Data Exploration for Human Freedom Index

Yuzhe Yang*, Siyuan Liu, Yuqian Dai, Laiyuan Zhang, and Lei Guo

***Abstract*—The Human Freedom Index (HFI) presents the state of human freedom in this world based on a broad measurement, which encompasses personal, civil, and economic freedom. In this project, we choose the Human Freedom Index-2019 as our raw data and we select the data in the year 2017 for further analysis. We first discover the basic information of dataset, such as the number of variables, observations, missing cells and etc. Second, we analyze the correlation among those variables and visualize the data on different levels. Third, we use both supervise and unsupervised learning approaches to analyze the data. Finally, we give our conclusions about the data. The related data can be downloaded from: https://www.cato.org/human-freedom-index-new. Our code for data exploration is available on: https://github.com/YuzheYang/AdvancedML2019-2020.**

***Index Terms*—Human Freedom Index, data visualization, data cleaning, correlation analysis**

I. Introduction

THE Human Freedom Index (HFI) is one of the most comprehensive freedom index created for the global countries. In this dataset, 162 countries are taken into account and each country contains 120-dimensional features, which include the HFI (hf score) and its rank (hf rank). The index ranks countries beginning in 2008, which is the most recent year when sufficient data are available, and it lasts until 2017.

A. Basic information

Basically, this fifth annual index uses 76 distinct indicators of personal and economic freedom in the following areas, such as rule of law, security and safety, religion, freedom to trade internationally, and etc. The range of each measurement are on a scale of 0 to 10, where 10 represents more freedom, the average human freedom rating for 162 countries in the 2017 was 6.89. Since year 2017 is the most up to date, we choose to make regression and clustering analysis based on this year's data. The main purpose of this work is to conduct numerical analysis and discussion of different aspects through the dataset provided and discover potential features for our future project.

In the first section, we will give a general introduction to HFI and its related features. Then in the second section, we will talk about related discoveries in data cleaning. Thrid, feature selection and regression analysis will be given based on the data in the year 2017. Forth, a cluster analysis will be given for analyzing the attributions of different countries. In the last section, a conclusion will be given to summarize the characteristics of our dataset and we will point out the potential problems, which we need to pay more attention to in our formal project.

Yang et al. are with the School of Electronics and Computer Science, University of Southampton, Southampton, e-mail: (yy1a19, sl18n19, yd6u19, lz4y19 and lg1y19) @soton.ac.uk, and this report is the data exploration for our group project in COMP6208: Advanced Machine Learning (2019-2020).

B. General discovery

Initially, we focus on discussing the distribution and setting of data labels and data analysis and conjecture between different countries and regions. For the division of the label, intuitively, the more complete data information is selected as the label class, to minimize the experimental error caused by the lack of data. The first chosen label is "region", because "region" not only provides a complete regional data division, but also a more intuitive and easier to understand data in the entire dataset. After manual selection, there are 10 types of model labels, which are Eastern Europe, Middle East and North Africa, Sub-Saharan Africa, Latin America and the Caribbean, Caucasus and Central Asia, Oceania, Western Europe, South Asia, North America, East Asia. However, due to the diversity of data, it can be found that the distribution of tags is not uniform. The specific distribution proportion graph is shown in Fig. 1.

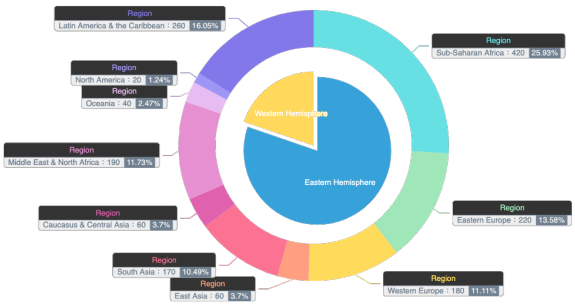

Fig. 1. Distribution proportion graph respects to region.

After visualizing the proportion, a preliminary conclusion can be drawn that Sub-Saharan Africa accounts for the largest proportion and North America accounts for the smallest proportion. Eastern Europe, Western Europe, Middle East and North Africa, and South Asia share a similar proportion. In other words, the country distribution is not quite balanced, however, it covers a large number of regions in this world.

To illustrate the function of the HFI score, we also view the characteristics of a certain sample. We selected Australia and France as the country and described the change of hf rank and hf score in 2008-2017. The results show that its ranking fluctuates within 4-8, and the score as a whole fluctuates upward. Based on this, we can speculate that some vicious events may have occurred in Australia in 2012, resulting in a particularly low hf score. As a comparison with France, it is impressive that France's overall indicators are worse than Australia's, France's scores continue to rise and the overall ranking also shows an upward trend.

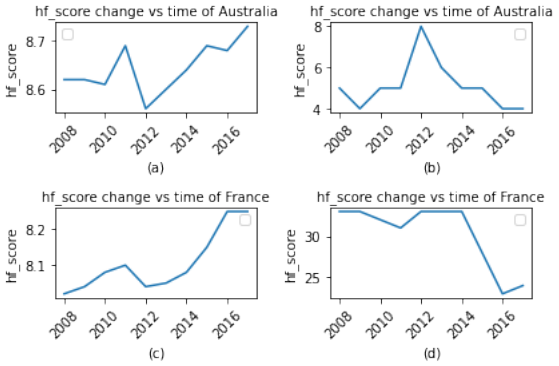

Fig. 2. Change of hf score and rank in Australia and France.

To summarize, in this section, we talk about some general discovery on our dataset and we point out the function of hf score and rank, which may mirror a country's overall status. In the next section, we will give an introduction to data cleaning and manipulate the missing data.

II. Data cleaning

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a recordset, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. In practice, a very common problem is that datasets often have missing values. A large amount of data is missing in the dataset. 96.66% of the columns have missing values where some features are with a high missing rate. To a certain extent, if the missing values are not dealt with, the future work will be difficult to continue. Generally speaking, there are three ways to eliminate missing data. First, we can directly abandon this data to avoid side affection. Second, we can fill the missing cell with zero or NaN. Third, we can use interpolation based on mean or mode value. In addition, we can use some machine learning models to fit the missing data with prior knowledge from similar data.

In this work, we first observe the missing data in our dataset and it turns out that the missing data cell is replaced with a special character "-". Specifically, according to the statistical result, 14 features contain a missing data over 20%. We delete those features from our dataset, since these features might cause great bias to the analysis result and they are unfillable. The visualized missing rate is shown in Fig. 3. It can be observed that most features contain miss data phenomenon. However, most of the missing data rates are still low, which indicates that the dataset is usable and with good quality.

In this work, we use zero-filling to those missing data, whose missing rate is under 20% for further processing. It is not only because the potential reasons behind the missing data, but also, after testing the corrected data's effectiveness and usability, it has been proved that zero-filling would do no harm to our analysis. Therefore, we choose to use zero fill operation to the missing value.

In Fig. 3, the y axis stands for the missing rate, and x axis stands for each feature. It can be observed intuitively that some

Fig. 3. Missing data rate in dataset.

features are totally missed and most features are in low missing rate which indicates that our dataset is usable and reliable.

To summarize, in this section, we first give an introduction to the data cleaning. Then, we choose the zero-filling operation to fill in the empty cells and we give explanations to this operation. In the next section, a regression analysis will be given for studying on the mapping correlation between hf score and its related features.

III. Regression analysis

In this section, we will give a regression analysis that can map fewer features into the HFI score. This work can help the survey designer to reduce their designing features and remove some unnecessary items to save both time and money. This section consists of two subsections. First, feature selection and correlation analysis will be illustrated for reducing the dimensionality. Then, we will make a comparison among different algorithms for the regression task, which include random forest, support vector regression (SVR), multi-layer perceptron (MLP) regression, AdaBoost and XGBoost.

As mentioned before, the dataset contains 120 dimensional features, which are all the accordance for calculating the HFI score. However, some features are highly correlated with the others, which may cause the regression model sensitive to parameters and also it will take up more weight parameters. Moreover, reducing the input features can make the relation between labels and features be highlighted. Hence, selecting features is vital to machine learning-based approaches.

In this work, we first compare hf score to all those qualified features. Then, we use a settled threshold to divided the features into high correlated and low correlated to hf score. After this classification, we use a scatter figure to observe the correlation between hf score and each feature. The result shows that there actually exist some features which are highly correlated with hf score and their cross-correlation are also high. Those features are labeled as redundant data, which should be resampled for reducing the dimensionality. Then, we use sklearn built-in feature selection approaches, such as SelectKBest, and Tree-based selections to select features. In addition, we also select one group hand-crafted features, which is set as a comparison group.

A. Correlation analysis

To select features, we first need to dig out which features are highly correlated with hf score, our regression target,

which in other words, these features are redundant and we need to eliminate those features. Hence, we first calculate the correlation index between hf score and features and we set a threshold which equals 0.7. If the Pearson coefficient is higher than that threshold, then we can tell that feature is highly correlated with hf score. The related calculation results is that there are 14 qualified features (missing data rate under 20%) are highly correlated with hf score and 84 features' coefficients are lower than 0.7.

B. Feature selection

We have three kind of feature selection, which are SelectKBest, Tree-based selection and our hand-crafted selection.

1) SelectKBest: SelectKBest selection is to remove all features but the k highest score features. The scoring approaches are various, where we choose the "f regression" to score the dataset and select 30 highest features for the regression task. The selected features and its correlation with hf score are visualized in Fig. 4.

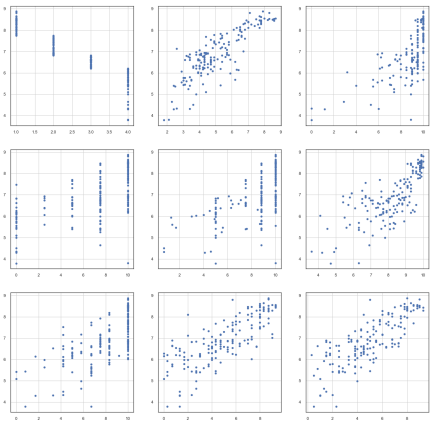

Fig. 4. Scatter of selected features with hf score via SelectKBest.

2) Tree-based feature selection: In addition, we also choose another approach named "Tree-based feature selection". Here, we use ExtraTreesRegressor as our regressor model with 30 estimators for selecting features. The selected features and their correlation with hf score are visualized in Fig. 5.

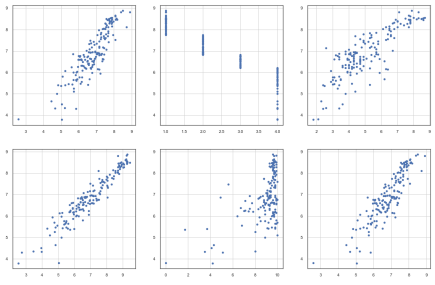

Fig. 5. Scatter of selected features with hf score via Tree-based selection.

From Fig. 4 and 5, it can be observed that these features are not in a common distribution or pattern with hf score, which promise the effectiveness of our selected features. Compared with highly correlated features, those features are more representative and can cover potential factors which may influence the final hf score. Moreover, our hand-crafted selection is totally based on the feature structure and their correlation on semantic level. Here, we have obtained our training data and in the next section, we will make regression analysis to build up the mapping correlation between our features and the HFI score.

C. Regression analysis

For analysing the regression results, we choose to divide the dataset into two parts, 130 sample as our training set and 32 sample as the testing set. In this work, we choose five regression algorithms, which are random forest, support vector regression (SVR), multi-layer perceptron (MLP) regression, AdaBoost and XGBoost for comparison. The result is shown in Table. III-C

Model	SelectKBest	Tree-based	Ours
Random Forest	0.87	**0.93**	0.78
SVR	**0.98**	0.97	0.94
MLPRegressor	0.97	**0.99**	0.86
AdaBoost	0.86	**0.94**	0.76
XGBoost	0.95	**0.96**	0.88

It can be observed that the Tree-based selection can outperform the others and select effective features for regression task and the average accuracy on test data is 0.96 which is quite high and it proves that our feature selection is successful and one good thing is the tree based selection reduces the features into 5 dimensions which is quite low.

IV. Cluster analysis

To better understand and cluster the data, we first analysis the feature structure and clean the data based on the structure graph. Then, we use k-means to cluster the data for analyzing the country HFI score and we direct the clustering result on the world map for observing the HFI cluster distribution.

A. Feature selection

The value of each feature as the upper-level father node is the mean of all its children nodes. Besides 38 columns of indices and overall information, there are still ten blank columns, one questionable column with obscure meaning, and 14 columns of data which are highly correlated with others. Through removing 63 columns of data mentioned above, there are still seven columns of non-basic features that are necessary because the basic features of them are partially or totally missing. Hence, 67 essential features are selected for further processing, which is also the data source in regression analysis "Ours". The features in the dataset are related and the feature structure is shown in Fig. 6.

B. Clustering analysis

As always, K-selection is a noteworthy question for K-required clustering algorithms. In cluster algorithms, the elbow method could be a reasonable heuristic to determine. Hence,

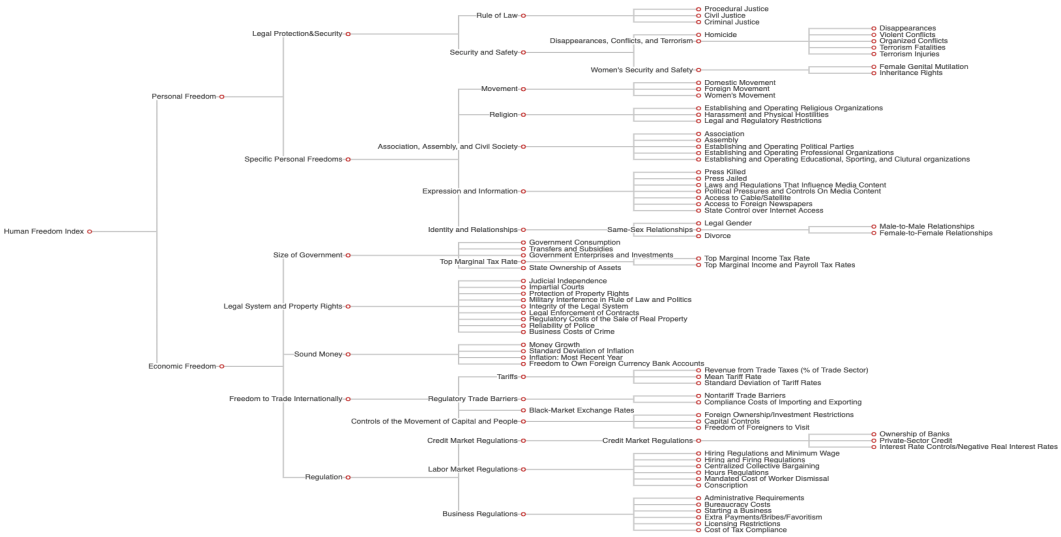

Fig. 6. Feature tree.

we select both the silhouette coefficient and the average distance of each point to cluster centres to measure the performance while consistently changing the K value. A related result is shown in Fig. 7. Through abandon overly small and overly significant K values, we can tell that there is an elbow point where K is 8. Then we use eight as the number of clusters of K-means.

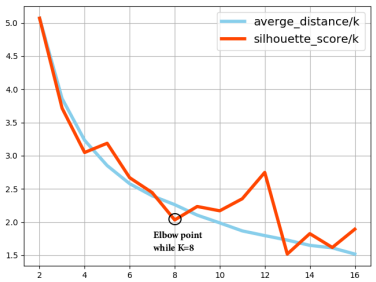

Fig. 7. K value selection via elbow method.

To decomposite 67 features into two-dimension coordinate-like data for K-means, we select three different decomposition algorithms, which are Non-negative matrix factorisation (NMF), t-distributed stochastic neighbour embedding (t-SNE) and the principal component analysis (PCA) as a comparison experiment. The result indicates that t-SNE can outperform the others. Therefore, t-SNE is chosen for decomposition.

We project that clustering result in Fig. 8. Here, each colour represents a cluster class, and some countries may have different colours for different regions, such as Taiwan and Hongkong of China. The white area with black boundary means no statistical data on these areas, and it also means that country has very few residents or rather small. And it's worth pointing out that some places and Antarctica are removed from this world map due to limited resolution. Interesting thing is, the adjacent countries are very apt to be in the same cluster while there is no geographic data that has been used in the selected 67 essential features. It might indicate that adjacent countries can share similar sense of freedom to some extent. Moreover, beyond the aforementioned phenomenons, some countries are surrounded by variant cluster which is contrary to the common laws.

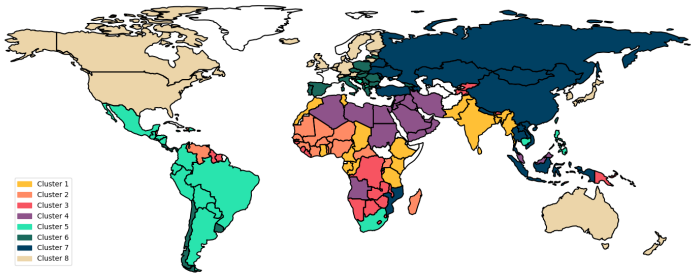

Fig. 8. Clustering result reflected on the world map.

V. Conclusion

In conclusion, in this data exploration, we first dig out some general features for intuitive observation. Then, we do data cleaning and feature selection through correlation analysis and we proposed our approach for feature selection. Moreover, regression analysis and clustering analysis are applied for understanding the country's freedom status. The regression analysis indicates that we can use Tree-based feature selection for predicting hf score and the number of selected features is 5 which is quite low and effective. The cluster analysis shows that adjacent countries are usually sharing a similar sense of freedom.

In the future, there are a few points we should discover. First, we need to figure out whether the selected features could be used to do regression prediction on the other index such as economic ,etc. Second, potential reasons should be discovered that the adjacent countries might usually share a similar sense of freedom. Third, beyond zero filling operation, what else can we do to improve the data quality. These problems are meaningful and remaining to be solved, which will be discussed in our project report.