



# **TDS2101**

## **Project Part B Report**

### **Dataset 1 - Education: Literacy rates**

**Group No.: <6>**

**<Raja Muhammad Darwisy>  
<Imran bin Zulkiflee>**

**<1191100792>  
<1191100756>**

# Table of Contents

1.	<b>Introduction</b>	3
1.1	Problem Statement	3
1.2	Motivation	3
2.	<b>Questions</b>	4
3.	<b>Data Collection</b>	7
4.	<b>Data Preprocessing</b>	8
5.	<b>Exploratory Data Analysis</b>	10
5.1	<b>Descriptive &amp; Causal Question</b>	10
5.1.1	Data Visualisation	10
5.2	<b>Mechanistic Question</b>	11
5.2.1	Data Visualisation	11
5.3	<b>Exploratory Question</b>	12
5.3.1	Data visualisation	14
5.4	<b>Inferential Question</b>	15
5.4.1	Data Visualisation	16
5.5	<b>Predictive Question</b>	17
5.5.1	Data Modelling	17
6.	<b>Challenges &amp; Conclusion</b>	20
7.	<b>References</b>	21

# 1. Introduction

## 1.1 Problem Statement

Poverty and poor health prevent youths and adults from getting education which affects the literacy rates, so a good starting point is comparing amongst the least developed countries and developed countries.

## 1.2 Motivation

In this modern day and age, literacy rates in developed and undeveloped countries are becoming a huge problem that affects poverty and poor health of youths and adults. Based on an article by the world population review, it stated that the global literacy rate among youth and adults is high. The literacy rate for all males and females that are at least 15 years old is 86.3%. Males aged 15 and over have a literacy rate of 90%, while females lag only slightly behind at 82.7%. However, massive country-to-country differences exist.

Besides that, poverty and poor health will also affect the literacy rate in developed and least developed countries. According to the website (world population review), it showed that poverty and illiteracy tend to go hand-in-hand, these problems are noticeable mostly in least developed countries that are shown in our data sets. Therefore, our team is curious about the relationship of poverty that affects the literacy rates amongst these countries.

Because of these problems, our team has the motive to find out the different literacy rates in developed and undeveloped countries, the difference between youth and adult literacy rates and the impact (whether there is any) of poverty and poor health towards literacy rates in these countries. The data set that we are going to use is mainly the amount of literacy rate amongst different countries. Our team is eager to find out whether living in poverty (least developed countries) will affect the literacy rate. By solving this problem, it will help the respective countries to take action on increasing the literacy rate of youth and adults and have a realisation that poverty and poor health has a huge impact on literacy rate.

## 2. Questions

### Type of question: Descriptive Question

A descriptive question attempts to summarise a feature of a collection of data. For example, in a set of data obtained from a group of people, identifying the proportion of males, the mean number of servings of fresh fruits and vegetables per day, or the incidence of viral infections are all examples. Because the outcome is a fact, an attribute of the set of data you're working with, there's no way to interpret it.

**“ What is the average number of youth and adults that have low literacy rate in a developed country and undeveloped country? ”**

Refined Descriptive question:

**“ What is the average number of youth and adult literacy rates in a developed country and undeveloped country? ”**

Additional question:

**“ What is the minimum number of youth and adult literacy rates in a developed country and undeveloped country? ”**

For Descriptive analysis, we refined the question because we decided to add a question to analyse the minimum number of youth and adult literacy rate in both types of rating of the country. Firstly, we locate and separate the data between the developed countries (DC) and underdeveloped countries (LDC) in both the youth and adult dataset. Next, we get the descriptive statistics for both dataset to plot the statistic to analyse the proposed questions.

### Type of question: Causal Question

A causal question examines if changing one factor in a population will impact another factor on average. By default, the basic nature of the data gathering sometimes permits you to pose a causal question.

**“ What are the major causes of low literacy rates in industrialised countries? ”**

**“ How is a developed country affecting high literacy rates in a community in the country? ”**

Refined Causal Question:

**“ Can the wealth status of a country influence literacy rates? ”**

We refined the question to a researchable question as for the previous questions we cannot make an analysis due to lack of data in this dataset. So for causal analysis we decided to make analysis together with descriptive analysis as from descriptive we can answer the refined causal question based on descriptive analysis.

Type of question: Mechanistic Question

A mechanistic question is one that focuses on how one element influences the outcome. (For example, how a healthy diet can reduce the amount of viral infections). A mechanistic question might be one that asks how a diet rich in fresh fruits and vegetables reduces the prevalence of viral diseases.

**“ How does sex gender and health condition affect to each other to get the different value of the literacy rates amongst the sex gender? ”**

Refined Mechanistic Question:

**“ Does the Developed and Undeveloped country has influences to get the different value of the literacy rates amongst the sex gender of the youth? ”**

As for the refined question, we just want to make it a clearer question synonym to the data that we have in both datasets. The reason only using the youth dataset as youth values is more consistent than the value of the adult. So we can make a clearer analysis about the influence of type of rating country amongst the sex gender. The two pairs of variables used in these questions are 'Rating of Country', 'Male' & 'Female'. The 'Rating of Country' are categorical variables while the 'Male' & 'Female' are numerical variables that show the percentage of literacy rate respectively. Hence, we plan to come up with a side by side bar chart as it is easier to display the influence by comparing the literacy rates amongst sex gender between the type rating of the country,

Type of question: Exploratory Question

An exploratory question examines the data to determine if any patterns, trends, or correlations between variables may be discovered. These types of analyses are often known as "hypothesis generating" analyses because it looks at patterns to generate hypotheses rather than testing them.

**“ What is the relationship of literacy rate amongst youth and adults in least developed and developed countries? ”**

Refined Exploratory Question:

**“ What is the relationship of literacy rate amongst male and female youth and male and female adults in Developed and Undeveloped countries? ”**

The reason why we refine the question is because when we make an early proposed question for an exploratory question, we only see the uncertainty of the data because the data has a lot of missing values and is totally messed up. So after the cleaning process we can see more clearly about the datasets and we refined it based on the values and objects to make this exploratory analysis. As for visualisation, we decided to use heatmap from package seaborn as it is the most common way to see the relationship between variables.

Type of question: Inferential Question

An inferential question would be a restatement of this proposed hypothesis as a question and would be answered by analysing a different set of data. These types of analyses are often known as "hypothesis generating" analyses because it looks at patterns to generate hypotheses rather than testing them.

**“ Hypothesis : The number of youth and adults that have low literacy rate lives in least developed countries is high. ”**

**“ Is this hypothesis also true for the youth and adults that live in developed countries? ”**

For this inferential question type, we want to observe whether the number of youth and adults that have low literacy rate lives in undeveloped countries is high. By doing this hypothesis, we can also be able to confirm whether it is also true for the youth and adults that live in developed countries. Then, we plan to get the total literacy rate of the developed and undeveloped countries and compare it to the total population of the respective countries.

Type of question: Predictive Question

A predictive inquiry is one in which you inquire about the set of predictors or factors that influence a specific behaviour.

**“ What would happen to the literacy rates of a country if the quality of lives amongst society in the country improved? ”**

Refined Predictive question

**“ What would happen to the literacy rate if the population of a country is high? ”**

Our team plans to do model training for the predictive question. We will use the total population and literacy rate columns to predict the value of literacy rate and the total population by using linear regression. We will also do a scatter plot as a visual representation of the data. The regression coefficients, regression intercept and regression score can be calculated to determine whether the regression can be used for prediction of literacy rate. In this model training, our goal is to find out whether linear regression is good for predicting the literacy rate.

### 3. Data Collection

The Education - Literacy rates Dataset was sourced from data.unicef.org, the dataset consists of the literacy rates of youth from 15 to 24 years old and literacy rates of adults 24 years and up. The dataset is composed of all Countries around the world with each region and subregion including all of the Africa region with each subregion. There are also separated literacy rates of males and females, and lastly the Total population of each country.

Although there are more students in school now than ever before, far too many of them are not learning. According to data from the Multiple Indicator Cluster Surveys (MICS), many children today lack core reading and numeracy skills that prepare them for life beyond school. Furthermore, because of the digital character of modern society, information and communications technology (ICT) skills are required for full social and economic involvement, however statistics suggest that many children and teens lack these critical abilities.

Unnamed: 0		Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10
0	NaN	Country	Region	Sub-region	Least developed countries (LDC)	Africa sub-regions	Africa region	Total	Male	Female	Total population
1	NaN	Afghanistan	SA	NaN	LDC	NaN	NaN	65.42055	74.0848	56.25475	8071.334
2	NaN	Albania	ECA	EECA	NaN	NaN	NaN	99.33	99.05	99.63	458.058
3	NaN	Algeria	MENA	NaN	NaN	Northern Africa	All	97.42652	97.59406	97.25216	6070.282
4	NaN	Andorra	ECA	WE	NaN		NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...
227	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
228	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
229	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
230	NaN	Literacy data: UIS 2018	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
231	NaN	Population data: UNPD wpp 2019 for year 2018	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

232 rows × 11 columns

Figure 3.1 Literacy rates of youth 15-24 years old

0	1	2	3	4	5	6	7	8	9	10	
0	NaN	Country	Region	Sub-region	Least developed countries (LDC)	Africa sub-regions	Africa region	Total	Male	Female	Total Population
1	NaN	Afghanistan	SA	NaN	LDC	NaN	NaN	43.01972	55.47545	29.80521	20193.661
2	NaN	Albania	ECA	EECA	NaN	NaN	NaN	98.14115	98.51362	97.76112	1977.051
3	NaN	Algeria	MENA	NaN	NaN	Northern Africa	All	81.40784	87.42296	75.32297	26810.322
4	NaN	Andorra	ECA	WE	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...
223	NaN	African sub-region - North Africa	NaN	NaN	NaN	Northern Africa	NaN	—	NaN	NaN	NaN
224	NaN	African sub-region - South Africa	NaN	NaN	NaN	Southern Africa	NaN	—	NaN	NaN	NaN
225	NaN	African sub-region - West Africa	NaN	NaN	NaN	Western Africa	NaN	58.55045	NaN	NaN	NaN
226	NaN	Africa region - All	NaN	NaN	NaN	NaN	All	—	NaN	NaN	NaN
227	NaN	World	NaN	NaN	NaN	NaN	NaN	85.580656	NaN	NaN	NaN

228 rows × 11 columns

Figure 3.2 Literacy rates of adult 24 years and up

As you can see from these two figures above, there are two types of dataset which is literacy rates of adults and youth that we need to merge together before starting the data science pipeline. There are also too many missing values that we have to clean and reposition all of the column names. This cleaning part will be discussed further in the Data preprocessing part.

## 4. Data Preprocessing

Cleaning the core data, as well as the various auxiliary datasets, was necessary on a number of levels. The majority of this process takes place in the beginning of the pipeline, before any substantial analysis takes place.

For all of these csv files, we started the data preprocessing by repositioning the name of each column that has been pushed down to index 0, so that we can easily access the column data by their name. As a result, we obtain a total of 231 rows of data with 10 columns for the Youth dataset and 227 rows with 10 columns for the Adult dataset.

