fulwell and Hampton Hill	Pub	Pizza Place	Convenience Store	Café	Fish & Chips Shop	Seafood Restaurant	Chinese Restaurant	Diner	Garden Center	Fast Food Restaurant
3	Ham, Petersham and Richmond Riverside	Pub	Coffee Shop	Italian Restaurant	Pharmacy	Indian Restaurant	Park	Grocery Store	Café	Bus Station	Bus Stop
4	Hampton	Park	Grocery Store	Café	Waterfront	Beer Garden	Pizza Place	Pub	Canal Lock	Seafood Restaurant	Soccer Stadium

The data frame is now ready for the clustering analysis. However, before we do the clustering, we will run the Silhouette Scoring algorithms to identify the optimal cluster size.

4. Analysis (Cont'd)

1.4 Silhouette Scoring

The second stage is to run the Silhouette Scoring on the data to understand and confirm the ideal cluster size. The following is the resultant output from this step.



This shows that the optimal cluster size would be 5. This is what will be used to complete the analysis .

4. Analysis (Cont'd)

1.5 Clustering and Analysing Output.

After running the clustering algorithms the following table shows the final data frame which I will be analysing.

	Neighborhood	Lat	Long	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Hampton Wick	51.412355	-0.31185	1	Pub	Park	Coffee Shop	Clothing Store	Department Store	Hotel	Thai Restaurant	Restaurant	Café	Plaza
18	Teddington	51.412046	-0.31724	1	Pub	Coffee Shop	Hotel	Italian Restaurant	Park	Café	Indian Restaurant	Mediterranean Restaurant	Grocery Store	Train Station
118	Ham, Petersham and Richmond Riverside	51.427081	-0.293606	1	Pub	Coffee Shop	Italian Restaurant	Pharmacy	Indian Restaurant	Park	Grocery Store	Café	Bus Station	Bus Stop
132	Hampton	51.406791	-0.348605	0	Park	Grocery Store	Café	Waterfront	Beer Garden	Pizza Place	Pub	Canal Lock	Seafood Restaurant	Soccer Stadium
152	Mortlake and Barnes Common	51.471836	-0.248994	0	Pub	Park	Grocery Store	Coffee Shop	Farmers Market	Gastropub	Café	Gym / Fitness Center	Platform	Pizza Place

I now start looking at each cluster starting with Cluster 1

	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
132	Hampton	Park	Grocery Store	Café	Waterfront	Beer Garden	Pizza Place	Pub	Canal Lock	Seafood Restaurant	Soccer Stadium
152	Mortlake and Barnes Common	Pub	Park	Grocery Store	Coffee Shop	Farmers Market	Gastropub	Café	Gym / Fitness Center	Platform	Pizza Place
178	Barnes	Food & Drink Shop	Café	Pub	Farmers Market	Park	Lake	Italian Restaurant	Community Center	Movie Theater	Coffee Shop
6119	Whitton	Rugby Stadium	Park	Convenience Store	Construction & Landscaping	Hotel	Sporting Goods Shop	Hotel Bar	Middle Eastern Restaurant	Museum	Gym / Fitness Center

4. Analysis (Cont'd)

Cluster 2

	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	Hampton Wick	Pub	Park	Coffee Shop	Clothing Store	Department Store	Hotel	Thai Restaurant	Restaurant	Café	Plaza
18	Teddington	Pub	Coffee Shop	Hotel	Italian Restaurant	Park	Café	Indian Restaurant	Mediterranean Restaurant	Grocery Store	Train Station
118	Ham, Petersham and Richmond Riverside	Pub	Coffee Shop	Italian Restaurant	Pharmacy	Indian Restaurant	Park	Grocery Store	Café	Bus Station	Bus Stop
779	East Sheen	Coffee Shop	Pub	Pizza Place	Stationery Store	Pharmacy	Creperie	Plaza	Dance Studio	Chinese Restaurant	Middle Eastern Restaurant
799	North Richmond	Coffee Shop	Supermarket	Pub	Food Truck	Seafood Restaurant	Middle Eastern Restaurant	Restaurant	Chinese Restaurant	Stationery Store	Plaza
812	Kew	Garden	Pub	Park	Botanical Garden	Restaurant	Hotel	Gym / Fitness Center	Art Gallery	Pier	Food Stand
1441	St Margarets and North Twickenham	Pub	Coffee Shop	Italian Restaurant	Indian Restaurant	Bus Stop	Pharmacy	Grocery Store	Vietnamese Restaurant	Japanese Restaurant	Farmers Market
2072	West Twickenham	Pub	Bus Stop	Supermarket	Pharmacy	Restaurant	Hardware Store	Health & Beauty Service	Italian Restaurant	Thai Restaurant	Fish & Chips Shop
2204	South Twickenham	Pub	Coffee Shop	Italian Restaurant	Pizza Place	Indian Restaurant	Grocery Store	Café	Supermarket	Convenience Store	Fast Food Restaurant
2762	South Richmond	Pub	Café	Italian Restaurant	Coffee Shop	Bakery	Bar	Restaurant	Train Station	Grocery Store	Burger Joint
3627	Fulwell and Hampton Hill	Pub	Pizza Place	Convenience Store	Café	Fish & Chips Shop	Seafood Restaurant	Chinese Restaurant	Diner	Garden Center	Fast Food Restaurant

4. Analysis (Cont'd)

Cluster 3

	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
5857	Heathfield	Playground	Soccer Field	Convenience Store	Park	Café	Chinese Restaurant	Pub	Grocery Store	Furniture / Home Store	Cricket Ground

Cluster 4

	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
1618	Twickenham Riverside	Thai Restaurant	Café	Pier	Italian Restaurant	Indian Restaurant	Liquor Store	Dance Studio	Cycle Studio	Deli / Bodega	Fish Market

Cluster 4

	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
4582	Hampton North	Jewelry Store	Gym / Fitness Center	Coffee Shop	Supermarket	Department Store	Fast Food Restaurant	Farmers Market	English Restaurant	Electronics Store	Diner

5. Conclusion

At the onset, the question that was framed was to identify some neighbourhoods for further investigation and research. The analysis will further be developed on two branches.

Branch 1 - Underserved Locations. This branch will look at neighbourhoods that, based on this analysis, appear to be underserved by a variety of food outlets be these coffee shops or proper restaurants. In this category, the business may look at the neighbourhoods in Cluster 1 or Cluster 3. To elaborate with examples:-

In Cluster 1, Whitton looks promising. It has a nearby Rugby stadium, a park, a hotel and a museum which suggests footfall. Yet it only has 1 restaurant. This presents an underserved area.

In Cluste3 , Heathfield has a similar scenario as above. There is a cricket ground, a soccer field, a park, some stores but only 1 food establishment in the 10 ten. There seems to be footfall albeit it could be time-based (busy on weekends etc)

Branch 2 – Gap in the Market. Cluster 2 offers the best options for further investigations. A lot of the neighbourhoods in this already have pubs as the most popular venues along with various types of restaurants. A casual scan suggests a concentration of Italian and Asian restaurants along with some fast food places. Is there a gap in the market for say something like a South American (say Mexican) ? Or even further analysis of the Asian restaurants suggests a gap for a Japanese (sushi or izakaya) establishment. Perhaps somewhere like Fulwell and Hampton Hill or North Richmond .

Of course, further investigation and analysis is need to progress this into a real business plan but this starts off this by helping the business focus within certain neighbourhoods.