Clustering the world's most visited countries

Nathan Smit

May 24, 2020

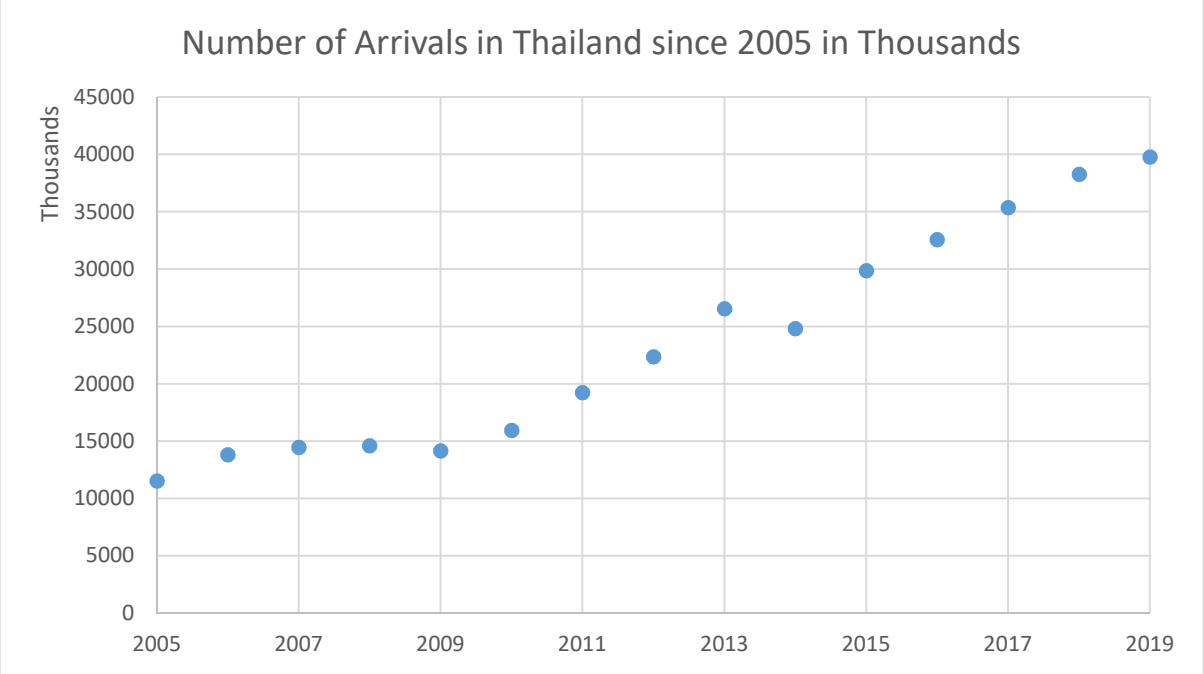
1. Introduction

The purpose of my assignment was to analyse the most visited cities in the world and cluster them based on popular venues in each city, such as museums, restaurants, cafés etc. We can use this analysis to determine the similarity of these cities to one another. We can then look at the most visited regions and see if these also have similar characteristics in terms of popular venues. Finally, we can overlay spending information for these cities to find cities

1.1 Business Problem

In recent years, Thailand has seen a large increase in its number of annual arrivals. In fact, since 2005, Thailand has seen a 245% increase in arrivals as can be seen in the above chart. As an assignment, I thought it'd be interesting to see if the most visited cities share characteristics and if less-visited cities might be able to replicate Thailand's success. This project therefore aims to determine if we can cluster the most-visited cities based on their most common venues.

Source: https://en.wikipedia.org/wiki/Tourism_in_Thailand



1.2 Interest

This would we be useful information for Travel companies when suggesting new locations to travel to. We can also cluster the income earned from Tourism in these countries (available from Wikipedia as Mastercard income is provided) to identify locations which are similar but which fall into different income clusters. This could be useful to governments looking to promote tourism in these cities.

2 Data acquisition and cleansing

2.1 Data sources

Data was sourced from a Wikipedia page listing most visited destinations as well as the Foursquare API.

The Wikipedia page (https://en.wikipedia.org/wiki/List_of_cities_by_international_visitors) consists of data in a table based on rankings provided by Euromonitor and Mastercard. These two entities have different definitions of "foreign visitor" and so these are ranked separately. For my analysis, I focused on the Mastercard data for 2016 as this enabled me to also make use of the Mastercard Income data.

To extract venues for each city, the Foursquare API was used. Specifically, I made use of the "Explore endpoint" (documented here: https://developer.foursquare.com/docs/api-reference/venues/explore/) which returns a list of recommended venues near the current location.

2.2 Data cleaning and feature engineering

I made use of the Pandas library to scrape data from the International Visitors Wikipedia page. As Mastercard and Euromonitor data were part of a single table, I made the decision to remove the Euromonitor data. I discovered at this point that there were only 99 entries for the Mastercard data as opposed to 100. Other than this, there didn't appear to be any other issues with the data. As I was going to be performing clustering, I decided not to remove any outliers. Each city had the "Income" amount stated in billions. I transformed this by multiplying each entry by 1 billion so that I could create an "Income per arrival" feature.

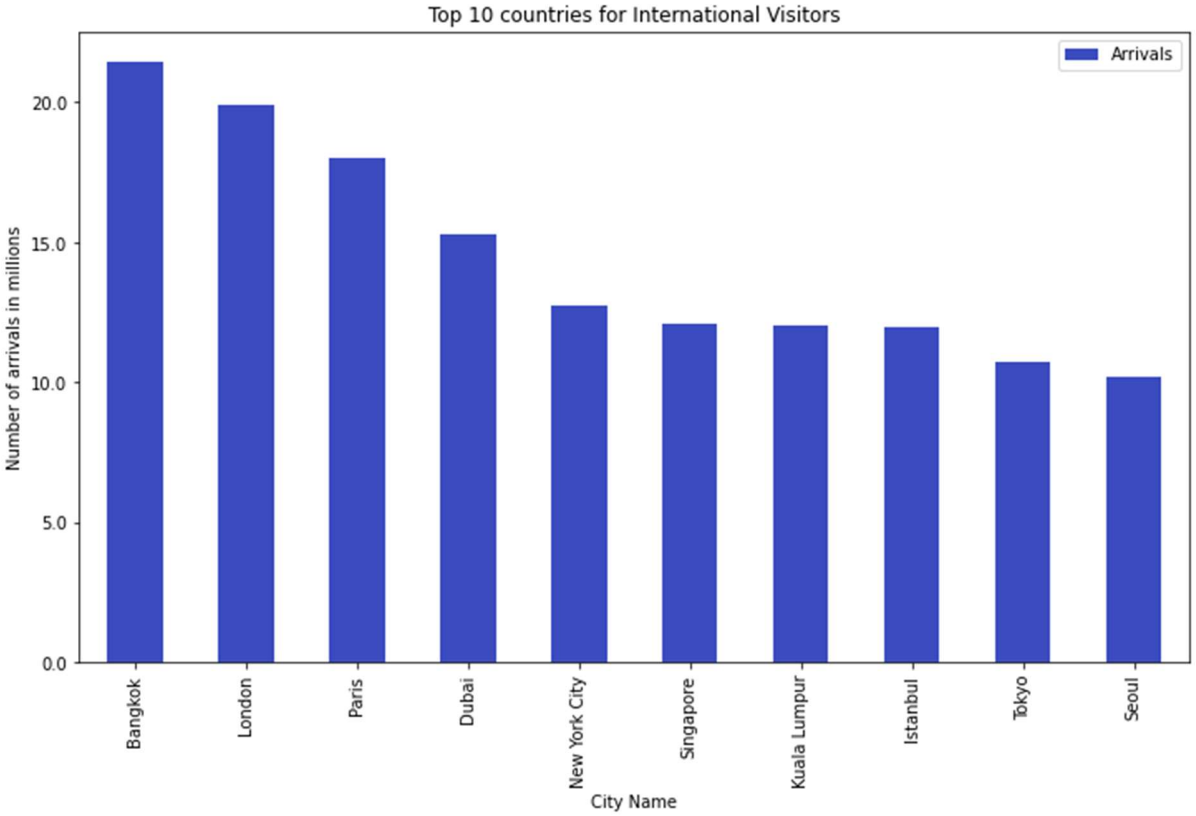
In order to retrieve data from the API, I required the latitude and longitude coordinates for each city. I did this geocoding using the geopy Nominatim library. A table was created that included the City, Country, Latitude, Longitude, Arrivals, Income and Income per Arrival.

Finally, data from the Foursquare "Explore" endpoint was extracted. The ten most common venues for each city was extracted and this data was the basis for additional clustering of the data.

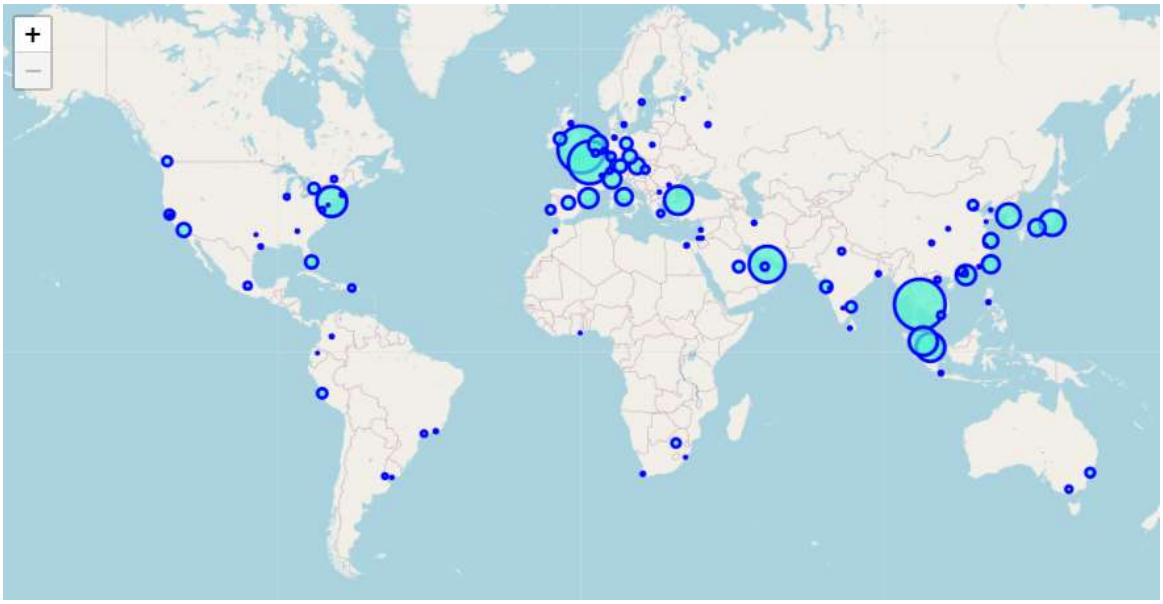
3 Exploratory data analysis

3.1 Top cities for international visitors

Exploring the Mastercard data, the cities with the most arrivals were large, well-known cities as can be seen from the top 10 countries chart below.

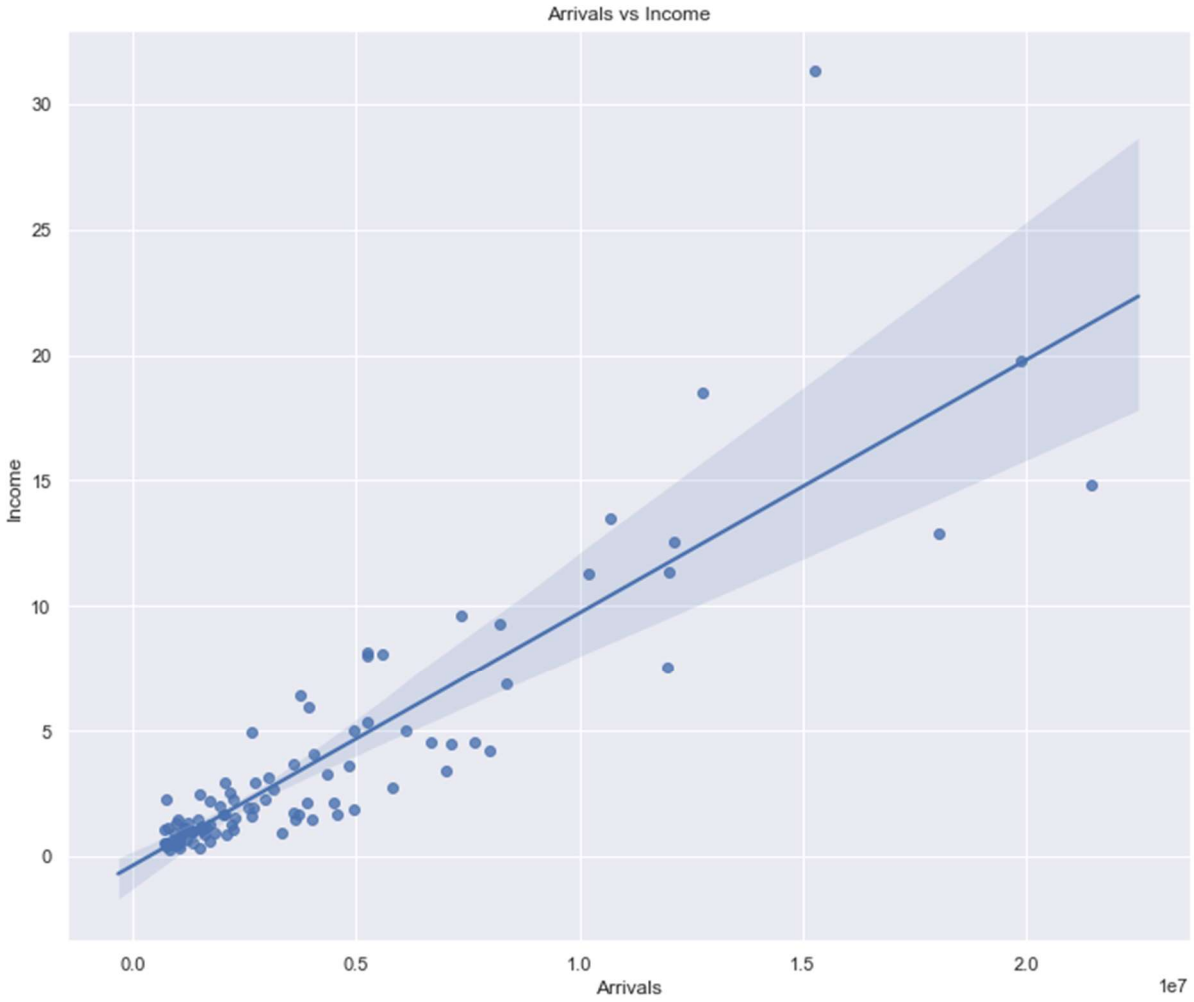


The number of arrivals for each city was plotted using the Folium library. This revealed that most of the visitors are concentrated in Europe and Asia, with New York being the only country outside of these continents where similarly large number of visitors were reported.



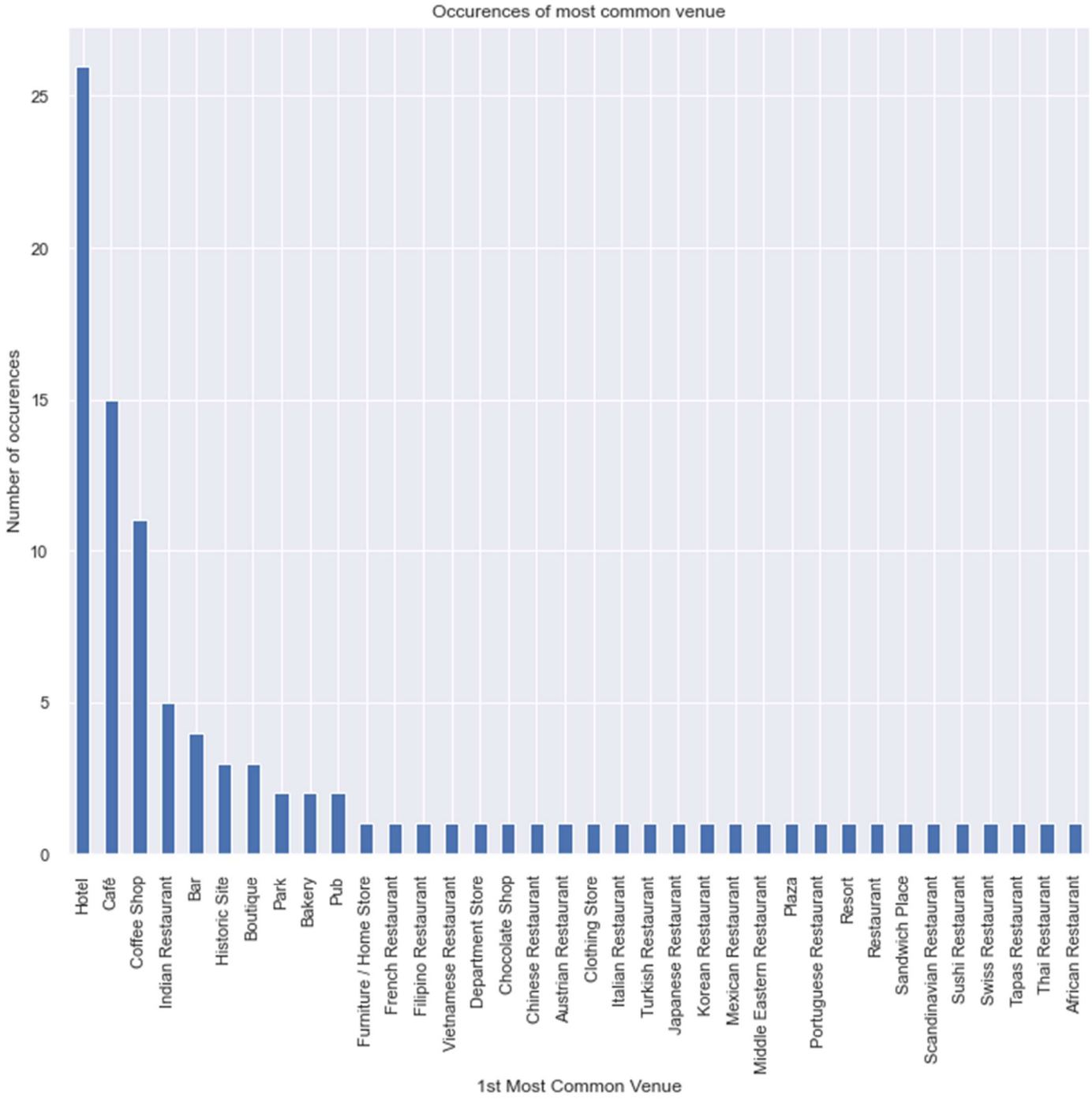
3.2 Arrivals vs Income

Plotting arrivals vs income, it appears there is a linear relationship between these two variables i.e. in order to drive more income from international visitors requires increasing the absolute number of arrivals.



3.3 1st most common venues

I plotted the most common occurrence for all the countries. "Hotel" appeared most often. This makes sense as these are common tourist destinations. Cafés and coffee shops also appear frequently.



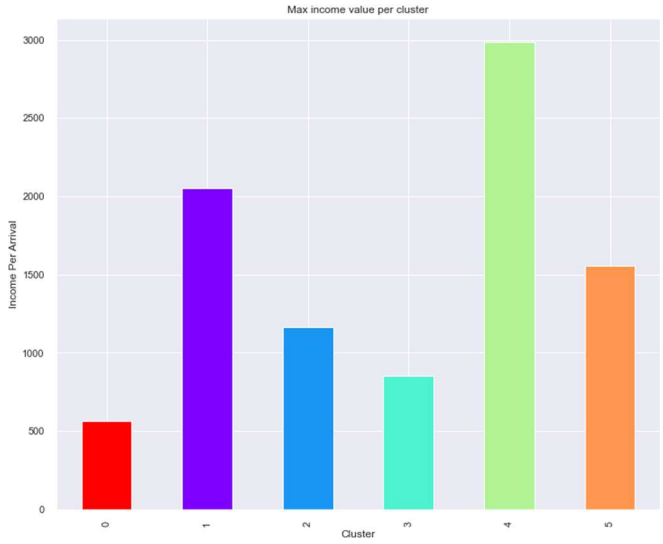
4 Methodology – Clustering with K-Means

For modelling I made use of unsupervised clustering using K-Means in order to group the cities. Analysis was done using the income for each country as well as the most common venues in each country based on data extracted using the Foursquare API. In both cases, the elbow method was used to determine the optimal number of clusters.

4.1 Clustering the countries based on income per arrival

I made use of the elbow method to determine that the optimal number of clusters for the data based on the income variable alone. This analysis suggested the optimal number of clusters was 6. We can visualise the calculated clusters on a world map.

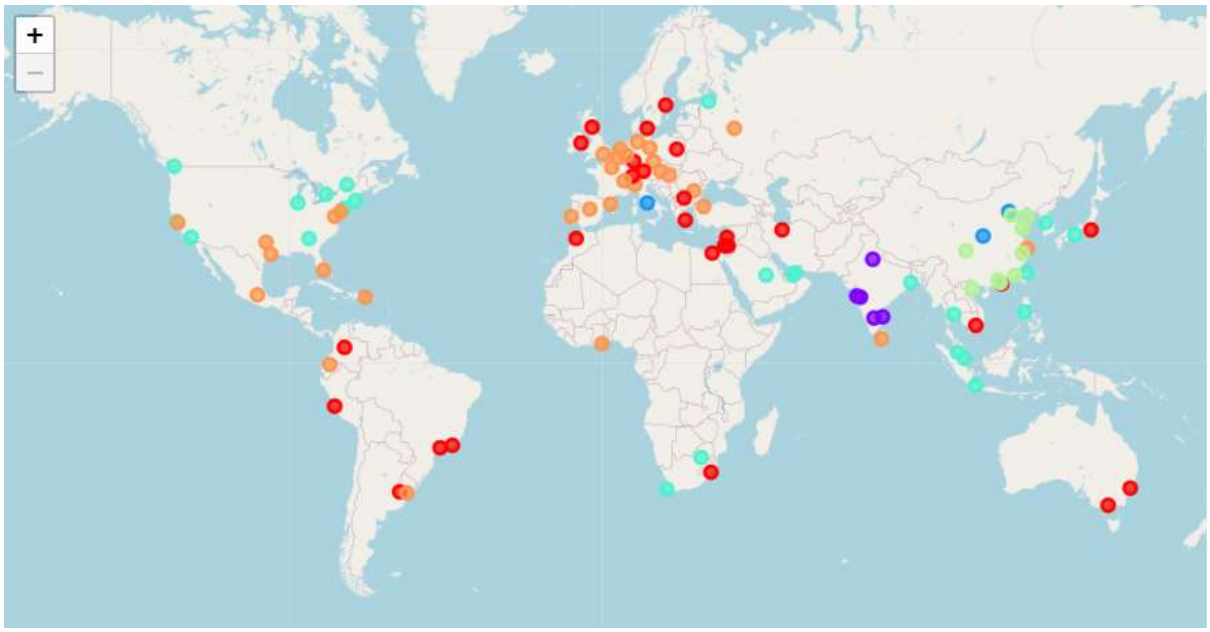
Taking a closer look at the composition of the clusters, cluster 4 is noteworthy as it contains just a single entry which is Tianjin in China. This is an outlier in the data as while the number of arrivals is relatively low the "Income Per Arrival" (IPA) is high for this city. Overall, the clusters for IPA are more spread out geographically compared to the absolute numbers of arrivals. This may suggest that spending patterns in certain cities are similar even if the absolute number of visitors is different, and therefore overall income, are quite different.



	Rank	City	Country	Arrivals	Income	Latitude	Longitude	Income Per Arrival	Cluster
94	96.000	Tianjin	China	750000.000	2.240	39.124	117.198	2986.667	4

4.2 Clustering the countries based on recommended venues

K-Means was used to group the cities based on the ten most common venues. Plotting this on the world map does seem to show some clusters concentrated in certain parts of the world.



Some of the generated clusters are quite small. For example, cluster 4 (the purple cluster on the map) appears only in India and appears to relate largely to cuisine in the city (i.e. Indian Restaurant).

composition of cluster 4 (purple cluster)

value	City	Cluster Labels	variable
Indian Restaurant	5	5	5
Hotel	4	4	4
Ice Cream Shop	3	3	3
Café	3	3	3
Chinese Restaurant	3	3	3
Coffee Shop	3	3	3
Sandwich Place	2	2	2
Fast Food Restaurant	2	2	2
Dessert Shop	2	2	2
Seafood Restaurant	2	2	2
Snack Place	2	2	2
Lounge	2	2	2
Middle Eastern Restaurant	1	1	1
Restaurant	1	1	1
South Indian Restaurant	1	1	1

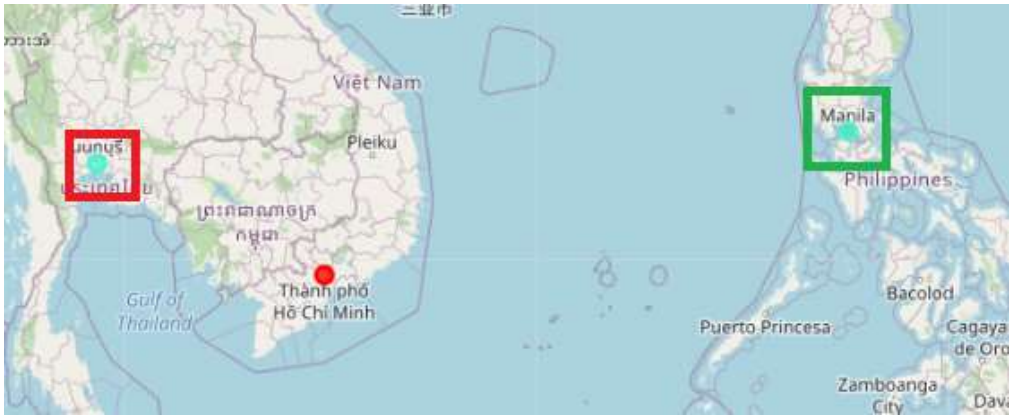
4.3 Analysing the venue clusters of the most-visited countries

Returning to our list of the top ten most-visited cities, 6 out of the 10 countries belong to the same cluster. This may suggest that the most-visited are in fact similar in terms of recommended venues.

	Rank	City	Cluster
0	1	Bangkok	3
0	2	London	0
0	3	Paris	5
0	4	Dubai	3
0	5	New York City	3
0	6	Singapore	3
0	7	Kuala Lumpur	3
0	8	Istanbul	5
0	9	Tokyo	0
0	10	Seoul	3

Looking at the created clusters, we come across cases where cities are similar in terms of income per arrival and nearby venues, but have very different values for arrivals. These are potentially cases where this similar destination can be recommended as a less-traveled-to location. Government of this city can also look to invest in infrastructure and find ways to further encourage tourism in the region.

Income per Arrival Clusters for Manila Philippines and Bangkok visualised



Nearby venues cluster for Manila Philippines and Bangkok visualised

Absolute number of arrivals for Manila Philippines and Bangkok visualised

5 Conclusion and Future directions

We've analysed popular travel destinations, focusing on number of visitors, money spent and similarity of each city based on recommended venues. We've see that there is a fairly linear relationship between number of arrivals and income earned and were able to cluster the cities based on average spend per arrival. We further split the 100 most visited cities based on an exploration of venues using the Foursquare API.

Based on the data retrieved and the clustering that was done, we saw that there is evidence to suggest that the most commonly visited cities are in fact similar in terms of the nearby venues generated using the Foursquare API. We also found that

In the future, further analysis can be done to find overlapping income and similarity clusters. For example, while spend in Australia is higher than in Durban, the cluster analysis suggests that these are similar regions. This may make Durban an attractive alternate destination for travelers in addition to "Manila" Philippines which was highlighted earlier in the previous sections. Additional metadata for these cities can also be obtained to dig deeper into what is similar and different about this destinations.