WEEKLY PROGRESS REPORT

Name: Sahil Sanjay Pokharkar

Domain: Data Science and Machine Learning

Date of Submission: 16/01/2026

Project Title: Prediction of Crop Production in India

Organization: Upskill Campus

Week Ending: Week 02

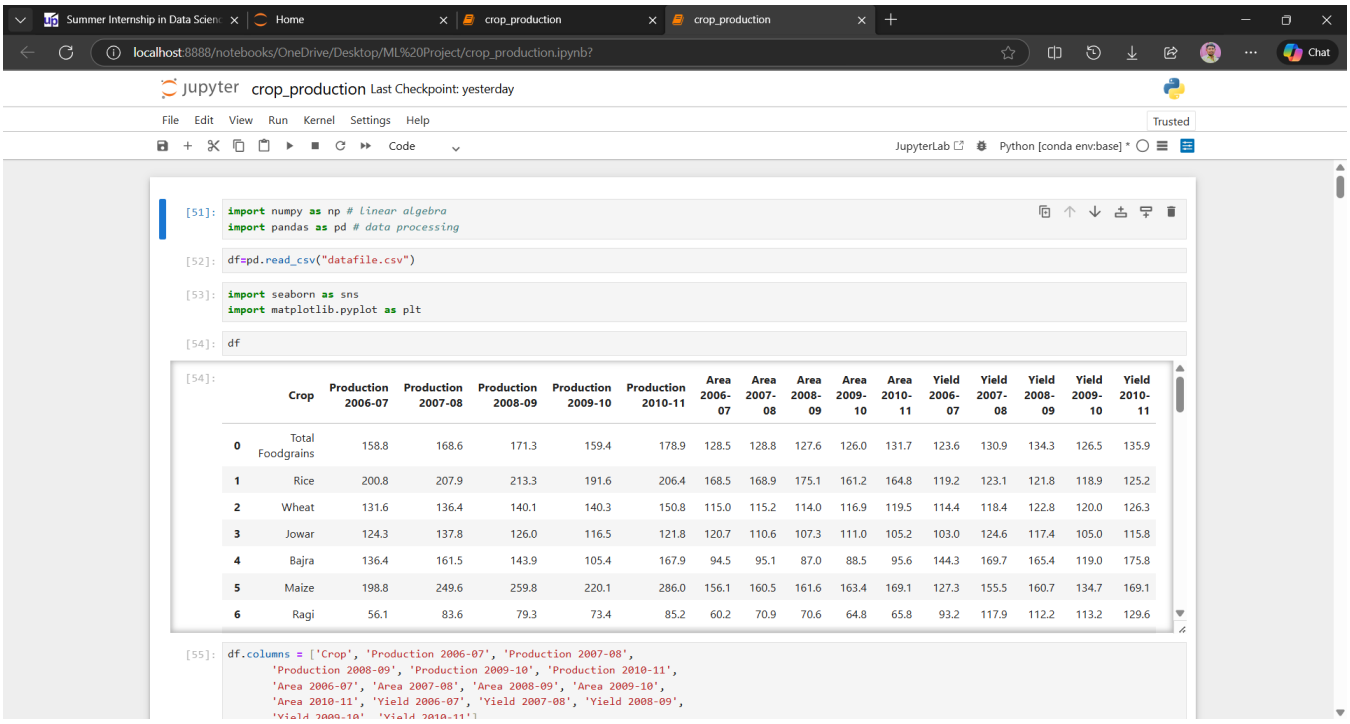
ACKNOWLEDGEMENT

I sincerely thank Upskill Campus and my mentors for their continued guidance and support throughout the second week of my internship. Their structured approach and timely feedback helped me strengthen my understanding of data preprocessing and exploratory data analysis. I am also grateful to my faculty members, family, and friends for their constant encouragement and motivation.

Table Of Content

Chapter 1: Overview

The second week of the internship was primarily focused on data cleaning, preprocessing, and performing detailed Exploratory Data Analysis (EDA) on the "Crop Production in India" dataset. The objective was to prepare a clean, structured dataset suitable for machine learning model development and to gain insights into patterns and trends related to crop production across different states, seasons, and crop types.

