TenStrandsData

August 3, 2023

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

```
[ ]: # from google.colab import drive
     # drive.mount('/content/drive')
```

1 Section 1: San Mateo County Equity Focus

```
[ ]: # Read in the data
     # url = '/content/drive/MyDrive/Ten Strands/Section 2: County level/San Mateo␣
      ↪data - section1.csv'
     # s1_san_mateo_data = pd.read_csv(url)
     # s1_san_mateo_data.info()
     # s1_san_mateo_data
```

```
[ ]: # Read in the data
     s1_san_mateo_data = pd.read_csv("San Mateo data - section1.csv")
     s1_san_mateo_data.info()
     s1_san_mateo_data
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
Data columns (total 3 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Demographic Indicators  10 non-null     object
 1   San Mateo County        10 non-null     object
 2   State Average           10 non-null     object
dtypes: object(3)
memory usage: 368.0+ bytes
```

```
[ ]:                               Demographic Indicators San Mateo County  \
     0                          Student Enrollment (2022)           86,422
     1                               # of School Districts               26
```

```
2                              Median Income (2021)        $136,837
3                                Expense ADA (2021)         $23,107
4                    % Unduplicated Students (2022)          34.42%
5  % Students Eligible Free or Reduced Lunch (FRL…         30.60%
6   % English Language Learner Students (ELL) (2022)          21.30%
7                        % Students of Color (2021)          73.40%
8      % Students Receiving Special Education (2019)           1.37%
9                                   Pollution Burden            35.6

  State Average
0     5,892,240
1           939
2       $84,097
3       $18,827
4       55.73 %
5        57.8 %
6        19.1 %
7        78.90%
8        13.80%
9            50
```

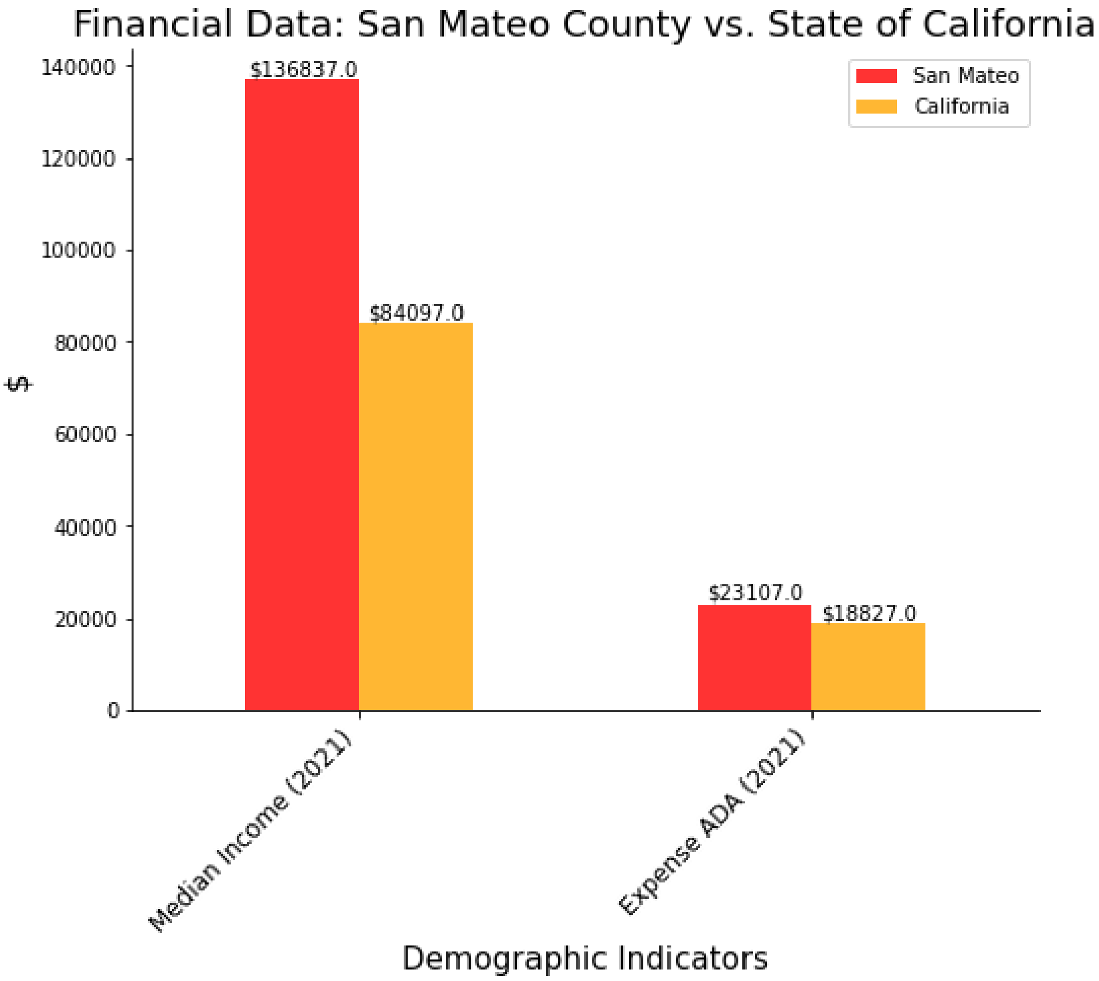
```
[ ]: financial_data = s1_san_mateo_data.copy()
     financial_data.set_index('Demographic Indicators', inplace=True)
     financial_data = financial_data.loc[['Median Income (2021)', 'Expense ADA␣
      ↪(2021)'], :]
     financial_data
```

```
[ ]:                        San Mateo County State Average
     Demographic Indicators
     Median Income (2021)           $136,837       $84,097
     Expense ADA (2021)              $23,107       $18,827
```

```
[ ]: # Remove $ and ,
     financial_data = financial_data.replace(r'[$,]', '', regex=True).astype(float)
     financial_data
```

```
[ ]:                         San Mateo County  State Average
     Demographic Indicators
     Median Income (2021)            136837.0        84097.0
     Expense ADA (2021)               23107.0        18827.0
```

```
[ ]: # Plot the comparative bar graph
     ax = financial_data.plot(kind='bar', figsize=(8, 6), rot=0, color=['r',␣
      ↪'orange'], alpha=0.8)
     plt.xlabel('Demographic Indicators', size=15)
     plt.xticks(size=12, rotation=45, ha='right')
     plt.ylabel('$', size=15)
     plt.title('Financial Data: San Mateo County vs. State of California', size=18)
     plt.legend(labels=['San Mateo', 'California'])

     # Adding data labels above each bar
     for i in ax.patches:
         ax.text(i.get_x() + i.get_width() / 2, i.get_height() + 0.5, '$' + str(i.
       ↪get_height()) ,
                 ha='center', va='bottom', fontsize=10)

     # Remove the top and right spines for a cleaner look
     ax.spines['top'].set_visible(False)
     ax.spines['right'].set_visible(False)

     plt.savefig('financial_indicators.png', bbox_inches='tight')
     plt.show()
```



```
[ ]: students_demographic = s1_san_mateo_data.copy()
     students_demographic = students_demographic[students_demographic['Demographic␣
      ↪Indicators'].str.startswith('%')]
     students_demographic.set_index('Demographic Indicators', inplace=True)
     students_demographic
```

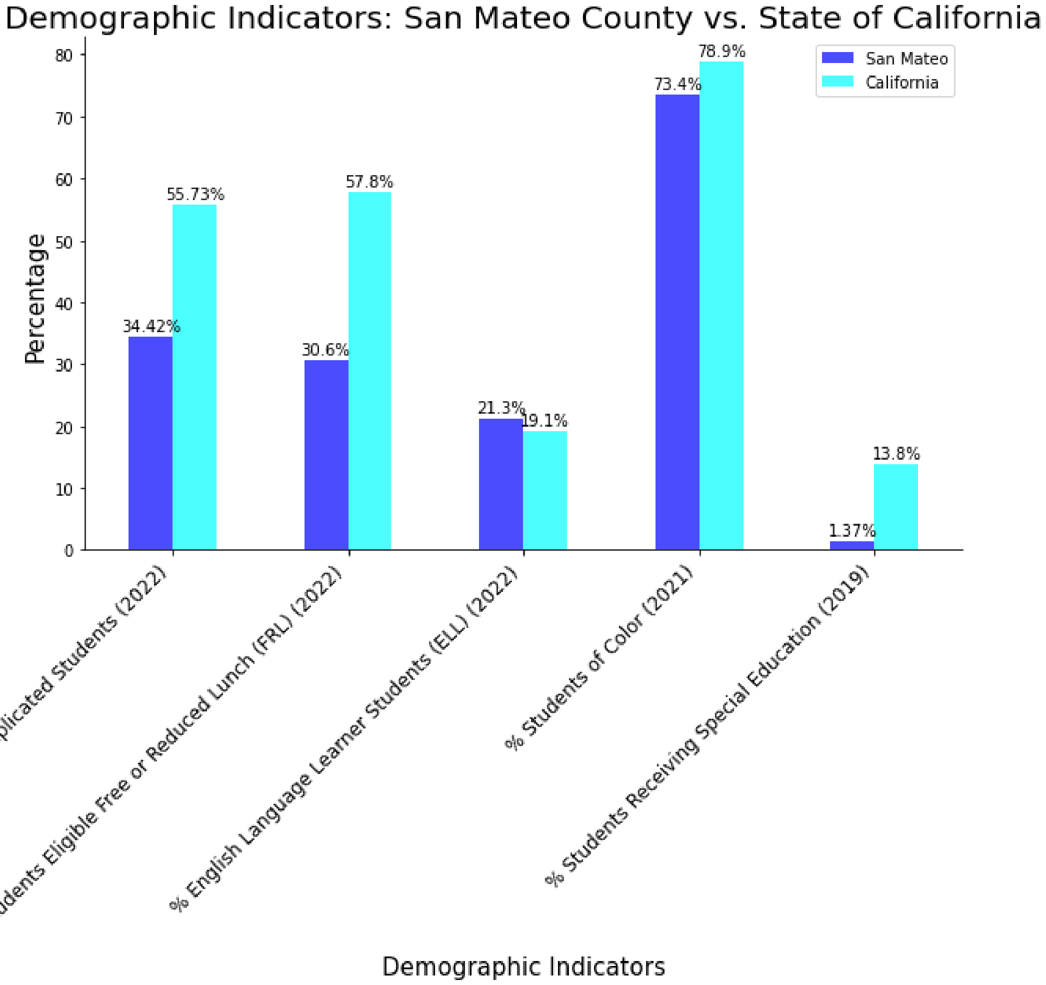
```
[ ]:                                                    San Mateo County  \
     Demographic Indicators
     % Unduplicated Students (2022)                               34.42%
     % Students Eligible Free or Reduced Lunch (FRL)…           30.60%
     % English Language Learner Students (ELL) (2022)             21.30%
     % Students of Color (2021)                                   73.40%
     % Students Receiving Special Education (2019)                 1.37%

                                                         State Average
     Demographic Indicators
     % Unduplicated Students (2022)                           55.73 %
     % Students Eligible Free or Reduced Lunch (FRL)…        57.8 %
     % English Language Learner Students (ELL) (2022)          19.1 %
     % Students of Color (2021)                                78.90%
     % Students Receiving Special Education (2019)             13.80%
```

```
[ ]: students_demographic['San Mateo County'] = students_demographic['San Mateo␣
      ↪County'].str.rstrip('%').astype(float)
     students_demographic['State Average'] = students_demographic['State Average'].
      ↪str.rstrip('%').astype(float)
     students_demographic
```

```
[ ]:                                                     San Mateo County  \
     Demographic Indicators
     % Unduplicated Students (2022)                                 34.42
     % Students Eligible Free or Reduced Lunch (FRL)…             30.60
     % English Language Learner Students (ELL) (2022)               21.30
     % Students of Color (2021)                                     73.40
     % Students Receiving Special Education (2019)                   1.37

                                                          State Average
     Demographic Indicators
     % Unduplicated Students (2022)                              55.73
     % Students Eligible Free or Reduced Lunch (FRL)…          57.80
     % English Language Learner Students (ELL) (2022)            19.10
     % Students of Color (2021)                                  78.90
     % Students Receiving Special Education (2019)               13.80
```

```
[ ]: # Plot the comparative bar graph
     ax = students_demographic.plot(kind='bar', figsize=(10, 6), rot=0,
                                    color=['blue', 'cyan'], alpha=0.7)
```

```python
plt.xlabel('Demographic Indicators', size=15)
plt.xticks(size=12, rotation=45, ha='right')
plt.ylabel('Percentage', size=15)
plt.title('Demographic Indicators: San Mateo County vs. State of California', size=20)
plt.legend(labels=['San Mateo', 'California'])

# Adding data labels above each bar
for i in ax.patches:
    ax.text(i.get_x() + i.get_width() / 2, i.get_height() + 0.5, str(i.get_height()) + '%',
            ha='center', va='bottom', fontsize=10)

# Remove the top and right spines for a cleaner look
ax.spines['top'].set_visible(False)
ax.spines['right'].set_visible(False)

plt.savefig('demographic_indicators.png', bbox_inches='tight')
plt.show()
```



```
[ ]: s1_san_mateo_data
```

```
[ ]:                               Demographic Indicators San Mateo County  \
     0                            Student Enrollment (2022)           86,422
     1                                # of School Districts               26
     2                                  Median Income (2021)         $136,837
     3                                    Expense ADA (2021)          $23,107
     4                      % Unduplicated Students (2022)           34.42%
     5  % Students Eligible Free or Reduced Lunch (FRL…         30.60%
     6   % English Language Learner Students (ELL) (2022)           21.30%
     7                         % Students of Color (2021)           73.40%
     8      % Students Receiving Special Education (2019)            1.37%
     9                                     Pollution Burden             35.6

       State Average
     0     5,892,240
     1           939
     2       $84,097
     3       $18,827
     4       55.73 %
     5        57.8 %
     6        19.1 %
     7        78.90%
     8        13.80%
     9            50
```

```
[ ]: numeric_data = s1_san_mateo_data.copy()
     numeric_data.set_index('Demographic Indicators', inplace=True)
     numeric_data = numeric_data.loc[['Student Enrollment (2022)', '# of School␣
      ↪Districts', 'Pollution Burden'], :]
     numeric_data
```

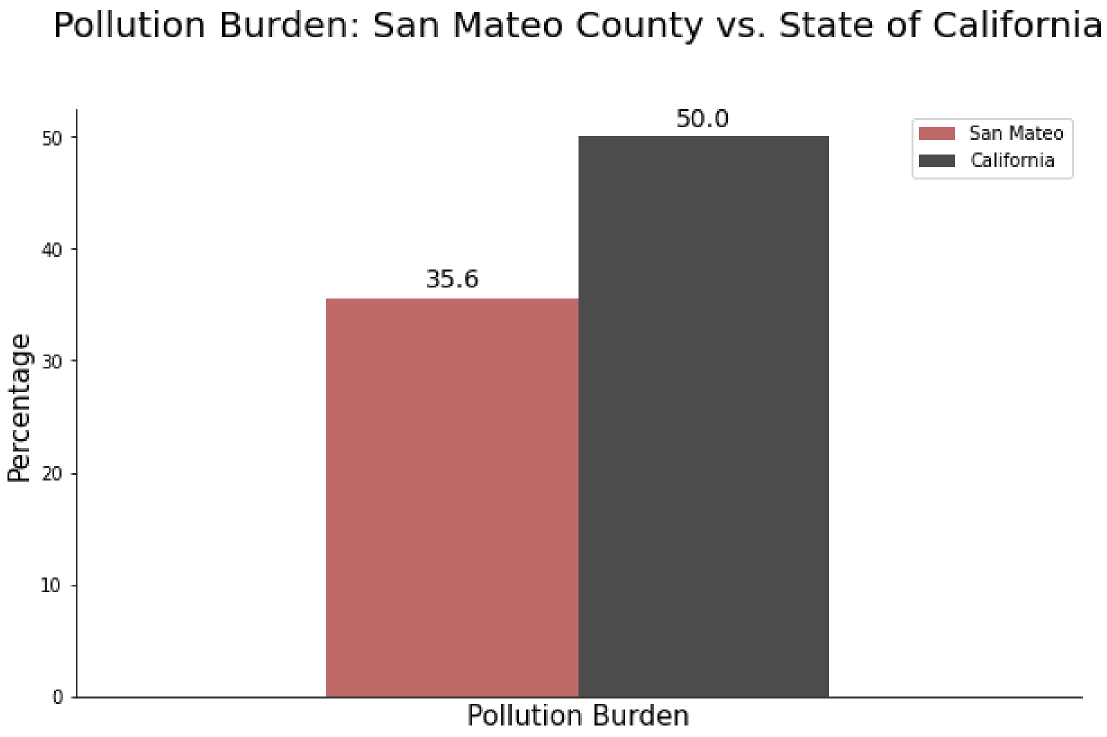
```
[ ]:                           San Mateo County State Average
     Demographic Indicators
     Student Enrollment (2022)           86,422     5,892,240
     # of School Districts                   26           939
     Pollution Burden                      35.6            50
```

```
[ ]: numeric_data = numeric_data.replace(r'[,]', '', regex=True).astype(float)
     numeric_data
```

```
[ ]:                            San Mateo County  State Average
     Demographic Indicators
     Student Enrollment (2022)           86422.0      5892240.0
     # of School Districts                  26.0          939.0
```

```
Pollution Burden                        35.6           50.0
```

[ ]:
```
# Plot the comparative bar graph
ax = numeric_data.loc[['Pollution Burden'], :].plot(kind='bar', figsize=(10, 6),
                                                    rot=0, color=['brown',␣
 ↪'black'], alpha=0.7)
plt.xlabel('Pollution Burden', size=15)
plt.xticks([])
plt.ylabel('Percentage', size=15)
plt.title('Pollution Burden: San Mateo County vs. State of California',␣
 ↪size=20, pad=40)
plt.legend(labels=['San Mateo', 'California'])

# Adding data labels above each bar
for i in ax.patches:
    ax.text(i.get_x() + i.get_width() / 2, i.get_height() + 0.5, str(i.
 ↪get_height()),
            ha='center', va='bottom', fontsize=14)

# Remove the top and right spines for a cleaner look
ax.spines['top'].set_visible(False)
ax.spines['right'].set_visible(False)

plt.savefig('pollution_burden.png', bbox_inches='tight')
plt.show()
```



2 Section 2: San Mateo County-Level Focus Visualization

[ ]:

```
# Read in the data
url = '/content/drive/MyDrive/Ten Strands/Section 2: County level/San Mateo␣
 ↪data - section2.csv'
s2_san_mateo_data = pd.read_csv(url)
s2_san_mateo_data
```

[ ]:

```
                             COE Investments  \
0              Climate Emergency Declaration   
1                  County Environmental Plan   
2                 County Climate Action Plan   
3           Local Hazard and Mitigation Plan   
4  County Greenhouse Gas Emissions Inventory   
5           Community Choice Aggregate (CCA)   
6        Clean Energy Infrastructure Project   

   San Mateo County \n# Within County  \
0                                   1   
1                                   1   
2                                   1   
3                                   1   
4                                   1   
5                                   1   
6                                   0   

  State Scorecard\n# of COEs with one or more investment
0                                      12/COEs = 21%
1                                      45 COEs = 77%
2                                      26 COEs = 45%
3                                     58 COEs = 100%
4                                      35 COEs = 60%
5                                      25 COEs = 43%
6                                      27 COEs = 46%
```

[ ]:

```
s2_san_mateo_data.columns
```

[ ]:

```
Index(['COE Investments', 'San Mateo County \n# Within County',
       'State Scorecard\n# of COEs with one or more investment'],
      dtype='object')
```

[ ]:

```
renamed_columns = {'San Mateo County \n# Within County': 'San Mateo County',
                  'State Scorecard\n# of COEs with one or more investment':␣
 ↪'State Scorecard'}
```

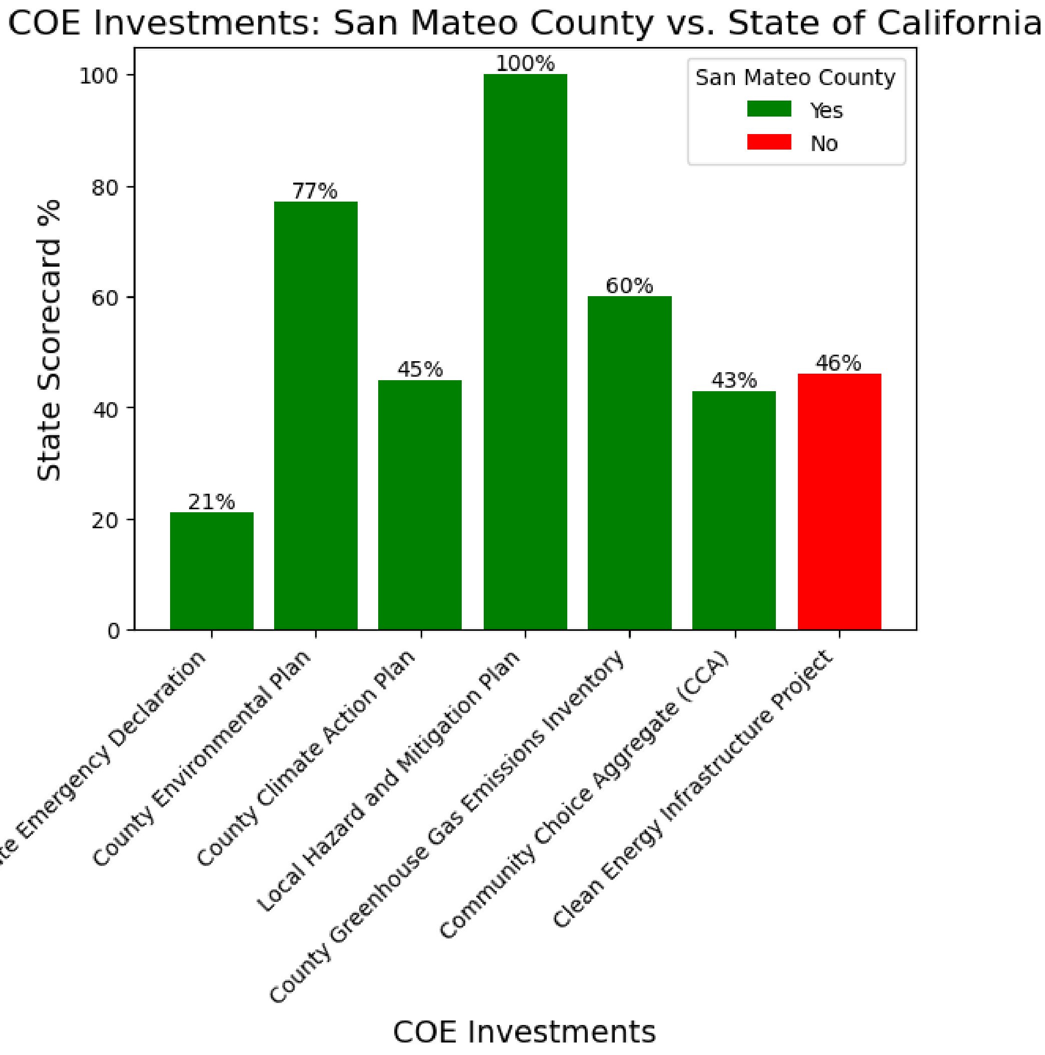
```
s2_san_mateo_data.rename(columns=renamed_columns, inplace=True)
#san_mateo_data
```

[ ]:
```
state_data = [12, 45, 26, 58, 35, 25, 27]
state_data_percent = [21, 77, 45, 100, 60, 43, 46]
s2_san_mateo_data['State Data'] = state_data
s2_san_mateo_data['State Data Percent'] = state_data_percent
```

[ ]:
```
s2_san_mateo_data
```

[ ]:
```
                             COE Investments  San Mateo County  \
0              Climate Emergency Declaration                 1
1                  County Environmental Plan                 1
2                 County Climate Action Plan                 1
3           Local Hazard and Mitigation Plan                 1
4  County Greenhouse Gas Emissions Inventory                 1
5           Community Choice Aggregate (CCA)                 1
6        Clean Energy Infrastructure Project                 0

  State Scorecard  State Data  State Data Percent
0   12/COEs = 21%          12                  21
1   45 COEs = 77%          45                  77
2   26 COEs = 45%          26                  45
3  58 COEs = 100%          58                 100
4   35 COEs = 60%          35                  60
5   25 COEs = 43%          25                  43
6   27 COEs = 46%          27                  46
```

[ ]:
```
investments = s2_san_mateo_data['COE Investments']
san_mateo_county = s2_san_mateo_data['San Mateo County']
state_percent = s2_san_mateo_data['State Data Percent']
colors = ['green' if val else 'red' for val in san_mateo_county]
plt.bar(investments, state_percent, color=colors)
plt.xlabel('COE Investments', size=14)
plt.xticks(rotation=45, ha='right')
plt.ylabel('State Scorecard %', size=14)
plt.title('COE Investments: San Mateo County vs. State of California', size=16)
# Display percentages on each bar
for i in range(len(state_percent)):
    v = state_percent[i]
    plt.text(i, v, str(v) + '%', ha='center', va='bottom')
# Create the legend
plt.legend(handles=[plt.bar(0, 0, color='green'), plt.bar(0, 0, color='red')],
           labels=['Yes', 'No'], title='San Mateo County', loc='upper right')
plt.show()
```



3 Section 3: COE-Level Focus

```
[ ]: s3_san_mateo_data = pd.read_csv('San Mateo data - section3.csv')
     s3_san_mateo_data
```

```
[ ]:                                         COE Investments  \
     0                     COE Environmental Coordinator
     1  COE Environmental Literacy and/or Sustainabili…
     2                               Climate Corps Fellow
     3              CAELI COE Innovation Hub Participants
     4                   CAELI COE Fellowship Participant
     5       CAELI COE Community of Practice Participants

       San Mateo County \r\n# of Staff or Initiatives  \
0                                       Yes (1+)
1                                       Yes (1+)
2                                        Yes (2)
3                                        Yes (1)
4                                        Yes (1)
5                                        Yes (1)

  State Scorecard\r\n# of COEs with one or more  \
0                                    7/58 = 12%
1                                   21/58 = 36%
2                                    3/58 =5.2%
3                                  8/58 = 13.8%
4                                 17/58 = 29.3%
5                                 36/58 = 62.1%

   % State Scorecard\r\n# of COEs with one or more
0                                             12.0
1                                             36.0
2                                              5.2
3                                             13.8
4                                             29.3
5                                             62.1
```

```
[ ]: s3_san_mateo_data.columns
```

```
[ ]: Index(['COE Investments', 'San Mateo County \r\n# of Staff or Initiatives',
            'State Scorecard\r\n# of COEs with one or more',
            '% State Scorecard\r\n# of COEs with one or more'],
           dtype='object')
```

```
[ ]: renamed_columns = {'San Mateo County \r\n# of Staff or Initiatives':'San Mateo␣
      ↪County # of Staff or Initiatives',
                        'State Scorecard\r\n# of COEs with one or more': 'State␣
      ↪Scorecard # of COEs with one or more',
                        '% State Scorecard\r\n# of COEs with one or more': '% State␣
      ↪Scorecard # of COEs with one or more'}
     s3_san_mateo_data.rename(columns=renamed_columns, inplace=True)
```

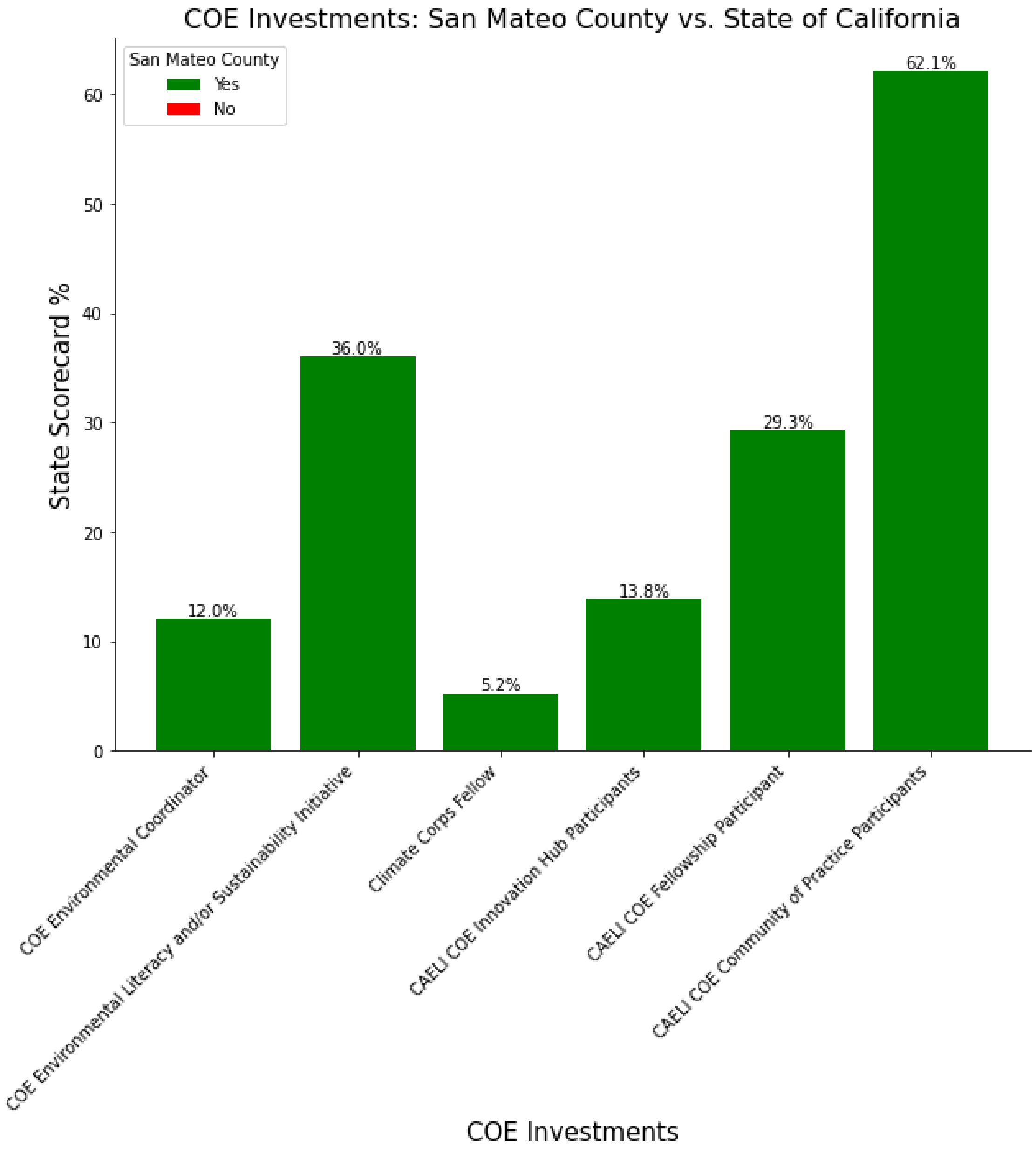
```
[ ]: s3_san_mateo_data['San Mateo County Staff or Initiatives'] = np.
      ↪where(s3_san_mateo_data['San Mateo County # of Staff or Initiatives'].str.
      ↪startswith('Yes'), 1, 0)
     s3_san_mateo_data
```

```
[ ]:                                    COE Investments  \
     0                       COE Environmental Coordinator
     1   COE Environmental Literacy and/or Sustainabili…
     2                                Climate Corps Fellow
```

```
3              CAELI COE Innovation Hub Participants
4                   CAELI COE Fellowship Participant
5       CAELI COE Community of Practice Participants

  San Mateo County # of Staff or Initiatives  \
0                                   Yes (1+)
1                                   Yes (1+)
2                                    Yes (2)
3                                    Yes (1)
4                                    Yes (1)
5                                    Yes (1)

  State Scorecard # of COEs with one or more  \
0                                 7/58 = 12%
1                                21/58 = 36%
2                                 3/58 =5.2%
3                               8/58 = 13.8%
4                              17/58 = 29.3%
5                              36/58 = 62.1%

   % State Scorecard # of COEs with one or more  \
0                                          12.0
1                                          36.0
2                                           5.2
3                                          13.8
4                                          29.3
5                                          62.1

   San Mateo County Staff or Initiatives
0                                      1
1                                      1
2                                      1
3                                      1
4                                      1
5                                      1
```

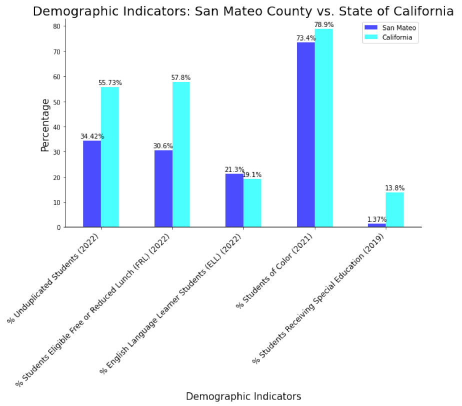
```
[ ]: investments = s3_san_mateo_data['COE Investments']
     san_mateo_county = s3_san_mateo_data['San Mateo County Staff or Initiatives']
     state_percent = s3_san_mateo_data['% State Scorecard # of COEs with one or␣
      ↪more']
     colors = ['green' if val else 'red' for val in san_mateo_county]
     plt.figure(figsize=(10, 8))
     plt.bar(investments, state_percent, color=colors)
     plt.xlabel('COE Investments', size=15)
     plt.xticks(rotation=45, ha='right')
     plt.ylabel('State Scorecard %', size=15)
     plt.title('COE Investments: San Mateo County vs. State of California', size=16)
```

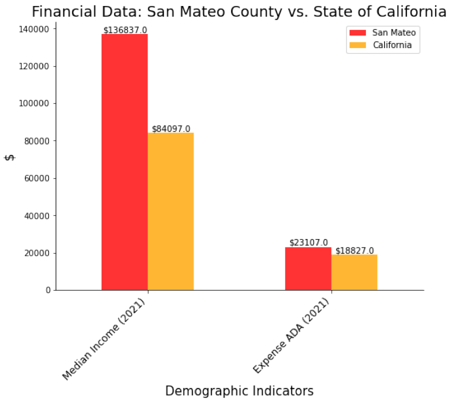
```
# Display percentages on each bar
for i in range(len(state_percent)):
    v = state_percent[i]
    plt.text(i, v, str(v) + '%', ha='center', va='bottom')
# Create the legend
plt.legend(handles=[plt.bar(0, 0, color='green'), plt.bar(0, 0, color='red')],
           labels=['Yes', 'No'], title='San Mateo County', loc='upper left')
plt.gca().spines['top'].set_visible(False)
plt.gca().spines['right'].set_visible(False)

plt.savefig('coe_initiatives.png', bbox_inches='tight')
plt.show()
```

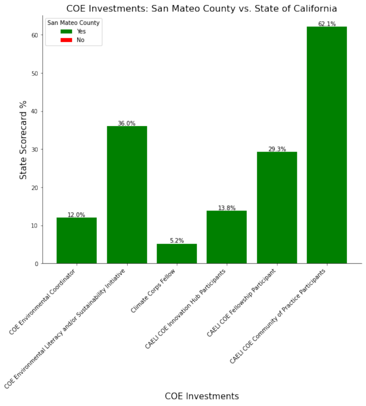


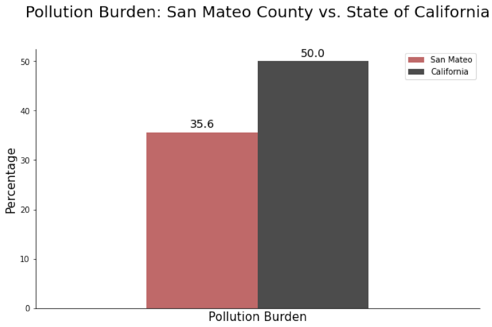
[ ]:

```
import matplotlib.image as mpimg
fig, axes = plt.subplots(nrows=2, ncols=2, figsize=(100, 80))

# Load the saved images for each subplot
img1 = mpimg.imread('demographic_indicators.png')
img2 = mpimg.imread('financial_indicators.png')
img3 = mpimg.imread('pollution_burden.png')
img4 = mpimg.imread('coe_initiatives.png')

axes[0, 0].imshow(img1)
axes[0, 0].axis('off')

axes[0, 1].imshow(img2)
axes[0, 1].axis('off')

axes[1, 0].imshow(img4)
axes[1, 0].axis('off')

axes[1, 1].imshow(img3)
axes[1, 1].axis('off')
# Adjust the layout to add padding and spacing between subplots
plt.subplots_adjust(wspace=0.3)

# Display the subplots together
plt.show()
```









4 Statistical Models

4.1 Problem 1: Which indicators contribute to the performance of a district in implementing ‘District-Wide Sustainability Initiatives’?

```
[ ]: # County data we have right now:
     # San Francisco, San Joaquin, San Mateo, Santa Cruz, Solano, San Diego
```

4.1.1 Data Cleaning

```
[ ]: san_francisco = pd.read_csv('San Francisco.csv')
     san_joaquin = pd.read_csv('San Joaquin.csv')
     san_mateo = pd.read_csv('San Mateo.csv')
     santa_cruz = pd.read_csv('Santa Cruz.csv')
     solano = pd.read_csv('Solano.csv')
     san_diego = pd.read_csv('San Diego.csv')
```

[ ]:
```
# Combine data
county_data = pd.concat([san_francisco, san_joaquin, san_mateo, santa_cruz,␣
 ↪solano, san_diego], axis=0)
county_data.shape
```

[ ]:
```
(104, 70)
```

[ ]:
```
# Drop districts that are no longer valid
county_data.dropna(subset=['District Type'], inplace=True)
```

[ ]:
```
county_data.shape
```

[ ]:
```
(98, 70)
```

[ ]:
```
county_data.columns
```

[ ]:
```
Index(['County', 'District Name', 'District Type', 'Grade Levels',
       'Number of Schools\n(2021-22)',
       'High School Partner District if Elementary',
       'Total # of Jurisdictions Per School District',
       'Jurisdiction Name\n(list on separate line for each jurisdiction)',
       'Student Enrollment \n(2021-22)',
       '# of Certificated Teachers\n(2018-19)',
       'Expense of Education per ADA \n(2020-21)',
       '% Unduplicated \n(2021-22)', '% FRM \n(2021-22)',
       '% English Learners \n(2021-22)',
       'Total GO Bonds and Parcel Taxes Attempted \n(2000 - 2029)',
       'Total GO Bonds and Parcel Taxes Passed \n(2000 - 2029)',
       'Most Recent Passed GO Bond Measure Year\n(2000-2029) ',
       'Total Amount of GO Bond Measure Funding ($)\n(2000-2029)',
       'Average CalEnviroScreen Pollution Burden',
       'Average Cal EnviroScreen Percentile',
       'Green Ribbon District\n1 (Yes) 0 (No)',
       'Green Ribbon Highest District Level Achievement',
       'Green Ribbon for Individual Schools within District\n1 (Yes) 0 (No)',
       'Board Policies Link',
       'Policies List \n(enter a separate line for each)',
       'BP: 3510 Green Schools Operations \n1 (Yes) 0 (No)',
       'BP: 3510 Year Adopted', 'BP: 3510 Most Recent Update/Revision',
       'BP: 3511 Energy And Water Management \n1 (Yes) 0 (No)',
       'BP: 3511 Year Adopted ', 'BP: 3511 Most Recent Update/Revision',
       'BP: 3511.1 Integrated Waste Management\n1 (Yes) 0 (No)',
       'BP: 3511.1 Year Adopted', 'BP: 3511.1 Most Recent Update/Revision',
       'BP: 3514 Environmental Safety\n1 (Yes) 0 (No)',
       'BP: 3514 Year Adopted', 'BP: 3514 Most Recent Update/Revision',
       'BP: 3514.1 Hazardous Substances\n1 (Yes) 0 (No)',
       'BP: 3514.1 Year Adopted', 'BP: 3514.1 Most Recent Update/Revision',
       'BP: 6142.5 Environmental Education\n1 (Yes) 0 (No)',
       'BP: 6142.5 Year Adopted', 'BP: 6142.5 Most Recent Update/Revision',
       'BP:7110 Facilities Master Plan\n1 (Yes) 0 (No)',
       'BP:7110 Year Adopted', 'BP: 7110 Most Recent Update/Revision',
       'Total Approved Policies',
       'Published Facilities Master Plan\n1 (Yes) 0 (No)', 'FMP Year Adopted',
       'FMP Year Most Recent Revision/Update', 'FMP Link',
       'Climate Change Resolutions / Climate Emergency Declarations\n1 (Yes) 0
(No)',
       'Climate Change Resolutions / Climate Emergency Declarations \nYear
Adopted ',
       'Climate Change Resolutions / Climate Emergency Declarations\nMost Recent
Update/Revision',
       'Climate Change Resolutions / Climate Emergency Declarations\nNotes',
       'Other Board Policies, Resolutions, and Declarations \n(list each as a
hyperlink and on a separate line)',
       'District-Wide Sustainability Initiatives\n1 (Yes) 0 (No)',
       'District-Wide Evidence of Campus Sustainability\n1 (Yes) 0 (No)',
       'District-Wide Evidence of Environmental Literacy Curriculum\n1 (Yes) 0
(No)',
       'District-Wide Evidence of Community and Culture Sustainability\n1 (Yes)
0 (No)',
       'District-Wide Sustainability Website\n1 (Yes) 0 (No)',
       'Sustainability Initiatives Notes\n(District and Site-Level)',
       'District-Wide Sustainability Staff\n1 (Yes) 0 (No)',
       'District-Wide Campus Sustainability Related Job\n1 (Yes) 0 (No)',
       'District-Wide Environmental Literacy Curriculum and Community and
Culture Related Job\n1 (Yes) 0 (No)',
       'Site-Level Environmental Literacy Curriculum and Community and Culture
Related Job \n1 (Yes) 0 (No)',
       'Sustainability Staff Notes\n(District and Site-Level)',
       'Most Recent Passed GO Bond Measure Year\n(2000-2029)',
       'BP: 3511 Year Adopted',
       'Climate Change Resolutions / Climate Emergency Declarations Year
Adopted'],
      dtype='object')
```

4.2 Logistic Regression Model

Manully select features for logitic regression model: 'District Type', 'Grade-Levels', 'Number of Schools', 'Total # of Jurisdictions Per School District', 'Student Enrollment', '# of Certificated Teachers', 'Expense of Education per ADA', '% Unduplicated (2021-22)', '% FRM', '% English Learners', 'Total GO Bonds and Parcel Taxes Passed (2000 - 2029)', 'Total Amount of GO Bond Measure Funding ($) (2000-2029)',

```
[ ]: selected_county_data = county_data[['District Type', 'Grade Levels',
            'Number of Schools\n(2021-22)',
            'Total # of Jurisdictions Per School District',
            'Student Enrollment \n(2021-22)',
            '# of Certificated Teachers\n(2018-19)',
            'Expense of Education per ADA \n(2020-21)',
            '% Unduplicated \n(2021-22)', '% FRM \n(2021-22)',
            '% English Learners \n(2021-22)',
            'Total GO Bonds and Parcel Taxes Passed \n(2000 - 2029)',
            'Total Amount of GO Bond Measure Funding ($)\n(2000-2029)',
            'Average CalEnviroScreen Pollution Burden',
            'Average Cal EnviroScreen Percentile',
            'Green Ribbon District\n1 (Yes) 0 (No)',
            'Green Ribbon for Individual Schools within District\n1 (Yes) 0 (No)',
            'Total Approved Policies',
            'Published Facilities Master Plan\n1 (Yes) 0 (No)',
            'District-Wide Sustainability Initiatives\n1 (Yes) 0 (No)',
            'District-Wide Sustainability Staff\n1 (Yes) 0 (No)']]
```

```
[ ]: selected_county_data.shape
```

```
[ ]: (98, 20)
```

```
[ ]: selected_county_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 98 entries, 0 to 41
Data columns (total 20 columns):
 #   Column                                                              Non-
Null Count  Dtype
---  ------
--------------  -----
 0   District Type                                                       98 non-
null     object
 1   Grade Levels                                                        98 non-
null     object
 2   Number of Schools
(2021-22)                                         98 non-null     float64
 3   Total # of Jurisdictions Per School District                        97 non-
null     float64
 4   Student Enrollment
(2021-22)                                       98 non-null     object
 5   # of Certificated Teachers
(2018-19)                                98 non-null     object
 6   Expense of Education per ADA
(2020-21)                             98 non-null     object
 7   % Unduplicated
(2021-22)                                           98 non-null     object
 8   % FRM
(2021-22)                                                    98 non-null
```

```
object
 9   % English Learners
(2021-22)                                       98 non-null     object
 10  Total GO Bonds and Parcel Taxes Passed
(2000 - 2029)                 98 non-null     object
 11  Total Amount of GO Bond Measure Funding ($)
(2000-2029)              89 non-null     object
 12  Average CalEnviroScreen Pollution Burden                            98 non-
null     object
 13  Average Cal EnviroScreen Percentile                                 98 non-
null     object
 14  Green Ribbon District
1 (Yes) 0 (No)                                98 non-null     float64
 15  Green Ribbon for Individual Schools within District
1 (Yes) 0 (No)  98 non-null     float64
 16  Total Approved Policies                                             96 non-
null     object
 17  Published Facilities Master Plan
1 (Yes) 0 (No)                      98 non-null     object
 18  District-Wide Sustainability Initiatives
1 (Yes) 0 (No)              98 non-null     float64
 19  District-Wide Sustainability Staff
1 (Yes) 0 (No)                    98 non-null     float64
dtypes: float64(6), object(14)
memory usage: 16.1+ KB
```

4.3 Data cleaning list:

1. N/A values: replace with average value
2. $ and , : remove the signs and convert to float
3. ’*’ values: remove the rows
4. %: remove ‘%’

[ ]:
```python
# Examine NULL values
#print(selected_county_data.isnull().sum())
```

[ ]:
```python
copy = selected_county_data.copy()
```

[ ]:
```python
def find_columns_with_star(dataframe):
    # Use the any() method to check for columns containing the value '*'
    columns_with_star = dataframe.columns[dataframe.eq('*').any()]

    return columns_with_star
```

4.4 Remove rows with ’*’

[ ]:
```python
find_columns_with_star(copy)
```

```
[ ]: Index(['Grade Levels', 'Student Enrollment \n(2021-22)',
       '# of Certificated Teachers\n(2018-19)',
       'Expense of Education per ADA \n(2020-21)',
       '% Unduplicated \n(2021-22)', '% FRM \n(2021-22)',
       '% English Learners \n(2021-22)',
       'Total GO Bonds and Parcel Taxes Passed \n(2000 - 2029)',
       'Total Amount of GO Bond Measure Funding ($)\n(2000-2029)',
       'Average CalEnviroScreen Pollution Burden',
       'Average Cal EnviroScreen Percentile', 'Total Approved Policies'],
      dtype='object')
```

```
[ ]: def remove_rows_with_star(dataframe):
         # Use boolean indexing to filter rows without the value '*'
         dataframe_no_star = dataframe[~dataframe.apply(lambda row: row.eq('*')).
      ↪any(axis=1)]

         return dataframe_no_star
```

```
[ ]: copy = remove_rows_with_star(copy)
```

```
[ ]: find_columns_with_star(copy)
```

```
[ ]: Index([], dtype='object')
```

4.5 Replace all $ and , with ’’

```
[ ]: # Replace all $ and , with ''
     copy = copy.replace(r'[$,]', '', regex=True)
```

```
[ ]: print(copy.isnull().sum())
```

```
District Type                                                          0
Grade Levels                                                           0
Number of Schools\n(2021-22)                                           0
Total # of Jurisdictions Per School District                           0
Student Enrollment \n(2021-22)                                         0
# of Certificated Teachers\n(2018-19)                                  0
Expense of Education per ADA \n(2020-21)                               0
% Unduplicated \n(2021-22)                                             0
% FRM \n(2021-22)                                                      0
% English Learners \n(2021-22)                                         0
Total GO Bonds and Parcel Taxes Passed \n(2000 - 2029)                 0
Total Amount of GO Bond Measure Funding ($)\n(2000-2029)               8
Average CalEnviroScreen Pollution Burden                               0
Average Cal EnviroScreen Percentile                                    0
Green Ribbon District\n1 (Yes) 0 (No)                                  0
Green Ribbon for Individual Schools within District\n1 (Yes) 0 (No)    0
Total Approved Policies                                                1
```

```
Published Facilities Master Plan\n1 (Yes) 0 (No)                  0
District-Wide Sustainability Initiatives\n1 (Yes) 0 (No)          0
District-Wide Sustainability Staff\n1 (Yes) 0 (No)                0
dtype: int64
```

```
[ ]: def find_null_rows(dataframe, column_name):
         # Use boolean indexing to filter rows with null values in the specified␣
      ↪column
         null_rows = dataframe[dataframe[column_name].isnull()]

         # Get the values of the null rows in the specified column
         null_values = null_rows[column_name]

         # Combine the null_rows DataFrame and the null_values Series to get the␣
      ↪result
         result = pd.concat([null_rows, null_values.rename('Null Values')], axis=1)

         return result
```

```
[ ]: find_null_rows(copy, 'Total Approved Policies')
```

```
[ ]:                  District Type Grade Levels  Number of Schools\n(2021-22)  \
     27  Elementary School District         K-08                          2.0

         Total # of Jurisdictions Per School District  \
     27                                           1.0

        Student Enrollment \n(2021-22) # of Certificated Teachers\n(2018-19)  \
     27                            589                                     66

        Expense of Education per ADA \n(2020-21) % Unduplicated \n(2021-22)  \
     27                                    24827                     8.83 %

        % FRM \n(2021-22) % English Learners \n(2021-22)  …  \
     27            4.6 %                           4.8 %  …

        Total Amount of GO Bond Measure Funding ($)\n(2000-2029)  \
     27                                           84800000

        Average CalEnviroScreen Pollution Burden  \
     27                                  36.5913

        Average Cal EnviroScreen Percentile Green Ribbon District\n1 (Yes) 0 (No)  \
     27                                41.0                                  0.0

         Green Ribbon for Individual Schools within District\n1 (Yes) 0 (No)  \
     27                                                0.0
```

```
    Total Approved Policies Published Facilities Master Plan\n1 (Yes) 0 (No)  \
27                    NaN                                             1.0

   District-Wide Sustainability Initiatives\n1 (Yes) 0 (No)  \
27                                                0.0

    District-Wide Sustainability Staff\n1 (Yes) 0 (No)  Null Values
27                                                0.0          NaN

[1 rows x 21 columns]
```

4.6 Replace NaN value with an average value

[ ]:

```python
# Replace the NaN value in 'Total Approved Policies' with a median value
def replace_nan_with_median(dataframe, column_name):
    # Calculate the median value of the column excluding NaN values
    median_value = dataframe[column_name].median(skipna=True)

    # Use fillna() to replace NaN values with the calculated median value
    dataframe[column_name].fillna(median_value, inplace=True)

    return dataframe
```

[ ]:

```python
copy = replace_nan_with_median(copy, 'Total Approved Policies')
```

[ ]:

```python
# Replace the NaN value in 'Total Amount of GO Bond Measure Funding␣
 ↪($)\n(2000-2029)' with a median value
def convert_string_to_float(dataframe, column_name):
    # Use to_numeric() to convert non-null string values to float, and skip␣
 ↪null (NaN) values
    dataframe[column_name] = pd.to_numeric(dataframe[column_name],␣
 ↪errors='coerce')

    return dataframe
```

[ ]:

```python
copy = convert_string_to_float(copy, 'Total Amount of GO Bond Measure Funding␣
 ↪($)\n(2000-2029)')
```

[ ]:

```python
copy = replace_nan_with_median(copy, 'Total Amount of GO Bond Measure Funding␣
 ↪($)\n(2000-2029)')
```

4.7 Remove %

```
[ ]: def remove_percent_sign(dataframe, column_name):
         # Use str.replace() to remove the '%' sign and then convert the values to␣
      ↪float
         dataframe[column_name] = dataframe[column_name].str.replace('%', '').
      ↪astype(float)

         return dataframe
```

```
[ ]: print(copy.isnull().sum())
```

```
District Type                                                         0
Grade Levels                                                          0
Number of Schools\n(2021-22)                                          0
Total # of Jurisdictions Per School District                          0
Student Enrollment \n(2021-22)                                        0
# of Certificated Teachers\n(2018-19)                                 0
Expense of Education per ADA \n(2020-21)                              0
% Unduplicated \n(2021-22)                                            0
% FRM \n(2021-22)                                                     0
% English Learners \n(2021-22)                                        0
Total GO Bonds and Parcel Taxes Passed \n(2000 - 2029)                0
Total Amount of GO Bond Measure Funding ($)\n(2000-2029)              0
Average CalEnviroScreen Pollution Burden                              0
Average Cal EnviroScreen Percentile                                   0
Green Ribbon District\n1 (Yes) 0 (No)                                 0
Green Ribbon for Individual Schools within District\n1 (Yes) 0 (No)   0
Total Approved Policies                                               0
Published Facilities Master Plan\n1 (Yes) 0 (No)                      0
District-Wide Sustainability Initiatives\n1 (Yes) 0 (No)              0
District-Wide Sustainability Staff\n1 (Yes) 0 (No)                    0
dtype: int64
```

```
[ ]: copy = remove_percent_sign(copy, '% Unduplicated \n(2021-22)')
     copy = remove_percent_sign(copy, '% FRM \n(2021-22)')
     copy = remove_percent_sign(copy, '% English Learners \n(2021-22)')
     # copy = remove_percent_sign(copy, 'Average Cal EnviroScreen Percentile')
```

```
[ ]: copy.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 81 entries, 0 to 39
Data columns (total 20 columns):
 #   Column                                                              Non-
Null Count  Dtype
---  ------
--------------  -----
 0   District Type                                                       81 non-
```

```
null     object
 1   Grade Levels                                                       81 non-
null     object
 2   Number of Schools
(2021-22)                                         81 non-null     float64
 3   Total # of Jurisdictions Per School District                       81 non-
null     float64
 4   Student Enrollment
(2021-22)                                       81 non-null     object
 5   # of Certificated Teachers
(2018-19)                                81 non-null     object
 6   Expense of Education per ADA
(2020-21)                             81 non-null     object
 7   % Unduplicated
(2021-22)                                           81 non-null     float64
 8   % FRM
(2021-22)                                                    81 non-null
float64
 9   % English Learners
(2021-22)                                       81 non-null     float64
 10  Total GO Bonds and Parcel Taxes Passed
(2000 - 2029)               81 non-null     object
 11  Total Amount of GO Bond Measure Funding ($)
(2000-2029)             81 non-null     float64
 12  Average CalEnviroScreen Pollution Burden                           81 non-
null     object
 13  Average Cal EnviroScreen Percentile                                81 non-
null     object
 14  Green Ribbon District
1 (Yes) 0 (No)                                81 non-null     float64
 15  Green Ribbon for Individual Schools within District
1 (Yes) 0 (No)  81 non-null     float64
 16  Total Approved Policies                                            81 non-
null     object
 17  Published Facilities Master Plan
1 (Yes) 0 (No)                     81 non-null     object
 18  District-Wide Sustainability Initiatives
1 (Yes) 0 (No)             81 non-null     float64
 19  District-Wide Sustainability Staff
1 (Yes) 0 (No)                   81 non-null     float64
dtypes: float64(10), object(10)
memory usage: 13.3+ KB
```

4.8 Convert values to int

```
[ ]: def replace_value_in_column(df, column_name, old_value, new_value):
         df[column_name] = df[column_name].replace(old_value, new_value)
         return df
```

```
[ ]: def convert_to_int(dataframe, column_name):
         dataframe[column_name] = dataframe[column_name].astype(int)
         return dataframe
```

```
[ ]: def convert_to_float(dataframe, column_name):
         dataframe[column_name] = dataframe[column_name].astype(float)
         return dataframe
```

```
[ ]: copy = replace_value_in_column(copy, 'Published Facilities Master Plan\n1 (Yes)␣
       ↪0 (No)', 'Yes', 1)
```

```
[ ]: copy = convert_to_int(copy, 'Student Enrollment \n(2021-22)')
     copy = convert_to_int(copy, '# of Certificated Teachers\n(2018-19)')
     copy = convert_to_float(copy, 'Expense of Education per ADA \n(2020-21)')
     copy = convert_to_int(copy, 'Total GO Bonds and Parcel Taxes Passed \n(2000 -␣
       ↪2029)')
     copy = convert_to_float(copy, 'Average CalEnviroScreen Pollution Burden')
     copy = convert_to_int(copy, 'Total Approved Policies')
     copy = convert_to_int(copy, 'Published Facilities Master Plan\n1 (Yes) 0 (No)')
     copy = convert_to_float(copy, 'Average Cal EnviroScreen Percentile')
```

```
[ ]: copy = convert_to_int(copy, 'District-Wide Sustainability Initiatives\n1 (Yes)␣
       ↪0 (No)')
```

```
[ ]: copy
```

```
[ ]:                  District Type Grade Levels  Number of Schools\n(2021-22)  \
     0      Unified School District         K-12                        126.0
     0      Unified School District        TK-12                         68.0
     1      Unified School District         K-12                         54.0
     3      Unified School District         K-12                         23.0
     4      Unified School District         K-12                         14.0
     ..                         …           …                           …
     34  Elementary School District         K-06                          7.0
     35  Elementary School District         K-08                         13.0
     37        High School District         K-12                         31.0
     38  Elementary School District         K-08                          1.0
     39     Unified School District         K-12                         33.0

         Total # of Jurisdictions Per School District  \
     0                                            1.0
     0                                            1.0
```

```
1                                            5.0
3                                            1.0
4                                            1.0
..                                           …
34                                           3.0
35                                           3.0
37                                           4.0
38                                           1.0
39                                           2.0

    Student Enrollment \n(2021-22)  # of Certificated Teachers\n(2018-19)  \
0                            55592                                   3886
0                            39803                                   1732
1                            30727                                   1614
3                            15398                                    729
4                             8967                                    442
..                             …                                    …
34                            2820                                    185
35                            6119                                    334
37                           38026                                   1893
38                             178                                     12
39                           22092                                   1191

    Expense of Education per ADA \n(2020-21)  % Unduplicated \n(2021-22)  \
0                                    29258.0                        52.22
0                                    20943.0                        72.41
1                                    18601.0                        63.40
3                                    15391.0                        52.24
4                                    15475.0                        57.54
..                                       …                         …
34                                   21793.0                        20.74
35                                   26453.0                        58.95
37                                   16947.0                        61.17
38                                   23672.0                        85.96
39                                   18238.0                        58.98

    % FRM \n(2021-22)  % English Learners \n(2021-22)  \
0                50.4                            26.3
0                79.2                            24.1
1                65.3                            20.2
3                47.9                            24.9
4                55.5                            12.1
..                …                             …
34               12.2                            11.4
35               65.7                            47.4
37               50.6                            22.9
38               83.7                            53.9
```

```
39              64.2                         17.3

   Total GO Bonds and Parcel Taxes Passed \n(2000 - 2029)  \
0                                                      8
0                                                      6
1                                                      3
3                                                      3
4                                                      2
..                                                   …
34                                                     1
35                                                     3
37                                                     3
38                                                     0
39                                                     2

   Total Amount of GO Bond Measure Funding ($)\n(2000-2029)  \
0                                        2.020250e+09
0                                        1.090880e+09
1                                        3.904000e+08
3                                        5.100000e+07
4                                        9.850000e+07
..                                                …
34                                       1.050000e+08
35                                       1.034000e+08
37                                       1.234000e+09
38                                       1.034000e+08
39                                       3.870000e+08

   Average CalEnviroScreen Pollution Burden  \
0                                   35.8232
0                                   51.1939
1                                   41.4561
3                                   39.2630
4                                   45.0943
..                                      …
34                                  29.6973
35                                  40.6387
37                                  39.8283
38                                  44.5441
39                                  36.2005

   Average Cal EnviroScreen Percentile  \
0                                37.82
0                                81.00
1                                54.15
3                                48.34
4                                65.22
```

```
..                                  …
34                                18.00
35                                52.49
37                                50.00
38                                63.56
39                                39.00

    Green Ribbon District\n1 (Yes) 0 (No)  \
0                                     1.0
0                                     0.0
1                                     0.0
3                                     0.0
4                                     0.0
..                                    …
34                                    0.0
35                                    0.0
37                                    0.0
38                                    0.0
39                                    0.0

    Green Ribbon for Individual Schools within District\n1 (Yes) 0 (No)  \
0                                                 1.0
0                                                 0.0
1                                                 0.0
3                                                 0.0
4                                                 0.0
..                                                …
34                                                0.0
35                                                0.0
37                                                0.0
38                                                0.0
39                                                1.0

    Total Approved Policies  Published Facilities Master Plan\n1 (Yes) 0 (No)  \
0                         2                                                 1
0                         7                                                 0
1                         3                                                 1
3                         4                                                 0
4                         3                                                 0
..                      …                                               …
34                        6                                                 1
35                        7                                                 1
37                        6                                                 1
38                        2                                                 0
39                        6                                                 1

    District-Wide Sustainability Initiatives\n1 (Yes) 0 (No)  \
0                                                   1
0                                                   1
1                                                   1
3                                                   0
4                                                   1
..                                                 …
34                                                  0
35                                                  0
37                                                  1
38                                                  0
39                                                  1

    District-Wide Sustainability Staff\n1 (Yes) 0 (No)
0                                                 1.0
0                                                 0.0
1                                                 0.0
3                                                 0.0
4                                                 1.0
..                                                 …
34                                                0.0
35                                                0.0
37                                                1.0
38                                                0.0
39                                                0.0

[81 rows x 20 columns]
```

```
[ ]: print(copy.isnull().sum())
```

```
District Type                                                           0
Grade Levels                                                            0
Number of Schools\n(2021-22)                                            0
Total # of Jurisdictions Per School District                            0
Student Enrollment \n(2021-22)                                          0
# of Certificated Teachers\n(2018-19)                                   0
Expense of Education per ADA \n(2020-21)                                0
% Unduplicated \n(2021-22)                                              0
% FRM \n(2021-22)                                                       0
% English Learners \n(2021-22)                                          0
Total GO Bonds and Parcel Taxes Passed \n(2000 - 2029)                  0
Total Amount of GO Bond Measure Funding ($)\n(2000-2029)                0
Average CalEnviroScreen Pollution Burden                                0
Average Cal EnviroScreen Percentile                                     0
Green Ribbon District\n1 (Yes) 0 (No)                                   0
Green Ribbon for Individual Schools within District\n1 (Yes) 0 (No)     0
Total Approved Policies                                                 0
Published Facilities Master Plan\n1 (Yes) 0 (No)                        0
District-Wide Sustainability Initiatives\n1 (Yes) 0 (No)                0
```

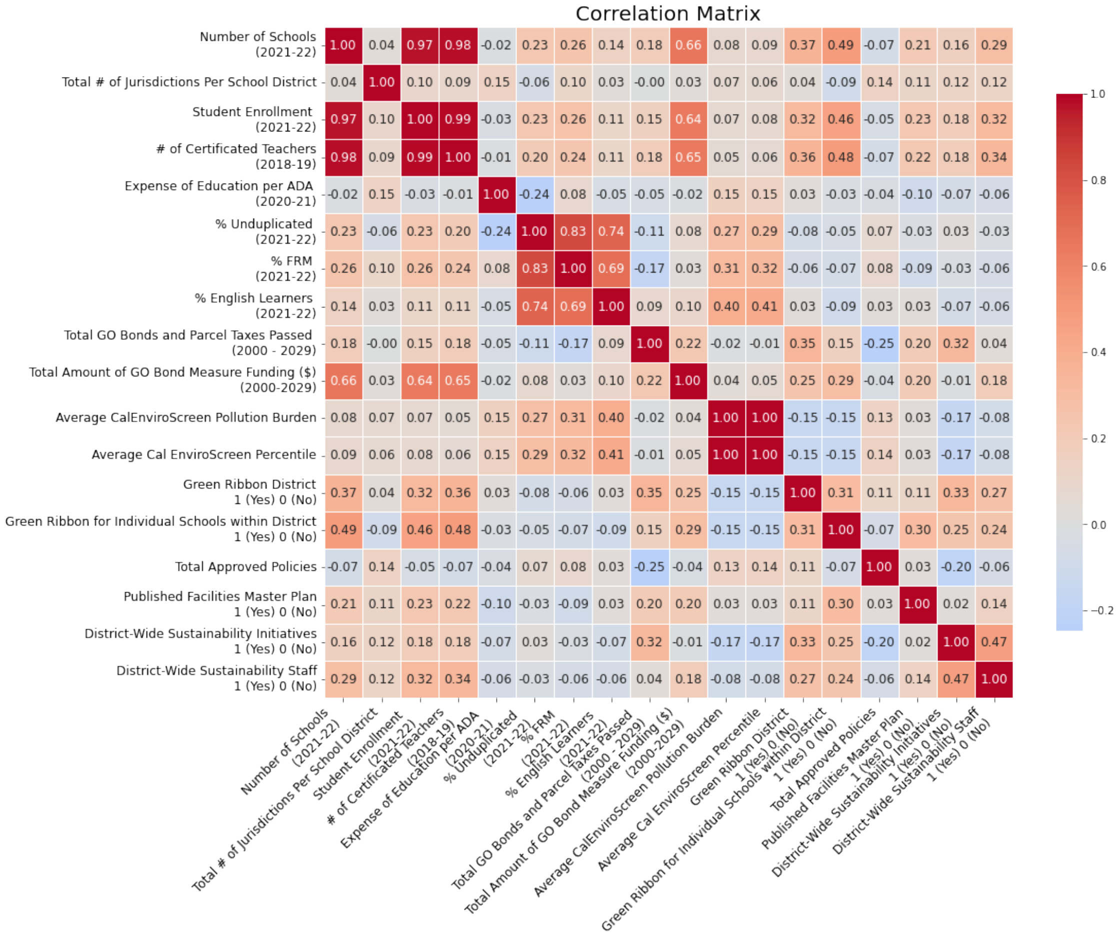
```
District-Wide Sustainability Staff\n1 (Yes) 0 (No)                     0
dtype: int64
```

4.9 EDA

```
[ ]: copy.columns
```

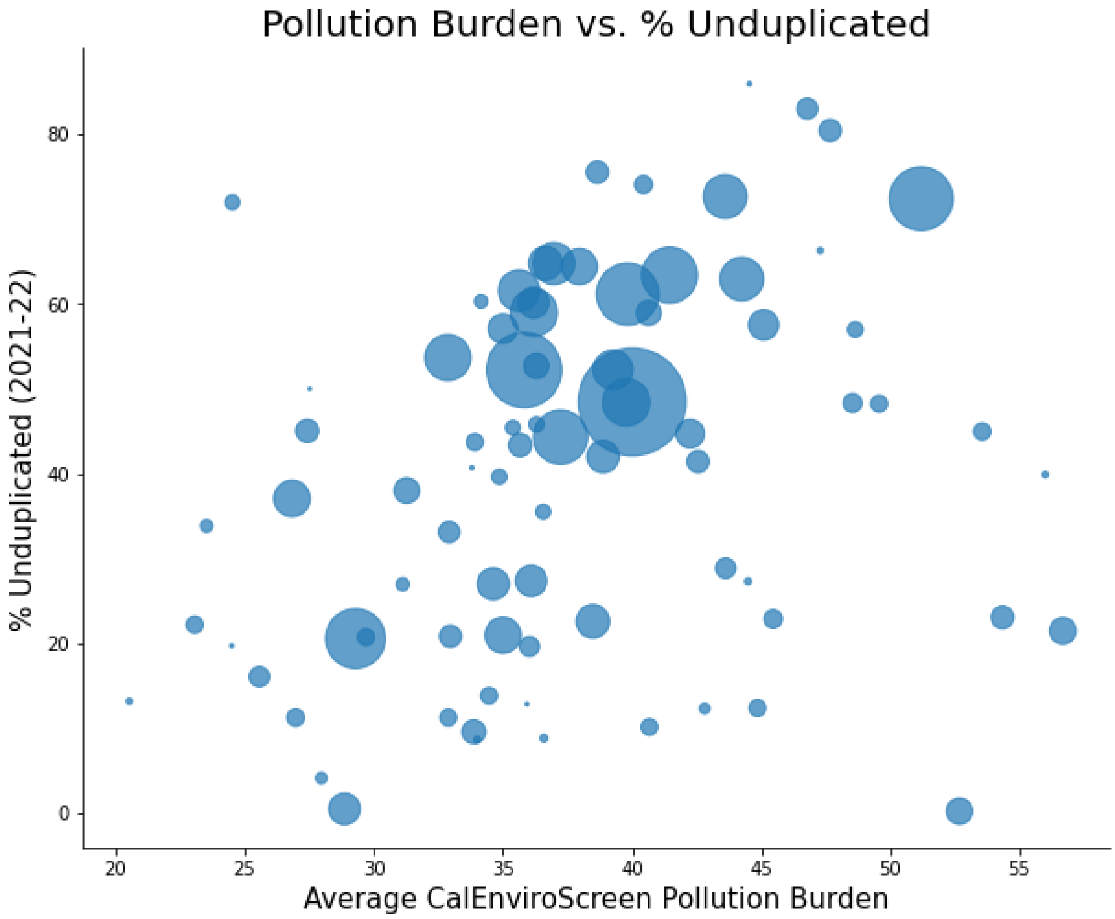
```
[ ]: Index(['District Type', 'Grade Levels', 'Number of Schools\n(2021-22)',
            'Total # of Jurisdictions Per School District',
            'Student Enrollment \n(2021-22)',
            '# of Certificated Teachers\n(2018-19)',
            'Expense of Education per ADA \n(2020-21)',
            '% Unduplicated \n(2021-22)', '% FRM \n(2021-22)',
            '% English Learners \n(2021-22)',
            'Total GO Bonds and Parcel Taxes Passed \n(2000 - 2029)',
            'Total Amount of GO Bond Measure Funding ($)\n(2000-2029)',
            'Average CalEnviroScreen Pollution Burden',
            'Average Cal EnviroScreen Percentile',
            'Green Ribbon District\n1 (Yes) 0 (No)',
            'Green Ribbon for Individual Schools within District\n1 (Yes) 0 (No)',
            'Total Approved Policies',
            'Published Facilities Master Plan\n1 (Yes) 0 (No)',
            'District-Wide Sustainability Initiatives\n1 (Yes) 0 (No)',
            'District-Wide Sustainability Staff\n1 (Yes) 0 (No)'],
           dtype='object')
```

```
[ ]: # Assuming you have a DataFrame named 'df'
     correlation_matrix = copy.corr()

     # Create a larger heatmap of the correlation matrix with rotated x-axis labels
     plt.figure(figsize=(15, 12))
     sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', center=0, fmt=".
      ↪2f", linewidths=0.5, xticklabels=correlation_matrix.columns,␣
      ↪yticklabels=correlation_matrix.columns, annot_kws={"size": 12},␣
      ↪cbar_kws={"shrink": 0.8})
     plt.title('Correlation Matrix', size=20)
     plt.xticks(rotation=45, ha='right', size=12)
     plt.yticks(rotation=0, size=12)
     plt.show()
```



```
[ ]: x_values = copy['Average CalEnviroScreen Pollution Burden']
     y_values = copy['% Unduplicated \n(2021-22)']
     dot_sizes = copy['Student Enrollment \n(2021-22)']
     scaling_factor = 0.03
     dot_sizes_scaled = dot_sizes * scaling_factor

     plt.figure(figsize=(10, 8))
     # Create the scatter plot
     plot = plt.scatter(x_values, y_values, s=dot_sizes_scaled, alpha=0.7)

     # Set labels and title
     plt.xlabel('Average CalEnviroScreen Pollution Burden', size=15)
     plt.ylabel('% Unduplicated (2021-22)', size=15)
     plt.title('Pollution Burden vs. % Unduplicated', size=20)
     #plt.legend(*plot.legend_elements("sizes", num=6), loc='best', fontsize=20,␣
      ↪prop={'size': 13})
```

```
plt.gca().spines['top'].set_visible(False)
plt.gca().spines['right'].set_visible(False)
# Show the plot
plt.show()
```



Use VIF to remove multicollinear (highly correlated) features

```
[ ]: # Feature selection: Calculate Varinace Inflation Factor for each feature
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import variance_inflation_factor

# The dataframe passed to VIF must include the intercept term. We add it the same way we did before.
def VIF(df, columns):
    values = sm.add_constant(df[columns]).values
    num_columns = len(columns)+1
    vif = [variance_inflation_factor(values, i) for i in range(num_columns)]
    return pd.Series(vif[1:], index=columns)
```

```
[ ]: copy.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 81 entries, 0 to 39
Data columns (total 20 columns):
 #   Column                                                              Non-
Null Count  Dtype
---  ------
--------------  -----
 0   District Type                                                       81 non-
null     object
 1   Grade Levels                                                        81 non-
null     object
 2   Number of Schools
(2021-22)                                         81 non-null     float64
 3   Total # of Jurisdictions Per School District                        81 non-
null     float64
 4   Student Enrollment
(2021-22)                                       81 non-null     int64
 5   # of Certificated Teachers
(2018-19)                                81 non-null     int64
 6   Expense of Education per ADA
(2020-21)                             81 non-null     float64
 7   % Unduplicated
(2021-22)                                           81 non-null     float64
 8   % FRM
(2021-22)                                                    81 non-null
float64
 9   % English Learners
(2021-22)                                       81 non-null     float64
 10  Total GO Bonds and Parcel Taxes Passed
(2000 - 2029)               81 non-null     int64
 11  Total Amount of GO Bond Measure Funding ($)
(2000-2029)             81 non-null     float64
 12  Average CalEnviroScreen Pollution Burden                            81 non-
null     float64
 13  Average Cal EnviroScreen Percentile                                 81 non-
null     float64
 14  Green Ribbon District
1 (Yes) 0 (No)                                81 non-null     float64
 15  Green Ribbon for Individual Schools within District
1 (Yes) 0 (No)  81 non-null     float64
 16  Total Approved Policies                                             81 non-
null     int64
 17  Published Facilities Master Plan
1 (Yes) 0 (No)                     81 non-null     int64
 18  District-Wide Sustainability Initiatives
1 (Yes) 0 (No)             81 non-null     int64
```

```
 19  District-Wide Sustainability Staff
1 (Yes) 0 (No)                   81 non-null     float64
dtypes: float64(12), int64(6), object(2)
memory usage: 13.3+ KB
```

[ ]:
```
features = ['Number of Schools\n(2021-22)',
       'Total # of Jurisdictions Per School District',
       'Student Enrollment \n(2021-22)',
       '# of Certificated Teachers\n(2018-19)',
       'Expense of Education per ADA \n(2020-21)',
       '% Unduplicated \n(2021-22)', '% FRM \n(2021-22)',
       '% English Learners \n(2021-22)',
       'Total GO Bonds and Parcel Taxes Passed \n(2000 - 2029)',
       'Total Amount of GO Bond Measure Funding ($)\n(2000-2029)',
       'Average CalEnviroScreen Pollution Burden',
       'Average Cal EnviroScreen Percentile',
       'Green Ribbon District\n1 (Yes) 0 (No)',
       'Green Ribbon for Individual Schools within District\n1 (Yes) 0 (No)',
       'Total Approved Policies',
       'Published Facilities Master Plan\n1 (Yes) 0 (No)',
       'District-Wide Sustainability Staff\n1 (Yes) 0 (No)']
```

[ ]:
```
VIF(copy, features)
```

[ ]:
```
Number of Schools\n(2021-22)
28.295810
Total # of Jurisdictions Per School District
1.263469
Student Enrollment \n(2021-22)
52.508443
# of Certificated Teachers\n(2018-19)
79.176331
Expense of Education per ADA \n(2020-21)
1.572123
% Unduplicated \n(2021-22)
6.231978
% FRM \n(2021-22)
5.466021
% English Learners \n(2021-22)
3.212513
Total GO Bonds and Parcel Taxes Passed \n(2000 - 2029)
1.515372
Total Amount of GO Bond Measure Funding ($)\n(2000-2029)
1.979108
Average CalEnviroScreen Pollution Burden
201.570215
Average Cal EnviroScreen Percentile
```

```
207.258485
Green Ribbon District\n1 (Yes) 0 (No)
1.618433
Green Ribbon for Individual Schools within District\n1 (Yes) 0 (No)
1.639390
Total Approved Policies
1.319583
Published Facilities Master Plan\n1 (Yes) 0 (No)
1.243427
District-Wide Sustainability Staff\n1 (Yes) 0 (No)
1.325354
dtype: float64
```

```
[ ]: # Remove 'Average Cal EnviroScreen Percentile'
     features = ['Number of Schools\n(2021-22)',
            'Total # of Jurisdictions Per School District',
            'Student Enrollment \n(2021-22)',
            '# of Certificated Teachers\n(2018-19)',
            'Expense of Education per ADA \n(2020-21)',
            '% Unduplicated \n(2021-22)', '% FRM \n(2021-22)',
            '% English Learners \n(2021-22)',
            'Total GO Bonds and Parcel Taxes Passed \n(2000 - 2029)',
            'Total Amount of GO Bond Measure Funding ($)\n(2000-2029)',
            'Average CalEnviroScreen Pollution Burden',
            'Green Ribbon District\n1 (Yes) 0 (No)',
            'Green Ribbon for Individual Schools within District\n1 (Yes) 0 (No)',
            'Total Approved Policies',
            'Published Facilities Master Plan\n1 (Yes) 0 (No)',
            'District-Wide Sustainability Staff\n1 (Yes) 0 (No)']
     VIF(copy, features)
```

```
[ ]: Number of Schools\n(2021-22)                                           28.086493
     Total # of Jurisdictions Per School District                            1.243287
     Student Enrollment \n(2021-22)                                         52.508412
     # of Certificated Teachers\n(2018-19)                                  79.090970
     Expense of Education per ADA \n(2020-21)                                1.537045
     % Unduplicated \n(2021-22)                                              6.063012
     % FRM \n(2021-22)                                                       5.375522
     % English Learners \n(2021-22)                                          3.171145
     Total GO Bonds and Parcel Taxes Passed \n(2000 - 2029)                  1.482648
     Total Amount of GO Bond Measure Funding ($)\n(2000-2029)                1.977600
     Average CalEnviroScreen Pollution Burden                                1.390051
     Green Ribbon District\n1 (Yes) 0 (No)                                   1.608846
     Green Ribbon for Individual Schools within District\n1 (Yes) 0 (No)     1.602997
     Total Approved Policies                                                 1.251052
     Published Facilities Master Plan\n1 (Yes) 0 (No)                        1.242995
     District-Wide Sustainability Staff\n1 (Yes) 0 (No)                      1.308100
```

```
dtype: float64
```

[ ]:
```
# Remove '# of Certificated Teachers\n(2018-19)'
features = ['Number of Schools\n(2021-22)',
       'Total # of Jurisdictions Per School District',
       'Student Enrollment \n(2021-22)',
       'Expense of Education per ADA \n(2020-21)',
       '% Unduplicated \n(2021-22)', '% FRM \n(2021-22)',
       '% English Learners \n(2021-22)',
       'Total GO Bonds and Parcel Taxes Passed \n(2000 - 2029)',
       'Total Amount of GO Bond Measure Funding ($)\n(2000-2029)',
       'Average CalEnviroScreen Pollution Burden',
       'Green Ribbon District\n1 (Yes) 0 (No)',
       'Green Ribbon for Individual Schools within District\n1 (Yes) 0 (No)',
       'Total Approved Policies',
       'Published Facilities Master Plan\n1 (Yes) 0 (No)',
       'District-Wide Sustainability Staff\n1 (Yes) 0 (No)']
VIF(copy, features)
```

[ ]:
```
Number of Schools\n(2021-22)                                           19.456841
Total # of Jurisdictions Per School District                            1.241360
Student Enrollment \n(2021-22)                                         18.125506
Expense of Education per ADA \n(2020-21)                                1.533057
% Unduplicated \n(2021-22)                                              5.863381
% FRM \n(2021-22)                                                       5.356284
% English Learners \n(2021-22)                                          3.112088
Total GO Bonds and Parcel Taxes Passed \n(2000 - 2029)                  1.479307
Total Amount of GO Bond Measure Funding ($)\n(2000-2029)                1.975609
Average CalEnviroScreen Pollution Burden                                1.368176
Green Ribbon District\n1 (Yes) 0 (No)                                   1.607082
Green Ribbon for Individual Schools within District\n1 (Yes) 0 (No)     1.585243
Total Approved Policies                                                 1.246469
Published Facilities Master Plan\n1 (Yes) 0 (No)                        1.241873
District-Wide Sustainability Staff\n1 (Yes) 0 (No)                      1.241047
dtype: float64
```

[ ]:
```
# Remove 'Number of Schools\n(2021-22)'
features = ['Total # of Jurisdictions Per School District',
       'Student Enrollment \n(2021-22)',
       'Expense of Education per ADA \n(2020-21)',
       '% Unduplicated \n(2021-22)', '% FRM \n(2021-22)',
       '% English Learners \n(2021-22)',
       'Total GO Bonds and Parcel Taxes Passed \n(2000 - 2029)',
       'Total Amount of GO Bond Measure Funding ($)\n(2000-2029)',
       'Average CalEnviroScreen Pollution Burden',
       'Green Ribbon District\n1 (Yes) 0 (No)',
       'Green Ribbon for Individual Schools within District\n1 (Yes) 0 (No)',
       'Total Approved Policies',
       'Published Facilities Master Plan\n1 (Yes) 0 (No)',
       'District-Wide Sustainability Staff\n1 (Yes) 0 (No)']
VIF(copy, features)
```

[ ]:

```
Total # of Jurisdictions Per School District                           1.199207
Student Enrollment \n(2021-22)                                         2.755124
Expense of Education per ADA \n(2020-21)                               1.532859
% Unduplicated \n(2021-22)                                             5.852168
% FRM \n(2021-22)                                                      5.351808
% English Learners \n(2021-22)                                         3.072474
Total GO Bonds and Parcel Taxes Passed \n(2000 - 2029)                 1.477940
Total Amount of GO Bond Measure Funding ($)\n(2000-2029)               1.890793
Average CalEnviroScreen Pollution Burden                               1.359594
Green Ribbon District\n1 (Yes) 0 (No)                                  1.534187
Green Ribbon for Individual Schools within District\n1 (Yes) 0 (No)    1.513521
Total Approved Policies                                                1.233343
Published Facilities Master Plan\n1 (Yes) 0 (No)                       1.229213
District-Wide Sustainability Staff\n1 (Yes) 0 (No)                     1.234292
dtype: float64
```

[ ]:

```
from sklearn.model_selection import train_test_split

y = copy['District-Wide Sustainability Initiatives\n1 (Yes) 0 (No)']
X = copy.loc[:, ['Total # of Jurisdictions Per School District',
       'Student Enrollment \n(2021-22)',
       'Expense of Education per ADA \n(2020-21)',
       '% Unduplicated \n(2021-22)', '% FRM \n(2021-22)',
       '% English Learners \n(2021-22)',
       'Total GO Bonds and Parcel Taxes Passed \n(2000 - 2029)',
       'Total Amount of GO Bond Measure Funding ($)\n(2000-2029)',
       'Average CalEnviroScreen Pollution Burden',
       'Green Ribbon District\n1 (Yes) 0 (No)',
       'Green Ribbon for Individual Schools within District\n1 (Yes) 0 (No)',
       'Total Approved Policies',
       'Published Facilities Master Plan\n1 (Yes) 0 (No)',
       'District-Wide Sustainability Staff\n1 (Yes) 0 (No)']]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,␣
 ↪random_state=10)
X_train.shape, X_test.shape
```

[ ]:

```
((56, 14), (25, 14))
```

[ ]:

```
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
import seaborn as sns
```

```
[ ]: logreg =  LogisticRegression(solver='liblinear')
     # fit the model with data

     logreg.fit(X_train,y_train)
```

```
[ ]: LogisticRegression(solver='liblinear')
```

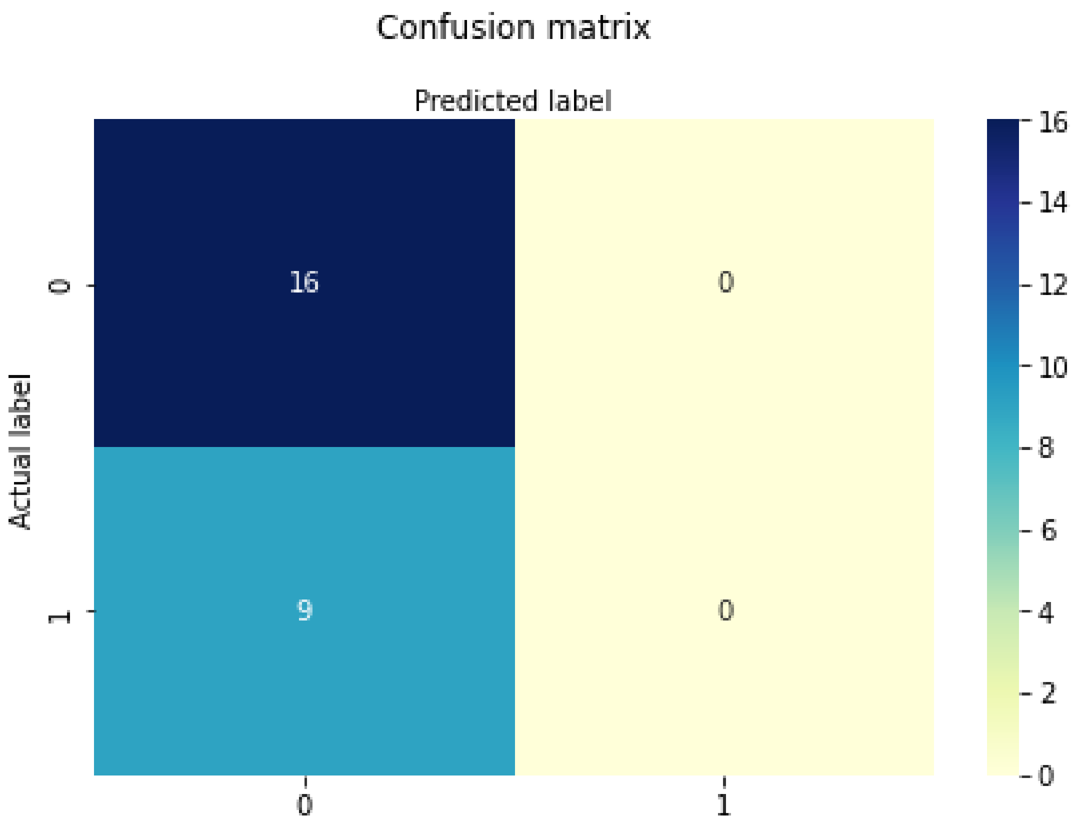
```
[ ]: y_pred=logreg.predict(X_test)
```

```
[ ]: cnf_matrix = metrics.confusion_matrix(y_test, y_pred)

     cnf_matrix
```

```
[ ]: array([[16,  0],
            [ 9,  0]])
```

```
[ ]: class_names = [0, 1]  # name  of classes

     fig, ax = plt.subplots()

     tick_marks = np.arange(len(class_names))

     plt.xticks(tick_marks, class_names)

     plt.yticks(tick_marks, class_names)

     # create heatmap

     sns.heatmap(pd.DataFrame(cnf_matrix), annot=True, cmap="YlGnBu", fmt='g')

     ax.xaxis.set_label_position("top")

     plt.tight_layout()

     plt.title('Confusion matrix', y=1.1)

     plt.ylabel('Actual label')

     plt.xlabel('Predicted label')
```

```
[ ]: Text(0.5, 257.44, 'Predicted label')
```



[ ]:
```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))

print("Precision:",metrics.precision_score(y_test, y_pred))

print("Recall:",metrics.recall_score(y_test, y_pred))
```

```
Accuracy: 0.64
Precision: 0.0
Recall: 0.0
```

```
/Users/michellelin/opt/anaconda3/lib/python3.9/site-
packages/sklearn/metrics/_classification.py:1318: UndefinedMetricWarning:
Precision is ill-defined and being set to 0.0 due to no predicted samples. Use
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

[ ]:
```
# from sklearn.model_selection import train_test_split

# # y = copy['District-Wide Sustainability Initiatives\n1 (Yes) 0 (No)']
# # X = copy.loc[:, ['Total # of Jurisdictions Per School District',
# #         'Student Enrollment \n(2021-22)',
# #         'Expense of Education per ADA \n(2020-21)',
# #         '% Unduplicated \n(2021-22)', '% FRM \n(2021-22)',
# #         '% English Learners \n(2021-22)',
# #         'Total GO Bonds and Parcel Taxes Passed \n(2000 - 2029)',
# #         'Total Amount of GO Bond Measure Funding ($)\n(2000-2029)',
# #         'Average CalEnviroScreen Pollution Burden',
# #         'Green Ribbon District\n1 (Yes) 0 (No)',
# #         'Green Ribbon for Individual Schools within District\n1 (Yes) 0␣
  ↪(No)',
# #         'Total Approved Policies',
# #         'Published Facilities Master Plan\n1 (Yes) 0 (No)',
# #         'District-Wide Sustainability Staff\n1 (Yes) 0 (No)']]
# data_train, data_test = train_test_split(copy, test_size=0.3,␣
  ↪random_state=142)
# data_train.shape, data_test.shape
```

[ ]:
```
# import statsmodels.formula.api as smf
# formula = 'Q("District-Wide Sustainability Initiatives\\n1 (Yes) 0 (No)") ~␣
  ↪Q("Total # of Jurisdictions Per School District") + Q("Student Enrollment␣
  ↪\\n(2021-22)") + Q("Expense of Education per ADA \\n(2020-21)") + Q("%␣
  ↪Unduplicated \\n(2021-22)") + Q("% FRM \\n(2021-22)") + Q("% English␣
  ↪Learners \\n(2021-22)") + Q("Total GO Bonds and Parcel Taxes Passed \\n(2000␣
  ↪- 2029)") + Q("Total Amount of GO Bond Measure Funding ($)\\n(2000-2029)") +␣
  ↪Q("Average CalEnviroScreen Pollution Burden") + Q("Green Ribbon District\\n1␣
  ↪(Yes) 0 (No)") + Q("Green Ribbon for Individual Schools within District\\n1␣
  ↪(Yes) 0 (No)") + Q("Total Approved Policies") + Q("Published Facilities␣
  ↪Master Plan\\n1 (Yes) 0 (No)") + Q("District-Wide Sustainability Staff\\n1␣
  ↪(Yes) 0 (No)")'
# logreg = smf.logit(formula=formula, data=X_train).fit()
# print(logreg.summary())
```

[ ]:
```
# # Predicting the probability of having district-wide sustainability intiatives
# y_prob = logreg.predict(X_test)

# # Predicting the label: 0 or 1?
# y_pred = pd.Series([1 if x > 1/2 else 0 for x in y_prob], index=y_prob.index)

# from sklearn.metrics import confusion_matrix
# y_test = X_test['District-Wide Sustainability Initiatives\n1 (Yes) 0 (No)']
# cm = confusion_matrix(y_test, y_pred)
# print ("Confusion Matrix : \n", cm)
```

[ ]:
```

```

[ ]:
```

```

[ ]:
```
copy
```

```
[ ]:                  District Type Grade Levels  Number of Schools\n(2021-22)  \
     0      Unified School District         K-12                         126.0
     0      Unified School District        TK-12                          68.0
     1      Unified School District         K-12                          54.0
     3      Unified School District         K-12                          23.0
     4      Unified School District         K-12                          14.0
     ..                         …          …                           …
     34  Elementary School District         K-06                           7.0
     35  Elementary School District         K-08                          13.0
     37        High School District         K-12                          31.0
     38  Elementary School District         K-08                           1.0
     39     Unified School District         K-12                          33.0

         Total # of Jurisdictions Per School District  \
     0                                            1.0
     0                                            1.0
     1                                            5.0
     3                                            1.0
     4                                            1.0
     ..                                           …
     34                                           3.0
     35                                           3.0
     37                                           4.0
     38                                           1.0
     39                                           2.0

         Student Enrollment \n(2021-22)  # of Certificated Teachers\n(2018-19)  \
     0                            55592                                  3886
     0                            39803                                  1732
     1                            30727                                  1614
     3                            15398                                   729
     4                             8967                                   442
     ..                             …                                    …
     34                            2820                                   185
     35                            6119                                   334
     37                           38026                                  1893
     38                             178                                    12
     39                           22092                                  1191

         Expense of Education per ADA \n(2020-21)  % Unduplicated \n(2021-22)  \
     0                                     29258.0                       52.22
     0                                     20943.0                       72.41
     1                                     18601.0                       63.40
     3                                     15391.0                       52.24
     4                                     15475.0                       57.54
     ..                                       …                         …
     34                                    21793.0                       20.74
```

```
35                                   26453.0                       58.95
37                                   16947.0                       61.17
38                                   23672.0                       85.96
39                                   18238.0                       58.98

    % FRM \n(2021-22)  % English Learners \n(2021-22)  \
0                50.4                           26.3
0                79.2                           24.1
1                65.3                           20.2
3                47.9                           24.9
4                55.5                           12.1
..                …                             …
34               12.2                           11.4
35               65.7                           47.4
37               50.6                           22.9
38               83.7                           53.9
39               64.2                           17.3

    Total GO Bonds and Parcel Taxes Passed \n(2000 - 2029)  \
0                                                   8
0                                                   6
1                                                   3
3                                                   3
4                                                   2
..                                                …
34                                                  1
35                                                  3
37                                                  3
38                                                  0
39                                                  2

    Total Amount of GO Bond Measure Funding ($)\n(2000-2029)  \
0                                        2.020250e+09
0                                        1.090880e+09
1                                        3.904000e+08
3                                        5.100000e+07
4                                        9.850000e+07
..                                                …
34                                       1.050000e+08
35                                       1.034000e+08
37                                       1.234000e+09
38                                       1.034000e+08
39                                       3.870000e+08

    Average CalEnviroScreen Pollution Burden  \
0                                    35.8232
0                                    51.1939
```

```
1                                    41.4561
3                                    39.2630
4                                    45.0943
..                                       …
34                                   29.6973
35                                   40.6387
37                                   39.8283
38                                   44.5441
39                                   36.2005

    Average Cal EnviroScreen Percentile  \
0                                 37.82
0                                 81.00
1                                 54.15
3                                 48.34
4                                 65.22
..                                  …
34                                18.00
35                                52.49
37                                50.00
38                                63.56
39                                39.00

    Green Ribbon District\n1 (Yes) 0 (No)  \
0                                     1.0
0                                     0.0
1                                     0.0
3                                     0.0
4                                     0.0
..                                    …
34                                    0.0
35                                    0.0
37                                    0.0
38                                    0.0
39                                    0.0

    Green Ribbon for Individual Schools within District\n1 (Yes) 0 (No)  \
0                                                 1.0
0                                                 0.0
1                                                 0.0
3                                                 0.0
4                                                 0.0
..                                                …
34                                                0.0
35                                                0.0
37                                                0.0
38                                                0.0
```

```
39                                                1.0

    Total Approved Policies  Published Facilities Master Plan\n1 (Yes) 0 (No)  \
0                         2                                                 1
0                         7                                                 0
1                         3                                                 1
3                         4                                                 0
4                         3                                                 0
..                      …                                               …
34                        6                                                 1
35                        7                                                 1
37                        6                                                 1
38                        2                                                 0
39                        6                                                 1

    District-Wide Sustainability Initiatives\n1 (Yes) 0 (No)  \
0                                                   1
0                                                   1
1                                                   1
3                                                   0
4                                                   1
..                                                …
34                                                  0
35                                                  0
37                                                  1
38                                                  0
39                                                  1

    District-Wide Sustainability Staff\n1 (Yes) 0 (No)
0                                                 1.0
0                                                 0.0
1                                                 0.0
3                                                 0.0
4                                                 1.0
..                                                …
34                                                0.0
35                                                0.0
37                                                1.0
38                                                0.0
39                                                0.0

[81 rows x 20 columns]
```

[ ]: