APPLIED DATA SCIENCE CAPSTONE PROJECT:

RELOCATION DILEMMA

Table of Contents

List of Figures

List of Tables

Introduction

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data which is then used to inform decision making [1]. Data science is finding application in different areas including solving location related problems. In this project, Data Science concepts are used to help a young man make an informed decision in relation to the neighbourhood he should choose to stay on moving to a new town for a job opportunity.

Background

Worley is a renowned consultancy firm specialising in chemicals, energy, and resources. It has over 50 000 employees and offices in over 60 countries all around the world [2]. John, a recent graduate, has been offered a position by the company as a process engineer. The employer has made it known to him that he will be working from the office in Stockton-on-Tees in United Kingdom. John needs to decide on which neighbourhood he will be living in and this is not an easy task as he has no prior knowledge of the quality of life in any of the neighbourhoods in Stockton-on-Tees.

John wants to live in a neighbourhood that suits his lifestyle and interests. He describes himself as a young professional who is a health enthusiast and enjoys going out and eating out. He also loves meeting new people from diverse backgrounds and learning about new cultures. John is willing to commute up to 3km to work every day. He understands that his life outside of work has a direct impact on his work performance, therefore, he is seriously considering which neighbourhood would best serve his needs and give him the best chance to succeed in his new role.

Project Aim

To determine the most suitable neighbourhood in Stockton-on-Tees that John should live based on his profile and distance from his workplace.

Project Objectives

1. To determine the neighbourhoods in Stockton-on-Tees Borough.
2. To determine the venues in the neighbourhoods in Stockton-on-Tees.
3. To cluster the neighbourhoods based on similarity in venues.
4. To determine the cluster with the most suitable neighbourhoods for John to live in.
5. To determine the most suitable neighbourhood for John to live in.

Stakeholder Interest

It is in the interest of John and his new employer that John chooses the right neighbourhood in Stockton-on-Tees to live in as this will ensure that he hits the ground running and performs well at the new job, therefore, contributing to Worley meeting its business targets. Also, John's parents will be interested in the outcome of this project as he is living the nest for the first time and they would be happy to know that he is comfortable.

Methodology

Firstly, relevant data is acquired from different sources. The acquired data is wrangled to present it in a way that makes it possible and easier for it to be analysed. Once the data is cleaned it is subjected to exploratory data analysis in order to obtain greater insights from it. Data visualization is used to emphasize the findings. The results are then recorded and discussed, and a conclusion drawn as to which neighbourhood is best suited for John. The illustration below summarizes the methodology to be followed.

