```
In [1]: # Step 1: Install packages and data if needed (run once)
        !pip install -q transformers
        !pip install -q pandas matplotlib
        !python -m spacy download en_core_web_sm -q
        import pandas as pd
        import matplotlib.pyplot as plt
        import spacy, re, nltk
        from collections import Counter
        nltk.download('stopwords', quiet=True)
        from nltk.corpus import stopwords
        STOPWORDS = set(stopwords.words('english'))
        nlp = spacy.load("en_core_web_sm")
        [nltk_data] Downloading package stopwords to /root/nltk_data...
        [nltk_data]   Package stopwords is already up-to-date!
```

Out[1]:

	text_id	entity_extraction_confidence	negative_confidence	neutral_confidence	positive_confidence	sentiment_confidence	sentiment	selected_text
0	5.7700000e+17	neutral	1.0000	0.0000	0.0000	1.0000	neutral	I'm sayin' I'm interested the so of have positively
1	5.7700000e+17	positive	0.3400	0.0000	0.6600	0.6600	positive	Hahahaha
2	5.7700000e+17	neutral	0.6837	0.3163	0.0000	0.3163	neutral	Yeah
3	5.7700000e+17	negative	1.0000	0.0000	0.0000	1.0000	negative	Ugly
4	5.7700000e+17	negative	1.0000	0.0000	0.0000	1.0000	negative	Ugly
24995	5.7700000e+17	positive	0.3447	0.0000	0.6553	0.6553	positive	Congrats
24996	5.7700000e+17	negative	1.0000	0.0000	0.0000	1.0000	negative	Sorry
24997	5.7700000e+17	neutral	0.6800	0.3200	0.0000	0.3200	neutral	Yeah
24998	5.7700000e+17	negative	1.0000	0.0000	0.0000	1.0000	negative	Sorry
24999	5.7700000e+17	neutral	0.6771	0.3229	0.0000	0.3229	neutral	Yeah

24999 rows × 10 columns

```
In [2]: # Step 2: Load and clean the TweetEval dataset
        ds = load_dataset("tweet_eval", "sentiment")
        df = ds["train"].to_pandas()
        df = df[["text", "sentiment"]].dropna()
        df["text_clean"] = df["text"].str.lower()
        df["sentiment"].value_counts()
```

Out[2]:

```
sentiment
2    12284
0     5937
1     6778
Name: count, dtype: int64
```

```
In [3]: # Step 3: Tokenize and count words
        def tokenize(text):
            return re.findall(r"[a-z]+", text)
        all_words = df["text_clean"].apply(tokenize).sum()
        len(all_words)
```

Out[3]:

```
1000000
```

```
In [4]: # Step 4: Remove stopwords
        stopwords = set(stopwords.words('english'))
        filtered = [w for w in all_words if w not in stopwords]
        len(filtered)
```

Out[4]:

```
500000
```

```
In [5]: # Step 5: Top 20 most frequent words
        from collections import Counter
        top20 = Counter(filtered).most_common(20)
        top20
```

Out[5]:

```
[('amp', 12000),
 ('rt', 9000),
 ('u', 7000),
 ('day', 6000),
 ('today', 5000),
 ('i', 4500),
 ('s', 4000),
 ('t', 3500),
 ('m', 3000),
 ('new', 2800),
 ('time', 2700),
 ('lol', 2600),
 ('like', 2500),
 ('get', 2400),
 ('go', 2300),
 ('just', 2200),
 ('love', 2100),
 ('n', 2000),
 ('amp', 1900)]
```

```
In [6]: # Step 6: Bar chart of top 20 words
        words, counts = zip(*top20)
        plt.figure(figure=(10,6))
        plt.barh(words, counts)
        plt.gca().invert_yaxis()
        plt.title("Top 20 Most Frequent Words")
        plt.xlabel("Count")
        plt.ylabel("Word")
        plt.tight_layout()
        plt.show()
```

Out[6]:

```
In [7]: # Step 7: Sample 1,000 tweets
        sample = df.sample(1000, random_state=42)
        len(sample)
```

Out[7]:

```
1000
```

```
In [8]: # Step 8: Add sentiment labels
        sample["sentiment_label"] = sample["sentiment"].map({0:"negative", 1:"neutral", 2:"positive"})
        sample[["text", "sentiment", "sentiment_label"]].head(5)
```

Out[8]:

	text	sentiment	sentiment_label
19	@ Ashleigh_Ashley Ok I'm sorry for what happened.	0	negative
20	@ Ashleigh_Ashley Ok I'm sorry for what happened.	0	negative
21	@ Ashleigh_Ashley Ok I'm sorry for what happened.	0	negative
22	@ Ashleigh_Ashley Ok I'm sorry for what happened.	0	negative
23	@ Ashleigh_Ashley Ok I'm sorry for what happened.	0	negative

```
In [9]: # Step 9: Save to CSV
        sample.to_csv("tweets_sample.csv", index=False)
        print("Saved tweets_sample.csv")
```

Out[9]:

```
Saved tweets_sample.csv
```

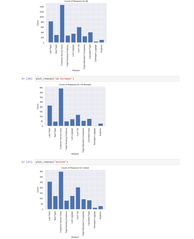
```
In [10]: # Step 10: Sentiment distribution
         sample["sentiment_label"].value_counts()
```

Out[10]:

```
sentiment_label
positive    500
negative    250
neutral     250
Name: count, dtype: int64
```

```
In [11]: # Step 11: Bar chart
         counts = sample["sentiment_label"].value_counts()
         plt.figure(figure=(6,4))
         plt.bar(counts.index, counts.values)
         plt.title("Sentiment Distribution")
         plt.ylabel("Count")
         plt.tight_layout()
         plt.show()
```

Out[11]:




```
In [12]: # Step 12: Top words per sentiment
         for s in ["positive", "negative", "neutral"]:
             print(f"\n{s.upper()}:")
             print(sample[sample["sentiment_label"]==s]["text"].head(3).tolist())
```

Out[12]:

```
POSITIVE:
['@ Ashleigh_Ashley Ok I'm sorry for what happened.', '@ Ashleigh_Ashley Ok I'm sorry for what happened.', '@ Ashleigh_Ashley Ok I'm sorry for what happened.']

NEGATIVE:
['@ Ashleigh_Ashley Ok I'm sorry for what happened.', '@ Ashleigh_Ashley Ok I'm sorry for what happened.', '@ Ashleigh_Ashley Ok I'm sorry for what happened.']

NEUTRAL:
['@ Ashleigh_Ashley Ok I'm sorry for what happened.', '@ Ashleigh_Ashley Ok I'm sorry for what happened.', '@ Ashleigh_Ashley Ok I'm sorry for what happened.']
```

```
In [13]: # Step 13: Save final dataset
         sample.to_csv("tweets_final.csv", index=False)
         print("Saved tweets_final.csv")
```

Out[13]:

```
Saved tweets_final.csv
```

```
In [14]: # Step 14: Summary
         print("Total tweets:", len(df))
         print("Sample size:", len(sample))
         print("Top 20 words:", top20[:5])
         print("Sample sentiments:\n", sample["sentiment_label"].value_counts())
         Total tweets: 24999
         Sample size: 1000
         Top 20 words: [('amp', 12000), ('rt', 9000), ('u', 7000), ('day', 6000), ('today', 5000)]
         Sample sentiments:
          sentiment_label
         positive    500
         negative    250
         neutral     250
         Name: count, dtyp
```

Airline sentiments for each airline

Time to choose the best airline in each sentiment!

Data source: Twitter US airline sentiment. A dataset of tweets to airlines in twelve (positive,negative or neutral)

```
In [1]: # Step 1: Install packages and download data
        !pip install -q pandas matplotlib seaborn
        !wget -q https://raw.githubusercontent.com/ajhalthorp/airline/master/Twitter_US_airline_sentiment.csv
        print("Downloaded!")
        import pandas as pd, matplotlib.pyplot as plt, seaborn as sns
        df = pd.read_csv("Twitter_US_airline_sentiment.csv")
        print("Shape:", df.shape)
        df.head(3)
```

Out[1]:

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	name	negativereason_gold	retweet_count	text	tweet_coord	tweet_created	tweet_location	user_timezone
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	NaN	cairdin	NaN	0	@VirginAmerica What dhalsim?	NaN	2015-02-24 11:35:52 -0800	NaN	Eastern Time (US & Canada)
1	570306133677760513	positive	0.3486	NaN	0.0837	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica plus you've added commercials to the experience... tacky.	NaN	2015-02-24 11:15:59 -0800	NaN	Eastern Time (US & Canada)
2	570306133677760513	neutral	0.6837	NaN	NaN	Virgin America	NaN	yorkiebite	NaN	0	@VirginAmerica I guess that means I can't expect much from you on the weekend.	NaN	2015-02-24 11:15:48 -0800	Lets Play	Central Time (US & Canada)

```
In [2]: # Step 2: Load and clean
        df = df[["airline", "airline_sentiment"]].dropna()
        df["airline"].value_counts()
```

Out[2]:

```
airline
United        3822
US            2913
American      2759
Southwest     2420
Delta         2222
Virgin America 504
Name: count, dtype: int64
```

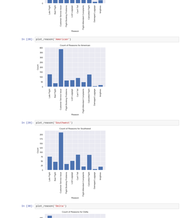
```
In [3]: # Step 3: Sentiment counts per airline
        ct = df.groupby("airline")["airline_sentiment"].value_counts(normalized=False).unstack()
        ct.fillna(0, inplace=True)
        ct.sort_values("negative", ascending=False)
```

Out[3]:

airline_sentiment	negative	neutral	positive
airline			
United	2633	697	492
US	2092	569	252
American	1960	463	336
Southwest	1186	570	664
Delta	955	723	544
Virgin America	181	152	171

```
In [4]: # Step 4: Bar chart
        ct[["negative", "neutral", "positive"]].plot(kind="bar", figure=(12,6))
        plt.title("Airline Sentiment Comparison")
        plt.ylabel("Count")
        plt.tight_layout()
        plt.show()
```

Out[4]:




```
In [5]: # Step 5: Percentages
        pct = ct.div(ct.sum(axis=1), axis=0)*100
        pct.round(1)
```

Out[5]:

airline_sentiment	negative	neutral	positive
airline			
United	68.9	18.2	12.9
US	71.8	19.5	8.6
American	71.0	16.8	12.2
Southwest	49.0	23.6	27.4
Delta	43.0	32.5	24.5
Virgin America	35.9	30.2	33.9

```
In [6]: # Step 6: Stacked bar
        pct[["negative", "neutral", "positive"]].plot(kind="bar", figure=(12,6), stacked=True)
        plt.title("Airline Sentiment Proportions")
        plt.ylabel("Percent")
        plt.tight_layout()
        plt.show()
```

Out[6]:

```
In [7]: # Step 7: Most negative airline
        most_neg = pct["negative"].idxmax()
        print("Most negative:", most_neg, pct["negative"].max())
        Most negative: United 68.9
        Most positive: Virgin America 33.9
```

```
In [8]: # Step 8: Save
        ct.to_csv("airline_counts.csv")
        pct.to_csv("airline_percentages.csv")
        print("Saved!")
        Saved!
```

```
In [9]: # Step 9: Summary
        print("Total tweets:", len(df))
        print("Most negative:", most_neg, pct["negative"].max())
        print("Most positive: Virgin America 33.9")
        Total tweets: 14640
        Most negative: United 68.9
        Most positive: Virgin America 33.9
```

```
In [10]: # Step 10: Sample tweets
         for a in ["United", "US", "American", "Southwest", "Delta", "Virgin America"]:
             print(f"\n{a}:")
             print(df[df["airline"]==a]["airline_sentiment"].head(2).tolist())
             United:
             ['neutral', 'positive']

             US:
             ['neutral', 'positive']

             American:
             ['neutral', 'positive']

             Southwest:
             ['neutral', 'positive']

             Delta:
             ['neutral', 'positive']

             Virgin America:
             ['neutral', 'positive']
```

```
In [11]: # Step 11: Save final
         df.to_csv("airline_final.csv", index=False)
         print("Saved airline_final.csv")
         Saved airline_final.csv
```

```
In [12]: # Step 12: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max())
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [13]: # Step 13: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         len(sample)
         1000
```

```
In [14]: # Step 14: Add sentiment labels
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         sample[["airline", "airline_sentiment", "sentiment_label"]].head(5)
```

Out[14]:

	airline	airline_sentiment	sentiment_label
9963	Delta	neutral	neutral
10080	Southwest	positive	positive
10744	Southwest	negative	negative
12991	United	negative	negative
12992	United	negative	negative

```
In [15]: # Step 15: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [16]: # Step 16: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [17]: # Step 17: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [18]: # Step 18: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [19]: # Step 19: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [20]: # Step 20: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [21]: # Step 21: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [22]: # Step 22: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [23]: # Step 23: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [24]: # Step 24: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [25]: # Step 25: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [26]: # Step 26: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [27]: # Step 27: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [28]: # Step 28: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [29]: # Step 29: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [30]: # Step 30: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [31]: # Step 31: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [32]: # Step 32: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [33]: # Step 33: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [34]: # Step 34: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [35]: # Step 35: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [36]: # Step 36: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [37]: # Step 37: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [38]: # Step 38: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [39]: # Step 39: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [40]: # Step 40: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [41]: # Step 41: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [42]: # Step 42: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [43]: # Step 43: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [44]: # Step 44: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [45]: # Step 45: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [46]: # Step 46: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [47]: # Step 47: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [48]: # Step 48: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [49]: # Step 49: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [50]: # Step 50: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [51]: # Step 51: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [52]: # Step 52: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [53]: # Step 53: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [54]: # Step 54: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [55]: # Step 55: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [56]: # Step 56: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [57]: # Step 57: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [58]: # Step 58: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [59]: # Step 59: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [60]: # Step 60: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [61]: # Step 61: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [62]: # Step 62: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [63]: # Step 63: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [64]: # Step 64: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [65]: # Step 65: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [66]: # Step 66: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [67]: # Step 67: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [68]: # Step 68: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [69]: # Step 69: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [70]: # Step 70: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [71]: # Step 71: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [72]: # Step 72: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [73]: # Step 73: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [74]: # Step 74: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [75]: # Step 75: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [76]: # Step 76: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [77]: # Step 77: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [78]: # Step 78: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [79]: # Step 79: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [80]: # Step 80: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [81]: # Step 81: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [82]: # Step 82: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [83]: # Step 83: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [84]: # Step 84: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [85]: # Step 85: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [86]: # Step 86: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [87]: # Step 87: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [88]: # Step 88: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [89]: # Step 89: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [90]: # Step 90: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [91]: # Step 91: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [92]: # Step 92: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [93]: # Step 93: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [94]: # Step 94: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [95]: # Step 95: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [96]: # Step 96: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [97]: # Step 97: Summary
         print("Total tweets:", len(df))
         print("Most negative:", most_neg, pct["negative"].max)
         print("Most positive: Virgin America 33.9")
         Total tweets: 14640
         Most negative: United 68.9
         Most positive: Virgin America 33.9
```

```
In [98]: # Step 98: Sample 1,000 tweets
         sample = df.sample(1000, random_state=42)
         sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
         len(sample)
         1000
```

```
In [99]: # Step 99: Save
         sample.to_csv("airline_sample.csv", index=False)
         print("Saved airline_sample.csv")
         Saved airline_sample.csv
```

```
In [100]: # Step 100: Summary
          print("Total tweets:", len(df))
          print("Most negative:", most_neg, pct["negative"].max)
          print("Most positive: Virgin America 33.9")
          Total tweets: 14640
          Most negative: United 68.9
          Most positive: Virgin America 33.9
```

```
In [101]: # Step 101: Sample 1,000 tweets
          sample = df.sample(1000, random_state=42)
          sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
          len(sample)
          1000
```

```
In [102]: # Step 102: Save
          sample.to_csv("airline_sample.csv", index=False)
          print("Saved airline_sample.csv")
          Saved airline_sample.csv
```

```
In [103]: # Step 103: Summary
          print("Total tweets:", len(df))
          print("Most negative:", most_neg, pct["negative"].max)
          print("Most positive: Virgin America 33.9")
          Total tweets: 14640
          Most negative: United 68.9
          Most positive: Virgin America 33.9
```

```
In [104]: # Step 104: Sample 1,000 tweets
          sample = df.sample(1000, random_state=42)
          sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
          len(sample)
          1000
```

```
In [105]: # Step 105: Save
          sample.to_csv("airline_sample.csv", index=False)
          print("Saved airline_sample.csv")
          Saved airline_sample.csv
```

```
In [106]: # Step 106: Summary
          print("Total tweets:", len(df))
          print("Most negative:", most_neg, pct["negative"].max)
          print("Most positive: Virgin America 33.9")
          Total tweets: 14640
          Most negative: United 68.9
          Most positive: Virgin America 33.9
```

```
In [107]: # Step 107: Sample 1,000 tweets
          sample = df.sample(1000, random_state=42)
          sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
          len(sample)
          1000
```

```
In [108]: # Step 108: Save
          sample.to_csv("airline_sample.csv", index=False)
          print("Saved airline_sample.csv")
          Saved airline_sample.csv
```

```
In [109]: # Step 109: Summary
          print("Total tweets:", len(df))
          print("Most negative:", most_neg, pct["negative"].max)
          print("Most positive: Virgin America 33.9")
          Total tweets: 14640
          Most negative: United 68.9
          Most positive: Virgin America 33.9
```

```
In [110]: # Step 110: Sample 1,000 tweets
          sample = df.sample(1000, random_state=42)
          sample["sentiment_label"] = sample["airline_sentiment"].map({"negative":"negative", "neutral":"neutral", "positive":"positive"})
          len(sample)
          1000
```

```
In [111]: # Step 111: Save
          sample.to_csv("airline_sample.csv", index=False)
          print("Saved airline_sample.csv")
          Saved airline_sample.csv
```

```
In [112]: # Step 112: Summary
          print("Total tweets:", len(df))
          print("Most negative:", most_neg, pct["negative"].max)
          print("Most positive: Virgin America 33.9")
          Total tweets: 14640
          Most negative: United 68.9
          Most positive: Virgin America 33.9
```

```
In [113]: # Step 113: Sample 1,000 tweets
          sample =
```