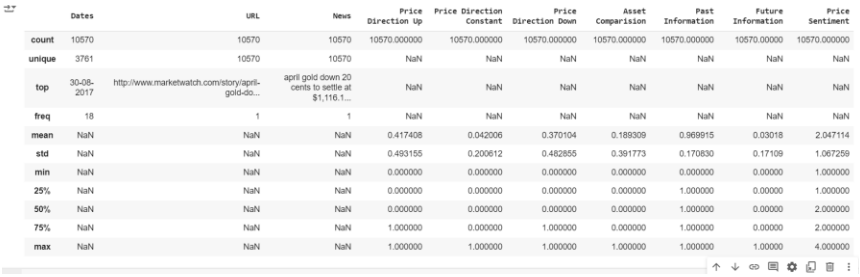
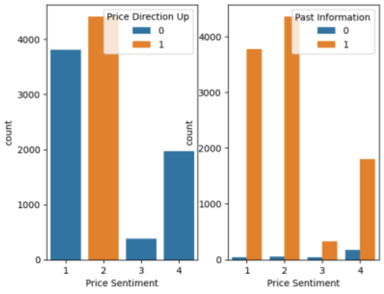
Data Collection and Preprocessing Phase

Date	14th July 2024
Team ID	739939
Project Title	Sentiment Analysis of Commodity News(Gold)
Maximum Marks	6 Marks

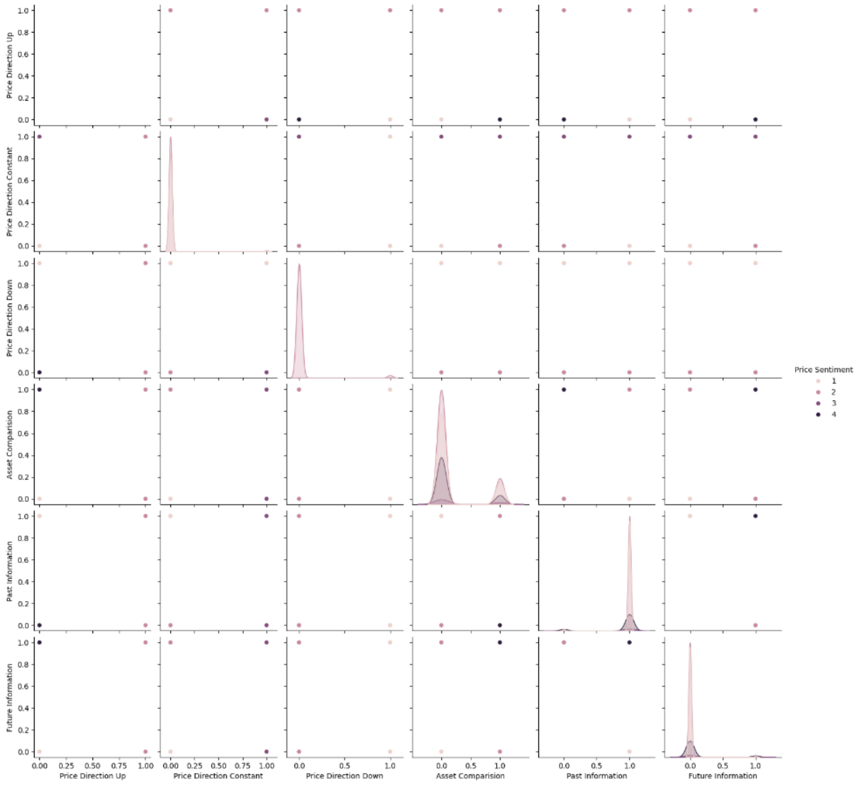
Data Exploration and Preprocessing Template

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

<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>Data Overview</td><td>Dimensions: (10570, 10)</td></tr>
<tr><td>Univariate Analysis</td><td>

<table>
<tr><th></th><th>Dates</th><th>URL</th><th>News</th><th>Price Direction Up</th><th>Price Direction Constant</th><th>Price Direction Down</th><th>Asset Comparision</th><th>Past Information</th><th>Future Information</th><th>Price Sentiment</th></tr>
<tr><td>count</td><td>10570</td><td>10570</td><td>10570</td><td>10570.000000</td><td>10570.000000</td><td>10570.000000</td><td>10570.000000</td><td>10570.000000</td><td>10570.00000</td><td>10570.000000</td></tr>
<tr><td>unique</td><td>3761</td><td>10570</td><td>10570</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td></tr>
<tr><td>top</td><td>30-08-2017</td><td>http://www.marketwatch.com/story/april-gold-do...</td><td>april gold down 20 cents to settle at $1,116.1...</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td></tr>
<tr><td>freq</td><td>18</td><td>1</td><td>1</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td></tr>
<tr><td>mean</td><td>NaN</td><td>NaN</td><td>NaN</td><td>0.417408</td><td>0.042006</td><td>0.370104</td><td>0.189309</td><td>0.969915</td><td>0.03018</td><td>2.047114</td></tr>
<tr><td>std</td><td>NaN</td><td>NaN</td><td>NaN</td><td>0.493155</td><td>0.200612</td><td>0.482855</td><td>0.391773</td><td>0.170830</td><td>0.17109</td><td>1.067259</td></tr>
<tr><td>min</td><td>NaN</td><td>NaN</td><td>NaN</td><td>0.000000</td><td>0.000000</td><td>0.000000</td><td>0.000000</td><td>0.000000</td><td>0.00000</td><td>1.000000</td></tr>
<tr><td>25%</td><td>NaN</td><td>NaN</td><td>NaN</td><td>0.000000</td><td>0.000000</td><td>0.000000</td><td>0.000000</td><td>1.000000</td><td>0.00000</td><td>1.000000</td></tr>
<tr><td>50%</td><td>NaN</td><td>NaN</td><td>NaN</td><td>0.000000</td><td>0.000000</td><td>0.000000</td><td>0.000000</td><td>1.000000</td><td>0.00000</td><td>2.000000</td></tr>
<tr><td>75%</td><td>NaN</td><td>NaN</td><td>NaN</td><td>1.000000</td><td>0.000000</td><td>1.000000</td><td>0.000000</td><td>1.000000</td><td>0.00000</td><td>2.000000</td></tr>
<tr><td>max</td><td>NaN</td><td>NaN</td><td>NaN</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.00000</td><td>4.000000</td></tr>
</table>
</td></tr>
<tr><td>Bivariate Analysis</td><td>


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

<table>
<tr><td>Multivariate Analysis</td><td>



</td></tr>
<tr><td>Outliers and Anomalies</td><td>-</td></tr>
<tr><td colspan="2">

Data Preprocessing Code Screenshots

</td></tr>
<tr><td>Loading Data</td><td>

```
df=pd.read_csv('/content/gold.csv')
```

</td></tr>
<tr><td>Handling Missing Data</td><td>

```
df.isnull().sum()
```

</td></tr>
<tr><td>Data Transformation</td><td>

```
def text_clean_1(text):
  text=text.lower()
  text=re.sub('\[.*?\]', '',text)
  text=re.sub('[%s]' %
re.escape(string.punctuation), '',text)
  text=re.sub('\w*\d\w','',text)
  text=re.sub('[‘’“”…]', '', text)
  text=re.sub('\n','',text)
  return text
```

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

	`Cleaned_News= lambda x: text_clean_1(x)`
Feature Engineering	NA

Save Processed Data	`import pickle`
	`pickle.dump(model,open('model.pkl','wb'))`