SESSION **2**

KING MONGKUT’S INSTITUTE OF TECHNOLOGY LATKRABANG
FACULTY OF ENGINEERING
DEPARTMENT OF ROBOTICS&AI ENGINEERING

01416516 – Robotic Laboratory 3
Assignment 1

INSTRUCTED BY: SUSHISH BARAL

64011740 – SAUNG HNIN PHYU

DATE OF SUBMISSION: 20/3/2023

Task 1: Data Acquistion & Preparation

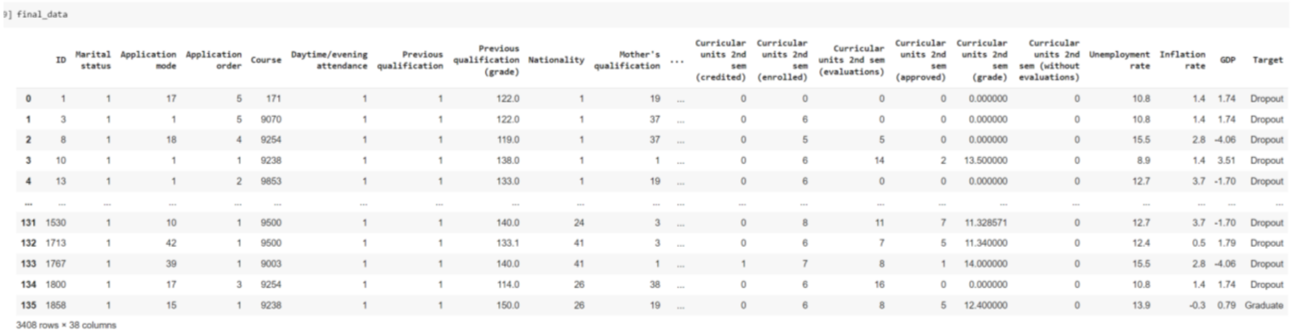
First of all, I imported pandas, numpy, matplot and seaborn libraries to the google collaboratory file. After that, I connected with google drive where I uploaded the data frames and loaded each data frame by using "pd.read_csv" command.

When merging the three data frames, I combined the first two data frames with "pd.merge()" function since they have the same set of students but two distinct sets of attributes. Then, I used "pd.concat()" function for the rest because in data 3, we have different student data sets. For this, I named the new data frame as "data".

When I imported all of the data that given by using pd.read_csv, the data didn’t show clearly so I added sep = ‘;’ to print these data and show all information. Then check the information data, type data, and missing values by data1.head() and data1.info().

While cleaning the data, I applied with "isnull" function first to check we have a value or not in our data. Afterward, I dropped duplicates data by using "drop_duplicate" command and null values by using "dropna" command. For removing impossible values, "loc" command is used. After all, I removed white spaces and checked the spelling.

This is my final data:



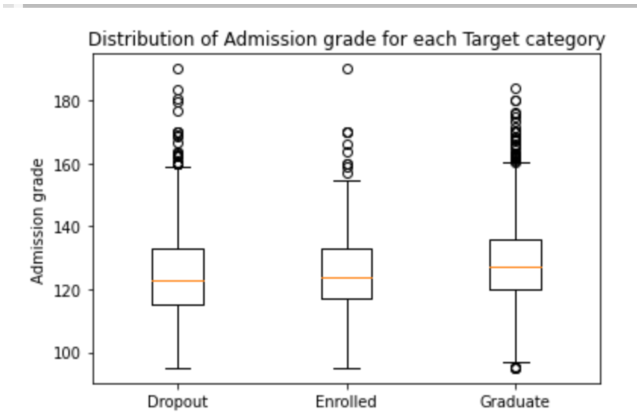
```
9] final_data
```

	ID	Marital status	Application mode	Application order	Course	Daytime/evening attendance	Previous qualification	Previous qualification (grade)	Nationality	Mother's qualification	...	Curricular units 2nd sem (credited)	Curricular units 2nd sem (enrolled)	Curricular units 2nd sem (evaluations)	Curricular units 2nd sem (approved)	Curricular units 2nd sem (grade)	Curricular units 2nd sem (without evaluations)	Unemployment rate	Inflation rate	GDP	Target
0	1	1	17	5	171	1	1	122.0	1	19	...	0	0	0	0	0.000000	0	10.8	1.4	1.74	Dropout
1	3	1	1	5	9070	1	1	122.0	1	37	...	0	6	0	0	0.000000	0	10.8	1.4	1.74	Dropout
2	8	1	18	4	9254	1	1	119.0	1	37	...	0	5	5	0	0.000000	0	15.5	2.8	-4.06	Dropout
3	10	1	1	1	9238	1	1	138.0	1	1	...	0	6	14	2	13.500000	0	8.9	1.4	3.51	Dropout
4	13	1	1	2	9853	1	1	133.0	1	19	...	0	6	0	0	0.000000	0	12.7	3.7	-1.70	Dropout
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
131	1530	1	10	1	9500	1	1	140.0	24	3	...	0	8	11	7	11.328571	0	12.7	3.7	-1.70	Dropout
132	1713	1	42	1	9500	1	1	133.1	41	3	...	0	6	7	5	11.340000	0	12.4	0.5	1.79	Dropout
133	1767	1	39	1	9003	1	1	140.0	41	1	...	1	7	8	1	14.000000	0	15.5	2.8	-4.06	Dropout
134	1800	1	17	3	9254	1	1	114.0	26	38	...	0	6	16	0	0.000000	0	10.8	1.4	1.74	Dropout
135	1858	1	15	1	9238	1	1	150.0	26	19	...	0	6	8	5	12.400000	0	13.9	-0.3	0.79	Graduate

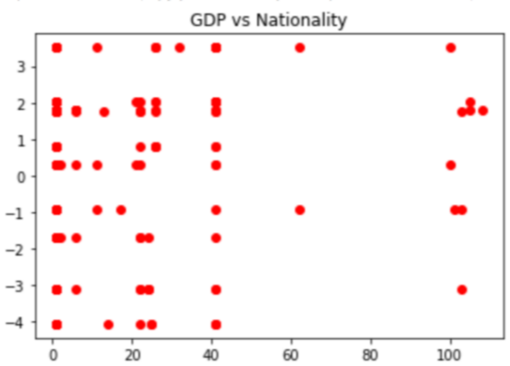
3408 rows × 38 columns

Task 2: Data Exploration

Sub Task 2.1:

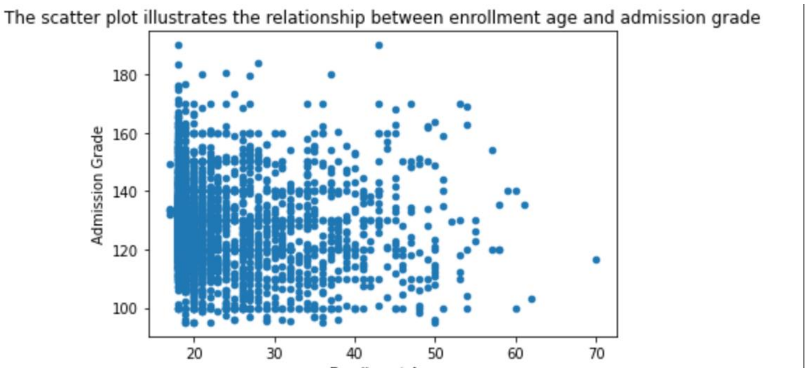


In the first question, we are asked to choose one categorical and numerical value each and visualize in the appropriate way. So, I chose 'Target' as 'Categorical' and 'Admission Grade' as 'Numerical' and constructed a box plot for that which I intended to see how the admission grade among the dropout, enrolled and graduated.

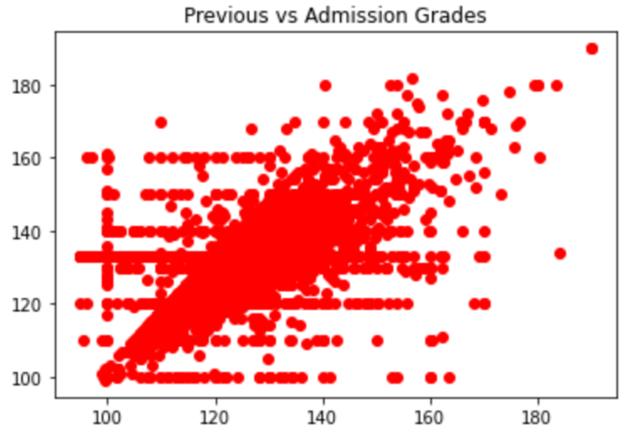


As for the second column pairs, I chose GDP vs Nationality. By seeing the graph, we can conclude which nation has higher GDP values. We can conclude that there is the most dense between 0 and 20 nationalities.

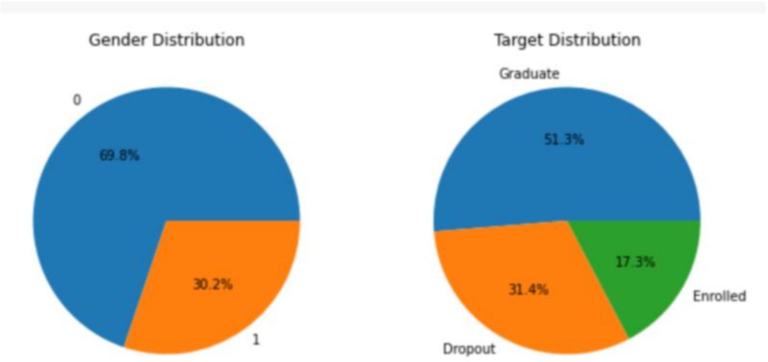
Sub Task 2.2:



I chose Age at Enrollment and Admission Grade for this task to see at which age has the more admission grade and the less admission grade. By looking at the scatter plot, we can see that there has the highest admission grade at the age of 20.

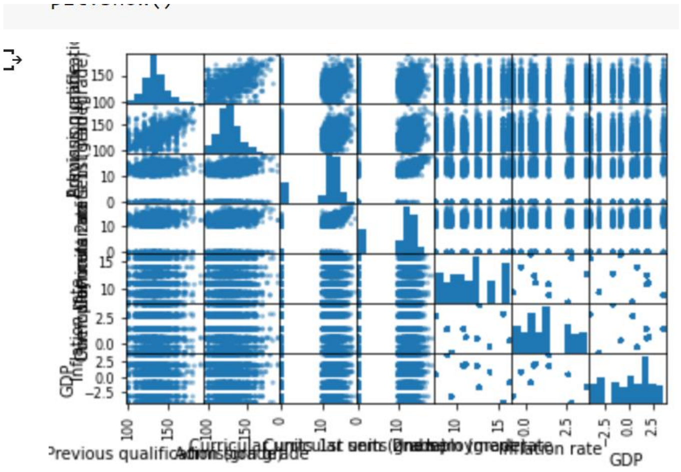


Secondly, I plotted the previous and admission grades to see the positive correlation between them.



I made a pie chart for the last column pair which are Gender and Target. So, we can know the percentage of gender and target distribution by looking at these charts. The reason why these two categories were chosen was because it was an easy way to see the capabilities of each student. We can see the capabilities of students between potential and score.

Sub Task 2.3:



In the question, it is asked to plot scatter matrix for all numerical values. According to the data, numerical values are 'Previous qualification (grade)', ' Admission grade', ' Curricular units 1st sem (grade)', ' Curricular units 2nd sem (grade)', ' Unemployment rate', ' Inflation rate' and ' GDP'. I plotted all these data by using 'pd.plotting.scatter_matrix' command. We can see histograms, scatterplot and line graph from these. We can conclude that this scatter matrix of this showed no correlation or if it were to have a positive correlation, it would be an extremely weak positive correlation of the data.