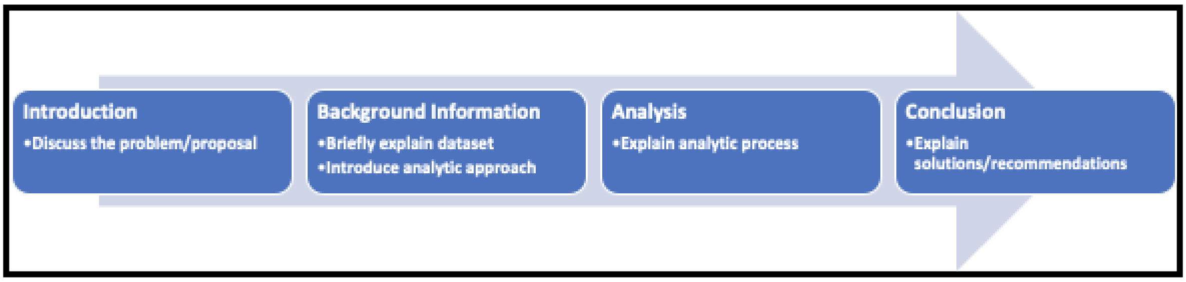
Executive Summary: Segmentation

Introduction

The Ministry of Economy in Spain has requested an analysis of all 52 provinces in Spain for an unknown year. In response, Team A from IE University School of Human Sciences and Technology have summarized their analysis of this dataset. Team A hopes that The Ministry of Economy in Spain reviews the analysis and strongly considers the resources provided to them prior to moving forward with a set of policies. Here is an illustration of the contents of this Executive Summary:



Introduction
- Discuss the problem/proposal

Background Information
- Briefly explain dataset
- Introduce analytic approach

Analysis
- Explain analytic process

Conclusion
- Explain solutions/recommendations

Background Information

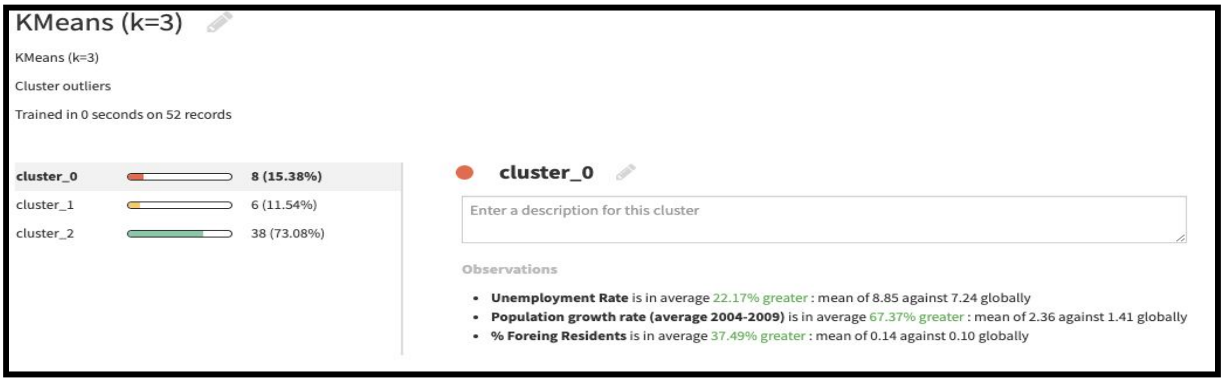
In an effort to help The Ministry of Economy in Spain learn which provinces are similar, the team analyzed data containing the economic and demographic data from all 52 provinces in Spain for an unknown year. This dataset includes economic information, such as indexes on retail and unemployment data, as well as demographic information, such as % Male and % of Foreign Residents. For a comprehensive repository of information about the economic and demographic data of all provinces, please refer to **Table 1: Data Dictionary** in the appendix.

The specific analysis used to understand which provinces are similar is called **Segmentation Analysis**. Segmentation Analysis allows groups to be created by analyzing the variables of individual items. In this case, because the team has economic and demographic data of each province (individual items) in Spain, the team can create groups by analyzing the economic and demographic characteristics of each province. More specifically, the team has leveraged **K-Means Clustering** to create groups. At a high-level, K-Means is the most simple clustering algorithm that identifies homogeneous subgroups **k** within a dataset.

Analysis

Design

After importing the data, cleaning the data, and completing an exploratory analysis of the data, the team decided to use most of the default settings for the K-Means model. The team did not perform dimensionality reduction or rescale the data. Although the team started with 5 clusters, there wasn't enough reason to keep 5 clusters given that some clusters had few provinces. As a result, 3 clusters were chosen - each with sufficient provinces - using the variables yielding the highest Silhouette.



