Project: Milestone 3

In [4]:
```
import pandas as pd
import numpy as np
from datetime import datetime
import matplotlib.pyplot as plt
import seaborn as sns
from fuzzywuzzy import fuzz
```

In [5]:
```
# Load CSV file
df = pd.read_csv('/content/amazon_vfl_reviews.csv')
```

In [6]:
```
# Replace null values with mean/average value
df.fillna(df.mean(), inplace=True)
```

```
<ipython-input-6-b8a133dcd53b>:2: FutureWarning: The default value of numeric_only in
DataFrame.mean is deprecated. In a future version, it will default to False. In addit
ion, specifying 'numeric_only=None' is deprecated. Select only valid columns or speci
fy the value of numeric_only to silence this warning.
  df.fillna(df.mean(), inplace=True)
```

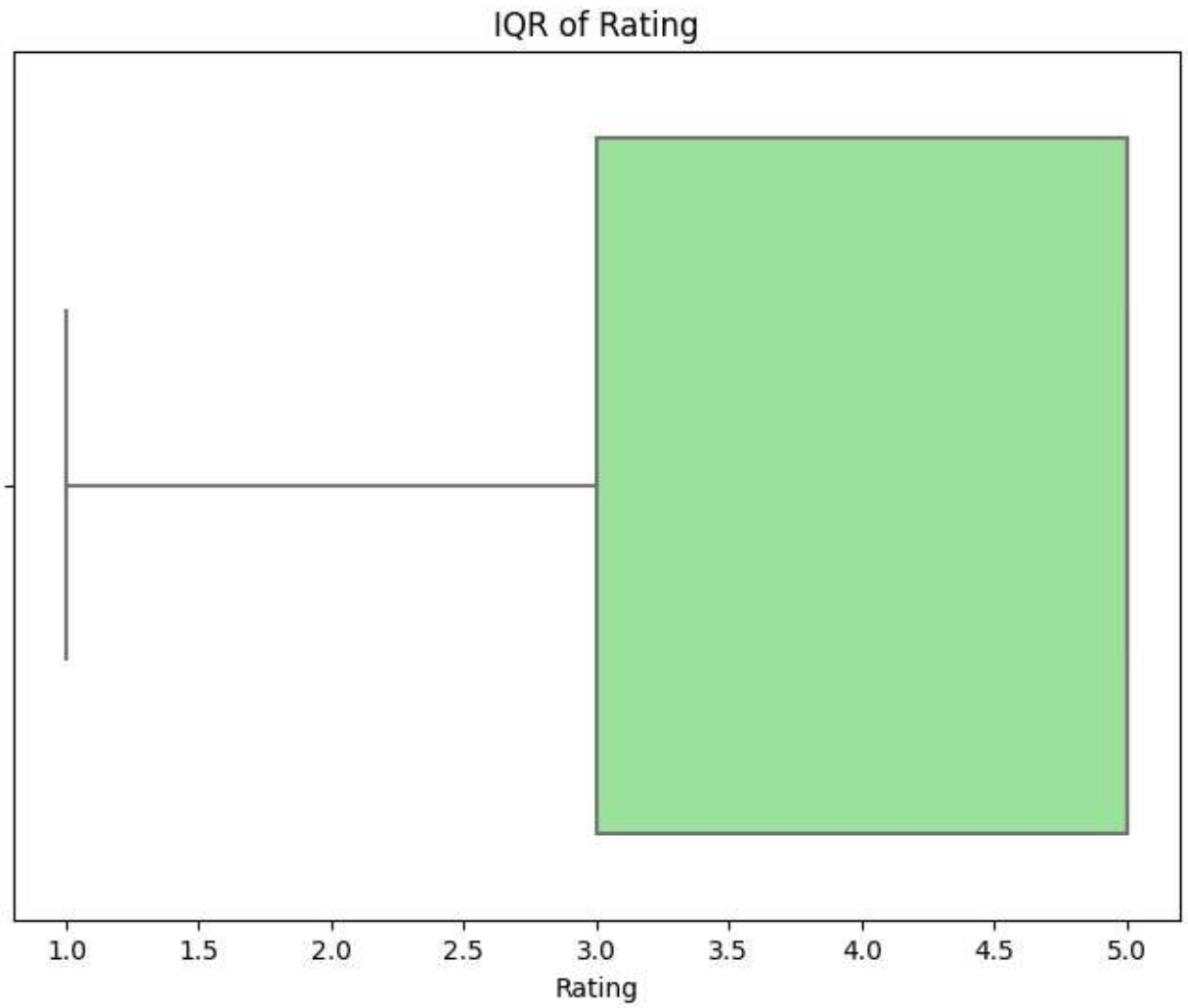
In [7]:
```
# Convert 'date' column to datetime
df['date'] = pd.to_datetime(df['date'], format='%Y-%m-%d')
```

In [8]:
```
# Check outliers using IQR
Q1 = df['rating'].quantile(0.25)
Q3 = df['rating'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
outliers = df[(df['rating'] < lower_bound) | (df['rating'] > upper_bound)]
print("Outliers:")
print(outliers)

# Remove outliers
df = df[(df['rating'] >= lower_bound) & (df['rating'] <= upper_bound)]
```

```
Outliers:
Empty DataFrame
Columns: [asin, name, date, rating, review]
Index: []
```

In [9]:
```
# Visualize IQR with box plot
plt.figure(figsize=(8, 6))
sns.boxplot(x=df['rating'], color='lightgreen')
plt.xlabel('Rating')
plt.title('IQR of Rating')
plt.show()
```

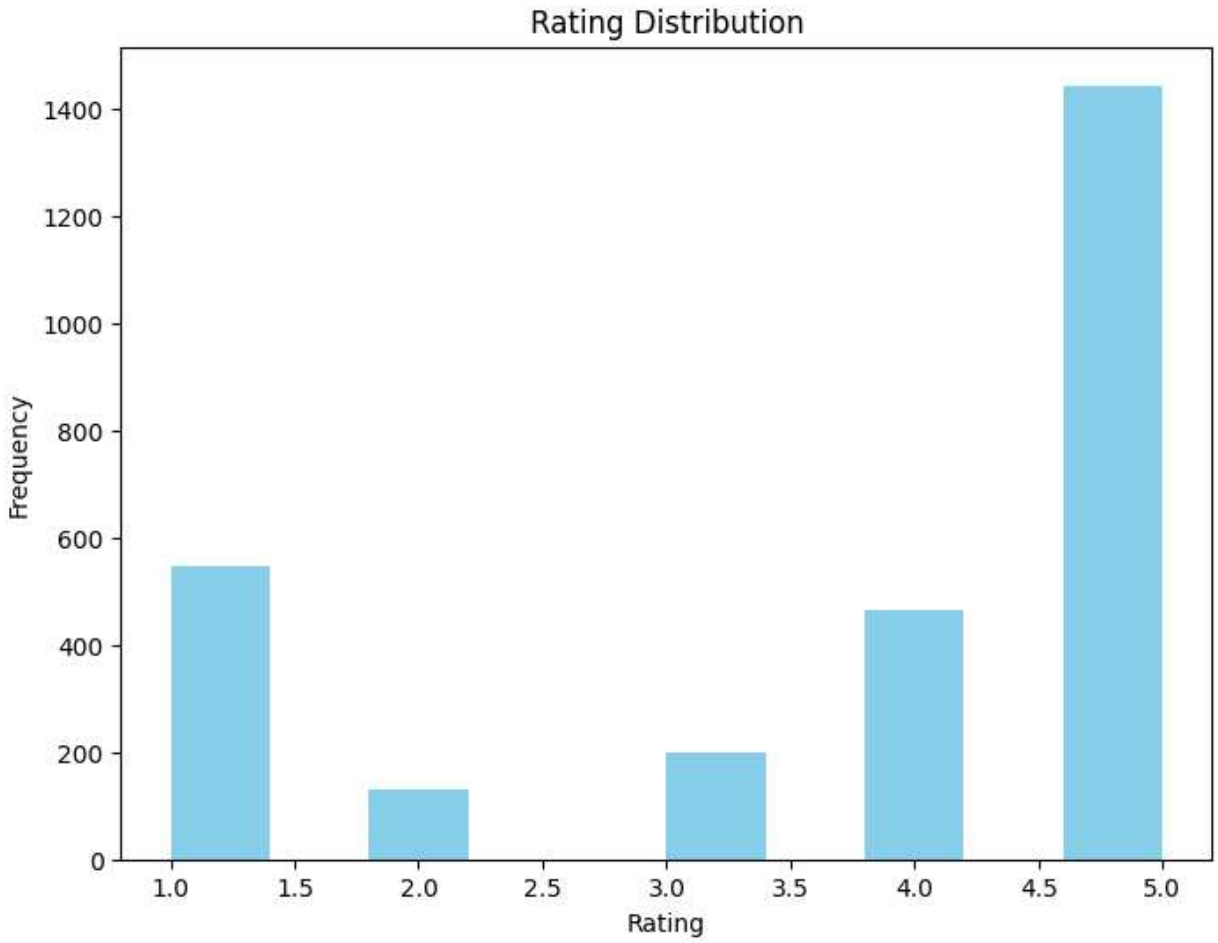


In [10]:
```python
# Data formatting
df['rating'] = df['rating'].astype(int)  # Convert rating to integers

# Fuzzy matching
def fuzzy_match(name, options):
    best_match = None
    highest_ratio = -1
    for option in options:
        ratio = fuzz.ratio(name, option)
        if ratio > highest_ratio:
            highest_ratio = ratio
            best_match = option
    return best_match

names = df['name'].unique()
df['fuzzy_match'] = df['name'].apply(lambda x: fuzzy_match(x, names))
```

In [11]:
```python
# Visualize rating with histogram
plt.figure(figsize=(8, 6))
plt.hist(df['rating'], bins=10, color='skyblue')
plt.xlabel('Rating')
plt.ylabel('Frequency')
plt.title('Rating Distribution')
plt.show()
```



In [12]:
```
# Print transformed DataFrame
print(df)
```

```
           asin                                        name       date  rating  \
0     B07W7CTLD1  Mamaearth-Onion-Growth-Control-Redensyl 2019-09-06       1
1     B07W7CTLD1  Mamaearth-Onion-Growth-Control-Redensyl 2019-08-14       5
2     B07W7CTLD1  Mamaearth-Onion-Growth-Control-Redensyl 2019-10-19       1
3     B07W7CTLD1  Mamaearth-Onion-Growth-Control-Redensyl 2019-09-16       1
4     B07W7CTLD1  Mamaearth-Onion-Growth-Control-Redensyl 2019-08-18       5
...          ...                                       ...        ...     ...
2777  B07MVHJ6CH            Mysore-Sandal-Soaps-Pack-Bars 2020-03-01       5
2778  B07MVHJ6CH            Mysore-Sandal-Soaps-Pack-Bars 2019-10-24       5
2779  B07MVHJ6CH            Mysore-Sandal-Soaps-Pack-Bars 2020-10-03       2
2780  B07MVHJ6CH            Mysore-Sandal-Soaps-Pack-Bars 2019-06-21       4
2781  B07MVHJ6CH            Mysore-Sandal-Soaps-Pack-Bars 2020-07-03       5

                                                 review  \
0     I bought this hair oil after viewing so many g...
1     Used This Mama Earth Newly Launched Onion Oil ...
2     So bad product...My hair falling increase too ...
3     Product just smells similar to navarathna hair...
4     I have been trying different onion oil for my ...
...                                                 ...
2777        Long lasting freshness throughout the day.
2778                                My preferred soap
2779                                     ठीक नहीं लगा
2780                                    Super Product
2781  Best soothing, cooling fragrance for hot summe...

                               fuzzy_match
0     Mamaearth-Onion-Growth-Control-Redensyl
1     Mamaearth-Onion-Growth-Control-Redensyl
2     Mamaearth-Onion-Growth-Control-Redensyl
3     Mamaearth-Onion-Growth-Control-Redensyl
4     Mamaearth-Onion-Growth-Control-Redensyl
...                                      ...
2777           Mysore-Sandal-Soaps-Pack-Bars
2778           Mysore-Sandal-Soaps-Pack-Bars
2779           Mysore-Sandal-Soaps-Pack-Bars
2780           Mysore-Sandal-Soaps-Pack-Bars
2781           Mysore-Sandal-Soaps-Pack-Bars

[2782 rows x 6 columns]
```