Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 1, 2024, 26-34
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

Factor Analysis on Possum Dataset to Simplify Many Independent Variables Into Fewer Factors.

Ayu Larasati[1*], Suminto[2]

[1,2] Universitas Pattimura, Indonesia
[*]corresponding author: ayul@gmail.com

Received April 17, 2024; Received in revised form April 27, 2024; Accepted April 27, 2024

Abstract. *This research aims to apply factor analysis to possum data with the aim of simplifying many independent variables into fewer factors. The factor analysis steps begin by grouping the variables to be analyzed and compiling a correlation matrix using the Bartlett test and the Kaiser-Meyer-Olkin (KMO) test. From the test results, it was found that the variables had sufficient correlation to proceed to factor analysis. After that, factor extraction was carried out using three criteria, namely eigenvalues, diversity percentage, and scree plot, which concluded that the number of factors formed was two. Next, factor rotation was carried out using the varimax method to simplify interpretation. The results show that certain variables have high loadings on certain factors, making it easier to identify patterns. In conclusion, factor analysis succeeded in simplifying the relationship between variables into two factors that can be interpreted more easily.*
Keywords*: factor analysis; possum dataset; simplifying factors*

INTRODUCTION

In this day and age where data analysis is increasingly becoming the primary basis for decision making in various fields, it is important to understand the techniques that can help perfect the complexity of large datasets (Sarker, 2021; Van Den Heuvel et al., 2020). One of the important techniques in statistics is factor analysis, which aims to reduce many independent variables into fewer but informative factors (Shrestha, 2021). In this context, research on the application of factor analysis to possum datasets has significant urgency. Possums, as one of the important animals in Australian ecology, have attracted the attention of researchers in understanding the various factors that influence their survival and population (Moseby et al., 2021). The possum dataset provides rich information on a variety of physical, environmental, and demographic attributes related to ringtail and bushtail possums (Neilly et al., 2022; Byrne, 2022).). However, the complexity of this dataset requires a careful analysis approach to reveal the underlying patterns of relationships between variables. Previous studies have used factor analysis in various contexts to relate complex datasets. However, the application of factor analysis to possum datasets is still relatively limited. Several studies have explored the relationship between environmental and demographic factors and possum populations (Patterson, 2021; Moore, 2023), but have not fully exploited the potential of factor analysis to understand these relationships in more depth.

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 1, 2024, 26-34
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

Although there has been some research attempting to identify the factors that influence possum populations, there are still gaps in our understanding of how these variables interact and how they can interact into more easily interpretable factors. By applying factor analysis to the possum dataset, this research seeks to fill this gap by providing new insights into the complex relationships between existing attributes. This research aims to apply factor analysis to the possum dataset with the aim of reducing many independent variables into fewer factors and representing variations in the data efficiently. The aim of this research is to identify the main factors influencing possum populations and understand the relationships that may exist between them.

METHOD

The data used in this research is possum data. Possums are a species of marsupial native to Australia, Papua New Guinea and Sulawesi (Wikipedia, 2022). From nine morphometric measurements on each of 104 mountain brushtail possums. This species is trapped in seven locations, namely Australia from South Victoria to Central Queensland. Several variables used to identify possum fishing areas are as follows:

Table 1. Independent variables of possum data

Indicator	**Description**
Age	Age
Head Length	Head Length
Skull Width	Skull Width
Total Length	Total Length
Taill Length	Taill Length
Foot Length	Foot Length
Ear Conch Length	Ear Conch Length
Eye	Distance from medial canthus to lateral canthus of right eye
Chest	Chest circumference (in cm)
Belly	Abdominal circumference (in cm)

Tabel 2. Summary data Possum

Summary/ Variabel	Age	Hdlgth	Skullw	Totlgth	Taill	Footlgth	Earconch	Eye	Chest	Belly
Min	1	82,50	50,00	75,00	32,00	60,30	40,30	12,80	22,0	25,0
Median	3	92,80	56,35	88,00	37,00	68,10	46,80	14,90	27,0	32,5
Max	9	103,10	68,60	96,50	43,00	77,90	56,20	17,80	32,0	40,0
Mean	3,8	92,6	56,88	87,09	37,01	68,49	48,13	15,05	27,0	32,5

Based on table 2, the average lifespan of this possum species is 3.8 years. When a possum is one year old, it will have a head length of 82.50 cm, while when it reaches a maximum age of 9 years the head length will be 103.10 cm. In contrast to the length of the possum's tail, the average is 37.01 cm. There is not much difference between 1 year old and 9 year old. Then if we look at the length of the legs, the average is 68.49 cm. Apart from that, we can also see based on the area where the possum is trapped along with its gender, which is visualized in the following image:

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 1, 2024, 26-34
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

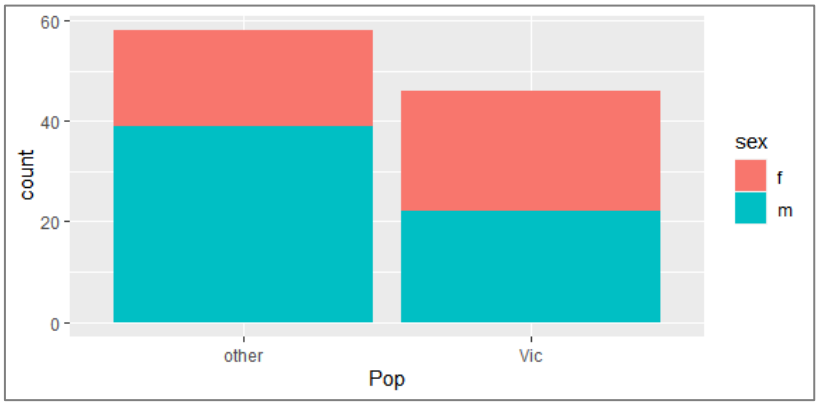

Figure 1. Number of Possums in certain areas by gender

In the Victoria area, 23 male and 24 female possums were detected. Meanwhile, apart from the Victoria location (6 other locations), the number of male possums was 37 and 19 females. This means that the number of possums in Victoria tends to be greater, compared to being spread across the other 6 locations.

RESULTS AND DISCUSSION

In this research, the correlation matrix examination was carried out in 2 ways, namely:

a. Bartlett Test (Bartlett Test of Sphericity)
   In the Bartlett's test, it can be determined whether these variables have a correlation or not by looking at the significance level values obtained (Alabdulkarim, 2022). If the sig value is below alpha 0.01, then the correlation between variables is low so factor analysis cannot continue. On the other hand, if the p-value is <0.01, then the correlation between variables is high and the factor analysis process can be continued (Zhang et al., 2022). The following is a visualization of the form of the correlation matrix formed:

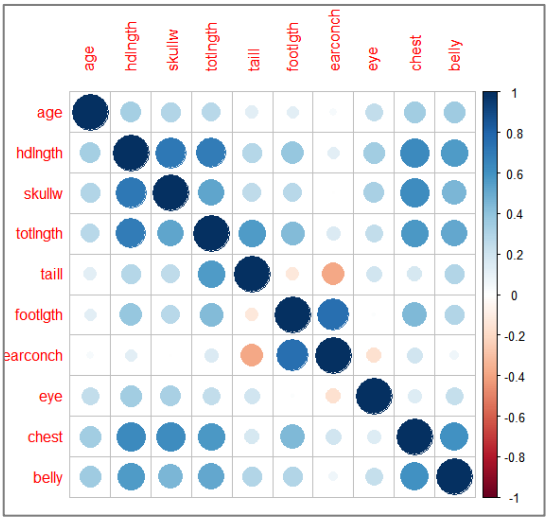

Figure 2. Corrplot of correlation between variables

Based on Figure 2, the variables have a good correlation, they can be identified from the color and size of the circle. If it goes towards dark blue, then the variables have a higher positive correlation, whereas conversely, if it is dark red, it indicates a higher negative correlation. Several variables have a fairly high positive correlation, namely between

Volume 2, No. 1, 2024, 26-34
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

head length and skull length, head length and total length, then between ear conch length and tail length, and several other variables. On average, all variables are correlated with each other, although there are some that are less correlated. Based on this, there is a possibility that these variables are in the same factor.

