Advances in Computer Science Research, volume 96

Proceedings of the International Conference on Mathematics, Geometry, Statistics, and Computation (IC-MaGeStiC 2021)

SHINY OFFICE-R: A Web-based Data Mining Tool for Exploring and Visualizing Company Profiles

I Made Tirta*, Mohamad Fatekurahman, Khairul Anam, Bayu Taruna Widjaja Putra

The University of Jember
**Corresponding author. Email: itirta.fmipa@unej.ac.id*

ABSTRACT
The profile of institutions or companies are often measured internally, nationally and internationally using several indicators that may be changed over time. We develop SHINY OFFICE-R a Web-GUI (Graphical User Interface) using R software to explore and visualize data on institution performance/ profile. Graphical visualization (such as bar-plot, line-plot, histogram, biplot and path diagram) can help a lot in gaining the insight of the data. The programs are flexible to accommodate different types of indicators that may be assigned for broad types of institutions and companies. In this paper we describe the main features of the program and illustrate application of the GUI using simulated data having various type of performance indicators (say local, national and international indicators). Various graphics visualizations are illustrated according to the objective of data explorations. Furthermore, our web GUI will be available online, so that it can be easily accessed and applied to explore and visualized the profile of users' institutions or companies that possibly have different types of indicators.

***Keyword**: Office statistics, Performance indicators, Graphical visualization, Cluster, Path analysis, Company profile, Structural Equation Model (SEM), Web.*

1. INTRODUCTION

1.1. *The Importance of Data for Supporting Decision Making*

Since the advanced of ICT (information Communication Technology) in the last two decades. Data (its sizes and complexity) has been growing rapidly and known as big data. The existences of big data were already taken into account to support policy making in organizations or business companies. Data-based or also known as evident-based decision making was proven to be more effective than merely based on experience and expert judgment of individual [1].

Recently many practitioners believe that the role of data (evident) in supporting decision or policy making can be categorized into three types namely, data-driven, data-informed and data-inspired decision making [2],[3]. If the rules of the decision makings are clear and they do not need human intervention (such as examinations or selections), then data-driven decision making is suitable and it can be made fully computerized (automated) processed. Ones may utilize data as main [2] or inspiring information or knowledge in making plans or policies, then combined with human experience, intuition, expert judgment and intelligence to draw final decision. In all situations, at least knowledge or information from data must be taken into account in making decisions, policies or plans for companies or institutions, so that the policies can be more realistic, beneficial, suitable and match the realistic needs of their companies or institutions.

In educational institutions, a more serious utilization of data mining (DME, data mining in education) to support decision making, apply data analytics. Applications of data analytic are categorized into two types namely, learning analytics and academic analytic. Learning analytic focuses more on learning, courses activities, while academic analytics focuses more on managerial aspects of educational institutions including support for decision making (such as optimizing resources, enhance reputations) [4].

1.2. *Simple Statistics to Describe Data*

Data can be just collections of many numbers and categories. In data mining principles, having data is not enough, we have to be able to get information, knowledge, insight, or even the wisdom of the data [5]. Data should be explored and described numerically or graphically. For numerical information we can apply simple or basic statistics summary including information about frequency, minimum, maximum, mean, median, variance, and correlation matrix of data. Comparison of mean and median values will give an idea if the data of the related variable is relatively symmetrical or noticeably skewed. A more advanced statistical method may be needed for special cases and purposes [5]. Clearer knowledge of the data may be grasped from the graphical visualizations of the data.

1.3. *Graphics Visualization to Get Insight of Data*

In addition to numerical description of data, we can proceed further to get the important and suitable knowledge, that become the insight and the wisdom of the data. Figure 1 illustrates clearly the different levels of understanding of data [6]. Graphical representations (visualizations) of the data can describe the insight of data more quickly and clearly [7]. Graphical representations or visualizations are also commonly utilized to support finding with statistical test or statistical models.

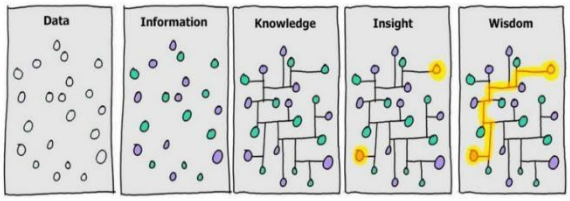


Figure 1 From row data to the insight and wisdom. (Source: https://www.theifactory.com/news/gaining-wisdom-from-data/)

The most common basic graphic visualization applied for exploring data include [7]: (i) bar diagram with various grouping or data selection, (ii) lines graph, (iii) histogram, (iv) scatter plot with density and grouping, (v) correlation plot. More advanced plot or graphics representations may include (vi) dendrogram (for clustering), (vii) PCA-biplot (for dimensions reduction), (viii) path and structural (SEM) plot (for visualizing the structural relationship). All types of the graphical visualizations may then be combined with group exploration representing different levels of working unit or division in the institutions or companies.

1.4. *Easily Accessible and User-Friendly Tool with Shiny*

To get the optimum benefits of data mining especially for educational institutions, the availability of special tools will help much. The tools should be user-friendly (in terms of operation and interpretation) and can be easily accessed. This kind of tools are commonly developed with Web-based interface so that user can access the program from anywhere, anytime with any device or gadget. One of widely chosen application for this kind of purposes is Shiny of R. Many web-based application have been developed using Shiny for various purposes or area including general decision support system [8], support system for environmental scientists [9], interactive program for penological study [10], toxicology [11]. We name our program as Shiny Office-R to indicate that the program is mainly designed to explore, visualize, institutional (Office) data to support decision making so that the company become brighter (shinier).

This paper describes and illustrate the features (capabilities, options and output) of the program (Web Interface), Shiny Office-R, that we have developed using Shiny and R Packages. We will also discuss future improvement, extension and accessibility of the program.

2. METHODS

The works consist of combining the suitable statistics [12,13], graphical visualization [14,15] which then they are structured in a user-friendly tool using R and shiny [16,17]. The main methods and steps in developing Web-Interfaces using Shiny and R Packages are as follow.

1. Determining the structure and scope of representation and visualization needed for decision making
2. Exploring related main statistics and graphical visualization
3. Exploring the most related and relevant R packages (related to import data, graphics, statistics and interface)
4. Defining and writing needed R functions
5. Developing ui.R and server.R
6. Enhancing input and output variability
7. Testing and illustrating the program
8. Improving performance (enhancing appearance, readability, user-friendliness, etc.)

3. RESULTS AND DISCUSSIONS

3.1. *Main Features*

At the time of preparing this manuscript, the main features related to data mining for supporting decision making include the following.

1. **Import/ Input data**. Several data format that can be directly uploaded are .xlc(x) and .csv
2. **Pra-Process**. Data pra-process (pre-process) mainly deal with (i) estimation methods of missing numerical observation, (ii) simple transformations (including standardization) and (iii) separating variables into factors, numerical variable and label for working unit level.

3. **Human Resources (HR) Exploration**. Exploration of human resources include mapping the distribution of the number of HR based on working unit and qualification on specified indicators/ variables. The outputs are presented in the form of bar plot, line plot, table and information of chi-square test of the frequency distributions.
4. **Achievement or Profile on selected Indicators.** At this stage user can explore the achievement on several selected indicators in the form of (i) overall correlation or correlation based of selected group levels (ii) density (histogram) plot of the achievement, (iii) scatter plot or correlation diagram for selected indicators (overall or group based).
5. **Specific or Advanced visualizations.** The advance visualizations include (i) the visualization using cluster analysis (clustering working unit in different level, for relatively big institutions), (ii) visualization of PCA-Biplot in combination with available factors or available clusters), (ii) Structural Equation and Path diagram to further explore the causal effect relationship of the indicators. To explore using advanced visualization, user must have appropriate knowledge about the corresponding statistics theory on them. However, Practical knowledge for the visualization actually can be trained quickly

