



CDS6214

Data Science Fundamentals

Project (40%)

Tutorial Section: TT7L

Group Number: G11

ID	Name	Email Address
1221304179	Mustafa Yousif	1221304179@student.mmu.edu.my

YouTube Link	<a href="https://youtu.be/dsFiweTweVc">https://youtu.be/dsFiweTweVc</a>
--------------	---

## Contribution table

Name	Questions	Contribution
Mustafa Yousif	All Questions	100%

<b>Problem Statement and Motivation</b>	<b>3</b>
<b>Process</b>	<b>4</b>
1. Acquiring Datasets	4
Structure	5
Problems faced	6
2. Data Wrangling/Cleaning	6
Problems Faced at this stage	8
<b>Question 1: Who are the top producers of Cocoa globally</b>	<b>9</b>
<b>Pipeline:</b>	<b>9</b>
Data Wrangling:	9
Exploratory Data Analysis:	10
Key Observations	10
<b>Question 2 : Is there a correlation between drought indicators on Cocoa production</b>	<b>14</b>
<b>Pipeline:</b>	<b>15</b>
<b>Question 3 : How does area harvested affect cocoa production?</b>	<b>19</b>
<b>References</b>	<b>22</b>

## **Studying trends in Cocoa production and its correlation with Drought Indicators**

### **Problem Statement and Motivation**

The global cocoa industry is facing a significant challenge due to an ongoing cocoa shortage, which has led to a dramatic increase in cocoa prices. This shortage is particularly concerning as it enters its fourth consecutive year, primarily affecting the world's top cocoa-producing countries, Côte d'Ivoire, the Congo and Ghana. Understanding the factors contributing to this shortage, particularly the role of environmental conditions such as droughts, is crucial for developing sustainable solutions to ensure the long-term viability of the chocolate industry.

### **Impact on Communities, Society, and the Nation**

Cocoa production and pricing strategies can have significant impacts on corporations, stakeholders, and have wide-reaching implications on global trade..

#### **Impact on Communities:**

- **Sustainable Cocoa Farming:** Insights into the effects of environmental factors on cocoa production can help farmers implement more sustainable and resilient farming practices, ensuring the long-term prosperity of cocoa-growing regions.
- **Economic Stability:** Cocoa is a vital export crop for many developing countries, particularly in Africa. Improving cocoa production and supply chain stability can contribute to economic growth, job creation, and poverty alleviation in these communities.
- **Food Security:** Cocoa is an essential ingredient in chocolate, a popular and widely consumed food product. Addressing cocoa shortages can help maintain the availability and affordability of this staple food, promoting food security.

#### **Impact on Society:**

- **Sustainable Chocolate Industry:** Understanding the drivers of cocoa shortages and developing strategies to mitigate them can help ensure the long-term viability of the global chocolate industry
- **Equitable Value Chain:** Analyzing the dynamics of the cocoa industry can inform policies and initiatives that promote fairness, transparency, and equity throughout the supply chain, empowering small-scale farmers and improving their livelihoods.
- **Environmental Sustainability:** Investigating the role of environmental factors, such as droughts, in cocoa production can lead to the development of more sustainable agricultural practices that minimize the industry's environmental footprint.

#### **Impact on the Nation:**

- **Economic Implications:** Malaysia is also an exporter of Cocoa beans, studying the effect of the shortages in cocoa can provide farmers here in Malaysia with insights about trends in prices and capture key market entries, allowing them to prioritize this crop over
- **Global Competitiveness:** Ensuring a reliable and sustainable cocoa supply can enhance Malaysia's competitiveness in the global chocolate industry, attracting investments, fostering innovation, and strengthening its international reputation, especially from heavy consumers of this commodity like the US and Germany.

By addressing the challenges facing the cocoa industry, particularly the effects of environmental factors on production, this analysis aims to contribute to the incorporation of insights into cocoa production everywhere and understanding what causes fluctuation in prices of the cocoa bean market

## Process

We will be following the standard data science process:

- **Acquiring Datasets**
- **Data Wrangling/Cleaning**

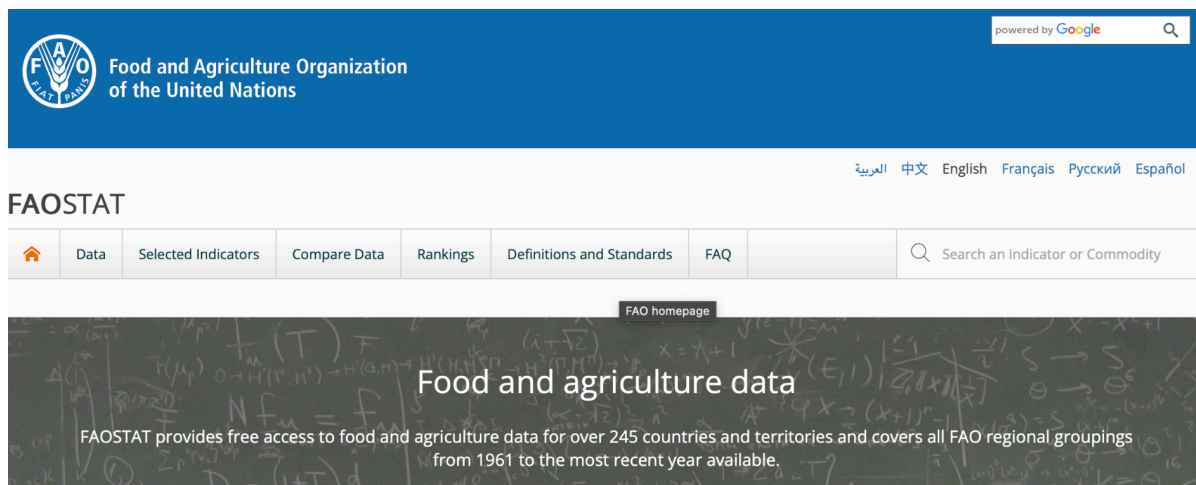
And for each question we will be doing:

- **EDA**
- **Modeling** (if applicable)

### 1. Acquiring Datasets

I aggregate data from a couple of trusted sources, some more than others (FAO for example), with hopes of combining all the data into a unified Dataframe.

