📊 Dataset Overview

- **Rows:** 13,100
- **Columns:** 10
- **Data Types:** 7 categorical (object), 3 numerical (float)
- **Columns:**
    1. `State` – State name
    2. `District` – District name
    3. `Market` – Market location
    4. `Commodity` – Type of commodity (e.g., Tomato, Maize, Onion)
    5. `Variety` – Variety of commodity (e.g., Hybrid, Local)
    6. `Grade` – Commodity grade (e.g., FAQ, Standard, Medium)
    7. `Arrival_Date` – Date of arrival (only **07/08/2025** is present → single date)
    8. `Min_x0020_Price` – Minimum price (float)
    9. `Max_x0020_Price` – Maximum price (float)
    10. `Modal_x0020_Price` – Most common (modal) price (float)

🧹 Data Characteristics

- **States:** 25 unique (e.g., Tamil Nadu, Andhra Pradesh, etc.)
- **Districts:** 369 unique
- **Markets:** 948 unique

- **Commodities:** 187 unique (Top = Onion with 490 entries)
- **Varieties:** 340 unique (Top = "Other")
- **Grades:** 6 unique (Top = "Local", then "FAQ")
- **Dates:** Only one (`07/08/2025`) → dataset is a snapshot of a single day's mandi prices

📌 Possible Analyses for Project

Here are **all statistical techniques you can apply**:

1. Exploratory Data Analysis (EDA)

- Commodity distribution by **State/District/Market**
- Top traded commodities & price variations
- Boxplots of price ranges per commodity
- Outlier detection (extremely high/low values)
- Correlation between `Min`, `Max`, `Modal` prices

2. Data Cleaning

- Handle anomalies (Min=0, unrealistic Max=120000)
- Standardize categorical features (e.g., "Local" vs "local")
- Possibly drop/encode **Arrival_Date** since it's only one day

3. Regression Analysis

- **Linear Regression:** Predict `Modal Price` from `Min` and `Max`
- **Multiple Regression:** Add categorical variables (commodity, state) using dummy encoding

- **Polynomial Regression:** If price trends are non-linear

4. ANOVA / Hypothesis Testing

- Compare **mean prices across states**
- Check if **variety significantly affects prices**
- Test price difference between **grades (FAQ vs Local vs Standard)**

5. Clustering & Classification

- Cluster commodities based on **price range patterns**
- Classify states/markets based on **average price levels**

6. Time-Series Analysis (Limited)

- Since only **one date** is given, no time-series forecasting is possible
- If historical data is added, you could model **seasonal trends in mandi prices**

7. Advanced Techniques

- **Principal Component Analysis (PCA):** Reduce high-cardinality categorical features (commodity, variety)
- **Outlier Detection:** Using Z-score or IQR on prices
- **Predictive Modeling:** Train models (Linear, Ridge, Random Forest) to estimate price ranges