a1

October 6, 2022

[ ]:
```
#importing req. Lib.
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import re
import nltk
from nltk.corpus import stopwords
from sklearn.model_selection import train_test_split
from mlxtend.plotting import plot_confusion_matrix
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import␣
 ↪accuracy_score,confusion_matrix,classification_report

from wordcloud import WordCloud
```

[ ]:
```
#load our data set
data = pd.read_csv('Tweets.csv')
```

[ ]:
```
# data.head()
data.columns#
data.info()

print("Looking at unique values")
data.nunique()

print("Looking at null values")
data.isnull().sum()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14640 entries, 0 to 14639
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   tweet_id                      14640 non-null  int64
 1   airline_sentiment             14640 non-null  object
 2   airline_sentiment_confidence  14640 non-null  float64
```

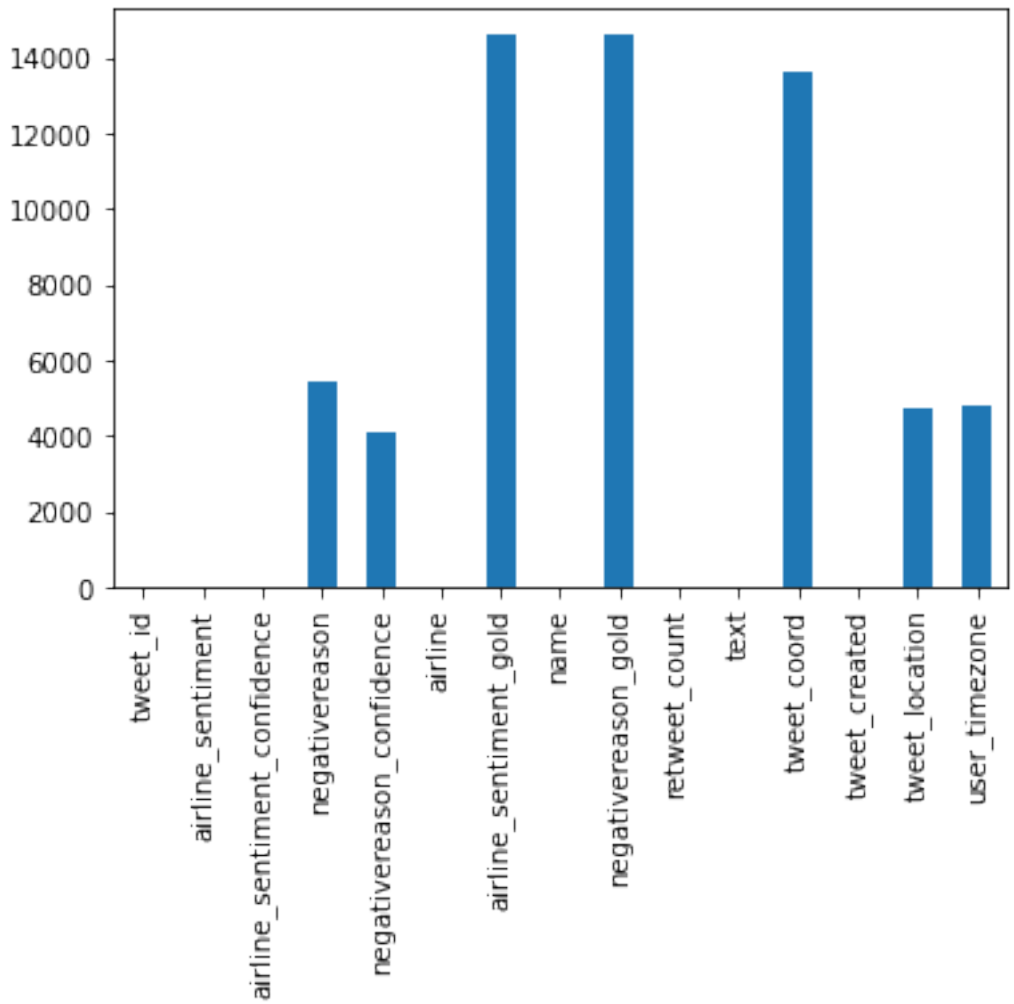
```
 3   negativereason                9178 non-null   object
 4   negativereason_confidence     10522 non-null  float64
 5   airline                       14640 non-null  object
 6   airline_sentiment_gold        40 non-null     object
 7   name                          14640 non-null  object
 8   negativereason_gold           32 non-null     object
 9   retweet_count                 14640 non-null  int64
 10  text                          14640 non-null  object
 11  tweet_coord                   1019 non-null   object
 12  tweet_created                 14640 non-null  object
 13  tweet_location                9907 non-null   object
 14  user_timezone                 9820 non-null   object
dtypes: float64(2), int64(2), object(11)
memory usage: 1.7+ MB
Looking at unique values
Looking at null values
```

[ ]:
```
tweet_id                          0
airline_sentiment                 0
airline_sentiment_confidence      0
negativereason                 5462
negativereason_confidence      4118
airline                           0
airline_sentiment_gold        14600
name                              0
negativereason_gold           14608
retweet_count                     0
text                              0
tweet_coord                   13621
tweet_created                     0
tweet_location                 4733
user_timezone                  4820
dtype: int64
```

[ ]:
```
print("Looking at null values")
data.isnull().sum().plot(kind='bar')
```

```
Looking at null values
```

[ ]:
```
<AxesSubplot:>
```



[ ]:
```
data['tweet_created'] = pd.to_datetime(data['tweet_created']).dt.date
data['tweet_created'] = pd.to_datetime(data['tweet_created'])
data.info()
data.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14640 entries, 0 to 14639
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   tweet_id                      14640 non-null  int64
 1   airline_sentiment             14640 non-null  object
 2   airline_sentiment_confidence  14640 non-null  float64
 3   negativereason                9178 non-null   object
 4   negativereason_confidence     10522 non-null  float64
 5   airline                       14640 non-null  object
```

```
 6   airline_sentiment_gold        40 non-null     object
 7   name                          14640 non-null  object
 8   negativereason_gold           32 non-null     object
 9   retweet_count                 14640 non-null  int64
 10  text                          14640 non-null  object
 11  tweet_coord                   1019 non-null   object
 12  tweet_created                 14640 non-null  datetime64[ns]
 13  tweet_location                9907 non-null   object
 14  user_timezone                 9820 non-null   object
dtypes: datetime64[ns](1), float64(2), int64(2), object(10)
memory usage: 1.7+ MB
```

[ ]:

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America

	airline_sentiment_gold	name	negativereason_gold	retweet_count	text	tweet_coord
0	NaN	cairdin	NaN	0	@VirginAmerica What @dhepburn said.	NaN
1	NaN	jnardino	NaN	0	@VirginAmerica plus you've added commercials t…	NaN
2	NaN	yvonnalynn	NaN	0	@VirginAmerica I didn't today… Must mean I n…	NaN
3	NaN	jnardino	NaN	0	@VirginAmerica it's really aggressive to blast…	NaN
4	NaN	jnardino	NaN	0	@VirginAmerica and it's a really big bad thing…	NaN

	tweet_created	tweet_location	user_timezone
0	2015-02-24	NaN	Eastern Time (US & Canada)
1	2015-02-24	NaN	Pacific Time (US & Canada)
2	2015-02-24	Lets Play	Central Time (US & Canada)
3	2015-02-24	NaN	Pacific Time (US & Canada)
4	2015-02-24	NaN	Pacific Time (US & Canada)

[ ]:
```
data['tweet_created'].nunique()
numberoftweets = data.groupby('tweet_created').size()
```

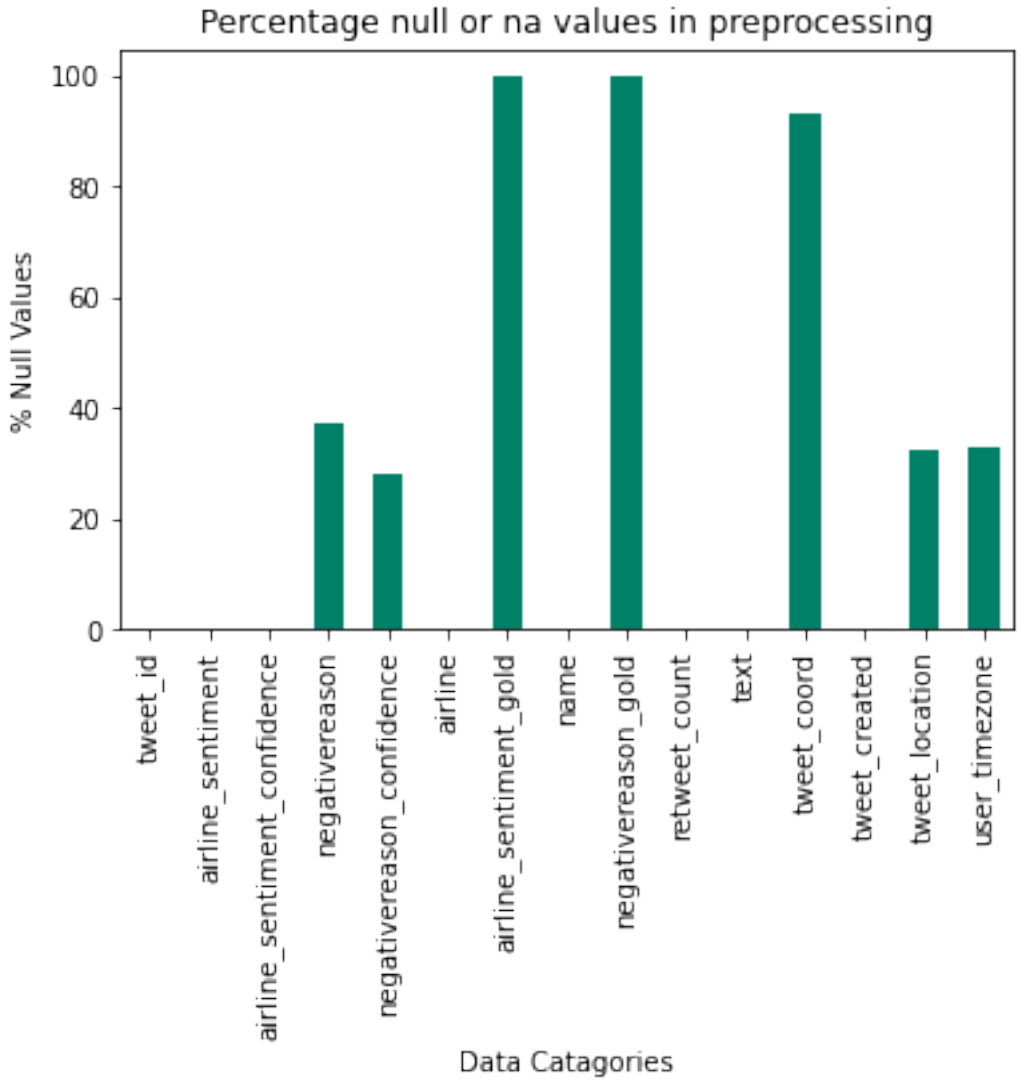
[ ]:
```
data.isna().sum()
```

[ ]:
```
tweet_id                          0
airline_sentiment                 0
airline_sentiment_confidence      0
negativereason                 5462
negativereason_confidence      4118
airline                           0
airline_sentiment_gold        14600
name                              0
negativereason_gold           14608
retweet_count                     0
text                              0
tweet_coord                   13621
tweet_created                     0
tweet_location                 4733
user_timezone                  4820
dtype: int64
```

[ ]:
```
print("Percentage null or na values in df")
nullvalue = ((data.isnull() | data.isna()).sum() * 100 / data.index.size).
 ↪round(2)
print(nullvalue)
nullvalue.plot(kind='bar', xlabel='Data Catagories', ylabel='% Null Values',␣
 ↪title='Percentage null or na values in preprocessing', colormap='summer')
```

```
Percentage null or na values in df
tweet_id                          0.00
airline_sentiment                 0.00
airline_sentiment_confidence      0.00
negativereason                   37.31
negativereason_confidence        28.13
airline                           0.00
airline_sentiment_gold           99.73
name                              0.00
negativereason_gold              99.78
retweet_count                     0.00
text                              0.00
tweet_coord                      93.04
tweet_created                     0.00
tweet_location                   32.33
user_timezone                    32.92
dtype: float64
```

[ ]: <AxesSubplot:title={'center':'Percentage null or na values in preprocessing'}, xlabel='Data Catagories', ylabel='% Null Values'>



[ ]:
```
del data['tweet_coord']
del data['airline_sentiment_gold']
del data['negativereason_gold']
data.head()
```

[ ]:
```
             tweet_id airline_sentiment  airline_sentiment_confidence  \
0  570306133677760513           neutral                        1.0000
1  570301130888122368          positive                        0.3486
2  570301083672813571           neutral                        0.6837
3  570301031407624196          negative                        1.0000
```

```
4  570300817074462722          negative                          1.0000

  negativereason  negativereason_confidence         airline        name  \
0            NaN                        NaN  Virgin America     cairdin
1            NaN                     0.0000  Virgin America    jnardino
2            NaN                        NaN  Virgin America  yvonnalynn
3     Bad Flight                     0.7033  Virgin America    jnardino
4     Can't Tell                     1.0000  Virgin America    jnardino

   retweet_count                                              text  \
0              0                 @VirginAmerica What @dhepburn said.
1              0  @VirginAmerica plus you've added commercials t…
2              0  @VirginAmerica I didn't today… Must mean I n…
3              0  @VirginAmerica it's really aggressive to blast…
4              0  @VirginAmerica and it's a really big bad thing…

  tweet_created tweet_location               user_timezone
0    2015-02-24            NaN  Eastern Time (US & Canada)
1    2015-02-24            NaN  Pacific Time (US & Canada)
2    2015-02-24      Lets Play  Central Time (US & Canada)
3    2015-02-24            NaN  Pacific Time (US & Canada)
4    2015-02-24            NaN  Pacific Time (US & Canada)
```

[ ]:
```
data.columns
data.head()
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14640 entries, 0 to 14639
Data columns (total 12 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   tweet_id                       14640 non-null  int64
 1   airline_sentiment              14640 non-null  object
 2   airline_sentiment_confidence   14640 non-null  float64
 3   negativereason                 9178 non-null   object
 4   negativereason_confidence      10522 non-null  float64
 5   airline                        14640 non-null  object
 6   name                           14640 non-null  object
 7   retweet_count                  14640 non-null  int64
 8   text                           14640 non-null  object
 9   tweet_created                  14640 non-null  datetime64[ns]
 10  tweet_location                 9907 non-null   object
 11  user_timezone                  9820 non-null   object
dtypes: datetime64[ns](1), float64(2), int64(2), object(7)
memory usage: 1.3+ MB
```

[ ]:
```
neg_df = data.negativereason.value_counts().to_frame()
wordcloud = WordCloud(min_font_size=10, width = 800, height = 800, colormap =␣
 ↪'Set2', background_color='white').generate_from_frequencies(neg_df.
 ↪negativereason)
plt.figure(figsize = (12,12))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
# neg_df.head()
```

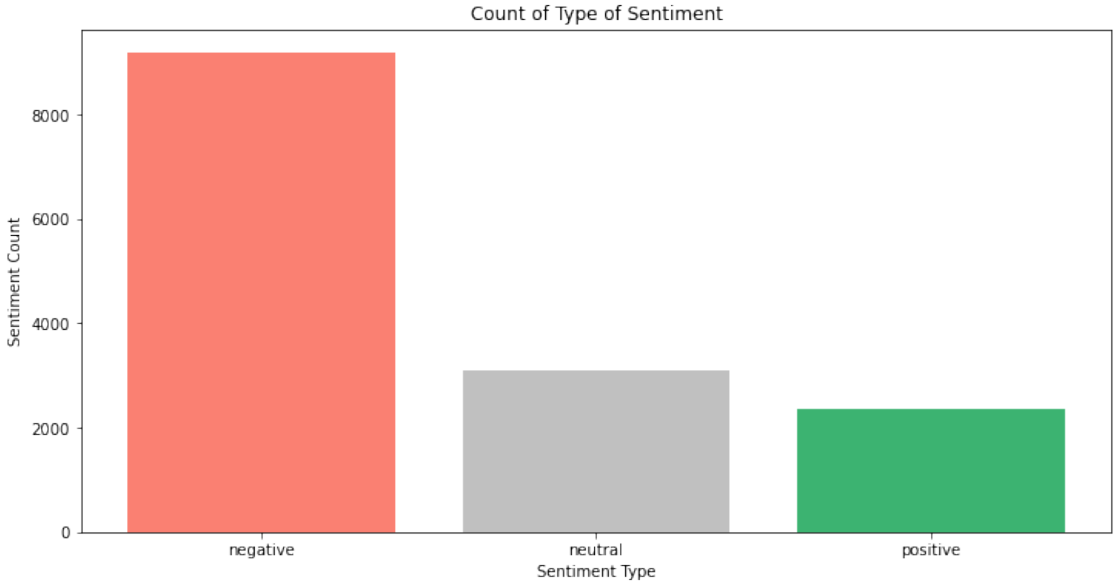
[ ]:
```
freq = data.groupby('negativereason').size()
freq
```

[ ]:
```
negativereason
Bad Flight                      580
Can't Tell                     1190
Cancelled Flight                847
Customer Service Issue         2910
Damaged Luggage                  74
Flight Attendant Complaints     481
Flight Booking Problems         529
Late Flight                    1665
Lost Luggage                    724
longlines                       178
dtype: int64
```

0.0.1 Beginning our EDA

[ ]:
```
counter = data.airline_sentiment.value_counts()
index = [1,2,3]
plt.figure(1,figsize=(12,6))
plt.bar(index,counter, color=['salmon','silver','mediumseagreen'])
plt.xticks(index,['negative','neutral','positive'],rotation=0)
plt.xlabel('Sentiment Type')
plt.ylabel('Sentiment Count')
plt.title('Count of Type of Sentiment')
```

[ ]:
```
Text(0.5, 1.0, 'Count of Type of Sentiment')
```

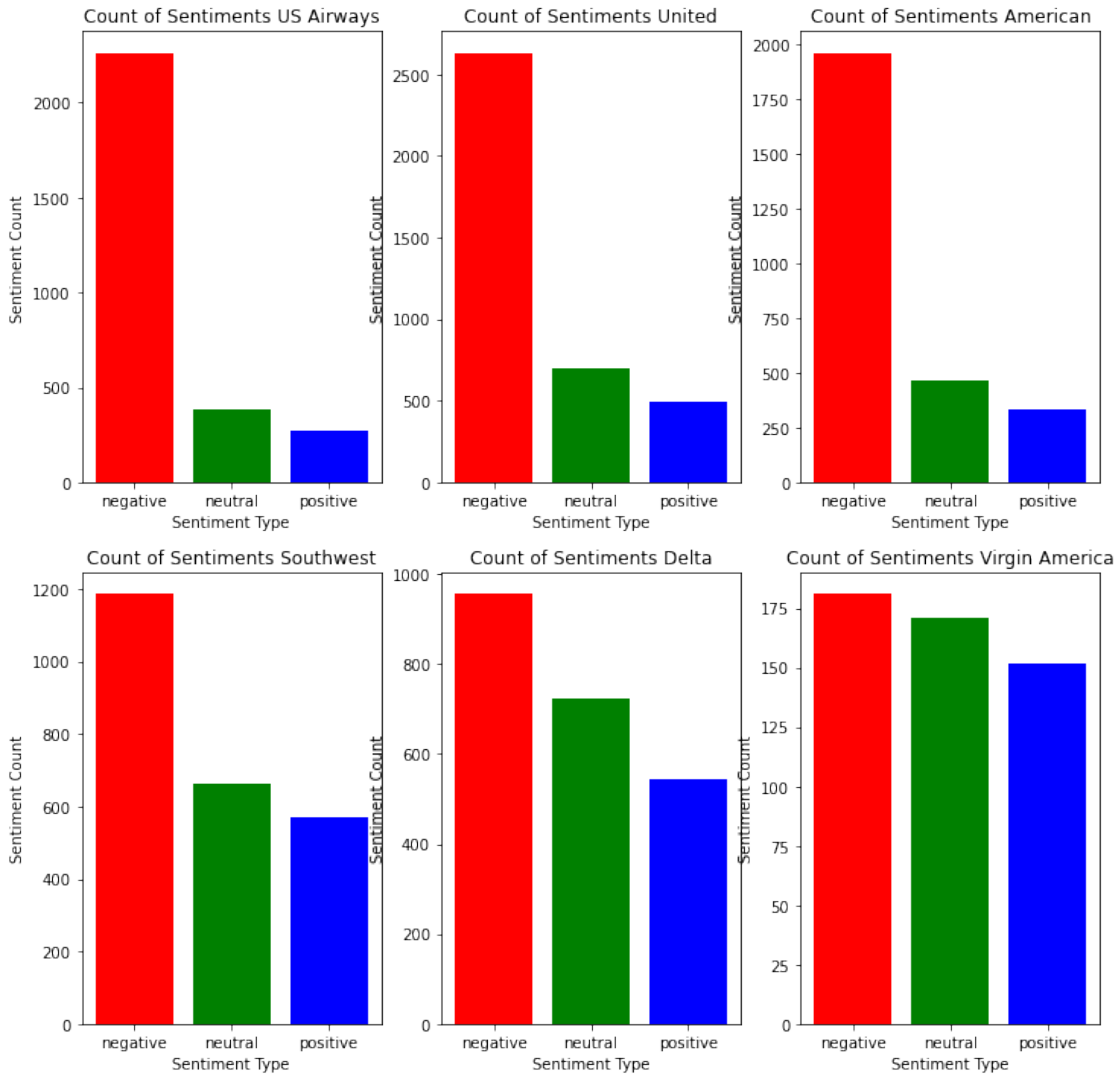


[ ]:
```
data['airline'].unique()
```

[ ]:
```
array(['Virgin America', 'United', 'Southwest', 'Delta', 'US Airways',
       'American'], dtype=object)
```

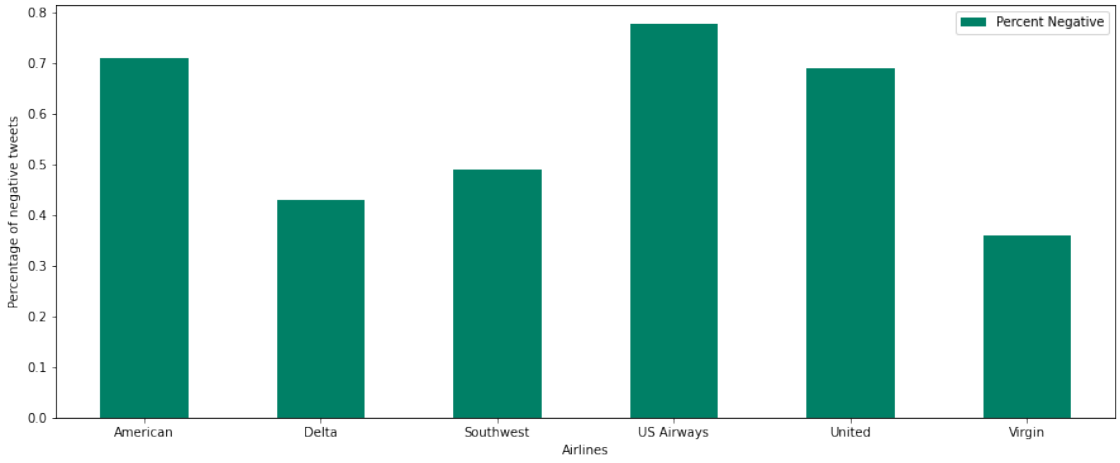
[ ]:
```
print("Total number of tweets for each airline \n ",data.
 ↪groupby('airline')['airline_sentiment'].count().sort_values(ascending=False))
airlines= ['US Airways','United','American','Southwest','Delta','Virgin␣
 ↪America']
plt.figure(1,figsize=(12, 12))
for i in airlines:
    indices= airlines.index(i)
    plt.subplot(2,3,indices+1)
    new_df=data[data['airline']==i]
    count=new_df['airline_sentiment'].value_counts()
    Index = [1,2,3]
    plt.bar(Index,count, color=['red', 'green', 'blue'])
    plt.xticks(Index,['negative','neutral','positive'])
    plt.ylabel('Sentiment Count')
    plt.xlabel('Sentiment Type')
    plt.title('Count of Sentiments '+i)
```

```
Total number of tweets for each airline
  airline
United            3822
US Airways        2913
American          2759
Southwest         2420
Delta             2222
Virgin America     504
Name: airline_sentiment, dtype: int64
```



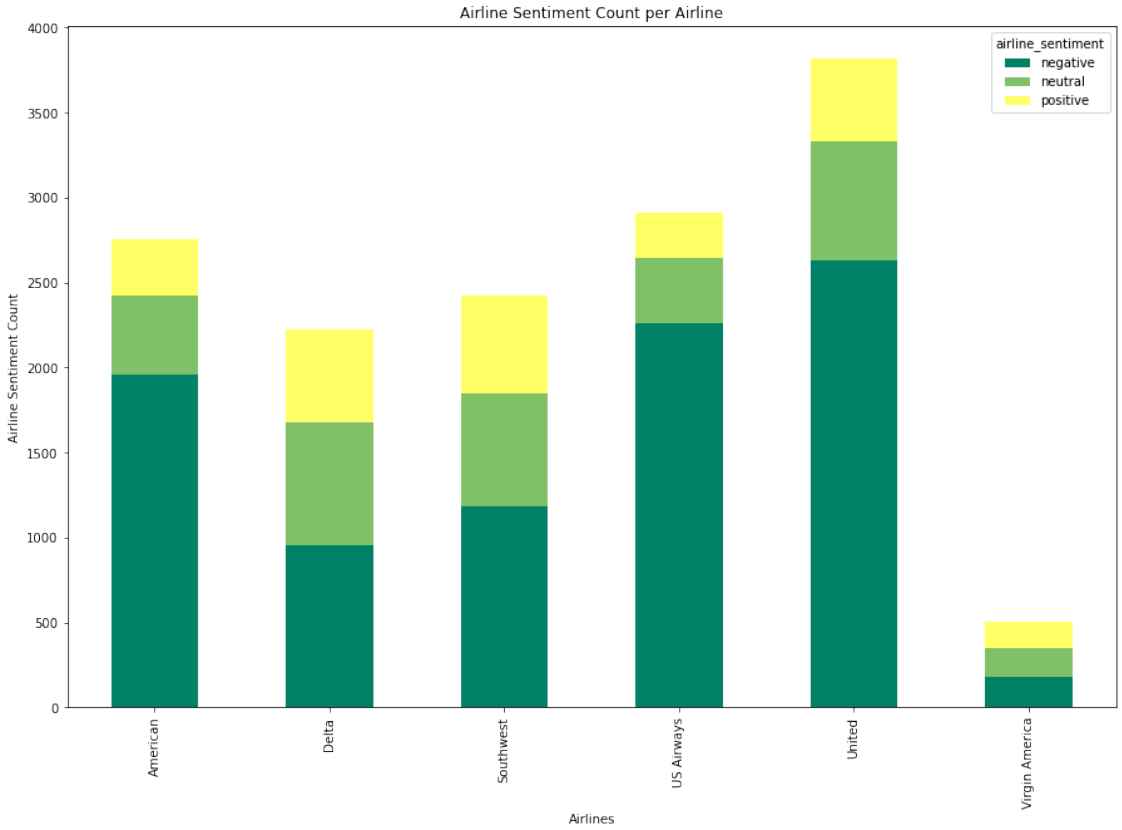
[ ]:
```
neg_tweets = data.groupby(['airline','airline_sentiment']).count().iloc[:,0]
total_tweets = data.groupby(['airline'])['airline_sentiment'].count()

my_dict = {'American':neg_tweets[0] / total_tweets[0],'Delta':neg_tweets[3] /␣
 ↪total_tweets[1],'Southwest': neg_tweets[6] / total_tweets[2],
'US Airways': neg_tweets[9] / total_tweets[3],'United': neg_tweets[12] /␣
 ↪total_tweets[4],'Virgin': neg_tweets[15] / total_tweets[5]}
perc = pd.DataFrame.from_dict(my_dict, orient = 'index')
perc.columns = ['Percent Negative']
print(perc)
ax = perc.plot(kind = 'bar', colormap='summer',rot=0, figsize = (15,6))
ax.set_xlabel('Airlines')
ax.set_ylabel('Percentage of negative tweets')
plt.show()
```

	Percent Negative
American	0.710402
Delta	0.429793
Southwest	0.490083
US Airways	0.776862
United	0.688906
Virgin	0.359127



[ ]:
```
from turtle import color


figure_2 = data.groupby(['airline', 'airline_sentiment']).size()
figure_2.unstack().plot(kind='bar', stacked=True, xlabel='Airlines',␣
 ↪ylabel='Airline Sentiment Count', title='Airline Sentiment Count per␣
 ↪Airline',figsize=(15,10), colormap = 'summer' )# color=['red', 'green',␣
 ↪'blue'])
```

[ ]:
```
<AxesSubplot:title={'center':'Airline Sentiment Count per Airline'},
xlabel='Airlines', ylabel='Airline Sentiment Count'>
```

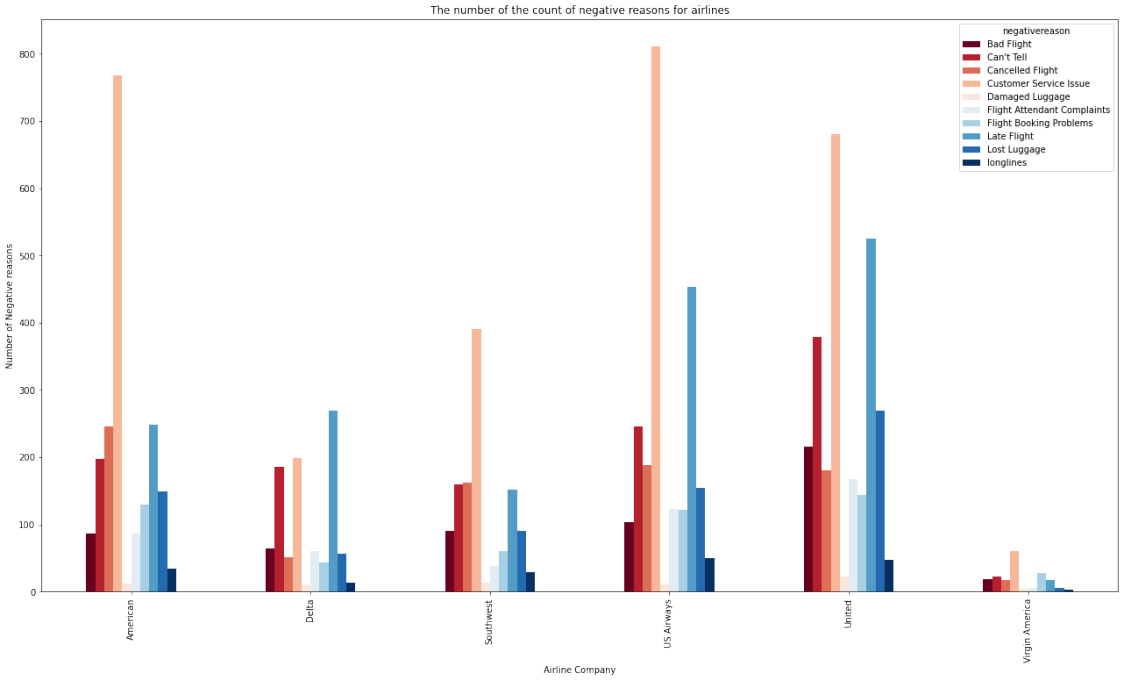


[ ]:
```
print(figure_2)
```

```
airline         airline_sentiment
American        negative             1960
                neutral               463
                positive              336
Delta           negative              955
                neutral               723
                positive              544
Southwest       negative             1186
                neutral               664
                positive              570
US Airways      negative             2263
                neutral               381
                positive              269
United          negative             2633
                neutral               697
                positive              492
Virgin America  negative              181
                neutral               171
                positive              152
```

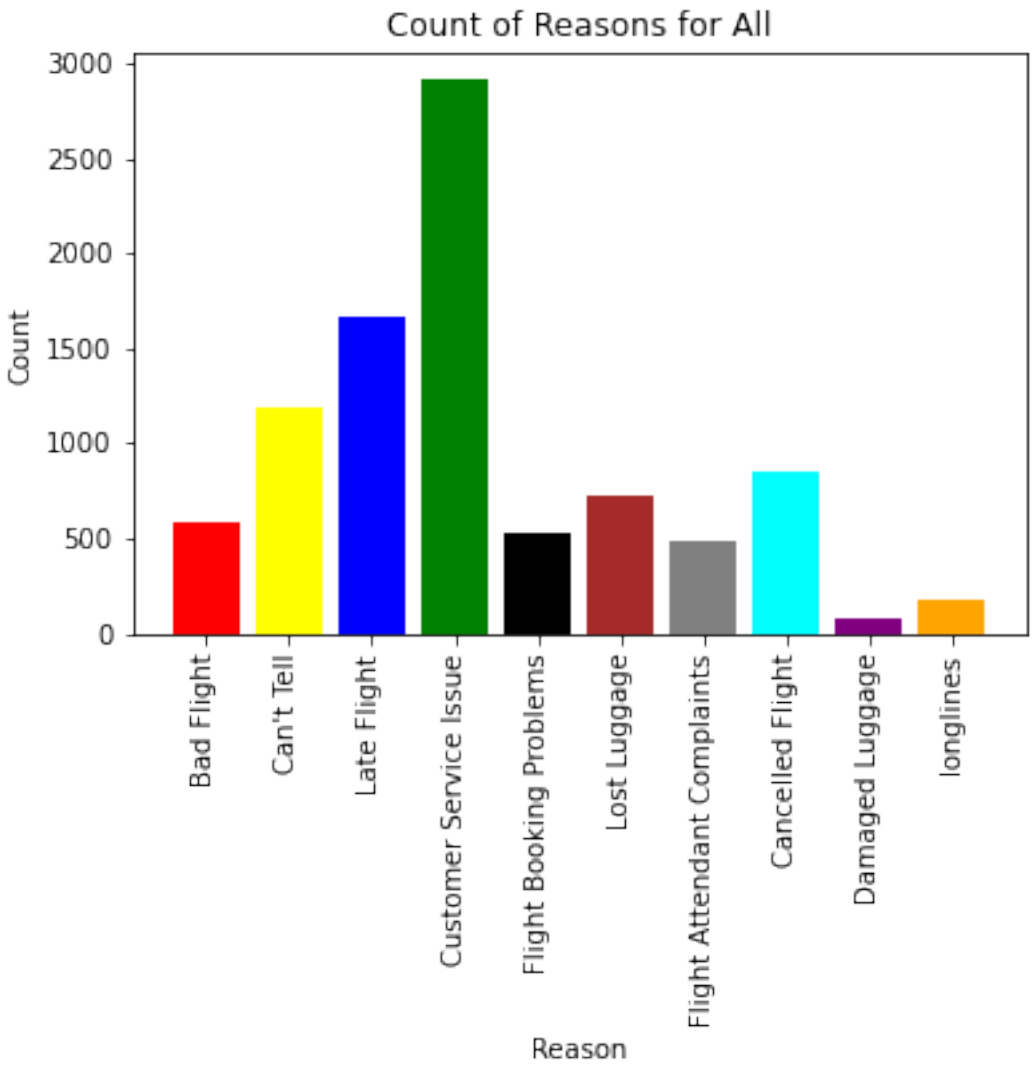
```
dtype: int64
```

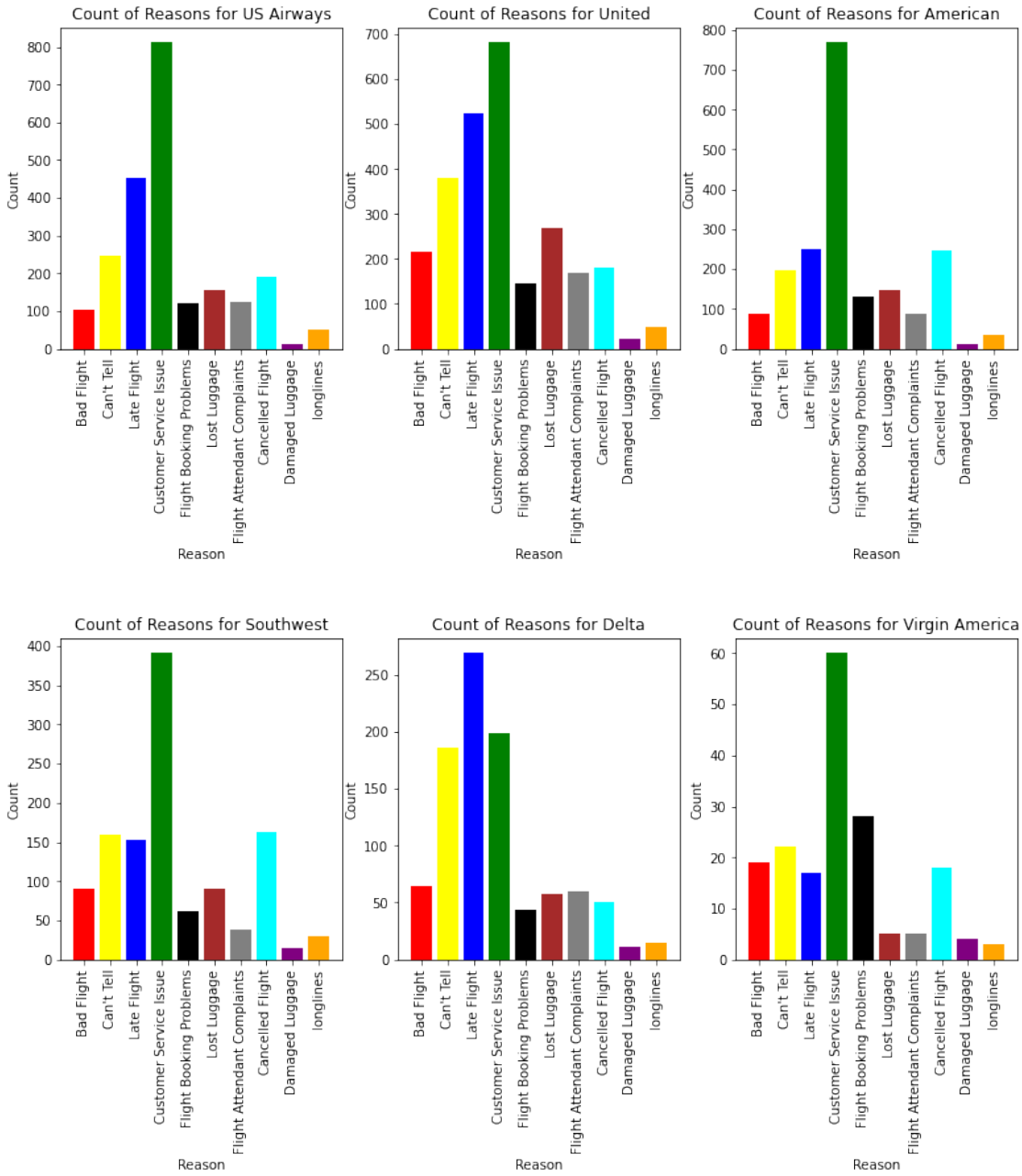
[ ]:
```
negative_reasons = data.groupby('airline')['negativereason'].
  ↪value_counts(ascending=True)
negative_reasons.groupby(['airline','negativereason']).sum().unstack().
  ↪plot(kind='bar',figsize=(22,12), colormap='RdBu')
plt.xlabel('Airline Company')
plt.ylabel('Number of Negative reasons')
plt.title("The number of the count of negative reasons for airlines")
plt.show()
```



[ ]:
```
#get the number of negative reasons
data['negativereason'].nunique()

NR_Count=dict(data['negativereason'].value_counts(sort=False))
def NR_Count(Airline):
    if Airline=='All':
        a=data
    else:
        a=data[data['airline']==Airline]
    count=dict(a['negativereason'].value_counts())
    Unique_reason=list(data['negativereason'].unique())
    Unique_reason=[x for x in Unique_reason if str(x) != 'nan']
    Reason_frame=pd.DataFrame({'Reasons':Unique_reason})
    Reason_frame['count']=Reason_frame['Reasons'].apply(lambda x: count[x])
```

```
    return Reason_frame
def plot_reason(Airline):

    a=NR_Count(Airline)
    count=a['count']
    Index = range(1,(len(a)+1))
    plt.bar(Index,count,␣
 ↪color=['red','yellow','blue','green','black','brown','gray','cyan','purple','orange'])
    plt.xticks(Index,a['Reasons'],rotation=90)
    plt.ylabel('Count')
    plt.xlabel('Reason')
    plt.title('Count of Reasons for '+Airline)

plot_reason('All')
plt.figure(2,figsize=(13, 13))
for i in airlines:
    indices= airlines.index(i)
    plt.subplot(2,3,indices+1)
    plt.subplots_adjust(hspace=0.9)
    plot_reason(i)
```


Count of Reasons for All
Count
3000
2500
2000
1500
1000
500
0
Bad Flight
Can't Tell
Late Flight
Customer Service Issue
Flight Booking Problems
Lost Luggage
Flight Attendant Complaints
Cancelled Flight
Damaged Luggage
longlines
Reason



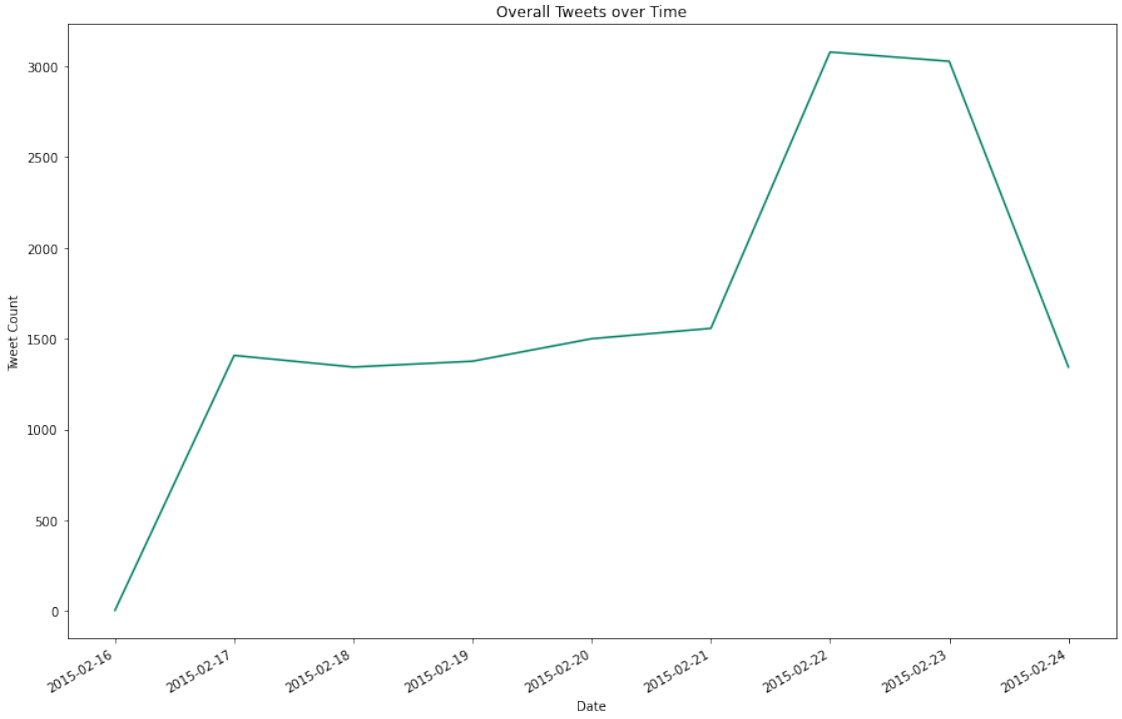
[ ]:
```
date = data.reset_index()
#convert the Date column to pandas datetime
date.tweet_created = pd.to_datetime(date.tweet_created)
#Reduce the dates in the date column to only the date and no time stamp using␣
 ↪the 'dt.date' method
date.tweet_created = date.tweet_created.dt.date
date.tweet_created.head()
df = date
```

```
# timeseries = data.groupby('tweet_created').size()

day_df = df.groupby(['tweet_created','airline','airline_sentiment']).size()
# day_df = day_df.reset_index()
day_df
```

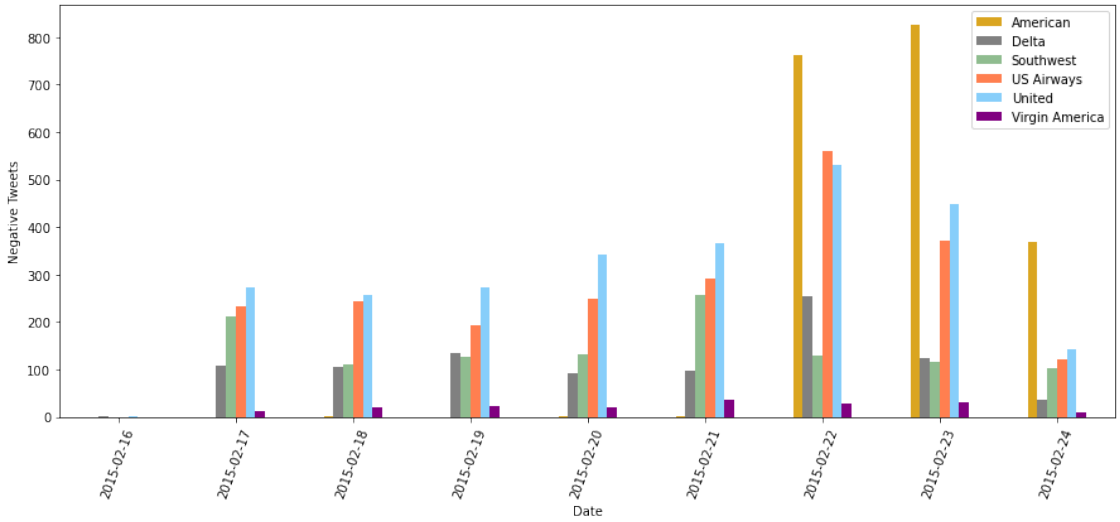
[ ]:
```
tweet_created  airline         airline_sentiment
2015-02-16     Delta           negative              1
                               neutral               1
               United          negative              2
2015-02-17     Delta           negative            108
                               neutral              86
                                                    ...
2015-02-24     United          neutral              49
                               positive             25
               Virgin America  negative             10
                               neutral               6
                               positive             13
Length: 136, dtype: int64
```

[ ]:
```
timeseries = data.tweet_created.value_counts()
timeseries.plot(figsize=(15,10), xlabel= "Date", ylabel = "Tweet␣
↪Count",title='Overall Tweets over Time', colormap='summer')
```

[ ]:
```
<AxesSubplot:title={'center':'Overall Tweets over Time'}, xlabel='Date',
ylabel='Tweet Count'>
```



[ ]:
```
day_df = day_df.loc(axis=0)[:,:,'negative']

#groupby and plot data
ax2 = day_df.groupby(['tweet_created','airline']).sum().unstack().plot(kind =␣
 ↪'bar', color=['goldenrod', 'grey',␣
 ↪'darkseagreen','coral','lightskyblue','purple'], figsize = (15,6), rot = 70)
labels = ['American','Delta','Southwest','US Airways','United','Virgin America']
ax2.legend(labels = labels)
ax2.set_xlabel('Date')
ax2.set_ylabel('Negative Tweets')
plt.show()
```



[ ]:
```
df = pd.read_csv('Tweets.csv')

display(df.shape, df.head(), df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14640 entries, 0 to 14639
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   tweet_id                      14640 non-null  int64
 1   airline_sentiment             14640 non-null  object
 2   airline_sentiment_confidence  14640 non-null  float64
 3   negativereason                9178 non-null   object
 4   negativereason_confidence     10522 non-null  float64
 5   airline                       14640 non-null  object
```

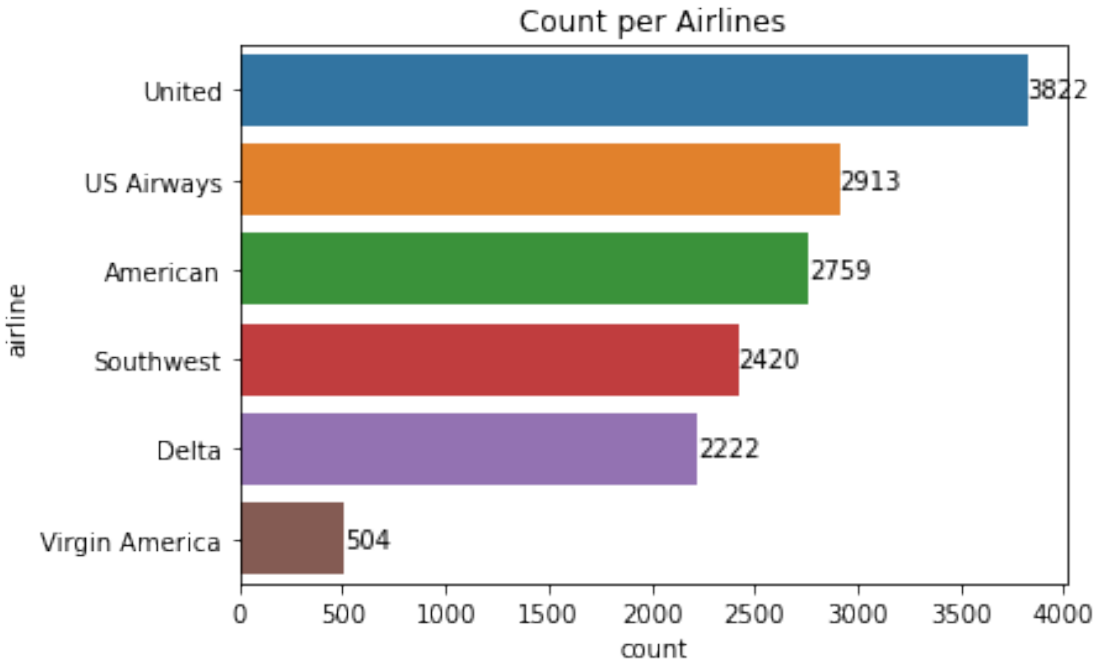
```
 6   airline_sentiment_gold        40 non-null     object
 7   name                          14640 non-null  object
 8   negativereason_gold           32 non-null     object
 9   retweet_count                 14640 non-null  int64
 10  text                          14640 non-null  object
 11  tweet_coord                   1019 non-null   object
 12  tweet_created                 14640 non-null  object
 13  tweet_location                9907 non-null   object
 14  user_timezone                 9820 non-null   object
dtypes: float64(2), int64(2), object(11)
memory usage: 1.7+ MB
```

```
(14640, 15)
```

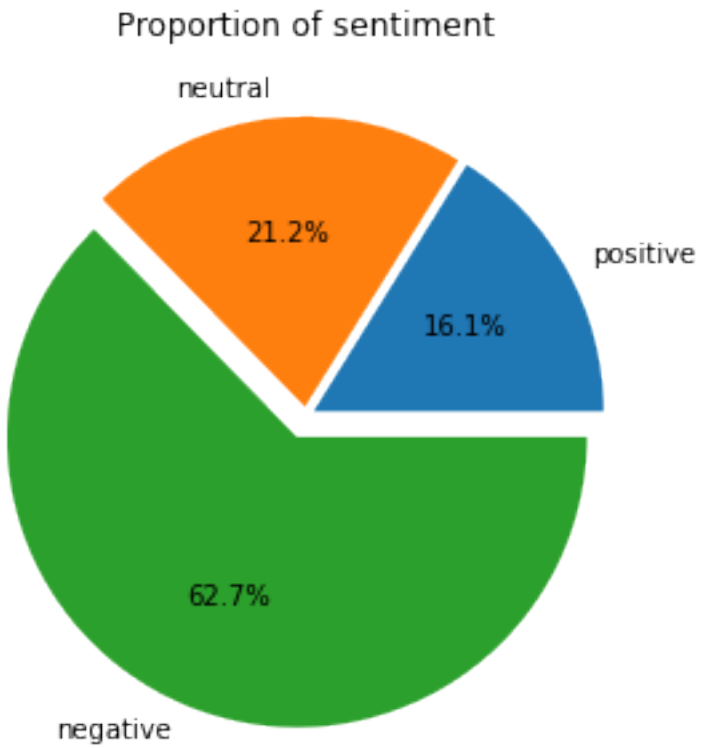
```
             tweet_id airline_sentiment  airline_sentiment_confidence  \
0  570306133677760513           neutral                        1.0000
1  570301130888122368          positive                        0.3486
2  570301083672813571           neutral                        0.6837
3  570301031407624196          negative                        1.0000
4  570300817074462722          negative                        1.0000

  negativereason  negativereason_confidence         airline  \
0            NaN                       NaN  Virgin America
1            NaN                    0.0000  Virgin America
2            NaN                       NaN  Virgin America
3     Bad Flight                    0.7033  Virgin America
4     Can't Tell                    1.0000  Virgin America

  airline_sentiment_gold        name negativereason_gold  retweet_count  \
0                    NaN     cairdin                 NaN              0
1                    NaN    jnardino                 NaN              0
2                    NaN  yvonnalynn                 NaN              0
3                    NaN    jnardino                 NaN              0
4                    NaN    jnardino                 NaN              0

                                                text tweet_coord  \
0                @VirginAmerica What @dhepburn said.         NaN
1  @VirginAmerica plus you've added commercials t…         NaN
2  @VirginAmerica I didn't today… Must mean I n…         NaN
3  @VirginAmerica it's really aggressive to blast…         NaN
4  @VirginAmerica and it's a really big bad thing…         NaN

              tweet_created tweet_location               user_timezone
0  2015-02-24 11:35:52 -0800            NaN  Eastern Time (US & Canada)
1  2015-02-24 11:15:59 -0800            NaN  Pacific Time (US & Canada)
2  2015-02-24 11:15:48 -0800      Lets Play  Central Time (US & Canada)
3  2015-02-24 11:15:36 -0800            NaN  Pacific Time (US & Canada)
4  2015-02-24 11:14:45 -0800            NaN  Pacific Time (US & Canada)
```

None

[ ]:
```
ax = sns.countplot(data = df, y = 'airline',
                   order = df.airline.value_counts().index)
ax.bar_label(ax.containers[0])
ax.set_title('Count per Airlines',)


plt.show()
```



[ ]:
```
x = df.airline_sentiment.value_counts().sort_values()

plt.figure(figsize=(5, 5))
ax = plt.pie(x = x, labels=x.index, autopct = '%1.1f%%',explode = [0.03, 0.03,␣
↪0.08])
plt.title('Proportion of sentiment')

plt.show()
```



[ ]:
```
a = df.groupby(['airline', 'airline_sentiment'])['airline_sentiment'].count().
 ↪unstack()
# a['total'] =  [a.values[x].sum() for x in range(0,6)]
a
```

[ ]:

airline_sentiment	negative	neutral	positive
airline			
American	1960	463	336
Delta	955	723	544
Southwest	1186	664	570
US Airways	2263	381	269
United	2633	697	492
Virgin America	181	171	152

[ ]:
```
fig, axes = plt.subplots(2, 3, figsize = (15, 8))
axes = axes.flatten()
for i, ax in zip(range(0, 6), axes):
    temp = a.iloc[i]
    ax.pie(x = temp, labels = temp.index, autopct = '%1.1f%%', explode = [0.08,␣
 ↪0.03, 0.03])
    ax.set_title(f"{a.index[i]}:{format(a.values[i].sum(),',')}")
```

```
plt.suptitle("Proportion of Sentiment", fontsize = 25)
plt.show()
```