National College of Ireland

Configuration Manual

MSc Research Project
MSc in Data Analytics

Binu Jemima Christopher Nadar
Student ID: x23323493

School of Computing
National College of Ireland

Supervisor: Dr. David Hamill

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student ID: Binu Jemima Christopher Nadar

Programme: MSc in Data Analytics **Year:** 2024-2025

Module: MSc Research Project

Lecturer: Dr. David Hamill
Submission Due Date: 15th September 2025

Project Title: Can We Predict Statistical Capacity? A Study on Predicting Statistical Performance Indicators

Word Count: 521 **Page Count:** 6

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Binu Jemima Christopher Nadar

Date: 15th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Binu Jemima Christopher Nadar
Student ID: x23323493

1 Hardware Details

The details about the machine used for this project are as follows -
OS: Windows 11
Processer: 12th Gen Intel(R) Core (TM) i5-1235U (1.30 GHz)
RAM: 16.0 GB
System type: 64-bit operating system, x64-based processor

2 Datasets

Two datasets were utilized in this project, both sourced from the World Bank, which provides open access to its data.

- Statistical Performance Indicators Dataset (World Bank Group)
  Author: The World Bank
  Dataset Name: Statistical Performance Indicators
  Dataset Source: https://datacatalog.worldbank.org/search/dataset/0037996/Statistical-Performance-Indicators

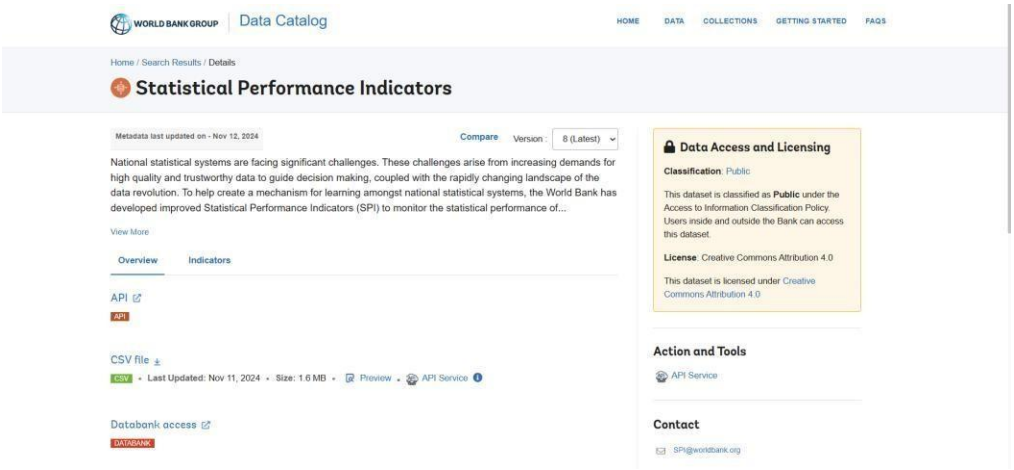

Figure 1: Webpage for downloading Statistical Performance Indicators dataset

- World Development Indicators Dataset (World Bank Group)
  Author: The World Bank
  Dataset Name: World Development Indicators
  Dataset Source: https://databank.worldbank.org/source/world-development-indicators/preview/on

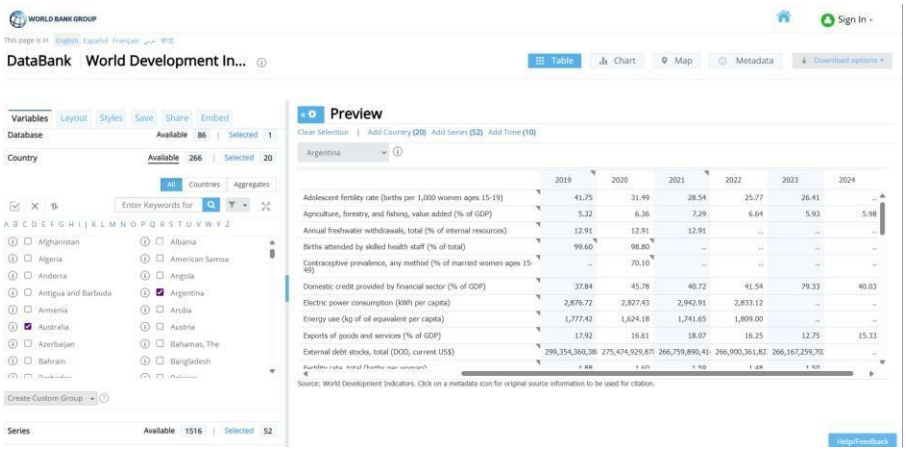

Figure 2: Webpage for downloading World Development Indicators dataset

3 Power BI

In this project, Power BI was utilized to create specific data visualizations.

- To import data, navigate to the Home tab and select 'Get Data' from the Data section.

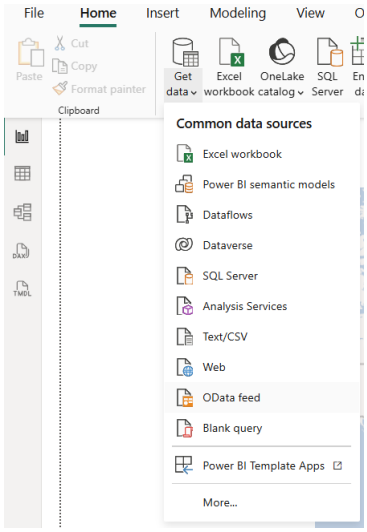

Figure 3: 'Get Data' - Power BI

- For data preprocessing and transformation, go to the Queries section, click on 'Transform data' and then again click on 'Transform data' from the dropdown, which will open the Power Query Editor in a new window. Power Query Editor is where we clean, transform and prepare data for visualizations.

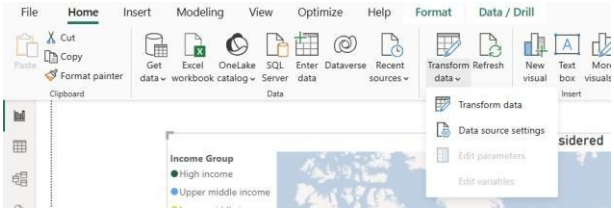

Figure 4: 'Transform data' - Power BI

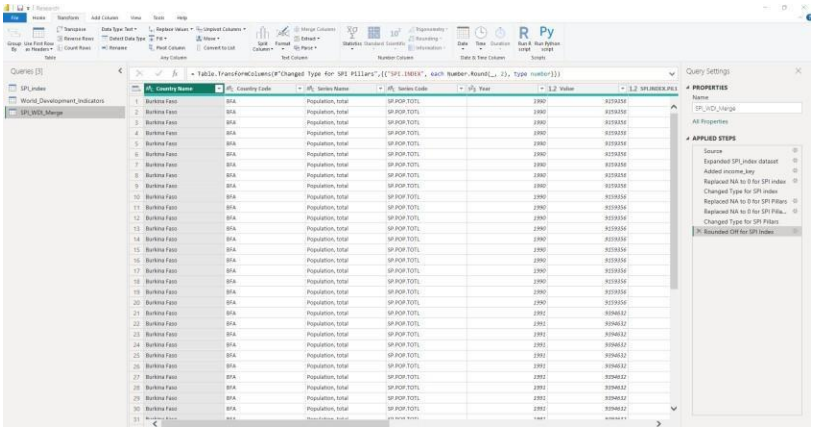

Figure 5: Power Query Editor – Power BI

- 'Measures' are calculations in Power BI that dynamically analyze data based on the current context, such as filters or selections. We create measures by click on 'New measure' in Calculations section under the Home tab.

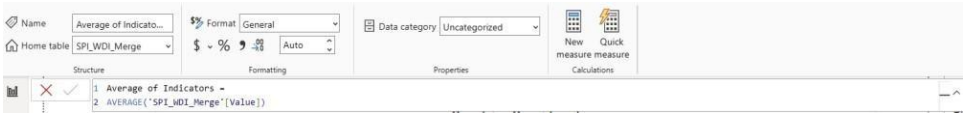

Figure 6: Measure for calculating Average of Indicators

- For data visualization, we use the ‘Visualizations’ pane on the right to select different types of charts and visuals, and then customize them using the ‘Format pane’ or the ‘Build Visual’ pane.

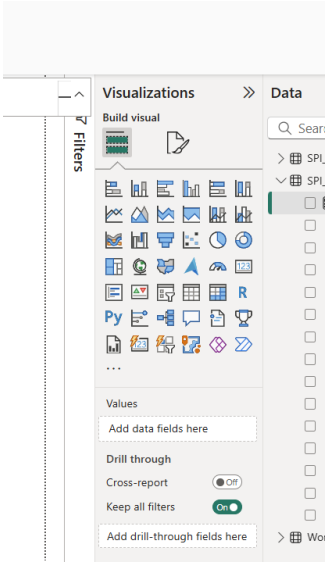

Figure 7: Visualizations pane in Power BI

Figure 8: Map visualization created in Power BI

4 Python & Google Collab

For this project, the Python programming language was used and executed using Google Colab, a free cloud-based service provided by Google.

To access Google Colab, you can either create a new account or simply sign in using your existing Google account.

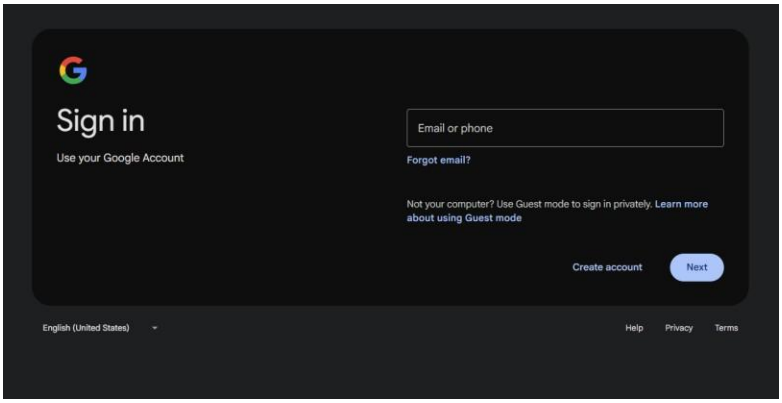

Figure 10: Sign in page for Google Colab

- To import data, the files must first be uploaded to Google Colab before using code to access them.

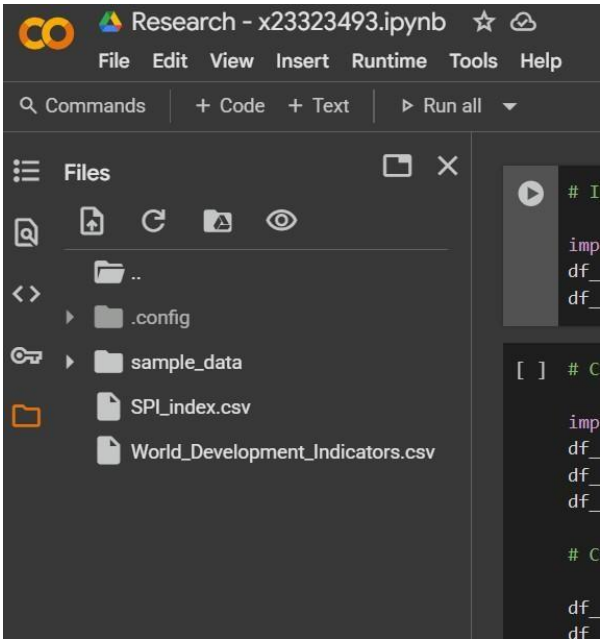

Figure 10: Uploading files in Google Colab

```
# Importing and reading datasets
df_spi = pd.read_csv('/content/SPI_index.csv') # SPI Index Dataset
df_wdi = pd.read_csv('/content/World_Development_Indicators.csv') # World Development Indicators Dataset
```

Figure 11: Using pandas in Python to read the uploaded files

- This project utilized the following libraries and module:
  - **pandas** - Library for data analysis.
  - **numpy** - Library for working with arrays, matrices and linear algebra.
  - **re** - Module for working with regular expressions.
  - **seaborn** - Library for data visualization.
  - **sklearn** - Library for machine learning.
  - **matplotlib** - Library for data visualization.

```python
# Libraries and modules used
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np
import re
```

Figure 12: Libraries and modules used

- Python was also used for data processing, transformation and visualization.

```python
# Removing [YR____] from year columns Eg: '1990 [YR1990]' to just '1990'
df_wdi.columns = [re.sub(r" \[YR\d+\]", "", col) for col in df_wdi.columns]
df_wdi = df_wdi.drop(columns=['Country Code'])
df_wdi.head(5)

# Changing column names
df_spi = df_spi.rename(columns={'country': 'Country', 'date': 'Year'})
df_wdi = df_wdi.rename(columns={'Country Name': 'Country'})
```

Figure 13: Data processing by renaming the columns – Python

```python
# Melting dataframe to convert year columns to rows
new_wdi = pd.melt(df_wdi,
                  id_vars=non_year_columns,
                  var_name='Year',
                  value_name='Value')
```

Figure 14: Data transformation by melting the table – Python

```python
# Plotting the heatmap for missing data
plt.figure(figsize=(20, 10))
sns.heatmap(pivot_year, annot=True, cmap="Reds",  fmt='.0f', cbar_kws={'label': 'Missing Count'})
plt.title('Missing Data per Year')
plt.xticks(rotation=45, ha='right')
plt.ylabel('Countries Considered for the Study')
plt.xlabel('Year')
plt.tight_layout()
plt.show()
```

Figure 15: Data visualization by creating a heatmap - Python

References

World Bank Group. (n.d.). *DataBank - World Development.* Retrieved from World Bank Group: https://databank.worldbank.org/source/world-development-indicators/preview/on

World Bank Group. (n.d.). *Data Catalog - Statistical Performance.* Retrieved from Data Catalog - World Bank Group: https://datacatalog.worldbank.org/search/dataset/0037996/Statistical-Performance-Indicators