# BURKINA FASO
# cereal production forecasting

*by Youssef SAWADOGO*

# table of contents.

- Introduction & Objectives
- Data Collection & Cleaning
- Exploratory Data Analysis
- Interactive visualization
- Prediction Modeling

# INTRODUCTION

The agricultural sector, one of the pillars of Burkina Faso's economy, relies heavily on the cultivation of cereals. Against a backdrop of increasing climatic variability, producers and decision-makers are facing **growing uncertainty about annual yields.**

To meet these challenges, this project aims to design **an interactive platform for visualizing and forecasting cereal production**, integrating data on production, climate and advanced modeling tools. This platform has been designed to be accessible to **farmers, cooperatives and public decision-makers**, with the aim of supporting agricultural planning, climate change adaptation and food security.

The overall aim of the project is to **develop a high-performance predictive model for forecasting cereal production.**
To guarantee scientific and operational reproducibility, all stages from data collection to prediction are encapsulated in a **Github repository: github.com/Youssef-Amine/cereal-production-forecast**

The specific objectives of the project are:

# OBJECTIVES

**Create a database on cereal production with:**
*At least 5 cereals*
*At least 5 dependent variables*
*Over at least 25 years*

Create a web application for interactive data visualization

Apply Random Forest and XGBoost prediction algorithms to build a model offering at least 85% accuracy

# DATA COLLECTION

For data collection, we began by researching studies on cereal production in Burkina Faso. From these studies, we found more than 20 variables that could explain changes in cereal production. We then scoured several sources to collect these data, while respecting the criteria defined by specific objective 1.

We were able to gather 8 different datasets from the **burkinafaso.opendataforafrica.org** website, grouping together the following variables:

- Région [***Region: there are 13 regions in the country***]
- Type de céréale [***Type of cereal: peanut, cotton, corn, millet, cowpea rice, sorghum***]
- Année [***Year: 1996 to 2022***]
- Production (en tonne) [***quantity produced in tons***]
- Superficie (en ha) [***Cultivated area in hectares***]
- Précipitations moyennes annuelles [***Average annual precipitation (mm)***]
- Nombre de jours de pluie [***Number of rainy days per year***
- Température moyenne annuelle [***Average annual temperature (°C)***]
- Humidité relative moyenne (%) [***Average relative humidity (%)***]
- Vent moyen annuel (km/h) [***Average annual wind (km/h)***]
- Durée d'ensoleillement [***Sunshine duration (hours/day)***]

***See these databases in the Github data folder.***

# DATA CLEANING

We performed a data cleaning operation to create a unique database suitable for our project.
The data cleaning steps are:

- **Merging the 8 datasets**

We merged them based on **Region, Type of cereal , and Year**, which are variables common to all 8.
We then filtered and sorted the data so that they were displayed primarily by region, then by type of cereal for all years, and all other variables present.
Since climate variables are not specific to cereals, they only differ by region and year.

- **Managing missing values**

We identified two types of missing values.
*Some missing values on a row*; for example, production values are missing two values for 2011 for the North Central region, or average annual precipitation is missing for the years 1996 to 2000 for the East Central region. To handle this, we **impute using the average of the values** in the corresponding row.

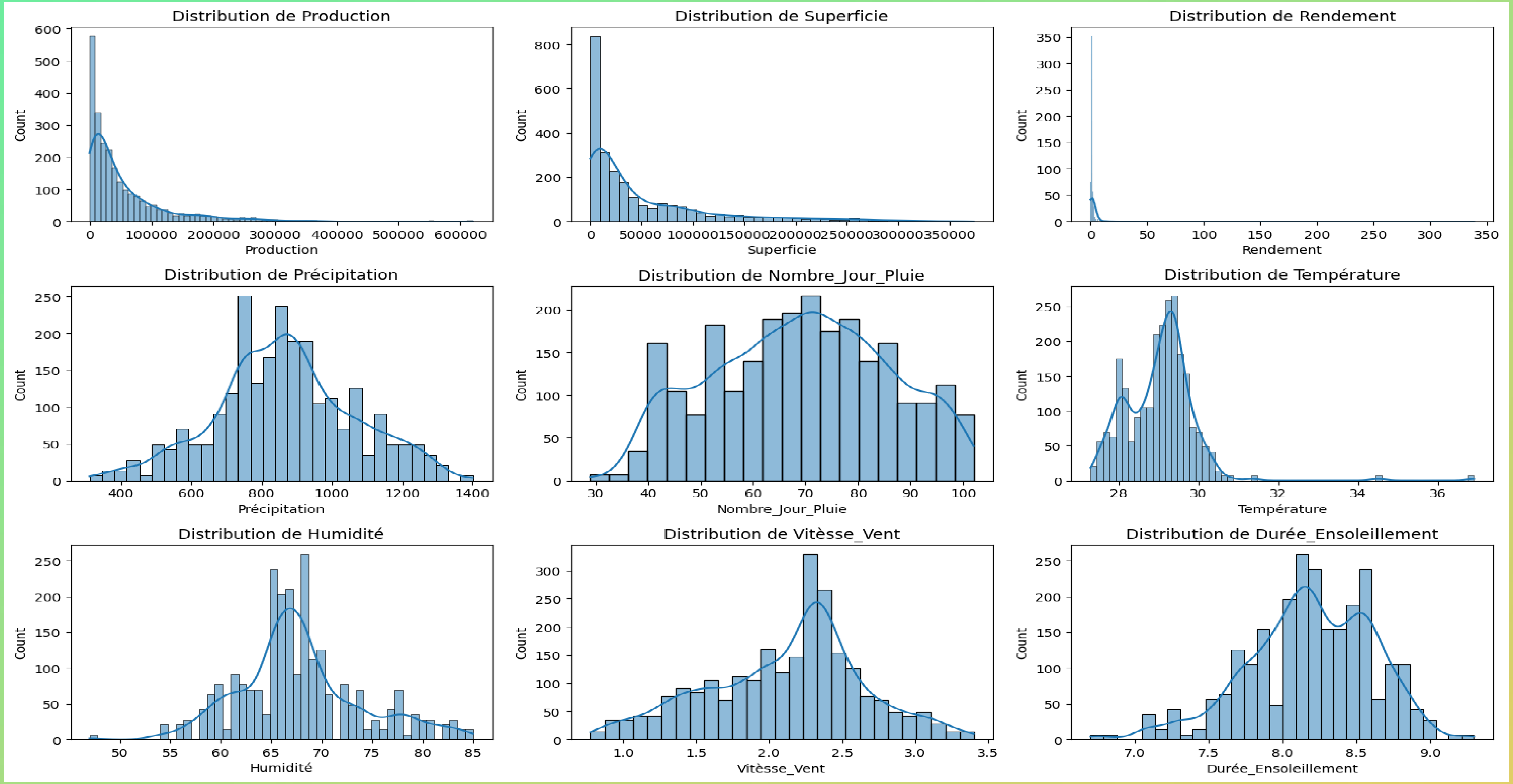
*The second type of missing value is, for example, relative humidity values, which are missing for the Cascades region for all years*. We process them using a method similar to **the k-nearest neighbors**. Indeed, the Cascades region is closer to the Hauts-Bassins and Southwest regions. Thus, for each year, the number of rainy days/year corresponding to the Cascades region will be equal to the average of the values for the Hauts-Bassins and Southwest regions.

**Creation of the yield variable**

This variable is created by calculating the **quotient between production and cultivated area**.

# EXPLORATORY DATA ANALYSIS

## Distribution histogram of each variable



❖ **Production and Area:** Distribution highly skewed to the right (positive),

❖ **Yield**: Distribution highly skewed to the right; the majority of yields are low, with very few high values.

❖ **All climate variables have an approximately normal or pseudo-normal distribution, with slightly skewed**

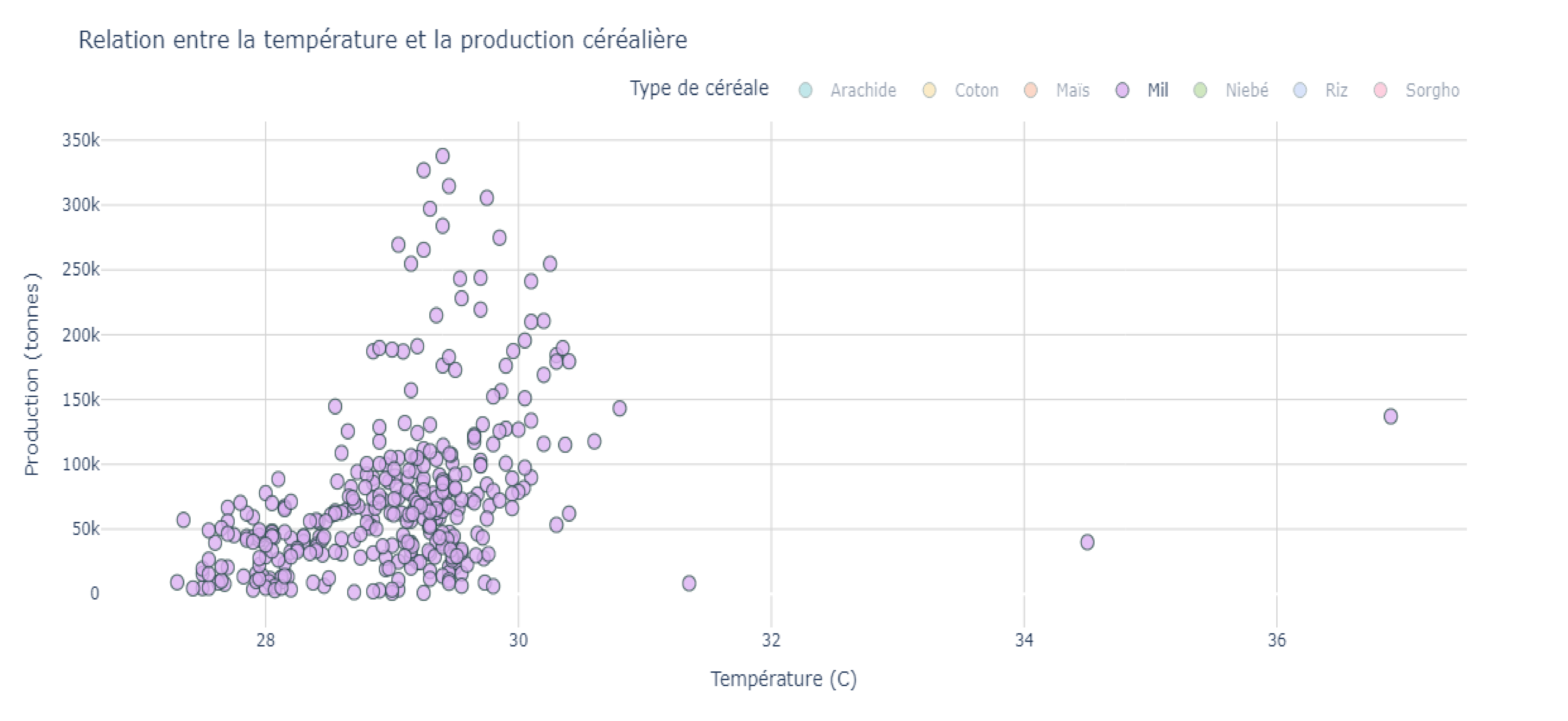
# EXPLORATORY DATA ANALYSIS

## Here we perform analysis to see the correlation between cereal production and temperature.

The graph below, for example, shows the distribution of millet production according to temperature.

We observe a continuous increase in millet production, **up to 350k tons, from 27°C to 30°C**. **Above 31°C, production decreases sharply**.

There appears to be a negative correlation above 30°C: as the temperature increases, millet production tends to decrease.



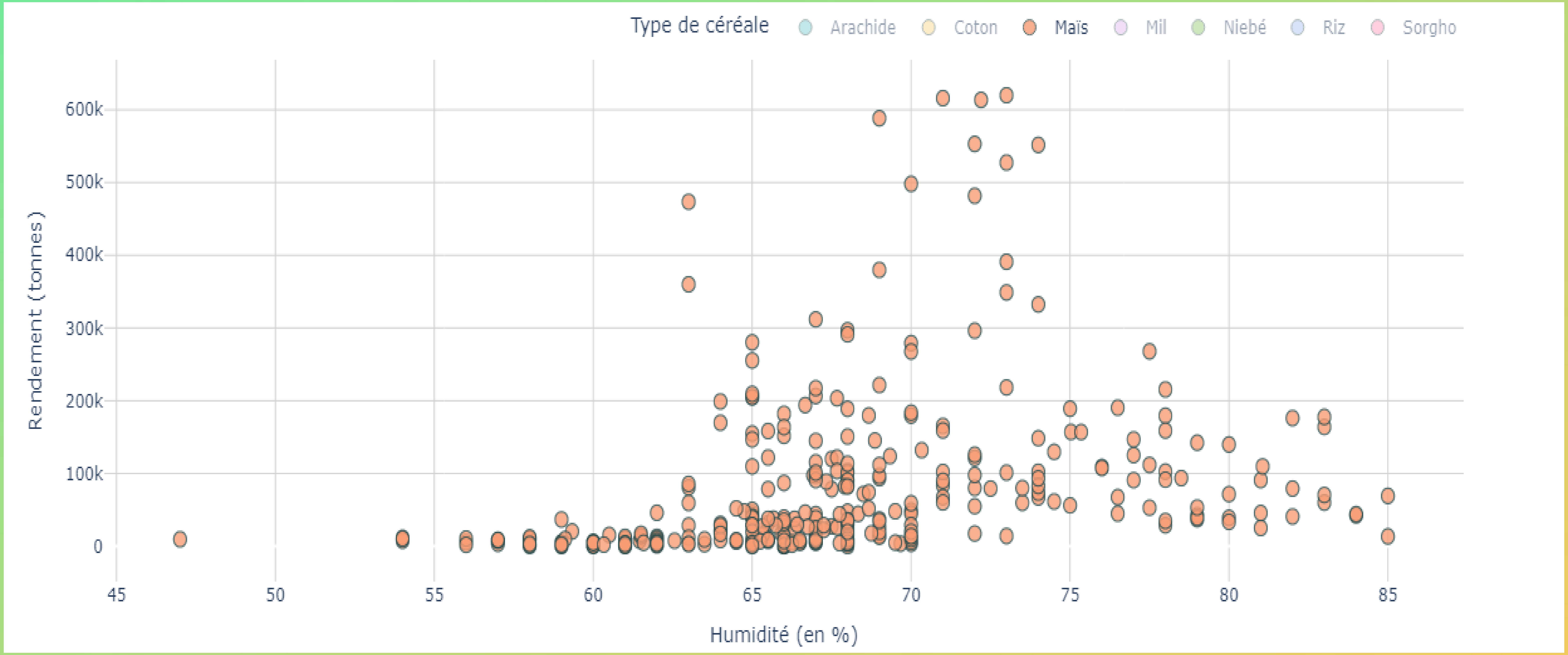
# EXPLORATORY DATA ANALYSIS

## Here we perform analysis to see the correlation between cereal production and humidity.

The graph below, for example, shows the distribution of corn production according to humidity.

General Trend: production generally increases with humidity up to a certain threshold (70%), beyond which it appear to either stagnate or decline. The majority of high productions (>200,000 tonnes) are concentrated between 65% and 72% humidity.

Above 75%, corn production tend to decline, which could indicate saturation or unfavorable conditions for corn growth.



# EXPLORATORY DATA ANALYSIS

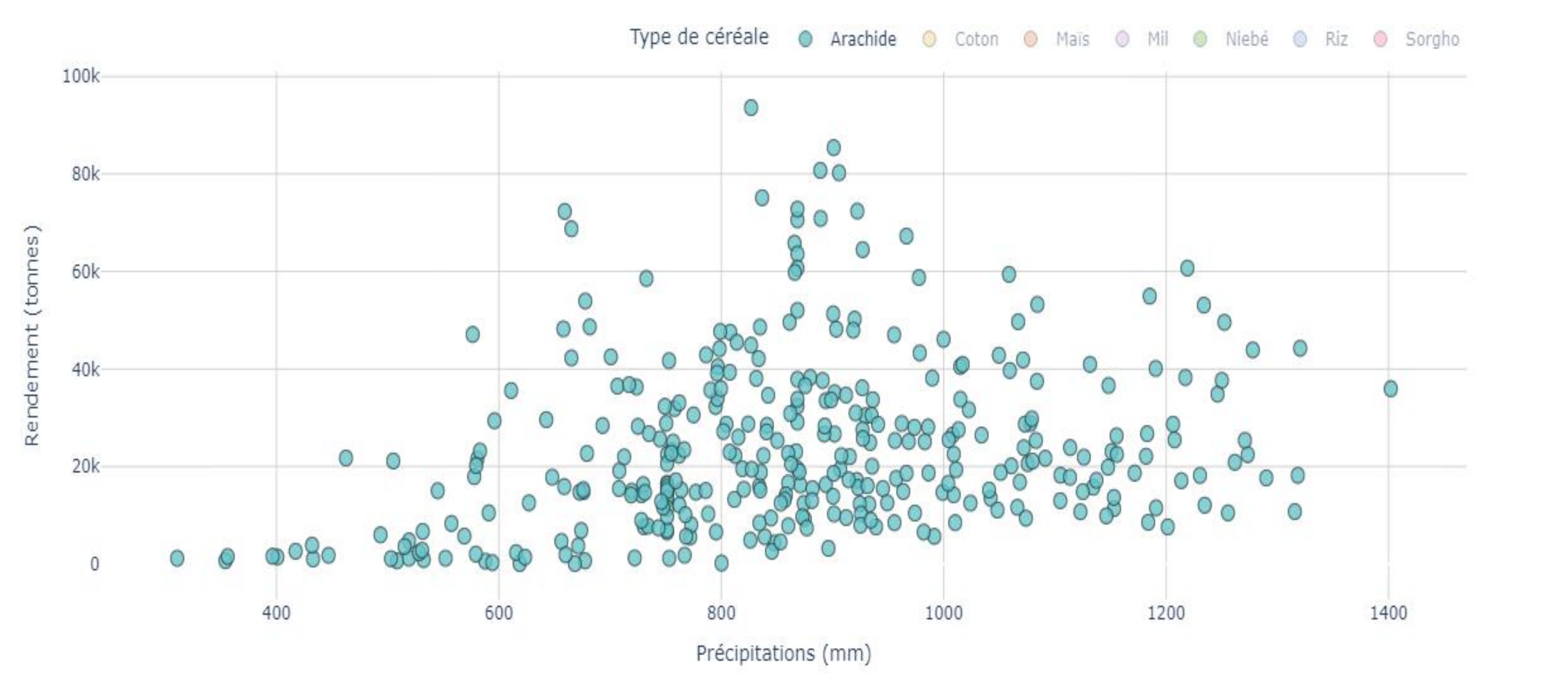
## Here we perform analysis to see the correlation between cereal production and precipitations.

The graph below, for example, shows the distribution of peanuts production according to precipitations.

Most peanut production fall between 600 mm and 1100 mm of rainfall. There is a wide dispersion of production, even for similar rainfall levels.

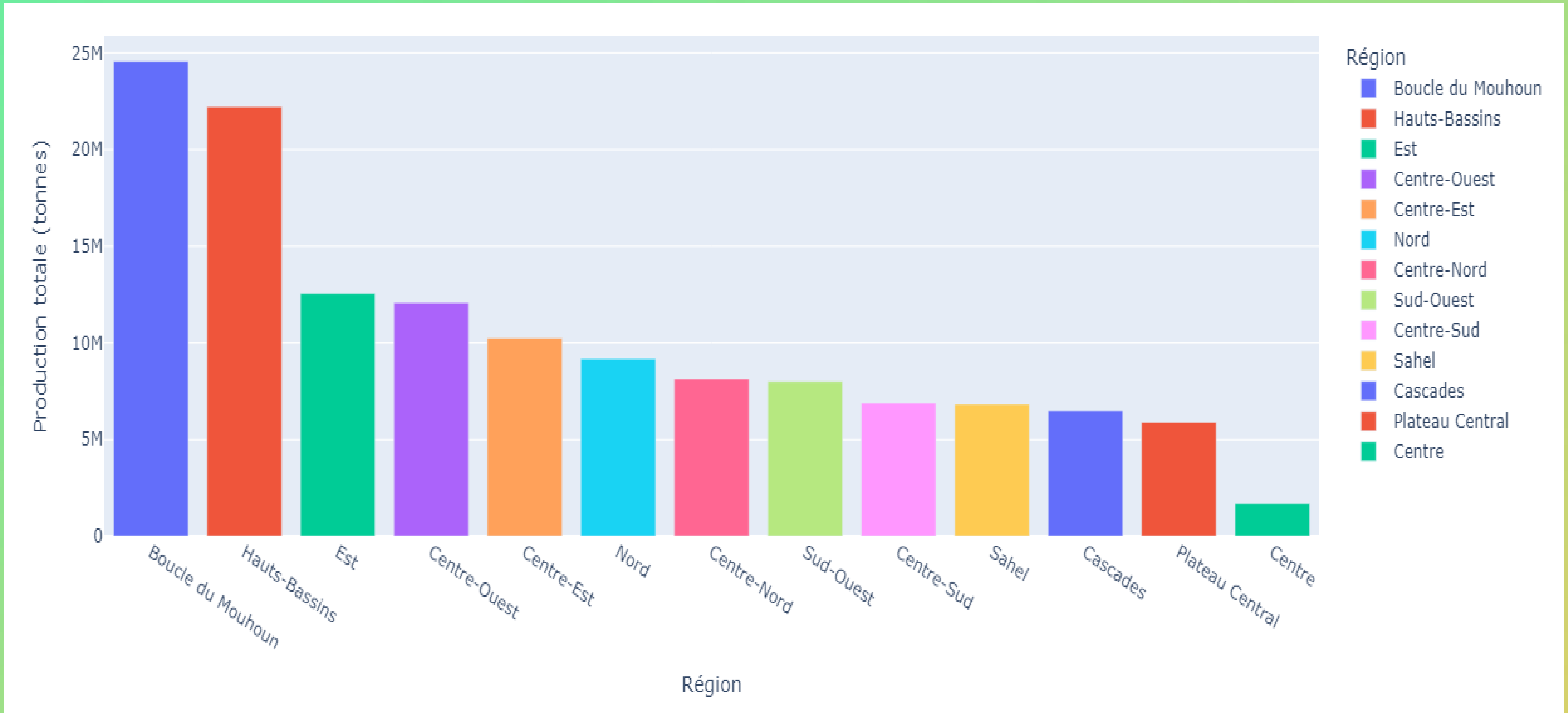
A peak (up to nearly 100k tons) is observed around 800 to 900 mm of rainfall. Below 600 mm, production are generally low.

Beyond 1100 mm, production tend to remain moderate, suggesting a possible negative effect of excess water.



# EXPLORATORY DATA ANALYSIS

## Quantity of cereal production by region



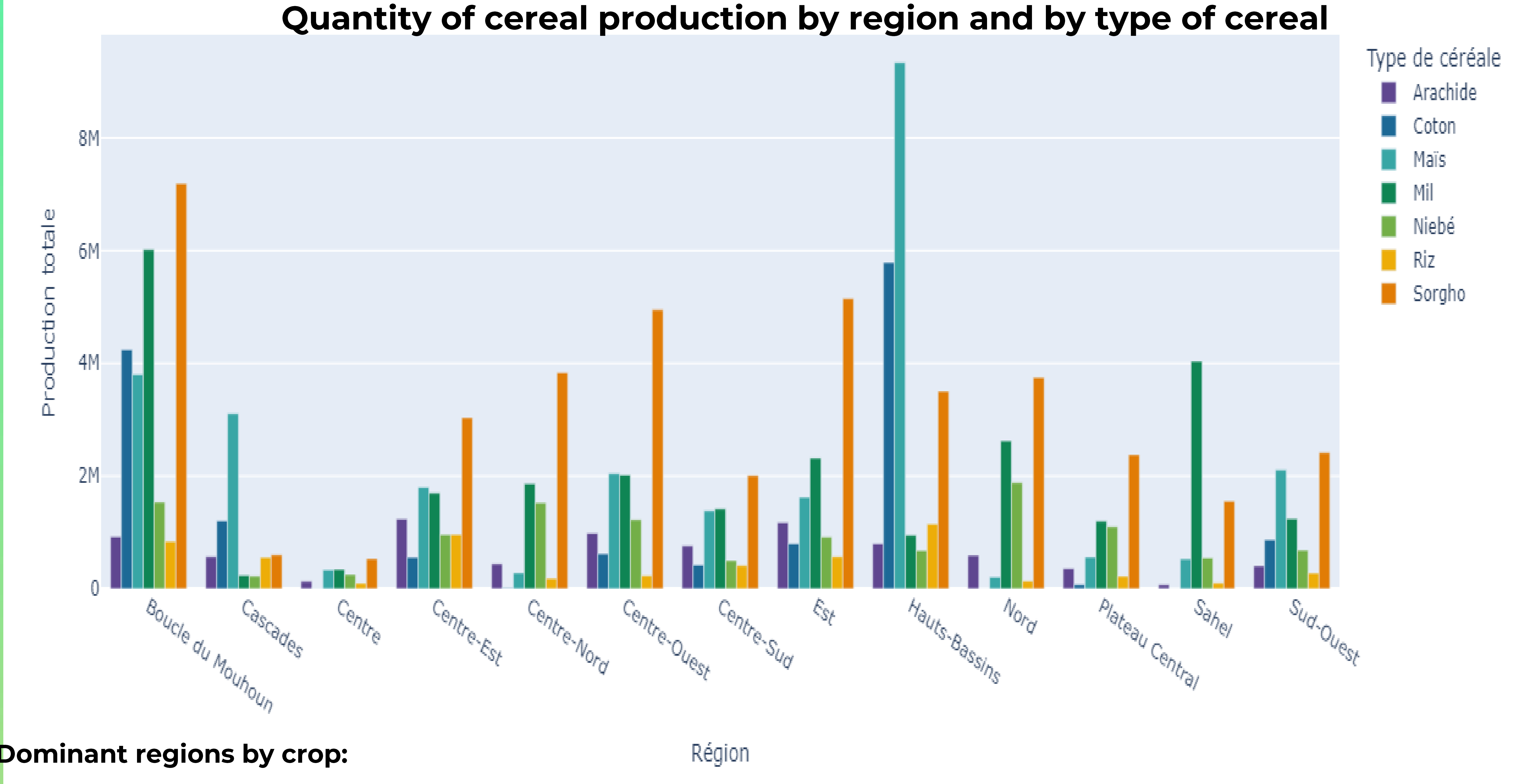
❖ **Boucle du Mouhoun is by far the most productive region, with production exceeding 24 million tonnes.**

❖ **Followed by Hauts-Bassins with approximately 22 million tonnes.**

❖ **East, Center-West, and Center-East are in the middle range (between 10 and 13 million tonnes).**

❖ **Center, Central Plateau, Cascades, Sahel, and Center-South have the lowest production volumes, particularly Center, which is significantly lower with less than 3 million tonnes.**

# EXPLORATORY DATA ANALYSIS



Quantity of cereal production by region and by type of cereal

**Dominant regions by crop:**

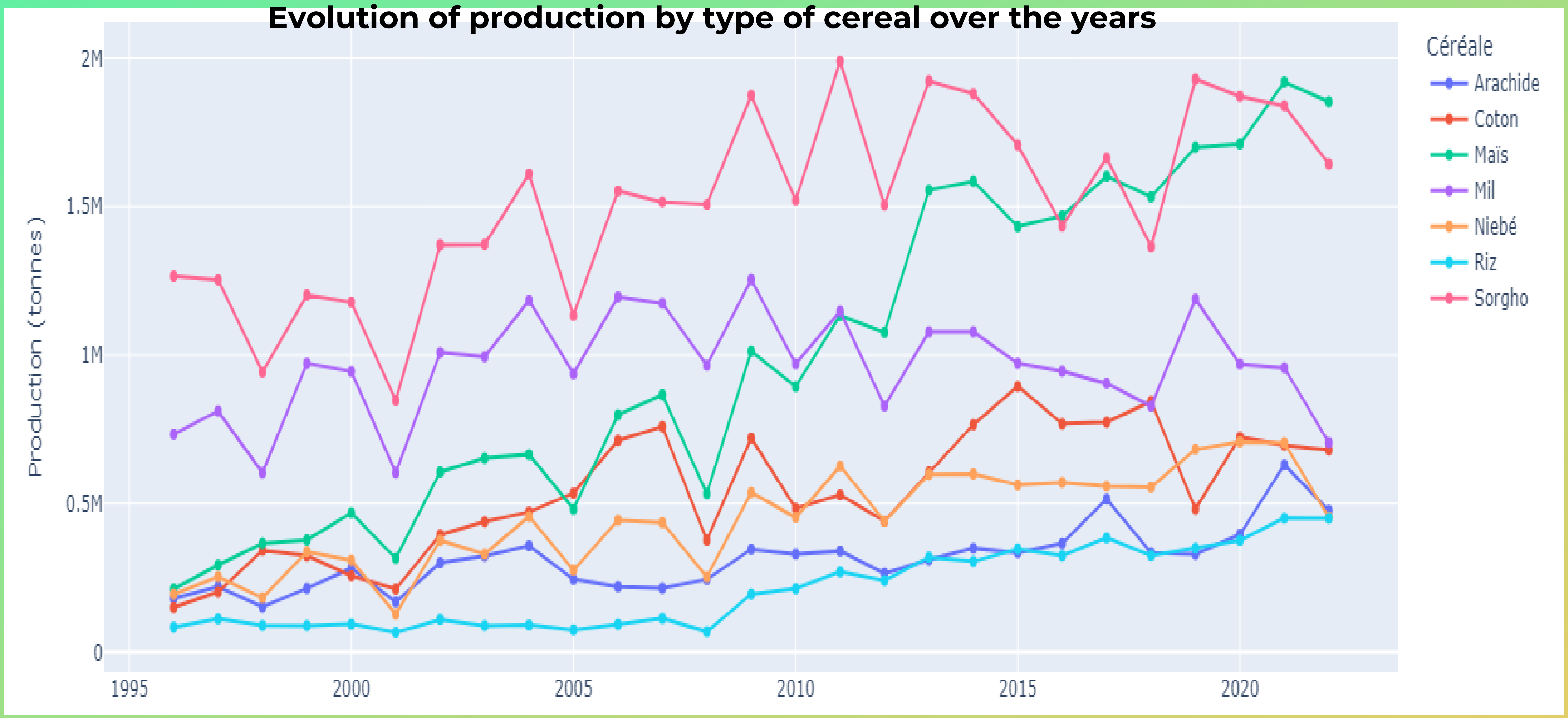
❖ Hauts-Bassins: Highly diversified, with peaks in corn, cotton, rice, and sorghum.

❖ Boucle du Mouhoun is productive in almost all cereals. East and North-Central are dominated by dry cereals (sorghum, millet).

❖ Sahel: Production oriented toward millet and sorghum.

❖ **General trends: Sorghum and corn are the most widely cultivated cereals.**

# EXPLORATORY DATA ANALYSIS



**Sorghum**: This is the most produced cereal throughout the period. Its production remained generally stable with moderate fluctuations.

**Corn**: Continued strong growth, especially from 2010 onwards, surpassing millet and approaching sorghum by the end of the period.