From the results obtained from the Bartlett's test, the chi-square value is 513.9327 with degrees of freedom of 47 and a p-value of 0.000. This value is less than 0.01 so it can be said that the variables used are correlated with each other and can be continued to the next process.

a. Uji KMO (*Kaiser-Meyer-Olkin*)
The test used to determine the feasibility of a factor analysis is on a scale between 0 and 1. If the calculated KMO value is <0.5 then factor analysis is not suitable for use. On the other hand, if the calculated KMO value is > 0.5 then factor analysis is feasible (Koçak et al., 2022). Meanwhile, based on Kaiser (1970), the KMO assessment is as follows
.

Table 3. KMO assessment

KMO Value Range	Assessment Category
0,9 ≤ KMO ≤ 1,0	Excellent data (marvelous) for factor analysis
0,8 ≤ KMO ≤ 0,9	Good (meritorious) data for factor analysis
0,7 ≤ KMO ≤ 0,8	The data is sufficient (middling) for factor analysis
0,6 ≤ KMO ≤ 0,7	Data is lacking (mediocre) for factor analysis
0,5 ≤ KMO ≤ 0,6	Bad (miserable) data for factor analysis
KMO ≤ 0,5	Data is unacceptable (unacceptable) for factor analysis

Based on the results obtained, the overall KMO value is greater than 0.5, namely 0.77. This means that the KMO test is fulfilled, so the data is good enough and factor analysis is suitable for use in this research. Next, to assess the suitability of each variable for factor analysis, the Measure of Sampling Adequacy (MSA) criterion was used. Another measure used to measure intercorrelation between variables and the suitability of factor analysis is MSA (Hair et al., 1998). According to Sharka et al. (2022) if the MSA value ≥ 0.5 indicates that the variable can still be predicted and analyzed further. In this study, the MSA value for each variable was more than 0.5 so the analysis could be continued. The following is a table of the MSA values for each variable.

Table 4. MSA value for each variable

Variable	MSA Values
Age	0,89
Head Length	0,84
Skull Width	0,83
Total Length	0,76
Tail Length	0,58
Foot Length	0,69
Ear Conch Length	0,55
Eye	0,80
Chest	0,85
Belly	0,88

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 1, 2024, 26-34
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

1. Determine the number of factors to be extracted
   The process of reducing a number of variables into a new set of variables so as to produce a smaller number of factors is called factor extraction (Taherdoost et al., 2004). There are several criteria for determining the number of factors that are formed, namely the characteristic root criterion, the diversity percentage criterion, and the scree plot criterion (Hair et al., 1998). In this research, these three criteria will be applied to determine the number of factors used.

   a. Characteristic root criteria (eigenvalues)
      One of the most frequently used techniques is looking at eigenvalues. Each variable has a contribution value of one to the total eigenvalue, this is the reason for using eigenvalues. For factors with an eigenvalue ≥ 1, they are considered significant, while for factors with an eigenvalue < 1, they are declared not significant and must be removed from the model.

Table 5. Eigenvalues and variance

Factor	**Eigen Values**		
	Total	**Variance (%)**	***Cumulative* (%)**
1	4,120	41,20	41,20
2	1,959	19,59	60,79
3	0,967	9,67	70,46
4	0,790	7,90	78,36
5	0,711	7,11	85,47
6	0,530	5,30	90,77
7	0,321	3,21	93,98
8	0,267	2,67	96,65
9	0,170	1,70	98,35
10	0,161	1,61	100,00

From the table above, factors 1 and 2 have an eigenvalue of more than 1. So with this criterion the number of factors used is 2 factors.

b. Diversity percentage criteria
   Determining the number of factors is seen from the specific value of the cumulative percentage of diversity that can be explained by the factors formed. This research will stop factor extraction if it reaches a total diversity of ≥ 60%. So based on table 3, the cumulative variance value between factor 1 and factor 2 is 60.79%, meaning that the 2 factors formed are able to represent 10 variables amounting to 60.79% of the possum data. Thus, the extraction of the 2 factors obtained can be stopped because they meet the percentage of diversity criteria.

c. Scree Plot Criteria
   A scree plot is a plot of eigenvalues against the number of factors extracted. The point at which the scree begins to occur indicates the exact number of factors. This point occurs when the scree begins to appear flat. Figure 2 shows that there are only 2 factors formed.

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 1, 2024, 26-34
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

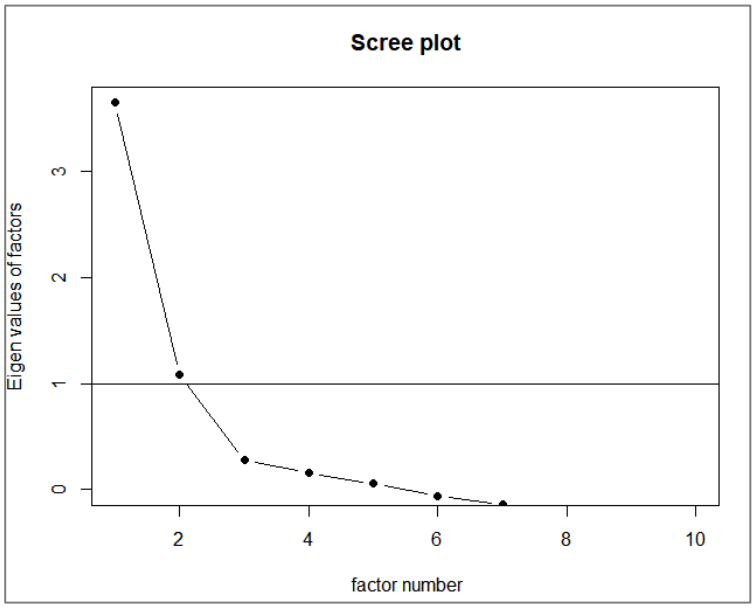

Figure 3. *Scree Plot*

Based on the combination of the three criteria for determining the number of factors above, it can be concluded that the most appropriate factor extraction is 2 factors.

1. Factor Rotation
   After obtaining the number of factors, a clearer relationship between individual factors and various variables will be sought. The factors obtained are generally difficult to interpret directly. Therefore, manipulation was carried out by rotating the loading L using the perpendicular rotation method or commonly known as Ortogonal Varimax. Some experts state that varimax rotation is closer to reality than others. The concept of varimax rotation is to maximize the weighting factors which produce correlations of variables with one factor that are close to one, and correlations with other factors that are close to zero, making it easy to interpret. The following is a table of comparison results of factor loadings before and after rotation using varimax rotation.

Table 6. Comparison of Factor Loadings Before and After Rotation

Variable	Before Rotation		After Rotation (Varimax)	
	Factor 1	Factor 2	Factor 1	Factor 2
Age	0,395		0,398	
Head Length	0,842	0,169	0,849	0,125
Skull Width	0,764		0,766	
Total Length	0,774	0,198	0,783	0,158
Tail Length	0,485	-0,362	0,465	-0,387
Foot Length	0,322	0,785	0,362	0,767
Ear Conch Length		0,986		0,988
Eye	0,385	-0,146	0,377	-0,166
Chest	0,732	0,246	0,743	0,208
Belly	0,668	0,105	0,673	

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 1, 2024, 26-34
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

After factor extraction, only two factor axes need to be shifted. Logically it can be said that rotating two axes is much easier than rotating eight axes at once, provided that the eight axes remain perpendicular to each other. After rotation, it is clear that now the distance between the variables age, head length, skull width, total length, foot length, chest and belly to the axis of factor 1 (rotated factor 1) is smaller than the distance to the axis of factor 2 (rotated factor 2). The term in factor analysis for this situation is that these variables have a high loading on factor 1, and the loading on factor 2 is close to 0. The opposite also happens for the tail length and eye variables. The image below shows loading before and after rotation.

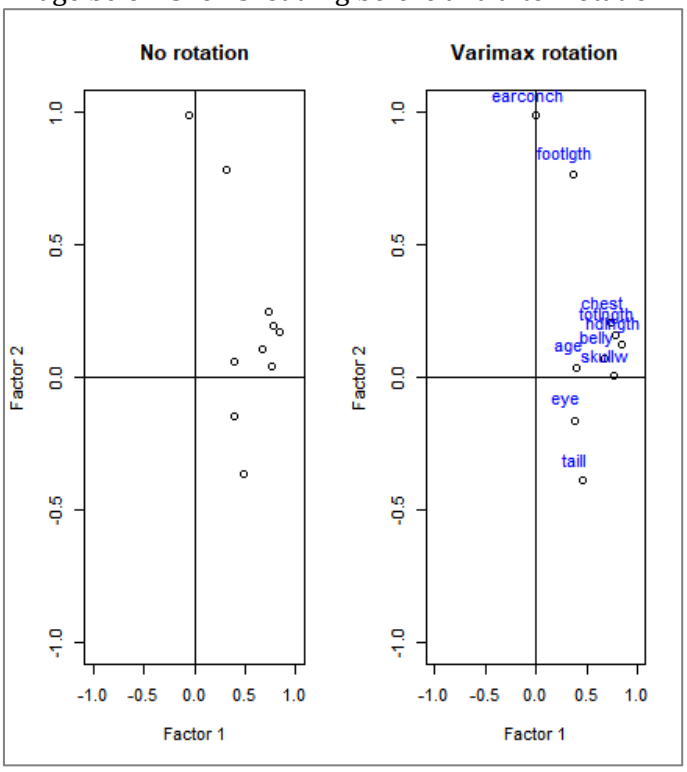

Figure 4. Plot after and before factor rotation

2. Factor Loading Matrix Diagram

Because the aim of this research is to simplify the form of relationship between several variables into a smaller number of factors in the variables studied, the following is a diagram of grouping variables into each factor.

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 1, 2024, 26-34
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

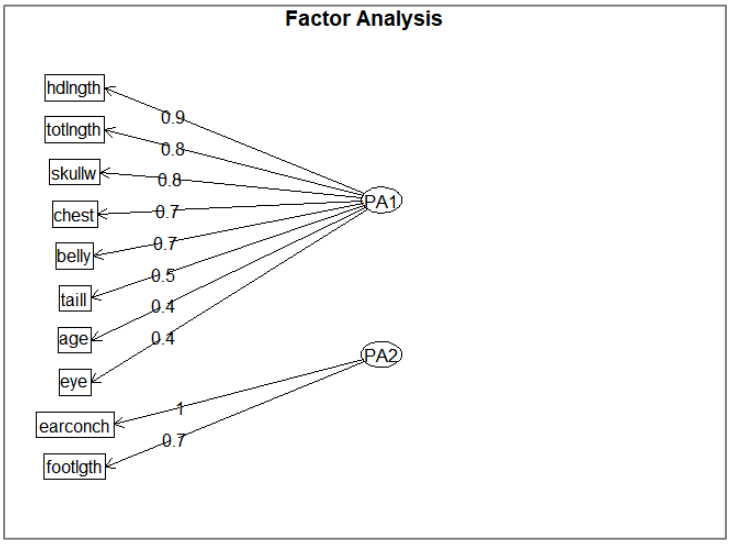

Figure 5. Plot after and before factor rotation

CONCLUSION

In accordance with the analysis above, after testing the KMO and Bartlett assumptions, it shows that the assumptions are met so that factor analysis is suitable for use in this research. Based on the aim of this research, namely simplifying the relationship of several variables into a smaller number of factors, the number of factors formed has been determined using the criteria of eigenvalues, proportion of diversity, and scree plot. As a result of the combination of these three variables, the number of factors formed is two factors. After that, factor rotation was carried out to make interpretation easier, namely using varimax rotation. Varimax rotation is one of the rotations that is quite popular, because it is closer to reality than others. After rotation, it is clear that now the distance between the variables age, head length, skull width, total length, foot length, chest and belly to the axis of factor 1 (rotated factor 1) is smaller than the distance to the axis of factor 2 (rotated factor 2). The term in factor analysis for this situation is that these variables have a high loading on factor 1, and the loading on factor 2 is close to 0. The opposite also happens for the tail length and eye variables. Then, if you look at it based on the factor loading diagram, it explains that factor 1 consists of the variables head length, total length, skull width, chest, belly, tail, age and eyes. Meanwhile, factor 2 consists of ear conch and foot length variables with each loading factor.

REFERENCE

Alabdulkarim, L. (2022). Development and validation of an Arabic pediatric sensorimotor development test. *International Journal of Pediatrics and Adolescent Medicine*, *9*(1), 36–40. https://doi.org/https://doi.org/10.1016/j.ijpam.2021.03.005

Byrne, R. W. (2022). Machiavellian Intelligence. In *Encyclopedia of Animal Cognition and Behavior* (pp. 4033-4038). Cham: Springer International Publishing.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed). NJ: Prentice-Hall.

Kaiser, H. F. (1970). *A second generation little jiffy.*

Neilly, H., McKenzie, T., Ward, M., Chaber, A., & Cale, P. (2022). Potential drivers of common brushtail possum (Trichosurus vulpecula) decline on a Murray River floodplain. *Australian Mammalogy*, *45*(1), 62-70.

Moseby, K., Hodgens, P., Bannister, H., Mooney, P., Brandle, R., Lynch, C., ... & Jensen, M. (2021). The ecological costs and benefits of a feral cat poison-baiting programme for

Volume 2, No. 1, 2024, 26-34
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

protection of reintroduced populations of the western quoll and brushtail possum. *Austral Ecology*, *46*(8), 1366-1382.

Koçak, C., Sandal, S., Çöl, M., Tanca, A. K., Kuloğlu, Z., & Kırsaçlıoğlu, C. T. (2022). Turkish Validity-Reliability Study of the Celiac Disease-Specific Pediatric Quality of Life Scale. *The Turkish Journal of Gastroenterology*, *33*(3), 248.

Moore, L. J., Petrovan, S. O., Bates, A. J., Hicks, H. L., Baker, P. J., Perkins, S. E., & Yarnell, R. W. (2023). Demographic effects of road mortality on mammalian populations: a systematic review. *Biological Reviews*, *98*(4), 1033-1050.

Patterson, C. R., Seddon, P. J., Wilson, D. J., & van Heezik, Y. (2021). Habitat-specific densities of urban brushtail possums. *New Zealand Journal of Ecology*, *45*(2), 1-9.

Sarker, I. H. (2021). Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. *SN Computer Science*, *2*(5), 377.

Sharka, R., San Diego, J., Nasseripour, M., & Banerjee, A. (2022). Faktor analysis of risk perceptions of using digital and social media in dental education and profession. *Journal of Dental Education*.

Shrestha, N. (2021). Factor analysis as a tool for survey analysis. *American Journal of Applied Mathematics and Statistics*, *9*(1), 4-11.

Taherdoost, H., Sahibuddin, S., & Jalaliyoon, N. (2004). Exploratory Faktor Analysis ; Concepts and Theory 2 Faktor Analysis 3 Types of Faktor Analysis 4 Exploratory Faktor Analyses. *Advances in Applied and Pure Mathematics*, 375–382.

Zhang, W., Jin, Y., Liu, N., Xiang, Z., Wang, X., Xu, P., Guo, P., Mao, M., & Feng, S. (2022). Predicting Physical Activity in Chinese Pregnant Women Using Multi-Theory Model: A Cross-Sectional Study. In *International Journal of Environmental Research and Public Health* (Vol. 19, Issue 20). https://doi.org/10.3390/ijerph192013383

Van Den Heuvel, L., Dorsey, R. R., Prainsack, B., Post, B., Stiggelbout, A. M., Meinders, M. J., & Bloem, B. R. (2020). Quadruple decision making for Parkinson's disease patients: combining expert opinion, patient preferences, scientific evidence, and big data approaches to reach precision medicine. *Journal of Parkinson's disease*, *10*(1), 223-231.