

# Análise Preditiva

Gabriel Vinícius Araújo Fonseca

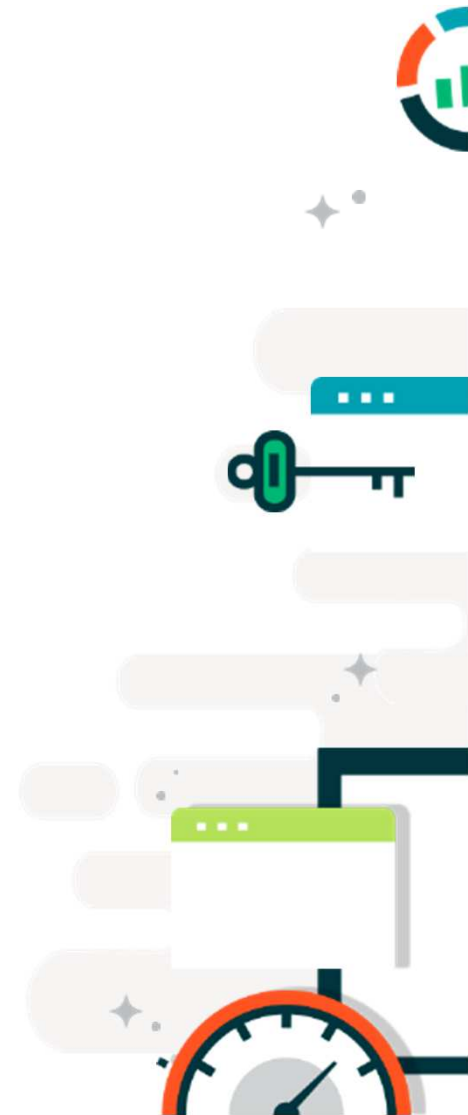


# **Apresentação da Disciplina**



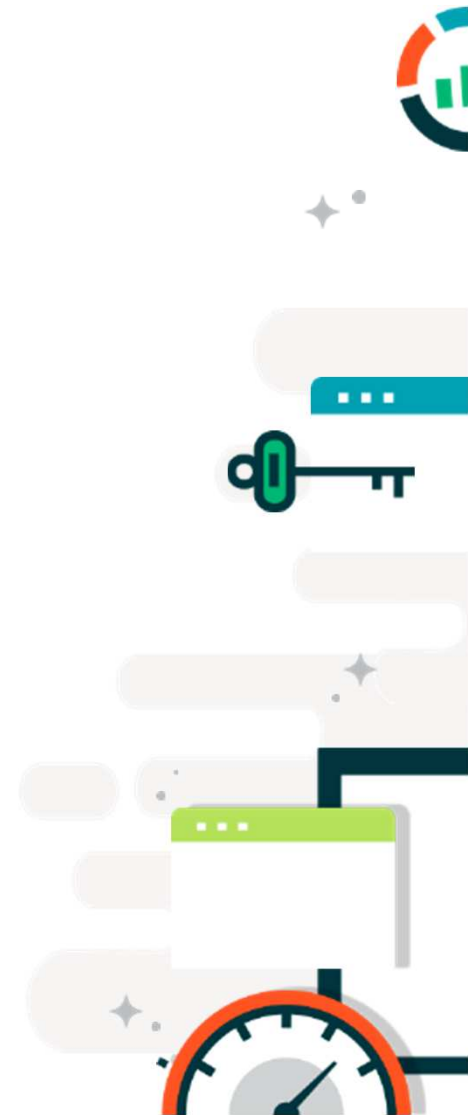
# Professor Gabriel

- Mestre em Estatística UFMG (2009)
- Graduação em Estatística UFMG (2006)
- Professor no Pitágoras, PUC e UNA (desde 2009).
- Analista de Dados – Grupo ANIMA Educação.
- Consultor em Análise de Dados.
- E-mail: [gabriel.fonseca@oi.com.br](mailto:gabriel.fonseca@oi.com.br)



# Aspectos gerais da disciplina

- Modelos Preditivos e tipos de análise
- Abordagens para análise preditiva
- Preparação da base
- Regressão Linear Simples e Múltipla
- Regressão Logística Simples e Múltipla
- Análise Multivariada de Dados
- Análise de Séries Temporais



# Análise Preditiva

O que é?

Por que usar?

O que você precisa para usar?

Ferramentas?



# Aplicações

- **Detecção de fraude e segurança:** diminuição de perdas ocorridas por atividades fraudulentas.
- **Marketing:** atrair, reter e desenvolver os clientes mais rentáveis e maximizar seus gastos.
- **Operações:** prever o estoque e gerenciar os recursos das fábricas.
- **Risco:** pontuação de crédito; probabilidade de inadimplência.



# Modelos Preditivos

- Análises Multivariadas de Dados
- Séries Temporais
- Regressão Linear
- Regressão Múltipla
- Regressão Logística
- Redes Neurais
- Árvores de decisão
- Algoritmos



**Prontos para a próxima  
etapa?**







# Análise Preditiva

Gabriel Vinícius Araújo Fonseca



# Estatística e suas Áreas



# Áreas da Estatística

## ■ Estatística Descritiva

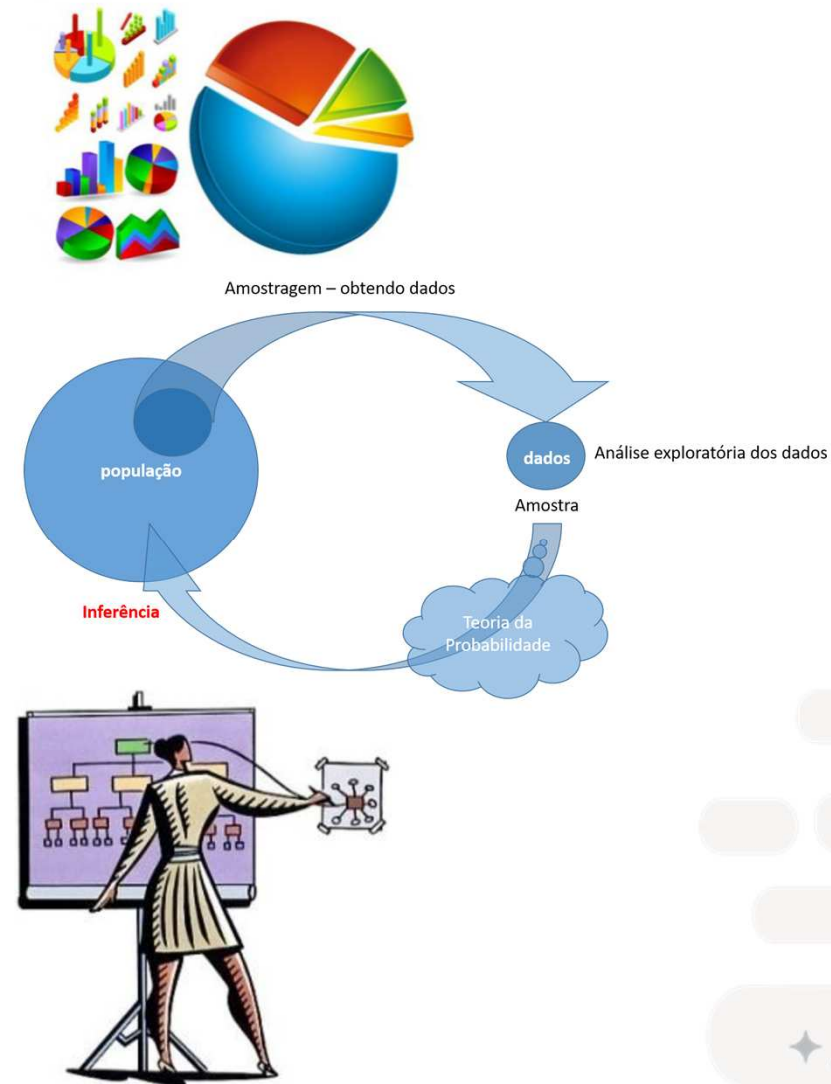
- Descrição, Exploração e Comparação de Dados: Medidas de **Tendência Central**, Medidas de **Variabilidade**, Medidas de **Associação**.

## ■ Estatística Inferencial

- Distribuição de **Probabilidade**, Teste de **Hipóteses** e Intervalos de **Confiança**.

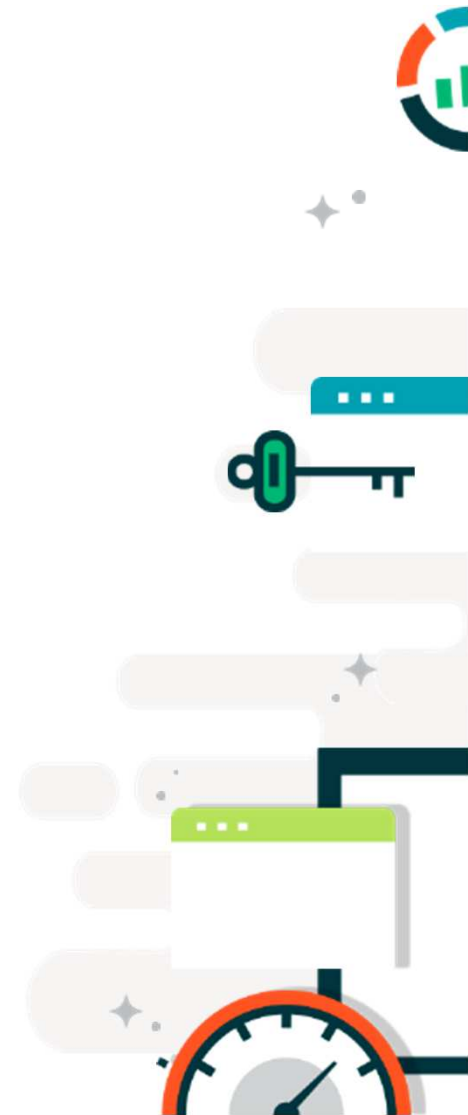
## ■ Modelagem Estatística

- Análise de Regressão, Séries Temporais e Estatística Multivariada.



# Modelos Preditivos – O que veremos?

- Regressão Linear Simples
- Regressão Linear Múltipla
- Regressão Logística
- Análises Multivariada
- Árvores de decisão
- Séries temporais



# Unidade 1

- Regressão Linear Simples
  - Diagrama de Dispersão
  - Coeficiente de Correlação Linear
  - Estimação do Modelo
  - Verificação do Ajuste
- Regressão Linear Múltipla
  - Matriz de Correlação
  - Ajuste do Modelo
  - Encontrar o Melhor Modelo
  - Verificação da Qualidade do Modelo



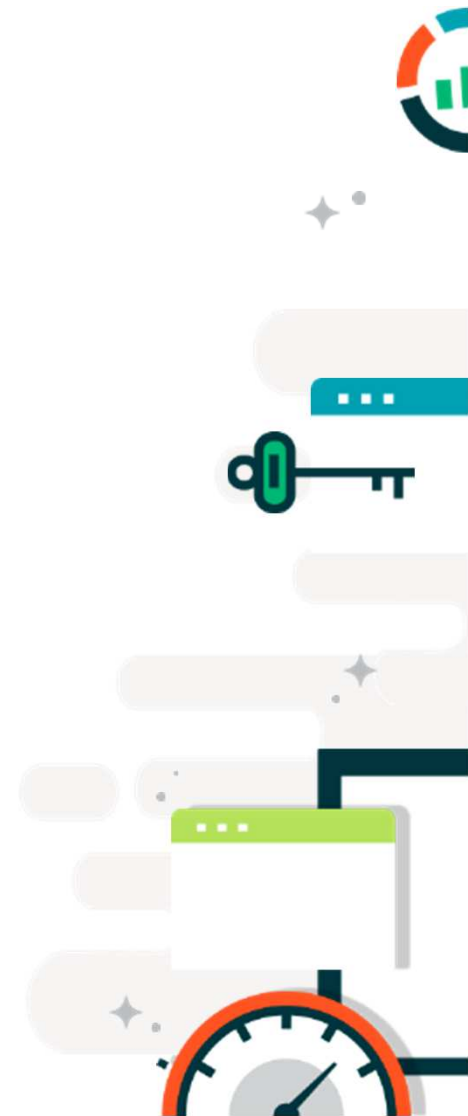
# Unidade 2

- Regressão Logística Simples
  - Escolha da Variável Resposta
  - Estimação do Modelo
  - Verificação do Ajuste
- Regressão Logística Múltipla
  - Escolha da Variável Resposta e das Variáveis Predictoras
  - Ajuste do Modelo
  - Encontrar o Melhor Modelo
  - Verificação da Qualidade do Modelo



# Unidade 3

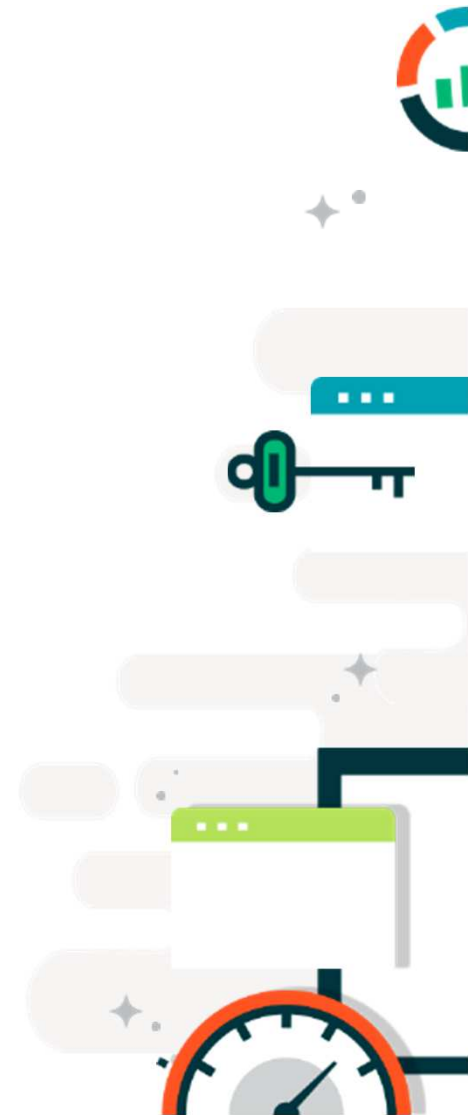
- Análises Multivariada de Dados
  - Análise Fatorial
  - Análise de Cluster (Agrupamento)
  - Análise Discriminante





# Unidade 4

- Análises de Séries Temporais
  - Análise da Série (Estacionaridade)
  - Análise de Tendência
  - Análise de Sazonalidade
  - Análise de Modelo ARIMA (p, d, q)



# Como preparar os Dados?





# Análise Preditiva

Gabriel Vinícius Araújo Fonseca

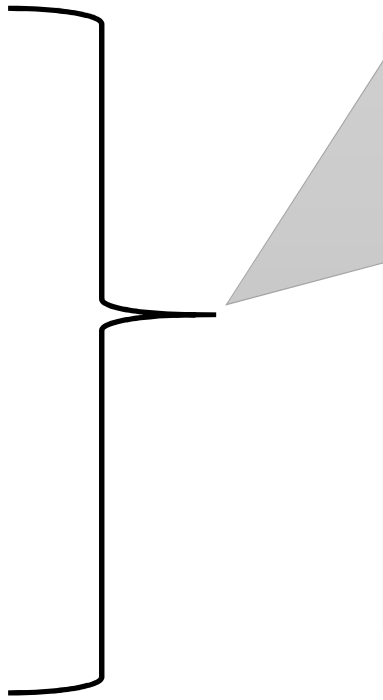


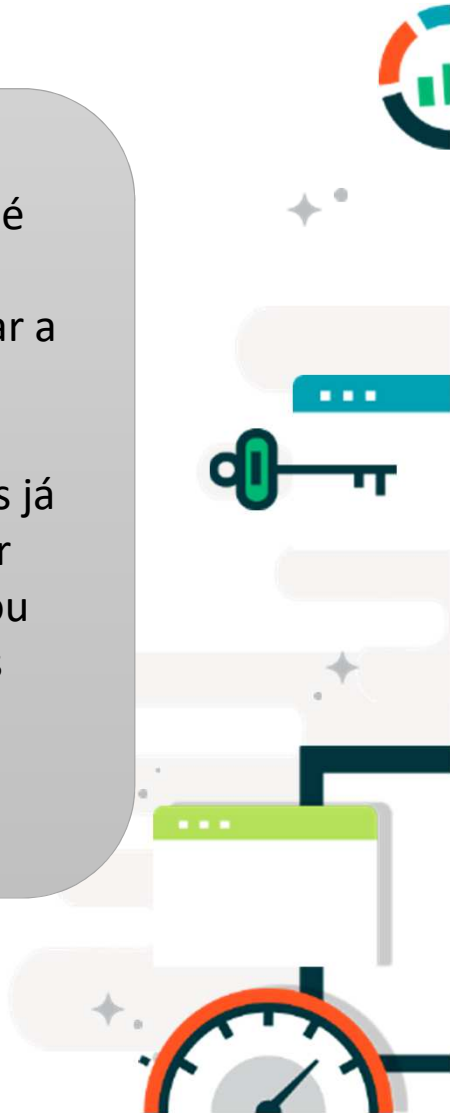
# Preparação do Dados



# Preparação dos Dados

- Tipos de variáveis
- Missing
- Outliers
- Transformações
- Padronizações
- Estrutura dos dados

- 
- ✓ A Estatística Descritiva é uma das principais ferramentas para ajudar a identificar dados com problemas.
  - ✓ Ferramentas avançadas já existem para identificar problemas de coletas ou mesmo casos de dados extremamente discrepantes.



# Padronização

- Uma das formas simples de identificar e trazer os dados para a mesma escala, é a padronização.
- Para cada observação, os dados podem ser transformados para a escala Z, dos quais são transformados em:

$$Z = \frac{x - \bar{x}}{s}$$



# Mudança de Escala (Normalização)

- A normalização consiste em deixar qualquer variável de estudo numa mesma escala, baseando-se exclusivamente nos valores mínimos e máximo observados.

$$N = \frac{x - x_{min}}{x_{max} - x_{min}}$$





# Outliers – Dados extremos

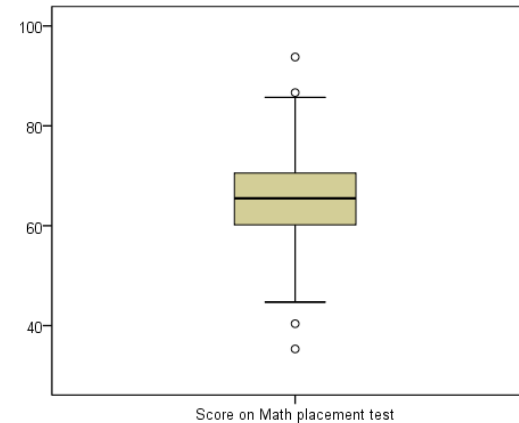
- **Dados extremos** ou **discrepantes**, mais conhecidos como ***Outliers*** são dados que fogem da realidade de uma determinada população de dados.
- Existem inúmeras maneiras de avaliar se **um dado** ou a **informação como toda (múltipla)** é fora do comum ou não.



# Outliers – Dados extremos

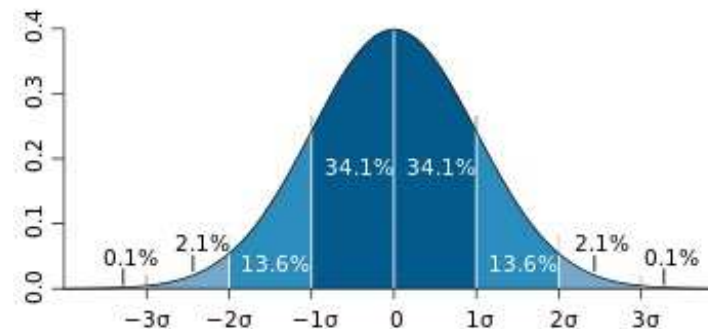
- **Forma 1:**

- Gráfico de Caixa (Box-plot): todo dado que estiver acima ou abaixo de 1,5 vezes a diferença interquartílica em relação a mediana pode ser considerada um ponto discrepante.



- **Forma 2:**

- Padronização: toda observação que ao ser padronizada pela média e desvio-padrão ter um valor absoluto superior a três deve ser considerado como observação discrepante.



**Depois do tratamento  
dos dados!  
Hora da modelagem?**





# Análise Preditiva

Gabriel Vinícius Araújo Fonseca

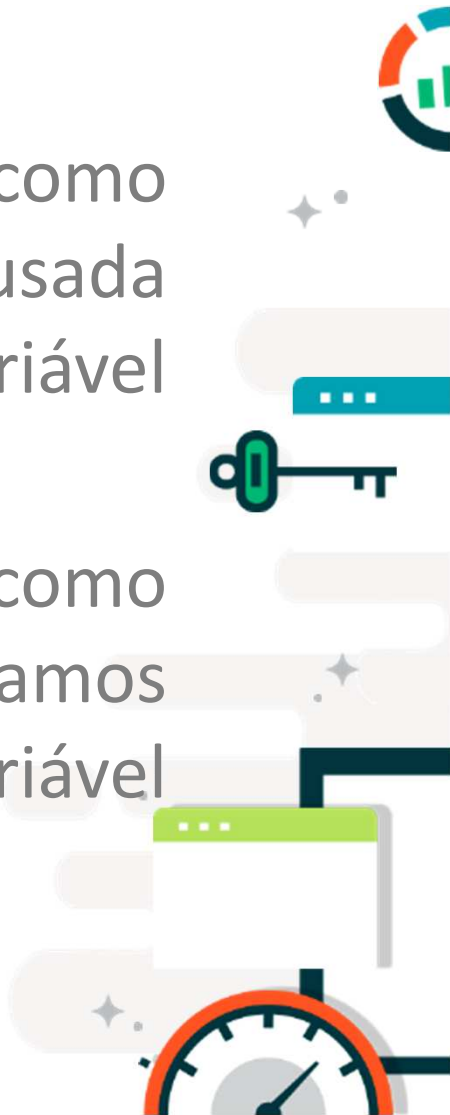


# Correlação entre duas variáveis



# Variáveis

- Variável Preditora (X): também conhecida como variável independente, é a variável que será usada como informação para se obter o valor da variável resposta (Y).
- Variável Resposta (Y): também conhecida como variável dependente, é a variável na qual desejamos buscar uma informação baseada na variável independente (X).



# Correlação entre duas variáveis

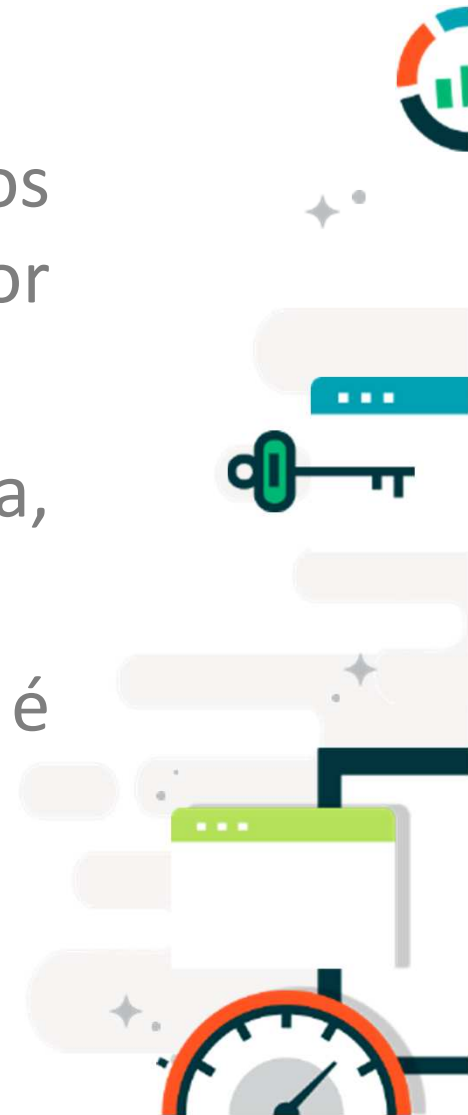
- Como avaliar a relação entre duas variáveis quantitativas?
- Diagrama de Dispersão
- Coeficiente de Correlação
- Teste de Hipóteses para o Coeficiente de Correlação



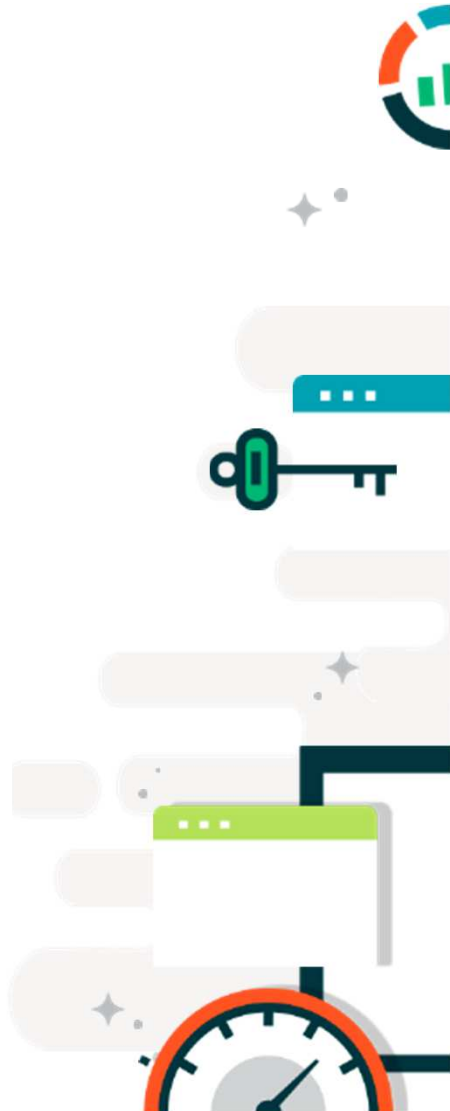
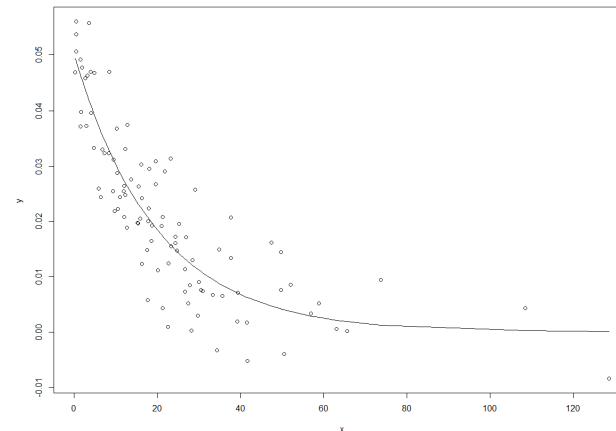
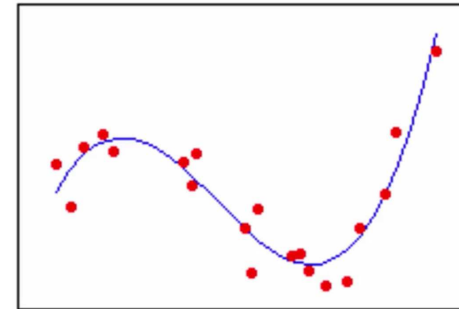
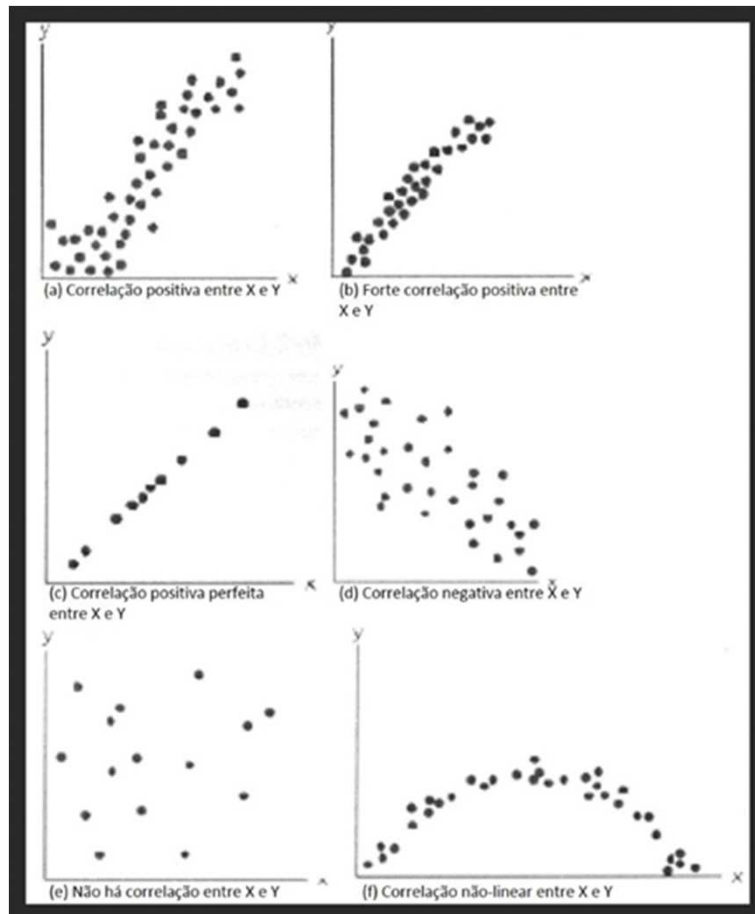


# Diagrama de Dispersão

- Gráfico que possibilita visualizar a distribuição dos pares de dados (X, Y) para encontrar a melhor relação dos pontos.
- As principais relações são: linear, quadrática, exponencial, logarítmica ou polinomiais.
- Em alguns, a relação não está bem definida e não é possível obter um modelo paramétrico.



