Data Analysis on crop production of India

Abstract:

Data is generally available in the raw form it can be unstructured and unorganised. Real world data can be Inconsistent, noisy, Incomplete because of the various circumstances while data preparation. data is useful if and only if it is transformed into Understandable, and efficient format. Data pre-processing is a technique which help us in transforming the data into clearer and understandable format. Data pre-processing is preliminary step, ignoring this can induce very high bias and variance to a machine learning or deep learning model.

Data pre-processing consists of various steps, which include data cleaning, data transformation, data discretization, and many more. The whole aim of these steps is to transform the given data into a form which meets the requirements of machine learning or deep learning algorithms. Basically, the output data after this process is input for machine learning, so how well a machine learning or deep learning model performs is very much dependent on how well the data is organised and structured. This paper majorly focuses on analysing the effect of these data pre-processing techniques on Agriculture data set. We also try to draw few meaningful insights from the given data.

Introduction:

The dataset used in this paper contains information on crop covered area in Hectares and production in tonnes for 122 different crops in 33 states of India across 14 years from 2000 to 2013.We will begin with handling missing values, then encoding the columns with categorical data such as crop name , then discretising the columns which consists of continuous numerical data such as area and production, after that we will move onto outlier handling ,here we try to get rid of the outliers present the various columns and finally, we will be concluding by discussing the applicable feature selection and transformation techniques on the data set.

On the other hand, we also analysed the data set to draw few meaning insights such as affect of the National policy for farmers (2007) on the production of the crops in India. The National Policy for farmers (NPF-2007) introduced in 2007 by government of India focus on supplying good quality seeds, disease-free planting material, issuing soil health passbooks to the farmers. We compared and analysed the area of cultivation and production of crop before and after introducing NPF-2007. We also compared and analysed the national statistics with State statistics of Andhra Pradesh before and after introducing NPF-2007 which are clearly mentioned in results section.

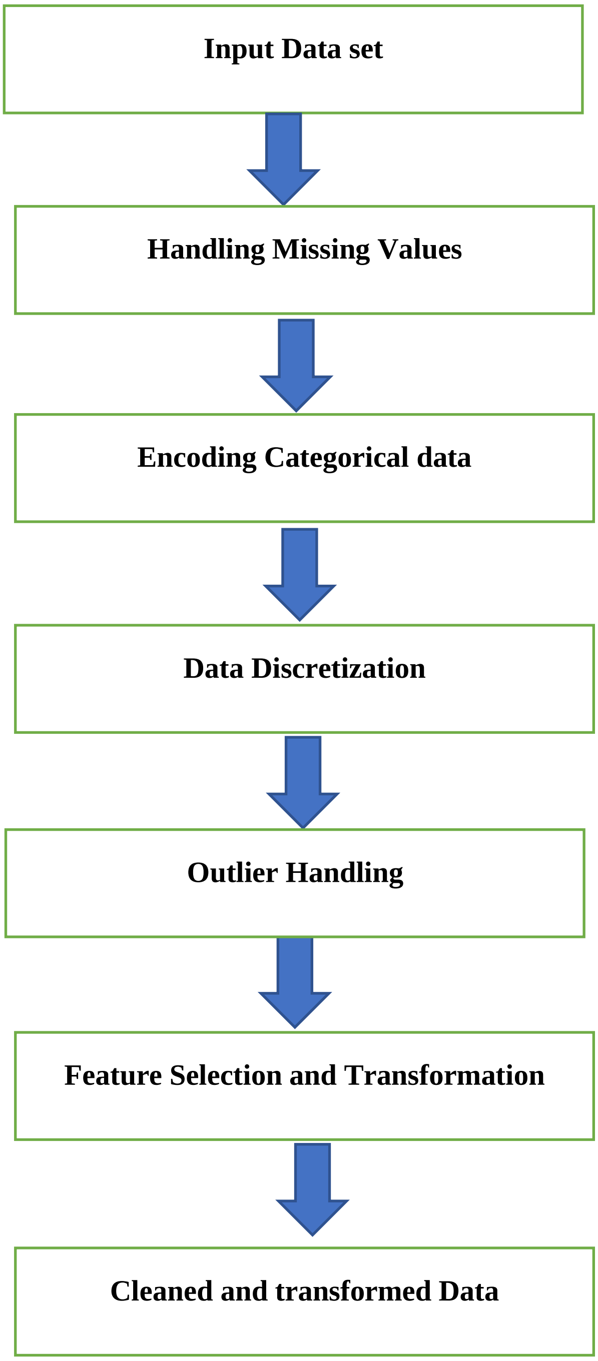
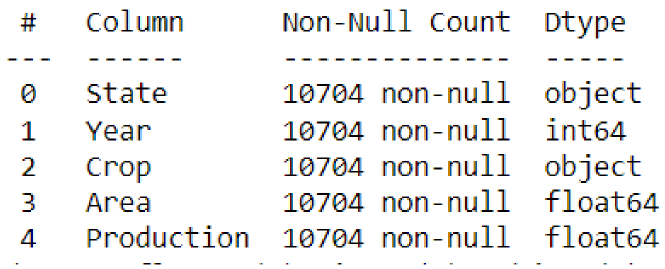

Input Data set
Handling Missing Values
Encoding Categorical data
Data Discretization
Outlier Handling
Feature Selection and Transformation
Cleaned and transformed Data

FIG (1): *Representing various steps in data pre-processing*

Data Pre-Processing Techniques	Description	Types
Complete Case Analysis (CCA)	• In this method, remove all the rows or records where any column or field contains a missing value. • Huge loss of data, advisable only on large data sets.	• List wise deletion • Column wise deletion
Handling missing numerical data	• Replaces all the null values with the mean or mode of the respective group • In arbitrary value imputation, a value is chosen arbitrarily to replace all the missing values. • For end of distribution imputation, chosen value from the end of the data accounts for the actual data which was missing	• Mean Imputation • Median Imputation • Mode Imputation • Arbitrary value Imputation • End of distribution Imputation
Handling missing categorical data	• In frequent category imputation, missing values are filled by the most frequently repeating value. It is also called as mode imputation. • Missing category imputation is similar to arbitrary value imputation. In the case of categorical value, missing value imputation adds an arbitrary category.	• Frequent Category Imputation • Missing Category Imputation
Encoding Categorical data	• The techniques that are used to convert numeric data into categorical data are called categorical data encoding schemes.	• One hot encoding • Label encoding • Frequency encoding • Ordinal encoding • Mean encoding
Discretization	• The process of converting continuous numeric values into discrete intervals is called discretization or binning. • It is very helpful to handle the outliers	• Equal Width Discretization • Equal Frequency Discretization • K-Means Discretization • Decision Tree Discretization • Custom Discretization
Outlier Handling	• An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.	• Outlier Trimming • Outlier Capping using IQR • Outlier Capping Using Mean and Std • Outlier Capping Using Quantiles • Outlier Capping using Custom Values
Feature Selection	• Feature selection is also known as Variable selection or Attribute selection. • Benefits of performing feature selection before modelling our data are, Reduces Overfitting Improves Accuracy, Reduces Training Time	• Filter Method • Wrapper Method
Transformation	• Standardization is the processing of centring the variable at zero and standardizing the • data variance to 1. • In min/max scaling, we subtract each value by the minimum value, and then divide the result by the difference of minimum and maximum value in the dataset.	• Z-Score Normalization • Min/Max Scaling

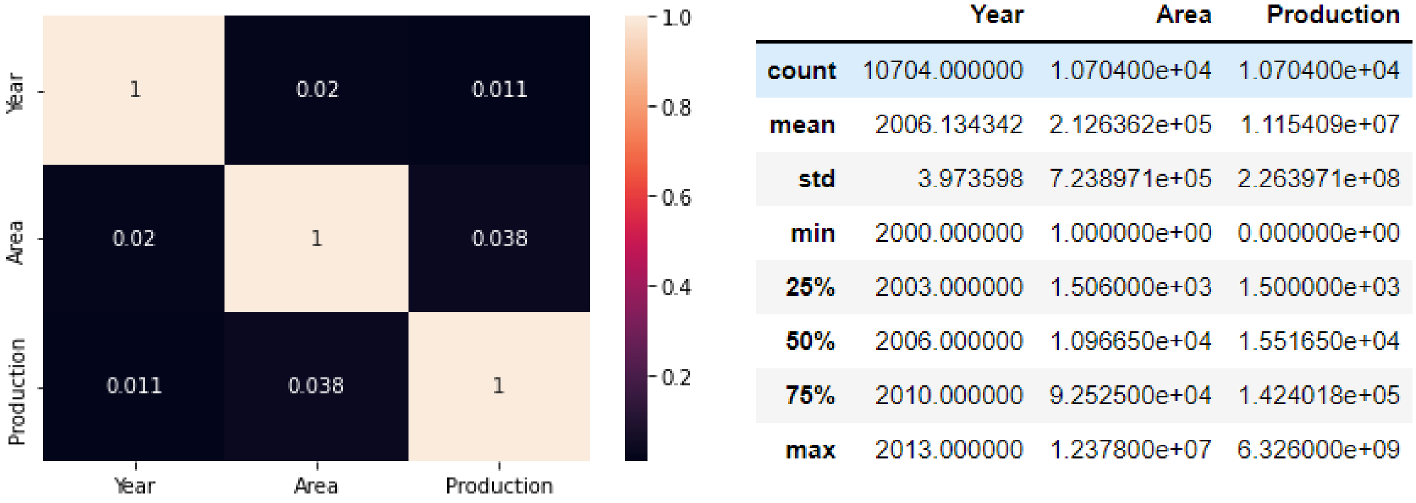
Experimental Results:

The chosen agriculture data set contains 10704 rows and 5 columns, each column representing Name of state, year, crop, area, production.



```
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   State       10704 non-null  object 
 1   Year        10704 non-null  int64  
 2   Crop        10704 non-null  object 
 3   Area        10704 non-null  float64
 4   Production  10704 non-null  float64
```

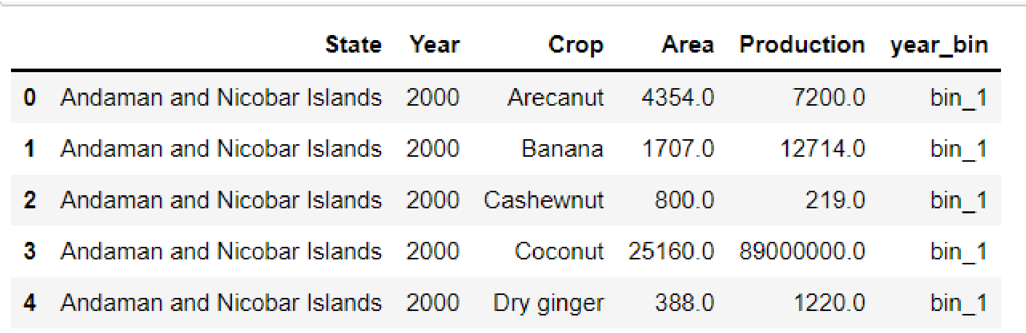
Fig (3): *Information of the data set.*



	Year	Area	Production
count	10704.000000	1.070400e+04	1.070400e+04
mean	2006.134342	2.126362e+05	1.115409e+07
std	3.973598	7.238971e+05	2.263971e+08
min	2000.000000	1.000000e+00	0.000000e+00
25%	2003.000000	1.506000e+03	1.500000e+03
50%	2006.000000	1.096650e+04	1.551650e+04
75%	2010.000000	9.252500e+04	1.424018e+05
max	2013.000000	1.237800e+07	6.326000e+09

Fig (4): *Representing Correlation and descriptive statistics.*

From fig (4) we can observe that no columns are heavily correlated. As stated above we will analyse based on year first i.e., before and after 2007. Now we will use binning technique and classify all the entries into 2 bins, namely before 2007 as bin 1 and after 2007 as bin 2.

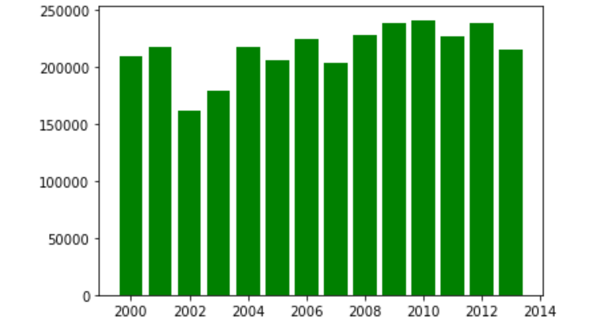
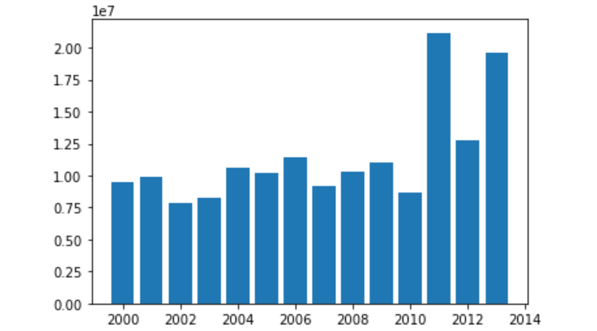
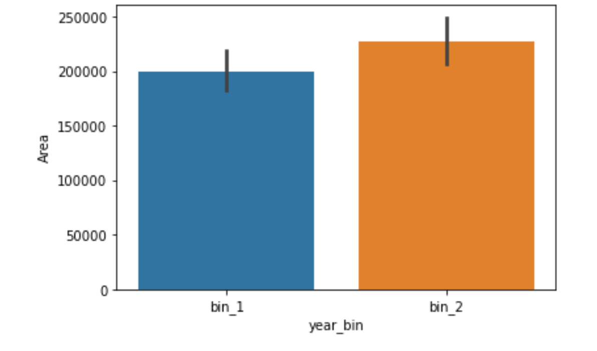
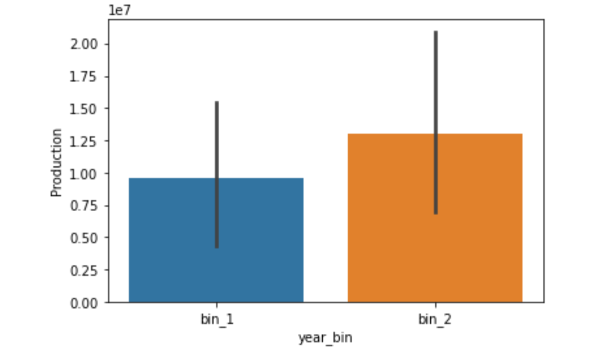
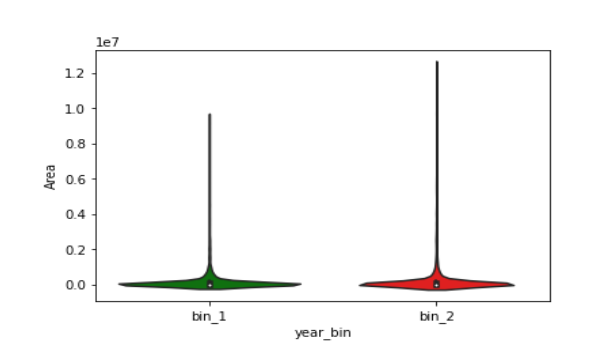
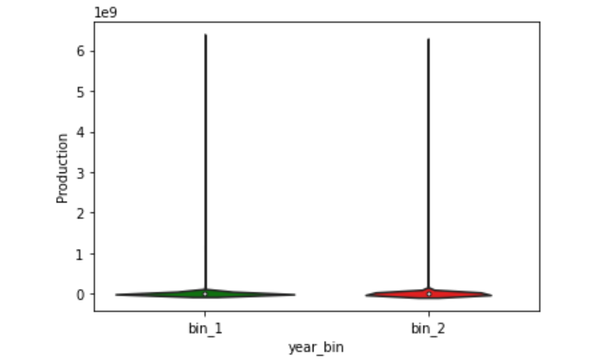


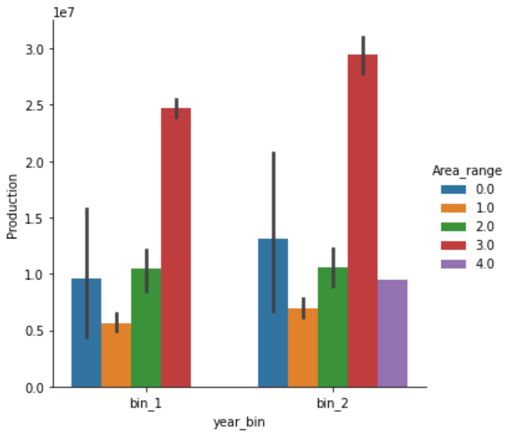
	State	Year	Crop	Area	Production	year_bin
0	Andaman and Nicobar Islands	2000	Arecanut	4354.0	7200.0	bin_1
1	Andaman and Nicobar Islands	2000	Banana	1707.0	12714.0	bin_1
2	Andaman and Nicobar Islands	2000	Cashewnut	800.0	219.0	bin_1
3	Andaman and Nicobar Islands	2000	Coconut	25160.0	89000000.0	bin_1
4	Andaman and Nicobar Islands	2000	Dry ginger	388.0	1220.0	bin_1

Fig (5): *After binning based on the year.*

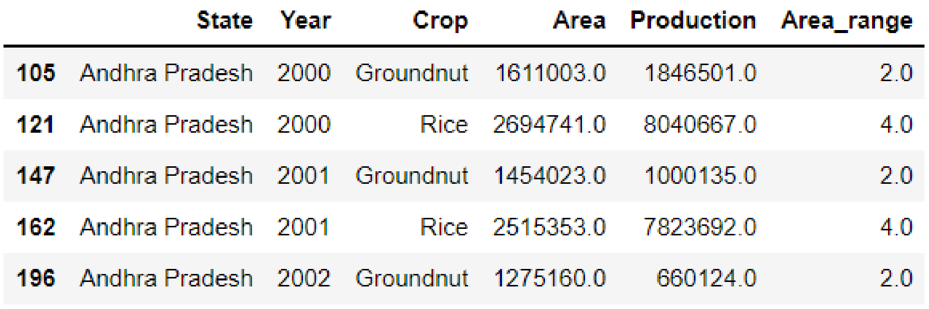
Average area of cultivation before 2007 (2000 to 2006) is 200200.53 hectares, but after 2007 (2007-2013) it is 227371.71 hectares. The average production before 2007 is 9551302.46 tonnes Which shoot up to 13053298.78 tonnes after 2007.

