Content

Introduction

- Problem statment
- Data
- Workfllow Describtion

Analysis

- EDA
- Infrence

insghts

Digtal Crope Estimation Problem

Introduction

Problem statment

Digtal crope yield estmation in India

- Smallholder farmers are crucial contributors to global food production, and in India often suffer most from poverty and malnutrition. These farmers face challenges such as limited access to modern agriculture, unpredictable weather, and resource constraints. To tackle this issue, Digital Green collected data via surveys, offering insights into farming practices, environmental conditions, and crop yields.
- **The objective of this challenge is to create a machine learning solution to predict the crop yield per acre of rice or wheat crops in India. Our goal is to empower these farmers and break the cycle of poverty and malnutrition.**

Data

- *The data was collected through a survey conducted across multiple districts in India. It consists of a variety of factors that could potentially impact the yield of rice crops.*
- The data source is Digital Green, Which hosted the data as a competition in zindi platform.
- Data structure '.csv' files
- Data Shape: `(3652, 50)`

- [ ] Variables defintion and dtype:
- [ ] Varibles type:
- [ ] DataFrame:

Workfllow Describtion

The data contain different types of features :

Numebr of datetime features : 5

Numebr of catogrical features : 24

Numebr of numerical features : 21

Hint: The catogrical data dosen't contain features of ordinal type.

📍 Note:

- **All what we see here is coming as a result of data analysis, dealing with missing data and dealing with outliers.**
- You can find all details in this Notebook.

1- Missing value :

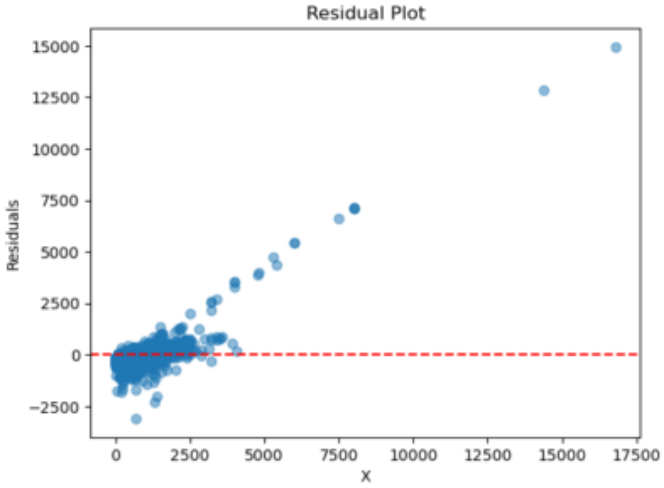
I used the following approach:

- **Drop any column with more than '40.0%' of missing values.**
- **Imput the catogrical columns with the mode.**
- **Impute the numerical columns with 'KNNImputer'.**

2- Outliers :

I use the following approach:

- **Detect some extrem data points and through them out**
- **Use the IQR to handel the outliers**
- **Outliers in the target variable 'Yield' with residual plot.**



Residual Plot

After Data cleainig we end up with

Numebr of datetime features : 5

Numebr of catogrical features : 17

Numebr of numerical features : 16

Analysis

EDA

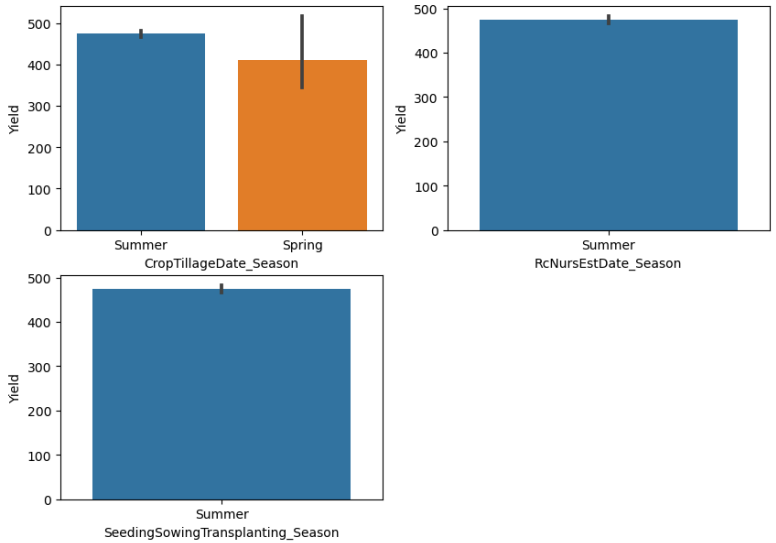
- Date time features :