3.2 *Illustrations*

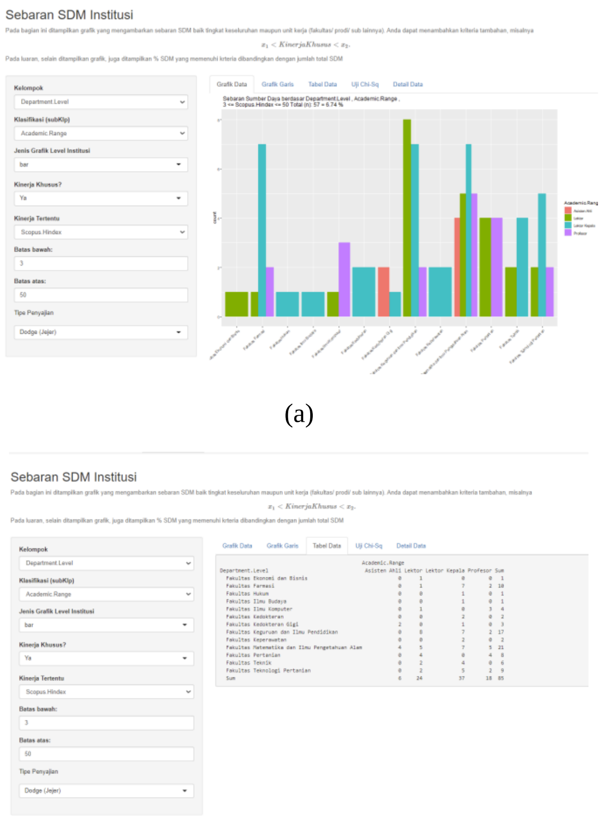
The main purpose of the article is to illustrate the feature of the programs. Therefore, in the illustration the graphical visualizations are captured together with the appearance of the features (such as pull-down menus of the programs)

3.2.1. HR *Distributions across Working Unit (All or Match Selected Criteria).*

User can visualize the distribution of their HR across working unit, with specific criteria. In higher education institutions, for example user may visualized distribution of lecture across faculties or departments with Academic Range or other existing grouping and having certain level of Scopus H-Index or other selected indicators. The user can then see the bar plot, line plot and the table of the selected data. In addition, there also information on the percentage of HR matching the criterion. The users can check the overall trend from the bar plot or line plot and check the exact number from the table (see Figure 2).

3.2.2. *Performance or Achievement of Indicators.*

Users can also explore the performance of selected indicators to see the distributions including the mean the mode and relationship among indicators across working unit or other existing grouping. In Higher Education institution, for example, user can check the correlation of Scopus and Sinta Score across Academic Range whether they are normally distributed, having different or similar behavior across the level of group (Figure 3).



Sebaran SDM Institusi

Pada bagian ini ditampilkan grafik yang mengambarkan sebaran SDM baik tingkat keseluruhan maupun unit kerja (fakultas/ prodi/ sub lainnya). Anda dapat menambahkan kriteria tambahan, misalnya

$$x_1 < KinerjaKhusus < x_2.$$

Pada luaran, selain ditampilkan grafik, juga ditampilkan % SDM yang memenuhi krteria dibandingkan dengan jumlah total SDM

Kelompok: Department.Level

Klasifikasi (subKlp): Academic.Range

Jenis Grafik Level Institusi: bar

Kinerja Khusus?: Ya

Kinerja Tertentu: Scopus.Hindex

Batas bawah: 3

Batas atas: 50

Tipe Penyajian: Dodge (Jejer)

Grafik Data
Grafik Garis
Tabel Data
Uji Chi-Sq
Detail Data
Sebaran Sumber Daya berdasar Department.Level , Academic.Range ,
3 <= Scopus.Hindex <= 50 Total (n): 57 = 6.74 %
count
Academic.Rang
Asisten Ahli
Lektor
Lektor Kepala
Profesor

(a)

