L.A (not so) Confidential – Examining the diversity of downtown L.A neighbourhoods

Introduction

With a population of four million residents, Los Angeles (L.A) is the second most populous city in the United States after New York. Due to its geographical location, appealing climate and strong links with the entertainment industry, Los Angeles is home to an incredibly diverse population living the American life.

For my Capstone project, I decided it would be an interesting idea to investigate the diversity of L.A neighbourhoods. This would provide a valuable insight into the kinds of venues that various L.A neighbourhoods have on offer as well as the kind of people that live in L.A and where abouts they are located.

The results of this project would be of use to anyone who would like to examine the trends in L.A neighbourhoods, i.e what are the prevalent characteristics of L.A neighbourhoods and where similar neighbourhoods are located. Such persons could be but are not limited to:

- A tourist visiting L.A who wants to see roughly how each area differs and what it contains.
- A businessman/businesswoman wanting to open up a new restaurant (for example) and would like to get a rough idea of each neighbourhood and what it contains.
- A researcher wanting to investigate the demographics of downtown LA by observing what kinds of businesses operate in the neighbourhoods.

Lastly, the other key point of this project is for me to successfully utilise the knowledge gained throughout my participation in the 'IBM Data Science' but applied to a genuine, real world problem.

Data

The points of interest for this project are the names of the L.A neighbourhoods. These can easily be obtained by webscraping a Wikipedia page for "L.A neighbourhoods" using the BeautifulSoup package as explained in the course and saved in a list ordered alphabetically.

The second data feature that must be obtained are the latitude and longitude values of each neighbourhood. These can be obtained through various python libraries (such as geopy) so should not be a problem. Of course, to obtain the correct latitude and longitude values of the neighbourhoods, the neighbourhood names need to already be obtained from the Wikipedia page. Care should be taken to obtain the latitude and longitude values of the neighbourhoods which are in L.A itself and not some other city as it is highly likely that the neighbourhood names are not unique to L.A. To overcome this, the string "Los Angeles" or something similar should be added to the neighbourhood names to specify the exact neighbourhoods we are looking for.

Lastly, information on each of the neighbourhoods needs to be obtained using Foursquare API (as shown in the data science course). For this project, only the venue category is of any relevance. It would also be helpful to obtain the names of the various venues as well as their latitude and longitude values incase further investigation would like to be carried out (for example, a venue may

appear twice but under different neighbourhoods, hence one of the instances should be deleted). Attributes such as price, reviews are completely out of the scope of this project and should not be obtained at all using the API.

Methodology

Obtaining neighbourhood names and coordinates:

Having downloaded the required libraries the first task was to webscrape the names of the neighbourhoods from the Wikipedia page "L.A neighbourhoods" (https://en.wikipedia.org/wiki/List_of_districts_and_neighborhoods_of_Los_Angeles) using the BeautifulSoup package. The neighbourhood names were saved to a list. The list contained many "None" values which were easily removed with a simple 'for loop' and 'if' statement, reducing the list length from 474 entries to 200.

Each entry in the list had the neighbourhood name as well as ', Los Angeles' or '(Los Angeles)' as part of the string. Rather than removing these parts immediately to clean the data set, they were kept and passed directly in to the 'geolocator' object (the 'geolocator' object is a feature of the geopy library which returns location coordinates of places around the world). This was done to specify that the neighbourhoods are in Los Angeles, as it is highly likely that the neighbourhood names are not unique to Los Angeles and exist in other U.S cities too. Once the latitude and longitude coordinates were obtained, the entries in the neighbourhood list were cleaned to remove the 'Los Angeles' string and a pandas data frame was created to display the information (Figure 1).

	Neighbourhoods	Latitude	Longitude
0	Angelino Heights	34.070289	-118.254796
1	Angeles Mesa	33.991402	-118.319520
2	Arleta	34.241327	-118.432205
3	Arlington Heights	34.043494	-118.321374
4	Arts District	34.041239	-118.234450

Figure 1: First 5 entries of the pandas data frame 'df' displaying the names of the neighbourhoods.

