International Journal of Latest Engineering and Management Research (IJLEMR)
ISSN: 2455-4847
www.ijlemr.com || Volume 08 – Issue 03 || March 2023 || PP. 139-152

Analyzing the data of the admitted applicants using Correspondence Analysis

Tran Thi Hue[1],
[1](*Faculty of International Training, Thai Nguyen University of Technology, Vietnam)*

Abstract: This article reveals the study on the data of admitted applicants who choose to take their diploma at Thai Nguyen University of Technology, one of the oldest and on the top-tier of the engineering and technology universities in Vietnam. The inside look of the data provides the managements and education scientists a useful view about the fact and the strength of the trademark of the university. This may help them in suggesting the idea of developing the trademark, the reputation as well as improving the quality of training the high labor quality and intellectual labor supplying to the manufacturing and industrial market. The study uses a quite advancing technology of Data Science which is suitable for the categorical data to reduce the dimension of the data set in order to find out the relationship between two variables in a contingency table of the answer about the research question which asks the admitted applicants the reason that they felt the most important to choose the university for completing their diploma in an engineering major.

Keywords: Admitted applicant data, Categorical data, Correspondence analysis, university trademark, education sciences.

I. INTRODUCTION

The paper approaches the enrollment problem at Thai Nguyen University of Technology based on the analysis of student data collected in the school year 2021 in order to provide topical and traditional factors which influences the choice of training institution. Analyze the admitted student data based on different aspects of information, including: hometown, academic achievement, personal orientation. We want to build a theoretical basis for the problem of analyzing enrollment data of Thai Nguyen University of Technology.Analyzing and comparing by using many different methods for different research questions about the relationships between factors affecting the decision of learners to choose a training institution. From here, conclusions are made on the basis of evaluating the positivity of the media for the school's enrollment. Here, the theoretical-based approaches of statistics (including the study of the ordinal logistic regression models in which the dependent variables are of the ordered data) and the Machine Learning theory of Correspondence analysis to make an assessment of the dependence between the determinants on the tendency of learners to choose a training institution.

We evaluate the applicability and effectiveness of the above data analysis methods. Statistical inferences and machine research inferences given provide a scientific basis for conducting research and answering the research questions.

Finally, we build the implementation for these of methods and checking the rationality and efficiency of the methods built on the R programming language.

II. DESCRIPTION OF THE DATA

The survey asking the admitted applicants (customers of the university) arrange in the decreasing order of importance for the reason that provokes them to choose the university for earning their diploma. The answers to this question (components of the variable *opinions*) and their notation is given in the Table 1.

Notation	Answer	Notation	Answer
job	Many opportunities for a good future career and position.	*passion*	Be interested in Technology and Engineering
tuition	Low cost of tuition and living (renting, food, customer goods)	*dorm*	Convenient dormitory
facility	Modern facility to study (libraries, classroom, laboratory, training workshop)	*lecturer*	Excellent/Conscientious lecturers
activity	Diverse outclass activity (English club, Japanese club, Korean club, Chinese club, Music club, gymnastics club, Volunteer Youth club, etc.)	*oversea*	Potential oversea internships/jobs (Engineers graduated have chance to visit and work in a developing country such as Japan, Korea, Taiwan, China, Germany, USA etc.)

other	Other answers

Table 1. Answers (which are considered as variables and the components of the variable opinions) and their notations.

The data was collected from 889 freshmen who have just admitted to university in the academic year 2021-2022. This survey asked students to find out what is the important order for the reasons by which they are persuaded to decide their study in the university. Each answer is assigned to a values *a, b, c, d, e, f, g, h, j* for the decreasing order of the important factors and the value *m* for an answer which has not been assigned such a value. The visualization of the data description is presented in the next section. The data is shown in Table 2.

		Values (components of the variable *options*)									
		a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*j*	*m*
opinions	*job*	380	91	57	23	16	13	9	9	0	291
	tuition	71	198	54	37	24	28	20	5	0	452
	facility	29	94	169	60	39	23	11	1	0	463
	activity	32	80	102	97	36	30	28	14	1	469
	passion	174	118	88	48	67	19	9	3	0	363
	dormitory	19	25	18	25	27	58	43	29	1	644
	lecturer	66	92	116	65	48	18	55	6	0	423
	oversea	13	27	36	41	17	21	23	95	1	615
	other	3	0	0	1	2	1	0	2	11	869

Table 2. Data about the frequency of the values assigned to a component of the variable opinions.

III. VISUALIZATION OF THE DATA

This section aims at representing the data in the descriptive statistics. This has a meaningful sense in drawing the conclusion on the overview of data. It allows us to have a great glimpse through the relationship between each answer and the most possible order of the importance to which it is assigned. This visual representation tells us the appearance of the story that data contains. We study here some basic plot on data provided in Table 2.

The tackle and un-tackle bar charts showing the frequency of each value of the importance to which the corresponding answer is assigned

The tackle and un-tackle bar chart in this section is classified into two groups: One has the appearance of the null value *m*, the other does not has it. This value may tell us the level of concerning of the student answering the research question. This also shows us the focus of that concerning or the importance of the answers to which a value is assigned. The Figure 1 and 2 are the bar charts showing the frequency with the value *m* taken into account while the Figure 3 and 4 are not. The Figure 5, 6, and 7 shows the frequency of the answers by each value of the importance. From each chart, we may get a useful fact about the data and the description of the relationship between an answer and a value. The charts conveys the fact that the answer *job* and the value *a*, *tuition* and *b*, *facility* and *c*, *activity* and *c*, *lecturer* and *e*, *dormitory* and *f*, *oversea* and *h* are most likely to be the matching order of the importance. For the value *m*, the less concern is most likely to be arrange in quite the same order except for the case of the answer *passion* which seems to be taken more care than the second class of the importance *b*, the answer *tuition*. This phenomenon can also be explained since the fact that the answer *passion* has quite high frequency of the class *a*. And to get a deeper look, we need more advance tool to analysis this data. And this is the great thank to the Correspondence Analysis. The powerful tool will help to reduce the dimension of the data and extract the intrinsic feature of the research question. This is presented in the next section.

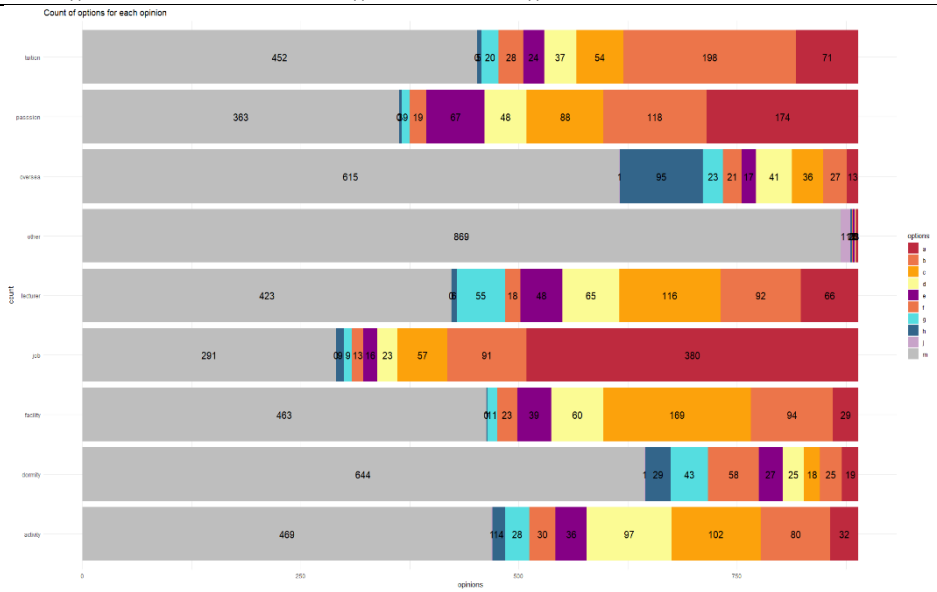

Figure 1. Tackle bar chart shows the frequency of the values including m which are assigned to each answer.

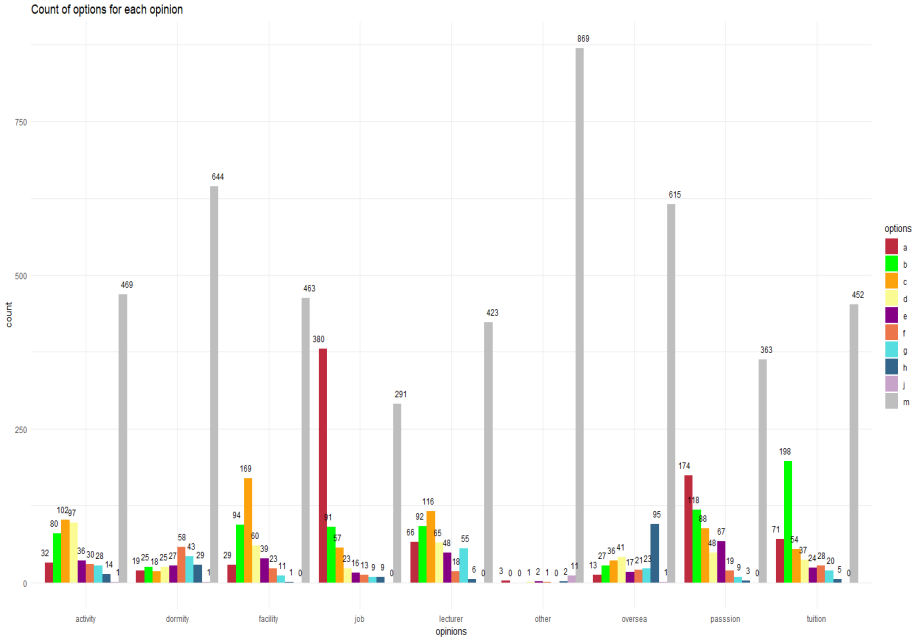

Figure 2.Un-tackle bar chart shows the frequency of the values including m for each answer.

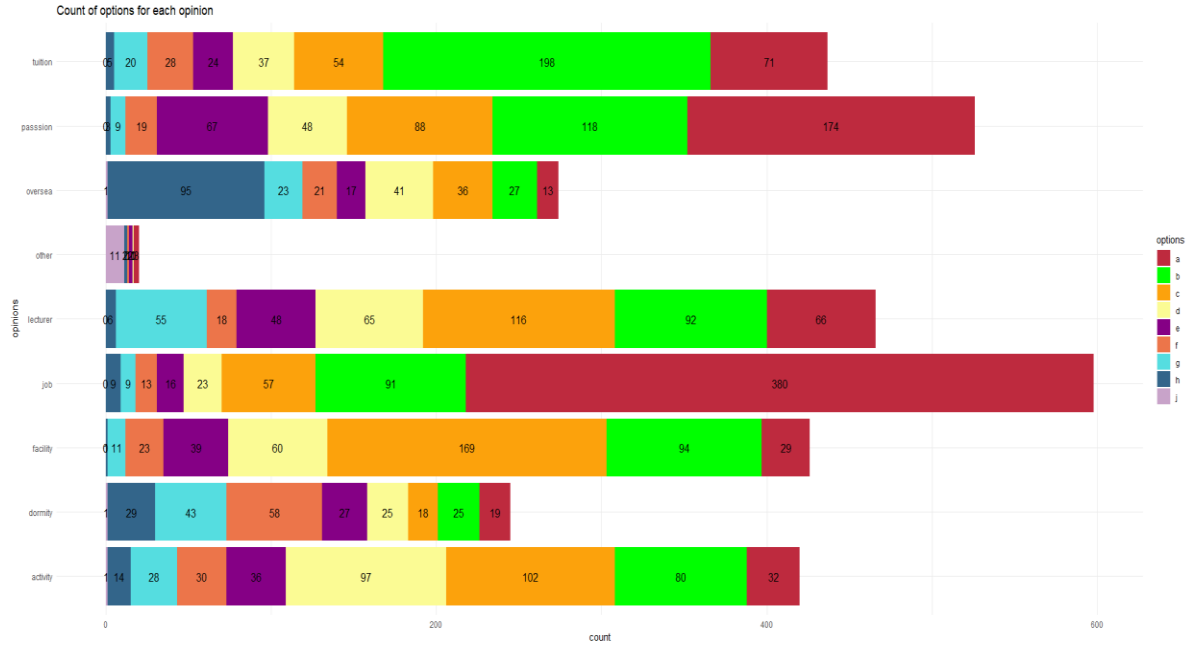

Figure 3. Tackle bar chart shows the frequency for the values of the importance including the value m which are assigned to each answer.

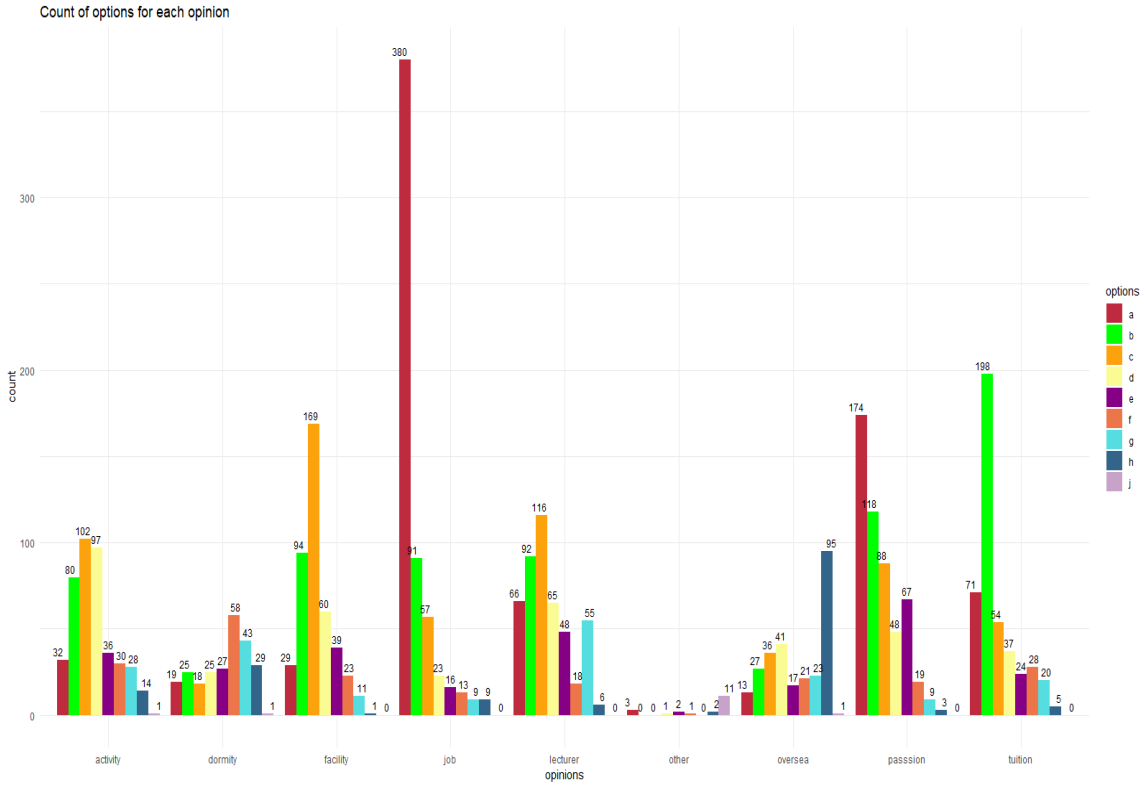

Figure 4. Un-tackle bar chart of the frequency for the values of the importance excluding the value m

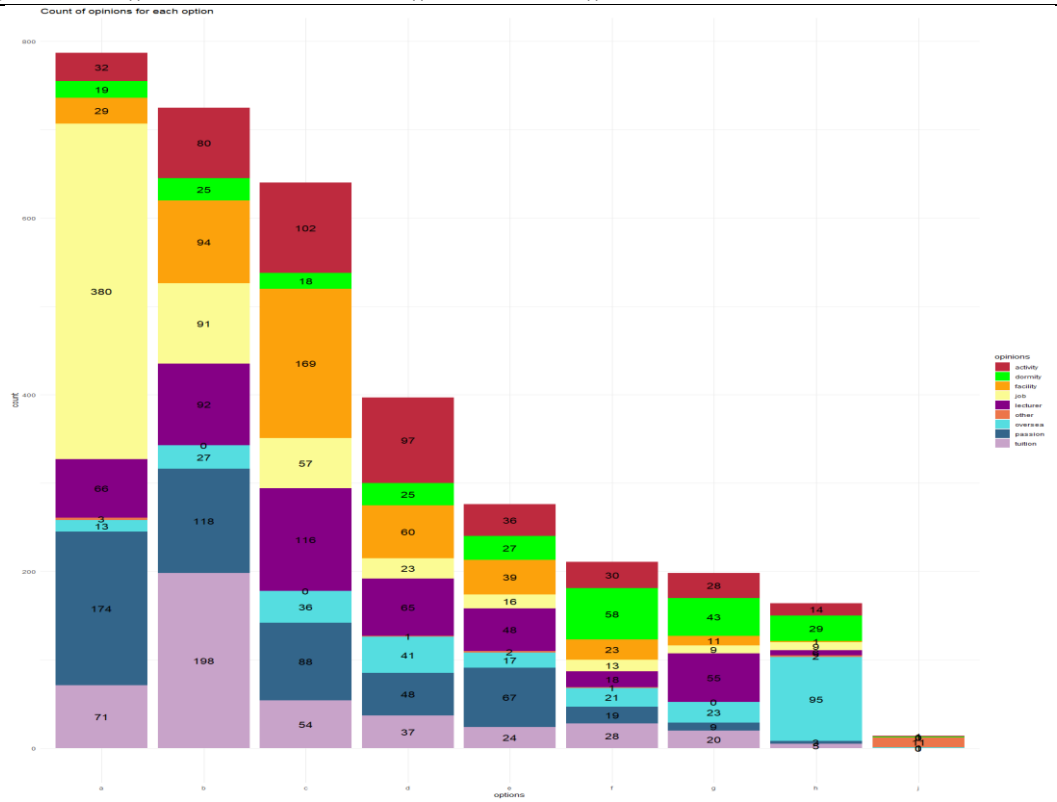

Figure 5. Tackle bar chart shows the answers to which each value of the importance is assigned excluding the value *m.*

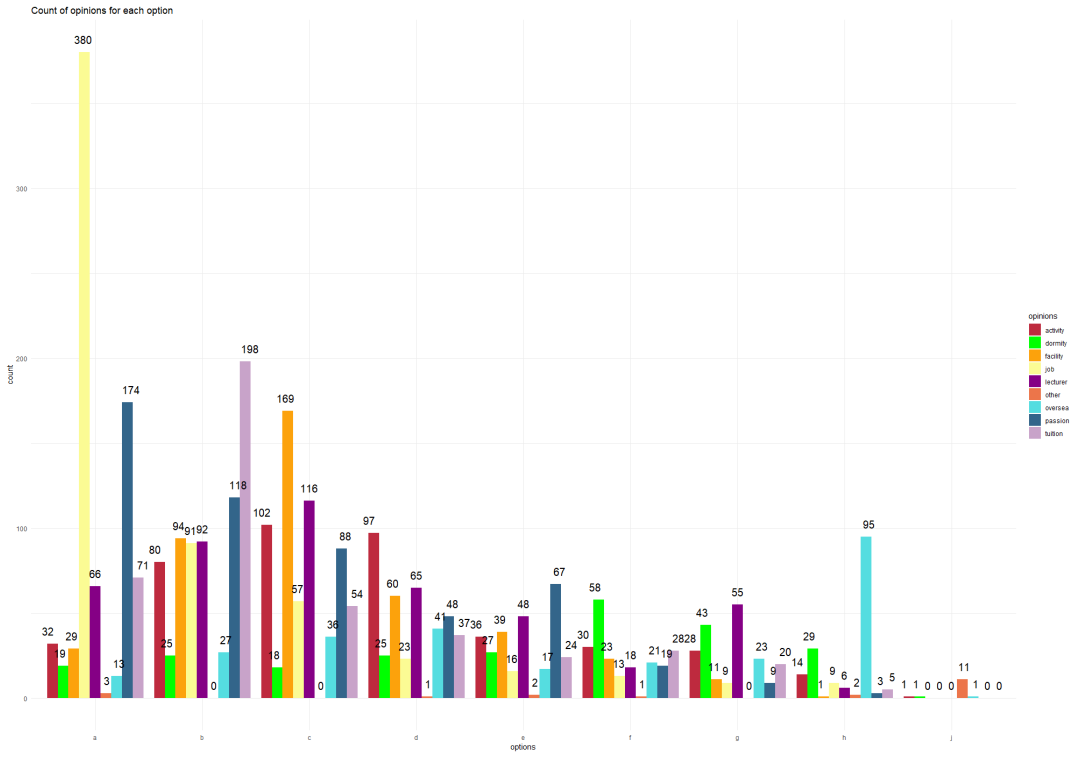

Figure 6. Un-tackle bar chart shows the frequency of the answers to which each value of the importance is assigned.

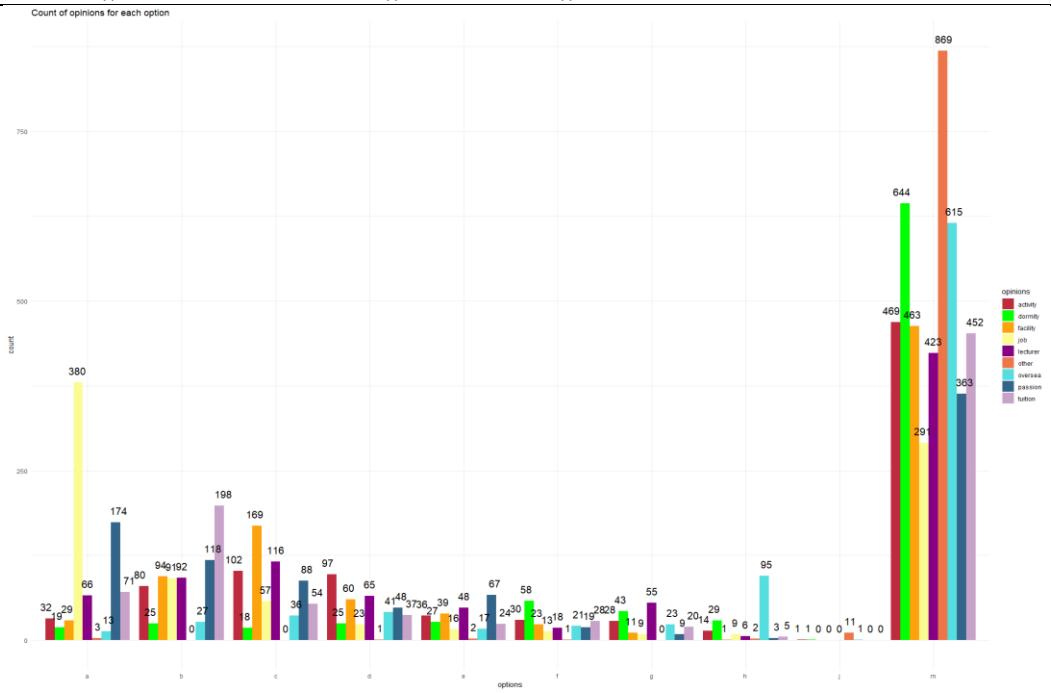

Figure 7. Un-tackle bar chart shows the frequency of the answers to which each value of the importance is assigned including the value *m.*

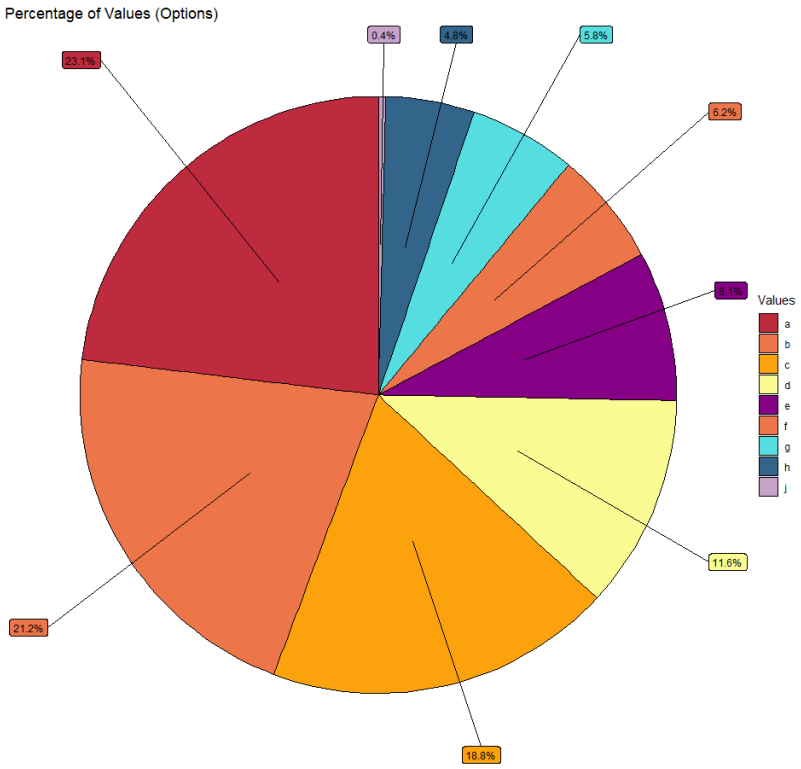

Figure 8. Pie chart represents the percentage of the values of the importance which are assigned to all over the answers.

Pie chart of percentage for the level of the importance: The Figure 8 represents the percentage of the values of the importance assigned to overall answers. This chart describes the percentage in Table 3, excluding the value *m.* This chart shows the three dominant levels of the importance which makes more than two-thirds of the total values assigned. After these three dominant levels, the next one falls only nearly a half of the previous dominant level. The strength of concern reduces quite high.

Values/ Options	a	b	c	d	e	f	g	h	j	sum
Count	787	725	640	397	276	211	198	164	14	3412
Percentage (%)	23	21.3	18.8	11.6	8.1	6.2	5.8	4.81	0.41	100

Table 5. Percentage of the values of the importance *a, b, c, d, e, f, g, h, j* assigned to overall answers.

IV. CORRESPONDENCE ANALYSIS

This section is served to present the achievement of the correspondence analysis ([1-5]) in order to extract an inside look through the data set which enables us to reveal the intrinsic relationship/ the similarity between an answer and a potential values of the importance which the answer may be assigned.

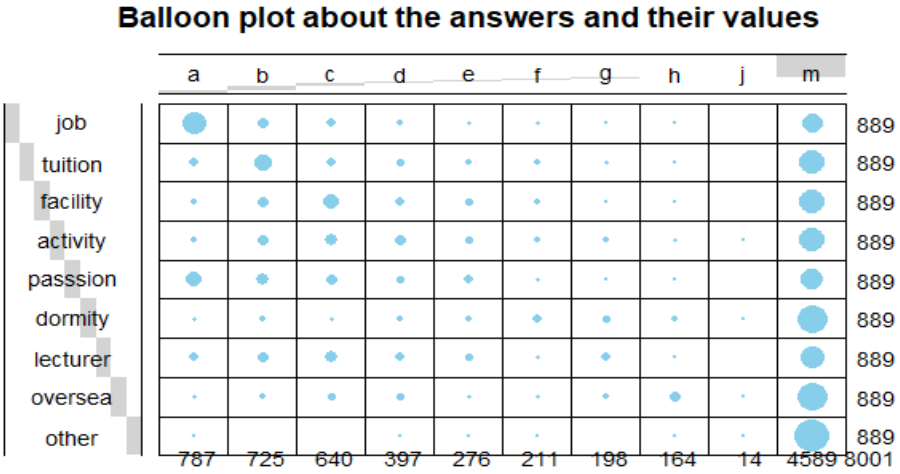

Figure 9. Balloon plot with the value m.

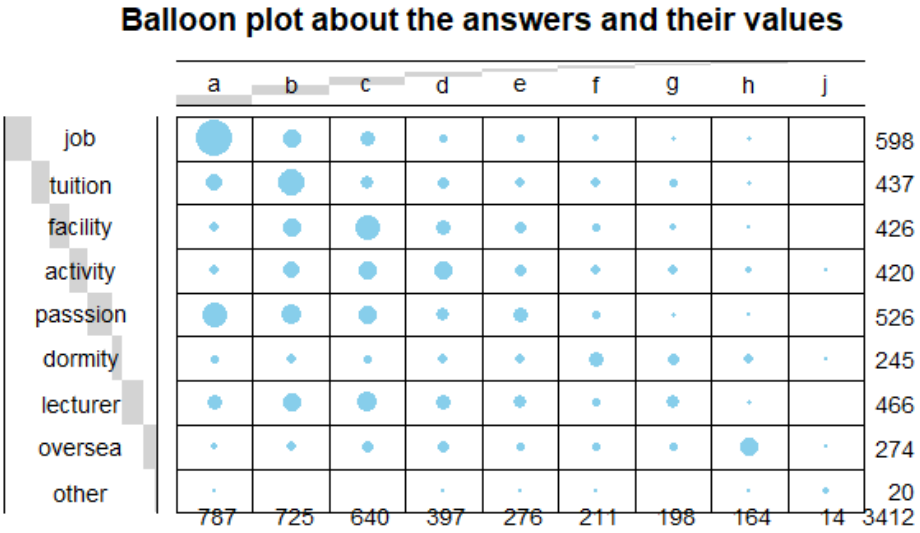

Figure 10.Balloon plot without the value m

We first construct two balloon plots in Figure 9 and 10 to show these relationship. Figure 9 considers the value *m* while Figure 10 does not. From these plots, we can observe a connection between the row variables (the answers) and the column variables (the values of importance). This connection is reduced in a decreasing direction along the main diagonal of the contingency table. The difference between Figure 9 and Figure 10 presented in the row margin values. These numbers reveals the fact of very high concern of the answers *job, passion, lecturer, tuition, facility* and *activity*. These values are assigned above 400 over total 889 students answering on the question. This proves that the students have seriously selected the university to follow the destiny of their future career basing on their personal perception about the society and the national economy, the development of the country and the local industrial regions.

Opinions	a	b	c	d	e	f	g	h	j
job	137.9	127.1	112.2	69.6	48.4	37	34.7	28.7	2.5
tuition	100.8	92.9	82	50.8	35.3	27	25.4	21	1.8
facility	98.3	90.5	79.9	49.6	34.5	26.3	24.7	20.5	1.7
activity	96.9	89.2	78.8	48.9	34	26	24.4	20.2	1.7
passion	121.3	111.8	98.7	61.2	42.5	32.5	30.5	25.3	2.2
dormitory	56.5	52.1	46	28.5	19.8	15.2	14.2	11.8	1
lecturer	107.5	99	87.4	54.2	37.7	28.8	27	22.4	1.9
oversea	63.2	58.2	51.4	31.9	22.2	16.9	15.9	13.2	1.1
other	4.6	4.2	3.8	2.3	1.6	1.2	1.2	1	0.1

Table 6. The expected frequency of the values of the importance assigned to each answer.

Opinions	a	b	c	d	e	f	g	h
job	138.34	127.93	112.93	69.88	48.35	37.06	34.94	28.59
tuition	101.09	93.49	82.53	51.06	35.33	27.08	25.53	20.89
facility	98.55	91.13	80.45	49.78	34.44	26.40	24.89	20.36
activity	96.93	89.64	79.13	48.96	33.88	25.96	24.48	20.03
passion	121.68	112.53	99.33	61.46	42.53	32.59	30.73	25.14
dormitory	56.45	52.20	46.08	28.51	19.73	15.12	14.26	11.66
lecturer	107.80	99.69	88.00	54.45	37.68	28.88	27.23	22.28
oversea	63.15	58.40	51.56	31.90	22.07	16.92	15.95	13.05

Table7. The expected frequency of the values of the importance except for the values j and m and the answer other from Table 6.

The expected frequency of the values of the importance are presented in Table 6 and 7.

Chi-squared test for the dependence between row and column in Figure 10: X-squared = 3474.3, df = 64 and p-value less than 2.2e-16. This test result shows the strong evidence of the dependence between the row and the column of the contingency table. It clarifies the fact stated in the previous section in which the conclusion is based on the visual view point only.
Chi-squared test can also be applied for the expected frequency table shown in Table 6 and 7, and these results also support the same conclusion on the dependence of the row and column variables. For example, this test presented for Table 7 show the result: X-squared = 2000.7, df = 49, p-value < 2.2e-16, which strongly proves that statement.

Contribution of the values of the importance to Chi-squared: This result is calculated as

$$contrib \coloneqq R^2/\chi^2, \chi^2 = \sum \frac{(A-E)^2}{E}.$$

The Figure 11 shows this contribution in percent. The size and darkness of the balloon are proportional to the percentage of the contribution. Light colors show a negative contribution (which effects in decreasing the Chi-squared value) and dark colors show a positive one.

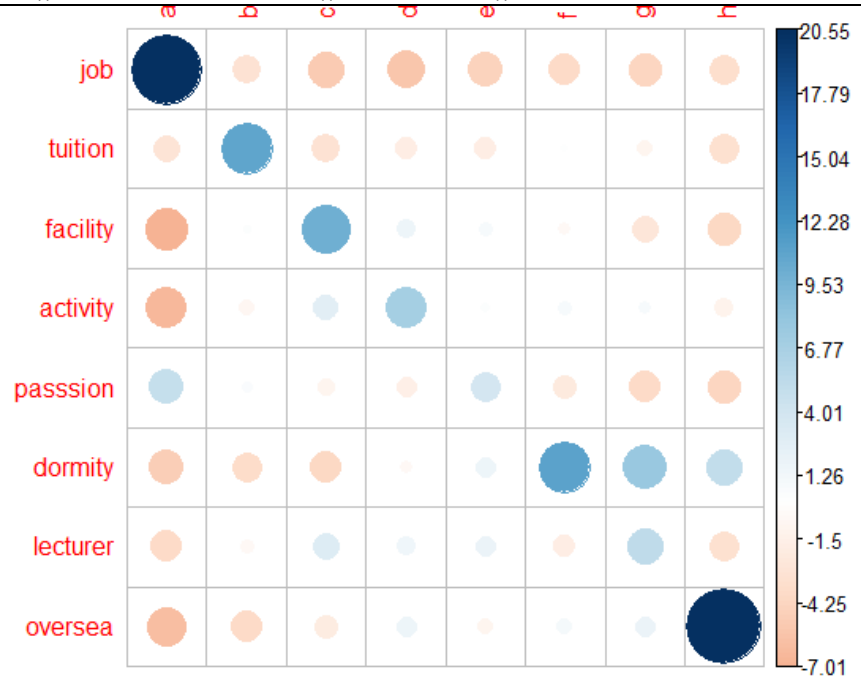

Figure 11. The contribution of each values of the importance to Chi-squared value.

Pearson's residual is another way to represent the impact of each value to the total residual of the data set. This table is shown below. The pair *job* and *a* makes a great positive contribution to the residual, but it still follows the pair *oversea* and *h.* But it is clear that the latter does not have a great sense in interpreting the answer to the research question in finding the most important reason for student to make his/her own decision of taking the study in the university. It is a great thank to this result that we could extract from here the positive effects which dominant the residual of the data set.

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*
job	20.546	-3.265	-5.263	-5.608	-4.652	-3.952	-4.388	-3.663
tuition	-2.993	10.809	-3.140	-1.968	-1.906	0.177	-1.095	-3.477
facility	-7.006	0.300	9.87	1.449	0.777	-0.661	-2.784	-4.291
activity	-6.595	-1.018	2.571	6.866	0.365	0.792	0.711	-1.347
passion	4.743	0.516	-1.137	-1.717	3.753	-2.381	-3.920	-4.416
dormitory	-4.984	-3.765	-4.136	-0.658	1.637	11.028	7.613	5.076
lecturer	-4.026	-0.770	2.985	1.430	1.682	-2.024	5.323	-3.448
oversea	-6.311	-4.109	-2.166	1.61	-1.080	0.993	1.765	22.685

Table 8. Pearson's residual of the answers and the values of the importance.

Alternative to Chi-squared test which can be used is the **G- test for the ratio likelihood**, which shows the test results are G = 1578.9, df = 49, p-value < 2.2e-16. This strongly supports the above conclusion on the dependence. The Figure 12 and Figure 13 represent the ratio likelihood between the row and column variables.

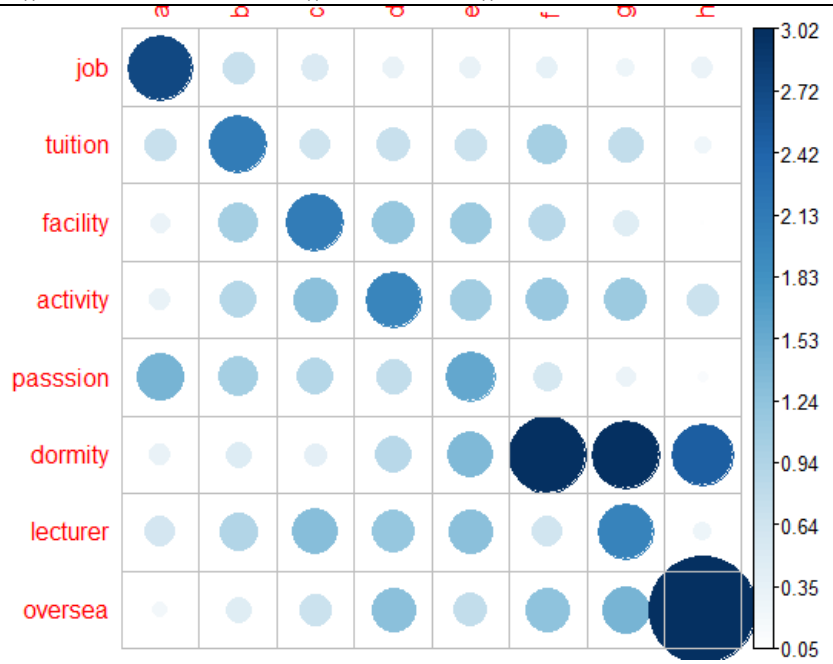

Figure 12. The ratio likelihood plot.

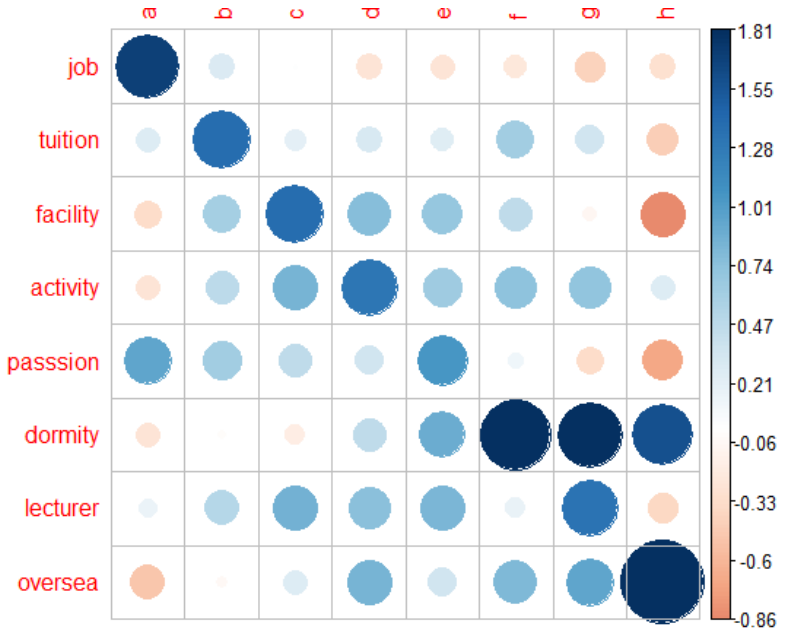

Figure 13. The logarithmic likelihood plot.

Eigenvalues in Correspondence Analysis CA

The eigenvalue in CA enables us estimate the amount of information retained in each axis. The dimensions arranged in a decreasing order of the explained information (i.e. the variance) which are kept by the corresponding axis. The following table will show these amount of information.

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.289	48.945	48.945
Dim.2	0.172	29.134	78.078
Dim.3	0.056	9.489	87.567
Dim.4	0.046	7.768	95.335
Dim.5	0.011	1.887	97.222

Dim.6	0.008	1.425	98.647
Dim.7	0.008	1.353	100.000

The Scree plot for CA

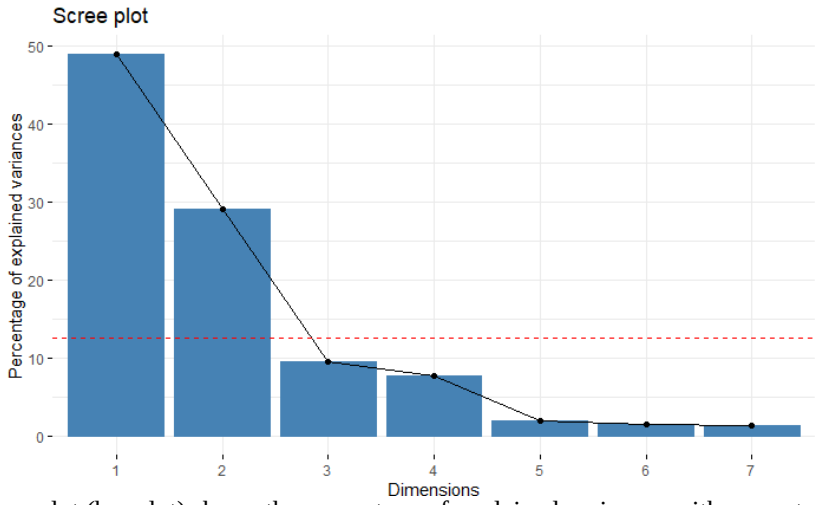

Figure 14.The Scree plot (bar plot) shows the percentage of explained variances with respect each dimension by CA.

The Scree plot shows the total explained variance by the first two dimensions to be about 78%, a quite large proportion. The third dimension falls below the dash red line which is about 12.5% standing for the barrier representing a significant level of the explained variance. This is very nice fact for representing the CA-Biplot of the data set in only two dimensions (two axes in the coordinate plane).

CA Biplot (Symmetric biplot)

In this plot, the column variables (standing for the values of the importance) and the row variable (standing for the answers) are show. The answer *other*, and the values *j*, *m* are supplementary variables. They are not be used to calculate the total explained variance. The percentage of the explained variance by axis 1 and axis 2 are show in the plot. The similarity between two answers or two values is shown by the distance between them, the closer the more similar. Also, we can draw an uncertain relationship amongst the answers and the values which are clustered by their similarity or the close dependence. The same conclusion can be drawn as stated in the section of the data visualization.

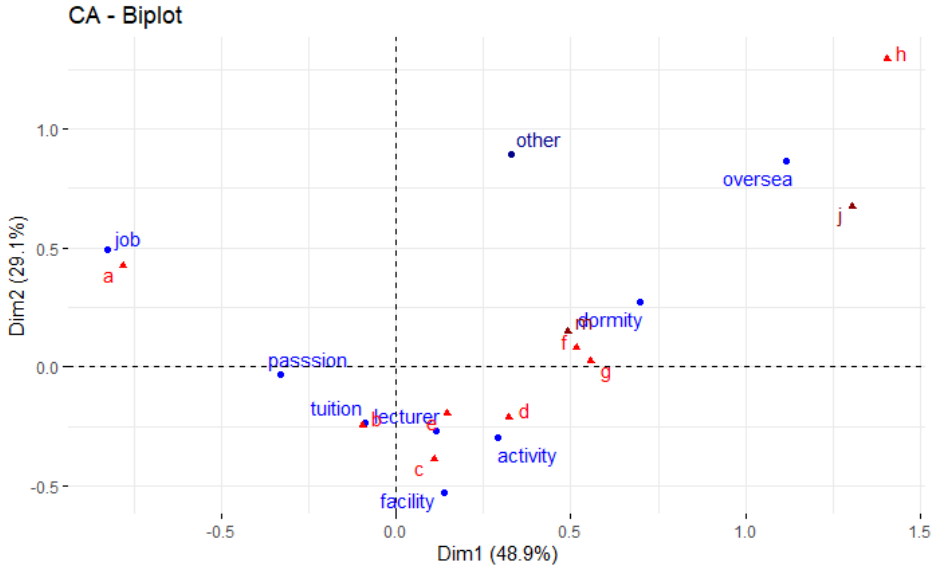

Figure 15. CA –biplot for the column (red triangle) and row (blue dot) variable. Supplementary column and row variables are shown in darkred triangle and darkblue dot respectively.

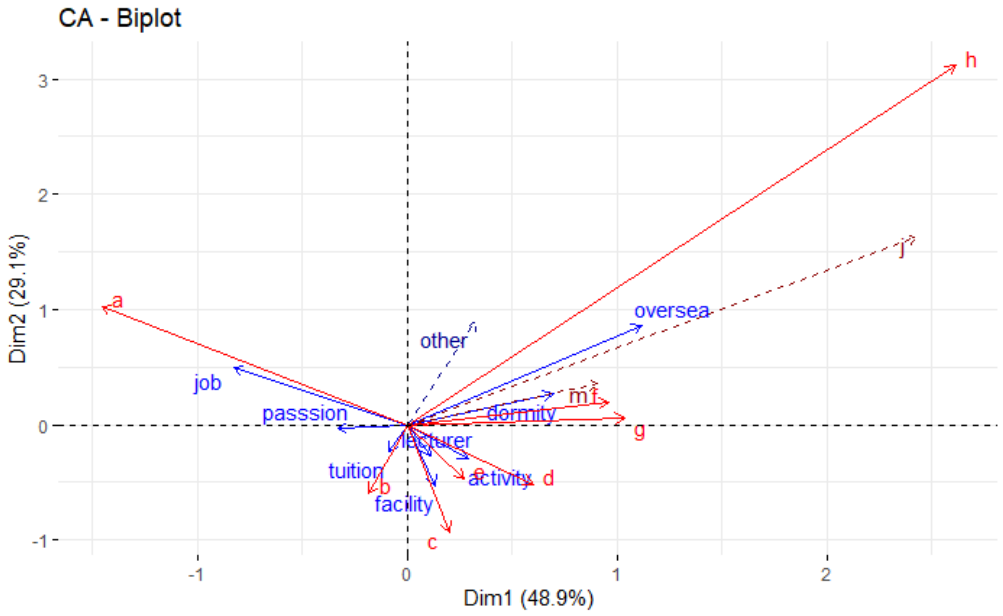

Figure 16. Asymmetric-biplot for the column and row variables

Asymmetric biplot

This plot allows us to interpret the distance between row and column variables. In the Figure 16, the row variable is represented in the principal coordinates while the column variable is in the standardized coordinate. Therefore, the plot is also called the preserving row variable. The meaning of the similarity between a column and row variable is shown by the inner product between two vectors originated at the origin and heading at the corresponding row and column points.

The coordinates of the column and row variables are shown in the following table:

For the row variable, the coordinates are given in the table below in each axis:

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
job	-0.827	0.492	-0.029	0.078	0.008
tuition	-0.089	-0.237	0.216	-0.496	0.012
facility	0.136	-0.528	-0.262	0.086	-0.173
activity	0.291	-0.300	-0.071	0.108	0.096
passion	-0.332	-0.031	-0.032	-0.001	-0.064
dormitory	0.696	0.272	0.661	0.270	-0.122
lecturer	0.113	-0.271	0.023	0.150	0.181
oversea	1.118	0.865	-0.332	-0.172	0.009

For the column variable, the coordinates are given below:

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
a	-0.783	0.426	-0.008	0.078	0.007
b	-0.097	-0.247	0.091	-0.377	0.011
c	0.109	-0.387	-0.276	0.139	-0.071
d	0.323	-0.214	-0.132	0.100	0.133
e	0.146	-0.196	0.048	0.137	-0.055
f	0.516	0.082	0.577	0.157	-0.249
g	0.558	0.023	0.477	0.251	0.260
h	1.406	1.293	-0.334	-0.225	-0.030

The Singular Value Decomposition allows us to construct these coordinates ([4-7]).

Besides, we could see the cos2 (cosine squared) to consider the quality of the representation of the CA-biplot. These quantities are calculated by

$$row.cos2 := \frac{row.coord^2}{d^2}, col.cos2 := \frac{col.coord^2}{d^2},$$

Where*row.coord, col.coord, d*are the row and column coordinates, and the distance from the average profile.

$$d^2 = d^2(row_i, average.profile) = \sum \frac{(row.profile_i - average.profile)^2}{average.profile},$$

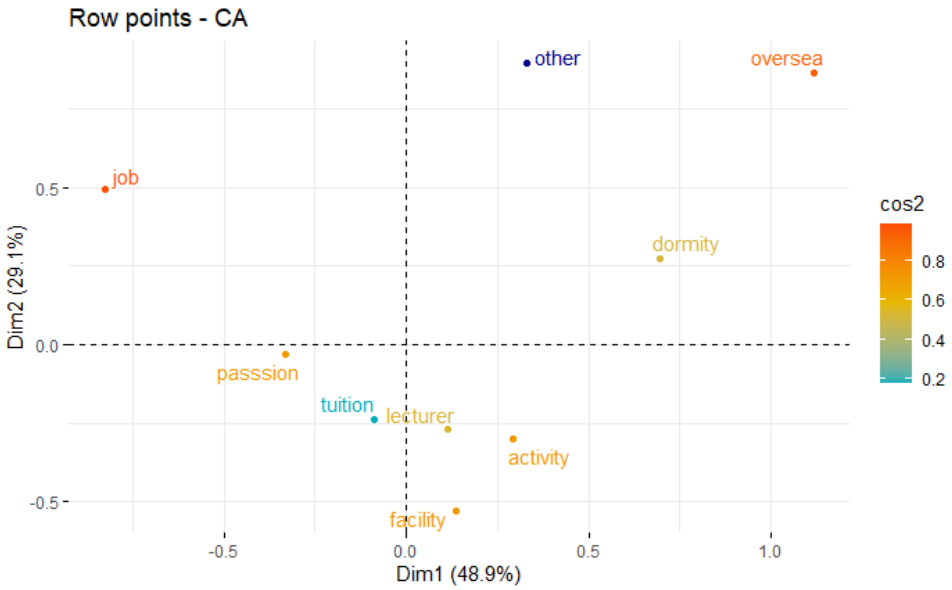

Figure 17.The cos2 plot for the row variable.

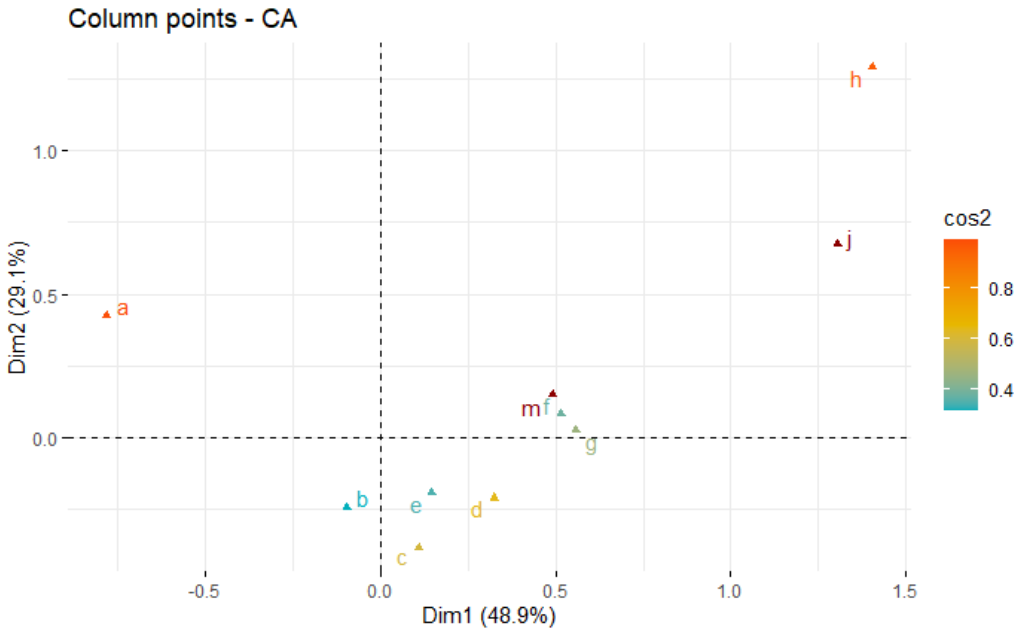

Figure 18. The cos2 plot for the column variable.

The Figure 17and 18 show the value of cos2 for the row and column variable respectively. These plots show the high quality of the CA-biplot for the answers *job, oversea passion, activity, facility* and the values *a, h, d, c.*

V. CONCLUSION

The article reveals the relationship between the reasons of choosing the university and their order of importance which are assigned to each reason by the accepted applicants. From this fact, we could see that the future career, the passion, the tuition fee and living-cost, the modern and innovative facilities, the interesting outclass activities, the excellent and conscientious lecturers, the comfortable dormitory and the chance of working oversea in a developed countries are the reasons in the decreasing order of the importance for the students to choose the university as their favorite trademark of a high level education. In fact, the first three

values *a, b, c* make a dominant meaning comparing to others. Therefore, the corresponding three first reasons: the future career, the passion and the low tuition fee and living-cost take their dominant part overall.

VI. ACKNOWLEDGEMENTS

This article is funded by Thai Nguyen University of Technology for thescientific project code T2022-ĐH01 of the university's contractsigned in 2022: "Application of Machine Learning in analyzing the role of the features of the media in the admission task at Thai Nguyen University of Technology".

REFERENCES

Books:

[1] J. E. Floyd, *Statistics for Economists: A Beginning, A manuscript of Lecture notes*, University of Toronto.

[2] Allan Bluman, *Elementary Statistics: A Step By Step Approach*, 10th Edition, Mc Craw Hill Education.

[3] Douglas C. Montgomery, George C. Runger, *Applied Statistics and Probability for Engineers*, John Wiley & Sons (2014).

[4] Ronald E. Walpole, Raymond H. Meyers, Sharon L. Meyers, Keying Ye (2012), *Probability & Statistics for Engineers & Scientists*, 9th edition, Prentice Hall, Person.

[5] Sheldon M. Ross (2010), *Introduction to Probability Models*, 10th edition, Elsevier.

[6] Abdi, Hervé, and Lynne J. Williams. 2010. "Principal Component Analysis." *John Wiley and Sons, Inc. WIREs Comp Stat* 2: 433–59. http://staff.ustc.edu.cn/~zwp/teach/MVA/abdi-awPCA2010.pdf.

[7] Bendixen, Mike. 2003. "A Practical Guide to the Use of Correspondence Analysis in Marketing Research." *Marketing Bulletin* 14. http://marketing-bulletin.massey.ac.nz/V14/MB_V14_T2_Bendixen.pdf.