

Data Collection and Preprocessing Phase

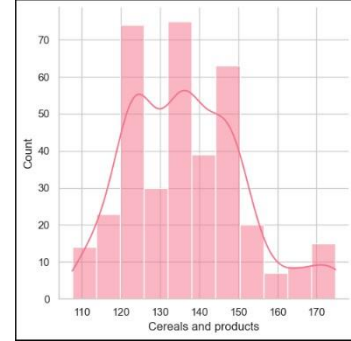
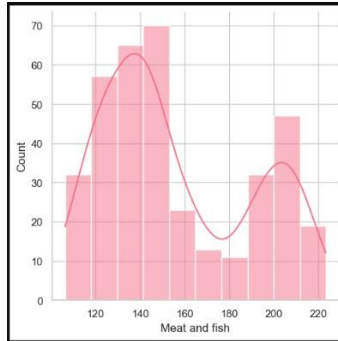
Date	22 June 2024
Team ID	Team - 740093
Project Title	To Predict Consumer Price Index
Maximum Marks	6 Marks

Data Exploration and Preprocessing Report

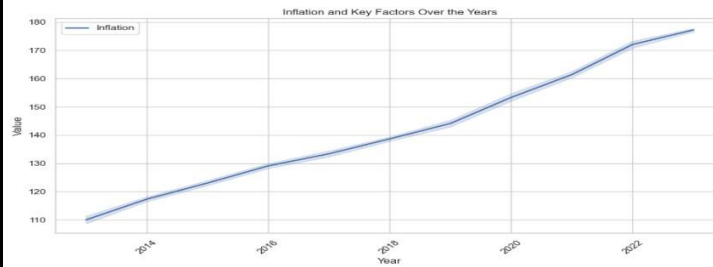
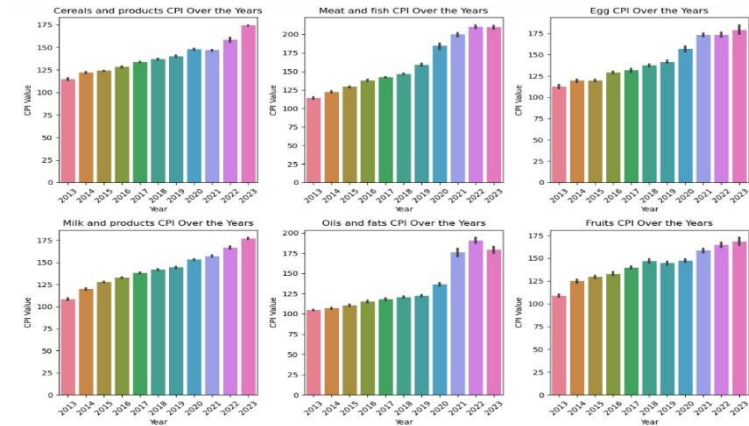
Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	Description																																																																	
Data Overview	<u>Descriptive statistics:</u>																																																																	
	<pre>#Load the dataset cpi_data=pd.read_csv("All_India_Index_july2019_20Aug2020_dec20_2.csv")</pre>																																																																	
	<pre>cpi_data</pre>																																																																	
	<table><tr><th></th><th>Sector</th><th>Year</th><th>Month</th><th>Cereals and products</th><th>Meat and fish</th><th>Egg</th><th>Milk and products</th><th>Oils and fats</th><th>Fruits</th><th>Vegetables</th></tr><tr><td>0</td><td>Rural</td><td>2013</td><td>January</td><td>107.5</td><td>106.3</td><td>108.1</td><td>104.9</td><td>106.1</td><td>103.9</td><td>101.9</td></tr><tr><td>1</td><td>Urban</td><td>2013</td><td>January</td><td>110.5</td><td>109.1</td><td>113.0</td><td>103.6</td><td>103.4</td><td>102.3</td><td>102.9</td></tr><tr><td>2</td><td>Rural+Urban</td><td>2013</td><td>January</td><td>108.4</td><td>107.3</td><td>110.0</td><td>104.4</td><td>105.1</td><td>103.2</td><td>102.2</td></tr><tr><td>3</td><td>Rural</td><td>2013</td><td>February</td><td>109.2</td><td>108.7</td><td>110.2</td><td>105.4</td><td>106.7</td><td>104.0</td><td>102.4</td></tr><tr><td>4</td><td>Urban</td><td>2013</td><td>February</td><td>112.9</td><td>112.9</td><td>116.9</td><td>104.0</td><td>103.5</td><td>103.1</td><td>104.9</td></tr></table>		Sector	Year	Month	Cereals and products	Meat and fish	Egg	Milk and products	Oils and fats	Fruits	Vegetables	0	Rural	2013	January	107.5	106.3	108.1	104.9	106.1	103.9	101.9	1	Urban	2013	January	110.5	109.1	113.0	103.6	103.4	102.3	102.9	2	Rural+Urban	2013	January	108.4	107.3	110.0	104.4	105.1	103.2	102.2	3	Rural	2013	February	109.2	108.7	110.2	105.4	106.7	104.0	102.4	4	Urban	2013	February	112.9	112.9	116.9	104.0	103.5	103.1
	Sector	Year	Month	Cereals and products	Meat and fish	Egg	Milk and products	Oils and fats	Fruits	Vegetables																																																								
0	Rural	2013	January	107.5	106.3	108.1	104.9	106.1	103.9	101.9																																																								
1	Urban	2013	January	110.5	109.1	113.0	103.6	103.4	102.3	102.9																																																								
2	Rural+Urban	2013	January	108.4	107.3	110.0	104.4	105.1	103.2	102.2																																																								
3	Rural	2013	February	109.2	108.7	110.2	105.4	106.7	104.0	102.4																																																								
4	Urban	2013	February	112.9	112.9	116.9	104.0	103.5	103.1	104.9																																																								

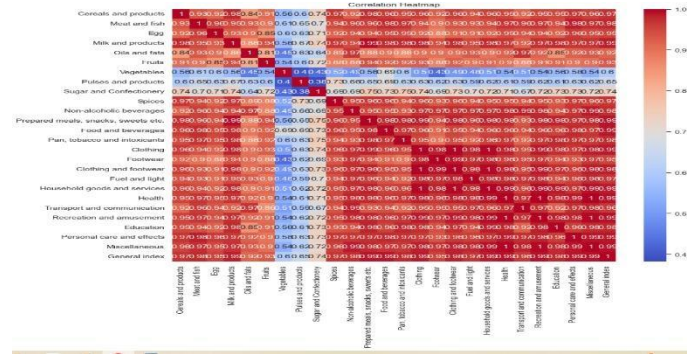
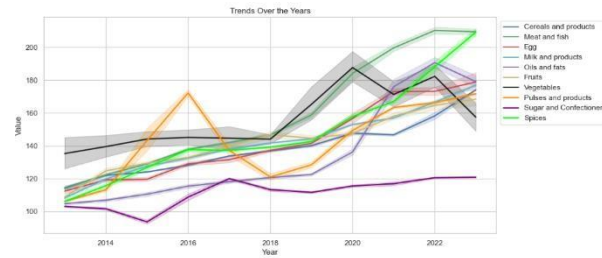
Univariate Analysis



Bivariate Analysis



Multivariate Analysis



Outliers and Anomalies

-

Data Preprocessing Code Screenshots

Loading Data

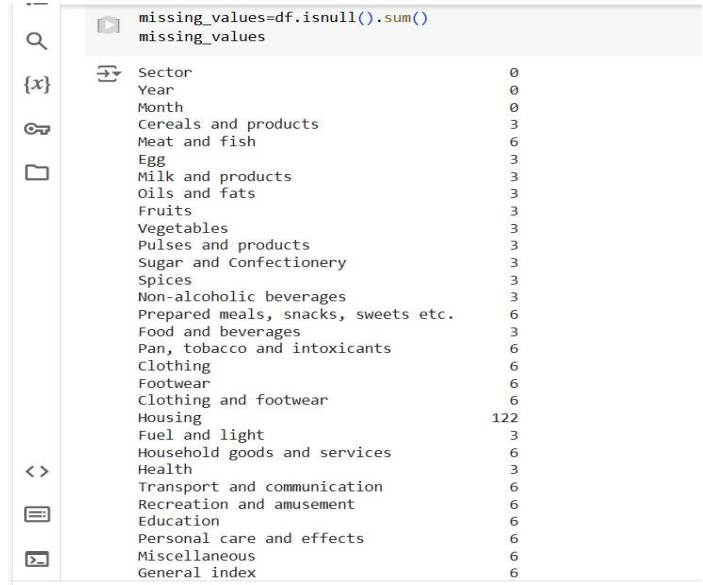
```
[1] df=pd.read_csv('content/archive (1) (1).zip')
```

Performing Different Operations To Understand Data

```
[1] df.head()
```

		Sector	Year	Month	Cereals and products	Meat and fish	Eggs	Milk and products	Oils and fats	Fruits	Vegetables	Housing	Fuel and light	Household goods and services	Health	Transport and communication	Recreation and amusement	Education	Personal care and effects	Miscellaneous
0	Rural	2013	January		107.5	106.3	108.1	104.9	106.1	103.9	101.9	...	NaN	105.5	104.8	104.0	103.3	103.4	103.8	104.7
1	Urban	2013	January		110.5	109.1	113.0	103.6	103.4	102.3	102.9	...	100.3	105.4	104.8	104.1	103.2	102.9	103.5	104.3
2	Rural-Urban	2013	January		108.4	107.3	110.0	104.4	105.1	103.2	102.2	...	100.3	105.5	104.8	104.0	103.2	103.1	103.6	104.5
3	Rural	2013	February		109.2	108.7	110.2	105.4	106.7	104.0	102.4	...	NaN	106.2	105.2	104.4	103.9	104.0	104.1	104.6
4	Urban	2013	February		112.9	112.9	116.9	104.0	103.5	103.1	104.9	...	100.4	105.7	105.2	104.7	104.4	103.3	103.7	104.3

Shows = 30 columns

Handling Missing Data	 <p>The screenshot shows a Jupyter Notebook cell with the following code and output:</p> <pre>missing_values=df.isnull().sum() missing_values</pre> <table border="1"> <thead> <tr> <th>Category</th> <th>Count</th> </tr> </thead> <tbody> <tr><td>Sector</td><td>0</td></tr> <tr><td>Year</td><td>0</td></tr> <tr><td>Month</td><td>0</td></tr> <tr><td>Cereals and products</td><td>3</td></tr> <tr><td>Meat and fish</td><td>6</td></tr> <tr><td>Egg</td><td>3</td></tr> <tr><td>Milk and products</td><td>3</td></tr> <tr><td>Oils and fats</td><td>3</td></tr> <tr><td>Fruits</td><td>3</td></tr> <tr><td>Vegetables</td><td>3</td></tr> <tr><td>Pulses and products</td><td>3</td></tr> <tr><td>Sugar and Confectionery</td><td>3</td></tr> <tr><td>Spices</td><td>3</td></tr> <tr><td>Non-alcoholic beverages</td><td>3</td></tr> <tr><td>Prepared meals, snacks, sweets etc.</td><td>6</td></tr> <tr><td>Food and beverages</td><td>3</td></tr> <tr><td>Pan, tobacco and intoxicants</td><td>6</td></tr> <tr><td>Clothing</td><td>6</td></tr> <tr><td>Footwear</td><td>6</td></tr> <tr><td>Clothing and footwear</td><td>6</td></tr> <tr><td>Housing</td><td>122</td></tr> <tr><td>Fuel and light</td><td>3</td></tr> <tr><td>Household goods and services</td><td>6</td></tr> <tr><td>Health</td><td>3</td></tr> <tr><td>Transport and communication</td><td>6</td></tr> <tr><td>Recreation and amusement</td><td>6</td></tr> <tr><td>Education</td><td>6</td></tr> <tr><td>Personal care and effects</td><td>6</td></tr> <tr><td>Miscellaneous</td><td>6</td></tr> <tr><td>General index</td><td>6</td></tr> </tbody> </table>	Category	Count	Sector	0	Year	0	Month	0	Cereals and products	3	Meat and fish	6	Egg	3	Milk and products	3	Oils and fats	3	Fruits	3	Vegetables	3	Pulses and products	3	Sugar and Confectionery	3	Spices	3	Non-alcoholic beverages	3	Prepared meals, snacks, sweets etc.	6	Food and beverages	3	Pan, tobacco and intoxicants	6	Clothing	6	Footwear	6	Clothing and footwear	6	Housing	122	Fuel and light	3	Household goods and services	6	Health	3	Transport and communication	6	Recreation and amusement	6	Education	6	Personal care and effects	6	Miscellaneous	6	General index	6
Category	Count																																																														
Sector	0																																																														
Year	0																																																														
Month	0																																																														
Cereals and products	3																																																														
Meat and fish	6																																																														
Egg	3																																																														
Milk and products	3																																																														
Oils and fats	3																																																														
Fruits	3																																																														
Vegetables	3																																																														
Pulses and products	3																																																														
Sugar and Confectionery	3																																																														
Spices	3																																																														
Non-alcoholic beverages	3																																																														
Prepared meals, snacks, sweets etc.	6																																																														
Food and beverages	3																																																														
Pan, tobacco and intoxicants	6																																																														
Clothing	6																																																														
Footwear	6																																																														
Clothing and footwear	6																																																														
Housing	122																																																														
Fuel and light	3																																																														
Household goods and services	6																																																														
Health	3																																																														
Transport and communication	6																																																														
Recreation and amusement	6																																																														
Education	6																																																														
Personal care and effects	6																																																														
Miscellaneous	6																																																														
General index	6																																																														
Feature Engineering	Enhance the accuracy and the robustness of the CPI predictions																																																														
Save Processed Data	-																																																														