Assignment 3 Report:

1.

The review data looks like this:

	overall	vote	verified	reviewTime	reviewerID	asin	style	reviewerName	reviewText	summary	unixReviewTime	image
0	5.0	67	True	09 18, 1999	AAP7PPBU72QFM	0151004714	{'Format:': ' Hardcover'}	D. C. Carrad	This is the best novel I have read in 2 or 3 y...	A star is born	937612800	NaN
1	3.0	5	True	10 23, 2013	A2E168DTVGE6SV	0151004714	{'Format:': ' Kindle Edition'}	Evy	Pages and pages of introspection, in the style...	A stream of consciousness novel	1382486400	NaN
2	5.0	4	False	09 2, 2008	A1ER5AYS3FQ9O3	0151004714	{'Format:': ' Paperback'}	Kcorn	This is the kind of novel to read when you hav...	I'm a huge fan of the author and this one did ...	1220313600	NaN
3	5.0	13	False	09 4, 2000	A1T17LMQABMBN5	0151004714	{'Format:': ' Hardcover'}	Caf Girl Writes	What gorgeous language! What an incredible wri...	The most beautiful book I have ever read!	968025600	NaN
4	3.0	8	True	02 4, 2000	A3QHJ0FXK33OBE	0151004714	{'Format:': ' Hardcover'}	W. Shane Schmidt	I was taken in by reviews that compared this b...	A dissenting view--In part.	949622400	NaN

A sample from metadata:

	category	tech1	description	fit	title	also_buy	tech2	brand	feature	rank	also_view	main_cat	similar_it
0	['Electronics', 'Camera & Photo', 'Video S...	NaN	['The following camera brands and models have ...	NaN	Genuine Geovision 1 Channel 3rd Party NVR IP S...	[]	NaN	GeoVision	['Genuine Geovision 1 Channel NVR IP Software'...	['>#3,092 in Tools & Home Improvement >...	[]	Camera & Photo	N
1	['Electronics', 'Camera & Photo']	NaN	['This second edition of the Handbook of Astro...	NaN	Books "Handbook of Astronomical Image Processi...	['0999470906']	NaN	33 Books Co.	['Detailed chapters cover these fundamental to...	['>#55,933 in Camera & Photo (See Top 100 ...	['0943396670', '1138055360', '0999470906']	Camera & Photo	N
2	['Electronics', 'eBook Readers & Accessori...	NaN	['A zesty tale. (Publishers Weekly) <br /...	NaN	One Hot Summer	['0425167798', '039914157X']	NaN	Visit Amazon's Carolina Garcia Aguilera Page	[]	3,105,177 in Books (	[]	Books	N
3	['Electronics', 'eBook Readers & Accessories',...	NaN	[]	NaN	Hurray for Hattie Rabbit: Story and pictures (...	['0060219521', '0060219580', '0060219394']	NaN	Visit Amazon's Dick Gackenbach Page	[]	2,024,298 in Books (	['0060219521', '0060219475', '0060219394']	Books	N
4	['Electronics', 'eBook Readers & Accessories',...	NaN	['“sex.lies.murder.fame. is brilllllli&#8...	NaN	sex.lies.murder.fame.: A Novel	[]	NaN	Visit Amazon's Lolita Files Page	[]	3,778,828 in Books (	[]	Books	N

Merged data based on asin keycolumn:

	overall	vote	verified	reviewTime	reviewerID	asin	style	reviewerName	reviewText	summary	unixReviewTime	image	title
0	5.0	67	True	09 18, 1999	AAP7PPBU72QFM	0151004714	{'Format:': ' Hardcover'}	D. C. Carrad	This is the best novel I have read in 2 or 3 y...	A star is born	937612800	NaN	The Last Life: A Novel
1	3.0	5	True	10 23, 2013	A2E168DTVGE6SV	0151004714	{'Format:': ' Kindle Edition'}	Evy	Pages and pages of introspection, in the style...	A stream of consciousness novel	1382486400	NaN	The Last Life: A Novel
2	5.0	4	False	09 2, 2008	A1ER5AYS3FQ9O3	0151004714	{'Format:': ' Paperback'}	Kcorn	This is the kind of novel to read when you hav...	I'm a huge fan of the author and this one did ...	1220313600	NaN	The Last Life: A Novel
3	5.0	13	False	09 4, 2000	A1T17LMQABMBN5	0151004714	{'Format:': ' Hardcover'}	Caf Girl Writes	What gorgeous language! What an incredible wri...	The most beautiful book I have ever read!	968025600	NaN	The Last Life: A Novel
4	3.0	8	True	02 4, 2000	A3QHJ0FXK33OBE	0151004714	{'Format:': ' Hardcover'}	W. Shane Schmidt	I was taken in by reviews that compared this b...	A dissenting view--In part.	949622400	NaN	The Last Life: A Novel

2. Product I choose: USB

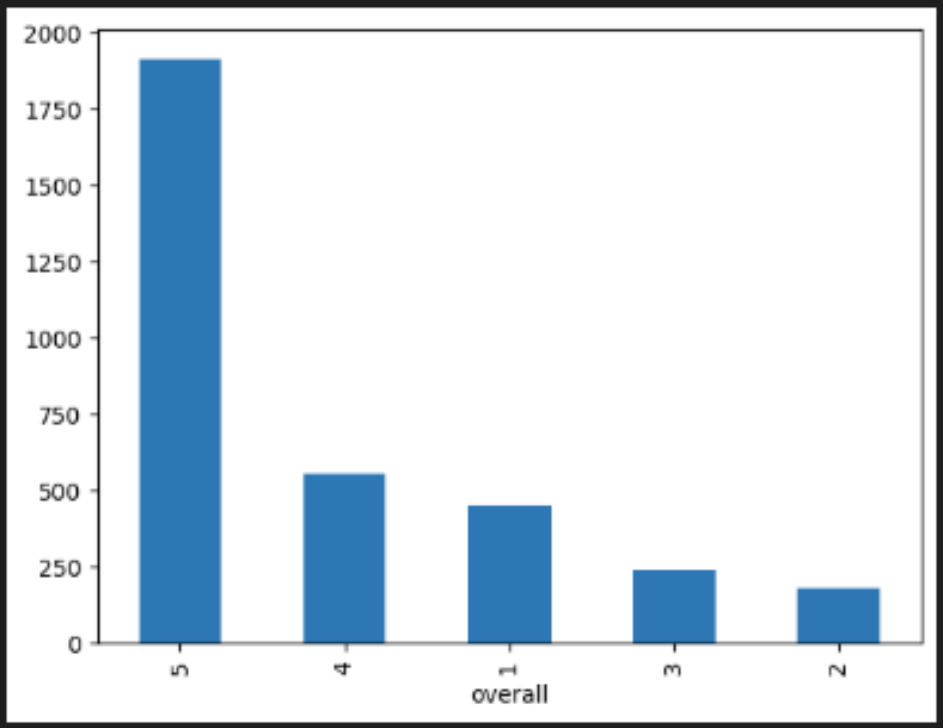
3. Total selected rows =3333
   Dropped NaN values and duplicated reviewIDs

Unnamed: 0	overall	vote	verified	reviewTime	reviewerID	asin	style	reviewerName	reviewText	summary	unixReviewTime	image	
10074	4	16.0	True	2016-05-13	A2665PMQ6QIEA7	B00000J1U5	{'Style:': ' DB15'}	erple2	i buy this cable to power an old microsoft sid...	After slight "hack", works brilliantly with Mi...	1463097600	['https://images-na.ssl-images-amazon.com/imag...	Bel Inc Jc Ac Side
10322	1	5.0	True	2015-10-30	A26UFBBEXEC4I7	B00000J1U5	{'Style:': ' DB15'}	Richfiddler11	i buy this thinking it have a circuit that wou...	Requires modding your MS Sidewinder Precision ...	1446163200	['https://images-na.ssl-images-amazon.com/imag...	Bel Inc Jc Ac Side
10754	5	6.0	True	2015-02-05	A1OU2FW26L47VV	B00000J1U5	{'Style:': ' VideoLink Powerline Internet'}	Amazon Customer	i receive these today and update the firmware ...	Well worth it...	1423094400	['https://images-na.ssl-images-amazon.com/imag...	Bel Inc Jc Ac Side
177122	5	2	True	2016-01-08	ALDM6DH1HFZE3	B00005108J	{'Color:': ' Red'}	Ginger Did It	do what the description say enjoy they immense...	Cool backlighting	1452211200	['https://images-na.ssl-images-amazon.com/imag...	Ad A Lig Red po
177158	5	2	True	2016-01-05	A3CLO0W8933D9N	B00005108J	{'Color:': ' Red'}	Melvin	these light be goooooood waaaay well than expe...	These lights are goooooood waaaay better than ...	1451952000	['https://images-na.ssl-images-amazon.com/imag...	Ad A Lig Red po

4.

```
No of reviews:  3333
Average rating:  3.990999099909991
No of unique products:  1426
No of  good ratings:  2706
No of  bad ratings:  627
```

```
<Axes: xlabel='overall'>
```

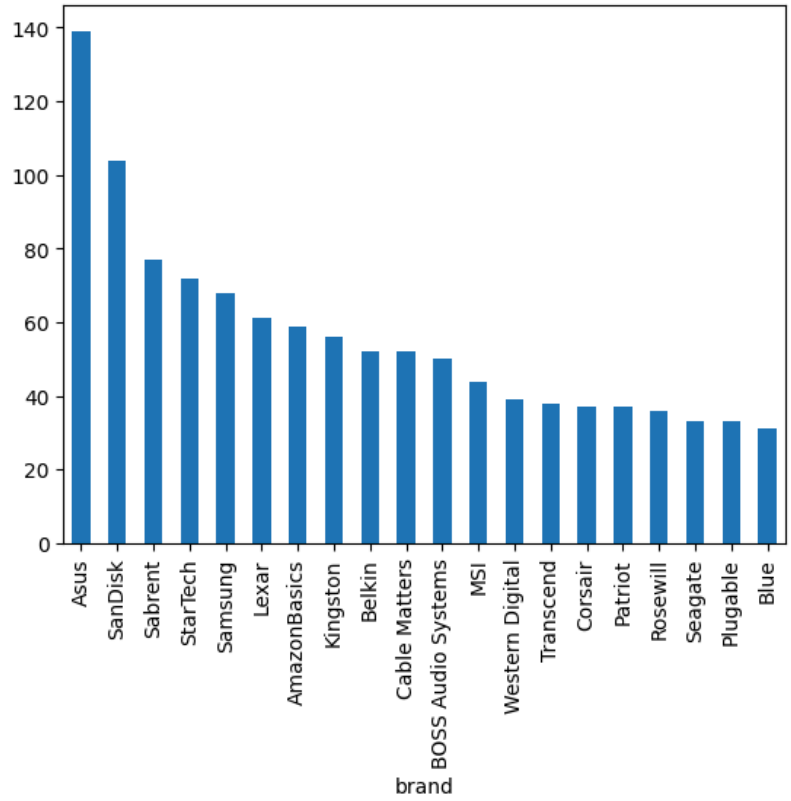


5. Text preprocessing

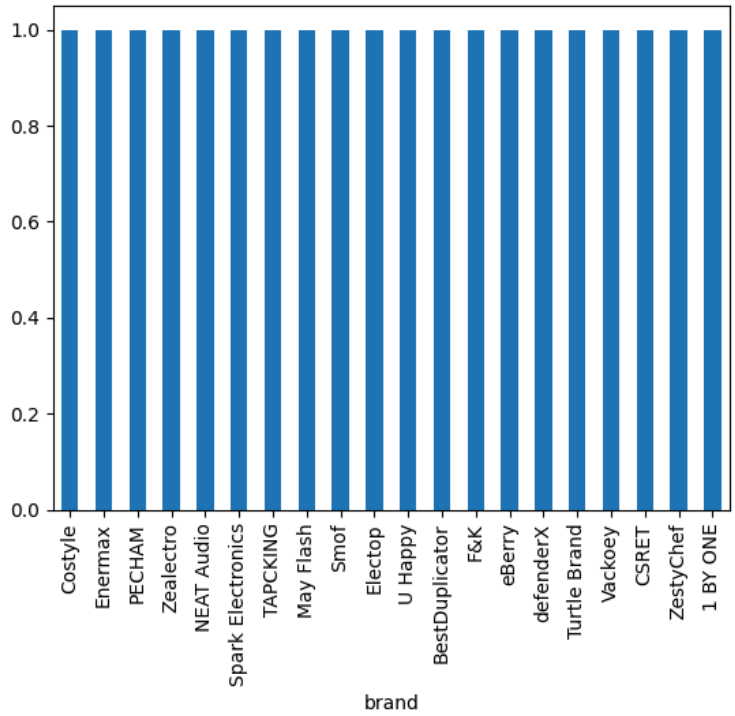
```
nltk.download('punkt')
nltk.download('wordnet')
lemmatizer = WordNetLemmatizer()


def lemmatize_text(text):
    tokens = word_tokenize(text)
    lemmatized_tokens = [lemmatizer.lemmatize(token) for token in tokens]
    return ' '.join(lemmatized_tokens)



lemmatizer = WordNetLemmatizer()


def remove_html_tags(text):
    return BeautifulSoup(text, 'html.parser').get_text()

def remove_accented_chars(text):
    return unicodedata.normalize('NFKD', text).encode('ascii', 'ignore').decode('utf-8', 'ignore')

def expand_acronyms(text):
    for acronym, full in acronyms:
        acronym=acronym.lower()
        text = text.lower().replace(acronym, full)
    return text


def expand_contractions(text):
    return contractions.fix(text)

def remove_special_characters(text):
    return re.sub(r'[^a-zA-Z0-9\s]', '', text)

# def lemmatize_text(text):
#     return ' '.join([lemmatizer.lemmatize(word) for word in word_tokenize(text)])

def normalize_text(text):
    text=text.lower()
    return unicodedata.normalize('NFKD', text).encode('ascii', 'ignore').decode('utf-8', 'ignore')

def preprocess_text(text):
    text = remove_html_tags(text)
    text = remove_accented_chars(text)
    text = expand_acronyms(text)
    text = expand_contractions(text)
    text = remove_special_characters(text)
    text = lemmatize_text(text)
    text = normalize_text(text)
    return text
```

6.

20 most reviewed brand:



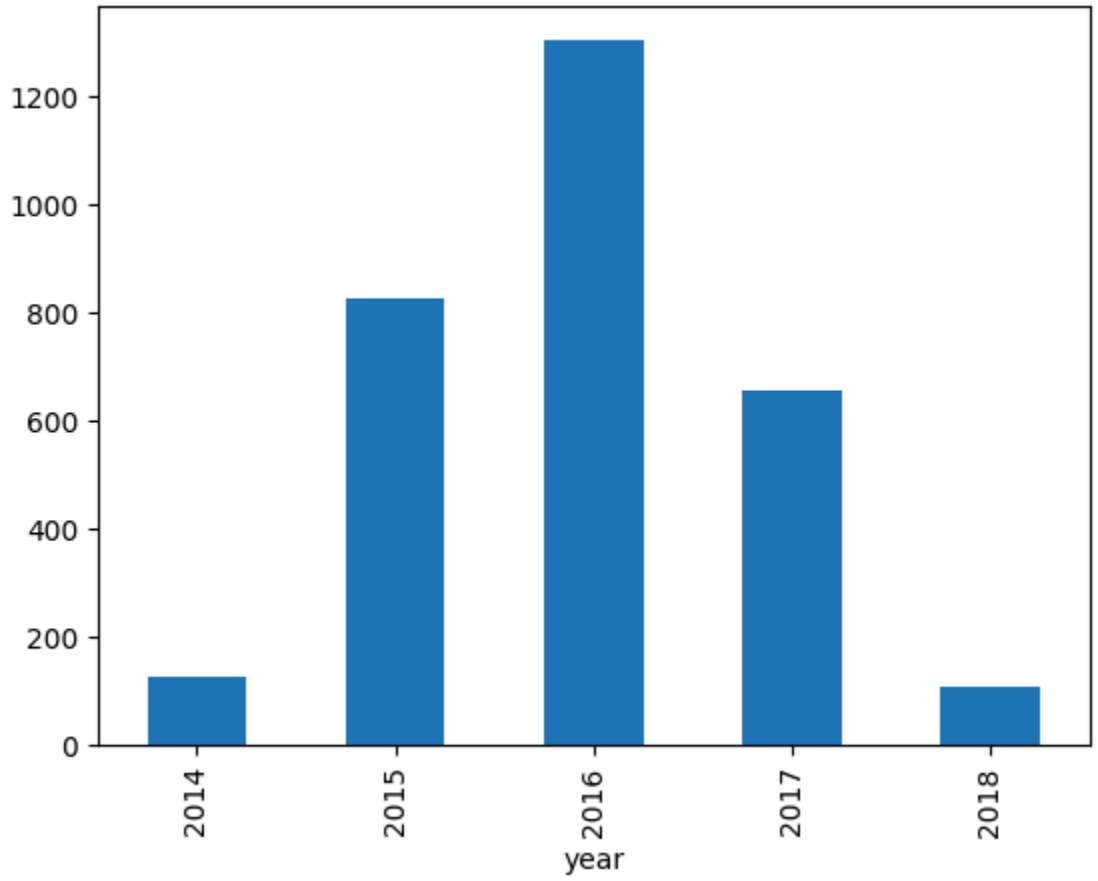
20 least reviewed brands:



c) highest rated usb product:

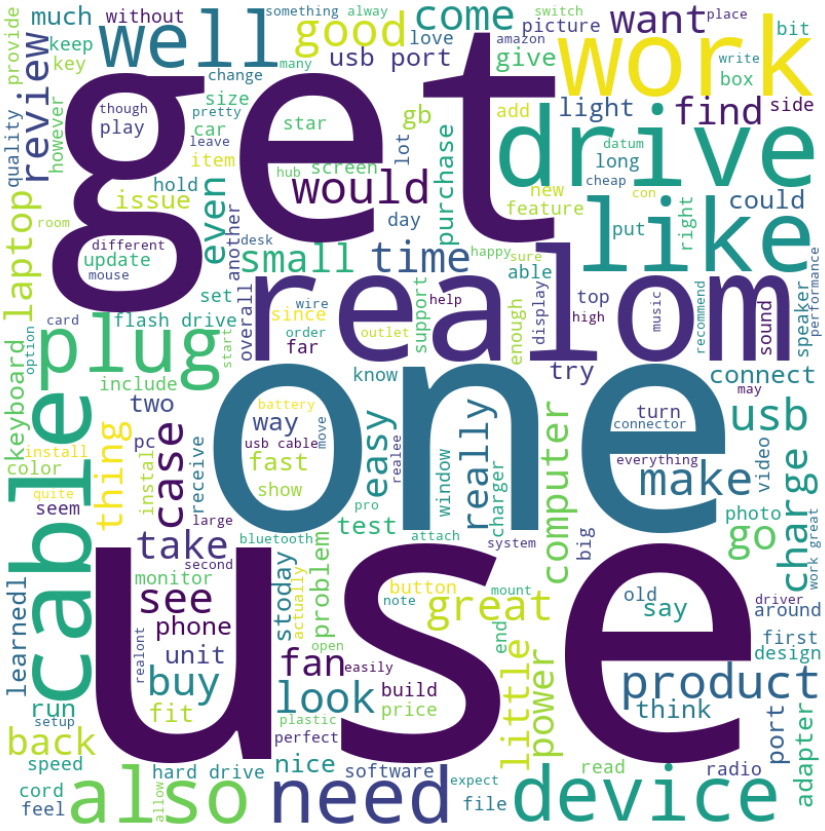
```
asin
B01HDUVMBU    5.0
Name: overall, dtype: float64
3277    Cooler Master MasterBox 5 White with Dark Mirr...
Name: title, dtype: object
```

d).

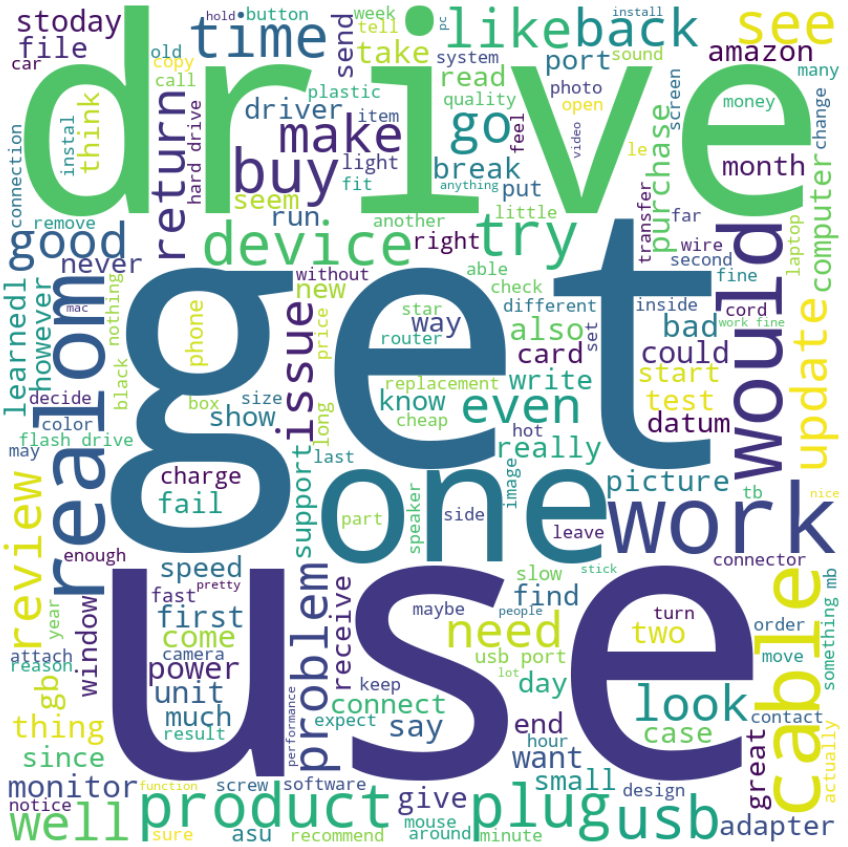


Review counts of past 5 years

e)

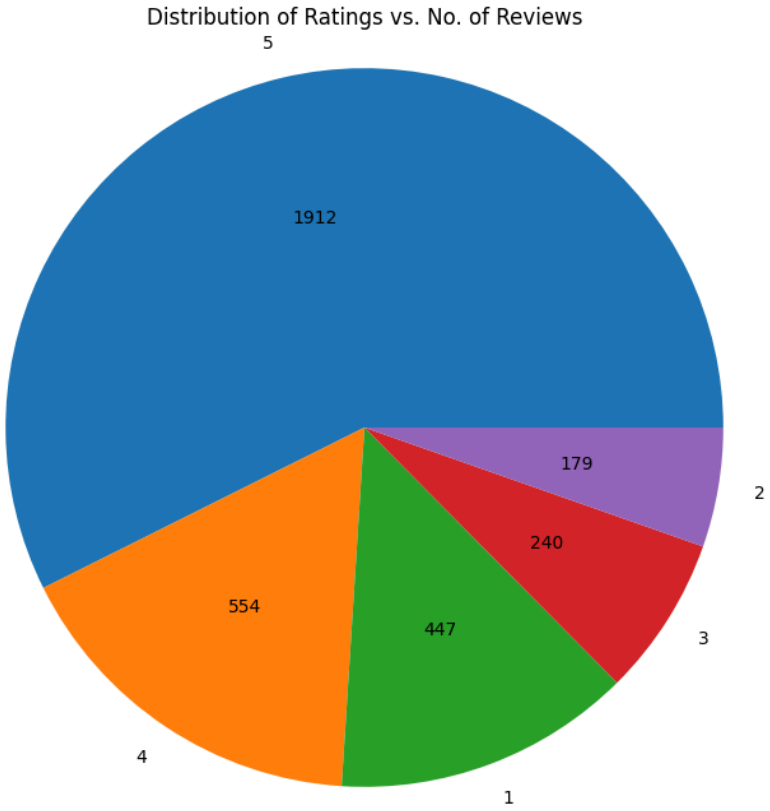


Good rated word cloud



Bad rated word cloud

f)



Pie chart of distribution of rationgs vs no of reviews

g)

The usb_df got maximum reviews in the year : 2016

h)

The year with the highest number of customers is: 2016

7.

```
TF-IDF representation:
[[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]
Vocabulary:
['175mbps', 'u1', 'lost', '975', 'wifiwifi', 'umthis', 'ff008001', 'fabulous', 'crackel', 'visiotek', 'driverpack', 'counterweight', 'name', 'excessivel
```

TF-IDF and Vectorizer

8 and 9).

```
usb_df["rating_class"] = usb_df["overall"].apply(lambda x: "Good" if x > 3 else "Average" if x == 3 else "Bad")
```

0.0s

```
# 9. From the dataset, take the Review Text as input feature and Rating Class as target
# variable. Divide the data into Train and Test Data in the ratio of 75:25.
from sklearn.model_selection import train_test_split

X = usb_df["reviewText"]
y = usb_df["rating_class"]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=1)
```

0.0s

10.

NAIVE BAYES

```
Accuracy: 0.7649880095923262
[[  0   0  54]
 [  0  19 139]
 [  0   3 619]]
              precision    recall  f1-score   support

     Average       0.00      0.00      0.00        54
         Bad       0.86      0.12      0.21       158
        Good       0.76      1.00      0.86       622

    accuracy                           0.76       834
   macro avg       0.54      0.37      0.36       834
weighted avg       0.73      0.76      0.68       834
```

LOGISTIC REGRESSION

```
Accuracy: 0.8105515587529976
0.7666345564209157
[[  0   8  46]
 [  0  66  92]
 [  0  12 610]]
              precision    recall  f1-score   support

     Average       0.00      0.00      0.00        54
         Bad       0.77      0.42      0.54       158
        Good       0.82      0.98      0.89       622

    accuracy                           0.81       834
   macro avg       0.53      0.47      0.48       834
weighted avg       0.75      0.81      0.77       834
```

SUPPORT VECTOR MACHINE

```
Accuracy: 0.8009592326139089
[[  0   3  51]
 [  0  49 109]
 [  0   3 619]]
              precision    recall  f1-score   support

     Average       0.00      0.00      0.00        54
         Bad       0.89      0.31      0.46       158
        Good       0.79      1.00      0.88       622

    accuracy                           0.80       834
   macro avg       0.56      0.44      0.45       834
weighted avg       0.76      0.80      0.75       834
```

RANDOM FOREST CLASSIFIER

```
Accuracy: 0.7973621103117506
[[  0   4  50]
 [  0  54 104]
 [  0  11 611]]
              precision    recall  f1-score   support

     Average       0.00      0.00      0.00        54
         Bad       0.78      0.34      0.48       158
        Good       0.80      0.98      0.88       622

    accuracy                           0.80       834
   macro avg       0.53      0.44      0.45       834
weighted avg       0.74      0.80      0.75       834
```

GBD For text classification

```
Accuracy: 0.7985611510791367
F1 Score: 0.7529966485964479
Confusion Matrix:
[[  0   6  48]
 [  2  58  98]
 [  2  12 608]]
Classification Report:
              precision    recall  f1-score   support

     Average       0.00      0.00      0.00        54
         Bad       0.76      0.37      0.50       158
        Good       0.81      0.98      0.88       622

    accuracy                           0.80       834
   macro avg       0.52      0.45      0.46       834
weighted avg       0.75      0.80      0.75       834
```

11.
Q11 is in the ipynb file:

12.