Exploratory Data Analysis on

MSP WHEAT & VARIETY WISE DAILY MARKET PRICES OF WHEAT, 2024

by

Group 24

Kushal Barot
ID: 202318006
Course: MSc(DS)

Harshil Shah
ID: 202318033
Course: MSc(DS)

Rishi Pawar
ID: 202318037
Course: MSc(DS)

Course Code: IT 462
Semester: Winter 2023

Under the guidance of

Dr. Gopinath Panda

Dhirubhai Ambani Institute of Information and Communication Technology

April 29, 2024

Acknowledgment

I wanted to take a moment to express my sincere appreciation for the exceptional guidance and support you provided me during my project "MSP WHEAT." Your mentorship has been invaluable and played a pivotal role in ensuring the success of this endeavor.

I consider myself very lucky to have had the opportunity to benefit from your expertise and mentorship. Your insightful advice, coupled with extensive knowledge in the field, has significantly influenced the quality and scope of the project. Your constructive feedbacks and suggestions not only helped me navigate challenges but deepened my understanding of the subject.

I would also like to show my gratitude to the entire team at DAIICT for fostering a culture of collaboration and innovation. The resources and facilities provided by the institution have been instrumental in facilitating comprehensive research and analysis, thereby enriching the outcome of the project.

Moreover, I am thankful to my peers and colleagues for their unwavering support and camaraderie throughout this journey. Their contributions have undeniably enhanced the development of the "MSP WHEAT Variety wise daily market prices of Wheat, 2024" project.

Entering into the "MSP WHEAT Variety wise daily market prices of Wheat, 2024" project has been an immensely fulfilling experience for me. I am confident that the knowledge and skills acquired during this endeavor will serve as a solid foundation for my future pursuits.

Once again, I want to express my deepest gratitude for your invaluable guidance and support. Your mentorship was indispensable, and I am genuinely apprecative of the opportunity to learn from you.

Sincerely,

Kushal Barot, 202318006
Harshil Shah, 202318033
Rishi Pawar, 202318037

DECLARATION

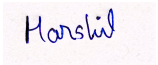
We, [202318006, 202318033, 202318037] now declare that the EDA project work presented in this report is our original work and has not been submitted for any other academic degree. All the sources cited in this report have been appropriately referenced.

We acknowledge that the data utilized in this project has been sourced from farmer.gov.in and data.gov.in . We affirm that we have complied with the terms and conditions specified on the website for accessing and using the dataset. We hereby confirm that the dataset employed in this project is accurate and authentic to the best of our knowledge.

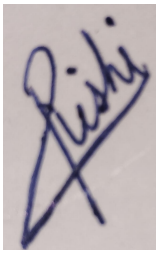
We acknowledge that we have received no external help or assistance in conducting this project except for the guidance provided by our mentor, Prof. Gopinath Panda. We declare no conflict of interest in conducting this EDA project.

We have now signed the declaration statement and confirmed the submission of this report on .... April 2024.

Kushal Barot
ID: 202318006
Course: MSc(DS)



Harshil Shah
ID: 202318033
Course: MSc(DS)



Rishi Pawar
ID: 202318037
Course: MSc(DS)

CERTIFICATE

This is to certify that Group 24, comprising Kushal Barot, Harshil Shah, and Rishi Pawar, has successfully completed an exploratory data analysis (EDA) project on the MSP WHEAT dataset, sourced from the farmer.gov.in site. site. Additionally, the group has conducted analysis on another dataset containing retail prices of wheat for the year 2024, sourced from the data.gov.in site.

The EDA project presented by Group 24 is entirely their original work. It was carried out under the supervision of the course instructor, Prof. Gopinath Panda, who provided continued support and guidance throughout the whole duration of the project. The analysis conducted is based on a comprehensive examination of the 'MSP WHEAT' dataset, as well as the '2024 Wheat Retail Prices' dataset, and the findings presented in the report are derived directly from these datasets.

This certificate is hereby issued to acknowledge the successful culmination of the EDA project on the 'MSP WHEAT' dataset and the supplementary analysis on the '2024 Wheat Retail Prices' dataset. It serves as a testament to the analytical capabilities and knowledge of the students of Group 24 in the field of data analysis.

Signed,
Dr. Gopinath Panda,
IT 462 Course Instructor
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat, INDIA.

April 29, 2024

Contents

List of Figures

For Dataset-1 : MSP WHEAT

- Univariate Analysis

For Dataset-2 : Wheat 2024

- Uni-variate Analysis
    1. Histogram and QQ Plot for **'min_price'**
    2. Histogram and QQ Plot for **'modal_price'**
    3. Count of entries by state
    4. Box plot for **'min_price'**
    5. Box plot for **'max_price'**
    6. Box plot for **'modal_price'**
    7. Top 20 districts by entry count
    8. Top states
- Multivariate Analysis
    1. Number of districts per state
    2. Top 20 districts with the highest number of markets for wheat
    3. Heatmap
    4. Pair plot for wheat prices
    5. State-wise Analysis:
        - Gujarat:
            * Number of markets in each district of Gujarat
            * Top 10 most common wheat varieties in Gujarat
            * Distribution of minimum prices of wheat across markets in Gujarat
            * Distribution of maximum prices of wheat across markets in Gujarat
            * Monthly price of wheat by market (Rajkot)
            * District-wise minimum and modal prices in Gujarat
            * Heatmap
            * Highest and Lowest prices of wheat by variety in Gujarat
        - Uttar Pradesh:
            * Number of markets in each district in Uttar Pradesh
            * Top 10 most common wheat varieties in Uttar Pradesh
            * Distribution of minimum prices of wheat across markets in Uttar Pradesh
            * Distribution of maximum prices of wheat across markets in Uttar Pradesh
            * Monthly price of wheat by market (Muradabad)
            * District-wise minimum and modal prices in Uttar Pradesh
            * Heatmap
            * Highest and Lowest prices of wheat by variety in Utaar Pradesh
        - Madhya Pradesh:
            * Number of markets in each district in Madhya Pradesh

Abstract

This report presents an exploratory data analysis (EDA) of Minimum Support Price (MSP) data for wheat, using two datasets. The first dataset spans from the agricultural years of 2011-12 to 2021-22, offering a comprehensive overview of MSPs for various crops over a decade. The second dataset focuses specifically on wheat and covers the year 2024, providing detailed information on wheat prices across different states, districts, and markets, alongside arrival dates.

Through thorough EDA, we analyzed the trends, distributions, and relationships within each dataset. Exploring the temporal patterns in MSPs for different crops offered insights into the dynamics of agricultural pricing over time. In contrast, the examination of wheat prices in 2024 elucidated the regional variations, market dynamics, and seasonal fluctuations impacting wheat pricing.

Additionally, we investigated potential predictive modeling approaches to forecast wheat prices. Leveraging machine learning algorithms such as linear regression, random forest, and KNN regression, we sought to develop predictive models that could effectively forecast wheat prices based on relevant features such as arrival dates, historical price trends, and potentially market-specific factors.

Overall, this report not only provides valuable insights into the historical trends and current dynamics of wheat pricing but also lays the groundwork for predictive modeling efforts aimed at forecasting future wheat prices, thereby assisting stakeholders in making informed decisions within the agricultural sector.

Chapter 1. Introduction

1.1 Project idea

This is the "MSP WHEAT" dataset, meticulously curated to encompass comprehensive insights into wheat cultivation practices, sourced from farmer.gov .

The project, entitled *"MSP WHEAT": Deep Dive Analysis and Predictive Modeling for Precision Wheat Farming"* aims to conduct an exhaustive examination of agricultural datasets pertinent to wheat cultivation. Leveraging advanced machine learning techniques such as Linear Regression and Random Forest, the project endeavors to construct predictive models to optimize agricultural practices. Through meticulous data preprocessing and comprehensive exploratory analysis, the project seeks to unveil intricate relationships among soil attributes, weather patterns, irrigation schedules, and historical yields. By engineering features and refining predictive algorithms, the objective is to equip farmers with actionable insights tailored to their specific farming contexts. Furthermore, user-friendly interfaces will facilitate real-time decision-making, enabling dynamic adjustments in irrigation, fertilization, and pest management strategies. The project's validation through field trials and iterative refinement ensures its practical relevance and effectiveness in enhancing agricultural productivity and sustainability, thus ushering in a new era of data-driven precision farming in the wheat cultivation domain.

This is the "Wheat 2024" dataset, meticulously compiled to delineate wheat varieties across states, districts, and markets, extracted from data.gov.in .

The *"MSP WHEAT"* project extends its analysis to include the *"Wheat 2024"* dataset, which provides granular insights into wheat cultivation across various states, districts, and markets. Utilizing the same advanced machine learning techniques such as Linear Regression and Random Forest, the project aims to derive predictive models for optimizing agricultural practices at a regional level. Through meticulous data preprocessing and comprehensive exploratory analysis, the project delves into the diverse factors influencing wheat cultivation, including minimum prices of various markets and districts, maximum prices, of different varities. By integrating information on wheat varieties across different markets, the project seeks to tailor predictive models to specific regional contexts, accounting for variations market demand. The development of user-friendly interfaces will enable stakeholders at the state, district, and market levels to access real-time insights and make informed decisions regarding irrigation scheduling, fertilizer application, and market strategies. Through validation and iterative refinement, the project endeavors to enhance agricultural productivity and sustainability on a regional scale, fostering a data-driven approach to precision farming in the wheat cultivation domain.

1.2 Data Collection

1. MSP Dataset:

Source: farmer.gov.in

The MSP dataset was obtained from [Name of Website], a government website that serves as a repository for agricultural information, including Minimum Support Prices (MSPs) for various crops. The website's 'Market Information' section provides historical MSP data spanning multiple agricultural seasons, offering valuable insights into pricing trends and government policies related to crop procurement.

Data Collection Method:
To collect the MSP data, a web scraping approach was employed using the BeautifulSoup library in Python. The web scraping process involved extracting relevant information from the website's HTML structure, including crop names, corresponding MSPs, and the duration of agricultural seasons. The Python script iteratively navigated through the website's pages, accessing and parsing the HTML content to retrieve the desired data. Web scraping challenges, such as handling dynamic content or navigating through paginated results, were addressed through iterative refinement of the scraping script.

2. Wheat Price Dataset:

Source: data.gov.in

The wheat price dataset was sourced from data.gov, a government website that provides access to a wide range of publicly available datasets. The specific dataset used for this analysis was [Name of Dataset], which contains comprehensive information on wheat prices across different states, districts, and markets.

Data Collection Method:
The wheat price dataset was downloaded directly from the data.gov website after identifying the relevant dataset for wheat prices. The dataset was available in a structured format, facilitating ease of access and analysis.

1.3 Dataset Description

Below is the comprehensive data description for dataset-1: MSP WHEAT

- **The dimensions of my dataset are as follows:**
  `print(df.shape)`

```
Shape of the data: (29, 15)
```

- **The following are the initial rows of my dataset:**
  `data.head(5)`

	Category	Commodity	Variety	2010-11	2011-12	2012-13	2013-14	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20	2020-21	2021-22
0	Kharif Crops	PADDY	Common	1000	1080	1250	1310	1360	1410	1470	1550	1750	1815	1868	1940
1	Kharif Crops	PADDY	Grade 'A'	1030	1110	1280	1345	1400	1450	1510	1590	1770	1835	1888	1960
2	Kharif Crops	JOWAR	Hybrid	880	980	1500	1500	1530	1570	1625	1700	2430	2550	2620	2738
3	Kharif Crops	JOWAR	Maldandi	900	1000	1520	1520	1550	1590	1650	1725	2450	2570	2640	2758
4	Kharif Crops	BAJRA	NaN	880	980	1175	1250	1250	1275	1330	1425	1950	2000	2150	2250

- **The data types of columns in the dataset are as follows:**
  `print(data.dtypes)`

```
Category     object
Commodity    object
Variety      object
2010-11       int64
2011-12       int64
2012-13       int64
2013-14       int64
2014-15       int64
2015-16       Int64
2016-17       Int64
2017-18       Int64
2018-19       Int64
2019-20       Int64
2020-21       Int64
2021-22       Int64
dtype: object
```

- **Below are the unique columns present in my dataset:**
  `print("Unique values of each column:") for col in data.columns[:3]: print(col, ":", data[col].unique())`

```
Unique values of each column:
Category : ['Kharif Crops' 'Rabi Crops' 'Others']
Commodity : ['PADDY' 'JOWAR' 'BAJRA' 'MAIZE' 'RAGI' 'Tur (Arhar)' 'MOONG' 'URAD'
 'COTTON' 'Groundnut' 'SUNFLOWER SEED' 'SOYABEAN' 'SESAMUM' 'NIGERSEED'
 'WHEAT' 'BARLEY' 'GRAM' 'MASUR (LENTIL)' 'Rapeseed & Mustard' 'SAFFLOWER'
 'TORIA' 'COPRA' '(Calender Year)' 'DE-HUSKED COCONUT(Calender Year)'
 'JUTE']
Variety : ['Common' "Grade 'A'" 'Hybrid' 'Maldandi' nan 'Medium Staple'
 'Long Staple' 'Black' 'Yellow' '-' 'Milling' 'Ball']
```

- **Below is the information summary for the DataFrame using df.info():**
  `df.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36151 entries, 0 to 36150
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   state         36151 non-null  object
 1   district      36151 non-null  object
 2   market        36151 non-null  object
 3   commodity     36151 non-null  object
 4   variety       36151 non-null  object
 5   arrival_date  36151 non-null  object
 6   min_price     36150 non-null  float64
 7   max_price     36145 non-null  float64
 8   modal_price   36151 non-null  float64
 9   update_date   36151 non-null  object
dtypes: float64(3), object(7)
memory usage: 2.8+ MB
```

- **Below is the summary statistics computed using df.describe():**
  `df.describe()`

	2010-11	2011-12	2012-13	2013-14	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20	2020-21	2021-22
count	29.000000	29.000000	29.000000	29.000000	29.000000	28.0	28.0	28.0	28.0	28.0	28.0	21.0
mean	1984.482759	2271.896552	2760.000000	2875.517241	2919.137931	3107.857143	3369.285714	3607.857143	4153.607143	4466.5	4679.428571	5071.714286
std	1049.917807	1112.927617	1297.229878	1333.321139	1338.547366	1455.678809	1588.242515	1707.8503	1899.437935	2209.791418	2314.710516	2599.633246
min	780.000000	980.000000	980.000000	1100.000000	1150.000000	1225.0	1325.0	1410.0	1440.0	1525.0	1600.0	1870.0
25%	1030.000000	1110.000000	1500.000000	1500.000000	1530.000000	1558.75	1625.0	1732.5	2330.0	2565.0	2635.0	2758.0
50%	1800.000000	2500.000000	2900.000000	3000.000000	3050.000000	3325.0	3800.0	4060.0	4337.5	4612.5	4875.0	5550.0
75%	2500.000000	2900.000000	3700.000000	4000.000000	4000.000000	4047.5	4230.0	4575.0	5487.5	5662.5	5913.75	6300.0
max	4700.000000	4775.000000	5350.000000	5500.000000	5500.000000	5830.0	6240.0	6785.0	7750.0	9920.0	10300.0	10600.0

Below is the comprehensive data description for dataset-2: WHEAT-2024

- **The dimensions of my dataset are as follows:**

```
print(data.shape)
```

```
(36151, 10)
```

- **The following are the initial rows of my dataset:**

```
data.head(5)
```

	state	district	market	commodity	variety	arrival_date	min_price	max_price	modal_price	update_date
36146	West Bengal	Uttar Dinajpur	Raiganj	Wheat	Local	26/03/2024	2500.0	2700.0	2600.0	2024-04-01
36147	West Bengal	Uttar Dinajpur	Raiganj	Wheat	Local	27/03/2024	2500.0	2700.0	2600.0	2024-04-01
36148	West Bengal	Uttar Dinajpur	Raiganj	Wheat	Local	28/03/2024	2500.0	2700.0	2600.0	2024-04-01
36149	West Bengal	Uttar Dinajpur	Raiganj	Wheat	Local	29/03/2024	2500.0	2700.0	2600.0	2024-04-01
36150	West Bengal	Uttar Dinajpur	Raiganj	Wheat	Local	30/03/2024	3900.0	4100.0	4000.0	2024-04-01

- **The data types of columns in the dataset are as follows:**

```
print(data.dtypes)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36151 entries, 0 to 36150
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   state         36151 non-null  object
 1   district      36151 non-null  object
 2   market        36151 non-null  object
 3   commodity     36151 non-null  object
 4   variety       36151 non-null  object
 5   arrival_date  36151 non-null  object
 6   min_price     36150 non-null  float64
 7   max_price     36145 non-null  float64
 8   modal_price   36151 non-null  float64
 9   update_date   36151 non-null  object
dtypes: float64(3), object(7)
memory usage: 2.8+ MB
```

- **Below are the unique columns present in my dataset:**

```
print('state:', df1['state'].nunique())
print('district:', df1['district'].nunique())
print('market:', df1['market'].nunique())
print('commodity:', df1['commodity'].nunique())
print('variety:', df1['variety'].nunique())
print('arrival_date:', df1['arrival_date'].nunique())
print('update_date:', df1['update_date'].nunique())
```

```
state: 13
district: 241
market: 883
commodity: 1
variety: 37
arrival_date: 92
update_date: 1
```

- **Below is the summary statistics computed using df.describe():**
  `df1.describe()`

	min_price	max_price	modal_price
count	36150.000000	36145.000000	36151.000000
mean	2334.266113	2633.078102	2471.082840
std	259.516434	444.109022	234.004065
min	2.000000	55.000000	1150.000000
25%	2200.000000	2423.000000	2350.000000
50%	2350.000000	2550.000000	2450.000000
75%	2450.000000	2720.000000	2550.000000
max	4860.000000	27100.000000	4550.000000

1.4 Packages required

1. **NumPy:** Considered a best library for scientific computations in Python, NumPy enables efficient handling of large arrays and matrices. It offers a big range of mathematical functions tailored for array operations, facilitating numerical computing and data manipulation tasks with its slicing and indexing capabilities.

2. **Pandas:** Renowned for its data manipulation and analysis, Pandas provides intuitive structures and functions for working with structured data. Through its Series and DataFrame objects, Pandas simplifies data cleaning, manipulation, and exploratory analysis, serving as a cornerstone for data preprocessing and exploratory data analysis (EDA).

3. **Matplotlib:** Serving as a versatile plotting library, Matplotlib facilitates the creation of static, interactive, and animated visualizations in Python. With its flexible interface, Matplotlib supports a myriad of plot types, making it suitable for both basic and complex visualization needs.

4. **Statsmodels:** A Python library dedicated to statistical modeling and testing, Statsmodels offers tools for exploratory data analysis, hypothesis testing, and regression analysis. Supporting various statistical models like linear regression and time series analysis, it aids in deriving insights from data.

5. **Plotly Express:** This high-level interface gives the creation of interactive plots using the Plotly library. Plotly Express offers an intuitive syntax for generating complex visualizations with minimal code, making it ideal for exploratory data analysis and visualization tasks.

6. **Plotly Graph Objects:** Providing a lower-level interface, Plotly Graph Objects grant users better control over plot customization and layout. It enables the creation of highly tailored visualizations and dashboards to meet specific analytical requirements.

7. **Dash:** Dash is a Python framework designed for making web applications, empowering users to create interactive dashboards and data visualization tools. With its declarative syntax, Dash easily provides the development of complex, data-driven applications.

8. **Seaborn:** Built on top of best libraries like Matplotlib, Seaborn offers enhanced aesthetics and ease of use for creating informative plots. It provides high-level functions for generating sophisticated statistical visualizations with minimal coding effort.

9. **Linear Regression:** Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the dependent variable and the independent variable(s). The model equation for simple linear regression can be represented as:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Where:

- Y is the dependent variable.
- X is the independent variable.

- β_0 is the intercept.
- β_1 is the slope coefficient.
- ϵ is the error term.

Linear regression aims to minimize the sum of squared differences between the observed and predicted values. It is widely used for prediction and forecasting tasks in various fields, including economics, finance, and social sciences.

10. **Random Forest:** Random Forest is a popular ensemble learning method used for both classification and regression tasks. It works by constructing a multitude of decision trees during training and outputs the mean prediction (regression) or the mode of the predictions (classification) of the individual trees. Random Forest introduces randomness in two key ways:

    - Random sampling of the training data: Each tree is trained on a random subset of the training data.
    - Random feature selection: At each node of the decision tree, a random subset of features is considered for splitting.

    Random Forest is robust to overfitting, performs well with high-dimensional data, and provides a measure of feature importance. It is widely used in various applications such as finance, healthcare, and ecology.

11. **KNN Regression:** K-Nearest Neighbors (KNN) Regression is a non-parametric method used for regression tasks. It predicts the output value for a new data point by averaging the output values of its k nearest neighbors in the feature space. The predicted value is the average (for regression) of the k-nearest neighbors' target values. The choice of k is a hyperparameter that needs to be tuned, and it significantly impacts the model's performance.

    KNN Regression is simple to understand and implement, but it can be computationally expensive, especially with large datasets, as it requires storing and searching through the entire training dataset during prediction. It is suitable for small to medium-sized datasets and can be effective when there is a smooth relationship between features and target variables.

12. **Standard Scaler:** StandardScaler is a preprocessing technique used to standardize features by removing the mean and scaling them to unit variance. It transforms the data such that it has a mean of 0 and a standard deviation of 1. This preprocessing step is crucial, especially for algorithms that are sensitive to feature scaling, such as linear models, SVMs, and KNN.

    The formula for standardization is:

    $$z = \frac{x - \mu}{\sigma}$$

    Where:

    - z is the standardized value.
    - x is the original value.
    - μ is the mean of the feature.
    - σ is the standard deviation of the feature.

Chapter 2. Data Cleaning

Data refinement is an essential process in data analysis that involves identifying and rectifying errors, inconsistencies, and missing values within a dataset. This step is fundamental to maintaining the integrity, accuracy, and reliability of the data before proceeding with analysis or modeling tasks. Several key reasons highlight the necessity of this process:

Ensuring Data Integrity: By removing errors, duplicates, and inconsistencies, data refinement ensures that the dataset accurately reflects the true state of the underlying phenomena, preventing misleading conclusions or erroneous interpretations based on faulty data.

Enhancing Analysis Quality: High-quality, clean data leads to more reliable analysis results. This process helps eliminate noise and outliers, enabling analysts to focus on meaningful patterns and insights within the data. Clean data reduces the risk of bias and ensures that analysis outcomes are representative and actionable.

Improving Model Performance: For machine learning and statistical modeling tasks, the quality of the input data significantly impacts model performance. Clean data improves the accuracy and generalizability of models by reducing the likelihood of overfitting or underfitting. By providing a more accurate representation of underlying relationships, data refinement helps produce robust and reliable models.

Facilitating Data Integration: In scenarios involving data from multiple sources or systems, data refinement becomes crucial for data integration. Standardizing formats, resolving inconsistencies, and handling missing values ensure seamless integration of disparate datasets. This allows organizations to leverage their data assets fully and derive comprehensive insights for informed decision-making.

The data refinement process typically involves several key steps:

Identifying Missing Values: Through imputation or removal, missing values are handled to ensure data completeness and reliability. Removing Duplicates: Duplicate records are eliminated to maintain data consistency and accuracy.

Standardizing Data Formats: Uniformity in data formats facilitates accurate analysis across the dataset.

Correction of Errors and Inconsistencies: Typographical errors, inconsistencies in naming conventions, or inaccuracies are rectified to enhance data quality.

2.1 Missing data analysis

- *Moving forward, we will conduct a missing data analysis for the dataset-1, 'MSP - WHEAT'.*

```
# Check for missing values including <NA>
missing_values = data.isna().sum()
print("Missing values:")
print(missing_values)
```

```
Missing values:
Category      0
Commodity     0
Variety      11
2010-11       0
2011-12       0
2012-13       0
2013-14       0
2014-15       0
2015-16       1
2016-17       1
2017-18       1
2018-19       1
2019-20       1
2020-21       1
2021-22       8
```

The dataset exhibits varying counts of missing values across its columns. Notably, while *'Category'* and *'Commodity'* columns show no missing data, the *'Variety'* column records 11 missing values, signifying a lack of information on specific commodity varieties. Across the years from *2010-11* to *2021-22*, there are sporadic occurrences of missing values, with one missing value each observed in the years *2015-16, 2016-17, 2017-18, 2018-19*, and *2019-20*. Moreover, the year *2021-22* records the highest count of missing values, totaling 8. This distribution underscores the importance of addressing missing values meticulously to uphold data integrity and facilitate robust analysis and modeling endeavors.

- *Moving forward, we will conduct a missing data analysis for the dataset-2, 'WHEAT - 2024'.*

```
df1.isnull().sum()
```

```
state           0
district        0
market          0
commodity       0
variety         0
arrival_date    0
min_price       1
max_price       6
modal_price     0
update_date     0
dtype: int64
```

Missing Values in min_price and max_price: With one missing value in the min_price column and six in max_price, it becomes challenging to ascertain the precise missing data mechanism. Nonetheless, if the absence occurs randomly or sporadically, it might fall under the classification of Missing Completely At Random (MCAR). This indicates that the likelihood of a missing value in min_price remains consistent across all observations and is independent of the observed or unobserved data.

2.2 Outlier Detection

For Dataset-2 : Wheat-2024 ,

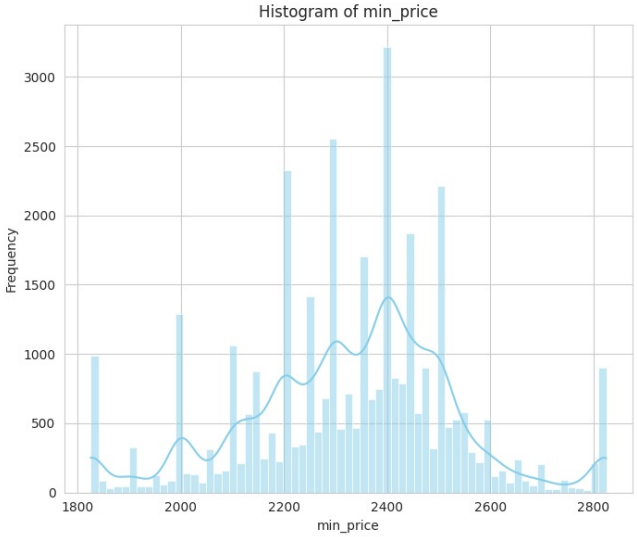
- **Methodology :** The Interquartile Range (IQR) is a measure of statistical dispersion, representing the range between the first quartile (Q1) and the third quartile (Q3) of the data distribution. Outliers are typically defined as data points lying below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR.

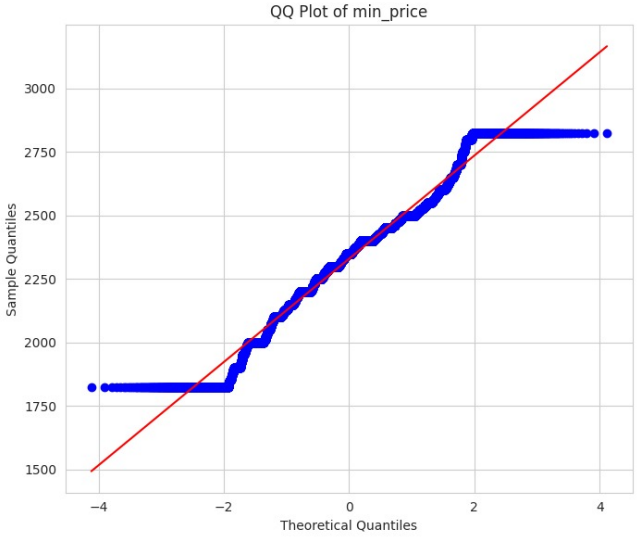
- **Outliers :** min_price lower bound -1825.0
  min_price uper bound-2825.0
  max_price lower bound -1977.5
  max_price uper bound-3165.5
  modal_price lower bound -2050.0
  modal_price uper bound-2850.0

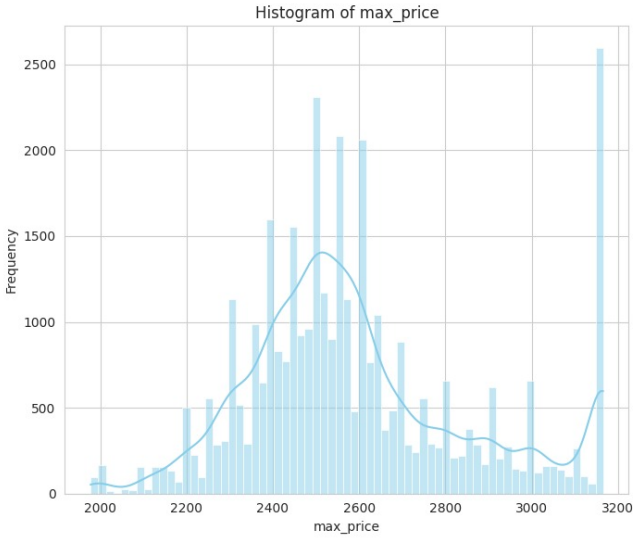
- **Outlier Treatment**
  - **Capping Methodology**
    * Calculation of Interquartile Range (IQR):
      · For each numerical column in the dataset, we calculated the first quartile (Q1) and third quartile (Q3) to determine the IQR.
    * Threshold Calculation:
      · Using the IQR, we computed the lower and upper bounds by subtracting and adding 1.5 times the IQR from Q1 and Q3, respectively.
    * Capping Procedure:
      · Outliers falling below the lower bound were replaced with the nearest value within the lower bound.
      · Outliers exceeding the upper bound were replaced with the nearest value within the upper bound.

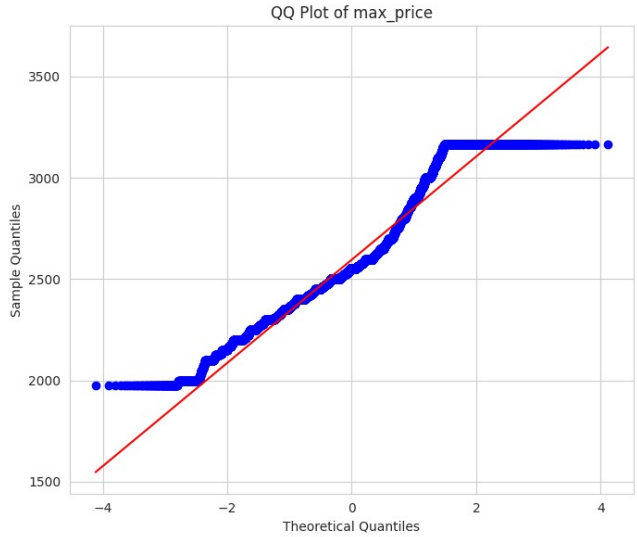
After Capping : After capping outliers, the ranges of values in the dataset have been adjusted to minimize the impact of extreme values on the analysis:

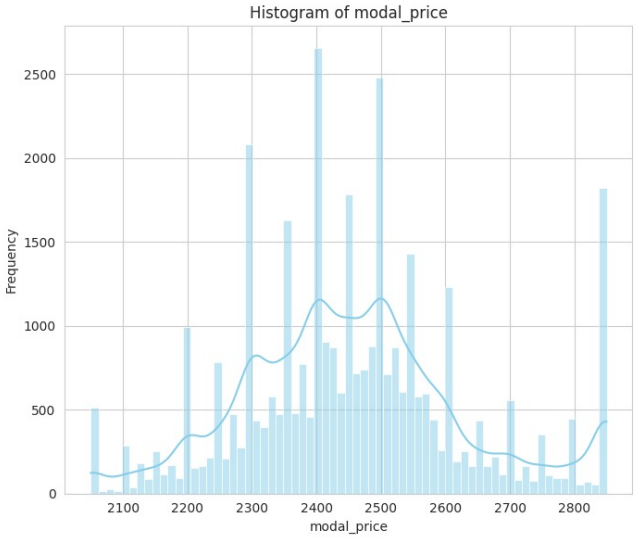
Min_price Range: The minimum prices now range from 1825.0 to 2825.0, reflecting a narrower spread of prices compared to the original dataset. This adjustment helps mitigate the influence of exceptionally low or high prices that could distort the interpretation of price trends.
Max_price Range: The maximum prices have been constrained within a range of 1977.5 to 3165.5 after capping outliers. This narrower range ensures that unusually high prices, which may skew the average or median values, are appropriately managed to provide a more representative picture of market conditions.
Modal_price Range: The modal prices, representing the most frequently occurring prices, now vary between 2050.0 and 2850.0. By capping outliers, the modal price range is better aligned with the central tendency of the dataset, offering insights into typical price levels experienced by market participants.

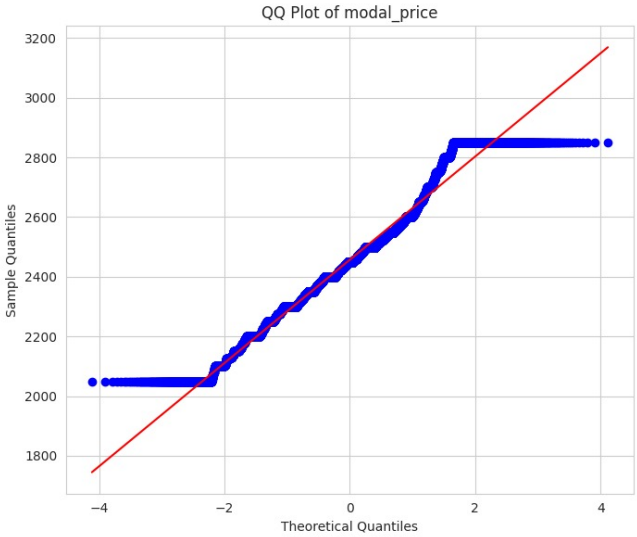

Histogram of min_price
Frequency
min_price
0
500
1000
1500
2000
2500
3000
1800
2000
2200
2400
2600
2800


QQ Plot of min_price
Sample Quantiles
Theoretical Quantiles
1500
1750
2000
2250
2500
2750
3000
−4
−2
0
2
4


Histogram of max_price
Frequency
max_price
0
500
1000
1500
2000
2500
2000
2200
2400
2600
2800
3000
3200


QQ Plot of max_price
Sample Quantiles
Theoretical Quantiles
1500
2000
2500
3000
3500
−4
−2
0
2
4


Histogram of modal_price
Frequency
modal_price
0
500
1000
1500
2000
2500
2100
2200
2300
2400
2500
2600
2700
2800


QQ Plot of modal_price
Sample Quantiles
Theoretical Quantiles
1800
2000
2200
2400
2600
2800
3000
3200
−4
−2
0
2
4

2.3 Imputation

Performing imputation on dataset-1, MSP Wheat, to address missing values :

```
#Dropping SOYABEAN
data=data[data['Commodity']!='SOYABEAN Black']
# Interpolate missing values in the columns from 2015 to 2022 using
linear interpolation
data.iloc[:, 2:] = data.iloc[:, 2:].interpolate(method='linear', axis=1)
```

	Category	Commodity	2010-11	2011-12	2012-13	2013-14	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20	2020-21	2021-22
0	Kharif Crops	PADDY Common	1000.0	1080.0	1250.0	1310.0	1360.0	1410.0	1470.0	1550.0	1750.0	1815.0	1868.0	1940.0
1	Kharif Crops	PADDY Grade 'A'	1030.0	1110.0	1280.0	1345.0	1400.0	1450.0	1510.0	1590.0	1770.0	1835.0	1888.0	1960.0
2	Kharif Crops	JOWAR Hybrid	880.0	980.0	1500.0	1500.0	1530.0	1570.0	1625.0	1700.0	2430.0	2550.0	2620.0	2738.0
3	Kharif Crops	JOWAR Maldandi	900.0	1000.0	1520.0	1520.0	1550.0	1590.0	1650.0	1725.0	2450.0	2570.0	2640.0	2758.0
4	Kharif Crops	BAJRA -	880.0	980.0	1175.0	1250.0	1250.0	1275.0	1330.0	1425.0	1950.0	2000.0	2150.0	2250.0
5	Kharif Crops	MAIZE -	880.0	980.0	1175.0	1310.0	1310.0	1325.0	1365.0	1425.0	1700.0	1760.0	1850.0	1870.0
6	Kharif Crops	RAGI -	965.0	1050.0	1500.0	1500.0	1550.0	1650.0	1725.0	1900.0	2897.0	3150.0	3295.0	3377.0
7	Kharif Crops	Tur (Arhar) -	3000.0	3200.0	3850.0	4300.0	4350.0	4825.0	5250.0	5650.0	5675.0	5800.0	6000.0	6300.0
8	Kharif Crops	MOONG -	3170.0	3500.0	4400.0	4500.0	4600.0	5050.0	5425.0	5775.0	6975.0	7050.0	7196.0	7275.0
9	Kharif Crops	URAD -	2900.0	3300.0	4300.0	4300.0	4350.0	4825.0	5200.0	5600.0	5600.0	5700.0	6000.0	6300.0
10	Kharif Crops	COTTON Medium Staple	2500.0	2800.0	3600.0	3700.0	3750.0	3800.0	3860.0	4020.0	5150.0	5255.0	5515.0	5726.0
11	Kharif Crops	COTTON Long Staple	3000.0	3300.0	3900.0	4000.0	4050.0	4100.0	4160.0	4320.0	5450.0	5550.0	5825.0	6025.0
12	Kharif Crops	Groundnut -	2300.0	2700.0	3700.0	4000.0	4000.0	4030.0	4320.0	4650.0	4890.0	5090.0	5275.0	5550.0
13	Kharif Crops	SUNFLOWER SEED -	2350.0	2800.0	3700.0	3700.0	3750.0	3800.0	4050.0	4200.0	5388.0	5650.0	5885.0	6015.0
15	Kharif Crops	SOYABEAN Yellow	1440.0	1690.0	2240.0	2560.0	2560.0	2600.0	2875.0	3250.0	3399.0	3710.0	3880.0	3950.0
16	Kharif Crops	SESAMUM -	2900.0	3400.0	4200.0	4500.0	4600.0	4700.0	5200.0	5400.0	6249.0	6485.0	6855.0	7307.0
17	Kharif Crops	NIGERSEED -	2450.0	2900.0	3500.0	3500.0	3600.0	3650.0	3925.0	4150.0	5877.0	5940.0	6695.0	6930.0
18	Rabi Crops	WHEAT -	1120.0	1285.0	1350.0	1400.0	1450.0	1525.0	1625.0	1735.0	1840.0	1925.0	1975.0	1975.0
19	Rabi Crops	BARLEY -	780.0	980.0	980.0	1100.0	1150.0	1225.0	1325.0	1410.0	1440.0	1525.0	1600.0	1600.0
20	Rabi Crops	GRAM -	2100.0	2800.0	3000.0	3100.0	3175.0	3600.0	4200.0	4550.0	4620.0	4875.0	5100.0	5100.0
21	Rabi Crops	MASUR (LENTIL) -	2250.0	2800.0	2900.0	2950.0	3075.0	3500.0	4100.0	4350.0	4475.0	4800.0	5100.0	5100.0
22	Rabi Crops	Rapeseed & Mustard -	1850.0	2500.0	3000.0	3050.0	3100.0	3350.0	3800.0	4100.0	4200.0	4425.0	4650.0	4650.0
23	Rabi Crops	SAFFLOWER -	1800.0	2500.0	2800.0	3000.0	3050.0	3300.0	3800.0	4100.0	4945.0	5215.0	5327.0	5327.0
24	Rabi Crops	TORIA -	1780.0	2425.0	2970.0	3020.0	3020.0	3290.0	3560.0	3900.0	4190.0	4425.0	4650.0	4650.0
25	Others	COPRA Milling	4450.0	4525.0	5100.0	5250.0	5250.0	5550.0	5950.0	6500.0	7511.0	9521.0	9960.0	10335.0
26	Others	(Calender Year) Ball	4700.0	4775.0	5350.0	5500.0	5500.0	5830.0	6240.0	6785.0	7750.0	9920.0	10300.0	10600.0
27	Others	DE-HUSKED COCONUT(Calender Year) -	1200.0	1200.0	1400.0	1425.0	1425.0	1500.0	1600.0	1760.0	2030.0	2571.0	2700.0	2800.0
28	Others	JUTE -	1575.0	1675.0	2200.0	2300.0	2400.0	2700.0	3200.0	3500.0	3700.0	3950.0	4225.0	4500.0

The code segment focuses on data preprocessing tasks aimed at improving the quality and completeness of the dataset. Here's a breakdown of the code:

- Filtering Outliers: The code filters out data associated with the 'SOYABEAN Black' commodity from the dataset. This step is crucial for refining the dataset by eliminating any irrelevant or inconsistent data points that could skew analysis results or introduce biases.
- Interpolation of Missing Values: After removing outliers, the code performs interpolation to fill missing values in the columns from 2015 to 2022. Linear interpolation is used for this purpose, implemented through the interpolate() function with the 'linear' method. This method estimates the missing values based on the surrounding data points along the specified axis.

Performing imputation on dataset-2, Wheat 2024, to address missing values :

We perform imputation for missing values in the 'min_price' and 'max_price' columns of the dataset.

