**DATA MANAGEMENT AND EXPLORATORY DATA ANALYSIS**

**(REFLECTIVE LOG)**

**Student Name: AKASH MOHANDOSS**

**Student ID: 220488118**

**Introduction**

The Dataset provided by the MOOC futurelearn platform was vast and too tidy. The first step to be done was that since it was a group of 8 datasets obtained from 7 runs, we merged the required dataset into a single file dataset to carry out our analysis. We must also use the CRISP-DM framework to analyze the data, identify the business objectives and then try to achieve the business goals and present them to the business.

**Data Preparation:**

The provided dataset is a bunch of 8 records of 7 runs each data representing unique information. The only common parameter/ key that connects us with every dataset is the learner\_id. Based on the business understanding we utilized the enrolment data to identify solutions for the business.

Firstly, I took the enrolment data of all 7 runs. Merged those into a single dataset and then further used this table to do my analysis for the business. The dataset contained a lot of tidy values such as NULL, NA, and Unknown. These values need to be removed to perform our analysis of the data and to get accurate results. The more tidiness presents the worst the output. the tidy values are removed very carefully and finally, we get a pre-processed dataset

**Interpretation of the modelling**

From the insights that I gained from the data, my chosen business objective was to predict the count of individuals who has enrolled in the course for a particular time frame. Then on the second cycle, we need to analyse and find the reason for which people have opted to drop the course.

Basic EDA has been applied to the dataset provided. A scatterplot is generated for the data obtained by combining all datasets. It can be used to find the relation between the features available which can be used to obtain any possible solution to the business objective.

I have a bar plot to display the count of individual who has enrolled in the course in the first plot and in the second plot I have used it to show the male-female proportion/ count who has enrolled for the time period, the course has been enrolled.

Then in the second cycle, I am trying to plot a bar plot showing the count of people who has given a particular reason. In the next plot, I have plotted the count of people giving a particular reason during the time period in which they have left the course. Thus we can say that people are dropping out of course works or taking the first step to joining the course during the time constraint so, the future learning platform can improve the learning structure that incorporates time constraints that helps working literates as well other students who are trying to complete course work simultaneously doing their job to easily complete their module within the module time frame.

**Information Obtained**

Insights obtained through applying the CRISP-DM framework are very accurate. Through performing normal analysis on the merged dataset, we were able to obtain useful information about the chosen columns how they are related, what information can we get through building a model and how can we use it for our future analysis. Further through CRISP-DM phases, we built a model to show the proportion of people enrolled in the course and how it varied during the time period for which the data was captured. This goes with the first cycle and on the second we used this information to frame our second business objective and displayed a bar plot that shows the count of people who gave a different reason for their unenrolment. Another mentioned the count of people opting out of the course during a particular time period was also displayed using a bar plot.

**Milestones**

Data preparation and data cleaning was the hardest part of all the phases of the CRISP-DM lifecycle. Since the data is authentic and is real-time data. There was a huge chunk of NA, Unknown, and NULL values in the data. This made the work too hard to remove the junk value. Each column needs to be analysed and column selection must be done very carefully in order to obtain authentic data. However, using the CRISP-DM framework I was able to complete the analysis in a very effective and was able to obtain the business objectives as planned