Task1 : Data Exploration and Preprocessing

- Explore the dataset and identify the number of rows and columns.
- Check for missing values in each column and handle them accordingly.
- Perform data type conversion if necessary.Analyze the distribution of the target variable ("Aggregate rating") and identify any class imbalances.

1.importing the python packages

In [94]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Read the dataset from CSV file

In [95]:
```
rdata_df=pd.read_csv(r'C:\Users\bhanuprasad\Desktop\cognify intenship\Dataset.cs
rdata_df
```

Out[95]:

	Restaurant ID	Restaurant Name	Country Code	City	Address	Locality
0	6317637	Le Petit Souffle	162	Makati City	Third Floor, Century City Mall, Kalayaan Avenu...	Century City Mall, Poblacion, Makati City
1	6304287	Izakaya Kikufuji	162	Makati City	Little Tokyo, 2277 Chino Roces Avenue, Legaspi...	Little Tokyo, Legaspi Village, Makati City
2	6300002	Heat - Edsa Shangri-La	162	Mandaluyong City	Edsa Shangri-La, 1 Garden Way, Ortigas, Mandal...	Edsa Shangri-La, Ortigas, Mandaluyong City
3	6318506	Ooma	162	Mandaluyong City	Third Floor, Mega Fashion Hall, SM Megamall, O...	SM Megamall, Ortigas, Mandaluyong City
4	6314302	Sambo Kojin	162	Mandaluyong City	Third Floor, Mega Atrium, SM Megamall, Ortigas...	SM Megamall, Ortigas, Mandaluyong City
...	...	...	...	...	...	...
9546	5915730	Naml\ Gurme	208	��stanbul	Kemanke�� Karamustafa Pa��a Mahallesi, Rlhtlm ...	Karak�_y
9547	5908749	Ceviz A��acl	208	��stanbul	Ko��uyolu Mahallesi, Muhittin ��st�_nda�� Cadd...	Ko��uyolu
9548	5915807	Huqqa	208	��stanbul	Kuru�_e��me Mahallesi, Muallim Naci Caddesi, N...	Kuru�_e��me
9549	5916112	A���k Kahve	208	��stanbul	Kuru�_e��me Mahallesi, Muallim Naci Caddesi, N...	Kuru�_e��me
9550	5927402	Walter's Coffee Roastery	208	��stanbul	Cafea��a Mahallesi, Bademaltl Sokak, No 21/B, ...	Moda

9551 rows × 21 columns

Data Exploration

In [96]:
```
rdata_df=pd.read_csv('Dataset.csv')
rdata_df
```

Out[96]:

	Restaurant ID	Restaurant Name	Country Code	City	Address	Locality
0	6317637	Le Petit Souffle	162	Makati City	Third Floor, Century City Mall, Kalayaan Avenu...	Century City Mall, Poblacion, Makati City
1	6304287	Izakaya Kikufuji	162	Makati City	Little Tokyo, 2277 Chino Roces Avenue, Legaspi...	Little Tokyo, Legaspi Village, Makati City
2	6300002	Heat - Edsa Shangri-La	162	Mandaluyong City	Edsa Shangri-La, 1 Garden Way, Ortigas, Mandal...	Edsa Shangri-La, Ortigas, Mandaluyong City
3	6318506	Ooma	162	Mandaluyong City	Third Floor, Mega Fashion Hall, SM Megamall, O...	SM Megamall, Ortigas, Mandaluyong City
4	6314302	Sambo Kojin	162	Mandaluyong City	Third Floor, Mega Atrium, SM Megamall, Ortigas...	SM Megamall, Ortigas, Mandaluyong City
...	...	...	...	...	...	...
9546	5915730	Naml\ Gurme	208	��stanbul	Kemanke�� Karamustafa Pa��a Mahallesi, Rlhtlm ...	Karak�_y
9547	5908749	Ceviz A��acl	208	��stanbul	Ko��uyolu Mahallesi, Muhittin ��st�_nda�� Cadd...	Ko��uyolu
9548	5915807	Huqqa	208	��stanbul	Kuru�_e��me Mahallesi, Muallim Naci Caddesi, N...	Kuru�_e��me
9549	5916112	A���k Kahve	208	��stanbul	Kuru�_e��me Mahallesi, Muallim Naci Caddesi, N...	Kuru�_e��me
9550	5927402	Walter's Coffee Roastery	208	��stanbul	Cafea��a Mahallesi, Bademaltl Sokak, No 21/B, ...	Moda

