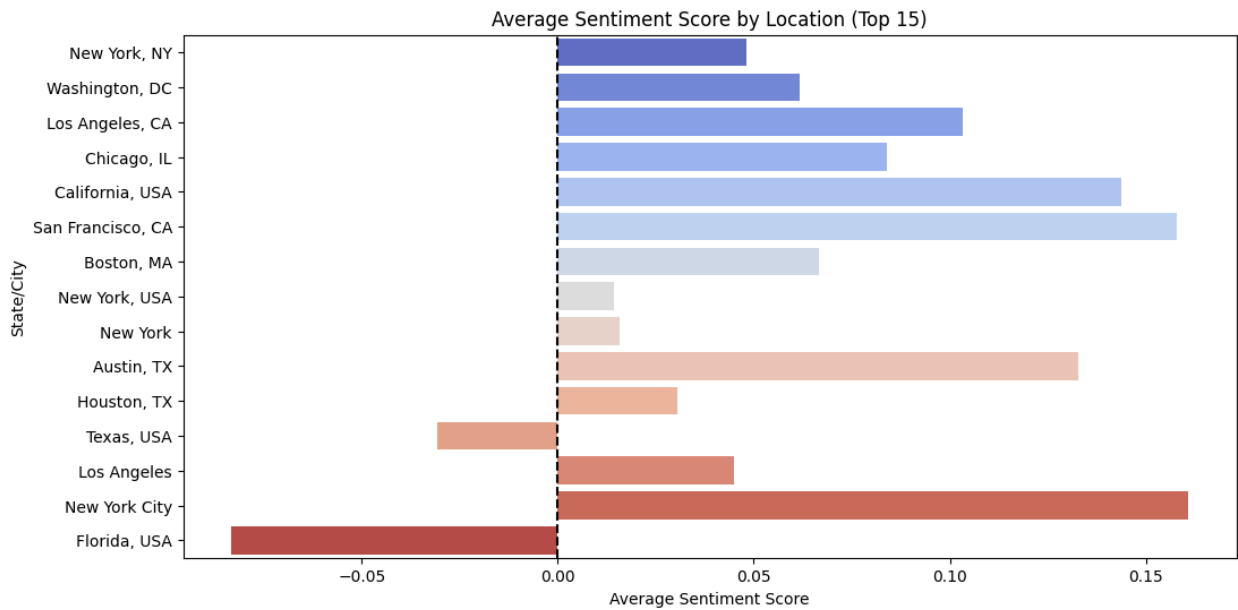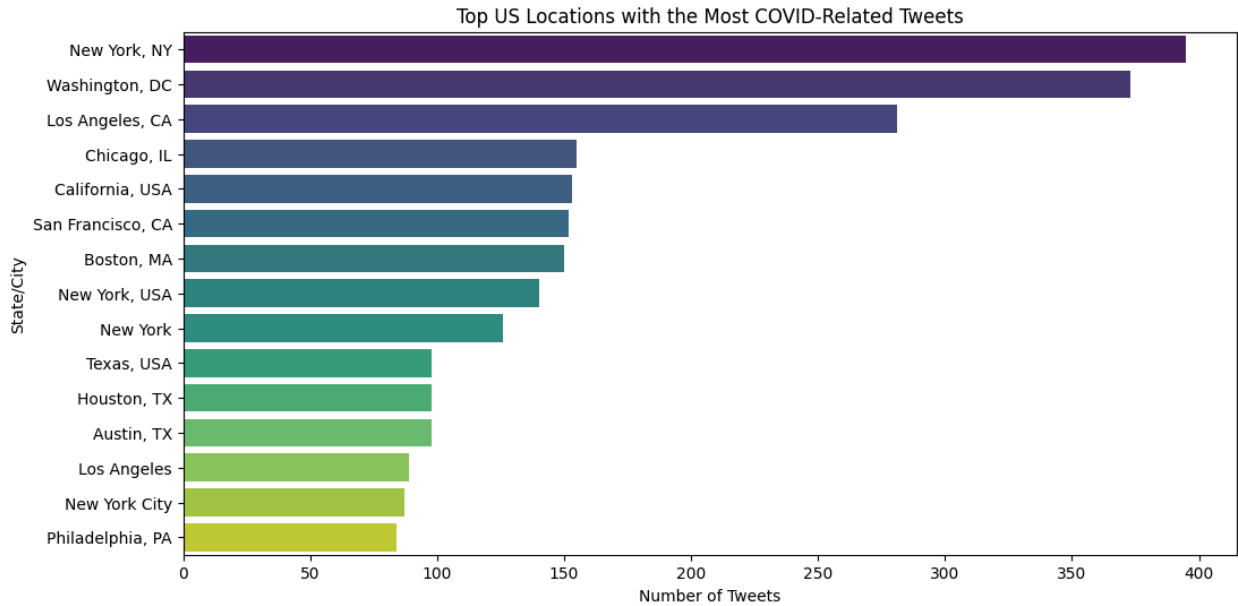