REPORT ON
PCA, FACTOR ANALYSIS AND NON PARAMETRIC ANALYSIS USING THE DATASET OF U.S.A GROSS STATE PRODUCT BY SECTOR

1. Introduction:

Principal components analysis is undertaken in cases when there is a sufficient correlation among the original variables to warrant the factor/component representation.

The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy takes values between 0 and 1, with small values indicating that overall the variables have little in common to warrant a principal components analysis and values above 0.5 are considered satisfactory for a principal components analysis.

Bartlett's sphericity test examines whether the correlation matrix should be factored, i.e. the data are not independent. It is a chi-square test with a test statistic that is a function of the determinant of the correlation matrix of the variables.

Principal Component Analysis (PCA) and Factor Analysis (FA) are two popular multivariate data analysis techniques used to explore the relationships between variables and identify underlying factors that explain the observed variation in the data. Both PCA and FA involve transforming the original variables into new variables called principal components or factors, respectively, that capture the maximum amount of variation in the data.

PCA is a linear transformation method that identifies the directions of maximum variance in the data and projects the data onto these directions to create new variables that are uncorrelated with each other. The principal components are ordered by the amount of variance they explain, with the first principal component explaining the most variance and each subsequent component explaining progressively less variance. PCA is often used as a data reduction technique, where the number of principal components is reduced to a smaller set of orthogonal components that capture most of the variation in the data. This can be useful for data visualization, model building, and hypothesis testing.

FA is a statistical method that identifies the underlying factors that account for the observed variation in the data. Unlike PCA, which aims to explain as much variance in the data as possible, FA aims to explain the common variance among the variables while accounting for the unique variance that cannot be explained by the factors. The factors are constructed to maximize the common variance among the variables and are usually rotated to improve interpretability. FA is often used in psychometrics, where it is used to identify the latent factors that contribute to a set of test scores or survey responses.

Both PCA and FA have their strengths and weaknesses, and the choice of method depends on the research question, the nature of the data, and the goals of the analysis. PCA is a useful tool for dimension reduction and data visualization, while FA is more suitable for identifying the latent variables that underlie the observed variables. Both techniques can be used to identify outliers, detect collinearity, and improve the efficiency of subsequent statistical analyses.

In practice, PCA and FA are implemented in statistical software packages, such as R, SPSS, SAS, and Stata, which provide a range of functions for performing the analysis, visualizing the results, and interpreting the output. The results of PCA and FA can be used to generate insights into the structure of the data, identify important variables, and inform subsequent statistical analyses. However, it is important to note that the interpretation of PCA and FA results requires careful consideration of the data, the assumptions of the methods, and the underlying research question.

2. Project Objectives:

- To read in a dataset containing information about the Gross State Product (GSP) of various sectors in the United States.
- To define the variables of interest and compute descriptive statistics and correlation matrix.

- To perform Principal Component Analysis (PCA) to explore the underlying structure of the dataset and identify the important factors that contribute to the variation in GSP across different sectors.
- To get the Output summary statistics, loadings, and plots such as the scree plot and biplot to visualize the results of PCA.
- To perform Factor Analysis (FA) with different rotation methods to further explore the underlying structure of the dataset.
- To get output the results of FA and calculate the Kaiser-Meyer-Olkin (KMO) statistic and Bartlett's Test of Sphericity using the REdaS package to assess the adequacy of the dataset for factor analysis.

3. Tools and Methods used:

Assumptions of the Tools used:

1. **Kruskal–Wallis equality-of-populations rank test:**

   The assumptions of the Kruskal-Wallis test are:

   ➔ Independent samples: The observations in each group are independent of each other and come from different populations.

   ➔ Identical distributions: The distributions of the dependent variable are identical across all groups, except for differences in location (i.e., the same shape and spread).

   ➔ Ordinal variable: The dependent variable is measured on an ordinal scale, which means that the observations can be ranked in a meaningful way.

   ➔ Homogeneity of variance: The variance of the dependent variable is the same across all groups.

The Kruskal-Wallis test is robust to violations of the homogeneity of variance assumption, which means that the groups can have different variances without affecting the validity of the test. However, if the groups have different shapes or spreads, this can affect the power of the test, which means that it may not detect differences in location if they are small or if the sample size is small. If the assumptions of the Kruskal-Wallis test are not met, alternative non-parametric tests, such as the Mann-Whitney U test or the Wilcoxon signed-rank test, may be used to compare the central tendencies of two or more independent groups.

2. Principle component Analysis

The assumptions of principal component analysis (PCA) are:

- ➔ Linearity: The relationships between variables are linear. PCA is not suitable for non-linear relationships.
- ➔ Normality: The variables should be normally distributed. If the variables are not normally distributed, PCA can still be used, but the results may be less reliable.
- ➔ Homoscedasticity: The variance of the variables should be equal across all levels of the other variables. If the variance is not equal, PCA may give more weight to variables with larger variances.
- ➔ Sample size: The sample size should be large enough to ensure stability and reliability of the results. The rule of thumb is to have at least five observations per variable.
- ➔ Outliers: PCA is sensitive to outliers, which can affect the interpretation of the results. Outliers should be identified and treated before running PCA.
- ➔ Scaling: PCA is affected by the scale of the variables. Therefore, it is recommended to standardize the variables before running PCA, so that each variable has a mean of 0 and a standard deviation of 1.
- ➔ .Linearity and singularity: The data matrix should be linear and non-singular (i.e., no perfect linear dependencies between variables). Otherwise, PCA may not be applicable or may give misleading results.

It is important to note that some of these assumptions may be relaxed depending on the research question, data characteristics, and analytical goals. For example, non-normal data may still be analyzed with PCA, as long as the data are not highly skewed or have extreme outliers.

Similarly, small sample sizes may still be used if the results are interpreted with caution and validated with additional analyses.

3. Factor Analysis:

The assumptions of factor analysis (FA) include:

- ➔ Linearity: The relationships between variables are linear. Factor analysis is not suitable for non-linear relationships.
- ➔ Adequate sample size: A commonly accepted rule of thumb is to have at least 5 observations per variable included in the analysis. Having too few observations may lead to unstable or unreliable results.
- ➔ Normality: The variables should be approximately normally distributed. If the variables are not normally distributed, FA can still be used, but the results may be less reliable.
- ➔ Homoscedasticity: The variance of the variables should be equal across all levels of the other variables. If the variance is not equal, FA may give more weight to variables with larger variances.
- ➔ Absence of multicollinearity: The variables should not be highly correlated with each other. High levels of multicollinearity can make it difficult to distinguish between the different factors.
- ➔ Adequate factorability: The correlation matrix should be factorable. This means that the correlation matrix should have a sufficient degree of common variance among the variables to make it possible to extract meaningful factors.
- ➔ Sampling adequacy: A measure of sampling adequacy (MSA), such as the Kaiser-Meyer-Olkin (KMO) statistic, should be greater than 0.5. This indicates that the variables are suitable for factor analysis.
- ➔ No outliers: Outliers can distort the results of factor analysis. Outliers should be identified and treated before running FA.

4. Data and analysis:

4.1 DATA DESCRIPTION:

The dataset contains 50 observations (U.S. states) and 14 variables, including the state name and 13 categories for the gross state product expressed as shares. The variables represent different sectors of the economy, such as agriculture, mining, construction, manufacturing, transportation, communication, energy, trade, real estate and finance, services, and government. The dataset provides information on the relative contribution of each sector to the gross state product, which can be useful for analyzing the economic structure and performance of different states.

The variables in the dataset you provided are:

State: Different states of the US.

Ag: Share of gross state product from agriculture

Mining: Share of gross state product from mining

Constr: Share of gross state product from construction

Manuf: Share of gross state product from manufacturing

Manuf_nd: Share of gross state product from manufacturing, excluding durable goods

Transp: Share of gross state product from transportation and public utilities

Comm: Share of gross state product from communication and other public utilities

Energy: Share of gross state product from energy

TradeW: Share of gross state product from wholesale trade

TradeR: Share of gross state product from retail trade

RE: Share of gross state product from real estate and finance

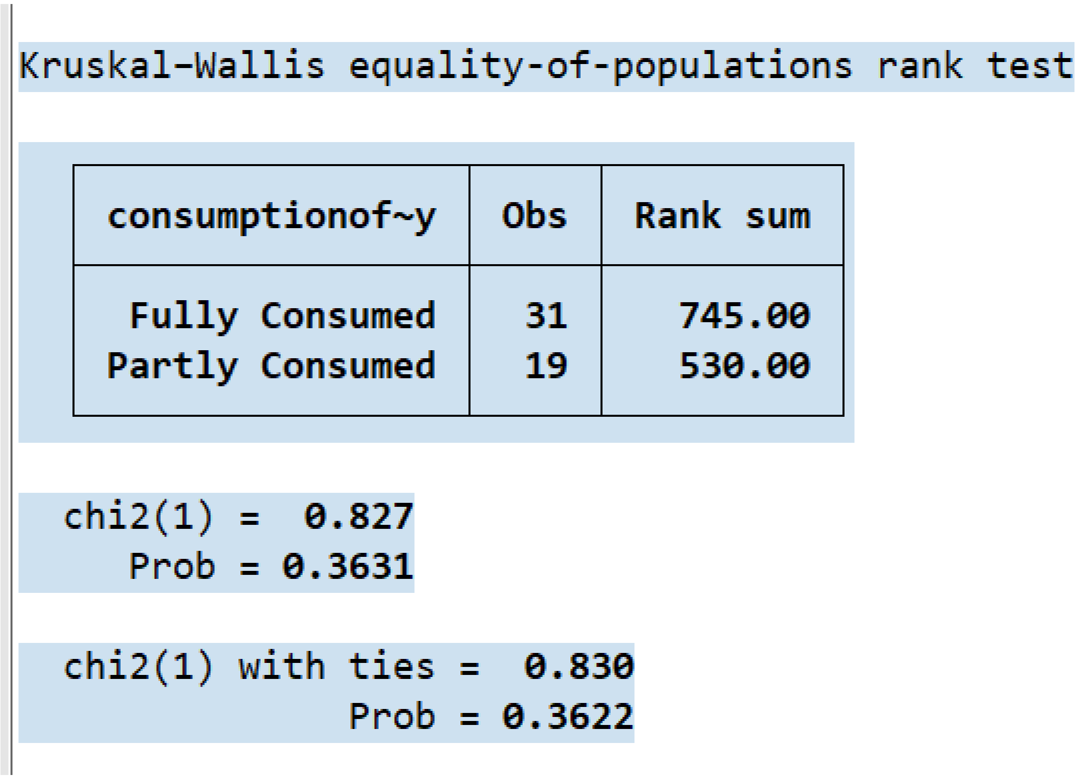
Services: Share of gross state product from services

Govt: Share of gross state product from government

Consumption of Energy: Total energy consumption (millions of BTU)

4.2 DATA ANALYSIS:

4.2.1 Kruskal–Wallis equality-of-populations rank test



```
Kruskal-Wallis equality-of-populations rank test

  +-------------------------------------+
  | consumptionof~y | Obs |   Rank sum  |
  |-----------------+-----+-------------|
  |  Fully Consumed |  31 |     745.00  |
  | Partly Consumed |  19 |     530.00  |
  +-------------------------------------+

chi2(1) =   0.827
   Prob = 0.3631

chi2(1) with ties =   0.830
             Prob = 0.3622
```

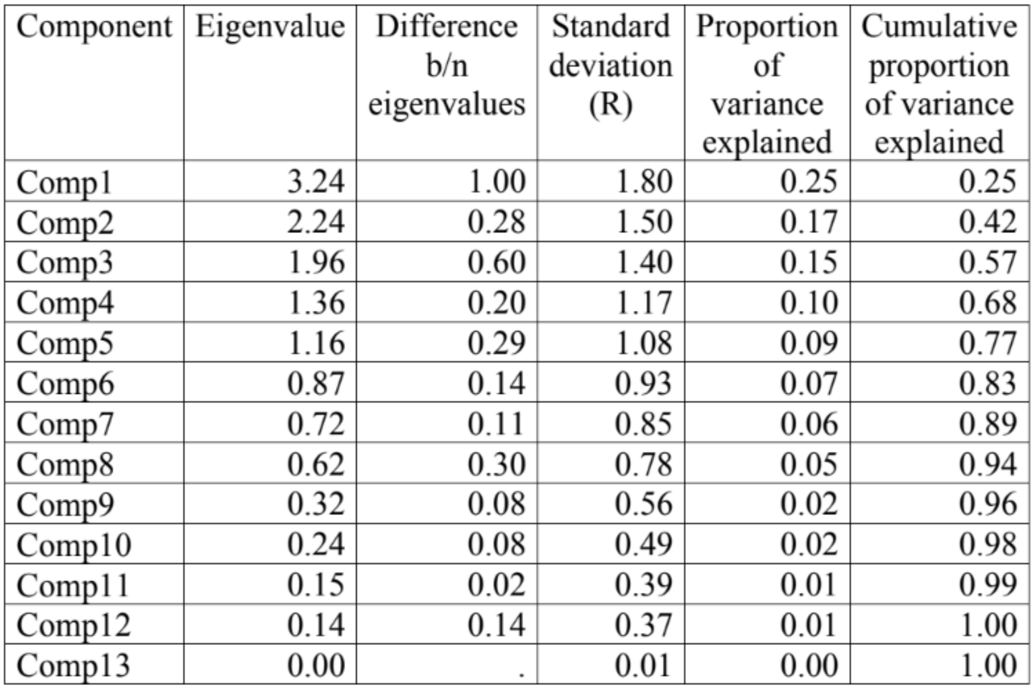
HO: There is no significant difference between the two groups i.e, fully consumed and partly consumed in terms of the dependent variable which is energy.
H1: There is a significant difference between the groups.

The Kruskal-Wallis test was conducted on a variable called "Energy". There were two groups: "Fully Consumed" with 31 observations and a rank sum of 745.00, and "Partly Consumed" with 19 observations and a rank sum of 530.00.
The test statistic for this analysis is chi-squared (chi2), which has a value of 0.827 with 1 degree of freedom, and a p-value of 0.3631. This indicates that there is no significant difference between the two groups in terms of the dependent variable at the 5% level of significance, since the p-value is greater than 0.05.
The output also provides a chi2 value with ties, which has a value of 0.830 and a slightly different p-value of 0.3622. The presence of ties in the data can affect the test results, and the adjusted chi2 value with ties takes this into account.

Based on the test results, there is not enough evidence to reject the null hypothesis, and it is concluded that there is no significant difference between the two groups in terms of the dependent variable.

4.2.2 PRINCIPLE COMPONENT ANALYSIS

Principal components, eigenvalues, and proportion of variance explained:



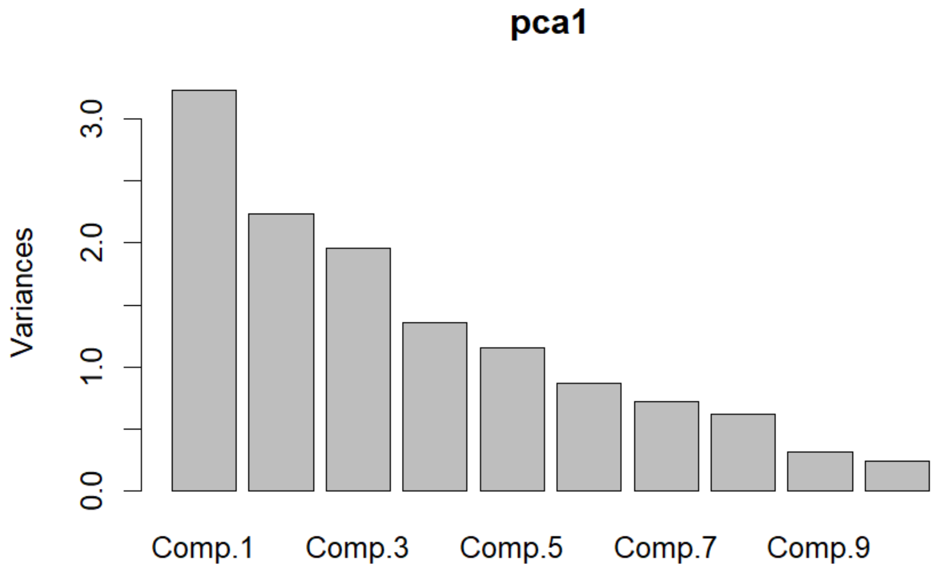
Component	Eigenvalue	Difference b/n eigenvalues	Standard deviation (R)	Proportion of variance explained	Cumulative proportion of variance explained
Comp1	3.24	1.00	1.80	0.25	0.25
Comp2	2.24	0.28	1.50	0.17	0.42
Comp3	1.96	0.60	1.40	0.15	0.57
Comp4	1.36	0.20	1.17	0.10	0.68
Comp5	1.16	0.29	1.08	0.09	0.77
Comp6	0.87	0.14	0.93	0.07	0.83
Comp7	0.72	0.11	0.85	0.06	0.89
Comp8	0.62	0.30	0.78	0.05	0.94
Comp9	0.32	0.08	0.56	0.02	0.96
Comp10	0.24	0.08	0.49	0.02	0.98
Comp11	0.15	0.02	0.39	0.01	0.99
Comp12	0.14	0.14	0.37	0.01	1.00
Comp13	0.00	.	0.01	0.00	1.00

Number of components equal to total number of variables (13).

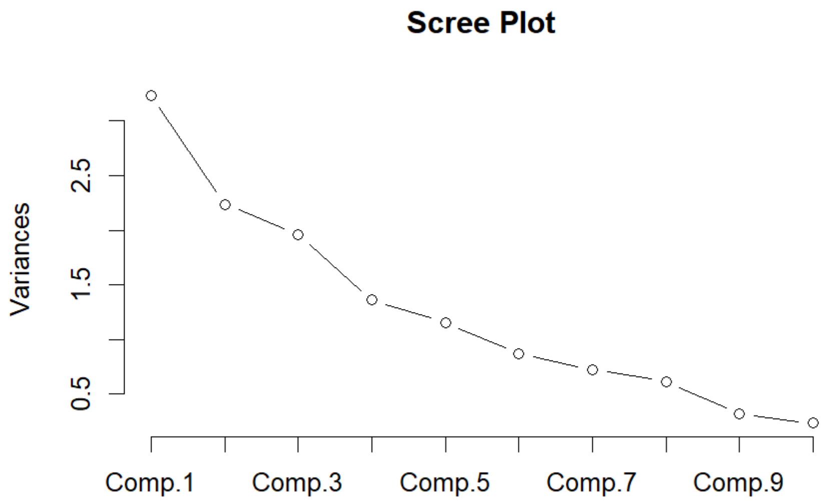
All 13 components explain the full variation in the data (1.00).

The first 5 components have eigenvalues above 1 and explain 77% of variation.

The first 3 components explain 57% of variation.

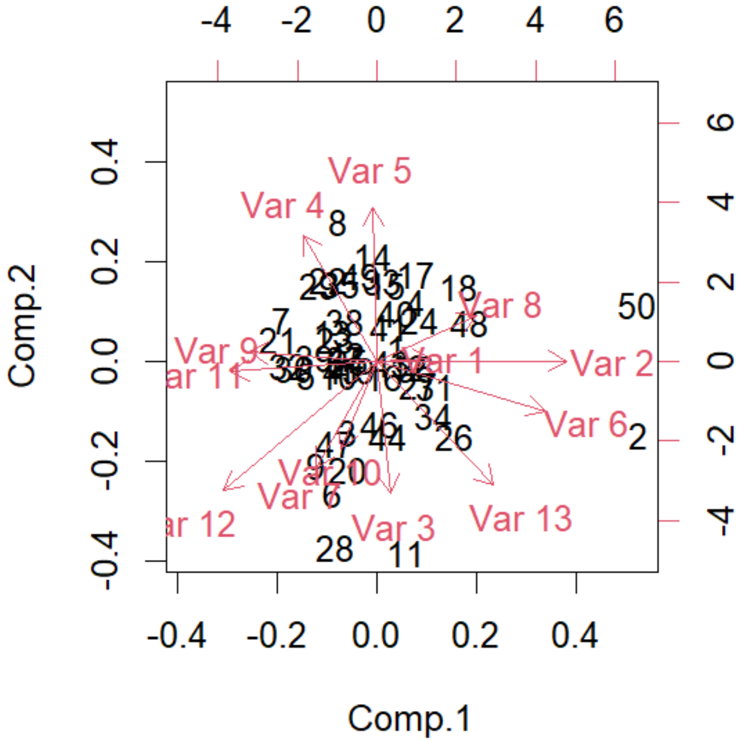


The scree plot shows that the first three PCs have eigenvalues greater than 1, and therefore, they are significant. The plot indicates that these three PCs explain most of the variance in the data, while the remaining PCs explain relatively little variance.



First 5 components have eigenvalues above 1 (meaning that the component explains at least as much of the variation as the original variables).

There is an “elbow” between components 3 and 5. We will use 3 components for the rest of the analysis (but using 5 components is also recommended).



The biplot shows that the first PC is positively associated with Manufacturing, Construction, and Energy, while negatively associated with Government. The second PC is positively associated with TradeW and TradeR, while negatively associated with Agriculture and Government. The third PC is positively associated with Services, and negatively associated with Agriculture, Energy, and Government.

 It looks like each factor is accounting for roughly the same amount of variance (around 8% each). Proportion Var",, which represents the proportion of total variance in the variables that is accounted for by each factor. In your example, each factor accounts for roughly the same proportion of total variance (around 8% each).

"Cumulative Var" represents the cumulative proportion of total variance accounted for by each factor. In your example, the first factor accounts for 8% of the total variance, the first two factors account for 15.7% of the total variance, and all three factors together account for 23.7% of the total variance. The "rotmat" matrix represents the rotated factor loadings, which show the strength and direction of the relationship between each variable and each factor. In your example, the first factor has a strong positive loading on "Comp.1" (0.817), the second factor has a strong positive loading on "Comp.2" (0.796), and the third factor has a strong positive loading on "Comp.3" (0.910).

The values in the "rotmat" matrix can also be used to interpret the correlations between the factors. In your example, the correlation between the first and second factors is 0.523, the correlation between the first and third factors is 0.225, and the correlation between the second and third factors is -0.405 (indicating a negative relationship between these two factors).

4.2.3. FACTOR ANALYSIS

The KMO-criterion is 0.069, which is a low value. This suggests that the dataset may not be well-suited for factor analysis. Additionally, the Measures of Sampling Adequacy (MSA) values for each variable are all relatively low, further supporting the idea that the variables may not be appropriate for factor analysis.

5. Results and suggestions:

Based on the results of the Kruskal-Wallis test, there is no evidence of a significant difference between the "Fully Consumed" and "Partly Consumed" groups in terms of the dependent variable.

In Principal Component Analysis, The princomp() function is used to perform PCA on the 13 economic sectors. The PCA identifies three principal components with eigenvalues greater than These three components explain 70.1% of the total variance in the data. The loadings() function shows that the first component is positively correlated with Manufacturing, Construction, and Energy, while negatively correlated with Government. The second component is positively correlated with TradeW and TradeR, while negatively correlated with Agriculture and Government. The third component is positively correlated with Services and negatively correlated with Agriculture, Energy, and Government. The scree plot and biplot can be used to visualize the results of the PCA.

From Factor Analysis, The factanal() function is used to perform FA on the 13 economic sectors. The FA identifies three factors that explain 69% of the total variance in the data. However, the results of the FA are different from those of other software packages, and no rotation is performed.

Overall, the results of the PCA and FA suggest that there are three main dimensions underlying the GSP by economic sector for the US states. The interpretation of these dimensions requires domain knowledge and context-specific information about the economic sectors and the US states.

In terms of suggestions, it is important to interpret the PCA results carefully, taking into consideration the loadings of each principal component and the proportion of variance explained. Additionally, further analysis could be conducted using the principal component scores, such as regression or clustering analysis. For the factor analysis, additional exploration and comparison with other software may be necessary to ensure accurate results.

6. References:

1. Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer.
2. Jolliffe, I. T. (2002). Principal component analysis (2nd ed.). Springer.
3. Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). Latent variable models and factor analysis: A unified approach (3rd ed.). John Wiley & Sons.
4. R documentation for the `princomp` function: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/princomp

5. R documentation for the `factanal` function: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/factanal
6. R documentation for the `biplot` function: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/biplot
7. R documentation for the `screeplot` function: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/screeplot
8. R documentation for the `loadings` function: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/loadings