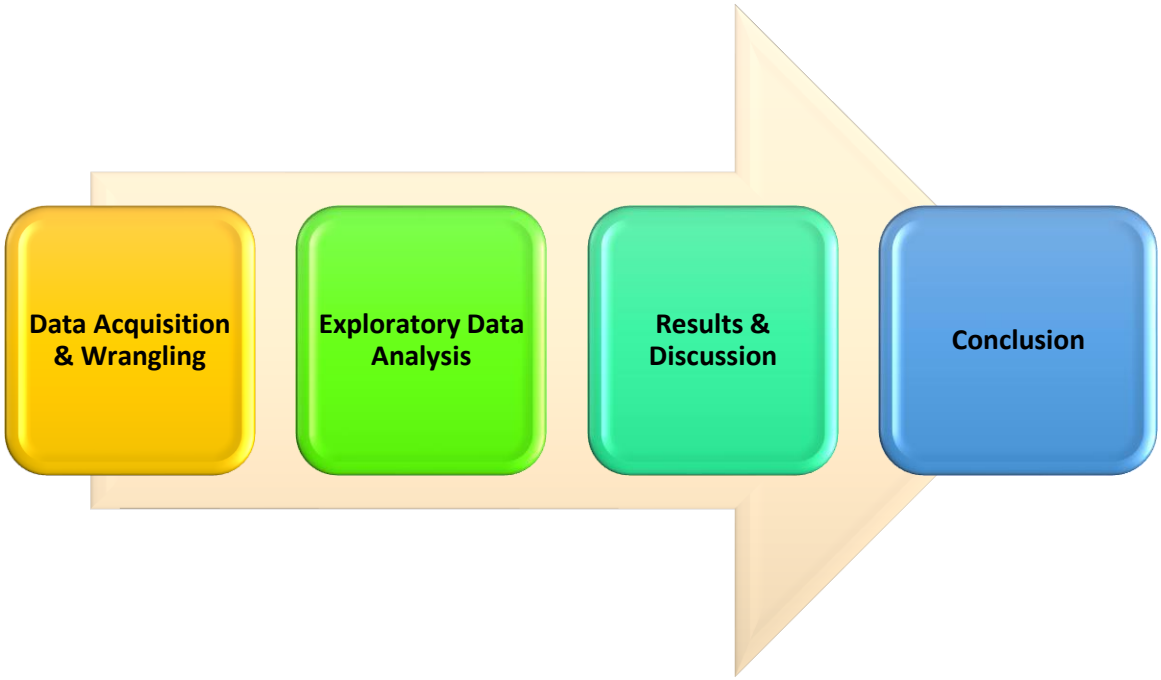

Figure 1: Project Methodology

Data acquisition and cleaning

The list of neighbourhoods in Stockton-on-Tees was retrieved from Wikipedia and scrapped from the internet by parsing data from the html into a beautifulsoup object [3]. The scrapped neighbourhoods were appended to an empty list from which a data frame containing the 13 neighbourhoods was created. A second data frame containing the neighbourhoods' latitude and longitude coordinates was also created. The two data frames were merged to form one data frame containing the 13 neighbourhoods and their coordinates.

The latitude and longitude coordinates of Worley offices, that is John's new work place, were obtained using the Worley offices address and geocoder via the Foursquare agent.

Lastly, the venues in the Stockton-on-Tees neighbourhoods which are within a 3km radius from the Worley offices were retrieved from the Foursquare database. A total of 317 venues and 37 unique venue categories in 13 neighbourhoods were retrieved thus marking the end of the data acquisition and wrangling stage. Below is a screenshot of the data frame containing the acquired data.

Table 1: Data Frame Containing Acquired and Wrangled Data.

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Category
0	Bishopsgarth	54.5780	-1.3590	Ropner Park	Park
1	Bishopsgarth	54.5780	-1.3590	The Masham	Pub
2	Bishopsgarth	54.5780	-1.3590	Co-op Food	Grocery Store
3	Bishopsgarth	54.5780	-1.3590	Co-op Food	Grocery Store
4	Bishopsgarth	54.5780	-1.3590	BP	Gas Station
...	...	...	...	...	...
312	Tilery	54.5744	-1.3068	Domino's Pizza	Pizza Place
313	Tilery	54.5744	-1.3068	Dunelm	Furniture / Home Store
314	Tilery	54.5744	-1.3068	Greggs	Bakery
315	Tilery	54.5744	-1.3068	B&M Store	Discount Store
316	Tilery	54.5744	-1.3068	Game	Video Game Store

Exploratory Data Analysis

The venues were grouped according to neighbourhood which allowed for determination of number of venues for each neighbourhood. The number of unique venue categories was found to be 37. One hot coding was then used to analyse each neighbourhood based on its venue categories. The rows from the resulting data frame were grouped by neighbourhood and the mean of the frequency of occurrence of each category calculated. The top 5 most common venues for each neighbourhood were printed. Next, a function was written to sort the venues in descending order and a new data frame showing the 10 most common venue categories for each neighbourhood was created. The code is shown below:

In [15]:
```
#write a function to sort the venues in descending order

def return_most_common_venues(row, num_top_venues):
    row_categories = row.iloc[1:]
    row_categories_sorted = row_categories.sort_values(ascending=False)

    return row_categories_sorted.index.values[0:num_top_venues]
```

In [16]:
```
#create the new dataframe and display the top 10 venues for each neighborhood

num_top_venues = 10

indicators = ['st', 'nd', 'rd']

# create columns according to number of top venues
columns = ['Neighbourhood']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

# create a new dataframe
neighbourhoods_venues_sorted = pd.DataFrame(columns=columns)
neighbourhoods_venues_sorted['Neighbourhood'] = stockton_grouped['Neighbourhood']

for ind in np.arange(stockton_grouped.shape[0]):
    neighbourhoods_venues_sorted.iloc[ind, 1:] = return_most_common_venues(stockton_grouped.iloc[ind, :], num_top_venues)

neighbourhoods_venues_sorted.head()
```

Figure 2: Code to List Top 10 Most Common Venue Categories for each Neighbourhood.

KMeans was then used to group the neighbourhoods into 5 clusters with similar venue categories. A new data frame containing cluster labels and the top 10 venues for each neighbourhood was then created. The 5 clusters (Cluster 1 – 5) were printed in separate tables for ease of inspection. The clustered were plotted on a map using folium. A screenshot of the map is given in Figure 3.

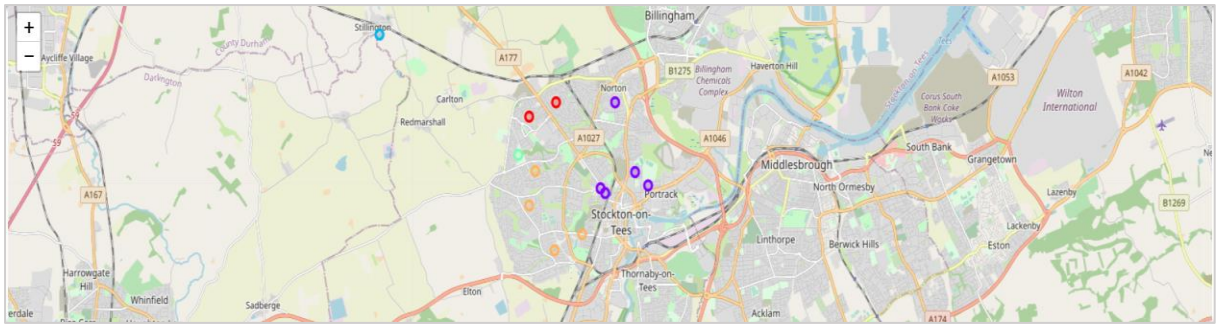

Figure 3: Map Showing 5 Clusters of Neighbourhoods in Stockton-on-Tees.

Cluster Selection

A data frame of the 37 unique venue categories was created and each venue category was assigned one of John's four profile categories, that is, Convenience, Health Enthusiast, Outgoing, and Pro-diversity. The 5 neighbourhood clusters were then be analysed against the 4 profile categories. Where a cluster satisfied one category of the profile it was awarded '1 point' otherwise it was awarded '0 points'. For Pro-diversity category, a cluster having one culturally diverse venue category was awarded '1 point', that having two culturally diverse venue categories was awarded '2points', and so on. All the points were added together. Cluster 2 had the highest mark and was selected at this stage for further analysis. See Figure 4.

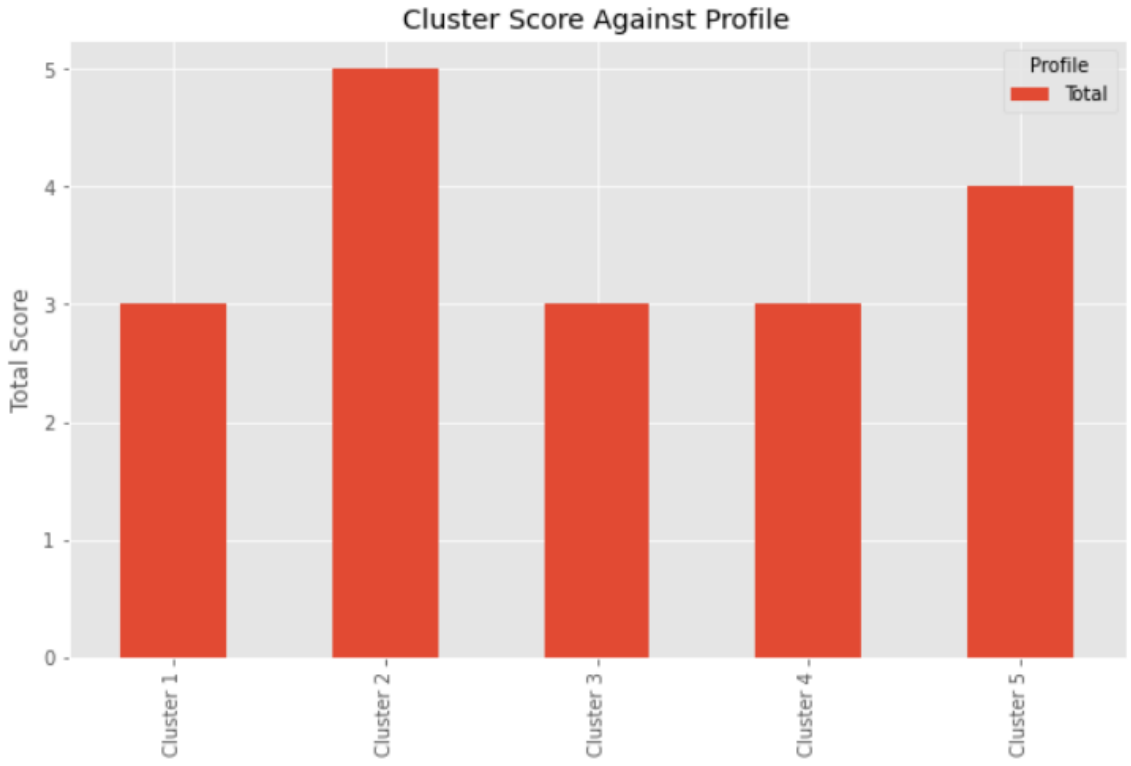

Figure 4: Cluster Score against John's Profile Categories.

Neighbourhood Selection

In order to select the most suitable neighbourhood for John within Cluster 2, the same procedure as for selection of the most suitable cluster was carried out. The results are shown in Figure 5 below.

