*Article*

**Title**

**Ana Beatriz Hernández-Lara 1, Antonia Terán-Bustamante 2,\*, Sandra Nelly Leyva-Hernández 1, and Vladimiro González-Zelaya 4**

| **Citation:** To be added by editorial staff during production.  Received: date  Revised: date  Accepted: date  Published: date    **Copyright:** © 2023 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/). |
| --- |

1 Facultad de Ingeniería y Negocios San Quintín, Universidad Autónoma de Baja California, Baja California, 22930; leyva.sandra@uabc.edu.mx

2 Facultad de Ciencias Económicas y Empresariales, Universidad Panamericana, Ciudad de México 03920; ateran@up.edu.mx

3 Facultad de Ingeniería y Negocios San Quintín, Universidad Autónoma de Baja California, Baja California, 22930; [leyva.sandra@uabc.edu.mx](mailto:leyva.sandra@uabc.edu.mx)

4 Facultad de Ciencias Económicas y Empresariales, Universidad Panamericana, Ciudad de México 03920; cvgonzalez@up.edu.mx

**\*** Correspondence: ateran@up.edu.mx

**Abstract:** A single paragraph of about 200 words maximum. For research articles, abstracts should give a pertinent overview of the work. We strongly encourage authors to use the following style of structured abstracts, but without headings: (1) Background: Place the question addressed in a broad context and highlight the purpose of the study; (2) Methods: briefly describe the main methods or treatments applied; (3) Results: summarize the article’s main findings; (4) Conclusions: indicate the main conclusions or interpretations. The abstract should be an objective representation of the article and it must not contain results that are not presented and substantiated in the main text and should not exaggerate the main conclusions.

**Keywords:** keyword 1; keyword 2; keyword 3 (List three to ten pertinent keywords specific to the article yet reasonably common within the subject discipline.)

**1. Introduction**

The introduction should briefly place the study in a broad context and highlight why it is important. It should define the purpose of the work and its significance. The current state of the research field should be carefully reviewed and key publications cited. Please highlight controversial and diverging hypotheses when necessary. Finally, briefly mention the main aim of the work and highlight the principal conclusions. As far as possible, please keep the introduction comprehensible to scientists outside your particular field of research. All the references mentioned in the text should be cited in the “Author-Date” format—e.g., (Baranwal and Munteanu [1921] 1955), (Berry and Smith 1999), (Cojocaru et al. 1999) or Driver et al. (2000). See the end of the document for further details on references.

**2. Materials and Methods**

Two data sets were used for our analyses: *Women, Business and the Law (WBL)* (World Bank 2023) and the *Entrepreneurship Database (WeData)* (World Bank 2021). Two experiments were performed:

1. Fit a Random Forest classifier to the WBL data to determine the most important predictors for a country’s level of income. This experiment is described in detail in subsection 2.2.
2. Perform K Means clustering on the data to detect similarities and differences across countries, going beyond income group and geographical region. This experiment is described in detail in subsection 2.3.

The data pre-processing and machine learning modeling were coded in *Python 3.10.12*, using *pandas 1.5.3*, *scikit-learn 1.2.2* and *seaborn 0.12.2* for data analysis, ML modeling and visualizations, respectively. Our code is available as a *Jupyter Notebook*, available online[[1]](#footnote-0) for reproducibility.

**2.1. Data Description and Pre-Processing**

This small section contains the technical details of the employed data sets, as well as the data pre-processing that was applied to them. Subsection 2.1.1. describes the two analyzed data sets, *WBL* and *WeData*, while Subsection 2.1.2. contains the relatively minor pre-processing that was performed on them before the analyses.

**2.1.1. Data Sets**

The *WBL* data set consists of data for 190 economies through the period from 1971 to 2023. The data features include two demographic variables (*Region* and *Income Group*), as well as 35 binary (Yes / No) questions regarding female equality by the different economies’ laws and 5 numerical variables quantifying both maternal and paternal leave after a child is born. The variables are grouped into 8 categories: *mobility*, *workplace*, *marriage*, *parenthood*, *entrepreneurship*, *assets* and *pension*. Finally, there are aggregated scores for each of these categories, as well as an overall aggregate, the *WBL Index*.

The *WeData* data set includes data for 68 economies, gathered from 2014 to 2020, regarding three key positions in entrepreneurship: number of business owners, number of business directors and number of sole proprietors of companies. These data are aggregated by sex, and therefore allow for an analysis of an economy’s female proportion for each of these categories throughout the years. The economies included in *WeData* are a subset of the economies in *WBL*, and therefore both datasets are not completely compatible, but clusters, regions and income levels may still be evaluated for the *WeData* economies.

**2.1.2. Data Pre-Processing**

For our experiments, we removed the data set’s aggregated columns (*WBL Index*, *Mobility*, *Workplace*, *Pay*, *Marriage*, *Parenthood*, *Entrepreneurship*, *Assets* and *Pension*), the identifier columns (*Economy*, *Economy Code*, *ISO Code*), as well as *Year* and *Region*, in order to keep only legal data. *Income* was chosen to be our target variable, and the legal columns —all being *No*-*Yes* questions— were re-encoded to *0-1*, respectively.

**2.2. Experiments**

Two separate experiments were performed on the data. The first experiment consisted of a classification task with *Income* as target and all of the legal columns as predictors, in order to detect the impact of the different limitations on women’s rights over the income of economies. The second experiment consisted of applying clustering techniques to the raw legal columns, to analyze the similarities and differences across groups. Both experiments’ descriptions follow in the subsequent sections.

**2.2.1. Income Prediction with Random Forest Model**

A classifier was fitted on the legal data (0-1 encoded) as predictors and *Income* set as target. For this experiment, the target was binarized, with *Low* and *Lower Middle* re-encoded to *0* and *Upper Middle* and *High* re-encoded to *1*. A random forest classification model was fit to the data due to random forests yielding the best predictive performance out of several tried-out models, as well as it providing an interpretability layer through the fitted model’s Gini feature importances.

In order to prevent overfitting the model, 10-fold cross-validation was performed, resulting in a mean accuracy of 70.4% and standard deviation of 9.7%. Finally, the variables were grouped by their *WBL* assigned category (*Mobility*, *Workplace*, *Pay*, *Marriage*, *Parenthood*, *Entrepreneurship*, *Assets* and *Pension*) and the sum of the Gini importances for each group’s predictors were averaged-out across the cross-validation folds, to detect the most relevant gender-related legal factors in predicting an economy’s income level.

**2.2.2. Data Clustering**

For the second experiment, the data set was aggregated by *Economy* across the recorded years (1971--2023). Afterwards, the data dimensions were reduced by applying an MCA transformation (Greenacre 2017). Finally, the *K-Means* algorithm was tested using *K* values ranging from 2 to 8, in order to find the optimal number of clusters. The inertia elbow point (Syakur et al. 2018) was graphically determined, and the Silhouette score (Shahapure et al. 2020) was calculated for each *K*. Additionally, these clusters were compared with grouping the economies by *Income* and *Region*. To improve the statistical significance of the experiment, the clusters were randomly generated 100 times for each *K*, and then averaged out.

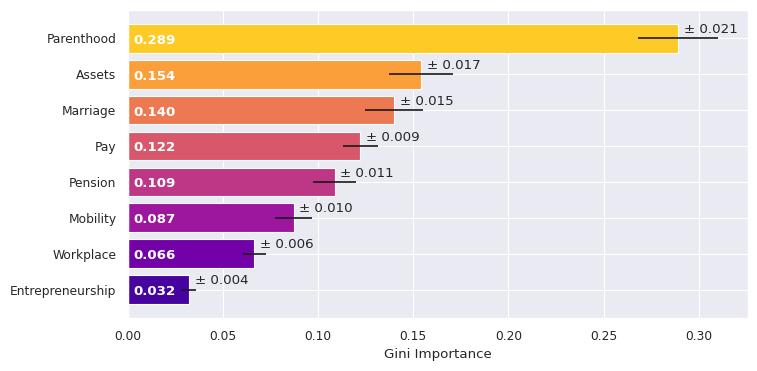
The generated cluster labels were then used to perform an exploratory analysis of both *WBL* and the *WeData*. Specifically, the clusters’ *WBL Index*, as well as the number of business owners, directors and sole proprietors were plotted across the timelapse of the recorded data.

**3. Results**

Having described the setting for both experiments, this section presents our findings in the same order as the previous one, by discussing the predictive model first and the data clusters afterwards.

**3.1. Income Prediction with Random Forest Model**

The Gini importances for the fitted random forest model indicate the most relevant features in predicting an Economy’s *Income* level. Figure 1 shows the sum of the Gini importances for each variable-group, with respect to the fitted random forest model. As may be seen, legislations regarding parenthood are by far the most important predictor for *Income*, accounting for almost 30% of the predictions, followed by women’s conditions to own assets (roughly 15%) and the different countries’ marriage laws (14%). Interestingly, legislation having to do with equality in entrepreneurship do not account for a big part of the prediction rule.



**Figure 1.** Gini importances for the different variable-groups in predicting

an economy’s income level.

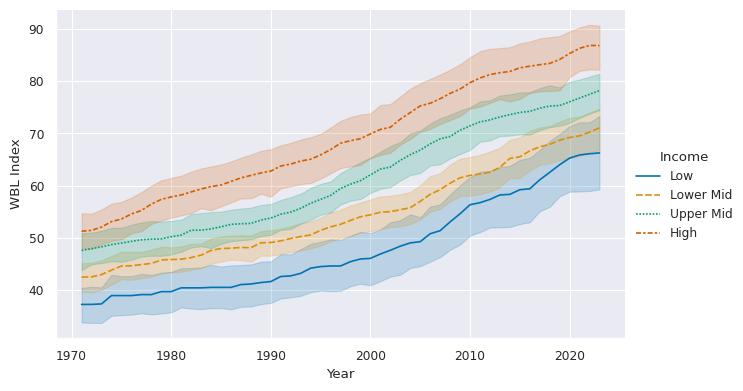
**3.2. Data Clustering**

Our second experiment consisted of analyzing the *WBL Index* behavior across time with economies grouped by:

1. *Income* level,
2. Geographical *Region*,
3. *Clusters* learnt through K-Means clustering.

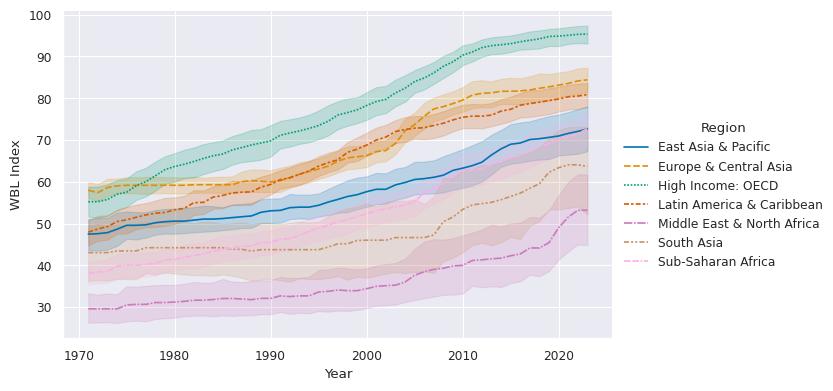
**3.2.1. WBL Index by Income and Region**

As may be seen in Figure 2, when grouped by income level, the different levels follow almost the same trajectory with an increasing vertical offset proportional to the income level. A one-way ANOVA analysis shows that the income groups are indeed well separated from each other, with an f-statistic of 617.385 and a p-value << 0.001.



**Figure 2.** WBL Index by income level across time with 95% confidence intervals.

When grouping by region the differences are not as apparent for some regions. Figure 3 shows that while the *High income: OECD* economies do display the most progressive laws of all regions and *Middle East & North Africa* is clearly the most lagging region in this regard, the trends for *Europe & Central Asia* and *Latin America & Caribbean* are easily confounded, as is the case for the *Sub-Saharan Africa* and *East Asia & Pacific* regions. In this case, the f-statistic for an ANOVA one-way test is even larger, at 1184.69 and a p-value << 0.01. Keep in mind, however, that this test only determines that there are at least two real groups, not that all groups are well separated.

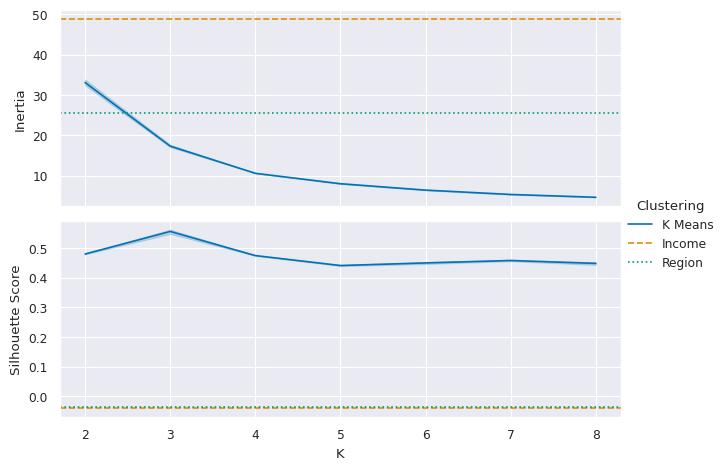


**Figure 3.** WBL Index by region across time with 95% confidence intervals.

**3.2.2. K-Means Clustering**

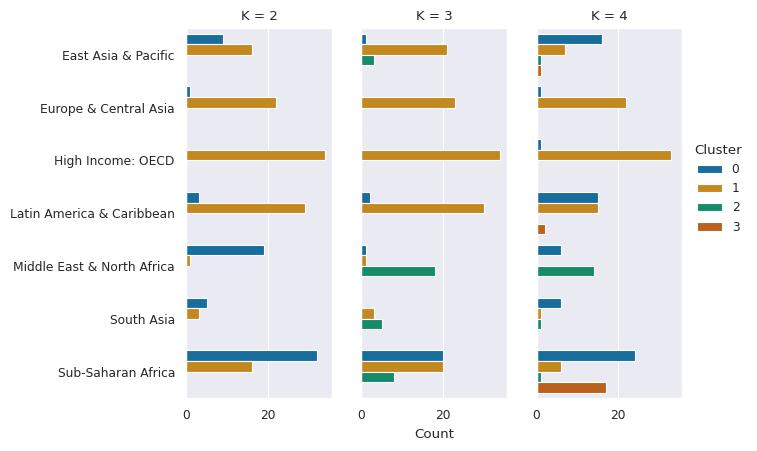
In order to determine the quality of the income and region separations, we compared them against learnt clusters of varying sizes, generated via the K-Means clustering algorithm. As may be seen in both plots of Figure 4, when measuring the group’s inertia (the sum of the distances from an economy to its cluster centroid), even two clusters provide better separation than income grouping, while three or more clusters already prove better than region separation.

When looking at the silhouette score —a measure of how similar an object is to its own cluster compared to other clusters— the results are even stronger, as both income and region grouping produce a negative score, indicating poor separation across groups. In contrast, K-Means positive silhouette scores for every *K*. Analysis of the silhouette score plot indicates that the optimal number of clusters is *K = 3*, but the *K = 2* and *K = 4* cases were also analyzed in our experiment as the elbow point for inertia is not obviously located at *K = 3*.



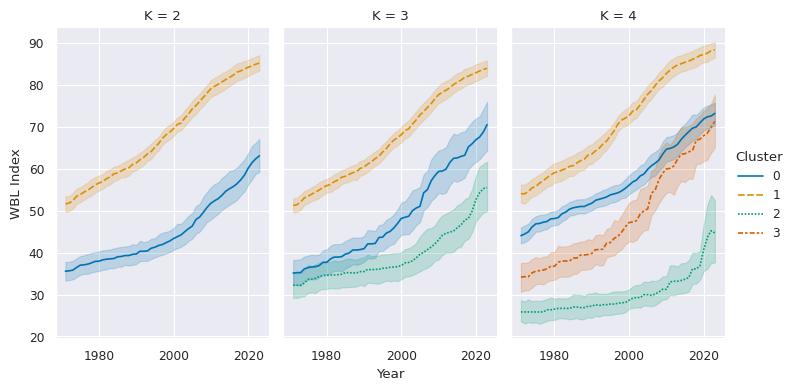
**Figure 4.** Inertia and Silhouette scores for *income* and *region* grouping vs. K-Means clustering for different *K* values.

Given the K-Means clusterings’ improved cohesion and separation over *Income* and *Region* grouping, we then proceeded to analyze the resulting clusters’ behavior. When looking at the regional composition of the clusters, shown in Figure 5, it becomes immediately apparent that the *High Income: OECD* economies and the *Middle East & North Africa* economies always end up in different clusters, regardless of *K*. *Europe & Central Asia* economies are clustered together with High Income in every case, same case as *East Asia & Pacific* and *Latin America & Caribbean* for *K = 2* and *K = 3*. These two regions get separated further as *K* increases to 4, though, part of them still clustered together with *High Income* and part of them belonging to a new, heterogeneous group. Interestingly, the fourth cluster in the *K = 4* case consists almost entirely of *Sub-Saharan Africa* countries.



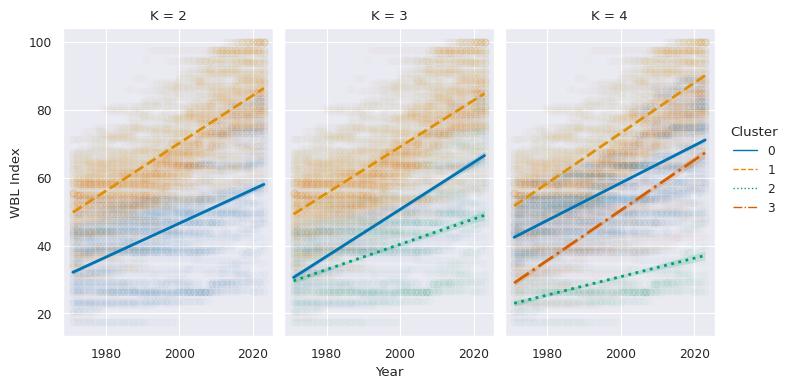
**Figure 5.** Regional composition of clusters for *K = 2*, *K = 3*, and *K = 4*.

When looking at the temporal trends of the different clusters, shown in Figure 6, we can see that the curves are clearly separated for all three *K*-values, with even the confidence intervals being disjoint most of the time. The resulting one-way ANOVA f-statistics for *K = 2*, *K = 3* and *K = 4* were 5847.44, 2937.81 and 2602.6, respectively. In every case, the p-value << 0.001.



**Figure 6.** WBL Index by cluster across time with 95% confidence intervals.

In order to further understand the temporal cluster trends, a simple linear regression of *WBL Index ~ Year* was performed for each cluster, resulting in the lines shown in Figure 7. These lines allowed us to predict the year by which each cluster will achieve the optimal *WBL Index* score of 100. The resulting predictions are presented in Table 1. As may be seen, there are large differences in the time windows by when this optimal condition may be expected. For the sake of completeness, a similar analysis was performed when grouping the economies by *Income* and *Region*. The resulting predictions are again vastly different across groups, and may be seen in Tables 2 and 3 for *Income* and *Region*, respectively.



