ATLANTE
Cuadernos de Educación y Desarrollo (2ª Época)

Marzo 2019 - ISSN: 1989-4155

ANÁLISE DE DESEMPENHO E PERFIS DE TIMES DO CAMPEONATO BRASILEIRO

Felipe Sheison Raimundo Santos,
Aluno de Bacharelado em Sistemas de Informação, Centro Universitário Padre Anchieta (Unianchieta), Jundiaí/SP, felipesheison@yahoo.com.br

Juliano Schimiguel,
Professor da Universidade Cruzeiro do Sul (São Paulo/SP), e do Centro Universitário Padre Anchieta – Unianchieta (Jundiaí/SP), schimiguel@gmail.com

Para citar este artículo puede utilizar el siguiente formato:

Felipe Sheison Raimundo Santos y Juliano Schimiguel (2019): "Análise de desempenho e perfis de times do campeonato brasileiro", Revista Atlante: Cuadernos de Educación y Desarrollo (marzo 2019). En línea:

https://www.eumed.net/rev/atlante/2019/03/desempenho-campeonato-brasileiro.html

Resumo

Para esse trabalho utilizamos análise de dados para descobrir os melhores e piores times, melhores e piores jogadores baseados em ações feitas dentro de cada partida de futebol. A partir dessa investigação podemos descobrir a causa que leva um jogador a ser um bom jogador ou um mal jogador e o que leva a um time ser campeão de um torneio. O método que foi utilizado são dados variados dos jogadores que são as ações que cada jogador faz em campo como: Gols, assistência, faltas e etc. Sendo assim esses dados serão coletados tratado, analisados para obter resultados e até mesmo fazer algoritmos para saber como um time pode chegar a uma vitória ou derrota em uma partida. O resultado dessas análises nos traz os melhores e piores times, melhores e piores jogadores, grupos de jogadores com mais gols na competição e média de vitórias e derrotas de cada time. Para chegar em todas essas resultados, foram estudados por meio de artigos, livros e sites, sobre metodologia de dados para saber como coletar e tratar os dados e utilizar as informações certas para chegar em resultados afirmativos, livros de análises de dados para saber como ler e interpretar as informações, foi utilizados livros para montar um banco de dados com a ferramenta SQL Server, com as informações dos times e jogadores, montar um DW para utilizar uma melhor forma de extração dos dados e por fim gerar gráficos e matrizes na ferramenta RStudio.

Palavra-chave: Análise dados; Linguagem RStudio; Banco de Dados.

Abstract

For this work we use a data analysis to find the best and worst times, the best and the most favored people based on actions taken within each football match. From the time of departure, you can find out what leads a player to be a good player or a player that leads to a tournament champion. The method that was used is the varied data of the players that are the actions that each one does in the field like: Goals, assist, faults and others. a team can reach a victory or defeat in a match. The games are more competitive than the best, the best times, the best games, the player games with the most goals in the competition and the wins and losses each time.

To get results in courses, through studies of books, books and websites, on the methodology of data to obtain information and to treat the data and use the information to arrive at affirmative actions, data analysis books to learn how to read and interpreted as information, was used to assemble a database with a SQL Server tool, with information about times and players, such as a database to calculate data extraction and generate graphs and arrays data.

Keywords: Analyze data; Language RStudio; Database.

1. INTRODUÇÃO

Dados, informação e conhecimento, esses são o essencial para começar um qualquer projeto de análise de dados. Ao inicio e final do projeto esses fatores foram utilizados desde planejamento ao termino de conclusões do projeto. Eles representam um ou mais significados de um sistema que isoladamente não pode transmitir uma mensagem ou representar algum conhecimento (de Silva, 2007).

O projeto obteve um processamento de dados para uma série de atividades ordenadas tendo em mente realizar, uma espécie de inúmeras informações, pois no início da atividade é feita a coleta de informações(dado), que passam a ser organizadas onde no final será passada para o usuário o dado limpo a sua busca (de Fernandes, 1992).

Após esse processamento foi feita a mineração de dados onde podemos encontrar anomalias, padrões e correlações em grandes conjuntos de dados para prever resultados, após essa utilização de foram feitas análises dos dados (AMO, 2010).

Nessa análise podemos transformar números em informações e solucionar problemas tanto de banco de dados como informações (RAMOS, 2015).

- Exploração: Investigar os dados e o processo, apenas para ver o que se pode aprender.
- Gerar hipóteses sobre as causas: Utilizar o novo conhecimento encontrado para identificar as causas mais prováveis para os defeitos.
- Testar as hipóteses ou eliminar as causas: Utilizar os dados, experimentos, ou mais análise do processo para verificar quais das causas geram os problemas.

Utilizando a técnica de Business Intelligence foi possível usar a coleta de dados, organização, análise, ação e monitoramento para tomar melhores decisões e saber se os investimentos feitos estão trazendo bons resultados, BI é um conjunto de processos que tem por objetivo entregar a informação certa, para a pessoa certa, na hora certa.

Com os seguintes pontos abaixo:

- coleta de dados: tudo o que acontece no negócio é analisado para determinar aspectos-chave, como produtividade, aproveitamento de oportunidades, gargalos, reputação no mercado, etc.
- organização e análise: todos os dados captados em cada ação da empresa são organizados em um banco de dados e apresentados de forma visual, para facilitar a análise dos tomadores de decisão;

- ação e monitoramento: os responsáveis tomam decisões com base nas informações analisadas, e monitoram seus resultados para ver se estão sendo bem-sucedidos.

Big Data é um conceito que descreve o grande volume de dados estruturados e não estruturados. Big Data está atrelado à possibilidade e oportunidade em cruzar esses dados por meio de diversas fontes para obtermos insights rápidos e preciosos (Junjie Wu).

O trunfo desse trabalho nos mostra a utilização de agrupamento k-means é um método de Clustering que objetiva particionar várias observações dentre grupos onde cada observação pertence ao grupo mais próximo da média. Isso resulta em uma divisão do espaço de dados em um Diagrama de Voronoi (Junjie Wu, 2014).

Com os conceitos acima foi construído um banco de dados com 20 times do campeonato Brasileiro 2013, cada time obtendo 11 jogadores, dos times foram coletas as seguintes informações vitórias e derrotas, já dos jogadores foram coletadas as seguintes informações gols, assistências, cartões, chutes ao gols e etc. Para análises futuras, essas informações foram armazenadas em um banco de dados (SGBD) utilizando a ferramenta SQL Server, onde foram extraídos e analisados pela ferramenta RStudio, baseados em gráficos e algoritmos.

Segundo “MAIA, Humberto Maia” o esporte mais popular do planeta descobriu que intuição e improviso não são suficientes.

O problema maior é descobrir o porquê os times e jogadores estão indo mal e o que faz aquele time ir mal em uma competição tão longa quanto o campeonato brasileiro, dentro de uma competição dessa a realização de trocas de jogadores são enormes, com isso queremos descobrir o desempenho por jogador e o desempenho do time em si com base em análises e estratégias de dados.

Justificamos que esse tema nos leva a querer estudar cada vez mais análise de dados, ou seja, seguir as regras de processos de dados! Para poder fazer uma coleta necessárias e organizar, reestruturar, analisar e utilizar de forma correta a parte de BI, onde devemos ler os dados de várias formas diferentes e interpretá-las da maneira mais correta possível. Sendo assim ganhamos a confiança do cliente e podemos utilizar essas informações corretas para melhorias constantes dentro de uma organização.

O objetivo principal é demonstrar a análise dos times e jogadores individuais, visando mostrar os melhores e piores times e identificando os jogadores, para ter análises concretas conforme resultados dos times e ações feitas por cada jogador em campo;

O objetivo secundário é demonstrar as inconsistências que existem na base de dados e apontar essas inconsistências para que sejam melhoradas e arrumadas de forma sucinta.

Uma metodologia própria para ocasiões que envolvam data mining é um ponto crucial para quem pretende ser profissional na área. Com isso utilizamos a metodologia Cross Industry Standard Process for Data Mining (CRISP-DM) pra solucionar e construir todo o projeto essa é uma metodologia especificamente desenhada para processos de mineração de dados. Essa técnica de nome gigante é uma das mais utilizados em Mineração de Dados. Suas principais vantagens é que

ela pode ser aplicada a qualquer tipo de negócio, e não tem dependência de ferramenta para ser executada, onde essa metodologia é feita em 6 passos: Entendimento do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e desenvolvimento (CRISP, 1999).

2. REFERENCIAS TEÓRICAS.

Para o desenvolvimento deste trabalho, foi realizado pesquisa na internet e livros, artigos. Assim foram utilizados conceitos de Big Data, B.I e data mining.

Para utilização de Big Data, foi construir um banco de dados robusto com muita informação e grande volume de dados para utilização de volume, velocidade, variedade e variabilidade dentro do mesmo (Junjie Wu). As técnicas que foram implementadas junto ao conceito de Big Data foi Data Mining que são minerar os dados é a prática de peneirar todas as evidências em busca de padrões anteriormente não reconhecidos com isso foi utilizado para organiza dados, encontrando neles padrões, associações, mudanças e anomalias relevantes conforme sugere (MJ Berry, G Linoff – 1997)

Uma das formas de alcançar esse diferencial pode ser por meio da ferramenta *Business Intelligence* (BI). Segundo Porter (1989), a tecnologia da informação é uma alavanca cada vez mais importante a disposição das empresas para criar vantagem competitiva. O conceito de *Business Intelligence* (BI), é o relacionamento entre as tecnologias da informação e os processos de negócio.

3. METODOLOGIA E TÉCNOLOGIA UTILIZADAS.

3.1. Tecnologias

As tecnologias utilizadas para o desenvolvimento dos dados e análises foram a linguagem de programação RStudio para analises e banco de dados SQL Server para construção do banco de dados. Cada tecnologia será descrita nas seções que seguem.

Segundo (EGEA, 2005) um SGBD (sistema de gerenciamento de banco de dados) é uma ferramenta de gerenciamento de dados, mais conhecida (SQL) e utilizada em uma linguagem. Segundo (EGEA, 2005) o SQL Server é uma SGBD da Microsoft, foi desenvolvido com a Sybase, na época de 1988, como um complemento do Windows NT, depois passou a aprimorado e vendido separados dos demais. No ano de 1994 a Sysbase terminou e então a Microsoft continuou a aperfeiçoar a ferramenta cada vez mais. Com isso utilizamos essa ferramenta para construir o banco de dados e gerenciar todas informações para o desenvolvimento do projeto.

Segundo (BRAUN, 2007) Linguagem R é baseado na linguagem de computador S, desenvolvida por John Chambers e outros no Bell Laboratories em 1976. Em 1993, Robert Gentleman e Ross Ihaka, da University of Auckland, quiseram experimentar a linguagem, desenvolveram a implementação e nomearam R. e centenas de pessoas em todo o mundo contribuíram para o seu desenvolvimento. S-PLUS é uma implementação comercial da linguagem S. Uma vez que ambos osRandS-PLUS se basearam na linguagem, muito do que é descrito no que se segue aplica-se sem alterações ao S-PLUS. Com isso utilizando o RStudio foram criados vários gráficos realizando análise exploratória e cruzando as informações para uma processo de melhoria de leitura dos dados, esse processo de análise exploratória foi feito a partir dos gráficos e matrizes utilizadas no projeto que são explicadas abaixo.

Segundo (BRAUN, 2007) diagramas: São Gráficos de duas ou mais dimensões e são construídos pelo plano cartesiano.

Segundo (BRAUN, 2007) os gráficos são classificados em horizontal que são apresentadas as variações geográficas e entre outros setores, já a parte vertical apresentam os valores para os fenômenos.

Os principais diagramas são:

Segundo (BRAUN, 2007) diagrama por pontos: são os dados são representados por pontos. É indicado para análises estatísticas.

Para o desenvolvimento deste trabalho, foi realizado pesquisa na internet e livros, artigos. Assim foram utilizados conceitos de Big Data, B.I e data mining.

Para utilização de Big Data, foi construir um banco de dados robusto com muita informação e grande volume de dados para utilização de volume, velocidade, variedade e variabilidade dentro do mesmo (Junjie Wu). As técnicas que foram implementadas junto ao conceito de Big Data foi Data

Mining que são minerar os dados é a prática de peneirar todas as evidências em busca de padrões anteriormente não reconhecidos com isso foi utilizado para organiza dados, encontrando neles padrões, associações, mudanças e anomalias relevantes conforme sugere (MJ Berry, G Linoff – 1997)

Uma das formas de alcançar esse diferencial pode ser por meio da ferramenta *Business Intelligence* (BI). Segundo Porter (1989), a tecnologia da informação é uma alavanca cada vez mais importante a disposição das empresas para criar vantagem competitiva. O conceito de *Business Intelligence* (BI), é o relacionamento entre as tecnologias da informação e os processos de negócio.

3.2. Metodologia

Pirâmide do Conhecimento: Esse modelo de hierarquia é muito usado em áreas como Ciência da Informação e também da Gestão de Conhecimento. Existem quatro conceitos usados nesse modelo sendo eles: Dados, informação, conhecimento e sabedoria.

Tudo se começa com os dados que serão utilizados no processo. Esse componente de dados pode ser números, palavras e imagens! Consegue esclarecer e facilitar a compreensão das informações. Dados na hierarquia DIKW correspondem à Data, o nível mais básico da pirâmide de conhecimento.

Essas informações são a junção das partes fornecidas pelos dados e acrescentam sentido e totalidade aos dados utilizados.

O conhecimento mostra a forma adequada de usar as informações coletadas, e a sabedoria consegue trabalhar o entendimento de quando usar as informações corretas e cria o seu próprio contexto.

Veja a estrutura da pirâmide de conhecimento.

Figura 1 Fonte:

Outro método a ser utilizado foi a Metodologia Crisp-DM: é uma metodologia desenhada para processos de mineração de dados

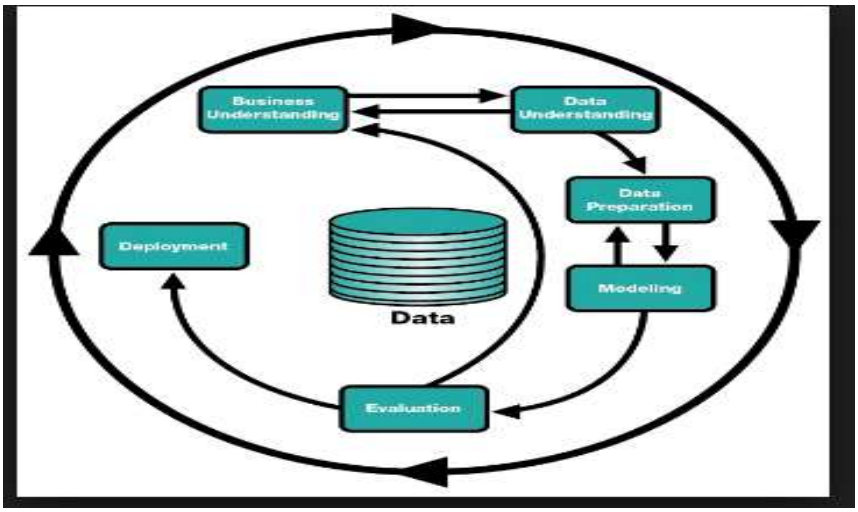

Figura 2 Fonte:

O CRISP-DM baseia-se em tentativas anteriores de definir metodologias de descoberta de conhecimento (Reinartz & Wirth, 1995; Adriaans & Zantinge, 1996; Brachman e Anand, 1996; Fayyad et al., 1996).

Esta seção fornece uma visão geral da metodologia CRISP-DM. Informações mais detalhadas podem

ser encontrado em (CRISP, 1999).

A metodologia CRISP-DM é descrita em termos de um modelo de processo hierárquico, compreendendo

quatro níveis de abstração (do geral ao específico): fases, tarefas genéricas, tarefas especializadas e instâncias de processo

Essa metodologia são estipuladas em 6 etapas:

- **1º - Entendimento do negócio:** Nesta etapa é preciso buscar detalhes especificos de como a questão afeta a organização e seus principais objetivos e expectativas em relação ao trabalho como um todo (CRISP, 1999).
- **2º - Compreensão dos dados:** O objetivo é organizar e descrever todos os dados disponíveis que foram coletados. Isso é feito para avaliar quais dados podem ser relevantes para decifrar o problema (CRISP, 1999).
- **3º - Preparação dos dados:** Com os dados definidos, organizados e bem inspecionados, os mesmo deveram ser conduzidos de forma técnica. É preciso preparar todas as databases, definir o formato que será necessário para a análise e ajustar demais questões técnicas(CRISP, 1999).