9551 rows × 21 columns

In [27]:
```
rdata_df.head()
```

Out[27]:

	Restaurant ID	Restaurant Name	Country Code	City	Address	Locality	Localit Verbos
0	6317637	Le Petit Souffle	162	Makati City	Third Floor, Century City Mall, Kalayaan Avenu...	Century City Mall, Poblacion, Makati City	Century Cit Ma Poblacio Makati Cit Mak
1	6304287	Izakaya Kikufuji	162	Makati City	Little Tokyo, 2277 Chino Roces Avenue, Legaspi...	Little Tokyo, Legaspi Village, Makati City	Little Toky Legasp Villag Makati Cit Ma
2	6300002	Heat - Edsa Shangri-La	162	Mandaluyong City	Edsa Shangri-La, 1 Garden Way, Ortigas, Mandal...	Edsa Shangri-La, Ortigas, Mandaluyong City	Edsa Shangr La, Ortiga Mandaluyon City, Ma
3	6318506	Ooma	162	Mandaluyong City	Third Floor, Mega Fashion Hall, SM Megamall, O...	SM Megamall, Ortigas, Mandaluyong City	SM Megama Ortiga Mandaluyon Cit Mandal
4	6314302	Sambo Kojin	162	Mandaluyong City	Third Floor, Mega Atrium, SM Megamall, Ortigas...	SM Megamall, Ortigas, Mandaluyong City	SM Megama Ortiga Mandaluyon Cit Mandal

5 rows × 21 columns

In [97]:
```
rdata_df.head(10)
```

Out[97]:

	Restaurant ID	Restaurant Name	Country Code	City	Address	Locality	Locali Verbo
0	6317637	Le Petit Souffle	162	Makati City	Third Floor, Century City Mall, Kalayaan Avenu...	Century City Mall, Poblacion, Makati City	Century C Ma Poblacio Makati Ci Mal
1	6304287	Izakaya Kikufuji	162	Makati City	Little Tokyo, 2277 Chino Roces Avenue, Legaspi...	Little Tokyo, Legaspi Village, Makati City	Little Toky Legas Villag Makati Ci Ma
2	6300002	Heat - Edsa Shangri-La	162	Mandaluyong City	Edsa Shangri-La, 1 Garden Way, Ortigas, Mandal...	Edsa Shangri-La, Ortigas, Mandaluyong City	Edsa Shang La, Ortiga Mandaluyo City, Ma
3	6318506	Ooma	162	Mandaluyong City	Third Floor, Mega Fashion Hall, SM Megamall, O...	SM Megamall, Ortigas, Mandaluyong City	S Megama Ortiga Mandaluyo Ci Manda
4	6314302	Sambo Kojin	162	Mandaluyong City	Third Floor, Mega Atrium, SM Megamall, Ortigas...	SM Megamall, Ortigas, Mandaluyong City	S Megama Ortiga Mandaluyo Ci Manda
5	18189371	Din Tai Fung	162	Mandaluyong City	Ground Floor, Mega Fashion Hall, SM Megamall, ...	SM Megamall, Ortigas, Mandaluyong City	S Megama Ortiga Mandaluyo Ci Manda
6	6300781	Buffet 101	162	Pasay City	Building K, SM By The Bay, Sunset Boulevard, M...	SM by the Bay, Mall of Asia Complex, Pasay City	SM by t Bay, Mall As Comple Pasay Ci
7	6301290	Vikings	162	Pasay City	Building B, By The Bay, Seaside Boulevard, Mal...	SM by the Bay, Mall of Asia Complex, Pasay City	SM by t Bay, Mall As Comple Pasay Ci

	Restaurant ID	Restaurant Name	Country Code	City	Address	Locality	Locali Verbo
8	6300010	Spiral - Sofitel Philippine Plaza Manila	162	Pasay City	Plaza Level, Sofitel Philippine Plaza Manila, ...	Sofitel Philippine Plaza Manila, Pasay City	Sofi Philippi Plaza Mani Pasay Ci I
9	6314987	Locavore	162	Pasig City	Brixton Technology Center, 10 Brixton Street, ...	Kapitolyo	Kapitoly Pasig C

10 rows × 21 columns

In [98]:
```
rdata_df.tail()
```

Out[98]:

	Restaurant ID	Restaurant Name	Country Code	City	Address	Locality	
9546	5915730	Naml\ Gurme	208	��stanbul	Kemanke�� Karamustafa Pa��a Mahallesi, R\ht\m ...	Karak�_y	
9547	5908749	Ceviz A��ac\	208	��stanbul	Ko��uyolu Mahallesi, Muhittin ��st�_nda�� Cadd...	Ko��uyolu	
9548	5915807	Huqqa	208	��stanbul	Kuru�_e��me Mahallesi, Muallim Naci Caddesi, N...	Kuru�_e��me	Ku
9549	5916112	A���k Kahve	208	��stanbul	Kuru�_e��me Mahallesi, Muallim Naci Caddesi, N...	Kuru�_e��me	Ku
9550	5927402	Walter's Coffee Roastery	208	��stanbul	Cafea��a Mahallesi, Bademalt\ Sokak, No 21/B, ...	Moda	

5 rows × 21 columns

In [99]:
```
len(rdata_df) #count of the data
```

Out[99]:
```
9551
```

In [100…
```
rdata_df.size
```

Out[100…
```
200571
```

In [101…
```
rdata_df.columns
```

Out[101…
```
Index(['Restaurant ID', 'Restaurant Name', 'Country Code', 'City', 'Address',
       'Locality', 'Locality Verbose', 'Longitude', 'Latitude', 'Cuisines',
       'Average Cost for two', 'Currency', 'Has Table booking',
       'Has Online delivery', 'Is delivering now', 'Switch to order menu',
       'Price range', 'Aggregate rating', 'Rating color', 'Rating text',
       'Votes'],
      dtype='object')
```

In [102…
```
rdata_df.dtypes
```

Out[102…
```
Restaurant ID           int64
Restaurant Name        object
Country Code            int64
City                   object
Address                object
Locality               object
Locality Verbose       object
Longitude             float64
Latitude              float64
Cuisines               object
Average Cost for two    int64
Currency               object
Has Table booking      object
Has Online delivery    object
Is delivering now      object
Switch to order menu   object
Price range             int64
Aggregate rating      float64
Rating color           object
Rating text            object
Votes                   int64
dtype: object
```

In [103…
```
rdata_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9551 entries, 0 to 9550
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   Restaurant ID         9551 non-null   int64  
 1   Restaurant Name       9551 non-null   object 
 2   Country Code          9551 non-null   int64  
 3   City                  9551 non-null   object 
 4   Address               9551 non-null   object 
 5   Locality              9551 non-null   object 
 6   Locality Verbose      9551 non-null   object 
 7   Longitude             9551 non-null   float64
 8   Latitude              9551 non-null   float64
 9   Cuisines              9542 non-null   object 
 10  Average Cost for two  9551 non-null   int64  
 11  Currency              9551 non-null   object 
 12  Has Table booking     9551 non-null   object 
 13  Has Online delivery   9551 non-null   object 
 14  Is delivering now     9551 non-null   object 
 15  Switch to order menu  9551 non-null   object 
 16  Price range           9551 non-null   int64  
 17  Aggregate rating      9551 non-null   float64
 18  Rating color          9551 non-null   object 
 19  Rating text           9551 non-null   object 
 20  Votes                 9551 non-null   int64  
dtypes: float64(3), int64(5), object(13)
memory usage: 1.5+ MB
```

