analyze_PR_Advisory_Tweets_Feb_2025

April 5, 2025

1 Analysis of `PR_Advisory_Tweets_Feb_2025.csv`

This notebook presents a structured analysis of Twitter data related to the Puerto Rico tsunami advisory alert in February 2025. It includes:

- **Data loading and preprocessing**
- **Data wrangling and feature engineering**
- **Exploratory data analysis (EDA)**
- **Visualizations** using libraries such as Pandas, Matplotlib, Seaborn, Plotly, Folium, PyWaffle, and WordCloud

The goal is to extract meaningful insights from tweet content and engagement patterns, with attention to language use, temporal trends, and textual features. This notebook serves both as an analytical report and as a reference for applying diverse Python tools in social media analysis.

1.1 1. Import Libraries

The following libraries will be used throughout the notebook.

[19]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Plotly
import plotly.express as px
import plotly.graph_objects as go

# For Waffle charts (PyWaffle)
from pywaffle import Waffle

# For WordCloud
from wordcloud import WordCloud
```

1.2 2. Data Loading

Load the CSV file into a Pandas DataFrame.

[20]:
```
# Load CSV with header inferred
df = pd.read_csv('PR_Advisory_Tweets_Feb_2025.csv')

print("Data Loaded. Number of rows:", df.shape[0])
df.head()
```

```
Data Loaded. Number of rows: 122
```

[20]:

	Query_Str	Post_URL	Author_Name	Author_Web_Page_URL	Author_Handle	Verified_Status	UTC_Time	Ads	Tweet_Content	Post_ID	…	Like_Count	View_Count	Bookmark_Count	\
0	Puerto Rico (tsunami OR sismo OR terremoto OR …	https://x.com/search?q=Puerto Rico (tsunami OR…	sia | fan	https://x.com/lalisalovemme	lalisalovemme	True	2025-02-07 00:08:19+00:00	False	THIS IS A 9.9 MAGNITUDE MOTHERSQUAKE!!!!\n\n#D…	1.887655e+18	…	4621	60693	66	
1	Puerto Rico (tsunami OR sismo OR terremoto OR …	https://x.com/search?q=Puerto Rico (tsunami OR…	Geól. Sergio Almazán	https://x.com/chematierra	chematierra	True	2025-02-07 14:07:39+00:00	False	AVISO \nEnjambre sísmico intenso cercano a #S…	1.887866e+18	…	221	8646	5	
2	Puerto Rico (tsunami OR sismo OR terremoto OR …	https://x.com/search?q=Puerto Rico (tsunami OR…	Jack Straw	https://x.com/JackStr42679640	JackStr42679640	True	2025-02-07 11:43:05+00:00	False	Guantanamo Bay Alert! (Watch till end) \n \nT…	1.887829e+18	…	427	15428	46	
3	Puerto Rico (tsunami OR sismo OR terremoto OR …	https://x.com/search?q=Puerto Rico (tsunami OR…	Belen Larchens	https://x.com/belenlarchens	belenlarchens	False	2025-02-07 23:07:14+00:00	False	El municipio de #ElBolsón informó : «se ordena…	1.888002e+18	…	467	9041	4	
4	Puerto Rico (tsunami OR sismo OR terremoto OR …	https://x.com/search?q=Puerto Rico (tsunami OR…	SkyAlert	https://x.com/SkyAlertMx	SkyAlertMx	True	2025-02-07 02:43:14+00:00	False	#Sismo magnitud 4.0 (SSN) ubicado a 14 km al s…	1.887694e+18	…				

```
4        229       15918                1

                                    Tweet_Image_URL  Replying_to  \
0  https://pbs.twimg.com/ext_tw_video_thumb/18876…        False
1    https://pbs.twimg.com/media/GjML8z7WEAAeZgK.jpg        False
2  https://pbs.twimg.com/ext_tw_video_thumb/18878…        False
3    https://pbs.twimg.com/media/GjOHcENXoAAlZnc.jpg        False
4  https://pbs.twimg.com/media/GjJvTPaXwAAxZCN.jp…        False

   Reply_to_Whom Reply_to_Whom_URL  Reply_to_Whom_Username  \
0            NaN               NaN                     NaN
1            NaN               NaN                     NaN
2            NaN               NaN                     NaN
3            NaN               NaN                     NaN
4            NaN               NaN                     NaN

  Reply_to_Whom_Handle Language
0                  NaN       en
1                  NaN       es
2                  NaN       en
3                  NaN       es
4                  NaN       es

[5 rows x 23 columns]
```

1.3 3. Quick Exploration

Let's quickly inspect the data.

[21]:
```
# View the first few and last few rows
print(df.head())
print(df.tail())

# Summary information
df.info()
print(df.describe(include='all'))

# Check for null values
print(df.isnull().sum())

# Data types
print(df.dtypes)
```

```
                                          Query_Str  \
0  Puerto Rico (tsunami OR sismo OR terremoto OR …
1  Puerto Rico (tsunami OR sismo OR terremoto OR …
2  Puerto Rico (tsunami OR sismo OR terremoto OR …
3  Puerto Rico (tsunami OR sismo OR terremoto OR …
```

```
4  Puerto Rico (tsunami OR sismo OR terremoto OR …

                                            Post_URL           Author_Name  \
0  https://x.com/search?q=Puerto Rico (tsunami OR…          sia   | fan
1  https://x.com/search?q=Puerto Rico (tsunami OR…  Geól. Sergio Almazán
2  https://x.com/search?q=Puerto Rico (tsunami OR…            Jack Straw
3  https://x.com/search?q=Puerto Rico (tsunami OR…    Belen Larchens
4  https://x.com/search?q=Puerto Rico (tsunami OR…              SkyAlert

            Author_Web_Page_URL    Author_Handle  Verified_Status  \
0    https://x.com/lalisalovemme    lalisalovemme             True
1      https://x.com/chematierra      chematierra             True
2  https://x.com/JackStr42679640  JackStr42679640             True
3    https://x.com/belenlarchens    belenlarchens            False
4       https://x.com/SkyAlertMx       SkyAlertMx             True

                    UTC_Time    Ads  \
0  2025-02-07 00:08:19+00:00  False
1  2025-02-07 14:07:39+00:00  False
2  2025-02-07 11:43:05+00:00  False
3  2025-02-07 23:07:14+00:00  False
4  2025-02-07 02:43:14+00:00  False

                                      Tweet_Content       Post_ID  …  \
0  THIS IS A 9.9 MAGNITUDE MOTHERSQUAKE!!!!\n\n#D…  1.887655e+18  …
1  AVISO  \nEnjambre sísmico intenso cercano a #S…  1.887866e+18  …
2   Guantanamo Bay Alert! (Watch till end) \n \nT…  1.887829e+18  …
3  El municipio de #ElBolsón informó : «se ordena…  1.888002e+18  …
4  #Sismo magnitud 4.0 (SSN) ubicado a 14 km al s…  1.887694e+18  …

  Like_Count  View_Count  Bookmark_Count  \
0       4621       60693              66
1        221        8646               5
2        427       15428              46
3        467        9041               4
4        229       15918               1

                                   Tweet_Image_URL  Replying_to  \
0  https://pbs.twimg.com/ext_tw_video_thumb/18876…        False
1    https://pbs.twimg.com/media/GjML8z7WEAAeZgK.jpg        False
2  https://pbs.twimg.com/ext_tw_video_thumb/18878…        False
3    https://pbs.twimg.com/media/GjOHcENXoAAlZnc.jpg        False
4  https://pbs.twimg.com/media/GjJvTPaXwAAxZCN.jp…        False

   Reply_to_Whom Reply_to_Whom_URL  Reply_to_Whom_Username  \
0            NaN               NaN                     NaN
1            NaN               NaN                     NaN
2            NaN               NaN                     NaN
```

```
3            NaN               NaN                      NaN
4            NaN               NaN                      NaN

  Reply_to_Whom_Handle Language
0                  NaN       en
1                  NaN       es
2                  NaN       en
3                  NaN       es
4                  NaN       es

[5 rows x 23 columns]
                                            Query_Str  \
117  Puerto Rico (tsunami OR sismo OR terremoto OR …
118  Puerto Rico (tsunami OR sismo OR terremoto OR …
119  Puerto Rico (tsunami OR sismo OR terremoto OR …
120  Puerto Rico (tsunami OR sismo OR terremoto OR …
121  Puerto Rico (tsunami OR sismo OR terremoto OR …

                                             Post_URL  \
117  https://x.com/search?q=Puerto Rico (tsunami OR…
118  https://x.com/search?q=Puerto Rico (tsunami OR…
119  https://x.com/search?q=Puerto Rico (tsunami OR…
120  https://x.com/search?q=Puerto Rico (tsunami OR…
121  https://x.com/search?q=Puerto Rico (tsunami OR…

                    Author_Name            Author_Web_Page_URL  \
117  Centro Sismológico Nacional  https://x.com/Sismos_Peru_IGP
118          The Spectator Index   https://x.com/spectatorindex
119               Alerta News 24     https://x.com/AlertaNews24
120               Alerta News 24     https://x.com/AlertaNews24
121               Alerta News 24     https://x.com/AlertaNews24

       Author_Handle  Verified_Status                   UTC_Time    Ads  \
117  Sismos_Peru_IGP             True  2025-02-08 14:19:12+00:00  False
118   spectatorindex             True  2025-02-08 23:40:45+00:00  False
119     AlertaNews24             True  2025-02-08 23:39:31+00:00  False
120     AlertaNews24             True  2025-02-08 23:40:50+00:00  False
121     AlertaNews24             True  2025-02-08 23:36:15+00:00  False

                                       Tweet_Content       Post_ID  …  \
117  REPORTE SÍSMICO\nIGP/CENSIS/RS 2025-0101\nFech…  1.888231e+18  …
118  BREAKING: Tsunami alert for coastlines along t…  1.888372e+18  …
119    | URGENTE: ALERTA DE TSUNAMI PARA ISLAS CAI…  1.888372e+18  …
120   | URGENTE: El Centro de Alerta de Tsunamis e…  1.888372e+18  …
121   | URGENTE: USGS: "Un terremoto de magnitud 8…  1.888371e+18  …

    Like_Count  View_Count  Bookmark_Count  Tweet_Image_URL  Replying_to  \
117        233       18668               4              NaN        False
```

```
118       2380      573696              63              NaN        False
119       3351      391537              53              NaN        False
120       3924      541184              80              NaN        False
121       7584      830556             160              NaN        False

     Reply_to_Whom Reply_to_Whom_URL  Reply_to_Whom_Username  \
117            NaN               NaN                     NaN
118            NaN               NaN                     NaN
119            NaN               NaN                     NaN
120            NaN               NaN                     NaN
121            NaN               NaN                     NaN

    Reply_to_Whom_Handle Language
117                  NaN       es
118                  NaN       en
119                  NaN       es
120                  NaN       es
121                  NaN       es

[5 rows x 23 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 122 entries, 0 to 121
Data columns (total 23 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Query_Str               122 non-null    object
 1   Post_URL                122 non-null    object
 2   Author_Name             122 non-null    object
 3   Author_Web_Page_URL     122 non-null    object
 4   Author_Handle           122 non-null    object
 5   Verified_Status         122 non-null    bool
 6   UTC_Time                122 non-null    object
 7   Ads                     122 non-null    bool
 8   Tweet_Content           122 non-null    object
 9   Post_ID                 122 non-null    float64
 10  Tweet_URL               122 non-null    object
 11  Reply_Count             122 non-null    float64
 12  Repost_Count            122 non-null    int64
 13  Like_Count              122 non-null    int64
 14  View_Count              122 non-null    int64
 15  Bookmark_Count          122 non-null    int64
 16  Tweet_Image_URL         84 non-null     object
 17  Replying_to             122 non-null    bool
 18  Reply_to_Whom           8 non-null      object
 19  Reply_to_Whom_URL       8 non-null      object
 20  Reply_to_Whom_Username  7 non-null      object
 21  Reply_to_Whom_Handle    8 non-null      object
 22  Language                122 non-null    object
```

```
dtypes: bool(3), float64(2), int64(4), object(14)
memory usage: 19.5+ KB
                                                Query_Str  \
count                                                 122
unique                                                  1
top     Puerto Rico (tsunami OR sismo OR terremoto OR …
freq                                                  122
mean                                                  NaN
std                                                   NaN
min                                                   NaN
25%                                                   NaN
50%                                                   NaN
75%                                                   NaN
max                                                   NaN

                                                Post_URL     Author_Name  \
count                                                 122             122
unique                                                  2              88
top     https://x.com/search?q=Puerto Rico (tsunami OR…  Alerta News 24
freq                                                  112               7
mean                                                  NaN             NaN
std                                                   NaN             NaN
min                                                   NaN             NaN
25%                                                   NaN             NaN
50%                                                   NaN             NaN
75%                                                   NaN             NaN
max                                                   NaN             NaN

              Author_Web_Page_URL Author_Handle Verified_Status  \
count                         122           122             122
unique                         88            88               2
top     https://x.com/AlertaNews24  AlertaNews24            True
freq                            7             7              98
mean                          NaN           NaN             NaN
std                           NaN           NaN             NaN
min                           NaN           NaN             NaN
25%                           NaN           NaN             NaN
50%                           NaN           NaN             NaN
75%                           NaN           NaN             NaN
max                           NaN           NaN             NaN

