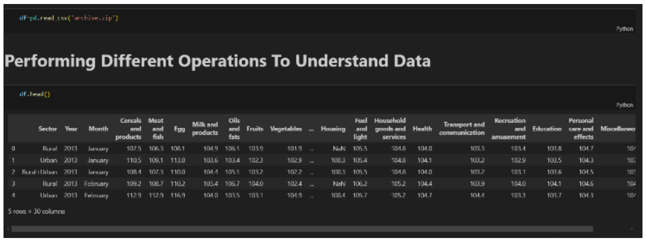
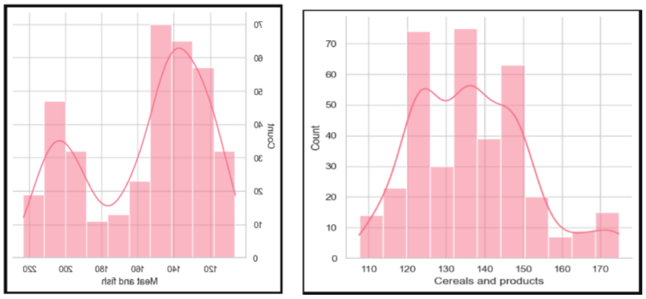
Data Collection and Preprocessing Phase

Date	05 June 2024
Team ID	739975
Project Title	To Predict Consumer Price Index
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	**Description**	**Screenshots**
Data Overview	**Basic Statistics:** Summarize the dataset with measures such as mean, median, standard deviation, and range for each category. **Dimensions:** Describe the size of the dataset (e.g., number of records, number of features). **Structure:** Explain the format and organization of the data, including any hierarchical structure (e.g., categories and subcategories of items).	
Univariate Analysis	Analyze each variable related to CPI, such as the price of vegetables, fruits, meat, housing, etc. Calculate and interpret measures of central tendency (mean, median, mode) and dispersion (variance, standard deviation) for each category. Visualize the distribution of each variable using histograms or box plots to understand their behavior over time	

```
df=pd.read_csv('archive.zip')
```

Python

Performing Different Operations To Understand Data

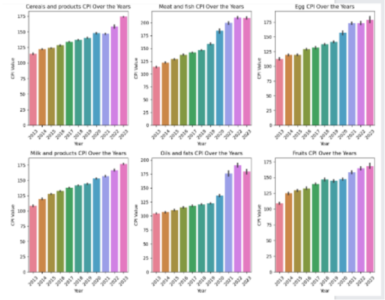
```
df.head()
```

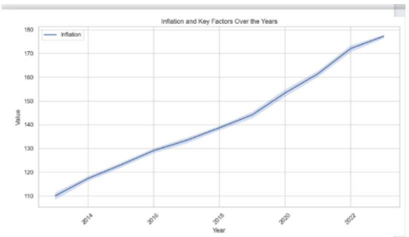
Python

	Sector	Year	Month	Cereals and products	Meat and fish	Egg	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	Miscellaneo
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	10
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	10
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	10
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	10
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	10

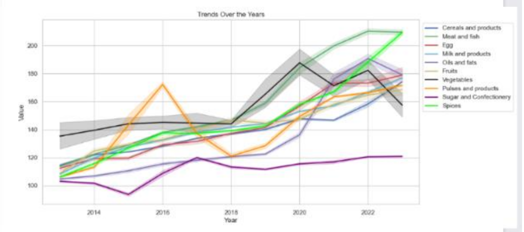
5 rows × 30 columns

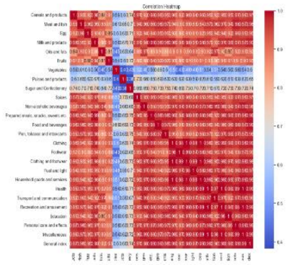
<table>
<tr>
<td>Bivariate Analysis</td>
<td>Investigate the relationships between pairs of variables, such as the price of pulses and the price of vegetables.
Calculate correlation coefficients to quantify the strength and direction of relationships between variables.
Use scatter plots to visualize these relationships and identify any patterns or trends.</td>
<td>



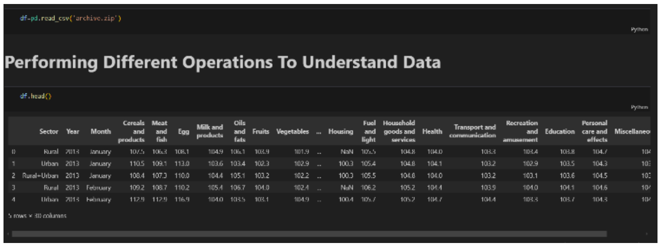


</td>
</tr>
<tr>
<td>Multivariate Analysis</td>
<td>Explore the interactions between multiple variables simultaneously, such as the combined effect of prices of different food items on the overall CPI.
Use techniques such as multiple regression analysis, principal component analysis (PCA), or cluster analysis to uncover underlying patterns and relationships.
Visualize multivariate relationships using pair plots, heatmaps, or 3D plots</td>
<td>





</td>
</tr>
<tr>
<td>Outliers and Anomalies</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td colspan="2"><b>Data Preprocessing Code Screenshots</b></td>
<td></td>
</tr>
<tr>
<td>Loading Data</td>
<td>Code to load the dataset into the preferred environment (e.g., Python, R).</td>
<td>



```
df-pd.read_csv('archive.zip')
```

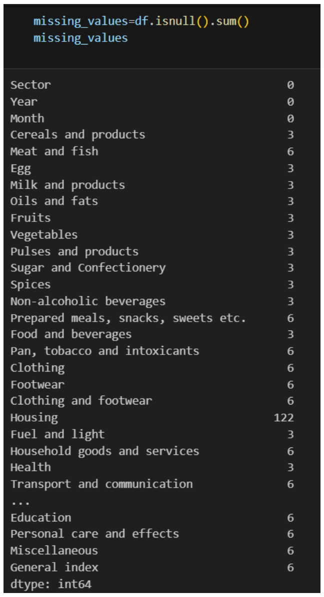
Performing Different Operations To Understand Data

```
df.head()
```

```
       Sector  Year     Month  Cereals   Meat    Egg  Milk and   Oils  Fruits  Vegetables  ...  Housing   Fuel  Household  Health  Transport and  Recreation  Education  Personal  Miscellaneo
                                   and    and         products    and                                         and  goods and          communication         and              care and
                              products   fish                    fats                                       light   services                          amusement               effects
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          104
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          103
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          103
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          104
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          104

5 rows × 30 columns
```

</td>
</tr>
</table>

<table>
<tr>
<td>Handling Missing Data</td>
<td>Code for identifying and handling missing values.</td>
<td>
<pre>
   missing_values=df.isnull().sum()
   missing_values

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
...
Education   6
Personal care and effects   6
Miscellaneous   6
General index   6
dtype: int64
</pre></td>
</tr>
<tr>
<td>Feature Engineering</td>
<td>Enhance the accuracy and the robustness of the CPI predictions</td>
<td></td>
</tr>
<tr>
<td>Save Processed Data</td>
<td>-</td>
<td></td>
</tr>
</table>