# 1ª Avaliação

## Análise de Dados via Métodos Multivariados

Edvaldo Sampaio & Filipe Costa

01 agosto,2022


UNIVERSIDADE
FEDERAL DO PIAUÍ

# Dados

- Iremos utilizar o banco de dados **mtcars**. Os dados foram extraídos da revista `Motor Trend US` de 1974 e se referem ao consumo de combustível e 10 características de 32 automóveis (modelos de 1973 a 1974).

```
df = dados::mtcarros[, 1:7]; df|> DT::datatable(options = list(pageLength = 4))
```

Show 4 entries Search:

| | milhas_por_galao | cilindros | cilindrada | cavalos_forca | eixo | peso | velocidade |
|---|---|---|---|---|---|---|---|
| Mazda RX4 | 21 | 6 | 160 | 110 | 3.9 | 2.62 | 16.46 |
| Mazda RX4 Wag | 21 | 6 | 160 | 110 | 3.9 | 2.875 | 17.02 |
| Datsun 710 | 22.8 | 4 | 108 | 93 | 3.85 | 2.32 | 18.61 |
| Hornet 4 Drive | 21.4 | 6 | 258 | 110 | 3.08 | 3.215 | 19.44 |

Showing 1 to 4 of 32 entries Previous 1 2 3 4 5 ... 8 Next

# Medidas descritivas

```
psych::describe(df)[,c(-1,-2)]|>round(3)|> knitr::kable()
```

| | mean | sd | median | trimmed | mad | min | max | range | skew | kurtosis | se |
|---|---|---|---|---|---|---|---|---|---|---|---|
| milhas_por_galao | 20.091 | 6.027 | 19.200 | 19.696 | 5.411 | 10.400 | 33.900 | 23.500 | 0.611 | -0.373 | 1.065 |
| cilindros | 6.188 | 1.786 | 6.000 | 6.231 | 2.965 | 4.000 | 8.000 | 4.000 | -0.175 | -1.762 | 0.316 |
| cilindrada | 230.722 | 123.939 | 196.300 | 222.523 | 140.476 | 71.100 | 472.000 | 400.900 | 0.382 | -1.207 | 21.909 |
| cavalos_forca | 146.688 | 68.563 | 123.000 | 141.192 | 77.095 | 52.000 | 335.000 | 283.000 | 0.726 | -0.136 | 12.120 |
| eixo | 3.597 | 0.535 | 3.695 | 3.579 | 0.704 | 2.760 | 4.930 | 2.170 | 0.266 | -0.715 | 0.095 |
| peso | 3.217 | 0.978 | 3.325 | 3.153 | 0.767 | 1.513 | 5.424 | 3.911 | 0.423 | -0.023 | 0.173 |
| velocidade | 17.849 | 1.787 | 17.710 | 17.828 | 1.416 | 14.500 | 22.900 | 8.400 | 0.369 | 0.335 | 0.316 |

# Teste de esferecidade de Bartlett

- O teste a hipótese que as variáveis não sejam relacionadas:

$$\chi^2 = -[(n-1) - \frac{2p-5}{6}]ln|R|, v = \frac{p(p-1)}{2}$$

- $H_0$ = a matriz de covariância é similar a uma matriz identidade (sem correlação)
- $H_1$ = a matriz de covariância não é similar a uma matriz identidade (possui correlação)

Construindo um função no R:

```
bartlett = function(M){
    n = dim(M)[1]; p = dim(M)[2]
    R = det(cor(M)); v = (p*(p-1))/2
    value = -((n-1) - (2*p+5)/6)*log(R)
    QuiQuadrado = 1-  pchisq(value, v)
    print(list(
    sprintf(paste('Estatísica de Barlett:', round(value,3))),
    sprintf(paste('p-valor:', round(QuiQuadrado,3)))
    ))
}
```

# Teste de esferecidade de Bartlett

- Com R com Função própria

```
## [[1]]
## [1] "Estatísica de Barlett: 244.187"
##
## [[2]]
## [1] "p-valor: 0"
```

