Data Visualization PSET 4

Roshni Vora

2001-02-07

Due 02/07 at 5:00PM Central.

"This submission is my work alone and complies with the 30538 integrity policy." Add your initials to indicate your agreement: **___**

Github Classroom Assignment Setup and Submission Instructions

1. **Accepting and Setting up the PS4 Assignment Repository**
   - Each student must individually accept the repository for the problem set from Github Classroom ("ps4") – https://classroom.github.com/a/hWhtcHqH
     - You will be prompted to select your cnetid from the list in order to link your Github account to your cnetid.
     - If you can't find your cnetid in the link above, click "continue to next step" and accept the assignment, then add your name, cnetid, and Github account to this Google Sheet and we will manually link it: https://rb.gy/9u7fb6
   - If you authenticated and linked your Github account to your device, you should be able to clone your PS4 assignment repository locally.
   - Contents of PS4 assignment repository:
     - `ps4_template.qmd`: this is the Quarto file with the template for the problem set. You will write your answers to the problem set here.
2. **Submission Process**:
   - Knit your completed solution `ps4.qmd` as a pdf `ps4.pdf`.
     - Your submission does not need runnable code. Instead, you will tell us either what code you ran or what output you got.
   - To submit, push `ps4.qmd` and `ps4.pdf` to your PS4 assignment repository. Confirm on Github.com that your work was successfully pushed.

Grading

- You will be graded on what was last pushed to your PS4 assignment repository before the assignment deadline
- Problem sets will be graded for completion as: {missing (0%); - (incomplete, 50%); + (excellent, 100%)}
  - The percent values assigned to each problem denote how long we estimate the problem will take as a share of total time spent on the problem set, not the points they are associated with.
- In order for your submission to be considered complete, you need to push both your `ps4.qmd` and `ps4.pdf` to your repository. Submissions that do not include both files will automatically receive 50% credit.

```
import pandas as pd
import altair as alt
import time
import requests
from bs4 import BeautifulSoup
import pandas as pd
import datetime as dt


import warnings
warnings.filterwarnings('ignore')
alt.renderers.enable("png")
```

```
RendererRegistry.enable('png')
```

Step 1: Develop initial scraper and crawler

```
url = "https://oig.hhs.gov/fraud/enforcement/"

response = requests.get(url)
response.raise_for_status()

text = response.text
```

```
soup = BeautifulSoup(text, 'lxml')
```

```
soup.find_all('<li>')
```

```
[]
```

```
lis = soup.find_all("li", class_='usa-card card--list pep-card--minimal
 ↪  mobile:grid-col-12')
print(len(lis))
```

```
20
```

```
for i, li in enumerate(lis):
    a_tag = li.find("a")
    print(a_tag.get_text(strip=True))
    print("https://oig.hhs.gov" + a_tag.get('href'))
    if i  == 3:
      break
```

```
Houston Transplant Doctor Indicted For Making False Statements In Patients'
Medical Records
https://oig.hhs.gov/fraud/enforcement/houston-transplant-doctor-indicted-for-making-false-sta
MultiCare Health System to Pay Millions to Settle Fraud Case
https://oig.hhs.gov/fraud/enforcement/multicare-health-system-to-pay-millions-to-settle-fraud
Brooklyn Banker Pleads Guilty to Laundering Proceeds of Medicare Fraud for
Transnational Criminal Organization
https://oig.hhs.gov/fraud/enforcement/brooklyn-banker-pleads-guilty-to-laundering-proceeds-of
Delafield Man Sentenced to 18 Months' Imprisonment for Conspiracy to Pay
Health Care Kickbacks
https://oig.hhs.gov/fraud/enforcement/delafield-man-sentenced-to-18-months-imprisonment-for-c
```

```
rows = []

for li in lis:
    a_tag = li.find("a")
    if a_tag is None:
        continue

    title = a_tag.get_text(strip=True)
    link = a_tag.get("href", "")
    if link.startswith("/"):
        link = "https://oig.hhs.gov" + link

    span = li.find("span")
    date = span.get_text(strip=True) if span else ""

    ul = li.find("ul")
    if ul:
        categories = [c.get_text(strip=True) for c in ul.find_all("li")]
        category = ", ".join(categories)
    else:
        category = ""

    rows.append({
```

```
        "title": title,
        "date": date,
        "category": category,
        "link": link
    })

rows
```

```
[{'title': 'Houston Transplant Doctor Indicted For Making False Statements In
Patients' Medical Records',
  'date': 'February 5, 2026',
  'category': 'Criminal and Civil Actions',
  'link':
  'https://oig.hhs.gov/fraud/enforcement/houston-transplant-doctor-indicted-for-making-false
 {'title': 'MultiCare Health System to Pay Millions to Settle Fraud Case',
  'date': 'February 4, 2026',
  'category': 'Criminal and Civil Actions',
  'link':
  'https://oig.hhs.gov/fraud/enforcement/multicare-health-system-to-pay-millions-to-settle-f
 {'title': 'Brooklyn Banker Pleads Guilty to Laundering Proceeds of Medicare
 Fraud for Transnational Criminal Organization',
  'date': 'February 3, 2026',
  'category': 'COVID-19',
  'link':
  'https://oig.hhs.gov/fraud/enforcement/brooklyn-banker-pleads-guilty-to-laundering-proceeds
 {'title': 'Delafield Man Sentenced to 18 Months' Imprisonment for Conspiracy
 to Pay Health Care Kickbacks',
  'date': 'February 3, 2026',
  'category': 'Criminal and Civil Actions',
  'link':
  'https://oig.hhs.gov/fraud/enforcement/delafield-man-sentenced-to-18-months-imprisonment-f
 {'title': 'Former NFL Player Convicted for $197M Medicare Fraud',
  'date': 'February 3, 2026',
  'category': 'Criminal and Civil Actions',
  'link':
  'https://oig.hhs.gov/fraud/enforcement/former-nfl-player-convicted-for-197m-medicare-fraud/
 {'title': 'Attorney General Hanaway Obtains Medicaid Fraud Conviction
 Against Couple Lying About Marital Status To Steal Funds',
  'date': 'February 3, 2026',
  'category': 'State Enforcement Agencies',
  'link':
  'https://oig.hhs.gov/fraud/enforcement/attorney-general-hanaway-obtains-medicaid-fraud-con
```

{'title': "AG's Office Secures Indictments Against Peabody Alcohol and Drug
Counselor and Her Businesses for More Than $850,000 in MassHealth Fraud",
 'date': 'February 2, 2026',
 'category': 'State Enforcement Agencies',
 'link':
 'https://oig.hhs.gov/fraud/enforcement/ags-office-secures-indictments-against-peabody-alcol
{'title': 'Florida Man Pleads Guilty to Conspiracy to Violate the
Anti-Kickback Statute',
 'date': 'January 30, 2026',
 'category': 'Criminal and Civil Actions',
 'link':
 'https://oig.hhs.gov/fraud/enforcement/florida-man-pleads-guilty-to-conspiracy-to-violate-t
{'title': 'Forefront Living Hospice Agreed to Pay $1.9 Million for Allegedly
Violating the Civil Monetary Penalties Law by Submitting Claims for Services
that Identified the Incorrect Provider or Were Performed by Non-enrolled or
Incorrect Providers',
 'date': 'January 30, 2026',
 'category': 'Fraud Self-Disclosures',
 'link':
 'https://oig.hhs.gov/fraud/enforcement/forefront-living-hospice-agreed-to-pay-19-million-fo
{'title': 'Attorney General Jeff Jackson Announces Health Care Fraud
Conviction and Settlement',
 'date': 'January 30, 2026',
 'category': 'State Enforcement Agencies',
 'link':
 'https://oig.hhs.gov/fraud/enforcement/attorney-general-jeff-jackson-announces-health-care-
{'title': 'Yadkinville Woman Sentenced in Connection with Multi-Million
Dollar Medicaid Fraud Scheme',
 'date': 'January 29, 2026',
 'category': 'Criminal and Civil Actions',
 'link':
 'https://oig.hhs.gov/fraud/enforcement/yadkinville-woman-sentenced-in-connection-with-multi
{'title': 'Attorney General Labrador Announces Sentencing of Kootenai County
Woman for Public Assistance Provider Fraud',
 'date': 'January 29, 2026',
 'category': 'State Enforcement Agencies',
 'link':
 'https://oig.hhs.gov/fraud/enforcement/attorney-general-labrador-announces-sentencing-of-ko
{'title': 'Attorney General Hanaway Obtains Medicaid Fraud Conviction For
Services Not Provided',
 'date': 'January 29, 2026',
 'category': 'State Enforcement Agencies',

```
 'link':
 'https://oig.hhs.gov/fraud/enforcement/attorney-general-hanaway-obtains-medicaid-fraud-conv
{'title': 'Holmes Regional Medical Center Agreed to Pay $113,000 for
Allegedly Violating Patient Dumping Statute by Failing to Provide an
Appropriate Medical Screening Examination',
 'date': 'January 28, 2026',
 'category': 'CMP and Affirmative Exclusions, EMTALA/Patient Dumping',
 'link':
 'https://oig.hhs.gov/fraud/enforcement/holmes-regional-medical-center-agreed-to-pay-113000-
{'title': 'Slidell Chiropractor Sentenced for Health Care Fraud',
 'date': 'January 28, 2026',
 'category': 'COVID-19, Criminal and Civil Actions',
 'link':
 'https://oig.hhs.gov/fraud/enforcement/slidell-chiropractor-sentenced-for-health-care-fraud
{'title': 'Repeat Health Care Fraud Offender Sentenced for Defrauding New
Hampshire Medicaid',
 'date': 'January 28, 2026',
 'category': 'Criminal and Civil Actions',
 'link':
 'https://oig.hhs.gov/fraud/enforcement/repeat-health-care-fraud-offender-sentenced-for-defr
{'title': 'Scranton Heart Institute Agrees To Pay $48,709.20 To Settle False
Claims Act Allegations',
 'date': 'January 28, 2026',
 'category': 'Criminal and Civil Actions',
 'link':
 'https://oig.hhs.gov/fraud/enforcement/scranton-heart-institute-agrees-to-pay-4870920-to-se
{'title': 'Rheumatologist Agrees To Resolve False Claims Act Allegations
Related To Unapproved Drugs',
 'date': 'January 28, 2026',
 'category': 'Criminal and Civil Actions',
 'link':
 'https://oig.hhs.gov/fraud/enforcement/rheumatologist-agrees-to-resolve-false-claims-act-al
{'title': 'Attorney General James Uthmeier Announces Arrests in Central
Florida Medicaid Fraud Scheme',
 'date': 'January 28, 2026',
 'category': 'State Enforcement Agencies',
 'link':
 'https://oig.hhs.gov/fraud/enforcement/attorney-general-james-uthmeier-announces-arrests-in
{'title': 'Cordell Memorial Hospital Agreed to Pay $40,000 for Allegedly
Violating Patient Dumping Statute by Failing to Provide an Appropriate
Medical Screening Examination',
 'date': 'January 27, 2026',
 'category': 'CMP and Affirmative Exclusions, EMTALA/Patient Dumping',
```

```
  'link':
  'https://oig.hhs.gov/fraud/enforcement/cordell-memorial-hospital-agreed-to-pay-40000-for-al
```

```
df_page1= pd.DataFrame(rows)
df_page1
```

	title	date	category
0	Houston Transplant Doctor Indicted For Making ...	February 5, 2026	Criminal and Civil Actions
1	MultiCare Health System to Pay Millions to Set...	February 4, 2026	Criminal and Civil Actions
2	Brooklyn Banker Pleads Guilty to Laundering Pr...	February 3, 2026	COVID-19
3	Delafield Man Sentenced to 18 Months' Imprison...	February 3, 2026	Criminal and Civil Actions
4	Former NFL Player Convicted for $197M Medicare...	February 3, 2026	Criminal and Civil Actions
5	Attorney General Hanaway Obtains Medicaid Frau...	February 3, 2026	State Enforcement Agencies
6	AG's Office Secures Indictments Against Peabod...	February 2, 2026	State Enforcement Agencies
7	Florida Man Pleads Guilty to Conspiracy to Vio...	January 30, 2026	Criminal and Civil Actions
8	Forefront Living Hospice Agreed to Pay $1.9 Mi...	January 30, 2026	Fraud Self-Disclosures
9	Attorney General Jeff Jackson Announces Health...	January 30, 2026	State Enforcement Agencies
10	Yadkinville Woman Sentenced in Connection with...	January 29, 2026	Criminal and Civil Actions
11	Attorney General Labrador Announces Sentencing...	January 29, 2026	State Enforcement Agencies
12	Attorney General Hanaway Obtains Medicaid Frau...	January 29, 2026	State Enforcement Agencies
13	Holmes Regional Medical Center Agreed to Pay $...	January 28, 2026	CMP and Affirmative Exclusio
14	Slidell Chiropractor Sentenced for Health Care...	January 28, 2026	COVID-19, Criminal and Civil
15	Repeat Health Care Fraud Offender Sentenced fo...	January 28, 2026	Criminal and Civil Actions
16	Scranton Heart Institute Agrees To Pay $48,709...	January 28, 2026	Criminal and Civil Actions
17	Rheumatologist Agrees To Resolve False Claims ...	January 28, 2026	Criminal and Civil Actions
18	Attorney General James Uthmeier Announces Arre...	January 28, 2026	State Enforcement Agencies
19	Cordell Memorial Hospital Agreed to Pay $40,00...	January 27, 2026	CMP and Affirmative Exclusio

Step 2: Making the scraper dynamic

1. Turning the scraper into a function

- a. Pseudo-Code FUNCTION (year, month, run_scraper) if run_scraper == False: LOAD enforcement_actions_startyear_startmonth.csv RETURN dataframe

IF start_year < 13: print "please enter a year after 2013" break

IF start_year >= 2013 run_scraper should loop: title date url category put into csv

break when csv stops compiling data

For this loop we will be using a nested loop with a while loop that will keep compiling data into the csv until there is no more data to put in the csv meaning the scraper is on the last page * b. Create Dynamic Scraper

In my final dataframe i got 3377 observations. the earliest date it scraped is

```
BASE = "https://oig.hhs.gov"
START_URL = "https://oig.hhs.gov/fraud/enforcement/"

def scrape_enforcement_actions(start_year, start_month, run_scraper=True):
    out_csv = f"enforcement_actions_{start_year}_{start_month:02d}.csv"

    if not run_scraper:
        return pd.read_csv(out_csv)

    if start_year < 2013:
        print("Please enter a year >= 2013 (only enforcement actions after
         ↪  2013 are listed).")
        return None

    cutoff = dt.date(start_year, start_month, 1)
    rows = []
    page = 1
    stop = False

    while not stop:
        url = f"{START_URL}?page={page}"
        response = requests.get(url)
        response.raise_for_status()
        soup = BeautifulSoup(response.text, "lxml")

        action_links = soup.select('h2 a[href^="/fraud/enforcement/"]')

        if len(action_links) == 0:
            break

        for a in action_links:
            title = a.get_text(strip=True)
            href = a.get("href", "")
            if not title or href == "/fraud/enforcement/":
                continue

            link = BASE + href
            li = a.find_parent("li")
```

```
            lines = [t.strip() for t in li.get_text("\n",
↪  strip=True).split("\n") if t.strip()] if li else []
            date_str = lines[1] if len(lines) > 1 else ""

            parsed = None
            for fmt in ("%B %d, %Y", "%b %d, %Y"):
                try:
                    parsed = dt.datetime.strptime(date_str, fmt).date()
                    break
                except ValueError:
                    pass

            if parsed and parsed < cutoff:
                stop = True
                break

            category = ""
            if li:
                ul = li.find("ul")
                if ul:
                    categories = [c.get_text(strip=True) for c in
↪  ul.find_all("li")]
                    category = ", ".join(categories)

            rows.append({
                "title": title,
                "date": date_str,
                "category": category,
                "link": link
            })

        if not stop:
            time.sleep(1)
            page += 1

    df = pd.DataFrame(rows).drop_duplicates(subset=["title", "date",
↪  "link"]).reset_index(drop=True)
    df.to_csv(out_csv, index=False)
    return df
```

- c. Test Your Code

#RUN_SCRAPER = True #df_all = scrape_enforcement_actions(2022, 1, run_scraper=RUN_SCRAPER) #df_all.shape #df_all.head() #df_all.tail()

```
RUN_SCRAPER = False
df_2024 = scrape_enforcement_actions(2024, 1, run_scraper=RUN_SCRAPER)
```

```
earliest_date = df_2024['date'].min()
earliest_date
```

```
'April 1, 2024'
```

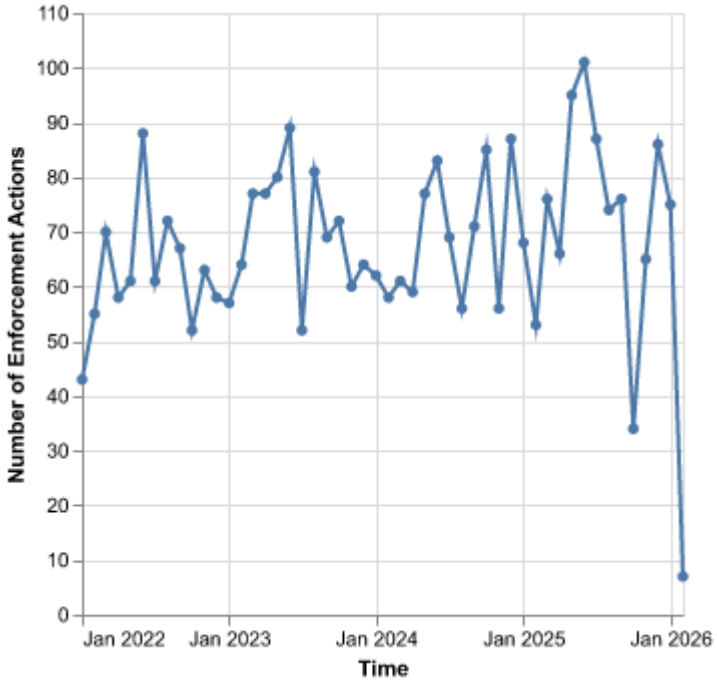
the earliest date is april 1 2024

Step 3: Plot data based on scraped data

1. Plot the number of enforcement actions over time

```
df_all = df = pd.read_csv("enforcement_actions_2022_01.csv")
```

```
enforcement_over_time = (
    alt.Chart(df_all)
    .mark_line(point=True)
    .encode(
        x=alt.X('yearmonth(date):T', title='Time'),
        y=alt.Y('count():Q', title='Number of Enforcement Actions')
    )
)

enforcement_over_time
```

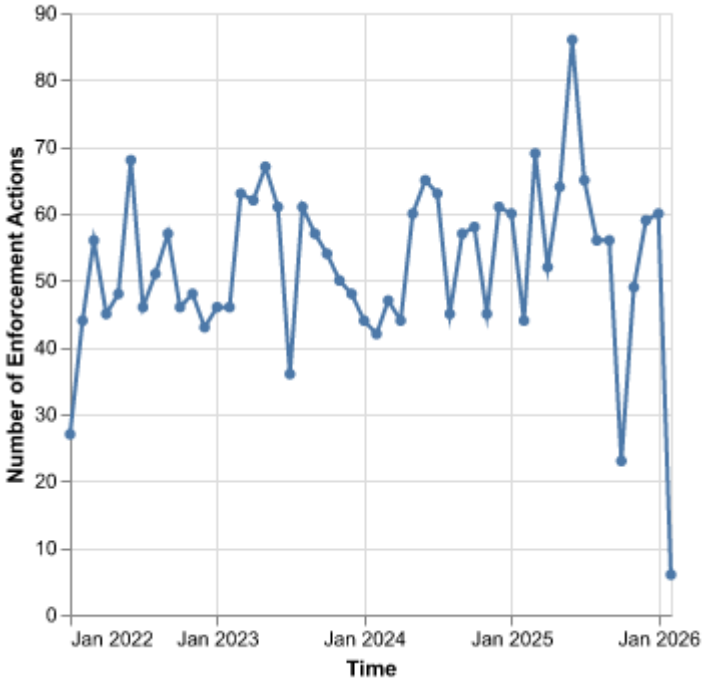


2. Plot the number of enforcement actions categorized:

- based on "Criminal and Civil Actions" vs. "State Enforcement Agencies"

```
df_split = df_all[df_all['category'].isin([
    'Criminal and Civil Actions',
    'State Enforcement Agencies'
])]
```

```
split_enforcement_over_time = (
    alt.Chart(df_split)
    .mark_line(point=True)
    .encode(
        x=alt.X('yearmonth(date):T', title='Time'),
        y=alt.Y('count():Q', title='Number of Enforcement Actions')
    )
)

split_enforcement_over_time
```



- based on five topics

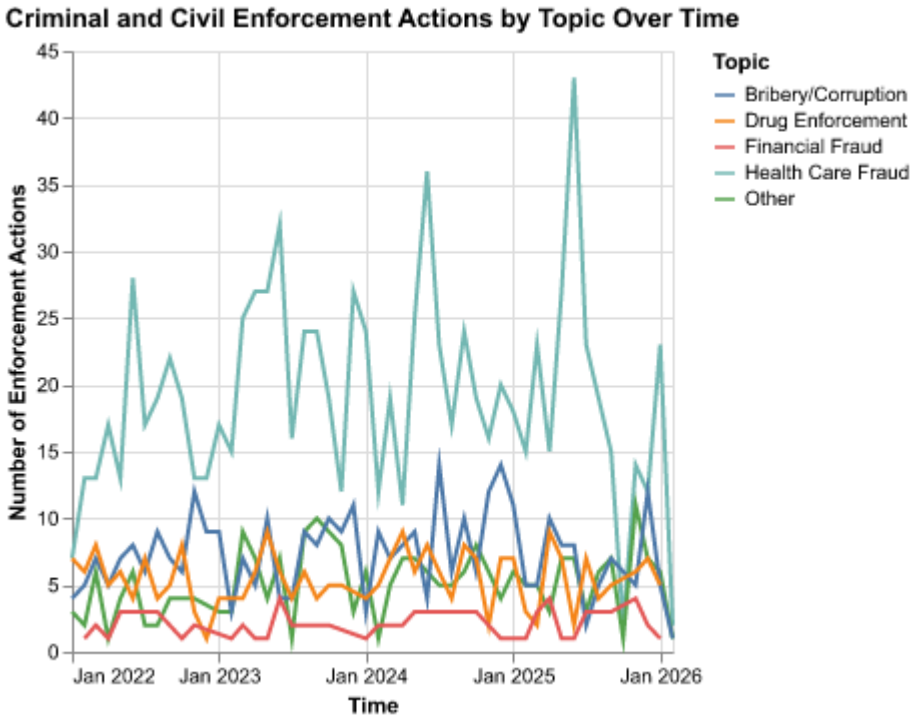
```
health_care_keywords = [
    "medicare", "medicaid", "health care", "healthcare", "physician",
    ↪  "doctor",
    "clinic", "hospital", "nursing", "home health", "billing", "claims"
]

financial_fraud_keywords = [
    "bank", "financial", "loan", "mortgage", "wire fraud", "securities",
    "money laundering", "launder", "tax fraud", "insurance fraud",
    ↪  "investment"
]

drug_enforcement_keywords = [
    "drug", "opioid", "fentanyl", "controlled substance", "prescription",
    "pharmacy", "dea", "traffick", "distribution"
]

bribery_corruption_keywords = [
    "bribe", "bribery", "corrupt", "kickback", "anti-kickback",
    "embezzl", "extort", "payoff", "conspiracy"
]
```

```
df_cca = df_all[
    df_all['category'].str.contains('Criminal and Civil Actions', na=False)
].copy()

df_cca['date'] = pd.to_datetime(df_cca['date'], errors='coerce')
titles = df_cca['title'].str.lower()
```

```
df_cca['topic'] = 'Other'  # default

df_cca.loc[
    titles.str.contains('|'.join(health_care_keywords), regex=True),
    'topic'
] = 'Health Care Fraud'

df_cca.loc[
    titles.str.contains('|'.join(financial_fraud_keywords), regex=True),
    'topic'
] = 'Financial Fraud'

df_cca.loc[
    titles.str.contains('|'.join(drug_enforcement_keywords), regex=True),
    'topic'
] = 'Drug Enforcement'

df_cca.loc[
    titles.str.contains('|'.join(bribery_corruption_keywords), regex=True),
    'topic'
] = 'Bribery/Corruption'
```

```
topic_chart = (
    alt.Chart(df_cca)
    .mark_line()
    .encode(
        x=alt.X('yearmonth(date):T', title='Time'),
        y=alt.Y('count():Q', title='Number of Enforcement Actions'),
        color=alt.Color('topic:N', title='Topic')
    )
    .properties(title='Criminal and Civil Enforcement Actions by Topic Over
    ↪ Time')
)
```

```
topic_chart
```


Criminal and Civil Enforcement Actions by Topic Over Time
Number of Enforcement Actions
45
40
35
30
25
20
15
10
5
0
Jan 2022
Jan 2023
Jan 2024
Jan 2025
Jan 2026
Time
Topic
Bribery/Corruption
Drug Enforcement
Financial Fraud
Health Care Fraud
Other