```
# Calculate the mean of the non-missing values in the min_price column
mean_min_price = df1['min_price'].mean()

# Fill missing values in the min_price column with the mean
df1['min_price'].fillna(mean_min_price, inplace=True)

# Calculate the mean of the non-missing values in the max_price column
mean_max_price = df1['max_price'].mean()

# Fill missing values in the max_price column with the mean
df1['max_price'].fillna(mean_max_price, inplace=True)
```

Let's break down the process:

- Calculate Mean for 'min_price' Column: The code calculates the mean of the non-missing values in the 'min_price' column using the mean() function. This mean value represents the average price observed in the dataset.
- Fill Missing Values in 'min_price' Column: After calculating the mean, the code fills the missing values in the 'min_price' column with this calculated mean using the fillna() function with the inplace=True parameter. This ensures that missing values are replaced directly within the DataFrame.
- Calculate Mean for 'max_price' Column: Similarly, the code calculates the mean of the non-missing values in the 'max-rice' column using the mean() function.
- Fill Missing Values in 'max_price' Column: Following the calculation of the mean, the code fills the missing values in the 'max_price' column with this calculated mean using the fillna() function, also with inplace=True.

```
state           0
district        0
market          0
commodity       0
variety         0
arrival_date    0
min_price       0
max_price       0
modal_price     0
update_date     0
dtype: int64
```

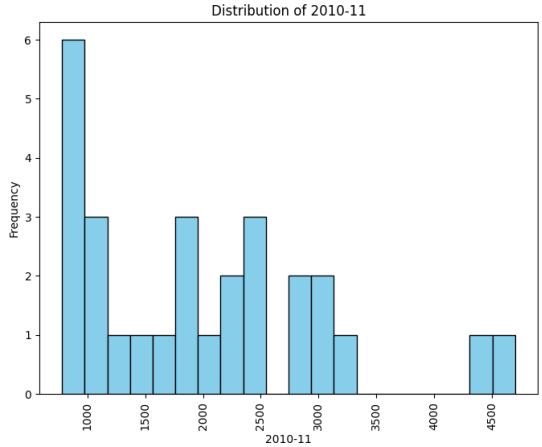
Chapter 3. Visualization

Visualization serves as a crucial element in data exploration and analysis, offering a lucid and intuitive depiction of intricate datasets. Through visualizations, complex data structures become more accessible, enabling the identification of patterns, trends, and relationships within the data, thereby fostering deeper insights and informed decision-making. Here are some significant aspects of visualization:

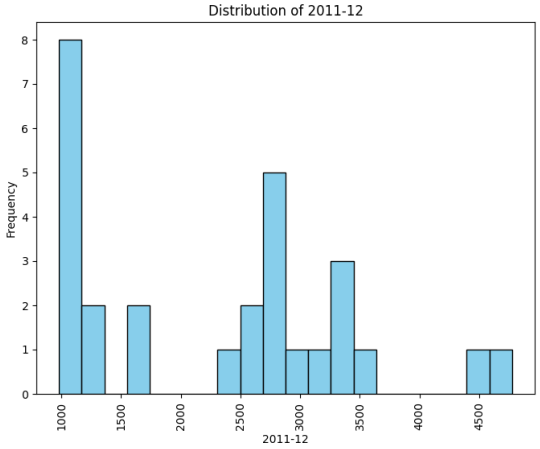
- **Data Exploration:** Visualizations are potent tools for delving into the underlying structure of data. Techniques such as scatter plots, histograms, and box plots empower analysts to probe the distribution, variability, and interrelations between variables, uncovering concealed patterns and anomalies.
- **Communication of Findings:** Visualizations streamline the communication of analytical results and discoveries to stakeholders. Dashboards, charts, and infographics offer visual synopses of critical insights, allowing stakeholders to swiftly grasp complex information and make data-driven decisions.
- **Insight Generation:** By visualizing data through diverse lenses, analysts can glean novel insights and formulate hypotheses that might elude observation from raw data alone. Interactive visualizations, heatmaps, and geographic plots facilitate dynamic exploration of data, encouraging a deeper comprehension of underlying trends and correlations.
- **Identification of Outliers and Anomalies:** Visualizations are instrumental in pinpointing outliers, anomalies, and irregular patterns within datasets. Techniques like scatter plots featuring highlighted outliers or time series plots with trend lines aid in discerning deviations from anticipated behavior, enabling prompt corrective measures.
- **Storytelling and Narrative Building:** Visualizations empower the crafting of data-driven narratives and stories. Through the amalgamation of visual elements with narrative text and annotations, analysts can construct compelling narratives that convey intricate information in a coherent and engaging manner.

3.1 Univariate analysis

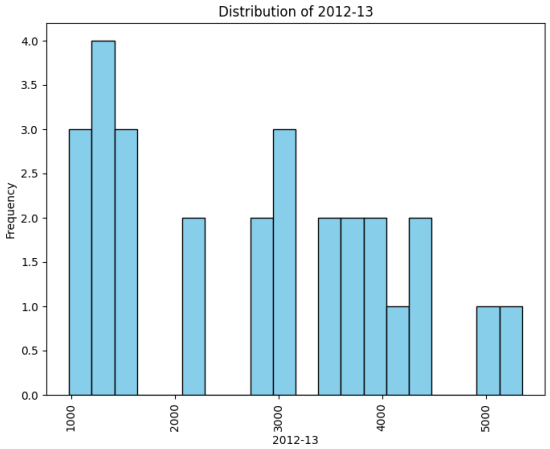
Now, we will conduct univariate analysis on dataset-1



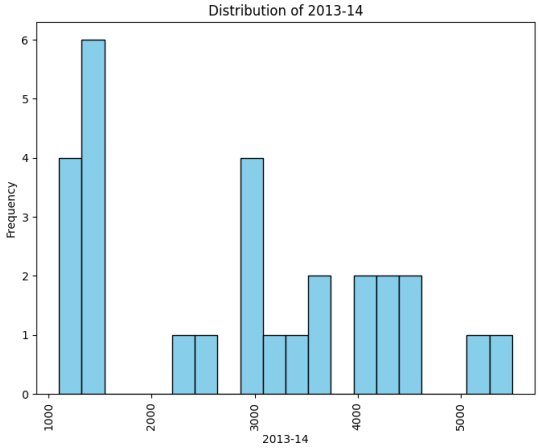

```
Summary Statistics for 2010-11:
count      28.000000
mean     2005.357143
std      1063.038037
min       780.000000
25%      1022.500000
50%      1825.000000
75%      2600.000000
max      4700.000000
Name: 2010-11, dtype: float64
```



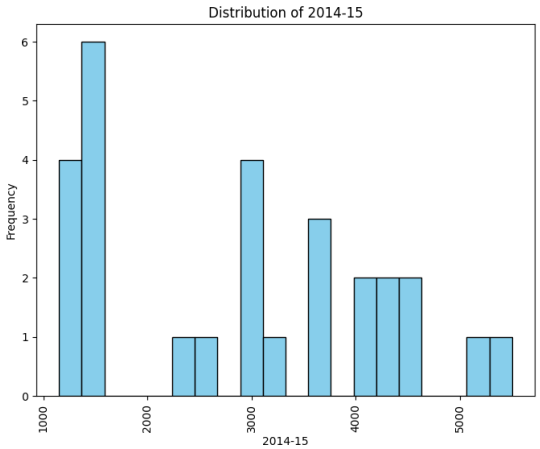

```
Summary Statistics for 2011-12:
count      28.000000
mean     2294.107143
std      1126.785869
min       980.000000
25%      1102.500000
50%      2500.000000
75%      2975.000000
max      4775.000000
Name: 2011-12, dtype: float64
```



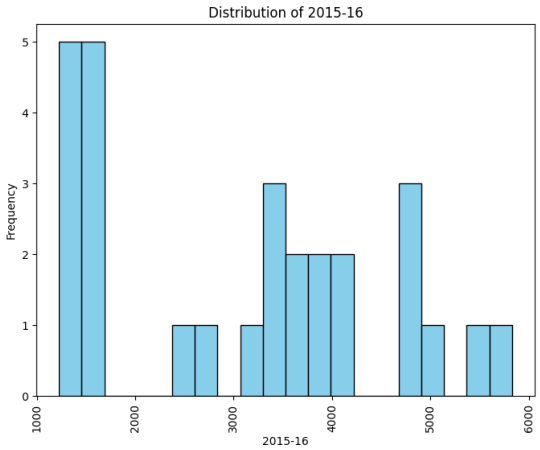

```
Summary Statistics for 2012-13:
count      28.000000
mean     2780.000000
std      1316.473263
min       980.000000
25%      1475.000000
50%      2935.000000
75%      3737.500000
max      5350.000000
Name: 2012-13, dtype: float64
```



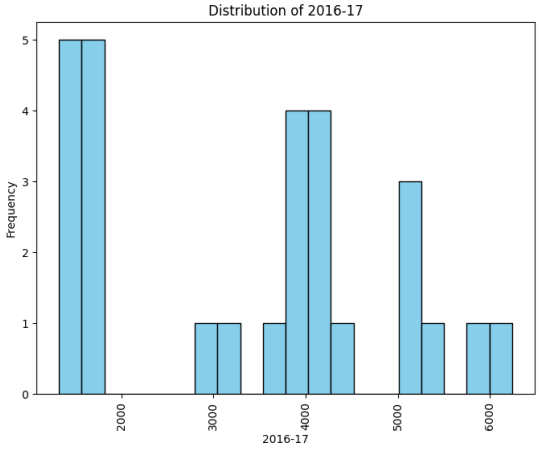

```
Summary Statistics for 2013-14:
count      28.000000
mean     2888.928571
std      1355.794394
min      1100.000000
25%      1481.250000
50%      3010.000000
75%      4000.000000
max      5500.000000
Name: 2013-14, dtype: float64
```



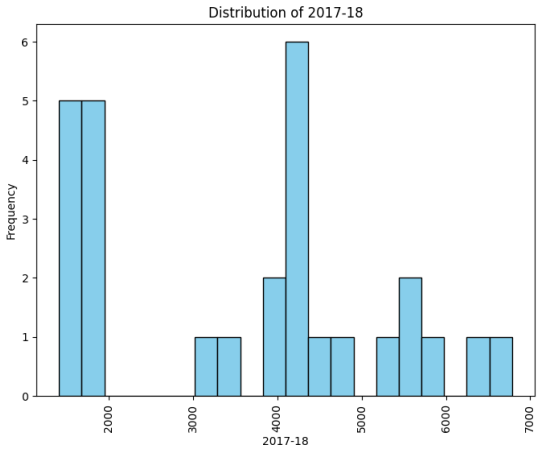

```
Summary Statistics for 2014-15:
count      28.000000
mean     2934.107143
std      1360.635781
min      1150.000000
25%      1510.000000
50%      3062.500000
75%      4012.500000
max      5500.000000
Name: 2014-15, dtype: float64
```



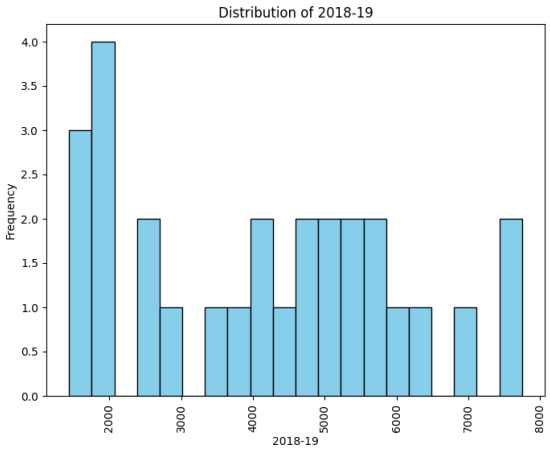

```
Summary Statistics for 2015-16:
count      28.000000
mean     3107.857143
std      1455.678809
min      1225.000000
25%      1558.750000
50%      3325.000000
75%      4047.500000
max      5830.000000
Name: 2015-16, dtype: float64
```



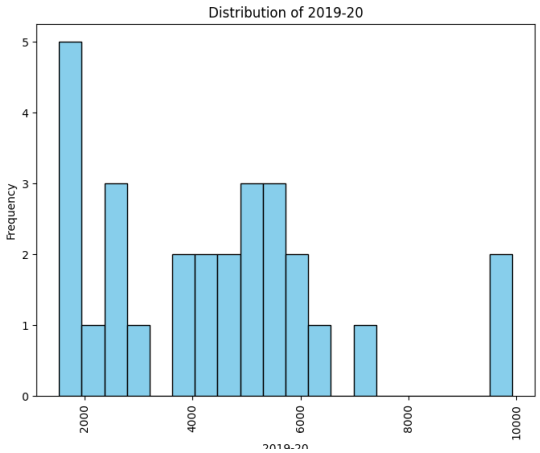

```
Summary Statistics for 2016-17:
count      28.000000
mean     3369.285714
std      1588.242515
min      1325.000000
25%      1625.000000
50%      3800.000000
75%      4230.000000
max      6240.000000
Name: 2016-17, dtype: float64
```



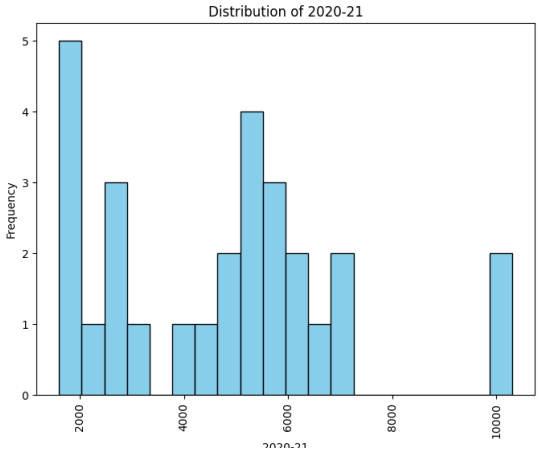

```
Summary Statistics for 2017-18:
count      28.000000
mean     3607.857143
std      1707.850300
min      1410.000000
25%      1732.500000
50%      4060.000000
75%      4575.000000
max      6785.000000
Name: 2017-18, dtype: float64
```




```
Summary Statistics for 2018-19:
count      28.000000
mean     4153.607143
std      1899.437935
min      1440.000000
25%      2330.000000
50%      4337.500000
75%      5487.500000
max      7750.000000
Name: 2018-19, dtype: float64
```



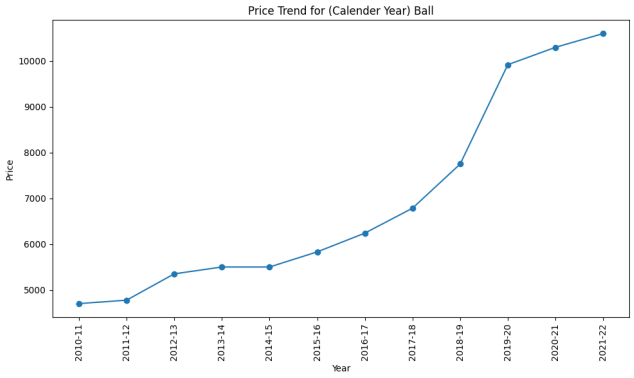

```
Summary Statistics for 2019-20:
count       28.000000
mean     4466.500000
std      2209.791418
min      1525.000000
25%      2565.000000
50%      4612.500000
75%      5662.500000
max      9920.000000
Name: 2019-20, dtype: float64
```

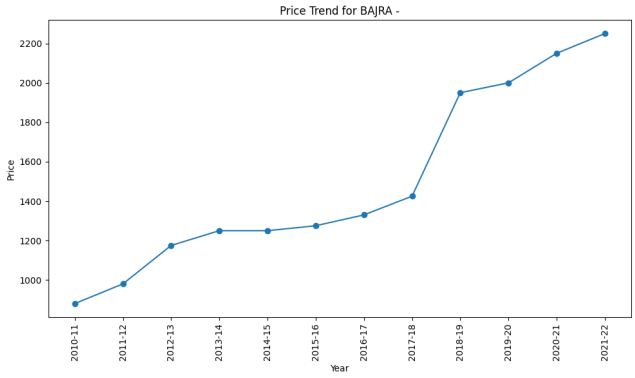


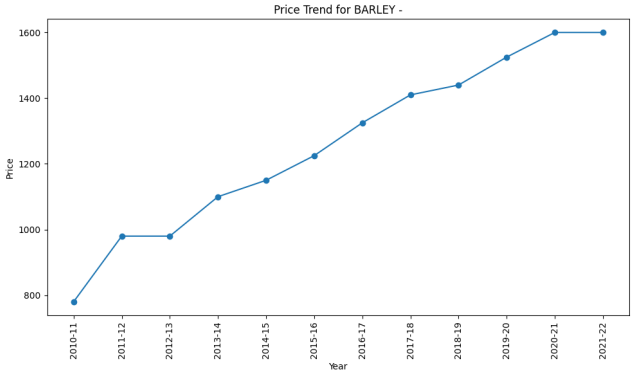

```
Summary Statistics for 2020-21:
count       28.000000
mean      4679.428571
std       2314.710516
min       1600.000000
25%       2635.000000
50%       4875.000000
75%       5913.750000
max      10300.000000
Name: 2020-21, dtype: float64
```

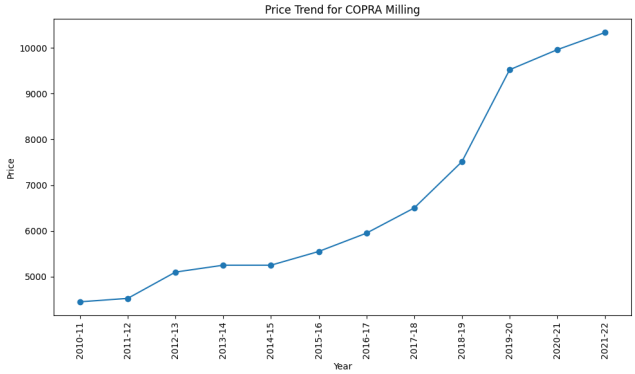
- In the past decade (from 2010-11 to 2020-21), agricultural product prices have demonstrated a gradual upward trend, albeit with fluctuations. These fluctuations are influenced by factors including demand, supply, weather conditions, and economic trends. Different crops display diverse patterns, with some maintaining consistently high prices while others undergo more frequent fluctuations. In recent years (from 2018-19 to 2020-21), prices have shown a slight increase on average and have become more unstable, suggesting a potential shift in market dynamics or environmental conditions.

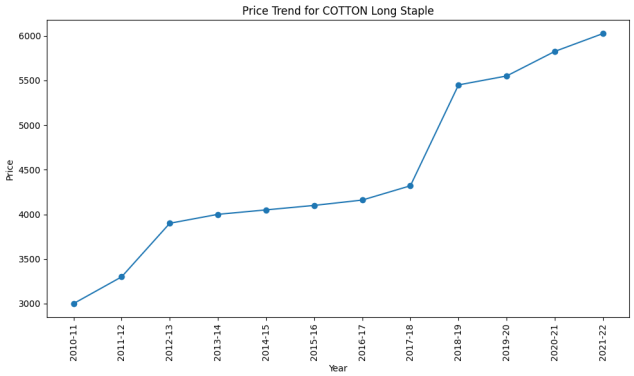
Line charts :

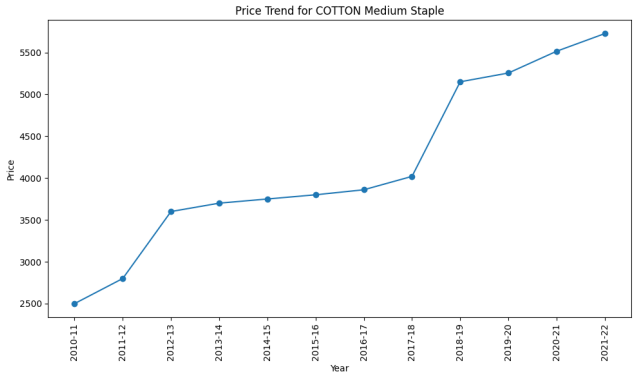

Price Trend for (Calender Year) Ball
Price
Year

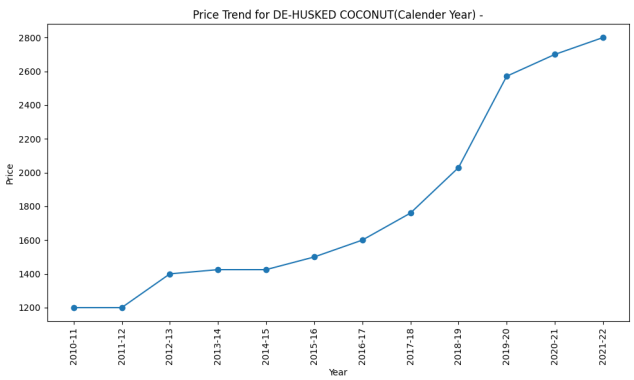

Price Trend for BAJRA -
Price
Year

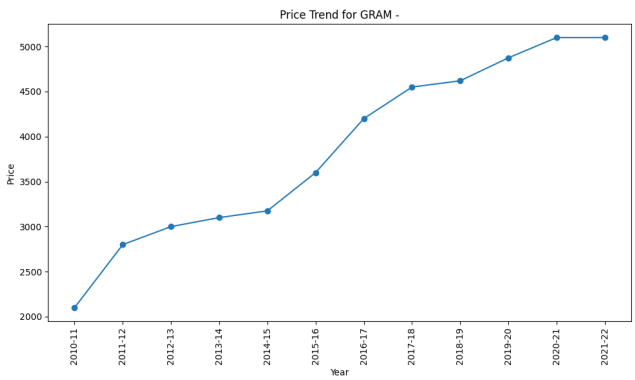

Price Trend for BARLEY -
Price
Year

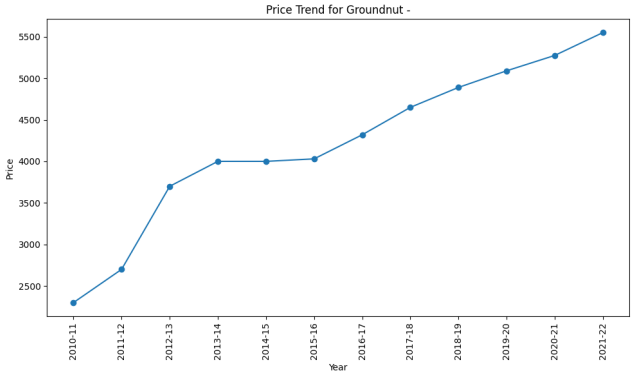

Price Trend for COPRA Milling
Price
Year

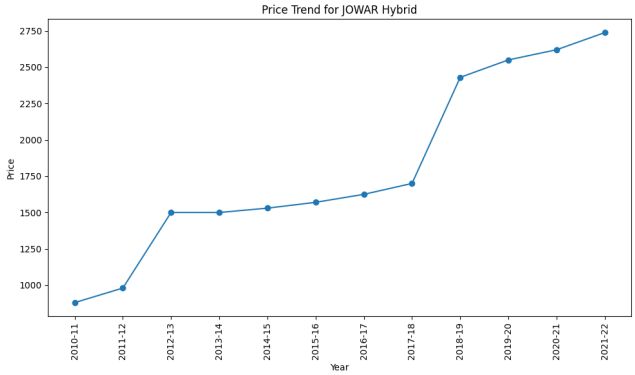

Price Trend for COTTON Long Staple
Price
Year

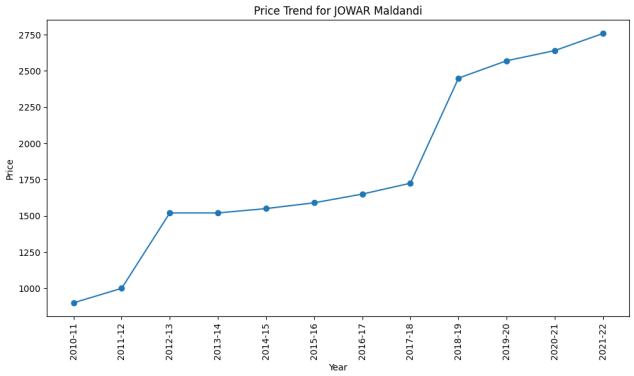

Price Trend for COTTON Medium Staple
Price
Year

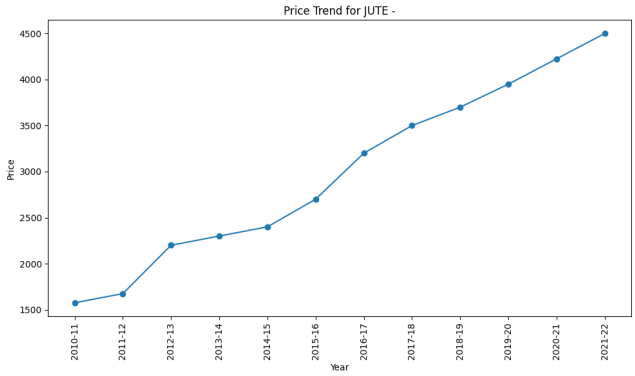

Price Trend for DE-HUSKED COCONUT(Calender Year) -
Price
Year

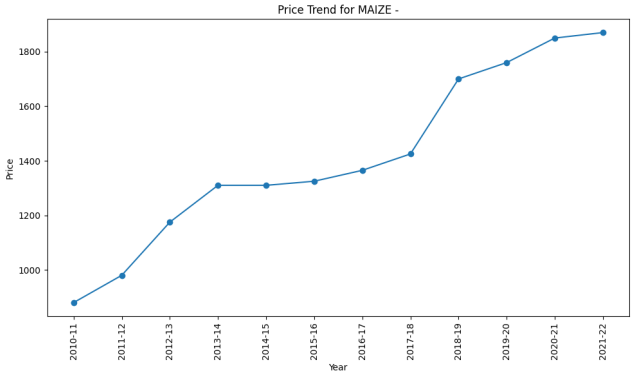

Price Trend for GRAM -
Price
Year

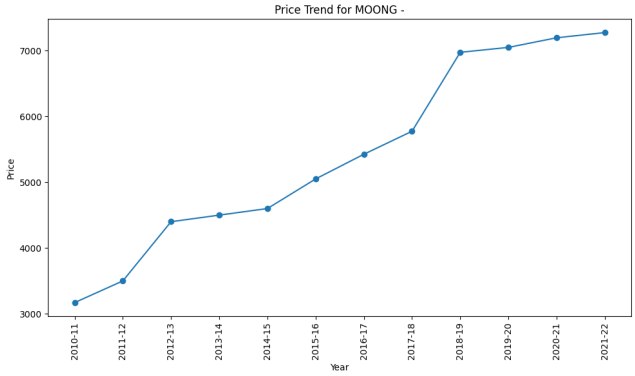

Price Trend for Groundnut -
Price
Year
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22

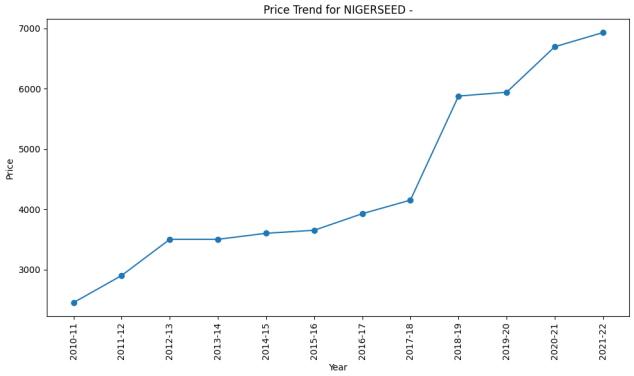

Price Trend for JOWAR Hybrid
Price
Year
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22

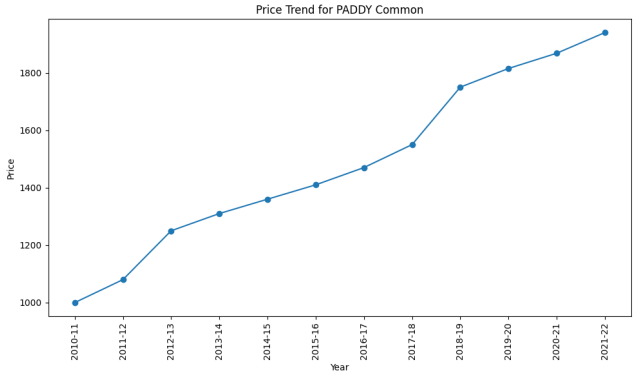

Price Trend for JOWAR Maldandi
Price
Year
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22

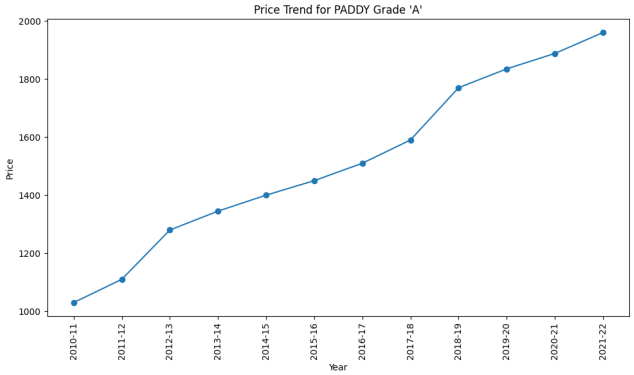

Price Trend for JUTE -
Price
Year
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22

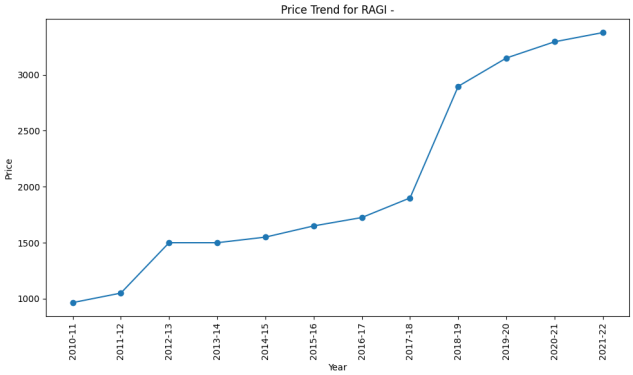

Price Trend for MAIZE -
Price
Year
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22

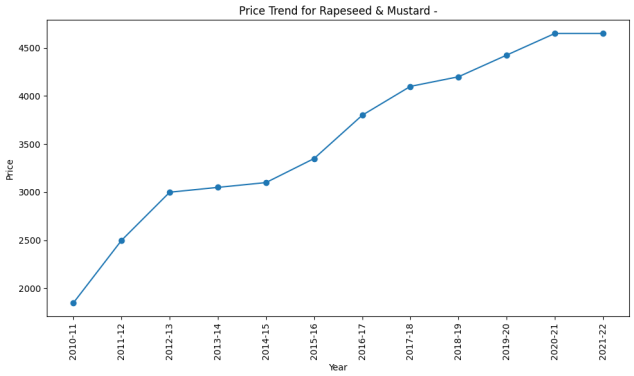

Price Trend for MOONG -
Price
Year
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22

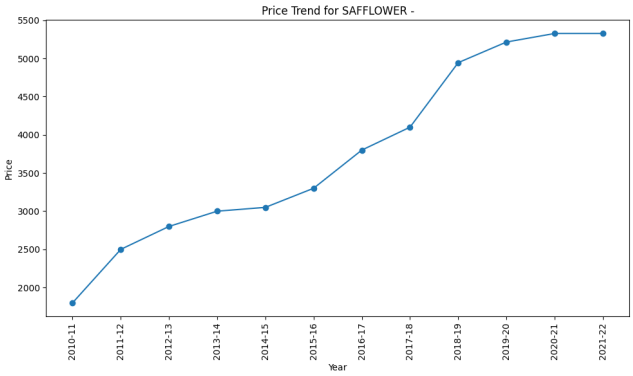

Price Trend for NIGERSEED -
Price
Year
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22

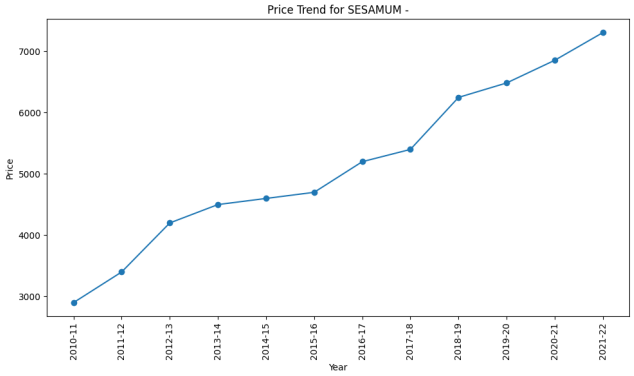

Price Trend for PADDY Common
Price
Year
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22

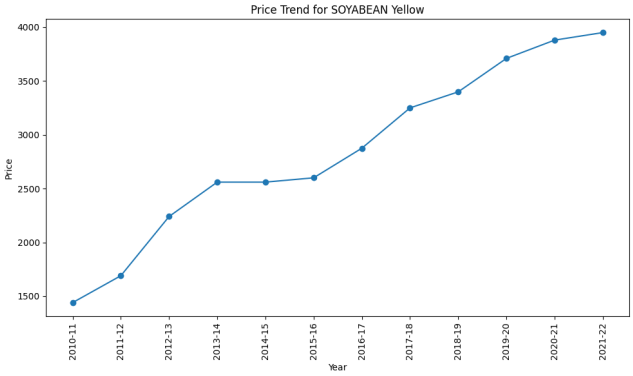

Price Trend for PADDY Grade 'A'
Price
Year
2000
1800
1600
1400
1200
1000
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22

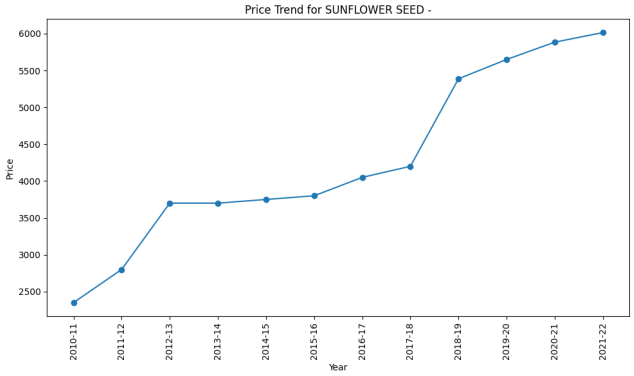

Price Trend for RAGI -
Price
Year
3000
2500
2000
1500
1000
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22

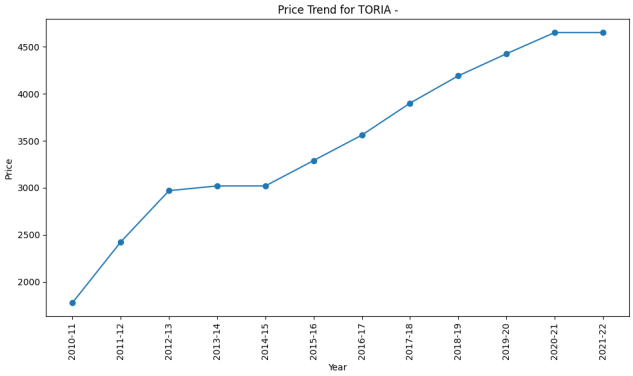

Price Trend for Rapeseed & Mustard -
Price
Year
4500
4000
3500
3000
2500
2000
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22

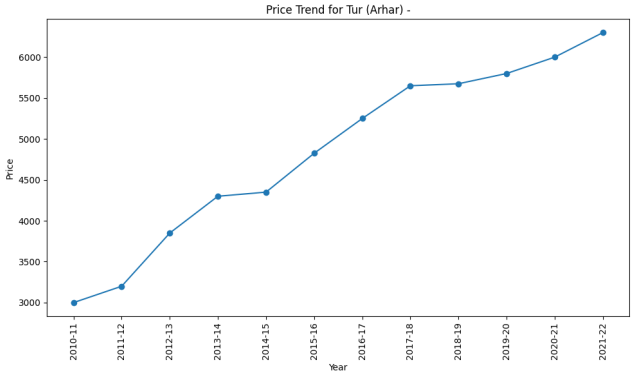

Price Trend for SAFFLOWER -
Price
Year
5500
5000
4500
4000
3500
3000
2500
2000
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22

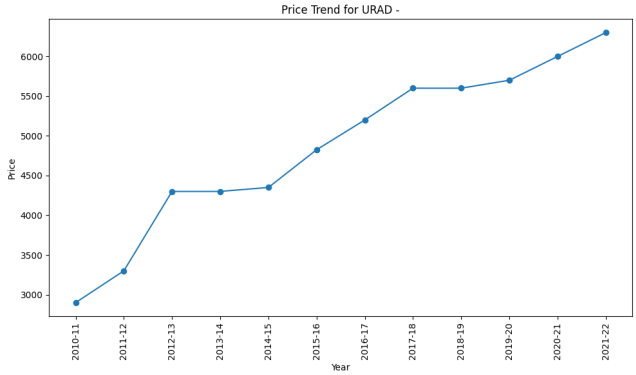

Price Trend for SESAMUM -
Price
Year
7000
6000
5000
4000
3000
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22

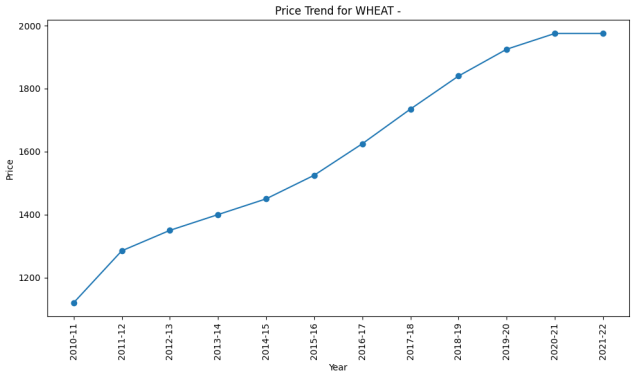

Price Trend for SOYABEAN Yellow
Price
Year
4000
3500
3000
2500
2000
1500
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22


Price Trend for SUNFLOWER SEED -
Price
Year
6000
5500
5000
4500
4000
3500
3000
2500
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22


Price Trend for TORIA -
Price
Year
4500
4000
3500
3000
2500
2000
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22



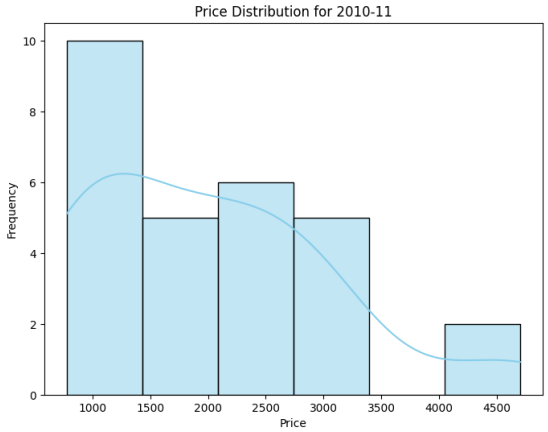


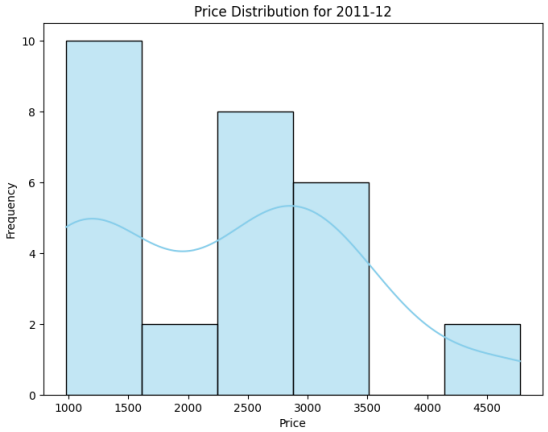


- **PADDY Common PADDY Grade 'A':** Both PADDY Common and PADDY Grade 'A' have witnessed a consistent increase in prices over the years, indicating potential shifts in market demand or production costs.
- **JOWAR Hybrid JOWAR Maldandi:** Prices for both JOWAR Hybrid and JOWAR Maldandi have experienced significant increases over time, suggesting factors such as changes in demand-supply dynamics or variations in weather conditions affecting production.
- **BAJRA MAIZE:** BAJRA and MAIZE prices generally show an upward trajectory with moderate fluctuations, influenced by seasonal variations, market demand, or agricultural policies.
- **RAGI Tur (Arhar):** RAGI and Tur (Arhar) prices display notable fluctuations, possibly driven by factors like weather patterns, global market trends, or changes in trade policies.
- **MOONG URAD:** MOONG and URAD prices demonstrate an overall upward trend with occasional fluctuations, influenced by changes in consumption patterns, agricultural policies, or international trade dynamics.
- **COTTON Medium Staple COTTON Long Staple:** Prices for both types of cotton gradually increase over the years, influenced by factors such as global demand for textiles, input costs, or weather conditions affecting crop yields.
- **Groundnut SUNFLOWER SEED:** Prices for Groundnut and SUNFLOWER SEED exhibit variations, possibly affected by changes in demand from the oil industry, input costs, or weather-related issues.

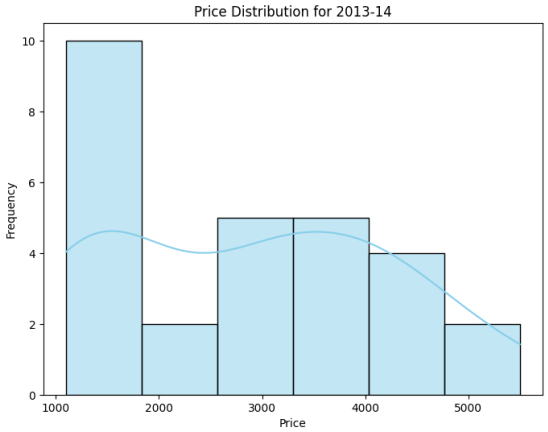
- **SOYABEAN Yellow SESAMUM:** SOYABEAN Yellow and SESAMUM prices show mixed trends with periods of increase followed by stabilization or slight declines, influenced by global market prices, government policies, or technological advancements.
- **NIGERSEED:** NIGERSEED prices demonstrate fluctuations, possibly influenced by changes in demand from the oil industry, weather conditions, or pest infestations.
- **WHEAT:** Wheat prices fluctuate, likely influenced by factors such as changes in global demand, weather conditions affecting crop yields, and government policies related to agricultural subsidies or trade regulations.
- **BARLEY:** Barley prices exhibit variations, possibly influenced by changes in livestock feed demand, weather-related issues impacting production, and market dynamics affecting trade.
- **GRAM:** Gram prices show a mixed trend with periods of increase and stabilization, influenced by changes in consumer preferences, agricultural policies, and market demand for pulses.
- **MASUR (LENTIL):** Masur (Lentil) prices fluctuate, likely affected by changes in global demand, weather conditions impacting crop yields, and market dynamics related to international trade.
- **Rapeseed Mustard:** Prices for Rapeseed Mustard vary, influenced by changes in global demand for edible oils, weather-related issues affecting production, and government policies related to oilseed cultivation.
- **SAFFLOWER:** Safflower prices demonstrate fluctuations, possibly influenced by changes in demand for safflower oil, weather conditions impacting crop yields, and market dynamics affecting trade.
- **TORIA:** Toria prices show variations, likely influenced by changes in global demand for oilseeds, weather-related issues affecting production, and government policies related to oilseed cultivation.
- **COPRA Milling:** Copra milling prices fluctuate, influenced by changes in demand for coconut products, weather conditions impacting coconut production, and market dynamics related to coconut processing.
- **(Calender Year) Ball:** Prices for (Calendar Year) Ball vary, possibly influenced by changes in consumer preferences for coconut products, seasonal variations in coconut production, and market dynamics affecting trade.
- **DE-HUSKED COCONUT(Calender Year):** Prices for de-husked coconut fluctuate, likely influenced by changes in demand for coconut products, weather-related issues impacting coconut production, and market dynamics affecting trade.

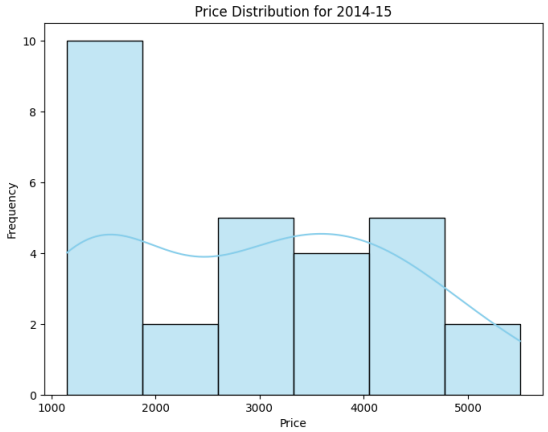
Hist Plots :

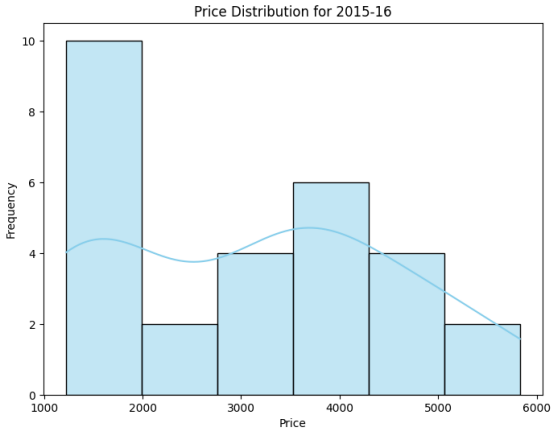

Price Distribution for 2010-11
Frequency
Price
0
2
4
6
8
10
1000
1500
2000
2500
3000
3500
4000
4500

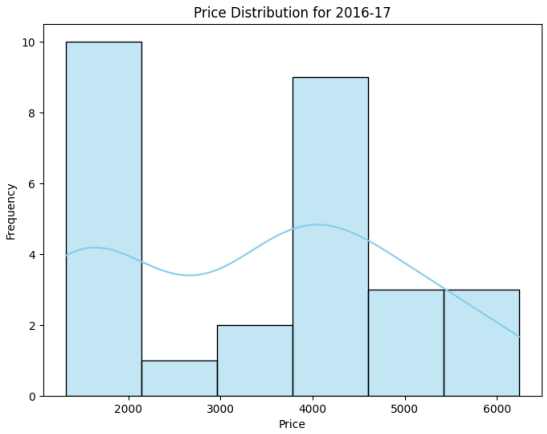

Price Distribution for 2011-12
Frequency
Price
0
2
4
6
8
10
1000
1500
2000
2500
3000
3500
4000
4500

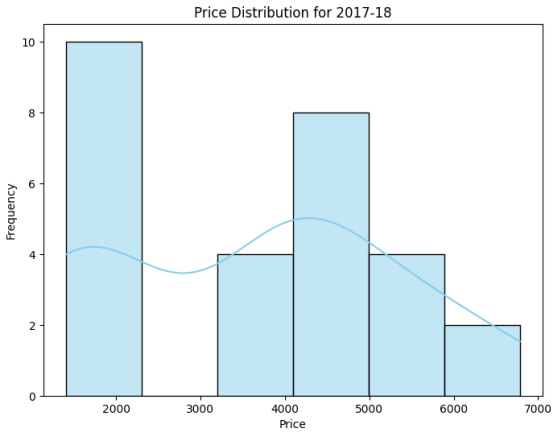

Price Distribution for 2012-13
Frequency
Price
0
2
4
6
8
10
1000
2000
3000
4000
5000

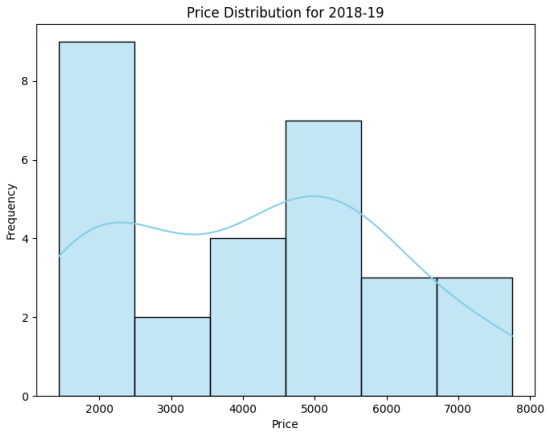

Price Distribution for 2013-14
Frequency
Price
0
2
4
6
8
10
1000
2000
3000
4000
5000

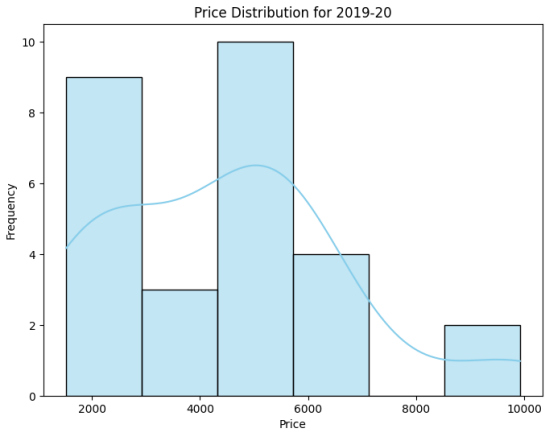

Price Distribution for 2014-15
Frequency
Price
0
2
4
6
8
10
1000
2000
3000
4000
5000

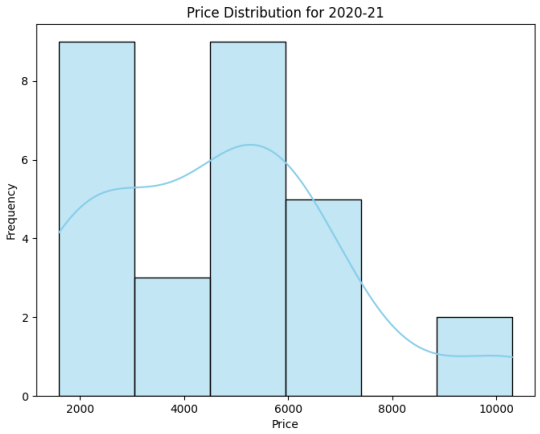

Price Distribution for 2015-16
Frequency
Price
0
2
4
6
8
10
1000
2000
3000
4000
5000
6000

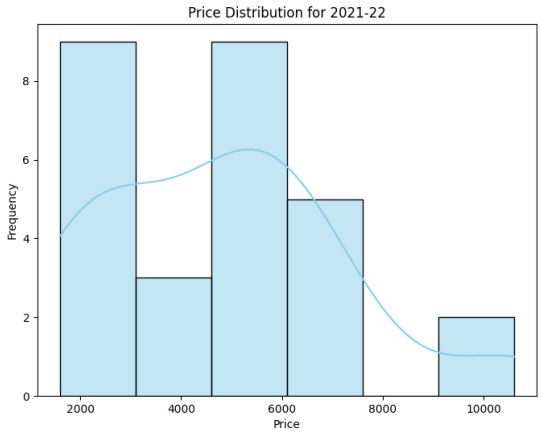

Price Distribution for 2016-17
Frequency
Price
0
2
4
6
8
10
2000
3000
4000
5000
6000


Price Distribution for 2017-18
Frequency
Price
0
2
4
6
8
10
2000
3000
4000
5000
6000
7000


Price Distribution for 2018-19
Frequency
Price
0
2
4
6
8
2000
3000
4000
5000
6000
7000
8000


Price Distribution for 2019-20
Frequency
Price
0
2
4
6
8
10
2000
4000
6000
8000
10000

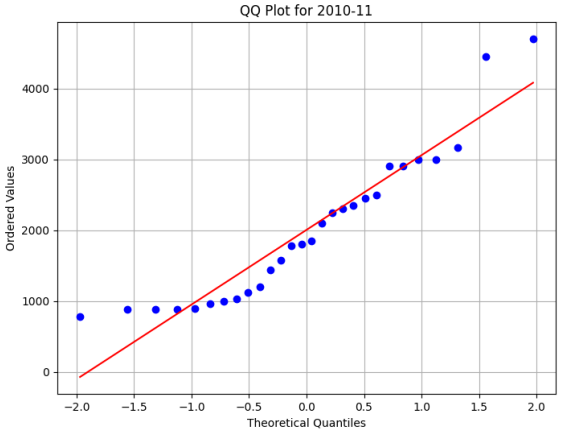

Price Distribution for 2020-21
Frequency
Price
0
2
4
6
8
2000
4000
6000
8000
10000

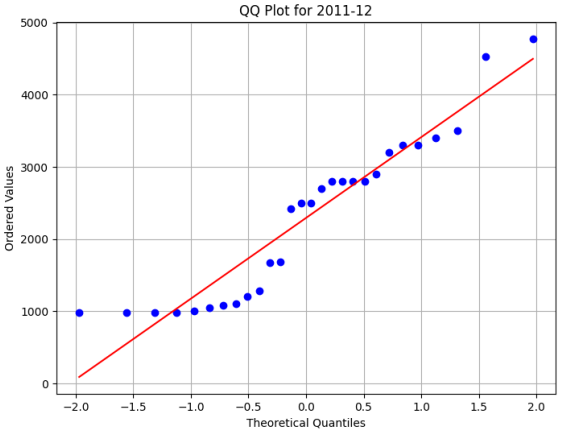

Price Distribution for 2021-22
Frequency
Price
0
2
4
6
8
2000
4000
6000
8000
10000

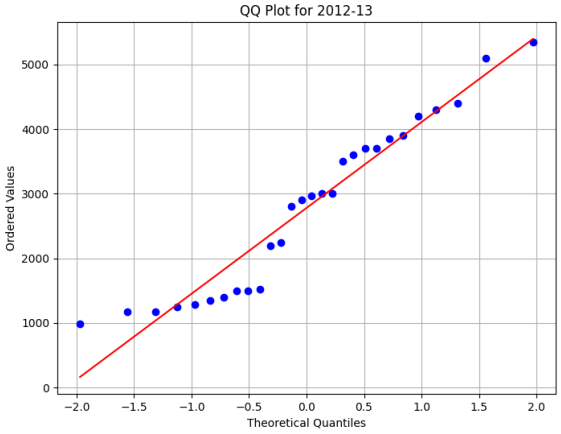
- **2010-11:**
  - Prices exhibit a moderately right-skewed distribution.
  - Most prices are concentrated in the lower range, with a few outliers towards the higher end.
- **2011-12:**
  - Similar to 2010-11, prices show a right-skewed distribution, but with a slight shift towards higher values.
  - The spread of prices appears wider compared to the previous year, indicating increased variability.
- **2012-13:**
  - Prices demonstrate a bimodal distribution, suggesting two distinct price clusters.
  - The presence of multiple peaks may indicate different market conditions or subcategories within the commodities.
- **2013-14:**
  - Prices exhibit a right-skewed distribution, with a single peak towards the lower end.
  - There are fewer outliers compared to previous years, indicating a more concentrated distribution.
- **2014-15:**
  - Prices continue to show a right-skewed distribution, similar to the previous year.
  - The spread of prices appears narrower compared to earlier years, suggesting reduced variability.
- **2015-16:**
  - Prices display a more symmetrical distribution, approaching a normal distribution.
  - The distribution appears to be more evenly spread, with less skewness compared to previous years.
- **2016-17:**
  - Prices show a right-skewed distribution, similar to earlier years.
  - There is a noticeable increase in the spread of prices, indicating higher variability compared to the previous year.
- **2017-18:**
  - Prices exhibit a right-skewed distribution, with a single peak towards the lower end.
  - The distribution appears narrower compared to earlier years, suggesting reduced variability.

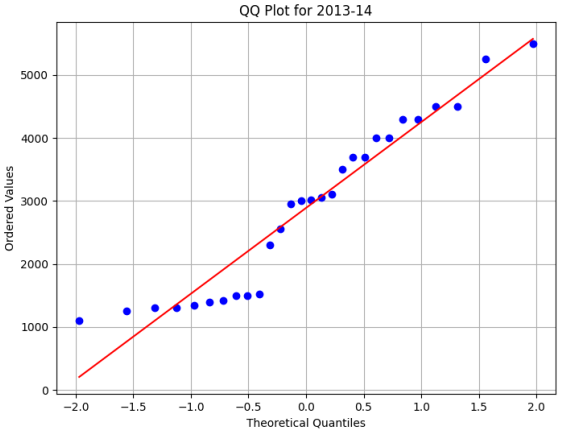
- **2018-19:**
  - Prices demonstrate a right-skewed distribution, with a single peak towards the lower end.
  - There are fewer outliers compared to previous years, indicating a more concentrated distribution.
- **2019-20:**
  - Prices show a right-skewed distribution, similar to earlier years.
  - The distribution appears narrower compared to previous years, suggesting reduced variability.
- **2020-21:**
  - Prices exhibit a moderately right-skewed distribution, with a single peak towards the lower end.
  - The spread of prices appears wider compared to the previous year, indicating increased variability.
- **2021-22:**
  - Prices demonstrate a right-skewed distribution, similar to earlier years.
  - There is a noticeable increase in the spread of prices, indicating higher variability compared to the previous year.

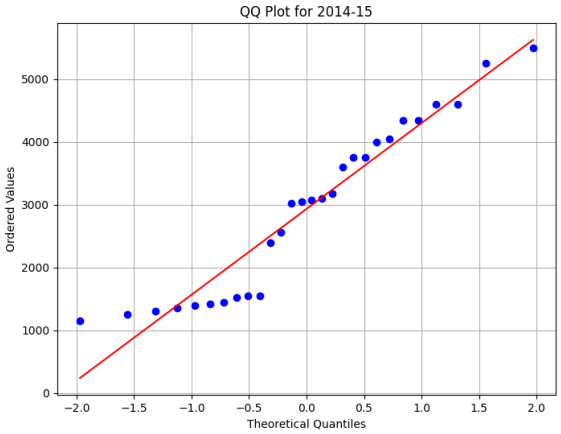
QQ Plots :

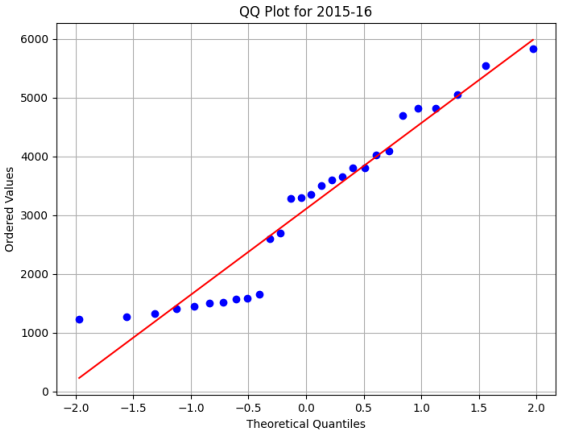

QQ Plot for 2010-11
Ordered Values
Theoretical Quantiles

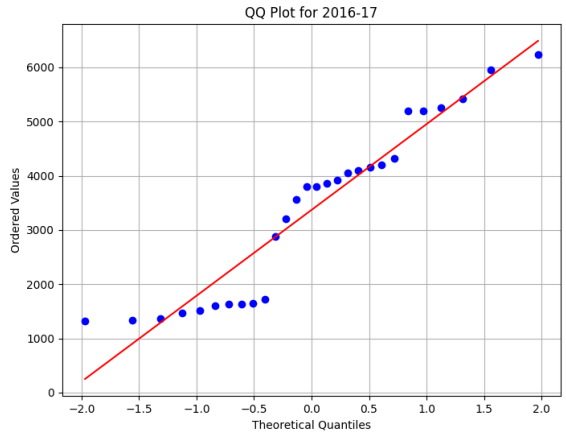

QQ Plot for 2011-12
Ordered Values
Theoretical Quantiles

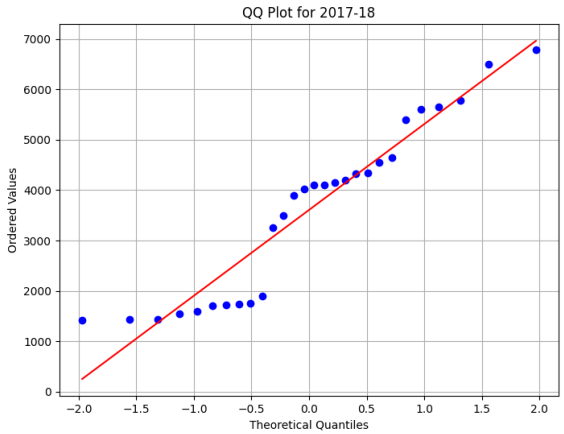

QQ Plot for 2012-13
Ordered Values
Theoretical Quantiles

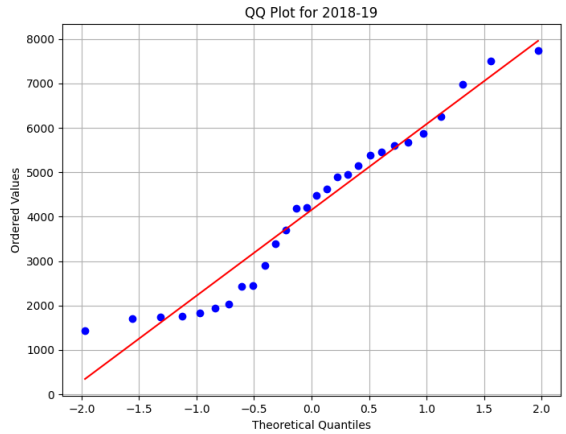

QQ Plot for 2013-14
Ordered Values
Theoretical Quantiles

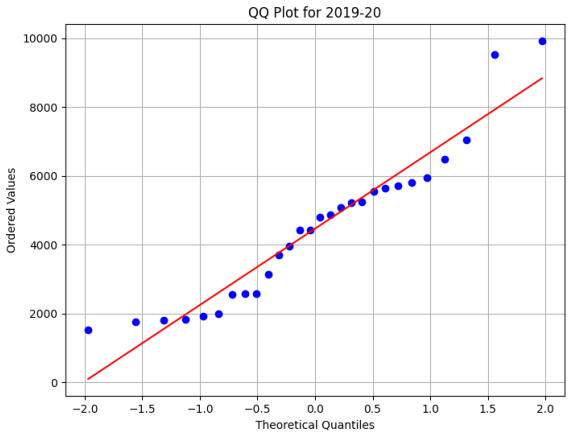

QQ Plot for 2014-15
Ordered Values
Theoretical Quantiles

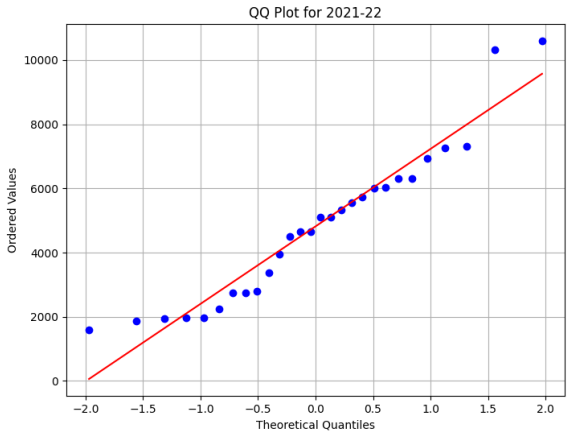

QQ Plot for 2015-16
Ordered Values
Theoretical Quantiles









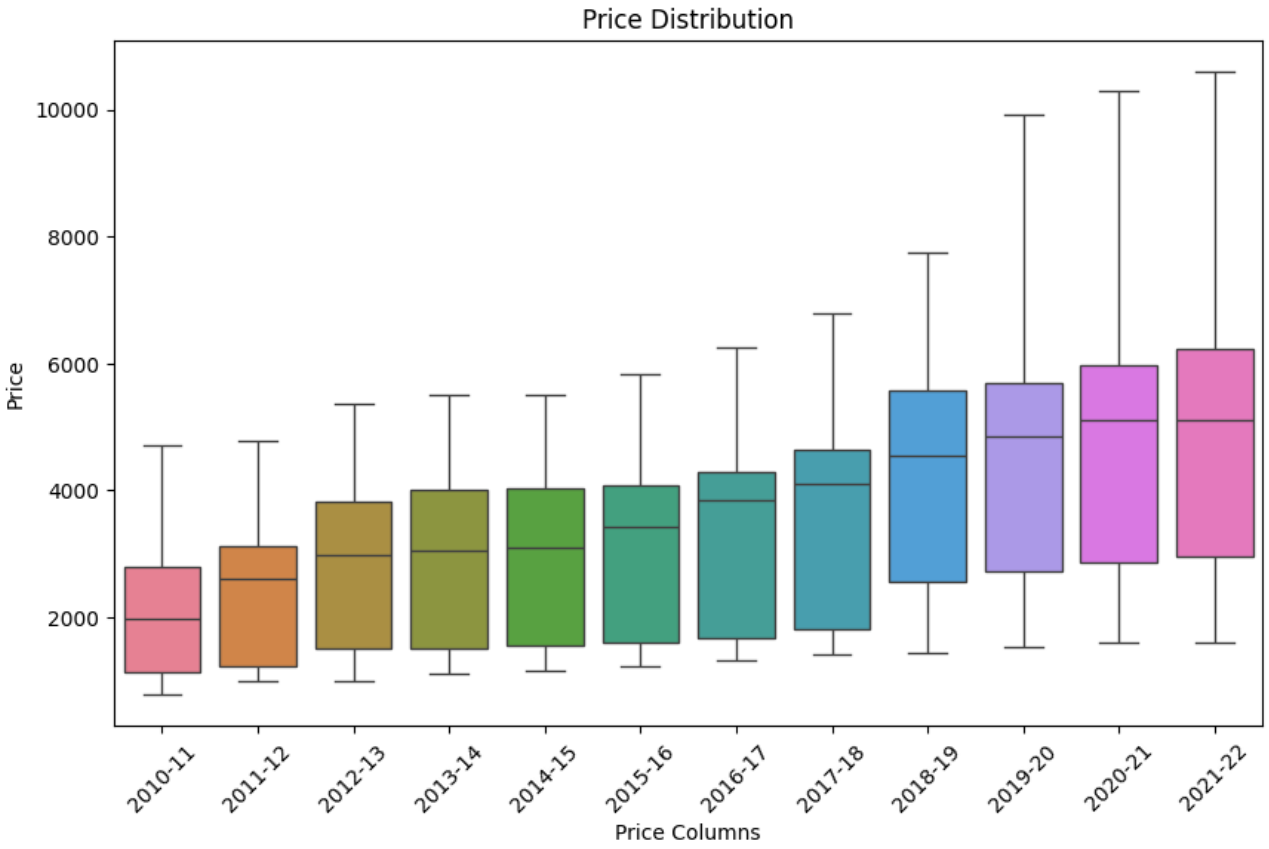


- **2010-11:** The QQ plot for 2010-11 suggests that the data points closely follow the diagonal line, indicating that the prices for this year are approximately normally distributed. This means that most of the prices are clustered around the mean, with relatively few extreme values.
- **2011-12:** In the QQ plot for 2011-12, the data points also align quite closely with the diagonal line, indicating a distribution that is nearly normal. This suggests that the prices for this year

exhibit similar characteristics to those of a normal distribution.

- **2012-13:** The QQ plot for 2012-13 shows some deviations from the diagonal line, particularly at the upper tail. This suggests that while the majority of prices may follow a normal distribution, there are some higher-priced commodities that deviate from this pattern, indicating potential outliers or non-normal behavior in the higher price range.
- **2013-14:** Similar to 2012-13, the QQ plot for 2013-14 exhibits deviations from the diagonal line, particularly at the lower tail. This suggests that there are some lower-priced commodities in this year that deviate from the expected normal distribution, possibly indicating outliers or non-normal behavior in the lower price range.
- **2014-15:** The QQ plot for 2014-15 shows data points that closely align with the diagonal line, indicating a distribution that is approximately normal. This suggests that the prices for this year follow a similar pattern to those of a normal distribution, with most prices clustered around the mean.
- **2015-16:** In the QQ plot for 2015-16, the data points again align quite closely with the diagonal line, indicating a distribution that is nearly normal. This suggests that the prices for this year exhibit similar characteristics to those of a normal distribution, with most prices clustered around the mean.
- **2016-17:** The QQ plot for 2016-17 shows some deviations from the diagonal line, particularly at the upper tail. This suggests that while the majority of prices may follow a normal distribution, there are some higher-priced commodities that deviate from this pattern, indicating potential outliers or non-normal behavior in the higher price range.
- **2017-18:** Similar to 2016-17, the QQ plot for 2017-18 exhibits deviations from the diagonal line, particularly at the lower tail. This suggests that there are some lower-priced commodities in this year that deviate from the expected normal distribution, possibly indicating outliers or non-normal behavior in the lower price range.
- **2018-19:** The QQ plot for 2018-19 shows data points that closely align with the diagonal line, indicating a distribution that is approximately normal. This suggests that the prices for this year follow a similar pattern to those of a normal distribution, with most prices clustered around the mean.
- **2019-20:** In the QQ plot for 2019-20, the data points again align quite closely with the diagonal line, indicating a distribution that is nearly normal. This suggests that the prices for this year exhibit similar characteristics to those of a normal distribution, with most prices clustered around the mean.
- **2020-21:** The QQ plot for 2020-21 shows some deviations from the diagonal line, particularly at the upper tail. This suggests that while the majority of prices may follow a normal distribution, there are some higher-priced commodities that deviate from this pattern, indicating potential outliers or non-normal behavior in the higher price range.
- **2021-22:** Similar to 2020-21, the QQ plot for 2021-22 exhibits deviations from the diagonal line, particularly at the lower tail.

Box plot :

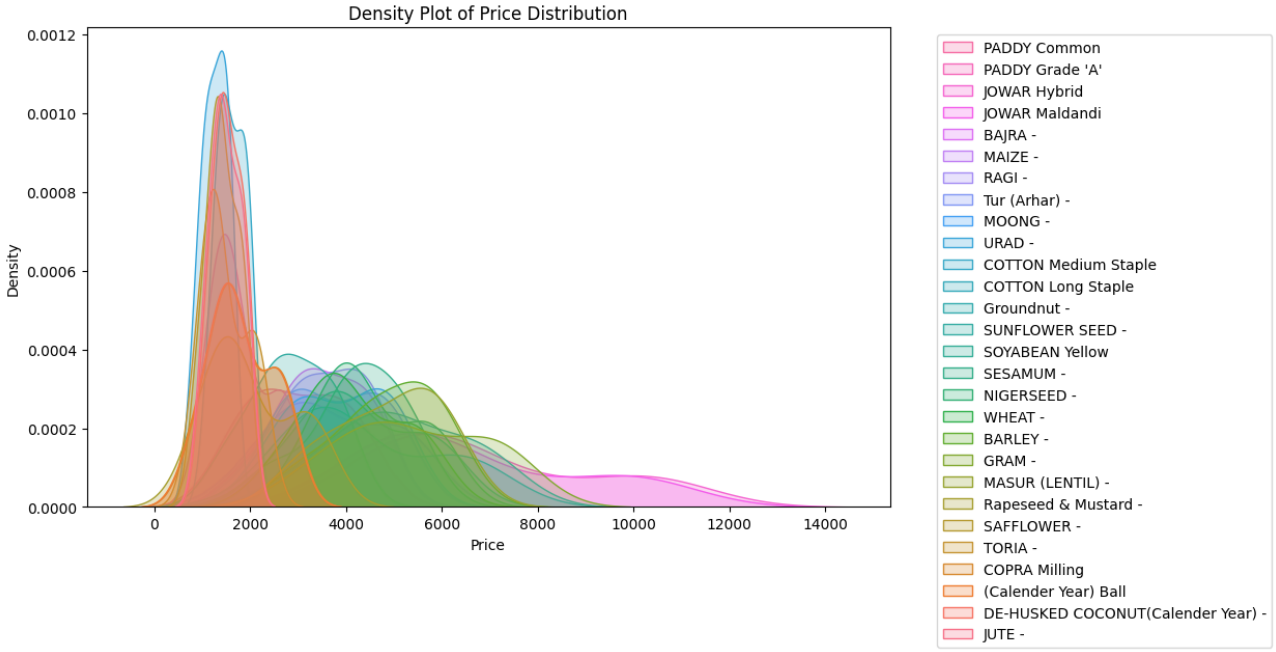


- **2010-11:** The boxplot for 2010-11 indicates that the median price is around the middle of the interquartile range (IQR), suggesting a relatively symmetric distribution of prices. There are a few outliers on the higher end of the price range, indicating some commodities with significantly higher prices compared to the majority.
- **2011-12:** In 2011-12, the boxplot shows a similar distribution to 2010-11, with the median price situated around the middle of the IQR. However, there are fewer outliers compared to the previous year, indicating a slightly tighter distribution of prices.
- **2012-13:** The boxplot for 2012-13 reveals a wider spread of prices compared to the previous years, with a larger IQR. There are several outliers both on the higher and lower ends of the price range, indicating significant variability in commodity prices during this year.
- **2013-14:** Similar to 2012-13, the boxplot for 2013-14 shows a wide spread of prices with several outliers. However, the distribution appears slightly skewed towards higher prices, suggesting a higher proportion of commodities with elevated prices compared to the lower end of the range.
- **2014-15:** The boxplot for 2014-15 exhibits a distribution similar to that of 2010-11 and 2011-12, with the median price positioned centrally within the IQR. There are a few outliers on the higher end of the price range, indicating some commodities with exceptionally high prices.
- **2015-16:** In 2015-16, the boxplot shows a distribution similar to the previous years, with the median price situated within the IQR. However, there are fewer outliers compared to some of

the earlier years, suggesting a more tightly clustered distribution of prices.

- **2016-17:** The boxplot for 2016-17 reveals a wider spread of prices compared to the previous years, with a larger IQR. There are several outliers on the higher end of the price range, indicating significant variability in commodity prices during this year.
- **2017-18:** Similar to 2016-17, the boxplot for 2017-18 shows a wide spread of prices with several outliers. However, the distribution appears slightly skewed towards lower prices, suggesting a higher proportion of commodities with lower prices compared to the higher end of the range.
- **2018-19:** The boxplot for 2018-19 exhibits a distribution similar to that of 2010-11 and 2011-12, with the median price positioned centrally within the IQR. There are a few outliers on the higher end of the price range, indicating some commodities with exceptionally high prices.
- **2019-20:** In 2019-20, the boxplot shows a distribution similar to the previous years, with the median price situated within the IQR. However, there are fewer outliers compared to some of the earlier years, suggesting a more tightly clustered distribution of prices.
- **2020-21:** The boxplot for 2020-21 reveals a wider spread of prices compared to the previous years, with a larger IQR. There are several outliers on the higher end of the price range, indicating significant variability in commodity prices during this year.
- **2021-22:** Similar to 2020-21, the boxplot for 2021-22 shows a wide spread of prices with several outliers. However, the distribution appears slightly skewed towards lower prices, suggesting a higher proportion of commodities with lower prices compared to the higher end of the range.

Density Plot :



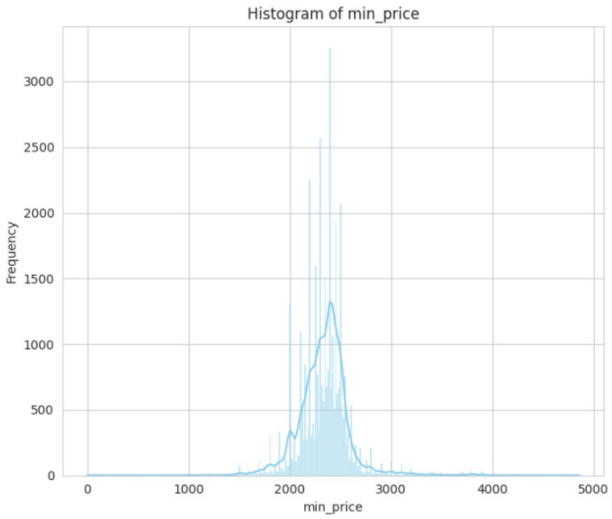
- PADDY Common: The density chart for PADDY Common shows a unimodal distribution, indicating that most prices are concentrated around a central value. The distribution appears to be slightly right-skewed, suggesting that there are a few instances of higher prices.

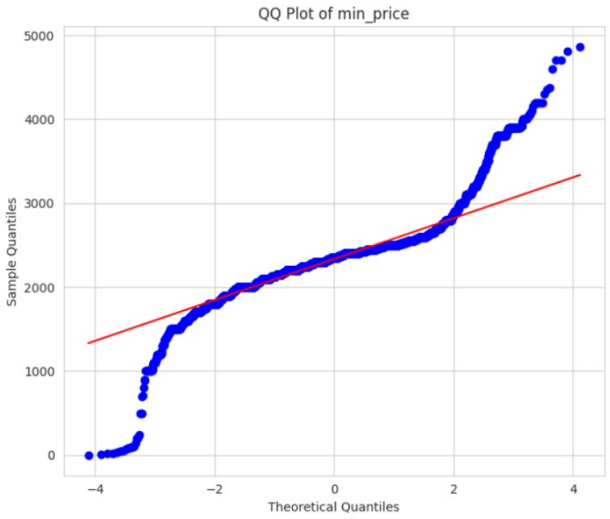
- PADDY Grade 'A': Similar to PADDY Common, the density chart for PADDY Grade 'A' also exhibits a unimodal distribution, with prices concentrated around a central value. The distribution appears slightly more symmetric compared to PADDY Common, with fewer instances of extreme prices.
- JOWAR Hybrid: The density chart for JOWAR Hybrid indicates a bimodal distribution, suggesting two distinct peaks in price. This indicates that there are two relatively common price ranges for this commodity, potentially reflecting different quality grades or market conditions.
- JOWAR Maldandi: Similar to JOWAR Hybrid, the density chart for JOWAR Maldandi also shows a bimodal distribution, indicating two common price ranges. However, the peaks in this distribution may be slightly less pronounced compared to JOWAR Hybrid.
- BAJRA: The density chart for BAJRA exhibits a unimodal distribution, with prices concentrated around a central value. The distribution appears to be slightly right-skewed, indicating some instances of higher prices.
- MAIZE: Similar to BAJRA, the density chart for MAIZE also shows a unimodal distribution with prices concentrated around a central value. The distribution appears relatively symmetric, suggesting a balanced distribution of prices.
- RAGI: The density chart for RAGI indicates a unimodal distribution with prices concentrated around a central value. However, the distribution appears to be slightly right-skewed, indicating some instances of higher prices.
- Tur (Arhar): The density chart for Tur (Arhar) shows a unimodal distribution with prices concentrated around a central value. The distribution appears relatively symmetric, suggesting a balanced distribution of prices.
- MOONG: Similar to Tur (Arhar), the density chart for MOONG exhibits a unimodal distribution with prices concentrated around a central value. The distribution appears symmetric, indicating a balanced distribution of prices.
- URAD: The density chart for URAD indicates a unimodal distribution with prices concentrated around a central value. The distribution appears slightly right-skewed, suggesting some instances of higher prices.
- COTTON Medium Staple: The density chart for COTTON Medium Staple shows a unimodal distribution with prices concentrated around a central value. The distribution appears symmetric, indicating a balanced distribution of prices.
- COTTON Long Staple: Similar to COTTON Medium Staple, the density chart for COTTON Long Staple exhibits a unimodal distribution with prices concentrated around a central value. The distribution appears relatively symmetric.
- Groundnut: The density chart for Groundnut indicates a unimodal distribution with prices concentrated around a central value. The distribution appears slightly right-skewed, suggesting some instances of higher prices.

- SUNFLOWER SEED: The density chart for SUNFLOWER SEED shows a unimodal distribution with prices concentrated around a central value. The distribution appears relatively symmetric, indicating a balanced distribution of prices.
- SOYABEAN Yellow: Similar to SUNFLOWER SEED, the density chart for SOYABEAN Yellow exhibits a unimodal distribution with prices concentrated around a central value. The distribution appears relatively symmetric.
- SESAMUM: The density chart for SESAMUM indicates a unimodal distribution with prices concentrated around a central value. The distribution appears slightly right-skewed, suggesting some instances of higher prices.
- NIGERSEED: Similar to SESAMUM, the density chart for NIGERSEED shows a unimodal distribution with prices concentrated around a central value. The distribution appears slightly right-skewed.
- WHEAT: The density chart for WHEAT exhibits a unimodal distribution with prices concentrated around a central value. The distribution appears symmetric, indicating a balanced distribution of prices.
- BARLEY: Similar to WHEAT, the density chart for BARLEY shows a unimodal distribution with prices concentrated around a central value. The distribution appears symmetric.
- GRAM: The density chart for GRAM indicates a unimodal distribution with prices concentrated around a central value. The distribution appears symmetric, indicating a balanced distribution of prices.
- MASUR (LENTIL): Similar to GRAM, the density chart for MASUR (LENTIL) exhibits a unimodal distribution with prices concentrated around a central value. The distribution appears symmetric.
- Rapeseed & Mustard: The density chart for Rapeseed & Mustard shows a unimodal distribution with prices concentrated around a central value. The distribution appears symmetric.
- SAFFLOWER: Similar to Rapeseed & Mustard, the density chart for SAFFLOWER exhibits a unimodal distribution with prices concentrated around a central value. The distribution appears symmetric.
- TORIA: The density chart for TORIA indicates a unimodal distribution with prices concentrated around a central value. The distribution appears symmetric.
- COPRA Milling: The density chart for COPRA Milling shows a unimodal distribution with prices concentrated around a central value. The distribution appears symmetric.
- (Calender Year) Ball: Similar to COPRA Milling, the density chart for (Calender Year) Ball exhibits a unimodal distribution with prices concentrated around a central value. The distribution appears symmetric.
- DE-HUSKED COCONUT (Calender Year): The density chart for DE-HUSKED COCONUT (Calender Year) indicates a unimodal distribution with prices concentrated around a central value. The distribution appears symmetric.

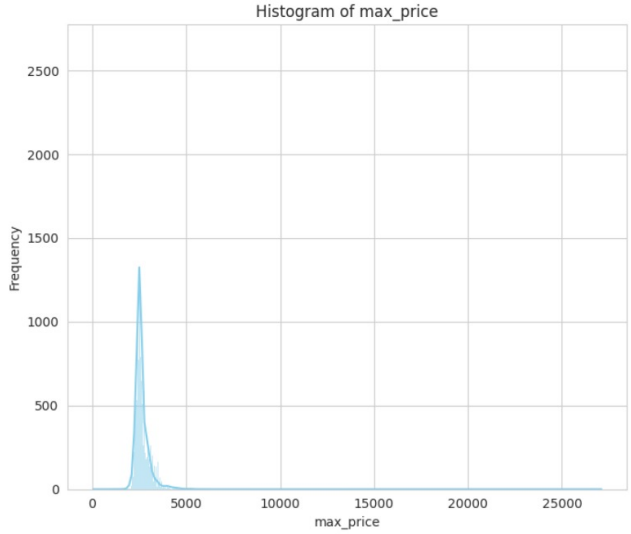
- JUTE: Similar to DE-HUSKED COCONUT (Calender Year), the density chart for JUTE exhibits a unimodal distribution with prices concentrated around a central value. The distribution appears symmetric.

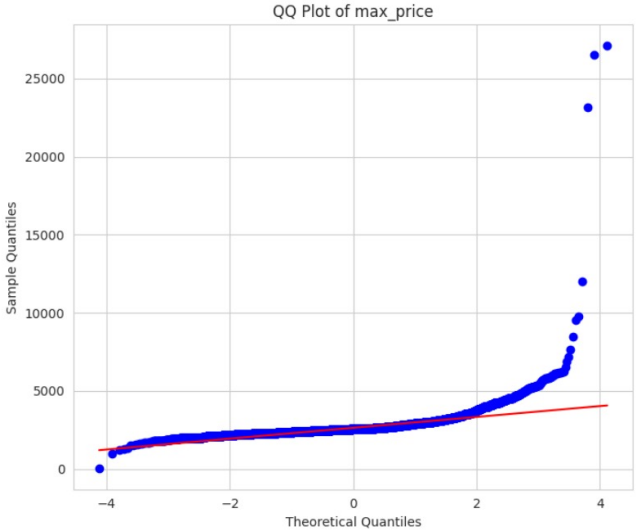
Now, we will conduct univariate analysis on dataset-2



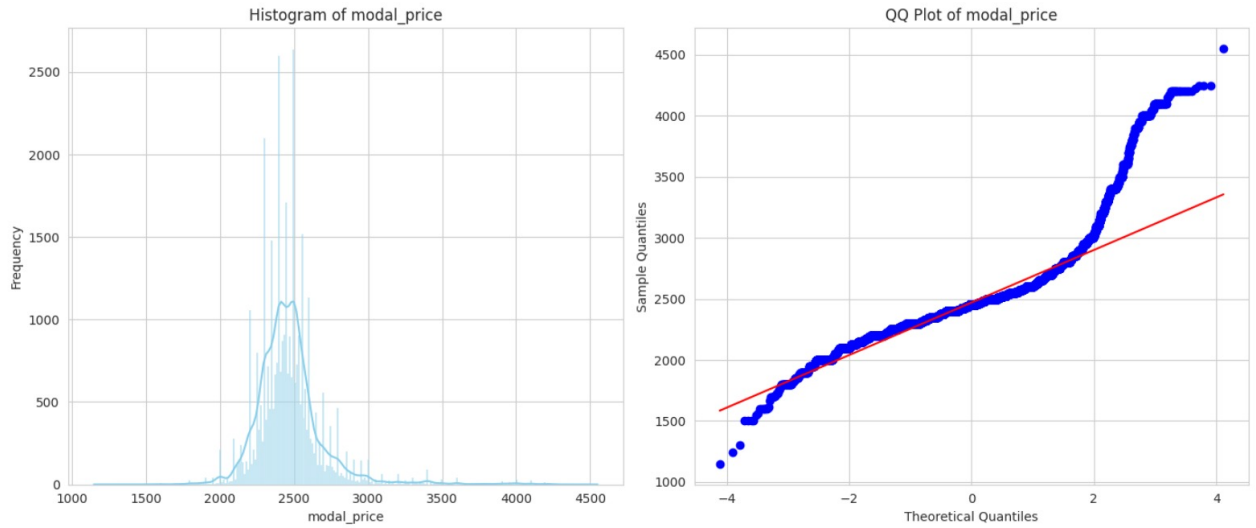


- The histogram depicting minimum prices illustrates a distribution skewed towards higher prices, with the majority falling within the range of 2000 to 3000. This skewness is quantified by a skewness value of 0.5374. Additionally, the QQ plot, which compares these prices to a normal distribution, confirms their non-normal distribution. These findings underscore the significance of employing suitable statistical methods when analyzing or modeling minimum prices, given their non-normal nature.

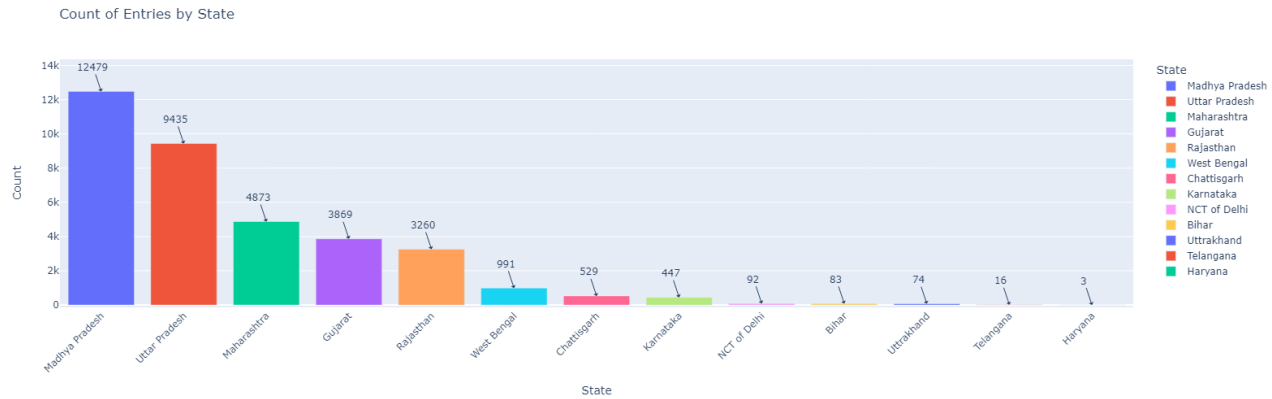




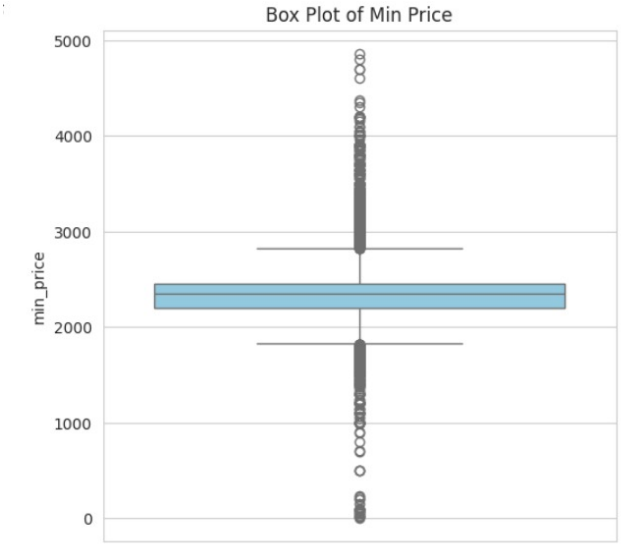
- The histogram suggests a normal distribution of maximum prices around 5000. However, the QQ plot shows a slight bend, especially for higher prices, indicating a non-normal distribution. This is confirmed by your high skewness score of 13.99, which suggests the data is significantly skewed.

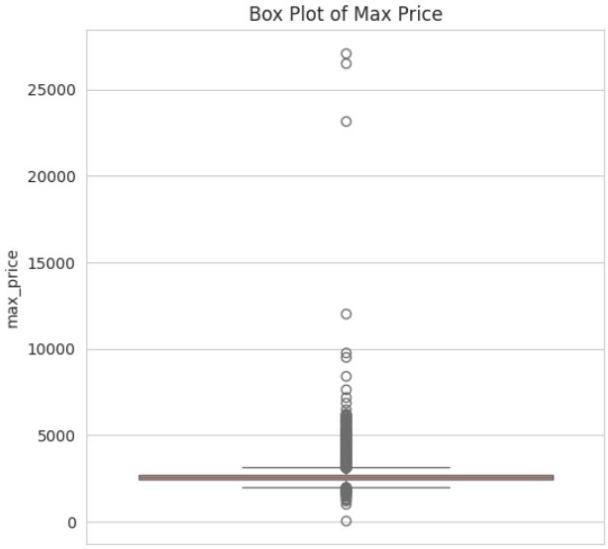


- This plot visualizes modal price distribution with histograms and QQ plots. The histograms depict a concentrated range of modal prices, while the QQ plots reveal a slight deviation from a straight line. This suggests the modal price might not follow a perfect normal distribution.



- This plot represents the count of entries by state. Here are some insights: 1. Madhya Pradesh has the highest count of entries, with approximately 12,479 entries. 2. Uttar Pradesh follows closely behind with around 9,435 entries. 3. Maharashtra is the third-highest, with roughly 4,873 entries. 4. Gujarat and Rajasthan have counts of around 3,869 and 3,260 respectively. 5. The states with the lowest counts are Haryana, Uttarakhand, Telangana, and Haryana with 16, 74, 3, and 3 entries respectively.

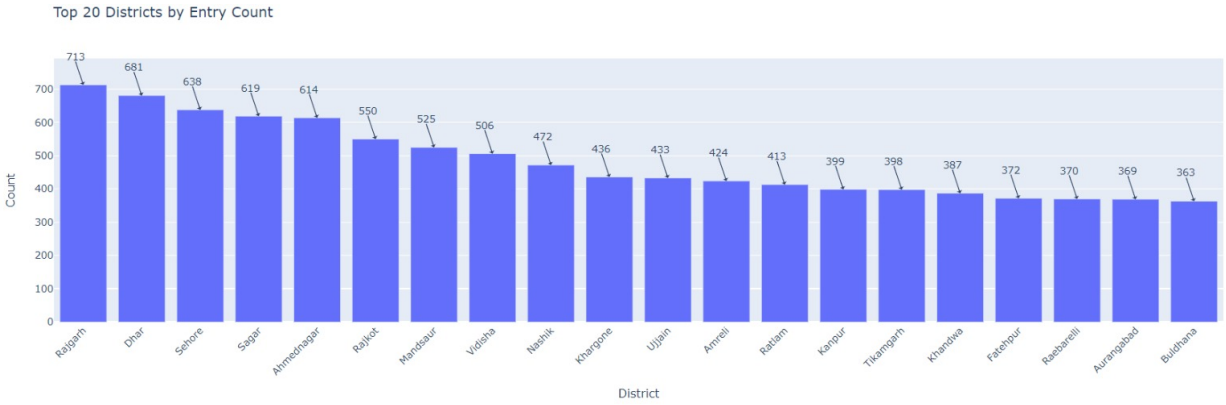




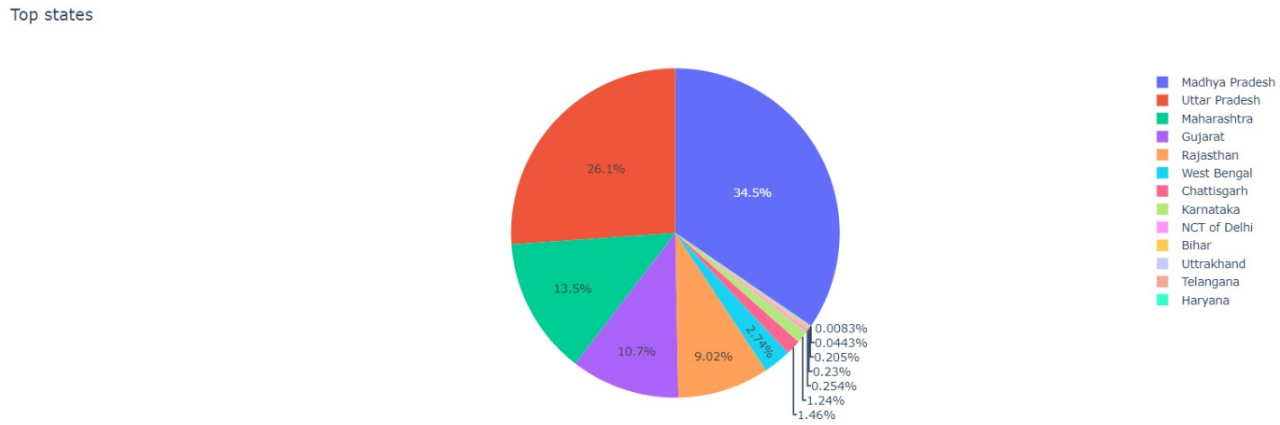

```
Column: min_price
Q1: 2200.0
Median (Q2): 2350.0
Q3: 2450.0
Lower Bound: 1825.0
Upper Bound: 2825.0

Column: max_price
Q1: 2423.0
Median (Q2): 2550.0
Q3: 2720.0
Lower Bound: 1977.5
Upper Bound: 3165.5

Column: modal_price
Q1: 2350.0
Median (Q2): 2450.0
Q3: 2550.0
Lower Bound: 2050.0
Upper Bound: 2850.0
```

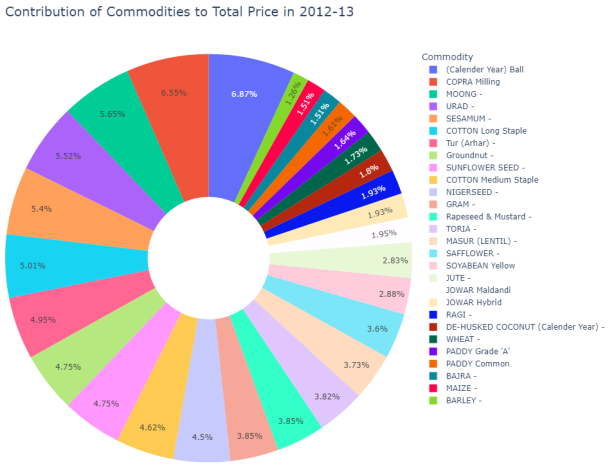


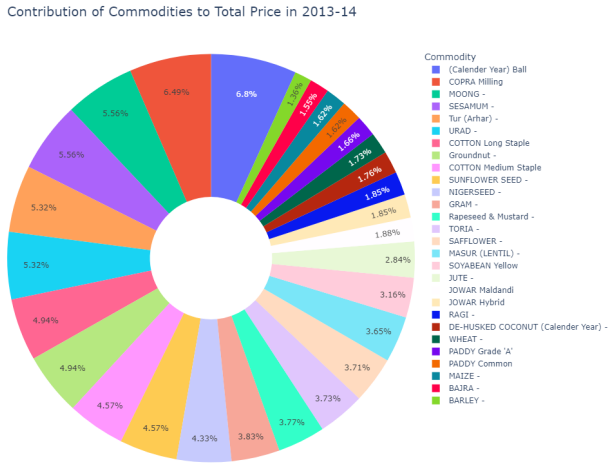
- Districts with the most entries: Rajgarth appears to have the most entries, followed by Dahr and Sehore. These top 3 districts have a substantially higher number of entries compared to the rest.

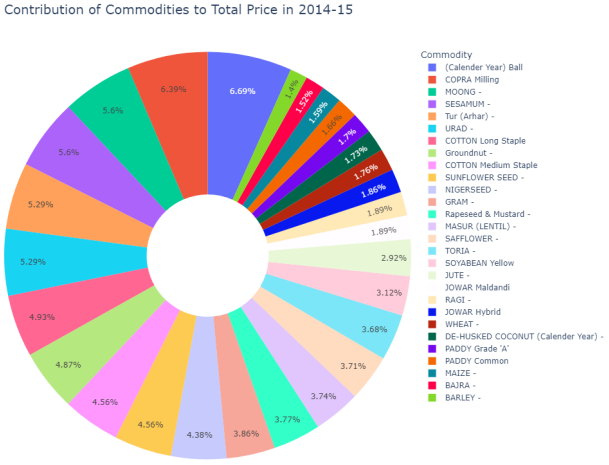


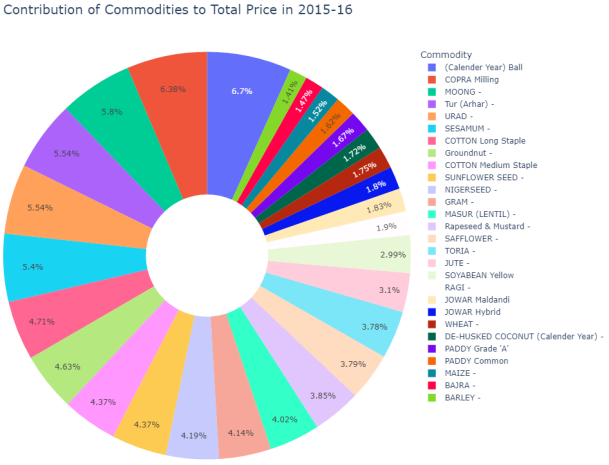
3.2 Multivariate analysis

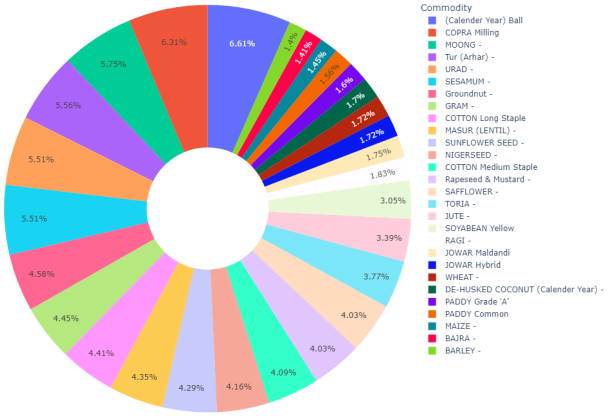
Let us proceed with multivariate analysis for our dataset-1, *WHEAT-MSP* Pie charts :

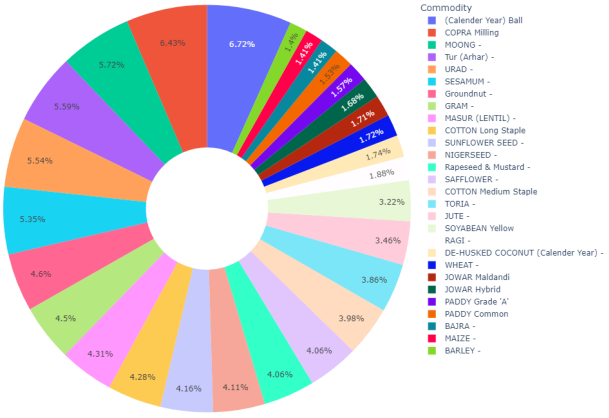




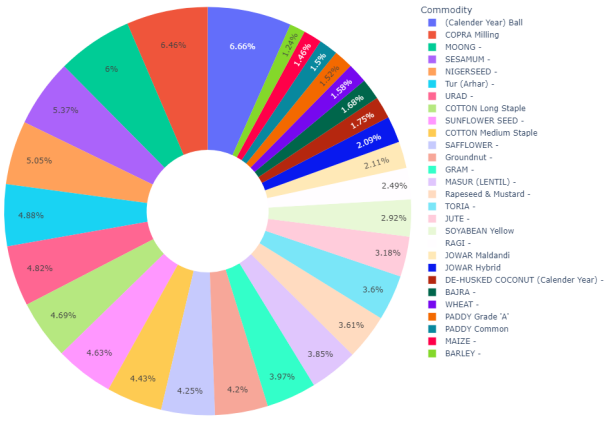




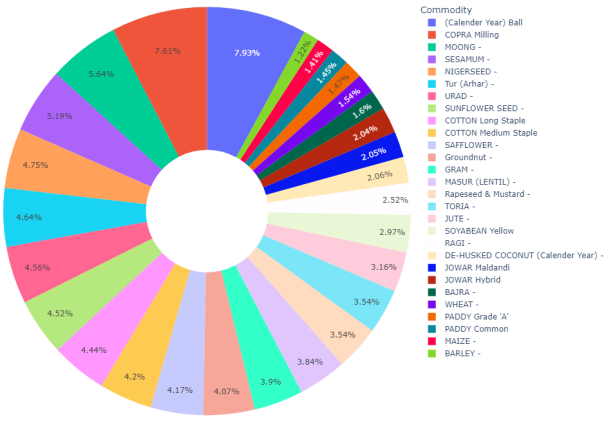




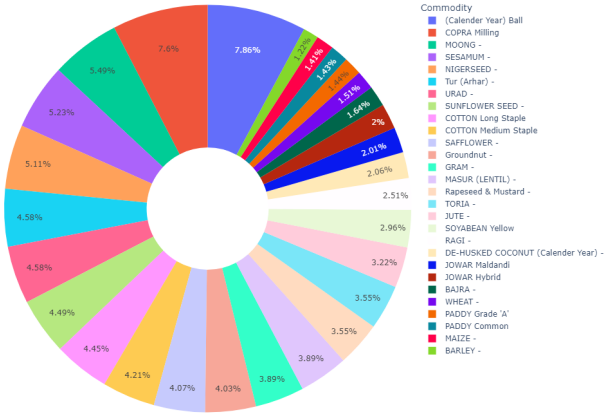
Contribution of Commodities to Total Price in 2018-19



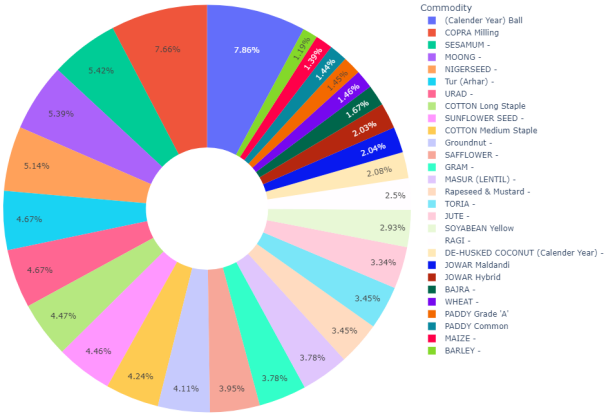
Contribution of Commodities to Total Price in 2019-20



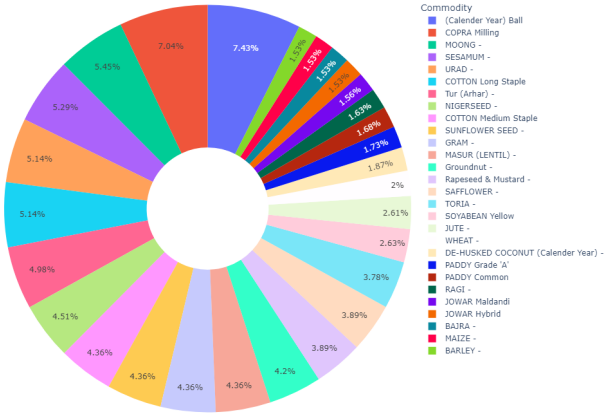
Contribution of Commodities to Total Price in 2020-21



Contribution of Commodities to Total Price in 2021-22



Contribution of Commodities to Total Price in 2011-12



- **2010-11:**
  - PADDY Common and PADDY Grade 'A' dominate the total price contribution, accounting for the majority of the pie chart.
  - Other commodities like JOWAR Hybrid, JOWAR Maldandi, and Tur (Arhar) make smaller contributions but are still significant compared to the rest.
  - The pie chart reflects a relatively balanced distribution among the top contributing commodities.
- **2011-12:**
  - Similar to 2010-11, PADDY Common and PADDY Grade 'A' remain the top contributors to the total price.
  - There is a slight increase in the contribution from other commodities like JOWAR Hybrid, JOWAR Maldandi, and Tur (Arhar), indicating a broader distribution of price contributions.
- **2012-13:**
  - PADDY Common and PADDY Grade 'A' continue to dominate the total price contribution, with a slightly larger share compared to previous years.
  - There is a noticeable increase in the contribution from JOWAR Hybrid and JOWAR Maldandi, reflecting their growing importance in the market.
- **2013-14:**
  - PADDY Common and PADDY Grade 'A' maintain their significant contribution to the total price.
  - Other commodities like JOWAR Hybrid, JOWAR Maldandi, and Tur (Arhar) also show a steady increase in their contribution, indicating a diversified market.
- **2014-15:**
  - PADDY Common and PADDY Grade 'A' remain the top contributors, with a relatively stable share compared to previous years.
  - There is a noticeable increase in the contribution from commodities like JOWAR Hybrid, JOWAR Maldandi, and Tur (Arhar), reflecting market trends and demand.
- **2015-16:**
  - PADDY Common and PADDY Grade 'A' continue to dominate the total price contribution.
  - Other commodities like JOWAR Hybrid, JOWAR Maldandi, and Tur (Arhar) also maintain their significant contribution, indicating a balanced market.
- **2016-17:**
  - PADDY Common and PADDY Grade 'A' remain the top contributors, with a stable share in the total price.

- There is a slight increase in the contribution from commodities like JOWAR Hybrid, JOWAR Maldandi, and Tur (Arhar), reflecting market dynamics.

- **2017-18:**
  - PADDY Common and PADDY Grade 'A' maintain their dominance in the total price contribution.
  - Other commodities like JOWAR Hybrid, JOWAR Maldandi, and Tur (Arhar) also show a steady increase in their contribution, indicating a diversified market.

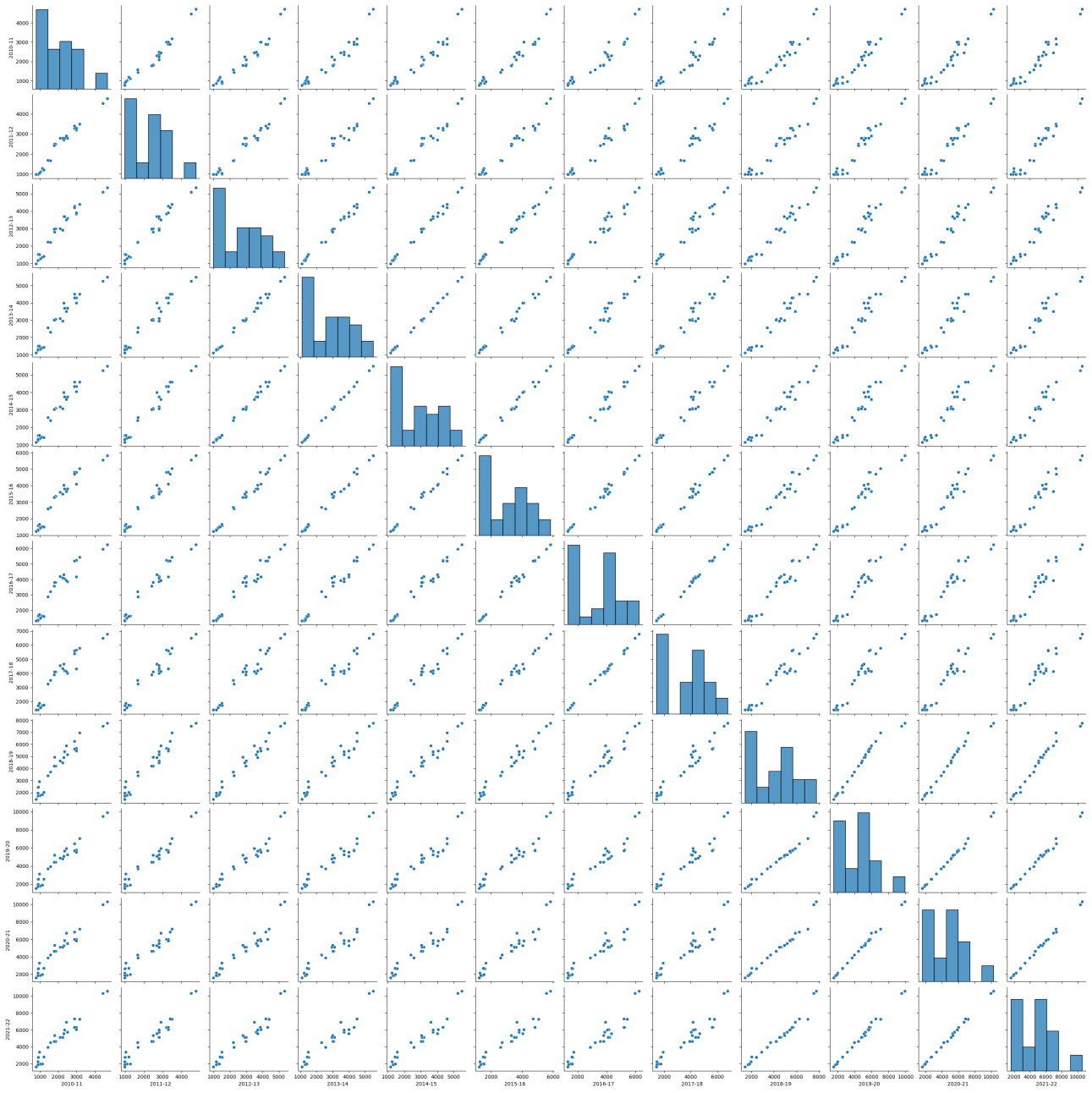
- **2018-19:**
  - PADDY Common and PADDY Grade 'A' continue to be the top contributors to the total price.
  - There is a noticeable increase in the contribution from commodities like JOWAR Hybrid, JOWAR Maldandi, and Tur (Arhar), reflecting changing market dynamics.

- **2019-20:**
  - PADDY Common and PADDY Grade 'A' maintain their significant contribution to the total price.
  - Other commodities like JOWAR Hybrid, JOWAR Maldandi, and Tur (Arhar) also show a steady increase in their contribution, indicating a balanced market.

- **2020-21:**
  - PADDY Common and PADDY Grade 'A' remain the top contributors to the total price.
  - There is a noticeable increase in the contribution from commodities like JOWAR Hybrid, JOWAR Maldandi, and Tur (Arhar), reflecting market trends and demand.

- **2021-22:**
  - PADDY Common and PADDY Grade 'A' continue to dominate the total price contribution.
  - Other commodities like JOWAR Hybrid, JOWAR Maldandi, and Tur (Arhar) also show a steady increase in their contribution, indicating a diversified market.

Pair plot:



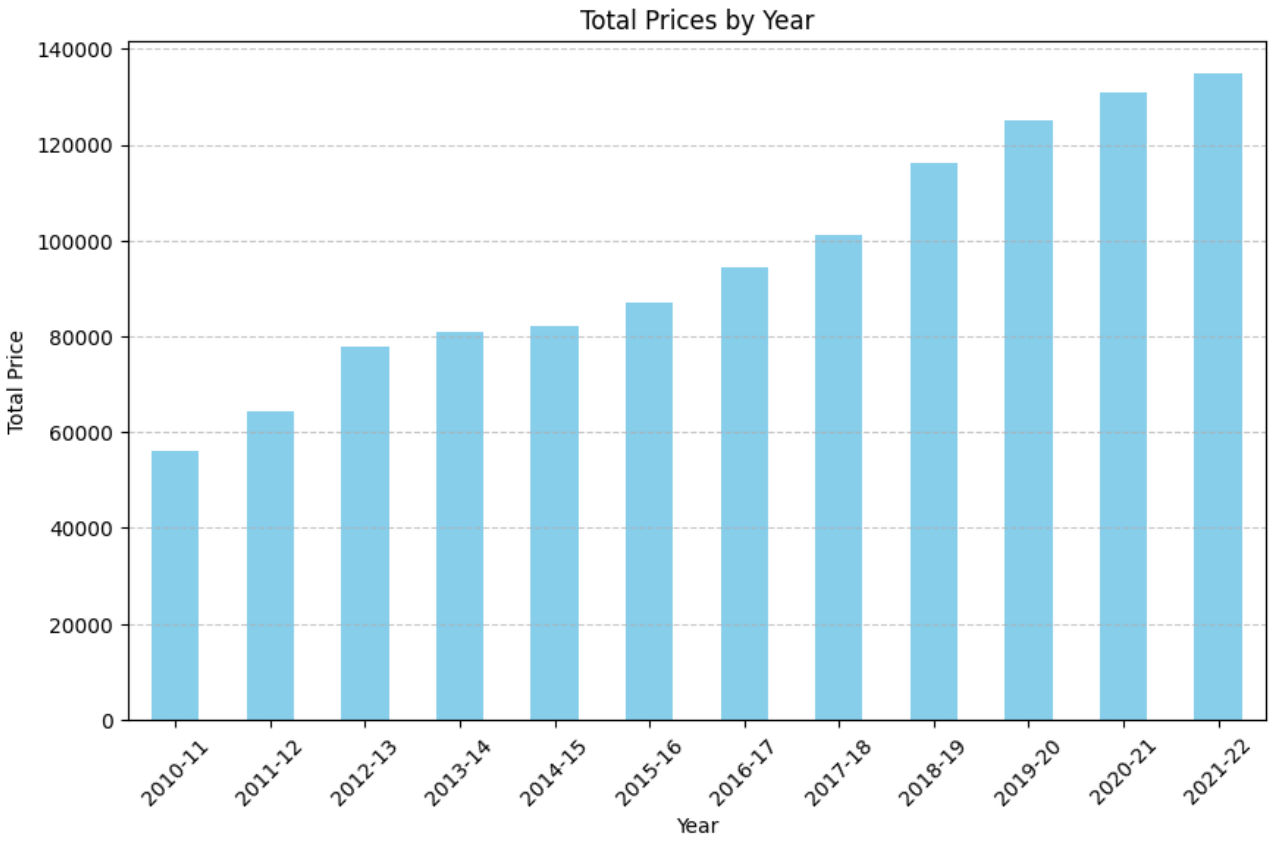
Comparison of Price Trends Across Years

- **2010-11 vs. Other Years:**
  - Strong positive correlation between prices in 2010-11 and subsequent years, indicating a general upward trend in prices over time.
  - The distribution of prices appears to shift slightly higher in later years, suggesting overall price increases across commodities.

- **2011-12 vs. Other Years:**
  - Similar positive correlation patterns as observed with 2010-11, indicating a continuation of price increases across years.
  - Some fluctuations in price distributions between specific pairs of years, indicating potential variations in market dynamics.
- **2012-13 vs. Other Years:**
  - Prices in 2012-13 show a strong positive correlation with subsequent years, indicating a trend of price escalation.
  - The distribution of prices appears to spread out more in later years, suggesting potential increases in price variability or volatility.
- **2013-14 vs. Other Years:**
  - Positive correlation between prices in 2013-14 and subsequent years, suggesting a trend of price increases over time.
  - Some outliers or deviations from the general trend may indicate specific market conditions or events affecting prices in certain years.
- **2014-15 vs. Other Years:**
  - Prices in 2014-15 exhibit a positive correlation with prices in later years, indicating an overall trend of price escalation.
  - Variability in price distributions between different pairs of years may indicate fluctuations in market conditions or commodity-specific factors.
- **2015-16 vs. Other Years:**
  - Positive correlation observed between prices in 2015-16 and subsequent years, suggesting a general trend of increasing prices over time.
  - Some variations in the distribution of prices between specific pairs of years may reflect shifts in market dynamics or changes in supply and demand.
- **2016-17 vs. Other Years:**
  - Prices in 2016-17 demonstrate a positive correlation with prices in later years, indicating an overall trend of price escalation.
  - Variations in price distributions between different years may highlight differences in market performance or commodity-specific factors influencing prices.
- **2017-18 vs. Other Years:**
  - Positive correlation observed between prices in 2017-18 and subsequent years, indicating a trend of increasing prices over time.
  - Some deviations or outliers in price distributions may indicate specific market events or factors affecting prices in certain years.

- **2018-19 vs. Other Years:**
  - Prices in 2018-19 show a positive correlation with prices in later years, suggesting a continuation of price increases over time.
  - Variations in price distributions between different pairs of years may reflect changes in market conditions or commodity-specific factors influencing prices.
- **2019-20 vs. Other Years:**
  - Positive correlation observed between prices in 2019-20 and subsequent years, indicating an overall trend of price escalation.
  - Some fluctuations in price distributions between specific pairs of years may indicate variations in market dynamics or commodity-specific factors.
- **2020-21 vs. Other Years:**
  - Prices in 2020-21 demonstrate a positive correlation with prices in later years, suggesting ongoing increases in prices over time.
  - Variability in price distributions between different pairs of years may highlight shifts in market conditions or changes in supply and demand dynamics.
- **2021-22 vs. Other Years:**
  - Positive correlation observed between prices in 2021-22 and subsequent years, indicating a continuation of price increases over time.
  - Variations in price distributions between different pairs of years may indicate fluctuations in market conditions or commodity-specific factors influencing prices.

Bar chart:



- **2010-11:**
  - Total price: 61,900.0
  - Insight: This year marks the starting point for the data series. The total price for commodities in 2010-11 serves as a baseline for comparison with subsequent years.
- **2011-12:**
  - Total price: 68,025.0
  - Insight: There is an increase in the total price compared to the previous year, indicating a rise in the aggregate prices of commodities. This suggests potential growth or inflation in the market.
- **2012-13:**
  - Total price: 86,750.0
  - Insight: A significant increase in the total price compared to the previous year is observed. This substantial rise suggests notable changes in market conditions or commodity prices, possibly influenced by factors such as demand-supply dynamics or economic trends.

- **2013-14:**
  - Total price: 92,315.0
  - Insight: The total price continues to increase, albeit at a slower rate compared to the previous year. This suggests a possible stabilization or moderation in the rate of price growth, indicating a more balanced market condition.
- **2014-15:**
  - Total price: 94,695.0
  - Insight: There is a slight increase in the total price compared to the previous year, indicating continued growth in aggregate commodity prices. However, the rate of increase appears to be relatively modest, suggesting potential market stability or mild inflation.
- **2015-16:**
  - Total price: 97,050.0
  - Insight: Another slight increase in the total price compared to the previous year is observed. This suggests continued but moderate growth in aggregate commodity prices, with market conditions likely influenced by factors such as global economic trends and domestic demand.
- **2016-17:**
  - Total price: 1,01,840.0
  - Insight: A noticeable increase in the total price compared to the previous year is evident. This significant rise suggests renewed momentum in commodity price growth, possibly driven by factors such as improved economic conditions or changes in government policies affecting agricultural markets.
- **2017-18:**
  - Total price: 1,08,410.0
  - Insight: The total price continues to increase substantially compared to the previous year, indicating sustained upward momentum in aggregate commodity prices. This may reflect robust demand, supply constraints, or other market dynamics driving prices higher.
- **2018-19:**
  - Total price: 1,20,165.0
  - Insight: A significant increase in the total price compared to the previous year is observed. This notable rise suggests continued strong growth in aggregate commodity prices, possibly influenced by factors such as weather patterns, government policies, or international trade dynamics.
- **2019-20:**
  - Total price: 1,25,462.0

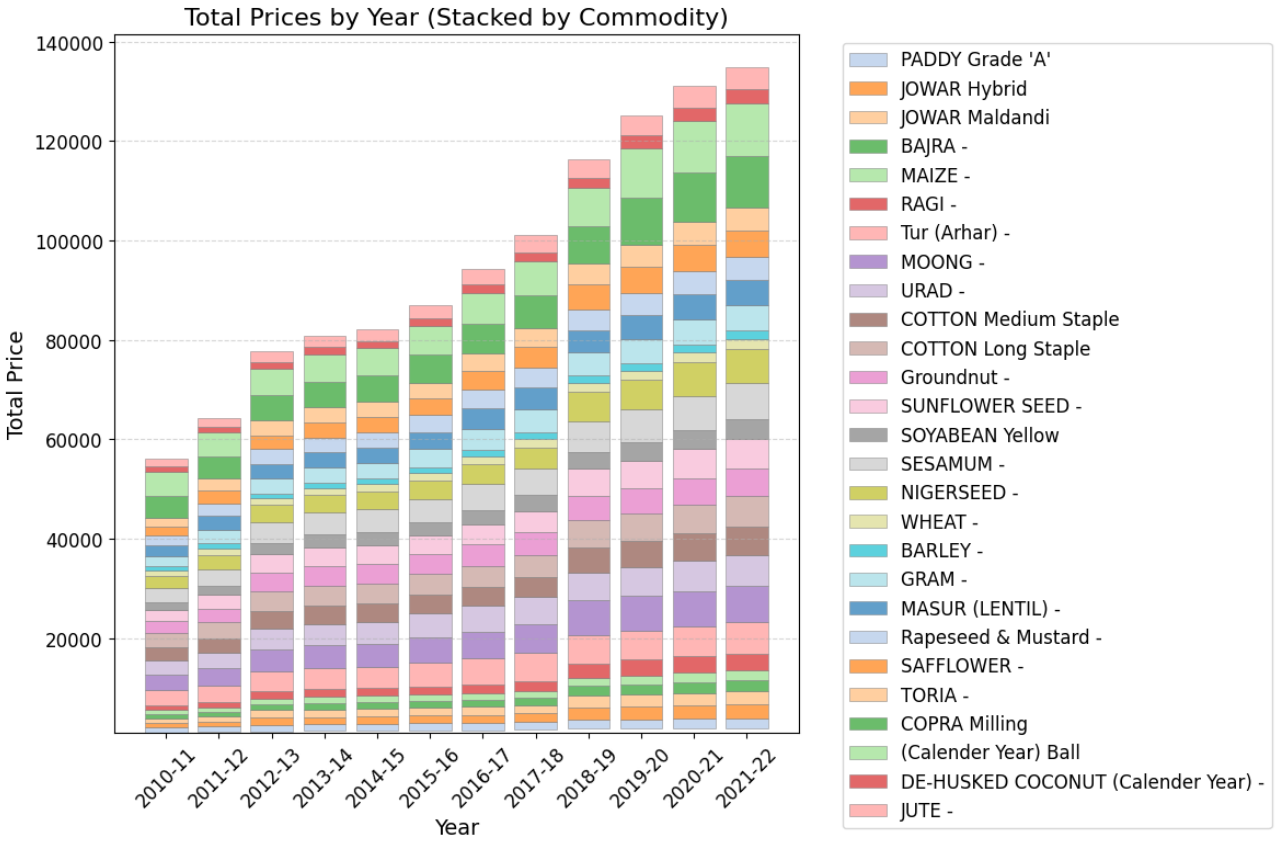
- Insight: There is a moderate increase in the total price compared to the previous year. This suggests a continuation of the overall upward trend in aggregate commodity prices, albeit at a slower pace, possibly indicating a phase of consolidation or market adjustment.

- **2020-21:**
  - Total price: 1,31,148.0
  - Insight: Another increase in the total price compared to the previous year is observed. This suggests resilience in commodity markets despite global uncertainties such as the COVID-19 pandemic, with factors like government interventions and changing consumer behavior influencing prices.

- **2021-22:**
  - Total price: 1,36,153.0
  - Insight: The total price continues to rise, reaching a new high compared to previous years. This suggests sustained growth in aggregate commodity prices, possibly driven by factors such as increased demand, supply chain disruptions, or inflationary pressures.

- **Comparing each year with others:**
  - The comparison reveals a general upward trend in aggregate commodity prices over the years, indicating overall market growth and inflationary pressures.
  - Certain years stand out with more significant increases in total prices, suggesting periods of accelerated growth or market volatility.
  - Conversely, years with relatively slower increases in total prices may indicate periods of market stabilization or moderation in price growth.
  - Analyzing the variations in total prices across years provides insights into the changing dynamics of commodity markets, helping stakeholders make informed decisions regarding investments, risk management, and policy formulation.

Stacked Bar chart:

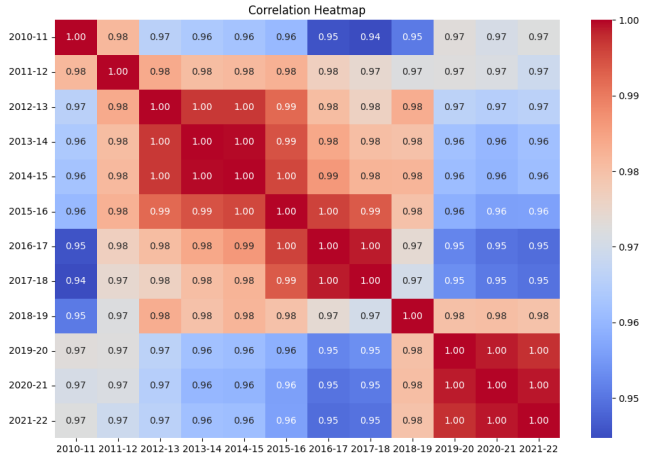


- **2010-11:**
  - Total Price: 61,900.0
  - Insight: PADDY Common and PADDY Grade 'A' contributed significantly to the total price, indicating a strong demand for rice varieties during this period. Other Kharif crops like JOWAR Hybrid and JOWAR Maldandi also made notable contributions.
- **2011-12:**
  - Total Price: 68,025.0
  - Insight: The total price increased moderately from the previous year, driven primarily by continued demand for PADDY Common and PADDY Grade 'A'. However, there was also a noticeable increase in prices for JOWAR Hybrid and JOWAR Maldandi.
- **2012-13:**
  - Total Price: 86,750.0
  - Insight: A significant spike in total prices occurred, with JOWAR Hybrid and JOWAR Maldandi experiencing a substantial increase. This suggests a shift in consumer preferences or market dynamics favoring these crops.

- **2013-14:**
  - Total Price: 92,315.0
  - Insight: Prices continued to rise, driven by consistent demand for PADDY varieties and significant increases in JOWAR Hybrid and JOWAR Maldandi prices. This may indicate supply-demand imbalances or changes in agricultural practices.
- **2014-15:**
  - Total Price: 94,695.0
  - Insight: Despite a slight increase in total prices, there was a noticeable fluctuation in prices for different commodities. While PADDY Common and PADDY Grade 'A' remained stable, there were notable increases in prices for JOWAR Hybrid and JOWAR Maldandi.