                        UTC_Time    Ads  \
count                        122    122
unique                       118      1
top     2025-02-08 23:49:19+00:00  False
freq                           2    122
mean                         NaN    NaN
std                          NaN    NaN
```

```
min                          NaN    NaN
25%                          NaN    NaN
50%                          NaN    NaN
75%                          NaN    NaN
max                          NaN    NaN

                                          Tweet_Content       Post_ID  …  \
count                                               122  1.220000e+02  …
unique                                              122           NaN  …
top     THIS IS A 9.9 MAGNITUDE MOTHERSQUAKE!!!!\n\n#D…           NaN  …
freq                                                  1           NaN  …
mean                                                NaN  1.888320e+18  …
std                                                 NaN  1.547304e+14  …
min                                                 NaN  1.887655e+18  …
25%                                                 NaN  1.888372e+18  …
50%                                                 NaN  1.888373e+18  …
75%                                                 NaN  1.888375e+18  …
max                                                 NaN  1.888377e+18  …

          Like_Count    View_Count  Bookmark_Count  \
count     122.000000  1.220000e+02      122.000000
unique           NaN           NaN             NaN
top              NaN           NaN             NaN
freq             NaN           NaN             NaN
mean     1593.196721  2.184285e+05       78.401639
std      4244.411236  6.021819e+05      232.226402
min         9.000000  1.726000e+03        0.000000
25%       222.500000  1.555050e+04        4.000000
50%       464.000000  6.182000e+04       18.500000
75%      1204.250000  1.419998e+05       56.750000
max     40906.000000  5.372795e+06     2017.000000

                                       Tweet_Image_URL  Replying_to  \
count                                               84          122
unique                                              81            2
top     https://pbs.twimg.com/amplify_video_thumb/1888…        False
freq                                                 3          114
mean                                               NaN          NaN
std                                                NaN          NaN
min                                                NaN          NaN
25%                                                NaN          NaN
50%                                                NaN          NaN
75%                                                NaN          NaN
max                                                NaN          NaN

        Reply_to_Whom         Reply_to_Whom_URL Reply_to_Whom_Username  \
count               8                         8                      7
unique              7                         7                      7
```

```
top        rawsalerts  https://x.com/rawsalerts            Nick Sortor
freq                2                         2                      1
mean              NaN                       NaN                    NaN
std               NaN                       NaN                    NaN
min               NaN                       NaN                    NaN
25%               NaN                       NaN                    NaN
50%               NaN                       NaN                    NaN
75%               NaN                       NaN                    NaN
max               NaN                       NaN                    NaN

       Reply_to_Whom_Handle Language
count                     8      122
unique                    7        2
top              rawsalerts       es
freq                      2       68
mean                    NaN      NaN
std                     NaN      NaN
min                     NaN      NaN
25%                     NaN      NaN
50%                     NaN      NaN
75%                     NaN      NaN
max                     NaN      NaN

[11 rows x 23 columns]
Query_Str                  0
Post_URL                   0
Author_Name                0
Author_Web_Page_URL        0
Author_Handle              0
Verified_Status            0
UTC_Time                   0
Ads                        0
Tweet_Content              0
Post_ID                    0
Tweet_URL                  0
Reply_Count                0
Repost_Count               0
Like_Count                 0
View_Count                 0
Bookmark_Count             0
Tweet_Image_URL           38
Replying_to                0
Reply_to_Whom            114
Reply_to_Whom_URL        114
Reply_to_Whom_Username   115
Reply_to_Whom_Handle     114
Language                   0
dtype: int64
```

```
Query_Str                  object
Post_URL                   object
Author_Name                object
Author_Web_Page_URL        object
Author_Handle              object
Verified_Status              bool
UTC_Time                   object
Ads                          bool
Tweet_Content              object
Post_ID                   float64
Tweet_URL                  object
Reply_Count               float64
Repost_Count                int64
Like_Count                  int64
View_Count                  int64
Bookmark_Count              int64
Tweet_Image_URL            object
Replying_to                  bool
Reply_to_Whom              object
Reply_to_Whom_URL          object
Reply_to_Whom_Username     object
Reply_to_Whom_Handle       object
Language                   object
dtype: object
```

1.4 4. Data Cleaning and Feature Engineering

The dataset was preprocessed to retain only the most relevant attributes for analysis. The following steps were performed:

- Removed columns containing only missing values.
- Dropped metadata and auxiliary fields not required for the analysis, including author information, image URLs, and reply targets.
- Filled missing values in engagement-related fields (`Reply_Count`, `Repost_Count`, `Like_Count`, `Bookmark_Count`) and converted them to integer type.
- Created a `Total_Engagement` column by summing the individual engagement metrics.
- Selected the key features for analysis: `Post_ID`, `Tweet_Content`, `Total_Engagement`, and `Language`.
- Added a `Tweet_Length` column to capture the number of characters in each tweet.

This results in a streamlined DataFrame suitable for content, engagement, and language-based analysis.

[22]:
```
# Step 1: Drop columns with all null values
df = pd.read_csv("PR_Advisory_Tweets_Feb_2025.csv")
df = df.dropna(axis=1, how='all')

# Step 2: Drop unnecessary metadata columns
drop_cols = [
```

```
    'Query_Str', 'Post_URL', 'Author_Name', 'Author_Web_Page_URL',␣
 ↪'Author_Handle',
    'Verified_Status', 'Tweet_URL', 'Tweet_Image_URL',
    'Replying_to', 'Reply_to_Whom', 'Reply_to_Whom_URL',
    'Reply_to_Whom_Username', 'Reply_to_Whom_Handle', 'Ads'
]
df = df.drop(columns=[col for col in drop_cols if col in df.columns])

# Step 3: Fill and convert engagement columns
engagement_cols = ["Reply_Count", "Repost_Count", "Like_Count",␣
 ↪"Bookmark_Count"]
for col in engagement_cols:
    df[col] = df[col].fillna(0).astype(int)

# Step 4: Create Total_Engagement column
df["Total_Engagement"] = df["Reply_Count"] + df["Repost_Count"] +␣
 ↪df["Like_Count"] + df["Bookmark_Count"]

# Step 5: Rename and convert UTC time for time-based analysis
df["Timestamp_UTC"] = pd.to_datetime(df["UTC_Time"])

# Step 6: Select relevant columns
df_selected = df[["Post_ID", "Tweet_Content", "Total_Engagement", "Language",␣
 ↪"Timestamp_UTC"]].copy()

# Step 7: Add Tweet_Length column
df_selected["Tweet_Length"] = df_selected["Tweet_Content"].str.len()

# Step 8: Add time-based features for later use
df_selected["Hour"] = df_selected["Timestamp_UTC"].dt.hour
df_selected["Weekday"] = df_selected["Timestamp_UTC"].dt.day_name()
df_selected["Date"] = df_selected["Timestamp_UTC"].dt.date

# Step 9: Add Tweet_Length_Category based on length bins
# Define tweet length bins and labels
length_bins = [0, 80, 140, 200, 280, df_selected["Tweet_Length"].max()]
length_labels = ["Very Short", "Short", "Medium", "Long", "Very Long"]

# Create a new categorical column
df_selected["Tweet_Length_Category"] = pd.cut(
    df_selected["Tweet_Length"],
    bins=length_bins,
    labels=length_labels,
    include_lowest=True
)

# Preview cleaned DataFrame
```

```
df_selected.head()
```

[22]:
```
         Post_ID                                      Tweet_Content  \
0  1.887655e+18  THIS IS A 9.9 MAGNITUDE MOTHERSQUAKE!!!!\n\n#D…
1  1.887866e+18  AVISO  \nEnjambre sísmico intenso cercano a #S…
2  1.887829e+18   Guantanamo Bay Alert! (Watch till end) \n \nT…
3  1.888002e+18  El municipio de #ElBolsón informó : «se ordena…
4  1.887694e+18  #Sismo magnitud 4.0 (SSN) ubicado a 14 km al s…

   Total_Engagement Language             Timestamp_UTC  Tweet_Length  Hour  \
0              6214       en 2025-02-07 00:08:19+00:00            97     0
1               319       es 2025-02-07 14:07:39+00:00           303    14
2               657       en 2025-02-07 11:43:05+00:00           302    11
3              1023       es 2025-02-07 23:07:14+00:00           300    23
4               255       es 2025-02-07 02:43:14+00:00           167     2

  Weekday        Date Tweet_Length_Category
0  Friday  2025-02-07                 Short
1  Friday  2025-02-07             Very Long
2  Friday  2025-02-07             Very Long
3  Friday  2025-02-07             Very Long
4  Friday  2025-02-07                Medium
```

1.5 5. Data analysis and exploration

1.5.1 SECTION A: Distribution Analysis

These plots show the shape and spread of the two main numeric features: - **Total_Engagement:** How much attention tweets received. - **Tweet_Length:** Number of characters in each tweet.

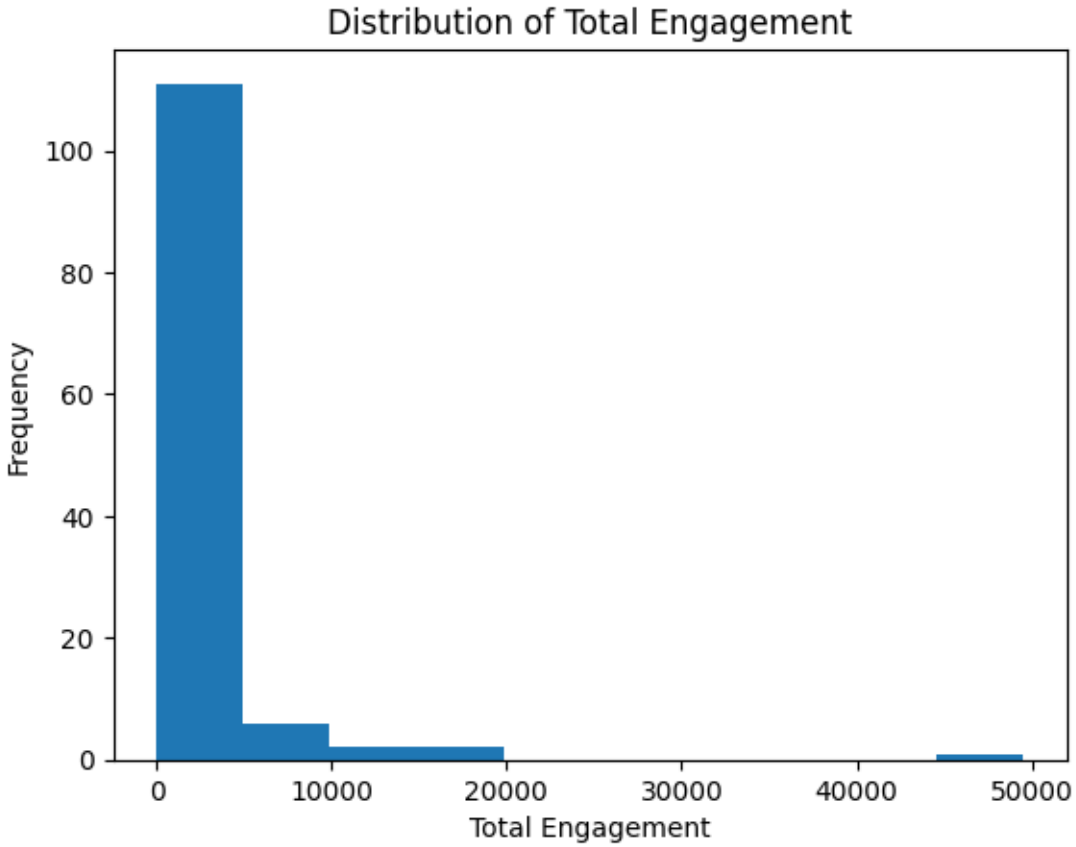
We use histograms to see distribution and boxplots to spot outliers.

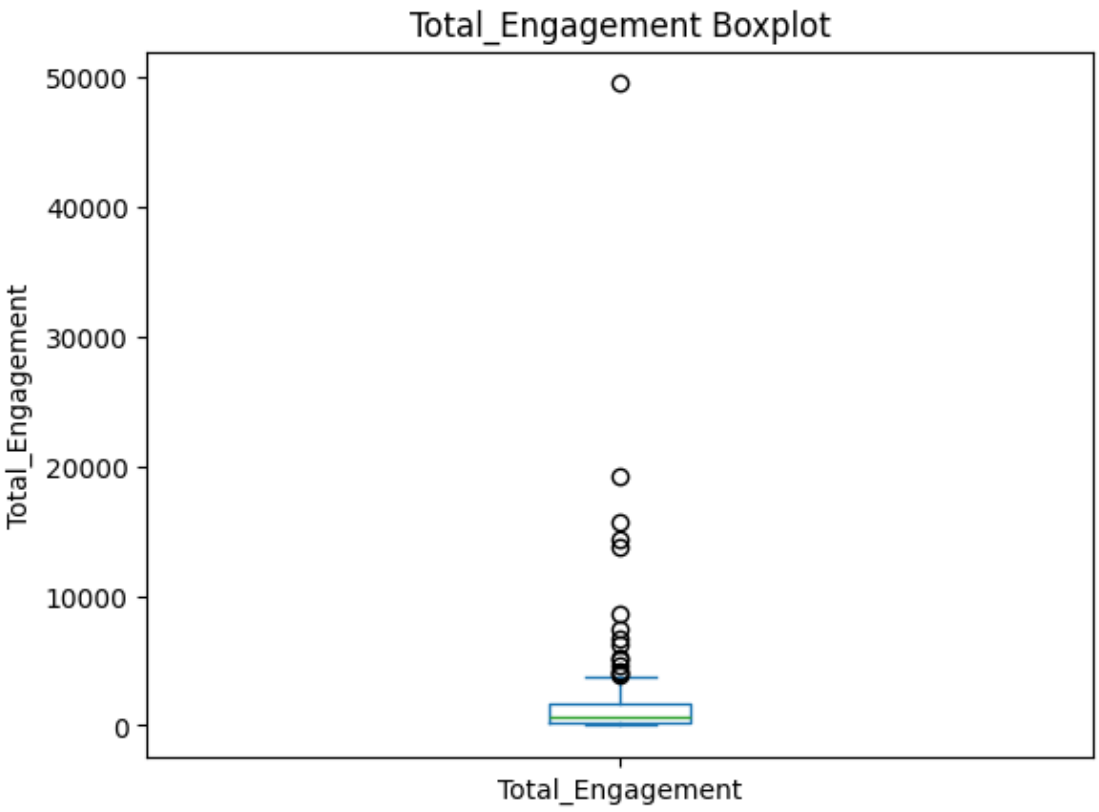
[23]:
```
# ========================================================
# SECTION A: TOTAL ENGAGEMENT DISTRIBUTION & INSIGHTS
# ========================================================

# Histogram: Raw Total Engagement
df_selected["Total_Engagement"].plot(kind="hist", bins=10, title="Distribution␣
 ↪of Total Engagement")
plt.xlabel("Total Engagement")
plt.ylabel("Frequency")
plt.show()

# Boxplot: Raw Total Engagement
df_selected["Total_Engagement"].plot(kind="box", title="Total_Engagement␣
 ↪Boxplot")
plt.ylabel("Total_Engagement")
plt.show()
```

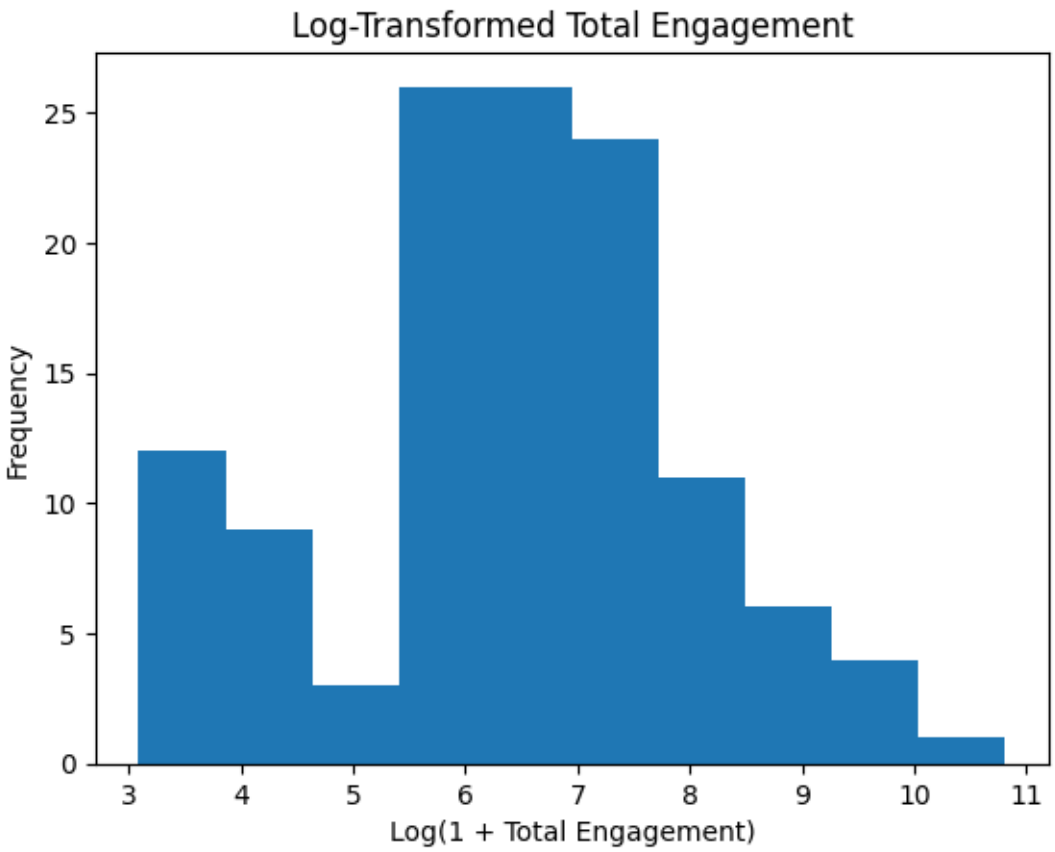
```
# Log-transform Total Engagement for better scale visibility
df_selected["Log_Total_Engagement"] = np.log1p(df_selected["Total_Engagement"])␣
 ↪ # log(1 + x)
df_selected["Log_Total_Engagement"].plot(kind="hist", bins=10,␣
 ↪title="Log-Transformed Total Engagement")
plt.xlabel("Log(1 + Total Engagement)")
plt.ylabel("Frequency")
plt.show()

# Boxplot: IQR-filtered Total Engagement (removing extreme outliers)
Q1 = df_selected["Total_Engagement"].quantile(0.25)
Q3 = df_selected["Total_Engagement"].quantile(0.75)
IQR = Q3 - Q1
filtered_df = df_selected[
    (df_selected["Total_Engagement"] >= Q1 - 1.5 * IQR) &
    (df_selected["Total_Engagement"] <= Q3 + 1.5 * IQR)
]

filtered_df["Total_Engagement"].plot(kind="box", title="Filtered Total␣
 ↪Engagement Boxplot (No Outliers)")
plt.ylabel("Total_Engagement")
plt.show()


# ========================================================
# SECTION B: TWEET LENGTH DISTRIBUTION & CATEGORIZATION
# ========================================================

# Histogram: Tweet Length
df_selected["Tweet_Length"].plot(kind="hist", bins=10, title="Distribution of␣
 ↪Tweet Length")
plt.xlabel("Tweet Length (characters)")
plt.ylabel("Frequency")
plt.show()

# Boxplot: Tweet Length
df_selected["Tweet_Length"].plot(kind="box", title="Tweet Length Boxplot")
plt.ylabel("Tweet Length (characters)")
plt.show()

# Count Plot: Tweet count by length category
sns.countplot(
    x="Tweet_Length_Category",
    hue="Tweet_Length_Category",  # same as x to apply palette correctly
    data=df_selected,
    order=["Very Short", "Short", "Medium", "Long", "Very Long"],
```

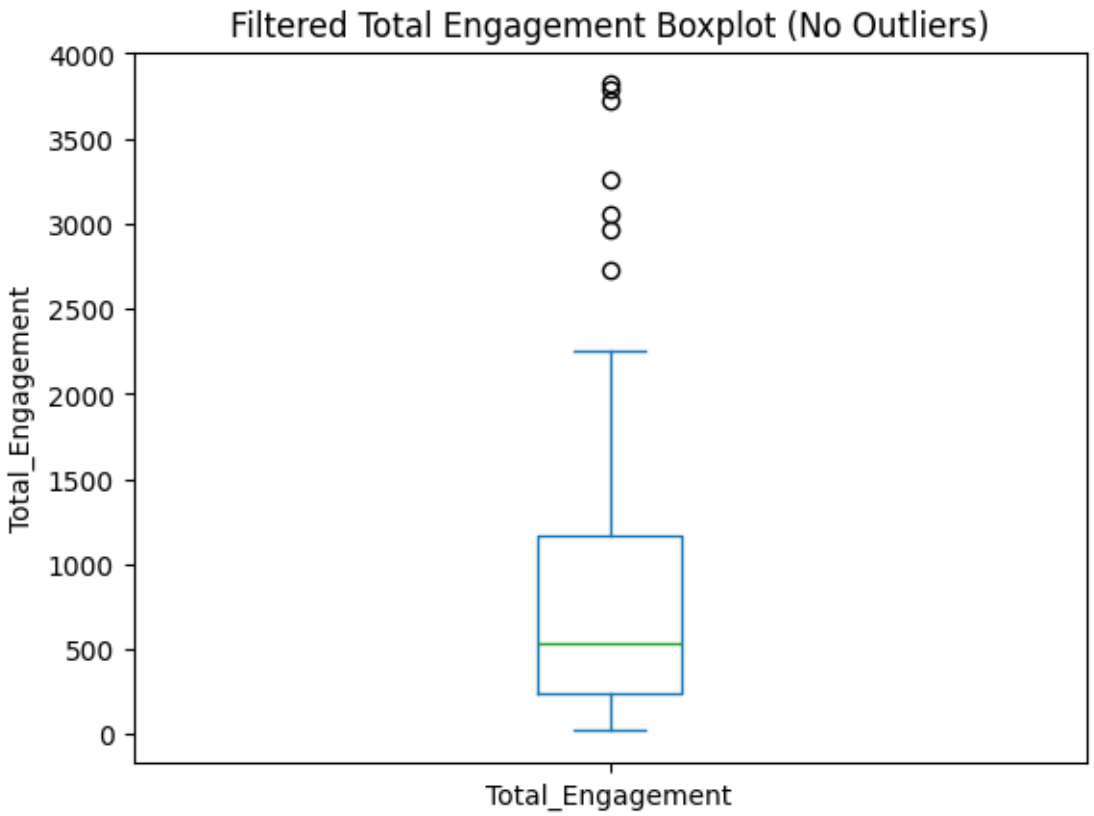
```
    palette="pastel",
    legend=False
)
plt.title("Tweet Count by Length Category")
plt.xlabel("Tweet Length Category")
plt.ylabel("Number of Tweets")
plt.show()

# Waffle Chart: Tweet length category distribution
length_counts = df_selected["Tweet_Length_Category"].value_counts().
 ↪sort_index().to_dict()

fig = plt.figure(
    FigureClass=Waffle,
    rows=5,
    values=length_counts,
    figsize=(10, 4),
    title={"label": "Tweet Length Distribution", "loc": "center"},
    legend={'loc': 'upper left', 'bbox_to_anchor': (1, 1)},
    colors=["#b3e2cd", "#fdcdac", "#cbd5e8", "#f4cae4", "#e6f5c9"],
    block_arranging_style='snake',
)
plt.show()


# ======================================================
# SECTION C: LANGUAGE DISTRIBUTION
# ======================================================

# Count Plot: Tweets by Language
if "Language" in df_selected.columns:
    sns.countplot(
        x="Language",
        hue="Language",
        data=df_selected,
        palette="pastel",
        legend=False
    )
    plt.title("Tweet Counts by Language")
    plt.xlabel("Language")
    plt.ylabel("Tweet Count")
    plt.show()

# Waffle Chart: Language Distribution
language_counts = df_selected["Language"].value_counts().to_dict()

fig = plt.figure(
```

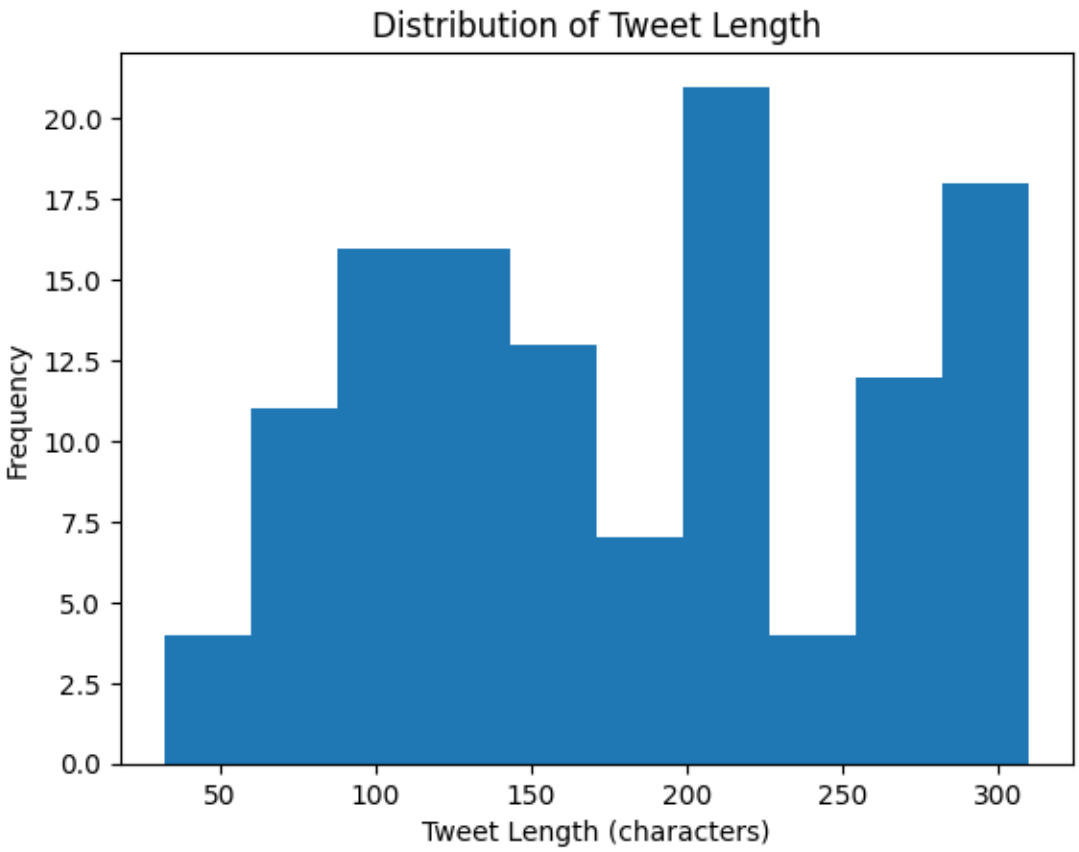
```
    FigureClass=Waffle,
    rows=5,
    values=language_counts,
    figsize=(8, 4),
    title={"label": "Tweet Language Distribution (en vs es)", "loc": "center"},
    legend={'loc': 'upper left', 'bbox_to_anchor': (1, 1)},
    colors=["#66c2a5", "#fc8d62"],
    block_arranging_style='snake',
)
plt.show()
```

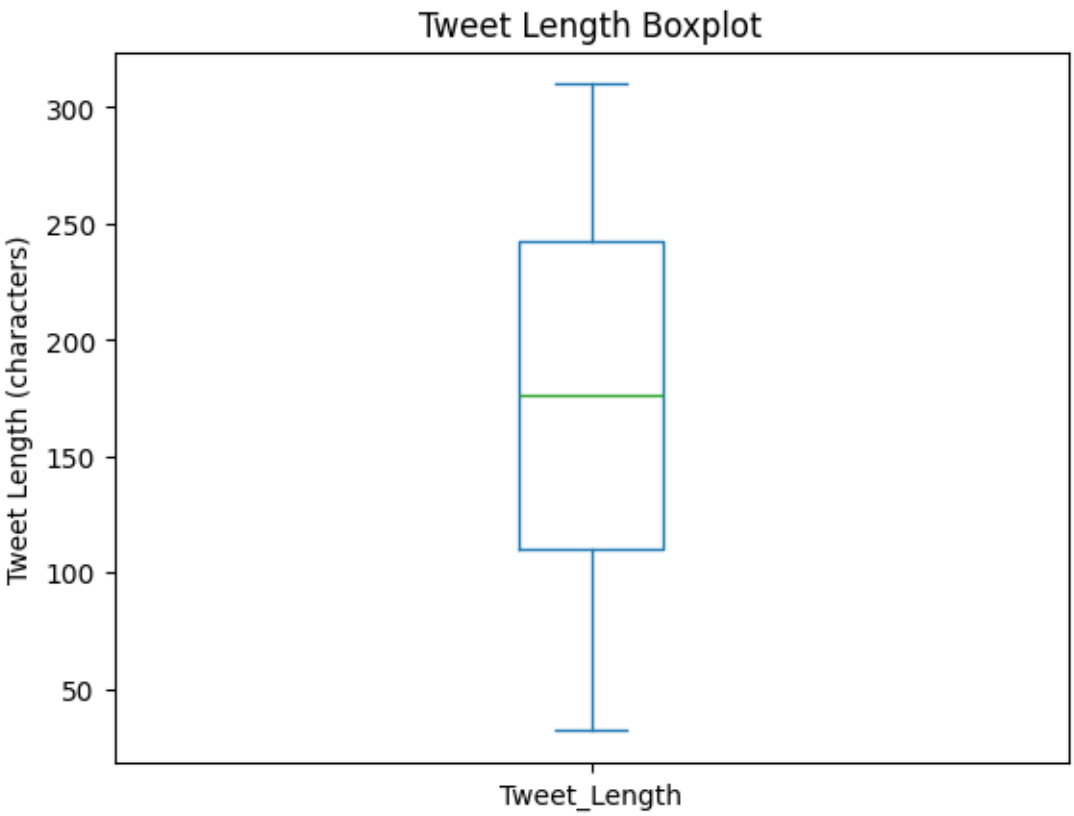


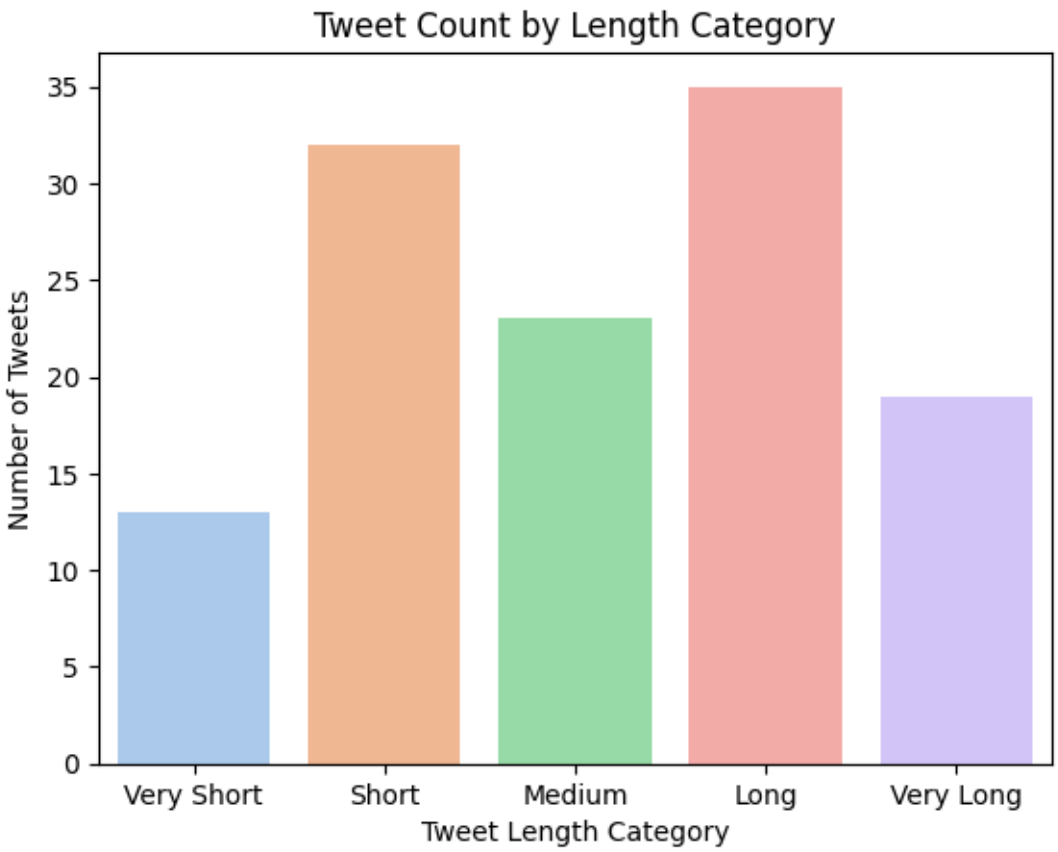

Total_Engagement Boxplot
50000
40000
30000
20000
10000
0
Total_Engagement
Total_Engagement

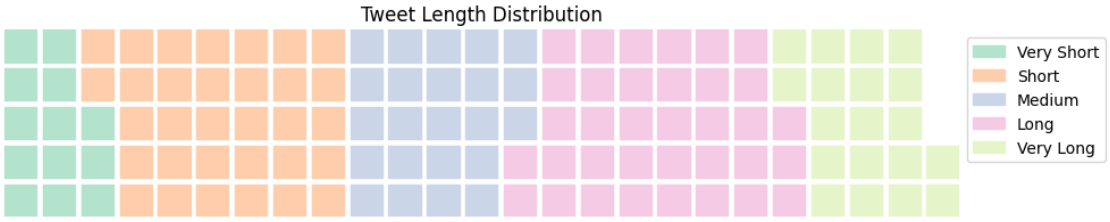

Log-Transformed Total Engagement
Frequency
0
5
10
15
20
25
3
4
5
6
7
8
9
10
11
Log(1 + Total Engagement)

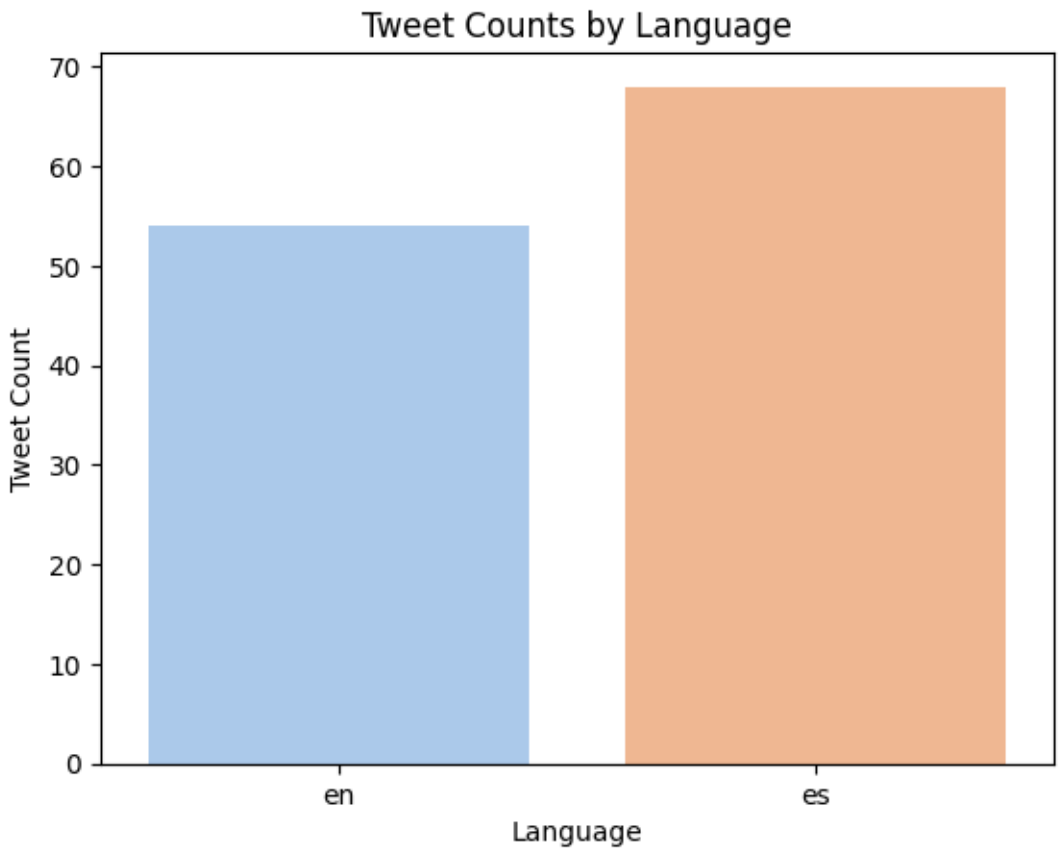

Filtered Total Engagement Boxplot (No Outliers)
Total_Engagement
4000
3500
3000
2500
2000
1500
1000
500
0
Total_Engagement

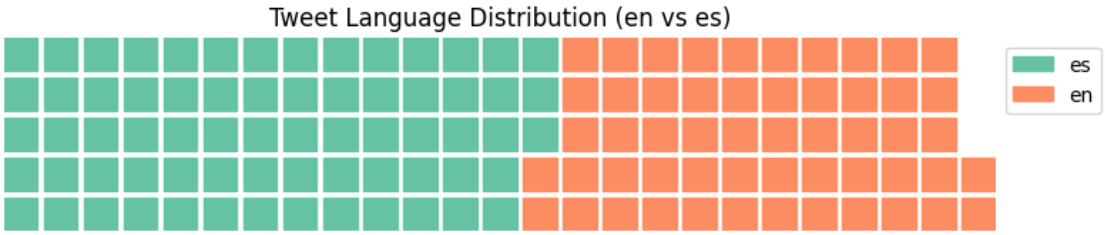

Distribution of Tweet Length
Frequency
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
50
100
150
200
250
300
Tweet Length (characters)


Tweet Length Boxplot
Tweet Length (characters)
300
250
200
150
100
50
Tweet_Length


Tweet Count by Length Category
Number of Tweets
0
5
10
15
20
25
30
35
Very Short
Short
Medium
Long
Very Long
Tweet Length Category


Tweet Length Distribution
Very Short
Short
Medium
Long
Very Long





1.6 Total Engagement Distribution

Raw Histogram

- The distribution of `Total_Engagement` is **extremely right-skewed**.
- Most tweets have low engagement, while a few go beyond **10,000 or even 50,000**, acting as extreme outliers.

Raw Boxplot

- The majority of tweets are clustered below **1,000** engagement.

- Clear presence of **extreme outliers** up to 50K that distort the scale.

Log-Transformed Histogram

- Log-transforming (`log(1 + x)`) yields a **more normalized, bell-shaped** distribution.
- Helps reveal patterns hidden in the long tail of low-engagement tweets.

Filtered Boxplot (No Outliers)

- After applying an IQR filter, the boxplot is **much more readable**.
- Typical tweets range from about **200 to 2,200** engagements.
- Reveals a **more interpretable core distribution**.

1.7 Insight: `Total_Engagement` needs transformation or outlier filtering for clearer analysis. Most tweets remain low in engagement, but a handful dominate the distribution.

1.8 Tweet Length Distribution

Histogram & Boxplot

- Tweet lengths range from about **30 to 310 characters**.
- A **uniform-like distribution** with slight peaks toward higher lengths.
- The median tweet is around **180–200 characters**.

Length Categories (Count + Waffle)

- **Short** and **Long** tweets dominate usage.
- **Very Short** and **Very Long** tweets are less common.
- The **Waffle Chart** and **Bar Plot** reinforce this pattern visually.

1.9 Insight: Tweet lengths are relatively balanced, with a lean toward medium-to-long messages, likely for sharing complete information or guidance.

1.10 Language Distribution

Bar Plot & Waffle Chart

- Spanish (`es`) tweets are **more common** than English (`en`) in this dataset.
- The language split is **consistent across visualizations**, suggesting more engagement or outreach in Spanish for this advisory event.

Insight: Language plays a key role in communication patterns—Spanish dominates this dataset, indicating localized outreach or audience targeting.

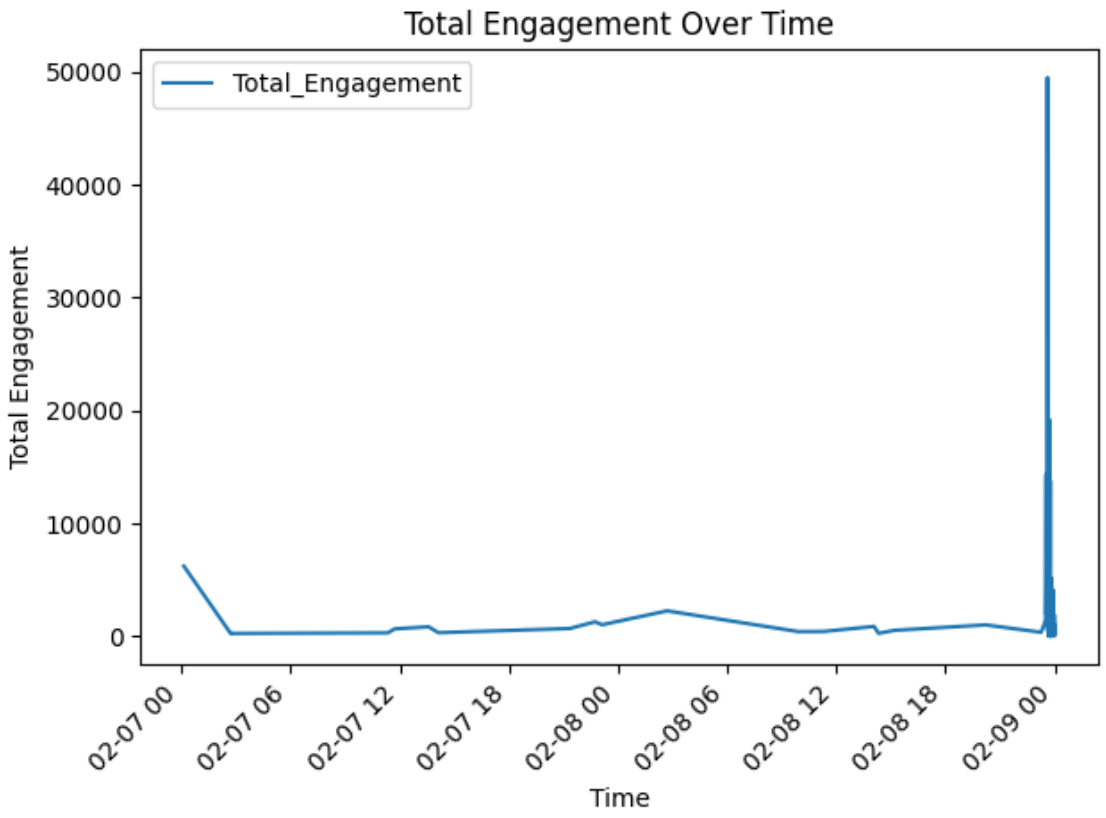
1.10.1 SECTION B: Time-Based Trends

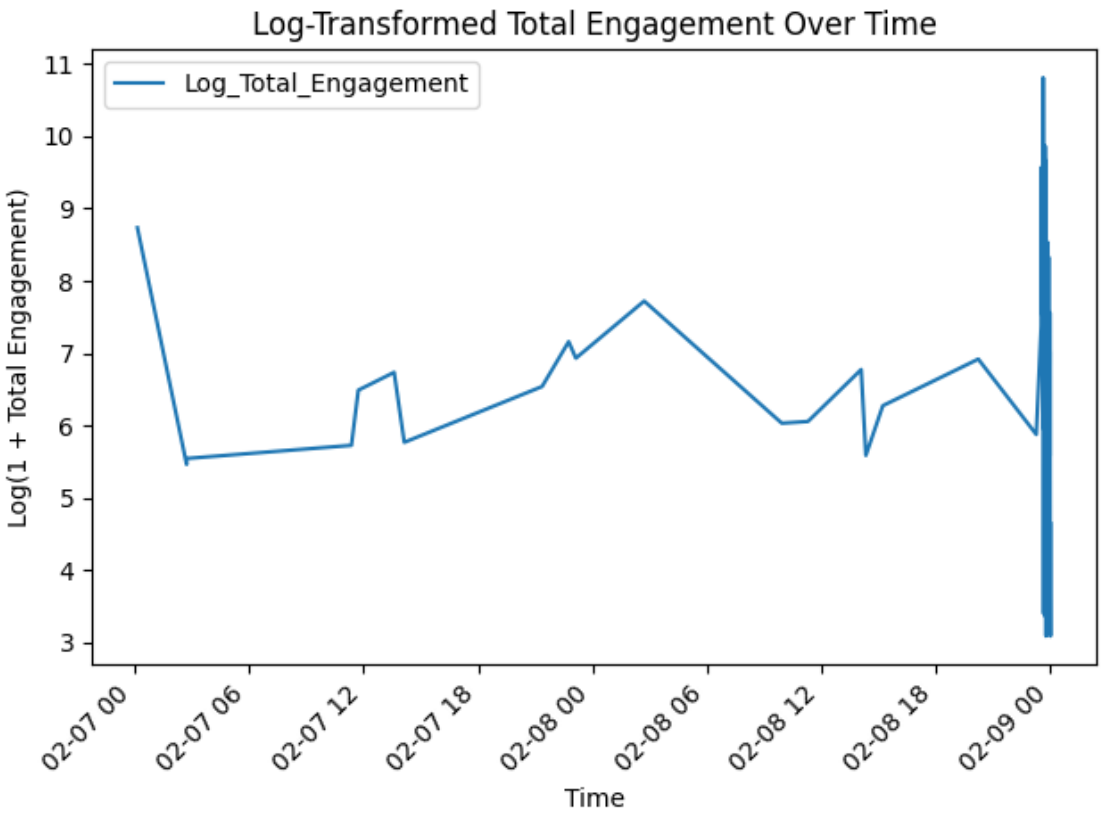
These plots track how `Total_Engagement` and `Tweet_Length` change over the sequence of tweets (approximated by index order).

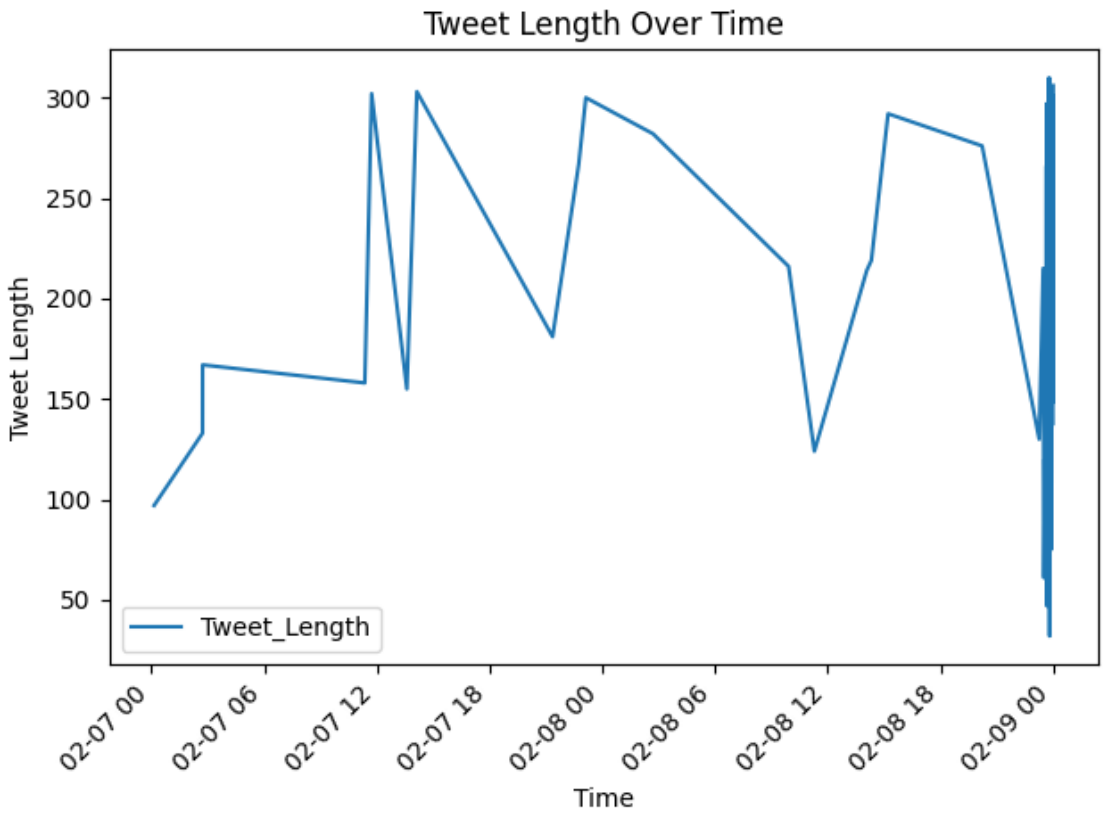
- **Line plot of engagement**: Detects bursts of public attention.
- **Line plot of tweet length**: Shows if users wrote longer or shorter tweets over time.

This helps analyze public behavior as the earthquake situation unfolded.

[24]:
```
# ----------------------------------------
# SECTION B: TIME-BASED TRENDS
# ----------------------------------------

# Line Plot: Total Engagement over time (raw)
df_selected.sort_values("Timestamp_UTC").plot(
    x="Timestamp_UTC", y="Total_Engagement", kind="line",
    title="Total Engagement Over Time"
)
plt.xlabel("Time")
plt.ylabel("Total Engagement")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

# Line Plot: Total Engagement over time (log-transformed)
df_selected["Log_Total_Engagement"] = np.log1p(df_selected["Total_Engagement"])␣
 ↪ # log(1 + x)
df_selected.sort_values("Timestamp_UTC").plot(
    x="Timestamp_UTC", y="Log_Total_Engagement", kind="line",
    title="Log-Transformed Total Engagement Over Time"
)
plt.xlabel("Time")
plt.ylabel("Log(1 + Total Engagement)")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

# Line Plot: Tweet Length over time
df_selected.sort_values("Timestamp_UTC").plot(
    x="Timestamp_UTC", y="Tweet_Length", kind="line",
    title="Tweet Length Over Time"
)
plt.xlabel("Time")
plt.ylabel("Tweet Length")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```


Total Engagement Over Time
Total_Engagement
50000
40000
30000
20000
10000
0
Total Engagement
02-07 00
02-07 06
02-07 12
02-07 18
02-08 00
02-08 06
02-08 12
02-08 18
02-09 00
Time


Log-Transformed Total Engagement Over Time
Log_Total_Engagement
Log(1 + Total Engagement)
11
10
9
8
7
6
5
4
3
02-07 00
02-07 06
02-07 12
02-07 18
02-08 00
02-08 06
02-08 12
02-08 18
02-09 00
Time



1.10.2 Time-Based Trends

Total Engagement Over Time (Raw)

- The engagement timeline reveals a **dramatic spike on February 9**, corresponding to the **earthquake/tsunami advisory** — a clear signal of **heightened public attention**.
- One post reaches nearly **50,000 interactions**, indicating a **viral advisory or critical update** during the event.
- Outside of this peak, engagement is generally modest, with **occasional smaller bursts** of interaction.

Total Engagement Over Time (Log-Transformed)

- Log-transformation smooths the scale, offering a clearer view of **consistent engagement patterns** beyond the February 9 surge.
- The February 9 spike remains a clear anomaly, but the rest of the activity becomes more visible, showing a **baseline rhythm** of advisory engagement.
- This format helps detect **trend shifts** without letting one viral moment dominate the view.

Tweet Length Over Time

- Around February 9, there's a noticeable **cluster of longer tweets**, many approaching the **character limit (280–310)**.
- This suggests users and agencies were sharing **detailed updates or emergency advisories** during the critical window.
- Shorter tweets are also interspersed, likely serving as **brief alerts or follow-ups**.

Insight

February 9 stands out as the **epicenter of public engagement**, triggered by the earthquake/tsunami advisory. Viral spikes, longer tweets, and elevated activity all align with the timeline of the event. **Log-transformed engagement** helps contextualize this moment within the broader communication pattern, offering insight into how urgency and message format evolved in real time.

1.10.3 SECTION C: Language-Based Comparisons

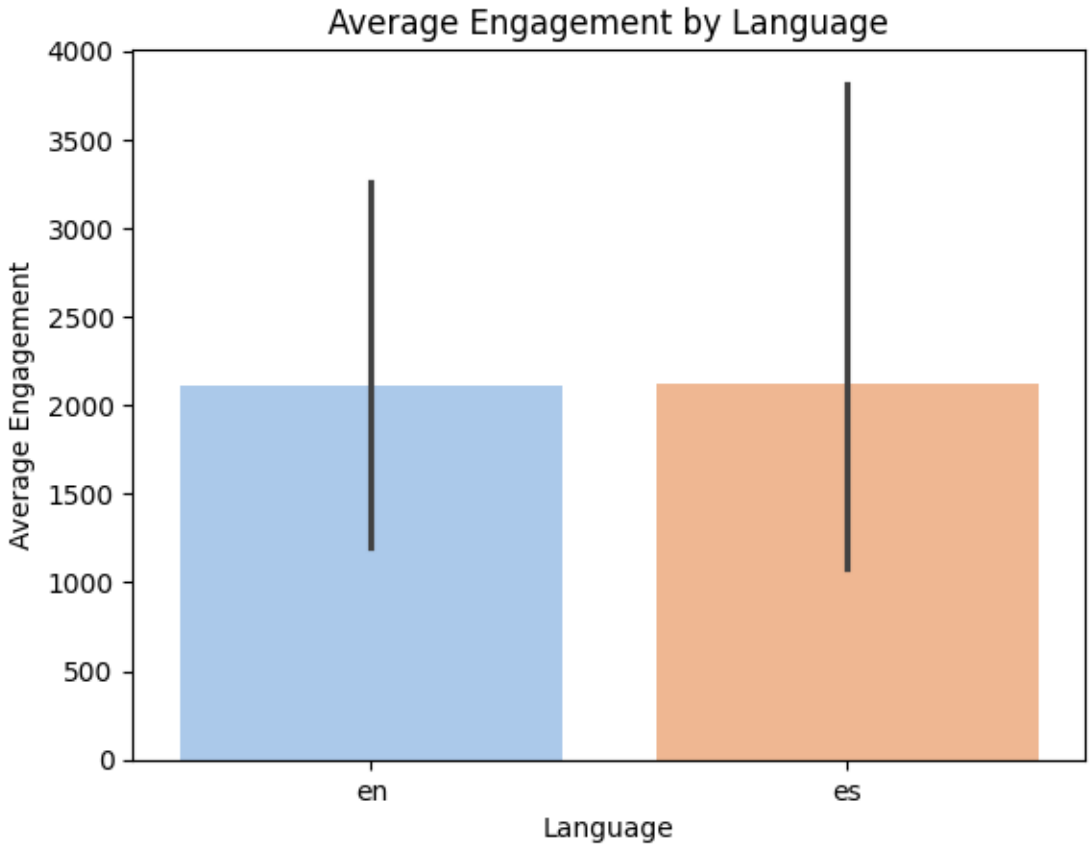
We compare tweet behavior by language:

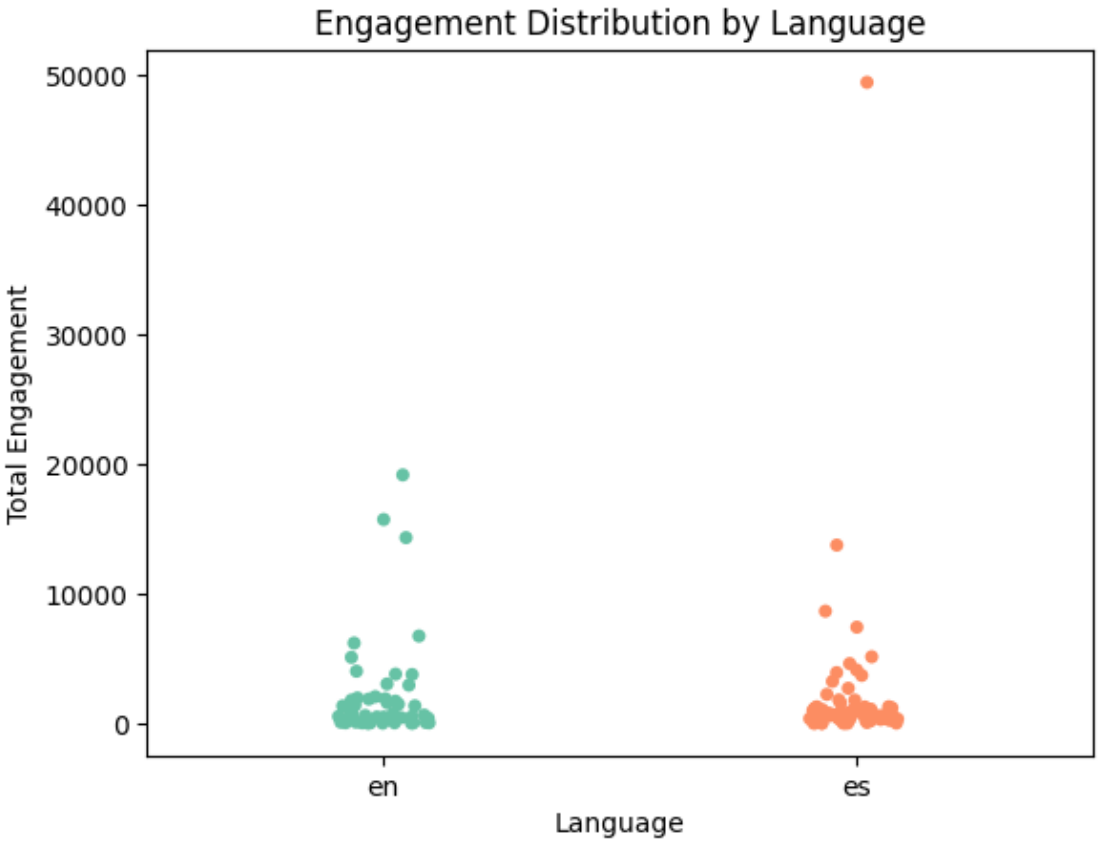
- **Count Plot**: Shows how many tweets were posted in each language.
- **Bar Plot**: Compares average engagement levels by language.
- **Strip Plot**: Reveals the spread and distribution of engagement for each language.

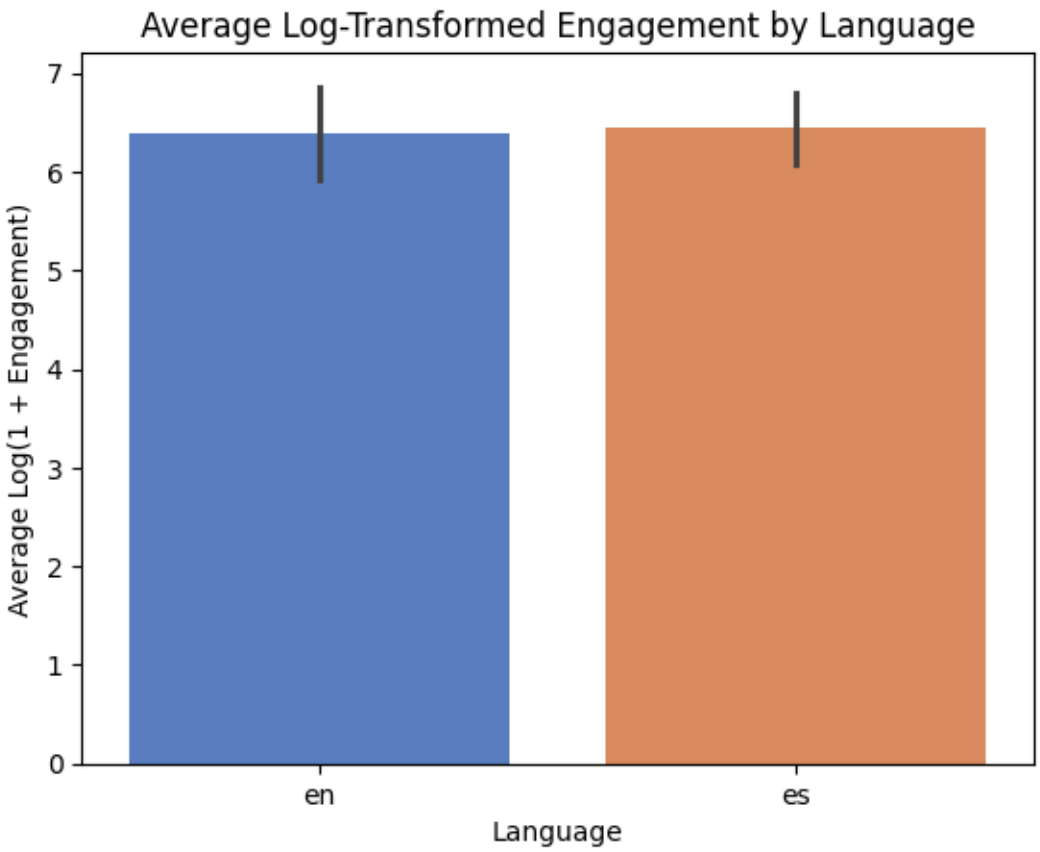
Together, these reveal which language communities were most active and engaging.

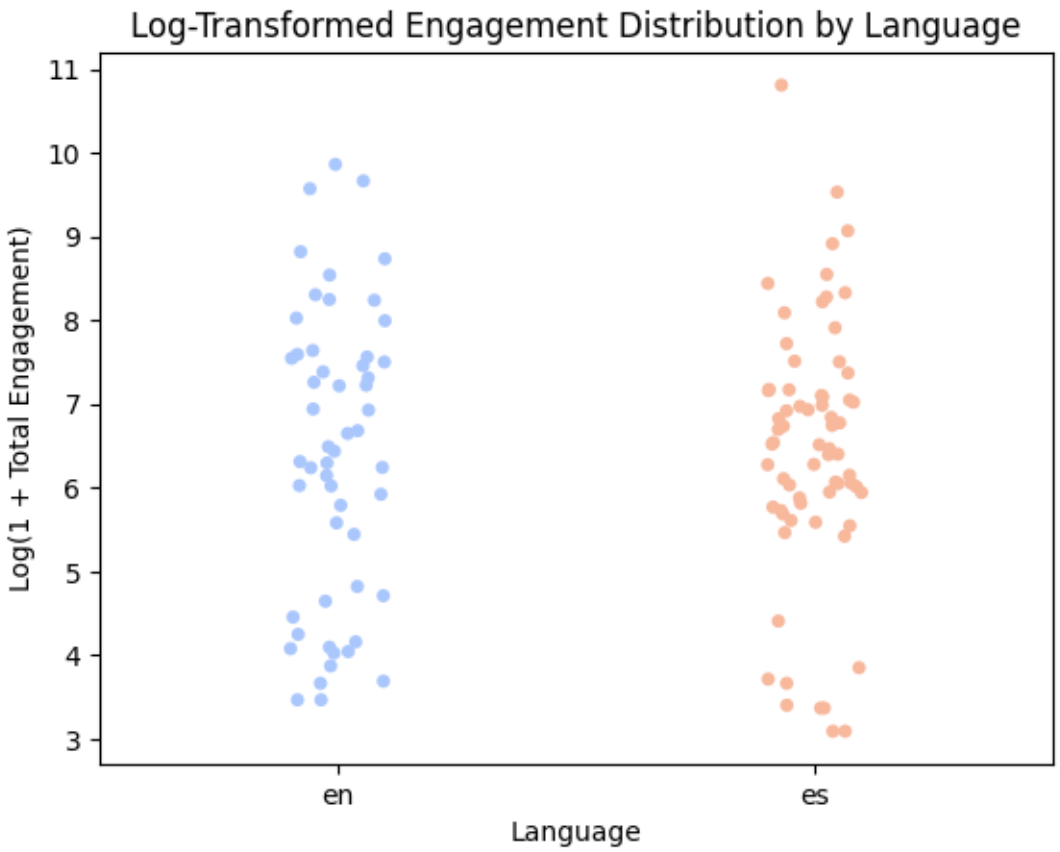
[25]:
```
# ----------------------------------------
# SECTION C: LANGUAGE-BASED COMPARISONS
# ----------------------------------------

# Seaborn Bar Plot: Average Total Engagement by Language
if "Language" in df_selected.columns:
    sns.barplot(x="Language", y="Total_Engagement", data=df_selected,␣
 ↪hue="Language", palette="pastel", legend=False)
    plt.title("Average Engagement by Language")
    plt.xlabel("Language")
    plt.ylabel("Average Engagement")
    plt.show()

# Seaborn Strip Plot: Raw engagement distribution by language
if "Language" in df_selected.columns:
    sns.stripplot(x="Language", y="Total_Engagement", data=df_selected,␣
 ↪hue="Language", palette="Set2", jitter=True, legend=False)
    plt.title("Engagement Distribution by Language")
    plt.xlabel("Language")
    plt.ylabel("Total Engagement")
    plt.show()

# ----------------------------------------
# NEW: Log-Transformed Engagement by Language
```

```
# ----------------------------------------

# Ensure the column is created
df_selected["Log_Total_Engagement"] = np.log1p(df_selected["Total_Engagement"])

# Bar Plot: Average log-transformed engagement by language
if "Language" in df_selected.columns:
    sns.barplot(x="Language", y="Log_Total_Engagement", data=df_selected,␣
 ↪hue="Language", palette="muted", legend=False)
    plt.title("Average Log-Transformed Engagement by Language")
    plt.xlabel("Language")
    plt.ylabel("Average Log(1 + Engagement)")
    plt.show()

# Strip Plot: Log-transformed engagement distribution
if "Language" in df_selected.columns:
    sns.stripplot(x="Language", y="Log_Total_Engagement", data=df_selected,␣
 ↪hue="Language", palette="coolwarm", jitter=True, legend=False)
    plt.title("Log-Transformed Engagement Distribution by Language")
    plt.xlabel("Language")
    plt.ylabel("Log(1 + Total Engagement)")
    plt.show()
```




Engagement Distribution by Language
Total Engagement
50000
40000
30000
20000
10000
0
en
es
Language


Average Log-Transformed Engagement by Language
Average Log(1 + Engagement)
0
1
2
3
4
5
6
7
en
es
Language



1.10.4 Language-Based Engagement Trends

Average Engagement by Language (Raw)

- **English (`en`) and Spanish (`es`) tweets show nearly identical average engagement**, both hovering just above 2,000.
- The **error bars** are wide, indicating significant variability — especially in Spanish tweets, suggesting more extreme highs and lows.
- This implies that while average visibility is similar, **Spanish tweets may contain both very low and extremely high engagement cases**.

Engagement Distribution by Language (Raw)

- The **scatterplot reveals several extreme outliers** — one Spanish tweet with nearly **50,000 engagements** stands out clearly.
- English tweets also show a few with 10k–20k range, but they are less extreme and less frequent.
- Overall, **Spanish tweets have a wider spread**, meaning engagement is less consistent and more polarized compared to English.

Average Log-Transformed Engagement by Language

- After applying a log scale, both languages **converge in average engagement**, with nearly identical bars around `log(1 + 6400)  6.4`.
- The **log transformation smooths the outliers**, allowing us to focus on general trends rather than viral anomalies.
- This suggests that day-to-day engagement levels are **consistently comparable across languages**, despite raw scale differences.

Log-Transformed Engagement Distribution

- Most tweets in both languages **cluster between log values of 5–7**, aligning with engagement between ~150 to ~1,000.
- Spanish tweets again display more upper-bound variability, with **several tweets exceeding log 9 (8,000+ engagements)**.
- The plot shows that **language is not a strong divider for engagement**, though **Spanish tweets exhibit more potential for virality**.

1.10.5 Insight

Raw metrics show Spanish tweets have higher extremes, while English tweets are more evenly distributed. Once log-transformed, both languages reveal nearly identical average engagement, suggesting that **message reach and interaction were similarly effective across both language groups**, but Spa

1.10.6 SECTION D: Relationship Between Features

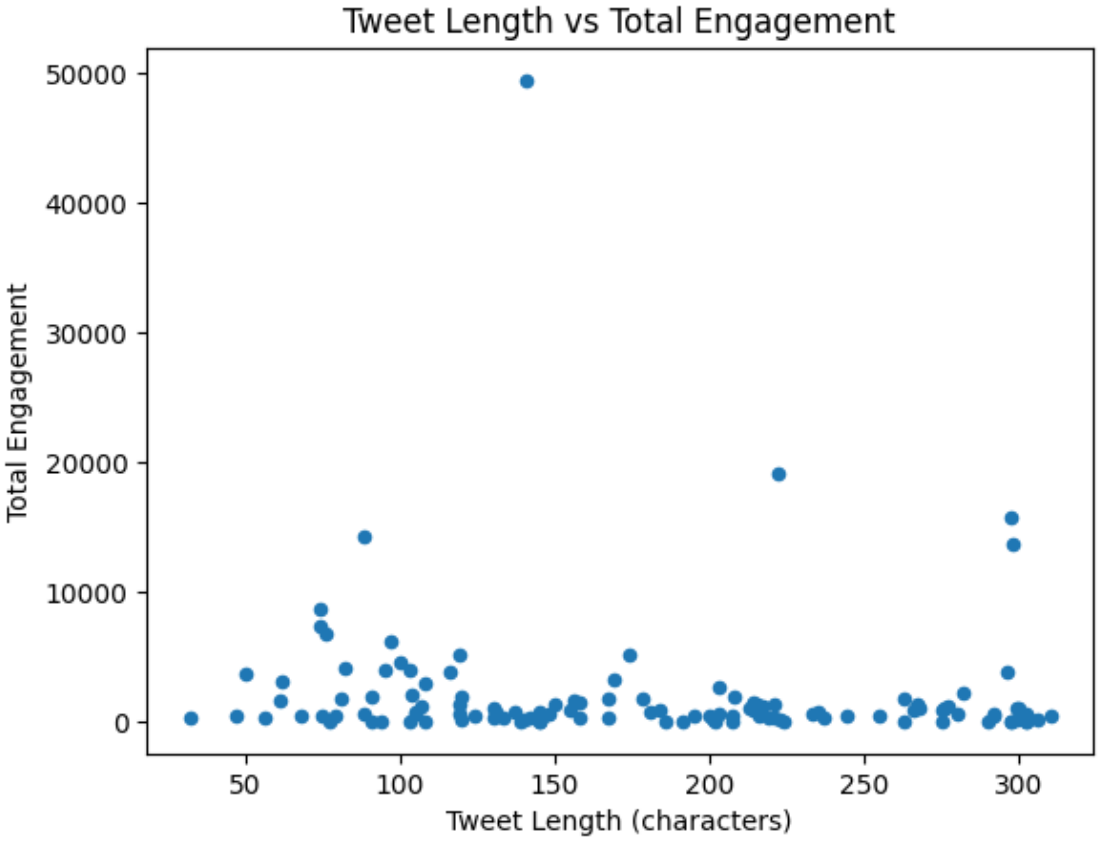
We use a **scatter plot** to test whether there's a relationship between:

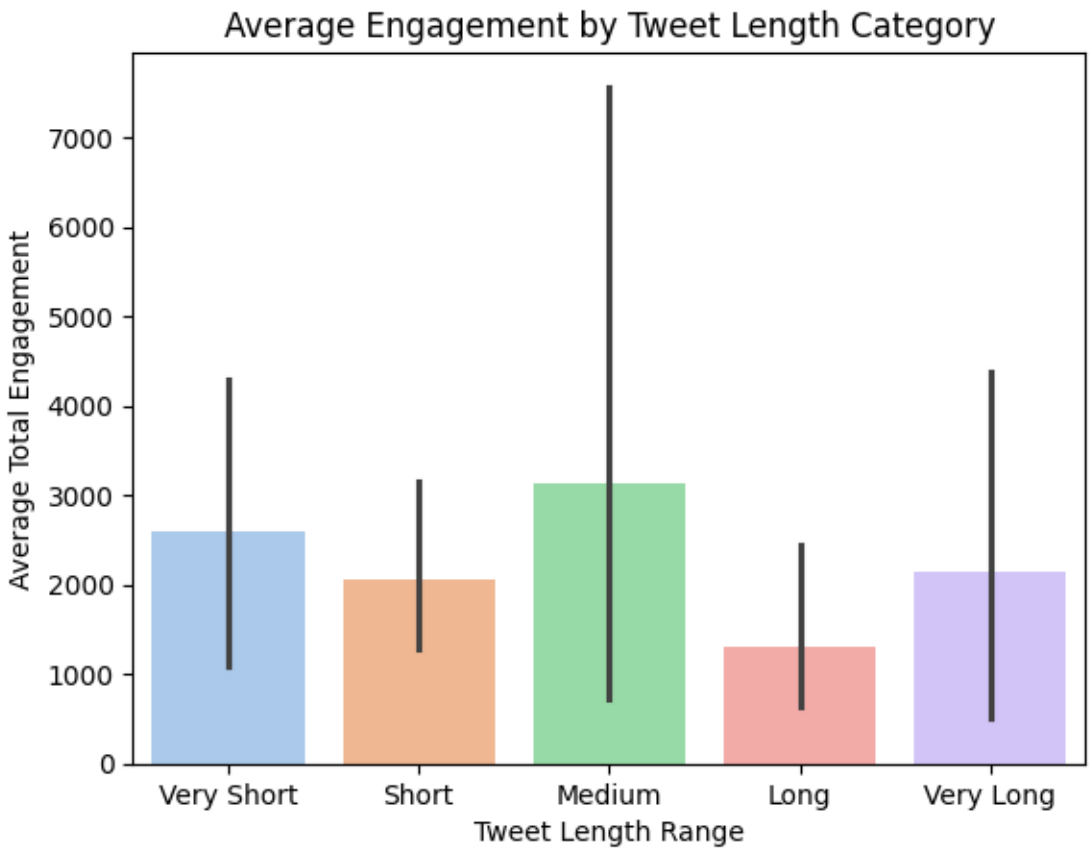
- **Tweet_Length** (x-axis)
- **Total_Engagement** (y-axis)

Helps answer: *Do longer tweets tend to get more attention?*

[26]:
```
# ----------------------------------------
# SECTION D: RELATIONSHIP BETWEEN FEATURES
# ----------------------------------------

# Scatter Plot: Tweet Length vs. Total Engagement
df_selected.plot(kind="scatter", x="Tweet_Length", y="Total_Engagement",␣
 ↪title="Tweet Length vs Total Engagement")
plt.xlabel("Tweet Length (characters)")
plt.ylabel("Total Engagement")
plt.show()

# Bar Plot: Average Engagement by Tweet Length Category
sns.barplot(x="Tweet_Length_Category", y="Total_Engagement",␣
 ↪hue="Tweet_Length_Category", data=df_selected, palette="pastel",␣
 ↪legend=False)
plt.title("Average Engagement by Tweet Length Category")
```

```
plt.xlabel("Tweet Length Range")
plt.ylabel("Average Total Engagement")
plt.show()
```





1.10.7 Relationship Between Tweet Length and Engagement

Tweet Length vs. Total Engagement (Scatter Plot)

- The scatter plot shows that **most tweets cluster at lower engagement levels**, regardless of length.
- However, there are a few **high-engagement outliers** — especially one tweet around **140 characters** that reached **nearly 50,000 interactions**.
- Additional peaks are visible in the 100–250 character range, suggesting that **concise-to-moderate-length tweets** may be optimal for gaining traction during urgent events.
- There's no strict linear relationship, but the spread indicates **engagement is not confined to very long or short tweets**.

Average Engagement by Tweet Length Category (Bar Plot)

- Tweets are grouped into categories like *Very Short*, *Short*, *Medium*, *Long*, and *Very Long*.
- **Medium-length tweets** (likely around 100–160 characters) show the **highest average engagement**, though the error bars are wide, reflecting variability.
- **Very Short** and **Very Long** tweets also perform reasonably well, possibly due to **quick alerts** or **in-depth advisories**.

- **Long tweets (just under the character limit)** have the lowest average engagement, possibly due to reduced readability or clarity during high-urgency moments.

1.10.8 Insight

Engagement appears to peak around **medium-length tweets**, which may strike the right balance between being **concise and informative**. Extremely short or long tweets can also perform well depending on context — such as **short alerts** or **detailed advisories**. However, **tweet length alone isn't a consistent predictor** of engagement; content clarity and timing likely play a larger role.

1.10.9 SECTION E: Correlation Matrix

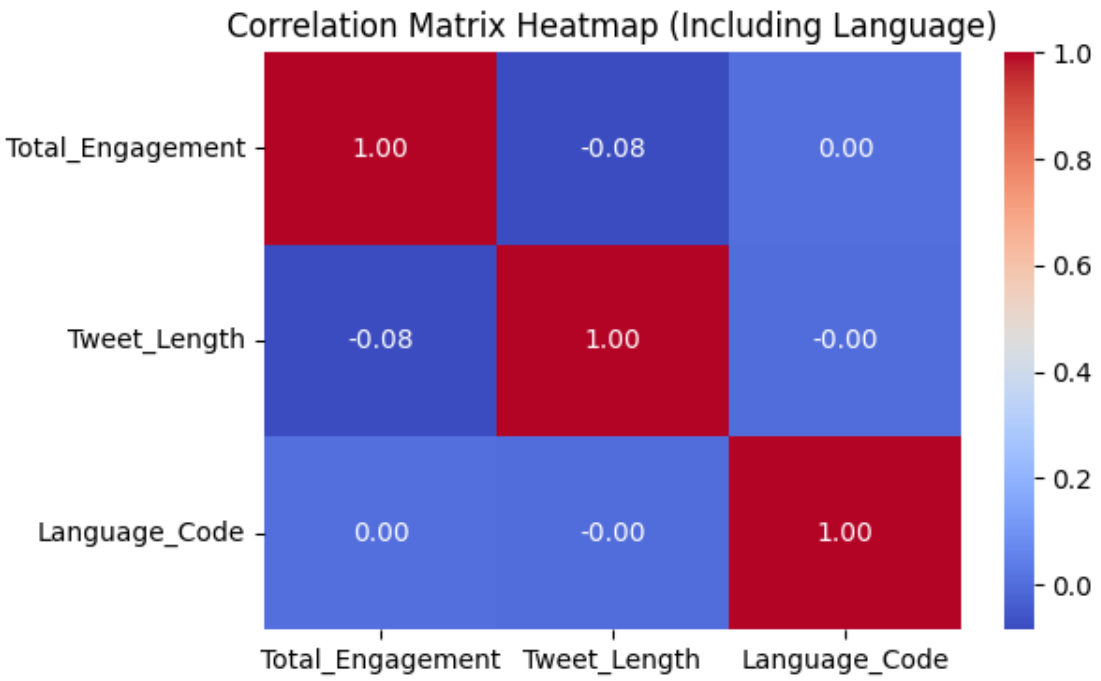
This heatmap shows the **correlation coefficient** between:

- `Total_Engagement`
- `Tweet_Length`

Correlation values range from: - **+1** = strong positive - **0** = no correlation - **−1** = strong negative

This confirms or challenges what we saw in the scatter plot.

[27]:
```
# ----------------------------------------
# SECTION E: CORRELATION MATRIX
# ----------------------------------------

# Heatmap: Correlation between numeric features

# Convert Language to numeric labels for correlation
df_selected["Language_Code"] = df_selected["Language"].map({"en": 0, "es": 1})

# Correlation matrix including language code
corr_matrix = df_selected[["Total_Engagement", "Tweet_Length",␣
 ↪"Language_Code"]].corr()

# Heatmap
plt.figure(figsize=(6, 4))
sns.heatmap(corr_matrix, annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Correlation Matrix Heatmap (Including Language)")
plt.show()
```



1.10.10 Correlation Matrix Insights

This matrix explores the relationships between: - **Total Engagement** - **Tweet Length** - **Language Code** (`en` = 0, `es` = 1)

Total Engagement vs. Tweet Length

- Correlation: **-0.08**
- This is a **very weak negative correlation**, suggesting that as tweet length increases, total engagement **slightly decreases**, but the relationship is not statistically meaningful.
- In practical terms, tweet length **doesn't strongly influence** how much engagement a tweet gets.

Total Engagement vs. Language

- Correlation: **0.00**
- There is effectively **no correlation** between language and total engagement.
- This confirms earlier findings: **both English and Spanish tweets perform similarly on average**, and language alone isn't a driver of engagement.

Tweet Length vs. Language

- Correlation: ~**0.00**
- Again, no meaningful relationship. Tweets in English and Spanish are **similar in length**, with no clear trend toward longer or shorter formats by language.

1.10.11 Insight

The correlation heatmap reveals that **none of the tested features (length or language)** have a strong direct influence on tweet engagement. This supports earlier findings that **context, timing, and content** are more critical drivers of public response than structural attributes like length or language.

1.10.12 SECTION F: Interactive Plotly Visuals

This section recreates key visualizations using **Plotly** for enhanced interactivity and insight. These plots allow:

- **Zooming and panning** to inspect engagement spikes over time
- **Hover tooltips** for precise data exploration
- **Side-by-side comparisons** of raw vs. log-transformed engagement
- Useful for presentations, dashboards, and deep dives into engagement behavior

[28]:
```
# ----------------------------------------
# SECTION F: INTERACTIVE PLOTLY VISUALS
# ----------------------------------------

import plotly.express as px
import plotly.graph_objects as go

# Interactive Line Plot: Total Engagement over Time
fig = px.line(
    df_selected.sort_values("Timestamp_UTC"),
    x="Timestamp_UTC",
    y="Total_Engagement",
    title="  Total Engagement Over Time (Interactive)",
    labels={"Timestamp_UTC": "Time", "Total_Engagement": "Total Engagement"}
)
fig.update_layout(xaxis_title="Time", yaxis_title="Total Engagement")
fig.show()

# Interactive Line Plot: Log-Transformed Engagement over Time
fig = px.line(
    df_selected.sort_values("Timestamp_UTC"),
    x="Timestamp_UTC",
    y="Log_Total_Engagement",
    title="  Log-Transformed Total Engagement Over Time (Interactive)",
    labels={"Timestamp_UTC": "Time", "Log_Total_Engagement": "Log(1 +␣
 ↪Engagement)"}
)
fig.update_layout(xaxis_title="Time", yaxis_title="Log(1 + Total Engagement)")
fig.show()
```

```
# Interactive Histogram: Total Engagement
fig = px.histogram(
    df_selected,
    x="Total_Engagement",
    nbins=30,
    title="  Total Engagement Distribution (Interactive)",
    labels={"Total_Engagement": "Total Engagement"},
)
fig.update_layout(yaxis_title="Tweet Count")
fig.show()

# Interactive Histogram: Log-Transformed Engagement
fig = px.histogram(
    df_selected,
    x="Log_Total_Engagement",
    nbins=30,
    title="  Log-Transformed Total Engagement Distribution (Interactive)",
    labels={"Log_Total_Engagement": "Log(1 + Engagement)"},
)
fig.update_layout(yaxis_title="Tweet Count")
fig.show()
```

1.11 Word Cloud: Frequent Tweet Words

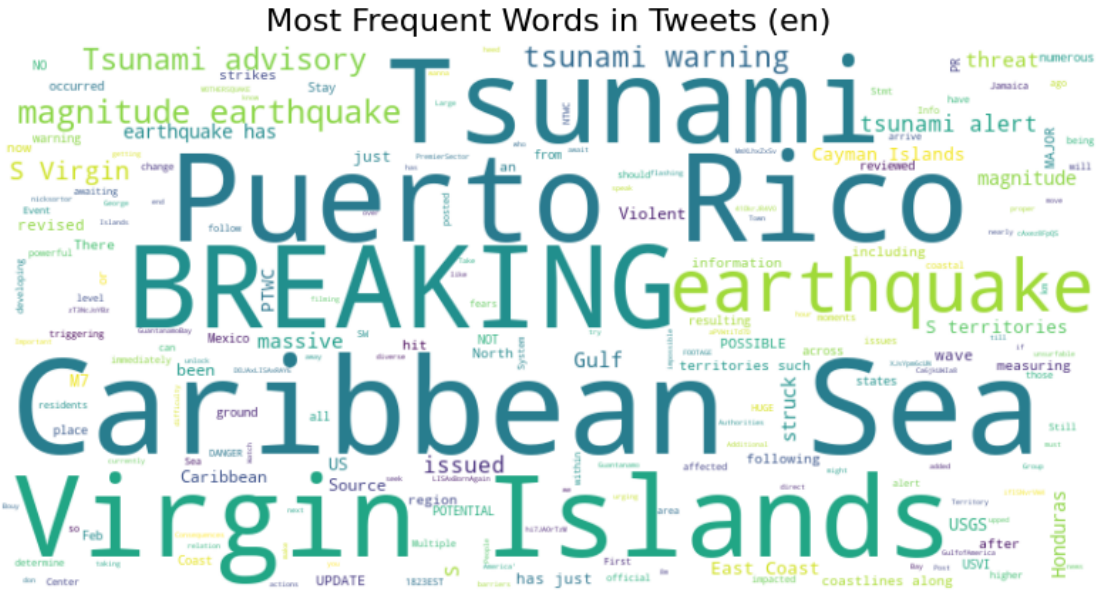
This WordCloud shows the **most common words** across all tweets.

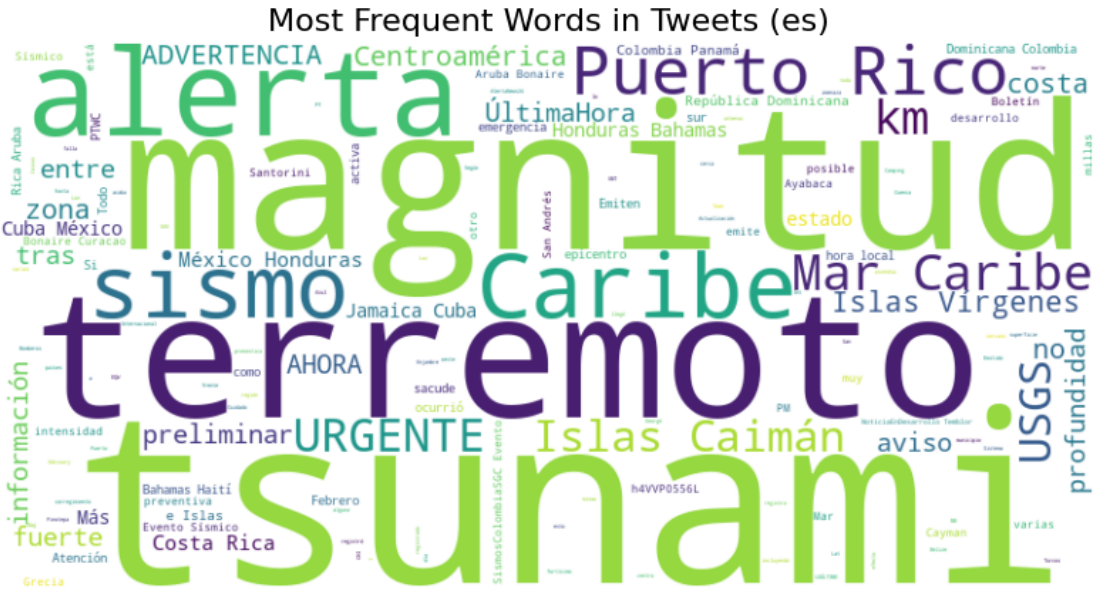
- **Larger words** = more frequent use.
- Useful for spotting **trending topics** or key terms.
- Common **stopwords** (in English and Spanish) and Twitter noise (like `RT`, `https`) were removed for clarity.

A quick visual summary of tweet content in both languages.

[29]:
```
from wordcloud import WordCloud
import matplotlib.pyplot as plt

# Define reusable stopwords set
custom_stopwords = {
    "https", "RT", "co", "amp",
    "de", "a", "t", "el", "que", "se", "la", "en", "por",
    "los", "las", "del", "al", "un", "una", "con", "para",
    "este", "esta", "estos", "estas", "ese", "esa", "esos", "esas",
    "y", "o", "u", "pero", "su", "sus", "porque", "son",
    "ser", "sido", "ha", "han", "hay", "qué", "etc", "PuertoRico",
    "the", "is", "to", "of", "and", "in", "for", "on", "at", "with", "as",␣
 ↪"this", "that", "it", "are",
}
```

```python
# Function to generate and show word cloud for given language
def generate_wordcloud_for_language(language_code):
    text = " ".join(
        df_selected[df_selected["Language"] == language_code]["Tweet_Content"].
 ↪dropna().astype(str)
    )
    wordcloud = WordCloud(
        width=800,
        height=400,
        background_color='white',
        colormap='viridis',
        max_words=200,
        stopwords=custom_stopwords
    ).generate(text)

    plt.figure(figsize=(10, 5))
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis("off")
    plt.title(f"Most Frequent Words in Tweets ({language_code})", fontsize=16)
    plt.show()

# Generate word clouds for English and Spanish
generate_wordcloud_for_language("en")
generate_wordcloud_for_language("es")
```





1.11.1 Top Terms in Tweets by Language

English Word Cloud (`en`)

- **Most prominent terms**:
  `BREAKING`, `Tsunami`, `Caribbean Sea`, `Puerto Rico`, `Virgin Islands`, `earthquake`, `magnitude`
- The language reflects an **urgent, news-style tone**, with emphasis on:
  - **Geographic focus**: `Caribbean`, `Puerto Rico`, `Virgin Islands`, `Cayman Islands`, etc.
  - **Event-driven terms**: `Tsunami`, `earthquake`, `warning`, `advisory`, `USGS`
  - **Communication framing**: Words like `BREAKING`, `UPDATE`, and `POTENTIAL` suggest a **news alert or real-time update format**.
- This points to the use of **English tweets for breaking news, alerts, and official advisories**, likely aimed at both international and regional audiences.

Spanish Word Cloud (`es`)

- **Most prominent terms**:
  `tsunami`, `terremoto`, `magnitud`, `alerta`, `Puerto Rico`, `Caribe`, `sismo`, `URGENTE`
- The dominant tone is **emergency and precision**, with:
  - Strong **scientific/informational focus**: `magnitud`, `USGS`, `profundidad`, `km`, `preliminar`
  - **Alert-oriented language**: `alerta`, `URGENTE`, `advertencia`, `aviso`
  - Geographic terms common to the region, like `Islas Caimán`, `Centroamérica`, `Costa Rica`, etc.
- Spanish tweets appear to emphasize **detailed seismic data and public safety alerts**, aimed at regional users seeking urgent local updates.

1.11.2 Insight

While both English and Spanish tweets center around the **earthquake-tsunami emergency**, their vocabularies reflect different communication styles. English tweets tend to highlight **newsworthiness and urgency**, while Spanish tweets emphasize **technical information and localized alerts**. Together, they showcase a bilingual flow of public communication during the February 9 seismic event.

1.12 Conclusion

This analysis provided a comprehensive overview of the `PR_Advisory_Tweets_Feb_2025.csv` dataset, focusing on tweet engagement, content characteristics, and language patterns. Through systematic data cleaning, feature engineering, and a variety of visual exploration techniques, we gained insights into:

- The distribution and structure of tweet content
- Temporal trends in tweet activity and engagement
- Differences in engagement across languages
- Common themes expressed by users through word frequency

A range of visualization tools—static and interactive—were employed to enhance interpretability and uncover patterns in the data. This approach demonstrates how Python's data science ecosystem can be effectively applied to social media analysis, especially in the context of real-time events like natural disasters.