Exploratory Data Analysis with Python or R - Tell Me About a Dataset

Pamulapati Harika (001266981)

Saint Louis University

ORES-5160 Data Management

October 26 2023

Introduction

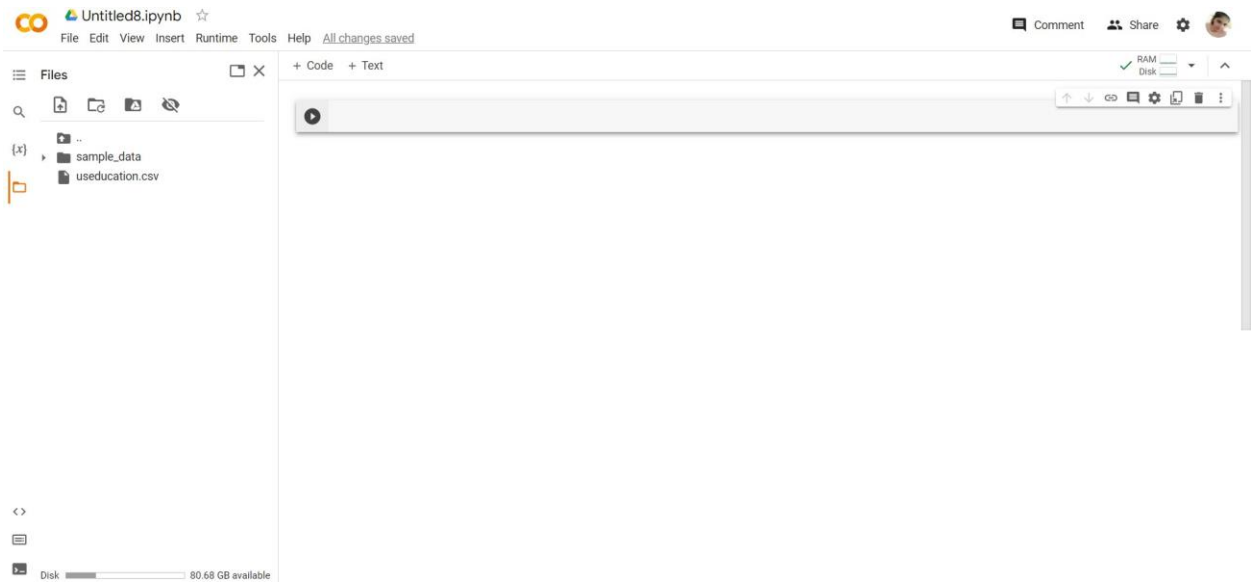
The steps involved in performing an Exploratory Data Analysis (EDA) on the **"US education"** dataset are outlined in this study. The dataset appears to be historical in nature and is organized to offer insights into how financing for education, enrollment, and other factors have changed over time in various U.S. states. In the field of education, this research is frequently utilized for policy analysis, decision-making, and research on education.

About the dataset

The U.S. Education Dataset is a thorough compilation of information that offers insights into many facets of the American educational system across a number of years. Usually, this dataset contains statistics on academic performance, student enrollment, school finance, and other important variables. This dataset is used by researchers, policymakers, educators, and analysts to track student performance, spot patterns and discrepancies in the U.S. education system, and better understand how educational resources are allocated. It is an invaluable tool for assessing the impact of policies and making well-informed choices to raise the standard of education in the nation.

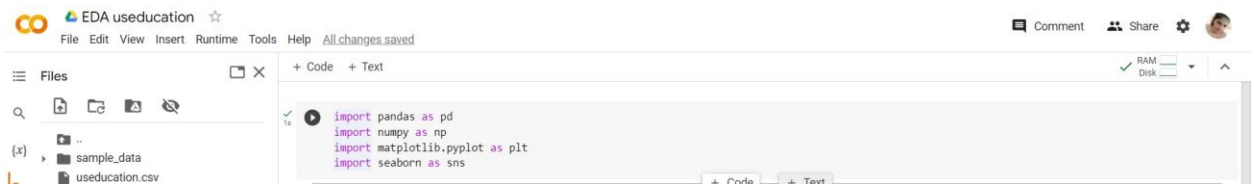
useducation dataset

I uploaded the useducation dataset into the Google colab that we utilised for the group project



2. Loading the dataset

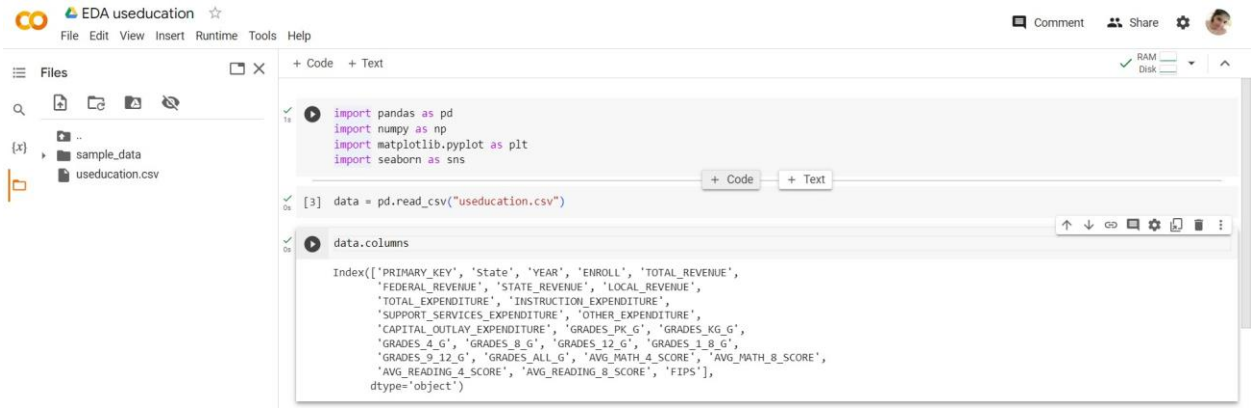
Then I started by importing the libraries that are required for data manipulation and visualization, such as pandas, numpy, matplotlib, and seaborn, in order to start our investigation. These libraries enable us to create intelligent visualizations and operate efficiently with our dataset.



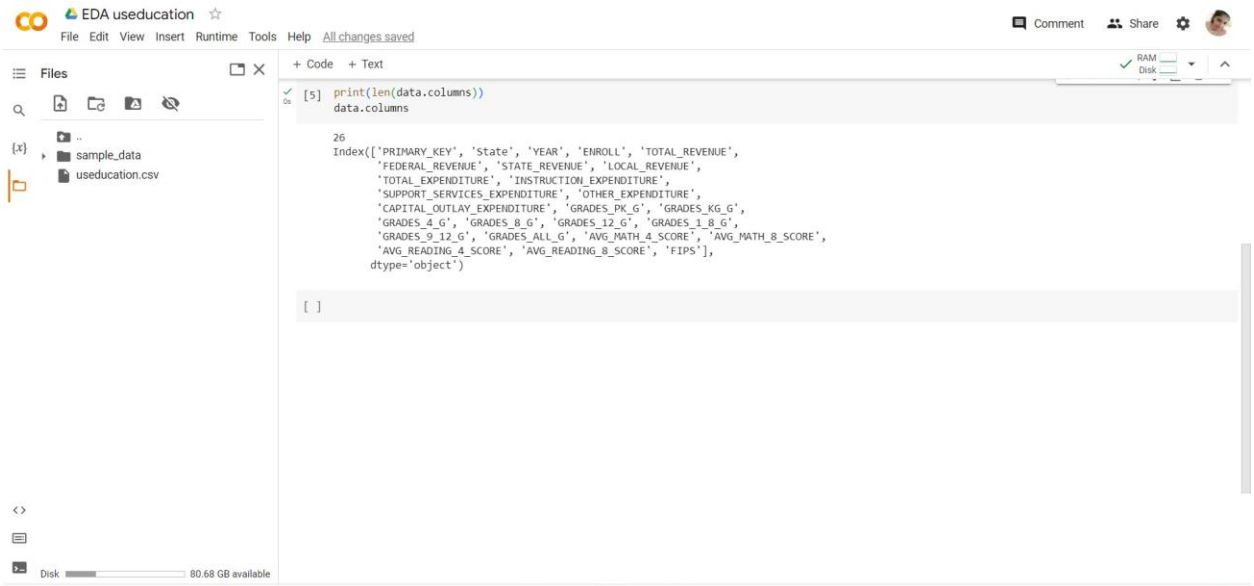
3. Once the file was read into Google Colab, I labeled it "data" and used the code below.

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

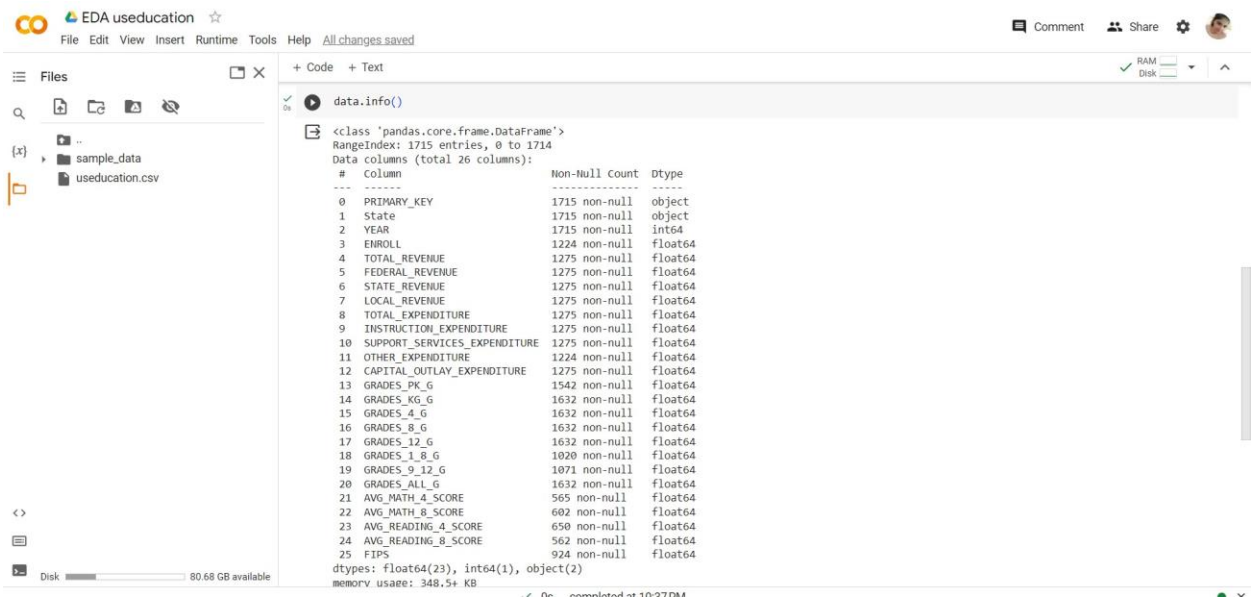
4. I used the code below to extract each column from the provided dataset.



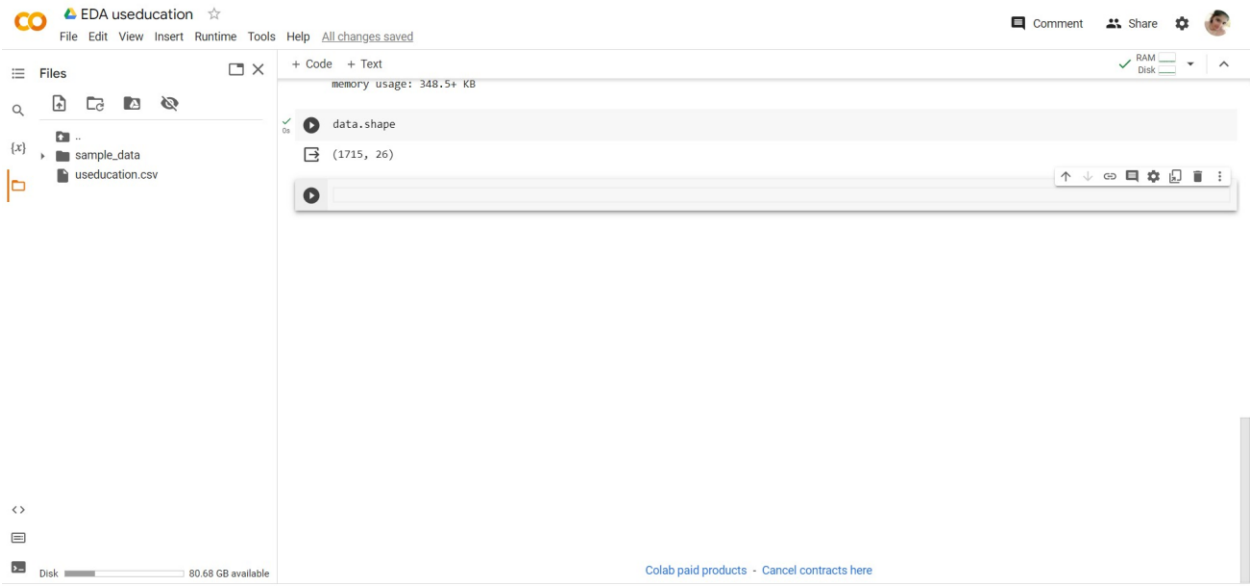
5. The useducation dataset contains total **26** columns and the code is **data.columns** used to get all of them.



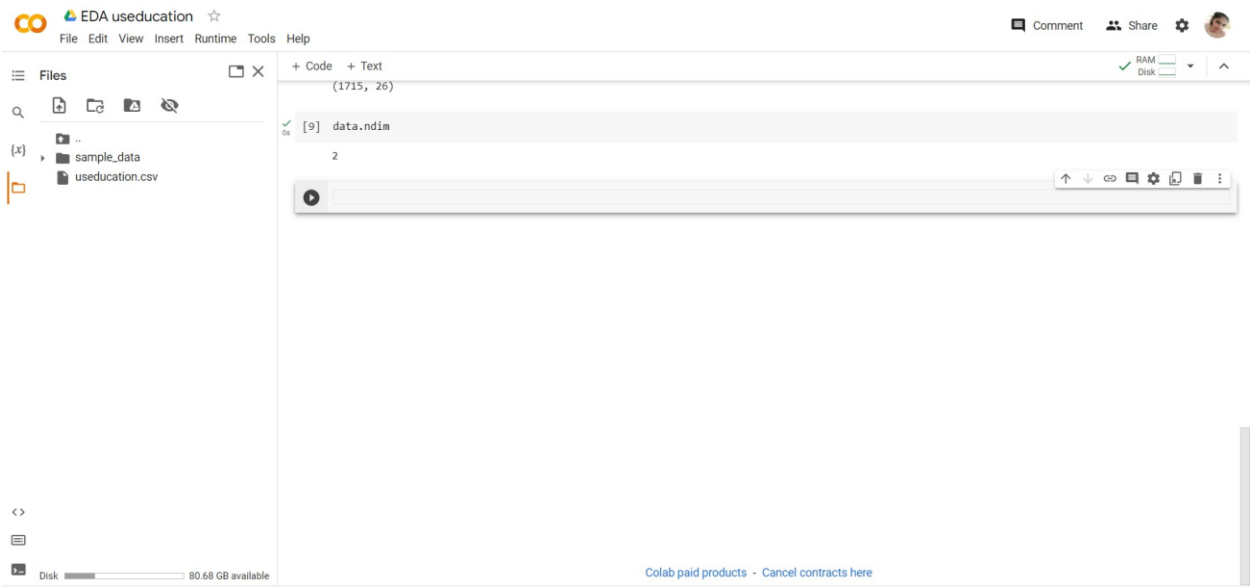
6. I was able to get the data information out of the file by using the code below. The **data.info( )** domain is used to generate general information about the useducation data.



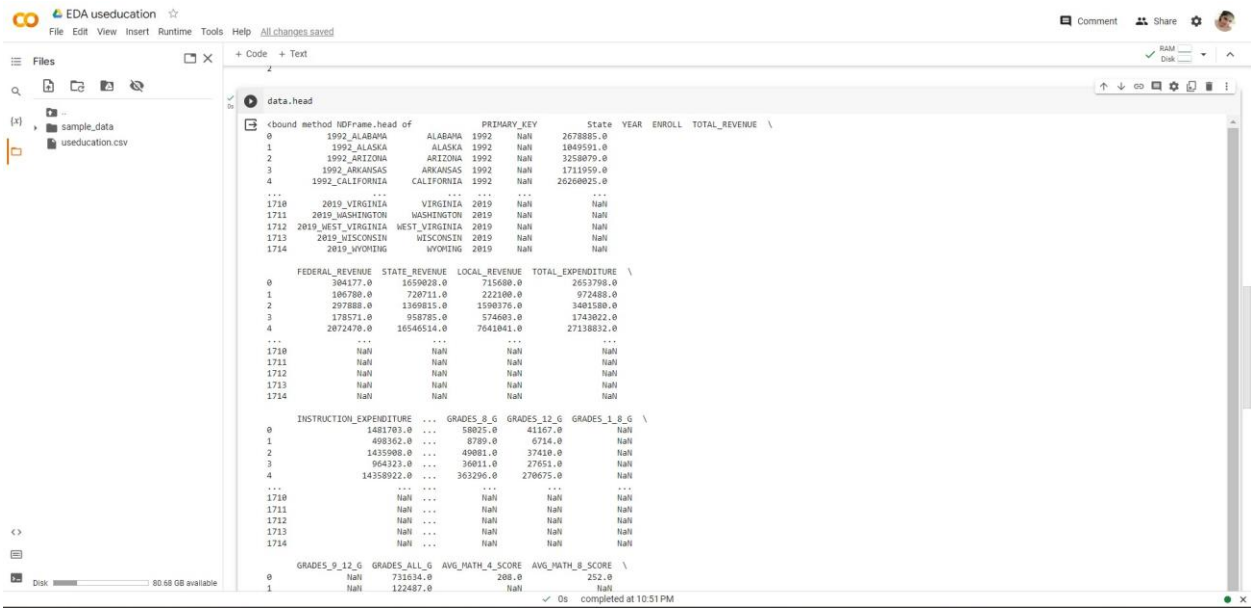
7. The total number of rows and columns in the file can be found using the following code. There are 1715 rows and 26 columns in the dataset.



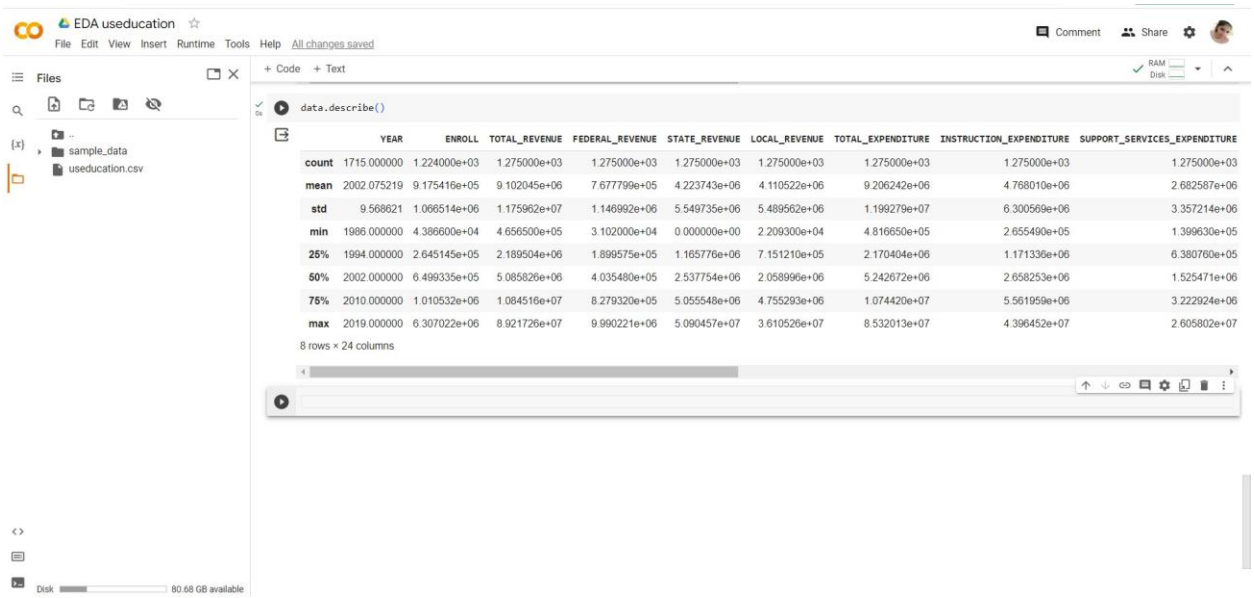
8. To determine the number of dimensions or axes in the data structure. The dataset has two dimensions and **data.ndim** is used to determine the number of axes or dimensions in the data structure.



9. The data.head extension is used to display the real values in the rows and columns**.**



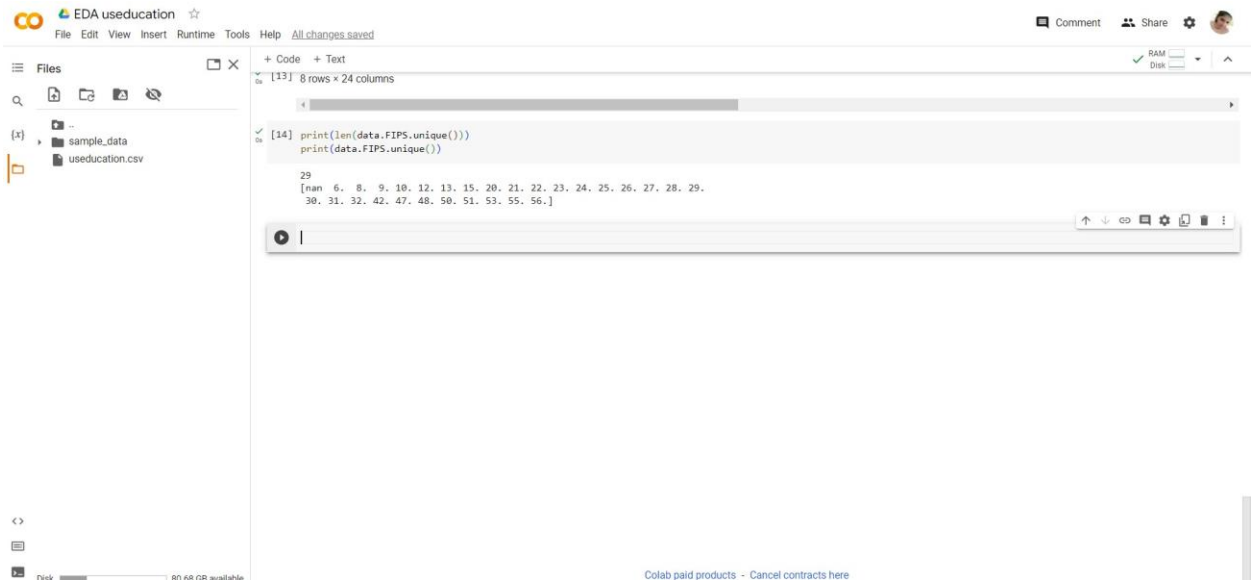
10. An summary of my dataset that is descriptive was produced using techniques such as **data.describe().**



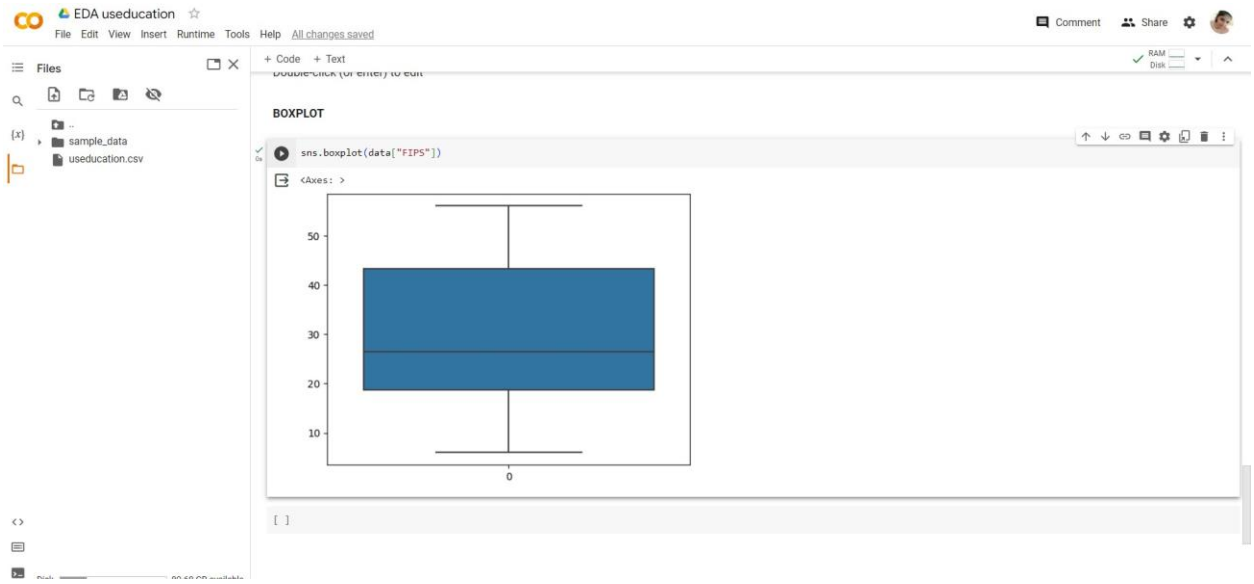
	YEAR	ENROLL	TOTAL_REVENUE	FEDERAL_REVENUE	STATE_REVENUE	LOCAL_REVENUE	TOTAL_EXPENDITURE	INSTRUCTION_EXPENDITURE	SUPPORT_SERVICES_EXPENDITURE
count	1715.000000	1.224000e+03	1.275000e+03	1.275000e+03	1.275000e+03	1.275000e+03	1.275000e+03	1.275000e+03	1.275000e+03
mean	2002.075219	9.175416e+05	9.102045e+06	7.677799e+05	4.223743e+06	4.110522e+06	9.206242e+06	4.768010e+06	2.682587e+06
std	9.568621	1.066514e+06	1.175962e+07	1.146992e+06	5.549735e+06	5.489562e+06	1.199279e+07	6.300569e+06	3.357214e+06
min	1986.000000	4.386600e+04	4.656500e+05	3.102000e+04	0.000000e+00	2.209300e+04	4.816650e+05	2.655490e+05	1.399630e+05
25%	1994.000000	2.645145e+05	2.189504e+06	1.899575e+05	1.165776e+06	7.151210e+05	2.170404e+06	1.171336e+06	6.380760e+05
50%	2002.000000	6.499335e+05	5.085826e+06	4.035480e+05	2.537754e+06	2.058996e+06	5.242672e+06	2.658253e+06	1.525471e+06
75%	2010.000000	1.010532e+06	1.084516e+07	8.279320e+05	5.055548e+06	4.755293e+06	1.074420e+07	5.561959e+06	3.222924e+06
max	2019.000000	6.307022e+06	8.921726e+07	9.990221e+06	5.090457e+07	3.610526e+07	8.532013e+07	4.396452e+07	2.605802e+07

8 rows × 24 columns

11. The code to find the total number of unique FIPS is provided below. A total of 29 FIPS with unique names are displayed.



12. To use Boxplot Python, you must import Matplotlib. Underneath the boxplot is the combination of the FIPS.



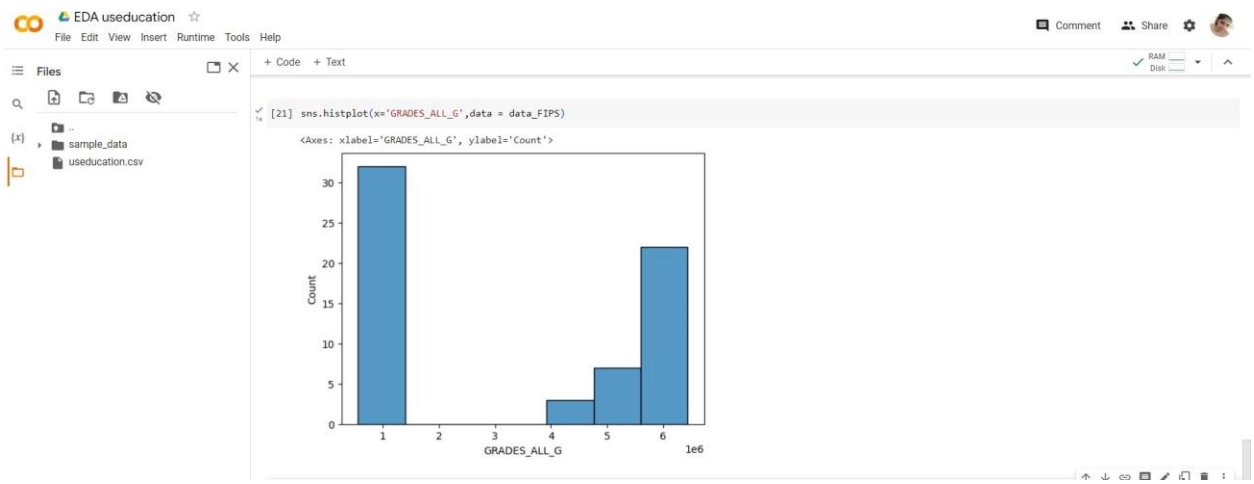
Interpretation:

The displayed boxplot provides insights into the distribution of the "FIPS" data. The central rectangle represents the interquartile range (IQR) with its lower and upper bounds indicating the 25th (Q1) and 75th (Q3) percentiles, respectively. The

horizontal line within the box signifies the median value, which offers a measure of central tendency. The whiskers, or the lines extending from the box, show the range within which the bulk of values fall, with outliers typically lying beyond these extremities. Based on the plot, the data appears to be roughly symmetric, given that the median is centrally located within the IQR. The absence of any points outside the whiskers suggests that there aren't noticeable outliers in the dataset. The overall range seems to span from a value slightly above 0 to about 50.

HISTOGRAM

In Python EDA, a histplot is a visual aid that displays data about a numerical variable's distribution, central trends, and spread by fusing a histogram and a kernel density plot. It's a great tool for figuring out patterns in data.



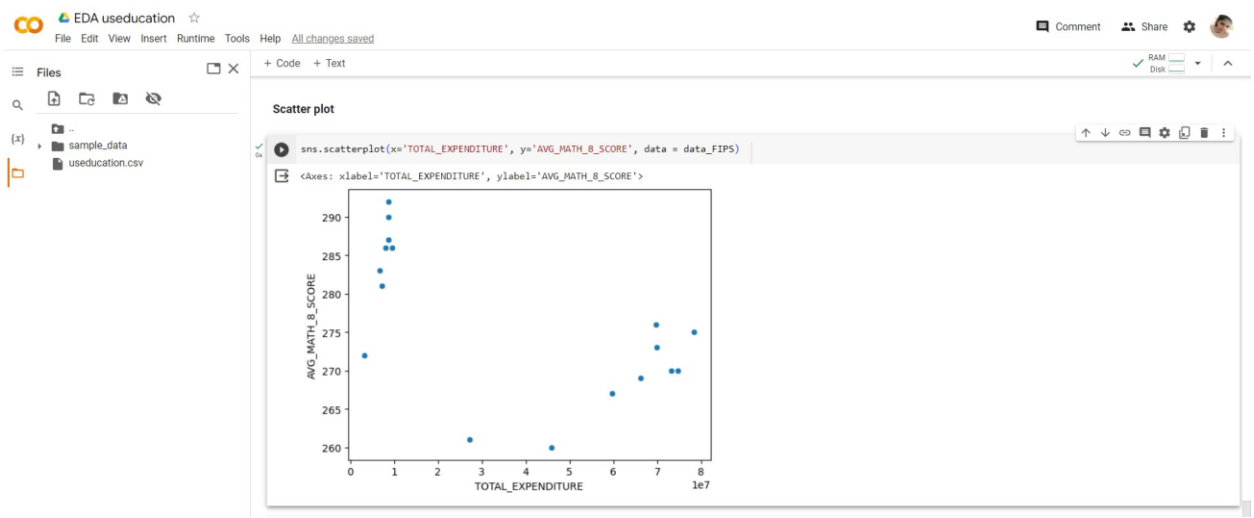
Interpretation:

The histogram showcases the distribution of data for the "GRADES_ALL_G" variable, which seems to represent the total grades for all students. The x-axis, labeled "GRADES_ALL_G," spans from around 2 million to slightly above 6 million. The y-axis, labeled "Count," indicates the frequency of observations. Most prominently, there's a significant peak at the leftmost side, suggesting that a substantial number of observations have grade totals close to 2 million. A smaller peak, or secondary mode, is noticeable between 5 million to slightly above 6 million. The region between these two peaks has fewer observations, with a particularly notable dip around the 4 million mark. In summary, the distribution

appears bimodal, with a significant number of entities having grade totals near 2 million and another notable group clustered just above 5 million.

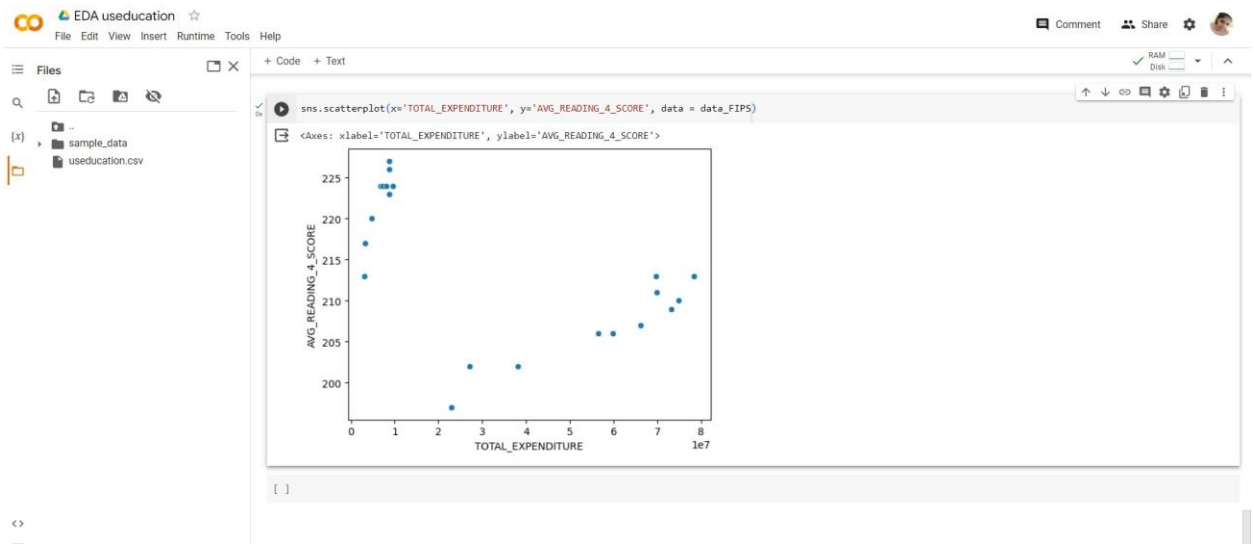
Scater Plot

In Python's EDA, a scatterplot is a graphical display that shows discrete data points as dots on a two-dimensional graph.it helps in the analysis of data patterns and correlations by revealing the relationship between two numerical variables.



Interpretation

The scatter plot displays the relationship between "TOTAL_EXPENDITURE" and "AVG_MATH_8_SCORE." The x-axis represents "TOTAL_EXPENDITURE" and spans from around 0 to slightly above 8 x 10^7. On the y-axis, the "AVG_MATH_8_SCORE" fluctuates between approximately 260 and 290. The plot suggests a mild positive correlation between the variables: as the "TOTAL_EXPENDITURE" increases, the "AVG_MATH_8_SCORE" generally seems to rise. However, this relationship is not very tight, as the data points are somewhat dispersed. While the plot implies that greater expenditure might be associated with higher math scores for 8th graders, the wide distribution of data points indicates other influencing factors or the possibility of outliers. In summary, there might be a link between expenditure and math scores, but the data suggests that other factors might also play a significant role in determining scores.



Interpretation

The scatter plot visualizes the relationship between "TOTAL_EXPENDITURE" and "AVG_READING_4_SCORE." On the x-axis, "TOTAL_EXPENDITURE" seems to be represented in a range that spans from approximately 0 to slightly above 8 x 10^7. On the y-axis, the "AVG_READING_4_SCORE" appears to range from around 200 to slightly above 225. There seems to be a modest positive correlation between the two variables: as the "TOTAL_EXPENDITURE" increases, the "AVG_READING_4_SCORE" generally appears to increase as well. However, the relationship is not perfectly linear, as there are data points scattered throughout, indicating potential outliers or other factors at play. Overall, while there's a suggestion that higher expenditure might be associated with better reading scores for 4th graders, the scatter of data points indicates that it's essential to consider other influencing factors or investigate the relationship further.

Citations:

1. U.S Education Datasets: https://www.kaggle.com/datasets/noriuk/us-education-datasets-unification-project

2.Google colab file

https://colab.research.google.com/drive/1l1YU50NgpZXope4yIrtA49k6KqvsYTfp#scrollTo=cmqDoXL4Wxbw

3.ORES_Group_dataset-GoogleDrive.(n.d.). https://drive.google.com/drive/folders/173tJ7JkxJeu9Pi62k9t00J40alAZfPP0?usp=share_link