DS311
Exploratory Data Analysis

Multivariate Relationships and Dimensionality Reduction in R: A Learning Portfolio

---

Overview

This learning portfolio covers how to explore and analyze relationships involving multiple variables in R, and how to apply dimensionality reduction techniques to simplify complex datasets. We will discuss strategies for visualizing multivariate data (and how to avoid common pitfalls like overplotting), introduce **Principal Component Analysis (PCA)** and **Factor Analysis** for reducing dimensionality, and emphasize best practices for interpreting these analyses in real-world contexts. Throughout, we use hands-on R examples to illustrate concepts, culminating in a case study of student academic performance. By the end, you should be able to conduct an exploratory multivariate analysis, apply PCA and factor analysis in R, and draw meaningful conclusions from the results.

Prerequisites

- **Basic R Knowledge:** You should be comfortable with basic R operations, including creating data frames, subsetting, and using functions. Familiarity with basic plotting and the concept of a correlation is helpful.
- **Statistical Concepts:** Understanding of terms like *mean*, *variance*, *correlation*, and *factor* will be useful. Prior exposure to the idea of principal components or eigenvalues is a plus but not required.
- **R Setup:** Make sure you have R (and ideally RStudio) installed. We'll use both base R and the **tidyverse** (especially `ggplot2` for plots and `dplyr` for data manipulation). You can install the tidyverse package with `install.packages("tidyverse")`. We will also use functions from base R's **stats** package (`prcomp()` for PCA and `factanal()` for factor analysis), which come pre-installed with

  R.

- **Data:** We will work with a publicly available *Student Performance* dataset (UCI Machine Learning Repository) covering student achievement in secondary school. The dataset includes multiple variables on student demographics, family background, study habits, and grades, which makes it ideal for practicing multivariate analysis. We will access this dataset directly in R during the examples and lab.

(If you plan to follow along, ensure your R environment has internet access to fetch the dataset or download it locally from the UCI repository. No other external data is required, and we will also demonstrate using built-in datasets when appropriate.)

Section 1: Exploring Multivariate Relationships

Understanding **multivariate relationships** means examining how more than two variables relate to each other simultaneously. This is an extension of bivariate analysis (just two variables at a time) to higher dimensions. For example, in a student dataset, a *bivariate* question might be "How do study time and exam score relate?" whereas a *multivariate* question could be "How do study time, class absences, and parental education **together** relate to exam scores?" Analyzing multivariate relationships can reveal deeper insights, but it is also more challenging—human beings cannot easily visualize data beyond three dimensions, so we need special techniques to explore these relationships effectively.

Visualizing Multiple Variables

A first step in multivariate analysis is often to look at **pairwise relationships** among variables as an approximation. We can create multiple bivariate plots or compute correlations for each pair of variables. One common tool is the **scatterplot matrix**, which shows scatterplots for every pair of variables in a grid layout. Another tool is the **correlation matrix**, which quantifies the linear relationship between every pair of variables.

Let's start with an example. We'll use a subset of the student performance data focusing on numeric variables for simplicity (e.g., exam scores, study time, absences, etc.). First, we load the data and select a few relevant columns:

```
# Load necessary package
library(tidyverse)
```

```
url <-
 ↪  "https://raw.githubusercontent.com/arunk13/MSDA-Assignments/master/IS607Fall2015/Assignment3/student-mat.csv"
student <- read.csv(url, sep = ";")

# Select a subset of numeric variables for exploration
numeric_vars <- c("G1", "G2", "G3", "studytime", "failures", "absences", "Medu", "Fedu")
student_sub <- student[ , numeric_vars]

# Quick peek at the data
dim(student_sub)        # dimensions of the subset
```

```
## [1] 395   8
```

```
head(student_sub, 3)    # first 3 rows
```

```
##   G1 G2 G3 studytime failures absences Medu Fedu
## 1  5  6  6         2        0        6    4    4
## 2  5  5  6         2        0        4    1    1
## 3  7  8 10         2        3       10    1    1
```

Running the above, we see our subset has 395 observations (students) and 8 variables. The variables are:

- **G1, G2, G3:** First, second, and final period grades (numeric, 0-20 scale).
- **studytime:** Weekly study time (ordinal, 1-4 scale where 1 = "<2 hours", 4 = ">10 hours").
- **failures:** Number of past class failures (numeric count).
- **absences:** Number of school absences.
- **Medu, Fedu:** Mother's and father's education level (0 = none, 4 = higher education).

These variables vary in scale and type (some are truly numeric, others are ordered categories encoded as numbers), but for exploration we'll treat them as numeric where it makes sense (e.g., `studytime` is ordinal but we can still compute correlations on its numeric codes).

Pairwise Correlations: A quick way to identify strong relationships is the correlation matrix. Let's compute the correlation matrix for these variables:

```
cor_matrix <- cor(student_sub, use = "complete.obs")
round(cor_matrix, 2)
```

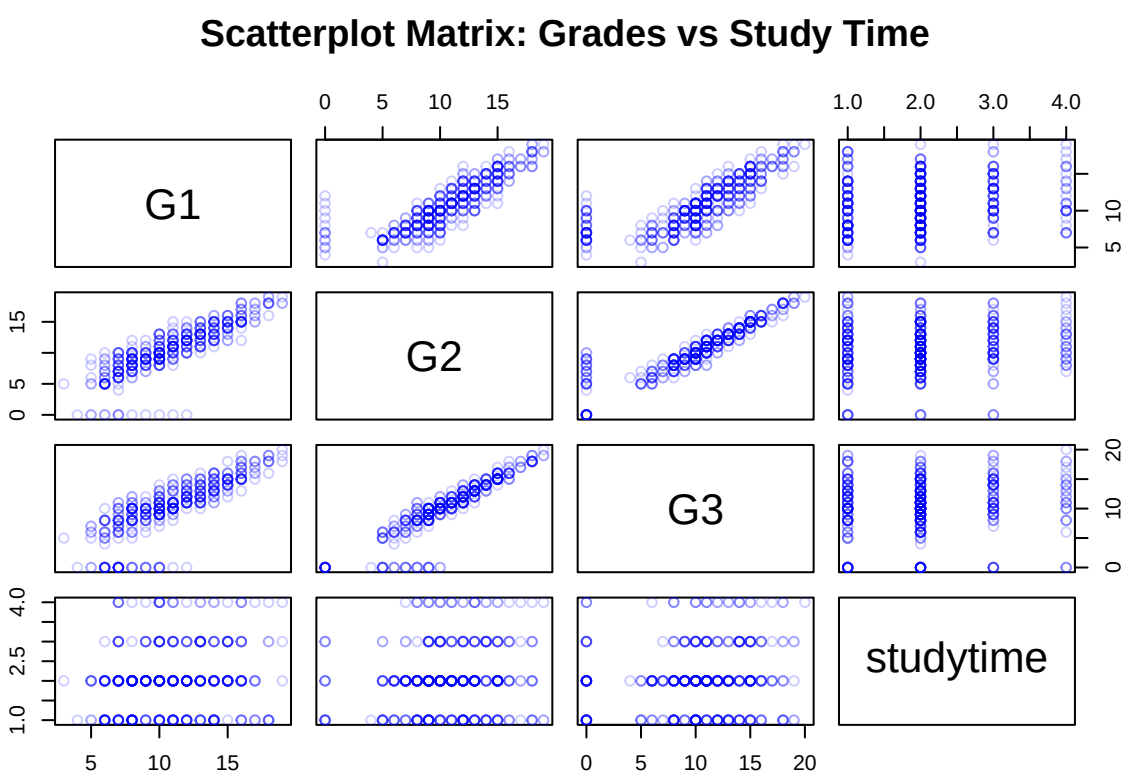
```
##               G1    G2    G3 studytime failures absences  Medu  Fedu
## G1          1.00  0.85  0.80      0.16    -0.35    -0.03  0.21  0.19
## G2          0.85  1.00  0.90      0.14    -0.36    -0.03  0.22  0.16
## G3          0.80  0.90  1.00      0.10    -0.36     0.03  0.22  0.15
## studytime   0.16  0.14  0.10      1.00    -0.17    -0.06  0.06 -0.01
## failures   -0.35 -0.36 -0.36     -0.17     1.00     0.06 -0.24 -0.25
## absences   -0.03 -0.03  0.03     -0.06     0.06     1.00  0.10  0.02
## Medu        0.21  0.22  0.22      0.06    -0.24     0.10  1.00  0.62
## Fedu        0.19  0.16  0.15     -0.01    -0.25     0.02  0.62  1.00
```

This gives a table of correlation coefficients (rounded to 2 decimals for clarity). For example, you might observe that `G1`, `G2`, and `G3` are all very highly correlated (coefficients often above 0.8). This makes sense: a student who does well in the first period tends to do well in the second and final periods too. Indeed, if we check, `G3` (final grade) has a strong positive correlation with `G2` and `G1` (e.g., around 0.85-0.90, meaning a linear relationship is quite strong). We also see that `failures` (number of past failures) is negatively correlated with grades (for instance, `failures` vs `G3` might be around -0.35 or so), indicating that more past failures tends to accompany lower final grades. Similarly, `absences` likely shows a negative correlation with grades (students who skip more classes generally perform worse). On the other hand, parental education (`Medu`, `Fedu`) may show a mild positive correlation with student grades — higher parental education could be associated with slightly better student performance, which aligns with common expectations..

> **Check-in:** Why do we look at pairwise correlations as a first step in multivariate analysis?
> **Answer:** It helps identify strong relationships between variables that we might want to investigate further. In a multivariate context, pairwise correlations are a quick diagnostic to spot patterns (or multicollinearity) among variables. However, they don't tell the whole story when more than two variables interact.

Scatterplot Matrix: Correlations give numeric summaries, but it's often useful to visualize the data. We can plot a scatterplot matrix for a subset of variables. For example, let's visualize the relationships among the three grade variables and study time:

```r
pairs(student_sub[c("G1", "G2", "G3", "studytime")],
      main = "Scatterplot Matrix: Grades vs Study Time",
      col = rgb(0,0,255,50, maxColorValue=255))  # semi-transparent blue points
```



(In the above, we used a semi-transparent color for points to help with overplotting, as explained shortly.)

This scatterplot matrix will show, for instance, `G1` vs `G2`, `G1` vs `G3`, `G2` vs `G3`, etc., as well as plots of grades vs study time. You will likely notice the grade-vs-grade plots have points roughly along a diagonal line (indicating strong positive relationship). The plots involving `studytime` might show a weaker trend or more spread (study time has only four discrete values, so points form vertical bands). Overall, such a matrix gives a quick visual summary of pairwise relations.

Interpreting Patterns: From these exploratory visuals, we start forming hypotheses. For example, if `G1` and `G3` are strongly related, final exam performance is largely reflected by first period performance (perhaps students consistent over time). If `studytime` has little correlation with grades, it might suggest simply counting hours isn't a clear indicator of performance (maybe quality of study or other factors matter). Keep in mind these are just associations; we are not concluding causation. In fact, it's crucial to consider other variables that might be influencing these relationships.

Caution – Simpson's Paradox and Context: When interpreting bivariate relationships, remember that a trend seen in an overall dataset can be misleading if there are subgroups with different trends. A **confounding variable** (a third variable) can affect the relationship between two others (statisticsbyjim.com). For instance, imagine we found that students with **higher** study time actually had slightly **lower** grades on average. Does studying more cause lower grades? A likely confounder could be that students who were struggling (hence had lower grades) started studying more to catch up, whereas naturally talented students might study less. In this hypothetical scenario, *within* each performance level the relationship might be positive (studying helps), but overall it appears negative because of the mix of groups — this is akin to **Simpson's paradox**. The lesson is: always interpret correlations in context and, when possible, check if

the relationship holds across relevant subgroups (like male/female, different schools, etc.). Failing to account for such factors can distort results, so you must consider confounders to understand the true relationship (statisticsbyjim.com).

> **Check-in:** What does it mean if two variables are strongly correlated (say r = 0.9)? And does a zero correlation mean no relationship?
> **Answer:** A strong correlation (close to 1 or -1) indicates a strong linear relationship — as one variable increases, the other tends to increase (positive correlation) or decrease (negative correlation) in a roughly linear fashion. However, correlation only measures linear association; r = 0 could mean *no linear relationship*, but the variables might still have a non-linear relationship. Also, correlation does not imply causation — two variables might both be influenced by a third factor (confounder) rather than directly affecting each other.

Best Practices: Avoiding Overplotting in Scatterplots

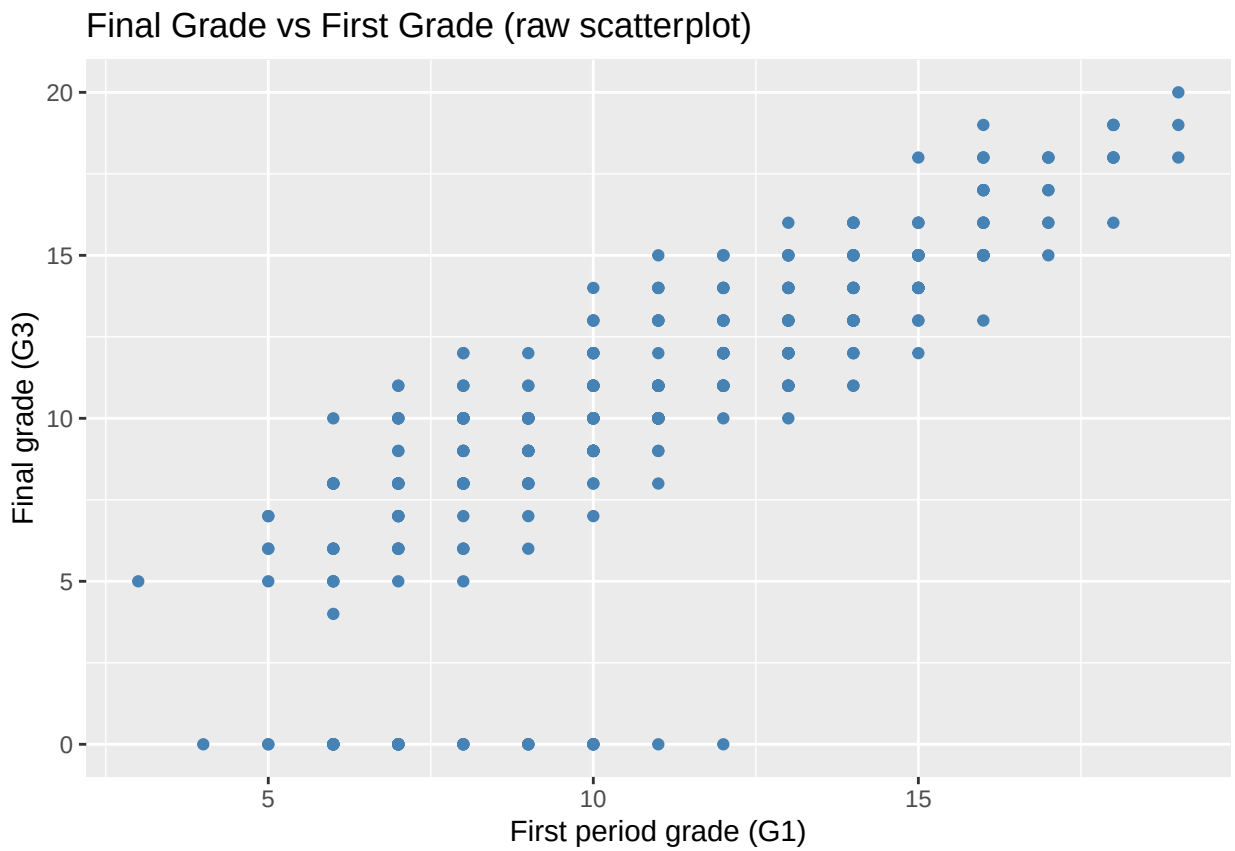
When visualizing multivariate data, especially with scatterplots, a common problem is **overplotting**. Overplotting occurs when many points fall on top of each other in a plot, which can obscure the true relationship between variables (bookdown.dongzhuoer.com). This often happens with large datasets or when data values are discrete or clustered. For example, in our student data, grades are recorded as whole numbers; if 50 students scored 15 in both G1 and G3, then 50 points will all sit exactly at (15,15) in a scatterplot — you'll just see what looks like one point.

Techniques to Address Overplotting: There are several strategies to mitigate this issue

- **Transparency (Alpha Blending):** Making points semi-transparent (`alpha` less than 1 in `ggplot2`) so that areas with many points appear darker. For instance, if you have 2000 points, setting `alpha = 0.3` means roughly 3 overplotted points are needed to get full opacity. This way, dense regions become visibly darker.
- **Jittering:** For discrete data, adding a small random noise to points so they don't overlap exactly. `geom_jitter()` in ggplot2 will spread points a little in the x and y directions. This is very useful for data like ours where many students share identical scores.
- **Smaller or Different Shapes:** Using smaller plotting characters or hollow outlines can help when the amount of overlap is not extreme. A tiny dot or a hollow circle won't obscure as much as a large solid dot.
- **Binning and Density:** For very large datasets, you might aggregate points into bins (e.g., using a 2D histogram or hexbin plot with `geom_bin2d()` or `geom_hex()`) or even plot density contours (`stat_density2d()`). These techniques visualize the distribution of points rather than each point, effectively handling heavy overplotting.

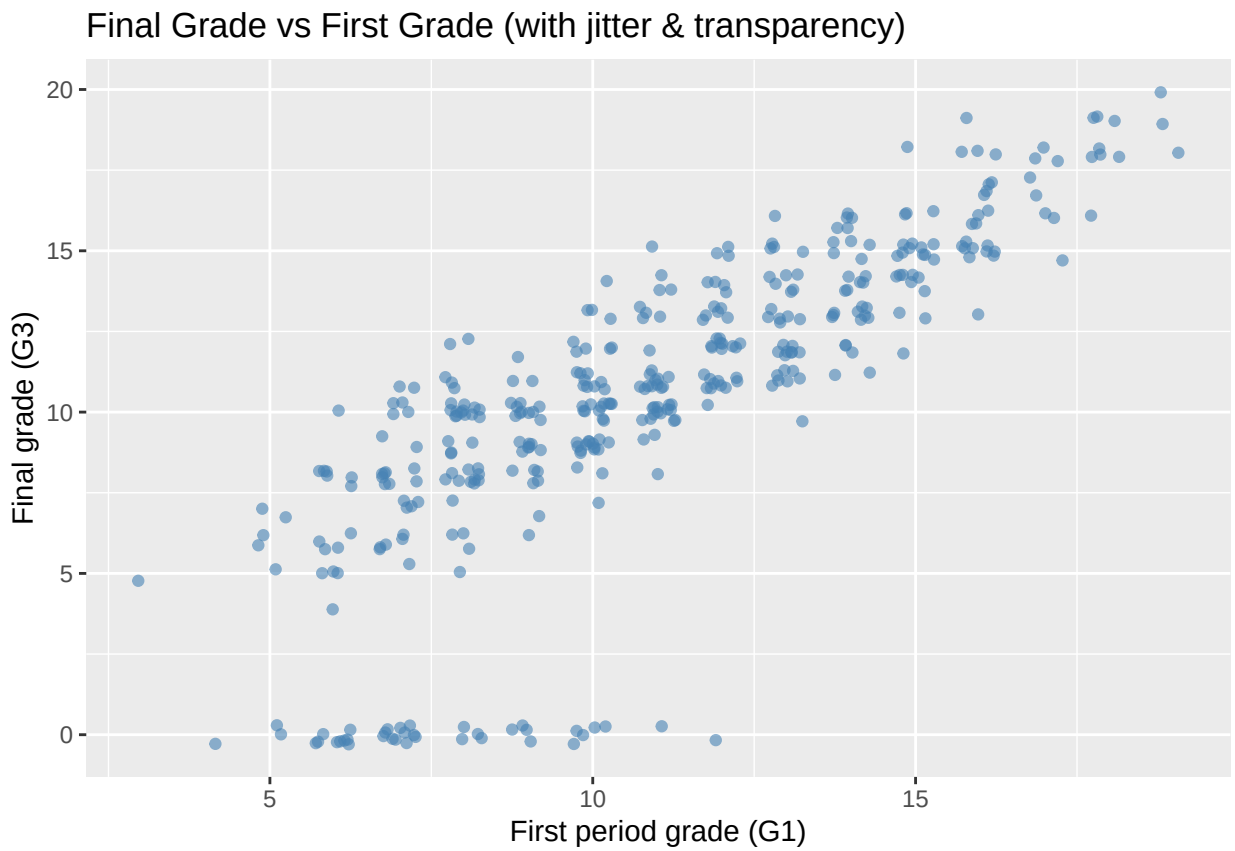
Let's apply a couple of these techniques to our example. We'll revisit the relationship between first period grade (`G1`) and final grade (`G3`). A basic scatterplot first:

```
ggplot(student, aes(x = G1, y = G3)) +
  geom_point(color = "steelblue") +
  labs(title = "Final Grade vs First Grade (raw scatterplot)",
       x = "First period grade (G1)", y = "Final grade (G3)")
```



This plot will show the general upward trend (as expected, students with higher first period grades tend to have higher final grades), but you might notice that many points are plotted over each other (since grades are integers 0–20). Now let's improve it:

```
ggplot(student, aes(x = G1, y = G3)) +
  geom_jitter(width = 0.3, height = 0.3, color = "steelblue", alpha = 0.6) +
  labs(title = "Final Grade vs First Grade (with jitter & transparency)",
       x = "First period grade (G1)", y = "Final grade (G3)")
```



Here we've added `geom_jitter()` with a small width and height (0.3) and set `alpha = 0.6` (60% opacity). The jitter spreads points a bit (0.3 is 30% of a unit, so e.g., a point at (15,15) might be randomly nudged to (15.1, 14.8), etc.). The transparency means if points do overlap, seeing a darker spot indicates multiple points. In the jittered plot, you'll discern the concentration of students along the diagonal and identify clusters (e.g., many students around 10-15 range for both grades). No single point completely hides another now.

You could further use shape adjustments (like `shape = 1` for a hollow circle) or binning for larger data. In our case, jitter+alpha suffices. The resulting pattern reaffirms a strong positive relationship between G1 and G3, with most points near the diagonal line. A few outliers may appear (e.g., a student with high G1 but much lower G3 or vice versa), which could prompt questions about what caused the change – something to possibly investigate with other variables or qualitative information.

> **Check-in:** What is *overplotting* and why is it a problem in data visualization? Name one method to address it.
> **Answer:** Overplotting is when data points overlap so much in a plot that it's hard or impossible to see patterns or density – you might just see a blob of ink. It obscures the true relationship between variables. One method to address it is using transparency (alpha blending) so that areas with many points become darker, revealing where points are concentrated. Other methods include jittering points to avoid exact overlap, using smaller or hollow point shapes, or binning points into heatmaps.

Interpreting Multivariate Patterns in Context

Visualization and correlation give us pieces of the puzzle. The next step is interpretation, which means translating patterns into real-world insights. Always ask: *Does this relationship make sense given what I*

know about the context? For example, if we find that **mother's education** (`Medu`) has a correlation of, say, 0.2 with final grade (`G3`), that suggests a slight trend that students with more educated mothers perform better. In context, this might be explained by socio-economic factors or emphasis on academics at home. However, that correlation is not destiny; there are many students with high grades whose parents did not have higher education, and vice versa. Contextual understanding helps prevent overinterpreting a correlation.