Dsecribtion	Defintions

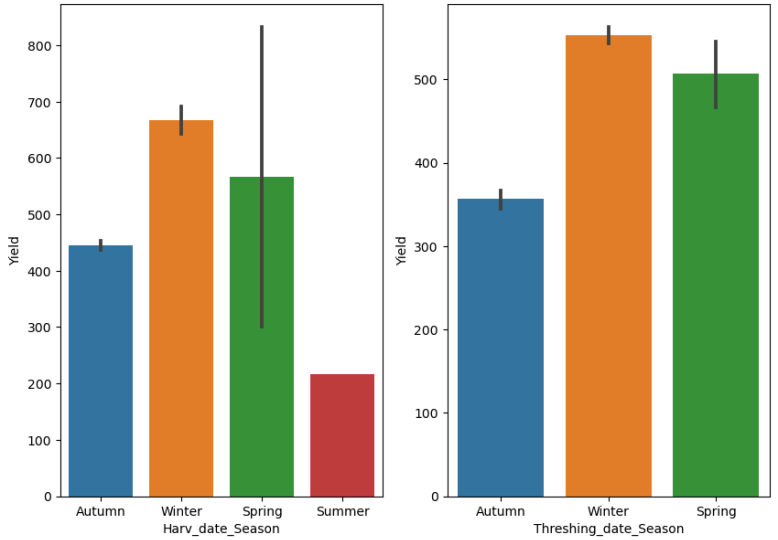
I devide the date time features to seasons, to see rice cultivation cycle.

Rice Cultivation seasons

Rice Harvrest seasons



Cultivation seasons



Harvrest seasons

- Catogrical features :

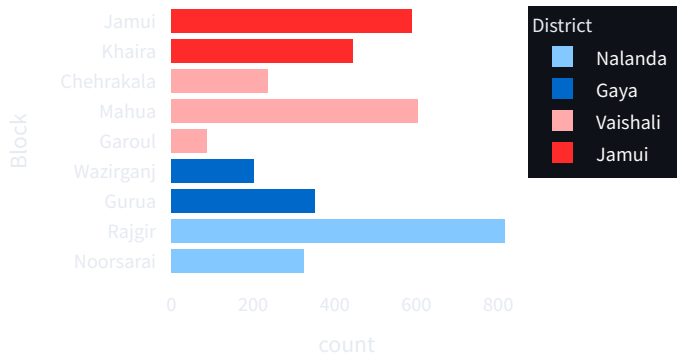
District and Block

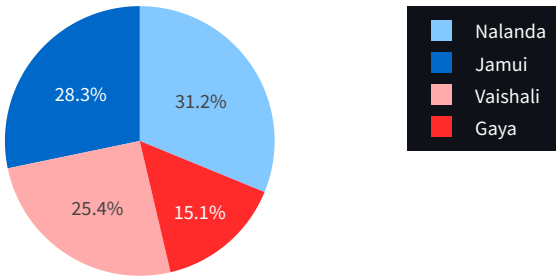
Dsecribtion

Defintions

Pie chart

Count plot

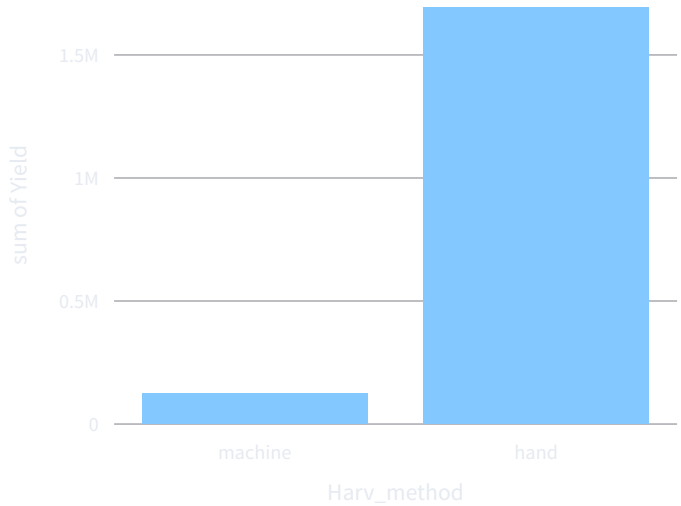


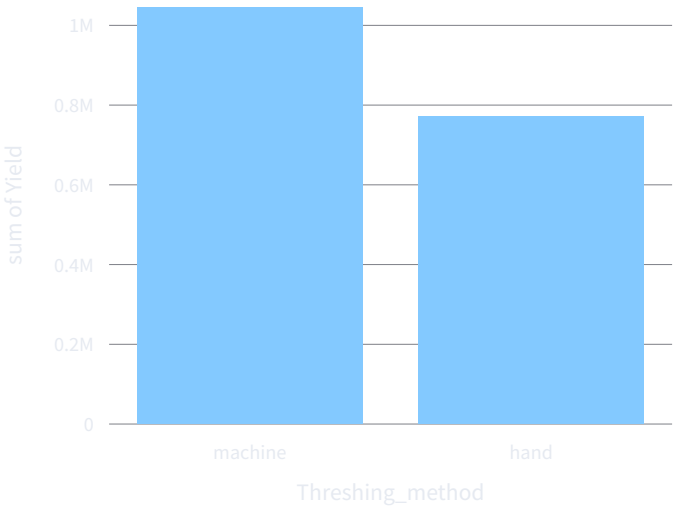


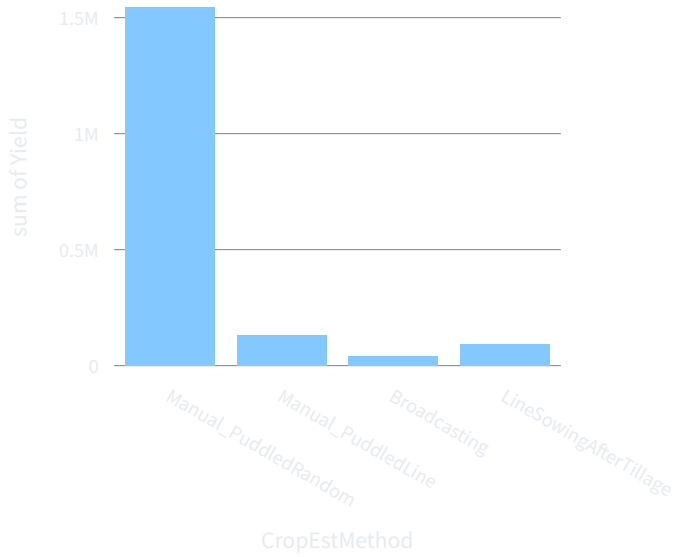
Transplantation, harvesting and threshing methods.

Dsecribtion

Defintions



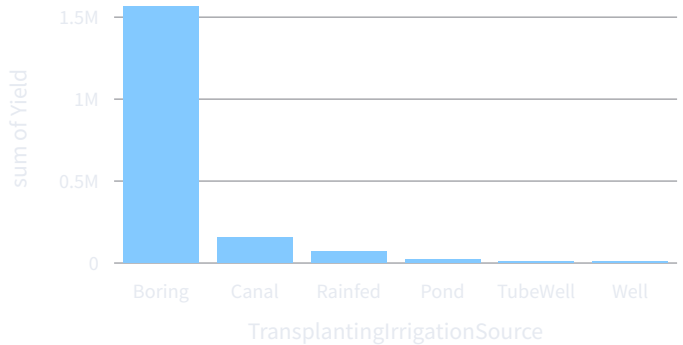


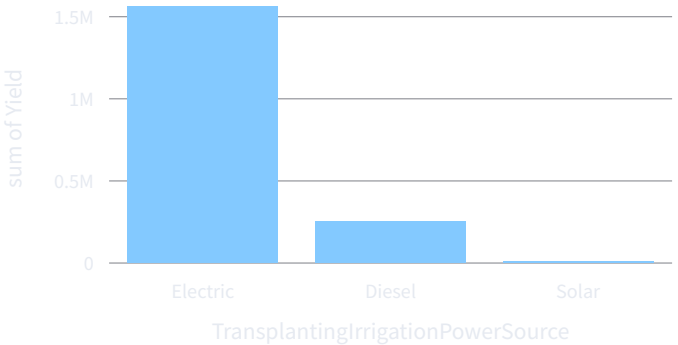


Source of water and Source of power

Dsecribtion

Defintions

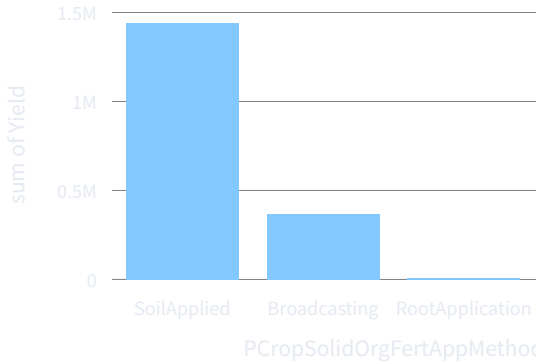


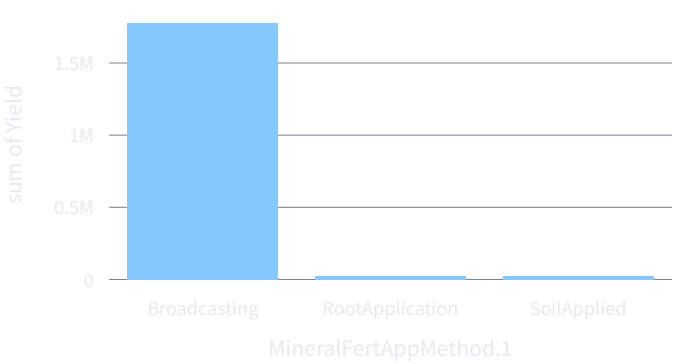


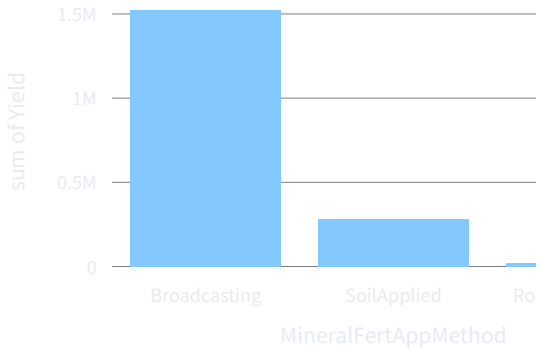
Methods of fertilization

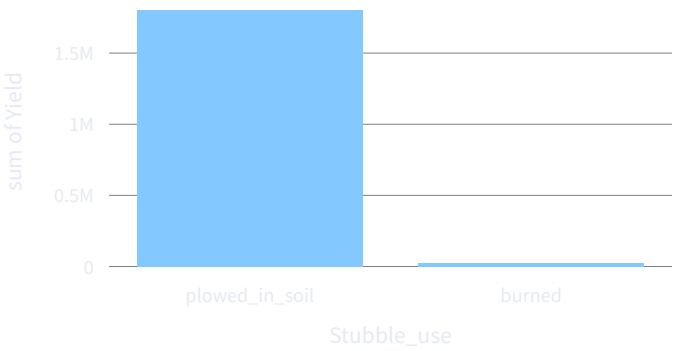
Dsecribtion

Defintions









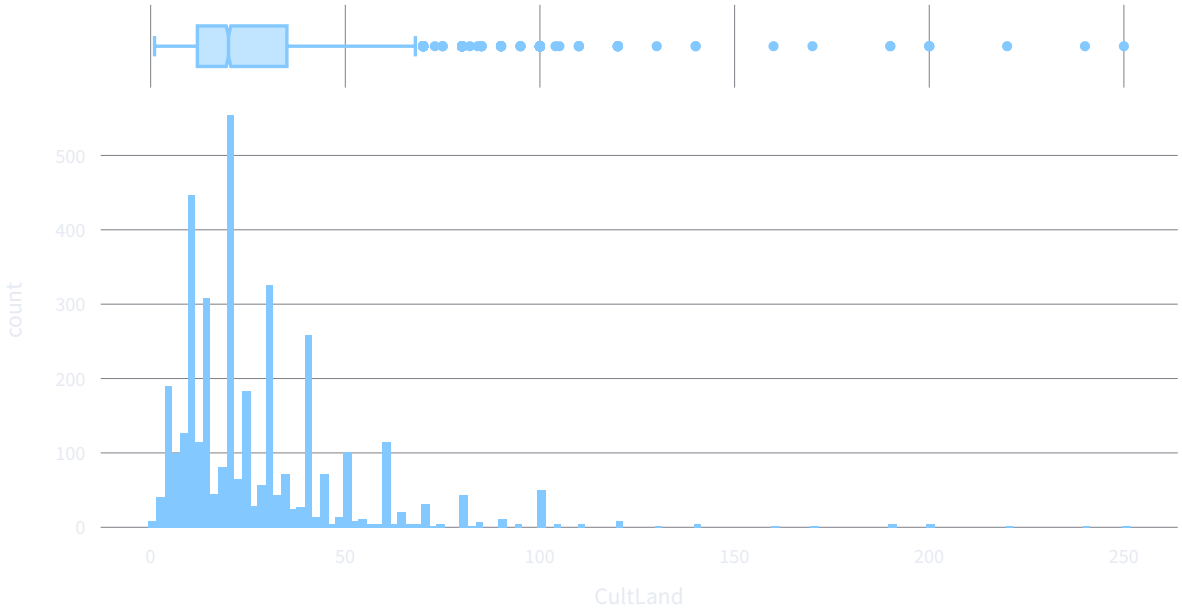
📍 **Note: We dealing with imbalance data.**

- Numerical features :

choose a variable to see it's distribution

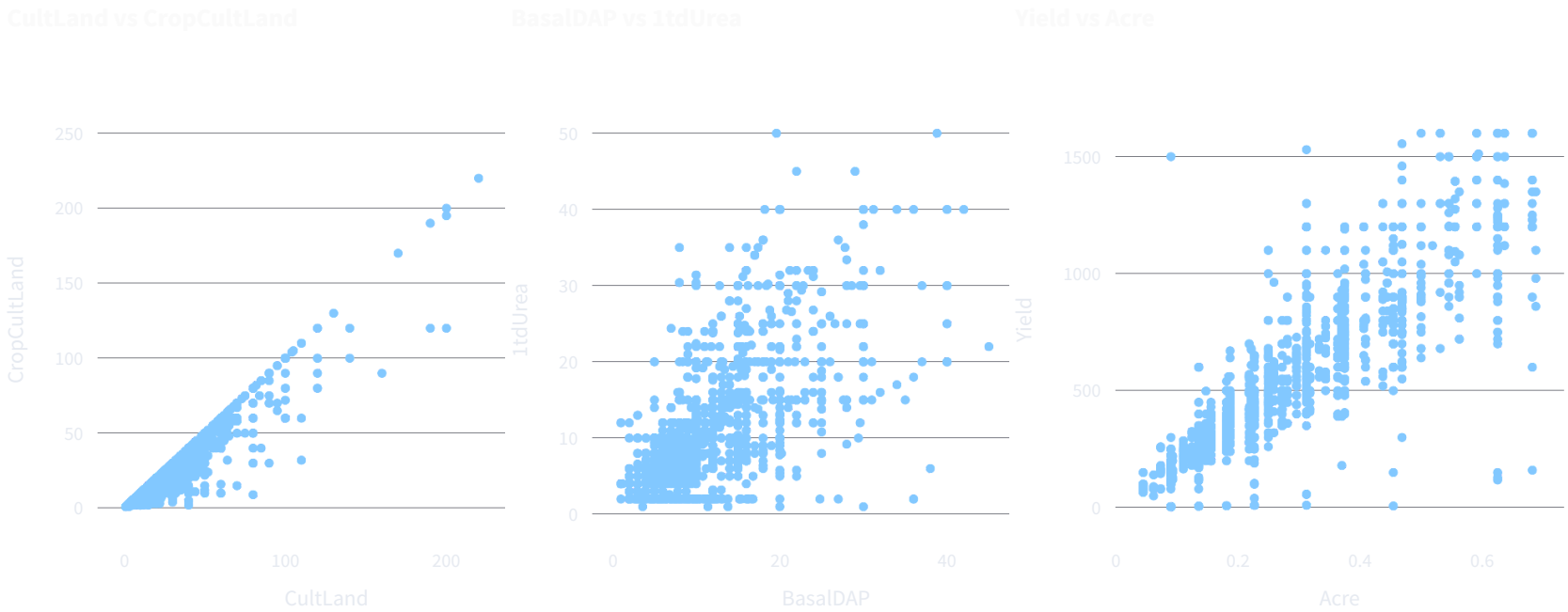
CultLand

	CultLand
Description	Area of total cultivated land



- *Bi variante* :

Correlated features



Infrence

- In this section we gonna use some statistics, hypothesis testing
- We will us noneparametric tests , due the data is not normaly distribuit

1- Mann_wetny U tset

Harv_method ('hand' , 'machine')

- Group1 n= `3642`
- Group2 n= `228`

Threshing_method ('hand' , 'machine')

- Group1 n= `1772`
- Group2 n= `2098`

Threshing_method ('hand' , 'machine')

- Group1 n= `3846`
- Group2 n= `24`

H0: Group1[Yield]median = Group2[Yeild]median

H1: Group1[Yield]median < Group2[Yeild]median

	MWU
U-val	286969.0
alternative	less
p-val	2.27041745003878e-15
RBC	0.3088215459020973
CLES	0.6544107729510487

- Descision : sice P-value < 0.05 level of significant , we rejct H0 and conclud that the median of yield given harv_method == 'hand' is less than the median of yield given harv_method == 'machine'

H0: Group1[Yield]median = Group2[Yeild]median

H1: Group1[Yield]median < Group2[Yeild]median

	MWU
U-val	1631694.5
alternative	less
p-val	2.6390400091343988e-11
RBC	0.1221917789058482
CLES	0.5610958894529241

- Descision : sice P-value < 0.05 level of significant , we rejct H0 and conclud that the median of yield given Threshing_method == 'hand' is less than the median of yield given Threshing_method == 'machine'

H0: Group1[Yield]median = Group2[Yeild]median

H1: Group1[Yield]median ≠ Group2[Yeild]median

	MWU
U-val	24109.5
alternative	two-sided
p-val	5.3012401101648294e-05
RBC	0.47760660426417056
CLES	0.2611966978679147

- Descision : sice P-value < 0.05 level of significant , we rejct H0 and conclud that the median of yield given Threshing_method == 'plowed_in_soil' not equal the median of yield given Threshing_method == 'burned'