**Table of Contents**

[Background of Data analysis 3](#_Toc126145767)

[Summary of data set 3](#_Toc126145768)

[Data pre-processing 4](#_Toc126145769)

[Simple Random Sampling 8](#_Toc126145770)

[Hypothesis testing 11](#_Toc126145771)

[Hypothesis 1 11](#_Toc126145772)

[Hypothesis 2 12](#_Toc126145773)

[Conclusion 12](#_Toc126145774)

[References 13](#_Toc126145775)

**MDA511**

**Mathematical and Statistical Methods**

# Background of Data analysis

This project is to perform data analysis for the Queensland Government Department of Education. By enhancing data analysis, the data set has been collected from Queensland Government. Data set is downloaded from the open portal of Queensland government. The analysis is carried with achievement of Mathematics by students during the term year of 2017. As a data analyst, mathematical and statistical principles has been applied with the collected data set. The report is structured by providing a summary of data set and its representation. Additionally, the description of each field of data set are explained effectively. Initially, the data analysis process like data cleaning has been performed inorder to maintain the accuracy of data for data analysis. Secondly, sampling method has been used to select the random dataset values for an analysis inorder to evaluate academic performance of students in mathematic subject. To show the variation, country of birth and year are considered. Finally, hypothesis analysis and testing is performed to identify potential variation in academic achievement based on country and year level.

# Summary of data set

The data set has been downloaded from open portal

**Link**: <https://www.data.qld.gov.au/dataset/academic-achievement-for-student-s-studying-maths-by-country-of-birth-and-year-level> .

The data set has been created from the student academic achievement report from the school where it contains different subjects like English, Mathematics and science along with the grades. The grades were represented in a common scale for the all the subjects. The data set also includes ID, learning Area, Year Level, Achievement result, Country of Birth Code and Country of Birth Name. The entire data set which is taken for an analysis contains 261,500 records of academic information. The description of each fields are explained below:

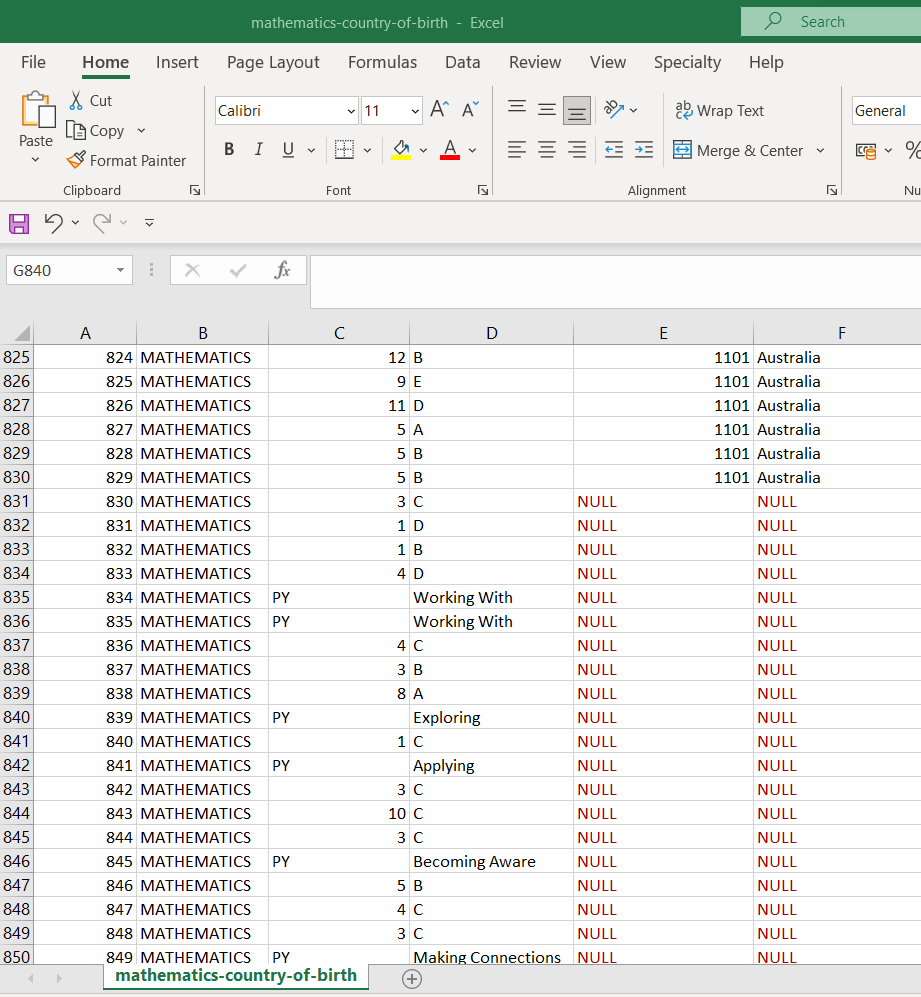
|  |  |  |
| --- | --- | --- |
| **Field Name** | **Description** | **Example** |
| ID | This is the unique field from the first column of data set. All the values in this field are not dependent towards other values. | Values initiate from 1, 2, 3 and so on. |
| LearningArea | Learning area specifies about the result where the student is achieved in particular subject. Here the mathematics subject is selected as a learning area inorder to perform data analysis. | Mathematics, English, Science, etc. |
| YearLevel | The year level indicate about the year where the student is achieved in particular year. This year level will be ranges between year 1 to year 12 but represented in numbers. | Year 1, 3, 4, etc. |
| AchievementResult | This is the student achievement result which is described in a common values. This is described as A, B, C, D and E. For the Prep Year, the achievement result are described as Applying, Making Connections, Working With, Exploring and Becoming Aware. | Grade A, B, C  PY Working With, Making Connection. |
| CountryOfBirthCode | This is the country code where the student belongs in achieving the result. This value is also unique for different country. | 1101 – Australia  2303 – France  3300 – Eastern Europe, nfd  2100 – United Kingdom, Channel Islands and Isle of Man, nfd. |
| CountryOfBirthName | This field belongs to country name where the student belongs in achieving the academic result. | Australia, Italy, Americas, Switzerland, United States of America and so on. |

# Data pre-processing

Data pre-processing is required for the collected data because data contains both missing and null values. The purpose of pre-processing is that it eliminates missing values, errors and if there are any duplicate value in the records [1]. The data analysis follows a data cleaning process which is the first step.

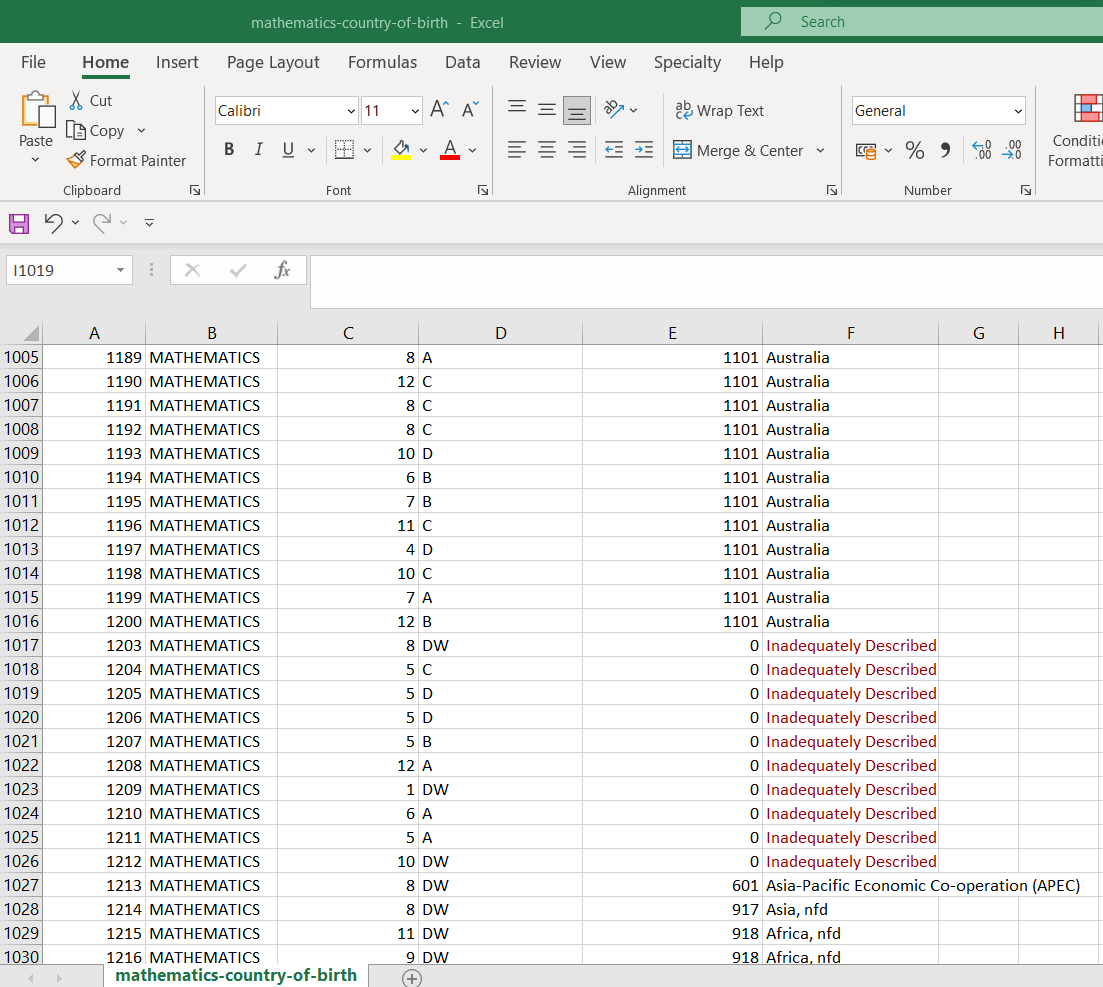
**Dataset cleaning**

**Removing “NULL” values**



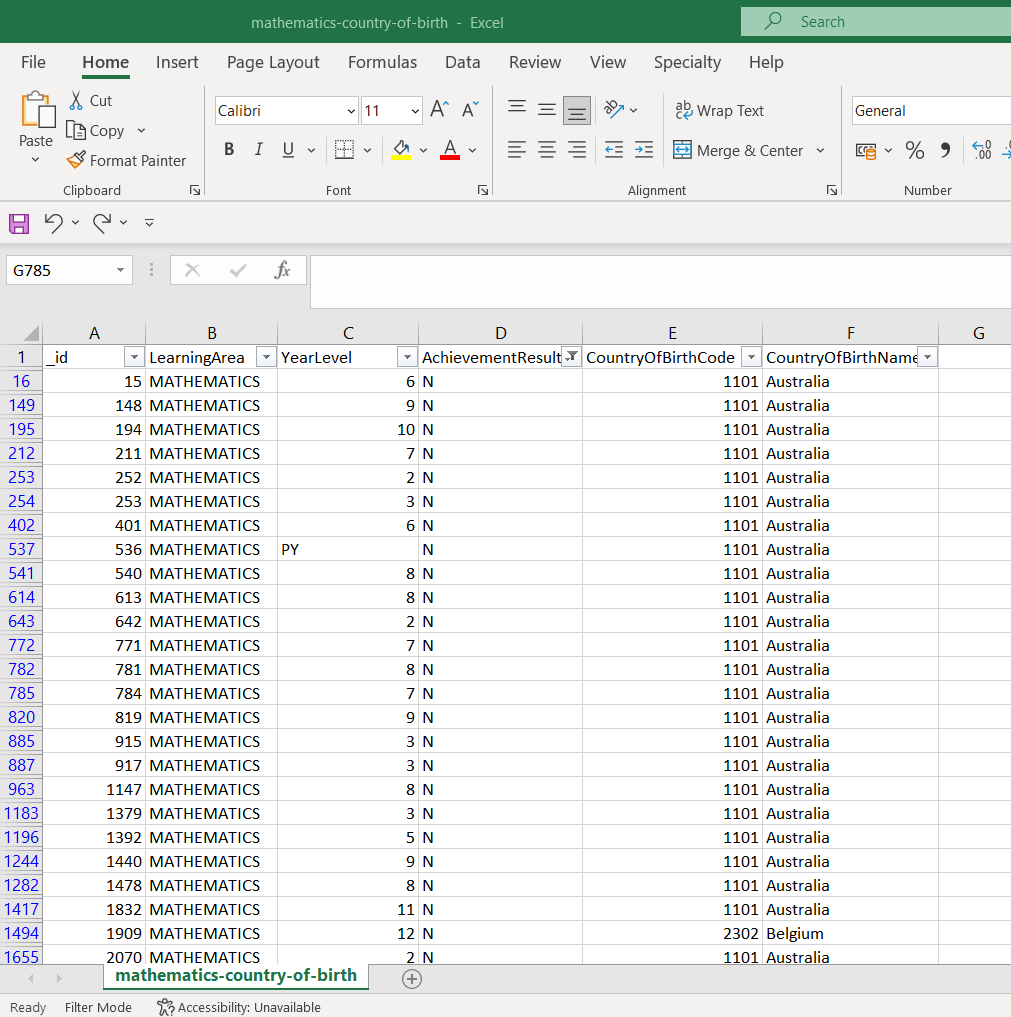
Using the data cleaning process, the NULL values which are available in the dataset is identified and removed successfully.

**Removing “Inadequate described” values**



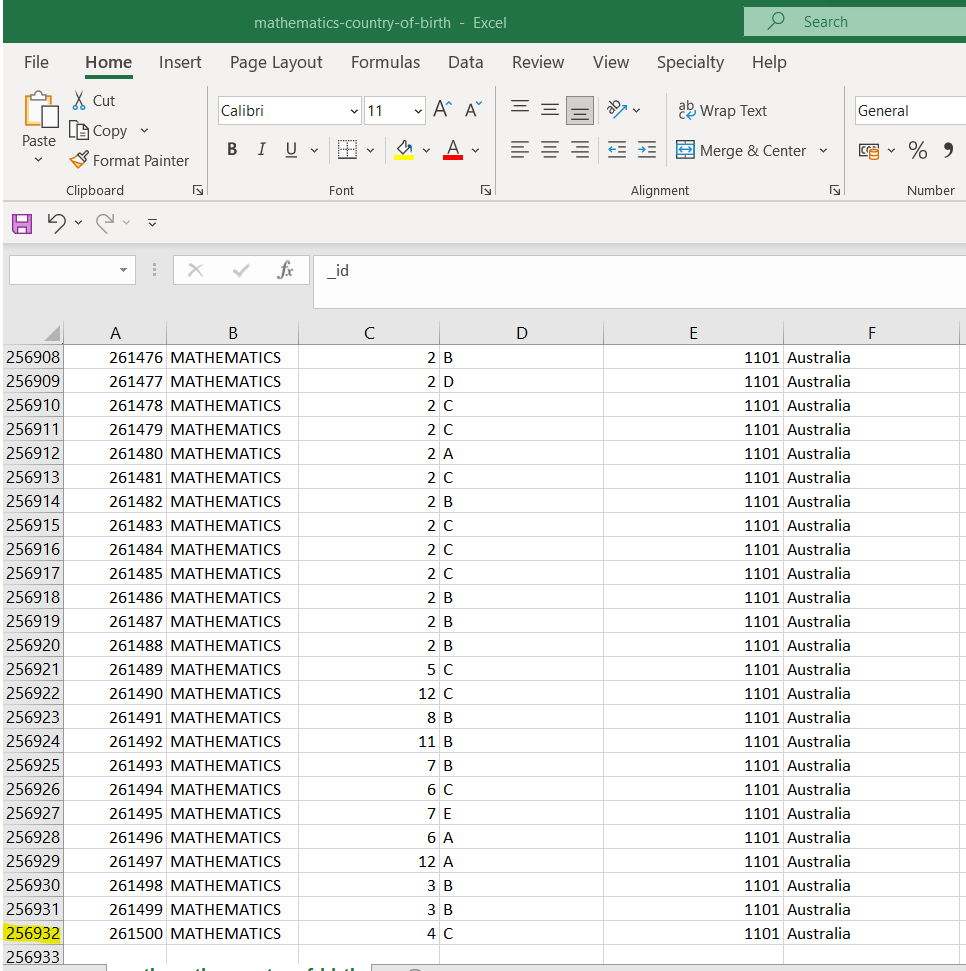
Data set also contains “inadequately Described” values in the CountryOfBirthName field which is highlighted above. These values are also removed from the data inorder to maintain accuracy of evaluation [2].

**Removing “N” values**



Few of the records in the data set include “N” values in the AchievementResult which are eliminated from the entire dataset.

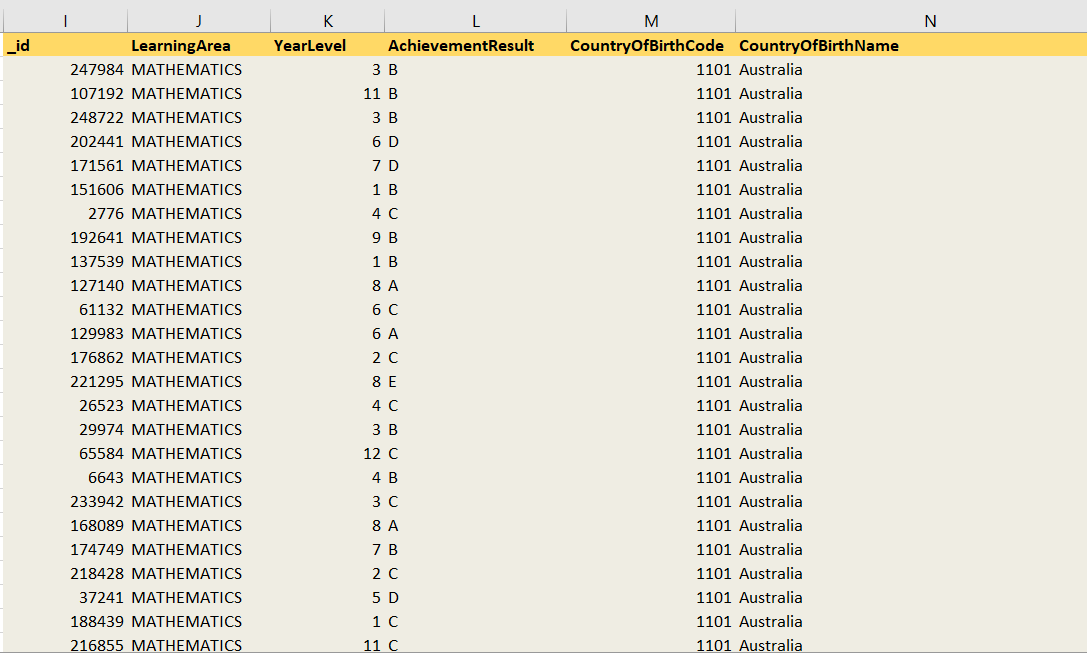
**Total number of instances**



After the successful completion of data cleaning process, error data are removed from the dataset. Finally, total of 256932 records were available for analysis academic achievement in mathematics by the students.

# Simple Random Sampling

Simple random sampling has been selected for Mathematic study analysis because it uses the same size of records for both sample and available data [3]. This technique seems to be simpler and easier to understand because statistical analysis can be performed effectively.



The above screen is the data has been selected randomly from the larger group of dataset. Representation of this selected individually to represent a new dataset from the entire dataset.

**Performing Z-test**

The Z test follows a normal distribution because it is performed based in the data that are collected based on samples. With the help of Z-test, the variation between the samples can be identified to show whether it is same or different.

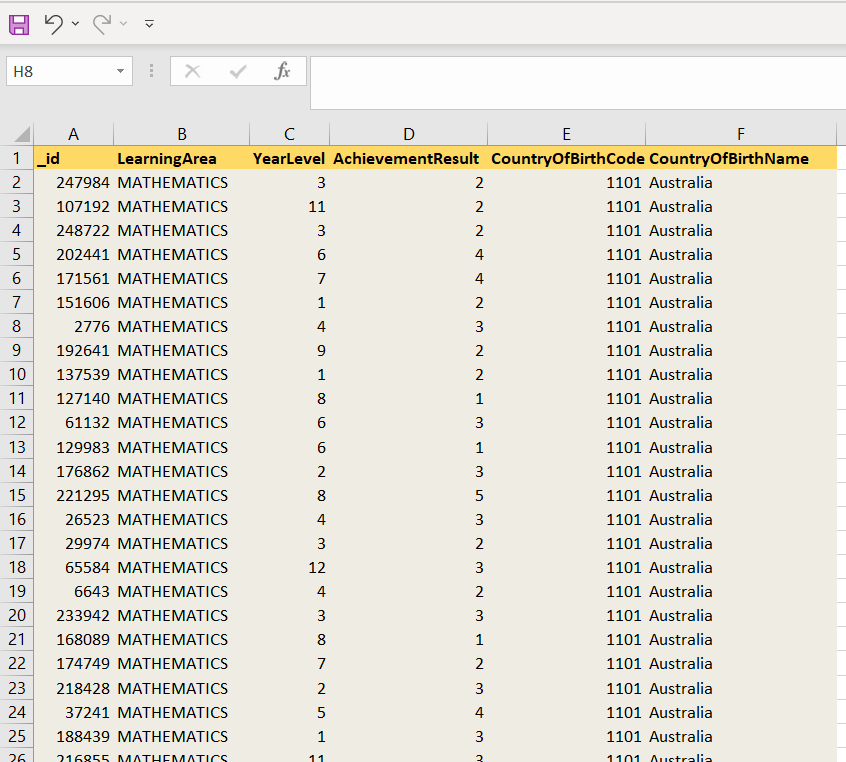
The “Z-test” is used to find out the variation in the academic performance of students studying in different countries and at different year levels. Here, academic performance of students is determined by their achievement result of students in Mathematics subject.

To use the Z-test, the value of AchievementResult is converted to numeric format. Because in Z-test, two input variables must be numbers.

**Conversion of grade to numbers**

Hence, the grades of the students in the AchievementResult are converted as,

|  |  |
| --- | --- |
| **AchievementResult – in Grade** | **AchievementResult – in numbers** |
| A | 1 |
| B | 2 |
| C | 3 |
| D | 4 |
| E | 5 |
| Applying | 1 |
| Making Connections | 2 |
| Working With | 3 |
| Exploring | 4 |
| Becoming Aware | 5 |



After performing Z test of the sampling dataset, the AchievementResult values are changed from grades to numbers.

# Hypothesis testing

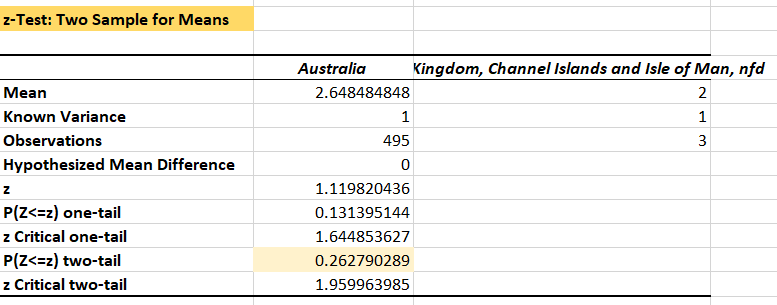
Hypothesis testing is used to identify whether the statistical experimentation result is valid or not. This is based on null and alternative hypothesis but both of them are mutually exclusive [4]. This means either one hypothesis will be true or another one will be false.

## Hypothesis 1

**Question: Whether there is variation between performance of students in two different countries?**

**Null hypothesis:** There is no difference in academic performance between students studying in Australia and the United Kingdom, Channel Islands and Isle of Man, nfd.

**Alternate hypothesis:** There is difference in academic performance between students studying in Australia and the United Kingdom, Channel Islands and Isle of Man, nfd.



Here the two samples were taken for hypothesis testing which are Australia and United Kingdom, Channel Islands and Isle of Man, nfd. The total observation is determined as 495 for Australia where the mean value is calculated as 2.64848 and variance of 1. Initially, the Z value is computed as 1.119820436 and based on z value, the P value for two sample is determined as 0.262790289 [5].

**Result:**

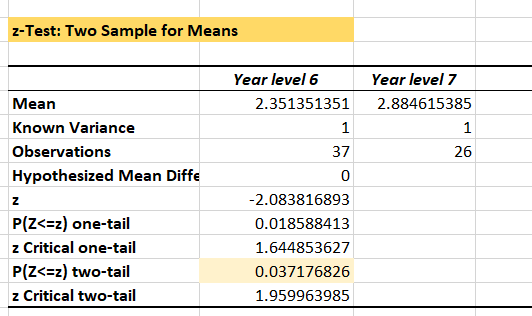
The result show that the p value that is calculated from the data set is 0.262790289 which is greater than the alpha value (0.05). This shows that there is no difference between the academic performance between the students studying in two different countries [6]. So, the null hypothesis is determined as a result through the computed estimation of mean and p values from the dataset.

## Hypothesis 2

**Question: Is there is any significant variation in academic performance during level year 6 and level year 7?**

**Null hypothesis:** There is no difference in academic performance between students studying in Year Level 6 and Year Level 7.

**Alternate hypothesis:** There is a difference in academic performance between students studying in Year Level 6 and Year Level 7.



The mean value is estimated as 2.351351351 for the year 6 based on the sample data of 37 records. Likewise, the year level 7 mean value is evaluated as 2.884615385. For the single tail, the p value is computed as 0.018588413 and for two tail, the value is 0.037176826. Based on the evaluation, AchievementResult of student during the year has the difference in academic performance [7].

**Result**

The result shows that the p value that is determined based on the year level 6 and year level 7 of Australia is 0.0371776826. Here, the p value is smaller than 0.05, hence, it follows the “alternative hypothesis” which shows there is a difference in academic performance of students between the year.

# Conclusion

The report explains about analysing of educational dataset by applying mathematical and statistical principles. The evaluation of dataset has been examined effectively with the fields and its purpose. Before undergoing data analysis, data pre-processing like data cleaning has been successfully made which are reported and explained in detail. Inorder to carry the evaluation of academic performance of students, the sampling technique has used to compute a new dataset where it does not contain any of duplicate or null values. Finally, hypothesis testing is carried with a question related to grade by Country of birth and grade by year level. The calculation shows that hypothesis 1 result in null hypothesis and hypothesis 2 attains alternative hypothesis.

# References

|  |  |
| --- | --- |
| [1] | I. H. Sarker1, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Computer Science,* 2021. |
| [2] | A. H. Yousssoufa Mohamadou, “A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19,” *Applied Intelligence,* p. 3913–3925, 2020. |
| [3] | B. U. a. O. M. Igwe N.O, “COMPARATIVE ANALYSIS OF THE EFFICIENCY OF SIMPLE RANDOM SAMPLING AND STRATIFIED RANDOM SAMPLING TECHNIQUES USING DATA FROM 2006 POPULATION FIGURES OF THE SIX SOUTH-SOUTH STATES OF NIGERIA,” *International Journal of Advanced Research (IJAR),* pp. 1056-1064, 2020. |
| [4] | A. S. Ayotunde Ola Kolawole, “Hypotheses and Hypothesis Testing,” in *Ph.D. Agricultural Economics Seminar*, Ekiti State University, Nigeria, 2017. |
| [5] | M. C. Joaquim Fernando Pinto da Costa, “Statistical Methods with Applications in Data Mining: A Review of the Most Recent Works,” *Mathematics ,* vol. 10, no. 6, p. 993, 2022. |
| [6] | A. B. Zorić, “Applied Statistics: Basic Principles and Application,” *International Journal of Innovation and Economic Development,* vol. 7, no. 3, pp. 27-33, 2021. |
| [7] | A. Saltelli, “A short comment on statistical versus mathematical modelling,” *Nature Communications ,* p. 3870, 2019. |