● **4º - Modelagem:** São selecionadas e aplicadas as técnicas de mineração de dados mais apropriadas, dependendo dos objetivos identificados na primeira fase (CRISP, 1999).

● **5º - Avaliação:** Considerada uma etapa de after-work, mas ainda assim extremamente importante para a vitalidade do ciclo, a quinta fase pede o acompanhamento dos resultados objetivos e a avaliação da aplicabilidade confiável dos insights e conhecimentos obtidos(CRISP, 1999).

● **6º - Desenvolvimento:** O conhecimento obtido por meio do trabalho de mineração e modelagem, deveram ser aplicado de forma prática. O ideal aqui é dar uma entrega mais palpável e aplicável ao cliente a partir das análises dos dados feitas pela equipe (CRISP, 1999).

3.4. Fases do Desenvolvimento

● **Coleta de Dados:** *Foi feito um banco de informações manualmente, sendo inseridos dados em todas as tabelas, essas informações foram retiradas com base nos times reais do campeonato brasileiro sendo elas: Nome de jogadores, nome dos clubes, estádios, locais dos clubes, ações realizadas nas partidas e etc..*

● **Estruturação:** *Segundo (DEL NERO, 2017) foi feito um estudo de como funciona o campeonato brasileiro e suas regras, criamos um banco de dados com todos os atributos conforme a coleta de dados acima.*

● **Análise Exploratória:** Nesta etapa foi feito uma análise exploratória para conhecimento de cada dado coletado (RAMOS, 2015).

● **Modelagem e Avaliação:** Após uma análise exploratória, vamos utilizar métodos de grupos para identificar grupos sendo eles: times e ações dos jogadores.

2.4.1 Coleta de Dados

Os dados foram coletados com base do campeonato brasileiro real do site: <https://globoesporte.globo.com/futebol/brasileirao-serie-a/>, foram retirados nome dos clubes, nome dos jogadores, estádios, cidades e posições dos atletas. Com base nessa coleta, foram inseridos 20 (vinte) clubes, 11 (onze) jogadores para cada clube, 15 (quinze) estádios, 15 (quinze) cidades, e 4 (quatro) tipos de posições dos jogadores.

Após as informações inseridas foi realizado com pesquisa da quantidade de jogos do campeonato sendo 38 jogos para cada clube sendo jogos apena 1 (um) jogo dentro de cada e 1 (um) jogo sendo fora de casa segundo o livro (DEL NERO, 2017). Com base nessas informações foram inseridas os 38 (trinta e oito) jogos para cada clube, obtendo um placar exato para cada jogo.

Com os jogos realizados foram coletadas as ações feitas por cada jogador em uma partida, essas ações foram inseridas manualmente com números relevantes para cada jogador.

3.4.2 Estruturação

Utilizando a ferramenta de banco de dados SQL Server, para montagem do banco em um modelo (Star Schema) com algumas dimensões em (Snow Flake) vamos utilizar os metódos segundo

(DEWSON, 2005), após essa modelagem podemos construir um banco de dados com tabelas entre ligadas sendo de forma mais organizada, utilizando os comandos básicos para construção desses dados (create table, constraints, primary key e foreign key) segundo (EGEA, 2005).

Figura 3 Fonte:

- ***Extração***

Vamos usar a ferramenta do SQL Server para extrair os dados e grava-los em formato txt. Segue as imagens do modo de extração:

Figura 4 Fonte:

Segundo (LEUDENSCHLAGER, 20/06/2017) os passos para exportação são:

1. No SQL Server Management Studio, conecte-se a uma instância do SQL Server Mecanismo de Banco de Dados.
2. Expanda os Bancos de dados.
3. Clique com o botão direito do mouse em um banco de dados.
4. Aponte para Tarefas.
5. Clique em uma das opções a seguir.

3.4.3 Análise Exploratória:

Foi utilizada as selects (SLC 1) abaixo para análise dos seguintes dados: Cartões, Passes e Faltas:

```
Select JP.JogosID as 'ID Jogo', CB.NomeClube as 'Nome do Time', JG.Nome as 'Nome do Jogador',
JP.Gols as 'Gols' , JP.Assistência as 'Assistências', JP.CartaoAmarelo as 'Cartão Amarelo',
JP.CartaoVermelho as 'Cartão Vermelho', JP.Desarmes as 'Desarmes',
JP.PassesCertos as 'Passes Certos', JP.PassesErrados as 'Passes Errados', JP.FaltasCometidas as
'Faltas Cometidas', JP.FaltasRecebidas as 'Faltas Recebidas', JP.DefesasDificeis as 'Defesas
Dificeis'

from Jogadores_Por_Jogo JP join jogadores JG
on JP.JogadoresID = JG.JogadoresID
join Contrato CT
on CT.JogadoresID = JG.JogadoresID
join Clubes CB
on CB.ClubeID = CT.ClubeID
```

Comandos em RStudio (RS 1) para gerar gráfico para análise dos gols:

```
# Carga dos Dados de Jogadores
Dados = read.csv("AcoesJogadores.txt", sep=";", dec=".");
```

- ***Explicação dos dados:***

Gol: Quantidade de Gols de cada jogador por partida;

Assistência: Quantidade de assistência de cada jogador por partida;

Cartão Amarelo: Quantidade de cartões amarelos de cada jogador por partida;

Cartão Vermelho: Quantidade de cartões vermelhos de cada jogador por partida;

Desarmes: Quantidade de desarmes de cada jogador por partida;

Passes Certos: Quantidade de passes certos de cada jogador por partida;

Passes Errados: Quantidade de passes errados de cada jogador por partida;

Faltas Cometidas: Quantidade de faltas cometidas de cada jogador por partida;

Faltas Recebidas: Quantidade de faltas recebidas de cada jogador por partida;

Defesas Difíceis: Quantidade de defesas difíceis de cada goleiro;

- ***Analise de Gols dos Jogadores:***

Conforme tabulação dos dados indicados no item 3.2.1 (Coletas de Dados), as variáveis das seguintes ações: gols, passes certos, faltas cometidas e cartões amarelos foram analisadas para visualização de seus comportamentos. As seguintes análises foram as identificadas como mais relevantes para garantir uma melhor seleção de variáveis:

```
> # Analisar os quantis
> summary(Dados$Gols);
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.1323  0.0000 40.0000
> quantile(Dados$Gols, seq(0,1,by=0.1));
  0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
   0    0    0    0    0    0    0    0    0    0   40
```

Mínimo de Gols por caso: 0.00

25% dos casos são até 0.00 gols por caso.

Mediana está entre 0.00 dos gols por caso.

Média está 0.1323 dos gols por caso.

75% dos casos são até 0.0 gols.

Máximo de gols: 40

Podemos analisar que cerca de 7500 casos estão com 0 (zero) gols, podemos visualizar também que há casos com 40 gols em uma única partida.

Utilizado SLC 1, RS 1 e comandos para análise de gols:

Histograma para ver a quantidade de jogadores que fizeram X Gols

ggplot(Dados, aes(x=Gols)) + geom_histogram(binwidth=1);

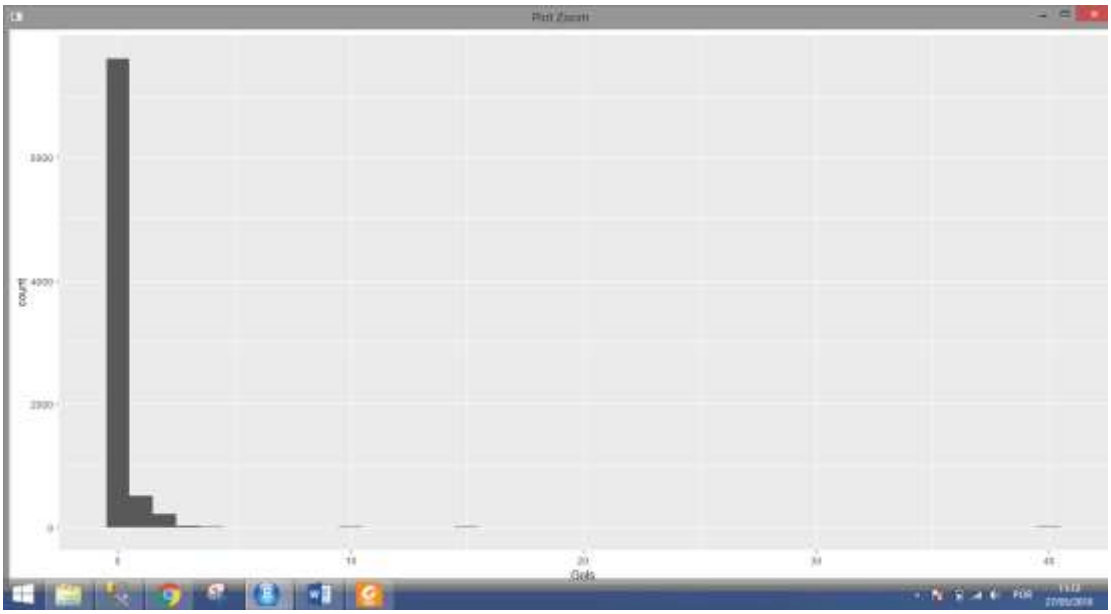

Figura 5 Fonte:

Para uma melhor visualização foi retirado os casos que estão com mais de 6 gols por partida, sendo assim visualizamos que cerca de 500 casos estão com 1 (um) gols, 250 casos estão com 2 (dois) e alguns casos estão entre 3 (três) ou mais gols por partida.

Utilizado SLC 1, RS 1 e comandos para análise de gols:

Gols = subset(Dados,Dados$Gols < 6)

Histograma para ver a quantidade de jogadores que fizeram X Gols

ggplot(Gols, aes(x=Gols)) + geom_histogram(binwidth=1);

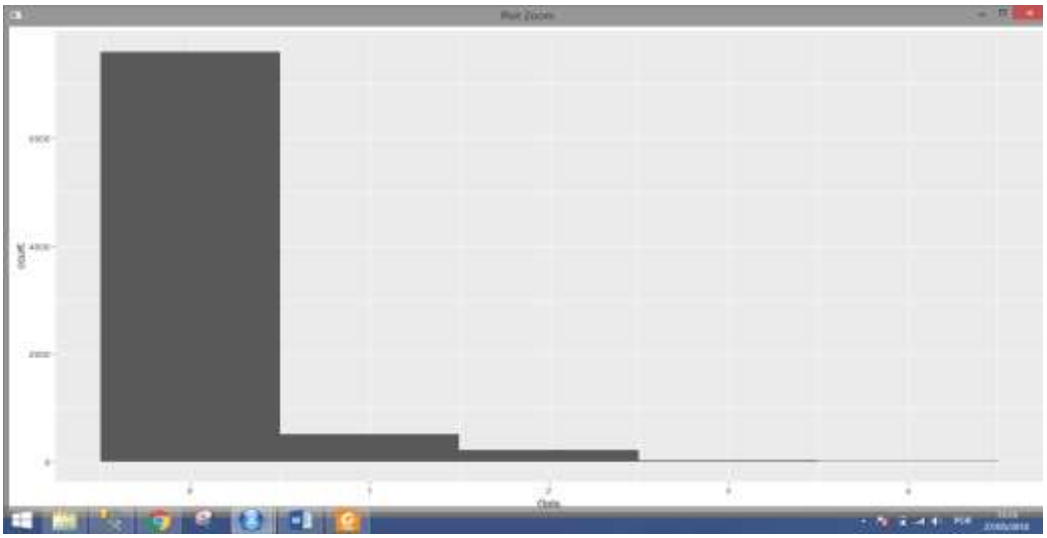

Figura 6 Fonte:

Utilizando gráfico de pontos, podemos observar a maior quantidade está na faixa de 0 (zero), obtendo a maior média de gols por jogo.

Utilizado SLC 1, RS 1 e comandos para análise de gols

```
Gols2 = subset(Dados,Dados$Gols < 6)
# Boxplot para ver onde está a maioria em relação a quantidade de Gols.
ggplot(Gols2, aes(x="", y=Gols)) + geom_boxplot();
```

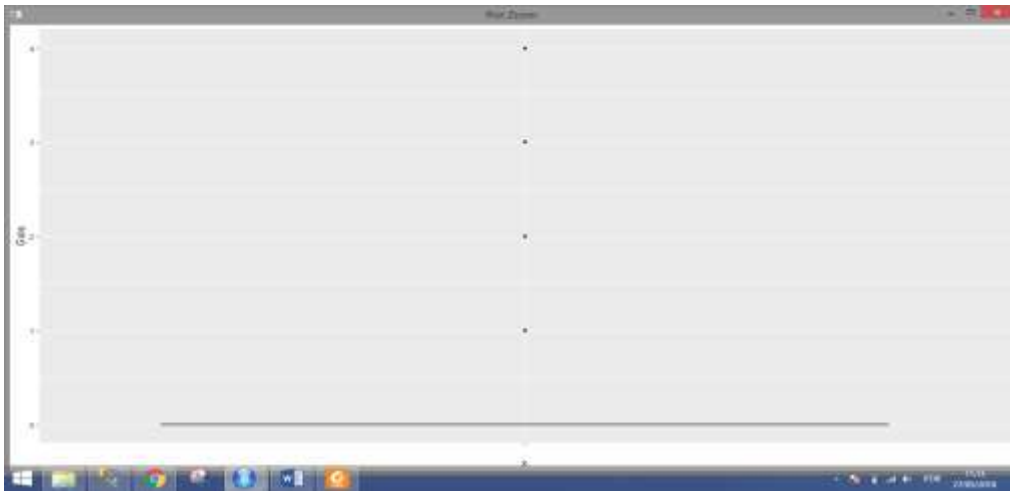

Figura 7 Fonte:

● ***Analise de Faltas Cometidas dos Jogadores:***

```
> summary(Dados$Faltas.Cometidas);
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   1.000   1.283   2.000  38.000
> quantile(Dados$Faltas.Cometidas, seq(0,1,by=0.1));
 0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%

  0    0    1    1    1    1    1    1    2    3   38
```

Mínimo de gols: 0.00

25% dos casos são até 1 faltas cometidas.

Mediana está entre 1 das faltas cometidas.

Média está 1.283 das faltas cometidas.

75% dos casos são até 2 faltas cometidas.

Máximo de faltas cometidas 38 em casos.

Nessa análise podemos análisar alguns pontos sendo eles: 4300 casos sendo a maior média com 2 (duas) faltas por jogo, alguns casos com 40 faltas por jogo, e uma distribuição de 0 a 4 faltas por jogo como mostra o diagráma abaixo.

Utilizado SLC 1, RS 1 e comandos para análise de faltas

Histograma para ver a quantidade de jogadores que fizeram X Faltas.Cometidas

ggplot(Dados, aes(x=Faltas.Cometidas)) + geom_histogram(binwidth=1);

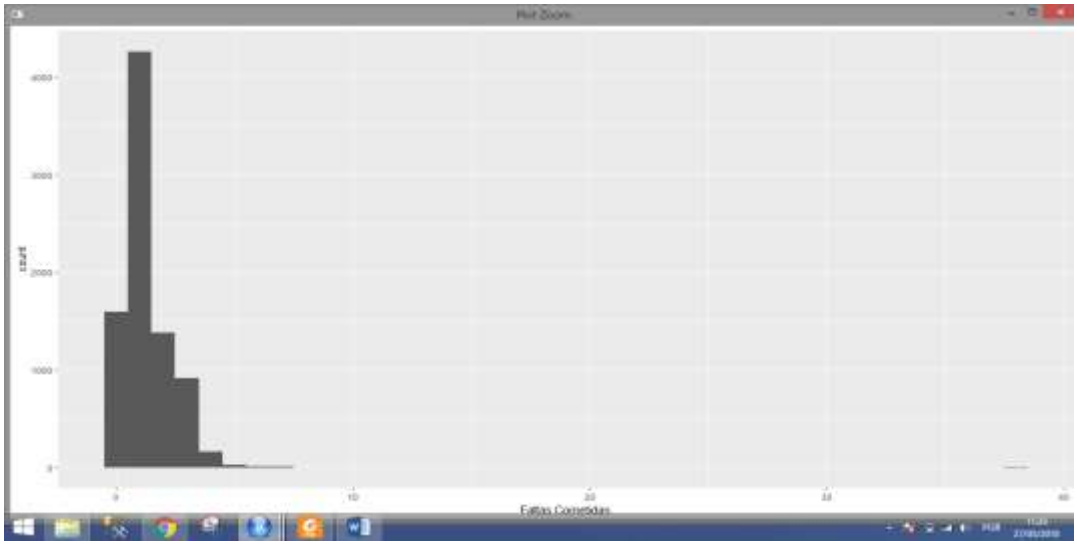

Figura 7 Fonte:

Foi retirado os casos com mais de 15 faltas por jogo, para uma melhor visualização conforme feita a análise acima.

Utilizado SLC 1, RS 1 e comandos para análise de gols

Faltas.Cometidas = subset(Dados,Dados$Faltas.Cometidas < 15)

Histograma para ver a quantidade de jogadores que fizeram X Faltas.Cometidas

ggplot(Faltas.Cometidas, aes(x=Faltas.Cometidas)) + geom_histogram(binwidth=1);

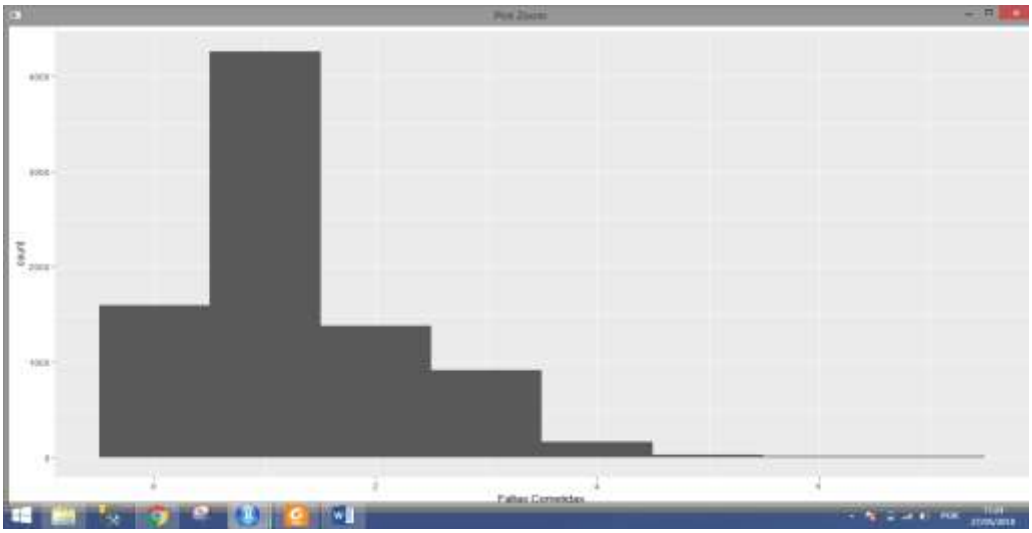

Figura 8 Fonte:

Foi retirado os casos com mais de 15 faltas por jogo, para uma melhor visualização em gráfico de pontos, tendo em vista que o maior numero de faltas está em 1 (uma) falta por jogo, a média está entre gols por partida está entre 1 a 2 faltas.

Utilizado SLC 1, RS 1 e comandos para análise de gols

Faltas.Cometidas2 = subset(Dados,Dados$Faltas.Cometidas < 15)

Boxplot para ver onde está a maioria em relação a quantidade de Faltas.Cometidas. ggplot(Faltas.Cometidas2, aes(x="", y=Faltas.Cometidas)) + geom_boxplot();

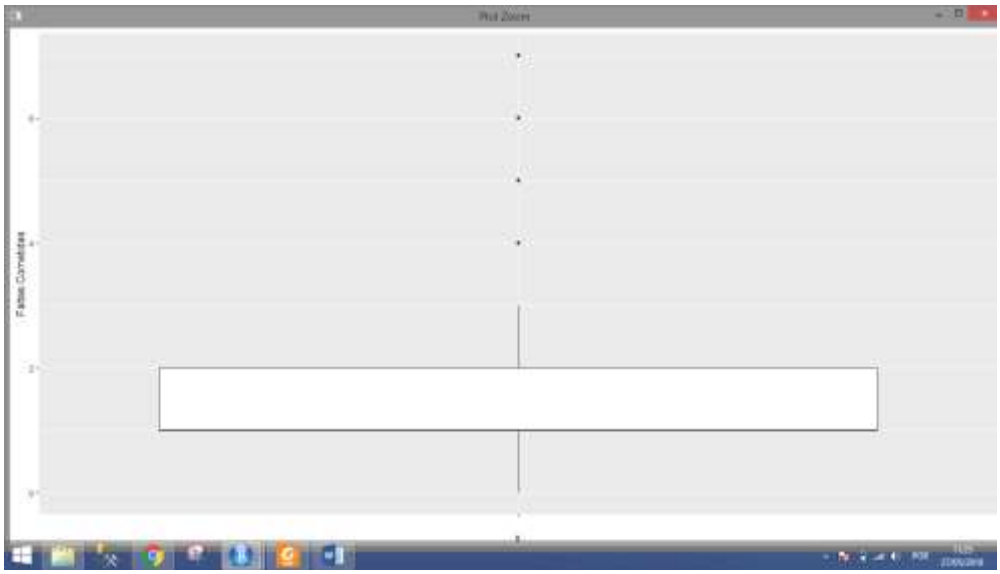

Figura 9 Fonte:

- ***Analise de Passes Certos dos Jogadores:***

  Conforme avaliado a quantidade de passes de cada jogador temos:

```
# Comandos para Analisar os quantis
summary(Dados$Passes.Certos);
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   19.00   28.00   23.55   30.00  200.00
quantile(Dados$Passes.Certos, seq(0,1,by=0.1));
  0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
   0    1   16   19   25   28   28   30   31   35  200
```

Quantis:

Mínimo de passes certos: 0.00

25% dos casos são até 19 passes certos.

Mediana está entre 28 dos passes certos.

Média está 23.55 dos passes certos.

75% dos casos são até 30 passes certos.

Máximo de passes certos: 200.

Podemos observar que gráfico está bem distribuido com dificil visualização de numeros concretos, podemos visualizar a a maior parte dos casos estão entre 1800 passes.

Utilizado SLC 1, RS 1 e comandos para análise de passes

```
# Histograma para ver a quantidade de jogadores que fizeram X Passes.Certos
ggplot(Dados, aes(x=Passes.Certos)) + geom_histogram(binwidth=1);
```

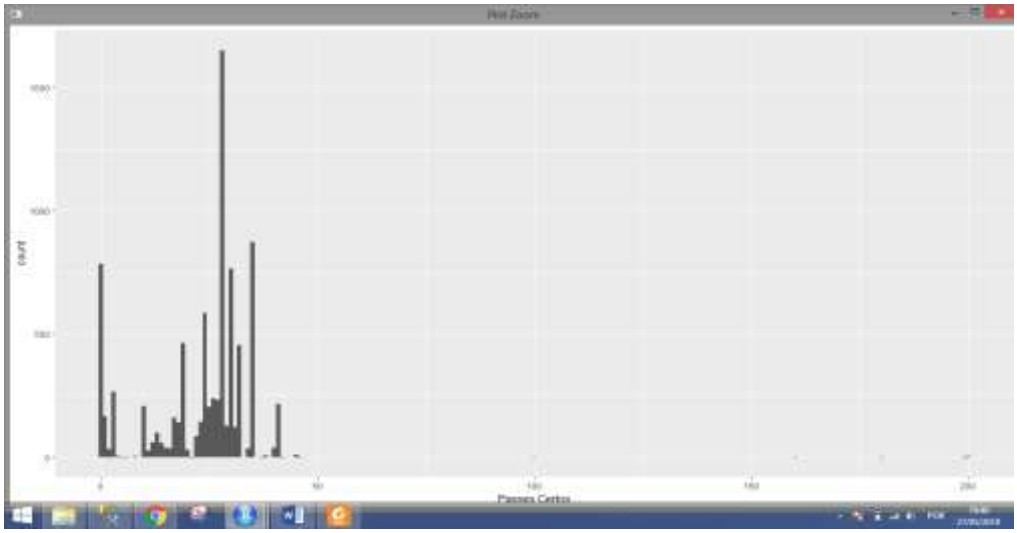

Figura 10 Fonte:

Foram retirados os casos acima de 50 passes certos por jogo para uma melhor visualização e leitura do gráfico, a maior quantidade de casos estão com 28 passes certo por jogo, existem casos com 0 passes por jogo e uma média por jogador está entre 18 a 30 passes por jogo.

Utilizado SLC 1, RS 1 e comandos para análise de passes

Passes = subset(Dados,Dados$Passes.Certos < 50)

Histograma para ver a quantidade de jogadores que fizeram X Passes.Certos

ggplot(Passes, aes(x=Passes.Certos)) + geom_histogram(binwidth=1);

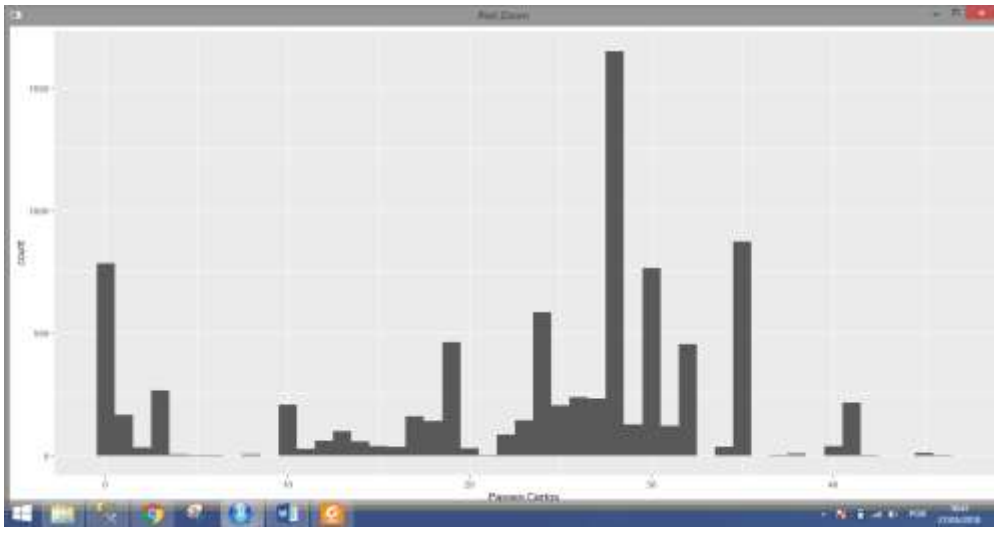

Figura 11 Fonte:

No gráfico abaixo, podemos oberservar a média com a faixa preta está em 28 passes por jogo, reafirmando a média de casos entre 18 a 30 passes por jogo.

Utilizado SLC 1, RS 1 e comandos para análise de gols

Passes2 = subset(Dados,Dados$Passes.Certos < 50) # Boxplot para ver onde está a maioria em relação a quantidade de Passes.Certos. ggplot(Passes2, aes(x="", y=Passes.Certos)) + geom_boxplot();

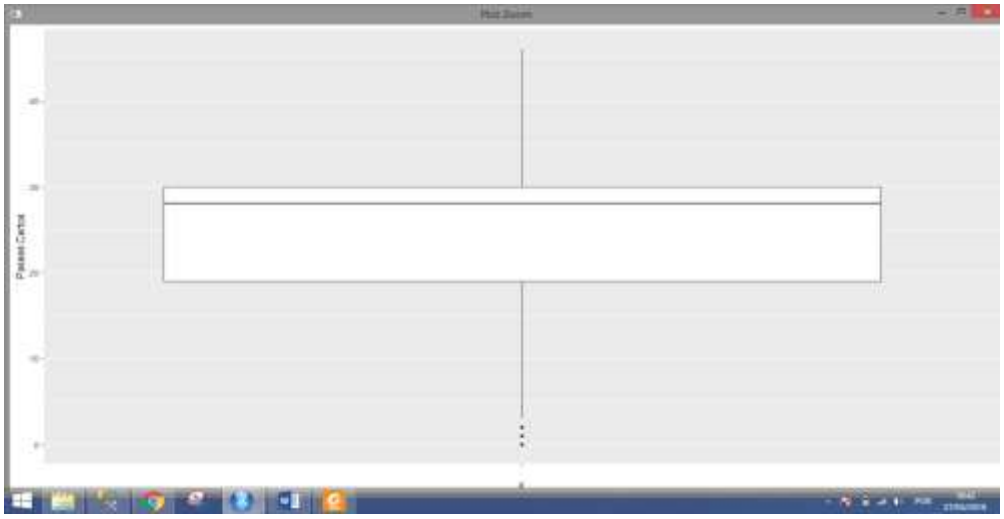

Figura 12 Fonte:

- ***Analise de Cartões Amarelos dos Jogadores:***

```
summary(Dados$Cartão.Amarelo);
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   0.236   0.000  30.000
> quantile(Dados$Cartão.Amarelo, seq(0,1,by=0.1));
  0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
   0    0    0    0    0    0    0    0    1    1   30
```

Mínimo de cartões: 0.00

25% dos casos são até 0 cartões por jogo.

Mediana está entre 0 dos cartões por jogo.

Média está 0.236 dos cartões por jogo.

75% dos casos são até 0.00 cartões por jogo.

Máximo de cartões: 30.

Podemos visualizar no gráfico abaixo que o maior número de casos tem 6300 cartões, alguns casos estão com 11 e 30 cartões.

Utilizado SLC 1, RS 1 e comandos para análise de gols

```
# Histograma para ver a quantidade de jogadores que fizeram X Cartão.Amarelo
ggplot(Dados, aes(x=Cartão.Amarelo)) + geom_histogram(binwidth=1);
```

Figura 13 Fonte:

Para uma melhor visualização foi retirado os casos acima de 6 cartões por jogo, podemos observar que a maior parte dos cartões estão em 0 (zero) por jogo, uma boa parte dos cartões estão em 1 (um) cartão por jogo e alguns casos estão 3 cartões por jogo.

Utilizado SLC 1, RS 1 e comandos para análise de golsCartoes = subset(Dados,Dados$Cartão.Amarelo < 6)

Histograma para ver a quantidade de jogadores que fizeram X Cartão.Amarelo

ggplot(Cartoes, aes(x=Cartão.Amarelo)) + geom_histogram(binwidth=1);

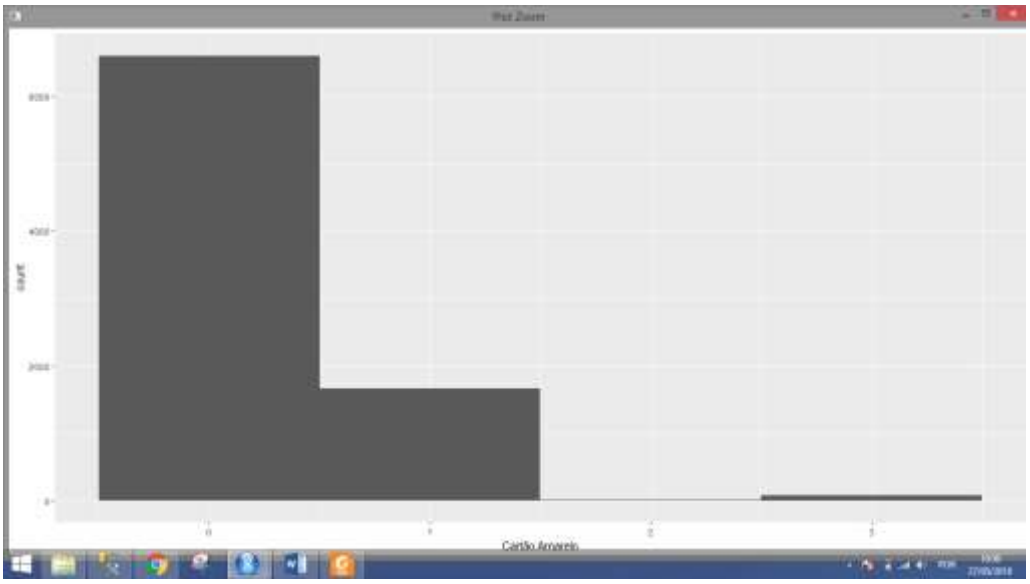

Figura 14 Fonte:

- **Analise dos clubes:**

Nesse diagrama será avaliado Vitórias, empates e derrotas dos clubes somente do Campeonato Brasileiro de 2013:

Segundo (DEL NERO, 2017), cada time deve disputar 19 jogos de ida e 19 jogos de volta.

Para análise dos seguintes dados de vitórias, derrotas e empates.

Foi utilizada a select (SLC 2) abaixo:

```
Select Campeonatos.NomeCampeonato as 'Campeonato', Clube1ID, Jogos.QtdGolsT1,
CLube2ID, Jogos.QtdGolsT2,

case
WHEN QtdGolsT1 > QtdGolsT2 then CL1.NomeClube
when QtdGolsT2 > QtdGolsT1 then CL2.NomeClube
else ''
end as 'Vencedor'
from Jogos join Campeonatos
on Jogos.CampeonatosID = Campeonatos.CampeonatosID
join Clubes CL1
on CL1.ClubeID = Jogos.Clube1ID
join Clubes CL2
on CL2.ClubeID = Jogos.CLube2ID
```

```
where Campeonatos.CampeonatosID = 1;
```

Comandos em RStudio (RS 2) para gerar gráfico para análise dos gols:

```
# Carga dos dados de vitorias
    Dados = read.csv("AnalisesClubes.txt", sep=";", dec=".");
    View(Dados);

#Analises de vitórias distribuídas

table(Dados$Vencedor);

América MG   Atletico MG   Atletico PR        Bahia      Botafogo        Ceará
        10            15            12           10            14           14
    Chapecoense   Corinthians  Cruzeiro       Empates      Flamengo    Fluminense
        10            27            30           76            24           10
    Gremio Internacional     Palmeiras       Paraná       Santos     São Paulo
        10            10            12           23            24           13
    Sport        Vasco       Vitória
        12           12            12

    #Gráfico em Barras
    ggplot(Dados, aes(x=Vencedor)) + geom_histogram(alpha=0.4, col="blue", binwidth=1);
```

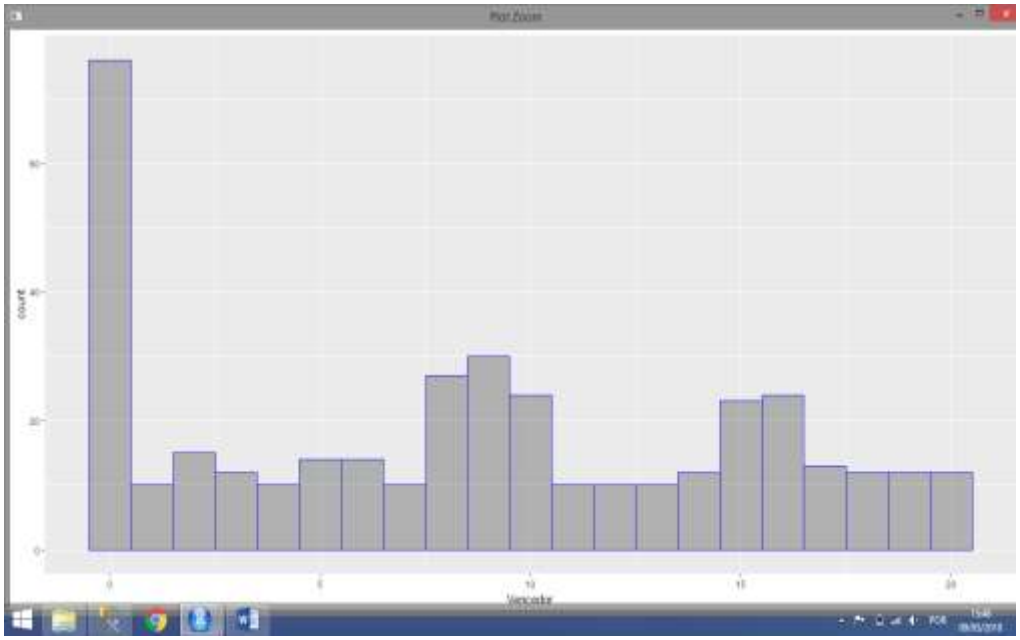

Figura 15 Fonte:

Matrizes:

Foi criado uma matriz para nos trazer o total de vitorias e empates e números de derrotas individuais para cada clube.

Para o número de derrotas individuais a matriz deve ser lida de forma horizontal, para total de empates deve ser lida de forma vertical e para o total de vitórias deve ser lida de forma diagonal.

Foi utilizado nesta etapa os comandos SLC 2 e RS 2.

Matriz de Vitórias Jogando em Casa

 # Tabela de Número de Vitorias por Times em Casa

table(Dados$Clube1ID, Dados$Vencedor);

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	7	6	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	0	1	0	0
2	4	0	8	1	0	0	0	0	1	0	1	0	0	0	0	1	1	0	0	1	1
3	3	0	1	6	0	1	1	0	0	1	1	0	0	0	1	1	1	1	1	0	0
4	4	0	0	0	7	0	0	0	1	1	0	0	1	1	0	0	1	1	0	1	1
5	4	0	1	1	0	9	0	1	0	1	0	0	0	0	1	0	0	1	0	0	0
6	3	0	0	1	1	0	11	1	1	1	0	0	0	0	0	0	0	0	0	0	0
7	3	0	1	0	0	1	1	7	0	0	0	1	0	1	1	0	1	1	0	0	1
8	2	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	1	0	0	0	0
10	3	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	1	0	0	0	0
11	2	0	1	0	0	0	0	0	1	1	1	9	0	1	0	1	0	0	1	1	0
12	7	1	0	0	0	0	0	0	1	0	0	0	7	0	0	0	1	1	0	1	0
13	6	0	1	0	1	0	1	0	0	0	1	0	0	6	0	1	1	0	0	0	1
14	4	0	1	1	0	1	0	0	2	1	1	0	0	0	7	0	0	0	0	1	0
15	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0
16	5	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	12	0	0	0	0
17	3	1	1	1	1	0	0	0	0	1	1	0	1	1	0	1	1	6	0	0	0
18	4	1	0	1	0	1	0	0	0	1	0	0	0	0	0	0	1	1	9	0	0
19	1	0	0	0	0	1	0	0	1	1	1	0	1	0	2	2	1	1	0	7	0
20	6	1	0	0	0	0	0	1	0	1	1	0	0	0	0	1	0	0	0	0	8

Foi utilizado nesta etapa os comandos SLC 2 e Box 2.

Matriz de vitórias jogando fora de casa.

Tabela de Número de Vitorias por Times Fora de Casa

table(Dados$CLube2ID, Dados$Vencedor);

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	2	4	0	0	1	1	0	0	1	1	3	0	0	0	0	2	1	1	0	1	1
2	5	0	7	0	1	0	0	0	1	1	1	0	0	0	0	1	1	0	0	1	0
3	4	0	0	6	1	0	0	1	1	1	1	1	0	0	0	1	1	0	0	0	1
4	4	1	0	1	3	1	0	1	1	1	1	0	0	0	2	1	1	0	0	1	0
5	2	1	0	0	1	5	0	0	1	1	1	1	1	1	1	0	1	1	0	0	1
6	5	0	0	1	0	0	3	0	2	1	0	1	1	0	2	0	0	0	2	1	0
7	4	0	3	1	1	0	1	3	0	0	0	1	1	1	0	1	2	0	0	0	1
8	4	0	0	0	0	0	0	1	0	10	1	1	0	0	1	0	0	0	0	0	1
9	4	0	0	0	0	0	0	0	0	1	12	1	0	1	0	0	0	0	0	0	0
10	5	0	0	0	0	0	1	0	0	0	1	9	0	1	0	1	1	0	0	0	0
11	4	0	1	1	1	1	1	0	1	1	0	1	1	0	1	0	1	1	2	1	0
12	2	2	1	0	0	1	2	1	1	1	1	1	3	1	0	1	0	0	0	0	1
13	6	0	1	1	0	1	1	0	1	1	0	0	0	4	0	0	1	0	1	1	0

14	2	0	2	0	1	0	1	0	3	1	2	1	0	0	5	0	0	0	0	0	0
15	5	0	0	0	0	0	0	1	0	0	1	0	0	0	0	9	0	0	3	0	0
16	4	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	12	0	0	0	0
17	4	1	0	0	0	0	0	0	1	1	0	1	0	0	0	1	1	7	1	0	1
18	3	0	0	0	0	1	3	0	0	3	0	1	0	1	0	3	0	1	3	0	0
19	3	0	0	1	0	1	1	0	0	1	1	0	0	1	0	2	1	1	0	5	1
20	4	1	0	0	0	1	0	2	1	1	1	0	1	0	0	0	1	1	0	1	4

3.5. Algoritmo Cluter com K-Means

Após uma análise completa dos dados acima, serão criados algoritmos de segmentação para unir 2 (dois) casos diferentes afim de localizar grupos, para que possamos identificar os grupos e destacar as posições de cada jogadores nos mesmo.

