FIAP

MBA+

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

FIAP

MBA+

PROGRAMANDO IA COM R

Prof. Elthon Manhas de Freitas
elthon@usp.br

2018

Revisão da última aula

- O que vimos na aula passada?

Plots

Análises Gráficas com o R

Por onde começar

- Você já conhece minimamente seus dados?
- Já sabe os agrupamentos principais?
- Quais perguntas quer responder?
- Que respostas ou idéias quer passar?

Como estar no controle?

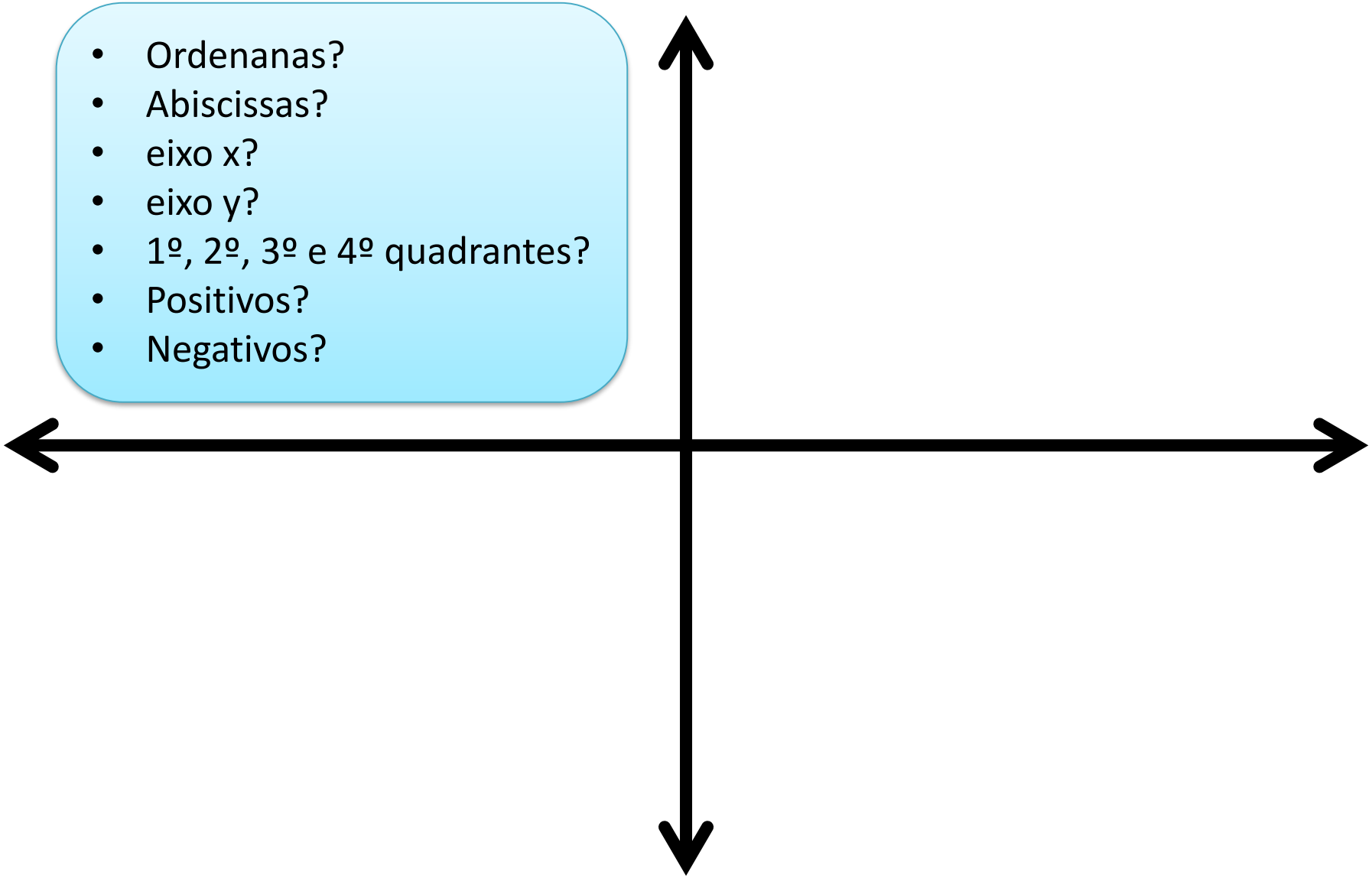
- Inteligência artificial não é uma máquina de moer carne.

Como se preparar

- Antes de obter respostas é preciso fazer perguntas.
- Para saber o que perguntar é preciso:
  - Conhecer o problema
  - Conhecer os dados disponíveis
- Que tal começar a esboçar o que você quer em um papel?

Plano Cartesiano

- Ordenanas?
- Abiscissas?
- eixo x?
- eixo y?
- 1º, 2º, 3º e 4º quadrantes?
- Positivos?
- Negativos?



Pacotes de plotagens mais conhecidos

FIAP

- Base Plot (pacote graphics)
  - Core do R, possui os comandos básicos
- lattice
  - Permite criar o plot em uma única função, ótimo para sub-plots
- ggplot2
  - Mistura elementos do core com do lattice. Este é o pacote mais usado
- plotly
  - Pacote mais novo, com gráficos interativos em HTML.
  - Possui versão para python, matlab, node.js, etc.
- plot3D
  - advinha

Base Plot – parabola quadrática

- Que tal explorar um pouco o plot básico

```
x =  -10:10
plot( x = x, y = x**2, main = Parábola')
```

Qual plot esperado?

Base Plot – parabola quadrática

FIAP

- Que tal explorar um pouco o plot básico

```
x =  -10:10
plot( x = x, y = x**2, main = Parábola')
```

- Experimente também com os parâmetros:
  - type = ‘l’
  - type = ‘o’
  - type = ‘p’
  - type = ‘h’
  - type = ‘b’
  - type = ‘s’

RMarkdown

Plot básico – Análise de tendência

FIAP

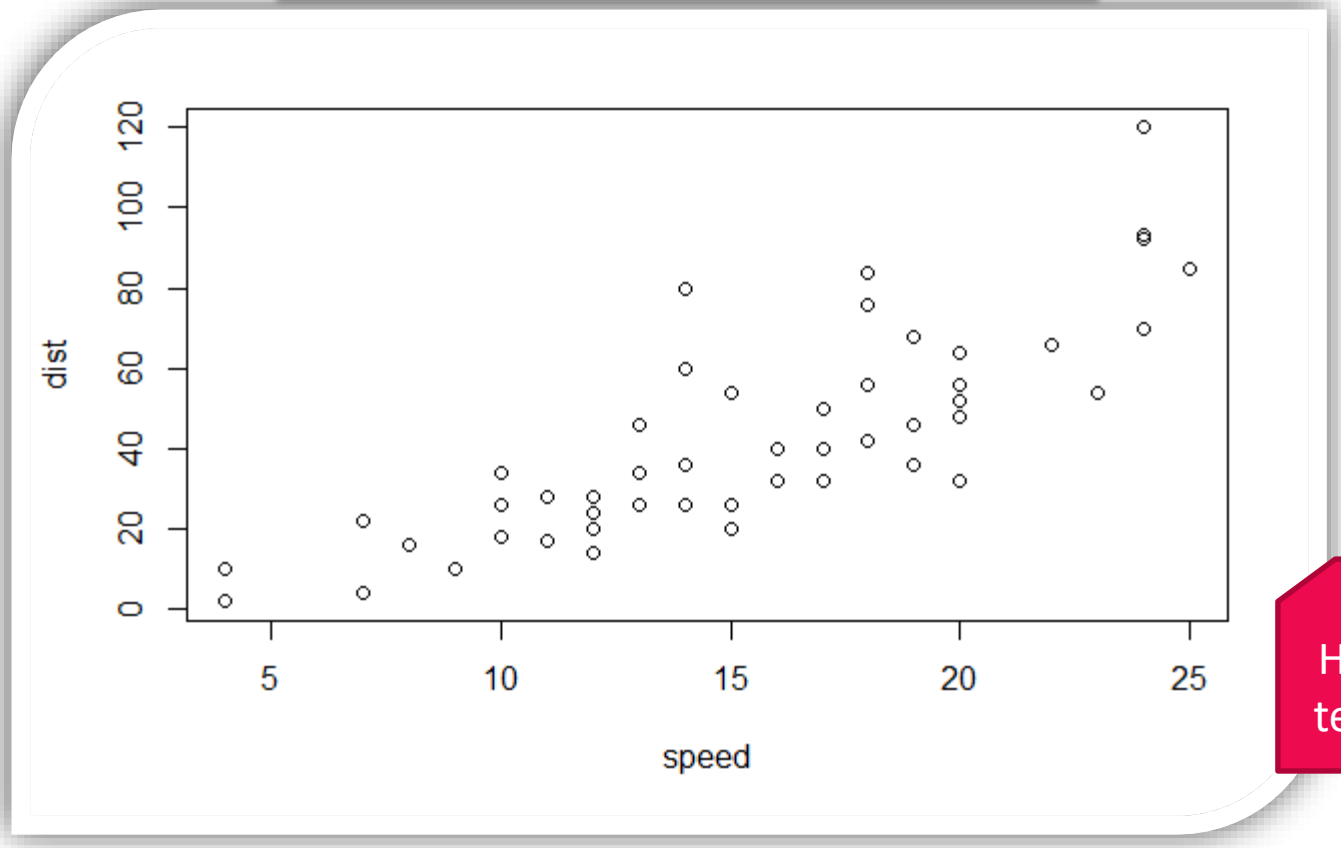
- Vamos explorar o dataset ‘cars’
  - Dataset gerado em 1920
  - 50 observações
  - Data uma velocidade registrada (mph) qual a distância percorrida até que o carro pare (ft)

```
head(cars)
summary(cars)
plot( x= cars$speed, y = cars$dist)
```

Plot básico – Análise de tendência

- Plot simples de dados nos eixos x e y

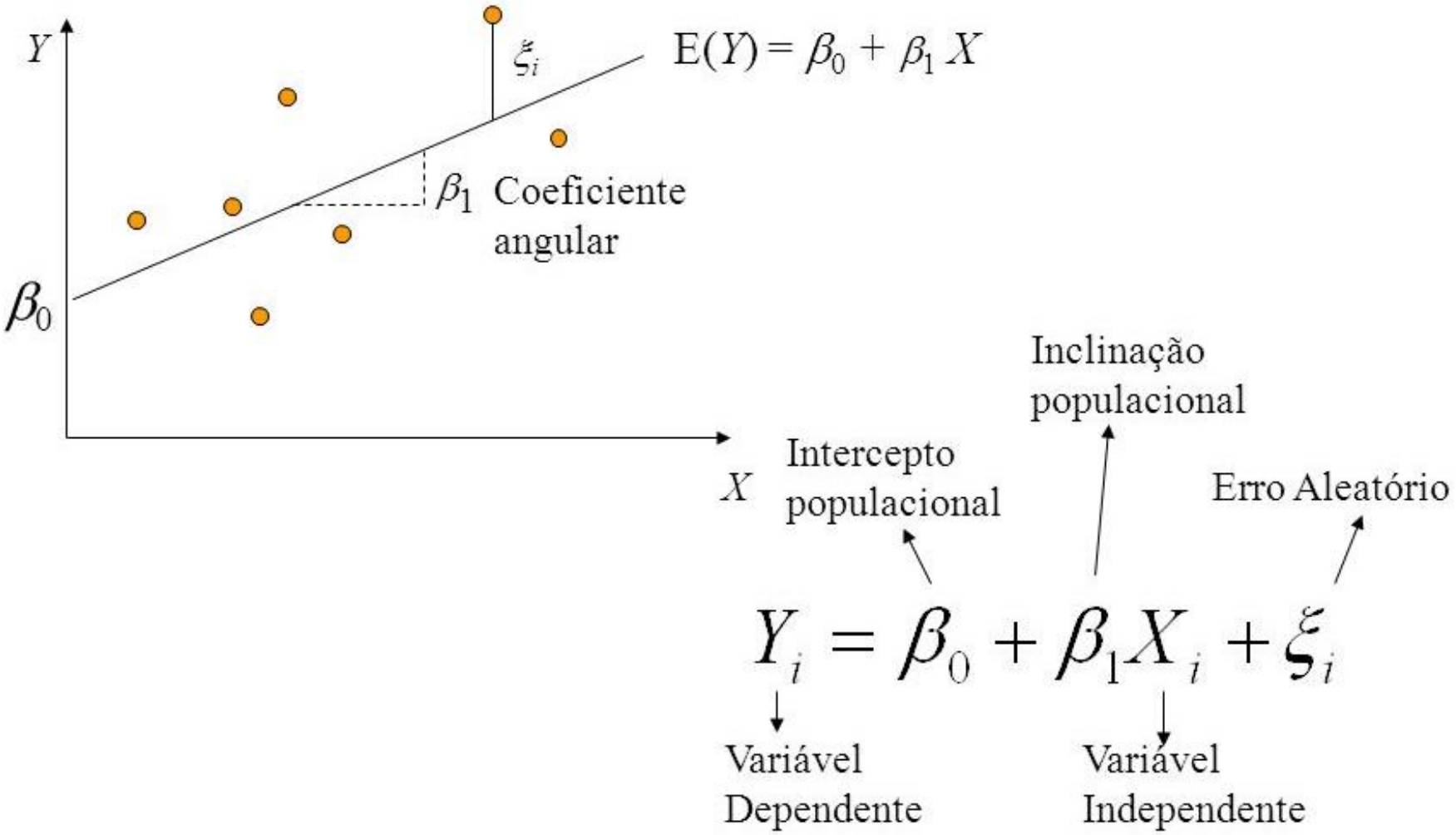
```
plot(cars$speed, cars$dist)
ou
plot(cars)
```



Há alguma tendência?

Rapidinha: Regressão Linear

* Será visto com mais detalhes nas aulas de Estatística



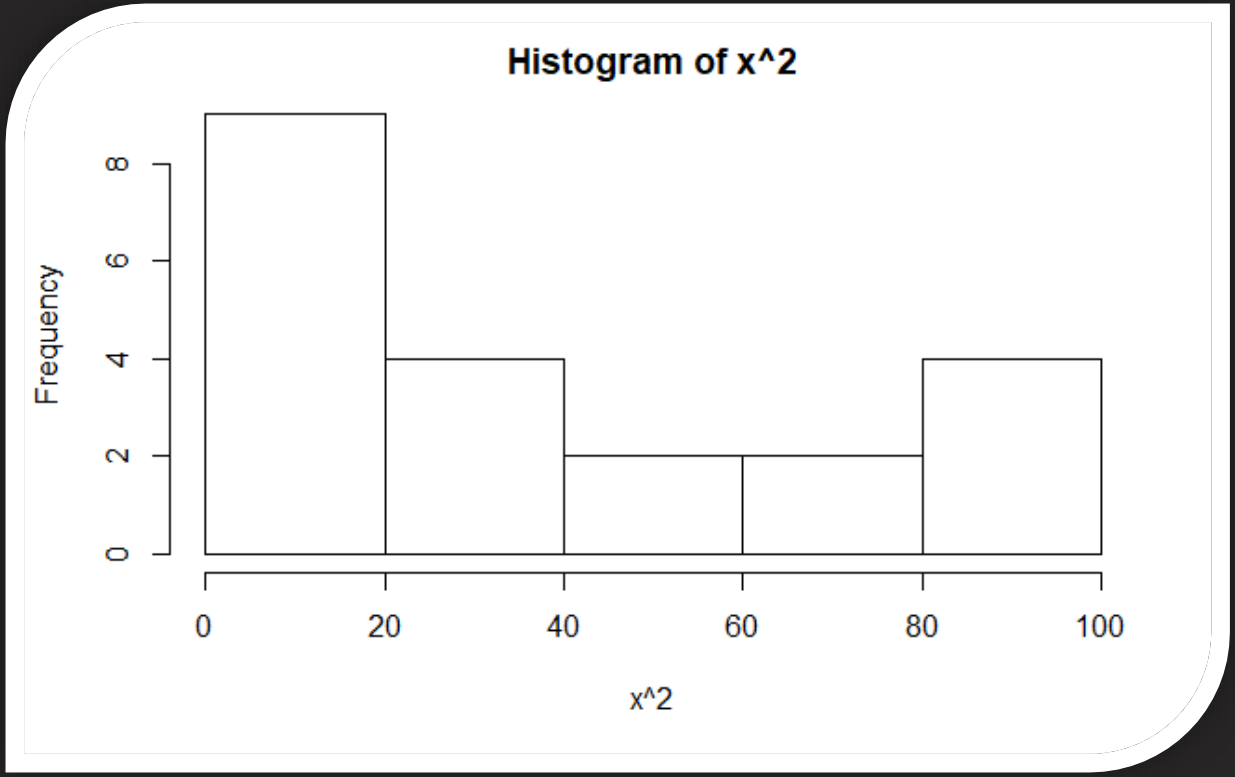
O sistema de plotagem

FIAP

- A área de plotagem é criada em uma única função (`plot, hist, boxplot, etc.`).
- Esta função cria uma nova área limpa e insere o plot.
- A partir daí, é possível manipular o plot com outros comandos (`lines`, por exemplo)
  - Um novo plot cria uma nova área de plotagem.

Plot Básico – Histograma

- O que é um histograma?
- Para que serve?



- Freqüencia × Densidade

Plot Básico – Caixas (Box Plot)

- Não confundir com Velas (Candle)
  - `summary(airquality)`

```
     Ozone           Solar.R          Wind             Temp           Month
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00   Min.
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00   1st Q
 Median : 31.50   Median :205.0   Median : 9.700   Median :79.00   Media
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88   Mean
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00   3rd Q
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00   Max.
 NA's   :37       NA's   :7
```

airquality:

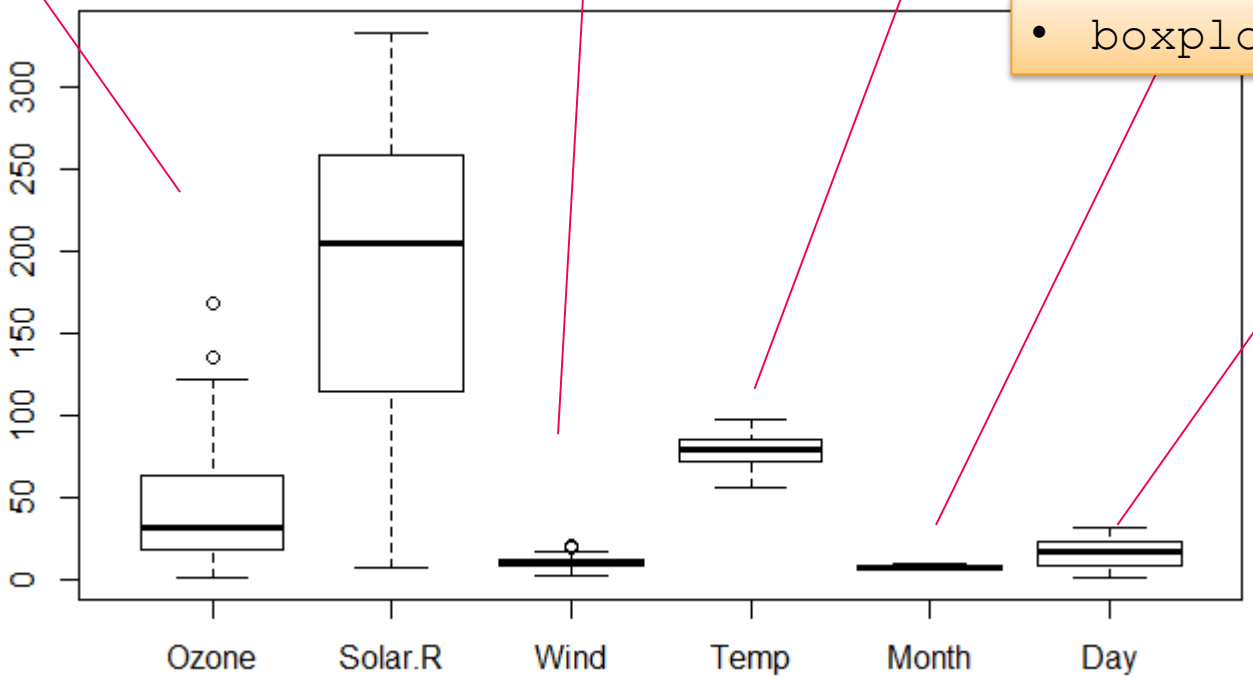
Medição diária da qualidade do ar de NY
De maio a setembro de 1973

Plot Básico – Caixas (Box Plot)

- `summary(airquality)`

```
 Ozone           Solar.R          Wind             Temp           Month            Day
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00   Min.   :5.000   Min.   : 1.0
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00   1st Qu.:6.000   1st Qu.: 8.0
 Median : 31.50   Median :205.0   Median : 9.700   Median :79.00   Median :7.000   Median :16.0
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88   Mean   :6.993   Mean   :15.8
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00   3rd Qu.:8.000   3rd Qu.:23.0
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00   Max.   :9.000   Max.   :31.0
 NA's   :37       NA's   :7
```

- `boxplot( airquality )`



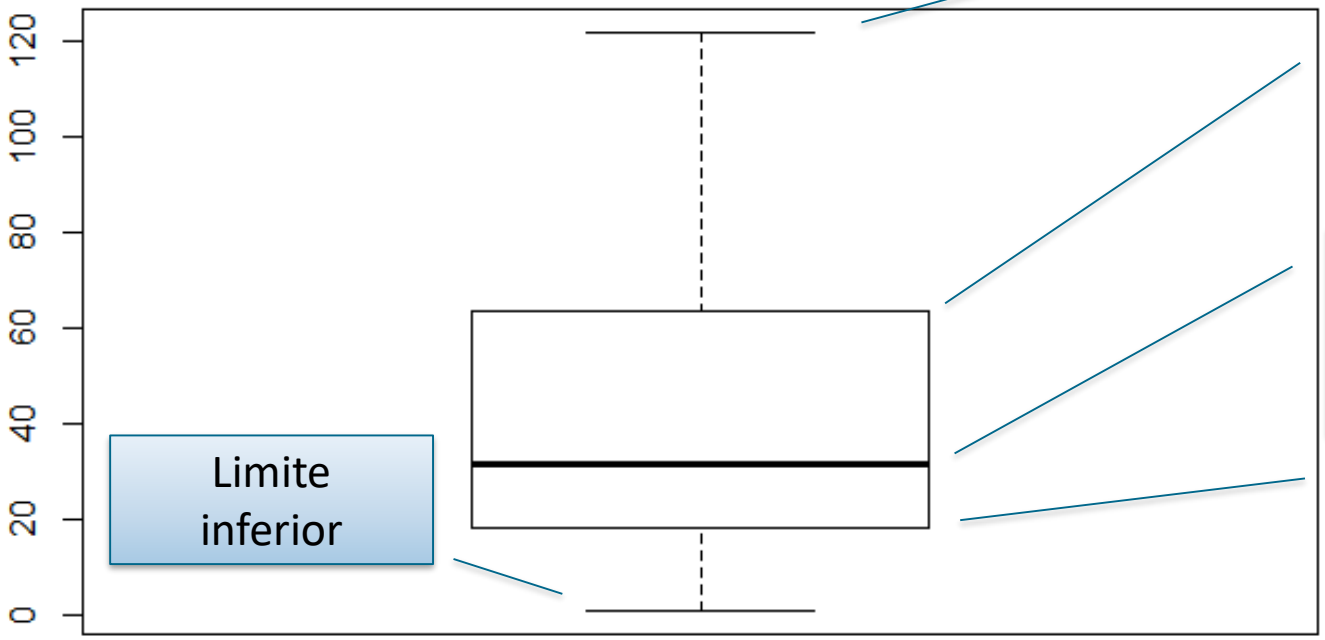
Plot Básico – Caixas (Box Plot)

- `summary( airquality$Ozone )`

```
Min.    :  1.00
1st Qu.: 18.00
Median : 31.50
Mean   : 42.13
3rd Qu.: 63.25
Max.   :168.00
NA's   :37
```

?

```
boxplot( airquality$Ozone, outline = F )
```



Limite superior

Terceiro quartil Q_3

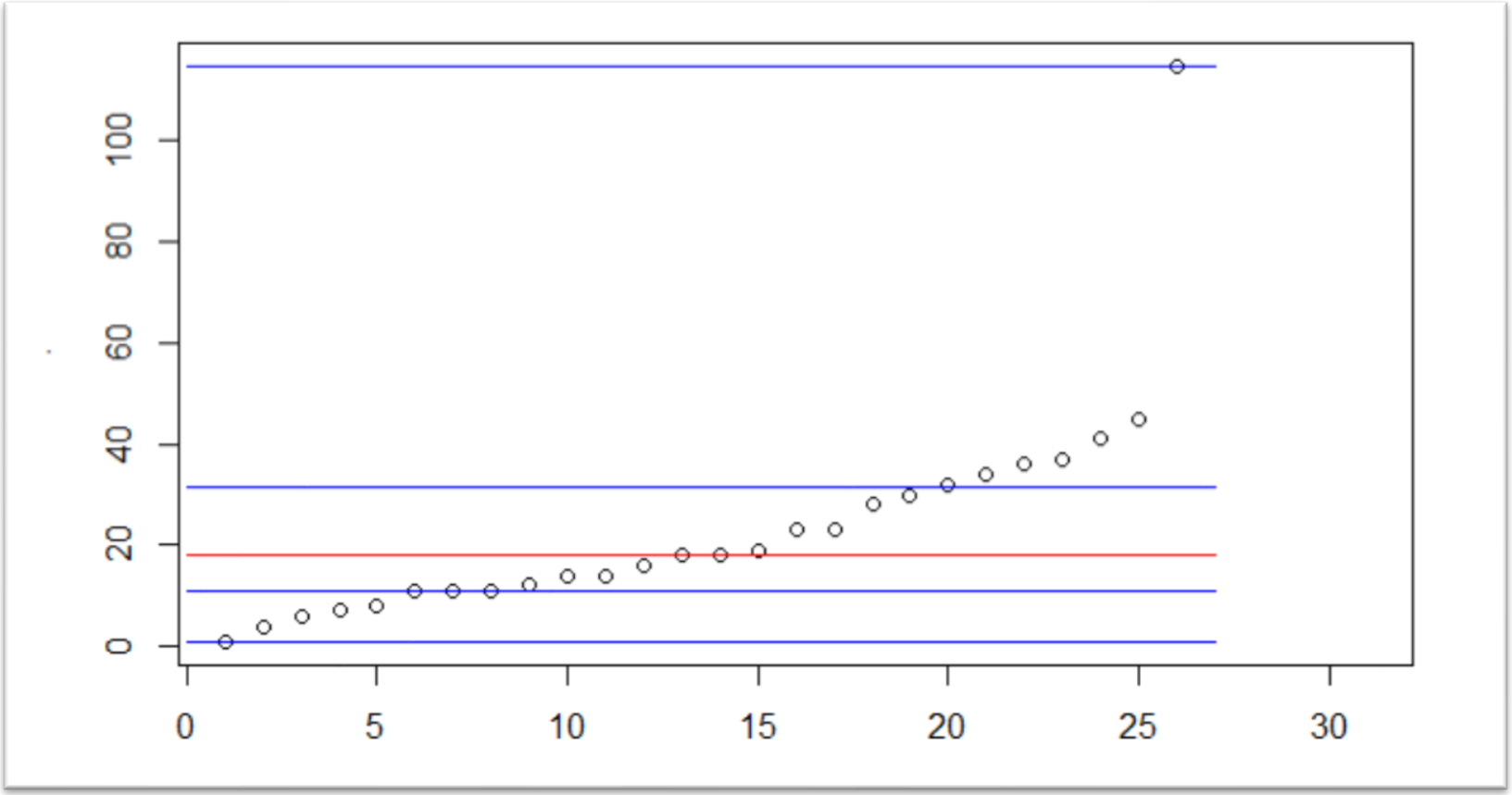
Mediana (segundo quartil) Q_2

Primeiro quartil Q_1

Limite inferior

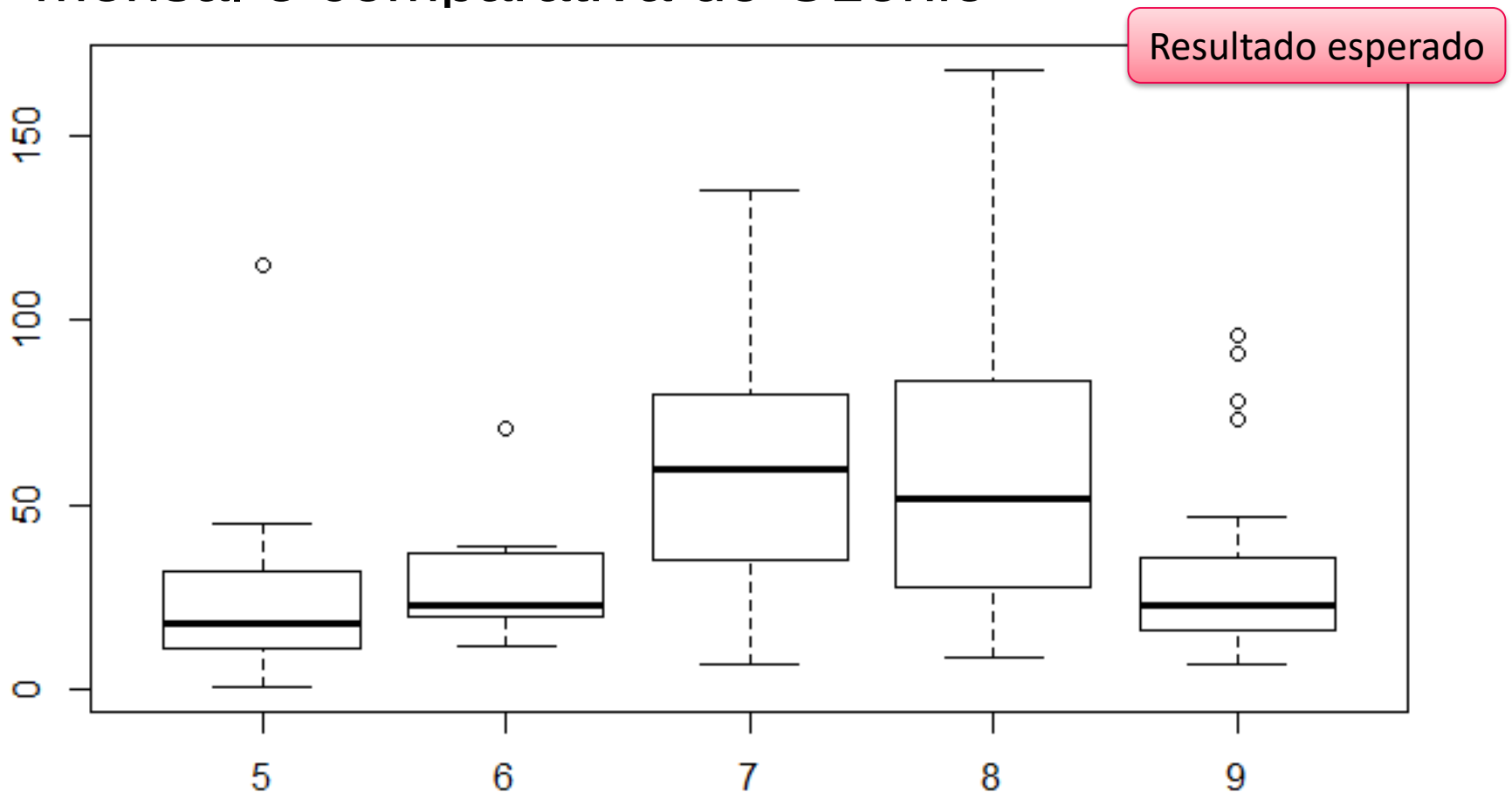
Exercício – Basic Plot

- Dataset airquality, medição de ozônio em **MAIO**
- Usando os comandos básicos de plotagem, faça o gráfico abaixo:



Exercício – BoxPlot

- Existem apenas 5 meses no dataset Airquality
- Fazer um boxplot que permita uma avaliação mensal e comparativa do Ozônio



Responda

- Qual mês possui maior mediana de O_3?
- Qual mês possui maior concentração de O_3?
- Qual mês apresenta a menor variância de O_3?
- E a maior?

- Qual o mês com a maior temperatura média?
- E qual tem a maior temperatura registrada?

- O mês com mais ventos é o mesmo mês que possui mais radiação solar?

Relação entre variáveis

FIAP

- Use o Plot Básico e responda:
- Há alguma relação aparente entre concentração de ozônio e vento?
  - Façam regressão linear com lm
- Desenhem a linha com lines()
  - Depois desenhem a mesma linha com a função abline.

Múltiplos gráficos

FIAP

- Experimente digitar os seguintes comandos, um por vez, no console:

```
par(mfrow=c(1,2))
plot(airquality$Wind, airquality$Ozone)
plot(airquality$Solar.R, airquality$Ozone)
```

- O que acontece?

Entendendo o escopo

- Os escopos, no R são similares ao das demais linguagens, entretanto são organizados em “Environments”

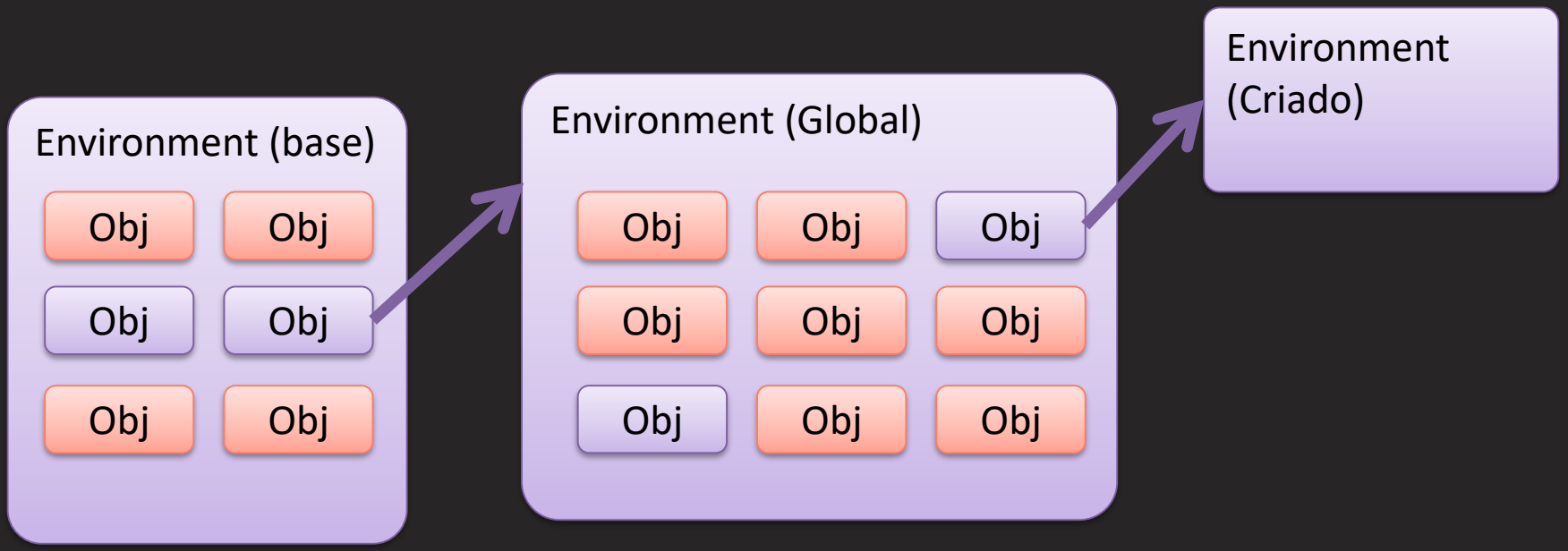
- Sua estrutura é praticamente a estrutura de uma lista.

- Todas as as variáveis são armazenadas em algum “Environment”
  - a raiz de tudo é o “Base Environment”, pois o cada environment também é uma variável. Não é aconselhável alterar nada no “Base Environment”,
  - O “Global Environment” é o nosso ponto de partida, seria nossa área de trabalho principal

Environments

- Environments notáveis

```
• .GlobalEnv
• globalenv()

• emptyenv()
• baseenv()
```



Environments

FIAP

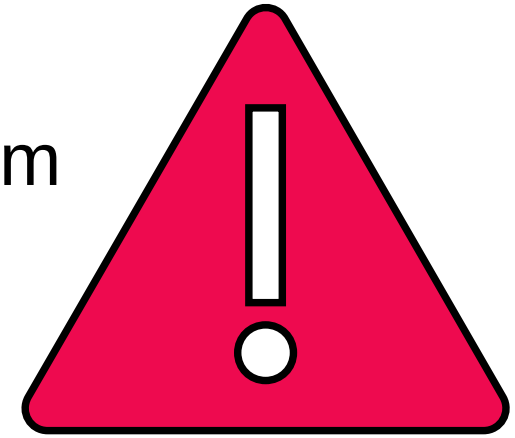
- Atribuição de variáveis entre Environments
  - Façam a seguinte execução e avaliem o resultado

```
e1 = new.env()

assign("var1", 1, envir = e1 )

get("var1", envir = e1)

exists("var1", envir = e1)

rm("var1", envir = e1)

exists("var1", envir = e1)
```

  - Onde está o environment e1?

Environment em funções

FIAP



- As funções em R sempre passam parâmetro por valor, nunca por referência.

- Cada execução de uma função cria um novo environment ligado à execução.

- Para compartilhar resultados entre funções é necessário compartilhar o environment

Exercício

- Criar uma função que:
  - Obtém o environment atual
  - Obtém o Global Environment
  - Imprima o environment atual
  - Imprima o Global Environment

- Executar esta função 5 vezes
  - O endereço do Environment foi o mesmo em todas as execuções?

Environment em funções

- Como falamos, o GlobalEnvironment é usado como área de trabalho dos programas R.

- O que fazem os seguintes operadores de atribuição?

  - <<-
  - ->>

(explicação em lousa)

Criando um environment com o `with`

FIAP

- Comando with
  - Cria um environment temporário
  - Desempacotas as variáveis de um objeto

```
• ls(envir = environment())

• with(mtcars, ls(envir = environment()) )

• with(cars, plot(speed, dist))
```

Mesmo plot, usando o with

FIAP

- Não é necessário digitar airquality

- Assim como o pipe %>%, faz parte das funções úteis do R.

```
par(mfrow = c(1,2))
with(airquality, {
  plot(Wind, Ozone,
       main = 'Ozonio pelo vento')
  plot(Solar.R, Ozone,
       main = 'Ozonio pela radiação solar')
})
```

LATTICE

- Mais usado para plotar Gráficos de tendência entre variáveis X e Y.

- Agrupa facilmente análises a partir de uma terceira variável Z

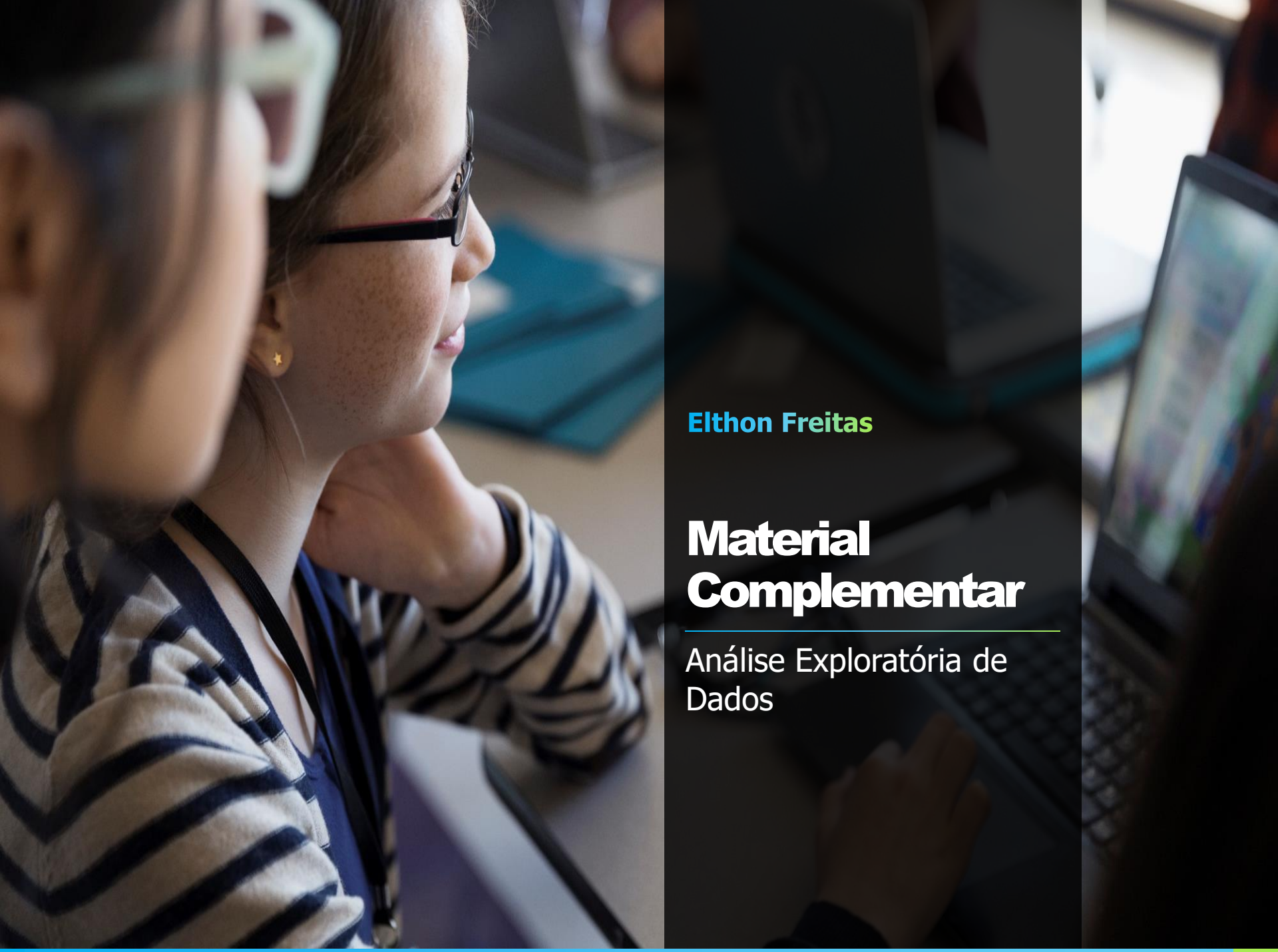
LATTICE - Básico

```
state <- data.frame(state.x77,
                    region = state.region)

xyplot(Life.Exp ~ Income | region,
       data = state,
       layout = c(4, 1))
```

- Qual gráfico resultante?
- Podemos dizer que há variação por estado quanto à relação de expectativa de vida e salário anual?
- A correlação entre assassinato e salário é positiva ou negativa?

Próxima aula: Apresentação, análise das EDA’s

MBA+



Elthon Freitas

Material Complementar

Análise Exploratória de Dados

Exploratory Data Analysis – EDA

Section 1 - Análise individual de campos

Dicionário de dados

- Não possui ainda um dicionário de dados?
- Eleger os identificadores únicos

Identificar série temporal

- O dataset possui série(s) temporal(is)?
- Qual a principal?
- Possíveis agregações

Identificar detalhes de cada variável

- Categórico vs Contínuo
- Volumetria de cada dado categórico
- Tipo de cada variável
- Concentração / distribuição
- Variância, média, etc.
- Outliers

Missing Values

- Quais os missing Values?
- NaN ou NaN+0
- Qual a quantidade de missing values?
- Determinar regra de preenchimento

Exploratory Data Analysis – EDA

Section 2 - Análise de dependência de dados

Correlação

- BE-A-BÁ:
- Tenha a matriz de correlação na ponta dos dedos
- Mapa de calor da matriz de correlação

Relacionamento

- Identificar chaves
- Identificar volumetria para falhas de relacionamento

Normalização

- Identificar melhor normalização que considere todo o dataset.

Transformação

- Identificar possíveis transformações para as variáveis
  - Lag / delay
  - Power / Log

Exploratory Data Analysis – EDA

Section 3 - Análise temporal de dados

Agregação

- Identificar o comportamento das agregações ao longo do tempo

Distribuição

- Identificar comportamentos de distribuição na linha do tempo

Tendências

- Quais as linhas de tendência temporal?
- Tendências lineares ou polinomiais?
- Quais as possíveis fórmulas de tendência temporal?

Missing Values

- Quais os missing Values?
- NaN ou NaN+0
- Qual a quantidade de missing values?
- Determinar regra de preenchimento

EDA + Data Manipulation

Section 4 - Conclusão

- Resumo do que foi encontrado;
- Principais insights identificados
- Pode ser integrado com data manipulation