EDA for Ease of Doing Business by Quality of Government

Karmanya Pathak

2020-09-07

Loading the packages

Research Question

The QoG dataset explores the quality of government data for the year 2020 from multiple data sources. From multiple categories and data sources, the focus of this research and report is on the Ease of doing business.

The research question for this project would be: "How do business regulations and enforcements across various economies affect their respective ease of doing business?"

Describe Data

Out of 1758 variables, this research focuses on 12 business factor varables and 2 variables identifying the economy or the country.

The following 2 dependent variables represent the country name and their respective country code to analyze their ease of doing business:

1. cname
2. ccode

The following independent variables are considered for analysing ease of business based on their rating system and the respective scores achieved by each country for the duration of the study:

1. eob_dcp16: Dealing with construction permits 1
2. eob_ec16: Enforcing contracts 2
3. eob_eob16: Ease of doing business score global 3
4. eob_gc15: Getting credit 4
5. eob_ge16: Getting electricity 5
6. eob_ldri: Land dispute resolution index 6
7. eob_pmi15: Protecting minority investors 7
8. eob_pt17: Paying taxes 8
9. eob_ri15: Resolving insolvency 9

10. eob_rp16: Registering property 10
11. eob_sab: Starting a business 11
12. eob_tab16: Trading across borders 12

The score for each of the above independent variables ranges from 0 to 100, where 0 represents the worst regulatory performance and 100 the best regulatory performance.

The dataset has 194 rows and 14 columns.

Numeric Variables: 13

ccode, eob_dcp16, eob_ec16, eob_eob16, eob_gc15, eob_ge16, eob_ldri, eob_pmi15, eob_pt17, eob_ri15, eob_rp16, eob_sab, eob_tab16

Categorical Variables: 0

Factor variables: 1

cname

Total of Missing values: 101 which is 3.718704 percent of data.

After dropping rows with missing values, the new total number of rows are 181

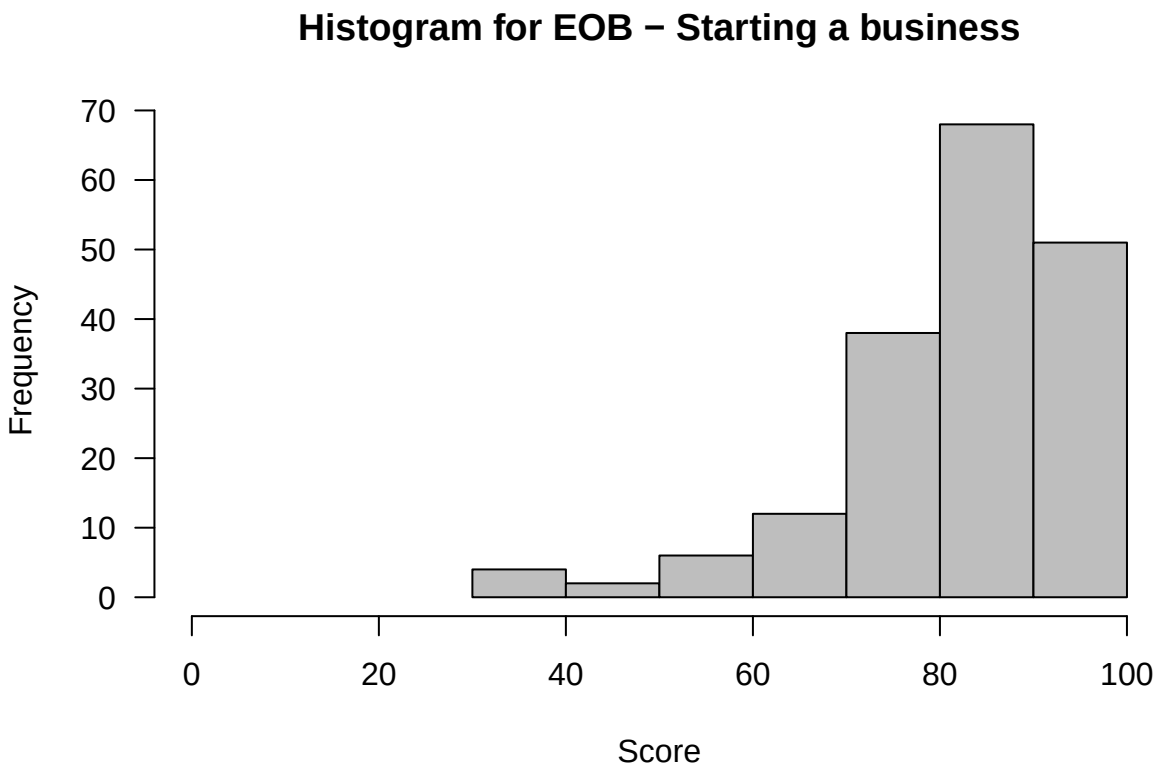
Since only 3.718704% of the observations have missing values, there is significant data to conduct further research by loading the data frame without missing values into a new data frame.

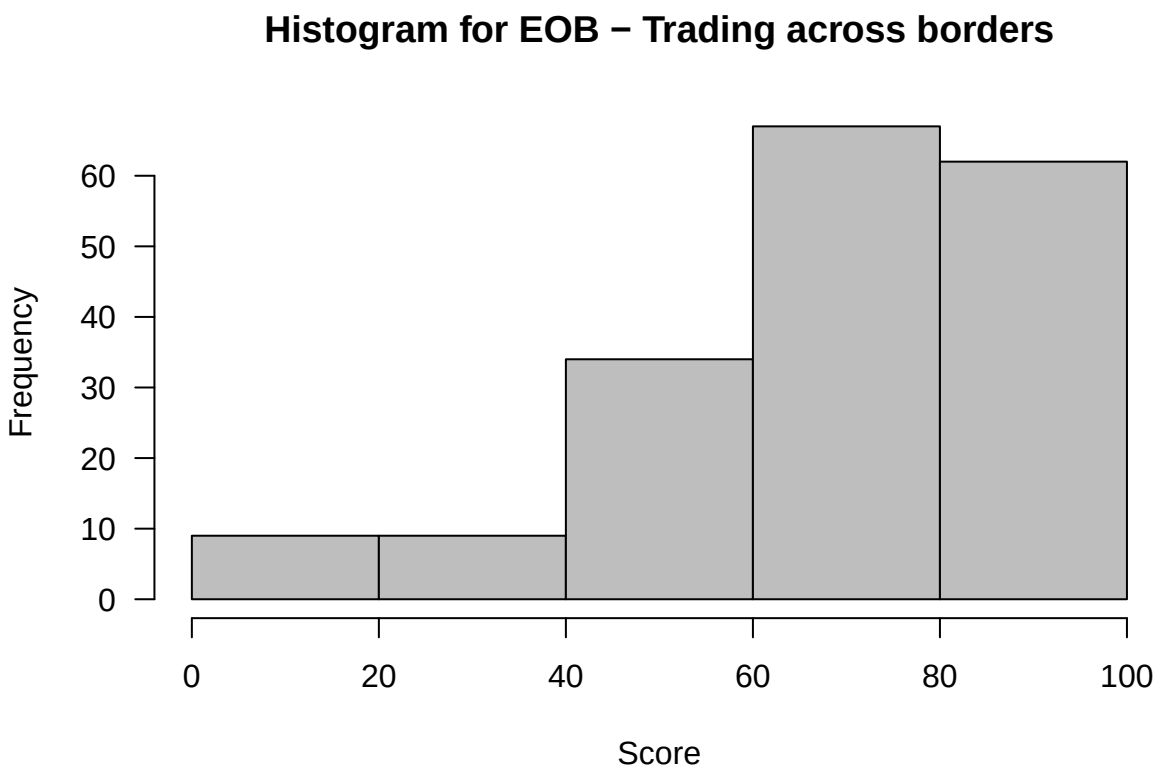
The dataset has 181 countries.

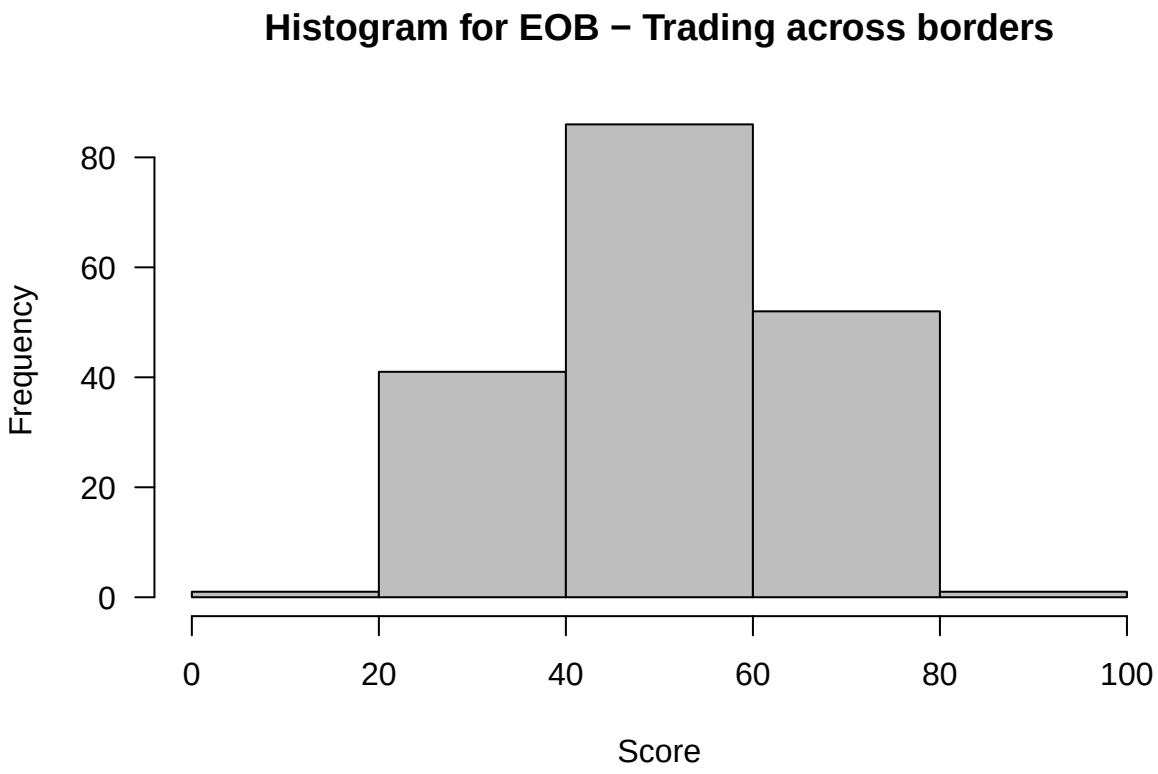
Exploratory Data Analysis

Within the scope of this report, only few variables are chosen for exploration.

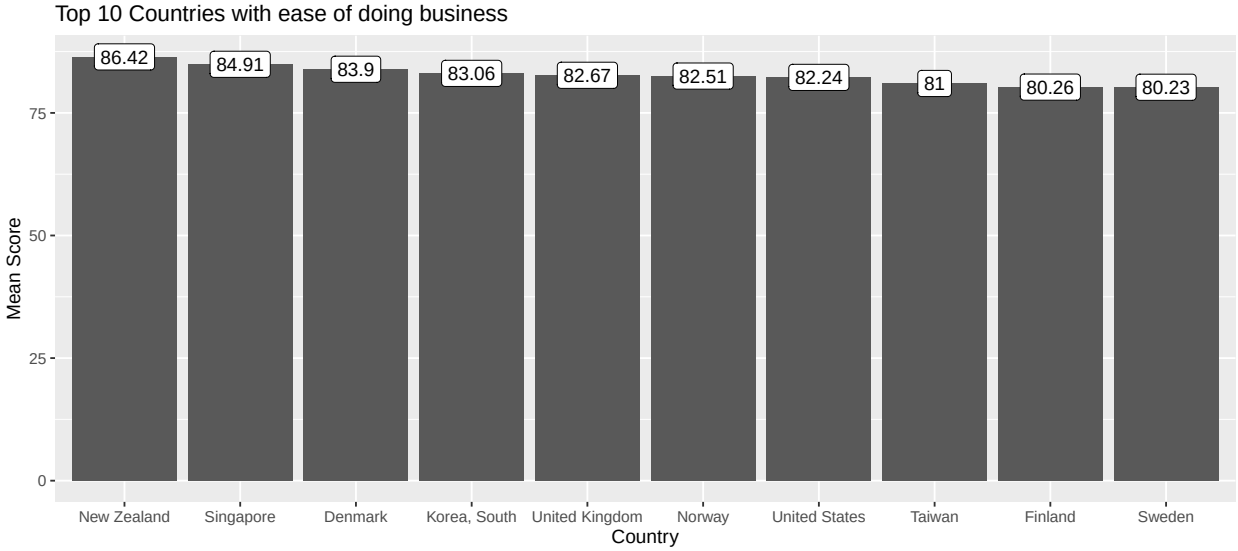
The following graphs show the score frequency for starting a business and trading across borders.


