In [3]:
```
# Import necessary libraries

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt



#Read the data

df = pd.read_csv(r"C:\Users\Vaish\Downloads\archive (1)\Reviews.csv", nrows=500)



# Look at the top 5 rows of the data

df.head(3)
```

Out[3]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessD
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 10 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   Id                      500 non-null    int64 
 1   ProductId               500 non-null    object
 2   UserId                  500 non-null    object
 3   ProfileName             500 non-null    object
 4   HelpfulnessNumerator    500 non-null    int64 
 5   HelpfulnessDenominator  500 non-null    int64 
 6   Score                   500 non-null    int64 
 7   Time                    500 non-null    int64 
 8   Summary                 500 non-null    object
 9   Text                    500 non-null    object
dtypes: int64(5), object(5)
memory usage: 39.2+ KB
```

In [7]:
```
#summary of reviews
df.Summary.head()
```

Out[7]:
```
0    Good Quality Dog Food
1        Not as Advertised
2    "Delight" says it all
3           Cough Medicine
4              Great taffy
Name: Summary, dtype: object
```

In [9]:
```
df.Text.head()
```

Out[9]:
```
0    I have bought several of the Vitality canned d...
1    Product arrived labeled as Jumbo Salted Peanut...
2    This is a confection that has been around a fe...
3    If you are looking for the secret ingredient i...
4    Great taffy at a great price.  There was a wid...
Name: Text, dtype: object
```

In [25]:
```
!pip install textblob
!python -m textblob.download_corpora
```

```
Collecting textblob
  Downloading textblob-0.19.0-py3-none-any.whl.metadata (4.4 kB)
Collecting nltk>=3.9 (from textblob)
  Using cached nltk-3.9.1-py3-none-any.whl.metadata (2.9 kB)
Requirement already satisfied: click in c:\users\vaish\anaconda3\lib\site-packages
(from nltk>=3.9->textblob) (8.1.7)
Requirement already satisfied: joblib in c:\users\vaish\anaconda3\lib\site-packages
(from nltk>=3.9->textblob) (1.4.2)
Requirement already satisfied: regex>=2021.8.3 in c:\users\vaish\anaconda3\lib\site-
packages (from nltk>=3.9->textblob) (2023.10.3)
Requirement already satisfied: tqdm in c:\users\vaish\anaconda3\lib\site-packages (f
rom nltk>=3.9->textblob) (4.66.4)
Requirement already satisfied: colorama in c:\users\vaish\anaconda3\lib\site-package
s (from click->nltk>=3.9->textblob) (0.4.6)
Downloading textblob-0.19.0-py3-none-any.whl (624 kB)
   ---------------------------------------- 0.0/624.3 kB ? eta -:--:--
   ---------------------------------------- 0.0/624.3 kB ? eta -:--:--
   ---------------------------------------- 0.0/624.3 kB ? eta -:--:--
   -- ------------------------------------- 41.0/624.3 kB 991.0 kB/s eta 0:00:01
   ---------------------------- ---------- 460.8/624.3 kB 5.8 MB/s eta 0:00:01
   ---------------------------------------- 624.3/624.3 kB 4.9 MB/s eta 0:00:00
Using cached nltk-3.9.1-py3-none-any.whl (1.5 MB)
Installing collected packages: nltk, textblob
  Attempting uninstall: nltk
    Found existing installation: nltk 3.8.1
    Uninstalling nltk-3.8.1:
      Successfully uninstalled nltk-3.8.1
Successfully installed nltk-3.9.1 textblob-0.19.0
Finished.
```

```
[nltk_data] Downloading package brown to
[nltk_data]     C:\Users\Vaish\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\brown.zip.
[nltk_data] Downloading package punkt_tab to
[nltk_data]     C:\Users\Vaish\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping tokenizers\punkt_tab.zip.
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\Vaish\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger_eng to
[nltk_data]     C:\Users\Vaish\AppData\Roaming\nltk_data...
[nltk_data]   Package averaged_perceptron_tagger_eng is already up-to-
[nltk_data]       date!
[nltk_data] Downloading package conll2000 to
[nltk_data]     C:\Users\Vaish\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\conll2000.zip.
[nltk_data] Downloading package movie_reviews to
[nltk_data]     C:\Users\Vaish\AppData\Roaming\nltk_data...
[nltk_data]   Package movie_reviews is already up-to-date!
```

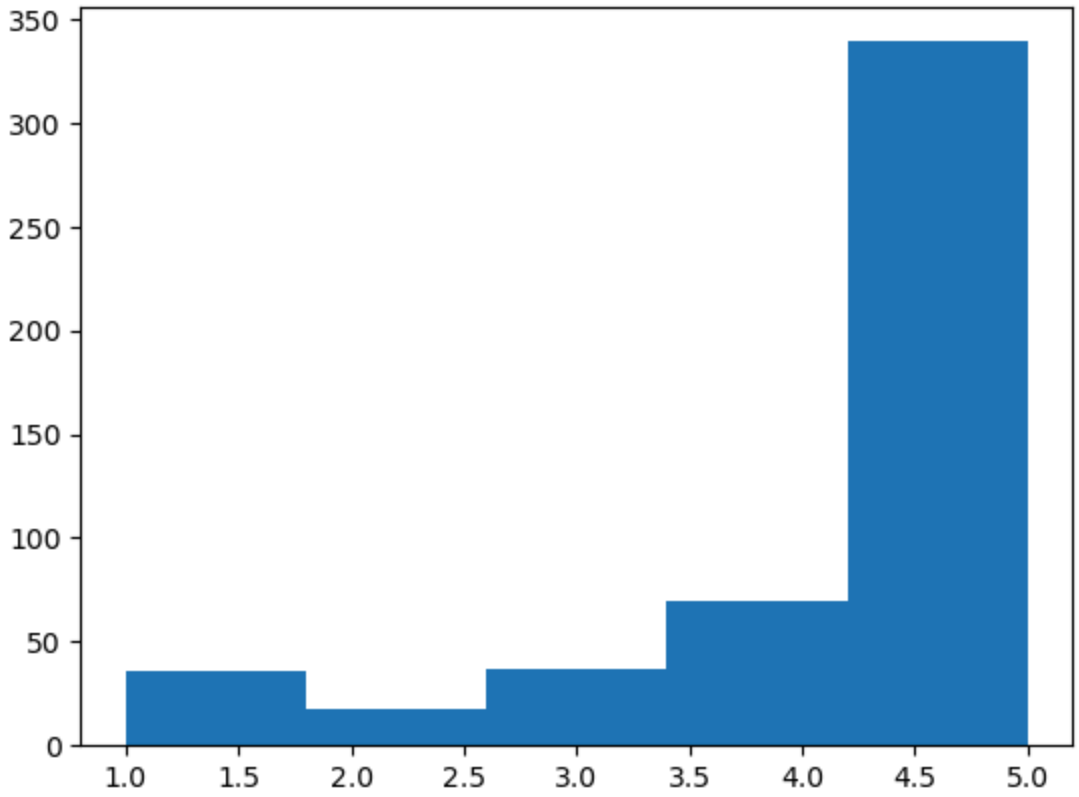
In [27]:

```python
# Import libraries
import pandas as pd
from nltk.corpus import stopwords
from textblob import TextBlob, Word

# Sample DataFrame
df = pd.DataFrame({'Text': ["This is an exmple sentence with some erors."]})
```

```
# Lower casing and removing punctuations
df['Text'] = df['Text'].apply(lambda x: " ".join(x.lower() for x in x.split()))
df['Text'] = df['Text'].str.replace(r'[^\w\s]', ' ', regex=True)

# Removal of stop words
stop = set(stopwords.words('english'))
df['Text'] = df['Text'].apply(lambda x: " ".join(x for x in x.split() if x not in s

# Spelling correction
df['Text'] = df['Text'].apply(lambda x: str(TextBlob(x).correct()))

# Lemmatization
df['Text'] = df['Text'].apply(lambda x: " ".join([Word(word).lemmatize() for word i

# Display first few rows
print(df.Text.head())
```

```
0    example sentence error
Name: Text, dtype: object
```

In [17]:
```
# Create a new data frame "reviews" to perform exploratory data analysis upon that

reviews = df



# Dropping null values

reviews.dropna(inplace=True)



# The histogram reveals this dataset is highly unbalanced towards high rating.

reviews.Score.hist(bins=5,grid=False)

plt.show()

print(reviews.groupby('Score').count().Id)
```




```
Score
1     36
2     18
3     37
4     70
5    339
Name: Id, dtype: int64
```

In [31]:
```
score_1 = reviews[reviews['Score'] == 1].sample(n=18)

score_2 = reviews[reviews['Score'] == 2].sample(n=18)

score_3 = reviews[reviews['Score'] == 3].sample(n=18)

score_4 = reviews[reviews['Score'] == 4].sample(n=18)

score_5 = reviews[reviews['Score'] == 5].sample(n=18)
```

In [36]:
```
# Here we recreate a 'balanced' dataset.

reviews_sample = pd.concat([score_1,score_2,score_3,score_4,score_5],axis=0)

reviews_sample.reset_index(drop=True,inplace=True)


# Printing count by 'Score' to check dataset is now balanced.

print(reviews_sample.groupby('Score').count().Id)
```

```
       Score
       1    18
       2    18
       3    18
       4    18
       5    18
       Name: Id, dtype: int64
```

In [ ]:

In [ ]:

In [ ]: