Management Based on Data Analysis. Part One. Data Visualization Analysis

Ilie Margareta
Ilie Constantin
"Ovidius" University of Constanta, Romania
ilie.marga@gmail.com

Abstract

Data visualization is associated with Business Intelligence and acts as the main base of data understanding for the decision making. The techniques and application used to apply the data visualization can be expensive or difficult to understand in accordance with the business targets. In this context, the present paper has the objective to demonstrate the possibility of applying an open source software (Python 3) for several types of data visualization. The method uses different data representation techniques, such as histograms, scatters, correlations, 3D surfaces, etc. Analyzed data refers to human resources evaluation, as the management can analyze and understand how data such as: satisfaction level of the employee, the average monthly hours worked, project number that the employee participated in, and time spent by company for employee care. The results delivered several types of representations that offer conclusions about the positioning of evaluated human resources in conjunction with other data.

Key words: management, data visualization, human resources, evaluation
J.E.L. classification: C15

1. Introduction

Data visualization plays a key role in data exploitation and decision making. Since about 2010, data visualization is enclosed in BI (Business Intelligence) as a tool. Given the influence of visualization, it is only expected to apply the powerful communication methods in the field of BI and analytics. Visualization has become more and more important to BI, where managers need technological support to understand and examine complex data sets and all kinds of information.

Representation of large amount of data (in spreadsheets and large tables) is complicated, difficult to interpret, and brings obstacles in extracting useful data from a database. Using techniques and methods for data visualization will help to solve problems, find how to use visuals in demonstrations and explanations, and make decisions.

Data visualization shows data and information graphically that feature pattern and tendencies and allow audiences to comprehend. Color, intensity, dimension, appearance, and flow of graphical objects are used to represent data, which facilitate analysis that just characters (text) or floats or simple graphs do not. Data visualization grows in importance as data become so large, that it is challenging to understand and apply the importance and consistence.

Different methods of data analysis and representation involve expensive algorithms and applications. Although relatively easy to apply, they must first be understood and customized according to the requirements of the organization. Free variants can be used, but their construction and application takes longer, if those who build the algorithm or application do not have the necessary knowledge.

Present paper shows the second solution. The authors must reveal that the preparation for the applying technique needed about one month for the development and application of algorithm in Google Colab Notebooks Python. The importance of the paper arises from the presentation and demonstration of the possibilities of applying free solutions of representation and visualization of data.

2. Theoretical background

An essential task for senior management today is to acquire decision support from accountants, developers, etc. by being assisted in the analysis of large, complex data sets. Data visualization simplifies this process by granting users permission to navigate, choose, and show data through an easy-to-use interface often used as a module of data analytics (Janvrin, 2014, p. 31).

As Wu et al. (2017) say, one of the most important features of data visualization is the "ubiquity" that makes interactive visualizations an important asset of the visualization process.

A general purpose of data visualization can be considered the delivering of improved technique of appearance and collaboration (Zheng, 2021, p. 11). Also, Zheng adds that more specific purposes (as different categories of data visualization will assist diverse purposes):

- **exhibitory**: displaying or monitoring activities, operations, and events.
- **exploratory**: information searching, browsing, examination, finding.
- **explanatory**: data analysis and insight generation, decision support.
- **communication** and **presentation**, for impression/persuasion.
- **understanding** and **cognition** (comprehension of abstract ideas and processes).
- **artistic** (attractiveness) expression and appreciation.
- **entertaining** and for fun.

With data visualization managers identify tendencies and patterns, structures and associations. Also, it helps concentrate on targets and interests.

3. Research methodology

The research methodology applies different functions and algorithms for the representation of 14,999 datasets consisting in data about human resource evaluations. Data was downloaded from www.kaggle.com. Data consist of 10 data sets, also presented in table 1, as the first 5 data rows:

Table no. 1 Database. First 6 rows

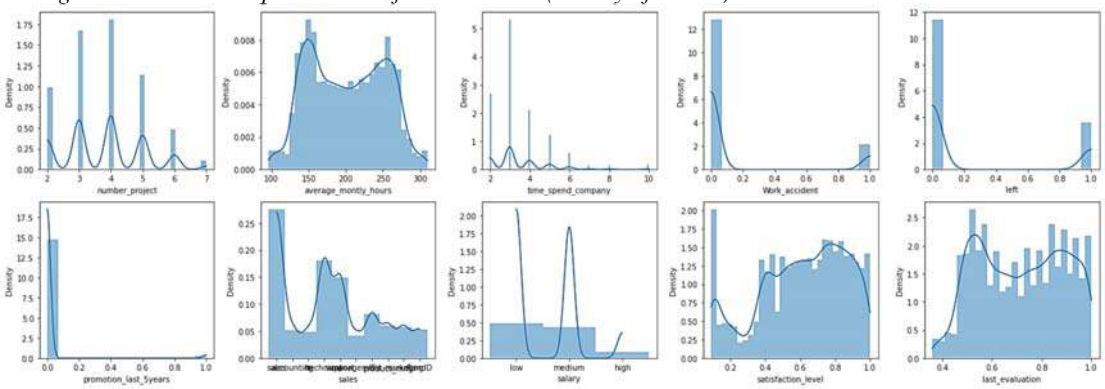
number _project	average _montly _hours	time_spent _company	work_ accident	left	promotion_ last_5years	sales	salary	Satisfaction _level	last_ evaluation
2	157	3	0	1	0	sales	low	0.38	0.53
5	262	6	0	1	0	sales	medium	0.8	0.86
7	272	4	0	1	0	sales	medium	0.11	0.88
5	223	5	0	1	0	sales	low	0.72	0.87
2	159	3	0	1	0	sales	low	0.37	0.52

Source: KAGGLE: https://www.kaggle.com/giripujar/hr-analytics

1. number _project – the number of the project in which the employees participate [no. from 2 to 7],
2. average _monthly _hours – the numbers of hours each employee worked [96÷310, hours),
3. time_spent _company – time spent by the company for employee [2÷7, hours],
4. work_ accident – if the employee had a work accident or not [0 or 1],
5. left – if the employee left the organization or not [0 or 1],
6. promotion_last_5years – promotion of the employee in last 5 years [0 or 1],
7. sales – if the employee is in sales or other department of organization [sales],
8. salary – the level of salary [low, medium, high],
9. satisfaction _level – the level of employee satisfaction [0.09÷1],
10. last_ evaluation – evaluation value of the employee [0.36÷1].

The data are represented in table 1, Also, the data evolution is represented in figure 1 with seaborne function histplot.

Figure no. 1. Initial representation for all 10 data (density of values)



Source: Authors' representation in Python 3.

After a simple analysis of the data, the categorical data were eliminated (sales and salary). The results are presented in table 2, after an analysis of the data:

Table no. 2 Database analysis.

	number_ project	average_ monthly_ hours	time_ spent_ company	Work_ accident	left	promotion_ last_5years	satisfaction_ level	last _evaluation
count	14999	14999	14999	14999	14999	14999	14999	14999
mean	3.803054	201.0503	3.498233	0.14461	0.238083	0.021268	0.612834	0.716102
std	1.232592	49.9431	1.460136	0.351719	0.425924	0.144281	0.248631	0.171169
min	2	96	2	0	0	0	0.09	0.36
25%	3	156	3	0	0	0	0.44	0.56
50%	4	200	3	0	0	0	0.64	0.72
75%	5	245	4	0	0	0	0.82	0.87
max	7	310	10	1	1	1	1	1

Source: Authors' analysis in Python.

