Group 11 Project EDA

By: Ethan Song, Raksha Sen, Isabella Nance, Pooja Jain

In [1]:

```
!pip3 install folium --upgrade
!pip3 install matplotlib
!pip3 install pandas
!pip3 install contextily
!pip3 install geopandas
```

```
Requirement already satisfied: folium in /opt/anaconda3/lib/python3.9/site-packages (0.14.0)
Requirement already satisfied: numpy in /opt/anaconda3/lib/python3.9/site-packages (from folium) (1.21.5)
Requirement already satisfied: requests in /opt/anaconda3/lib/python3.9/site-packages (from folium) (2.28.1)
Requirement already satisfied: jinja2>=2.9 in /opt/anaconda3/lib/python3.9/site-packages (from folium) (2.11.3)
Requirement already satisfied: branca>=0.6.0 in /opt/anaconda3/lib/python3.9/site-packages (from folium) (0.6.0)
Requirement already satisfied: MarkupSafe>=0.23 in /opt/anaconda3/lib/python3.9/site-packages (from jinja2>=2.9->foliu
m) (2.0.1)
Requirement already satisfied: certifi>=2017.4.17 in /opt/anaconda3/lib/python3.9/site-packages (from requests->folium)
(2022.9.24)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /opt/anaconda3/lib/python3.9/site-packages (from requests->foli
um) (1.26.11)
Requirement already satisfied: charset-normalizer<3,>=2 in /opt/anaconda3/lib/python3.9/site-packages (from requests->f
olium) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in /opt/anaconda3/lib/python3.9/site-packages (from requests->folium) (3.3)
Requirement already satisfied: matplotlib in /opt/anaconda3/lib/python3.9/site-packages (3.5.2)
Requirement already satisfied: numpy>=1.17 in /opt/anaconda3/lib/python3.9/site-packages (from matplotlib) (1.21.5)
Requirement already satisfied: pillow>=6.2.0 in /opt/anaconda3/lib/python3.9/site-packages (from matplotlib) (9.2.0)
Requirement already satisfied: fonttools>=4.22.0 in /opt/anaconda3/lib/python3.9/site-packages (from matplotlib) (4.25.
0)
Requirement already satisfied: python-dateutil>=2.7 in /opt/anaconda3/lib/python3.9/site-packages (from matplotlib) (2.
8.2)
Requirement already satisfied: cycler>=0.10 in /opt/anaconda3/lib/python3.9/site-packages (from matplotlib) (0.11.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /opt/anaconda3/lib/python3.9/site-packages (from matplotlib) (1.4.
2)
Requirement already satisfied: packaging>=20.0 in /opt/anaconda3/lib/python3.9/site-packages (from matplotlib) (21.3)
Requirement already satisfied: pyparsing>=2.2.1 in /opt/anaconda3/lib/python3.9/site-packages (from matplotlib) (3.0.9)
Requirement already satisfied: six>=1.5 in /opt/anaconda3/lib/python3.9/site-packages (from python-dateutil>=2.7->matpl
otlib) (1.16.0)
Requirement already satisfied: pandas in /opt/anaconda3/lib/python3.9/site-packages (1.4.4)
Requirement already satisfied: python-dateutil>=2.8.1 in /opt/anaconda3/lib/python3.9/site-packages (from pandas) (2.8.
2)
Requirement already satisfied: pytz>=2020.1 in /opt/anaconda3/lib/python3.9/site-packages (from pandas) (2022.1)
Requirement already satisfied: numpy>=1.20.0 in /opt/anaconda3/lib/python3.9/site-packages (from pandas) (1.21.5)
Requirement already satisfied: six>=1.5 in /opt/anaconda3/lib/python3.9/site-packages (from python-dateutil>=2.8.1->pan
das) (1.16.0)
Requirement already satisfied: contextily in /opt/anaconda3/lib/python3.9/site-packages (1.3.0)
Requirement already satisfied: joblib in /opt/anaconda3/lib/python3.9/site-packages (from contextily) (1.1.0)
Requirement already satisfied: matplotlib in /opt/anaconda3/lib/python3.9/site-packages (from contextily) (3.5.2)
Requirement already satisfied: requests in /opt/anaconda3/lib/python3.9/site-packages (from contextily) (2.28.1)
Requirement already satisfied: rasterio in /opt/anaconda3/lib/python3.9/site-packages (from contextily) (1.3.6)
Requirement already satisfied: mercantile in /opt/anaconda3/lib/python3.9/site-packages (from contextily) (1.2.1)
Requirement already satisfied: pillow in /opt/anaconda3/lib/python3.9/site-packages (from contextily) (9.2.0)
Requirement already satisfied: geopy in /opt/anaconda3/lib/python3.9/site-packages (from contextily) (2.3.0)
Requirement already satisfied: xyzservices in /opt/anaconda3/lib/python3.9/site-packages (from contextily) (2023.2.0)
Requirement already satisfied: geographiclib<3,>=1.52 in /opt/anaconda3/lib/python3.9/site-packages (from geopy->contex
tily) (2.0)
Requirement already satisfied: packaging>=20.0 in /opt/anaconda3/lib/python3.9/site-packages (from matplotlib->contexti
ly) (21.3)
Requirement already satisfied: python-dateutil>=2.7 in /opt/anaconda3/lib/python3.9/site-packages (from matplotlib->con
textily) (2.8.2)
Requirement already satisfied: numpy>=1.17 in /opt/anaconda3/lib/python3.9/site-packages (from matplotlib->contextily)
(1.21.5)
Requirement already satisfied: fonttools>=4.22.0 in /opt/anaconda3/lib/python3.9/site-packages (from matplotlib->contex
tily) (4.25.0)
Requirement already satisfied: pyparsing>=2.2.1 in /opt/anaconda3/lib/python3.9/site-packages (from matplotlib->context
ily) (3.0.9)
Requirement already satisfied: kiwisolver>=1.0.1 in /opt/anaconda3/lib/python3.9/site-packages (from matplotlib->contex
tily) (1.4.2)
Requirement already satisfied: cycler>=0.10 in /opt/anaconda3/lib/python3.9/site-packages (from matplotlib->contextily)
(0.11.0)
Requirement already satisfied: click>=3.0 in /opt/anaconda3/lib/python3.9/site-packages (from mercantile->contextily)
(8.0.4)
Requirement already satisfied: click-plugins in /opt/anaconda3/lib/python3.9/site-packages (from rasterio->contextily)
(1.1.1)
Requirement already satisfied: certifi in /opt/anaconda3/lib/python3.9/site-packages (from rasterio->contextily) (2022.
9.24)
Requirement already satisfied: affine in /opt/anaconda3/lib/python3.9/site-packages (from rasterio->contextily) (2.4.0)
Requirement already satisfied: snuggs>=1.4.1 in /opt/anaconda3/lib/python3.9/site-packages (from rasterio->contextily)
(1.4.7)
Requirement already satisfied: setuptools in /opt/anaconda3/lib/python3.9/site-packages (from rasterio->contextily) (6
3.4.1)
Requirement already satisfied: attrs in /opt/anaconda3/lib/python3.9/site-packages (from rasterio->contextily) (21.4.0)
Requirement already satisfied: cligj>=0.5 in /opt/anaconda3/lib/python3.9/site-packages (from rasterio->contextily) (0.
7.2)
Requirement already satisfied: idna<4,>=2.5 in /opt/anaconda3/lib/python3.9/site-packages (from requests->contextily)
(3.3)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /opt/anaconda3/lib/python3.9/site-packages (from requests->cont
extily) (1.26.11)
Requirement already satisfied: charset-normalizer<3,>=2 in /opt/anaconda3/lib/python3.9/site-packages (from requests->c
ontextily) (2.0.4)
Requirement already satisfied: six>=1.5 in /opt/anaconda3/lib/python3.9/site-packages (from python-dateutil>=2.7->matpl
otlib->contextily) (1.16.0)
Requirement already satisfied: geopandas in /opt/anaconda3/lib/python3.9/site-packages (0.12.2)
Requirement already satisfied: pyproj>=2.6.1.post1 in /opt/anaconda3/lib/python3.9/site-packages (from geopandas) (3.5.
Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js
Requirement already satisfied: packaging in /opt/anaconda3/lib/python3.9/site-packages (from geopandas) (21.3)
Requirement already satisfied: pandas>=1.0.0 in /opt/anaconda3/lib/python3.9/site-packages (from geopandas) (1.4.4)
Requirement already satisfied: shapely>=1.7 in /opt/anaconda3/lib/python3.9/site-packages (from geopandas) (2.0.1)
Requirement already satisfied: fiona>=1.8 in /opt/anaconda3/lib/python3.9/site-packages (from geopandas) (1.9.3)
Requirement already satisfied: munch>=2.3.2 in /opt/anaconda3/lib/python3.9/site-packages (from fiona>=1.8->geopandas)
(2.5.0)
Requirement already satisfied: importlib-metadata in /opt/anaconda3/lib/python3.9/site-packages (from fiona>=1.8->geopa
ndas) (4.11.3)
Requirement already satisfied: click-plugins>=1.0 in /opt/anaconda3/lib/python3.9/site-packages (from fiona>=1.8->geopa
ndas) (1.1.1)
Requirement already satisfied: click~=8.0 in /opt/anaconda3/lib/python3.9/site-packages (from fiona>=1.8->geopandas)
(8.0.4)
Requirement already satisfied: certifi in /opt/anaconda3/lib/python3.9/site-packages (from fiona>=1.8->geopandas) (202
2.9.24)
Requirement already satisfied: cligj>=0.5 in /opt/anaconda3/lib/python3.9/site-packages (from fiona>=1.8->geopandas)
(0.7.2)
Requirement already satisfied: attrs>=19.2.0 in /opt/anaconda3/lib/python3.9/site-packages (from fiona>=1.8->geopandas)
(21.4.0)
Requirement already satisfied: python-dateutil>=2.8.1 in /opt/anaconda3/lib/python3.9/site-packages (from pandas>=1.0.0
->geopandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /opt/anaconda3/lib/python3.9/site-packages (from pandas>=1.0.0->geopanda
s) (2022.1)
Requirement already satisfied: numpy>=1.20.0 in /opt/anaconda3/lib/python3.9/site-packages (from pandas>=1.0.0->geopand
as) (1.21.5)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /opt/anaconda3/lib/python3.9/site-packages (from packaging->
geopandas) (3.0.9)
Requirement already satisfied: six in /opt/anaconda3/lib/python3.9/site-packages (from munch>=2.3.2->fiona>=1.8->geopan
das) (1.16.0)
Requirement already satisfied: zipp>=0.5 in /opt/anaconda3/lib/python3.9/site-packages (from importlib-metadata->fiona>
=1.8->geopandas) (3.8.0)
```

In [2]:

```
#from datascience import *
import matplotlib
%matplotlib inline
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import geopandas as gpd
import folium
import json
import os
from branca.colormap import linear
import branca.colormap
import requests
import chardet
import requests
from io import StringIO
```

Ethan's EDA Code

Import Relevant Datasets. We will be using publically available data from the California Department of Education. Key datasets we will examine include expulsion data, absence data, as well as demographic data, all based on schools. As for our research question: **"what factors are associated with student expulsions in California schools?"**, we will analyze expulsion school by school.

NOTE: Keep in mind that this data is re-downloaded at every runtime, so if the websites are down, the tables may not be imported properly.

In [3]:

```
expulsion_data_url = 'https://www3.cde.ca.gov/demo-downloads/discipline/expulsion22-v3.txt'
absence_data_url = 'https://www3.cde.ca.gov/demo-downloads/attendance/chronicabsenteeism22-v2.txt'
enrollment_data_url = 'https://dq.cde.ca.gov/dataquest/dlfile/dlfile.aspx?cLevel=School&cYear=2022-23&cCat=Enrollment&cl

response_expulsion = requests.get(expulsion_data_url)
response_expulsion.raise_for_status()

response_absence = requests.get(absence_data_url)
response_absence.raise_for_status()

expulsion_data = pd.read_csv(StringIO(response_expulsion.text), sep='\t')
absence_data = pd.read_csv(StringIO(response_absence.text), sep='\t')
```

In [4]:

```
# expulsion_data_path = os.path.join('data', 'expulsion22-v3-csv.csv')
# expulsion_data = pd.read_csv(expulsion_data_path)
# absence_data = pd.read_csv('data/chronicabsenteeism22.csv')
# enroll_data = pd.read_csv('data/enr21.csv')
```

Clean up expulsion data by dealing with non-numeric '*'s that may appear, as well as ensuring numeric data is the correct data type (float or int). Finally, we change cryptic codes such as 'RB' and 'RI' to more recognizable and easily interpretable names.

In [5]:

```
expulsion_data_s = expulsion_data[expulsion_data['AggregateLevel']=='S']
                                            inplace=True)
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

```python
expulsion_data_s_ta = expulsion_data_s[expulsion_data_s['ReportingCategory'] == 'TA']

columns_to_convert = [
    'Total Expulsions',
    'Unduplicated Count of Students Expelled (Total)',
    'Unduplicated Count of Students Expelled (Defiance-Only)',
    'Expulsion Rate (Total)',
    'Expulsion Count Violent Incident (Injury)',
    'Expulsion Count Violent Incident (No Injury)',
    'Expulsion Count Weapons Possession',
    'Expulsion Count Illicit Drug-Related',
    'Expulsion Count Defiance-Only',
    'Expulsion Count Other Reasons'
]

for column in columns_to_convert:
    expulsion_data_s[column] = expulsion_data_s[column].astype(float)

replacement_dict_expulsions = {
    'RB': 'African American',
    'RI': 'American Indian or Alaska Native',
    'RA': 'Asian',
    'RF': 'Filipino',
    'RH': 'Hispanic or Latino',
    'RD': 'Not Reported',
    'RP': 'Pacific Islander',
    'RT': 'Two or More Races',
    'RW': 'White',
    'GM': 'Male',
    'GF': 'Female',
    'GX': 'Non-Binary Gender (Beginning 2019–20)',
    'GZ': 'Missing Gender',
    'SE': 'English Learners',
    'SD': 'Students with Disabilities',
    'SS': 'Socioeconomically Disadvantaged',
    'SM': 'Migrant',
    'SF': 'Foster',
    'SH': 'Homeless'
}

expulsion_data_s['ReportingCategory'] = expulsion_data_s['ReportingCategory'].replace(replacement_dict_expulsions)
expulsion_data_s
```

```
/var/folders/64/69_lqqy93wz576gl39ljxz5h0000gn/T/ipykernel_70886/3125572364.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a
-view-versus-a-copy
  expulsion_data_s.replace('*', '0', inplace=True)
/var/folders/64/69_lqqy93wz576gl39ljxz5h0000gn/T/ipykernel_70886/3125572364.py:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a
-view-versus-a-copy
  expulsion_data_s[column] = expulsion_data_s[column].astype(float)
/var/folders/64/69_lqqy93wz576gl39ljxz5h0000gn/T/ipykernel_70886/3125572364.py:43: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a
-view-versus-a-copy
  expulsion_data_s['ReportingCategory'] = expulsion_data_s['ReportingCategory'].replace(replacement_dict_expulsions)
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Out[5]:

	AcademicYear	AggregateLevel	CountyCode	DistrictCode	SchoolCode	CountyName	DistrictName	SchoolName	CharterYN	Reporting
44473	2021-22	S	1	10017.0	130419.0	Alameda	Alameda County Office of Education	Alameda County Community	No	
44474	2021-22	S	1	10017.0	130419.0	Alameda	Alameda County Office of Education	Alameda County Community	No	
44475	2021-22	S	1	10017.0	130419.0	Alameda	Alameda County Office of Education	Alameda County Community	No	
44476	2021-22	S	1	10017.0	130419.0	Alameda	Alameda County Office of Education	Alameda County Community	No	African
44477	2021-22	S	1	10017.0	130419.0	Alameda	Alameda County Office of Education	Alameda County Community	No	Not
...	...	...	...	...	...	...	...	...	...	
225289	2021-22	S	58	10587.0	117242.0	Yuba	Yuba County Office of Education	Yuba Environmental Science Charter Academy	Yes	
225290	2021-22	S	58	10587.0	117242.0	Yuba	Yuba County Office of Education	Yuba Environmental Science Charter Academy	Yes	Stuc D
225291	2021-22	S	58	10587.0	117242.0	Yuba	Yuba County Office of Education	Yuba Environmental Science Charter Academy	Yes	English
225292	2021-22	S	58	10587.0	117242.0	Yuba	Yuba County Office of Education	Yuba Environmental Science Charter Academy	Yes	Socioeco Disac
225293	2021-22	S	58	10587.0	117242.0	Yuba	Yuba County Office of Education	Yuba Environmental Science Charter Academy	Yes	

180821 rows × 21 columns

Absence Data cleaned in a very similar way to the expulsion data.

In [6]:

```
absence_data_s = absence_data[absence_data['Aggregate Level'] == 'S']

replacement_dict_absence_data = {
    'RB': 'African American',
    'RI': 'American Indian or Alaska Native',
    'RA': 'Asian',
    'RF': 'Filipino',
    'RH': 'Hispanic or Latino',
    'RD': 'Did not Report',
    'RP': 'Pacific Islander',
    'RT': 'Two or More Races',
    'RW': 'White',
    'GM': 'Male',
    'GF': 'Female',
    'GX': 'Non-Binary Gender (Beginning 2019–20)',
    'GZ': 'Missing Gender',
    'SE': 'English Learners',
    'SD': 'Students with Disabilities',
    'SS': 'Socioeconomically Disadvantaged',
    'SM': 'Migrant',
    'SF': 'Foster',
    'SH': 'Homeless',
    'GRKN': 'Kindergarten (GRK prior to 2020–21)',
    'GR13': 'Grades 1–3',
    'GR46': 'Grades 4–6',
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

```
    'GRK8': 'Grades K–8',
    'GR912': 'Grades 9–12',
    'GRUG': 'Ungraded Elementary and Secondary (Retired in 2017–18)'
}

absence_data_s['Reporting Category'] = absence_data_s['Reporting Category'].replace(replacement_dict_absence_data)
absence_data_s_condensed = absence_data_s[['School Name','Reporting Category', 'School Code','Charter School', 'Chronic
absence_data_s_condensed['Charter School'] = absence_data_s_condensed['Charter School'].replace({'Yes': 1, 'No ': 0})
absence_data_s_condensed = absence_data_s_condensed.astype({'School Code': 'int64'})
absence_data_s_condensed
```

```
/var/folders/64/69_lqqy93wz576gl39ljxz5h0000gn/T/ipykernel_70886/4137370627.py:32: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a
-view-versus-a-copy
  absence_data_s['Reporting Category'] = absence_data_s['Reporting Category'].replace(replacement_dict_absence_data)
/var/folders/64/69_lqqy93wz576gl39ljxz5h0000gn/T/ipykernel_70886/4137370627.py:34: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a
-view-versus-a-copy
  absence_data_s_condensed['Charter School'] = absence_data_s_condensed['Charter School'].replace({'Yes': 1, 'No ': 0})
```

Out[6]:

	School Name	Reporting Category	School Code	Charter School	ChronicAbsenteeismEligibleCumulativeEnrollment	ChronicAbsenteeismCount	ChronicAb
57342	Urban Montessori Charter	Female	125567	1	157	50	
57343	Opportunity Academy	Female	136226	1	189	105	
57344	Aurum Preparatory Academy	Female	137448	1	81	17	
57345	Yu Ming Charter	Female	124172	1	346	0	
57346	Cox Academy	Female	6001788	1	271	113	
...	...	...	...	...	...	...	
264857	Wheatland Union High	Socioeconomically Disadvantaged	5838305	0	779	180	
264858	District Office	TA	0	0	*	*	
264859	Edward P. Duplex	TA	133751	0	63	55	
264860	Wheatland Union High	TA	5838305	0	1058	212	
264861	Wheatland Community Day High	TA	123570	0	11	3	

207520 rows × 7 columns

Join expulsion and absence dataframes on Reporting category (male, female, etc.) and School Name

In [7]:

```
expulsion_absence_data = pd.merge(expulsion_data_s,
                                  absence_data_s_condensed,
                                  how='inner',
                                  left_on=['ReportingCategory','SchoolName'],
                                  right_on = ['Reporting Category','School Name']
                                )
expulsion_absence_data.drop(['School Name', 'CharterYN', 'Reporting Category'], axis=1, inplace=True)
```

Relevant data analysis

In [8]:

```
#total expulsion and expulsion rates among groups of California students
expulsion_absence_data_grouped_by_category = expulsion_absence_data.groupby('ReportingCategory').agg('mean').reset_inde
expulsion_absence_data_grouped_by_category
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Out[8]:

	ReportingCategory	CountyCode	DistrictCode	SchoolCode	Total Expulsions	Unduplicated Count of Students Expelled (Total)	Unduplicated Count of Students Expelled (Defiance-Only)	Expulsion Rate (Total)	Expulsion Count Violent Incident (Injury)	Expulsion Count Violent Incident (No Injury)	Ex W Pos
0	African American	30.480270	67246.408173	1.731878e+06	0.015914	0.015875	0.000000	0.019285	0.007354	0.004824	0
1	American Indian or Alaska Native	30.480270	67246.408173	1.731878e+06	0.000613	0.000613	0.000000	0.002572	0.000277	0.000198	0.
2	Asian	30.480270	67246.408173	1.731878e+06	0.002214	0.002214	0.000000	0.002232	0.000791	0.000791	0.
3	English Learners	29.137073	66730.018113	3.534280e+06	0.049420	0.049297	0.000082	0.028542	0.016514	0.013236	0
4	Female	29.421187	67089.955305	2.784724e+06	0.042914	0.042723	0.000032	0.010708	0.019564	0.010180	0
5	Filipino	30.480270	67246.408173	1.731878e+06	0.000751	0.000751	0.000000	0.001637	0.000198	0.000455	0.
6	Foster	27.847506	66267.772654	4.135189e+06	0.006681	0.006681	0.000000	0.032918	0.002789	0.001816	0.
7	Hispanic or Latino	30.480270	67246.408173	1.731878e+06	0.078918	0.078701	0.000059	0.017027	0.025245	0.021232	0
8	Homeless	28.137064	66569.467840	4.641275e+06	0.022194	0.022069	0.000063	0.050670	0.007042	0.006162	0
9	Male	30.235908	67100.353005	1.949917e+06	0.101759	0.101469	0.000156	0.025788	0.031055	0.030120	0
10	Migrant	29.423285	67042.500589	4.675510e+06	0.011077	0.011077	0.000000	0.020387	0.004949	0.002357	0
11	Non-Binary Gender (Beginning 2019–20)	29.155227	65221.058771	3.299807e+06	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.
12	Pacific Islander	30.480270	67246.408173	1.731878e+06	0.000455	0.000455	0.000000	0.002117	0.000198	0.000138	0.
13	Socioeconomically Disadvantaged	29.678144	66923.767759	2.570981e+06	0.147352	0.146853	0.000235	0.028153	0.051808	0.041388	0.
14	Students with Disabilities	30.468808	67233.170155	1.789166e+06	0.026748	0.026605	0.000082	0.018240	0.009831	0.007332	0.
15	TA	30.480270	67246.408173	1.731878e+06	0.117903	0.117527	0.000158	0.016215	0.039894	0.033390	0.
16	Two or More Races	30.480270	67246.408173	1.731878e+06	0.002946	0.002906	0.000020	0.006255	0.001087	0.000949	0.
17	White	30.480270	67246.408173	1.731878e+06	0.014234	0.014155	0.000040	0.009007	0.003954	0.004329	0.

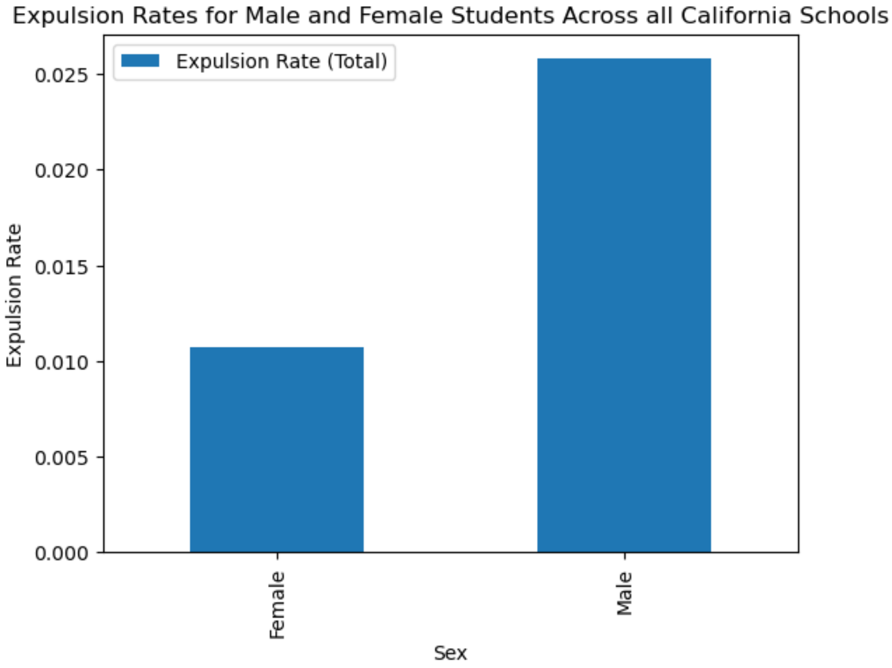
Per our research question, we explore different factors that may influence expulsion rates.

In [9]:

```
expulsion_absence_data_grouped_by_category[
    (expulsion_absence_data_grouped_by_category['ReportingCategory']=='Female')
    | (expulsion_absence_data_grouped_by_category['ReportingCategory']=='Male')
].plot.bar(
    x = 'ReportingCategory',
    y = 'Expulsion Rate (Total)'
)
plt.title("Expulsion Rates for Male and Female Students Across all California Schools")
plt.xlabel("Sex")
plt.ylabel("Expulsion Rate")
```

Out[9]:

```
Text(0, 0.5, 'Expulsion Rate')
```



We can see here that males are far more likely to be expelled

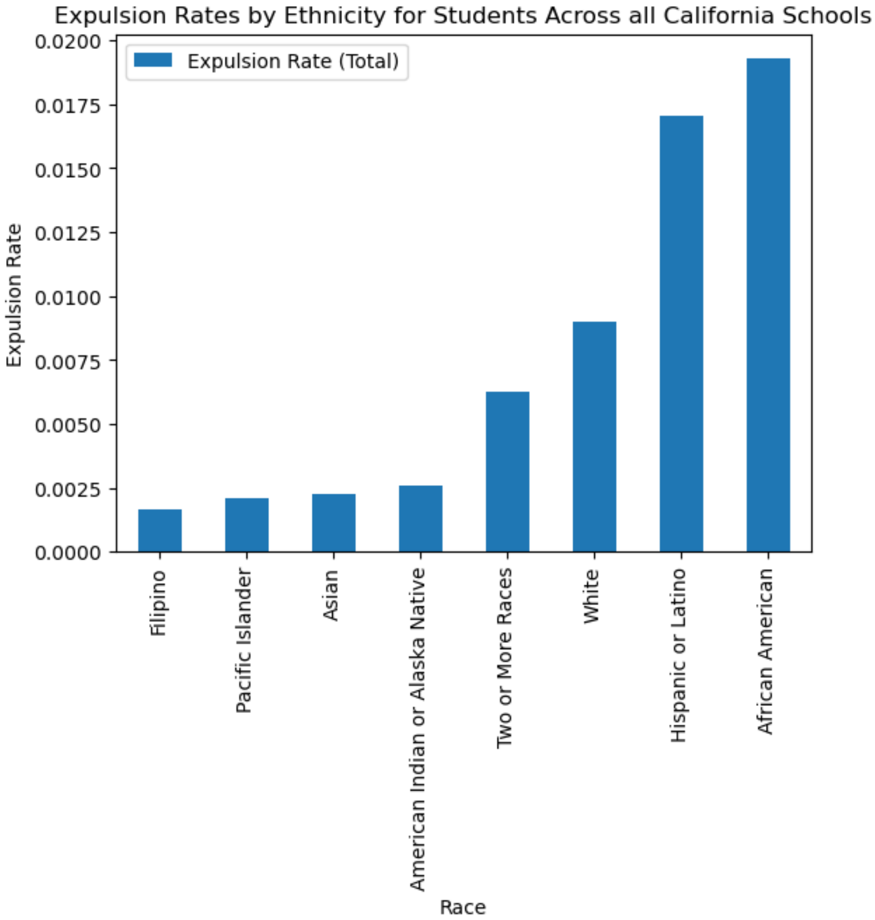
In [10]:

```
by_race = expulsion_absence_data_grouped_by_category[
    (expulsion_absence_data_grouped_by_category['ReportingCategory']=='White')
    | (expulsion_absence_data_grouped_by_category['ReportingCategory']=='African American')
    | (expulsion_absence_data_grouped_by_category['ReportingCategory']=='Asian')
    | (expulsion_absence_data_grouped_by_category['ReportingCategory']=='Pacific Islander')
    | (expulsion_absence_data_grouped_by_category['ReportingCategory']=='Hispanic or Latino')
    | (expulsion_absence_data_grouped_by_category['ReportingCategory']=='Filipino')
    | (expulsion_absence_data_grouped_by_category['ReportingCategory']=='American Indian or Alaska Native')
    | (expulsion_absence_data_grouped_by_category['ReportingCategory']=='Two or More Races')
][['ReportingCategory','Expulsion Rate (Total)']]
by_race.sort_values('Expulsion Rate (Total)').plot.bar(x = 'ReportingCategory')
plt.title("Expulsion Rates by Ethnicity for Students Across all California Schools")
plt.xlabel("Race")
plt.ylabel("Expulsion Rate")
```

Out[10]:

```
Text(0, 0.5, 'Expulsion Rate')
```

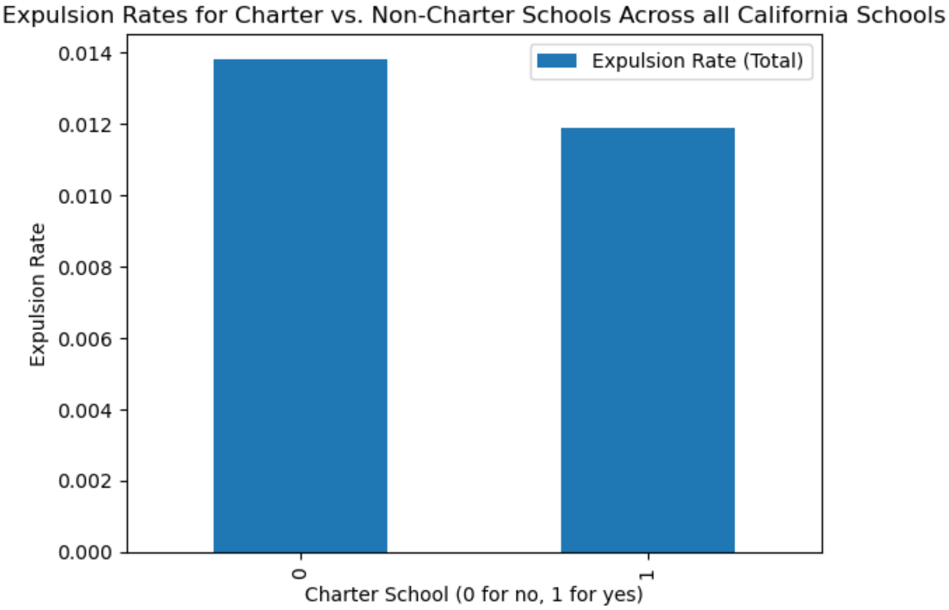
Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js



Here, we can see the variation in expulsion rates based on ethnicity.

In [11]:
```
expulsion_absence_data.groupby('Charter School').agg('mean').reset_index().plot.bar(
    x = 'Charter School',
    y = 'Expulsion Rate (Total)'
)
plt.title("Expulsion Rates for Charter vs. Non-Charter Schools Across all California Schools")
plt.xlabel("Charter School (0 for no, 1 for yes)")
plt.ylabel("Expulsion Rate")
```

Out[11]:
```
Text(0, 0.5, 'Expulsion Rate')
```



Now here, we can see the variation in expulsion rates based on whether the school is a charter school or not.

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

In [12]:

```
expulsion_absence_data_ta = expulsion_absence_data[expulsion_absence_data['ReportingCategory'] == 'TA']
cols_to_check = ['ChronicAbsenteeismEligibleCumulativeEnrollment',
                 'ChronicAbsenteeismCount',
                 'ChronicAbsenteeismRate']
expulsion_absence_data_ta = expulsion_absence_data_ta[~expulsion_absence_data_ta[cols_to_check].apply(
    lambda x: x.str.contains('*', regex=False)).any(axis=1)]
expulsion_absence_data_ta[cols_to_check] = expulsion_absence_data_ta[cols_to_check].astype(float)
expulsion_absence_data_ta
```

Out[12]:

	AcademicYear	AggregateLevel	CountyCode	DistrictCode	SchoolCode	CountyName	DistrictName	SchoolName	ReportingCategory	C
15	2021-22	S	1	10017.0	130419.0	Alameda	Alameda County Office of Education	Alameda County Community	TA	
31	2021-22	S	1	10017.0	130401.0	Alameda	Alameda County Office of Education	Alameda County Juvenile Hall/Court	TA	
47	2021-22	S	1	10017.0	130625.0	Alameda	Alameda County Office of Education	Alternatives in Action	TA	
63	2021-22	S	1	10017.0	137448.0	Alameda	Alameda County Office of Education	Aurum Preparatory Academy	TA	
79	2021-22	S	1	10017.0	123968.0	Alameda	Alameda County Office of Education	Community School for Creative Education	TA	
...	...	...	...	...	...	...	...	...	...	
676639	2021-22	S	58	10587.0	5830047.0	Yuba	Yuba County Office of Education	Harry P B Carden	TA	
676655	2021-22	S	58	10587.0	113274.0	Yuba	Yuba County Office of Education	Thomas E. Mathews Community	TA	
676671	2021-22	S	58	10587.0	5830112.0	Yuba	Yuba County Office of Education	Yuba County Career Preparatory Charter	TA	
676687	2021-22	S	58	10587.0	6069249.0	Yuba	Yuba County Office of Education	Yuba County Special Education	TA	
676701	2021-22	S	58	10587.0	117242.0	Yuba	Yuba County Office of Education	Yuba Environmental Science Charter Academy	TA	

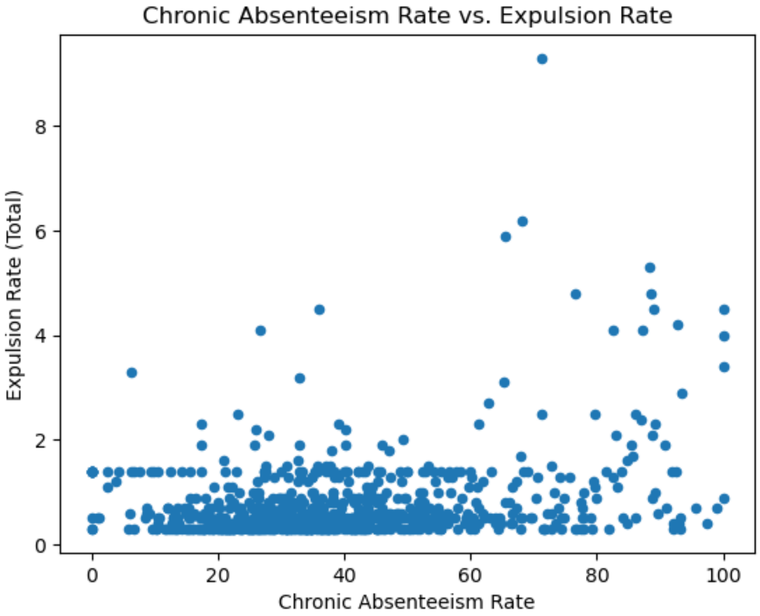
27592 rows × 25 columns

In [13]:

```
ax1 = expulsion_absence_data_ta[expulsion_absence_data_ta['Expulsion Rate (Total)'] > 0.25].plot.scatter(
    y = 'Expulsion Rate (Total)',
    x = "ChronicAbsenteeismRate"
)
plt.xlabel("Chronic Absenteeism Rate")
plt.title("Chronic Absenteeism Rate vs. Expulsion Rate")
```

Out[13]:

```
Text(0.5, 1.0, 'Chronic Absenteeism Rate vs. Expulsion Rate')
```



Here, we examine any possible correlation between the absence rate of a school and the expulsion rate of a school. We joined a table containing absence data with our original expulsion data.

Raksha's EDA Code

In [14]:

```
# with open(expulsion_data_path, 'rb') as f:
#     result = chardet.detect(f.read())

# df = pd.read_csv(expulsion_data_path)
# df = df[df['CountyName'] != 'State']
# df.head(5)
df = expulsion_data
```

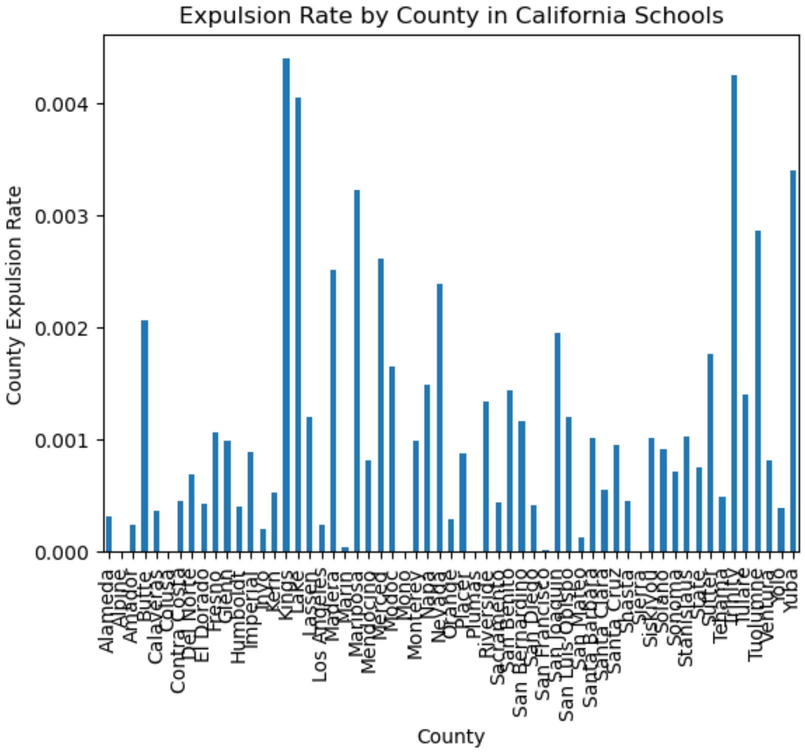
In [15]:

```
#CLEAN DATA
# Replace "*" with NaN (Not a Number)
df = df.replace('*', pd.NA)

# Convert the columns to numeric values and drop any rows with missing values
df[['CumulativeEnrollment', 'Total Expulsions']] = df[['CumulativeEnrollment', 'Total Expulsions']].apply(pd.to_numeric
df_select = df.dropna(subset=['CumulativeEnrollment', 'Total Expulsions'])
df_select
#ADD COLUMN COUNTY EXPULSION RATE
sum_T = df_select.groupby('CountyName')['Total Expulsions'].sum()
sum_C = df_select.groupby('CountyName')['CumulativeEnrollment'].sum()
data = sum_T/sum_C
data = sum_T/sum_C
data = pd.DataFrame(data=data, columns=['County Expulsion Rate']).rename_axis('CountyName')
```

In [16]:

```
# Calculate the mean expulsion rate for each county
expulsion_rate_by_county = data.groupby('CountyName')['County Expulsion Rate'].mean()

# Plot the results
expulsion_rate_by_county.plot(kind='bar', x='CountyName', y='County Expulsion Rate')

plt.xlabel('County')
plt.ylabel('County Expulsion Rate')
plt.title('Expulsion Rate by County in California Schools')
plt.show()
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js



This graph is the distribution of expulsion rate by county as a percentage. Expelling students is rare so the rate per district is small but some counties such as Lake, Kings, and Trinity have higher rates than others. Identifying this will help us look into features of such high-expulsion rate districts over the course of the project.

In [17]:

```
#CHLOROPLETH MAP BY COUNTY

m = folium.Map(location=[36.7783, -119.4179], zoom_start=5.5)

#Import geojson
county = gpd.read_file('data/California_County_Boundaries.geojson')
#folium.GeoJson(districts, name='geojson').add_to(m)

folium.GeoJson(county, name="geojson").add_to(m)
```

Out[17]:

```
<folium.features.GeoJson at 0x11e7f8670>
```

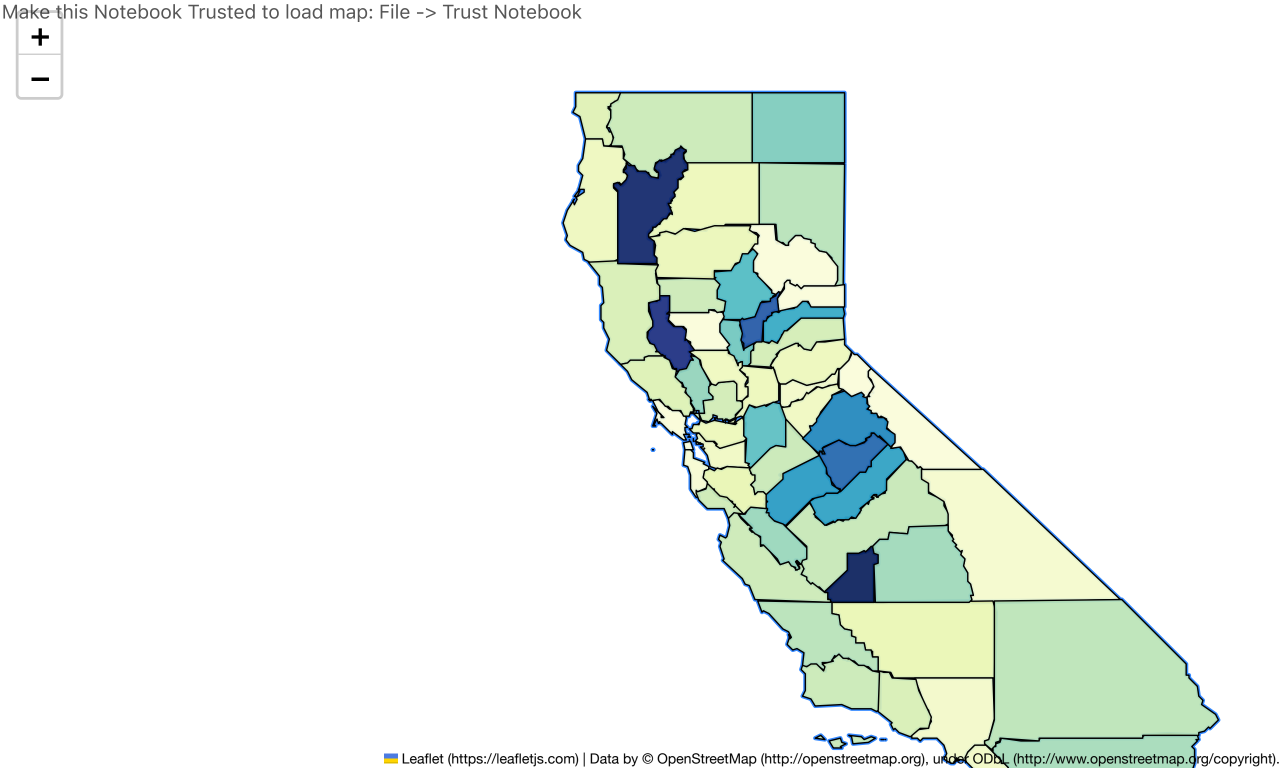
In [18]:

```
# Merge the GeoDataFrame and the DataFrame using the county FIPS code
merged_df = pd.merge(county, data, on='CountyName')

#step1 colormap scale
colormap = linear.YlGnBu_09.scale(merged_df['County Expulsion Rate'].min(), merged_df['County Expulsion Rate'].max())
colormap

#step2 create dictionary
merged_dict = merged_df.set_index('CountyName')['County Expulsion Rate'].to_dict()
merged_dict

# Create a chloropleth map of the expulsion rate
folium.GeoJson(
    county,
    name='geojson',
    style_function=lambda feature: {
        'fillColor': colormap(merged_dict[feature['properties']['CountyName']]),
        'color': 'black',
        'weight': 1,
        'fillOpacity': 0.9},
        tooltip=folium.features.GeoJsonTooltip(
            fields=['CountyName'],
            aliases=['County:'])
).add_to(m)

print(county.keys())
m

#merged_dict
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

```
Index(['OBJECTID', 'CountyName', 'AdminRegion', 'FireMAR', 'LawMAR',
       'State_FIPS_ID', 'County_FIPS_ID', 'Shape__Area', 'Shape__Length',
       'geometry'],
      dtype='object')
```

Out[18]: Make this Notebook Trusted to load map: File -> Trust Notebook



🇺🇦 Leaflet (https://leafletjs.com) | Data by © OpenStreetMap (http://openstreetmap.org), under ODbL (http://www.openstreetmap.org/copyright).

This choloropleth map identifies the districts that have much higher expulsion rates than others in dark blue (ex. Lake, Trinity, Kings). The districts in pale green have significantly lower expulsion rates such as Alpine and Mono. Using this district variation, we will be able to train our model to identify the specific factors/features that make expulsion rates higher in some schools rather than others.

Bella's EDA Code

In [19]:
```
#list of other relevant variables:
other_reporting_variables = ['SE', 'SD', 'SS', 'SM', 'SF', 'SH']

#make a table that sums up the reasons for expulsion for students in the above reporting categories
schools = expulsion_data[(expulsion_data['AggregateLevel'] == 'S')]
only_s_variables = schools[(schools['ReportingCategory'].isin(other_reporting_variables))]
only_s_variables.groupby(by = 'ReportingCategory').agg(sum)
```

Out[19]:

	CountyCode	DistrictCode	SchoolCode
ReportingCategory			
SD	305007	694006533.0	4.201300e+10
SE	290612	664445511.0	4.147692e+10
SF	233227	540143143.0	3.412033e+10
SH	249202	573212758.0	3.684165e+10
SM	89871	200871300.0	1.318313e+10
SS	302850	689481572.0	4.213636e+10

Pooja's EDA Code

In [20]:
```
by_category = expulsion_data[(expulsion_data['AggregateLevel'] == 'S') & (expulsion_data['ReportingCategory'] == 'TA')]
by_category
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Out[20]:

	AcademicYear	AggregateLevel	CountyCode	DistrictCode	SchoolCode	CountyName	DistrictName	SchoolName	CharterYN	Reporting
44489	2021-22	S	1	10017.0	130419.0	Alameda	Alameda County Office of Education	Alameda County Community	No	
44506	2021-22	S	1	10017.0	130401.0	Alameda	Alameda County Office of Education	Alameda County Juvenile Hall/Court	No	
44523	2021-22	S	1	10017.0	130625.0	Alameda	Alameda County Office of Education	Alternatives in Action	Yes	
44540	2021-22	S	1	10017.0	137448.0	Alameda	Alameda County Office of Education	Aurum Preparatory Academy	Yes	
44557	2021-22	S	1	10017.0	123968.0	Alameda	Alameda County Office of Education	Community School for Creative Education	Yes	
...	...	...	...	...	...	...	...	...	...	
225226	2021-22	S	58	10587.0	5830047.0	Yuba	Yuba County Office of Education	Harry P B Carden	No	
225243	2021-22	S	58	10587.0	113274.0	Yuba	Yuba County Office of Education	Thomas E. Mathews Community	No	
225261	2021-22	S	58	10587.0	5830112.0	Yuba	Yuba County Office of Education	Yuba County Career Preparatory Charter	Yes	
225278	2021-22	S	58	10587.0	6069249.0	Yuba	Yuba County Office of Education	Yuba County Special Education	No	
225293	2021-22	S	58	10587.0	117242.0	Yuba	Yuba County Office of Education	Yuba Environmental Science Charter Academy	Yes	

10652 rows × 21 columns

In [21]:

```python
by_category = by_category.set_index('DistrictName')
by_category.loc[:, "Expulsion Count Violent Incident (Injury)":]
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Out[21]:

DistrictName	Expulsion Count Violent Incident (Injury)	Expulsion Count Violent Incident (No Injury)	Expulsion Count Weapons Possession	Expulsion Count Illicit Drug-Related	Expulsion Count Defiance-Only	Expulsion Count Other Reasons
Alameda County Office of Education	0	0	0	0	0	0
Alameda County Office of Education	0	0	0	0	0	0
Alameda County Office of Education	0	0	0	0	0	0
Alameda County Office of Education	0	0	0	0	0	0
Alameda County Office of Education	0	0	0	0	0	0
...	...	...	...	...	...	...
Yuba County Office of Education	0	0	0	0	0	0
Yuba County Office of Education	0	0	0	0	0	0
Yuba County Office of Education	0	0	0	0	0	0
Yuba County Office of Education	0	0	0	0	0	0
Yuba County Office of Education	0	0	0	0	0	0

10652 rows × 6 columns

In [22]:

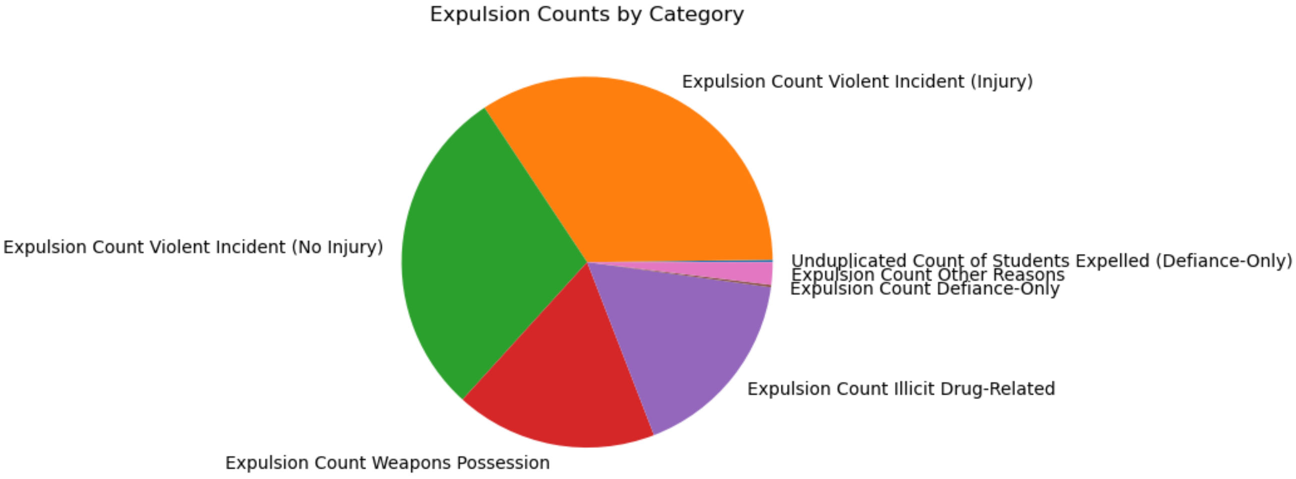
```
by_category2 = expulsion_data[(expulsion_data['AggregateLevel'] == 'S')]
by_category2.groupby(by='AggregateLevel').agg(sum)
```

Out[22]:

AggregateLevel	CountyCode	DistrictCode	SchoolCode
S	5207673	1.185798e+10	7.224763e+11

In [23]:

```
import seaborn as sns
import matplotlib.pyplot as plt

data = [38.0, 6225.0, 5280.0, 3208.0, 3095.0, 38.0, 357.0]
labels = ['Unduplicated Count of Students Expelled (Defiance-Only)', 'Expulsion Count Violent Incident (Injury)', 'Expul
plt.pie(data, labels = labels)
plt.title('Expulsion Counts by Category')
plt.show()
```



Above, this pie chart lets us see what reasons students get expelled for

In [ ]:

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js