

```
In [59]: # Basic libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import re
from collections import Counter
from wordcloud import WordCloud
from sklearn.feature_extraction.text import ENGLISH_STOP_WORDS

# Configure plot style
sns.set(style="whitegrid", palette="muted", font_scale=1.1)
%matplotlib inline
```

```
In [60]: # Load Amazon Fine Food Reviews dataset
df = pd.read_csv("Reviews.csv")

# Select relevant columns
df = df[["Text", "Score", "Time", "Summary", "ProductId", "UserId"]].dropna()

# Show first 5 rows
df.head()
```

```
Out[60]:
```

	Text	Score	Time	Summary	ProductId	UserId
0	I have bought several of the Vitality canned d...	5	1303862400	Good Quality Dog Food	B001E4KFG0	A3SGXH7AUHU8GW
1	Product arrived labeled as Jumbo Salted Peanut...	1	1346976000	Not as Advertised	B00813GRG4	A1D87F6ZCVE5NK
2	This is a confection that has been around a fe...	4	1219017600	"Delight" says it all	B000LQOCH0	ABXLMWJIXXAIN
3	If you are looking for the secret ingredient i...	2	1307923200	Cough Medicine	B000UA0QIQ	A395BORC6FGVXV
4	Great taffy at a great price. There was a wid...	5	1350777600	Great taffy	B006K2ZZ7K	A1UQRSCLF8GW1T

```
In [61]: # Dataset info
df.info()

# Basic statistics
df.describe()

# Check for missing values
df.isnull().sum()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 568427 entries, 0 to 568453
Data columns (total 6 columns):
#   Column      Non-Null Count  Dtype
---  -
0   Text         568427 non-null object
1   Score        568427 non-null int64
2   Time         568427 non-null int64
3   Summary      568427 non-null object
4   ProductId    568427 non-null object
5   UserId       568427 non-null object
dtypes: int64(2), object(4)
memory usage: 30.4+ MB
```

```
Out[61]: Text         0
Score         0
Time          0
Summary       0
ProductId     0
UserId        0
dtype: int64
```

```
In [62]: # Clean text: lowercase, remove special characters, numbers, extra spaces
def clean_text(text):
    text = str(text).lower()
    text = re.sub(r"^[^a-z\s]", " ", text) # remove non-alpha
    text = re.sub(r"\s+", " ", text).strip()
    return text

df["clean_text"] = df["Text"].apply(clean_text)

# Review length (number of words)
df["review_length"] = df["clean_text"].apply(lambda x: len(x.split()))

df.head()
```

Out [62]:

		Text	Score	Time	Summary	ProductId	UserId	clean_text	review_length
0		I have bought several of the Vitality canned d...	5	1303862400	Good Quality Dog Food	B001E4KFG0 A3SGXH7AUHU8GW		i have bought several of the vitality canned d...	48
1		Product arrived labeled as Jumbo Salted Peanut...	1	1346976000	Not as Advertised	B00813GRG4 A1D87F6ZCVE5NK		product arrived labeled as jumbo salted peanut...	32
2		This is a confection that has been around a fe...	4	1219017600	"Delight" says it all	B000LQOCH0 ABXLMWJIXXAIN		this is a confection that has been around a fe...	93
3		If you are looking for the secret ingredient i...	2	1307923200	Cough Medicine	B000UA0QIQ A395BORC6FGVXV		if you are looking for the secret ingredient i...	41
4		Great taffy at a great price. There was a wid...	5	1350777600	Great taffy	B006K2ZZ7K A1UQRSCLF8GW1T		great taffy at a great price there was a wide ...	27

In [63]:

```
# Map score to sentiment
def score_to_sentiment(score):
    if score <= 2:
        return "Negative"
    elif score == 3:
        return "Neutral"
    else:
        return "Positive"

df["Sentiment"] = df["Score"].apply(score_to_sentiment)

# Count of each sentiment
df["Sentiment"].value_counts()
```

Out [63]:

Sentiment
Positive 443777
Negative 82012
Neutral 42638
Name: count, dtype: int64

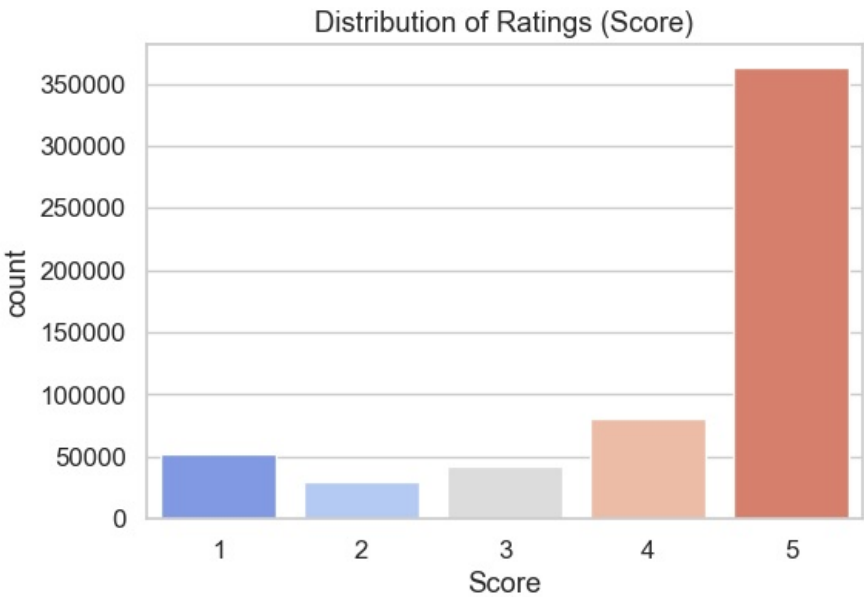
In [64]:

```
plt.figure(figsize=(6,4))
sns.countplot(x="Score", data=df, palette="coolwarm")
plt.title("Distribution of Ratings (Score)")
plt.show()
```

C:\Users\Vedant\AppData\Local\Temp\ipykernel\_2548\1139816463.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

sns.countplot(x="Score", data=df, palette="coolwarm")



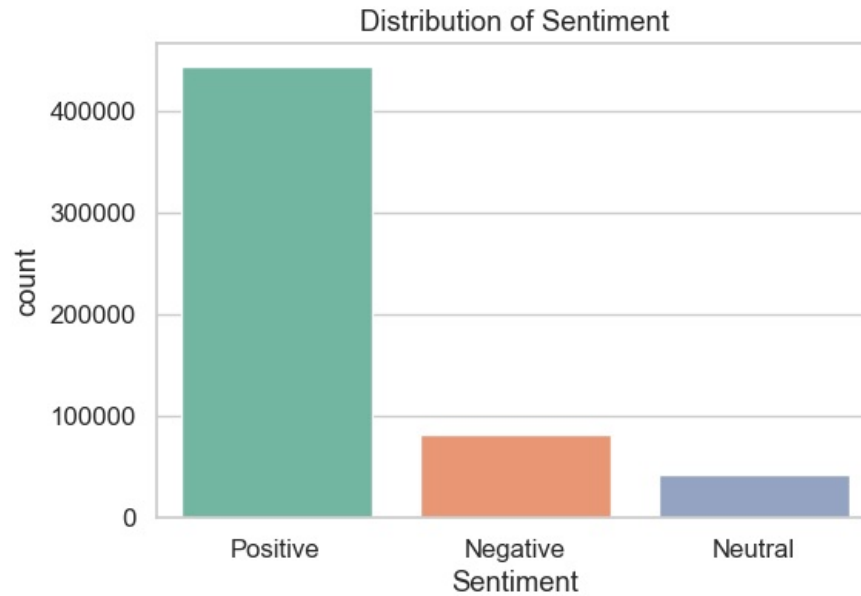
In [65]:

```
plt.figure(figsize=(6,4))
sns.countplot(x="Sentiment", data=df, palette="Set2")
plt.title("Distribution of Sentiment")
plt.show()
```

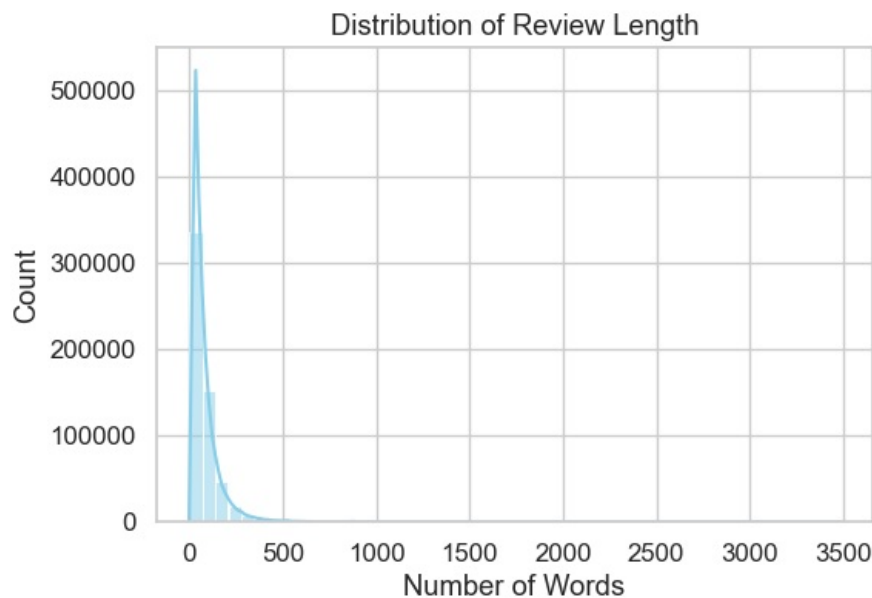
C:\Users\Vedant\AppData\Local\Temp\ipykernel\_2548\952997905.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

```
sns.countplot(x="Sentiment", data=df, palette="Set2")
```



```
In [66]: plt.figure(figsize=(6,4))
sns.histplot(df["review_length"], bins=50, kde=True, color="skyblue")
plt.title("Distribution of Review Length")
plt.xlabel("Number of Words")
plt.show()
```

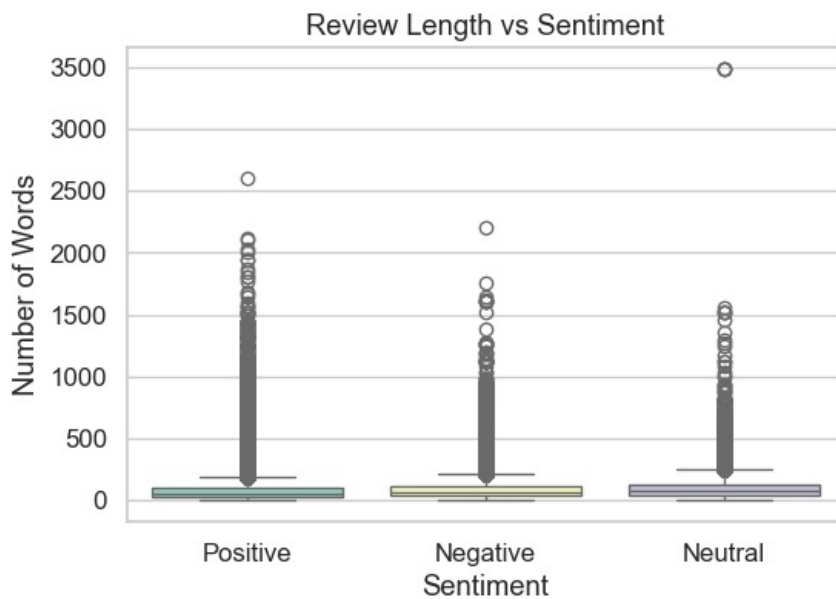


```
In [67]: plt.figure(figsize=(6,4))
sns.boxplot(x="Sentiment", y="review_length", data=df, palette="Set3")
plt.title("Review Length vs Sentiment")
plt.ylabel("Number of Words")
plt.show()
```

