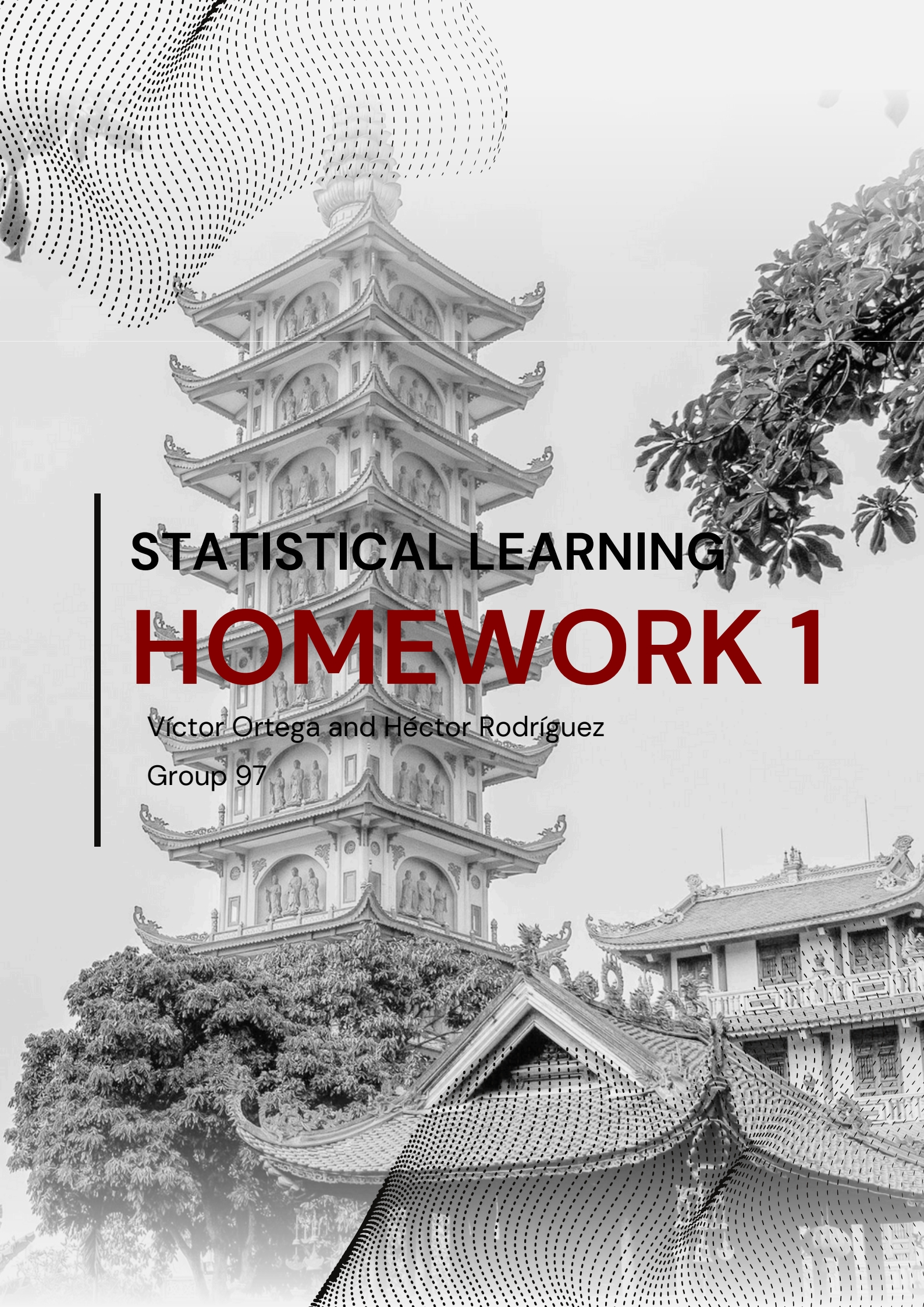




STATISTICAL LEARNING **HOMEWORK 1**

Víctor Ortega and Héctor Rodríguez

Group 97



INDEX

01

Data preprocessing

02

Visualization tools

03

Regression models
and resampling tools

04

Principal Component
Analysis

05

Factor Analysis

06

Clustering tools



“

*Understanding a part of
Vietnam through data*

Introduction

During the entire homework we will be using the dataset "Vietnam1.csv", which contains demographic, socioeconomic and health information about almost 28.000 Vietnamese people.