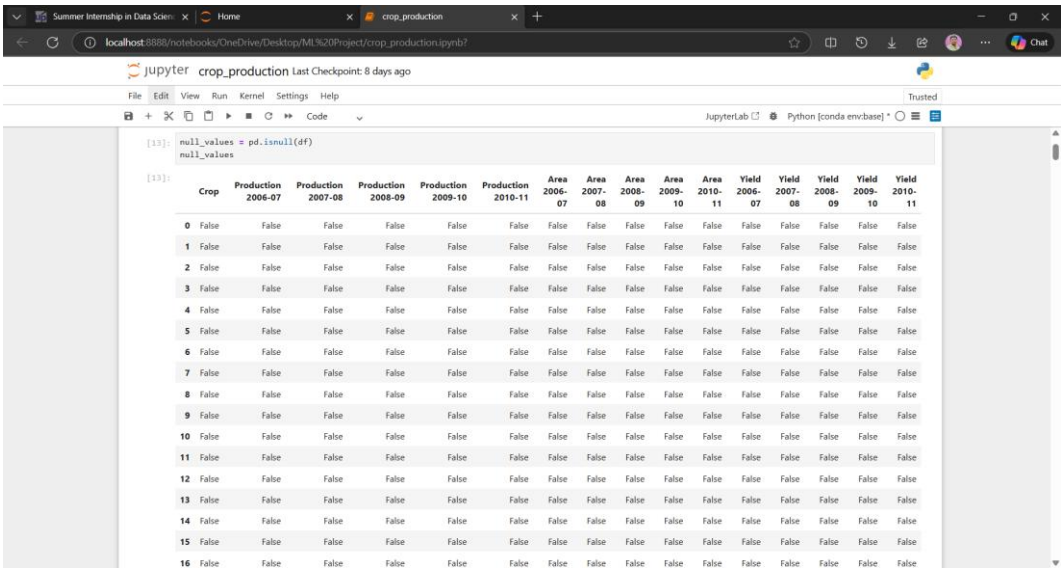


	Crop	Production 2006-07	Production 2007-08	Production 2008-09	Production 2009-10	Production 2010-11	Area 2006-07	Area 2007-08	Area 2008-09	Area 2009-10	Area 2010-11	Yield 2006-07	Yield 2007-08	Yield 2008-09	Yield 2009-10	Yield 2010-11
0	Total Foodgrains	158.8	168.6	171.3	159.4	178.9	128.5	128.8	127.6	126.0	131.7	123.6	130.9	134.3	126.5	135.9
1	Rice	200.8	207.9	213.3	191.6	206.4	168.5	168.9	175.1	161.2	164.8	119.2	123.1	121.8	118.9	125.2
2	Wheat	131.6	136.4	140.1	140.3	150.8	115.0	115.2	114.0	116.9	119.5	114.4	118.4	122.8	120.0	126.3
3	Jowar	124.3	137.8	126.0	116.5	121.8	120.7	110.6	107.3	111.0	105.2	103.0	124.6	117.4	105.0	115.8
4	Bajra	136.4	161.5	143.9	105.4	167.9	94.5	95.1	87.0	88.5	95.6	144.3	169.7	165.4	119.0	175.8
5	Maize	198.8	249.6	259.8	220.1	286.0	156.1	160.5	161.6	163.4	169.1	127.3	155.5	160.7	134.7	169.1
6	Ragi	56.1	83.6	79.3	73.4	85.2	60.2	70.9	70.6	64.8	65.8	93.2	117.9	112.2	113.2	129.6

Chapter 2: Tasks Completed & Achievements

1. Data Cleaning and Preprocessing

- Identified and handled missing and null values in critical columns such as *Area* and *Production*.
- Removed or corrected inconsistent entries and standardized categorical values for crops and seasons.
- Converted data types where necessary to ensure numerical columns were suitable for analysis and modelling.