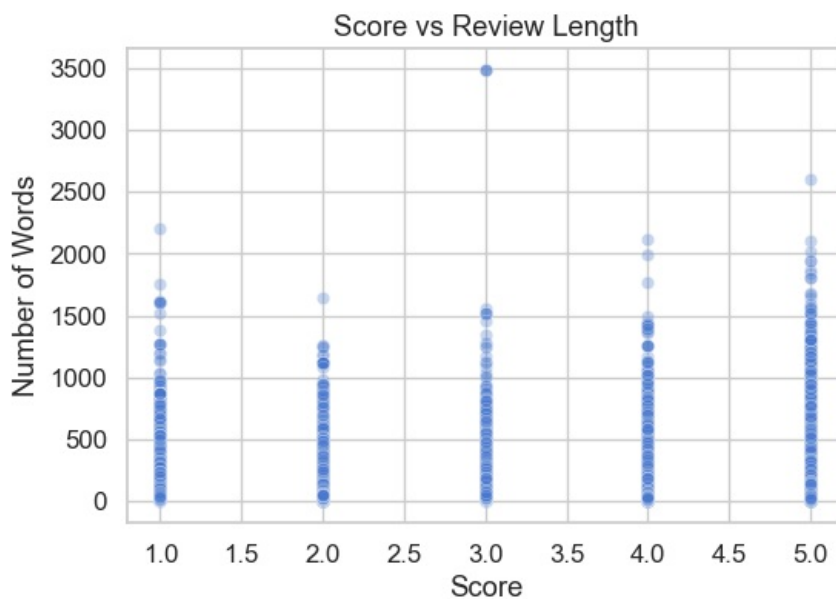
C:\Users\Vedant\AppData\Local\Temp\ipykernel\_2548\2116653056.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

```
sns.boxplot(x="Sentiment", y="review_length", data=df, palette="Set3")
```



```
In [68]: plt.figure(figsize=(6,4))
sns.scatterplot(x="Score", y="review_length", data=df, alpha=0.3)
plt.title("Score vs Review Length")
plt.ylabel("Number of Words")
plt.show()
```



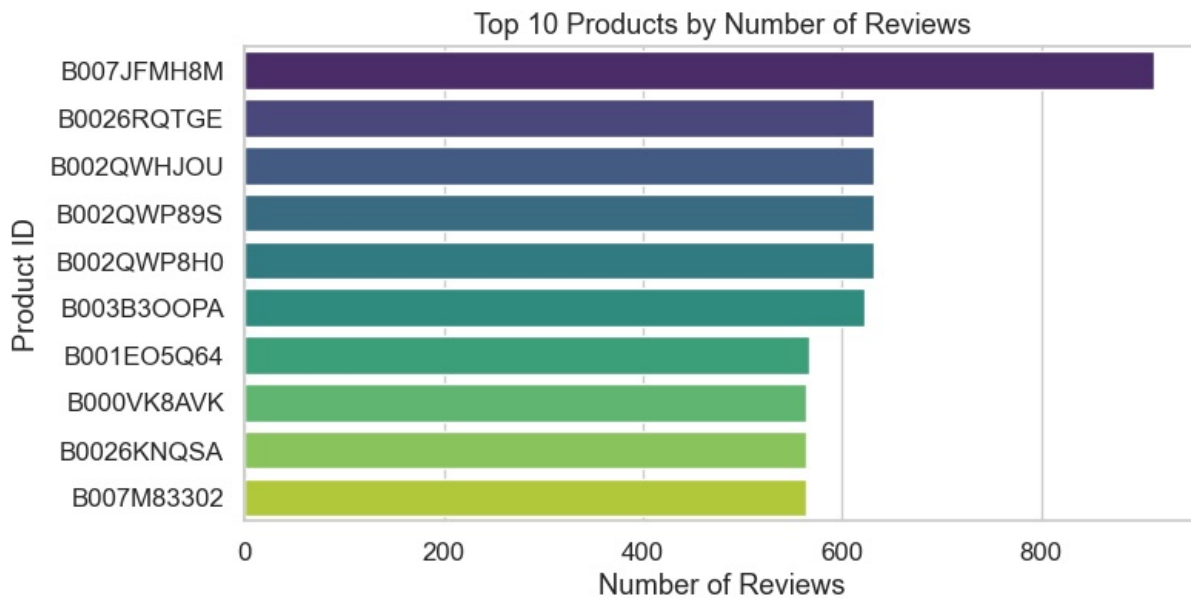
```
In [69]: top_products = df["ProductId"].value_counts().head(10)

plt.figure(figsize=(8,4))
sns.barplot(x=top_products.values, y=top_products.index, palette="viridis")
plt.xlabel("Number of Reviews")
plt.ylabel("Product ID")
plt.title("Top 10 Products by Number of Reviews")
plt.show()
```

C:\Users\Vedant\AppData\Local\Temp\ipykernel\_2548\3143326071.py:4: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same effect.

```
sns.barplot(x=top_products.values, y=top_products.index, palette="viridis")
```