KMeans (k=3)

KMeans (k=3)

Cluster outliers

Trained in 0 seconds on 52 records

cluster_0	8 (15.38%)
cluster_1	6 (11.54%)
cluster_2	38 (73.08%)

cluster_0

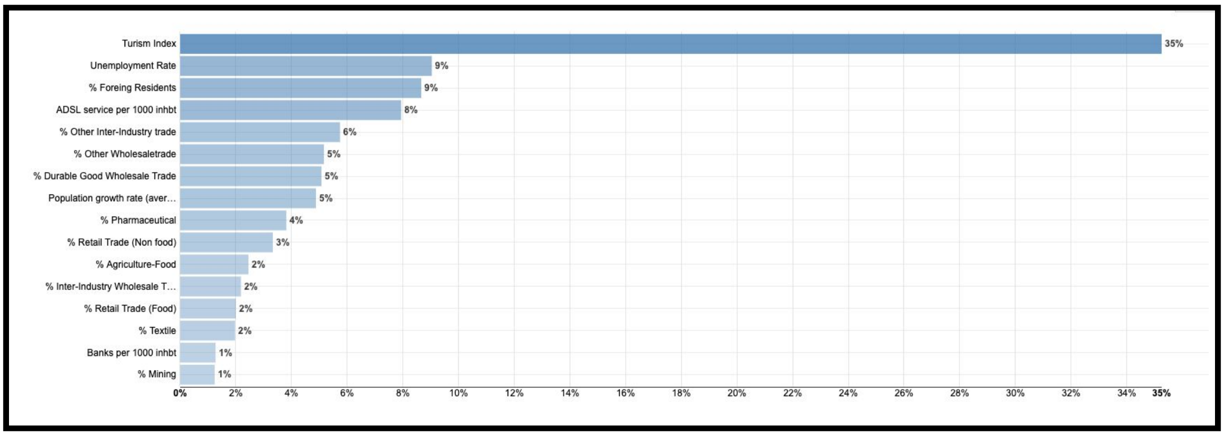
Enter a description for this cluster

Observations

- **Unemployment Rate** is in average 22.17% greater : mean of 8.85 against 7.24 globally
- **Population growth rate (average 2004-2009)** is in average 67.37% greater : mean of 2.36 against 1.41 globally
- **% Foreing Residents** is in average 37.49% greater : mean of 0.14 against 0.10 globally

Results

Unsurprisingly, the variables with the highest importance included **tourism, unemployment, and foreign residency** - these three variables make up more than 50% of the variability among the clusters. Given how popular of a travel destination some cities in Spain are, the team felt comfortable with the variables importance. Additionally, the provinces that shared clusters aligned with teams' basic knowledge about the provinces, especially Madrid and Barcelona. For example, when analyzing the relationship between tourism and % of foreign residents, the cities one would expect are frequented more often are indeed separate from those cities that are not frequented often.



Evaluation Metrics

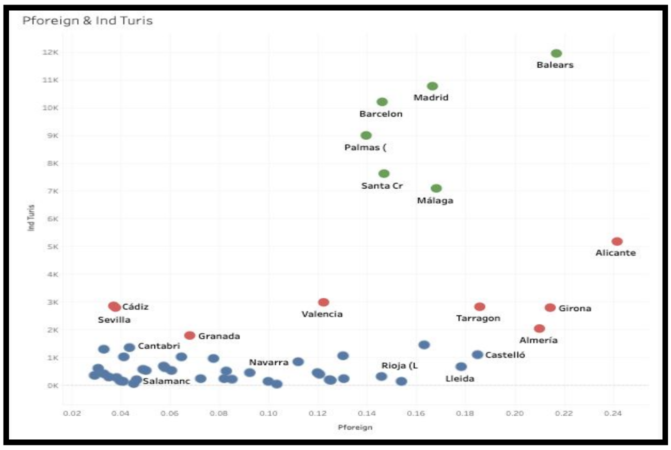
Using the evaluation metrics native to Dataiku, the team learned about the quality of the model. With a **Silhouette** of roughly .74, the model shows a high degree of separation between the clusters. To have good clusters, a model should have a Silhouette Value as close to 1 as possible. Furthermore, the model has an **Inertia** of roughly 3.05e+7 meaning there is much dispersion within the clusters. In summary, the model shows that although there is a high degree of separation between the clusters, there is also much dispersion within the clusters.

Conclusion

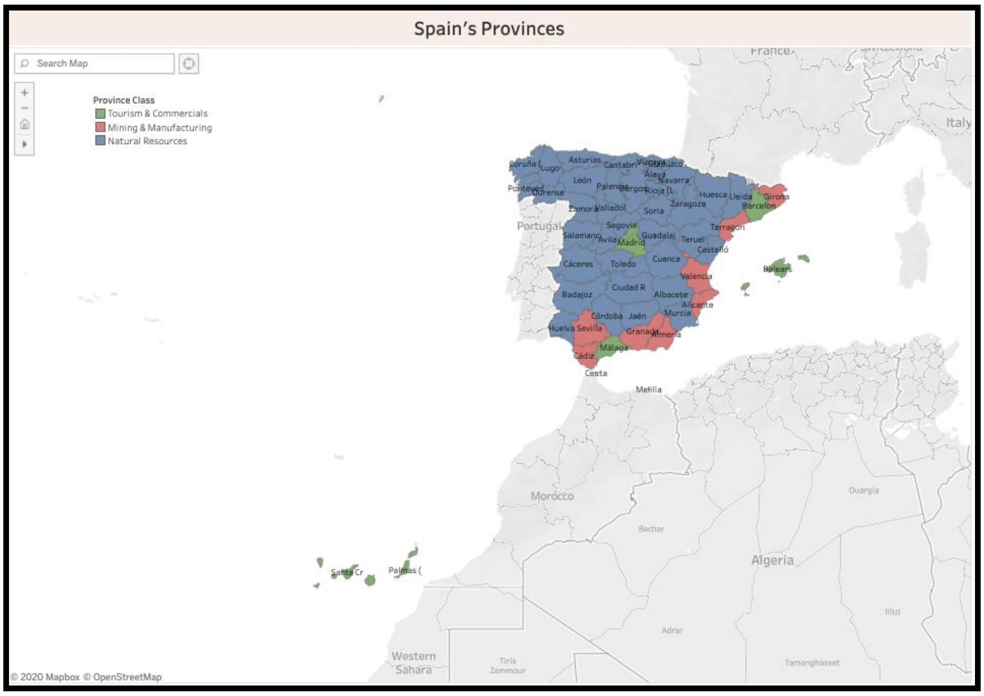
Given the general knowledge known about Spain and the results of the K-Means model that supports the belief that there are indeed differences among all provinces in Spain, Team A recommends that The Ministry of Economy in Spain group all provinces in the following cluster profiles:

Cluster's Name	**Brief Description**
Tourism & Commercials	This cluster consists of 6 provinces that shows a distinguishing degree of factors such as **tourism**, **commerce**, **wholesales** and **retails**.
Mining & Manufacturing	This cluster consists of 8 provinces that shows a distinguishing degree of factors such as **mining**, **manufacturing** and **metals**.
Natural Resources	This cluster consists of 38 provinces that shows a distinguishing degree of factors such as **agriculture**, **farming** and **energy**.

Here is the relationship between two variables amongst the clusters in a scatter plot and a geographical representation of the cluster profiles:

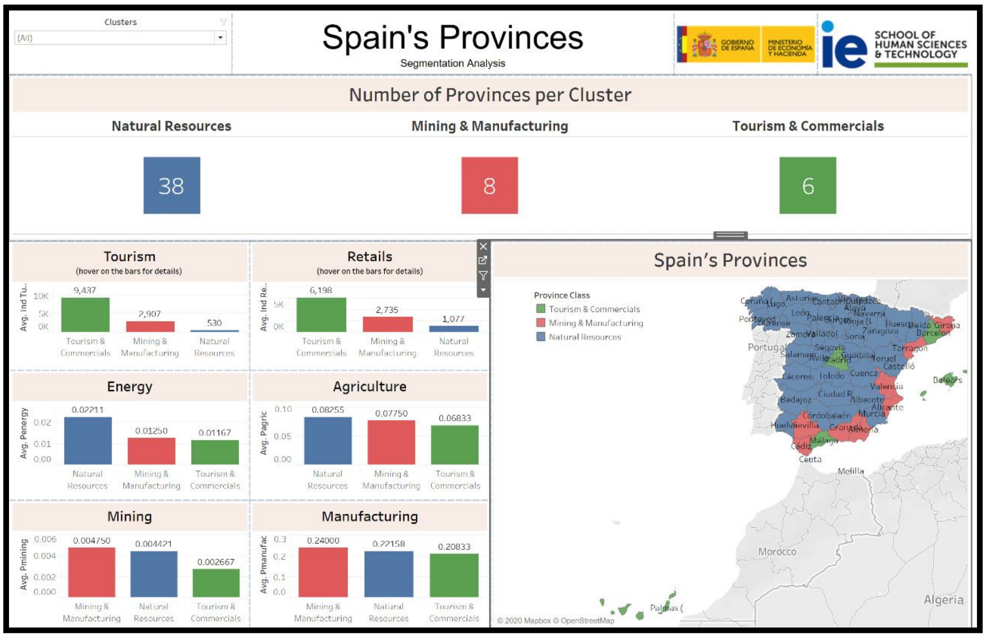


Scatter Plot: Pforeign & Ind Turis



Geo Map: Spain's Provinces

While The Ministry of Economy in Spain deliberates over novel policies to increase the prosperity of the country, Team A would suggest that The Ministry of Economy in Spain tailor the proposed policies not just to the country at large but to the aforementioned cluster profiles. Along with the resources already provided, Team A built a Tableau Dashboard for The Ministry of Economy in Spain: Segmentation Analysis - Spain provinces. This dashboard empowers The Ministry of Economy in Spain with the means to meticulously study the provinces in Spain prior to executing policies.



The Ministry's Dashboard to Explore the Spain's Provinces After Segmentation Analysis