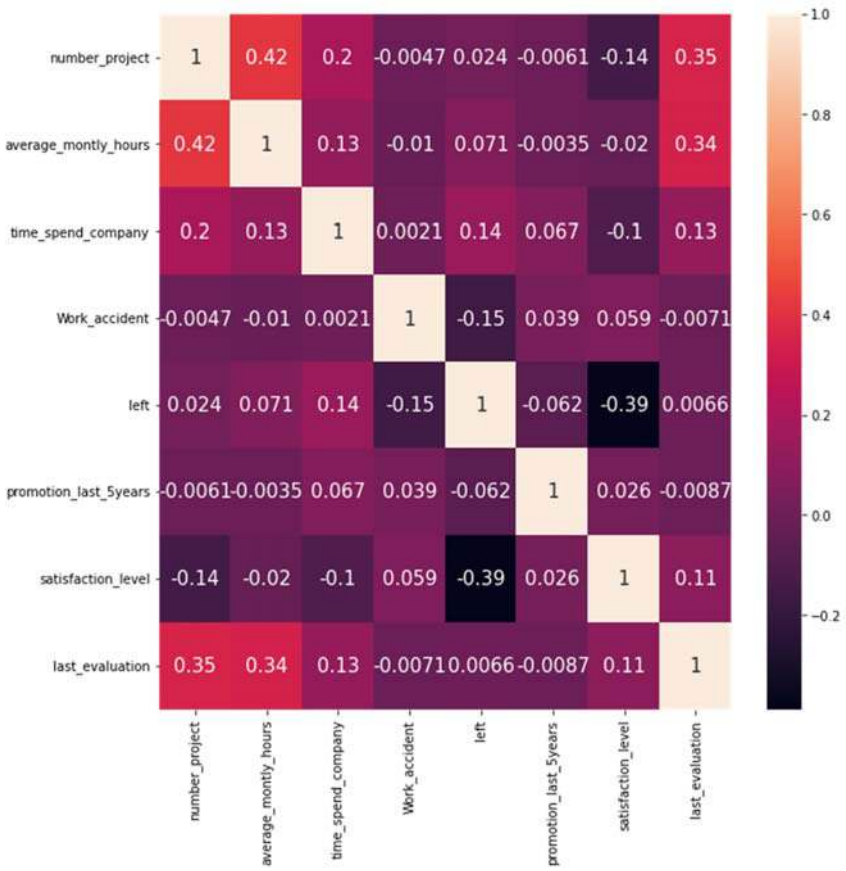
For further analysis and visual representation of data, the following functions were used:

- pandas.DataFrame.corr - Compute pairwise correlation of columns, excluding NA/null values,
- sns.heatmap - Plot rectangular data as a color-encoded matrix.
- sns.jointplot - Draw a plot of two variables with bivariate and univariate graphs.
- plot_trisurf - creates a surface by first finding a set of triangles formed between adjacent points.
- scatter3D - is a mathematical diagram, the most basic version of three-dimensional plotting used to display the properties of data as three variables of a dataset using the cartesian coordinates.
- scatter_3d - create a 3D scatter plot.

4. Findings

The research first application of the correlation function is shown in figure 2, with values between [-0.39; 0.42].

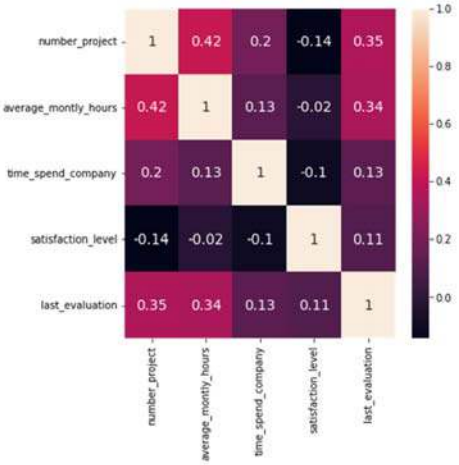
Figure no. 2. Initial correlation between all 10 data.



Source: Authors' representation in Python 3.

Further, the eight data were diminished to five by choosing the data with correlation values bigger than 1 in accordance with last_evaluation. The result to minimizing the correlation is showed in figure 3. Only four data met these criteria: satisfaction level, average monthly hours, number of project, and time spent by company, against the last_evaluation.

Figure no. 3. Final correlation. Chosen accordingly to correlation values min. 0.1

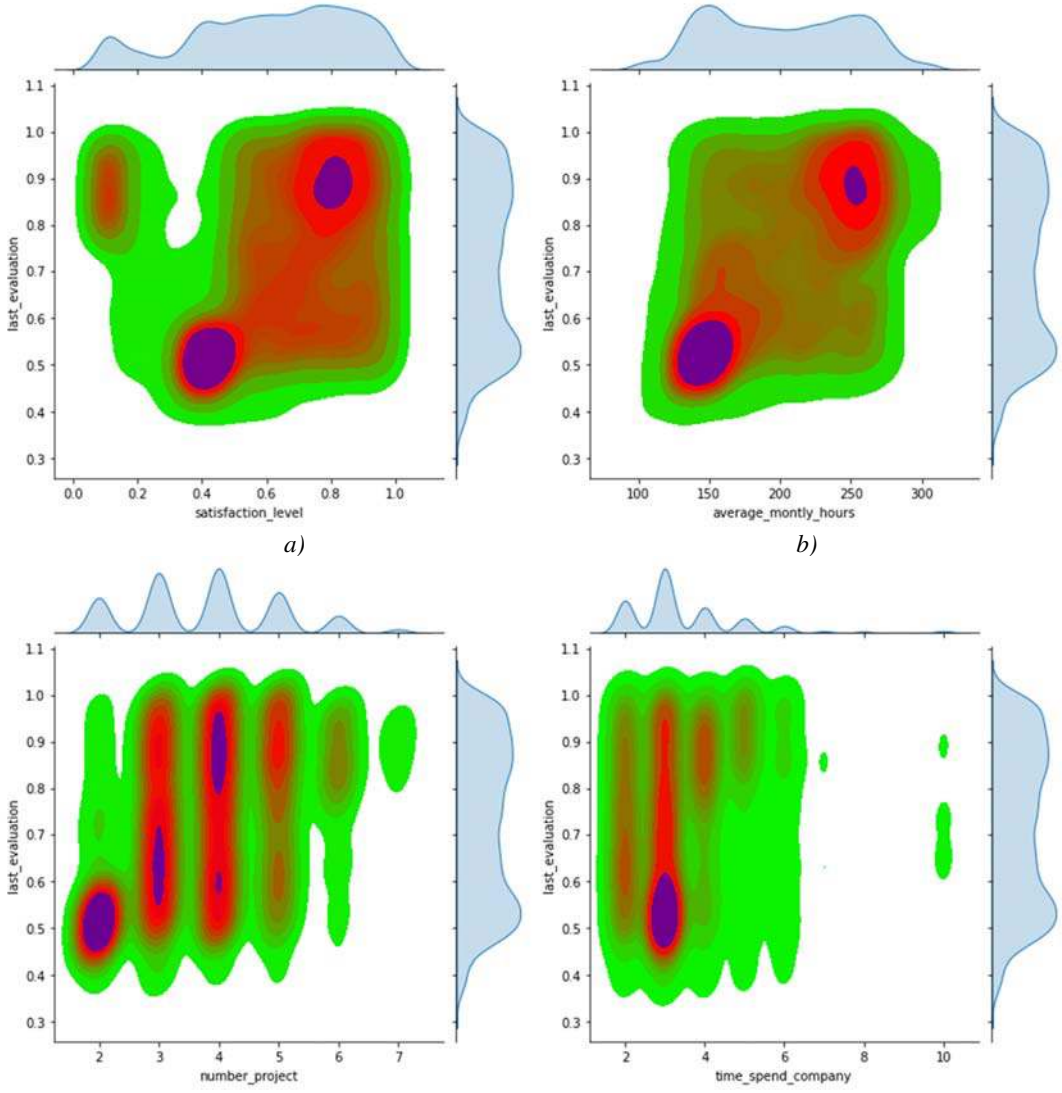


Source: Authors' representation in Python 3.

From this point only these five data will be used, and, also, the last_evaluation will be considered the target data, against which all the other four data sets will be correlated. The last_evaluation is the target of the management, and it will be used for the next visual representations.

In other to understand how the data are correlated, the last_evaluation was represented in figure 4 against each of the other four data sets.

Figure no. 4. Plot of last evaluation against and each of other 4 variables: a) satisfaction level; b) average monthly hours; c) number of project; d) time spent in company.



Source: Authors' representation in Python 3.

Figure 4.a shows that most employees have a satisfaction level divided in 2 groups, first between [0.3; 0.5], and second [0.75; 0.85]. The group with the highest evaluation is the second one, which reveals that there are two groups of employees: the first has a low satisfaction level, yet does not work for better evaluation, and second that demonstrates that employees with high satisfaction score also high in evaluation.

The graphic in figure 4.b proves that employees with more average monthly hours of work score high on evaluation. Also shows two groups: one that works less than 175 hours monthly and the other group that works more than 250 hours monthly.

Figure 4.c shows that most employee worked at project number 2, followed by project 4 and then 3. From the three projects, project number 4 has the employees with the highest evaluations.

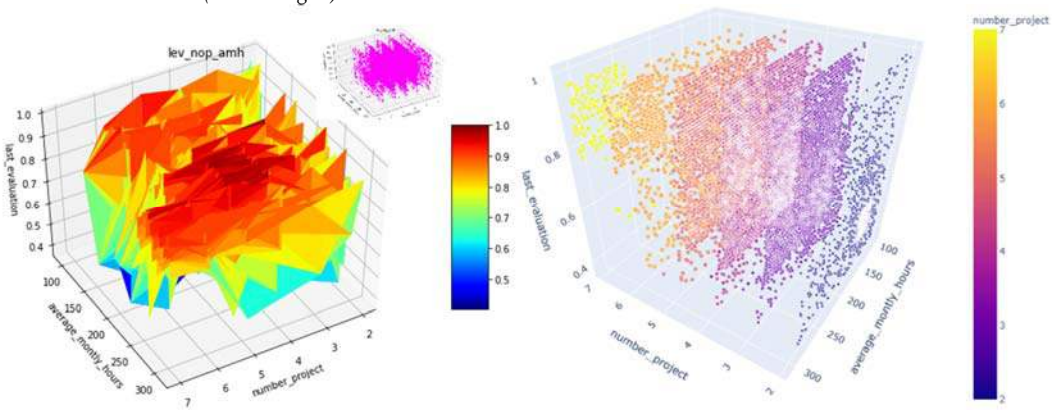
In figure 4.d one can see that the employees that organization spent 3 days on training/tutoring are the most of them. Also, they are most appreciated in the evaluation.

The first conclusion is that the employees with the highest evaluation scores have a satisfaction level between [0.75; 0.85], work 250 hours monthly, worked especially in the 4th project, and were helped by the organization with programs about 3 hours.

After establishing an initial group that scores the highest evaluation points, another visual analysis was made by representing in 3D the values of last_evaluation in accordance with each of the other three data considered 2 by two. This kind of representation is shown in figures 5-10. For a better understanding of the graphics the following information must be followed:

- the 3D surface representation (on the left of the figures) has the highest point marked in red, and those are the areas that we are interesting in, as they represent the highest evaluation scores,
- to the 3D surface were added a smaller image of the scatter plot,
- on the right there is the scatter plot of the data,
- the largest numbers of employees clusters are represented in white areas.

Figure no. 5. Variation of data evaluation accordingly to average monthly hours and number of projects (on the left) and scatter data points for average monthly hours and number of projects in the last evaluation 3D view (on the right).



Source: Authors' representation in Python 3.

Analyzing figure 5 the largest numbers of employees pertain to projects numbers 3 and 4 and works between 250 and 300 hours monthly. The data with the highest evaluation scores belong to values between 3 and 4 on projects and monthly hours between 200 and 250. As a intersection of data the best evaluation score belongs to employees that worked to project 4 and worked 250 hours monthly.

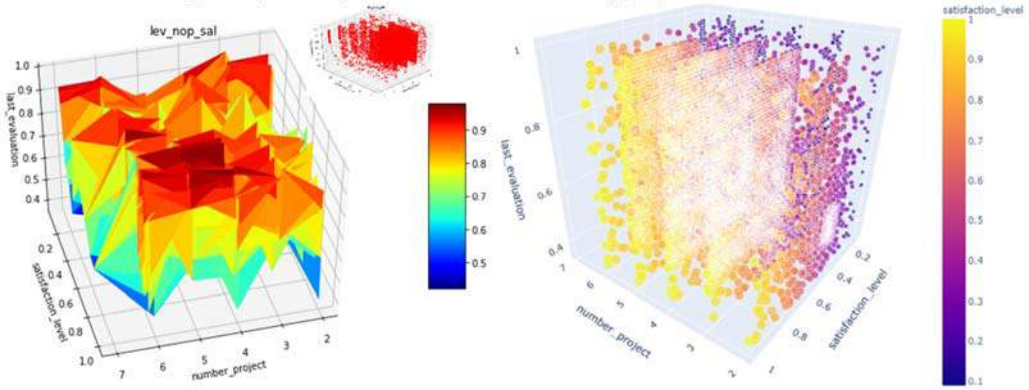
In figure 6 the most employees worked on projects 3 to 5 and have a satisfaction level between 0.5 to 0.8. The best evaluation score belongs to projects number 4 and 5 and a satisfaction level between 0.6 and 0.8.

Figure 7 sets the main employees from 0.6 to 0.9 as satisfaction level, and monthly hours worked from 150 to 275. The highest evaluation score appears in the 200-250 hours worked monthly, for satisfaction levels between 0.8 and 1.00.

The graphic in figure 8 shows that most employees belong to the crowd in limits between 0.5 and 0.9 for satisfaction level and time spent by the organization between 2 and 4 days. The highest score of evaluation belongs to satisfaction level of 0.9 and 4 days of time spent by organization.

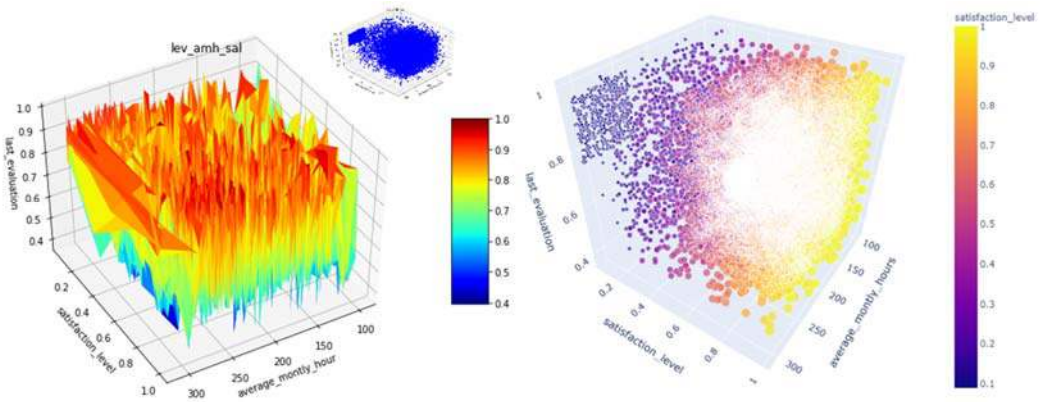
In figure 9, the larger numbers of employees fit in time spent by the company among 2 and 5 as time spent by the company and 150-275 average monthly hours. The highest evaluation scores can be found between time spent from 3 to 5 and average monthly hours between 200 and 250 hours.

Figure no. 6. Variation of data evaluation accordingly to satisfaction level and number of project (on the left) and scatter data points for satisfaction level and number of project in the last evaluation 3D view.



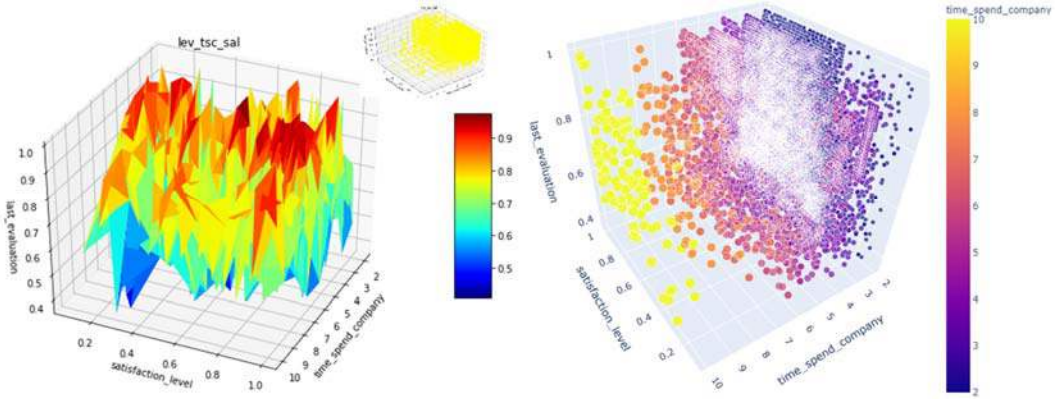
Source: Authors' representation in Python 3.

Figure no. 7. Variation of data evaluation accordingly to satisfaction level and average monthly hours (on the left) and scatter data points for satisfaction level and average monthly hours in the last evaluation 3D view.



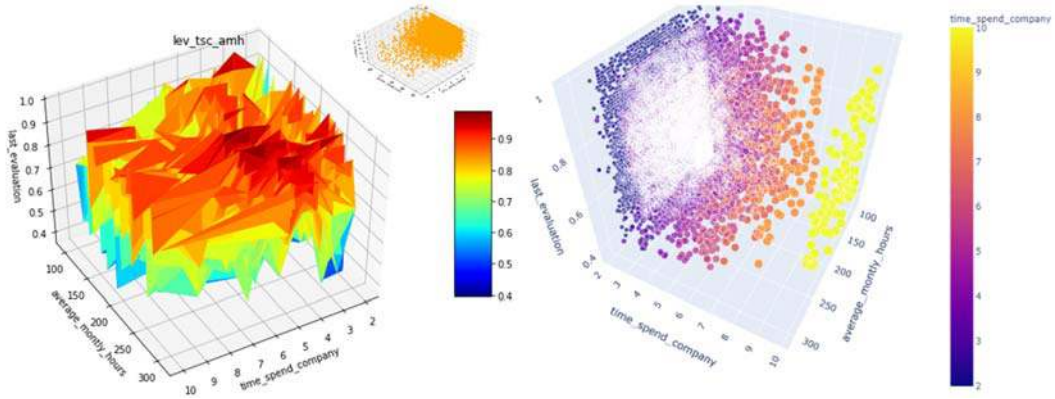
Source: Authors' representation in Python 3.

Figure no. 8. Variation of data evaluation accordingly to satisfaction level and time spent in company (on the left) and scatter data points for satisfaction level and time spent in company in the last evaluation 3D view.



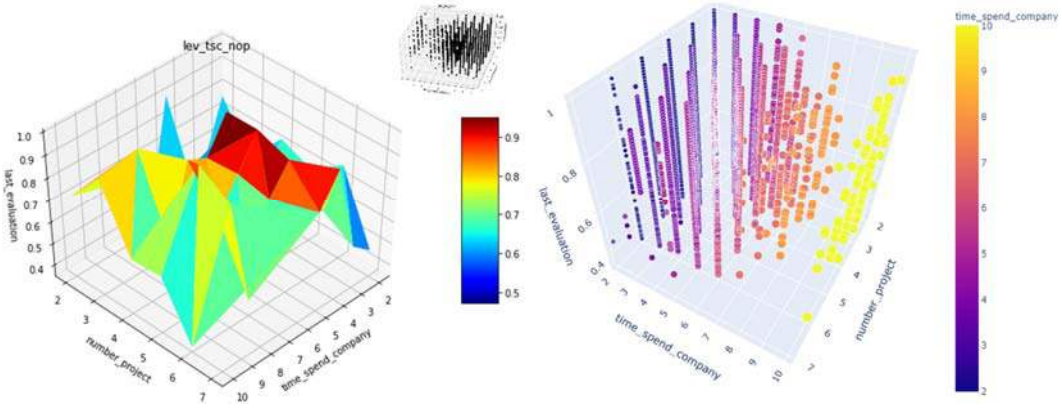
Source: Authors' representation in Python 3.

Figure no. 9. Variation of data evaluation accordingly to average monthly hours and time spent in company (on the left) and scatter data points for average monthly hours and time spent in company in the last evaluation 3D view.



Source: Authors' representation in Python 3.

Figure no. 10. Variation of data evaluation accordingly to number of projects and time spent in company (on the left) and scatter data points for number of projects and time spent in company in the last evaluation 3D view.



Source: Authors' representation in Python 3.

Figure 10 reveals that the project that most employees participated in are 3 to 5, and time spent by the company 4 and 5 days.

5. Conclusions

The data visualization can be used to represent large amount of data to analyze their evolution and make decisions in accordance with the evaluated data. Using open-source software, managers can develop algorithms and tools of representation of data.

In the present paper after applying Google Colab Notebooks Python, the authors developed 2D and 3D figures that show the relations between the number of the project in which employees participated, the numbers of hours each employee worked, time spent by the company for the employee, the level of employee satisfaction, and evaluation value of the employee. The results of the visual analysis showed that the employees with the best evaluation score participated in the projects 3 and 4, worked between 250 and 275 hours monthly, were included in 4 days of company time and have a satisfaction level of 0.5 to 0.8.

Starting from these results the management should concentrate in implementing the above data values and sets to maximize the evaluation values.

6. References

- Janvrin, D., Raschke, R. Dilla, W., 2014, *Making sense of complex data using interactive data visualization*, Journal of Accounting Education, Volume 32, Issue 4, pp. 31-48.
- KAGGLE: https://www.kaggle.com/giripujar/hr-analytics
- Wu, E. et al. 2017, *Combining Design and Performance in a Data Visualization Management System*, Conference on Innovative Data Systems Research (CIDR), Available at http://www.cs.columbia.edu/~fotis/pubs/papers/dvms-cidr17.pdf [10.12.2021].
- Zheng, J, G, 2021, *Data Visualization for Analytics and Business Intelligence: A Comprehensive Overview*, Conference: IT 7113, Available at https://www.researchgate.net/publication/327578825_Data_Visualization_for_Analytics_and_Business_Intelligence_A_Comprehensive_Overview [10.12.2021].
- https://matplotlib.org/stable/#.
- https://pandas.pydata.org/docs/index.html.
- https://seaborn.pydata.org/index.html.