Project Step 3: Data Cleaning

ABD Project Team 7:

Anagha Nitin Navale

Harsh Ketan Khona

Jayesh Ramesh Borkar

Stuti Pandey

Information Technology, Arizona State University

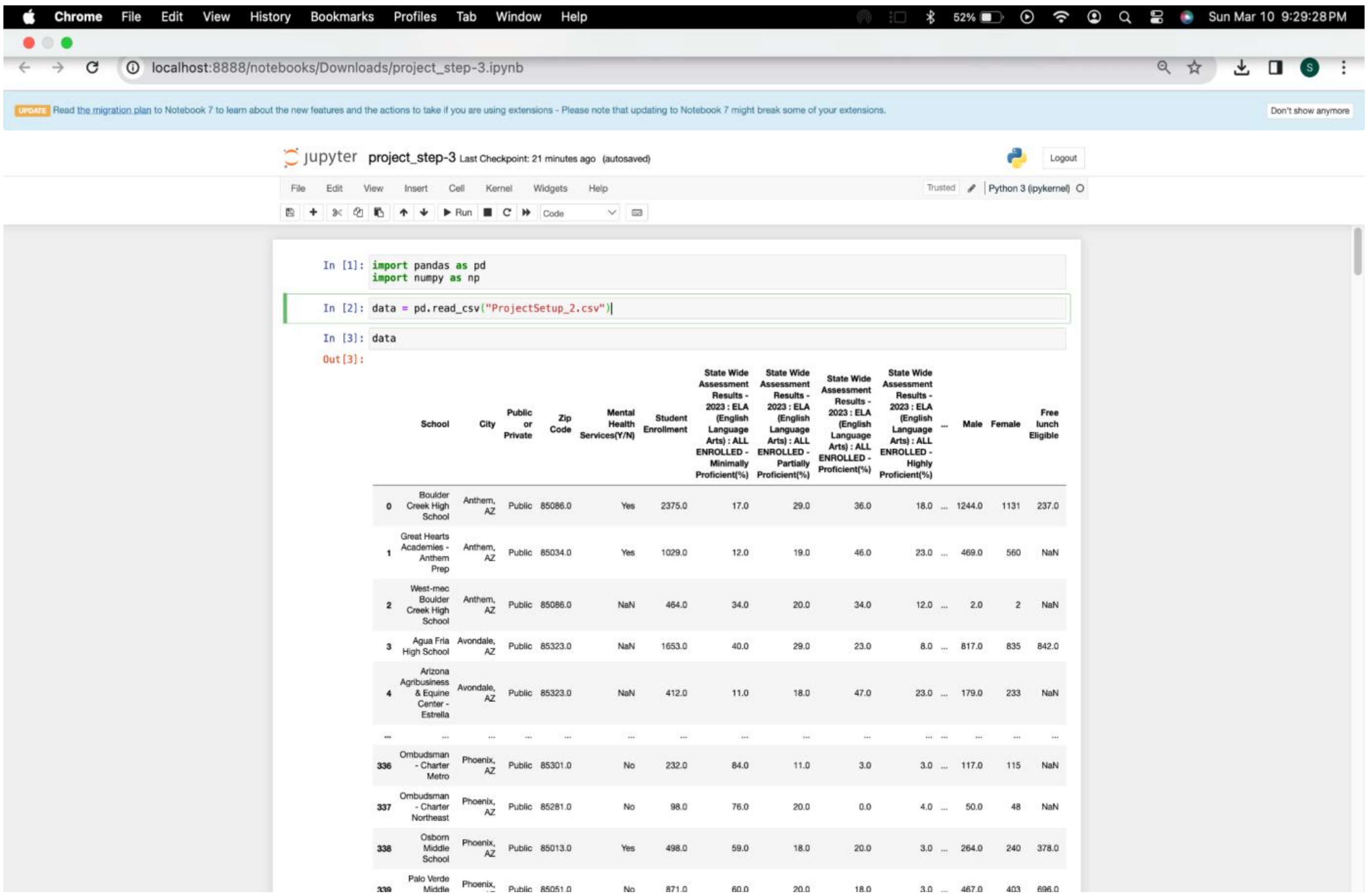
IFT 511: Analyzing Big Data

Professor: Asmaa Elbadrawy

March 10, 2024

1. Data cleaning steps:

1) The provided Python code uses the pandas library to perform comprehensive data cleaning and transformation on a dataset read from a CSV file named "ProjectSetup_2.csv."



2) The initial exploration of the data includes checking its shape, information, and descriptive statistics. Subsequently, missing values are addressed by first identifying columns with more than 50% null values, which are then dropped to retain valuable data points. Rows with any remaining null values are removed as well.

Chrome File Edit View History Bookmarks Profiles Tab Window Help 52% Sun Mar 10 9:30:17 PM

localhost:8888/notebooks/Downloads/project_step-3.ipynb

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter project_step-3 Last Checkpoint: 22 minutes ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

In [7]:
```
#Checking Columns and their with missing values
data.isnull().sum()
```

Out[7]:
```
School                                                                                       0
City                                                                                         0
Public or Private                                                                            0
Zip Code                                                                                     5
Mental Health Services(Y/N)                                                                242
Student Enrollment                                                                          56
State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Minimally Proficient(%)    135
State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Partially Proficient(%)    135
State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Proficient(%)              135
State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Highly Proficient(%)       135
State Wide Assessment Results - 2023 All:MATH : ALL ENROLLED - Minimally Proficient(%)                         135
State Wide Assessment Results - 2023 All:MATH : ALL ENROLLED - Partially Proficient(%)                         135
State Wide Assessment Results - 2023 All:MATH : ALL ENROLLED - Proficient(%)                                   135
State Wide Assessment Results - 2023 All:MATH : ALL ENROLLED - Highly Proficient(%)                            135
Graduation Rate: 2023 (Cohort 2022) (%)                                                    174
Dropout Rate: 2023 (%)                                                                     171
End of Year Promotion (%)                                                                  133
College or Career Readiness (CCRI) Points                                                  181
American Indian/ Alaska Native                                                              42
Asian                                                                                       42
Black                                                                                       42
Hispanic                                                                                    42
Native Hawaiian/ Pacific Islander                                                           42
White                                                                                       42
Two or More Races                                                                           42
Male                                                                                        64
Female                                                                                      63
Free lunch Eligible                                                                        203
Reduced-price Lunch eligible                                                               203
Free lunch eligible by Direct Certification                                                209
Classroom Teachers (FTE)                                                                   205
 Student Teacher Ratio                                                                     192
Grade Levels                                                                                51
County                                                                                      23
Phoenix, AZ Metro Area High Schools Ranking - US News                                      221
dtype: int64
```

In [8]:
```
#Checking rows with null values and showing their specific counts
#Number of Null Values within that row - Number of rows corresponding to that
data.isnull().sum(axis=1).value_counts().sort_index()
```

Out[8]:
```
0    25
1    37
2     7
3    23
4    13
5    40
6    34
7    17
```

Chrome File Edit View History Bookmarks Profiles Tab Window Help 52% Sun Mar 10 9:30:34 PM

localhost:8888/notebooks/Downloads/project_step-3.ipynb

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter project_step-3 Last Checkpoint: 22 minutes ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Since we only have 14 data points with non-null values, we would drop columns initially as you can see from above, many rows have just 1-6 null values within them which would become non-null when we remove columns first so we would be able to retain that data point by dropping columns first

In [9]:
```
col_null  = data.isnull().sum()
col_null
```

Out[9]:
```
School                                                                                       0
City                                                                                         0
Public or Private                                                                            0
Zip Code                                                                                     5
Mental Health Services(Y/N)                                                                242
Student Enrollment                                                                          56
State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Minimally Proficient(%)    135
State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Partially Proficient(%)    135
State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Proficient(%)              135
State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Highly Proficient(%)       135
State Wide Assessment Results - 2023 All:MATH : ALL ENROLLED - Minimally Proficient(%)                         135
State Wide Assessment Results - 2023 All:MATH : ALL ENROLLED - Partially Proficient(%)                         135
State Wide Assessment Results - 2023 All:MATH : ALL ENROLLED - Proficient(%)                                   135
State Wide Assessment Results - 2023 All:MATH : ALL ENROLLED - Highly Proficient(%)                            135
Graduation Rate: 2023 (Cohort 2022) (%)                                                    174
Dropout Rate: 2023 (%)                                                                     171
End of Year Promotion (%)                                                                  133
College or Career Readiness (CCRI) Points                                                  181
American Indian/ Alaska Native                                                              42
Asian                                                                                       42
Black                                                                                       42
Hispanic                                                                                    42
Native Hawaiian/ Pacific Islander                                                           42
White                                                                                       42
Two or More Races                                                                           42
Male                                                                                        64
Female                                                                                      63
Free lunch Eligible                                                                        203
Reduced-price Lunch eligible                                                               203
Free lunch eligible by Direct Certification                                                209
Classroom Teachers (FTE)                                                                   205
 Student Teacher Ratio                                                                     192
Grade Levels                                                                                51
County                                                                                      23
Phoenix, AZ Metro Area High Schools Ranking - US News                                      221
dtype: int64
```

In [10]:
```
col_list = col_null[col_null > ((50/100)*data.shape[0])].index.to_list()
#Dropping those columns where the number of null values is greater than
col_list
```

Out[10]:
```
['Mental Health Services(Y/N)',
 'Graduation Rate: 2023 (Cohort 2022) (%)',
 'Dropout Rate: 2023 (%)',
```

Chrome File Edit View History Bookmarks Profiles Tab Window Help 52% Sun Mar 10 9:30:53 PM

localhost:8888/notebooks/Downloads/project_step-3.ipynb

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter project_step-3 Last Checkpoint: 22 minutes ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

In [10]:
```
col_list = col_null[col_null > ((50/100)*data.shape[0])].index.to_list()
#Dropping those columns where the number of null values is greater than
col_list
```

Out[10]:
```
['Mental Health Services(Y/N)',
 'Graduation Rate: 2023 (Cohort 2022) (%)',
 'Dropout Rate: 2023 (%)',
 'College or Career Readiness (CCRI) Points',
 'Free lunch Eligible',
 'Reduced-price Lunch eligible',
 'Free lunch eligible by Direct Certification',
 'Classroom Teachers (FTE)',
 ' Student Teacher Ratio',
 'Phoenix, AZ Metro Area High Schools Ranking – US News']
```

These above columns don't add much value to our data as majority of their values are null

In [11]:
```
data = data.drop(columns=col_list, axis=1)
data
```

Out[11]:

	School	City	Public or Private	Zip Code	Student Enrollment	State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Minimally Proficient(%)	State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Partially Proficient(%)	State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Proficient(%)	State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Highly Proficient(%)	State Wide Assessment Results - 2023 All:MATH : ALL ENROLLED - Minimally Proficient(%)	...	Asian	Black	Hispanic	H
0	Boulder Creek High School	Anthem, AZ	Public	85086.0	2375.0	17.0	29.0	36.0	18.0	26.0	...	37.0	38.0	347.0	
1	Great Hearts Academies - Anthem Prep	Anthem, AZ	Public	85034.0	1029.0	12.0	19.0	46.0	23.0	10.0	...	86.0	16.0	157.0	
2	West-mec Boulder Creek High School	Anthem, AZ	Public	85086.0	464.0	34.0	20.0	34.0	12.0	29.0	...	0.0	0.0	0.0	
3	Agua Fria High School	Avondale, AZ	Public	85323.0	1653.0	40.0	29.0	23.0	8.0	52.0	...	31.0	177.0	1184.0	
4	Arizona Agribusiness & Equine Center - Estrella	Avondale, AZ	Public	85323.0	412.0	11.0	18.0	47.0	23.0	18.0	...	25.0	16.0	209.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	

Chrome File Edit View History Bookmarks Profiles Tab Window Help 49% Sun Mar 10 9:52:44 PM

localhost:8888/notebooks/Downloads/project_step-3.ipynb

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter project_step-3 Last Checkpoint: an hour ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

After dropping those columns, now we are left with 196 data points which are non null

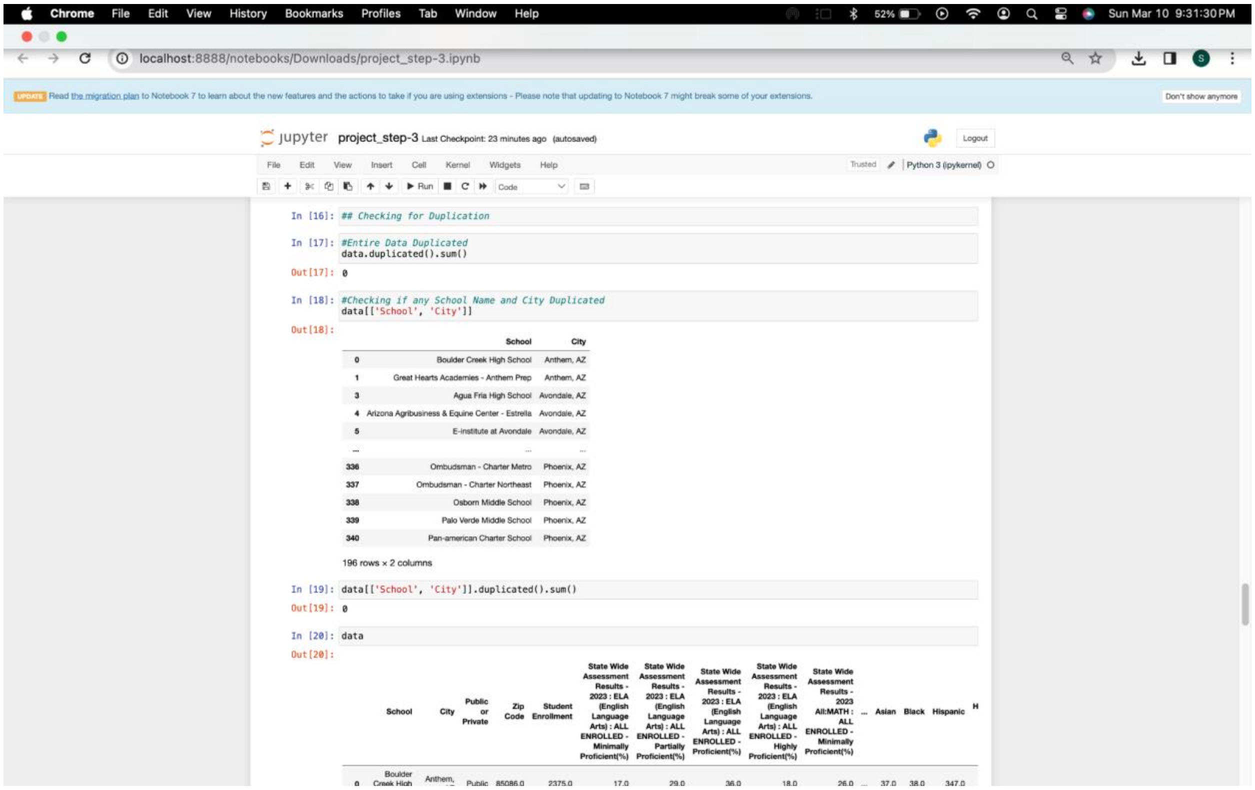
In [13]:
```
#Dropping rows with null values
data.dropna(inplace = True)
```

In [14]:
```
data.isnull().sum()
```

Out[14]:
```
School                                                                                            0
City                                                                                              0
Public or Private                                                                                 0
Zip Code                                                                                          0
Student Enrollment                                                                                0
State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Minimally Proficient(%)   0
State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Partially Proficient(%)   0
State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Proficient(%)   0
State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Highly Proficient(%)   0
State Wide Assessment Results - 2023 All:MATH : ALL ENROLLED - Minimally Proficient(%)            0
State Wide Assessment Results - 2023 All:MATH : ALL ENROLLED - Partially Proficient(%)            0
State Wide Assessment Results - 2023 All:MATH : ALL ENROLLED - Proficient(%)                      0
State Wide Assessment Results - 2023 All:MATH : ALL ENROLLED - Highly Proficient(%)               0
End of Year Promotion (%)                                                                         0
American Indian/ Alaska Native                                                                    0
Asian                                                                                             0
Black                                                                                             0
Hispanic                                                                                          0
Native Hawaiian/ Pacific Islander                                                                 0
White                                                                                             0
Two or More Races                                                                                 0
Male                                                                                              0
Female                                                                                            0
Grade Levels                                                                                      0
County                                                                                            0
dtype: int64
```

In [15]:
```
data.isnull().sum(axis=1).value_counts()
```

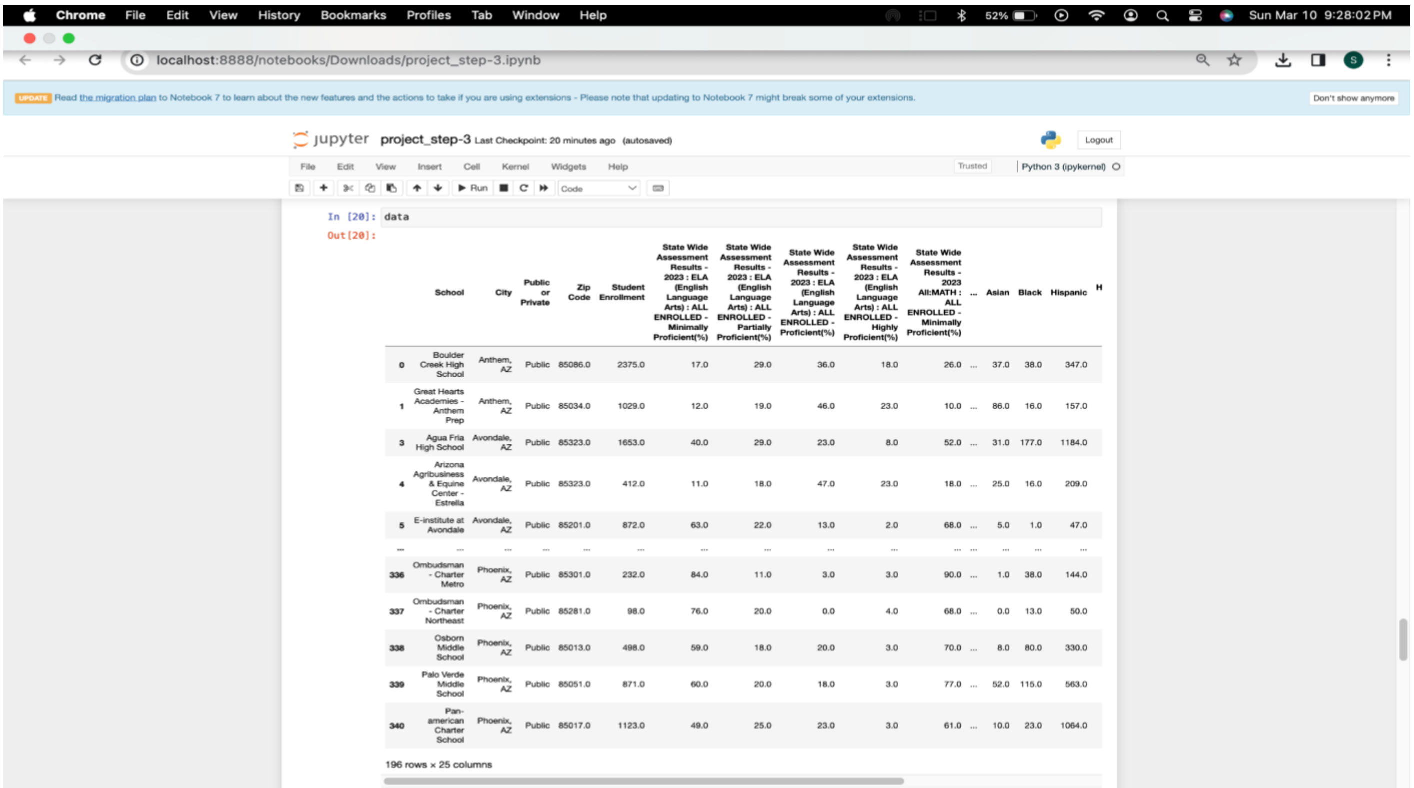
3) Then we proceed further and investigate and handle duplicate records in the dataset, checking for duplicates in both the entire dataset and specific columns (for example - ‘'School’ and ‘City’). Duplicate records are addressed by combining them into a single record if multiple entries exist for a particular school. But in this dataset, as we can see there are no duplicate entries.



4) In the next step, we move on to data transformation. This phase focuses on cleaning and standardizing specific columns. For instance, the 'City' column has trailing ", AZ" removed, and numeric columns like 'Zip Code,' 'Student Enrollment,' and demographic categories are appropriately converted to integers after removing non-numeric characters.

5) The final data-cleaning procedure involves systematically addressing missing values, handling duplicates, and ensuring uniformity in data types. The code's outcome is a cleaned

and transformed dataset ready for further analysis. After performing all the data cleaning steps, we are left with **196 rows * 25 columns** of data. Attaching screenshot below shows data after data cleaning:



	School	City	Public or Private	Zip Code	Student Enrollment	State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Minimally Proficient(%)	State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Partially Proficient(%)	State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Proficient(%)	State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Highly Proficient(%)	State Wide Assessment Results - 2023 All:MATH : ALL ENROLLED - Minimally Proficient(%)	...	Asian	Black	Hispanic	H
0	Boulder Creek High School	Anthem, AZ	Public	85086.0	2375.0	17.0	29.0	36.0	18.0	26.0	...	37.0	38.0	347.0	
1	Great Hearts Academies - Anthem Prep	Anthem, AZ	Public	85034.0	1029.0	12.0	19.0	46.0	23.0	10.0	...	86.0	16.0	157.0	
3	Agua Fria High School	Avondale, AZ	Public	85323.0	1653.0	40.0	29.0	23.0	8.0	52.0	...	31.0	177.0	1184.0	
4	Arizona Agribusiness & Equine Center - Estrella	Avondale, AZ	Public	85323.0	412.0	11.0	18.0	47.0	23.0	18.0	...	25.0	16.0	209.0	
5	E-institute at Avondale	Avondale, AZ	Public	85201.0	872.0	63.0	22.0	13.0	2.0	68.0	...	5.0	1.0	47.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
336	Ombudsman - Charter Metro	Phoenix, AZ	Public	85301.0	232.0	84.0	11.0	3.0	3.0	90.0	...	1.0	38.0	144.0	
337	Ombudsman - Charter Northeast	Phoenix, AZ	Public	85281.0	98.0	76.0	20.0	0.0	4.0	68.0	...	0.0	13.0	50.0	
338	Osborn Middle School	Phoenix, AZ	Public	85013.0	498.0	59.0	18.0	20.0	3.0	70.0	...	8.0	80.0	330.0	
339	Palo Verde Middle School	Phoenix, AZ	Public	85051.0	871.0	60.0	20.0	18.0	3.0	77.0	...	52.0	115.0	563.0	
340	Pan-american Charter School	Phoenix, AZ	Public	85017.0	1123.0	49.0	25.0	23.0	3.0	61.0	...	10.0	23.0	1064.0	

196 rows × 25 columns

2. **Data description and transformation methods used:**

After Cleaning, we have 25 columns. Here's a description of attributes for each column:

Categorical: School, City, Public or Private, Zip Code, County

Ordinal: Grade Levels

All the remaining columns are of Ratio type.

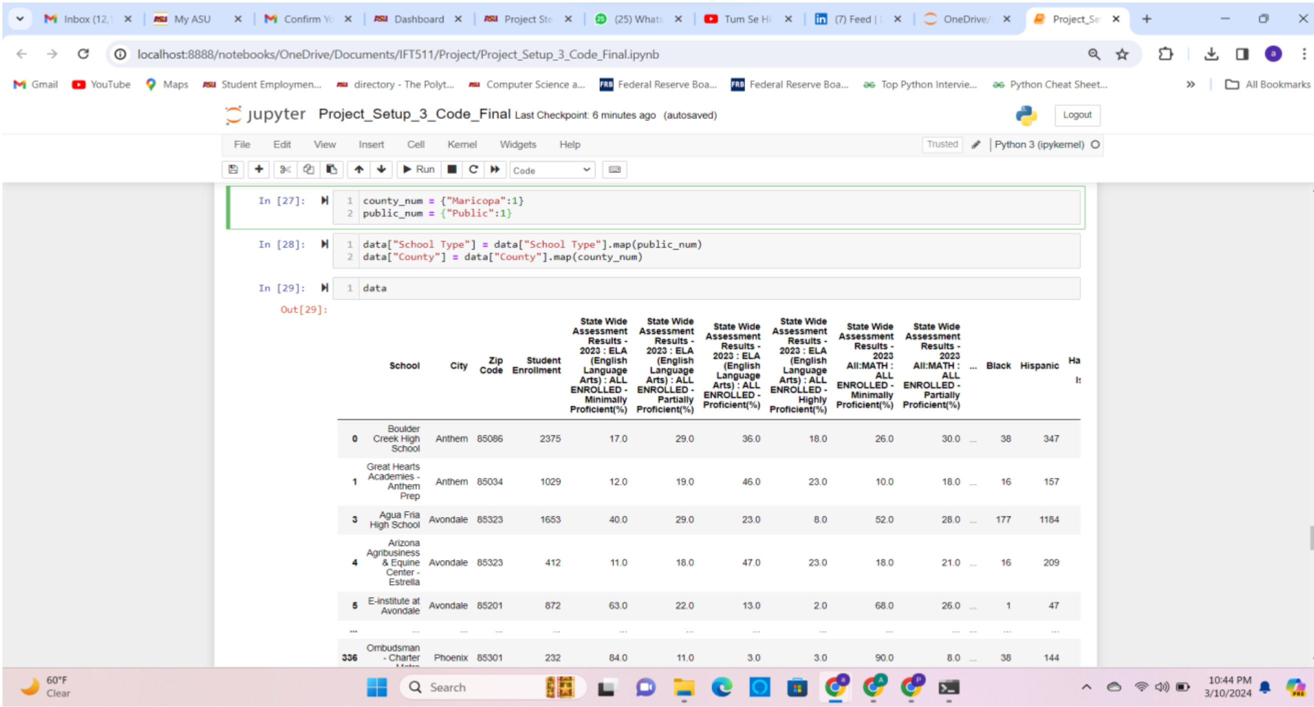
Ratio: Student Enrollment, State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Minimally Proficient(%), State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Partially Proficient(%), State Wide Assessment

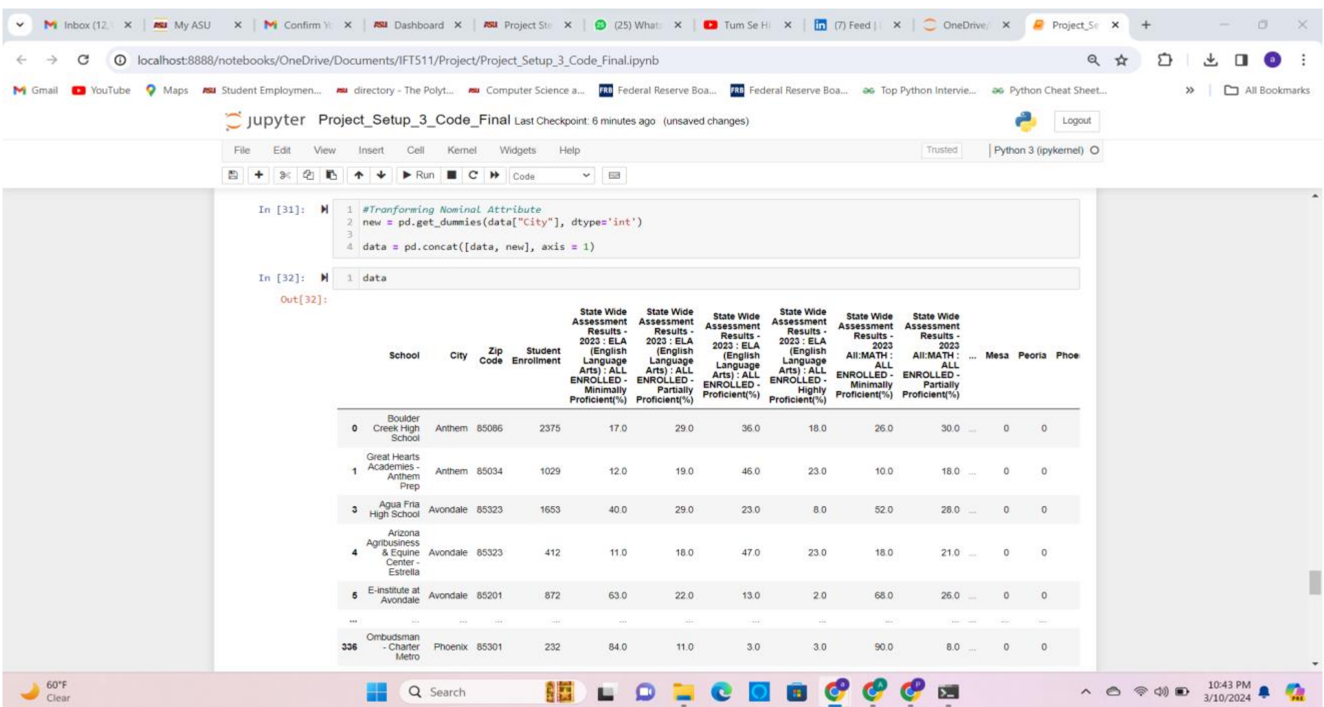
Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Proficient(%), State Wide Assessment Results - 2023 : ELA (English Language Arts) : ALL ENROLLED - Highly Proficient(%), State Wide Assessment Results - 2023 All:MATH : ALL ENROLLED - Minimally Proficient(%), State Wide Assessment Results - 2023, All:MATH : ALL ENROLLED - Partially Proficient(%), State Wide Assessment Results - 2023 All:MATH : ALL ENROLLED - Proficient(%), State Wide Assessment Results - 2023 All:MATH : ALL, ENROLLED - Highly Proficient(%), End of Year Promotion (%), American Indian/ Alaska Native, Asian, Black, Hispanic, Native Hawaiian/ Pacific Islander, White, Two or More Races Races, Male, Female.

We have opted not to undergo any transformations on School and Zip codes as we believe their numerical values are unnecessary for model training at this stage. Should the need arise, any required transformations will be performed in subsequent steps.

Transformation Methods for each type:

1. Ratio attributes do not need any transformation, as it is already in a numerical form and we can perform operations on it easily.
2. Categorical attributes need to be converted using One-Hot Encoding. By creating binary columns for every category and designating if a category exists with a 1 or 0, one-hot encoding is done.





After the transformation, we have 196 rows and 47 columns.

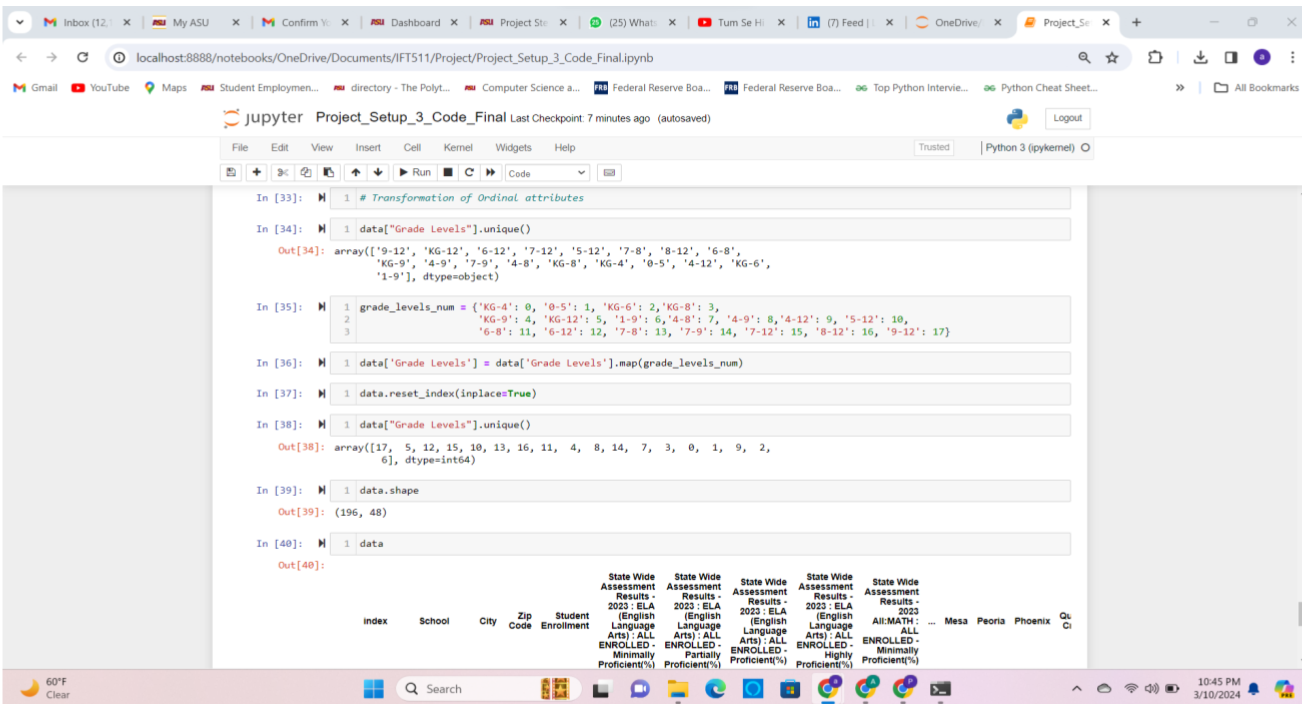
25 previous columns and 22 cities, which we have added.

3. For transforming ordinal attributes, we can do Label Encoding. Label encoding with a customized mapping can be used if the ordinal attribute has a distinct order. Refer following screenshot for mentioning the order for each category explicitly.

```
In [36]:  data["Grade Levels"].unique()

Out[36]: array(['9-12', 'KG-12', '6-12', '7-12', '5-12', '7-8', '8-12', '6-8',
                'KG-9', '4-9', '7-9', '4-8', 'KG-8', 'KG-4', '0-5', '4-12', 'KG-6',
                '1-9'], dtype=object)

In [37]:  grade_levels_num = {'KG-4': 0, '0-5': 1, 'KG-6': 2,'KG-8': 3,
                             'KG-9': 4, 'KG-12': 5, '1-9': 6,'4-8': 7, '4-9': 8,'4-12': 9, '5-12': 10,
                             '6-8': 11, '6-12': 12, '7-8': 13, '7-9': 14, '7-12': 15, '8-12': 16, '9-12': 17}
```



The code for the Cleaning and Transformation:

```
import pandas as pd

import numpy as np

data = pd.read_csv("Project_Setup_Data.csv")

data

data.shape

data.info()

data

data.describe()

#Checking Columns and their with missing values

data.isnull().sum()

#Checking rows with null values and showing their specific counts

#Number of Null Values within that row - Number of rows corresponding to that

data.isnull().sum(axis=1).value_counts().sort_index()

col_null  = data.isnull().sum()

col_null

col_list = col_null[col_null > ((50/100)*data.shape[0])].index.to_list()

#Dropping those columns where the number of null values is greater than

col_list

data = data.drop(columns=col_list, axis=1)

data

data.isnull().sum(axis=1).value_counts().sort_index()#Dropping rows with null values

data.dropna(inplace = True)
```

```
data.isnull().sum()

data.isnull().sum()

data.isnull().sum(axis=1).value_counts()

# Checking whether there are any duplication

data.duplicated().sum()

# Checking if any School Name and City has duplicates

data[['School', 'City']]

data[['School', 'City']].duplicated().sum()

data["School Type"] =  data['Public or Private'].str.replace(" ","")

data.drop(columns=["Public or Private"], axis=1, inplace=True)

print(data['County'].unique())

print(data['School Type'].unique())

print(data['City'].unique())

data['City'] = data['City'].str.replace(", AZ","")

data['Zip Code'] = data['Zip Code'].astype("int").astype("str")

int_con = ['Student Enrollment','American Indian/ Alaska Native', 'Asian', 'Black', 'Hispanic',
'Native Hawaiian/ Pacific Islander', 'White', 'Two or More Races', 'Male', 'Female']

for v_ in int_con:

    data[v_] = data[v_].astype("int")

data.columns

county_num = {"Maricopa":1}

public_num = {"Public":1}

data["School Type"] = data["School Type"].map(public_num)
```

```
data["County"] = data["County"].map(county_num)

data

data["City"].unique()

#Transforming Nominal Attribute

new = pd.get_dummies(data["City"], dtype='int')


data = pd.concat([data, new], axis = 1)

data

data["Grade Levels"].unique()grade_levels_num = {'KG-4': 0, '0-5': 1, 'KG-6': 2,'KG-8': 3,

                 'KG-9': 4, 'KG-12': 5, '1-9': 6,'4-8': 7, '4-9': 8,'4-12': 9, '5-12': 10,

                 '6-8': 11, '6-12': 12, '7-8': 13, '7-9': 14, '7-12': 15, '8-12': 16, '9-12': 17}

data.reset_index(inplace=True)

data["Grade Levels"].unique()

data.shape

data
```