**Theory Assignment No 1**

Name: Dhammanand Babanrao Kawle

PRN: 202401050045

Roll No :CC-21

Topic: witter US Airline Sentiment

1. Find the number of tweets for each sentiment (positive, neutral, negative).

df['airline\_sentiment'].value\_counts()

2. Calculate the percentage of each sentiment.

df['airline\_sentiment'].value\_counts(normalize=True) \* 100

3. Find out which airline has the most negative tweets.

df[df['airline\_sentiment'] == 'negative']['airline'].value\_counts()

4. Find the average length of tweets (in characters).

df['tweet\_length'] = df['text'].apply(len)

df['tweet\_length'].mean()

5. Find the shortest and longest tweets.

shortest\_tweet = df.loc[df['tweet\_length'].idxmin()]

longest\_tweet = df.loc[df['tweet\_length'].idxmax()]

6. How many tweets are from each user location? (Top 10 locations)

df['user\_timezone'].value\_counts().head(10)

7. Find the ratio of negative to positive tweets.

neg = df['airline\_sentiment'].value\_counts()['negative']

pos = df['airline\_sentiment'].value\_counts()['positive']

neg\_pos\_ratio = neg / pos

neg\_pos\_ratio

8. Find the average confidence for each sentiment.

df.groupby('airline\_sentiment')['airline\_sentiment\_confidence'].mean()

9. Filter tweets with sentiment confidence below 50%.

low\_confidence\_tweets = df[df['airline\_sentiment\_confidence'] < 0.5]

10. Find the top 5 users who tweeted the most.

df['name'].value\_counts().head(5)

11. Get the count of tweets containing the word 'help'.

help\_tweets = df['text'].str.contains('help', case=False).sum()

12. Get the number of tweets for each airline.

df['airline'].value\_counts()

13. Which airline has the highest proportion of positive tweets?

airline\_positive = df[df['airline\_sentiment'] == 'positive']['airline'].value\_counts()

airline\_total = df['airline'].value\_counts()

(airline\_positive / airline\_total).sort\_values(ascending=False)

14. Find missing values in each column.

df.isnull().sum()

15. Drop rows where tweet text is missing.

df = df.dropna(subset=['text'])

16. Find correlation between tweet length and sentiment confidence.

df['tweet\_length'].corr(df['airline\_sentiment\_confidence'])

17. Find airlines with an average tweet length greater than 140 characters.

df.groupby('airline')['tweet\_length'].mean()[df.groupby('airline')['tweet\_length'].mean() > 140]

18. Classify tweets into 'short' (<100), 'medium' (100-200), 'long' (>200) based on length.

df['length\_category'] = pd.cut(df['tweet\_length'], bins=[0, 100, 200, np.inf], labels=['short', 'medium', 'long'])

df['length\_category'].value\_counts()

19. Create a pivot table of airline vs sentiment counts.

pd.pivot\_table(df, index='airline', columns='airline\_sentiment', aggfunc='size', fill\_value=0)

20. Identify airlines with the most frequent 'late flight' complaints (tweets containing "late" or "delay").

late\_tweets = df[df['text'].str.contains('late|delay', case=False, na=False)]

late\_tweets['airline'].value\_counts()