Crop Production Analysis in India

Project Report By: Aadithya Ram

Project Details

Project Title	Crop Production Analysis in India
Technologies	Data Science
Domain	Agriculture
Project Difficulties level	Advanced

Objectives and Problem Statement

Objectives

01

ETL & EDA

Extract – Transform - Load

02

Key Indicators of Production

Crop Production Factors & Analysis

03

Crop Production Predictions

Regression Models, Plots

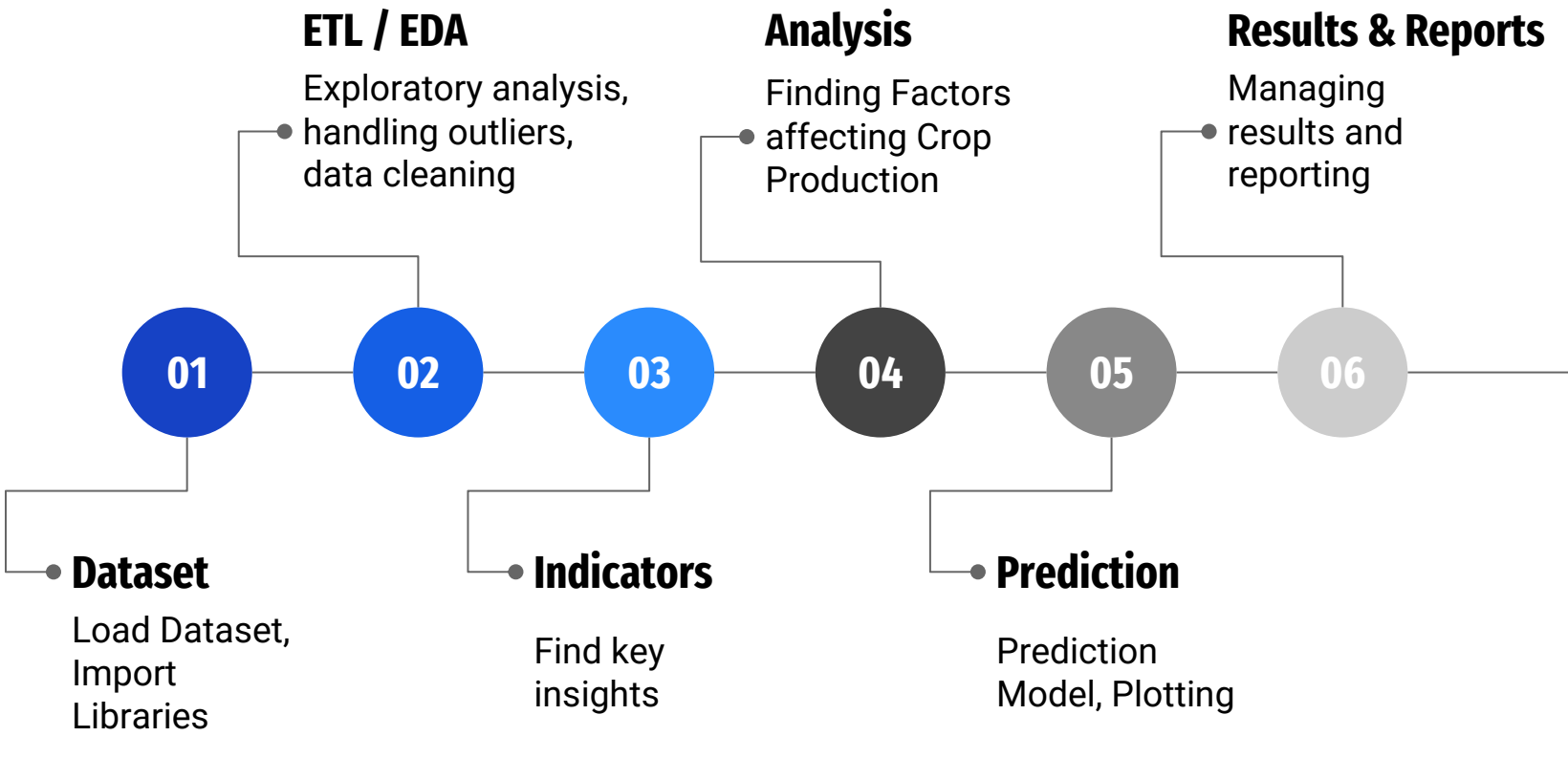
Problem Statement

The Agriculture business domain, as a vital part of the overall supply chain, is expected to highly evolve in the upcoming years via the developments, which are taking place on the side of the Future Internet. This paper presents a novel Business-to-Business collaboration platform from the agri-food sector perspective, which aims to facilitate the collaboration of numerous stakeholders belonging to associated business domains, in an effective and flexible manner.

This dataset provides a huge amount of information on crop production in India ranging from several years. Based on the Information the ultimate goal would be to predict crop production and find important insights highlighting key indicators and metrics that influence crop production.

Make views and dashboards first and also make a story out of it.

Project Architecture



Dataset

[56]:
```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.ticker as ticker
import plotly as px
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# Display plots inline
%matplotlib inline
# Load the dataset
file_path = '/Users/aadithyaram/Desktop/cropanalysis.csv'
data = pd.read_csv(file_path)

# Display the first few rows of the dataset
data.head()
```

[56]:

	State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
0	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254.0	2000.0
1	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Other Kharif pulses	2.0	1.0
2	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Rice	102.0	321.0
3	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Banana	176.0	641.0
4	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Cashewnut	720.0	165.0

ETL

[57]:
```
data.shape
```

[57]:
```
(246091, 7)
```

[58]:
```
data.columns
```

[58]:
```
Index(['State_Name', 'District_Name', 'Crop_Year', 'Season', 'Crop', 'Area',
       'Production'],
      dtype='object')
```

[59]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 246091 entries, 0 to 246090
Data columns (total 7 columns):
 #   Column         Non-Null Count   Dtype  
---  ------         --------------   -----  
 0   State_Name     246091 non-null  object 
 1   District_Name  246091 non-null  object 
 2   Crop_Year      246091 non-null  int64  
 3   Season         246091 non-null  object 
 4   Crop           246091 non-null  object 
 5   Area           246091 non-null  float64
 6   Production     242361 non-null  float64
dtypes: float64(2), int64(1), object(4)
memory usage: 13.1+ MB
```

[60]:
```
data.isnull().sum()
```

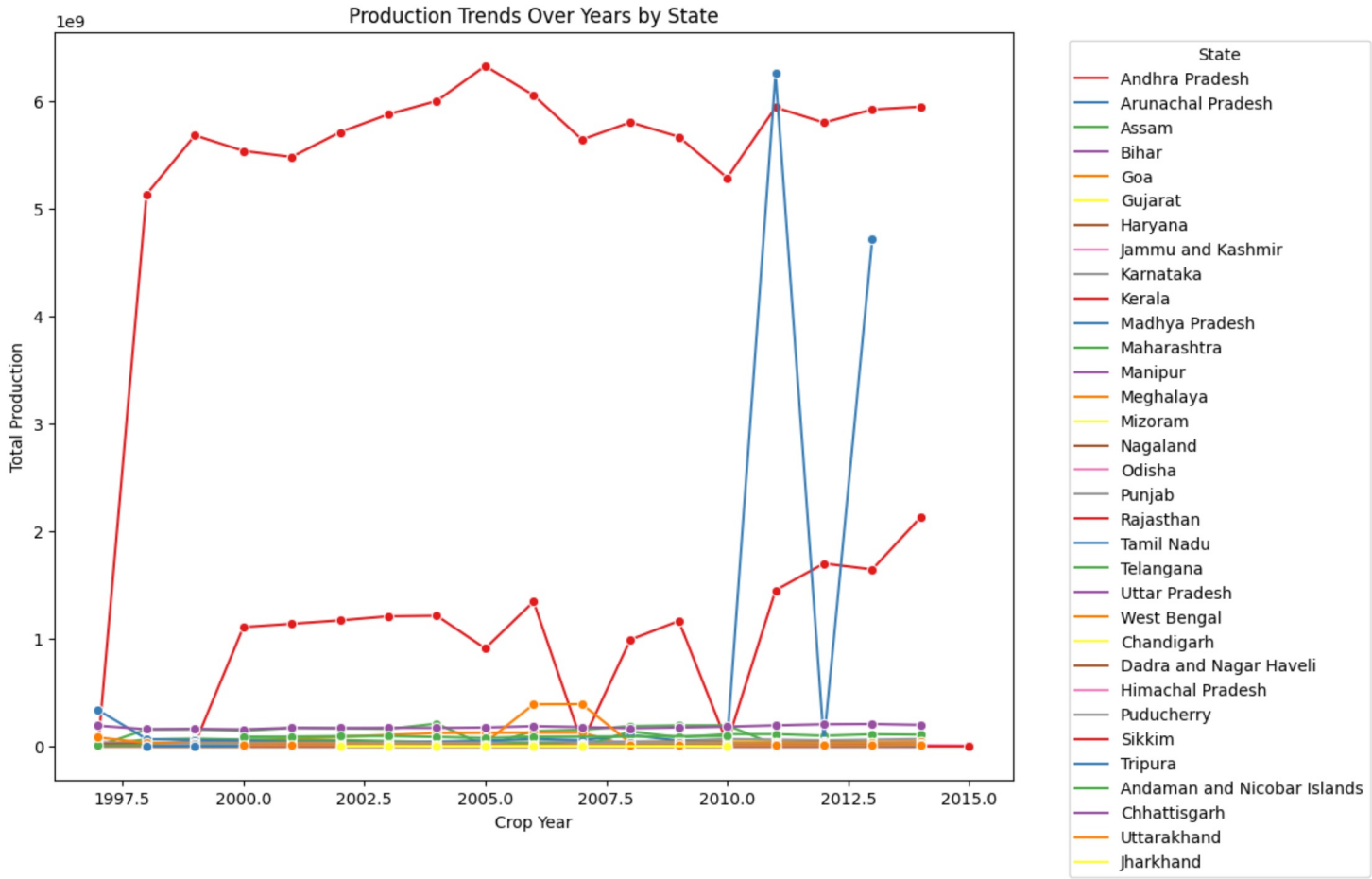
[60]:
```
State_Name          0
District_Name       0
Crop_Year           0
Season              0
Crop                0
Area                0
Production       3730
dtype: int64
```

[61]:
```
data.describe()
```

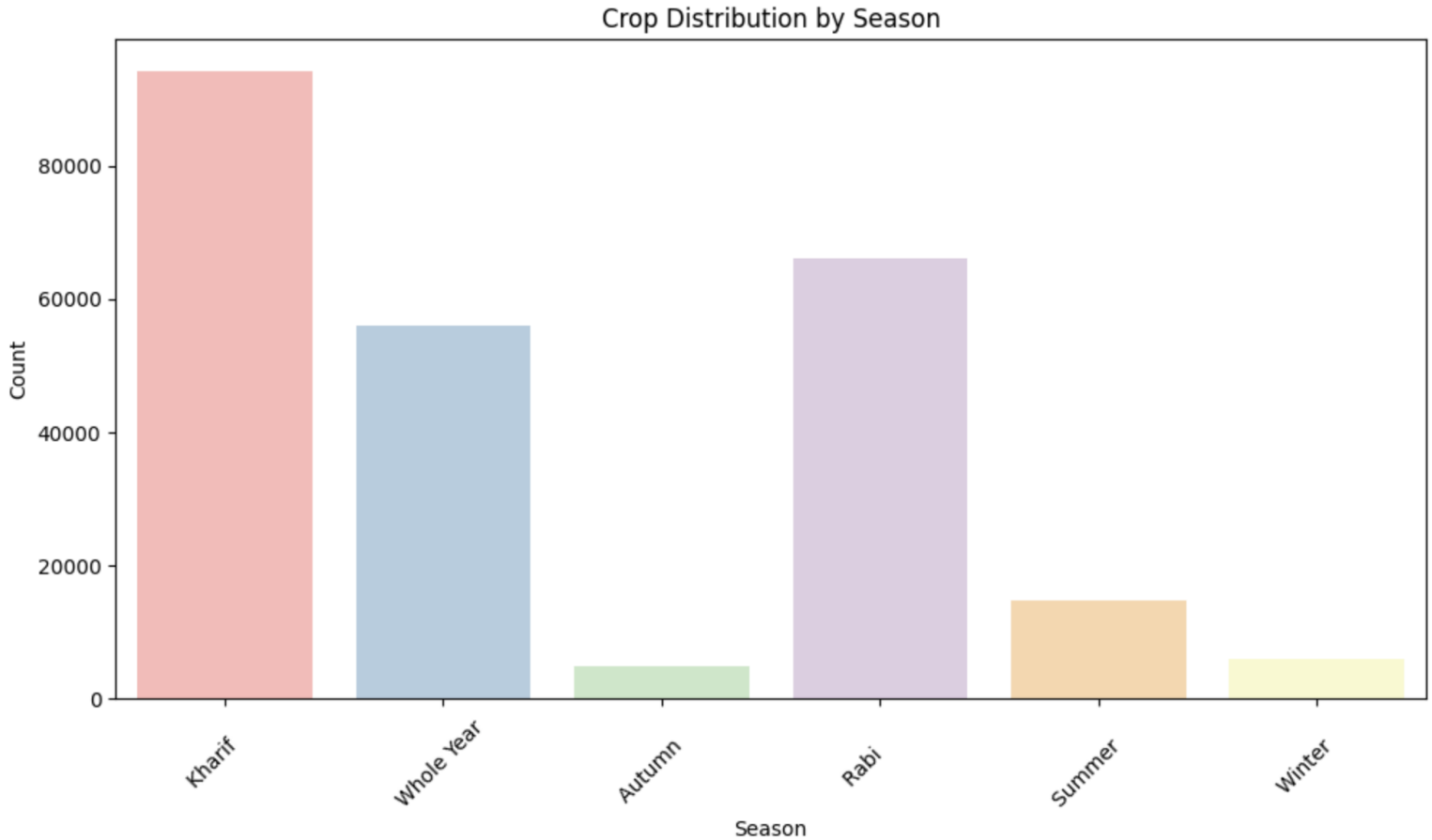
[61]:

	Crop_Year	Area	Production
count	246091.000000	2.460910e+05	2.423610e+05
mean	2005.643018	1.200282e+04	5.825034e+05
std	4.952164	5.052340e+04	1.706581e+07
min	1997.000000	4.000000e-02	0.000000e+00
25%	2002.000000	8.000000e+01	8.800000e+01
50%	2006.000000	5.820000e+02	7.290000e+02
75%	2010.000000	4.392000e+03	7.023000e+03
max	2015.000000	8.580100e+06	1.250800e+09

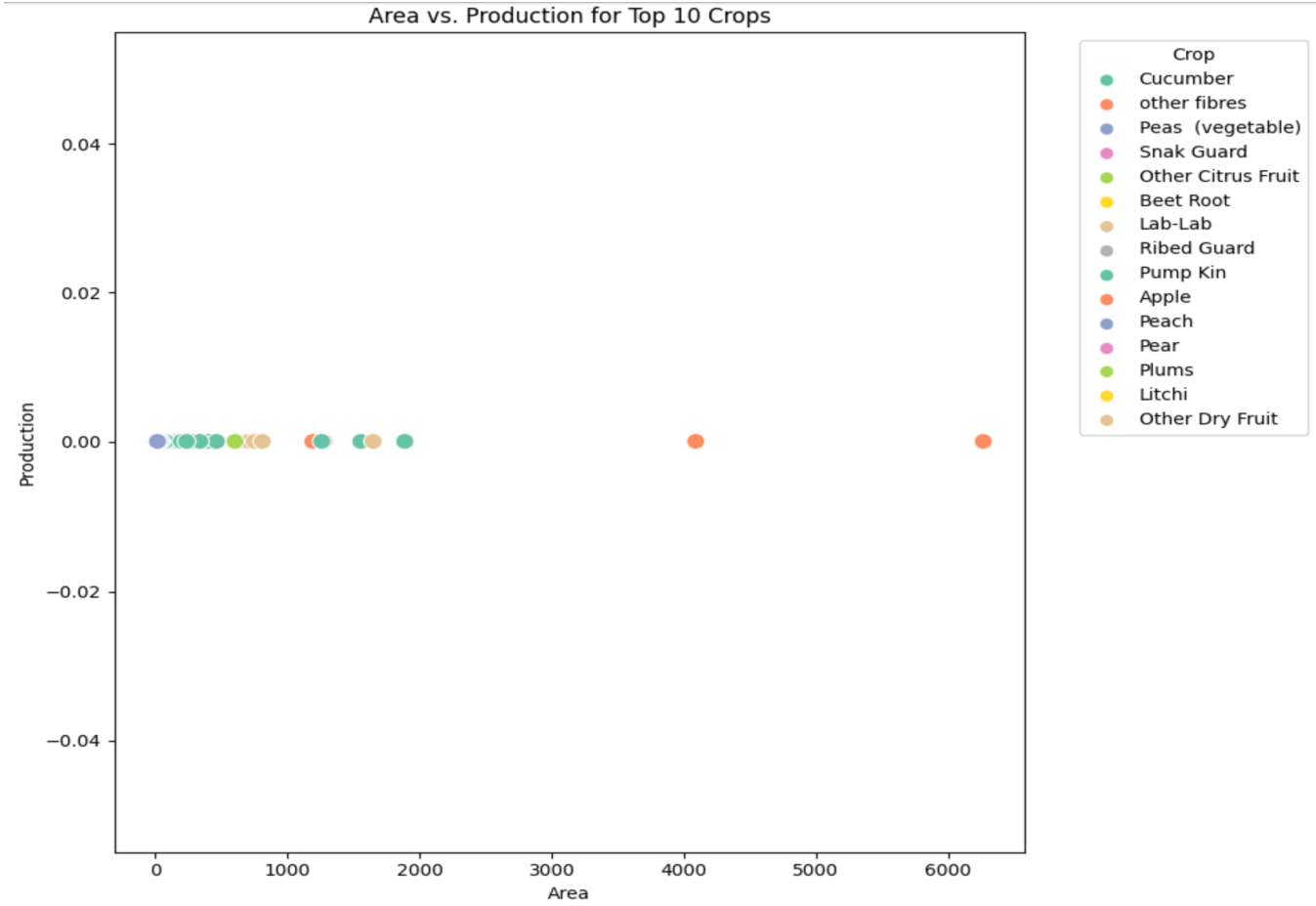
Production Trends



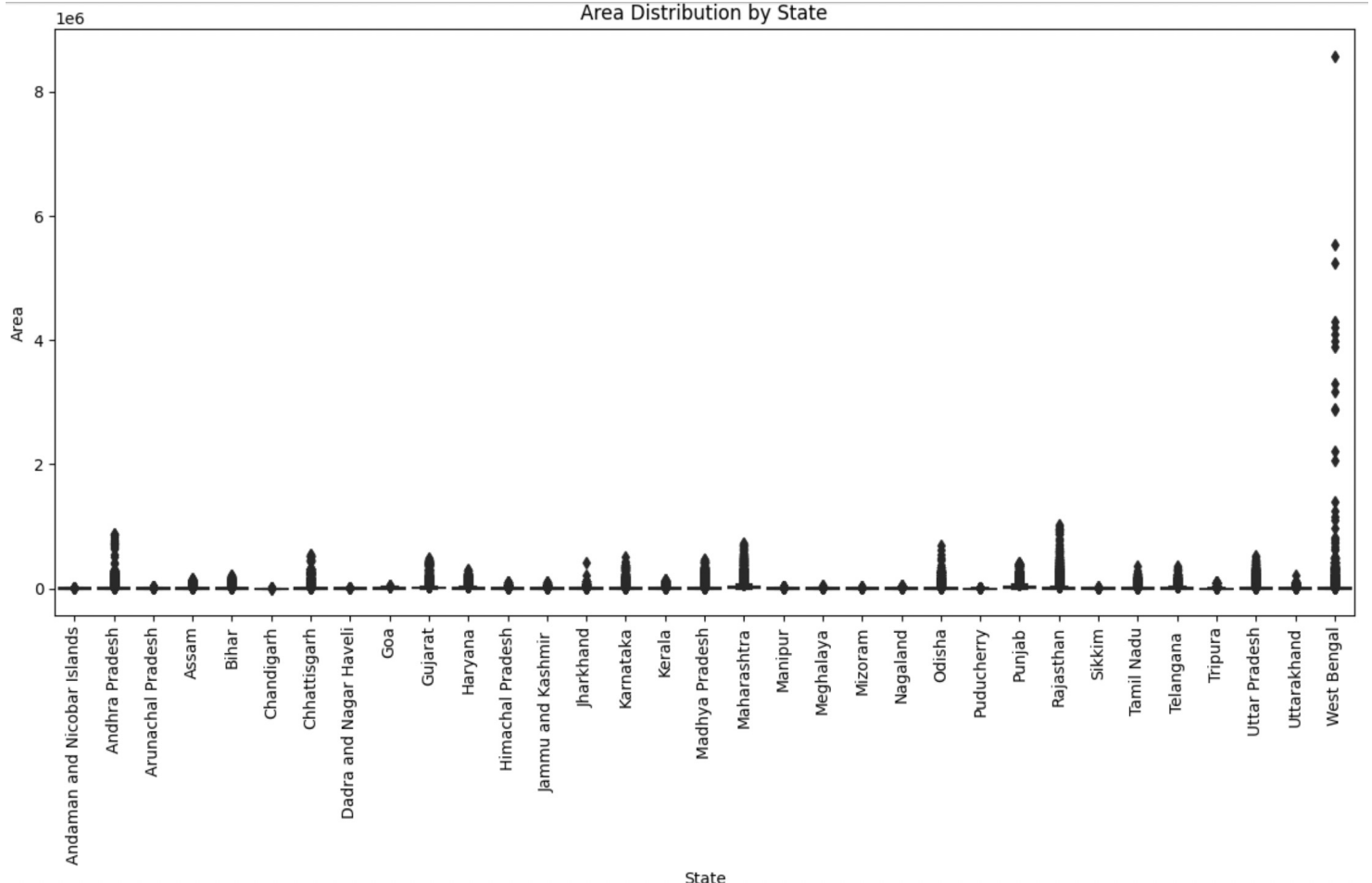
Crop Distribution



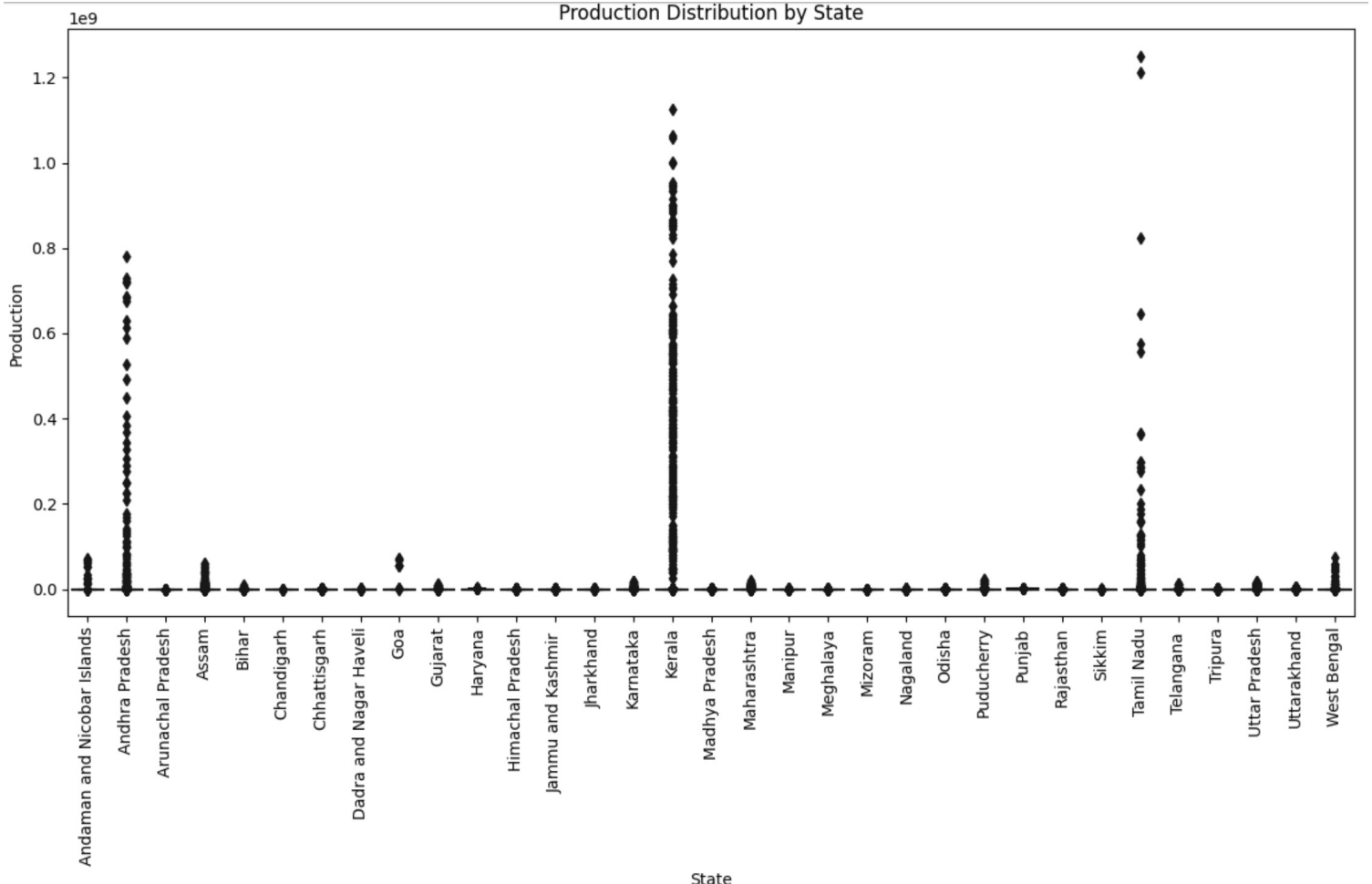
Crop Production



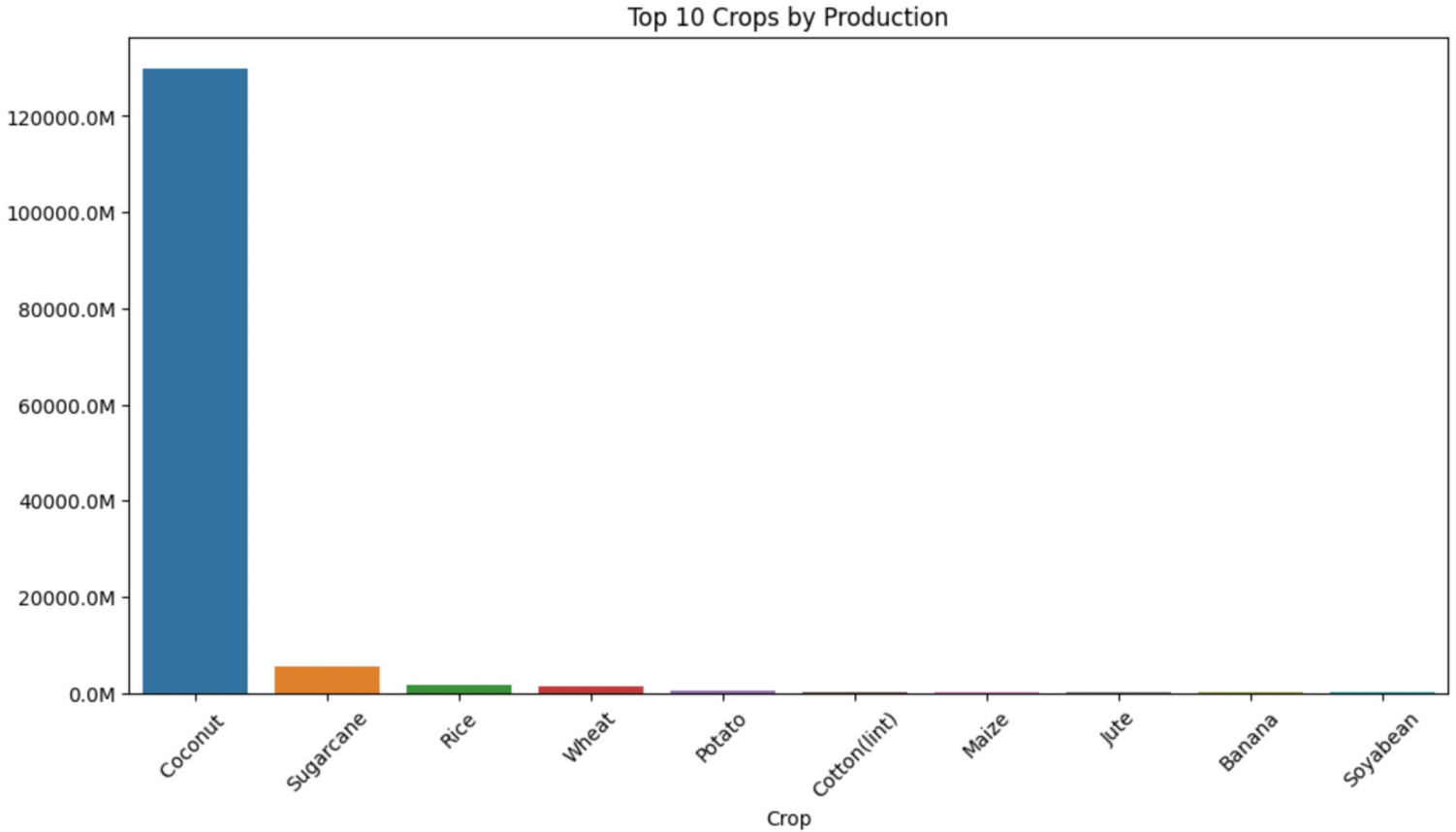
State Wise Distribution



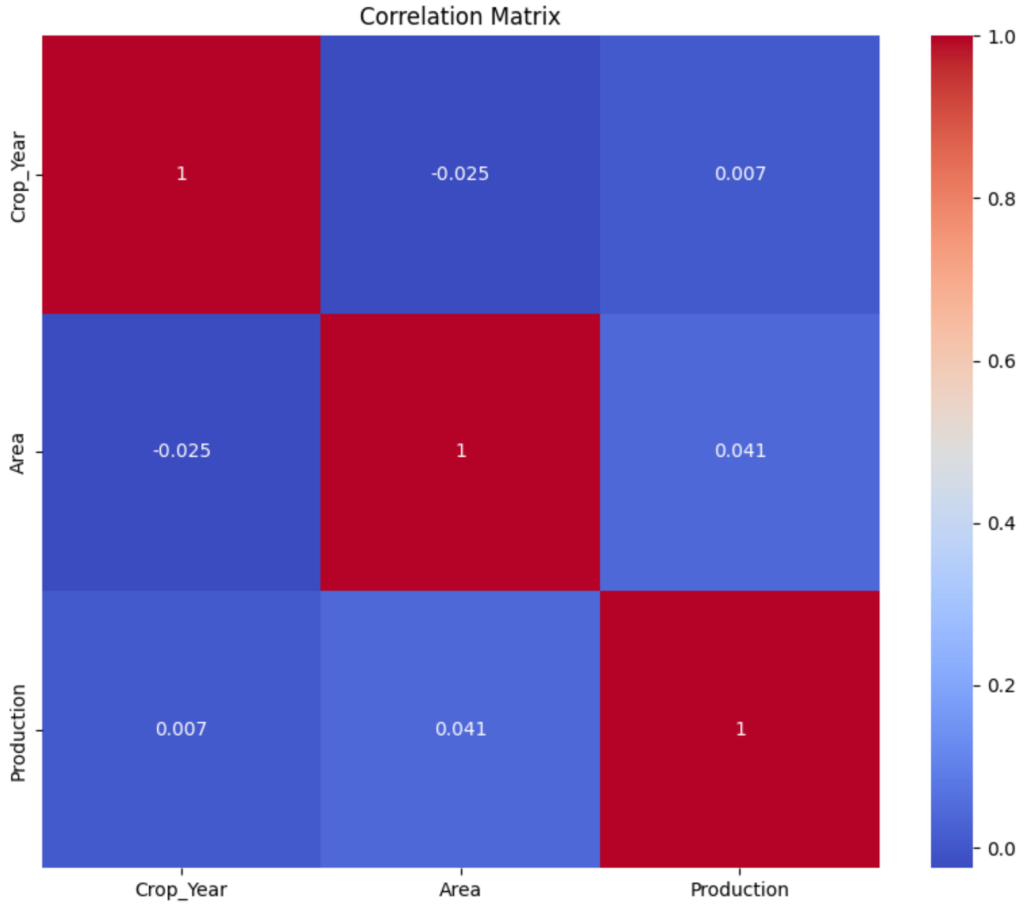
State Wise Distribution



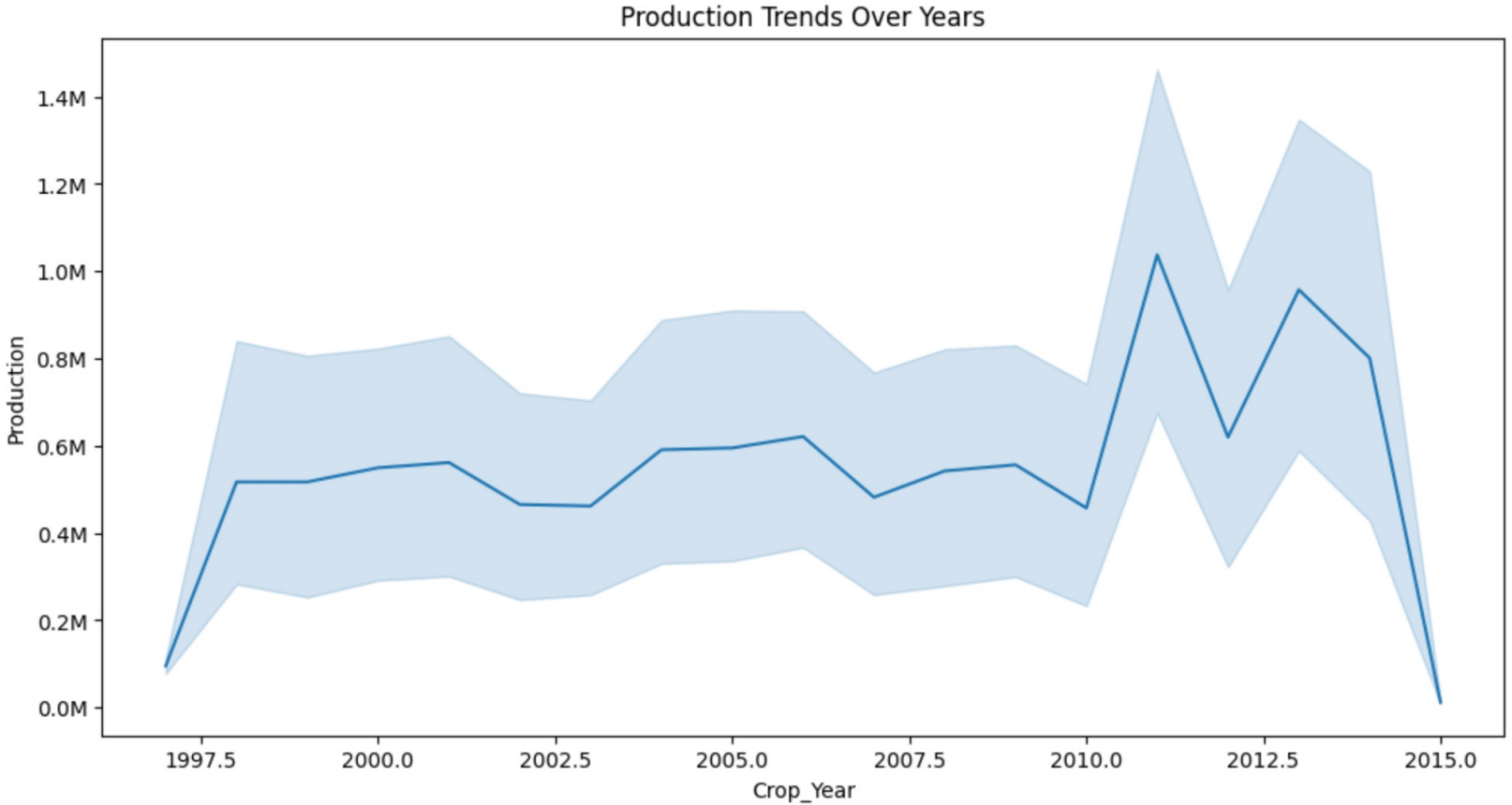
Top Productions



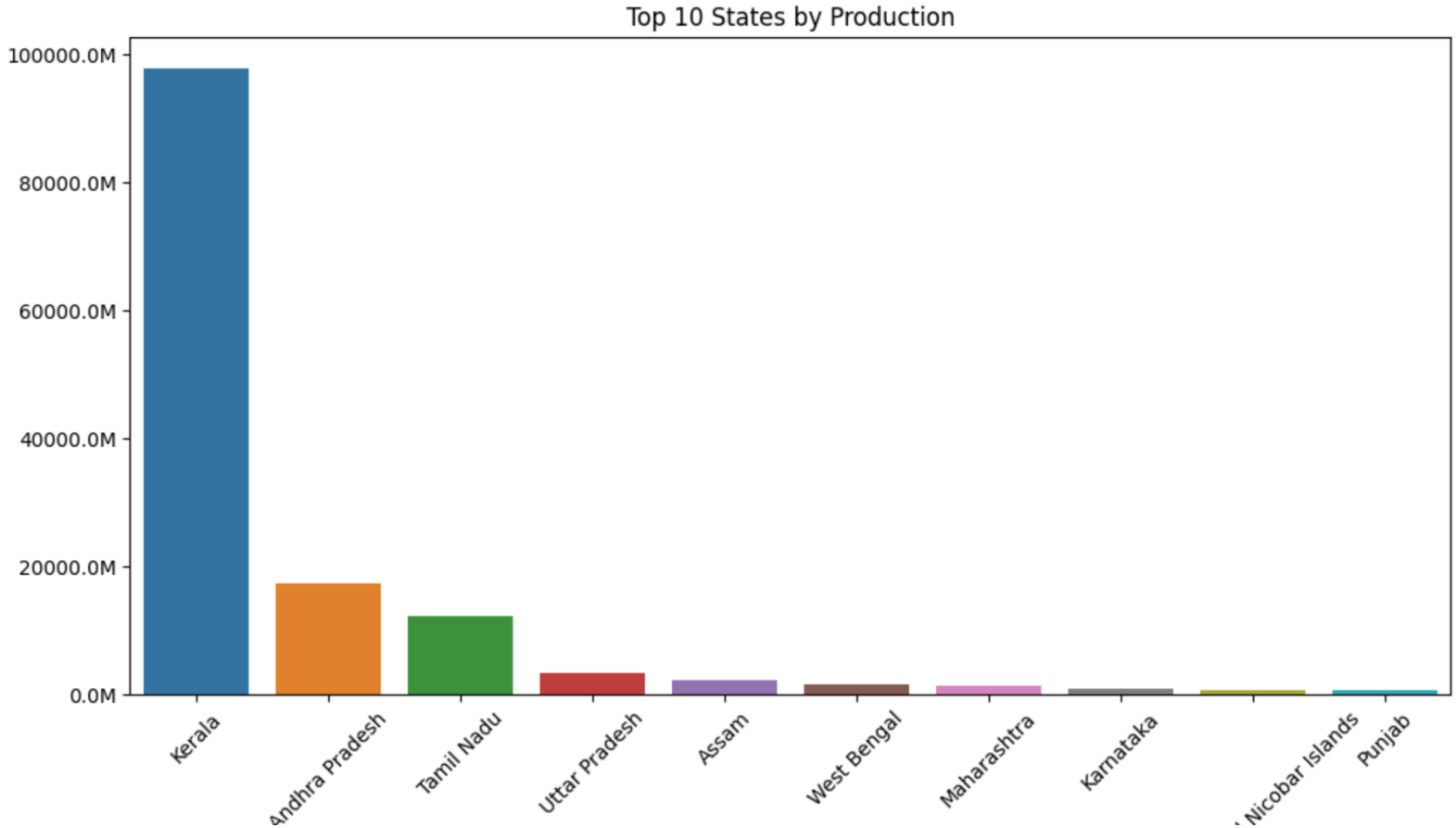
Correlations



Production Trends



Top Productions



Crop Prediction

[74]:
```
# Drop rows with missing values
data = data.dropna(subset=['Production'])
missing_values = data.isnull().sum()
print(missing_values)
```

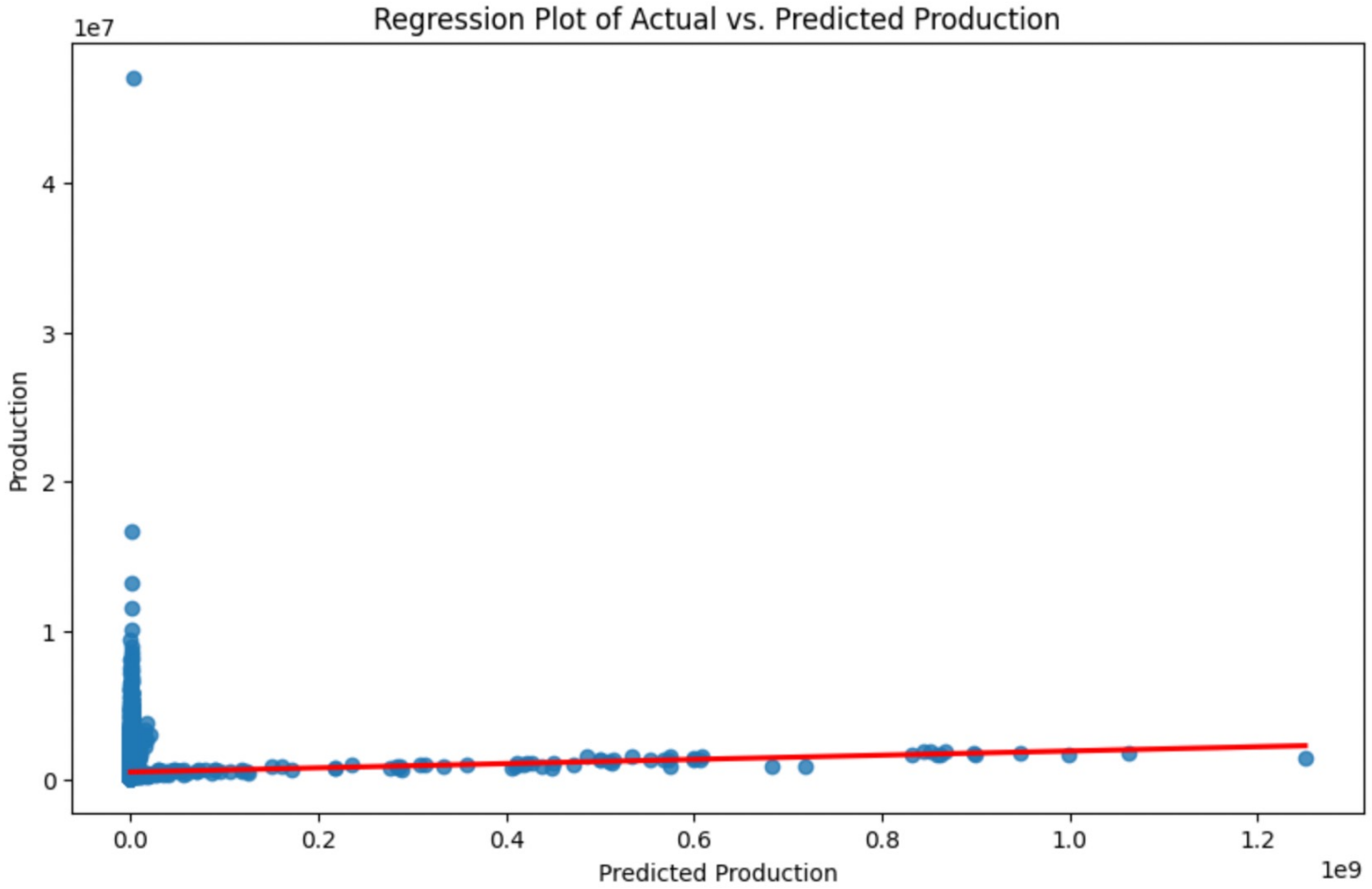
```
State_Name       0
District_Name    0
Crop_Year        0
Season           0
Crop             0
Area             0
Production       0
dtype: int64
```

[75]:
```
X = data[['Crop_Year', 'Area']]
y = data['Production']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

#Regression Model
model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

#Model Eval
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print(f'Mean Squared Error: {mse}')
print(f'R-squared: {r2}')
```

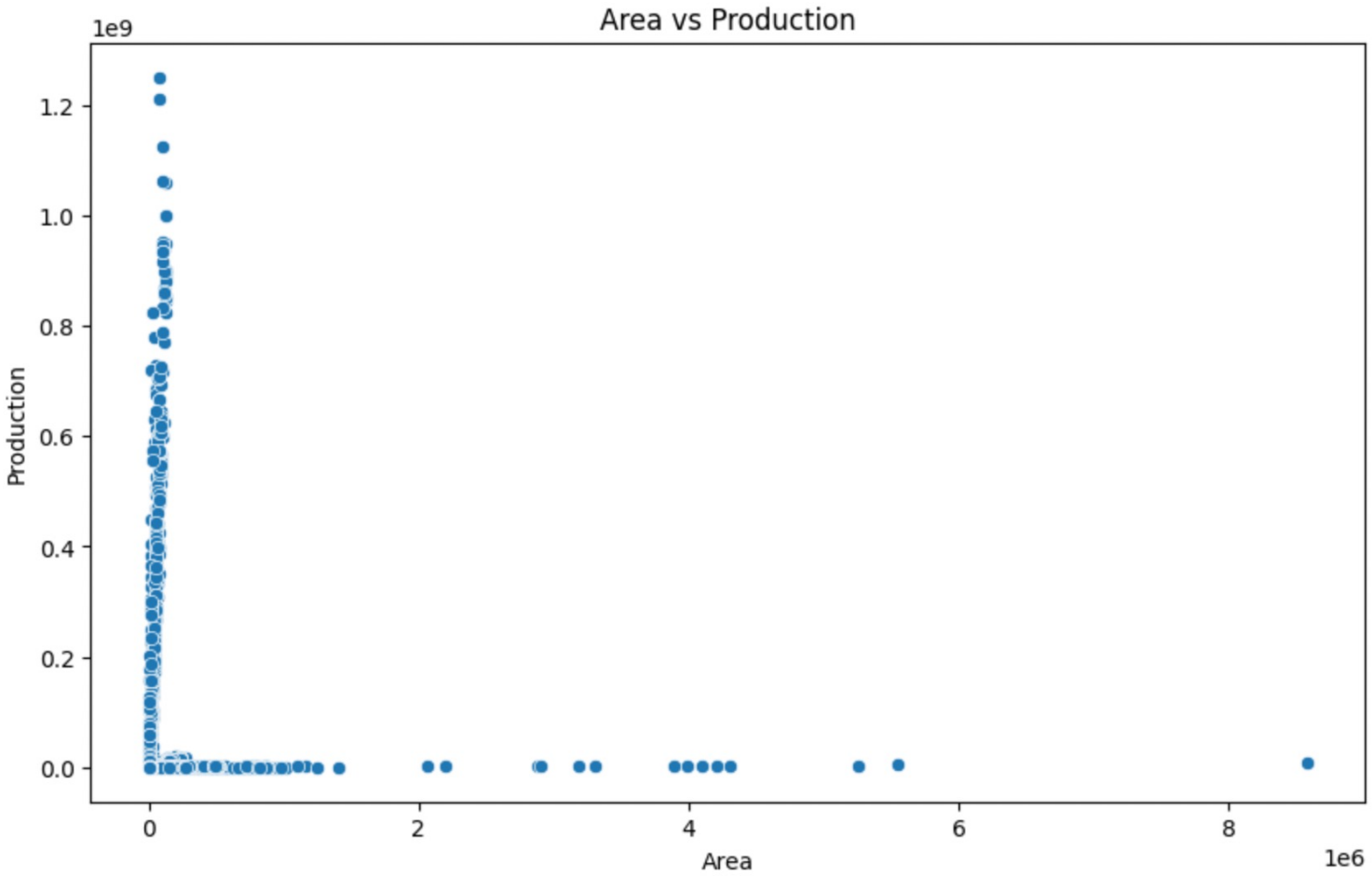
```
Mean Squared Error: 401500837552061.3
R-squared: 0.00208025972940562
```

[76]:
```
# Regression plot of Actual vs. Predicted Production
plt.figure(figsize=(10, 6))
sns.regplot(x=y_test, y=y_pred, line_kws={"color": "red"})
plt.xlabel('Predicted Production')
plt.ylabel('Production')
plt.title('Regression Plot of Actual vs. Predicted Production')
plt.show()
```

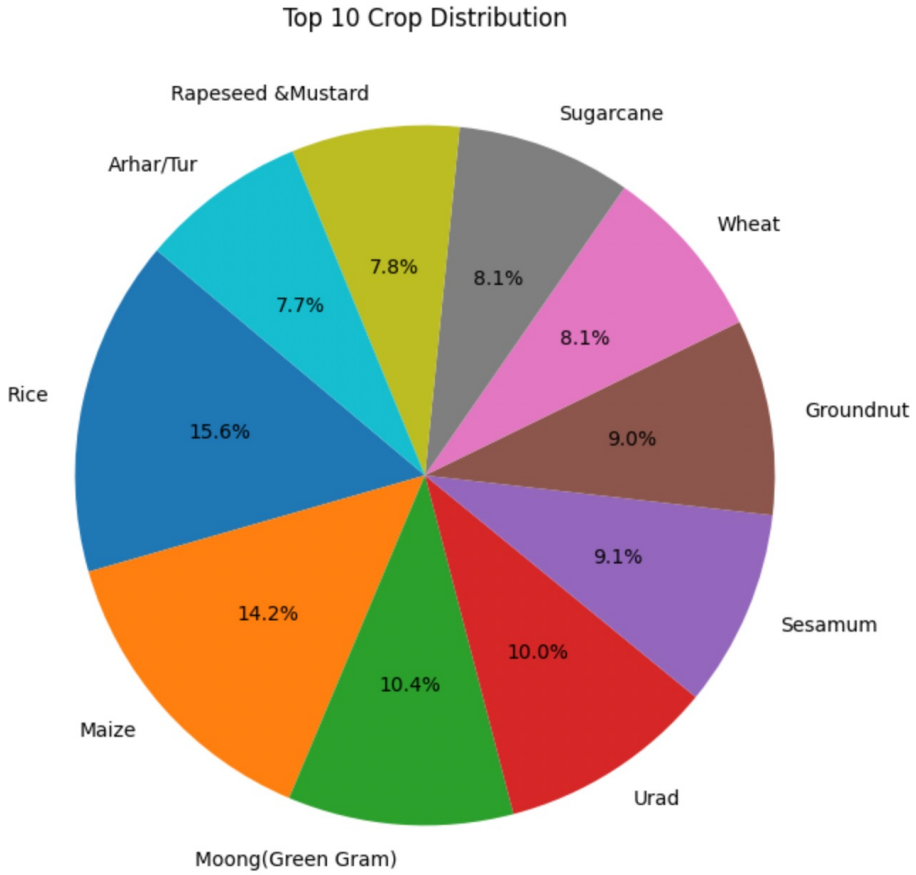
Crop Prediction



Crop Prediction



Crop Distribution



THANK YOU

Report by - Aadithya Ram

Full Code - https://github.com/Aadithya-4010002/Amazon-Sales-Data-Analytics
LinkedIn – linkedin.com/in/aadiithyyya