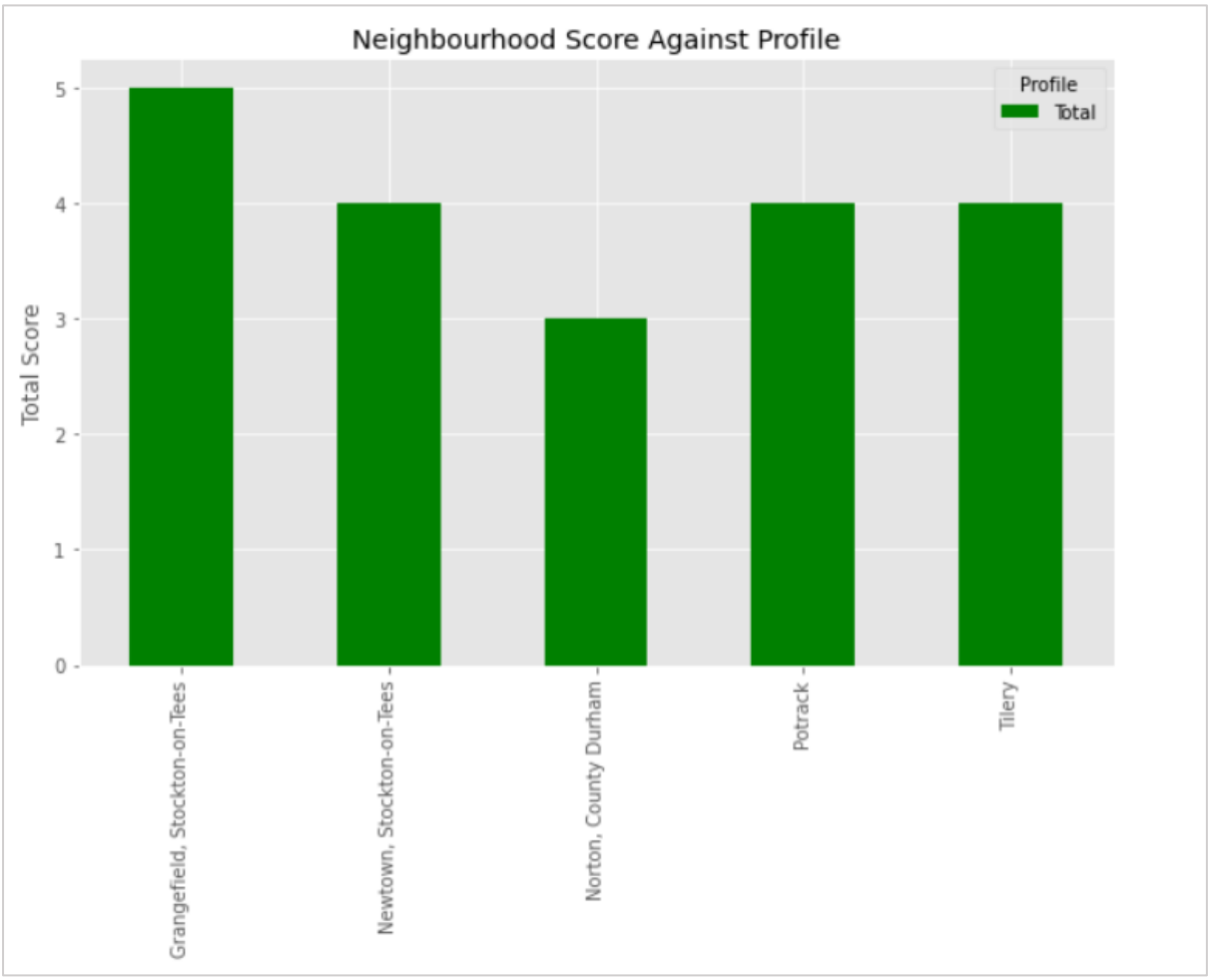

Figure 5: Cluster 2 Neighbourhood Score against John's Profile Categories.

Grangefield Stockton-on-Tees gave the highest mark of 5 showing that it best suits John's profile compared to the other neighbourhoods in Cluster 2. However, Newtown Stockton-on-Tees, Potrack and Tilery, also satisfied all the aspects of John's profile although they do not offer as much cultural diversity as Grangefield. A calculation of distance between each of the four neighbourhoods to John's workplace was, therefore, carried out to determine which one is closest hence is the overall best neighbourhood. The distance between the neighbourhoods and Worley offices was calculated using the Haversine formula. A function was created for the distance calculation and then applied to location coordinates of the four neighbourhoods. The function is given below.

```
#create function to calculate distance between Worley offices and neighbourhood

import math

def distance (lat1, lat2, lon1, lon2):

    R = 6373.0     #radius of the Earth

    #calculating distance between Elm Tree Farm and Worley Offices

    lat1 = math.radians(lat1)
    lon1 = math.radians(lon1)
    lat2 = math.radians(lat2)
    lon2 = math.radians(lon2)

    dlon = lon2 - lon1
    dlat = lat2 - lat1

    a = math.sin(dlat / 2)**2 + math.cos(lat1) * math.cos(lat2) * math.sin(dlon / 2)**2

    c = 2 * math.atan2(math.sqrt(a), math.sqrt(1 - a))
    distance = R * c

    return distance
```

Figure 6: Code to Create Function for Distance Calculation.

The calculated distances were plotted in a bar chart. See Figure 7.