Histogram for EOB − Starting a business
Frequency
70
60
50
40
30
20
10
0
0
20
40
60
80
100
Score

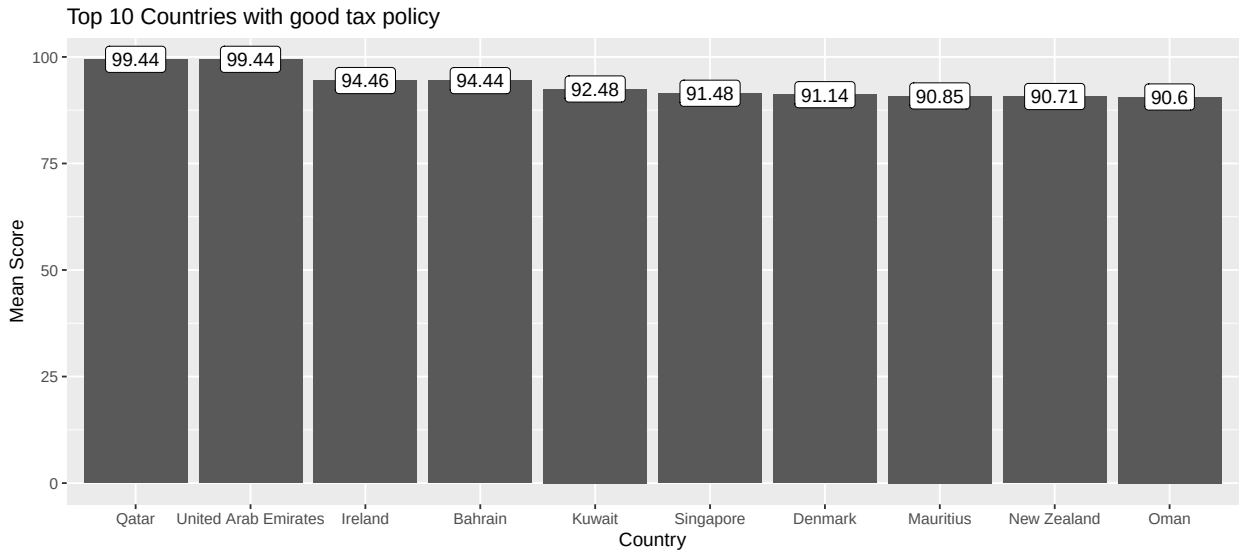

Histogram for EOB − Trading across borders
Frequency
60
50
40
30
20
10
0
0
20
40
60
80
100
Score



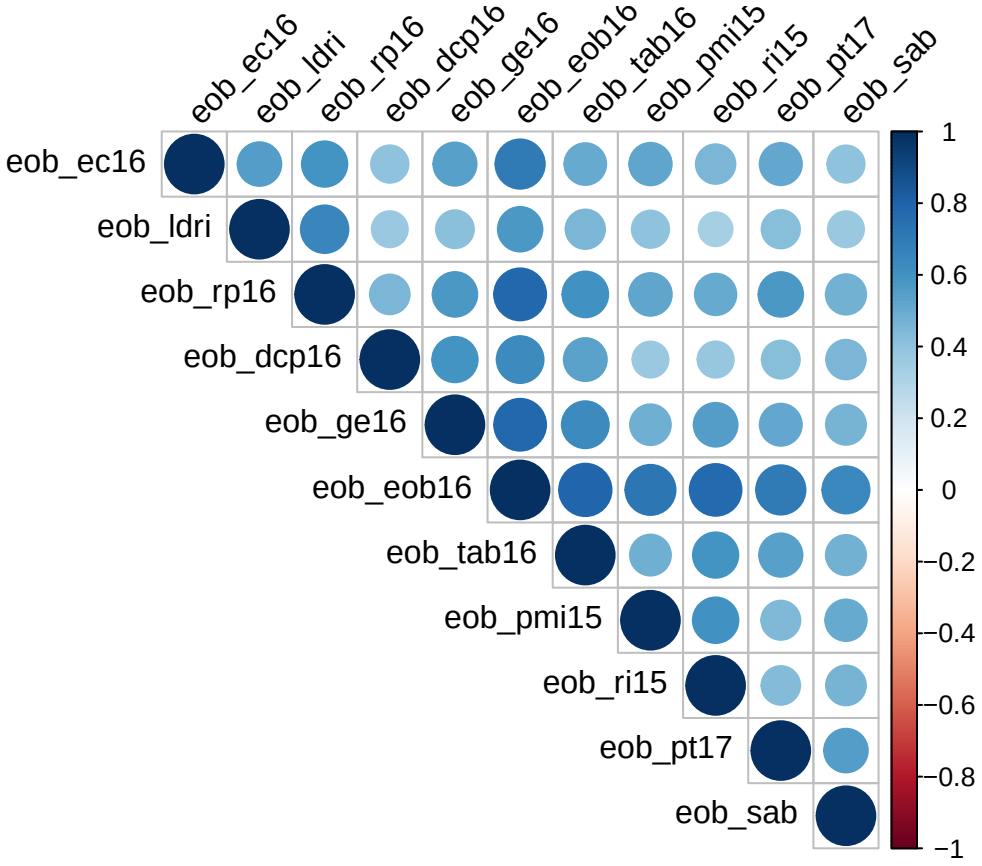
The following graph shows the top 10 countries by ease of doing business.



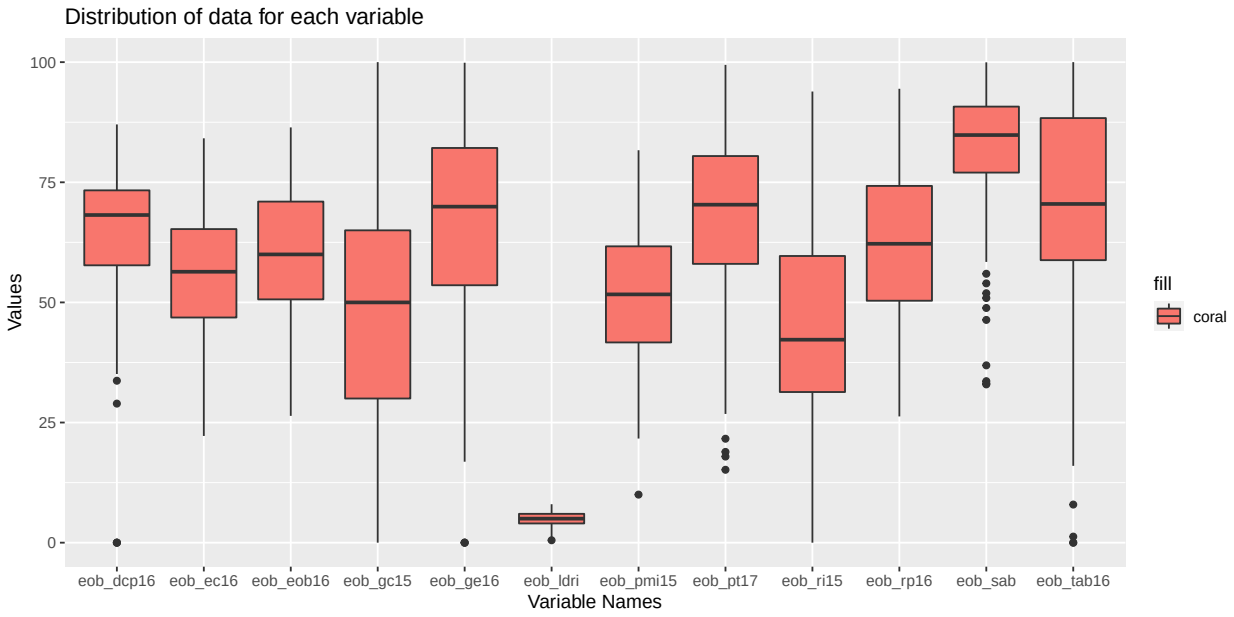
The following graph shows the top 10 countries that have good policies for taxes.



The following plot shows the correlation between the variables



The following graph shows the distribution of data for each variable.



The data exploration reveals that there are no missing values and the variable eob_ldri Land dispute resolution index has a different data range.

This column will not be used in future analysis in this research.

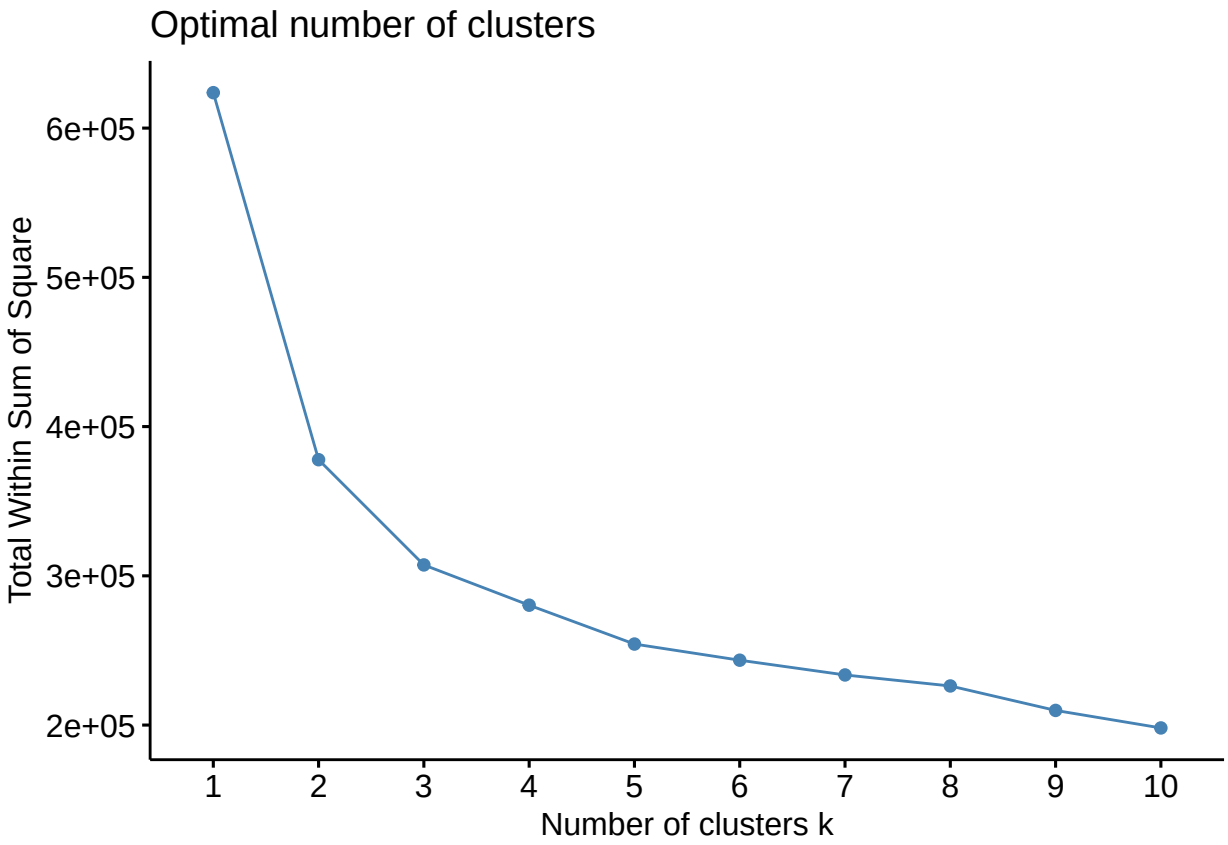
Also, variables eob_sab, eob_pt17, eob_tab16 and eob_dcp16 have significant frequencies of outliers and decreasing in the said order.

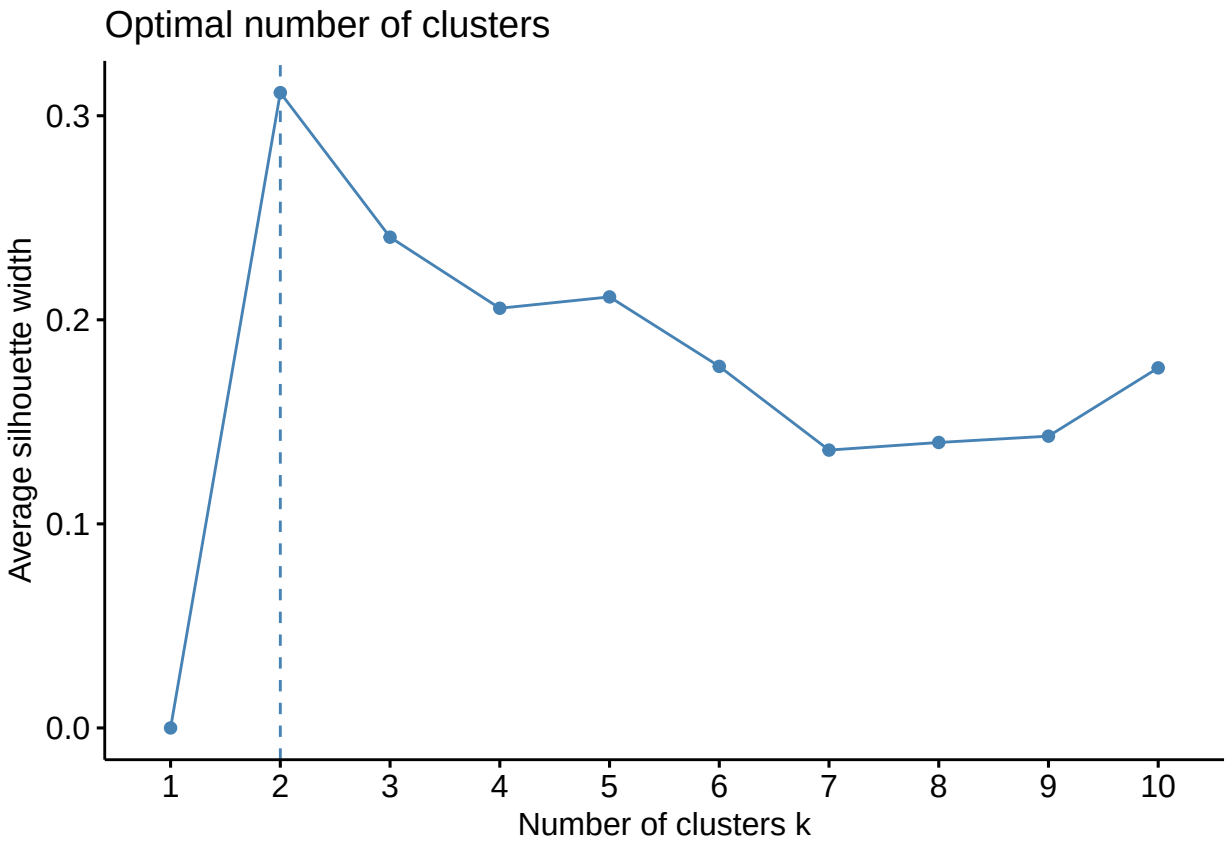
Clustering

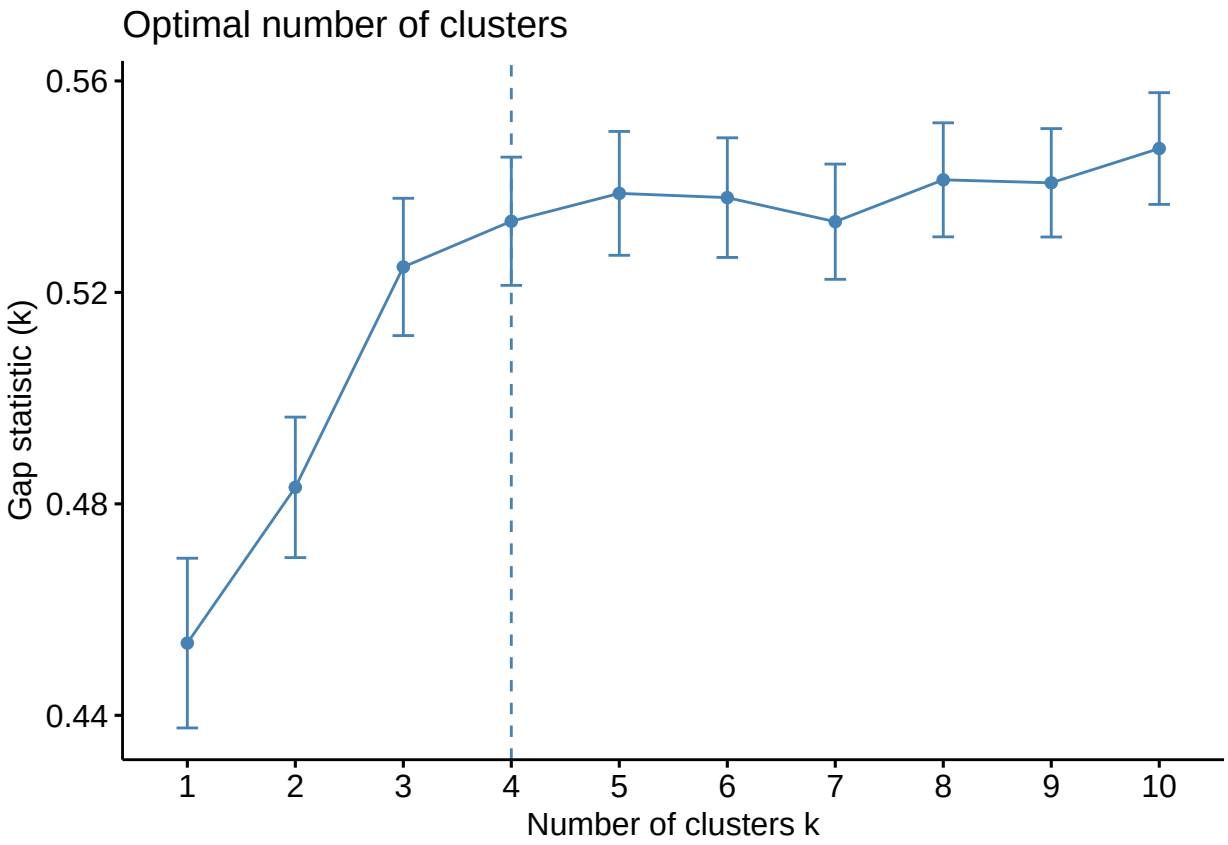
The aim of cluster analysis is to group countries with similar scores of ease of business and to see which countries fall in each group.

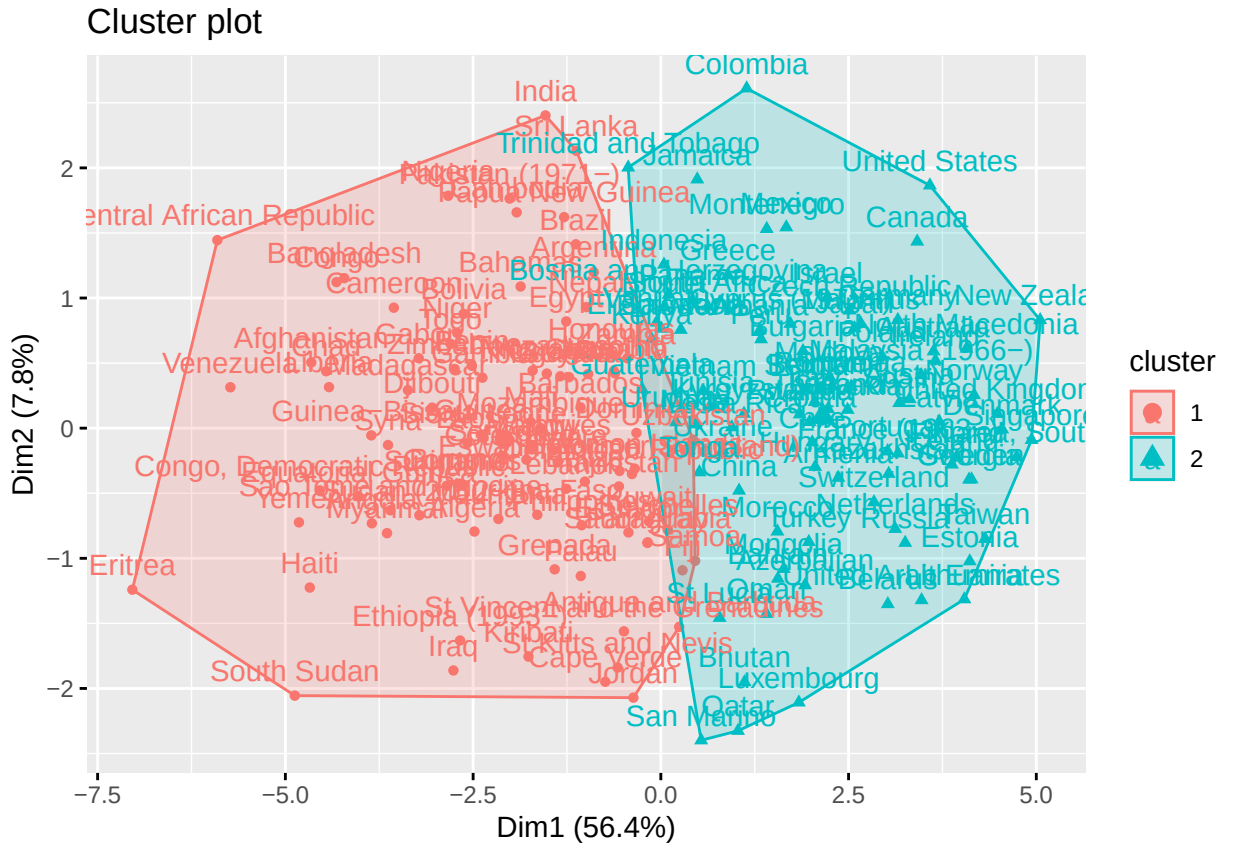
Since the data is already standardized on a scale, there was no need to scale the data using scale().

Using elbow, silhouette and gap statistics methods to find the optimal number of clusters, it can be concluded that it is optimal to have 3 clusters for this data frame.


Optimal number of clusters
Total Within Sum of Square
6e+05
5e+05
4e+05
3e+05
2e+05
1
2
3
4
5
6
7
8
9
10
Number of clusters k


Optimal number of clusters
Average silhouette width
0.0
0.1
0.2
0.3
1
2
3
4
5
6
7
8
9
10
Number of clusters k

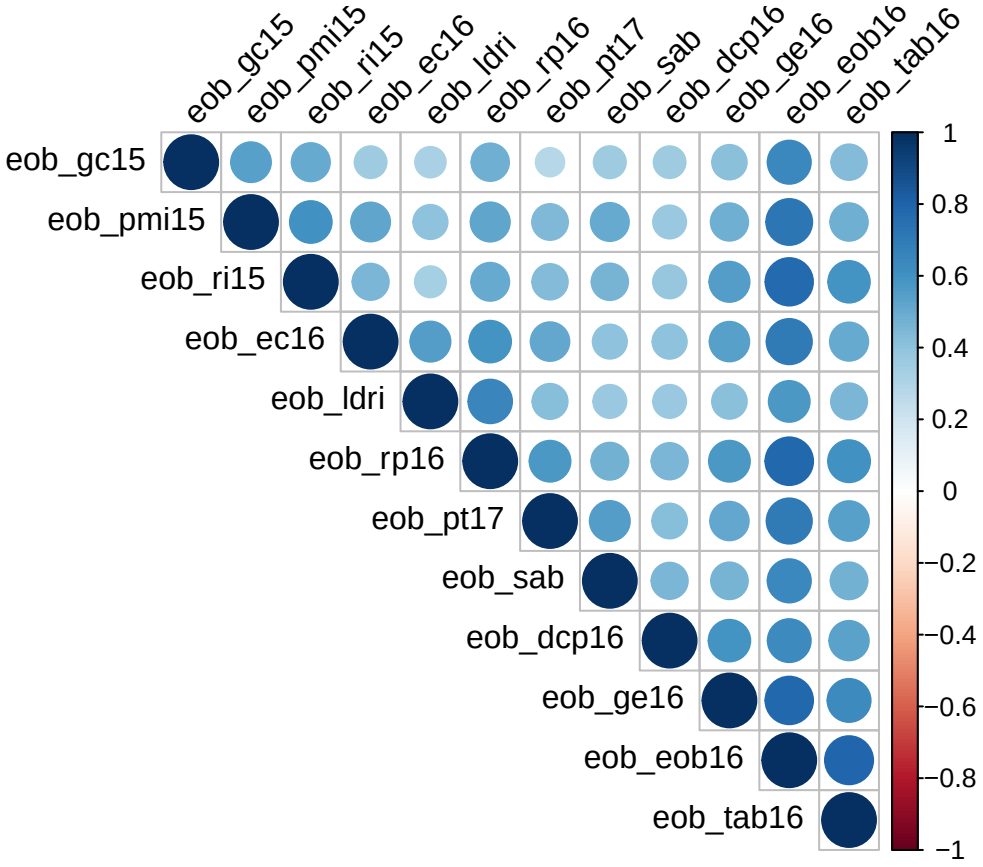

Optimal number of clusters
Gap statistic (k)
0.56
0.52
0.48
0.44
1
2
3
4
5
6
7
8
9
10
Number of clusters k



PCA

Looking at the data, it is observed that there are more variable than can be used to form a good analysis. To reduce the number of variables, Prinicipal Component Analysis is utilized.

For PCA, first the correlation between variables are plotted.



eob_eob16 has a strong positive linear relationship (+.070) with 7 out of 10 other numeric variables, namely:

eob_eob16 and eob_ec16: 0.71 eob_eob16 and eob_ge16: 0.78 eob_eob16 and eov_pmi15: 0.7284136 eob_eob16 and eob_pt17: 0.71 eob_eob16 and eob_ri15: 0.7769863 eob_eob16 and eob_rp16: 0.7848144 eob_eob16 and eob_tab16: 0.7983523