	Country	Region	Sub-region	Least developed countries (LDC)	Africa sub-regions	Africa region	Total	Male	Female	Total population
1	Afghanistan	SA	NaN	LDC	NaN	NaN	65.42055	74.0848	56.25475	8071.334
2	Albania	ECA	EECA	NaN	NaN	NaN	99.33	99.05	99.63	458.058
3	Algeria	MENA	NaN	NaN	Northern Africa	All	97.42652	97.59406	97.25216	6070.282
4	Andorra	ECA	WE	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	Angola	SSA	ESA	LDC	Southern Africa	All	NaN	NaN	NaN	5958.112
...	...	...	...	...	...	...	...	...	...	...
227	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
228	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
229	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
230	Literacy data: UIS 2018	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
231	Population data: UNPD wpp 2019 for year 2018	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 4.1

As mentioned in the data collection section, all necessary data has been recorded, but it has a lot of incompleteness. So we need to fill any missing values or maybe drop the rows that have all missing values for all columns. Next, we dealt with the inconsistent data as we need to reformat the column which is Least Developed Countries (LDC) and Africa sub-regions columns as we can see in the figure above. For the LDC column, We renamed it to the Rating of Country as we find it easier to fill the missing value with 'DC' string values as for developed countries in this column. Next, as for missing values in the Africa sub-regions column, we decided to fill it with 'Not Africa Region' string values so we do not need to drop this column as these two columns are important for us to do analysis and answer some analysis question that include the type of country and african region. Then, we checked and dropped duplicates which are Sub-region and Africa region columns to make sure of the uniqueness of Region and Africa sub-regions in these two dataset. As a result, this dataset has 8 columns for both dataset. Lastly, after dropping rows that have all missing values in all Total, Male, Female or Total Population, the dataset becomes 69 rows with 8 columns as shown in two figures below for Youth dataset and Adult dataset. Finally, our team decided not to merge both dataset as it is easier to make a comparison and an analysis due to both dataset having different values of total literacy rate, male literacy rate, female literacy rate and the population of a country.



	Country	Region	Rating of Country	Africa sub-regions	Total	Male	Female	Total population
1	Afghanistan	SA	LDC	Not Africa region	65.42055	74.0848	56.25475	8071.334
2	Albania	ECA	DC	Not Africa region	99.33	99.05	99.63	458.058
3	Algeria	MENA	DC	Northern Africa	97.42652	97.59406	97.25216	6070.282
8	Argentina	LAC	DC	Not Africa region	99.50552	99.24183	99.75765	7017.353
14	Bahrain	MENA	DC	Not Africa region	99.6872	99.97614	99.30134	192.456
...	...	...	...	...	...	...	...	...
189	Uganda	SSA	LDC	Eastern Africa	89.39631	88.8269	89.95064	8861.896
195	Uruguay	LAC	DC	Not Africa region	98.88027	98.60475	99.16599	508.816
197	Vanuatu	EAP	LDC	Not Africa region	96.28212	95.97657	96.59661	53.104
199	Viet Nam	EAP	DC	Not Africa region	98.4081	98.46012	98.35314	14052.786
201	Zambia	SSA	LDC	Southern Africa	92.09157	92.55964	91.62819	3604.646

69 rows × 8 columns

Figure 4.2 Youth literacy rates

	Country	Region	Rating of Country	Africa sub-regions	Total	Male	Female	Total Population
1	Afghanistan	SA	LDC	Not Africa region	43.01972	55.47545	29.80521	20193.661
2	Albania	ECA	DC	Not Africa region	98.14115	98.51362	97.76112	1977.051
3	Algeria	MENA	DC	Northern Africa	81.40784	87.42296	75.32297	26810.322
8	Argentina	LAC	DC	Not Africa region	99.00387	98.93774	99.06204	28444.915
14	Bahrain	MENA	DC	Not Africa region	97.46419	98.76317	94.94888	1229.245
...	...	...	...	...	...	...	...	...
189	Uganda	SSA	LDC	Eastern Africa	76.5275	82.65601	70.83806	21846.758
195	Uruguay	LAC	DC	Not Africa region	98.70386	98.37136	99.00774	2227.456
197	Vanuatu	EAP	LDC	Not Africa region	87.50631	88.30635	86.70609	168.324
199	Viet Nam	EAP	DC	Not Africa region	95.00038	96.45678	93.59671	66454.47
201	Zambia	SSA	LDC	Southern Africa	86.74796	90.60118	83.08344	9190.19

69 rows × 8 columns

Figure 4.3 Adult literacy rates

Based on Wikipedia's list of countries by literacy rate by UNESCO, for the countries that have been dropped, it said that particular countries did not report their literacy rate to them. As a result there are no records for all the countries that have been dropped from the dataset and that is the reason why the rows for both dataset become 69 from 231 and 227 .

## 5. Exploratory Data Analysis (EDA)

Our data exploratory data analysis stage consists of 4 main parts: (i) Descriptive & Causal inquiry, (ii) Mechanistic inquiry, (iii) Exploratory Mining inquiry and (iv) Predictive Modelling.

The main format of our exploration involves answering the proposed questions using basic statistical analysis techniques and visualisations. Some of the questions come directly from part A of our project and some of them were new ones that we added later. Further questions or ideas to pursue that may arise from the results of such analysis are then explored later in this section.

### 5.1 Descriptive & Causal Questions

Refined Descriptive question: “ **What is the average number of youth and adult literacy rates in a developed country and undeveloped country? ”**

Additional Descriptive question: “ **What is the minimum number of youth and adult literacy rates in a developed country and undeveloped country? ”**

Refined Causal Question: “ **Can the wealth status of a country influence literacy rates? ”**

#### 5.1.1 Data Visualisation

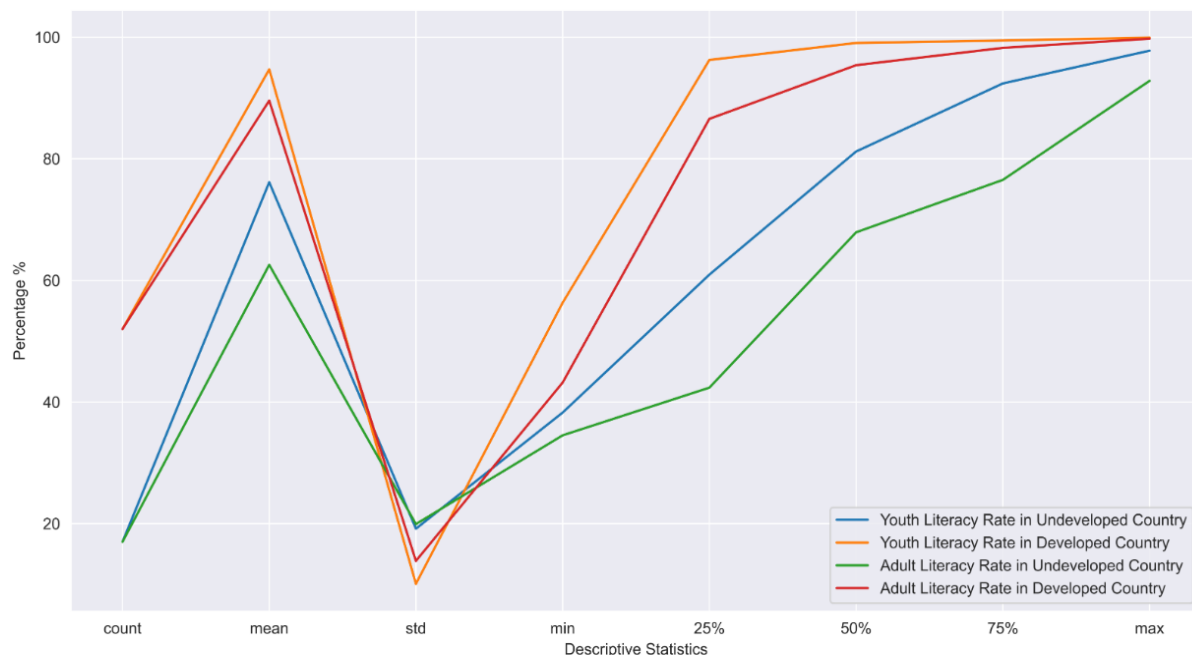


Figure 5.1.1.1 Descriptive statistics of Literacy rate of Youth and Adult in two type rating of country

Based on the Figure 5.1.1 above, we can see that the average number of youth literacy rates in Developed Country is the highest amongst all four variables with a value of 94.71%. Followed by the average adult literacy rate in Developed Country with a number of 89.56%. For Undeveloped Country, the average youth literacy rate is 76.14% and the lowest amongst all four variables, the average adult literacy rate in Undeveloped Country is 62.56%.

For the minimum number of literacy rate, first of for Developed Country, the minimum number of youth literacy rate is 56.34% while the minimum number of adult literacy rate is

43.21%. Next for Undeveloped Country, the minimum number of literacy rates of youth and adults is 38.27% and 34.52% respectively.

To answer the causal question, Based on the Figure 5.1.1, as you can see both youth and adults literacy rates for Developed Country is higher than youth and adults literacy rates for Undeveloped Country for all types of statistics. If we compare the mean, first quartile, median and 3rd quartile of the literacy rates of youth in Developed countries to the literacy rates of youth in Undeveloped Countries the values are much different as those values for youth in Developed country is reached above 90% rate, 94.71% the mean, 99% for first quartile, median and third quartile. While the value for youth in an Undeveloped country is only 76.14% for the mean, 61% first quartile, 81.2% median, and 92.4% third quartile. It is just enough to compare only the youth literacy rates as the adult is as the same stats as the youth but a bit lower for both countries. From that we can say that the wealth status of a country really influences the literacy rates from youth to adults in a country.

## 5.2 Mechanistic Question

Refined Mechanistic question: “ **Does the Developed and Undeveloped country has influences to get the different value of the literacy rates amongst the sex gender of the youth? ”**

### 5.2.1 Data Visualisation of Mechanistic question

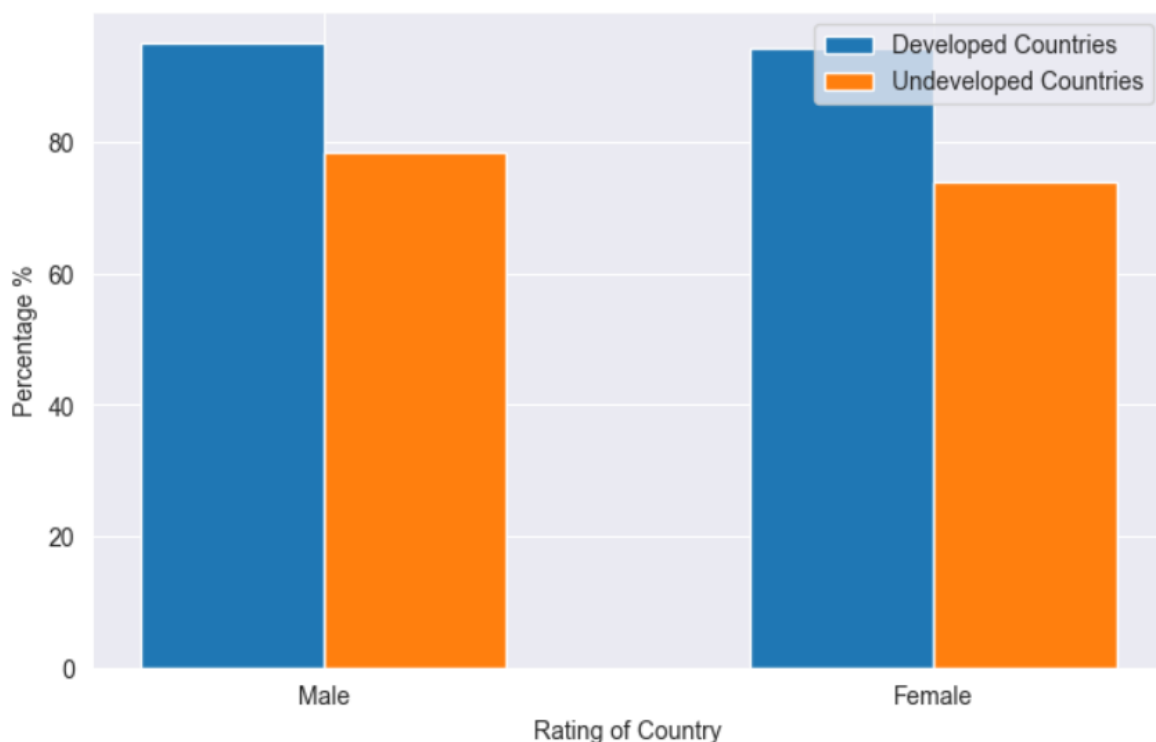


Figure 5.2.1.1 Column side by side Bar chart of Developed and Undeveloped country amongst sex gender. From the figure we can see that there is not much of a difference in the values of Male and Female literacy rate between the developed country and Undeveloped Country. So from that we can conclude that the type rating of country does not really influence the literacy rate amongst the sex gender.

### 5.3 Exploratory Question

Refined Exploratory Question: “ **What is the relationship of literacy rate amongst male and female youth and male and female adults in Developed and Undeveloped countries?** ”

First and foremost, we separate the Youth dataset in both developed countries and the underdeveloped countries by using the .loc function in pandas to locate the specific countries as shown in the two figures below.

	Country	Region	Rating of Country	Africa sub-regions	Total	Male	Female	Total population
1	Albania	ECA	DC	Not Africa region	99.33000	99.05000	99.63000	458.058
2	Algeria	MENA	DC	Northern Africa	97.42652	97.59406	97.25216	6070.282
3	Argentina	LAC	DC	Not Africa region	99.50552	99.24183	99.75765	7017.353
4	Bahrain	MENA	DC	Not Africa region	99.68720	99.97614	99.30134	192.456
6	Belarus	ECA	DC	Not Africa region	99.85000	99.83000	99.87000	926.642

Figure 5.3.1

	Country	Region	Rating of Country	Africa sub-regions	Total	Male	Female	Total population
0	Afghanistan	SA	LDC	Not Africa region	65.42055	74.08480	56.25475	8071.334
5	Bangladesh	SA	LDC	Not Africa region	93.29644	91.80372	94.91259	30778.440
7	Benin	SSA	LDC	Western Africa	60.94808	69.76118	51.94362	2279.108
10	Burkina Faso	SSA	LDC	Western Africa	58.29000	61.79000	54.67000	3960.255
12	Central African Republic	SSA	LDC	Central Africa	38.26865	47.80493	28.70666	1009.933

Figure 5.3.2

Then we do the same as the youth dataset for adults and the two data frames are as shown below.

	Country	Region	Rating of Country	Africa sub-regions	Total	Male	Female	Total Population
1	Albania	ECA	DC	Not Africa region	98.14115	98.51362	97.76112	1977.051
2	Algeria	MENA	DC	Northern Africa	81.40784	87.42296	75.32297	26810.322
3	Argentina	LAC	DC	Not Africa region	99.00387	98.93774	99.06204	28444.915
4	Bahrain	MENA	DC	Not Africa region	97.46419	98.76317	94.94888	1229.245
6	Belarus	ECA	DC	Not Africa region	99.75656	99.79398	99.72518	6455.084

Figure 5.3.3

	Country	Region	Rating of Country	Africa sub-regions	Total	Male	Female	Total Population
0	Afghanistan	SA	LDC	Not Africa region	43.01972	55.47545	29.80521	20193.661
5	Bangladesh	SA	LDC	Not Africa region	73.91220	76.66926	71.18194	108341.186
7	Benin	SSA	LDC	Western Africa	42.36240	53.97703	31.07165	6236.226
10	Burkina Faso	SSA	LDC	Western Africa	41.22445	50.07459	32.68762	10398.147
12	Central African Republic	SSA	LDC	Central Africa	37.39582	49.51459	25.75638	2467.571

Figure 5.3.4

Secondly, in order to see the relationship of male and female, we add the ‘Total’, ‘Male’, ‘Female’ variables together in the adult of developed countries and the youth of developed

countries data and the same goes for the adult of undeveloped countries and the youth of undeveloped countries data. This is because to answer the exploratory question, we need to combine the adult and youth values from the same type of rating of a country to see the correlation amongst the youth and adults in different types of rating country. The same procedure for the undeveloped countries. Figure below shows the values of the dataframe after combining the dataset together.

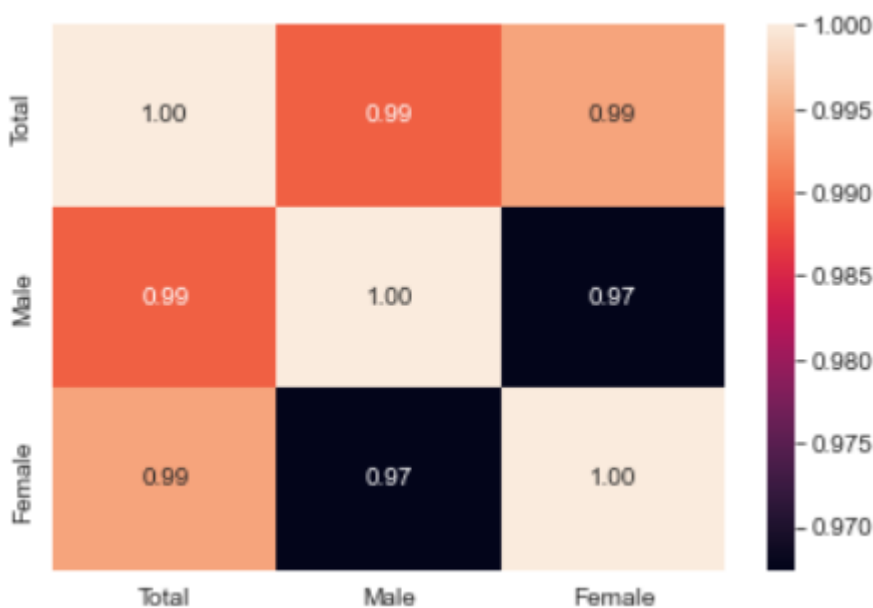
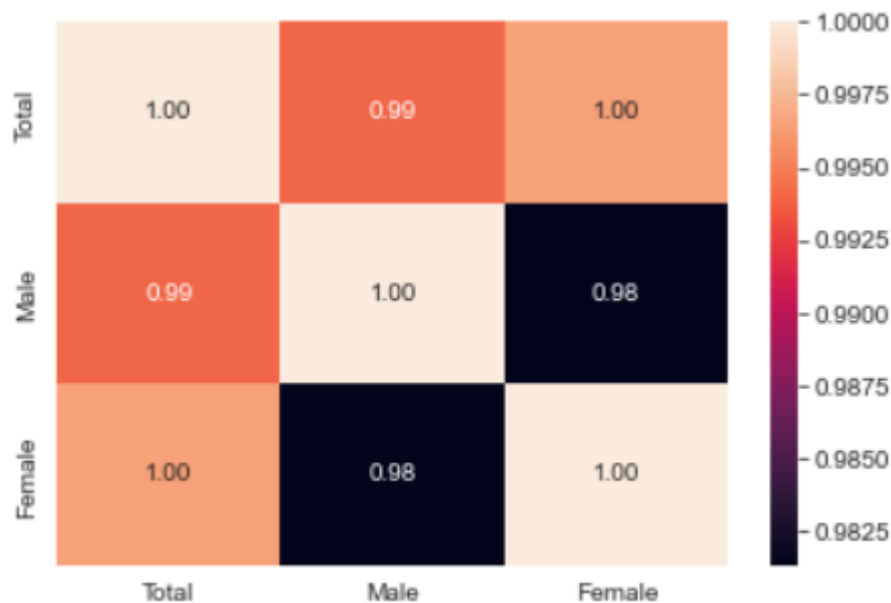
	Total	Male	Female		Total	Male	Female
<b>1</b>	197.47115	197.56362	197.39112	<b>0</b>	108.44027	129.56025	86.05996
<b>2</b>	178.83436	185.01702	172.57513	<b>5</b>	167.20864	168.47298	166.09453
<b>3</b>	198.50939	198.17957	198.81969	<b>7</b>	103.31048	123.73821	83.01527
<b>4</b>	197.15139	198.73931	194.25022	<b>10</b>	99.51445	111.86459	87.35762
<b>6</b>	199.60656	199.62398	199.59518	<b>12</b>	75.66447	97.31952	54.46304

Figure 5.3.5 Developed countries

Figure 5.3.Undeveloped countries

Lastly before visualising, to find the relationship, we use Pearson correlation to get a value that is valid for comparisons between them.

### 5.3.1 Data Visualisation of Exploratory question



The two figures above show the heatmap of both Developed countries and Undeveloped countries literacy rates of Youth and Adults. It said that it is a perfect positive relationship if the correlation values is +1, while a strong positive relationship if the values  $>0.75$ . As we can see, The value of both developed and undeveloped countries is mostly within 0.9 and 1.00. So from that we can conclude that the relationship of literacy rate amongst male and female youth and male and female adults in Developed and Undeveloped countries is a strong positive relationship.

## 5.4 Inferential Question

### Inferential Question:

“ Hypothesis : The number of youth and adults that have low literacy rate lives in least developed countries is high. ”

“ Is this hypothesis also true for the youth and adults that live in developed countries? ”

Firstly, we separate the youth and adult dataset into a new data set. The new data set will include the columns of Country, Rating of Country, Total and Total Population.

	Country	Rating of Country	Total	Total Population
0	Afghanistan	LDC	43.01972	20193.661
1	Albania	DC	98.14115	1977.051
2	Algeria	DC	81.40784	26810.322
3	Argentina	DC	99.00387	28444.915
4	Bahrain	DC	97.46419	1229.245
...	...	...	...	...
64	Uganda	LDC	76.52750	21846.758
65	Uruguay	DC	98.70386	2227.456
66	Vanuatu	LDC	87.50631	168.324
67	Viet Nam	DC	95.00038	66454.470
68	Zambia	LDC	86.74796	9190.190

	Country	Rating of Country	Total	Total population
0	Afghanistan	LDC	65.42055	8071.334
1	Albania	DC	99.33000	458.058
2	Algeria	DC	97.42652	6070.282
3	Argentina	DC	99.50552	7017.353
4	Bahrain	DC	99.68720	192.456
...	...	...	...	...
64	Uganda	LDC	89.39631	8861.896
65	Uruguay	DC	98.88027	508.816
66	Vanuatu	LDC	96.28212	53.104
67	Viet Nam	DC	98.40810	14052.786
68	Zambia	LDC	92.09157	3604.646

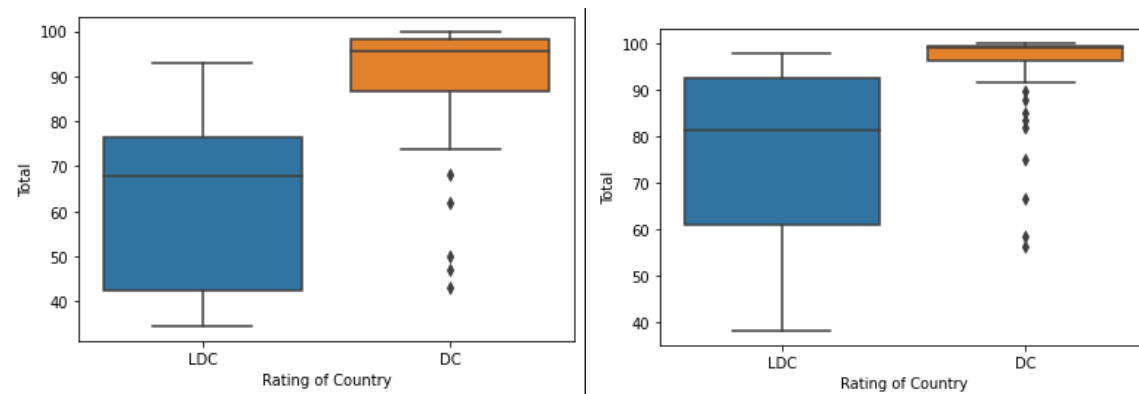
Figures 5.4.1 shows the new data set for adult and youth

Next, we will group the dataset according to their rating which is DC that stands for Developed Countries and LDC for least developed countries. This is to calculate the mean literacy rate of these countries and the total population of these countries.

<code>adultDCInfo['Total'].mean()</code>	<code>youthDCInfo['Total'].mean()</code>
✓ 0.3s	✓ 0.2s
89.56666384615386	94.70823499999999

<code>adultDCInfo['Total Population'].sum()</code>	<code>youthDCInfo['Total population'].mean()</code>
✓ 0.4s	✓ 0.5s
2996304.2980000004	13197.069884615381

### 5.4.1 Data Visualisation of Inferential question



Boxplot for Adult and Youth

As we can observe from the boxplot, we know that it is left-skewed for both boxplot of adults and youth. The median for the boxplot for adults is DC is higher compared to LDC and there are outliers for DC. Whereas the median for the boxplot for youths in DC is also much higher compared to LDC and there are outliers for DC.

To answer the inferential question, the mean for Developed Countries adults is 89.57% with a total population of 2996304 while Least Developed Countries of adults has 62.56% with a total population of 259969. The mean for Developed Countries youth is 94.71% with a total population of 686247 while Least Developed Countries of adults has 76.14% with a total population of 88021. To answer this question, the hypothesis is True, that the number of youth and adults that have low literacy rate lives in least developed countries is high. The youth and adults that live in developed countries have a low literacy rate.



## 5.5 Predictive Question

Refined Predictive question: “ **What would happen to the literacy rate if the population of a country is high?** ”

For this predictive question, after data cleaning and grouping by the attributes that we need for this predictive analysis, the dataset for youth and adults now has 69 rows with 2 columns with columns of Population and Total enrollment. It can be shown in the figure below.+

	Total population	Total
0	8071.334	65.42055
1	458.058	99.33000
2	6070.282	97.42652
3	7017.353	99.50552
4	192.456	99.68720
...	...	...
64	8861.896	89.39631
65	508.816	98.88027
66	53.104	96.28212
67	14052.786	98.40810
68	3604.646	92.09157

	Total Population	Total
0	20193.661	43.01972
1	1977.051	98.14115
2	26810.322	81.40784
3	28444.915	99.00387
4	1229.245	97.46419
...	...	...
64	21846.758	76.52750
65	2227.456	98.70386
66	168.324	87.50631
67	66454.470	95.00038
68	9190.190	86.74796

Figure 5.5.1 Youth and Adult total population and literacy rate

### 5.5.1 Data Modelling

Before beginning model training, we must first divide our data into two parts: training and test sets. The training set, also known as X train and y train, is the data that is utilised to make the model. This data set should not be utilised to assess the model's actual performance. As a result, the test set is required. The test set, also known as the X test and y test, is a piece of the data that will be held out to assess how well the model can predict unknown variables.

Now, obviously, all of the data we have is "seen," but we construct a situation in which there is "unseen" data— data that the machine did not see during the training phase. This can help us figure out how effectively the model generalises to "new" data.

By default, 'train test split' divides the data in \*X\* into a train and test section depending on the parameter 'test size=0.25', or 25% of samples in the test size and the remaining (75%) in the training set. This results in a 3:1 training-to-test ratio. The 'random state' parameter accepts a seed value to ensure that the splitting is consistent (for reproducibility purposes in this exercise). It is acceptable to omit this argument, which will result in completely random splits in each run.

Total population		Total		Total Population		Total	
63	20.134	63	99.44112	63	60.503	63	99.41437
31	1212.291	31	99.75182	31	4589.084	31	95.06944
26	5837.183	26	99.93000	26	38752.393	26	99.15576
36	22125.480	36	99.31845	36	83562.914	36	95.37991
65	508.816	65	98.88027	65	2227.456	65	98.70386
47	1075.275	47	99.66392	47	6624.167	47	96.13759
35	193.155	35	99.04203	35	896.305	35	91.32539
56	722.700	56	99.92963	56	4390.579	56	97.34486
49	13883.633	49	99.69845	49	98240.474	49	99.73006
2	6070.282	2	97.42652	2	26810.322	2	81.40784
38	82.261	38	99.11000	38	419.471	38	98.84718
48	2044.819	48	99.42919	48	12898.841	48	98.84450
34	48.675	34	99.30071	34	287.161	34	94.50319
53	1061.457	53	98.30000	53	5806.686	53	98.28920
27	1926.351	27	99.34374	27	6169.326	27	98.22711
61	240.617	61	95.46937	61	657.592	61	88.41938
51	34.948	51	99.12348	51	111.556	51	99.09577
52	40.495	52	97.78255	52	115.130	52	92.81664

Figure 5.5.1.1 of X\_test and y\_test for youths

Figure 5.5.1.2 of X\_test and y\_test for adults

After dividing the data, we conduct linear regression by using the values of X train and y train. . The regression coefficients, regression intercept and regression score can be found to determine whether the regression can be used for prediction of total literacy rate.

<code>reg.coef_</code>	<code>adultreg.coef_</code>
✓ 0.4s array([4.17529789e-05])	✓ 0.5s array([1.02631875e-05])
<code>reg.intercept_</code>	<code>adultreg.intercept_</code>
✓ 0.4s 86.43889920305413	✓ 0.3s 77.80118675481685
<code>reg.score(X_test,y_test)</code>	<code>adultreg.score(adultX_test, adulty_test)</code>
✓ 0.5s -132.34944876077523	✓ 0.3s -14.705242757183749

As we can observe that the regression score for youth and adult is both negative. Scikit-learn's LinearRegression scores use  $R^2$  score. A negative  $R^2$  means that the model fitted in data extremely badly. Since  $R^2$  compares the fit of the model with that of the null hypothesis( a horizontal straight line ), then  $R^2$  is negative when the model fits worse than a horizontal line based on the website stack overflow.

Besides that, we took the first 5 entries of the test data and used that model to create predictions on the total literacy value and do a comparison and get the difference for both adults and youths data sets.

```
some_data = X_test.iloc[:5]
predicted_enroll_values = reg.predict(some_data)
predicted_enroll_values
✓ 0.3s
array([86.43973986, 86.48951596, 86.68261898, 87.3627039 , 86.46014379])

actual_enroll_values = y_test.iloc[:5].values
actual_enroll_values
✓ 0.3s
array([99.44112, 99.75182, 99.93 , 99.31845, 98.88027])

np.abs(predicted_enroll_values-actual_enroll_values)
✓ 0.3s
array([13.00138014, 13.26230404, 13.24738102, 11.9557461 , 12.42012621])
```

Youth data set

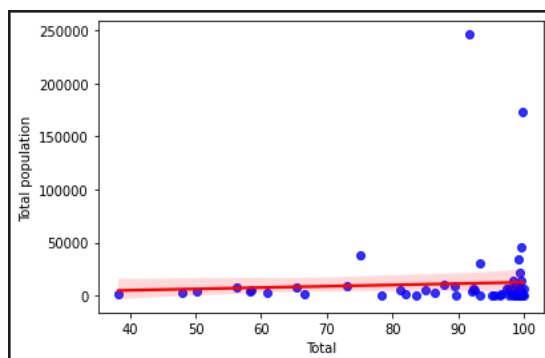
```
adult_some_data = adultX_test.iloc[:5]
adult_predicted_literacy_values = adultreg.predict(adult_some_data)
adult_predicted_literacy_values
✓ 0.3s
array([77.80180771, 77.84828538, 78.19890983, 78.65880861, 77.82404755])

adult_actual_literacy_values = adulty_test.iloc[:5].values
adult_actual_literacy_values
✓ 0.3s
array([99.41437, 95.06944, 99.15576, 95.37991, 98.70386])

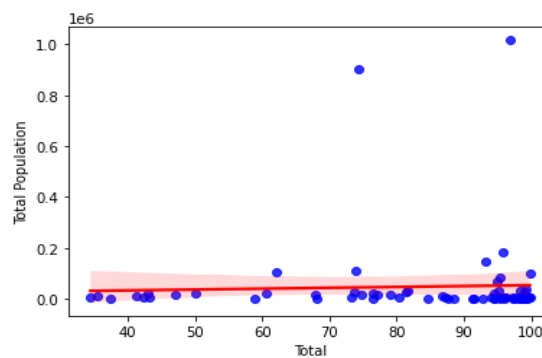
np.abs(adult_predicted_literacy_values-adult_actual_literacy_values)
✓ 0.2s
array([21.61256229, 17.22115462, 20.95685017, 16.72110139, 20.87981245])
```

Adult data set

Our team is using the model to forecast the enrollment value based on the first five entries from the test data. The 'predict()' function is able to predict the first five values of X\_test and the output is 86.44, 86.49, 86.68, 87.36 , 86.46 for youth data set and 77.80, 77.85, 78.20, 78.66, 77.82 for adult data set Then, the first five values of y\_test are taken for the actual enrollment values which are 99.44, 99.75, 99.93, 99.31, 98.88 for youth data set and 99.41 95.07, 99.16, 95.38, 98.70 for adult data set. The difference between the calculated value is shown and the result is concluded that the predicted enrollment values are near to the actual enrollment values. Hence, it performs a positive correlation.



Youth regplot



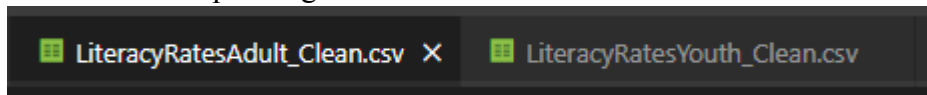
Adult regplot

A regression plot is made by using 'regplot' and it shows a positive correlation. Hence, it clearly shows that in both dataset if the number of population is high, the total literacy rate will also be high.

## 6. Challenges & Conclusion

First and foremost, the first challenge that we faced is separating and using the data set that is given. The data set was stuck together. For example the data of youth and adult is in terms of sheets instead of a separate excel file and we needed to separate it in order to obtain the value and do the analysis properly. The solution that we have come up with is to google on how to separate the sheets so that it becomes two separate csv files, from there on we manage to organise the data.

Solution after separating the csv files:



Besides that, our next problem is the data set is full of null value. There are a lot of empty spaces without info in our dataset, and we needed to drop some of the data because of insufficient data in the csv file. Therefore, the solution that our team has come out with is to drop the data and clean it so that we will be able to use the relevant data in the data set.

Furthermore, the challenges that our team faced is the workload for our team is not separated properly. Our team consists of only 2 members and puts us at a disadvantage because of the workload. When our team is doing our respective tasks separately, our jupyter notebook becomes a mess, when we try to combine it together. This is because when we try to combine the codes together, some of the variables become the same and the work that has to be done is redundant. Therefore, the solution of our problem is to separate and plan out our team's work in a more efficient and organised way so that the problem would not repeat itself in the future.

In addition, our future proposals firstly is to choose a data set wisely by researching the data set, for example checking the variable and content inside of it. Besides that, in the future, we will improve by learning in depth about how to do data visualisation, data mining, doing prediction and so on, so that we can apply it in our future job . Other than that, our team will learn and adapt on how to handle our work sufficiently without any problems.

In conclusion, our team has done the analysis of the data set of regional-aggregate Literacy rate. We are able to see the relationship between least developed countries and developed countries' literacy rate of youth and adults , male and female. Our team concluded that the number of youth and adults in developed countries has a higher literacy rate compared to undeveloped countries.

## 7. References

1. [https://en.wikipedia.org/wiki/List\\_of\\_countries\\_by\\_literacy\\_rate](https://en.wikipedia.org/wiki/List_of_countries_by_literacy_rate)
2. <https://www.google.com/search?client=opera-gx&q=what+happen+if+the+country+have+0%25+of+literacy+rates&sourceid=opera&ie=UTF-8&oe=UTF-8>
3. <https://worldpopulationreview.com/country-rankings/literacy-rate-by-country>
4. <https://stackoverflow.com/questions/63757258/negative-accuracy-in-linear-regression#:~:text=Sckit%2Dlearn's%20LinearRegression%20scores%20uses,worse%20than%20a%20horizontal%20line.>



# **TDS2101**

## **Project Part B Report**

### **Dataset 2 - Colleges and Universities in US**

**Group No.: <6>**

**<Raja Muhammad Darwisy>  
<Imran bin Zulkiflee>**

**<1191100792>  
<1191100756>**

# Table of Contents

1.	<b>Introduction</b>	3
1.1	Problem Statement	3
1.2	Motivation	3
2.	<b>Questions</b>	4
3.	<b>Data Collection</b>	7
4.	<b>Data Preprocessing</b>	8
5.	<b>Exploratory Data Analysis</b>	9
5.1	<b>Descriptive &amp; Causal Question</b>	9
5.1.1	Data Visualisation of Descriptive	8
5.1.2	Data Visualisation of Causal	11
5.2	<b>Mechanistic Question</b>	12
5.2.1	Data Visualisation	12
5.3	<b>Exploratory Question</b>	14
5.3.1	Data Visualisation	16
5.4	<b>Inferential Question</b>	17
5.4.1	Data Visualisation	18
5.5	<b>Predictive Question</b>	19
5.5.1	Data Modelling	19
6.	<b>Challenges &amp; Conclusion</b>	22
7.	<b>References</b>	23

# 1. Introduction

## 1.1 Problem Statement

Unincorporated countries or Islands in North America tend to have high enrollment in a college or an university which increase the Student financial aids, so comparing population amongst the countries in NA is a good starting point.

## 1.2 Motivation

Enrolment in a university or college is becoming a must in every single part of the world, especially the population of North America. According to the article from Melanie Hanson, enrollment to a university or college amongst North America peaked in 2010 at 21.0 million where 12.0 million or 60.9% of all students, graduate and undergraduate, are enrolled full-time. Table 1 below shows the rate of students that enrol from the year 2017 to 2021.

Year	Total Enrollment	Public Institution Enrollment Rate
2017	19,778,151	74.0%
2018	19,651,412	74.0%
2019	19,637,499	73.8%
2020**=	19,744,000	74.0%
2021*	19,778,000	74.0%

Table 1

Besides that, unincorporated countries or Islands in North America tend to have high enrollment in a college or a university which will increase the Student financial aid. This has become a trend because the smaller population in unincorporated countries or Islands in North America are able to support the population by giving the population to enrol in higher education. Therefore, our team is curious about the relationship of unincorporated and corporated countries or Islands that provide financial support affects the rate of enrolment in universities or colleges in North America.

Because of these problems, our team has the motive to find out the the rate of enrolment between unincorporated and corporated countries or Islands in North America, the difference between enrolment rate and the impact (whether there is any) of providing financial aid to the population that will affect the enrolment rate. The data set that we are going to use is mainly about the enrolment rate amongst different countries. Our team is eager to find out whether living in unincorporated countries or Islands will affect the enrollment rate in a university or college . By solving this problem, it will have a huge impact on increasing the enrolment in a corporated country in North America.



## 2. Questions

### Type of question: Descriptive Question

A descriptive question attempts to summarise a feature of a collection of data. For example, in a set of data obtained from a group of people, identifying the proportion of males, the mean number of servings of fresh fruits and vegetables per day, or the incidence of viral infections are all examples. Because the outcome is a fact, an attribute of the set of data you're working with, there's no way to interpret it.

**“ What is the comparison of the population of Unincorporated Countries and Corporated Countries enrollment in enrolling to a college or an university in North America? ”**

Refined Descriptive Question:

**“ What is the maximum average of the population of Unincorporated Countries and Corporated Countries enrolling in a college or a university in North America? ”**

For this question, we refined the descriptive question as a comparison word is not suitable for the descriptive type of question. Firstly we decided to locate the corporate country and unincorporated country in the dataset as it is easier for us to compare the average of both types of countries. Next we use the numpy package to use the aggregate function to get the mean of the variables. Then plot a horizontal bar graph to make an analysis of the question.

### Type of question: Causal Question

A causal question examines if changing one factor in a population will impact another factor on average. By default, the basic nature of the data gathering sometimes permits you to pose a causal question.

**“ How is the population of a country or county affecting the enrollment of students into college and university? ”**

Refined Causal Question:

**“ Does higher enrollment rate of students into college and universities affect the employment rate? ”**

Our team plans to do data visualisation for the Causal question. We wanted to find out the relationship between high enrollment rate and employment rate whether both of them will increase or decrease. The variable we will be using for this question is TOT\_ENROLL and TOT\_EMPLOY. We were planning to do a scatter plot since both of the variables are continuous. Besides that, we will do a correlation analysis to find the value between enrollment and employment, then a heatmap will be plotted as a visual representation.

Type of question: Mechanistic Question

A mechanistic question is one that focuses on how one element influences the outcome. (For example, how a healthy diet can reduce the amount of viral infections.) A mechanistic question might be one that asks how a diet rich in fresh fruits and vegetables reduces the prevalence of viral diseases.

**“ How does the number of employment play a role in determining the number of enrolled students? ”**

Refined Mechanistic Question:

**“ Does the number of employment play a role in determining the number of enrolled students? ”**

For this question, we refined the question to a more convincing mechanistic question. Firstly, we decided to get a descriptive stats of the the specific variables which in this question is TOT\_ENROLL which represent the number of employment and TOT\_EMPLOY represent the number of enrolled students and plot a line graph of it to analyse the central tendency between the variables and to check if any dispersion of data to make an early conclusion of any significant role between these two variables. After that we decided to plot a stacked bar graph to strengthen the conclusion from the previous graph.

Type of question: Exploratory Question

An exploratory question examines the data to determine if any patterns, trends, or correlations between variables may be discovered.

**“ What is the relationship between enrollment in college or university in corporate countries and unincorporated countries? ”**

For this exploratory question type, we want to observe the exact relationship of the type of country with the enrollment into college and university in this dataset. At first we decided to do some data mining for this type of analysis but due to the question that we want to analyse, it is not suitable and it would make the analysis redundant. So we decided to do a quick correlational analysis to evaluate the relationship between these variables. The correlational analysis we will use Pearson Correlation of each type of country. Then we use heatmap to see a clear description to make an analysis of the relationship between the enrollment and both corporate countries and unincorporated countries whether strong or weak correlation between these variables.

Type of question: Inferential Question

An inferential question would be a restatement of this proposed hypothesis as a question and would be answered by analysing a different set of data. These types of analyses are often known as "hypothesis generating" analyses because it looks at patterns to generate hypotheses rather than testing them.

**“ Hypothesis : The number of enrolment in a college or an university of unincorporated country or islands is higher. ”**

**“ Is this hypothesis accepted? ”**

For this inferential question type, we want to observe whether the number of enrolment in a college or university of an unincorporated country is higher. By doing this hypothesis, we can also check whether it also applies to corporate countries. Then, we plan to get the total of enrollment of the countries and compare it to the total population of the respective countries.

Type of question: Predictive Question

A predictive inquiry is one in which you inquire about the set of predictors or factors that influence a specific behaviour.

**“ Will a student accept a financial aid award offered to them for the financial need to enrol into college or university? ”**

Refined Predictive question:

**“ What would happen to the number of enrollment to college or university if the population of the country is high? ”**

Our team plans to do model training for the predictive question. We will use the POPULATION and TOT\_ENROLL columns to predict the value of population and the total enrollment by using linear regression. The regression coefficients, regression intercept and regression score can be calculated to determine whether the regression can be used for prediction of enrollment rate. In this model training, our goal is to find out whether linear regression is good for predicting enrollment rate.

### 3. Data Collection

The colleges and universities in US Dataset was sourced from Kaggle.com, The colleges and university dataset is composed of all **Post Secondary Education** facilities as defined by the **Integrated Post Secondary Education System (IPEDS)**, **National Centre for Education Statistics, US Department of Education**.

Included are Doctoral/Research Universities, Masters Colleges and Universities, Baccalaureate Colleges, Associates Colleges, Theological seminaries, Medical Schools and other health care professions, Schools of engineering and technology, business and management, art, music, design, Law schools, Teachers colleges, Tribal colleges, and other specialised institutions. Overall, this data layer covers all 50 states, as well as Puerto Rico and other assorted U.S. territories. This feature class contains all MEDS as approved by National Geospatial-Intelligence Agency (NGA).

ESS	ADDRESS2	CITY	STATE	ZIP	...	ALIAS
604 Locust St	NOT AVAILABLE	N Little Rock	AR	72114	...	NOT AVAILABLE
7518 Baird Way	NOT AVAILABLE	Citris Heights	CA	95610	...	NOT AVAILABLE
1670 Hillhurst Avenue	NOT AVAILABLE	Los Angeles	CA	90027	...	NOT AVAILABLE
1271 North Main Street	NOT AVAILABLE	Salinas	CA	93906	...	NOT AVAILABLE
1 Education Street	NOT AVAILABLE	Cambridge	MA	02141	...	NOT AVAILABLE
...	...	...	...	...	...	...
1154 Poquonnock Rd	NOT AVAILABLE	Groton	CT	06340	...	NOT AVAILABLE
1201 Locust Ave	NOT AVAILABLE	Fairmont	WV	26554	...	NOT AVAILABLE

7735 rows × 46 columns [Open in new tab](#)

Figure 3.1

For each field the 'NOT AVAILABLE' designations in the Figure 1 above are used to indicate that the data for the particular record and field is currently unavailable and will be populated when and if that data becomes available. But for this project we will drop it as we will not use the specific attributes for our analysis.

## 4. Data Preprocessing

Cleaning the core data, as well as the other auxiliary datasets, was required across a wide range of dimensions. The majority of this step occurs at the start of the pipeline, before any meaningful analysis is performed.

The first and most obvious problem was the poor data quality, that is, the large number of missing values for certain attributes such as 'ADDRESS2', 'ALIAS' and 'SHELTER\_ID'. First and foremost when dealing with missing values we need to replace values with missing values (NA). This is useful in cases when we know the origin of the data and can be certain which values should be missing. So we replace the 'NOT AVAILABLE' values to 'Nan' Values by importing the nan as NA from Numpy Packages in Python. We then check if there are any other missing Values for example values of 'N/A' or maybe '-99' in other attributes to further clean the dataset by dropping those missing values. Plus there is some duplicated data for example there is 'X' and 'Y' which is already recorded as the columns called 'LATITUDE' and 'LONGITUDE'.

There are also irrelevant attributes to our analysis purposes such as 'ADDRESS2', 'ZIP', 'ZIP4', 'TELEPHONE', 'SOURCE', 'SOURCE\_DAT', 'VAL\_METHOD', 'WEBSITE', 'VAL\_DATE', 'CLOSE\_DATE', 'MERGE\_ID', 'ALIAS', 'SIZE\_SET', 'INST\_SIZE', 'SHELTER\_ID'. So we decided to drop all of these columns.

After those cleaning processes, we then save the clean dataset into a new csv file named 'Colleges\_and\_Universities\_Clean.csv' to make the analysis process easier when we move to different stages of the pipeline.

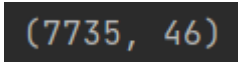
A black rectangular box with the text "(7735, 46)" in a light blue, monospace-style font.

Figure 4.1: Before Cleaning

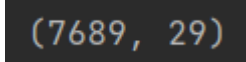
A black rectangular box with the text "(7689, 29)" in a light blue, monospace-style font.

Figure 4.2: After cleaning

Figure 2 and 3 shown that the dataset's shape Before and After cleaning process

## 5. Exploratory Data Analysis (EDA)

Our data exploratory data analysis stage consists of 4 main parts: (i) Descriptive & Causal inquiry, (ii) Mechanistic inquiry, (iii) Exploratory Mining inquiry and (iv) Predictive Modelling.

The main format of our exploration involves answering the proposed questions using basic statistical analysis techniques and visualisations. Some of the questions come directly from part A of our project and some of them were new ones that we added later. Further questions or ideas to pursue that may arise from the results of such analysis are then explored later in this section.

### 5.1 Descriptive & Causal Questions

Descriptive refined question: “ *What is the maximum average of the population of Unincorporated Countries and Corporated Countries enrolling in a college or a university in North America? ”*

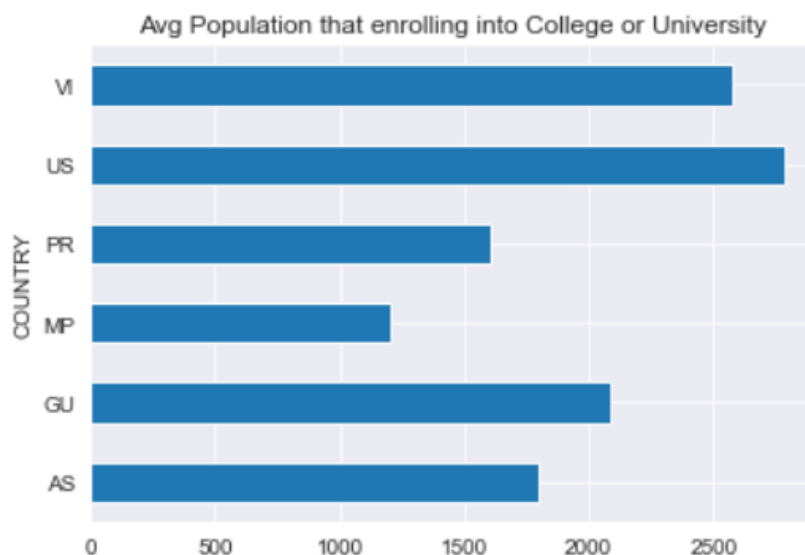


Figure 5.1.1:  
Horizontal Bar chart of  
Average Population of  
each country in North  
America enrolling into  
college or university

The following Figure 4 compares the Maximum Average Population that enrolls in a college or university of each country in North America, it is shown that the United States (US) has the largest maximum average Population enrolling, 2785.57 thousands people to be exact. But not far enough behind is the U.S Virgin Islands (VI) with Average population enrollment of 2579 thousands people. The other countries that follow are Puerto Rico (PR), Northern Mariana Islands (MP), Guam (GU), American Samoa (AS). The counts range from about 1795 thousands to 2785.57 thousands people.

The five major unincorporated U.S territories are American Samoa, Puerto Rico, Guam, the U.S Virgin Islands and the Northern Mariana Islands. We noticed that only the U.S Virgin Islands has a large average population amongst the five unincorporated countries. But as we can see the United States is undoubtedly one of the corporated countries that has the highest Population of enrolment into college or university in North America.

Causal refined question: “ ***Does higher enrollment rate of students into college and universities affect the employment rate?*** ”

For this causal question, after data cleaning and grouping by the attributes that we need for this causal analysis, the dataset now has 7689 rows with 4 columns with columns of Name, Country, Total enrollment and Total employment as shown in the figure 5 below. For this causal question, we will be needing two variables which are total enrollment of the country and total employment

Data visualisation with the data frame Figure 5, we used scatter plot by plotting total enrollment and total employment to observe how total enrollment affects total employment.

	NAME	COUNTRY	TOT_ENROLL	TOT_EMPLOY
0	Shorter College	US	52	18
1	Citrus Heights Beauty College	US	30	9
2	Joe Blasco Makeup Artist Training Center	US	24	11
3	Waynes College of Beauty	US	34	9
4	Hult International Business School	US	2243	143
...	...	...	...	...
7684	Connecticut Center for Massage Therapy-Groton	US	138	25
7685	Pierpont Community and Technical College	US	3187	283
7686	Universal College of Beauty Inc-Compton	US	0	0
7687	ITT Technical Institute-Duluth	US	668	84
7688	ITT Technical Institute-Hilliard	US	484	83

7689 rows × 4 columns

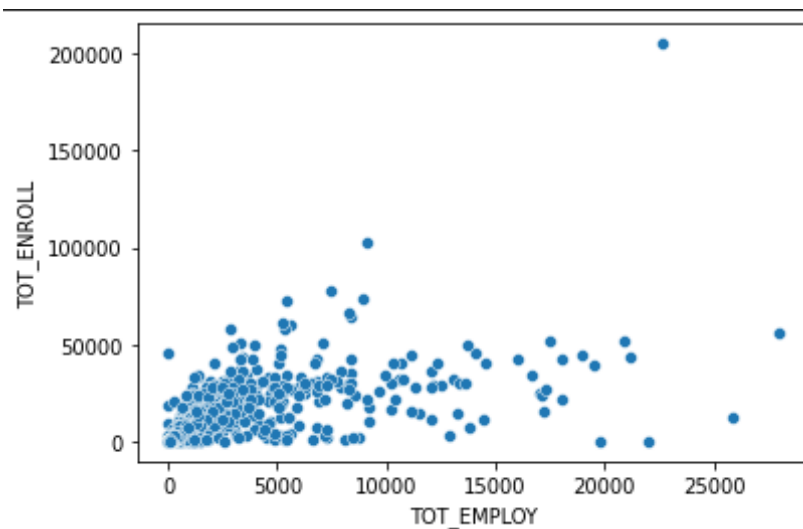
Figure 5.1.2

From the perspective of statistical analysis of the dataframe, we would be able to know that the maximum number of total enrollment is 204920 while the maximum number of total employment is 28018. The average number of total enrollments for a country is 2760.338275 whereas the average number of total employment is 471.724932.

	TOT_ENROLL	TOT_EMPLOY
count	7689.000000	7689.000000
mean	2760.338275	471.724932
std	6677.184788	1474.543436
min	0.000000	0.000000
25%	117.000000	21.000000
50%	437.000000	75.000000
75%	2244.000000	384.000000
max	204920.000000	28018.000000

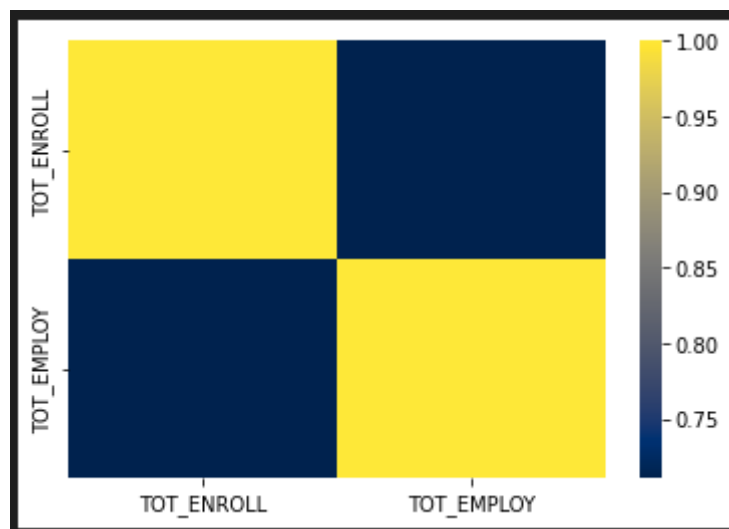
Figure 5.1.3: Statistical summary of total enrollment and total employment

### 5.1.2 Data Visualisation of Causal



The scatter plot shows that it has a moderate positive relationship. We can comment that there is a strong linear association between enrollment and total employment because the data points are near the average points.

By getting in depth regarding the relationship of total enrollment and total employment, we calculated that the correlation value between the two variables is 0.71105 which means that it has a high correlation. The correlation value is then visualised in a heatmap.



As a conclusion, we can conclude from the above information that the enrollment rate does affect the employment rate. This can be shown through the scatter plot where the correlation value is high. Based on an article by Ricardo Nogales, Pamela Cordova and Manuel Urquidi, it says that educational institutions such as colleges and universities are increasingly paying attention to the labour market success of their graduates as a means to boost their reputation among employers. By referencing this article, it clearly shows that if students enrol in a college or university (higher education), they will have a high chance to get employed by companies. Besides that, there is also another article by the National Centre for education statistics, In March 2020, the employment rate was higher for those who attended higher education institutions. For example, the employment rate was highest for students who have a bachelor's or higher degree which is 86% .

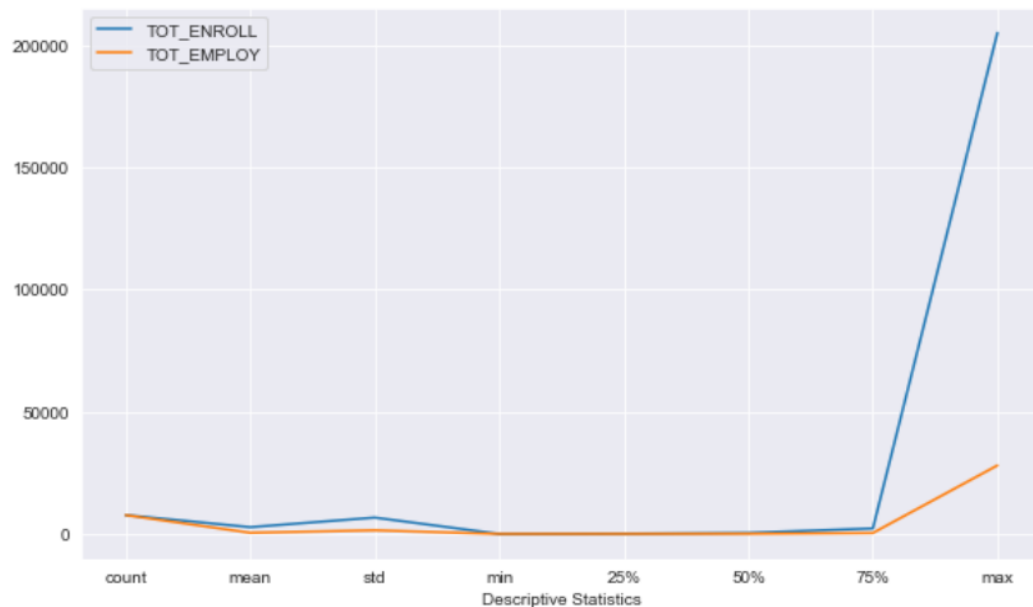


## 5.2 Mechanistic Question

Mechanistic refined Question: “ Does the number of employment play a role in determining the number of enrolled students? ”

### 5.2.1 Data Visualisation of Mechanistic Question

Figure 5.2.1.1 descriptive statistics of total enrollment and total employment



We use descriptive statistics to measure the central tendency of the data for attributes total enrollment and total employment as the frequency of both attributes are the same. From Figure 5.2.1.1, we can see that average and standard deviation of total enrollment is slightly higher than total employment but the middle values of the data of both attributes is the same value, whereas the measures of variability is dispersed on the max value as total enrollment value is skyrocketing far away from total employment. From that we can conclude that the number of employment does not really determine the number of enrolled students. It is because when total employment increases the total number of enrolled students also should increase in order to determine that these two variables are playing a role to each other.

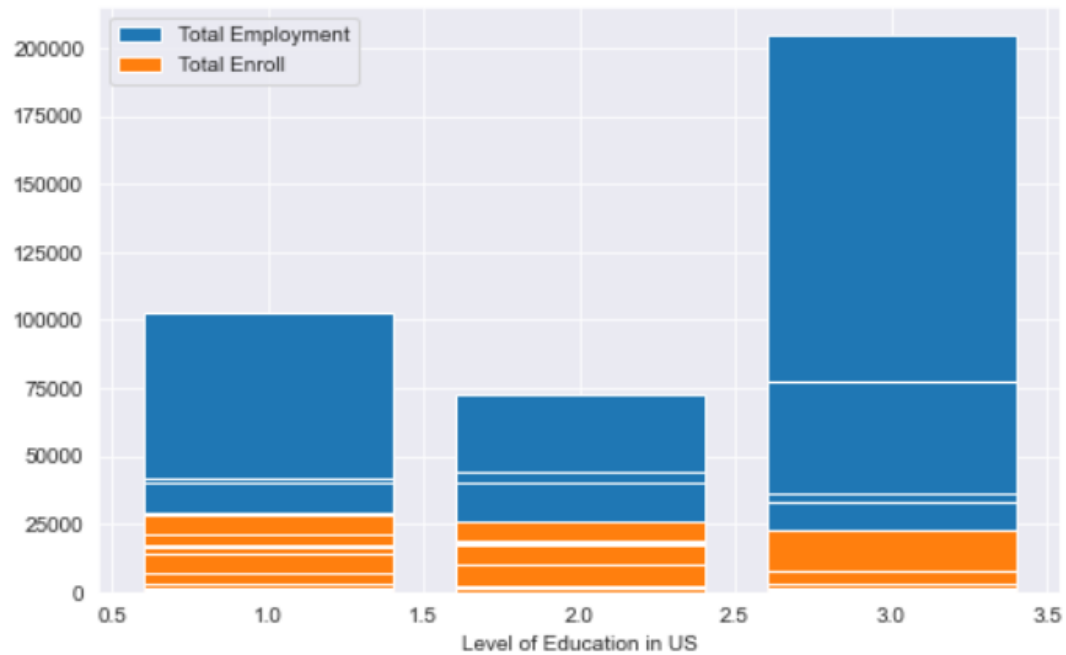


Figure 5.2.1.2 Stacked bar graph of Total Employment and Total Enrollment

Figure 5.2.1.2 is showing the clearer description of why the number of employment does not play a significant role to the number of enrolled students. Based on the figure above, the total enrollment still has a similar value for every level of education in the US, while the number of total employment has changed values for every level of education. From that we can strongly back up the previous conclusion that the number of employment does not have any significant role to the total number of enrolled students.

### 5.3 Exploratory Question

Exploratory question: “ **What is the relationship between enrollment in college or university in corporate countries and unincorporated countries?** ”

First and foremost, we locate all the corporate countries which in this dataset are only US, so we use the .loc function in pandas to locate the specific countries as shown in the figure below.

	POPULATION	COUNTRY	PT_ENROLL	FT_ENROLL	TOT_ENROLL
0	70	US	24	28	52
1	39	US	6	24	30
2	35	US	0	24	24
3	43	US	18	16	34
4	2386	US	0	2243	2243
...	...	...	...	...	...
7684	163	US	119	19	138
7685	3470	US	772	2415	3187
7686	0	US	0	0	0
7687	752	US	227	441	668
7688	567	US	221	263	484

7523 rows × 5 columns

Figure 5.3.1 : Dataframe of Corporate country US

	POPULATION	PT_ENROLL	FT_ENROLL	TOT_ENROLL
POPULATION	1.000000	0.729548	0.927415	0.991087
PT_ENROLL	0.729548	1.000000	0.445526	0.769674
FT_ENROLL	0.927415	0.445526	1.000000	0.914483
TOT_ENROLL	0.991087	0.769674	0.914483	1.000000

Figure 5.3.2 : Dataframe of Correlation between Population and Enrollment of Corporate country

Next, We then use Pearson correlation, by using the corr() function to see the relationship between enrollment in college or university and the corporated countries' population.

For Unincorporated countries also we do the same step as Corporate countries by firstly locating all the Unincorporated countries in the dataset then using the same step and method to find the correlation in Unincorporated Countries as shown in the two figures below.

	POPULATION	COUNTRY	PT_ENROLL	FT_ENROLL	TOT_ENROLL
43	38	PR	0	28	28
71	504	PR	140	295	435
86	93	PR	0	81	81
87	483	PR	0	267	267
99	96	PR	0	7	7
...	...	...	...	...	...
7614	640	PR	0	579	579
7615	317	PR	0	285	285
7616	288	PR	0	259	259
7624	96	PR	0	82	82
7659	193	PR	0	152	152

166 rows × 5 columns

Figure 5.3.3: Dataframe of Unincorporate Country

	POPULATION	PT_ENROLL	FT_ENROLL	TOT_ENROLL
POPULATION	1.000000	0.761946	0.987104	0.997139
PT_ENROLL	0.761946	1.000000	0.663796	0.777453
FT_ENROLL	0.987104	0.663796	1.000000	0.986464
TOT_ENROLL	0.997139	0.777453	0.986464	1.000000

Figure 5.3.4: Dataframe of Correlation between Population and Enrollment of Unincorporated Countries

Using Pearson correlation, It is the most common way to quantify a relationship between two variables which is a measure of the association between two variables. It has a value between -1 and 1 where -1 indicates a perfectly negative correlation, 0 indicates no linear correlation, and 1 indicates a perfectly positive linear correlation between two variables.

### 5.3.1 Data Visualisation of Exploratory Question

For Data visualisation, we use heatmap from seaborn package as we can see clearer relationships for two different variables that we want to analyse by using the previous Pearson correlation values and build the heatmap.

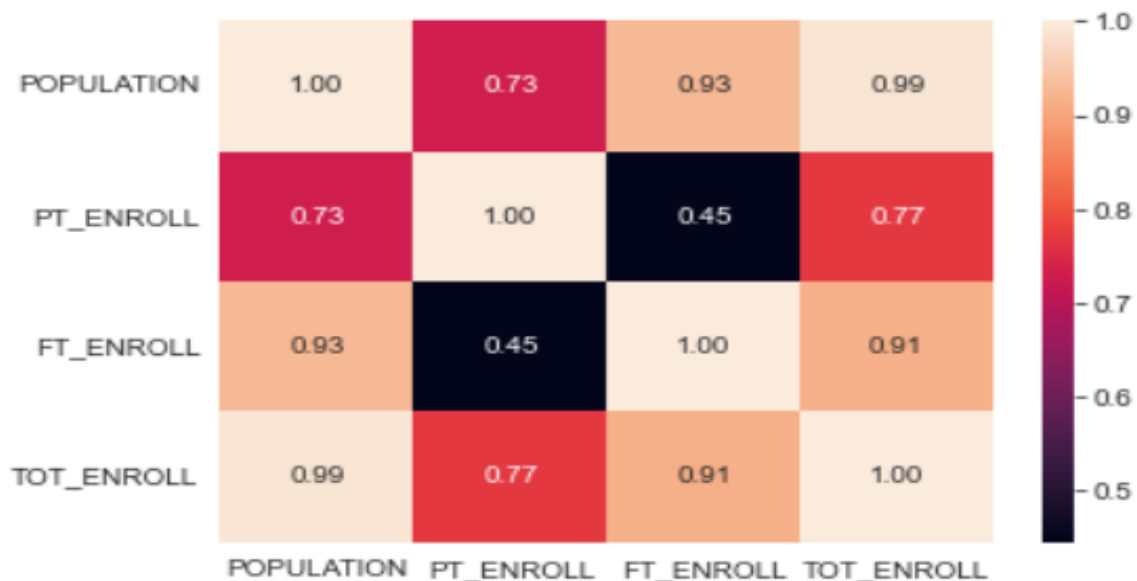


Figure 5.3.1.1: Correlation of Corporate country US

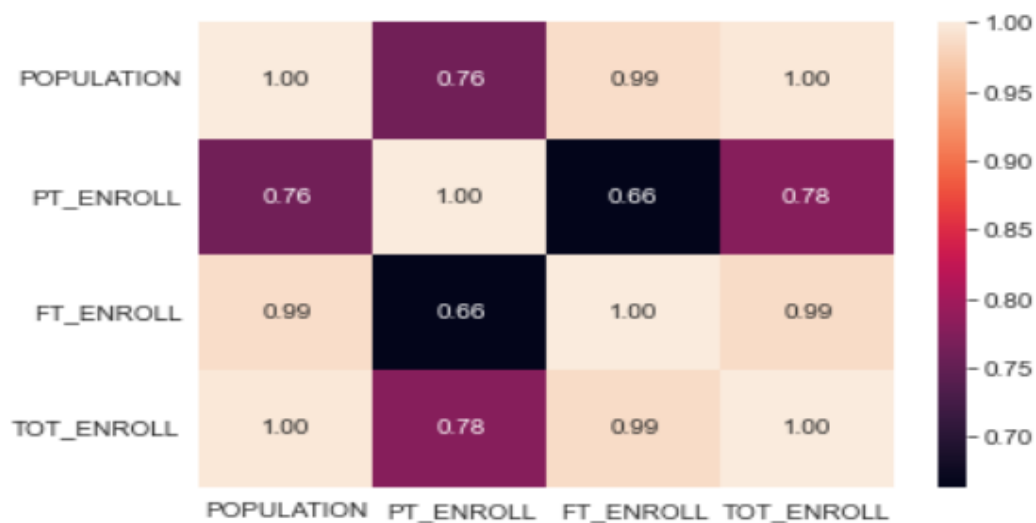


Figure 5.3.1.2: Correlation of Unincorporated Countries

Based on the two figures above, we can assume that the relationship between Enrollment and Population for both Corporate Country US and Unincorporated Countries has a positive correlation as all of the values are above 0 and even larger than 0.7. As we can see, the Relationship between total enrollment and population for both is larger than 0.75 with the value of 0.99 for Corporate country and 1.00 for the unincorporated country. This means that for both types of country, the higher the population of the country, the higher the enrollment into college and university. So we can conclude that Enrollment and Population country have a strong positive correlation. But the difference is that for part-time and full-time enrollment for a Corporate country is slightly lower than for an unincorporated country. From that, We can say that the financial aid for an unincorporated country is slightly higher than a corporate country.

## 5.4 Inferential Question

### Inferential Question:

“ Hypothesis : The number of enrolment in a college or an university of an unincorporated country or islands is higher compared to a corporated country. ”

“ Is this hypothesis accepted? ”

First and Foremost, we need to separate the corporated and unincorporated countries into two different dataset. By doing so, we will be able to obtain the value of total enrollment of the different countries.

Figure below shows the head which is 5 of the data set of unincorporated countries

	FID	IPEDSID	NAME	ADDRESS	CITY	STATE	TYPE	STATUS	POPULATION	COUNTY	...	LEVEL	HI_OFFER	DEG_GRANT	LOCALE	PT_ENROLL	FT_ENROLL	TOT_ENROLL	
43	7044	475644	Global Institute	Ave Luis Munoz Marin Esq Georgetti	Caguas	PR	3	N	38	Caguas	...	3	2		2	13	0	28	28
71	7072	468723	National University College-Caguas	190 Ave. Gautier Ben?tez esquina Ave. Federico...	Caguas	PR	3	A	504	Caguas	...	1	5		1	13	140	295	435
86	7087	469391	Dewey University	Carr# 182 Km 0.2, Catalina Morales Street	Yabucoa	PR	2	A	93	Yabucoa	...	3	2		2	21	0	81	81
87	7088	469407	Dewey University	State. Road #2 Km. 86.9, Bo. Pueblo	Hatillo	PR	2	A	483	Hatillo	...	3	2		2	41	0	267	267
99	7100	475811	Universidad Internacional Iberoamericana	Carr. 658, Km. 1.3, Barrio Arenalejos, S?ctor ...	Arecibo	PR	2	N	96	Arecibo	...	1	7		1	41	0	7	7

Figure below shows the head which is 5 of the data set of corporated countries

	FID	IPEDSID	NAME	ADDRESS	CITY	STATE	TYPE	STATUS	POPULATION	COUNTY	...	LEVEL	HI_OFFER	DEG_GRANT	LOCALE	PT_ENROLL	FT_ENROLL	TOT_ENROLL
0	7001	107840	Shorter College	604 Locust St	N Little Rock	AR	2	R	70	Pulaski	...	2	3	1	13	24	28	52
1	7002	112181	Citrus Heights Beauty College	7518 Baird Way	Citrus Heights	CA	3	R	39	Sacramento	...	3	2	2	21	6	24	30
2	7003	116660	Joe Blasco Makeup Artist Training Center	1670 Hillhurst Avenue	Los Angeles	CA	3	R	35	Los Angeles	...	3	1	2	11	0	24	24
3	7004	125310	Waynes College of Beauty	1271 North Main Street	Salinas	CA	3	R	43	Monterey	...	3	2	2	12	18	16	34
4	7005	164368	Hult International Business School	1 Education Street	Cambridge	MA	2	R	2386	Middlesex	...	1	7	1	12	0	2243	2243

After differentiating the data set, we will be able to obtain the total value of both enrollment of respective countries which is 20955813 for corporated countries and 268428 for unincorporated countries

	Corperated	Uncorperated
0	20955813	268428

To answer our inferential question, we would have to compare the enrollment rate and the population rate of the countries.

	Population of Corperated	Population of Uncorperated
0	24543493	307841

The difference for Corported countries is 1.1712 and unincorporated countries is 1.1468. Therefore, the hypothesis is accepted which is that the number of enrolment in a college or an university of an unincorporated country or islands is higher compared to a corporated country.

## 5.5 Predictive Question

Refined Predictive question: “ **What would happen to the number of enrollment to college or university if the population of the country is high? ”**

For this predictive question, after data cleaning and grouping by the attributes that we need for this predictive analysis, the dataset now has 7689 rows with 2 columns with columns of Population and Total enrollment. It can be shown in the figure below.

	POPULATION	TOT_ENROLL
0	70	52
1	39	30
2	35	24
3	43	34
4	2386	2243
...	...	...
7684	163	138
7685	3470	3187
7686	0	0
7687	752	668
7688	567	484

### 5.5.1 Data Modelling

Before beginning model training, we must first divide our data into two parts: training and test sets. The training set, also known as X train and y train, is the data that is utilised to make the model. This data set should not be utilised to assess the model's actual performance. As a result, the test set is required. The test set, also known as the X test and y test, is a piece of the data that will be held out to assess how well the model can predict unknown variables.

Now, obviously, all of the data we have is "seen," but we construct a situation in which there is "unseen" data— data that the machine did not see during the training phase. This can help us figure out how effectively the model generalises to "new" data.

```
X = dataset2.drop(["TOT_ENROLL"], axis =1)
y = dataset2["TOT_ENROLL"].copy()
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=1)
```

By default, 'train test split' divides the data in \*X\* into a train and test section depending on the parameter 'test size=0.25', or 25% of samples in the test size and the remaining (75%) in the training set. This results in a 3:1 training-to-test ratio. The 'random state' parameter accepts a seed value to ensure that the splitting is consistent (for reproducibility purposes in



this exercise). It is acceptable to omit this argument, which will result in completely random splits in each run.

POPULATION		TOT_ENROLL		POPULATION		TOT_ENROLL	
3527	771	3527	664	2444	188	2444	154
2083	184	2083	166	1412	304	1412	253
6926	189	6926	159	1190	113	1190	92
2968	752	2968	672	1273	1448	1273	1303
4670	6372	4670	5439	6662	119	6662	103
...	...	...	...	...	...	...	...
7644	2411	7644	2021	905	310	905	266
6476	11623	6476	9830	5192	3118	5192	2336
5577	278	5577	203	3980	33	3980	31
2270	25300	2270	21500	235	102	235	88
1653	2907	1653	2631	5157	118	5157	107

Dataframe for X\_test, y\_test, X\_train, y\_train

After dividing the data, we conduct linear regression by using the values of X train and y train. . The regression coefficients, regression intercept and regression score can be found to determine whether the regression can be used for prediction of enrollment rate.

```

reg.coef_
✓ 0.3s
array([0.84131372])

reg.intercept_
✓ 0.3s
41.98919880719541

reg.score(X_test,y_test)
✓ 0.3s
0.9817094074402414

```

For the function `reg.score()` which uses the linear regression fit of `X_train` and `y_train`, it returns the  $r^2$  determination coefficient which measures how well is the fitting. For example, if the value of the  $r^2$  is closer to 1.0, it will result as a perfect fit while -1.0 results in a poorest fit, any value around 0.0 basically means the model predicts something disregarding the input features, which is as good as random. The `reg.score()` in the regression is 0.9817 is a good fit of the model because the value is closer to 1. So, the linear regression model is recommended to predict the number of employment if a population of a country is high.

```

some_data = X_test.iloc[:5]
predicted_enroll_values = reg.predict(some_data)
predicted_enroll_values
✓ 0.3s

array([ 690.64207765, 196.79092346, 200.99749207, 674.65711696,
        5402.84022865])

actual_enroll_values = y_test.iloc[:5].values
actual_enroll_values
✓ 0.4s

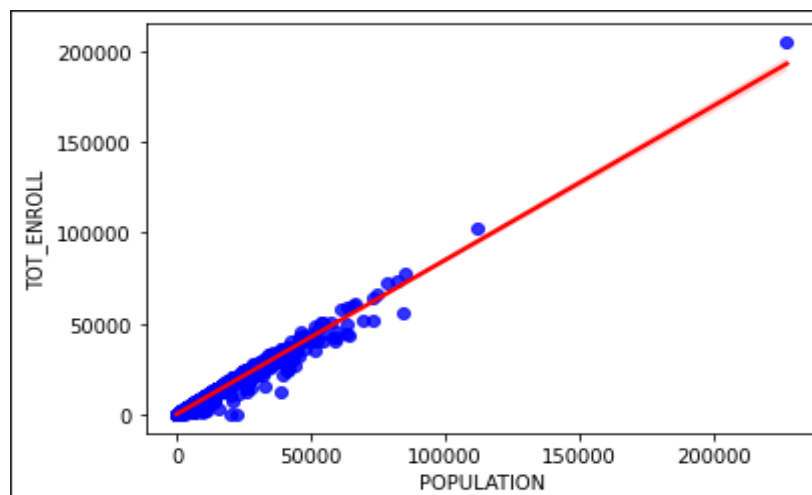
array([ 664, 166, 159, 672, 5439], dtype=int64)

np.abs(predicted_enroll_values-actual_enroll_values)
✓ 0.9s

array([26.64207765, 30.79092346, 41.99749207, 2.65711696, 36.15977135])

```

Our team is using the model to forecast the enrollment value based on the first five entries from the test data. The 'predict()' function is able to predict the first five values of X\_test and the output is 690.64, 196.79, 200.99, 674.66, 5402.84. Then, the first five values of y\_test are taken for the actual enrollment values which are 664, 166, 159, 672 and 5439. The difference between the calculated value is shown and the result is concluded that the predicted enrollment values are near to the actual enrollment values. Hence, it performs a positive correlation.



A regression plot is made by using 'regplot' and it shows a positive correlation. Hence, it clearly shows that if the number of population is high, the total number of enrollment is also high. To answer this predictive question the number of enrollment to college or university will be high if the population of the country is high.

## 6. Challenges & Conclusion

First and foremost, the first challenge that we faced is that the data set contains a number of unknown attributes such as from the column TYPE, STATUS, SECTOR and LEVEL. The dataset that we have chosen lacks the description for our team to further understand and fully utilise the data set. The solution that we have come up with is to google and search thoroughly to understand this sector and for some of the attributes that cannot be found online, our team decided to remove the column by cleaning the data.

FID	IPEDSID	NAME	ADDRESS	CITY	STATE	TYPE	STATUS	POPULATION	COUNTY	LEVEL	HI_OFFER	DEG_GRANT	LOCALE	PT_ENROLL	FT_ENROLL	TOT_ENROLL	HOUSING	DORM_CAP	TOT_EMPLOY	
7001	107840	Shorter College	604 Locust St	N Little Rock	AR	2	R	70	Pulaski	...	2	3	1	13	24	28	52	2	0	18
7002	112181	Citrus Heights Beauty College	7518 Baird Way	Citrus Heights	CA	3	R	39	Sacramento	...	3	2	2	21	6	24	30	2	0	9
7003	116660	Joe Blasco Makeup Artist Training Center	1670 Hillhurst Avenue	Los Angeles	CA	3	R	35	Los Angeles	...	3	1	2	11	0	24	24	2	0	11
7004	125310	Waynes College of Beauty	1271 North Main Street	Salinas	CA	3	R	43	Monterey	...	3	2	2	12	18	16	34	2	0	9
7005	164368	Hult International Business School	1 Education Street	Cambridge	MA	2	R	2386	Middlesex	...	1	7	1	12	0	2243	2243	2	0	143

The red circle represents one of the unknown attributes.

Besides that, the challenges that our team faced is the workload for our team is not separated properly. Our team consists of only 2 members and puts us at a disadvantage because of the workload. When our team is doing our respective tasks separately, our jupyter notebook becomes a mess, when we try to combine it together. This is because when we try to combine the codes together, some of the variables become the same and the work that has to be done is redundant. Therefore, the solution of our problem is to separate and plan out our team's work in a more efficient and organised way so that the problem would not repeat itself in the future.

Furthermore, there is another big challenge which is that the relationship that we received when we are plotting our graph does not have solid evidence for the solution. For example, for the predictive question, : “ What would happen to the number of enrollment to college or university if the population of the country is high? ”, the population and the total number of enrollment are very far apart from each other and the range is considered to be very high.

In addition, our future proposals firstly is to choose a data set wisely by researching the data set, for example checking the variable and content inside of it. Besides that, in the future, we will improve by learning in depth more about the linear regression model and also apply other supervised learning or unsupervised learning algorithms, for example by applying the method in the predictive question. Other than that, our team will learn and adapt on how to handle our work sufficiently without any problems.

In conclusion, our team has done the analysis of the data set of colleges and universities in the US. We are able to see the relationship of enrollment in college or university in corporate countries and unincorporated countries. Our team concluded that Both unincorporated and incorporated countries have high enrollment rates but unincorporated countries have a slightly higher enrollment rate because of the financial aid provided by the respective countries which causes the increase of financial aid in unincorporated countries.

## 7. References

1. [https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143\\_013697](https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143_013697)
2. [https://nces.ed.gov/programs/edge/docs/EDGE\\_LOCALE17\\_ZCTA\\_FILESPEC.pdf](https://nces.ed.gov/programs/edge/docs/EDGE_LOCALE17_ZCTA_FILESPEC.pdf)
3. <https://nces.ed.gov/ipeds/>
4. <https://journals.sagepub.com/doi/full/10.1177/1035304620962265>
5. <https://nces.ed.gov/programs/coe/indicator/cbc>