Analyzing Time Series Data with Socrata and Python

In [147…
```
import os
os.getcwd()
```

Out[147…
```
'/Users/markraskin/Downloads'
```

Loading datasets into our notebook

We'll start by loading a city of Chicago school dataset into a Pandas DataFrame.

In [148…
```
import pandas as pd
import numpy as np
# load Chicago permits data
cps_progress_reports = pd.read_csv("/Users/markraskin/Downloads/Chicago_Publ
```

In the next few cells we'll do some exploration of our datasets using the `len`, `head`, and `value_counts` functions. We'll start by getting a sense of how many rows are in each of our datasets with the `len` function.

Now let's see have a peek at the first 10 rows in each of those dataset using the `head` method. You can optionally pass a parameter for the number of rows you want to print if 5 isn't enough.

In [149…
```
print(len(cps_progress_reports))
cps_progress_reports.head(10)
```

```
650
```

Out[149…

	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 27TI 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 × 182 columns

Printing out the data types of each column in cps_progress_reports.

In [150…

```
print(cps_progress_reports.dtypes)
```

```
School_ID                                  int64
Short_Name                                object
Long_Name                                 object
School_Type                               object
Primary_Category                          object
                                           ...
Growth_PSAT_Math_Grade_10_School_Lbl      float64
Growth_SAT_Reading_Grade_11_School_Pct    float64
Growth_SAT_Reading_Grade_11_School_Lbl    float64
Growth_SAT_Math_Grade_11_School_Pct       float64
Growth_SAT_Math_Grade_11_School_Lbl       float64
Length: 182, dtype: object
```

TODO 2: Deal with missing values

In [151…
```
cps_progress_reports.shape
```

Out[151… (650, 182)

In [152…
```
cps_progress_reports.info()
```

```
<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
```

In [153…
```
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_
```

```
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)
```

In [154…
```
cps_progress_reports_cleaned.columns
```

Out[154…
```
Index(['School_ID', 'Short_Name', 'Long_Name', 'School_Type',
       'Primary_Category', 'Address', 'City', 'State', 'Zip', 'Phone',
       ...
       'SAT_Grade_11_Score_School_Avg', 'SAT_Grade_11_Score_CPS_Avg',
       'Attainment_PSAT_Grade_9_School_Pct',
       'Attainment_PSAT_Grade_9_School_Lbl',
       'Attainment_PSAT_Grade_10_School_Pct',
       'Attainment_PSAT_Grade_10_School_Lbl',
       'Attainment_SAT_Grade_11_School_Pct',
       'Attainment_SAT_Grade_11_School_Lbl',
       'Attainment_All_Grades_School_Pct', 'Attainment_All_Grades_School_Lb
l'],
      dtype='object', length=102)
```

In [155…
```
columns_to_drop = [col for col in cps_progress_reports_cleaned.columns if co
cps_progress_reports_cleaned = cps_progress_reports_cleaned.drop(columns=col
cps_progress_reports_cleaned.columns
cps_progress_reports_cleaned['Student_Attendance_Avg_Pct'] = cps_progress_re
cps_avg = cps_progress_reports_cleaned['Student_Attendance_Avg_Pct']
```

In [156…
```
print(cps_progress_reports_cleaned.head(10))
```

	School_ID	Short_Name	Long_Name	\
0	610125	RUIZ	Irma C Ruiz Elementary School	
1	609728	ROOSEVELT HS	Theodore Roosevelt High School	
2	610040	LLOYD	Henry D Lloyd Elementary School	
3	609983	HEDGES	James Hedges Elementary School	
4	610225	WHISTLER	John Whistler Elementary School	
5	610016	KELLOGG	Kate S Kellogg Elementary School	
6	610081	SHERIDAN	Mark Sheridan Math & Science Academy	
7	610073	MITCHELL	Ellen Mitchell Elementary School	
8	609839	CARROLL	Carroll-Rosenwald Specialty Elementary School	
9	400112	ACERO - IDAR	Acero Charter Schools - Jovita Idar	

	School_Type	Primary_Category	Address	City	State	\
0	Neighborhood	ES	2410 S LEAVITT ST	Chicago	Illinois	
1	Neighborhood	HS	3436 W WILSON AVE	Chicago	Illinois	
2	Neighborhood	ES	2103 N LAMON AVE	Chicago	Illinois	
3	Neighborhood	ES	4747 S WINCHESTER AVE	Chicago	Illinois	
4	Neighborhood	ES	11533 S ADA ST	Chicago	Illinois	
5	Neighborhood	ES	9241 S LEAVITT ST	Chicago	Illinois	
6	Magnet	ES	533 W 27TH ST	Chicago	Illinois	
7	Neighborhood	ES	2233 W OHIO ST	Chicago	Illinois	
8	Neighborhood	ES	2929 W 83RD ST	Chicago	Illinois	
9	Charter	ES	5050 S HOMAN AVE	Chicago	Illinois	

	Zip	Phone	...	SAT_Grade_11_Score_School_Avg	\
0	60608	773-535-4825	...	0.0	
1	60625	773-534-5000	...	850.0	
2	60639	773-534-3070	...	0.0	
3	60609	773-535-7360	...	0.0	
4	60643	773-535-5560	...	0.0	
5	60643	773-535-2590	...	0.0	
6	60616	773-534-9120	...	0.0	
7	60612	773-534-7655	...	0.0	
8	60652	773-535-9414	...	0.0	
9	60632	312-455-5450	...	0.0	

	SAT_Grade_11_Score_CPS_Avg	Attainment_PSAT_Grade_9_School_Pct	\
0	939.0	0.0	
1	939.0	29.2	
2	939.0	0.0	
3	939.0	0.0	
4	939.0	0.0	
5	939.0	0.0	
6	939.0	0.0	
7	939.0	0.0	
8	939.0	0.0	
9	939.0	0.0	

	Attainment_PSAT_Grade_9_School_Lbl	Attainment_PSAT_Grade_10_School_Pct	\
0	0.0	0.0	
1	29.2	11.8	
2	0.0	0.0	
3	0.0	0.0	
4	0.0	0.0	

```
5                                    0.0                                    0.0
6                                    0.0                                    0.0
7                                    0.0                                    0.0
8                                    0.0                                    0.0
9                                    0.0                                    0.0

   Attainment_PSAT_Grade_10_School_Lbl  Attainment_SAT_Grade_11_School_Pct
\
0                                  0.0                                    0.0
1                                 11.8                                    9.9
2                                  0.0                                    0.0
3                                  0.0                                    0.0
4                                  0.0                                    0.0
5                                  0.0                                    0.0
6                                  0.0                                    0.0
7                                  0.0                                    0.0
8                                  0.0                                    0.0
9                                  0.0                                    0.0

   Attainment_SAT_Grade_11_School_Lbl  Attainment_All_Grades_School_Pct  \
0                                 0.0                               0.0
1                                 9.9                              17.1
2                                 0.0                               0.0
3                                 0.0                               0.0
4                                 0.0                               0.0
5                                 0.0                               0.0
6                                 0.0                               0.0
7                                 0.0                               0.0
8                                 0.0                               0.0
9                                 0.0                               0.0

  Attainment_All_Grades_School_Lbl
0                              0.0
1                             17.1
2                              0.0
3                              0.0
4                              0.0
5                              0.0
6                              0.0
7                              0.0
8                              0.0
9                              0.0

[10 rows x 97 columns]
```

Print descriptive statistics

Use info() to get some basic summary about the dataframe, and also describe() helps us to get some descriptive statistics about columns containing the numeric values.

In [157…

```
cps_progress_reports_cleaned.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 650 entries, 0 to 649
Data columns (total 97 columns):
 #   Column                                        Non-Null Count  Dtype  
---  ------                                        --------------  -----  
 0   School_ID                                     650 non-null    int64  
 1   Short_Name                                    650 non-null    object 
 2   Long_Name                                     650 non-null    object 
 3   School_Type                                   648 non-null    object 
 4   Primary_Category                              650 non-null    object 
 5   Address                                       650 non-null    object 
 6   City                                          650 non-null    object 
 7   State                                         650 non-null    object 
 8   Zip                                           650 non-null    int64  
 9   Phone                                         650 non-null    object 
 10  Fax                                           644 non-null    object 
 11  CPS_School_Profile                            650 non-null    object 
 12  Website                                       650 non-null    object 
 13  Progress_Report_Year                          650 non-null    int64  
 14  Blue_Ribbon_Award_Year                        24 non-null     float64
 15  Excelerate_Award_Gold_Year                    294 non-null    float64
 16  Spot_Light_Award_Year                         22 non-null     float64
 17  Improvement_Award_Year                        62 non-null     float64
 18  Excellence_Award_Year                         30 non-null     float64
 19  Student_Growth_Rating                         611 non-null    object 
 20  Student_Attainment_Rating                     648 non-null    object 
 21  Culture_Climate_Rating                        648 non-null    object 
 22  School_Survey_Student_Response_Rate_Pct       628 non-null    float64
 23  School_Survey_Student_Response_Rate_Avg_Pct   648 non-null    float64
 24  School_Survey_Teacher_Response_Rate_Pct       643 non-null    float64
 25  School_Survey_Teacher_Response_Rate_Avg_Pct   648 non-null    float64
 26  Healthy_School_Certification                  650 non-null    object 
 27  Creative_School_Certification                 646 non-null    object 
 28  School_Survey_Involved_Families               648 non-null    object 
 29  School_Survey_Supportive_Environment          648 non-null    object 
 30  School_Survey_Ambitious_Instruction           648 non-null    object 
 31  School_Survey_Effective_Leaders               648 non-null    object 
 32  School_Survey_Collaborative_Teachers          648 non-null    object 
 33  School_Survey_Safety                          648 non-null    object 
 34  Suspensions_Per_100_Students_Year_1_Pct       29 non-null     float64
 35  Suspensions_Per_100_Students_Year_2_Pct       380 non-null    float64
 36  Suspensions_Per_100_Students_Avg_Pct          648 non-null    float64
 37  Misconducts_To_Suspensions_Year_1_Pct         29 non-null     float64
 38  Misconducts_To_Suspensions_Year_2_Pct         380 non-null    float64
 39  Misconducts_To_Suspensions_Avg_Pct            648 non-null    float64
 40  Average_Length_Suspension_Year_1_Pct          29 non-null     object 
 41  Average_Length_Suspension_Year_2_Pct          380 non-null    object 
 42  Average_Length_Suspension_Avg_Pct             648 non-null    object 
 43  Behavior_Discipline_Year_1                    648 non-null    float64
 44  Behavior_Discipline_Year_2                    648 non-null    float64
 45  Student_Attendance_Year_1_Pct                 640 non-null    float64
 46  Student_Attendance_Year_2_Pct                 641 non-null    float64
 47  Student_Attendance_Avg_Pct                    641 non-null    float64
 48  Teacher_Attendance_Year_1_Pct                 511 non-null    float64
 49  Teacher_Attendance_Year_2_Pct                 512 non-null    float64
 50  Teacher_Attendance_Avg_Pct                    648 non-null    float64
```

```
 51  One_Year_Dropout_Rate_Year_1_Pct            161 non-null    float64
 52  One_Year_Dropout_Rate_Year_2_Pct            164 non-null    float64
 53  One_Year_Dropout_Rate_Avg_Pct               168 non-null    float64
 54  Other_Metrics_Year_1                        648 non-null    float64
 55  Other_Metrics_Year_2                        648 non-null    float64
 56  Freshmen_On_Track_School_Pct_Year_2         130 non-null    float64
 57  Freshmen_On_Track_CPS_Pct_Year_2            130 non-null    float64
 58  Freshmen_On_Track_School_Pct_Year_1         129 non-null    float64
 59  Freshmen_On_Track_CPS_Pct_Year_1            131 non-null    float64
 60  Graduation_4_Year_School_Pct_Year_2         138 non-null    float64
 61  Graduation_4_Year_CPS_Pct_Year_2            140 non-null    float64
 62  Graduation_4_Year_School_Pct_Year_1         139 non-null    float64
 63  Graduation_4_Year_CPS_Pct_Year_1            143 non-null    float64
 64  Graduation_5_Year_School_Pct_Year_2         141 non-null    float64
 65  Graduation_5_Year_CPS_Pct_Year_2            143 non-null    float64
 66  Graduation_5_Year_School_Pct_Year_1         138 non-null    float64
 67  Graduation_5_Year_CPS_Pct_Year_1            139 non-null    float64
 68  College_Enrollment_School_Pct_Year_2        161 non-null    float64
 69  College_Enrollment_CPS_Pct_Year_2           163 non-null    float64
 70  College_Enrollment_School_Pct_Year_1        161 non-null    float64
 71  College_Enrollment_CPS_Pct_Year_1           163 non-null    float64
 72  College_Persistence_School_Pct_Year_2       135 non-null    float64
 73  College_Persistence_CPS_Pct_Year_2          135 non-null    float64
 74  College_Persistence_School_Pct_Year_1       141 non-null    float64
 75  College_Persistence_CPS_Pct_Year_1          141 non-null    float64
 76  Progress_Toward_Graduation_Year_1           648 non-null    float64
 77  Progress_Toward_Graduation_Year_2           648 non-null    float64
 78  State_School_Report_Card_URL                646 non-null    object
 79  Mobility_Rate_Pct                           595 non-null    float64
 80  Chronic_Truancy_Pct                         616 non-null    float64
 81  Empty_Progress_Report_Message               2 non-null      object
 82  Supportive_School_Award                     648 non-null    object
 83  Supportive_School_Award_Desc                648 non-null    object
 84  Parent_Survey_Results_Year                  648 non-null    float64
 85  School_Latitude                             650 non-null    float64
 86  School_Longitude                            650 non-null    float64
 87  SAT_Grade_11_Score_School_Avg               648 non-null    float64
 88  SAT_Grade_11_Score_CPS_Avg                  648 non-null    float64
 89  Attainment_PSAT_Grade_9_School_Pct          648 non-null    float64
 90  Attainment_PSAT_Grade_9_School_Lbl          648 non-null    float64
 91  Attainment_PSAT_Grade_10_School_Pct         648 non-null    float64
 92  Attainment_PSAT_Grade_10_School_Lbl         648 non-null    float64
 93  Attainment_SAT_Grade_11_School_Pct          648 non-null    float64
 94  Attainment_SAT_Grade_11_School_Lbl          648 non-null    float64
 95  Attainment_All_Grades_School_Pct            648 non-null    float64
 96  Attainment_All_Grades_School_Lbl            648 non-null    float64
dtypes: float64(65), int64(3), object(29)
memory usage: 492.7+ KB
```

Milestone Part 1

Follow a similar procedure and use your chosen dataset from the final project to answer the following questions. Your result should be similar to the results shown on the code boxes.

Q1

TODO 1:

1. Load the dataset that you chose for your final project and save it as a data frame.
2. Print the length of your data frame and the first 10 rows to understand the structure of your data.
3. Print the data types of each column to understand what kind of data is stored.

In [158…
```
## TODO:Q1
import pandas as pd
import numpy as np
# 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)
cps_progress_reports.shape
```

```
650
```

Out[158… (650, 182)

Now lets clean the data by removing empty columns and removing unecessary string description columns.

Q2

You can learn other methods to deal with missing values here: https://www.analyticsvidhya.com/blog/2021/05/dealing-with-missing-values-in-python-a-complete-guide/

Answer the following questions:

2.1 Handle missing values. What variables have missing values? What types/forms of missing values are they? (e.g blank, NA, N/A, -, etc.). After dealing with missing values, show the dimensions of the data.

2.2 Please briefly describe how you deal will with these missing values and justify why you chose these methods (Hint: common imputation methods include impute by mean/median/mode, keep and ignore the NAs, drop the observations with NAs).

2.3 Identify and remove any duplicate records in the dataset. Then, eliminate any invalid or extreme outliers or values and explain how you identified and handled theme. Provide the total number of rows before and after these changes and print the first 10 rows of the cleaned dataset.

2.4 Generate summary statistics for your cleaned data (e.g., mean, median, standard deviation, sum, etc.). Compare these statistics to the ones from the raw dataset, and discuss how the data cleaning process affected the results.

2.5 What were the major data cleaning tasks you performed? Provide statistics that show how the data was cleaned (e.g., missing values handled, duplicates removed, outliers dealt with). Summarize your findings from the cleaned data and highlight any notable changes or insights you gained from the data cleaning process.

In [159…

```
## TODO: Q2 2.1
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()
```

```
<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
```

2.2: The data was cleaned by replacing missing values with NaN, dropping empty columns, dropping columns that contained redundant descriptions, as well as creating a new column which has the average attendance percentage of students. I chose these methods to decrease the overall size of the dataset, as well as get rid of unecessary data.

In [160…

```
columns_to_drop = [col for col in cps_progress_reports_cleaned.columns if cc
cps_progress_reports_cleaned = cps_progress_reports_cleaned.drop(columns=col
cps_progress_reports_cleaned.columns
cps_progress_reports_cleaned.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 650 entries, 0 to 649
Data columns (total 97 columns):
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   School_ID                                     650 non-null    int64
 1   Short_Name                                    650 non-null    object
 2   Long_Name                                     650 non-null    object
 3   School_Type                                   648 non-null    object
 4   Primary_Category                              650 non-null    object
 5   Address                                       650 non-null    object
 6   City                                          650 non-null    object
 7   State                                         650 non-null    object
 8   Zip                                           650 non-null    int64
 9   Phone                                         650 non-null    object
 10  Fax                                           644 non-null    object
 11  CPS_School_Profile                            650 non-null    object
 12  Website                                       650 non-null    object
 13  Progress_Report_Year                          650 non-null    int64
 14  Blue_Ribbon_Award_Year                        24 non-null     float64
 15  Excelerate_Award_Gold_Year                    294 non-null    float64
 16  Spot_Light_Award_Year                         22 non-null     float64
 17  Improvement_Award_Year                        62 non-null     float64
 18  Excellence_Award_Year                         30 non-null     float64
 19  Student_Growth_Rating                         611 non-null    object
 20  Student_Attainment_Rating                     648 non-null    object
 21  Culture_Climate_Rating                        648 non-null    object
 22  School_Survey_Student_Response_Rate_Pct       628 non-null    float64
 23  School_Survey_Student_Response_Rate_Avg_Pct   648 non-null    float64
 24  School_Survey_Teacher_Response_Rate_Pct       643 non-null    float64
 25  School_Survey_Teacher_Response_Rate_Avg_Pct   648 non-null    float64
 26  Healthy_School_Certification                  650 non-null    object
 27  Creative_School_Certification                 646 non-null    object
 28  School_Survey_Involved_Families               648 non-null    object
 29  School_Survey_Supportive_Environment          648 non-null    object
 30  School_Survey_Ambitious_Instruction           648 non-null    object
 31  School_Survey_Effective_Leaders               648 non-null    object
 32  School_Survey_Collaborative_Teachers          648 non-null    object
 33  School_Survey_Safety                          648 non-null    object
 34  Suspensions_Per_100_Students_Year_1_Pct       29 non-null     float64
 35  Suspensions_Per_100_Students_Year_2_Pct       380 non-null    float64
 36  Suspensions_Per_100_Students_Avg_Pct          648 non-null    float64
 37  Misconducts_To_Suspensions_Year_1_Pct         29 non-null     float64
 38  Misconducts_To_Suspensions_Year_2_Pct         380 non-null    float64
 39  Misconducts_To_Suspensions_Avg_Pct            648 non-null    float64
 40  Average_Length_Suspension_Year_1_Pct          29 non-null     object
 41  Average_Length_Suspension_Year_2_Pct          380 non-null    object
 42  Average_Length_Suspension_Avg_Pct             648 non-null    object
 43  Behavior_Discipline_Year_1                    648 non-null    float64
 44  Behavior_Discipline_Year_2                    648 non-null    float64
 45  Student_Attendance_Year_1_Pct                 640 non-null    float64
 46  Student_Attendance_Year_2_Pct                 641 non-null    float64
 47  Student_Attendance_Avg_Pct                    641 non-null    float64
 48  Teacher_Attendance_Year_1_Pct                 511 non-null    float64
 49  Teacher_Attendance_Year_2_Pct                 512 non-null    float64
 50  Teacher_Attendance_Avg_Pct                    648 non-null    float64
```

```
 51  One_Year_Dropout_Rate_Year_1_Pct            161 non-null    float64
 52  One_Year_Dropout_Rate_Year_2_Pct            164 non-null    float64
 53  One_Year_Dropout_Rate_Avg_Pct               168 non-null    float64
 54  Other_Metrics_Year_1                        648 non-null    float64
 55  Other_Metrics_Year_2                        648 non-null    float64
 56  Freshmen_On_Track_School_Pct_Year_2         130 non-null    float64
 57  Freshmen_On_Track_CPS_Pct_Year_2            130 non-null    float64
 58  Freshmen_On_Track_School_Pct_Year_1         129 non-null    float64
 59  Freshmen_On_Track_CPS_Pct_Year_1            131 non-null    float64
 60  Graduation_4_Year_School_Pct_Year_2         138 non-null    float64
 61  Graduation_4_Year_CPS_Pct_Year_2            140 non-null    float64
 62  Graduation_4_Year_School_Pct_Year_1         139 non-null    float64
 63  Graduation_4_Year_CPS_Pct_Year_1            143 non-null    float64
 64  Graduation_5_Year_School_Pct_Year_2         141 non-null    float64
 65  Graduation_5_Year_CPS_Pct_Year_2            143 non-null    float64
 66  Graduation_5_Year_School_Pct_Year_1         138 non-null    float64
 67  Graduation_5_Year_CPS_Pct_Year_1            139 non-null    float64
 68  College_Enrollment_School_Pct_Year_2        161 non-null    float64
 69  College_Enrollment_CPS_Pct_Year_2           163 non-null    float64
 70  College_Enrollment_School_Pct_Year_1        161 non-null    float64
 71  College_Enrollment_CPS_Pct_Year_1           163 non-null    float64
 72  College_Persistence_School_Pct_Year_2       135 non-null    float64
 73  College_Persistence_CPS_Pct_Year_2          135 non-null    float64
 74  College_Persistence_School_Pct_Year_1       141 non-null    float64
 75  College_Persistence_CPS_Pct_Year_1          141 non-null    float64
 76  Progress_Toward_Graduation_Year_1           648 non-null    float64
 77  Progress_Toward_Graduation_Year_2           648 non-null    float64
 78  State_School_Report_Card_URL                646 non-null    object
 79  Mobility_Rate_Pct                           595 non-null    float64
 80  Chronic_Truancy_Pct                         616 non-null    float64
 81  Empty_Progress_Report_Message               2 non-null      object
 82  Supportive_School_Award                     648 non-null    object
 83  Supportive_School_Award_Desc                648 non-null    object
 84  Parent_Survey_Results_Year                  648 non-null    float64
 85  School_Latitude                             650 non-null    float64
 86  School_Longitude                            650 non-null    float64
 87  SAT_Grade_11_Score_School_Avg               648 non-null    float64
 88  SAT_Grade_11_Score_CPS_Avg                  648 non-null    float64
 89  Attainment_PSAT_Grade_9_School_Pct          648 non-null    float64
 90  Attainment_PSAT_Grade_9_School_Lbl          648 non-null    float64
 91  Attainment_PSAT_Grade_10_School_Pct         648 non-null    float64
 92  Attainment_PSAT_Grade_10_School_Lbl         648 non-null    float64
 93  Attainment_SAT_Grade_11_School_Pct          648 non-null    float64
 94  Attainment_SAT_Grade_11_School_Lbl          648 non-null    float64
 95  Attainment_All_Grades_School_Pct            648 non-null    float64
 96  Attainment_All_Grades_School_Lbl            648 non-null    float64
dtypes: float64(65), int64(3), object(29)
memory usage: 492.7+ KB
```

2.3 Deleting duplicates

In [161…

```
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
```

Deleted columns with less than 2 unique values. Cant delete outliers, each row represents a CPS school. This greatly filtered the data

In [162…

```
cps_progress_reports_cleaned.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 650 entries, 0 to 649
Data columns (total 68 columns):
 #   Column                                        Non-Null Count  Dtype  
---  ------                                        --------------  -----  
 0   School_ID                                     650 non-null    int64  
 1   Short_Name                                    650 non-null    object 
 2   Long_Name                                     650 non-null    object 
 3   School_Type                                   648 non-null    object 
 4   Primary_Category                              650 non-null    object 
 5   Address                                       650 non-null    object 
 6   State                                         650 non-null    object 
 7   Zip                                           650 non-null    int64  
 8   Phone                                         650 non-null    object 
 9   Fax                                           644 non-null    object 
 10  CPS_School_Profile                            650 non-null    object 
 11  Website                                       650 non-null    object 
 12  Blue_Ribbon_Award_Year                        24 non-null     float64
 13  Spot_Light_Award_Year                         22 non-null     float64
 14  Improvement_Award_Year                        62 non-null     float64
 15  Excellence_Award_Year                         30 non-null     float64
 16  Student_Growth_Description                    648 non-null    object 
 17  Student_Attainment_Rating                     648 non-null    object 
 18  Student_Attainment_Description                648 non-null    object 
 19  Culture_Climate_Rating                        648 non-null    object 
 20  School_Survey_Student_Response_Rate_Pct       628 non-null    float64
 21  School_Survey_Teacher_Response_Rate_Pct       643 non-null    float64
 22  Creative_School_Certification                 646 non-null    object 
 23  Creative_School_Certification_Description     648 non-null    object 
 24  School_Survey_Involved_Families               648 non-null    object 
 25  School_Survey_Supportive_Environment          648 non-null    object 
 26  School_Survey_Ambitious_Instruction           648 non-null    object 
 27  School_Survey_Effective_Leaders               648 non-null    object 
 28  School_Survey_Collaborative_Teachers          648 non-null    object 
 29  School_Survey_Safety                          648 non-null    object 
 30  Suspensions_Per_100_Students_Year_1_Pct       29 non-null     float64
 31  Suspensions_Per_100_Students_Year_2_Pct       380 non-null    float64
 32  Misconducts_To_Suspensions_Year_1_Pct         29 non-null     float64
 33  Misconducts_To_Suspensions_Year_2_Pct         380 non-null    float64
 34  Average_Length_Suspension_Year_1_Pct          29 non-null     object 
 35  Average_Length_Suspension_Year_2_Pct          380 non-null    object 
 36  Student_Attendance_Year_1_Pct                 640 non-null    float64
 37  Student_Attendance_Year_2_Pct                 641 non-null    float64
 38  Teacher_Attendance_Year_1_Pct                 511 non-null    float64
 39  Teacher_Attendance_Year_2_Pct                 512 non-null    float64
 40  One_Year_Dropout_Rate_Year_1_Pct              161 non-null    float64
 41  One_Year_Dropout_Rate_Year_2_Pct              164 non-null    float64
 42  Freshmen_On_Track_School_Pct_Year_2           130 non-null    float64
 43  Freshmen_On_Track_School_Pct_Year_1           129 non-null    float64
 44  Graduation_4_Year_School_Pct_Year_2           138 non-null    float64
 45  Graduation_4_Year_School_Pct_Year_1           139 non-null    float64
 46  Graduation_5_Year_School_Pct_Year_2           141 non-null    float64
 47  Graduation_5_Year_School_Pct_Year_1           138 non-null    float64
 48  College_Enrollment_School_Pct_Year_2          161 non-null    float64
 49  College_Enrollment_School_Pct_Year_1          161 non-null    float64
 50  College_Persistence_School_Pct_Year_2         135 non-null    float64
```

```
 51  College_Persistence_School_Pct_Year_1      141 non-null    float64
 52  State_School_Report_Card_URL               646 non-null    object
 53  Mobility_Rate_Pct                          595 non-null    float64
 54  Chronic_Truancy_Pct                        616 non-null    float64
 55  Supportive_School_Award                    648 non-null    object
 56  Supportive_School_Award_Desc               648 non-null    object
 57  School_Latitude                            650 non-null    float64
 58  School_Longitude                           650 non-null    float64
 59  SAT_Grade_11_Score_School_Avg              648 non-null    float64
 60  Attainment_PSAT_Grade_9_School_Pct         648 non-null    float64
 61  Attainment_PSAT_Grade_9_School_Lbl         648 non-null    float64
 62  Attainment_PSAT_Grade_10_School_Pct        648 non-null    float64
 63  Attainment_PSAT_Grade_10_School_Lbl        648 non-null    float64
 64  Attainment_SAT_Grade_11_School_Pct         648 non-null    float64
 65  Attainment_SAT_Grade_11_School_Lbl         648 non-null    float64
 66  Attainment_All_Grades_School_Pct           648 non-null    float64
 67  Attainment_All_Grades_School_Lbl           648 non-null    float64
dtypes: float64(39), int64(2), object(27)
memory usage: 345.4+ KB
```

2.4 Summary statistics

In [163…
```
cps_progress_reports_cleaned['Student_Attendance_Avg_Pct'] = cps_progress_re
print(cps_progress_reports_cleaned['Student_Attendance_Avg_Pct'])
```

```
0      91.05
1      81.00
2      91.60
3      89.50
4      87.75
       ...
645    85.70
646    45.40
647    78.00
648    88.40
649    88.30
Name: Student_Attendance_Avg_Pct, Length: 650, dtype: float64
```

Created a column of the avg percentage of student attendance for each school across grade level.

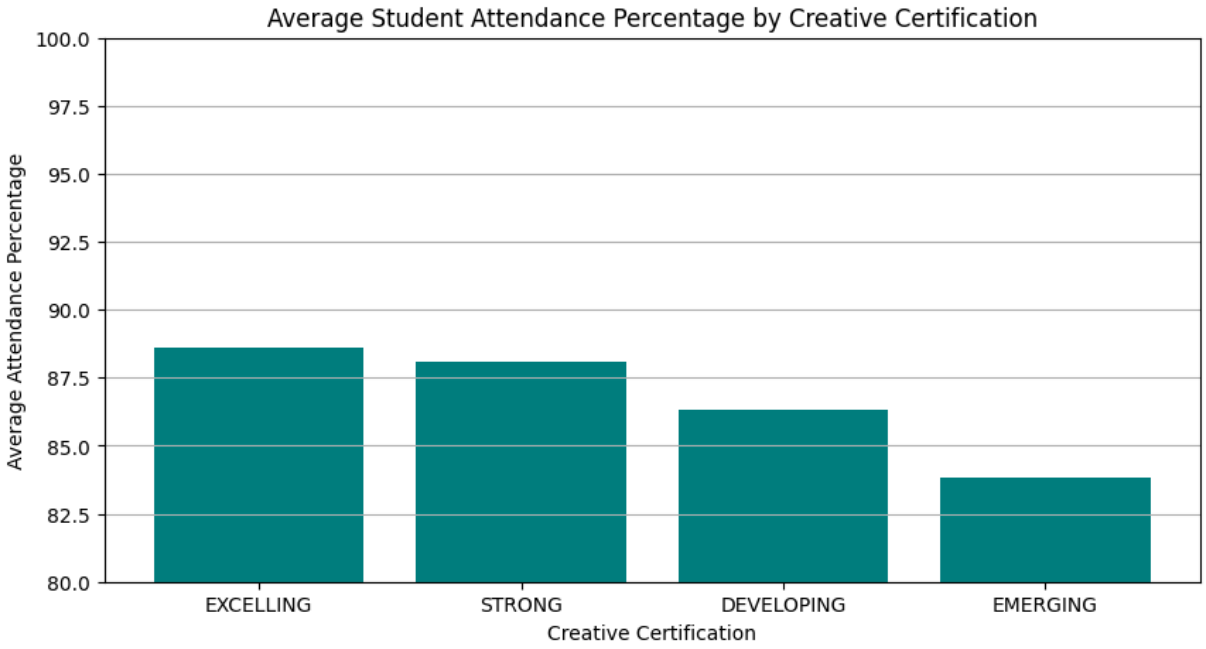
Q3

Using your own dataset, create two different visualizations to analyze the data, using Python and any plotting library of your choice (e.g., Matplotlib, Seaborn, Plotly). For each dataset, spend some time observing the visualizations, and identify any significant patterns or trends you found. Describe the insights or trends revealed, and discuss how these findings can help address the research question you identified.

Analyzing creative certification strength in the arts vs attendance

In [164…

```
import datetime
import matplotlib.pyplot as plt



cps_progress_reports_cleaned['Creative_School_Certification'] = cps_progress

cps_progress_reports_cleaned.dropna(subset=['Student_Attendance_Avg_Pct'], i
category_order = ['EXCELLING', 'STRONG', 'DEVELOPING', 'EMERGING']

average_attendance = (cps_progress_reports_cleaned.groupby('Creative_School_


plt.figure(figsize=(10, 5))
plt.bar(average_attendance['Creative_School_Certification'], average_attenda
plt.title('Average Student Attendance Percentage by Creative Certification')
plt.xlabel('Creative Certification')
plt.ylabel('Average Attendance Percentage')
plt.ylim(80, 100)
plt.grid(axis='y')
plt.show()
```

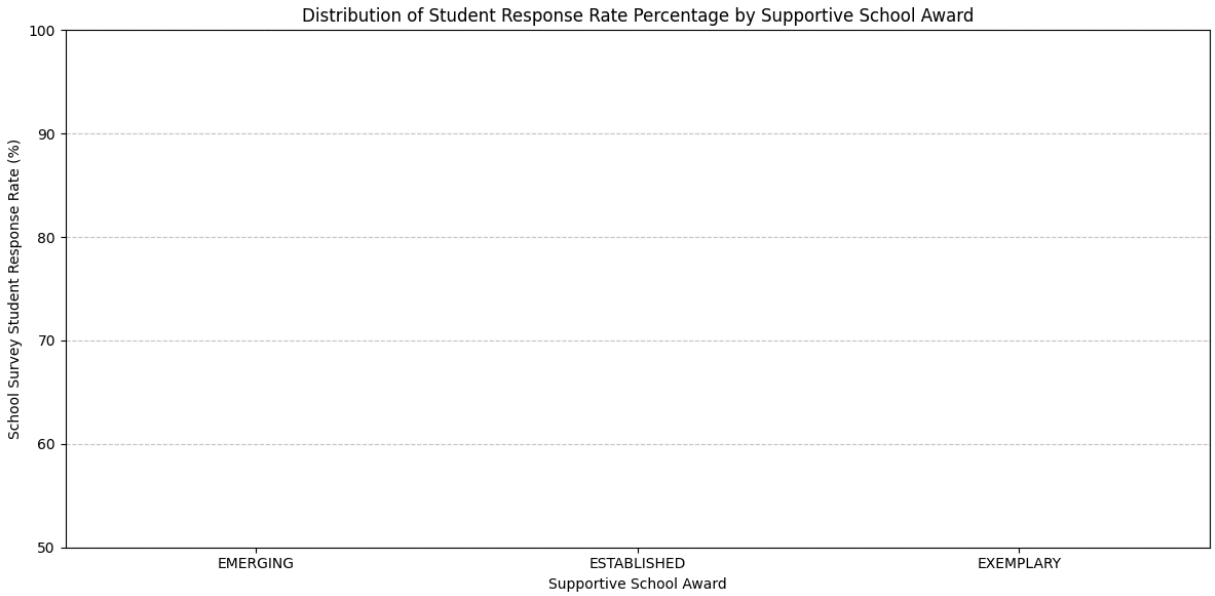


The bar chart shows the average student attendance percentages for different creative school certifications: EXCELLING, STRONG, DEVELOPING, and EMERGING .Each category displays a distinct average attendance rate, with EXCELLING schools showing the highest average attendance and DEVELOPING schools displaying the lowest. This trend suggests a positive correlation between the type of creative certification and student attendance. Schools recognized for their excellence in creative education might implement practices that lead to higher student engagement.

Analyzing student survey response rate percentage vs Supportive School Award standing

In [171…

```
school_data = cps_progress_reports_cleaned[cps_progress_reports_cleaned['Sup


plt.figure(figsize=(12, 6))
plt.boxplot(
    [school_data[school_data['Supportive_School_Award'] == award]['School_Su
     for award in ['EMERGING', 'ESTABLISHED', 'EXEMPLARY']],
    labels=['EMERGING', 'ESTABLISHED', 'EXEMPLARY'],
    patch_artist=True
)

plt.title('Distribution of Student Response Rate Percentage by Supportive Sc
plt.xlabel('Supportive School Award')
plt.ylabel('School Survey Student Response Rate (%)')
plt.ylim(50, 100)
plt.grid(axis='y', linestyle='--', alpha=0.7)

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



The box plot illustrates the distribution of student response rates for schools categorized as EMERGING, ESTABLISHED, and EXEMPLARY. EXEMPLARY schools not only have the highest median response rate but also a narrower interquartile range, suggesting that these schools have more consistent student engagement in surveys. EMERGING and ESTABLISHED schools show more variability in their response rates, indicating that while some may perform well, others in these categories may struggle

with student engagement. The findings indicate that improving the conditions that lead to higher supportive school awards could enhance student engagement.