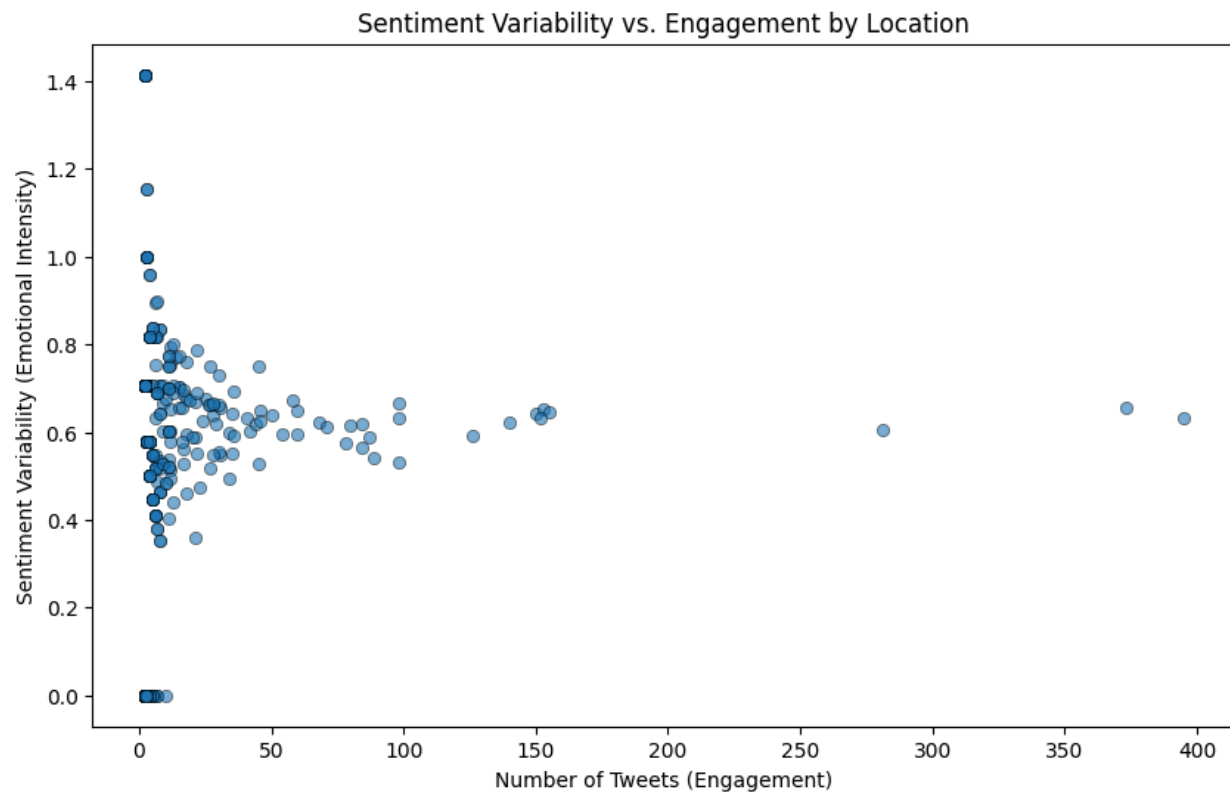# Data Appendix

Questions the Data Answers:

- Does the severity of Tweet sentiments by location correlate with more drastic state and local shutdown protocols?
- What states produced the most COVID-related tweets? What was the average sentiment score for each state or major city?

Top US Locations with the Most COVID-Related Tweets

A horizontal bar chart titled "Top US Locations with the Most COVID-Related Tweets" showing Number of Tweets on the x-axis (0 to 400) and State/City on the y-axis. Locations from top to bottom: New York, NY (~395); Washington, DC (~375); Los Angeles, CA (~280); Chicago, IL (~155); California, USA (~153); San Francisco, CA (~152); Boston, MA (~150); New York, USA (~140); New York (~125); Texas, USA (~98); Houston, TX (~98); Austin, TX (~97); Los Angeles (~88); New York City (~86); Philadelphia, PA (~84).

Average Sentiment Score by Location (Top 15)

A horizontal bar chart titled "Average Sentiment Score by Location (Top 15)" showing Average Sentiment Score on the x-axis (approximately -0.05 to 0.15) and State/City on the y-axis. Locations from top to bottom: New York, NY (~0.045 positive); Washington, DC (~0.06 positive); Los Angeles, CA (~0.10 positive); Chicago, IL (~0.085 positive); California, USA (~0.14 positive); San Francisco, CA (~0.155 positive); Boston, MA (~0.065 positive); New York, USA (~0.015 positive); New York (~0.018 positive); Austin, TX (~0.13 positive); Houston, TX (~0.03 positive); Texas, USA (~-0.03 negative); Los Angeles (~0.045 positive); New York City (~0.16 positive); Florida, USA (~-0.08 negative).

Sentiment Variability vs. Engagement by Location

*Figures 5-7

Dataset 1: Corona_NLP_train.csv

Variable	Definition	Total number of observations/missing observations (n(m))
UserName	Data type: int64 Unique identifier for each user	41157(0)
ScreenName	Data type: int64 Another unique identifier for the user's account	41157(0)
Location	Data type: object The location provided by the user	41157(8590)
TweetAt	Data type: object The date the tweet was posted (DD-MM-YYYY)	41157(0)
OriginalTweet	Data type: object The full text of the tweet	41157(0)

| Sentiment | Data type: object
The sentiment label assigned to the tweet | 41157(0) |

Dataset 2: Corona_NLP_test.csv

Variable	Definition	Total number of observations/missing observations (n(m))
UserName	Data type: int64 Unique identifier for each user	3798(0)
ScreenName	Data type: int64 Another unique identifier for the user's account	3798(0)
Location	Data type: object The location provided by the user	3798(834)
TweetAt	Data type: object The date the tweet was posted (DD-MM-YYYY)	3798(0)
OriginalTweet	Data type: object The full text of the tweet	3798(0)
Sentiment	Data type: object The sentiment label assigned to the tweet	3798(0)

Dataset 3: df_us_with_sentiment.csv

Variable	Definition	Total number of observations/missing observations (n(m))
UserName	Data type: int64 Unique identifier for each user	7224(0)
ScreenName	Data type: int64 Another unique identifier for the user's account	7224(0)
Location	Data type: object The location provided by the user	7224(0)

TweetAt	Data type: object The date the tweet was posted (DD-MM-YYYY)	7224(0)
OriginalTweet	Data type: object The full text of the tweet	7224(0)
Sentiment	Data type: object The sentiment label assigned to the tweet	7224(0)
VADER Score	Data type: float VADER sentiment score calculated	7224(0)