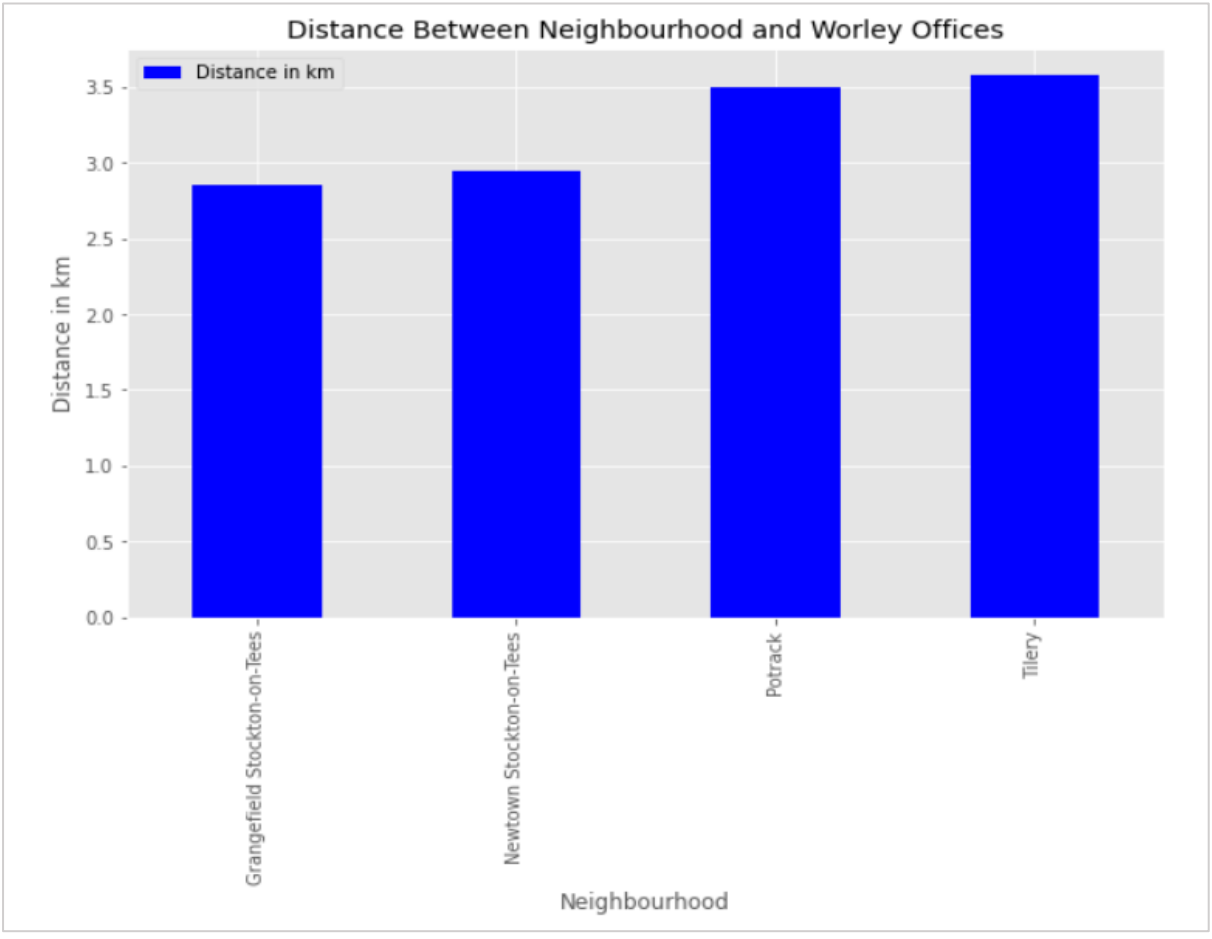

Figure 7: Distance between Cluster 2 Candidate Neighbourhoods and Worley Offices.

Grangefield Stockton-on-Tees was found to be closest to the Worley Offices being 2.85km away from it. Grangefield is the most suitable neighbourhood in terms of venues in the neighbourhood which suit John's profile closely, and also because of its proximity to John's workplace.

