In [1]:
```
# Import necessary libraries

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

#Read the data

df = pd.read_csv('Reviews.csv', nrows=500)

# Look at the top 5 rows of the data

df.head(3)
```

Out[1]:

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

In [2]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 10 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   Id                      1000 non-null   int64 
 1   ProductId               1000 non-null   object
 2   UserId                  1000 non-null   object
 3   ProfileName             1000 non-null   object
 4   HelpfulnessNumerator    1000 non-null   int64 
 5   HelpfulnessDenominator  1000 non-null   int64 
 6   Score                   1000 non-null   int64 
 7   Time                    1000 non-null   int64 
 8   Summary                 1000 non-null   object
 9   Text                    1000 non-null   object
dtypes: int64(5), object(5)
memory usage: 78.2+ KB
```

In [3]:
```
df.Summary.head()
```

Out[3]:
```
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 [4]:
```
df.Text.head()
```

Out[4]:
```
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 [5]:
```
3!pip install TextBlob
```

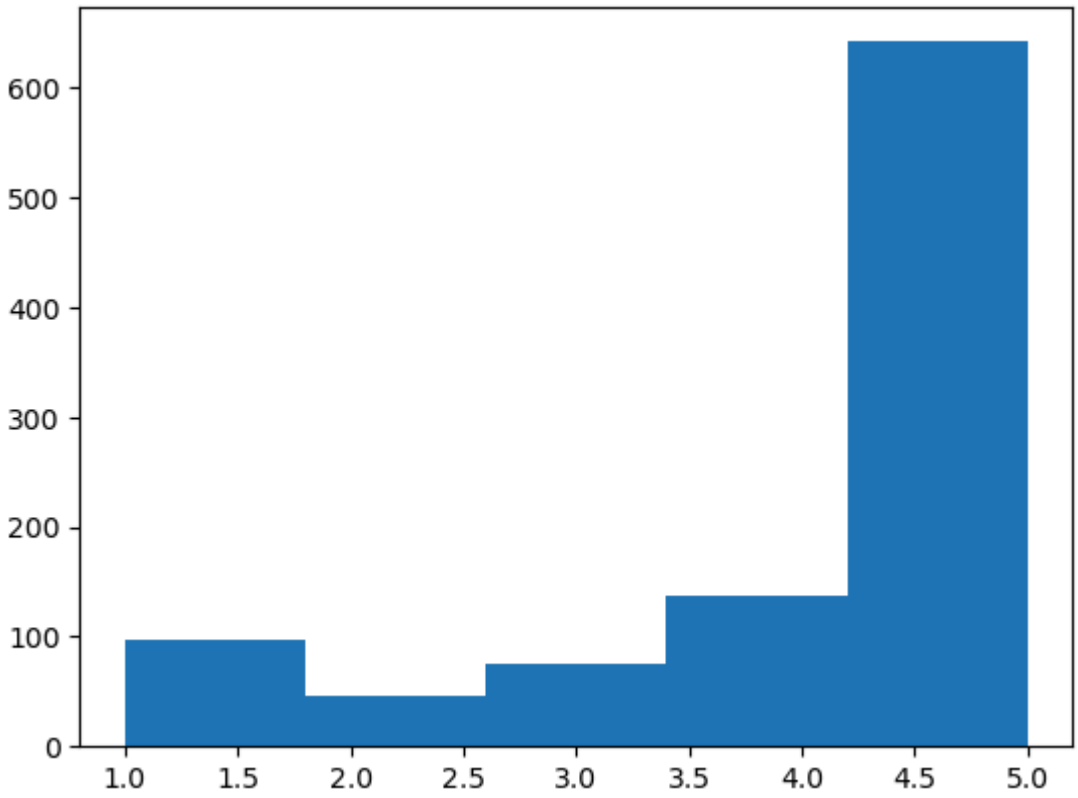
```
Requirement already satisfied: TextBlob in c:\users\91939\anaconda3\lib\site-packa
ges (0.19.0)
Requirement already satisfied: nltk>=3.9 in c:\users\91939\anaconda3\lib\site-pack
ages (from TextBlob) (3.9.1)
Requirement already satisfied: click in c:\users\91939\anaconda3\lib\site-packages
(from nltk>=3.9->TextBlob) (8.0.4)
Requirement already satisfied: tqdm in c:\users\91939\anaconda3\lib\site-packages
(from nltk>=3.9->TextBlob) (4.64.1)
Requirement already satisfied: joblib in c:\users\91939\anaconda3\lib\site-package
s (from nltk>=3.9->TextBlob) (1.1.0)
Requirement already satisfied: regex>=2021.8.3 in c:\users\91939\anaconda3\lib\sit
e-packages (from nltk>=3.9->TextBlob) (2022.7.9)
Requirement already satisfied: colorama in c:\users\91939\anaconda3\lib\site-packa
ges (from click->nltk>=3.9->TextBlob) (0.4.6)
```

In [6]:
```
# Import libraries
from nltk.corpus import stopwords
from textblob import TextBlob

from textblob import Word

# 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('[^\w\s]',' ')

# Removal of stop words

stop = 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

df.Text.head()
```

```
C:\Users\91939\AppData\Local\Temp\ipykernel_10716\368340856.py:13: FutureWarning:
The default value of regex will change from True to False in a future version.
  df['Text'] = df['Text'].str.replace('[^\w\s]',' ')
```

Out[6]:
```
0    bought several vitality canned dog food produc...
1    product arrived labelled lumbo halted peanut p...
2    connection around century light pillow city ge...
3    looking secret ingredient robitussin believe f...
4    great staff great price wide assortment mummy ...
Name: Text, dtype: object
```

In [7]:
```
# 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     98
2     47
3     75
4    138
5    642
Name: Id, dtype: int64
```

In [12]:
```
# To make it balanced data, we sampled each score by the lowest n-count from above.

score_1 = reviews[reviews['Score'] == 1].sample(n=47)
score_2 = reviews[reviews['Score'] == 2].sample(n=47)
score_3 = reviews[reviews['Score'] == 3].sample(n=47)
score_4 = reviews[reviews['Score'] == 4].sample(n=47)
score_5 = reviews[reviews['Score'] == 5].sample(n=47)
```

In [9]:
```
# 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    47
2    47
3    47
4    47
5    47
Name: Id, dtype: int64
```

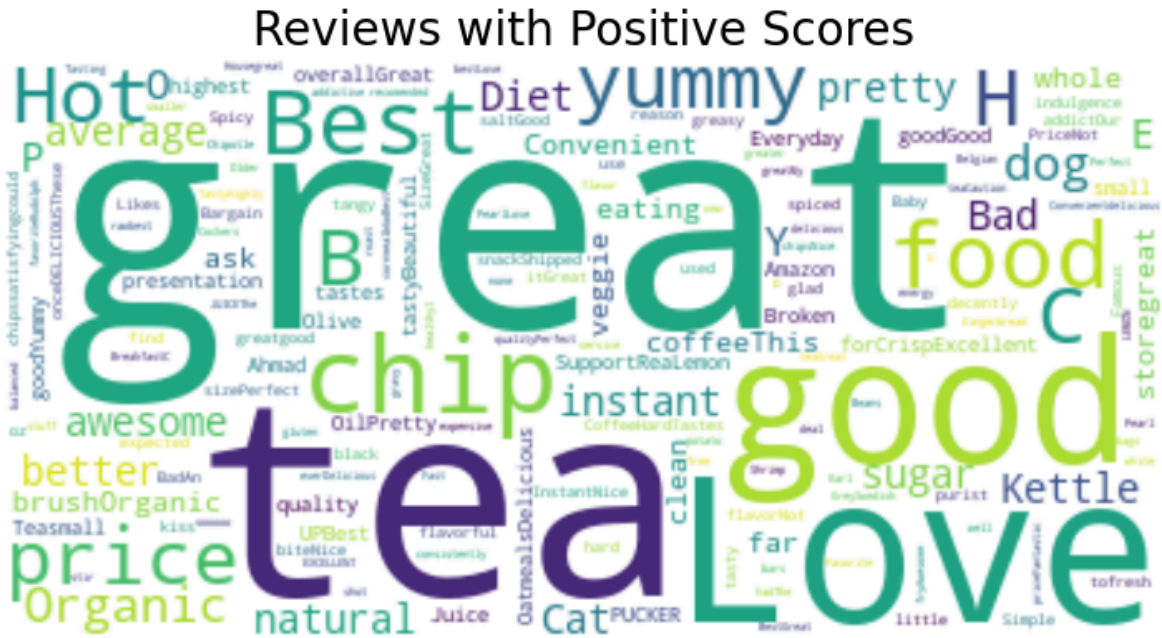
In [10]:
```
3!pip install WordCloud
```

```
Requirement already satisfied: WordCloud in c:\users\91939\anaconda3\lib\site-pack
ages (1.9.4)
Requirement already satisfied: numpy>=1.6.1 in c:\users\91939\anaconda3\lib\site-p
ackages (from WordCloud) (1.21.5)
Requirement already satisfied: matplotlib in c:\users\91939\anaconda3\lib\site-pac
kages (from WordCloud) (3.5.2)
Requirement already satisfied: pillow in c:\users\91939\anaconda3\lib\site-package
s (from WordCloud) (9.2.0)
Requirement already satisfied: packaging>=20.0 in c:\users\91939\anaconda3\lib\sit
e-packages (from matplotlib->WordCloud) (21.3)
Requirement already satisfied: pyparsing>=2.2.1 in c:\users\91939\anaconda3\lib\si
te-packages (from matplotlib->WordCloud) (3.0.9)
Requirement already satisfied: cycler>=0.10 in c:\users\91939\anaconda3\lib\site-p
ackages (from matplotlib->WordCloud) (0.11.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\91939\anaconda3\lib\s
ite-packages (from matplotlib->WordCloud) (1.4.2)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\91939\anaconda3\li
b\site-packages (from matplotlib->WordCloud) (2.8.2)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\91939\anaconda3\lib\s
ite-packages (from matplotlib->WordCloud) (4.25.0)
Requirement already satisfied: six>=1.5 in c:\users\91939\anaconda3\lib\site-packa
ges (from python-dateutil>=2.7->matplotlib->WordCloud) (1.16.0)
```

In [11]:
```
# Let's build a word cloud looking at the 'Summary'  text

from wordcloud import WordCloud
reviews_str = " ".join(reviews_sample["Summary"].to_numpy())
#reviews_str = reviews_sample.Summary.str.cat()
wordcloud = WordCloud(background_color='white').generate(reviews_str)

plt.figure(figsize=(10,10))

plt.imshow(wordcloud,interpolation='bilinear')

plt.axis("off")

plt.show()
```

In [20]:
```
# Now let's split the data into Negative (Score is 1 or 2) and Positive (4 or #5) R

negative_reviews = reviews_sample[reviews_sample['Score'].isin([1,2]) ]

positive_reviews = reviews_sample[reviews_sample['Score'].isin([4,5]) ]



# Transform to single string

negative_reviews_str = negative_reviews.Summary.str.cat()

positive_reviews_str = positive_reviews.Summary.str.cat()
```

In [22]:
```
# Create wordclouds

wordcloud_negative = WordCloud(background_color='white').generate(negative_reviews_

wordcloud_positive = WordCloud(background_color='white').generate(positive_reviews_

# Plot

fig = plt.figure(figsize=(10,10))

ax1 = fig.add_subplot(211)

ax1.imshow(wordcloud_negative,interpolation='bilinear')

ax1.axis("off")

ax1.set_title('Reviews with Negative Scores',fontsize=20)

ax2 = fig.add_subplot(212)

ax2.imshow(wordcloud_positive,interpolation='bilinear')

ax2.axis("off")

ax2.set_title('Reviews with Positive Scores',fontsize=20)

plt.show()
```



In [23]:
```
!pip install vaderSentiment
```

```
Collecting vaderSentiment
  Downloading vaderSentiment-3.3.2-py2.py3-none-any.whl (125 kB)
     -------------------------------------- 126.0/126.0 kB 3.6 MB/s eta 0:00:00
Requirement already satisfied: requests in c:\users\91939\anaconda3\lib\site-packa
ges (from vaderSentiment) (2.28.1)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\91939\anaconda3\l
ib\site-packages (from requests->vaderSentiment) (1.26.11)
Requirement already satisfied: charset-normalizer<3,>=2 in c:\users\91939\anaconda
3\lib\site-packages (from requests->vaderSentiment) (2.0.4)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\91939\anaconda3\lib
\site-packages (from requests->vaderSentiment) (2022.9.14)
Requirement already satisfied: idna<4,>=2.5 in c:\users\91939\anaconda3\lib\site-p
ackages (from requests->vaderSentiment) (3.3)
Installing collected packages: vaderSentiment
Successfully installed vaderSentiment-3.3.2
```

In [24]:
```
import seaborn as sns

from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
```

```
plt.style.use('fivethirtyeight')

# Function for getting the sentiment

cp = sns.color_palette()


analyzer = SentimentIntensityAnalyzer()


# Generating sentiment for all the sentence present in the dataset

emptyline=[]

for row in df['Text']:

    vs=analyzer.polarity_scores(row)

    emptyline.append(vs)
```

In [25]:
```
# Creating new dataframe with sentiments

df_sentiments=pd.DataFrame(emptyline)

df_sentiments.head()
```

Out[25]:

	neg	neu	pos	compound
0	0.000	0.503	0.497	0.9413
1	0.258	0.644	0.099	-0.5719
2	0.134	0.602	0.264	0.7880
3	0.000	0.854	0.146	0.4404
4	0.000	0.455	0.545	0.9186

In [ ]:
```

```