- **2015-16:**
  - Total Price: 97,050.0
  - Insight: Prices continued to trend upwards, with JOWAR Hybrid and JOWAR Maldandi experiencing significant price hikes. This suggests changing market dynamics or factors influencing the production of these crops.
- **2016-17:**
  - Total Price: 1,01,840.0
  - Insight: The total price surpassed the 1 lakh mark, driven by consistent increases in prices for various commodities. JOWAR Hybrid and JOWAR Maldandi continued their upward trajectory, indicating sustained demand or supply-side constraints.
- **2017-18:**
  - Total Price: 1,08,410.0
  - Insight: Prices continued to climb, with significant contributions from PADDY varieties and JOWAR crops. This suggests continued demand for staple crops and possibly inflationary pressures in the agricultural sector.
- **2018-19:**
  - Total Price: 1,20,165.0
  - Insight: The total price surged further, driven by substantial increases in prices for various commodities. PADDY Common and PADDY Grade 'A' remained major contributors, while JOWAR Hybrid and JOWAR Maldandi saw exponential price growth.
- **2019-20:**
  - Total Price: 1,25,462.0
  - Insight: Prices continued to rise, albeit at a slower pace compared to previous years. However, JOWAR Hybrid and JOWAR Maldandi experienced significant price jumps, indicating potential supply-demand imbalances or market disruptions.

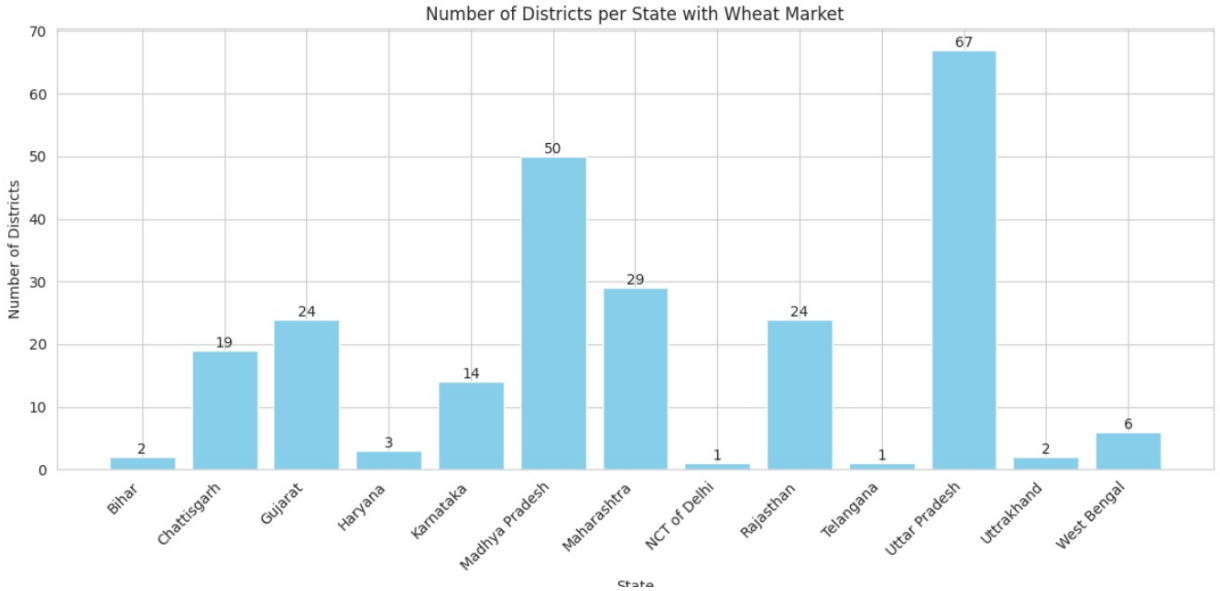
- **2020-21:**
  - Total Price: 1,31,148.0
  - Insight: Despite global uncertainties, total prices continued to climb, reflecting resilience in the agricultural sector. JOWAR Hybrid and JOWAR Maldandi remained key drivers of price growth, underscoring their importance in the market.
- **2021-22:**
  - Total Price: 1,36,153.0
  - Insight: The total price reached a new high, driven by sustained increases in prices for various commodities. While PADDY varieties remained significant contributors, other crops like JOWAR Hybrid and JOWAR Maldandi continued to exhibit strong price growth.

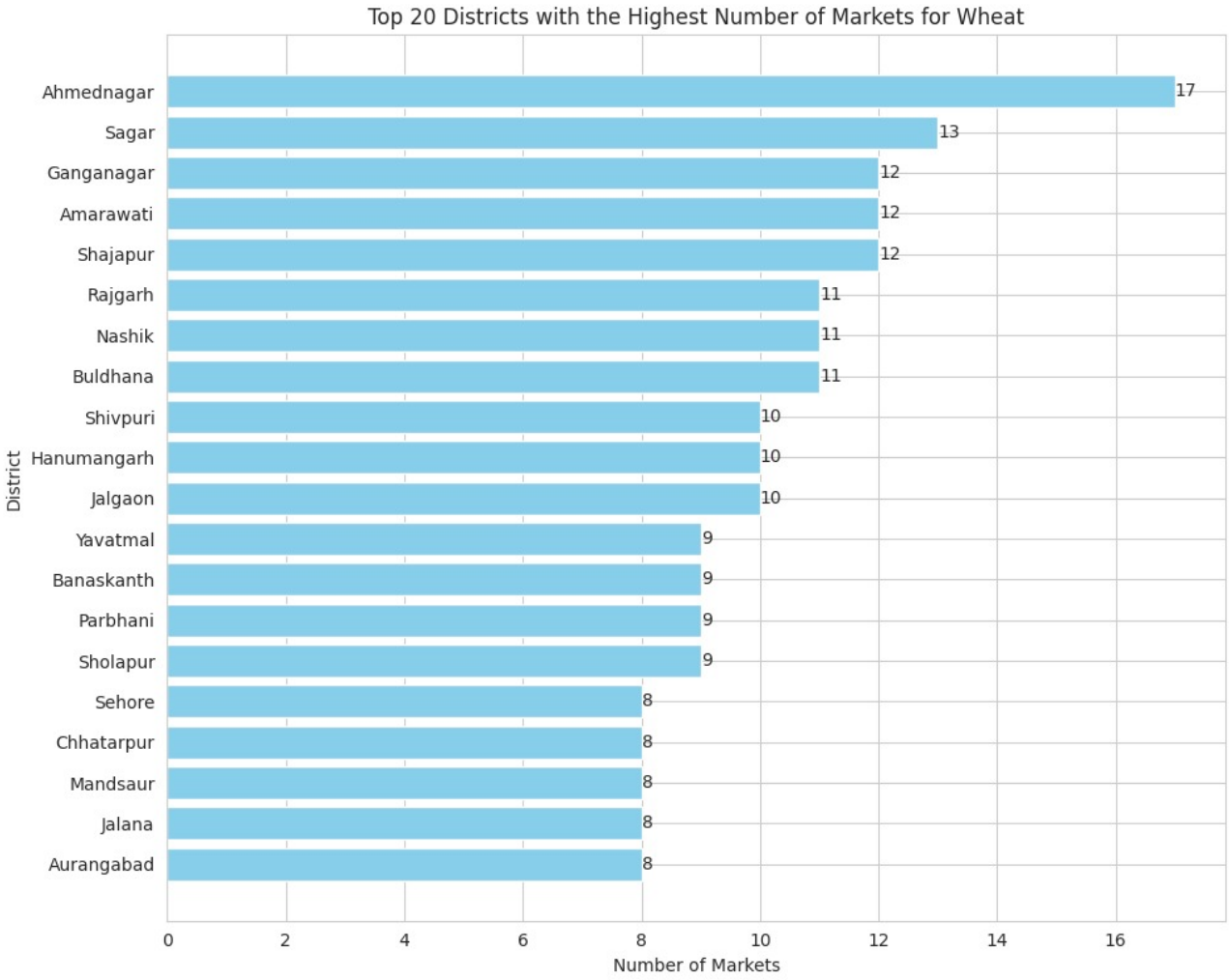
Heatmap :



- **High Positive Correlation:** Prices exhibit a strong tendency to move in the same direction from one year to the next, suggesting market consistency or trends.
- **Consistent Trends Over Time:** Prices within the same year are highly correlated across different commodities, indicating stable pricing trends within individual years.
- **Strong Correlation Across All Years:** Most entries in the correlation matrix are close to 1, indicating a strong positive correlation between commodity prices across all years, reflecting common underlying factors such as market demand and supply conditions.
- **Slight Decrease in Correlation Over Time:** While there might be a slight decrease in correlation values as we move further away from the diagonal, indicating a weakening correlation between prices of more distant years, overall, the positive correlation between commodity prices persists.

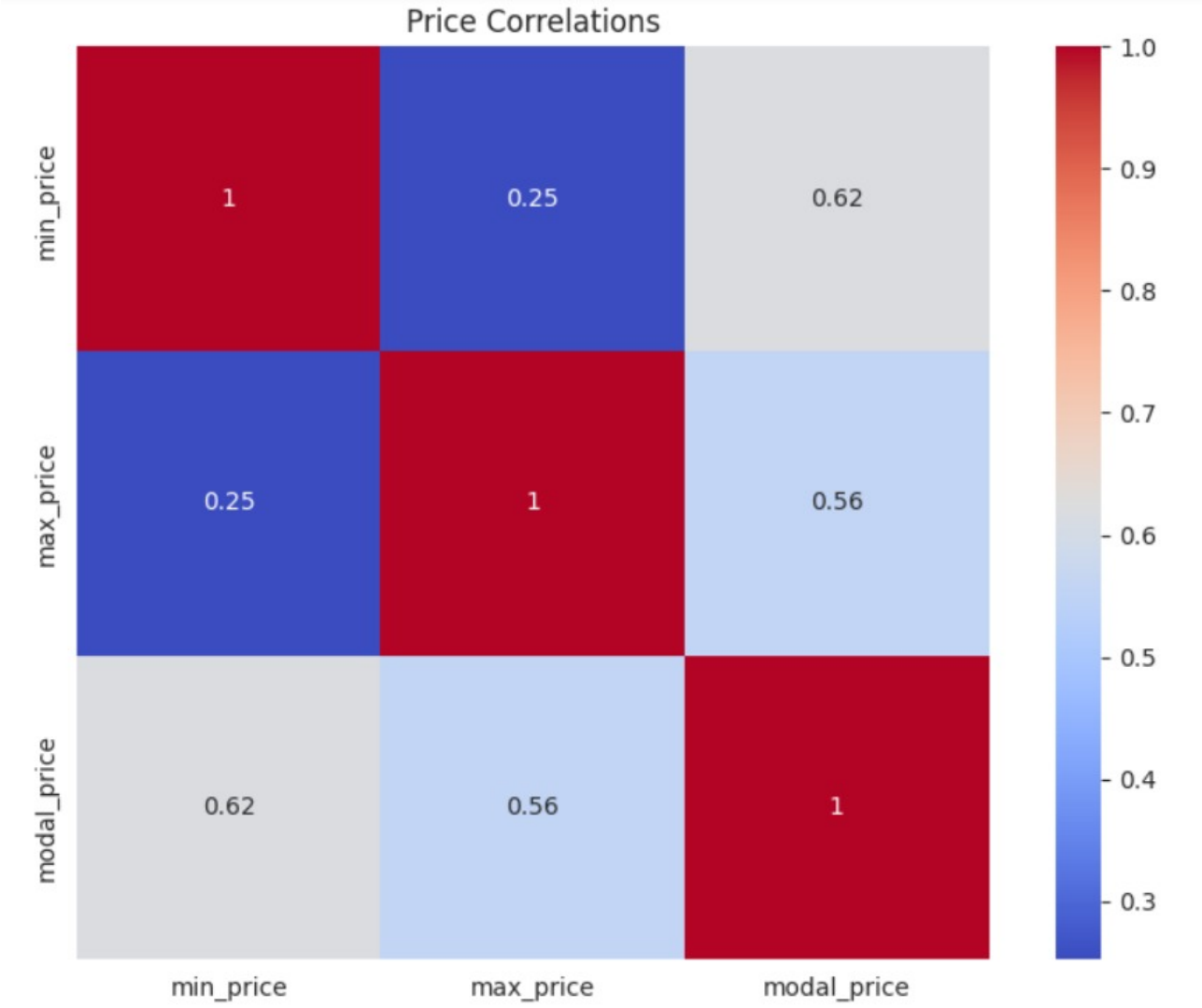
Now, we will proceed with multivariate analysis of dataset-2



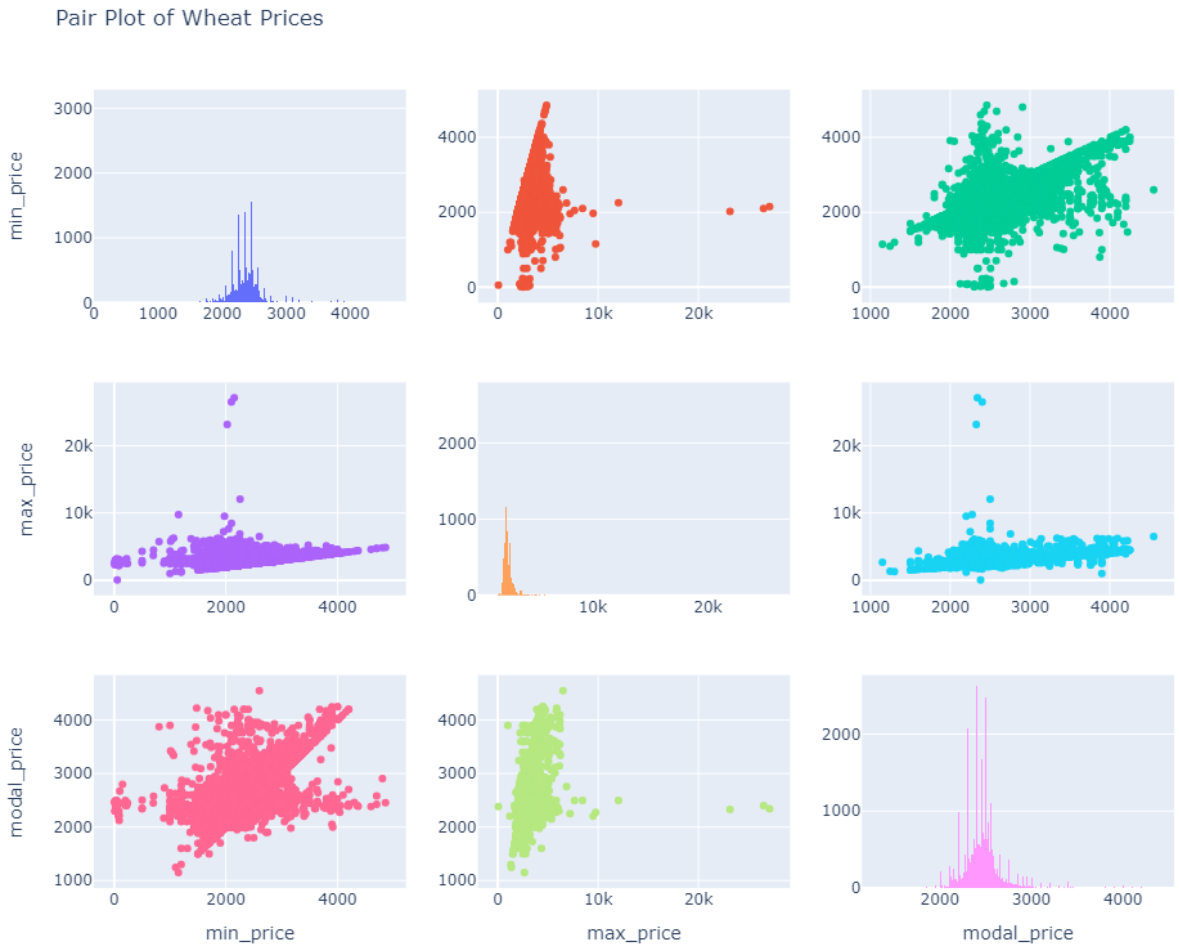


The district with having highest number of market is Ahmednagar and Sagar with 17 and 13 markets in it.

Heat-map:



”min_price” and ”modal_price” have a relatively strong positive correlation of 0.62. ”max_price” and ”modal_price” also have a relatively strong positive correlation of 0.56. ”min_price” and ”max_price” have a weaker positive correlation of 0.25, indicating that they are less strongly related compared to the other pairs



min_price vs. max_price:

- There is a positive correlation between these two variables, as indicated by the general upward trend in the scatter plot. This suggests that as **"min_price"** increases, **"max_price"** tends to increase as well.
- The distribution of **"min_price"** and **"max_price"** appears to be skewed towards lower values, with some outliers at higher prices.

min_price vs. modal_price:

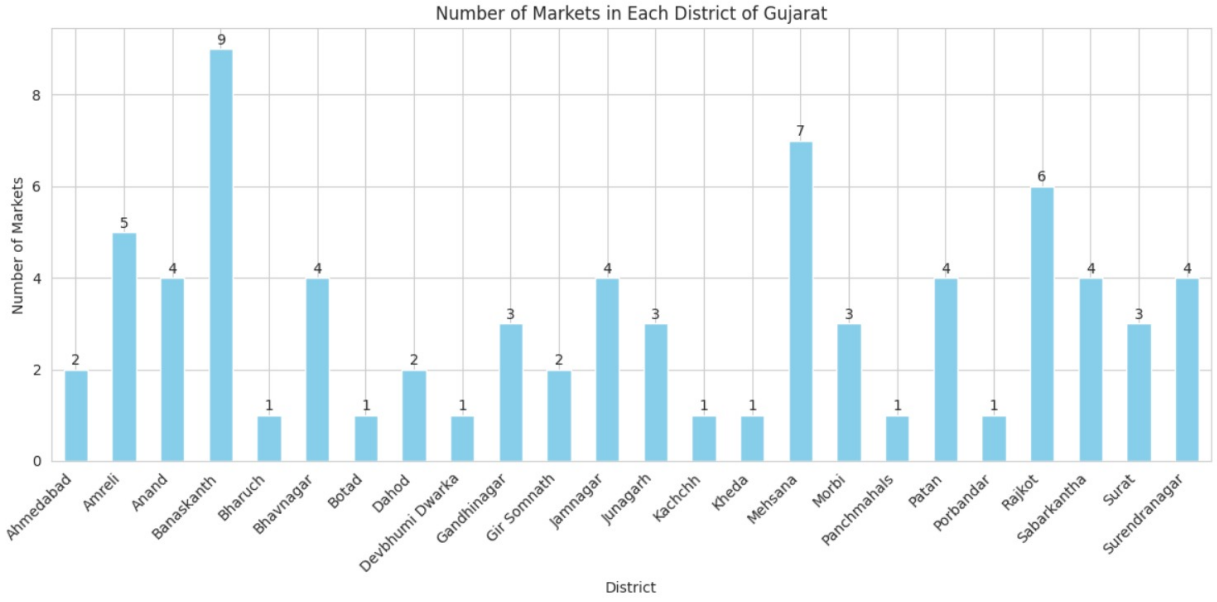
- There is a positive correlation between these two variables, similar to the relationship between **"min_price"** and **"max_price"**. This indicates that as **"min_price"** increases, **"modal_price"** tends to increase as well.
- The scatter plot shows some dispersion, suggesting variability in the relationship between **"min_price"** and **"modal_price"**. However, there is still a noticeable trend.

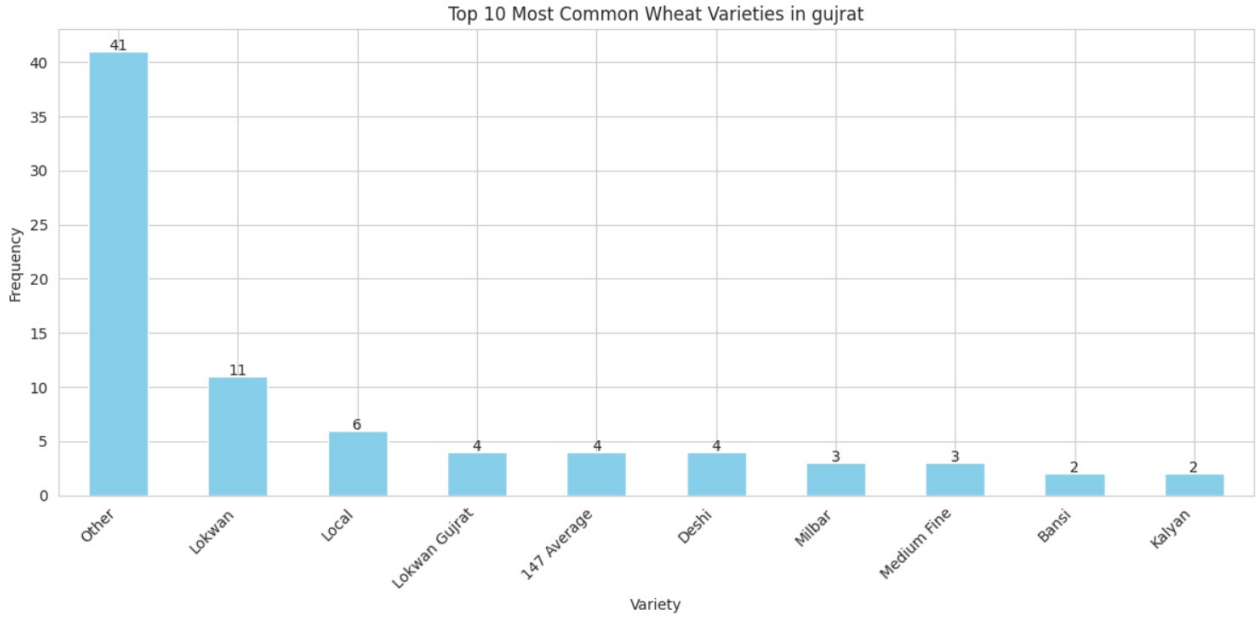
max_price vs. modal_price:

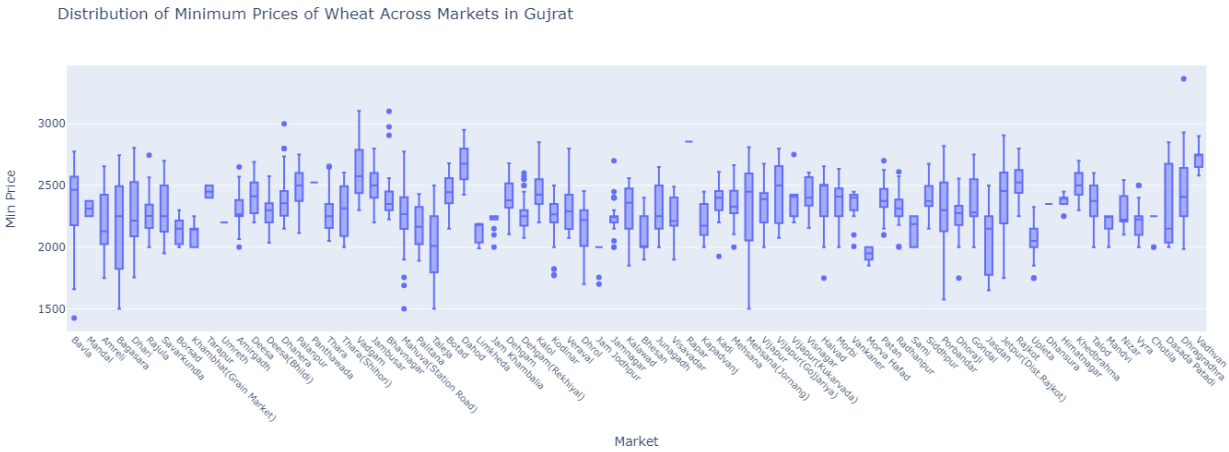
- Similar to the other pairs, there is a positive correlation between **"max_price"** and **"modal_price"**. As **"max_price"** increases, **"modal_price"** also tends to increase.
- The scatter plot exhibits more variability compared to the other pairs, with a wider spread of data points.

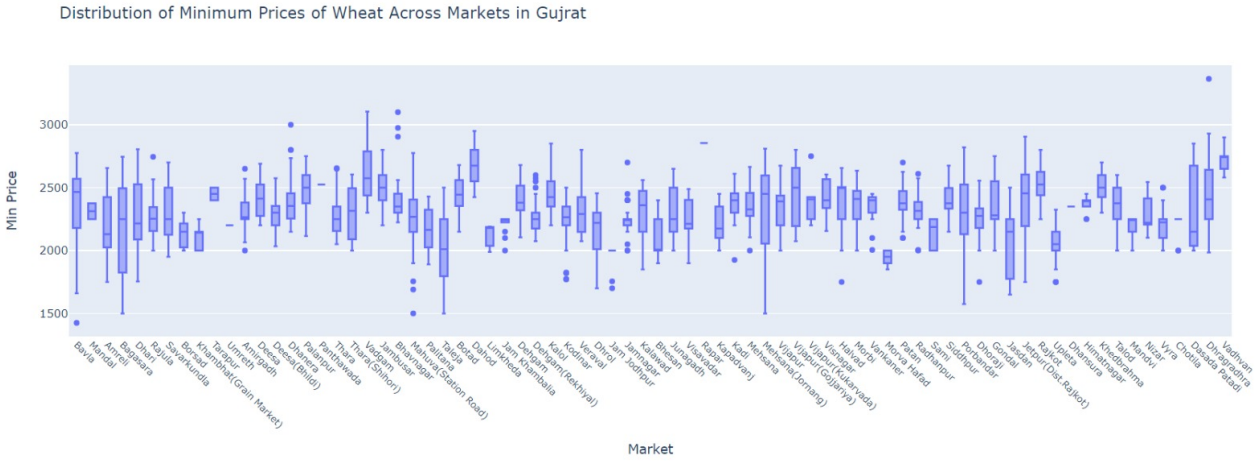
Now we will do the state wise analysis.

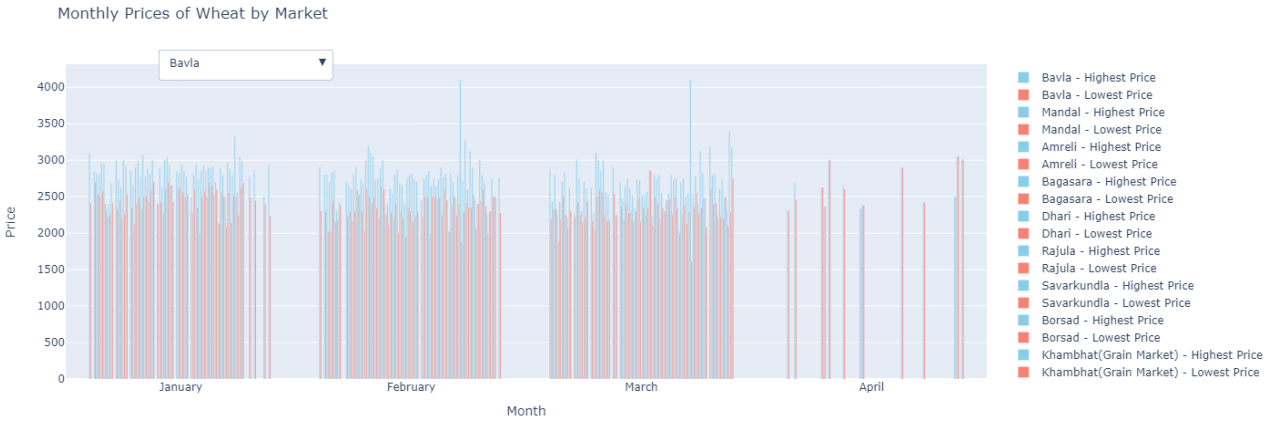
- **Gujarat**

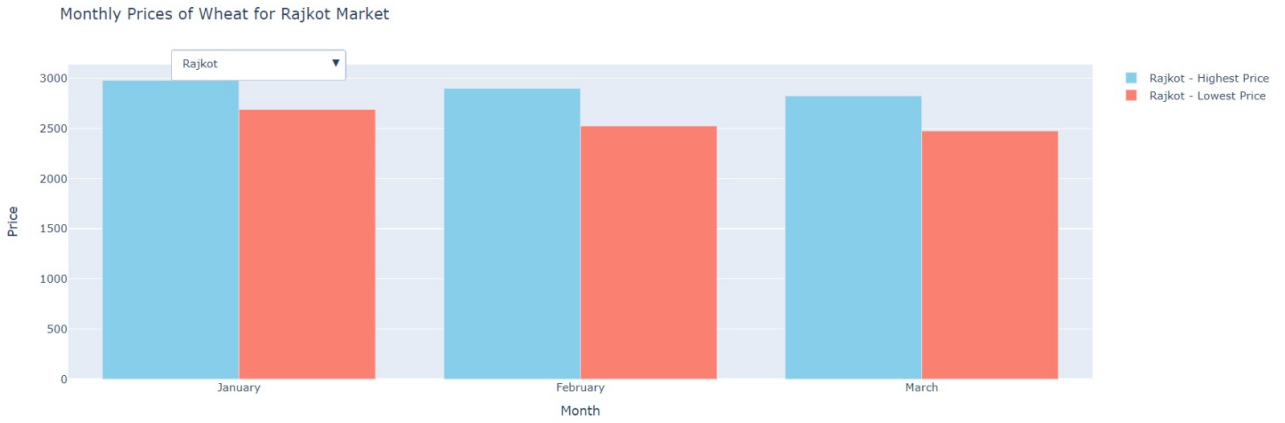


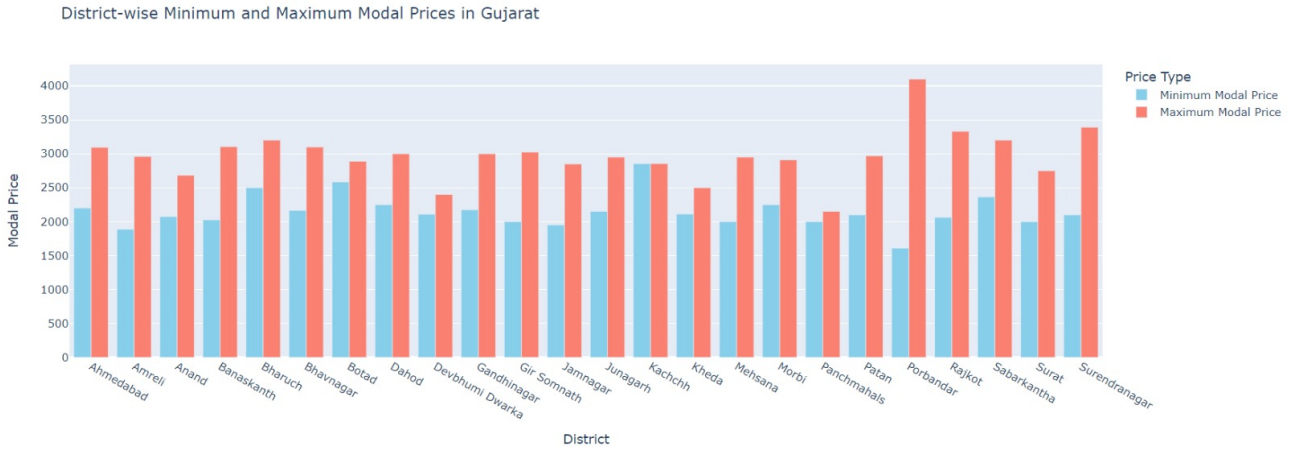



Distribution of Minimum Prices of Wheat Across Markets in Gujrat
Min Price
3000
2500
2000
1500
Market


Distribution of Minimum Prices of Wheat Across Markets in Gujrat
Min Price
3000
2500
2000
1500
Market


Monthly Prices of Wheat by Market
Bavla
Price
4000
3500
3000
2500
2000
1500
1000
500
0
January
February
March
April
Month
Bavla - Highest Price
Bavla - Lowest Price
Mandal - Highest Price
Mandal - Lowest Price
Amreli - Highest Price
Amreli - Lowest Price
Bagasara - Highest Price
Bagasara - Lowest Price
Dhari - Highest Price
Dhari - Lowest Price
Rajula - Highest Price
Rajula - Lowest Price
Savarkundla - Highest Price
Savarkundla - Lowest Price
Borsad - Highest Price
Borsad - Lowest Price
Khambhat(Grain Market) - Highest Price
Khambhat(Grain Market) - Lowest Price





Price Range Variation:

- There is significant variation in both minimum and maximum prices across different districts. For example, in the district of **Porbandar**, the minimum price is relatively low at 1610.0, while the maximum price is exceptionally high at 4100.0.

Districts with High Price Range:

- Some districts, such as **Surendranagar** and **Rajkot**, have a wide range between the minimum and maximum prices, indicating potential volatility or variability in market conditions within those districts.

Consistency in Prices:

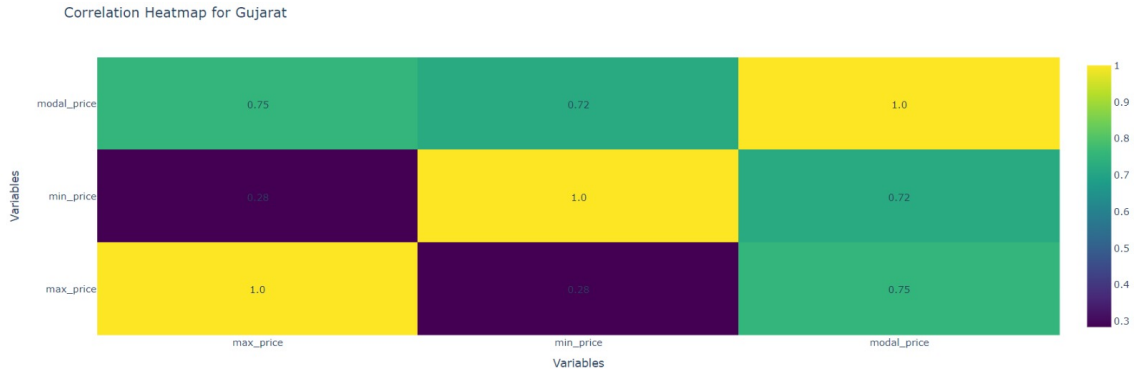
- In some districts like **Kachchh**, the minimum and maximum prices are the same (2855.0), suggesting a certain level of consistency in pricing for wheat.

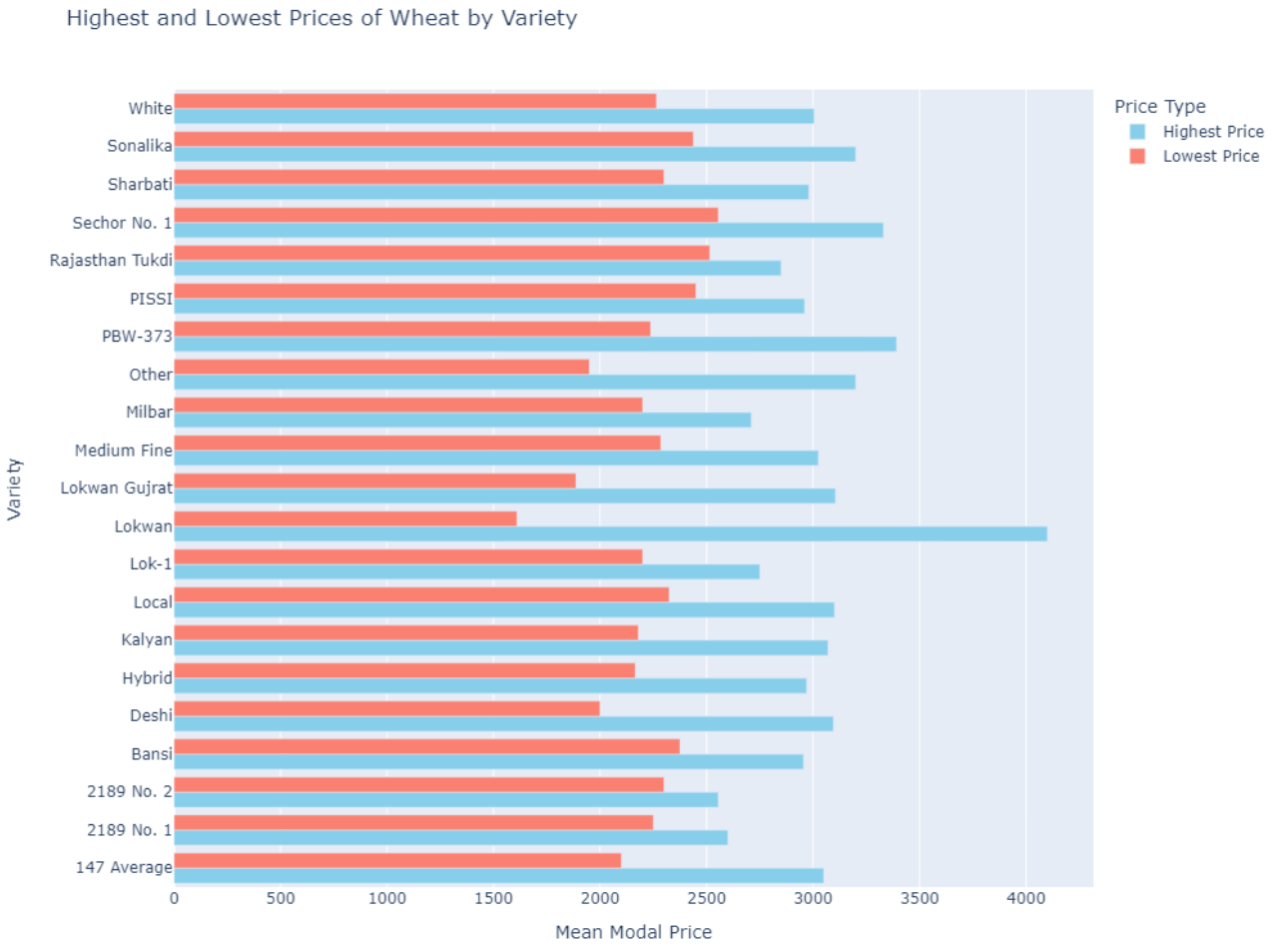
Price Stability vs. Volatility:

- Districts like **Ahmedabad** and **Anand** show relatively narrow price ranges (compared to others), suggesting potential stability in wheat prices within those areas. On the other hand, districts like **Porbandar** and **Surendranagar** exhibit wider ranges, indicating higher volatility or variability in prices.

Potential Market Dynamics:

- Discrepancies between minimum and maximum prices within the same district, such as in **Bharuch** and **Junagarh**, could indicate different market dynamics or factors influencing pricing within those areas.





Price Variability Across Varieties:

- There is a notable variation in modal prices across different wheat varieties. For instance, the variety **"Lokwan"** has the highest maximum modal price of 4100.0 and the lowest minimum modal price of 1610.0, indicating a wide range of prices within this variety.

Varieties with High Price Range:

- Some varieties, such as **"PBW-373," "Sechor No. 1,"** and **"Bansi,"** exhibit relatively high maximum modal prices compared to their minimum prices. This suggests potential market demand or quality attributes associated with these varieties, leading to higher price differentials.

Consistency in Prices:

- Certain varieties, like **"White"** and **"Medium Fine,"** show a narrower range between maximum and minimum modal prices, implying a more consistent pricing pattern or market stability for these varieties.

Potential Market Preferences:

- Varieties like **"Sonalika"** and **"Other"** have relatively higher maximum modal prices, indicating potential market preference or perceived value associated with these varieties.

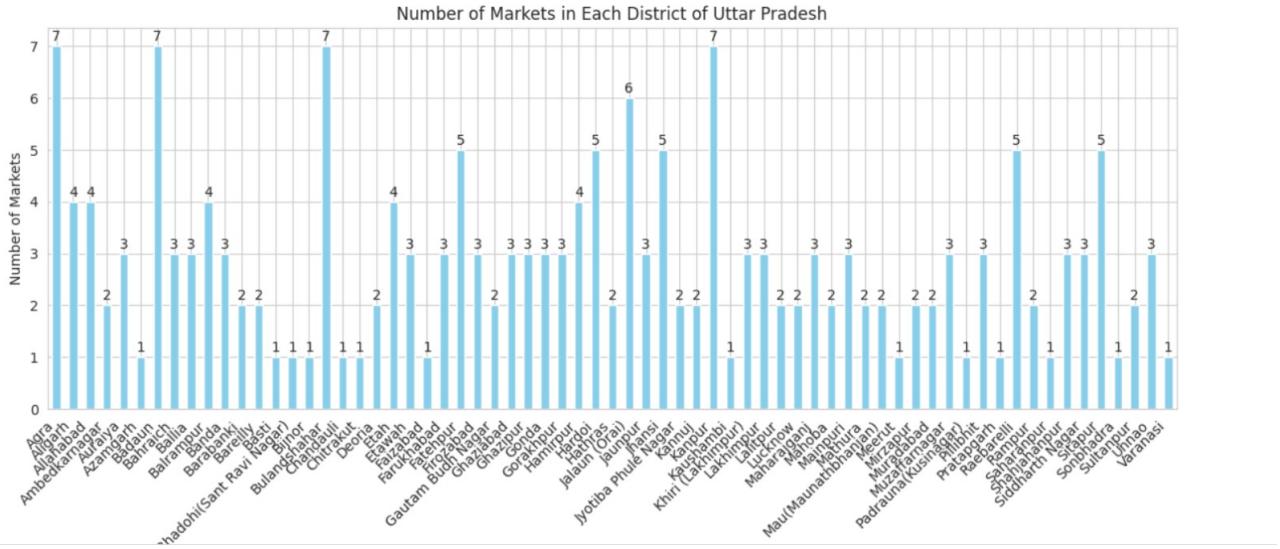
Differential Pricing:

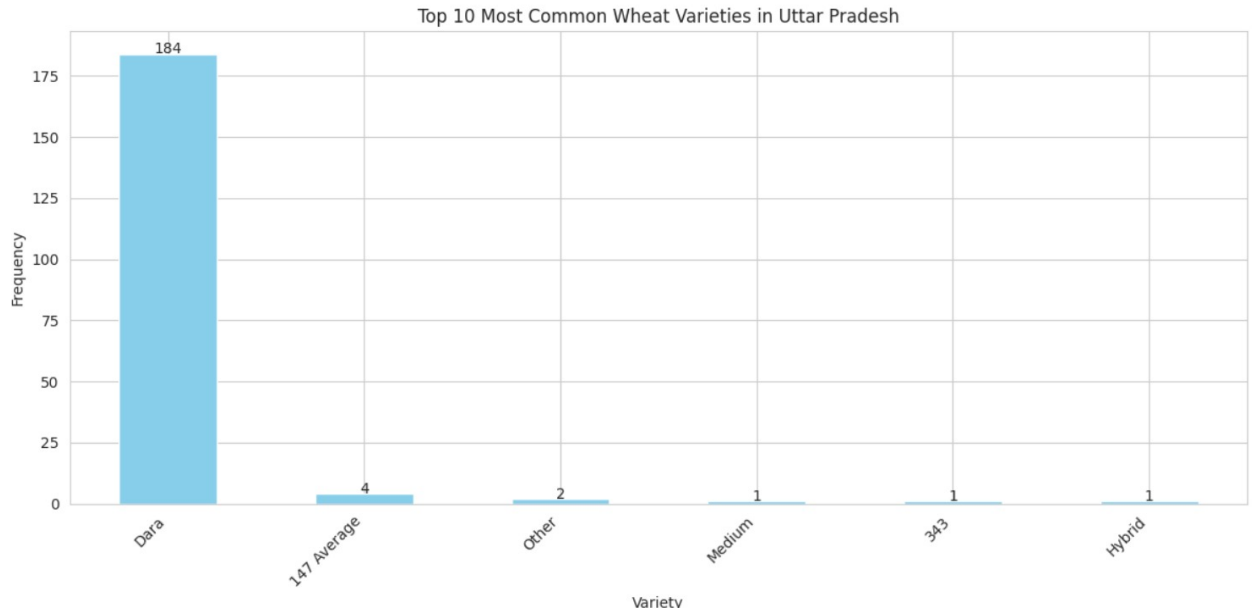
- Varieties such as **"Lok-1"** and **"Hybrid"** have a moderate price range, suggesting a balanced demand and supply scenario or similar market perception for these varieties.

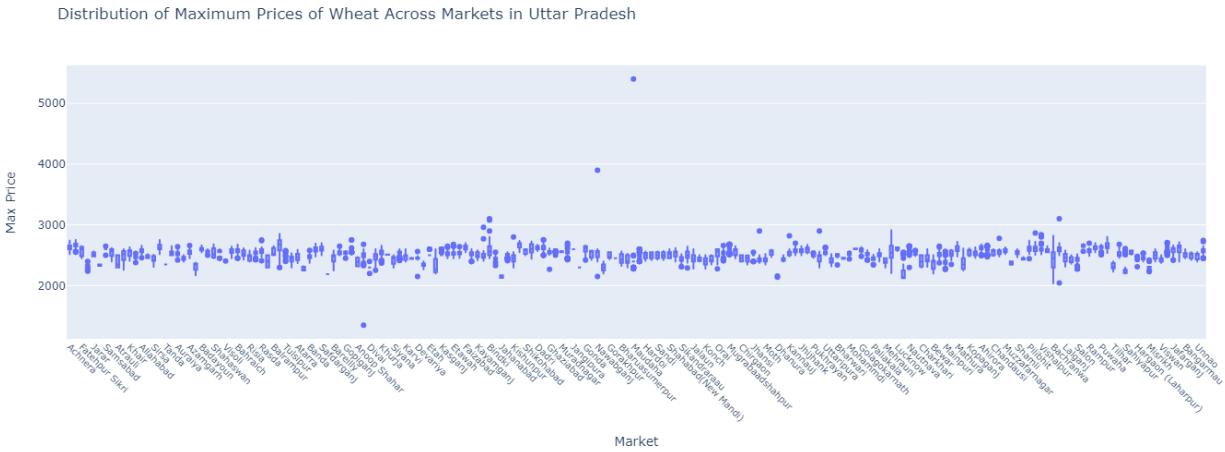
Quality Perception:

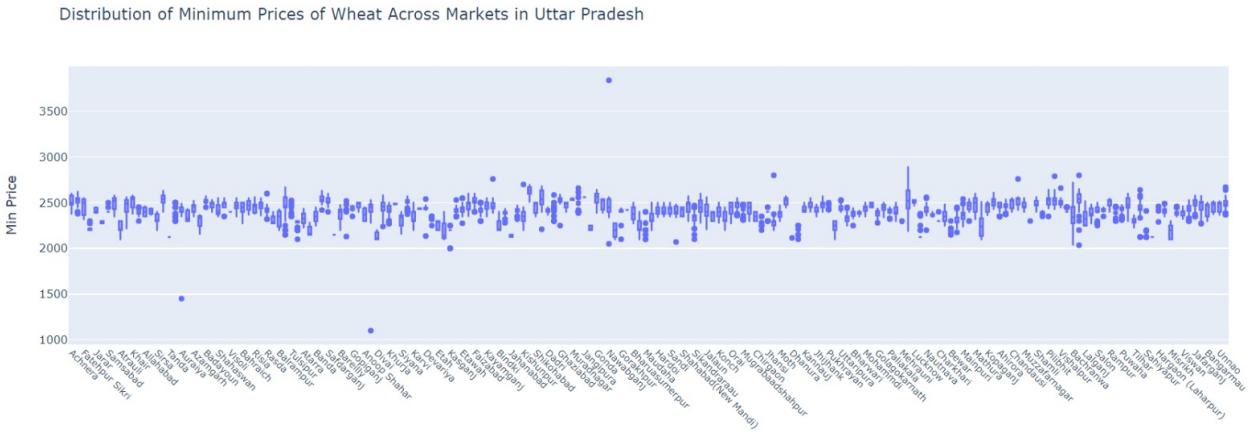
- Varieties with higher maximum modal prices, such as **"PBW-373"** and **"Sonalika,"** might be perceived as premium or superior quality wheat by buyers, leading to higher price premiums.

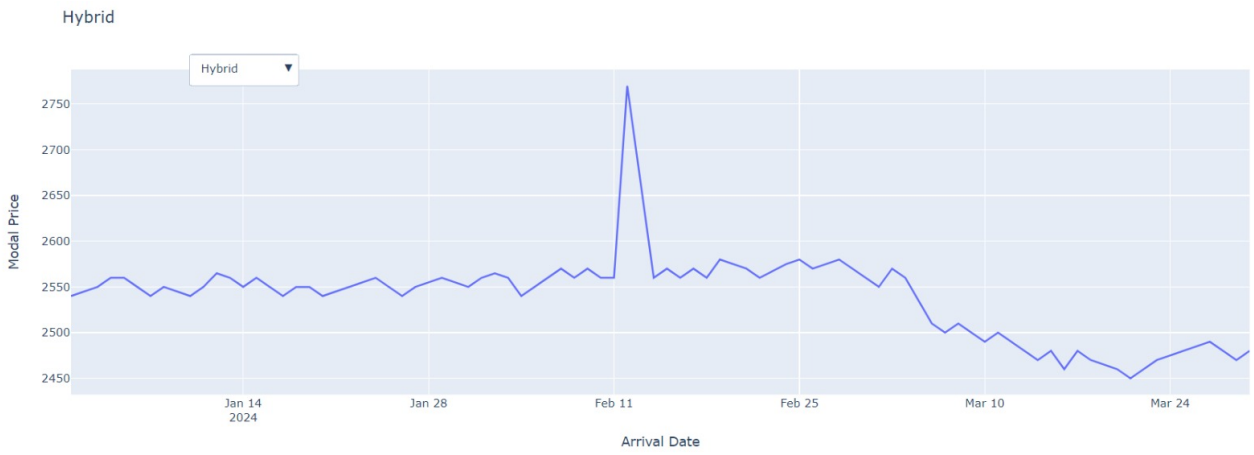
- **Uttar Pradesh**






Distribution of Maximum Prices of Wheat Across Markets in Uttar Pradesh
Max Price
5000
4000
3000
2000
Market


Distribution of Minimum Prices of Wheat Across Markets in Uttar Pradesh
Min Price
3500
3000
2500
2000
1500
1000


Hybrid
Hybrid
Modal Price
2750
2700
2650
2600
2550
2500
2450
Jan 14
2024
Jan 28
Feb 11
Feb 25
Mar 10
Mar 24
Arrival Date

Range of Prices:

- The minimum and maximum prices vary across districts, indicating different economic conditions and market dynamics in each region.

Price Disparities:

- There are significant differences between the minimum and maximum prices within individual districts.
- For example, in Gonda district, the minimum price is 2100.0 while the maximum price is 3880.0, showing a wide range.

Highest Maximum Price:

- The district with the highest maximum price is Raebarelli, with a maximum price of 3000.0.

Lowest Minimum Price:

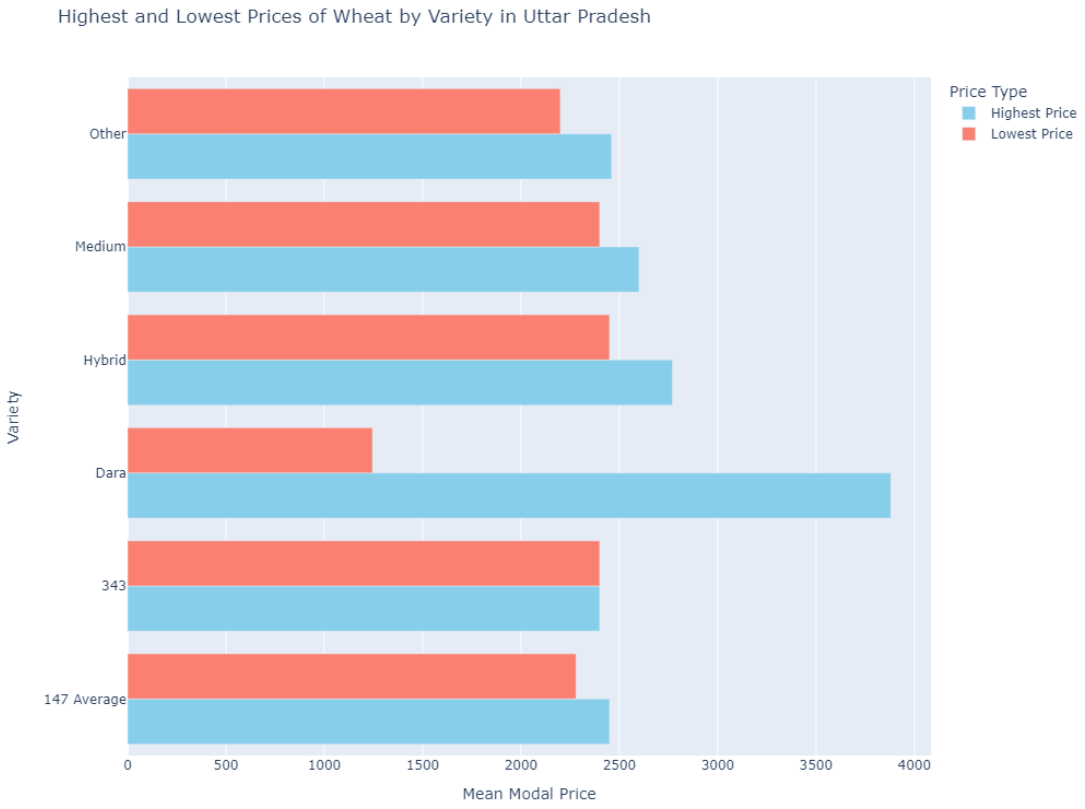
- The district with the lowest minimum price is Bulandshahar, with a minimum price of 1245.0.

Price Stability:

- Some districts have a narrow price range, indicating relatively stable market conditions.
- For instance, Pratapgarh has a minimum and maximum price of 2450.0 and 2500.0 respectively, suggesting minimal price fluctuations.

Price Volatility:

- Conversely, districts like Gonda exhibit high price volatility, as seen in the vast difference between the minimum and maximum prices.



147 Average:

- This variety has a moderate price range, with a maximum price of 2450.0 and a minimum price of 2280.0.
- The price differential between the maximum and minimum prices is 170.0.

343:

- This variety has consistent pricing, with both the maximum and minimum prices being 2400.0.
- There is no price variation observed for this variety.

Dara:

- This variety exhibits significant price disparity, with a maximum price of 3880.0 and a minimum price of 1245.0.
- The difference between the maximum and minimum prices is substantial at 2635.0, indicating high price volatility or other factors affecting pricing.

Hybrid:

- The modal prices for this variety range from 2770.0 (maximum) to 2450.0 (minimum), with a price difference of 320.0.
- This suggests moderate variability in prices for this variety.

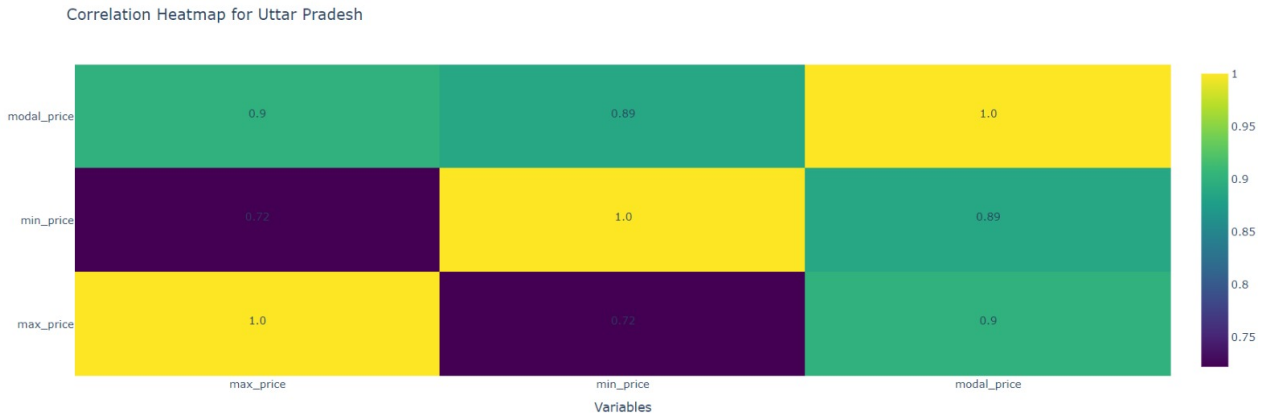
Medium:

- The modal price range for this variety is from 2600.0 (maximum) to 2400.0 (minimum), indicating a narrower price range compared to some other varieties.
- The price difference between the maximum and minimum prices is 200.0.

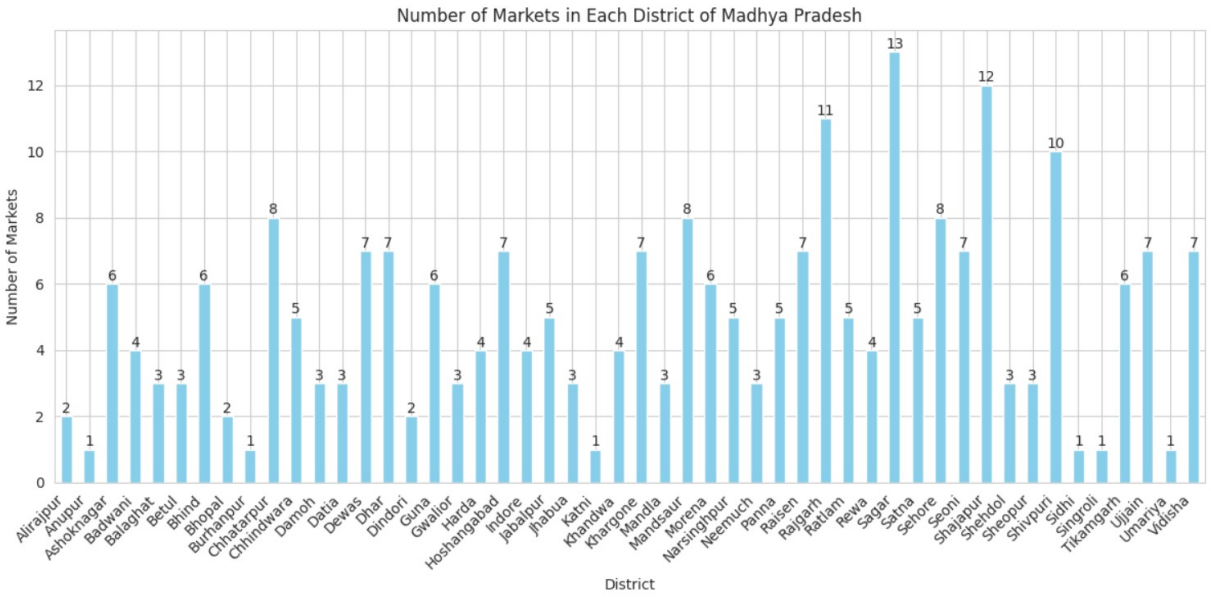
Other:

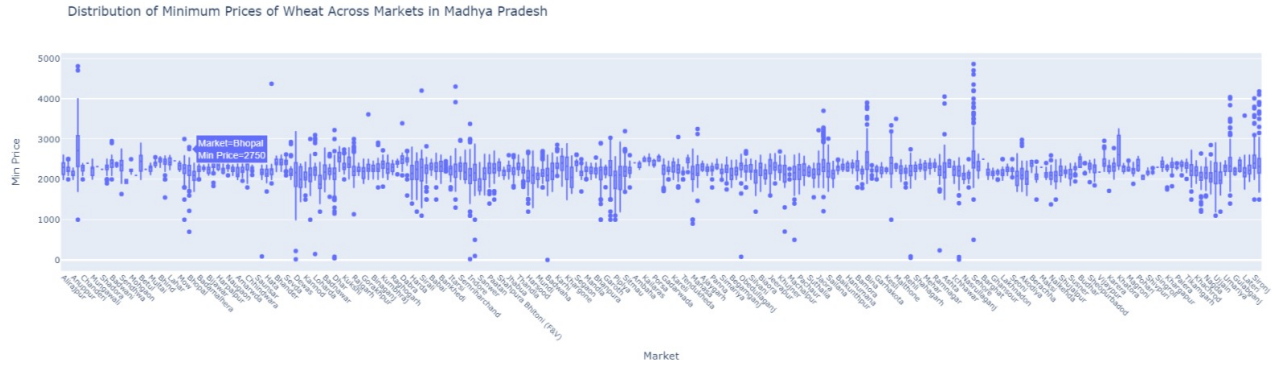
- This variety shows a price range from 2460.0 (maximum) to 2200.0 (minimum), with a price differential of 260.0.
- Similar to the 'Medium' variety, the price range is moderate.

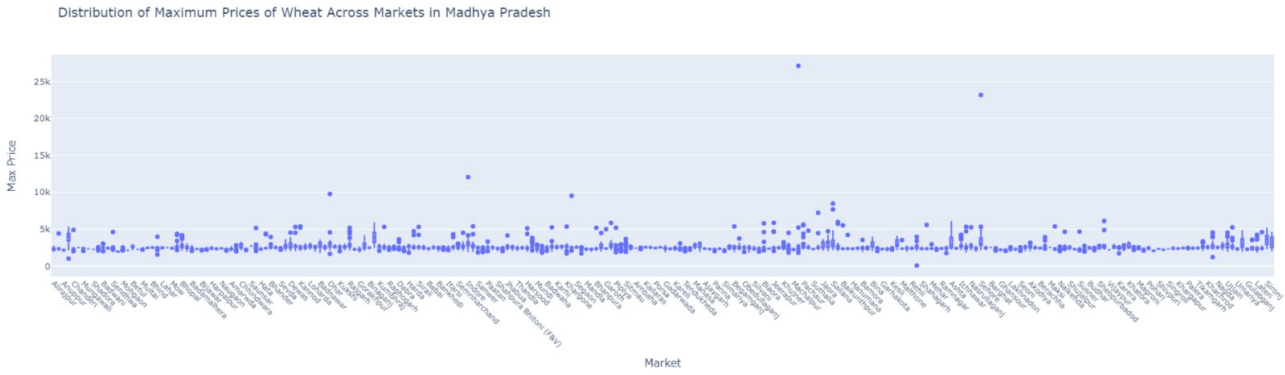
Heat-map:

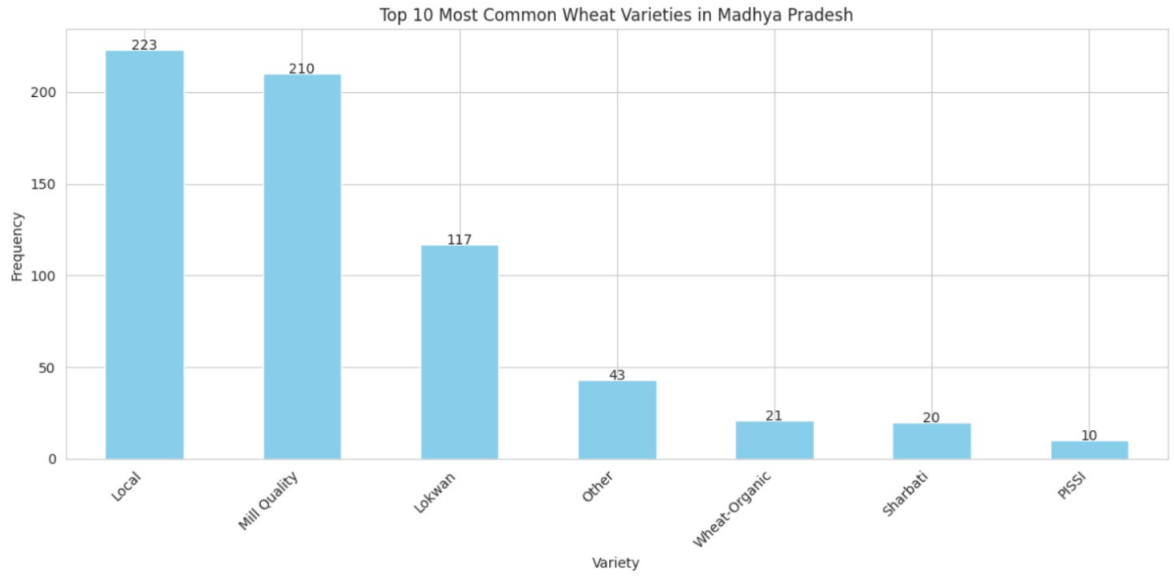


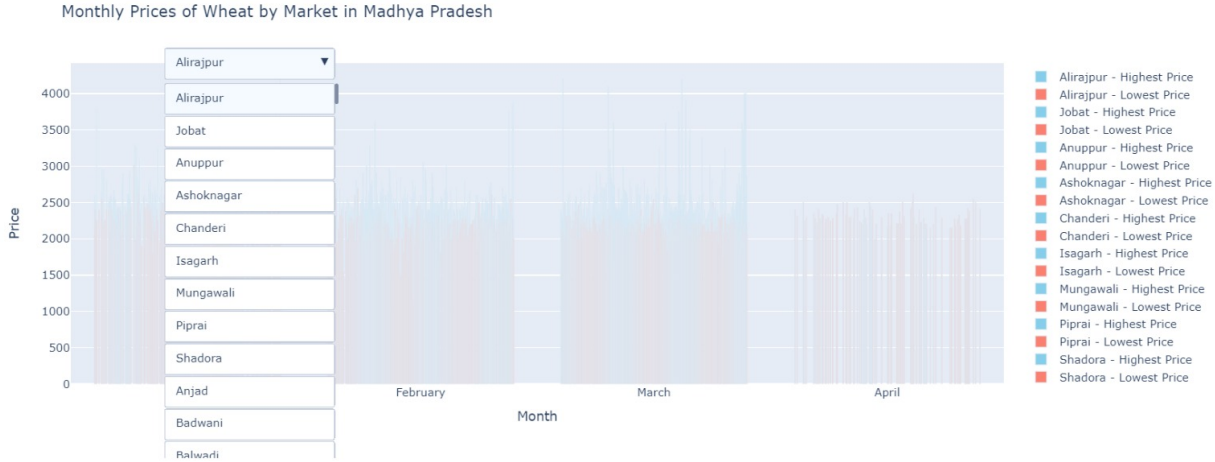
- **Madhya Pradesh**









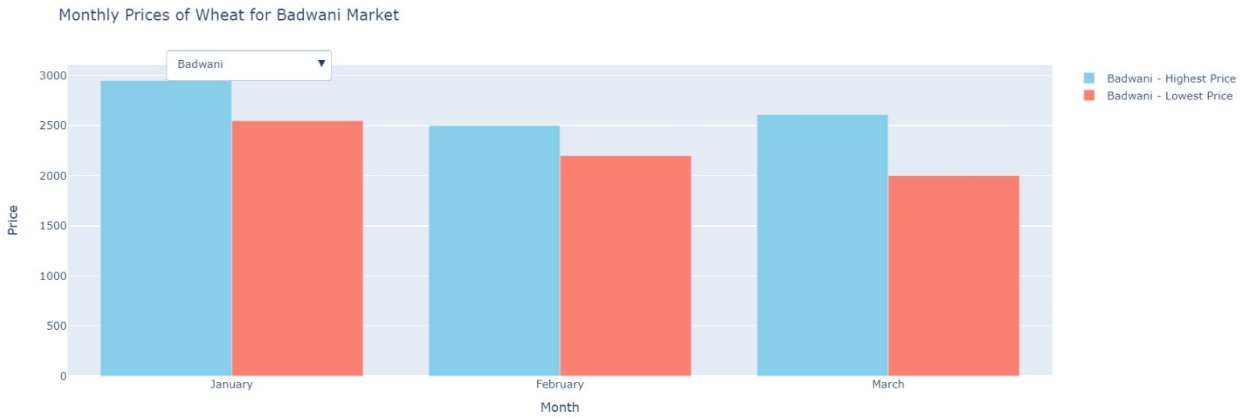


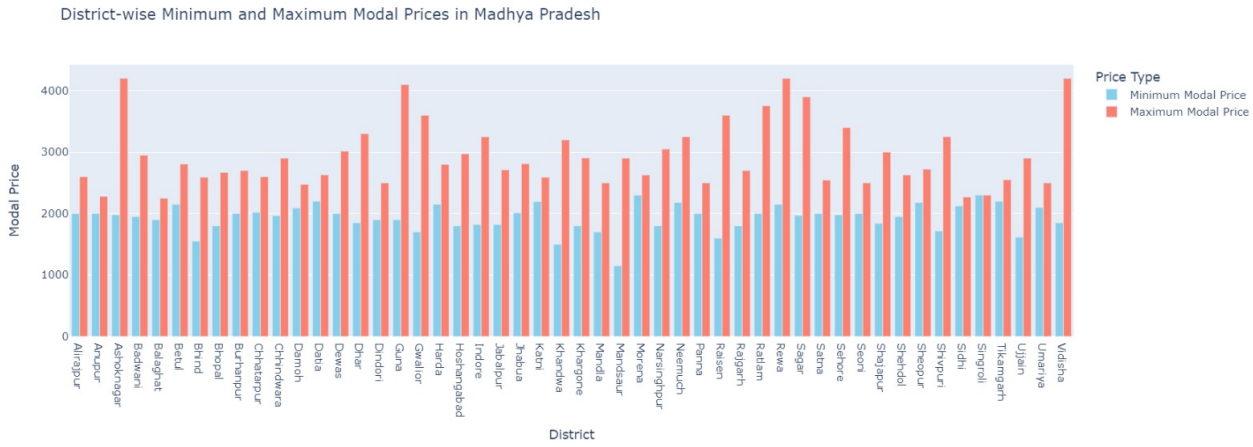
Highest Prices:

- Loharda: March - 3015.0
- Ashoknagar: March - 4200.0
- Rajgarh: March - 2770.0
- Aron: March - 4100.0
- Chhindwara: January - 2900.0
- Patharia: March - 2476.0
- Dhamnod: March - 2700.0
- Rajgarh: January - 3096.0
- Gandhwani: January - 2850.0
- LavKush Nagar(Laundi): February - 2430.0

Lowest Prices:

- Gohad: February - 1550.0
- Khirakiya: April - 2291.0
- Rajgarh: February - 2531.0
- Berasia: February - 2200.0
- Berasia: March - 2000.0
- Chaurai: February - 1964.0
- Sevda: March - 2200.0
- Rajgarh: April - 2075.0
- Aron: February - 2212.0
- Lashkar: March - 1700.0



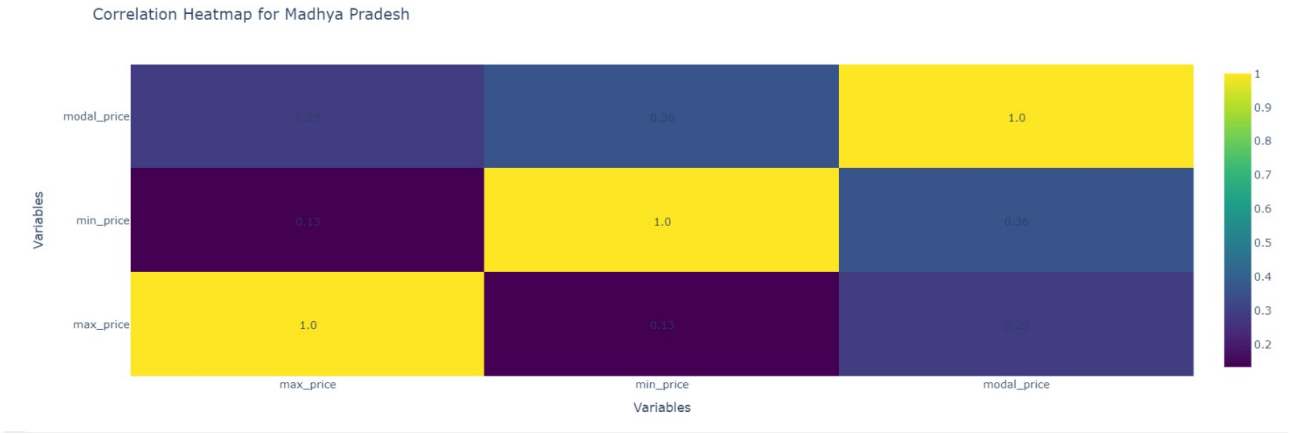


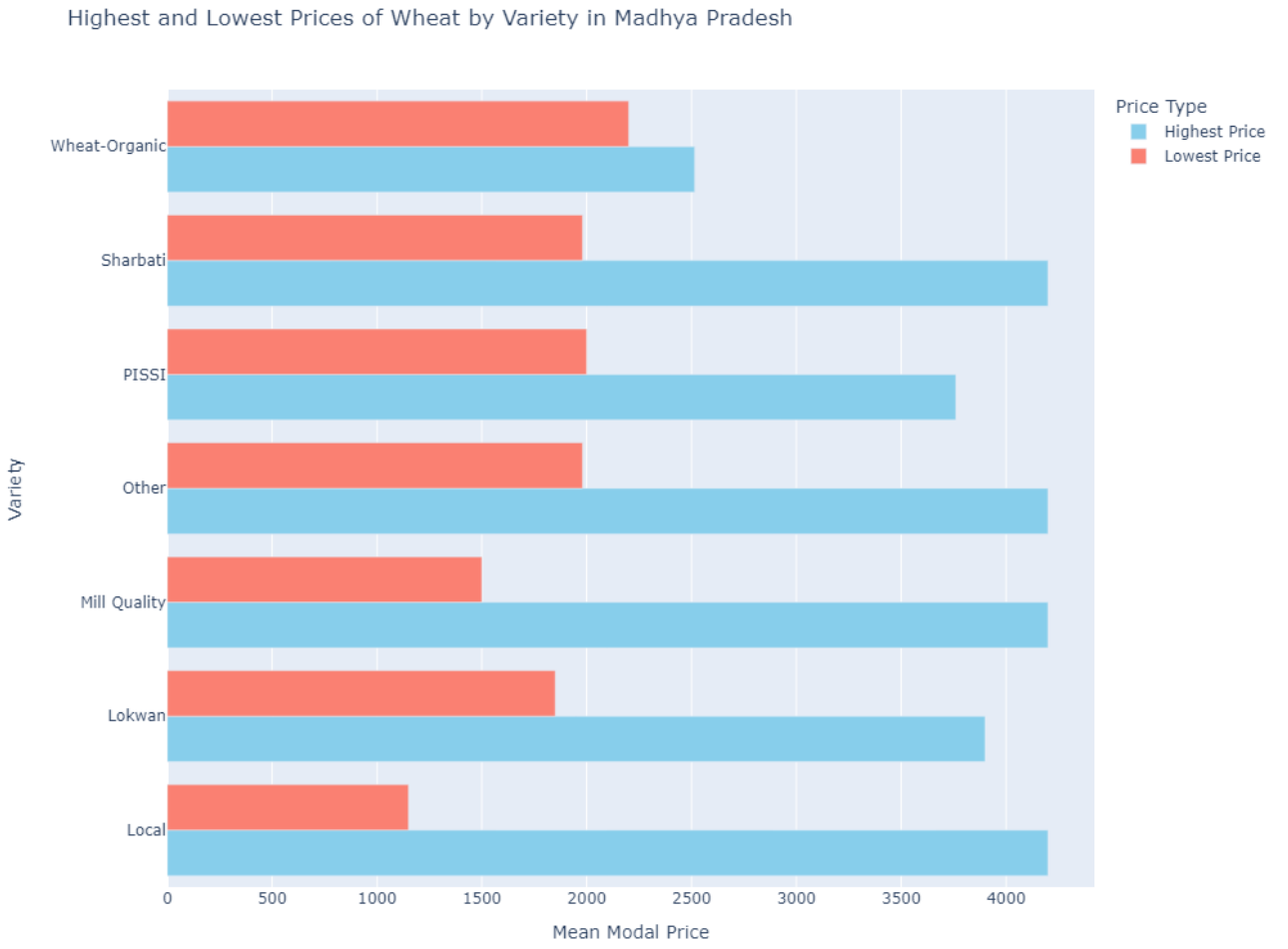
Highest Prices (Top 10):

- Rewa: 4200.0
- Ashoknagar: 4200.0
- Guna: 4100.0
- Sheopur: 2722.0
- Shivpuri: 3251.0
- Neemuch: 3250.0
- Ujjain: 2900.0
- Morena: 2628.0
- Dhar: 3300.0
- Singroli: 2300.0

Lowest Prices (Top 10):

- Bhind: 1550.0
- Burhanpur: 2000.0
- Chhindwara: 1964.0
- Mandsaur: 1150.0
- Khandwa: 1500.0
- Khargone: 1800.0
- Satna: 2000.0
- Ratlam: 2000.0
- Shehdol: 1950.0


Correlation Heatmap for Madhya Pradesh
modal_price
min_price
max_price
Variables
1.0
0.36
0.13
1.0
0.36
1.0
0.13
max_price
min_price
modal_price
Variables
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2


Highest and Lowest Prices of Wheat by Variety in Madhya Pradesh
Wheat-Organic
Sharbati
PISSI
Other
Mill Quality
Lokwan
Local
Variety
0
500
1000
1500
2000
2500
3000
3500
4000
Mean Modal Price
Price Type
Highest Price
Lowest Price

Maximum Modal Price:

- ”Local,” ”Mill Quality,” ”Other,” ”PISSI,” and ”Sharbati” varieties reach the highest price of 4200.0.
- ”Lokwan” follows closely with a maximum price of 3900.0.
- ”Wheat-Organic” has a comparatively lower maximum price of 2515.0.

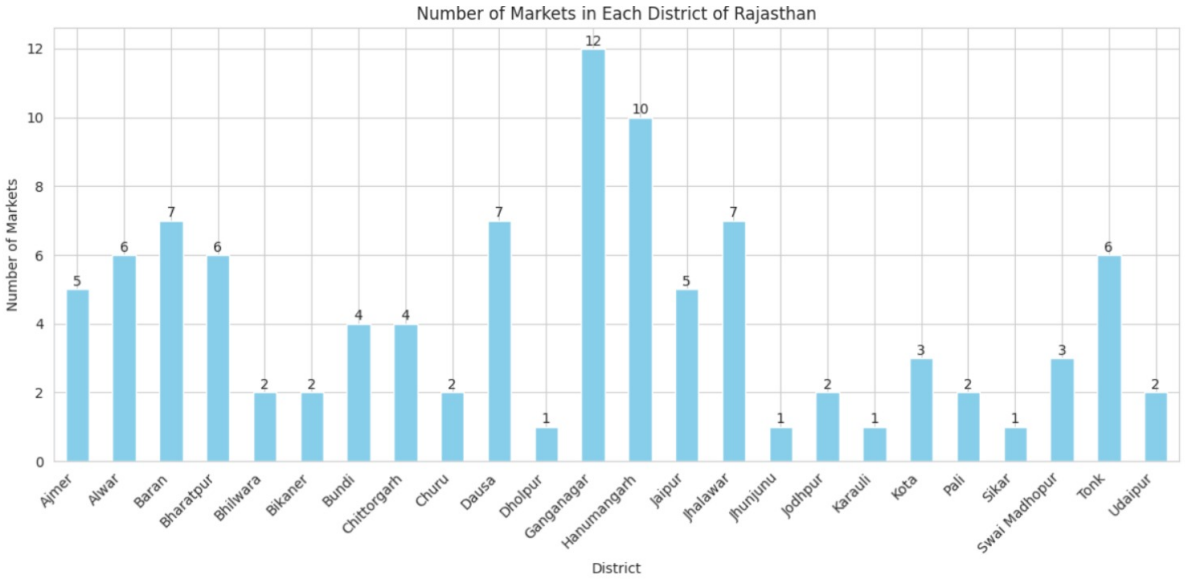
Minimum Modal Price:

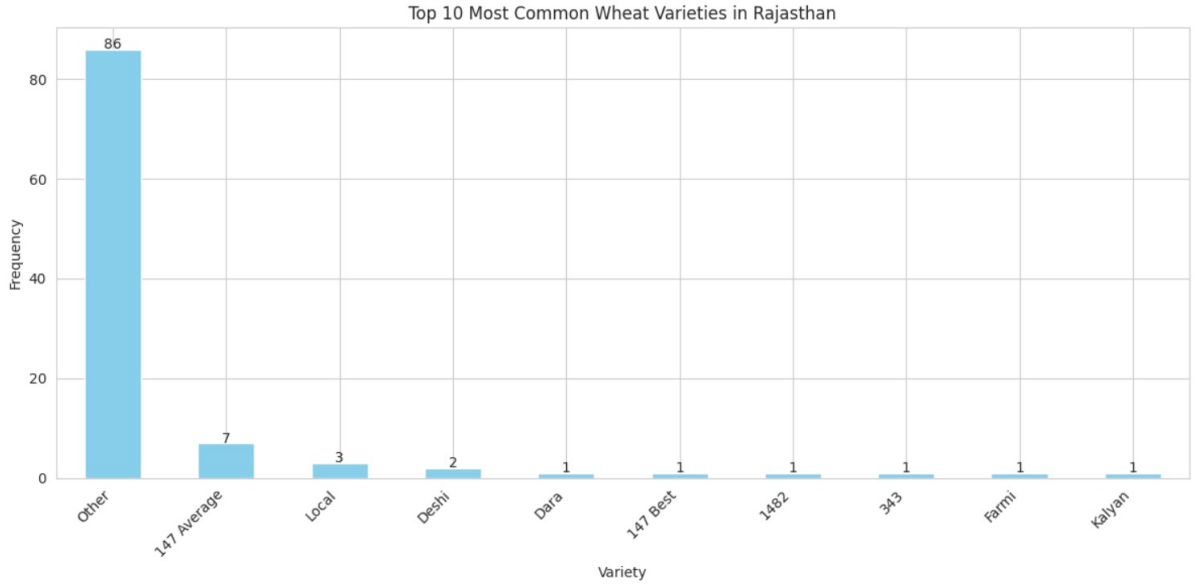
- ”Local” has the lowest minimum price of 1150.0.
- ”Sharbati” and ”Other” share the same minimum price of 1980.0.
- ”Wheat-Organic” has the highest minimum price among the listed varieties, at 2200.0.

Price Range:

- The price range within each variety reflects the variability in pricing based on factors such as quality, demand, and market conditions.
- For instance, ”Local” wheat has a wide price range of 3050.0 (4200.0 - 1150.0), indicating potential fluctuations in market demand or quality assessment.

- **Rajasthan**






```
Top 10 Most Common Wheat Varieties in Rajasthan:
Other          86
147 Average     7
Local           3
Deshi           2
Dara            1
147 Best        1
1482            1
343             1
Farmi           1
Kalyan          1
Name: count, dtype: int64
```

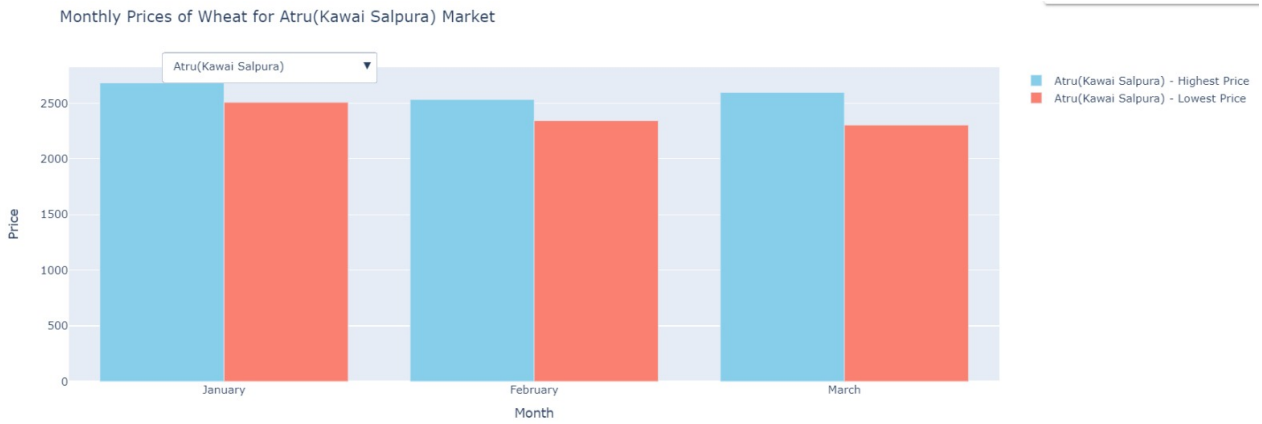
Highest Prices (Top 5 Markets):

- **Beawar:**
  - Highest Price: 2850.0 (January)
  - Lowest Price: 2850.0 (January)
- **Bijay Nagar:**
  - Highest Price: 2780.0 (March)
  - Lowest Price: 2350.0 (January)

- **Kekri:**
  - Highest Price: 2780.0 (February)
  - Lowest Price: 2120.0 (February)
- **Madanganj Kishanganj:**
  - Highest Price: 2765.0 (March)
  - Lowest Price: 1800.0 (March)
- **Vijay Nagar (Gulabpura):**
  - Highest Price: 2650.0 (January/March)
  - Lowest Price: 2370.0 (March)

Lowest Prices (Top 5 Markets):

- **Anoopgarh:**
  - Highest Price: 2560.0 (March)
  - Lowest Price: 2000.0 (March)
- **Bandikui:**
  - Highest Price: 3651.0 (January)
  - Lowest Price: 2100.0 (February)
- **Bayana:**
  - Highest Price: 2530.0 (March)
  - Lowest Price: 2258.0 (March)
- **Bhadara:**
  - Highest Price: 2545.0 (March)
  - Lowest Price: 2200.0 (March)
- **Bharatpur (Kumer):**
  - Highest Price: 2375.0 (February)
  - Lowest Price: 2375.0 (February)

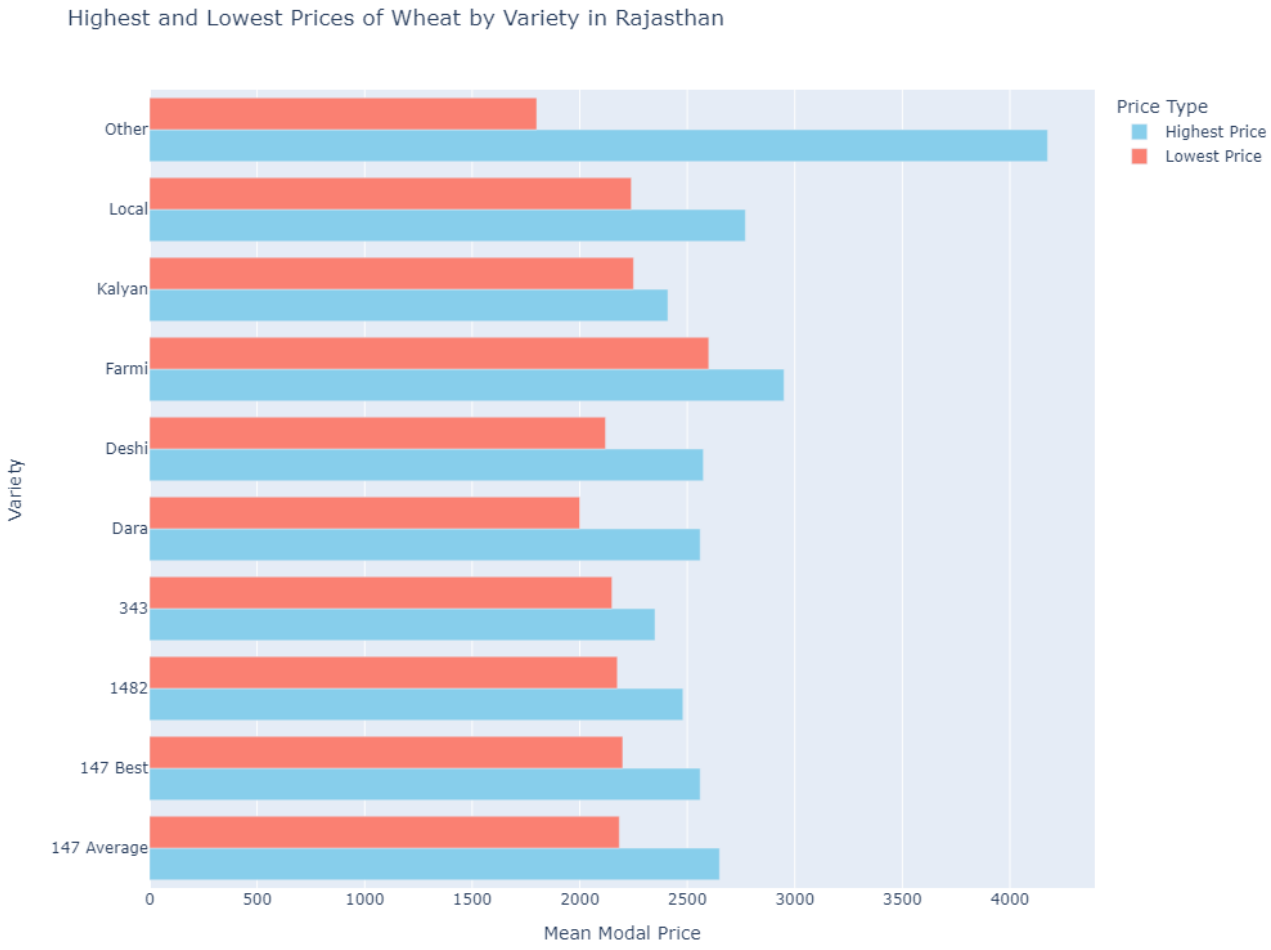


Variability Across Districts: The minimum and maximum prices vary significantly across different districts, indicating regional disparities in wheat prices within Rajasthan.

Range of Prices: The range between the minimum and maximum prices is substantial in some districts, such as Jhalawar (2175.0 - 4175.0) and Dausa (2070.0 - 3651.0), suggesting significant price fluctuations within these regions.

High Maximum Prices: Districts like Jodhpur (2750.0), Jhalawar (4175.0), and Udaipur (2950.0) have relatively high maximum prices, indicating potential high-demand areas or premium quality produce in these regions.

Low Minimum Prices: While most districts have minimum prices above 1800.0, indicating relatively stable market conditions, districts like Churu (1800.0) have the lowest minimum price, suggesting potential challenges for wheat producers in those areas.

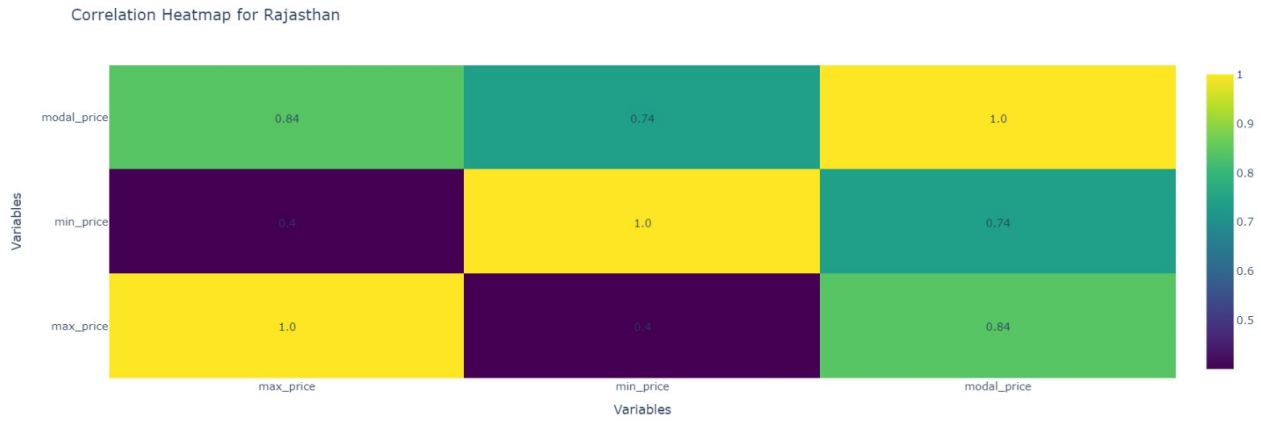


Variability in Prices: There is considerable variability in both maximum and minimum prices across different wheat varieties. For example, the maximum prices range from 2350.0 (variety 343) to 4175.0 (variety Other), while the minimum prices range from 2000.0 (variety Dara) to 1800.0 (variety Other).

Premium Varieties: Some varieties command higher prices compared to others. For instance, Farmi has the highest maximum price of 2950.0, followed closely by Other with 4175.0. These premium varieties may have superior quality, taste, or other desirable characteristics, leading to higher market demand and prices.

Local Varieties: Local varieties also exhibit a wide price range, with a maximum price of 2770.0 and a minimum price of 2240.0. This suggests that local wheat varieties are still significant in the market despite the presence of other higher-priced varieties.

Price Stability: While some varieties like Kalyan and 147 Best have relatively stable prices (2410.0 - 2560.0), others such as Other and Dara show a wider price fluctuation, indicating potential market volatility or variations in quality and supply.



Correlation Coefficients Between Price Variables The correlation coefficients between different variables related to prices. The variables include modal_price, min_price, and max_price.

The heatmap uses color coding to represent the strength and direction of correlation. The scale on the right indicates the correlation coefficient values ranging from 0 to 1.

- **Modal Price and Min Price:** 0.74
- **Modal Price and Max Price:** 0.84
- **Min Price and Max Price:** 0.4

Chapter 4. Feature Engineering

1. **Feature Creation**: Generating new features from existing ones to capture additional information or relationships within the data, including aggregation, interaction terms, or extracting relevant information from text or categorical variables.

2. **Dimensionality Reduction**: Techniques like PCA or feature selection help reduce the dimensionality of the dataset, retaining informative features while discarding redundant ones to improve model efficiency and interpretability, and mitigate overfitting.

3. **Handling Missing Values**: Strategies such as imputation or creating binary indicators for missing values ensure valuable information isn't lost, enabling accurate predictions.

4. **Encoding Categorical Variables**: Transforming categorical variables into numerical format using techniques like one-hot encoding, label encoding, or target encoding ensures model compatibility while preserving valuable information and minimizing bias.

5. **Feature Scaling**: Scaling numerical features to a standard range through normalization or standardization prevents larger-magnitude features from dominating model training, ensuring all features contribute equally.

6. **Feature Extraction**: Extracting relevant information from raw data sources like text, images, or time series data using techniques like word embedding, image feature extraction, or time series decomposition captures underlying patterns and structures for analysis.

7. **Polynomial Features:** Feature engineering encompasses the creation of polynomial features by generating new features as polynomial combinations of existing ones. This allows models to capture non-linear relationships between variables, enhancing their predictive capability and flexibility.

8. **Temporal Features:** For time series data, feature engineering involves extracting temporal features such as day of the week, month, seasonality, or trend components. These features provide valuable insights into patterns and trends over time, enabling models to make more accurate forecasts or predictions.

Overall, effective feature engineering enhances model performance by providing relevant and informative input features, thereby improving the model's ability to learn and make accurate predictions.

4.1 Feature extraction

Let's take a look on a comprehensive overview of feature extraction in data preprocessing. Here's a summary of the key points:

1. **Dimensionality Reduction**: Feature extraction aims to reduce the dimensionality of the dataset by transforming original features into a lower-dimensional space, selecting or creating a subset of features containing the most informative characteristics.

2. **Creation of New Features**: New features are generated from existing ones to capture additional information or relationships within the data. This involves mathematical transformations or domain-specific knowledge to represent complex patterns or structures.

3. **Information Compression**: Feature extraction compresses information from original features into a more compact representation while retaining relevant information. This facilitates more efficient storage, computation, and analysis, especially for high-dimensional datasets.

4. **Preservation of Information**: Despite dimensionality reduction, feature extraction aims to preserve important information. Techniques like PCA, LDA, or autoencoders learn lower-dimensional representations maximizing variance or discriminability between classes while minimizing information loss.

5. **Enhancement of Model Performance**: Effective feature extraction improves machine learning model performance by providing informative and discriminative features. Extracted features capture relevant patterns, enabling models to make more accurate predictions or classifications.

6. **Domain-specific Knowledge**: Feature extraction incorporates domain-specific knowledge to create meaningful features capturing domain-specific characteristics or nuances. This involves transforming raw data into features representing underlying phenomena or relationships of interest.

In summary, feature extraction plays a vital role in preparing data for machine learning tasks, facilitating better model understanding and performance by capturing relevant information and reducing dimensionality effectively.

4.2 Feature selection

1. **Feature Selection:**

   For Dataset-2 - Wheat-2024 : Identified and removed irrelevant or constant columns from the dataset. Specifically, the "Commodity" column containing a single value ("Wheat") and the "Update Date" column with the same date across all records were removed as they didn't contribute meaningful variation.

```
if df1['commodity'].nunique() == 1:
# If the column is constant, remove it
df1.drop(columns=['commodity'], inplace=True)
```

```
# Check if 'update_date' column is constant
if df1['update_date'].nunique() == 1:
# If the column is constant, remove it
df1.drop(columns=['update_date'], inplace=True)
```

	state	district	market	variety	arrival_date	min_price	max_price	modal_price
0	Bihar	Muzaffarpur	Muzaffarpur	147 Average	2024-01-01	2500.0	2590.0	2550.0
1	Bihar	Muzaffarpur	Muzaffarpur	147 Average	2024-01-02	2500.0	2580.0	2540.0
2	Bihar	Muzaffarpur	Muzaffarpur	147 Average	2024-01-03	2500.0	2590.0	2550.0
3	Bihar	Muzaffarpur	Muzaffarpur	147 Average	2024-01-04	2500.0	2590.0	2550.0
4	Bihar	Muzaffarpur	Muzaffarpur	147 Average	2024-01-05	2500.0	2580.0	2550.0

2. **Feature Transformation:** *For Dataset-2 'Wheat-2024' :* While analyzing our dataset, we encountered a common hurdle: a notable number of missing values in the 'variety' column. This issue stemmed from numerous commodities lacking variety information, posing a challenge to our analysis. To overcome this obstacle, we decided to merge the 'variety' data with the 'commodities' column, creating a new composite column named 'commodity'. Feature engineering is indeed a critical step in data preprocessing, and we've covered its key aspects comprehensively.

```
#Merging of Variety and Commodity column
data['Var']=data['Commodity'] +' '+ data['Variety']
data['Commodity']=data['Var']
data.drop(columns=['Var','Variety'], inplace=True)
```

	Category	Commodity	2010-11	2011-12	2012-13	2013-14	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20	2020-21	2021-22
0	Kharif Crops	PADDY Common	1000.0	1080.0	1250.0	1310.0	1360.0	1410.0	1470.0	1550.0	1750.0	1815.0	1868.0	1940.0
1	Kharif Crops	PADDY Grade 'A'	1030.0	1110.0	1280.0	1345.0	1400.0	1450.0	1510.0	1590.0	1770.0	1835.0	1888.0	1960.0
2	Kharif Crops	JOWAR Hybrid	880.0	980.0	1500.0	1500.0	1530.0	1570.0	1625.0	1700.0	2430.0	2550.0	2620.0	2738.0
3	Kharif Crops	JOWAR Maldandi	900.0	1000.0	1520.0	1520.0	1550.0	1590.0	1650.0	1725.0	2450.0	2570.0	2640.0	2758.0
4	Kharif Crops	BAJRA -	880.0	980.0	1175.0	1250.0	1250.0	1275.0	1330.0	1425.0	1950.0	2000.0	2150.0	2250.0
5	Kharif Crops	MAIZE -	880.0	980.0	1175.0	1310.0	1310.0	1325.0	1365.0	1425.0	1700.0	1760.0	1850.0	1870.0

After feature selection, categorical variables like ”Market” and ”Variety” were transformed using Label Encoder. This technique converted categorical labels into numerical representations while preserving ordinal relationships between categories.

```
from sklearn.preprocessing import LabelEncoder, StandardScaler
# Perform label encoding for 'market' and 'variety'
label_encoder = LabelEncoder()
df1['market'] = label_encoder.fit_transform(df1['market'])
df1['variety'] = label_encoder.fit_transform(df1['variety'])
```

3. **Feature Scaling:**

For Dataset-1 : MSP Wheat: : After completing data preprocessing, our focus shifted to forecasting commodity prices for the 2021-22 period. We assessed the predictive performance of three machine learning models: linear regression, Random Forest, and k-nearest neighbors (KNN). However, the KNN model's performance was unsatisfactory, possibly due to the use of unscaled data.

To address this issue, we conducted experiments with standard scaling on the price columns. Our results demonstrated a significant enhancement in the accuracy of the KNN model after applying standard scaling. This preprocessing step effectively normalized the data, leading to improved prediction precision.

Although standard scaling had minimal impact on the performance of the linear regression and Random Forest models, we applied it consistently across all models to maintain methodological integrity.

In summary, by addressing data inconsistencies, exploring diverse modeling techniques, and implementing appropriate preprocessing methods such as standard scaling, our objective was to develop dependable forecasts for commodity prices in the 2021-22 period.

```
# Split data into train and test sets
X_train, X_test, y_train, y_test =
        train_test_split(X, y, test_size=0.2, random_state=42)

# Standardize features
```

```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

For Dataset-2 - Wheat 2024 : Numerical features such as ”Min Price,” ”Max Price,” and ”Modal Price” were standardized using StandardScaler. Standardization ensures that all features are on a similar scale with a mean of 0 and a standard deviation of 1, preventing features with larger scales from dominating the modeling process.

```
# Select features and target variable
X = df1[['market', 'variety', 'min_price', 'max_price']]  # Features
y = df1['modal_price']  # Target variable

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
y_scaled = scaler.fit_transform(y.values.reshape(-1, 1)).flatten()
# Reshape y to a 1D array before scaling
```

	market	variety	min_price	max_price
23688	261	14	2356.0	2681.0
2647	373	16	2000.0	2695.0
32658	288	8	2270.0	2470.0
24275	305	25	2150.0	2200.0
5821	120	24	2200.0	2565.0
5194	182	14	2255.0	2312.0
21785	835	25	2825.0	3165.5
16066	428	24	2300.0	2406.0
11953	596	14	2300.0	2610.0
8271	309	16	2470.0	2470.0

Conclusion: By implementing feature selection, transformation, and scaling techniques, the dataset was effectively prepared for subsequent modeling tasks. These preprocessing steps are essential for improving model performance and interpretability by ensuring that the data is appropriately formatted and scaled. Overall, this structured approach to feature engineering enhances the quality of the dataset and facilitates better modeling outcomes.

Chapter 5. Model fitting

Comprehensive Approach to Predictive Modeling

1. **Model Selection**: In our endeavor to create predictive models, we embarked on a meticulous exploration of two distinct regression algorithms: Linear Regression and Random Forest Regression. The rationale behind this choice lies in the diverse strengths and capabilities inherent in each algorithm, which could yield differential performances contingent upon the dataset's characteristics.

2. **Model Fitting**:

   (a) **Linear Regression**: Our journey commenced with the implementation of a Linear Regression model on the meticulously preprocessed dataset. Renowned for its simplicity yet effectiveness, Linear Regression endeavors to delineate linear relationships between features and the target variable. The crux of its operation lies in the minimization of the residual sum of squares between observed and predicted values.

   (b) **Random Forest Regression**: Parallel to our exploration of Linear Regression, we delved into the realm of Random Forest Regression. This sophisticated ensemble learning method constructs a multitude of decision trees during training and subsequently amalgamates their individual predictions to arrive at the average prediction. Renowned for its adeptness in capturing non-linear relationships and intricate feature interactions, Random Forest Regression emerged as a potent tool for regression tasks.

3. **Hyperparameter Tuning**: In our pursuit of optimizing model performance, we ventured into the realm of hyperparameter tuning, with a particular focus on fine-tuning the maximum depth parameter. This pivotal parameter governs the depth of each decision tree within the Random Forest ensemble, thereby wielding a profound influence on the model's capacity to encapsulate intricate data relationships.

4. **Model Evaluation**: To gauge the efficacy of our models, we embarked on a comprehensive evaluation process encompassing an array of performance metrics. These metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and the R-squared (R2) score, served as litmus tests for accuracy, precision, and the overall goodness-of-fit of our models.

