PCA Case Study - Life Expectancy and Starbucks Satisfaction

Grace Bianchi

2025-11-21

Load Libraries

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(factoextra)
```

```
## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```r
library(FactoMineR)
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```r
library(ggplot2)
```

Analysis of Life Expectancy

Context:

This dataset contains indicators related to global health outcomes across countries. The data describes the association between each of the 183 countries and each of the 9 health variables. PCA is applied to identify underlying dimensions of health and development by reducing the large number of correlated variables into a few principal components.

Variable Description:

Life Expectancy: Life Expectancy in age

Adult Mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)

Infant Deaths: Number of Infant Deaths per 1000 population

Percentage Expenditure: Expenditure on health as a percentage of Gross Domestic Product per capita (%)

Measles: Measles - number of reported cases per 1000 population

Under-Five Deaths: Number of under-five deaths per 1000 population

Polio: Polio (Pol3) immunization coverage among 1-year-olds (%)

Diphtheria: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)

HIV/AIDS: Deaths per 1000 live births HIV/AIDS (0-4 years)

Reading and Preparing the Data

```
#Reading the data
life_expectance_data <- read.csv("Life Expectancy Data (1).csv")

#Select data from 2015
life_expectance_data <- filter(life_expectance_data, Year == 2015)

#Keep only numeric columns
life_expectance_data <- select_if(life_expectance_data, is.numeric)

#Remove columns with NA values and Zero Variance
life_expectance_data  <- life_expectance_data  %>% select_if(~ !any(is.na(.)))
life_expectance_data <- life_expectance_data %>% select_if(~ var(.) > 0)

dim(life_expectance_data)
```

```
## [1] 183   9
```

```
str(life_expectance_data)
```

```
## 'data.frame':    183 obs. of  9 variables:
##  $ Life.expectancy       : num  65 77.8 75.6 52.4 76.4 76.3 74.8 82.8 81.5 72.7 ...
##  $ Adult.Mortality       : int  263 74 19 335 13 116 118 59 65 118 ...
##  $ infant.deaths         : int  62 0 21 66 0 8 1 1 0 5 ...
##  $ percentage.expenditure: num  71.3 365 0 0 0 ...
##  $ Measles               : int  1154 0 63 118 0 0 33 74 309 0 ...
##  $ under.five.deaths     : int  83 0 24 98 0 9 1 1 0 6 ...
##  $ Polio                 : int  6 99 95 7 86 93 96 93 93 98 ...
##  $ Diphtheria            : int  65 99 95 64 99 94 94 93 93 96 ...
##  $ HIV.AIDS              : num  0.1 0.1 0.1 1.9 0.2 0.1 0.1 0.1 0.1 0.1 ...
```

PCA Analysis and Biplots

```
#Standardize data for PCA Analysis
pca_result_life <- prcomp(life_expectance_data, scale = T)

summary(pca_result_life)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.8560 1.5505 1.0019 0.9615 0.66416 0.57346 0.51123
## Proportion of Variance 0.3827 0.2671 0.1115 0.1027 0.04901 0.03654 0.02904
## Cumulative Proportion  0.3827 0.6499 0.7614 0.8641 0.91312 0.94966 0.97870
##                            PC8     PC9
## Standard deviation     0.43342 0.06234
## Proportion of Variance 0.02087 0.00043
## Cumulative Proportion  0.99957 1.00000
```

```
#Scree plot
fviz_eig(pca_result_life, addlabels = TRUE)
```

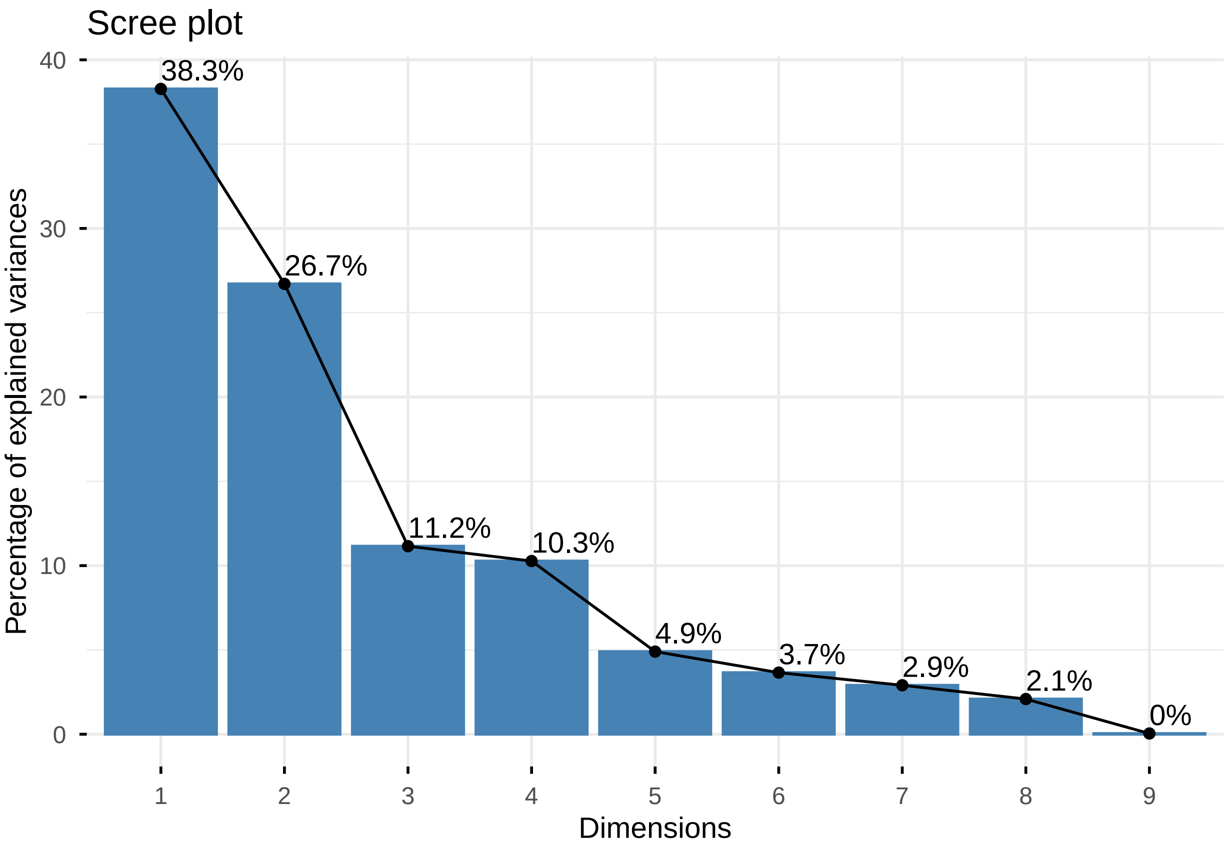

Figure 1.1: Scree Plot Showing Variance Explained by Each Principal Component (Life Expectancy Data)

```
#FactoMineR PCA
life.pca = PCA(life_expectance_data,scale.unit = T , ncp = 6 , graph=F)

#Biplot from FactoMineR PCA
fviz_pca_biplot(life.pca, repel = TRUE, geom = "point")
```

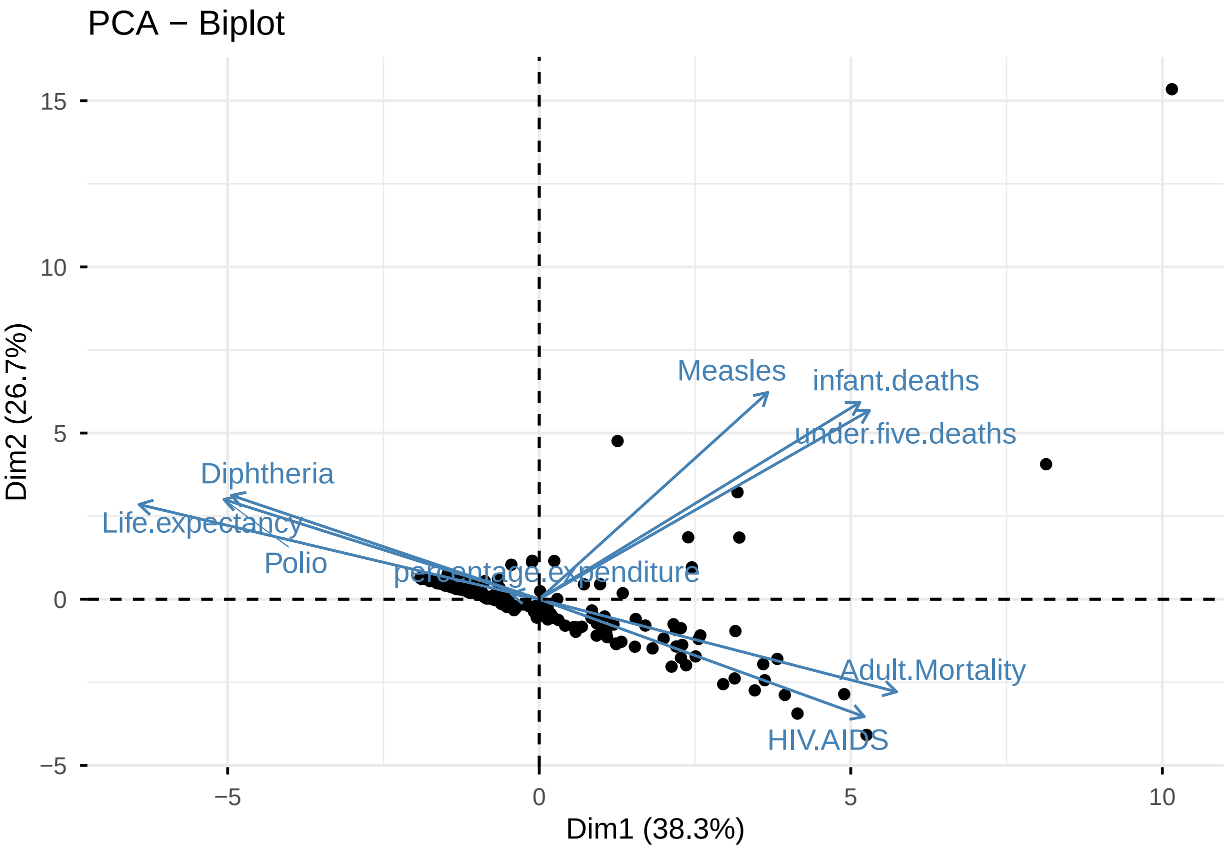

Figure 1.2: PCA Biplot Displaying Relationships Among Countries and Health Variables

Interpretation:

The first principal component explains about 38% of the total variance, the second principal component explains about 27%, and the third principal component explains about 11%. Together the three capture about 76% of the total variance, capturing most of the meaningful structure of the data. The biplot illustrates how these components structure the data: countries with high life expectancy and strong immunization coverage cluster on the negative side of PC1, while those with higher adult mortality, HIV/AIDS prevalence, and child deaths appear on the positive side. PC2 further separates countries where health challenges are concentrated in early childhood (higher measles and under-five deaths) from those with relatively lower child disease burden. PC3 introduces an orthogonal dimension tied primarily to variation in health spending (Percentage Expenditure), capturing financial differences that do not align with the main health-outcome patterns. Together, the first three componenets reveal clear groupings of countries based on overall health conditions, disease patterns, and investment in healthcare.

Analysis of Eigenvalues and Eigenvectors

