Coursera_Courses

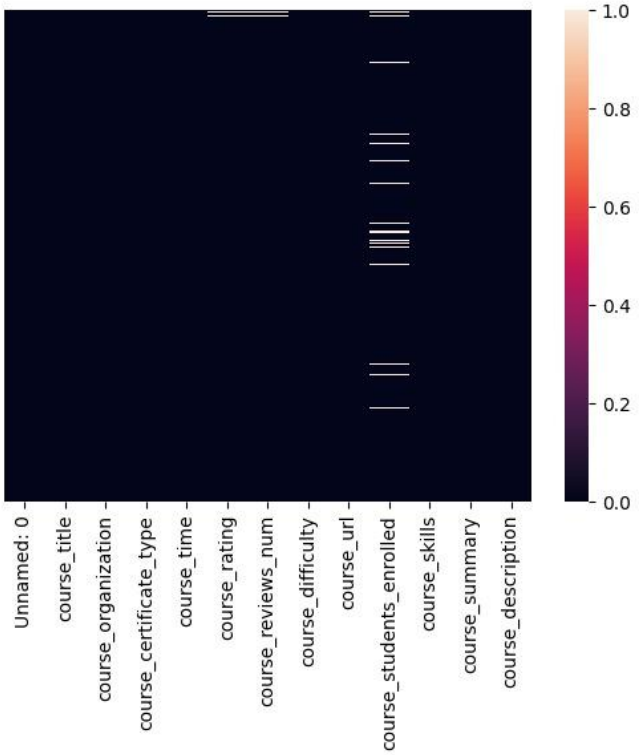
Name: Muhammad Umer Mehmood
Sudent ID: 23102319

Github: https://github.com/UmerCheena/Applied-Data-Science

Introduction:

This report includes a data set with details of coursera courses and gives us analysis to uncover different results, such as course_organization, course_certificate_type, course_time, course_rating, and course_difficulty. We have done course segmentation to understand course's behavior. We will also explore clustering techniques and make the fitting process in data.

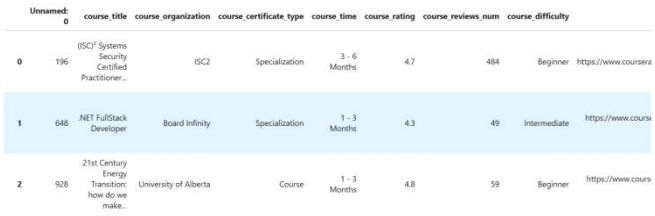
Data Processing and Cleaning



The code snippet uses to identify missing values in a DataFrame and creates a heatmap with to visualize them, where brighter colors indicate missing values. The hides row labels for a cleaner plot, and displays the visualization, making it easy to spot missing data patterns in the dataset.

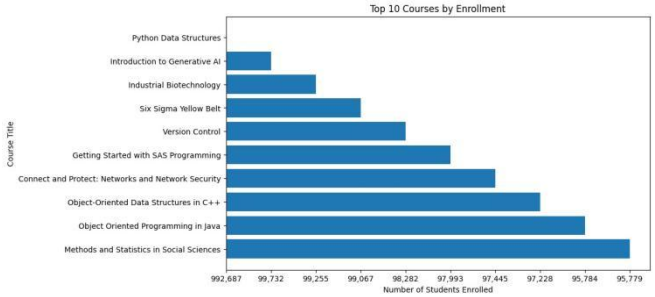
EDA Explanatory Data Analysis:

This data set shows considerable variability across different metrics. Course_organization and course_certificate_type are relatively more concentrated around their means, showing less variation among the majority of courses. On the other hand, course_time, course_difficulty.

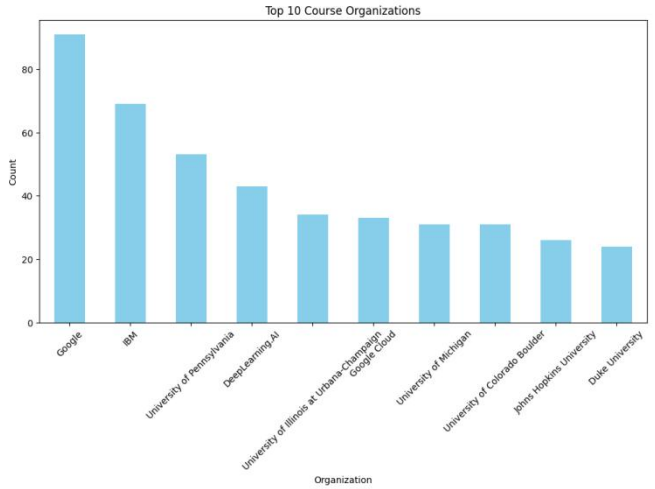


	Unnamed: 0	course_title	course_organization	course_certificate_type	course_time	course_rating	course_reviews_num	course_difficulty	
0	196	(ISC)² Systems Security Certified Practitioner...	ISC2	Specialization	3 - 6 Months	4.7	484	Beginner	https://www.coursera
1	648	.NET FullStack Developer	Board Infinity	Specialization	1 - 3 Months	4.3	49	Intermediate	https://www.course
2	928	21st Century Energy Transition: how do we make...	University of Alberta	Course	1 - 3 Months	4.8	59	Beginner	https://www.cours

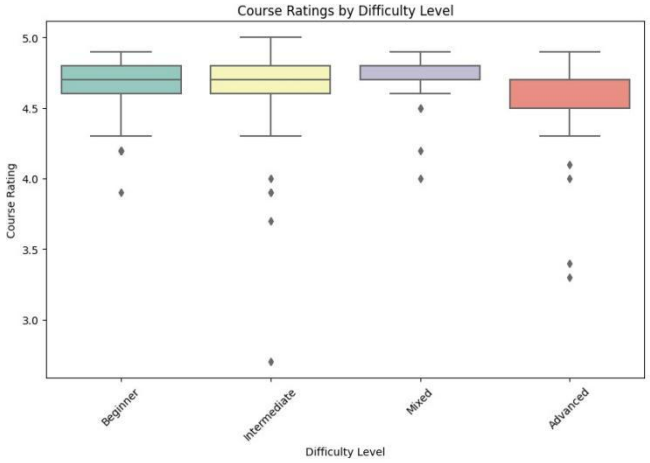
The code identifies the top 10 courses with the highest student enrollments by sorting the dataset in descending order. A horizontal bar plot is created using to display the top courses and their enrollment numbers, with labels and a title for clarity. The y-axis is inverted to display the courses in descending order of enrollment for better visualization.

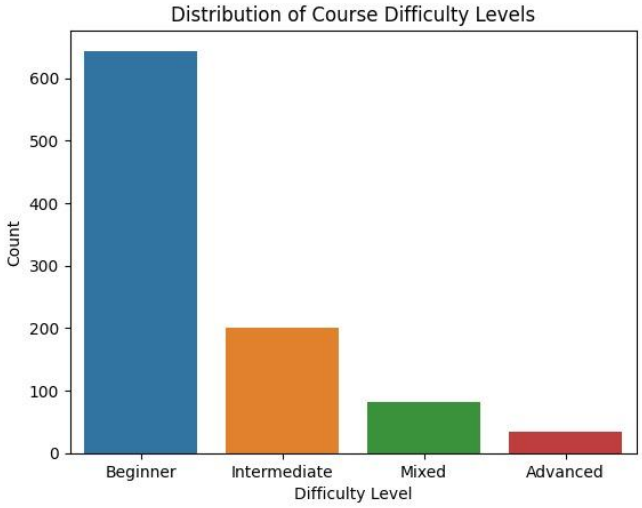


The code identifies the top 10 organizations offering the most courses by counting occurrences in the column and displays the results. A bar plot is then generated to visualize the counts of these top organizations, with appropriate labels, a title, and rotated x-axis labels for readability. This provides insight into the most active organizations in the dataset.

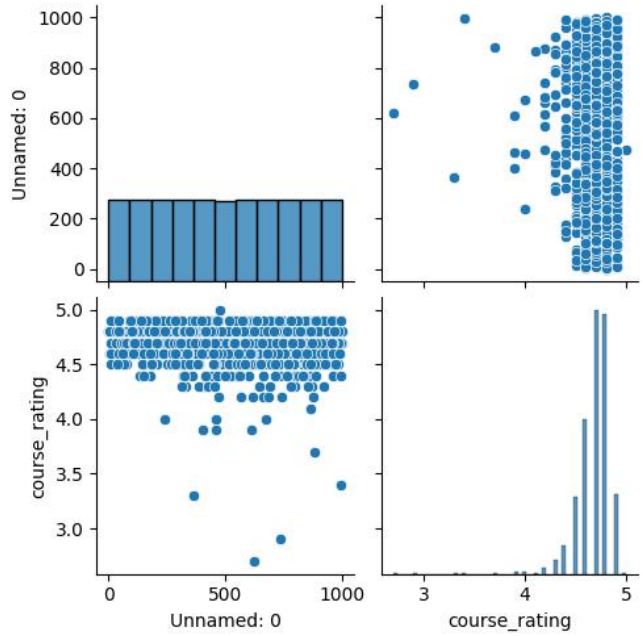


The code creates a box plot using to visualize the distribution of for each level in the dataset. The plot includes a title, axis labels, and uses a color palette, with the x-axis labels rotated for better readability. This helps compare course ratings across different difficulty levels, highlighting central tendencies and variability.





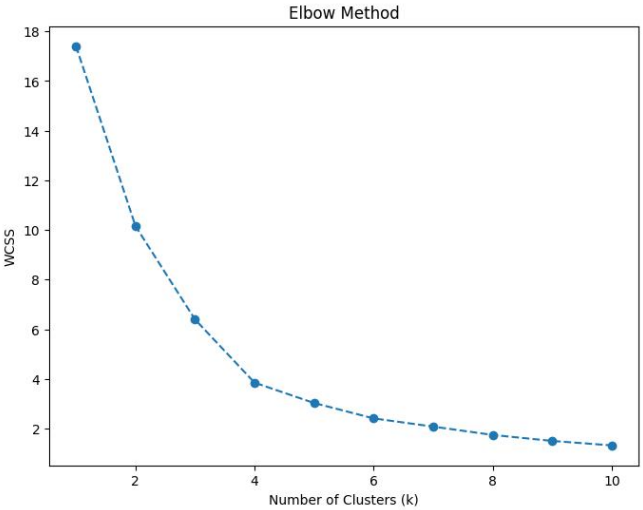
The code creates a grid of scatterplots for each pair of numeric columns in the `data` DataFrame. It visually explores relationships and correlations between variables, helping identify patterns, trends, or outliers. The diagonal typically shows histograms or kernel density plots for individual variables.



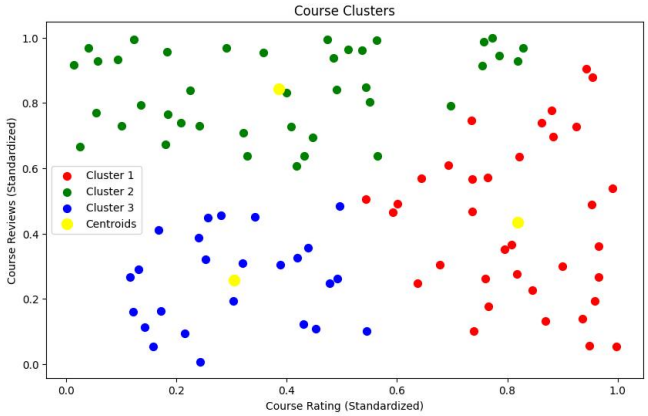
The code uses to create a bar plot showing the distribution of values in the column of the DataFrame. The x-axis represents the difficulty levels, and the y-axis shows the count of each level. Labels and a title are added to make the plot more informative, providing a clear visualization of the frequency of different difficulty levels in the dataset.

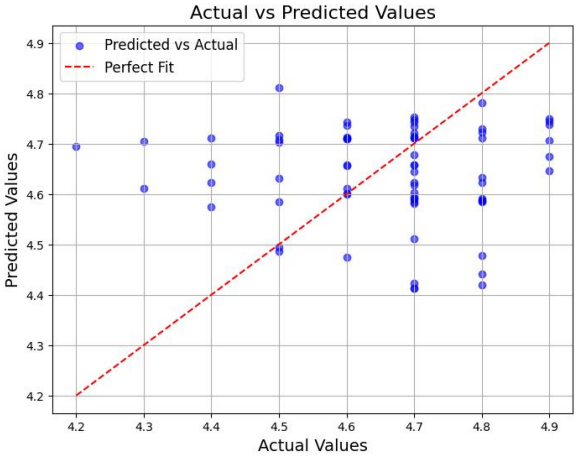
Elbow Method:

The code defines a function, to apply the Elbow Method for determining the optimal number of clusters in a K-Means clustering algorithm. It iterates over different values of calculates the within-cluster sum of squares (WCSS), and plots it against to identify the point where WCSS decreases more slowly the elbow. Using randomly generated 2D data, the function is called with visualizing the clustering performance for values from 1 to 10.



The code visualizes K-Means clustering results by plotting data points in each cluster with different colors and labeling them accordingly. It uses a loop to scatter plot data points for each cluster, with centroids marked in yellow for distinction. Labels and a legend are added to enhance interpretability, while the x and y axes represent standardized features like course ratings and reviews, making the clusters visually identifiable.





Regression Process:

The code creates a scatter plot to compare the actual and predicted values of a regression model, highlighting the model's performance. Data points are plotted in blue, while a red dashed diagonal line represents a "perfect fit," where predictions match the actual values. The plot includes labels, a legend, and a grid to improve readability, making it easy to assess how closely the predictions align with the actual values.

Conclusion:

The analysis of the Coursera dataset provides insights into course characteristics such as ratings, enrollments, difficulty levels, and certification types. Beginner courses often attract more learners, while advanced courses are more common in professional certificates and specializations. Most courses are designed for completion within 1-3 months, balancing accessibility and depth. Segmentation and clustering reveal patterns in learner preferences and engagement, offering actionable insights for optimizing course design and marketing strategies. This analysis highlights key trends shaping online learning.