CS534: EXPLORING THE AGRICULTURE IN INDIA USING SPARK FINISHED

~~Omkar Asawale (osa20)~~

Pranita Eugena Burani (peb63)

Introduction

Agriculture plays a vital role in Indian Economy.
It contributes about 18% towards Annual GDP, 70% of jobs in rural areas are agriculture dependend.
Crop Yield in any country mainly depends on technological advancements and weather variability. Technological advancements do not involve uncertainities, they only contibute towards the increase in production, therefore some parameter of time can be used used to study the effect of technology on yield, whereas the weather conditions have always been an uncontrolled source for the prediction of yield.
But due to the latest advancements, weather forecast has been reliable and accurate. This makes the whole idea of predicting agricultural yield acceptable.

Data Collection/ Source of the Data

We wanted to know much India has progressed since its independence and the formation of Republic of India in 1950.
The Indian Agriclture and Climate Data set we are using is from the dataset compiled by Duke university. “Information about data for development research”.
The database provides district level data on agriculture and climate in India from 1957/58 through 1986/87. The dataset includes information on
1. Area planted, production and farm harvest prices for five major and fifteen minor crops.
2. Areas under irrigated and high-yielding varieties (HYV) for major crops.
3. Data on agricultural inputs, such as, fertilizers, bullocks and tractors - in both quantity and price terms
4. Agricultural labor, cultivators, wages and factory earnings, rural population and literacy proportion.
5. Meteorological station level climate data (average climate over 30 year period)
6. Soil data

Objectives

1. Predict the crop yield for major crops in one state based on weather conditions.

2. Recommend high yield variety for every district in the state.

Data Description

Apart from basic ID's the important features in the dataset we used in this analysis are
1. STATENAM - Name of the state
2. DISTNAME - Name of the district
3. ACROP - Area of the crop planted(number*1000 ha)
4. QCROP - Production of the crop(number*1000 tonnes)
5. PCROP - Price of the crop(Rupees/quintal)
6. RNMONTH - Rainfall in mm for every month
7. TNMONTH - Temperature in degree celcius for every month
8. YEAR - 1956-1987

For more detailed description of the data https://ipl.econ.duke.edu/dthomas/dev_data/datafiles/india_agric_climate.htm (https://ipl.econ.duke.edu/dthomas/dev_data/datafiles/india_agric_climate.htm)

Preprocessing

The dataset provided was in cfm format(Coldfusion Markup Language) and each cfm file is a year's worth of data.
Each file contains a contains a continuous list of space-separated values; these are observations for 271 districts and 227 variables per year.

cfm files were converted to text files. With the help of STATA software application, we converted .txt files to .dta files, finally used pandas library to convert .dta to .csv files.

STATA produced csv files based on the year of data. With the help of a small python script, we were able to merge all the files into a single csv file. This file was loaded into spark for analysis.

We have used Pyspark to remove unwanted data from the csv file.
Due to inconsitencies in data from 1950 to 1955. We decided not use it for analysis in this report.

Data Analysis

Analysis 1:

Which States to Consider for the chosen Crops?

Basic objective is to maximize the yield production.
In our dataset we have data from 1956-87
So we chose to plot the above graphs for 1956, 1966,1976 and 1986

- In 1956 - Top 3 highest Rice Yield is from the states - West Bengal,Tamil Nadu and Bihar
  Lowest 3 Yield from Punjab, Haryana, Rajastan

Similarly for the years 1966, 1976,1986 The top and lowest yields are as follows
1966- High - West Bengal, Tamil Nadu, Orissa and Low - Punjab, Gujarat, Rajasthan
1976- High - West Bengal, Tamil Nadu, Bihar and Low - Madhya Pradesh, Gujarat, Rajasthan
1988- High - West Bengal, Punjab, Tamil Nadu and Low - Maharashtra ,Gujarat, Rajasthan

Not only should we cold we consider the Production but also the should consider the Area the crops were planted in.

For this the results for the above years is as follows.
1956 - High - Orissa, West Bengal, Bihar and Low - Punjab, Haryana, Rajasthan
1966 - High - Orissa, West Benagl, Bihar and Low - Gujarat, Punjab, Rajasthan
1976 - High - Orissa, West Bengal, Bihar and Low - MadhyaPradesh, Gujarat, Rajasthan
1986 - High - West Bengal, Orissa, Bihar and Low - Maharashtra, Gujarat, Rajasthan.

From the above analysis it is clear that West Bengal, Orissa, Bihar have highest area allotted for Rice plantation and also had highest yield among others states.

Similar Analysis was done for other crops and decided to use.
West_Bengal - For Rice
Uttar_Pradesh - For Sugar
Punjab - For Cotton
Punjab - For Wheat

Plots for the above analysis are in the folder with file names as crop names

Analysis 2

The cost per quintal values are plotted against the years for crops corresponding to the states of their high production and low production, no big difference in the price of the crop for that particular year, but with the increase in years the price increased(with some dips, because of few nulls in the data).

Took 0 sec. Last updated by anonymous at October 24 2019, 9:23:44 PM.

READY

Analysis 3

Comparing production of wheat:

We find that Uttar Pradesh has the largest wheat production amongst all the other states. It has a very large lead over other states.

settings

READY

Analysis 4

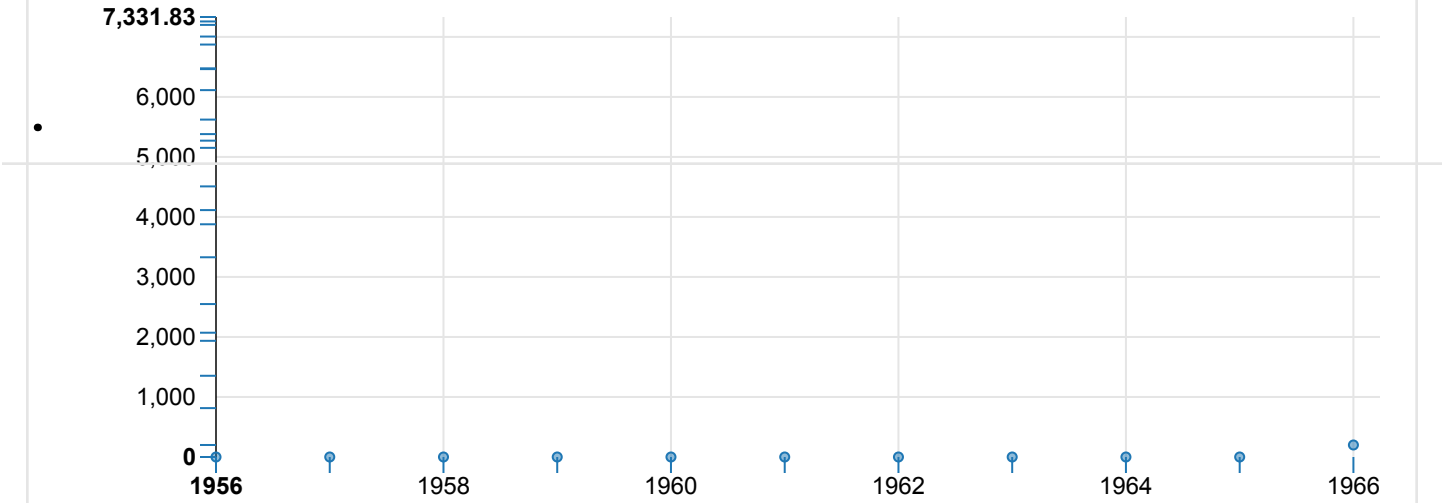
READY

Exploring the growth of wheat in Uttar Pradesh:

We find that there was very little production of wheat in the starting year of this dataset but after 1965 there was a growth spurt. The growth is almost linear except for the year 1979. We initally anticipated that rainfall might be a key factor in this. But soon after doing some research, we found that India had expreinced a slight recession. Two of the prime ministers had resigned in a quick succession.

settings

READY



Analysis 5

READY

Mansoon season in India:

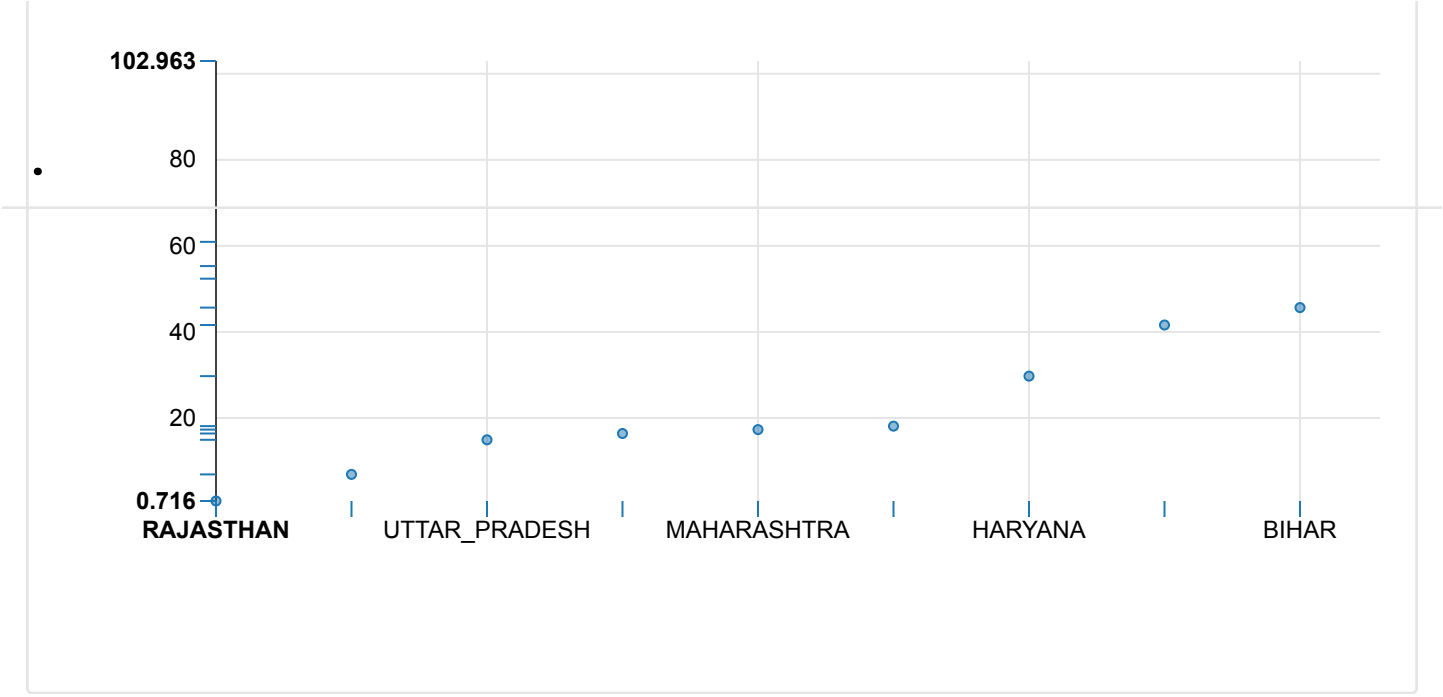
India experiences the most rainfall in the month of September. We have plotted the rainfall data for all the states. West Bengal experienes the most rainfall. It is located in Nort-West region of the country. As per common knowlegde, rice usually grows in regions where there is a lot of rainfall.

We were suprised to learn that Tamil Nadu is largest producer of rice. Although, it is located in the opposite part of the country where it does not rain as much as West Bengal.

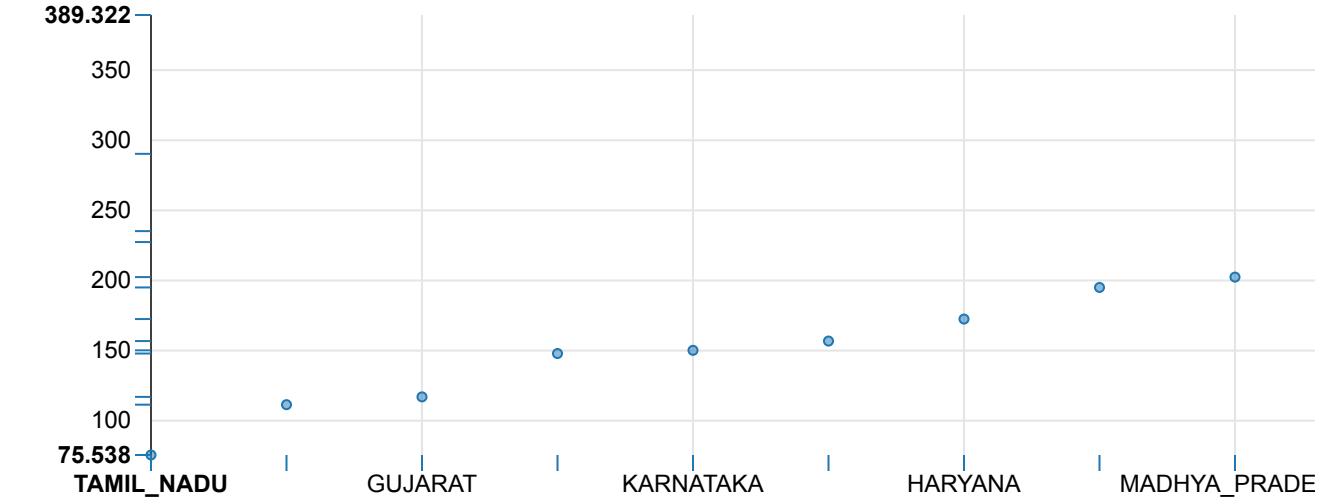
While in the rainfall chart Tamil Nadu is at the bottom. This suggests that there are some descrepancies in the dataset.

settings

READY



settings READY



Analyis 6

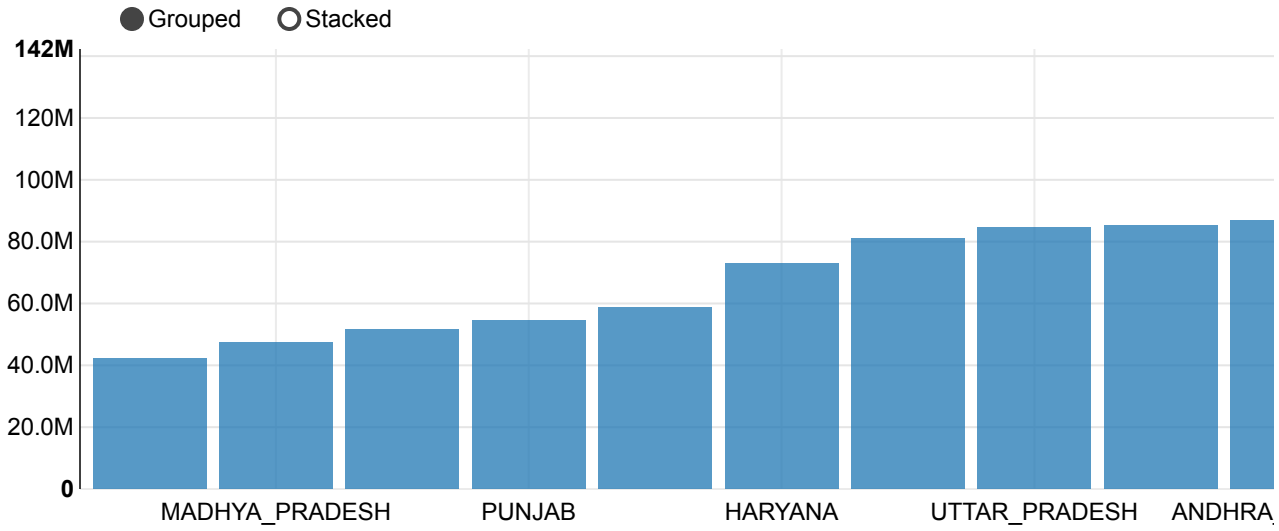
READY

Comparing mechanical labor and manual labor:

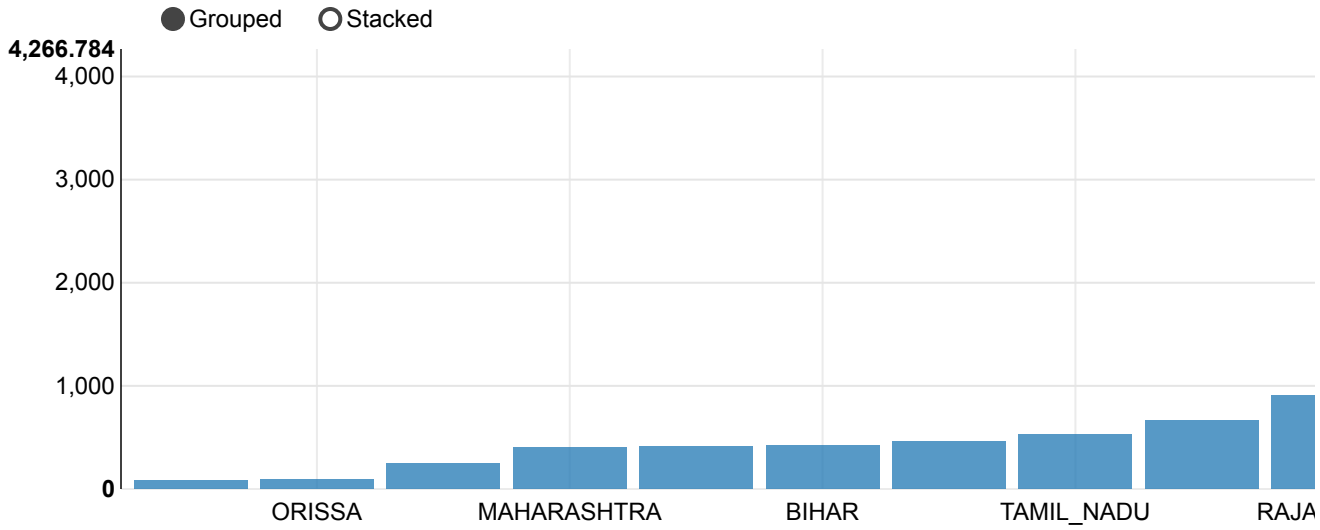
We find that Tamil Nadu has the most number of people working in fields as compared to Punjab where more people own tractors. West bengal is at the bottom beacuse there are a lot of paddy fields where the rice is cultivated so tractor does not improve

efficency.

settings READY



settings READY



Future Development

READY

We were able to find various insights with the help of data. There were a lot of instances where the data conflicted the reality. We would like to analyze this dataset even further. There are a lot of other variables in the dataset which will help us to improve the accuracy of our work. There is a lot of data on soil which we need to study and understand in more detail.

- The objective of this project is to predict the yield of the crop, This can be done by training a regression model on a significant dataset. The regression model will be used as the output of yield is a numerical value. In the next phase we plan to implement a regression model and train it on our pre-processed dataset.
Also try implementing ANN's to predict crop yield, by giving weather variables as input and get the predicted values of crop yield as output.

READY