Dataset 3 - Glass Identification

Phan Thị Ngọc Linh - 22280052

2025-08-14

Dataset 3: Glass Identification

Dataset Description

The **Glass Identification** dataset is sourced from the UCI Machine Learning Repository. It contains chemical and physical information of glass samples. The goal is to predict the type of glass based on the given features. The dataset consists of 214 observations and 10 attributes (excluding the `Id` column), with the target variable being the glass type (7 discrete classes).

Attribute Information

Attribute Name	Description	Unit
`RI`	Refractive Index	-
`Na`	Sodium content (Na)	% oxide by mass
`Mg`	Magnesium content (Mg)	% oxide by mass
`Al`	Aluminum content (Al)	% oxide by mass
`Si`	Silicon content (Si)	% oxide by mass
`K`	Potassium content (K)	% oxide by mass
`Ca`	Calcium content (Ca)	% oxide by mass
`Ba`	Barium content (Ba)	% oxide by mass
`Fe`	Iron content (Fe)	% oxide by mass

Note: All chemical components are measured in **percentage by mass of the corresponding oxide**.

Label (class attribute) information:

`1`: Building windows (float processed)

- Glass produced by the float method (floating on molten metal), typically has high flatness and transparency.

2: Building windows (non-float processed)

- Glass not produced by the float method, usually lower quality and commonly used in construction.

3: Vehicle windows (float processed)

- Glass for vehicles produced by the float method, typically has high durability and impact resistance.

4: Vehicle windows (non-float processed) (no samples in this dataset)

5: Containers

- Glass used for bottles, containers, offers good strength and medium transparency.

6: Tableware

- Glass used for cups, plates, typically with high durability and aesthetics.

7: Headlamps

- Glass used in vehicle headlights, capable of withstanding high heat and intense light.

Source: https://www.kaggle.com/datasets/uciml/glass

Analysis Objectives

The main aim is to **analyze the relationship** between chemical and physical components of various glass types, thereby supporting classification according to usage, and providing a scientific basis for optimizing production formulas suited to specific requirements regarding physical, optical, or mechanical properties.

Specifically, my analysis focuses on:

1. Understanding the chemical and physical characteristics of glass samples through **descriptive statistics**.
2. **Dimensionality reduction** and **extracting statistically significant latent factors** that strongly affect glass properties, using Principal Component Analysis (PCA).
3. **Classifying** glass types based on similarity in chemical composition.
4. Supporting manufacturers in **optimizing formulas and production processes** based on physical and chemical properties for specific uses such as construction, optics, headlamps, tableware, etc.

Data Exploration

Load data

```
dat_glass <- read.csv("datasets/glass.csv", header = TRUE)
head(dat_glass)
```

```
##        RI    Na   Mg   Al    Si    K   Ca Ba   Fe Type
## 1 1.52101 13.64 4.49 1.10 71.78 0.06 8.75  0 0.00    1
## 2 1.51761 13.89 3.60 1.36 72.73 0.48 7.83  0 0.00    1
## 3 1.51618 13.53 3.55 1.54 72.99 0.39 7.78  0 0.00    1
## 4 1.51766 13.21 3.69 1.29 72.61 0.57 8.22  0 0.00    1
## 5 1.51742 13.27 3.62 1.24 73.08 0.55 8.07  0 0.00    1
## 6 1.51596 12.79 3.61 1.62 72.97 0.64 8.07  0 0.26    1
```

```
dim(dat_glass)
```

```
## [1] 214  10
```

Check missing value

```
# Count the number of missing values per column
colSums(is.na(dat_glass))
```

```
##   RI   Na   Mg   Al   Si    K   Ca   Ba   Fe Type
##    0    0    0    0    0    0    0    0    0    0
```

Comment: The dataset contains no missing values.

Descriptive statistics

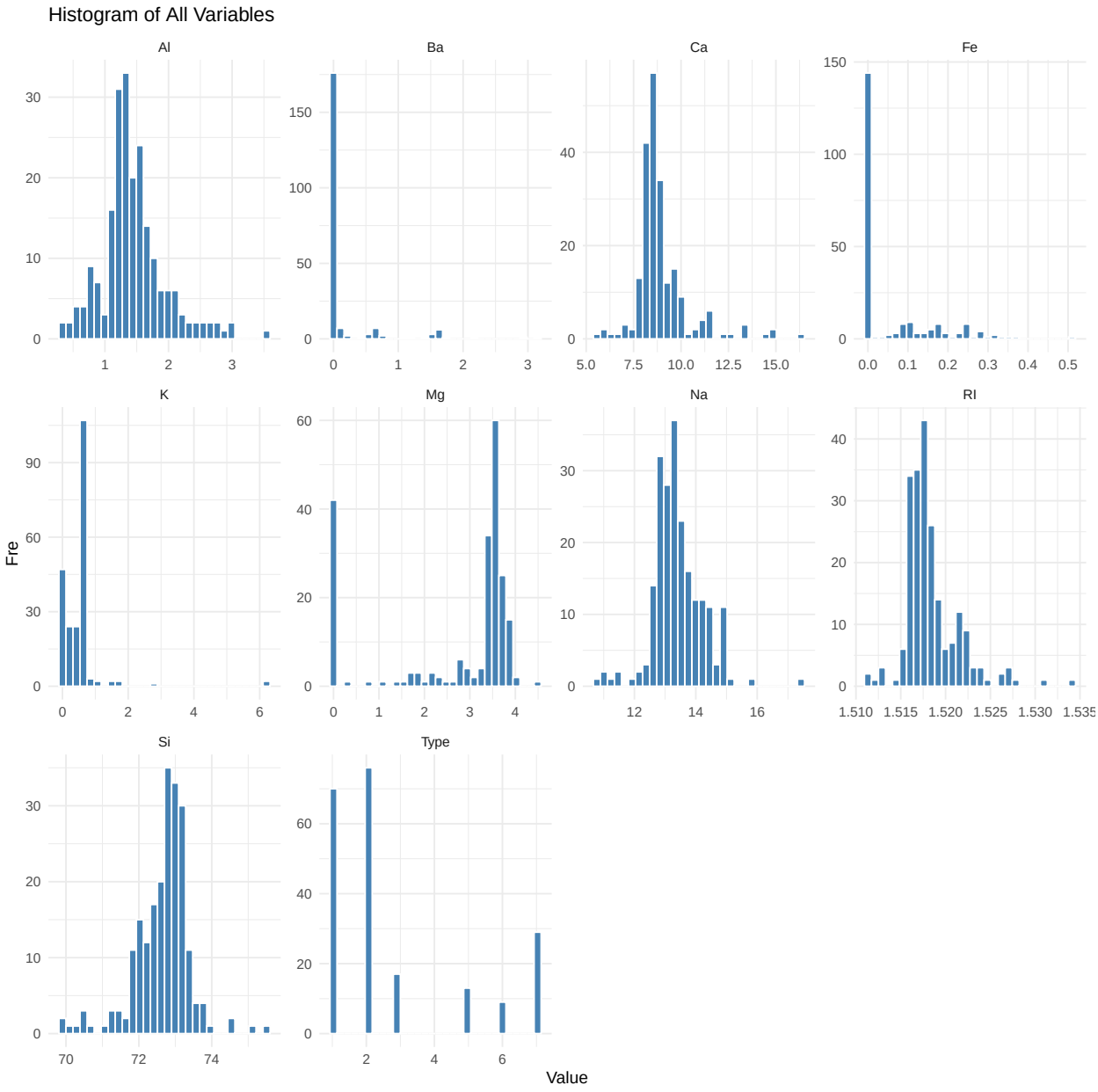
```
summary(dat_glass)
```

```
##        RI              Na              Mg              Al
##  Min.   :1.511   Min.   :10.73   Min.   :0.000   Min.   :0.290
##  1st Qu.:1.517   1st Qu.:12.91   1st Qu.:2.115   1st Qu.:1.190
##  Median :1.518   Median :13.30   Median :3.480   Median :1.360
##  Mean   :1.518   Mean   :13.41   Mean   :2.685   Mean   :1.445
##  3rd Qu.:1.519   3rd Qu.:13.82   3rd Qu.:3.600   3rd Qu.:1.630
##  Max.   :1.534   Max.   :17.38   Max.   :4.490   Max.   :3.500
##        Si              K               Ca              Ba
```

```
##  Min.   :69.81   Min.   :0.0000   Min.   : 5.430   Min.   :0.000
##  1st Qu.:72.28   1st Qu.:0.1225   1st Qu.: 8.240   1st Qu.:0.000
##  Median :72.79   Median :0.5550   Median : 8.600   Median :0.000
##  Mean   :72.65   Mean   :0.4971   Mean   : 8.957   Mean   :0.175
##  3rd Qu.:73.09   3rd Qu.:0.6100   3rd Qu.: 9.172   3rd Qu.:0.000
##  Max.   :75.41   Max.   :6.2100   Max.   :16.190   Max.   :3.150
##        Fe               Type
##  Min.   :0.00000   Min.   :1.00
##  1st Qu.:0.00000   1st Qu.:1.00
##  Median :0.00000   Median :2.00
##  Mean   :0.05701   Mean   :2.78
##  3rd Qu.:0.10000   3rd Qu.:3.00
##  Max.   :0.51000   Max.   :7.00
```

Comment: The percentage of oxides in `Ba` and `Fe` is almost zero. These may be important factors for glass identification.

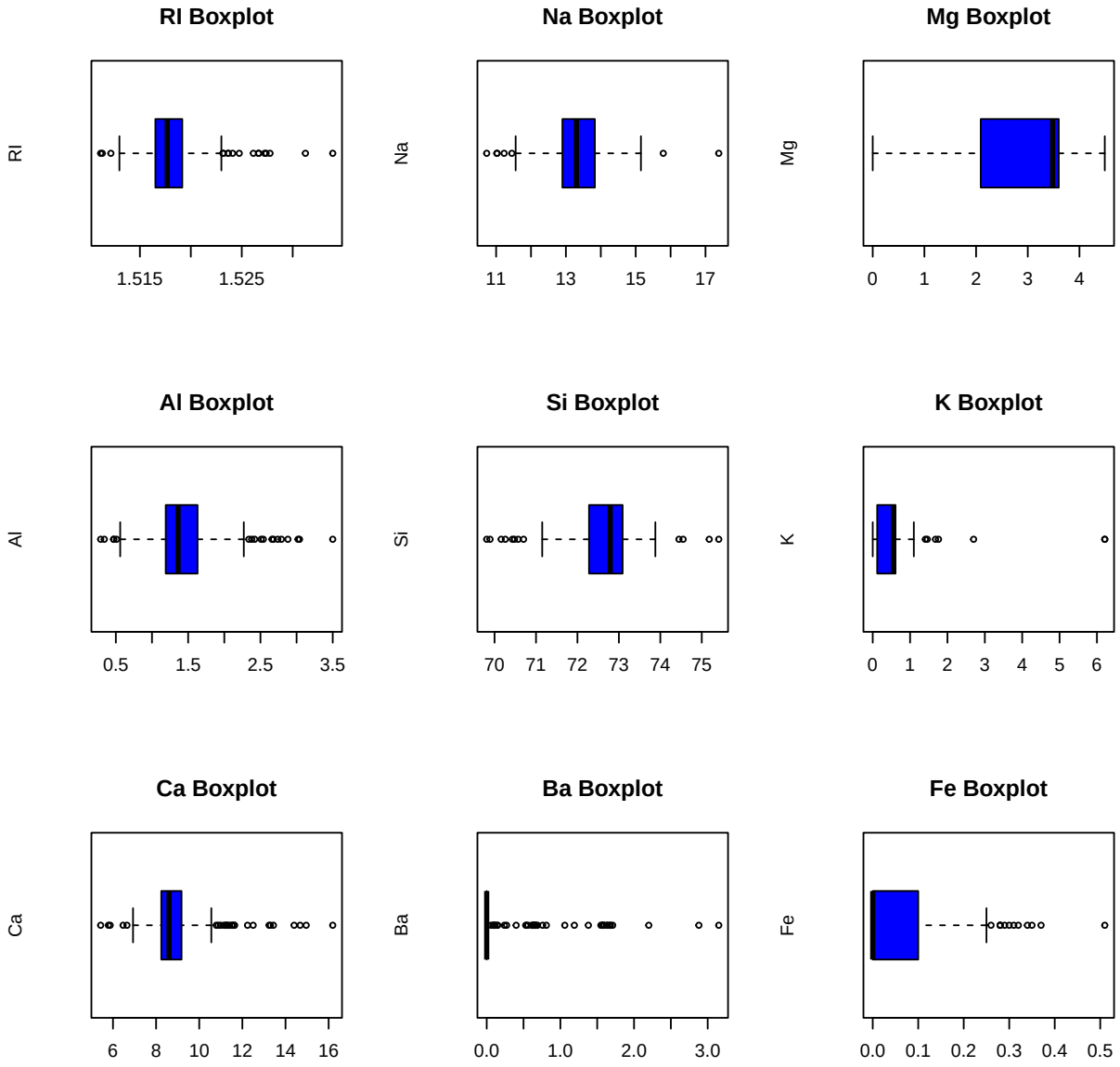
Histogram of variable distributions



Comments

- Variables `RI`, `Na`, `Al`, `Ca`, `Si` are nearly normally distributed.
- `Ba` and `Fe` have many zero values.
- Type (glass class): 6 discrete values, clear classification. Some types (e.g., 1, 2) appear frequently, type 4 does not exist in the dataset. The data is unbalanced across groups.

Checking for outliers



General comments

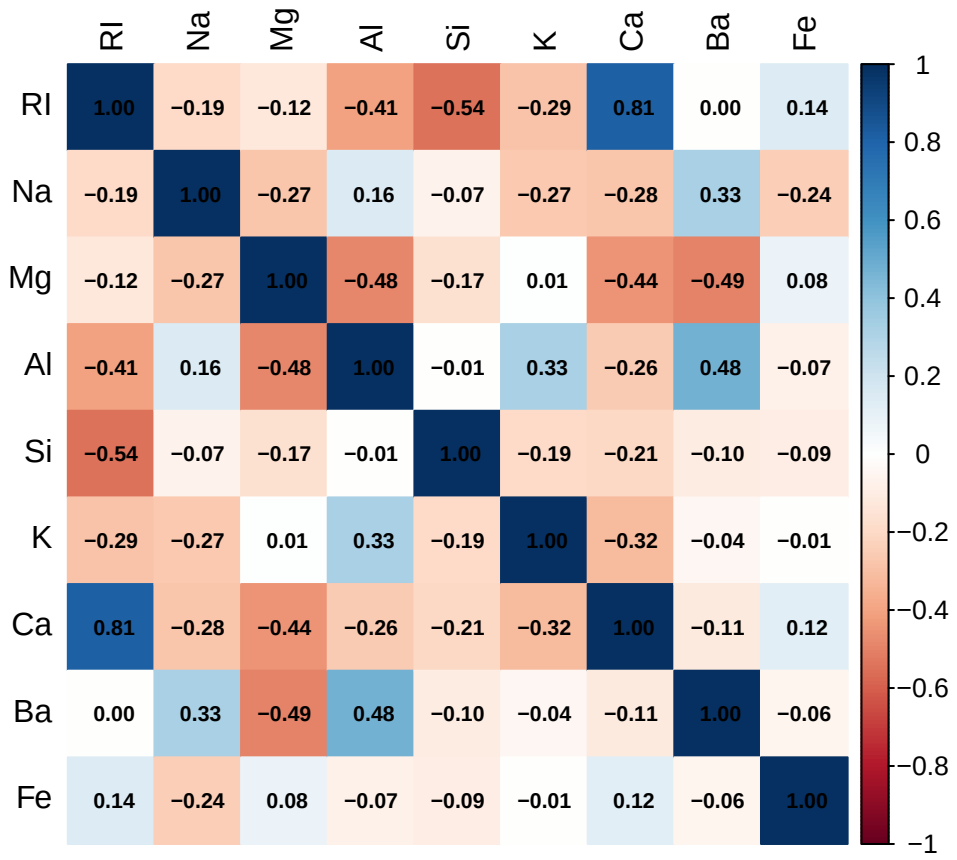
Chemical variables like `RI`, `Na`, `Al`, `Si`, `Ca` have fairly even distributions, centered around the median with few outliers. Conversely, variables like `K`, `Ba`, `Fe`, `Mg` display many outliers and skewed distributions, indicating significant sample diversity. Some substances are present only in certain **glass types**. These characteristics are important for distinguishing groups during clustering or classification.

Remove type column from data

```r
data_glass <- dat_glass |> select(-Type)
head(data_glass)
```

```
##        RI    Na   Mg   Al    Si    K   Ca Ba   Fe
## 1 1.52101 13.64 4.49 1.10 71.78 0.06 8.75  0 0.00
## 2 1.51761 13.89 3.60 1.36 72.73 0.48 7.83  0 0.00
## 3 1.51618 13.53 3.55 1.54 72.99 0.39 7.78  0 0.00
## 4 1.51766 13.21 3.69 1.29 72.61 0.57 8.22  0 0.00
## 5 1.51742 13.27 3.62 1.24 73.08 0.55 8.07  0 0.00
## 6 1.51596 12.79 3.61 1.62 72.97 0.64 8.07  0 0.26
```

```r
R <- cor(data_glass)
corrplot(R, method = "color", tl.col = "black",
         addCoef.col = "black", number.cex = 0.7)
```

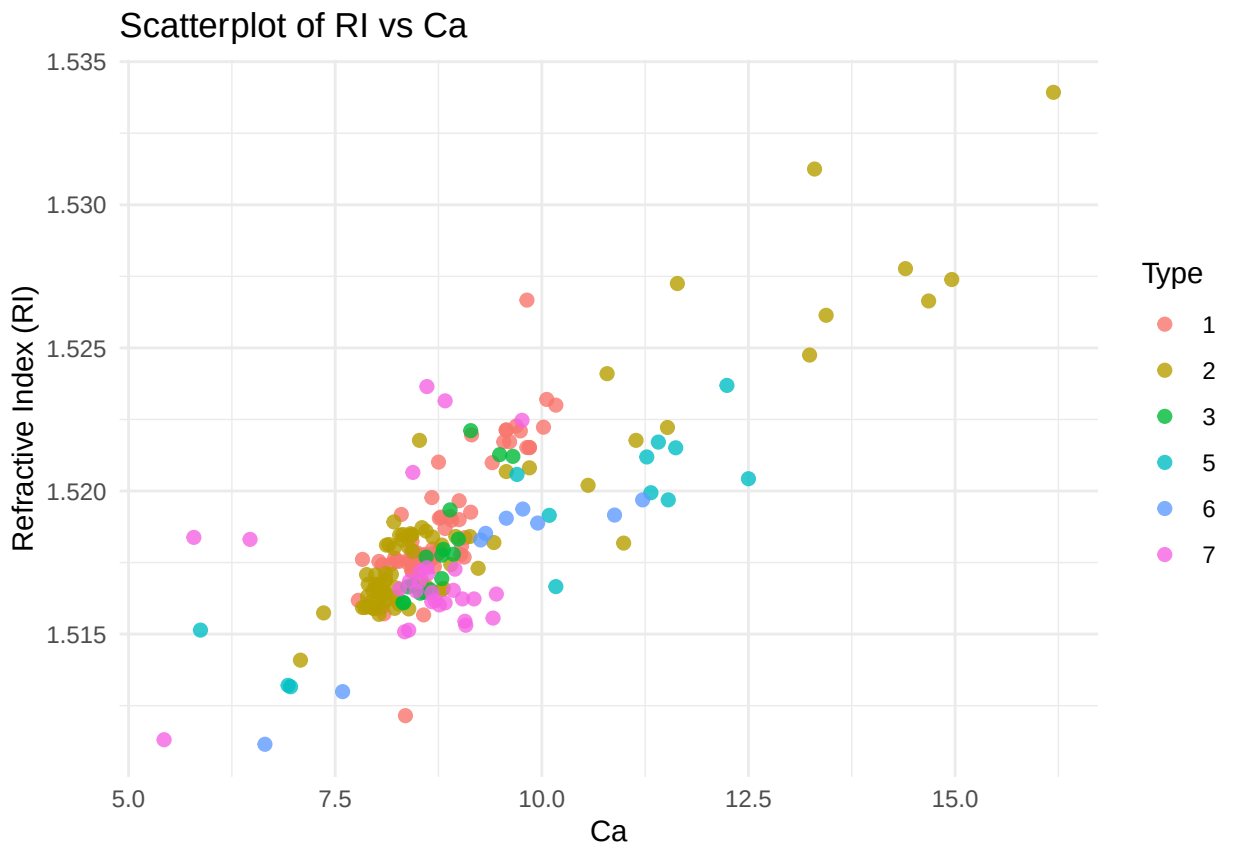


Comments

The correlation matrix shows linear relationships between pairs of attributes in the `Glass Identification` dataset. `Ca` and `RI` have the strongest correlation, at **0.81**.

Let's draw a scatter plot to examine the relationship between `RI` and `Ca`:

Scatter plot of Ca versus RI



Observation: `Ca` and `RI` have a positive linear relationship. When Calcium content increases, so does the refractive index.

Chemistry aspect: Calcium is a key component in glass, especially soda-lime glass. Calcium content affects durability and transparency. The refractive index (`RI`) reflects how the glass bends light, and `Ca` plays a major role in this property.

Practical application: Glass with high calcium content and refractive index is often used in applications requiring high durability or heat resistance, such as building or automotive glass.

From the plot: High calcium and high refractive index glass samples tend to belong to Type 2 (building windows, non-float processed).

However, other chemical variables may also be closely related or jointly influence glass properties (as seen in the correlation plot), so examining individual variables is not sufficient to fully understand their relationships. Using too many variables can be noisy and difficult to visualize. Therefore, I apply Principal Component Analysis (PCA) to reduce dimensionality while retaining the most important information.

Principal Component Analysis (PCA)

*Note: **`dat_glass`** (with **`type`** column), **`data_glass`** (type column removed)*

As variable scales differ greatly, data normalization is performed.

```
data_glass_scale <- scale(data_glass)
```

Use `princom()` for PCA on normalized data

```
pc.data_glass <- princomp(data_glass_scale)
summary(pc.data_glass,loadings=TRUE)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## Standard deviation     1.5809584 1.4284581 1.1824886 1.0735231 0.9537983
## Proportion of Variance 0.2790182 0.2277858 0.1560938 0.1286514 0.1015558
## Cumulative Proportion  0.2790182 0.5068040 0.6628978 0.7915492 0.8931050
##                            Comp.6     Comp.7      Comp.8       Comp.9
## Standard deviation     0.72468587 0.60599863 0.252100316 0.0400162465
## Proportion of Variance 0.05862613 0.04099538 0.007094772 0.0001787575
## Cumulative Proportion  0.95173109 0.99272647 0.999821242 1.0000000000
##
## Loadings:
##    Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## RI  0.545  0.286         0.147         0.115         0.752
## Na -0.258  0.270 -0.385  0.491 -0.154 -0.558  0.149  0.128 -0.312
## Mg  0.111 -0.594         0.379 -0.124  0.308 -0.206        -0.577
## Al -0.429  0.295  0.329 -0.138                -0.699  0.274 -0.192
## Si -0.229 -0.155 -0.459 -0.653                 0.216  0.380 -0.298
## K  -0.219 -0.154  0.663         0.307 -0.244  0.504  0.110 -0.261
## Ca  0.492  0.345        -0.276  0.188 -0.149        -0.399 -0.579
## Ba -0.250  0.485         0.133 -0.251  0.657  0.352 -0.145 -0.198
## Fe  0.186         0.284 -0.230 -0.873 -0.243
```

Observations

Comp.1: Three prominent factors: `RI` (0.545), `Ca` (0.492) positively, and `Al` (-0.429) negatively. This indicates refractive index and calcium content together enhance glass optical and durability properties, while aluminum has an opposite effect. In practice, calcium improves strength and optical transparency; aluminum is added for mechanical strength and corrosion resistance, but excess aluminum may reduce refractive index and optical properties. **Comp.1** reflects the balance between optical (`RI`, `Ca`) and mechanical (`Al`) properties, crucial for distinguishing construction glass (which prioritizes clarity and durability) from other types.

Comp.2: `Mg` (-0.594), and `Ba` (0.485) are key. These have opposing effects. `Mg` increases durability and heat resistance, often reducing brightness (strong glass is usually less bright). `Ba` enhances brightness and optical properties but may reduce strength. Notably, `Mg` and `Ba`

are important for special glass types like safety or insulating glass, balancing mechanical and optical/aesthetic properties. This opposition helps classify glass according to use: durable glass favors Mg, aesthetic glass favors Ba.

Comp.3: `K` (0.663) and `Si` (-0.459) have opposite signs, showing the relationship between flexibility (potassium) and durability (silicon). Potassium increases flexibility but reduces hardness; silicon increases durability and especially **optical transparency**. These compete in defining glass properties. For high flexibility and easy processing, potassium is increased, but this can lower overall strength and transparency (by reducing silicon's role). Examples:

- Optical glass: May need high potassium for improved physical properties, possible lower silicon.
- Construction/durable glass: Higher silicon for strength and transparency, potassium plays a minor role.

Comp.4: `Na` (0.491), `Si` (-0.653), `Mg` (0.379) are notable. `Na` lowers melting point, reducing energy costs; `Si` enhances strength and transparency; `Mg` increases durability. This component balances processability (`Na`) and strength (`Si`, `Mg`), suitable for **tableware, bottles, containers** needing both manufacturability and durability.

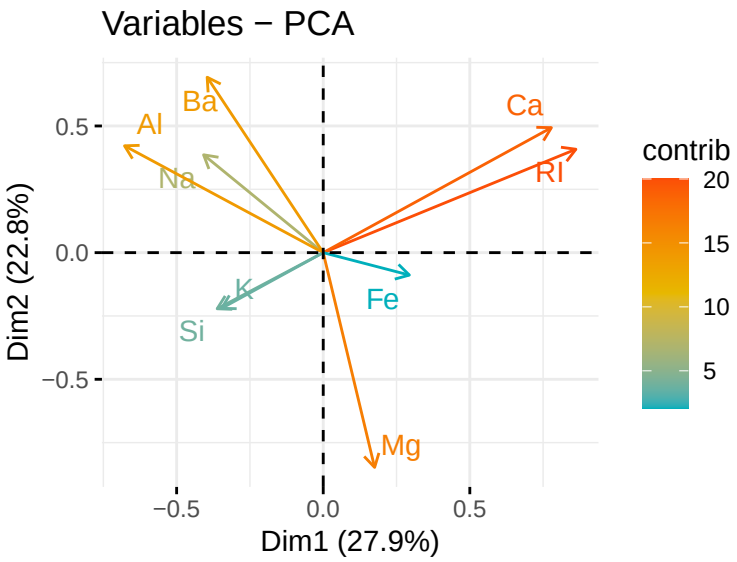
Comp.5: `Fe` (-0.873) is dominant, nearly alone in representing iron variability. `Fe` strongly affects color and heat resistance, and contributes to UV filtering, which is crucial for colored glass, headlamp glass, or UV-protective glass.

Other **PCs** from **Comp.6** onward mainly reflect minor variations or weak relationships among chemical components, with no clearly important latent factors.

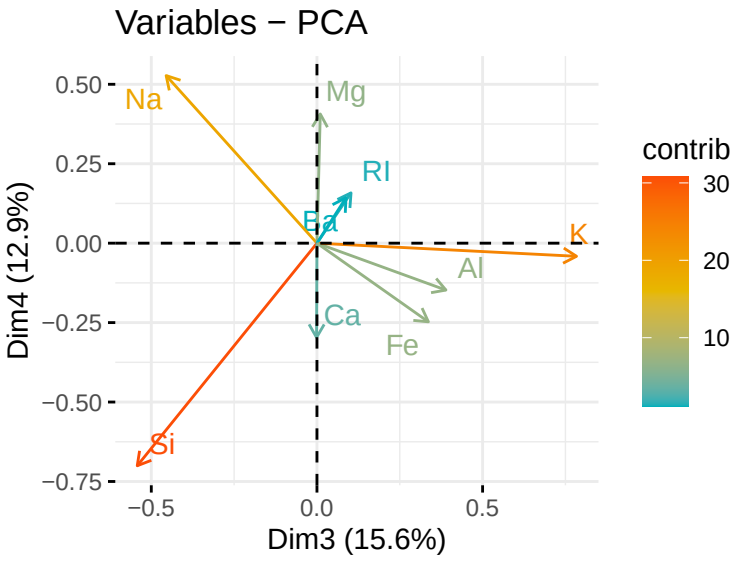
Thus, the PCs show strong relationships and oppositions among key chemical components in glass. Adjusting these proportions determines not just physical and chemical properties but also allows for optimizing glass for specific applications: construction glass (needs high `RI`, `Ca`, `Si`), optical or decorative glass (prefers `Ba`, `K`), durable glass (more `Mg`, `Si`), colored/headlamp glass (`Fe`), and tableware (balance among `Na`, `Mg`, `Si`).

From PCA, manufacturers can identify which factors to prioritize in glass formulation, optimizing processes and costs while meeting technical standards for specialized glass products.

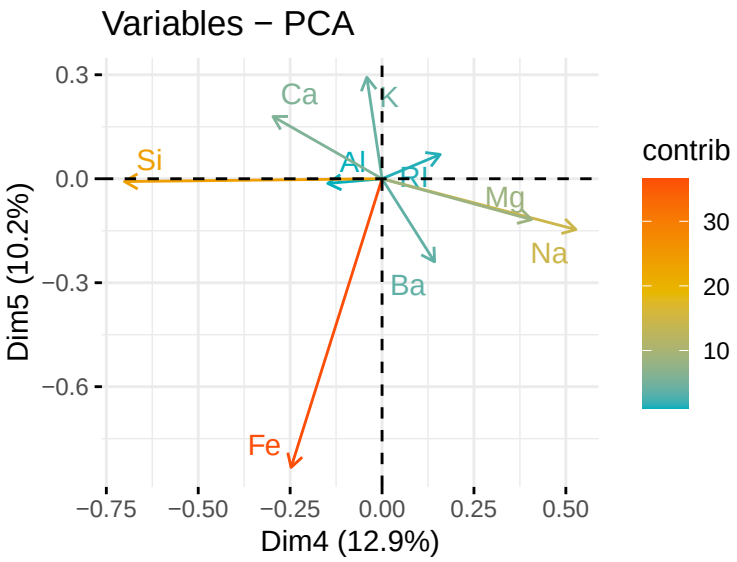
Biplot for PC1 and PC2



Biplot for PC3 and PC4



Biplot for PC4 and PC5



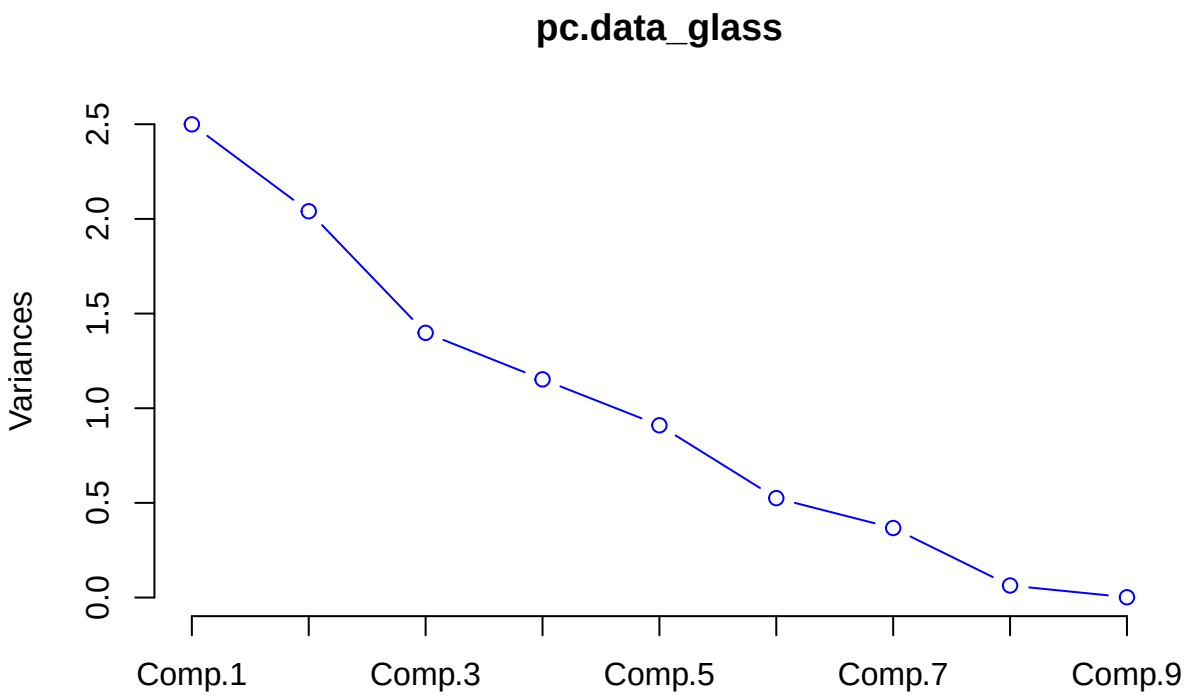
Comments:

Based on the loading analysis and contribution plots, the roles and relationships between chemical and physical components for each principal component are clear.

- In the first **Dim1-Dim2** plot, `RI` and `Ca` have long vectors and deep orange color, indicating major contribution to variance, showing their importance for **optical and mechanical properties**.
- In the **Dim3-Dim4** plot, `K` stands out, while `Si` is in an opposite direction. These factors are key for distinguishing glass types by intended use (tableware, durability, decorative…).
- In the **Dim4-Dim5** plot, `Fe` is dominant on the `Dim5` axis, crucial for color and light filtering. This highlights the need to retain the fifth principal component to capture important latent features.

These visualizations clearly illustrate the contributions and properties of `RI`, `Ca`, `Mg`, `K`, `Na`, `Si`, and `Fe` (key determinants of material characteristics and glass type classification).

Scree plot



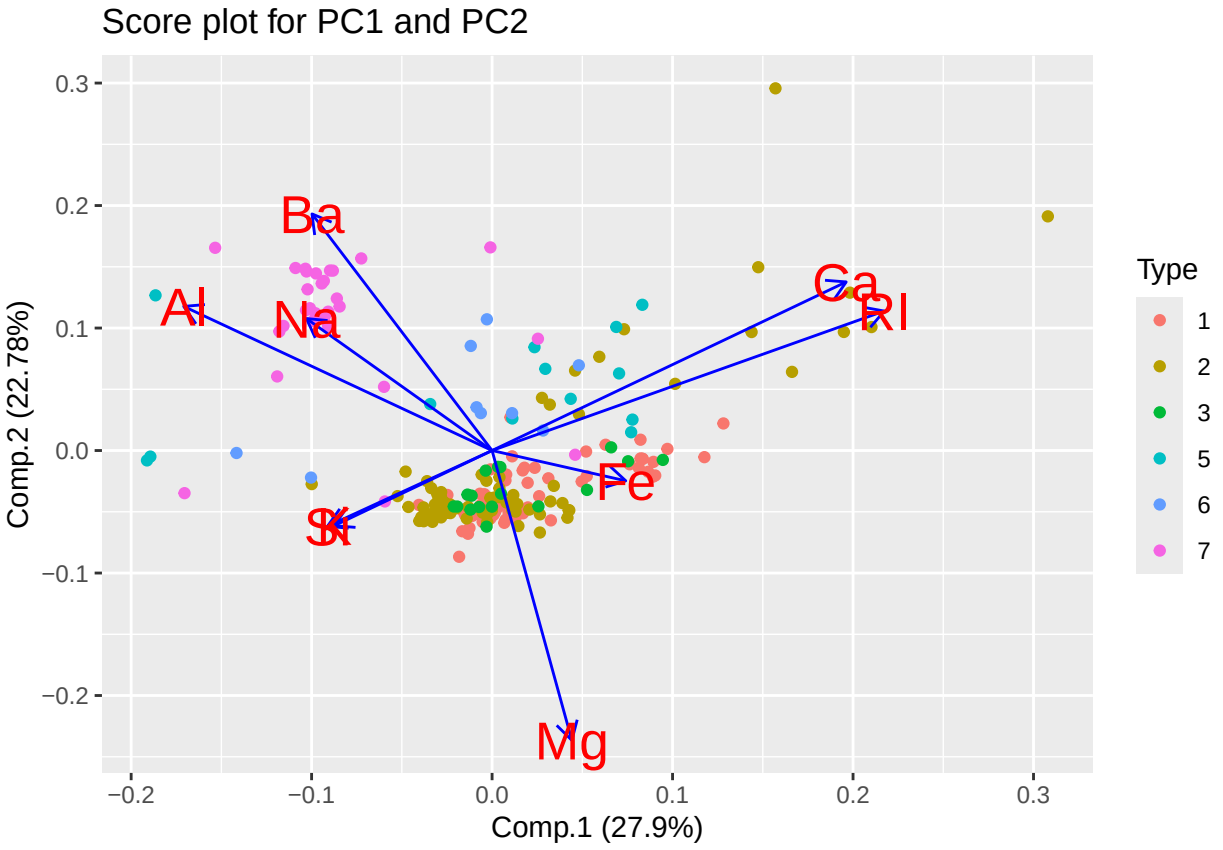
Choosing number of principal components to retain

Selecting the number of dimensions in PCA should consider both the elbow plot and the practical significance of components. The elbow plot shows clear bends at **Comp.3**, **Comp.5**, **Comp.6**, and **Comp.7**, with the first five components explaining about **89.3%** of total variance. Besides variance, consider the **latent value** of each component. **Comp.5** is notable for its high loading of `Fe` (-0.873), which is crucial for color and light filtering. Retaining **Comp.5** ensures important latent features related to Fe are not lost, aiding effective exploration of special glass types like colored glass, headlamp glass, and those needing special optical properties.

In summary, retaining five principal components preserves nearly 90% of information and helps discover and analyze valuable latent features that later components may overlook if considering only variance.

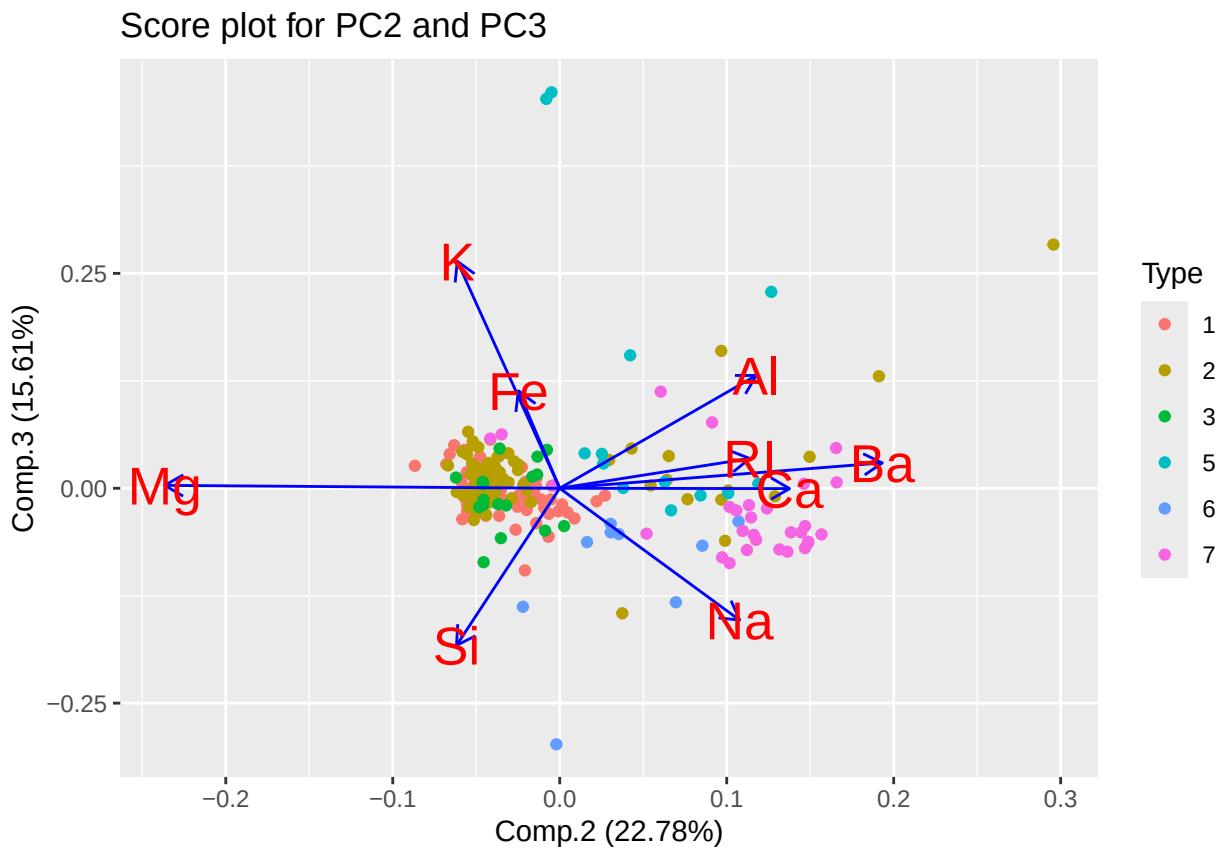
Next, plot the distribution of data on different PCs to observe clustering or interesting distribution features.

Plot data distribution on PCA components



Comments

- **Class 7 (heatlamps)**: Shows clear separation, located on the left of the plot, distinct from other classes. This indicates unique chemical properties, especially `Al`, `Ba`, and `Na`.
- **Other classes** `1 (building_windows_float_processed)`, `2 (building_windows_non_float_pro` `3 (vehicle_windows_float_processed)`, `5 (containers)`, `6 (tableware)`: Fairly overlapping distributions, with no clear clustering. These glass types have similar chemical and optical characteristics.



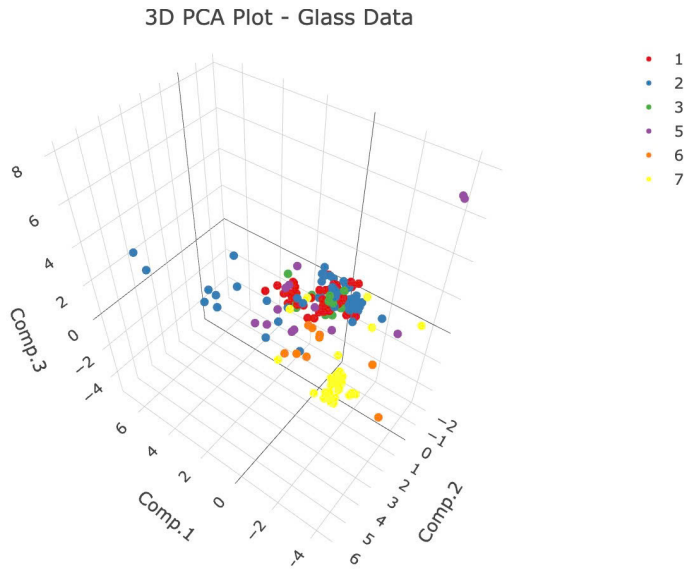
Comments:

- **Class 7 (heatlamps)** continues to be clearly separated, located in the lower right of the plot, meaning PCs 2 and 3 also reflect its distinctiveness.
- **Other classes:** Still heavily overlapping, especially among `1 (building_windows_float_processed)`, `2 (building_windows_non_float_processed)`, `3 (vehicle_windows_float_processed)`, `5 (containers)`, `6 (tableware)`, indicating PCs 2 and 3 do not help distinguish these classes well.

Since clustering is unclear, try visualizing in three dimensions.

3D plot on Comp.1, Comp.2, Comp.3

```
dat_glass_3D <- as.data.frame(pc.data_glass$scores)
dat_glass_3D$Type <- as.factor(dat_glass$Type)
```



From the plot:

- **Class 7 (heatlamps):** In 3D space, class 7 remains clearly separate, confirming its chemical distinctiveness.
- Other classes such as **`1 (building_windows_float_processed)`**, **`2 (building_windows_non_flo`**
  `3 (vehicle_windows_float_processed)`, **`5 (containers)`**, **`6 (tableware)`** still lack clear clustering, suggesting these glass types have highly similar chemical properties.

Conclusion: PCA captures various aspects of the dataset, but specific clustering is not evident. Therefore, factor analysis is applied for deeper understanding of latent structure and variable relationships.

Factor Analysis (FA)

Factor analysis with 5 factors

```
# Phân tích Factor với 2 nhân tố
glass.ft <- factanal(covmat = R, factors = 5,
                     rotation = "none", n.obs = 214)
glass.ft
```

```
##
## Call:
## factanal(factors = 5, covmat = R, n.obs = 214, rotation = "none")
##
```

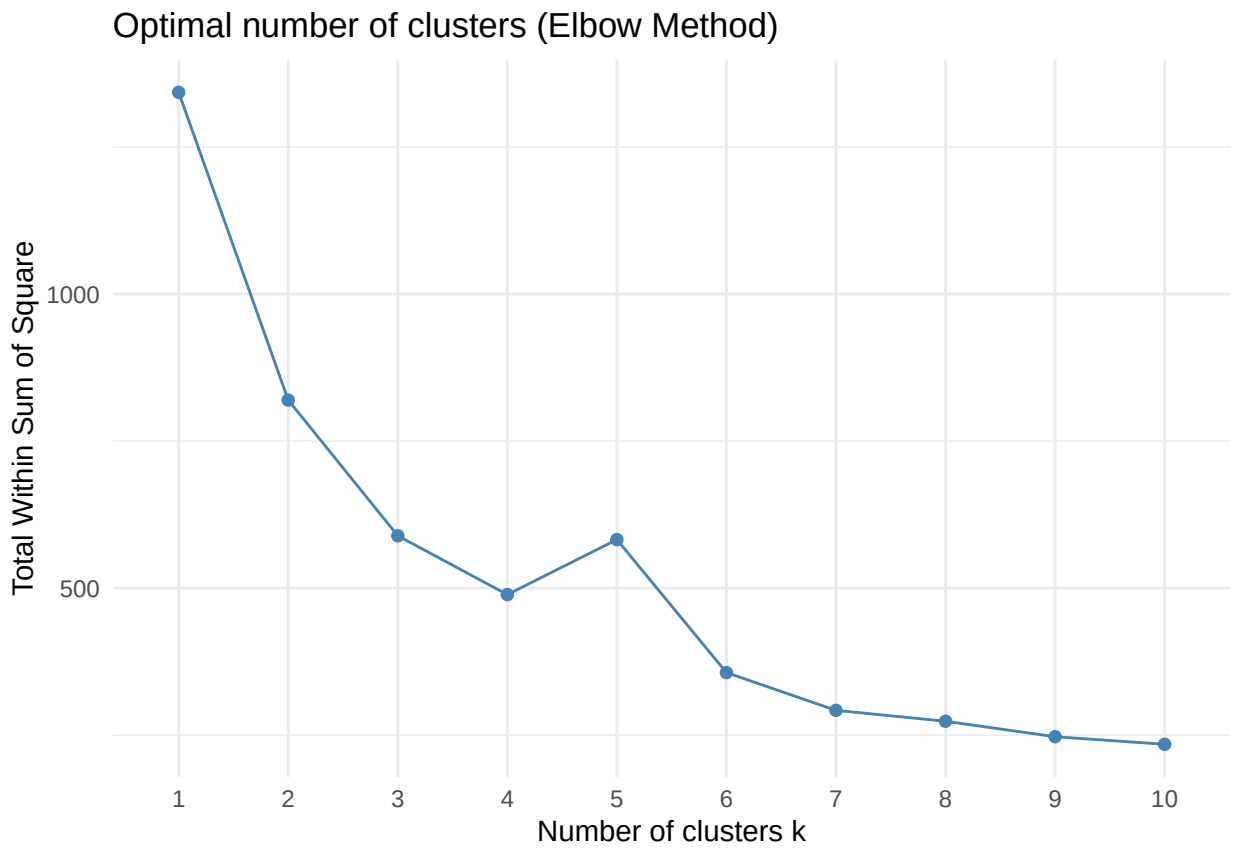
```
## Uniquenesses:
##    RI    Na    Mg    Al    Si     K    Ca    Ba    Fe
## 0.115 0.005 0.005 0.302 0.005 0.184 0.005 0.005 0.926
##
## Loadings:
##    Factor1 Factor2 Factor3 Factor4 Factor5
## RI  0.759          -0.350   0.419
## Na -0.518   0.599           0.215  -0.559
## Mg -0.351  -0.844  -0.239   0.314
## Al -0.188   0.483   0.583           0.293
## Si -0.242          -0.125  -0.959
## K          -0.142   0.823   0.110   0.313
## Ca  0.950   0.169  -0.183          -0.177
## Ba -0.192   0.799  -0.151   0.137   0.527
## Fe  0.179  -0.137                   0.142
##
##                Factor1 Factor2 Factor3 Factor4 Factor5
## SS loadings      2.041   2.018   1.278   1.276   0.835
## Proportion Var   0.227   0.224   0.142   0.142   0.093
## Cumulative Var   0.227   0.451   0.593   0.735   0.828
##
## Test of the hypothesis that 5 factors are sufficient.
## The chi square statistic is 472.85 on 1 degree of freedom.
## The p-value is 7.69e-105
```

After trying factor analysis with 2 to 5 factors, no reliable results are obtained: `p_value` is always very small.

To clarify grouping in the data, the next step is to use **K-means clustering** to visualize separation among data groups.

Kmeans clustering

```r
set.seed(123)
fviz_nbclust(data_glass, kmeans, method = "wss") +
  theme_minimal() +
  ggtitle("Optimal number of clusters (Elbow Method)")
```

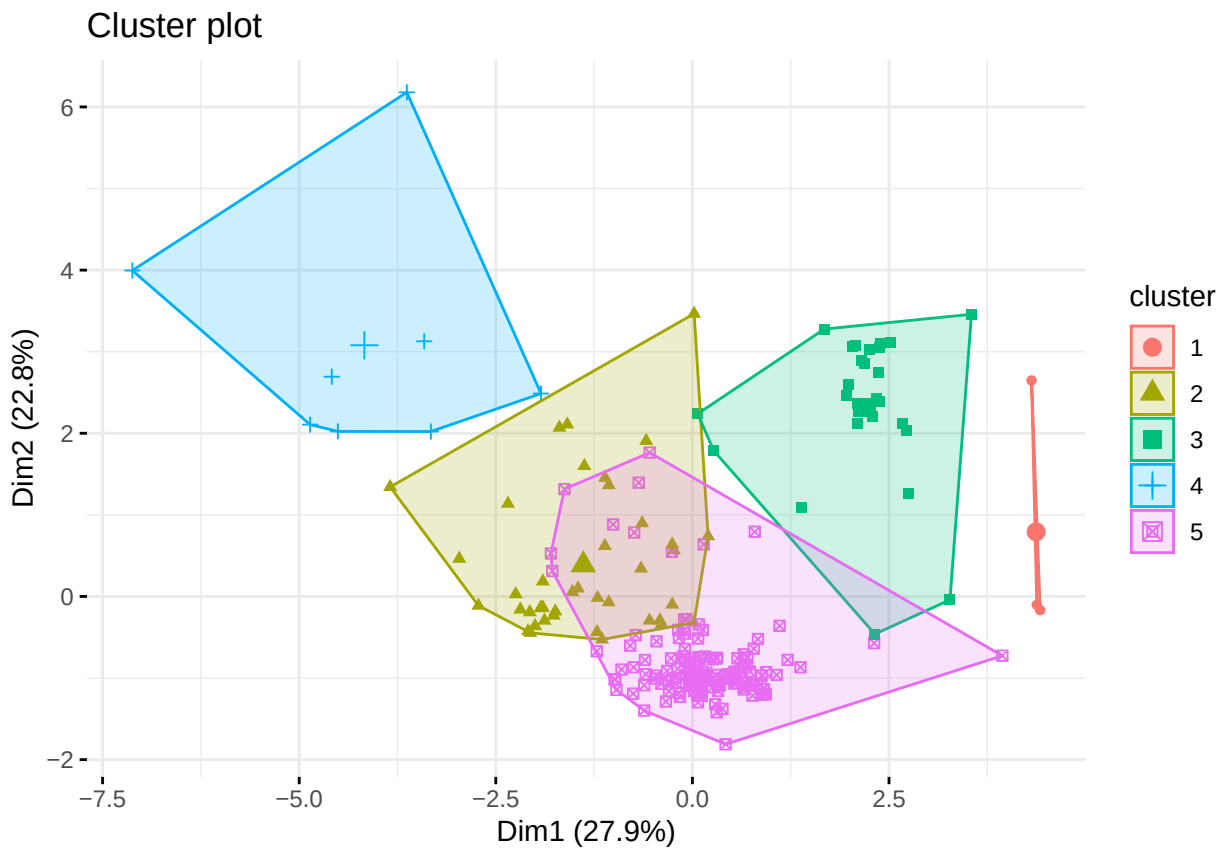


Comment: The plot suggests choosing `k = 5`.

```
pca_data <- pc.data_glass$scores[, 1:5]
```

```
# K-means clustering
km.res <- kmeans(pca_data, centers = 5, nstart = 25)

fviz_cluster(km.res, data = data_glass,
             geom = "point",
             ellipse.type = "convex",
             repel = TRUE,
             ggtheme = theme_minimal(),
             axes = c(1, 2),
             main = "Cluster plot")
```



Based on clustering plots, the following comments are made:

- **Cluster 4** (blue) and **Cluster 1** (red): In PC1 and PC2 space, these clusters are clearly separated. Data points in **Cluster 4** and **Cluster 1** occupy distinct regions, not overlapping with other clusters. Glass types in these clusters have unique chemical or optical properties.
- **Cluster 2** (yellow triangle), **Cluster 3** (green square), and **Cluster 5** (pink square): These three clusters overlap significantly in the center of the plot (PC1, PC2). Data points are intermingled, making distinction difficult. This overlap may result from similar chemical or optical properties among these glass types.

Testing other PC pairs (PC2-PC3, PC3-PC4, PC4-PC5): Results indicate clustering on these PC pairs is not promising. Clusters overlap heavily, with no clear separation. Thus, only PC1 and PC2 space is optimal for analyzing cluster differences.

Conclusion

- PC1 and PC2 space is best for clustering, allowing clear separation of **Cluster 4** and **Cluster 1**.
- Other clusters (**Cluster 2, Cluster 3, Cluster 5**) require further analysis or additional data for improved differentiation.

Summary

In this analysis, I examined the **Glass Identification** dataset to explore the relationships between chemical composition and glass type, providing support for manufacturers in optimizing formulas and production processes based on physical and chemical characteristics. The main steps included:

1. **Descriptive analysis** of physical and chemical features like `RI`, `Ca`, `Mg`, `Si`, `Fe`, etc. to assess the roles and influence of each component on glass properties.

2. **Principal Component Analysis (PCA)** was used for **dimensionality reduction** and **extracting statistically significant latent factors**. The first principal components (PC1–PC5) explained most data variance and highlighted contrasting dimensions in glass chemistry. These principal components revealed differences among factors shaping optical properties, mechanical strength, processing flexibility, and UV filtering or coloration. These key axes showed how features like `RI`, `Ca`, `Mg`, `Ba`, `Si`, `Fe`… cooperate or oppose to define the properties and applications of each glass type.

   This helps identify which factors to prioritize or adjust depending on intended use: construction glass, optical glass, headlamps, or tableware.

   - **Class 1, 2 (building_windows_float_processed, building_windows_non_float_pro** Construction glass should prioritize strength and heat resistance.
     - Related variables: `Mg`, `Al`, `Ca`, `Ba` (increase strength, heat resistance), `Si` (increase hardness, thermal stability).
   - **Class 3 (vehicle_windows_float_processed):** Vehicle glass focuses on safety and impact resistance.
     - Related variables: `Ca`, `Mg`, `Si` (increase robustness and resistance), `Na` (maintains stable chemical-physical properties).
   - **Class 7 (headlamp glass):** Headlamp glass requires optimal refractive index, high transparency, and heat resistance.
     - Related variables: `RI` (high refractive index for light transmission), `Fe`, `Si` (low content for clarity), `Ba`, `Al` (increase heat resistance).

3. I also tried **factor analysis (FA)** to extract latent factors but results were unclear, so did not pursue this direction.

4. Finally, I used **K-means** to cluster data in **PCA space**. Results showed some glass groups have distinct features, but some clusters overlap due to similar chemical composition.

In conclusion:

Through these analytical steps, I gained insight into how chemical components interact to create distinct glass properties. This not only supports classification but can also guide manufacturers in creating glass suited for specific applications, optimizing both costs and production processes.