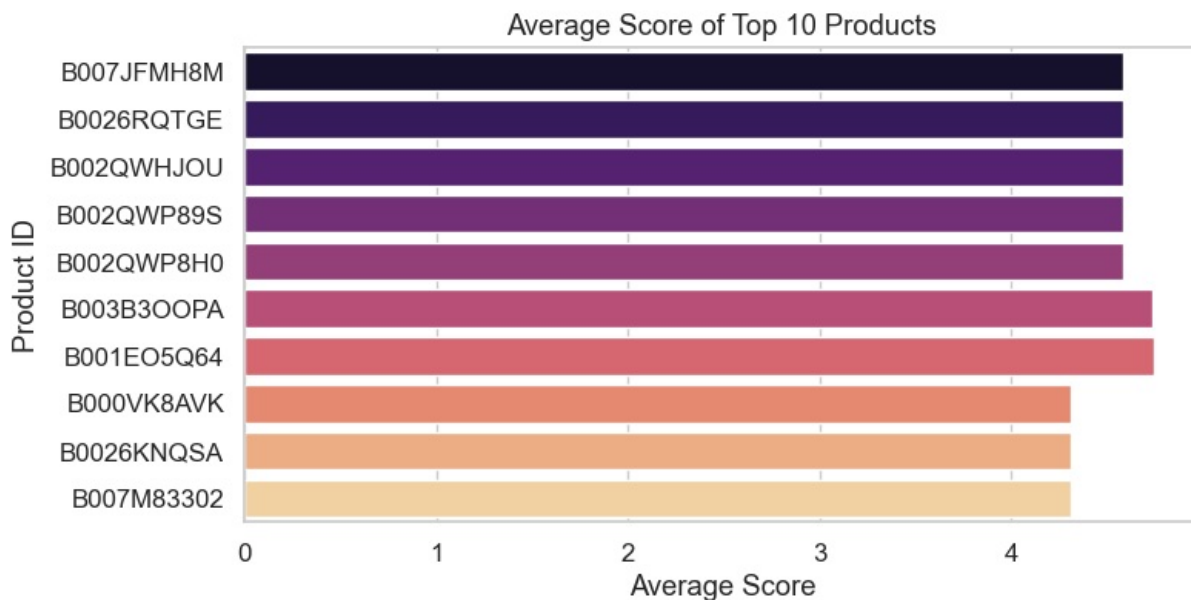
```
In [70]: avg_score_products = df.groupby("ProductId")["Score"].mean().loc[top_products.index]

plt.figure(figsize=(8,4))
sns.barplot(x=avg_score_products.values, y=avg_score_products.index, palette="magma")
plt.xlabel("Average Score")
plt.ylabel("Product ID")
plt.title("Average Score of Top 10 Products")
plt.show()
```

C:\Users\Vedant\AppData\Local\Temp\ipykernel\_2548\2800932052.py:4: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same effect.

```
sns.barplot(x=avg_score_products.values, y=avg_score_products.index, palette="magma")
```



```
In [71]: # Clean text: lowercase, remove HTML tags, special characters, numbers, extra spaces
def clean_text(text):
    text = str(text).lower()
    # remove HTML tags like <br />
    text = re.sub(r"<.*?>", " ", text)
    # remove special chars and numbers
    text = re.sub(r"[^a-z\s]", " ", text)
    # remove extra spaces
    text = re.sub(r"\s+", " ", text).strip()
    return text

df["clean_text"] = df["Text"].apply(clean_text)

# Review length (number of words)
df["review_length"] = df["clean_text"].apply(lambda x: len(x.split()))
```

```
df.head()
```

Out[71]:

	Text	Score	Time	Summary	ProductId	UserId	clean_text	review_length	Sentiment
0	I have bought several of the Vitality canned d...	5	1303862400	Good Quality Dog Food	B001E4KFG0	A3SGXH7AUHU8GW	i have bought several of the vitality canned d...	48	Positive
1	Product arrived labeled as Jumbo Salted Peanut...	1	1346976000	Not as Advertised	B00813GRG4	A1D87F6ZCVE5NK	product arrived labeled as jumbo salted peanut...	32	Negative
2	This is a confection that has been around a fe...	4	1219017600	"Delight" says it all	B000LQOCH0	ABXLMWJIXXAIN	this is a confection that has been around a fe...	93	Positive
3	If you are looking for the secret ingredient i...	2	1307923200	Cough Medicine	B000UA0QIQ	A395BORC6FGVXV	if you are looking for the secret ingredient i...	41	Negative
4	Great taffy at a great price. There was a wid...	5	1350777600	Great taffy	B006K2ZZ7K	A1UQRSCLF8GW1T	great taffy at a great price there was a wide ...	27	Positive

In [72]:

```
# Remove stopwords and single letters
def remove_stopwords_and_shortwords(text):
    words = text.split()
    words = [w for w in words if w not in ENGLISH_STOP_WORDS and len(w) > 2]
    return words

df["tokens"] = df["clean_text"].apply(remove_stopwords_and_shortwords)

# Flatten list of all tokens
all_tokens = [word for tokens in df["tokens"] for word in tokens]

# Count frequency
word_counts = Counter(all_tokens)

# Top 20 words
most_common_words = word_counts.most_common(20)
most_common_words
```

Out[72]:

```
[('like', 256215),
 ('good', 200644),
 ('just', 172973),
 ('taste', 172831),
 ('great', 167175),
 ('coffee', 166784),
 ('product', 151884),
 ('flavor', 148028),
 ('tea', 138153),
 ('food', 128525),
 ('love', 127520),
 ('really', 101076),
 ('don', 91874),
 ('amazon', 90504),
 ('time', 84769),
 ('use', 83908),
 ('little', 83499),
 ('buy', 76916),
 ('best', 76837),
 ('tried', 76486)]
```

In [73]:

```
words, counts = zip(*most_common_words)

plt.figure(figsize=(10,5))
sns.barplot(x=list(counts), y=list(words), palette="coolwarm")
plt.title("Top 20 Most Frequent Words (Cleaned, Stopwords Removed)")
plt.xlabel("Frequency")
plt.show()
```

C:\Users\Vedant\AppData\Local\Temp\ipykernel\_2548\3736262562.py:4: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same effect.

```
sns.barplot(x=list(counts), y=list(words), palette="coolwarm")
```

Word	Frequency
like	260000
good	200000
just	175000
taste	170000
great	165000
coffee	165000
product	150000
flavor	145000
tea	135000
food	125000
love	125000
really	100000
don	90000
amazon	90000
time	85000
use	85000
little	85000
buy	75000
best	75000
tried	75000