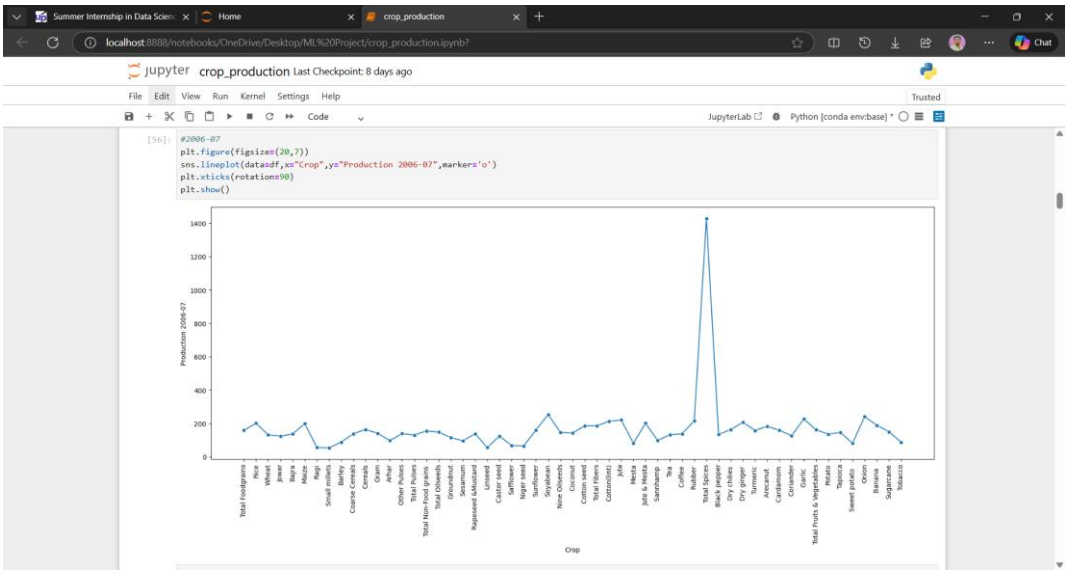


2. Exploratory Data Analysis (EDA)

- Analysed crop production trends across different states and union territories.
- Studied seasonal variations (Kharif, Rabi, Zaid) and their impact on production levels.
- Explored relationships between cultivated area and production using scatter plots and correlation analysis.
- Identified top crops in terms of production and area under cultivation.

