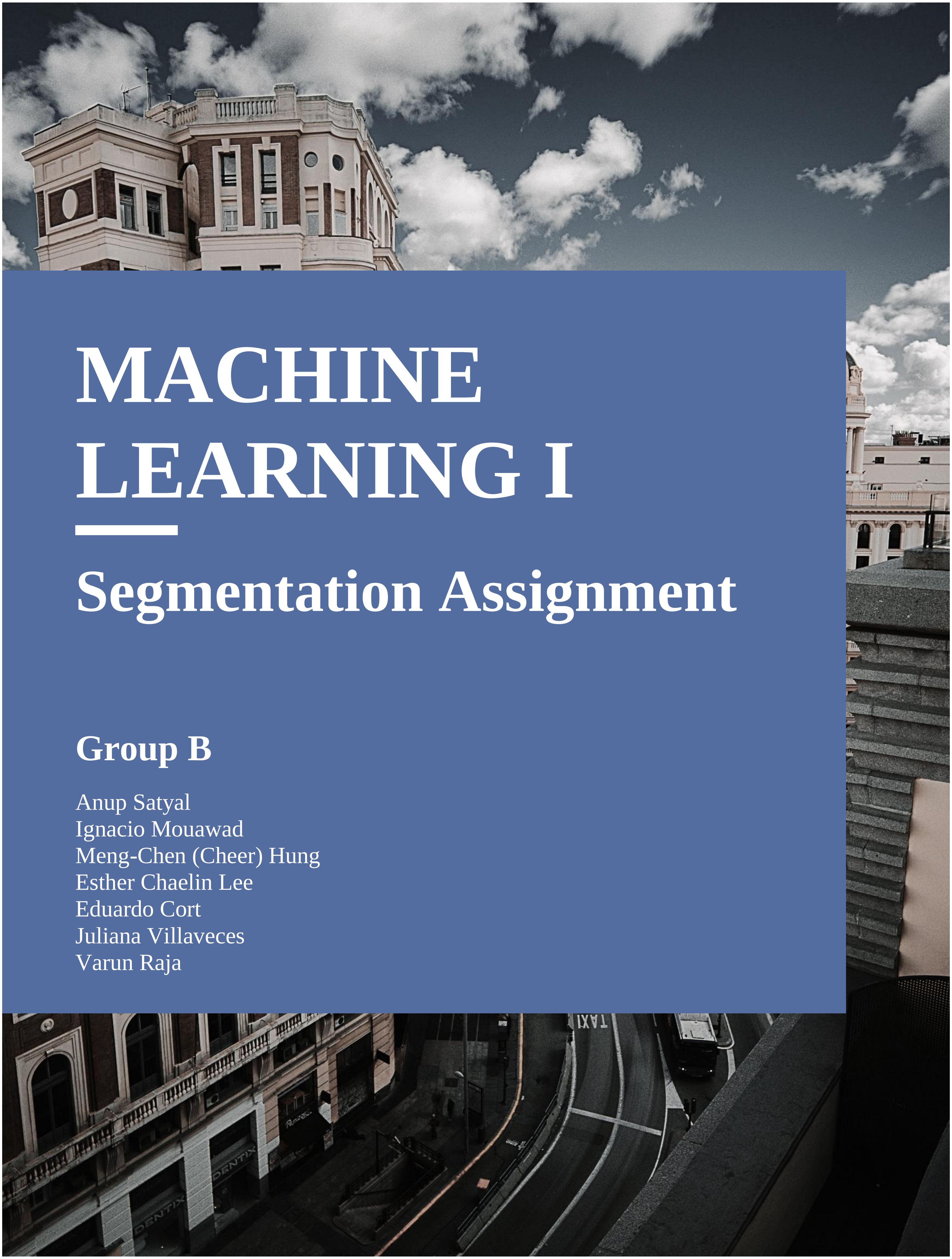


MACHINE LEARNING I

Segmentation Assignment

Group B

Anup Satyal
Ignacio Mouawad
Meng-Chen (Cheer) Hung
Esther Chaelin Lee
Eduardo Cort
Juliana Villaveces
Varun Raja

Executive Summary

a K-mean unsupervised clustering model, 52 provinces in Spain were divided into three clusters. These clusters, which are labelled as Banking and Industrial Stronghold, Tourist Metropolitans and Upcoming Cosmopolitans, have some defining characteristics.