A word cloud of negative reviews for a product. The most prominent words are 'product', 'taste', 'flavor', 'used', 'bought', 'tried', 'say', 'think', 'better', 'smell', 'waste', 'money', 'hard', 'look', 'got', 'expensive', 'one's', 'received', 'looked', 'really', 'lot', 'cat', 'food', 'case', 'bottle', 'container', 'tasted', 'actually', 'purchased', 'don't', 'know', 'gave', 'amazon', 'disappointed', 'big', 'little', 'feel', 'good', 'drink', 'great', 'fine', 'jar', 'bit', 'problem', 'issue', 'soup', 'sauce', 'packaging', 'dog', 'bad', 'treat', 'science', 'diet', 'water', 'sweet', 'cookie', 'snack', 'minute', 'leave', 'coffee', 'isn't', 'purchase', 'try', 'work', 'ones', 'fact', 'green', 'tea', 'piece', 'tea', 'smell', 'hard', 'waste', 'money', 'add', 'type', 'ingredient', 'probably', 'small', 'maybe', 'wrong', 'need', 'enjoy', 'company', 'know', 'best', 'mix', 'gluten', 'free', 'stick', 'cereal', 'wanted', 'package', 'expiration', 'date', 'kind', 'grocery', 'store', 'cost', 'terrible', 'month', 'sure', 'buy', 'people', 'eat', 'cup', 'didn't', 'expected', 'instead', 'far', 'ordered', 'used', 'thought', 'want', 'mean', 'guess', 'texture', 'left', 'year', 'price', 'thing', 'come', 'looking', 'love', 'brand', 'make', 'order', 'bar', 'going', 'using', 'came', 'customer', 'service', 'unfortunately', 'definitely', 'said', 'nice', 'bo', 'to', 'a', 'of', 'the', 'this', 'that', 'it', 'he', 'she', 'we', 'you', 'they', 'me', 'him', 'her', 'us', 'them', 'one', 'two', 'three', 'four', 'five', 'six', 'seven', 'eight', 'nine', 'ten', 'eleven', 'twelve', 'thirteen', 'fourteen', 'fifteen', 'sixteen', 'seventeen', 'eighteen', 'nineteen', 'twenty', 'twenty-one', 'twenty-two', 'twenty-three', 'twenty-four', 'twenty-five', 'twenty-six', 'twenty-seven', 'twenty-eight', 'twenty-nine', 'thirty', 'thirty-one', 'thirty-two', 'thirty-three', 'thirty-four', 'thirty-five', 'thirty-six', 'thirty-seven', 'thirty-eight', 'thirty-nine', 'forty', 'forty-one', 'forty-two', 'forty-three', 'forty-four', 'forty-five', 'forty-six', 'forty-seven', 'forty-eight', 'forty-nine', 'fifty', 'fifty-one', 'fifty-two', 'fifty-three', 'fifty-four', 'fifty-five', 'fifty-six', 'fifty-seven', 'fifty-eight', 'fifty-nine', 'sixty', 'sixty-one', 'sixty-two', 'sixty-three', 'sixty-four', 'sixty-five', 'sixty-six', 'sixty-seven', 'sixty-eight', 'sixty-nine', 'seventy', 'seventy-one', 'seventy-two', 'seventy-three', 'seventy-four', 'seventy-five', 'seventy-six', 'seventy-seven', 'seventy-eight', 'seventy-nine', 'eighty', 'eighty-one', 'eighty-two', 'eighty-three', 'eighty-four', 'eighty-five', 'eighty-six', 'eighty-seven', 'eighty-eight', 'eighty-nine', 'ninety', 'ninety-one', 'ninety-two', 'ninety-three', 'ninety-four', 'ninety-five', 'ninety-six', 'ninety-seven', 'ninety-eight', 'ninety-nine', 'hundred', 'hundred-one', 'hundred-two', 'hundred-three', 'hundred-four', 'hundred-five', 'hundred-six', 'hundred-seven', 'hundred-eight', 'hundred-nine', 'hundred-ten', 'hundred-eleven', 'hundred-twelve', 'hundred-thirteen', 'hundred-fourteen', 'hundred-fifteen', 'hundred-sixteen', 'hundred-seventeen', 'hundred-eighteen', 'hundred-nineteen', 'hundred-twenty', 'hundred-twenty-one', 'hundred-twenty-two', 'hundred-twenty-three', 'hundred-twenty-four', 'hundred-twenty-five', 'hundred-twenty-six', 'hundred-twenty-seven', 'hundred-twenty-eight', 'hundred-twenty-nine', 'hundred-thirty', 'hundred-thirty-one', 'hundred-thirty-two', 'hundred-thirty-three', 'hundred-thirty-four', 'hundred-thirty-five', 'hundred-thirty-six', 'hundred-thirty-seven', 'hundred-thirty-eight', 'hundred-thirty-nine', 'hundred-forty', 'hundred-forty-one', 'hundred-forty-two', 'hundred-forty-three', 'hundred-forty-four', 'hundred-forty-five', 'hundred-forty-six', 'hundred-forty-seven', 'hundred-forty-eight', 'hundred-forty-nine', 'hundred-fifty', 'hundred-fifty-one', 'hundred-fifty-two', 'hundred-fifty-three', 'hundred-fifty-four', 'hundred-fifty-five', 'hundred-fifty-six', 'hundred-fifty-seven', 'hundred-fifty-eight', 'hundred-fifty-nine', 'hundred-sixty', 'hundred-sixty-one', 'hundred-sixty-two', 'hundred-sixty-three', 'hundred-sixty-four', 'hundred-sixty-five', 'hundred-sixty-six', 'hundred-sixty-seven', 'hundred-sixty-eight', 'hundred-sixty-nine', 'hundred-seventy', 'hundred-seventy-one', 'hundred-seventy-two', 'hundred-seventy-three', 'hundred-seventy-four', 'hundred-seventy-five', 'hundred-seventy-six', 'hundred-seventy-seven', 'hundred-seventy-eight', 'hundred-seventy-nine', 'hundred-eighty', 'hundred-eighty-one', 'hundred-eighty-two', 'hundred-eighty-three', 'hundred-eighty-four', 'hundred-eighty-five', 'hundred-eighty-six', 'hundred-eighty-seven', 'hundred-eighty-eight', 'hundred-eighty-nine', 'hundred-ninety', 'hundred-ninety-one', 'hundred-ninety-two', 'hundred-ninety-three', 'hundred-ninety-four', 'hundred-ninety-five', 'hundred-ninety-six', 'hundred-ninety-seven', 'hundred-ninety-eight', 'hundred-ninety-nine', 'one thousand', 'one thousand one', 'one thousand two', 'one thousand three', 'one thousand four', 'one thousand five', 'one thousand six', 'one thousand seven', 'one thousand eight', 'one thousand nine', 'one thousand ten', 'one thousand eleven', 'one thousand twelve', 'one thousand thirteen', 'one thousand fourteen', 'one thousand fifteen', 'one thousand sixteen', 'one thousand seventeen', 'one thousand eighteen', 'one thousand nineteen', 'one thousand twenty', 'one thousand twenty-one', 'one thousand twenty-two', 'one thousand twenty-three', 'one thousand twenty-four', 'one thousand twenty-five', 'one thousand twenty-six', 'one thousand twenty-seven', 'one thousand twenty-eight', 'one thousand twenty-nine', 'one thousand thirty', 'one thousand thirty-one', 'one thousand thirty-two', 'one thousand thirty-three', 'one thousand thirty-four', 'one thousand thirty-five', 'one thousand thirty-six', 'one thousand thirty-seven', 'one thousand thirty-eight', 'one thousand thirty-nine', 'one thousand forty', 'one thousand