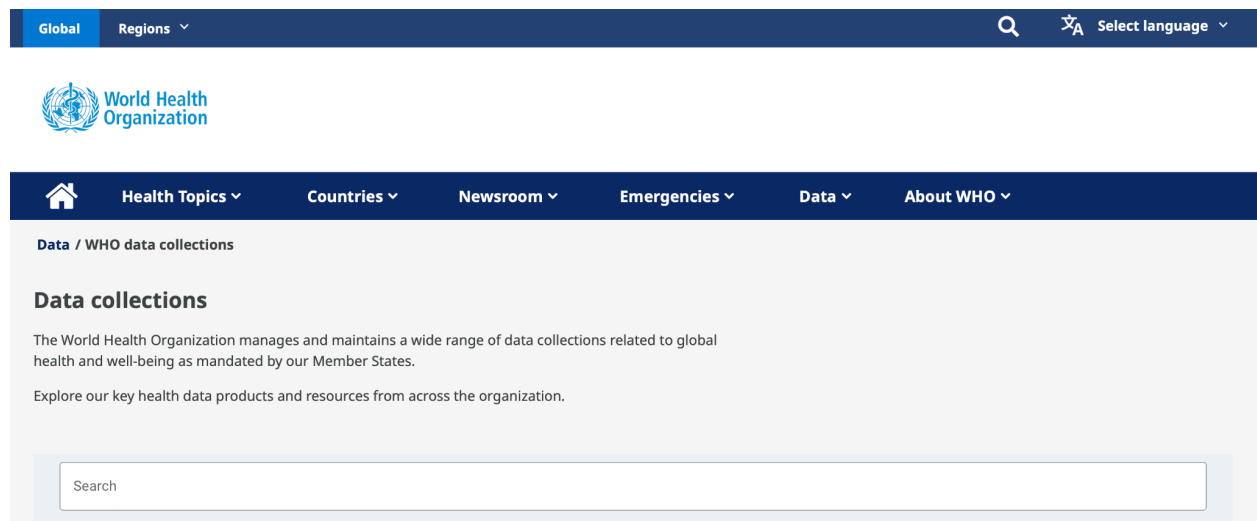
## FAO



The screenshot shows the FAO (Food and Agriculture Organization of the United Nations) website. The header is blue with the FAO logo and text. Below the header, there are language options: العربية, 中文, English, Français, Русский, Español. The main section is titled "FAOSTAT" and features a navigation bar with links: Home, Data, Selected Indicators, Compare Data, Rankings, Definitions and Standards, and FAQ. A search bar is also present. Below the navigation bar, there is a large banner with the text "Food and agriculture data" and a subtext: "FAOSTAT provides free access to food and agriculture data for over 245 countries and territories and covers all FAO regional groupings from 1961 to the most recent year available."

[FAO](#) is a specialized agency of the United Nations that leads international efforts to defeat hunger and improve nutrition and food security. The FAO has a dedicated mandate to collect, analyze, and disseminate data on agricultural production, including cocoa, across the world. As a global authority on food and agriculture, the FAO is recognized for its expertise and the reliability of the data it provides. It also adheres well-established methodologies and protocols for data collection. I also chose this source especially as they have a lot of other datasets besides drought and cocoa bean production datasets.

## WHO



WHO is a trusted source for the same reasons FAO is, and it is quite reputable. The citation for the specific dataset by Rustemeier Elke is found in the list of references..

## Structure

All the datasets are stored in the ``/datasets/`` folder:



In the following structure:

<input type="checkbox"/> <a href="#">asi</a>	2 days ago	
<input type="checkbox"/> <a href="#">development_indices</a>	2 days ago	
<input type="checkbox"/> <a href="#">ndvi</a>	a day ago	
<input type="checkbox"/> <a href="#">un_exports</a>	2 days ago	
<input type="checkbox"/> <a href="#">vhi</a>	a day ago	
<input type="checkbox"/> <a href="#">annual_surface_temperature_change.csv</a>	a day ago	189 kB
<input type="checkbox"/> <a href="#">atmospheric_co_concentrations.csv</a>	a day ago	906 kB
<input type="checkbox"/> <a href="#">change_in_mean_sea_level.csv</a>	a day ago	15.3 MB
<input type="checkbox"/> <a href="#">cleaned.csv</a>	3 hours ago	354 kB
<input type="checkbox"/> <a href="#">climate_disasters.csv</a>	a month ago	34.7 kB
<input type="checkbox"/> <a href="#">climate_related_disasters_frequency.csv</a>	a day ago	482 kB
<input type="checkbox"/> <a href="#">cocoa_production_by_continent</a>	a month ago	1.26 MB
<input type="checkbox"/> <a href="#">cocoa_production_fao.csv</a>	a month ago	1.72 MB
<input type="checkbox"/> <a href="#">commodity_prices.csv</a>	a month ago	11.4 kB
<input type="checkbox"/> <a href="#">country_income_group.csv</a>	7 days ago	59.4 kB
<input type="checkbox"/> <a href="#">drought_events.csv</a>	a month ago	1.74 MB
<input type="checkbox"/> <a href="#">forrests_and_carbon.csv</a>	a day ago	831 kB
<input type="checkbox"/> <a href="#">...</a>	a month ago	1.92 MB

ASI, NDVI, and VHI are all indicators, either for crop health or drought intensity, and since there are a lot of related datasets, I store them in the same folder for ease of access.

**Notes:** For some datasets, I had to write a script to fetch and download set of datasets and read them and clean them

## Problems faced

- **Aggregate many datasets from different sources:** In trying to get all the data about as much countries as I can meant I have to get different datasets from different sources, and try to merge them together. I faced problems trying to standardise everything, sometimes because different sources measure metrics differently and even have different names for the same countries, etc, meaning I had to do a lot of manual data cleaning and rely on pip libraries like pycountry.
- **Tedious to download:** Some of these datasets can only be downloaded for each country, one such example is the FAO's VHI index dataset. I could not find a way to bulk download, and I considered writing a script that downloads all the needed datasets.
- **GeoTIFF format:** A lot of these datasets are downloaded in geoTIFF format, which is image based, and to convert that into dataframes was hard and tedious
- **Broken Links:** A lot of the download links didn't work, which meant I had to spend a lot of time looking for alternatives.

## 2. Data Wrangling/Cleaning

In this stage the focus is data cleaning and preparation process. By the end of this stage we should have a cleaned dataset that is processed, ready for the next stage. This is a crucial step to

ensure the integrity and reliability of the subsequent analyses. I'll do it in such a way that I only have to do it once, by including all the columns and only selecting them later as I need:

1. **Data Loading:** I have script reads the 'cocoa\_production\_fao.csv' file from the 'datasets' directory and loads it into a Pandas DataFrame.
2. **Data Inspection:** The code prints the shape of the DataFrame, the data types of each column, and a concise summary of the data, including column names, non-null counts, and data types.
3. **Handling Missing Data:** The script identifies and removes the empty rows from the DataFrame, as well as the 'Value Footnotes' column, which is not necessary for the analysis.
4. **Data Type Conversion:** The 'Year' column is converted from a float to an integer data type to ensure consistent handling of the years.
5. **Data Sorting:** The DataFrame is sorted in ascending order by the 'Year' column to maintain a chronological order of the data.
6. **Duplicate Handling:** The script checks for and removes any duplicate rows in the DataFrame.
7. **Data Reshaping:** The DataFrame is pivoted to create a wide format, where each unique combination of 'Element' and 'Unit' is a separate column, and the 'Region' and 'Year' are the index.
8. **Encoding Detection:** The script handles the encoding issue encountered when reading the 'asi\_annual\_season\_1.csv' file by using the `chardet` library to detect the encoding. It then defines a function to read and append the data to the main DataFrame.
9. **NDVI Data Integration:** The script reads the 'cote\_dvoire.csv' file from the 'datasets/ndvi' directory, pivots the DataFrame, and adds the 'Normalized Difference Vegetation Index (NDVI)' column to the main DataFrame.
10. **Climate-Related Disasters Data Integration:** The script reads the 'climate\_related\_disasters\_frequency.csv' file, renames the columns, and joins the data with the main DataFrame.
11. **Sea Level Data Integration:** The script reads the 'change\_in\_mean\_sea\_level.csv' file, processes the 'Date' column, and joins the sea level data with the main DataFrame.
12. **Atmospheric CO2 Data Integration:** The script reads the 'atmospheric\_co\_concentrations.csv' file, processes the 'Date' column, and joins the atmospheric CO2 data with the main DataFrame.
13. **Surface Temperature Data Integration:** The script reads the 'annual\_surface\_temperature\_change.csv' file, processes the data, and joins the surface temperature data with the main DataFrame.
14. **Country and Continent Separation:** The script separates the data into two DataFrames: one for countries and one for continents, based on the 'Region' column.
15. **Continent Assignment:** The script uses the `pycountry_convert` library to assign a continent to each country in the 'countries' DataFrame.

**16. Filling in Missing Values:** The script fills in the missing values in the 'countries' DataFrame using the mean or median of the respective columns.

Once done with this stage we are sure that the subsequent analysis is performed on a high-quality, integrated dataset. We don't have to visit this stage again for further EDA.

Problems Faced at this stage

- **Many indicators and no clear consensus:** An issue is that there are various indicators, and there's no information about the weighting of each of the indicators, their reliability or their accuracy.
- **Developing Countries :** Another issue is that as I am investigating countries that are in low on the development index, so this may lead to inaccurate and incomplete datasets, as those countries tend to not have systems in place to track those metrics and report them. Most of the data is collected by NGOs such as the FAO, who don't have datapoints for all countries across all years. This means there are a lot of rows in the datasets that have NAN in them, to the point that replacing those non-values with the mean alters the data too much.
- **Different encodings and corrupt datasets::**

```
File /opt/anaconda3/lib/python3.11/site-packages/pandas/io/parsers/c_parser_wrapper.py:93, in CParserWrapper.__init__
    __ (self, src, **kwargs)
    90 if kwds["dtype_backend"] == "pyarrow":
    91     # Fail here loudly instead of in cython after reading
    92     import_optional_dependency("pyarrow")
--> 93 self._reader = parsers.TextReader(src, **kwargs)
    95 self.unnamed_cols = self._reader.unnamed_cols
    97 # error: Cannot determine type of 'names'

File parsers.pyx:579, in pandas._libs.parsers.TextReader.__cinit__()
File parsers.pyx:668, in pandas._libs.parsers.TextReader._get_header()
File parsers.pyx:879, in pandas._libs.parsers.TextReader._tokenize_rows()
File parsers.pyx:890, in pandas._libs.parsers.TextReader._check_tokenize_status()
File parsers.pyx:2050, in pandas._libs.parsers.raise_parser_error()

UnicodeDecodeError: 'utf-8' codec can't decode byte 0xf4 in position 75346: invalid continuation byte
```

A lot of the datasets had obscure encodings and since many of them came from different sources, they had different formats that meant I spent some time converting them into a unified format or discarding them altogether.



## Question 1: Who are the top producers of Cocoa globally

Identify key players in the global cocoa market, understand the concentration of production and potential risks or vulnerabilities in the global cocoa supply if production is heavily skewed toward a few countries. It also investigates the factors that contribute to the high productivity of the top-producing countries, such as climate, and weather conditions, percentage of land utilized, etc.

Key Points Addressed:

### 1. **Dominant Producers:**

- Identify the top cocoa-producing countries globally
- Identify factors that affect those countries' cocoa production capability

### 2. **Continental Focus:**

- Identify the major continents which dominate global production and analyze environmental conditions there
- Consider secondary focuses on South America and Asia for inferential analysis

### 3. **Temporal Trends:**

- Analyze the exponential growth in African cocoa production since the 1980s
- Investigate any historical dips in production and their potential link to droughts

### 4. **Environmental Factors:**

- Focus on drought effects using various indicators, but also consider other climate change-related factors
- Explore the potential for developing drought-resistant cocoa varieties or improved farming practices

By focusing on these aspects and countries, we provide insights into the potential impacts of droughts on major producing regions, from the top producing continents, and then the ranking of countries in terms of their production within that continent, in this case it was Africa, and the 3 countries, Ghana, Congo and Cote Dvoire.

Pipeline:

Data Wrangling:

We'll be using the same dataset that we have produced in the cleaning stage, `'datasets/cleaned.csv'`, and drop any unnecessary columns.

Exploratory Data Analysis:

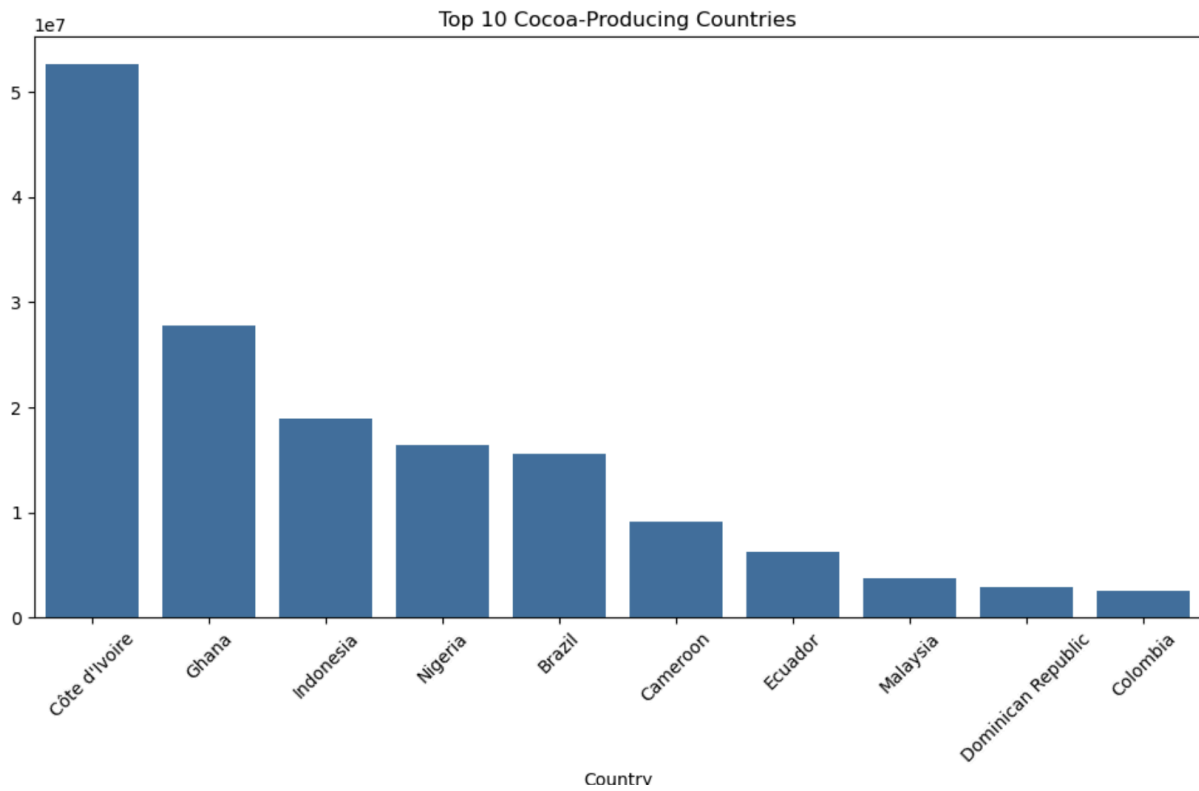
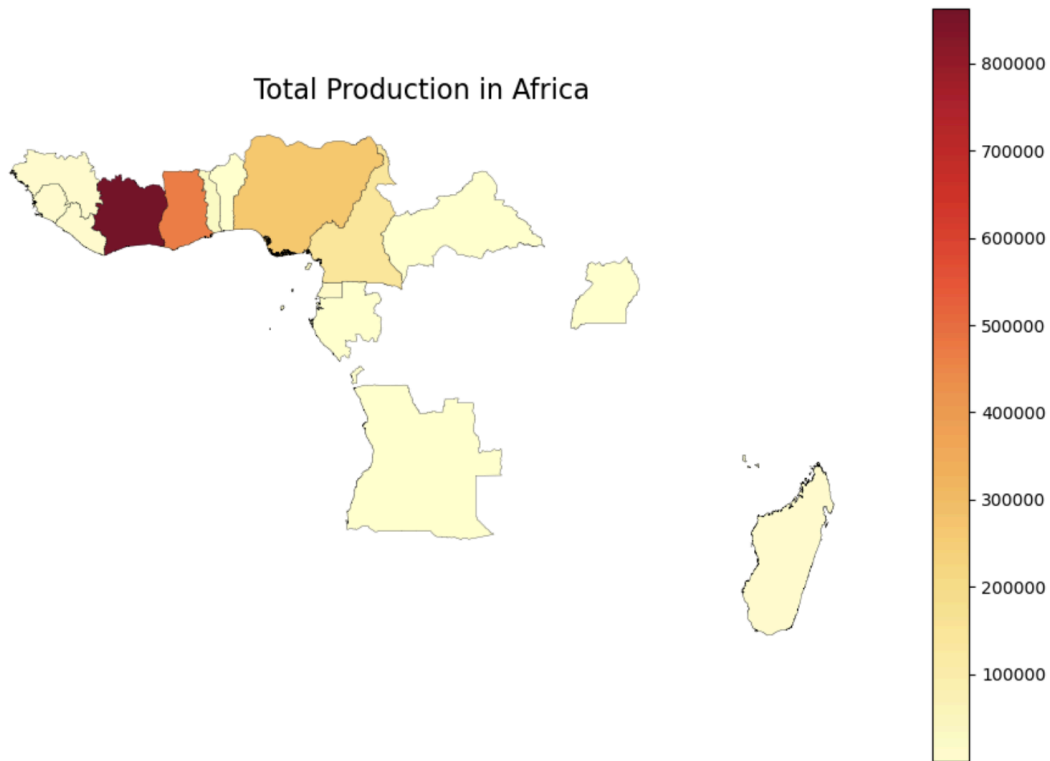
Visualizations: Choropleth, Bar Chart, Line Chart, Pie Charts, Heatmaps, Boxplots

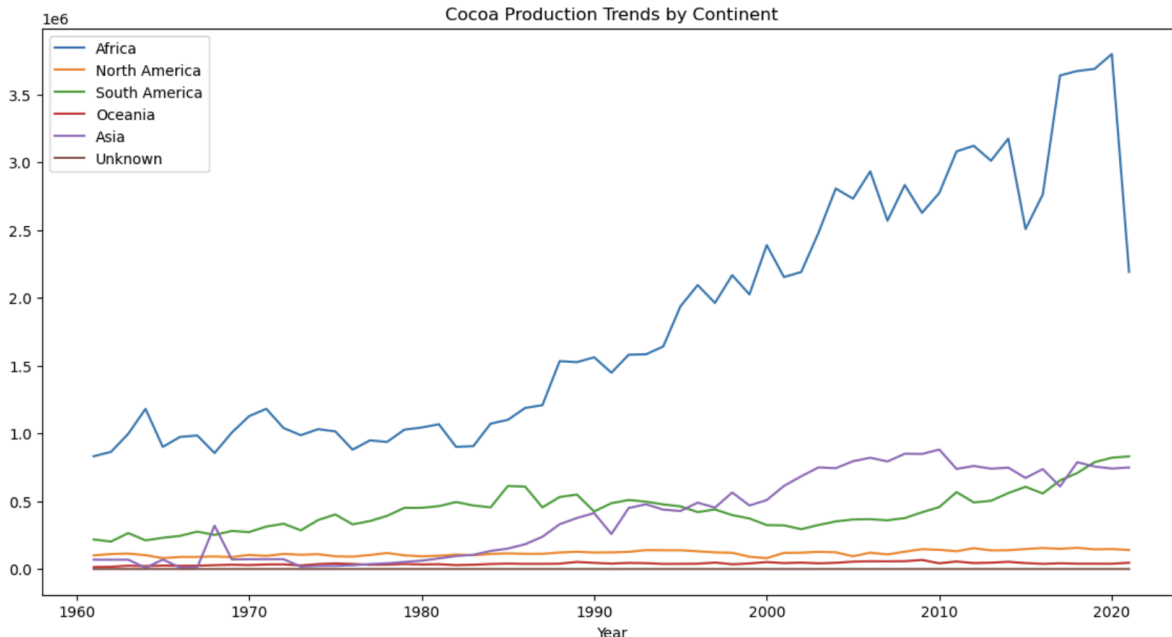
## Findings:

1. **Overwhelming Market Share:** The graph clearly shows that Africa has been the dominant continent in cocoa production since the 1960s, and its lead has grown significantly over time.
2. **Exponential Growth:** Africa's cocoa production has seen exponential growth, especially from the 1980s onward. The production has increased from about 1 million tonnes in 1980 to over 3.5 million tonnes by 2020.
3. **Widening Gap:** While other continents have shown some growth in cocoa production, the gap between Africa and the rest of the world has widened considerably over the decades.
4. **Market Share Increase:** Africa's share of global cocoa production has increased substantially. By 2020, it appears to account for well over 70% of the world's cocoa production.
5. **Consistent Leadership:** Despite some fluctuations, Africa has maintained its leadership position throughout the entire period shown in the graph (1960-2020).
6. **Resilience:** The graph shows some dips in production, possibly due to environmental or economic factors, but Africa's cocoa industry has consistently rebounded and continued its upward trend.

## Key Observations

1. **Clear Leadership:** Côte d'Ivoire stands out prominently as the top cocoa-producing country, with its bar towering over all others in the graph.
2. **Production Scale:** The y-axis indicates that Côte d'Ivoire's production is over 5 million tonnes (5e7 on the scale), which is a massive volume compared to other producers.
3. **Margin of Dominance:** The gap between Côte d'Ivoire and the second-largest producer (likely Ghana, based on typical cocoa production statistics) is substantial. Côte d'Ivoire's production appears to be nearly twice that of the next country.
4. **Relative Scale:** Côte d'Ivoire's production dwarfs that of most other countries in the top 10. Many of the lower-ranked countries produce less than 1 million tonnes, highlighting the scale of Côte d'Ivoire's output.
5. **African Dominance:** Given that Côte d'Ivoire is in Africa, this graph reinforces the continent's dominance in global cocoa production, with at least two African countries (Côte d'Ivoire and likely Ghana) at the top of the list.





## Question 2 : Is there a correlation between drought indicators on Cocoa production

Explore the relationship between drought indicators, such as the Vegetation Health Index (VHI) and Normalized Difference Vegetation Index (NDVI), and cocoa production metrics, including yield and total production. The goal is to understand how these drought-related factors correlate with and potentially impact cocoa cultivation across different continents:

### 1. VHI Correlation:

- Examine the correlation between VHI % below 35 (an indicator of drought conditions) and cocoa yield.
- Assess the strength and direction of this relationship to understand the potential impact of drought on yields.

### 2. NDVI Regression:

- Investigate the linear relationship between NDVI (a measure of vegetation health) and cocoa yield using a regression model.
- Evaluate the model's performance and interpretability to determine the viability of using NDVI as a predictor of yields.

### 3. Continental Perspectives:

- Analyze the average VHI % below 35 by continent to identify regional differences in drought conditions.
- Explore how these continental variations in drought indicators may be related to cocoa production patterns.

## **Pipeline:**

### **1. Data preprocessing**

We'll be using the same dataset that we have produced in the cleaning stage, `datasets/cleaned.csv`, and drop any unnecessary columns.

### **2. Exploratory Data Analysis:**

- Visualizations: Listings map, neighborhood stats, price distribution

Based on the provided data and analyses, we can conclude that both the Vegetation Health Index (VHI) and the Normalized Difference Vegetation Index (NDVI) show weak correlations with cocoa yield. This conclusion is supported by the following observations:

#### **1. VHI Correlation:**

- The correlation heatmap showed a weak negative correlation (-0.051) between VHI % below 35 and yield.
- This suggests that drought conditions, as indicated by low VHI values, have only a minimal impact on cocoa yields.

#### **2. NDVI Linear Regression:**

- The linear regression between NDVI and yield produced a negative R-squared score (-0.50), indicating a very poor fit.
- The scatter plot revealed high variability and no clear linear relationship between NDVI and yield.