**Figure 7.** Linear regression fits *WBL Index ~ Year* for each cluster.

**Table 1.** Year prediction for each cluster to achieve *WBL Index = 100*.

|  | Year | | |
| --- | --- | --- | --- |
| Cluster | K = 2 | K = 3 | K = 4 |
| 0 | 2107 | 2071 | 2255 |
| 1 | 2042 | 2045 | 2036 |
| 2 | — | 2170 | 2075 |
| 3 | — | — | 2067 |

**Table 2.** Year prediction for each income level to achieve *WBL Index = 100*.

| Income | Year |
| --- | --- |
| High | 2040 |
| Upper Mid | 2056 |
| Lower Mid | 2079 |
| Low | 2089 |

**Table 3.** Year prediction for each region to achieve *WBL Index* = 100.

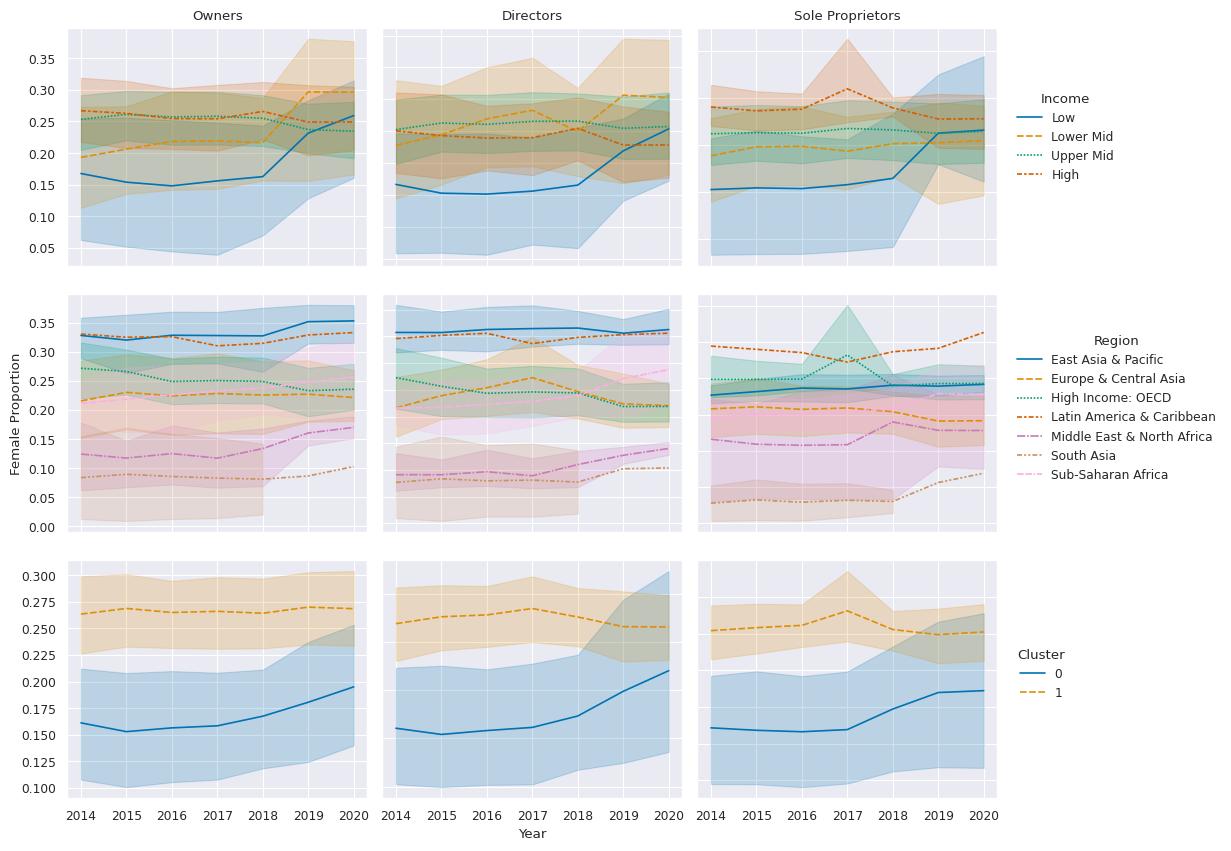
| Region | Year |
| --- | --- |
| High Income: OECD | 2024 |
| Latin America & Caribbean | 2047 |
| Europe & Central Asia | 2049 |
| Sub-Saharan Africa | 2067 |
| East Asia & Pacific | 2083 |
| South Asia | 2134 |
| Middle East & North Africa | 2167 |

