African Institute for Mathematical Sciences (AIMS)–Senegal

High-Dimensional Data Analysis (HDDA)

Scroll, Tap, Repeat

A Data-Driven Look at Student Phone Behavior

Group 1 members:
Adama Telly Ba
Wandiya James
Fatou Binetou Mbaye
Nguemtchueng Tsemo Danielle

Supervisor: Prof. Sophie Dabo Niang

Date: January 22, 2026

Contents

Abstract

This report analyses student phone behaviour using survey responses from 369 AIMS–Senegal students. We first recall the theoretical background of Singular Value Decomposition (SVD), Correspondence Analysis (CA), Multiple Correspondence Analysis (MCA), and Principal Component Analysis (PCA), and explain how these methods are connected. We then apply PCA to five quantitative variables (Phone, SocialNetworks, Happiness, Walk, InstagramRatio) and interpret the main behavioural dimensions. Finally, we fit a multiple linear regression and perform backward selection to identify the most important predictors of phone time. The results consistently show that social network time is the strongest driver of phone time, while happiness has a protective (negative) association.

1 Introduction

Smartphones are essential tools for communication and learning, but they can also fragment attention and reduce deep study time. At AIMS–Senegal, the academic pace is intense, so repeated small distractions can accumulate into a real performance cost.

Instead of giving opinions about "phone addiction", we focus on what the data say. Using a survey of 369 students, we study relationships between total phone time and other behaviours (social networks, happiness, walking, and Instagram profile). Our workflow follows three steps: (i) exploratory structure with PCA, (ii) interpretation of behavioural profiles, and (iii) explanation using linear regression.

2 SVD, CA, MCA and PCA

This chapter summarises the core ideas needed for the rest of the report.

2.1 Singular Value Decomposition (SVD)

SVD is a matrix factorisation that decomposes a data matrix into orthogonal directions and singular values. It is widely used for dimensionality reduction and numerical stability, and it is the mathematical foundation behind many multivariate methods. For a real matrix $M \in \mathbb{R}^{I \times J}$ of rank r, SVD writes

$$M = U \Lambda V^{\top},$$

where U and V have orthonormal columns and $\Lambda = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ contains singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$. A key property is optimal low-rank approximation: keeping the first $s < r$ components gives the closest rank-s approximation in Frobenius norm.

2.2 Correspondence Analysis (CA)

CA is an exploratory method for qualitative data arranged as contingency tables. It represents row and column categories in a low-dimensional factorial space: categories that co-occur frequently appear close together, and weakly related categories appear farther apart.

2.3 Multiple Correspondence Analysis (MCA)

MCA extends CA to several qualitative variables (typical survey questionnaires). It helps identify profiles of individuals and groupings of modalities: individuals close in the map tend to share similar responses, and nearby modalities are often observed together.

2.4 Relationship between SVD, CA and MCA

Although they apply to different data types, SVD, CA, and MCA are closely related. SVD provides the common mathematical foundation, while CA and MCA adapt the same philosophy to qualitative data: the goal is to structure and understand the data before more advanced modelling.

2.5 Principal Component Analysis (PCA)

PCA is the main tool used in the applied part of this project because our core variables are quantitative. PCA seeks new orthogonal axes (principal components) that maximise the variance of projected data, and it is closely connected to SVD on a centred (and often scaled) data matrix [**?**]. In practice, PCA provides:

- a low-dimensional representation of individuals,
- a way to interpret relationships between variables,
- a basis for describing dominant behavioural patterns.

3 Data and methods

3.1 Dataset

The dataset contains responses from 369 AIMS students and includes five quantitative variables: **Phone** (minutes/day), **SocialNetworks** (minutes/day), **Happiness** (score 1–10), **Walk** (km/day), and **InstagramRatio** (followers/following).

A first descriptive summary indicates strong outliers for **Phone** and **InstagramRatio**. For example, the maximum phone time reaches 1440 minutes (24 hours), and InstagramRatio can reach about 51.

3.2 Step 1: Check missing values

Before running PCA, we verify that the dataset does not contain missing values. We compute the number of missing values per variable and the proportion relative to the sample size. This step ensures that PCA is not affected by incomplete observations.

3.3 Step 2: Center and reduce the data

For PCA, variables must be on a comparable scale. We therefore apply centering and standardisation:

- **Centering** subtracts the mean of each variable, so each variable has mean 0.
- **Reducing (standardising)** divides by the standard deviation, so each variable has variance 1.

This is important because otherwise variables with large numerical values (e.g., Phone minutes or InstagramRatio) would dominate the PCA and the other variables would be almost ignored.

In the HTML workflow, this is implemented using `scale(dataset, center=TRUE, scale=TRUE)`.

3.4 Exploratory analysis

We use boxplots to inspect distributions and outliers, then a correlation matrix to check which variables are positively or negatively related.

3.5 PCA workflow

We apply PCA to the centred and reduced variables and interpret: (i) the scree plot (explained variance), (ii) the correlation circle (loadings), (iii) contributions and $\cos^2$ (representation quality), and (iv) the individual map and biplot.

3.6 Regression workflow

We fit a multiple linear regression model with Phone as the response and candidate predictors including SocialNetworks, Happiness, Walk, and Instagram status. We then apply backward stepwise selection (5% threshold) to keep a simpler model with significant predictors.

4 PCA results and interpretation

4.1 Distributions

Figure 4.1 summarises distributions. Phone time and social network time are right-skewed, with a small group reporting very high values.

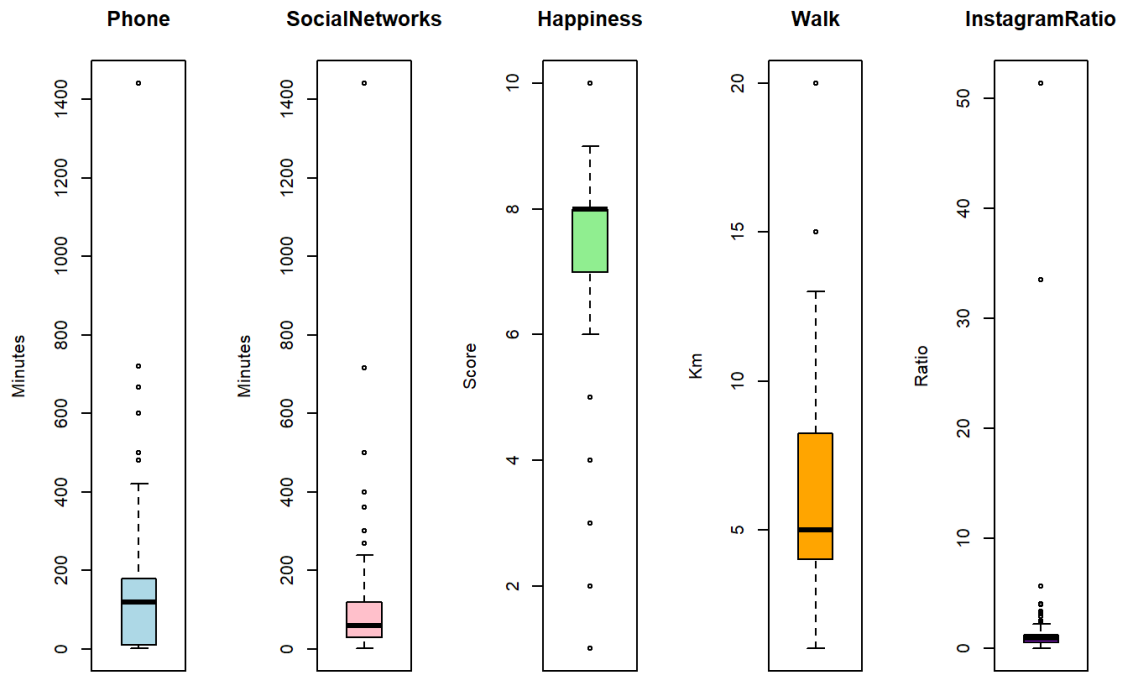

Figure 4.1: Boxplots for Phone, SocialNetworks, Happiness, Walk, and InstagramRatio.

4.2 Correlation structure

The correlation matrix (Figure 4.2) shows a strong positive link between Phone and SocialNetworks, and a negative association between Happiness and phone time (moderate).

4.3 Choosing the number of components

From the scree plot (Figure 4.3), the first two dimensions capture most of the structure (Dim1 $\approx$ 35%, Dim2 $\approx$ 22.9%), so we focus on a 2D interpretation.

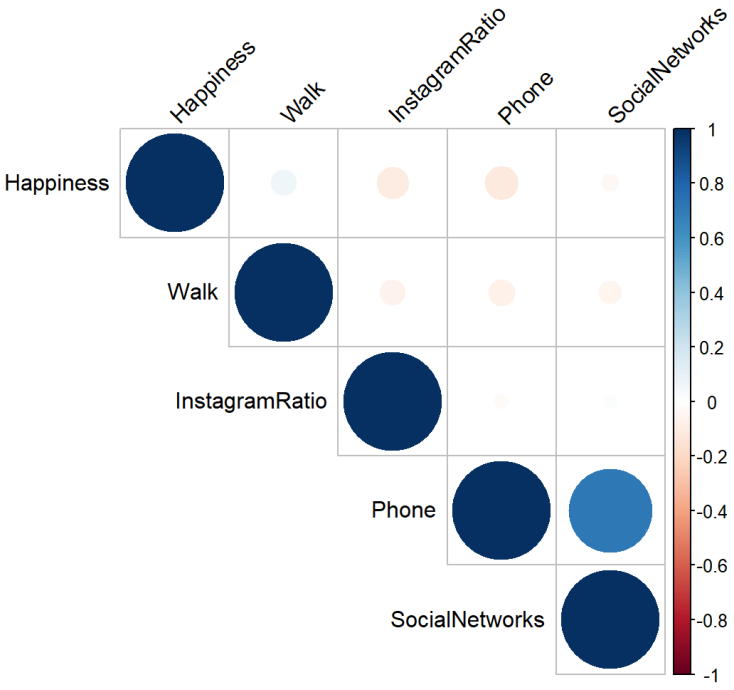

Figure 4.2: Correlation matrix of the main variables.

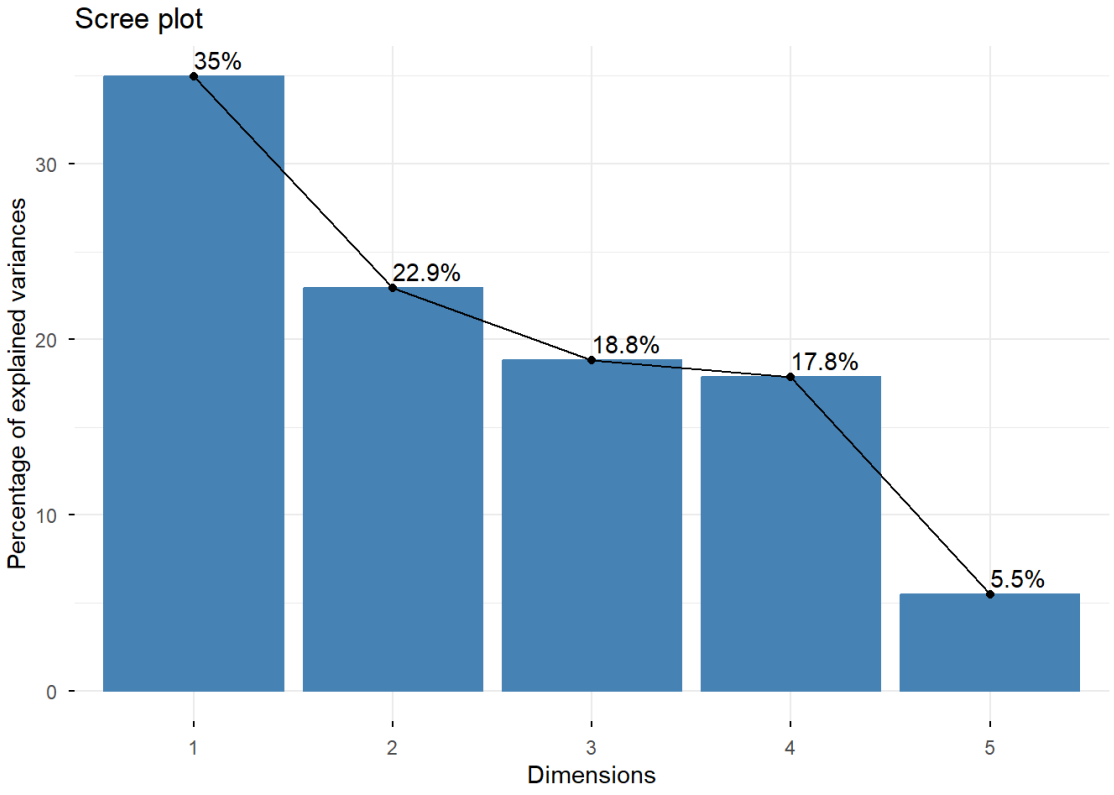

Figure 4.3: Scree plot of explained variance.

4.4 Variable structure (loadings)

The correlation circle (Figures 4.4 and 4.5) indicates that Phone and SocialNetworks align strongly with Dim1, while Happiness points in the opposite direction. InstagramRatio

contributes more to Dim2.

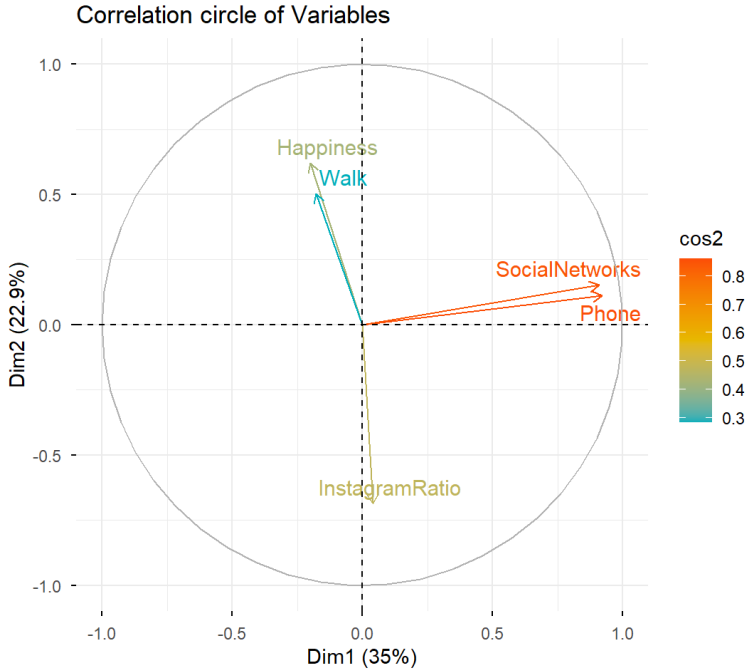

Figure 4.4: Correlation circle coloured by $\cos^2$ (quality of representation).

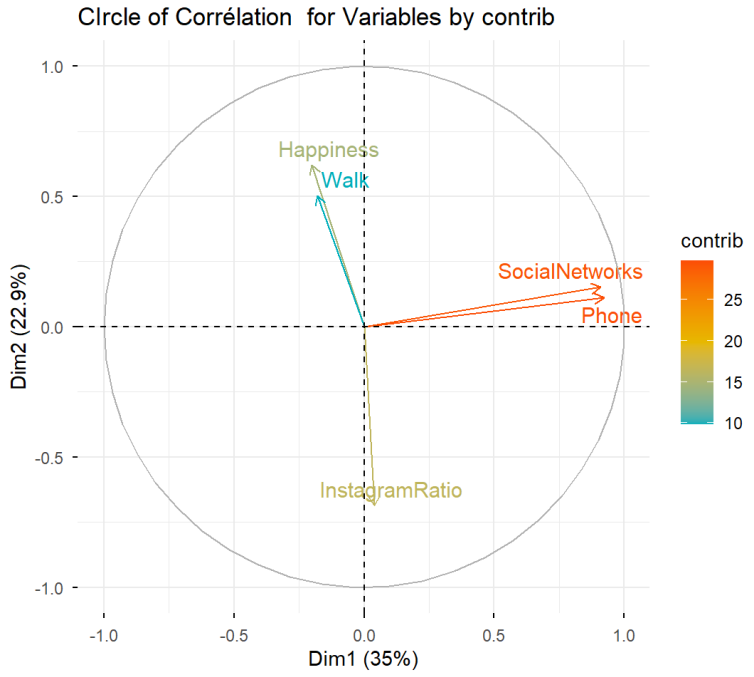

Figure 4.5: Correlation circle coloured by contribution.

4.5 Contributions and cos^2

Figures 4.6–4.9 confirm that Dim1 is dominated by Phone and SocialNetworks, while Dim2 is driven mainly by InstagramRatio and Happiness.

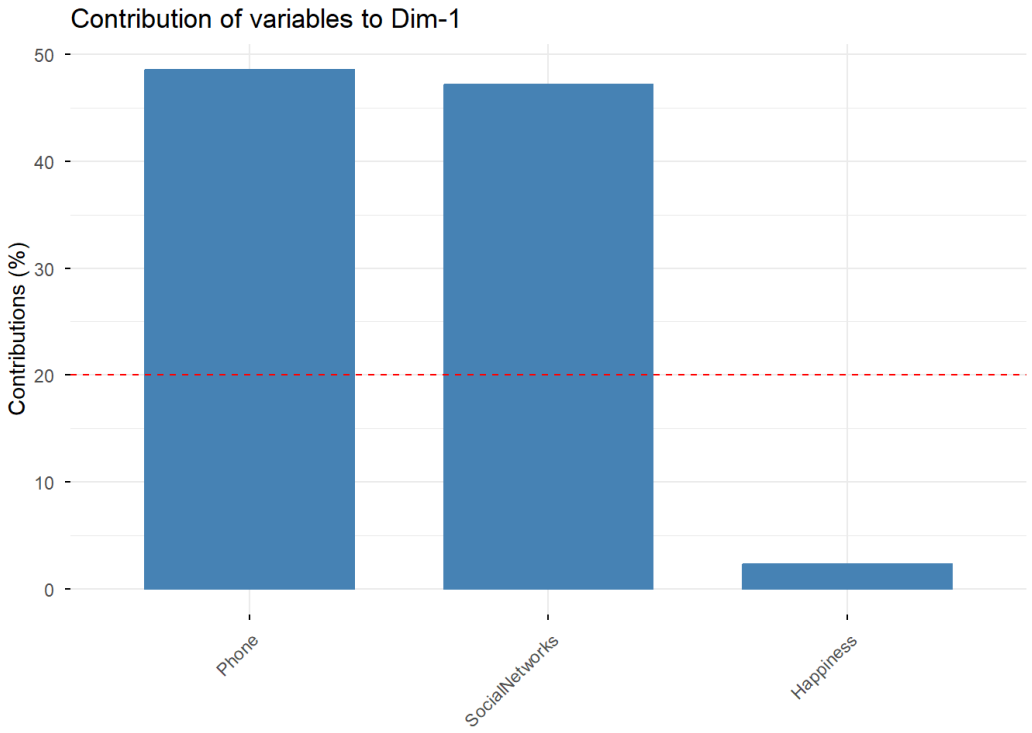

Figure 4.6: Contribution of variables to Dim1.

4.6 Individuals and behavioural profiles

The individuals map (Figure 4.10) shows a dense cluster of typical students and a few extreme users. Combining this with the loadings, Dim1 can be read as an "always-connected" axis (high phone + social networks) versus a more balanced behaviour.

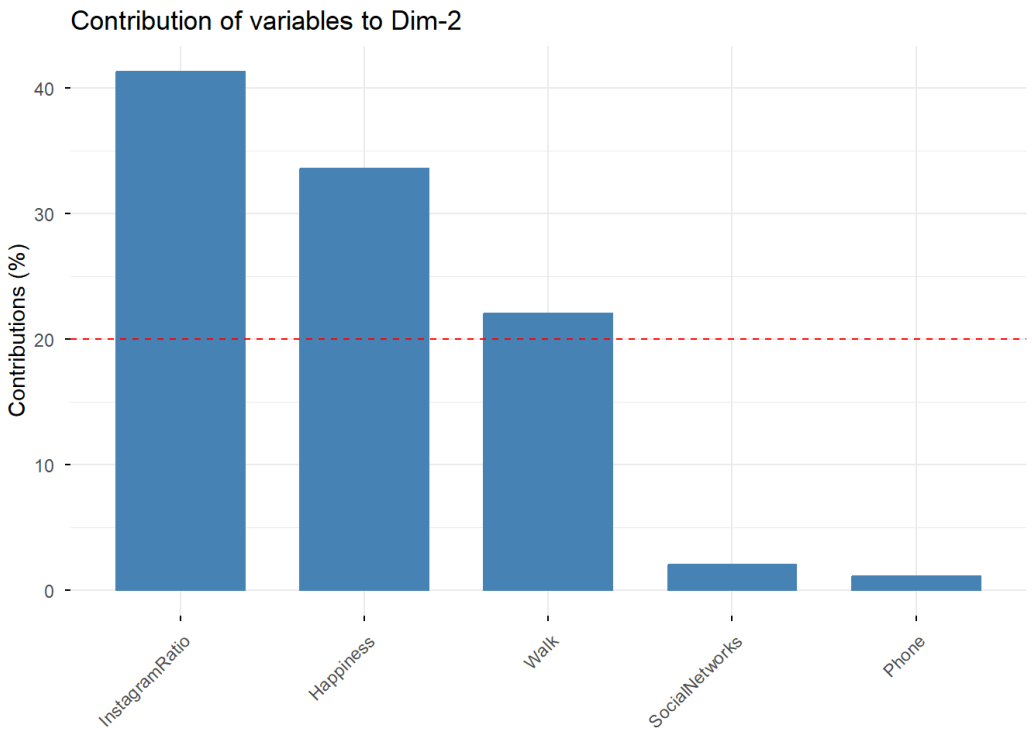

Figure 4.7: Contribution of variables to Dim2.

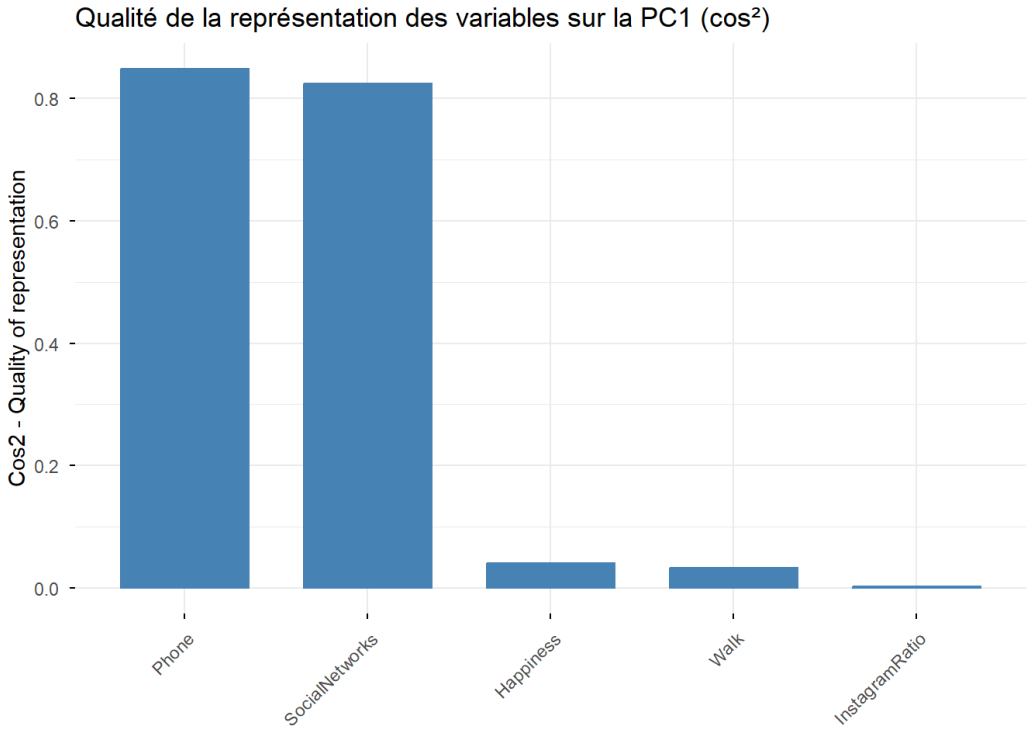

Figure 4.8: Cos^2 of variables on PC1.

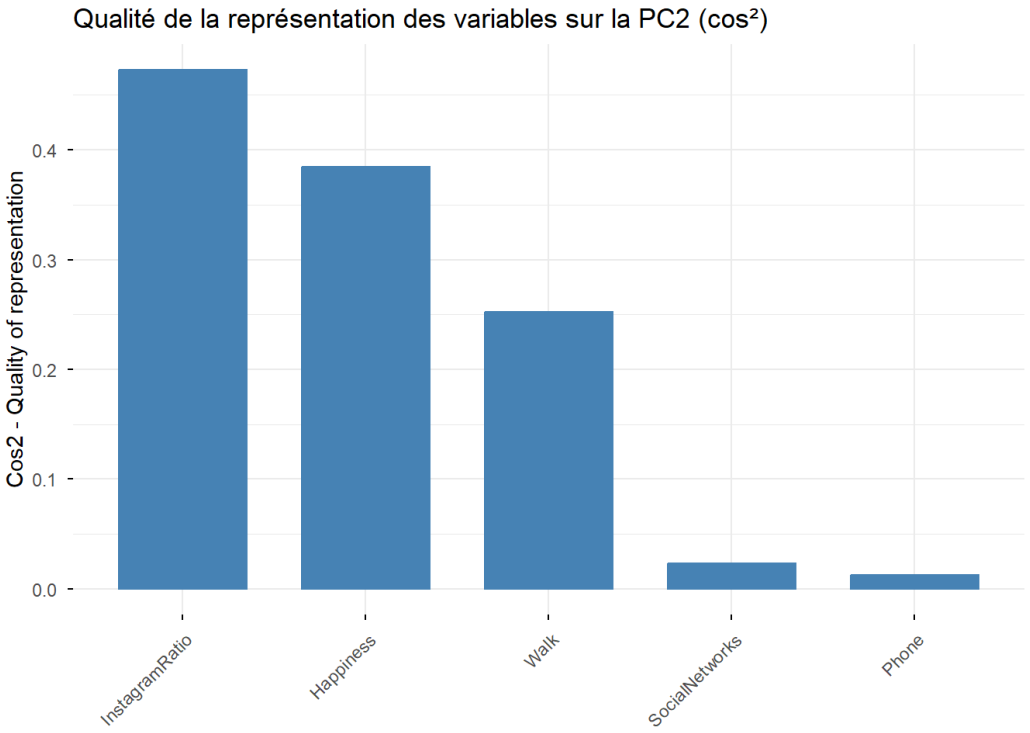

Figure 4.9: Cos^2 of variables on PC2.

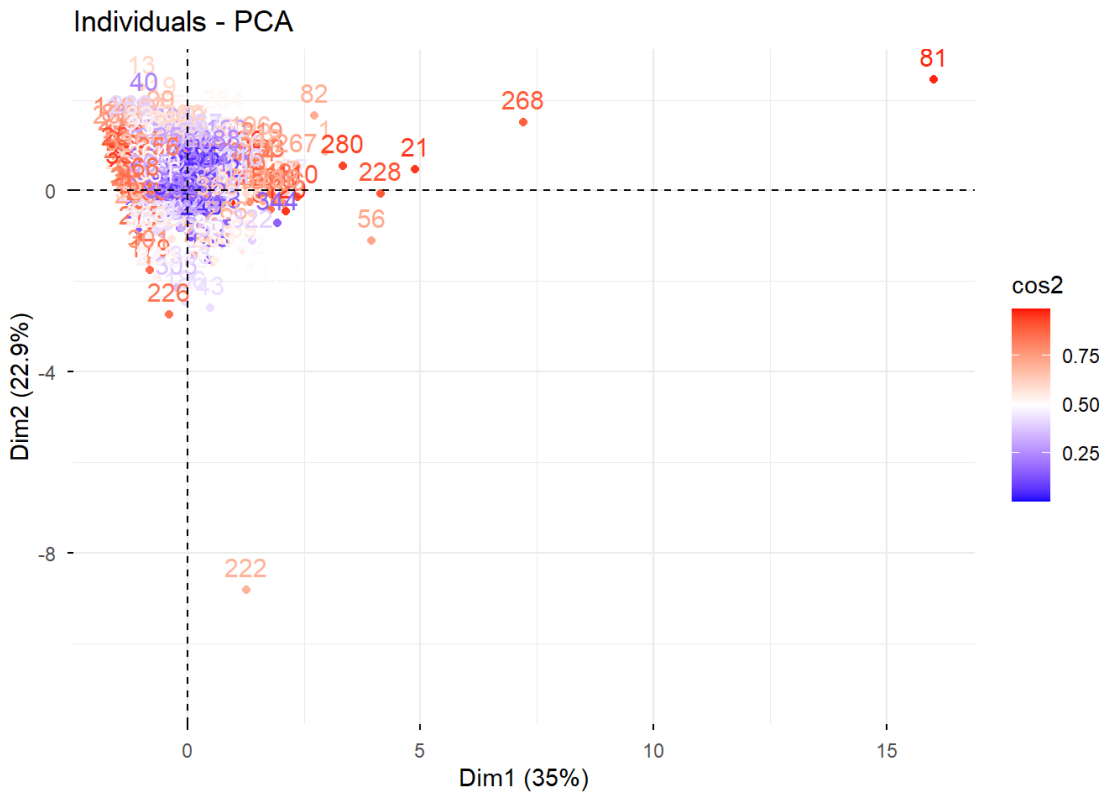

Figure 4.10: Individuals projected on the first two PCA dimensions.

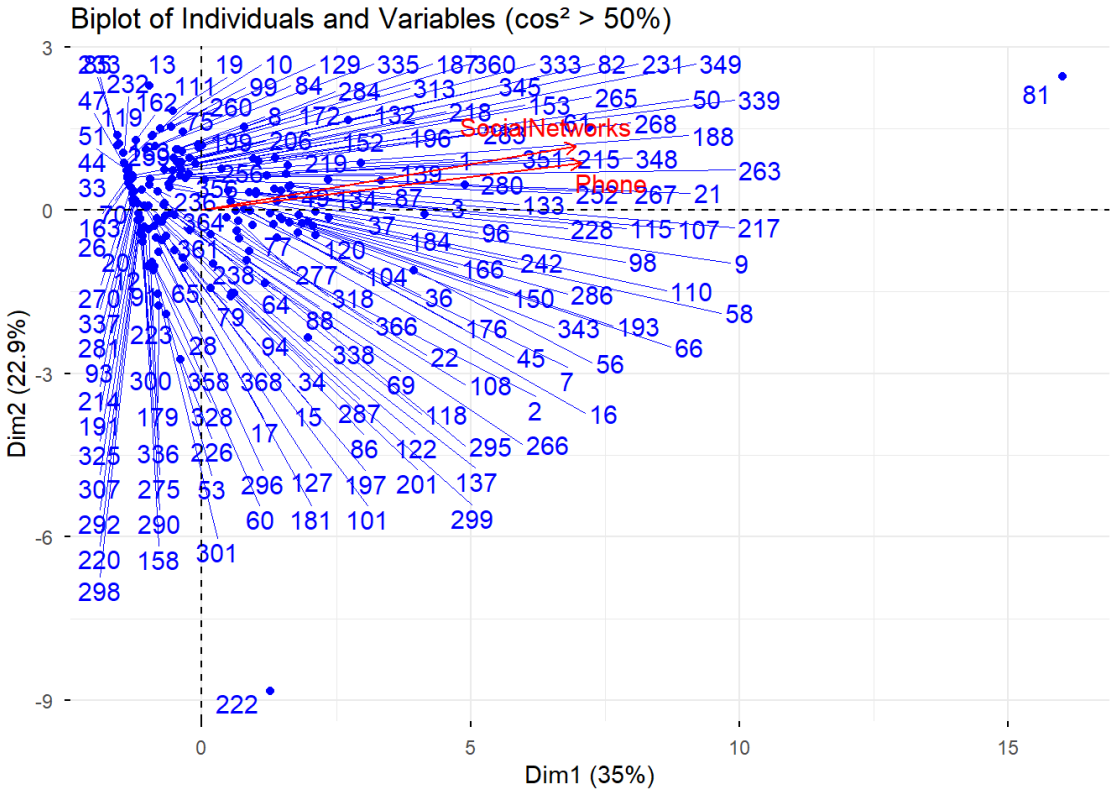

Figure 4.11: Biplot of individuals and variables ($\cos^2 > 50\%$ in the original analysis).

5 Multiple linear regression

5.1 Full model

We fit the following model:

$$\text{Phone} = \beta_0 + \beta_1\,\text{SocialNetworks} + \beta_2\,\text{Happiness} + \beta_3\,\text{Walk} + \beta_4\,\text{LowInstagram} + \beta_5\,\text{HighInstagram} + \varepsilon.$$

The model is strongly significant overall (F-statistic = 83.38 on 5 and 362 df, $p < 2.2 \times 10^{-16}$) and explains about 53% of the variance (Multiple $R^2 = 0.5352$, Adjusted $R^2 = 0.5288$).

Table 5.1: Full regression model estimates.

Predictor	Estimate	Std. Error	t value	p-value
Intercept	111.88	25.25	4.43	1.24e-05
SocialNetworks	0.9108	0.0456	20.00	$< 2e$-16
Happiness	-7.3952	2.9765	-2.49	0.0134
Walk	-0.9574	1.4848	-0.65	0.5194
LowInstagram	17.9076	11.6051	1.54	0.1237
HighInstagram	-34.7791	19.9727	-1.74	0.0825

Interpretation (what the coefficients mean)

- **SocialNetworks** is the strongest factor: each additional minute spent on social networks adds about 0.91 minutes of total phone use. fileciteturn2file1L23-L24
- **Happiness** has a negative effect: an increase of one point in happiness is associated with about 7–8 fewer minutes of phone time per day.
- **Walk** is not significant in this model.
- Instagram profile variables show weaker effects (borderline p-values), so we check model selection.

5.2 Backward stepwise selection (5% threshold)

Backward selection removes variables that are not significant. In the final model, the remaining significant predictors are SocialNetworks, Happiness, and HighInstagram. The adjusted R^2 remains close (Adjusted $R^2 = 0.5275$), meaning we simplify without losing much explanatory power.

Table 5.2: Final model after backward selection.

Predictor	Estimate	Std. Error	t value	p-value
Intercept	115.81	23.05	5.03	7.92e-07
SocialNetworks	0.9116	0.0455	20.02	$< 2e$-16
Happiness	-8.1423	2.9497	-2.76	0.00606
HighInstagram	-39.4431	19.7607	-2.00	0.04667

Main takeaway

Both the full model and the selected model tell the same story: phone time is mainly explained by social networks (increase phone time) and reduced by happiness (protective effect).

6 Conclusion

This study combines theory and application to analyse student phone behaviour at AIMS–Senegal. After reviewing the links between SVD, CA, MCA, and PCA, we used PCA to reveal the main behavioural structure in five quantitative variables. The first dimension separates an "always-connected" pattern (high phone and social networks) from a more balanced behaviour associated with higher happiness.

Linear regression confirms the same message quantitatively: social network time is the dominant predictor of total phone time, while happiness is negatively associated with phone use. A practical implication is that interventions should target social media habits first (limits, notifications, fixed checking windows) and support routines that increase well-being.

Bibliography

The Elements of Statistical Learning, , by Trevor Hastie, Robert Tibshirani, Jerome Friedman,2009,