**Introduction**

1. **Dataset selection**

We have selected a dataset called [World University Ranking](https://www.kaggle.com/mylesoneill/world-university-rankings) from Kaggle to perform analytics using KNIME. This dataset has 6 tables, however we have shortlisted two that had relational elements and would serve a purpose in our analytics; Education Expenditure (as a percentage of GDP) and Times Data (University Ranking).

The Education Expenditure table contains a list of countries along with their spending on education as a percentage of GDP. This is further bifurcated based on level of education (higher education and elementary education) and type of spending (total, public, private). This is available for multiple years, however, we have focused on the year 2011 as its corresponding data is available in the Times Data table.

The education expenditure table is relevant as it represents expenditure as a percentage of gross domestic product for each country. Times Data table highlights the world university ranking. Times higher education ranking is widely regarded as one of the most influential ratings providers for universities globally. It uses multiple parameters to analyze a university’s performance, which are part of the times data table as well, such as students to staff ratio, percentage of international students, and male to female ratio.

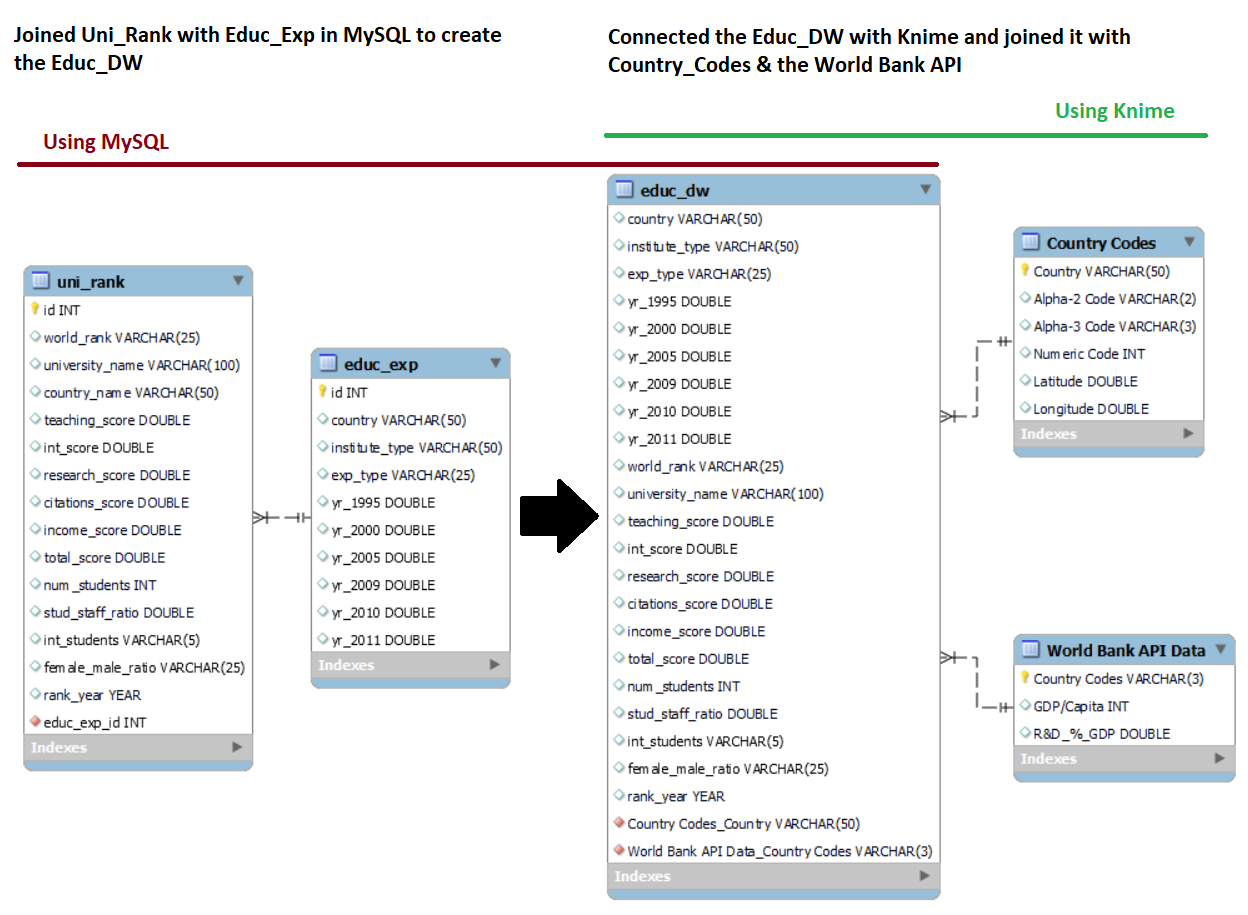
One question these tables help answer is the proportions of public and private spending towards elementary and higher education in the country. We have created an interactive Sun Chart in Knime that helps visualize it easily.

