Generated at Sun Mar 3 05:19:05 2024.

Source at https://www.kaggle.com/code/ravishahi/exploring-deforestation-via-unsupervised-ml-0ef0c0/exploring-deforestation-via-unsupervised-ml.

Notebook by Ravi Shah. Dataset by Global Forest Watch. Inspiration by Jacqueline Ellis. **Insert the formal** for the remaining, select formal + Enter. Change cell type to Markdown using Cell menu or Ctrl+M, M.

Make sure you fill in your data for `NAME`, `YEAR_START`, `YEAR_END`, `YEAR_SHORTENED` and `AREA` as well as your name and collaborators below.

```
NAME = "Kenya Project"
YEAR_START = "2010"
```

Project presentation

This notebook presents our project and goal for our Data Science project for the presentation days. You can sign up for a presentation slot at

https://docs.google.com/spreadsheets/d/1ZqqJ-Ln_Ln_Ln_Ln_Ln_Ln_Ln_Ln/edit#gid=0

Presentations should be ~10 minutes long, and have ~10 min. 5 minutes for questions. You should introduce the topic, show your data and your approach. You can also include a conclusion if applicable.

Note that presentations are not long, so keep it short. Sharing your screen and that you are a data scientist, you can do the presentation during your data day.

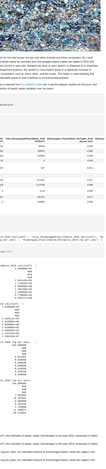
You will turn in a PDF version of your presentation.

Title: Exploratory Data Analysis of Plastic Waste Using K-Means Clustering

In recent a popular towards digitization and in machine learning and data analysis. It is an unsupervised learning algorithm that combines a dataset into k clusters, where each data point belongs to the cluster with the nearest mean.

Clustering of data is performed for various reasons:

- **Data Organization**: Clustering helps in organizing extensive datasets and structure within a dataset. By grouping similar data points together, patterns and relationships that may not be easily recognized and understood.
- **Data Compression**: Clustering can be used for data compression, where instead of representing each individual data point, a cluster can be represented by its centroid. This reduces the amount of data, allowing for more efficient storage and processing of large datasets.
- **Pattern Recognition and Summarization**: Clustering helps in uncovering the underlying structure of large datasets. It provides a way to identify patterns, trends, and relationships within the data that may not be apparent when looking at individual data points.
- **Anomaly Detection**: Clustering is useful in anomaly detection and outlier detection. It can identify data points that do not belong to any cluster, indicating potential anomalies or outliers in the dataset.
- **Dimensionality Reduction**: Clustering can be used as a pre-processing step for dimensionality reduction. Instead of dealing with a high-dimensional dataset, clustering can reduce the dimensionality of the data points based on their membership to specific clusters, capturing the dimensionality of the dataset.



Microplastic Plastic Waste is one serious problem for not only human but also any other animals and whole ecosystem. Our main goal is to explore the distribution of plastic waste in different countries using K-Means clustering. We will use the data from Global Forest Watch (GFW) dataset which provides open access to tracks deforestation and land use, changes and carbon emissions. We will use the data from GFW to explore the distribution of plastic waste in different countries using K-Means clustering. We will also use the data from GFW to explore the distribution of microplastic waste. This data is produced by microplastic waste. This data is produced by the scientists at the ecosystems such as rivers, lakes, and the ocean. This helps in understanding the distribution and impact of microplastic waste on different ecosystems and environments.

- **Data Collection and Data Exploration**: Data is obtained from Global Forest Watch. We will explore the dataset, examine its structure, and check for missing values. Visualize the distribution of plastic waste variables over the years.
- **Data Preprocessing**: Clean the data by handling missing values, standardizing or normalizing the data, and ensuring data consistency. Prepare the data for clustering by selecting relevant features and potentially applying dimensionality reduction techniques if needed.
- **Feature Engineering**: Consider creating new features or transformations of existing features that may provide additional insights or improve the clustering performance. Feature engineering can enhance the ability of the clustering algorithm to identify meaningful patterns.
- **Clustering Algorithm Selection**: Choose the K-Means clustering algorithm as the primary clustering algorithm. Consider other clustering algorithms for comparison or as alternatives if K-Means does not yield satisfactory results.
- **Determining the Optimal Number of Clusters**: Use the elbow method, silhouette analysis, or other techniques to determine the optimal number of clusters (k) for the K-Means algorithm. Evaluate the performance of different values of k and select the optimal value based on the results.
- **Cluster Analysis and Interpretation**: Analyze the clusters formed by the K-Means algorithm. Examine the characteristics of each cluster, such as the average plastic waste generation, the geographical distribution, and any other relevant features. Interpret the clusters to gain insights into the patterns and trends in plastic waste generation.
- **Visualization and Reporting**: Create visualizations to present the results of the clustering analysis. Use charts, graphs, and maps to visualize the clusters, their characteristics, and the distribution of plastic waste across different regions. Prepare a report or presentation to summarize the findings, methodology, and conclusions.

```
# Run this cell to install geopandas
!pip install geopandas
!apt-get install -qq -y libspatialindex-dev
!pip install -q -U git+git://github.com/geopy/geopy
!pip install -q -U git+git://github.com/gtsa/geoplot
!pip install -q -U git+git://github.com/jtmiclat/cartopy
```

```
pip3 [-h] [-V] [-vvv] [-q] [-y] [-c CONFIG_FILE]
                  [--log-level LOG_LEVEL] [--log-file LOG_FILE]
                  [--no-cache-dir] [--no-input] [--proxy PROXY]
                  [--retries RETRIES] [--timeout TIMEOUT] [--exists-action
                  EXISTS_ACTION] [--trusted-host TRUSTED_HOST] [--cert CERT]
                  [--client-cert CLIENT_CERT] [--cache-dir CACHE_DIR]
                  [--no-cache-dir] [--disable-pip-version-check]
                  [--no-color] [--no-python-version-warning]
                  [--use-feature feature] [--use-deprecated deprecated]
                  [--keyring-provider {auto,disabled,import,subprocess}]
                  [--break-system-packages] [--pre]
                  [--dry-run] [--prefer-binary] [--force-reinstall]
                  [--upgrade] [--ignore-installed ignore_installed]
                  [--no-build_isolation] [--no-deps]
                  [--no-binary :none: | :all: | package_name[,...]]
                  [--only-binary :none: | :all: | package_name[,...]]
                  [--prefer-binary] [--require-virtualenv]
                  [--no-compile] [--no-warn-script-location]
                  [--no-warn-conflicts] [--no-clean] [--no-index]
                  [--quiet] [--success-exit-status status]
                  [--root-user-action ROOT_USER_ACTION]
                  [--report report] [--progress-bar progress_bar]
                  [--root root] [--prefix prefix] [--target dir]
                  [--user] [--no-setuptools] [--no-wheel]
                  [--no-pyproject-build] [--no-cache] [--no-deps]
                  [--no-binary :none: | :all: | package_name[,...]]
                  [--only-binary :none: | :all: | package_name[,...]]
                  [--prefer-binary] [--force-reinstall]
                  [--upgrade] [--ignore-installed ignore_installed]
                  [--no-build_isolation
```