

Data Collection and Preprocessing Phase

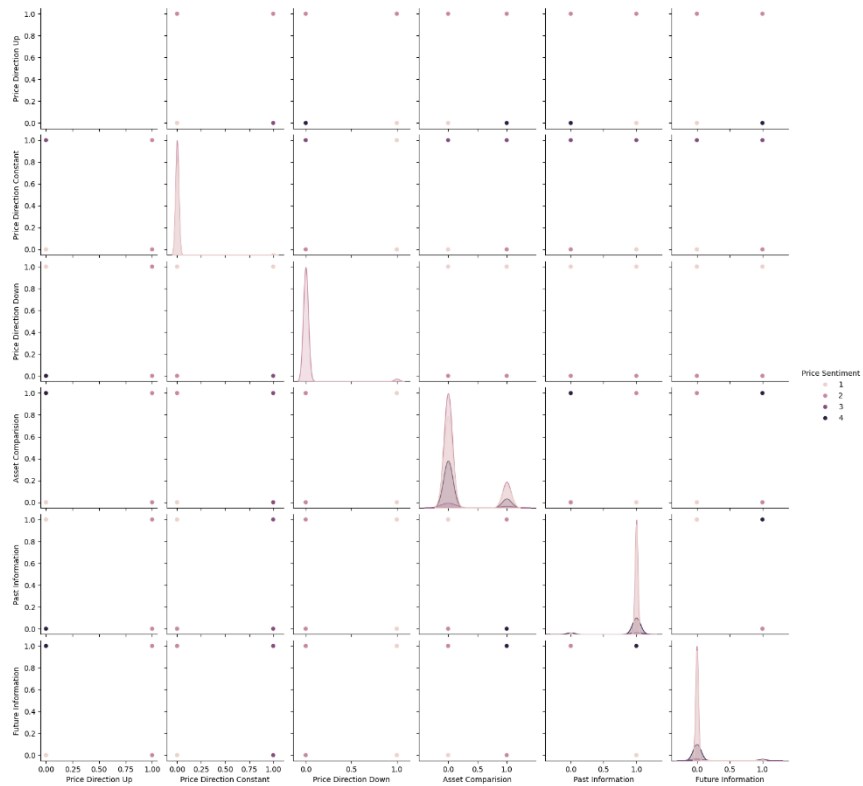
Date	15 th July 2024
Team ID	739691
Project Title	Sentiment Analysis of Commodity News(Gold)
Maximum Marks	6 Marks

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.

Section	Description																																																																																																																																				
Data Overview	Dimensions: (10570, 10)																																																																																																																																				
Univariate Analysis	<div><div><div></div><table><thead><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 Comparison</th><th>Past Information</th><th>Future Information</th><th>Price Sentiment</th></tr></thead><tbody><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>10570.000000</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>30-05-2017</td><td>http://www.marketwatch.com/story/april-gold-down-20-cents-to-settle-at-\$1,116.1...</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>0.042008</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>0.000000</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>0.000000</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>0.000000</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>0.000000</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>1.000000</td><td>4.000000</td></tr></tbody></table></div><div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div></div>		Dates	URL	News	Price Direction Up	Price Direction Constant	Price Direction Down	Asset Comparison	Past Information	Future Information	Price Sentiment	count	10570	10570	10570	10570.000000	10570.000000	10570.000000	10570.000000	10570.000000	10570.000000	10570.000000	unique	3761	10570	10570	NaN	NaN	NaN	NaN	NaN	NaN	NaN	top	30-05-2017	http://www.marketwatch.com/story/april-gold-down-20-cents-to-settle-at-\$1,116.1...	april gold down 20 cents to settle at \$1,116.1...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	freq	18	1	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	mean	NaN	NaN	NaN	0.417408	0.042008	0.370104	0.189309	0.969915	0.03018	2.047114	std	NaN	NaN	NaN	0.493155	0.200612	0.482855	0.391773	0.170830	0.17109	1.067259	min	NaN	NaN	NaN	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	25%	NaN	NaN	NaN	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	1.000000	50%	NaN	NaN	NaN	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	2.000000	75%	NaN	NaN	NaN	1.000000	0.000000	1.000000	0.000000	1.000000	0.000000	2.000000	max	NaN	NaN	NaN	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	4.000000
	Dates	URL	News	Price Direction Up	Price Direction Constant	Price Direction Down	Asset Comparison	Past Information	Future Information	Price Sentiment																																																																																																																											
count	10570	10570	10570	10570.000000	10570.000000	10570.000000	10570.000000	10570.000000	10570.000000	10570.000000																																																																																																																											
unique	3761	10570	10570	NaN	NaN	NaN	NaN	NaN	NaN	NaN																																																																																																																											
top	30-05-2017	http://www.marketwatch.com/story/april-gold-down-20-cents-to-settle-at-\$1,116.1...	april gold down 20 cents to settle at \$1,116.1...	NaN	NaN	NaN	NaN	NaN	NaN	NaN																																																																																																																											
freq	18	1	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN																																																																																																																											
mean	NaN	NaN	NaN	0.417408	0.042008	0.370104	0.189309	0.969915	0.03018	2.047114																																																																																																																											
std	NaN	NaN	NaN	0.493155	0.200612	0.482855	0.391773	0.170830	0.17109	1.067259																																																																																																																											
min	NaN	NaN	NaN	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000																																																																																																																											
25%	NaN	NaN	NaN	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	1.000000																																																																																																																											
50%	NaN	NaN	NaN	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	2.000000																																																																																																																											
75%	NaN	NaN	NaN	1.000000	0.000000	1.000000	0.000000	1.000000	0.000000	2.000000																																																																																																																											
max	NaN	NaN	NaN	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	4.000000																																																																																																																											
Bivariate Analysis	<div><div><div></div><div><div>Price Direction Up</div><div><div>0</div><div>1</div></div></div><div><div>count</div><div><div>4000</div><div>3000</div><div>2000</div><div>1000</div><div>0</div></div><div><div>1</div><div>2</div><div>3</div><div>4</div></div><div>Price Sentiment</div></div><div><div></div><div><div>Past information</div><div><div>0</div><div>1</div></div></div><div><div>count</div><div><div>4000</div><div>3000</div><div>2000</div><div>1000</div><div>0</div></div><div><div>1</div><div>2</div><div>3</div><div>4</div></div><div>Price Sentiment</div></div></div></div></div>																																																																																																																																				

Multivariate Analysis



Outliers and Anomalies

-

Data Preprocessing Code Screenshots

Loading Data

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

Handling Missing Data

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

Data Transformation

```
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
```

	<pre>Cleaned_News= lambda x: text_clean_1(x)</pre>
Feature Engineering	<pre>NA</pre>
Save Processed Data	<pre>import pickle pickle.dump(model,open('model.pkl','wb'))</pre>