E.M. trabalho pessoal

Emanuele Vivoli

November 16, 2021

1 Dataset Identification

In this work I'll put my attention to an interesting dataset from Kaggle called "World Happiness Report" (https://www.kaggle.com/unsdsn/world-happiness). The dataset is composed by "Happiness scored according to economic production, social support, etc." and I'll use some of the methods that we studied in class. First I'll see if there are attributes that are not statistical significants using ANOVA test, then I'll use the PCA to apply dimensionality reduction and importance attributes studies. Then I'll apply some clustering technique such K-Means.

2 Dataset Description

The dataset is composed by just 12 columns, and it makes me think that maybe a PCA can be useful but not too much (for dimentionality reduction a twelve dimension dataset is not such big that need PCA reduction). The first two columns are composed by the name of the Country and the Region the Country belongs to. The other 10 columns are, instead, all numerical and are, the first two, the Happiness Score and the Happiness Rank, then we have 8 columns where are represented the extent to which every attribute contributes to the calculation of the Happiness Score.

3 Goal

The goal of this project is to use most of the instrument we saw in the lectures, but first of all I'd like to see PCA and K-Means for a real context in which I didn't see them before. It is true what I wrote before, so one of the goal is also to confirm or not my idea that maybe a PCA here is not essential.

4 Methodology

By an analysis of the data, use ANOVA to make sure that the interaction effect between two attributes and the main effect of just one of those is not statistically significant. This can be use for cut away one attribute if the effect of it is not significant. Then we can proceed using PCA in order to make a dimensionality reduction and see if it's worth it.