Coursera_Courses

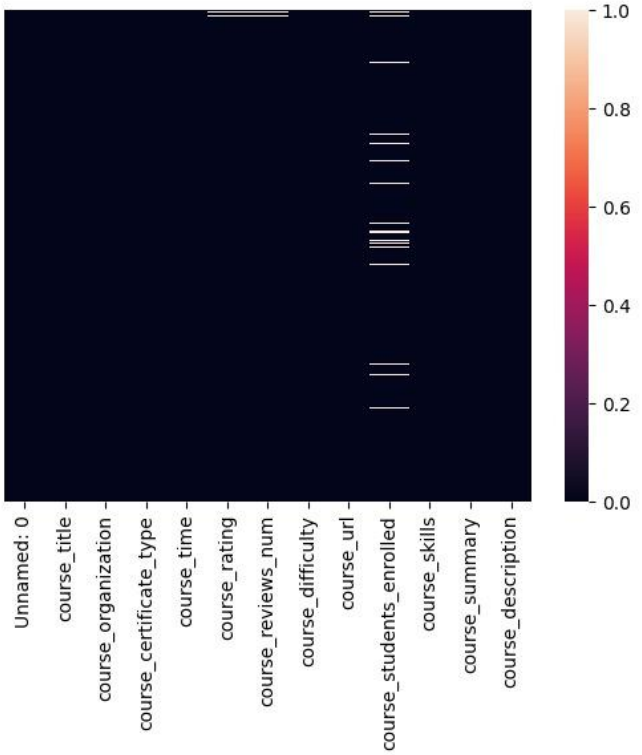
Name: Muhammad Umer Mehmood
Sudent ID: 23102319

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 courses' behaviour. We will also explore clustering techniques and make the fitting process in data.

Data Processing and Cleaning



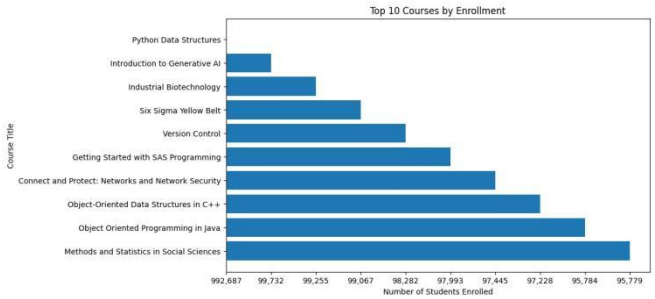
The code scrap employments to recognize lost values in a DataFrame and makes a heatmap with to imagine them, where brighter colours show lost values. The covers up push names for a cleaner plot, and shows the visualization, making it simple to spot missing information designs within the dataset.

EDA Explanatory Data Analysis:

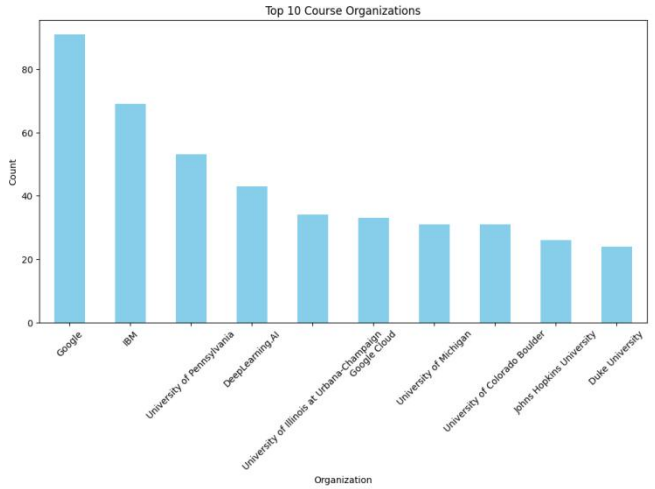
This information set appears significant inconstancy over diverse measurements.Course_organization and course_certificate_type are moderately more concentrated around their implies, appearing less variety among the larger part of courses. On the other hand, course_time, course_difficulty.

	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

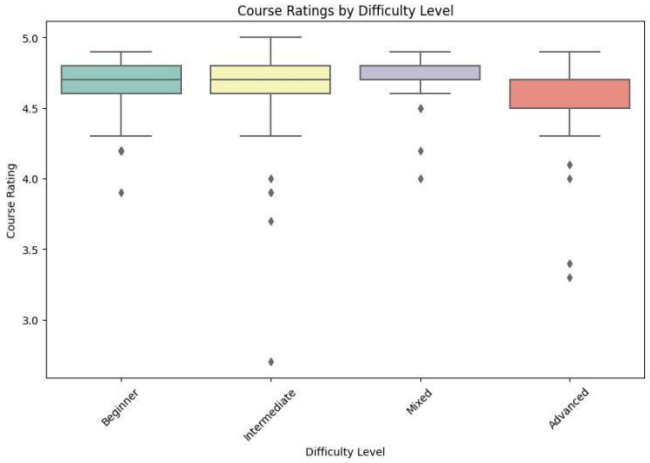
The code recognizes the beat 10 courses with the most noteworthy understudy enrollments by sorting the dataset in plummeting arrange. A level bar plot is made utilizing to show the beat courses and their enrollment numbers, with names and a title for clarity. The y-axis is modified to show the courses in plummeting arrange of enrollment for way better visualization.



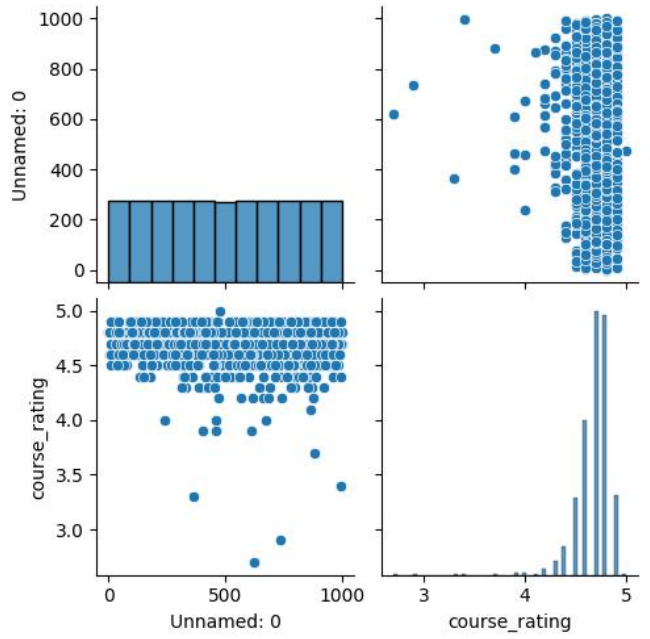
The code recognizes the best 10 organizations advertising the foremost courses by checking events within the column and shows the results. A bar plot is at that point produced to imagine the tallies of these best organizations, with suitable names, a title, and turned x-axis labels for coherence. This gives understanding into the foremost dynamic organizations within the dataset.



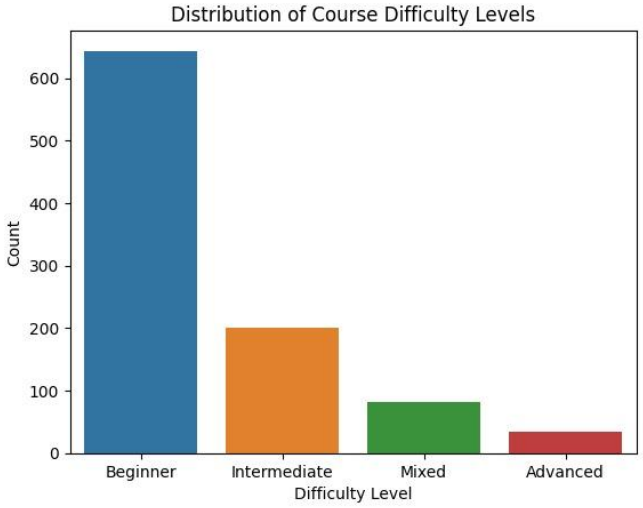
The code makes a box plot utilizing to imagine the conveyance of for each level within the dataset. The plot incorporates a title, pivot names, and employments a color palette, with the x-axis names turned for superior meaningfulness. This makes a difference compare course appraisals over diverse trouble levels, highlighting central inclinations and changeability.



The code makes a network of scatterplots for each match of numeric columns within the DataFrame. It outwardly investigates connections and relationships between factors, making a difference recognize designs, patterns, or exceptions. The corner to corner ordinarily appears histograms or part thickness plots for person factors.

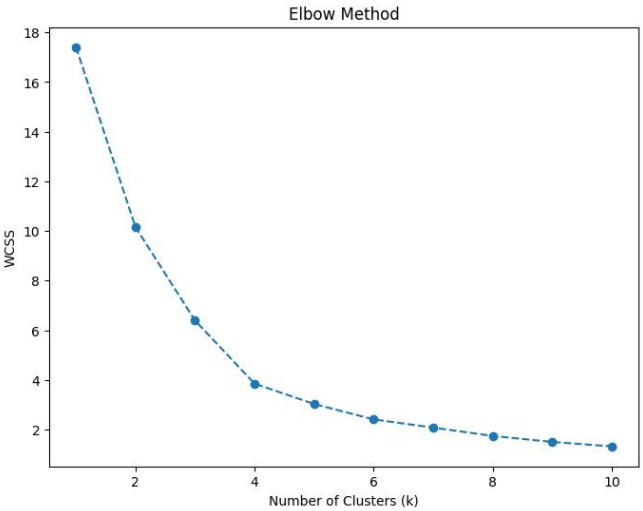


The code employments to form a bar plot showing the dissemination of values within the column of the DataFrame. The x-axis speaks to the trouble levels, and the y-axis appears the number of each level. Names and a title are included to create the plot more instructive, giving a internationalization recurrence of distinctive trouble levels within the dataset.

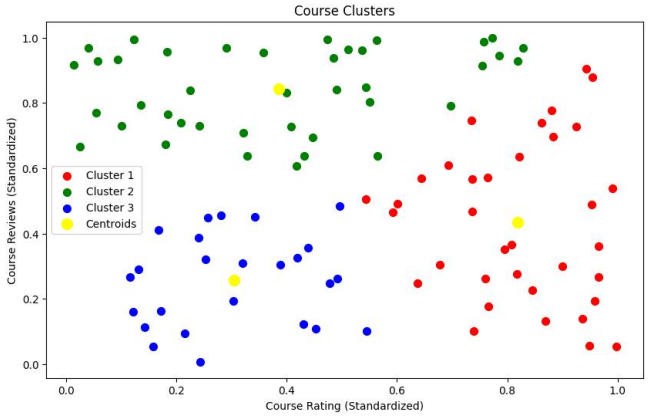


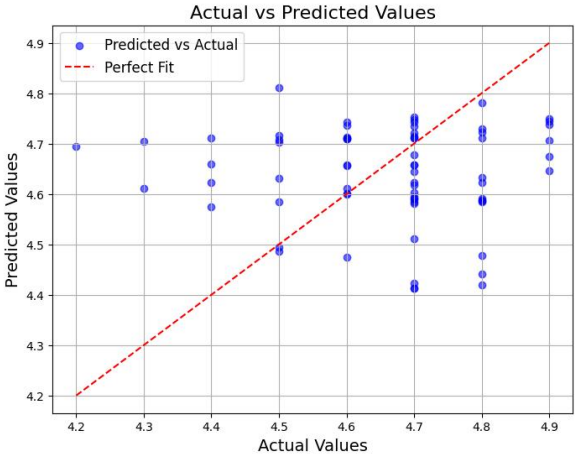
Elbow Method:

The code characterizes a work, to apply the Elbow Strategy for deciding the ideal number of clusters in a K-Means clustering calculation. It emphasizes over diverse values of calculates the within-cluster entirety of squares (WCSS), and plots it against to distinguish the point where WCSS diminishes more gradually the elbow. Utilizing haphazardly created 2D information, the work is called with visualizing the clustering execution 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 investigation of the Coursera dataset gives experiences into course characteristics such as appraisals, enrollments, trouble levels, and certification sorts. Apprentice courses frequently draw in more learners, whereas progressed courses are more common in proficient certificates and specializations. Most courses are outlined for completion inside 1-3 months, adjusting openness and profundity. Division and clustering uncover designs in learner inclinations and engagement, advertising significant bits of knowledge for optimizing course plan and promoting techniques. This investigation highlights key patterns forming online learning.