INDIVIDUAL ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

AICT009-4-2-IDA

INTRODUCTION TO DATA ANALYTICS

UCDF2005(1) ICT(DI)

IDA DOCUMENTATION

Name: Chan Ming Li

TP Number: TP060774

Hand out Date: 09 August 2021

Hand in Date: 15 November 2021

Weightage: 40%

Lecturer Name: Ms. Hema Latha Krishna Nair

Table of Contents

1.0 Introduction

An environment is the surrounding and nature in which the non-living and living things live around. The environment is composed of a variety of essential elements such as air, water, and land. These elements are important to maintain a healthy and balanced nature in an environment. It is undeniable that the environment has been exposed to several threats and causing environmental issues. These environmental issues have enormously increased due to unbalanced nature such as changes of climate, natural disasters, polluted waters, air, land etc. (Vedantu Learn LIVE Online, 2021) . Due to these negative impacts on the environment, organisms have reduced available sources to live such as air, water, and land. In addition, the extinction of biodiversity, lack of minerals, as well as countries would be affected economically. If these problems continue for the following decades, the environment that we live with for years will cease to be exist. Yet we still ignore despite the public awareness, media attention and few of research has done to prevent environmental issues.

As the severity of the environmental issues increases, many problems and challenges have emerged and becoming serious which has a huge impact on our lives and environment. One of the problems caused by environmental issues is human health as it increases the risk of getting diseases such as lung cancer, asthma and many more. Dirty and poisonous gas emissions which causes air pollution have a similar effect as smoking which will destroy our respiratory system (World Health Organization, 2021). Besides, environmental issues threaten the habitat of flora and fauna by habitat fragmentation and habitat degradation. For example, a forest which is mainly the habitat of flora and fauna is destroyed due to deforestation caused by agricultural expansion and so on. Moreover, we found that environmental issues have a huge impact on the economy although some companies have started sustainable development which will mitigate the effects of pollution. As the availability of non-renewable resources is declining, production or business which highly dependent on it is affected (Rinkesh, n.d.).

2.0 Business Goal & Objective

2.1 Business Goal

The business goal of our group is to conduct analysis on discovering analytic solution that is suitable to solves and handles impacts of global environmental issues effectively. Whether individually or as a group, we designed some solution that should reduce the impacts of global environment by referencing real world solutions that has been implemented. The solution that we identified in reducing impacts of global environment issues is enhancing the health of human by identifying the health issues, controlling the habitat index to be in the normal level and develop plans for sustainable environment and economy growth. We are planning to achieve this solution using different strategies such as predictive modelling to predict and forecast future trends or performance in different aspects. Besides, we will use dashboard solution to visualize significant impacts of environmental issues. By using these strategies, BI report and OLAP is able to carry out activity of discovering patterns and trends in datasets collected.

2.2 Objectives

1. To evaluate the effect of global environment issues on human health based on locations and utilize a clustering model to reduce environmental issue and health disease.
2. To identify the loss of forest due to environmental issue and implement a statistical model to determine the relationship between the possible drivers of deforestation and loss of forest based on years and locations. **(Chan Ming Li)**
3. To analyze how economic growth contribute to environmental issues and develop predictive model to support plan for sustainable environment and economy.
4. To study the loss of population affected by environmental issues and design a predictive model based on different categories in a type of environmental issues based on years.

2.3 Scopes

2.3.1 Scope 1: OLAP Dashboard on The Change of Forest Loss Impacted by Deforestation

Deliverables: To analyse and visualize the change of forest loss impacted by deforestation globally based on years by using Analytical Dashboard (OLAP).

- OLAP is known as Online Analytical Processing which is a technology for data discovery includes complex analytical calculations, trend analysis and more. OLAP performs multidimensional analysis which allows extraction of information can be done in a consistence, interactive and fast way. For instance, the severity of the forest loss can be measured and determined by processing the change of forest loss across the years and countries with other factors.
- Layman or other users will have a better understanding on the analysis done as OLAP provides a clear, user-friendly and interactive dashboard which contains information from different aspects to them. Hence, they are able to understand the current situation or the severity of forest loss which leads them to several compliance actions.
- There are several basic operations like roll up, drill down, slice and dice and pivot can be performed by implementing OLAP. All these operations will help users to gain detailed information from different perspectives or dimensions. For example, the regions of forest loss can be drilled down into different countries by fragmenting the data into smaller parts.

2.3.2 Scope 2: Predictive Modelling on Forest Loss and Possible Drivers of Deforestation

Deliverables: To develop a linear regression model to determine the relationship between the forest loss and the possible drivers of deforestation in order to implement suitable conservation planning.

- Predictive modelling which is also known as predictive analytics is commonly utilized to predict or forecast the future events or possible outcomes by analysing the patterns from the related input data. For instance, the forest loss can be predicted by analysing the relationship between forest loss and the possible drivers of deforestation.
- The predictive results provided by the model can be used to make fast and critical decisions by users as the results provide an overview on the most likely happen scenario or outcome.
- Based on the results from the prediction model, public or different organisation can provide and implement different solutions to avoid or mitigate the forest loss which is highly affecting our environment.

3.0 Data Analytics Life Cycle & Methodologies

3.1 Fayyad's KDD Methodology

There are many methodologies often implemented for data analytics and data mining such as Fayyad's KDD, Crisp-DM, SEMMA and many more. Fayyad's knowledge discovery in databases (KDD) is a process or procedure of discovering knowledge that find and transform patterns from a large amount of data by implementing different data mining techniques and algorithms. It is often utilized by researchers from different fields such as machine learning, databases, data visualization, artificial intelligence and many more. KDD methodology plays an important role in data mining field as the main backbone of its process is data mining which allows recognizing and exploring understandable patterns from a vast amount of data (Usama Fayyad, 1996).

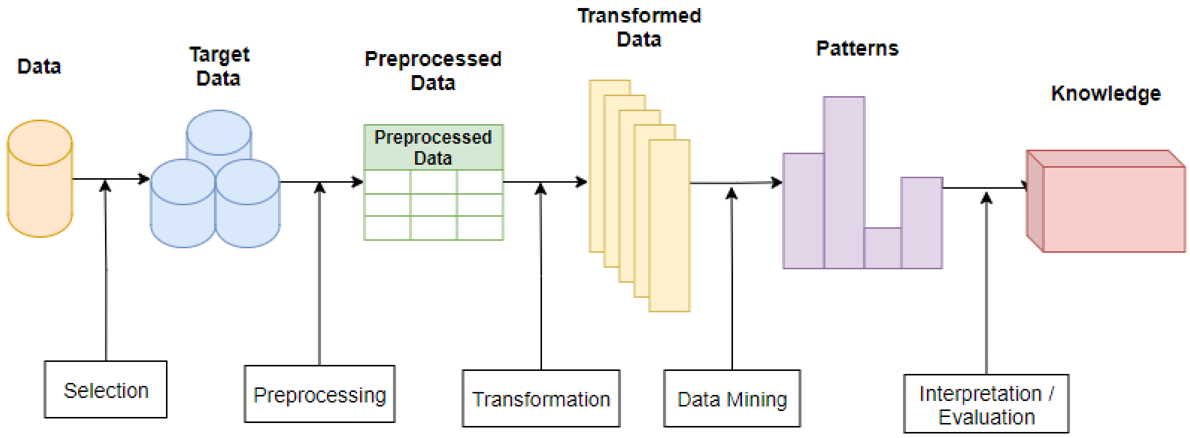


Figure 3(a): Fayyad's KDD Process Model

3.2 Fayyad's KDD Process

Fayyad's KDD is an iterative and interactive process which can be different in a range of 5 to 9 steps based on different situation and the decisions made by users. Firstly, the initial phase in the process is fundamental as problem statement is needed to be identify. The user should have some relevant knowledge and understand the application domain in order to avoid false interpretations and identify their goal (Sharma, 2020). For instance, the understanding of our group domain and identifying the objectives and scopes are conducted in 2.0 part.

After that, data selection takes places as the second phase in the Fayyad's KDD process where includes discovering or finding the relevant, accessible, and critical dataset. As the objectives and goal are identified earlier, users will have a better understanding on which kind of dataset and variables are needed for the data models (Usama Fayyad, 1996). For example, the resources required for the project such as datasets are determined and selected from different sources which links are provided in 4.0 part. Besides, the usability of the dataset or attributes are determined and selected after exploring and reviewing in 5.1 part to ensure all datasets or attributes are related and helpful based on the objectives and scopes.

Furthermore, data pre-processing and cleaning is carried out as the third phase in the methodology which includes some basic operations to improve the data quality and integrity (Sharma, 2020). Strategies on handling missing or null values as well as noisy data are determined to improve the reliability of the data chosen. For instance, some cleaning operations like removing null values are conducted in 5.2 part based on the necessity of the following analysis and modelling. There is a data cleaning report includes the steps taken in the data cleaning process shown in 5.2 part.

Moreover, the fourth phase of the process is data transformation which prepares data for data mining and modelling. The data are integrated and consolidated by merging datasets, converting data types and many more according to the usage of following analysis and modelling (Sharma, 2020). For example, the datasets are merged based on the matching columns and several columns are aggregated which helps to build OLAP dashboard and Linear Regression model much easier and effective in 5.3 part.

In addition, data mining is the following phase in the process which is the root phase of the Fayyad's KDD methodology. Different approaches and algorithms are executed in order to extract and reveal information by discovering understandable patterns and trends from the cleaned and transformed data. For instance, a correlation plot is built in order to determine the correlation coefficient of different attributes for linear regression model.

Besides, interpretation or evaluation will be carried out in the sixth phase of the process which interprets the mined patterns or trends obtained from the previous phase. These patterns will be visualized in different forms such as bar graphs, line graphs, pie charts and many more so that the patterns or trends can be clearly viewed by users. For instance, OLAP dashboard that includes different visualization like line graphs, area charts and many more is built and evaluated in 7.0 part.

Lastly, the final phase of Fayyad's KDD process is using the knowledge discovered from the previous phases. The users are able to measures the impact and make decisions based on the knowledge gained. The knowledge discovered will be documented in a report so that the users can better evaluate the output (Sharma, 2020).

3.3 The Advantages & Drawbacks of Fayyad's KDD

There are some advantages and drawbacks when Fayyad's KDD methodology is implemented. Firstly, one of the advantages of KDD is that it could be used for forecasting purposes. For examples, market forecasting which is predicting the customer behaviour and consumer trends by using KDD methodology. Furthermore, the process in KDD is iterative which increases impact of the data by refining the data and gain new knowledge from it. The knowledge that gained from any phase of process can be applied back to the beginning of the process (Data Science Process Alliance, 2021).

However, there are some drawbacks in the process of Fayyad's KDD too. First of all, time consuming is one of the drawbacks of the methodology. As the methodology has an iterative process, the knowledge gained is applied back to the process in the next iteration which leads to time consuming. While returning the process, there is a possible that additional data will be used which needed to run through the whole process again. Hence, the process is considered as one of the time-consuming processes. Besides, the security and privacy of the process is vulnerable which is definitely a drawback of the methodology. When there is a need of customer information for analysis, the data is needed to be collect as much as possible in order to gain more knowledge from it. Therefore, there will be a huge calamity if there is any malicious attacks or data breaches occurs (Data Science Process Alliance, 2021).

4.0 Data Understanding

Dataset Name: Agricultural Area Dataset

Data Source: https://data.worldbank.org/indicator/AG.LND.AGRI.K2

- There are 65 columns and 266 rows about the agricultural land area in square kilometer for different countries from 1960 to 2020.
- The columns that exist in the dataset is "Country Name", "Country Code", "Indicator Name", "Indicator Code", Year columns (1960 – 2020).
- All data in the dataset is in a text data type.
- There are errors like null values, unstandardized, unmatched data types, blank columns, blank rows.

Column	Problems Identified
Year (1960 - 2020)	Null values, unstandardized, unmatched data types,

Dataset Name: Land Area Dataset

Data Source: https://data.worldbank.org/indicator/AG.LND.TOTL.K2

- There are 65 columns and 266 rows about the land area of country globally from 1960 to 2020.
- The columns that exist in the dataset is "Country Name", "Country Code", "Indicator Name", "Indicator Code", Year columns (1960 – 2020).
- All data in the dataset is in a text data type.
- There are errors like null values, unstandardized, unmatched data types, duplicate columns, blank rows.

Column	Problems Identified
Year (1960 - 2020)	Null values, unstandardized, unmatched data types,
Country Name	Duplicate columns

Dataset Name: Forest Area Dataset

Data Source: https://data.worldbank.org/indicator/AG.LND.FRST.K2

- There are 65 columns and 266 rows about the forest area in square kilometers for different countries and years.

- The columns that exist in the dataset is “Country Name”, “Country Code”, “Indicator Name”, “Indicator Code”, Year columns (1960 – 2020).
- All data in the dataset is in a text data type.
- There are errors like null values, unstandardized, unmatched data types, blank columns, blank rows.

Column	Problems Identified
Year (1960 - 2020)	Null values, unstandardized, unmatched data types,

Dataset Name: Annual Deforestation Dataset

Data Source: https://ourworldindata.org/grapher/annual-deforestation?tab=table

- There are 4 columns and 495 rows about the global deforestation rate annually for certain years which is 1990, 2000, 2020 and 2015.
- The columns that exist in the dataset is “Country”, “Code”, “Year”, “Deforestation”, Year columns (1990, 2000, 2020, 2015).
- All data in the dataset is in a text data type.
- There are errors like null values, unstandardized, unmatched data types, unmatched unit of measurement.

Column	Problems Identified
Code	Null values
Deforestation	Null values, unstandardized, unmatched data types, unmatched unit of measurement

Dataset Name: Forest Loss Dataset

Data Source: https://gfw.global/3kAcKgX

- There are 4 columns and 3968 rows about forest loss for different countries from 2001 to 2020.
- The columns that exist in the dataset is “iso”, “umd_tree_cover_loss__year”, “umd_tree_cover_loss__ha”, “gfw_gross_emissions_co2e_all_gases__Mg”.
- All data in the dataset is in a text data type.
- There are errors like null values, unstandardized, unmatched data types, unmatched unit of measurement.

Column	**Problems Identified**
umd_tree_cover_loss__year	Null values
umd_tree_cover_loss__ha	Null values, unmatched data type, unmatched unit of measurement

Dataset Name: Global Forest Loss by Dominant Drivers

Data Source: https://gfw.global/3Dr9WtK

- There are 3 columns and 252 rows about the forest loss caused by dominant drivers globally.
- The columns that exist in the dataset is "tsc_tree_cover_loss_drivers__type", "umd_tree_cover_loss__year", "umd_tree_cover_loss__ha", "gfw_gross_emissions_co2e_all_gases__Mg"
- All data in the dataset is in a text data type.
- It will be used to replace the missing countries name from other datasets with the countries code as a matching column.
- The columns in the dataset are unstandardized.

Dataset Name: iso_metadata

Data Source: https://gfw.global/3kAcKgX

- There are 2 columns and 252 rows about countries and their countries codes.
- The columns that exist in the dataset is "name", "iso"
- All data in the dataset is in a text data type.
- It will be used to replace the missing countries name from other datasets with the countries code as a matching column.

Dataset Name: Countries Data

Data Source: https://www.kaggle.com/fernandol/countries-of-the-world

- There are 20 columns and 228 rows about the details of different countries.
- Example columns that exist in the dataset is "Country", "Region", "Population", "Area".
- All data in the dataset is in a text data type.
- The countries and its region is required for the analysis as the other columns are not related to the objectives and scopes.

5.0 Data Preparation

Data preparation is a process that includes data cleaning, data transforming and data consolidating which is commonly utilized in business intelligence (BI), machine learning, predictive analytics and more. Data preparation is essential for data analysis or modelling as it helps to ensure data integrity, data quality and the accuracy of the analysis or modelling. The data preparation process is conducted in Power BI which is a user-friendly analytic service or platform provided by Microsoft.

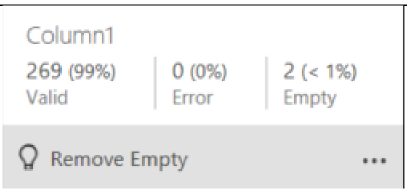
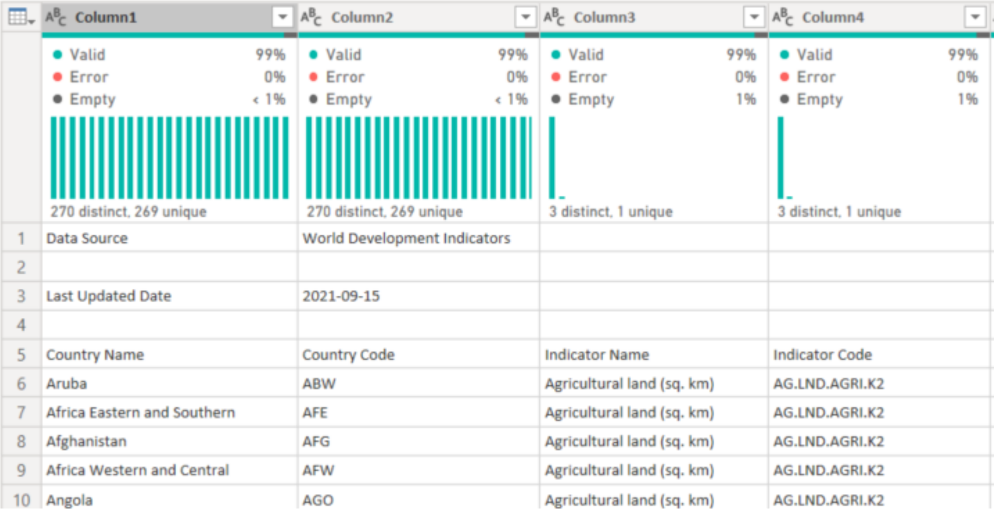
5.1 Data Selection

After exploring and discovering the datasets above, there are a total of 228 columns to be selected for the analysis and modelling. The variables which is best related to the objectives and scopes will be selected to remove unused variables and increase the effectiveness of the analysis. The columns that are needed for the further analysis are “countries name”, “countries code”, “year”, “land area”, “agricultural land area”, “forest area”, “annual deforestation”, and “forest loss”. The other columns that are not selected or unused will be deleted or removed in the data cleaning process. As they are datasets from different sources, the columns from different datasets might have a different unit of measurement. The vary of unit of measurement will causes inaccurate analysis result and reduce the data quality. Therefore, the unit of measurement that is planned to utilize in the analysis and modelling is square kilometer (sq.km). The unit of measurement of certain columns such as “Annual Deforestation” will be changed and converted to square kilometer (sq.km) in the data transformation process. In addition, some aggregated columns like “average agricultural area” are required for the analysis and the columns will be calculated in the data transformation process.

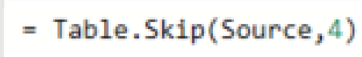
5.2 Data Cleaning

The data cleaning and transformation process will be repeated to some selected datasets which contains the same problem. Therefore, the data cleaning process will be documented once so that there will no repetitive and redundant steps in the documentation. Besides, datasets that have involved in certain steps will be recorded in the dataset column and the description will be provided in the table. The illustration from an example dataset will be provided in the table.

