Exemplos de Aplicação

Case:
Air Pollution in US Cities

Exemplo

- Estudo de poluição do ar em 41 cidades dos EUA
  - √ Ano: 1970
- Dados: *Usairpollution*{*MVA*}

- Variáveis:
  - √ SO2: conteúdo de dióxido de enxofre no ar, em $\mu g/m^3$.
  - √ Temp: temperatura média anual (ºF)
  - √ Indust: quantidade de empresas manufatureiras empregando pelo menos 20 empregados.
  - √ Pop: população (censo 1970), em milhares.
  - √ Vento: velocidade média anual de vento, em milhas/h
  - √ Precip: precipitação média anual, em polegadas
  - √ Dias: número médio anual de dias com precipitação

Variáveis

- √ 'Resposta': SO2
- √ Ambientais: Vento, Precip, Dias, Temp
- √ Demográficas: Indust, Pop

Comandos em R:

- √ Carregamento dos dados:

```
> library(MVA)
> library(HSAUR2, ADGofTest)# carrega os pacotes
> data(USairpollution)# carrega o banco de dados
> poluicao <- USairpollution
> colunas<- c("SO2", "Temp", "Indust", "Pop",
+ "Vento", "Precip", "Dias")
> colnames(poluicao)<- colunas
>
> poluicao$negtemp <- poluicao$Temp * (-1)
> poluicao$Temp <- NULL
```

- √ Comentários:
  - Extrair componentes da matriz de correlações:
    - Variáveis estão em escalas muito distintas
  - Ignora a variável SO2 (considera só as 'explicativas')
  - Uso da temperatura negativa
    - As 6 variáveis têm valores altos, de maneira que representam um ambiente menos atrativo

- √ Matriz de correlações:

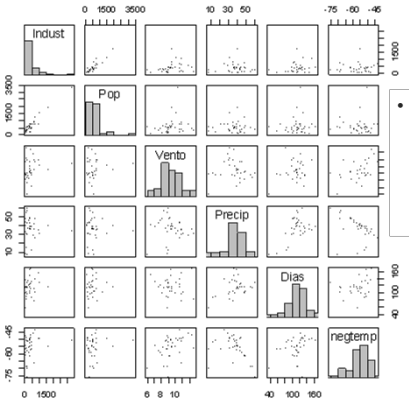
```
> # Descrição do conjunto de dados
>
> poluicao.corr <- cor(poluicao[,-1])
> poluicao.corr
              Indust         Pop       Vento      Precip       Dias     negtemp
Indust    1.00000000  0.95526935  0.23794683 -0.03241688 0.13182930  0.19004216
Pop       0.95526935  1.00000000  0.21264375 -0.02611873 0.04208319  0.06267813
Vento     0.23794683  0.21264375  1.00000000 -0.01299438 0.16410559  0.34973963
Precip   -0.03241688 -0.02611873 -0.01299438  1.00000000 0.49609671 -0.38625342
Dias      0.13182930  0.04208319  0.16410559  0.49609671 1.00000000  0.43024212
negtemp   0.19004216  0.06267813  0.34973963 -0.38625342 0.43024212  1.00000000
```

- √ Comentário:
  - Valores altos de correlação entre Indust e Pop

- √ Matrix plot:
  - Comandos em R:

```
> panel.hist <- function(x, ...){
+ usr <- par("usr"); on.exit(par(usr))
+ par(usr = c(usr[1:2], 0, 1.5) )
+ h <- hist(x, plot = FALSE)
+ breaks <- h$breaks; nB <- length(breaks)
+ y <- h$counts; y <- y/max(y)
+ rect(breaks[-nB], 0, breaks[-1], y, col = "grey", ...)
+     }
>
> pairs(poluicao[,-1], pch = ".", cex = 1.5)
> pairs(poluicao[,-1], diag.panel = panel.hist, pch = ".", cex = 1.5)
```

√ Gráfico matricial dos diagramas de dispersão



- Há pelo menos um outlier
  - √ Provavelmente mais de um
  - √ Chicago tem cerca do dobro de fábricas do que a 2ª. cidade com maior número

√ Análise de componentes principais:

```
> poluicao.acp <- princomp(poluicao[,-1], cor = TRUE)
> summary(poluicao.acp, loadings = TRUE)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
Standard deviation     1.4819456 1.2247218 1.1809526 0.8719099 0.33848287
Proportion of Variance 0.3660271 0.2499906 0.2324415 0.1267045 0.01909511
Cumulative Proportion  0.3660271 0.6160177 0.8484592 0.9751637 0.99425879
                          Comp.6
Standard deviation     0.185599752
Proportion of Variance 0.005741211
Cumulative Proportion  1.000000000
```

√ 3 primeiras componentes têm variâncias maiores que 1
- Compreendem 85% da variância total

√ Escores das componentes podem ser usados para visualização gráfica, com pouca perda de informação.

√ Interpretação das componentes principais:

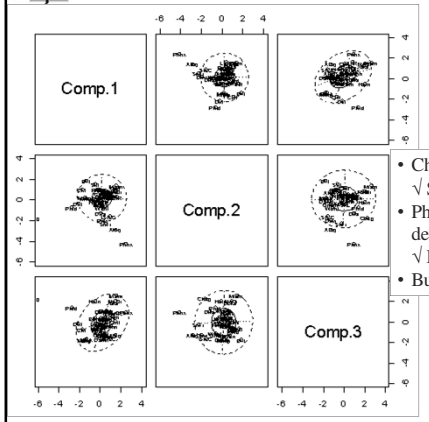
```
Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
Indust   -0.612  0.168 -0.273 -0.137  0.102  0.703
Pop      -0.578  0.222 -0.350                -0.695
Vento    -0.354 -0.131  0.297  0.869 -0.113
Precip          -0.623 -0.505  0.171  0.568
Dias     -0.238 -0.708         -0.311 -0.580
negtemp  -0.330 -0.128  0.672 -0.306  0.558 -0.136
>
```

√ Componente 1
- Relaciona-se com algum índice de qualidade de vida
- Altos valores indicam um ambiente com menos qualidade

√ Componente 2
- Coeficientes altos para Precip e Dias
- Relaciona-se com chuva (clima úmido)

√ Componente 3
- Contraste entre Precip e negtemp
- Separa cidades c/ altas temperaturas e muita chuva (tipo de clima)

√ Boxplot bivariado para as 3 primeiras componentes

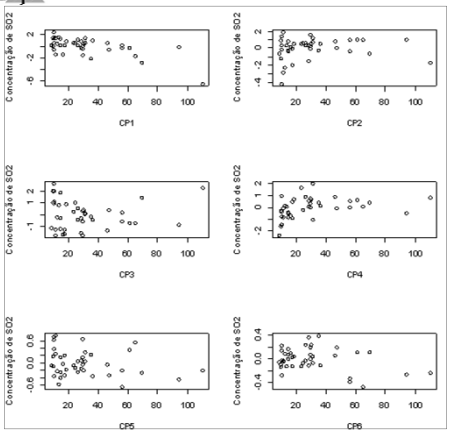


- Chicago é outlier
  - √ Suspeita-se de Phoenix e Philadelphia
- Phoenix aparenta ter a melhor qualidade de vida
  - √ Baseando-se nas 6 variáveis
- Bufallo tem clima mais seco

Questão Interessante

- Quais dentre as variáveis climáticas e ambientais são as melhores preditoras do grau de poluição do ar (concentração de SO_2)?
  - √ Esta questão é tratada com regressão linear múltipla
  - √ Potencial problema para aplicação dessa técnica:
    - Alta correlação entre Indust e Pop
  - √ Solução:
    - Retirar uma das variáveis
  - √ Alternativa:
    - Fazer regressão dos níveis de SO_2 com as componentes principais derivadas das 6 variáveis originais
    - Pode ser melhor regredir com todas as 6 componentes

- √ SO2 dependendo das componentes principais



- √ Regressão com as 6 componentes principais:

```
> poluicao.reg <- lm(SO2 ~ poluicao.acp$scores,
+  data = poluicao)
> summary(poluicao.reg)

Call:
lm(formula = SO2 ~ poluicao.acp$scores, data = poluicao)

Residuals:
    Min      1Q  Median      3Q     Max
-23.004  -8.542  -0.991   5.758  48.758

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                30.049      2.286  13.146 6.91e-15 ***
poluicao.acp$scoresComp.1  -9.942      1.542  -6.446 2.28e-07 ***
poluicao.acp$scoresComp.2   2.240      1.866   1.200  0.23845
poluicao.acp$scoresComp.3   0.375      1.935   0.194  0.84752
poluicao.acp$scoresComp.4   8.549      2.622   3.261  0.00253 **
poluicao.acp$scoresComp.5 -15.176      6.753  -2.247  0.03122 *
poluicao.acp$scoresComp.6 -39.271     12.316  -3.189  0.00306 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.64 on 34 degrees of freedom
Multiple R-squared:  0.6695,    Adjusted R-squared:  0.6112
F-statistic: 11.48 on 6 and 34 DF,  p-value: 5.419e-07
```

- √ Escores da 1ª. componente predizem mais a resposta
- √ Componentes com menor variância não têm necessariamente as menores correlações com a resposta

Fonte

- EVERITT, B.; HOTHORN, T. *An Introduction to Applied Multivariate Analysis with R*. Springer, 2011.
  - √ Sec. 3.10.3: Air pollution in US cities, pg. 86.