Another point is considering **multiple variables together**. Perhaps neither study time nor absences alone has a very strong correlation with grades, but students who both study a lot *and* rarely miss class do significantly better. Such an interaction would be a multivariate pattern you won't see by looking at each pair of variables separately. Sometimes building a multiple regression model or using color/shape in plots for a third variable can uncover these interactions. For instance, coloring the G1 vs G3 scatterplot by whether the student has extra educational support (`schoolsup`) might show that among students with low first-period grades, those with extra support improve more by final grade than those without (a hypothetical scenario). The key is that *multivariate* exploration often involves looking at combinations of conditions.

Finally, remember that statistics can hint at relationships, but **real-world validation** is crucial. We use domain knowledge to decide if a pattern is plausible or if it might be spurious. And we must be cautious not to infer causality from correlation alone. For example, a strong association between two variables could be due to a third factor influencing both (confounding) or even just chance, especially in smaller samples. Always consider performing a deeper analysis or experiment to test causal hypotheses.

Having explored and visualized our data, we might have some specific findings (e.g., "*Grades are highly correlated with each other, indicating consistency; past failures and absences are associated with lower grades; no obvious linear relationship between study time and grades…*"). Now, we will turn to dimensionality reduction techniques to see the data from another angle: instead of examining variables one pair at a time, can we summarize multiple variables into a few composite dimensions that capture the essence of the data?

Section 2: Dimensionality Reduction Techniques

When faced with datasets that have many variables, it becomes hard to understand the data as a whole. **Dimensionality reduction** techniques aim to simplify the dataset by reducing the number of variables while retaining most of the important information. This not only helps with visualization (for example, reducing data to two dimensions so we can make a scatterplot) but also can help with modeling by removing redundant features.

In this section, we'll cover two popular methods: **Principal Component Analysis (PCA)** and **Exploratory Factor Analysis (EFA)**. Both techniques find combinations of the original variables, but they have different goals and interpretations, which we'll discuss.

Principal Component Analysis (PCA)

Concept: PCA is a statistical technique used to analyze high-dimensional data and capture its most important patterns of variation (datacamp.com). It does so by transforming the original variables into new variables called **principal components (PCs)**. Each principal component is a linear combination of the original variables. The first principal component (PC1) is the combination that explains the largest amount of variance in the data; the second principal component (PC2) is the combination that explains the next largest amount of variance *subject to being uncorrelated with PC1*, and so on. In essence, PCA finds the "directions" in the data where the data spread (variance) is maximized, and these directions are orthogonal (perpendicular) to each other.

Some key points about PCA:

- PCA is an *unsupervised* method (it doesn't use any outcome variable; it just looks at the structure of the data).

- The principal components are ordered by how much variance they explain. Typically, a few of the first PCs explain most of the variance in the original data.
- We often **standardize** variables before PCA (i.e., subtract mean and divide by standard deviation for each variable), especially if they are on different scales. Otherwise, variables with larger scales or higher variance will dominate the first components. (For example, if one variable is "income in dollars" ranging in the thousands and another is "years of education" ranging 0-20, the income variable's variance is much larger and would overpower the PCA if not scaled).
- The results of PCA include **eigenvalues** (the variance explained by each PC) and **loadings** (the coefficients of each original variable in each principal component). An eigenvalue corresponds to a principal component's variance; if you divide it by the number of variables, you get the proportion of total variance that component accounts for.
- A common criterion to decide how many components to keep is **Kaiser's rule**: retain components with eigenvalue > 1 (meaning the component explains more variance than an average single original variable in standardized data). Another way is to look at a **scree plot** (plot of eigenvalues) and find the "elbow" where additional components have diminishing returns.

Let's apply PCA to our student data subset. We will include several variables related to student performance and background. For demonstration, we'll use: `G1, G2, G3, failures, absences, Medu, Fedu` (we exclude `studytime` here to keep example interpretable, but you could include it too). We'll perform PCA on these 7 variables.

```
# Select variables for PCA
pca_vars <- c("G1", "G2", "G3", "failures", "absences", "Medu", "Fedu")
pca_data <- student[ , pca_vars]

# Perform PCA with scaling (center=TRUE, scale.=TRUE by default in prcomp when scale.=TRUE)
pca_res <- prcomp(pca_data, scale. = TRUE, na.action = na.omit)
# Print summary of PCA results
summary(pca_res)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.7645 1.2154 1.0094 0.8551 0.61038 0.44885 0.29212
## Proportion of Variance 0.4448 0.2110 0.1456 0.1045 0.05322 0.02878 0.01219
## Cumulative Proportion  0.4448 0.6558 0.8013 0.9058 0.95903 0.98781 1.00000
```

The `summary(pca_res)` output shows the importance of each principal component. For example, it might output something like:

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7
Standard deviation     1.7645 1.2154 1.0094 0.8551 0.61038 0.44885 0.29212
Proportion of Variance 0.4448 0.2110 0.1456 0.1045 0.05322 0.02878 0.01219
Cumulative Proportion  0.4448 0.6558 0.8013 0.9058 0.95903 0.98781 1.00000
```

Variance Explained

- **PC1** explains about **44.5%** of the variance.
- **PC2** explains about **21.1%**.
- **PC3** explains about **14.6%**.
- **PC4** explains about **10.5%**.
- The remaining PCs (PC5–PC7) each contribute less than ~5%.

So together:

- The **first two PCs** capture about **65.6%** of the variance.
- The **first three PCs** already cover **80.1%**.
- By PC4, you've explained **90.6%** of the dataset variance.

This means most of the structure in your data can be summarized in just **2–3 components**.

Kaiser's Rule (Eigenvalue > 1)

Eigenvalues are simply the square of the standard deviations:

- **PC1**: 1.7645^2 **3.11**
- **PC2**: 1.2154^2 **1.48**
- **PC3**: 1.0094^2 **1.02**
- **PC4**: 0.8551^2 **0.73**
- The rest are < 1

By **Kaiser's criterion**, you'd keep **PC1, PC2, and PC3**, since their eigenvalues are > 1.

Practical Takeaway

- **PC1–PC3** are the main drivers of variation, accounting for ~80% of the information.
- Beyond PC3, the gain in explained variance is small (diminishing returns).
- For visualization or dimensionality reduction, keeping **2 PCs** is often enough (good balance between simplification and variance retained).
- For modeling or interpretation, you might justify keeping **3 PCs** to capture >80% variance.

Interpretation in words: "This analysis shows that student performance data can be effectively reduced to about 2–3 principal components without losing much information. PC1 is the most dominant factor, capturing nearly half of the total variation, while PC2 and PC3 add meaningful but smaller contributions. Beyond that, additional components do not explain enough variance to be practically useful."

Now let's examine the **loadings** (the makeup of each principal component in terms of original variables):

```
round(pca_res$rotation, 2)
```

```
##              PC1   PC2   PC3   PC4   PC5   PC6   PC7
## G1          0.50  0.21  0.04 -0.13 -0.12 -0.79  0.21
## G2          0.52  0.24  0.06 -0.15  0.01  0.22 -0.78
## G3          0.51  0.23  0.12 -0.10  0.05  0.56  0.59
## failures -0.32  0.11  0.28 -0.90 -0.03  0.01  0.02
## absences  0.00 -0.16  0.94  0.28 -0.09 -0.04 -0.04
## Medu      0.26 -0.63  0.01 -0.19  0.71 -0.08  0.00
## Fedu      0.24 -0.64 -0.12 -0.17 -0.69  0.10  0.00
```

This will show a matrix of coefficients (loadings) for each variable on each PC, for example:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
G1	0.50	0.21	0.04	-0.13	-0.12	-0.79	0.21
G2	0.52	0.24	0.06	-0.15	0.01	0.22	-0.78
G3	0.51	0.23	0.12	-0.10	0.05	0.56	0.59
failures	-0.32	0.11	0.28	-0.90	-0.03	0.01	0.02
absences	0.00	-0.16	0.94	0.28	-0.09	-0.04	-0.04
Medu	0.26	-0.63	0.01	-0.19	0.71	-0.08	0.00
Fedu	0.24	-0.64	-0.12	-0.17	-0.69	0.10	0.00

Interpretation of PCA Loadings

Each principal component (PC) is a **weighted combination** of your original standardized variables. The numbers in the table are the *loadings* (weights). Larger absolute values mean the variable contributes more to that component.

PC1

- **Strong positive loadings**: G1 (0.50), G2 (0.52), G3 (0.51)
- **Negative loading**: failures (-0.32)
- Minor contributions: Medu (0.26), Fedu (0.24), absences (0.00)

Interpretation: PC1 is an **academic performance axis**. Students with high PC1 scores have high grades and few failures, while low scores indicate poor grades and more failures. This dimension essentially captures **overall student achievement**.

PC2

- **Strong negative loadings**: Medu (-0.63), Fedu (-0.64)
- Moderate positives: G1–G3 (~0.21–0.24)
- Small others

Interpretation: PC2 mainly contrasts **parental education** levels. A high PC2 score (less negative) reflects lower parental education, while a low PC2 score reflects higher parental education. It is somewhat independent of the grades dimension, separating **background factors from performance**.

PC3

- **Strong positive loading**: absences (0.94)
- Moderate positive: failures (0.28)
- Small positives for G3 (0.12)

Interpretation: PC3 reflects **attendance/engagement**. Students with high PC3 scores have many absences, regardless of grades. This component isolates **unique variance in attendance behavior**.

PC4

- **Very strong negative**: failures (-0.90)
- Moderate positives: absences (0.28), G1–G3 small negatives (~-0.10)

Interpretation: PC4 distinguishes students specifically by **failures**. Even though failures already appeared in PC1, here it is picked up as a **separate independent factor**.

PC5

- **Strong positive**: Medu (0.71)
- **Strong negative**: Fedu (-0.69)

Interpretation: PC5 separates **mother's vs father's education**. It suggests there is some independent contrast between the two parental education levels.

PC6

- **Strong negative**: G1 (-0.79)
- **Positive**: G3 (0.56), G2 (0.22)

Interpretation: PC6 highlights a **grade pattern contrast** — particularly separating G1 from G2 and G3. This could reflect differences in performance across grading periods.

PC7

- **Strong negative**: G2 (-0.78)
- **Strong positive**: G3 (0.59)

Interpretation: PC7 isolates a **specific distinction between G2 and G3** performance. This is a minor pattern since it explains very little variance.

Big Picture Summary

- **PC1**: Overall **academic success** (grades vs failures).
- **PC2**: **Parental education** background.
- **PC3**: **Attendance/absences** behavior.
- **PC4**: Purely **failures** as a unique dimension.
- **PC5**–**PC7**: Subtler contrasts (mother vs father education, differences among G1–G3 scores).

Together, PC1–PC3 already explain ~80% of the variance, meaning these three are the most interpretable and useful for downstream analysis or visualization.

Check-in: Why did we choose to standardize (scale) the variables before PCA in this example?
Answer: Because our variables had different units and variances (e.g., grades 0-20, failures 0-5, parental ed 0-4). PCA is influenced by the scale of variables — without scaling, a variable with a larger range or variance would dominate the principal components. Scaling puts them on equal footing (mean 0, SD 1 each), so each variable contributes based on its correlations, not its raw magnitude.

Check-in: If the first principal component (PC1) of a PCA has high positive loadings on G1, G2, G3 and high negative loading on failures, what would you interpret this component as?
Answer: I would interpret PC1 as an **overall academic performance** dimension. High values of PC1 mean high grades and few failures (strong student), while low PC1 means low grades and more failures (weak academic performance).

A note on usage: PCA is great for reducing many variables into a few summary indices. For instance, instead of using three exam scores in a regression model, we might use PC1 as a single "performance index". However, PCA doesn't differentiate between "cause" and "effect" variables; it merely finds patterns in the data. It's also sensitive to outliers (an extreme outlier can skew components). Always examine the data and perhaps try PCA with and without certain outliers if suspect. Additionally, PCA assumes linear relationships among variables (it captures linear combinations). If relationships are highly non-linear, PCA might not capture them well.

Factor Analysis

Concept: Factor Analysis is another technique to reduce dimensionality, but with a different perspective. It assumes that observed variables are influenced by a smaller number of unobserved latent variables called **factors**. The idea is that there are underlying factors which cannot be directly measured, but which cause the patterns of correlations we see among the observed variables (library.virginia.edu). For example, in education data, one might hypothesize factors like "learning aptitude" or "family support" that are not directly measured but manifest in multiple observed variables (grades, test scores, parental involvement, etc.). Exploratory Factor Analysis (EFA) tries to identify such factors from the data itself, without prior specification.

Key points about factor analysis (particularly EFA):

- It explains the covariance/correlation between variables via a few factors. Each factor is a latent construct, and each observed variable has a loading on each factor (similar to PCA loadings, but interpreted as how strongly that factor influences the variable).
- Unlike PCA, factor analysis separates out unique variance (variance in a variable not explained by common factors). In PCA, all variance is used. In factor analysis, we assume each observed variable = common part (from factors) + unique part (error or specificity).
- We need to decide the number of factors to extract. This can be guided by eigenvalues of the correlation matrix as well (looking at how many eigenvalues are >1, similar to PCA) or other criteria, but often it's guided by theory or interpretability as well.
- Factor analysis often benefits from rotation of factors. Rotation (e.g., varimax rotation) is a technique to make the factor structure clearer by simplifying loadings (it doesn't change the overall explained variance, but distributes it differently among factors). A varimax rotation, for example, tries to make each variable load high on one factor and low on others, to aid interpretation.
- The end result is a set of factors with loadings, and we interpret each factor by looking at which variables have high loadings on it. We often give the factor a name representing the concept those variables share. *This naming is subjective and requires domain knowledge; the analysis won't automatically label a factor for you*

Let's perform an exploratory factor analysis on our student data subset. Based on our PCA results and domain intuition, let's try extracting **2 factors** from the variables `G1, G2, G3, failures, absences, Medu, Fedu`. Our expectation (from PCA) is that one factor might capture "academic performance" (grades vs failures/absences) and another "family background" (parental education). We'll use R's `factanal()` function for EFA.

```
# Perform Factor Analysis with 2 factors, using varimax rotation
fact_res <- factanal(student[ , pca_vars], factors = 2, rotation = "varimax")
print(fact_res, digits = 2, cutoff = 0.3)
```

```
##
## Call:
## factanal(x = student[, pca_vars], factors = 2, rotation = "varimax")
##
## Uniquenesses:
##       G1       G2       G3 failures absences     Medu     Fedu
##     0.24     0.04     0.15     0.81     1.00     0.45     0.28
##
## Loadings:
##          Factor1 Factor2
## G1        0.87
## G2        0.98
## G3        0.92
## failures -0.37
## absences
## Medu               0.71
## Fedu               0.83
##
##                Factor1 Factor2
## SS loadings       2.78    1.24
## Proportion Var    0.40    0.18
## Cumulative Var    0.40    0.58
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 20.65 on 8 degrees of freedom.
## The p-value is 0.00813
```

Interpreting the Factor Loadings

- **Factor1**
  - High positive loadings: **G1 (0.87), G2 (0.98), G3 (0.92)**
  - Moderate negative loading: **failures (-0.37)**
  - Absences didn't load meaningfully here (uniqueness = 1.00).

Interpretation: Factor1 represents an **Academic Achievement factor**. Students with high scores on Factor1 perform well on all grade variables and have fewer failures. This latent factor is essentially capturing overall **student academic success**.

- **Factor2**
  - High positive loadings: **Medu (0.71), Fedu (0.83)**
  - No meaningful contributions from grades or failures.

Interpretation: Factor2 represents a **Parental Education factor**, reflecting the educational background of the parents. Students with higher Factor2 scores come from families with higher parental education.

This matches our earlier PCA interpretation — indeed, a well-known result is that if certain assumptions hold, the factors from EFA can align with principal components. The difference is that factor analysis is framing it as *there is a latent trait (academic ability) influencing all these grade outcomes*, whereas PCA was just a mathematical combination. In factor analysis, `G1, G2, G3` are high on Factor1 because we think there is some underlying skill or trait causing a student to do well across all exams.

Uniquenesses

- **G2 (0.04)** and **G3 (0.15)** have very low uniqueness → their variance is almost entirely explained by the factors (mostly Factor1).
- **Absences (1.00)** and **Failures (0.81)** have high uniqueness → most of their variance is not explained by the two retained factors. This suggests attendance and failure patterns may reflect other dimensions not captured in the 2-factor model.

Bottomline: The factor analysis output also provides uniqueness values for each variable, which indicate the proportion of variance of that variable *not* explained by the factors. For instance, if `G1` has uniqueness 0.24, that means 76% of its variance is explained by the common factors (which is high, consistent with G1 being mostly driven by the academic performance factor). Variables that didn't fit perfectly (like `absences` might have higher uniqueness, meaning the two factors don't capture as much of what absences measure).

Variance Explained

- **Factor1** explains **40%** of the variance.
- **Factor2** explains **18%** of the variance.
- Together, the two factors explain **58%** of the total variance.

Hypothesis Test

- The chi-square test for sufficiency of 2 factors yields:
  - $^2 = 20.65$, df $= 8$, p $= 0.00813$
  - Since p < 0.05, we reject the null hypothesis that 2 factors are sufficient.
  - This means that while 2 factors capture much of the structure, there is still meaningful unexplained variance — additional factors might be needed to fully represent the data.

Summary in Plain Words

- **Factor1 ("Academic Performance")**: Captures grades and failures, aligning with how well a student is doing in school.
- **Factor2 ("Parental Education")**: Captures family background in terms of parental education levels.
- Absences don't fit neatly into either factor, meaning student attendance behavior is not well-explained by these two latent traits.
- The test suggests 2 factors provide a reasonable summary, but they may not fully capture all patterns (especially attendance).

Check-in: If an exploratory factor analysis yields one factor with high loadings on G1, G2, G3 and a second factor with high loadings on Medu and Fedu, what real-world concepts could these factors represent?

Answer: The first factor is likely representing **overall academic performance** (since it ties together all the exam scores and related measures). The second factor represents **parental education background** (a socio-demographic factor related to the family). These names aren't given by the analysis; we infer them based on which variables cluster on each factor.

Rotation note: We used varimax rotation which is orthogonal (factors remain uncorrelated). There are other rotations (including oblique rotations that allow factors to correlate). If we allowed factors to correlate, we might find, for example, that academic performance and parental education factors are mildly correlated (which in reality they might be). But keeping it simple, varimax gave us a neat structure to interpret. The decision of rotation and number of factors is often guided by both statistical criteria and interpretability. It's common to test a few different counts of factors and see which yields meaningful and distinct factors. In our case, a 2-factor solution made sense. A 3-factor solution might have pulled out another factor (perhaps separating absences into its own "engagement" factor or something), but if that third factor is weak or hard to explain, one might prefer the more parsimonious 2-factor model.

One advantage of factor analysis is the clear *conceptual* interpretation if done well. For instance, identifying a factor as "academic ability" can be very useful in educational research or psychology (it connects to theory, can be used to create composite scores, etc.). PCA, by contrast, will always create components that are orthogonal and explain variance, but those components might be harder to interpret if the data doesn't have clear underlying constructs. In our example, both methods ended up telling a similar story, which is reassuring.

Finally, keep in mind factor analysis is exploratory here. If we had a hypothesis about factors (say we believe there are exactly two factors: academic and family background), we could also perform a **confirmatory factor analysis (CFA)** to statistically test that structure. CFA is more advanced and requires specifying the factor structure in advance, so it's beyond our scope here.

PCA vs Factor Analysis – When to Use Which?

Both PCA and EFA reduce dimensionality, but they are used in slightly different contexts:

- **PCA** is often used as a data preprocessing or summarization step. It's purely mathematical, with no assumption of an underlying causal model. Use PCA when you want to compress data, remove multicollinearity, or visualize high-dimensional data. For example, in machine learning pipelines, PCA can be used to reduce feature count while preserving variance.

- **Factor Analysis** is used when you suspect latent concepts are driving your data and you want to discover or confirm what those are. It's common in social sciences, psychology, marketing research, etc., where you might design a questionnaire and then use factor analysis to see which questions group together to form latent factors (e.g., a "satisfaction" factor composed of several survey items). Use EFA when your goal is to uncover hidden constructs and you have some theoretical reason to believe they exist.

In practice, if your goal is purely to reduce predictors in a predictive model, PCA is straightforward. If your goal is to **interpret** and name the underlying dimensions in the data, factor analysis provides a framework to do so (with rotation aiding interpretation). In our student case study, we care about understanding the data's structure (not just reducing for a model), so interpreting factors/components was a key part of our analysis.

Having learned the concepts and seen examples, it's time for you to get hands-on. In the next section, you'll find a lab activity with exercises to practice these techniques on the student performance dataset. Try to solve them on your own, and then you can check the provided solutions.

Section 3: Lab Activity

In this lab, you'll apply what we've covered to the student performance data. The dataset we use is from a study of secondary school students from two Portuguese schools, with attributes on student background and grades. We will use the "Math" class subset (395 students) for these exercises.

1. Data Understanding: Load the student performance dataset into R (you can use the code provided in the examples). Examine its structure using functions like `dim()`, `str()`, or `summary()`. How many observations and variables are there? Identify which variables are numerical and which are categorical. (Hint: Variables like `school`, `sex`, etc. are categorical; `age` is numeric; some, like `studytime`, are ordered factors coded as numeric.)

2. Exploring Relationships: Pick at least five numeric variables and compute a correlation matrix for them (for example, you might include the three grade columns `G1, G2, G3` and a couple of others like `studytime`, `absences`, `failures`). Which pair of variables has the highest correlation? Which pair has a strong negative correlation (if any)? Describe what those correlations imply in the context of student performance.

3. Visualizing Multivariate Patterns: Create a scatter plot to visualize the relationship between two interesting variables from the dataset. A good choice is **first period grade (`G1`) vs final grade (`G3`)** as we did above. Plot these using `ggplot2::geom_point()`. Do you notice any pattern or outliers? Now improve the plot by addressing overplotting: for example, use `geom_jitter()` or set `alpha` transparency as demonstrated. How does this help in seeing the data distribution? Briefly describe the trend you observe and any deviations.

(Optional extension: Color the points by a categorical variable such as `sex` (gender) or `schoolsup` (extra school support) to see if the relationship differs by subgroup.)

4. Principal Component Analysis: Perform PCA on a set of numerical variables related to student performance. A suggested set: `c("G1","G2","G3","failures","absences","Medu","Fedu")` (as we used in the example). Ensure you scale the variables (`prcomp(scale.=TRUE)`). How many principal components have eigenvalues greater than 1? Look at the proportion of variance explained by the first two PCs. Now inspect the PCA loadings (use `pca_result$rotation` or `print(pca_result)`). Based on the loadings, what does PC1 seem to represent? What about PC2? (In other words, which variables heavily influence PC1 vs PC2?)

5. Factor Analysis: Using the same set of variables as in #4, conduct an exploratory factor analysis. Try extracting 2 factors with `factanal(..., factors=2, rotation="varimax")`. Examine the factor loadings. Which variables load strongly on Factor1, and which on Factor2? Name each factor in plain language (e.g., "Factor1: _____ factor, Factor2: _____ factor"). How well does this factor solution make sense, and how does it compare to the PCA results from #4?

Write down your findings and interpretations for each step.