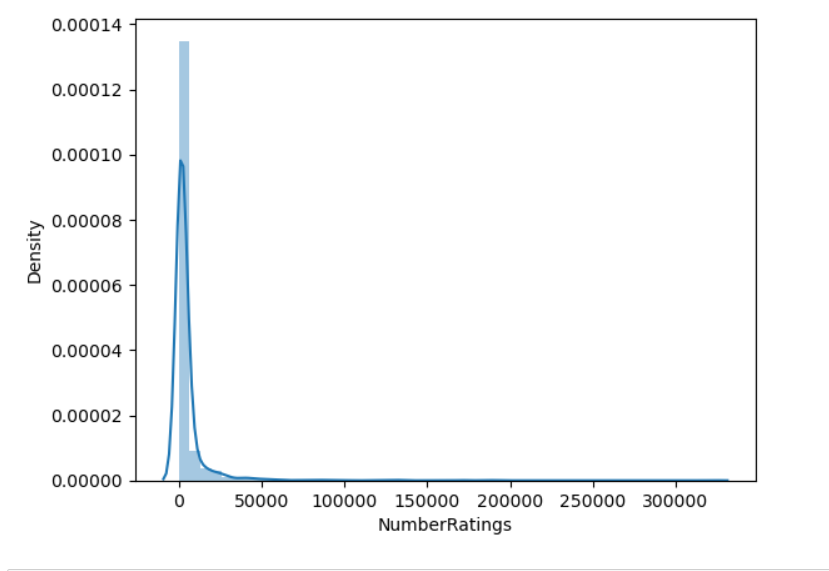
**Title and Introduction**

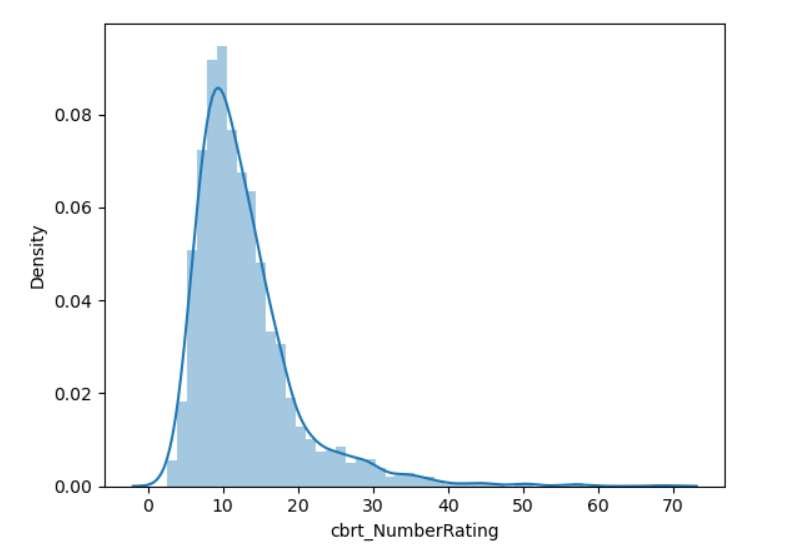
Machine Learning project for predicting the Number Rating using following features. About the dataset it contains information about different online courses. Each row represents a course we have some data which have Binary variables indicating the course category from following courses A value of 1 indicates course belong to following category.

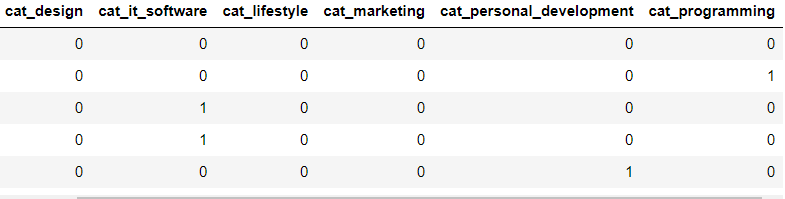
In this project we use two regression models which are Linear Regression and Lasso Regression using different feature selection methods which are using correlation, Variance Threshold and Select Best then in these different features Data Frame we also applied log transformation to all these features Log transformation helps to normalize skewed data and make it more suitable for modelling. and then performed scaling on all the features which is Robust Scaling method then we tested our model score on and RMSE (mean square error).

**Dataset Analysis**

In Lab02 we check the shape of Number Ratings and it is right skewed. Then we perform cube root transformation to make it normalise. Which we can achieve close to that. Then we remove values which are greater than 40 after that we get close to normalise shape. Then we performed cube root transformation on Enrolment and deleted value which are greater than 60. And same for Student and Review and we deleted index which are less than or equal to 3.75. Then we created dummy features for the CourseCategory and then drop the column Course Category.

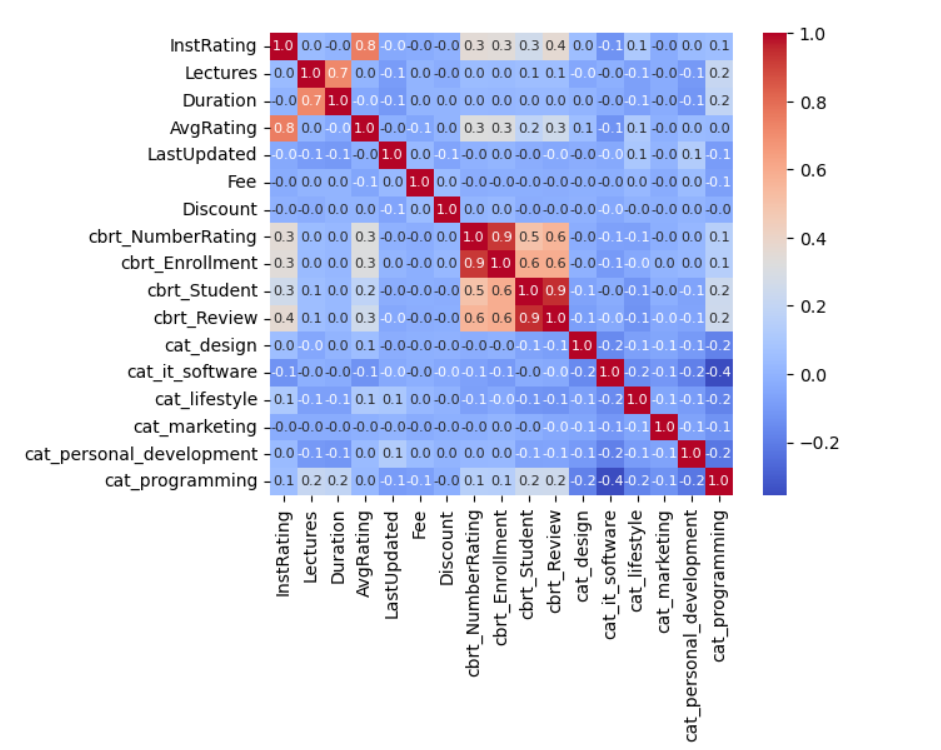






Shape of Data (1883 rows and 18 columns):

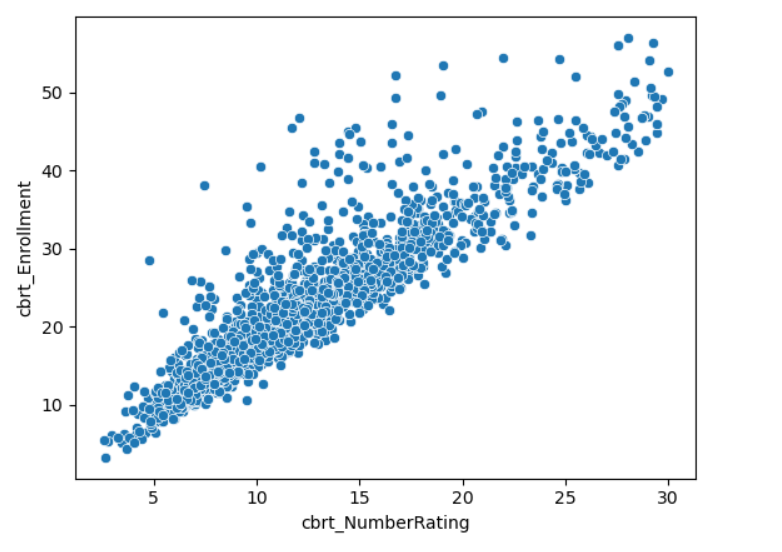
**EDA**



Among all the features, "cbrt\_Enrollment" exhibits a strong positive correlation of approximately 90% with the target variable.

This indicates that there is a significant association between the number of enrolments (cbrt\_Enrollment) and the target variable (cbrt\_NumberRating). When the number of enrolments is high, it is likely that the cbrt\_NumberRating is also high. This positive correlation suggests that as the enrolment increases, there is a tendency for the number rating to increase as well.

Furthermore, the features "cbrt\_Student" and "cbrt\_Review" also show moderate positive correlations of 48% and 53%, respectively, with the target variable. As shown in figure below



**Feature observation and hypothesis**

features such as "InstRating," "Lectures," "Duration," "AvgRating," "LastUpdated," "Fee," "Discount," "cbrt\_Enrollment," "cbrt\_Student," "cbrt\_Review," "cat\_design," "cat\_it\_software," "cat\_lifestyle," "cat\_marketing," "cat\_personal\_development," and "cat programming." "InstRating" represents the institution's rating, "Lectures" denotes the number of lectures, "Duration" indicates the duration of the course, "AvgRating" represents the average rating, "LastUpdated" signifies the time of the last update, "Fee" denotes the course fee, and "Discount" represents the available discount. The variables prefixed with "cbrt\_" is the cube root transformations. Hypothesis for regression model is that cbrt\_NumberRating can be predicted by using some of these features or all of these in further steps we performed feature selection. The regression model would aim to estimate the relationship between the independent variables and the dependent variable.

**Simple Linear Regression Report**

We used 3 type of feature selection technique which are Correlation method, Variance Threshold method and SelectKBest for **correlation** selecting only those features which have correlation with cbrt\_NumberRating above 0.1 below are the features we get.



**For Variance Threshold** we pass threshold of 3



And for **SelectKBest** we need 7 features so we pass value 7 and below are 7 features which we get.

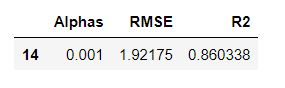


**Linear Regression with Lasso Report**

We use 15 alpha values between 1 and 0.001 as shown below and for alpha value 0.001 we got the highest R2 which is 86% and least RMSE score which is 1.92



Selected Alphas



**Analysis**

After checking all the models, we selected df\_corr where no log transformation is performed because we get the highest R2 below is the table and coefficient values for the following. We can increase the score by try using further models for regression like Decision Trees Regression and Random Forest Trees Regression. And by continuous improvement regularly re evaluate model performance as new data becomes available and update and refine model accordingly