Another question this dataset answers is the link between the percentage of international students in a university and its World ranking. To answer this question, we have created a Bar Chart in Knime.

1. **MySQL Data Warehouse**

The project uses MySQL Workbench to load the Times Data and Education Expenditure data into two separate SQL tables. These two tables are then joined to create the Educ\_DW data warehouse that is linked to Knime for further analysis.

An EER diagram is created based on these linkages, as shown below. The Educ\_DW EER diagram is supplemented by two more tables (Country Codes, World Bank API) that are joined with the Educ\_DW in Knime. Hence, the following EER summarizes the overall relationship between the tables used in the project.



1. **World Bank API**

We have used the World Bank [API](https://data.worldbank.org/indicator/GB.XPD.RSDV.GD.ZS?view=chart) that pulls Research & Development spending of a country as a percentage of its GDP, for the year 2011. The year 2011 is specifically chosen to maintain correspondence with the information available in our main table (World University Ranking). This is the API [URL](http://api.worldbank.org/v2/country/all/indicator/GB.XPD.RSDV.GD.ZS?format=json&date=2011).

The goal of using this API data is to check if there exists any relationship between the university rankings of a country with its R&D spending as a percentage of the GDP.

1. **ETL in KNIME**

The reader will find below all the information regarding each node used by the team. It includes the reason for choosing the node, any configurations made, how is the node serving a purpose in our analytics. Certain nodes are used to construct a set of nodes that form a workflow annotation. Each workflow annotation has a couple of nodes that are part of a process, for example ‘Data Cleaning.’

1. **Reading the dataset from MYSQL**

The first workflow annotation which includes two nodes is the starting point of our ETL pipeline in KNIME. To be able to perform analytics we are using the *MySQL connector*node to connect with MySQL. The next node which is the DB Query Reader enabled us to read the data table (Educ\_DW) from MySQL to KNIME. The screenshot below gives a preview of the data table that was previously joined in MySQL. It includes many relevant columns such as expenditure type, years, and university name but also some poorly formatted information for some variables that we will clean in the Data Cleaning workflow annotation.

Table

Description automatically generated

We have used the *File Reader* node to upload the Country Codes csv file which we obtained from [GitHub](https://gist.github.com/tadast/8827699). This file contains the codes for each country as the ‘Alpha-2 code’ column and we add it as another column in our main data table (Educ\_dw) alongside country names. These were joined using a left outer join.

Table

Description automatically generated

1. **Data Cleaning**

**Daniel, please enter some notes on what cleaning you have done.**

1. **Joiner**

This node joins the Country codes table with the Educ\_DW table sourced from the SQL database. The reason of joining these two tables is to allow us to join the Educ\_DW with the World Bank API. The two tables are joined on the country name using a left outer join.

1. **String Manipulation**

Created a [URL](http://api.worldbank.org/v2/country/all/indicator/GB.XPD.RSDV.GD.ZS?format=json&date=2011) containing the [API](https://data.worldbank.org/indicator/GB.XPD.RSDV.GD.ZS?view=chart) with proper country codes. This node joins together the above joined Educ\_DW table with the API on the 2-digit country codes.

1. **Group By**

This node groups the rows based on the 2-digit country codes and the URL. Following are the grouped columns.



1. **Missing Value**

Since we had a few missing values, we used the Missing Value node to remove the rows that contained missing values.

1. **Get request**

This node loads the API in KNIME. We now have the country codes and the URL for the API for each country.

1. **Container Output**

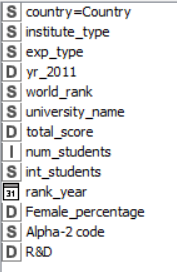
This node is created to be able to view the entire JSON file and if needed, export the loaded data.

1. **JSON path**

This node is used to extract the values of R&D as a percentage of GDP.

1. **Column filter**

This node is used to filter out the columns that are not required for our analysis. Following is the list of columns that we have carried forward.



**Visualization 1: University World Ranking by % of International Students**

The purpose of this workflow is to answer the question of what the link is between the percentage of international students in a university and its World ranking?

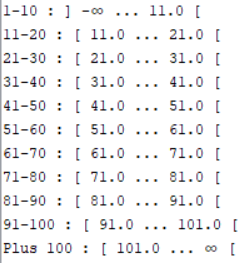
1. **String Manipulation**

The column int\_students was stored as a string variable as it contained the symbol ‘%’ in the values. The first string manipulation node removes the ‘%’ sign and converts the column into an integer variable so that it can be used in visualization ahead.

The second string manipulation node converts the world\_rank column into a double as it is also stored as a string. The purpose of this is to utilize this in the visualization ahead by creating bins.

1. **Numeric Binner**

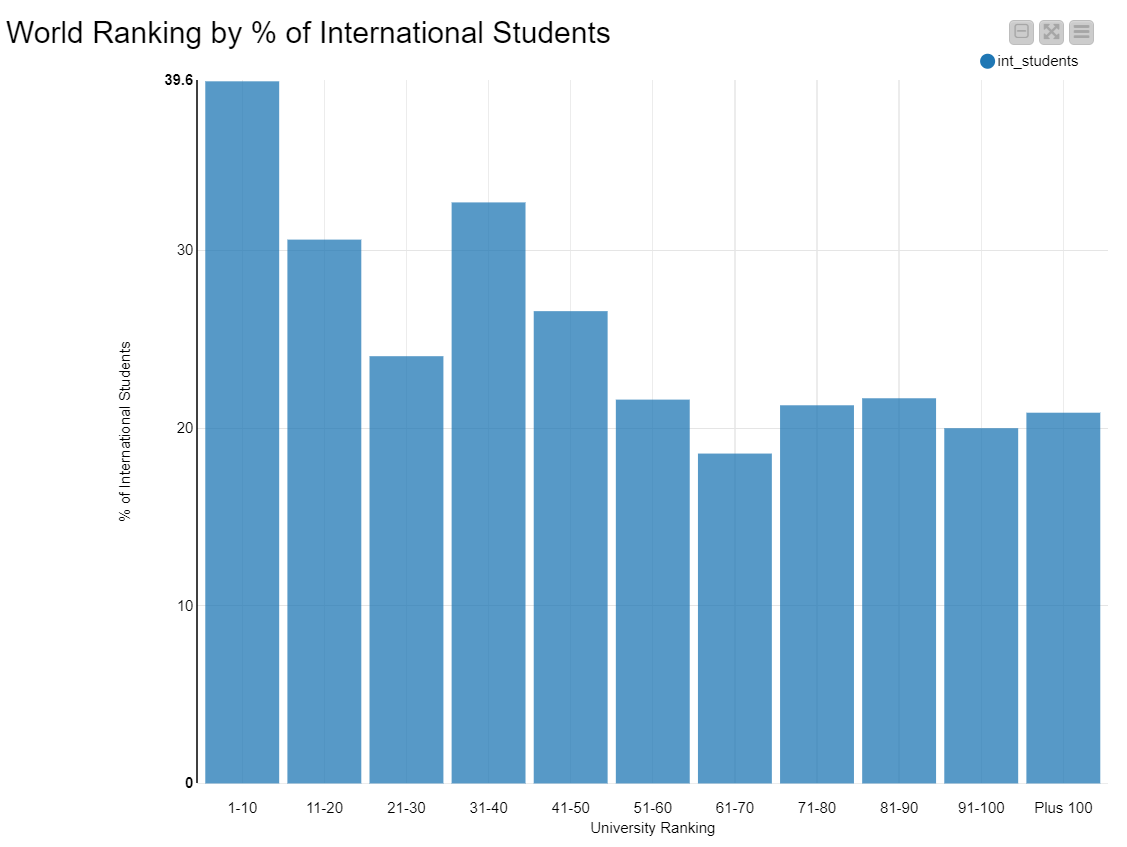
This node is used to create manual bins for the world\_rank column. It is done to make sure the final bar chart is visually appealing to the viewer. Following are the bins.



1. **Bar Chart**

This node creates a bar chart of percentage of international students in a university vs its World ranking.

A general trend is visible from the chart where top universities have more international students compared to low rated universities. However, whether there is a correlation between these two variables cannot be defined from this chart as further analysis will be required to determine such a relationship.



**Visualization 2: Government Spending on Education**

The purpose of this workflow is to answer the question of, depending on the type of institution, what was the percentage of public or private expenditure by each country in the year 2011?

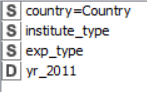
1. **Row Filter**

There are two row filter nodes in this workflow. This first node filters out the rows that contain data for ‘All Institutions’ category and the second node filters out the ‘Total’ category.

The first *row filter* excludes the ‘Total Expenditure’ so we can select either ‘Private Expenditure’ or ‘Public Expenditure’ in the final chart. The second *row filter* node excludes the ‘total’ so we can select ‘Higher Education’ or ‘Elementary or Secondary Education’ in our chart.

1. **Group By**

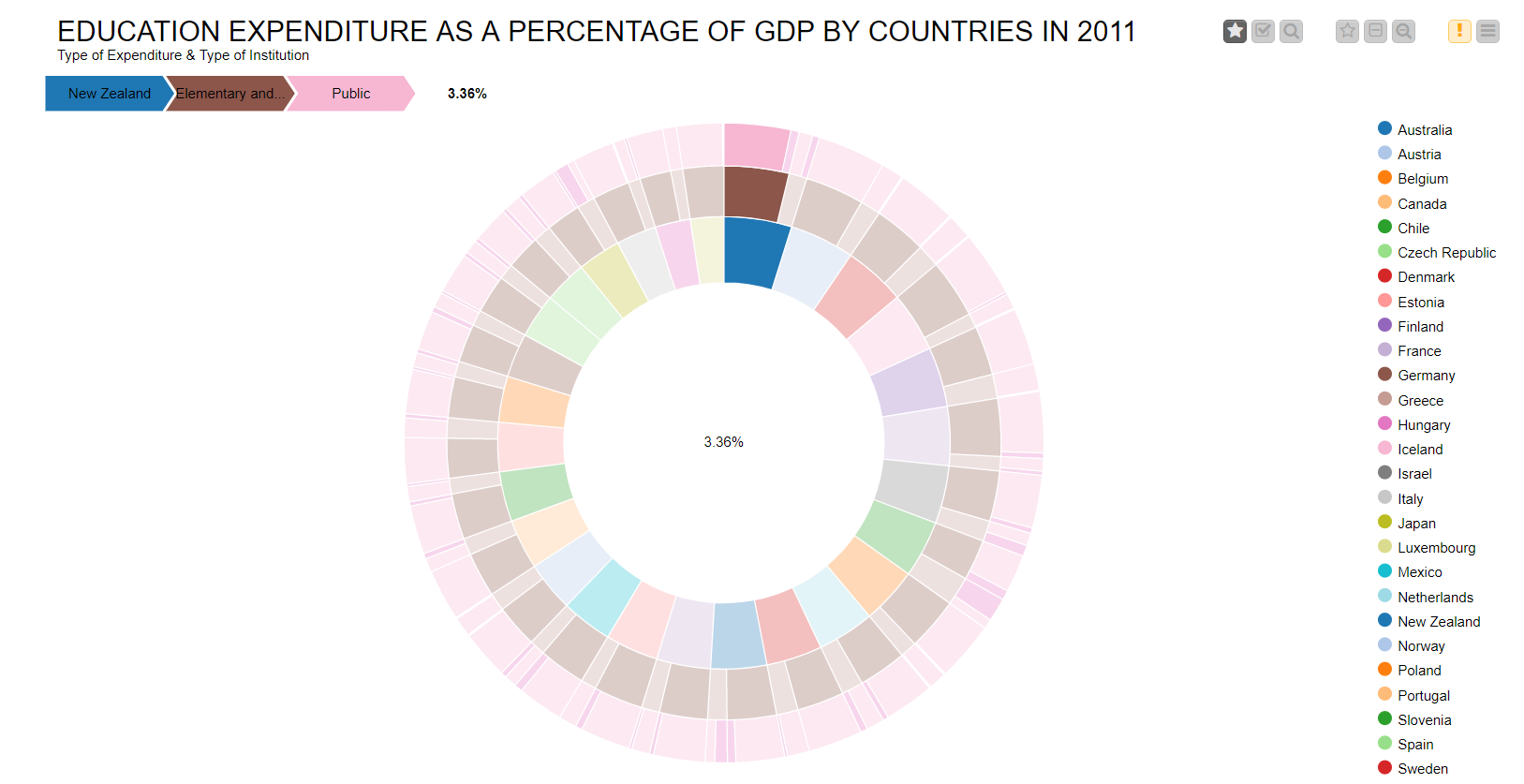
This node groups the following columns for the visualization ahead.



1. **Sunburst Chart**

This node creates a Sunburst Chart to visualize the percentage of public or private expenditure by each country in the year 2011. This is an interactive chart to let the viewer have freedom in bisecting the public vs private spending in elementary vs higher education.

The innermost circle shows all the countries. Starting from the countries that spent the most in a clockwise direction. Then we can select the type of institution and finally either public or private spending as a percentage of the country’s GDP. This is all done by hovering the mouse over each fragment and based on the hover, the results are displayed on the top left corner. For example, the following image shows the results for public spending in New Zealand for elementary schools and secondary schools.

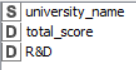


**Visualization 3: Government spending on R&D as a % of GDP to Total Scores of Universities**

The purpose of this node is to answer the question, how are the research and development expenditure of the country and the university score connected?

1. **Group By**

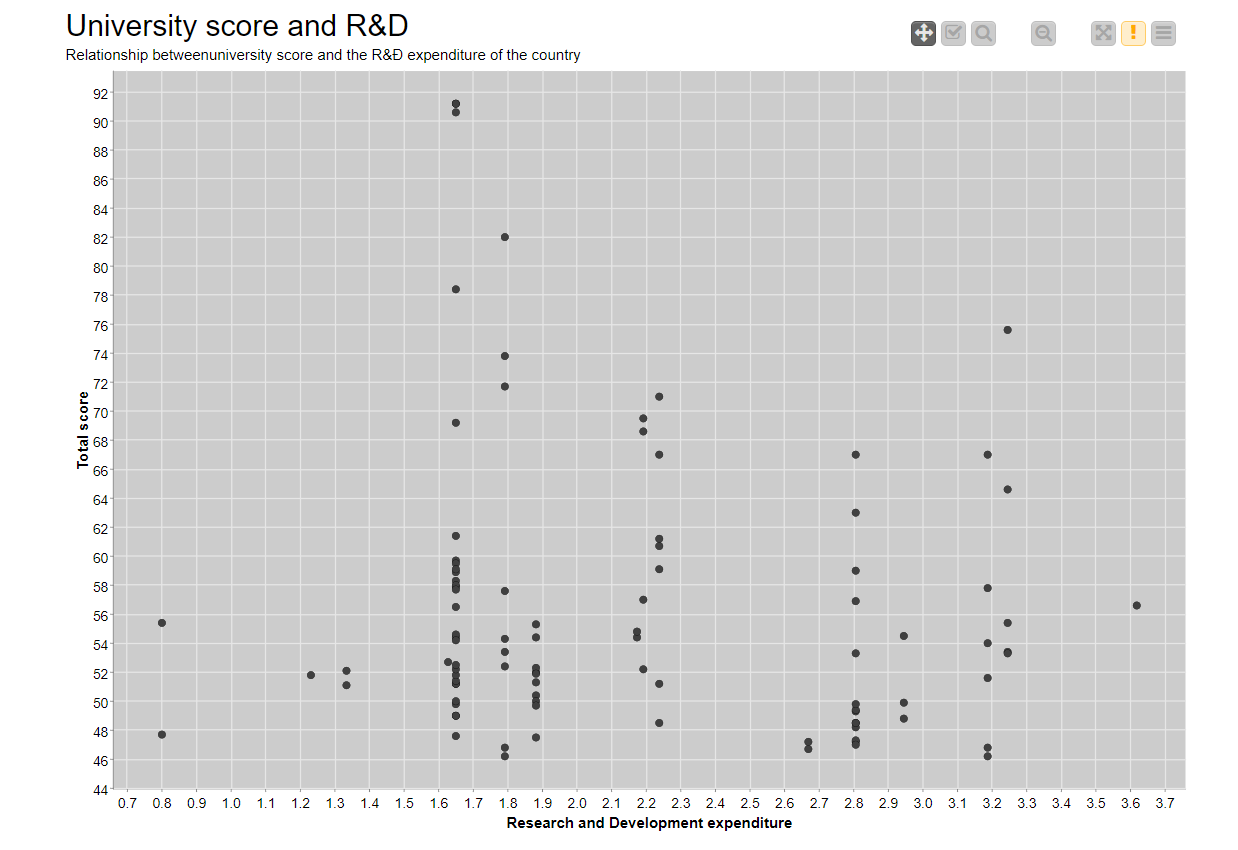
This node groups the columns shown below.



1. **Scatter Plot**

This node is used to create a quick scatter plot of the two above-mentioned variables before running a regression. It seems difficult to extract a clear relationship from the scatter plot as total, hence, we went forward with the regression analysis.

The scatter plot below has Total Score on the y-axis where a higher total score translates into a better ranking. The x-axis gives the R&D expenditure as a percentage of GDP.



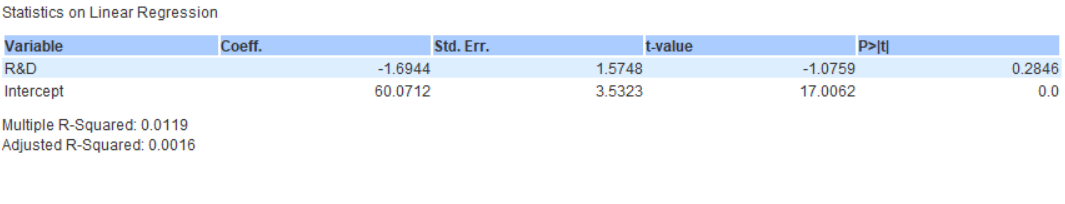
1. **Linear Regression Learner**

We want to see if there is a relationship between a country’s R&D expenditure as a percentage of GDP and the total score received by universities from Times.

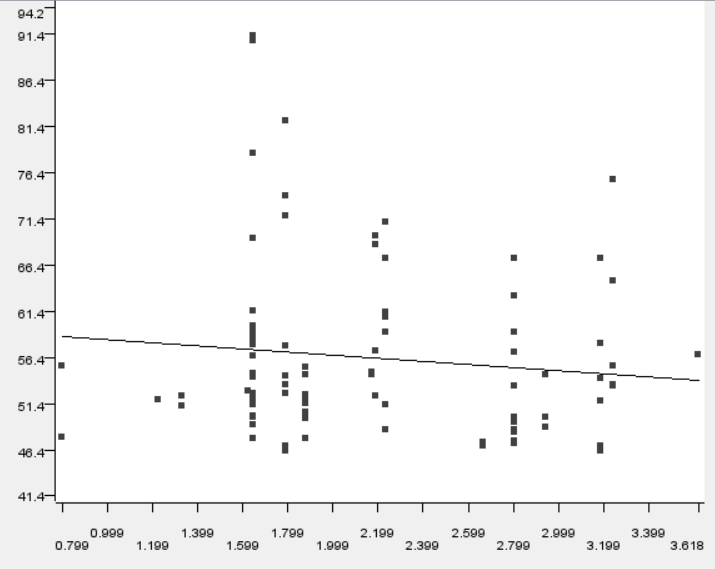
We ran the following equation in the simple OLS regression.

Total score = alpha + beta \* R&D

As we can see from the table, the model’s R-square is low. The beta coefficient is –1.69 which suggest a negative relationship contrary to our hypothesis, however, it is not significant (P-value ~0.28) so we cannot reject that the coefficient is zero. Thus, there is no meaningful relationship between the R&D expenditure as a percentage of GDP and the total score received by a university.



In the regression plot below, we have Total Score on the y-axis where a higher total score translates into a better ranking. The x-axis gives the R&D expenditure as a percentage of GDP.



**Team Members’ Contributions:**

1. **Aftab Alam**

* Loaded the datasets in SQL and created the data warehouse to be used in Knime
* Data visualization number 1: University World Ranking by % of International Students
* Project documentation
* Project compilation and pushing it to Git

1. **Daniel Gyebnar**
   * Data cleaning in Knime to fix the female to male ratio column
   * Project presentation slides
2. **Gyongyver Kamenar**
   * Connected the SQL data warehouse to Knime
   * Found the country codes on Git and connected it to the data warehouse in Knime
   * Found the World Bank API and loaded it in Knime
   * Data visualization number 3: Government spending on R&D as a % of GDP to Total Scores of Universities
3. **Nawal Zehra Hasan**
   * Found the university ranking dataset
   * Connected the SQL data warehouse to Knime
   * Connected the country codes to the data warehouse in Knime
   * Loaded the API in Knime
   * Data visualization number 2: Government Spending on Education