Results and Discussion

Five clusters of neighbourhoods in Stockton-on-Tees were created using KMeans and similar neighbourhoods were assigned to each cluster. A map showing the five clusters was created using folium. The five clusters were analysed to determine which one was better suited to John's profile. Firstly, John's profile was split into four categories, that is, Convenience, Health Enthusiast, Outgoing, and Pro-diversity. The 37 venue categories in Stockton-on-Tees which were obtained from Foursquare database were matched to a category of John's profile. The Clusters were then analysed against the profile categories. If a cluster satisfied one category of the profile it was awarded '1' otherwise it was awarded '0'. Under Pro-diversity, a cluster having one culturally diverse venue category was awarded '1', that having two culturally diverse venue categories was awarded '2', and so on. Cluster 2 had the highest mark of 5. The neighbourhoods in the cluster have Gym/Fitness Centres where John can go and workout whenever he wants as he is a health enthusiast. The cluster has a wide selection of hangout spots, some of which are more suited for professional meet ups like the coffeeshop and restaurants, and some which are a bit more casual like the pub, casino, bowling alley, and theatre. Cluster 2 also has several venues that are convenient and would make John's life easier, these include, supermarkets, shopping malls, discount stores, bakeries, and furniture/home stores. Lastly, Cluster 2 shows cultural diversity which is an important factor of John's profile. It has Italian Restaurant and American Restaurant venue categories. Cluster 2 satisfies John's profile the most and was, therefore, selected for further analysis.

Cluster 2 was further analysed in order to determine the neighbourhood in the cluster that best suits John's profile. the same procedure as for selection of cluster was carried out. Grangefield Stockton-on-Tees satisfied all aspects of John's profile and had the highest overall score of 5. However, Newtown Stockton-on-Tees, Potrack and Tilery, also satisfied all of John's profile categories although they did not offer as much cultural diversity as Grangefield. A calculation of distance between each of the four remaining candidate neighbourhoods and the Worley Offices was carried out in order to confirm which neighbourhood was in overall most suitable for John to live in. The distance was calculated using the location coordinates in the Haversine formula. The best neighbourhood would be one that suits John's profile and is the shortest distance to his workplace. Analysis of the obtained results showed that Grangefield Stockton-on-Tees was the nearest to the Worley offices with a distance of 2.85km. It was, therefore, selected as the most suitable neighbourhood for John to live in.

Conclusion

The purpose of the project was to determine the most suitable neighbourhood for John to live in, when he located to Stockton-on-Tees in the UK, which would make him feel comfortable and allow him to perform at his best at his new job. The list of neighbourhoods in Stockton-on-Tees was scrapped from Wikipedia together with their coordinates. Foursquare was used to determine the neighbourhoods within a 3km radius from Worley offices as well as retrieve the venues in these candidate neighbourhoods. KMeans was then used to cluster similar neighbourhoods with the resulting 5 clusters being visualized using folium. The clusters were analysed to determine the one most suited to John's profile based on the venue categories. Cluster 2 proved to be the most suited,

therefore, it was further analysed to pick a neighbourhood within the cluster which would be recommended to John. The neighbourhood was selected based on venues as well as distance from Worley offices. The neighbourhood giving the best results was Grangefield Stockton-on-Tees which is 2.85km from John's workplace and satisfies all of his profile categories whilst also offering greater cultural diversity.

Recommendation

It is recommended that John resides in Grangefield Stockton-on-Tees as it most suited to his profile and is within his preferred commuting distance to work being only 2.85km from Worley Offices.

References

[1] IBM, Composer, *What is Data Science.* [Sound Recording]. Coursera. 2021.

[2] Worley, "Who we are," Worley, [Online]. Available: https://www.worley.com/who-we-are. [Accessed 28 March 2021].

[3] Wikipedia, "Category: Areas of Stockton-on-Tees," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Category:Areas_of_Stockton-on-Tees. [Accessed 28 March 2021].