

## Data Science Assignment 4

Student Name: Minahil Irfan

### Overview

This assignment focused on applying statistical concepts to analyze relationships between features in a real-world dataset. Using descriptive statistics and correlation analysis, the goal was to identify which three features are most closely related to the target sentiment variable.

### Statistical Concepts Applied:

- **Mean, Median, Mode:** Describe central tendency.
- **Variance & Standard Deviation:** Indicate data spread.
- **Correlation:** Measures the strength and direction of relationships between variables.

### Task Overview

The objective was to calculate descriptive statistics and determine the top three features most correlated with the sentiment target variable in the **Twitter Sentiment Dataset**.

### Activity

#### Step 1: Load Dataset

- Loaded `twitter_sentiment_cleaned.csv` into Jupyter Notebook using Pandas.
- Renamed columns for clarity and mapped sentiment labels (0 = Negative, 4 = Positive).

#### Step 2: Feature Engineering

- Created numeric features for analysis:
  - `tweet_length`: number of characters.
  - `word_count`: number of words.
  - `avg_word_length`: average word size.

#### Step 3: Descriptive Statistics

- Computed **mean, median, mode**, and **variance** for all numeric features.
- Summarized dataset behavior and spread.

#### Step 4: Correlation Analysis

- Generated a correlation matrix using `df.corr()`.
- Visualized results with a **Seaborn heatmap** to show feature relationships.

#### Step 5: Results Interpretation

Correlation with target sentiment variable:

```
===== CORRELATION WITH TARGET =====
target          1.000000
avg_word_length 0.157307
tweet_length    -0.005860
word_count      -0.058335
```

### Top 3 features most related to target:

1. Average Word Length 0.1573
2. Tweet Length -0.0058
3. Word Count -0.0583

### Insights:

- Average word length shows the strongest (though weak) positive correlation with positive sentiment.
- Tweet length and word count have minor negative correlations, indicating longer tweets don't necessarily imply positivity.

### Summary

This analysis demonstrated how basic statistical and correlation methods reveal relationships within text-based data. Although correlations were weak, the process established a foundation for identifying influential features for predictive sentiment modeling.