# Tipos de Relações e exemplos



# Coeficiente de Correlação Linear

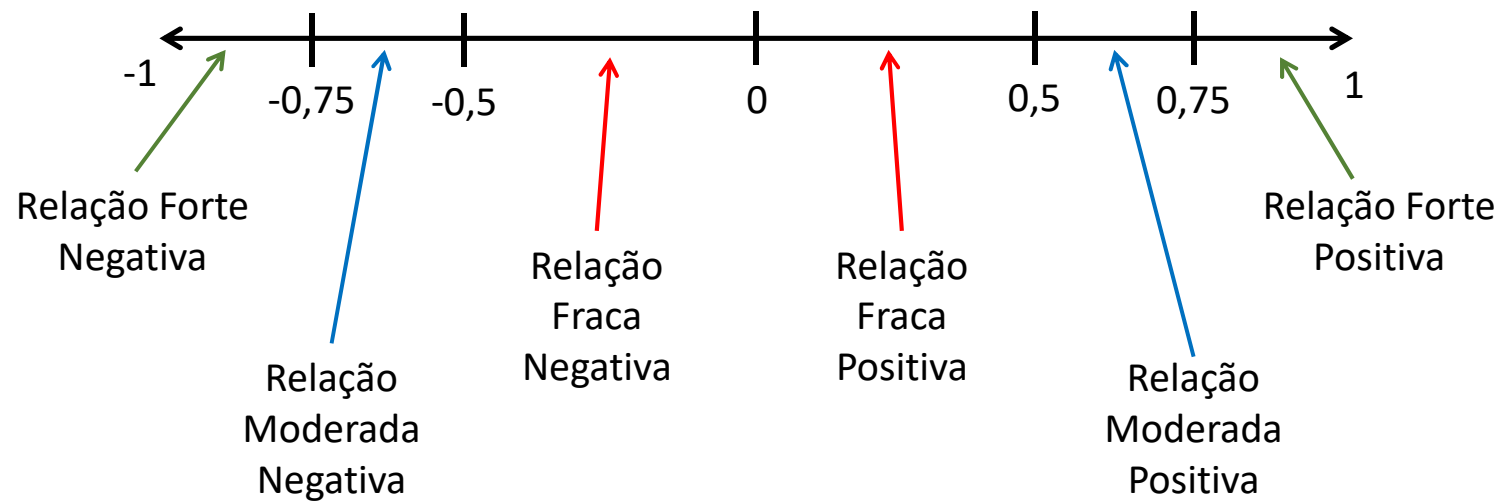
- Para medir o grau de relação linear entre duas variáveis, podemos usar a medida estatística chamada coeficiente de correlação linear (ou de Pearson).
- O valor estará sempre entre -1 e 1.

$$r = \frac{(\sum_{i=1}^n x_i y_i) - n\bar{x}\bar{y}}{\sqrt{\left((\sum_{i=1}^n x_i^2) - n\bar{x}^2\right) \left((\sum_{i=1}^n y_i^2) - n\bar{y}^2\right)}}$$



# Interpretação

- Como interpretar o coeficiente de correlação?



# Exemplos – Casas.xlsx

- O banco de dados se trata de quarenta casas vendidas no Condado de Dutchess, Nova York – EUA.
- Elas foram avaliadas em **Preço de Venda** (mil US\$), **Preço Anunciado** (mil US\$), **Área Útil** (m<sup>2</sup>), **Terreno** (acre), **Idade** (anos) e pelo número de **Cômodos**, **Quartos** e **Banheiros** (cada uma).



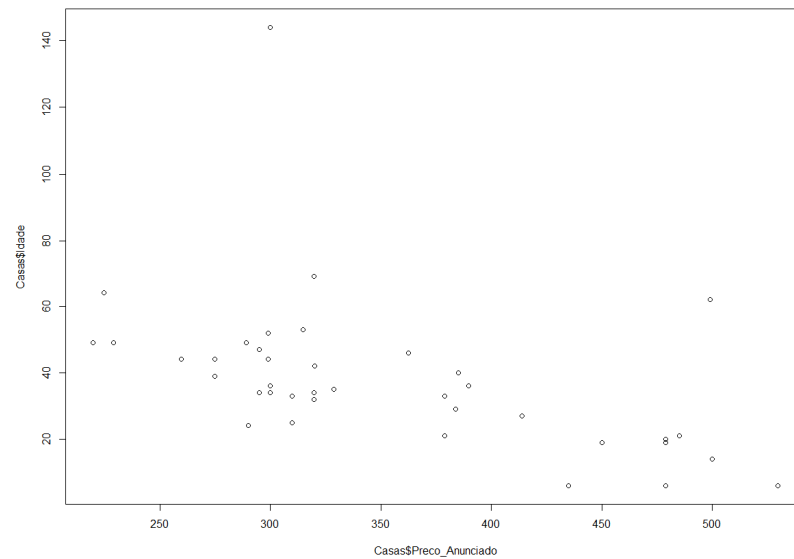
# Exemplos

- Utilizando os dados de 40 casas anunciadas e vendidas nos EUA, será que existe correlação entre o preço anunciado e a idade do imóvel?

## Script

```
r1 <- cor(Casas$Preco_Anunciado,  
Casas$Idade)  
r1  
plot(Casas$Preco_Anunciado,  
Casas$Idade)
```

```
> r1 <- cor(Casas$Preco_Anunciado,  
Casas$Idade)  
> r1 [1] -0.4928441  
> plot(Casas$Preco_Anunciado,  
Casas$Idade)
```



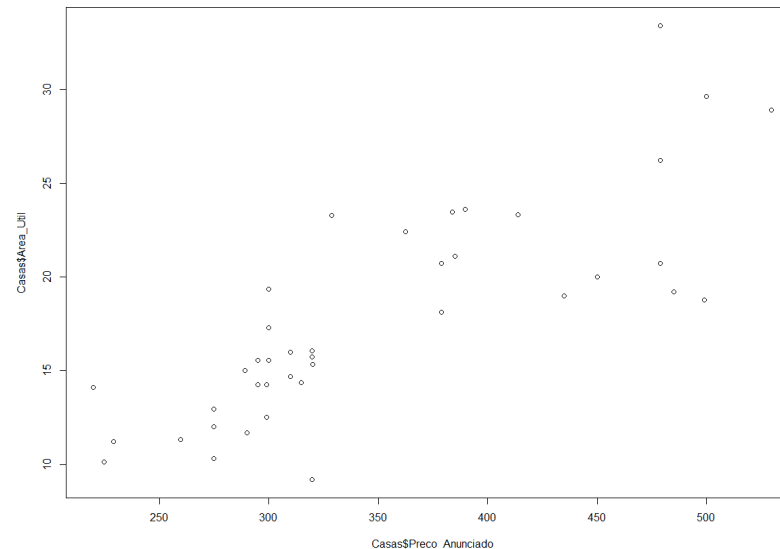
# Exemplos

- Utilizando os dados de 40 casas anunciadas e vendidas nos EUA, será que existe correlação entre o preço anunciado e a área útil?

## Script

```
r2<- cor(Casas$Preco_Anunciado,  
Casas$Area_Util)  
r2  
plot(Casas$Preco_Anunciado,  
Casas$Area_Util)
```

```
> r2<- cor(Casas$Preco_Anunciado,  
Casas$Area_Util)  
> r2 [1] 0.8083902  
> plot(Casas$Preco_Anunciado,  
Casas$Area_Util)
```



# Exemplos

- Testando se as duas correlações calculadas são ou não iguais a zero.

## Script

```
teste1 <- cor.test(Casas$Preco_Anunciado, Casas$Idade)
```

```
teste1
```

```
teste2 <- cor.test(Casas$Preco_Anunciado, Casas$Area_Util)
```

```
teste2
```

```
> teste1 <- cor.test(Casas$Preco_Anunciado, Casas$Idade)
> teste1

Pearson's product-moment correlation

data: Casas$Preco_Anunciado and Casas$Idade
t = -3.4916, df = 38, p-value = 0.001234
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6973000 -0.2142235
sample estimates:
cor
-0.4928441
```

```
> teste2 <- cor.test(Casas$Preco_Anunciado, Casas$Area_Util)
> teste2

Pearson's product-moment correlation

data: Casas$Preco_Anunciado and Casas$Area_Util
t = 8.4656, df = 38, p-value = 2.821e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6641203 0.8946163
sample estimates:
cor
0.8083902
```





**Como obter o modelo  
que descreve essa  
relação linear?**





# Análise Preditiva

Gabriel Vinícius Araújo Fonseca



# Regressão Linear Simples



# Regressão Linear Simples

## Equação do modelo

Intercepto Y  
Populacional

Coefficiente  
angular  
População

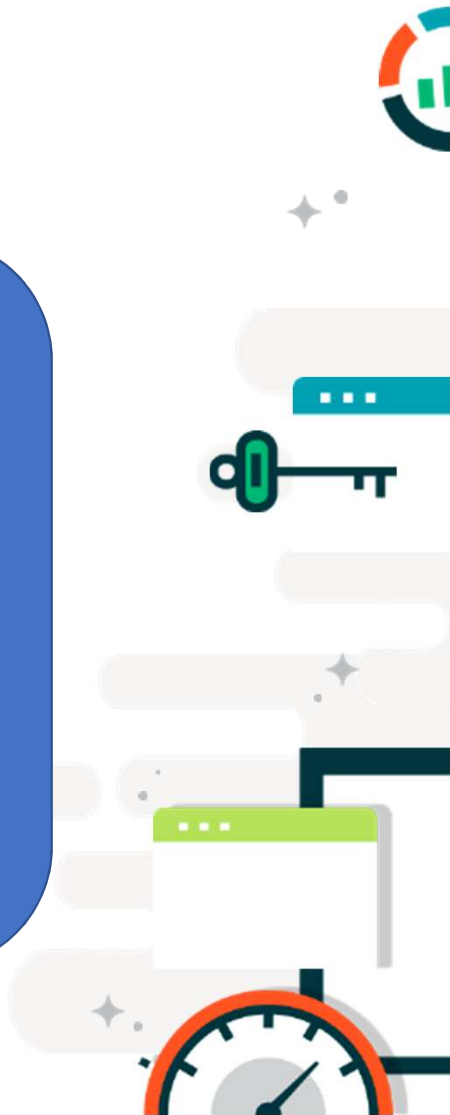
Erro  
aleatório

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Variável  
dependente  
(Resposta)

Variável independente  
(Preditora)

The diagram shows the equation  $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$  centered within a blue rounded rectangle. Five labels with arrows point to the components of the equation: 'Intercepto Y Populacional' points to  $\beta_0$ , 'Coefficiente angular População' points to  $\beta_1$ , 'Erro aleatório' points to  $\varepsilon_i$ , 'Variável dependente (Resposta)' points to  $Y_i$ , and 'Variável independente (Preditora)' points to  $X_i$ .

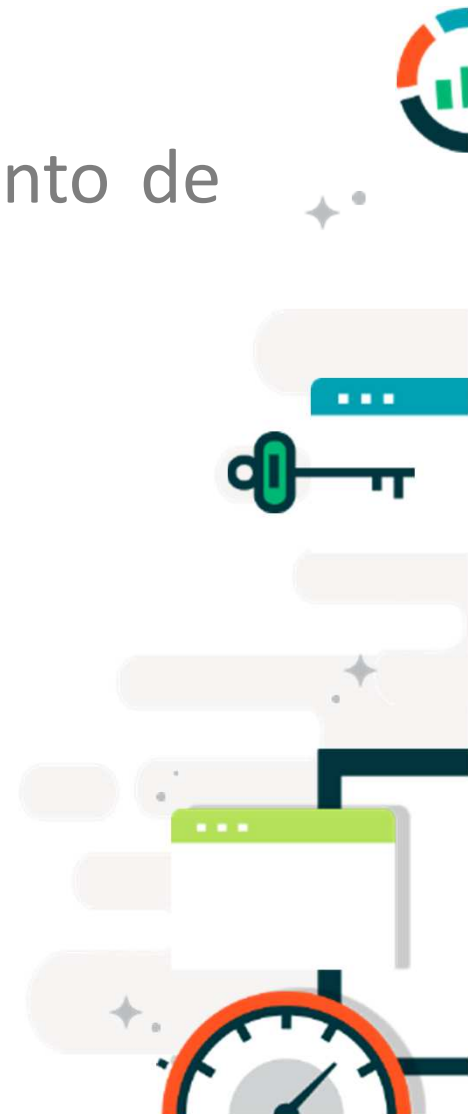


# Regressão Linear Simples

- Como obter os valores de a e b para um conjunto de dados X e Y?

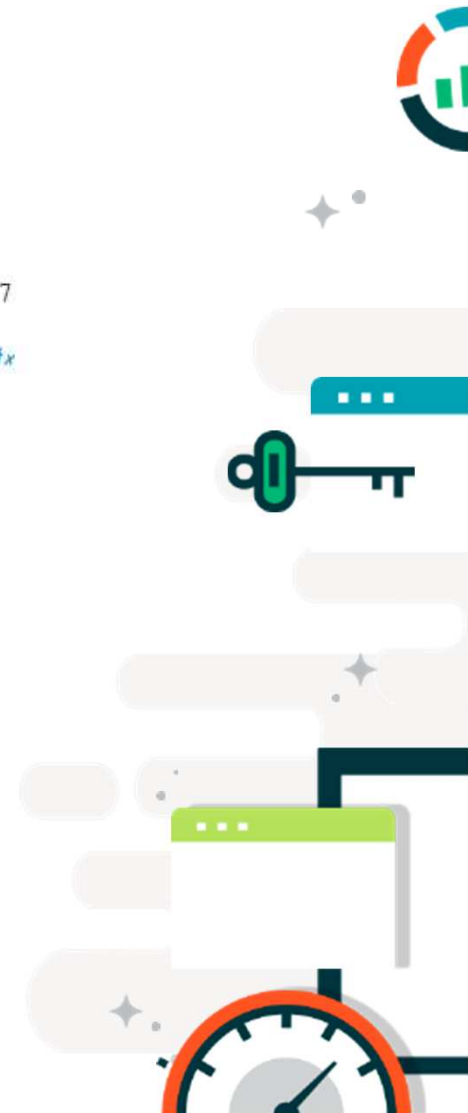
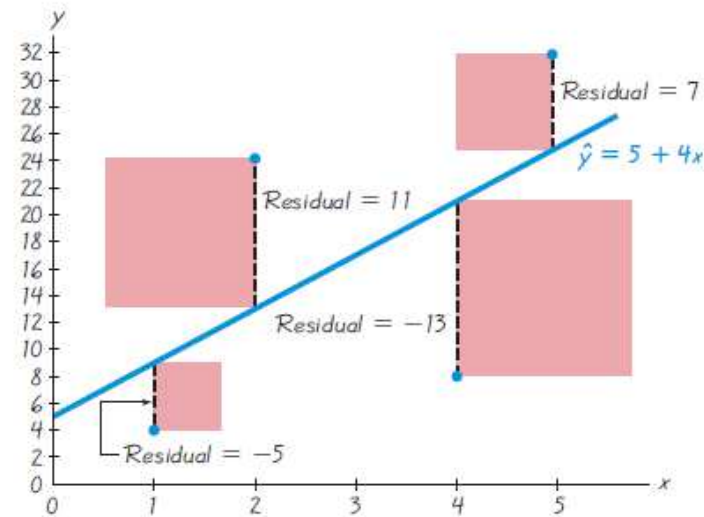
$$\hat{a} = \frac{(\sum_{i=1}^n x_i y_i) - n\bar{x}\bar{y}}{\left( (\sum_{i=1}^n x_i^2) - n\bar{x}^2 \right)}$$
$$\hat{b} = \bar{y} - \hat{a}\bar{x}$$

- Logo meu modelo estimado será:  $\hat{y} = \hat{a}x + \hat{b}$ .



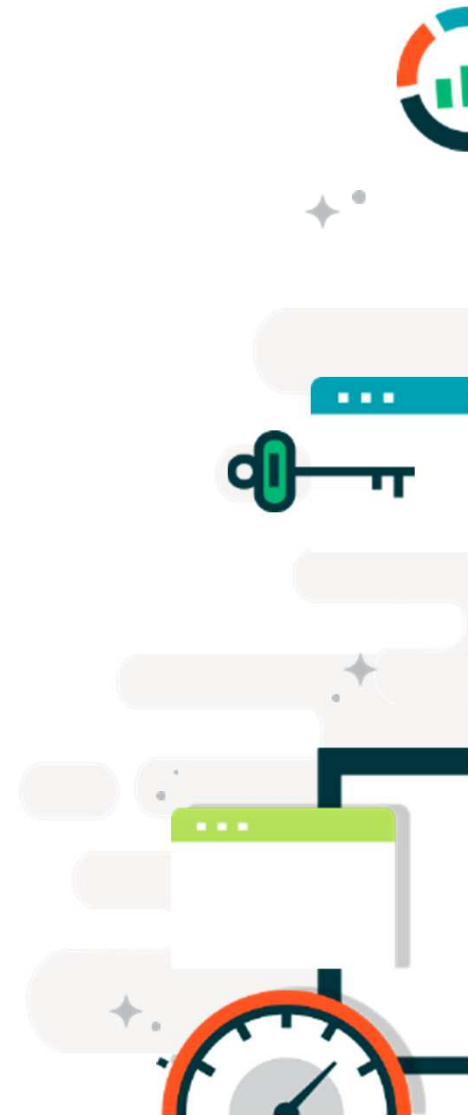
# Método de Mínimos Quadrados

- Distâncias verticais entre os dados originais e a reta (resíduos)
- A soma dos quadrados dos resíduos é a menor possível (Propriedade dos Mínimos Quadrados)



# Interpretando o Modelo

- $\beta_0$  (Intercepto): representa qual será o valor de Y quando X for igual a zero.
- $\beta_1$  (coeficiente angular): representa o quanto irá afetar Y a cada unidade em X.





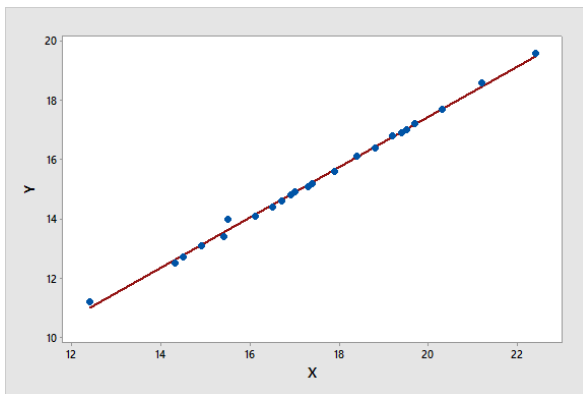
# Procedimentos para uso de um modelo de Regressão

- Definir o problema
- Selecionar as variáveis (preditoras e predita)
- Diagrama de dispersão
- Gerar o modelo de regressão
- Verificar a existência de *Outliers*
- Verificação do ajuste
- Uso do modelo

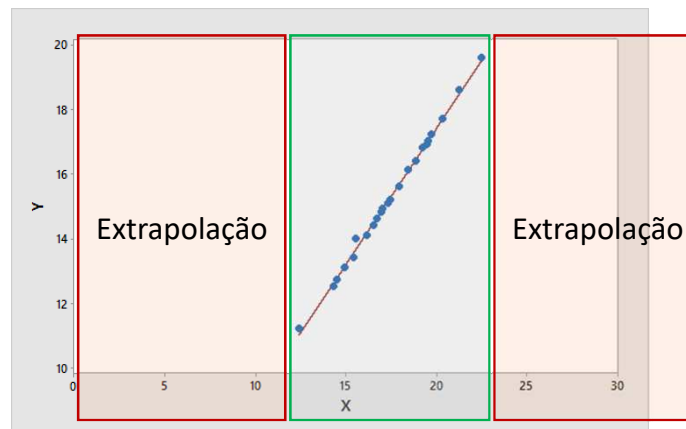


# Predições

Modelo de Regressão



Intervalo para predição



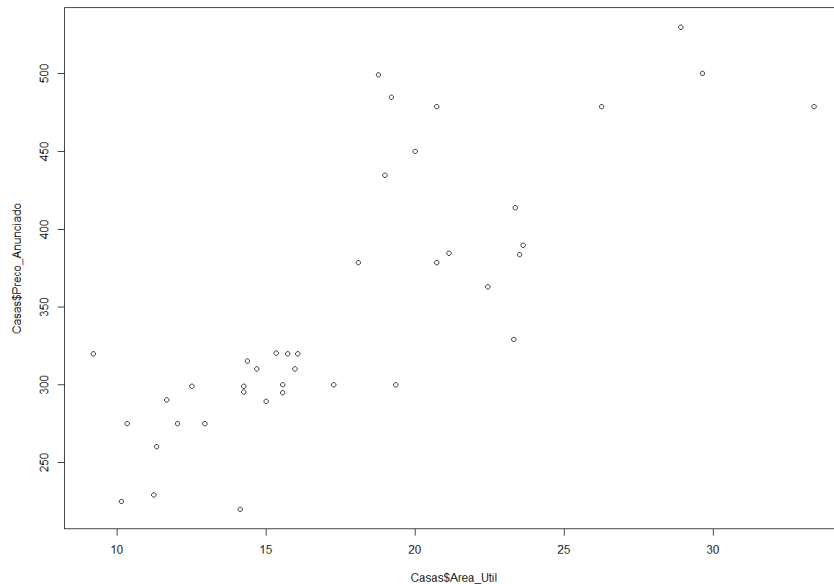
As predições (Y) não devem extrapolar o intervalo dos dados da variável que foi utilizada (X) para gerar o modelo .

Corre-se o risco de fazer uma estimativa errada uma vez que não se sabe o comportamento dos dados em intervalos menores ou maiores do que os da amostra usada.



# Exemplo 1

- Utilizando os dados de 40 casas anunciadas e vendidas nos EUA, qual o modelo que descreve a relação entre a área útil e o preço anunciado?



```
> r2<- cor(Casas$Area_Util, Casas$Preco_Anunciado)
> r2 [1] 0.8083902
> plot(Casas$Area_Util, Casas$Preco_Anunciado)
```



# Exemplo 1

- Obtendo o ajuste do modelo utilizando o comando “lm”

## Script

```
ajuste <- lm(Preco_Anunciado ~ Area_Util, data = Casas)
summary(ajuste)
lines(Casas$Area_Util, ajuste$fitted.values, col = 2)
```

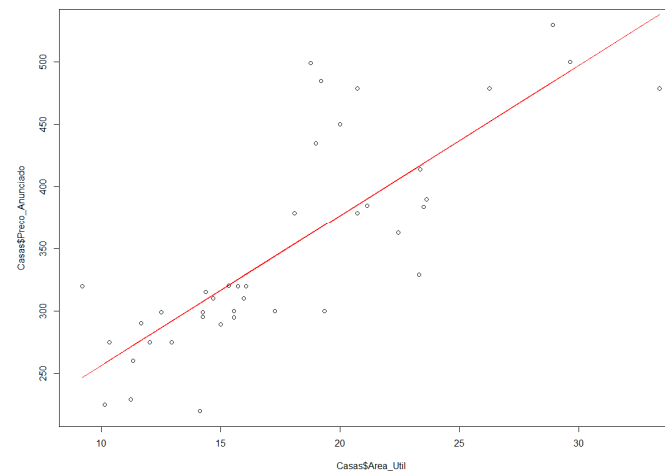
```
Call:
lm(formula = Preco_Anunciado ~ Area_Util, data = Casas)

Residuals:
    Min       1Q   Median       3Q      Max
-87.726 -28.780  -6.163  13.784 137.045

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  135.997     26.501   5.132 8.79e-06 ***
Area_Util     12.042      1.422   8.466 2.82e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

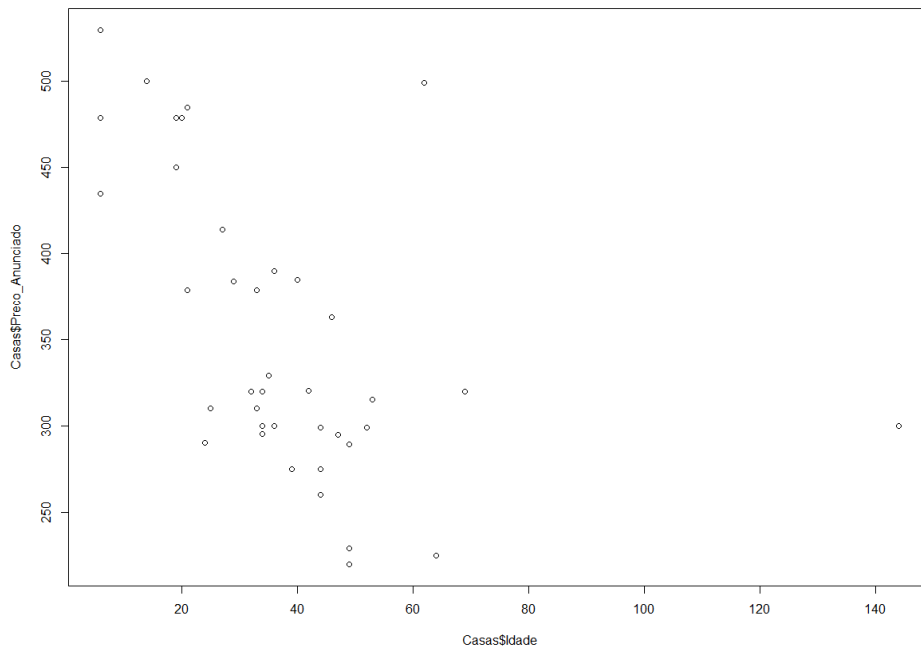
Residual standard error: 50.45 on 38 degrees of freedom
Multiple R-squared:  0.6535,    Adjusted R-squared:  0.6444
F-statistic: 71.67 on 1 and 38 DF, p-value: 2.821e-10
```

$$\hat{y} = 135,997 + 12,042x$$



# Exemplo 2

- E qual seria o modelo entre as variáveis Preço Anunciado e a Idade do imóvel?



```
> r1 <- cor(Casas$Idade, Casas$Preco_Anunciado)
> > r1 [1] -0.4928441
> > plot(Casas$Idade, Casas$Preco_Anunciado)
```



# Exemplo 2

- Obtendo o ajuste do modelo utilizando o comando “lm”

## Script

```
ajuste <- lm(Preco_Anunciado ~ Idade, data = Casas)
summary(ajuste)
lines(Casas$Idade, ajuste$fitted.values, col = 2)
```

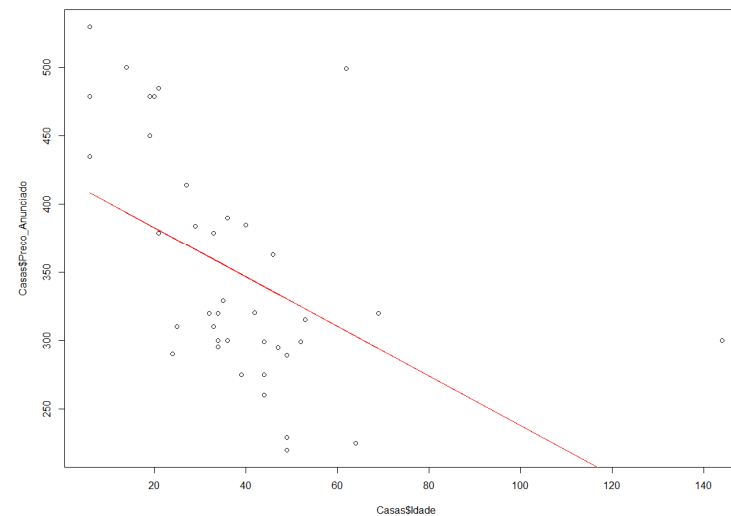
```
call:
lm(formula = Preco_Anunciado ~ Idade, data = Casas)

Residuals:
    Min       1Q   Median       3Q      Max
-110.28  -58.76  -24.31   39.81  192.44

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  419.2104    23.0788   18.164 < 2e-16 ***
Idade        -1.8170     0.5204   -3.492  0.00123 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 74.57 on 38 degrees of freedom
Multiple R-squared:  0.2429,    Adjusted R-squared:  0.223
F-statistic: 12.19 on 1 and 38 DF,  p-value: 0.001234
```

$$\hat{y} = 419,2104 - 1,8170x$$



**E se desejarmos  
adicionar mais de uma  
variável preditora?**







# Análise Preditiva

Gabriel Vinícius Araújo Fonseca

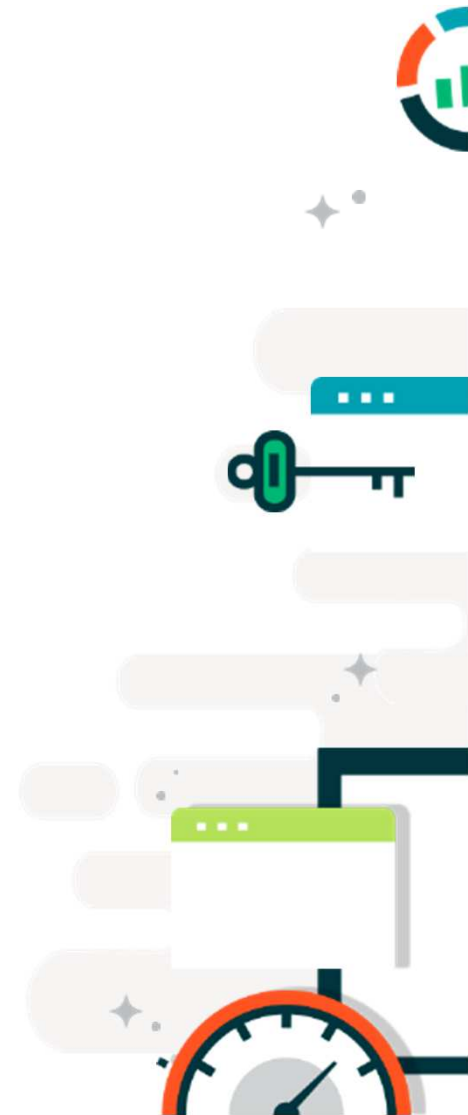


# Exemplo de Regressão Linear Simples



# Procedimentos para uso de um modelo de Regressão

- Definir o problema
- Selecionar as variáveis (preditoras e predita)
- Diagrama de dispersão
- Gerar o modelo de regressão
- Verificar a existência de *Outliers*
- Verificação do ajuste
- Uso do modelo



# Conjunto de Dados

- Foram coletados 45 bezerros em uma fazenda e medido em cm o perímetro torácico com fita métrica e depois o peso do mesmo. Seria possível criar uma fita padrão para prever o peso através do valor medido na fita?

[http://www.guaporepecuaria.com.br/pecuaria/trabalhos/04\\_nem\\_fitametrica\\_pn/artigo\\_fita.jpg](http://www.guaporepecuaria.com.br/pecuaria/trabalhos/04_nem_fitametrica_pn/artigo_fita.jpg)

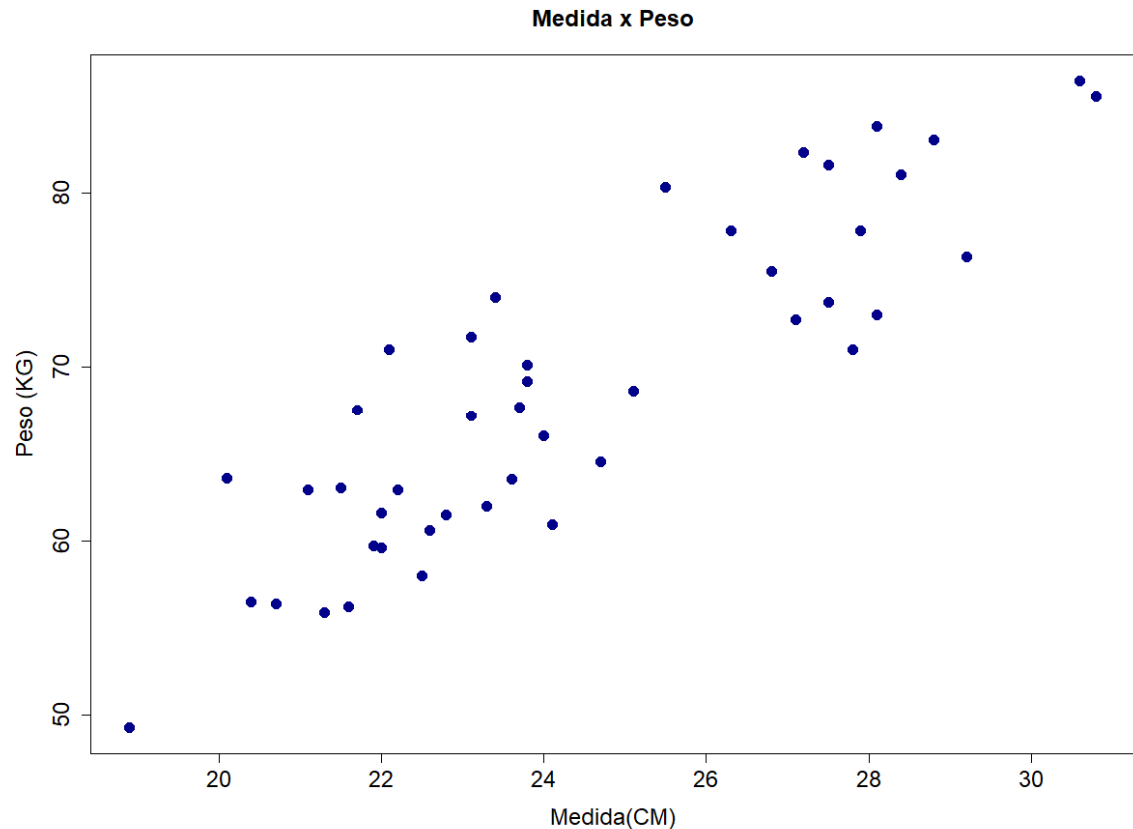


# Diagrama de Dispersão

## Script

```
attach(RLS_BEZERRO)
```

```
plot(`Medida(CM)`,'Peso (KG)',  
     pch = 19, col = "Dark Blue",  
     main = "Medida x Peso",  
     cex.axis = 1.5, cex.main = 1.5,  
     cex.lab = 1.5, cex = 1.5)
```



# Coeficiente de Correlação e Teste

## Script

```
cor.test(`Medida(CM)`, `Peso (KG)`)
```

```
> cor.test(`Medida(CM)`, `Peso (KG)`)

Pearson's product-moment correlation

data: Medida(CM) and Peso (KG)
t = 12.236, df = 43, p-value = 1.342e-15
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7930667 0.9334432
sample estimates:
      cor 
0.8814144
```



# Modelo de Regressão

## Script

```
ajuste <- lm(`Peso (KG)` ~ `Medida(CM)`)
```

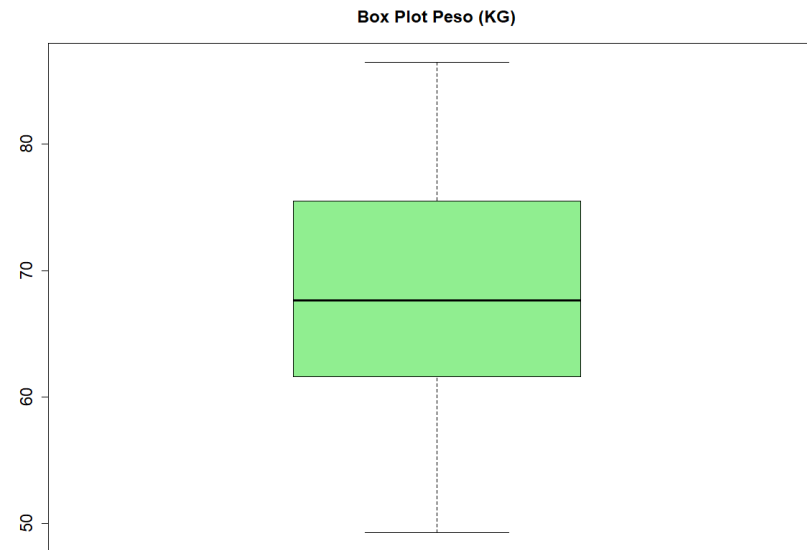
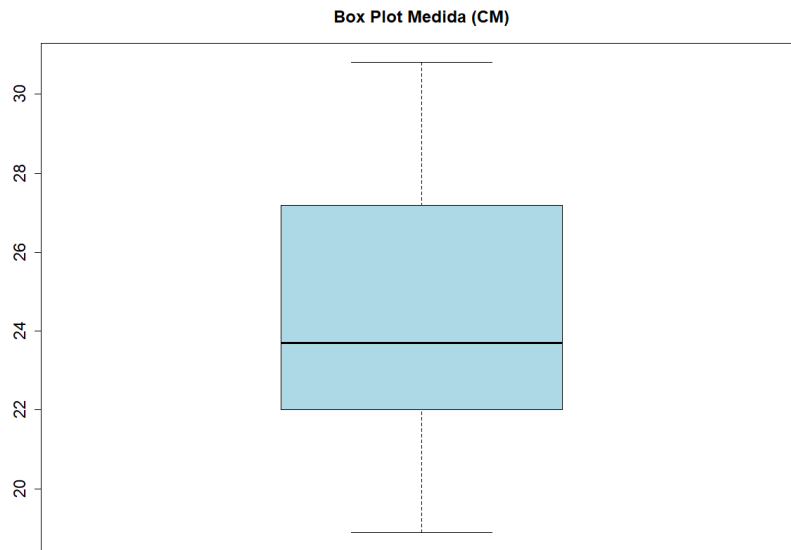
```
summary(ajuste)
```

$$\hat{y} = 2,7893 + 2,7012 \cdot Medida$$

```
> ajuste <- lm(`Peso (KG)` ~ `Medida(CM)`)  
> summary(ajuste)  
  
Call:  
lm(formula = `Peso (KG)` ~ `Medida(CM)`)  
  
Residuals:  
    Min       1Q   Median       3Q      Max   
-6.9876 -3.2912 -0.4855  3.0227  8.6307   
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)      
(Intercept)   2.7893     5.4299   0.514    0.61      
`Medida(CM)`   2.7012     0.2207  12.236 1.34e-15 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 4.426 on 43 degrees of freedom  
Multiple R-squared:  0.7769,    Adjusted R-squared:  0.7717   
F-statistic: 149.7 on 1 and 43 DF,  p-value: 1.342e-15
```

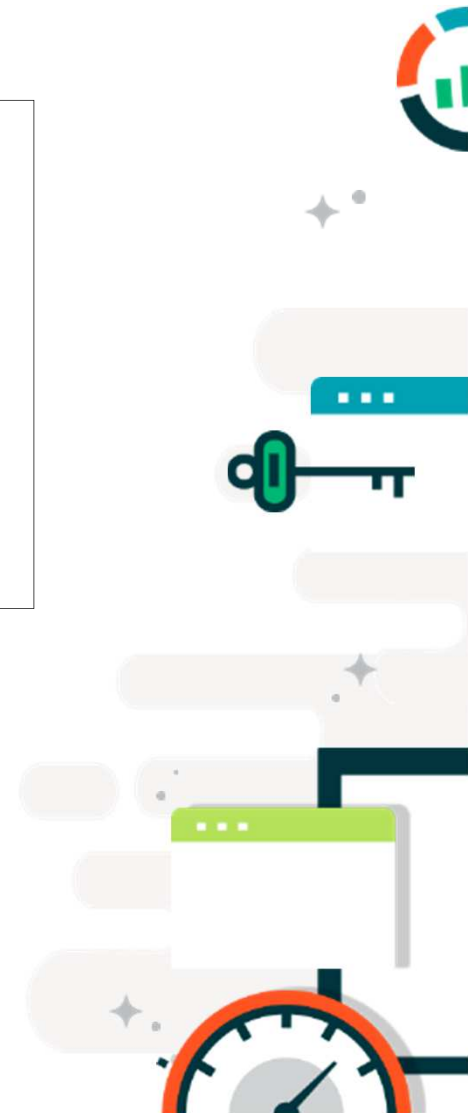


# Existe algum Outlier?



## Script

```
boxplot(`Medida(CM)` , col = "Light Blue", main = "Box Plot Medida (CM)",  
        cex.axis = 1.5, cex.main = 1.5, cex.lab = 1.5)  
boxplot(`Peso (KG)` , col = "Light Green", main = "Box Plot Peso (KG)",  
        cex.axis = 1.5, cex.main = 1.5, cex.lab = 1.5)
```





# Existe algum Outlier?

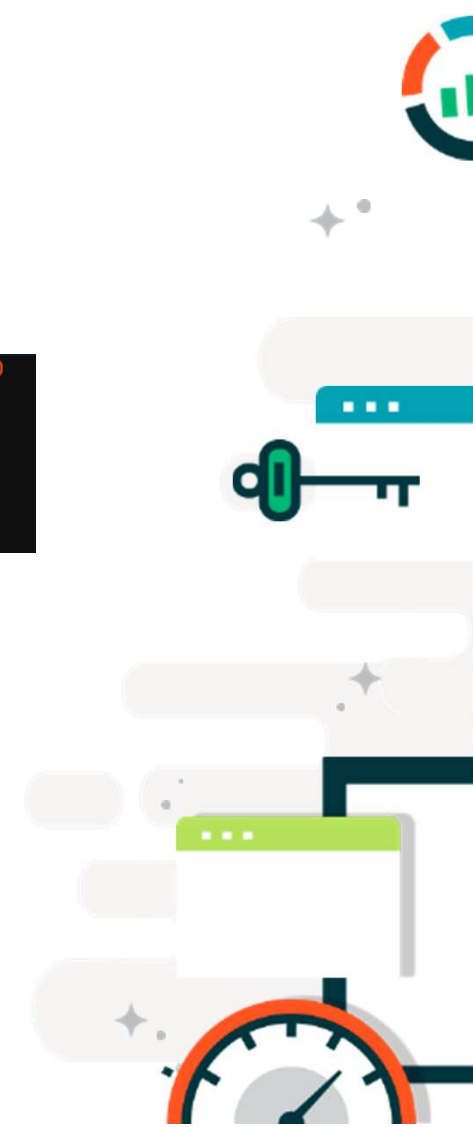
Padronização (score)

$$Z = \frac{x - \bar{x}}{s}$$

## Script

```
RLS_BEZERRO$Z_Medida <- scale(`Medida(CM)`)
RLS_BEZERRO$Z_Peso <- scale(`Peso (KG)`)
range(RLS_BEZERRO$Z_Medida)
range(RLS_BEZERRO$Z_Peso)
```

```
> RLS_BEZERRO$Z_Medida <- scale(`Medida(CM)`)
> RLS_BEZERRO$Z_Peso <- scale(`Peso (KG)`)
> range(RLS_BEZERRO$Z_Medida)
[1] -1.824839  2.112314
> range(RLS_BEZERRO$Z_Peso)
[1] -2.098742  1.906573
>
```



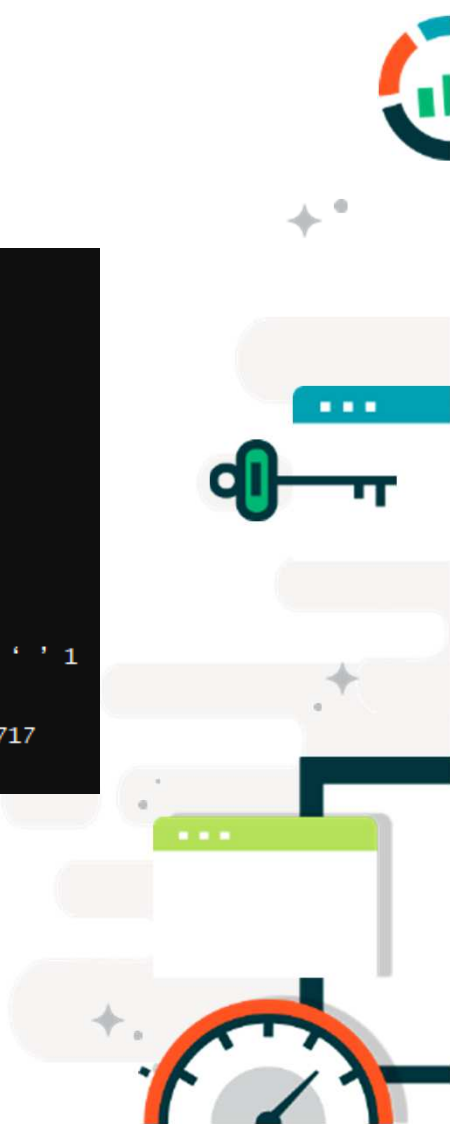
# Verificação do Ajuste e Suposições

## Script

```
ajuste <- lm(`Peso (KG)` ~ `Medida(CM)`)
```

```
summary(ajuste)
```

```
> ajuste <- lm(`Peso (KG)` ~ `Medida(CM)`)  
> summary(ajuste)  
  
Call:  
lm(formula = `Peso (KG)` ~ `Medida(CM)`)  
  
Residuals:  
    Min       1Q   Median       3Q      Max   
-6.9876 -3.2912 -0.4855  3.0227  8.6307   
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)      
(Intercept)   2.7893     5.4299   0.514    0.61      
`Medida(CM)`   2.7012     0.2207  12.236 1.34e-15 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 4.426 on 43 degrees of freedom  
Multiple R-squared:  0.7769,    Adjusted R-squared:  0.7717   
F-statistic: 149.7 on 1 and 43 DF,  p-value: 1.342e-15
```

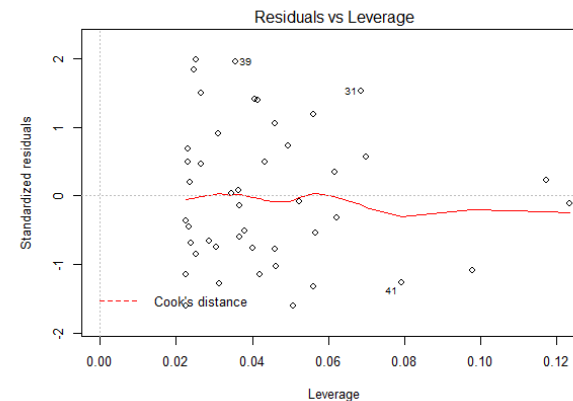
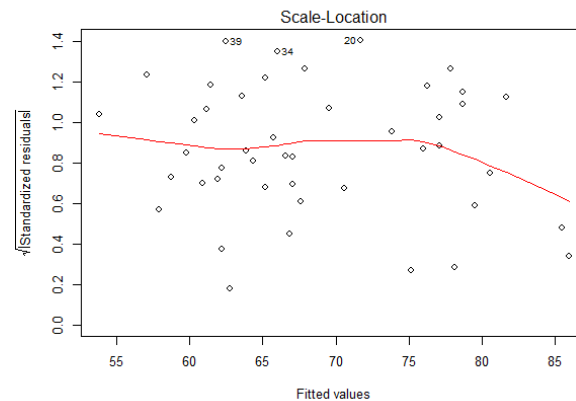
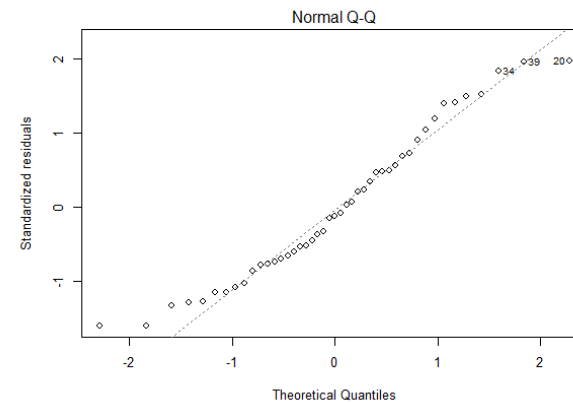
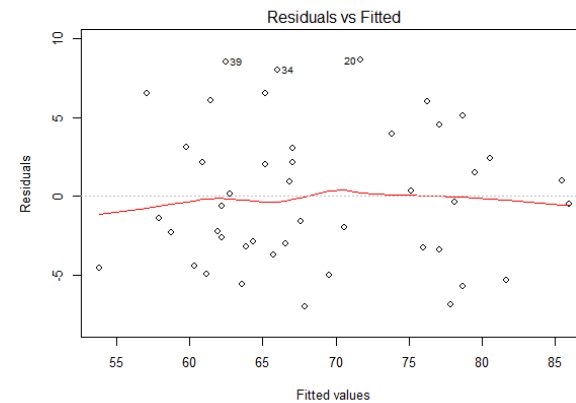


# Verificação do Ajuste e Suposições

## Script

```
par(mfrow=c(2,2))
```

```
plot(ajuste)
```



# Uso do Modelo (Predição)

## Script

```
predicao <- data.frame(Peso_previsto = c(20, 28))
```

```
coef_ajuste <- coefficients(ajuste)
```

```
predicao$Peso_Previsto <- coef_ajuste[1]+coef_ajuste[2]*predicao
```

```
colnames(predicao) <- c("Medida", "Peso_previsto")
```

predicao

```
> predicao <- data.frame(Peso_previsto = c(20, 28))
> coef_ajuste <- coefficients(ajuste)
> predicao$Peso_Previsto <- coef_ajuste[1]+coef_ajuste[2]*predicao
> colnames(predicao) <- c("Medida", "Peso_previsto")
> predicao
  Medida Peso_previsto
1     20      56.81281
2     28      78.42221
```





# Análise Preditiva

Gabriel Vinícius Araújo Fonseca



# Regressão Linear Composta ou Multivariada

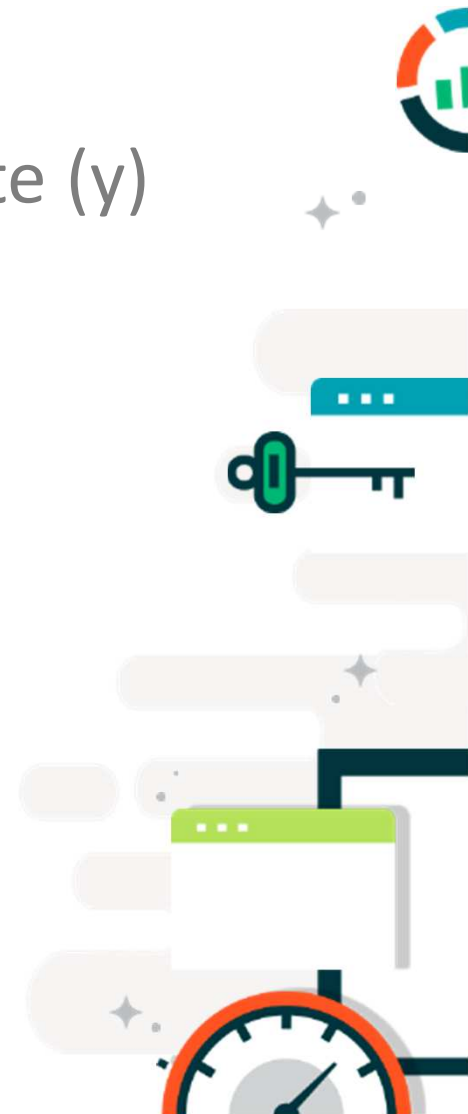


# Regressão Multivariada

- Expressa a relação entre uma variável dependente ( $y$ ) e duas ou mais variáveis preditoras ( $x_i$ )

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + E$$

- $K$  = número de variáveis preditoras ( $i = 1, \dots, k$ ).
- $Y$  = variável resposta.
- $\beta_0$  = intercepto do modelo.
- $\beta_i$  = coeficiente das variáveis preditoras.





# Etapas de Validação do Modelo

- Verificação dos Coeficientes das Variáveis Preditoras
- Verificação da ANOVA do Modelo
- $R^2$  ajustado
- Multicolinearidade (FIV)
- Método de Seleção de Variáveis (Stepwise)
- Análise de Resíduos - Suposições



# Análise de Variância - ANOVA

- Na análise de variância, testamos a hipótese se todos os coeficientes são iguais a zero versus existe pelo menos um coeficiente significativo no modelo.
- Ou seja:
- $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$
- $H_1: \text{pelo menos um } \beta_i \text{ diferente de zero}$



# Teste de Hipóteses para os Coeficientes

- Para cada coeficiente, é feito o seguinte teste de hipóteses:
- $H_0: \beta_i = 0$
- $H_1: \beta_i \neq 0$
- Ao verificar o valor p, caso seja inferior a 0,05, devemos rejeitar a hipótese nula e dizer que o efeito da variável preditora, se for caso, é significativo.

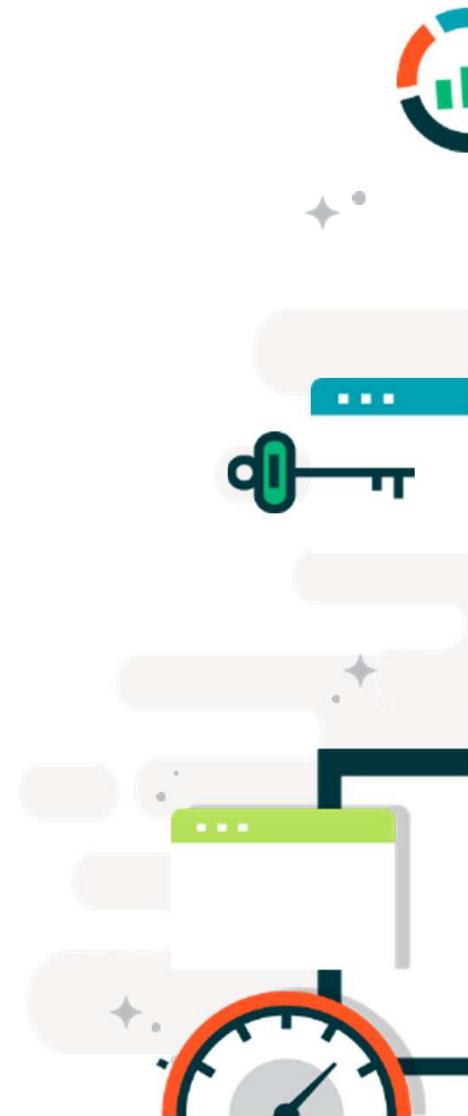


# R<sup>2</sup> Ajustado

- O coeficiente de determinação ajustado leva em consideração o número de variáveis no ajuste do modelo final

$$\text{adjusted } R^2 = 1 - \frac{(n - 1)}{[n - (k + 1)]} (1 - R^2)$$

- $n$  = tamanho da amostra
- $k$  = número de variáveis do modelo
- $R^2$  = coeficiente de determinação



# Multicolinearidade

- Informações semelhantes para explicar a variável dependente
- Reduz a capacidade de previsão do modelo
- Interpretações incoerentes dos coeficientes estimados
- Identificar através da matriz de correlação entre as variáveis e Fator Inflacionário da Variância



# Fator Inflacionário da Variância (FIV)

$$FIV_j = \frac{1}{1 - R_j^2}$$

- Onde  $R_j^2$  é o coeficiente de determinação de um modelo com  $X_j$  sendo a variável resposta e as demais como independentes

FIV > 5 significa que  $X_j$  é correlacionada com as outras variáveis dependentes



# Stepwise

- Método de seleção do melhor modelo baseado no testes estatístico sobre a variável preditora

	VAR 1	VAR 2	VAR 3	VAR 4	VAR 5	VAR 6	VAR 7	VAR 8	VAR 9	VAR 10	VAR 11	VAR 12
Passo 1												
Passo 2												
Passo 3												
Passo 4												
Passo 5												
Passo 6												
Passo 7												



**Como podemos obter o  
ajuste de um modelo no  
R?**







# Análise Preditiva

Gabriel Vinícius Araújo Fonseca



# Exemplo de Regressão Linear Composta



# Etapas de Validação do Modelo

- Verificação dos Coeficientes das Variáveis Predictoras
- Verificação da ANOVA do Modelo
- $R^2$  ajustado
- Multicolinearidade (FIV)
- Método de Seleção de Variáveis (Stepwise)
- Análise de Resíduos - Suposições



# Conjunto de Dados

- O arquivo HOUSES\_EUA.xlsx contém dados sobre 128 vendas recentes. Para cada venda, o arquivo mostra o bairro (1, 2 ou 3) em que a casa está localizada, o número de ofertas feitas na casa, a metragem quadrada, se a casa é feita principalmente de tijolos, o número de banheiros, o número de quartos, e o preço de venda.



# Matriz de Correlação

## Script

```
attach(HOUSES_EUA)

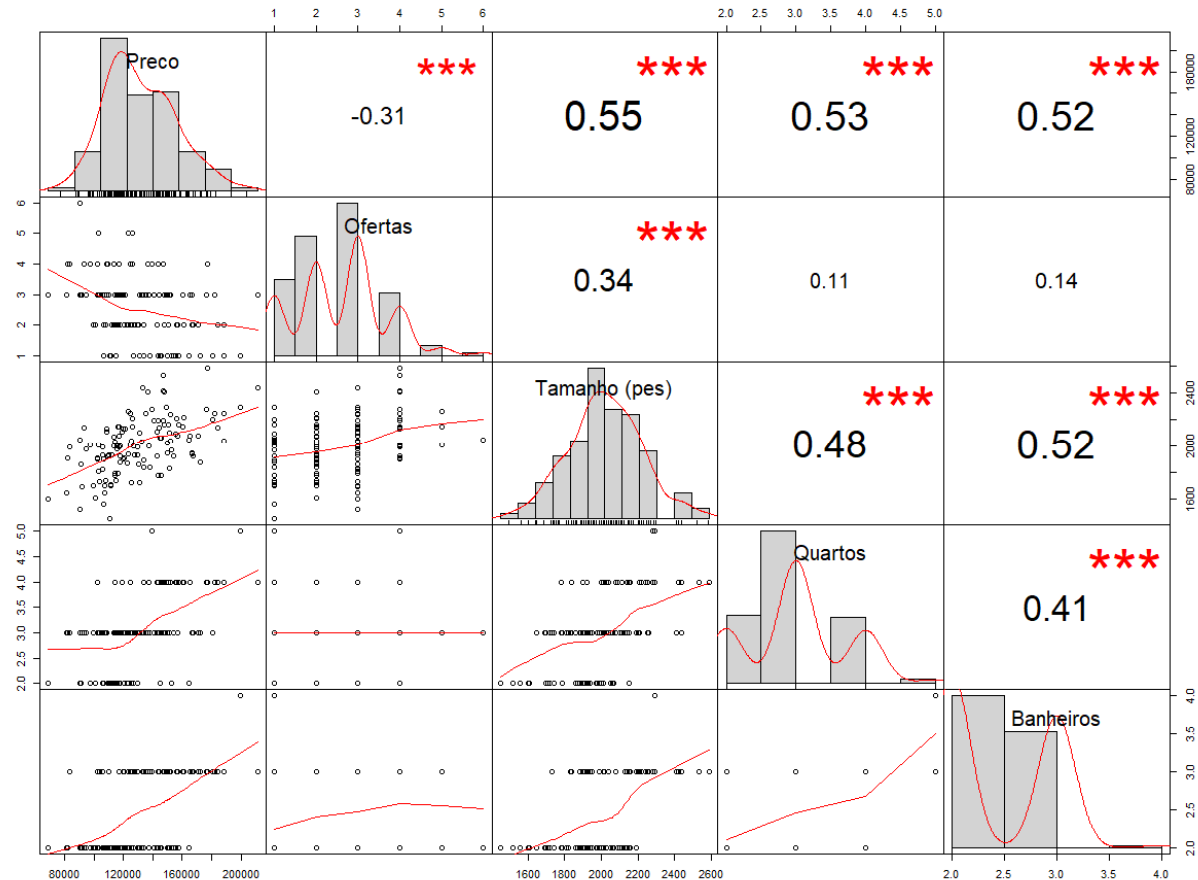
install.packages("PerformanceAnalytics")

library(PerformanceAnalytics)

HOUSES_Numeric <- HOUSES_EUA[,-2]

HOUSES_Numeric <- HOUSES_Numeric[,-4]

chart.Correlation(HOUSES_Numeric)
```



# Ajuste do Modelo e Verificações

## Script

```
ajuste <- lm(Preco ~ Ofertas + `Tamanho (pes)` + Quartos + Banheiros + factor(Bairro) + Feita_tijolos)
```

```
summary(ajuste)
```

```
> ajuste <- lm(Preco ~ Ofertas + `Tamanho (pes)` + Quartos + Banheiros + factor(Bairro) + Feita_tijolos)
> summary(ajuste)

Call:
lm(formula = Preco ~ Ofertas + `Tamanho (pes)` + Quartos + Banheiros +
    factor(Bairro) + Feita_tijolos)

Residuals:
    Min       1Q   Median       3Q      Max
-27337.3  -6549.5   -41.7    5803.4   27359.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2159.498    8877.810   0.243  0.80823
Ofertas       -8267.488    1084.777  -7.621 6.47e-12 ***
`Tamanho (pes)`    52.994     5.734   9.242 1.10e-15 ***
Quartos       4246.794    1597.911   2.658  0.00894 **
Banheiros     7883.278    2117.035   3.724  0.00030 ***
factor(Bairro)2 -1560.579    2396.765  -0.651  0.51621
factor(Bairro)3  20681.037    3148.954   6.568 1.38e-09 ***
Feita_tijolosYes 17297.350    1981.616   8.729 1.78e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10020 on 120 degrees of freedom
Multiple R-squared:  0.8686,    Adjusted R-squared:  0.861
F-statistic: 113.3 on 7 and 120 DF,  p-value: < 2.2e-16
```



# Multicolinearidade (FIV)

## Script

```
install.packages("car")
```

```
library(car)
```

```
vif(ajuste)
```

```
> vif(ajuste)
```

	GVIF	Df	GVIF <sup>1/(2*Df)</sup>
ofertas	1.702392	1	1.304757
`Tamanho (pes)`	1.862215	1	1.364630
Quartos	1.702472	1	1.304788
Banheiros	1.500976	1	1.225143
factor(Bairro)	2.129593	2	1.208020
Feita_tijolos	1.103908	1	1.050670





# Seleção do Modelo - Stepwise

## Script

```
install.packages("MASS")
```

```
library(MASS)
```

```
stepAIC(ajuste, direction = "both")
```

```
> stepAIC(ajuste, direction = "both")
Start: AIC=2366.07
Preco ~ Ofertas + `Tamanho (pes)` + Quartos + Banheiros + factor(Bairro) +
  Feita_tijolos
```

	Df	Sum of Sq	RSS	AIC
<none>			1.2046e+10	2366.1
- Quartos	1	709024419	1.2755e+10	2371.4
- Banheiros	1	1391878768	1.3437e+10	2378.1
- Ofertas	1	5830560660	1.7876e+10	2414.6
- factor(Bairro)	2	7844585461	1.9890e+10	2426.3
- Feita_tijolos	1	7648281675	1.9694e+10	2427.0
- `Tamanho (pes)`	1	8573167475	2.0619e+10	2432.9

```
Call:
lm(formula = Preco ~ ofertas + `Tamanho (pes)` + Quartos + Banheiros +
  factor(Bairro) + Feita_tijolos)
```

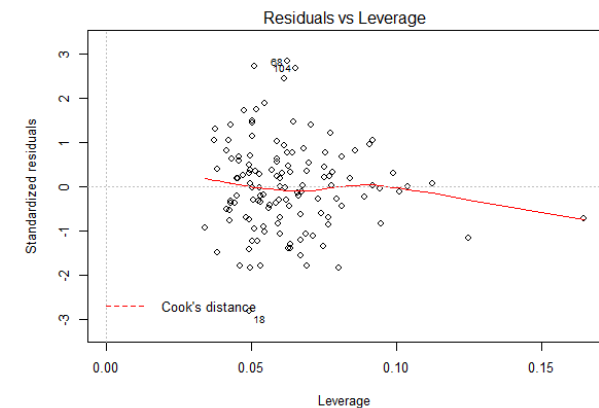
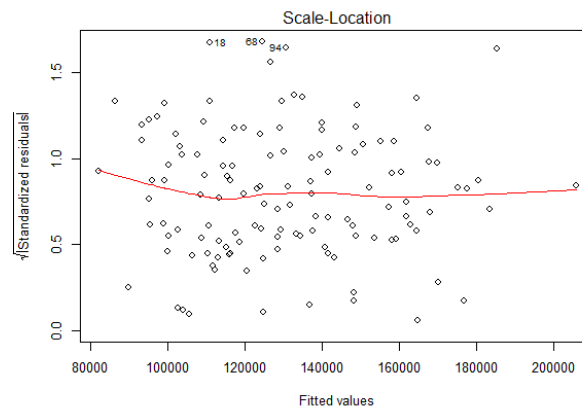
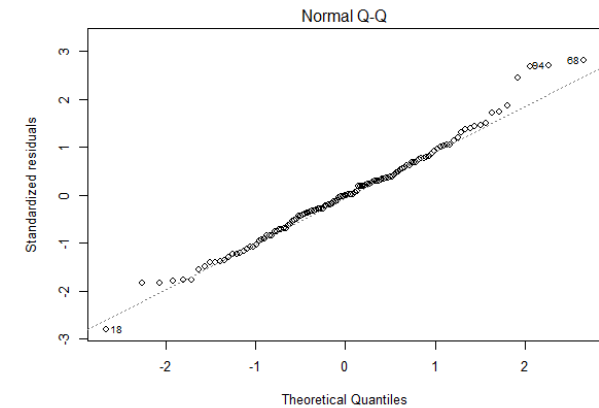
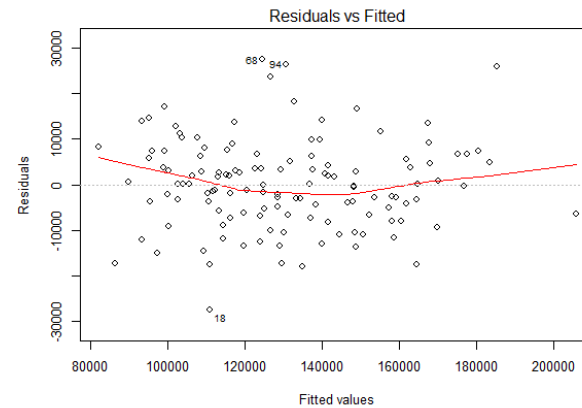
Coefficients:

	ofertas	`Tamanho (pes)`	Quartos	Banheiros	factor(Bairro)2
(Intercept)	2159.50	-8267.49	52.99	4246.79	7883.28
factor(Bairro)3	20681.04	17297.35			-1560.58



# Análise dos Resíduos e Suposições

Script  
`par(mfrow=c(2,2))`  
`plot(ajuste)`



# Problemas de Escala

## Script

```
ajuste <- lm(scale(Preco) ~ Ofertas + scale(`Tamanho (pes)`) + Quartos + Banheiros + factor(Bairro) + Feita_tijolos)
```

```
summary(ajuste)
```

```
vif(ajuste)
```

```
stepAIC(ajuste, direction = "both")
```

```
par(mfrow=c(2,2))
```

```
plot(ajuste)
```

```
> summary(ajuste)

Call:
lm(formula = scale(Preco) ~ Ofertas + scale(`Tamanho (pes)`) +
    Quartos + Banheiros + factor(Bairro) + Feita_tijolos)

Residuals:
    Min       1Q   Median       3Q      Max
-1.01744 -0.24376 -0.00155  0.21599  1.01826

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.82738    0.24895  -3.323  0.00118 **
Ofertas        -0.30770    0.04037  -7.621 6.47e-12 ***
scale(`Tamanho (pes)`) 0.41729    0.04515   9.242 1.10e-15 ***
Quartos         0.15806    0.05947   2.658  0.00894 **
Banheiros       0.29340    0.07879   3.724  0.00030 ***
factor(Bairro)2 -0.05808    0.08920  -0.651  0.51621
factor(Bairro)3  0.76971    0.11720   6.568 1.38e-09 ***
Feita_tijolosYes 0.64377    0.07375   8.729 1.78e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3729 on 120 degrees of freedom
Multiple R-squared:  0.8686,    Adjusted R-squared:  0.861
F-statistic: 113.3 on 7 and 120 DF, p-value: < 2.2e-16
```



# Problemas de Escala

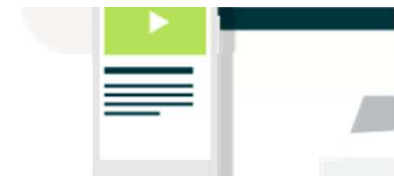
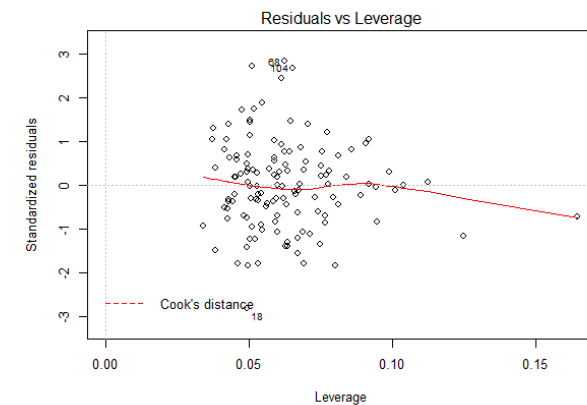
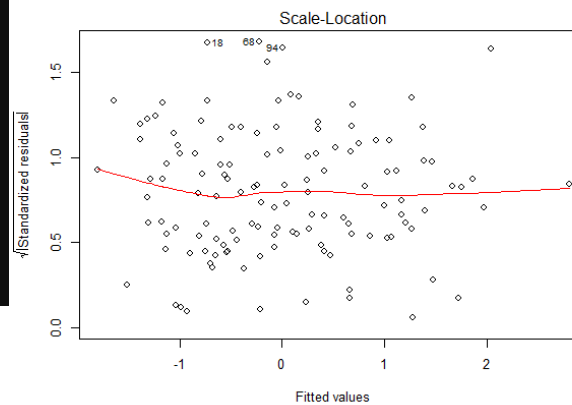
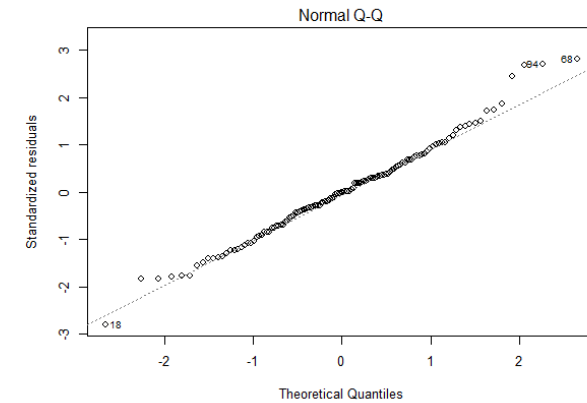
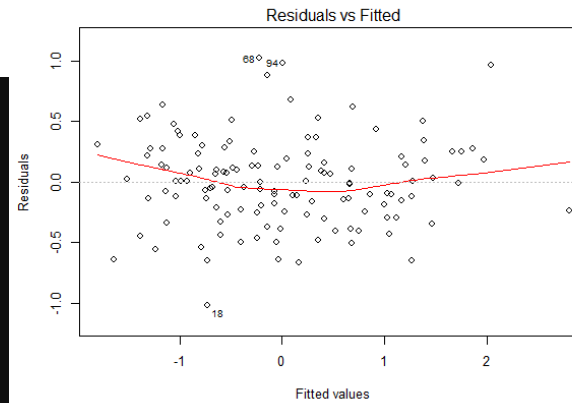


```
> stepAIC(ajuste, direction = "both")
Start: AIC=-244.8
scale(Preco) ~ ofertas + scale('Tamanho (pes)') + Quartos + Banheiros +
  factor(Bairro) + Feita_tijolos
```

	Df	Sum of Sq	RSS	AIC
<none>			16.685	-244.80
- Quartos	1	0.9821	17.667	-239.48
- Banheiros	1	1.9280	18.613	-232.81
- ofertas	1	8.0763	24.762	-196.27
- factor(Bairro)	2	10.8661	27.551	-184.61
- Feita_tijolos	1	10.5942	27.279	-183.88
- scale('Tamanho (pes)')	1	11.8753	28.561	-178.00

```
Call:
lm(formula = scale(Preco) ~ ofertas + scale('Tamanho (pes)') +
  quartos + Banheiros + factor(Bairro) + Feita_tijolos)

Coefficients:
(Intercept)      ofertas scale('Tamanho (pes)')
-0.82738      -0.30770      0.41729
0.15806        0.29340
factor(Bairro)2 factor(Bairro)3 Feita_tijolosYes
-0.05808        0.76971      0.64377
```





# Análise Preditiva

Gabriel Vinícius Araújo Fonseca

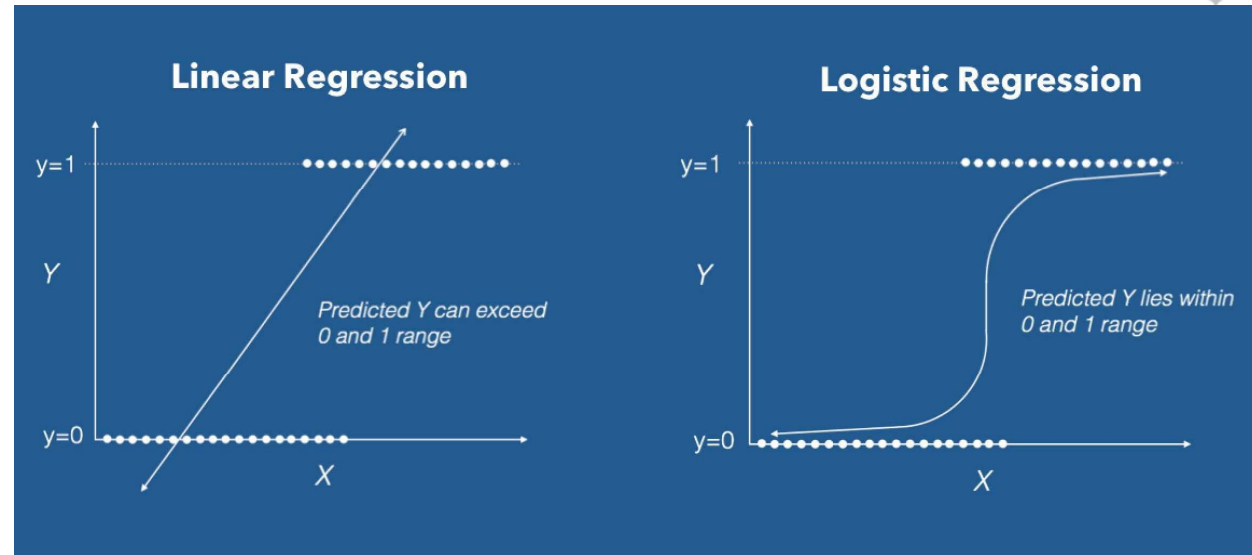


# Regressão Logística Simples



# Regressão Logística

- Método muito usado para classificar indivíduos através de variáveis exploratórias
- Estima a probabilidade de ocorrência do evento



<https://www.datacamp.com/community/tutorials/logistic-regression-R>



# Regressão Logística

- Variável resposta categórica
  - Binária: duas categorias
  - Multinomial: 3 ou mais
- Usada para criar score de probabilidade de evento estudado
- O exponencial dos coeficientes dá o Odds Ratio (Risco)



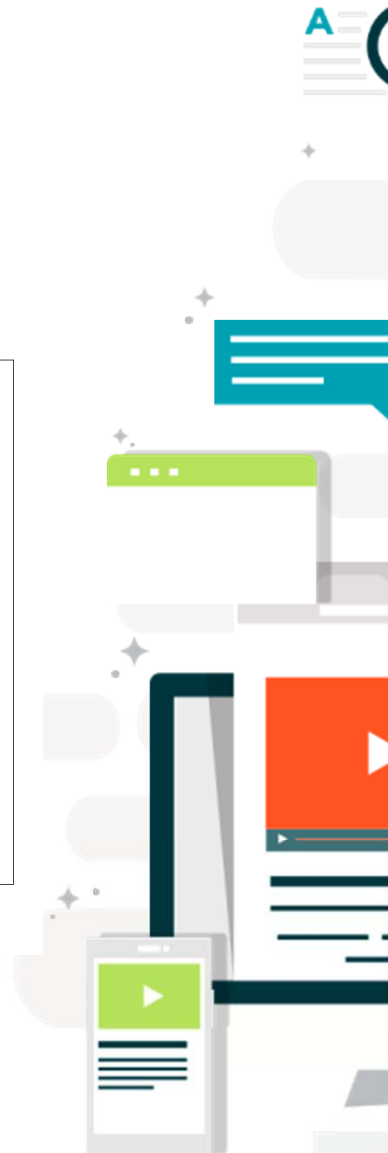
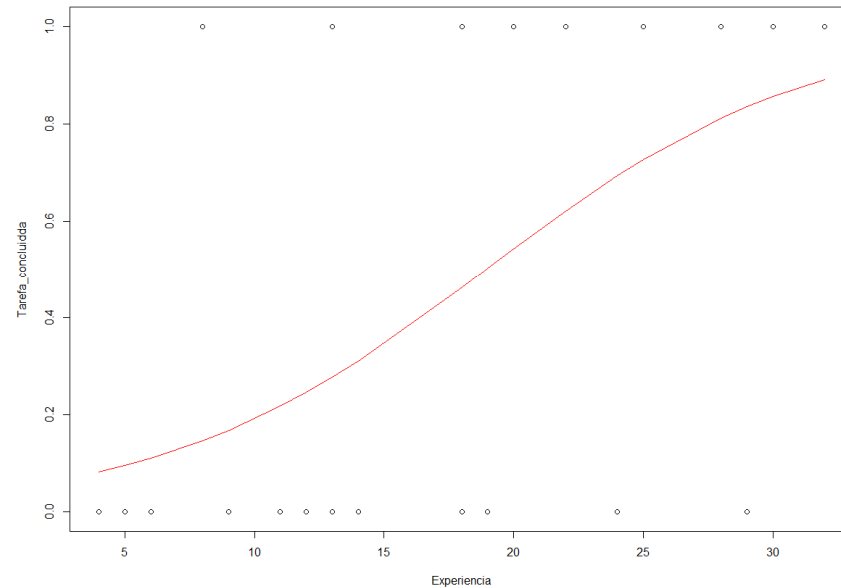
# Regressão Logística Simples

- Equação de Regressão Logística Simples

- $\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1$

- $p = \frac{e^{\beta_0 + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_1}}$

- Odds Ratio:  $\frac{p}{1-p}$



# Procedimentos

- Definir corretamente a variável resposta do tipo binária (0 ou 1).
- Definir a sua variável preditora quantitativa contínua.
- Ajuste do modelo.
- Coeficiente significativos e do modelo.
- Avaliação das Suposições e do Resíduo.



# Exemplos

- Será que os anos de experiência aumenta as chances de contratação em uma determinada empresa?
- Será que o excesso de peso aumenta as chances de sofrer um infarto?
- Será que ingerir menos açúcares diminui as chances de ter diabetes?



**E como podemos fazer  
isso no Rstudio?**





# Análise Preditiva

Gabriel Vinícius Araújo Fonseca



# Exemplo de Regressão Logística Simples





# Procedimentos

- Definir corretamente a variável resposta do tipo binária (0 ou 1).
- Definir a sua variável preditora quantitativa contínua.
- Ajuste do modelo.
- Coeficiente significativos e do modelo.
- Avaliação das Suposições e do Resíduo.



# Exemplo

- Um professor está selecionando alguns alunos que trabalham na área de computação para desenvolver um programa para o seu novo projeto. Ele queria saber se o tempo de experiência influencia no término de uma tarefa (um programa teste) a ser feito pelos alunos.
- **Arquivo Experiencia\_Tarefa.xlsx**



# Exemplo

- A função no R que fará o ajuste do modelo chama “glm” já pré-instalado.

## Script

```
attach(Experiencia_Tarefa)
ajuste <- glm(Tarefa_concluida ~ Experiencia,
family = binomial)
summary(ajuste)
```

Modelo para estimativa da probabilidade de concluir a tarefa:

$$\hat{p} = \frac{e^{-3,06+0,16X_1}}{1 + e^{-3,06+0,16X_1}}$$

```
> ajuste <- glm(Tarefa_concluida ~ Experiencia, family = binomial)
> summary(ajuste)

Call:
glm(formula = Tarefa_concluida ~ Experiencia, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8992  -0.7509  -0.4140   0.7992   1.9624

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.05970    1.25935  -2.430   0.0151 *
Experiencia  0.16149    0.06498   2.485   0.0129 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.296  on 24  degrees of freedom
Residual deviance: 25.425  on 23  degrees of freedom
AIC: 29.425

Number of Fisher Scoring iterations: 4
```

# Exemplo

- Avaliando a qualidade do ajuste e a razão de chances (*odds*).

## Script

```
anova(ajuste, test = 'Chisq')
require(MASS)
exp(cbind(coef(ajuste), confint.default(ajuste)))
```

$$odds = 1,1753$$

Aumento de 17,5% nas chances de concluir a tarefa a cada ano a mais de experiência.

```
> anova(ajuste, test = 'Chisq')
Analysis of Deviance Table

Model: binomial, link: logit
Response: Tarefa_concluida

Terms added sequentially (first to last)

              Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL              24      34.296
Experiencia  1    8.8719      23      25.425 0.002896 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> require(MASS)
Carregando pacotes exigidos: MASS
Warning message:
package 'MASS' was built under R version 3.5.1
> exp(cbind(coef(ajuste), confint.default(ajuste)))
              2.5 %      97.5 %
(Intercept) 0.04690196 0.003974024 0.5535432
Experiencia 1.17525591 1.034716464 1.3348840
```

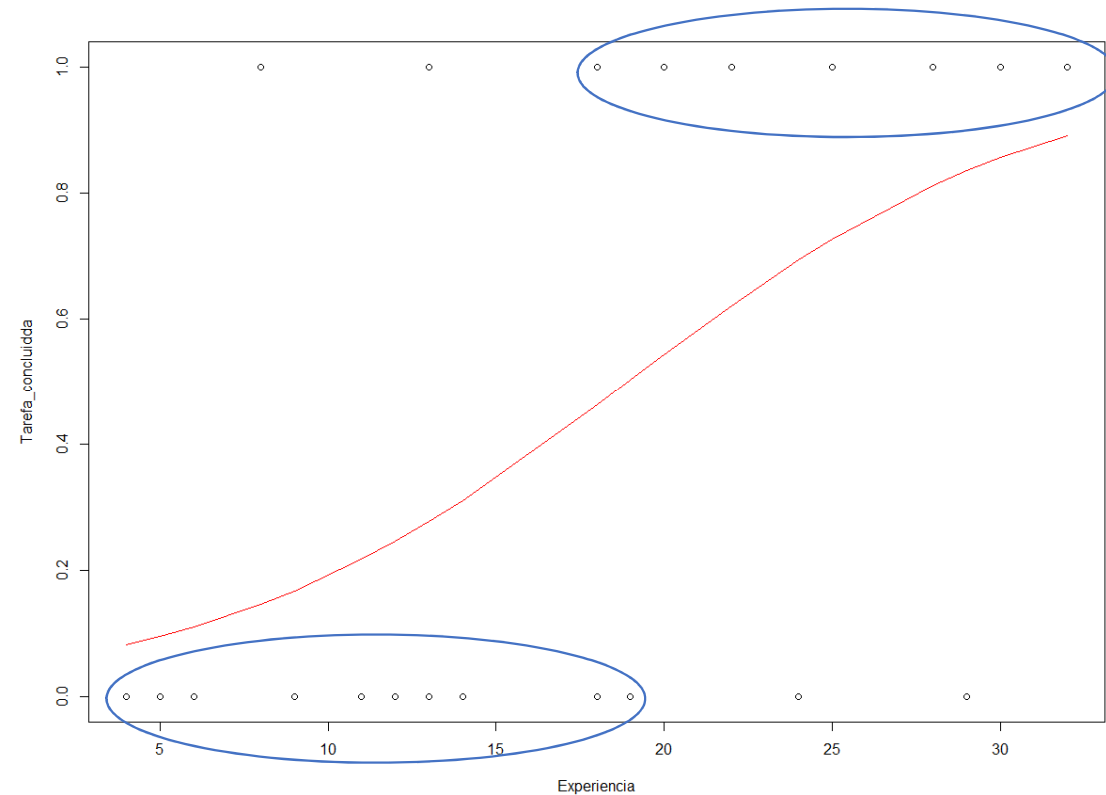
# Exemplo

- Gráfico dos valores preditos

## Script

```
plot(Experiencia_Tarefa)  
y_chapeu <- data.frame(  
  X = Experiencia_Tarefa$Experiencia,  
  y_chapeu = ajuste$fitted.values)  
y_chapeu <- y_chapeu[order(y_chapeu$X),]  
lines(y_chapeu, col = 2)  
y_chapeu
```

$$\hat{p} = \frac{e^{-3,06+0,16X_1}}{1 + e^{-3,06+0,16X_1}}$$

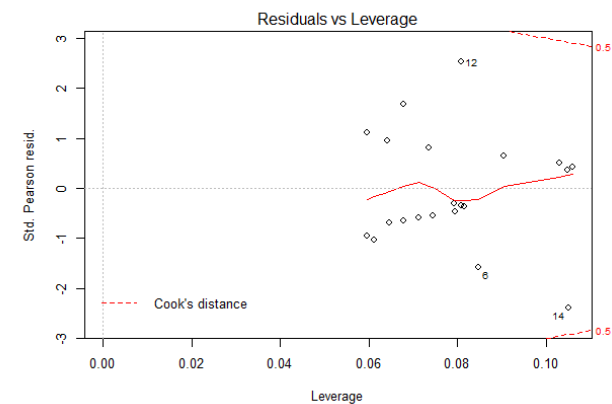
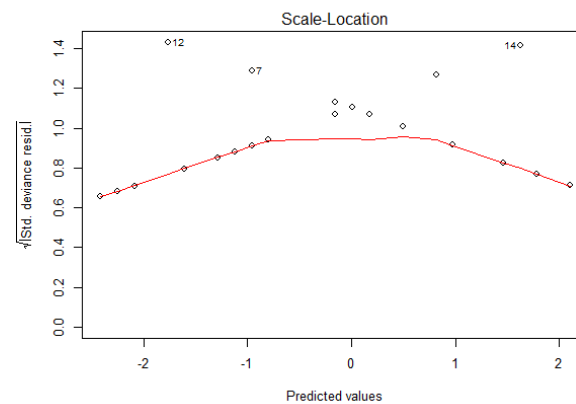
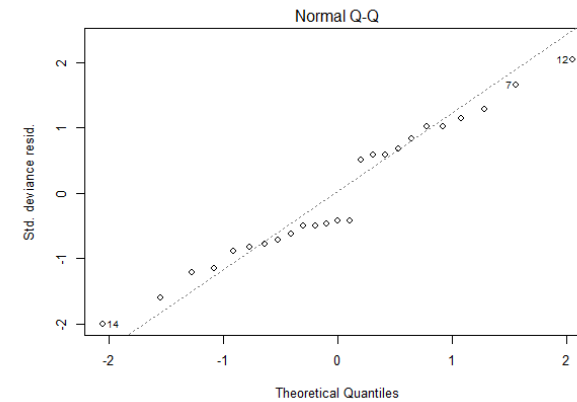
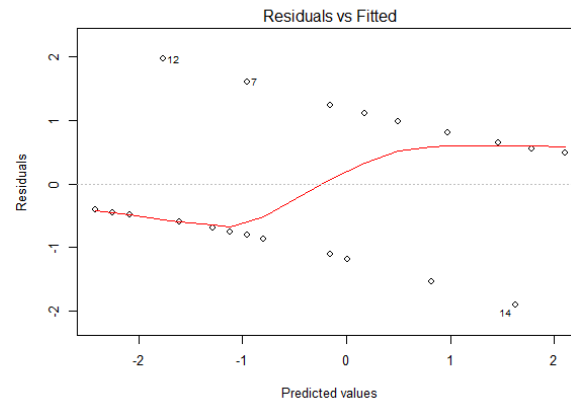


# Exemplo

- Verificação das Suposições e Resíduos

## Script

```
par(mfrow=c(2,2))  
plot(ajuste)
```



# Exemplo

- Nível de Acertos nos valores preditos ( $\hat{p}$ )

## Script

```
y_chapeu <- data.frame(X = Experiencia_Tarefa$Experiencia, y_chapeu =  
ajuste$fitted.values)  
y_chapeu$conclusão <- ifelse(y_chapeu$y_chapeu > .5, 1, 0)  
y_chapeu$y <- Tarefa_concluida  
table(y_chapeu[,3:4])  
mean(y_chapeu$conclusão == y_chapeu$y)
```

```
> y_chapeu <- data.frame(X = Experiencia_Tarefa$Experiencia, y_chapeu = ajuste$fitted.values)  
> y_chapeu$conclusão <- ifelse(y_chapeu$y_chapeu > .5, 1, 0)  
> y_chapeu$y <- Tarefa_concluida  
> table(y_chapeu[,3:4])  
      y  
conclusão 0  1  
      0 11  3  
      1  3  8  
> mean(y_chapeu$conclusão == y_chapeu$y)  
[1] 0.76
```

**E se caso tivéssemos  
mais de uma variável  
preditora?**







# Análise Preditiva

Gabriel Vinícius Araújo Fonseca



# Regressão Logística Composta ou Multivariada



# Regressão Logística Composta

- Equação de Regressão Logística Composta
- $\ln \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 X_1 \cdots + \beta_k X_k$
- $p = \frac{e^{\beta_0 + \beta_1 X_1 \cdots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 \cdots + \beta_k X_k}}$
- Odds Ratio:  $\frac{p}{1-p}$  para cada variável X ( $\exp(\beta_k)$ ).



# Tipos de Variáveis Preditoras

- Quantitativas
  - Contínuas
  - Discretas
- Qualitativas
  - Transformar em variáveis *Dummys*
  - Se tivermos 4 categorias -> 3 variáveis *Dummys*
  - Exemplo: Bom, Regular e Ruim (variável X qualquer de satisfação).
  - *Dummy 1: 0 para não for Bom e 1 para for Bom.*
  - *Dummy 2: 0 para não for Regular e 1 para for Regular.*
  - Não precisamos da terceira *Dummy*, pois quando a *Dummy 1* e 2 forem ambas iguais a 0, automaticamente teremos a categoria Ruim.



# Procedimentos

- Definir corretamente a variável resposta do tipo binária (0 ou 1).
- Definir as variáveis preditoras.
- Ajuste do modelo.
- Coeficiente significativos e do modelo.
- Interpretação dos coeficientes.
- Verificação da Multicolinearidade.
- Avaliação das Suposições e do Resíduo.



# Exemplos

- Considere os seguintes dados de amostra selecionada aleatoriamente de estudantes que pretendem ingressar em uma determinada universidade, que incluem sexo (gênero) codificado 0 para indivíduos do sexo feminino e 1 para indivíduos do sexo masculino, média de notas no ensino médio (Hsavg) medida em típica escala de 100 pontos e resultado (decisão) da admissão codificado como 0 para rejeitado e 1 para admitido.
- A probabilidade (chance) de ser ou não admitido está atrelado ao gênero e à pontuação no ensino médio?



# Exemplo

- Os clientes potenciais de uma empresa foram acompanhados por uma pesquisa para saber o perfil dos compradores de uma lasanha congelada.
- A empresa quer entender porque alguns clientes potenciais são compradores e outros não. O gênero faz alguma diferença? A renda faz a diferença? Em geral, o que distingue significativamente os clientes de compradores e não compradores?
- Foram usados os dados demográficos sobre esses clientes para construir um modelo preditivo sobre a chance de comprarem ou não a sua lasanha.





# Como construir um modelo Logístico Preditivo no Rstudio?





# Análise Preditiva

Gabriel Vinícius Araújo Fonseca



# Exemplo de Regressão Logística Composta



# Procedimentos

- Definir corretamente a variável resposta do tipo binária (0 ou 1).
- Definir as variáveis preditoras.
- Ajuste do modelo.
- Coeficiente significativos e do modelo.
- Interpretação dos coeficientes.
- Verificação da Multicolinearidade.
- **Stepwise**
- Avaliação das Suposições e do Resíduo.



# Exemplo

- Os clientes potenciais de uma empresa foram acompanhados por uma pesquisa para saber o perfil dos compradores de uma lasanha congelada.
- A empresa quer entender porque alguns clientes potenciais são compradores e outros não. O gênero faz alguma diferença? A renda faz a diferença? Em geral, o que distingue significativamente os clientes de compradores e não compradores?
- Foram usados os dados demográficos sobre esses clientes para construir um modelo preditivo sobre a chance de comprarem ou não a sua lasanha (**Lasagna Triers.xlsx**).



# Definição das Variáveis

Nome da Coluna	Descrição
Person	Identificação da Pessoa
Age	Idade em Anos
Weight	Peso em Libras
Income	Rendimentos em Dólares
Pay Type	Tipo de Recebimento (Horista ou Salariado)
Car Value	Valor do Carro em Dólares
CC Debt	Dívida no Cartão de Crédito em Dólares
Gender	Sexo (Masculino e Feminino)
Live Alone	Se vive ou não sozinho
Dwell Type	Tipo de Moradia (Apartamento, Casa ou)
Mall Trips	Número de Visitas ao Shopping
Nbhd	Região do Bairro (Oeste, Leste e Sul)
Have Tried	Se adquiriram ou não a Lasanha

Total de 856 indivíduos



# Leitura dos Dados

- Primeiro, faremos a leitura dos dados

## Script

```
library(readxl)
```

```
Lasagna_Triers <- read_excel("Local no seu PC/Lasagna Triers.xlsx")
```

```
View(Lasagna_Triers)
```

	Person	Age	Weight	Income	Pay Type	Car Value	CC Debt	Gender	Live Alone	Dwell Type	Mall Trips	Nbhd	Have Tried
1	1	48	175	65500	Hourly	2190	3510	Male	No	Home	7	East	No
2	2	33	202	29100	Hourly	2110	740	Female	No	Condo	4	East	Yes
3	3	51	188	32200	Salaried	5140	910	Male	No	Condo	1	East	No
4	4	56	244	19000	Hourly	700	1620	Female	No	Home	3	West	No
5	5	28	218	81400	Salaried	26620	600	Male	No	Apt	3	West	Yes
6	6	51	173	73000	Salaried	24520	950	Female	No	Condo	2	East	No
7	7	44	182	66400	Salaried	10130	3500	Female	Yes	Condo	6	West	Yes
8	8	29	189	46200	Salaried	10250	2860	Male	No	Condo	5	West	Yes
9	9	28	200	61100	Salaried	17210	3180	Male	No	Condo	10	West	Yes



# Ajustando o Modelo

## Script

```
attach(Lasagna_Triers)
Lasagna_Triers$Compra <- ifelse(Lasagna_Triers$`Have Tried`=="Yes", 1, 0)
ajuste <- glm(Compra ~ Age + Weight + Income + `Pay Type` +
              `Car Value` + `CC Debt` + Gender + `Live Alone` +
              `Dwell Type` + `Mall Trips` + Nbhd, family = binomial)
```

```
> attach(Lasagna_Triers)
The following objects are masked from Lasagna_Triers (pos = 3):
  Age, Car Value, CC Debt, Dwell Type, Gender, Have Tried, Income, Live Alone, Mall Trips, Nbhd, Pay
  Type, Person, weight

The following objects are masked from Lasagna_Triers (pos = 4):
  Age, Car Value, CC Debt, Dwell Type, Gender, Have Tried, Income, Live Alone, Mall Trips, Nbhd, Pay
  Type, Person, weight

> Lasagna_Triers$Compra <- ifelse(Lasagna_Triers$`Have Tried`=="Yes", 1, 0)
> ajuste <- glm(Compra ~ Age + Weight + Income + `Pay Type` +
+               `Car Value` + `CC Debt` + Gender + `Live Alone` +
+               `Dwell Type` + `Mall Trips` + Nbhd, family = binomial)
```

# Ajustando o Modelo

## Script

summary(ajuste)

```
> summary(ajuste)

call:
glm(formula = Compra ~ Age + Weight + Income + `Pay Type` + `Car Value` +
  `CC Debt` + Gender + `Live Alone` + `Dwell Type` + `Mall Trips` +
  Nbhd, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6076  -0.5016   0.1274   0.5000   2.5009

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.311e+00  9.888e-01  -3.348  0.000814 ***
Age          -6.896e-02  1.150e-02  -5.995  2.03e-09 ***
Weight        5.195e-03  4.195e-03   1.238  0.215631
Income        4.004e-06  4.032e-06   0.993  0.320748
`Pay Type`Salaried  1.391e+00  2.383e-01   5.838  5.28e-09 ***
`Car Value`   -2.388e-05  2.153e-05  -1.109  0.267491
`CC Debt`     6.282e-05  1.012e-04   0.621  0.534727
GenderMale    3.719e-01  2.069e-01   1.798  0.072187 .
`Live Alone`Yes  1.282e+00  3.066e-01   4.181  2.90e-05 ***
`Dwell Type`Condo -8.187e-02  2.937e-01  -0.279  0.780459
`Dwell Type`Home  1.691e-01  2.625e-01   0.644  0.519359
`Mall Trips`    6.927e-01  6.350e-02  10.908  < 2e-16 ***
NbhdSouth      8.628e-01  2.620e-01   3.293  0.000990 ***
Nbhdwest       2.108e+00  2.478e-01   8.509  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1165.60  on 855  degrees of freedom
Residual deviance:  604.92  on 842  degrees of freedom
AIC: 632.92

Number of Fisher Scoring iterations: 6
```



# Stepwise

- Visto que temos algumas variáveis que não foram significativa no modelo, rodaremos o stepwise para a seleção do melhor modelo.

## Script

```
library(MASS)  
stepAIC(ajuste, direction = "both")
```

```
Call: glm(formula = Compra ~ Age + `Pay Type` + Gender + `Live Alone` +  
  `Mall Trips` + Nbhd, family = binomial)  
  
Coefficients:  
      (Intercept)              Age  `Pay Type`Salaried      GenderMale  `Live Alone`Yes  
          -2.2433          -0.0667           1.3302           0.3885           1.2247  
  `Mall Trips`      NbhdSouth      NbhdWest  
           0.7045           0.8647           2.1346  
  
Degrees of Freedom: 855 Total (i.e. Null); 848 Residual  
Null Deviance: 1166  
Residual Deviance: 609.3      AIC: 625.3
```

# Ajustando o novo Modelo

## Script

```
ajuste2 <- glm(Compra ~ Age +  
  `Pay Type` +  
  Gender +  
  `Live Alone` +  
  `Mall Trips` +  
  Nbhd, family = binomial)  
summary(ajuste2)
```

```
> ajuste2 <- glm(Compra ~ Age + `Pay Type` + Gender + `Live Alone` +  
+ `Mall Trips` + Nbhd, family = binomial)  
> summary(ajuste2)  
  
Call:  
glm(formula = Compra ~ Age + `Pay Type` + Gender + `Live Alone` +  
  `Mall Trips` + Nbhd, family = binomial)  
  
Deviance Residuals:  
    Min       1Q   Median       3Q      Max   
-2.6084  -0.5182   0.1244   0.5045   2.4211  
  
Coefficients:  
                Estimate Std. Error z value Pr(>|z|)      
(Intercept)    -2.24326    0.55379  -4.051 5.11e-05 ***  
Age             -0.06670    0.01135  -5.878 4.16e-09 ***  
`Pay Type`Salaried  1.33020    0.20859   6.377 1.81e-10 ***  
GenderMale       0.38853    0.20556   1.890 0.058745 .  
`Live Alone`Yes  1.22474    0.29571   4.142 3.45e-05 ***  
`Mall Trips`     0.70455    0.05837  12.070 < 2e-16 ***  
NbhdSouth       0.86471    0.25939   3.334 0.000857 ***  
NbhdWest        2.13458    0.24674   8.651 < 2e-16 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 1165.60 on 855 degrees of freedom  
Residual deviance: 609.34 on 848 degrees of freedom  
AIC: 625.34
```

# Interpretando os coeficientes

- Calculando os *odds* e interpretando os valores

## Script

```
require(MASS)  
exp(cbind(coef(ajuste2), confint.default(ajuste2)))
```

```
> exp(cbind(coef(ajuste2), confint.default(ajuste2)))  
                                2.5 %    97.5 %  
(Intercept)      0.1061125 0.03584092 0.3141624  
Age               0.9354730 0.91489586 0.9565129  
'Pay Type'`Salaried` 3.7818079 2.51272666 5.6918532  
GenderMale        1.4748123 0.98574019 2.2065360  
'Live Alone'`Yes`   3.4032695 1.90627663 6.0758459  
'Mall Trips'        2.0229283 1.80423729 2.2681268  
NbhdSouth         2.3743144 1.42806722 3.9475515  
Nbhdwest          8.4534751 5.21205594 13.7107588
```



# Qualidade do Ajuste do Modelo

- Teste Chi-quadrado para verificação da qualidade do modelo

## Script

```
pchisq(ajuste2$deviance, ajuste2$df.residual, lower.tail = F) #Teste Chi-quadrado do Deviance  
pchisq(ajuste2$null.deviance - ajuste2$deviance,  
       ajuste2$df.null - ajuste2$df.residual, lower.tail = F) #Teste Chi-quadrado da Regressão
```

```
> pchisq(ajuste2$deviance, ajuste2$df.residual, lower.tail = F) #Teste Chi-quadrado do Deviance  
[1] 1  
> pchisq(ajuste2$null.deviance - ajuste2$deviance,  
+       ajuste2$df.null - ajuste2$df.residual, lower.tail = F) #Teste Chi-quadrado da Regressão  
[1] 1.588155e-110
```



# Verificação de Multicolinearidade

- Usando o vif para identificação da multicolinearidade.

## Script

```
library(car)  
vif(ajuste2)
```

```
> library(car)  
Carregando pacotes exigidos: carData  
Warning message:  
package 'car' was built under R version 3.5.1  
> vif(ajuste2)
```

	GVIF	Df	GVIF <sup>1/(2*Df)</sup>
Age	1.026665	1	1.013245
`Pay Type`	1.044361	1	1.021940
Gender	1.013654	1	1.006804
`Live Alone`	1.040894	1	1.020242
`Mall Trips`	1.089323	1	1.043706
Nbhd	1.062575	2	1.015290



# Verificação do Coeficiente de Determinação

- Uma alternativa para a estimação do coeficiente de determinação ( $R^2$ ), utilizamos um pacote chamado pscl.

## Script

```
install.packages('pscl')  
library(pscl)  
pR2(ajuste2)
```

```
> library(pscl)  
> pR2(ajuste2)
```

	llh	llhNull	G2	McFadden	r2ML	r2CU
	-304.6687217	-582.8024065	556.2673697	0.4772350	0.4778733	0.6424997

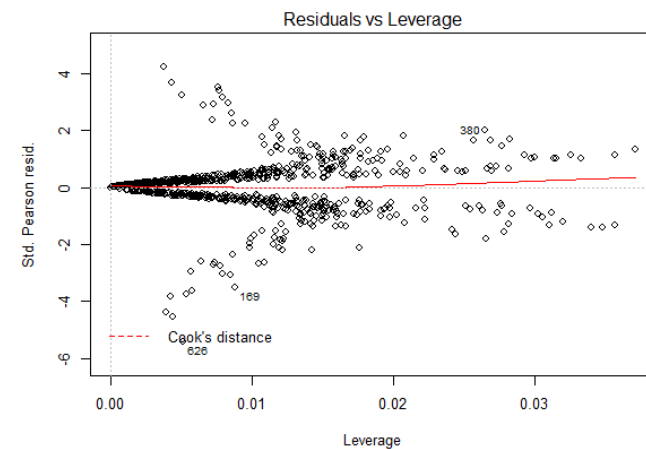
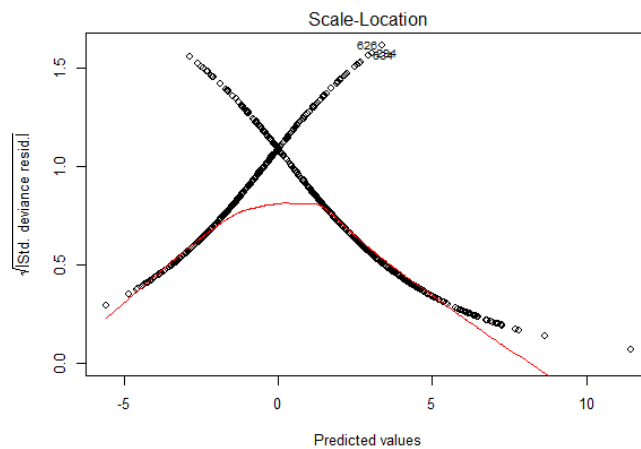
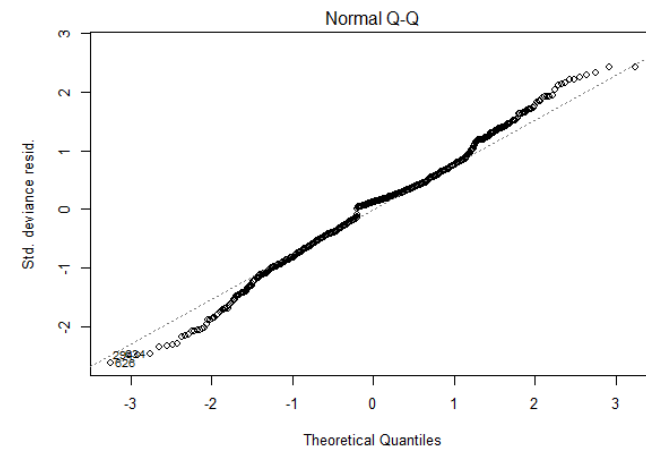
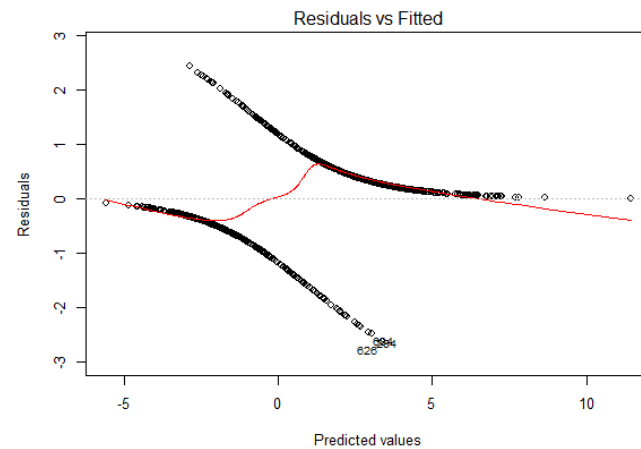




# Verificação das Suposições e Resíduos

## Script

```
par(mfrow=c(2,2))  
plot(ajuste2)
```



# Estimando uma probabilidade

- Suponha que tenho um indivíduo, de 20 anos, salariado, sexo masculino, que vive sozinho, chega a ir 10 vezes ao shopping e mora na região oeste. Qual a probabilidade de comprar a lasanha do fabricante?

## Script

```
coef2 <- data.frame(coef(ajuste2))
p_hat <- (exp(coef2[1,] + coef2[2,]*20 + coef2[3,] + coef2[4,] +
  coef2[5,] + coef2[6,]*10 + coef2[8,]))/
  (1+exp(coef2[1,] + coef2[2,]*20 + coef2[3,] + coef2[4,] +
  coef2[5,] + coef2[6,]*10 + coef2[8,]))
p_hat
```

```
> coef2 <- data.frame(coef(ajuste2))
> p_hat <- (exp(coef2[1,] + coef2[2,]*20 + coef2[3,] + coef2[4,] +
+   coef2[5,] + coef2[6,]*10 + coef2[8,]))/
+   (1+exp(coef2[1,] + coef2[2,]*20 + coef2[3,] + coef2[4,] +
+   coef2[5,] + coef2[6,]*10 + coef2[8,]))
> p_hat
[1] 0.9998058
> |
```

**Quais outros métodos  
de predição?**





# Análise Preditiva

Gabriel Vinícius Araújo Fonseca



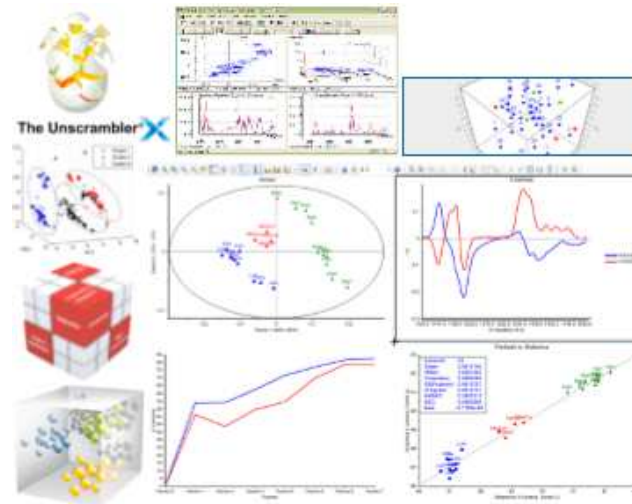
# Análise Multivariada de Dados



**A**

- 

An illustration featuring a laptop and a smartphone. The laptop screen shows a video player with a large orange play button. The smartphone screen also shows a video player with a green play button. The background is white with light gray rounded rectangles and small gray stars, suggesting a clean, modern digital environment.



# O que é Análise Multivariada?

- Se antes utilizamos apenas uma variável resposta sendo dependente de uma ou mais variáveis preditoras, para realizar uma previsão, agora o ponto forte da análise multivariada é encontrar relacionamento entre inúmeras variáveis ou encontrar quais variáveis não se correlacionam.





# Tipos de Análise Multivariadas

- Análise Fatorial
- Análise de Agrupamentos (Cluster)
- Análise Discriminante
- Análise de Correspondência
- Modelo de Equações Estruturais
- Entre outras



# Análise Fatorial

- **Objetivo**

- Analisar inter-relações entre um grande número de variáveis e explicar essas variáveis em termos de suas dimensões inerentes comuns (fatores).
- Encontrar um modo de condensar a informação contida em diversas variáveis originais em conjunto menor de novas dimensões, com perdas mínimas de informação.

- **Variáveis**

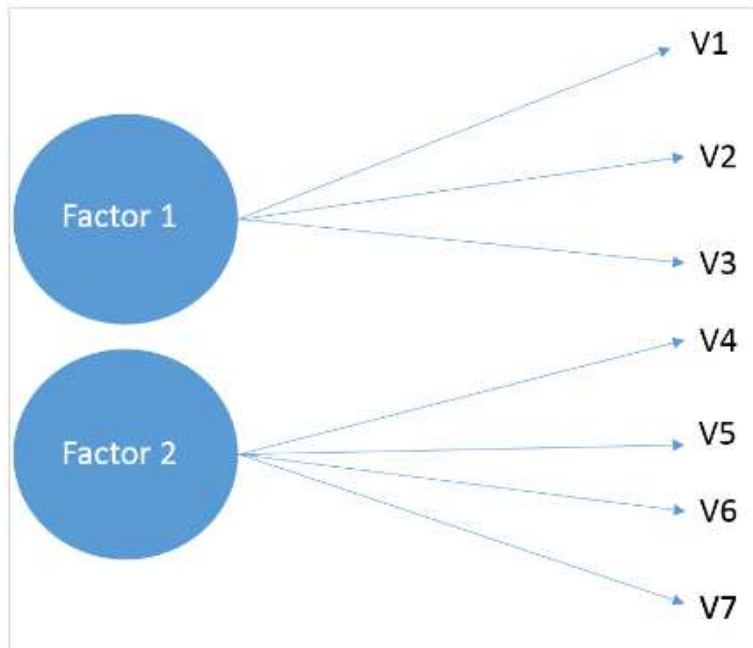
- Variáveis métricas ou quantitativas.
- Uma estrutura mínima de cinco variáveis por fator.

- **Utilizações**

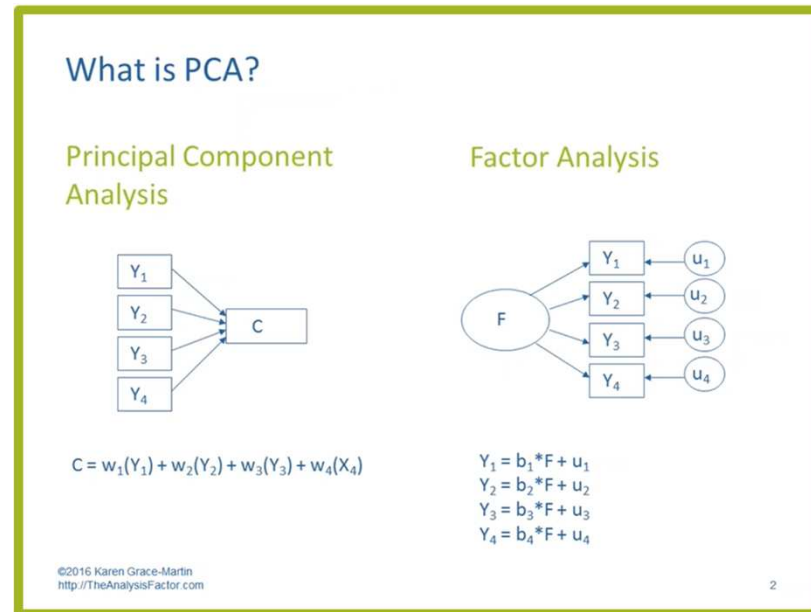
- Entender as relações entre as avaliações de clientes de uma lanchonete. O questionário pode ser dividido em sabor da comida, temperatura da comida, estética da comida, tempo de espera, limpeza e atendimento.



# Análise Fatorial e Componente Principal



<https://www.promptcloud.com/blog/exploratory-factor-analysis-in-r/>



<https://www.theanalysisfactor.com/factor-analysis-1-introduction/>

# Análise de Cluster

- Objetivo

- Definir a estrutura de dados colocando as observações mais parecidas em grupos (Cluster de Observações) ou colocando as variáveis mais parecidas em um grupo (Cluster de Variáveis).

- Variáveis

- Variáveis métricas ou quantitativas.
- Somente variáveis que se relacionam especificamente (Cluster de Observações).

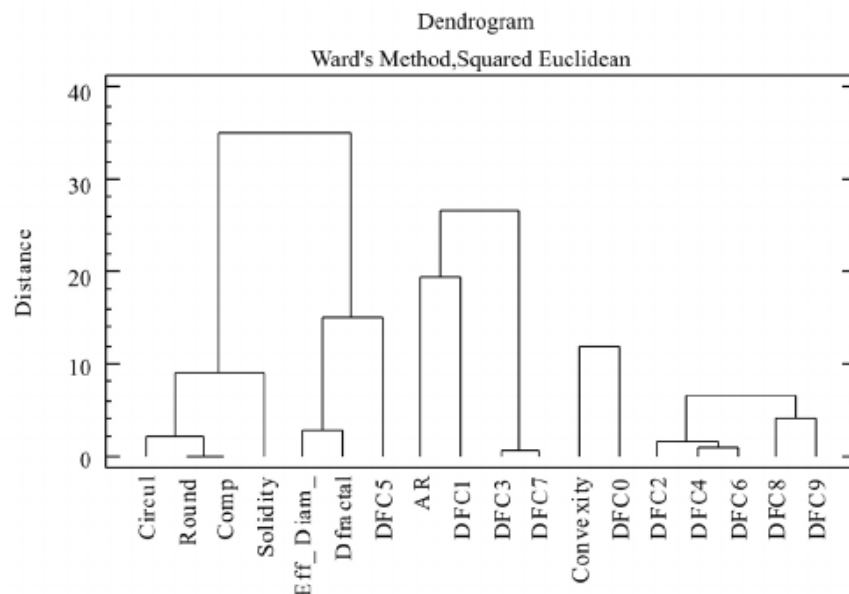
- Utilizações

- Pesquisa de Marketing que queira determinar segmentos de mercado em uma comunidade com base em padrões de lealdade a marcas e lojas.



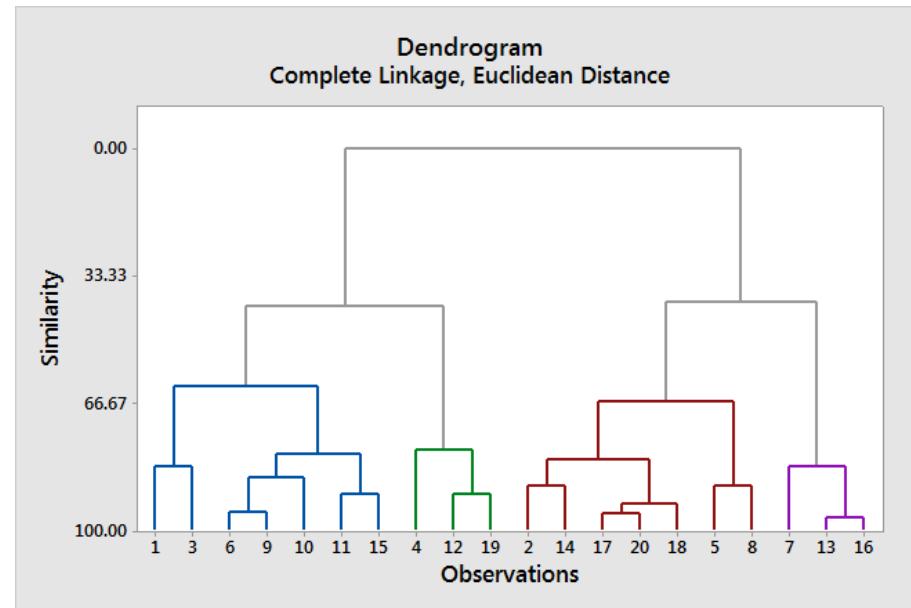
# Análise de Cluster

Por Variáveis



[https://www.researchgate.net/figure/Dendrogram-showing-groupings-of-the-variables-on-the-conglomerate-analysis-by-the-Ward\\_fig4\\_262612886](https://www.researchgate.net/figure/Dendrogram-showing-groupings-of-the-variables-on-the-conglomerate-analysis-by-the-Ward_fig4_262612886)

Por Observações



<https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-observations/interpret-the-results/all-statistics-and-graphs/dendrogram/>



# Análise Discriminante

- **Objetivo**

- Estabelecer procedimentos para classificar objetos (indivíduos, firmas , produtos, etc) em grupos, com base em seus escores em um conjunto de variáveis independentes.

- **Variáveis**

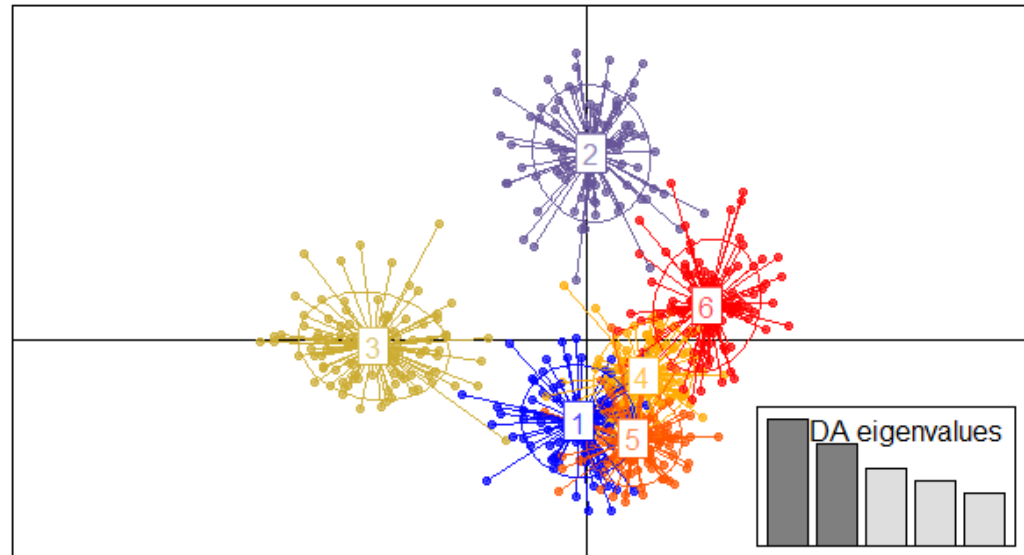
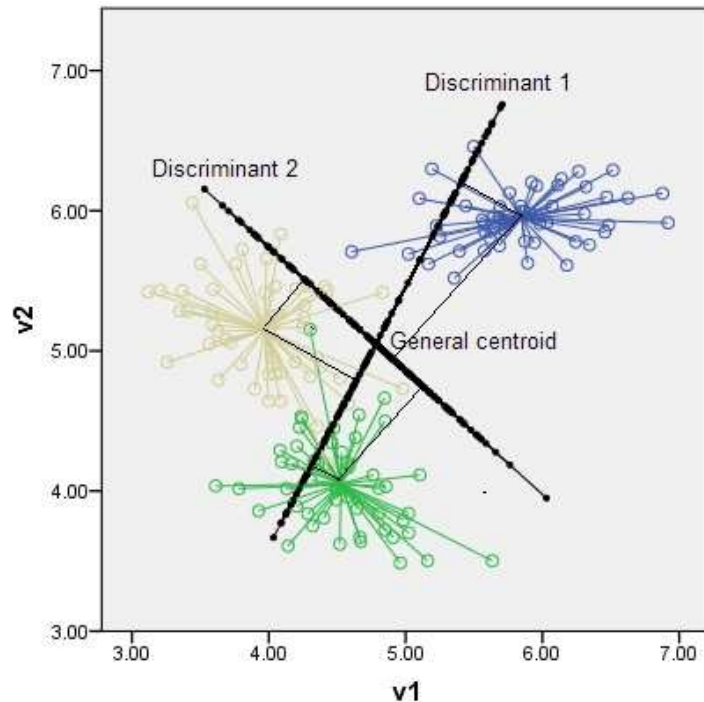
- Necessita de uma variável dependente categóricas excludentes (mesmo objeto em duas categorias).
- Variáveis independentes tanto categóricas quanto métricas que consiga diferenciar as categorias proposta na variável dependente.

- **Utilizações**

- Conseguir identificar as melhores características que distinguem um comprador de um não comprador.



# Análise Discriminante



<https://stats.stackexchange.com/questions/74098/r-package-to-make-a-linear-discriminant-analysis-scatter-plot>



# Como criar análises multivariadas no Rstudio?







# Análise Preditiva

Gabriel Vinícius Araújo Fonseca



# Análise Fatorial



# Análise Fatorial

- **Objetivo**

- Analisar inter-relações entre um grande número de variáveis e explicar essas variáveis em termos de suas dimensões inerentes comuns (fatores).
- Encontrar um modo de condensar a informação contida em diversas variáveis originais em conjunto menor de novas dimensões, com perdas mínimas de informação.

- **Variáveis**

- Variáveis métricas ou quantitativas.
- Uma estrutura mínima de cinco variáveis por fator.

- **Utilizações**

- Entender as relações entre as avaliações de clientes de uma lanchonete. O questionário pode ser dividido em sabor da comida, temperatura da comida, estética da comida, tempo de espera, limpeza e atendimento.



# Exemplo

- O arquivo EFA.xlsx corresponde a 90 respostas de 14 perguntas considerando a compra de um carro. As respostas foram computadas numa escala Likert de cinco pontos sendo 1 para muito baixo e 5 muito alto. As variáveis são:

Variável	Descrição
Price	Preço
Safety	Segurança
Exterior looks	Exterior
Space and comfort	Espaço e Conforto
Technology	Tecnologia
After sales service	Pós-venda
Resale value	Revenda

Variável	Descrição
Fuel type	Tipo de Combustível
Fuel efficiency	Consumo
Color	Cor
Maintenance	Manutenção
Test drive	Testdrive
Product reviews	Reviews
Testimonials	Depoimentos



# Leitura dos Dados

## Script

```
library(readxl)
EFA <- read_excel("D:/Google Drive/PUC Pós/PUC Virtual/Análise Preditiva/EFA.xlsx")
View(EFA)
```

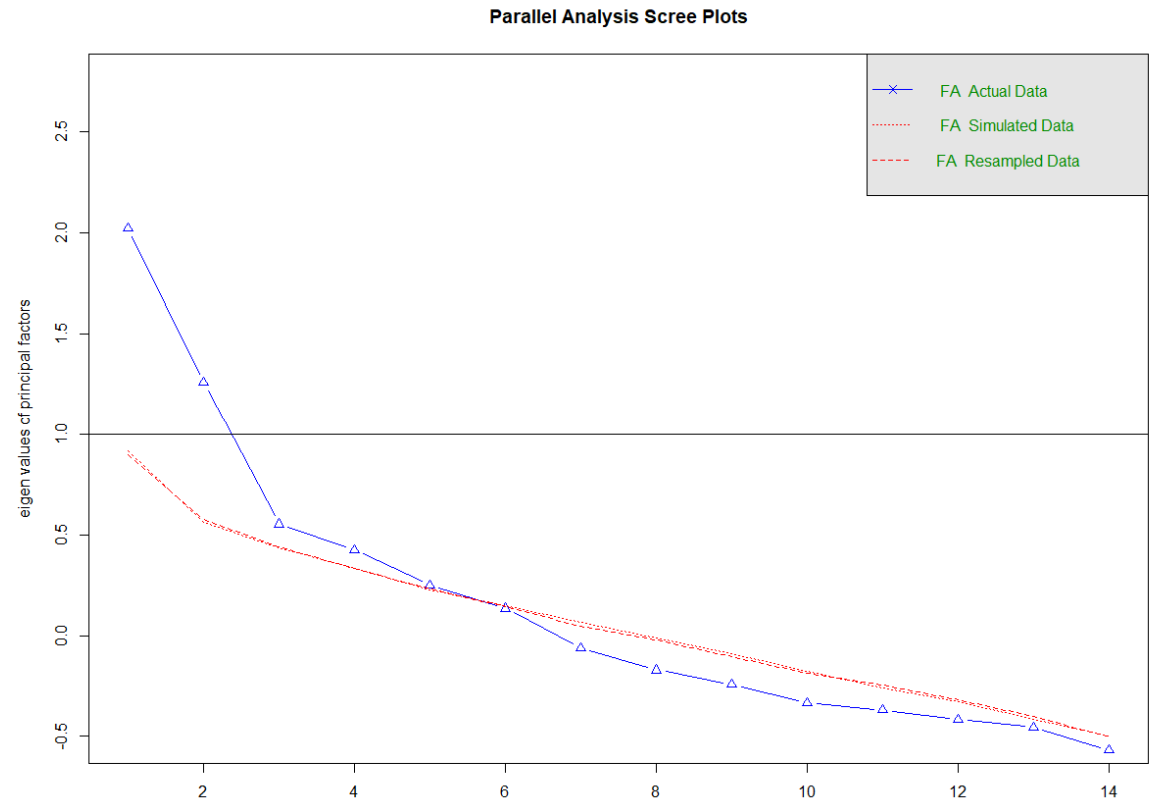
	Price	Safety	Exterior_Looks	Space_comfort	Technology	After_Sales_Service	Resale_Value
1	4	4	5	4	3	4	5
2	3	5	3	3	4	4	3
3	4	4	3	4	5	5	5
4	4	4	4	3	3	4	5
5	5	5	4	4	5	4	5



# Escolha do número de fatores

## Script

```
install.packages('psych')  
install.packages('GPArotation')  
library(psych)  
library(GPArotation)  
parallel <- fa.parallel(EFA, fm = 'pa', fa = 'fa')
```



Parallel analysis suggests that the number of factors = 5 and the number of components = NA

# Análise Fatorial

## Script

```
tresfatores <- fa(EFA, nfactors = 3, rotate = "oblimin", fm="pa")
tresfatores
```

## Rotação

Varimax => correlação entre os fatores

Oblimin => não existe correlação entre os fatores

```
> tresfatores <- fa(EFA, nfactors = 3, rotate = "oblimin", fm="pa")
> tresfatores
Factor Analysis using method = pa
Call: fa(r = EFA, nfactors = 3, rotate = "oblimin", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
```

	PA1	PA2	PA3	h2	u2	com
Price	0.44	0.12	-0.19	0.25	0.75	1.5
Safety	-0.23	0.31	-0.11	0.14	0.86	2.1
Exterior_Looks	-0.16	0.18	0.05	0.06	0.94	2.2
Space_comfort	-0.03	0.83	0.04	0.70	0.30	1.0
Technology	0.09	0.34	0.02	0.13	0.87	1.1
After_Sales_Service	0.25	0.46	0.00	0.29	0.71	1.5
Resale_Value	0.60	-0.16	-0.29	0.48	0.52	1.6
Fuel_Type	0.03	0.57	-0.13	0.32	0.68	1.1
Fuel_Efficiency	0.65	0.13	0.16	0.49	0.51	1.2
Color	0.46	-0.18	0.25	0.27	0.73	1.9
Maintenance	0.67	0.01	-0.06	0.45	0.55	1.0
Test_drive	0.19	0.14	0.33	0.19	0.81	2.0
Product_reviews	0.42	0.13	0.27	0.29	0.71	1.9
Testimonials	-0.03	-0.01	0.73	0.53	0.47	1.0





# Análise Fatorial

## Script

```
print(tresfatores$loadings, cutoff = .3)
```

```
> print(tresfatores$loadings, cutoff = .3)

Loadings:
          PA1    PA2    PA3
Price      0.444
Safety                    0.311
Exterior_Looks
Space_comfort                0.832
Technology                  0.342
After_Sales_Service        0.459
Resale_value      0.601
Fuel_Type                0.574
Fuel_Efficiency    0.654
Color              0.463
Maintenance        0.668
Test_drive                                0.330
Product_reviews    0.423
Testimonials                                0.732

          PA1    PA2    PA3
SS loadings 2.015 1.604 0.965
Proportion Var 0.144 0.115 0.069
Cumulative Var 0.144 0.259 0.327
```



# Análise Fatorial

## Script

```
quatrofatores <- fa(EFA, nfactors = 4, rotate = "oblimin", fm="pa")  
quatrofatores
```

```
> quatrofatores <- fa(EFA, nfactors = 4, rotate = "oblimin", fm="pa")  
> quatrofatores  
Factor Analysis using method = pa  
Call: fa(r = EFA, nfactors = 4, rotate = "oblimin", fm = "pa")  
Standardized loadings (pattern matrix) based upon correlation matrix
```

	PA1	PA2	PA4	PA3	h2	u2	com
Price	0.54	0.12	-0.08	-0.05	0.30	0.70	1.2
Safety	-0.33	0.36	0.12	-0.25	0.23	0.77	3.0
Exterior_Looks	0.11	0.07	-0.55	0.27	0.34	0.66	1.6
Space_comfort	-0.04	0.78	-0.13	0.08	0.67	0.33	1.1
Technology	0.02	0.36	0.06	0.02	0.13	0.87	1.1
After_Sales_Service	0.10	0.54	0.18	-0.05	0.33	0.67	1.3
Resale_value	0.73	-0.14	-0.02	-0.16	0.57	0.43	1.2
Fuel_Type	0.01	0.58	-0.01	-0.12	0.32	0.68	1.1
Fuel_Efficiency	0.43	0.23	0.31	0.16	0.47	0.53	2.8
color	0.08	-0.06	0.72	0.14	0.60	0.40	1.1
Maintenance	0.56	0.08	0.21	0.02	0.44	0.56	1.3
Test_drive	0.11	0.16	0.01	0.37	0.20	0.80	1.6
Product_reviews	0.34	0.15	0.05	0.37	0.32	0.68	2.4
Testimonials	-0.20	-0.01	0.06	0.68	0.51	0.49	1.2



# Análise Fatorial

## Script

```
print(quatrofatores$loadings, cutoff = .3)
```

```
Loadings:
          PA1    PA2    PA4    PA3
Price      0.544
Safety    -0.329  0.359
Exterior_Looks      -0.555
Space_comfort      0.782
Technology      0.357
After_Sales_Service 0.537
Resale_value  0.730
Fuel_Type      0.575
Fuel_Efficiency 0.434  0.307
Color          0.720
Maintenance  0.564
Test_drive           0.366
Product_reviews 0.343  0.366
Testimonials           0.682

          PA1    PA2    PA4    PA3
ss loadings 1.642 1.637 1.044 0.969
Proportion Var 0.117 0.117 0.075 0.069
Cumulative Var 0.117 0.234 0.309 0.378
```



# Análise Fatorial

## Script

```
quatrofatores <- fa(EFA, nfactors = 4, rotate =  
"oblimin", fm="pa")  
quatrofatores
```

```
Mean item complexity = 1.6  
Test of the hypothesis that 4 factors are sufficient.  
  
The degrees of freedom for the null model are 91 and the objective function was 2.97 with chi square of 247.71  
The degrees of freedom for the model are 41 and the objective function was 0.57  
  
The root mean square of the residuals (RMSR) is 0.05  
The df corrected root mean square of the residuals is 0.07  
  
The harmonic number of observations is 90 with the empirical chi square 38.46 with prob < 0.58  
The total number of observations was 90 with Likelihood Chi Square = 46.24 with prob < 0.26  
  
Tucker Lewis Index of factoring reliability = 0.922  
RMSEA index = 0.052 and the 90 % confidence intervals are 0 0.085  
BIC = -138.25  
Fit based upon off diagonal values = 0.94  
Measures of factor score adequacy  


|                                                 | PA1  | PA2  | PA4  | PA3  |
|-------------------------------------------------|------|------|------|------|
| Correlation of (regression) scores with factors | 0.87 | 0.88 | 0.83 | 0.80 |
| Multiple R square of scores with factors        | 0.76 | 0.77 | 0.69 | 0.64 |
| Minimum correlation of possible factor scores   | 0.52 | 0.55 | 0.39 | 0.28 |


```

The root mean square of residuals (**RMSR**)

- Deve estar próximo de zero

**RMSEA** (root mean square error of approximation) index

- Deve ser inferior a 0,05

Tucker-Lewis Index (**TLI**)

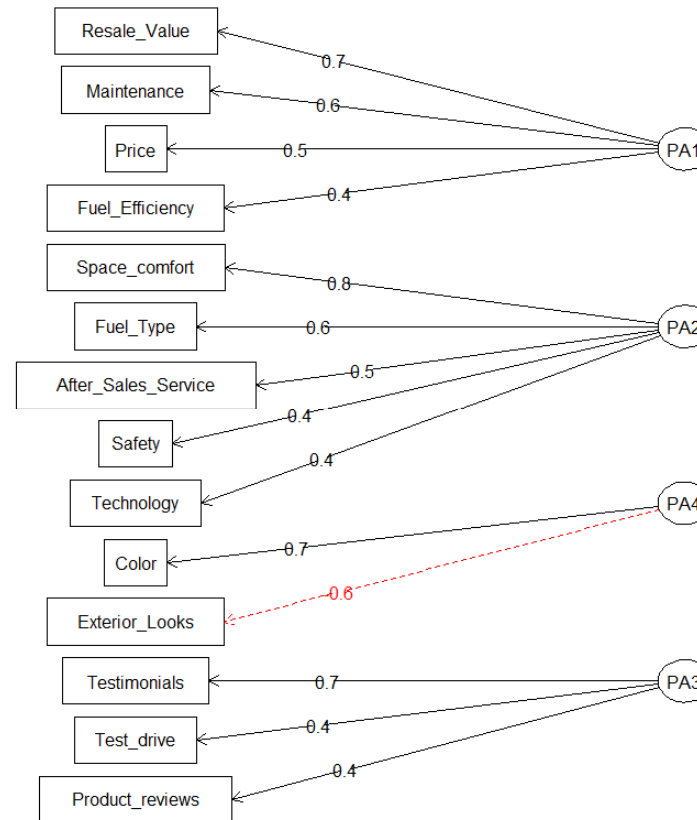
- Valores aceitáveis acima de 0,900

# Análise Fatorial

## Factor Analysis

### Script

```
fa.diagram(quatrofatores)
```



# Como construir uma análise de cluster?





# Análise Preditiva

Gabriel Vinícius Araújo Fonseca





# Análise de Cluster



# Análise de Cluster

- **Objetivo**

- Definir a estrutura de dados colocando as observações mais parecidas em grupos (Cluster de Observações) ou colocando as variáveis mais parecidas em um grupo (Cluster de Variáveis).

- **Variáveis**

- Variáveis métricas ou quantitativas.
- Somente variáveis que se relacionam especificamente (Cluster de Observações).

- **Utilizações**

- Pesquisa de Marketing que queira determinar segmentos de mercado em uma comunidade com base em padrões de lealdade a marcas e lojas.



# Exemplo

- O arquivo Protein.xlsx considera o consumo de proteína em 25 países europeus. Foram considerados no total nove tipos de alimentos que contêm proteína.

Variável	Descrição
RedMeat	Carne Vermelha
WhiteMeat	Carne Branca
Eggs	Ovos
Milk	Leite
Fish	Peixe
Cereals	Cereais
Starch	Amido
Nuts	Nozes
Fr&Veg	Frutas e Verduras



# Leitura dos Dados

## Script

```
library(readxl)
Proteina <- read_excel("D:/Google Drive/PUC Pós/PUC Virtual/Análise Preditiva/Protein.xlsx")
View(Proteina)
```

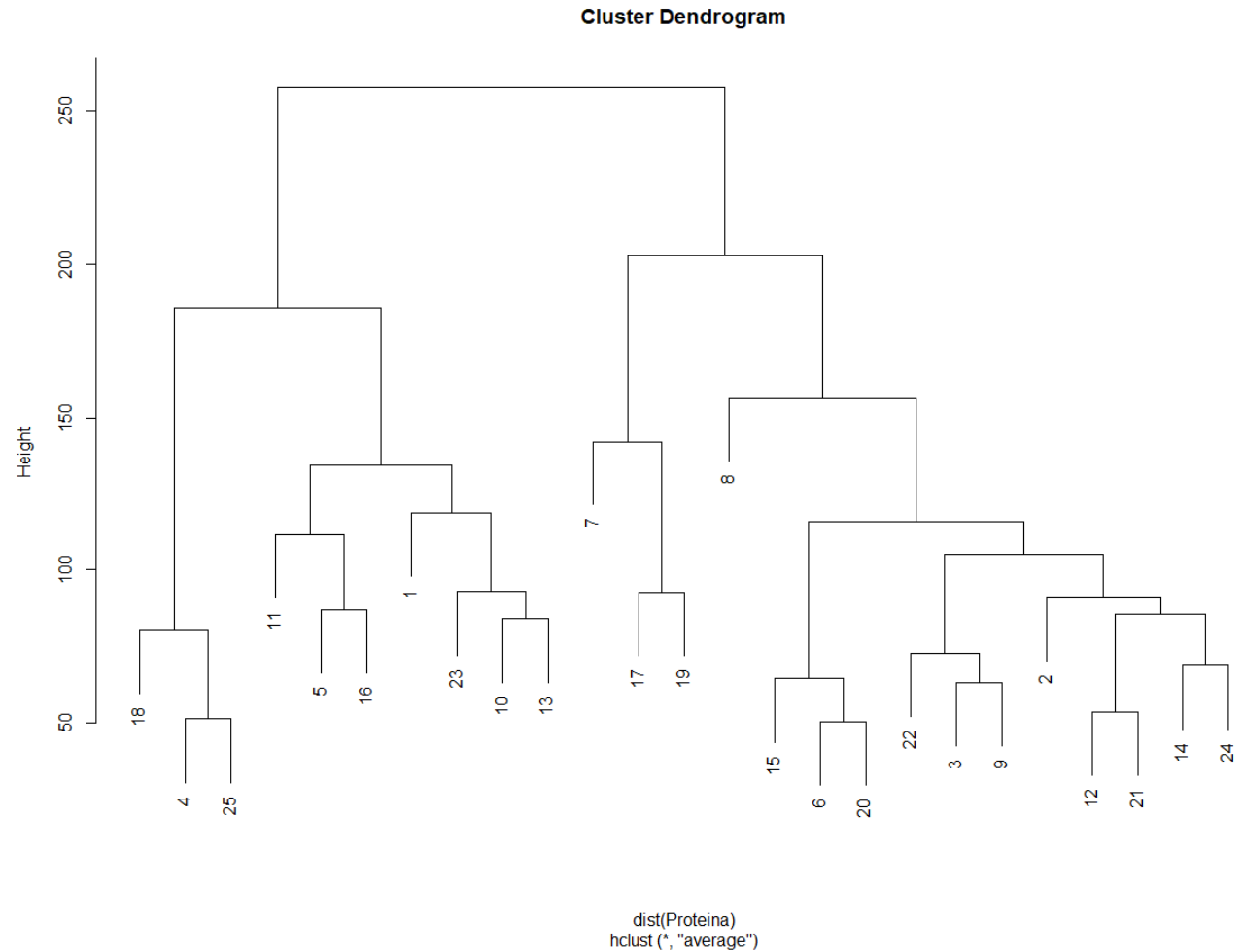
	Country	RedMeat	WhiteMeat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fr&Veg
1	Albania	101	14	5	89	2	423	6	55	17
2	Austria	89	140	43	199	21	280	36	13	43
3	Belgium	135	93	41	175	45	266	57	21	40
4	Bulgaria	78	60	16	83	12	567	11	37	42
5	Czechoslovakia	97	114	28	125	20	343	50	11	40
6	Denmark	106	108	37	250	99	219	48	7	24
7	E Germany	84	116	37	111	54	246	65	8	36



# Escolha do número Clusters

## Script

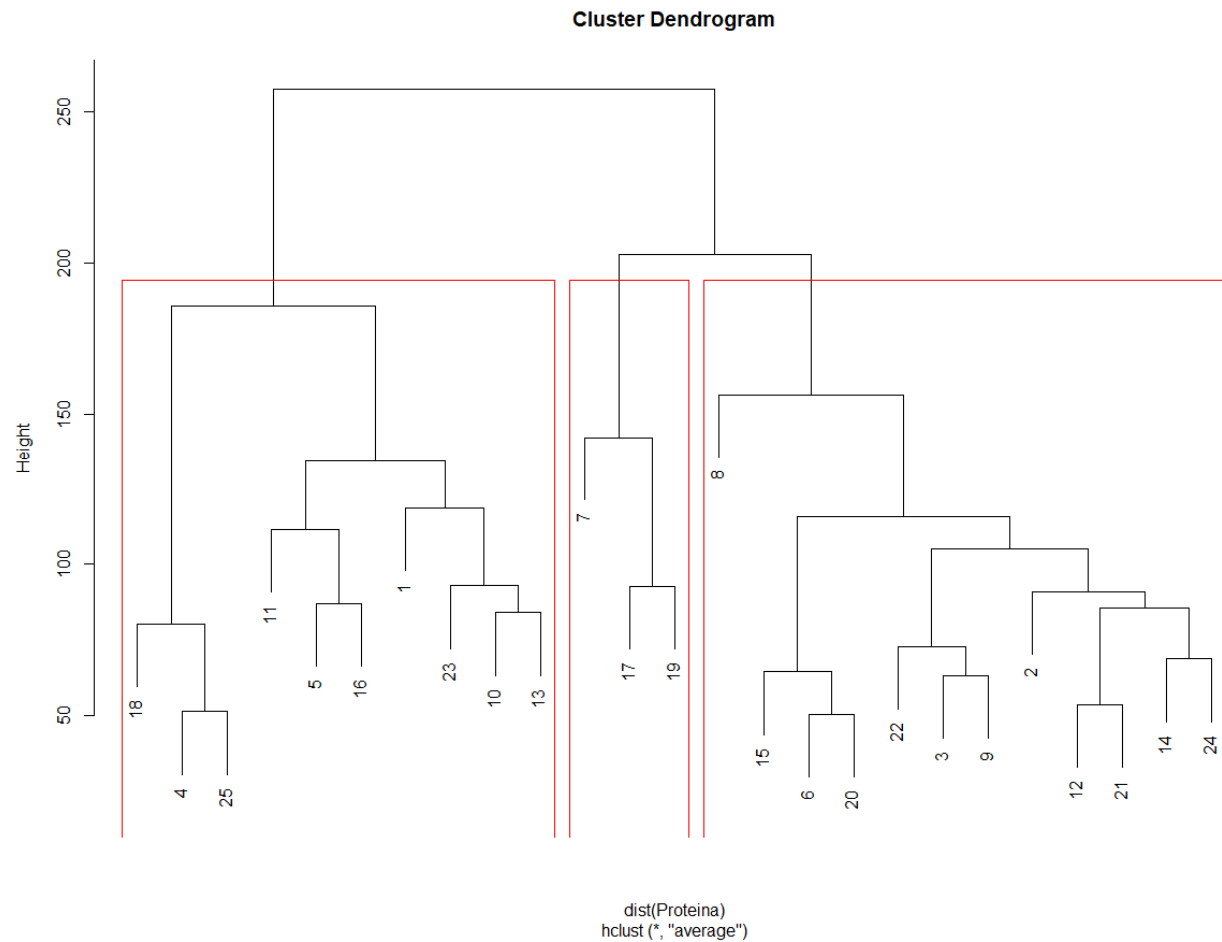
```
install.packages("cluster")  
library(cluster)  
hc <- hclust(dist(Proteina[,-1]), method =  
'average')  
plot(hc)
```



# Construindo os Clusters

## Script

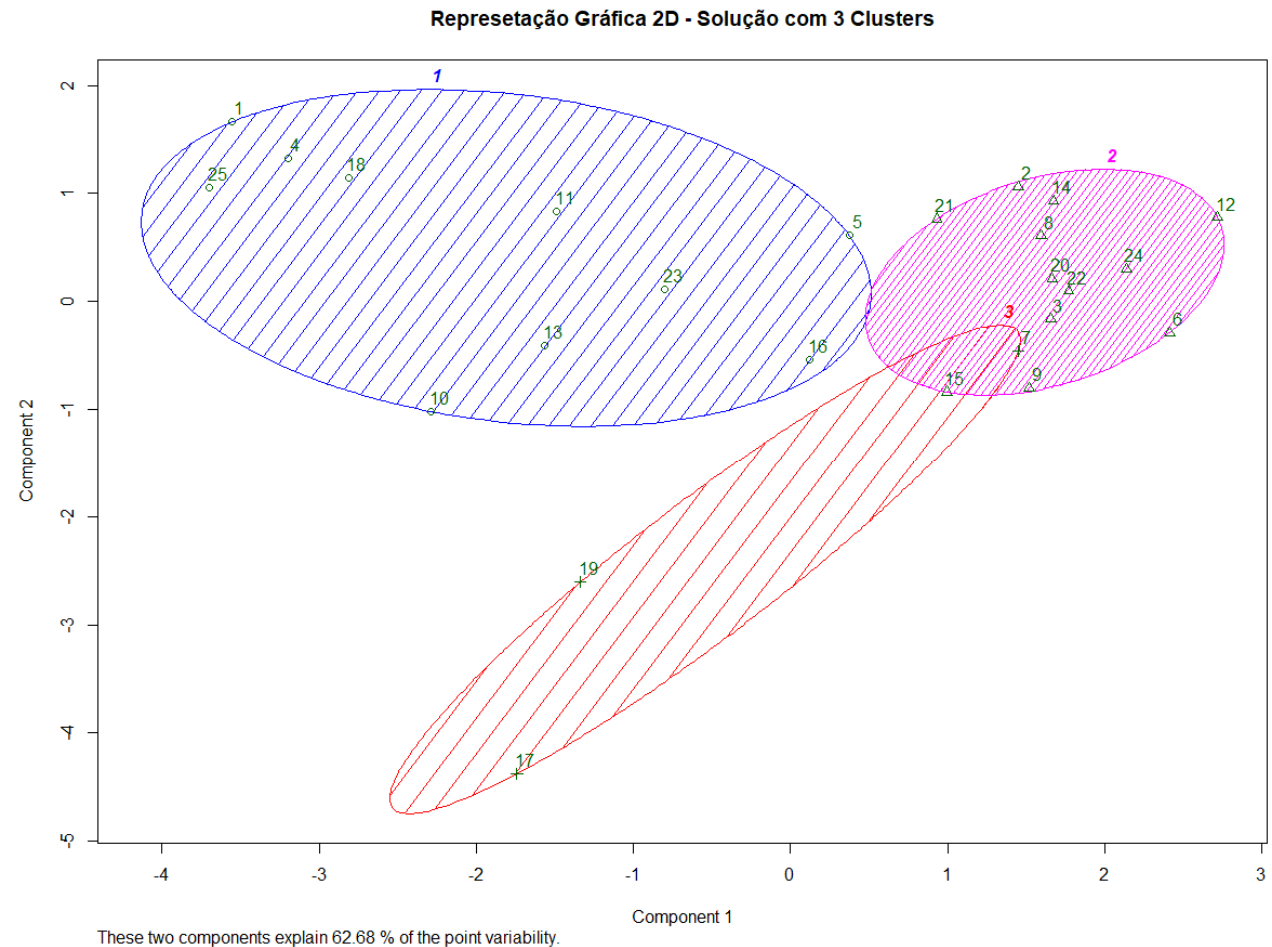
```
clusterCut <- cutree(hc, 3)  
rect.hclust(hc, k=3, border="red")
```



# Visualização dos Cluster e as Componentes

## Script

```
clusplot(Proteina[,-1], clusterCut,  
         main='Represetação Gráfica 2D - Solução  
         com 3 Clusters',  
         color=TRUE, shade=TRUE, labels=2,  
         lines=0)
```



# Análise Final

## Script

```
Proteina$Grupos <- clusterCut  
Grupo_ordenado <-  
Proteina[order(Proteina$Grupos),]  
somente_grupo <- subset(Grupo_ordenado,select =  
c(Country, Grupos))  
View(somente_grupo)
```

	Country	Grupos
1	Albania	1
2	Bulgaria	1
3	Czechoslovakia	1
4	Greece	1
5	Hungary	1
6	Italy	1
7	Poland	1
8	Romania	1
9	USSR	1
10	Yugoslavia	1
11	Austria	2
12	Belgium	2
13	Denmark	2
14	Finland	2
15	France	2
16	Ireland	2
17	Netherlands	2
18	Norway	2
19	Sweden	2
20	Switzerland	2
21	UK	2
22	W Germany	2
23	E Germany	3
24	Portugal	3
25	Spain	3





# Como construir uma análise discriminante?





# Análise Preditiva

Gabriel Vinícius Araújo Fonseca



# Análise Discriminante



# Análise Discriminante

- **Objetivo**

- Estabelecer procedimentos para classificar objetos (indivíduos, firmas , produtos, etc) em grupos, com base em seus escores em um conjunto de variáveis independentes.

- **Variáveis**

- Necessita de uma variável dependente categóricas excludentes (mesmo objeto em duas categorias).
- Variáveis independentes tanto categóricas quanto métricas que consiga diferenciar as categorias proposta na variável dependente.

- **Utilizações**

- Conseguir identificar as melhores características que distinguem um comprador de um não comprador.



# Exemplo

- O arquivo Candidatos.xlsx possui a informação da de 63 candidatos a um curso de pós-graduação. Temos os dados da nota técnica no exame e do histórico escolar, além, é claro, se foi aprovado, ficou na lista de espera ou foi reprovado no concurso.

(Mingoti, Sueli Aparecida. *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. Página 234)

Encontrado no site: <http://www.portalaction.com.br/en/node/2092>



# Leitura dos Dados

## Script

```
library(readxl)
Candidatos <- read_excel("D:/Google Drive/PUC Pós/PUC Virtual/Análise Preditiva/Candidatos.xlsx")
View(Candidatos)
```

	Candidato	Grupo	Nota_tecnica	Historico
1	1	1	19.0	9.0
2	2	1	17.5	8.5
3	3	1	18.2	8.2
4	4	1	17.8	9.2
5	5	1	17.6	9.9
6	6	1	18.2	8.3
7	7	1	19.4	8.2
8	8	1	19.4	8.4
9	9	1	17.3	9.1
10	10	1	18.4	8.8



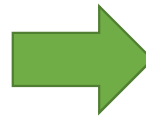
# Tratamento dos Dados

## Script

```
ds_candidatos <- Candidatos[,-1]  
View(ds_candidatos)
```

```
ds_candidatos$Grupo <- factor(ds_candidatos$Grupo, levels = c(1, 2, 3),  
                             labels = c("Aprovado", "Espera", "Reprovado"))  
View(ds_candidatos)
```

	Grupo	Nota_tecnica	Historico
1	1	19.0	9.0
2	1	17.5	8.5
3	1	18.2	8.2
4	1	17.8	9.2
5	1	17.6	9.9
6	1	18.2	8.3



	Grupo	Nota_tecnica	Historico
1	Aprovado	19.0	9.0
2	Aprovado	17.5	8.5
3	Aprovado	18.2	8.2
4	Aprovado	17.8	9.2
5	Aprovado	17.6	9.9
6	Aprovado	18.2	8.3



# Análise Discriminante Linear

## Script

```
attach(ds_candidatos)
require(MASS)
install.packages("klaR")
library(klaR)
ajuste <- lda(Grupe ~ Nota_tecnica + Historico)
ajuste
```

$$LD1 = -0,499 \cdot NotaTecnica - 0,706 \cdot Historico$$

$$LD2 = -0,761 \cdot NotaTecnica + 1,667 \cdot Historico$$

```
> ajuste <- lda(Grupe ~ Nota_tecnica + Historico)
> ajuste
call:
lda(Grupe ~ Nota_tecnica + Historico)

Prior probabilities of groups:
Aprovado      Espera Reprovado
0.3225806 0.2419355 0.4354839

Group means:
              Nota_tecnica Historico
Aprovado      18.14000    8.800000
Espera        16.22000    7.686667
Reprovado     12.31481    6.096296

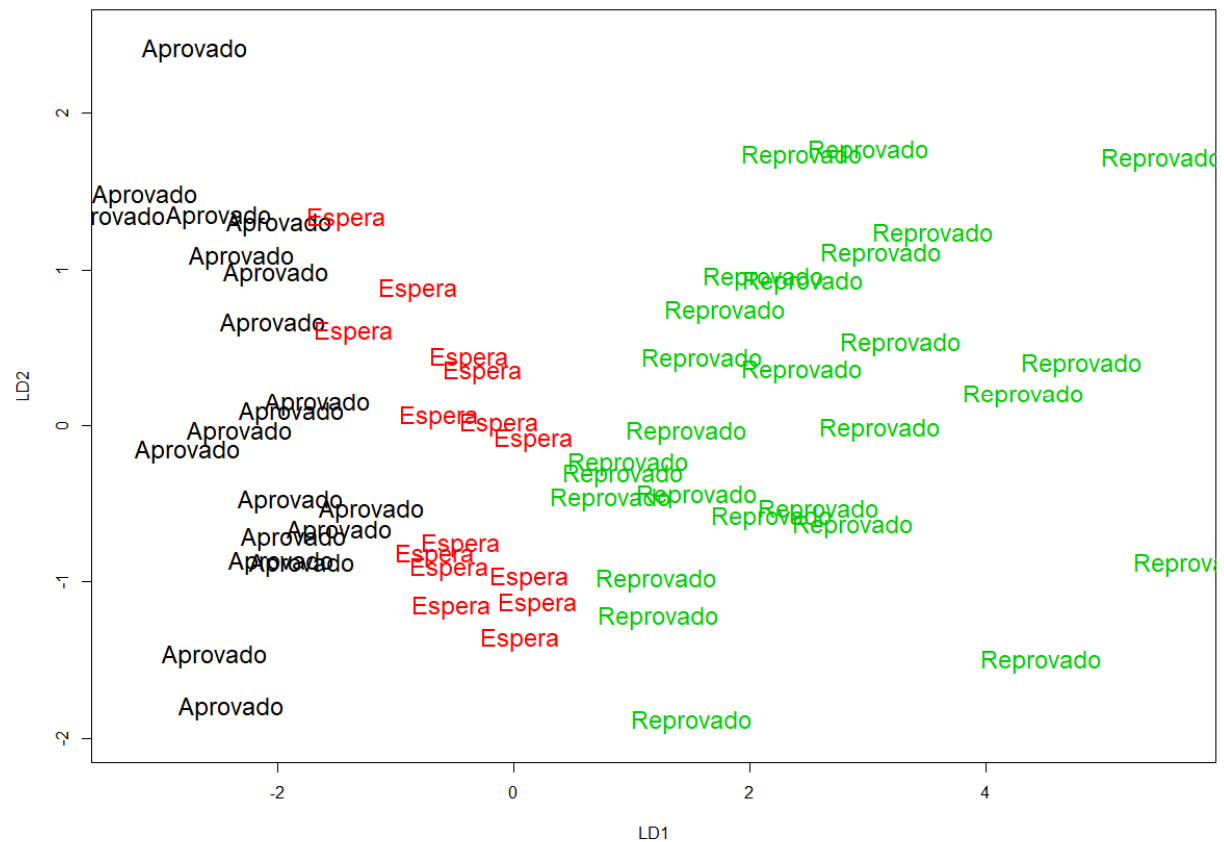
Coefficients of linear discriminants:
              LD1      LD2
Nota_tecnica -0.4989489 -0.7610307
Historico    -0.7060304  1.6670302

Proportion of trace:
  LD1  LD2
0.9947 0.0053
```

# Visualização das Observações em relação a LD1 e LD2

## Script

```
plot(ajuste, col = as.integer(Grupo))
```



# Acurácia do Modelo

## Script

```
table(Grupo, Predito = predict(ajuste, Candidatos[,3:4])$class)
mean(Grupo == predict(ajuste, Candidatos[,3:4])$class)
```

```
> table(Grupo, Predito = predict(ajuste, Candidatos[,3:4])$class)
      Grupo      Predito
      Aprovado Espera Reprovado
Aprovado      19       1        0
Espera         2      13        0
Reprovado      0       2       25
> mean(Grupo == predict(ajuste, Candidatos[,3:4])$class)
[1] 0.9193548
```

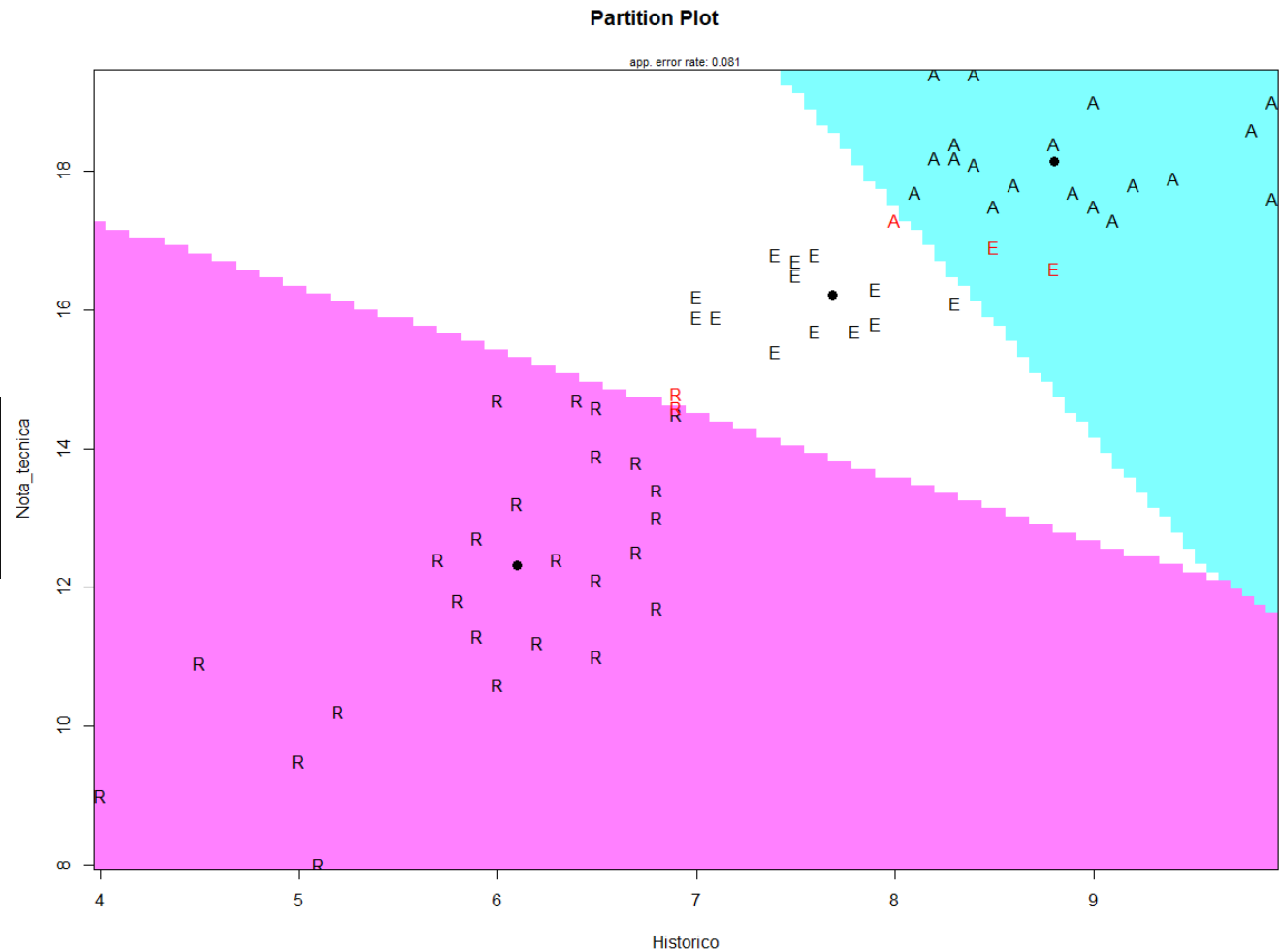


# Visualização da Acurácia do Modelo

## Script

```
partimat(Grupo ~ Nota_tecnica  
+ Historico, method="lda")
```

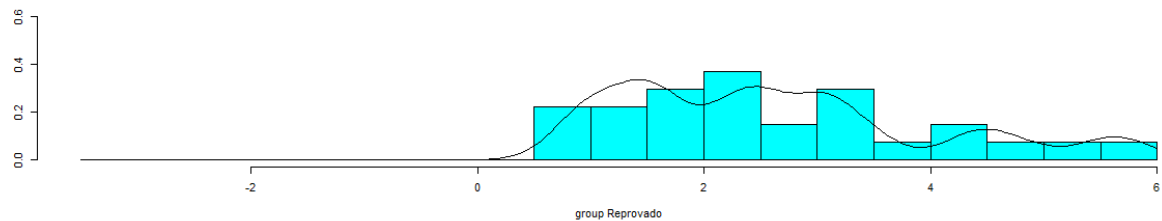
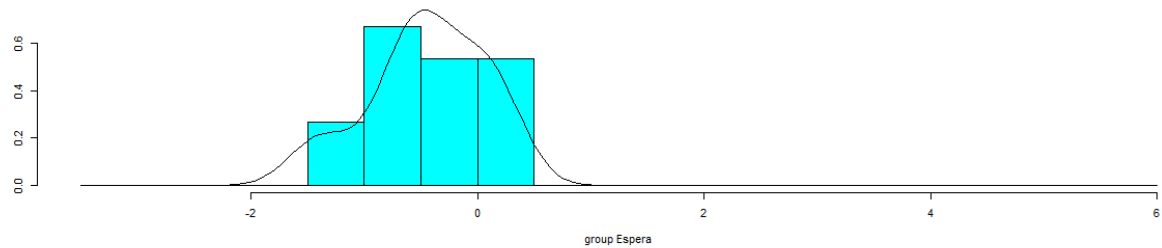
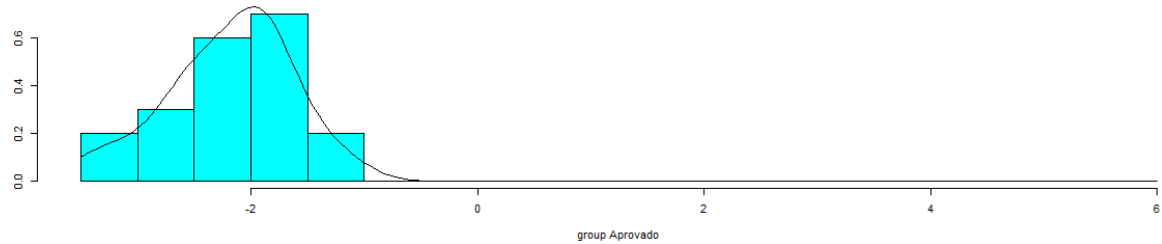
Grupo	Predito		
	Aprovado	Espera	Reprovado
Aprovado	19	1	0
Espera	2	13	0
Reprovado	0	2	25



# Distribuição dos Dados - Normalidade

## Script

```
plot(ajuste, dimen = 1, type = "b")
```



# Análise Discriminante Quadrática

## Script

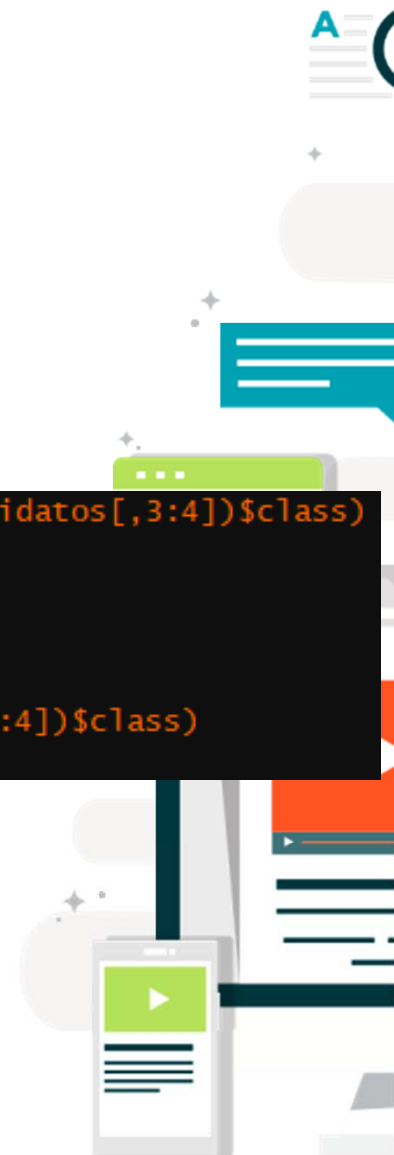
```
ajuste2 <- qda(Grupo ~ Nota_tecnica + Historico)
ajuste2
table(Grupo, Predito = predict(ajuste2, Candidatos[,3:4])$class)
mean(Grupo == predict(ajuste2, Candidatos[,3:4])$class)
```

```
> ajuste2 <- qda(Grupo ~ Nota_tecnica + Historico)
> ajuste2
Call:
qda(Grupo ~ Nota_tecnica + Historico)

Prior probabilities of groups:
  Aprovado   Espera Reprovado 
0.3225806 0.2419355 0.4354839 

Group means:
      Nota_tecnica Historico
Aprovado    18.14000    8.800000
Espera      16.22000    7.686667
Reprovado   12.31481    6.096296
```

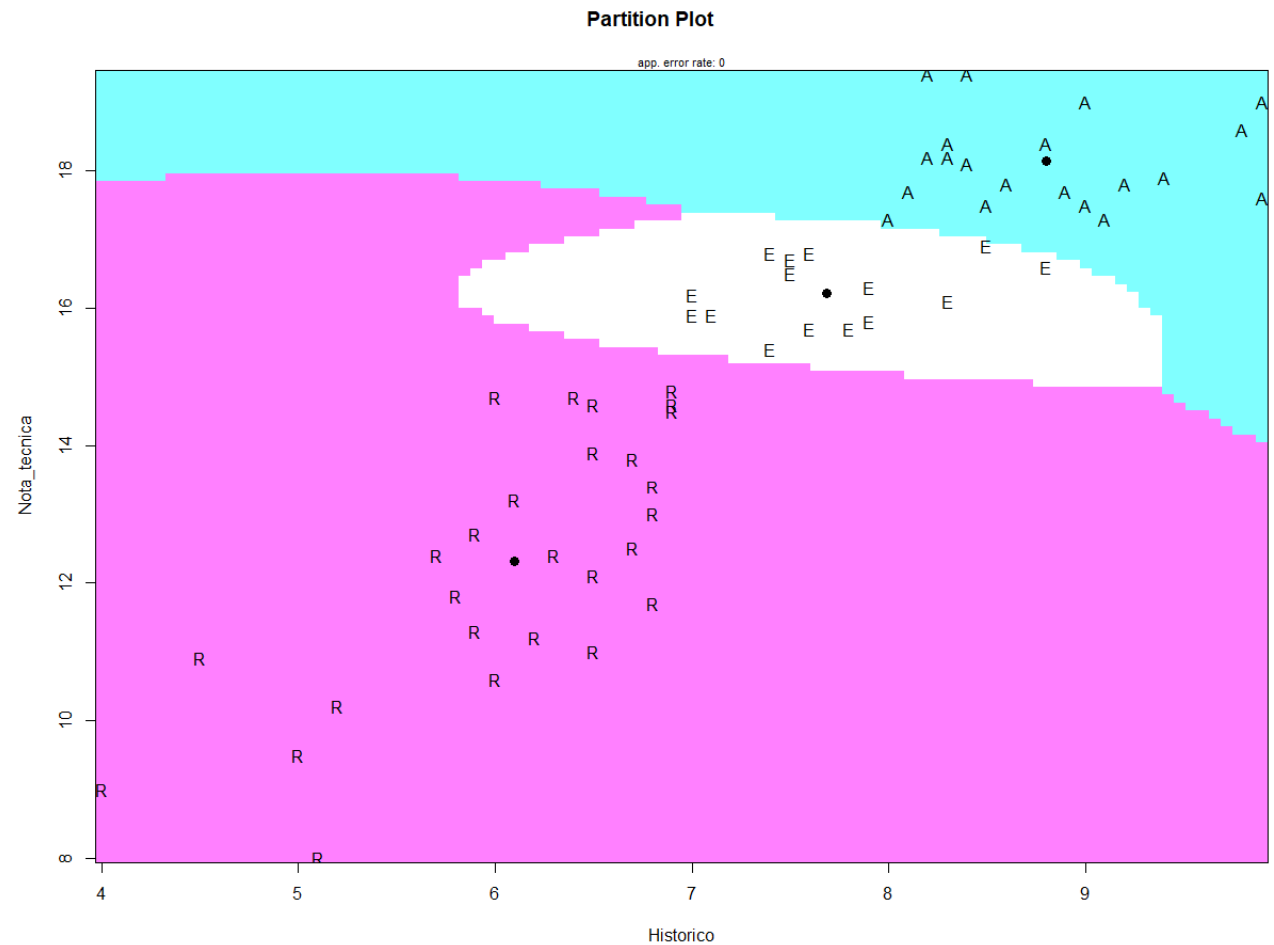
```
> table(Grupo, Predito = predict(ajuste2, Candidatos[,3:4])$class)
      Predito
Grupo   Aprovado  Espera  Reprovado
Aprovado      20       0         0
Espera         0      15         0
Reprovado      0       0        27
> mean(Grupo == predict(ajuste2, Candidatos[,3:4])$class)
[1] 1
```



# Visualização da Acurácia do Modelo

## Script

```
partimat(Grupo ~ Nota_tecnica +  
Historico, method="qda")
```



# Resumo

- **Linear Discriminant Analysis (LDA)**
  - **Análise Discriminante Linear:** Usado quando **se assume** que a covariância das variáveis independentes são iguais entre todos os grupos.
- **Quadratic Discriminant Analysis (QDA)**
  - **Análise Discriminante Quadrática:** Usado quando **não se assume** que a covariância das variáveis independentes são iguais entre todos os grupos.





**E o que tem mais para  
análise preditiva?**





# Análise Preditiva

Gabriel Vinícius Araújo Fonseca



# Análise de Séries Temporais



# O que é Análise de Séries Temporais?

- Série temporal é qualquer conjunto de observações ordenadas no tempo.
- As observações vizinhas têm dependência e estamos interessados em analisar e modelar esta dependência.
- Tem que haver uma dependência dos dados no tempo.



# Aplicações em Séries Temporais

- **Economia:** preços diários de ações; taxa de desemprego.
- **Medicina:** níveis de eletrocardiograma ou eletroencefalograma.
- **Epidemiologia:** casos semanais de uma doença; casos mensais de AIDS.
- **Meteorologia:** temperatura diária; registro de marés.
- **Mercado:** Predição de consumo.



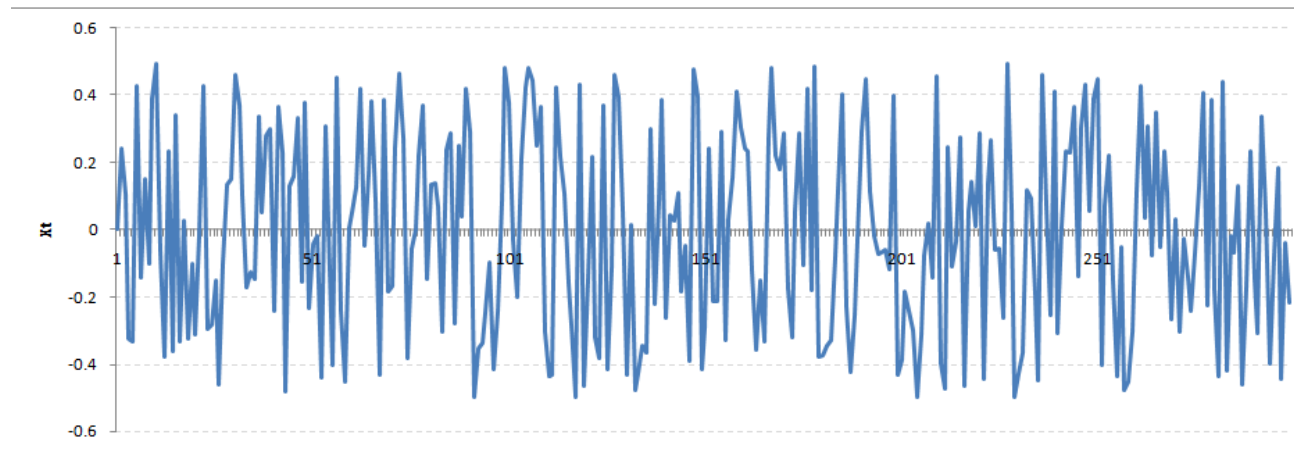
# Objetivos

1. Compreender o mecanismo gerador da série
2. Predizer o comportamento futuro da série
3. Descrever o comportamento da série
4. Procurar periodicidades relevantes nos dados



# Estacionaridade

- **Significado:** a série se desenvolve no tempo aleatoriamente ao redor de uma média constante, refletindo alguma forma de equilíbrio estável.

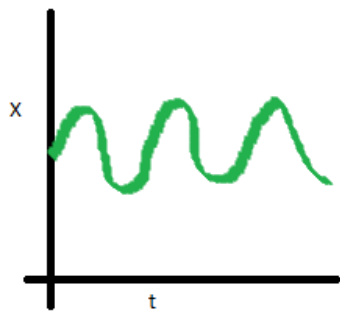


<https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/>

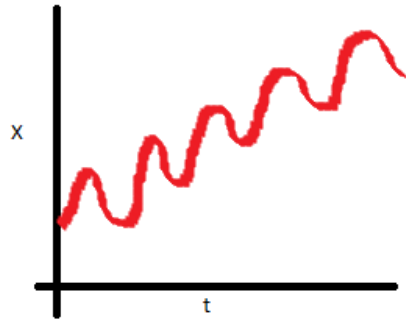




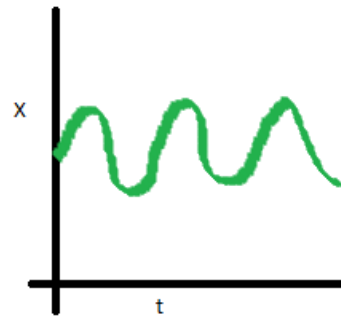
# Exemplos de Não Estacionaridade



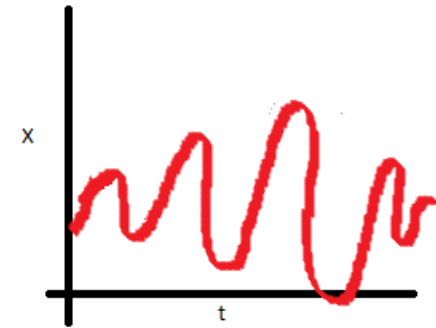
Stationary series



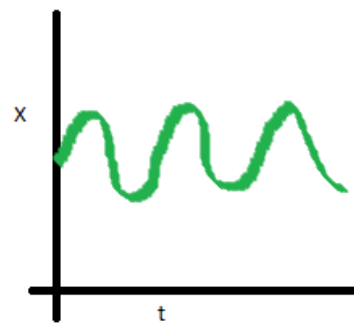
Non-Stationary series



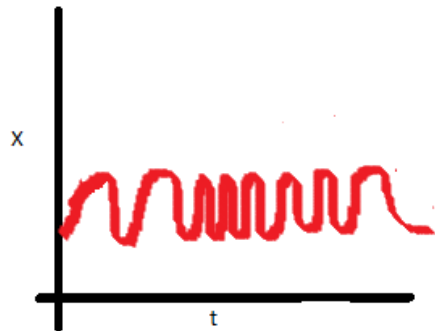
Stationary series



Non-Stationary series



Stationary series

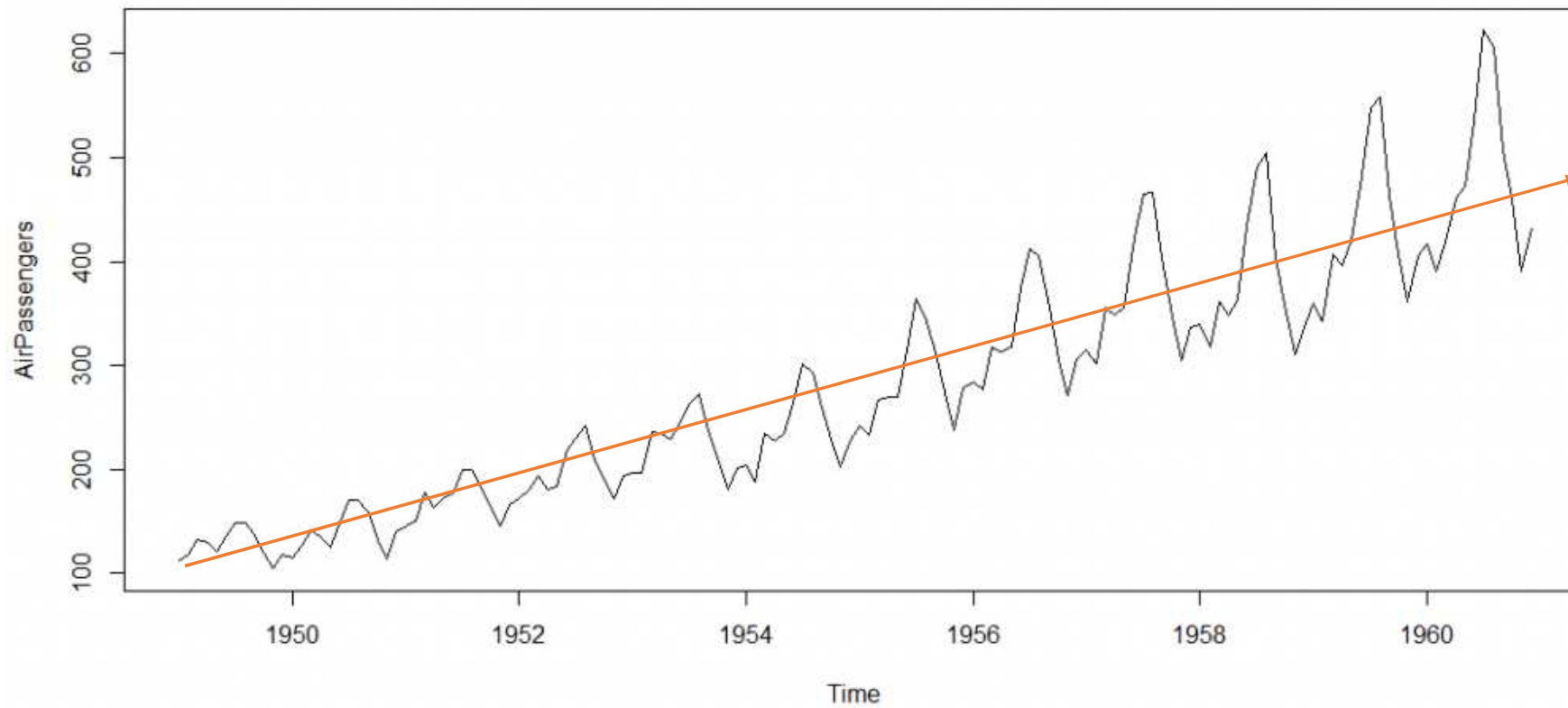


Non-Stationary series

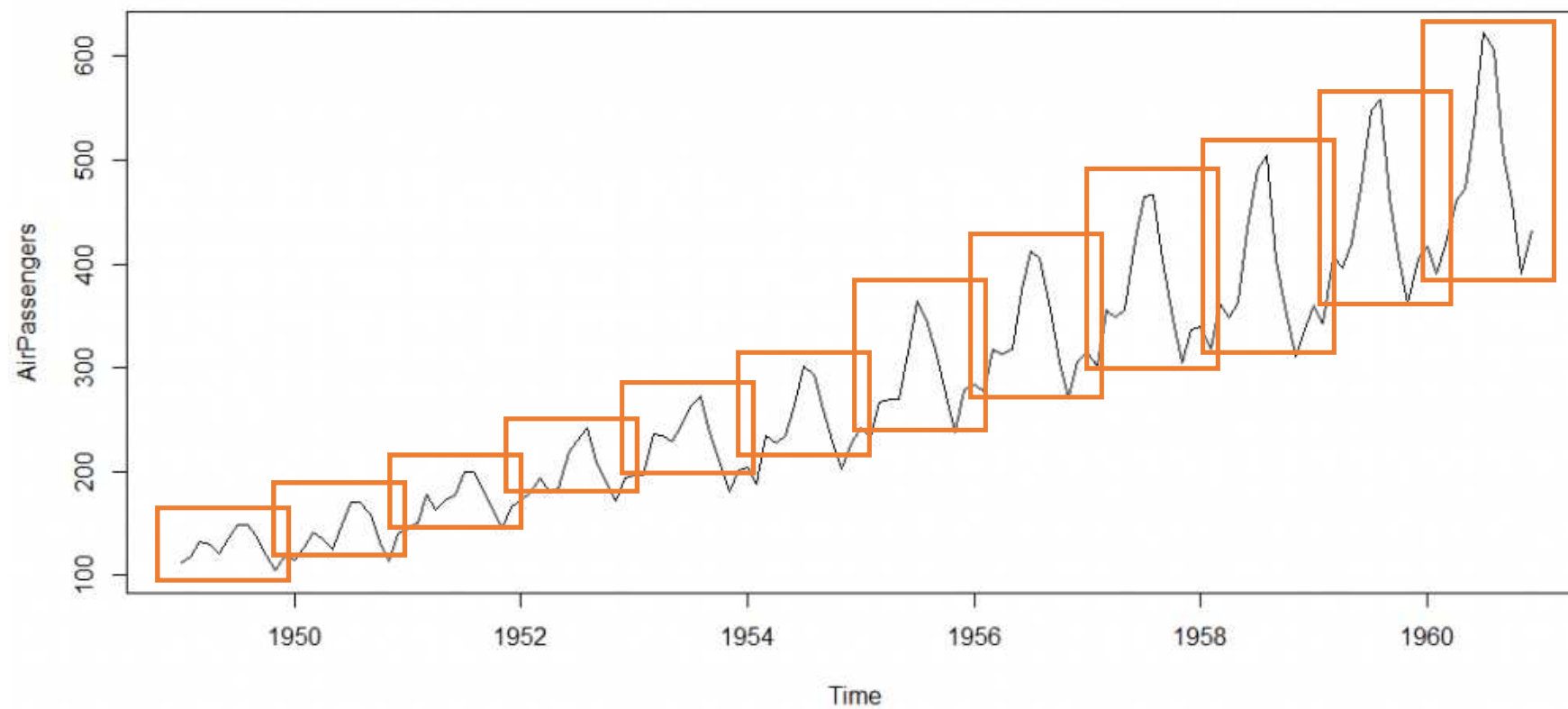
<https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/>



# Tendência



# Sazonalidade



# Modelos ARIMA

- **Nome:** Modelo Autoregressivo (AR) Integrado (I) de Médias Móveis (MA).
- Notação Estatística: ARIMA ( $p$ ,  $d$ ,  $q$ ).
- Para a construção do modelo, os parâmetros são estimados pelos dados amostrais.
- $p$  indica o nível da Autoregressão.
- $d$  indica o nível da Integração (diferenciação).
- $q$  indica o nível de Médias Móveis.

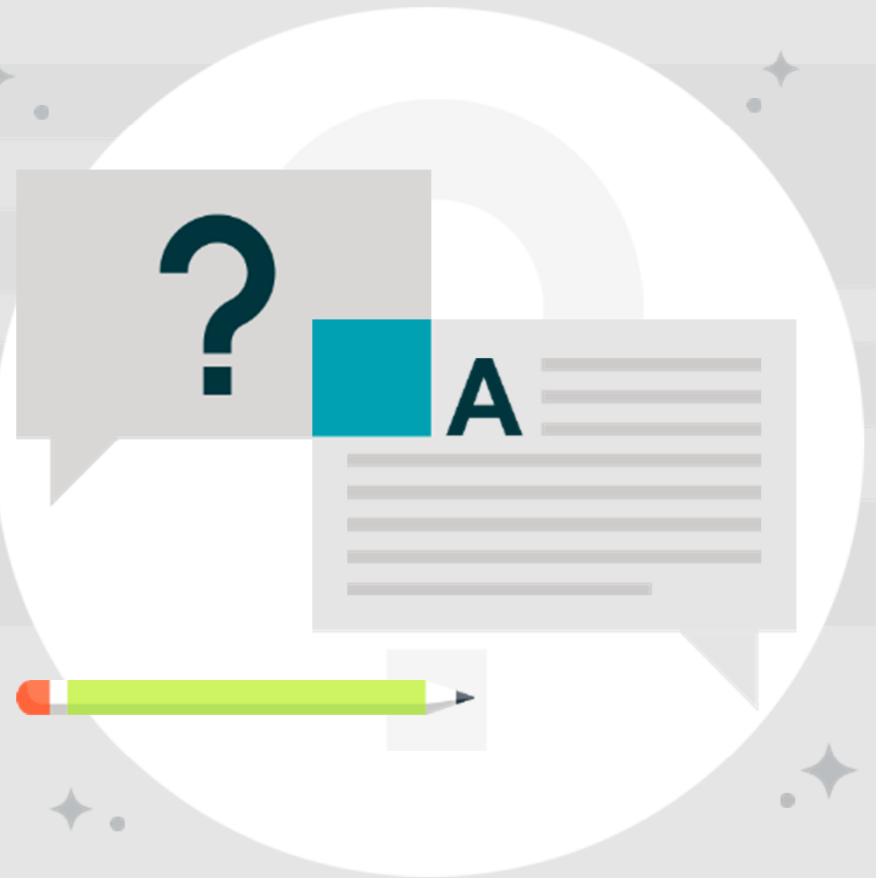


# Etapas da Modelagem ARIMA

- Faça um gráfico para identificação da série – **Tendência, Ciclo e Sazonalidade**.
- Tornar a série estacionária – Por diferenciação.
- Determinar se os termos **autorregressivos (AR)** ou de **médias móveis (MA)** são necessários para corrigir qualquer autocorrelação que permaneça na série diferenciada.
- Testar o modelo que funciona melhor – **Menor AIC**.
- Analisar os resíduos - Deve apresentar os resíduos estacionários, com média zero e variância constante = **Modelo correto**.



# Como criar modelos de Análise de Séries Temporais no Rstudio?





# Análise Preditiva

Gabriel Vinícius Araújo Fonseca





# Exemplo de Análise de Séries Temporais



# Etapas da Modelagem ARIMA

- Faça um gráfico para identificação da série – **Tendência, Ciclo e Sazonalidade**.
- Tornar a série estacionária – Por diferenciação.
- Determinar se os termos **autorregressivos (AR)** ou de **médias móveis (MA)** são necessários para corrigir qualquer autocorrelação que permaneça na série diferenciada.
- Testar o modelo que funciona melhor – **Menor AIC**.
- Analisar os resíduos - Deve apresentar os resíduos estacionários, com média zero e variância constante = **Modelo correto**.



# Exemplo

- No arquivo AirPassengers consiste no número mensal de passageiros internacionais entres os anos de 1949 a 1960, um total de 12 anos.
- Esse conjunto de dados está contido dentro do próprio R. A estrutura dos dados já está no formato de uma série temporal (ts).
- Assim algumas informações estão ocultas como as datas, mas elas estão contida no dataset.



# Exploração dos Dados

		Filter
1	112	
2	118	
3	132	
4	129	
5	121	
6	135	
7	148	
8	148	
9	136	
10	119	

## Script

```
data("AirPassengers")  
View(AirPassengers)
```

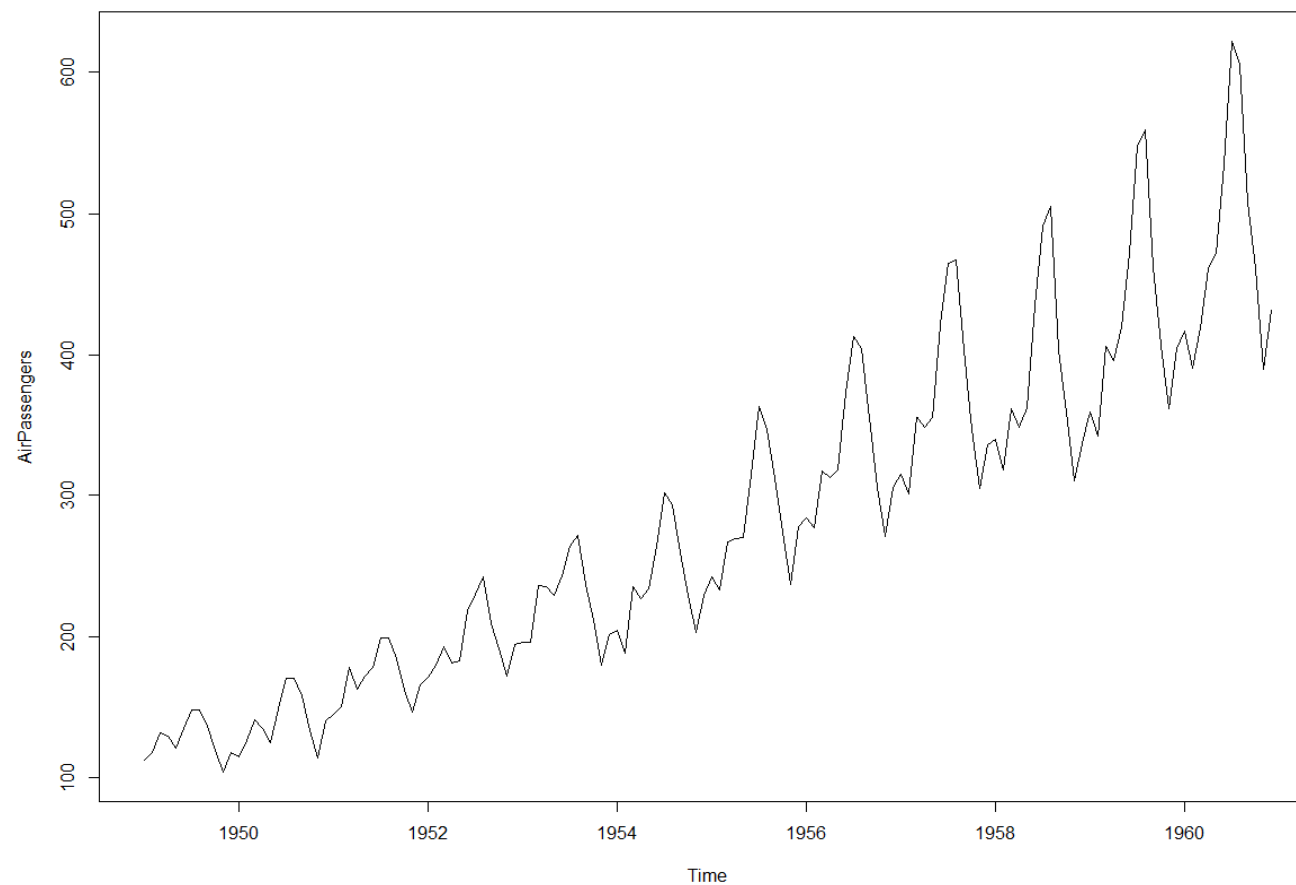
Environment	History	Connections
Import Dataset		
Global Environment		
Values		
AirPassengers	Time-Series [1:144] from 1949 to 1961: 112 118 132 129 121 135 148 148 136 1...	



# Plot da Série

## Script

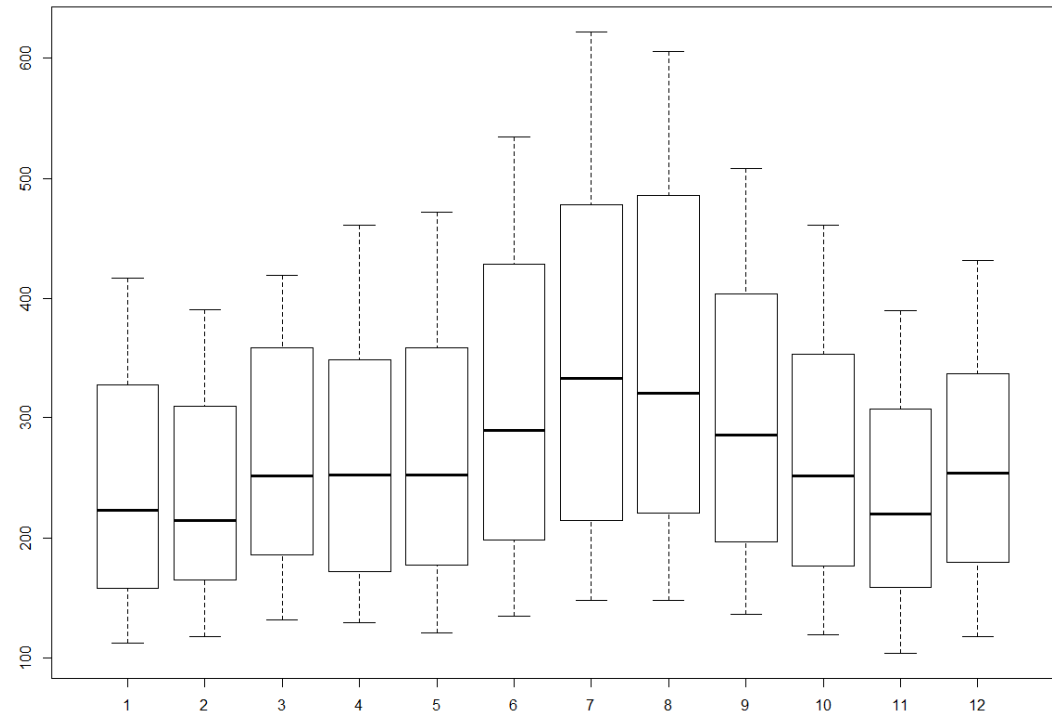
```
plot(AirPassengers)
```



# Box-Plot da Série

## Script

```
boxplot(AirPassengers~cycle(AirPassengers))
```



# Observações sobre essa série

- Tendência

- A cada ano, o número médio de passageiros aumenta.

- Sazonalidade

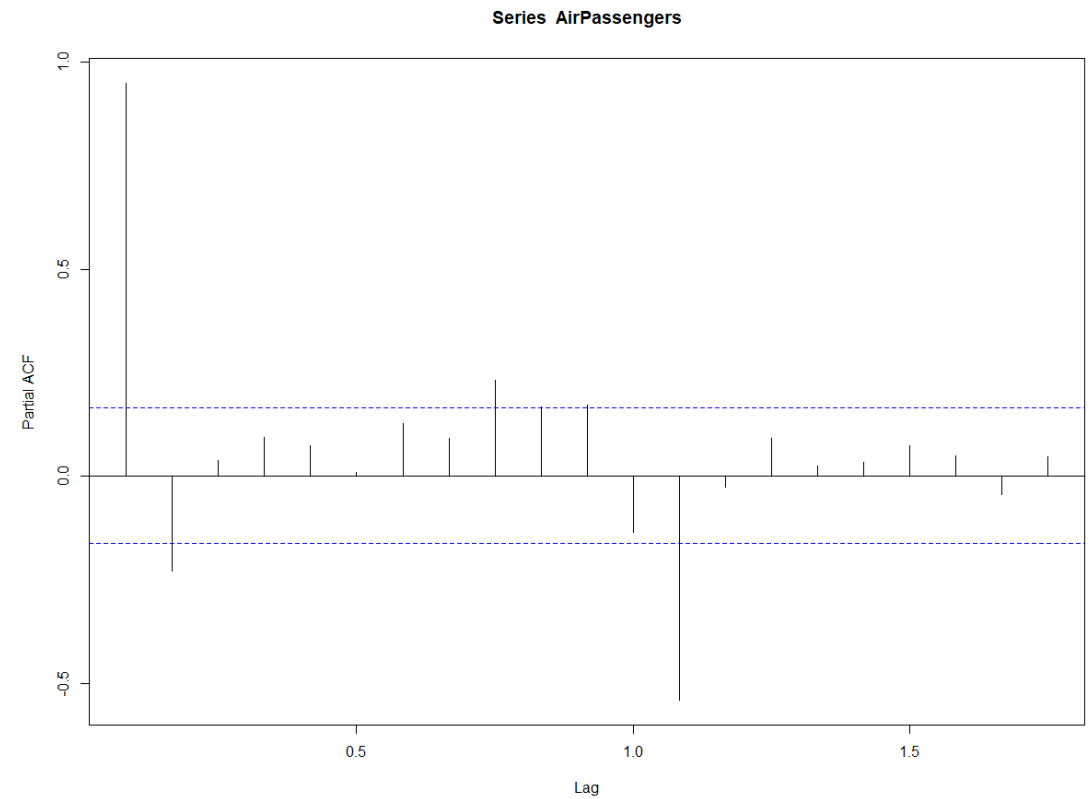
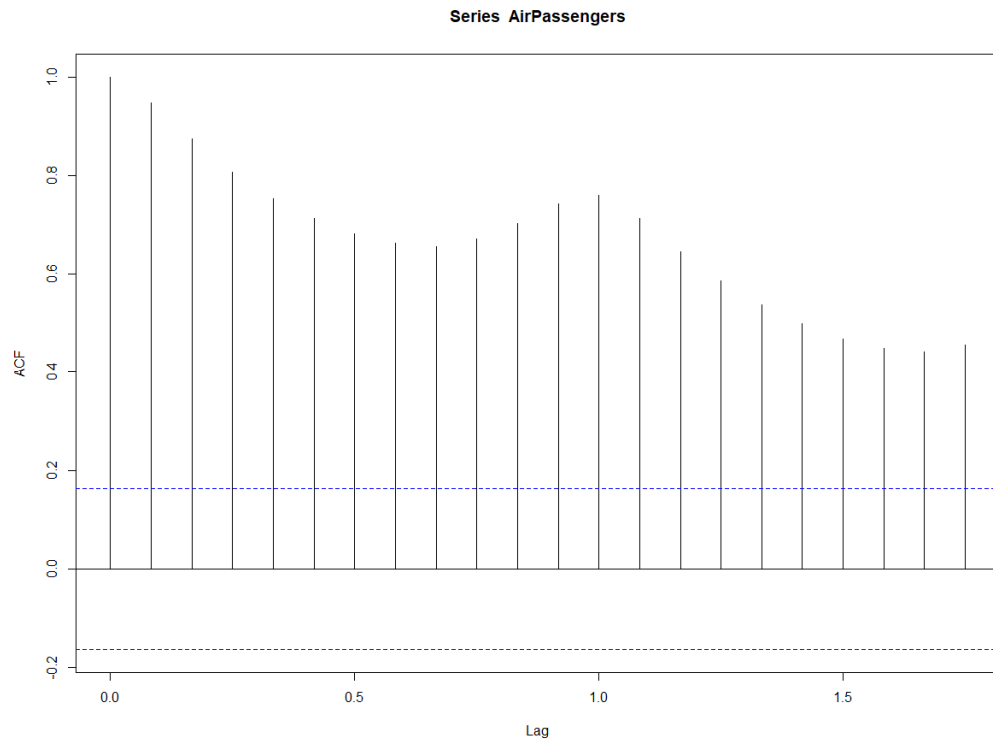
- O comportamento da série é bem parecida a cada 12 meses.

- Variação

- Vejam que a variação da série em cada ciclo de 12 meses aumenta ao longo da série.



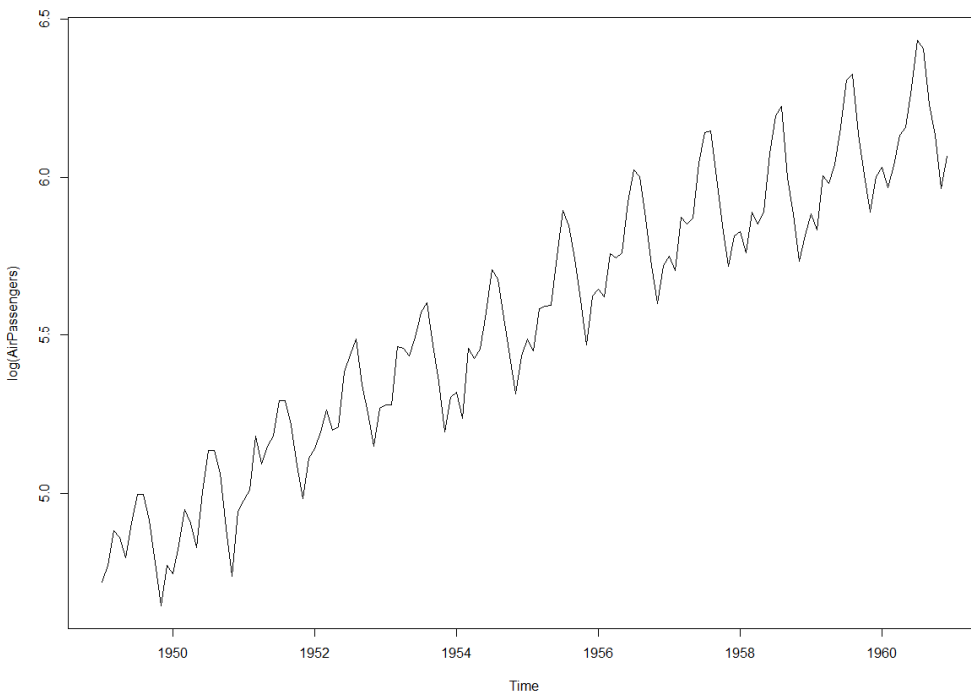
# Autocorrelações



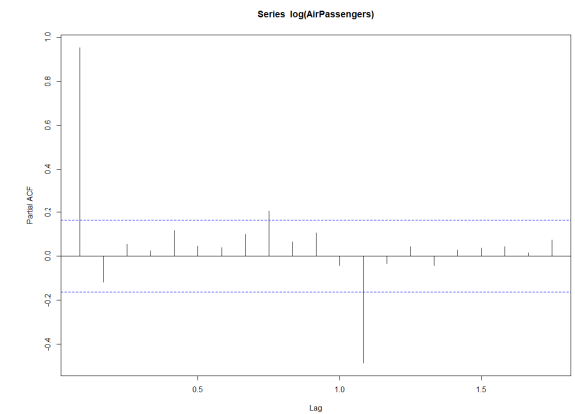
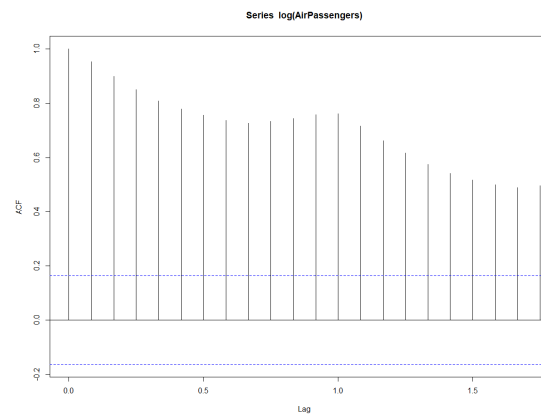


# Transformação da Série

```
plot(log(AirPassengers))
```

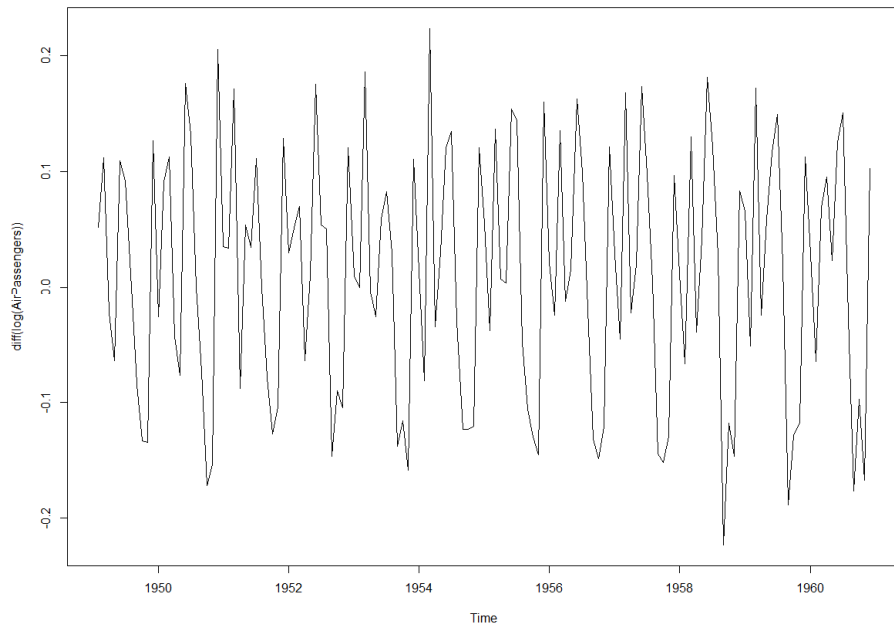


```
acf(log(AirPassengers))  
pacf(log(AirPassengers))
```

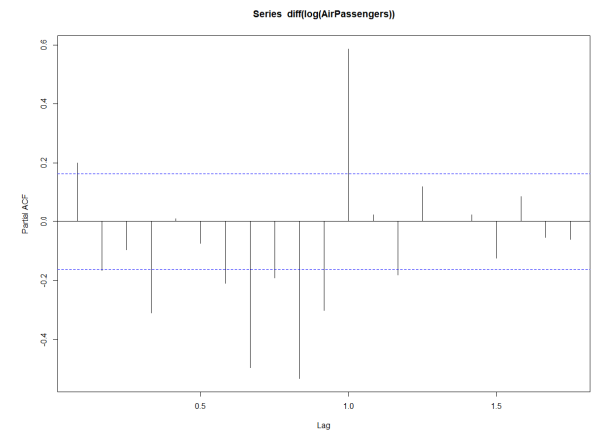
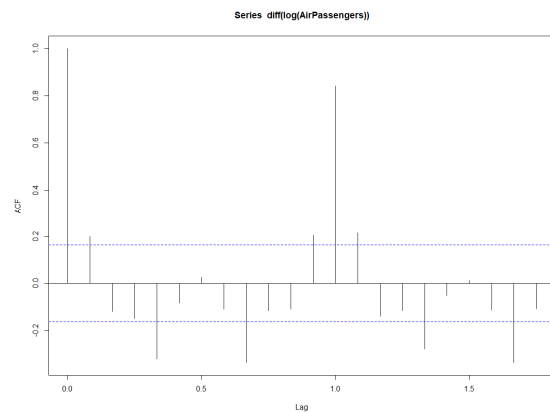


# Estacionaridade da Série

```
plot(diff(log(AirPassengers)))
```



```
acf(diff(log(AirPassengers)))  
pacf(diff(log(AirPassengers)))
```



```
install.packages("tseries")  
library(tseries)  
adf.test(diff(log(AirPassengers)), alternative="stationary", k=0)
```

```
> adf.test(diff(log(AirPassengers)), alternative="stationary", k=0)
```

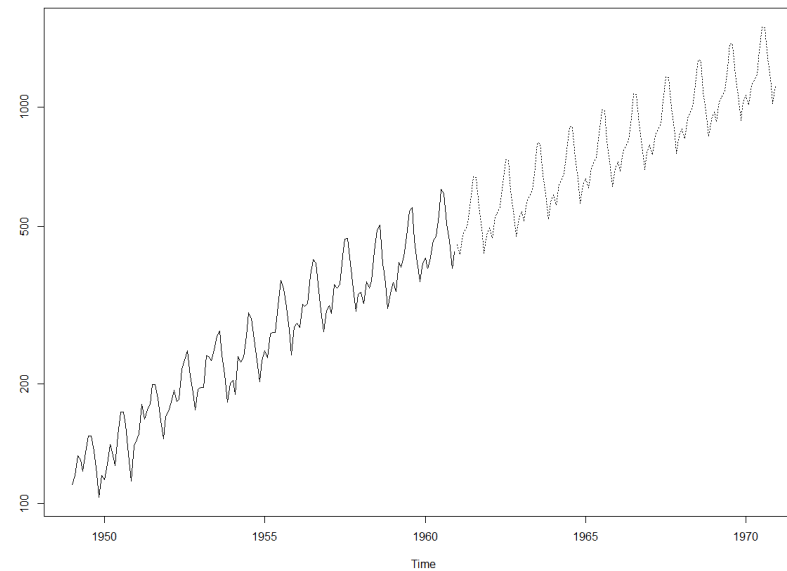
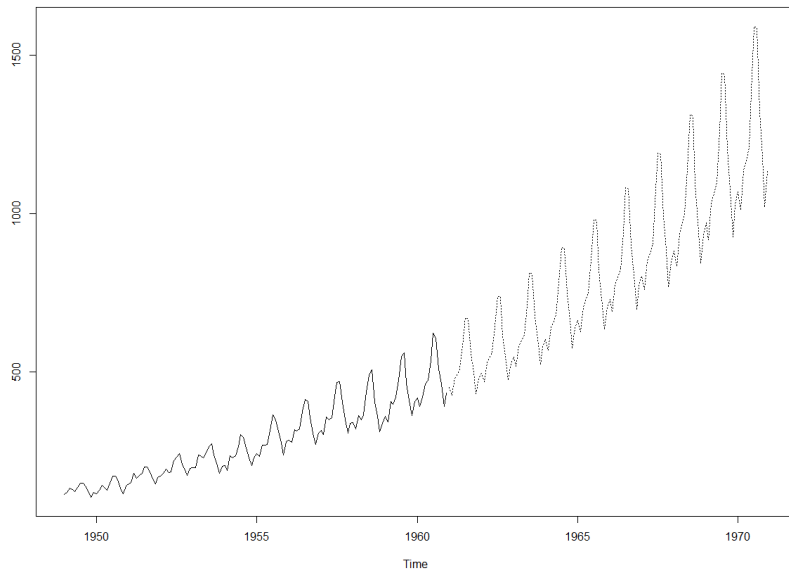
Augmented Dickey-Fuller Test

```
data: diff(log(AirPassengers))  
Dickey-Fuller = -9.6003, Lag order = 0, p-value = 0.01  
alternative hypothesis: stationary
```

# Estimando Modelo ARIMA

## Script

```
ajuste <- arima(log(AirPassengers), c(0, 1, 1), seasonal = list(order = c(0, 1, 1), period = 12))  
pred <- predict(ajuste, n.ahead = 10*12)  
ts.plot(AirPassengers, exp(pred$pred), lty = c(1,3))  
ts.plot(AirPassengers, exp(pred$pred), log = "y", lty = c(1,3))
```



**E existe mais opções  
para Análise de  
Predição?  
Claro!!!!!!**



