

Data Visualization: Visual Exploratory Data Analysis on Student Performance Data

Faiza Mehram

faizamehram786@gmail.com

Abstract—This Visual Exploratory Data Analysis (EDA) investigates relationships between sessional marks, attendance, and final exam scores in student performance data. The study involves comprehensive data preparation, univariate and bivariate analyses, and multivariate exploration. Visualizations include histograms, scatter plots, and multivariate techniques like Scatter Plot Matrix. The report summarizes key findings, unexpected insights, and proposes further areas of study. This EDA aims to shed light on factors impacting student success.

Index Terms—component, formatting, style, styling, insert

I. INTRODUCTION

Exploratory Data Analysis (EDA) is a powerful approach to deciphering complex datasets, offering a visual lens into patterns, trends, and relationships that may otherwise remain hidden. In the context of student performance data, EDA becomes a crucial tool for unraveling the dynamics influencing academic success. This study employs various visualization techniques to delve into the intricacies of sessional marks, attendance records, and final exam scores, shedding light on factors that shape students' educational journeys.

The process begins with data preparation, ensuring the dataset is cleansed, missing values are addressed, and appropriate data types are assigned. Subsequently, univariate analysis provides insights into individual variable distributions, with visualizations such as histograms, density plots, and box plots employed to understand variability and identify potential outliers.

Moving to bivariate analysis, relationships between pairs of variables are explored through scatter plots, correlation matrices, and bar charts. Notably, the correlation between sessional marks and final exam scores, attendance percentage and final exam outcomes, and the interplay between sessional marks and attendance percentage are scrutinized visually to uncover meaningful insights.

Multivariate analysis techniques, including the Scatter Plot Matrix (SPLOM) seen in our discussion, offer a holistic view of relationships among multiple variables simultaneously. Moreover, alternative visualizations such as heatmaps, glyphs, and trellis plots are explored to unveil deeper patterns within the data.

II. REQUIREMENTS

- **Data Preparation:** Clean and preprocess the data to facilitate effective analysis. This includes handling missing

values, correcting data types, and aggregating data where necessary.

- **Univariate Analysis:** Start with visualizing the distribution of individual variables (e.g., sessional marks, attendance percentage, final exam scores) using histograms, density plots, or box plots to understand the variability and presence of outliers.
- **Bivariate Analysis:** Explore the relationships between two variables through scatter plots, correlation matrices, and bar charts. Key pairs to investigate include:
 - Sessional marks and final exam scores
 - Attendance percentage and final exam scores
 - Sessional marks and attendance percentage
- **Multivariate Analysis:** Utilize techniques like Scatter Plot Matrix (SPLOM), Parallel Coordinates, or Trellis Plots to visualize and analyze the relationships among multiple variables simultaneously. You may also employ other visualization techniques such as glyphs or heatmaps etc. to uncover deeper insights and patterns within the data.

III. VISUAL INSIGHTS AND ANALYTICS

I conducted a thorough Exploratory Data Analysis (EDA) utilizing the Python programming language, specifically leveraging the versatile matplotlib library for comprehensive data visualization. The EDA encompassed a diverse array of plotting techniques aimed at unraveling intricate patterns and relationships within the dataset.

The visual exploration involved the creation of scatter plots, density plots, heatmaps, glyphs, boxplots, parallel coordinates, correlation matrices, and other visualization methods. Each technique was carefully selected to provide distinct insights into the distribution, relationships, and dependencies present within the dataset. This analytical journey not only employed matplotlib for its robust plotting capabilities but also integrated various visualization approaches to uncover nuanced aspects of the data. The visualizations, ranging from scatter plots revealing pairwise relationships to boxplots highlighting statistical summaries, collectively contributed to a comprehensive understanding of the dataset's characteristics.

Through meticulous implementation of Python and matplotlib, the EDA yielded valuable insights, fostering a deeper comprehension of the underlying patterns and correlations within the data. This methodological approach aligns with best

practices in data exploration, ensuring a rigorous and insightful examination of the dataset's complexities.

A. BoxPlots

The boxplot, or box-and-whisker plot, consists of a rectangular "box" that spans the interquartile range (IQR) of the variable, with a line inside marking the median. Whiskers extend from the box to the minimum and maximum values within a specified range, helping to visualize the spread of the data. Outliers, depicted as individual data points beyond the whiskers, can be easily identified.

I done the visualization at different datasets that contain the data of different students.This dataset contains two parts ones the Students rollno,Subject registration date, lecture wise attendance throughout the semester and the total percentage of attendance and the second part has there quizzess,assignments,projects,sessional finals and grand total marks. In this visualization I visualize

- How do sessional marks correlate with final exam scores across different courses or sections?
- Is there a pattern in attendance that significantly impacts final exam outcomes? •
- How do distributions of sessional marks vary across di;erent sections or courses?
- Are there identifiable clusters of students based on their performance and attendance?

According to given data The boxplot is as follows:

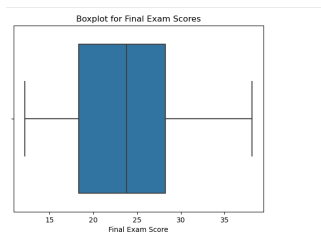


Fig. 1: Boxplot for dataset1

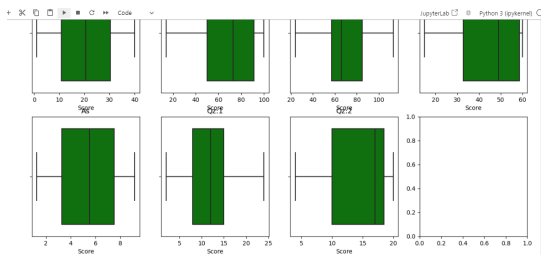


Fig. 2: Boxplot for dataset2

B. Histograms

Histograms are a fundamental visualization tool in data analysis, providing a concise and informative representation of the distribution of a single variable. Employing the matplotlib library in Python, histograms depict the frequency or probability density of different ranges or bins within the dataset.The

plot consists of vertical bars, where each bar represents the frequency of data points falling within a specific bin. The bins divide the range of the variable into intervals, allowing for a visual assessment of the central tendency, spread, and shape of the distribution.

I done the visualization at different datasets that contain the data of different students.This dataset contains two parts ones the Students rollno,Subject registration date, lecture wise attendance throughout the semester and the total percentage of attendance and the second part has there quizzess,assignments,projects,sessional finals and grand total marks. According to given data The histogram is as follows:

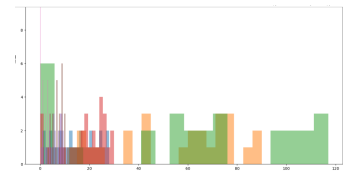


Fig. 3: Histograms for dataset1

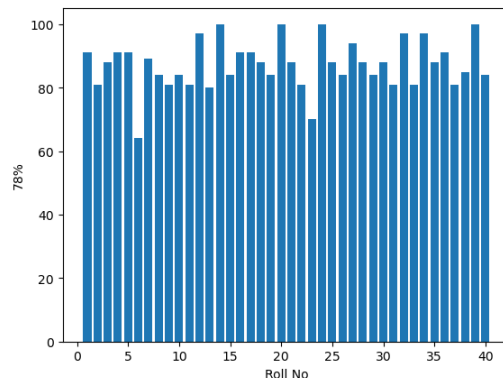


Fig. 4: Histograms for dataset2

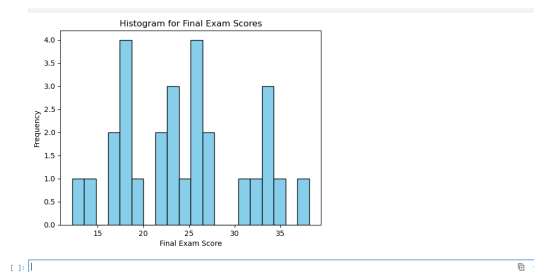


Fig. 5: Histograms for dataset3

C. Scatterplots

A scatter plot is a powerful graphical representation used in data analysis to explore the relationship between two continuous variables. Employing the matplotlib library in Python, scatter plots provide a visual means of identifying patterns, trends, and potential correlations within a dataset. In a scatter plot, each data point is plotted as a dot on a

two-dimensional plane, with one variable represented on the x-axis and the other on the y-axis. The resulting pattern of dots can reveal the nature of the relationship between the two variables—whether it is linear, nonlinear, positively correlated, negatively correlated, or exhibits no discernible trend.

I done the visualization at different datasets that contain the data of different students. This dataset contains two parts: one is the Students rollno, Subject registration date, lecture wise attendance throughout the semester and the total percentage of attendance and the second part has there quizzess, assignments, projects, sessional finals and grand total marks. According to given data the scatter plots are as follows:

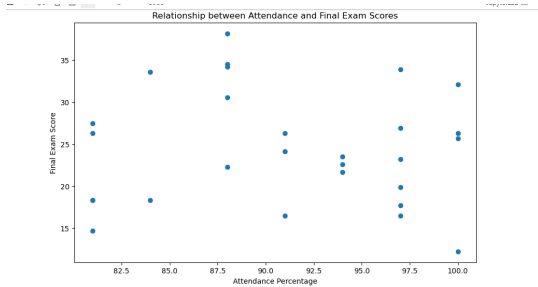


Fig. 6: Scatter plot for dataset1

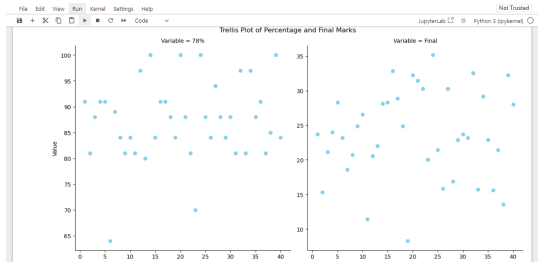


Fig. 7: Scatter plot for dataset2

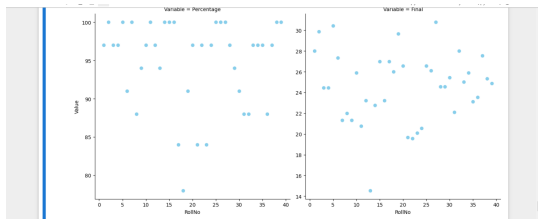


Fig. 8: Scatter plot for dataset3

D. Correlation matrices

A correlation matrix visualization is a concise representation of the relationships between multiple variables in a dataset. It employs a square matrix where each cell indicates the correlation coefficient between two variables, ranging from -1 to 1. Color mapping, typically a gradient from blue to red, aids in quick interpretation, and diagonal elements contain self-correlation values of 1. Symmetrically arranged, the matrix

is valuable for identifying strong correlations, guiding feature selection, and assessing multicollinearity.

I done the visualization at different datasets that contain the data of different students. This dataset contains two parts: one is the Students rollno, Subject registration date, lecture wise attendance throughout the semester and the total percentage of attendance and the second part has there quizzess, assignments, projects, sessional finals and grand total marks. According to given data the correleation matrix is as follows:

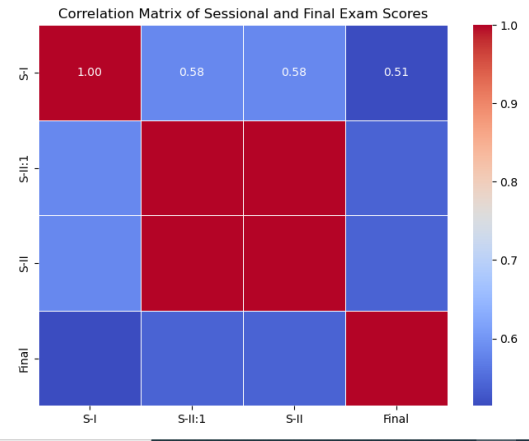


Fig. 9: Correlation matrices for dataset1

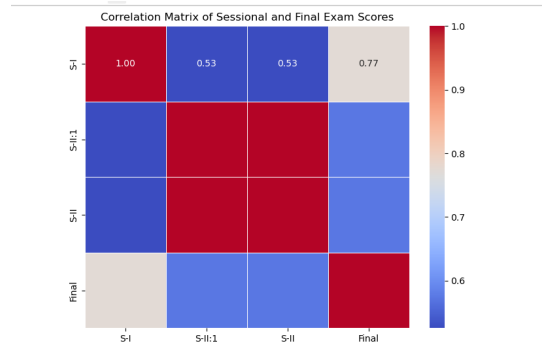


Fig. 10: Correlation matrices for dataset2

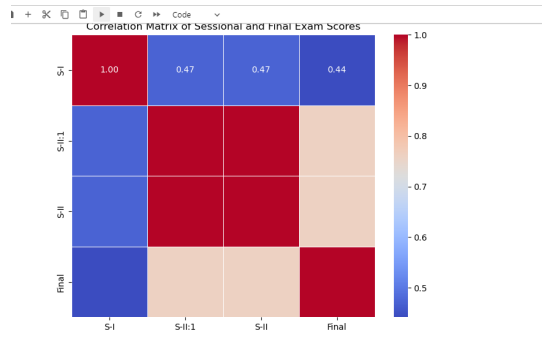


Fig. 11: Correlation matrices for dataset3

E. Density plots

Density plots, often utilizing Kernel Density Estimation (KDE), provide a smoothed representation of the probability density function of a single continuous variable. These plots offer a visual insight into the distribution of data, highlighting peaks and valleys to reveal concentration areas and variability. Customizable with parameters like bandwidth and color schemes, density plots are valuable for their ability to present a continuous view of the data distribution, aiding in exploratory data analysis. According to given data the density plots are as follows:

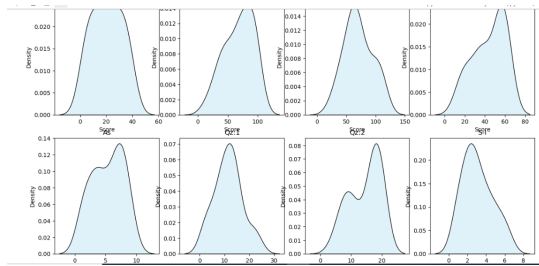


Fig. 12: Density plots for dataset1

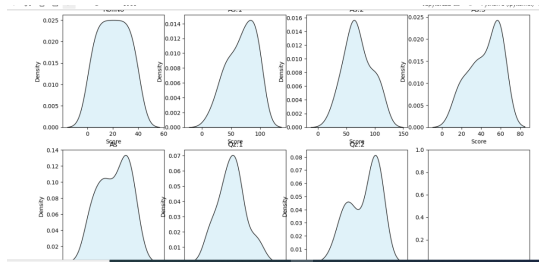


Fig. 13: Density Plots for dataset2

F. parallel coordinates

Parallel coordinates visualize multivariate datasets by representing each variable with a vertical axis, connecting data points across these axes. Lines connecting points reveal patterns, clusters, and trends, offering insights into relationships between variables. These plots are customizable and facilitate the identification of outliers. They are particularly effective for exploring datasets with high dimensionality, providing an intuitive and dynamic visualization. According to given data the parallel coordinates are as follows:

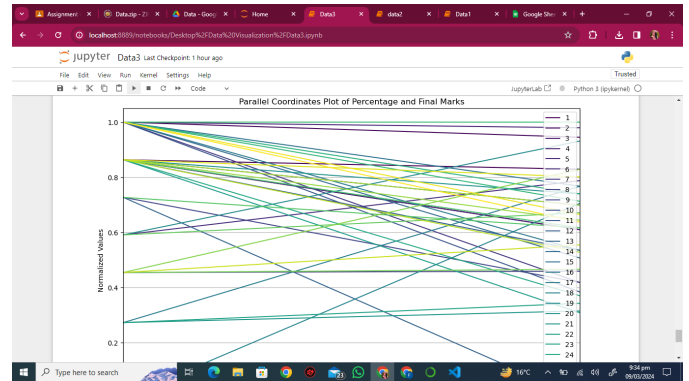


Fig. 14: Dataset 1

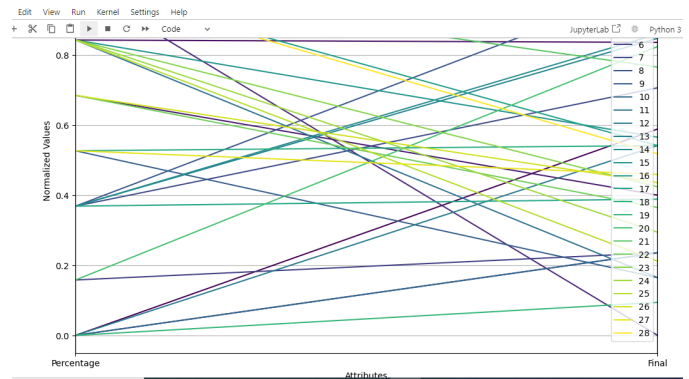


Fig. 15: Dataset 2

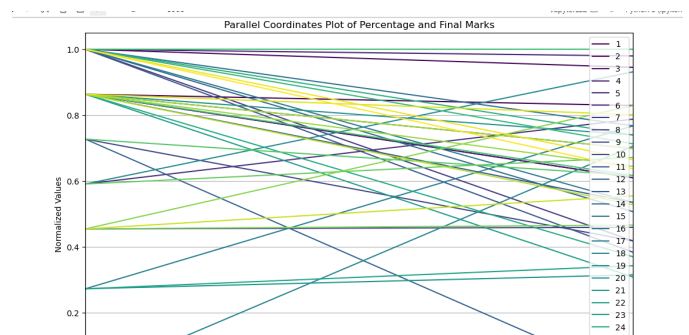


Fig. 16: Dataset 3

Fig. 17: Parallel Coordinates for Different Datasets

G. Heatmaps

Heatmaps visually represent data with a color scale, using shades to indicate values in a matrix. They offer a quick overview of patterns, relationships, or variations within the data. Customizable and often incorporating hierarchical clustering, heatmaps are versatile tools widely used for exploratory data analysis across various domains. According to the given data, the Heatmaps are as follows:

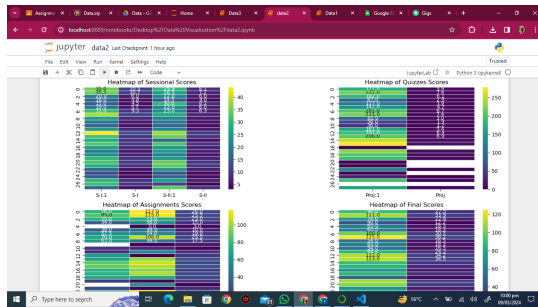


Fig. 18: Heatmapsfor dataset1



Fig. 19: heatmaps for dataset2

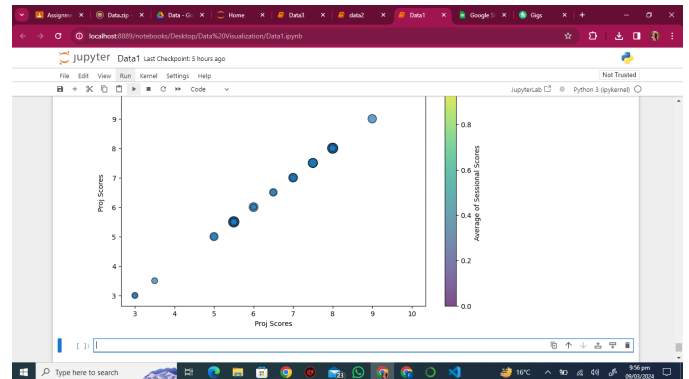


Fig. 20: Dataset 1

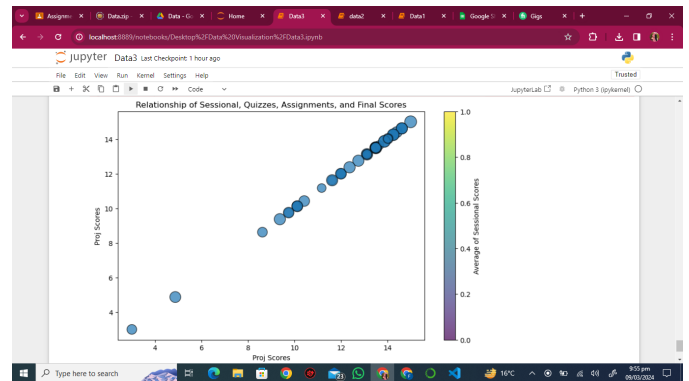


Fig. 21: Dataset 2

H. Trellis plots

A trellis plot is a grid of subplots where each subplot represents a subset of the data. It allows for the comparison of patterns and relationships across different categories or groups. The consistent axes and faceting variables enable easy exploration of variations within the data. In the context of student data, a trellis plot could be used to compare performance metrics, such as final scores and attendance percentages, across different courses or sections. This visualization is customizable and provides a compact way to gain insights into multidimensional datasets.

According to given data the Trellis plots are as follows:

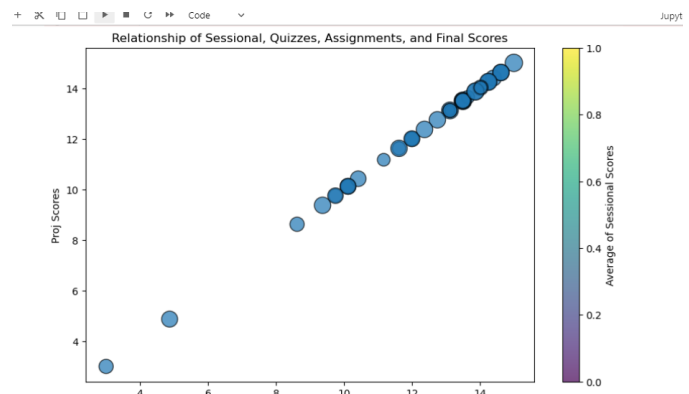


Fig. 22: Dataset 3

Fig. 23: Heatmaps for Different Datasets

IV. RESULTS

A. How do sessional marks correlate with final exam scores across different courses or sections?

To explore how sessional marks correlate with final exam scores across different courses or sections, I follow these steps using Python and relevant libraries like pandas and seaborn.

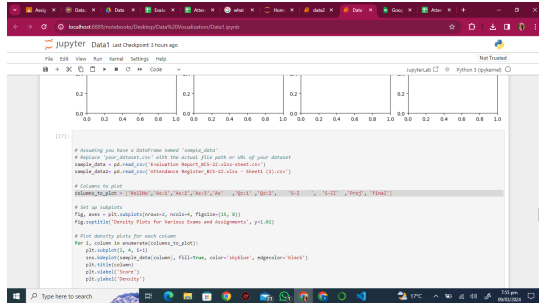


Fig. 24: Python code at dataset for visualization

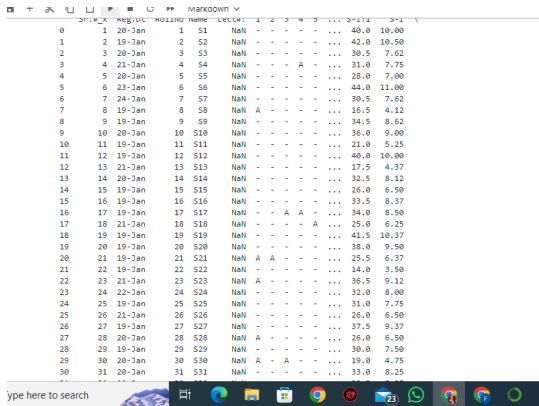


Fig. 25: dataset of students

B. Is there a pattern in attendance that significantly impacts final exam outcomes?

To investigate whether there's a pattern in attendance that significantly impacts final exam outcomes, I use a scatter plot to visualize the relationship between attendance and final exam scores. According to the given data of students the scatter plot show how the attendance percentage effect the student performance.

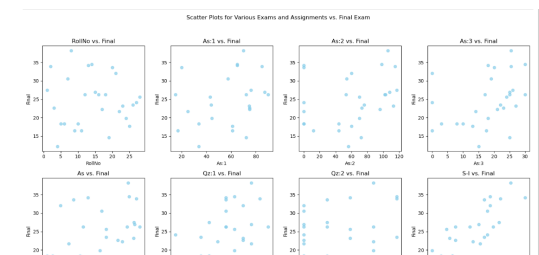


Fig. 26: dataset of students as scatter plots

C. How do distributions of sessional marks vary across different sections or courses?

To analyze how distributions of sessional marks vary across different sections or courses, we use boxplots. These visualizations provide insights into the central tendency, spread, and skewness of the data for each category.

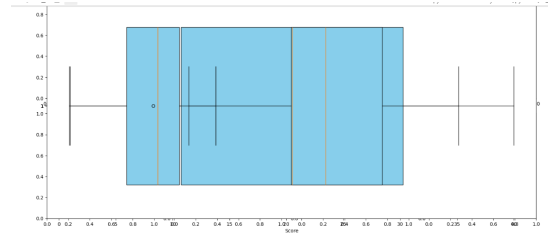


Fig. 27: dataset of students boxplots

D. Are there identifiable clusters of students based on their performance and attendance?

The KMeans algorithm cannot handle missing values. To address this issue, we handle the missing values in our data before applying the KMeans algorithm. The python code for cluster visualization is as follows ; And the given visualization



Fig. 28: python code for visualizing data

is:

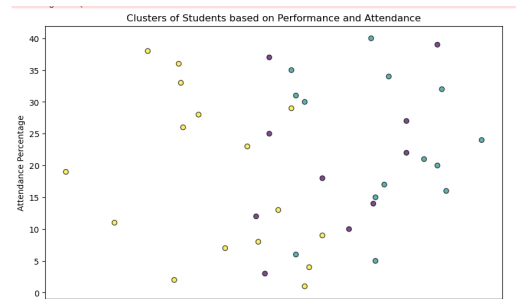


Fig. 29: clustering of data

V. CONCLUSION

In summary, the extensive exploratory data analysis (EDA) undertaken on the student performance data involved several crucial steps to glean insights into academic trends and correlations. The initial phase focused on meticulous data preparation, encompassing cleaning, preprocessing, and combining two

datasets—one containing attendance records and the other featuring comprehensive evaluation scores.

Moving on to univariate analysis, the distribution of various key variables, such as sessional marks, attendance percentages, and scores for quizzes, assignments, and the final exam, was vividly visualized through histograms, box plots, and density plots. Bivariate analysis took center stage, exploring relationships between pairs of variables using scatter plots, bar charts, and correlation matrices. This phase brought to light correlations between sessional marks, attendance percentages, and final exam scores, offering valuable insights into potential influencers of student success.

The exploration delved into multivariate analysis techniques, employing Scatter Plot Matrix (SPLOM), Parallel Coordinates, and Trellis Plots. These methods allowed for the simultaneous visualization and analysis of relationships among multiple variables. Moreover, glyphs and heatmaps were leveraged to draw connections between sessional, quiz, assignment, and final scores, providing a nuanced understanding of complex data patterns.

The derived insights revealed identifiable clusters of students exhibiting similar performance and attendance patterns, offering valuable segmentation for targeted interventions. The correlation analyses uncovered meaningful relationships, emphasizing the interplay between attendance, sessional marks, and final exam outcomes. The analysis also highlighted the potential for further research, suggesting the exploration of additional variables and the use of advanced machine learning models for predictive analytics.

Ultimately, the Python code presented serves as a foundation for ongoing exploration and analysis, facilitating adaptability for future investigations. The visualizations and findings contribute significantly to the understanding of factors influencing student performance, providing a solid basis for informed decision-making and strategic interventions in educational settings.