**Data Analysis for Provisional\_COVID-19\_Deaths\_by\_Sex\_and\_Age Dataset. As a Data Scientist i want to calculate the mean of 'Total Deaths', the median of 'COVID-19 Deaths', the mode of the 'Sex' with higher death occurance, range of 'Influenza Deaths', the variance of 'Influenza Death' spread of the dataset, the Standard Deviation of 'Influenza Death' deviation score, the PMF(probability mass function) of 'Influenza Death' mass function, the PDF(probability density function) of 'Influenza Death' and visualize the distributions of the dataset to derive useful insight from 'Provisional\_COVID-19\_Deaths\_by\_Sex\_and\_Age' dataset.**

**DataSet Overview.**

All data come from one source which was csv file shared and contains details in columns as following.

index: just index.

Data As Of: Initial data point.

Start Date: Start date.

End Date: End date.

Group: The group order.

Year: The year.

Month: The month.

State: The state.

Sex: The gender(Sex).

Age Group: The age group order.

COVID-19 Deaths: The COVID-19 Deaths scores.

Total Deaths: Total death scores.

Pneumonia Deaths: The Pneumonia Deaths scores.

Pneumonia and COVID-19 Deaths: The Pneumonia and COVID-19 Deaths scores.

Influenza Deaths: The Influenza Deaths scores.

Pneumonia, Influenza, or COVID-19 Deaths: The Pneumonia, Influenza, or COVID-19 Deaths scores.

Footnote: The Footnote records.

**Table of Content**

We will go through several tasks to archieve our goals and purposes:

Task1: Problem statement(goals and purpose) for Provisional\_COVID-19\_Deaths\_by\_Sex\_and\_Age Dataset for Analysis with an overview of the whole Dataset.

Task2: Importing libraries.

Task3: Loading data from my google drive, i used 'COLAB' for my python programming.

Task4: Viewing the data, the head view of the 1st 5 rows of the data.

Task5: Viewing the data, the bottom view of the 1st 5 rows of the train data.

Task6: To view the shape of the data.

Task7: Checking for NAN/NULL values(missing values) of the data.

Task8: To get the details of the data colums nature information.

Task9: To get the full descriptive statistics chart table for 'Year' colum from the data with missing values.

Task10: Computing mean value Year to the missing values in Year colum

Task11: Computing mean value Month to the missing values in Month colum

Task12: Computing mean value COVID-19 Deaths to the missing values in COVID-19 Deaths colum

Task13: Computing mean value Total Deaths to the missing values in Total Deaths colum

Task14: Computing mean value Pneumonia Deaths to the missing values in Pneumonia Deaths colum

Task15: Computing mean value Pneumonia and COVID-19 Deaths to the missing values in Pneumonia and COVID-19 Deaths colum

Task16: Computing mean value Influenza Deaths to the missing values in Influenza Deaths colum

Task17: Computing mean value Pneumonia, Influenza, or COVID-19 Deaths to the missing values in Pneumonia, Influenza, or COVID-19 Deaths colum

Task18: To calculate the mean of Total Deaths and the mean report Total Deaths.

Task19: To calculate for median of COVID-19 Deaths colum and report

Task20: Calculating for the mode of the Sex with higher death occurance and report.

Task21: To calculate the range of Influenza Deaths in the dataset and report.

Task22: To calculate the variance of Influenza Death spread of the dataset and report.

Task23: To calculate the Standard Deviation of Influenza Death , Deviation score the dataset and report.

Task24: To calculate the PMF of Influenza Death , the mass function value and report.

Task25: To calculate the PDF of Influenza Death , the distribution influenza death function values and report.

References: The sources i took the code from

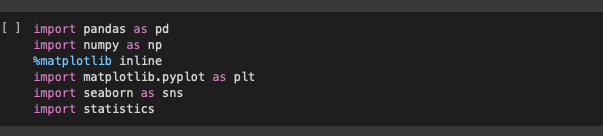
**Task1**

**Problem Statement(Aim):**

Data Analysis for Provisional\_COVID-19\_Deaths\_by\_Sex\_and\_Age. As a Data Scientist i want to calculate the mean of 'Total Deaths', the median of 'COVID-19 Deaths', the mode of the 'Sex' with higher death occurance, range of 'Influenza Deaths', the variance of 'Influenza Death' spread of the dataset, the Standard Deviation of 'Influenza Death' Deviation score, the PMF(probability mass function) of 'Influenza Death' mass function, the PDF(probability density function) of 'Influenza Death' and visualize the distributions of the dataset to derive useful insight from 'Provisional\_COVID-19\_Deaths\_by\_Sex\_and\_Age' data.

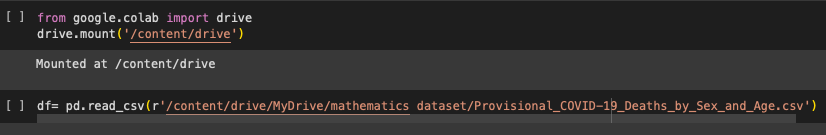
**Task2:**

Importing libraries



**Task3:**

Loading data from my google drive, i used 'COLAB' for my python programming.



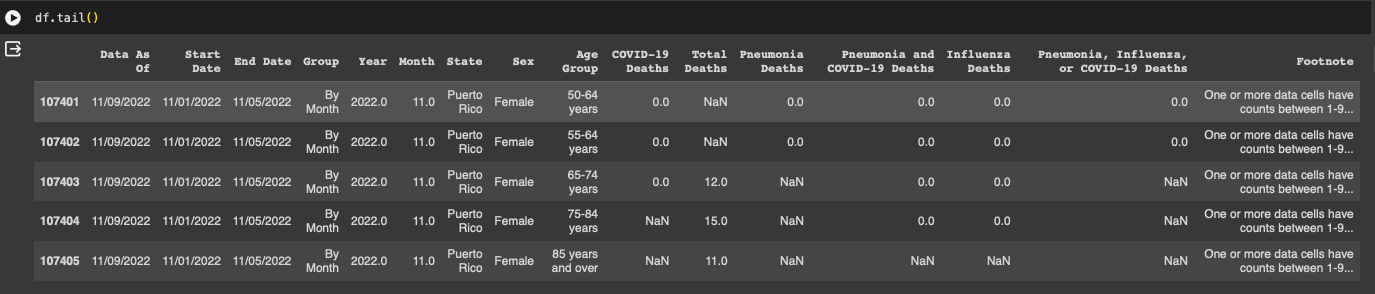
**Task4:**

Viewing the data, the head view of the 1st 5 rows of the data.



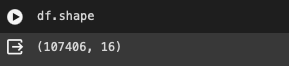
**Task5:**

Viewing the data, the bottom view of the 1st 5 rows of the train data.



**Task6:**

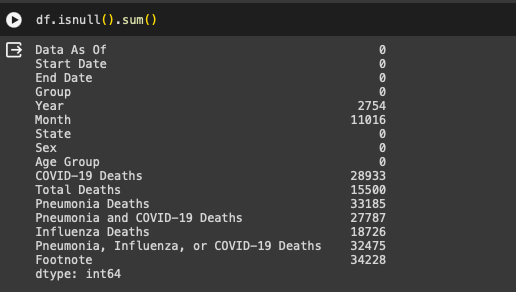
To view the shape of the data.



The data contains 107406 files and 16 colums

**Task7:**

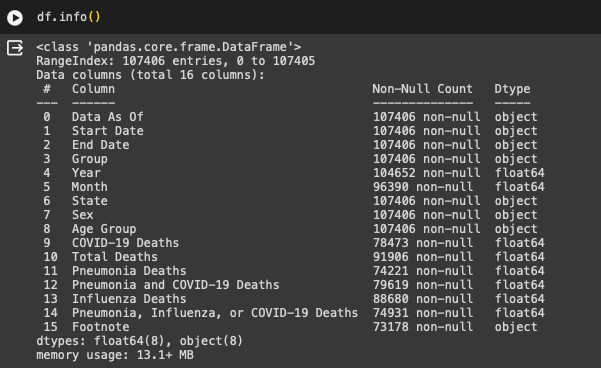
Checking for NAN/NULL values(missing values) of the data.



There are 2754 missing values in Year colume, 11016 missing values in Month, 28933 missing values in COVID-19 Deaths,15500 missing values in Total Deaths,33185 missing values in Pneumonia Deaths,27787 missing values in Pneumonia and COVID-19 Deaths,18726 missing values in Influenza Deaths,32475 missing values in Pneumonia, Influenza, or COVID-19 Deaths,34228 missing values in Footnote. I have to fill the missing values using mean method for numerical missing values and mode method for categorical values.

**Task8:**

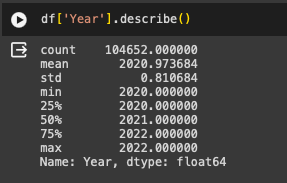
To get the details of the data colums nature information.



There are 107406 entries, ranges from 0 to 107405 Data columns (total 16 columns) dtypes is float64(8), object(8) of the data. I will convert the "Object" to "Categorical" for proper data cleaning order.

**Task9:**

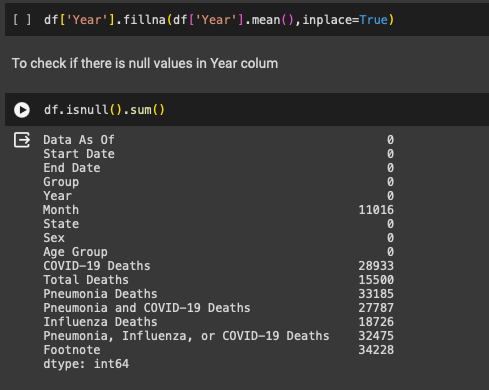
To get the full descriptive statistics chart table for 'Year' colum from the data with missing values.



Year is numerical colum so i fill it with Mean Imputation. To remove null values in Year by computing the mean value of the column since Year is a numerical values.

**Task10:**

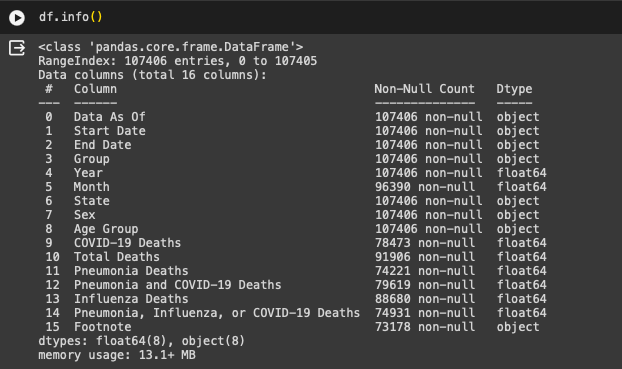
Computing mean value Year to the missing values in Year colum.



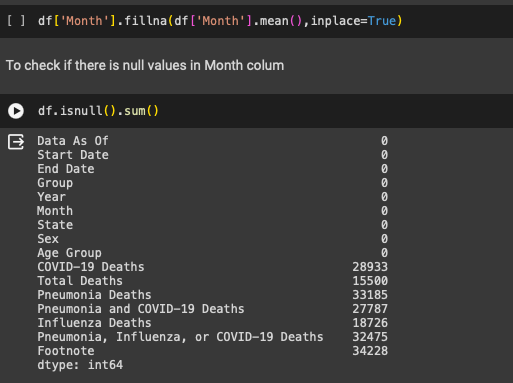
We can see that all the missing values in year colum has been filled with the mean value of year colum.

**Task11:**

Computing mean value Month to the missing values in Month colum.



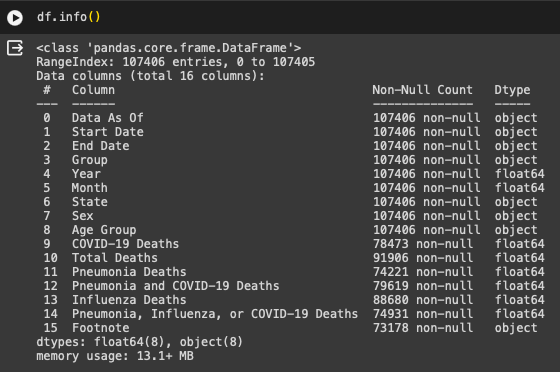
Month is numerical column so i fill it with Mean Inputation. To remove null values in Month by computing the mean value of the column since Month is a numerical values.



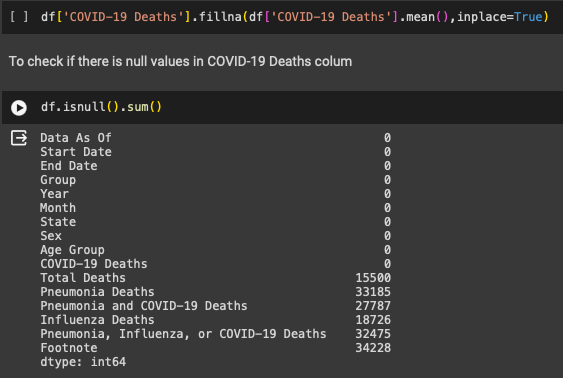
We can see that all the missing values in Month colum has been filled with the mean value of Month colum.

**Task12:**

Computing mean value COVID-19 Deaths to the missing values in COVID-19 Deaths colum.



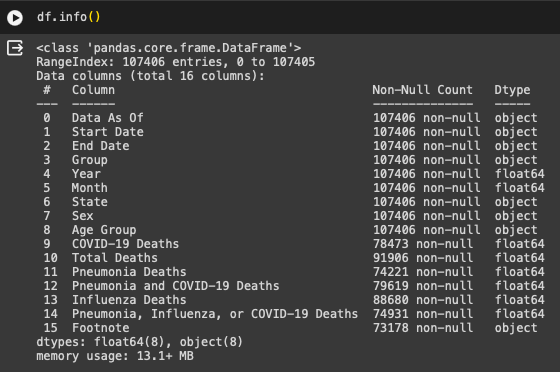
COVID-19 Deaths is numerical column so i fill it with Mean Imputation. To remove null values in COVID-19 Deaths by computing the mean value of the column since COVID-19 Deaths is a numerical values.



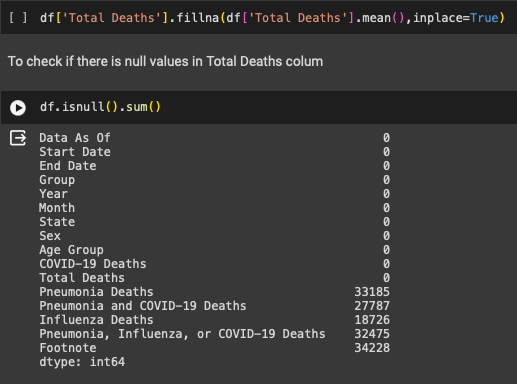
We can see that all the missing values in COVID-19 Deaths colum has been filled with the COVID-19 Deaths value of COVID-19 colum.

**Task13:**

Computing mean value Total Deaths to the missing values in Total Deaths colum.



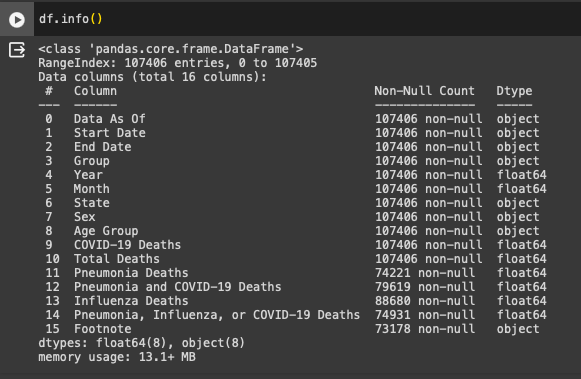
Total Deaths is numerical column so i fill it with Mean Imputation. To remove null values in Total Deaths by computing the mean value of the column since Total Deaths is a numerical values.



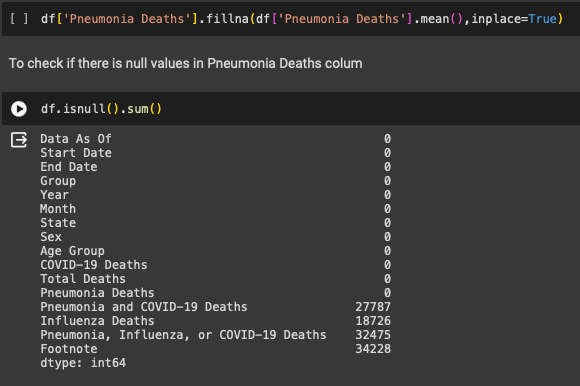
We can see that all the missing values in Total Deaths colum has been filled with the Total Deaths value of Total Deaths colum.

**Task14:**

Computing mean value Pneumonia Deaths to the missing values in Pneumonia Deaths colum.



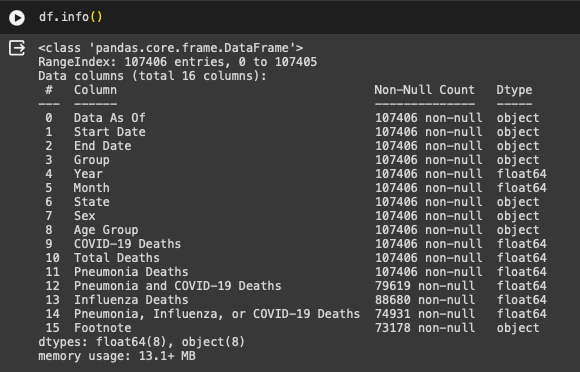
Pneumonia Deaths is numerical column so i fill it with Mean Imputation. To remove null values in Pneumonia Deaths by computing the mean value of the column since Pneumonia Deaths is a numerical values.



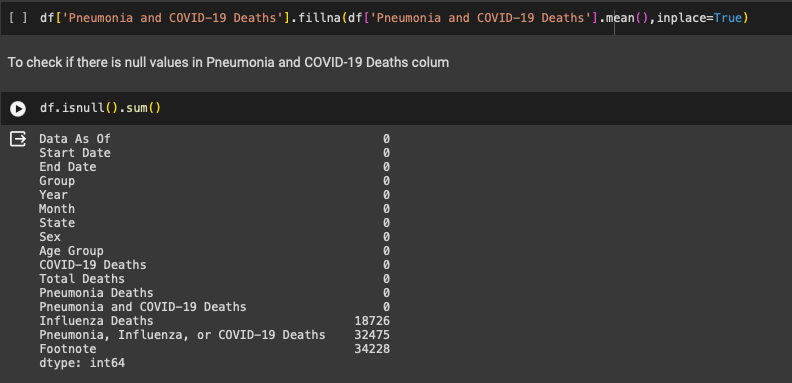
We can see that all the missing values in Pneumonia Deaths colum has been filled with the Pneumonia Deaths value of Pneumonia Deaths colum.

**Task15:**

Computing mean value Pneumonia and COVID-19 Deaths to the missing values in Pneumonia and COVID-19 Deaths colum.



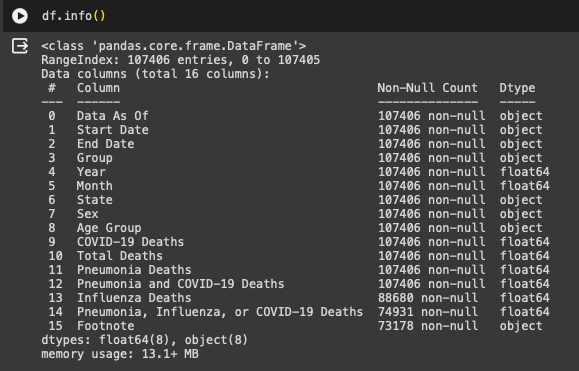
Pneumonia and COVID-19 Deaths is numerical column so i fill it with Mean Imputation. To remove null values in Pneumonia and COVID-19 Deaths by computing the mean value of the column since Pneumonia and COVID-19 Deaths is a numerical values.



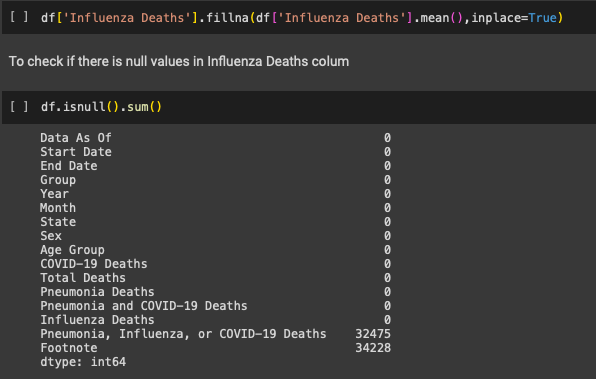
We can see that all the missing values in Pneumonia and COVID-19 Deaths colum has been filled with the Pneumonia and COVID-19 Deaths value of Pneumonia and COVID-19 Deaths colum.

**Task16:**

Computing mean value Influenza Deaths to the missing values in Influenza Deaths colum.



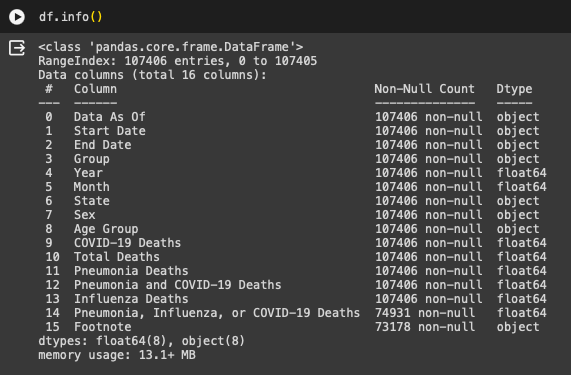
Influenza Deaths is numerical column so i fill it with Mean Imputation. To remove null values in Pneumonia and Influenza Deaths by computing the mean value of the column since Pneumonia and Influenza Deaths is a numerical values.



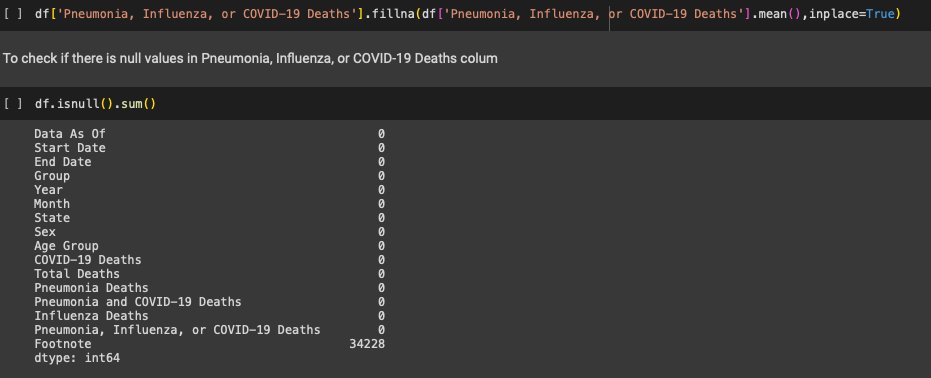
We can see that all the missing values in Influenza Deaths colum has been filled with the Influenza Deaths value of Influenza Deaths colum.

**Task17:**

Computing mean value Pneumonia, Influenza, or COVID-19 Deaths to the missing values in Pneumonia, Influenza, or COVID-19 Deaths colum.

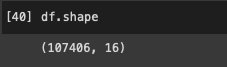


Pneumonia, Influenza, or COVID-19 Deaths is numerical column so i fill it with Mean Imputation. To remove null values in Pneumonia, Influenza, or COVID-19 Deaths by computing the mean value of the column since Pneumonia, Influenza, or COVID-19 Deaths is a numerical values.



We can see that all the missing values in Pneumonia, Influenza, or COVID-19 Deaths colum has been filled with the Pneumonia, Influenza, or COVID-19 Deaths value of Pneumonia, Influenza, or COVID-19 Deaths colum.

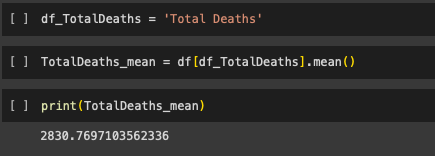
Our data is totally clean and contains no missing value for the colums features that we are using for our mathematical calculations and reports.



The total values of the dataset are 107406 with 16 colums

**Task18:**

To calculate the mean of Total Deaths and the mean report Total Deaths.

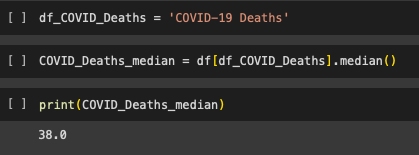


**Report:**

The 'Provisional\_COVID-19\_Deaths\_by\_Sex\_and\_Age' dataset reveals that the mean of the Total Deaths is approximately 2830.77. This statistical measure provides insight into the average mortality across different demographic groups, emphasizing the significance of understanding COVID-19's impact on various age and gender categories.

**Task19:**

To calculate for median of COVID-19 Deaths colum and report.

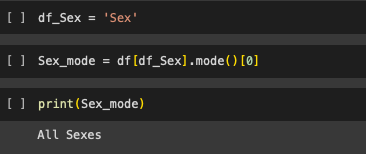


**Report:**

In the 'Provisional\_COVID-19\_Deaths\_by\_Sex\_and\_Age' dataset, the median for 'COVID-19 Deaths' is 38.0. This central tendency measure provides a representative value, indicating the middle point of the distribution. Understanding the median contributes to a more comprehension of mortality patterns across different demographics.

**Task20:**

Calculating for the mode of the Sex with higher death occurance and report.

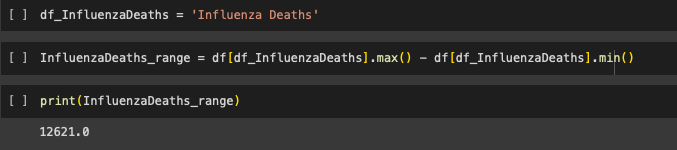


**Report:**

The 'Provisional\_COVID-19\_Deaths\_by\_Sex\_and\_Age' dataset reveals that the mode for 'Sex' is "All Sexes." This statistical mode indicates the most frequently occurring category, emphasizing the predominant representation of data related to all sexes in the context of COVID-19 deaths.

**Task21:**

To calculate the range of Influenza Deaths in the dataset and report.

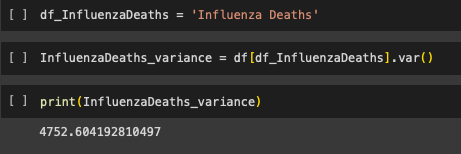


**Report:**

The 'Provisional\_COVID-19\_Deaths\_by\_Sex\_and\_Age' dataset shows a range score of 12621.0 for Influenza Deaths. This range, representing the difference between the maximum and minimum values, highlights the variability in Influenza Deaths across different demographic categories, underscoring the dataset's diverse nature.

**Task22:**

To calculate the variance of Influenza Death spread of the dataset and report.

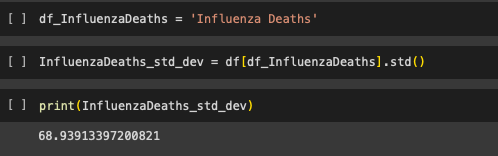


**Report:**

The 'Provisional\_COVID-19\_Deaths\_by\_Sex\_and\_Age' dataset indicates a variance score of 4752.60 for Influenza Deaths. This statistical measure quantifies the spread or dispersion of Influenza Deaths data, providing valuable insights into the dataset's variability across diverse demographic categories.

**Task23:**

To calculate the Standard Deviation of Influenza Death , Deviation score the dataset and report.

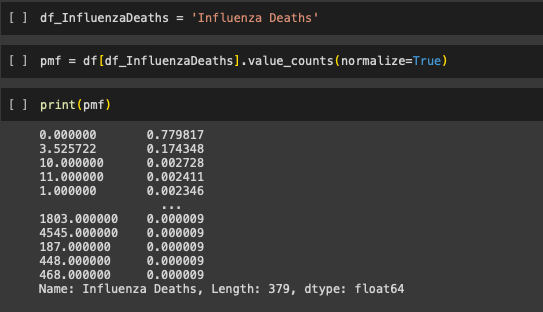


**Report:**

For the 'Provisional\_COVID-19\_Deaths\_by\_Sex\_and\_Age' dataset, the Standard Deviation score is 68.94 for Influenza Deaths. This measure of dispersion showcases the extent to which Influenza Deaths deviate from the mean, providing a understanding of the dataset's variability across demographic categories.

**Task24:**

To calculate the PMF of Influenza Death , the mass function value and report.

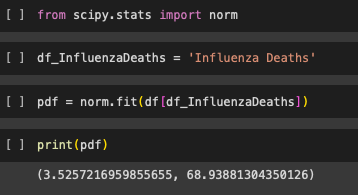


**Report:**

The probability mass function for Influenza Deaths in the 'Provisional\_COVID-19\_Deaths\_by\_Sex\_and\_Age' dataset provides insights into the likelihood of different death counts. The probabilities range from 0.779817 for a count of 0 to smaller probabilities for higher counts. This distribution reveals the statistical likelihood of encountering specific Influenza Death counts, contributing to a comprehensive understanding of the dataset's variability and potential patterns. It serves as a valuable tool for assessing the probability of observing different outcomes related to Influenza Deaths within diverse demographic categories.

**Task25:**

To calculate the PDF of Influenza Death , the distribution influenza death function values and report.

****

The probability density function (PDF) for 'Influenza Deaths' in the 'Provisional\_COVID-19\_Deaths\_by\_Sex\_and\_Age' dataset is defined by a mean of 3.53 and a standard deviation of 68.94. This PDF provides a statistical distribution of Influenza Deaths, indicating that the majority of values are expected to lie within one standard deviation of the mean. The mean and standard deviation serve as central parameters, offering valuable insights into the dataset's central tendency and spread. Understanding the PDF aids in gauging the likelihood of observing specific Influenza Death values, contributing to interpretation of the dataset.

**References:**

1. **<https://github.com/pydeveloperashish/BigMart-Sales-Prediction-With-Deployment/blob/main/BigMart%20Sales%20Prediction%20-%20Updated.ipynb>**
2. **<https://github.com/TahsinNakibTalukder/Python-Sales-data-analysis-using-pandas/blob/main/Sales%20analysis.ipynb>**
3. **Data\_Analytics\_lecture\_note <https://moodle.roehampton.ac.uk/mod/resource/view.php?id=1568523>**
4. **Machine\_Learning\_Lecture\_Note <https://moodle.roehampton.ac.uk/course/view.php?id=15797>**
5. **Mathematics for Data Science Lecture Note**