The results show that the majority of provinces in Spain are Banking and Industrial strongholds with low populations (533,000 on average vs 2,153,500 in others), of which a significant portion are local (91.15% vs 84.81% in the other clusters) and have relatively lower unemployment rate (6.8% in this cluster on average vs 8.85% in the others).

By contrast, the Tourist Metropolitans hold the largest provinces in the Country, namely Madrid and Barcelona. Perhaps surprisingly, rather than being completely separated from the rest of Spain, these provinces are in fact similar to other tourist destinations like Las Palmas.

The Upcoming Cosmopolitans can be currently classified as Tier 2 tourist destinations, but have a strong growth in population that might indicate that the economy in these provinces is improving.

The report briefly discusses the approach followed in getting to these conclusions. It concludes with policy recommendations that the Ministry of Economy may follow to improve the overall economic situation of each of the clusters. These policies have the potential to help diversify the economy of the country as a whole.

Approach

MODEL SELECTION AND DATA PREPARATION

An unsupervised K-mean model, which aggregates data points based on similarities, was chosen for the problem. Prior to training the model, certain data preparation processes were carried out to fit its needs. For instance, the Ministry understands well that people tend to migrate where they perceive there are better opportunities. All else being equal, population rarely determines the economic growth of a province. Rather, economic growth determines the increase or decrease in the population. Thus, population can be classified as a profiling variable as it assists the interpretation of the clusters provided by the chosen model. During the preparation phase, in addition to the population, the percentage of males, foreigners and unemployment rates of each of the provinces were chosen as profiling variables.

Once the profiling variables were defined and separated, an analysis was run on the remaining variables to identify any correlations. Since correlations are essentially the interrelationship between variables, if two or more variables are correlated with each other, it is virtually impossible to separate the impact of each individual variable in determining the clusters. For example, in the dataset provided by the Ministry, the wholesale index shares a relatively high 66% correlation with the percentage of GDP contributed by the textile sector. If both variables were included in the model, it would have been difficult to explain which variable contributed more towards the differentiation. With an initial correlation threshold of 60%, one of any two highly correlated variables were removed from the dataset to be used for the modelling process. With a reduced threshold of 50%, an additional reduction process was carried out. The remaining variables, which are shown on a matrix (Exhibit 1), had the lowest correlation (interrelationship) with the other variables and were used to train the model.

Analysis

Once the training phase was completed, a model with a *silhouette* ratio of 0.55 was chosen. In essence, a silhouette ratio (ranging from 0 to 1) indicates the similarities within and differences across clusters. A higher ratio would indicate high similarity within and high difference across the clusters. The chosen model divided the provinces into three clusters with a ratio of 40:6:6. Based on the differentiating factors, the clusters were labelled as follows:

a) Banking and Industrial Strongholds: This cluster includes 40 provinces with an average contribution of 22.6% of GDP by the banking sector, 4% higher than the other two clusters. The 40 provinces also have, on average, 21.75% of GDP coming from the energy sector, 86.38% higher than the other two clusters and 10.2% of GDP coming from inter-industry trade, 32.99% higher than the other two clusters.

b) Tourist Metropolitans: With an average tourism index of 9,437, compared to 1,918 for the other two clusters, the 6 provinces in this cluster generate very high tourist interest. This cluster also accounts for 35% of the total population of the country and has, on average 16.4% foreigners. It is interesting to note that provinces in this cluster generate, on average, 16.17% of its GDP from the textile industry, which is 129% higher than that in the Banking and Industrial Strongholds and 20% higher than the Upcoming Cosmopolitans.

c) Upcoming Cosmopolitans: While this cluster of 6 provinces also attracts tourists, the average index value of 3,238 indicates that the interest is not as strong as for Tourist Metropolitans. The provinces in this city, which host historical sites such as “Plaza de España” in Seville and “La Ciudad de las Artes y las Ciencias” in Valencia, may be more interesting to the culturally inclined but perhaps not as interesting as the beaches of Barcelona. In that sense, this cluster could be classified as Tier 2 tourist destinations that may not have as much thriving tourism industry as the Tourist Metropolitans. The most differentiating factor for this cluster; however, is population growth, which averaged 2.3% between 2004-2009 compared to 1.63% for the other two clusters. This might be an indication of a growing economy in this cluster of provinces, which requires further investigation in part of the Ministry.

