Reddit Data Cleaning - r/TSLA

```
import pandas as pd
```

```
csv_url = "/content/reddit_tsla_Jan1_Mar12_with_sentiment.csv"

df = pd.read_csv(csv_url)

df_cleaned = df[df['content'] != '[removed]']

df_cleaned.to_csv("reddit_tsla_Jan1_Mar12_with_sentiment_cleaned.csv", index=False)

print("Rows before cleaning:", len(df))
print("Rows after cleaning:", len(df_cleaned))
```

```
Rows before cleaning: 12472
Rows after cleaning: 11241
```

Reddit Data Cleaning - r/tslamotors & r/elonmusk

```
csv_url = "/content/reddit_tslamotors_elonmusk_Jan1_Mar12_with_sentiment.csv"

df = pd.read_csv(csv_url)

df_cleaned = df[df['content'] != '[removed]'] #remving  the deleted comments

df_cleaned.to_csv("/content/reddit_tslamotors_elonmusk_Jan1_Mar12_with_sentiment_cleaned.csv", index=False)

print("Rows before cleaning:", len(df))
print("Rows after cleaning:", len(df_cleaned))
```

```
Rows before cleaning: 26556
Rows after cleaning: 25310
```

X.com Data Cleaning - @elonmusk

```
csv_path = "/content/xcom_elonmusk_Jan1_Mar12_with_sentiment.csv"
# removing "@elonmusk" from text
df = pd.read_csv(csv_path)
print("Rows before cleaning:", len(df))

print("\n--- Content (head) BEFORE cleaning ---")
print(df['content'].head())

df['content'] = (df['content']
    .str.replace('@elonmusk', '', regex=False)
)

print("\nRows after cleaning:", len(df))

print("\n--- Content (head) AFTER cleaning ---")
print(df['content'].head())

df.to_csv("xcom_elonmusk_Jan1_Mar12_with_sentiment_cleaned.csv", index=False)
```

```
Rows before cleaning: 8507

--- Content (head) BEFORE cleaning ---
0    This is why the Democrats want to destroy me\n...
1                               @elonmusk We love Elon
2    @elonmusk @terrymu74847907 It is also that the...
3     @elonmusk Hang in there for the rest of America!
4    @elonmusk Democrats are Globalists.   They are...
Name: content, dtype: object

Rows after cleaning: 8507

--- Content (head) AFTER cleaning ---
0    This is why the Democrats want to destroy me\n...
1                                        We love Elon
```

```
2    @terrymu74847907 It is also that they don't t...
3               Hang in there for the rest of America!
4    Democrats are Globalists.   They are despicab...
Name: content, dtype: object
```

X.com Data Cleaning - @realDonaldTrump, @WhiteHouse, @Trump

```
csv_path = "/content/xcom_trumpWH_Jan1_Mar12_with_sentiment.csv"
# removing "@realDonaldTrump, @Trump, @WhiteHouse" from text
df = pd.read_csv(csv_path)
print("Rows before cleaning:", len(df))

print("\n--- Content (head) BEFORE cleaning ---")
print(df['content'].head())

df['content'] = (df['content']
    .str.replace('@realDonaldTrump', '', regex=False)
    .str.replace('@Trump', '', regex=False)
    .str.replace('@WhiteHouse', '', regex=False)
)

print("\nRows after cleaning:", len(df))

print("\n--- Content (head) AFTER cleaning ---")
print(df['content'].head())

df.to_csv("xcom_trumpWH_Jan1_Mar12_with_sentiment_cleaned.csv", index=False)
```

```
Rows before cleaning: 7968

--- Content (head) BEFORE cleaning ---
0    Our Country is a disaster, a laughing stock al...
1    @realDonaldTrump Close the border! https://t.c...
2    @realDonaldTrump It has been six months and we...
3    @realDonaldTrump The Biden Body Count is highe...
4    @realDonaldTrump On January 20th the Traitors ...
Name: content, dtype: object

Rows after cleaning: 7968

--- Content (head) AFTER cleaning ---
0    Our Country is a disaster, a laughing stock al...
1            Close the border! https://t.co/UqaksgCGQA
2     It has been six months and we still know noth...
3     The Biden Body Count is higher than ANYONE ca...
4     On January 20th the Traitors must go. https:/...
Name: content, dtype: object
```

Fixing the TimeStamp

```
import glob
```

```
files = glob.glob("cleaned_timestamp/*.csv")
for f in files:
    df = pd.read_csv(f)
    print("File:", f)
    print("Before:\n", df['timestamp'].head())
    df['timestamp'] = pd.to_datetime(df['timestamp'], errors='coerce').dt.strftime('%Y-%m-%d %H:%M:%S') #fixing timeformats to just one
    print("After:\n", df['timestamp'].head())
    df.to_csv(f, index=False)
```

```
File: cleaned_timestamp/reddit_tsla_Jan1_Mar12_with_sentiment_cleaned.csv
Before:
 0    1/1/2025 0:55
1    1/1/2025 1:17
2    1/1/2025 1:18
3    1/1/2025 1:34
4    1/1/2025 1:34
Name: timestamp, dtype: object
After:
 0    2025-01-01 00:55:00
1    2025-01-01 01:17:00
2    2025-01-01 01:18:00
3    2025-01-01 01:34:00
4    2025-01-01 01:34:00
```

```
Name: timestamp, dtype: object
File: cleaned_timestamp/xcom_trumpWH_Jan1_Mar12_with_sentiment_cleaned.csv
Before:
 0    Fri Jan 03 05:22:11 +0000 2025
1    Fri Jan 03 05:22:33 +0000 2025
2    Fri Jan 03 05:24:34 +0000 2025
3    Fri Jan 03 05:28:08 +0000 2025
4    Fri Jan 03 05:25:42 +0000 2025
Name: timestamp, dtype: object
<ipython-input-9-3e6dcf12fb61>:6: UserWarning: Could not infer format, so each element will be parsed individually, falling back to `
  df['timestamp'] = pd.to_datetime(df['timestamp'], errors='coerce').dt.strftime('%Y-%m-%d %H:%M:%S')
After:
 0    2025-01-03 05:22:11
1    2025-01-03 05:22:33
2    2025-01-03 05:24:34
3    2025-01-03 05:28:08
4    2025-01-03 05:25:42
Name: timestamp, dtype: object
File: cleaned_timestamp/reddit_tslamotors_elonmusk_Jan1_Mar12_with_sentiment_cleaned.csv
Before:
 0    1/1/2025 0:01
1    1/1/2025 0:02
2    1/1/2025 0:03
3    1/1/2025 0:08
4    1/1/2025 0:10
Name: timestamp, dtype: object
After:
 0    2025-01-01 00:01:00
1    2025-01-01 00:02:00
2    2025-01-01 00:03:00
3    2025-01-01 00:08:00
4    2025-01-01 00:10:00
Name: timestamp, dtype: object
File: cleaned_timestamp/xcom_elonmusk_Jan1_Mar12_with_sentiment_cleaned.csv
Before:
 0    Thu Mar 13 00:40:48 +0000 2025
1    Fri Mar 14 23:30:48 +0000 2025
2    Sat Mar 15 16:32:48 +0000 2025
3    Fri Mar 14 18:55:01 +0000 2025
4    Fri Mar 14 21:58:23 +0000 2025
Name: timestamp, dtype: object
<ipython-input-9-3e6dcf12fb61>:6: UserWarning: Could not infer format, so each element will be parsed individually, falling back to `
  df['timestamp'] = pd.to_datetime(df['timestamp'], errors='coerce').dt.strftime('%Y-%m-%d %H:%M:%S')
```

YouTube Data Cleaning

```
import pandas as pd

f = "/content/cleaned_timestamp_unzipped/youtube_teslaNews_Jan1_Mar12_with_sentiment_cleaned.csv"
df = pd.read_csv(f)
df.loc[df['type'] == 'news', 'type'] = 'post' #changing news with post
df.to_csv(f, index=False)
```

```
import pandas as pd

file_path = "/content/cleaned_timestamp_unzipped/youtube_teslaNews_Jan1_Mar12_with_sentiment_cleaned.csv"

df = pd.read_csv(file_path)
df['sentiment_score'] = pd.to_numeric(df['sentiment_score'], errors='coerce') #removing some rows that had issue
df = df.dropna(subset=['sentiment_score'])
df.to_csv(file_path, index=False)
```

Stats

```
files = glob.glob("cleaned_timestamp/*.csv")
tot_c = 0
tot_p = 0

for f in files:
    df = pd.read_csv(f)
    print("File:", f)
    c = len(df[df['type'].str.lower() == 'comment'])
    p = len(df[df['type'].str.lower().isin(['post','news'])])
    tot_c += c
    tot_p += p
```

```
    df.to_csv(f, index=False)
print("---------------------------------")
print("Total comments:", tot_c)
print("Total posts/news:", tot_p)
print("Combined:", tot_c + tot_p)
print("---------------------------------")
```

```
File: cleaned_timestamp/reddit_tsla_Jan1_Mar12_with_sentiment_cleaned.csv
File: cleaned_timestamp/xcom_trumpWH_Jan1_Mar12_with_sentiment_cleaned.csv
File: cleaned_timestamp/reddit_tslamotors_elonmusk_Jan1_Mar12_with_sentiment_cleaned.csv
File: cleaned_timestamp/xcom_elonmusk_Jan1_Mar12_with_sentiment_cleaned.csv
<ipython-input-10-48e7739ff736>:6: DtypeWarning: Columns (5,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,3
  df = pd.read_csv(f)
File: cleaned_timestamp/youtube_teslaNews_Jan1_Mar12_with_sentiment_cleaned.csv
File: cleaned_timestamp/dailymail_tesla_Jan1_Mar12_with_sentiment_cleaned.csv
---------------------------------
Total comments: 122503
Total posts/news: 1061
Combined: 123564
---------------------------------
```

```
import shutil
shutil.make_archive('cleaned_timestamp', 'zip', 'cleaned_timestamp')
```

```
'/content/cleaned_timestamp.zip'
```