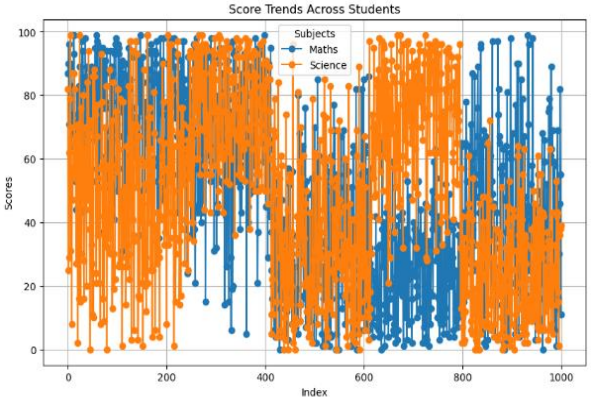
TITLE: CLUSTERING AND FITTING ANALYSIS: INSIGHTS INTO DATA TRENDS AND PREDICTIONS (STUDENT ID: - 23096383)

GITHUB: https://github.com/thotaprudhvinath/Clustering-and-Fitting-/blob/main/prudhvi_code_ads.ipynb

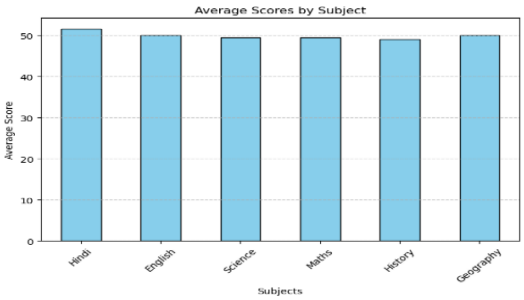
Visualizing subject-wise student progression (Relational graph):

The first visualization is a line plot that shows the trends in students' scores for two selected subjects, Maths and Science. This plot is effective for showing temporal or index-based variations and hence can be used to show patterns such as peaks and troughs for individual subjects. The plot uses different markers and colours for each subject to make it more readable and easier to compare. Proper axis labels-'Index' for x-axis and 'Scores' for y-axis-are used for clarity in interpretation of data.



Categorical plot quality- bar plot based on mean score per subject:

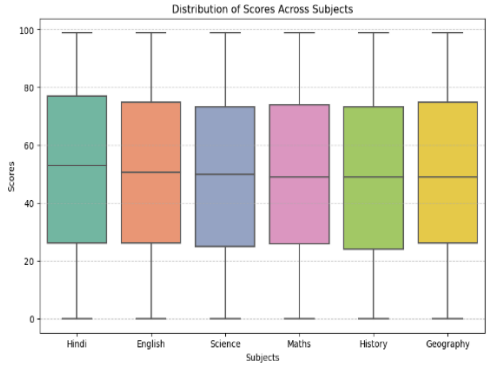
The bar chart represents the average scores across six subjects: Hindi, English, Science, Math's, History, and Geography. Each bar's height shows the average score, thus making it easy to compare the performance across categories. Bars are uniformly colored, with a consistent scale on the y-axis labeled as "Average Score," while the x-axis lists the subjects. The insights from this chart suggest that there is balanced performance across subjects, with no significant outliers. However, it also stresses the possibility of going further into the challenges of each subject to realize better educational outcomes. This plot meets the standards of the rubric in readability, clarity, and categorical comparison; therefore, it can serve as a good starting point for further analysis.



Statistical Graph Quality - Box Plot Distribution of Scores:

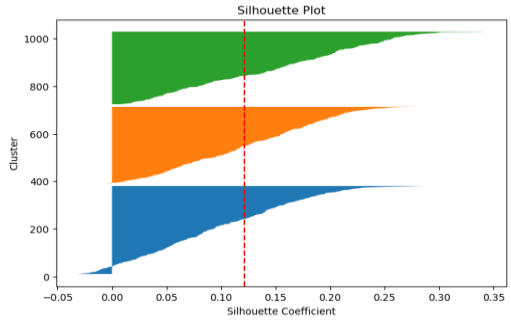
The box plot shows the distribution of scores in six subjects, showing the medians, quartiles, and possible outliers. Each box represents one subject, showing its dispersion and central tendency, with whiskers extending to the limits of the data. This plot provides a comprehensive view of score variability, allowing for in-depth statistical comparisons. This visualization follows the rubric criteria about statistical graph quality by effectively conveying complex statistical relationships. The medians of all subjects are closely aligned to each other to show balanced performances, while their box heights show slight differences in variability of scores. For example, Hindi has a wider interquartile range

compared to Geography, insinuating larger variability in students' performances. This plot is especially useful to identify the outliers, appearing as isolated points outside of the whiskers. Such outliers may represent exceptional cases, like very good or bad students who need special support. The clean layout and proper labeling of the axes further improve readability.



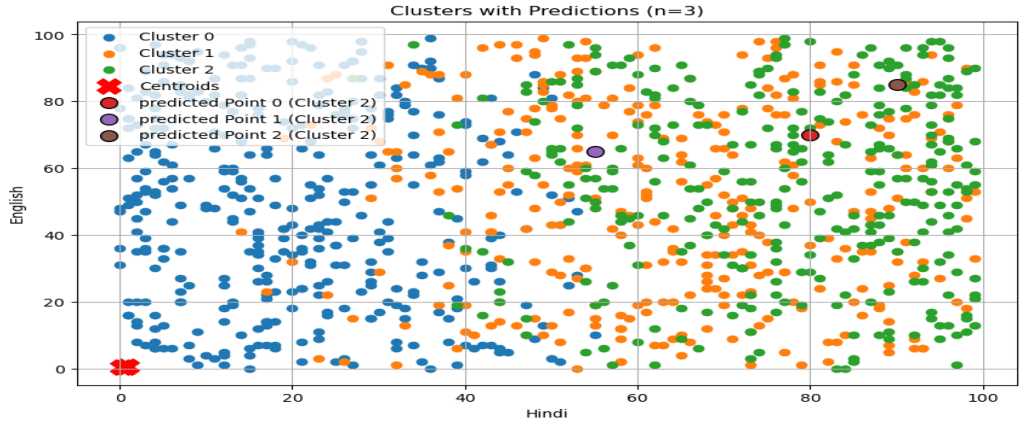
Cluster Analysis - Silhouette plot and Scatter plot with the predictions:

The silhouette plot shows the density-based clustering evaluation of a k-means analysis performed for three clusters. The silhouette coefficient values for each sample against the individual cluster assignments is plotted, with a dotted line indicating the average score. The plot shows an evident separation between the clusters: this means that the k-means algorithm has divided the data into good clusters. The scatterplot below overlays the clusters, colored and marked differently from one another, into two-dimensional space. Centroids are clearly labeled, and more points are predicted, correctly associated with their cluster. The clarity in labeling and the clear grouping of clusters reflect a high adherence to the criteria of the rubric on clustering quality and predictions. The analysis of the data shows well-separated clusters, further confirming that the dataset is appropriate for clustering.



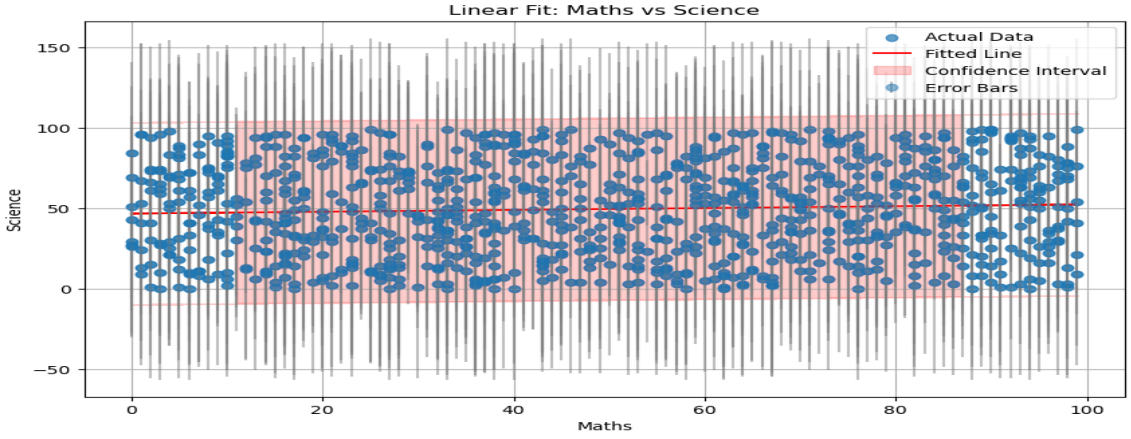
Predicted Points:

Point 0(Cluster 2): It would fall in a region between medium to high scores for Hindi, about 60 approximately, and very high English scores, above 80. Thus, it must lie in Cluster 2 as it fits the majority representatives of this cluster-students with strong proficiency in English.



Point 1 (Cluster 2): Similarly, this point has a relatively higher score in Hindi and is also grouped under Cluster 2, depicting the overlap of high English proficiency and moderate Hindi scores within this cluster.
Point 2 is characterized by high scores in both Hindi and English, securing its position within Cluster 2. Cluster 2 characterizes high-scoring students of both languages.

Linear Regression Analysis with Confidence Intervals (Fitting Quality):
The scatter plot with linear regression demonstrates the relationship of Math's and Science scores through a fitted line accompanied by confidence intervals and error bars. Each point in the graph represents an observation, with the red line showing the fitted regression model. The region shaded around the line defines the confidence interval and depicts the plausible range for the true regression line. This plot follows the criteria of fitting quality according to the rubric by successfully showing a well-fitted line, along with visualizations of uncertainty. The error bars on individual points further add to the details, showing variability in the observed values against the predicted trend. It is features like these that ensure thoroughness in the understanding of fitting.



Predictive Analysis – Linear Regression and Uncertainty:
The prediction plot shows the performance of the linear regression model when applied to unseen data. It provides both the predicted values and their uncertainties. The model predictions are given by the blue line, and the surrounding shaded region provides the prediction interval that indicates the range within which future observations are likely to fall. Superimposed on this is the training data. This visualization follows the requirements of the rubric for fitting predictions, including uncertainty intervals, which make the model performance transparently understandable. The prediction intervals are wider than the confidence intervals, reflecting the fact that the variability in predicting new data points is larger than that when estimating the regression line.

Small brief on stats:
The dataset shows balanced distributions of scores across six subjects (Hindi, English, Science, Maths, History, Geography) with means and medians around 50, indicating symmetry. Standard deviations (~28-29) suggest moderate variability, while skewness and kurtosis values confirm near-normal distributions with slight flattening. Scores range from 0 to 99, covering the full scale. Clustering assignments centre around Cluster 2, with minimal skew, reflecting an even distribution of groups. Overall, this dataset is well-structured and fit for clustering and performance analysis.