Exploratory data analysis:

It is standard data science practice to explore your raw data to understand exactly what it contains and what changes need to be made before applying the machine learning model. For the neighbourhood locations two forms of data analysis were performed, a box plot and a visualization using the Folium library.

The box plot was used on both the latitude and longitude coordinates to see which neighbourhoods can be considered outliers, with the intention of removing these neighbourhoods as they contribute

misleading information and hence reduce the effectiveness of the machine learning model. The two box plots are shown below in Figure 2:

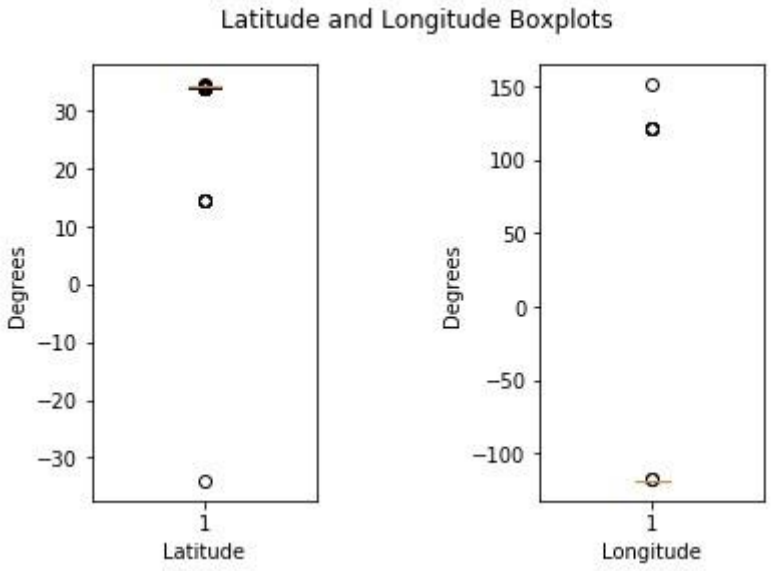

Figure 2: Initial box plots of latitude and longitude values

It is immediately apparent that there are some very extreme outliers in the data set, so much so that the actual box plot is barely visible. These were probably neighbourhoods in entirely different cities which the geopy library returned instead of the specified Los Angeles neighbourhoods. A folium map was also produced which showed a few outliers within Los Angeles itself which were also worth removing (Figure 3):

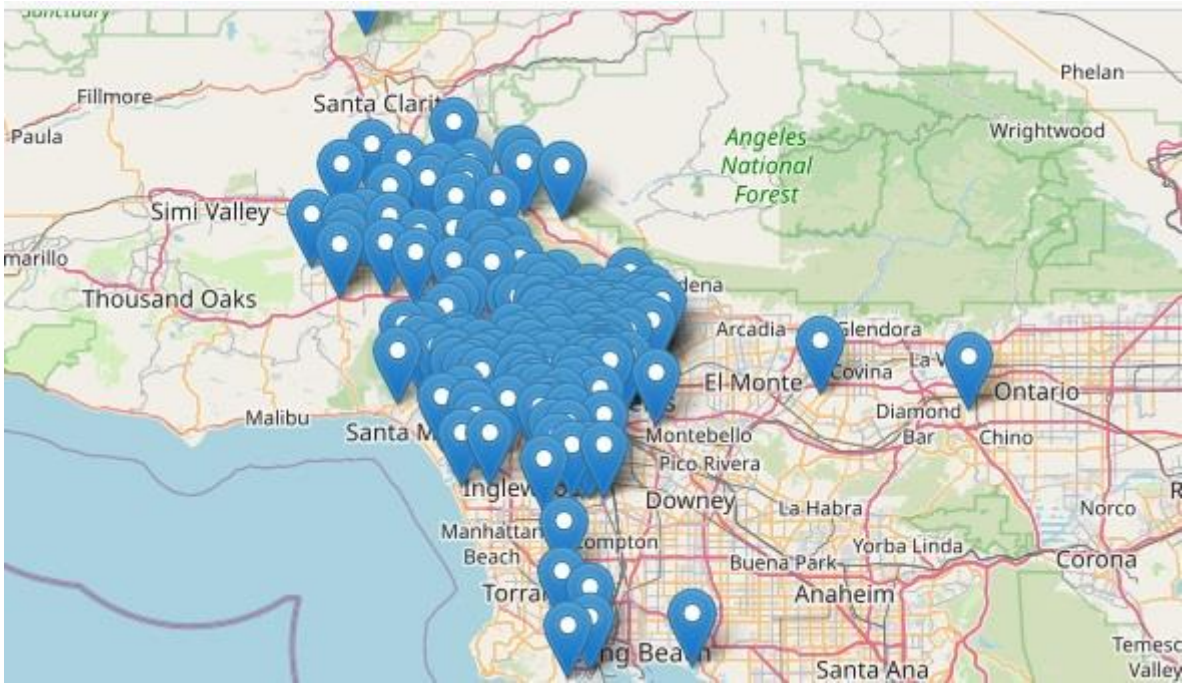

Figure 3: Neighbourhood locations within and around L.A

As there are only a few outliers in the data set, it was easier to outright remove all outlier entries rather than try and re-run the geopy module to find the neighrbourhoods that were attributed to other cities.

By trial and error using the numpy.percentile() function, the outliers were completely removed. The percentile values used are shown in Table 1:

Table 1: Percentiles used to remove outliers

Percentiles	
Latitude upper	89
Latitude lower	7
Longitude upper	95
Longitude lower	3

The result of the data cleaning on the box plots is shown below in Figure 4:

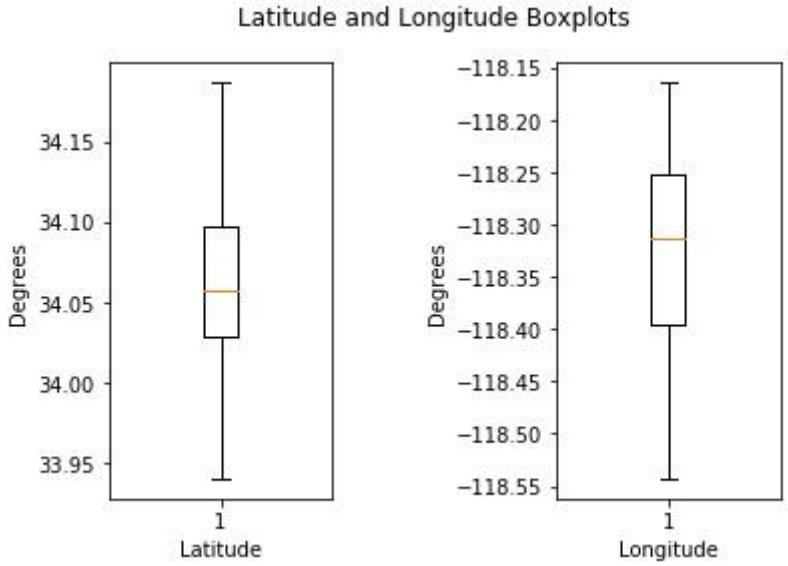

Figure 4: Box plots for Latitude and Longitude after outliers were removed

It is apparent that all outliers were removed and the data is evenly spread, showing neither a positive nor a negative skew. A visual aid using the folium library was also produced to complement the box plots, shown in Figure 5:

Note that the process of removing outliers reduced the number of neighbourhoods in the dataset from 175 to 138.

