Theory Activity No. 1

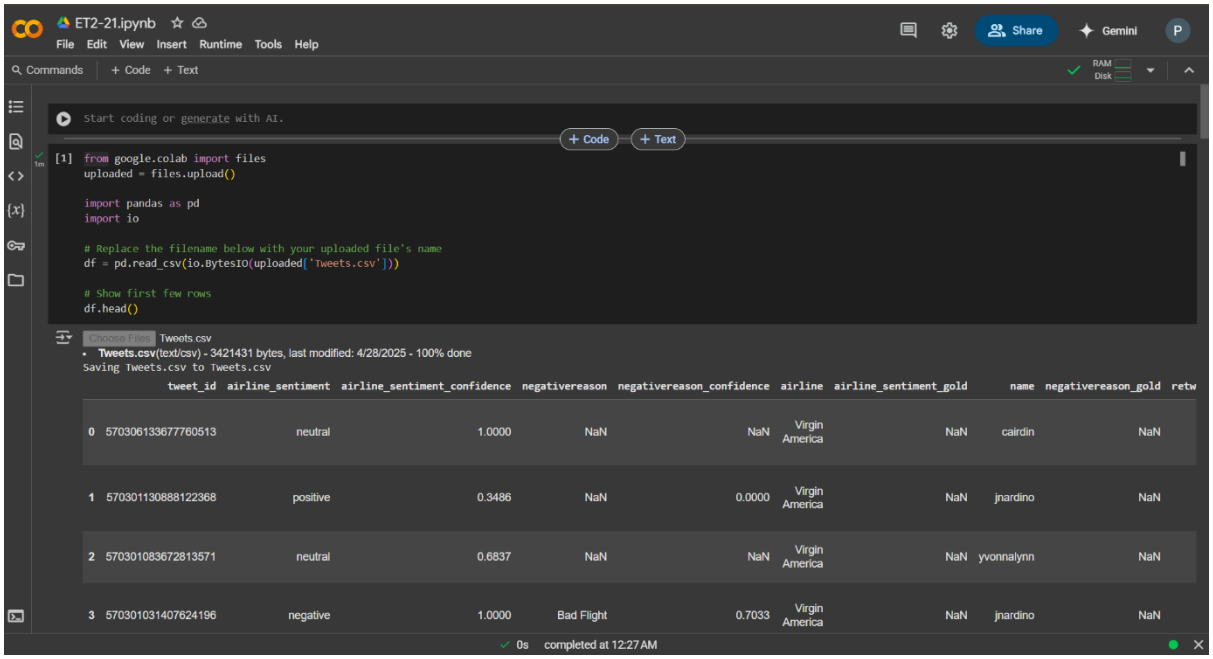
Name: Priti Sutar

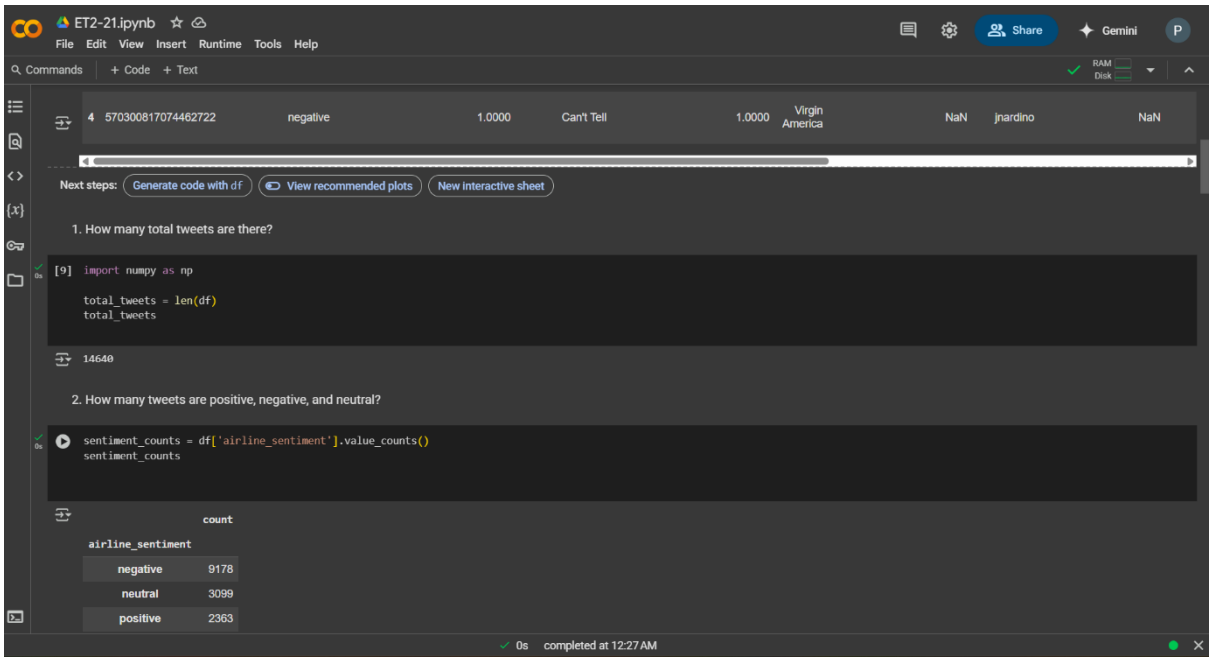
Batch: ET22

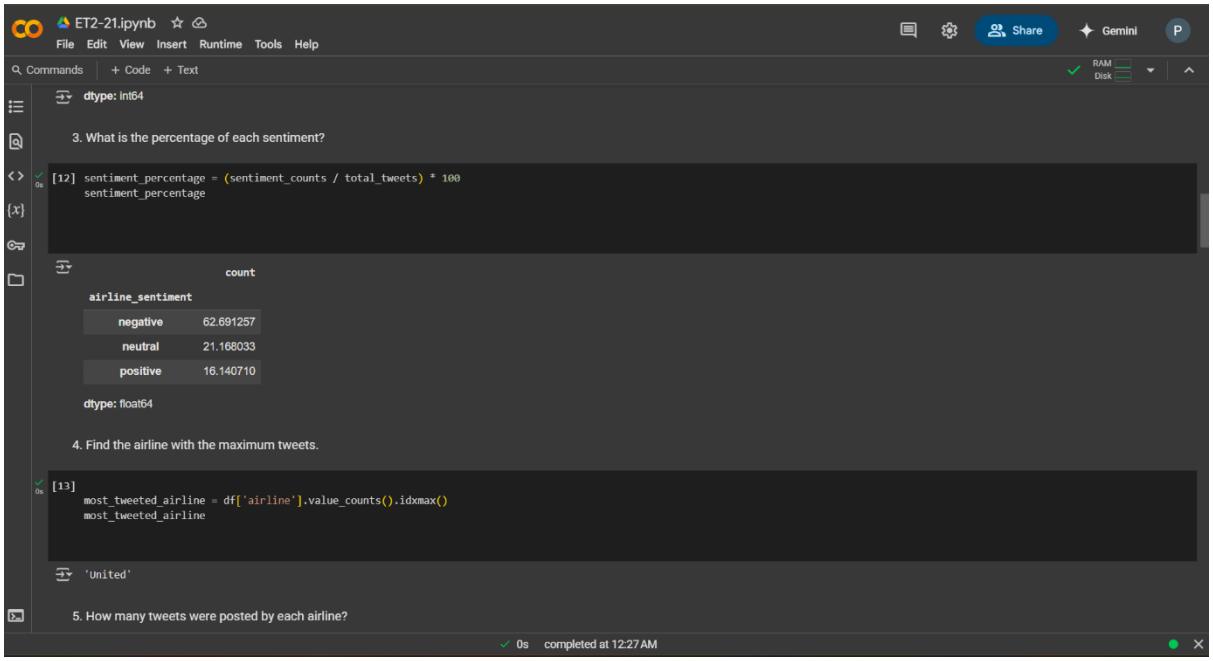
Roll No: ET2-21

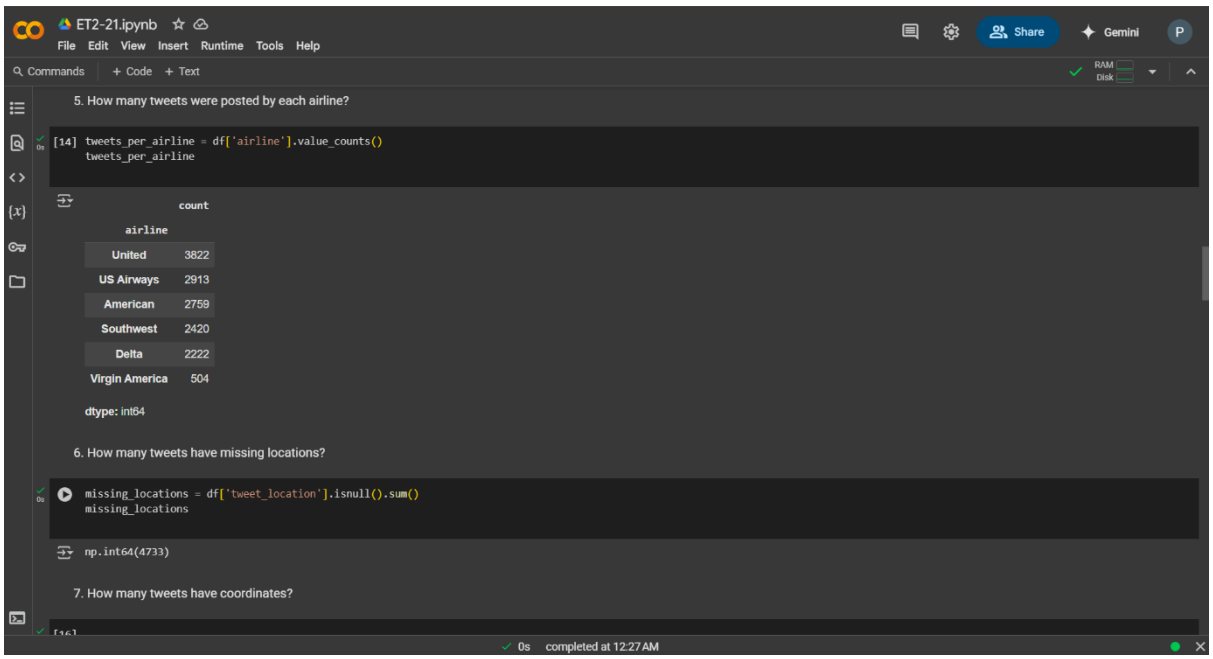
PRN: 202401070075

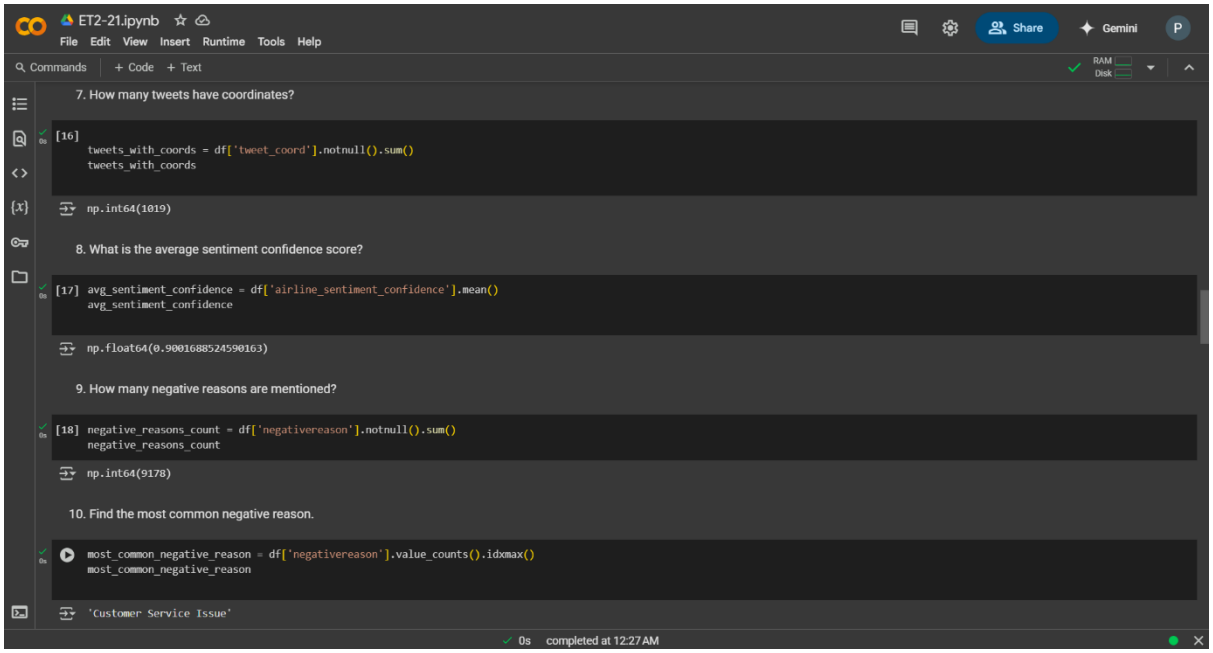
Colab Link: https://colab.research.google.com/drive/1mxGb7-oGpNmOKFHziJTlPZDZzhnmwGSR?usp=sharing

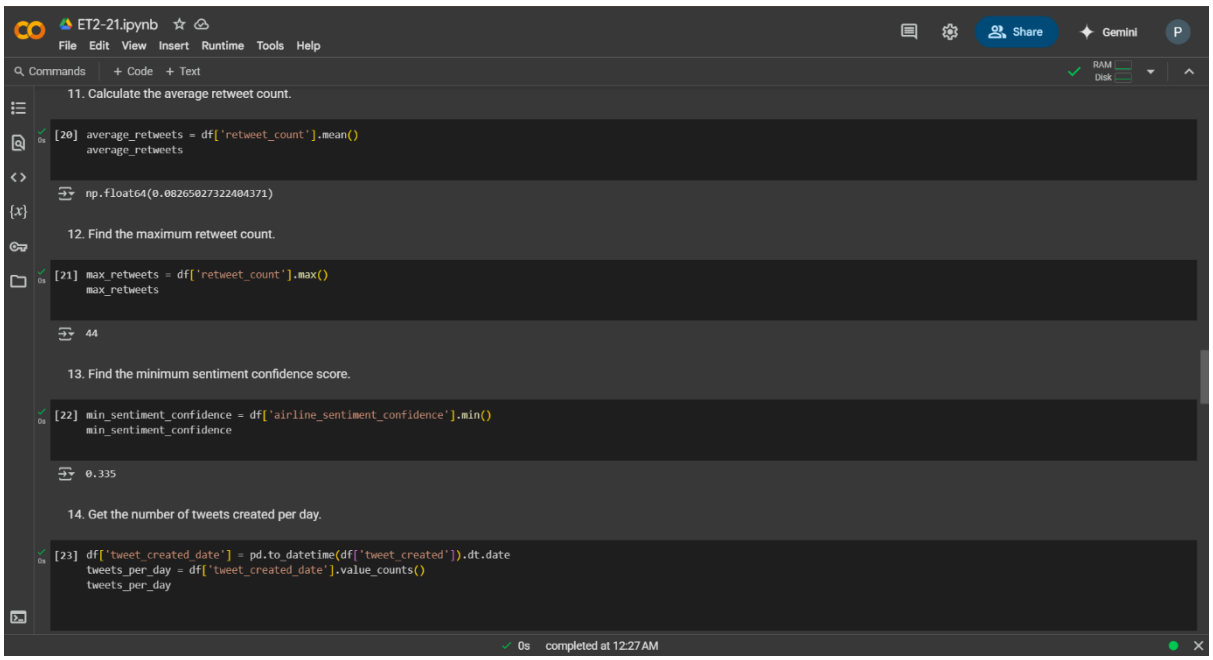

ET2-21.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
Start coding or generate with AI.
[1] from google.colab import files
uploaded = files.upload()
import pandas as pd
import io
# Replace the filename below with your uploaded file's name
df = pd.read_csv(io.BytesIO(uploaded['Tweets.csv']))
# Show first few rows
df.head()
Choose Files
Tweets.csv
Tweets.csv(text/csv) - 3421431 bytes, last modified: 4/28/2025 - 100% done
Saving Tweets.csv to Tweets.csv
tweet_id airline_sentiment airline_sentiment_confidence negativereason negativereason_confidence airline airline_sentiment_gold name negativereason_gold retw
0 570306133677760513 neutral 1.0000 NaN NaN Virgin America NaN cairdin NaN
1 570301130888122368 positive 0.3486 NaN 0.0000 Virgin America NaN jnardino NaN
2 570301083672813571 neutral 0.6837 NaN NaN Virgin America NaN yvonnalynn NaN
3 570301031407624196 negative 1.0000 Bad Flight 0.7033 Virgin America NaN jnardino NaN
0s completed at 12:27 AM

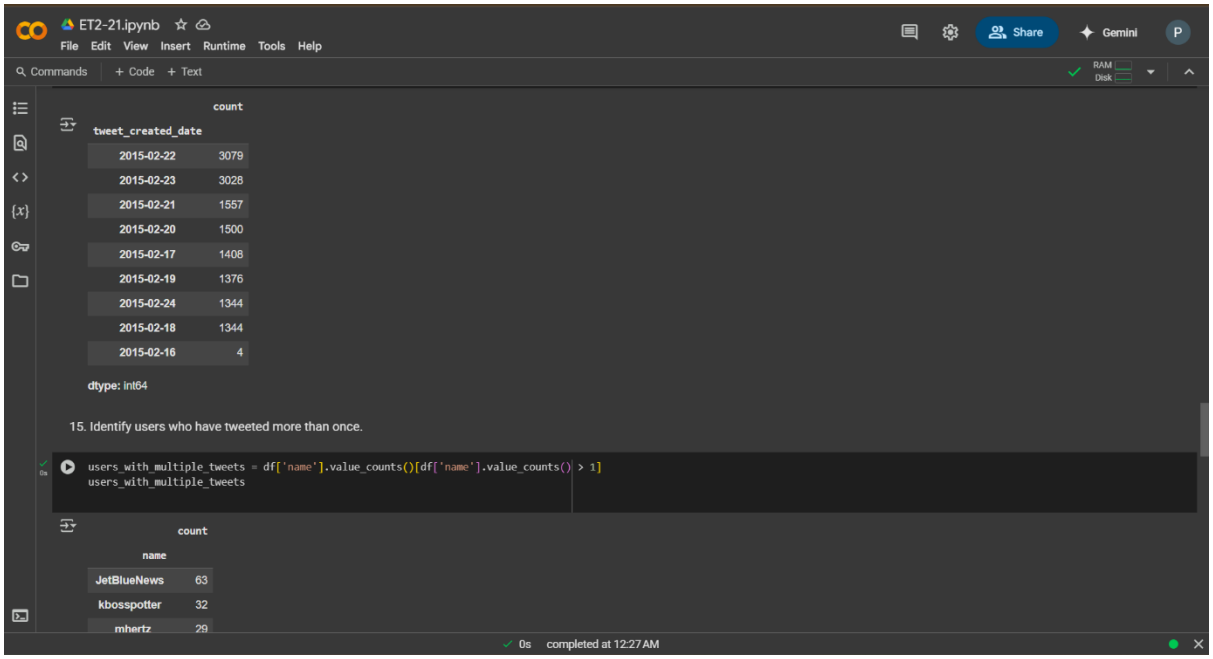

ET2-21.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
4 570300817074462722 negative 1.0000 Can't Tell 1.0000 Virgin America NaN jnardino NaN
Next steps: Generate code with df
View recommended plots
New interactive sheet
1. How many total tweets are there?
[9] import numpy as np
total_tweets = len(df)
total_tweets
14640
2. How many tweets are positive, negative, and neutral?
sentiment_counts = df['airline_sentiment'].value_counts()
sentiment_counts
count
airline_sentiment
negative 9178
neutral 3099
positive 2363
0s completed at 12:27 AM


ET2-21.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
RAM
Disk
dtype: int64
3. What is the percentage of each sentiment?
[12] sentiment_percentage = (sentiment_counts / total_tweets) * 100
sentiment_percentage
count
airline_sentiment
negative 62.691257
neutral 21.168033
positive 16.140710
dtype: float64
4. Find the airline with the maximum tweets.
[13]
most_tweeted_airline = df['airline'].value_counts().idxmax()
most_tweeted_airline
'United'
5. How many tweets were posted by each airline?
0s completed at 12:27 AM


ET2-21.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
RAM
Disk
5. How many tweets were posted by each airline?
[14] tweets_per_airline = df['airline'].value_counts()
tweets_per_airline
count
airline
United 3822
US Airways 2913
American 2759
Southwest 2420
Delta 2222
Virgin America 504
dtype: int64
6. How many tweets have missing locations?
missing_locations = df['tweet_location'].isnull().sum()
missing_locations
np.int64(4733)
7. How many tweets have coordinates?
0s completed at 12:27 AM


ET2-21.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
RAM
Disk
7. How many tweets have coordinates?
[16]
tweets_with_coords = df['tweet_coord'].notnull().sum()
tweets_with_coords
np.int64(1019)
8. What is the average sentiment confidence score?
[17]
avg_sentiment_confidence = df['airline_sentiment_confidence'].mean()
avg_sentiment_confidence
np.float64(0.9001688524590163)
9. How many negative reasons are mentioned?
[18]
negative_reasons_count = df['negativereason'].notnull().sum()
negative_reasons_count
np.int64(9178)
10. Find the most common negative reason.
most_common_negative_reason = df['negativereason'].value_counts().idxmax()
most_common_negative_reason
'Customer Service Issue'
0s completed at 12:27 AM


ET2-21.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
RAM
Disk
11. Calculate the average retweet count.
[20]
average_retweets = df['retweet_count'].mean()
average_retweets
np.float64(0.08265027322404371)
12. Find the maximum retweet count.
[21]
max_retweets = df['retweet_count'].max()
max_retweets
44
13. Find the minimum sentiment confidence score.
[22]
min_sentiment_confidence = df['airline_sentiment_confidence'].min()
min_sentiment_confidence
0.335
14. Get the number of tweets created per day.
[23]
df['tweet_created_date'] = pd.to_datetime(df['tweet_created']).dt.date
tweets_per_day = df['tweet_created_date'].value_counts()
tweets_per_day
0s completed at 12:27 AM



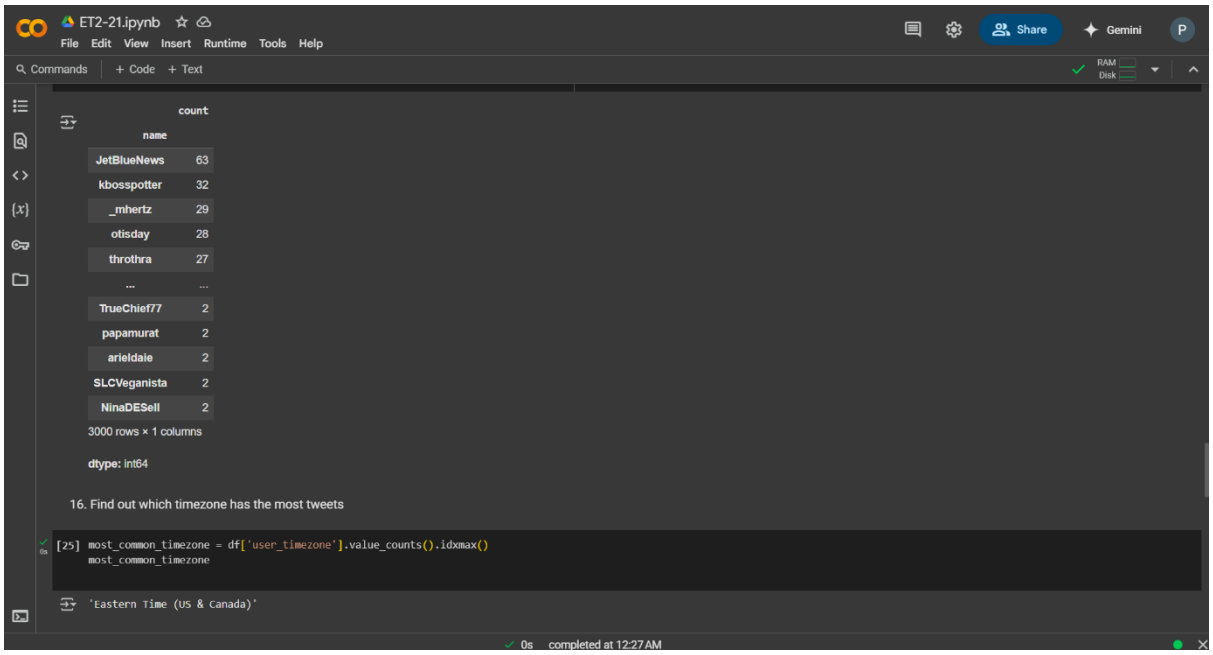
tweet_created_date	count
2015-02-22	3079
2015-02-23	3028
2015-02-21	1557
2015-02-20	1500
2015-02-17	1408
2015-02-19	1376
2015-02-24	1344
2015-02-18	1344
2015-02-16	4

dtype: int64

15. Identify users who have tweeted more than once.

```
users_with_multiple_tweets = df['name'].value_counts()[df['name'].value_counts() > 1]
users_with_multiple_tweets
```

name	count
JetBlueNews	63
kbosspotter	32
_mhertz	29



name	count
JetBlueNews	63
kbosspotter	32
_mhertz	29
otisday	28
throthra	27
...	...
TrueChief77	2
papamurat	2
arieldaie	2
SLCVeganista	2
NinaDESell	2

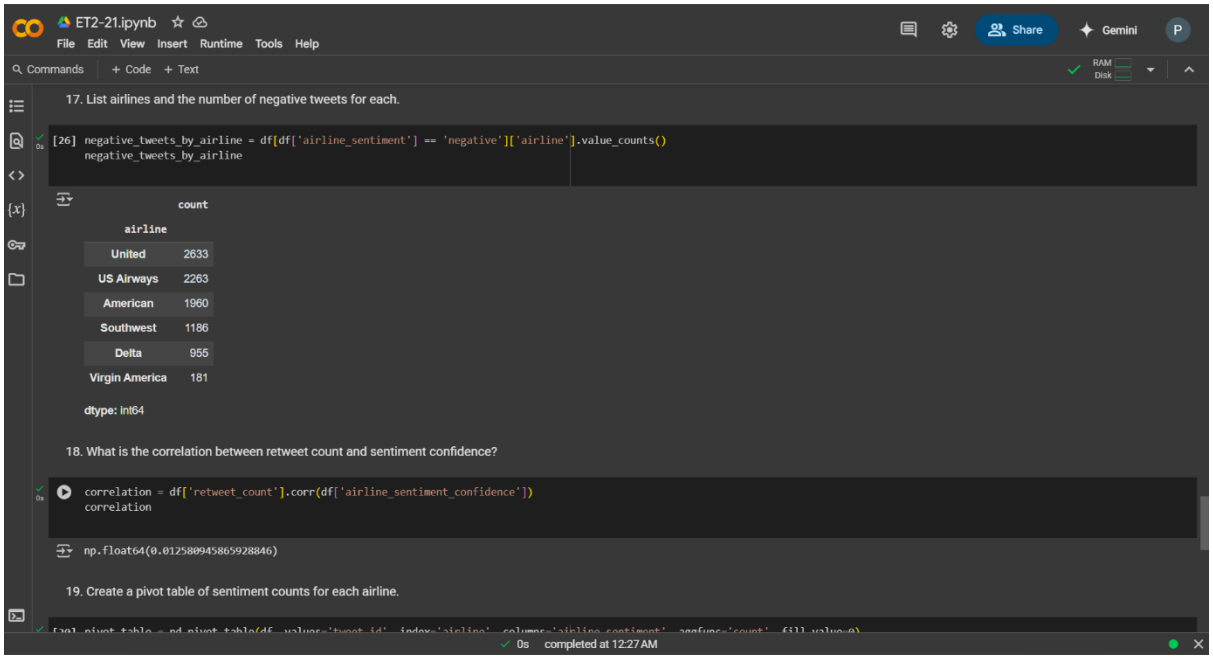
3000 rows × 1 columns

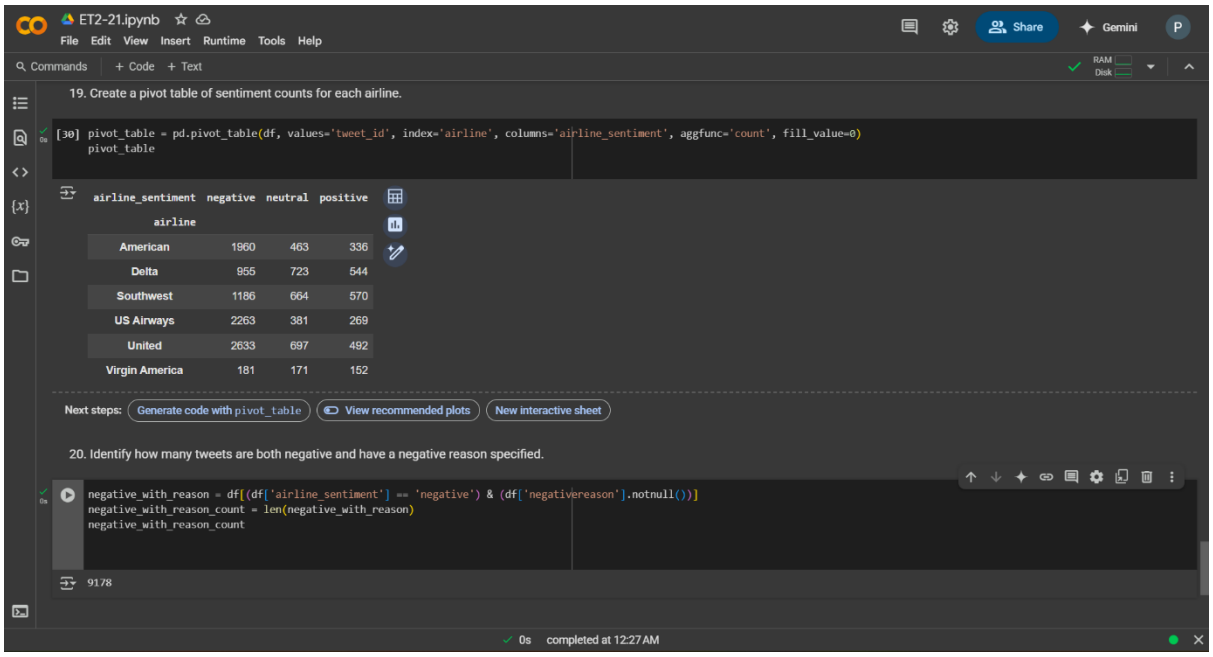
dtype: int64

16. Find out which timezone has the most tweets

```
[25] most_common_timezone = df['user_timezone'].value_counts().idxmax()
     most_common_timezone
```

```
'Eastern Time (US & Canada)'
```


ET2-21.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands + Code + Text
17. List airlines and the number of negative tweets for each.
[26] negative_tweets_by_airline = df[df['airline_sentiment'] == 'negative']['airline'].value_counts()
negative_tweets_by_airline
count
airline
United 2633
US Airways 2263
American 1960
Southwest 1186
Delta 955
Virgin America 181
dtype: int64
18. What is the correlation between retweet count and sentiment confidence?
correlation = df['retweet_count'].corr(df['airline_sentiment_confidence'])
correlation
np.float64(0.012580945865928846)
19. Create a pivot table of sentiment counts for each airline.
0s completed at 12:27 AM


ET2-21.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands + Code + Text
19. Create a pivot table of sentiment counts for each airline.
[30] pivot_table = pd.pivot_table(df, values='tweet_id', index='airline', columns='airline_sentiment', aggfunc='count', fill_value=0)
pivot_table
airline_sentiment negative neutral positive
airline
American 1960 463 336
Delta 955 723 544
Southwest 1186 664 570
US Airways 2263 381 269
United 2633 697 492
Virgin America 181 171 152
Next steps: Generate code with pivot_table View recommended plots New interactive sheet
20. Identify how many tweets are both negative and have a negative reason specified.
negative_with_reason = df[(df['airline_sentiment'] == 'negative') & (df['negativereason'].notnull())]
negative_with_reason_count = len(negative_with_reason)
negative_with_reason_count
9178
0s completed at 12:27 AM