DE2 Term Project

Introduction

The business environment is an integral part of any economy and has a widespread impact on it. In this project, we aim to uncover how business dynamics are constantly changing across various regions in the world, and how the regulatory settings of a country affect the extent of business disclosure in terms of starting and operating a new local firm in a country.

Trying to identify the specifics of the business establishment in different countries can help to form a reliable strategy and an efficient working plan. Opening a new business project must be difficult but could be easier thanks to being aware of the local factors.

The project tends to evaluate the way regions' index changes and whether it is correlated with the costs of business establishment. The index demonstrates a scale of complexity in trying to start a successful business. It can vary depending on time and location. An attempt to consider the index of easiness and future spending on the new project can help to make righteous business decisions and save both time and financial resources as a result. The project reveals the genuine impact of numerous business procedures as well. Their implementation can depend on regional business traditions, too.

To check the trustworthiness of the developed index and relevant findings, it is advantageous to apply the project's implications to the status of the wealthy and other countries. It is significant to understand whether the easiness or complexity of business development is interdependent with the countries' wealth or poverty. Using the projects' results for dealing with poverty issues can help to improve the material progress in general. Project benefits include an objective purpose and a useful tool, which can be helpful in any business sphere.

Project Scope

The objective of this analysis is to find out if starting up a new business is easy across the world by answering the following questions:

1. How ease of business varies across different geographies?
2. How index changes with different regions? And does lower index translate to less time to start a business?
3. How is ease business index distributed across different income groups?
4. How has the ease of business index changed over the years for each region?
5. What are best and worst performing countries by ease of business improvement?
6. Do countries with higher GDP per capita have lower cost of procedures?

All team members contributed to all parts of the project; however, each member was responsible for a part.

Data source – Mariam Bazzi

Data cleaning – Karola Takacs

Data analytics – Tamas Stahl

The Dataset

The dataset combines key statistics to evaluate the business environment in a country. First, we would like to describe some of our variables. Ease of doing business is an index published by the World Bank. It is an aggregate figure that includes different parameters which define the ease of doing business in a country.

API

- GDP per capita (PPP) (World Bank)
- Countries (World Bank)

Data (World Bank) - *Downloaded on 5 December 2020*

- Ease Business Index
- Ease Business Indicator

Data Cleaning & Transformation

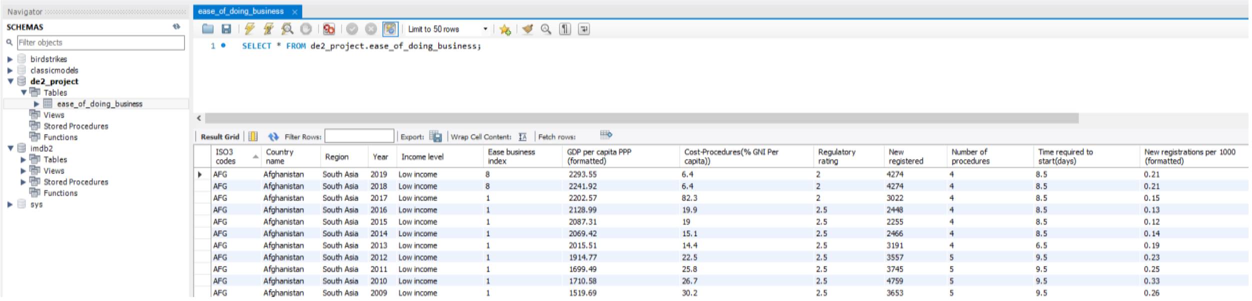
In order to clean the data and prepare it for analysis, we have applied data transformations of the following types:

- Column filter
  - First step was to remove duplicate columns that resulted from the joins of the 4 data sources. These were mostly the ones related to country codes and country names, and indicator codes and indicator names.
  - We also used this node to exclude the years prior to 2009 from the Ease-business-index.csv: there were no records earlier than 2005 and there were more missing values before 2009.
- Row filter
  - The World Bank dataset contained a lot of aggregate regions, therefore, they had to be excluded in order not to have some countries counted multiple times. The column value matching criteria was to test the region column and if it equals to 'Aggregates' then exclude those rows.
  - Another bulk filtering was applied to remove observations where the ease business index had missing values for a country. This was necessary for the simple reason that our research questions were mainly focused on this variable.
  - We have noticed that Switzerland had an Ease business index of 0, which is most probably an error in the dataset (we would expect it to have a higher index), due to failure in data collection or a typing error from 10. Since we cannot source the error or fix it, we have decided to filter this country out.
- The Column rename node was used to have all the column names unified.
- With the help of the Column resorter node, we could create a more logical column order that was useful before writing the cleaned data table to the SQL database. The columns' orders were also re-arranged as they were mixed up as a result of some transformations. For example, GDP PPP – variable was ordered last in the column's arrangement, but this was an important variable.
- String to number conversion was applied to columns Year and GDP PPP.

- After having all necessary variables in integer or double data type, we applied the Number formatter node for rounding up the numbers to 2 decimal places. This was also a concern in terms of readability.
- String replacer node was used seven times for countries we could identify have a somewhat reverse order in their names such as Venezuela, Rep. To Venezuela, or are too long and can be shortened, for example, Russian Federation to Russia).

Storing the data

After cleaning and transforming, data was persisted in SQL. We have decided to proceed with SQL rather than NoSQL since we were not working with an enormous dataset. Our final cleaned dataset was table structured and was easily written to the database in MySQL Workbench through a MySQL Connector, a DB Table creator and populated with DB Insert node. However, the schema had to be created in MySQL using the 'CREATE SCHEMA de2_project' statement, prior to executing those nodes in KNIME.

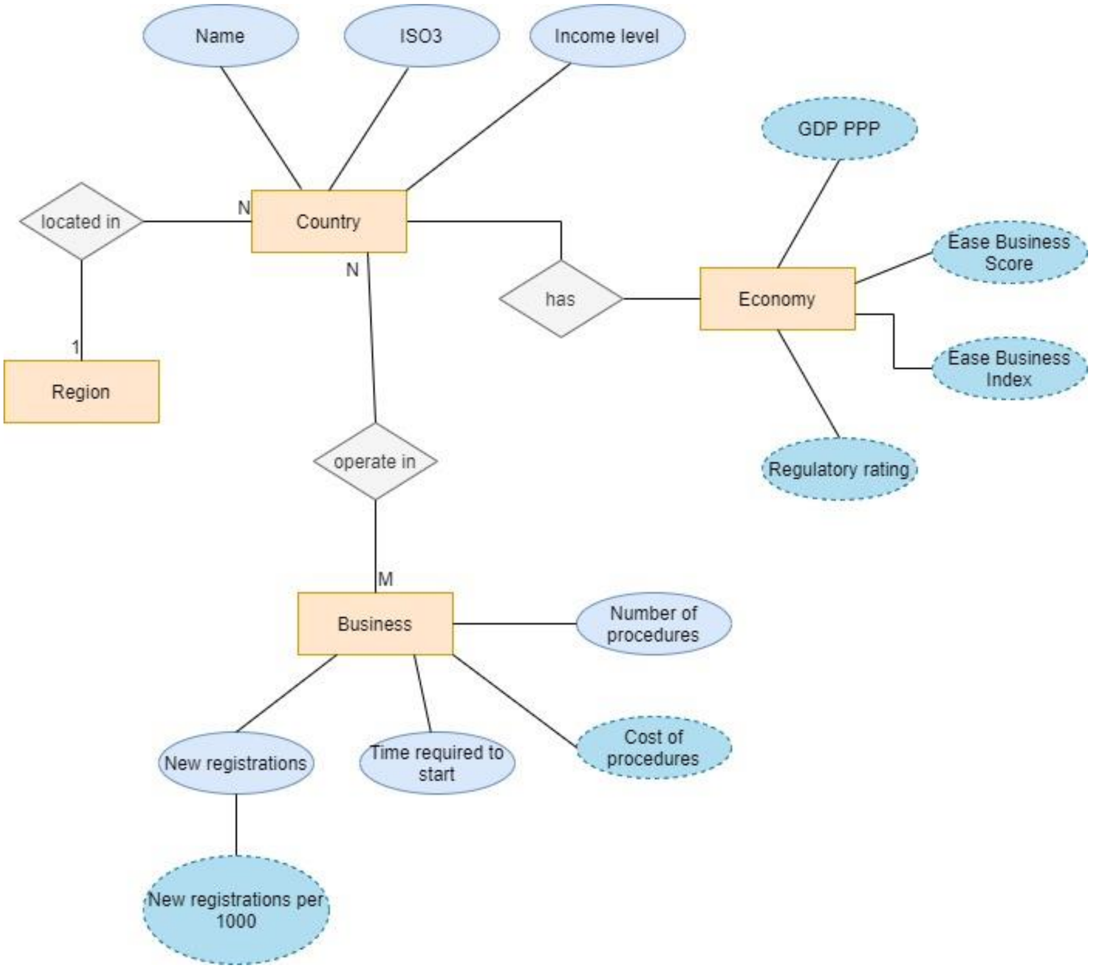


```
SELECT * FROM de2_project.ease_of_doing_business;
```

ISO3 codes	Country name	Region	Year	Income level	Ease business index	GDP per capita PPP (formatted)	Cost-Procedures(% GNI Per capita))	Regulatory rating	New registered	Number of procedures	Time required to start(days)	New registrations per 1000 (formatted)
AFG	Afghanistan	South Asia	2019	Low income	8	2293.55	6.4	2	4274	4	8.5	0.21
AFG	Afghanistan	South Asia	2018	Low income	8	2241.92	6.4	2	4274	4	8.5	0.21
AFG	Afghanistan	South Asia	2017	Low income	1	2202.57	82.3	2	3022	4	8.5	0.15
AFG	Afghanistan	South Asia	2016	Low income	1	2128.99	19.9	2.5	2448	4	8.5	0.13
AFG	Afghanistan	South Asia	2015	Low income	1	2087.31	19	2.5	2255	4	8.5	0.12
AFG	Afghanistan	South Asia	2014	Low income	1	2069.42	15.1	2.5	2466	4	8.5	0.14
AFG	Afghanistan	South Asia	2013	Low income	1	2015.51	14.4	2.5	3191	4	6.5	0.19
AFG	Afghanistan	South Asia	2012	Low income	1	1914.77	22.5	2.5	3557	5	9.5	0.23
AFG	Afghanistan	South Asia	2011	Low income	1	1699.49	25.8	2.5	3745	5	9.5	0.25
AFG	Afghanistan	South Asia	2010	Low income	1	1710.58	26.7	2.5	4759	5	9.5	0.33
AFG	Afghanistan	South Asia	2009	Low income	1	1519.69	30.2	2.5	3653	5	9.5	0.26

ER Diagram

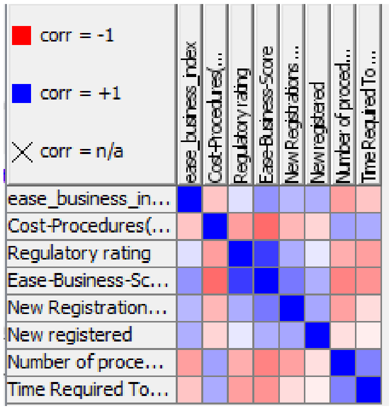
Identified entities in this data are region, country, business and a country's economy. Attributes with the dashed line represent derived variables. We could say that a country is in a certain region and can be characterized by its name, iso3 code and income level. There can be many businesses operating in a country with characteristics of the number of registrations, the time needed to set up a business, how many procedures/steps are needed before starting-up a new business etc. Our variables connected to economy are all derived ones since they are the outcome of various other measures.



Analysis

In our analysis, we would like to answer our questions asked under point Project scope with the visualization tools provided in KNIME.

Correlation matrix



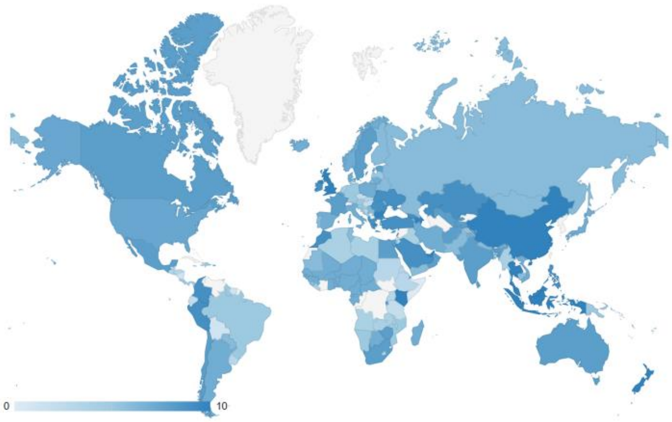
We were interested in the correlation (negative or positive) between our numerical variables. As you could see below in Figure X that the highest correlation is between Ease of Business Score and Regulatory rating with 0.76. Besides that, Ease of Business Score and Cost of procedure are significantly negatively correlated with a value of –0.58.

Dashboard

In order to visualize our findings more easily we decided to prepare a dashboard, as after some research it turned out that it is relatively easy to prepare such in KNIME.

1. *How ease of business varies across different geographies?*

For this question the most logical form of visualization is a map. For this we downloaded an extension node called Choropleth World Map (please do download it if you do not have the node, as some error messages would be given otherwise). Also, we chose the year 2019 as we would get the most information from the most recent numbers. For external validity other years could be filtered all the way to 2009. Although it should be noted that this index does not change frequently. Using this interactive world map in KNIME we can see that China, the UK, Thailand, Malaysia, Indonesia and New Zealand are the countries where the business environment is most incentive.

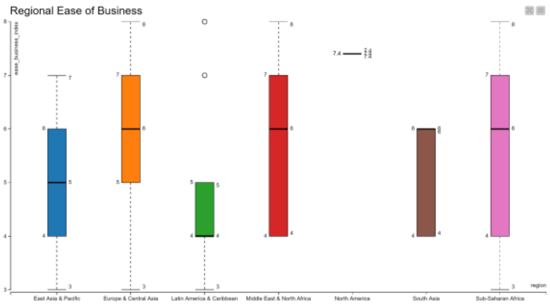


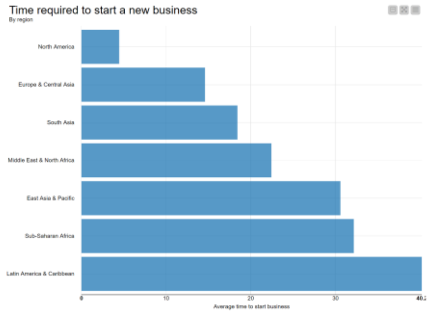
2. *How index changes with different regions? And does lower index translate to less time to start a business?*

The easiest way to visualize the first part of the second question is to use a box plot. For this we created an interactive conditional box plot, where the x-axis shows the regions, and we can change our y variables, which are the following: Ease of Business Index, Cost of Procedures, Time required to Start a New Business and Number of procedures. The best performing region is North America based on the average score of the countries in that region with an average Index of 7.4.

On the box plot the outliers and extreme values are visualized with an ‘o’ and an ‘x’ respectively, therefore in order to have more meaningful plots we removed most of the outliers. For this removal we downloaded the extension node called Outlier Removal. In this node we could easily set the range for the outliers to be removed.

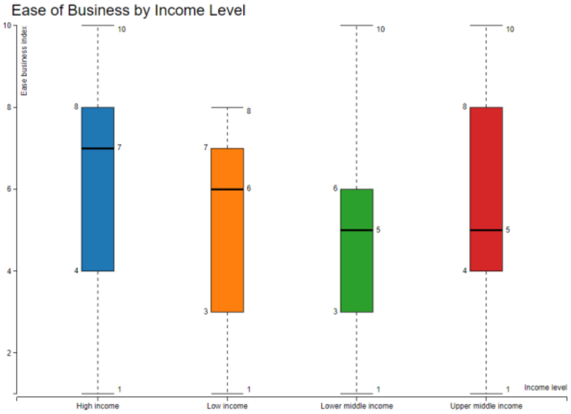
For the second part we also included a bar chart, which shows the average number of days required to start a new business by region in an ascending order. Again, it is not so surprising that North America is the best again, with just an average of 4.5 days. Although, we must mention that the North American region is made up of only 2 countries compared to the 47 countries mentioned in this report for the region Europe & Central Asia. From a larger set of countries, it is easier to have underperforming ones which would significantly increase the average for the region.





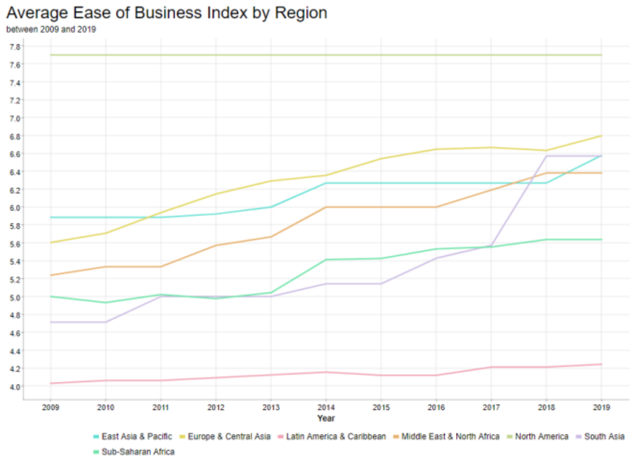
3. How is ease of business index distributed across different income level groups?

Based on the below plot we can spot immediately that for low-income countries the index values only span from 1 to 8 while others also have values 9 and 10. The middle values in the dataset for both lower-middle and upper-middle income countries are 5, but with a different skewness: lower income countries' distribution is more skewed to the left while upper-middle countries' rather to the right. It is also clear that upper-middle income countries resemble high-income countries having the same middle number for Q1 and Q3 but the overall median is higher by two for the high-income countries.



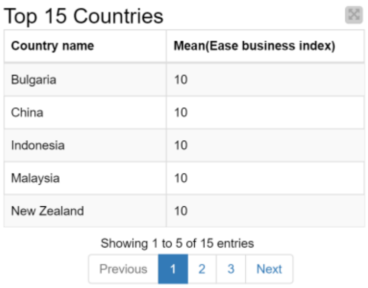
4. How has the ease of business index changed over the years for each region?

As mentioned previously average is not always the best summary statistic to use since it masks the differences within groups but in this question, we decided to focus on the region aggregate. Since North America has only two countries in its group, and their score is not changing, that stays constant across the years. Most of the regions have an upward movement signaling that their indexes are increasing. The line with the biggest difference between the starting- and ending values refers to South Asia. This was possible since several of the countries belonging to this region increased their index to 8 (e.g. Afghanistan from index 1 to 8) over the years. Europe & Central Asia took the leading position early in 2011 and could retain it so far, but East Asia, Middle East & North Africa are closing the gap. The Caribbean region has not been improving a lot in the past 10 years and is a definite laggard.



5. What are best and worst performing countries by ease of business improvement?

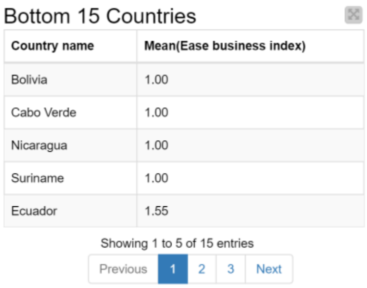
Do countries with higher GDP per capita have lower cost of procedures? To answer this question, we simply listed the 15 best performing and worst performing countries by calculating the mean of the ease of business index over the last 10 years. China, Singapore and New Zealand were among the top; meaning that they have less strict business regulations.



Top 15 Countries

Country name	Mean(Ease business index)
Bulgaria	10
China	10
Indonesia	10
Malaysia	10
New Zealand	10

Showing 1 to 5 of 15 entries



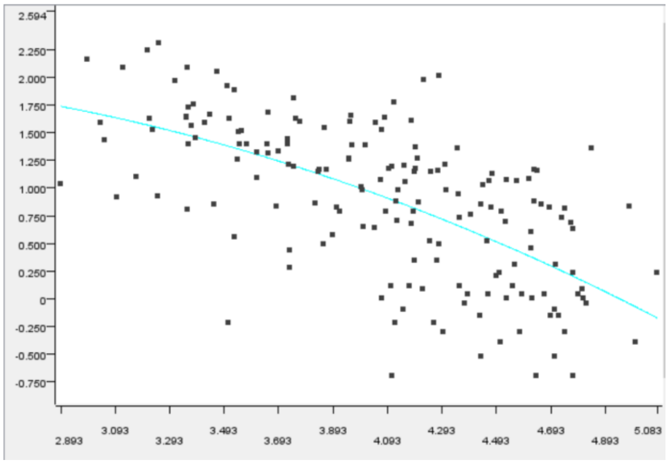
Bottom 15 Countries

Country name	Mean(Ease business index)
Bolivia	1.00
Cabo Verde	1.00
Nicaragua	1.00
Suriname	1.00
Ecuador	1.55

Showing 1 to 5 of 15 entries

6. Do countries with higher GDP per capita have lower cost of procedures?

The scatterplot is a great way to visualize the correlation between GDP per capita and Cost of Procedures to start a business. We have transformed the GDP and the GNI to its logarithmic forms to uncover a better association between those 2 variables. AS we can see countries with higher GDP tend to have lower cost of procedures (x axis – log_GDP , y axis log_cost of procedures).


2.594
2.250
2.000
1.750
1.500
1.250
1.000
0.750
0.500
0.250
0
-0.250
-0.500
-0.750
2.893
3.093
3.293
3.493
3.693
3.893
4.093
4.293
4.493
4.693
4.893
5.083