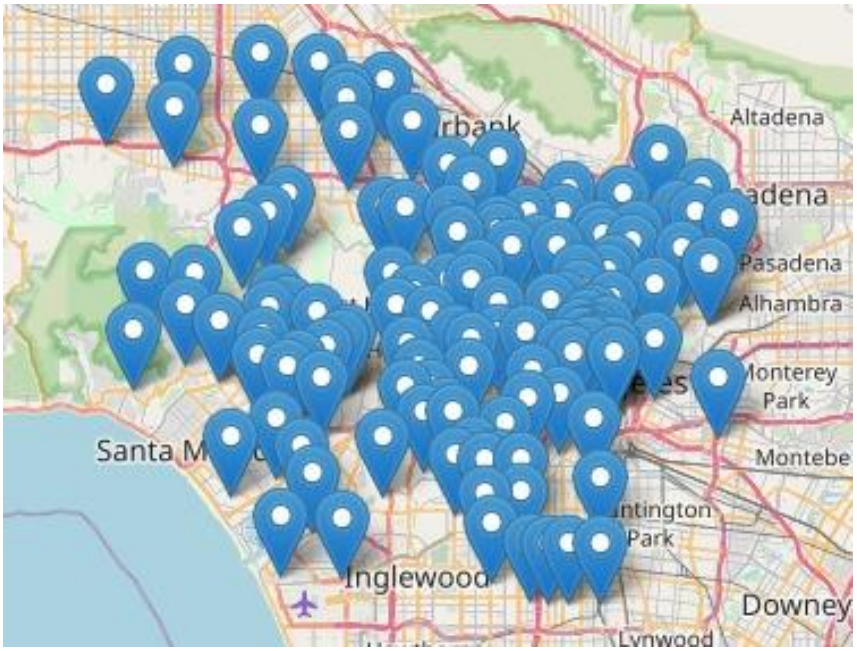

Figure 5: Neighbourhood locations after removing outliers

Initial test:

Having obtained the correct neighbourhoods, the next task was to use the Foursquare API to start importing the features that each neighbourhood has and start analysing them.

A test run was performed on the first neighbourhood entry in the list of neighbourhoods, 'Angelino Heights'. The neighbourhood latitude and longitude coordinates were passed to the Foursquare API (with limit set to 100 and radius set to 500) and a JSON file was returned from which the venue name, category and coordinates were obtained. A pandas dataframe was then created to display all the results for Angelino Heights (Figure 6):

	venue.name	venue.categories	venue.location.lat	venue.location.lng
0	Halliwell Manor	Performing Arts Venue	34.069329	-118.254165
1	Guisados	Taco Place	34.070262	-118.250437
2	Eightfold Coffee	Coffee Shop	34.071245	-118.250698
3	Ototo	Sake Bar	34.072659	-118.251740
4	Michael Jackson's "Thriller" House (and Tree)	Historic Site	34.069557	-118.254599

Figure 6: Top 5 features of 'Angelino Heights' as a pandas dataframe, using '.head()' function.

Obtaining all venues:

Having successfully passed this test a function 'getNearbyVenues' was created based off the test run code where each neighbourhood name, latitude and longitude values would be passed to the Foursquare API with a JSON file being returned. Like the test run, only the venue name, category, latitude and longitude locations would taken from the JSON file.

The function ‘getNearbyVenues’ was successfully called for all 138 neighbourhoods. Like the test run, the venue’s key characteristics were saved to a pandas dataframe (called ‘LA_venues’) along with the neighbourhood name and the neighborhood’s latitude and longitude coordinates. An example of the dataframe is shown in Figure 7:

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Angelino Heights	34.070289	-118.254796	Halliwell Manor	34.069329	-118.254165	Performing Arts Venue
1	Angelino Heights	34.070289	-118.254796	Guisados	34.070262	-118.250437	Taco Place
2	Angelino Heights	34.070289	-118.254796	Eightfold Coffee	34.071245	-118.250698	Coffee Shop
3	Angelino Heights	34.070289	-118.254796	Ototo	34.072659	-118.251740	Sake Bar

Figure 7: Dataframe ‘LA_venues’ containing all venues from all 138 L.A neighbourhoods

The dimensions of ‘LA_venues’ are 3357 by 7, meaning there are 3357 venue entries and 7 columns.

Of course, the neighbourhoods have different numbers of venue categories due to their size and diversity. To visualize this disparity, a histogram was plotted with number of venues on the x-axis and number of neighbourhoods on the y-axis, shown in Figure 8:

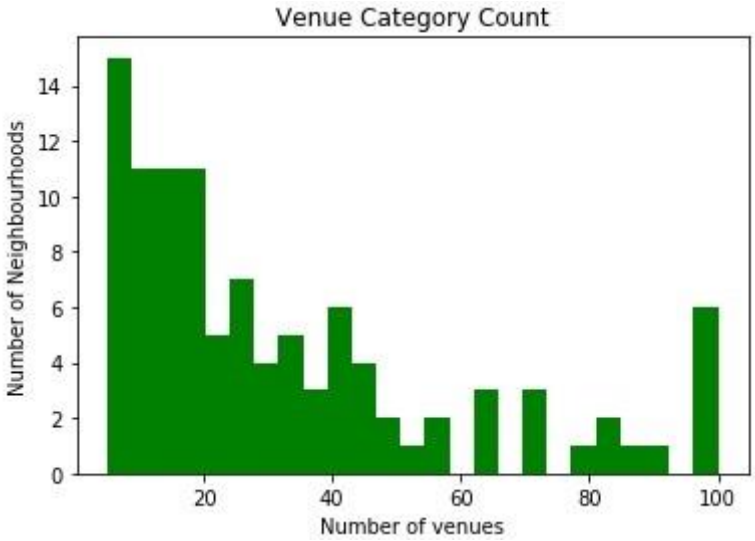

Figure 8: Histogram showing frequency of number of venues per neighbourhood

As can be seen from the histogram, many neighbourhoods contain have less than 4 different types of venues. Recall that the goal of this exercise is to use a clustering algorithm to assign each neighbourhood to a cluster. For this to be successful it is preferable that each neighbourhood be as diverse as possible, i.e contain many different types of venues. As such, the neighbourhoods containing 4 or less venues were omitted from the dataset as they were not considered diverse enough and so would skew the results of the machine learning model.

Once these neighbourhoods were eliminated, the amount of neighbourhoods in the dataset decreased from 138 to 104.

The final step before applying the machine learning model was to apply one hot encoding to the trimmed dataset, grouping the data by neighbourhood name. This is a necessary step as it groups the data by neighbourhood name and displays the frequency of each venue category in that

neighbourhood. Only like this can the machine learning algorithm (k means clustering) act on the dataset and group each neighbourhood according to the venues it contains.

Neighbourhood	Latitude	Longitude	ATM	Accessories Store	Adult Boutique	Airport Lounge	Airport Terminal	American Restaurant	Amphitheater	Aquarium	...	Video Game Store	Video Store	Vietnamese Restaurant
Angeles Mesa	33.991402	-118.319520	0.000000	0.000000	0.000	0.00000	0.000000	0.000000	0.0	0.0	...	0.000000	0.000000	0.000000
Angelino Heights	34.070289	-118.254796	0.000000	0.000000	0.000	0.00000	0.000000	0.000000	0.0	0.0	...	0.000000	0.000000	0.000000
Arlington Heights	34.043494	-118.321374	0.000000	0.000000	0.000	0.00000	0.000000	0.000000	0.0	0.0	...	0.000000	0.000000	0.000000
Arts District	34.041239	-118.234450	0.000000	0.000000	0.000	0.00000	0.000000	0.000000	0.0	0.0	...	0.000000	0.000000	0.000000
Atwater Village	34.118698	-118.262392	0.000000	0.000000	0.000	0.00000	0.000000	0.023810	0.0	0.0	...	0.000000	0.000000	0.047619
Baldwin Hills	34.010989	-118.337071	0.000000	0.025641	0.000	0.00000	0.000000	0.000000	0.0	0.0	...	0.025641	0.000000	0.000000
Baldwin Village	34.019456	-118.345910	0.000000	0.000000	0.000	0.00000	0.000000	0.000000	0.0	0.0	...	0.000000	0.000000	0.000000

Figure 9: One hot encoding of all venues, grouped by neighbourhood name

K-means clustering:

The machine learning algorithm selected for classifying the neighbourhoods was k-means clustering, imported from the 'scikit learn' library. K-means was selected as it is the 'go to choice' amongst the data science community for simple clustering models, due to its simplicity, speed and variety of uses. Selecting the number of clusters was a somewhat arbitrary decisions, but by experimenting with different numbers of clusters, it was deemed that 4 clusters was the most suitable. Anything less than 4 would have been restrictive and not shown the true diversity of the data and anything more than 4 produced groups that were quite similar and made displaying the data more difficult.

The results of the machine learning algorithm were saved to a list and were then assigned to each neighbourhood. Using Folium, a map was crated of the neighbourhoods, with the colour of the marker indicating the group that the neighbouhood belongs to. The colour – group pairing is shown in Table 2 and the Folium map in Figure 10:

Table 2: Cluster – Colour pairings

Cluster	Colour
0	Red
1	Blue
2	Green
3	Purple

Table 3: Number of neighbourhoods per cluster

Cluster	No. of Neigh.
0	93
1	9
2	1
3	1

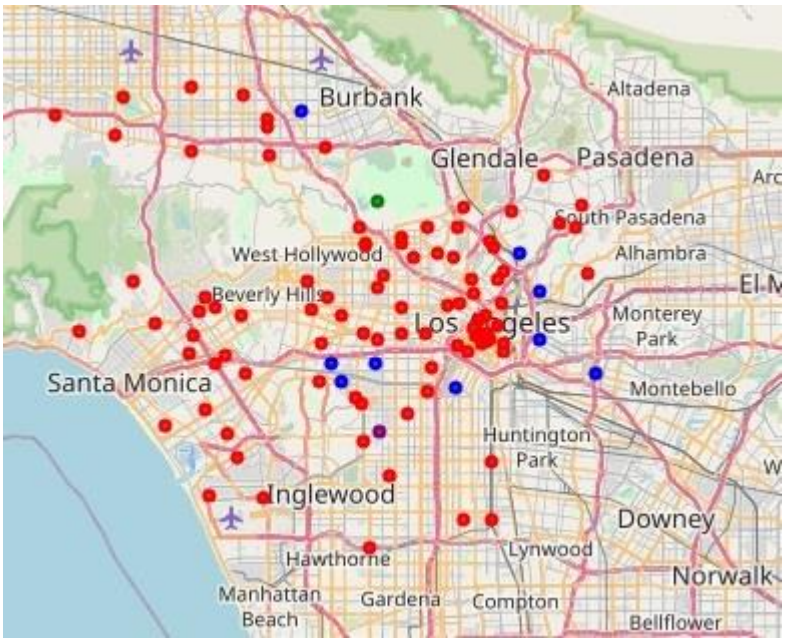

Figure 10: Neighbourhoods with cluster assignment

Following these plots, the dataframe 'LA_common_venues' was created which ranks the 10 most common venues in each neighbourhood and displays the cluster number for each neighbourhood. From this dataset, a histogram was created for each cluster with the venue frequency on the x-axis and the venue category on the y-axis (Figure 11).

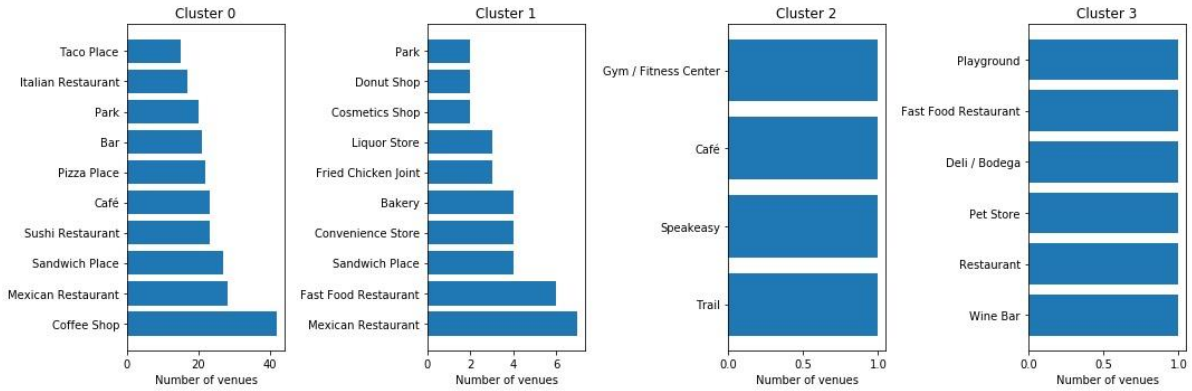

Figure 11: Number and type of venues per cluster group

This is the final plot of the project and achieves the project objective by showing what each cluster group contains.

Results & Discussion

As a brief reminder, the original objectives of the project were to investigate the diversity of LA neighbourhoods by splitting them into groups and seeing how these groups look on a map of LA. This was done by using k-means clustering to assign each neighbourhood to a cluster based on the type of venues located within that neighbourhood. For this project, each neighbourhood was assigned to one for 4 clusters.

The key results of the project are displayed in Table 3 and Figures 10, 11.

Table 3 shows how many neighbourhoods appear in each cluster, with the total number of neighbourhoods being 104.

Figure 10 shows the location of each neighbourhood in L.A with the colour of the marker indicating which group the neighbourhood belongs to. The colour - group pairing is explained in Table 2.

Lastly, Figure 11 shows 4 histograms (one for each group) which displays the top 10 most popular venues for the group and their total number for that group.

The most striking result is how unevenly distributed the cluster sizes are. Clusters 2 & 3 (green & purple) for instance only contain 1 neighbourhood while cluster 1 (blue) contains 9, with all the rest belonging to cluster 0 (red). Each cluster is analysed below:

Green cluster: The easiest cluster to analyse is the green cluster, which judging by its position on the map and venues is a park with amenities such as a trail and café. While there are other parks in L.A, perhaps they do not contain any amenities or had fewer than 4 and so were filtered out before clustering. Indeed, a park is a unique feature of a city, so it is hardly surprising that this formed a cluster on its own.

Purple cluster: The second easiest cluster to analyse is the purple cluster – which only has one neighbourhood. The position of the cluster is in a southern part of L.A and is close to a few green areas and is surrounded by red neighbourhoods. As for venues, the purple cluster contains a unique mix of up-market venues like a wine bar and deli as well as more surprising results like a playground and pet shop. Given the location and venues in the cluster, it seems this is an upper class and unique neighbourhood, that combines expensive dining with more child friendly venues like a pet store and playground (perhaps influenced by the presence of the park nearby). Further analysis would need to be done to find out why exactly this neighbourhood is in a cluster of its own and does not follow a similar trend to the encircling red neighbourhoods.

Blue cluster: Figure 10 indicates that the blue clustered neighbourhoods are mostly in pairs and surround the central, most densely populated area of L.A and are far from the sea or any parks. In terms of venues, the blue cluster is not too dissimilar to the extremely popular red cluster, with both clusters sharing 2 of their top 3 venues. The defining characteristic of the blue clustered neighbourhoods is the popularity of Mexican Restaurants in these neighbourhoods, indicating that these neighbourhoods have a high number of Mexican inhabitants. People of the same ethnicity generally like to live close together in foreign countries, which explains why the blue neighbourhoods mostly come in groups and (with one exception) are close to the center of L.A.

Red cluster: By far the most popular cluster of the 4 is the red cluster, which makes up around 90% of all L.A neighbourhoods. Figure 10 is not particularly insightful in this case as the red neighbourhoods are dispersed all over L.A due to their sheer number, so it is difficult to spot any trends in location. The histogram in Figure 11 indicates that by far the most popular venue in these areas are coffee shops with the following 9 (excluding parks) being food & drink venues. In terms of demographics, there seems to again be a heavy Mexican influence due to the number of Mexican restaurants in these areas as well as a strong Italian presence (around 20 restaurants) and a strong Asian influence (around 25 sushi restaurants). Unfortunately, due to the sheer size of the Red cluster it is difficult to pick out any unique features of these neighbourhoods. Instead, the red cluster represents L.A as a whole, depicting it as a fast moving city with many ‘grab and go’ food & drink venues (such as coffee shops and sandwich places) with a strong Mexican, Italian and Asian influence.

Overall Discussion: The results of the machine learning model are not what I expected. Considering cultural and economic diversity of L.A I was surprised to see 90% of the neighbourhoods fall under the same cluster, with 2 of the other clusters containing only 1 neighbourhood each. I was expecting more diversity between with clusters, with some clusters containing more cinemas and entertainment venues than others due to the world-famous filming industry that exists in L.A. In addition, I would have expected to see a greater variety of nationalities, but instead only saw a Mexican, Italian, and Asian presence. Lastly, the most surprising fact of all was seeing 90% of the neighbourhoods fall under the same category, indicating that L.A is not as diverse as I had anticipated.

A possible improvement on my data science project would be to increase the number of clusters with the aim of breaking down the red cluster in to smaller, more insightful clusters which would better represent the diversity of L.A.

In terms of demographics, the Mexican, Italian and Asian population were observed by the number of restaurants that exist in the neighbourhoods. An insightful test would be to eliminate all venues other than restaurants of specific nationalities and then clustering them. This would be effective at

highlighting the demographic differences of L.A which are otherwise overshadowed by generic venues like coffee houses.

Lastly, since the red cluster is vastly dominated by food & drink venues, it would be helpful to eliminate these entirely so that less popular venues such as cinemas or bowling alleys can be examined as they were overshadowed by the food and drink venues.

Conclusion

The original goal of the project was to examine the diversity of L.A neighbourhoods by clustering the neighbourhoods according to the venues they contain and seeing the position of these clusters on a map of L.A.

The results indicated that L.A is not as diverse as one would expect, with 90% of neighbourhoods falling under the same cluster, with 2 of the 4 clusters containing 1 L.A neighbourhood. The most popular venues of the most popular cluster (red) were food and drink venues. These were either small 'grab and go' establishments like sandwich places and coffee shops, or ethnic restaurants such as Mexican, Italian, or Asian. Likewise, the blue cluster was dominated by Mexican restaurants, indicating that Mexicans make up a large part of L.A's population. The remaining 2 clusters were very different; one being very nature and exercise themed and the other having a mix of high-class eating venues and child friendly establishments.

Due to the disappointing outcome of the project, a few key changes were proposed that would make better use of the data and would better highlight LA's diversity. The first change would be to simply increase the number of clusters in the hope of breaking down the red cluster into smaller, more insightful clusters. Secondly, all venues should be eliminated other than ethnic restaurants in order to better highlight the demographics of L.A. Alternatively, all food and rink venues could be eliminated (as these make up the vast majority of venues) so that the entertainment venues and other amenities can be examined.

Despite the shortcomings, the project proved an invaluable learning experience which tested many key skills that a modern data scientist needs.