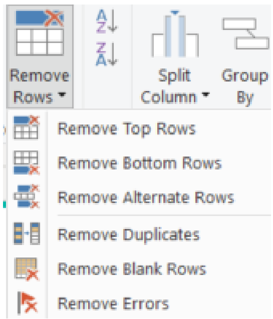
5.2.1 Delete Unnecessary Rows

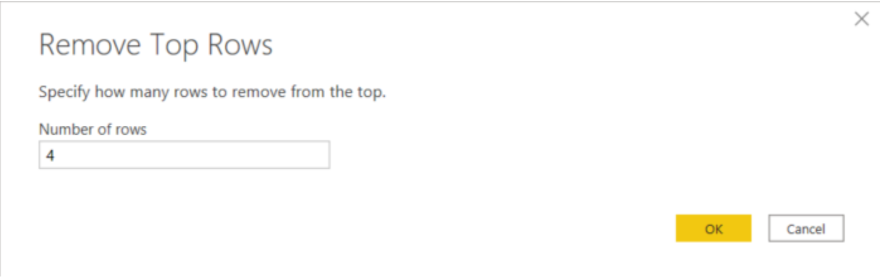
<table>
<tr><td>Dataset</td><td>Agriculture Land, Land Area, Forest Area</td></tr>
<tr><td>Before Cleaning</td><td>

Column1
269 (99%) Valid | 0 (0%) Error | 2 (&lt; 1%) Empty
Remove Empty ...

<table>
<tr><th></th><th>Column1</th><th>Column2</th><th>Column3</th><th>Column4</th></tr>
<tr><td></td><td>Valid 99%
Error 0%
Empty &lt; 1%
270 distinct, 269 unique</td><td>Valid 99%
Error 0%
Empty &lt; 1%
270 distinct, 269 unique</td><td>Valid 99%
Error 0%
Empty 1%
3 distinct, 1 unique</td><td>Valid 99%
Error 0%
Empty 1%
3 distinct, 1 unique</td></tr>
<tr><td>1</td><td>Data Source</td><td>World Development Indicators</td><td></td><td></td></tr>
<tr><td>2</td><td></td><td></td><td></td><td></td></tr>
<tr><td>3</td><td>Last Updated Date</td><td>2021-09-15</td><td></td><td></td></tr>
<tr><td>4</td><td></td><td></td><td></td><td></td></tr>
<tr><td>5</td><td>Country Name</td><td>Country Code</td><td>Indicator Name</td><td>Indicator Code</td></tr>
<tr><td>6</td><td>Aruba</td><td>ABW</td><td>Agricultural land (sq. km)</td><td>AG.LND.AGRI.K2</td></tr>
<tr><td>7</td><td>Africa Eastern and Southern</td><td>AFE</td><td>Agricultural land (sq. km)</td><td>AG.LND.AGRI.K2</td></tr>
<tr><td>8</td><td>Afghanistan</td><td>AFG</td><td>Agricultural land (sq. km)</td><td>AG.LND.AGRI.K2</td></tr>
<tr><td>9</td><td>Africa Western and Central</td><td>AFW</td><td>Agricultural land (sq. km)</td><td>AG.LND.AGRI.K2</td></tr>
<tr><td>10</td><td>Angola</td><td>AGO</td><td>Agricultural land (sq. km)</td><td>AG.LND.AGRI.K2</td></tr>
</table>
</td></tr>
<tr><td>Delete unnecessary rows</td><td>

- Delete the first four rows of the dataset which is unnecessary for our findings by using the remove top rows function in Power BI.

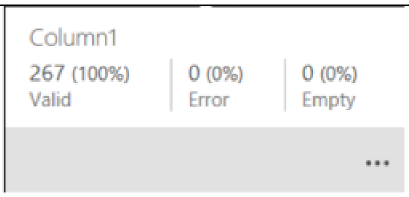
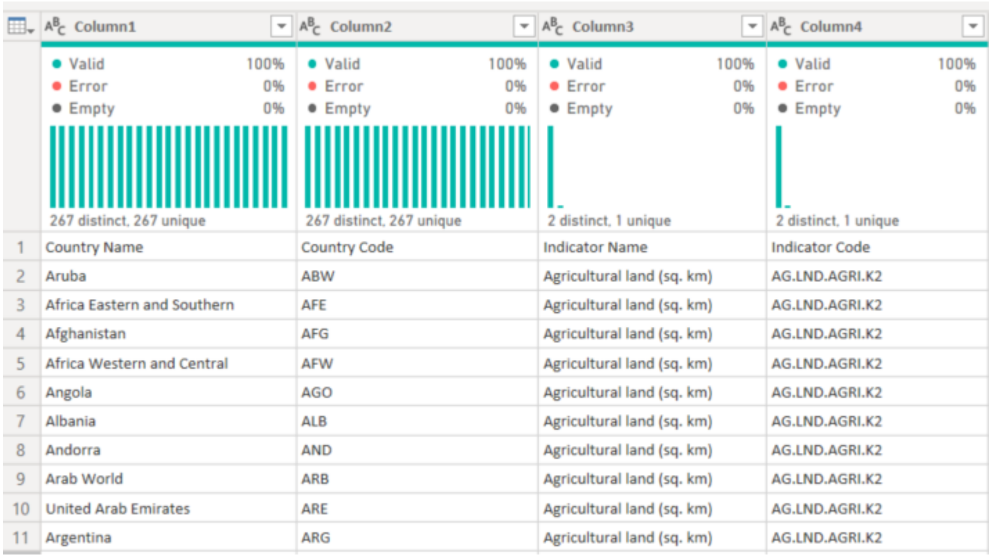


```
= Table.Skip(Source,4)
```

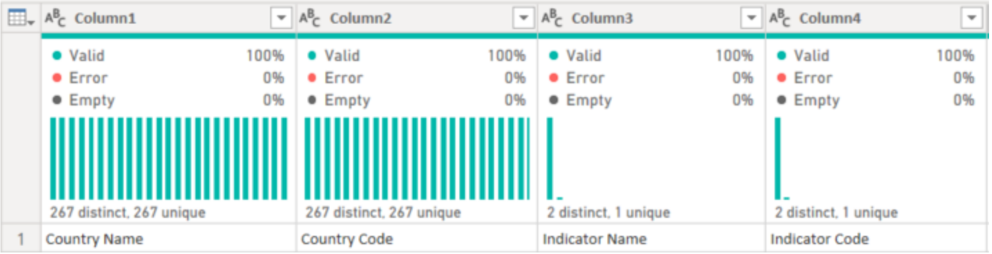

Remove Rows
Split Column
Group By
Remove Top Rows
Remove Bottom Rows
Remove Alternate Rows
Remove Duplicates
Remove Blank Rows
Remove Errors

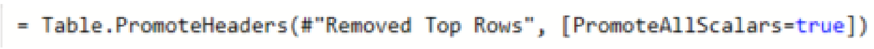

Remove Top Rows
Specify how many rows to remove from the top.
Number of rows: 4

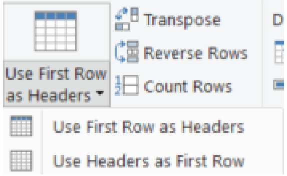
</td></tr>
<tr><td>Explanation</td><td>The first four rows of the dataset are needed to delete as they are just dataset title and last updated date. Hence, deleting the rows allows the dataset to be</td></tr>
</table>

<table>
<tr><td></td><td>more standardise and does not cause much significant to the dataset for analysis.</td></tr>
<tr><td>Results</td><td>

<table>
<tr><th colspan="3">Column1</th></tr>
<tr><td>267 (100%)
Valid</td><td>0 (0%)
Error</td><td>0 (0%)
Empty</td></tr>
</table>

<table>
<tr><th></th><th>Column1</th><th>Column2</th><th>Column3</th><th>Column4</th></tr>
<tr><td></td><td>● Valid 100%
● Error 0%
● Empty 0%</td><td>● Valid 100%
● Error 0%
● Empty 0%</td><td>● Valid 100%
● Error 0%
● Empty 0%</td><td>● Valid 100%
● Error 0%
● Empty 0%</td></tr>
<tr><td></td><td>267 distinct, 267 unique</td><td>267 distinct, 267 unique</td><td>2 distinct, 1 unique</td><td>2 distinct, 1 unique</td></tr>
<tr><td>1</td><td>Country Name</td><td>Country Code</td><td>Indicator Name</td><td>Indicator Code</td></tr>
<tr><td>2</td><td>Aruba</td><td>ABW</td><td>Agricultural land (sq. km)</td><td>AG.LND.AGRI.K2</td></tr>
<tr><td>3</td><td>Africa Eastern and Southern</td><td>AFE</td><td>Agricultural land (sq. km)</td><td>AG.LND.AGRI.K2</td></tr>
<tr><td>4</td><td>Afghanistan</td><td>AFG</td><td>Agricultural land (sq. km)</td><td>AG.LND.AGRI.K2</td></tr>
<tr><td>5</td><td>Africa Western and Central</td><td>AFW</td><td>Agricultural land (sq. km)</td><td>AG.LND.AGRI.K2</td></tr>
<tr><td>6</td><td>Angola</td><td>AGO</td><td>Agricultural land (sq. km)</td><td>AG.LND.AGRI.K2</td></tr>
<tr><td>7</td><td>Albania</td><td>ALB</td><td>Agricultural land (sq. km)</td><td>AG.LND.AGRI.K2</td></tr>
<tr><td>8</td><td>Andorra</td><td>AND</td><td>Agricultural land (sq. km)</td><td>AG.LND.AGRI.K2</td></tr>
<tr><td>9</td><td>Arab World</td><td>ARB</td><td>Agricultural land (sq. km)</td><td>AG.LND.AGRI.K2</td></tr>
<tr><td>10</td><td>United Arab Emirates</td><td>ARE</td><td>Agricultural land (sq. km)</td><td>AG.LND.AGRI.K2</td></tr>
<tr><td>11</td><td>Argentina</td><td>ARG</td><td>Agricultural land (sq. km)</td><td>AG.LND.AGRI.K2</td></tr>
</table>
</td></tr>
</table>

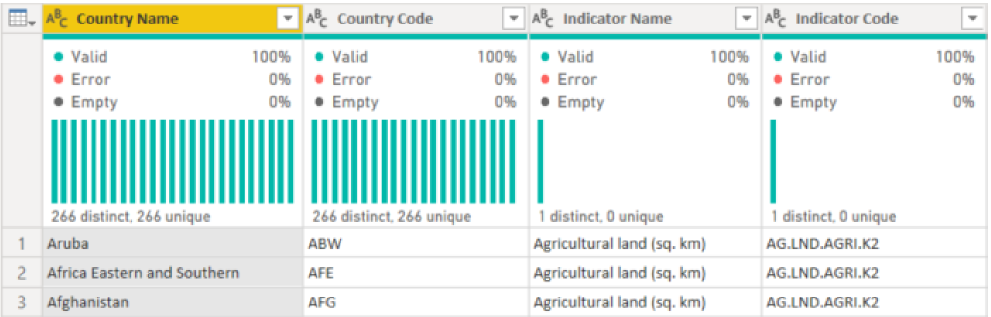
5.2.2 Use First Row as Header

<table>
<tr><td>Dataset</td><td>Agriculture Land, Land Area, Forest Area,</td></tr>
<tr><td>Before Cleaning</td><td>Example from Agricultural Land Dataset:

<table>
<tr><th></th><th>Column1</th><th>Column2</th><th>Column3</th><th>Column4</th></tr>
<tr><td></td><td>● Valid 100%
● Error 0%
● Empty 0%</td><td>● Valid 100%
● Error 0%
● Empty 0%</td><td>● Valid 100%
● Error 0%
● Empty 0%</td><td>● Valid 100%
● Error 0%
● Empty 0%</td></tr>
<tr><td></td><td>267 distinct, 267 unique</td><td>267 distinct, 267 unique</td><td>2 distinct, 1 unique</td><td>2 distinct, 1 unique</td></tr>
<tr><td>1</td><td>Country Name</td><td>Country Code</td><td>Indicator Name</td><td>Indicator Code</td></tr>
</table>
</td></tr>
<tr><td>Use first row as header</td><td>The first row of the dataset is used as the header by using the function provided.

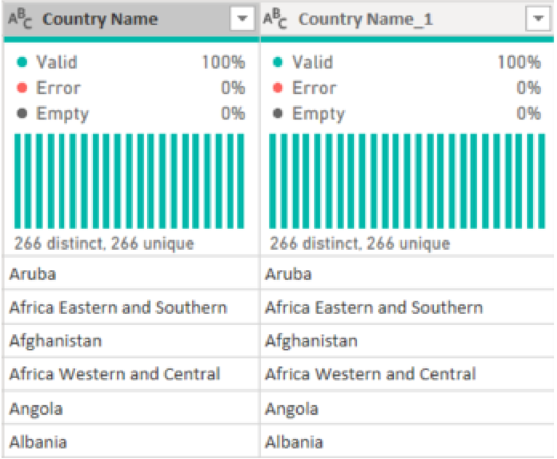
Example from Agricultural Land Dataset:

<pre>= Table.PromoteHeaders(#"Removed Top Rows", [PromoteAllScalars=true])</pre>
</td></tr>
</table>

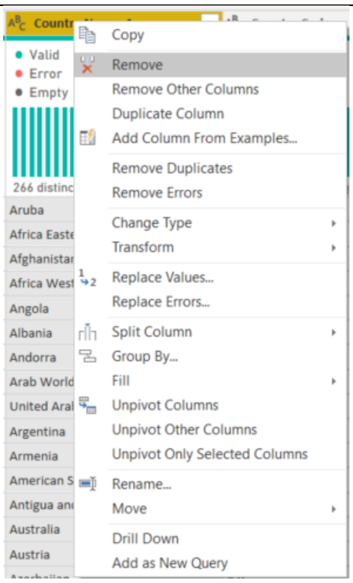
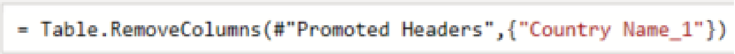
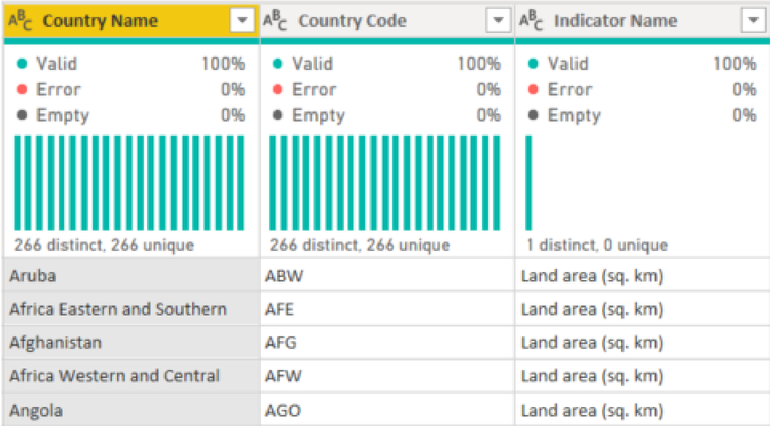
<table>
<tr><td></td><td>

Use First Row as Headers
Transpose
Reverse Rows
Count Rows
Use First Row as Headers
Use Headers as First Row

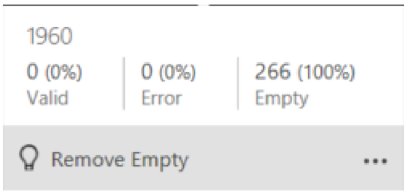
</td></tr>
<tr><td>Explanation</td><td>The header for every column is needed as it will provides a brief message on what is the column about. The data is organized and can be read easily when the header is identified and added.</td></tr>
<tr><td>Results</td><td>Example from Agricultural Land Dataset:

<table>
<tr><th></th><th>Country Name</th><th>Country Code</th><th>Indicator Name</th><th>Indicator Code</th></tr>
<tr><td></td><td>Valid 100%
Error 0%
Empty 0%
266 distinct, 266 unique</td><td>Valid 100%
Error 0%
Empty 0%
266 distinct, 266 unique</td><td>Valid 100%
Error 0%
Empty 0%
1 distinct, 0 unique</td><td>Valid 100%
Error 0%
Empty 0%
1 distinct, 0 unique</td></tr>
<tr><td>1</td><td>Aruba</td><td>ABW</td><td>Agricultural land (sq. km)</td><td>AG.LND.AGRI.K2</td></tr>
<tr><td>2</td><td>Africa Eastern and Southern</td><td>AFE</td><td>Agricultural land (sq. km)</td><td>AG.LND.AGRI.K2</td></tr>
<tr><td>3</td><td>Afghanistan</td><td>AFG</td><td>Agricultural land (sq. km)</td><td>AG.LND.AGRI.K2</td></tr>
</table></td></tr>
</table>

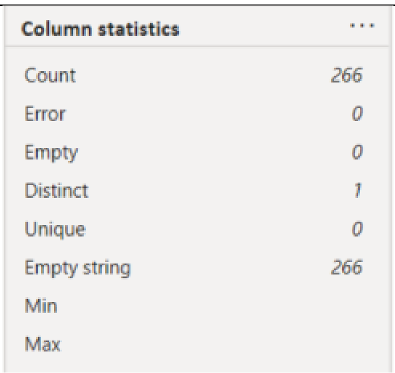
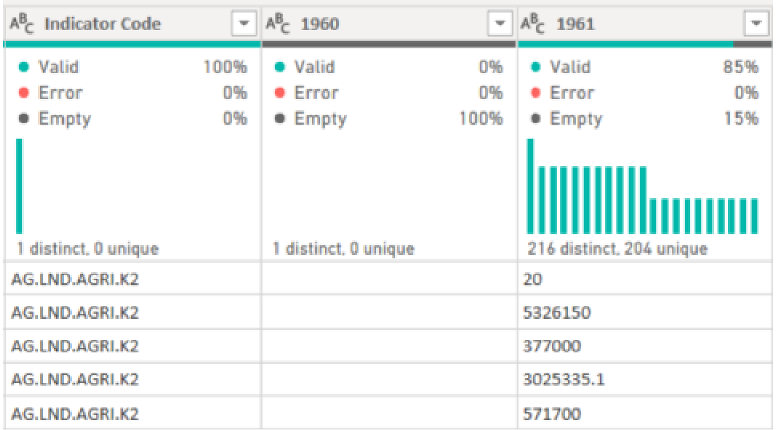
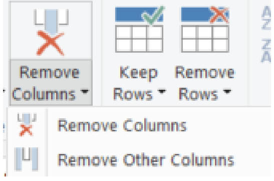
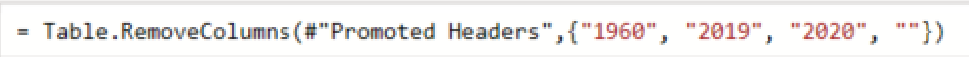
5.2.3 Remove Duplicate Columns

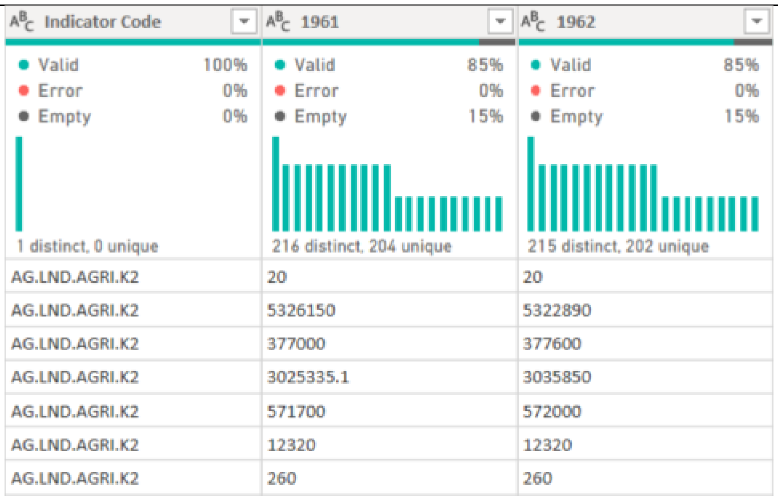
<table>
<tr><td>Dataset</td><td>Land Area</td></tr>
<tr><td>Before Cleaning</td><td>Example from Land Area Dataset:

<table>
<tr><th>Country Name</th><th>Country Name_1</th></tr>
<tr><td>Valid 100%
Error 0%
Empty 0%
266 distinct, 266 unique</td><td>Valid 100%
Error 0%
Empty 0%
266 distinct, 266 unique</td></tr>
<tr><td>Aruba</td><td>Aruba</td></tr>
<tr><td>Africa Eastern and Southern</td><td>Africa Eastern and Southern</td></tr>
<tr><td>Afghanistan</td><td>Afghanistan</td></tr>
<tr><td>Africa Western and Central</td><td>Africa Western and Central</td></tr>
<tr><td>Angola</td><td>Angola</td></tr>
<tr><td>Albania</td><td>Albania</td></tr>
</table></td></tr>
<tr><td>Remove Duplicate Columns</td><td>• Remove duplicate columns which is “Country Name_1”</td></tr>
</table>

	 Valid Error Empty Copy Remove Remove Other Columns Duplicate Column Add Column From Examples... Remove Duplicates Remove Errors Change Type Transform Replace Values... Replace Errors... Split Column Group By... Fill Unpivot Columns Unpivot Other Columns Unpivot Only Selected Columns Rename... Move Drill Down Add as New Query  `= Table.RemoveColumns(#"Promoted Headers",{"Country Name_1"})`
Explanation	The duplicated columns are removed as it will provide redundant data to the following analysis and modelling, and no information can be extracted from the column.
Results	 Country Name | Country Code | Indicator Name Valid 100% | Valid 100% | Valid 100% Error 0% | Error 0% | Error 0% Empty 0% | Empty 0% | Empty 0% 266 distinct, 266 unique | 266 distinct, 266 unique | 1 distinct, 0 unique Aruba | ABW | Land area (sq. km) Africa Eastern and Southern | AFE | Land area (sq. km) Afghanistan | AFG | Land area (sq. km) Africa Western and Central | AFW | Land area (sq. km) Angola | AGO | Land area (sq. km)

5.2.4 Delete Null Columns

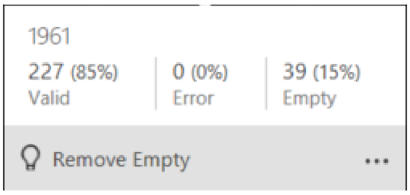
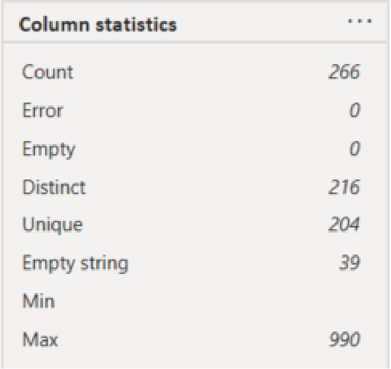
Dataset	Agricultural Land, Land Area, Forest Area
Before Cleaning	Example from Agricultural Land Dataset:  1960 0 (0%) Valid 0 (0%) Error 266 (100%) Empty

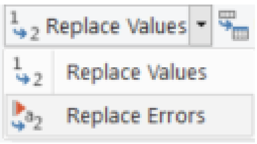
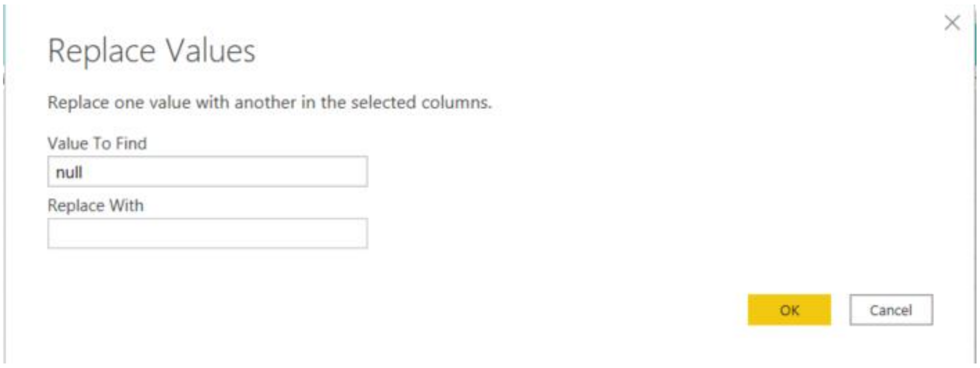
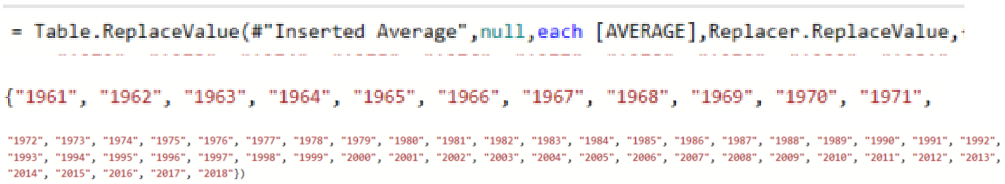
	 Column statistics ... Count 266 Error 0 Empty 0 Distinct 1 Unique 0 Empty string 266 Min Max  Indicator Code: ● Valid 100% ● Error 0% ● Empty 0% — 1 distinct, 0 unique 1960: ● Valid 0% ● Error 0% ● Empty 100% — 1 distinct, 0 unique 1961: ● Valid 85% ● Error 0% ● Empty 15% — 216 distinct, 204 unique AG.LND.AGRI.K2 | | 20 AG.LND.AGRI.K2 | | 5326150 AG.LND.AGRI.K2 | | 377000 AG.LND.AGRI.K2 | | 3025335.1 AG.LND.AGRI.K2 | | 571700
Delete Null Columns	• For Agricultural Land dataset, delete the columns in the dataset that are 100% null (1960, 2019, 2020, a blank column). • For Land Area dataset, delete the columns in the dataset that are 100 % null (1960, a blank column). • For Forest Area dataset, delete the columns in the dataset that are 100 % null (1960 – 1989, a blank column). Example from Agricultural Land Dataset:  Remove Columns Keep Rows Remove Rows Remove Columns Remove Other Columns  `= Table.RemoveColumns(#"Promoted Headers",{"1960", "2019", "2020", ""})`
Explanation	The columns that have 100% missing values are deleted as they will adversely affect the performance of the analysis and do not provide any values to the analysis.
Results	Example from Agricultural Land Dataset:

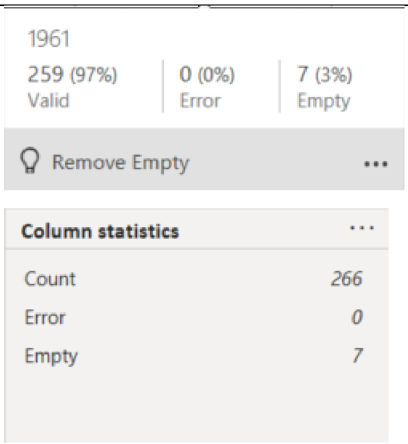
Indicator Code	1961	1962
● Valid 100%	● Valid 85%	● Valid 85%
● Error 0%	● Error 0%	● Error 0%
● Empty 0%	● Empty 15%	● Empty 15%
1 distinct, 0 unique	216 distinct, 204 unique	215 distinct, 202 unique
AG.LND.AGRI.K2	20	20
AG.LND.AGRI.K2	5326150	5322890
AG.LND.AGRI.K2	377000	377600
AG.LND.AGRI.K2	3025335.1	3035850
AG.LND.AGRI.K2	571700	572000
AG.LND.AGRI.K2	12320	12320
AG.LND.AGRI.K2	260	260

5.2.5 Replace Null Values

Dataset	Agriculture Land, Land Area, Forest Area, Annual Deforestation, Forest Loss
Before Cleaning	Example from Agricultural Land Dataset:  1961 227 (85%) Valid | 0 (0%) Error | 39 (15%) Empty Remove Empty ...  Column statistics ... Count 266 Error 0 Empty 0 Distinct 216 Unique 204 Empty string 39 Min Max 990
Replace null values with the average value across the years for different countries	• For Agricultural Land dataset, replace the missing values which consists of 3% - 15% in each selected column (1961 – 2018) with average value of agricultural land area calculated in 6.2.2. • For Land Area dataset, replace the missing values which consists of 1% - 3% in each selected column (1961 – 2020) with average value of land area calculated in 6.2.2. • For Forest Area dataset, replace the missing values which consists of 2% - 3% in each selected column (1990 – 2020) with average value of forest area calculated in 6.2.2.

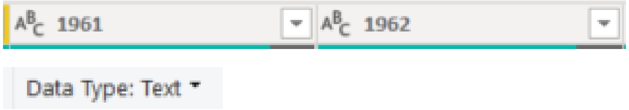
	• For Annual Deforestation dataset, replace the missing values which consists of 3% - 20% in each selected column (1990. 2000, 2010 and 2015) with average value of forest area calculated in 6.2.2 and "Code" column with "N/A".
	• For Forest Loss dataset, replace the missing values which consists of 3% - 10% in each selected column (2001 - 2020) with average value of forest loss calculated in 6.2.2.
	
	Example from Agricultural Land Dataset:
	
	Replace Values
	Replace one value with another in the selected columns.
	Value To Find
	null
	Replace With
	
	`= Table.ReplaceValue(#"Inserted Average",null,each [AVERAGE],Replacer.ReplaceValue,`
	`{"1961", "1962", "1963", "1964", "1965", "1966", "1967", "1968", "1969", "1970", "1971",`
	`"1972", "1973", "1974", "1975", "1976", "1977", "1978", "1979", "1980", "1981", "1982", "1983", "1984", "1985", "1986", "1987", "1988", "1989", "1990", "1991", "1992",`
	`"1993", "1994", "1995", "1996", "1997", "1998", "1999", "2000", "2001", "2002", "2003", "2004", "2005", "2006", "2007", "2008", "2009", "2010", "2011", "2012", "2013",`
	`"2014", "2015", "2016", "2017", "2018"})`

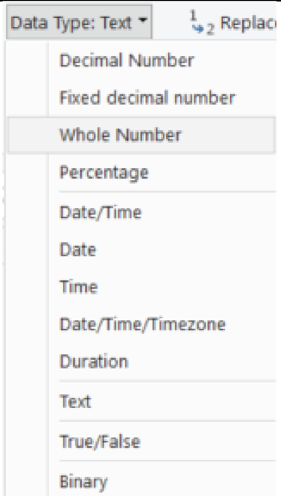
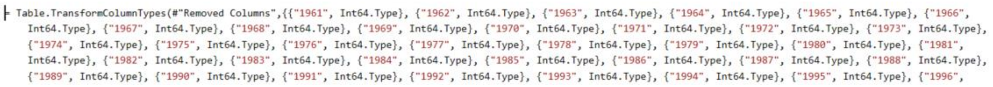
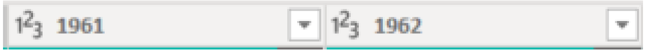
Explanation	The datasets consist of 3% - 15 % missing values in some columns which is a big number of missing values. If removing them, it will cause the sample data size smaller and information loss which is unbeneficial to the analysis. Therefore, the average of the values across the years for different countries is used to replace the null values in the selected columns. However, there are rows that are 100 % null which causes average cannot be calculated. Hence, the remaining null rows will be deleted in the following data cleaning process.
Results	Example from Agricultural Land Dataset:

	 1961 259 (97%) Valid 0 (0%) Error 7 (3%) Empty Remove Empty Column statistics Count 266 Error 0 Empty 7

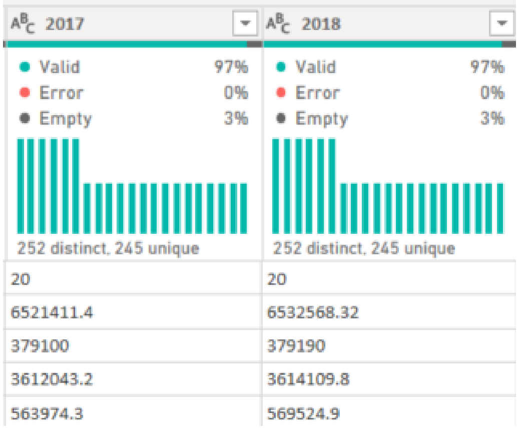
5.3 Data Transformation

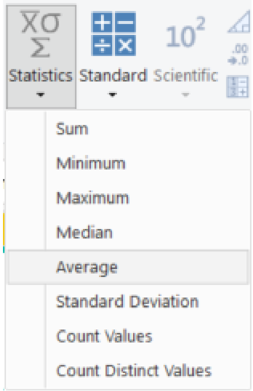
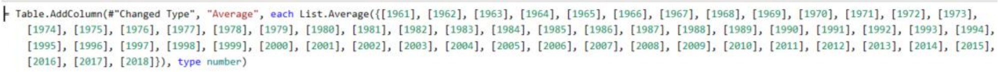
5.3.1 Change Data Type

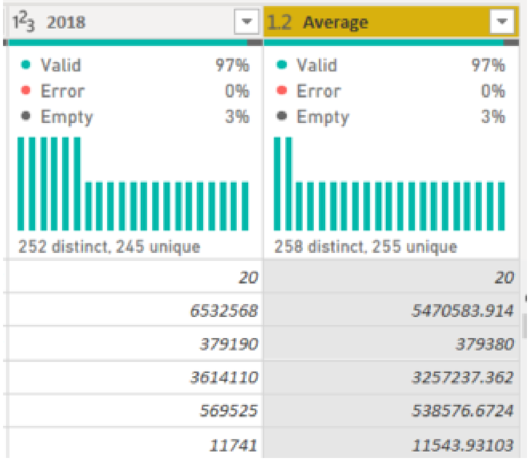
Dataset	Agricultural Land, Land Area, Forest Area, Annual Deforestation,
Before Transformation	Example from Agricultural Land Dataset:  1961 1962 Data Type: Text
Change Data Type	• For Agricultural Land dataset, the data type of columns (1961 – 2018) are changed from “text” to “whole number”. • For Land Area dataset, the data type of columns (1961 – 2020) are changed from “text” to “whole number”. • For Forest Area dataset, the data type of columns (1990 – 2020) are changed from “text” to “whole number”. • For Annual Deforestation dataset, the data type of columns (1990, 2000, 2010, 2015) are changed from “text” to “whole number”. Example from Agricultural Land Dataset:

	 Data Type: Text Replac Decimal Number Fixed decimal number Whole Number Percentage Date/Time Date Time Date/Time/Timezone Duration Text True/False Binary `Table.TransformColumnTypes(#"Removed Columns",{{"1961", Int64.Type}, {"1962", Int64.Type}, {"1963", Int64.Type}, {"1964", Int64.Type}, {"1965", Int64.Type}, {"1966", Int64.Type}, {"1967", Int64.Type}, {"1968", Int64.Type}, {"1969", Int64.Type}, {"1970", Int64.Type}, {"1971", Int64.Type}, {"1972", Int64.Type}, {"1973", Int64.Type}, {"1974", Int64.Type}, {"1975", Int64.Type}, {"1976", Int64.Type}, {"1977", Int64.Type}, {"1978", Int64.Type}, {"1979", Int64.Type}, {"1980", Int64.Type}, {"1981", Int64.Type}, {"1982", Int64.Type}, {"1983", Int64.Type}, {"1984", Int64.Type}, {"1985", Int64.Type}, {"1986", Int64.Type}, {"1987", Int64.Type}, {"1988", Int64.Type}, {"1989", Int64.Type}, {"1990", Int64.Type}, {"1991", Int64.Type}, {"1992", Int64.Type}, {"1993", Int64.Type}, {"1994", Int64.Type}, {"1995", Int64.Type}, {"1996",`
Explanation	The data type is changed because the default data type which is text does not match the value which is a whole number. This may cause errors and the columns cannot be aggregated for the following cleaning and transformation. Besides, the data is rounded off automatically when the data type is changed to “whole number”.
Results	Example from Agricultural Land Dataset: 

5.3.2 Create New Column with Average Value by Year and Country

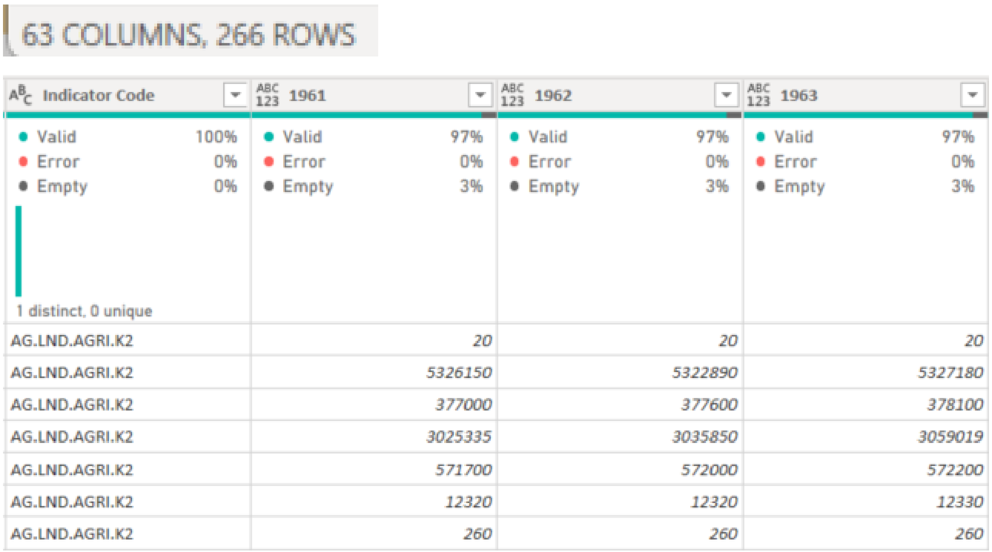
Dataset	Agricultural Land, Land Area, Forest Area, Annual Deforestation, Forest Loss
Before Transformation	Example from Agricultural Land Dataset:  ABC 2017 ABC 2018 Valid 97% Error 0% Empty 3% Valid 97% Error 0% Empty 3% 252 distinct, 245 unique 252 distinct, 245 unique 2017 | 2018 20 | 20 6521411.4 | 6532568.32 379100 | 379190 3612043.2 | 3614109.8 563974.3 | 569524.9

Create New Column	• For Agricultural Land dataset, create a new column by calculating the average value from 1961 to 2018 for different countries. • For Land Area dataset, create a new column by calculating the average value from 1961 to 2020 for different countries. • For Forest Area dataset, create a new column by calculating the average value from 1990 to 2020 for different countries. • For Annual Deforestation dataset, create a new column by calculating the average value from 1990, 2000, 2010 and 2015 for different countries. • For Forest Loss dataset, create a new column by calculating the average value from 2001 to 2020 for different countries. Example from Agricultural Land Dataset:  Statistics Standard Scientific Sum Minimum Maximum Median Average Standard Deviation Count Values Count Distinct Values  `Table.AddColumn(#"Changed Type", "Average", each List.Average({[1961], [1962], [1963], [1964], [1965], [1966], [1967], [1968], [1969], [1970], [1971], [1972], [1973], [1974], [1975], [1976], [1977], [1978], [1979], [1980], [1981], [1982], [1983], [1984], [1985], [1986], [1987], [1988], [1989], [1990], [1991], [1992], [1993], [1994], [1995], [1996], [1997], [1998], [1999], [2000], [2001], [2002], [2003], [2004], [2005], [2006], [2007], [2008], [2009], [2010], [2011], [2012], [2013], [2014], [2015], [2016], [2017], [2018]}), type number)`
Explanation	The average of values across the year for different countries will be used to fill in the null or missing values respectively in different datasets. Besides, the average values for each country can also be used in the following analysis.
Results	Example from Agricultural Land Dataset:

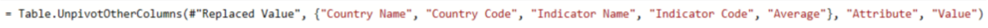
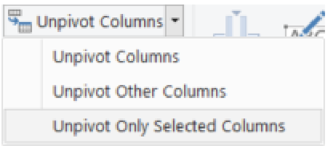
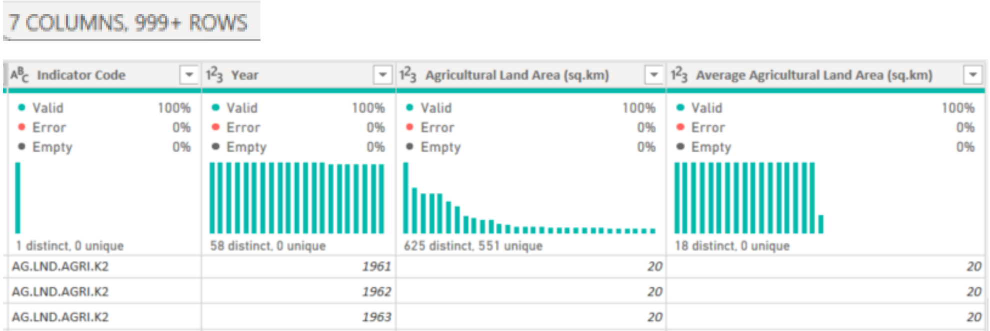
2018	Average
Valid 97%	Valid 97%
Error 0%	Error 0%
Empty 3%	Empty 3%
252 distinct, 245 unique	258 distinct, 255 unique
20	20
6532568	5470583.914
379190	379380
3614110	3257237.362
569525	538576.6724
11741	11543.93103

5.3.3 Unpivot Columns

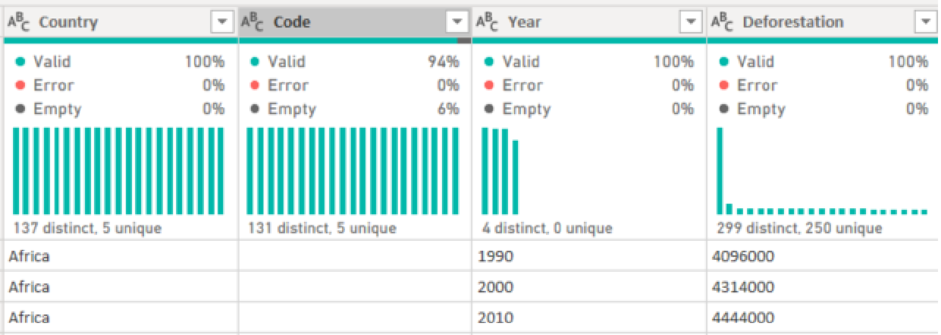
Dataset	Agricultural Land, Land Area, Forest Area, Annual Deforestation, Forest Loss
Before Transformation	Example from Agricultural Land Dataset: 
Unpivot Columns, Rename the columns	• For Agricultural Land dataset, there are 58 columns (1961 – 2018) are selected to unpivot. • For Land Area dataset, there are 60 columns (1961 – 2020) are selected to unpivot. • For Forest Area dataset, there are 31 columns (1990 – 2020) are selected to unpivot. • For Annual Deforestation dataset, there are 4 columns (1990, 2000, 2010 and 2015) are selected to unpivot.

63 COLUMNS, 266 ROWS

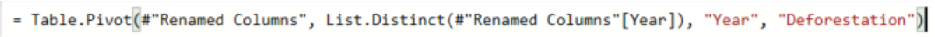
Indicator Code	1961	1962	1963
Valid 100%	Valid 97%	Valid 97%	Valid 97%
Error 0%	Error 0%	Error 0%	Error 0%
Empty 0%	Empty 3%	Empty 3%	Empty 3%
1 distinct, 0 unique			
AG.LND.AGRI.K2	20	20	20
AG.LND.AGRI.K2	5326150	5322890	5327180
AG.LND.AGRI.K2	377000	377600	378100
AG.LND.AGRI.K2	3025335	3035850	3059019
AG.LND.AGRI.K2	571700	572000	572200
AG.LND.AGRI.K2	12320	12320	12330
AG.LND.AGRI.K2	260	260	260

	• For Forest Loss dataset, there are 20 columns (2001 - 2020) are selected to unpivot. • Rename the “Attribute” as “Year”, “Value” as “Agricultural Land Size (sq.km)” and “Average” as “Average Agricultural Land Size (sq.km)”. • Rename the “Attribute” as “Year”, “Value” as “Land Area (sq.km)” and “Average” as “Average Land Area (sq.km)”. • Rename the “Attribute” as “Year”, “Value” as “Forest Area (sq.km)” and “Average” as “Average Forest Area (sq.km)”. • Rename the “Attribute” as “Year”, “Value” as “Annual Deforestation (ha)” and “Average” as “Average Annual Deforestation (ha)”. • Rename the “Attribute” as “Year”, “Value” as “Forest Loss (ha)” and “Average” as “Average Forest Loss (ha)”. Example from Agricultural Land Dataset:  = Table.UnpivotOtherColumns(#"Replaced Value", {"Country Name", "Country Code", "Indicator Name", "Indicator Code", "Average"}, "Attribute", "Value")  Unpivot Columns Unpivot Columns Unpivot Other Columns Unpivot Only Selected Columns
Explanation	When the columns are all unpivot, the values are now transformed from columns into rows which allows us to evaluate and generate charts or graphs more easily. When unpivot columns is done, the rows that have 100 % null values will not include in the new column. Therefore, all the columns have no null values now as others null values is replaced by the average value across the years for different countries. Besides, the two columns generated is renamed as the column name provides us a better understanding on the data.
Results	Example from Agricultural Land Dataset:  7 COLUMNS, 999+ ROWS Indicator Code | Year | Agricultural Land Area (sq.km) | Average Agricultural Land Area (sq.km) Valid 100% | Valid 100% | Valid 100% | Valid 100% Error 0% | Error 0% | Error 0% | Error 0% Empty 0% | Empty 0% | Empty 0% | Empty 0% 1 distinct, 0 unique | 58 distinct, 0 unique | 625 distinct, 551 unique | 18 distinct, 0 unique AG.LND.AGRI.K2 | 1961 | 20 | 20 AG.LND.AGRI.K2 | 1962 | 20 | 20 AG.LND.AGRI.K2 | 1963 | 20 | 20

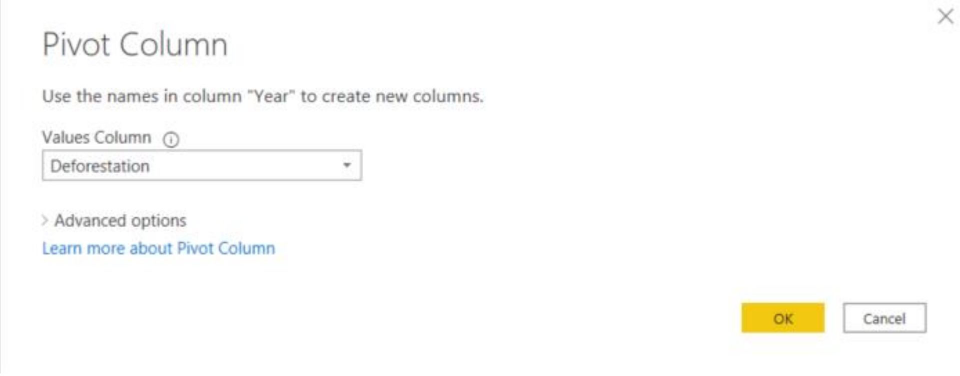
5.3.4 Pivot Columns

<table>
<tr><td>Dataset</td><td>Annual Deforestation, Forest Loss</td></tr>
<tr><td>Before Transformation</td><td>Example from Annual Deforestation Dataset:

<table>
<tr><th>Country</th><th>Code</th><th>Year</th><th>Deforestation</th></tr>
<tr><td>● Valid 100%
● Error 0%
● Empty 0%</td><td>● Valid 94%
● Error 0%
● Empty 6%</td><td>● Valid 100%
● Error 0%
● Empty 0%</td><td>● Valid 100%
● Error 0%
● Empty 0%</td></tr>
<tr><td>137 distinct, 5 unique</td><td>131 distinct, 5 unique</td><td>4 distinct, 0 unique</td><td>299 distinct, 250 unique</td></tr>
<tr><td>Africa</td><td></td><td>1990</td><td>4096000</td></tr>
<tr><td>Africa</td><td></td><td>2000</td><td>4314000</td></tr>
<tr><td>Africa</td><td></td><td>2010</td><td>4444000</td></tr>
</table></td></tr>
<tr><td>Pivot Columns</td><td>

- For Annual Deforestation Dataset, the values in “Year” column are used to create new columns and the values of “Deforestation” are used as values in the columns.

Example from Annual Deforestation Dataset:


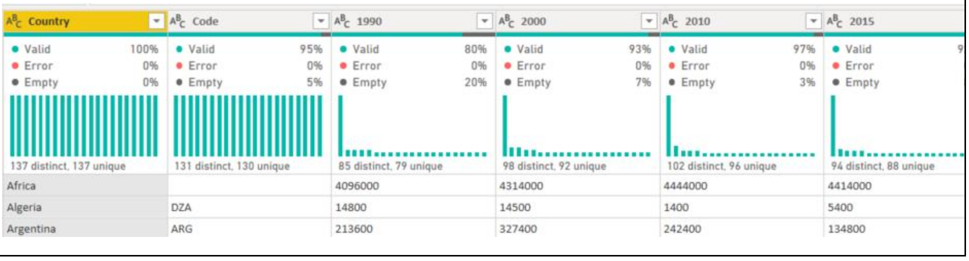
```
= Table.Pivot(#"Renamed Columns", List.Distinct(#"Renamed Columns"[Year]), "Year", "Deforestation")
```



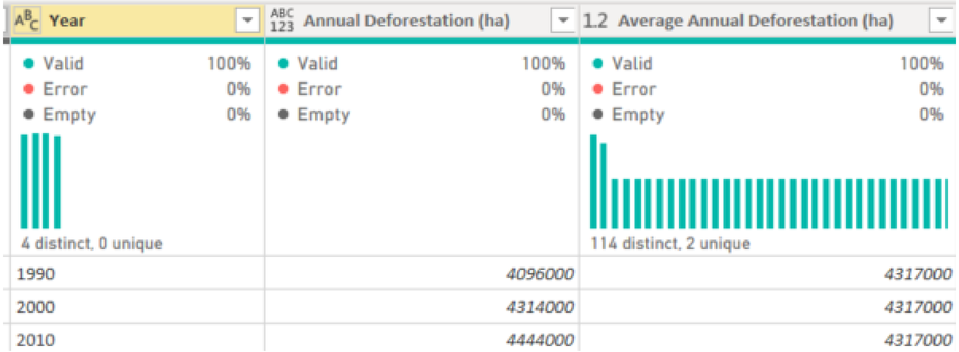
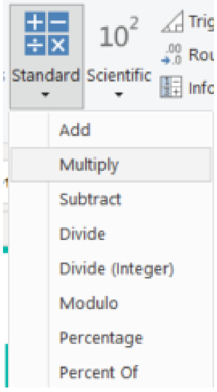
Pivot Column

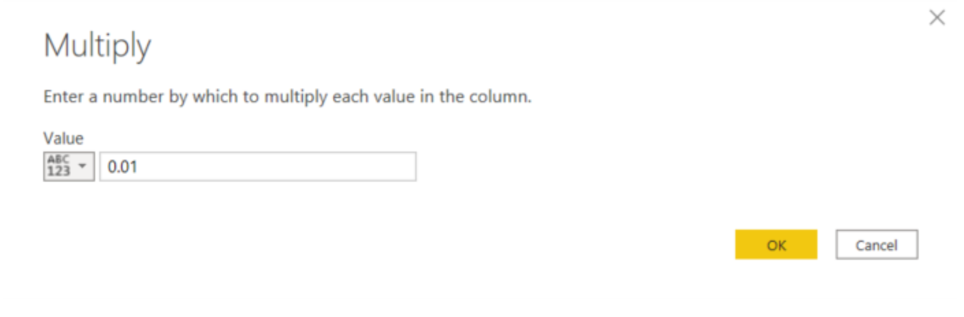
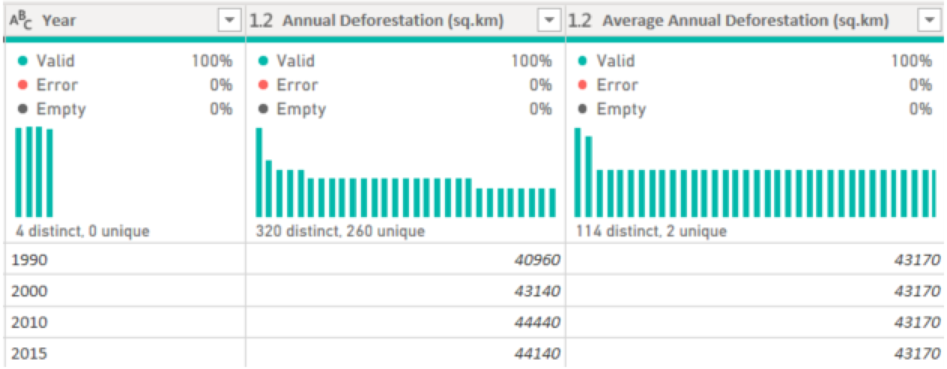
Use the names in column "Year" to create new columns.

Values Column: Deforestation

</td></tr>
<tr><td>Explanation</td><td>Pivot columns is required to calculate the average of the value across the years for different countries in order to replace the null values in the datasets.</td></tr>
<tr><td>Results</td><td>Example from Annual Deforestation Dataset:

<table>
<tr><th>Country</th><th>Code</th><th>1990</th><th>2000</th><th>2010</th><th>2015</th></tr>
<tr><td>● Valid 100%
● Error 0%
● Empty 0%</td><td>● Valid 95%
● Error 0%
● Empty 5%</td><td>● Valid 80%
● Error 0%
● Empty 20%</td><td>● Valid 93%
● Error 0%
● Empty 7%</td><td>● Valid 97%
● Error 0%
● Empty 3%</td><td>● Valid 9
● Error
● Empty</td></tr>
<tr><td>137 distinct, 137 unique</td><td>131 distinct, 130 unique</td><td>85 distinct, 79 unique</td><td>98 distinct, 92 unique</td><td>102 distinct, 96 unique</td><td>94 distinct, 88 unique</td></tr>
<tr><td>Africa</td><td></td><td>4096000</td><td>4314000</td><td>4444000</td><td>4414000</td></tr>
<tr><td>Algeria</td><td>DZA</td><td>14800</td><td>14500</td><td>1400</td><td>5400</td></tr>
<tr><td>Argentina</td><td>ARG</td><td>213600</td><td>327400</td><td>242400</td><td>134800</td></tr>
</table></td></tr>
</table>

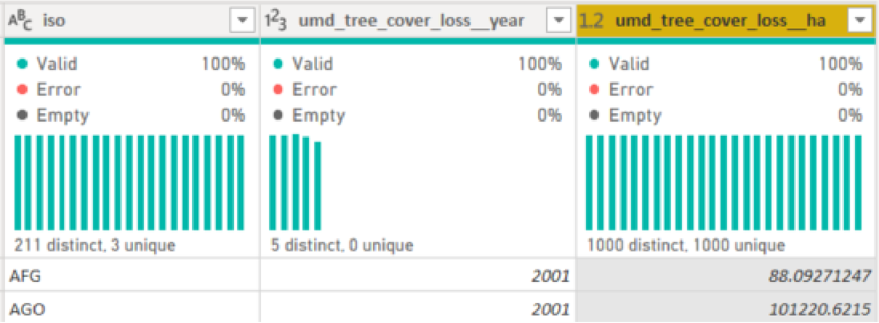
5.3.5 Convert Unit of Measurement

Dataset	Annual Deforestation, Forest Loss, Global Forest Loss by Dominant Drivers
Before Transformation	Example from Annual Deforestation Dataset:  Year | Annual Deforestation (ha) | Average Annual Deforestation (ha) Valid 100% | Valid 100% | Valid 100% Error 0% | Error 0% | Error 0% Empty 0% | Empty 0% | Empty 0% 4 distinct, 0 unique | | 114 distinct, 2 unique 1990 | 4096000 | 4317000 2000 | 4314000 | 4317000 2010 | 4444000 | 4317000
Convert Unit of Measurement	• For Annual Deforestation dataset, convert the values from hectares (ha) to square kilometre (sq.km) in selected columns (Annual Deforestation (ha), Average Annual Deforestation (ha)). • For Forest Loss dataset, convert the values from hectares (ha) to square kilometre (sq.km) in selected columns (Forest Loss (ha), Average Forest Loss (ha)). • For Global Forest Loss by Dominant Drivers dataset, convert the values from hectares (ha) to square kilometre (sq.km) in selected columns (Forest Loss (ha)). • The formula used to convert hectares (ha) to square kilometre (sq.km) is [sq.km = ha/100]. Example from Annual Deforestation Dataset:  Standard Scientific Trig Rou Info Add Multiply Subtract Divide Divide (Integer) Modulo Percentage Percent Of

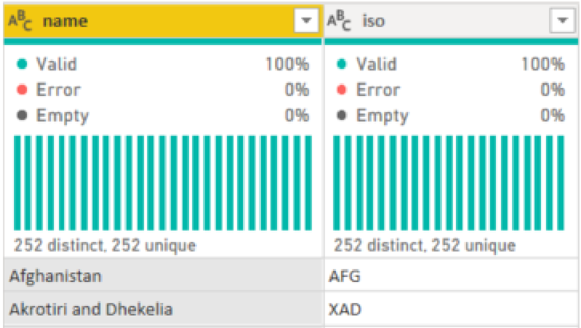
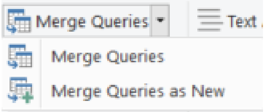
	 Multiply Enter a number by which to multiply each value in the column. Value 0.01 OK Cancel
Explanation	The unit of measurement is converted from hectares to square kilometre to standardize the dataset and avoid bias in the analysis and modelling. The result of analysis or modelling will inaccurate if the unit of measurement of the values in dataset is different.
Results	Example from Annual Deforestation Dataset: 

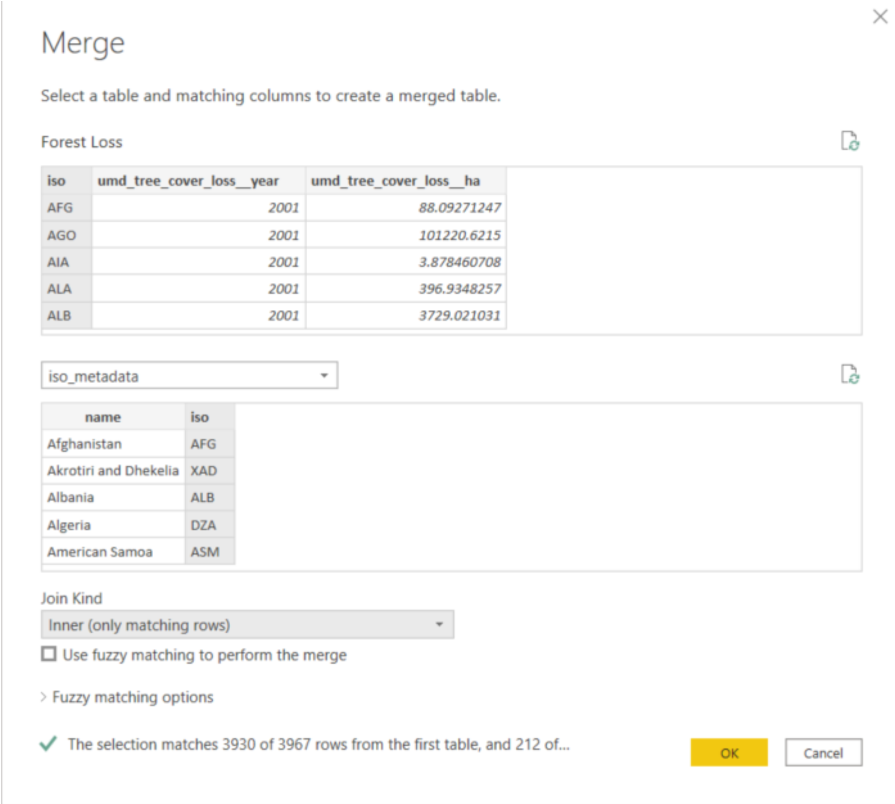
Year	Annual Deforestation (sq.km)	Average Annual Deforestation (sq.km)
Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%
4 distinct, 0 unique	320 distinct, 260 unique	114 distinct, 2 unique
1990	40960	43170
2000	43140	43170
2010	44440	43170
2015	44140	43170

5.3.6 Merge Datasets

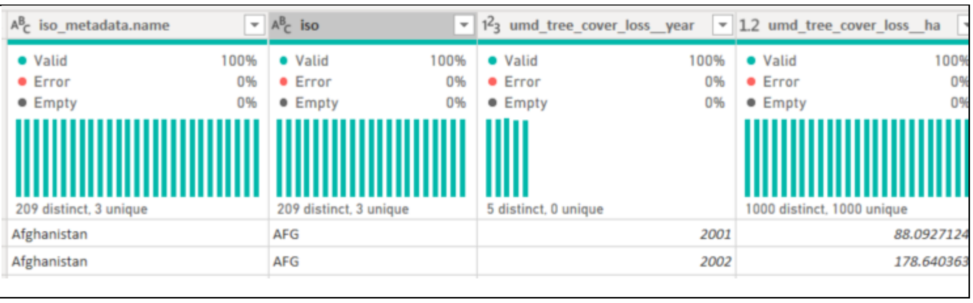
Dataset	Forest Loss, iso_metadata,
Before Transformation	Example from Forest Loss Dataset: 

iso	umd_tree_cover_loss__year	umd_tree_cover_loss__ha
Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%
211 distinct, 3 unique	5 distinct, 0 unique	1000 distinct, 1000 unique
AFG	2001	88.09271247
AGO	2001	101220.6215

<table>
<tr><td></td><td>Example from iso_metadata dataset:

<table>
<tr><th>name</th><th>iso</th></tr>
<tr><td>● Valid 100%
● Error 0%
● Empty 0%</td><td>● Valid 100%
● Error 0%
● Empty 0%</td></tr>
<tr><td>252 distinct, 252 unique</td><td>252 distinct, 252 unique</td></tr>
<tr><td>Afghanistan</td><td>AFG</td></tr>
<tr><td>Akrotiri and Dhekelia</td><td>XAD</td></tr>
</table>
</td></tr>
<tr><td>Merge Dataset</td><td>
<ul><li>Inner join the Forest Loss dataset and iso_metadata dataset with a matching column, “iso”.</li></ul>
Example from Forest Loss & iso_metadata Dataset:


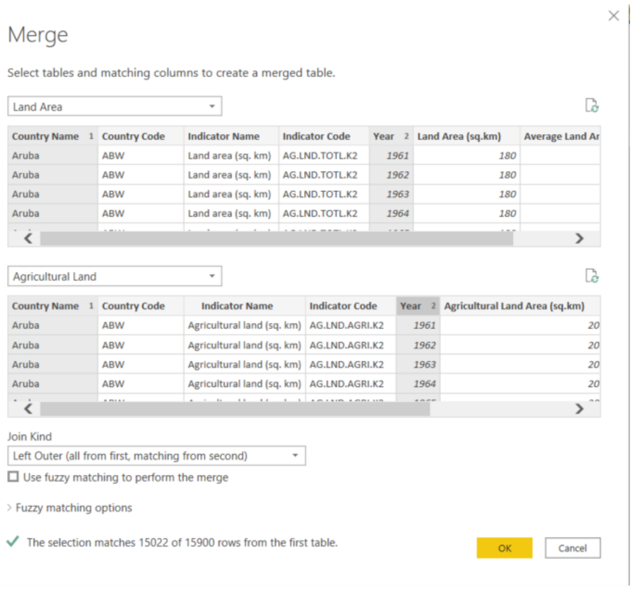

Merge
Select a table and matching columns to create a merged table.
Forest Loss
<table>
<tr><th>iso</th><th>umd_tree_cover_loss__year</th><th>umd_tree_cover_loss__ha</th></tr>
<tr><td>AFG</td><td>2001</td><td>88.09271247</td></tr>
<tr><td>AGO</td><td>2001</td><td>101220.6215</td></tr>
<tr><td>AIA</td><td>2001</td><td>3.878460708</td></tr>
<tr><td>ALA</td><td>2001</td><td>396.9348257</td></tr>
<tr><td>ALB</td><td>2001</td><td>3729.021031</td></tr>
</table>
iso_metadata
<table>
<tr><th>name</th><th>iso</th></tr>
<tr><td>Afghanistan</td><td>AFG</td></tr>
<tr><td>Akrotiri and Dhekelia</td><td>XAD</td></tr>
<tr><td>Albania</td><td>ALB</td></tr>
<tr><td>Algeria</td><td>DZA</td></tr>
<tr><td>American Samoa</td><td>ASM</td></tr>
</table>
Join Kind
Inner (only matching rows)
□ Use fuzzy matching to perform the merge
› Fuzzy matching options
✓ The selection matches 3930 of 3967 rows from the first table, and 212 of...

</td></tr>
<tr><td>Explanation</td><td>Both datasets have a matching column which allows merging together so that the following analysis would be easier as the variables are now stored in the same table.</td></tr>
<tr><td>Results</td><td>Example from Forest Loss Dataset:</td></tr>
</table>

iso_metadata.name	iso	umd_tree_cover_loss__year	umd_tree_cover_loss__ha
● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%
209 distinct, 3 unique	209 distinct, 3 unique	5 distinct, 0 unique	1000 distinct, 1000 unique
Afghanistan	AFG	2001	88.0927124
Afghanistan	AFG	2002	178.640363

The five cleaned and transformed datasets which are Land Area, Agricultural Land, Forest Area, Annual Deforestation and Forest Loss datasets are ready to merge and integrate in order to provide an overview on all variable in one table. This allows us to perform more efficient searches and the following analysis would be much easier to carry out. Firstly, Land Area dataset is merged with Agricultural Land dataset based on "Country Name" and "Year" column as illustrated in Figure 5.3.6(a) and Figure 5.3.6(c). After that, the columns like "Agricultural Land (sq.km)" and "Average Agricultural Land (sq.km)" from Agricultural Land (sq.km) is expanded in the table.



Merge

Select tables and matching columns to create a merged table.

Land Area

Country Name 1	Country Code	Indicator Name	Indicator Code	Year 2	Land Area (sq.km)	Average Land Ar
Aruba	ABW	Land area (sq. km)	AG.LND.TOTL.K2	1961	180	
Aruba	ABW	Land area (sq. km)	AG.LND.TOTL.K2	1962	180	
Aruba	ABW	Land area (sq. km)	AG.LND.TOTL.K2	1963	180	
Aruba	ABW	Land area (sq. km)	AG.LND.TOTL.K2	1964	180	

Agricultural Land

Country Name 1	Country Code	Indicator Name	Indicator Code	Year 2	Agricultural Land Area (sq.km)
Aruba	ABW	Agricultural land (sq. km)	AG.LND.AGRI.K2	1961	20
Aruba	ABW	Agricultural land (sq. km)	AG.LND.AGRI.K2	1962	20
Aruba	ABW	Agricultural land (sq. km)	AG.LND.AGRI.K2	1963	20
Aruba	ABW	Agricultural land (sq. km)	AG.LND.AGRI.K2	1964	20

Join Kind

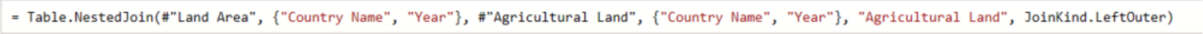
Left Outer (all from first, matching from second)

□ Use fuzzy matching to perform the merge

› Fuzzy matching options

✓ The selection matches 15022 of 15900 rows from the first table.

Figure 5.3.6(a): GUI for Merging Land Area Dataset and Agricultural Land Dataset

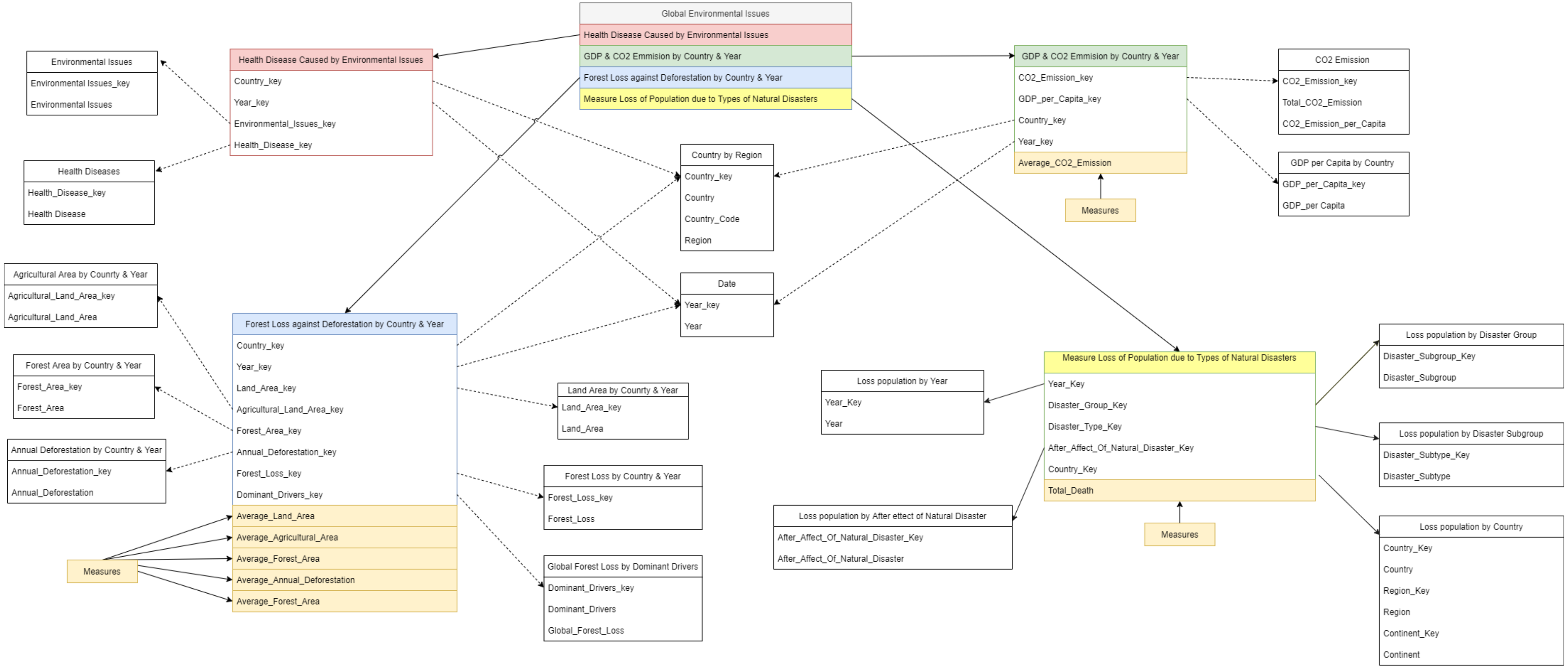


```
= Table.NestedJoin(#"Land Area", {"Country Name", "Year"}, #"Agricultural Land", {"Country Name", "Year"}, "Agricultural Land", JoinKind.LeftOuter)
```

Figure 5.3.6(b): Code for Merging Land Area Dataset and Agricultural Land Dataset

After that, other datasets are merged by using the method shown above. As we are using left join for merging datasets, there are some null values or missing values exists in numerical columns. Therefore, replacing the null values with "N/A" would be good approach to avoid null values in the analysis as average and median could not be aggregated from the columns.

5.4 Data Schema



6.0 Prediction & Modelling Techniques

6.1 OLAP Report

OLAP Report: Forest Loss and Deforestation against Agricultural, Forest and Land Area by Regions, Countries and Year.

Attributes: Land Area (sq.km), Forest Area (sq.km), Agricultural Land Area (sq.km), Average Agricultural Land Area (sq.km), Average Forest Area (sq.km), Average Land Area (sq.km), Forest Loss, Dominant Drivers, Annual Deforestation (sq.km), Regions, Countries, Year.

Purpose: OLAP Dashboard allows users to visualize on the summary of forest loss and deforestation across the years for different countries. They are able to compare the forest area, land area and agricultural land area from the visualisation done in order to determine the severity of the forest loss and deforestation. The basic operations such as drill down can be done to view the forest loss, agricultural land area, forest area, annual deforestation by regions, countries and years. The users can have a better understanding and make decisions on mitigating the forest loss and deforestation by viewing the relationship between different attributes based on the visualization in the dashboard.

Advantages:

There are several advantages when OLAP is chosen to implement in data analysis to explore and discover understandable trends and patterns. Firstly, OLAP provides multidimensional data representation includes drill down, roll up, slice and dice and pivot which allows users to have a better understanding on the relationships between different attributes. For instance, users can view the more detailed forest loss by regions and also countries by drilling down. Furthermore, OLAP allows users to create measures and filters in order to review the information from a specific aspect. Besides, the processing speed of the information is another advantage of OLAP report. The execution time of a single query is short so that users does not have to spend much time and extract the information in a shorter time.

Disadvantages:

OLAP report provides many advantages to users but there are some disadvantages too. First of all, one of the disadvantages of OLAP is high dependency on IT as it requires users to have at least basic IT skills in order to create the report. For some cases, SQL scripts or any programming codes are needed to create OLAP report. Furthermore, it requires a vast amount of data to create an informative and insightful report as there are operations like drill down can be done.

6.2 Predictive Modelling

6.2.1 Linear Regression Model

In this project, the prediction technique used is Linear Regression model which is a model that shows the relationship between an independent variable and dependent variable and predict the future outcome.

Attributes: Agricultural Land (sq.km), Annual Deforestation (sq.km), Regions

By implementing linear regression model, the relationship between agricultural land area (sq.km) and annual deforestation (sq.km) by regions can be determined and the future annual deforestation in different regions will be predicted. The annual deforestation will be predicted by entering a value of agricultural land area. Hence, the users will have a better understanding on the severity of deforestation and how agricultural land area will impact the deforestation in different regions. After that, users are able to make decisions on different solutions to mitigate the deforestation rate with the predicted results.

Advantages:

Linear regression model is widely used by different analyst or users as it is very helpful in data analysis or predictive modelling. Firstly, linear regression is user-friendly, and it only requires simple implementation for analysis and modelling. Users are able to train models efficiently and effectively by using a lower computational power compared to other modelling algorithms. Furthermore, overfitting can be avoided by implementing regularization which can be implement easily.

Disadvantages:

One of the disadvantages of linear regression model is that the model is sensitive to noisy data or outliers as they might affect the performance or accuracy of the model. Besides, the model will assume that there is a relationship between the independent and dependent variables which is not accurate and not reliable.

6.2.2 Correlation Plot

In this project, a correlation plot is utilized to determine and investigate the strength of relationship between variables. There is a numerical measure between two variables known as correlation coefficient shown in the correlation plot. The correlation plot will help us to determine which variable have relation with another so that the two variables can be used to fit into the linear regression model. The correlations of the variables are mapped in the plot as illustrated in Figure 6.2.2(a). The positive and negative correlations can be determined by the colours which is blue to red or the numerical values which is range between -1 to 1. In our correlation plot, the blue colour indicates positive correlation while red colour indicates the negative correlation. Based on the correlation plot, there are several strong correlations includes agricultural land area and annual deforestation, land area and annual deforestation and more. Hence, agricultural land area and annual deforestation is selected for the linear regression model as they have strong correlation and agricultural land area is one of the main possible drivers to deforestation.

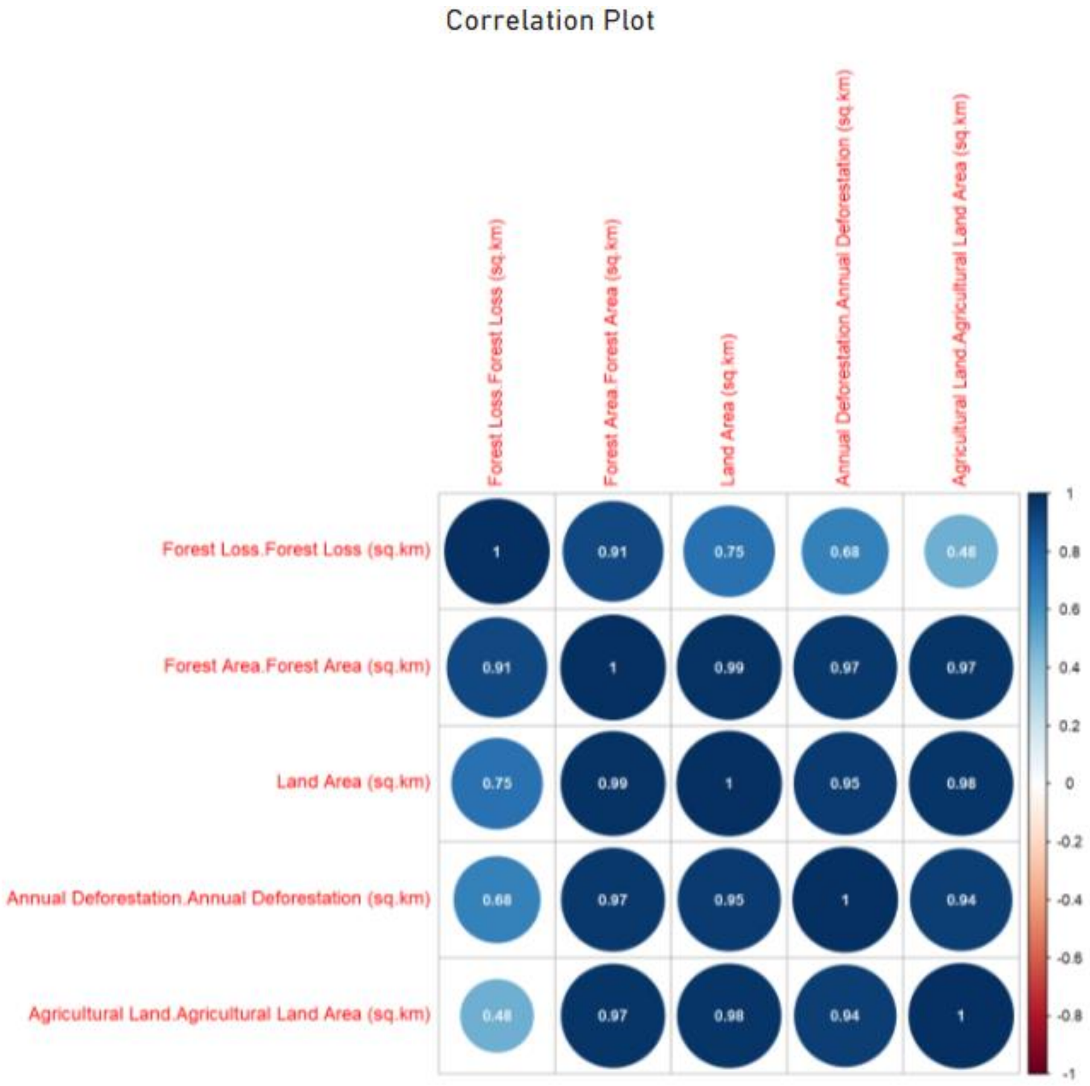


Figure 6.2.2(a): Correlation Plot

7.0 Analysis and Recommendations

7.1 OLAP Report

The data that was selected in data selection process related to forest loss and deforestation will be used to create a OLAP dashboard. Hence, the dashboard can be viewed by different users and understand the severity of deforestation.

7.1.1 Building OLAP Dashboard

In this project, OLAP Dashboard is created by using Power BI which is an analytical service provided by Microsoft. The dashboard provides several visualizations such as graphs and charts to analyse and extract information from the data. There are some aggregated data are used in the dashboard to allows users to visualize in another aspect.

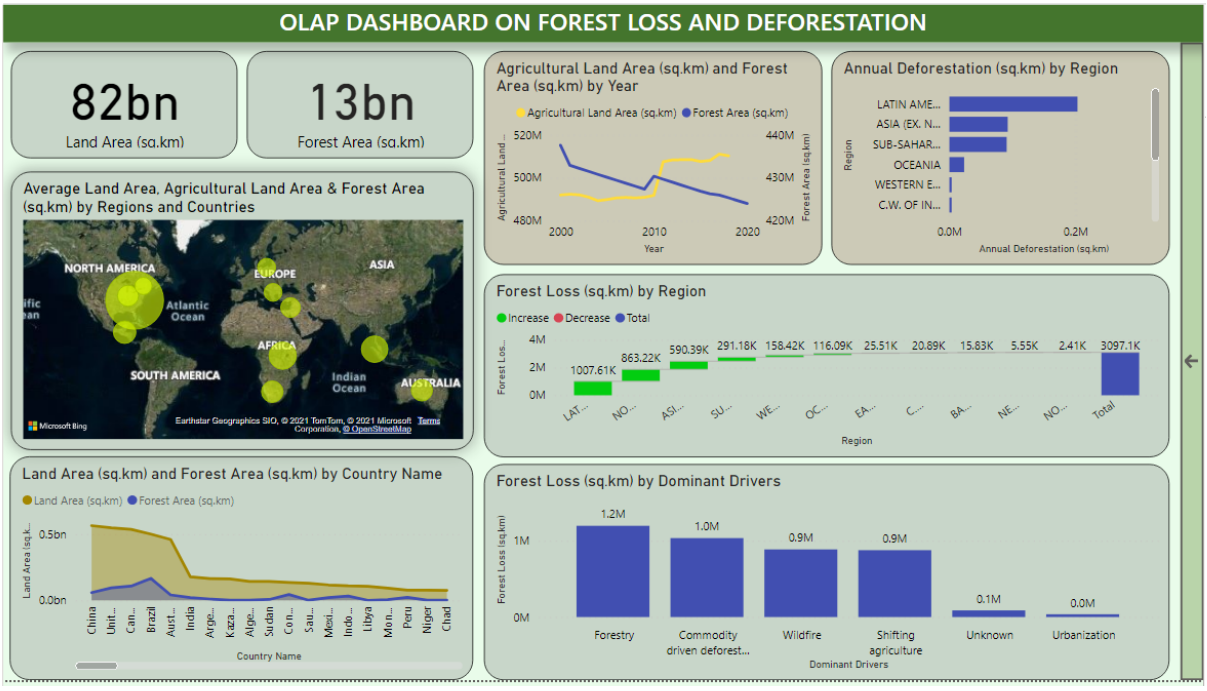


Figure7.1.1(a): Main Dashboard on Forest Loss and Deforestation

The main objective of the main dashboard as illustrated in Figure 7.1.1(a) is to provides insights and descriptive analysis of several variables includes forest loss, annual deforestation, land area, agricultural land area, forest area and the dominant drivers of deforestation. The dashboard consists of 2 cards, a map, an area chart, a line chart, a clustered bar chart, a waterfall chart and a stacked column chart.

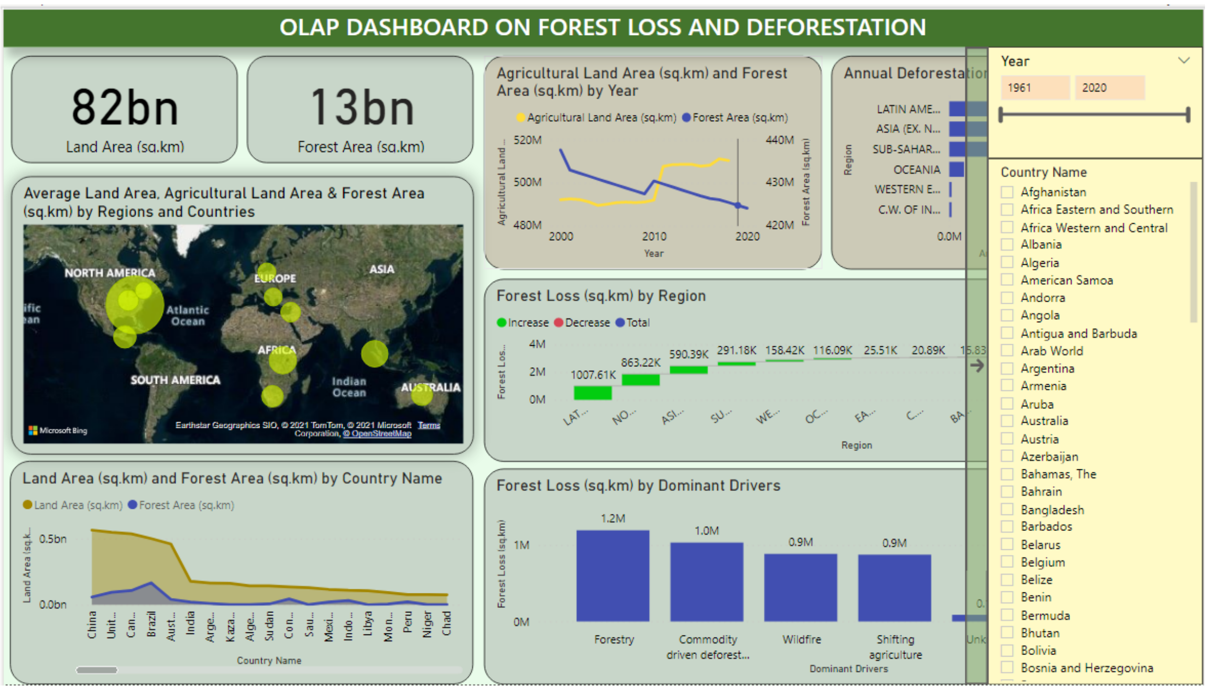

OLAP DASHBOARD ON FOREST LOSS AND DEFORESTATION
82bn
Land Area (sq.km)
13bn
Forest Area (sq.km)
Agricultural Land Area (sq.km) and Forest Area (sq.km) by Year
Agricultural Land Area (sq.km) Forest Area (sq.km)
520M 500M 480M
440M 430M 420M
2000 2010 2020
Year
Annual Deforestation
LATIN AME... ASIA (EX. N... SUB-SAHAR... OCEANIA WESTERN E... C.W. OF IN...
0.0M
Region
Year
1961 2020
Country Name
Afghanistan
Africa Eastern and Southern
Africa Western and Central
Albania
Algeria
American Samoa
Andorra
Angola
Antigua and Barbuda
Arab World
Argentina
Armenia
Aruba
Australia
Austria
Azerbaijan
Bahamas, The
Bahrain
Bangladesh
Barbados
Belarus
Belgium
Belize
Benin
Bermuda
Bhutan
Bolivia
Bosnia and Herzegovina
Average Land Area, Agricultural Land Area & Forest Area (sq.km) by Regions and Countries
NORTH AMERICA
EUROPE
ASIA
Atlantic Ocean
AFRICA
SOUTH AMERICA
Indian Ocean
AUSTRALIA
Microsoft Bing
Earthstar Geographics SIO, © 2021 TomTom, © 2021 Microsoft Corporation, © OpenStreetMap Terms
Forest Loss (sq.km) by Region
Increase Decrease Total
4M 2M 0M
1007.61K 863.22K 590.39K 291.18K 158.42K 116.09K 25.51K 20.89K 15.83
LAT... NO... ASI... SU... WE... OC... EA... C... BA
Region
Land Area (sq.km) and Forest Area (sq.km) by Country Name
Land Area (sq.km) Forest Area (sq.km)
0.5bn 0.0bn
China Unit... Can... Brazil Aust... India Arge... Kaza... Alge... Sudan Con... Sau... Mexi... Indo... Libya Mon... Peru Niger Chad
Country Name
Forest Loss (sq.km) by Dominant Drivers
1M 0M
1.2M 1.0M 0.9M 0.9M
Forestry
Commodity driven deforest...
Wildfire
Shifting agriculture
Dominant Drivers

Figure 7.1.1(b): Filter by Year and Country is Provided in the Main Dashboard

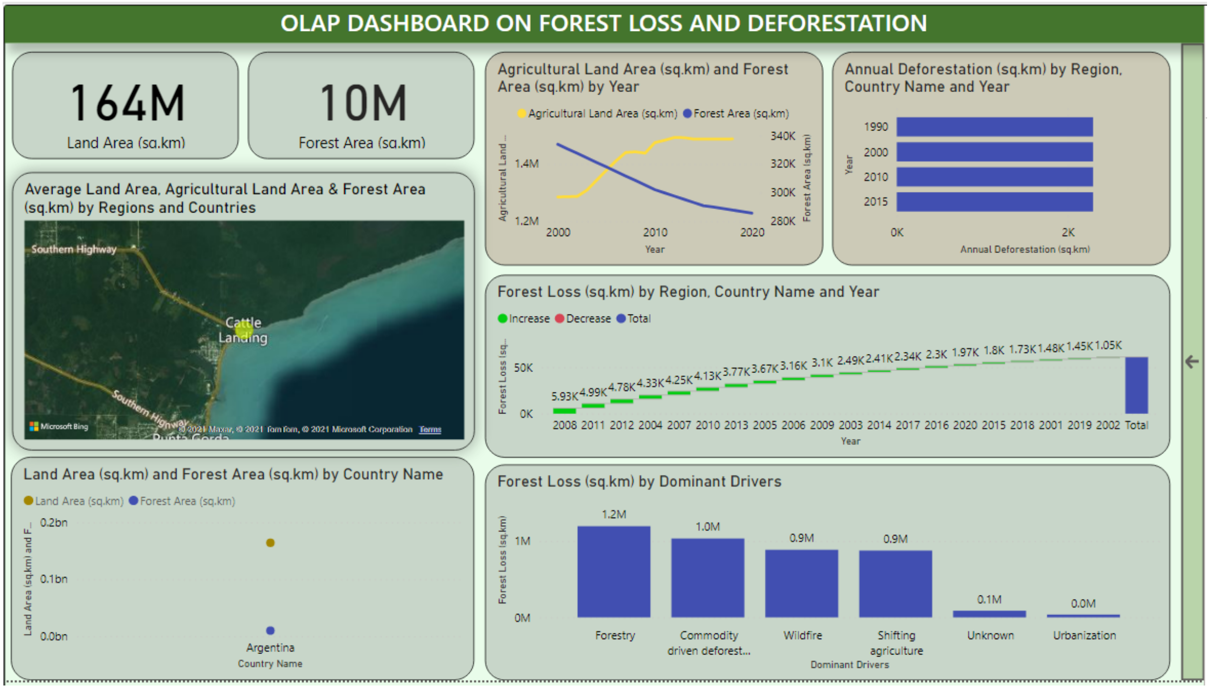

OLAP DASHBOARD ON FOREST LOSS AND DEFORESTATION
164M
Land Area (sq.km)
10M
Forest Area (sq.km)
Agricultural Land Area (sq.km) and Forest Area (sq.km) by Year
Agricultural Land Area (sq.km) Forest Area (sq.km)
1.4M 1.2M
340K 320K 300K 280K
2000 2010 2020
Year
Annual Deforestation (sq.km) by Region, Country Name and Year
1990 2000 2010 2015
0K 2K
Annual Deforestation (sq.km)
Average Land Area, Agricultural Land Area & Forest Area (sq.km) by Regions and Countries
Southern Highway
Cattle Landing
Southern Highway
Punta Gorda
Microsoft Bing
© 2021 Maxar, © 2021 TomTom, © 2021 Microsoft Corporation Terms
Forest Loss (sq.km) by Region, Country Name and Year
Increase Decrease Total
50K 0K
5.93K 4.99K 4.78K 4.33K 4.25K 4.13K 3.77K 3.67K 3.16K 3.1K 2.49K 2.41K 2.34K 2.3K 1.97K 1.8K 1.73K 1.48K 1.45K 1.05K
2008 2011 2012 2004 2007 2010 2013 2005 2006 2009 2003 2014 2017 2016 2020 2015 2018 2001 2019 2002 Total
Year
Land Area (sq.km) and Forest Area (sq.km) by Country Name
Land Area (sq.km) Forest Area (sq.km)
0.2bn 0.1bn 0.0bn
Argentina
Country Name
Forest Loss (sq.km) by Dominant Drivers
1M 0M
1.2M 1.0M 0.9M 0.9M 0.1M 0.0M
Forestry
Commodity driven deforest...
Wildfire
Shifting agriculture
Unknown
Urbanization
Dominant Drivers

Figure 7.1.1(c): Filter Result on the Dashboard

Besides, the main dashboard provides a filter function shown in Figure 7.1.1(b) where users can filter the graphs and charts by year and countries. After choosing a specific year or country, the dashboard will be sliced into the selected year or country. For instance, “Argentina” is selected as the country on the filter panel and the dashboard will show the related information of Argentina on the main dashboard as illustrated in Figure 7.1.1(c).

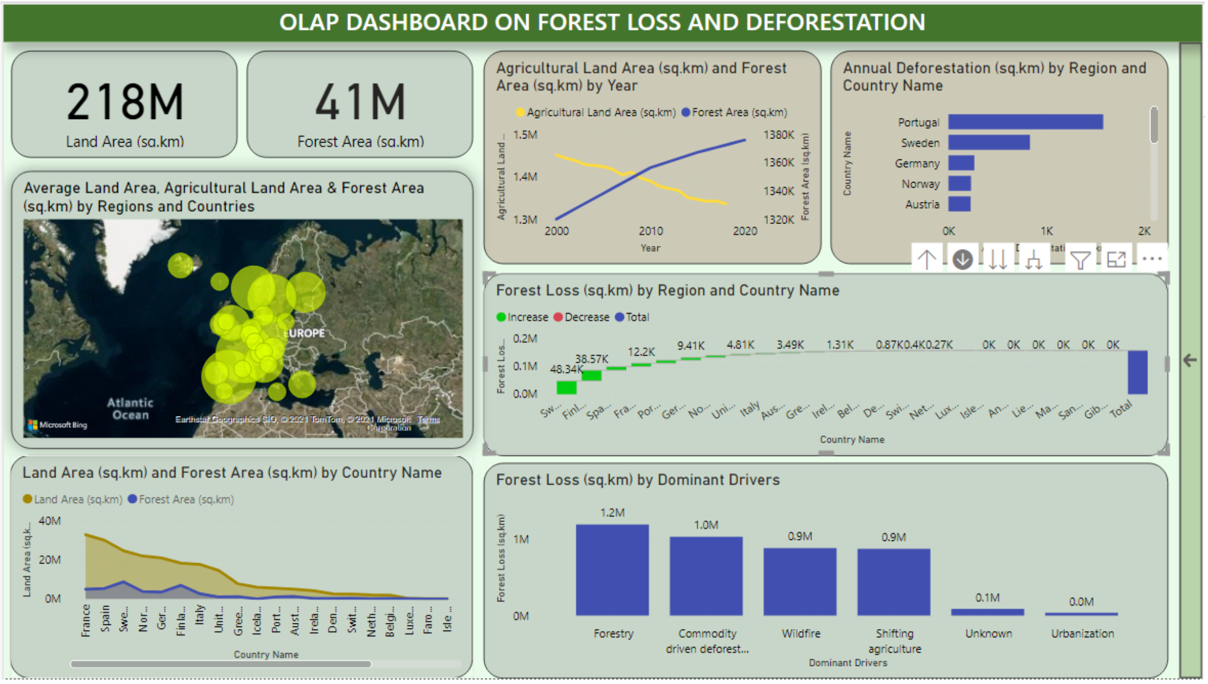


Figure7.1.1(d): Example of Drilling Down in the Main Dashboard

Furthermore, the main dashboard can be drilled down into different layers directly from the graphs or charts. For instance, users can drill down into a specific region which is “Western Europe” to visualize different information in that specific region as shown in Figure 7.1.1(d). The others graphs and charts will also show the information of the specific region automatically.

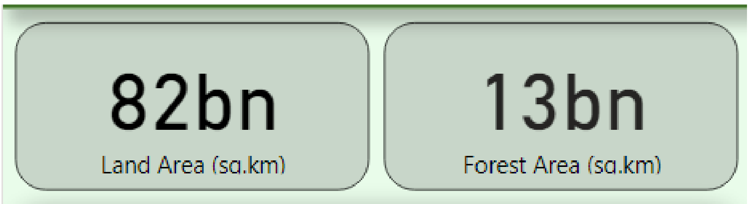


Figure 7.1.1(e): 2 Cards Showing Land Area and Forest Area in sq.km

The cards illustrated in Figure 7.1.1(e) is showing the total land area and forest area in square kilometer (sq.km). When there is any operation done such as drilling down or slicing, the value of the cards will turn accordingly. Users can view the value of the cards immediately as the land area and forest area will provide insights in determining the difference between the land area and the forest area in the region or country.

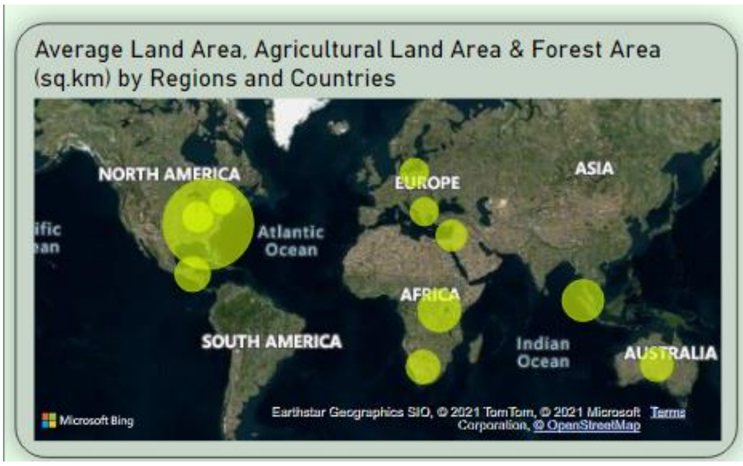

Average Land Area, Agricultural Land Area & Forest Area (sq.km) by Regions and Countries
NORTH AMERICA
EUROPE
ASIA
Atlantic Ocean
AFRICA
SOUTH AMERICA
Indian Ocean
AUSTRALIA
Microsoft Bing
Earthstar Geographics SIO, © 2021 TomTom, © 2021 Microsoft Corporation, © OpenStreetMap Terms

Figure 7.1.1(f): Map that shows the Average Land Area, Agricultural Land Area & Forest Area (sq.km) with the locations.

As shown in Figure 7.1.1(f), the regions exist in our dataset are plotted in the map to provide an overview on where is the regions located for users. The country will be plotted when the map is drilled down into a specific region. For instance, countries like Sweden, Norway, Belgium and many more will be shown if the map is drilled down into "Western Europe". The values of average land area, agricultural land area and forest area will appear when cursor is pointing at the region or country.

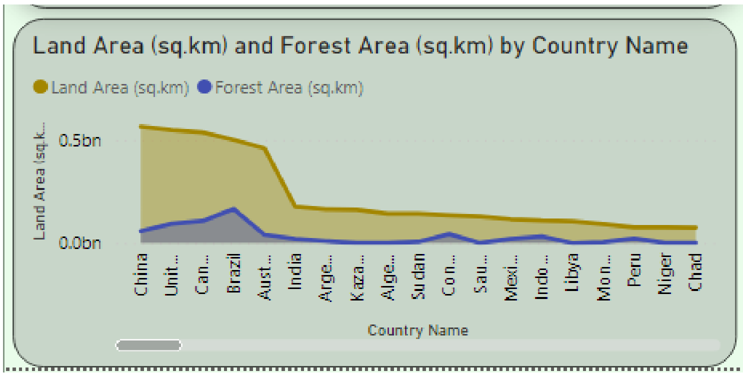
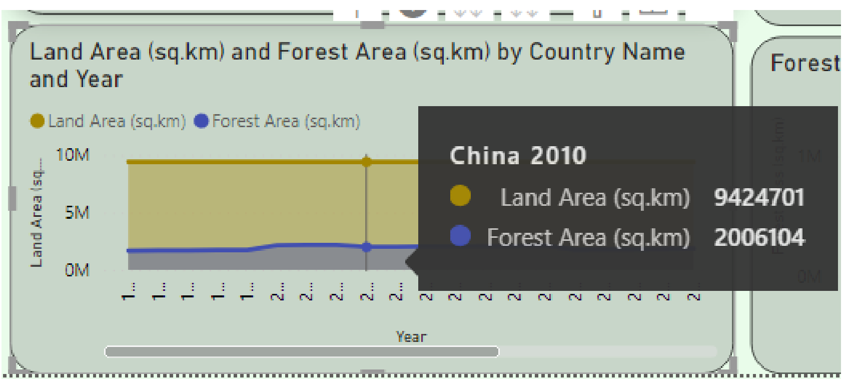


Figure 7.1.1(g): Area Chart Showing Land Area and Forest Area by Country


Land Area (sq.km) and Forest Area (sq.km) by Country Name and Year
● Land Area (sq.km) ● Forest Area (sq.km)
Land Area (sq...
10M
5M
0M
Year
Forest

China 2010
● Land Area (sq.km) 9424701
● Forest Area (sq.km) 2006104

Figure7.1.1(h): Area Chart After Drilled Down

The area chart in Figure 7.1.1(g) shows the land area (sq.km) which is in brown colour and forest area (sq.km) which is in blue colour of different countries. This chart allows users to compare the land area and forest area of different countries visually so that the relationship between them can be determined easily. The values can be viewed by users when the cursor is pointing in the area chart. The chart can be drilled down into different years of different countries. For example, users can view the land area and forest area of a country such as "China" from year 1990 to 2020 as illustrated in Figure 7.1.1(h).

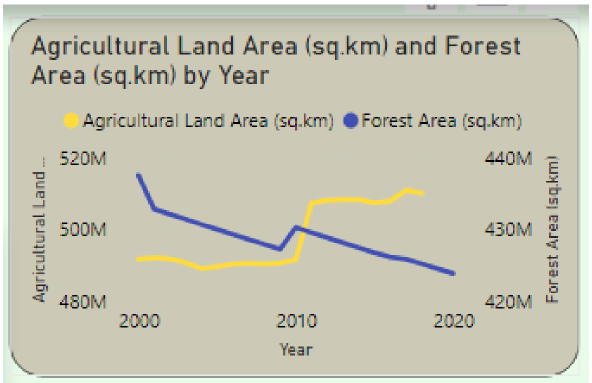


Figure 7.1.1(i): Line Graph Showing Agricultural Land Area and Forest Area (sq.km) by Year

The line graph illustrated in Figure 7.1.1(i) demonstrates the relationship between agricultural land area and forest area (sq.km) across the years. The yellow line indicates the agricultural land area whereas the blue line indicates the forest area. For instance, we can see that the forest area is declining when the agricultural land area is increasing between the year 2000 and 2020. Hence, users can analyse some meaningful patterns and extract information from the graph above.

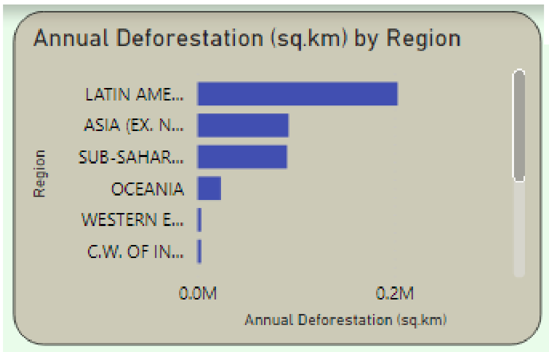


Figure 7.1.1(j): Clustered Bar Chart Showing Annual Deforestation by Region

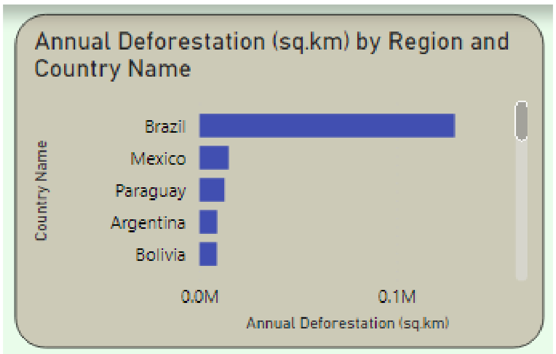


Figure 7.1.1(k): Clustered Bar Graph After Drilled Down

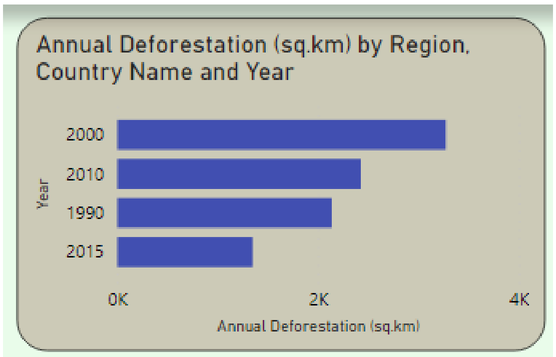


Figure7.1.1(l): Clustered Bar Graph After Second Drilled Down

As illustrated in Figure 7.1.1(j), the clustered bar chart shows the annual deforestation for different regions. Hence, users can identify which region has the highest annual deforestation by looking at the chart. In this chart, users can also drill down the chart the show a more detailed information by countries. For example, when users drill down into a region such as "Latin America", the information of the countries in Latin America such as "Brazil", "Mexico" and many more will show up as illustrated in Figure 7.1.1(k). After that, users can drill down one more layer which is year in order to visualize the deforestation across years for the country as shown in Figure 7.1.1(l). Therefore, a more detailed analysis can be carried out by the users.

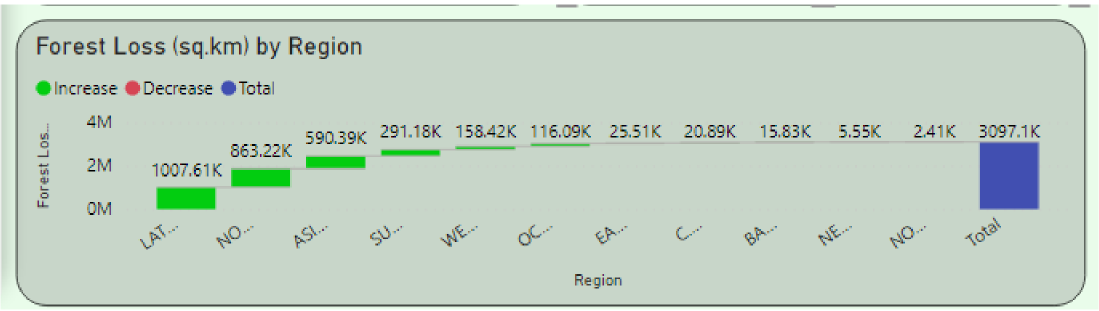


Figure 7.1.1(m): Waterfall Chart Showing Forest Loss (sq.km) by region

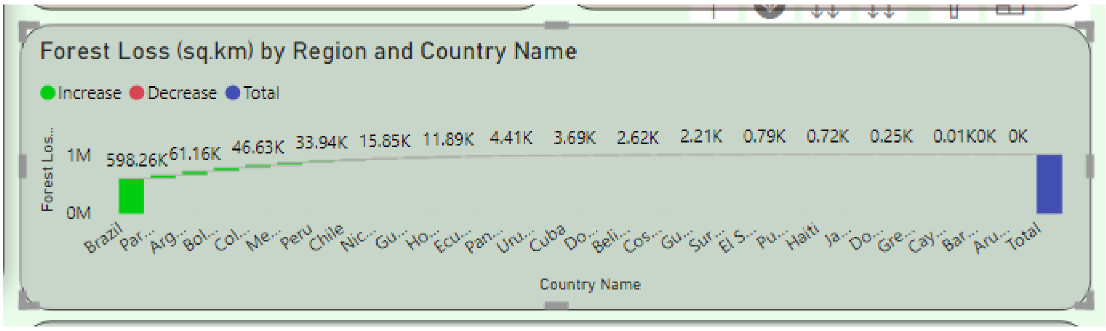


Figure 7.1.1(n): Waterfall Chart After Drilling Down

The waterfall chart as shown in Figure 7.1.1(m) demonstrates the forest loss (sq,km) in different regions. The green bar indicates that the forest loss has increased while the red bar indicates that the forest loss has decreased, and the total amount of forest loss is illustrated with a blue bar. The chart can also be drilled down into different countries within a particular region in order to analyse the forest loss across the countries which is located in the same region. For example, the forest loss of countries such as "Brazil", "Argentina" and more will be shown when "Latin America" has been drilled down.

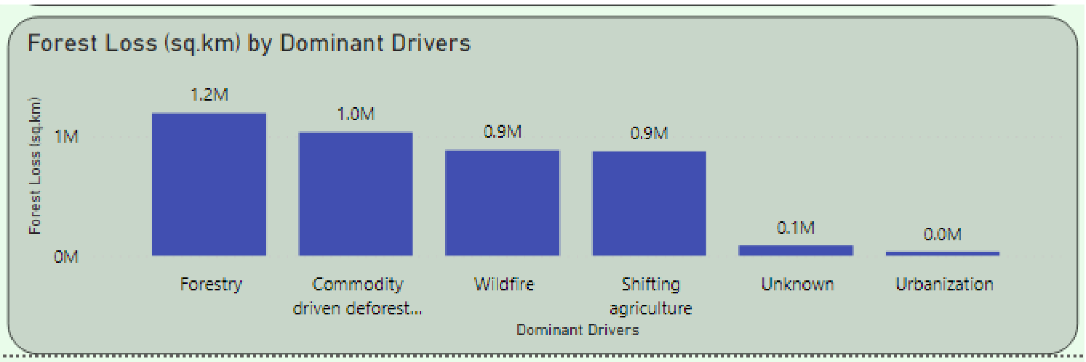


Figure 7.1.1(o): Forest Loss by Dominant Drivers

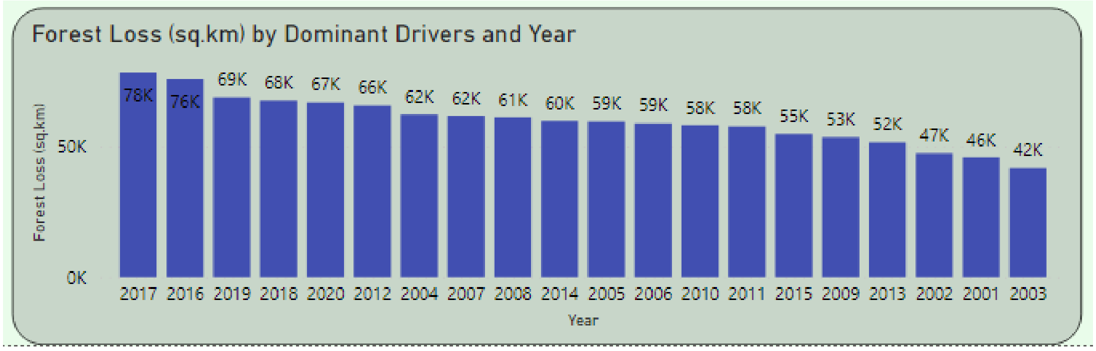


Figure 7.11(p): Stacked Column Chart After Drilled Down

Based on the stacked column chart shown in Figure 7.1.1(o), there are 6 dominant drivers includes forestry, commodity driven deforestation, wildfire, shifting agriculture, unknown and urbanization are used in the chart. This allows users to have an overview and determine which dominant drivers causes the most forest loss globally. Besides, the users can drill down into each dominant drivers to view the forest loss caused by the specific dominant driver across the years. For instance, Figure 7.1.1(p) demonstrates the forest loss caused by 'Forestry" globally across the years.

7.1.2 Analysis on OLAP Dashboard

After building OLAP dashboard, it provides a ton of insights and information on deforestation and forest loss which matches the scope of the project. The visualizations gained from the OLAP dashboard will be analysed to extract meaningful information which will help in decision making and achieve the scope of the project. Firstly, the countries are having a larger land area compared to the forest area in an overall basis as illustrated in Figure 7.1.1(g). After drilling down as illustrated in 7.1.1(h), the forest area in most of the countries increases between the year 1990 to 2000 whereas it decreases after 2000 year. This indicates that the forest area in most of the countries are reducing when the forestry in 1990 to 2000 is not as developed as now. Before 2000s, the global development is slow compared to the current situation, there is still a large amount of land that has not been developed. Hence, there is no need to remove or develop the forest area. However, the global development after 2000s is growing rapidly and the land area that can be developed in some countries are getting scarce and the forestry has a rising demand. Therefore, the forest area in some countries is declining after 2000s as the deforestation has increases as illustrated in Figure 7.1.1(l).

Besides, the agricultural land area increases while the forest area decreases from year 2000 to 2020 as shown in the Figure 7.1.1(i). We can know that agricultural land may be one of the possible drivers of deforestation based on the graph. This is because when the demand of agricultural industry increases, the use of agricultural land will expand to the forest area but not land use area as the land use area are usually utilized for urbanization purpose. Furthermore, we can see that the forest area in Brazil is larger compared to other countries in Latin America and the Brazil has the highest annual deforestation in Latin America by slicing down to Brazil. Based on the scenario, we can assume that countries that have larger forest area may have a higher deforestation rate. When the country has a large forest area as their natural resources, the government or organisation of the country might want to fully utilize the natural resources to increase the economy of the country. However, there might be other factors or possible drivers such as different policy of countries will affect the annual deforestation of the country.

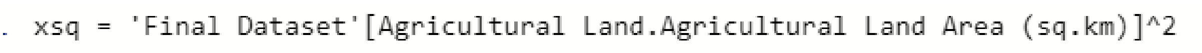
Moreover, the forest loss caused by commodity driven deforestation globally is lower before year 2010 in an overall basis. This indicates that the global demand of commodity that drives deforestation from human increases as the time goes on. Therefore, there are more commodity that requires the resources from forest which will increase the forest loss globally. With this information, the public or users will know that the deforestation will increases time by time as the global development is still growing rapidly, Besides, they will also understand that the expansion of agricultural land and commodity driven deforestation might have a large impact on the forest loss. The countries that are having large forest area should be alert on the deforestation rate of their countries. Lastly, the users or public should take action on solutions based on forest loss and deforestation that has been analysed and visualized according to locations and time.

7.2 Linear Regression Modelling

The data that was selected from the data selection process related to the annual deforestation and its possible drivers will be used for linear regression model. Based on the correlation plot done previously, agricultural land area (sq.km) and annual deforestation (sq.km) will be utilized in the linear regression model.

7.2.1 Building Linear Regression Model

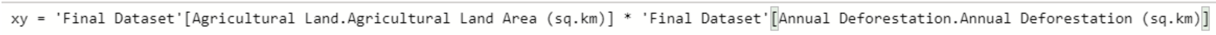
In this project, linear regression model is created and developed by using Power BI. The relationship between agricultural land area (sq.km) and annual deforestation (sq.km) will be modelled by fitting into linear equation which is y = mx + b. In order to fit the linear equation, some values should be calculated by using different formula using DAX in Power BI.



```
xsq = 'Final Dataset'[Agricultural Land.Agricultural Land Area (sq.km)]^2
```

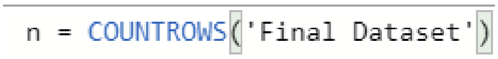
Figure 7.2.1(a): Calculate x square

In our case, x represents the agricultural land area whereas y represents the annual deforestation. Agricultural land area is selected as it is one of the possible drivers of deforestation. Besides, the annual deforestation is the variable chosen to be predicted as it provides us the future deforestation which meets the scope of this project. Firstly, x square is calculated as it is required for the following calculations to determine the scope and b-intercept.



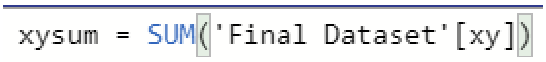
```
xy = 'Final Dataset'[Agricultural Land.Agricultural Land Area (sq.km)] * 'Final Dataset'[Annual Deforestation.Annual Deforestation (sq.km)]
```

*Figure 7.2.1(b): Calculate x * y*



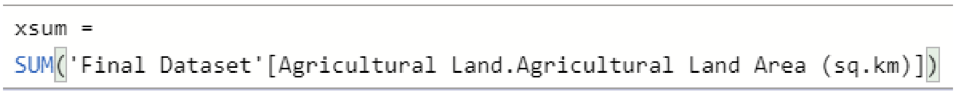
```
n = COUNTROWS('Final Dataset')
```

Figure 7.2.1(c): Count Rows of the Dataset Used



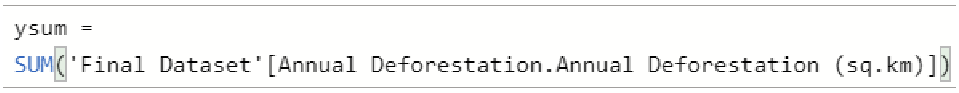
```
xysum = SUM('Final Dataset'[xy])
```

Figure 7.2.1(d): Sum of xy



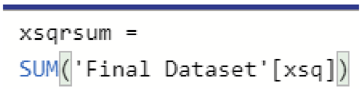
```
xsum =
SUM('Final Dataset'[Agricultural Land.Agricultural Land Area (sq.km)])
```

Figure 7.2.1(e): Sum of x



```
ysum =
SUM('Final Dataset'[Annual Deforestation.Annual Deforestation (sq.km)])
```

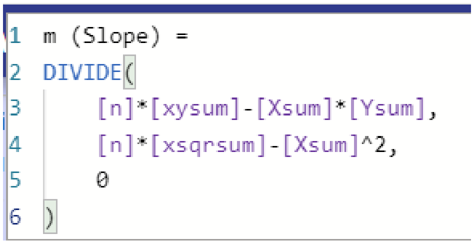
Figure 7.2.1(f): Sum of y



```
xsqrsum =
SUM('Final Dataset'[xsq])
```

Figure 7.2.1(g): Sum of xsq

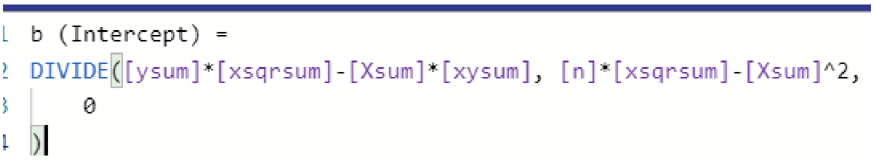
After that, x times y which represents agricultural land area times annual deforestation is also required for the following calculations as illustrated in Figure 7.2.1(b). Besides, the total number of rows are needed to calculate as it will be used in calculating slope and b-intercept formula as illustrated in Figure 7.2.1(c). The total of xy, total of x and total of y is also be calculated by using the in-built formula provided by Power BI.



```
m (Slope) =
DIVIDE(
    [n]*[xysum]-[Xsum]*[Ysum],
    [n]*[xsqrsum]-[Xsum]^2,
    0
)
```

Figure 7.2.1(h): Calculate slope

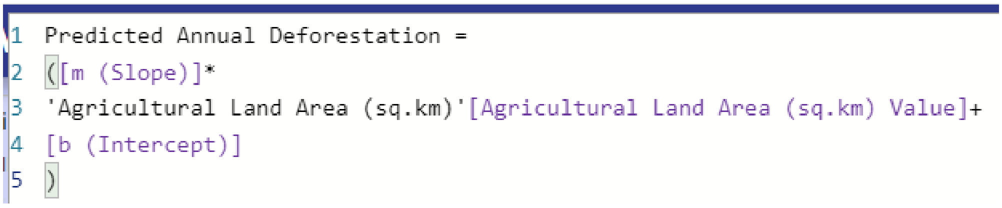
Then, the slope is calculated as illustrated in Figure 7.2.1(h) by using the formula and the measures that are calculated previously as they are needed to fit into the linear equation which is y = mx + b.



```
b (Intercept) =
DIVIDE([ysum]*[xsqrsum]-[Xsum]*[xysum], [n]*[xsqrsum]-[Xsum]^2,
    0
)
```

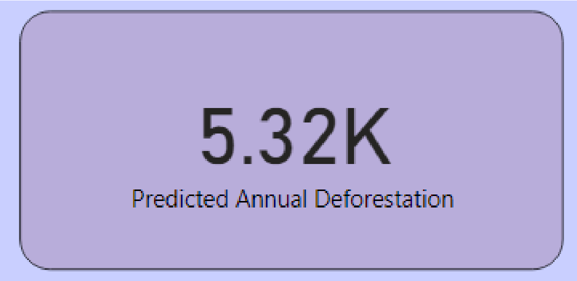
Figure 7.2.1(i): Calculate b-intercept

Next, b-intercept is calculated too by using the formula as shown in Figure 7.2.1(i).



```
Predicted Annual Deforestation =
([m (Slope)]*
'Agricultural Land Area (sq.km)'[Agricultural Land Area (sq.km) Value]+
[b (Intercept)]
)
```

Figure 7.2.1 (j): Calculate the Predicted Annual Deforestation

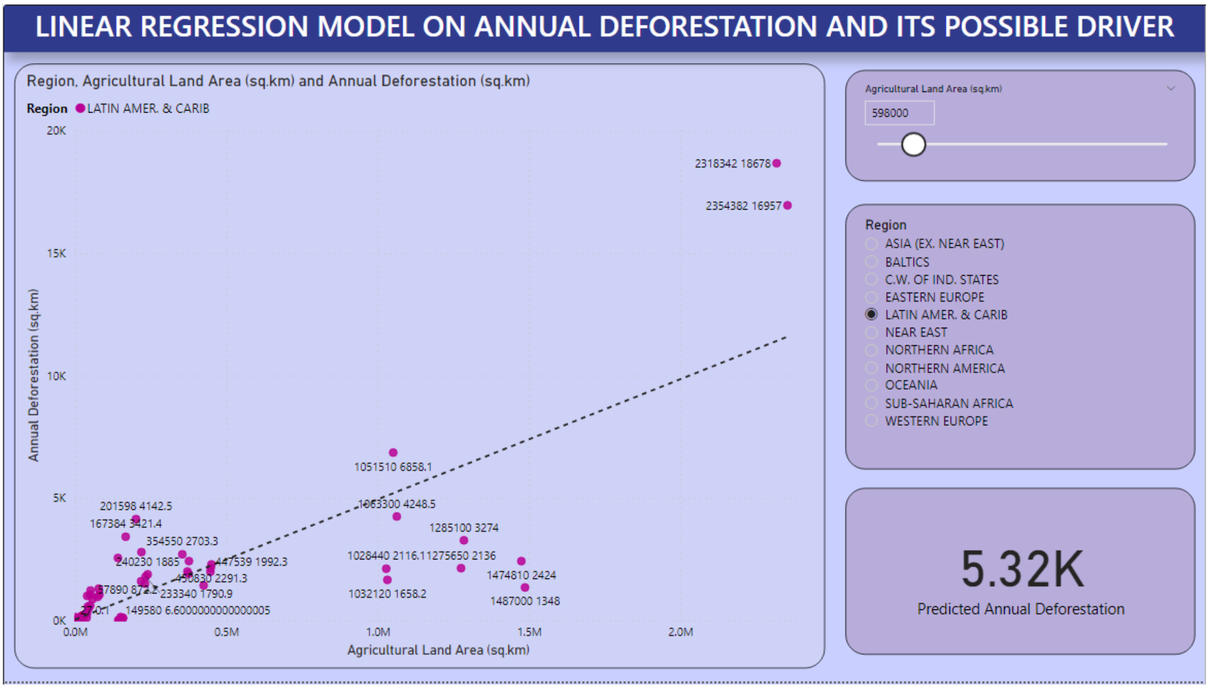


5.32K

Predicted Annual Deforestation

Figure 7.2.1(k): Card Visualization for Predicted Annual Deforestation

Lastly, the predicted annual deforestation will be calculated by using the forumla as shown in Figure 7.2.1(j) and the slope and b-intercept calculated previously. The predidcted value will be calculated based on the agriculturcal land area value entered by users as the formula consists of the value of agricultural land area from the what-if parameter. The value of predicted annual deforestation will be shown in a card visualization as illustrated as Figure 7.2.1(k) which helps users to determine the value more easier.



LINEAR REGRESSION MODEL ON ANNUAL DEFORESTATION AND ITS POSSIBLE DRIVER

Agricultural Land Area (sq.km): 598000

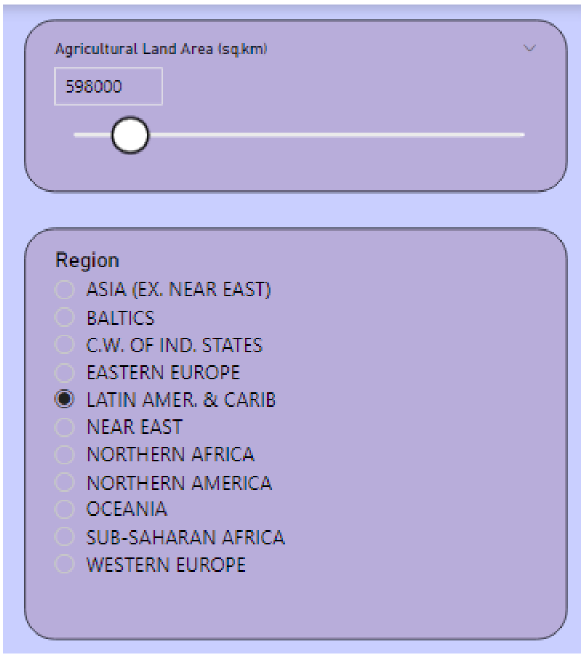
Region:
○ ASIA (EX. NEAR EAST)
○ BALTICS
○ C.W. OF IND. STATES
○ EASTERN EUROPE
● LATIN AMER. & CARIB
○ NEAR EAST
○ NORTHERN AFRICA
○ NORTHERN AMERICA
○ OCEANIA
○ SUB-SAHARAN AFRICA
○ WESTERN EUROPE

5.32K

Predicted Annual Deforestation

Figure 7.2.1(l): Visualization of Linear Regression Model

Linear regression model as shown in Figure 7.2.1(l) is developed to shows the relationship between agricultural land area and annual deforestation and the future value of annual deforestation will be predicted based on the value of agricultural land. This model gives an overview on the future annual deforestation value to the users so that they could have a better understanding on the probable future situation and make decision on different solutions.



Agricultural Land Area (sq.km)

598000

Region

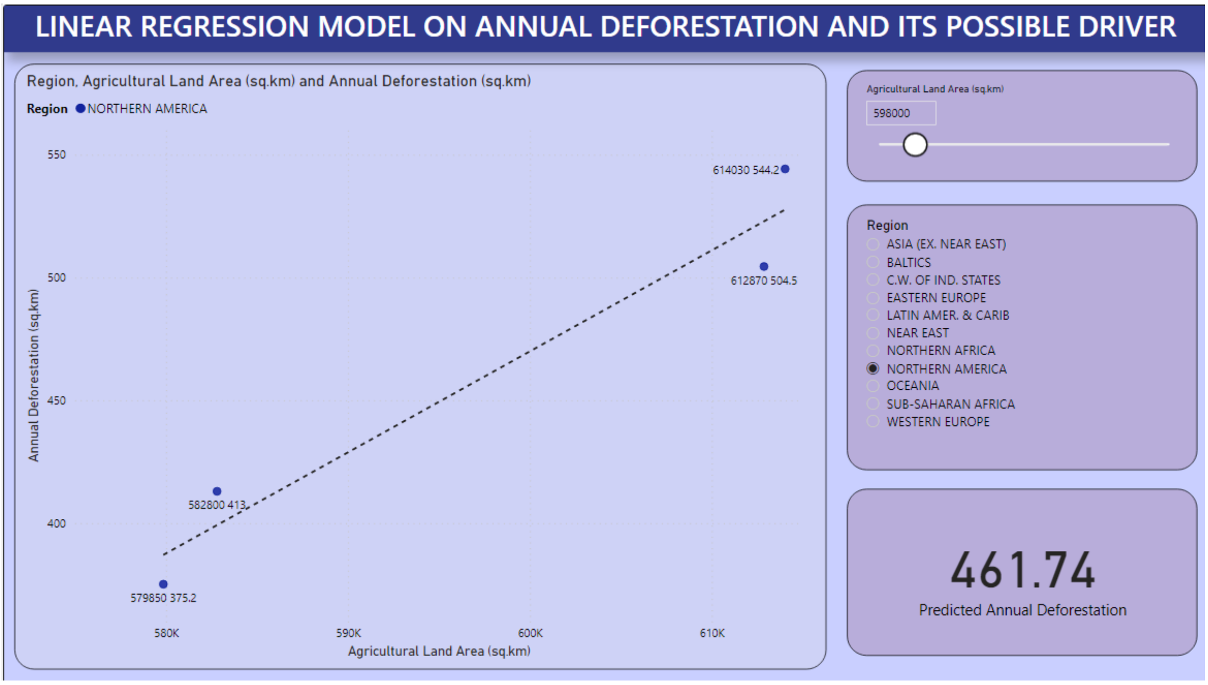
- ○ ASIA (EX. NEAR EAST)
- ○ BALTICS
- ○ C.W. OF IND. STATES
- ○ EASTERN EUROPE
- ● LATIN AMER. & CARIB
- ○ NEAR EAST
- ○ NORTHERN AFRICA
- ○ NORTHERN AMERICA
- ○ OCEANIA
- ○ SUB-SAHARAN AFRICA
- ○ WESTERN EUROPE

Figure 7.2.1(m): What-If Parameter and Slicer

What-if parameter as shown in Figure 7.2.1(m) is utilized as the assumption or condition of the model in order to determine and view the predicted value. When the value of agricultural land area is entered into the what-if parameter, the predicted annual deforestation will appear accordingly in the card visualization. Besides, regions for the linear regression model can be selected by using the slicer. This is because different regions might have different annual deforestation and agricultural land area which will affect the predicted future value of annual deforestation. Therefore, users could view the predicted value of annual deforestation for different regions in order to extract the more accurate results so that different approaches can be made for different regions.

7.2.2 Analysis of Linear Regression Model

From the linear regression model above, there are several visualization and prediction results based on the regions. The visualization and predicted results will be analysed in order to meet the scope and objectives of this project. Firstly, the linear regression model for “Latin America & Carib” as illustrated in Figure 7.2.1(l) is showing that it is positive correlation as it has a positive gradient. This indicates that when the agricultural land area increases, the annual deforestation also increases. Hence, we can know that expansion of agricultural land area most probably is one of main drivers of deforestation based on the plotted graph. Besides, we can know that if the agricultural land area is around 0.5 million, the predicted annual deforestation value will be 5.32k. Based on the prediction result, users are able to know the probable value of annual deforestation so that actions can be made to mitigate or reduce the annual deforestation. Users can also have a better understanding on the severity on annual deforestation which may caused by the expansion of agricultural land area according to the prediction results. Based on the linear regression model, when the value is entered into the what-if parameter which is the agricultural land area increases, the value of annual deforestation will also increase. This indicates that the expansion of agricultural land area is highly affecting the annual deforestation rate. In addition, the agricultural is expanding as the need of global food and energy is increasing rapidly due to population growth of the world. As some of the countries have reserved land area for future development, the agricultural land can only be expanded to the forest area which leads to deforestation.



LINEAR REGRESSION MODEL ON ANNUAL DEFORESTATION AND ITS POSSIBLE DRIVER

Agricultural Land Area (sq.km)
598000

Region
○ ASIA (EX. NEAR EAST)
○ BALTICS
○ C.W. OF IND. STATES
○ EASTERN EUROPE
○ LATIN AMER. & CARIB
○ NEAR EAST
○ NORTHERN AFRICA
● NORTHERN AMERICA
○ OCEANIA
○ SUB-SAHARAN AFRICA
○ WESTERN EUROPE

461.74
Predicted Annual Deforestation

Region, Agricultural Land Area (sq.km) and Annual Deforestation (sq.km)
Region ● NORTHERN AMERICA
550
500
450
400
Annual Deforestation (sq.km)
614030 544.2
612870 504.5
582800 413
579850 375.2
580K
590K
600K
610K
Agricultural Land Area (sq.km)

Figure 7.2.2(a): Visualization on Northern America

For Northern America linear regression model as illustrated in Figure 7.2.2(a), the relationship between agricultural land area and annual deforestation is positive correlation where the gradient is positive. Hence, we know that the expansion of agricultural land area is also affecting the annual deforestation in Northern America based on the graph. Both Northern America and Latin America might have the same issues as the gradient showed in the model is almost the same which is also increasing linearly. In addition, if the agricultural land area is around 600k, the predicted annual deforestation will be around 462 sq.km. Therefore, dedicated organisation or department can limit the agricultural land area to be expanded to mitigate the annual deforestation by using the predicted results.

7.3 Recommendations

7.3.1 Social Issues and Ethical Issues

There are a possibility that social and ethical issues will be raised when data mining is carried out. During the process of the project, we utilized the datasets that engage with the public includes the agricultural land area datasets and so on which may raise some social and ethical issues. Firstly, privacy is one of the main social and ethical issues arises in research or data mining process. For instance, the data collected should always stored in a protected and secure location to avoid data breach and malicious attacks which loses the confidentiality of the data. In our case, all data includes agricultural land area, annual deforestation and many more should protect and handle with care in the data preparation process like data cleaning and data transformation.

Furthermore, another social and ethical issue is that getting permission from the owner or sources before processing the data. In our project, all datasets have the permission to be collected for analysis and modelling as the data collected from the sources are open -source or available for public download. Hence, the possible social and ethical issue can be avoided if actions are taken carefully before conducting the analysis and modelling.

8.0 Conclusion

As deforestation is one of the most concerning global environmental issues, the analysis and modelling in the project will help to determine the relationship between the possible driver and deforestation. Based on the project, agricultural land area is considered as one of the possible drivers that have a large impact on deforestation and forest loss. The objectives and scopes identified earlier are achieved successfully. The public or dedicated department could propose solutions and make decisions in order to mitigate the forest loss and deforestation.

8.1 Workload Matrix Report

Name & TP Number	Tasks
Chan Ming Li – TP060774	• Linear Regression Predictive Model • OLAP
Cheong Jia Yao – TP060827	• Time-series Predictive Model • OLAP
Emily Kuan Yu Xuan – TP060808	• Clustering Model • OLAP
Praveenaraaj Sangar – TP061061	• Distribution Graph • OLAP

8.2 Personal Reflection Report

In a nutshell, I have achieved some learning outcomes by working on this project such as data cleaning, data mining techniques, knowledge discovery and more. Firstly, we have to identify the objectives and scopes based on the domain chosen so that analysis can be conducted well. After identifying the objectives and scopes, data analytics methodologies are researched, and we have selected Fayyad’s KDD as our group’s methodology. I have learned what is Fayyad’s KDD methodology and its phases includes data selection, data prepressing, data transformation, data mining, interpretation or evaluation and applying knowledge gained. The methodology phases are implemented in our group’s project as well. Data understanding, selection, cleaning and transformation has done by using the techniques in Power BI taught in class and also online resources. I have used power query featured by Power BI to clean and transform the null and noisy data in order to improve the data integrity and improve the performance of the analysis.

Moreover, I have learned and implemented OLAP Dashboard and Linear Regression Model in Power BI to achieve my objectives and scopes. After the models are developed, analysis is carried out in order to extract different information which will help users to have a better understanding on forest loss and deforestation and make decisions on the solutions which will mitigate the deforestation rate. In addition, I have researched the social or ethical issues that may faced in our case. In short, the objectives, scopes, learning outcomes are achieved by implementing data analytical skills that I have learned.

9.0 References

Arciniega, D. (17 9, 2021). *How to do Simple Linear Regression in Power BI*. Retrieved from IterationInsights: https://iterationinsights.com/article/linear-regression-in-power-bi/

Brownlee, J. (25 3, 2016). *Linear Regression for Machine Learning*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/linear-regression-for-machine-learning/

Data Science Process Alliance. (2021). *KDD and Data Mining*. Retrieved from Data Science Process Alliance: https://www.datascience-pm.com/kdd-and-data-mining/

O'Connor, E. (26 5, 2021). *CAN I DO DATA CLEANING IN POWER BI?* Retrieved from EPCGroup: https://www.epcgroup.net/can-i-do-data-cleaning-in-power-bi/

Rinkesh. (n.d.). *What is Habitat Loss and Destruction*. Retrieved from Conserve Energy Future: https://www.conserve-energy-future.com/causes-effects-solutions-for-habitat-loss-and-destruction.php

Sharma, R. (23 11, 2020). *KDD Process in Data Mining: What You Need To Know?* Retrieved from upGrad: https://www.upgrad.com/blog/kdd-process-data-mining/#What_is_KDD_in_Data_Mining

Usama Fayyad, G. P.-S. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*, 37 - 47.

Vedantu Learn LIVE Online. (8 September, 2021). *(https://www.vedantu.com/english/environmental-issue-essay)*. Retrieved from (https://www.vedantu.com: (https://www.vedantu.com/english/environmental-issue-essay)

World Health Organization. (2021). *How air pollution is destroying our health*. Retrieved from World Health Organization: https://www.who.int/news-room/spotlight/how-air-pollution-is-destroying-our-health