**3.2.3. *WeData* Data Set Analysis**

Finally, we merged the *WeData* data set with *WBL* to analyze the trends for each group regarding the three variables registered in *WeData*: proportion of female ownership, proportion of female business directors and proportion of female sole-proprietorship. It is important to note that the economies in *WeData* consist of a subset of the economies in *WBL*. Therefore, while it is possible to assign each economy in *WeData* an income level, a region and cluster labels, the resulting trends will be incomplete, as many countries are not present in *WeData*.

Still, it is possible to differentiate the groups with some clarity, as may be seen in Figure 8, especially in the 2-group cluster separation presented in the bottom row of the plot grid. When split by income level, the only group that shows an increase in female participation in businesses are the *low* income economies, specifically after 2018. Interestingly, the trends for most regions and clusters are quite flat, signifying that these proportions are very culturally embedded and are specific to each region.

When analyzed by region, the middle row in Figure 8, *Latin America & Caribbean* and *Europe & Central Asia* lead the female participation in these strategic positions, even outperforming the *High Income: OECD* economies, while *South Asia* lags behind every other region.



**Figure 8.** Female proportion of ownership, directing and being sole-proprietors

by *Income*, *Region*, and *Cluster (K = 2)*, with 95% confidence intervals.

**4. Discussion**

Author

**5. Conclusions**

This

**Author Contributions:** Conceptualization, Ana Beatriz Hernandez-Lara, Antonia Terán-Bustamante and Sandra Nelly Leyva-Hernández and; Data curation, Sandra Nelly Leyva-Hernández, and Vladimiro González-Zelaya; Formal analysis, Sandra Nelly Leyva-Hernández, Antonia Terán-Bustamante, Paola Miriam Arango-Ramírez and Vladimiro González-Zelaya; Investigation, Sandra Nelly Leyva-Hernández and Antonia Terán-Bustamante; Methodology, Antonia Terán-Bustamante, Paola Miriam Arango-Ramírez and Vladimiro González-Zelaya; Project administration, Sandra Nelly Leyva-Hernández and Antonia Terán-Bustamante; Resources, Sandra Nelly Leyva-Hernández and Antonia Terán-Bustamante; Software, Sandra Nelly Leyva-Hernández, Paola Miriam Arango-Ramírez and Vladimiro González-Zelaya; Supervision, Antonia Terán-Bustamante; Validation, Sandra Nelly Leyva-Hernández, Paola Miriam Arango-Ramírez and Vladimiro González-Zelaya; Visualization, Sandra Nelly Leyva-Hernández and Vladimiro González-Zelaya; Writing – original draft, Sandra Nelly Leyva-Hernández, Antonia Terán-Bustamante, Paola Miriam Arango-Ramírez and Vladimiro González-Zelaya; Writing – review & editing, Sandra Nelly Leyva-Hernández and Antonia Terán-Bustamante.

**Funding:** This research received no external funding

**Institutional Review Board Statement:** Not applicable

**References**

(Greenacre 2017) Greenacre, M. (2017). Correspondence Analysis in Practice. CRC Press.

(Shahapure et al. 2020) Shahapure, K. R., & Nicholas, C. (2020, October). Cluster Quality Analysis Using Silhouette Score. In 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 747-748). IEEE.

(Syakur et al. 2018) Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster. In IOP Conference Series: Materials Science and Engineering (Vol. 336, p. 012017). IOP Publishing.

(World Bank 2021) World Bank. (2021). Entrepreneurship Database 2021. Washington, DC: World Bank.

(World Bank 2023) World Bank. (2023). Women, Business and the Law 2023. Washington, DC: World Bank.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

1. Our code is available at <https://github.com/vladoxNCL/WBL_paper/blob/main/WBL_Analysis.ipynb> [↑](#footnote-ref-0)