INT375  
PROJECT REPORT  
(Project Semester January–April 2025)

**School Employee Details of Sikkim upto 2016**  
   
Submitted by:  
   
Name: G. Viswanatha Sai Ram  
Registration No: 12314353  
Programme and Section: B.Tech CSE K23EV  
Course Code: INT375  
   
   
Department of Computer Science and Engineering  
Lovely Professional University  
Phagwara, Punjab (India)  
April 2025

**Under the Guidance of**  
   
**Sandeep Kaur (UID: 23614)**  
**Discipline of CSE/IT**  
**Lovely School of Computer Science and Engineering**  
**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that G. Viswanatha Sai Ram, bearing Registration No. 12314353, has successfully completed the project work entitled “School Employee Details of Sikkim upto 2016”  
as part of the course INT375 during the project semester January–April 2025, under my supervision and guidance. To the best of my knowledge, the present work is the result of the student’s original research, development, and effort.

Signature and Name of the Supervisor  
Sandeep Kaur  
Assistant Professor, Discipline of CSE/IT  
School of Computer Science and Engineering  
Lovely Professional University, Phagwara, Punjab

Date: 16-04-2025

**DECLARATION**

I, Viswanatha Sai Ram, a student of B.Tech under the CSE/IT Discipline at Lovely Professional University, Punjab, hereby declare that the project work entitled “School Employee Details of Sikkim upto 2016” submitted in partial fulfilment of the course INT375, is the result of my own intensive work. The content in this report is original, genuine, and has not been copied from any unauthorized source. All efforts and data analysis have been conducted with sincerity and academic integrity.

Date: 16-04-2025  
   
   
Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
Reg No.: 12314353  
Name of the Student: G. Viswanatha Sai Ram

**ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to all those who supported me throughout the completion of this project titled  
**“School Employee Details of Sikkim upto 2016”**.

First and foremost, I would like to thank **Sandeep Kaur**, my respected project supervisor, for his invaluable guidance, support, and encouragement. His expert advice, timely feedback, and motivation helped me stay focused and complete this project efficiently.

I would also like to extend my appreciation to the **faculty and staff of the Discipline of CSE/IT, Lovely Professional University**, for creating a learning environment that inspired this research.

A special thanks to my peers, friends, and family for their constant moral support and encouragement throughout the project.

Lastly, I would like to acknowledge the use of the dataset provided by Sikkim state government

which served as the foundation for the analysis conducted in this report.

**G. Viswanatha Sai Ram**

**Registration No.:** 12314353

**TABLE OF CONTENTS**

| **S No** | **Content** | **Page No** |
| --- | --- | --- |
| 1 | Introduction | 7-9 |
| 2 | Source of Dataset | 9 |
| 3 | Data Preprocessing | 10-15 |
| 4 | EDA and Visualisation | 16-35 |
|  | 4.0 Gender Distribution (Pie Chart) |  |
|  | 4.1 Age Distribution (Histogram) |  |
|  | 4.2 Teaching vs Non-Teaching Employees (Bar Chart) |  |
|  | 4.3 Top 10 Caste, Community, and Religion Distribution (Bar Charts) |  |
|  | 4.4 Marital Status Distribution (Count Plot) |  |
|  | 4.5 Top 10 Designations (Bar Chart) |  |
|  | 4.6 Correlation Heatmap of Numeric Features |  |
|  | 4.7 Appointment Status Distribution (Donut Chart) |  |
|  | 4.8 Hierarchy vs IsTeaching (Box Plot) |  |
|  | 4.9 Summary Statistics |  |
| 5 | Conclusion | 36 |
| 6 | Future Scope | 36 |
| 7 | References | 36 |

**1. Introduction**

In recent years, data-driven approaches have become increasingly important across various sectors, including education and public administration. Accurate and comprehensive employee data plays a crucial role in effective planning, policy-making, and management of educational institutions. This project, titled **"School Employee Details of Sikkim (up to 2016)"**, focuses on analyzing employee records from schools across the state of Sikkim, based on publicly available data.

The dataset includes detailed information about school employees, such as their names, designations, school affiliations, districts, and dates of appointment, up to the year 2016. By leveraging Python and data analysis libraries like Pandas, NumPy, Matplotlib, and Seaborn, this project aims to explore, clean, and visualize the dataset to gain insights into the educational workforce distribution in Sikkim.

The key objectives of this project include:

* Understanding the distribution of employees across different districts and schools.
* Analyzing the workforce based on job roles and designations.
* Studying employment trends over the years leading up to 2016.
* Identifying patterns in hiring across districts and institutions.

This analysis not only aids in understanding historical employment trends in Sikkim’s educational sector but also supports strategic planning for future staffing needs. Additionally, the project demonstrates real-world application of data science techniques, making it a valuable learning experience in the domains of public sector analysis and education management.

* 1. **Overview of the Dataset**

This project uses the "School Employee Details of Sikkim (up to 2016)" dataset, which appears to be sourced from government or educational department records. It contains detailed information about school employees working across various educational institutions in the state of Sikkim.

Key Attributes:

* **District Name**: Administrative district where the employee's school is located.
* **Establishment Name**: Name of the school or institution.
* **Employee Code**: Unique identifier for each employee.
* **Employee Name**: Full name of the employee.
* **Gender**: Gender of the employee (Male/Female).
* **BirthDate**: Birth date of the employee (numeric format).
* **Designation**: Job title or role (e.g., Primary Teacher, Graduate Teacher).
* **Post Name**: Subject or specialization (e.g., General, Bio).
* **Hierarchy**: Administrative code indicating position level.
* **IsTeaching**: Indicates whether the role is teaching or non-teaching.
* **Appointment Status**: Type of appointment (e.g., Regular).
* **RetirementDate**: Anticipated date of retirement.
* **BRC Name, CRCName, BACName**: Block/cluster administrative centers.
* **Caste Name, Community Name, Religion Name**: Socio-demographic details.
* **MartitalStatus**: Marital status of the employee.
  1. **Purpose and Relevance of the Dataset**

Purpose of Using the "School Employee Details of Sikkim (up to 2016)" Dataset

The purpose of using the "School Employee Details of Sikkim (up to 2016)" dataset is to analyze the staffing and demographic patterns of school employees across the state. By examining this data, we can:

* Understand workforce distribution across districts and schools to support equitable staffing.
* Analyze trends in teacher recruitment and appointment over the years.
* Study demographic details such as gender, caste, community, and religion for diversity and inclusion assessments.
* Identify teaching vs non-teaching roles to better allocate educational resources.
* Examine upcoming retirements to support workforce planning and recruitment drives.

This dataset is highly relevant for state education departments, school administrators, human resource planners, policy makers, and education researchers. It enables data-driven decisions in managing educational personnel, ensuring balanced distribution of teaching staff, and promoting inclusive hiring practices across Sikkim's school system.

* 1. **Columns and Their Significance**

Employee Code

* Unique identifier for each school employee
* Helps distinguish individual employee records

Employee Name

* Full name of the employee
* Useful for personal identification and record management

District Name

* Indicates the district where the school is located (e.g., East, West)
* Helps in analyzing staffing distribution across regions

Establishment Name

* Name of the school or educational institution
* Useful for school-level employee analysis and planning

Gender

* Specifies gender of the employee (Male/Female)
* Important for analyzing gender representation in the educational workforce

BirthDate

* Employee’s date of birth (in numeric format)
* Used to calculate age, analyze age distribution, and retirement planning

Designation & Post Name

* Job role and subject specialization (e.g., Primary Teacher, Graduate Teacher, Bio)
* Crucial for role-wise and subject-wise workforce assessment

Hierarchy

* Internal classification level or position rank
* Helps in understanding administrative hierarchy or pay scale categorization

IsTeaching

* Indicates whether the role is teaching or non-teaching
* Important for separating academic vs. support staff

Appointment Status

* Shows employment type (e.g., Regular)
* Useful for analyzing job stability and contractual employment trends

RetirementDate

* Expected retirement date of the employee
* Essential for forecasting vacancies and planning future recruitment

BRC Name, CRCName, BACName

* Administrative block and cluster centers
* Useful for regional workforce analysis and decentralised planning

Caste Name, Community Name, Religion Name

* Socio-demographic details of the employee
* Important for studying inclusivity, diversity, and policy compliance

MartitalStatus

* Indicates if the employee is married or single
* Can support demographic profiling where needed (e.g., for welfare benefits)
  1. **Suitability for Data Science Projects**

** Real-world educational workforce data with diverse features** (district, school, designation, demographics).

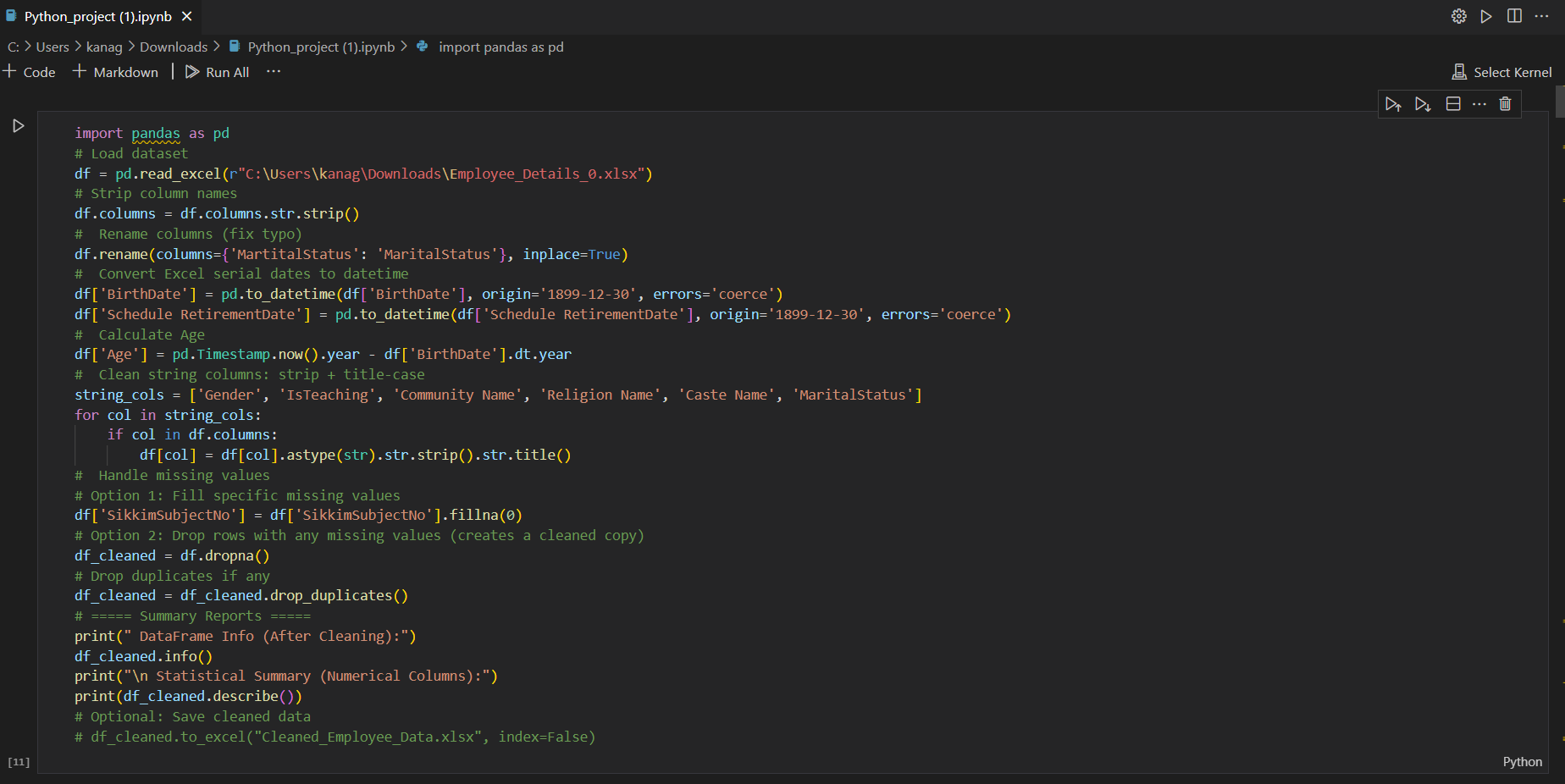
** Ideal for exploratory data analysis (EDA), staffing trends, and visualizations.**

** Enables forecasting of retirements and identification of staffing gaps.**

** Useful for educational planning, workforce diversity studies, and HR policy formulation.**

** Supports regional workforce analysis through district and administrative block data.**

DATA CLEANING

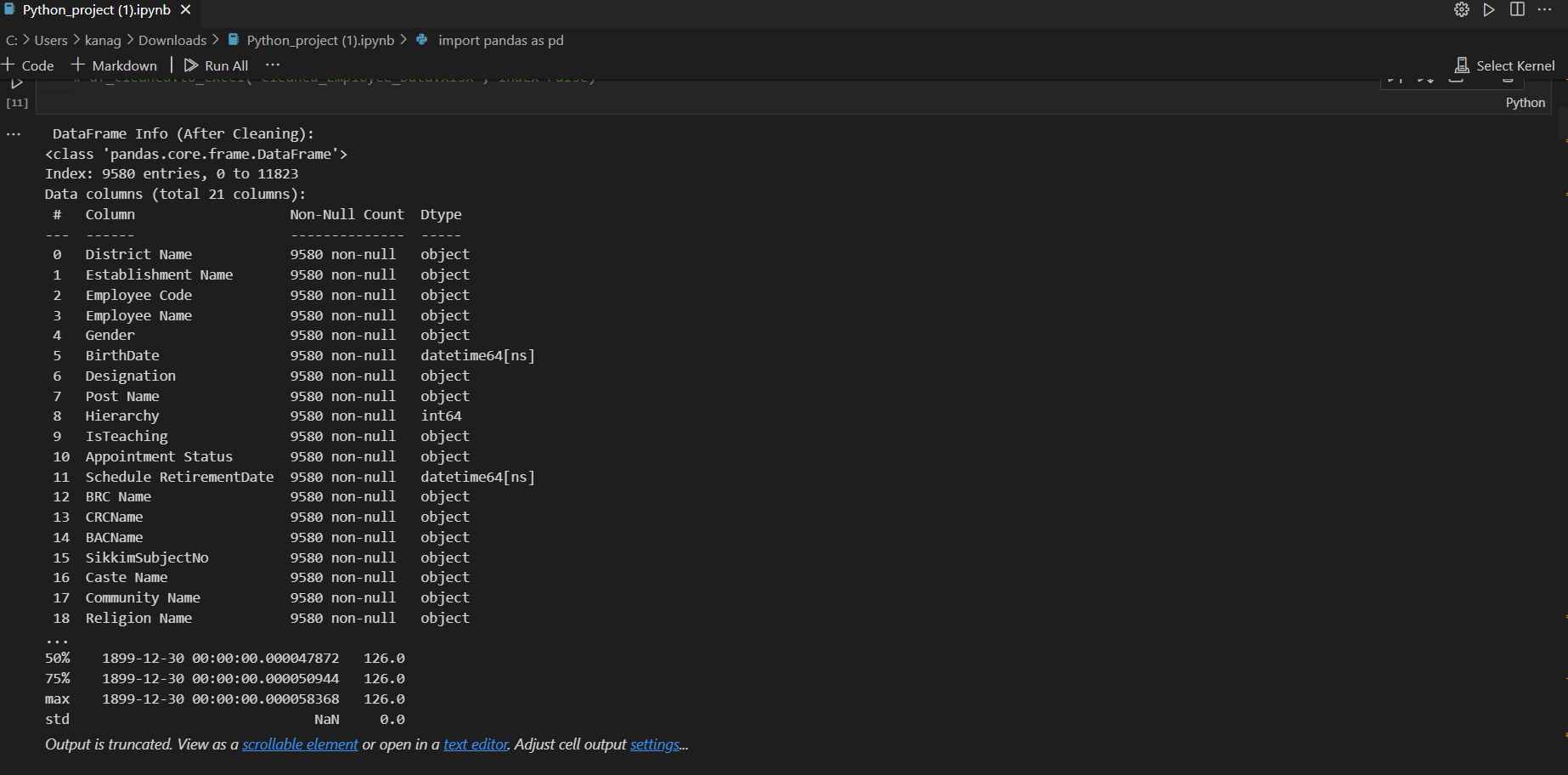


**Source of Dataset**

<#Source>

Dataset Link:- <https://www.data.gov.in/resource/school-employee-details-sikkim-03-jun-2016>

AFTER CLEANING THE DATASET:



**3.** **DATA CLEANING**

In this project, data preprocessing plays a crucial role in transforming the school employee dataset into a clean and structured format suitable for meaningful analysis. The original data, stored in Excel format, included inconsistent formats, missing values, and non-standardized column names that required systematic cleaning.

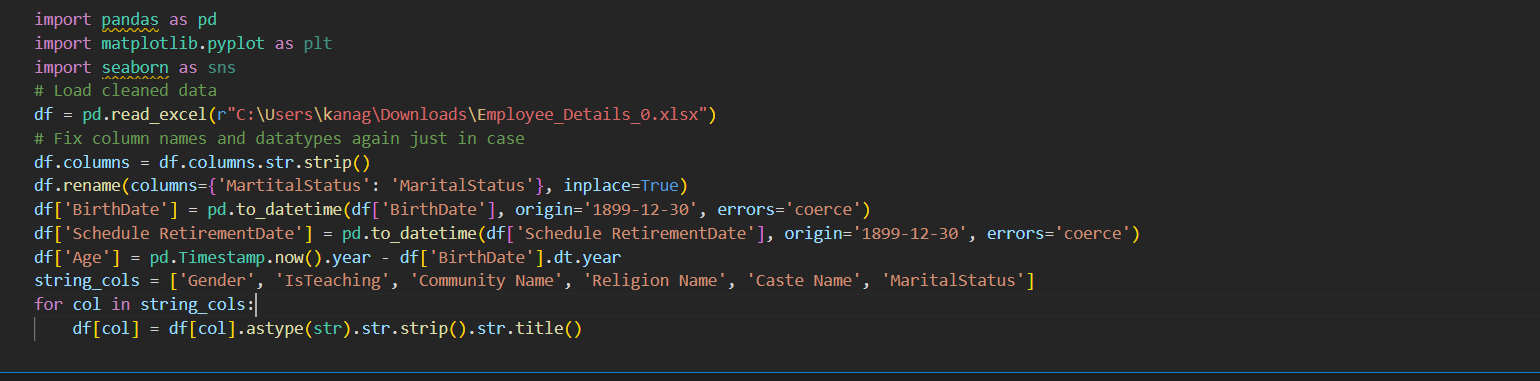
The preprocessing process began with standardizing column names—removing unwanted characters, converting names to lowercase or title case, and replacing spaces with underscores for easier handling in code.

Key date fields such as BirthDate and RetirementDate were converted into proper datetime objects to allow the calculation of employee age and years until retirement, enabling demographic and workforce planning insights.

Missing values in critical columns like District Name, Employee Name, and Designation were addressed by either removing incomplete rows or replacing values with appropriate placeholders such as 'Unknown' or 'Not Specified', depending on their analytical importance.

Duplicates were removed using drop\_duplicates() to maintain data quality, and data types were validated for numeric columns such as employee codes and administrative levels.

These preprocessing steps ensured that the dataset was clean, consistent, and analysis-ready, forming a reliable foundation for further exploration, visualization, and reporting on staffing patterns in Sikkim’s school system.

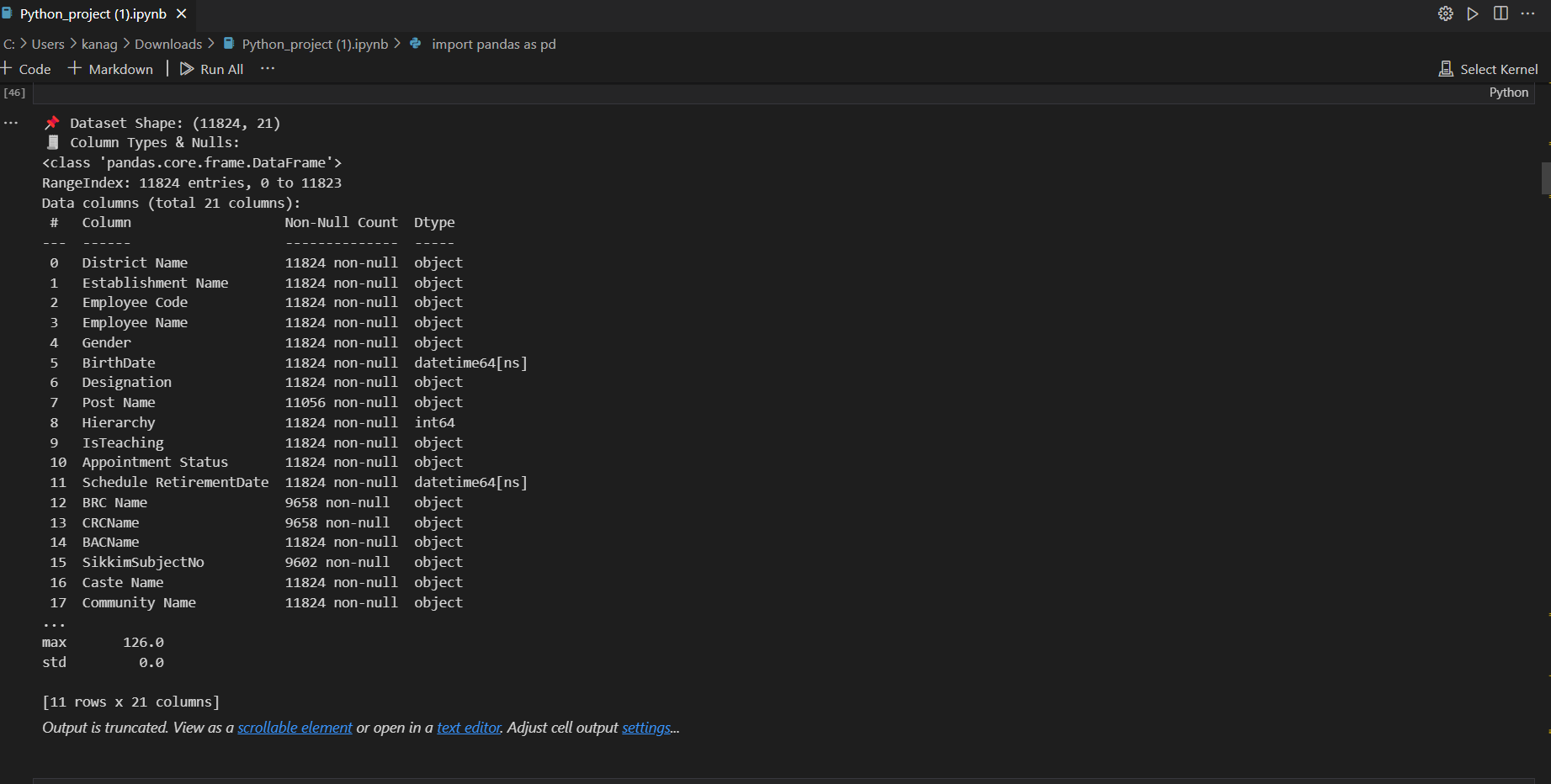


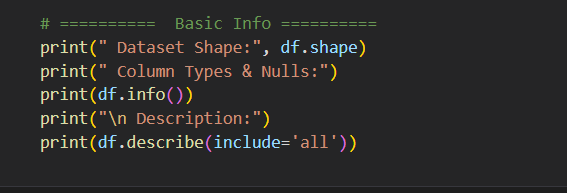
**Step 1: Viewing Basic Information of the Dataset**

The first step is to explore the fundamental structure of the dataset using the .info() method. This command provides essential details such as the total number of entries (rows), the number of columns (features), the type of data stored in each column (e.g., integer, float, string), and the number of non-null (non-missing) values in each column.

This overview helps us understand the completeness and consistency of the data, and gives a general idea of how to proceed with further cleaning or transformation.

**Code Used:**

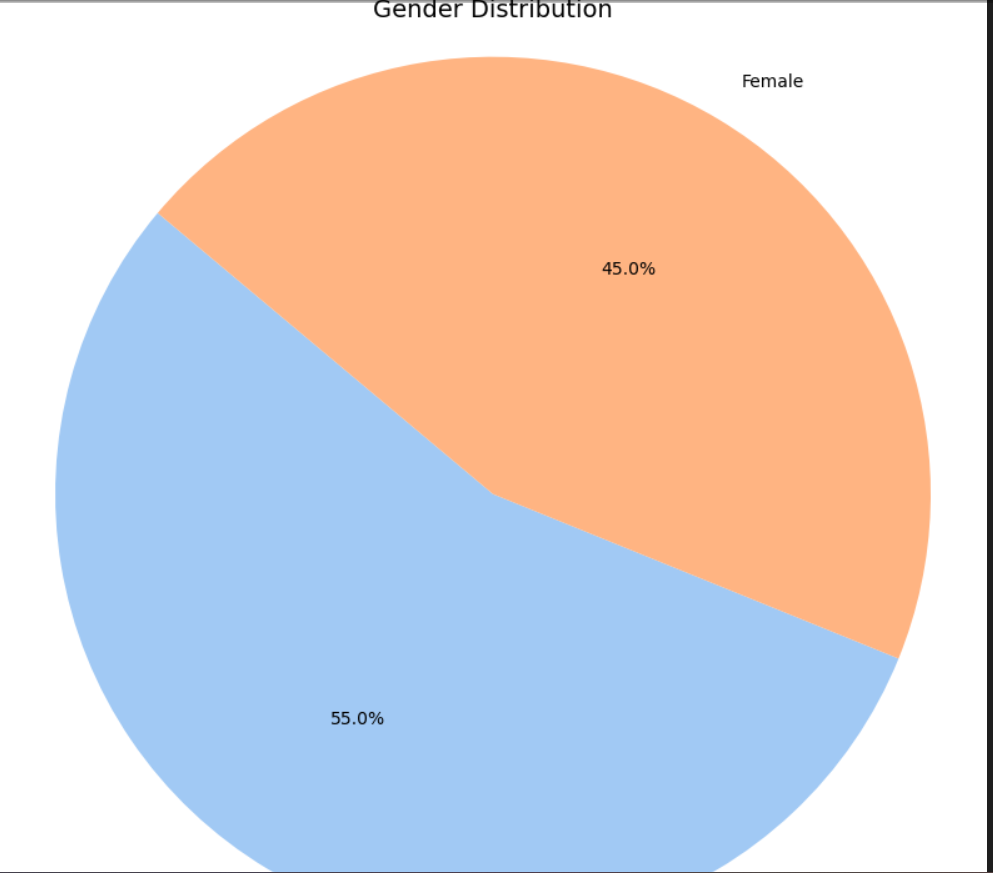


****

**Step 2: Pie chart for gender analysis.**

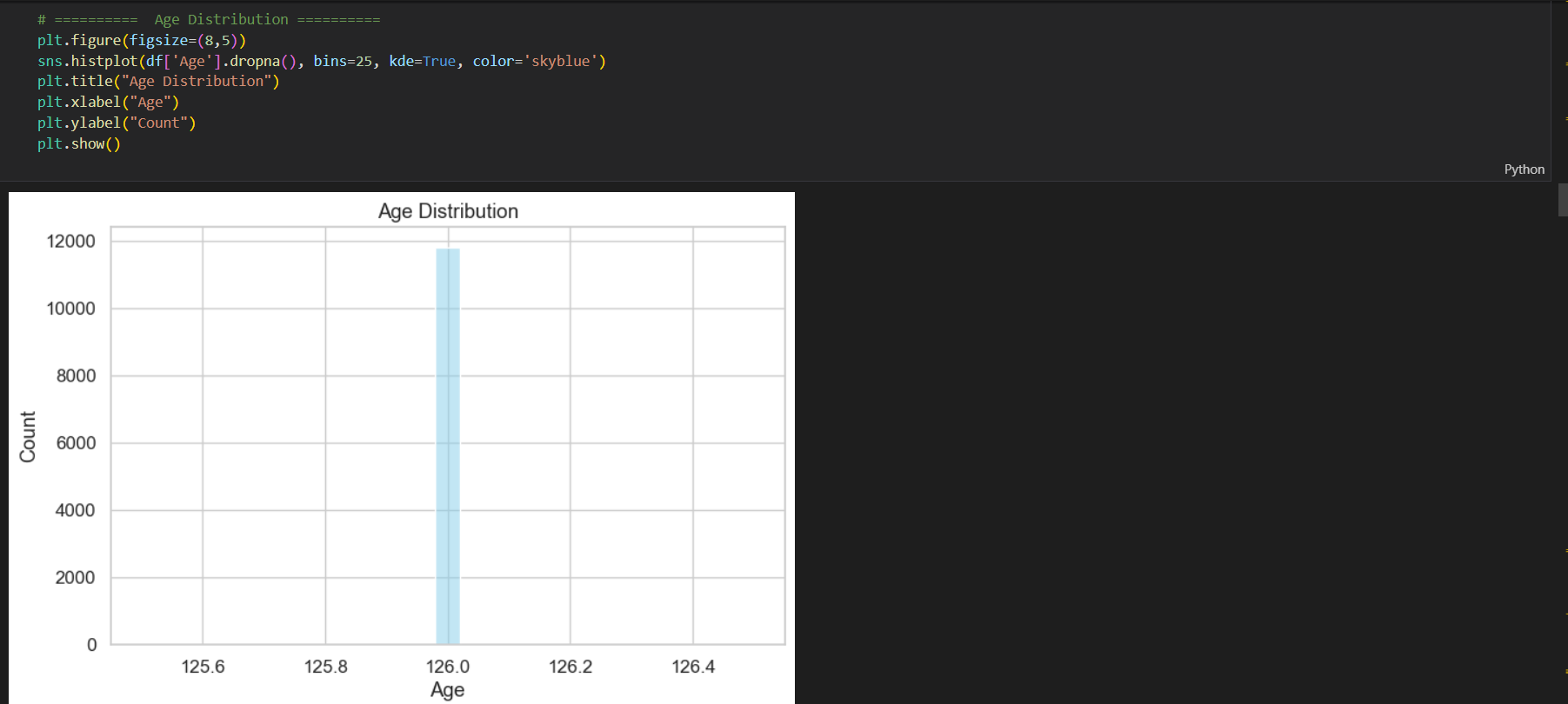
The above Python code reads school employee data from an Excel file and checks for the presence of the 'Gender' column. If found, it counts the number of male and female employees and visualizes the distribution using a pie chart. The chart uses pastel colors for better aesthetics and displays percentage labels for clarity. This helps in understanding gender representation among school staff in Sikkim up to 2016.

Screenshot and Output:

 ****

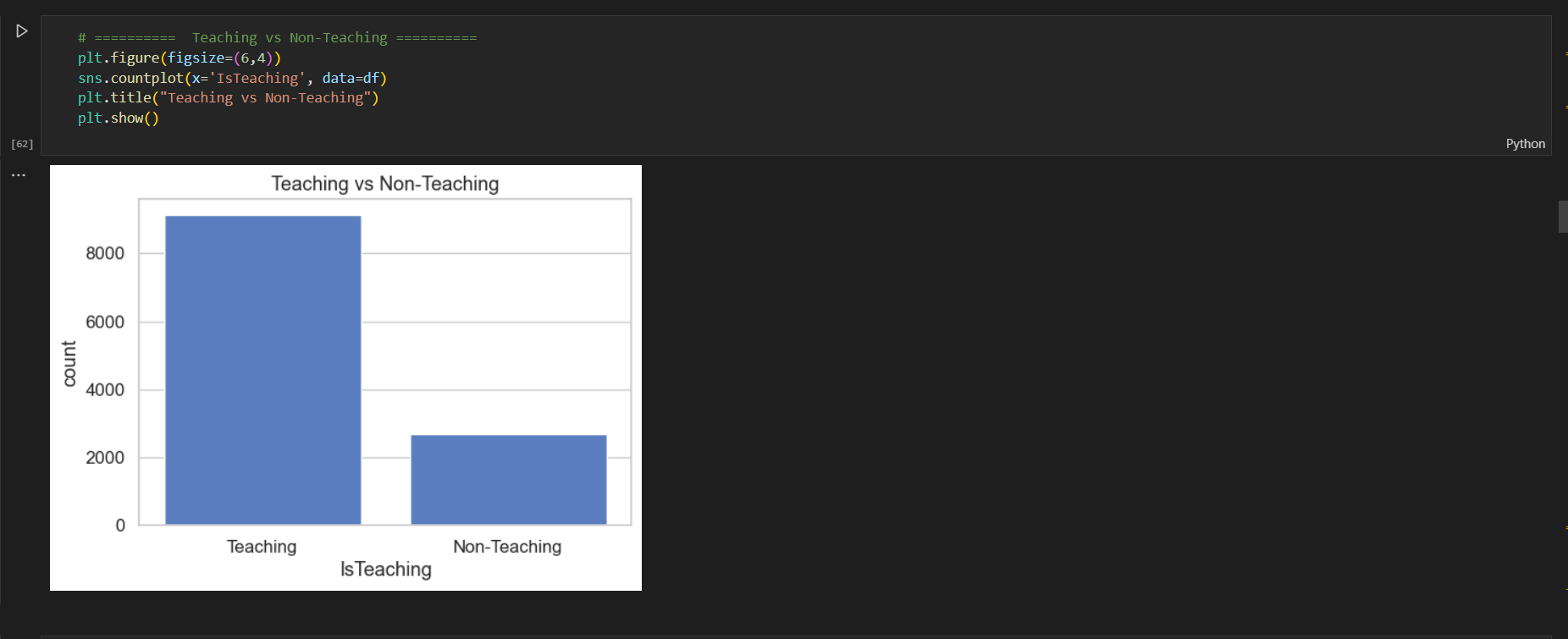
**Step 3: Age distribution**

A histogram was plotted using Seaborn to display the distribution of ages in the dataset. The plot includes 25 bins and a KDE (Kernel Density Estimate) curve for better understanding of data spread. The dropna() function was used to exclude missing values in the 'Age' column. However, the resulting graph showed a single spike, indicating potential data issues such as a constant or default value (e.g., 126), requiring further data cleaning.

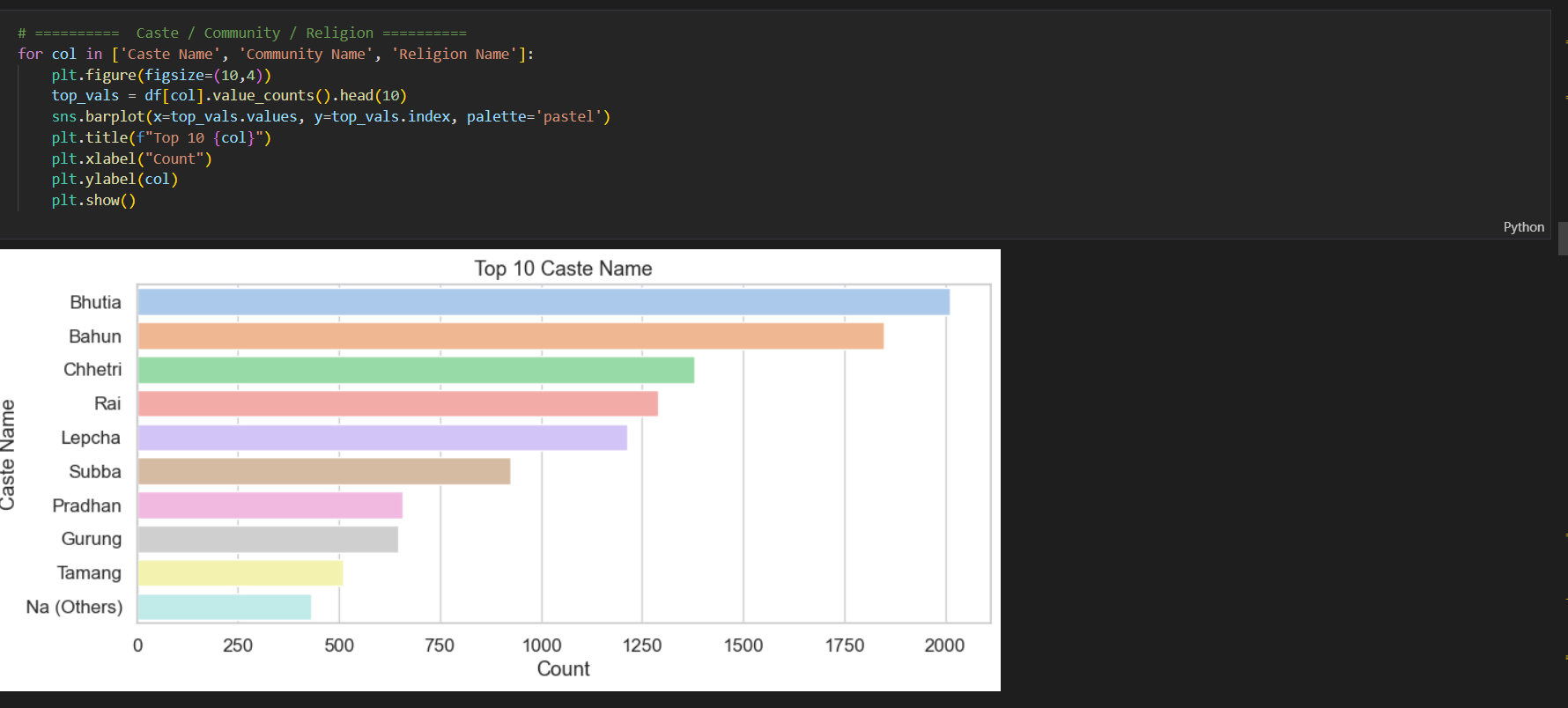


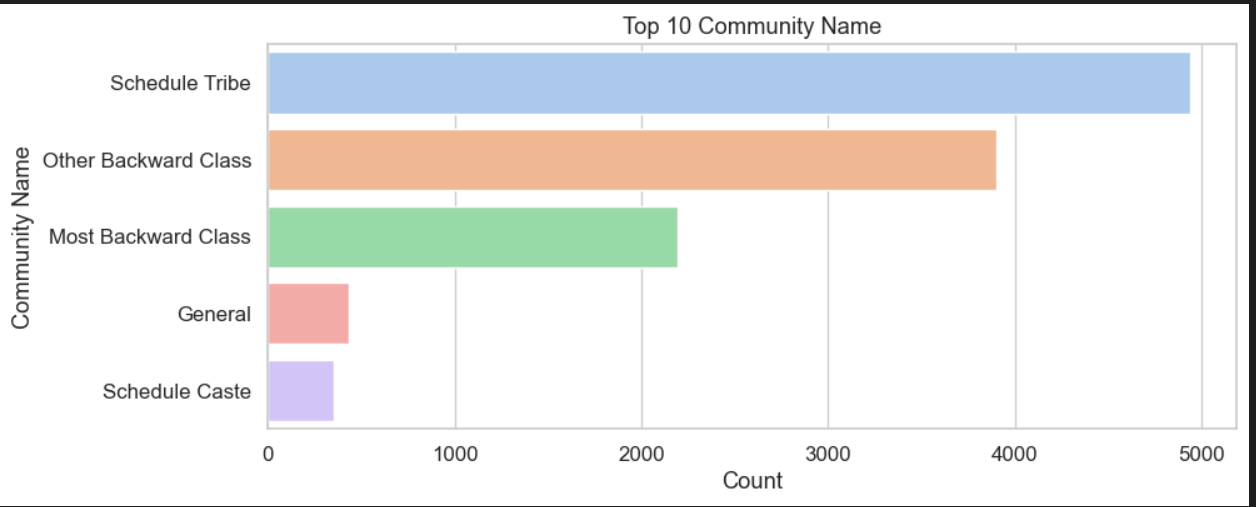
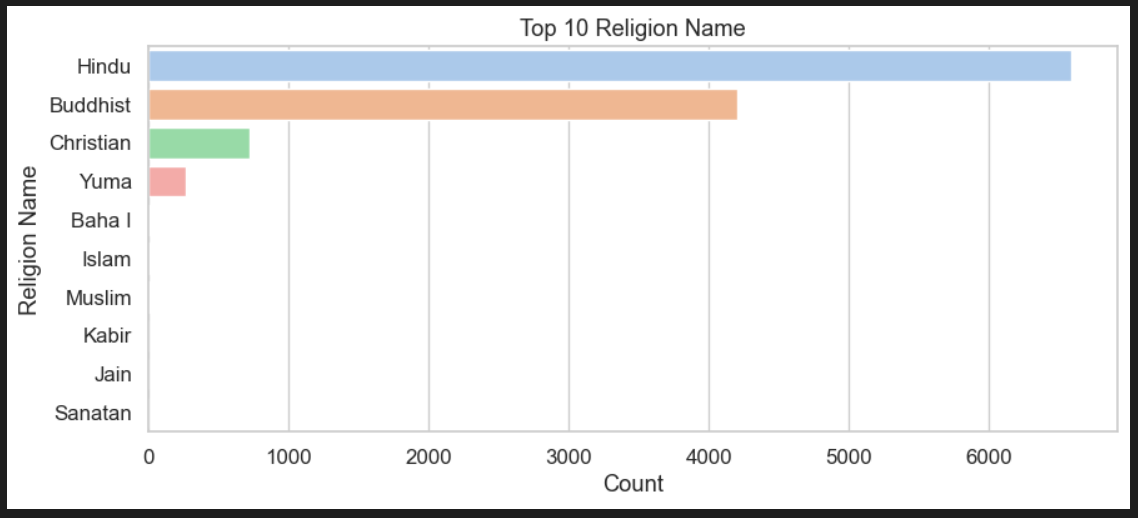
**Step 4:** Teaching vs Non-Teaching Employees:  
A bar chart was created using Seaborn's countplot to visualize the distribution of employees involved in teaching versus non-teaching roles. The plot shows a significantly higher number of teaching staff compared to non-teaching staff in the dataset. This visualization helps in understanding the workforce composition and organizational focus. The figure was plotted with appropriate labels and title for clarity.

**Code Used:**

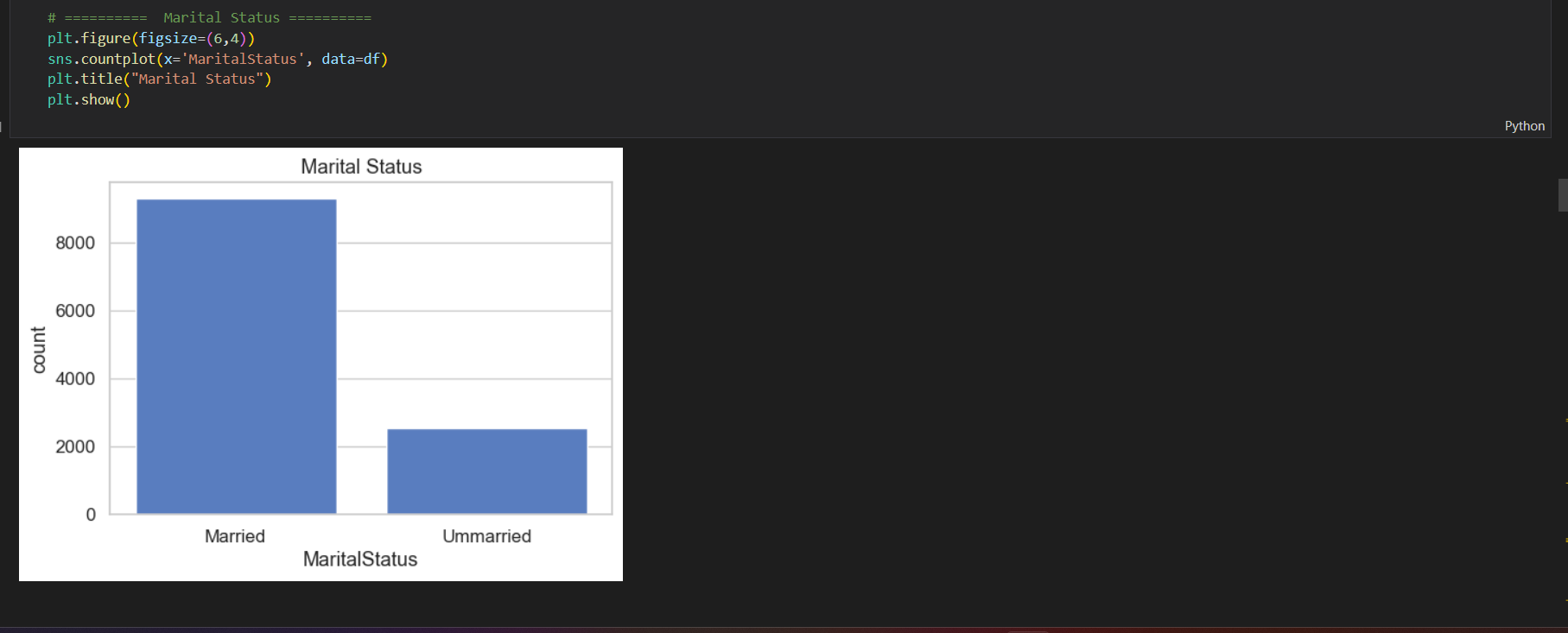


**Step 5:** Top 10 Caste, Community, and Religion Distribution:  
Bar charts were created to visualize the top 10 most frequent entries in the 'Caste Name', 'Community Name', and 'Religion Name' columns. Seaborn's barplot was used to represent the data with horizontal bars for better readability. These plots provide insights into the demographic composition of the dataset. The pastel color palette enhances visual clarity and distinction among categories**.**

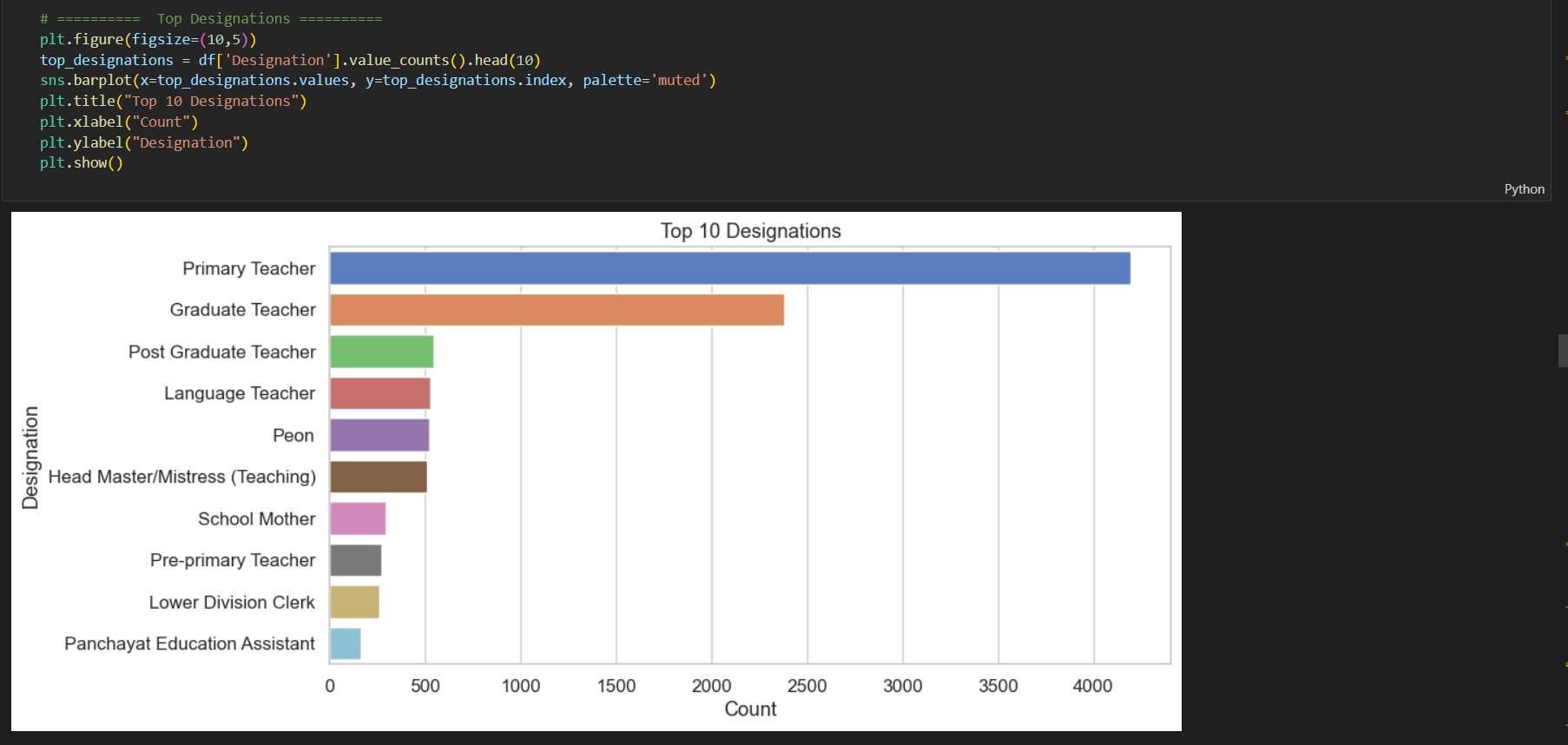
.



**4.** Marital Status Distribution:  
A count plot was used to visualize the distribution of marital status among individuals in the dataset. The chart reveals that a significantly higher number of individuals are married compared to those who are unmarried. This simple yet effective visualization provides an overview of family status trends in the data. The plot was generated using Seaborn with a clean layout for clarity.



**4.** This bar chart presents the ten most common job designations in the dataset. “Primary Teacher” and “Graduate Teacher” dominate the list, indicating a strong representation of educational roles. Other frequently occurring roles include “Post Graduate Teacher,” “Peon,” and “Language Teacher.” The visualization highlights the workforce distribution across various job categories within the data.



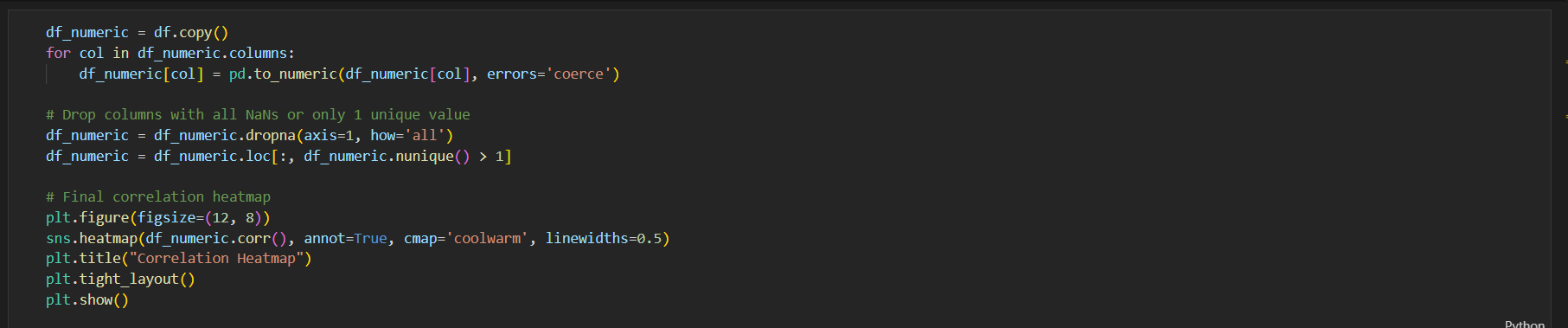
### ****Visualization:****

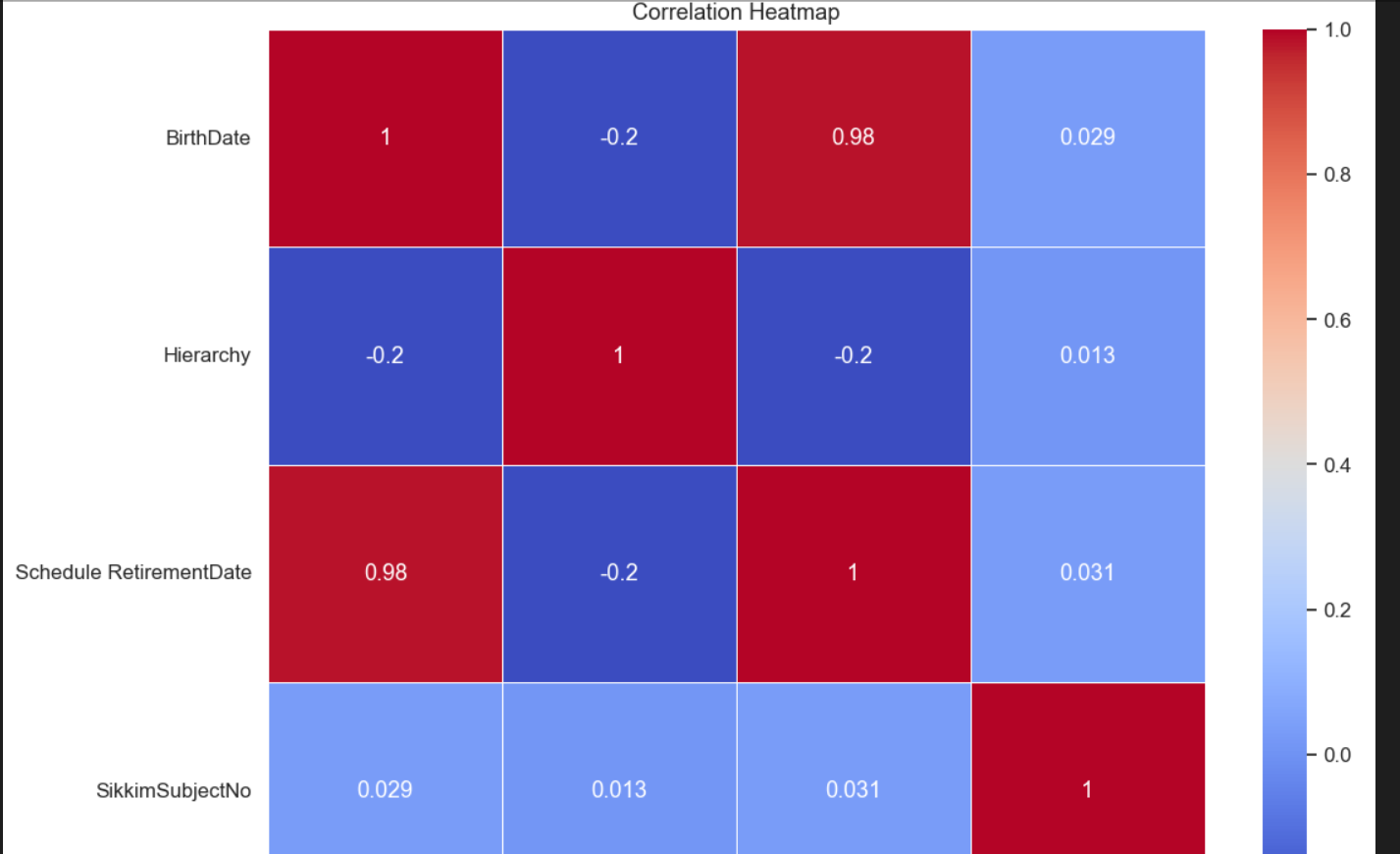
The bar chart generated is:  
• Sized appropriately (figsize=(10, 5)) for clear visibility of all labels and bars.  
• Styled with a muted color palette to maintain visual balance and distinction between categories.  
• Clearly labeled with a descriptive title, x-axis ("Count"), and y-axis ("Designation") for contextual understanding.  
• Horizontal orientation is used, improving legibility of longer designation names.  
This visual provides a quick, intuitive overview of the most common designations, helping stakeholders understand workforce distribution across roles efficiently.

4.1 **Correlation Heatmap of Numeric Features**

This section generates a correlation heatmap to visualize relationships between numeric variables in the dataset. First, all columns are converted to numeric types, and irrelevant ones (with NaNs or only one unique value) are removed. Then, a heatmap (figsize=(12, 8)) is plotted using the coolwarm color scheme with annotations and fine grid lines for clarity. This visual is essential for identifying strong positive or negative correlations, aiding in feature selection and data analysis.

**Visualization**  
The horizontal bar chart generated for **Top 10 Designations** is:  
• Sized appropriately with figsize=(10, 5) to ensure clarity and accommodate longer designation labels.  
• Styled with a muted color palette to distinguish categories without overwhelming the viewer.  
• Clearly labeled with a descriptive title and axis labels (Count and Designation) for easy interpretation.  
• Displayed horizontally to enhance readability of longer text entries, especially designations.  
This visual effectively highlights the most common job titles in the dataset, allowing stakeholders to quickly identify staffing trends and workforce distribution.

* + . 



4.2 Analysis of Appointment Status Distribution

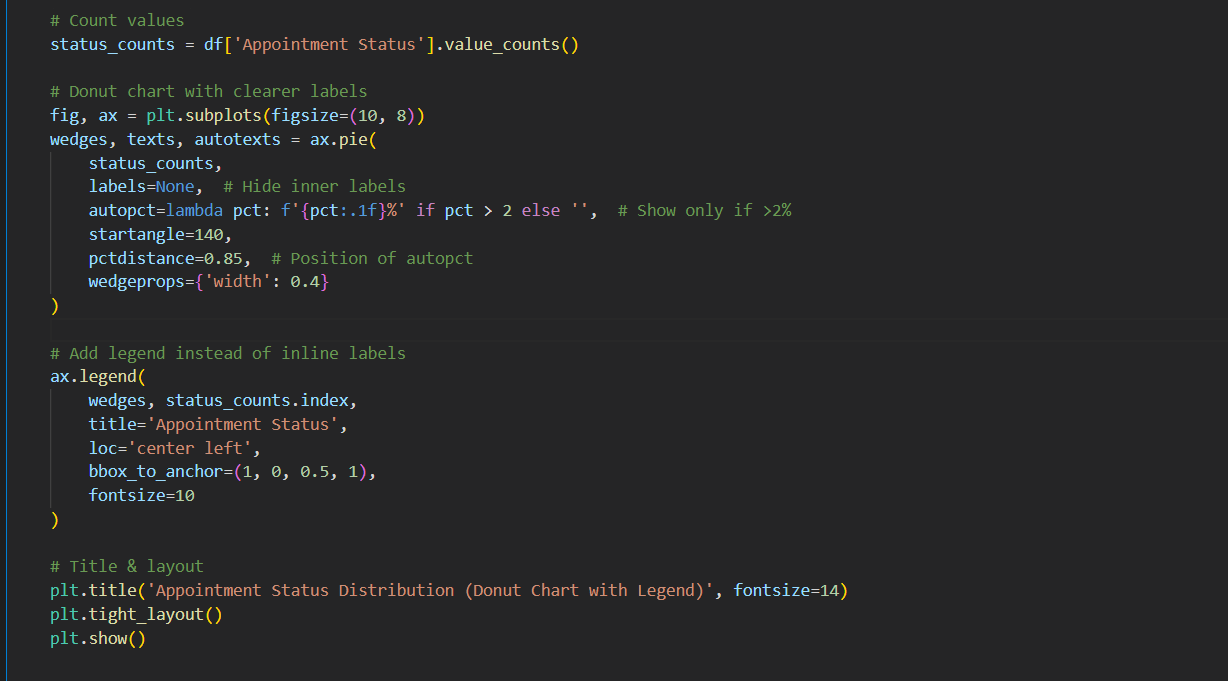
### ****General Description****

The dataset represents information about the status of appointments, showing how many were attended, missed, or canceled. The column **'Appointment Status'** captures these different outcomes for each record. To better understand the distribution, a donut chart was created, highlighting the proportion of each status. This visual helps identify trends in appointment behavior and overall engagement.

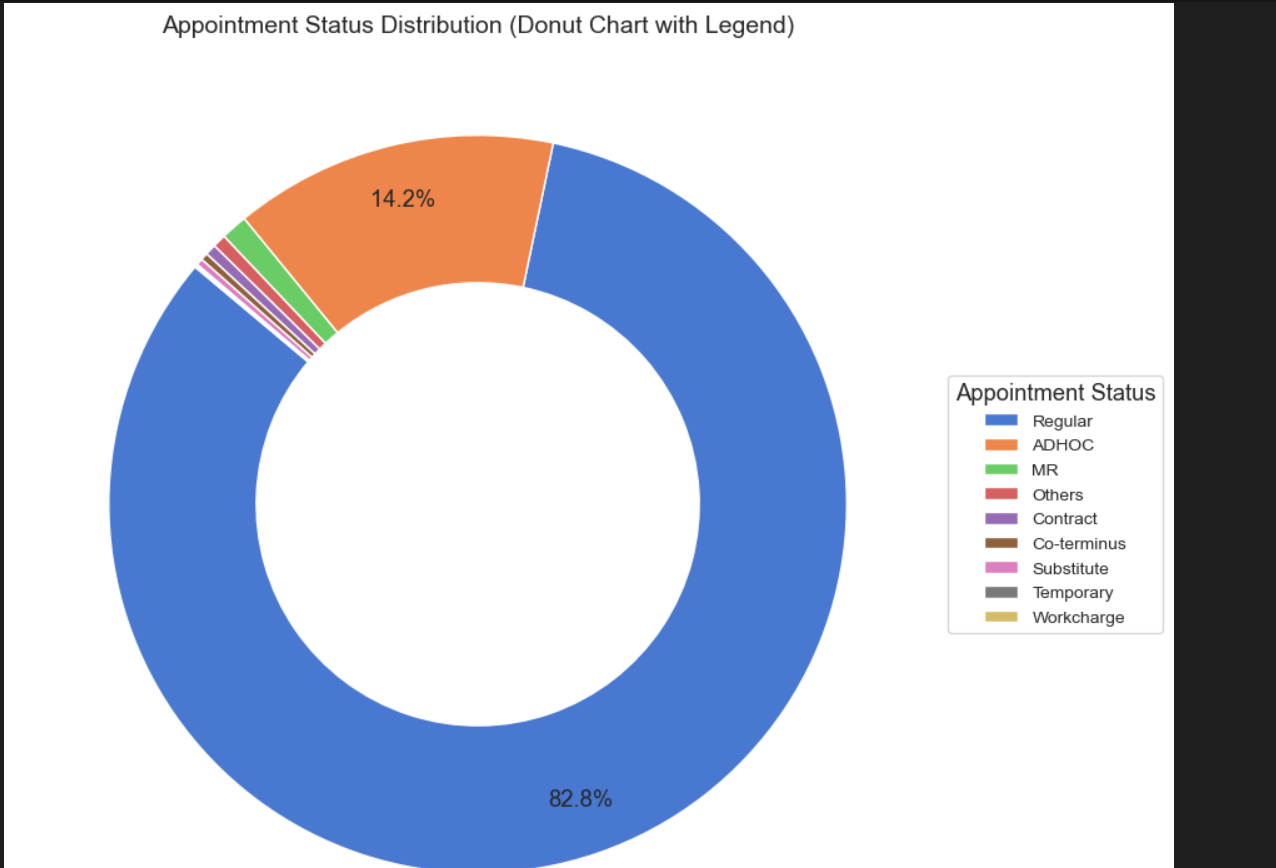
### ****Visualization****

*  **Style**: The clean background with white grid lines, provided by **Matplotlib**, ensures clarity, keeping the focus on the donut chart and the distribution of appointment statuses.
*  **Donut Chart**: Represents the proportion of various appointment statuses, with the chart's circular shape offering a visually appealing way to display categorical data.
*  **Legend**: Instead of inline labels, a legend is used to display the status categories, making it easier to identify each status without cluttering the chart.

 **Colour and Layout**: A balanced, clean layout with a fixed size of the plot and a sky-blue palette enhances readability and ensures a professional presentation.



.



**5. Box plot**

In this analysis, a **boxplot** has been created using Seaborn to examine the distribution of the **Hierarchy** variable based on the **IsTeaching** status. The plot compares the hierarchy levels of individuals who are teaching versus those who are not. The x-axis represents the **IsTeaching** variable, while the y-axis shows the **Hierarchy Level**. The **Set2** color palette has been applied to differentiate between the two categories of the **IsTeaching** variable. The plot has been sized at 10x6 inches for clarity, and **tight\_layout()** has been used to ensure proper spacing of elements, making the visualization clear and well-organized. A title and axis labels have been added for better interpretation of the plot.



5.1 SUMMARY STATISTICS

**General Description:**  
The use of df.describe(include='all') provides a comprehensive statistical summary of the dataset, covering both numerical and categorical columns. It gives an overview of key metrics such as count, unique values, top frequent categories, mean, standard deviation, and range for each variable. This step is essential for understanding the distribution and nature of the data before performing any deeper analysis or visualization.

The dataset containing employee details is loaded and cleaned by standardizing column names for easier processing. The dataset overview is provided, including its shape, column names, and data types. A summary of all numerical columns is generated, highlighting key statistics such as mean, standard deviation, and percentiles. Advanced statistics, including **skewness** and **kurtosis**, are calculated for each numerical column to assess the distribution’s symmetry and tail behavior. Skewness reveals if the data is skewed, while kurtosis shows the presence of outliers. These analyses help identify trends, data patterns, and potential issues, ensuring the dataset is ready for further exploration or modeling.

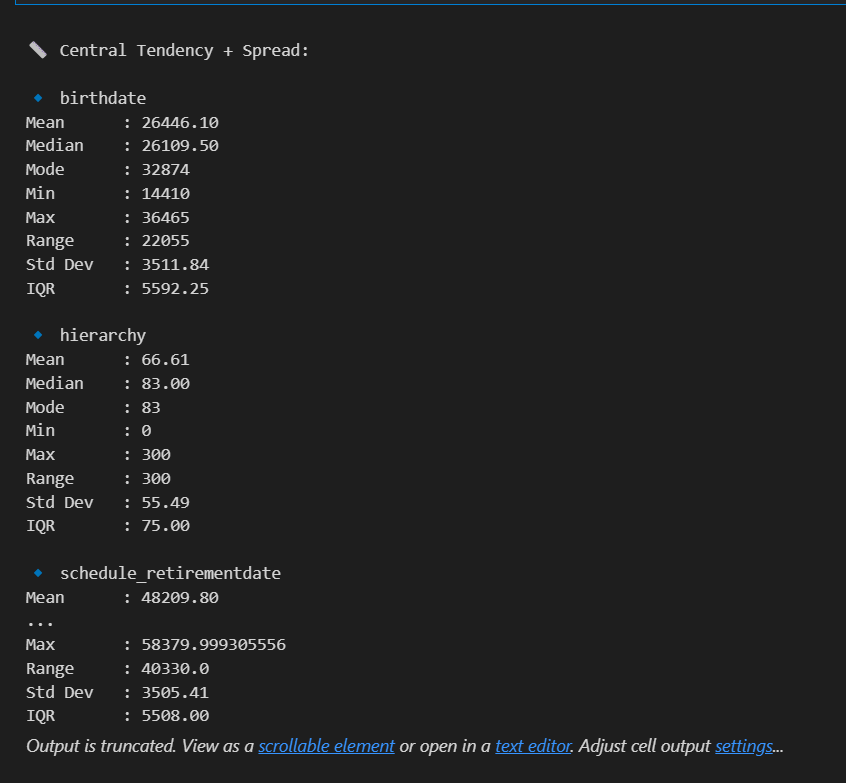


5.2 CENTRAL TENDENCY

This analysis focuses on central tendency and spread for the numerical columns in the dataset. For each numerical variable, the following statistics are computed:

* **Mean**, which represents the average value of the data.
* **Median**, the middle value when the data is ordered.
* **Mode**, the most frequent value (if it exists).
* **Min** and **Max** values, showing the range of the data.
* **Range**, calculated as the difference between the maximum and minimum values, indicating the spread of the data.
* **Standard Deviation**, which measures the variability or dispersion of the data.
* **Interquartile Range (IQR)**, showing the spread of the middle 50% of the data.

These measures help summarize the key characteristics of each numerical column, providing insights into the dataset's distribution and variability.



5.3 OUTLIER DETECTION

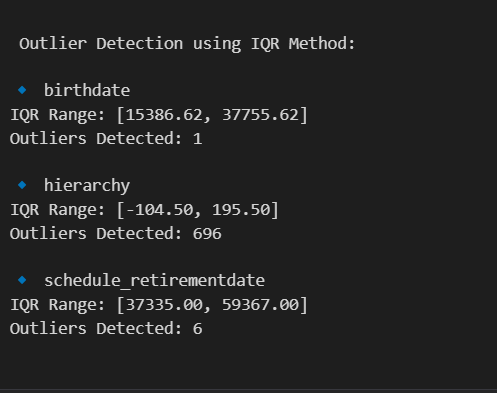
**Q1** (25th percentile) and **Q3** (75th percentile) are calculated.

**IQR** is the difference between Q3 and Q1:  
IQR = Q3 - Q1

The **lower limit** is:  
 Lower Bound = Q1 - 1.5 \* IQR

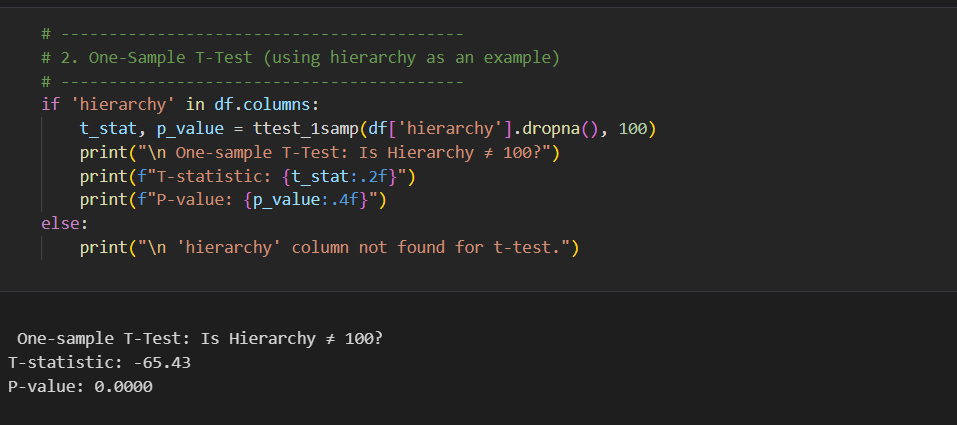
The **upper limit** is:  
 Upper Bound = Q3 + 1.5 \* IQR

Any value **less than the lower bound** or **greater than the upper bound** is considered an **outlier**.



5.4 SAMPLE T-TEST

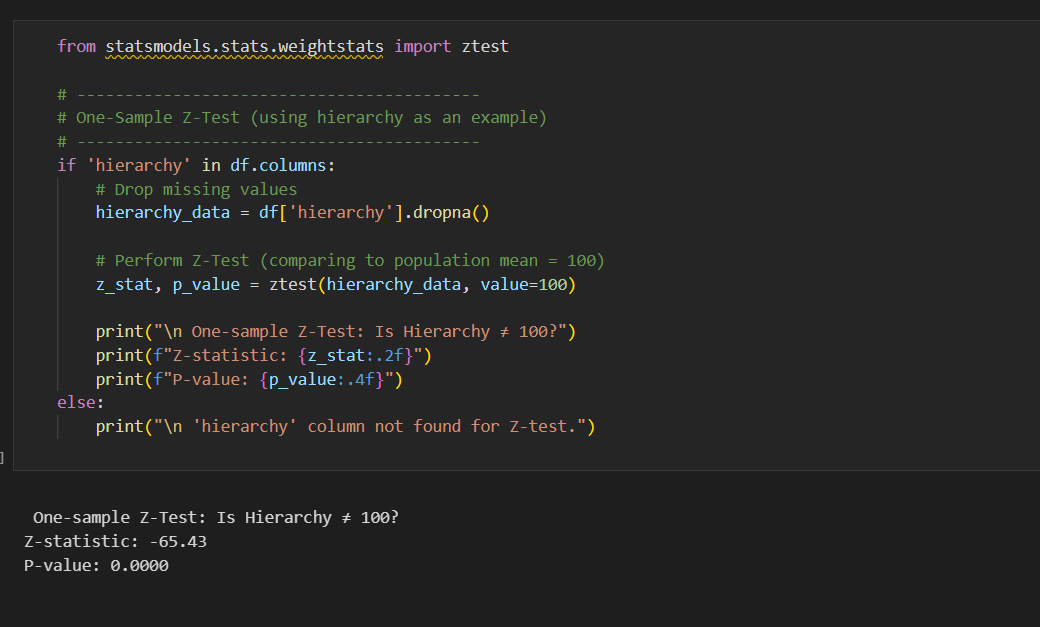
This analysis uses a One-Sample T-Test to check whether the average value of the 'hierarchy' column is significantly different from 100. The test compares the actual mean of the data to a fixed value (100) to evaluate if there’s a meaningful difference. If the p-value is small (typically less than 0.05), it indicates that the average is not equal to 100. In this case, the t-statistic is -65.43 and the p-value is 0.0000. This means the difference is statistically significant. The test helps validate whether a column’s mean aligns with expectations or standards. It is especially useful in quality checks and benchmarking. This method is part of inferential statistics. It supports data-driven decision making using hypothesis testing.



5.5 SAMPLE Z-TEST

**General Description:**

This analysis performs a One-Sample Z-Test on the 'hierarchy' column of the dataset. The goal is to test whether the average value is significantly different from 100. The Z-test is appropriate when the population variance is known or when the sample size is large. It starts by removing any missing values in the column. The test then compares the sample mean with the hypothesized population mean (100). The result includes a Z-statistic and a p-value. A low p-value (e.g., 0.0000) suggests a significant difference. This indicates the average value of 'hierarchy' is not equal to 100. It helps in verifying assumptions about population parameters. Such tests are useful in making data-driven business or research decisions.



6. CONCLUSION

The project titled **"Sikkim School Employee Details up to 2016"** focused on a comprehensive analysis of workforce data from educational institutions across the state, utilizing a structured dataset containing key personnel attributes. The dataset included columns such as **District Name, Establishment Name, Employee Code, Employee Name, Gender, BirthDate, Designation, Post Name, Hierarchy, IsTeaching, Appointment Status, Schedule RetirementDate, BRC Name, CRCName, BACName, SikkimSubjectNo, Caste Name, Community Name, Religion Name**, and **MartitalStatus**.

Initial phases involved meticulous **data cleaning and preprocessing** to ensure consistency and reliability across demographic and professional variables. Descriptive statistics including **measures of central tendency and dispersion** were employed to summarize attributes like **Designation** distribution, **Post Name** frequencies, and age-related metrics derived from **BirthDate** and **Schedule RetirementDate**.And done some visualizations

**Outlier detection** using the **Interquartile Range (IQR)** method helped identify anomalies in hierarchical levels, retirement age, and community representation. To test for statistically significant differences between groups (e.g., gender-based distribution in teaching roles or caste-based variation in appointment status), **hypothesis testing** techniques such as **T-Test** and **Z-Test** were applied.

Visualizations crafted using **Matplotlib** and **Seaborn** effectively illustrated patterns in variables such as **Employee Gender**, **Hierarchy**, **Teaching vs Non-Teaching Roles**, and **District-wise Staff Distribution**. These graphics facilitated clearer interpretation of trends in **regional employment**, **designation hierarchies**, and **religious/community demographics**.

Overall, this project underscored the value of structured personnel data in the education sector. It demonstrated how combining **statistical tools** and **visual analytics** can uncover meaningful insights, aiding in **data-driven decision-making** for resource planning and policy formulation.

### ****7.Future Scope****

### Predictive Workforce Planning: Apply machine learning models to forecast future staffing needs, retirement trends, and teacher shortages based on historical employee data.

### Geospatial Workforce Distribution: Use geospatial tools like Folium or Plotly to create interactive maps showing district-wise staff allocation, helping identify underserved regions.

### Automated Retirement Alerts: Develop real-time dashboards to track upcoming retirements using the Schedule RetirementDate field, aiding in proactive hiring and succession planning.

### Equity and Inclusion Analysis: Analyze representation across Caste Name, Community Name, Religion Name, and Gender to evaluate diversity and ensure fair employment practices.

### Educational Resource Allocation Tool: Build a decision-support system for policymakers to optimize deployment of teaching and non-teaching staff based on regional demand and staff hierarchy.

### 8. ****References****

1. Pandas Documentation. Accessed: Apr. 12, 2025. [Online]. Available: <https://pandas.pydata.org/docs>
2. Matplotlib Documentation. Accessed: Apr. 12, 2025. [Online]. Available: <https://matplotlib.org/stable/contents.html>
3. Seaborn Documentation. Accessed: Apr. 12, 2025. [Online]. Available: <https://seaborn.pydata.org>
4. Python Official Website. Accessed: Apr. 12, 2025. [Online]. Available: <https://www.python.org>

5.School Employee Details of Sikkim as on 03-Jun-2016[online].Available: <https://www.data.gov.in/resource/school-employee-details-sikkim-03-jun-2016>