Sebaran SDM Institusi

Pada bagian ini ditampilkan grafik yang mengambarkan sebaran SDM baik tingkat keseluruhan maupun unit kerja (fakultas/ prodi/ sub lainnya). Anda dapat menambahkan kriteria tambahan, misalnya

$$x_1 < KinerjaKhusus < x_2.$$

Pada luaran, selain ditampilkan grafik, juga ditampilkan % SDM yang memenuhi krteria dibandingkan dengan jumlah total SDM

Kelompok: Department.Level

Klasifikasi (subKlp): Academic.Range

Jenis Grafik Level Institusi: bar

Kinerja Khusus?: Ya

Kinerja Tertentu: Scopus.Hindex

Batas bawah: 3

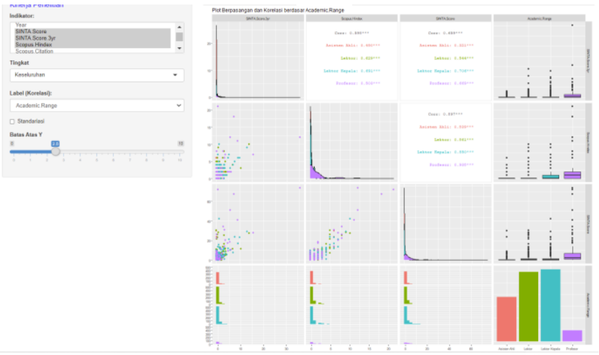
Batas atas: 50

Tipe Penyajian: Dodge (Jejer)

```
                                              Academic.Range
Department.Level                               Asisten Ahli Lektor Lektor Kepala Profesor Sum
  Fakultas Ekonomi dan Bisnis                             0      1             0        0   1
  Fakultas Farmasi                                        0      1             7        2  10
  Fakultas Hukum                                          0      0             1        0   1
  Fakultas Ilmu Budaya                                    0      0             1        0   1
  Fakultas Ilmu Komputer                                  0      1             0        3   4
  Fakultas Kedokteran                                     0      0             2        0   2
  Fakultas Kedokteran Gigi                                2      0             1        0   3
  Fakultas Keguruan dan Ilmu Pendidikan                   0      8             7        2  17
  Fakultas Keperawatan                                    0      0             2        0   2
  Fakultas Matematika dan Ilmu Pengetahuan Alam           4      5             7        5  21
  Fakultas Pertanian                                      0      4             0        4   8
  Fakultas Teknik                                         0      2             4        0   6
  Fakultas Teknologi Pertanian                            0      2             5        2   9
  Sum                                                     6     24            37       18  85
```

(b)

Figure 2 Bar plot (a) and Table (b) of HR Across Working Unit (Faculties and Academic.Range) with certain criterion (Scopus.H-Idex).



Indikator: Year, SINTA.Score, SINTA.Score.3yr, Scopus.Hindex, Scopus.Citation

Tingkat: Keseluruhan

Label (Korelasi): Academic.Range

☐ Standarisasi

Batas Atas Y: 2.5

Figure 3 Scatter Plot and Correlation information of selected indicators (publication index) across selected grouping (academic level).

3.2.3. *Clustering Working Units and Characteristics of Cluster*

The next visualizations are clustering visualization using dendrogram and combined with the mapping of indicators performance using PCA-biplot. These visualizations are useful if institutions have relatively many working units or many indicators to achieve. User can quickly see the cluster of working units with similar conditions. Figure 4 illustrate that the working units can be can be divided into 3 groups, by which each group has similar properties based on the selected indicators. However, dendrogram can not specify the characteristics of each group. The characteristics of group and working unit can be visualized using PCA-biplot (Figure 5). PCA-biplot can map the strength and weakness related to the indicators using vector representation. Therefore, dendrogram combined with PCA-biplot can quickly represent (i) interdependency among indicators, (ii) the position and characteristics of working units regarding to the selected indicators (for example one working unit is very strong on several indicators but quite weak in other indicators). Policy maker then can utilize the information to decide best treatment for each group.

Further visualization can be done to explore possible causal-effect among indicators using path or Structural Equation Model (SEM). This information will be very beneficial in deciding what indicator should be selected so it can effectively and efficiently improve the majority of other indicators. However, to interpret PCA-biplot, path and SEM diagrams users need to understand medium or even higher level of statistics methods (cluster analysis, PCA, regressions, Path analysis or SEM). Therefore, more specific and deeper study may be needed as research topics using more advanced statistics methods.

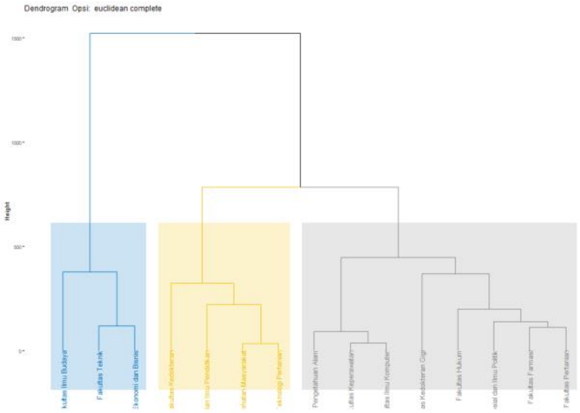


Figure 4 Dendrogram of working units grouped into three clusters.

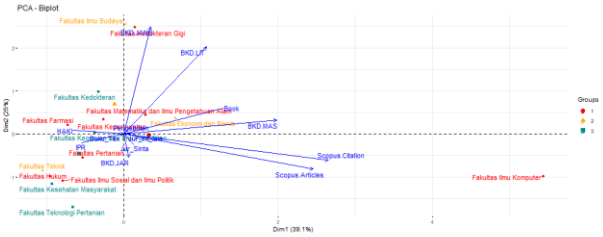


Figure 5 Biplot two-dimensional representation of indicators combined with position of working unit with resulting clusters.

3.2.4. *Advanced Model for The Achievements on Indicators*

The last visualizations that can be performed by the programs are related to modeling the achievements on selected indicators. Two main visualizations are parts of path analysis and structural equation model. These visualizations show how each indicator related to or affect other indicators directly or indirectly. Figure 6 shows that some indicators have significantly positif or negatif effect to other indicators and some do not have significant effect. We may also treat some indicators as measurement of latent factors (in our ilustration are classified as local, national, international indexes). In this case, relationship of indicators achievement can be visualized using SEM plot (Figure 7). SEM plot and SEM analysis are very beneficial to picture institution as a system measured by several type of indicators (including in gereral, at least input, process, facilities and output/outcome). In addition to path or SEM diagram, user may also see the output of the model. However, deeper studies may be needed to model indicators using Path Analysis, SEM and other complex statistical modeling.

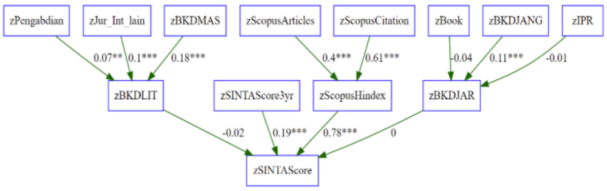


Figure 6 Path diagram visualizing the cause effect relationship of indicators.

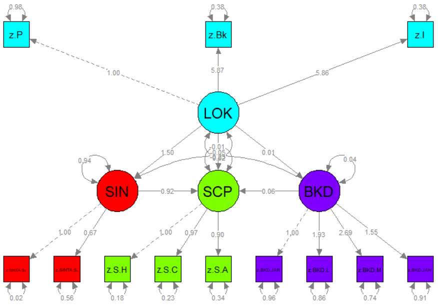


Figure 7 SEM plot visualizing the cause effect relationship of latent variables with indicators.

By doing various visualizations, user may already have enough insight of the data that can be utilized as supports in making plan or decision. In some situation further or deeper investigation or study may be needed to get more detail and accurate results.

For the programs (Web-GUI interface), further improvement may be needed for better appearance or lay out. Related to the future anticipation of visualizing more complex of company data with very large number of various type of indicators (Big Data), Shiny Office-R still need to be improved related to its speed and capacity. Furthermore, it can be officially deployed to the shiny-server so it can be utilized by any individual or working unit who needs it.

4. CONCLUSION

Our proposed Interface Program (Shiny Office-R) already has various features needed for importing data, exploring and visualizing company profile on selected indicators.

The available or resulting visualizations are very useful in order to get insight or wisdom of the data for supporting decision or policy making. Shiny Office-R also provides some complex visualizations and its numerical output related to complex modeling indicators. However special study may be needed to get more specific model and deeper conclusion.

Our Shiny Office-R is flexible for visualizing different data with different indicators especially for higher education institutions. In fact, our Shiny Office-R is readily applicable for companies other than educational institutions. In near future Shiny Office-R may be improved or extended for its speed, appearance and capacity to handle big data.

AUTHORS' CONTRIBUTIONS

The first author mainly contributed on writing the shiny interface program, the second author contributed of preparing type of statistics and R-packages needed. The third and the fourth authors contributed mainly on testing the program using university's data and deploying it to University of Jember Web. All authors involved in preparing the manuscript.

ACKNOWLEDGMENTS

This paper is part of the research supported by internal University of Jember Grant 2021.

REFERENCES

[1] J. Young. 2021. Evidence-based practice for effective decision making. *CIPD*. 15 April 2021. https://www.cipd.co.uk/knowledge/strategy/analytics/evidence-based-practice-factsheet

[2] K.L. Webber & H. Zheng. 2019. *Data Analytics and the* Imperatives *for Data-Informed Decision-Making in Higher Education*. (Institute of Higher Education Research Projects Series, 2019-004)

[3] N.A.B Iskandar. 2021. Data-driven, data-informed and data-inspired. *Sapera Studios.* https://sapera.com/en/blog/data-inspired-data-informed-or-data-driven

[4] A. Nguyen, L. Gardner, & D. Sheridan. 2020. Data Analytics in Higher Education: An Integrated View. *Journal of Information Systems Education*, 31(1), 61-71

[5] S. Toeffery 2011. *Data Mining and Statistics for Decision Making*. John Willey & Son

[6] Factory. *Data to Wisdom via Information, Knowledge and Insight*. https://www.theifactory.com/news/ gaining-wisdom-from-data/

[7] G. Aisch. Using Data Visualization to Find Insights in Data. *DataJournalism**.com*** https:// datajournalism.com/read/handbook/one/understanding-data/using-data-visualization-to-find-insights-in-data [accessed oct 2021]

[8] PHM. Albuqerque, and G. Menteiro. 2020. RMCriteria: a decision making support system package for R.*Communication in Statistics-Simulation and Computation* **50(1):**1

[9] Y. Li. 2020. Towards fast prototyping of cloud-based environmental decision support systems for environmental scientists using R Shiny and Docker. *Environmental Modelling & Software*. Volume 132, October 2020, 104797

[10] M. Möllera, L. Boutarfaab, J. Strassemeyera. 2020. PhenoWin – An R Shiny application for visualization and extraction of phenological windows in Germany. Computers and Electronics in Agriculture Volume 175, August 2020, 105534

[11] T. A. Holland-Letz, Kopp-Schneider. 2021. An R-shiny application to calculate optimal designs for single substance and interaction trials in dose response experiments. *Toxicology Letters.* Volume **337**, 1 February 2021: 18-27

[12] M. Sullivan III. *STATISTICS: Informed Decisions Using Data*. 2018. Pearson.

[13] W. Beaty. 2017. Decision Support Using Non-Parametric Statistics.

[14] A. Kassambara. 2013. *Guide to Create Beautiful Graphics in R.* STHDA.

[15] H. Wickham. 2016. Ggplot2: Elegant Graphics for Data Analysis. Springer (UseR! series)

[16] K-W Moon. 2016. Learn ggplot2 Using Shiny App. Springer (UseR! Series)

[17] C. Beeley. 2013. *Web Application Development with R Using Shiny*. PACKT Open Source