The objective is to extract as many information as possible from the dataset, which will help us to answer many questions about relationships between different aspects of the Vietnamese society (socioeconomic context, health situation, etc.)

In order to do that, we will firstly clean the data and visualize it (preprocessing and visualization/exploratory analysis). Then, we will use both supervised and unsupervised learning tools in order to extract and interpret information from the dataset.

01

Data preprocessing

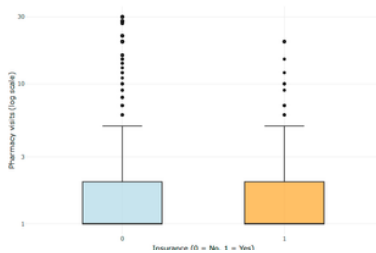
To begin with this project, we started by cleaning and transforming the dataset before the analysis. Missing values, duplicated, and unnecessary rows were checked and removed. We converted the categorical variables into factors and brought back to real scale those that were in logarithmic form. We continued studying the outliers using the 3-sigma and the IQR methods. However, we decided to keep them as we thought they could represent

realistic extreme cases. Finally, we decided to add new variables to the dataset to make it more useful and interesting, used to separate the dataset by age, education, or wealth (age_group, education_level, wealth_group) or to study other factors like high pharmacy visits (high_pharvis) or illness presence (is_ill). This part will help make the next parts of this project easier.

02

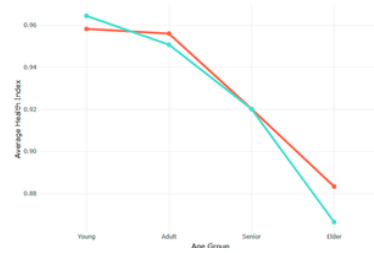
Visualization tools

Secondly, we used several visualization tools to better understand the dataset and explore relationships between the different variables. With ggplot2 and plotly, we created descriptive graphics showing how health, wealth, insurance, and demographic factors interacted in the Vietnamese population. Each plot helped us understand the data in more depth and guided the next parts of the project.



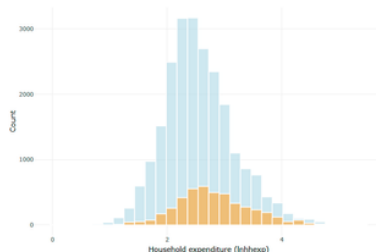
1° Does having insurance reduce pharmacy visits?

The first visualization compares pharmacy visits between insured and uninsured individuals using a boxplot on a logarithmic scale. The results show that having insurance does not significantly change the usual number of pharmacy visits. However, people without insurance present more extreme values, meaning some of them go to the pharmacy much more often



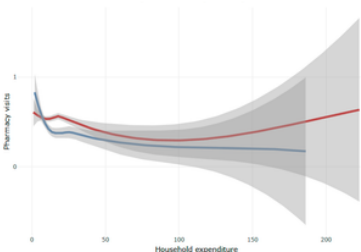
5° How does health change with age and gender?

The fifth visualization shows how the health index changes by age and gender. Health clearly declines with age for both men and women, but women maintain slightly better health in older groups. This pattern is common worldwide and consistent with their longer life expectancy



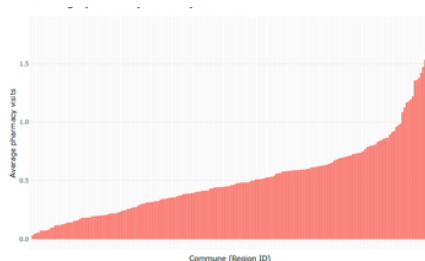
2° Are insured households economically wealthier?

The second visualization shows the distribution of household expenditure by insurance status using a histogram on a log scale. Both insured and uninsured households follow a similar normal distribution. However, among the highest spenders, there are slightly more insured households, suggesting that wealthier families are more likely to have insurance



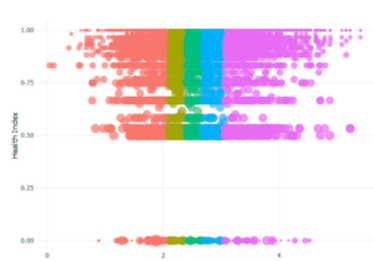
6° Does higher income lead to higher pharmacy visits?

The sixth visualization shows the relationship between household expenditure and pharmacy visits. The trend remains mostly flat, meaning higher income does not lead to more visits. However, people without insurance tend to go slightly more often, especially among high spender



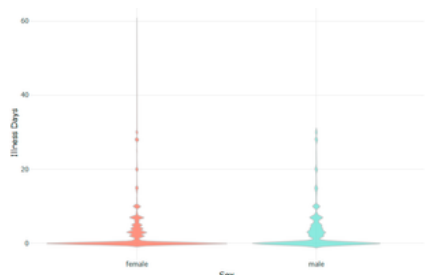
3° Do pharmacy visits vary across communes (regions)?

The third visualization shows the average number of pharmacy visits across communes. The results reveal large regional differences: some communes have very low averages, while others show much higher values. This suggests significant disparities in healthcare access or local health habits



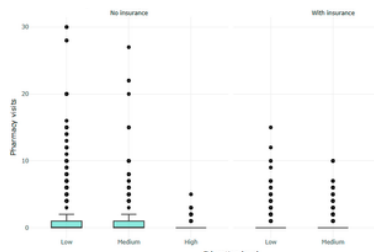
7° How do health, wealth, and illness interact across regions?

The seventh visualization explores how health, wealth, and illness interact across income groups. The results show that illness affects all groups similarly, but wealthier households tend to maintain a better health index, suggesting they cope better with illness than poorer ones



4° Are there gender differences in illness duration?

The fourth visualization compares the distribution of illness days between men and women. The violin plot shows that both genders have almost identical patterns, suggesting there are no clear gender differences in illness duration or access to healthcare



8° How do insurance and education influence health and spending?

The eighth visualization compares pharmacy visits by education level and insurance status. The results show that people with insurance go to the pharmacy less often across all education levels. Education itself does not change the pattern much, but having insurance clearly reduces visits

03

Regression models and resampling tools

Later on, we used supervised learning to study the impact that the sanitary house expenses (lnhhexp) has on the demographic and health situation of the individuals. To do so, we created different regression models and resampled the data with the help of K-Fold cross validation, so that we could also evaluate the performance of the models and study the balance between bias and variance. By computing this bias/variance trade-off, we avoided underfitting and overfitting models.

We wanted to answer questions such as: Does household healthcare spending increase with age? Do people with higher education levels have higher medical expenses? Does having health insurance increase healthcare spending (by facilitating access) or reduce it (by covering costs)?

Firstly, we selected the most important variables and discarded those with irrelevant data. Then, we used 80% of the data (train) to predict the other 20% (test) and observe its performance. In this case, we used a linear regression model to predict the variable 'lnhhexp'. With the summary() function we observed the variables that have the greatest impact on 'lnhhexp', such as education (educ) or insurance. Then, in order to evaluate the performance of the model, we computed the mean squared error (MSE) and square root (RMSE). The small RMSE value suggests that the model is able to explain a considerably big amount of the variability of 'lnhhexp', while still having a little error.

```
summary(lm_fit)

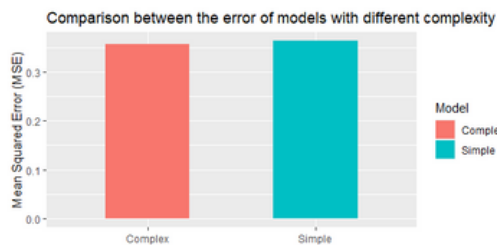
Call:
lm(formula = lnhhexp ~ age + educ + married + illness + illdays +
    pharvis + insurance + sex, data = train)

Residuals:
    Min       1Q   Median       3Q      Max
-2.44675  -0.40189  -0.05502   0.34976   2.55604

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.1942394   0.0163576 134.142 < 2e-16 ***
age          0.0886674   0.0033070  27.126 < 2e-16 ***
educ        0.0759185   0.0021699  34.987 < 2e-16 ***
married1    -0.0920670   0.0104319  -8.826 < 2e-16 ***
illness     -0.0820563   0.0037381 -21.815 < 2e-16 ***
illdays     -0.0031768   0.0009194  -3.455 0.000551 ***
pharvis      0.0140068   0.0031702   4.396 3.25e-05 ***
insurance    0.1310579   0.0113333  11.529 < 2e-16 ***
sexmale     -0.0021600   0.0080789  -0.267 0.789189

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5977 on 22134 degrees of freedom
Multiple R-squared:  0.09045, Adjusted R-squared:  0.09012
F-statistic: 275.1 on 8 and 22134 DF, p-value: < 2.2e-16
```



Afterwards, we applied a K-Fold Cross Validation with K = 10, to check many different combinations with train and test, which gives solid foundation to the conclusions, as instead of checking one sample we are checking many. During the cross validation, we created two models, one simpler than the other one, to study the bias/variance trade-off and avoid underfitting and overfitting models. In this case, we can observe that the simpler model has a higher error value, which means a higher bias and, therefore, the model is underfitting. At the same time, the complex model has a lower error value, but not too much, as it would be overfitting. We assume that the complex model is balanced in terms of bias/variance.

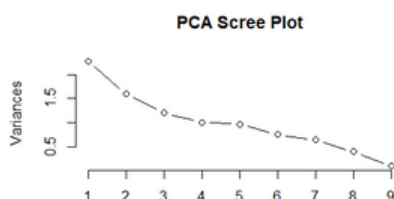
In conclusion, the regression model and resampling suggest that variables such as 'educ' and 'insurance' have a great impact on the 'lnhhexp' variable. This is logical, as richer people who spend more money on insurance and education will also spend money on their homes. Also, in this case it is better to use a complex model with all of the selected variables, because the simpler one (with just a few variables) loses important data. However, increasing the complexity too much is not good, as it will produce an overfitting model.

04

Principal Component Analysis (PCA)

In this part of the homework, we developed a Principal Component Analysis to reduce dimensionality, in order to study better the hidden structures between the most important variables of the dataset. PCA allows us to group similar variables with the help of linear combinations, making way easier both the interpretation and visualization of the data.

We wanted to answer questions such as: Can we reduce the dimensionality of the dataset without losing too much information? Which variables are the most influential in the variability of the data? Can we detect patterns or groupings among individuals (for example, between men and women)? Are there relationships between health, economy, and age within the principal components?



Firstly, we selected the most important (numeric) variables. Then, we computed all of the principal components. In this case, we chose to study the first three principal components, which contain almost the 60% of the total variance. This is more than enough to study the most part of the data. We confirmed and supported this decision by computing the eigenvalues and by applying the elbow method.

The loadings analysis allowed us to identify the most important variables of each component. Thanks to this, we were able to relate PC1 to "health situation", PC2 to "age" specially, and PC3 to "socioeconomic context". In order to visualize the relation between the different components, we created different plots and added the sex factor to get even more information. This graphics helped us understand a lot of information about the dataset.

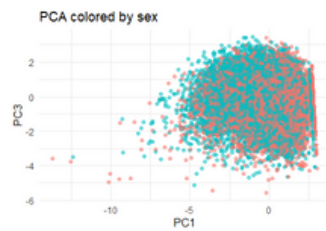
```
sort(abs(principal.components[,1]), decreasing = TRUE)
sort(abs(principal.components[,2]), decreasing = TRUE)
sort(abs(principal.components[,3]), decreasing = TRUE)

# we can observe that the first two components are more related to health
# and the third one is more related to socioeconomic context

age_real  illdays  pharvis  insurance  sex_num  acldays  lnhhexp  educ
0.5014269  0.4509702  0.4112782  0.1274962  0.3024056  0.0121245  0.0049051  0.0476723  0.0372493

age  age_real  illdays  pharvis  insurance  educ
0.127817408  0.460150418  0.401903234  0.117182875  0.321664007  0.284117591  0.216115709
0.0149999  0.0149999  0.0149999  0.0149999  0.0149999  0.0149999  0.0149999

educ  lnhhexp  pharvis  illdays  age_real  age  illdays  sex_num  acldays
0.48022902  0.48022902  0.48022902  0.48022902  0.48022902  0.48022902  0.48022902  0.48022902  0.48022902
```



There is a slight ascending inclination, which makes sense as older people tend to get ill more easily than younger people.

Individuals with lower socioeconomic levels (left side) do not present data for health records. The less fortunate do get ill but cannot attend hospitals, so the data does not exist. Also, we can see that the upper half of the graphic presents more data than the lower one, because of the same reason.

There is not a lot of information to extract out of a dense, uniform, round cloud of data. As well as in the first plot, the inclination is slightly ascending. This makes sense as older people tend to be richer, as they have had more time to earn money than the youth.

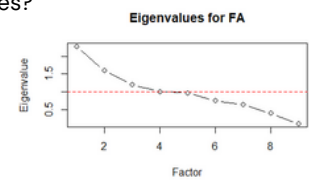
05

Factor Analysis (FA)

We used Factor Analysis to identify the factors that explain the correlations between the different variables, which is the same as studying the reason behind the creation of the principal components. It is important to distinguish between reducing dimensionality to study data easily (PCA) and studying the reason why different variables are more similar than others (FA).

We wanted to answer questions such as: Are there latent factors that explain the internal structure of the variables? How many factors should we retain to adequately represent the information? Which variables are most related to each other and how do they group into factors? How are individuals distributed according to their factor scores?

Firstly, as well as in the PCA, we will select the most important (numeric) variables. We will compute the correlation matrix and the eigenvalues, in order to determine the optimal number of factors with the Kaiser rule (values > 1). The graphic shows 3 values above 1, so we will study the first 3 factors.



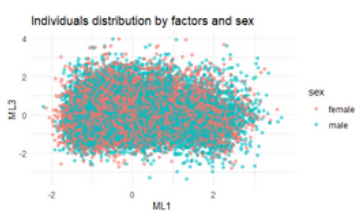
Then, when checking the most influential variables of each factor, we gathered the following information:

- Factor 1: Social and economic context of the individual (variables such as Inhhexp, insurance and education). A high score in this factor implies a great social and economic condition (e.g.: great insurance and education)
- Factor 2: Health situation of the individual (variables such as illness, illdays and pharvis). A high score in this factor implies a great health condition (e.g.: no recent illnesses)
- Factor 3: General information about the individual (variables such as age and sex, specially age). A high score in this factor implies that the person is old.

Afterwards, we studied and visualized the relations between the different factors, and got a lot of information that matched the information shown by the PCA.



The right half of the graphic presents more data than the left one; it is denser. This makes sense, as people with a great social and economic situation regularly attend hospitals and healthcare, while people from a lower social class find more difficulties. Also, the individuals with lower socioeconomic level do not present data for health records. This means that the less fortunate do get ill but cannot attend hospitals, so the data does not exist.



This graphic does not provide a lot of information. The result, a round cloud of uniform data, suggests that there are not many socioeconomic differences in terms of age. We could say that the society in Vietnam is equally distributed in terms of socioeconomic context. There are young rich people, old rich people, young poor people, and old poor people.

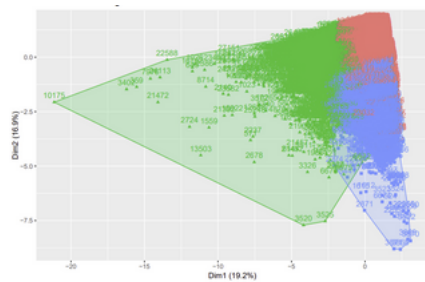


The inclination of the graphic, which looks like the shape of an arrow pointing to the right side, makes perfect sense, as both the elders and the kids tend to get ill regularly, while adults are less likely to get ill. Also, when a kid or an elder gets ill, the hospital is usually the first destination, while many adults avoid going to the hospital unless it is a serious illness or injury.

Clustering tools

Finally, we applied several clustering methods to identify distinct profiles within the population and compare their performance. The goal was to find natural groupings based on demographic, economic, and health variables and see which technique gave the clearest and most useful results. We used five approaches, K-means, PAM, Kernel K-means, Hierarchical, and EM, after selecting and scaling the variables and checking

correlations. Because of the computation time, some functions used a dataset sample, though the full data could be used with a better computer, we also used an anomalously reduced n_{boots} in all bootstraps, ideally it would be higher. After testing all models, K-means and Hierarchical clustering gave the most interpretable and well-separated groups, both showing three clear profiles in the data. PAM, Kernel K-means, and EM presented too much overlap, so overall K-means and Hierarchical offered the best balance between simplicity, clarity, and interpretability.

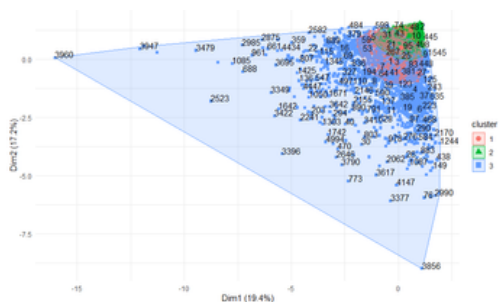


1° K-means

We applied the K-means algorithm to group individuals by their demographic, economic, and health characteristics. To get all this information, we used several functions to test, visualize, and interpret the results, including methods to find the optimal number of clusters and to analyze their centers and PCA dimensions. The elbow, silhouette, and gap methods were used to determine the best number of clusters, and we finally kept three for simplicity and interpretability. The groups represented: the average population, younger and moderately educated, people with higher pharmaceutical and household expenses, and a wealthier, better-educated, and insured group. Despite some overlap, K-means offered a balanced and interpretable result.

2° PAM

We then applied the PAM (Partitioning Around Medoids) algorithm to compare the results with K-means and check if it could provide better group separation. Several functions were used to test different cluster numbers using the elbow, silhouette, and gap methods. The results suggested between 3 and 5 clusters, and we finally selected 4 as the most balanced option. However, the visualization showed strong overlap between clusters, meaning PAM could not separate the data clearly. When comparing both methods, the results were quite different, confirming that PAM created less defined groups than K-means.

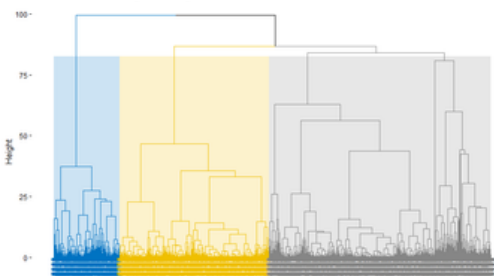


3° Kernel k-means

Next, we tested the Kernel K-means algorithm using an RBF kernel to see if it could improve the separation of the groups found by K-means. We kept three clusters for consistency and used functions to inspect the model, check compactness, and visualize the results. However, the method did not improve the clustering: the second group was almost completely nested inside the first, and the third overlapped both. Therefore, Kernel K-means did not provide any clearer structure than the standard version.

4° Hierarchical Clustering

We then applied Hierarchical Clustering using Ward's method to explore the data structure more visually. The dendrogram showed three clearly separated groups, confirming the adequacy of this approach. The clusters represented: older, lower-education, and uninsured individuals, young and healthy people, and educated, insured, and wealthier individuals. When compared with K-means, the results were partially similar, but the hierarchical model provided a clearer and more interpretable structure overall.



5° Expectation Maximization

Finally, we used the EM algorithm with a Gaussian Mixture Model to try a probabilistic approach. The model found six clusters, but most of them were highly overlapped, and one small cluster seemed to be outliers. When we compared it with the other methods, EM looked similar to PAM and only partly matched the hierarchical one, confirming that the groups were not very well defined.