- Com função específica do pacote MVTests

```
results = MVTests::Bsper(data = df)
results$Chisq
```

```
## [1] 244.1867
```

```
results$p.value
```

```
## [1] 6.020416e-40
```

- Temos evidências para afirmar que existe correlação entre as variáveis

# KMO

- Medida de adequabilidade: é um teste estatístico que sugere a proporção de variância dos itens que pode estar sendo explicada por uma variável latente, indicando se é adequado aplicação de PCA e Fatorial nos dados.

```
KmoDF = psych::KMO(df); KmoDF
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: psych::KMO(r = df)
## Overall MSA =  0.83
## MSA for each item =
## milhas_por_galao        cilindros       cilindrada    cavalos_forca
##             0.87             0.87             0.85             0.90
##             eixo             peso       velocidade
##             0.85             0.77             0.61
```

- O índide do `KMO` é de 0.83, indicando um bom resultado, dessa forma temos a possibilidade de aplicar de métodos multivariados.

# Matriz de Correlação

```
c = cor(df); c|>round(3)|>knitr::kable()
```

| | milhas_por_galao | cilindros | cilindrada | cavalos_forca | eixo | peso | velocidade |
|---|---|---|---|---|---|---|---|
| milhas_por_galao | 1.000 | -0.852 | -0.848 | -0.776 | 0.681 | -0.868 | 0.419 |
| cilindros | -0.852 | 1.000 | 0.902 | 0.832 | -0.700 | 0.782 | -0.591 |
| cilindrada | -0.848 | 0.902 | 1.000 | 0.791 | -0.710 | 0.888 | -0.434 |
| cavalos_forca | -0.776 | 0.832 | 0.791 | 1.000 | -0.449 | 0.659 | -0.708 |
| eixo | 0.681 | -0.700 | -0.710 | -0.449 | 1.000 | -0.712 | 0.091 |
| peso | -0.868 | 0.782 | 0.888 | 0.659 | -0.712 | 1.000 | -0.175 |
| velocidade | 0.419 | -0.591 | -0.434 | -0.708 | 0.091 | -0.175 | 1.000 |

Esta tabela será a base para os cálculos de `Componentes Principais` e `Análise Fatorial`

# Medidas algébricas

```
autos = eigen(c); autos$values|>round(3) # autovalores
```

```
## [1] 5.086 1.157 0.345 0.158 0.129 0.076 0.049
```

```
autos$vectors|>round(3)|>knitr::kable() # autovetores
```

| | | | | | | |
|---|---|---|---|---|---|---|
| 0.413 | -0.083 | 0.242 | 0.767 | -0.213 | -0.090 | -0.351 |
| -0.425 | -0.078 | 0.188 | 0.194 | 0.238 | 0.781 | -0.273 |
| -0.423 | 0.082 | -0.118 | 0.588 | 0.149 | -0.162 | 0.638 |
| -0.388 | -0.337 | -0.203 | -0.007 | -0.831 | 0.046 | -0.039 |
| 0.331 | -0.449 | -0.755 | 0.117 | 0.222 | 0.233 | 0.037 |
| -0.391 | 0.322 | -0.441 | 0.107 | 0.167 | -0.366 | -0.613 |
| 0.240 | 0.749 | -0.294 | 0.061 | -0.328 | 0.407 | 0.131 |

# Componentes Principais

Técnica da estatística multivariada que consiste em transformar um conjunto de variáveis originais em outro conjunto de variáveis de mesma dimensão denominadas de `componentes principais`.

- Cada componente principal é uma combinação linear de todas as variáveis originais.
- Os componente são independentes entre si e estimados com o propósito de reter, em ordem de estimação, o máximo de informação, em termos da variação total contida nos dados

Os Objetivos são:

- **Identificação de padrões ocultos dos dados**;
- **Redução de dimensionalidade, pela diminuição da redundância nos dados**;
- **Identificar variáveis correlacionadas**.

# Medidas algébricas - análise

## Componentes Principais

- $Y1 = 0.413X1 - 0.425X2 - 0.423X3 - 0.388X4 + 0.331X5 - 0.391X6 + 0.240X7$
- $Y2 = -0.083X1 - 0.078X2 + 0.082X3 - 0.337X4 - 0.449X5 + 0.322X6 + 0.749X7$
- $Y3 = 0.242X1 + 0.188X2 - 0.118X3 - 0.203X4 - 0.755X5 - 0.441X6 - 0.294X7$
- $Y4 = 0.767X1 + 0.194X2 + 0.588X3 - 0.007X4 + 0.117X5 + 0.107X6 - 0.061X7$
- $Y5 = -0.213X1 + 0.238X2 + 0.149X3 - 0.831X4 - 0.222X5 - 0.167X6 - 0.328X7$
- $Y6 = -0.090X1 + 0.781X2 - 0.162X3 + 0.046X4 + 0.233X5 - 0.366X6 + 0.407X7$
- $Y7 = -0.351X1 - 0.273X2 + 0.638X3 - 0.039X4 + 0.037X5 - 0.613X6 + 0.131X7$

## Variancia explicada a partir do maior autovalor(componente)

```
PVTE1 = round(autos$values,3)/sum(round(autos$values,3));round(PVTE1*100,3)
```

```
## [1] 72.657 16.529  4.929  2.257  1.843  1.086  0.700
```

# Análise por Componentes Principais

**Usando a função `prcomp`, do pacote base do `r`**

```
PCAdf = princomp(df, cor = TRUE); PCAdf$sdev|>round(3)
```

```
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
##  2.255  1.075  0.587  0.397  0.360  0.275  0.222
```

```
summary(PCAdf)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     2.2552383 1.0754374 0.5872406 0.39740859 0.35985282
## Proportion of Variance 0.7265857 0.1652236 0.0492645 0.02256194 0.01849915
## Cumulative Proportion  0.7265857 0.8918093 0.9410738 0.96363579 0.98213494
##                           Comp.6     Comp.7
## Standard deviation     0.27542160 0.22180708
## Proportion of Variance 0.01083672 0.00702834
## Cumulative Proportion  0.99297166 1.00000000
```

# Análise por Componentes Principais

**Escores**

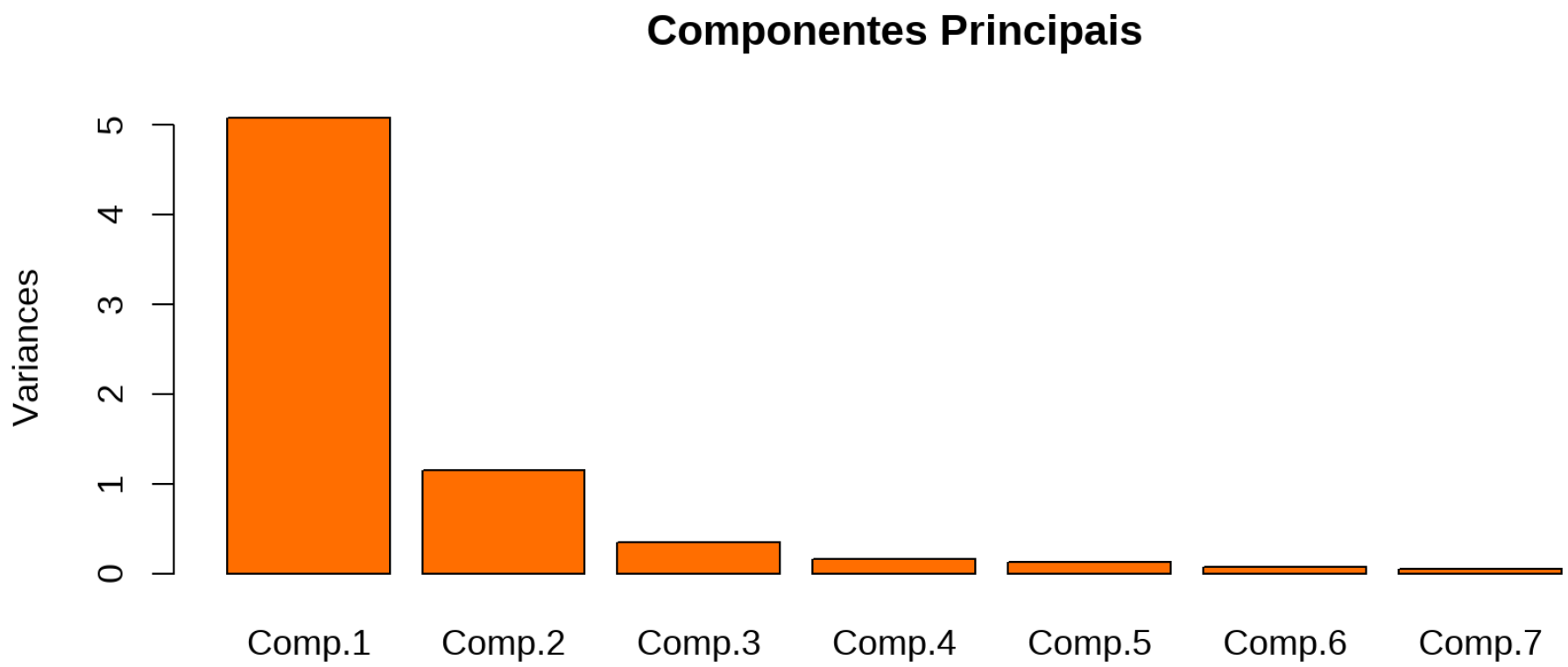
```
PCAdf$scores[,1:2]|>round(3)|> DT::datatable(options = list(pageLength = 6))
```

Show 6 entries Search:

| | Comp.1 | Comp.2 |
|---|---|---|
| Mazda RX4 | 0.809 | 0.919 |
| Mazda RX4 Wag | 0.781 | 0.595 |
| Datsun 710 | 2.079 | -0.053 |
| Hornet 4 Drive | 0.146 | -1.309 |
| Hornet Sportabout | -1.63 | 0.013 |
| Valiant | -0.135 | -2.045 |

Showing 1 to 6 of 32 entries Previous 1 2 3 4 5 6 Next

# Análise por Componentes Principais



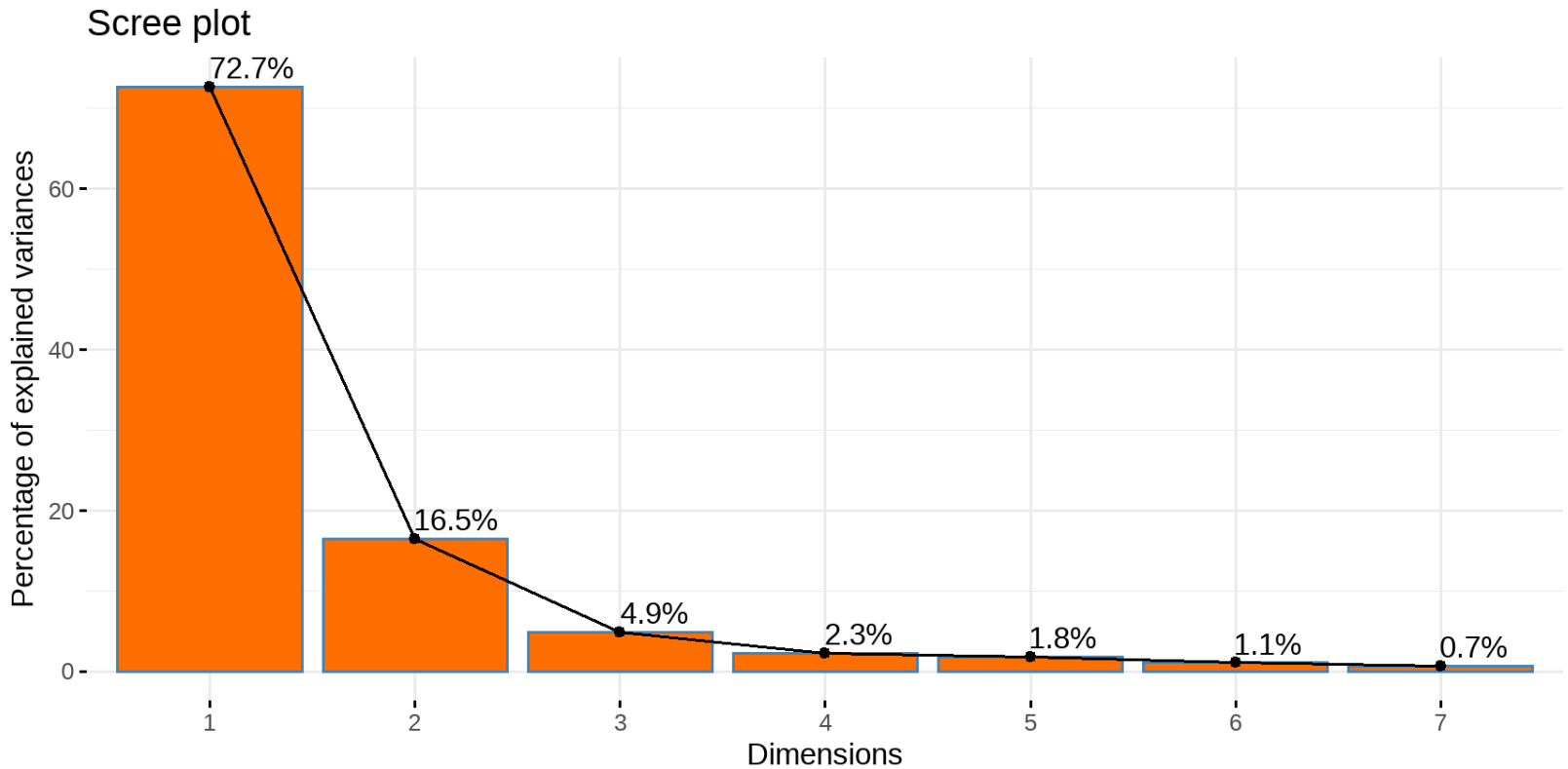
# Análise por Componentes Principais

# Componentes Principais - Pacote `FactoMineR`

```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 32 individuals, described by 7 variables
## *The results are available in the following objects:
##
##    name                description
## 1  "$eig"              "eigenvalues"
## 2  "$var"              "results for the variables"
## 3  "$var$coord"        "coord. for the variables"
## 4  "$var$cor"          "correlations variables - dimensions"
## 5  "$var$cos2"         "cos2 for the variables"
## 6  "$var$contrib"      "contributions of the variables"
## 7  "$ind"              "results for the individuals"
## 8  "$ind$coord"        "coord. for the individuals"
## 9  "$ind$cos2"         "cos2 for the individuals"
## 10 "$ind$contrib"      "contributions of the individuals"
## 11 "$call"             "summary statistics"
## 12 "$call$centre"      "mean of the variables"
## 13 "$call$ecart.type" "standard error of the variables"
## 14 "$call$row.w"       "weights for the individuals"
## 15 "$call$col.w"       "weights for the variables"
```

# Componentes Principais - Pacote `FactoMineR`

```
factoextra::fviz_eig(PCAdf1, addlabels = TRUE, barfill = "#ff6e00")
```



# Análise por Componentes Principais - Interpretação

**Quantos componentes podemos usar?**

- Analisando o percentual de variância explicada, vemos que o primeiro componente explica `72,65%` da variância, o já é um índice aceitável para análise, e o segundo componente explica `16,56%`. Somados estes dois componentes temos um total de `94,10%` da variância explicada. Dessa forma, podemos definir o uso de 2 componentes principais.

**Quais as possíveis interpretações**

- Podemos entender que através da componente 1 que há uma influencia positiva e negativa de acordo com a variável escolhida, entretanto não há nenhuma com peso muito alto, maior está em cilindos num valor de `-0.413`. Na segunda compenente, temos o maior valor de influencia, de `- 0.749` para velocidade.

# Análise Fatorial

Este método aborda o problema de analisar a estrutura das inter-relações (correlações) entre um grande número de variáveis (escores de testes, itens de testes, respostas de questionários), definindo um conjunto de dimensões latentes comuns, chamados fatores. Então, a análise fatorial, permite primeiro identificar as dimensões separadas da estrutura e então determinar o grau em que cada variável é explicada por cada dimensão. Uma vez que essas dimensões e a explicação da cada variável estejam determinadas, os dois principais usos da análise fatorial podem ser conseguidos:

- **Resumo**: ao resumir os dados, a análise fatorial obtém dimensões latentes que, quando interpretadas e compreendidas, descrevem os dados em um número muito menor de conceitos do que as variáveis individuais originais.

- **Redução de dados**: pode ser obtida calculando escores para cada dimensão latente e substituindo as variáveis originais pelos mesmos.

Fonte: https://smolski.github.io/livroavancado/analisf.html

# Análise Fatorial

- Comunalidade - Proporção de variabilidade de cada variável que é explicada pelos fatores.
- Especifidade - o erro ou parcela da variância que não pode ser explicada pelos fatores

```
L = cbind(sqrt(autos$values[1])*autos$vectors[
    sqrt(autos$values[2])*autos$vectors[,2])
aux = L%*%t(L);aux|>round(3)|>knitr::kable()
```

| | | | | | | |
|---|---|---|---|---|---|---|
| 0.874 | -0.884 | -0.895 | -0.782 | 0.738 | -0.852 | 0.432 |
| -0.884 | 0.925 | 0.905 | 0.868 | -0.675 | 0.816 | -0.586 |
| -0.895 | 0.905 | 0.916 | 0.801 | -0.754 | 0.872 | -0.444 |
| -0.782 | 0.868 | 0.801 | 0.896 | -0.478 | 0.646 | -0.765 |
| 0.738 | -0.675 | -0.754 | -0.478 | 0.791 | -0.826 | 0.015 |
| -0.852 | 0.816 | 0.872 | 0.646 | -0.826 | 0.899 | -0.198 |
| 0.432 | -0.586 | -0.444 | -0.765 | 0.015 | -0.198 | 0.942 |

```
aux = L%*%t(L)
qsi = diag(diag(c - aux));qsi|>round(3)|>
knitr::kable()
```

| | | | | | | |
|---|---|---|---|---|---|---|
| 0.126 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0.000 | 0.075 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0.000 | 0.000 | 0.084 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0.000 | 0.000 | 0.000 | 0.104 | 0.000 | 0.000 | 0.000 |
| 0.000 | 0.000 | 0.000 | 0.000 | 0.209 | 0.000 | 0.000 |
| 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.101 | 0.000 |
| 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.058 |

```
## [1] 0.7573346
```

# Análise Fatorial

## Correlação estimada

```
cor_est = aux + qsi; #cor_est|>round(3)|>knitr::kable()

cor_est|>round(3)|>knitr::kable()
```

| | | | | | | |
|---|---|---|---|---|---|---|
| 1.000 | -0.884 | -0.895 | -0.782 | 0.738 | -0.852 | 0.432 |
| -0.884 | 1.000 | 0.905 | 0.868 | -0.675 | 0.816 | -0.586 |
| -0.895 | 0.905 | 1.000 | 0.801 | -0.754 | 0.872 | -0.444 |
| -0.782 | 0.868 | 0.801 | 1.000 | -0.478 | 0.646 | -0.765 |
| 0.738 | -0.675 | -0.754 | -0.478 | 1.000 | -0.826 | 0.015 |
| -0.852 | 0.816 | 0.872 | 0.646 | -0.826 | 1.000 | -0.198 |
| 0.432 | -0.586 | -0.444 | -0.765 | 0.015 | -0.198 | 1.000 |

# Análise Fatorial

**Escores**

```
esc = t(t(L)%*%t(df));esc|>round(3) |> DT::datatable(options = list(pageLength = 6))
```

Show 6 entries Search:

| | V1 | V2 |
|---|---|---|
| Mazda RX4 | -225.342 | -15.772 |
| Mazda RX4 Wag | -225.264 | -15.233 |
| Datsun 710 | -155.945 | -12.559 |
| Hornet 4 Drive | -317.873 | -4.121 |
| Hornet Sportabout | -477.791 | -20.47 |
| Valiant | -285.162 | -4.071 |

Showing 1 to 6 of 32 entries Previous 1 2 3 4 5 6 Next

# Análise Fatorial

- o critério VARIMAX se concentra na simplificação das colunas da matriz fatorial

```
AFdf = factanal(df, 3, scores = "regression", rotation = "varimax");AFdf$loadings
```

```
##
## Loadings:
##                   Factor1 Factor2 Factor3
## milhas_por_galao -0.718  -0.374  -0.437
## cilindros         0.528   0.569   0.596
## cilindrada        0.703   0.392   0.514
## cavalos_forca     0.536   0.711   0.249
## eixo             -0.435          -0.756
## peso              0.899   0.102   0.420
## velocidade                -0.934
##
##                 Factor1 Factor2 Factor3
## SS loadings       2.578   2.006   1.622
## Proportion Var    0.368   0.287   0.232
## Cumulative Var    0.368   0.655   0.886
```

# Análise Fatorial

```
AFdf$uniquenesses # variancia explicada
```

```
## milhas_por_galao         cilindros       cilindrada    cavalos_forca
##       0.15398115       0.04292221       0.08850617       0.14521043
##             eixo             peso       velocidade
##       0.23715006       0.00500000       0.12198292
```

```
apply(AFdf$loadings^2,1,sum) #comunalidade
```

```
## milhas_por_galao         cilindros       cilindrada    cavalos_forca
##        0.8460192        0.9570776        0.9114944        0.8547901
##             eixo             peso       velocidade
##        0.7628478        0.9950170        0.8780173
```

```
1 - apply(AFdf$loadings^2,1,sum) #especifidade
```

```
## milhas_por_galao         cilindros       cilindrada    cavalos_forca
##      0.153980812      0.042922429      0.088505596      0.145209884
##             eixo             peso       velocidade
##      0.237152215      0.004982954      0.121982659
```

# Análise Fatorial

- sem nenhum critério de rotatividade

```
AFdf2 = factanal(df, 2, scores = "regression", rotation = 'none');AFdf2$loadings
```

```
##
## Loadings:
##                   Factor1 Factor2
## milhas_por_galao -0.776  -0.490
## cilindros         0.687   0.656
## cilindrada        0.813   0.508
## cavalos_forca     0.498   0.757
## eixo             -0.768  -0.160
## peso              0.922   0.257
## velocidade               -0.994
##
##                 Factor1 Factor2
## SS loadings       3.432   2.580
## Proportion Var    0.490   0.369
## Cumulative Var    0.490   0.859
```

# Análise Fatorial

```
AFdf2$uniquenesses # variancia explicada
```

```
## milhas_por_galao         cilindros       cilindrada    cavalos_forca
##       0.15753638        0.09885065        0.08018598        0.17881319
##             eixo              peso        velocidade
##       0.38515603        0.08289216        0.00500000
```

```
apply(AFdf2$loadings^2,1,sum) #comunalidade
```

```
## milhas_por_galao         cilindros       cilindrada    cavalos_forca
##        0.8424632         0.9011493         0.9198137         0.8211877
##             eixo              peso        velocidade
##        0.6148465         0.9171078         0.9950017
```

```
1 - apply(AFdf2$loadings^2,1,sum) #especifidade
```

```
## milhas_por_galao         cilindros       cilindrada    cavalos_forca
##      0.157536755       0.098850736       0.080186253       0.178812272
##             eixo              peso        velocidade
##      0.385153549       0.082892153       0.004998266
```

# Obrigado!!!