RMIT UNIVERSITY

Associate Degree Pathway to Undergraduate Degree

Name	Student ID
Mutum Deva Neil Singh	S3773370

Contents

1. Introduction 3

2. Background 3

3. Data preparation 4

3.1. Demographic data 4

3.2. Student enrolment data 4

4. Performance analysis 6

4.1. Overall performance 7

4.1.1. Descriptive statistics and visualisation 7

4.1.2. Bootstrapping method 8

4.1.3. Hypothesis testing 9

4.1.4. Result 10

4.2. Performance in each year 10

4.2.1. Descriptive statistics and visualisation 11

4.2.2. Hypothesis testing 11

4.2.3. Result 12

4.3. Performance of students commencing in each year 12

4.3.1. Descriptive statistics and visualisation 13

4.3.2. Hypothesis testing 14

4.3.3. Result 15

4.4. Performance of students in selected programs 15

4.4.1. Descriptive statistics and visualisation 16

4.4.2. Hypothesis test 16

4.4.3. Result 17

4.5. Analysis 17

5. Visualisation proposal 17

6. Insights 19

6.1. Performance of students from each AD program in undergraduate program 19

6.2. AD program to undergraduate program transition 19

6.3. Insights from demographic data 20

7. Conclusion 24

8. Appendix 25

8.1. Project team 25

8.2. Self-reflection 25

1. Introduction

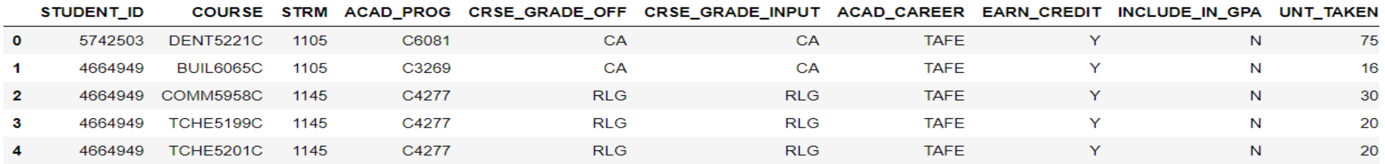
The partner organisation of this project is Royal Melbourne Institute of Technology (RMIT) University. The university is located in Melbourne, Australia and it is one of the oldest and top universities in the country. The learning analytics acts as a custodian of RMIT's data for education, who gather, analyse and present data to enhance learning and teaching experience. The university is very active in the field of data science with many industrial connections.

The aim of this project is to determine the performance of students taking the associate degree pathway to undergraduate degree. Performance analysis is to be done by comparing the undergraduate program scores of students taking the pathway and those students who enrol directly to an undergraduate program. Many useful insights are often overlooked therefore, another aim of this project is to determine meaningful insights that may help in making the student experience better. The project is data exploration based which is one of the most important aspect of data science. Therefore, the deliverables of this project are the statistical findings and visualisation to show the performance of AD and Non-AD students. Descriptive statistics along with hypothesis testing is required if we are to compare the central tendency of a data. Visualisation also help viewers understand the problem better therefore, appropriate visualisation technique is necessary for a good story-telling. The final aim of this project is to propose a visualisation that will show the pathway students' performance. A good visualisation is one that place meaning into complicated datasets so that their message is clear and concise. The visualisation proposed in this task will make the viewers understand the results obtained from main objective in a clear and concise way.

2. Background

According to the university, as associate degree is a two-year full-time qualification given to students at an undergraduate level. The aim of associate degree is to give students the basic technical and academic knowledge and transferable skills they need to go on to employment or further study in their field of interest. This means that a student who completes an associate degree has an option to undertake an undergraduate program. Students who do not have an associate degree can directly enrol for an undergraduate degree. So, in an undergraduate program, there may be two types of students – one who takes the pathway (AD students) and others who directly enrol (Non-AD students).

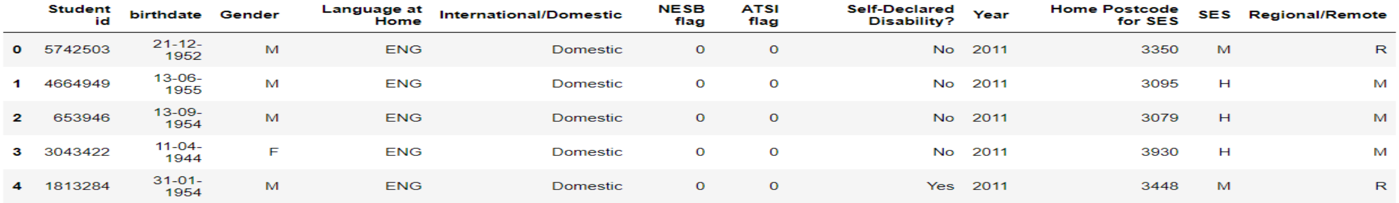
Two datasets were provided for this project – the student enrolment data and the demographic data. The student enrolment data contains information like - student id, program code, course code, grades and enrolment status. This data contains the information of both AD and Non-AD students. This data has a total of 17 attributes and 5,156,045 rows.



	STUDENT_ID	COURSE	STRM	ACAD_PROG	CRSE_GRADE_OFF	CRSE_GRADE_INPUT	ACAD_CAREER	EARN_CREDIT	INCLUDE_IN_GPA	UNT_TAKEN
0	5742503	DENT5221C	1105	C6081	CA	CA	TAFE	Y	N	75
1	4664949	BUIL6065C	1105	C3269	CA	CA	TAFE	Y	N	16
2	4664949	COMM5958C	1145	C4277	RLG	RLG	TAFE	Y	N	30
3	4664949	TCHE5199C	1145	C4277	RLG	RLG	TAFE	Y	N	20
4	4664949	TCHE5201C	1145	C4277	RLG	RLG	TAFE	Y	N	20

Figure 1: The student enrolment dataset.

The demographic data contains information of each students like – birthdate, gender, socio economic status (SES) and international/domestic. This data has a total of 12 attributes and 1,048,575 rows.



	Student id	birthdate	Gender	Language at Home	International/Domestic	NESB flag	ATSI flag	Self-Declared Disability?	Year	Home Postcode for SES	SES	Regional/Remote
0	5742503	21-12-1952	M	ENG	Domestic	0	0	No	2011	3350	M	R
1	4664949	13-06-1955	M	ENG	Domestic	0	0	No	2011	3095	H	M
2	653946	13-09-1954	M	ENG	Domestic	0	0	No	2011	3079	H	M
3	3043422	11-04-1944	F	ENG	Domestic	0	0	No	2011	3930	H	M
4	1813284	31-01-1954	M	ENG	Domestic	0	0	Yes	2011	3448	M	R

Figure 2: The demographic dataset.

3. Data preparation

Data preparation is an important data science technique which refers to the process of cleaning and transforming the data before performing any analysis. The process involves steps like imputing missing values, reformatting the data and combining the data. There were a lot of issues in the data that was provided for performance analysis. This process was performed in two phases in this project – one for the demographic data and the other for the student enrolment data.

3.1. Demographic data

The gender and SES attributes of the demographic data contains values that could confuse someone not familiar with the data. Due to this, the values had to be renamed for better understanding. Values of gender attribute were changed from M, F, X and U to "Male", "Female" and "Other" while values of SES attribute were changed from H, M, L and ND to "High", "Medium", "Low" and "No Data". Missing values are generally imputed with values provided, meaningful value or by the mean, median or mode. SES attribute contained some missing values which were imputed to "No Data" because missing values of SES means that we do not have the data for that particular student which is equivalent to "No Data". Finally, duplicate records were removed and attribute selection was done to select only those attributes that would be useful during the analysis process. Gender, SES and International/Domestic attributes were selected for analysis.

3.2. Student enrolment data

The student enrolment data contains information of all the students enrolled from 2011 to 2020. It also contains information of postgraduate students which is irrelevant for this project as we are interested in undergraduate performance only. The first step of preparing the student enrolment data was to filter out students in undergraduate program until the year 2019 as 2020 is the ongoing session due to which no grades are registered. Further filtration was done to only select students that were enrolled in a program and to only include courses whose grades would be included in the GPA. Although the "STRM" attribute of the data contains information about the year and semester, both the records are combined in a single attribute. For example, a student in second semester of 2014 has a value of 1452 in the STRM attribute. The first two digits represent the year and the last two digits represent the semester. The second step of preparing the student enrolment data was to create two new attributes "Year" and "Term". These attributes contain information about the year and the semester the student is in. The values of these two attributes depends on the value in the STRM column. For example, 1452 STRM value means that the year attribute would have a value of 2014 and the term attribute would have a value of "Term 2". Having two new attributes for year and term rather than one attribute containing the combined information of year and term is better as it is easy to understand.

The analysis to be done was on the scores achieved by students in each course so it was necessary that we knew the scores achieved in each course. The "CRSE_GRADE_INPUT" contains the scores achieved but, in some cases, instead of the score, grades like HD, DI etc were given. A new attribute "SCORE" was created according to the CRSE_GRADE_INPUT attribute. For cases, where grades were given in place of score, the grades were converted to score using information provided. The table shown below shows the grades and their equivalent score which were used to covert the grades to score

Grade	Equivalent score
HD	80
DI	70
CC	60
NN	0

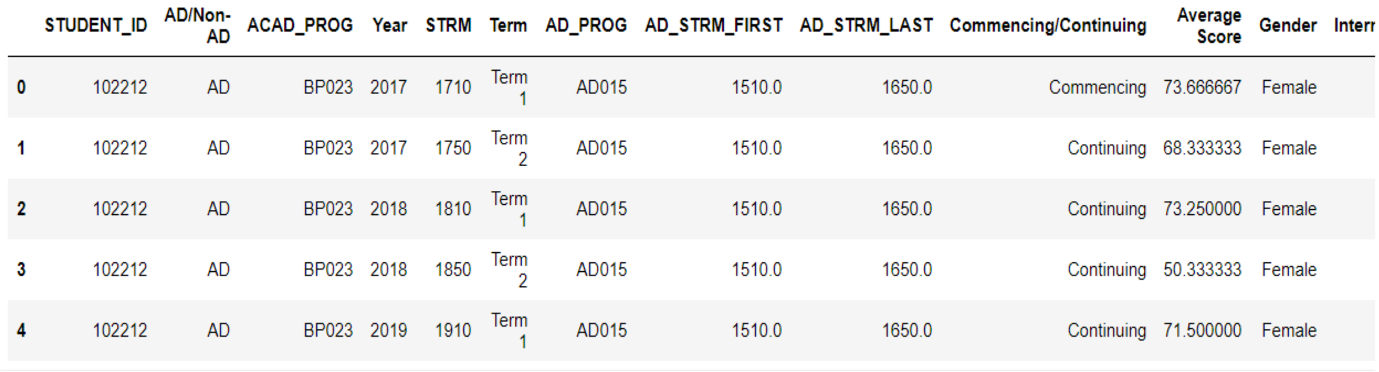
Table 1: Grades and their equivalent score.

The datatype of the new attribute SCORE had to be changed from object to integer so that it could be used to finding the measure of central tendency during analysis. From the 20 attributes, only 6 attributes were selected because only these selected attributes were required for further analysis. The selected attributes were – student id, year, term, program code, course code and score.

The student enrolment data contains information of students of both AD and Non-AD cohort. To compare the performances of AD and Non-AD, categorising each student to either AD or Non-AD was required. Record of each individual had to be checked before categorising them. From the data, two new data were created using filter – one data contained the records of AD program and the other contained the records of undergraduate program. The idea of creating these two data was that – if a student took the AD pathway, the record of that student would be present in both the data. This was achieved by performing set intersection of both data. The output of the intersection was the set of student id of students who were present in both the data. A new data was created by filtering the students whose id were present in the set achieved after performing intersection. This new data created contained the information of all those students taking the AD pathway. Another data was created that contained information of students who enrolled directly. From the dataset containing information about AD pathway students, results of AD program were removed because the analysis is to be done on these students' performance in undergraduate program. A new attribute was created and named "AD program" that contain the program code of the AD program from which the student came.

A new attribute "Commencing/Continuing" was created in both AD and Non-AD datasets. The idea of creating this attribute is to find out students who belong to the same batch. A student commencing in a program in the term 1 of 2014 will have 'Commencing' as its value in the "Commencing/Continuing" attribute and for subsequent terms, the value of the attribute will be "Continuing". "AD/Non-AD" attribute was created in both AD and Non-AD datasets with values "AD" for AD dataset and "Non-AD" for Non-AD dataset. With this attribute, grouping students who took the pathway and who did not will be easier. The AD and Non-AD datasets were concatenated to form a single dataset. This new dataset was then merged with the demographic data to create the final dataset that will be used for analysis.

The data preparation consumed majority of the project duration because of the complex procedures. Many iterative loops were required which took time because of the large size of the data. Many complex logical implementations were also required to perform the filter and also to create the new attributes. Implementation of set operation also made this process more complex.



	STUDENT_ID	AD/Non-AD	ACAD_PROG	Year	STRM	Term	AD_PROG	AD_STRM_FIRST	AD_STRM_LAST	Commencing/Continuing	Average Score	Gender	Interr
0	102212	AD	BP023	2017	1710	Term 1	AD015	1510.0	1650.0	Commencing	73.666667	Female	
1	102212	AD	BP023	2017	1750	Term 2	AD015	1510.0	1650.0	Continuing	68.333333	Female	
2	102212	AD	BP023	2018	1810	Term 1	AD015	1510.0	1650.0	Continuing	73.250000	Female	
3	102212	AD	BP023	2018	1850	Term 2	AD015	1510.0	1650.0	Continuing	50.333333	Female	
4	102212	AD	BP023	2019	1910	Term 1	AD015	1510.0	1650.0	Continuing	71.500000	Female	

Figure 3: The final dataset after data preparation.

4. Performance analysis

To achieve the aim of this project, comparison of performance of AD pathway students and Non-AD students was required. The performance comparison in this project was done by comparing the scores of students. Descriptive statistics, bootstrapping method, statistical hypothesis tests and visualisation techniques were used to perform each analysis in this project.

Descriptive statistics include measure like mean, median and standard deviation. Mean and median can be used to get some idea about the distribution of the data. Moreover, means of two different cohort can give us some information about which group performed better. Incorporating visualisations with descriptive statistics helps while comparing the cohort. Normal bar plots and grouped bar plots were used in this project to visualise the difference in score of AD and Non-AD students. Bootstrapping method is a resampling technique used to estimate the statistics on a data by performing sampling with replacement. In this project, bootstrapping method was used to determine the mean score.

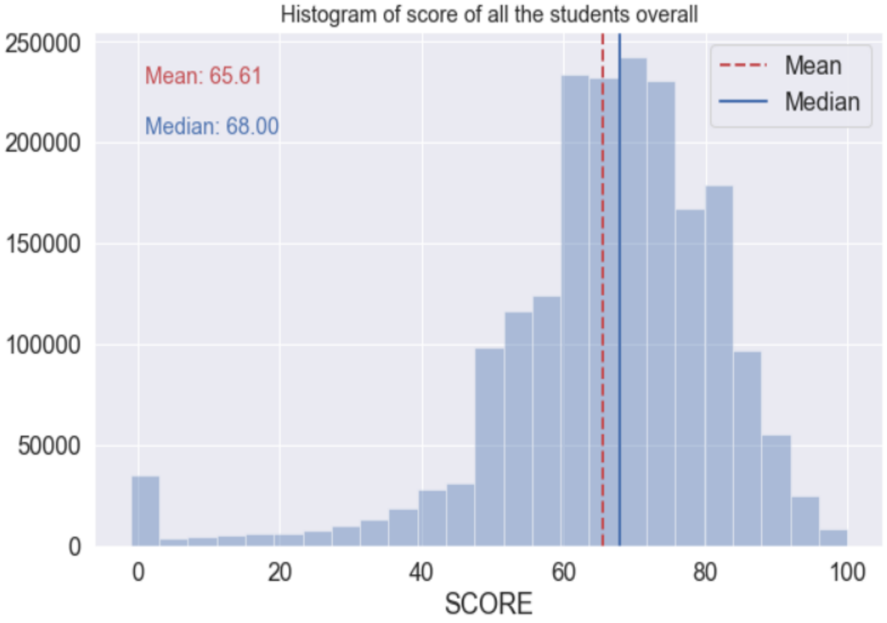


Figure 4: Overall distribution of score.

Figure 4 shows the distribution of overall score of all the students. The score attribute in this dataset is continuous due to which histogram was chosen to show the distribution of score. It can be seen that the histogram is skewed to the left which suggests that the data is not normally distributed. Also, the fact that the mean is lower than median suggests that the data is indeed not normal.

Hypothesis test can be used to determine whether the difference between two or groups is statistically significant. There are two types of hypothesis testing – parametric and non-parametric tests. Parametric tests assume that the data is normal which is not the case in this project. Non-parametric tests do not assume about data normality due to which it can work on non-normal data. Therefore, in this project non-parametric tests shall be used. From the various non-parametric tests available, Kruskal-Wallis test was chosen because this test is used for comparing two or more independent samples of equal or different sample size.

Four analyses were performed to determine whether a student from AD pathway performed better than Non-AD students. The four analyses performed to achieve the aim are:

- Overall performance comparison.
- Performance of AD and Non-AD students in each year.
- Performance of AD and Non-AD students commencing in each year.
- Performance of AD and Non-AD students selected AD programs.

4.1. Overall performance

Overall performance comparison was done by taking into account all the scores of AD and Non-AD students from the year 2011 until 2019. Using this data, three different techniques were applied to determine whether students from AD pathway perform better than Non-AD students.

4.1.1. Descriptive statistics and visualisation

The descriptive statistics used for analysing the overall performance are: mean, median and standard deviation. Visualisations used are countplot and barplot. The count plot in figure 5 shows the count of unique students AD and Non-AD students. It is clearly visible that the difference in the number of unique students in AD and Non-AD is quite high due to which comparison of performance on the basis of score is better than comparison using the number of courses passed or failed by students of each group.

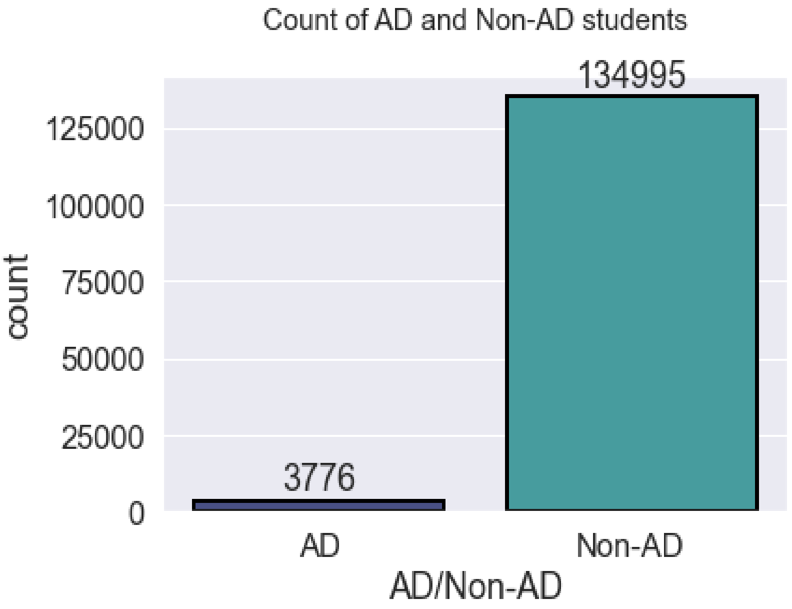


Figure 5: Count of unique students in each group.

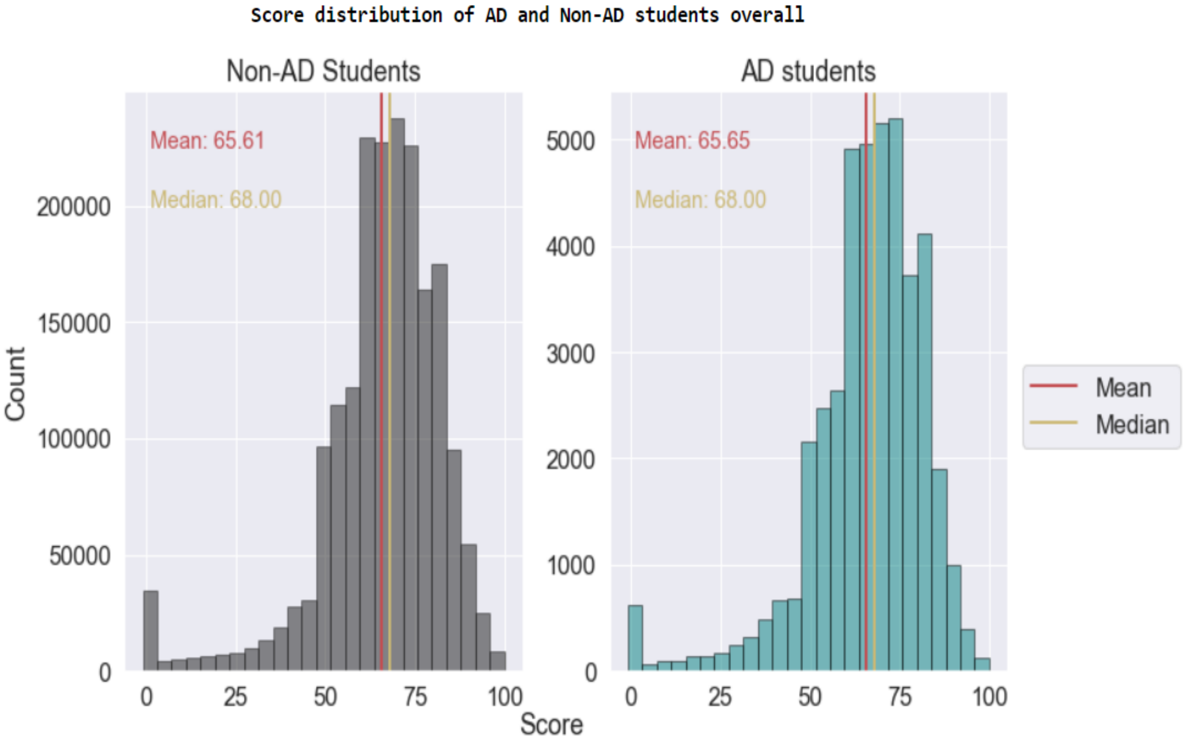
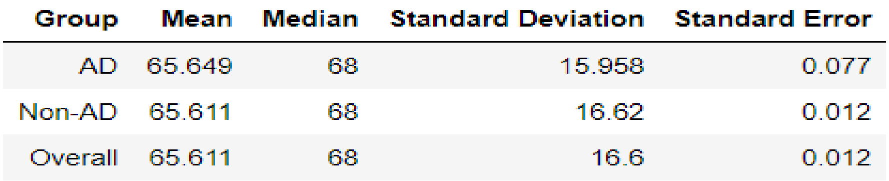


Figure 6: Score distribution of AD and Non-AD.



Group	Mean	Median	Standard Deviation	Standard Error
AD	65.649	68	15.958	0.077
Non-AD	65.611	68	16.62	0.012
Overall	65.611	68	16.6	0.012

Figure 7: Descriptive statistics.

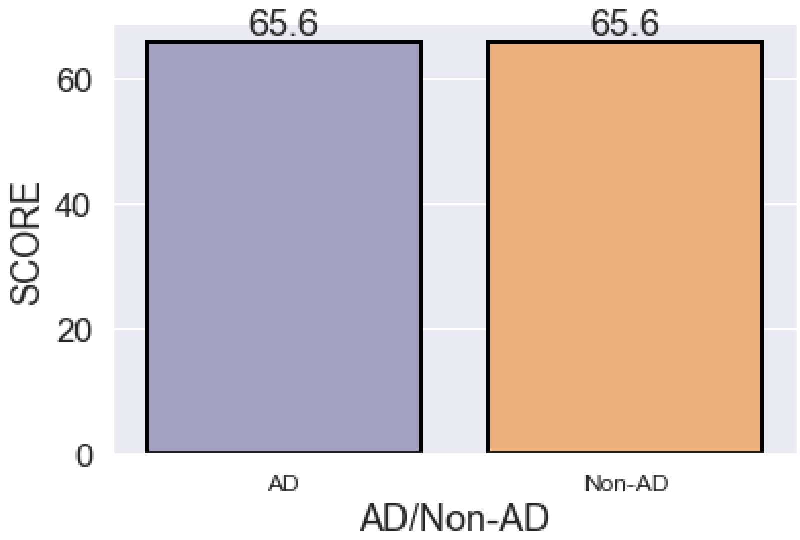


Figure 8: Mean score visualisation of AD and Non-AD students.

From the overall analysis using descriptive statistics, the result obtained shows that there is very small difference in the mean score of AD and Non-AD students. The median score of both the cohort is same due to which we can assume that the score distribution of AD and Non-AD may be similar. The standard deviation of Non-AD students is higher than the standard deviation of AD students.

4.1.2. Bootstrapping method

For bootstrapping method, the final dataset was split into AD and Non-AD cohort and bootstrap was applied on both the cohort. An iterative loop was used to specify the number of samples for bootstrapping method. The number of samples were initialized from 1 to 40,000 samples. Sampling with replacement was done for each number of samples and a new dataset is created that stores the number of samples and the mean of that sample group.

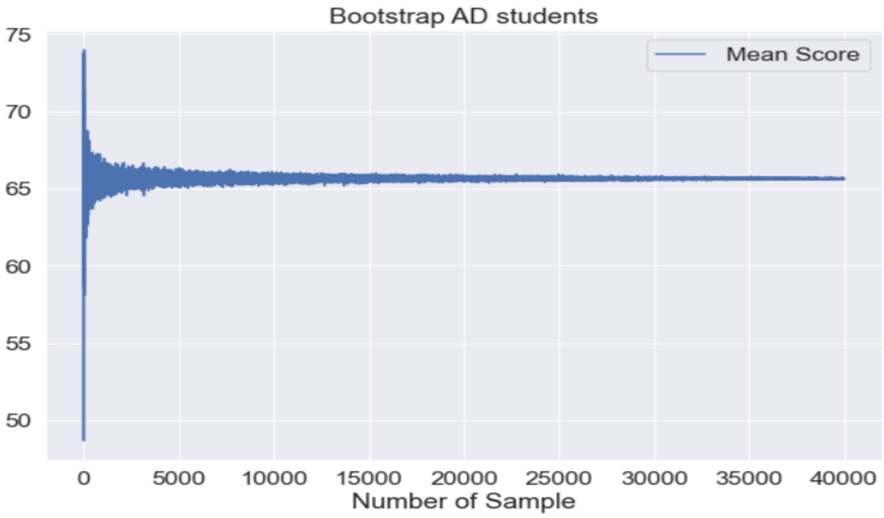


Figure 9: Bootstrap for AD students.

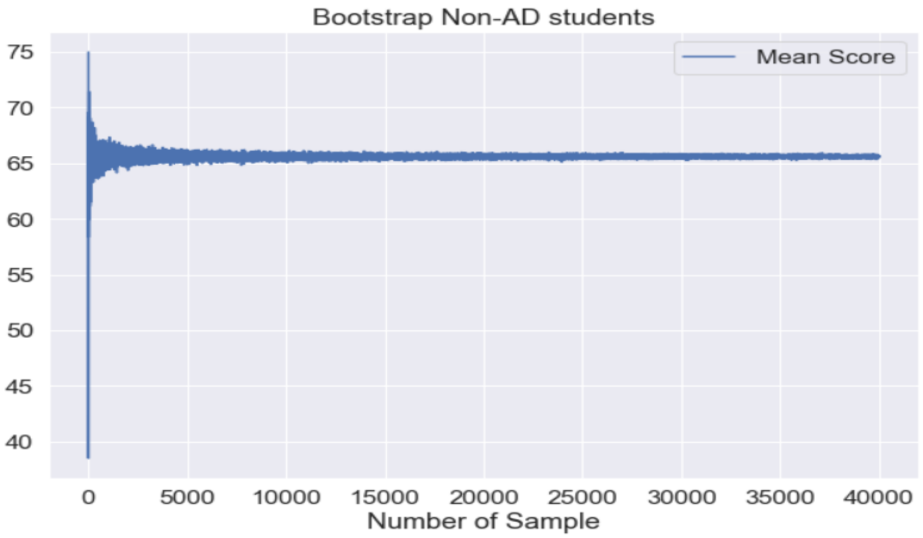
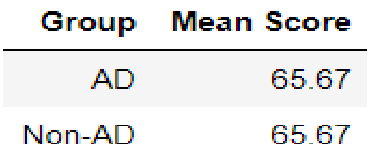


Figure 10: Bootstrap for Non-AD students.



Group	Mean Score
AD	65.67
Non-AD	65.67

Figure 11: Mean score of the last sample cohort.

From the bootstrap plots, the mean score for both the cohort looks same. The mean score of the last sample group shown in figure 11 shows that there is no difference between the scores of AD and Non-AD students. Applying bootstrapping method for each group takes long time. It took around 30 minutes for each group which is a big disadvantage in terms of run time.

4.1.3. Hypothesis testing

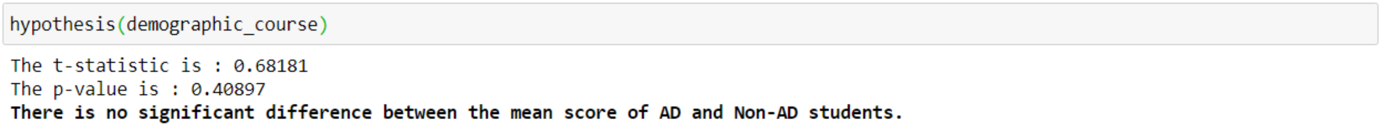
Kruskal-Wallis test in a non-parametric test that was chosen for hypothesis testing because of the non-normal distribution of the score. Another advantage of Kruskal-Wallis test is that it can compare two or more cohort with different sample sizes. The test was performed using the “Kruskal” function present in “scipy.stats” module in python.

A hypothesis testing has a null hypothesis (H0) and an alternate hypothesis (HA). For this project, the null and alternate hypotheses are:

Null hypothesis (H0): There is no statistically significant difference between the scores of AD and Non-AD students.

Alternate hypothesis (HA): There is statistically significant difference between the scores of AD and Non-AD students.

In hypothesis testing, if the p-value obtained is less than the confidence level, then the test in statistically significant and if it is greater than the confidence level, the test in not statistically significant.



```
hypothesis(demographic_course)
```

```
The t-statistic is : 0.68181
The p-value is : 0.40897
There is no significant difference between the mean score of AD and Non-AD students.
```

Figure 12: Result of Kruskal-Wallis test.

The p-value obtained from Kruskal-Wallis test is greater than the confidence level which shows that the difference between the scores of AD and Non-AD students is statistically insignificant.

4.1.4. Result

The overall analysis was done using three techniques – descriptive statistics and visualisation, bootstrap method and hypothesis testing. The finding of descriptive statistics and visualisation shows that there was almost no difference between the scores of AD and Non-AD. The bootstrap method also showed no difference in the scores. The p-value obtained from the hypothesis proves the findings of the previous techniques that the difference between the scores of AD and Non-AD is statistically insignificant. Therefore, we can say that, both AD and Non-AD students performed equally overall.

4.2. Performance in each year

The difference in sample size of AD and Non-AD students is very large. The aim of this analysis is to compare the performance of students in each year and also decreasing the difference in sample sizes. In this task, descriptive statistics and visualisations along with hypothesis testing were performed to analyse the performance in each year. Bar plots and grouped bar plots were used to compare the means of the two cohort. Bootstrap method was not applied because of the time taken to find the mean for each group. From the count plots in figure 13, we can see that the difference in sample size is still large. The number of AD students in each year is very less as compared to Non-AD students. Number of AD students seems to have increased over the years and number of Non-AD students decreased in the later years.

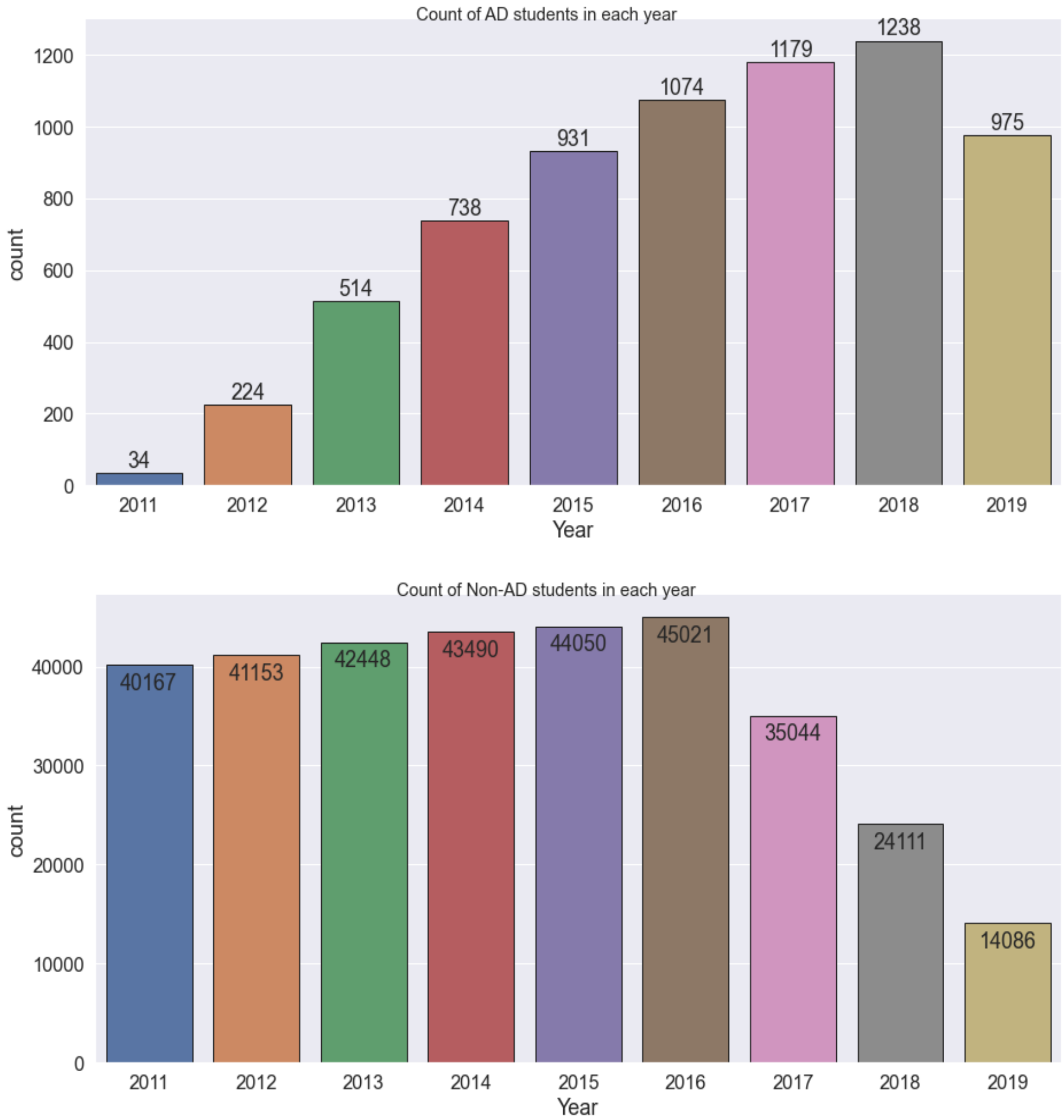
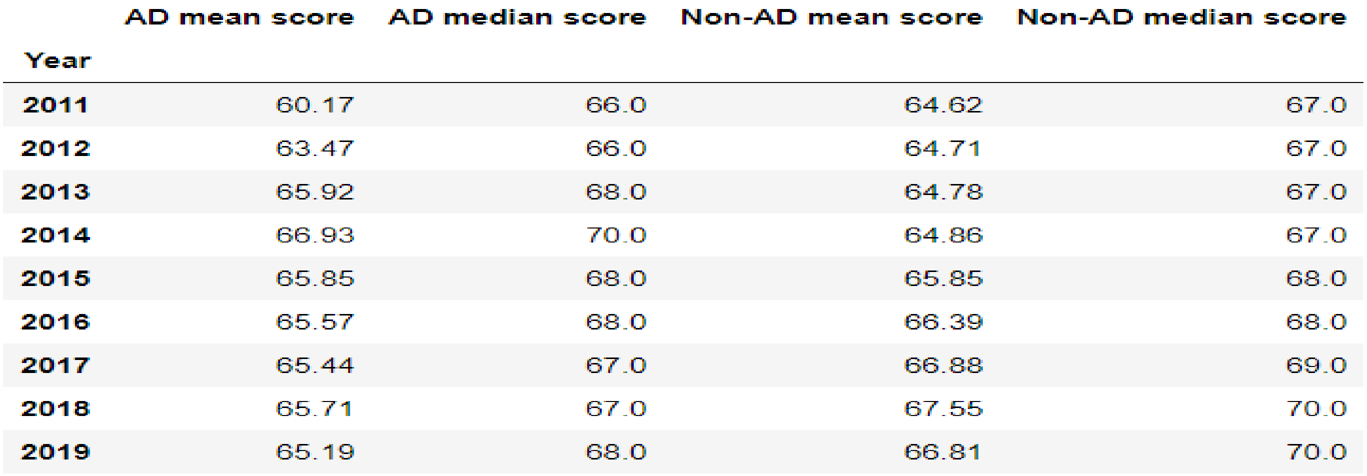


Figure 13: Count of AD and Non-AD students in each year.

4.2.1. Descriptive statistics and visualisation

Mean and median along with mean score visualisation using bar plots were used to compare the performance of AD and Non-AD students. To find the descriptive statistics of each year, iterative loop was run for each unique year in the dataset and for a unique year, the statistics were determined using mean () and median () functions in python. The year and the statistics were then used to create a dataset shown in figure 14.



	AD mean score	AD median score	Non-AD mean score	Non-AD median score
Year				
2011	60.17	66.0	64.62	67.0
2012	63.47	66.0	64.71	67.0
2013	65.92	68.0	64.78	67.0
2014	66.93	70.0	64.86	67.0
2015	65.85	68.0	65.85	68.0
2016	65.57	68.0	66.39	68.0
2017	65.44	67.0	66.88	69.0
2018	65.71	67.0	67.55	70.0
2019	65.19	68.0	66.81	70.0

Figure 14: Descriptive statistics for each year.

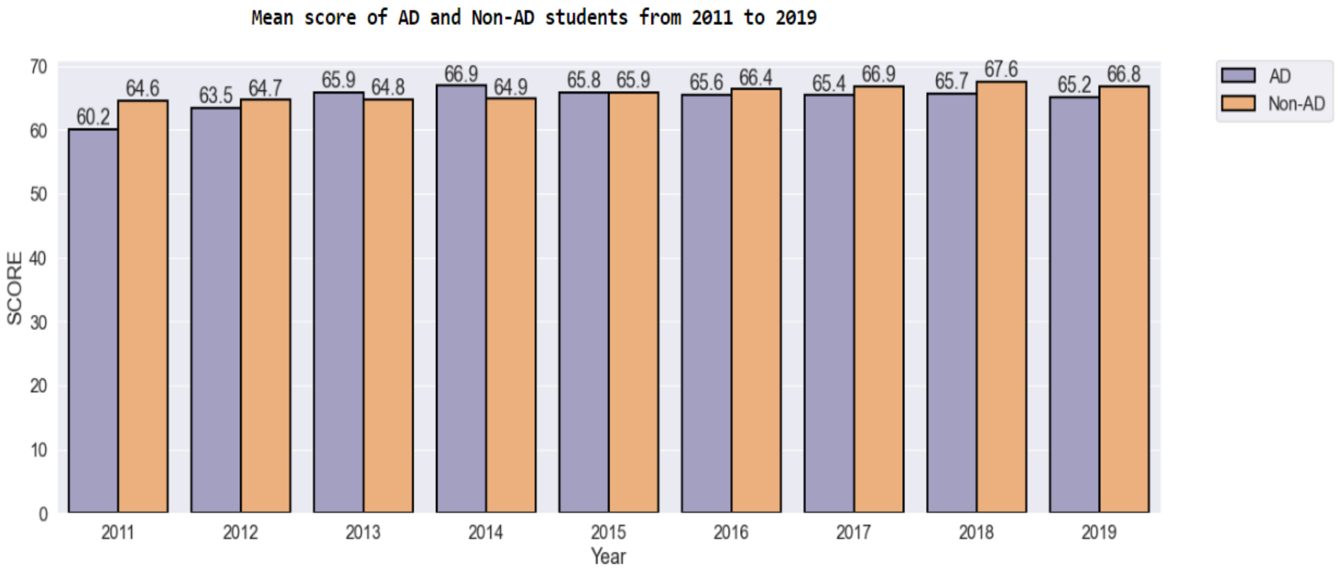
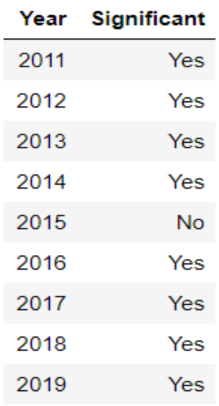


Figure 15: Mean score visualisation.

The results obtained from this technique shows that in the year 2015, AD and Non-AD students' mean scores were similar and AD students had higher mean score than Non-AD students in 2013 and 2014.

4.2.2. Hypothesis testing

Hypothesis test is required to prove the findings of descriptive statistics. In figure 6, we can see that the score distribution of AD and Non-AD do not have a normal distribution. Generally, sample taken from a population have the same distribution thus, the samples for this analysis do not have normal distribution. Therefore, parametric tests cannot be performed for hypothesis testing. Thus, non-parametric test Kruskal-Wallis was used for this purpose. Hypothesis test for statistically significant difference needs to be performed for each year in the dataset. The null hypothesis states that there is no statistically significant difference in the scores of AD and Non-AD students and the alternate hypothesis states that there is statistically significant difference in the scores of AD and Non-AD students.



Year	Significant
2011	Yes
2012	Yes
2013	Yes
2014	Yes
2015	No
2016	Yes
2017	Yes
2018	Yes
2019	Yes

Figure 16: Hypothesis test result.

Figure 16 shows the year and whether the difference of scores of AD and Non-AD students for that year is statistically significant or not. The result from hypothesis testing also shows that there was no statistically significant difference in the scores of AD and Non-AD students in the year 2015.

4.2.3. Result

Both the techniques applied in this analysis supported the finding of each other. Comparing the performance of AD and Non-AD students in each year is inconclusive because AD students had higher mean scores than Non-AD students in some years and Non-AD students had higher mean scores than AD students in some yeas.

4.3. Performance of students commencing in each year

The idea of this analysis is to compare the performance of students from the same batch rather than comparing the performance of all the students across all batches. Students who commence in a program in the same term of a year are considered batchmates. The count of AD and Non-AD students commencing in each year is shown in figure 17. The difference in sample size is still very large but the difference decreased with progression of year.

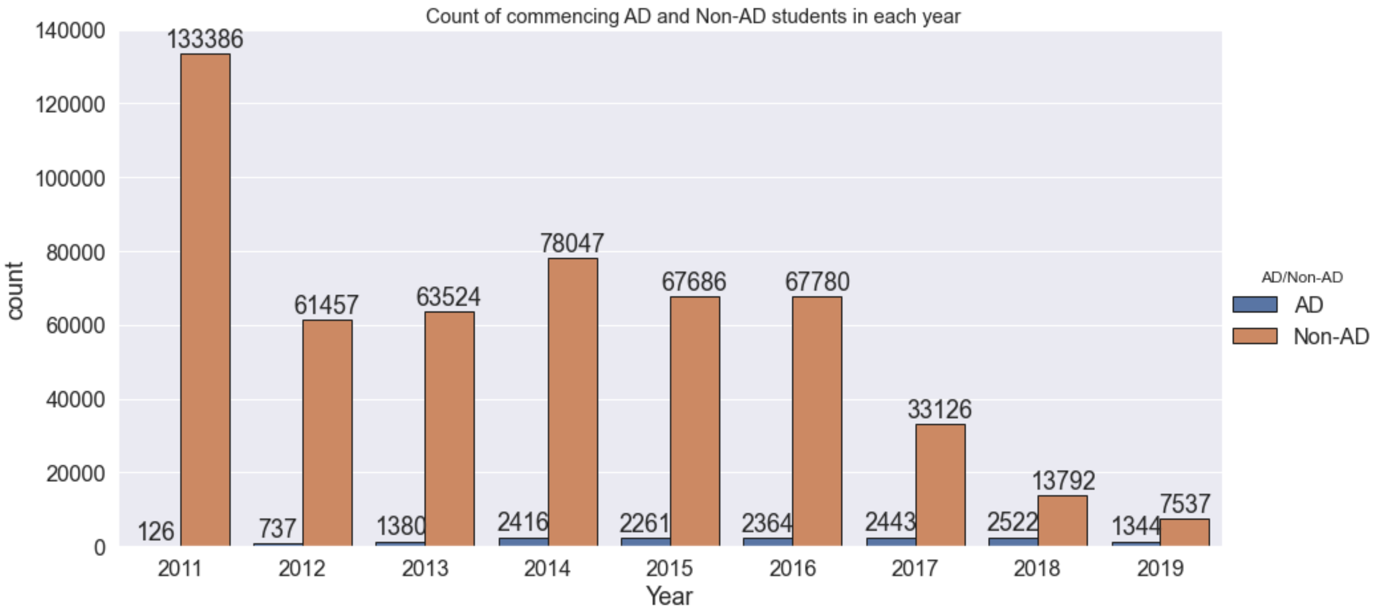
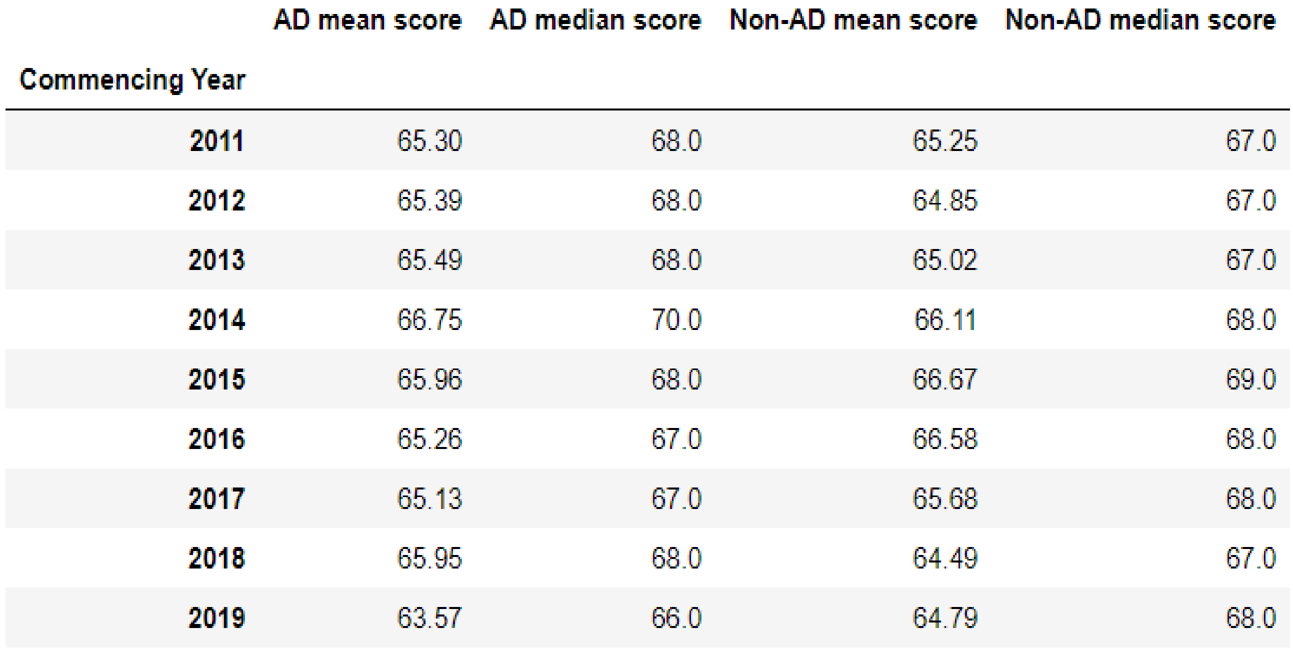


Figure 17: Count of commencing students in each year.

Performance analysis belonging to the same batch was done using descriptive statistics, visualisations and hypothesis test for the difference in the scores of AD and Non-AD students of the same batch. Bootstrapping was not performed because of the time complexity.

4.3.1. Descriptive statistics and visualisation

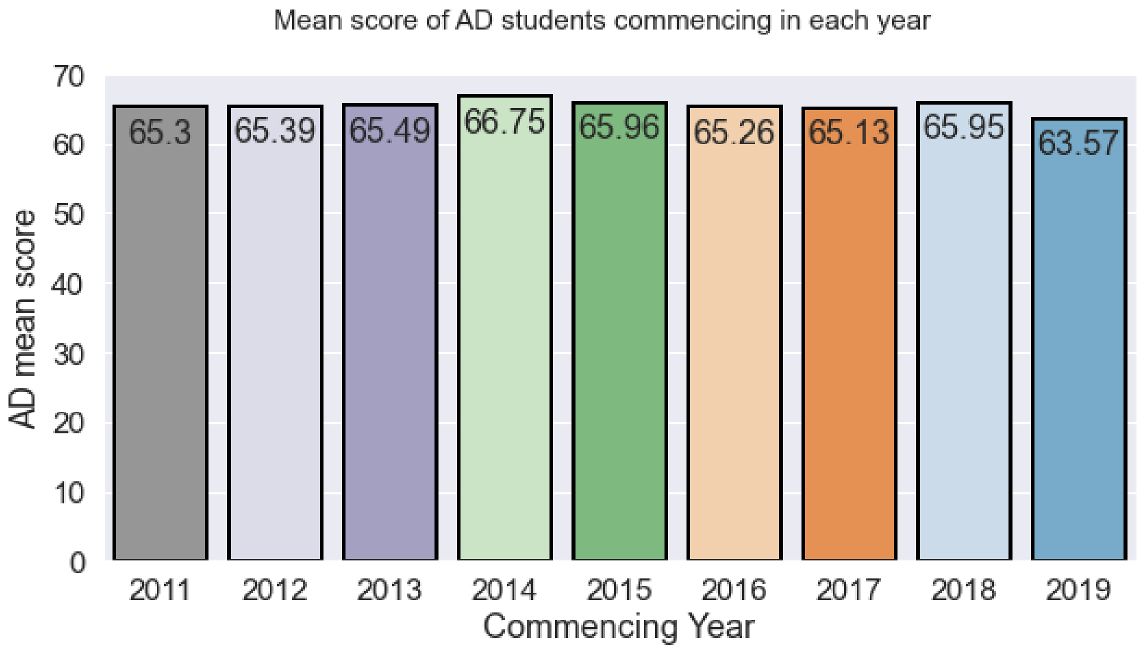
To find the descriptive statistics of students belonging to the same batch, logical implementation and iterative loops were applied to the dataset to filter students belonging to the same batch. For each batch, the descriptive statistics were found using the mean () and median () functions in python. Using the statistics found, a new dataset shown in figure 18 was created.



	AD mean score	AD median score	Non-AD mean score	Non-AD median score
Commencing Year				
2011	65.30	68.0	65.25	67.0
2012	65.39	68.0	64.85	67.0
2013	65.49	68.0	65.02	67.0
2014	66.75	70.0	66.11	68.0
2015	65.96	68.0	66.67	69.0
2016	65.26	67.0	66.58	68.0
2017	65.13	67.0	65.68	68.0
2018	65.95	68.0	64.49	67.0
2019	63.57	66.0	64.79	68.0

Figure 18: Descriptive statistics of students commencing in each year.

Comparing the mean scores using bar plots is easier to understand which is why bar plots were used to visualise the descriptive statistics of students belonging to the same batch.



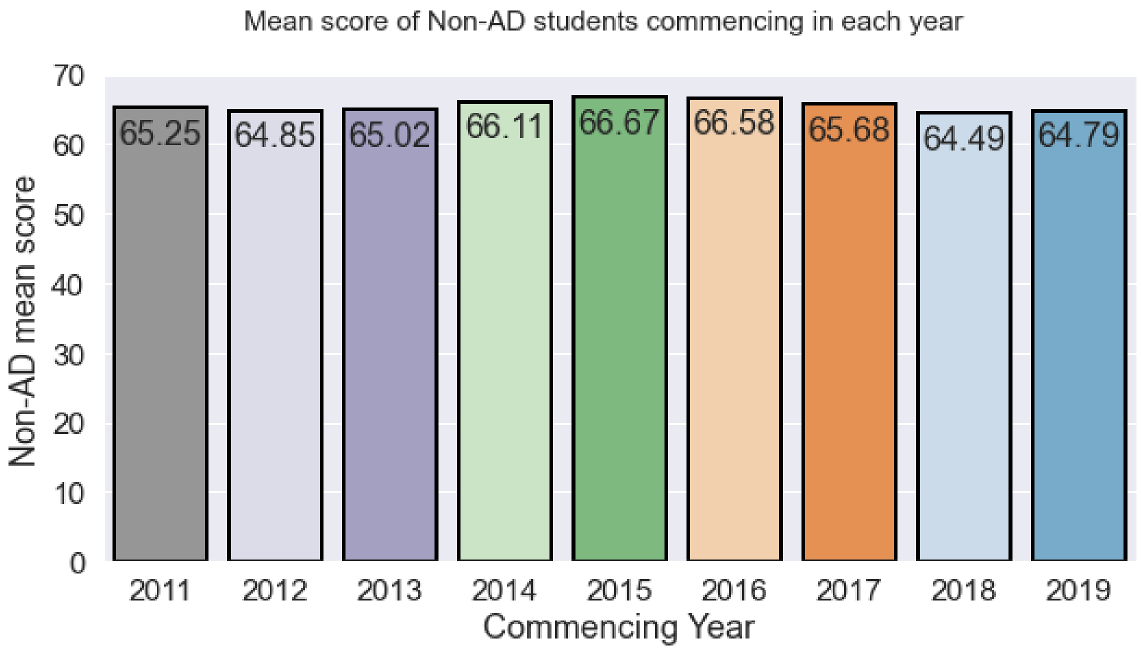
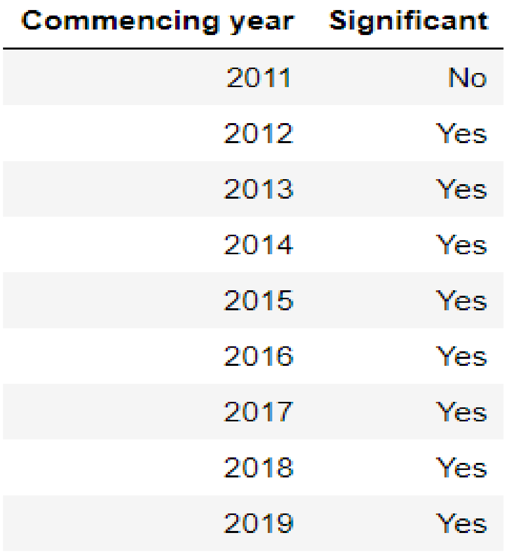


Figure 19: Mean score of students commencing in each year.

The results obtained shows that there was no difference in the mean score of AD and Non-AD students in the year 2011. AD students had higher mean score than Non-AD students in 2012, 2013, 2014 and 2018 whereas Non-AD students had higher mean score than AD students in 2015, 2016, 2017 and 2019.

4.3.2. Hypothesis testing

Hypothesis test is required to determine whether the difference in scores of AD and Non-AD students of students belonging to the same is statistically significant. The distribution of score of the cohort is non-normal as these are the samples from a population with non-normal distribution. Therefore, non-parametric test Kruskal-Wallis was performed for all the batch cohort.



Commencing year	Significant
2011	No
2012	Yes
2013	Yes
2014	Yes
2015	Yes
2016	Yes
2017	Yes
2018	Yes
2019	Yes

Figure 20: Hypothesis test result.

Kruskal-Wallis test also confirmed the findings from descriptive statistics. The difference in score of AD and Non-AD students was not statistically significant for the 2011 batch but were statistically significant for rest of the batches.

4.3.3. Result

Performance analysis of students belonging to the same batch gave inconclusive result as AD students in some batches had higher score than Non-AD students and, in some batches, Non-AD students had higher mean score than AD students. Further analysis is required to see which group of students performed better.

4.4. Performance of students in selected programs

It was observed that not all undergraduate programs had students from AD pathway therefore, including those programs for analysis is not suitable. It was also seen that there were many undergraduate programs with very less number of AD students. Therefore, it is logical to only include those undergraduate programs that had the most number of AD students. Count plot shown in figure 21 shows the count of unique AD student in each undergraduate programs. There were only 10 undergraduate programs with count of AD students close to 100 which is why for this analysis, the dataset was filtered to contain information of these 10 programs. The analysis was done on the overall (considering records from 2011-2019) score of the students.

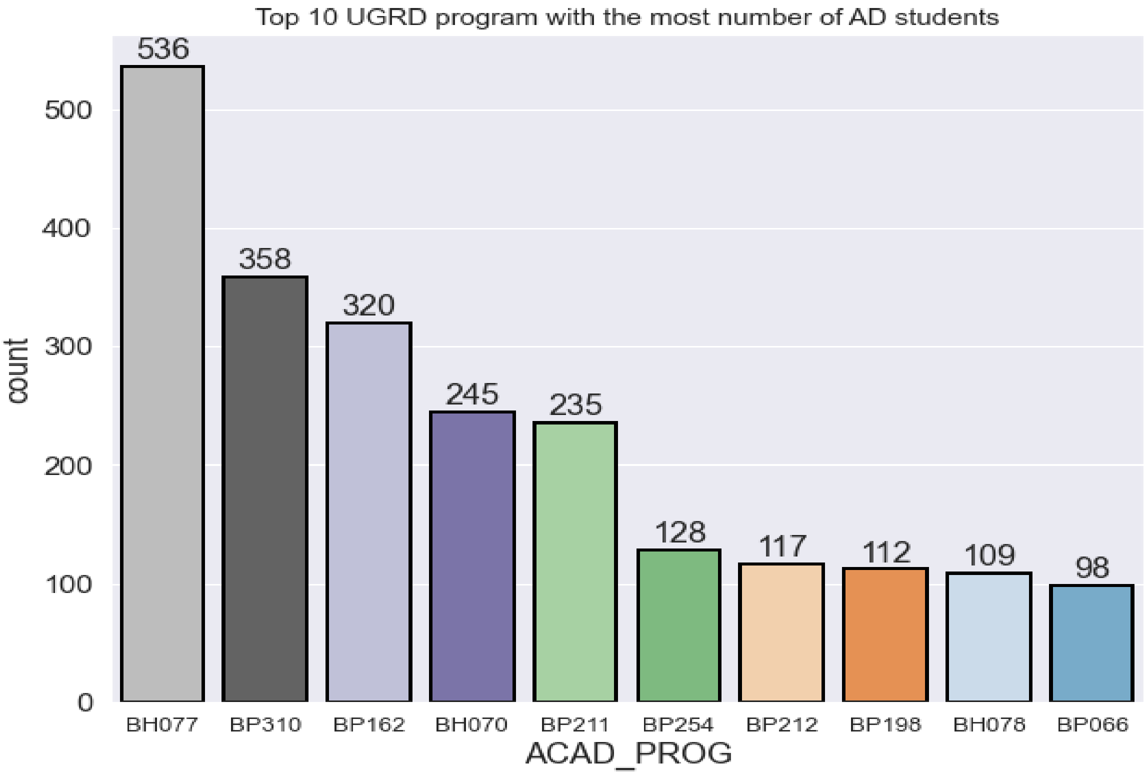


Figure 21: Count of unique AD students in undergraduate programs.

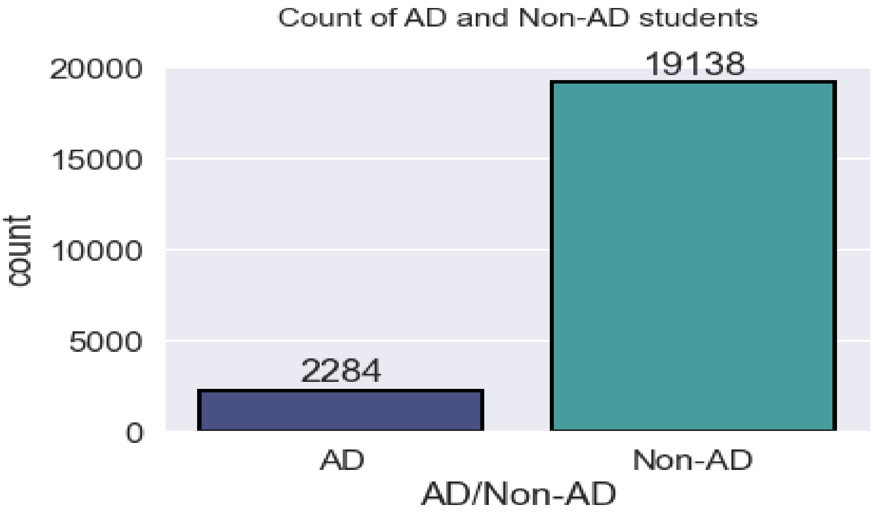
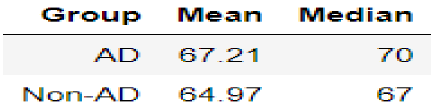


Figure 22: Count of unique AD and Non-AD students in the selected programs.

From the count plot in figure 22, it is evident that the difference in the sample size of AD and Non-AD students in these selected programs is very small as compared to sample differences observed in previous analyses and it can be assumed that the analysis would be better because of this. Performance analysis was done using the same techniques applied in the previous analyses – descriptive statistics with visualisation and hypothesis testing to support the findings.

4.4.1. Descriptive statistics and visualisation

The mean and median were determined using mean () and median () python function and bar plots along with histogram were used to visualise the mean score.



Group	Mean	Median
AD	67.21	70
Non-AD	64.97	67

Figure 23: Descriptive statistics.

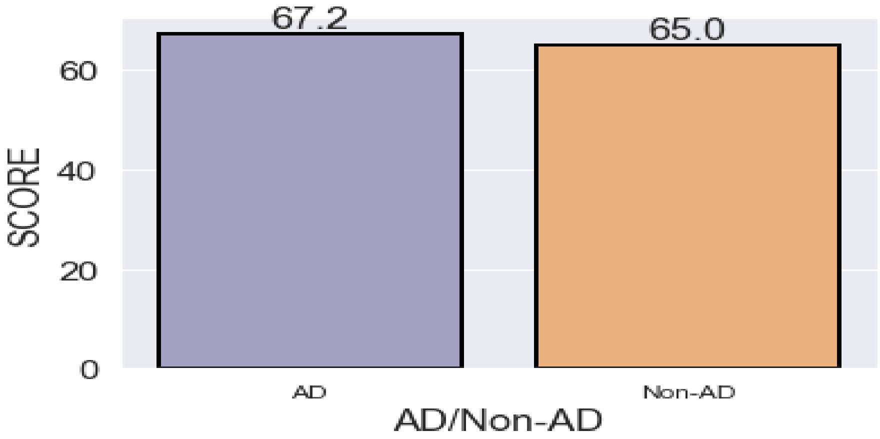


Figure 24: Mean score visualisation of AD and Non-AD students.

Result of descriptive statistics shows that AD students have higher mean score as well as median score than Non-AD students in the selected programs. The difference seems to be significant but it can only be confirmed after performing hypothesis test.

4.4.2. Hypothesis test

The distribution of scores of AD and Non-AD students in the selected undergraduate programs seems to be non-normal as figure 25 shows that distribution is left skewed.

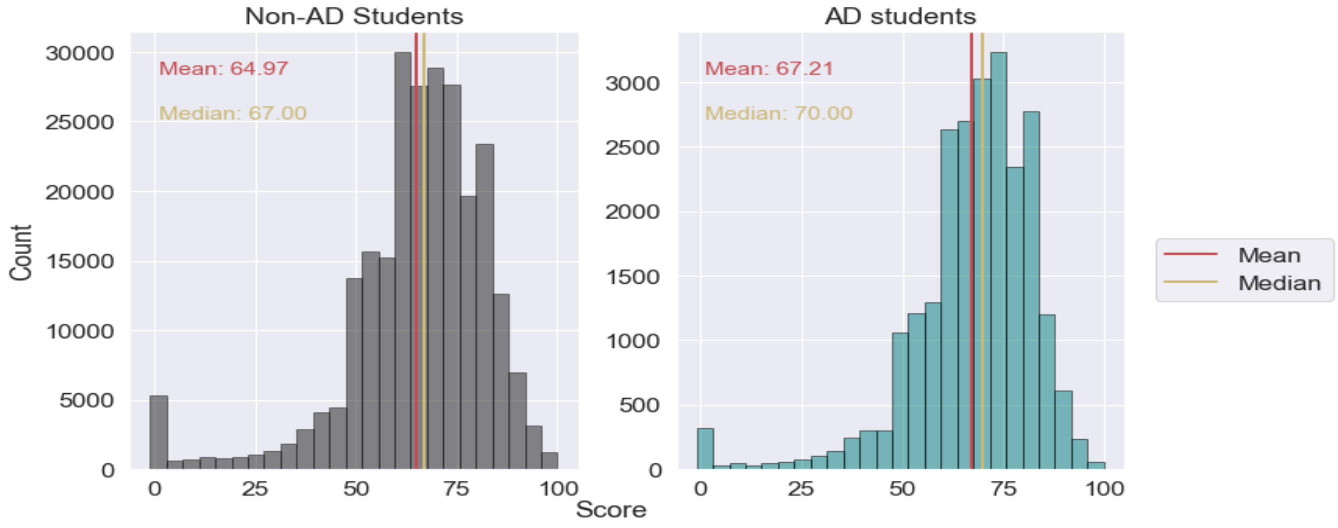
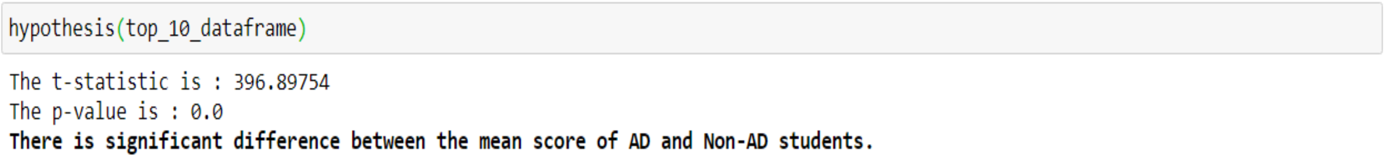


Figure 25: Score distribution of AD and Non-AD students.

Due to the non-normal distribution, non-parametric Kruskal-Wallis test was performed. The null hypothesis for the test is that there is no significant difference in the score of AD and Non-AD students and the alternate hypothesis is that there is statistically significant difference in the score of AD and Non-AD students.



```
hypothesis(top_10_dataframe)

The t-statistic is : 396.89754
The p-value is : 0.0
There is significant difference between the mean score of AD and Non-AD students.
```

Figure 26: Kruskal-Wallis test result.

The output of Kruskal-Wallis test shown in figure 26 supports the findings from the descriptive statistics analysis. The result proves that the difference in the score of AD and Non-AD students seen is statistically significant.

4.4.3. Result

The result obtained from analysing the performance of students in these selected programs shows that AD students performed better than Non-AD students. The mean score of AD students 67.2 and the mean score of Non-AD students was 65. The median score of AD students was also higher than the median score of Non-AD students therefore, we can assume that most of the AD students belong to the higher spectrum of score as compared to Non-AD students.

4.5. Analysis

The main aim of this project was to compare the performance of AD and Non-AD students in undergraduate programs. Four analyses were performed to determine which cohort of students performed better. Three of the analyses performed yielded inconclusive result but analysis done on selected undergraduate programs shows that students belonging to AD cohort perform better than students belonging to Non-AD cohort. One possible for this could be that AD students have prior knowledge of the undergraduate program they enrol in as associate degree is about specialising in a field.

5. Visualisation proposal

Another aim of this project is to propose a visualisation that could be used to show the performance of students taking the associate degree pathway to undergraduate degree. The visualisation needs to have a proper flow of records from the time the student commences in a program until the end of the program. A network plot was proposed for this task it can be clear and more informative about how a student progresses in a program.

The visualisation shown in figure 27 was proposed as part of this project. Considering a batch of students commencing in an undergraduate program in term 1 of the year 2015, the visualisation shows the progression of students from the start of the program until the end of the program. The primary nodes will contain the information of term and year. In the scenario considered above, the primary node will start from term 1 of 2015 and will end at term 2 of 2019 which is the duration of the undergraduate program. The primary nodes are connected to each other to represent the progression of students in the program. The thickness of lines connecting the nodes represent the number of students thus, thicker the line, more the number of students and vice versa.

Each primary node is connected to nodes containing the information of students coming from different AD programs (Non-AD node for students directly enrolling in the undergraduate program). These secondary nodes are colour coded according to fail rate of the students. Rate is considered rather than the count of fail because the number of students from each associate degree may vary. If the fail-rate of an associate degree program is less than 5%, the node will be green in colour. Programs with fail-rate greater than 5% but less than 10% will be colour coded to yellow and for programs with fail rate more than 10%, the node will be colour coded to red.

After the end of a term, two rounded corners rectangle nodes are connected to the primary node progression arrow. One of these nodes will store the information of students who withdraw from the program and the other node will store the information of students who commence to the next term of the program.

At the end of the program duration, the secondary node is connected to tertiary rectangle nodes which will store information of students by grouping them according to their overall grades. The tertiary rectangle nodes will also store the information of students who failed or withdrew from the program over the duration of the program.

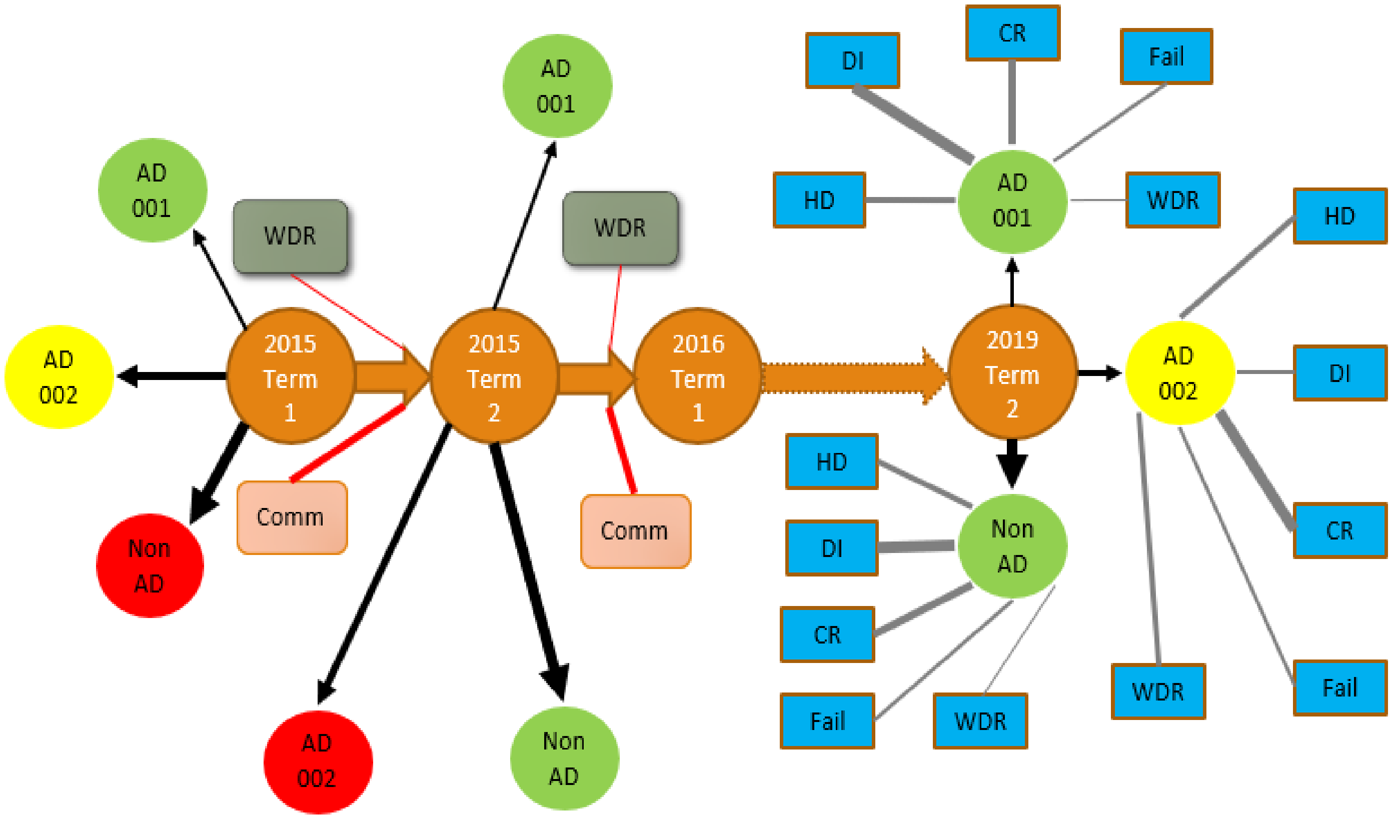


Figure 27: Network plot showing the progression of students in an undergraduate program.

6. Insights

Finding meaningful insights was necessary in this project because further analysis could be done on those insights to make the student experience better. Access to demographic data also helped in finding insights that could be useful in the future.

6.1. Performance of students from each AD program in undergraduate program

To find this insight, the overall mean score achieved by students belonging to each AD program were visualised using bar plots. From the visualisation in figure 28, we can see that students from only 3 associate degree programs (AD022, AD013 and AD016) manged to achieve a mean score above 70 or DI grades. Further analysis can be done and acted upon to improve the performance of students belonging to AD programs with low mean scores.

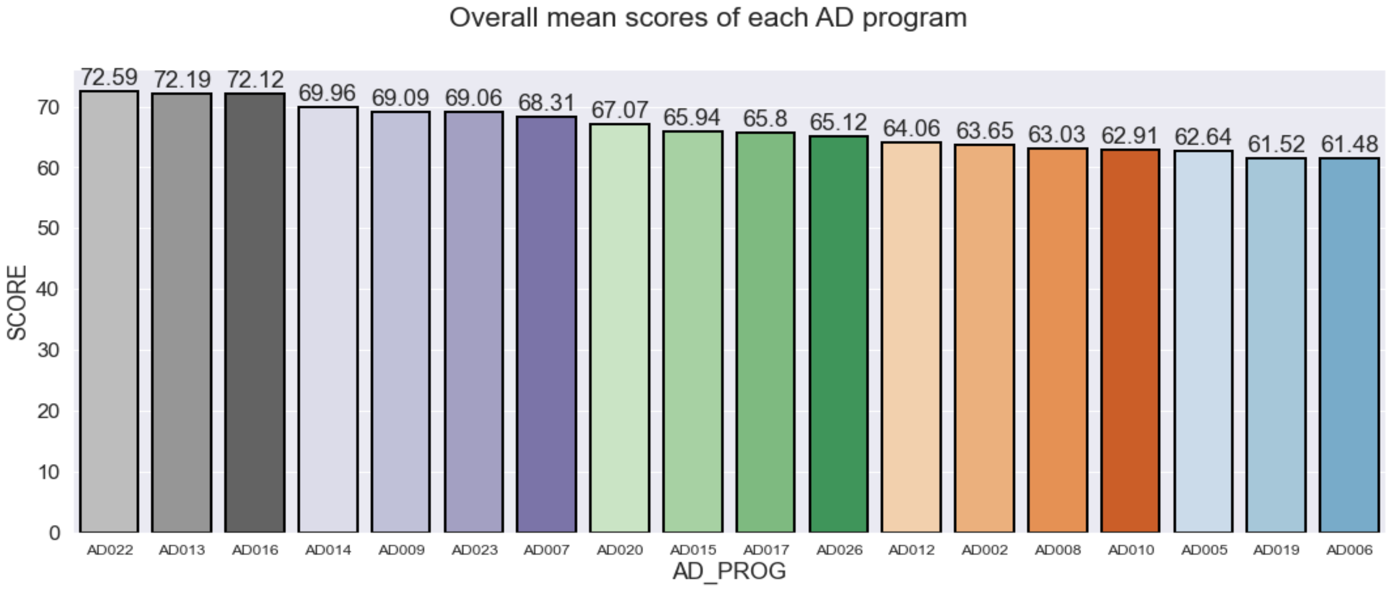


Figure 28: Overall mean score visualisation of students from each AD program.

6.2. AD program to undergraduate program transition

The information given by this insight is the undergraduate program a student from a specific AD program may move to if he/she decides to take the associate degree pathway to undergraduate degree. Sankey diagram was used to visualise this insight.

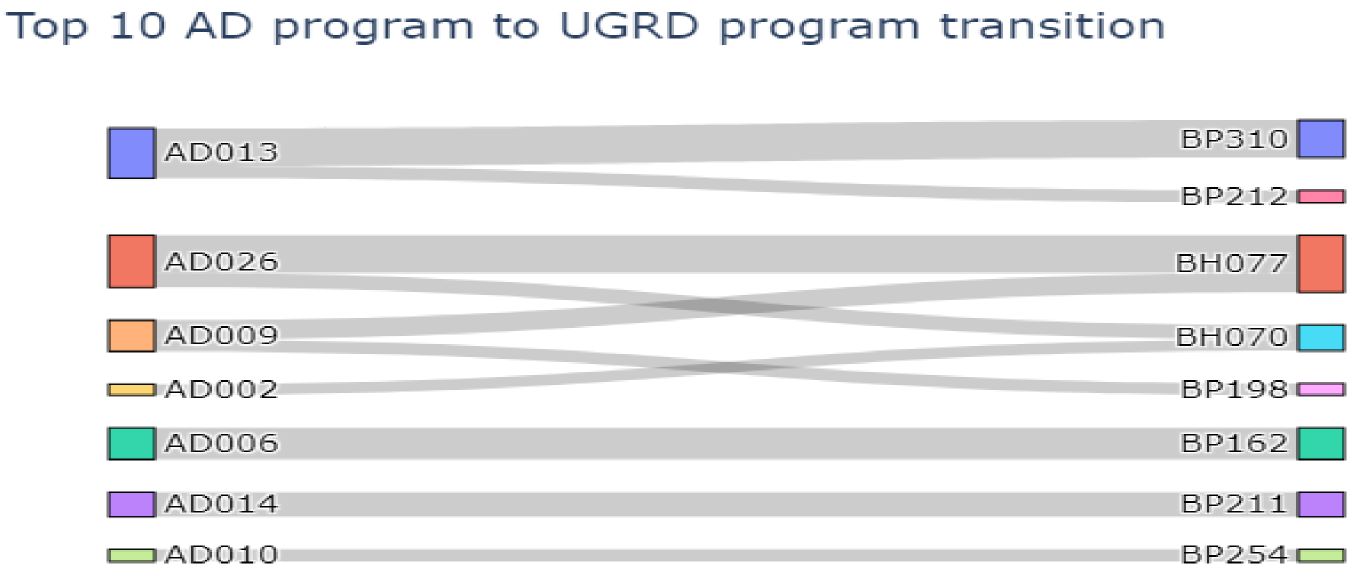


Figure 29: Top 10 AD program to undergraduate program pathway.

AD Program	Undergraduate Program	Count
AD013	BP310	351
AD026	BH77	350
AD006	BP162	295
AD014	BP211	231
AD009	BH077	182
AD026	BH070	137
AD013	BP212	117
AD010	BP254	114
AD009	BP198	111
AD002	BH070	104

Table 2: AD program to undergraduate program pathway count.

Table 2 contains the count of instances when a student from a specific AD program took the pathway to enrol in a specific undergraduate program. From the visualisation and the table above, we can see that generally, students from AD program AD013 take the pathway to enrol in undergraduate program BP310.

Analysis can be done on this insight by doing performance analysis on the scores achieved by students from a specific AD program in an undergraduate program. By doing this, recommendations of which undergraduate program to enrol in can be made to a student deciding to take the pathway. Also, if a particular AD program performs poorly in an undergraduate program, the university can make changes accordingly.

6.3. Insights from demographic data

Three demographic attributes – gender, socio economic status and international/domestic were merged in the dataset. Many insights can be gathered using these attributes.

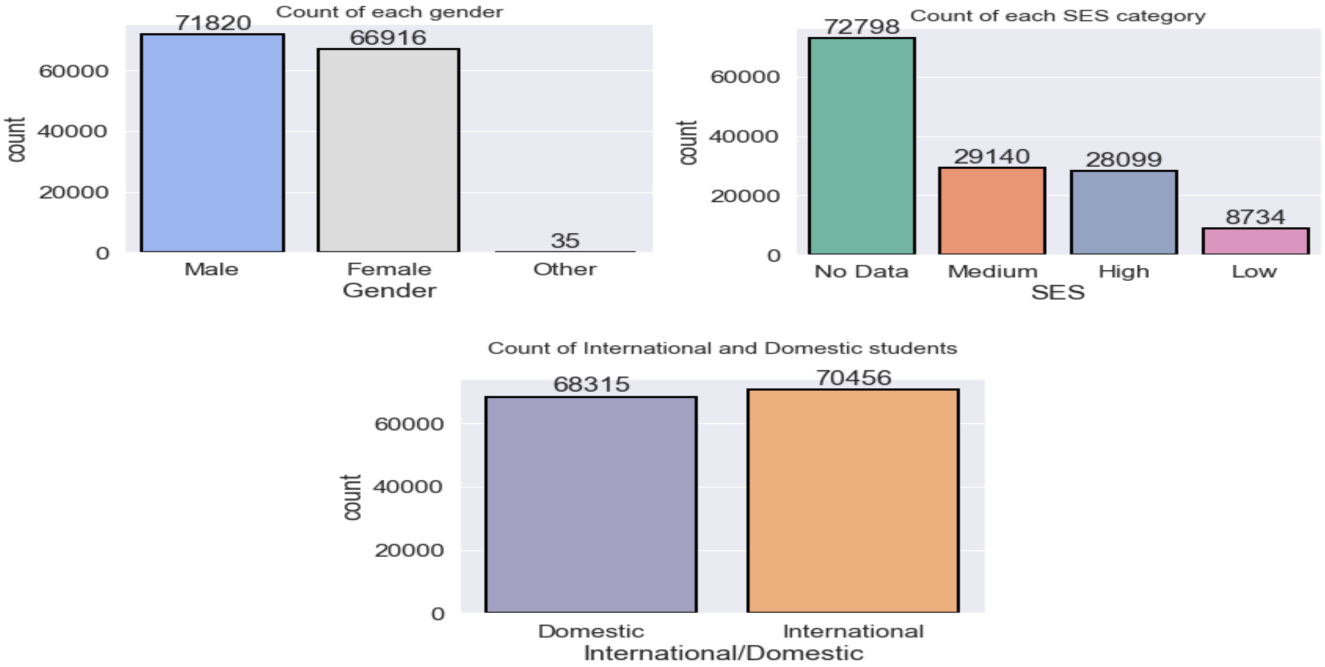
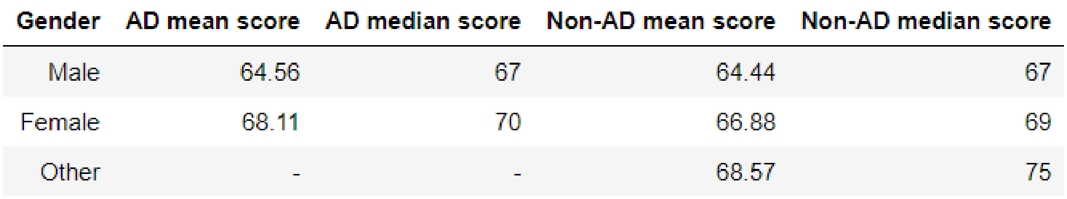


Figure 30: Count of students in each demographic groups.

The performance of each demographic was compared using descriptive statistics and visualisations. Bar plots and grouped bar plots were used to compare the performance. The comparison was made on the scores achieved by students belonging to each demographic groups.



Gender	AD mean score	AD median score	Non-AD mean score	Non-AD median score
Male	64.56	67	64.44	67
Female	68.11	70	66.88	69
Other	-	-	68.57	75

Figure 31: Descriptive statistics of gender.

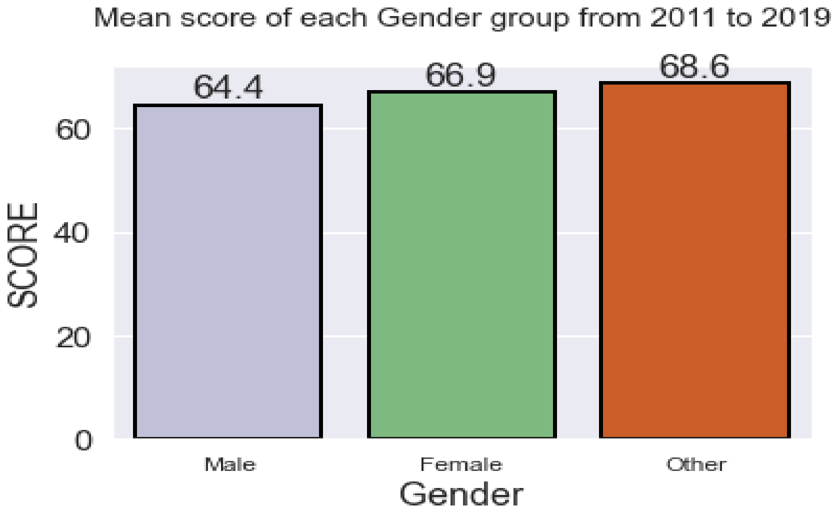


Figure 32: Overall mean score of each gender category.

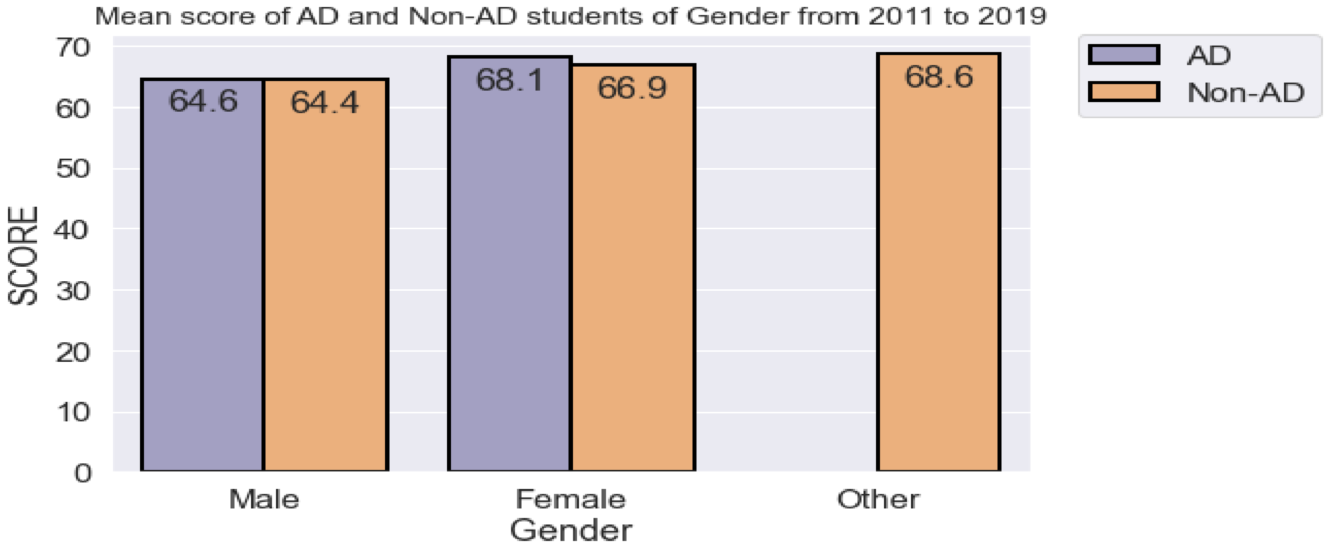
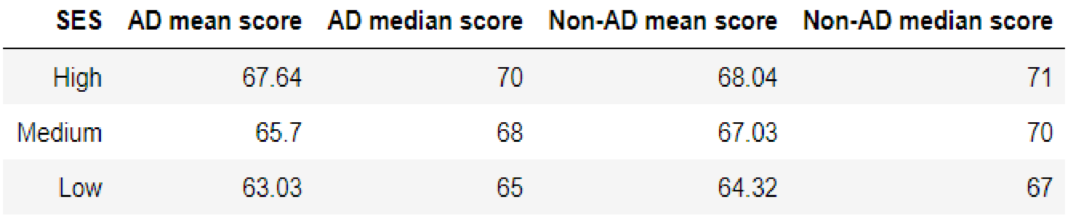


Figure 33: Mean score of AD and Non-AD students belonging to each gender category.

From the visualisation above, we can see that female students have higher mean score than male students. Female AD and Non-AD students also had higher mean score than their male counterpart. The median score of female students is also higher in both AD and Non-AD cohort due to which we can assume that female students achieve higher score than male students. No students with "Other" as gender was present in AD program.



SES	AD mean score	AD median score	Non-AD mean score	Non-AD median score
High	67.64	70	68.04	71
Medium	65.7	68	67.03	70
Low	63.03	65	64.32	67

Figure 34: Descriptive statistics of socio-economic status.

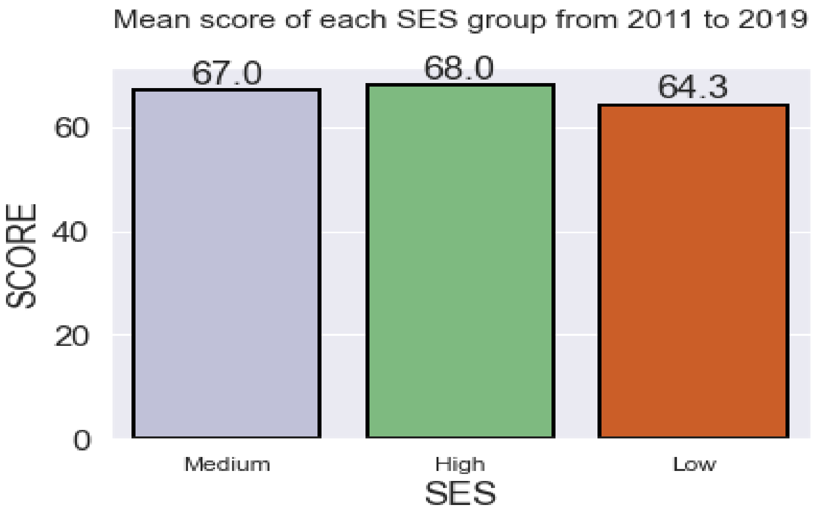


Figure 35: Overall mean score of socio-economic status category.

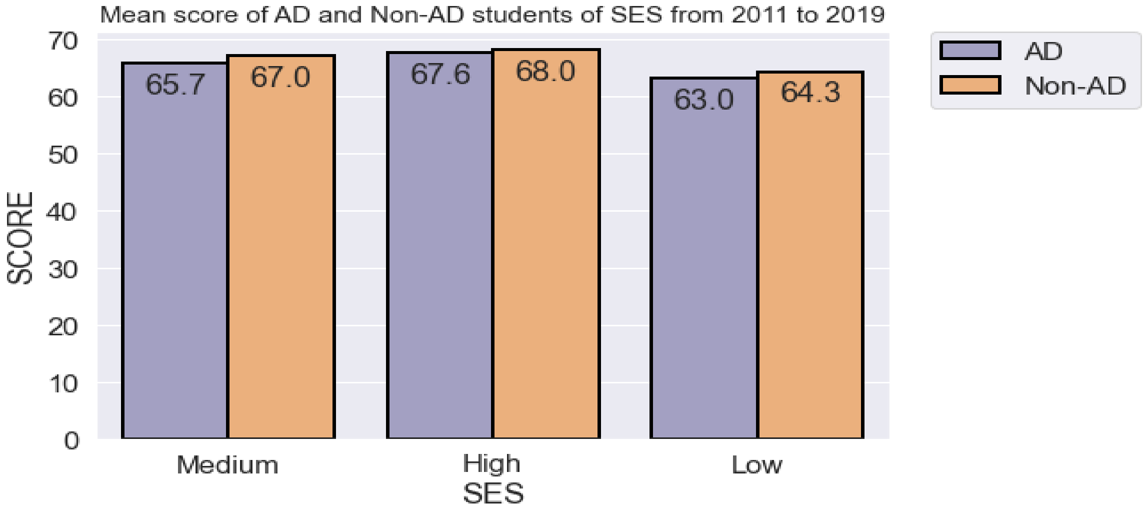
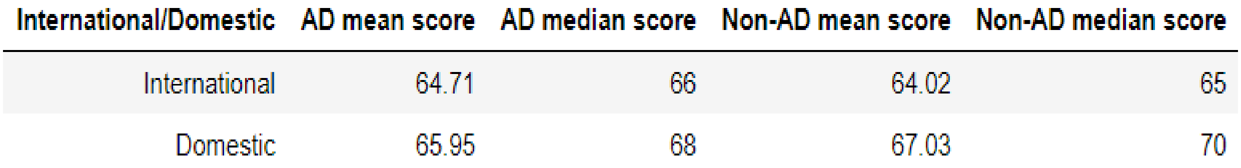


Figure 36: Mean score of AD and Non-AD students belonging to each socio-economic status category.

From the above visualisations, we can say that students who have high socio-economic status have higher mean as well as median score than students who have from medium and low socio-economic status. AD and Non-AD students with high socio-economic status also have higher mean score than the students who have low and medium socio-economic status. Students with low socio-economic status have the lowest mean and median score among the three categories. Further analysis could be done and appropriate actions can be taken to improve the performance of students belonging to each socio-economic status category.



International/Domestic	AD mean score	AD median score	Non-AD mean score	Non-AD median score
International	64.71	66	64.02	65
Domestic	65.95	68	67.03	70

Figure 36: Descriptive statistics of international/domestic.

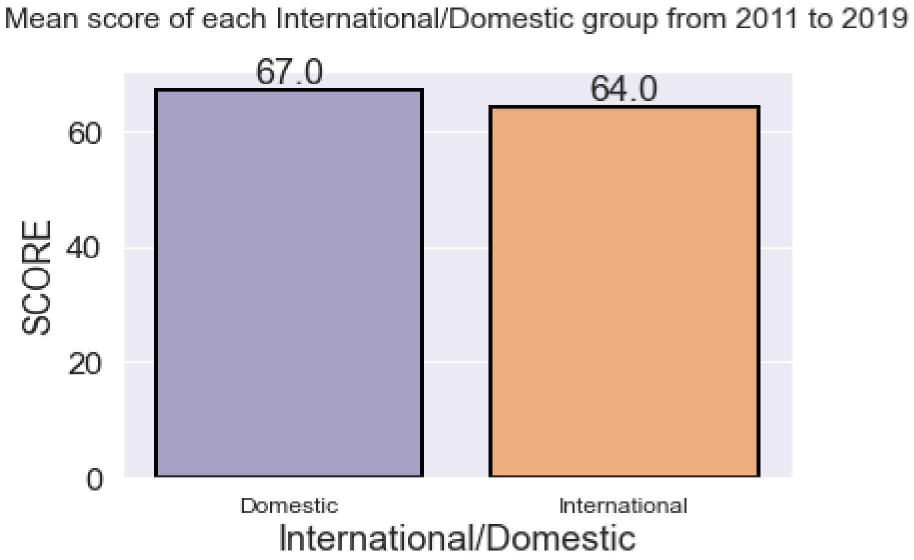


Figure 37: Overall mean score of international and domestic students.

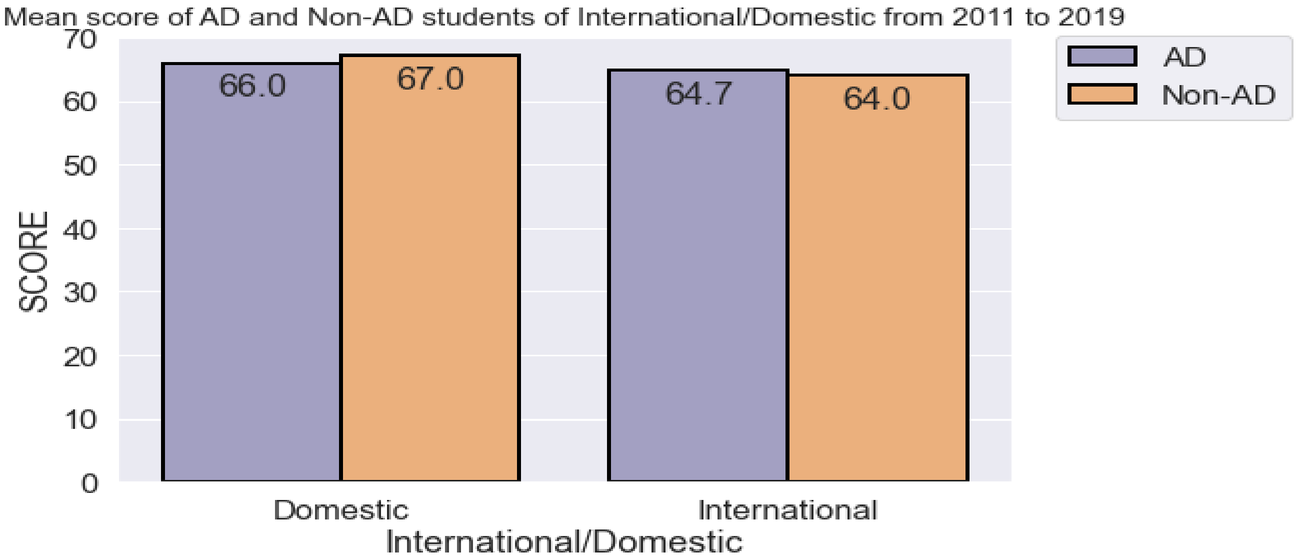


Figure 38: Mean score of international and domestic students in AD and Non-AD cohort.

From the visualisations above, we can see that domestic students have higher mean as well as median score than international students. Domestic students from AD and Non-AD groups also perform better than international students.

7. Conclusion

The project deals with the process of data preparation and exploration which are two of the most important aspects of data science. The objective of the project was to compare the performance of students taking the associate degree pathway to undergraduate degree and students who enrol directly in an undergraduate program. Descriptive statistics along with visualisations and hypothesis tests were used to compare the performance of students. Data preparation consumed the most time in this project because of the very complex solutions needed to prepare the data before analysis could be performed.

The objective of the project was achieved by performing four different analyses. Performance analysis done on top 10 undergraduate programs with the most number of students from AD pathway provided us with the evidence that students from associate degree pathway performed slightly better than student who directly enrol to the program. Although the analysis gave significant result, few things like considering students from the same batch could be incorporated. Another aim was to propose a visualisation to show the progression of students coming from associate degree pathway. The proposal could be worked on in the future by implementing it practically. Networkx module available in python could be used to design the proposed visualisation.

8. Appendix

8.1. Project team

This project was done individually with the help of mentors from the company. Work on this project was done on python programming language. Jupyter notebook was used to develop the solutions of this project. Functions available in Pandas and NumPy libraries were used for the purpose of data preparation. Visualisation were achieved using matplotlib, seaborn and plotly libraries.

Two meetings were held every week (Wednesday and Friday) with the project mentors to discuss the progress of the project. The meetings were held online on Microsoft Teams. With the guidance of my mentors who were very helpful, I could achieve the aim of this project. They were there whenever I had any doubts regarding the project.

8.2. Self-reflection

Many analyses and visualisations were explored which were not included in the project. Exploration of different techniques was necessary so that we could decide which visualisations were appropriate for this project. Although all the visualisation gave some story, only those which could be easily understood were chosen to include in this project.

The codes in data preparation had to be optimised due to which it took a lot of time. In the early version of the codes, many iterative loops were used to achieve the desired dataset. Using iterative loops takes a lot of time especially if the number of records is in thousands which was the case in this project. Optimising the codes taught me new techniques that could be use in data preparation process.

I also tried to normalise the score distribution using Box-Cox transformation so that I could perform parametric hypothesis test like Two-sample T Test. Performing Shapiro-Wilk test on the normalised data gave inaccurate results which upon research I learned that these tests have limitation of sample size (maximum of 3000 samples otherwise the results would be inaccurate). Since I could not prove the normality of the score attribute using this test, I decided to only perform non-parametric test.

While researching on visualisation techniques, I came across new libraries and techniques that could be very helpful in the future. Time management was a major issue for me which could be improved next time. Overall, working on this project was very exciting.