**Construction of an index to evaluate the democratic level of countries**

**I.** **Abstract**

**II.Introduction**

**III.Methods and Materials   
 A.Studies of several datasets**

1. The Democracy index from The Economist
2. Dataset from last year (dataset 2)
3. Construction of a new dataset using new variables (dataset 3)
4. A dataset containing the most and the least democratic countries

**B. Statistical tools used**

      1.   A first approach to understand the dataset with Principal Component Analysis (PCA)  
                   2.   Group the individuals according to their characteristics with clustering  
       3.   Linear and logistic models  
       4.   Regression and classification trees  
       5.   Understand the importance of variables with Random Forest

**C. Missing data management**

**IV.** **Selection of the variables which are the most relevant to explain the democratic level**

**A. Studies of the link between the different variables with descriptive analysis and PCA.**

**B.** **Clustering to group countries according to their democratic level**

**C.** **Logistic and linear model to explain the democratic level of the Economist**

**D.Regression/Classification tree and Random Forest to find the most important variables to explain Gscore**

**E. Comparison between the most and the least democratic countries**

**V.** **Conclusion  
VI.Acknowledgement  
VII.References**

**features or variables ? => pour lier au fait que ce sont des “caractéristiques” qu’on regarde dans chaque pays. Parce qu’une variable ça peut être tout ou rien ?**

**Construction of an index to evaluate the democratic level of countries**

**I.** **Abstract**

**II.Introduction**

**III.Methods and Materials   
 A.Studies of several datasets**

1. Democracy index from The Economist
2. Dataset building with variables from *Wikipedia*
3. Dataset building with variables from the website *Our World in Data*
4. Retrieval of the most and the least democratic countries // récupération, extraction, isolation (expliquer que c’est étudier des comportements extrêmes)

**B. Statistical tools used**

      1.   Descriptive statistics and Principal Component Analysis (PCA)  
                   2.  Clustering   
       3.   Linear and logistic models  
       4.   Regression and classification trees  
       5.   RandomForest

**C. Management of missing data**

**IV.** **Selection of the variables which are the most relevant to explain the democratic level**

Je pense que les parties en orange peuvent être facultatives et plutôt décrites directement dans les paragraphes. Et puis toutes les méthodes sont déjà citées au-dessus

**A. Link between the different features with descriptive analysis**

**B. Reduction of the problem’s dimension thanks to PCA**

**B. Grouping of countries with similar behavior using clustering**

**C.** **Modelisation/Explanation of the democratic level (Gscore) according to multiple features with linear and logistic regression/model**

1.

2. Focus on the most and least democratic countries

Est-ce que c’est nécessaire de faire des sous parties ou juste une phrase de transition qui dit qu’on va faire un focus suffirait ?

**D. Highlighting the features which significantly explain the Gscore using Regression/Classification trees and RandomForest**

1.

2. Focus on the most and least democratic countries

Même question précédemment

**E. Comparison between the most and the least democratic countries**

**V.** **Conclusion  
VI.Acknowledgement  
VII.References**