Visualisations:

Area of cultivation	Production
	
	
	



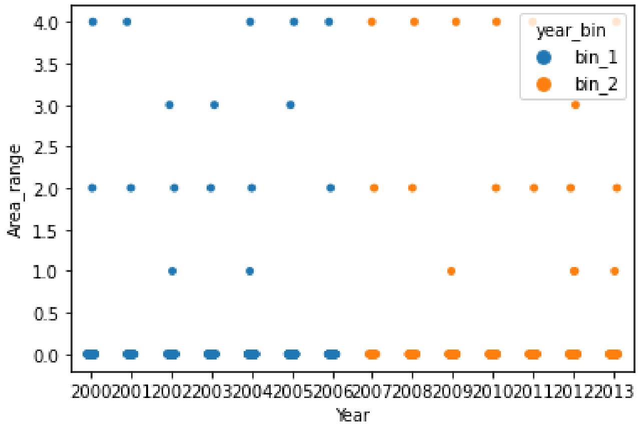
Andhra Pradesh State Analysis:



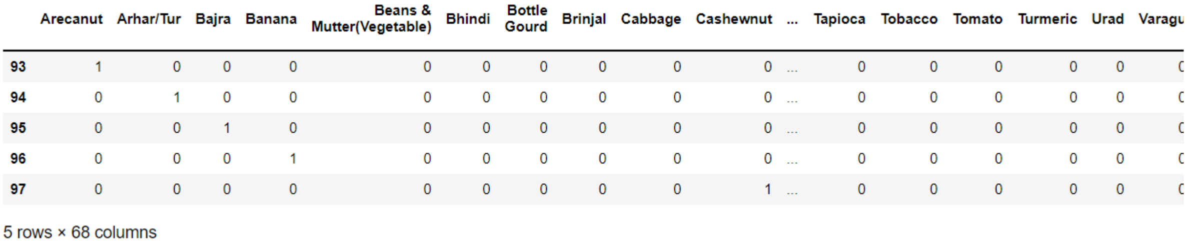
	State	Year	Crop	Area	Production	Area_range
105	Andhra Pradesh	2000	Groundnut	1611003.0	1846501.0	2.0
121	Andhra Pradesh	2000	Rice	2694741.0	8040667.0	4.0
147	Andhra Pradesh	2001	Groundnut	1454023.0	1000135.0	2.0
162	Andhra Pradesh	2001	Rice	2515353.0	7823692.0	4.0
196	Andhra Pradesh	2002	Groundnut	1275160.0	660124.0	2.0

Discretisation of Area column

Performed the data discretisation technique on the area column using KBinsDiscretizer from sklearn. preprocessing, the discretized area is grouped into 5 groups as shown in the Area_range column.



Visualisation after binning and discretisation



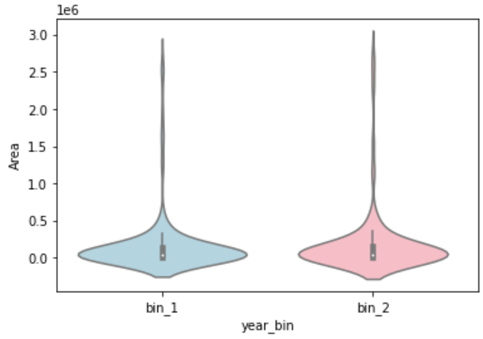
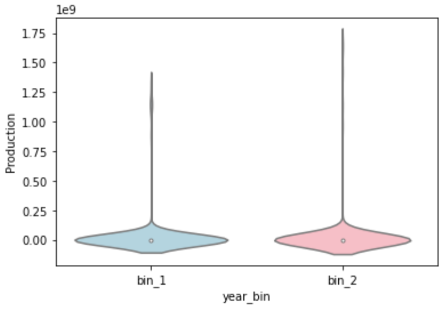
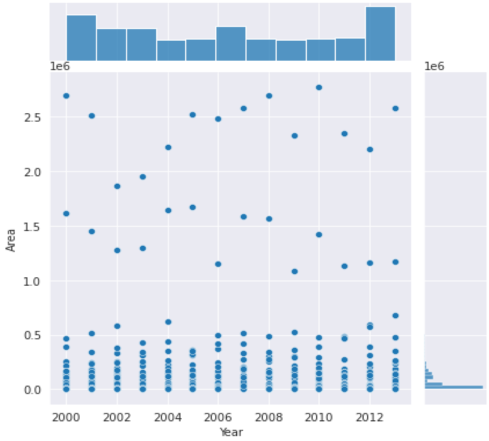
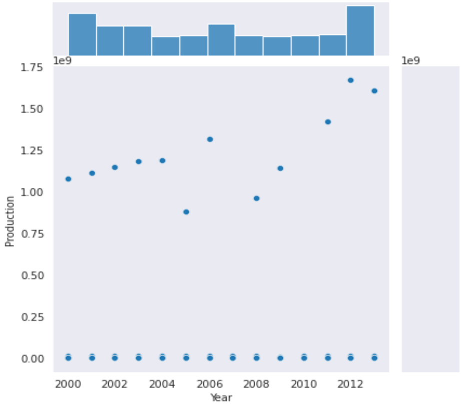
	Arecanut	Arhar/Tur	Bajra	Banana	Beans & Mutter(Vegetable)	Bhindi	Bottle Gourd	Brinjal	Cabbage	Cashewnut	...	Tapioca	Tobacco	Tomato	Turmeric	Urad	Varagu
93	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
94	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
95	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0
96	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0
97	0	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0

5 rows × 68 columns

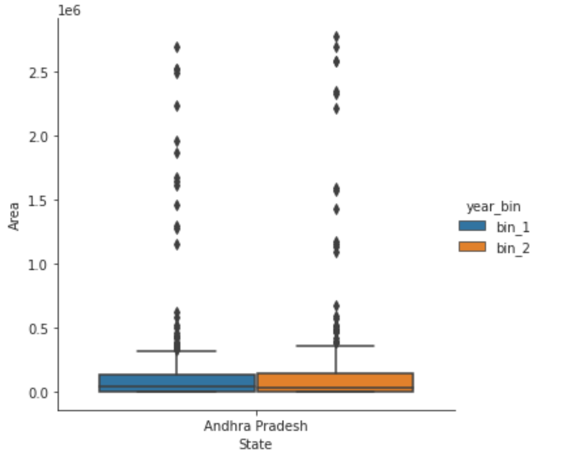
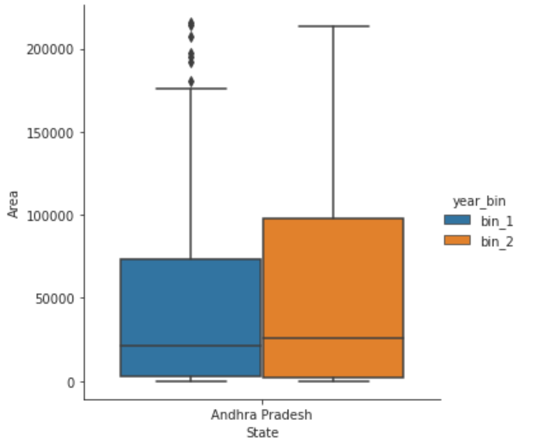
One hot encoding of crop column

As our major point of analysis is not based on the crop, we don’t worry much about the categorical encoding part, but if we want, we can try label encoding or one -hot encoding.

Visualizations:

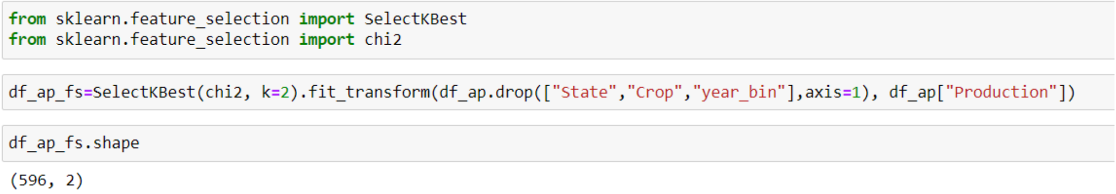
Area of cultivation	Production
	
	

Outlier Analysis:

Original Data	After outlier handling
	

Feature selection:

We used the univariate feature selection technique which select the best features based on the univariate statistical tests. We made use of the chi-square test.



```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
```

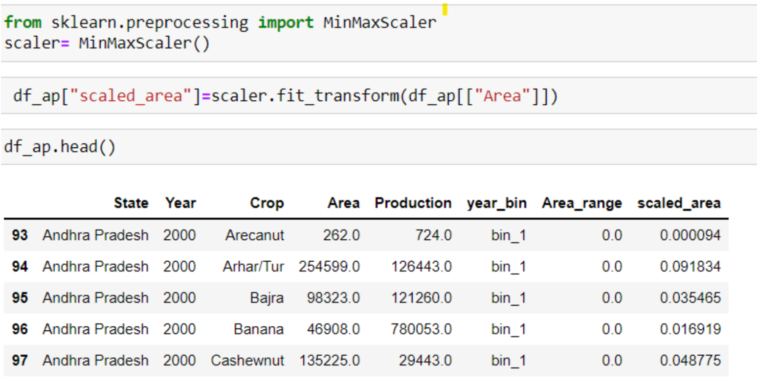
```
df_ap_fs=SelectKBest(chi2, k=2).fit_transform(df_ap.drop(["State","Crop","year_bin"],axis=1), df_ap["Production"])
```

```
df_ap_fs.shape
```

```
(596, 2)
```

Transformation:

We apply Min max transformation on area column as our main point of analysis is area of cultivation, if the whole data of this column is adjusted such that the range is in between 0 and 1 it is better to analyse

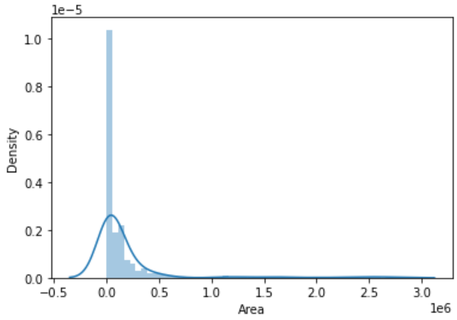
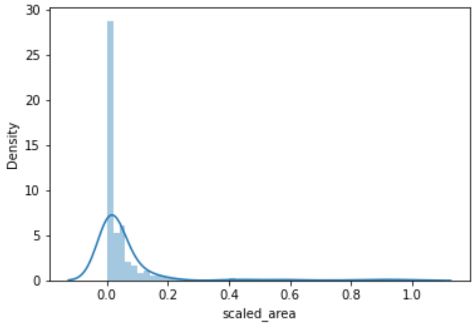


```
from sklearn.preprocessing import MinMaxScaler
scaler= MinMaxScaler()
```

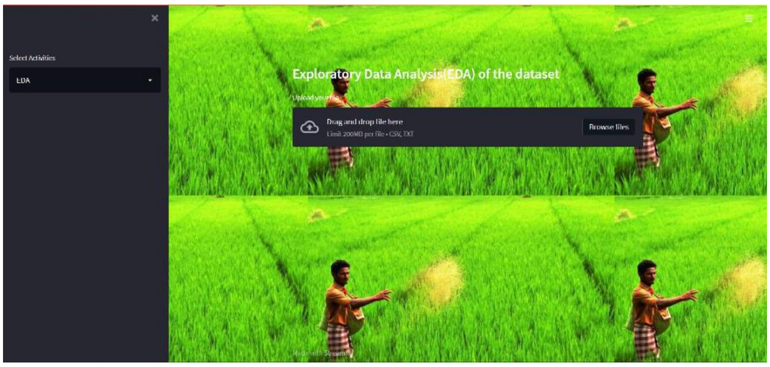
```
df_ap["scaled_area"]=scaler.fit_transform(df_ap[["Area"]])
```

```
df_ap.head()
```

	State	Year	Crop	Area	Production	year_bin	Area_range	scaled_area
93	Andhra Pradesh	2000	Arecanut	262.0	724.0	bin_1	0.0	0.000094
94	Andhra Pradesh	2000	Arhar/Tur	254599.0	126443.0	bin_1	0.0	0.091834
95	Andhra Pradesh	2000	Bajra	98323.0	121260.0	bin_1	0.0	0.035465
96	Andhra Pradesh	2000	Banana	46908.0	780053.0	bin_1	0.0	0.016919
97	Andhra Pradesh	2000	Cashewnut	135225.0	29443.0	bin_1	0.0	0.048775

Before Transformation	After Transformation
	

Graphical User Interface (GUI):



Select Activities

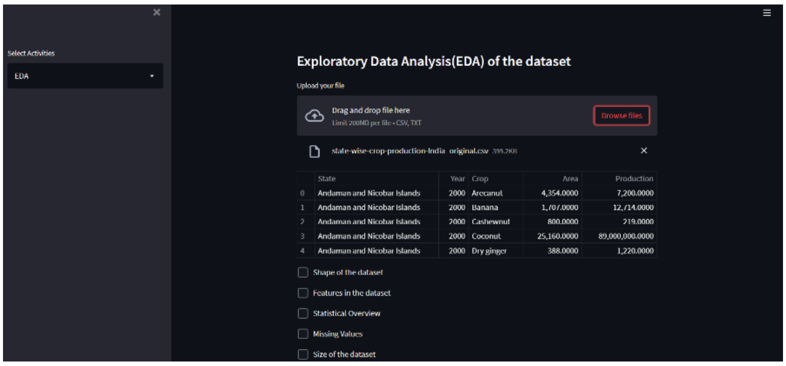
EDA

Exploratory Data Analysis(EDA) of the dataset

Upload your file

Drag and drop file here

Limit 200MB per file • CSV, TXT



Select Activities

EDA

Exploratory Data Analysis(EDA) of the dataset

Upload your file

Drag and drop file here

Limit 200MB per file • CSV, TXT

state-wise-crop-production-India_original.csv 399.2KB

	State	Year	Crop	Area	Production
0	Andaman and Nicobar Islands	2000	Arecanut	4,354.0000	7,200.0000
1	Andaman and Nicobar Islands	2000	Banana	1,707.0000	12,714.0000
2	Andaman and Nicobar Islands	2000	Cashewnut	800.0000	219.0000
3	Andaman and Nicobar Islands	2000	Coconut	25,160.0000	89,000,000.0000
4	Andaman and Nicobar Islands	2000	Dry ginger	388.0000	1,220.0000

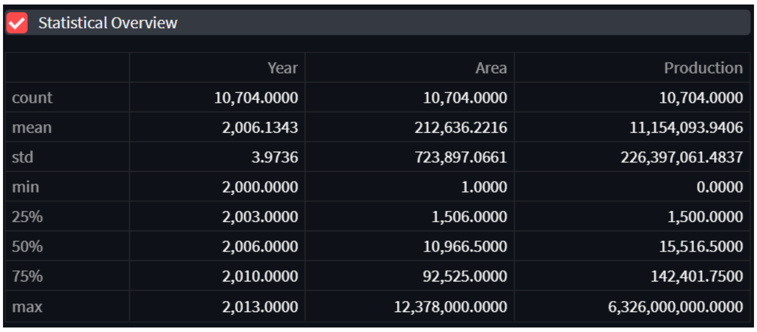
□ Shape of the dataset

□ Features in the dataset

□ Statistical Overview

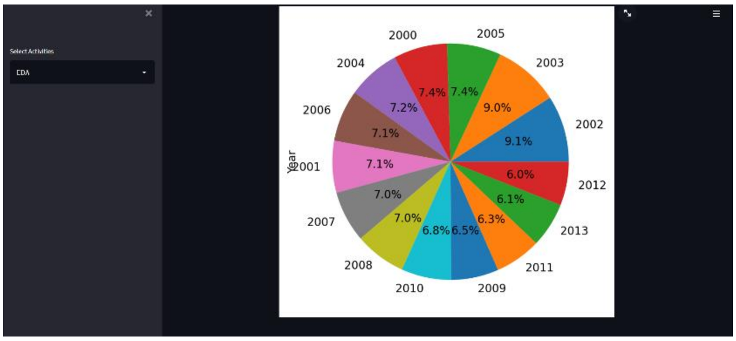
□ Missing Values

□ Size of the dataset



☑ Statistical Overview

	Year	Area	Production
count	10,704.0000	10,704.0000	10,704.0000
mean	2,006.1343	212,636.2216	11,154,093.9406
std	3.9736	723,897.0661	226,397,061.4837
min	2,000.0000	1.0000	0.0000
25%	2,003.0000	1,506.0000	1,500.0000
50%	2,006.0000	10,966.5000	15,516.5000
75%	2,010.0000	92,525.0000	142,401.7500
max	2,013.0000	12,378,000.0000	6,326,000,000.0000



Select Activities

EDA

Conclusion:

This paper Summarizes the effects of various data preprocessing techniques on the data set, this paper also visualises and compare the distribution of the data before and after every technique applied. It is very evident from the above results that, data preprocessing is a very important first step for anyone dealing with data sets. It leads to better data sets, that are clearer, more organised, more structured, and more manageable. In this paper we also tried to draw few insights regarding how NPF-2007 affected the agriculture sector of India, although the observations can be improved by improving the data cleaning part. Designing a model to based on various machine learning and deep learning algorithms to predict the production each year can be possible enhancements.