29 April 2025

EDS Theory Activity 1:

BY Mahek Chaurasia
Roll No: CS2-63
PRN: 202401040370

You can check the 20 problem statements by clicking on the given link

Google Colab

google.com

https://colab.research.google.com/drive/1jF0OndIQjX3yTm_oQAVtFbXKlB1A3q-d#scrollTo=e5M64-91Xf5N

This is the dataset that i have imported from the kaggle.

```
[5] import kagglehub

    # Correct dataset handle format: {username}/{dataset_name}
    path = kagglehub.dataset_download('crowdflower/twitter-airline-sentiment')

    print("Path to dataset files:", path)
```

```
Path to dataset files: /kaggle/input/twitter-airline-sentiment
```

Twitter US Airline Sentiment Dataset

Using Numpy

1. Load the dataset and print the shape of the data.

```
[ ] import pandas as pd

    df = pd.read_csv('Tweets.csv')

    import numpy as np

    # Get shape using numpy
    shape = np.array(df).shape
    print(f"Dataset shape: {shape}")
```

```
Dataset shape: (14640, 15)
```

2. Calculate the percentage of missing values in each column.

```
[ ]  # Calculate missing values
     missing_percent = (df.isnull().sum().values / len(df)) * 100
     print("Missing Value Percentage:\n", missing_percent)
```

```
Missing Value Percentage:
 [ 0.          0.          0.         37.30874317 28.1284153   0.
 99.72677596  0.         99.78142077  0.          0.         93.03961749
  0.         32.32923497 32.92349727]
```

3. Find the number of unique sentiments in the dataset.

```
[ ]  # Unique sentiments
     sentiments = np.unique(df['airline_sentiment'].values)
     print("Unique sentiments:", sentiments)
     print("Count:", len(sentiments))
```

```
Unique sentiments: ['negative' 'neutral' 'positive']
Count: 3
```

4. Find the tweet with the maximum length (characters).

```
[ ] # Lengths of tweets
    tweet_lengths = np.array([len(str(tweet)) for tweet in df['text']])
    max_length_idx = np.argmax(tweet_lengths)

    print("Tweet with maximum length:")
    print(df['text'].iloc[max_length_idx])
    print(f"Length: {tweet_lengths[max_length_idx]}")
```

```
Tweet with maximum length:
@USAirways Eyyyy! Cancelled Flightlations, Flight Booking Problemss, reFlight Booking Problemss, but y'all got me on the same flight
Length: 186
```

5. Calculate the average tweet length.

```
[ ] avg_length = np.mean(tweet_lengths)
    print(f"Average tweet length: {avg_length:.2f} characters")
```

```
Average tweet length: 103.82 characters
```

6. Count number of tweets per sentiment.

```
[ ] unique_sentiments, counts = np.unique(df['airline_sentiment'], return_counts=True)
    sentiment_counts = dict(zip(unique_sentiments, counts))
    print("Tweet counts per sentiment:", sentiment_counts)
```

```
Tweet counts per sentiment: {'negative': np.int64(9178), 'neutral': np.int64(3099), 'positive': np.int64(2363)}
```

7. Find standard deviation of tweet lengths.

```
[ ] std_dev_length = np.std(tweet_lengths)
    print(f"Standard Deviation of tweet lengths: {std_dev_length:.2f}")
```

```
Standard Deviation of tweet lengths: 36.28
```

8. Find tweets posted by a specific airline (e.g., "United").

```
[ ] # Filter by airline
    airline_filter = (df['airline'].values == 'United')
    united_tweets = df['text'].values[airline_filter]

    print(f"Number of tweets for 'United': {len(united_tweets)}")
    print("Sample tweet:", united_tweets[0])
```

```
Number of tweets for 'United': 3822
Sample tweet: @united thanks
```

9. Encode sentiments numerically: (negative → 0, neutral → 1, positive → 2).

```
[ ] sentiment_map = {'negative': 0, 'neutral': 1, 'positive': 2}
    encoded_sentiments = np.array([sentiment_map[sent] for sent in df['airline_sentiment'].values])

    print("Encoded Sentiments (first 10):", encoded_sentiments[:10])
```

```
Encoded Sentiments (first 10): [1 2 1 0 0 0 2 1 2 2]
```

10. Find correlation between tweet length and sentiment label.

```
[ ] # Use encoded_sentiments from previous step
    correlation = np.corrcoef(tweet_lengths, encoded_sentiments)[0, 1]
    print(f"Correlation between tweet length and sentiment: {correlation:.4f}")
```

```
Correlation between tweet length and sentiment: -0.3358
```

Using Pandas

11.Load the dataset and display the first 5 rows.

[30]
```python
import pandas as pd

# Load dataset
df = pd.read_csv('/Tweets.csv')

# Display first 5 rows
print(df.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  \
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
```

✓ 0s completed at 11:23 PM

```
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)
```

✓ 0s completed at 11:23 PM

12. Show the column names of the dataset.

[31]
```python
# Column names
print(df.columns.tolist())
```

```
['tweet_id', 'airline_sentiment', 'airline_sentiment_confidence', 'negativereason', 'negativereason_confidence', 'airline', 'airline_sentiment_gold', 'name', 'ne
```

13.Get a summary (count, mean, std, min, max) of numerical columns.

[32]
```
# Summary statistics
print(df.describe())
```

```
           tweet_id  airline_sentiment_confidence  negativereason_confidence  \
count  1.464000e+04                  14640.000000                10522.000000
mean   5.692184e+17                      0.900169                    0.638298
std    7.791112e+14                      0.162830                    0.330440
min    5.675883e+17                      0.335000                    0.000000
25%    5.685592e+17                      0.692300                    0.360600
50%    5.694779e+17                      1.000000                    0.670600
75%    5.698905e+17                      1.000000                    1.000000
max    5.703106e+17                      1.000000                    1.000000

       retweet_count
count   14640.000000
mean        0.082650
std         0.745778
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max        44.000000
```

14. Find how many tweets are positive, neutral, or negative.

[33]
```
# Sentiment counts
sentiment_counts = df['airline_sentiment'].value_counts()
print(sentiment_counts)
```

```
airline_sentiment
negative    9178
neutral     3099
positive    2363
Name: count, dtype: int64
```

15. Find the top 5 airlines based on the number of tweets.

[34]
```python
# Airline counts
top_airlines = df['airline'].value_counts().head(5)
print(top_airlines)
```

```
airline
United        3822
US Airways    2913
American      2759
Southwest     2420
Delta         2222
Name: count, dtype: int64
```

16. Extract only the tweets that are labeled as "positive".

[35]
```
# Positive tweets
positive_tweets = df[df['airline_sentiment'] == 'positive']
print(positive_tweets[['airline', 'text']].head())
```

```
           airline                                               text
1   Virgin America  @VirginAmerica plus you've added commercials t...
6   Virgin America  @VirginAmerica yes, nearly every time I fly VX...
8   Virgin America    @virginamerica Well, I didn't…but NOW I DO! :-D
9   Virgin America  @VirginAmerica it was amazing, and arrived an ...
11  Virgin America  @VirginAmerica I &lt;3 pretty graphics. so muc...
```

17. Create a new column with the length of each tweet.

[38]
```
# New column 'tweet_length'
df['tweet_length'] = df['text'].apply(lambda x: len(str(x)))

print(df[['text', 'tweet_length']].head())
```

```
                                                text  tweet_length
0                @VirginAmerica What @dhepburn said.            35
1  @VirginAmerica plus you've added commercials t...            72
2  @VirginAmerica I didn't today... Must mean I n...            71
3  @VirginAmerica it's really aggressive to blast...           126
4  @VirginAmerica and it's a really big bad thing...            55
```

18. Find the airline with the most negative tweets.

[39]
```python
# Filter negative tweets
negative_tweets = df[df['airline_sentiment'] == 'negative']

# Count by airline
most_negative_airline = negative_tweets['airline'].value_counts().idxmax()
print(f"Airline with most negative tweets: {most_negative_airline}")
```

```
Airline with most negative tweets: United
```

19. Group the tweets by airline and find the average tweet length for each airline.

[40]
```python
# Group by airline and mean tweet length
avg_length_per_airline = df.groupby('airline')['tweet_length'].mean()

print(avg_length_per_airline)
```

```
airline
American          108.630301
Delta              92.501800
Southwest         103.212810
US Airways        109.261586
United            103.817373
Virgin America     98.930556
Name: tweet_length, dtype: float64
```

20. Find all tweets that mention "customer service" (case insensitive).

[41]

```python
# Search 'customer service' in text
customer_service_tweets = df[df['text'].str.contains('customer service', case=False, na=False)]

print(customer_service_tweets[['airline', 'text']].head())
```

```
            airline                                                text
71   Virgin America  @VirginAmerica I emailed your customer service...
78   Virgin America  @VirginAmerica what is going on with customer ...
80   Virgin America  @VirginAmerica why can't you supp the biz trav...
122  Virgin America  @VirginAmerica I like the customer service but...
124  Virgin America  @VirginAmerica you have the absolute best team...
```

Thank you!