E.M. trabalho pessoal

Emanuele Vivoli

December 20, 2019

1 Introduction

In this work is used a dataset from Kaggle called *World Happiness Report*. The data are collected by an association called *Gallup World Poll* (GWP) that has the goal to became an indispensable tool for global thought leaders and decision-makers. They collect data in order to help powerful people to make decision by following global agglomerated data about poverty, happiness, education and other type of interesting topics. Every year a *World Happiness Report* is published by the United Nations Sustainable Development Solutions Network using the GWP data, and some sort of computation on those data.

This work is structured as follow. In Section 2 we will describe our data, and what they represent. In Section 3 we will focus our attention on understanding how the data were taken and what every variable means. We'll navigate the data using some Visualization tools and we will examine the data itself by looking at correlation information and statistical independence between variables and happiness score. We will apply, in Section 4, features selection techniques such as *Filtering* (Uni-variate filtering) as well as *Principal Component Analysis* (PCA) for dimensionality reduction. Finally, in Section 5, we will apply clustering method such *K-Means*, and we'll evaluate the clusters by using a Score that will be introduce in that Section. The conclusion will be done in the Section 6 were we will examine results from the prototype clustering method K-Means in a visual way and with the score from the previous evaluation.

2 Dataset Description

Each dataset we use (every files from 2015 to 2019 in the WHR [Net19]) is composed by few columns in which are represented a composition of the 27 columns from the GWP dataset [Lau17]. Because we are using dataset of 5 different years, some column changed in this period, so the 11 columns that are common in all the 2015-2019 dataset are the following:

- **Country** , it is the name of the Country.
- **Region** , it is the name of the Region the Country belongs to.
- **Happiness Rank** , it is the rank of the Country based on the Happiness Score.
- **Happiness Score** , it is the national average response to the question: "Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?".

- **Economy GDP per Capita** , (variable name gdp) in purchasing power parity (PPP) at constant 2011 international dollar prices are from the August 10, 2016 release of the World Development Indicators (WDI). The GDP data are missing in some Country and so substitute with some early release and adjusted with a moltiplicator [JHS17].
- **Family** or Social support (or having someone to count on in times of trouble) is the national average of the binary responses (either 0 or 1) to the GWP question "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?".
- **Health Life Expectancy** , (HLE) is the time series of healthy life expectancy at birth are calculated by the authors based on data from the World Health Organization (WHO), the World Development Indicators (WDI), and statistics published in journal articles.
- **Freedom** to make life choices is the national average of responses to the GWP question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?".
- **Generosity** , is the residual of regressing national average of response to the GWP question "Have you donated money to a charity in the past month?" on GDP per capita.
- **Trust Government Corruption** is the national average of the survey responses to two questions in the GWP: "Is corruption widespread throughout the government or not" and "Is corruption widespread within businesses or not?". The overall perception is just the average of the two 0-or-1 responses. In case the perception of government corruption is missing, we use the perception of business corruption as the overall perception. The corruption perception at the national level is just the average response of the overall perception at the individual level.
- **Dystopia Residual** is the Dystopia Happiness Score(1.85) + the Residual value or the unexplained value for each country. The residuals differ for each country, reflecting the extent to which the past six variables either over- or under-explain average 2014-2016 life evaluations. These residuals have an average value of approximately zero over the whole set of countries.

3 Dataset Analysis

The first step is to visualize the data. Figure 1 shows the Vionin Plot of our data, referring to the year 2016, of the Happiness Score grouped by Region. We can see how Middle East and North Africa (MNA) have a very spread graph while Australia and New Zealand (ANZ) have really low variance.

The question that comes with this figure is if only the Happiness Score have this difference in the variance between Regions or also other features shows a

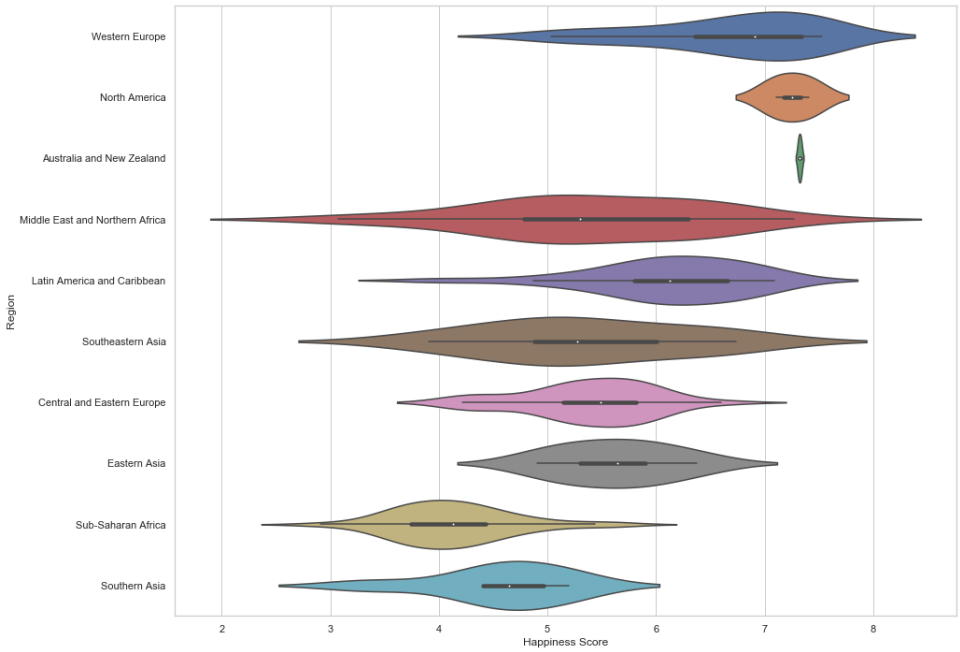

Figure 1: Violin Plot: abscissas "Happiness Score" and ordinates "Regions"

high variance. To answer to that question we can use Box-plot visualization of some other attribute. The Figure 2 shown respectively the Box-plot of Economy, Trust (in government) and Health attributes over Regions. We can see also from this figures that ANZ and NA have always low variance while MNA and SEA have high variance almost in all the three pictures and WE change a lot between the first two figures and the lastone (in term of variance).

So there is either something different in the data between those group or some unbalanced proportion of Countries in Regions. This second thing can be seen in the Figure 3 that shows the percentage of Country (count of countries over the total number of countries) that belongs to one Region. Then the variances of ANZ and NA are little over the features because both of them have only 1.27% of the total number of Countries, and it makes them have low variances.

Now we understood quite good the composition of our data, is time to study if there are some relationship between our features. This can be done, as first approach, by looking at the correlation matrix in Figure 4 that shows the correlation matrix between features. We can clearly see the high correlation between Happiness Score and Economy (GDP), Family and Health (or Life) attributes, but also a high correlation between those attribute them-self. We can also see that the Trust is low in every correlation, but almost 0.5 in Freedom and lower in Happiness Score.

Another visualization method used to understand the feature we are working

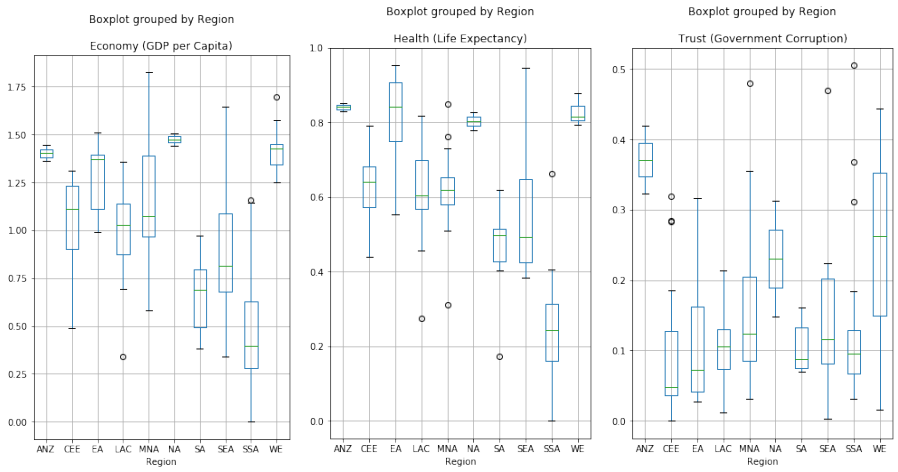

Figure 2: (left) economy; (central) health; (right) trust.

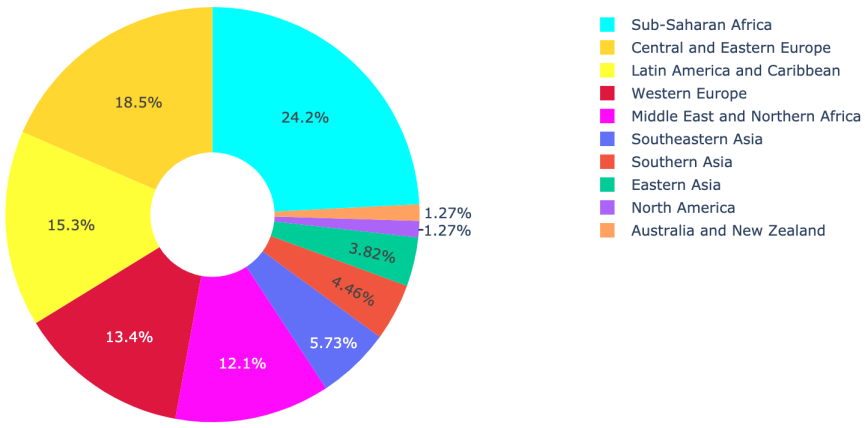

Figure 3: Pie graph about count of Countries belonging to a Region over the total number of Countries.

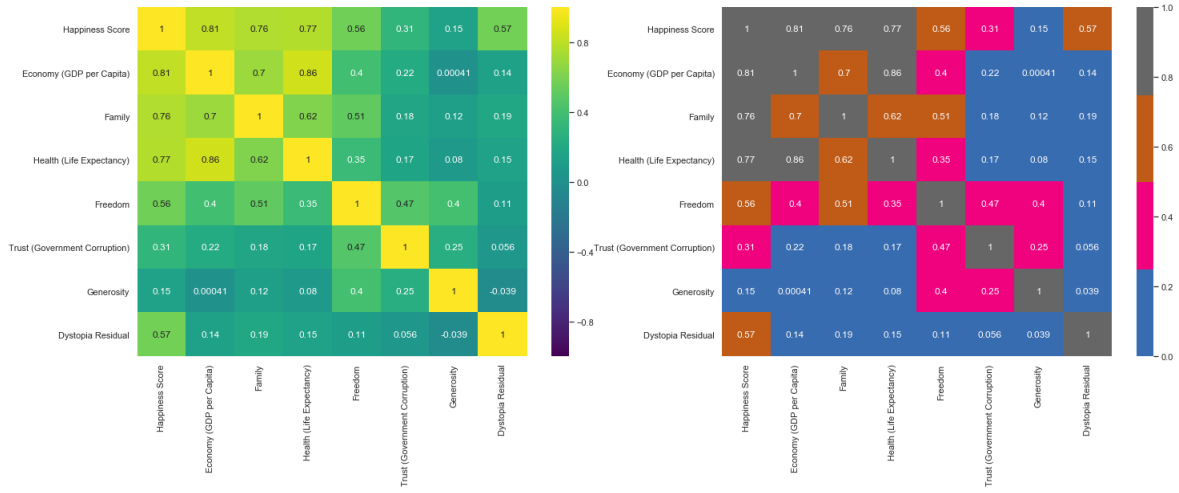

Figure 4: Correlation matrixes, (left) with sign, (right) after quantization.

with is the scatter matrix, shown in Figure 5. The picture colors represent level of Happiness *Low, Low-Mid, Top-Mid, Top* that are 4 sets equally big obtained by partitioning the Happiness Rank feature in 4. From this picture we can see that GDP, Family and Life are linearly correlated with Happiness Score and also between them.

The last method used is called Parallel Plot, or parallel coordinates, and is shown in Figure 6. This shows two things: the Top class stays over the others for GDP, Family and Life and the Low class stays under the others for the first 4 features. Unfortunately we cannot understand much from this plot. In Generosity and Trust nothing is distinguishable.

4 Features Selection

Once we have a good understanding of the relationships between data, we can think about some features that is redundant in order to reduce the dimensions of our dataset (even if working with 6 features is already an easy problem for cluster algorithms). From the previous study, one features between GDP and Life can be removed because the correlation is high and just one of those is sufficient to express, with approximation, the other. So following the one that have higher correlation with Happiness Score we remove Life from the dataset.

Another thing we can notice is that in the scatter matrix (Figure 5) Generosity has a non useful distribution of data over the 4 classes. We can try an ANOVA one-way test in order to see if the means on different classes are statistically equal depending on the features classes. So for the ANOVA test the H0 is accepted (means equals for different classes, or, indipendence between feature and classes), so we need to reject the variable.

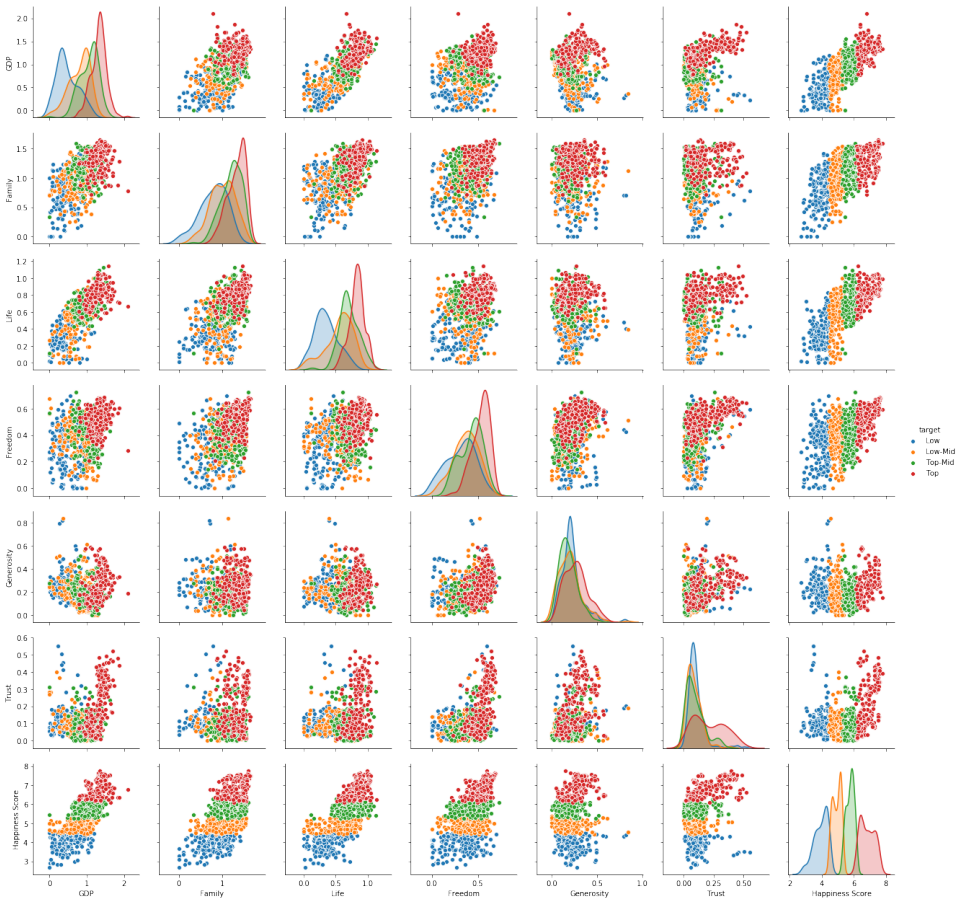

Figure 5: Scatter Matrix

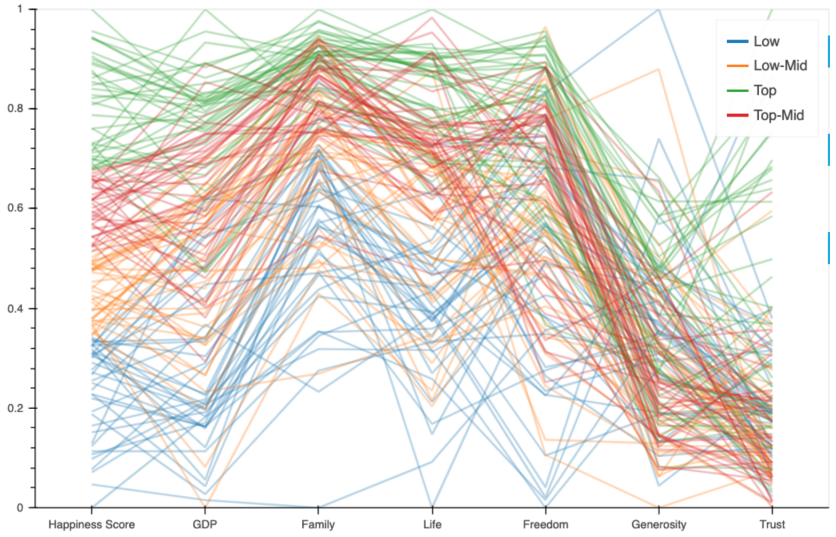

Figure 6: Parallel plot

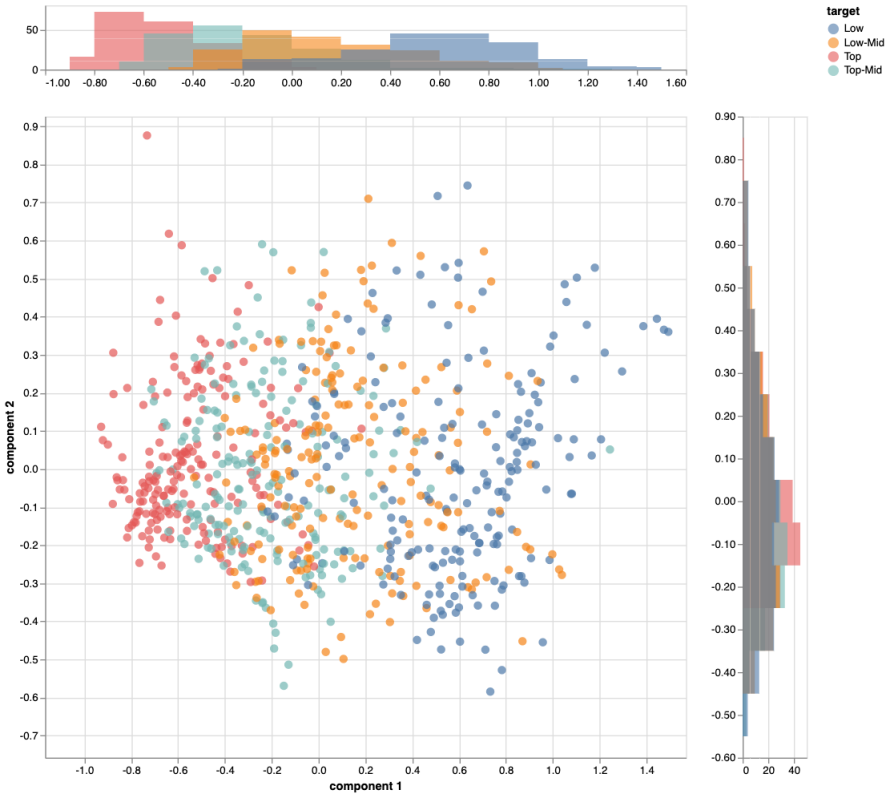

Figure 7: PCA: PC1 on the abscissas and PC2 on the ordinates.

Now that we have the importants features, if we still want to reduce our dataset (for dimentionality reduction or just for 2D visualization) we can apply the PCA and retain only the first two principal components. Doing so we obtain Figure 7 that has on the abscissas the PC1 and PC2 on the ordinates.

We can see from this figure that the different Happiness Scores are not very isolated. We can try, before applying the clustering method in the next Section, to use also the third principal component, so we use PCA dimentionality reduction on the features, retaining only the firsts 3 Principal components.

5 Clustering

In this section we apply the K-Means algorithm on the 3 principal components with K=4 in order to recreate the cluster for *Top, Top-Mid, Low-Mid, Low* classes.

To visualize the results (because the 3D plot visualization is not well understandable) we create two plots of the Globe, one with the real Target value for

each Country and another with the clustering assignment given back from the K-Means algorithm. The Figure 8 shown respectively the Real Target class and the K-Means clustering.

6 Conclusion

In conclusion, to evaluate the result from K-Means, we applied the Adjusted Random Index as comparison method for clustering algorithm. The method is a variant of the Random index (RI) as follow:

$$ARI = \frac{(RI - Expected\_RI)}{(max(RI) - Expected\_RI)}$$

where the RI is defined as:

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}} = \frac{TP+TN}{TP+FP+FN+TN}$$

with TP (true positive), TN (true negative), FP (false positive), FN (false negative), and the following definition:

"the number of pairs of elements in S that are in"

- a, the same subset in X and in the same subset in Y
- b, different subsets in X and in different subsets in Y
- c, the same subset in X and in different subsets in Y
- d, different subsets in X and in the same subset in Y

Now we have defined the score, we can proceed to calculate this ARI on our K-Means result obtaining a score of *0.2893405747114898*. It depend on the data set, but in general, 0.20 is not very good because ARI adapts the rand index (RI) to have score 0 when is putting everything in random buckets. So 0.20 and lower are results quite close to randomness.

In our case, then, we can deduct that a PCA can not be useful, due to the fact that number of features that we had at the beginning wasn't that high. So as alternative can be better to not do the first features selection step, when we removed the Life feature, or also retain much more features for the PCA. Another problem can also be the Clustering algorithm and an alternative can be to use Hierarchical clustering or some clustering that is density-based such as DBSCAN or Gaussian Mixture models.

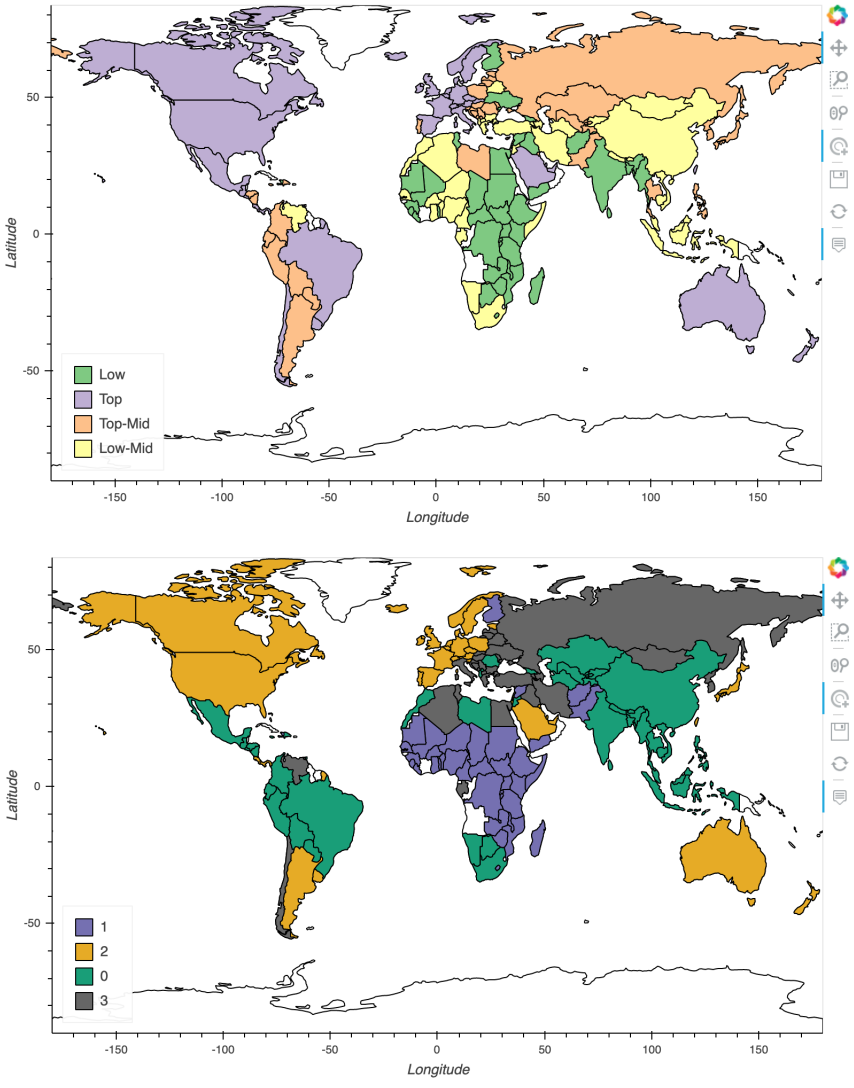

Figure 8: (top) Real Target class; (bottom) K-Means Target class.

References

[JHS17] F. Helliwell John, Huang Haifang, and Wang Shun. "Statistical Appendix for 'The social foundations of world happiness'". In: *Data.World* (2017), p. 2. DOI: https://data.world/laurel/world-happiness-report-data.

[Lau17] Hanscom Laurel. *World Happiness Report Data*. Data.World, 2017. URL: https://data.world/laurel/world-happiness-report-data.

[Net19] Sustainable Development Solutions Network. *World Happiness Report*. Kaggle, 2019. URL: https://www.kaggle.com/unsdsn/world-happiness/data.