In [22]:

```python
def explore_dataset(table):

    print("top 5 rows - using head")
    print(table.head())
    print()

    print("bottom 5 rows using tail")
    print(table.tail())
    print()

    print("numbers of samples and columns")
    print(table.shape)
    print()

    print("numbers of samples ")
    print(len(table))
    print()

    print("numbers of entries in the data frame")
    print(table.size)
    print()

    print("Columns Names")
    print(table.columns)
    print()

    print("Columns dtypes")
    print(table.dtypes)
    print()

    print("Dataframe info")
```

```
    print(table.info())
    print()

explore_dataset(rdata_df)
```

top 5 rows - using head

	Restaurant ID	Restaurant Name	Country Code	City	Address	Locality	Locality Verbose	Longitude	Latitude	Cuisines	...	Currency	Has Table booking	Has Online delivery	Is delivering now	Switch to order menu	Price range	Aggregate rating	Rating color	Rating text	Votes
0	6317637	Le Petit Souffle	162	Makati City	Third Floor, Century City Mall, Kalayaan Avenu...	Century City Mall, Poblacion, Makati City	Century City Mall, Poblacion, Makati City, Mak...	121.027535	14.565443	French, Japanese, Desserts	...	Botswana Pula(P)	Yes	No	No	No	3	4.8	Dark Green	Excellent	314
1	6304287	Izakaya Kikufuji	162	Makati City	Little Tokyo, 2277 Chino Roces Avenue, Legaspi...	Little Tokyo, Legaspi Village, Makati City	Little Tokyo, Legaspi Village, Makati City, Ma...	121.014101	14.553708	Japanese	...	Botswana Pula(P)	Yes	No	No	No	3	4.5	Dark Green	Excellent	591
2	6300002	Heat - Edsa Shangri-La	162	Mandaluyong City	Edsa Shangri-La, 1 Garden Way, Ortigas, Mandal...	Edsa Shangri-La, Ortigas, Mandaluyong City	Edsa Shangri-La, Ortigas, Mandaluyong City, Ma...	121.056831	14.581404	Seafood, Asian, Filipino, Indian	...	Botswana Pula(P)	Yes	No	No	No	4	4.4	Green	Very Good	270
3	6318506	Ooma	162	Mandaluyong City	Third Floor, Mega Fashion Hall, SM Megamall, O...	SM Megamall, Ortigas, Mandaluyong City	SM Megamall, Ortigas, Mandaluyong City, Mandal...	121.056475	14.585318	Japanese, Sushi	...	Botswana Pula(P)	No	No	No	No	4	4.9	Dark Green	Excellent	365
4	6314302	Sambo Kojin	162	Mandaluyong City	Third Floor, Mega Atrium, SM Megamall, Ortigas...	SM Megamall, Ortigas, Mandaluyong City	SM Megamall, Ortigas, Mandaluyong City, Mandal...	121.057508	14.584450	Japanese, Korean	...	Botswana Pula(P)	Yes	No	No	No	4	4.8	Dark Green	Excellent	229

[5 rows x 21 columns]

bottom 5 rows using tail

	Restaurant ID	Restaurant Name	Country Code	City
9546	5915730	Naml\ Gurme	208	��stanbul
9547	5908749	Ceviz A��ac\	208	��stanbul
9548	5915807	Huqqa	208	��stanbul
9549	5916112	A���k Kahve	208	��stanbul
9550	5927402	Walter's Coffee Roastery	208	��stanbul

```
                                                Address     Locality  \
9546  Kemanke�� Karamustafa Pa��a Mahallesi, R\ht\m ...     Karak�_y
9547  Ko��uyolu Mahallesi, Muhittin ��st�_nda�� Cadd...    Ko��uyolu
9548  Kuru�_e��me Mahallesi, Muallim Naci Caddesi, N...  Kuru�_e��me
9549  Kuru�_e��me Mahallesi, Muallim Naci Caddesi, N...  Kuru�_e��me
9550  Cafea��a Mahallesi, Bademalt\ Sokak, No 21/B, ...         Moda

           Locality Verbose  Longitude   Latitude  \
9546     Karak�_y, ��stanbul  28.977392  41.022793
9547    Ko��uyolu, ��stanbul  29.041297  41.009847
9548  Kuru�_e��me, ��stanbul  29.034640  41.055817
9549  Kuru�_e��me, ��stanbul  29.036019  41.057979
9550         Moda, ��stanbul  29.026016  40.984776

                           Cuisines  ...          Currency  \
9546                         Turkish  ...  Turkish Lira(TL)
9547  World Cuisine, Patisserie, Cafe  ...  Turkish Lira(TL)
9548          Italian, World Cuisine  ...  Turkish Lira(TL)
9549                 Restaurant Cafe  ...  Turkish Lira(TL)
9550                            Cafe  ...  Turkish Lira(TL)

     Has Table booking Has Online delivery Is delivering now  \
9546                No                  No                No
9547                No                  No                No
9548                No                  No                No
9549                No                  No                No
9550                No                  No                No

     Switch to order menu Price range  Aggregate rating  Rating color  \
9546                   No           3               4.1         Green
9547                   No           3               4.2         Green
9548                   No           4               3.7        Yellow
9549                   No           4               4.0         Green
9550                   No           2               4.0         Green

     Rating text Votes
9546   Very Good   788
9547   Very Good  1034
9548        Good   661
9549   Very Good   901
9550   Very Good   591

[5 rows x 21 columns]

numbers of samples and columns
(9551, 21)

numbers of samples
9551

numbers of entries in the data frame
200571

Columns Names
Index(['Restaurant ID', 'Restaurant Name', 'Country Code', 'City', 'Address',
       'Locality', 'Locality Verbose', 'Longitude', 'Latitude', 'Cuisines',
       'Average Cost for two', 'Currency', 'Has Table booking',
       'Has Online delivery', 'Is delivering now', 'Switch to order menu',
       'Price range', 'Aggregate rating', 'Rating color', 'Rating text',
       'Votes'],
```

```
      dtype='object')

Columns dtypes
Restaurant ID             int64
Restaurant Name          object
Country Code              int64
City                     object
Address                  object
Locality                 object
Locality Verbose         object
Longitude               float64
Latitude                float64
Cuisines                 object
Average Cost for two      int64
Currency                 object
Has Table booking        object
Has Online delivery      object
Is delivering now        object
Switch to order menu     object
Price range               int64
Aggregate rating        float64
Rating color             object
Rating text              object
Votes                     int64
dtype: object

Dataframe info
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9551 entries, 0 to 9550
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Restaurant ID         9551 non-null   int64
 1   Restaurant Name       9551 non-null   object
 2   Country Code          9551 non-null   int64
 3   City                  9551 non-null   object
 4   Address               9551 non-null   object
 5   Locality              9551 non-null   object
 6   Locality Verbose      9551 non-null   object
 7   Longitude             9551 non-null   float64
 8   Latitude              9551 non-null   float64
 9   Cuisines              9551 non-null   object
 10  Average Cost for two  9551 non-null   int64
 11  Currency              9551 non-null   object
 12  Has Table booking     9551 non-null   object
 13  Has Online delivery   9551 non-null   object
 14  Is delivering now     9551 non-null   object
 15  Switch to order menu  9551 non-null   object
 16  Price range           9551 non-null   int64
 17  Aggregate rating      9551 non-null   float64
 18  Rating color          9551 non-null   object
 19  Rating text           9551 non-null   object
 20  Votes                 9551 non-null   int64
dtypes: float64(3), int64(5), object(13)
memory usage: 1.5+ MB
None
```

identify the number of rows and columns

In [105…
```
print('the number of rows in restaurnt dataset',rdata_df.shape[0])
print('the number of rows in restaurnt dataset',rdata_df.shape[1])
```

```
the number of rows in restaurnt dataset 9551
the number of rows in restaurnt dataset 21
```

Check for missing values in each column and handle them accordingly.

In [106…
```
rdata_df.isnull() # shows the missing values in data set
```

Out[106…

	Restaurant ID	Restaurant Name	Country Code	City	Address	Locality	Locality Verbose	Longitude	L
0	False	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	False	
…	…	…	…	…	…	…	…	…	
9546	False	False	False	False	False	False	False	False	
9547	False	False	False	False	False	False	False	False	
9548	False	False	False	False	False	False	False	False	
9549	False	False	False	False	False	False	False	False	
9550	False	False	False	False	False	False	False	False	

9551 rows × 21 columns

In [107…
```
rdata_df.isnull().sum()
# it will be shows the columns and related null values.
```

Out[107…

```
Restaurant ID           0
Restaurant Name         0
Country Code            0
City                    0
Address                 0
Locality                0
Locality Verbose        0
Longitude               0
Latitude                0
Cuisines                9
Average Cost for two    0
Currency                0
Has Table booking       0
Has Online delivery     0
Is delivering now       0
Switch to order menu    0
Price range             0
Aggregate rating        0
Rating color            0
Rating text             0
Votes                   0
dtype: int64
```

In [108…

```
rdata_df.isnull().sum()
```

Out[108…

```
Restaurant ID           0
Restaurant Name         0
Country Code            0
City                    0
Address                 0
Locality                0
Locality Verbose        0
Longitude               0
Latitude                0
Cuisines                9
Average Cost for two    0
Currency                0
Has Table booking       0
Has Online delivery     0
Is delivering now       0
Switch to order menu    0
Price range             0
Aggregate rating        0
Rating color            0
Rating text             0
Votes                   0
dtype: int64
```

In [109…

```
categorical=rdata_df.select_dtypes(include='object').columns
numerical=rdata_df.select_dtypes(exclude='object').columns
```

In [46]:

```
categorical
```

Out[46]:

```
Index(['Restaurant Name', 'City', 'Address', 'Locality', 'Locality Verbose',
       'Cuisines', 'Currency', 'Has Table booking', 'Has Online delivery',
       'Is delivering now', 'Switch to order menu', 'Rating color',
       'Rating text'],
      dtype='object')
```

In [47]:
```
numerical
```

Out[47]:
```
Index(['Restaurant ID', 'Country Code', 'Longitude', 'Latitude',
       'Average Cost for two', 'Price range', 'Aggregate rating', 'Votes'],
      dtype='object')
```

In [5]:
```
# here completed the handiling missing values.
import pandas as pd
rdata_df=pd.read_csv('Dataset.csv')
```

In [6]:
```
rdata_df.fillna(rdata_df.mean(numeric_only=True), inplace=True)
rdata_df.fillna(rdata_df.mode().iloc[0], inplace=True)
```

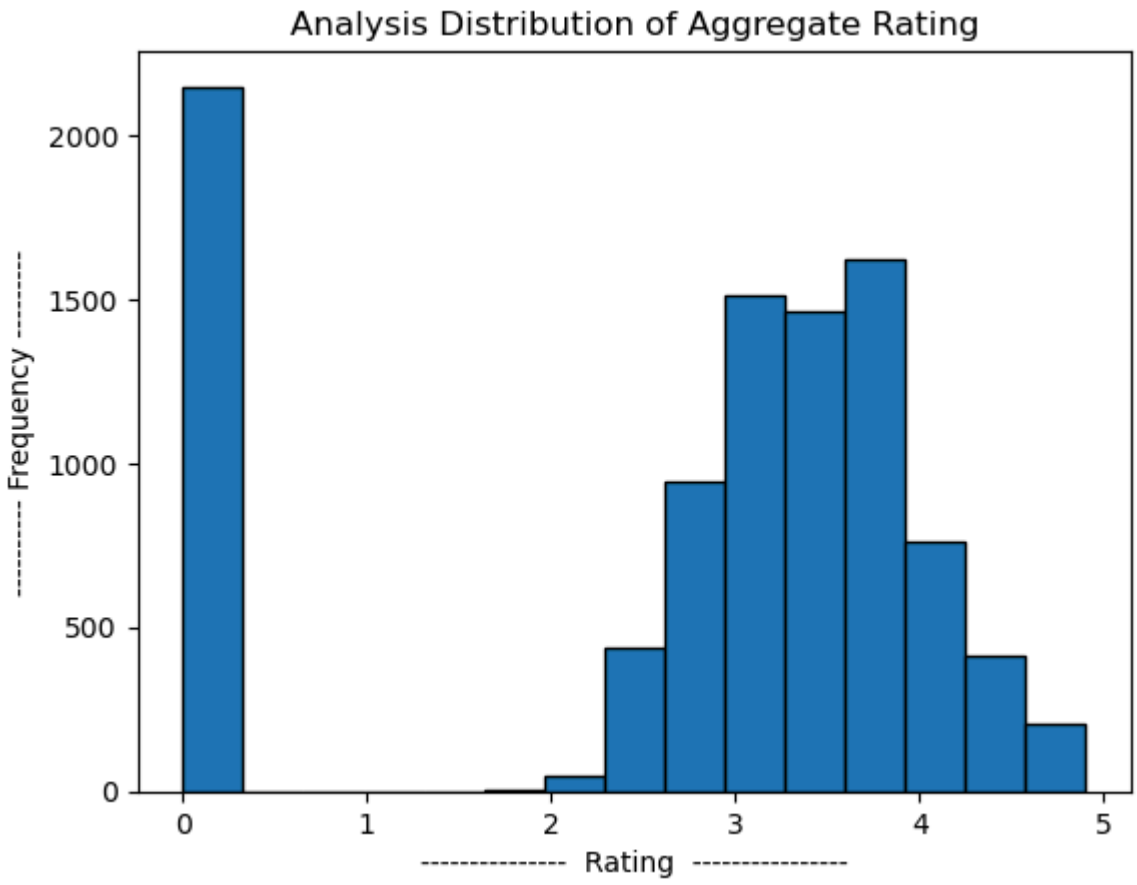
In [7]:
```
rdata_df.isnull().sum().sum() # count of the null values in dataset.
```

Out[7]:
```
0
```

Perform data type conversion if necessary.Analyze the distribution of the target variable

("Aggregate rating") and identify any class imbalances.

In [15]:
```
rdata_df['Aggregate rating'] = pd.to_numeric(rdata_df['Aggregate rating'], error
plt.hist(rdata_df['Aggregate rating'].dropna(), bins=15, edgecolor='black')
plt.title('Analysis Distribution of Aggregate Rating')
plt.xlabel('---------------  Rating  ----------------')
plt.ylabel('---------- Frequency ---------')
plt.show()
```

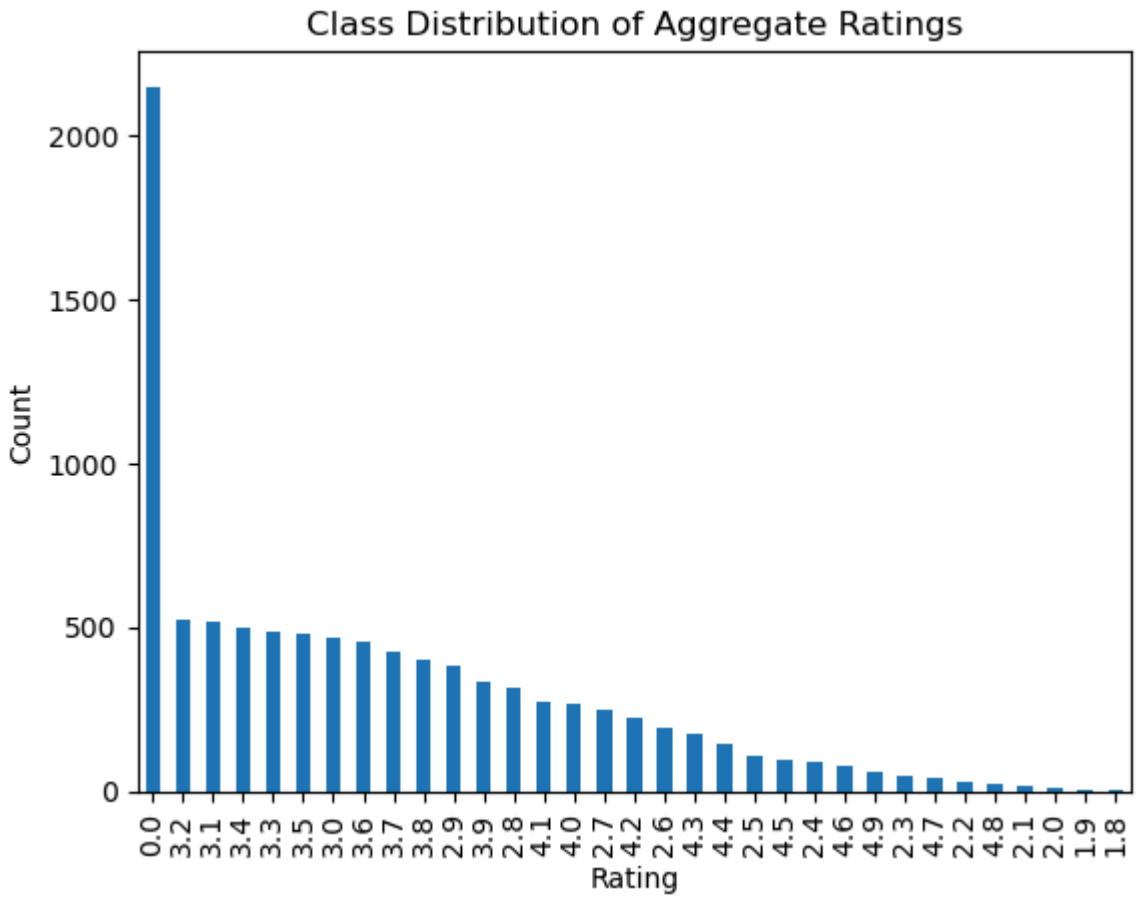


In [20]:
```
# Identify class imbalances ,distribution of each rating value
class_distribution = rdata_df['Aggregate rating'].value_counts()
print(class_distribution)
```

```
Aggregate rating
0.0    2148
3.2     522
3.1     519
3.4     498
3.3     483
3.5     480
3.0     468
3.6     458
3.7     427
3.8     400
2.9     381
3.9     335
2.8     315
4.1     274
4.0     266
2.7     250
4.2     221
2.6     191
4.3     174
4.4     144
2.5     110
4.5      95
2.4      87
4.6      78
4.9      61
2.3      47
4.7      42
2.2      27
4.8      25
2.1      15
2.0       7
1.9       2
1.8       1
Name: count, dtype: int64
```

In [21]:

```python
# here  class imbalance detection
class_distribution.plot(kind='bar')
plt.title('Class Distribution of Aggregate Ratings')
plt.xlabel('Rating')
plt.ylabel('Count')
plt.show()
```


Class Distribution of Aggregate Ratings
Count
2000
1500
1000
500
0
0.0
3.2
3.1
3.4
3.3
3.5
3.0
3.6
3.7
3.8
2.9
3.9
2.8
4.1
4.0
2.7
4.2
2.6
4.3
4.4
2.5
4.5
2.4
4.6
4.9
2.3
4.7
2.2
4.8
2.1
2.0
1.9
1.8
Rating

In [ ]: