
ANÁLISE DE
DESEMPENHO
NO FUTEBOL
Analytics
Eduardo Cecconi


Conceitos Básicos

Business Intelligence

- processo de coleta, organização, análise, compartilhamento e monitoramento de informações que oferecem **suporte à gestão** de negócios

- Inclui os aplicativos, a infraestrutura, as ferramentas e as melhores práticas

- Análises:

    - **Descritiva** (o que aconteceu; quando aconteceu)
    - **Diagnóstica** (Por que aconteceu)

- Abordagens:

    - Indicadores
    - Dashboards
    - Relatórios
    - ScoreCards

Business Analytics

- processo de descoberta, interpretação e comunicação de padrões em dados através da **análise computacional** - o que é **potencializado através de Big Data**

- Análises:

    - **Preditiva** (o que vai acontecer)
    - **Prescritiva** (o que fazer)

- Abordagens:

    - Data Mining
    - Machine Learning
    - Variados algoritmos de apoio à decisão

Big Data

- grandes bancos de dados analisados para revelar padrões, tendências e associações
- **Advanced Analytics**: algoritmos matemáticos e estatísticos avançados
- **Data Science**: conjunto de técnicas de programação e estatística/matemática voltadas a coletar, tratar, manipular, organizar, analisar, extrair e apresentar de dados, para subsidiar a tomada de decisão
- O fundamento da Ciência de Dados é a habilidade de **armazenar e processar grandes quantidades de dados** e extrair valores segundo uma instrução

Algoritmo

- Sequência explícita, literal, limitada e sistêmica de instruções e operações direcionadas à consecução de um objetivo

- **Regras para se executar uma ação** (resolver um problema)

- **Algoritmo para beber água**:
  pegar um copo -> pegar uma garrafa d'água -> servir a água no copo -> beber a água

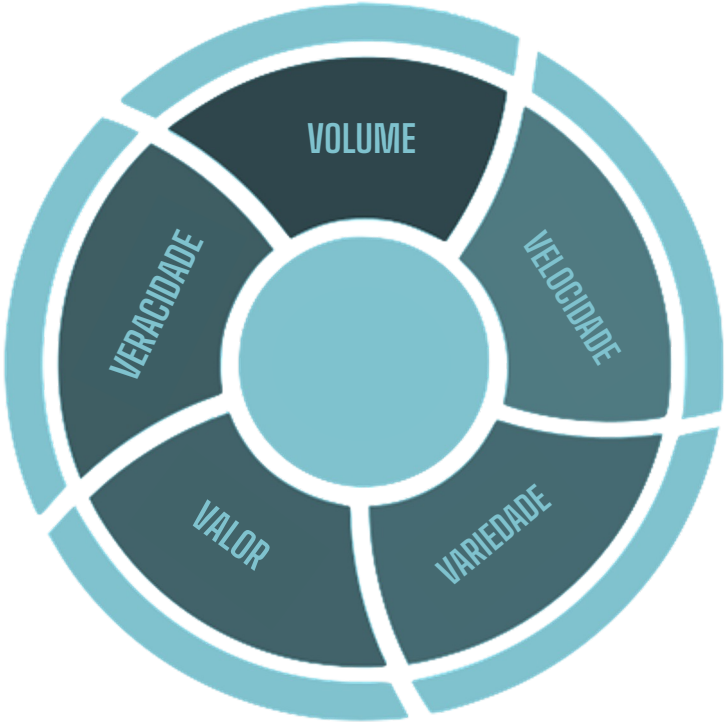
- **Algoritmo para beber água**:
  pegar um copo vazio -> colocar o copo, com a abertura para cima, sobre a mesa -> pegar uma garrafa d'água -> abrir a tampa da garrafa -> inclinar a garrafa direcionando a abertura ao copo -> derramar o volume d'água desejado no copo -> recolher a garrafa à posição -> fechar a garrafa -> colocar a garrafa sobre a mesa -> pegar o copo -> levar o copo até a boca -> abrir a boca -> inclinar o copo em direção à boca -> derramar a água...


Big Data

Os 5 V's do Big Data

- **Volume**: capacidade de armazenamento evolui rapidamente. 1 Petabyte = 1 milhão de Gigabytes
- **Velocidade**: taxa de transferência e latência (tempo de transferência)
- **Variedade**: diversidade das fontes de dados
- **Veracidade**: autenticidade dos dados; precisão
- **Valor**: extrair valor dos dados

Requer alta capacidade computacional e de armazenamento



CLOUD COMPUTING

- **Nuvem**: conjunto de recursos tecnológicos disponibilizados através da internet
- **Serviços de Nuvem**: diversos provedores oferecem serviços na nuvem auxiliando em toda a arquitetura do processo, desde a estratégia até a implantação, conforme as necessidades do cliente e o modelo de serviço escolhido

- **Camadas da Nuvem**: Armazenamento, Capacidade Computacional, Segurança e Ambientes de Rede

CLOUD COMPUTING

- **Vantagens da Nuvem**:
  Agilidade/Flexibilidade
  Redução de custos operacionais
  Maximização dos investimentos em T.I
  Controle da segurança
  Administração, análise e compartilhamento dos dados
  Inovação - acesso a tecnologia e serviços inovadores

- **Barreiras de Implantação**:
  Custos de migração
  Custos dos serviços na nuvem
  Desafios de integração com a tecnologia substituída
  Requer profissionais altamente capacitados para funções específicas (arquitetura de nuvem, engenharia de dados, ciência de dados)

E.T.L

- **EXTRACT, TRANSFORM & LOAD**
- Processo utilizado para coletar e combinar dados de várias fontes, transformá-los conforme regras do negócio e carregá-los em um banco de dados
- **Extrair**: obtenção dos dados estruturados brutos, de diversas fontes e formatos
- **Transformar**: organização e limpeza dos dados
- **Carregar**: depósito para os dados tratados, permitindo a modelagem do banco de dados

PROCESSO DE E.T.L NA NUVEM

FONTES DE DADOS

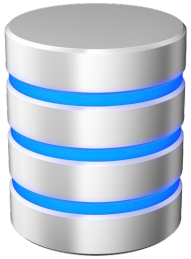
EXTRAIR

TRANSFORMAR

CARREGAR

STAGING AREA

Organização e limpeza dos dados em um armazenamento provisório



DATA WAREHOUSE

Modelagem do Banco de Dados
Integrado (coerente)
Não-volátil (dados consolidados)
Variável em relação ao tempo
Orientado ao assunto (Data Mart's)

ANÁLISE


Smart Data

Smart Data

- É o processamento do Big Data
- Filtragem/Seleção de um conjunto reduzido e relevante de dados para análise
- Reduz a demanda por armazenamento e capacidade computacional
- Cenário mais comum nos clubes de futebol brasileiro:
  Assinatura de provedores que utilizam Big Data
  Seleção customizada dos dados para cada demanda de análise

PROCESSO DE E.T.L CLIENT-SIDE

StatsBomb
hudlsportscode
STATS PERFORM
CATAPULT
InStat
METRICA SPORTS
wyscout
LONGOMATCH

FONTES DE DADOS

EXTRAIR

API's
Download
Transferência
Bancos de Dados
Bibliotecas

TRANSFORMAR

Power BI
python
R

Tidy Data
Data Wrangling

CARREGAR

Data Center
HD Externo
Pastas locais
Armazenamento na nuvem (limitado):
Google Drive, DropBox

ANÁLISE

python
R
Power BI
tableau

Data Wrangling

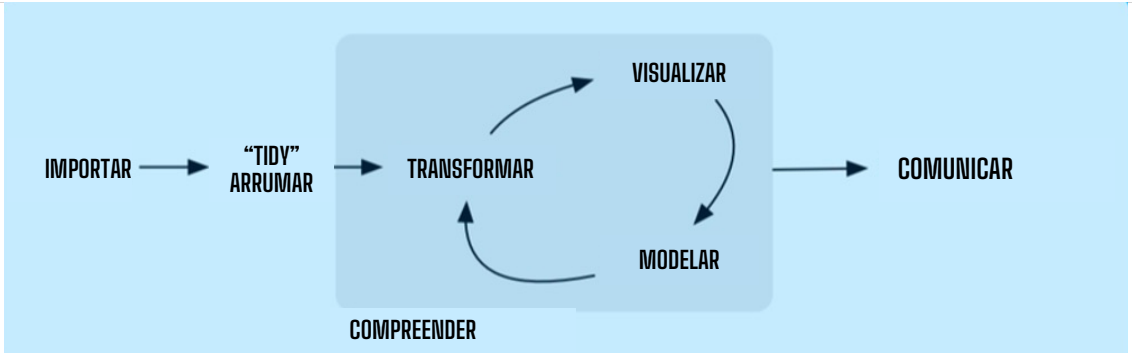
Processo de limpeza, estruturação e enriquecimento dos dados.

Tidy Data:
Variáveis nas **colunas**
Observações nas **linhas**
Valores nas **intersecções**

Padronização de **nomes/grafias**

Atenção a **espaços, caracteres especiais, maiúsculas/minúsculas, acentos**

Adequação dos **tipos de variáveis**



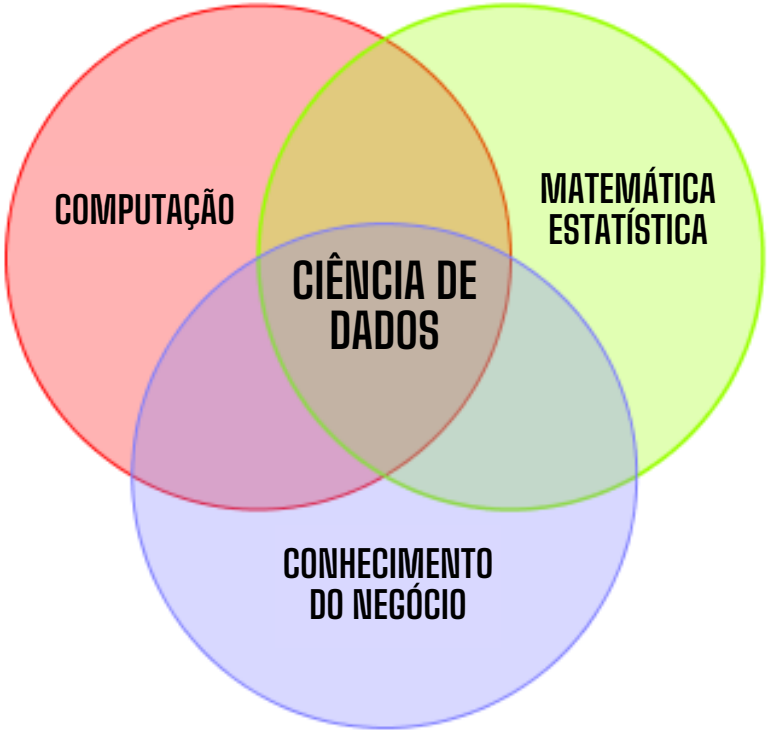

Análise

Requisitos Profissionais (Hard Skills)

- **COMPUTAÇÃO**:
  Excel (VBA, Macros, Funções, Ferramentas de Análise)
  Power B.I (DAX, Medidas, Power Query, Cardinalidade)
  Python/R (sintaxe, bibliotecas, operadores, funções...)
  SQL (consultas em bancos de dados relacionais)

- **MATEMÁTICA/ESTATÍSTICA**:
  Cálculo; Álgebra linear
  Modelagem estatística
  Análise exploratória
  Análise de regressão
  Clusterização
  Machine Learning

- **NEGÓCIO**:
  Momentos do Jogo; Princípios e sub-princípios
  Métricas



Requisitos Profissionais (Soft Skills)

- Criatividade e Curiosidade
- Trabalho em equipe
    - Colaboração
    - Empatia
    - Respeito
- Liderança e gestão de pessoas
- Resolução de problemas e de conflitos
- Comunicação
    - Storytelling (saber contar histórias com início, meio e fim)
    - Visualização (gráficos, mapas, cores, textos)
    - Apresentação (consolidar o produto; falar em público; escrever com correção)
- Adaptabilidade
- Velocidade de reação
- Ética de trabalho e boas práticas

Primeiros Passos pós-E.T.L

- **SABER O QUE SE QUER: entendimento claro do problema**
  Definir objetivos -> Explorar os dados -> Selecionar os dados
  Técnica dos 5 W's: Who, What, Where, When, Why

- **Qual o tipo de variável selecionada?**
  Numérica? Categórica?

- **Qual o tipo de análise?**
  Descritiva? Diagnóstica? Preditiva? Prescritiva?

- **Qual o propósito da análise?**
  Comparação? Relação? Previsão?

- **Qual o método estatístico mais apropriado** para atender ao propósito escolhido e às variáveis selecionadas? Estatística Descritiva? Análise de Associação/Correlação? Testes de Hipóteses? Técnicas Multivariadas Exploratórias/Confirmatórias? Aprendizado de Máquina? Redes Neurais Artificiais?

Estatística Descritiva

Permitem extrair informação das variáveis:

- **Tabelas de Frequência**
  quantidade de ocorrências (absoluta ou relativa), aplicadas em variáveis categóricas

- **Medidas de Posição**
  média, mediana, moda, percentis - variáveis numéricas

- **Medidas de Dispersão**
  amplitude, variância, desvio padrão, coeficiente de variação

Técnicas Multivariadas Exploratórias

Não há intenção de se fazer previsões/inferências. Estudam a relação entre variáveis para:

- Redução ou simplificação dos dados
- Classificação/agrupamento de variáveis categóricas
- **Extração de Fatores**
- Elaboração de **rankings** de desempenho (**ScoreCards**)
- Diagnóstico
- Exemplos: **Análise Fatorial - Exploratória, Confirmatória ou por Componentes Principais; Análise de Agrupamentos (Clusters)**

Técnicas Multivariadas Confirmatórias – Modelos de Regressão

Servem para **prever comportamentos** com base na associação entre **uma variável dependente** e uma ou mais variáveis explicativas. Utiliza a correlação entre as variáveis para:

- Determinar quais variáveis estão relacionadas à variável dependente
- Entender o relacionamento entre as variáveis dependente e explicativas
- Prever valores desconhecidos da variável dependente (prognóstico)

Machine Learning

- Aprendizagem de Máquina
- Utilizada para automatizar tarefas com base na descoberta sistemática de padrões nos conjuntos de dados disponíveis
- Modelos **Supervisionados**: regressão, classificação, árvores de decisão, redes neurais, entre outros
- Modelos **Não-Supervisionados**: agrupamento (cluster), componentes principais (fatorial), redução de dimensionalidade, entre outros
- Divisão dos dados em treinamento (80%) e teste (20%), com diferentes técnicas de recombinação
- **Deep Learning**: várias camadas; ajuste de pesos, viés e erros.

Data Mining

- Aplicação de técnicas e ferramentas de aprendizado de máquina e classificação para minerar os dados **à procura de padrões e relacionamentos despercebidos** dentro de um grande volume

- **Knowledge Discovery in Database (KDD)**:

  Dados → Pré-processamento → Transformação → Mineração → Interpretação → Conhecimento

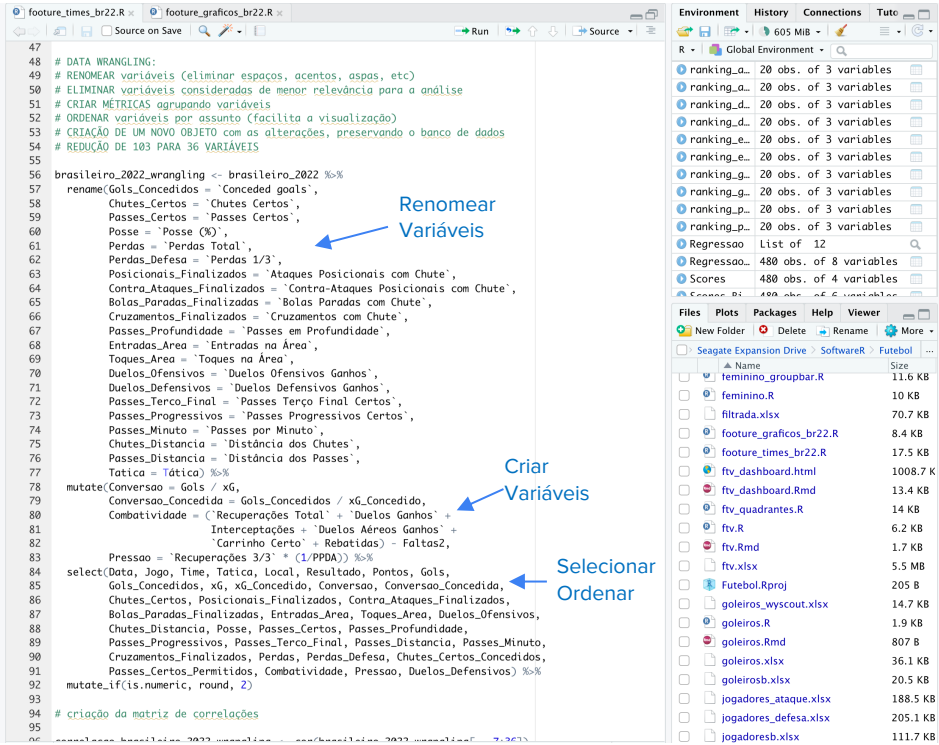
- **Etapas**:

  - **Exploração**
  - **Construção do Modelo**
  - **Validação/Verificação**


Exemplos

Análise Fatorial Exploratória

- Dados do **Brasileiro Série A 2022**, até a **24ª rodada**, extraídos por download na área “stats” dos clubes no site do provedor **Wyscout** (com **97 variáveis e 480 observações**)
- Uso da IDE **RStudio** para processar a análise em linguagem **R**
- **Primeiros passos**:
  - Download do arquivo .xlsx
  - Pré-processamento no **Excel** (como se fosse a “Staging Area” do processo)
  - Carregamento dos dados em um **R Script**
  - Data Wrangling dos dados utilizando funções das bibliotecas **dplyr** e **tidyr**



Análise Fatorial Exploratória

- Mesmo criadas 4 variáveis (função mutate), o processo de transformação do banco de dados reduziu de 97 iniciais para 36 variáveis

  - **Segundo passo**:
    Testar se o **conjunto de dados é compatível** com essa técnica

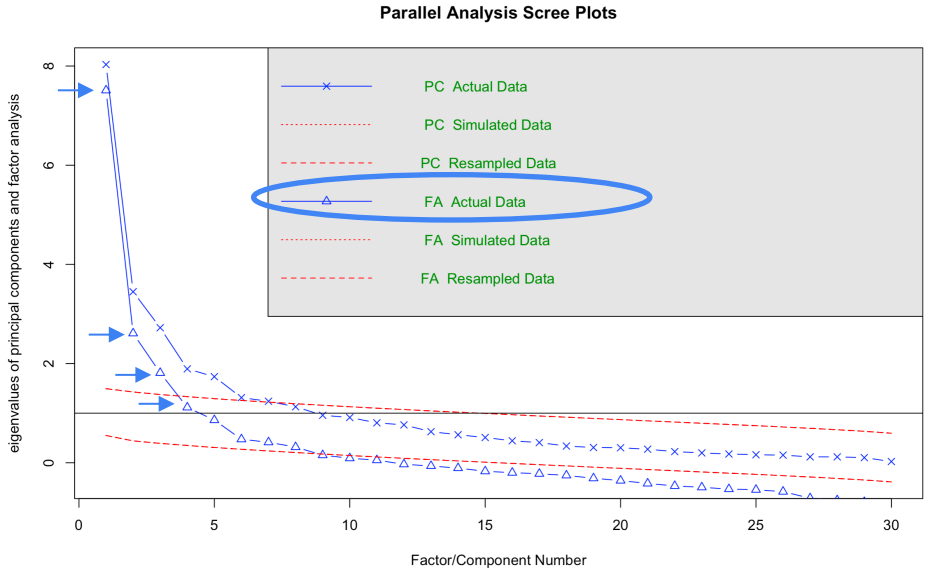
  - Resultados:
    Teste de Esfericidade de Bartlett (p-valor < 5):
    p-valor = 0 (é compatível)

    Teste Kayser-Meyer-Olkin (MSA > 0.5):
    KMO = 0.78 (é compatível)

```
> cortest.bartlett(correlacao_brasileiro_2022_wrangling, n = 480)
$chisq
[1] 10057.32

$p.value
[1] 0

$df
[1] 435

> KMO(brasileiro_2022_wrangling[ , 7:36])
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = brasileiro_2022_wrangling[, 7:36])
Overall MSA =  0.78
MSA for each item =
                    Pontos                       Gols           Gols_Concedidos
                      0.72                       0.64                       0.61
                        xG                xG_Concedido                  Conversao
                      0.78                       0.68                       0.64
       Conversao_Concedida             Chutes_Certos  Posicionais_Finalizados
                      0.53                       0.88                       0.79
Contra_Ataques_Finalizados  Bolas_Paradas_Finalizadas            Entradas_Area
                      0.60                       0.78                       0.92
              Toques_Area          Duelos_Ofensivos            Chutes_Distancia
                      0.87                       0.82                       0.65
                    Posse             Passes_Certos       Passes_Profundidade
                      0.74                       0.73                       0.87
       Passes_Progressivos         Passes_Terco_Final           Passes_Distancia
                      0.88                       0.91                       0.63
           Passes_Minuto    Cruzamentos_Finalizados                      Perdas
                      0.64                       0.90                       0.60
           Perdas_Defesa  Chutes_Certos_Concedidos  Passes_Certos_Permitidos
                      0.86                       0.82                       0.73
           Combatividade                   Pressao          Duelos_Defensivos
                      0.64                       0.92                       0.63
```

Análise Fatorial Exploratória

- **Terceiro passo**:
  - Determinar o **número de fatores** com o uso da função **fa.parallel** (biblioteca **psych**)
  - Critério: auto-valor (eigenvalue, linha contínua do gráfico) superior a 1
  - **Quatro fatores** atendem ao critério, ou seja, quatro é um número **suficiente** de fatores para **explicar o conjunto de dados** composto pelas **36 variáveis**



Análise Fatorial Exploratória

- Utilizando o método “principal fator (**pa**)” como um dos atributos da função **fa** (exploratory fator analysis) da biblioteca **psych**, as variáveis são carregadas nos fatores (algumas em mais de um fator)

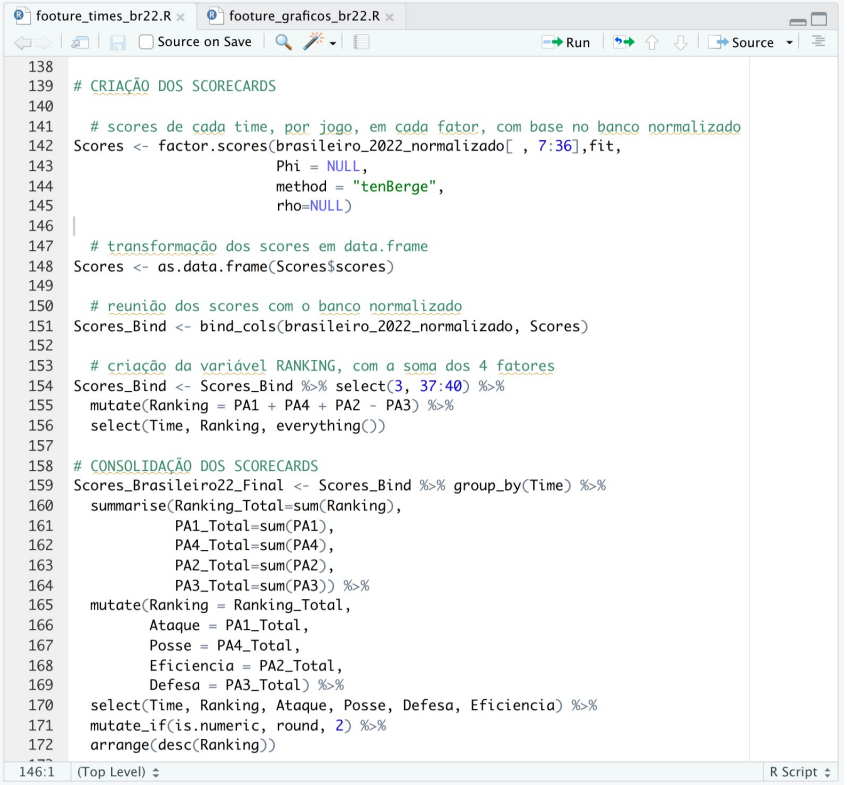
- O fator **PA1**, que explica 42.9% dos dados, agrupa principalmente variáveis de **Ataque**
- O fator **PA4**, que explica 22.7% dos dados, agrupa principalmente variáveis de **Posse**
- O fator **PA2**, que explica 19% dos dados, agrupa diferentes variáveis, com um viés à **Eficiência**
- O fator **PA1**, que explica 15.4% dos dados, agrupa principalmente variáveis de **Defesa**

```
Factor Analysis using method =  pa
Call: fa(r = brasileiro_2022_normalizado[, 7:36], nfactors = 4, rotate = "Promax",
    scores = TRUE, SMC = FALSE, fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
                          item   PA1   PA4   PA2   PA3   h2     u2  com
Entradas_Area               12  0.853                    0.782 0.2180 1.05
Toques_Area                 13  0.853                    0.792 0.2082 1.17
Cruzamentos_Finalizados     23  0.748                    0.517 0.4827 1.07
Posicionais_Finalizados      9  0.689                    0.461 0.5395 1.02
xG                           4  0.642         0.434      0.608 0.3919 2.07
Passes_Profundidade         18  0.641                    0.539 0.4606 1.30
Passes_Progressivos         19  0.625  0.358             0.663 0.3373 1.63
Pressao                     29  0.562                    0.505 0.4954 1.69
Bolas_Paradas_Finalizadas   11  0.468                    0.235 0.7648 1.07
Passes_Certos_Permitidos    27 -0.437 -0.389             0.502 0.4980 2.56
Perdas_Defesa               25 -0.375 -0.336             0.394 0.6060 2.66
Chutes_Distancia            15 -0.333                    0.211 0.7886 2.55
Duelos_Ofensivos            14  0.318                    0.153 0.8473 1.53
Passes_Certos               17         0.853             0.918 0.0819 1.16
Posse                       16  0.460  0.666             0.877 0.1234 1.95
Passes_Minuto               22         0.595             0.379 0.6205 1.13
Passes_Terco_Final          20  0.509  0.559             0.757 0.2432 2.02
Passes_Distancia            21        -0.456             0.193 0.8071 1.34
Gols                         2                0.837      0.687 0.3127 1.01
Pontos                       1                0.789  0.667 0.829 0.1713 2.04
Chutes_Certos                8  0.431         0.513      0.477 0.5230 1.98
Conversao                    6 -0.336         0.412      0.227 0.7731 2.18
Perdas                      24               -0.380  0.339 0.357 0.6430 2.64
Contra_Ataques_Finalizados  10                0.338      0.131 0.8693 1.45
Gols_Concedidos              3                      -0.850 0.688 0.3118 1.20
Combatividade               28        -0.301        0.519 0.524 0.4759 2.31
Chutes_Certos_Concedidos    26                      -0.507 0.329 0.6708 1.95
Conversao_Concedida          7                      -0.414 0.192 0.8075 1.53
xG_Concedido                 5                      -0.397 0.232 0.7677 2.26
Duelos_Defensivos           30                             0.144 0.8555 1.88

                       PA1   PA4   PA2   PA3
SS loadings          6.139 3.250 2.714 2.201
Proportion Var       0.205 0.108 0.090 0.073
Cumulative Var       0.205 0.313 0.403 0.477
Proportion Explained 0.429 0.227 0.190 0.154
Cumulative Proportion 0.429 0.656 0.846 1.000
```

Análise Fatorial Exploratória

- As funções das diferentes bibliotecas estatísticas do R executam a criação do ScoreCard no processo de análise exploratória sem a necessidade de criar um modelo

- Feitos os cálculos, basta ajustar os dados resultantes, com os fatores renomeados, sumarizados pelo total e agrupados por time



Análise Fatorial Exploratória

Posicao	Time	Ranking
1	Atlético-MG	43.05
2	Flamengo	41.27
3	Fluminense	35.25
4	São Paulo	21.59
5	Internacional	20.48
6	**Palmeiras**	**18.76**
7	Atlético-GO	10.06
8	Corinthians	4.27
9	RB Bragantino	1.00
10	Coritiba	0.85
11	Ceará	-8.68
12	Santos	-10.64
13	Avaí	-13.75
14	Botafogo	-15.11
15	Athletico	-16.11
16	Juventude	-19.73
17	Fortaleza	-24.74
18	Cuiabá	-25.38
19	América-MG	-31.03
20	Goiás	-31.43

Posicao	Time	Ataque
1	Flamengo	13.48
2	Atlético-MG	12.90
3	**Palmeiras**	**11.41**
4	São Paulo	7.61
5	Atlético-GO	5.14
6	RB Bragantino	4.91
7	América-MG	3.56
8	Internacional	2.81
9	Fluminense	2.64
10	Santos	-1.01
11	Ceará	-2.00
12	Fortaleza	-2.07
13	Coritiba	-3.91
14	Botafogo	-5.16
15	Cuiabá	-5.46
16	Juventude	-7.12
17	Corinthians	-8.11
18	Avaí	-8.35
19	Athletico	-9.68
20	Goiás	-11.58

Posicao	Time	Defesa
1	América-MG	8.59
2	**Palmeiras**	**5.54**
3	Corinthians	5.32
4	Fortaleza	5.21
5	Santos	3.38
6	Botafogo	2.85
7	Flamengo	1.91
8	RB Bragantino	1.86
9	Cuiabá	1.39
10	Fluminense	0.54
11	Ceará	0.48
12	Atlético-MG	-0.10
13	Avaí	-1.60
14	Athletico	-2.03
15	Internacional	-2.13
16	Goiás	-3.20
17	São Paulo	-3.45
18	Juventude	-6.66
19	Atlético-GO	-6.97
20	Coritiba	-10.92

Posicao	Time	Eficiencia
1	Internacional	15.10
2	Flamengo	14.75
3	**Palmeiras**	**10.40**
4	Fluminense	7.37
5	Athletico	4.47
6	Atlético-MG	4.43
7	Santos	2.34
8	São Paulo	1.67
9	Corinthians	0.55
10	Coritiba	-0.39
11	Goiás	-0.81
12	RB Bragantino	-1.79
13	Ceará	-3.77
14	Atlético-GO	-5.60
15	América-MG	-5.98
16	Cuiabá	-6.51
17	Fortaleza	-6.99
18	Botafogo	-7.99
19	Avaí	-8.93
20	Juventude	-12.33

Posicao	Time	Posse
1	Fluminense	25.78
2	Atlético-MG	25.62
3	Corinthians	17.14
4	Flamengo	14.95
5	São Paulo	8.86
6	Atlético-GO	3.55
7	**Palmeiras**	**2.50**
8	Avaí	1.93
9	Botafogo	0.89
10	Internacional	0.43
11	RB Bragantino	-0.26
12	Ceará	-2.42
13	Coritiba	-5.76
14	Juventude	-6.94
15	Santos	-8.58
16	Fortaleza	-10.47
17	Cuiabá	-12.02
18	Athletico	-12.93
19	América-MG	-20.02
20	Goiás	-22.23

O **Palmeiras**, líder do Campeonato na 24ª rodada, foi destacado nas tabelas (geradas com a biblioteca **kableExtra**) divididas por ranking total e os quatro fatores

No **ranking geral**, o Palmeiras aparece apenas em **6º**, mas fica evidente que é o **fator Posse (7º) o responsável pelo contraste** com sua pontuação no campeonato

O Palmeiras ficou em 2º no fator Defesa, e em 3º nos fatores Ataque e Eficiência

Análise Fatorial Confirmatória

- Ao contrário da técnica exploratória, na Análise Fatorial Confirmatória o pesquisador faz a modelagem dos fatores e submete o modelo à validação

- Foram mantidos os 4 fatores encontrados na técnica exploratória, com a criação de 2 critérios:

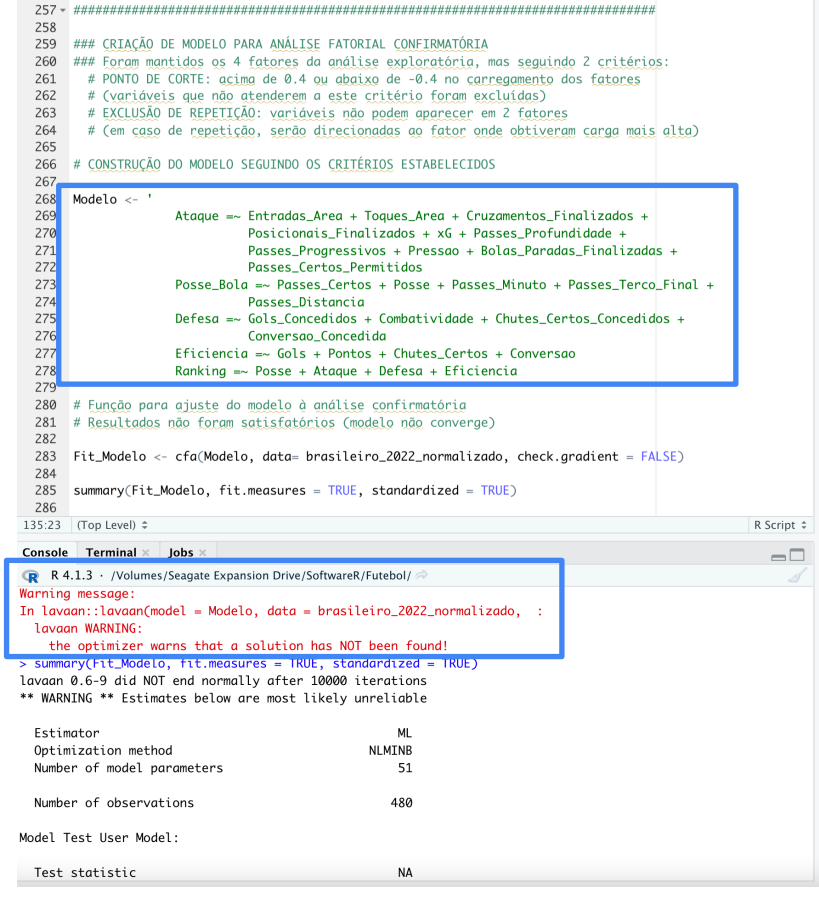
  - Variáveis com carregamentos entre -0.4 e 0.4 foram excluídas
  - Variáveis carregadas originalmente em dois fatores foram direcionadas ao fator de maior carga

```
Factor Analysis using method =  pa
Call: fa(r = brasileiro_2022_normalizado[, 7:36], nfactors = 4, rotate = "Promax",
    scores = TRUE, SMC = FALSE, fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
                          item   PA1    PA4    PA2    PA3    h2     u2  com
Entradas_Area               12  0.853                      0.782 0.2180 1.05
Toques_Area                 13  0.853                      0.792 0.2082 1.17
Cruzamentos_Finalizados     23  0.748                      0.517 0.4827 1.07
Posicionais_Finalizados      9  0.689                      0.461 0.5395 1.02
xG                           4  0.642         0.434        0.608 0.3919 2.07
Passes_Profundidade         18  0.641                      0.539 0.4606 1.30
Passes_Progressivos         19  0.625  0.358               0.663 0.3373 1.63
Pressao                     29  0.562                      0.505 0.4954 1.69
Bolas_Paradas_Finalizadas   11  0.468                      0.235 0.7648 1.07
Passes_Certos_Permitidos    27 -0.437 -0.389               0.502 0.4980 2.56
Perdas_Defesa               25 -0.375 -0.336               0.394 0.6060 2.66
Chutes_Distancia            15 -0.333                      0.211 0.7886 2.55
Duelos_Ofensivos            14  0.318                      0.153 0.8473 1.53
Passes_Certos               17         0.853               0.918 0.0819 1.16
Posse                       16  0.460  0.666               0.877 0.1234 1.95
Passes_Minuto               22         0.595               0.379 0.6205 1.13
Passes_Terco_Final          20  0.509  0.559               0.757 0.2432 2.02
Passes_Distancia            21        -0.456               0.193 0.8071 1.34
Gols                         2                0.837        0.687 0.3127 1.01
Pontos                       1                0.789  0.667 0.829 0.1713 2.04
Chutes_Certos                8  0.431         0.513        0.477 0.5230 1.98
Conversao                    6 -0.336         0.412        0.227 0.7731 2.18
Perdas                      24                -0.380  0.339 0.357 0.6430 2.64
Contra_Ataques_Finalizados  10                 0.338        0.131 0.8693 1.45
Gols_Concedidos              3                       -0.850 0.688 0.3118 1.20
Combatividade               28        -0.301         0.519 0.524 0.4759 2.31
Chutes_Certos_Concedidos    26                       -0.507 0.329 0.6708 1.95
Conversao_Concedida          7                       -0.414 0.192 0.8075 1.53
xG_Concedido                 5                       -0.397 0.232 0.7677 2.26
Duelos_Defensivos           30                              0.144 0.8555 1.88

                       PA1   PA4   PA2   PA3
SS loadings          6.139 3.250 2.714 2.201
Proportion Var       0.205 0.108 0.090 0.073
Cumulative Var       0.205 0.313 0.403 0.477
Proportion Explained 0.429 0.227 0.190 0.154
Cumulative Proportion 0.429 0.656 0.846 1.000
```

Análise Fatorial Confirmatória

- A função **cfa** (Confirmatory Factor Analysis) da biblioteca **lavaan** emitiu um aviso que os dados não são confiáveis, ou seja, não recomendou que a pesquisa tivesse sequência - o que não impede a continuidade do processo (não é uma falha)

- Esse tipo de aviso costuma surgir quando o modelo é submetido a bancos de dados com pouca amostragem, o que é o caso



```
257 ##############################################################################
258
259 ### CRIAÇÃO DE MODELO PARA ANÁLISE FATORIAL CONFIRMATÓRIA
260 ### Foram mantidos os 4 fatores da análise exploratória, mas seguindo 2 critérios:
261   # PONTO DE CORTE: acima de 0.4 ou abaixo de -0.4 no carregamento dos fatores
262   # (variáveis que não atenderem a este critério foram excluídas)
263   # EXCLUSÃO DE REPETIÇÃO: variáveis não podem aparecer em 2 fatores
264   # (em caso de repetição, serão direcionadas ao fator onde obtiveram carga mais alta)
265
266 # CONSTRUÇÃO DO MODELO SEGUINDO OS CRITÉRIOS ESTABELECIDOS
267
268 Modelo <- '
269               Ataque =~ Entradas_Area + Toques_Area + Cruzamentos_Finalizados +
270                         Posicionais_Finalizados + xG + Passes_Profundidade +
271                         Passes_Progressivos + Pressao + Bolas_Paradas_Finalizadas +
272                         Passes_Certos_Permitidos
273               Posse_Bola =~ Passes_Certos + Posse + Passes_Minuto + Passes_Terco_Final +
274                         Passes_Distancia
275               Defesa =~ Gols_Concedidos + Combatividade + Chutes_Certos_Concedidos +
276                         Conversao_Concedida
277               Eficiencia =~ Gols + Pontos + Chutes_Certos + Conversao
278               Ranking =~ Posse + Ataque + Defesa + Eficiencia
279
280 # Função para ajuste do modelo à análise confirmatória
281 # Resultados não foram satisfatórios (modelo não converge)
282
283 Fit_Modelo <- cfa(Modelo, data= brasileiro_2022_normalizado, check.gradient = FALSE)
284
285 summary(Fit_Modelo, fit.measures = TRUE, standardized = TRUE)
286
135:23  (Top Level)                                                    R Script

Console  Terminal  Jobs
R 4.1.3 · /Volumes/Seagate Expansion Drive/SoftwareR/Futebol/
Warning message:
In lavaan::lavaan(model = Modelo, data = brasileiro_2022_normalizado,  :
  lavaan WARNING:
    the optimizer warns that a solution has NOT been found!
> summary(Fit_Modelo, fit.measures = TRUE, standardized = TRUE)
lavaan 0.6-9 did NOT end normally after 10000 iterations
** WARNING ** Estimates below are most likely unreliable

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        51

  Number of observations                           480

Model Test User Model:

  Test statistic                                    NA
```

Análise Fatorial Confirmatória

Posicao	Time	Ranking
1	Atlético-MG	233.12
2	Flamengo	202.59
3	Fluminense	186.83
4	Palmeiras	131.16
5	Internacional	90.93
6	São Paulo	82.61
7	RB Bragantino	33.03
8	Corinthians	22.34
9	Atlético-GO	-4.57
10	Ceará	-12.24
11	Santos	-32.38
12	América-MG	-45.07
13	Botafogo	-51.29
14	Fortaleza	-61.65
15	Avaí	-79.49
16	Coritiba	-98.37
17	Cuiabá	-107.72
18	Athletico	-136.95
19	Juventude	-149.96
20	Goiás	-202.94

Posicao	Time	Ataque
1	Atlético-MG	113.11
2	Flamengo	97.13
3	Palmeiras	83.07
4	São Paulo	56.89
5	Fluminense	51.29
6	RB Bragantino	39.22
7	Atlético-GO	29.78
8	Internacional	16.60
9	América-MG	9.27
10	Ceará	-10.89
11	Santos	-18.32
12	Coritiba	-26.87
13	Botafogo	-33.27
14	Fortaleza	-34.14
15	Cuiabá	-43.95
16	Corinthians	-51.56
17	Juventude	-55.54
18	Avaí	-69.41
19	Goiás	-75.50
20	Athletico	-76.90

Posicao	Time	Defesa
1	Palmeiras	22.16
2	Santos	20.02
3	América-MG	13.72
4	Fortaleza	10.81
5	Internacional	9.97
6	Cuiabá	7.04
7	Flamengo	7.01
8	Corinthians	5.93
9	Ceará	4.29
10	Atlético-MG	4.04
11	Botafogo	0.99
12	RB Bragantino	-2.85
13	Avaí	-3.74
14	Goiás	-4.38
15	Athletico	-5.72
16	Fluminense	-7.73
17	São Paulo	-10.46
18	Atlético-GO	-21.61
19	Coritiba	-22.04
20	Juventude	-27.45

Posicao	Time	Eficiencia
1	Palmeiras	38.71
2	Fluminense	28.24
3	Flamengo	26.91
4	Internacional	24.01
5	Atlético-MG	17.79
6	RB Bragantino	12.18
7	São Paulo	4.82
8	Corinthians	4.12
9	Athletico	3.18
10	Fortaleza	0.92
11	Santos	-4.79
12	Coritiba	-8.00
13	Goiás	-8.85
14	Ceará	-9.14
15	Botafogo	-11.80
16	Atlético-GO	-14.09
17	América-MG	-14.41
18	Avaí	-24.99
19	Juventude	-29.82
20	Cuiabá	-34.99

Posicao	Time	Posse
1	Fluminense	115.03
2	Atlético-MG	98.18
3	Flamengo	71.54
4	Corinthians	63.86
5	Internacional	40.36
6	São Paulo	31.36
7	Avaí	18.64
8	Ceará	3.50
9	Atlético-GO	1.34
10	Botafogo	-7.20
11	Palmeiras	-12.79
12	RB Bragantino	-15.51
13	Santos	-29.29
14	Cuiabá	-35.82
15	Juventude	-37.15
16	Fortaleza	-39.24
17	Coritiba	-41.45
18	América-MG	-53.65
19	Athletico	-57.52
20	Goiás	-114.21

Os ajustes nos fatores fizeram o Palmeiras **subir de 6º para 4º no ranking geral**.

Também houve melhora em **Defesa** (de **2º para 1º**) e **Eficiência** (de **3º para 1º**), mantendo-se **3º em Ataque**.

No entanto, houve queda em **Posse**, de **7º para 11º**. A diferença para o 1º deste fator - em ambas técnicas o Fluminense - passou de 23.28 para 127.82.

Ou seja, o resultado acentua a percepção sobre o fator **Posse tratar-se mais de estilo do que de performance**, enquanto os demais fatores se mostram determinantes.

