Data Transformation with Python or R – Clean Your Data

Pamulapati Harika (001266981)
Saint Louis University
ORES-5160 Data Management
November 7 2023

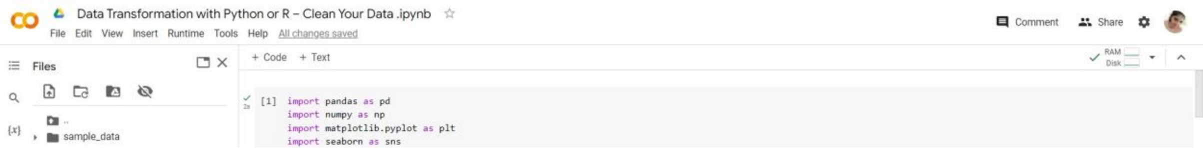
DATA CLEANING REPORT

Introduction:

The report's objective is to provide documentation of the "useducation.csv" dataset's data cleaning procedure. This report describes how the "useducation.csv" transforms the reasoning behind each decision.

1. Importing Required Libraries for Data Visualization and Analysis

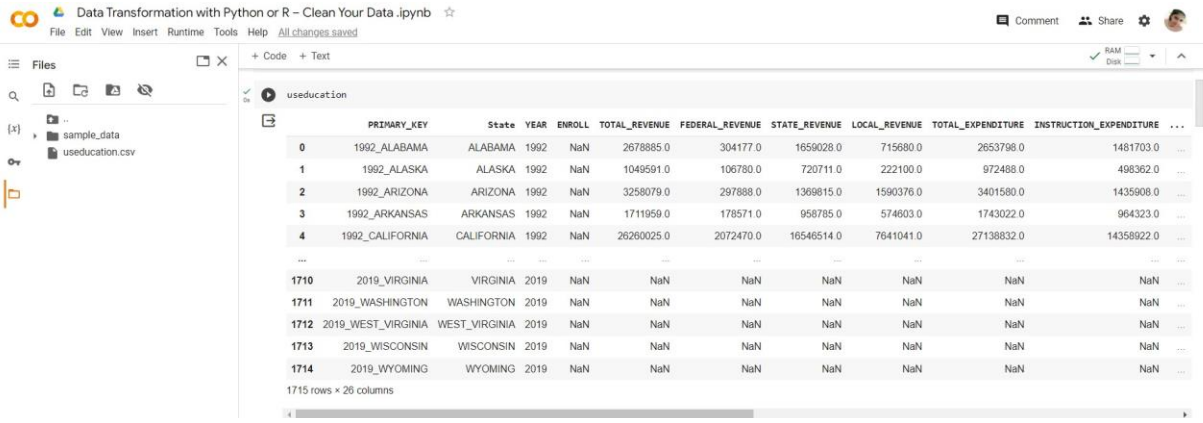
To start our search, I first loaded the libraries needed for data manipulation and visualization, such as matplotlib, pandas, numpy, and seaborn. These packages allow us to work efficiently with our dataset and create insightful visuals.



2. Importing the dataset

I labeled the file "useducation" when it was read into Google Colab using the code below.

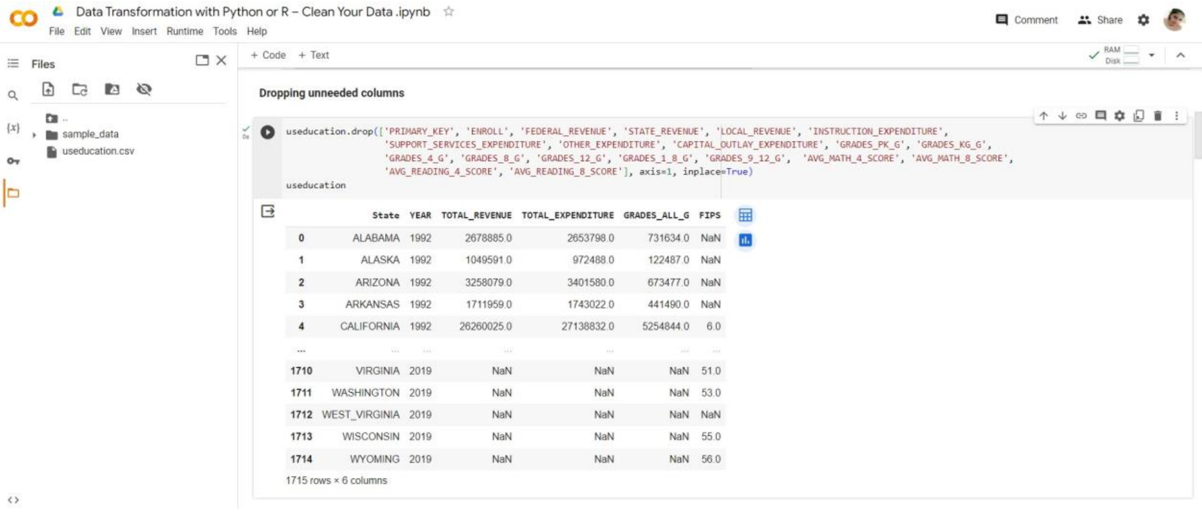
3. useducation dataset preview



	PRIMARY_KEY	State	YEAR	ENROLL	TOTAL_REVENUE	FEDERAL_REVENUE	STATE_REVENUE	LOCAL_REVENUE	TOTAL_EXPENDITURE	INSTRUCTION_EXPENDITURE	...
0	1992_ALABAMA	ALABAMA	1992	NaN	2678885.0	304177.0	1659028.0	715680.0	2653798.0	1481703.0	...
1	1992_ALASKA	ALASKA	1992	NaN	1049591.0	106780.0	720711.0	222100.0	972488.0	498362.0	...
2	1992_ARIZONA	ARIZONA	1992	NaN	3258079.0	297888.0	1369815.0	1590376.0	3401580.0	1435908.0	...
3	1992_ARKANSAS	ARKANSAS	1992	NaN	1711959.0	178571.0	958785.0	574603.0	1743022.0	964323.0	...
4	1992_CALIFORNIA	CALIFORNIA	1992	NaN	26260025.0	2072470.0	16546514.0	7641041.0	27138832.0	14358922.0	...
...	...	...	...	...	...	...	...	...	...	...	...
1710	2019_VIRGINIA	VIRGINIA	2019	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
1711	2019_WASHINGTON	WASHINGTON	2019	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
1712	2019_WEST_VIRGINIA	WEST_VIRGINIA	2019	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
1713	2019_WISCONSIN	WISCONSIN	2019	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
1714	2019_WYOMING	WYOMING	2019	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...

1715 rows × 26 columns

4. Dropping unneeded columns



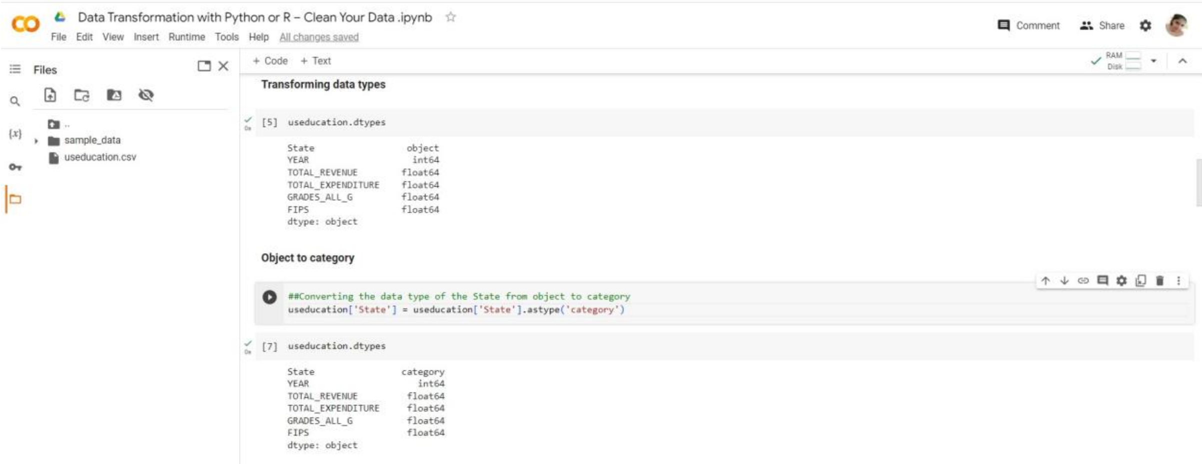
	State	YEAR	TOTAL_REVENUE	TOTAL_EXPENDITURE	GRADES_ALL_G	FIPS
0	ALABAMA	1992	2678885.0	2653798.0	731634.0	NaN
1	ALASKA	1992	1049591.0	972488.0	122487.0	NaN
2	ARIZONA	1992	3258079.0	3401580.0	673477.0	NaN
3	ARKANSAS	1992	1711959.0	1743022.0	441490.0	NaN
4	CALIFORNIA	1992	26260025.0	27138832.0	5254844.0	6.0
...	...	...	...	...	...	...
1710	VIRGINIA	2019	NaN	NaN	NaN	51.0
1711	WASHINGTON	2019	NaN	NaN	NaN	53.0
1712	WEST_VIRGINIA	2019	NaN	NaN	NaN	NaN
1713	WISCONSIN	2019	NaN	NaN	NaN	55.0
1714	WYOMING	2019	NaN	NaN	NaN	56.0

The Python code in the cell attempts to eliminate specified columns from the dataset, namely the 'PRIMARY_KEY', 'ENROLL', 'FEDERAL_REVENUE', 'STATE_REVENUE','LOCAL_REVENUE','INSTRUCTION_EXPENDITURE', 'SUPPORT_SERVICES_EXPENDITURE', 'OTHER_EXPENDITURE', 'CAPITAL_OUTLAY_EXPENDITURE','GRADES_PK_G', 'GRADES_KG_G', 'GRADES_4_G', 'GRADES_8_G', 'GRADES_12_G', 'GRADES_1_8_G', 'GRADES_9_12_G', 'AVG_MATH_4_SCORE', 'AVG_MATH_8_SCORE', 'AVG_READING_4_SCORE' and 'AVG_READING_8_SCORE' columns.

This is done by passing the inplace =True parameter to the drop method on the dataframe useducation, telling it to modify the original data frame directly instead of allocating the outcome to a new variable. The columns for 'State', 'YEAR ', 'TOTAL_REVENUE ', 'TOTAL_EXPENDITURE ', 'GRADES_ALL_G ', and 'FIPS ' are displayed in the data frame that results below the code cell.

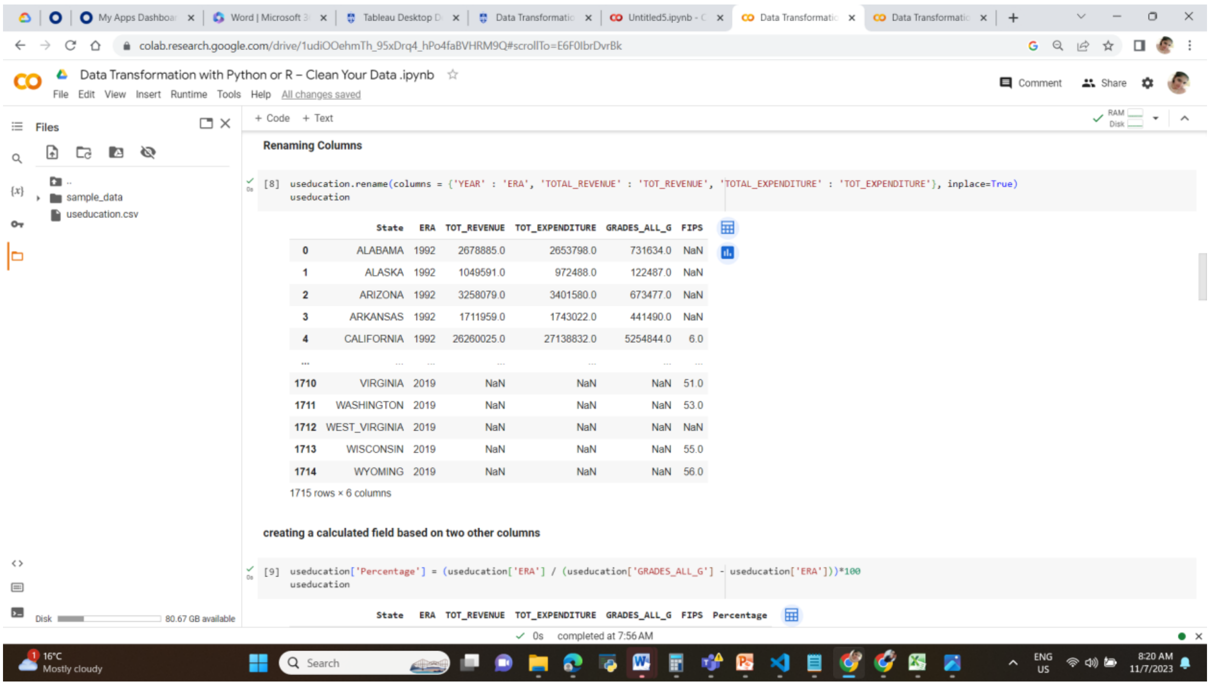
The decision to remove unnecessary columns is indicative of a focus on relevant data and data simplification for analysis.

5. Transforming Data Types



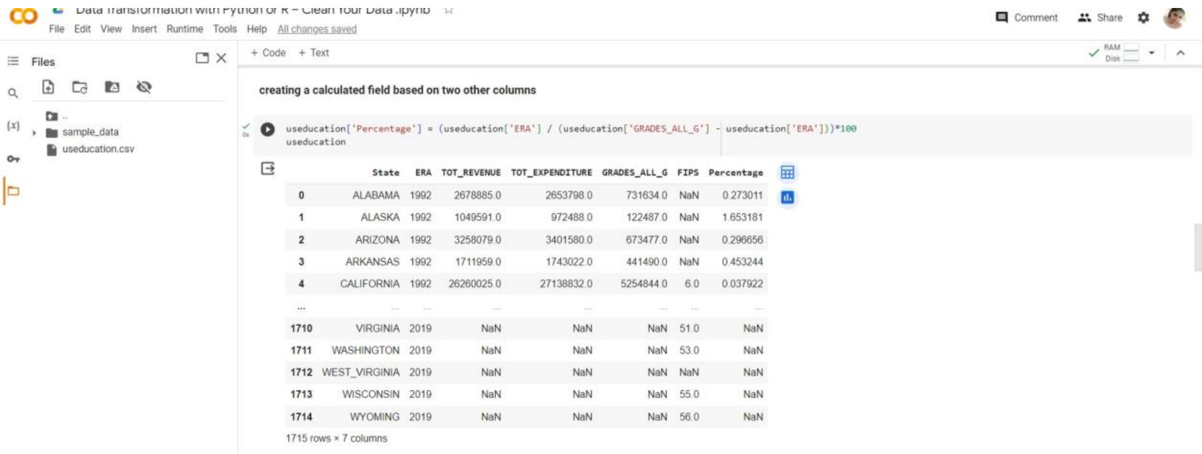
The 'State' column in the useducation Data Frame underwent a data type modification by me. I displayed the data types of each column in the Data Frame using the dtypes attribute. The output verified that 'State' was successfully converted to Category.

6. Renaming the columns



During this round of data processing, I modified the column names in the useducation Data Frame to be more concise and clear. I modified the 'YEAR', 'TOTAL_REVENUE', and 'TOTAL_EXPENDITURE' columns to 'ERA,' TOT_REVENUE,' and TOT_EXPENDITURE,' using the rename () method. The changes are applied, therefore the original Data Frame remains unchanged. This is indicated by the input in place=True. Shorter column names are sometimes more appropriate for data handling and coding, thus after renaming the columns, the Data Frame shows the new column headings, improving data accessibility and making the Data Frame easier to deal with in further analysis.

7. Creating a calculated field based on two other columns



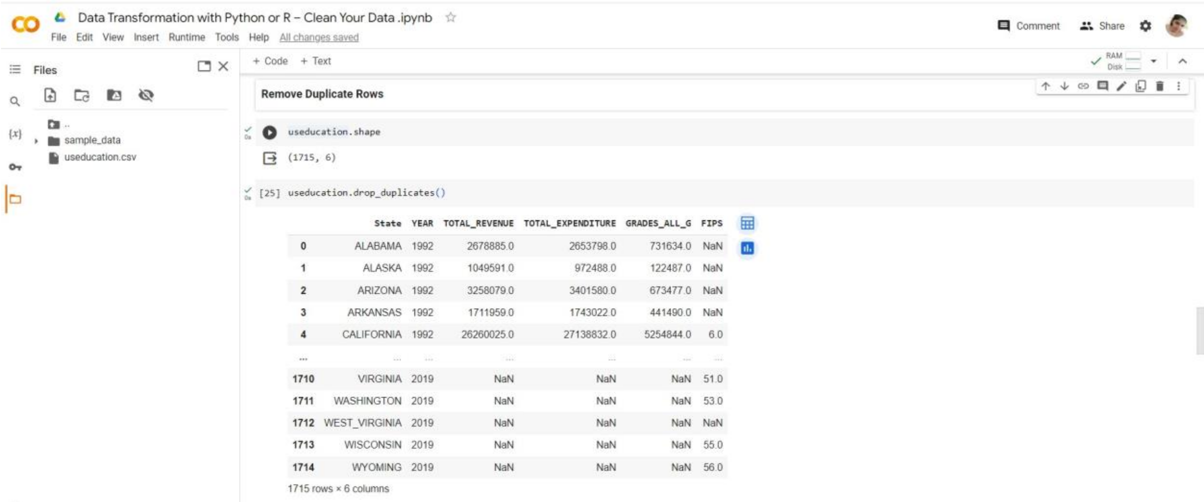
A significant data transformation procedure is introduced in the given code snippet inside the framework of educational data analysis. The code adds a new column called "Percentage" to a DataFrame called "useducation." A core educational metric is represented by the estimated values that fill this newly created column.

The percentage of pupils who have attained a given educational milestone is the main focus of the computation, which is based on the idea of educational attainment. It is the ratio of the 'ERA' (the number of students who reach the educational achievement) to the discrepancy between the 'ERA' count and the total number of students in the 'GRADES_ALL_G' column. The outcome is then expressed as a percentage by multiplying it by 100.

Measuring the percentage of students who have attained a particular educational outcome in relation to the total number of students, the 'Percentage' column provides insightful information on educational progress and success rates. Using this data to evaluate the efficacy of schooling, identify patterns, and formulate wise policy decisions can be very helpful.

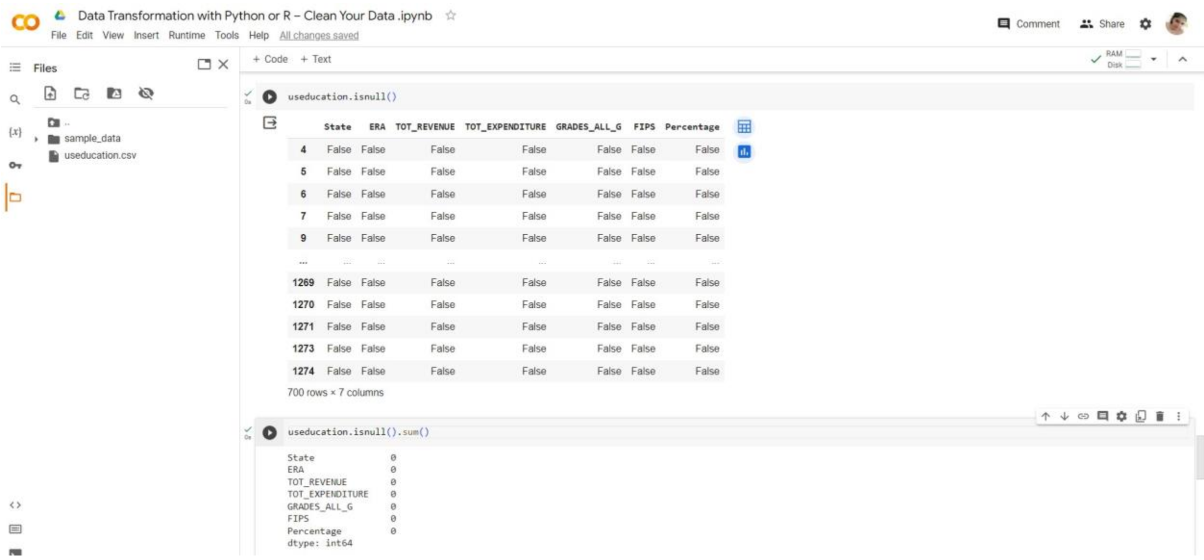
The execution of the code and the subsequent presentation of the 'useducation' DataFrame represent a critical stage in the analysis of educational data, providing researchers and data analysts with a refined dataset enhanced by this informative 'Percentage' column. The addition of this column expands the dataset's analytical possibilities and facilitates the extraction of significant conclusions about student performance and educational outcomes.

8. Remove duplicate rows



The Data Frame ‘useducation’ has 1715 rows and 6 columns based on the output of the `shape` function. The shape of the Data Frame does not change after using the `drop_duplicates()` method, indicating that there were no duplicate rows in the dataset that needed to be removed.

9. Evaluating presence of NULL values in the useducation dataset

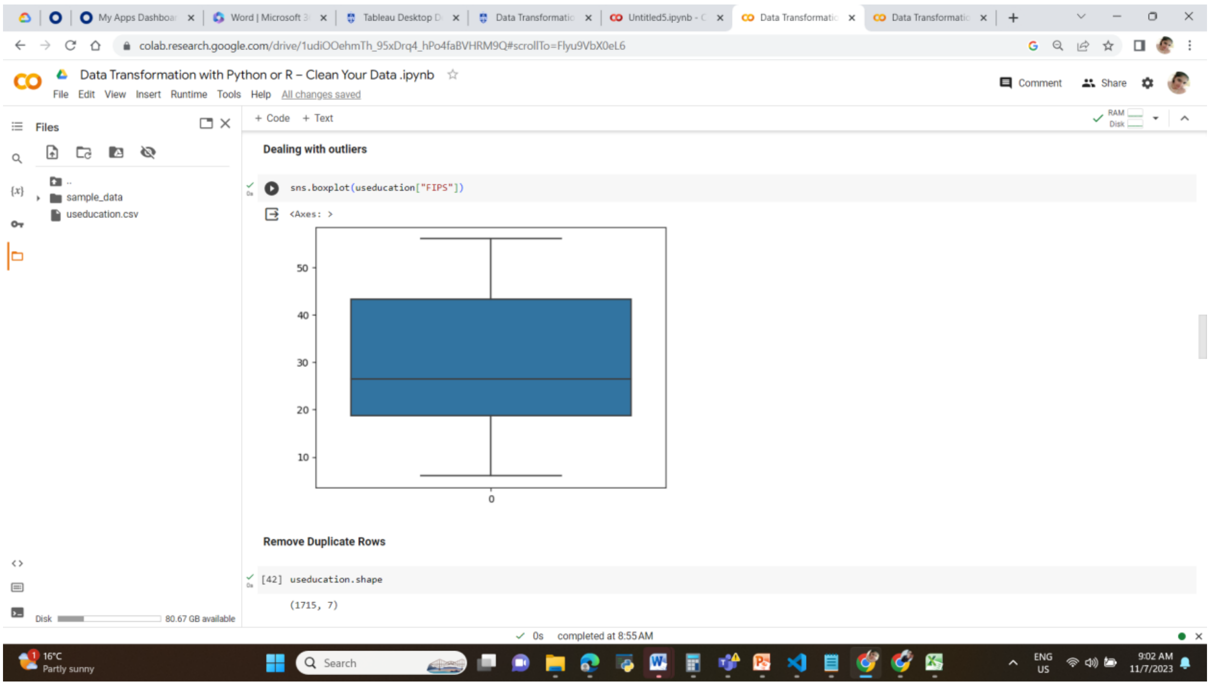


A basic evaluation of data quality in the context of educational data analysis is provided by the provided code snippet. 'useducation' DataFrame's missing values are identified and quantified by means of the useducation.isnull().sum() function.

This operation's output counts the number of null or missing values in each column of the dataset. This information is essential for preprocessing and evaluating the quality of the data. It clarifies which columns may contain missing data and to what degree, hence illuminating the dataset's completeness. The total amount of null values in every column indicates how many data points require attention, imputation, or additional research.

One crucial element in the data analysis process is figuring out what missing data is and how to fill it in. It influences decisions on data imputation, exclusion, and refinement and provides researchers and data analysts with information about the dataset's integrity. This code's method of quantifying missing values makes it easier to clean and prepare data in an orderly and structured manner, which is essential for accurate and trustworthy analysis and interpretation in the field of education.

10. Dealing with outliers



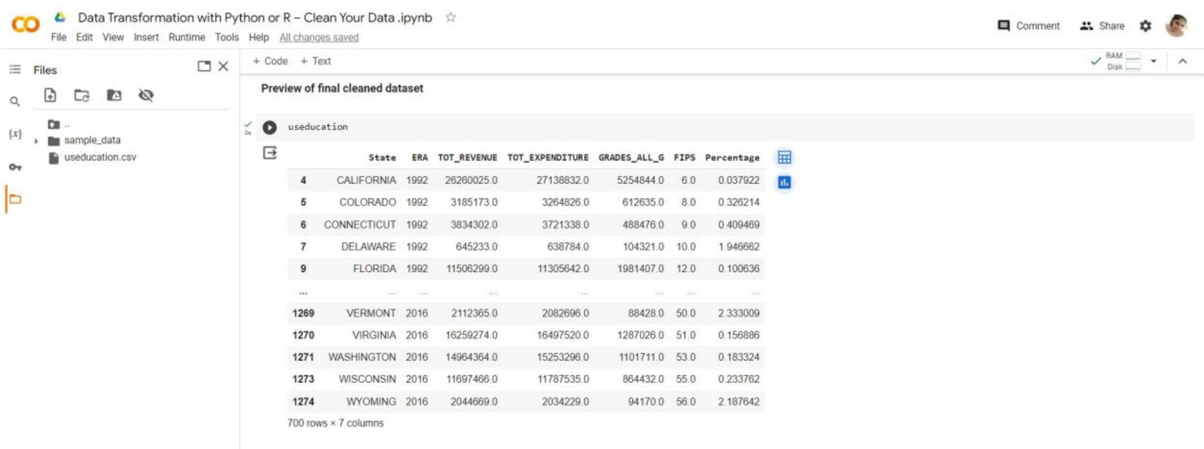
The code sample that is included presents a box plot as a data visualization method in the context of educational data analysis. The code primarily focuses on the 'FIPS' column of a DataFrame called 'useducation,' which normally contains regional or geographic identifiers like Federal Information Processing Standards (FIPS) codes.

The program "sns.boxplot(useducation["FIPS"]"), which generates the box plot using the Seaborn library, is an effective visual aid for comprehending the variability and distribution of the 'FIPS' data. Essential statistical details about the dataset, like the median, quartiles, and any outliers, are shown in a box plot.

Since geographic factors can have a big impact on educational outcomes and policies, this visual approach is especially useful for educational data. Data analysts and researchers can use the box plot to quickly and easily summarize the distribution of 'FIPS' values, which helps them to make well-informed judgments on educational initiatives, resources, and interventions at the regional level.

This code's execution and the box plot's construction help to provide a more thorough grasp of the geographical dimension contained in the educational dataset, which in turn makes data-driven insights and educational decision-making easier.

11. Preview of the final cleaned dataset



	State	ERA	TOT_REVENUE	TOT_EXPENDITURE	GRADES_ALL_G	FIPS	Percentage
4	CALIFORNIA	1992	26260025.0	27138832.0	5254844.0	6.0	0.037922
5	COLORADO	1992	3185173.0	3264826.0	612635.0	8.0	0.326214
6	CONNECTICUT	1992	3834302.0	3721338.0	488476.0	9.0	0.409469
7	DELAWARE	1992	645233.0	638784.0	104321.0	10.0	1.946662
9	FLORIDA	1992	11506299.0	11305642.0	1981407.0	12.0	0.100636
...	...	...	...	...	...	...	...
1269	VERMONT	2016	2112365.0	2082696.0	88428.0	50.0	2.333009
1270	VIRGINIA	2016	16259274.0	16497520.0	1287026.0	51.0	0.156886
1271	WASHINGTON	2016	14964364.0	15253296.0	1101711.0	53.0	0.183324
1273	WISCONSIN	2016	11697466.0	11787535.0	864432.0	55.0	0.233762
1274	WYOMING	2016	2044669.0	2034229.0	94170.0	56.0	2.187642

700 rows × 7 columns

The dataset seems to be an organized compilation of demographic and financial information about schooling across several states between 1992 and 2016. 'State', 'ERA' (presumably the year), 'TOT_REVENUE' (total revenue), 'TOT_EXPENDITURE' (total expenditure), 'GRADES_ALL_G' (possibly the total number of students across all grades), 'FIPS' (geographic codes used to identify U.S. states), and 'Percentage' (which may represent some sort of ratio or rate relevant to the other data points) are among the columns that are present. The uniform structure and the inclusion of computed fields like 'Percentage' suggest that the data has been cleansed for ease of analysis. The 700 rows in the dataset indicate a thorough collection spanning 25 years for several states, which could be helpful for longitudinal fiscal and educational analysis.

Conclusion

In conclusion, this overview addressed popular data cleaning methods such as eliminating duplicate rows, renaming columns, handling null or infinite values, changing data types, eliminating unnecessary columns, creating a calculated field based on two other columns and dealing with outliers. The creation of an excellent analytical basis table provided the motivation for every action. Decisions were made to balance simplifying the data for core analysis with preserving important information.

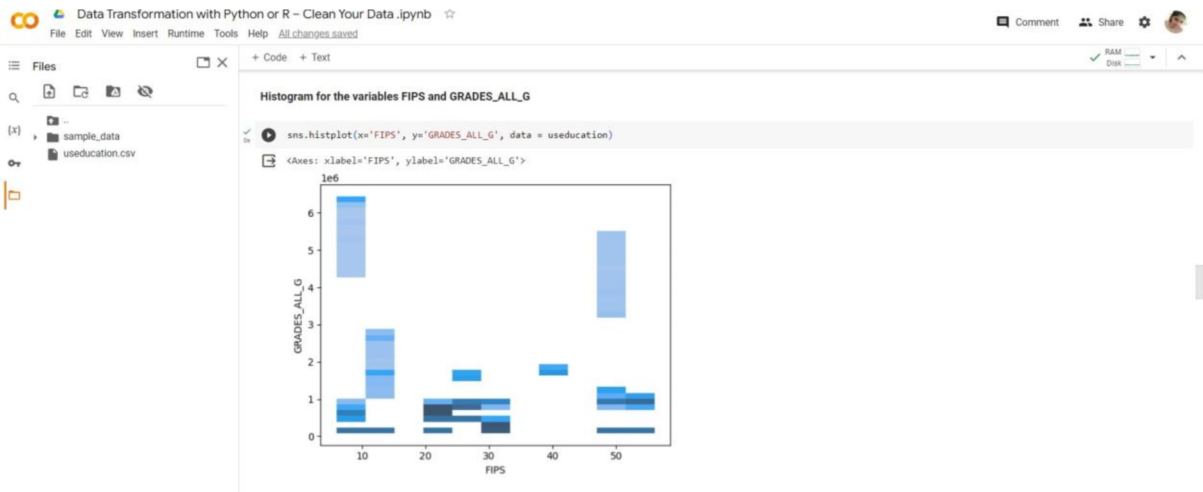
References

- **U.S. Education Datasets:** Unification Project. (n.d.). U.S. Education Datasets:UnificationProject|Kaggle. https://www.kaggle.com/datasets/noriuk/us-education-datasets-unification-project
- **ORES_Group_datasets–GoogleDrive.(n.d.).** https://drive.google.com/drive/folders/173tJ7JkxJeu9Pi62k9t00J40alAZfPP0?usp=share_link
- **Google colab link of cleaned dataset**

  https://github.com/HarikaPamulapati/ores/tree/0c0c7f818d3cca03628d0574e096113c039da00d
- https://www.youtube.com/playlist?list=PLjNQtX45f0dRONMZZKkCzn_EjdzUXc8Rx
- https://chat.openai.com/c/688b6964-5497-4ddf-80d0-44a05b592a5a

Appendix


Data Transformation with Python or R – Clean Your Data .ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
Files
sample_data
useducation.csv
+ Code + Text
Histogram
sns.histplot(data=useducation, x="FIPS", y="ERA")
<Axes: xlabel='FIPS', ylabel='ERA'>
ERA
FIPS


Data Transformation with Python or R – Clean Your Data .ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
Files
sample_data
useducation.csv
+ Code + Text
Histogram for the variables FIPS and GRADES_ALL_G
sns.histplot(x='FIPS', y='GRADES_ALL_G', data = useducation)
<Axes: xlabel='FIPS', ylabel='GRADES_ALL_G'>
1e6
GRADES_ALL_G
FIPS


Data Transformation with Python or R – Clean Your Data .ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
Files
sample_data
useducation.csv
+ Code + Text
Distplot for the variable FIPS
sns.displot(useducation['FIPS'], bins=5)
<seaborn.axisgrid.FacetGrid at 0x7f610f543f70>
Count
FIPS

Data Transformation with Python or R – Clean Your Data .ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

Files

sample_data
useducation.csv

+ Code + Text

Scatterplot for the variables FIPS and ERA

```
sns.scatterplot(x='FIPS', y='ERA', data = useducation)
```

```
<Axes: xlabel='FIPS', ylabel='ERA'>
```