Skills Assessment - Data Analyst -Implimentation Office

Cleveland Metropolitan School District

Pradnya Patil

In [1]:

```
# Importing the Python libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
import plotly.graph_objects as go
from plotly.subplots import make_subplots
```

Cleaning and Merging Enrollment data

Enrollment 2015-2016 Data Cleaning & Preprocessing

In [2]:

```
#Importing data enrollment data 2015-2016
df_BR15_16 = pd.read_excel(r'C:/Users/patil/Desktop/CMDS Skills Assessment/Data/BUILDING_RATING_2015_2016
df_BR15_16.head()
```

Out[2]:

	District IRN	District Name	Building IRN	Building Name	County	Region	Address	City, State, Zip	Phone	Principal	...	4 Year Grad Rate 2015	Letter Grade of 5 Year Grad Rate 2014	Y G R 2
0	138	Pathway School of Discovery	138	Pathway School of Discovery	Montgomery	Region 10	173 Avondale Dr	Dayton, OH, 45404-2123	(937) 235-5498	Keith B. Colbert	...	NC	NR	
1	139	Alliance Academy of Cincinnati	139	Alliance Academy of Cincinnati	Hamilton	Region 13	1712 Duck Creek Rd	Cincinnati, OH, 45207-1644	(513) 751-5555	Elizabeth L. King	...	NC	NR	
2	222	Wildwood Environmental Academy	222	Wildwood Environmental Academy	Lucas	Region 1	1546 Dartford Rd	Maumee, OH, 43537-1374	(419) 868-9885	Elizabeth A. Lewin	...	NC	NR	
3	236	Ohio Connections Academy, Inc	236	Ohio Connections Academy, Inc	Cuyahoga	Region 3	3740 Euclid Ave Ste 101	Cleveland, OH, 44115-2229	(513) 234-4900	NaN	...	71.7	F	6
4	296	Summit Academy Community School-Columbus	296	Summit Academy Community School-Columbus	Franklin	Region 11	2521 Fairwood Ave Ste 100	Columbus, OH, 43207-2712	(614) 237-5497	Cheryl L. Elliott	...	NC	NR	

5 rows × 44 columns

In [3]:

```
#find out the shape of data
df_BR15_16.shape
```

Out[3]:

```
(3387, 44)
```

In [4]:

```
#find out the column names
df_BR15_16.columns
```

Out[4]:

```
Index(['District IRN', 'District Name', 'Building IRN', 'Building Name',
       'County', 'Region', 'Address', 'City, State, Zip', 'Phone', 'Principal',
       'Enrollment 2015-2016', 'Letter Grade of Achievement Component',
       'Letter Grade of Percent Standards',
       'Gifted Indicator Met/Not Met Status', 'Percent of Standards Met',
       'Letter Grade of Performance Index', 'Performance Index Percent',
       'Letter Grade of AMO', 'AMO Points', 'Letter Grade of K3 Literacy',
       'K3 Literacy Percent', 'Letter Grade of Progress Component',
       'Letter Grade of Overall Value Added', 'Overall Value Added Gain Index',
       'Letter Grade of Gifted Value Added', 'Gifted Value Added Gain Index',
       'Letter Grade of Students with Disabilities Value Added',
       'Students with Disabilities Value Added Gain Index',
       'Letter Grade of Lowest 20% Value Added',
       'Lowest 20% Value Added Gain Index',
       'Letter Grade of High Mobility Value Added',
       'High Mobility Value Added Gain Index',
       'Letter Grade of Grad Rate Component',
       'Letter Grade of 4Year Grad Rate 2015', '4 Year Grad Rate 2015',
       'Letter Grade of 5 Year Grad Rate 2014', '5 Year Grad Rate 2014',
       'Letter Grade of Prepared for Success Component',
       'Percent of Prepared for Success Component',
       'Attendance Rate 2015-2016', 'Attendance Rate 2014-2015',
       'Attendance Rate 2013-2014', 'Chronic Absenteeism Percent 2015-2016',
       'Watermark'],
      dtype='object')
```

In [5]:

```
#Adding School Year Column
df_BR15_16['School_Year'] = '2015-2016'
```

In [6]:

```
#Rename enrollment column
df_BR15_16.rename(columns={'Enrollment 2015-2016': 'Enrollment'}, inplace=True)
```

In [7]:

```
# filtering the data for cleaveland district
df_BR15_16 = df_BR15_16.loc[df_BR15_16['District IRN'] == 43786]
df_BR15_16.head()
```

Out[7]:

	District IRN	District Name	Building IRN	Building Name	County	Region	Address	City, State, Zip	Phone	Principal	...	Letter Grade of 5 Year Grad Rate 2014	5 Year Grad Rate 2014	L Grac Prep Succ Compo
447	43786	Cleveland Municipal City	224	Adlai Stevenson School	Cuyahoga	Region 3	18300 Woda Avenue	Cleveland, OH, 44122-6441	(216) 482-2950	Christopher T. Wyland	...	NR	NC	
448	43786	Cleveland Municipal City	318	Menlo Park Academy	Cuyahoga	Region 3	14440 Triskett Rd	Cleveland, OH, 44111-2263	(440) 925-6365	NaN	...	NR	NC	
449	43786	Cleveland Municipal City	489	Almira	Cuyahoga	Region 3	3375 W 99th St	Cleveland, OH, 44102-4642	(216) 838-6150	Laverne Hooks	...	NR	NC	
450	43786	Cleveland Municipal City	729	Andrew J Rickoff	Cuyahoga	Region 3	3500 E 147th St	Cleveland, OH, 44120-4834	(216) 838-4150	Gloriane R. Smith	...	NR	NC	
451	43786	Cleveland Municipal City	828	Anton Grdina	Cuyahoga	Region 3	2955 E 71st St	Cleveland, OH, 44104-4101	(216) 812-1543	Harold S. Booker	...	NR	NC	

5 rows × 45 columns

In [8]:

```
# Selecting only 5 columns thats required for furture steps
df_BR15_16 = df_BR15_16[['District IRN', 'Building IRN', 'Building Name', 'Enrollment', 'School_Year']]
df_BR15_16.head()
```

Out[8]:

	District IRN	Building IRN	Building Name	Enrollment	School_Year
447	43786	224	Adlai Stevenson School	430	2015-2016
448	43786	318	Menlo Park Academy	367	2015-2016
449	43786	489	Almira	499	2015-2016
450	43786	729	Andrew J Rickoff	477	2015-2016
451	43786	828	Anton Grdina	371	2015-2016

In [9]:

```
#Find out the shape
df_BR15_16.shape
```

Out[9]:

```
(117, 5)
```

In [10]:

```
#Find out the info about the data
df_BR15_16.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 117 entries, 447 to 563
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   District IRN   117 non-null    int64 
 1   Building IRN   117 non-null    int64 
 2   Building Name  117 non-null    object
 3   Enrollment     117 non-null    object
 4   School_Year    117 non-null    object
dtypes: int64(2), object(3)
memory usage: 5.5+ KB
```

In [11]:

```
#Checking the unique count for building IRN
df_BR15_16['Building IRN'].unique
```

Out[11]:

```
<bound method Series.unique of 447       224
448       318
449       489
450       729
451       828
        ...  
559     86306
560    133215
561    133520
562    147389
563    147397
Name: Building IRN, Length: 117, dtype: int64>
```

In [12]:

```
#Checking the unique count for building Name
df_BR15_16['Building Name'].unique
```

Out[12]:

```
<bound method Series.unique of 447                    Adlai Stevenson School
448                        Menlo Park Academy
449                                    Almira
450                          Andrew J Rickoff
451                              Anton Grdina
                        ...                  
559         Health Careers Center High School
560             Intergenerational School, The
561                          Citizens Academy
562                SuccessTech Academy School
563    Cleveland School of Science & Medicine
Name: Building Name, Length: 117, dtype: object>
```

In [ ]:

In [ ]:

Enrollment 2016-2017 Data Cleaning & Preprocessing

In [13]:

```
#Importing data enrollment 2016-2017 data
df_BR16_17 = pd.read_excel(r'C:/Users/patil/Desktop/CMDS Skills Assessment/Data/BUILDING_RATING_2016_2017
df_BR16_17.head()
```

Out[13]:

	District IRN	District Name	Building IRN	Building Name	County	Region	Address	City, State, Zip	Phone	Principal	...	4 Year Grad Rate 2016	Letter Grade of 5 Year Grad Rate 2015	Y G R 2
0	138	Pathway School of Discovery	138	Pathway School of Discovery	Montgomery	Region 10	173 Avondale Dr	Dayton, OH, 45404-2123	(937) 235-5498	Keith B. Colbert	...	NC	NR	
1	139	Alliance Academy of Cincinnati	139	Alliance Academy of Cincinnati	Hamilton	Region 13	1712 Duck Creek Rd	Cincinnati, OH, 45207-1644	(513) 751-5555	Elizabeth L. King	...	NC	NR	
2	222	Wildwood Environmental Academy	222	Wildwood Environmental Academy	Lucas	Region 1	1546 Dartford Rd	Maumee, OH, 43537-1374	(419) 868-9885	Elizabeth A. Lewin	...	90.9	NR	
3	236	Ohio Connections Academy, Inc	236	Ohio Connections Academy, Inc	Cuyahoga	Region 11	3615 Superior Ave E	Cleveland, OH, 44114-2229	(513) 486-9120	NaN	...	67.6	F	7
4	296	Summit Academy Community School-Columbus	296	Summit Academy Community School-Columbus	Franklin	Region 11	2521 Fairwood Ave Ste 100	Columbus, OH, 43207-2712	(614) 237-5497	Cheryl L. Elliott	...	NC	NR	

5 rows × 44 columns

In [14]:

```
#Adding School Year Column
df_BR16_17['School_Year'] = '2016-2017'
```

In [15]:

```
#Rename enrollment column
df_BR16_17.rename(columns={'Enrollment 2016-2017': 'Enrollment'}, inplace=True)
```

In [16]:

```python
# filtering the data for cleaveland district
df_BR16_17 = df_BR16_17.loc[df_BR16_17['District IRN'] == 43786]
df_BR16_17.head()
```

Out[16]:

	District IRN	District Name	Building IRN	Building Name	County	Region	Address	City, State, Zip	Phone	Principal	...	Letter Grade of 5 Year Grad Rate 2015	5 Year Grad Rate 2015	L Grac Prep Succ Compo
441	43786	Cleveland Municipal City	224	Adlai Stevenson School	Cuyahoga	Region 3	18300 Woda Avenue	Cleveland, OH, 44122-6441	(216) 838-5300	Christopher T. Wyland	...	NR	NC	
442	43786	Cleveland Municipal City	318	Menlo Park Academy	Cuyahoga	Region 3	14440 Triskett Rd	Cleveland, OH, 44111-2263	(440) 925-6365	Stacy J. Stuhldreher	...	NR	NC	
443	43786	Cleveland Municipal City	489	Almira	Cuyahoga	Region 3	3375 W 99th St	Cleveland, OH, 44102-4642	(216) 838-6150	James Greene	...	NR	NC	
444	43786	Cleveland Municipal City	729	Andrew J Rickoff	Cuyahoga	Region 3	3500 E 147th St	Cleveland, OH, 44120-4834	(216) 838-4150	SHELBY R. SCHUTT	...	NR	NC	
445	43786	Cleveland Municipal City	828	Anton Grdina	Cuyahoga	Region 3	2955 E 71st St	Cleveland, OH, 44104-4101	(216) 838-1150	Harold S. Booker	...	NR	NC	

5 rows × 45 columns

In [17]:

```
# Selecting only 5 columns thats required for furture steps
df_BR16_17 = df_BR16_17[['District IRN', 'Building IRN', 'Building Name', 'Enrollment', 'School_Year']]
df_BR16_17.head()
```

Out[17]:

	District IRN	Building IRN	Building Name	Enrollment	School_Year
441	43786	224	Adlai Stevenson School	445	2016-2017
442	43786	318	Menlo Park Academy	405	2016-2017
443	43786	489	Almira	491	2016-2017
444	43786	729	Andrew J Rickoff	457	2016-2017
445	43786	828	Anton Grdina	361	2016-2017

In [18]:

```
#Find out the shape
df_BR16_17.shape
```

Out[18]:

```
(116, 5)
```

In [19]:

```
#Checking the unique count for building IRN
df_BR16_17['Building IRN'].unique
```

Out[19]:

```
<bound method Series.unique of 441       224
442       318
443       489
444       729
445       828
        ...
552     68221
553     86306
554    133215
555    147389
556    147397
Name: Building IRN, Length: 116, dtype: int64>
```

In [20]:

```
#Checking the unique count for Building Name
df_BR16_17['Building Name'].unique
```

Out[20]:

```
<bound method Series.unique of 441                    Adlai Stevenson School
442                        Menlo Park Academy
443                                    Almira
444                          Andrew J Rickoff
445                              Anton Grdina
                          ...                
552                         Kenneth W Clement
553          Health Careers Center High School
554              Intergenerational School, The
555                 SuccessTech Academy School
556     Cleveland School of Science & Medicine
Name: Building Name, Length: 116, dtype: object>
```

In [ ]:

In [ ]:

Enrollment 2017-2018 Data Cleaning & Preprocessing

In [21]:

```
#Importing data for enrollment 2017-2018
df_BR17_18 = pd.read_excel(r'C:/Users/patil/Desktop/CMDS Skills Assessment/Data/BUILDING_OVERVIEW_2017_20
df_BR17_18.head()
```

Out[21]:

	District IRN	District Name	Building IRN	Building Name	County	Region	Address	City, State, Zip	Phone	Principal	...	4 Year Grad Rate 2017	Letter Grade of 5 Year Grad Rate 2016	5 Year Grad Rate 2016	Cc
0	45187	Ada Exempted Village	59	Ada Elementary School	Hardin	Region 6	435 Grand Ave	Ada, OH, 45810-1013	(419) 634-2341	Robin E. Vanbuskirk	...	NC	NR	NC	
1	45187	Ada Exempted Village	67	Ada High School	Hardin	Region 6	435 Grand Ave	Ada, OH, 45810-1013	(419) 634-2746	Robin E. Vanbuskirk	...	98.4	B	94.3	
2	44743	Sandusky City	83	Sandusky Middle School	Erie	Region 2	318 Columbus Ave	Sandusky, OH, 44870-2616	(419) 984-1180	Timothy P. Kozak	...	NC	NR	NC	
3	48520	Meigs Local	102	Meigs Primary School	Meigs	Region 16	36871 State Route 124	Middleport, OH, 45760-9717	(740) 742-3000	Kristin C. Baer	...	NC	NR	NC	
4	48520	Meigs Local	105	Meigs Intermediate School	Meigs	Region 16	36871 State Route 124	Middleport, OH, 45760-9717	(740) 742-2666	IRENE C. Murphy	...	NC	NR	NC	

5 rows × 43 columns

In [22]:

```
#Adding School Year Column
df_BR17_18['School_Year'] = '2017-2018'
```

In [23]:

```python
#Changing the column name
df_BR17_18.rename(columns={'Enrollment 2017-2018': 'Enrollment'}, inplace=True)
```

In [24]:

```python
# Filtering the data for cleaveland district
df_BR17_18 = df_BR17_18.loc[df_BR17_18['District IRN'] == 43786]
df_BR17_18.head()
```

Out[24]:

	District IRN	District Name	Building IRN	Building Name	County	Region	Address	City, State, Zip	Phone	Principal	...	Letter Grade of 5 Year Grad Rate 2016	5 Year Grad Rate 2016	Prep Suc Compo G
24	43786	Cleveland Municipal	224	Adlai Stevenson School	Cuyahoga	Region 3	18300 Woda Avenue	Cleveland, OH, 44122-6441	(216) 838-5300	Christopher T. Wyland	...	NR	NC	
49	43786	Cleveland Municipal	318	Menlo Park Academy	Cuyahoga	Region 3	2149 W 53rd St	Cleveland, OH, 44102-2263	(440) 925-6365	Stacy J. Stuhldreher	...	NR	NC	
88	43786	Cleveland Municipal	489	Almira	Cuyahoga	Region 3	3375 W 99th St	Cleveland, OH, 44102-4642	(216) 838-6150	James Greene	...	NR	NC	
143	43786	Cleveland Municipal	729	Andrew J Rickoff	Cuyahoga	Region 3	3500 E 147th St	Cleveland, OH, 44120-4834	(216) 838-4150	SHELBY R. SCHUTT	...	NR	NC	
160	43786	Cleveland Municipal	828	Anton Grdina	Cuyahoga	Region 3	2955 E 71st St	Cleveland, OH, 44104-4101	(216) 838-1150	Harold S. Booker	...	NR	NC	

5 rows × 44 columns

In [25]:

```
#Selecting only 5 columns thats required for furture steps
df_BR17_18 = df_BR17_18[['District IRN', 'Building IRN', 'Building Name', 'Enrollment', 'School_Year']]
df_BR17_18.head()
```

Out[25]:

	District IRN	Building IRN	Building Name	Enrollment	School_Year
24	43786	224	Adlai Stevenson School	443	2017-2018
49	43786	318	Menlo Park Academy	418	2017-2018
88	43786	489	Almira	547	2017-2018
143	43786	729	Andrew J Rickoff	441	2017-2018
160	43786	828	Anton Grdina	396	2017-2018

In [26]:

```
#Find out the shape
df_BR17_18.shape
```

Out[26]:

```
(123, 5)
```

In [ ]:

Dataframe Merging for Enrollment 2015-2016, 2016-2017, 2017-2018

In [27]:

```
#Appending the enrollment dataset using concat function
frames = [df_BR15_16, df_BR16_17, df_BR17_18]

df_enroll = pd.concat(frames)
df_enroll
```

Out[27]:

	District IRN	Building IRN	Building Name	Enrollment	School_Year
447	43786	224	Adlai Stevenson School	430	2015-2016
448	43786	318	Menlo Park Academy	367	2015-2016
449	43786	489	Almira	499	2015-2016
450	43786	729	Andrew J Rickoff	477	2015-2016
451	43786	828	Anton Grdina	371	2015-2016
...	...	...	...	...	...
3042	43786	86306	Martin Luther King Jr. Campus	356	2017-2018
3198	43786	133215	Intergenerational School, The	247	2017-2018
3214	43786	133520	Citizens Academy	410	2017-2018
3219	43786	133629	Horizon Science Acad Cleveland	440	2017-2018
3390	43786	147397	Cleveland School of Science & Medicine	401	2017-2018

356 rows × 5 columns

In [28]:

```
#Find out the info about the data
df_enroll.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 356 entries, 447 to 3390
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   District IRN   356 non-null    int64 
 1   Building IRN   356 non-null    int64 
 2   Building Name  356 non-null    object
 3   Enrollment     356 non-null    object
 4   School_Year    356 non-null    object
dtypes: int64(2), object(3)
memory usage: 16.7+ KB
```

In [29]:

```
#Find out the shape
df_enroll.shape
```

Out[29]:

```
(356, 5)
```

In [30]:

```
#Find out the missing values
df_enroll.isnull().sum()
```

Out[30]:

```
District IRN     0
Building IRN     0
Building Name    0
Enrollment       0
School_Year      0
dtype: int64
```

In [31]:

```
#Checking the unique count for building Name
df_enroll['Building Name'].unique
```

Out[31]:

```
<bound method Series.unique of 447                           Adlai Stevenson School
448                        Menlo Park Academy
449                                    Almira
450                          Andrew J Rickoff
451                              Anton Grdina
                         ...
3042             Martin Luther King Jr. Campus
3198             Intergenerational School, The
3214                         Citizens Academy
3219              Horizon Science Acad Cleveland
3390     Cleveland School of Science & Medicine
Name: Building Name, Length: 356, dtype: object>
```

In [32]:

```
#Checking the unique count for building IRN
df_enroll['Building IRN'].unique
```

Out[32]:

```
<bound method Series.unique of 447        224
448         318
449         489
450         729
451         828
          ...
3042     86306
3198    133215
3214    133520
3219    133629
3390    147397
Name: Building IRN, Length: 356, dtype: int64>
```

In [ ]:

In [ ]:

Cleaning Building Value Added Grade Data for 2015-2016, 2016-2017, 2017-2018

Value Added Grade 2015-2016 Data Cleaning & Preprocessing

In [33]:

```
#Importing Value Added data for 2015-2016
df_BVA15_16 = pd.read_excel(r'C:/Users/patil/Desktop/CMDS Skills Assessment/Data/BUILDING_VA_2015_2016.xl
df_BVA15_16.head()
```

Out[33]:

	District IRN	District Name	Building IRN	Building Name	County	Region	Overall Value Added Grade	Overall Composite	Gifted Value Added Grade	Gifted Composite	Students with Disabilities Value Added Grade	Stu Disabi comp
0	138	Pathway School of Discovery	138	Pathway School of Discovery	Montgomery	Region 10	C	0.87	NR	NC	C	
1	139	Alliance Academy of Cincinnati	139	Alliance Academy of Cincinnati	Hamilton	Region 13	F	-6.45	NR	NC	D	
2	222	Wildwood Environmental Academy	222	Wildwood Environmental Academy	Lucas	Region 1	A	4.85	NR	NC	C	
3	236	Ohio Connections Academy, Inc	236	Ohio Connections Academy, Inc	Cuyahoga	Region 3	F	-4.60	NR	NC	C	
4	296	Summit Academy Community School-Columbus	296	Summit Academy Community School-Columbus	Franklin	Region 11	C	-0.68	NR	NC	A	

In [34]:

```
#Find out the shape
df_BVA15_16.shape
```

Out[34]:

```
(3387, 17)
```

In [35]:

```
#Adding School Year Column
df_BVA15_16['School_Year'] = '2015-2016'
```

In [36]:

```
# Filtering the data for cleaveland district
df_BVA15_16 = df_BVA15_16.loc[df_BVA15_16['District IRN'] == 43786]
df_BVA15_16.head()
```

Out[36]:

	District IRN	District Name	Building IRN	Building Name	County	Region	Overall Value Added Grade	Overall Composite	Gifted Value Added Grade	Gifted Composite	Students with Disabilities Value Added Grade	Students with Disabilities composite
447	43786	Cleveland Municipal City	224	Adlai Stevenson School	Cuyahoga	Region 3	F	-6.56	NR	NC	F	-3.48
448	43786	Cleveland Municipal City	318	Menlo Park Academy	Cuyahoga	Region 3	C	-0.54	C	-0.49	NR	NC
449	43786	Cleveland Municipal City	489	Almira	Cuyahoga	Region 3	F	-6.16	NR	NC	F	-5.10
450	43786	Cleveland Municipal City	729	Andrew J Rickoff	Cuyahoga	Region 3	F	-6.40	NR	NC	F	-5.68
451	43786	Cleveland Municipal City	828	Anton Grdina	Cuyahoga	Region 3	F	-7.53	NR	NC	D	-1.53

In [37]:

```
#Checking the column names
df_BVA15_16.columns
```

Out[37]:

```
Index(['District IRN', 'District Name', 'Building IRN', 'Building Name',
       'County', 'Region', 'Overall Value Added Grade', 'Overall Composite',
       'Gifted Value Added Grade', 'Gifted Composite',
       'Students with Disabilities Value Added Grade',
       'Students with Disabilities composite', 'Lowest 20% Value Added Grade',
       'Lowest 20% Value Added Composite', 'High Mobility Value Added Grade',
       'High Mobility Composite', 'Watermark', 'School_Year'],
      dtype='object')
```

In [38]:

```
# Selecting only 5 columns thats required for furture steps
df_BVA15_16 = df_BVA15_16[['District IRN', 'Building IRN', 'Building Name', 'Overall Value Added Grade',
df_BVA15_16.head()
```

Out[38]:

	District IRN	Building IRN	Building Name	Overall Value Added Grade	School_Year
447	43786	224	Adlai Stevenson School	F	2015-2016
448	43786	318	Menlo Park Academy	C	2015-2016
449	43786	489	Almira	F	2015-2016
450	43786	729	Andrew J Rickoff	F	2015-2016
451	43786	828	Anton Grdina	F	2015-2016

In [39]:

```
#Find out the shape
df_BVA15_16.shape
```

Out[39]:

```
(117, 5)
```

In [ ]:

Value Added Grade 2016-2017 Data Cleaning & Preprocessing

In [40]:

```
#Importing Value Added data for 2016-2017
df_BVA16_17 = pd.read_excel(r'C:/Users/patil/Desktop/CMDS Skills Assessment/Data/BUILDING_VA_2016_2017.xl
df_BVA16_17.head()
```

Out[40]:

	District IRN	District Name	Building IRN	Building Name	County	Region	Overall Value Added Grade	Overall Composite	Gifted Value Added Grade	Gifted Composite	Students with Disabilities Value Added Grade	Stuc Disabi comp
0	138	Pathway School of Discovery	138	Pathway School of Discovery	Montgomery	Region 10	C	0.18	NR	NC	C	
1	139	Alliance Academy of Cincinnati	139	Alliance Academy of Cincinnati	Hamilton	Region 13	F	-5.34	NR	NC	F	
2	222	Wildwood Environmental Academy	222	Wildwood Environmental Academy	Lucas	Region 1	A	2.11	NR	NC	F	
3	236	Ohio Connections Academy, Inc	236	Ohio Connections Academy, Inc	Cuyahoga	Region 11	F	-13.89	NR	NC	F	
4	296	Summit Academy Community School-Columbus	296	Summit Academy Community School-Columbus	Franklin	Region 11	F	-2.13	NR	NC	C	

In [41]:

```
#Find out the shape
df_BVA16_17.shape
```

Out[41]:

```
(3374, 17)
```

In [42]:

```
#Adding School Year Column
df_BVA16_17['School_Year'] = '2016-2017'
```

In [43]:

```
# Filtering the data for cleaveland district
df_BVA16_17 = df_BVA16_17.loc[df_BVA16_17['District IRN'] == 43786]
df_BVA16_17.head()
```

Out[43]:

	District IRN	District Name	Building IRN	Building Name	County	Region	Overall Value Added Grade	Overall Composite	Gifted Value Added Grade	Gifted Composite	Students with Disabilities Value Added Grade	Students with Disabilities composite
439	43786	Cleveland Municipal City	224	Adlai Stevenson School	Cuyahoga	Region 3	F	-8.05	NR	NC	F	-2.79
440	43786	Cleveland Municipal City	318	Menlo Park Academy	Cuyahoga	Region 3	F	-10.14	F	-10.13	NR	NC
441	43786	Cleveland Municipal City	489	Almira	Cuyahoga	Region 3	F	-9.32	NR	NC	F	-4.95
442	43786	Cleveland Municipal City	729	Andrew J Rickoff	Cuyahoga	Region 3	F	-7.93	NR	NC	F	-5.60
443	43786	Cleveland Municipal City	828	Anton Grdina	Cuyahoga	Region 3	F	-8.68	NR	NC	C	-0.44

In [44]:

```
# Selecting only 5 columns thats required for furture steps
df_BVA16_17 = df_BVA16_17[['District IRN', 'Building IRN', 'Building Name', 'Overall Value Added Grade',
df_BVA16_17.head()
```

Out[44]:

	District IRN	Building IRN	Building Name	Overall Value Added Grade	School_Year
439	43786	224	Adlai Stevenson School	F	2016-2017
440	43786	318	Menlo Park Academy	F	2016-2017
441	43786	489	Almira	F	2016-2017
442	43786	729	Andrew J Rickoff	F	2016-2017
443	43786	828	Anton Grdina	F	2016-2017

In [45]:

```
#Find out the shape
df_BVA16_17.shape
```

Out[45]:

```
(119, 5)
```

In [ ]:

Value Added Grade 2017-2018 Data Cleaning & Preprocessing

In [46]:

```
#Importing Value Added data for 2017-2018
df_BVA17_18 = pd.read_excel(r'C:/Users/patil/Desktop/CMDS Skills Assessment/Data/BUILDING_VA_2017_2018.xl
df_BVA17_18.head()
```

```
C:\Users\patil\anaconda3\lib\site-packages\openpyxl\worksheet\header_footer.py:48: UserWarning: Cannot p
arse header or footer so it will be ignored
  warn("""Cannot parse header or footer so it will be ignored""")
```

Out[46]:

	District IRN	District Name	Building IRN	Building Name	County	Region	Overall Value Added Grade	Overall Composite	Gifted Value Added Grade	Gifted Composite	Students with Disabilities Value Added Grade	Stu Disabi Comp
0	131	Glass City Academy	131	Glass City Academy	Lucas	Region 1	F	-3.47	NR	NC	NR	
1	138	Pathway School of Discovery	138	Pathway School of Discovery	Montgomery	Region 10	F	-5.67	NR	NC	C	
2	139	Alliance Academy of Cincinnati	139	Alliance Academy of Cincinnati	Hamilton	Region 13	F	-8.56	NR	NC	F	
3	222	Wildwood Environmental Academy	222	Wildwood Environmental Academy	Lucas	Region 1	B	1.32	NR	NC	F	
4	236	Ohio Connections Academy, Inc	236	Ohio Connections Academy, Inc	Cuyahoga	Region 11	F	-25.65	NR	NC	F	

In [47]:

```
#Find out the shape
df_BVA17_18.shape
```

Out[47]:

```
(3423, 17)
```

In [48]:

```
#Adding School Year Column
df_BVA17_18['School_Year'] = '2017-2018'
```

In [49]:

```
# Filtering the data for cleaveland district
df_BVA17_18 = df_BVA17_18.loc[df_BVA17_18['District IRN'] == 43786]
df_BVA17_18.head()
```

Out[49]:

	District IRN	District Name	Building IRN	Building Name	County	Region	Overall Value Added Grade	Overall Composite	Gifted Value Added Grade	Gifted Composite	Students with Disabilities Value Added Grade	Students with Disabilities Composite
467	43786	Cleveland Municipal	224	Adlai Stevenson School	Cuyahoga	Region 3	F	-8.39	NR	NC	C	-1.00
468	43786	Cleveland Municipal	318	Menlo Park Academy	Cuyahoga	Region 3	F	-4.99	F	-4.97	NR	NC
469	43786	Cleveland Municipal	489	Almira	Cuyahoga	Region 3	F	-7.71	NR	NC	F	-5.33
470	43786	Cleveland Municipal	729	Andrew J Rickoff	Cuyahoga	Region 3	F	-5.56	NR	NC	F	-5.06
471	43786	Cleveland Municipal	828	Anton Grdina	Cuyahoga	Region 3	F	-7.57	NR	NC	D	-1.45

In [50]:

```
# Selecting only 5 columns thats required for furture steps
df_BVA17_18 = df_BVA17_18[['District IRN', 'Building IRN', 'Building Name', 'Overall Value Added Grade',
df_BVA17_18.head()
```

Out[50]:

	District IRN	Building IRN	Building Name	Overall Value Added Grade	School_Year
467	43786	224	Adlai Stevenson School	F	2017-2018
468	43786	318	Menlo Park Academy	F	2017-2018
469	43786	489	Almira	F	2017-2018
470	43786	729	Andrew J Rickoff	F	2017-2018
471	43786	828	Anton Grdina	F	2017-2018

In [51]:

```
#Find out the shape
df_BVA17_18.shape
```

Out[51]:

```
(123, 5)
```

Dataframe Merging for All Value Added Grade 2015-2016, 2016-2017, 2017-2018

In [52]:

```
#Appending the enrollment dataset using concat function
frames = [df_BVA15_16, df_BVA16_17, df_BVA17_18]

df_VA = pd.concat(frames)
df_VA.head()
```

Out[52]:

	District IRN	Building IRN	Building Name	Overall Value Added Grade	School_Year
447	43786	224	Adlai Stevenson School	F	2015-2016
448	43786	318	Menlo Park Academy	C	2015-2016
449	43786	489	Almira	F	2015-2016
450	43786	729	Andrew J Rickoff	F	2015-2016
451	43786	828	Anton Grdina	F	2015-2016

In [53]:

```
#Find out the shape
df_VA.shape
```

Out[53]:

```
(359, 5)
```

In [54]:

```
#Find out the column names
df_VA.columns
```

Out[54]:

```
Index(['District IRN', 'Building IRN', 'Building Name',
       'Overall Value Added Grade', 'School_Year'],
      dtype='object')
```

In [55]:

```
#Checking the unique count for building Name
df_VA['Building Name'].unique
```

Out[55]:

```
<bound method Series.unique of 447                    Adlai Stevenson School
448                        Menlo Park Academy
449                                    Almira
450                          Andrew J Rickoff
451                              Anton Grdina
                        ...
585             Martin Luther King Jr. Campus
586             Intergenerational School, The
587                          Citizens Academy
588            Horizon Science Acad Cleveland
589    Cleveland School of Science & Medicine
Name: Building Name, Length: 359, dtype: object>
```

In [ ]:

In [ ]:

In [ ]:

Performance Index Percent Data Cleaning & merging 2015-2016, 2016-2017, 2017-2018

Performance Index Percent 2015-2016 Data Cleaning & Preprocessing

```
In [56]:
#Importing perfomance index data for 2015-2016
df_BPIP15_16 = pd.read_excel(r'C:/Users/patil/Desktop/CMDS Skills Assessment/Data/BUILDING_ACHIEVEMENT_20
df_BPIP15_16.head()
```

Out[56]:

	Building IRN	Building Name	District IRN	District Name	County	Region	Address	City and Zip Code	Phone #	Principal	...	Percent of Gifted Students Not Tested	Percent of Gifted Students Below	P of St
0	59	Ada Elementary School	45187	Ada Exempted Village	Hardin	Region 6	435 Grand Ave	Ada, OH, 45810-1013	(419) 634-2341	Robin E. Vanbuskirk	...	0	0	
1	67	Ada High School	45187	Ada Exempted Village	Hardin	Region 6	435 Grand Ave	Ada, OH, 45810-1013	(419) 634-2746	Robin E. Vanbuskirk	...	0	0	
2	83	Sandusky Middle School	44743	Sandusky City	Erie	Region 2	318 Columbus Ave	Sandusky, OH, 44870-2616	(419) 984-1180	Marie A. Prieto	...	0	2.8	
3	102	Meigs Primary School	48520	Meigs Local	Meigs	Region 16	36871 State Route 124	Middleport, OH, 45760-9717	(740) 742-3000	Kristin C. Baer	...	NC	NC	
4	105	Meigs Intermediate School	48520	Meigs Local	Meigs	Region 16	36871 State Route 124	Middleport, OH, 45760-9717	(740) 742-2666	IRENE C. Murphy	...	0	5.7	

5 rows × 32 columns

```
In [57]:
#Find out the shape
df_BPIP15_16.shape
```

```
Out[57]: (3387, 32)
```

In [58]:

```python
#Adding School Year Column
df_BPIP15_16['School_Year'] = '2015-2016'
```

In [59]:

```python
#Changing the column name
df_BPIP15_16.rename(columns={'Performance Index Percent 2015-16': 'Performance_Index_Percent'}, inplace=T
```

In [60]:

```python
# Filtering the data for cleaveland district
df_BPIP15_16 = df_BPIP15_16.loc[df_BPIP15_16['District IRN'] == 43786]
df_BPIP15_16.head()
```

Out[60]:

	Building IRN	Building Name	District IRN	District Name	County	Region	Address	City and Zip Code	Phone #	Principal	...	Percent of Gifted Students Below	Percent of Gifted Students Basic
20	224	Adlai Stevenson School	43786	Cleveland Municipal	Cuyahoga	Region 3	18300 Woda Avenue	Cleveland, OH, 44122-6441	(216) 482-2950	Christopher T. Wyland	...	NC	NC
83	489	Almira	43786	Cleveland Municipal	Cuyahoga	Region 3	3375 W 99th St	Cleveland, OH, 44102-4642	(216) 838-6150	Laverne Hooks	...	NC	NC
137	729	Andrew J Rickoff	43786	Cleveland Municipal	Cuyahoga	Region 3	3500 E 147th St	Cleveland, OH, 44120-4834	(216) 838-4150	Gloriane R. Smith	...	NC	NC
154	828	Anton Grdina	43786	Cleveland Municipal	Cuyahoga	Region 3	2955 E 71st St	Cleveland, OH, 44104-4101	(216) 812-1543	Harold S. Booker	...	NC	NC
192	1040	Artemus Ward	43786	Cleveland Municipal	Cuyahoga	Region 3	4315 W 140th St	Cleveland, OH, 44135-2128	(216) 920-7055	Chris P. Myslenski	...	NC	NC

5 rows × 33 columns

In [61]:

```
#Find out the column names
df_BPIP15_16.columns
```

Out[61]:

```
Index(['Building IRN', 'Building Name', 'District IRN', 'District Name',
       'County', 'Region', 'Address', 'City and Zip Code', 'Phone #',
       'Principal', 'Performance Index Score 2015-16',
       'Performance_Index_Percent', 'Letter Grade of Performance Index',
       'Percent of Students Not Tested', 'Percent of Students Below',
       'Percent of Students Basic', 'Percent of Students Proficient',
       'Percent of Students Accelerated', 'Percent of Students Advanced',
       'Percent of Students Advanced Plus',
       'Gifted Performance Index Score 2015-16',
       'Gifted Performance Index 2015-16',
       'Percent of Gifted Students Not Tested',
       'Percent of Gifted Students Below', 'Percent of Gifted Students Basic',
       'Percent of Gifted Students Proficient',
       'Percent of Gifted Students Accelerated',
       'Percent of Gifted Students Advanced',
       'Percent of Gifted Students Advanced Plus',
       'Performance Index Score 2014-15', 'Performance Index Score 2013-14',
       'Watermark', 'School_Year'],
      dtype='object')
```

In [62]:

```
# Selecting only 5 columns thats required for furture steps
df_BPIP15_16 = df_BPIP15_16[['District IRN', 'Building IRN', 'Building Name', 'Performance_Index_Percent'
df_BPIP15_16.head()
```

Out[62]:

	District IRN	Building IRN	Building Name	Performance_Index_Percent	School_Year
20	43786	224	Adlai Stevenson School	36.6	2015-2016
83	43786	489	Almira	38.7	2015-2016
137	43786	729	Andrew J Rickoff	36.9	2015-2016
154	43786	828	Anton Grdina	32.4	2015-2016
192	43786	1040	Artemus Ward	47.3	2015-2016

In [63]:

```
#Find out the shape
df_BPIP15_16.shape
```

Out[63]:

```
(101, 5)
```

In [ ]:

```
```

Performance Index Percent 2016-2017 Data Cleaning & Preprocessing

In [64]:

```
#Importing perfomance index data for 2016-2017
df_BPIP16_17 = pd.read_excel(r'C:/Users/patil/Desktop/CMDS Skills Assessment/Data/BUILDING_ACHIEVEMENT_20
df_BPIP16_17.head()
```

Out[64]:

	Building IRN	Building Name	District IRN	District Name	County	Region	Address	City and Zip Code	Phone #	Principal	...	Percent of Gifted Students Not Tested	Percent of Gifted Students Below	F of St
0	59	Ada Elementary School	45187	Ada Exempted Village	Hardin	Region 6	435 Grand Ave	Ada, OH, 45810-1013	(419) 634-2341	Robin E. Vanbuskirk	...	0.0	0.0	
1	67	Ada High School	45187	Ada Exempted Village	Hardin	Region 6	435 Grand Ave	Ada, OH, 45810-1013	(419) 634-2746	Robin E. Vanbuskirk	...	0.0	2.0	
2	83	Sandusky Middle School	44743	Sandusky City	Erie	Region 2	318 Columbus Ave	Sandusky, OH, 44870-2616	(419) 984-1180	Marie A. Prieto	...	0.0	1.4	
3	102	Meigs Primary School	48520	Meigs Local	Meigs	Region 16	36871 State Route 124	Middleport, OH, 45760-9717	(740) 742-3000	Kristin C. Baer	...	NC	NC	
4	105	Meigs Intermediate School	48520	Meigs Local	Meigs	Region 16	36871 State Route 124	Middleport, OH, 45760-9717	(740) 742-2666	IRENE C. Murphy	...	0.0	1.5	

5 rows × 32 columns

In [65]:

```
#Find out the shape
df_BPIP16_17.shape
```

Out[65]:

```
(3374, 32)
```

In [66]:

```
#Adding School Year Column
df_BPIP16_17['School_Year'] = '2016-2017'
```

In [67]:

```
#Changing the column name
df_BPIP16_17.rename(columns={'Performance Index Percent 2016-17': 'Performance_Index_Percent'}, inplace=T
```

In [68]:

```
# Filtering the data for cleaveland district
df_BPIP16_17 = df_BPIP16_17.loc[df_BPIP16_17['District IRN'] == 43786]
df_BPIP16_17.head()
```

Out[68]:

	Building IRN	Building Name	District IRN	District Name	County	Region	Address	City and Zip Code	Phone #	Principal	...	Percent of Gifted Students Below	Percent of Gifted Students Basic
20	224	Adlai Stevenson School	43786	Cleveland Municipal	Cuyahoga	Region 3	18300 Woda Avenue	Cleveland, OH, 44122-6441	(216) 838-5300	Christopher T. Wyland	...	33.3	11.1
44	318	Menlo Park Academy	43786	Cleveland Municipal	Cuyahoga	Region 3	14440 Triskett Rd	Cleveland, OH, 44111-2263	(440) 925-6365	NaN	...	1.9	8.5
83	489	Almira	43786	Cleveland Municipal	Cuyahoga	Region 3	3375 W 99th St	Cleveland, OH, 44102-4642	(216) 838-6150	James Greene	...	14.3	21.4
137	729	Andrew J Rickoff	43786	Cleveland Municipal	Cuyahoga	Region 3	3500 E 147th St	Cleveland, OH, 44120-4834	(216) 838-4150	SHELBY R. SCHUTT	...	0.0	0.0
153	828	Anton Grdina	43786	Cleveland Municipal	Cuyahoga	Region 3	2955 E 71st St	Cleveland, OH, 44104-4101	(216) 838-1150	Harold S. Booker	...	33.3	0.0

5 rows × 33 columns

In [69]:

```
# Selecting only 5 columns thats required for furture steps
df_BPIP16_17 = df_BPIP16_17[['District IRN', 'Building IRN', 'Building Name', 'Performance_Index_Percent'
df_BPIP16_17.head()
```

Out[69]:

	District IRN	Building IRN	Building Name	Performance_Index_Percent	School_Year
20	43786	224	Adlai Stevenson School	41.1	2016-2017
44	43786	318	Menlo Park Academy	88.7	2016-2017
83	43786	489	Almira	39.0	2016-2017
137	43786	729	Andrew J Rickoff	39.3	2016-2017
153	43786	828	Anton Grdina	33.6	2016-2017

In [70]:

```
#Find out the shape
df_BPIP16_17.shape
```

Out[70]:

```
(119, 5)
```

In [ ]:

Performance Index Percent 2017-2018 Data Cleaning & Preprocessing

In [71]:

```
#Importing perfomance index data for 2017-2018
df_BPIP17_18 = pd.read_excel(r'C:/Users/patil/Desktop/CMDS Skills Assessment/Data/BUILDING_ACHIEVEMENT_20
df_BPIP17_18.head()
```

Out[71]:

	Building IRN	Building Name	District IRN	District Name	County	Region	Address	City and Zip	Phone Number	Principal	...	Gifted Performance Index Score 2017-18	Per
0	59	Ada Elementary School	45187	Ada Exempted Village	Hardin	Region 6	435 Grand Ave	Ada, OH, 45810-1013	(419) 634-2341	Robin E. Vanbuskirk	...	117.722	
1	67	Ada High School	45187	Ada Exempted Village	Hardin	Region 6	435 Grand Ave	Ada, OH, 45810-1013	(419) 634-2746	Robin E. Vanbuskirk	...	116.346	
2	83	Sandusky Middle School	44743	Sandusky City	Erie	Region 2	318 Columbus Ave	Sandusky, OH, 44870-2616	(419) 984-1180	Timothy P. Kozak	...	115.392	
3	102	Meigs Primary	48520	Meigs	Meigs	Region	36871 State	Middleport, OH,	(740) 742	Kristin C.		NC	

In [72]:

```
#Find out the shape
df_BPIP17_18.shape
```

Out[72]:

```
(3423, 32)
```

In [73]:

```
#Adding School Year Column
df_BPIP17_18['School_Year'] = '2017-2018'
```

In [74]:

```
#Changing the column name
df_BPIP17_18.rename(columns={'Performance Index Percent 2017-18': 'Performance_Index_Percent'}, inplace=T
```

In [75]:

```
# Filtering the data for cleaveland district
df_BPIP17_18 = df_BPIP17_18.loc[df_BPIP17_18['District IRN'] == 43786]
df_BPIP17_18.head()
```

Out[75]:

	Building IRN	Building Name	District IRN	District Name	County	Region	Address	City and Zip	Phone Number	Principal	...	Gifted Performance Index Percent 2017-18	Perc of Gif Stude l Tes
24	224	Adlai Stevenson School	43786	Cleveland Municipal	Cuyahoga	Region 3	18300 Woda Avenue	Cleveland, OH, 44122-6441	(216) 838-5300	Christopher T. Wyland	...	NC	
49	318	Menlo Park Academy	43786	Cleveland Municipal	Cuyahoga	Region 3	2149 W 53rd St	Cleveland, OH, 44102-2263	(440) 925-6365	Stacy J. Stuhldreher	...	90.7	
88	489	Almira	43786	Cleveland Municipal	Cuyahoga	Region 3	3375 W 99th St	Cleveland, OH, 44102-4642	(216) 838-6150	James Greene	...	66.7	
143	729	Andrew J Rickoff	43786	Cleveland Municipal	Cuyahoga	Region 3	3500 E 147th St	Cleveland, OH, 44120-4834	(216) 838-4150	SHELBY R. SCHUTT	...	NC	
160	828	Anton Grdina	43786	Cleveland Municipal	Cuyahoga	Region 3	2955 E 71st St	Cleveland, OH, 44104-4101	(216) 838-1150	Harold S. Booker	...	NC	

5 rows × 33 columns

In [76]:

```
# Selecting only 5 columns thats required for furture steps
df_BPIP17_18 = df_BPIP17_18[['District IRN', 'Building IRN', 'Building Name', 'Performance_Index_Percent'
df_BPIP17_18.head()
```

Out[76]:

	District IRN	Building IRN	Building Name	Performance_Index_Percent	School_Year
24	43786	224	Adlai Stevenson School	40.1	2017-2018
49	43786	318	Menlo Park Academy	90.7	2017-2018
88	43786	489	Almira	41.6	2017-2018
143	43786	729	Andrew J Rickoff	41.6	2017-2018
160	43786	828	Anton Grdina	35.6	2017-2018

In [77]:

```
#Find out the shape
df_BPIP17_18.shape
```

Out[77]:

```
(123, 5)
```

In [ ]:

In [ ]:

Dataframe Merging for All Performance Index Percent 2015-2016, 2016-2017, 2017-2018

In [78]:

```
#Appending the Performance Index Percent dataset using concat function
frames = [df_BPIP15_16, df_BPIP16_17, df_BPIP17_18]

df_PIP = pd.concat(frames)
df_PIP
```

Out[78]:

	District IRN	Building IRN	Building Name	Performance_Index_Percent	School_Year
20	43786	224	Adlai Stevenson School	36.6	2015-2016
83	43786	489	Almira	38.7	2015-2016
137	43786	729	Andrew J Rickoff	36.9	2015-2016
154	43786	828	Anton Grdina	32.4	2015-2016
192	43786	1040	Artemus Ward	47.3	2015-2016
...	...	...	...	...	...
3042	43786	86306	Martin Luther King Jr. Campus	38.2	2017-2018
3198	43786	133215	Intergenerational School, The	65.7	2017-2018
3214	43786	133520	Citizens Academy	64.5	2017-2018
3219	43786	133629	Horizon Science Acad Cleveland	50.1	2017-2018
3390	43786	147397	Cleveland School of Science & Medicine	80	2017-2018

343 rows × 5 columns

In [79]:

```
#Find out the shape
df_PIP.shape
```

Out[79]:

```
(343, 5)
```

In [80]:

```
#Checking the unique count for building Name
df_PIP['Building Name'].unique
```

Out[80]:

```
<bound method Series.unique of 20                      Adlai Stevenson School
83                                      Almira
137                           Andrew J Rickoff
154                               Anton Grdina
192                               Artemus Ward
                         ...                  
3042             Martin Luther King Jr. Campus
3198             Intergenerational School, The
3214                          Citizens Academy
3219            Horizon Science Acad Cleveland
3390    Cleveland School of Science & Medicine
Name: Building Name, Length: 343, dtype: object>
```

In [ ]:

In [ ]:

Merging Enrollment, Value Added Grade & Performance Index Percent Dataframes

In [81]:

```
#Merging enrollment, Value added and performance index dataframe togeather
df_CMSD = (df_enroll.merge(df_VA,on=['District IRN', 'Building IRN', 'Building Name', 'School_Year'],how=
df_CMSD.head()
```

Out[81]:

	District IRN	Building IRN	Building Name	Enrollment	School_Year	Overall Value Added Grade	Performance_Index_Percent
0	43786	224	Adlai Stevenson School	430	2015-2016	F	36.6
1	43786	318	Menlo Park Academy	367	2015-2016	C	NaN
2	43786	489	Almira	499	2015-2016	F	38.7
3	43786	729	Andrew J Rickoff	477	2015-2016	F	36.9
4	43786	828	Anton Grdina	371	2015-2016	F	32.4

In [82]:

```
# rename the column without space for easy use
df_CMSD.rename(columns={'District IRN': 'District_IRN',
                        'Building IRN': 'Building_IRN',
                        'Building Name': 'Building_Name',
                        'Overall Value Added Grade': 'Overall_Value_Added_Grade' },
                         inplace=True)
```

In [83]:

```
#Find out the shape
df_CMSD.shape
```

Out[83]:

```
(359, 7)
```

In [84]:

```
#Find out the datatypes
df_CMSD.dtypes
```

Out[84]:

```
District_IRN                   int64
Building_IRN                   int64
Building_Name                 object
Enrollment                    object
School_Year                   object
Overall_Value_Added_Grade     object
Performance_Index_Percent     object
dtype: object
```

In [85]:

```
#Converting enrollment dtype from object to numeric
df_CMSD[['Enrollment']] = df_CMSD[['Enrollment']].apply(pd.to_numeric)
print(df_CMSD.dtypes)
```

```
District_IRN                    int64
Building_IRN                    int64
Building_Name                  object
Enrollment                    float64
School_Year                    object
Overall_Value_Added_Grade      object
Performance_Index_Percent      object
dtype: object
```

In [86]:

```
#Converting Performance Index Percent dtype from object to numeric
df_CMSD["Performance_Index_Percent"] = pd.to_numeric(df_CMSD["Performance_Index_Percent"], errors='coerce
print(df_CMSD.dtypes)
```

```
District_IRN                    int64
Building_IRN                    int64
Building_Name                  object
Enrollment                    float64
School_Year                    object
Overall_Value_Added_Grade      object
Performance_Index_Percent     float64
dtype: object
```

In [87]:

```
#Checking the info about the data
df_CMSD.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 359 entries, 0 to 358
Data columns (total 7 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   District_IRN                 359 non-null    int64  
 1   Building_IRN                 359 non-null    int64  
 2   Building_Name                359 non-null    object 
 3   Enrollment                   356 non-null    float64
 4   School_Year                  359 non-null    object 
 5   Overall_Value_Added_Grade    359 non-null    object 
 6   Performance_Index_Percent    339 non-null    float64
dtypes: float64(2), int64(2), object(3)
memory usage: 19.8+ KB
```

In [88]:

```
#Describe the data
df_CMSD.describe()
```

Out[88]:

	District_IRN	Building_IRN	Enrollment	Performance_Index_Percent
count	359.0	359.000000	356.000000	339.000000
mean	43786.0	25837.111421	366.676966	47.737758
std	0.0	27294.566571	156.060352	12.385176
min	43786.0	224.000000	48.000000	25.900000
25%	43786.0	11291.000000	265.000000	38.350000
50%	43786.0	15722.000000	357.000000	44.300000
75%	43786.0	32060.000000	431.750000	54.100000
max	43786.0	147397.000000	1276.000000	90.700000

In [89]:

```
#Filling the missing values with 0
df_CMSD = df_CMSD.fillna(0)
```

In [90]:

```
#df_CMSD.fillna(0)
df_CMSD.isnull().sum()
```

Out[90]:

```
District_IRN                0
Building_IRN                0
Building_Name               0
Enrollment                  0
School_Year                 0
Overall_Value_Added_Grade    0
Performance_Index_Percent    0
dtype: int64
```

In [91]:

```
df_CMSD
```

Out[91]:

	District_IRN	Building_IRN	Building_Name	Enrollment	School_Year	Overall_Value_Added_Grade	Performance_Index_Percent
0	43786	224	Adlai Stevenson School	430.0	2015-2016	F	36.6
1	43786	318	Menlo Park Academy	367.0	2015-2016	C	0.0
2	43786	489	Almira	499.0	2015-2016	F	38.7
3	43786	729	Andrew J Rickoff	477.0	2015-2016	F	36.9
4	43786	828	Anton Grdina	371.0	2015-2016	F	32.4
...	...	...	...	...	...	...	...
354	43786	133629	Horizon Science Acad Cleveland	440.0	2017-2018	D	50.1
355	43786	147397	Cleveland School of Science & Medicine	401.0	2017-2018	A	80.0
356	43786	12852	Citizens Academy East	0.0	2016-2017	F	68.1
357	43786	15261	Citizens Academy Southeast	0.0	2016-2017	NR	65.7
358	43786	133520	Citizens Academy	0.0	2016-2017	F	66.1

359 rows × 7 columns

In [92]:

```
#Rearranging the column names
df_CMSD = df_CMSD[['District_IRN', 'Building_IRN', 'Building_Name', 'School_Year', 'Enrollment', 'Overall
df_CMSD.head()
```

Out[92]:

	District_IRN	Building_IRN	Building_Name	School_Year	Enrollment	Overall_Value_Added_Grade	Performance_Index_Percent
0	43786	224	Adlai Stevenson School	2015-2016	430.0	F	36.6
1	43786	318	Menlo Park Academy	2015-2016	367.0	C	0.0
2	43786	489	Almira	2015-2016	499.0	F	38.7
3	43786	729	Andrew J Rickoff	2015-2016	477.0	F	36.9
4	43786	828	Anton Grdina	2015-2016	371.0	F	32.4

In [ ]:

In [ ]:

Download the Combined Dataframe as Excel File

In [ ]:

```
#Giving the file name to dataframe
CMSD_Final_Data = pd.DataFrame(df_CMSD)
file_name = 'CMSD_Final_Data.xls'

#Saving the dataframe into excel
CMSD_Final_Data.to_excel(file_name)
print('DataFrame is written to Excel File successfully.')
```

In [ ]:

2nd Part of Assessment

Aggregate the data

```
In [93]:  #Aggregating Building IRN for enrollment for last 3 years
          df_CMSD.groupby(df_CMSD["Building_IRN"]).Enrollment.agg(["min", "max", "sum", "count", "mean"])
```

```
Out[93]:
                 min    max     sum  count        mean
  Building_IRN
           224  430.0  445.0  1318.0      3  439.333333
           318  367.0  418.0  1190.0      3  396.666667
           489  491.0  547.0  1537.0      3  512.333333
           729  441.0  477.0  1375.0      3  458.333333
           828  361.0  396.0  1128.0      3  376.000000
           ...    ...    ...     ...    ...         ...
        133215  241.0  250.0   738.0      3  246.000000
        133520    0.0  446.0   856.0      3  285.333333
        133629  440.0  440.0   440.0      1  440.000000
        147389   48.0   85.0   133.0      2   66.500000
        147397  382.0  404.0  1187.0      3  395.666667

  128 rows × 5 columns
```

```
In [94]:  #Aggregating Building IRN for PIP for last 3 years
          df_CMSD.groupby(df_CMSD["Building_IRN"]).Performance_Index_Percent.agg(["min", "max", "sum", "count", "me
```

```
Out[94]:
                   min   max    sum  count       mean
     Building_IRN
              224  36.6  41.1  117.8      3  39.266667
              318   0.0  90.7  179.4      3  59.800000
              489  38.7  41.6  119.3      3  39.766667
              729  36.9  41.6  117.8      3  39.266667
              828  32.4  35.6  101.6      3  33.866667
              ...   ...   ...    ...    ...        ...
           133215   0.0  72.2  137.9      3  45.966667
           133520   0.0  66.1  130.6      3  43.533333
           133629  50.1  50.1   50.1      1  50.100000
           147389   0.0  25.9   25.9      2  12.950000
           147397  76.0  80.0  234.8      3  78.266667

     128 rows × 5 columns
```

```
In [95]:  #Aggregating enrollment for last 3 school years
          df_CMSD.groupby(df_CMSD["School_Year"]).Enrollment.agg(["min", "max", "sum", "count", "mean"])
```

```
Out[95]:
                  min     max      sum  count        mean
     School_Year
       2015-2016  85.0  1252.0  43643.0    117  373.017094
       2016-2017   0.0  1276.0  42695.0    119  358.781513
       2017-2018  48.0   956.0  44199.0    123  359.341463
```

In [123]:

```
#Calculating sum of enrollment foe each year
total_by_year = df_CMSD.groupby('School_Year')['Enrollment'].sum().reset_index()
print(total_by_year)
```

```
  School_Year  Enrollment
0   2015-2016     43643.0
1   2016-2017     42695.0
2   2017-2018     44199.0
```

In [96]:

```
#Aggregating PIP for last 3 school years
df_CMSD.groupby(df_CMSD["School_Year"]).Performance_Index_Percent.agg(["min", "max", "sum", "count", "mea
```

Out[96]:

	min	max	sum	count	mean
School_Year					
2015-2016	0.0	76.0	4362.7	117	37.288034
2016-2017	0.0	88.7	5692.4	119	47.835294
2017-2018	27.6	90.7	6128.0	123	49.821138

In [97]:

```
# Calculate the sum of categories by year using groupby function
result = df_CMSD.groupby('School_Year')['Overall_Value_Added_Grade'].value_counts().unstack(fill_value=0)
print(result)
```

```
Overall_Value_Added_Grade School_Year   A   B   C   D   F  NR
0                           2015-2016   8   2  10  13  78   6
1                           2016-2017  10   3  16   8  78   4
2                           2017-2018  14  12  14   9  72   2
```

In [ ]:

In [ ]:

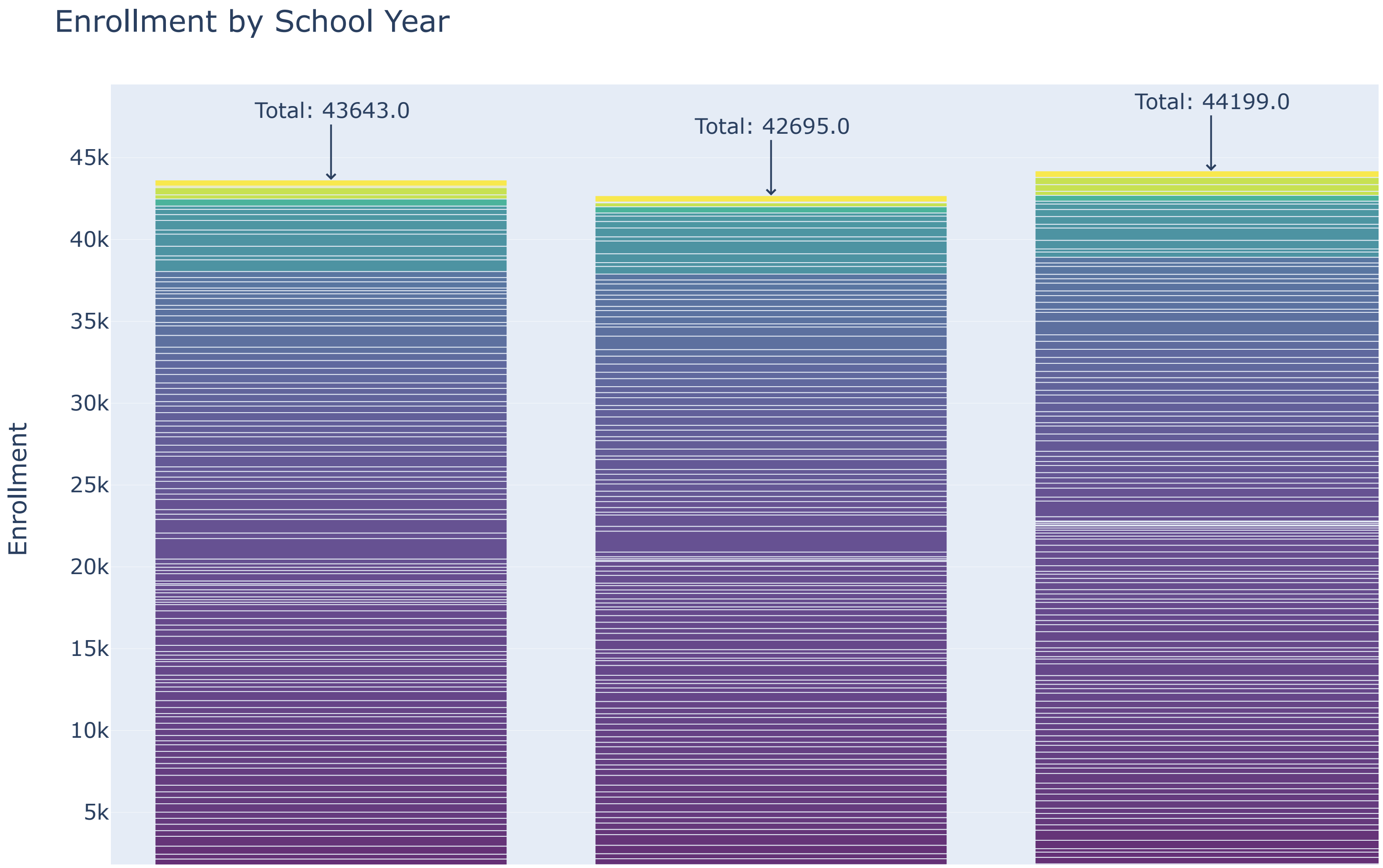
Data Visualization

Interactive Chart for total number of Enrollment by school year

In [122]:

```
#Interactive Chart for total number of Enrollment by school year
#Calculating sum of enrollment foe each year
total_by_year = df_CMSD.groupby('School_Year')['Enrollment'].sum().reset_index()
print(total_by_year)

# Plot the graph
fig = px.bar(df_CMSD,
             x="School_Year",
             y="Enrollment",
             color="Building_IRN",
             color_continuous_scale='Viridis',
             opacity=0.8,
             height=600
            )

fig.update_layout(
            title='Enrollment by School Year',
            xaxis_title='School Year',
            yaxis_title='Enrollment'
)

# Annotate the plot with total values
for i, row in total_by_year.iterrows():
    fig.add_annotation(
        x=row['School_Year'],
        y=row['Enrollment'],
        text=f"Total: {row['Enrollment']}",
        showarrow=True,
        arrowhead=4,
        ax=0,
        ay=-40
    )

fig.show()
```

```
  School_Year  Enrollment
0   2015-2016     43643.0
1   2016-2017     42695.0
2   2017-2018     44199.0
```

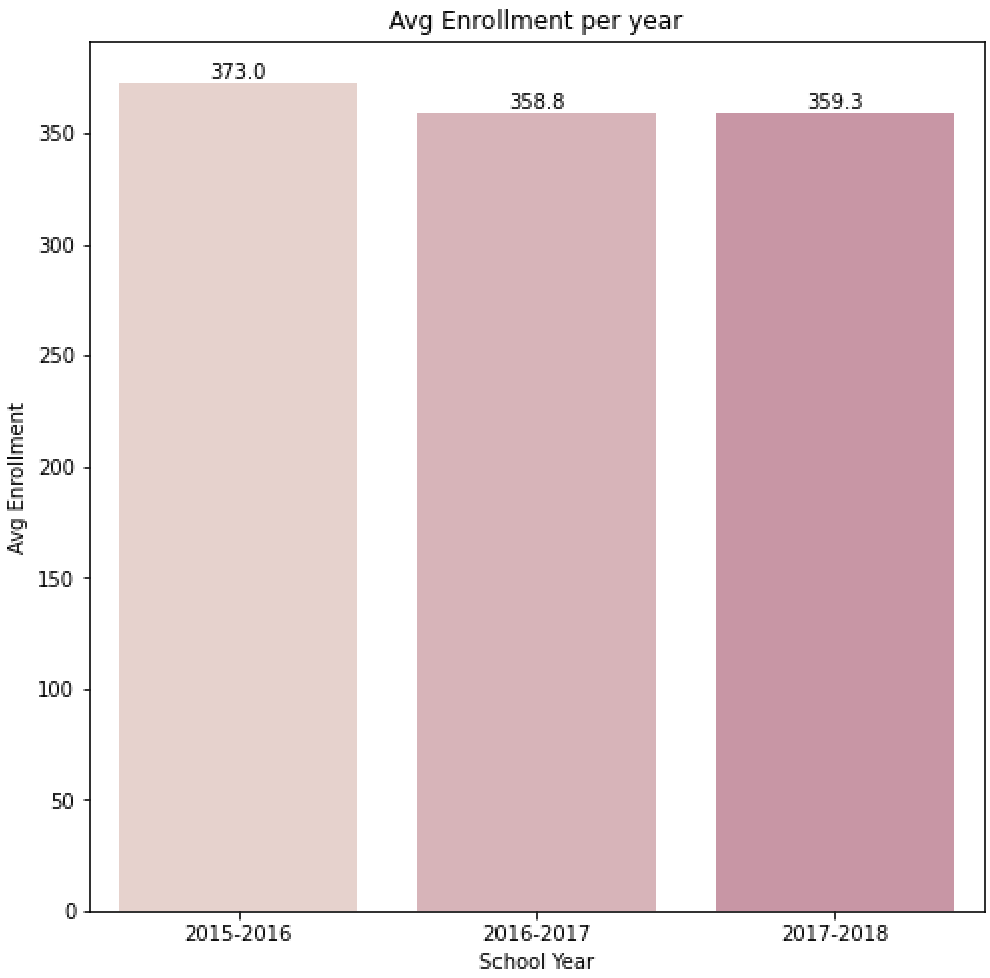


Bar plot for Average Number of Enrollment per year

In [101]:

```
# Average Number of Enrollment per year

# Calculate mean of total enrollment
Mean = df_CMSD.groupby('School_Year')['Enrollment'].mean().reset_index()

# Create a bar plot
fig, ax = plt.subplots(figsize=(8, 8))
Enrollment_plot = sns.barplot(
                             x='School_Year',
                             y='Enrollment',
                             data=Mean,
                             palette=sns.cubehelix_palette(8),
                             ax=ax)

# Add total values on top of each bar
for p in Enrollment_plot.patches:
    height = p.get_height()
    ax.text(p.get_x() + p.get_width() / 2., height + 2, f'{height:.1f}', ha="center")

# Show the plot
plt.title('Avg Enrollment per year')
plt.xlabel('School Year')
plt.ylabel('Avg Enrollment')
plt.show()
```

```
<ipython-input-101-b5fb51b327d3>:8: FutureWarning:



Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` v
ariable to `hue` and set `legend=False` for the same effect.


<ipython-input-101-b5fb51b327d3>:8: UserWarning:

The palette list has more values (8) than needed (3), which may not be intended.
```


Avg Enrollment per year
373.0
358.8
359.3
350
300
250
200
150
100
50
0
Avg Enrollment
2015-2016
2016-2017
2017-2018
School Year

Ploting the graph of enrollment percentage changes over the school year 2015-2016

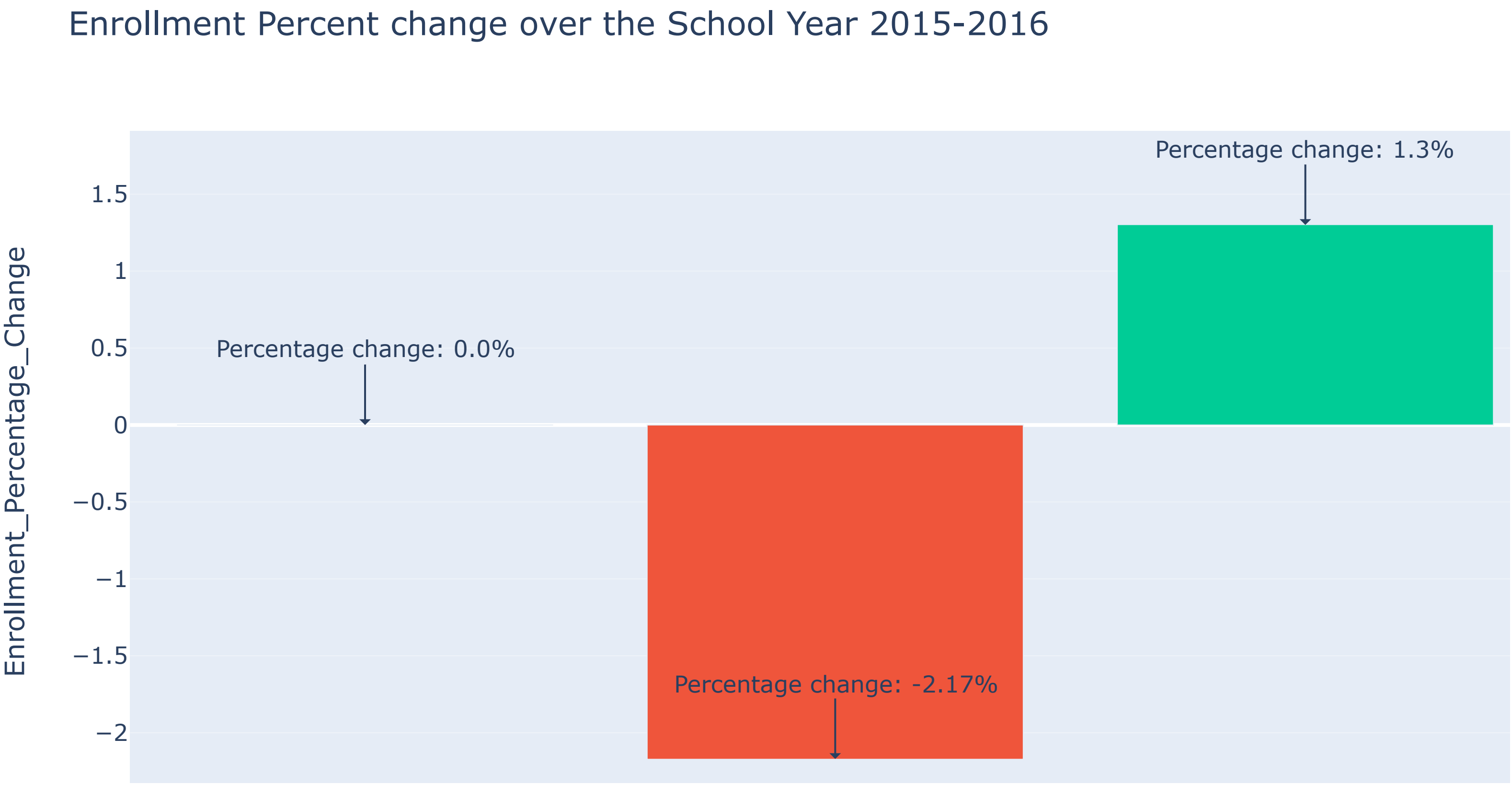
In [129]:

```
# Calculating the sum of enrollment for each year
total_by_year = df_CMSD.groupby('School_Year')['Enrollment'].sum().reset_index()

# Filter data for the year 2015-2016 and later
filtered_data = total_by_year[total_by_year['School_Year'] >= '2015-2016']

# Calculate the difference in enrollment relative to 2015-2016
filtered_data['Enrollment_Difference'] = filtered_data['Enrollment'] - filtered_data.loc[filtered_data['S

# Calculating the percentage change per year
filtered_data['Enrollment_Percentage_Change'] = (filtered_data['Enrollment_Difference'] / filtered_data['

#Filling Missing values with 0's
filtered_data = filtered_data.fillna(0)

# Print the results
print("Enrollment Changes in terms of school Year 2015-2016:")
print(filtered_data)
```

```
Enrollment Changes in terms of school Year 2015-2016:
  School_Year  Enrollment  Enrollment_Difference  Enrollment_Percentage_Change
0   2015-2016     43643.0                    0.0                      0.00000
1   2016-2017     42695.0                 -948.0                     -2.17217
2   2017-2018     44199.0                  556.0                      1.30226
```

In [104]:

```
# Plot the percentage change in enrollment
fig3 = px.bar(filtered_data,
              x='School_Year',
              y='Enrollment_Percentage_Change',
              color='School_Year',
              title='Percentage Change in Enrollment per Year',
              labels={'Enrollment_Percentage_Change': 'Percentage Change', 'School_Year': 'School Year'})

fig3.update_layout(
            title='Enrollment Percent change over the School Year 2015-2016',
            xaxis_title='School Year',
            yaxis_title='Enrollment_Percentage_Change')

# Rounding off the 'Enrollment_Percentage_Change' values to 2 decimal places
filtered_data['Enrollment_Percentage_Change'] = round(filtered_data['Enrollment_Percentage_Change'], 2)

# Annotating the plot
for i, row in filtered_data.iterrows():
    fig3.add_annotation(
        x=row['School_Year'],
        y=row['Enrollment_Percentage_Change'],
        text=f"Percentage change: {row['Enrollment_Percentage_Change']}%",
        showarrow=True,
        arrowhead=1,
        ax=0,
        ay=-40
    )

fig3.show()
```


Enrollment Percent change over the School Year 2015-2016
Enrollment_Percentage_Change
1.5
1
0.5
0
−0.5
−1
−1.5
−2
Percentage change: 0.0%
Percentage change: -2.17%
Percentage change: 1.3%

Calculation of number of school & Building name which has increased enrollment over 3 consecutive years

In [105]:

```
# Create a new column for increased enrollment
df_CMSD['Increased_Enrollment_3_Years'] = False

# For loop to Iterate the data and check for three consecutive increases per school
for school_name in df_CMSD['Building_Name'].unique():
    school_df = df_CMSD[df_CMSD['Building_Name'] == school_name]

    # Check if there are enough data points for the school
    if len(school_df) >= 3:
        mask = (
            (school_df['Enrollment'].shift(2) < school_df['Enrollment'].shift(1)) &
            (school_df['Enrollment'].shift(1) < school_df['Enrollment']) &
            school_df['Enrollment'].notna()
        )

        df_CMSD.loc[school_df.index[mask], 'Increased_Enrollment_3_Years'] = True

# Filter schools with increased enrollment for three consecutive years
schools_with_increased_3_years = df_CMSD[df_CMSD['Increased_Enrollment_3_Years']]
print("Schools with increased enrollment for three consecutive years:")
print(schools_with_increased_3_years[['Building_IRN','Building_Name']])

# Calculating the sum of schools
schools_with_increased_3_years_sum = df_CMSD['Increased_Enrollment_3_Years'].sum()
print(f"Number of schools with increased enrollment: {schools_with_increased_3_years_sum}")
```

```
Schools with increased enrollment for three consecutive years:
     Building_IRN                                      Building_Name
234           318                                 Menlo Park Academy
238           930      Cleveland Entrepreneurship Preparatory School
257          9285                                  Douglas MacArthur
261         10201                        Design Lab @ Health Careers
267         12030                 Near West Intergenerational School
268         12031  Entrepreneurship Preparatory School - Woodland...
269         12350                        Campus International School
271         12353                        New Technology HS@East Tech
272         12355          Facing History High School@Charles Mooney
275         12898                         Garfield Elementary School
276         13034  Village Preparatory School:: Woodland Hills Ca...
281         14913                 Lakeshore Intergenerational School
282         14918         Cleveland High School for the Digital Arts
283         14919                                         PACT @ JFK
284         14920                       Bard Early College Cleveland
285         15039                                     E3agle Academy
287         15239                              Stonebrook Montessori
290         15598                John Marshall School of Engineering
291         15599  John Marshall School of Business and Civic Lea...
292         15600     John Marshall School of Information Technology
306         18408                       Cleveland Early College High
312         23085                                     Mary M Bethune
319         26443                                 Nathan Hale School
320         27102                              Newton D Baker School
322         28720                                     Orchard School
324         29413                    Paul L Dunbar Elementary School
328         33902                                    Scranton School
330         37101                            Thomas Jefferson School
335         39149                                      Walton School
336         39206                    Warner Girls Leadership Academy
348         65565                 Marion C Seltzer Elementary School
349         65573                  Marion-Sterling Elementary School
Number of schools with increased enrollment: 32
```

```
<ipython-input-105-1e5e86b7635d>:2: 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.h
tml#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.htm
l#returning-a-view-versus-a-copy)
```

Calculation of number of school & Building name which has decreased enrollment over 3 consecutive years

In [106]:

```python
# Convert 'Enrollment' to numeric (if not already)
df_CMSD['Enrollment'] = pd.to_numeric(df_CMSD['Enrollment'], errors='coerce')

# Create a new column to flag schools with decreased enrollment for three consecutive years
df_CMSD['Decreased_Enrollment_3_Years'] = False

# Iterate through the data and check for three consecutive decreases per school
for school_name in df_CMSD['Building_Name'].unique():
    school_df = df_CMSD[df_CMSD['Building_Name'] == school_name]

    # Check if there are enough data points for the school
    if len(school_df) >= 3:
        mask = (
            (school_df['Enrollment'].shift(2) > school_df['Enrollment'].shift(1)) &
            (school_df['Enrollment'].shift(1) > school_df['Enrollment']) &
            school_df['Enrollment'].notna()
        )

        df_CMSD.loc[school_df.index[mask], 'Decreased_Enrollment_3_Years'] = True

# Filter schools with decreased enrollment for three consecutive years
schools_with_decreased_3_years = df_CMSD[df_CMSD['Decreased_Enrollment_3_Years']]
print("Schools with decreased enrollment for three consecutive years:")
print(schools_with_decreased_3_years[['Building_IRN','Building_Name']])

# Calculating the sum of schools
schools_with_decreased_3_years_sum = df_CMSD['Decreased_Enrollment_3_Years'].sum()
print(f"Number of schools with decreased enrollment: {schools_with_decreased_3_years_sum}")
```

```
<ipython-input-106-5644730189e5>:2: 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.h
tml#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.htm
l#returning-a-view-versus-a-copy)

<ipython-input-106-5644730189e5>:5: 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.h
tml#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.htm
l#returning-a-view-versus-a-copy)
```

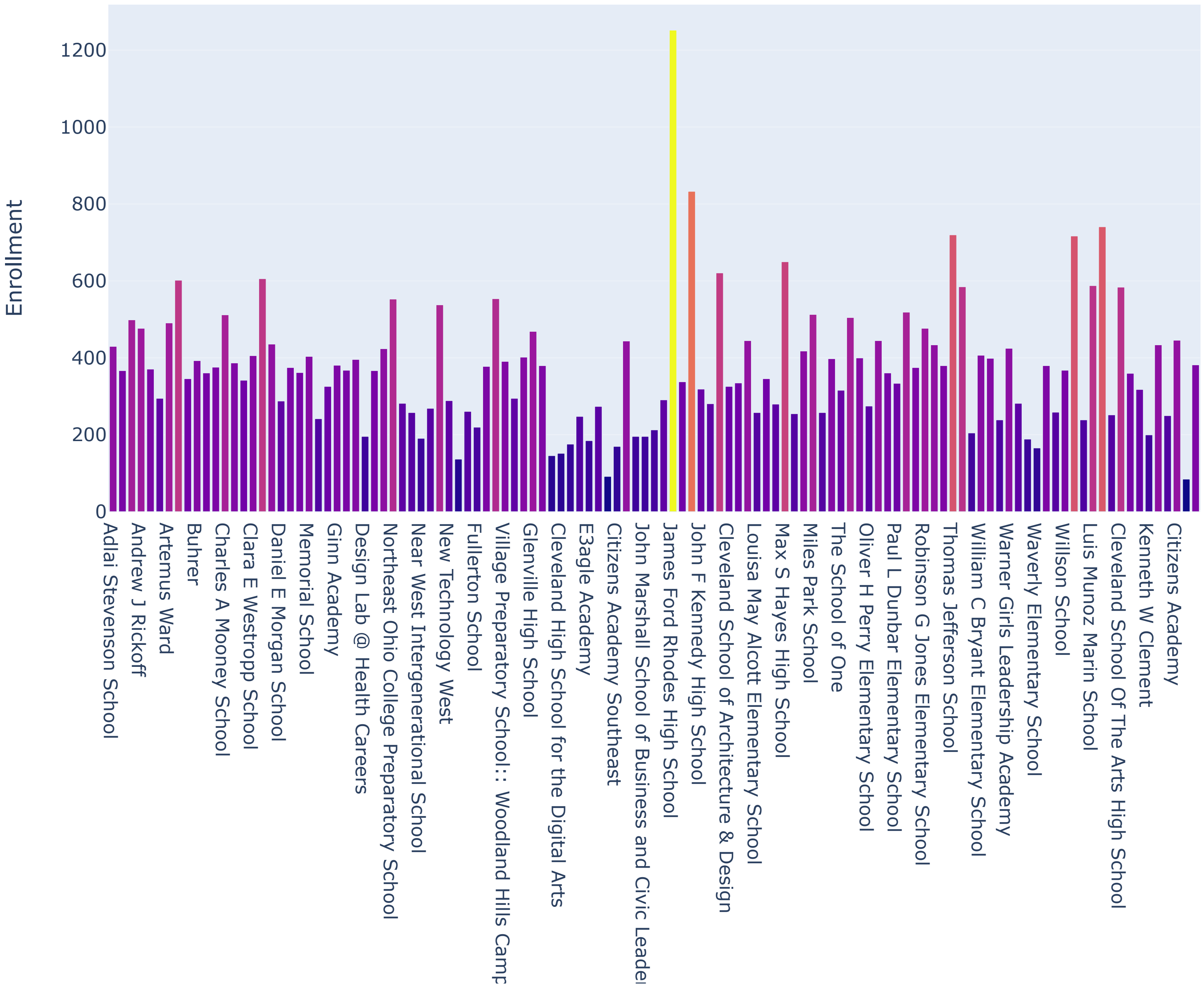
```
Schools with decreased enrollment for three consecutive years:
     Building_IRN                                  Building_Name
236           729                               Andrew J Rickoff
244          5066                                           Case
245          5637                                 Alfred Benesch
246          5892                        Charles A Mooney School
250          6429                                   Clark School
251          6940                         Collinwood High School
252          8060                         Daniel E Morgan School
253          8383                                        Denison
259          9555                     East Technical High School
262         10801                  Euclid Park Elementary School
277         13148                              Stepstone Academy
278         13292                       George Washington Carver
279         13680                          Glenville High School
286         15073  Hannah Gibbons-Nottingham Elementary School
302         17467                   Iowa-Maple Elementary School
304         17863       Jane Addams Business Careers High School
305         18325                         John Adams High School
308         21527                           Louis Agassiz School
309         21543                          Franklin D. Roosevelt
310         21550            Louisa May Alcott Elementary School
311         23069                           Mary B Martin School
316         24703                               Michael R. White
318         25874                              The School of One
323         29371                           Patrick Henry School
325         31963                               Riverside School
331         37457                      Tremont Montessori School
340         41517                                  Willow School
341         41541                                 Willson School
342         62315                       Lincoln-West High School
343         62323                           Whitney Young School
344         62760                         Luis Munoz Marin School
347         64576       Cleveland School Of The Arts High School
350         68221                              Kenneth W Clement
358        133520                               Citizens Academy
Number of schools with decreased enrollment: 34
```

Interactive bar plot for each years Enrollment distribution for each school

In [138]:

```
# Filter data for the year 2015-2016
df_2015_2016 = df_CMSD[df_CMSD['School_Year'] == '2015-2016']

# Plot with Plotly Express
fig = px.bar(
    df_2015_2016,
    x ='Building_Name',
    y ='Enrollment',
    color='Enrollment',
    title='Enrollment for each School in 2015-2016',
    labels={'Enrollment': 'Enrollment', 'School_Name': 'School Name'},
    height=800,
    width=900,
)

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

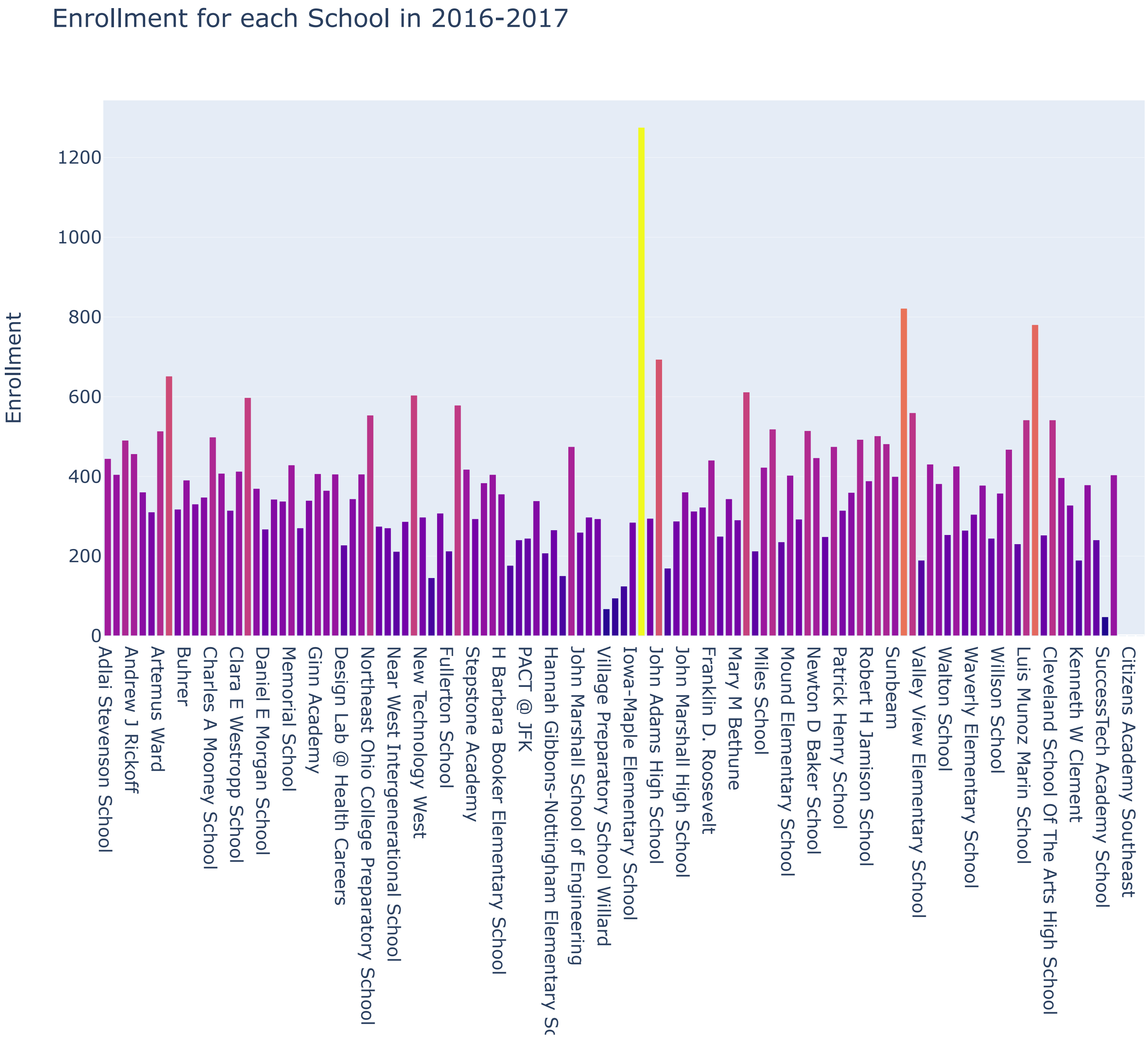
Enrollment for each School in 2015-2016



us
rship
Building_Name

In [137]:

```
# Filter data for the year 2015-2016
df_2016_2017 = df_CMSD[df_CMSD['School_Year'] == '2016-2017']

# Plot with Plotly Express
fig = px.bar(
    df_2016_2017,
    x ='Building_Name',
    y ='Enrollment',
    title='Enrollment for each School in 2016-2017',
    color='Enrollment',
    labels={'Enrollment': 'Enrollment', 'School_Name': 'School Name'},
    height=800,
    width=900,
)

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

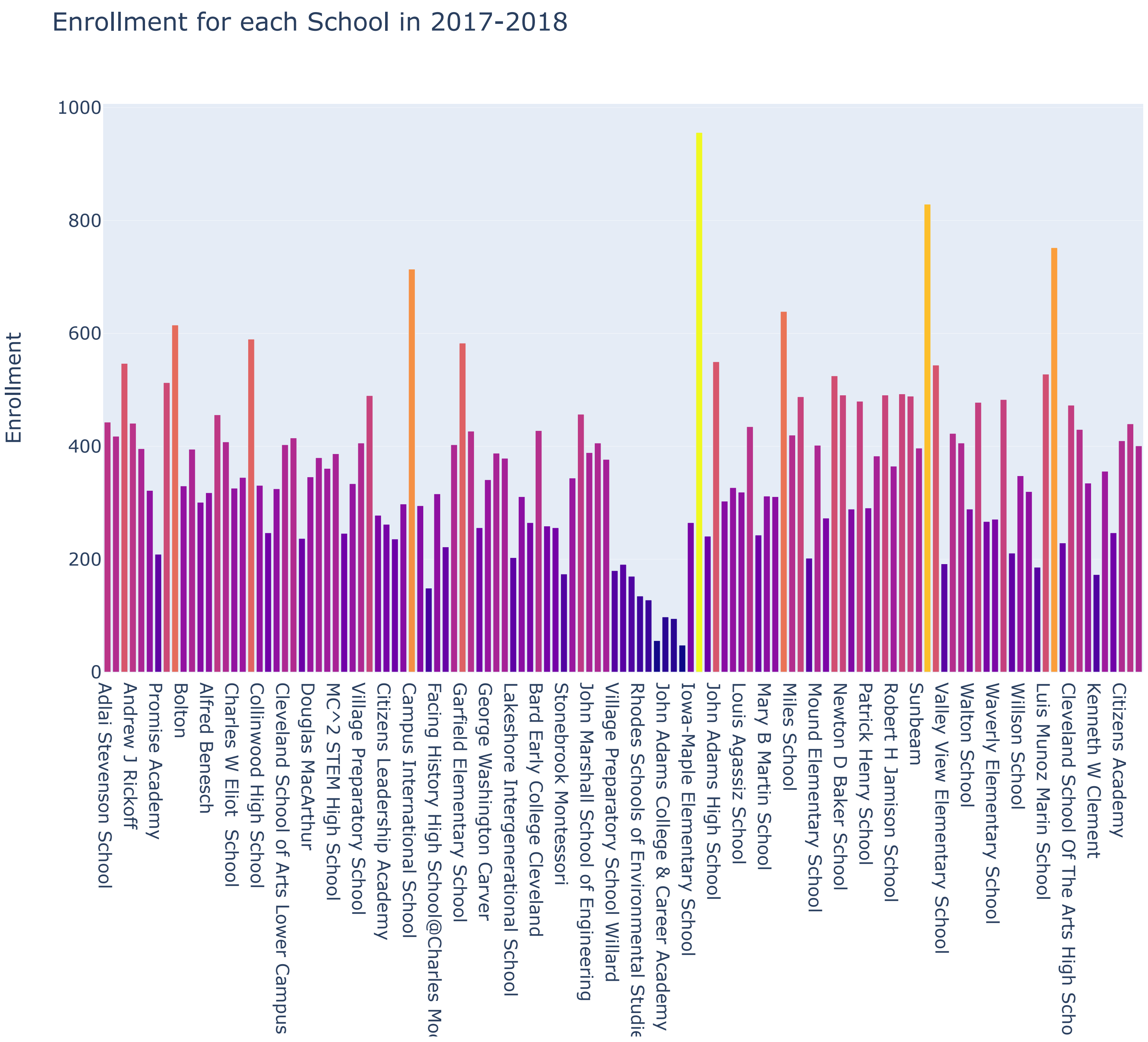

Enrollment for each School in 2016-2017
Enrollment
0
200
400
600
800
1000
1200
Adlai Stevenson School
Andrew J Rickoff
Artemus Ward
Buhrer
Charles A Mooney School
Clara E Westropp School
Daniel E Morgan School
Memorial School
Ginn Academy
Design Lab @ Health Careers
Northeast Ohio College Preparatory School
Near West Intergenerational School
New Technology West
Fullerton School
Stepstone Academy
H Barbara Booker Elementary School
PACT @ JFK
Hannah Gibbons-Nottingham Elementary Sc
John Marshall School of Engineering
Village Preparatory School Willard
Iowa-Maple Elementary School
John Adams High School
John Marshall High School
Franklin D. Roosevelt
Mary M Bethune
Miles School
Mound Elementary School
Newton D Baker School
Patrick Henry School
Robert H Jamison School
Sunbeam
Valley View Elementary School
Walton School
Waverly Elementary School
Willson School
Luis Munoz Marin School
Cleveland School Of The Arts High School
Kenneth W Clement
SuccessTech Academy School
Citizens Academy Southeast

chool

Building_Name

In [139]:

```
# Filter data for the year 2015-2016
df_2017_2018 = df_CMSD[df_CMSD['School_Year'] == '2017-2018']

# Plot with Plotly Express
fig = px.bar(
    df_2017_2018,
    x ='Building_Name',
    y ='Enrollment',
    color='Enrollment',
    title='Enrollment for each School in 2017-2018',
    labels={'Enrollment': 'Enrollment', 'School_Name': 'School Name'},
    height=800,
    width=900,
)

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

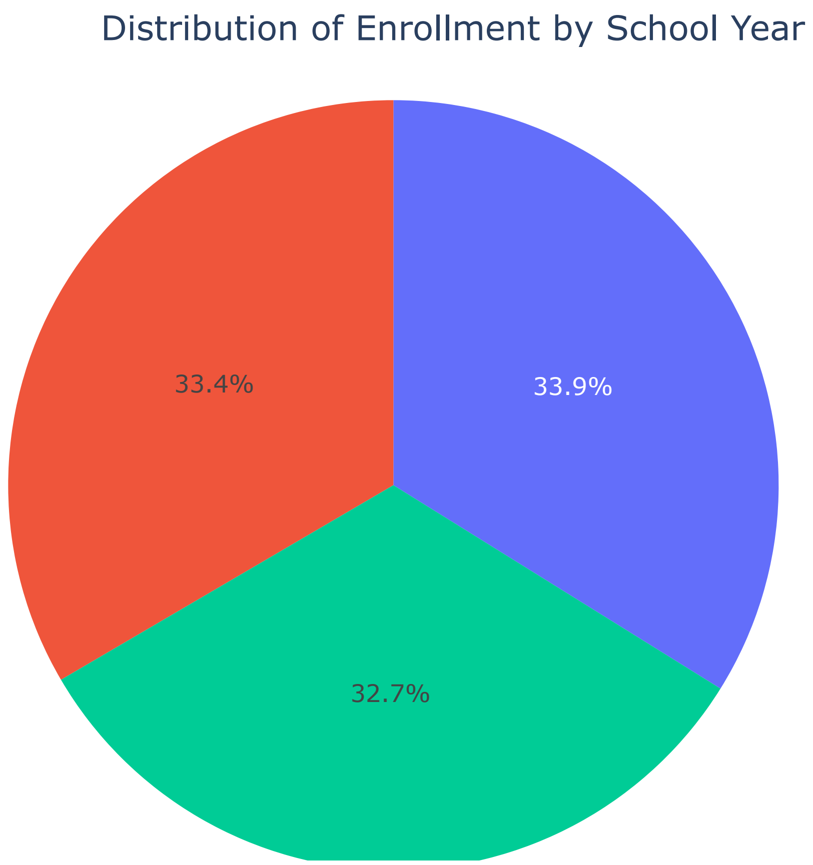

Enrollment for each School in 2017-2018
Enrollment
1000
800
600
400
200
0
Adlai Stevenson School
Andrew J Rickoff
Promise Academy
Bolton
Alfred Benesch
Charles W Eliot School
Collinwood High School
Cleveland School of Arts Lower Campus
Douglas MacArthur
MC^2 STEM High School
Village Preparatory School
Citizens Leadership Academy
Campus International School
Facing History High School@Charles Mo
Garfield Elementary School
George Washington Carver
Lakeshore Intergenerational School
Bard Early College Cleveland
Stonebrook Montessori
John Marshall School of Engineering
Village Preparatory School Willard
Rhodes Schools of Environmental Studie
John Adams College & Career Academy
Iowa-Maple Elementary School
John Adams High School
Louis Agassiz School
Mary B Martin School
Miles School
Mound Elementary School
Newton D Baker School
Patrick Henry School
Robert H Jamison School
Sunbeam
Valley View Elementary School
Walton School
Waverly Elementary School
Willson School
Luis Munoz Marin School
Cleveland School Of The Arts High Scho
Kenneth W Clement
Citizens Academy

oney
s
ol
Building_Name

Interactive plot for Enrollment distribution percentage over the year

In [147]:

```
#Interactive plot for Enrollment distribution percentage over the year
fig = px.pie(df_CMSD,
             values="Enrollment",
             names="School_Year",
            )
fig.update_layout(title_text='Distribution of Enrollment by School Year',title_x=0.5)
fig.show()
```



```
In [ ]:
```

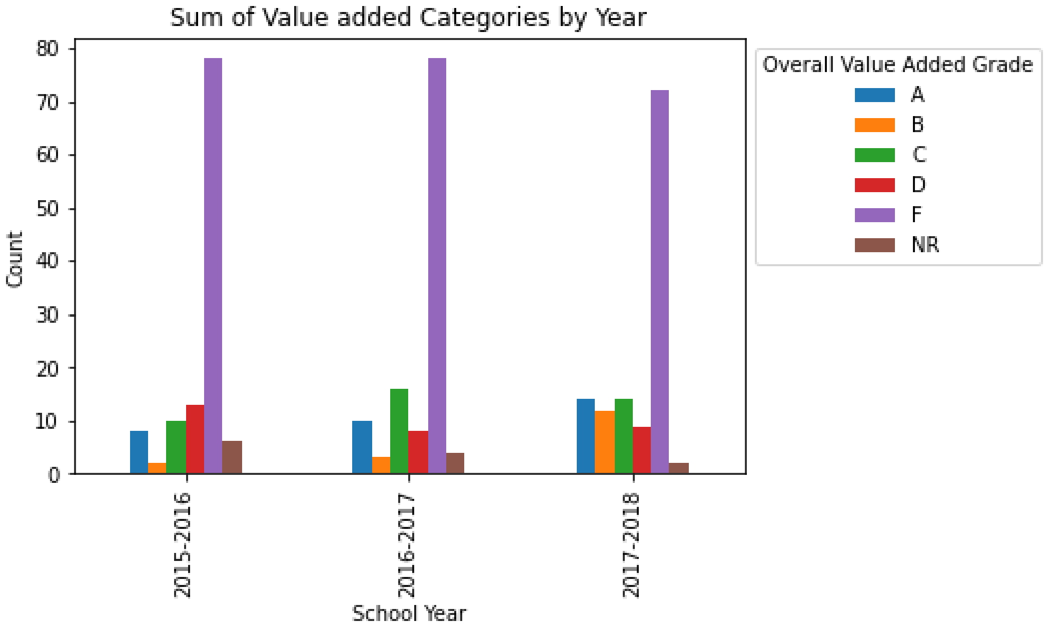
Data Visualization of Value Added Grades

Creat a visualization that shows how value added grade changed over the 3 years.

In [110]:

```
# Calculate the sum of categories by year
result = df_CMSD.groupby('School_Year')['Overall_Value_Added_Grade'].value_counts().unstack(fill_value=0)

# Print the result
print(result)

# Plot the graph
result.plot(x='School_Year', kind='bar')
plt.title('Sum of Value added Categories by Year')
plt.xlabel('School Year')
plt.ylabel('Count')
plt.legend(title='Overall Value Added Grade', bbox_to_anchor=(1, 1))

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

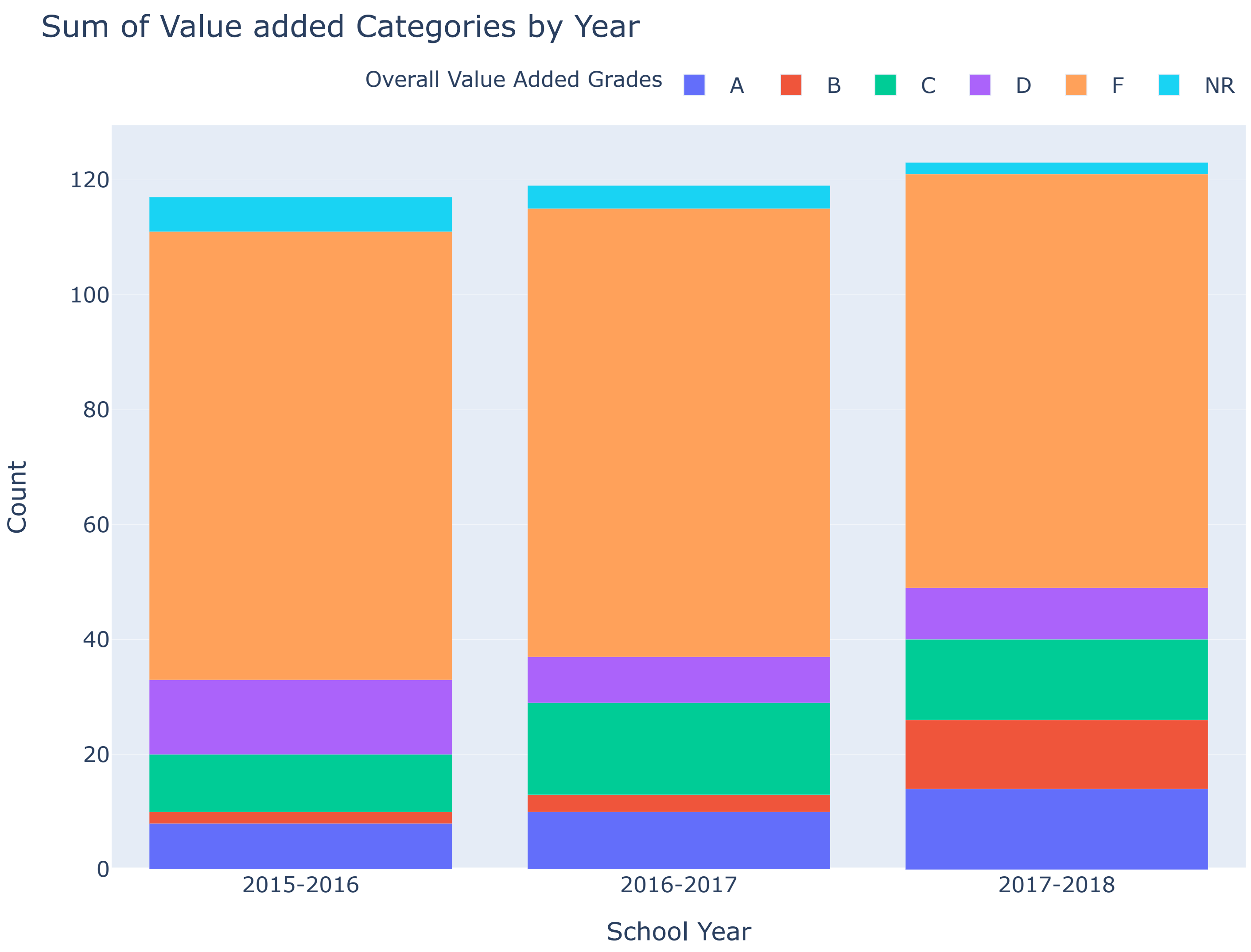
```
Overall_Value_Added_Grade School_Year   A   B   C   D   F  NR
0                           2015-2016   8   2  10  13  78   6
1                           2016-2017  10   3  16   8  78   4
2                           2017-2018  14  12  14   9  72   2
```


Sum of Value added Categories by Year
Count
0
10
20
30
40
50
60
70
80
2015-2016
2016-2017
2017-2018
School Year
Overall Value Added Grade
A
B
C
D
F
NR

Interactive stacked bar plot for how Overall value added grade is changing over the years

In [150]:

```python
# Calculate the sum of categories by year
result = df_CMSD.groupby('School_Year')['Overall_Value_Added_Grade'].value_counts().unstack(fill_value=0)

# Plot with Plotly Express
fig = px.bar(
    result,
    x='School_Year',
    y=result.columns[1:],
    title='Sum of Value added Categories by Year',
    labels={'value': 'Count', 'School_Year': 'School Year'},
    barmode='stack',
    height=600,
    width=800,
)

# Add legend
fig.update_layout(
    legend=dict(
        title='Overall Value Added Grades',
        orientation='h',
        yanchor='bottom',
        y=1.02,
        xanchor='right',
        x=1
    )
)

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


Sum of Value added Categories by Year
Overall Value Added Grades
A
B
C
D
F
NR
Count
0
20
40
60
80
100
120
2015-2016
2016-2017
2017-2018
School Year

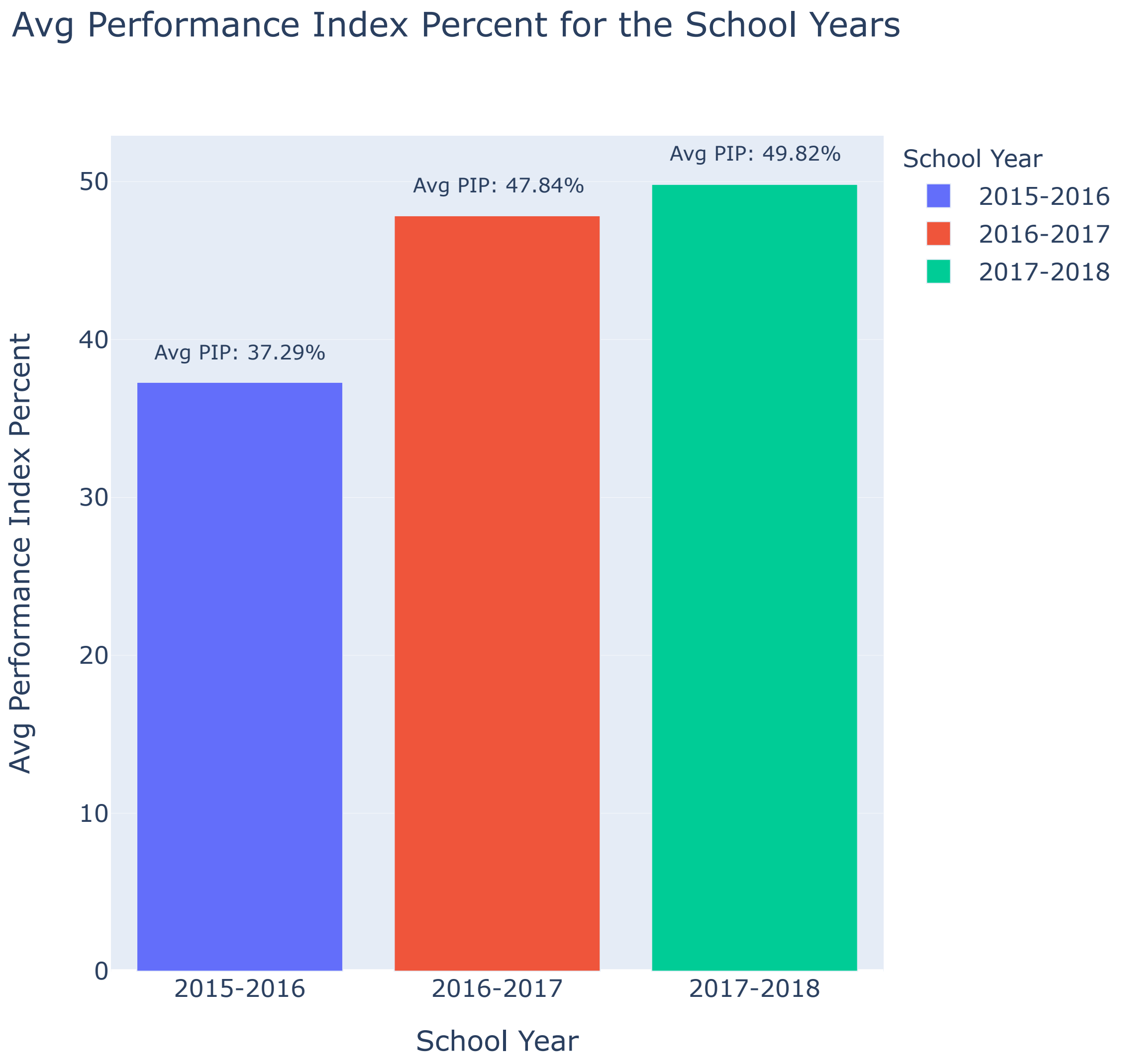
In [ ]:

Data visualization for Performance Index Percent

Interactive Plot thats shows Average Percentage of performance index for each year

In [141]:

```
# Filter data for the year 2015-2016
Mean = df_CMSD.groupby('School_Year')['Performance_Index_Percent'].mean().reset_index()

# Plot bar plot for avg PIP over the years
fig = px.bar(
    Mean,
    x ='School_Year',
    y ='Performance_Index_Percent', color = 'School_Year',
    title='Avg Performance Index Percent for the School Years',
    labels={'Performance_Index_Percent': 'Avg Performance Index Percent', 'School_Year': 'School Year'},
    height=600,
    width=600,)

# Adding mean values
for index, row in Mean.iterrows():
    fig.add_annotation(
        x=row['School_Year'],
        y=row['Performance_Index_Percent']+2,
        text=f'Avg PIP: {row["Performance_Index_Percent"]:.2f}%',
        showarrow=False,
        font=dict(size=10),)

fig.show()
```


Avg Performance Index Percent for the School Years
Avg PIP: 37.29%
Avg PIP: 47.84%
Avg PIP: 49.82%
School Year
2015-2016
2016-2017
2017-2018
Avg Performance Index Percent
0
10
20
30
40
50
2015-2016
2016-2017
2017-2018
School Year

Ploting the graph of how Average of Performance Index Percentage is changed over the school year 2015-2016 in Percentage

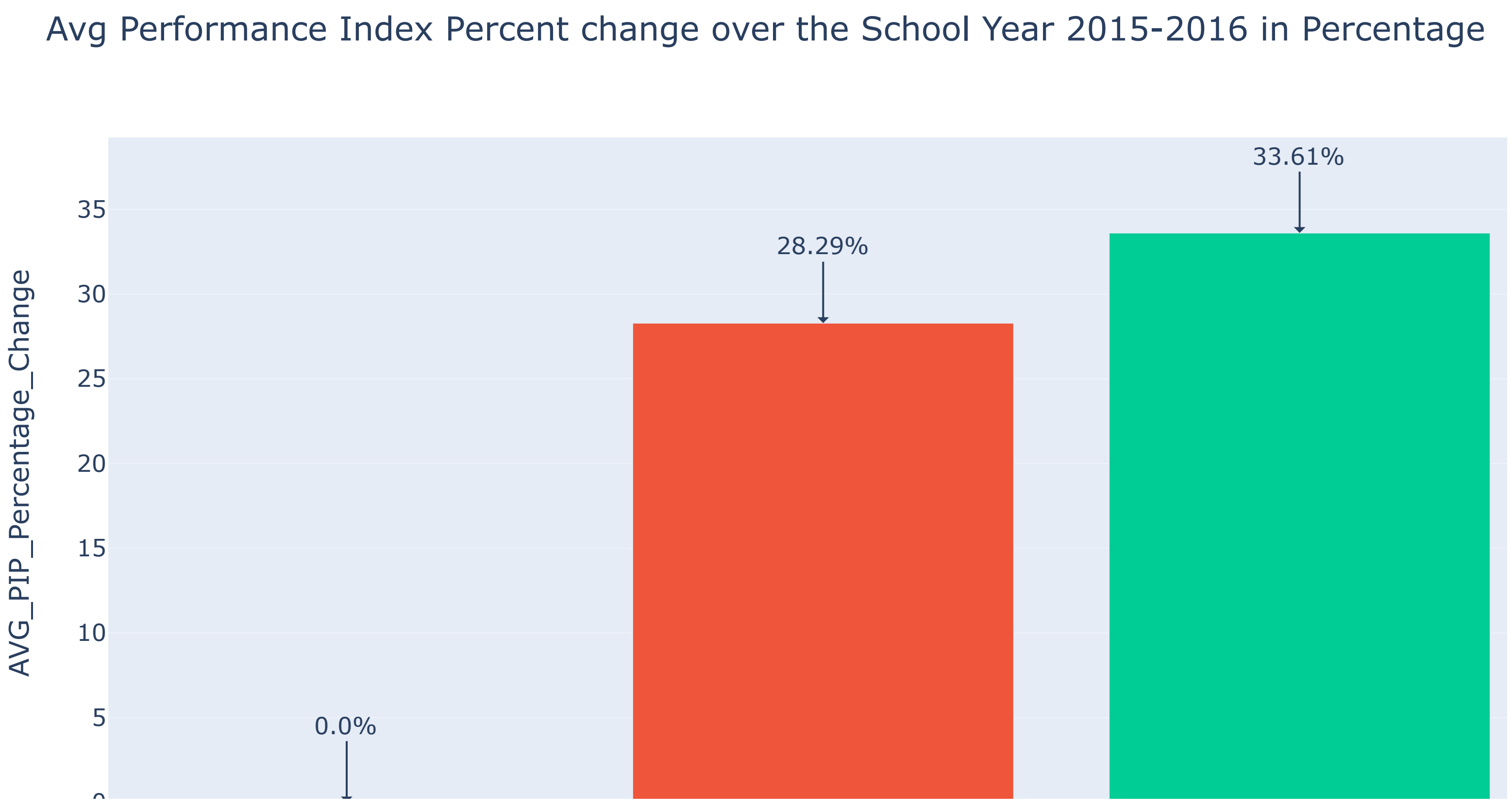
In [113]:

```
# Calculating the sum of PIP for each year
total_by_year_PIP = df_CMSD.groupby('School_Year')['Performance_Index_Percent'].mean().reset_index()

# Filter data for the year 2015-2016 and later
filtered_data_PIP = total_by_year_PIP[total_by_year_PIP['School_Year'] >= '2015-2016']

# Calculate the difference in enrollment relative to 2015-2016
filtered_data_PIP['Avg_PIP_Difference'] = filtered_data_PIP['Performance_Index_Percent'] - filtered_data_

# Calculating the percentage change per year
filtered_data_PIP['AVG_PIP_Percentage_Change'] = (filtered_data_PIP['Avg_PIP_Difference'] / 37.288034) *

print(filtered_data_PIP)
```

```
  School_Year  Performance_Index_Percent  Avg_PIP_Difference  \
0   2015-2016                  37.288034            0.000000
1   2016-2017                  47.835294           10.547260
2   2017-2018                  49.821138           12.533104

   AVG_PIP_Percentage_Change
0                   0.000000
1                  28.285911
2                  33.611598
```

In [143]:

```
# Plot the percentage change in PIP
fig = px.bar(filtered_data_PIP,
              x='School_Year',
              y='AVG_PIP_Percentage_Change',
              color='School_Year',
              title='Percentage Change in PIP per Year',
              labels={'AVG_PIP_Percentage_Change': 'PIP Percentage Change', 'School_Year': 'School Year'}

fig.update_layout(
    title='Avg Performance Index Percent change over the School Year 2015-2016 in Percentage',
    xaxis_title='School Year',
    yaxis_title='AVG_PIP_Percentage_Change'
)

# Rounding off the 'PIP_Percentage_Change' values to 2 decimal places
filtered_data_PIP['AVG_PIP_Percentage_Change'] = round(filtered_data_PIP['AVG_PIP_Percentage_Change'], 2)

# Annotating the plot
for i, row in filtered_data_PIP.iterrows():
    fig.add_annotation(
        x=row['School_Year'],
        y=row['AVG_PIP_Percentage_Change'],
        text=f"{row['AVG_PIP_Percentage_Change']}%",
        showarrow=True,
        arrowhead=1,
        ax=0,
        ay=-40
    )

fig.show()
```


Avg Performance Index Percent change over the School Year 2015-2016 in Percentage
AVG_PIP_Percentage_Change
35
30
25
20
15
10
5
0
0.0%
28.29%
33.61%

Calculating number of schools & buildings name which has increased Performance Index Percent over 3 consecutive years

In [116]:

```
# Create a new column for increased PIP
df_CMSD['Increased_PIP_3_Years'] = False

# For loop to Iterate the data and check for three consecutive increases per school
for school_name in df_CMSD['Building_Name'].unique():
    school_df = df_CMSD[df_CMSD['Building_Name'] == school_name]

    # Check if there are enough data points for the school
    if len(school_df) >= 3:
        mask = (
            (school_df['Performance_Index_Percent'].shift(2) < school_df['Performance_Index_Percent'].shi
            (school_df['Performance_Index_Percent'].shift(1) < school_df['Performance_Index_Percent']) &
            school_df['Performance_Index_Percent'].notna()
        )

        df_CMSD.loc[school_df.index[mask], 'Increased_PIP_3_Years'] = True

# Filter schools with increased enrollment for three consecutive years
schools_with_increased_PIP_3_years = df_CMSD[df_CMSD['Increased_PIP_3_Years']]
print("Schools with increased Performance index pecent for three consecutive years:")
print(schools_with_increased_PIP_3_years[['Building_IRN','Building_Name']])

# Calculating the sum of schools
schools_with_increased_PIP_3_years_sum = df_CMSD['Increased_PIP_3_Years'].sum()
print(f"Number of schools with increased PIP: {schools_with_increased_PIP_3_years_sum}")
```

```
<ipython-input-116-5703869049ee>:2: 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.h
tml#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.htm
l#returning-a-view-versus-a-copy)
```

```
Schools with increased Performance index pecent for three consecutive years:
     Building_IRN                                  Building_Name
234           318                             Menlo Park Academy
235           489                                         Almira
236           729                               Andrew J Rickoff
237           828                                   Anton Grdina
238           930  Cleveland Entrepreneurship Preparatory School
240          1040                                   Artemus Ward
244          5066                                           Case
245          5637                                 Alfred Benesch
246          5892                        Charles A Mooney School
249          6353                        Clara E Westropp School
251          6940                         Collinwood High School
252          8060                         Daniel E Morgan School
253          8383                                        Denison
255          8680                                Memorial School
256          8987                                     East Clark
257          9285                              Douglas MacArthur
259          9555                     East Technical High School
260         10200                          MC^2 STEM High School
263         11291                     Village Preparatory School
269         12350                    Campus International School
271         12353                    New Technology HS@East Tech
272         12355      Facing History High School@Charles Mooney
275         12898                     Garfield Elementary School
278         13292                       George Washington Carver
279         13680                          Glenville High School
282         14918     Cleveland High School for the Digital Arts
283         14919                                     PACT @ JFK
285         15039                                 E3agle Academy
286         15073    Hannah Gibbons-Nottingham Elementary School
303         17830                  James Ford Rhodes High School
306         18408                   Cleveland Early College High
308         21527                           Louis Agassiz School
309         21543                          Franklin D. Roosevelt
314         24687                                   Miles School
315         24695                              Miles Park School
316         24703                               Michael R. White
317         25650                        Mound Elementary School
319         26443                             Nathan Hale School
322         28720                                 Orchard School
323         29371                           Patrick Henry School
324         29413                Paul L Dunbar Elementary School
```

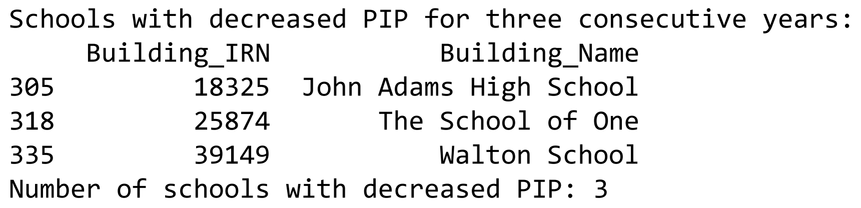
```
325         31963                               Riverside School
328         33902                                Scranton School
329         36475                                        Sunbeam
332         38182                  Valley View Elementary School
333         38604             William C Bryant Elementary School
340         41517                                  Willow School
343         62323                           Whitney Young School
344         62760                        Luis Munoz Marin School
345         62778                      Joseph M Gallagher School
346         63461          Garrett Morgan Schl Of Science School
350         68221                              Kenneth W Clement
355        147397         Cleveland School of Science & Medicine
356         12852                          Citizens Academy East
358        133520                               Citizens Academy
Number of schools with increased PIP: 55
```

Calculating number of schools & buildings name which has decreased Performance Index Percent over 3 consecutive years

In [117]:

```
# Create a new column to with decreased PIP for three consecutive years
df_CMSD['Decreased_PIP_3_Years'] = False

# Iterate through the data and check for three consecutive decreases per school
for school_name in df_CMSD['Building_Name'].unique():
    school_df = df_CMSD[df_CMSD['Building_Name'] == school_name]

    # Check if there are enough data points for the school
    if len(school_df) >= 3:
        mask = (
                (school_df['Performance_Index_Percent'].shift(2) > school_df['Performance_Index_Percent']
                (school_df['Performance_Index_Percent'].shift(1) > school_df['Performance_Index_Percent']
                school_df['Performance_Index_Percent'].notna()
               )

        df_CMSD.loc[school_df.index[mask], 'Decreased_PIP_3_Years'] = True

# Filter schools with decreased enrollment for three consecutive years
schools_with_decreased_PIP_3_years = df_CMSD[df_CMSD['Decreased_PIP_3_Years']]
print("Schools with decreased PIP for three consecutive years:")
print(schools_with_decreased_PIP_3_years[['Building_IRN','Building_Name']])

# Calculating the sum of schools
schools_with_decreased_PIP_3_years_sum = df_CMSD['Decreased_PIP_3_Years'].sum()
print(f"Number of schools with decreased PIP: {schools_with_decreased_PIP_3_years_sum}")
```

```
<ipython-input-117-e96ea351e971>:2: 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.h
tml#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.htm
l#returning-a-view-versus-a-copy)
```

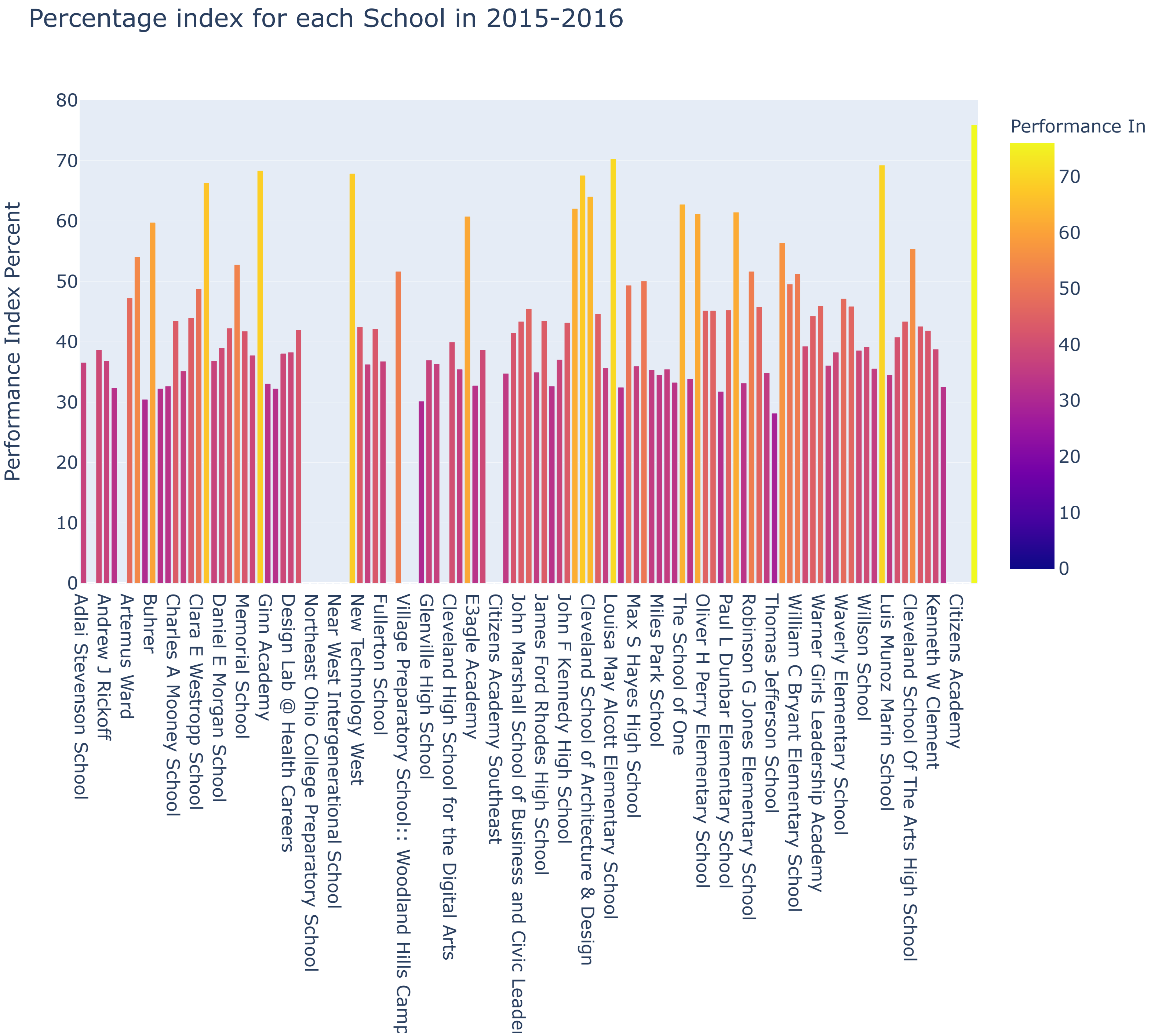


```
Schools with decreased PIP for three consecutive years:
     Building_IRN           Building_Name
305         18325  John Adams High School
318         25874       The School of One
335         39149           Walton School
Number of schools with decreased PIP: 3
```

Interactive bar plot for each years percentage index Percent distribution for each school

In [144]:

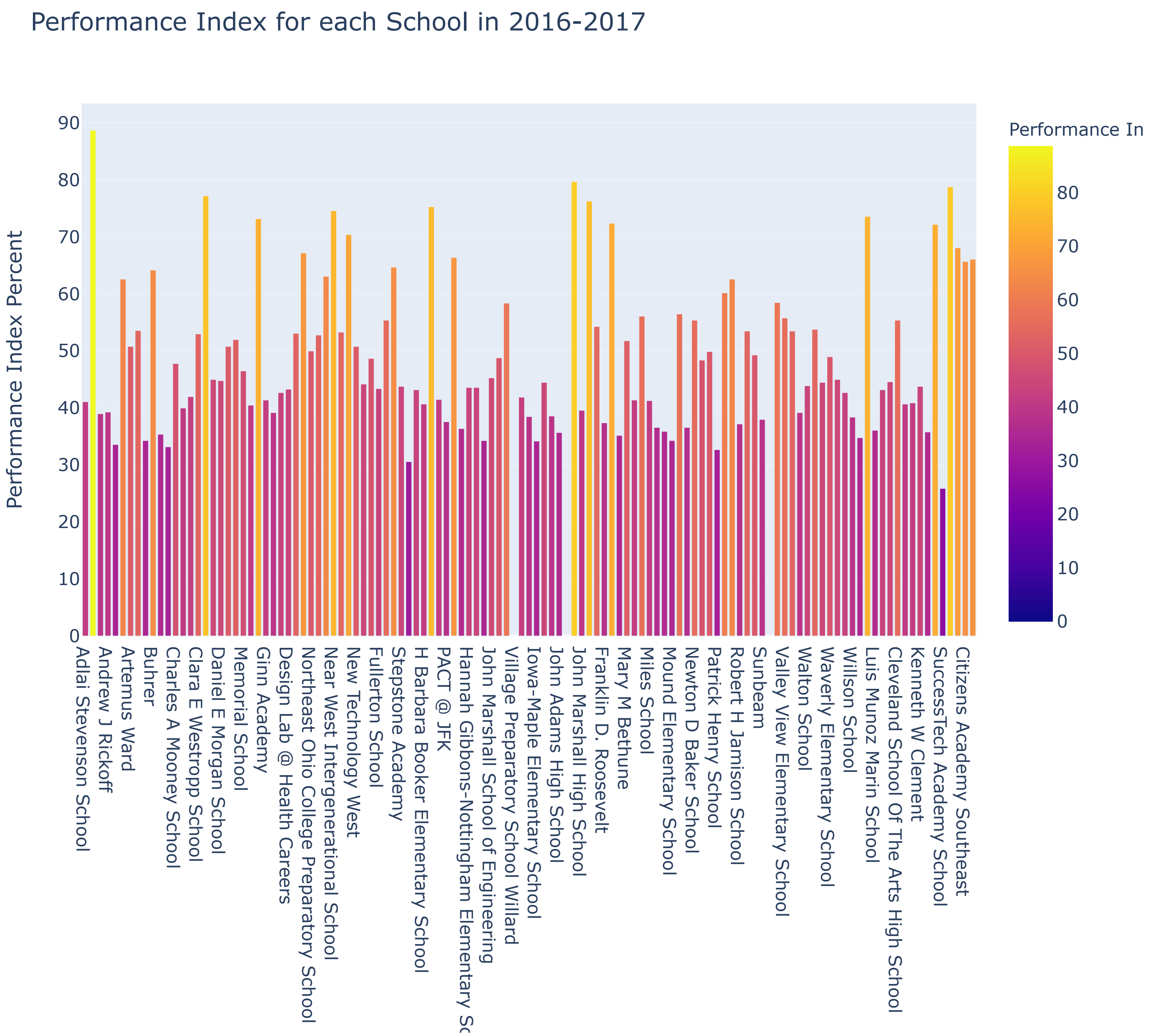
```
# Filter data for the year 2015-2016
df_2015_2016 = df_CMSD[df_CMSD['School_Year'] == '2015-2016']

# Plot with Plotly Express
fig = px.bar(
    df_2015_2016,
    x ='Building_Name',
    y ='Performance_Index_Percent',
    color='Performance_Index_Percent',
    title='Percentage index for each School in 2015-2016',
    labels={'Performance_Index_Percent': 'Performance Index Percent', 'School_Name': 'School Name'},
    height=800,  # Set the height of the figure
    width=900,
)

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


Percentage index for each School in 2015-2016
Performance Index Percent
80
70
60
50
40
30
20
10
0
Adlai Stevenson School
Andrew J Rickoff
Artemus Ward
Buhrer
Charles A Mooney School
Clara E Westropp School
Daniel E Morgan School
Memorial School
Ginn Academy
Design Lab @ Health Careers
Northeast Ohio College Preparatory School
Near West Intergenerational School
New Technology West
Fullerton School
Village Preparatory School:: Woodland Hills Camp
Glenville High School
Cleveland High School for the Digital Arts
E3agle Academy
Citizens Academy Southeast
John Marshall School of Business and Civic Leade
James Ford Rhodes High School
John F Kennedy High School
Cleveland School of Architecture & Design
Louisa May Alcott Elementary School
Max S Hayes High School
Miles Park School
The School of One
Oliver H Perry Elementary School
Paul L Dunbar Elementary School
Robinson G Jones Elementary School
Thomas Jefferson School
William C Bryant Elementary School
Warner Girls Leadership Academy
Waverly Elementary School
Willson School
Luis Munoz Marin School
Cleveland School Of The Arts High School
Kenneth W Clement
Citizens Academy
Performance In
70
60
50
40
30
20
10
0

us
rship
Building_Name

In [145]:

```
# Filter data for the year 2015-2016
df_2016_2017 = df_CMSD[df_CMSD['School_Year'] == '2016-2017']

# Plot with Plotly Express
fig = px.bar(
    df_2016_2017,
    x ='Building_Name',
    y ='Performance_Index_Percent',
    color='Performance_Index_Percent',
    title='Performance Index for each School in 2016-2017',
    labels={'Performance_Index_Percent': 'Performance Index Percent', 'School_Name': 'School Name'},
    height=800,  # Set the height of the figure
    width=900,
)

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

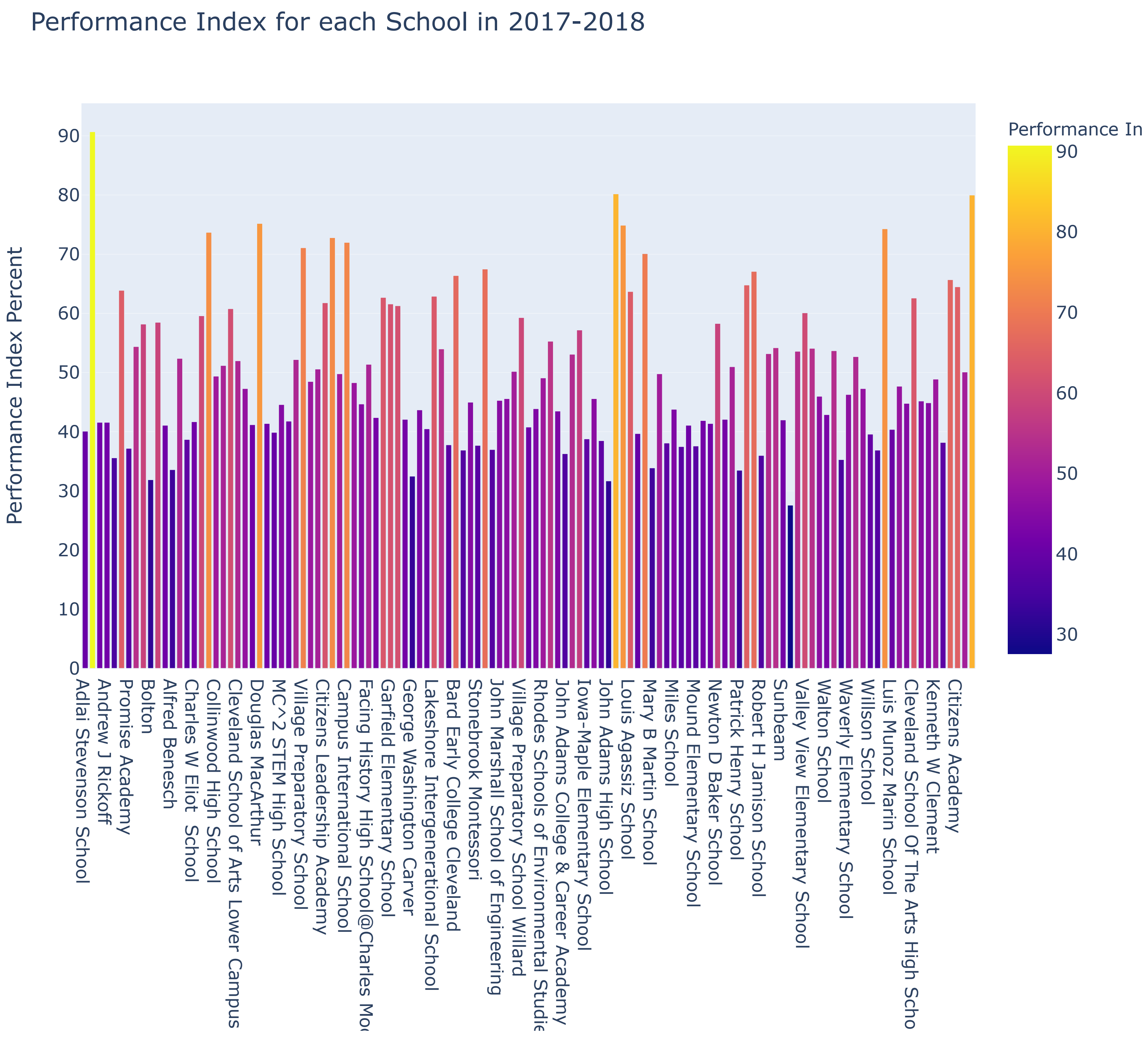


Performance Index for each School in 2016-2017 chool

Building_Name

In [146]:

```
# Filter data for the year 2015-2016
df_2017_2018 = df_CMSD[df_CMSD['School_Year'] == '2017-2018']

# Plot with Plotly Express
fig = px.bar(
    df_2017_2018,
    x ='Building_Name',
    y ='Performance_Index_Percent',
    color='Performance_Index_Percent',
    title='Performance Index for each School in 2017-2018',
    labels={'Performance_Index_Percent': 'Performance Index Percent', 'School_Name': 'School Name'},
    height=800,  # Set the height of the figure
    width=900,
)

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


Performance Index for each School in 2017-2018
Performance Index Percent
0
10
20
30
40
50
60
70
80
90
Performance In
30
40
50
60
70
80
90
Adlai Stevenson School
Andrew J Rickoff
Promise Academy
Bolton
Alfred Benesch
Charles W Eliot School
Collinwood High School
Cleveland School of Arts Lower Campus
Douglas MacArthur
MC^2 STEM High School
Village Preparatory School
Citizens Leadership Academy
Campus International School
Facing History High School@Charles Mo
Garfield Elementary School
George Washington Carver
Lakeshore Intergenerational School
Bard Early College Cleveland
Stonebrook Montessori
John Marshall School of Engineering
Village Preparatory School Willard
Rhodes Schools of Environmental Studie
John Adams College & Career Academy
Iowa-Maple Elementary School
John Adams High School
Louis Agassiz School
Mary B Martin School
Miles School
Mound Elementary School
Newton D Baker School
Patrick Henry School
Robert H Jamison School
Sunbeam
Valley View Elementary School
Walton School
Waverly Elementary School
Willson School
Luis Munoz Marin School
Cleveland School Of The Arts High Scho
Kenneth W Clement
Citizens Academy

In [ ]:

In [ ]:

In [ ]: