EXPLORATORY DATA ANALYSIS

AGRICULTURE DATA

Summary

Given agriculture data refers to three states production. This data contains a total of 28824 records. These records are taken from year 1997 to 2014 (17 Years). In this data, we have State name, District Name, Crop year, Season, Crop, Area and Production as our variables. The three states are Andhra Pradesh, Telangana, and Tamil Nadu. There are a total of 54 districts in which Andhra Pradesh has 13 districts, Telangana has 10 districts and Tami Nadu has 31 districts. And the three seasons are Kharif, Rabi and Whole year.

Data

State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
Andhra Pradesh	ANANTAPUR	1997	Kharif	Arhar/Tur	21400	2600
Andhra Pradesh	ANANTAPUR	1997	Kharif	Bajra	1400	500
Andhra Pradesh	ANANTAPUR	1997	Kharif	Castor seed	1000	100
Andhra Pradesh	ANANTAPUR	1997	Kharif	Cotton(lint)	7300	9400
Andhra Pradesh	ANANTAPUR	1997	Kharif	Dry chillies	3700	7100

Describing Data

	Crop_Year	Area	Production
count	28418.000000	28418.000000	2.841800e+04
mean	2005.334753	10846.467521	1.046385e+06
std	4.953926	34525.760259	2.169548e+07
min	1997.000000	1.000000	0.000000e+00
25%	2002.000000	110.000000	1.220000e+02
50%	2005.000000	994.000000	1.313000e+03
75%	2010.000000	6041.750000	1.240000e+04
max	2014.000000	877029.000000	1.250800e+09

Describing Data

Above we can observe there are only three variables are taken into consideration because the rest are categorical variables. In the table, we can see the maximum, minimum, Standard deviation, count, mean and quartiles. The production has a maximum value of 1250800000 and a minimum value of 0. For the Area, the maximum value is 877029 and the minimum value is 1

OBSERVATIONS

Special Character and zeros in the target variable

- In the data, our target attribute is production. After checking all the data types of the variables we found that production is in object type (Categorical). To build a model the target variable should be numeric. But, while converting into numeric we found 406 rows in production has special character '='. So, after analysing we decided to remove the rows because 406 samples do not affect too much the data which contains 28824 samples.
- For the same production column, we found there are 1067 records that contain zero. But zero is also a value and it is important. So, we do not want to remove them.

Correlation

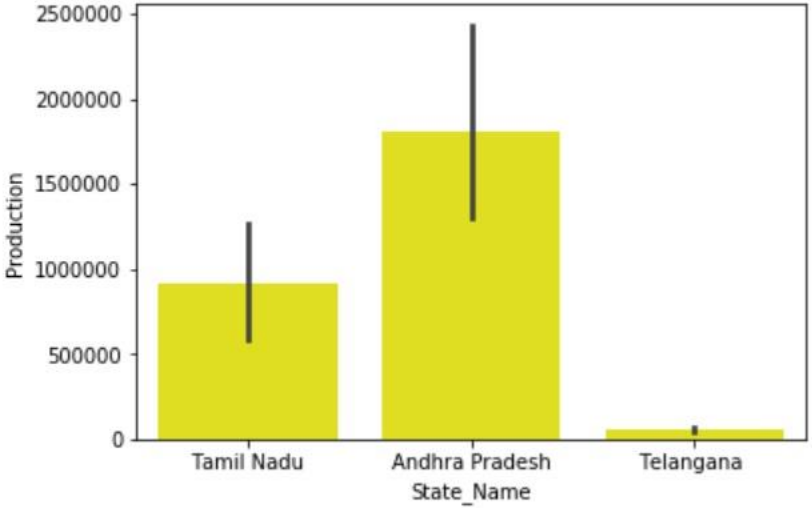
As our target attribute is the production we want to check the relation between the data for that we performed a correlation test. From this test, we can infer how much the independent attributes correlating with the target attribute.

State_Name	0.366
District_Name	0.391
Crop_Year	-0.000
Season	-0.037
Crop	-0.161
Area	0.717

From the above values, we can see that the Area is highly correlating with the target attribute (production) and the least correlated variables are Crop_year and Season.

State which has overall highest and lowest production

Andhra Pradesh has overall highest production and Telangana has the lowest production from 1997 - 2014.



Top 5 Districts which has highest Production from 1997 - 2014

Top five districts which have maximum production and we observed that **Coimbatore**, **Thanjavur**, and **Tiruppur** from Tamil Nadu. **East Godavari** and **West Godavari** from Andhra Pradesh have taken the top five places when compared with all the years.

Top 5 Districts which has Lowest Production from 1997 – 2014

Top five districts which have minimum production and we observed that **Adilabad** from Telangana. **Ramanathapuram**, **Puddukotta**, and **Perambalur** from Tamil Nadu. **Prakasam** from Andhra Pradesh have taken taken top five place when compared with all the years.

Maximum area used for cultivation of the crop

In 2005, **Anantapur** from Andhra Pradesh has used maximum area of 877029 for cultivation of groundnut.

Which State used maximum overall area for cultivation

In 17years, **Andhra Pradesh** is number one for using maximum overall area for cultivation.

From 1997 – 2014 which year has high overall production

In **2011**, the overall production is high.

From 1997 – 2014 which year has low overall production

In **2010**, the overall production is low.

Highest production season from 1997 – 2014

In **Whole year** season the production is high.

Lowest production season from 1997 – 2014

In **Rabi season** the production is low

Top 15 Crops which have maximum production

Coconut	2.784942e+10
Sugarcane	1.137392e+09
Rice	3.130608e+08
Banana	7.436034e+07
Maize	6.028835e+07
Tapioca	5.762065e+07
Cotton(lint)	5.338759e+07
Groundnut	4.196935e+07
Onion	1.475339e+07
Jowar	1.390801e+07
Dry chillies	1.237079e+07
Mango	9.982900e+06
Gram	9.181453e+06
Total foodgrain	9.121209e+06
Urad	7.533542e+06

Top 15 Crops which have minimum production.

Apple	0
Other Fresh Fruits	0
Snak Guard	0
Ribed Guard	0
Redish	0
Rapeseed &Mustard	0
Lab-Lab	0
Pump Kin	0
Linseed	0
Litchi	0
Pome Granet	0
Plums	0
Peas (vegetable)	0
Pear	0
Moong(Green Gram)	0

Top 5 crops which have used maximum area for cultivation

Groundnut	877029
Rice	409286
Total foodgrain	367554
Cotton(lint)	361169
Gram	239387

Top 5 crops which have used minimum area for cultivation

Apple	1
Pump Kin	1
Potato	1
Pome Granet	1
Plums	1

Which season utilized maximum and minimum area overall

Kharif	198532196
Rabi	60268470
Whole Year	49434248

Here **Kharif** has utilised maximum area and **Whole year** has utilised minimum area from 1997-2014

Top 5 crops which have used overall maximum area in 17 years

Crops utilized maximum area from 1997 – 2014

Rice	102741891
Groundnut	38568960
Cotton(lint)	25860232
Maize	14954632
Jowar	13513919

Top 5 crops which have used overall minimum area in 17 years

Crops utilized minimum area from 1997 – 2014

Other Dry Fruit	7
Apple	9
Litchi	25
Peach	42
Turnip	62

Thank You