forty-one', 'one thousand forty-two', 'one thousand forty-three', 'one thousand forty-four', 'one thousand forty-five', 'one thousand forty-six', 'one thousand forty-seven', 'one thousand forty-eight', 'one thousand forty-nine', 'one thousand five hundred', 'one thousand five hundred one', 'one thousand five hundred two', 'one thousand five hundred three', 'one thousand five hundred four', 'one thousand five hundred five', 'one thousand five hundred six', 'one thousand five hundred seven', 'one thousand five hundred eight', 'one thousand five hundred nine', 'one thousand five hundred ten', 'one thousand five hundred eleven', 'one thousand five hundred twelve', 'one thousand five hundred thirteen', 'one thousand five hundred fourteen', 'one thousand five hundred fifteen', 'one thousand five hundred sixteen', 'one thousand five hundred seventeen', 'one thousand five hundred eighteen', 'one thousand five hundred nineteen', 'one thousand five hundred twenty', 'one thousand five hundred twenty-one', 'one thousand five hundred twenty-two', 'one thousand five hundred twenty-three', 'one thousand five hundred twenty-four', 'one thousand five hundred twenty-five', 'one thousand five hundred twenty-six', 'one thousand five hundred twenty-seven', 'one thousand five hundred twenty-eight', 'one thousand five hundred twenty-nine', 'one thousand five hundred thirty', 'one thousand five hundred thirty-one', 'one thousand five hundred thirty-two', 'one thousand five hundred thirty-three', 'one thousand five hundred thirty-four', 'one thousand five hundred thirty-five', 'one thousand five hundred thirty-six', 'one thousand five hundred thirty-seven', 'one thousand five hundred thirty-eight', 'one thousand five hundred thirty-nine', 'one thousand five hundred forty', 'one thousand five hundred forty-one', 'one thousand five hundred forty-two', 'one thousand five hundred forty-three', 'one thousand five hundred forty-four', 'one thousand five hundred forty-five', 'one thousand five hundred forty-six', 'one thousand five hundred forty-seven', 'one thousand five hundred forty-eight', 'one thousand five hundred forty-nine', 'one thousand five hundred fifty', 'one thousand five hundred fifty-one', 'one thousand five hundred fifty-two', 'one thousand five hundred fifty-three', 'one thousand five hundred fifty-four', 'one thousand five hundred fifty-five', 'one thousand five hundred fifty-six', 'one thousand five hundred fifty-seven', 'one thousand five hundred fifty-eight', 'one thousand five hundred fifty-nine', 'one thousand five hundred sixty', 'one thousand five hundred sixty-one', 'one thousand five hundred sixty-two', 'one thousand five hundred sixty-three', 'one thousand five hundred sixty-four', 'one thousand five hundred sixty-five', 'one thousand five hundred sixty-six', 'one thousand five hundred sixty-seven', 'one thousand five hundred sixty-eight', 'one thousand five hundred sixty-nine', 'one thousand five hundred seventy', 'one thousand five hundred seventy-one', 'one thousand five hundred seventy-two', 'one thousand five hundred seventy-three', 'one thousand five hundred seventy-four', 'one thousand five hundred seventy-five', 'one thousand five hundred seventy-six', 'one thousand five hundred seventy-seven', 'one thousand five hundred seventy-eight', 'one thousand five hundred seventy-nine', 'one thousand five hundred eighty', 'one thousand five hundred eighty-one', 'one thousand five hundred eighty-two', 'one thousand five hundred eighty-three', 'one thousand five hundred eighty-four', 'one thousand five hundred eighty-five', 'one thousand five hundred eighty-six', 'one thousand five hundred eighty-seven', 'one thousand five hundred eighty-eight', 'one thousand five hundred eighty-nine', 'one thousand five hundred ninety', 'one thousand five hundred ninety-one', 'one thousand five hundred ninety-two', 'one thousand five hundred ninety-three', 'one thousand five hundred ninety-four', 'one thousand five hundred ninety-five', 'one thousand five hundred ninety-six', 'one thousand five hundred ninety-seven', 'one thousand five hundred ninety-eight', 'one thousand five hundred ninety-nine', 'two thousand', 'two thousand one', 'two thousand two', 'two thousand three', 'two thousand four', 'two thousand five', 'two thousand six', 'two thousand seven', 'two thousand eight', 'two thousand nine', 'two thousand ten', 'two thousand eleven', 'two thousand twelve', 'two thousand thirteen', 'two thousand fourteen', 'two thousand fifteen', 'two thousand sixteen', 'two thousand seventeen', 'two thousand eighteen', 'two thousand nineteen', 'two thousand twenty', 'two thousand twenty-one', 'two thousand twenty-two', 'two thousand twenty-three', 'two thousand twenty-four', 'two thousand twenty-five', 'two thousand twenty-six', 'two thousand twenty-seven', 'two thousand twenty-eight', 'two thousand twenty-nine', 'two thousand thirty', 'two thousand thirty-one', 'two thousand thirty-two', 'two thousand thirty-three', 'two thousand thirty-four', 'two thousand thirty-five', 'two thousand thirty-six', 'two thousand thirty-seven', 'two thousand thirty-eight', 'two thousand thirty-nine', 'two thousand forty', 'two thousand forty-one', 'two thousand forty-two', 'two thousand forty-three', 'two thousand forty-four', 'two thousand forty-five', 'two thousand forty-six', 'two thousand forty-seven', 'two thousand forty-eight', 'two thousand forty-nine', 'two thousand five hundred', 'two thousand five hundred one', 'two thousand five hundred two', 'two thousand five hundred three', 'two thousand five hundred four', 'two thousand five hundred five', 'two thousand five hundred six', 'two thousand five hundred seven', 'two thousand five hundred eight', 'two thousand five hundred nine', 'two thousand five hundred ten', 'two thousand five hundred eleven', 'two thousand five hundred twelve', 'two thousand five hundred thirteen', 'two thousand five hundred fourteen', 'two thousand five hundred fifteen', 'two thousand five hundred sixteen', 'two thousand five hundred seventeen', 'two thousand five hundred eighteen', 'two thousand five hundred nineteen', 'two thousand five hundred twenty', 'two thousand five hundred twenty-one', 'two thousand five hundred twenty-two', 'two thousand five hundred twenty-three', 'two thousand five hundred twenty-four', 'two thousand five hundred twenty-five', 'two thousand five hundred twenty-six', 'two thousand five hundred twenty-seven', 'two thousand five hundred twenty-eight', 'two thousand five hundred twenty-nine', 'two thousand five hundred thirty', 'two thousand five hundred thirty-one', 'two thousand five hundred thirty-two', 'two thousand five hundred thirty-three', 'two thousand five hundred thirty-four', 'two thousand five hundred thirty-five', 'two thousand five hundred thirty-six', 'two thousand five hundred thirty-seven', 'two thousand five hundred thirty-eight', 'two thousand five hundred thirty-nine', 'two thousand five hundred forty', 'two thousand five hundred forty-one', 'two thousand five hundred forty-two', 'two thousand five hundred forty-three', 'two thousand five hundred forty-four', 'two thousand five hundred forty-five', 'two thousand five hundred forty-six', 'two thousand five hundred forty-seven', 'two thousand five hundred forty-eight', 'two thousand five hundred forty-nine', 'two thousand five hundred fifty', 'two thousand five hundred fifty-one', 'two thousand five hundred fifty-two', 'two thousand five hundred fifty-three', 'two thousand five hundred fifty-four', 'two thousand five hundred fifty-five', 'two thousand five hundred fifty-six', 'two thousand five hundred fifty-seven', 'two thousand five hundred fifty-eight', 'two thousand five hundred fifty-nine', 'two thousand five hundred sixty', 'two thousand five hundred sixty-one', 'two thousand five hundred sixty-two', 'two thousand five hundred sixty-three', 'two thousand five hundred sixty-four', 'two thousand five hundred sixty-five', 'two thousand five hundred sixty-six', 'two thousand five hundred sixty-seven', 'two thousand five hundred sixty-eight', 'two thousand five hundred sixty-nine', 'two thousand five hundred seventy', 'two thousand five hundred seventy-one', 'two thousand five hundred seventy-two', 'two thousand five hundred seventy-three', 'two thousand five hundred seventy-four', 'two thousand five hundred seventy-five', 'two thousand five hundred seventy-six', 'two thousand five hundred seventy-seven', 'two thousand five hundred seventy-eight', 'two thousand five hundred seventy-nine', 'two thousand five hundred eighty', 'two thousand five hundred eighty-one', 'two thousand five hundred eighty-two', 'two thousand five hundred eighty-three', 'two thousand five hundred eighty-four', 'two thousand five hundred eighty-five', 'two thousand five hundred eighty-six', 'two thousand five hundred eighty-seven', 'two thousand five hundred eighty-eight', 'two thousand five hundred eighty-nine', 'two thousand five hundred ninety', 'two thousand five hundred ninety-one', 'two thousand five hundred ninety-two', 'two thousand five hundred ninety-three', 'two thousand five hundred ninety-four', 'two thousand five hundred ninety-five', 'two thousand five hundred ninety-six', 'two thousand five hundred ninety-seven', 'two thousand five hundred ninety-eight', 'two thousand five hundred ninety-nine', 'three thousand', 'three thousand one', 'three thousand two', 'three thousand three', 'three thousand four', 'three thousand five', 'three thousand six', 'three thousand seven', 'three thousand eight', 'three thousand nine', 'three thousand ten', 'three thousand eleven', 'three thousand twelve', 'three thousand thirteen', 'three thousand fourteen', 'three thousand fifteen', 'three thousand sixteen', 'three thousand seventeen', 'three thousand eighteen', 'three thousand nineteen', 'three thousand twenty', 'three thousand twenty-one', 'three thousand twenty-two', 'three thousand twenty-three', 'three thousand twenty-four', 'three thousand twenty-five', 'three thousand twenty-six', 'three thousand twenty-seven', 'three thousand twenty-eight', 'three thousand twenty-nine', 'three thousand thirty', 'three thousand thirty-one', 'three thousand thirty-two', 'three thousand thirty-three', 'three thousand thirty-four', 'three thousand thirty-five', 'three thousand thirty-six', 'three thousand thirty-seven', 'three thousand thirty-eight', 'three thousand thirty-nine', 'three thousand forty', 'three thousand forty-one', 'three thousand forty-two', 'three thousand forty-three', 'three thousand forty-four', 'three thousand forty-five', 'three thousand forty-six', 'three thousand forty-seven', 'three thousand forty-eight', 'three thousand forty-nine', 'three thousand five hundred', 'three thousand five hundred one', 'three thousand five hundred two', 'three thousand five hundred three', 'three thousand five hundred four', 'three thousand five hundred five', 'three thousand five hundred six

```
In [75]: from sklearn.feature_extraction.text import TfidfVectorizer

# Function to get top n keywords for a set of documents
def get_top_keywords(corpus, n=20):
    vectorizer = TfidfVectorizer(max_features=5000, stop_words='english', ngram_range=(1,2))
    X = vectorizer.fit_transform(corpus)
    feature_names = np.array(vectorizer.get_feature_names_out())
    # Sum TF-IDF scores across all documents
    scores = X.sum(axis=0).A1
    top_indices = scores.argsort()[-n:][::-1]
    top_features = feature_names[top_indices]
    top_scores = scores[top_indices]
```



```

return list(zip(top_features, top_scores))

# Prepare text by joining tokens back
df["processed_text"] = df["tokens"].apply(lambda x: " ".join(x))

negative_corpus = df[df["Sentiment"]=="Negative"]["processed_text"].tolist()

top_negative_keywords = get_top_keywords(negative_corpus, n=20)

print("\nTop Negative Keywords:\n", top_negative_keywords)

