Personalized Travel Destination

Capstone Project Report

From: John Harrop

Date: 2020-04-09

Introduction

A hypothetical, but realistic problem is being investigated for the Capstone Project in the IBM Data Science Specialization on Coursera. This report is presented in conjunction with the supporting work that was conducted using Python in a Jupyter Notebook. A business problem is presented and then translated into a science/engineering question that a research and development team can address. Data and methods are selected to address this problem discussed by an interpretation and recommendation based on the findings.

Business Problem

A travel agency is interested in developing a niche market identifying less well-known destinations that fit with individual client's profiles. The concept would be used both to target marketing to their clients but would also be available to their agents when discussing options with clients. The company feels that there is a huge amount of information available on the Internet about under utilized, lower profile destinations. Based on feedback from their clients they feel that making these kind of connections could provide a significant competitive edge for their business.

Following this boardroom level discussion we have suggested that the scope of this initial test be to determine if destinations can be differentiated using open data sources. A positive answer to this question would be an important green light for further development. This characterization would need to systematically group destinations in a way that corresponds to clients likes (or dislikes). This is a complex problem that could benefit from a number of sources with different kinds of data (venues, topography, climate, etc.). An initial test of the concept has been proposed using only venue data.

Data

A set of locations were selected that represent a number of different regions and various size communities. Locations were chosen from regions in western Canada, Ireland, UK, Spain and Argentina. A fourth location is being considered. Each region is sampled in several locations that include cities, towns and villages. For cities (Vancouver, Dublin, Buenos Aires, Salta and Newcastle upon Tyne), multiple sampling centres are used but this is not possible for smaller locations. A mixture of well-known destinations that

attract tourists, industrial or commercial areas and smaller "off the beaten track" were included in the study. A total of 36 sites were used in the study.

Geographical coordinates (longitude and latitude) were determined for each location using Google Maps for input into the venue search engine. In addition, elevation, temperature (lowest monthly minimum, highest monthly maximum and annual average), monthly rainfall (minimum monthly, maximum monthly and total annual precipitation) and the population of the entire community (not just the venue search radius) was also determined. These are considered simple proxies for topography and climate. Due to the small number of locations in this study these parameters were determined manually from Wikipedia (population), FreeMapTools elevation finder (https://www.freemaptools.com/elevation-finder.htm) and Climate-Data.org was used for temperature and precipitation data. Two sites (Lahinch and Howth) did not have stations reporting on Climate-Data.org so the closest similar location was used. Some anomalous values were encountered in elevation data from FreeMapTools with elevation data that suggested a significant change from sea level in coastal locations when in fact that was known to not be the case.

Venue data was collected from searches on Foursquare

Topographic and climate data was assembled from the web sites in Excel and checked for consistency. This was time consuming but acceptable for the test case in this study, but would need to be replaced by web scraping versions in a production system.

Locations were selected primarily from Argentina (8), Canada (10) and Ireland (11) with a few for comparison from England (4) and Spain (3). The following list the locations by country. Names in bold are inland locations while the remainder are coastal. Some major cities have multiple locations (shown in brackets).

Argentina: Buenos Aires (3), **Mendoza**, **San Juan**, **Salta** (3)

Canada: **Kelowna**, **Smithers**, Vancouver (6), **Vernon**, Victoria

England: **Matlock**, Newcastle (3)

Ireland: **Athlone**, Cork, Dublin (3), Howth, Lahinch, **Limerick** (2), Waterford, Wexford

Spain: Huelva, **Madrid**, **Sevilla**

Methodology

The process of the study was iterative with locations being rejected and additional sites added during the study. Iterations also provided an opportunity to evaluate how well

the classification process was working. Several adjustments to the process were made after these evaluations.

No attempt was made to implement a cross validation method due to the relatively small number of locations used in the data (n=36). However, further studies should be based on larger data sets, which would warrant the use of cross validation.

When locations were tested on FourSquare some were found to contain less than five venues within a 500m radius. Search results with less than five venues were considered too thin to be useful in characterizing the location venues and diversity. Two solutions were considered: 1) increase the search radium until at least 5 venues were encountered, and 2) drop the low venue density locations and replace them. It seems reasonable that in a location with low density in venues you would naturally go further to find what you are looking for. But there were concerns that changing the search radius would result in other parameters being less easily compared. The second option was chosen and three locations were dropped and replaced by higher population locations.

Feature Engineering and Preparation

Location **population** and **elevation** data was noted to span several orders of magnitude. The reduce bias in this information both were log transformed and then normalized.

Feature engineering of the **climate** data and **venue** data was included in the data preparation. Climate data originally consisted of three measurements each for **temperature** and **precipitation**. Using mean temperature and the range between lowest minimum monthly temperature and highest maximum monthly temperature reduced this to two parameters. Both were then normalized. Similar reduction was done for the precipitation parameters.

Venue data was also summarized to diversity and density measurements. **Venue density** was calculated from number of venues returned in the search divided by 100 (the maximum number allowed to return). This resulted in a range from 0.05 to 1.00 and does not need normalization. **Venue diversity** was calculated by dividing the number of unique kinds of venues by the number of venues returned. A low value of 0.05 would result from only 1 type of venue in 100 search results, while 1.00 would indicate all of the venues returned in the search were different types. Although this is not perfectly representing diversity across different populations sizes (and venue densities) it was considered a reasonable estimate and also did not need normalization.

Venue types were converted to proportion by type, which results in the sum of proportions of all types being 1. These values were not normalized since they function as a related set of values and that meaning would be changed by individually normalizing each parameter. In addition, the values fall within 0 to 1 range and should not result in any scale bias. Further discussion of this part of the data is in the *Dimensional Reduction and Cluster Analysis* heading.

EDA – Or Lack Of…

Previous work with Foursquare had provided experience with venue data generated by the system and EDA of venue data was not used. Instead, the venue diversity and density indices were examined. Only four locations (Buenos Aires1, Buenos Aires2, Cork and Mendoza) returned the maximum 100 venues. The locations with the greatest venue diversity (Newcastle3 and Smithers) also have some of the lowest counts of venues returned by the search. Although improvements for venue characterization parameters can be made, the data appears to be sufficient for the scope of this study. Climate, topography and population parameters appear to cover the ranges they were selected to represent.

K-means and PCA

K-means was selected at the primary method of generating groups or clusters of locations. Selection of how many clusters included the use elbow method. The initial pass used 214 parameters. Following the K-means process a PCA was also done to evaluate how the clusters plot on the first two principal components. This has been useful in the past on other projects and is essentially an iterative use of PCA as an EDA or visualization tool that can assist in interpreting results.

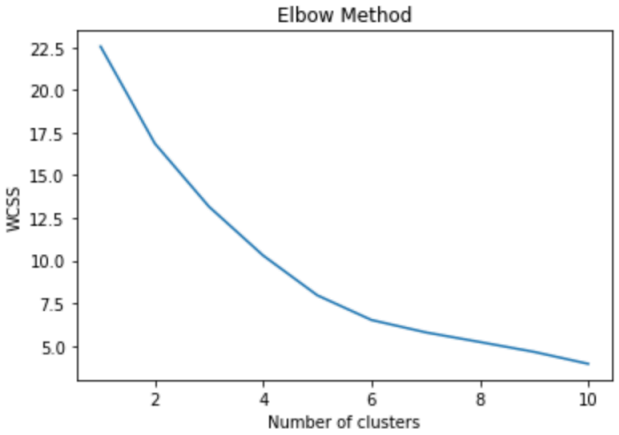
Dimensional Reduction and Cluster Analysis

The large number of venue categories was reduced using PCA from 206 unique categories. A scree plot was inspected to view how many components contribute significant information.

Results

First Pass

The elbow plot for K-means clustering for k=1 to 10 does not show a strong break in slope that could be used to select an optimal number of clusters. Perhaps there is a weak break at k=6, but that is very subtle. Based on the interpretability and size of the test data k=5 was selected.



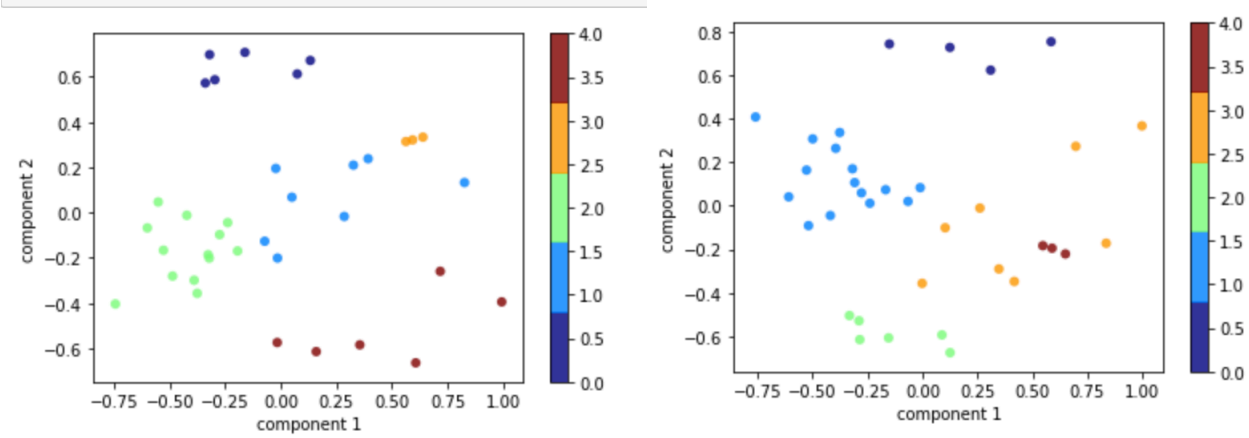
The flowing table of the clusters defined by K-means in the first pass has yellow highlights for the locations that moved to a different cluster in the second pass. In addition, locations that are inland are shown in bold. The increasing prevalence of inland locations to the right side of the table is clear.

0	1	2	3	4
Vancouver1	Buenos Aires1	**Athlone**	**Salta1**	**Kelowna**
Vancouver2	Buenos Aires2	Dublin2	**Salta2**	**Madrid**
Vancouver3	Buenos Aires3	Dublin3	**Salta3**	**Mendoza**
Vancouver4	Cork	Howth		**San Juan**
Vancouver5	Dublin1	Lahinch		**Smithers**
Vancouver6	Huelva	**Limerick1**		**Vernon**
	Newcastle1	**Limerick2**		
	Sevilla	**Matlock**		
		Newcastle2		
		Newcastle3		
		Victoria		
		Waterford		
		Wexford		

PCA was performed on the same data used for K-means clustering with the first two components accounting for 20.0% and 15.3% of the variance respectively. Inspection of the eigenvectors to understand the loadings indicated that their major influences are:

PC1: venue density and homogeneity, average and range of temperature and to a lesser extent elevation.

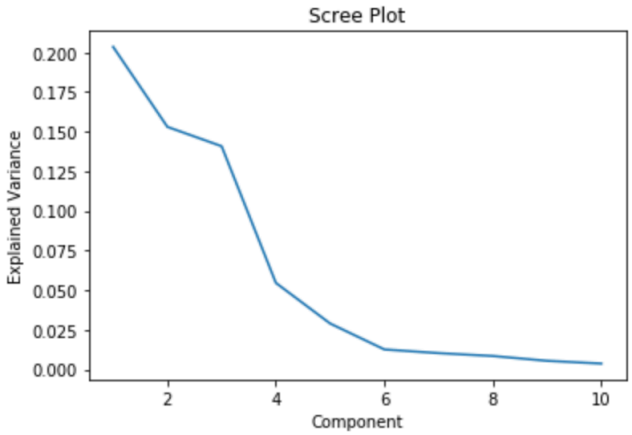
PC2: strongly influenced by precipitation range more than precipitation total and to a lesser extent by venue and average temperature parameters.



Clustering by K-means shows coherence when plotted on a PC1 versus PC2 and assists in interpreting the meaning of clusters. (The left side plot uses homogeneity of venues while the right side uses diversity.)

Second Pass

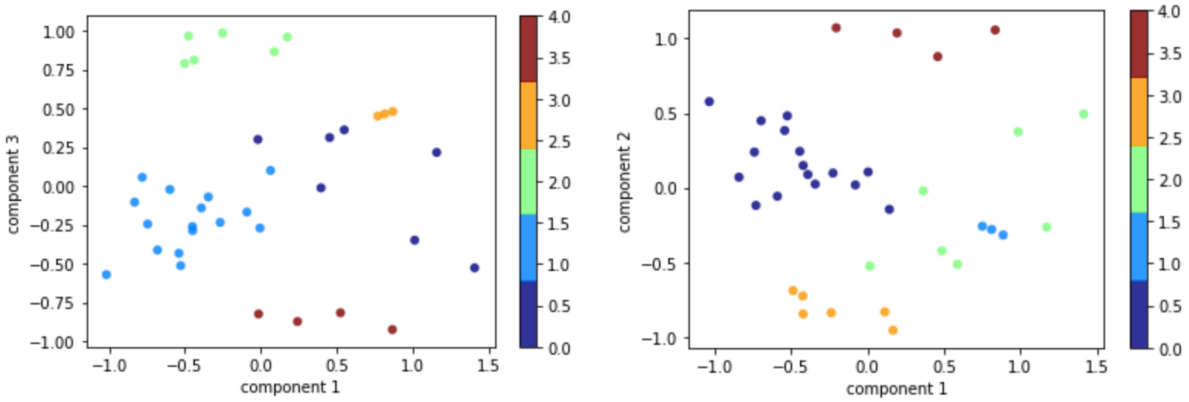
Several adjustments were made to determine the sensitivity of the methods used. One result was that changing venue diversity to homogeneity (homogeneity is 1 – diversity) affected changes to K-means clusters between the with homogeneity supporting what was interepreted to be more meaningful clustering. The reason for this is not obvious. Plots and clusters presented have been generated using the homogeneity rather than diversity index.



A scree plot of explained variance shows that at 6 the contribution of additional components do not contribute much and are likely reflecting forms of noise. The venue data was transformed and the first 10 components used to replace the original venue data in the second pass. This was done to determine how much the consolidation of venue information through dimensional reduction would affect the outcome of K-means clustering. The changes were not large and the five locations that changed cluster presumably indicate affinity for both groups.

0 (2)	1 (0)	2 (1)	3 (3)	4 (4)
Vancouver1	Buenos Aires1	**Athlone**	**Salta1**	**Kelowna**
Vancouver2	Buenos Aires2	Buenos Aires3	**Salta2**	**San Juan**
Vancouver3	Cork	Dublin1	**Salta3**	**Smithers**
Vancouver4	Huelva	Dublin2		**Vernon**
Vancouver5	**Madrid**	Dublin3		
Vancouver6	**Mendoza**	Howth		
	Sevilla	Lahinch		
		Limerick1		
		Limerick2		
		Matlock		
		Newcastle1		
		Newcastle2		
		Newcastle3		
		Victoria		
		Waterford		
		Wexford		

Unfortunately cluster labels changed between the two passes. The label numbers given in brackets are the new labels reflected below in the PC1 versus PC2 scatter plot. In general, the clusters have not changed much.



The left side plot uses homogeneity of venues while the right side uses diversity. Although the right side plot does show better clustering, only one location changed cluster – so even though the diversity clustering is probably better than homogeneity we will discuss the significance of the cluster changes using homogeneity.

Note that after switching to PCA reduction of venue data the variance explained by PC1 and PC2 to 39.2% and 30.6% respectively.

Discussion

Cluster Meaning

Identifying meaning for the clusters is important if there is any potential to recommend further development of this approach to recommending travel locations to clients. We will start by looking at the clusters generated in the first pass.

First Pass Clusters

Review of the PCA scatter plot and the contributions to the first and second components suggests that the y-axis is reflecting climate, especially precipitation. The lower end of the axis reflect dry and little range in monthly precipitation. The lowest cluster is dryer, even desert locations while the upper cluster is the wettest with the most variation in precipitation. The middle group moderate climate and clusters separate based on changes in venue density and diversity.

Label 0: all Vancouver locations were grouped here. Locations around Vancouver included business/shopping downtown, tourist downtown, a popular beach area in an urban residential area, a beach on the southern edge of greater Vancouver and two communities on the eastern side of greater Vancouver with recently gentrified and

strongly Asian immigrant characteristics respectively. These cover locations that would be seen as quite different within the context of Vancouver, but the climate overrides all these differences to isolate the cluster. Compare with label 3 which is an Argentine city where the three locations chosen for different social status and business differences do not differentiate very much. All Vancouver locations are considered coastal.

Label 1: this cluster is dominated by coastal locations and is a mixture of Spanish and a few Irish and English sites. Coastal locations dominate this group. Note that locations in Buenos Aires range from the central tourist and shopping district, more Argentine hotel and park area in Palermo and La Boca (centred beside the Boca Juniors stadium). All three of these group together. But Newcastle and Dublin locations split between two groups suggesting perhaps that there is grater variation in diversity characteristics for these cities (since climate, population and elevation do not change between locations).

Label 2: mixed inland and coastal locations with moderate climate and Irish or English cultural affiliation. Interestingly, Victoria a coastal city in Canada not far from Vancouver associates with this group. Victoria is proud of its English heritage and sometime is described as more *English than England*.

Label 3: as previously mentioned Salta groups very tightly in the space of PC1 and PC2. This likely has created a slightly artificial group which had it been a single sample would have grouped with 1 and further increased its Spanish cultural affiliation.

Label 4: inland locations with warm and typically dry climate. Locations have equally Canadian and Spanish cultural affiliation. This cultural range is reflected in the linear and dispersed shape of the cluster. In fact, this is the most widely dispersed of the clusters.

Second Pass Cluster Changes

There were no dramatic changes between the clustering generated by the first and second pass. The primary purpose of the second pass was to evaluate if dimensional reduction of the venue data by PCA would result in stronger classification of locations by venue results, which are interpreted to reflect more cultural characteristics than the climate, topography and population parameters can. The changes can be summarized as two reassignments:

1) Madrid and Mendoza moved from label 4 (inland, mixed Canadian/Spanish *culture*) to label 1 (mostly coastal, mixed culture).
2) Buenos Aires 3, Dublin 1 and Newcastle 1 moved from label 1 (mostly coastal, mixed culture) to label 2 (mixed coastal/inland, Irish and English culture).

This does appear to strengthen the cultural component of classification. Label 4 continues to be strongly inland locations but is mainly Canadian. San Juan, in Argentina, is interesting since it was rebuilt after a devastating earthquake in 1944 thus making it a *new* city and as such potentially more similar to Canadian cities.

The movement of Madrid and Mendoza to label 1 combined with the removal of Newcastle 1 and Dublin 1 from that group significantly strengthens the Spanish cultural component of label 1. That leaves Cork as an interesting anomaly in that group.

Dublin 1 and Newcastle 1 moving to label 2 unites them with the rest of their city's locations, and is consistent with a strongly Irish and English signature to that group. But Buenos Aires 1 (the shopping and tourist area) moving to this group is an anomaly.

In conclusion, labels 1 and 4 are primarily controlled by climate. Labels 1 and 3 can logically combine into a Spanish cultural influence group while label 2 is strongly identified with Irish and English culture.

FourSquare Limitations

A couple of comments about FourSquare as a data source are worth mentioning in case it is used in future work. Variability in venue reporting choices by users and the density of FourSquare users must contribute to variability in information density in different countries. Perhaps a scale factor could be used to compensate for this to reduce venue density bias for those places better covered by FourSquare.

Diversity and count by category may be reducing the influence of some similar types of venues. Consolidation of very similar venues (for example: Beer Bar, Bar and Pub) would help reduce this. Consolidation of venue results in data preparation and feature engineering could improve this. Foursquare may already support reporting total numbers in broad categories (for example *all restaurants*). This would enable conversion into proportion of restaurants by type rather than restaurant type as a proportion of total venues.

Conclusions

Initial results suggest that FourSquare venue data combined with supplemental climate and topographic data can classify locations into groups with important similarities.

Further work with this is warranted and should include 1) a much larger set of locations, 2) preferred locations from a number of clients, 3) cross validation should be implemented, 4) further work identifying feature engineering of FourSquare data.

PCA (or another dimensional reduction method) should continue to be used to strengthen the cultural component of the clustering

Lastly, switch to doing this in R Studio and R Markdown, my experience with those tools is that they produce better professional quality reports than Jupyter Notebooks.