Regressão Linear Múltipla

- Dando sequência à exploração deste banco de dados com variadas técnicas estatísticas através da linguagem R, a **Regressão Linear Múltipla** reforça os insights resultantes da comparação entre as análises fatoriais exploratória e confirmatória

- O código é muito simples e objetivo:
  - Primeiro se deve apresentar as variáveis dependente e explicativas
  - Na fórmula atribuída à função **lm** (fitting linear models) da biblioteca **stats**, a variável **Pontos é a dependente**, enquanto os **fatores Ataque, Posse, Defesa e Eficiência são as variáveis explicativas**
  - A função **summary** permite visualizar o **relatório da análise**
  - A função **step**, da biblioteca **stats**, executa o método **Stepwise**

```
399 ##############################################################################
400
401 ### REGRESSÃO LINEAR UTILIZANDO O MODELO ELABORADO
402
403 # Seleção da variável Pontos, sumarizada pela soma e agrupada por time
404
405 Pontos <- brasileiro_2022_wrangling %>% select(Time, Pontos) %>%
406                                         group_by(Time) %>%
407                                         summarise(Pontos=sum(Pontos))
408
409 # Combinação da variável pontos com o ranking consolidado do modelo CFA
410
411 Regressao_Normalizado <- left_join(modelo_cfa_brasileiro22, Pontos, by = "Time")
412
413 # Função para executar a Regressão Linear
414
415 Regressao <- lm(formula = Pontos ~ Ataque + Posse + Defesa + Eficiencia,
416                 data = Regressao_Normalizado)
417
418 #Visualização da análise do algoritmo
419
420 summary(Regressao)
421
422 # Método Stepwise
423 # usado para selecionar quais variáveis mais influenciam o conjunto de saída,
424 # o que pode diminuir o número de variáveis que compõem a equação de regressão.
425
426 step(Regressao)
427
427:1  (Untitled)  R Script
```

Regressão Linear Múltipla

- O relatório apresenta diversas medidas de ajuste da fórmula, que **justificam aplicar a técnica neste banco de dados**
- Destaca-se o Adjusted R-squared de 0.74 (ou seja, **74.3% da variação dos dados pode ser explicado pelas variáveis preditoras**). A proximidade dele com o Multiple R-squared (0.79) é considerada um sinal positivo
- Também chama atenção que, entre os fatores, apenas Posse tem p-valor alto (0.41), enquanto Defesa e Eficiência são menores que 0,05 (parâmetro de referência), e Ataque está próximo (0,08) - indicando **menor relevância da Posse** em comparação com os demais fatores

```
Call:
lm(formula = Pontos ~ Ataque + Posse + Defesa + Eficiencia, data = Regressao_Normalizado)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1277 -2.4661 -0.3175  3.1717  6.4168

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.39980    0.96555  33.556 1.59e-15 ***
Ataque      -0.05209    0.02824  -1.845 0.084924 .
Posse        0.01906    0.02271   0.839 0.414632
Defesa       0.30942    0.07666   4.036 0.001077 **
Eficiencia   0.48304    0.10668   4.528 0.000401 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.318 on 15 degrees of freedom
Multiple R-squared:  0.7972,    Adjusted R-squared:  0.7431
F-statistic: 14.74 on 4 and 15 DF,  p-value: 4.442e-05
```

Regressão Linear Múltipla

- O método Stepwise é utilizado para selecionar as variáveis que mais influenciam no conjunto de saída, o que pode levar à exclusão de variáveis da fórmula
- Neste caso, o Stepwise excluiu a variável Posse e manteve as demais

```
Start:  AIC=62.76
Pontos ~ Ataque + Posse + Defesa + Eficiencia

             Df Sum of Sq    RSS    AIC
- Posse       1     13.13 292.81 61.676
<none>                    279.69 62.759
- Ataque      1     63.45 343.13 64.848
- Defesa      1    303.75 583.43 75.464
- Eficiencia  1    382.25 661.94 77.989

Step:  AIC=61.68
Pontos ~ Ataque + Defesa + Eficiencia

             Df Sum of Sq    RSS    AIC
<none>                    292.81 61.676
- Ataque      1     50.72 343.53 62.871
- Defesa      1    293.24 586.05 73.553
- Eficiencia  1    418.31 711.12 77.422

Call:
lm(formula = Pontos ~ Ataque + Defesa + Eficiencia, data = Regressao_Normalizado)

Coefficients:
(Intercept)       Ataque       Defesa   Eficiencia
   32.39977     -0.04304      0.30198      0.49810
```

Conclusão

- Ao final da exploração dos dados, cabe ao pesquisador/cientista/analista **tomar decisões sobre os próximos passos**, sempre **orientado à solução do problema de negócio** que originou o processo (o que queremos?)

- É preciso ter **profundo conhecimento do problema e do usuário-alvo** da análise (dirigente? treinador? jogador?) para escolher entre as diversas técnicas, interpretar os resultados e entregar o produto-final

- Neste caso, se estamos orientados à construção do **monitoramento dos modelos de jogo**, uma boa solução é **explorar maneiras de melhor aproveitar as variáveis de posse** (quem sabe com outras métricas? Ou então voltando à construção do objeto de **data wrangling** para selecionar outras variáveis de posse que foram descartadas?)

- Por outro lado, se a intenção é aumentar a **força preditora dos dados**, o ideal é **manter a exclusão da posse** e aplicar técnicas de **machine learning** para aumentar a capacidade de prever pontos baseado nos fatores, o que depende do **aumento da amostragem**

Análise Fatorial Exploratória com dados de jogadores

```r
jogadores_brasileiro_2022 <- read_excel("wyscout_jogadores_br22.xlsx")

# função para extrair apenas a 1ª posição dos jogadores listados com mais de uma posição

posicoes <- word(jogadores_brasileiro_2022$Posição, 1)

# visualização do total de posições descritas
# há diversas duplicatas de conceitos e duplicatas de grafias seguidas com vírgulas

unique(posicoes)

# funções para reajuste da grafia e dos conceitos das posições

posicoes[posicoes %in% c('LCB', 'RCB,', 'RCB', 'LCB,', 'CB,')] <- 'CB' #zagueiro
posicoes[posicoes %in% c('RB,', 'RWB', 'RWB,')] <- 'RB' #lateral-direito
posicoes[posicoes %in% c('LB,', 'LWB', 'LWB,')] <- 'LB' #lateral-esquerdo
posicoes[posicoes %in% c('RDMF', 'RDMF,', 'DMF', 'LDMF,', 'LDMF', 'DMF,')] <- 'DM' #volante
posicoes[posicoes %in% c('RCMF,', 'LCMF,', 'RCMF', 'LCMF' )] <- 'CM' #meia-central
posicoes[posicoes %in% c('AMF,', 'RAMF', 'AMF', 'LAMF,', 'LAMF', 'RAMF,')] <- 'OM' #meia-ofensiv
posicoes[posicoes %in% c('LWF,', 'LW,', 'LW', 'LM')] <- 'LM' #extremo-esquerdo
posicoes[posicoes %in% c('RW,', 'RWF,', 'RW', 'RWF')] <- 'RM' #extremo-direito
posicoes[posicoes == 'CF,'] <- 'CF' #centroavante

# substituição da variável original pela variável ajustada

jogadores_brasileiro_2022$Posição <- posicoes

# Há ocorrência de apenas uma observação NA na posição
# O jogador é João Victor, do Botafogo. Em breve pesquisa encontra-se que ele é CF

jogadores_brasileiro_2022$Posição[is.na(jogadores_brasileiro_2022$Posição)] <- 'CF'

unique(jogadores_brasileiro_2022$Posição)
```

Limpeza dos dados com variáveis string

48:1 (Top Level) R Script

Dados extraídos do Wyscout com **674 observações** (jogadores) e **97 variáveis** (médias por jogo ou totais).

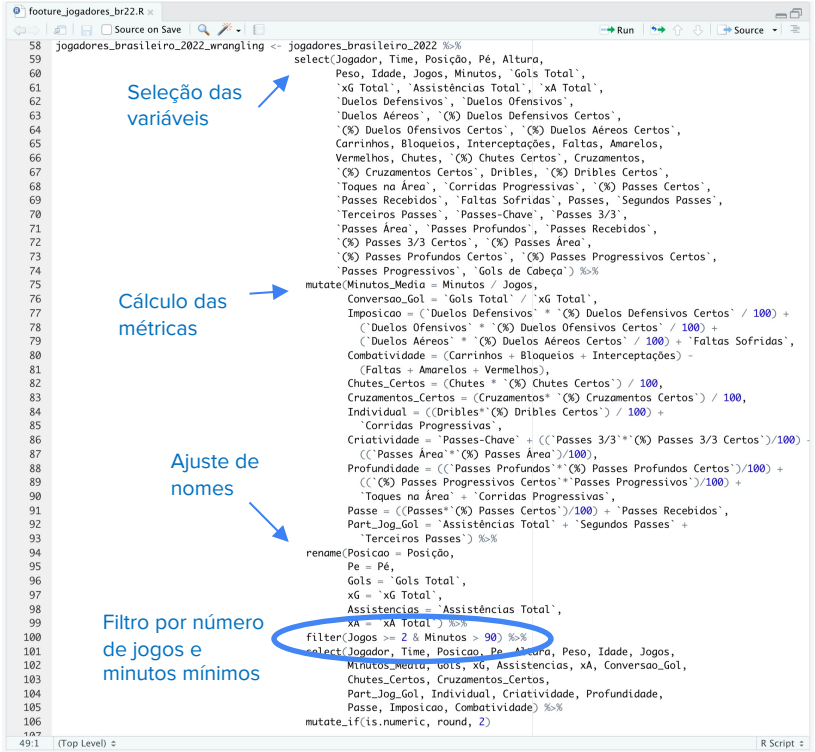
Recorte: Campeonato Brasileiro 2022 até a 24ª rodada.

Técnica de **Análise Fatorial Exploratória** para extrair **rankings por função**.

Wyscout cataloga 9 funções de linha:

zagueiro, laterais direito e esquerdo, volante, meia-central, meia-atacante, extremos esquerdo e direito, e centroavante.

Análise Fatorial Exploratória com dados de jogadores




```
jogadores_brasileiro_2022_wrangling <- jogadores_brasileiro_2022 %>%
                                       select(Jogador, Time, Posição, Pé, Altura,
                                              Peso, Idade, Jogos, Minutos, `Gols Total`,
                                              `xG Total`, `Assistências Total`, `xA Total`,
                                              `Duelos Defensivos`, `Duelos Ofensivos`,
                                              `Duelos Aéreos`, `(%) Duelos Defensivos Certos`,
                                              `(%) Duelos Ofensivos Certos`, `(%) Duelos Aéreos Certos`,
                                              Carrinhos, Bloqueios, Interceptações, Faltas, Amarelos,
                                              Vermelhos, Chutes, `(%) Chutes Certos`, Cruzamentos,
                                              `(%) Cruzamentos Certos`, Dribles, `(%) Dribles Certos`,
                                              `Toques na Área`, `Corridas Progressivas`, `(%) Passes Certos`,
                                              `Passes Recebidos`, `Faltas Sofridas`, Passes, `Segundos Passes`,
                                              `Terceiros Passes`, `Passes-Chave`, `Passes 3/3`,
                                              `Passes Área`, `Passes Profundos`, `Passes Recebidos`,
                                              `(%) Passes 3/3 Certos`, `(%) Passes Área`,
                                              `(%) Passes Profundos Certos`, `(%) Passes Progressivos Certos`,
                                              `Passes Progressivos`, `Gols de Cabeça`) %>%
                                       mutate(Minutos_Media = Minutos / Jogos,
                                              Conversao_Gol = `Gols Total` / `xG Total`,
                                              Imposicao = (`Duelos Defensivos` * `(%) Duelos Defensivos Certos` / 100) +
                                                (`Duelos Ofensivos` * `(%) Duelos Ofensivos Certos` / 100) +
                                                (`Duelos Aéreos` * `(%) Duelos Aéreos Certos` / 100) + `Faltas Sofridas`,
                                              Combatividade = (Carrinhos + Bloqueios + Interceptações) -
                                                (Faltas + Amarelos + Vermelhos),
                                              Chutes_Certos = (Chutes * `(%) Chutes Certos`) / 100,
                                              Cruzamentos_Certos = (Cruzamentos* `(%) Cruzamentos Certos`) / 100,
                                              Individual = ((Dribles*`(%) Dribles Certos`) / 100) +
                                                `Corridas Progressivas`,
                                              Criatividade = `Passes-Chave` + ((`Passes 3/3`*`(%) Passes 3/3 Certos`)/100) -
                                                ((`Passes Área`*`(%) Passes Área`)/100),
                                              Profundidade = ((`Passes Profundos`*`(%) Passes Profundos Certos`)/100) +
                                                ((`(%) Passes Progressivos Certos`*`Passes Progressivos`)/100) +
                                                `Toques na Área` + `Corridas Progressivas`,
                                              Passe = ((Passes*`(%) Passes Certos`)/100) + `Passes Recebidos`,
                                              Part_Jog_Gol = `Assistências Total` + `Segundos Passes` +
                                                `Terceiros Passes`) %>%
                                       rename(Posicao = Posição,
                                              Pe = Pé,
                                              Gols = `Gols Total`,
                                              xG = `xG Total`,
                                              Assistencias = `Assistências Total`,
                                              xA = `xA Total`) %>%
                                       filter(Jogos >= 2 & Minutos > 90) %>%
                                       select(Jogador, Time, Posicao, Pe, Altura, Peso, Idade, Jogos,
                                              Minutos_Media, Gols, xG, Assistencias, xA, Conversao_Gol,
                                              Chutes_Certos, Cruzamentos_Certos,
                                              Part_Jog_Gol, Individual, Criatividade, Profundidade,
                                              Passe, Imposicao, Combatividade) %>%
                                       mutate_if(is.numeric, round, 2)
```

Data Wrangling com **filtro** determinando valores mínimos de jogos (2) e minutos (90) reduziu **de 674 para 579** observações (jogadores);

A criação de **11 métricas** reduziu **de 97 para 23 variáveis**.

Exemplos

Análise Fatorial Exploratória com dados de jogadores

```
Factor Analysis using method =  pa
Call: fa(r = zagueiros_fatorial[, 5:23], nfactors = 6, rotate = "Promax",
    scores = TRUE, fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
                   item   PA2   PA3   PA1   PA4   PA5   PA6    h2    u2  com
Criatividade         15 0.926                                0.871 0.1294 1.08
Profundidade         16 0.922                                0.842 0.1584 1.05
Passe                17 0.657                                0.408 0.5919 1.11
Part_Jog_Gol         13        1.003                         0.976 0.0241 1.01
Assistencias          8        1.001                         0.983 0.0170 1.00
xG                    7              0.829                   0.714 0.2860 1.33
Jogos                 4              0.768                   0.616 0.3843 1.14
Minutos_Media         5              0.490                   0.270 0.7295 1.23
xA                    9              0.395                   0.435 0.5646 2.75
Peso                  2                    0.951             0.934 0.0665 1.02
Altura                1                    0.916             0.825 0.1746 1.03
Conversao_Gol        10                          0.831       0.655 0.3445 1.15
Gols                  6              0.414       0.682       0.705 0.2953 1.70
Chutes_Certos        11                          0.666       0.453 0.5469 1.03
Individual           14 0.361                                0.517 0.431 0.5694 2.09
Cruzamentos_Certos   12 0.335                                0.483 0.454 0.5460 2.31
Imposicao            18                                      0.401 0.177 0.8229 1.39
Idade                 3                                      0.145 0.8546 2.70
Combatividade        19                                      0.244 0.7560 4.40
```

```
Factor Analysis using method =  pa
Call: fa(r = volantes_fatorial[, 5:23], nfactors = 3, rotate =
    scores = TRUE, fm = "pa")
Standardized loadings (pattern matrix) based upon correlation m
                   item   PA1   PA2   PA3    h2    u2  com
Part_Jog_Gol         13 0.808             0.7054 0.295 1.25
Assistencias          8 0.794             0.6772 0.323 1.23
xA                    9 0.731             0.6282 0.372 1.24
xG                    7 0.723       0.332 0.7742 0.226 1.69
Jogos                 4 0.645             0.4505 0.550 1.22
Minutos_Media         5 0.552             0.3291 0.671 1.19
Peso                  2                   0.0583 0.942 1.49
Criatividade         15       0.822       0.6875 0.312 1.03
Profundidade         16       0.783       0.7022 0.298 1.18
Passe                17       0.749       0.5760 0.424 1.09
Combatividade        19      -0.405       0.1640 0.836 1.01
Altura                1                   0.0957 0.904 1.48
Idade                 3                   0.0606 0.939 2.15
Conversao_Gol        10             0.731 0.5158 0.484 1.08
Gols                  6 0.429       0.671 0.7564 0.244 1.80
Chutes_Certos        11             0.611 0.4652 0.535 1.26
Cruzamentos_Certos   12             0.582 0.3682 0.632 1.31
Individual           14             0.311 0.1928 0.807 2.19
Imposicao            18                   0.0609 0.939 2.19
```

```
Factor Analysis using method =  pa
Call: fa(r = meias_fatorial[, 5:23], nfactors = 4, rotate = "Promax",
    scores = TRUE, fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
                   item   PA1   PA3   PA2   PA4    h2    u2  com
xG                    7 0.948                   0.86283 0.137 1.37
xA                    9 0.771                   0.72600 0.274 1.15
Jogos                 4 0.699                   0.43497 0.565 1.19
Assistencias          8 0.688                   0.72281 0.277 1.43
Part_Jog_Gol         13 0.687                   0.72988 0.270 1.46
Minutos_Media         5 0.526                   0.27852 0.721 1.18
Criatividade         15       0.939             0.81111 0.189 1.05
Profundidade         16       0.926             0.89048 0.110 1.02
Passe                17       0.752             0.49303 0.507 1.20
Cruzamentos_Certos   12       0.384             0.26809 0.732 1.98
Combatividade        19                         0.00461 0.995 1.59
Peso                  2              0.941       0.86250 0.138 1.03
Altura                1              0.937       0.88315 0.117 1.01
Idade                 3              0.338       0.14152 0.858 1.61
Conversao_Gol        10                    0.785 0.60796 0.392 1.06
Gols                  6 0.443             0.686 0.75523 0.245 1.76
Chutes_Certos        11                   0.463 0.30812 0.692 1.42
Imposicao            18                   0.309 0.12474 0.875 1.84
Individual           14                         0.30003 0.700 2.94
```

```
Factor Analysis using method =  pa
Call: fa(r = meias_ofensivos_fatorial[, 5:23], nfactors = 3, rot
    scores = TRUE, fm = "pa")
Standardized loadings (pattern matrix) based upon correlation ma
                   item   PA1   PA2   PA3    h2    u2  com
Assistencias          8 0.899             0.7868 0.2132 1.00
Part_Jog_Gol         13 0.898             0.7947 0.2053 1.00
xG                    7 0.871             0.6868 0.3132 1.04
xA                    9 0.854             0.7689 0.2311 1.01
Gols                  6 0.804             0.6109 0.3891 1.03
Jogos                 4 0.722             0.5363 0.4637 1.00
Minutos_Media         5 0.665             0.4562 0.5438 1.13
Chutes_Certos        11                   0.1289 0.8711 2.51
Combatividade        19                   0.0481 0.9519 1.68
Criatividade         15       0.902       0.8089 0.1911 1.08
Passe                17       0.853       0.7074 0.2926 1.01
Idade                 3       0.647       0.4272 0.5728 1.07
Imposicao            18                   0.1711 0.8289 2.33
Individual           14             0.887 0.7758 0.2242 1.09
Profundidade         16 0.540 0.808       0.9264 0.0736 1.75
Cruzamentos_Certos   12       0.349       0.2108 0.7892 1.73
Peso                  2                   0.0915 0.9085 1.46
Conversao_Gol        10                   0.1415 0.8585 2.09
Altura                1                   0.0614 0.9386 1.14
```

Dados contextualizados: cada função é filtrada para ser analisada em separado, o que gera diferentes resultados.
Zagueiros: 6 fatores; Peso e Altura separados em um fator, carregamento alto; indicadores ofensivos altos.
Volantes: 3 fatores; Altura e Peso desconsiderados; indicadores de criatividade se destacam.
Meias-Centrais: 4 fatores; Idade carrega com Peso e Altura em um fator; xG alto.
Meias-Atacantes: 3 fatores; Individual aparece, com valor alto; Idade aumenta a influência; Assistências alto.

Análise Fatorial Exploratória com dados de jogadores

Ranking	Zagueiro	Time
1	G. Gómez	Palmeiras
2	Manoel	Fluminense
3	V. Cuesta	Botafogo
4	J. Alonso	Atlético Mineiro
5	Natan	Red Bull Bragantino
6	Willian Arão	Fenerbahçe
7	Joaquim Henrique	Cuiabá
8	Éder Ferreira	América Mineiro
9	Léo Ortiz	Red Bull Bragantino
10	Igor Rabello	Atlético Mineiro

Ranking	Lateral Direito	Time
1	Yago Pikachu	Shimizu S-Pulse
2	Marcos Rocha	Palmeiras
3	Rodrigo Soares	Juventude
4	Rafinha	São Paulo
5	F. Bustos	Internacional
6	Kevin	Avaí
7	Samuel Xavier	Fluminense
8	Mayke	Palmeiras
9	Matheuzinho	Flamengo
10	Mariano	Atlético Mineiro

Ranking	Lateral Esquerdo	Time
1	Reinaldo	São Paulo
2	Luan Cândido	Red Bull Bragantino
3	Welington	São Paulo
4	Ayrton Lucas	Flamengo
5	Fábio Santos	Corinthians
6	Caio Paulista	Fluminense
7	Guilherme Arana	Atlético Mineiro
8	Moisés	Internacional
9	Daniel Borges	Botafogo
10	Filipe Luis	Flamengo

Análise Fatorial Exploratória com dados de jogadores

Ranking	Volante	Time
1	Marlon Freitas	Atlético GO
2	Andrey	Coritiba
3	Jair	Atlético Mineiro
4	E. Atuesta	Palmeiras
5	Zé Rafael	Palmeiras
6	Fernandinho	Athletico Paranaense
7	Lucas Evangelista	Red Bull Bragantino
8	Zanocelo	Santos
9	Val	Coritiba
10	Gabriel Menino	Palmeiras

Ranking	Meia	Time
1	C. de Pena	Internacional
2	Lima	Ceará
3	Alê	América Mineiro
4	Andreas Pereira	Fulham
5	Jean Pyerre	Avaí
6	Rodrigo Nestor	São Paulo
7	Fellipe Bastos	Goiás
8	A. Vidal	Flamengo
9	André	Fluminense
10	Hércules	Fortaleza

Ranking	Meia Ofensivo	Time
1	Dudu	Palmeiras
2	G. De Arrascaeta	Flamengo
3	I. Fernández	Atlético Mineiro
4	J. Arias	Fluminense
5	Gustavo Scarpa	Palmeiras
6	Taison	Internacional
7	Igor Paixão	Coritiba
8	Ganso	Fluminense
9	Robinho	Coritiba
10	Éverton Ribeiro	Flamengo

Análise Fatorial Exploratória com dados de jogadores

Ranking	Extremo	Time
1	Ângelo	Santos
2	S. Mendoza	Ceará
3	Pedro Henrique	Internacional
4	Wanderson	Internacional
5	Edenílson	Internacional
6	William Pottker	Avaí
7	Patrick	São Paulo
8	Luiz Henrique	Real Betis
9	Pedrinho	América Mineiro
10	Éverton	Flamengo

Ranking	Centroavante	Time
1	Pedro Raul	Goiás
2	Hulk	Atlético Mineiro
3	Pedro	Flamengo
4	Vina	Ceará
5	Gabriel Barbosa	Flamengo
6	Alemão	Internacional
7	Cléber	Ceará
8	J. Calleri	São Paulo
9	G. Cano	Fluminense
10	Moisés	Fortaleza


Visualização

Tópicos Relevantes

- Ajuda a encontrar padrões e a explica-los, **orientando a tomada de decisão**

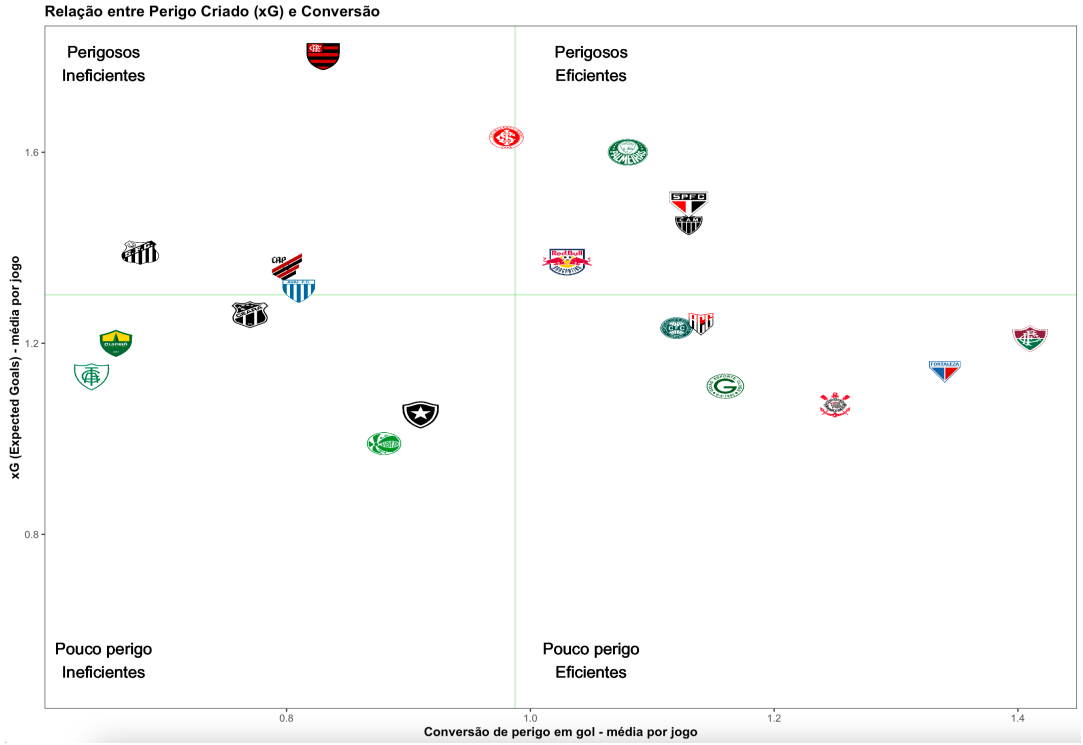
- Softwares como **Power B.I e Tableau** facilitam a análise de forma intuitiva (“**drag and drop**”)
  - **Dimensões**: perspectivas da análise (filtros)
  - **Medidas**: métricas.

- **Python e R**, cada uma com suas peculiaridades, são linguagens de programação **orientadas a objetos e voltadas a estatística e ciência de dados**, oferecendo **algoritmos e representações gráficas** de alto nível

- **Cor, posição e tamanho** são atributos que, além de estéticos, podem ajudar na compreensão dos dados

- Existem **gráficos apropriados** para diferentes situações:
  - **Linha**: variação de valores ao longo do tempo
  - **Dispersão** (Scatterplot): comparação de variáveis quantitativas
  - **Radar**: correlação entre variáveis
  - **Colunas agrupadas** (Stacked): comparação de variáveis categóricas

Gráfico de Dispersão – Ataque (xG x Conversão)

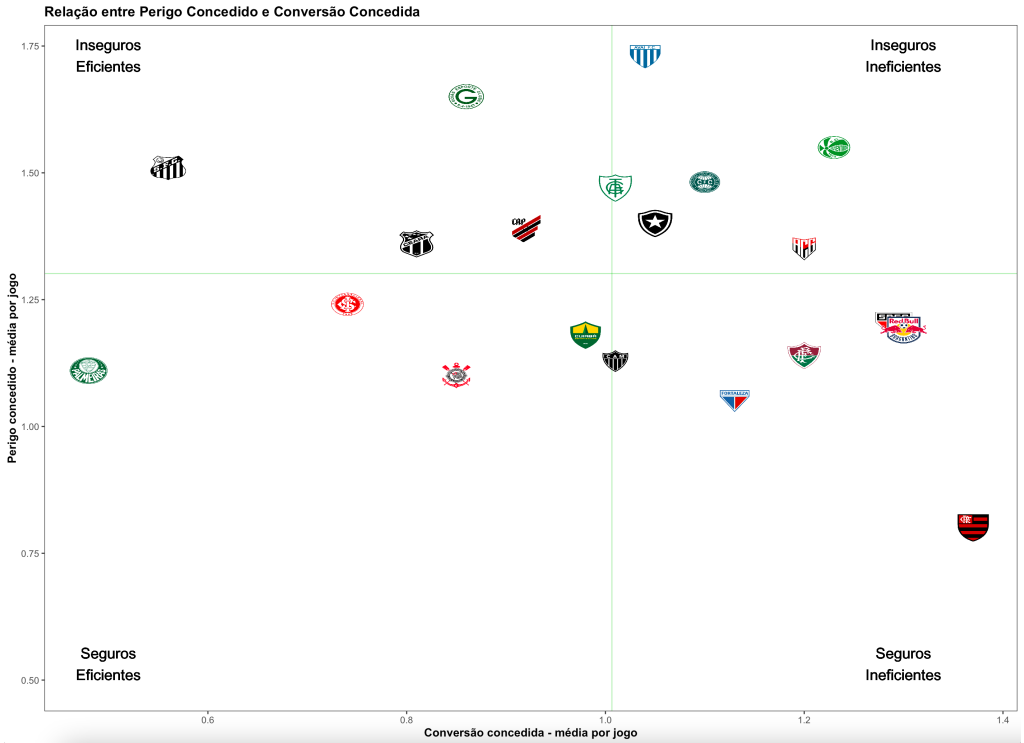


O scatterplot pode ser **dividido em quadrantes através de linhas com as médias das duas variáveis comparadas**, o que permite ao pesquisador contextualizar cada zona.

Neste caso, xG e Conversão altos identificam equipes **perigosas e eficientes** (como o Palmeiras), enquanto xG e Conversão baixos identificam equipes **pouco perigosas e ineficientes** (como Juventude e Botafogo).

Há ainda os **perigosos e ineficientes** (como o Flamengo), e os **pouco perigosos e eficientes** (como Fortaleza e Fluminense)

Gráfico de Dispersão – Defesa (xG e Conversão Concedidos)

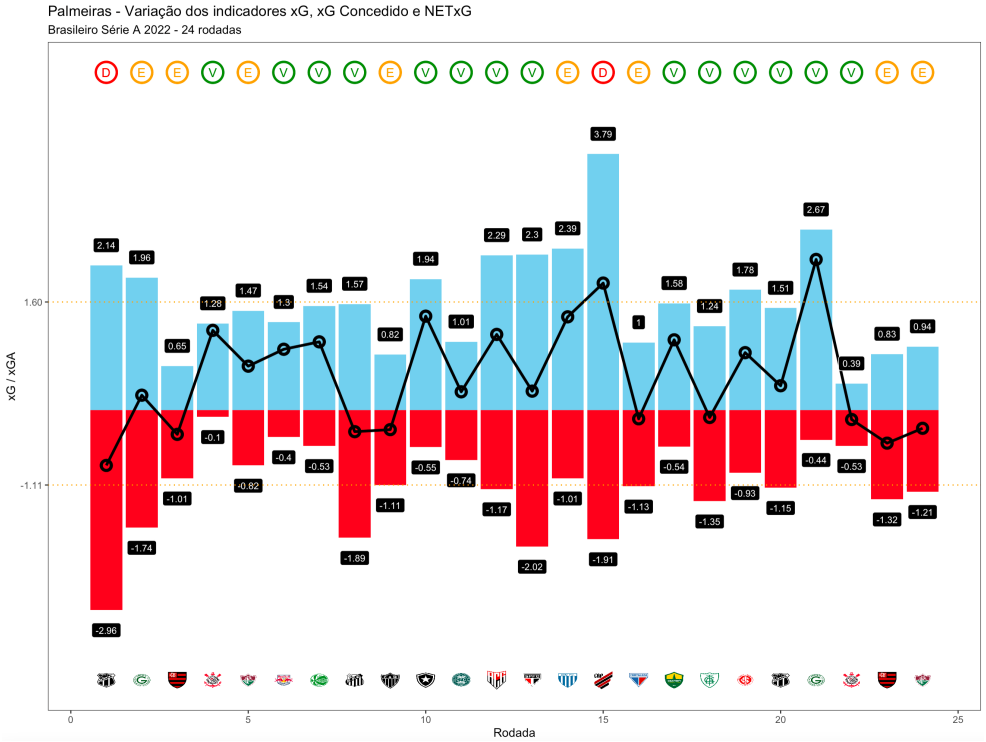


Também elaborado no **RStudio** e seguindo o mesmo princípio, as equipes agora são comparadas quanto à **eficiência defensiva** (perigo concedido vs. conversão concedida).

Percebe-se que utilizar os logos dos clubes, embora divertido, pode ser um problema quando os valores estão próximos (**São Paulo e Bragantino sobrepostos**).

Outro ponto negativo poderia ser entregar este gráfico para um **cliente que não está familiarizado com os logos** dos clubes analisados, dificultando a identificação.

Gráfico de Barras Duplas e Linha



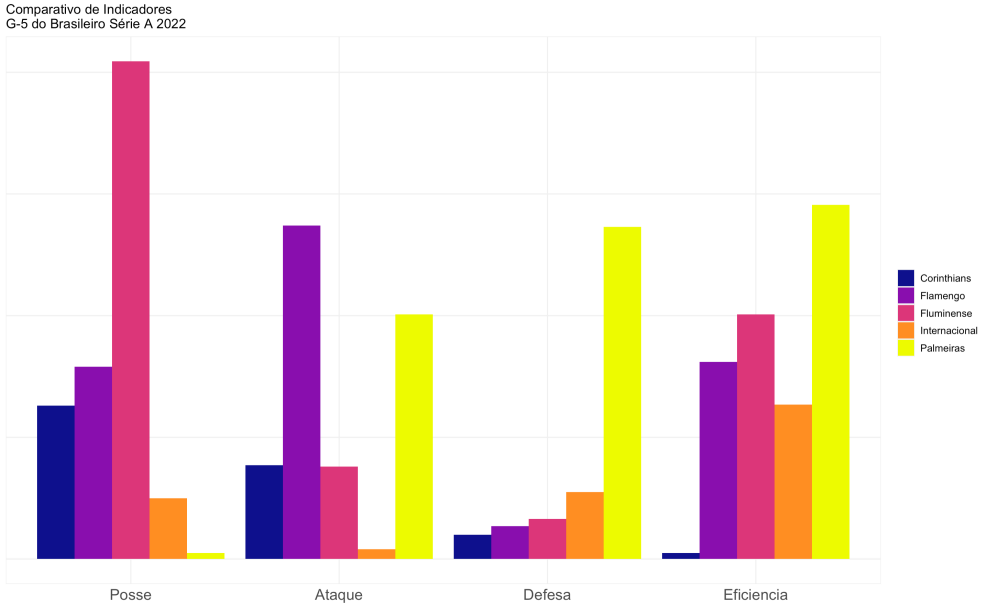
Este gráfico agrupa 3 variáveis (**xG, xG Concedido e NET xG**) de uma equipe analisada (no exemplo, o Palmeiras) ao longo do tempo (da **1ª à 24ª rodada** do Brasileiro Série A 2022).

Apresentar este gráfico requer **clareza nas legendas explicando os conceitos** (seja em relatório, seja em aplicação):

xG nas barras azuis;
xG Concedido (xGA) nas barras vermelhas;
NETxG (xG - xGA) na linha preta com pontos;
Resultado (V, E ou D) no topo;
Adversário (logos) na base.

É possível acompanhar a **evolução das variáveis contextualizadas ao adversário, ao resultado da partida e à sequência**.

Gráfico de Colunas Agrupadas

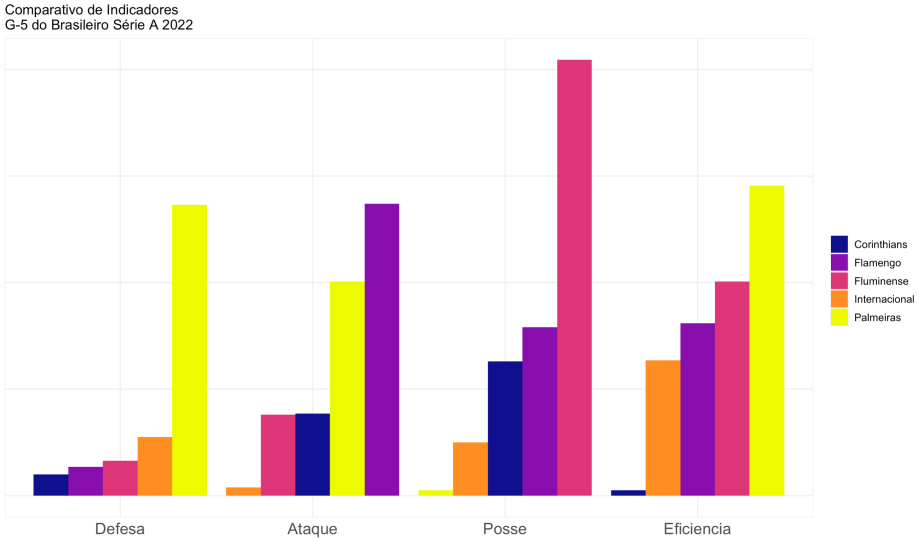


As colunas agrupadas permitem **fácil comparação entre variáveis categóricas**.

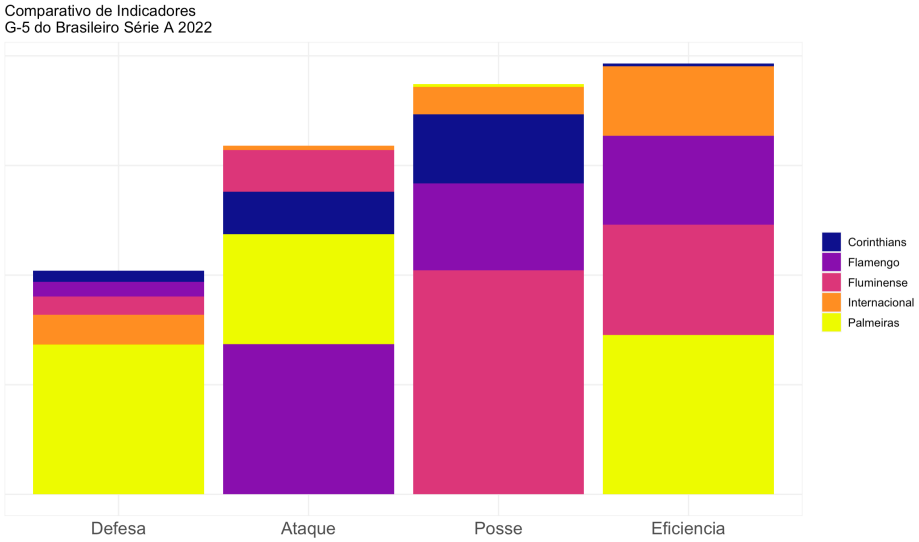
No exemplo foram **selecionados os 5 primeiros colocados do Brasileiro** Série A 2022 até a 24ª rodada para comparação dentro dos **quatro fatores resultantes da análise fatorial confirmatória**.

Fica muito evidente o contraste do Palmeiras, extremamente abaixo em posse, mas acima em defesa (principalmente) e eficiência, assim como superior a 3 dos 4 adversários selecionados (abaixo apenas do Flamengo).

MESMOS DADOS, MESMO GRÁFICO BASE, VÁRIAS POSSIBILIDADES

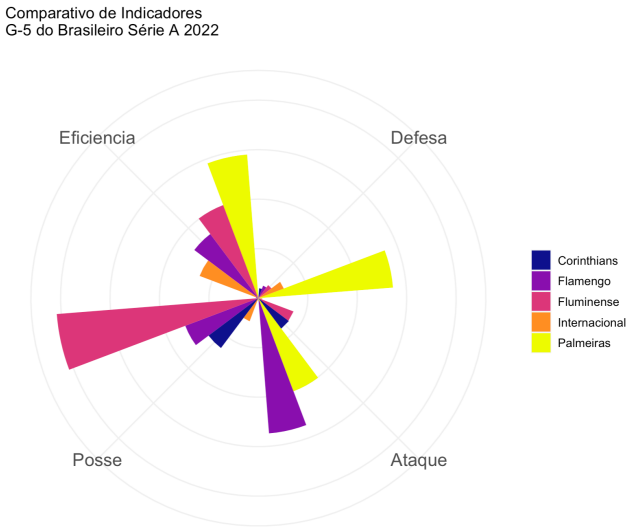


Colunas Agrupadas - position “dodge”

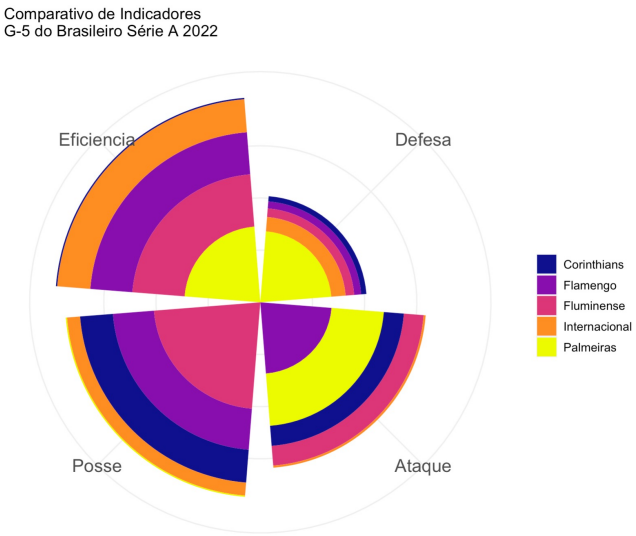


Colunas Agrupadas - position “stack”

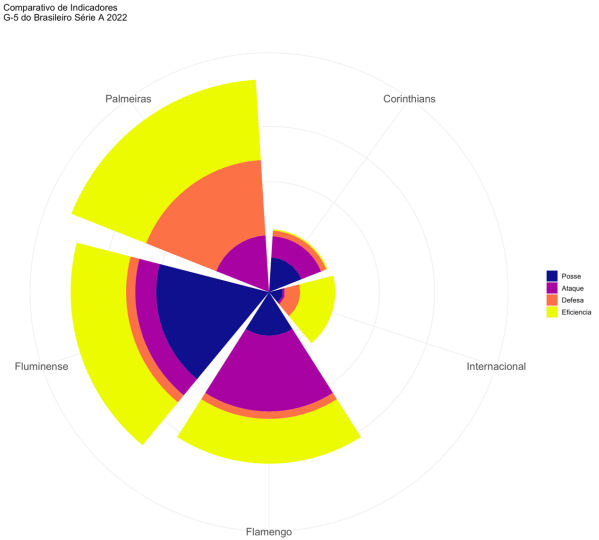
MESMOS DADOS, MESMO GRÁFICO BASE, VÁRIAS POSSIBILIDADES



Position “dodge” e COORD_POLAR



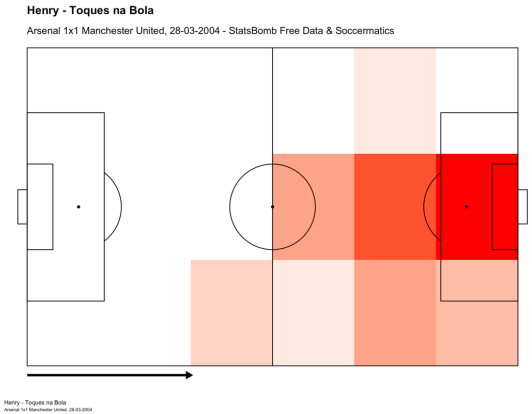
Position “stack” e COORD_POLAR

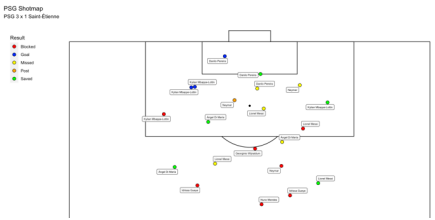


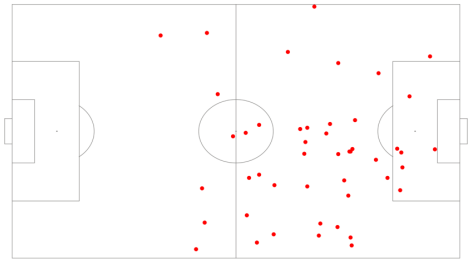
Position “dodge” e COORD_POLAR

Inversão das variáveis nos eixos

Coordenadas x/y







Bases de dados com as coordenadas x/y das ações permitem representações gráficas em campos, seja em quadrantes, com pontos, flechas, traços, etc.

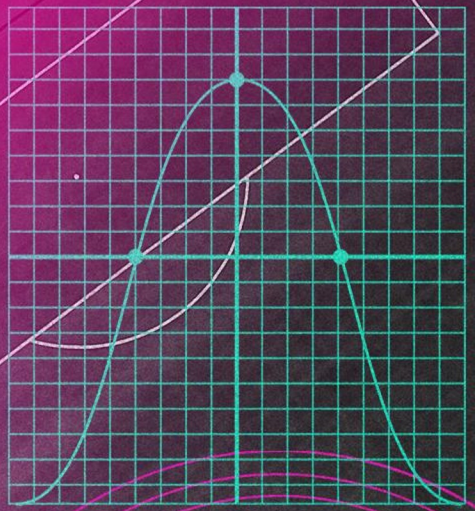
Poucos provedores fornecem dados de futebol com coordenadas x/y confiáveis, e os serviços têm preço muito mais elevado.

Nos dois primeiros exemplos ao lado, foram utilizados dados tornados públicos pela empresa **StatsBomb** na biblioteca **StatsBombR**, enquanto o mapa de chutes combina as bibliotecas **worldfootballR** e **ggsoccer**.

Melhores Amigos dos Programadores

stackoverflow.com

QR Code para acessar **repositório no Git Hub com os Scripts em R e os dados em .xlsx** utilizados na aula



Informações para
contato/dúvidas
QR para o perfil no Linkedin

Muito obrigado!