In [ ]:

```
import pandas as pd
import seaborn as sns
```

In [ ]:

```
df=pd.read_csv("/content/drive/MyDrive/Colab Notebooks/DSBDA/AmazonAlexa_Reviews.
```

In [ ]:

```
df
```

Out[3]:

	rating	date	variation	verified_reviews	feedback
0	5	31-Jul-18	Charcoal Fabric	Love my Echo!	1
1	5	31-Jul-18	Charcoal Fabric	Loved it!	1
2	4	31-Jul-18	Walnut Finish	Sometimes while playing a game, you can answer...	1
3	5	31-Jul-18	Charcoal Fabric	I have had a lot of fun with this thing. My 4 ...	1
4	5	31-Jul-18	Charcoal Fabric	Music	1
...	...	...	...	...	...
3145	5	30-Jul-18	Black Dot	Perfect for kids, adults and everyone in betwe...	1
3146	5	30-Jul-18	Black Dot	Listening to music, searching locations, check...	1
3147	5	30-Jul-18	Black Dot	I do love these things, i have them running my...	1
3148	5	30-Jul-18	White Dot	Only complaint I have is that the sound qualit...	1
3149	4	29-Jul-18	Black Dot	Good	1

3150 rows × 5 columns

In [ ]:

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

Out[4]:

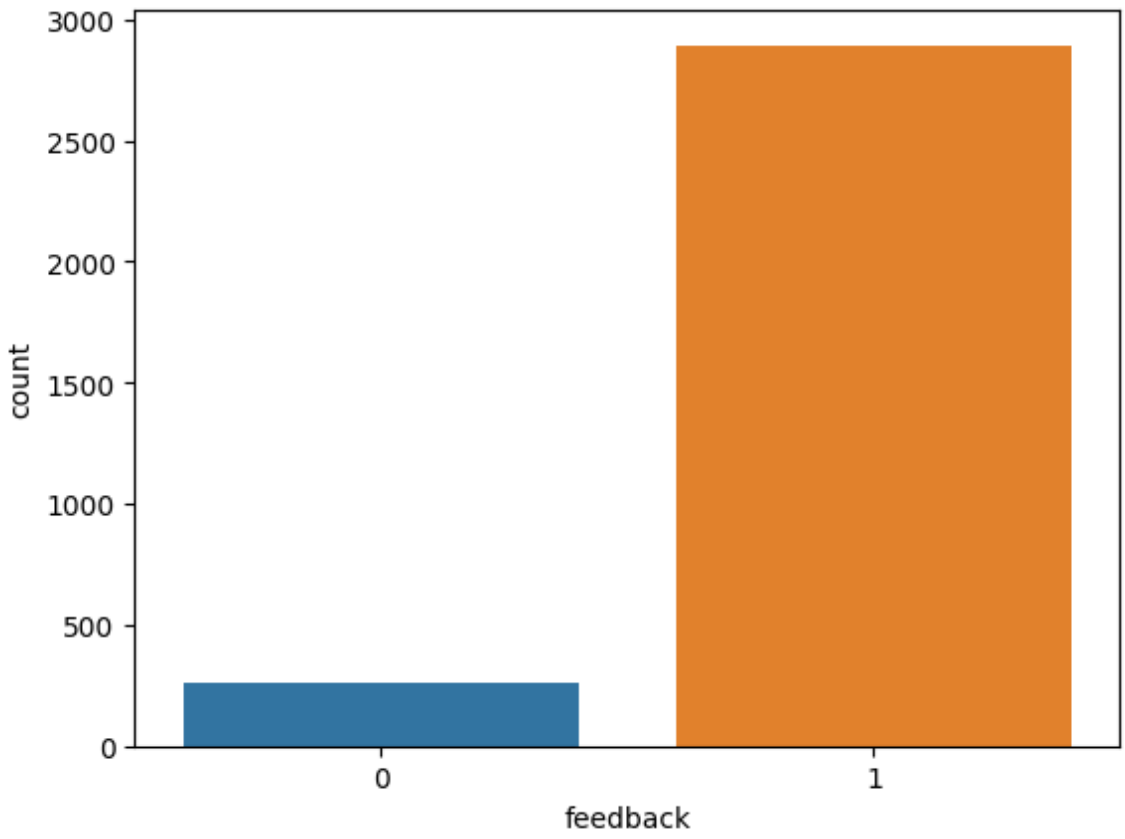
```
rating              0
date                0
variation           0
verified_reviews    0
feedback            0
dtype: int64
```

In [ ]:

```
sns.countplot(x='feedback', data=df)
```

Out[5]:

```
<Axes: xlabel='feedback', ylabel='count'>
```

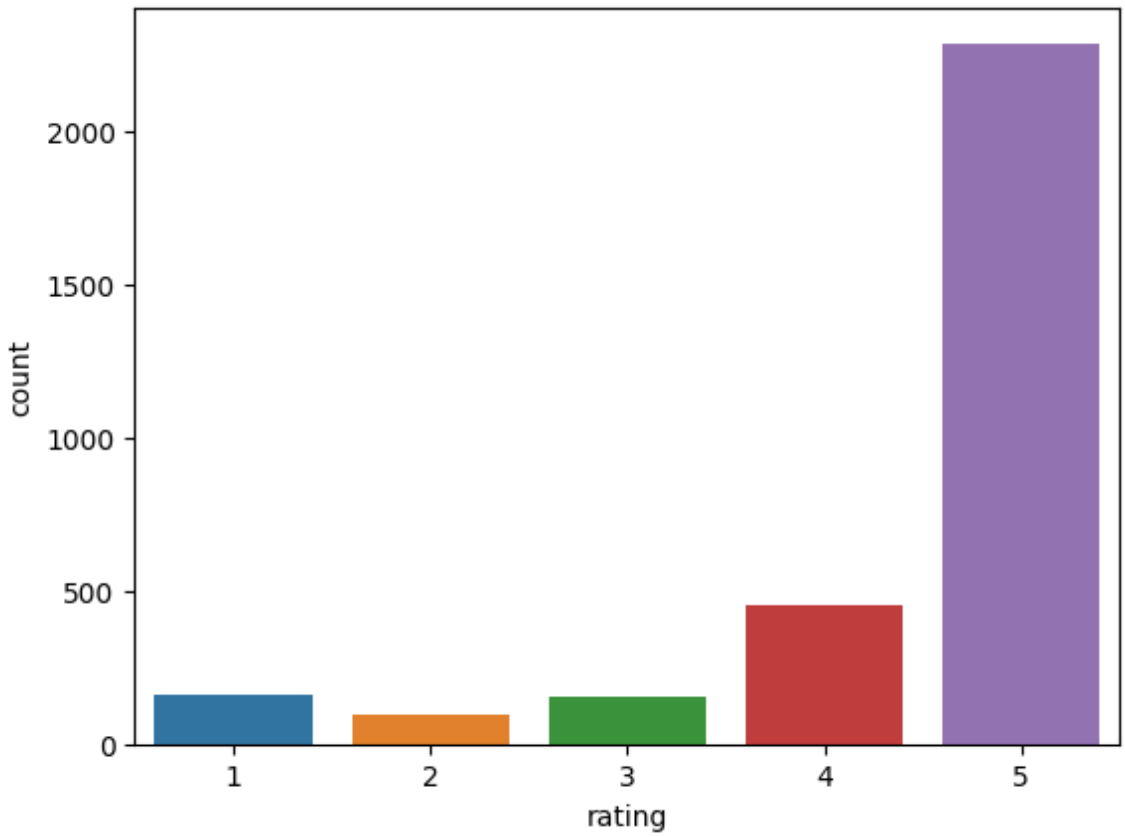


In [ ]:

```
sns.countplot(x='rating', data=df)
```

Out[6]:

```
<Axes: xlabel='rating', ylabel='count'>
```



Transform To Lowercase

In [ ]:

```
df['lower_reviews'] = df['verified_reviews'].str.lower()
```

In [ ]:

```
df[['verified_reviews','lower_reviews']].head(10)
```

Out[8]:

	verified_reviews	lower_reviews
0	Love my Echo!	love my echo!
1	Loved it!	loved it!
2	Sometimes while playing a game, you can answer...	sometimes while playing a game, you can answer...
3	I have had a lot of fun with this thing. My 4 ...	i have had a lot of fun with this thing. my 4 ...
4	Music	music
5	I received the echo as a gift. I needed anothe...	i received the echo as a gift. i needed anothe...
6	Without having a cellphone, I cannot use many ...	without having a cellphone, i cannot use many ...
7	I think this is the 5th one I've purchased. I'...	i think this is the 5th one i've purchased. i'...
8	looks great	looks great
9	Love it! I've listened to songs I haven't hear...	love it! i've listened to songs i haven't hear...

In [ ]:

```
def clean_text(text):
  for char in text:
    if char in string.punctuation:
      text=text.replace(char, '')
      text=text.strip()
  return text
```

In [ ]:

```
import string
df['reviews_pun']=df['lower_reviews'].apply(lambda x:clean_text(x))
```

In [ ]:

```
df.head(5)
```

Out[11]:

	rating	date	variation	verified_reviews	feedback	lower_reviews	reviews_pun
0	5	31-Jul-18	Charcoal Fabric	Love my Echo!	1	love my echo!	love my echo
1	5	31-Jul-18	Charcoal Fabric	Loved it!	1	loved it!	loved it
2	4	31-Jul-18	Walnut Finish	Sometimes while playing a game, you can answer...	1	sometimes while playing a game, you can answer...	sometimes while playing a game you can answer ...
3	5	31-Jul-18	Charcoal Fabric	I have had a lot of fun with this thing. My 4 ...	1	i have had a lot of fun with this thing. my 4 ...	i have had a lot of fun with this thing my 4 y...
4	5	31-Jul-18	Charcoal Fabric	Music	1	music	music

In [ ]:

```
pip install emoji
```

```
Looking in indexes: https://pypi.org/simple, (https://pypi.org/simpl
e,) https://us-python.pkg.dev/colab-wheels/public/simple/ (https://u
s-python.pkg.dev/colab-wheels/public/simple/)
Collecting emoji
  Downloading emoji-2.2.0.tar.gz (240 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 240.9/240.9 kB 5.7 MB/s
eta 0:00:00
  Preparing metadata (setup.py) ... done
Building wheels for collected packages: emoji
  Building wheel for emoji (setup.py) ... done
  Created wheel for emoji: filename=emoji-2.2.0-py3-none-any.whl siz
e=234926 sha256=31626d2f9e55e9ba898d66a252232026084b71e7eb0bc7c19869
1cc1f6011157
  Stored in directory: /root/.cache/pip/wheels/9a/b8/0f/f580817231cb
f59f6ade9fd132ff60ada1de9f7dc85521f857
Successfully built emoji
Installing collected packages: emoji
Successfully installed emoji-2.2.0
```

In [ ]:

```
import emoji
df['reviews_emoji'] = df['reviews_pun'].apply(lambda s: emoji.replace_emoji(s, ''
```

In [ ]:

```
df.head(5)
```

Out[14]:

	rating	date	variation	verified_reviews	feedback	lower_reviews	reviews_pun	reviews_en
0	5	31-Jul-18	Charcoal Fabric	Love my Echo!	1	love my echo!	love my echo	love my e
1	5	31-Jul-18	Charcoal Fabric	Loved it!	1	loved it!	loved it	love
2	4	31-Jul-18	Walnut Finish	Sometimes while playing a game, you can answer...	1	sometimes while playing a game, you can answer...	sometimes while playing a game you can answer ...	sometir while playin game you answe
3	5	31-Jul-18	Charcoal Fabric	I have had a lot of fun with this thing. My 4 ...	1	i have had a lot of fun with this thing. my 4 ...	i have had a lot of fun with this thing my 4 y...	i have had a of fun with thing my 4
4	5	31-Jul-18	Charcoal Fabric	Music	1	music	music	mu

In [ ]:

```
import nltk
nltk.download('punkt')
import nltk
nltk.download('stopwords')
import nltk
from nltk.corpus import stopwords
print(stopwords.words('english'))
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.

['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you',
"you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself',
'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her',
'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 't
heir', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'thi
s', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was',
'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do',
'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or',
'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with',
'about', 'against', 'between', 'into', 'through', 'during', 'befor
e', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'ou
t', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'onc
e', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'b
oth', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no',
'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very',
's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should'v
e", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "are
n't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'h
adn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "is
n't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "need
n't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'w
eren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]

[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

In [ ]:

```
#Remove stopwords 'n punctuation
sw = stopwords.words('english')

def transform_text(s):



    # remove punctuation
    # remove stopwords
    tokens = nltk.word_tokenize(s)

    new_string = []
    for w in tokens:
        # remove words with len = 2 AND stopwords
        if len(w) > 2 and w not in sw:
            new_string.append(w)

    s = ' '.join(new_string)
    s = s.strip()

    exclude = set(string.punctuation)
    s = ''.join(ch for ch in s if ch not in exclude)

    return s.strip()
```

In [ ]:

```
df['reviews_stop']=df['reviews_emoji'].apply(transform_text)
```

In [ ]:

```
import re
def tokenize(text):
    split=re.split("\W+",text)
    return split
df['reviews_split']=df['reviews_stop'].apply(lambda x: tokenize(x.lower()))
df.head(5)
```

Out[18]:

	rating	date	variation	verified_reviews	feedback	lower_reviews	reviews_pun	reviews_en
0	5	31-Jul-18	Charcoal Fabric	Love my Echo!	1	love my echo!	love my echo	love my e
1	5	31-Jul-18	Charcoal Fabric	Loved it!	1	loved it!	loved it	love
2	4	31-Jul-18	Walnut Finish	Sometimes while playing a game, you can answer...	1	sometimes while playing a game, you can answer...	sometimes while playing a game you can answer ...	sometir while playin game you answe
3	5	31-Jul-18	Charcoal Fabric	I have had a lot of fun with this thing. My 4 ...	1	i have had a lot of fun with this thing. my 4 ...	i have had a lot of fun with this thing my 4 y...	i have had a of fun with thing my 4
4	5	31-Jul-18	Charcoal Fabric	Music	1	music	music	mu

In [ ]:

```
df['reviews_stop']=str(df['reviews_stop'])
```

In [ ]:

```
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
```

Out[30]:

```
True
```

In [ ]:

```
import os
import re
from string import punctuation
from textblob import Word
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud
import nltk
from nltk import word_tokenize
from nltk.corpus import stopwords
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn.tree import DecisionTreeClassifier
from sklearn.feature_extraction.text import CountVectorizer
```

In [ ]:

```
def clean(df):

    #Lemmatization
    df['reviews_lem'] = df['reviews_split'].apply(lambda words: " ".join([Word(x)
    return df
```

In [ ]:

```
df = clean(df)
df.head()
```

Out[38]:

	rating	date	variation	verified_reviews	feedback	lower_reviews	reviews_pun	reviews_en
0	5	31-Jul-18	Charcoal Fabric	Love my Echo!	1	love my echo!	love my echo	love my e
1	5	31-Jul-18	Charcoal Fabric	Loved it!	1	loved it!	loved it	love
2	4	31-Jul-18	Walnut Finish	Sometimes while playing a game, you can answer...	1	sometimes while playing a game, you can answer...	sometimes while playing a game you can answer ...	sometir while playin game you answe
3	5	31-Jul-18	Charcoal Fabric	I have had a lot of fun with this thing. My 4 ...	1	i have had a lot of fun with this thing. my 4 ...	i have had a lot of fun with this thing my 4 y...	i have had a of fun with thing my 4
4	5	31-Jul-18	Charcoal Fabric	Music	1	music	music	mu

In [ ]:

```
from nltk.stem.porter import PorterStemmer

stemmer = PorterStemmer()
def stem_words(text):
    return " ".join([stemmer.stem(word) for word in text.split()])

df["reviews_stem"] = df["verified_reviews"].apply(lambda text: stem_words(text))
df["reviews_stem"]
```

Out[40]:

```
0                                          love my echo!
1                                               love it!
2       sometim while play a game, you can answer a qu...
3       i have had a lot of fun with thi thing. my 4 y...
4                                                 music
                              ...
3145     perfect for kids, adult and everyon in between!!
3146    listen to music, search locations, check time,...
3147    i do love these things, i have them run my ent...
3148    onli complaint i have is that the sound qualit...
3149                                               good
Name: reviews_stem, Length: 3150, dtype: object
```

In [ ]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [ ]:

```
#tf idf
tf_idf = TfidfVectorizer()
#applying tf idf to data
data_tf = tf_idf.fit_transform(df['reviews_split'])
```

In [ ]:

```
print("n_samples: %d, n_features: %d" % data_tf.shape)
```

```
n_samples: 3150, n_features: 3903
```

In [ ]:

```
```