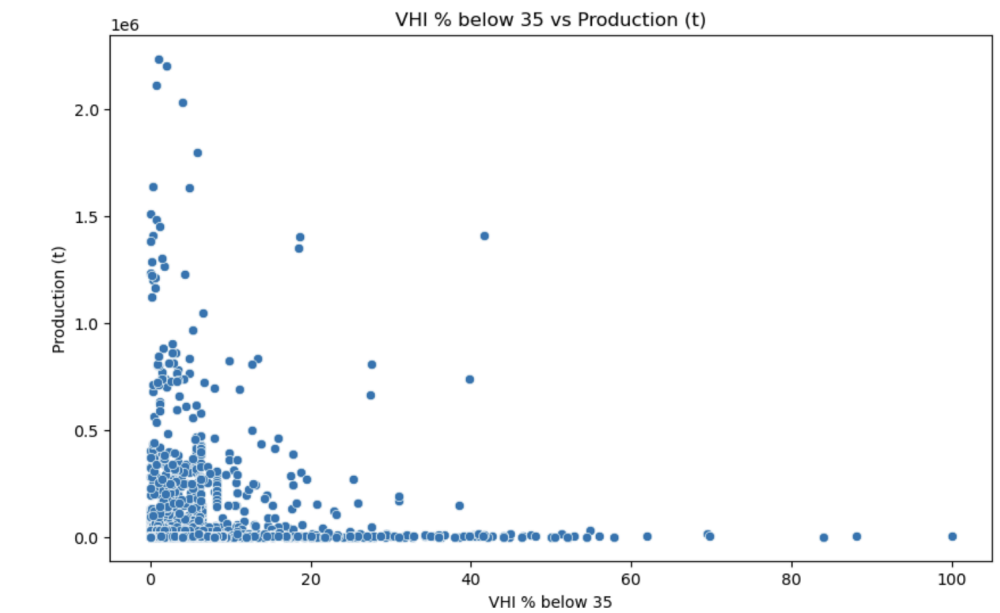
#### **3. Implications:**

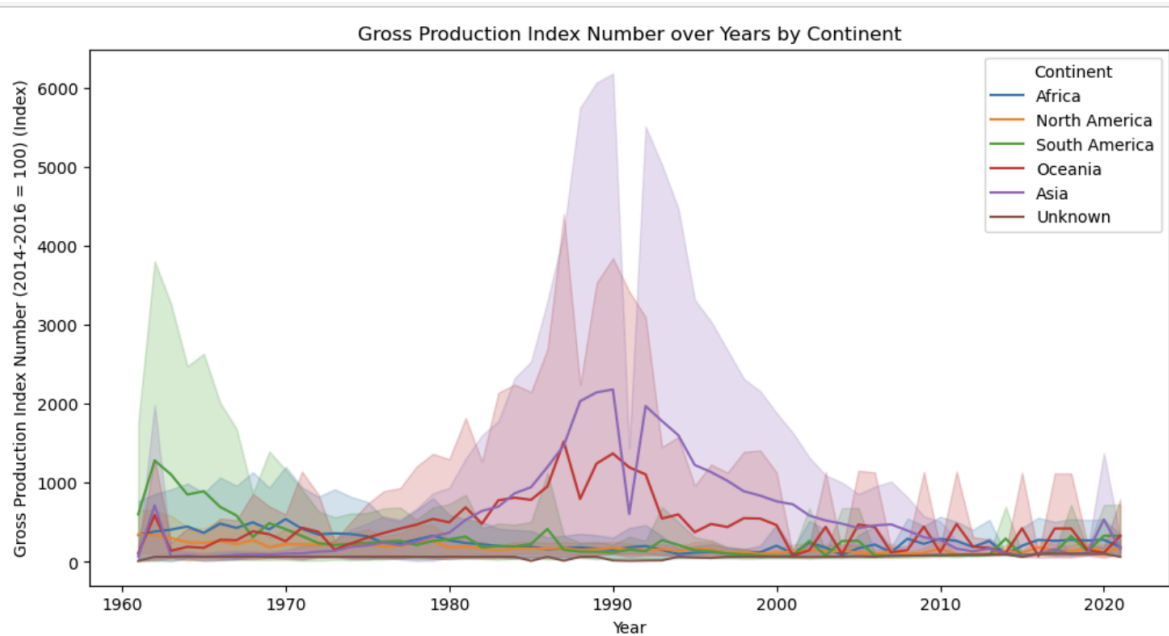
- These weak correlations suggest that using VHI or NDVI alone may not be reliable predictors of cocoa yield.
- Other factors, such as agricultural practices, soil conditions, or more complex interactions of environmental variables, likely play more significant roles in determining cocoa yields.

#### **4. Future Directions:**

- More sophisticated multivariate analyses or non-linear models may be necessary to accurately predict cocoa yields.
- Incorporating additional environmental and agronomic factors could potentially improve yield prediction models.

In summary, while vegetation indices like VHI and NDVI provide valuable information about plant health and environmental conditions, their direct relationship with cocoa yield appears to be limited based on the available data.

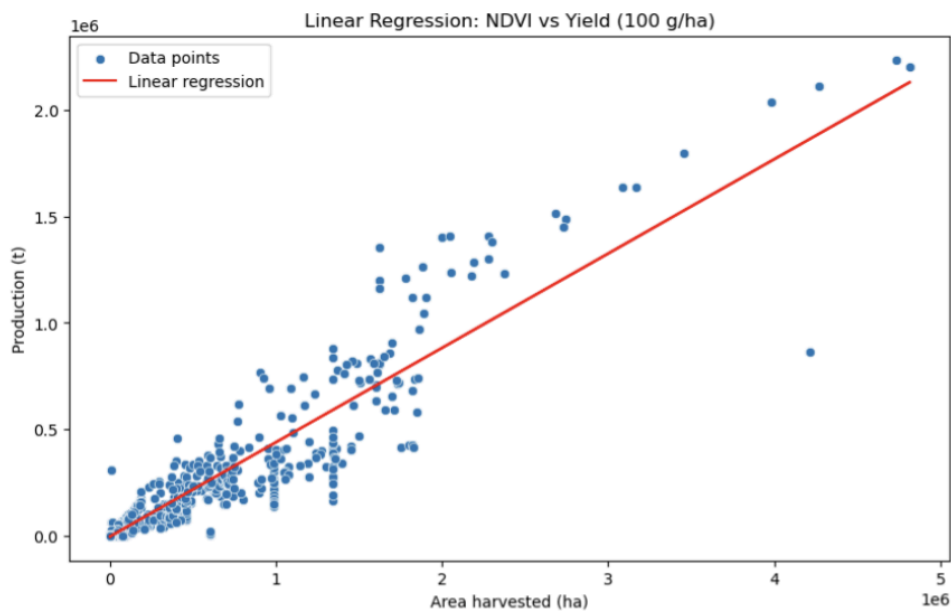




### Question 3 : Predict production based on area harvested affect cocoa production?

Measure the correlation between area harvested and cocoa production, see how strong that relationships is and how good of an indicator it is to predict area harvested.

Mean squared error: 3558165949.37  
R-squared score: 0.82



Intercept: -3983.89  
Coefficient: 0.44  
Number of data points used: 3133

**Model Development:**

### Algorithm: Logistic Regression

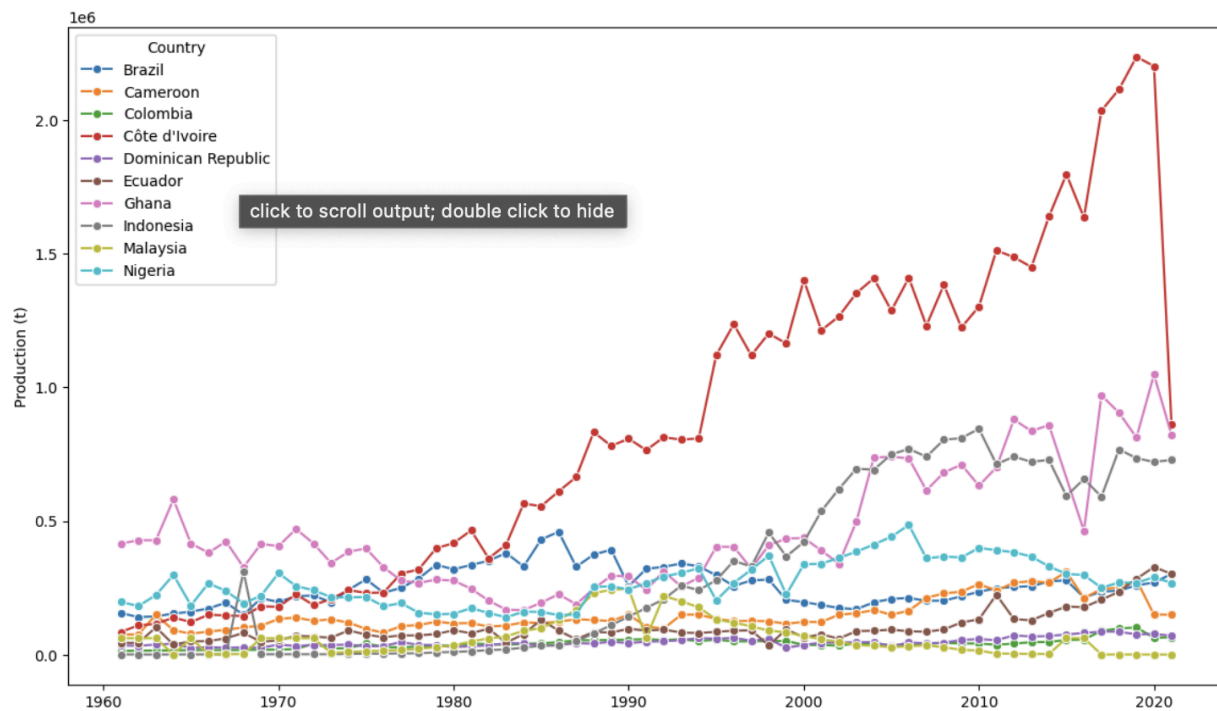
Predict Production based on Area Harvested

**Feature** : Area Harvested

- normalization of numerical data
- Production based on Area harvested linear regression has a R-score of 0.82, which is quite decent and suggests there is a strong correlation.

### Question 4: What are the trends in Cocoa production?

Covering a wide range of factors, we explore projections of cocoa production and overall trends over the past decade.



### Analysis and Conclusion

From the plot showing cocoa production over time in the top-producing countries, we can draw several conclusions:

1. Côte d'Ivoire Dominance: Côte d'Ivoire is the leading cocoa producer by a significant margin, especially noticeable from the late 1980s onwards. Its production has shown a steady and significant increase over time, with occasional fluctuations.



2. Growth in Ghana and Indonesia: Ghana and Indonesia are also major cocoa producers. While Ghana shows consistent growth with some fluctuations, Indonesia's production seems to peak around the 1990s and then stabilizes with minor variations.
3. Fluctuating Trends: Countries like Brazil, Nigeria, and Ecuador show more fluctuation in their production trends. Brazil's production shows growth till around 2010, followed by a decline. Nigeria and Ecuador have less pronounced trends but do show variations in production.
4. Stable or Declining Trends: Malaysia and the Dominican Republic exhibit relatively stable or declining trends in production, indicating possible stagnation or reduction in cocoa farming activities.

## References

- © FAO. FAOLEX Database. Seasonal 2 . License: CC BY-NC-SA 3.0 IGO. Extracted from: [www.fao.org/faolex/opendata](http://www.fao.org/faolex/opendata). Date of Access: 22/04/2024
- Rustemeier Elke; Hänsel, Stephanie; Finger, Peter; Schneider, Udo; Ziese, Markus (2022): GPCC Climatology Version 2022 at 0.5°: Monthly Land-Surface Precipitation Climatology for Every Month and the Total Year from Rain-Gauges built on GTS-based and Historical Data. DOI: [10.5676/DWD\\_GPCC/CLIM\\_M\\_V2022\\_050](https://doi.org/10.5676/DWD_GPCC/CLIM_M_V2022_050)