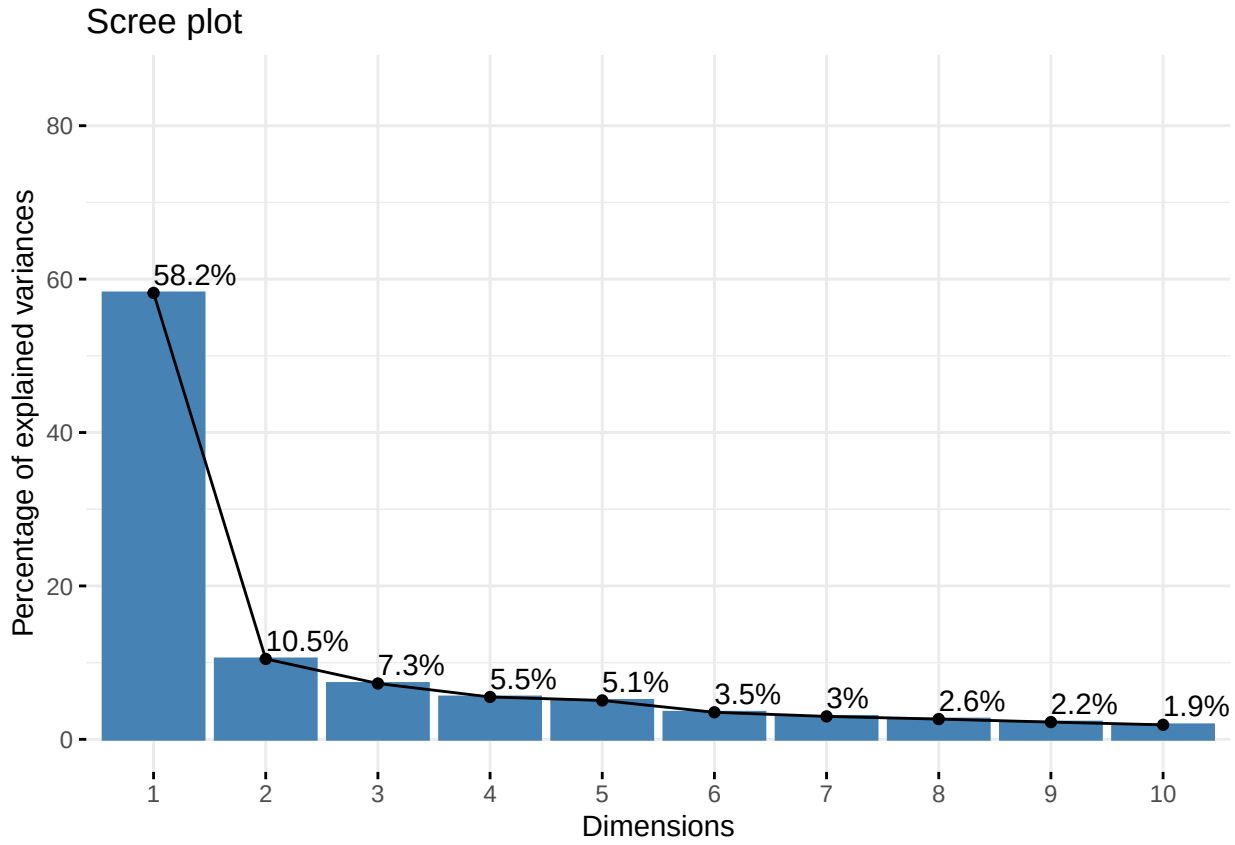
The above graph clearly shows the correlation between some variables are high whereas others are moderate to weak.

The next steps is to reduce the variable by fitting the data using PCA.

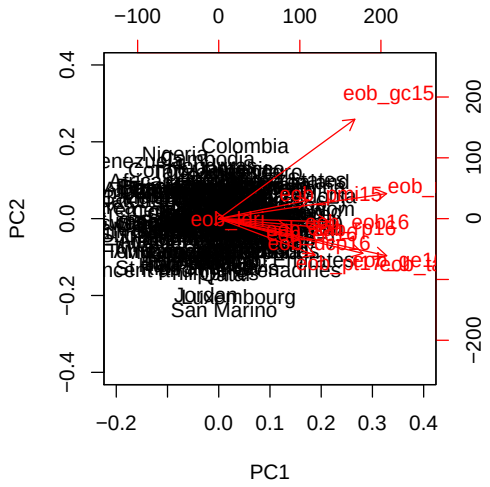
```
## Importance of components:
##                             PC1     PC2      PC3      PC4      PC5      PC6
## Standard deviation     44.908 19.0583 15.88461 13.83389 13.25680 11.05088
## Proportion of Variance  0.582  0.1048  0.07281  0.05523  0.05071  0.03524
## Cumulative Proportion   0.582  0.6868  0.75961  0.81484  0.86555  0.90080
##                            PC7     PC8     PC9    PC10    PC11    PC12
## Standard deviation     10.1788 9.56585 8.82618 8.0933 2.02390 1.07883
## Proportion of Variance  0.0299 0.02641 0.02248 0.0189 0.00118 0.00034
## Cumulative Proportion   0.9307 0.95710 0.97958 0.9985 0.99966 1.00000

##                              PC1       PC2        PC3        PC4
## Afghanistan          -71.710135  18.71333  4.3152229   2.255178
## Albania               23.400133  14.89497 18.8338792   4.109742
## Algeria              -48.663274 -11.99608 26.8998168  -8.907164
## Angola               -81.784466 -16.38438 -7.7797557 -25.668689
## Antigua and Barbuda   -8.585865 -23.07723  0.3957049   5.926952
## Azerbaijan            17.614810 -15.30373 -5.4683719 -23.049409
```

From the above importance of components, it is deduced that the 1st 4 have a cumulative proportion of 81.48% of the variance.

This can be further explained using the graphs below.



The Biplot clearly shows that the variable: eob_ri15, eob_tab16, eob_ge16 and eob_eob16 have the highest contribution to both PC1 and PC2.



Conclusion

From the above analysis, it can be concluded that business regulations and enforcements across various economies affect their respective ease of doing business and they can be grouped into 3 different clusters of countries from the methods based on the scores of the variables considered.

The PCA shows us that the whole analysis can be reduced to just 4 components or variables that would reproduce the information with similar information making the computation faster and easier analysis.

Reference

1) https://qog.pol.gu.se/data/datadownloads/qogstandarddata
2) Teorell, Jan, Stefan Dahlberg, Sören Holmberg, Bo Rothstein, Natalia Alvarado Pachon & Sofia Axelsson. 2020. The Quality of Government Standard Dataset, version Jan20. University of Gothenburg: The Quality of Government Institute, http://www.qog.pol.gu.se doi:10.18157/qogstdjan20
3) The QoG Standard Dataset 2020 Codebook

i) Teorell, Jan, Stefan Dahlberg, Sören Holmberg, Bo Rothstein, Natalia Alvarado Pachon & Sofia Axelsson. 2020. The Quality of Government Standard Dataset, version Jan20. University of Gothenburg: The Quality of Government Institute, http://www.qog.pol.gu.se doi:10.18157/qogstdjan20 http://www.qogdata.pol.gu.se/data/qog_std_jan20.pdf
ii) http://www.doingbusiness.org/en/doingbusiness (The World Bank Group, 2019)

Appendix

1 {#1} eob_dcp16: Score-Dealing with Construction Permits (DB16-19 methodology) measures the gap between an economy's performance and the regulatory best practice on the Dealing with Construction permits indicator components. It is calculated as the simple average of the scores for Procedures (number), Time (days), Cost (a percentage of the warehouse value), and the Building Quality Control Index. The score ranges from 0 to 100, where 0 represents the worst regulatory performance and 100 the best regulatory performance.

2 {#2} eob_ec16: Score-Enforcing contracts (DB16 methodology) measures the gap between an economy's performance and the regulatory best practice on the Enforcing Contracts indicator components. It is calculated as the simple average of the scores for Time (days), Cost (% of claim value) and Quality of judicial processes index. The score ranges from 0 to 100, where 0 represents the worst regulatory performance and 100 the best regulatory performance.

3 {#3} eob_eob16: Ease of doing business score (DB16 methodology) captures the gap between an economy's performance and a measure of best practice across the entire sample of 41 indicators for 10 Doing Business topics. The score ranges from 0 to 100, where 0 represents the worst regulatory performance and 100 the best regulatory performance. Calculating the ease of doing business score for each economy involves two main steps. In the first step individual component indicators are normalized to a common unit where each of the 41 component indicators y (except for the total tax and contribution rate) is rescaled using the linear transformation (worst - y)/(worst - best). In this formulation the highest score represents the best regulatory performance on the indicator across all economies since 2005 or the third year in which data for the indicator were collected. Both the best regulatory performance and the worst regulatory performance are established every five years based on the Doing Business data for the year in which they are established and remain at that level for the five years regardless of any changes in data in interim years. In the second step for calculating the ease of doing business score, the scores obtained for individual indicators for each economy are aggregated through simple averaging into one score, first for each topic and then across all 10 topics. For the ease of doing business score (DB16 methodology), the specific topic scores used are: Score-Starting a business, Score-Dealing with construction permits (DB16-19 methodology), Score-Getting electricity (DB16-19 methodology), Score-Registering property (DB16 methodology), Score-Getting credit

(DB15-19 methodology), Score-Protecting minority investors (DB15-19 methodology), Score-Paying taxes (DB06-16 methodology), Score-Trading across borders (DB16-19 methodology), Score-Enforcing contracts (DB16 methodology), Score-Resolving insolvency (DB15-19 methodology).

4 {#4} eob_gc15: Score-Getting credit (DB15-19 methodology) measures the gap between an economy's performance and the regulatory best practice on the Getting Credit indicator components. The sub-indicators are weighted proportionally, according to their contribution to the total score, with a weight of 60% assigned to the strength of legal rights index and 40% to the depth of credit information index. The score ranges from 0 to 100, where 0 represents the worst regulatory performance and 100 the best regulatory performance.

5 {#5} eob_ge16: Score-Getting electricity (DB16-19 methodology) measures the gap between an economy's performance and the regulatory best practice on the Getting Electricity indicator components. It is calculated as the simple average of the scores for Procedures (number), Time (days), Cost (% of income per capita), and Reliability of supply and transparency of tariff index. The score ranges from 0 to 100, where 0 represents the worst regulatory performance and 100 the best regulatory performance.

6 {#6} eob_ldri: Land dispute resolution index (0-8) (DB16-19 methodology) measures the accessibility of conflict resolution mechanisms and the extent of liability for entities or agents recording land transactions.

7 {#7} eob_pmi15: Score-Protecting minority investors (DB15-19 methodology) measures the gap between an economy's performance and the regulatory best practice on the Protecting Minority Investors indicator components. It is calculated as the simple average of the scores for Extent of conflict of interest regulation index (0-10) (DB15-19 methodology) and Extent of shareholder governance index (0-10) (DB15-19 methodology). The score ranges from 0 to 100, where 0 represents the worst regulatory performance and 100 the best regulatory performance.

8 {#8} eob_pt17: Score-Paying taxes (DB17-19 methodology) measures the gap between an economy's performance and the regulatory best practice on the Paying Taxes indicator components. It is calculated as the simple average of the scores for Payments (number per year), Time (hours), Total Tax and Contribution Rate (% of profits), and Postfiling index (0-100) (DB17-19 methodology). The score ranges from 0 to 100, where 0 represents the worst regulatory performance and 100 the best regulatory performance.

9 {#9} eob_ri15: Score-Resolving insolvency (DB15-19 methodology) measures the gap between an economy's performance and the regulatory best practice on the Resolving Insolvency indicator components. It is calculated as the simple average of the scores for the Recovery Rate (cents on the dollar) and the Strength of Insolvency Framework Index (0-16). The score ranges from 0 to 100, where 0 represents the worst regulatory performance and 100 the best regulatory performance.

10 {#10} eob_rp16: Score-Registering Property (DB16 methodology) measures the gap between an economy's performance and the regulatory best practice on the Registering Property indicator components. It is calculated as the simple average of the scores for Procedures (number), Time (days), Cost (% of property value), and Quality of land administration index (0-30) (DB16 methodology). The score ranges from 0 to 100, where 0 represents the worst regulatory performance and 100 the best regulatory performance.

11 {#11} eob_sab: Score-Starting a business measures the gap between an economy's performance and the regulatory best practice on the Starting a Business indicator components. It is calculated as the simple average of the scores for Procedures (number), Time (calendar days), Cost (% of income per capita), and Paid-in Minimum capital (% of income per capita). The scores for the following components are obtained as such: the score for Procedures (number) is calculated based on the average of scores for Procedures - Men (number) and Procedures - Women (number); the score for Time (calendar days) is calculated based on the average of scores for Time - Men (calendar days) and Time - Women (calendar days); and the score for Cost (% of income per capita) is calculated based on the average of scores for Cost - Men (% of income per capita) and Cost - Women (% of income per capita). The score ranges from 0 to 100, where 0 represents the worst regulatory performance and 100 the best regulatory performance.

12 {#12} eob_tab16: Score-Trading across Borders (DB16-19 methodology) measures the gap between an economy's performance and the regulatory best practice on the Trading across Borders indicator components. It is calculated as the simple average of the scores for Time to export: Border compliance (hours), Cost to export: Border compliance (US dollar), Time to export: Documentary compliance (hours), Cost to export:

Documentary compliance (US dollar), Time to import: Border compliance (hours), Cost to import: Border compliance (US dollar), Time to import: Documentary compliance (hours) and Cost to import: Documentary compliance (US dollar). The score ranges from 0 to 100, where 0 represents the worst regulatory performance and 100 the best regulatory performance.