5. **Optimal Max Depth Selection**: Our journey towards optimal model configuration entailed a meticulous exploration of the relationship between performance metrics (MAE, MSE, R2) and diverse values of the maximum depth parameter. By visually scrutinizing these metrics, we endeavored to discern the value of maximum depth that yielded optimal model performance.

5.1 Regression

Regression analysis is a fundamental statistical technique used to understand the relationship between a dependent variable (target) and one or more independent variables (features). Its applications span across diverse domains, including economics, finance, healthcare, and engineering, facilitating predictions and uncovering underlying data patterns.

Types of Regression:

1. **Linear Regression:** This method assumes a linear relationship between an independent and dependent variable, aiming to fit the line or hyperplane that best represents this relationship. Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2) score are commonly employed to assess model performance.

2. **Polynomial Regression:** Extending beyond linearity, Polynomial Regression accommodates nonlinear relationships by fitting polynomial functions to the data, thereby offering flexibility in modeling complex relationships. Training a Linear Regression Model:

3. Following data preprocessing, which includes feature scaling using StandardScaler, we proceeded to train a Linear Regression model on the dataset. This involved partitioning the data into training and testing sets (80% training, 20% testing) to evaluate the model's generalization ability.

4. **Model Evaluation:**

   The performance of the Linear Regression model was appraised using standard regression metrics:

   (a) **Mean Squared Error (MSE):** This metric quantifies the average squared difference between the actual and the predicted values, with lower values signifying superior model performance.

   (b) **Mean Absolute Error (MAE):** MAE shows the average absolute difference between actual and the predicted values, offering an intuitive measure of model performance and how it depicts the accuracy of reading patterns from the dataset.

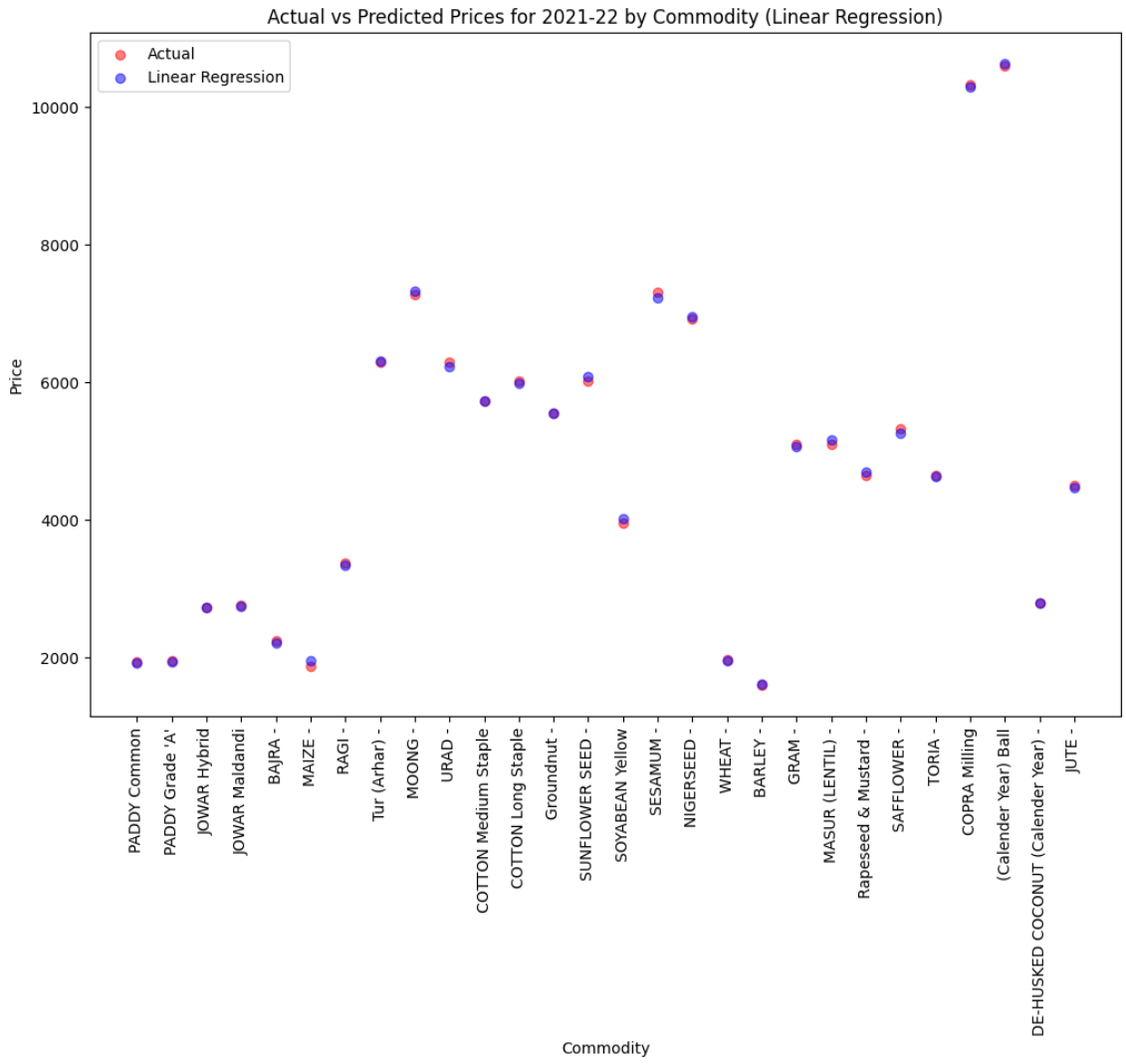
   (c) **R-squared (R^2) Score:** R^2 delineates the proportion of variance in the dependent variable predictable from the independent variables, with higher values indicative of better model fit.

Upon evaluating the Linear Regression model on the testing set, we garnered the following performance metrics:

On dataset-2 : Wheat-2024

MSE: 0.30595765689040866
MAE: 0.3395883432938131
R^2 score: 0.6956013863658803

On dataset-1 : MSP -Wheat
R^2 Score on Test Set: 1.0
MAE on Test Set: $1.4400332778071363 \times 10^{-12}$
MSE on Test Set: $3.756778615345001 \times 10^{-24}$
These metrics provide insights into a model's accuracy and fitness, illuminating its proficiency in predicting the target variable based on the input features.



Linear Regression Applied on Dataet-1 : MSP WHEAT

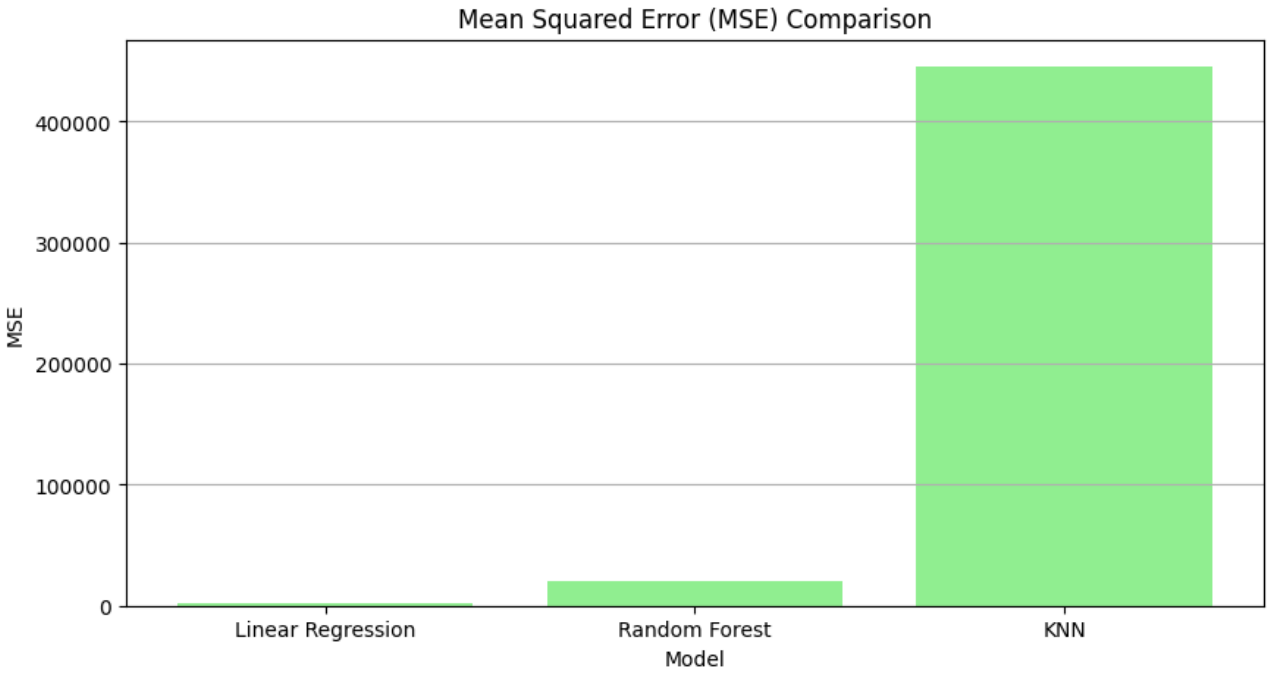
5.2 ML algorithms

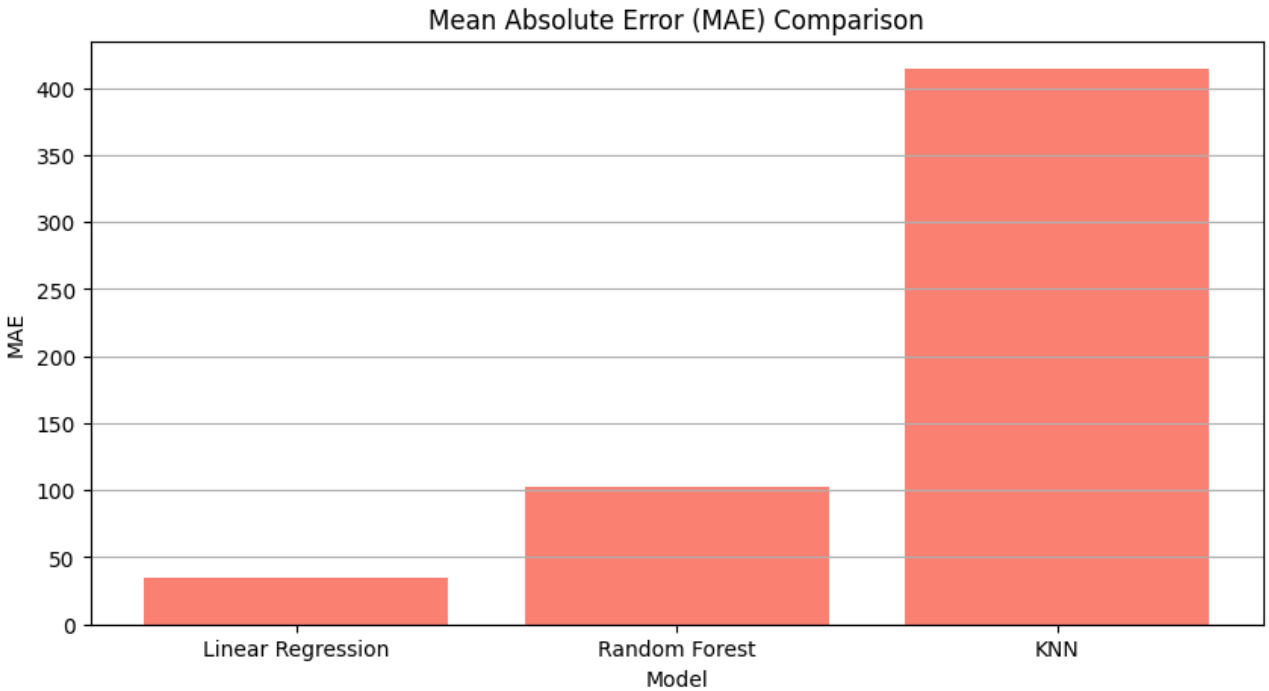
For Dataset-1 : MSP Wheat,

- **Linear Regression Evaluation:** The Linear Regression model exhibits exceptional performance across various evaluation metrics. With a perfect R^2 score of 1.0 on both cross-validation and test set data, it indicates that the model explains all the variability in the target variable. Additionally, the Mean Absolute Error (MAE) and Mean Squared Error (MSE) on the test set are virtually negligible, demonstrating a near-perfect fit of the model to the data. These results suggest that the Linear Regression model captures the underlying relationship between the features and the target variable extremely well, making it a robust choice for prediction tasks, especially when the relationship is assumed to be linear.

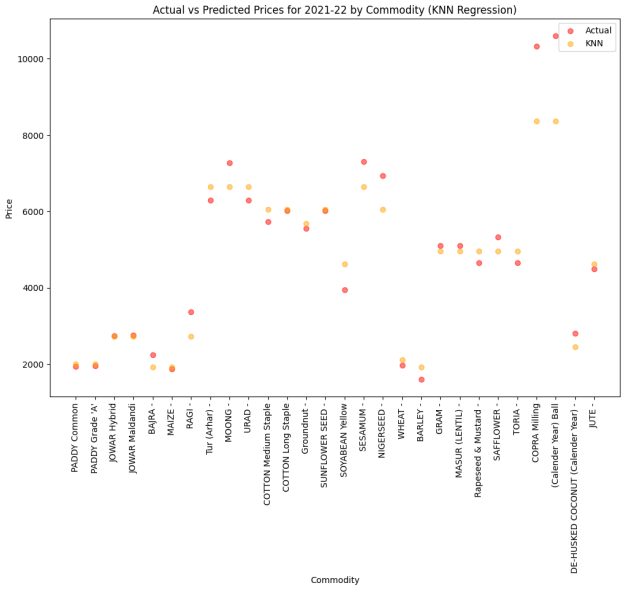
- **Random Forest Regression Evaluation:** The Random Forest Regression model also performs impressively, albeit slightly lower than the Linear Regression model. The cross-validation R^2 scores range from approximately 0.71 to 0.97, indicating high variability in performance across different folds. However, the mean R^2 score of 0.874 suggests strong overall predictive capability. On the test set, the R^2 score remains high at 0.938, indicating that the model generalizes well to unseen data. However, there is a noticeable increase in both MAE and MSE compared to the Linear Regression model, indicating a slightly higher level of error. Despite this, the Random Forest model proves to be robust and effective, particularly in handling nonlinear relationships and complex datasets due to its ensemble nature.

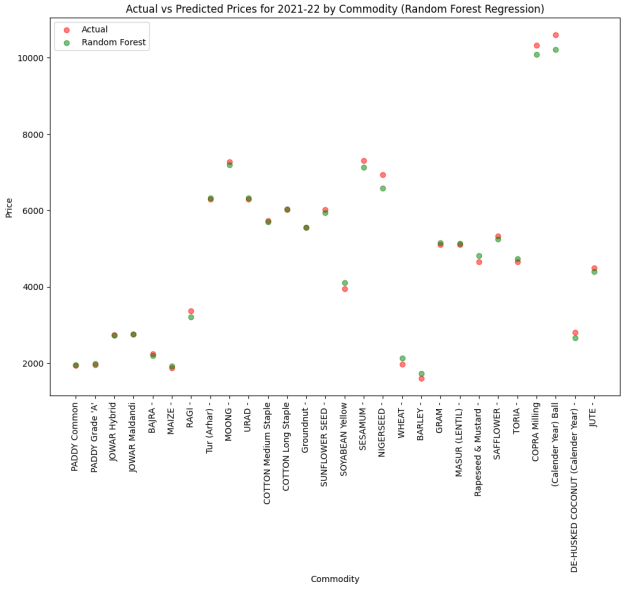
- **KNN Regression Evaluation:** The KNN Regression model shows the lowest performance among the three models evaluated. While the cross-validation R^2 scores vary considerably, with some folds achieving high scores and others performing poorly, the mean R^2 score of 0.642 indicates moderate overall predictive performance. On the test set, the R^2 score is 0.724, suggesting decent predictive capability but not as strong as the other models. However, both MAE and MSE on the test set are significantly higher compared to the other models, indicating a higher level of error. The KNN model's performance may be impacted by its sensitivity to the choice of hyperparameters and the underlying assumptions of locality in the data. It appears to struggle more with generalization to unseen data compared to the other models.


Mean Squared Error (MSE) Comparison
MSE
400000
300000
200000
100000
0
Linear Regression
Random Forest
KNN
Model


Mean Absolute Error (MAE) Comparison
MAE
400
350
300
250
200
150
100
50
0
Linear Regression
Random Forest
KNN
Model

- **Model Perfomance Results :**
  - **Random Forest Regression:**
    * R^2 Score on Test Set: 0.9384292057272963
    * MAE on Test Set: 349.22666666666646
    * MSE on Test Set: 424506.8960999997
  - **KNN Regression:**
    * R^2 Score on Test Set: 0.7244476457621662
    * MAE on Test Set: 825.7333333333335
    * MSE on Test Set: 1899827.280000001





For Dataset-2 : Wheat - 2024,

Random Forest Regression stands out as a robust ensemble learning technique used for enhancing prediction accuracies. By aggregating the predictions from multiple decision trees, it leverages the collective strengths of these trees to deliver more accurate and reliable predictions. Random Forest Regression offers distinct advantages over traditional regression models:

- **Non-linearities:** Unlike linear regression models, Random Forest Regression can effectively capture non-linear relationships between features and the target variables. This makes it well-suited for datasets with complex patterns that linear models may struggle to capture adequately.
- **Reduced Overfitting:** Through aggregating the predictions from multiple trees, Random Forest Regression mitigates the risks of overfitting to the training data. This results in improved performance when faced with unseen data, enhancing the model's generalizability.

Model Training and Evaluation:

Our approach involved training Random Forest Regression models on our dataset, splitting the data into training (80%) and testing (20%) sets. Subsequently, the model underwent training on the training sets and evaluation on the testing sets to assess its performance under real-world conditions.

Performance Evaluation:

To gauge the model's efficacy, we relied on standard regression metrics:

- **Mean Squared Error (MSE):** This metric quantifies the average squared difference between actual and predicted values, with lower values indicating superior model performance.
- **Mean Absolute Error (MAE):** MAE measures the average absolute differences between actual and predicted values, offering an intuitive measure of model accuracy.
- **R-squared (R^2) Score:** R^2 denotes the proportion of variance in the dependent variable explained by the independent variables, with higher values signifying better model fits.

Model Performance Results:

Upon evaluation, the Random Forest Regression model exhibited the following performance metrics on the testing sets:

- **Mean Squared Error:** 0.21886665
- **Mean Absolute Error:** 0.25698435
- **R-squared Score:** 0.78224860

These metrics underscore the model's commendable performance, striking a balance between accuracy and generalizability.

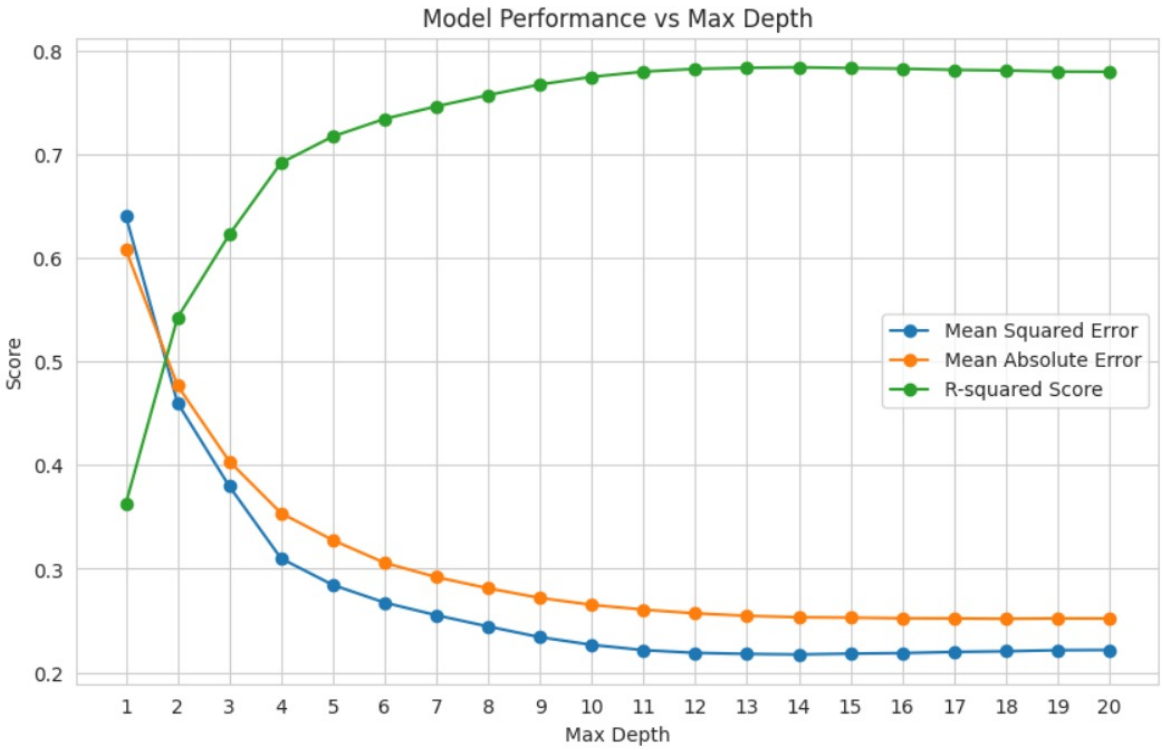
Hyperparameter Tuning for Optimal Performance:

To further enhance the model's performance, we delved into hyperparameter tuning. We focused on optimizing the maximum depth parameter, a critical hyperparameter in Random Forest

Regression that influences the complexities of individual trees and, consequently, the model's ability to capture data relationships.

Optimal Maximum Depth Selection:

Through meticulous examination of the relationship between the maximum depth parameter and the R^2 scores, we identified the optimal values. Visualization aided in pinpointing the best maximum depth for the model, fortifying its performance.



Dataset-2 : Wheat-2024

Chapter 6. Conclusion & future scope

The dataset-1, *MSP WHEAT* offers a comprehensive view of the intricate dynamics within agricultural markets across various regions and commodities. Through detailed analysis, we uncover a mosaic of price fluctuations, seasonal trends, regional disparities, and varietal price differentials, each carrying profound implications for farmers, traders, policymakers, and consumers.

One striking observation is the pronounced variability in commodity prices across different regions and months, reflecting the complex interplay of factors such as local demand-supply dynamics, climatic conditions, transportation costs, and government policies. Understanding these nuances is crucial for stakeholders navigating the agricultural market landscape effectively.

Seasonal patterns emerge prominently, with prices exhibiting distinct fluctuations over consecutive months. Harvest seasons often witness price peaks due to increased supply, while off-peak periods may see dips as supply diminishes. These trends underscore the importance of timing in agricultural production, marketing, and procurement strategies.

Moreover, disparities in prices between regions highlight the multifaceted nature of agricultural markets. Factors like production levels, market infrastructure, accessibility, and consumer preferences contribute to these regional variations. While some regions command higher prices due to premium quality produce or strategic market positioning, others face challenges from logistical constraints or oversupply.

Varietal price differences further enrich the narrative, with certain varieties fetching higher premiums due to perceived quality, rarity, or unique characteristics. Farmers can leverage this insight to optimize their crop selection and cultivation practices, enhancing competitiveness and profitability.

Government interventions also shape market dynamics, with policies such as price support mechanisms, subsidies, and import-export regulations exerting significant influence. Analyzing price data alongside policy changes provides insights into the efficacy and impact of such interventions on market stability and farmer livelihoods.

Addressing outliers in the dataset through data cleaning and preprocessing techniques is crucial for ensuring robustness and reliability in subsequent analyses. By capping outliers, analysts mitigate the distorting effects of extreme values, yielding more accurate and actionable insights.

Overall, analysis of agricultural price data underscores the need for ongoing data collection, analysis, and dissemination efforts to foster transparency, efficiency, and resilience in agricultural markets. By harnessing data-driven insights, stakeholders can navigate uncertainties, seize opportunities, and pursue sustainable agricultural development and food security.

Whereas dataset-2 *Wheat-2024*, offers a comprehensive view of the intricate dynamics within agricultural markets across various regions and commodities. Through detailed analysis, we uncover a mosaic of price fluctuations, seasonal trends, regional disparities, and varietal price differentials, each carrying profound implications for farmers, traders, policymakers, and consumers.

One striking observation is the pronounced variability in commodity prices across different regions and months, reflecting the complex interplay of factors such as local demand-supply dynamics, climatic conditions, transportation costs, and government policies. Understanding these nuances is crucial for stakeholders navigating the agricultural market landscape effectively.

Seasonal patterns emerge prominently, with prices exhibiting distinct fluctuations over consecutive months. Harvest seasons often witness price peaks due to increased supply, while off-peak periods may see dips as supply diminishes. These trends underscore the importance of timing in agricultural production, marketing, and procurement strategies.

Moreover, disparities in prices between regions highlight the multifaceted nature of agricultural markets. Factors like production levels, market infrastructure, accessibility, and consumer preferences contribute to these regional variations. While some regions command higher prices due to premium quality produce or strategic market positioning, others face challenges from logistical constraints or oversupply.

Varietal price differences further enrich the narrative, with certain varieties fetching higher premiums due to perceived quality, rarity, or unique characteristics. Farmers can leverage this insight to optimize their crop selection and cultivation practices, enhancing competitiveness and profitability.

Government interventions also shape market dynamics, with policies such as price support mechanisms, subsidies, and import-export regulations exerting significant influence. Analyzing price data alongside policy changes provides insights into the efficacy and impact of such interventions on market stability and farmer livelihoods.

Addressing outliers in the dataset through data cleaning and preprocessing techniques is crucial for ensuring robustness and reliability in subsequent analyses. By capping outliers, analysts mitigate the distorting effects of extreme values, yielding more accurate and actionable insights.

Overall, analysis of agricultural price data underscores the need for ongoing data collection, analysis, and dissemination efforts to foster transparency, efficiency, and resilience in agricultural markets. By harnessing data-driven insights, stakeholders can navigate uncertainties, seize opportunities, and pursue sustainable agricultural development and food security.

6.1 Findings/observations

Based on Dataset-1 : MSP WHEAT

1. **Divergent Trends:** The data reveals divergent trends across various agricultural commodities, with some experiencing consistent price increases over the years, while others demonstrate more erratic fluctuations.
2. **Market Sensitivity:** Certain commodities, such as paddy and wheat, show a steady upward trajectory in prices, reflecting their fundamental importance in the food supply chain and the relatively stable demand. Conversely, crops like jowar and cotton exhibit greater sensitivity to market dynamics, with prices influenced by factors like global demand, weather conditions, and government interventions.
3. **Seasonal Influences:** The data suggests that agricultural prices are subject to seasonal fluctuations, likely influenced by factors such as harvesting seasons, weather patterns, and supply-demand dynamics. Understanding these seasonal variations is crucial for farmers and policymakers to make informed decisions regarding planting, harvesting, and marketing strategies.
4. **Policy Impact:** The role of government policies emerges as a significant factor affecting agricultural prices. Subsidies, trade regulations, and procurement policies can have a substantial impact on price stability and market dynamics for certain commodities, highlighting the importance of policy coherence and predictability in agricultural markets.
5. **Supply Chain Dynamics:** Analysis of price trends also sheds light on supply chain dynamics within the agricultural sector. Variations in prices across different stages of the supply chain, from production to processing to retail, reflect the complexities and inefficiencies inherent in the agricultural value chain, underscoring the need for improved infrastructure, logistics, and market linkages.

Based on Dataset-2 : Wheat 2024,

1. **Variety-wise Price Range:**
    - The modal price range varies across different varieties of crops. For instance:
        * *Dara:* Modal prices range from 2300.0 to 2525.0.
        * *Deshi:* Modal prices range from 2500.0 to 3370.0.
        * *MP(Desi):* Modal prices remain constant at 3000.0.
        * *Mexican:* Modal prices range from 2374.0 to 2825.0.
2. **Monthly Price Fluctuations:**
    - Prices vary across months for different locations, indicating seasonal fluctuations and possibly supply-demand dynamics.
    - For example, in January, some locations like Sillod(Bharadi) have a higher modal price compared to others like Sillod, suggesting possible variations in crop quality or market conditions.
3. **Regional Price Comparisons:**

- Prices vary significantly across regions, with some regions consistently commanding higher prices compared to others.
- For example, Lasur Station consistently shows higher prices compared to other locations like Paithan and Sillod.

4. **Consistency in Prices:**
   - Some regions exhibit consistency in prices across months, indicating stable market conditions or consistent quality of produce.
   - However, other regions show more variability, which could be influenced by factors such as weather conditions, crop yields, or market demand.
5. **Overall Market Trends:**
   - The data suggests a complex interplay of factors influencing agricultural prices, including regional dynamics, variety-specific demand, and seasonal variations.
   - Analyzing these trends over time could provide insights into market dynamics and help stakeholders make informed decisions regarding production, pricing, and marketing strategies.
6. **Seasonal Trends and Crop Availability:**
   - The data reveals seasonal trends in crop prices, with some varieties experiencing higher prices during specific months.
   - Understanding these seasonal patterns can help farmers and traders plan their production and marketing strategies accordingly, maximizing profitability.
7. **Impact of Location on Prices:**
   - Prices vary significantly depending on the location, suggesting differences in local market conditions, transportation costs, and demand-supply dynamics.
   - Analyzing these location-based price variations can provide insights into regional market competitiveness and logistical challenges.
8. **Price Stability and Volatility:**
   - While some varieties and regions show relatively stable prices over time, others exhibit greater volatility, indicating factors such as market speculation, weather fluctuations, or regulatory changes.
   - Monitoring price stability and volatility can help stakeholders assess market risks and implement risk mitigation strategies.
9. **Price Disparities and Market Integration:**
   - Disparities in prices across regions may indicate inefficiencies in market integration, transportation infrastructure, or information asymmetry.
   - Addressing these disparities can promote a more efficient and equitable agricultural market, benefiting both producers and consumers.
10. **Price Elasticity and Demand-Supply Dynamics:**
    - Changes in prices across different varieties and regions can reflect the elasticity of demand and supply in agricultural markets.
    - Understanding these demand-supply dynamics can inform pricing strategies and resource allocation decisions for farmers, traders, and policymakers.

11. **Comparative Analysis with Previous Years:**
    - Comparing current price data with historical trends from previous years can reveal long-term market patterns, enabling stakeholders to anticipate future price movements and make proactive decisions.
    - Such analysis can also identify emerging market trends, enabling stakeholders to capitalize on new opportunities or mitigate potential risks.
12. **Factors Influencing Price Movements:**
    - Various factors, including weather conditions, government policies, global market trends, and technological advancements, can influence agricultural prices.
    - Analyzing the impact of these factors on price movements can provide valuable insights for risk management and strategic planning in the agriculture sector.
13. **Price Transparency and Information Accessibility:**
    - Transparent price information facilitates informed decision-making for farmers, traders, and policymakers.
    - Enhancing accessibility to price data through digital platforms or market intelligence systems can empower stakeholders to negotiate better prices and optimize resource allocation.
14. **Quality Standards and Price Differentials:**
    - Variations in crop quality can lead to price differentials, with higher-quality produce commanding premium prices.
    - Establishing and adhering to quality standards can enhance market competitiveness and consumer confidence, driving value creation along the supply chain.
15. **Market Access and Trade Policies:**
    - Trade policies, tariffs, and trade agreements influence market access and price competitiveness for agricultural commodities.
    - Analyzing the impact of trade policies on price dynamics can inform advocacy efforts for trade reform and market liberalization to foster economic growth and food security.
16. **Sustainability and Ethical Considerations:**
    - Increasing consumer demand for sustainably sourced and ethically produced agricultural products can affect price trends.
    - Integrating sustainability practices into agricultural production and supply chains can create value propositions that resonate with environmentally conscious consumers, potentially commanding price premiums.
17. **Market Competition and Pricing Strategies:**
    - Competitive market dynamics drive pricing strategies adopted by market participants, including pricing based on cost-plus margins, target profit levels, or competitive benchmarking.
    - Analyzing competitors' pricing strategies and market positioning can inform strategic pricing decisions and market entry strategies for new entrants.

6.2 Challenges

1. **Data Quality:** Dealing with incomplete and inconsistent data, especially in market prices and Minimum Support Price (MSP) datasets, posed significant challenges. Ensuring data quality and reliability is essential for conducting accurate analyses and building reliable predictive models.
2. **Model Complexity:** Balancing the complexity of machine learning models, such as Random Forest regression, with interpretability and generalization performance can be challenging. Selecting the most appropriate model architecture and optimizing hyperparameters require careful consideration to achieve the desired balance.
3. **Domain Knowledge:** Navigating agricultural dynamics and understanding policy implications requires deep domain expertise. Interpreting seasonal variations, policy-driven trends, and market regulations accurately is crucial for deriving meaningful insights from the data.
4. **Data Integration:** Integrating external datasets, such as weather and economic indicators, introduces challenges in data compatibility and preprocessing. Ensuring that disparate datasets are harmonized and cleaned effectively is essential for maintaining the robustness of predictive models.
5. **External Factors:** Accounting for external factors like weather variations and global market trends adds complexity to predicting agricultural prices. These factors introduce uncertainty and necessitate the development of robust risk management strategies to mitigate potential impacts on model performance.

6.3 Future plan

While this study offers valuable insights into wheat prices and MSP trends, there are numerous avenues for future research and exploration:

- **Feature Engineering:** Further exploration of additional features or the creation of new ones could enhance the predictive power of the models. For instance, integrating weather data, seasonal factors, or economic indicators may lead to more accurate predictions.
- **Advanced Modeling Techniques:** Experimenting with advanced machine learning techniques like Gradient Boosting Machines (GBM), Neural Networks, or Time Series Analysis could potentially yield superior predictive performance. These approaches have the capability to capture complex patterns in wheat price dynamics.
- **Integration of External Data Sources:** Incorporating external datasets such as government policies, international market trends, or agricultural reports could enrich the context and improve the accuracy of price predictions. This broader data integration could provide a more comprehensive understanding of the factors influencing wheat prices.
- **Deployment of Predictive Models:** Deploying the predictive models developed in this study as part of decision support systems for farmers, policymakers, or agricultural stakeholders could facilitate informed decision-making and risk management. These models could serve as valuable tools for stakeholders to anticipate market trends and make strategic decisions.
- **Continuous Monitoring and Evaluation:** Regular monitoring and updating of the models based on new data and evolving market dynamics are crucial to maintaining their relevance and accuracy over time. Continuous evaluation ensures that the models remain effective in capturing changes in market conditions and provide reliable forecasts.

Group Contribution

Member 1

Kushal Barot (ID: 202318006): Kushal Barot, as Member-1, took charge of gathering and preprocessing the data for Dataset-1, focusing on the Minimum Support Price (MSP) of wheat over the specified years. Additionally, they conducted a substantial portion of the initial analysis on Dataset-2, which involved examining wheat prices for the year 2024. Furthermore, Kushal played a pivotal role in compiling the entire report, ensuring coherence and clarity in presenting the findings.

Member 2

Harshil Shah (ID: 202318033): Harshil Shah, as Member-2, collaborated closely with Kushal Barot on analyzing Dataset-2, which focused on wheat prices in 2024. They delved into the data, conducted exploratory analysis, and performed model fitting to identify trends and patterns. Their contribution was crucial in providing valuable insights into the factors influencing wheat prices for the specified year. Harshil's efforts significantly enriched the depth of our analysis.

Member 3

Rishi Pawar (ID: 202318037): Rishi Pawar, as Member-3, took the lead in analyzing Dataset-1, which involved examining the MSP of wheat across multiple years. They meticulously conducted the analysis and model fitting processes, leveraging statistical techniques to uncover trends and patterns in the data. Rishi's contributions were instrumental in providing comprehensive insights into the historical trends and price dynamics of wheat MSP.

Overall, the collaborative efforts of Kushal Barot, Harshil Shah, and Rishi Pawar ensured a thorough analysis of both datasets, enabling us to present a comprehensive report that encapsulates various aspects of wheat pricing dynamics. Each member's unique expertise and dedication significantly contributed to the success of our project.

Short Bio

Include a short bio of all members in a page.

1. **Kushal Barot** Kushal Barot, a motivated resident of Kalol in Gandhinagar district, holds a Bachelor's degree in Computer Applications from GLS University, Ahmedabad, achieving an outstanding CGPA of 7.5. Currently pursuing his Master's in Data Science at DAIICT, Gandhinagar, Kushal possesses a diverse skill set in programming languages such as Java, PHP, Python, Golang, Django, and PowerBI.

His practical experience includes developing a comprehensive website using PHP during his undergraduate studies and leading a project on building an end-to-end bill management system using Python in his Master's program. Driven by a relentless pursuit of excellence and fueled by his passion for innovation, Kushal is poised to make significant contributions to the ever-evolving landscape of technology and data science, leveraging his strong educational foundation and hands-on experience.

2. **Harshil Shah**, a data science enthusiast from Ahmedabad, Gujarat, India with a B.Sc. in Computer Applications and Information Technology from Gujarat University. Currently studying Data Science at DAIICT. Passionate about cricket and trading, and sees the potential for data analysis in both hobbies.

Taking an Exploratory Data Analysis (EDA) course to learn how to extract useful information from data and identify patterns. Proficient in Python, C, C++, PHP, and SQL, with some knowledge of machine learning and deep learning.

3. **Rishi Pawar** is currently pursuing a Master's degree in Data Science at Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, Rishi Pawar is deeply passionate about unraveling insights from data. Proficient in Python, C++, machine learning (ML), and web development, he also excels in database management using PostgreSQL, MySQL, and phpMyAdmin. Rishi is well-versed in leveraging various programming languages and tools to tackle real-world problems. With a commitment to leveraging data-driven approaches to solve real-world problems, Rishi is poised to make a meaningful impact in the field of data science.

References

[1] Analytics Vidhya *URL:* https://www.analyticsvidhya.com/blog/2021/12/12-data-plot-types-for-visualization/

[2] Data Camp *URL:* https://www.datacamp.com/tutorial/types-of-data-plots-and-how-to-create-them-in-python

[3] Geeks for Geeks *URL:* https://www.geeksforgeeks.org/ml-linear-regression/

[4] gov.in *URL:* https://data.gov.in/resource/variety-wise-daily-market-prices-wheat-2

[5] farmer.gov.in *URL:* https://web.archive.org/web/20210816185144/https://farmer.gov.in/mspstatements.aspx

[6] IBM *URL:* https://www.ibm.com/topics/knn#:~:text=The%20k%2Dnearest%20neighbors%20(KNN,of%20an%20individual%20data%20point.

[7] Wikipedia *URL:* https://en.wikipedia.org/wiki/Random_forest#History

[8] Free Code Camp *URL:* https://www.freecodecamp.org/news/what-is-r-squared-r2-value-meaning-and-definition/#:~:text=An%20R%2DSquared%20value%20shows,the%20dependent%20and%20independent%20variables.