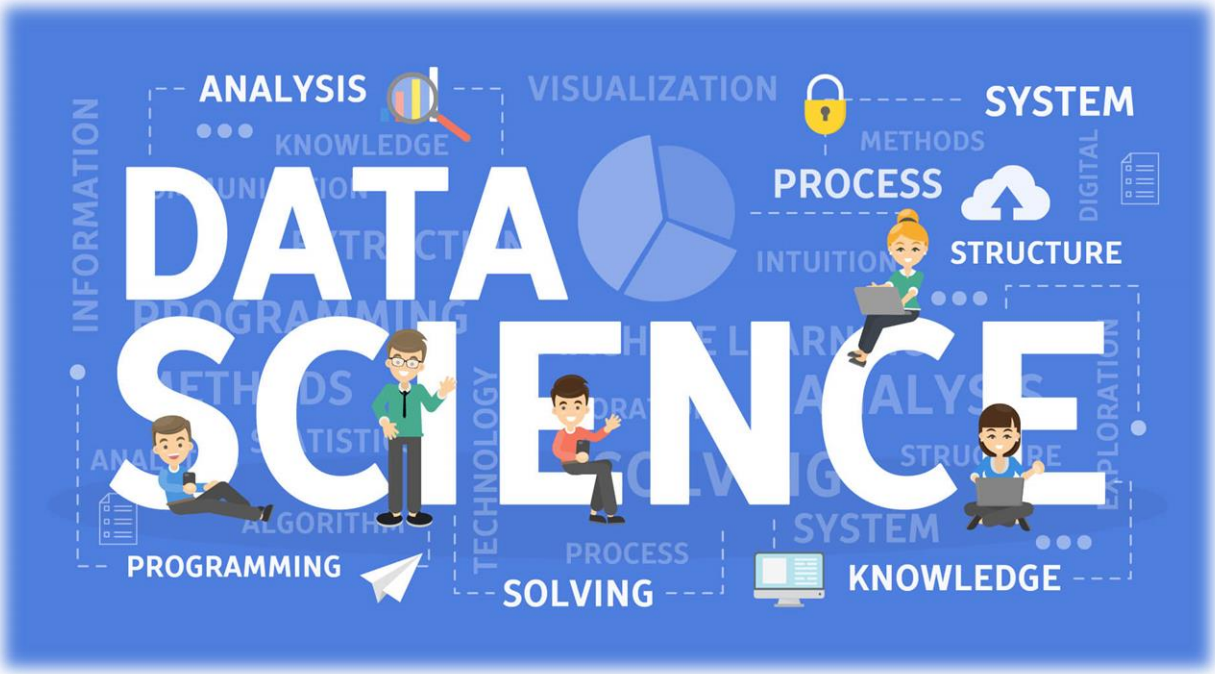
Data clustering report

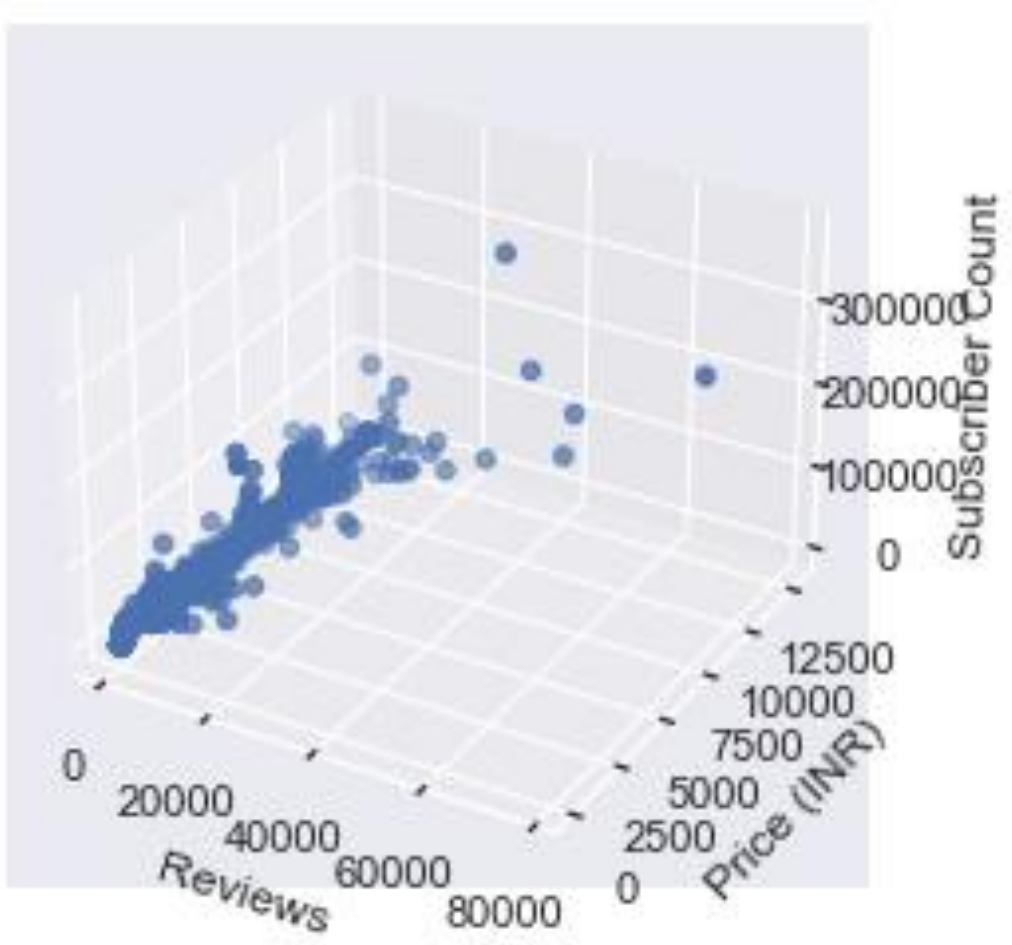


Faculty of Computers and Data Science - General

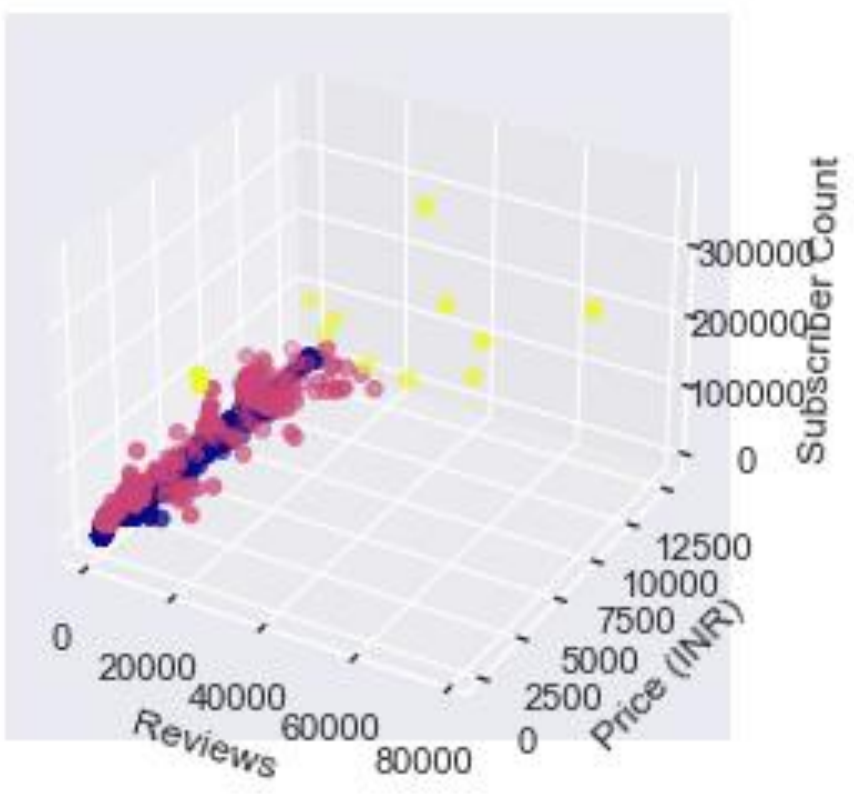
Data science methodology project

We wanted to divide the courses into clusters in order to analyze them:

First, we plotted the data in a 3d scatter plot with Price, number of reviews and the number of subscribers being the axes.



After that, we used K-means clustering package from sklearn and set the number of clusters to 3 which results in the plot below:



- The first cluster (blue) represents most courses on the website which has lower than 50000 subscribers and less than 10000 reviews with varying prices.
- The second cluster (red) represents the courses that have moderate popularity (less than 100000 subscribers) but has higher number of lectures on average than the first cluster with very few courses exceeding 10000 INR in price.
- The third and the smallest cluster containing only 12 courses represents the most popular courses (over 100000 subscribers and up to 350000 subscribers!) with 2 of the courses priced less than 2000 INR and the rest of the courses falling approximately within the price range 9000 to 10000 INR with no free courses in the cluster and having at least 10000 reviews.