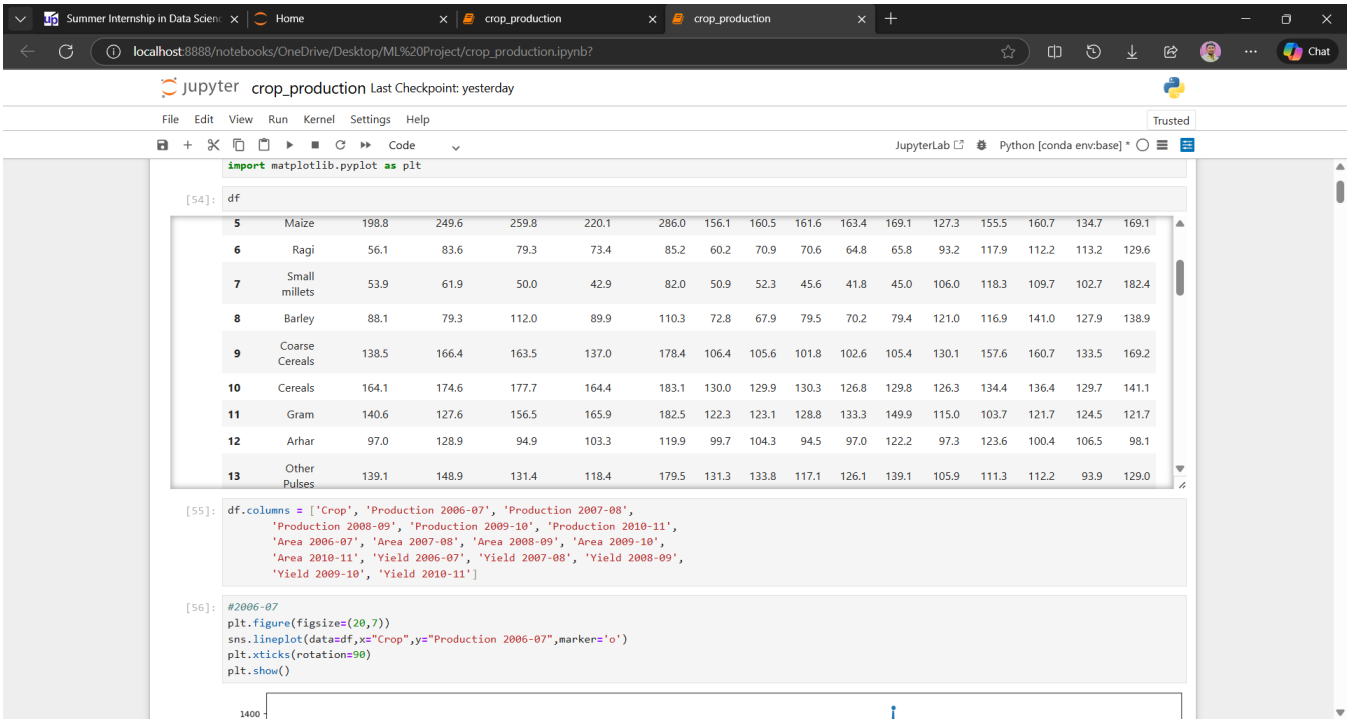
3. Data Visualization

- Created bar charts and line plots to visualize state-wise and crop-wise production trends.
- Used histograms and box plots to understand data distribution and detect outliers.
- Developed comparative visualizations to analyse production patterns across seasons.



4. Feature Understanding and Selection

- Gained clarity on which features (state, district, crop, season, area) are most influential for predicting production.
- Prepared a refined set of features for upcoming machine learning model training.

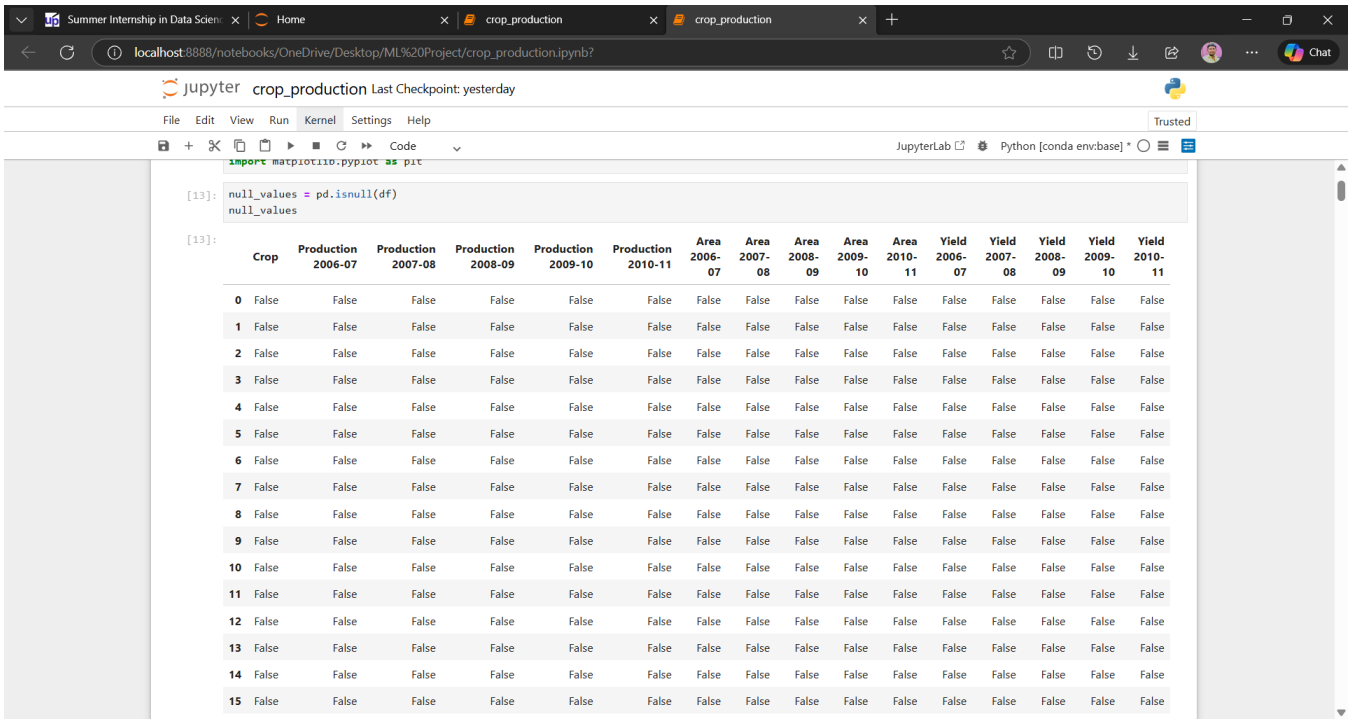


5	Maize	198.8	249.6	259.8	220.1	286.0	156.1	160.5	161.6	163.4	169.1	127.3	155.5	160.7	134.7	169.1
6	Ragi	56.1	83.6	79.3	73.4	85.2	60.2	70.9	70.6	64.8	65.8	93.2	117.9	112.2	113.2	129.6
7	Small millets	53.9	61.9	50.0	42.9	82.0	50.9	52.3	45.6	41.8	45.0	106.0	118.3	109.7	102.7	182.4
8	Barley	88.1	79.3	112.0	89.9	110.3	72.8	67.9	79.5	70.2	79.4	121.0	116.9	141.0	127.9	138.9
9	Coarse Cereals	138.5	166.4	163.5	137.0	178.4	106.4	105.6	101.8	102.6	105.4	130.1	157.6	160.7	133.5	169.2
10	Cereals	164.1	174.6	177.7	164.4	183.1	130.0	129.9	130.3	126.8	129.8	126.3	134.4	136.4	129.7	141.1
11	Gram	140.6	127.6	156.5	165.9	182.5	122.3	123.1	128.8	133.3	149.9	115.0	103.7	121.7	124.5	121.7
12	Arhar	97.0	128.9	94.9	103.3	119.9	99.7	104.3	94.5	97.0	122.2	97.3	123.6	100.4	106.5	98.1
13	Other Pulses	139.1	148.9	131.4	118.4	179.5	131.3	133.8	117.1	126.1	139.1	105.9	111.3	112.2	93.9	129.0

Chapter 3: Challenges

- **Handling Large Dataset:** Managing and processing a large dataset caused performance issues during visualization and analysis.
- **Data Imbalance:** Certain crops and states dominated the dataset, making it challenging to draw balanced insights.
- **Outliers and Anomalies:** Presence of extreme values in production data required careful handling to avoid misleading conclusions.
- **Understanding Agricultural Context:** Interpreting why certain crops performed better in specific regions required additional domain research.

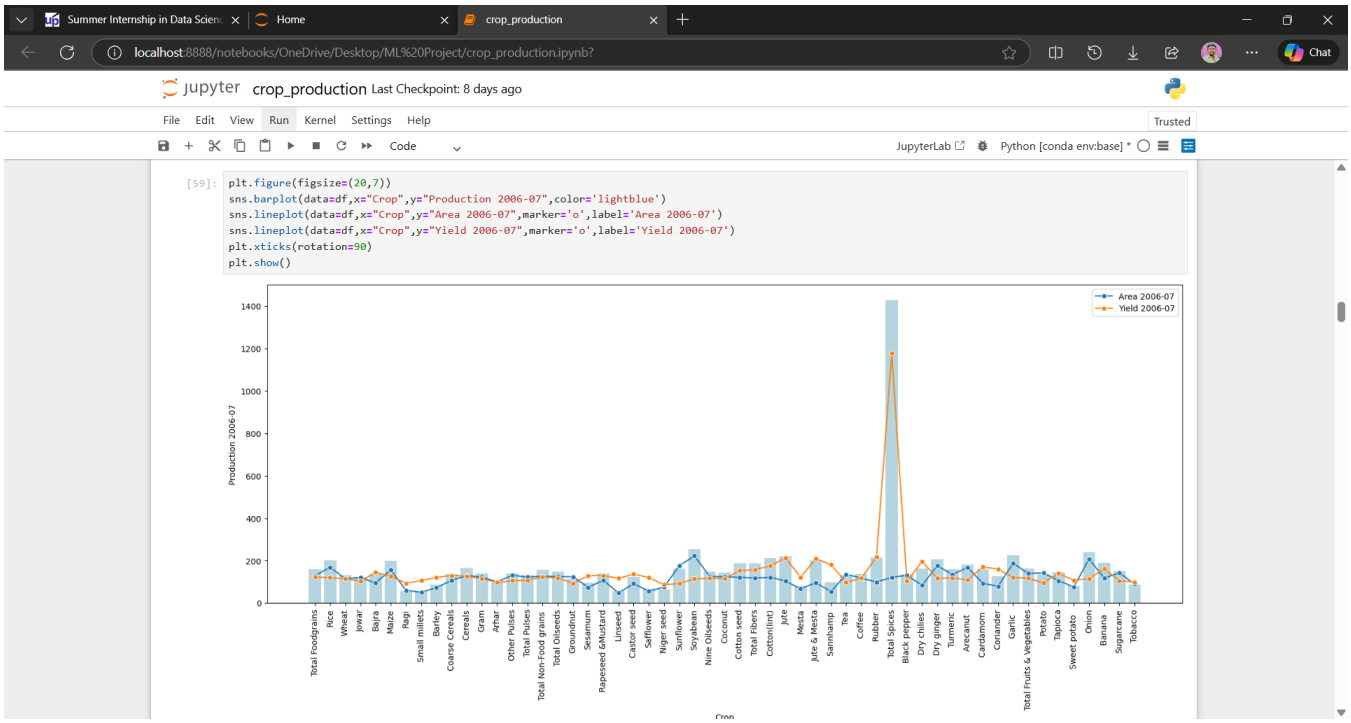
```
[7]: df.shape

[7]: (55, 16)
```



	Crop	Production 2006-07	Production 2007-08	Production 2008-09	Production 2009-10	Production 2010-11	Area 2006-07	Area 2007-08	Area 2008-09	Area 2009-10	Area 2010-11	Yield 2006-07	Yield 2007-08	Yield 2008-09	Yield 2009-10	Yield 2010-11
0	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
5	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
6	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
7	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
8	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
9	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
10	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
11	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
12	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
13	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
14	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
15	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False

Chapter 4: Lessons Learned

- Learned effective techniques to handle missing data and outliers using Pandas and NumPy.
- Gained practical experience in performing structured EDA to extract meaningful insights from raw data.
- Improved understanding of how domain knowledge (agriculture) complements data analysis for better interpretation.
- Enhanced data visualization skills to communicate insights clearly and professionally.
- Developed problem-solving skills by optimizing code to handle large datasets efficiently.



Chapter 5: Learning Resources

- Official documentation of Pandas, NumPy, Matplotlib, and Seaborn.
- Online tutorials on data preprocessing and EDA techniques.
- Upskill Campus video lectures and project guidelines.
- Articles and blogs on agricultural data analysis and crop yield prediction.

Chapter 6: Next Week's Goals

- Perform feature encoding for categorical variables.
- Split the dataset into training and testing sets.
- Implement initial machine learning models such as Linear Regression and Decision Tree.
- Evaluate model performance using appropriate metrics.
- Fine-tune preprocessing steps based on model requirements.

Chapter 7: Additional Comments

Week 02 was highly productive and focused on strengthening the data foundation of the project. The hands-on experience with data cleaning and EDA significantly improved my confidence in handling real-world datasets. The insights gained this week will play a crucial role in building accurate and reliable machine learning models in the upcoming phases of the project.