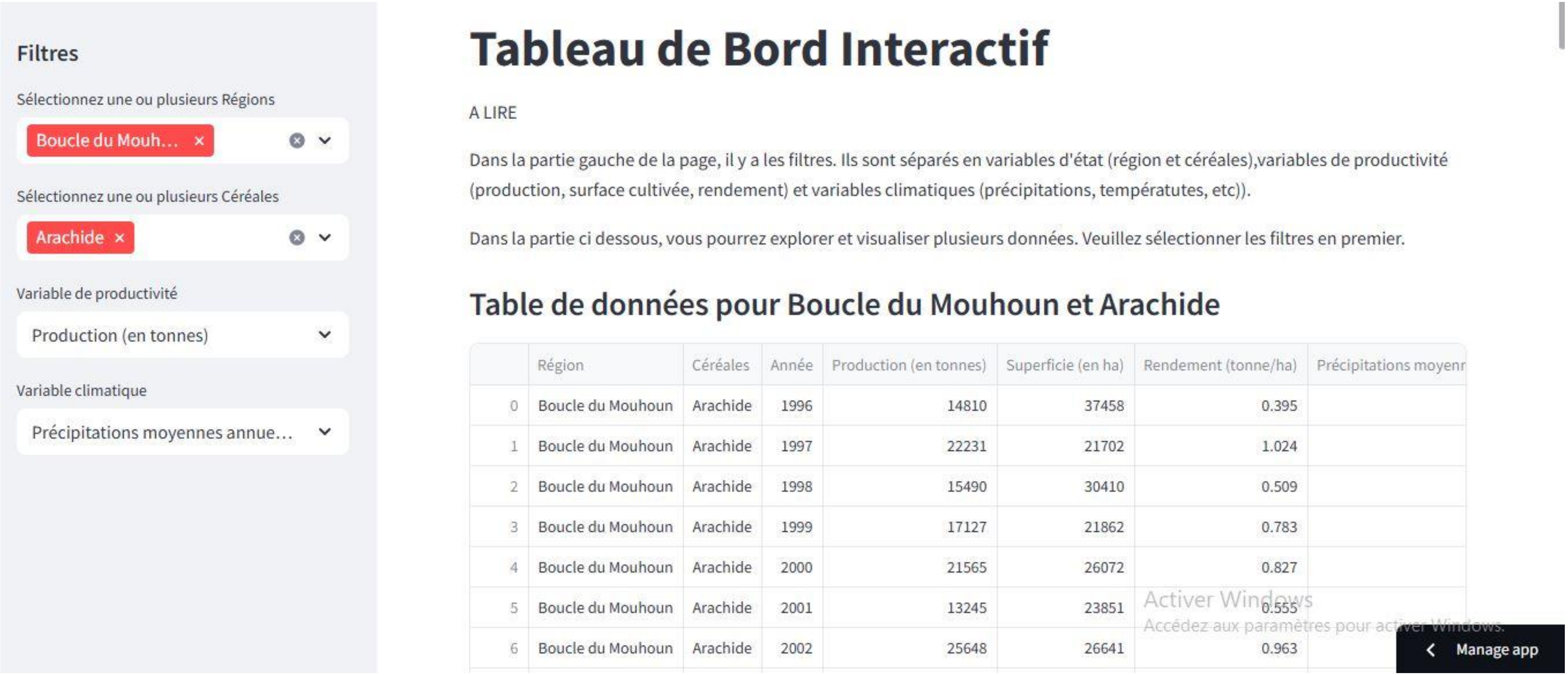
**Millet**: Relatively stable production, with some variations but no marked upward or downward trend.

**Peanuts, cotton, cowpeas, and rice** show similar trends. An upward trend after 2010, but they remain among the least produced.

# INTERACTIVE VISUALIZATION

We create an interactive visualization app that can allow to any user to visualize an explore some of the graphic analysis we’ve done above.

***Here the link of the app (click and follow the instructions): https://cerealbf-forecast-visualization.streamlit.app/***



Filtres

Sélectionnez une ou plusieurs Régions

Boucle du Mouh... ×

Sélectionnez une ou plusieurs Céréales

Arachide ×

Variable de productivité

Production (en tonnes)

Variable climatique

Précipitations moyennes annue...

# Tableau de Bord Interactif

A LIRE

Dans la partie gauche de la page, il y a les filtres. Ils sont séparés en variables d'état (région et céréales),variables de productivité (production, surface cultivée, rendement) et variables climatiques (précipitations, températutes, etc)).

Dans la partie ci dessous, vous pourrez explorer et visualiser plusieurs données. Veuillez sélectionner les filtres en premier.

## Table de données pour Boucle du Mouhoun et Arachide

| | Région | Céréales | Année | Production (en tonnes) | Superficie (en ha) | Rendement (tonne/ha) | Précipitations moyenr |
|---|---|---|---|---|---|---|---|
| 0 | Boucle du Mouhoun | Arachide | 1996 | 14810 | 37458 | 0.395 | |
| 1 | Boucle du Mouhoun | Arachide | 1997 | 22231 | 21702 | 1.024 | |
| 2 | Boucle du Mouhoun | Arachide | 1998 | 15490 | 30410 | 0.509 | |
| 3 | Boucle du Mouhoun | Arachide | 1999 | 17127 | 21862 | 0.783 | |
| 4 | Boucle du Mouhoun | Arachide | 2000 | 21565 | 26072 | 0.827 | |
| 5 | Boucle du Mouhoun | Arachide | 2001 | 13245 | 23851 | 0.555 | |
| 6 | Boucle du Mouhoun | Arachide | 2002 | 25648 | 26641 | 0.963 | |

# PREDICTION MODELING

Before applying Random Forest and Xgboost algorithms, here are the keys steps for modeling:

- **Separation of data into factors (X) and target (y)**

The first step will consist of structuring the data into two sets:
X, containing the explanatory (or dependent) variables, such as humidity, temperature, precipitations, area, etc
y, representing the target variable, namely cereal production .

- **Avoiding data leakage**

We need remove the yield variable to avoid data leakage. Indeed, calculating yield requires production and area values, so having yield in the model's dependent variables induces information leakage.

- **Model evaluation**

To evaluate and choose the best model we will use evaluation metrics the coefficient of determination ($R^2$) and Root Mean Squared Error (RMSE) .
The best model will be the use with the lowest RMSE and the highest $R^2$.
Reminder: A model is acceptable if $R^2$>= 85%.

## Model selection

**Random forest is the best model for our data, it gives us 92% of precision.**