Problem Statement - Clustering Countries for providing Tourism Services

Context

Tourism is now recognized as a directly measurable activity, enabling more accurate analysis and more effective policies can be made for tourism. Whereas previously the sector relied mostly on approximations from related areas of measurement (e.g. Balance of Payments statistics), tourism nowadays is a productive activity that can be analyzed using factors like economic indicators, social indicators, environmental & infrastructure indicators, etc. As a Data Scientist in a leading tours and travels company, you have been assigned the task of analyzing several of these factors and group countries based on them to help understand the key locations where the company can invest in tourism services.

Objective

To explore the data and identify different groups of countries based on important factors to find key locations where investments can be made to promote tourism services.

Key Questions

. Which variables should be used for clustering?
. How many different groups/clusters of countries can be found from the data?
. How do the different clusters vary?
. How to use PCA to retain the components which explain 90% variance?
. How to perform clustering using the components obtained from PCA?

Data Description

This dataset contains key statistical indicators of the countries. It covers sections like general information, economic indicators, social indicators, environmental & infrastructural indicators.

Data Dictionary

country: country

Region: region of the country

Surface area (km2): Surface area in sq. km

Population in thousands (2017): Population of the country, in thousands, as in the year 2017

Population density (per km2, 2017): Population density per km2, as in the year 2017

Sex ratio (m per 100 f, 2017): Number of males per 100 female, as in the year 2017
GDP: Gross domestic product (million current US\\$): GDP of the country in million USD
Economy: Agriculture (% of GVA): Contribution of agriculture to the economy as a percentage of Gross Value Added
Economy: Industry (% of GVA): Contribution of the industry to the economy as a percentage of Gross Value Added
Economy: Services and other activity (% of GVA): Contribution of services and other activities to the economy as a percentage of Gross Value Added
International trade: Exports (million US dollar): Amount, in million USD, of international exports
International trade: Imports (million US dollar): Amount, in million USD, of international imports
International trade: Balance (million US dollar): Amount, in million USD, of balance between international exports and imports
Fertility rate, total (live births per woman): Fertility rate of the country computed as the no. of live births per woman
Infant mortality rate (per 1000 live births): Infant mortality rate of the country computed as the no. of dead infants per 1000 live births
Health: Total expenditure (% of GDP): Total expenditure on healthcare facilities as a percentage of GDP
Education: Government expenditure (% of GDP): Total expenditure on education as a percentage of GDP
Mobile-cellular subscriptions (per 100 inhabitants): no. of mobile/cellular subscriptions per 100 people
Individuals using the Internet (per 100 inhabitants): no. of individuals using the Internet per 100 people
Threatened species (number): Number of threatened species
CO2 emission estimates (million tons/tons per capita): CO2 emission estimates in million tons
Energy production, primary (Petajoules): energy production in petajoules

Learning Outcomes

Exploratory Data Analysis
Clustering
Cluster Profiling
Dimensionality Reduction
Business Recommendations

Steps and Tasks

Import libraries and load dataset
EDA

Perform clustering and cluster profiling

Perform dimensionality reduction

Perform clustering and cluster profiling on lower-dimensional data

Provide business recommendations and conclusion