```
#Eigenvalues table
eig_val_life = get_eigenvalue(life.pca)
eig_val_life
```

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1 3.444655571      38.27395079                    38.27395
## Dim.2 2.403962985      26.71069984                    64.98465
## Dim.3 1.003811559      11.15346176                    76.13811
## Dim.4 0.924508671      10.27231857                    86.41043
## Dim.5 0.441112639       4.90125154                    91.31168
## Dim.6 0.328853488       3.65392764                    94.96561
## Dim.7 0.261356348       2.90395943                    97.86957
## Dim.8 0.187852285       2.08724761                    99.95682
## Dim.9 0.003886454       0.04318282                   100.00000
```

```
#Variable results: coordinates, correlations, cos2, contributions
var_life = get_pca_var(life.pca)

var_life$coord
```

```
##                              Dim.1       Dim.2        Dim.3        Dim.4
## Life.expectancy        -0.80428084  0.35699637  0.048899608 -0.222156549
## Adult.Mortality         0.71829936 -0.34896505 -0.114484580  0.464403644
## infant.deaths           0.64464128  0.74199750  0.010870517 -0.011285678
## percentage.expenditure -0.05662137  0.02050781  0.965972810  0.250631563
## Measles                 0.45949768  0.77922143  0.003440692 -0.038936459
## under.five.deaths       0.66390596  0.71132579  0.010524392 -0.003808886
## Polio                  -0.63384742  0.37622815 -0.179763144  0.492596357
## Diphtheria             -0.61857563  0.39129203 -0.138295999  0.532552601
## HIV.AIDS                0.65326715 -0.44264165 -0.059404184  0.262196199
##                              Dim.5       Dim.6
## Life.expectancy         0.24422318  0.13305482
## Adult.Mortality        -0.21860296 -0.01053028
## infant.deaths           0.01810643 -0.03884866
## percentage.expenditure  0.00813246 -0.01765748
## Measles                 0.03747834  0.11292580
## under.five.deaths       0.01713804 -0.05299334
## Polio                   0.18406466 -0.37931334
## Diphtheria             -0.13820088  0.35481206
## HIV.AIDS                0.52783408  0.15455508
```

Interpretation:

Based on the Kaiser criterion (eigenvalues > 1), the first three principal components were retained for interpretation, as they capture the most meaningful variance. The loadings indicate that PC1 reflects overall health outcomes. Life Expectancy, Polio, and Diphtheria load negatively, while Adult Mortality, Infant Deaths, Under-Five Deaths, and HIV/AIDS load positively. This means that higher values on PC1 represent poorer health outcomes and lower life expectancy, while lower values indicate better overall health and longer life expectancy. PC2 differentiates disease patterns by age groups. Measles, Infant Deaths, and Under-Five Deaths load positively, while Adult Mortality and HIV/AIDS show negative loadings. In other words, PC2 separates countries where health challenges are concentrated in childhood from those where disease burden is more prominent in adulthood. PC3 is driven primarily by financial variation, with Percentage Expenditure loading strongly and positive. This componenet captures differences in health-related spending across countries, separate from the broader health and disease patterns described by PC1 and PC2.

```
#Correlation Circle Plot
fviz_pca_var(life.pca, col.var = "steelblue", repel = TRUE)
```

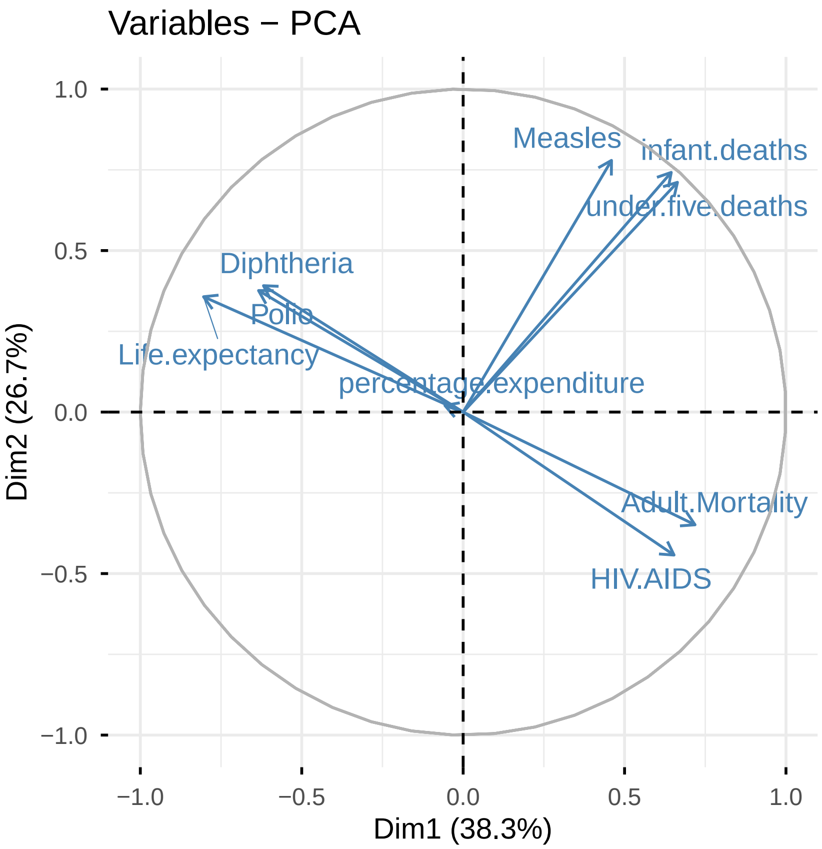

Figure 1.3: Correlation Circle of Variables on the First Two Principal Components

Interpretation:

The correlation circle shows that variables pointing in similar directions are positively correlated, while those on opposite sides are negatively correlated. Life Expectancy, Polio, and Diphtheria cluster together on the left, indicating that countries with higher life expectancy also tend to have higher vaccination rates. On the opposite side, Adult Mortality and HIV/AIDS point in the opposite direction, showing their strong negative relationship with those health indicators. Infant Deaths, Under-Five Deaths, and Measles cluster toward the upper right, meaning they are closely related to each other and represent another aspect of poor health conditions. Percentage Expenditure is found near the origin, showing it is not well represented on the factor map. Overall, the circle highlights two main patterns: one associated with better health outcomes and vaccination coverage, and another linked to higher mortality and disease prevalence.

```r
#Quality of representation heatmap
fviz_cos2(life.pca, choice = "var", axes = 1:2)
```

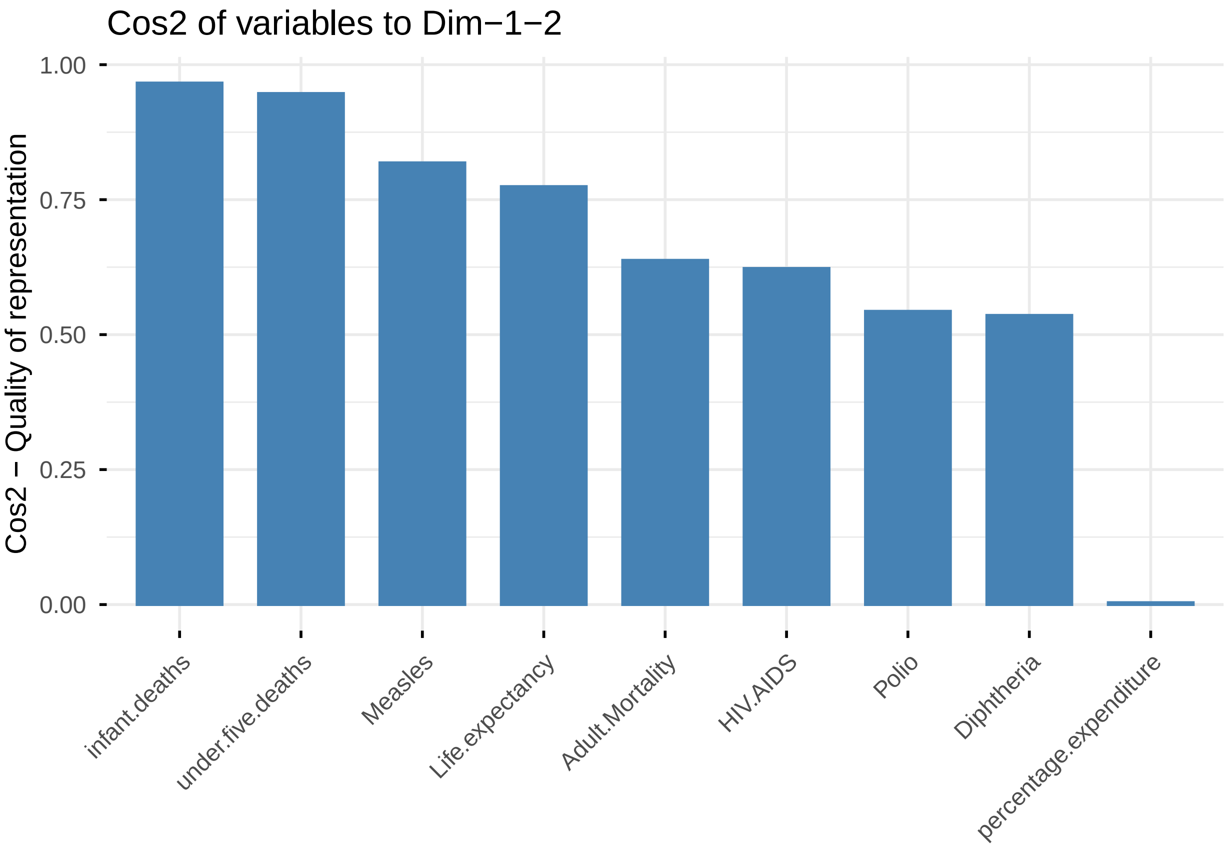

Figure 1.4: Quality of Representation $\cos^2$ **of Variables on PC1 and PC2**

Interpretation:

The $\cos^2$ (quality of representation) of each variable on the first two principal components is displayed in this plot. Variables with higher $\cos^2$ values are better represented in the reduced PCA space. Infant Deaths and Under-Five Deaths have the strongest representation, followed by Measles and Life Expectancy, meaning most of their variation is captured by PC1 and PC2. In contrast, Percentage Expenditure is very weakly represented, suggesting that it contributes more to later components rather than the main health and mortality dimensions.

```r
corrplot(var_life$cos2, is.corr = FALSE)
```

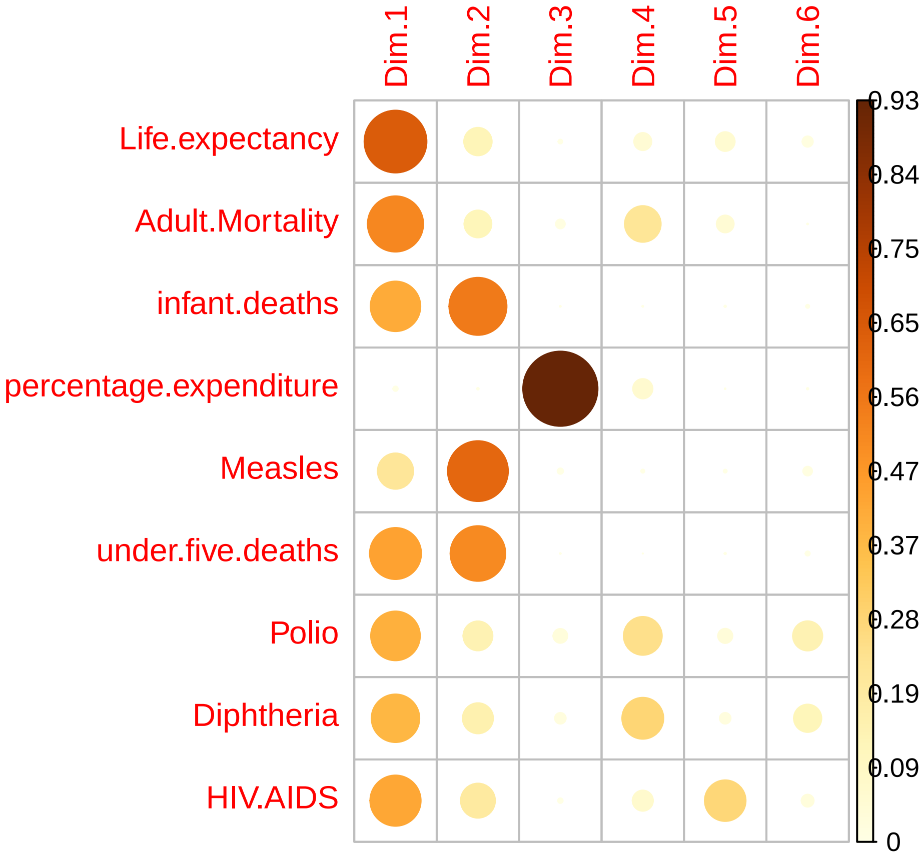

Figure 1.5: Heatmap of Variable Representation Quality $\cos^2$ for the First Two Principal Components

Interpretation:

This heatmap shows the $\cos^2$ values across all principal components, showing how each variable is represented by each dimension. Life Expectancy, Adult Mortality, HIV/AIDS, and immunization indicators (Polio and Diphtheria) are captured mainly by the first component, which represents general health outcomes. Infant Deaths, Under-Five Deaths, and Measles are associated with both the first and second components, indicating a related but distinct dimension of child mortality and disease burden. Percentage Expenditure is almost entirely captured by the third component, reflecting differences in healthcare spending. Overall, the first two dimensions summarize health and mortality patterns, while the third captures financial variation across countries

```r
#Top contributing variables to PC1
fviz_contrib(life.pca, choice = "var", axes = 1, top = 10)
```

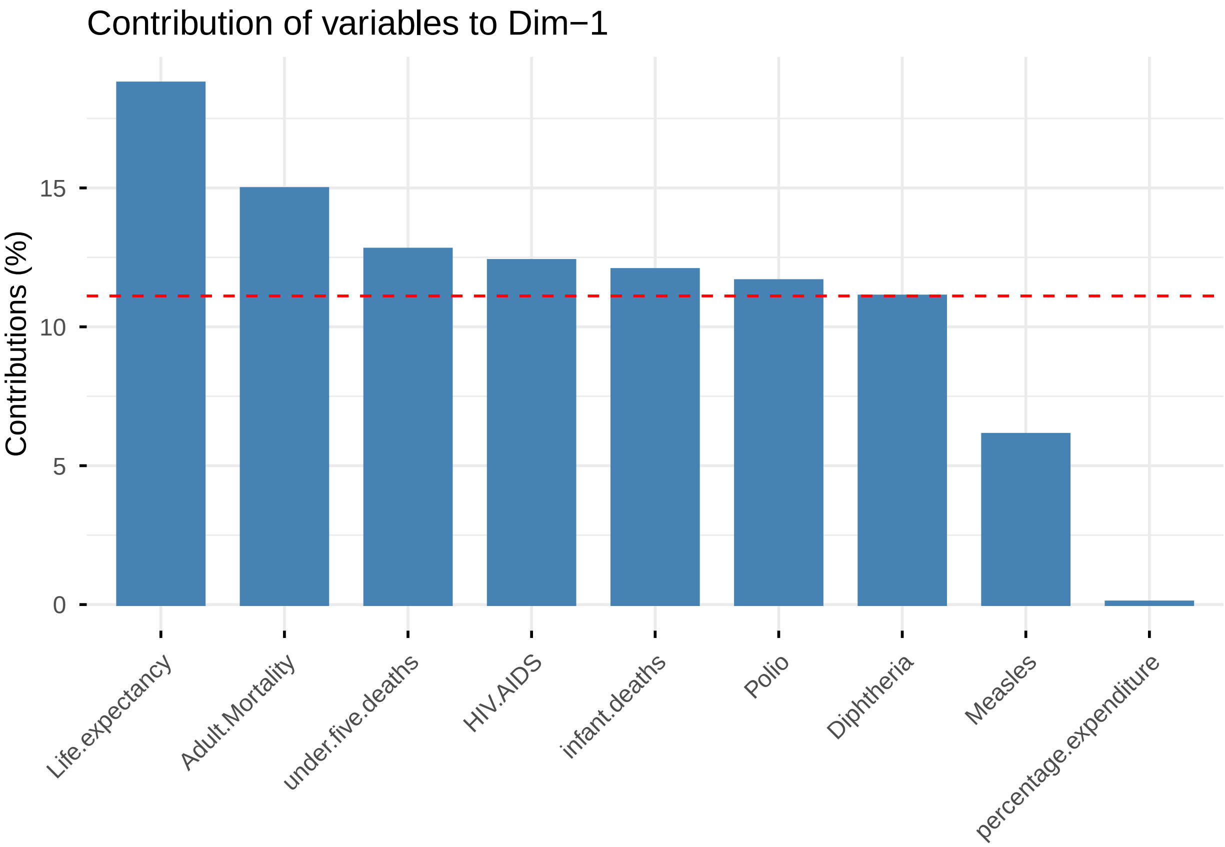

Figure 1.6: Top Variable Contributions to the First Principal Component (PC1)

```r
#Top contributing variables to PC2
fviz_contrib(life.pca, choice = "var", axes = 2, top = 10)
```

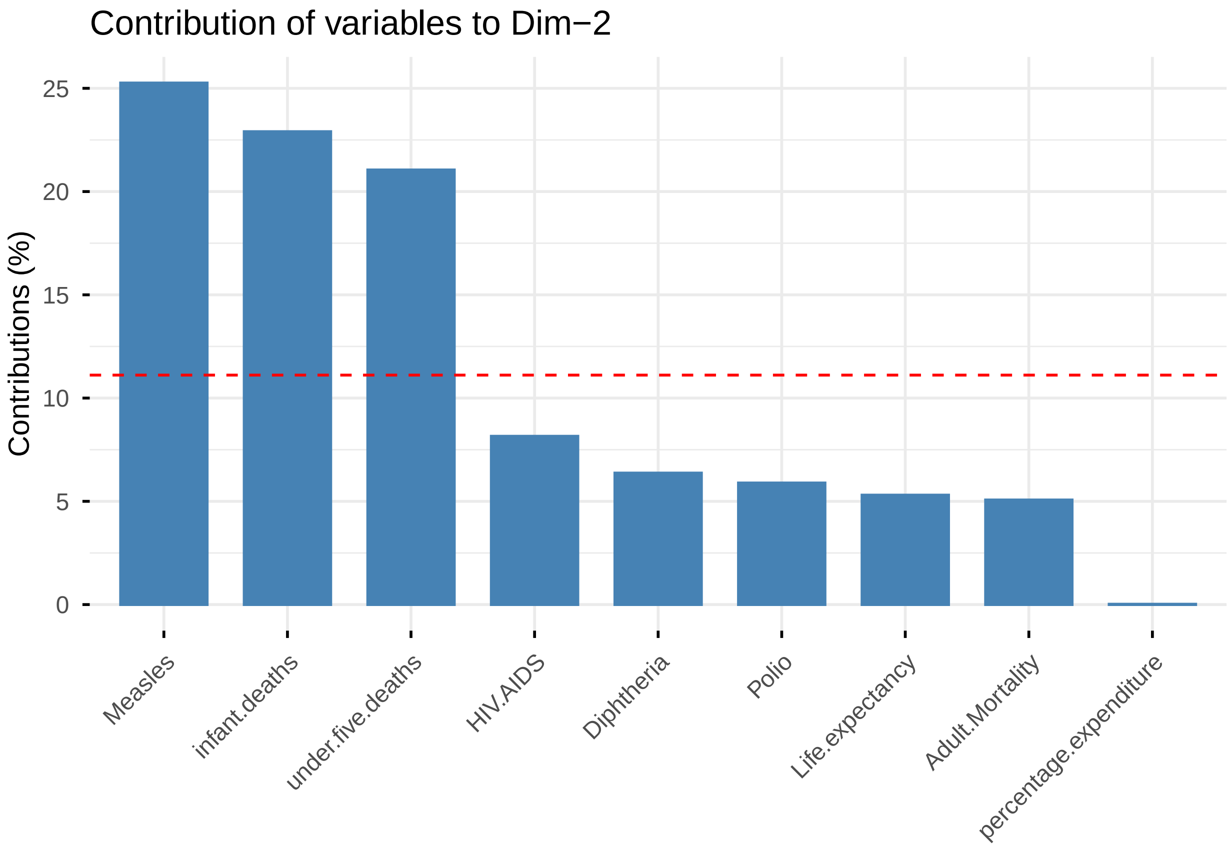

Figure 1.7: Top Variable Contributions to the Second Principal Component (PC2)

```r
#Top contributing variables to PC3
fviz_contrib(life.pca, choice = "var", axes = 3, top = 10)
```

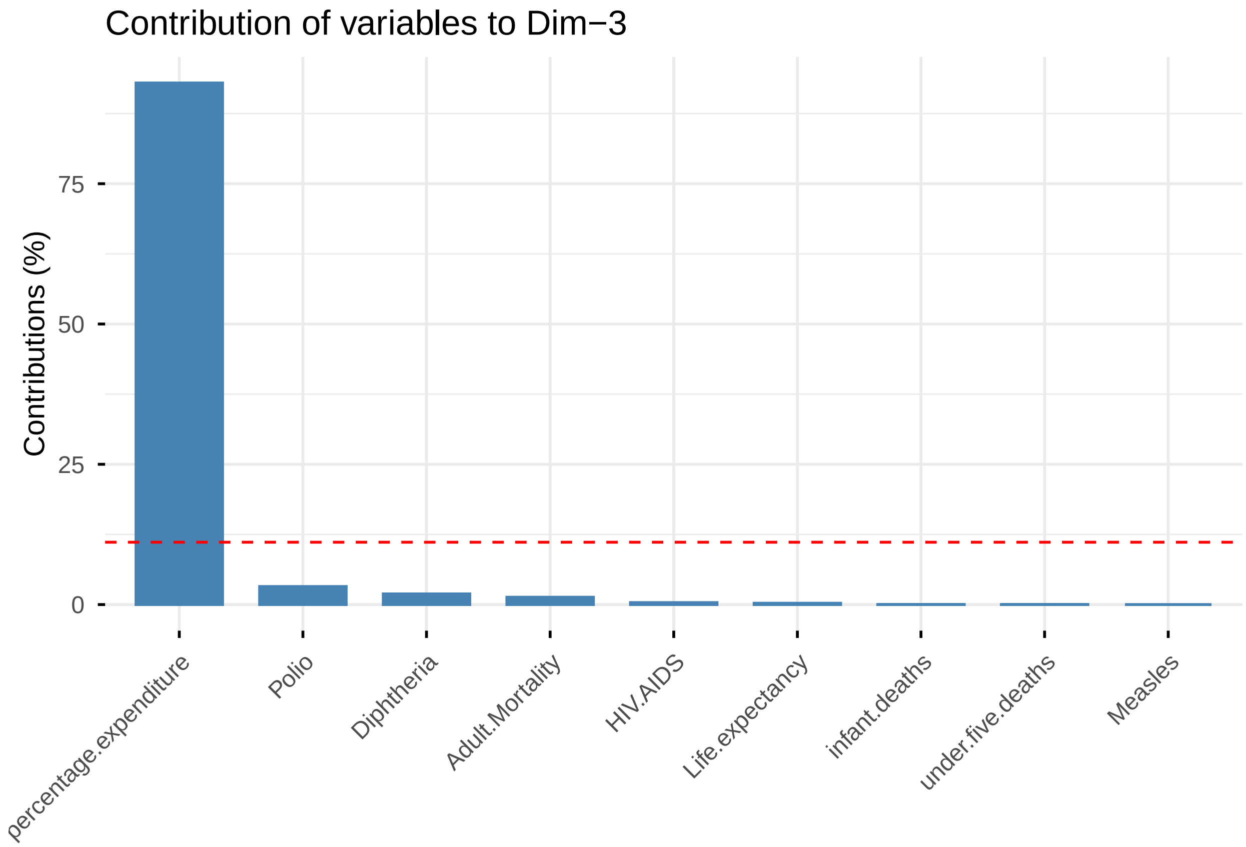

Figure 1.8: Top Variable Contributions to the Third Principal Component (PC3)

Interpretation:

The contribution plots show that Life Expectancy and Adult Mortality contribute the most to PC1, making them the main variables driving this overall health dimension. The remaining mortality and disease indicators contribute at fairly similar levels, while Percentage Expenditure contributes very little. For the second component, Measles, Infant Deaths, and Under-Five Deaths stand out as the main contributors, suggesting that PC2 captures variation related to childhood disease and early-life outcomes. For the third componenet, Percentage Expenditure stands out as the main contributor, suggesting that PC3 captures variation related to financial variation across countries.

Conclusion

Overall, the PCA shows that life expectancy differences across countries in 2015 can be largely explained by a small group of strongly correlated health indicators. The first three components account for about 76% of the total variation: PC1 captures a broad health and development dimension, PC2 reflects variation in child mortality and infectious disease, and PC3 represents financial differences in health spending. Countries with higher life expectancy and stronger vaccination coverage tend to score lower on PC1, while those with higher HIV/AIDS prevalence, infant deaths, and overall mortality score higher. These components together highlight how a few key variables can summarize the major global patterns in population health.

Discussion of PCA Assumptions and Limitations

PCA assumes linear relationships among variables and that the components explaining the most variance are the most meaningful. Since all variables were standardized, each contributed equally to the analysis, regardless of their original scales of measurement units. However, PCA is sensitive to outliers and depends heavily on the correlation structure of the data, meaning that highly correlated variables can dominate the results. In the context of the life expectancy data, this is especially relevant because many health indicators, such as mortality rates, tend to move together. Despite these limitations, PCA still proved useful for reducing dimensionality and uncovering clear, interpretable patterns in global health outcomes.

Analysis of Starbucks Customer Satisfaction: Using First Dataset Provided

Context:

This dataset contains survey responses from 113 customers regarding their purchasing behavior at Starbucks. It includes 20 variables measuring satisfaction, demographics, spending patterns, and interaction preferences. PCA is applied to identify the underlying dimensions that structure these responses, revealing the latent factors that shape overall customer satisfaction and loyalty.

```r
#Reading the data
starbucks_data <- read.csv("Starbucks satisfactory survey encode cleaned (1).csv")

#Keep only numeric columns
starbucks_data <- select_if(starbucks_data, is.numeric)

#Remove columns with NA values and Zero Variance
starbucks_data  <- starbucks_data  %>% select_if(~ !any(is.na(.)))
starbucks_data <- starbucks_data %>% select_if(~ var(.) > 0)

dim(starbucks_data)
```

```
## [1] 113  20
```

```r
str(starbucks_data)
```

```
## 'data.frame':    113 obs. of  20 variables:
##  $ Id               : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ gender           : int  1 1 0 1 0 1 1 0 1 0 ...
##  $ age              : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ status           : int  0 0 2 0 0 0 0 2 0 2 ...
##  $ income           : int  0 0 0 0 0 0 0 2 0 0 ...
##  $ visitNo          : int  3 3 2 3 2 3 3 3 3 2 ...
##  $ method           : int  0 2 0 2 2 0 0 0 1 2 ...
##  $ timeSpend        : int  1 0 1 0 1 1 0 1 0 0 ...
##  $ location         : int  0 1 2 2 1 2 0 2 2 2 ...
##  $ membershipCard   : int  0 0 0 1 1 1 0 0 0 1 ...
##  $ spendPurchase    : int  1 1 1 1 2 1 2 1 3 2 ...
##  $ productRate      : int  4 4 4 2 3 4 5 4 5 4 ...
##  $ priceRate        : int  3 3 3 1 3 3 5 2 4 3 ...
##  $ promoRate        : int  5 4 4 4 4 5 5 3 4 3 ...
##  $ ambianceRate     : int  5 4 4 3 2 5 5 3 4 4 ...
##  $ wifiRate         : int  4 4 4 3 2 4 3 3 4 3 ...
##  $ serviceRate      : int  4 5 4 3 3 5 5 3 4 3 ...
##  $ chooseRate       : int  3 2 3 3 3 4 5 3 4 4 ...
##  $ promoMethodOthers: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ loyal            : int  0 0 0 1 0 0 0 0 0 0 ...
```

PCA and Biplots

```r
#PCA Analysis
pca_result_starbucks <- prcomp(starbucks_data, scale = T)

summary(pca_result_starbucks)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.9373 1.6519 1.32748 1.27678 1.12219 1.07027 0.92854
## Proportion of Variance 0.1877 0.1364 0.08811 0.08151 0.06297 0.05727 0.04311
## Cumulative Proportion  0.1877 0.3241 0.41220 0.49371 0.55668 0.61395 0.65706
##                            PC8     PC9   PC10   PC11    PC12    PC13    PC14
## Standard deviation     0.91293 0.89002 0.8820 0.8173 0.77750 0.75813 0.71009
## Proportion of Variance 0.04167 0.03961 0.0389 0.0334 0.03023 0.02874 0.02521
## Cumulative Proportion  0.69873 0.73834 0.7772 0.8106 0.84086 0.86960 0.89481
##                           PC15    PC16    PC17   PC18    PC19    PC20
## Standard deviation     0.67112 0.65157 0.62203 0.5568 0.52945 0.50165
## Proportion of Variance 0.02252 0.02123 0.01935 0.0155 0.01402 0.01258
## Cumulative Proportion  0.91733 0.93856 0.95790 0.9734 0.98742 1.00000
```

```r
#Scree Plot
fviz_eig(pca_result_starbucks, addlabels = TRUE)
```

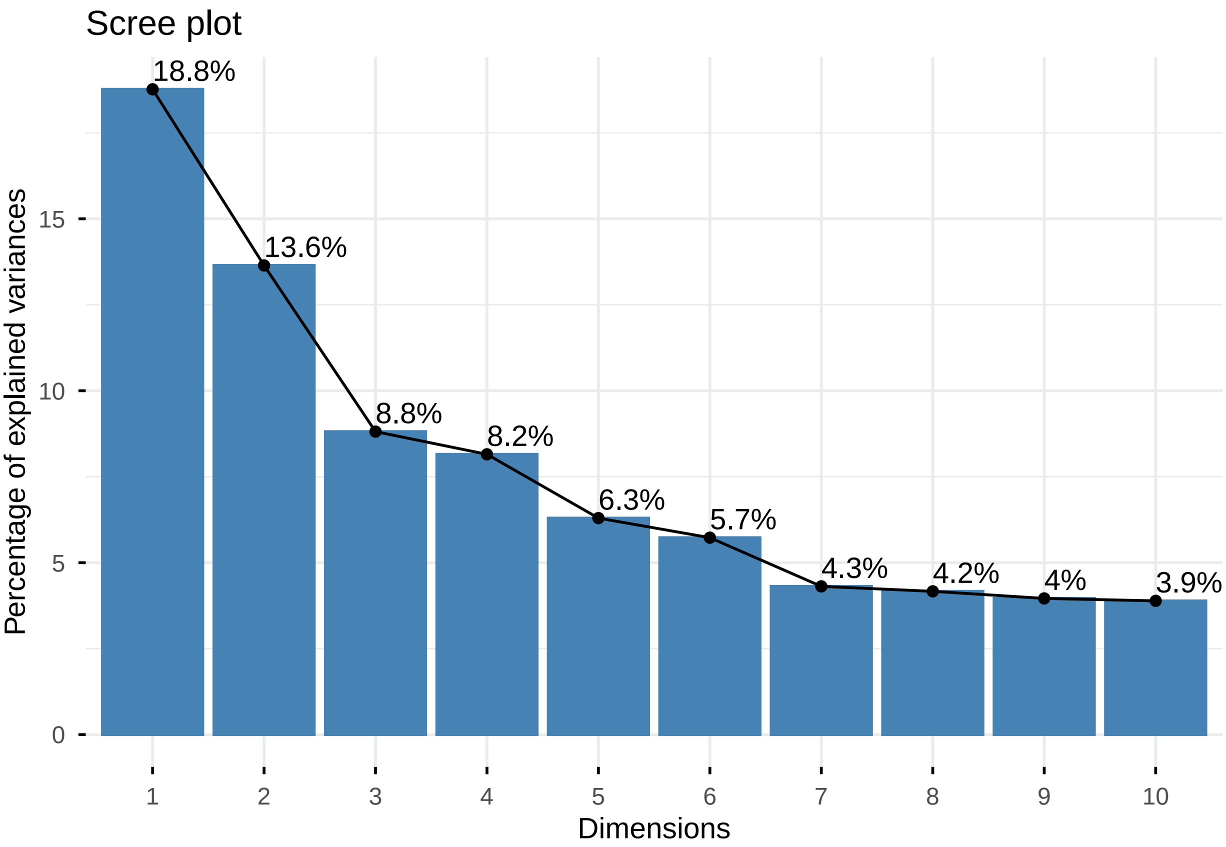

Figure 2.1: Scree Plot Showing Variance Explained by Each Principal Component (Starbucks Data)

```r
#FactoMineR PCA
starbucks.pca = PCA(starbucks_data,scale.unit = T , ncp = 6 , graph=F)

#Biplot from FactoMineR PCA
fviz_pca_biplot(starbucks.pca, repel = TRUE, geom = "point")
```

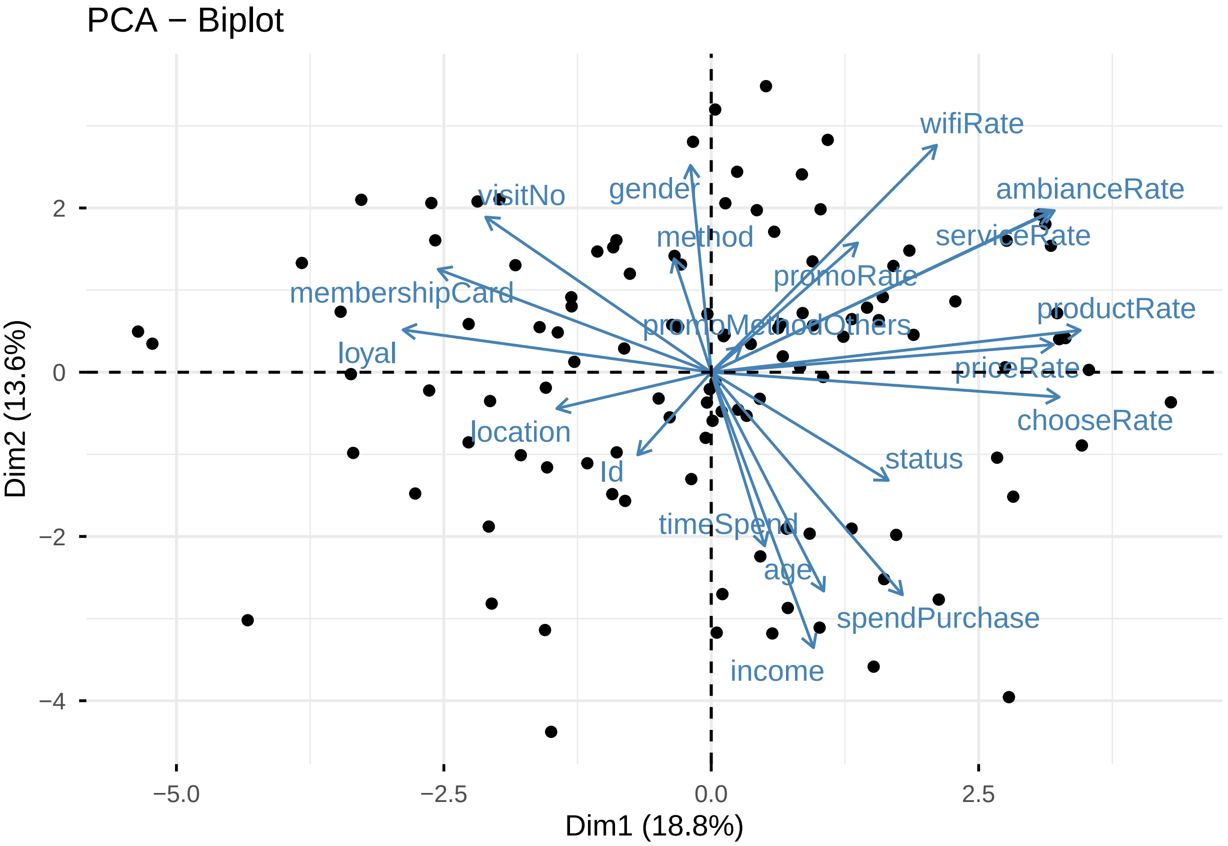

Figure 2.2: PCA Biplot Showing Relationships Among Satisfaction Dimensions

Interpretation:

From the scree plot, the first three components explain about 42% of the total variance (~19%, ~14%, and ~9% respectively). While a full analysis would retain the first six components (which together account for about 61% of the variance), the first three are sufficient for summarizing the main structure in the data. PC1 reflects overall customer satisfaction: serviceRate, chooseRate, ambianceRate, productRate, and priceRate all load strongly and positively, indicating that customers tend to rate these aspects together. PC2 contrasts demographic and behavioral factors. Income, Age, and spendPurchase load negatively, while wifiRate loads positively, separating customers with higher spending and income from those whose satisfaction depends more on store amenities. PC3 captures contextual and lifestyle-related variation, driven mainly by variables such as Status, Method, and Location. These factors do not measure satisfaction directly but instead reflect differences in customers' employment context, how far they travel to the store, and their preferred purchasing method. This component represents situational patterns in how customers engage with Starbucks, which operate independently from the main satisfaction dimensions captured by PC1 and PC2.

Analysis of Eigenvalues and Eigenvectors

```
#Eigenvalues table
eig_val_starbucks = get_eigenvalue(starbucks.pca)
eig_val_starbucks
```

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1   3.7530858        18.765429                    18.76543
## Dim.2   2.7287682        13.643841                    32.40927
## Dim.3   1.7622015         8.811008                    41.22028
## Dim.4   1.6301762         8.150881                    49.37116
## Dim.5   1.2593005         6.296503                    55.66766
## Dim.6   1.1454718         5.727359                    61.39502
## Dim.7   0.8621908         4.310954                    65.70597
## Dim.8   0.8334334         4.167167                    69.87314
## Dim.9   0.7921358         3.960679                    73.83382
## Dim.10  0.7779670         3.889835                    77.72365
## Dim.11  0.6679341         3.339671                    81.06333
## Dim.12  0.6045072         3.022536                    84.08586
## Dim.13  0.5747645         2.873822                    86.95968
## Dim.14  0.5042283         2.521141                    89.48083
## Dim.15  0.4504035         2.252018                    91.73284
## Dim.16  0.4245475         2.122737                    93.85558
## Dim.17  0.3869250         1.934625                    95.79021
## Dim.18  0.3099910         1.549955                    97.34016
## Dim.19  0.2803155         1.401578                    98.74174
## Dim.20  0.2516525         1.258262                   100.00000
```

```
#Variable results: coordinates, correlations, cos2, contributions
var_starbucks = get_pca_var(starbucks.pca)

var_starbucks$coord
```

```
##                        Dim.1       Dim.2        Dim.3       Dim.4        Dim.5
## Id                -0.14061758 -0.20604102  0.362004597 -0.04880276  0.554356728
## gender            -0.03974587  0.51573131 -0.342382163 -0.19979933 -0.241807404
## age                0.21528526 -0.54546397 -0.284715155  0.21216475  0.280244098
## status             0.33879117 -0.26952286 -0.509022515  0.15368726  0.165872807
## income             0.19570046 -0.68666825 -0.311241478  0.30279685  0.058192100
## visitNo           -0.43158763  0.38668221 -0.124923190  0.25977212 -0.096034909
## method            -0.07099052  0.28270923 -0.537833049 -0.40291844  0.370260444
## timeSpend          0.10209369 -0.43287095  0.430457350  0.42749423 -0.320667886
## location          -0.29558460 -0.09051646  0.471126259 -0.01911577  0.389686413
## membershipCard    -0.52268489  0.25643090  0.275729322  0.01684988  0.388926131
## spendPurchase      0.36588443 -0.55465889 -0.049407867 -0.36815290  0.009442139
## productRate        0.70620319  0.10441413  0.196618104 -0.04041420 -0.108877084
## priceRate          0.65457284  0.06878637  0.276961537 -0.21862964  0.116138636
## promoRate          0.28004497  0.32227901  0.046452725  0.42782101 -0.038872687
## ambianceRate       0.65663016  0.40278957  0.104583476  0.23152780  0.072393697
## wifiRate           0.43105665  0.56577223  0.051664123  0.17303028  0.302814748
## serviceRate        0.64747349  0.39924515 -0.137861657  0.28037120  0.127781737
## chooseRate         0.66587861 -0.06185744  0.227720517 -0.04408794  0.027431208
## promoMethodOthers  0.05650765  0.06343201  0.290204456 -0.48394650 -0.298980874
## loyal             -0.58969396  0.10557609  0.009330162  0.49802555 -0.038321979
##                         Dim.6
## Id                   0.430256273
## gender               0.157182007
## age                  0.100253368
## status               0.392588693
## income              -0.167018649
## visitNo              0.039239108
## method               0.131788489
## timeSpend            0.256655304
## location            -0.085674242
## membershipCard      -0.065981219
## spendPurchase       -0.041698613
## productRate         -0.290226565
## priceRate           -0.172880450
## promoRate            0.412861817
## ambianceRate        -0.034209441
## wifiRate            -0.055744572
## serviceRate          0.057405771
## chooseRate           0.120783499
## promoMethodOthers    0.581912131
## loyal                0.001873974
```

Interpretation:

Based on the Kaiser criterion (eigenvalues > 1), the first six principal components should be retained, but for interpretability the first three are examined. The loadings indicate that PC1 reflects overall customer satisfaction, with productRate, serviceRate, ambianceRate, chooseRate, and priceRate all loading strongly and positively. This means higher PC1 values correspond to customers who rate the Starbucks experience more favorably across multiple dimensions. PC2 captures demographic and behavioral variation, with income, age, and spendPurchase loading negatively, while wifiRate loads positively. In other words, PC2 separates higher-income or higher-spending customers from those whose evaluations are more influenced by amenities. PC3 is shaped by variables such as status, method, and location, representing differences in customers' circumstances and store-use patterns rather than differences in satisfaction. Together, the first three components summarize the main patterns underlying customer ratings and behaviors.

```
#Correlation Circle Plot
fviz_pca_var(starbucks.pca, col.var = "steelblue", repel = TRUE)
```

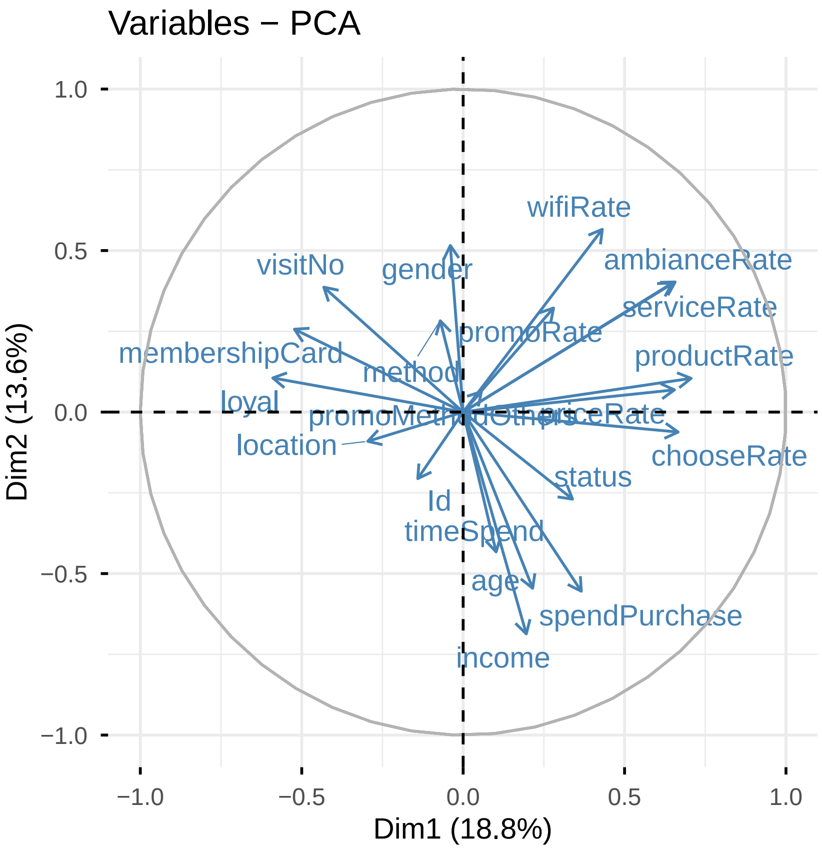

Figure 2.3: Correlation Circle of Customer Satisfaction Variables on PC1 and PC2

Interpretation:

The correlation circle shows that variables pointing in similar directions are positively correlated, while those pointing in opposite or far-separated directions are negatively or weakly related. ServiceRate, ambianceRate, productRate, priceRate, and chooseRate cluster tightly together on the right, indicating that they capture a common satisfaction dimension. Income, age, and spendPurchase form a separate grouping oriented roughly 90 degrees away, suggesting that demographic and spending behavior vary largely independently from satisfaction ratings. Variables such as status, method, and location appear near the center of the plot, indicating weak associations with PC1 and PC2—consistent with their stronger contributions to PC3. This suggests that PC3 represents an additional organizational or behavioral dimension that is not well captured in the first two components. Overall, the circle highlights two main patterns in the data: one driven by customer satisfaction ratings and another shaped by demographic and spending behavior, with a third subtler pattern emerging from membership and usage characteristics.

```
#Quality of representation heatmap
fviz_cos2(starbucks.pca, choice = "var", axes = 1:2)
```

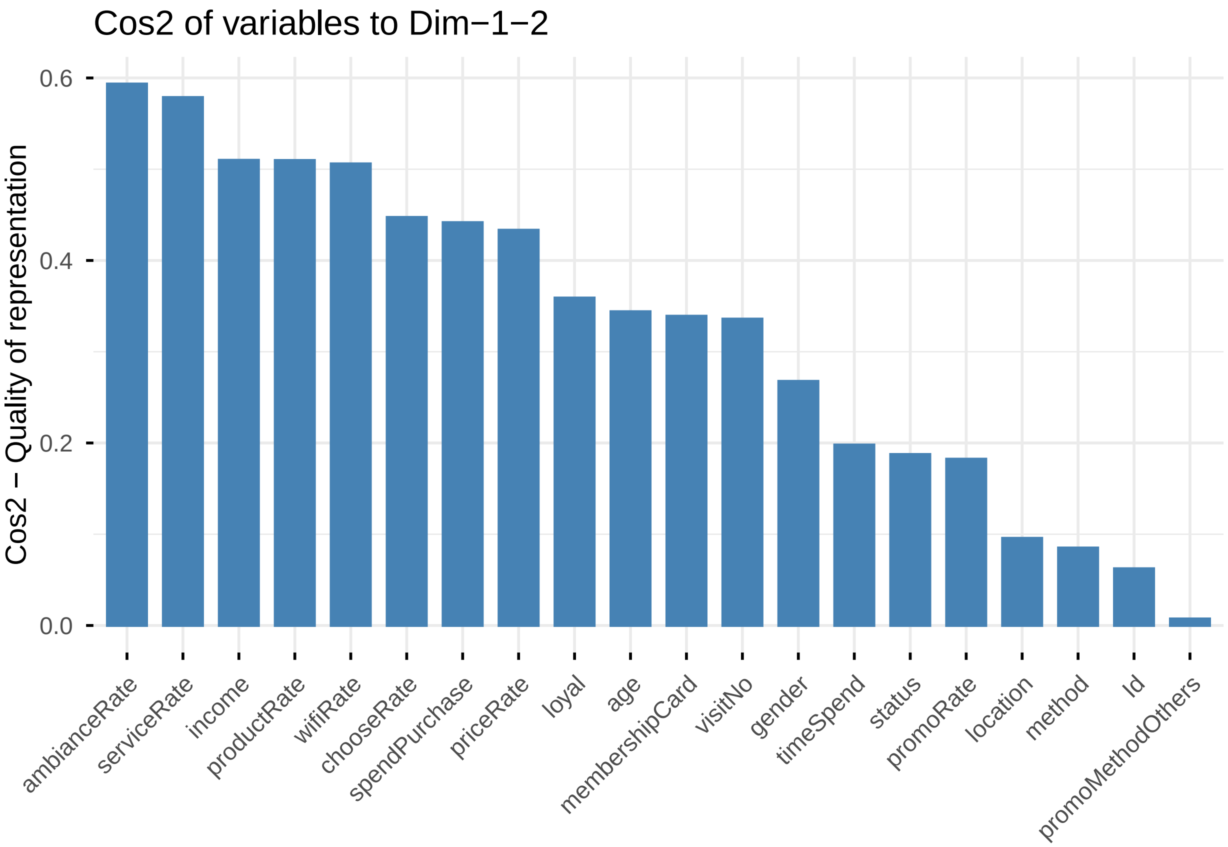

Figure 2.4: Quality of Representation $\cos^2$ of Variables on PC1 and PC2

Interpretation:

The $\cos^2$ (quality of representation) of each variable on the first two principal components is displayed in this plot. Variables with higher $\cos^2$ are better represented in the reduced PCA space. AmbianceRate and serviceRate have the strongest representation, followed by income, productRate, and wifiRate, meaning most of their variation is captured by PC1 and PC2. In contrast, status, location, and method are very weakly represented, suggesting that they contribute more to later components rather than the main customer satisfaction and behavior dimensions.

```r
corrplot(var_starbucks$cos2, is.corr = FALSE)
```

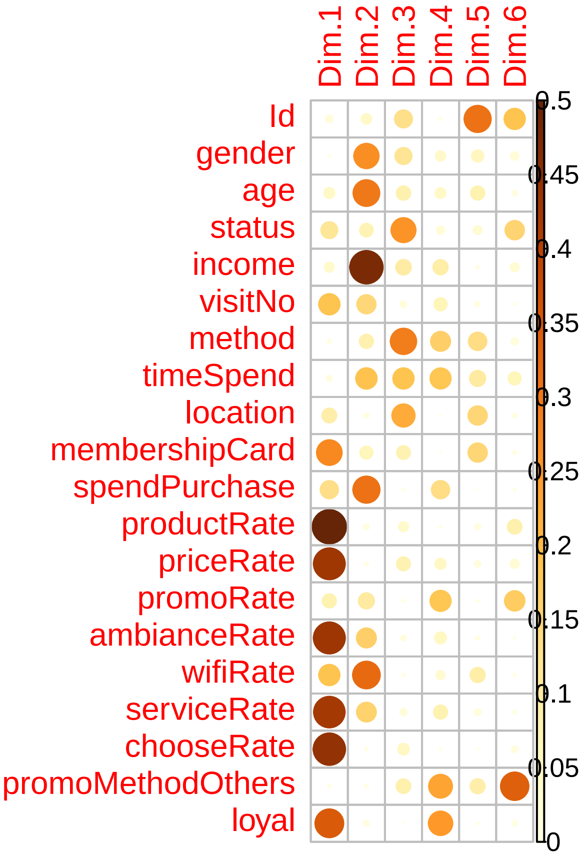

Figure 2.5: Heatmap of Variable Representation Quality $\cos^2$ **for the First Two Principal Components**

Interpretation:

The heatmap shows the $\cos^2$ values across all principal components, showing how each variable is represented by each dimension. ProductRate, priceRate, ambianceRate, serviceRate, and chooseRate are captured mainly by the first component, reflecting a shared customer satisfaction dimension. Income, spendPurchase, age, and wifiRate are mainly captured by the second components, indicating a demographic and spending-behavior dimension. Finally, status, location, and method show their strongest representation on the third component, indicating a separate contextual dimension related to customers' circumstances and store-use patterns.

```
#Top contributing variables to PC1
fviz_contrib(starbucks.pca, choice = "var", axes = 1, top = 10)
```

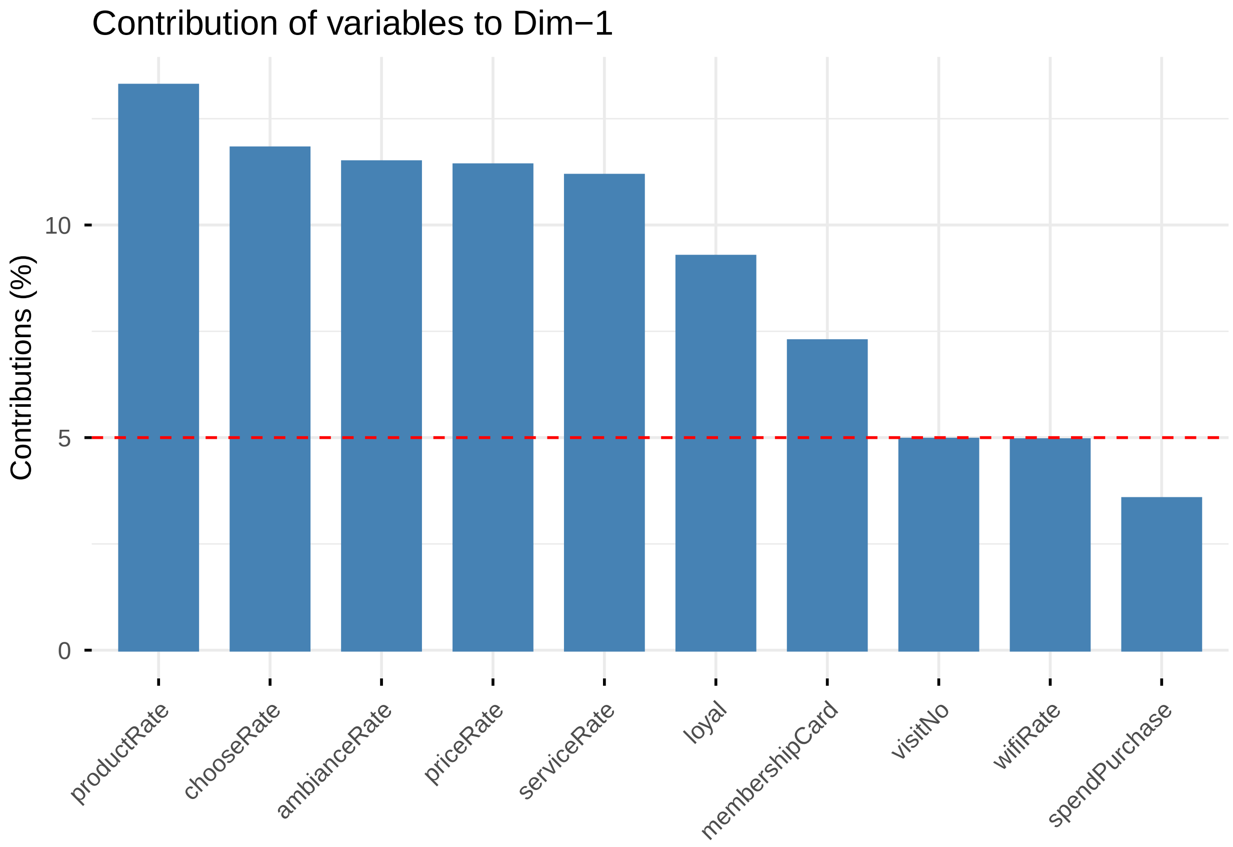

Figure 2.6: Top Variable Contributions to the First Principal Component (PC1)

```
#Top contributing variables to PC2
fviz_contrib(starbucks.pca, choice = "var", axes = 2, top = 10)
```

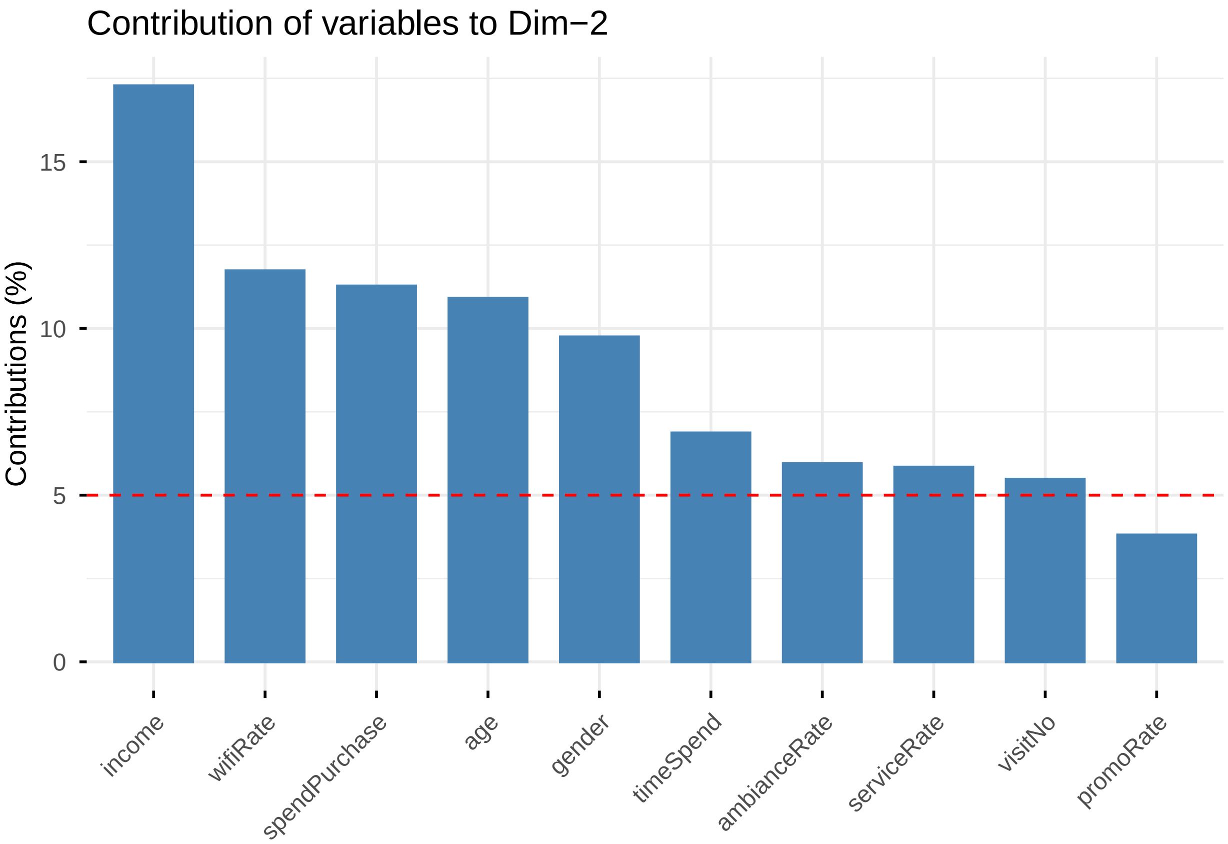

Figure 2.7: Top Variable Contributions to the Second Principal Component (PC2)

```
#Top contributing variables to PC2
fviz_contrib(starbucks.pca, choice = "var", axes = 3, top = 10)
```

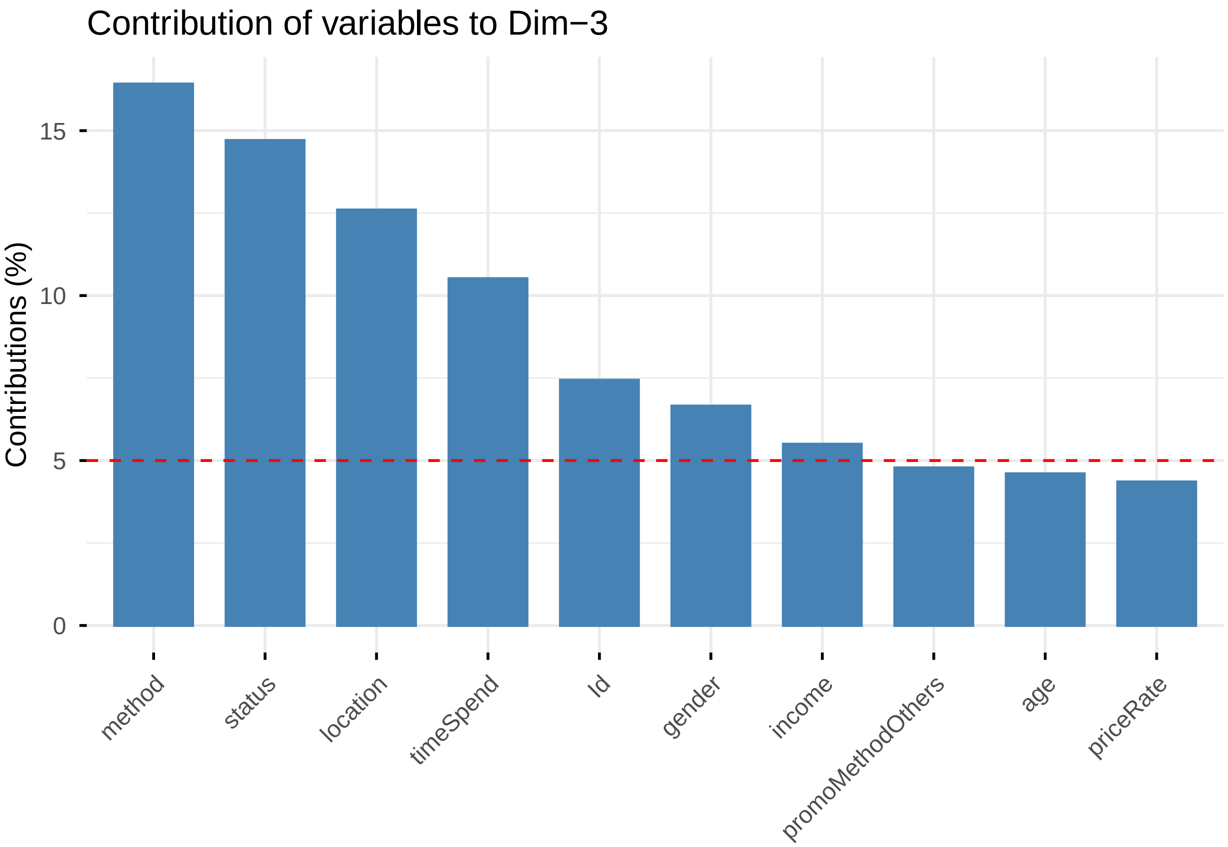

Figure 2.8: Top Variable Contributions to the Third Principal Component (PC3)

Interpretation:

The contribution plots show that productRate, chooseRate, ambianceRate, priceRate, and serviceRate contribute most strongly to the first principal component, confirming that PC1 represents overall customer satisfaction and the core in-store experience. PC2 is driven mainly by income, wifiRate, spendPurchase, and age, indicating that it captures demographic and spending-related variation rather than satisfaction itself. PC3 is dominated by Method, Status, and Location, reflecting a third dimension related to customers' circumstances and store-use patterns rather than their satisfaction ratings. Together, these components show that PCA separates satisfaction, demographics/behavior, and engagement patterns into three distinct sources of variation in the Starbucks dataset.

Conclusion

The Starbucks PCA results show that customer satisfaction is shaped by multiple overlapping dimensions rather than a single factor. The first three principal components together capture the most meaningful structure in the data. PC1 reflects overall satisfaction, driven by ratings of service, product quality, ambiance, price, and choice. PC2 captures demographic and spending-related variation, distinguishing higher-income or higher-spending customers from those more influenced by amenities. PC3 reflects customer circumstances and store-use patterns, including ordering method, distance from the store, and employment status. Overall, the analysis suggests that while the in-store experience is the strongest driver of satisfaction, demographic differences and engagement behaviors also meaningfully shape how customers evaluate their Starbucks visits.

Discussion of PCA Assumptions and Limitations

PCA assumes that relationships among variables are linear and that the components explaining the greatest variance are the most meaningful. Because all variables were standardized, each contributed equally to the analysis, regardless of its original scale. However, the method is sensitive to outliers and assumes that the underlying relationships are continuous and additive, which may not fully capture the complexity of customer preferences. Despite these limitations, PCA effectively reduced dimensionality and revealed clear, interpretable patterns in customer satisfaction and behavior.