Analyzing Time Series Data and perform regressional analysis with Socrata and Python

In [21]:

```
import os
os.getcwd()
import numpy as np
import pandas as pd
```

Milestone 2

Q 1.1

Load your dataset and print the first 10 rows and any summary statistics that provide context about the data.

In [22]:

```
# load Chicago permits data
cps_progress_reports = pd.read_csv("/Users/markraskin/Downloads/Chicago_Publ
print(len(cps_progress_reports))
cps_progress_reports.head(10)
```

```
650
```

Out[22]:

	School_ID	Short_Name	Long_Name	School_Type	Primary_Category	Addres
0	610125	RUIZ	Irma C Ruiz Elementary School	Neighborhood	ES	2410 LEAVITT S
1	609728	ROOSEVELT HS	Theodore Roosevelt High School	Neighborhood	HS	3436 WILSON AV
2	610040	LLOYD	Henry D Lloyd Elementary School	Neighborhood	ES	2103 LAMON AV
3	609983	HEDGES	James Hedges Elementary School	Neighborhood	ES	4747 WINCHESTE AV
4	610225	WHISTLER	John Whistler Elementary School	Neighborhood	ES	11533 S AD S
5	610016	KELLOGG	Kate S Kellogg Elementary School	Neighborhood	ES	9241 LEAVITT S
6	610081	SHERIDAN	Mark Sheridan Math & Science Academy	Magnet	ES	533 W 27T S
7	610073	MITCHELL	Ellen Mitchell Elementary School	Neighborhood	ES	2233 W OHI S
8	609839	CARROLL	Carroll-Rosenwald Specialty Elementary School	Neighborhood	ES	2929 83RD S
9	400112	ACERO - IDAR	Acero Charter Schools - Jovita Idar	Charter	ES	5050 HOMAN AV

10 rows × 182 columns

Q 1.2

Prepare the data for analysis by performing the necessary cleaning, transformation, and preprocessing steps.

In [26]:

```
cps_progress_reports.info()

cps_progress_reports.replace(["", "NA", "N/A", "-"], np.nan, inplace=True)

empty_columns = cps_progress_reports.columns[cps_progress_reports.isnull().a

print("Completely empty columns:\n", empty_columns)

cps_progress_reports_cleaned = cps_progress_reports.drop(columns=empty_colum

print("Data dimensions after dropping empty columns:", cps_progress_reports_

cps_progress_reports_cleaned.info()

cols_to_drop = [col for col in cps_progress_reports.columns if cps_progress_
cps_progress_reports_cleaned = cps_progress_reports.drop(columns=cols_to_drc
cps_progress_reports_cleaned.head(10)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 650 entries, 0 to 649
Columns: 182 entries, School_ID to Growth_SAT_Math_Grade_11_School_Lbl
dtypes: float64(145), int64(3), object(34)
memory usage: 924.3+ KB
Completely empty columns:
 Index(['Growth_Reading_Grades_Tested_Pct_ES',
       'Growth_Reading_Grades_Tested_Label_ES',
       'Growth_Math_Grades_Tested_Pct_ES',
       'Growth_Math_Grades_Tested_Label_ES', 'Attainment_Reading_Pct_ES',
       'Attainment_Reading_Lbl_ES', 'Attainment_Math_Pct_ES',
       'Attainment_Math_Lbl_ES', 'School_Survey_Parent_Response_Rate_Pct',
       'School_Survey_Parent_Response_Rate_Avg_Pct',
       'NWEA_Reading_Growth_Grade_3_Pct', 'NWEA_Reading_Growth_Grade_3_Lbl',
       'NWEA_Reading_Growth_Grade_4_Pct', 'NWEA_Reading_Growth_Grade_4_Lbl',
       'NWEA_Reading_Growth_Grade_5_Pct', 'NWEA_Reading_Growth_Grade_5_Lbl',
       'NWEA_Reading_Growth_Grade_6_Pct', 'NWEA_Reading_Growth_Grade_6_Lbl',
       'NWEA_Reading_Growth_Grade_7_Pct', 'NWEA_Reading_Growth_Grade_7_Lbl',
       'NWEA_Reading_Growth_Grade_8_Pct', 'NWEA_Reading_Growth_Grade_8_Lbl',
       'NWEA_Math_Growth_Grade_3_Pct', 'NWEA_Math_Growth_Grade_3_Lbl',
       'NWEA_Math_Growth_Grade_4_Pct', 'NWEA_Math_Growth_Grade_4_Lbl',
       'NWEA_Math_Growth_Grade_5_Pct', 'NWEA_Math_Growth_Grade_5_Lbl',
       'NWEA_Math_Growth_Grade_6_Pct', 'NWEA_Math_Growth_Grade_6_Lbl',
       'NWEA_Math_Growth_Grade_7_Pct', 'NWEA_Math_Growth_Grade_7_Lbl',
       'NWEA_Math_Growth_Grade_8_Pct', 'NWEA_Math_Growth_Grade_8_Lbl',
       'NWEA_Reading_Attainment_Grade_2_Pct',
       'NWEA_Reading_Attainment_Grade_2_Lbl',
       'NWEA_Reading_Attainment_Grade_3_Pct',
       'NWEA_Reading_Attainment_Grade_3_Lbl',
       'NWEA_Reading_Attainment_Grade_4_Pct',
       'NWEA_Reading_Attainment_Grade_4_Lbl',
       'NWEA_Reading_Attainment_Grade_5_Pct',
       'NWEA_Reading_Attainment_Grade_5_Lbl',
       'NWEA_Reading_Attainment_Grade_6_Pct',
       'NWEA_Reading_Attainment_Grade_6_Lbl',
       'NWEA_Reading_Attainment_Grade_7_Pct',
       'NWEA_Reading_Attainment_Grade_7_Lbl',
       'NWEA_Reading_Attainment_Grade_8_Pct',
       'NWEA_Reading_Attainment_Grade_8_Lbl',
       'NWEA_Math_Attainment_Grade_2_Pct', 'NWEA_Math_Attainment_Grade_2_Lb
l',
       'NWEA_Math_Attainment_Grade_3_Pct', 'NWEA_Math_Attainment_Grade_3_Lb
l',
       'NWEA_Math_Attainment_Grade_4_Pct', 'NWEA_Math_Attainment_Grade_4_Lb
l',
       'NWEA_Math_Attainment_Grade_5_Pct', 'NWEA_Math_Attainment_Grade_5_Lb
l',
       'NWEA_Math_Attainment_Grade_6_Pct', 'NWEA_Math_Attainment_Grade_6_Lb
l',
       'NWEA_Math_Attainment_Grade_7_Pct', 'NWEA_Math_Attainment_Grade_7_Lb
l',
       'NWEA_Math_Attainment_Grade_8_Pct', 'NWEA_Math_Attainment_Grade_8_Lb
l',
       'School_Survey_School_Community',
       'School_Survey_Parent_Teacher_Partnership',
       'School_Survey_Quality_Of_Facilities',
       'School_Survey_Rating_Description', 'PSAT_Grade_9_Score_School_Avg',
       'PSAT_Grade_10_Score_School_Avg', 'Growth_PSAT_Grade_9_School_Pct',
       'Growth_PSAT_Grade_9_School_Lbl',
       'Growth_PSAT_Reading_Grade_10_School_Pct',
       'Growth_PSAT_Reading_Grade_10_School_Lbl',
       'Growth_SAT_Grade_11_School_Pct', 'Growth_SAT_Grade_11_School_Lbl',
       'Growth_PSAT_Math_Grade_10_School_Pct',
       'Growth_PSAT_Math_Grade_10_School_Lbl',
       'Growth_SAT_Reading_Grade_11_School_Pct',
       'Growth_SAT_Reading_Grade_11_School_Lbl',
       'Growth_SAT_Math_Grade_11_School_Pct',
       'Growth_SAT_Math_Grade_11_School_Lbl'],
      dtype='object')
Data dimensions after dropping empty columns: (650, 102)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 650 entries, 0 to 649
Columns: 102 entries, School_ID to Attainment_All_Grades_School_Lbl
dtypes: float64(65), int64(3), object(34)
memory usage: 518.1+ KB
```

Out[26]:

	School_ID	Short_Name	Long_Name	School_Type	Primary_Category	Addres
0	610125	RUIZ	Irma C Ruiz Elementary School	Neighborhood	ES	2410 LEAVITT S
1	609728	ROOSEVELT HS	Theodore Roosevelt High School	Neighborhood	HS	3436 V WILSON AV
2	610040	LLOYD	Henry D Lloyd Elementary School	Neighborhood	ES	2103 I LAMON AV
3	609983	HEDGES	James Hedges Elementary School	Neighborhood	ES	4747 WINCHESTE AV
4	610225	WHISTLER	John Whistler Elementary School	Neighborhood	ES	11533 S AD S
5	610016	KELLOGG	Kate S Kellogg Elementary School	Neighborhood	ES	9241 LEAVITT S
6	610081	SHERIDAN	Mark Sheridan Math & Science Academy	Magnet	ES	533 W 27T S
7	610073	MITCHELL	Ellen Mitchell Elementary School	Neighborhood	ES	2233 W OHI S
8	609839	CARROLL	Carroll-Rosenwald Specialty Elementary School	Neighborhood	ES	2929 V 83RD S
9	400112	ACERO - IDAR	Acero Charter Schools - Jovita Idar	Charter	ES	5050 HOMAN AV

10 rows × 68 columns

Q2.1

Submit at least two examples where you apply analytical techniques, such as regression, clustering, or machine learning algorithms to your datasets.

In [ ]:

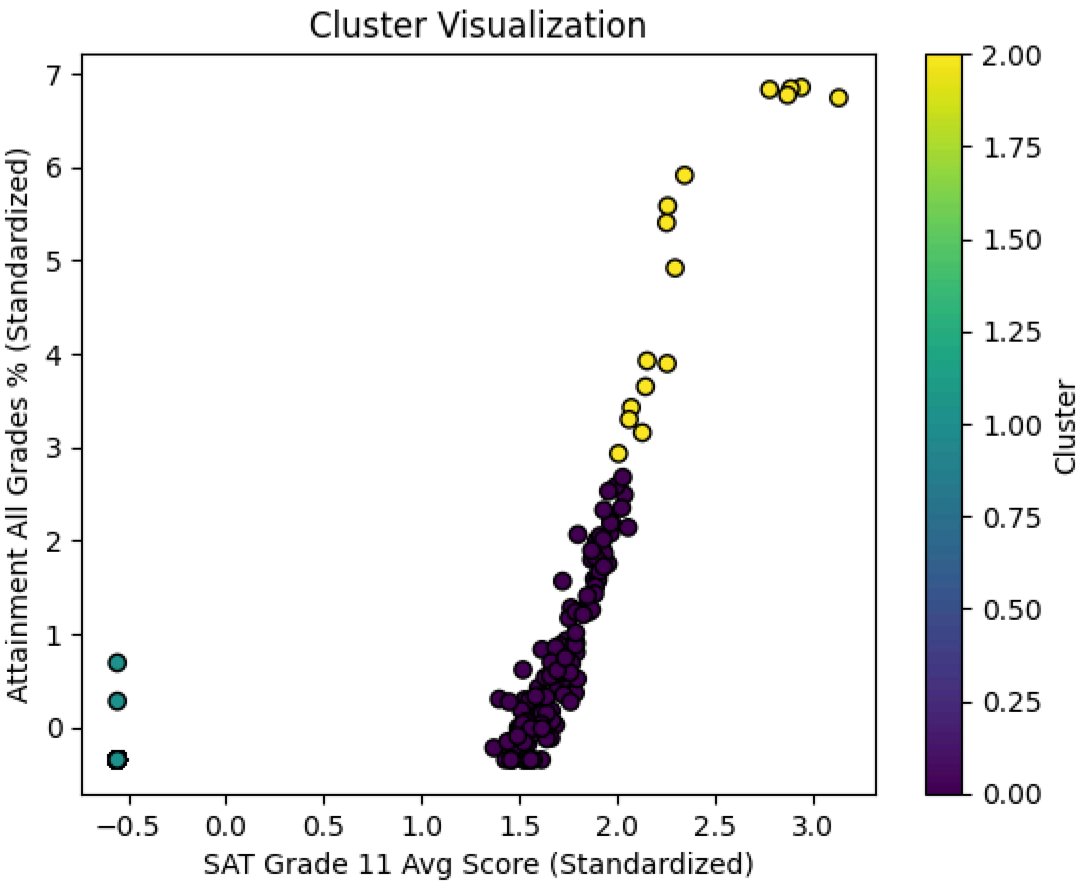
```
```

In [27]:

```
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt

features = cps_progress_reports_cleaned[['SAT_Grade_11_Score_School_Avg', 'A
cleaned_data = cps_progress_reports_cleaned.dropna(subset=['SAT_Grade_11_Sco

scaler = StandardScaler()
scaled_features = scaler.fit_transform(features.dropna())

# Elbow Method to determine optimal number of clusters
kmeans = KMeans(n_clusters=3, random_state=42)
cleaned_data['Cluster'] = kmeans.fit_predict(scaled_features)

# Display the dataset with cluster labels
print(cleaned_data[['SAT_Grade_11_Score_School_Avg', 'Attainment_All_Grades_

plt.scatter(
    scaled_features[:, 0], scaled_features[:, 1],
    c=cleaned_data['Cluster'], cmap='viridis', marker='o', edgecolor='k'
)

# Add labels and title
plt.title('Cluster Visualization')
plt.xlabel('SAT Grade 11 Avg Score (Standardized)')
plt.ylabel('Attainment All Grades % (Standardized)')
plt.colorbar(label='Cluster')
plt.show()
```

```
     SAT_Grade_11_Score_School_Avg  Attainment_All_Grades_School_Pct  Cluste
r
0                              0.0                               0.0
1
1                            850.0                              17.1
0
2                              0.0                               0.0
1
3                              0.0                               0.0
1
4                              0.0                               0.0
1
..                             ...                               ...
...
645                          949.0                              43.9
2
646                          758.0                               3.4
0
647                          848.0                              14.6
0
648                            0.0                               0.0
1
649                         1074.0                              83.8
2

[648 rows x 3 columns]
```

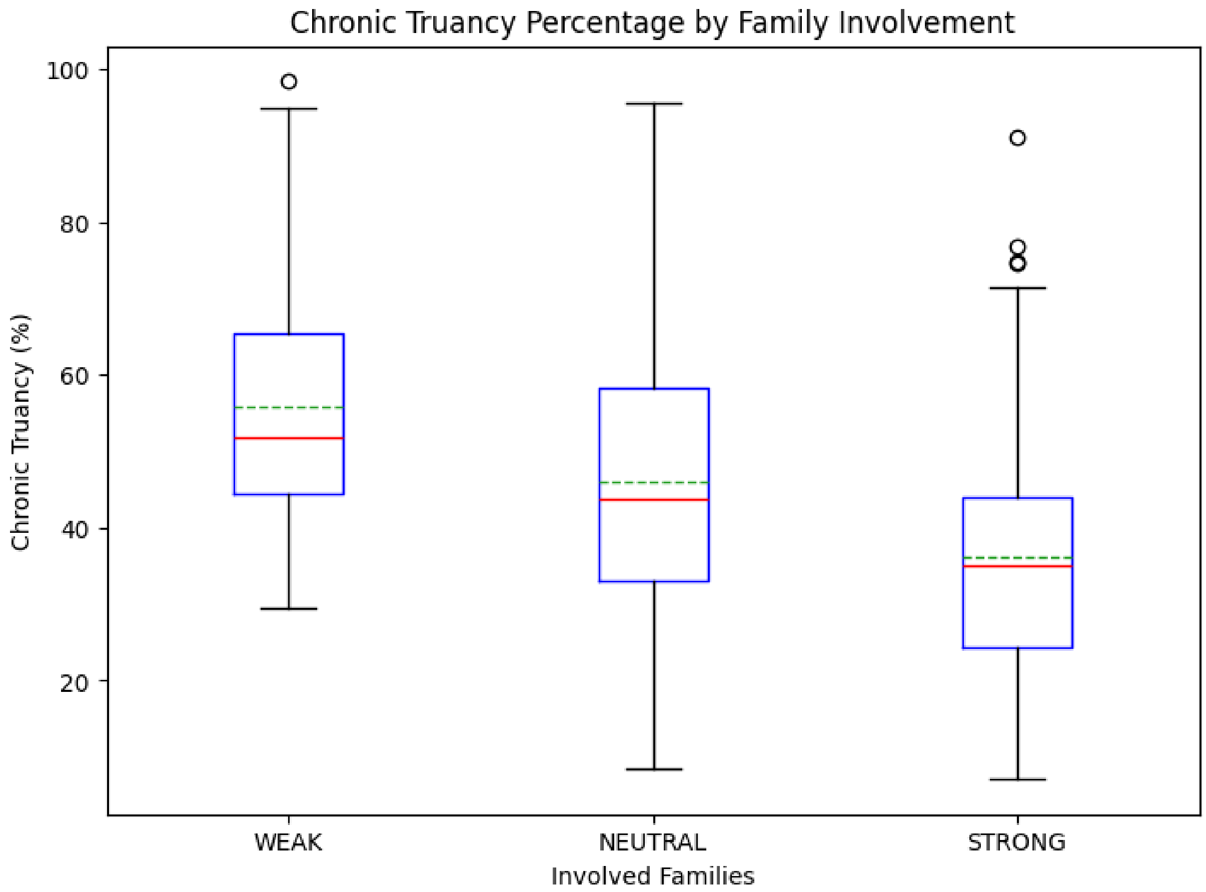
```
/var/folders/4g/y8wxgxyx3475d1ky77l34fn40000gn/T/ipykernel_49483/3882814400.
py:13: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/
stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  cleaned_data['Cluster'] = kmeans.fit_predict(scaled_features)
```



In [25]:

```
# Subset the relevant columns
subset = cleaned_data[['School_Survey_Involved_Families', 'Chronic_Truancy_P
subset = subset[subset['School_Survey_Involved_Families'] != 'NOT ENOUGH DAT

# Prepare data
categories = ['WEAK', 'NEUTRAL', 'STRONG']
subset['School_Survey_Involved_Families'] = pd.Categorical(
    subset['School_Survey_Involved_Families'], categories=categories, ordere
)

# Create the box plot
subset.boxplot(
    column='Chronic_Truancy_Pct',
    by='School_Survey_Involved_Families',
    grid=False,
    showmeans=True,
    meanline=True,
    color=dict(boxes='blue', whiskers='black', medians='red', caps='black'),
    figsize=(8, 6)
)

# Add titles and labels
plt.title('Chronic Truancy Percentage by Family Involvement')
plt.suptitle('')  # Removes automatic Matplotlib subtitle
plt.xlabel('Involved Families')
plt.ylabel('Chronic Truancy (%)')
```

```
# Show the plot
plt.show()
```



Q2.2

For each example, describe how you applied the analytical method (e.g., regression, clustering, machine learning) to your dataset. Explain any data cleaning, feature selection, or transformation processes you used. How did you ensure that the method was applied correctly and that the results were reliable?

Q2.3

Identify any patterns or insights you discovered through the analysis. Explain how these findings contribute to understanding the research challenge.

For the first visualization, I removed rows with missing values in the selected features (SAT_Grade_11_Score_School_Avg and Attainment_All_Grades_School_Pct). This ensured that clustering algorithms wouldn't face errors due to NaNs. I chose these two numerical features because they represent key indicators of academic performance and overall attainment, which are relevant for grouping schools. I applied the Elbow Method to determine the optimal number of clusters by analyzing the Within-Cluster Sum of

Squares (WCSS). This ensured that I chose an appropriate number of clusters that balanced simplicity and accuracy. Although clustering wasn't super good for identifying new trends, it showed clusters around the 1.5 Standardized SAT grade, as well as by the 3.0, showing that higher SAT and higher Attainment formed clusters.

For the second method, I applied Correlation Analysis between School_Survey_Involved_Families and Chronic_Truancy_Pct. Rows with missing values or categories labeled NOT ENOUGH DATA in the School_Survey_Involved_Families feature were removed to ensure meaningful analysis. I computed the correlation coefficient and visualized the results using box plots to confirm consistency in trends. By using both visualizations, I ensured the analysis was strong. I was able to see that schools rated STRONG for involved parents had lower chronic truancy, but schools marked WEAK had overall higher.

Q 2.4

What are some challenges you encountered during the process and how did you address them.

Some challenges that I encountered was identifying the right data visualizations tools to use, as well as identifying which algorithms to use on data that isn't highly and is mostly categorical. How I addressed these challenges was by finding techninques that would be effective at offering insight into my data specifically, even if it meant not using the most sophisticated techniques. I also faced challenges implementing the code for the techniques and visualizations, for which I turned to the internet.

Q 3

Based on the insights derived from the analysis, what potential areas for improvement did your analysis reveal and what specific actions would you suggest to address these issues? How can the findings from your analysis be used to develop strategies to improve the current situation.

Based on the insights derived from the analysis, there are a few potential areas for improvement that my analysis revealed. First off, regarding the clustering technique, it is visible that there were clusters on the lower end of the standardized SAT score, the same cluster having lower standardized attainment scores, while there was another cluster on the higher end of the SAT score range that also was on the higher end of the attainment. Although the clusters don't necessarily prove correlation, we can assume that if a schools average SAT score increases significantly to fit into another cluster, then that school's attainment will also increase, or vice versa.

Some insights that we can get from the second technique/model is that strong involved families correlate to smaller chronic truancy rates. This was to be expected, as more involved families are more presents in a childs life, ensuring that the child goes to school. However, it is good to visualize these effects in the box plot to see the discrepancies visually. This confirmation also reinforces the fact that as much as the education system is responsible for a students development, ultimatley, a student's home enviornment has a great influence into their performance as well.