Recommendations

The results of our analysis indicate that the tourism industry is clearly a differentiating factor amongst the provinces in Spain. Based on the evidence provided by the analysis, some specific policy recommendations that may assist in the betterment of the economic situation of each of the clusters are as follows:

In the analysis, it was evident that the wholesale, retail and restaurant indices, with a 72% positive correlation with the tourism index, were highly correlated. While correlation does not imply causation, it is recommended that the Ministry find ways to leverage the impact of tourism in the Tourist Metropolitans and Upcoming Cosmopolitans clusters to improve the overall economic situation of the provinces.

If it is found to be the case that tourism can indeed be leveraged to increase the economic contribution of wholesale, retail and restaurant industries, the Ministry should take extra caution not to lead to a situation where the Banking and Industrial Strongholds suffer. This recommendation is supported by the fact that the Banking sector is negatively correlated with wholesale, retail and restaurants. In order to counteract the impact of the policies being put in place in Tourist Metropolitans and Upcoming Cosmopolitans clusters, the agriculture and building industries may be supported as these two industries have a positive correlation of 61% and 58% respectively with the Banking sector.

On the one hand, the Banking and Industrial stronghold has a well-developed mining industry that contributes 4.75% of the GDP of the provinces on average, which is 44% higher than the average of the provinces in the other two clusters. On the other hand, the Tourist Metropolitans and Upcoming Cosmopolitans clusters account for 56% of the population of the entire country. This is not considering the growing population of the Upcoming Cosmopolitans, which will contribute to a higher share of the clusters driven by the Tourism industry. In the unfortunate event of an economic slowdown, 56% of the country's population will have difficulties as the tourism industry is positively correlated (68%) with overall economic activity (economic activity index). However, mining has a negative, albeit low, correlation with overall economic activity.

By focusing on possible policy initiatives in the mining industry that may lead to economic equilibrium, the Ministry might not only support the Banking and Industrial Strongholds, it may also find ways to diversify the economy away from its overreliance on tourism.

This report has only highlighted the similarities between the various provinces in Spain. It is recommended that the ministry look further into the cause-effect relationships between the various factors in order to optimize policies.

Appendix

EXHIBIT 1. CORRELATION HEATMAP

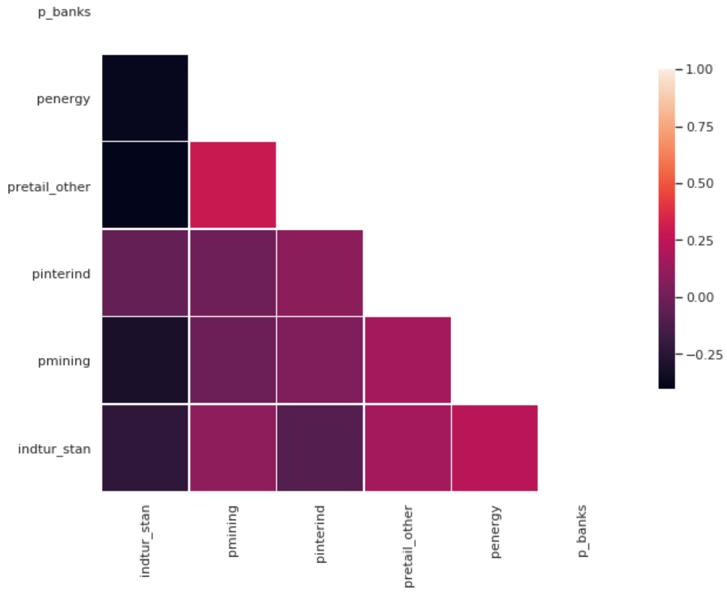


EXHIBIT 2. CORRELATION HEATMAP