```

Top Negative Keywords:

```

[('like', 2518.952694463791), ('product', 2179.958248342551), ('taste', 2179.891925612207), ('coffee', 1993.978
713269519), ('just', 1736.7172527856949), ('flavor', 1605.3815041938315), ('good', 1559.8365575581413), ('tea',
1510.4296645448028), ('food', 1401.423855418517), ('buy', 1315.9309286582904), ('don', 1242.202368882589), ('ama
zon', 1216.0404654025765), ('really', 1126.777196284686), ('box', 1111.5371631866772), ('dog', 1071.988871884558
9), ('bought', 1039.9068516424252), ('tried', 1015.2221749030888), ('did', 954.3797367764944), ('eat', 954.01499
79351871), ('bad', 951.6739031326804)]

```

```

In [76]: def plot_top_keywords(keywords, title, color="blue"):
words, scores = zip(*keywords)
plt.figure(figsize=(10,6))
sns.barplot(x=list(scores), y=list(words), palette=color)
plt.xlabel("TF-IDF Score")
plt.title(title)
plt.show()

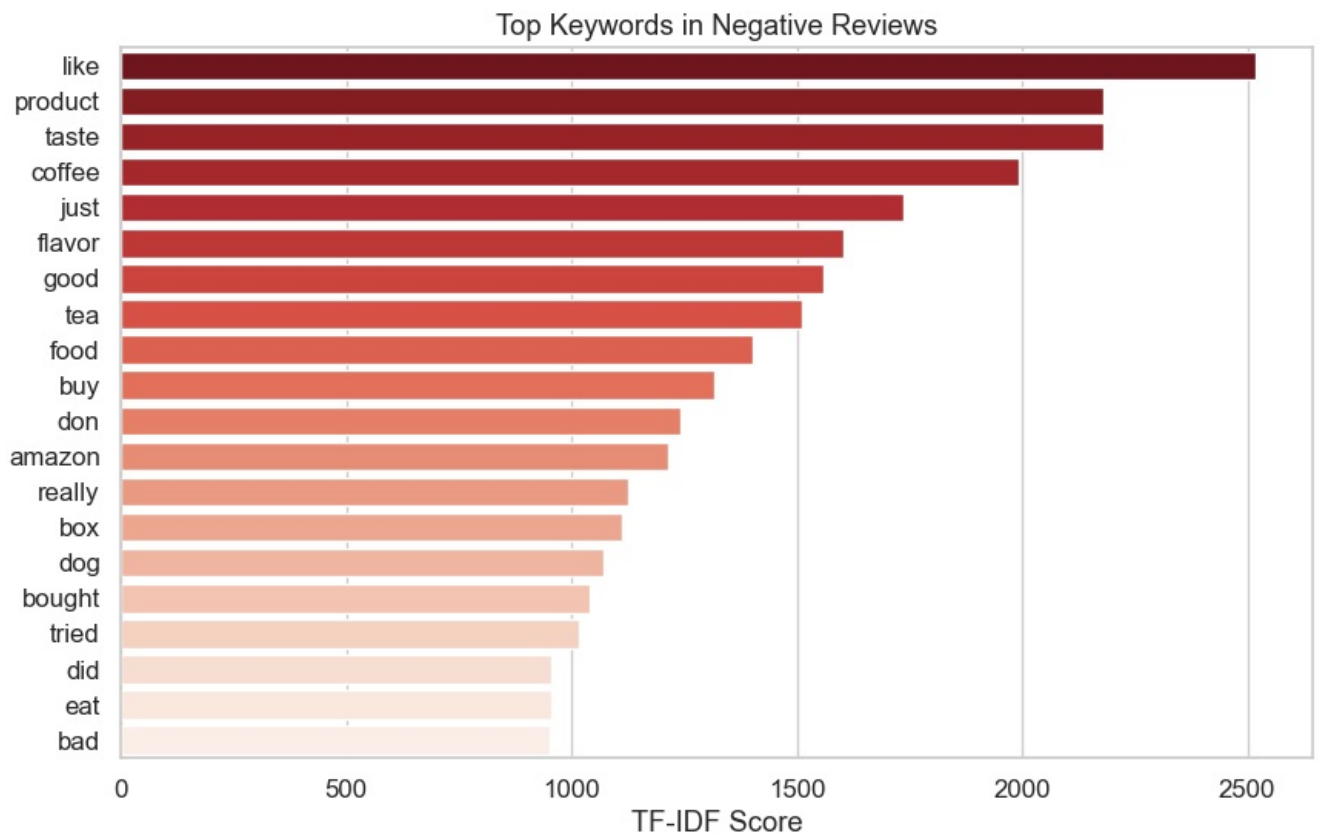
# Negative
plot_top_keywords(top_negative_keywords, "Top Keywords in Negative Reviews", color="Reds_r")

```

C:\Users\Vedant\AppData\Local\Temp\ipykernel\_2548\2597827445.py:4: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same effect.

```
sns.barplot(x=list(scores), y=list(words), palette=color)
```



In [ ]: