Twitter Sentiment Prediction of US Airlines
Milestone: Data Collection and Processing

Group 12

Student-1 Name: Amitoj Singh Kohli

Student-2 Name: Aryaman Kakad

857-313-0650 (Tel of Student 1)

857-313-5234 (Tel of Student 2)

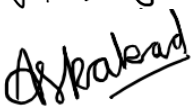
kohli.am@northeastern.edu

kakad.a@northeastern.edu

Percentage of Effort Contributed by Student1: 100%

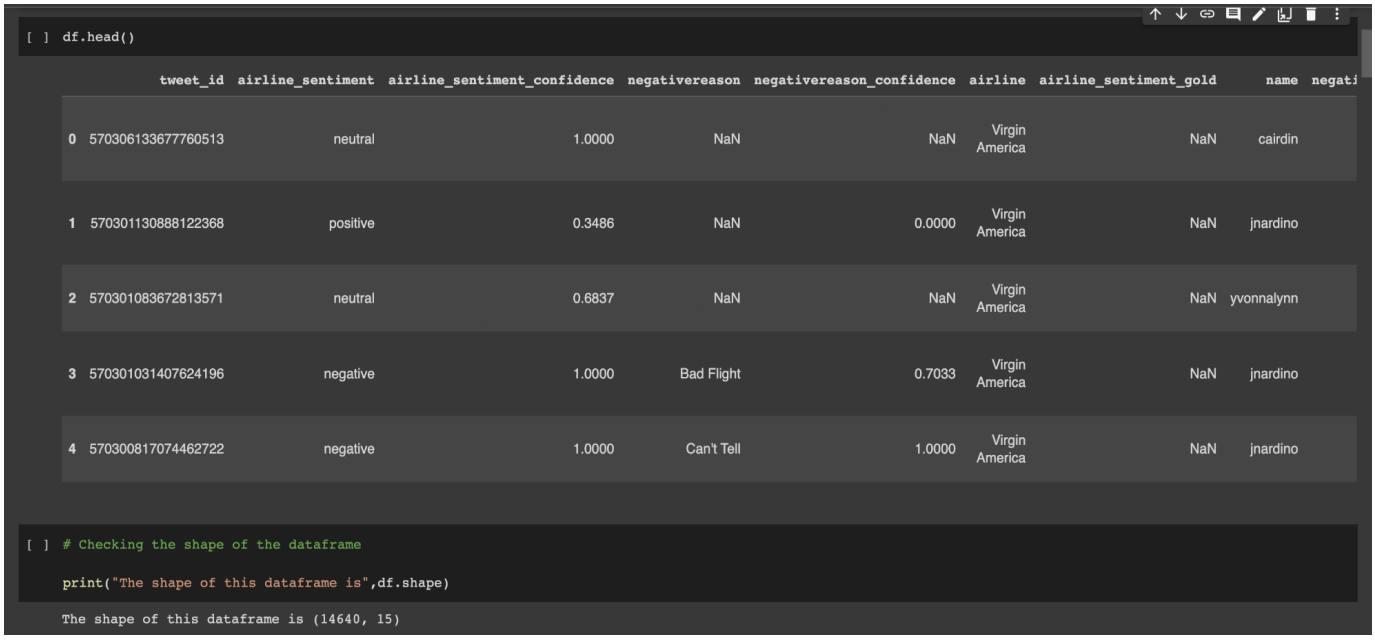
Percentage of Effort Contributed by Student2: 100%

Signature of Student 1: 

Signature of Student 2: 

Submission Date: 04 / 08 / 23

The dataset was collected from Kaggle and it contains tweets of US airlines for 9 days in the year 2015.

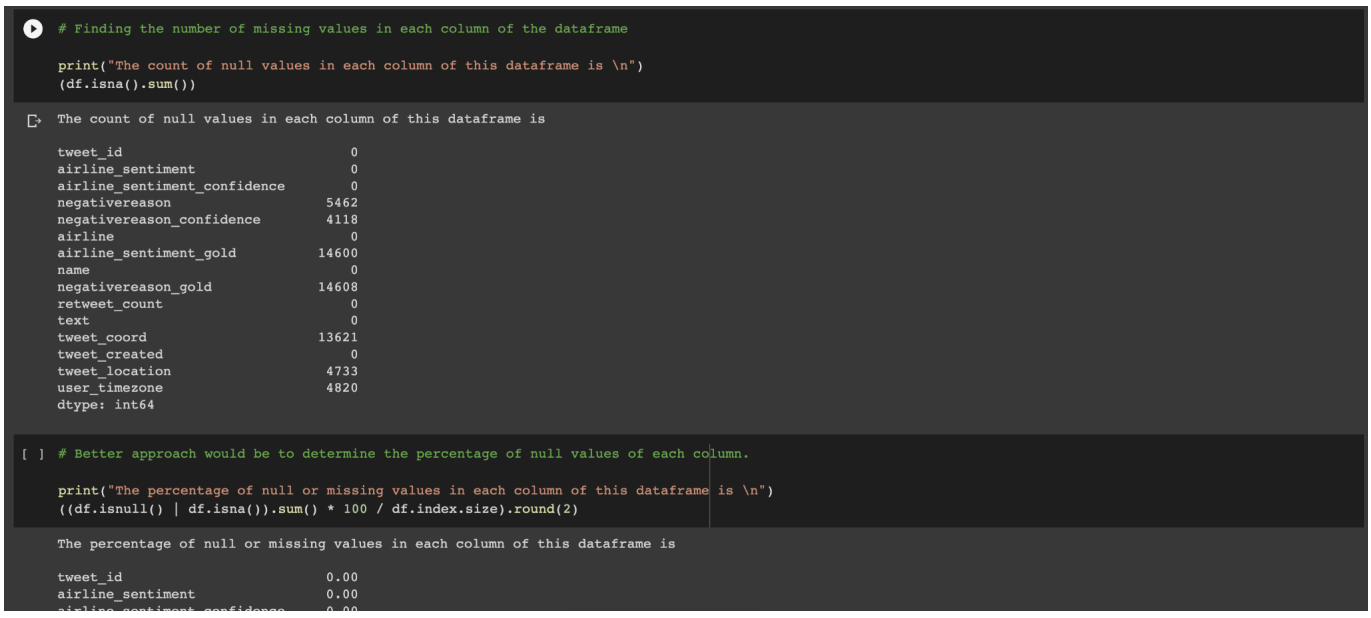


```
[ ] df.head()
```

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

```
[ ] # Checking the shape of the dataframe

    print("The shape of this dataframe is",df.shape)

    The shape of this dataframe is (14640, 15)
```

We calculated the percentage of missing values in the dataset.



```
# Finding the number of missing values in each column of the dataframe

print("The count of null values in each column of this dataframe is \n")
(df.isna().sum())

The count of null values in each column of this dataframe is

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

[ ] # Better approach would be to determine the percentage of null values of each column.

    print("The percentage of null or missing values in each column of this dataframe is \n")
    ((df.isnull() | df.isna()).sum() * 100 / df.index.size).round(2)

    The percentage of null or missing values in each column of this dataframe is

    tweet_id                      0.00
    airline_sentiment             0.00
```

```
# Better approach would be to determine the percentage of null values of each column.

print("The percentage of null or missing values in each column of this dataframe is \n")
((df.isnull() | df.isna()).sum() * 100 / df.index.size).round(2)
```

```
The percentage of null or missing values in each column of this dataframe is

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
```

Insights: The three columns having more than 90% of missing values i.e 'airline_sentiment_gold', 'negativereason_gold', 'tweet_coord' will not contribute to determining the dependent variable 'airline_sentiment' and therefore should be deleted.

```
df.drop(columns=['airline_sentiment_gold', 'negativereason_gold', 'tweet_coord'], axis=1, inplace=True)
```

```
df.head()
```

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	name	retweet_count	text	t
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	cairdin	0	@VirginAmerica What @dhepburn said.	
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	jnardino	0	@VirginAmerica plus you've added commercials t...	
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	yvonnalynn	0	@VirginAmerica I didn't today... Must mean I n...	
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	jnardino	0	@VirginAmerica it's really aggressive to blast...	
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	jnardino	0	@VirginAmerica and it's a really big bad thing...	

```
df.shape
```

```
(14640, 12)
```

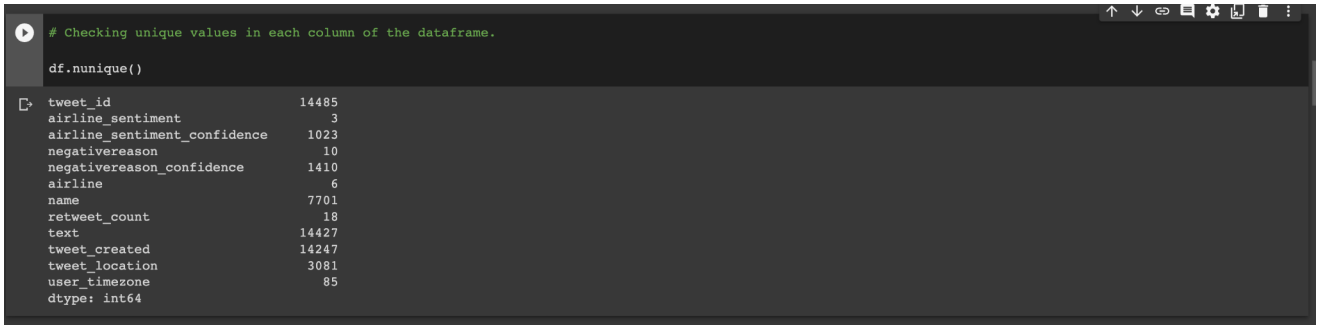
The three columns which had 90% missing values have been dropped from the dataset. Now we check the data types of the variables and the unique values of each of the features.

```
# Checking data type of each column in our dataframe.

df.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  object 
 10  tweet_location                9907 non-null   object 
 11  user_timezone                 9820 non-null   object 
dtypes: float64(2), int64(2), object(8)
memory usage: 1.3+ MB
```

```
[ ] # Checking unique values in each column of the dataframe.

df.nunique()
```

```
tweet_id                      14485
airline_sentiment                 3
airline_sentiment_confidence   1023
negativereason                   10
negativereason_confidence      1410
airline                           6
```



'tweet_created' column has a datatype of object when it contains values regarding the time and date of the tweet created and hence this column has to be converted to a datetime data type with only the date information retained and not the time.

```
[ ] df['tweet_created'] = pd.to_datetime(df['tweet_created'], format='%Y-%m-%d').dt.date

[ ] df['tweet_created'] = pd.to_datetime(df['tweet_created'])

df.head()
```

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	name	retweet_count	text	t
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	cairdin	0	@VirginAmerica What @dhepburn said.	
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	jnardino	0	@VirginAmerica plus you've added commercials t...	
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	yvonnalynn	0	@VirginAmerica I didn't today... Must mean I n...	
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	jnardino	0	@VirginAmerica it's really aggressive to blast...	
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	jnardino	0	@VirginAmerica and it's a really big bad thing...	

Now we will be checking the earliest and latest date of the tweets for the US airlines.

```
df.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
```

```
[ ] # Checking the earliest and latest date of a tweet in our dataframe.

    print(df['tweet_created'].min())
    print(df['tweet_created'].max())
```

```
2015-02-16 00:00:00
2015-02-24 00:00:00
```

We have tweets of the US Airlines from 16th February 2015 to 24th February 2015 i.e. 9 days.

We have tweets from 16th February 2015 to 24th February 2015.

```
[ ] # To confirm our above statement, can find the unique values in the 'tweet_created' column.

    df['tweet_created'].nunique()

    9

    # Finding the number of tweets on each of the nine days.

    tweets_count = df.groupby('tweet_created').size()
    tweets_count

    tweet_created
    2015-02-16       4
    2015-02-17    1408
    2015-02-18    1344
    2015-02-19    1376
    2015-02-20    1500
    2015-02-21    1557
    2015-02-22    3079
    2015-02-23    3028
    2015-02-24    1344
    dtype: int64
```

We also calculated the number of tweets on these 9 days.