EDA on Student Enrolment Data

COS60008 Introduction to Data Science, 2023, Semester 1, Assignment 1 Report

Name: Raj Kishor Singh Naruka
Student ID: 104216694
Email: 104216694@student.swin.edu.au
Submission Date: 17-04-2023

Introduction

This report aims to communicate the findings during the analysis of university undergraduate student enrolment datasets. The data source for this analysis was three separate data files “data1.csv”, “data2.csv”, and “data3.csv containing students’ data. The whole process consists of data acquisition, data cleaning, data processing and exploratory data analysis separated into two different tasks. This report details the data acquisition and preparation process, data cleaning issues and methods used, and the data exploration process, including the selected visualisation methods and interesting findings. The report aims to provide a clear and concise overview of the project, highlighting key insights.

Task 1: Data Acquisition & Preparation

The data source for this analysis was three separate data files “data1.csv”, “data2.csv”, and “data3.csv containing students’ data. The files “data1.csv” and “data2.csv” contain the same set of students but distinct sets of attributes for describing the student, where each student has its unique ID. The file “data3.csv” contains a different set of students with each student described by all attributes from both “data1.csv” and “data2.csv”.

Approach

- Loaded all three data files into separate dataframes and then create one dataframe, which contains information of all the students.
- Handled scenarios with data entry acquisition errors
- Handled missing values
- Handled data duplicacy

Observation:

- The files were separated by semi-colons **';'** and not by commas **','**. Therefore, I had to update our code to read the file correctly

Task 1.1 – Data Acquisition:

- df1 and df2 had different attributes of same students. Joined these two datasets on their IDs
- df3 consisted data for students that are not present in df. So I concatenated these records in df and get one dataframe that had all the data available

Task 1.2 – Data Cleaning

Observations

- Observed column datatypes as a starting point for data cleaning process
- Observed Datatype mismatch between "Mother's qualification" and "Father's qualification" columns
- Observed columns with 'object' datatype can give us better insight
- Observed that "Daytime/evening attendance" column had inconsistency in entered values. Values 1 and 0 were entered as number as well as text data. Also, the data had records with 'Y' value which should mean present/1
- Observed that "Father's qualification" column had inconsistency in entered values. Few values were entered as numbers where as we have some values as text
- Observed that "Father's qualification" column values consisted of whitespaces
- Observed that "Target" column had data inconsistency in entered values. Values were misspelled
- Checked numerical columns with skewd distribution and check if it is justified or is a data entry issue
- Following attributes had high skew: Nationality, Educational special needs, Age at enrollment, Curricular units 1st sem (without evaluations), Curricular units 2nd sem (without evaluations)
- Observed the value counts distribution for each columns

- Observed that most students had nationality 1, which is a possibility if the institution has a lot of students from home country. (This was a good insight however, was not considered as a data inconsistency)
- Observed that 51 out of 4424 students were with Educational special needs which is a general scenario for most institutions
- Observed that few students were more than 100 years old. Which was definitely an anomaly. Decided to handle this data inconsistency.
- Due to the lack of information on curricular columns. Assumption was made that the distribution was expected
- Observed that there are very few students that are older than 35 yrs. This again was a good observation but not an anomaly
- Checked for number of null values in each column
- Two columns, 'Tuition fees up to date' and 'International' had missing values
- Another column 'Nationality' was present in the dataset which could be related with column 'International' and be used to compute and impute missing values
- Both of these columns had high number of records with missing values. These records could not be dropped
- 'Tuition fees up to date' 'Tuition fees up to date' had no relation with any other attribute

Solutions

- Mapped the "Daytime/evening attendance" column values to respective numerical value
- Mapped the "Father's qualification" column values yo respective numerical value
- Removed Whitespaces from "Father's qualification" column
- Mapped the "Target" column values to respective original/correct spelling/value
- Looked at the data distribution graph for the Age at enrollment column (Figure 1)

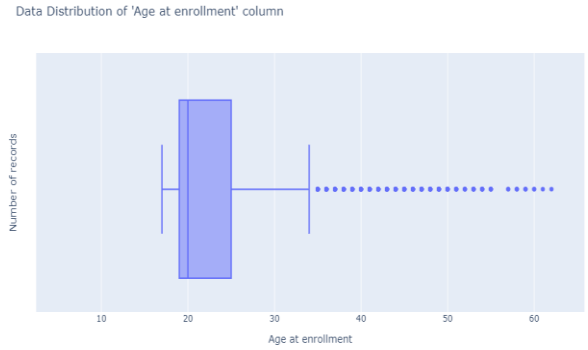

Figure 1: Data distribution of 'Age at enrollment' attribute

- For the Age at enrollment column, decided whether to impute the data inconsistency or drop the records based on the number of such records. If it were to be too high then we would impute else we would drop the records
- Such records were less than 1% of total records and were dropped.
- Checked the intuition that any student who had different nationality than the institution's home nation would have been be considered as an international student.
- The intuition was true. Students with different Nationality than 1 were considered International. Also, missing values were only present for students with Nationality 1
- Imputed International column based on Nationality
- Decided to keep the missing values in 'Tuition fees up to date' column as the number of records with None values were too high

Task 2: Data Exploration

Approach:

- Visualise at Data Distribution of some descriptive columns to better understand the data
- Visualise Look at Data Distribution of some descriptive columns to get an idea about students

Task 2.1 – Data Distributions

2.1.1

- Pie chart (Figure 2) to Observe Marital Status distribution. Decided to visualise pie chart because we were already aware of number of classes in this column which are not many and pie chart is a good visualization

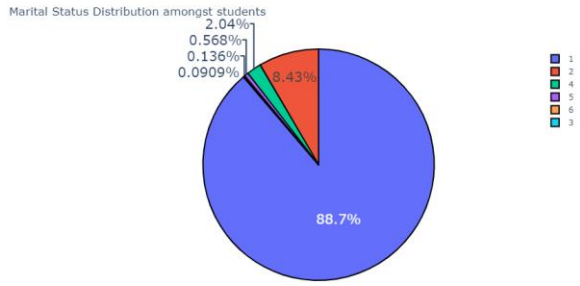

Figure 2: Marital Status Distribution amongst students

Observation

- Tere was a huge class imbalance in marital status
- 88.7% students had marital status 1. Which was expected as we do observe that most of the students pursuing undergraduate degrees are unmarried

Insight

- We can look at any relation with age as in general probability of a person being married increases as he/she ages(Figure 3)

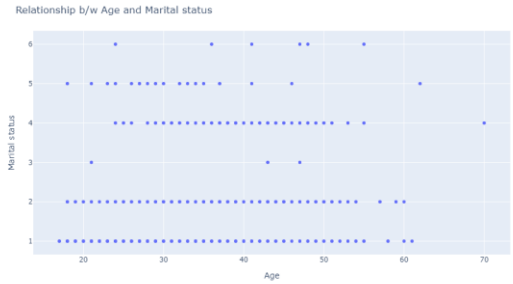

Figure 3: Relationship between Age and Marital Status

Observations

- There was no relationship between Age and Marital status of a student
- Will had to stop our analysis for this attribute as we do not have further information about the data/attribute

2.1.2

- Bar chart(Figure 4) to see the popularity of courses taken. Bar chart is the best chart for this purpose as we do have more classes than what a pie chart is suited for

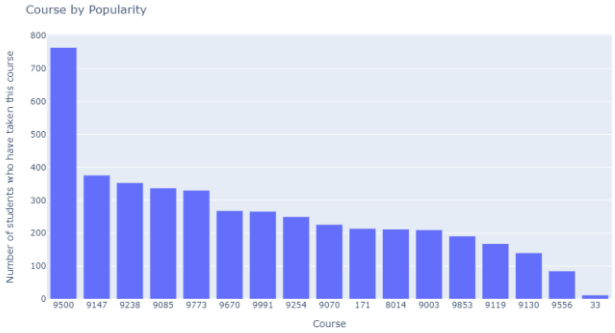

Figure 4: Course by Popularity

Observations

- The most and least popular courses were: course 9500 and course 33
- Course 33 only had 12 students enrolled and was an outlier to the distribution. This could be a newly introduced course
- Handled the datatype issue in Course attribute

2.1.3

- Pie chart(Figure 5) to look at Attendance distribution. Pie chart is suited for this task as we were already aware of number of classes in this column which are not many and pie chart is a good visualization

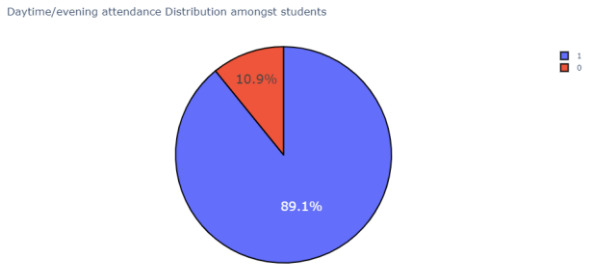

Figure 5: Attendance distribution amongst students

Observations

- Most of the students are present in class 1 (89.1)

2.1.4

- Comparing distribution of nationality with domestic/international students distribution(Figure 6) with bar chart and donut chart. These charts serve the purpose. (Highly encourage to zoom in the visualization in .ipynb file to verify the observations)

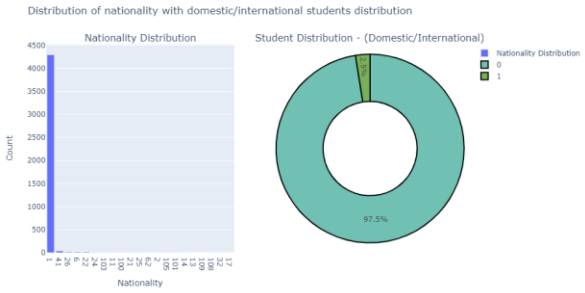

Figure 6: Nationality Distribution with Domestic/International Distribution

Observations

- Majority of the students (97.5%) were domestic students which have nationality 1
- Very few students (2.5%) were International students
- Naionality distribution among these international students showed that most students were from nation 41 (Please zoom in the graph to take a look)
- Handled the datatype issue with Nationality Attribute

2.1.5

- Visualized data distribution for the following attributes: Displaced, Educational special needs, Debtor, Scholarship holder. (Figure 7)

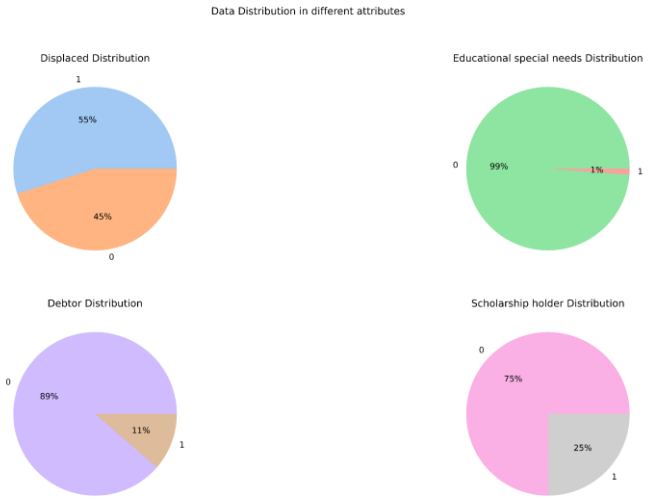

Figure 7: Data Distribution of descriptive categorical attributes

Observations:

- Very few students(1%) had Education Special Needs. which can be observed in real world as well
- 5% of students hold scholarship. Which is a fairly general scenario

2.1.6

- Visualized data distribution for Gender and Target Attributes(Figure 8)

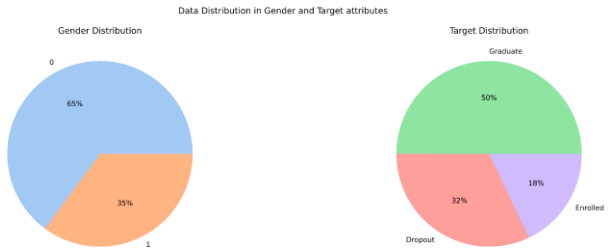

Figure 8: Data distribution of Gender and Target Attributes

Observations:

- Most students are of gender 1 (65%)
- Half of the students graduate followed by 32 of the students who drooped out and 18% which have status as enrolled

Task 2.2 – Relationships among attributes

2.2.1

- Checked countries with highest/lowest gdp as well as employment rate and Inflation rate.(Figure 9- 10).
- Generally we can expect a country with low gdp to have a higher unemployment rate or high Inflation rate
- Bar chart provides information about their metrics and the scatter_matrix gives insight about the relationship

Figure 9: GDP, Inflation Rate and Unemployment rate by Country

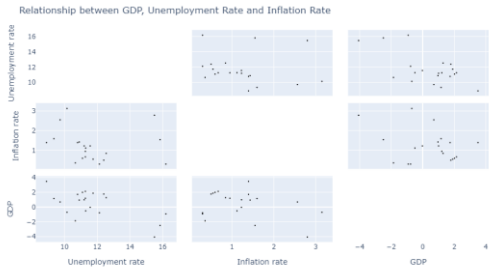

Figure 10: Relationship between GDP, Unemployment rate and Inflation rate

Observations

- Highest GDP student Nationality: 32
- Lowest GDP student Nationality: 14
- Found a negative correlation between GDP and Unemployment Rate(-0.7). This means that as the GDP increases, Unemplyment rate decreses
- Found a slight negative correlation between GDP and Inflation Rate(-0.43). This means that as the GDP increases, Inflation rate decreses
- Found No correlation between Unemplyment rate and Inflation Rate

2.2.2

- Distribution of Age with Gender(Figure 11). Histogram combined with boxplot gives us a great insight about the data spread

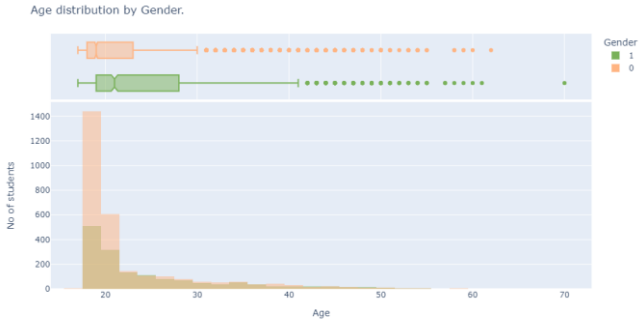

Figure 11: Age distribution by Gender

Observations:

- Both gender 0 and 1 had more number young students (18-21 years of age)
- Gnder 0 had more students than gender 1 in the young age group(18-21 yrs of age)
- Gender 1 had student age spread over long spectrum asin, the IRQ for gander 1 was greater than IQR of gender 0

2.2.3

- Performance trends of students through semesters(Figure 12)
- One can presume that a student who has scored good marks in first semester will do good in next semester as well
- Scatterplot is a good visualization to see the trend and relationship between grades of two semester

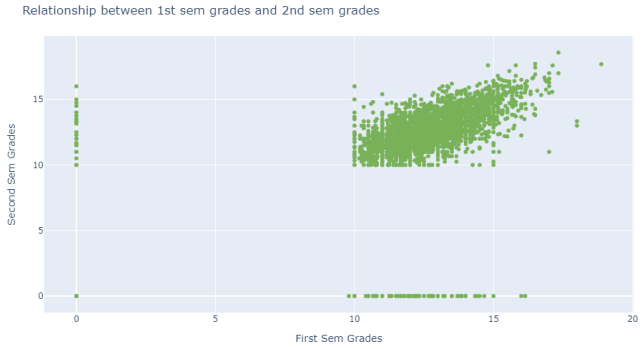

Figure 12: Relationship between 1st sem grades and 2nd sem grades

Observations:

- Few students had performed well in first semester but could not do well in second semester
- Few students could not perform well in first semester yet did good in second semester
- The general trend is that if a student had performed well in first semester, the performance is likely to persist through second semester as well(Correlation : 0.84)
- One student had scored 0 in both the semester he/she might need some special guidance

2.2.4

- Checking whether the same pattern applied to previous qualification as well.(Figure 13).
- Students have this fear of toppers who have performed good in their previous qualifications. It is good to check whether this holds true in the dataset or not
- Scatter matrix gives us ability to check relationship among more than two attributes at a time.

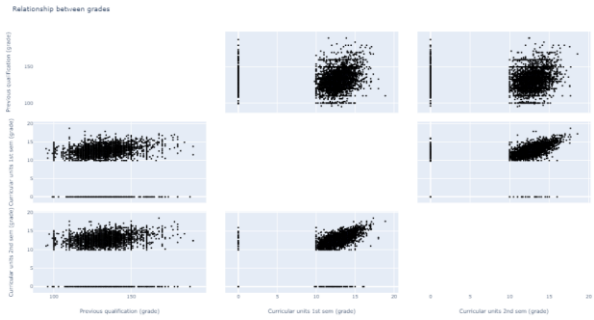

Figure 13: Relationship of university grades with previous qualification grades

Observation

- No such relation is present between (previous qualification grades - 1st semester grades) or (previous qualification grades - 2nd semester grades). (Correlation: <0.1)

Task 2.3 – Scatter Matrix

Chose the following six numerical columns to check for some relation:

1. 'Unemployment rate'
2. 'Inflation rate'
3. 'GDP'
4. 'Previous qualification (grade)'
5. 'Curricular units 1st sem (grade)'
6. 'Curricular units 2nd sem (grade)']

Scatter matrix(Figure 14) is used to check for relationship and heatmap(Figure 15) is used to visualise the correlation values.

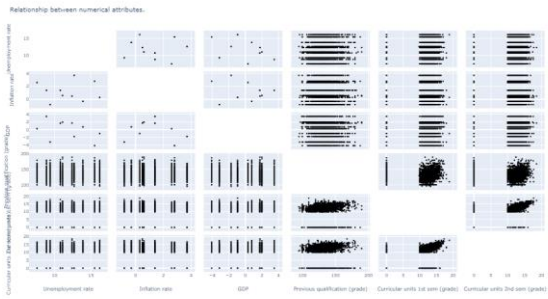

Figure 14: relationship between numerical variables

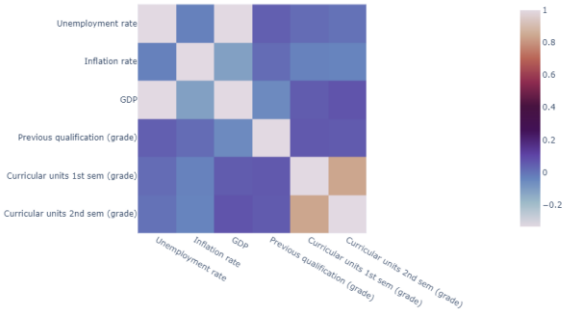

Figure 15: Heatmap for numerical variables' correlation scores

Observations:

- As we had seen previously, For the student performace attributes, 1st sem grades and 2nd sem grades have a positive correlation and previous ualification grades does not have any relationship
- As we had seen previously, GDP has negative correlation with Unemplyment rate and slight negative correlation with Inflation rate. Unemployment rate and Inflation rate share no such relationship

Conclusions

The data has some useful and interesting insights, which can help teachers to interact with students and motivate them to score better or encourage them on doing so, or target specific students in case they might need help. In addition, help the organisation take better decisions.
There is scope for more in depth analysis if the metadata is available of these data sources, which can help in making sense of some classes, which are not very descriptive.

References

- https://stackoverflow.com/questions/8924173/how-can-i-print-bold-text-in-python
- https://www.kaggle.com/code/rajnaruka0698/how-things-work-in-new-york
- https://www.kaggle.com/code/rajnaruka0698/zomato-hyderabad-analysis-and-prediction
- https://plotly.com/python/pie-charts/
- https://community.plotly.com/t/axis-labels-on-scatter-matrix/41865/3#fromHistory