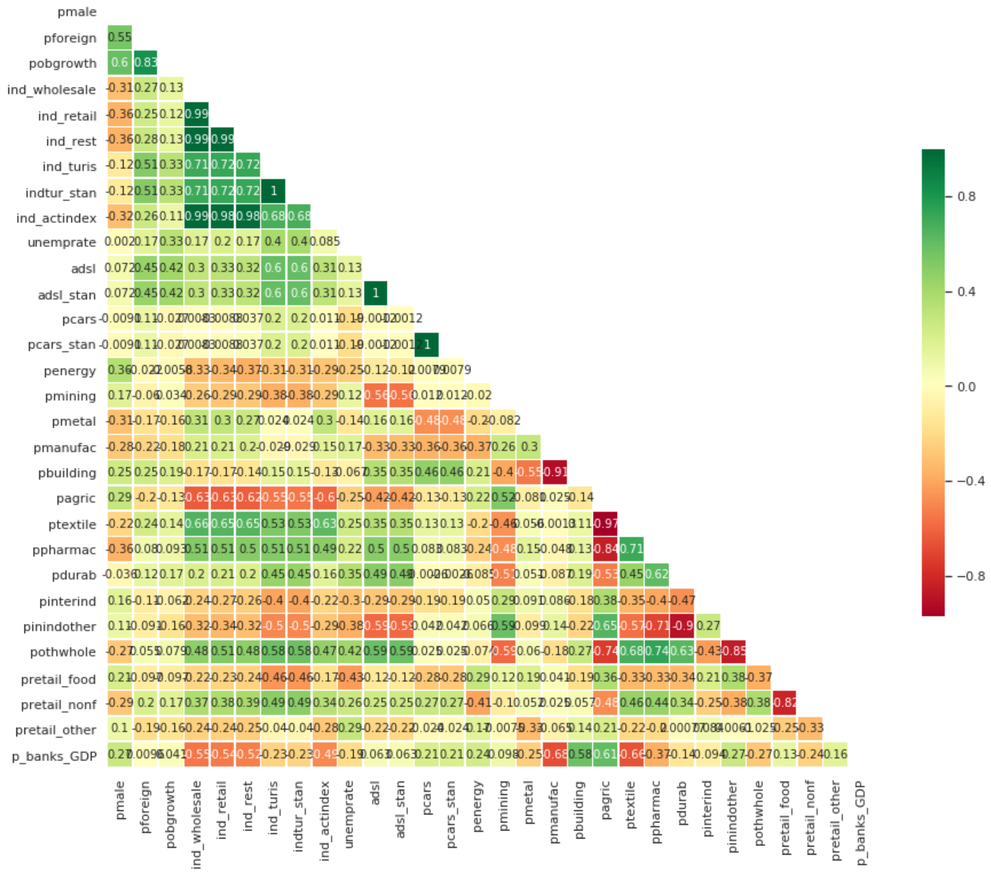


EXHIBIT 3. CLUSTER MAP

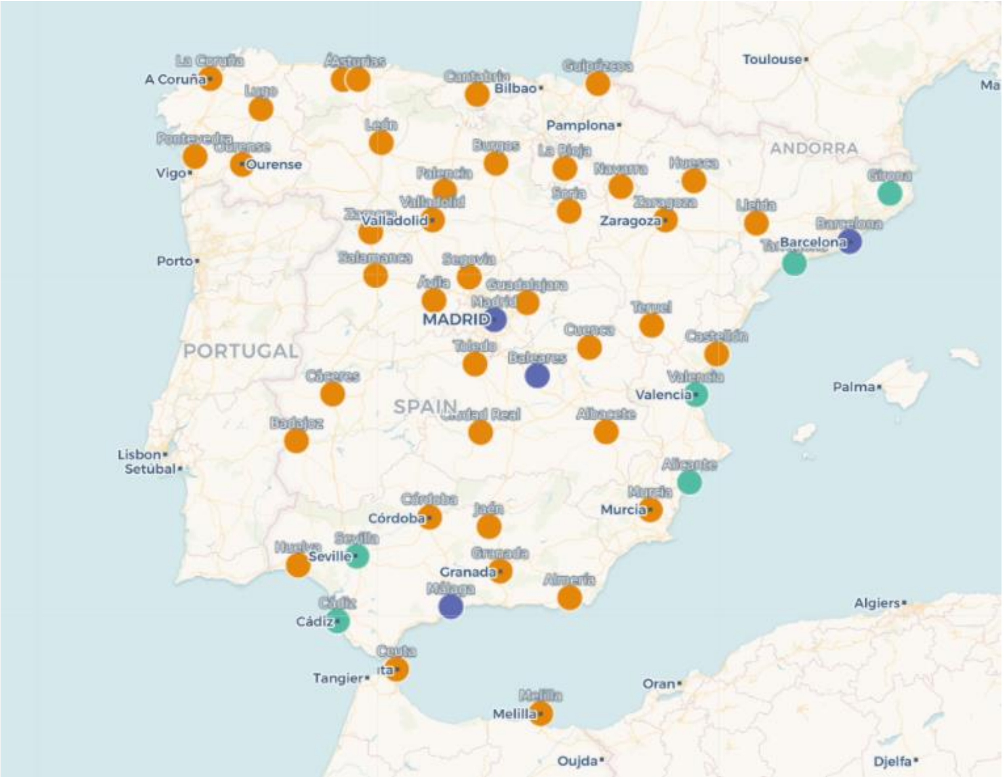


EXHIBIT 4. AVERAGE TOURISM INDEX BY CLUSTER

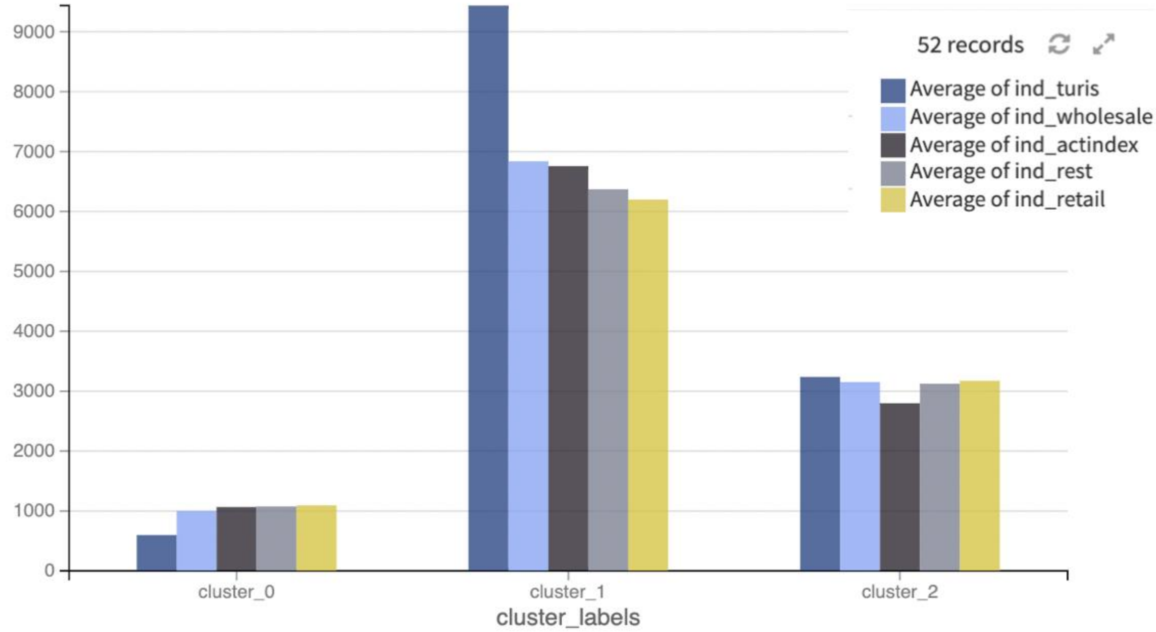


EXHIBIT 5

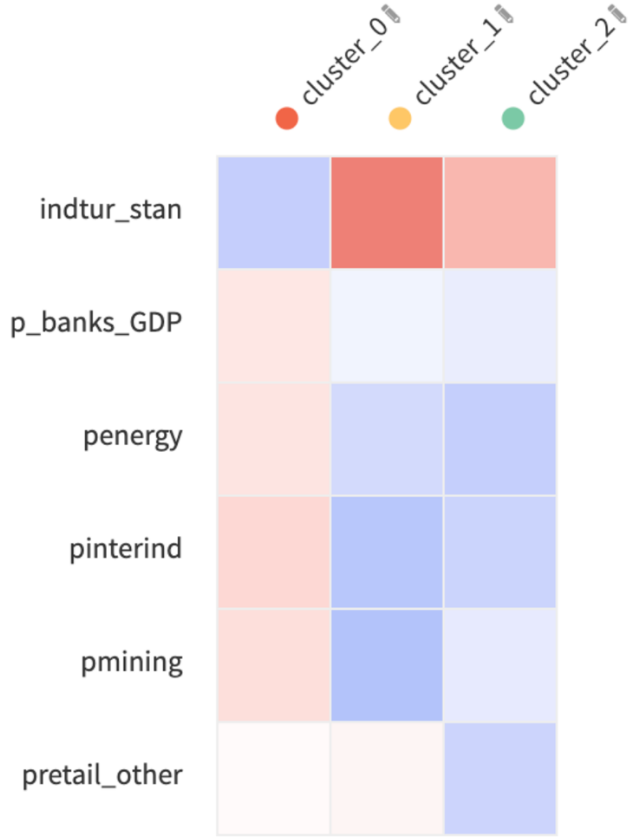
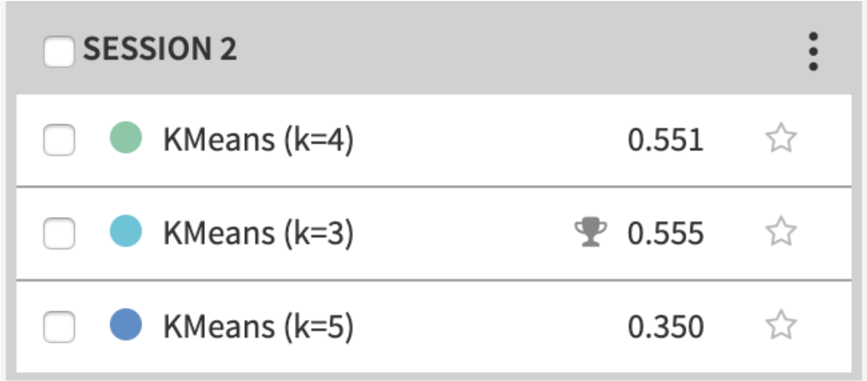


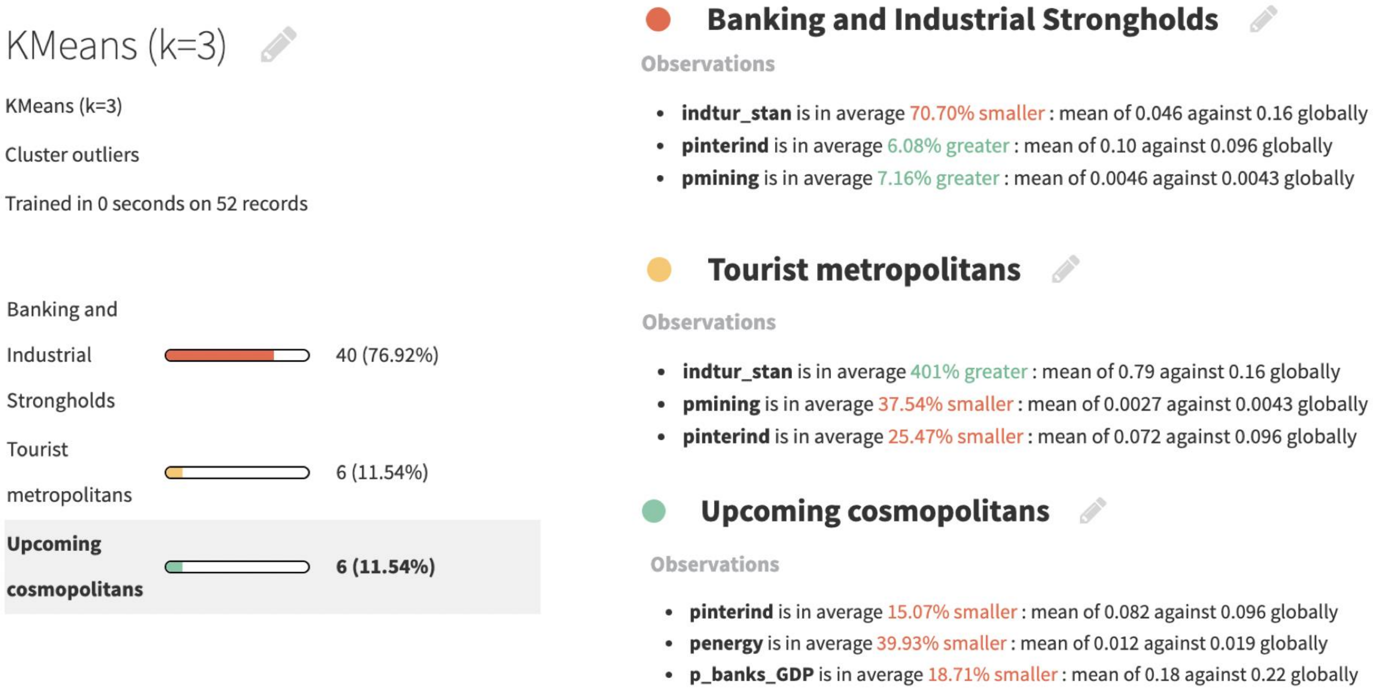
EXHIBIT 6



□ SESSION 2

□	●	KMeans (k=4)	0.551	☆
□	●	KMeans (k=3)	🏆 0.555	☆
□	●	KMeans (k=5)	0.350	☆

EXHIBIT 7



KMeans (k=3)

KMeans (k=3)

Cluster outliers

Trained in 0 seconds on 52 records

Banking and Industrial Strongholds	40 (76.92%)
Tourist metropolitans	6 (11.54%)
Upcoming cosmopolitans	**6 (11.54%)**

● **Banking and Industrial Strongholds**

Observations

- **indtur_stan** is in average 70.70% smaller : mean of 0.046 against 0.16 globally
- **pinterind** is in average 6.08% greater : mean of 0.10 against 0.096 globally
- **pmining** is in average 7.16% greater : mean of 0.0046 against 0.0043 globally

● **Tourist metropolitans**

Observations

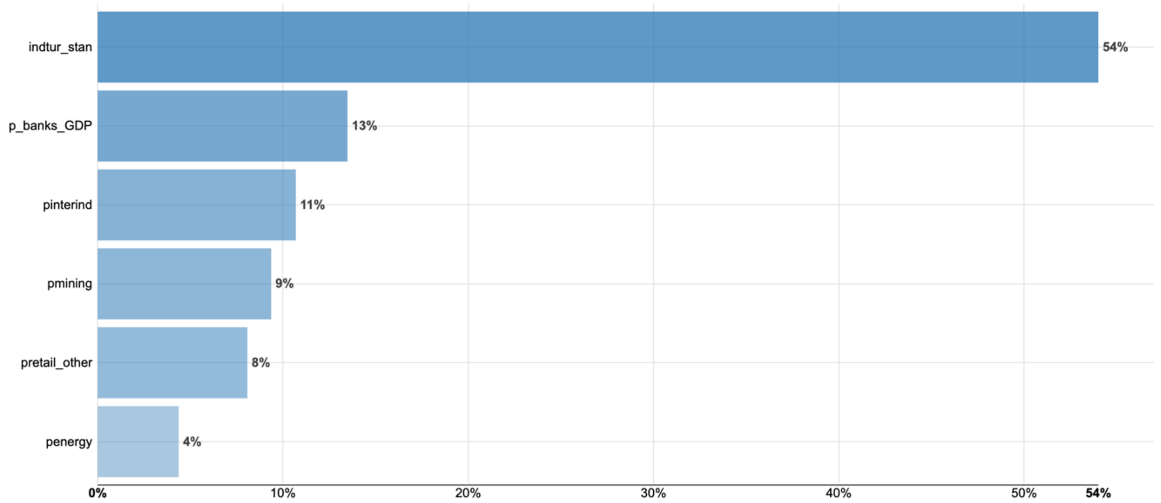
- **indtur_stan** is in average 401% greater : mean of 0.79 against 0.16 globally
- **pmining** is in average 37.54% smaller : mean of 0.0027 against 0.0043 globally
- **pinterind** is in average 25.47% smaller : mean of 0.072 against 0.096 globally

● **Upcoming cosmopolitans**

Observations

- **pinterind** is in average 15.07% smaller : mean of 0.082 against 0.096 globally
- **penergy** is in average 39.93% smaller : mean of 0.012 against 0.019 globally
- **p_banks_GDP** is in average 18.71% smaller : mean of 0.18 against 0.22 globally

EXHIBIT 8


indtur_stan
p_banks_GDP
pinterind
pmining
pretail_other
penergy