- **K-means**

K-means indica pontos aleatórios para que deles sejam calculadas as distâncias e detectados os grupos.

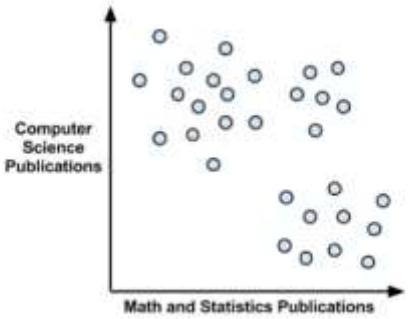

Gráfico 1 Fonte:

Para buscar 3 clusters, 3 pontos aleatórios são definidos.

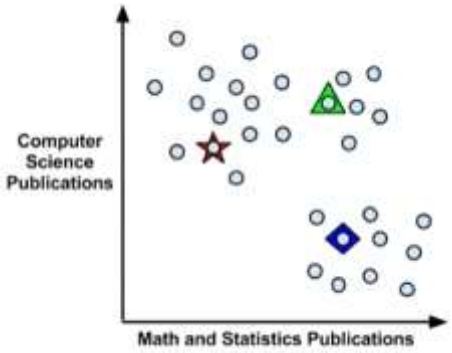

Gráfico 2 Fonte:

Tradicionalmente usa distância Euclidiana para identificar os grupos, mas também podem utilizar Manhattan ou Monkowski.

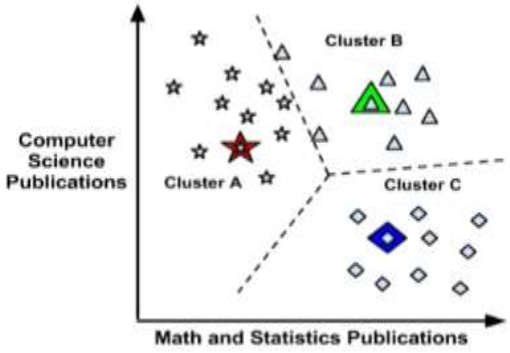

Gráfico 3 Fonte:

Definidos os pontos são atualizadas as posições dos pontos com o conceito de centóides, centralizando assim os pontos sorteados nos dados para poder refinar os elementos de cada cluster

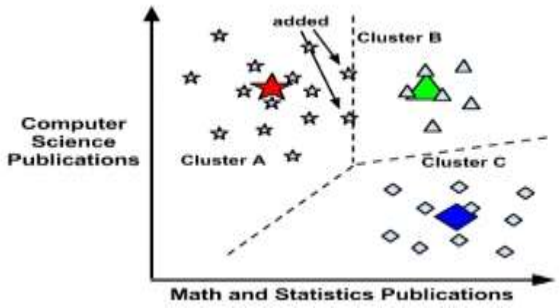

Gráfico 4 Fonte:

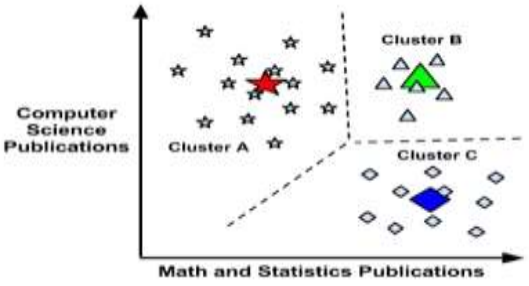

Gráfico 5 Fonte:

- **Normalização e formula**

A primeira normalização discutido aqui é o z-score, é uma estratégia bastante utilizada em estatística. A expressão a seguir representa a transformação dos dados x. s x x z − = Onde x é a

media das amostras, e s é o desvio padrão dos dados. Nesta normalização, a nova variável z têm média zero e variância unitária. Quando a transformação z-score é usada no contexto de clustering, é importante que seja aplicada de uma forma global em todas as observações. Caso não quisermos o centro dos dados em zero, removemos a média das amostras e teremos a seguinte expressão

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

Onde x é a media das amostras, e s é o desvio padrão dos dados. Nesta normalização, a nova variável z têm média zero e variância unitária. Quando a transformação z-score é usada no contexto de clustering, é importante que seja aplicada de uma forma global em todas as observações. Caso não quisermos o centro dos dados em zero, removemos a média das amostras e teremos a seguinte expressão:

$$z = \frac{x}{s}$$

Esta transformação tornará a variância igual a um, as transformações são lineares.

Calculando K-means em RStudio:

```
# Descoberta de 5 grupos usando algoritmo KMEANS
Dados.Algoritmo = DadosCortados[,c(4,9)];
Dados.Modelo5Grupos = kmeans(Dados.Algoritmo, 5);
DadosCortados$Grupo = factor(Dados.Modelo5Grupos$cluster);
```

- **Análise de grupos**

Serão analisados os gráficos por grupos, podendo identificar os grupos de cada posições dos jogadores e as relações das ações dos mesmo.

```
ggplot(DadosCortados, aes(x=Passes.Certos, y=Gols, col=Grupo)) + geom_jitter();
```

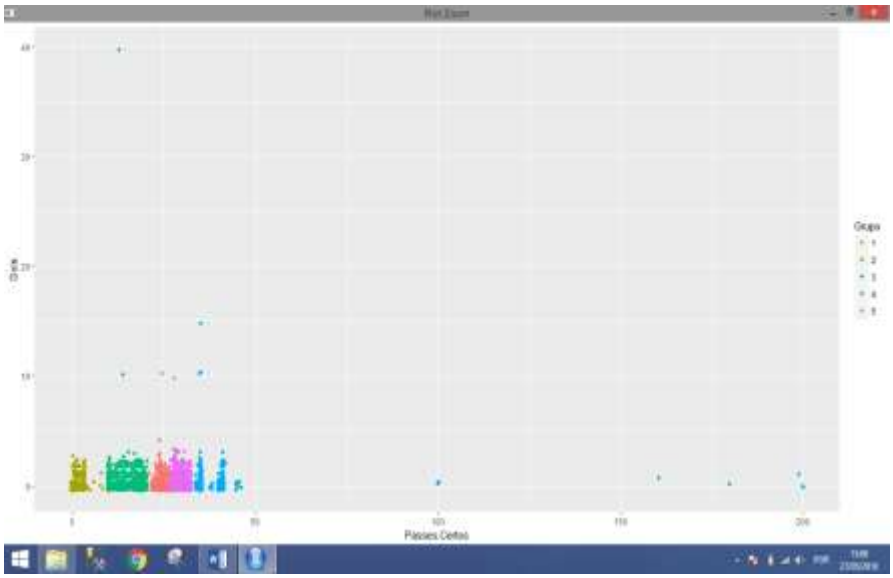

Figura 16 Fonte:

```
# Descoberta de 5 grupos usando algoritmo KMEANS
Dados.Algoritmo = DadosCortados[,c(4,9)];
Dados.Modelo5Grupos = kmeans(Dados.Algoritmo, 5);
DadosCortados$Grupo = factor(Dados.Modelo5Grupos$cluster);
ggplot(DadosCortados, aes(x=Cartão.Amarelo, y=Faltas.Cometidas, col=Grupo)) +
geom_jitter();
```

4.CONCLUSÕES

Concluímos que a coleta de dados foi essencial para para o trabalho como um todo, a importância desses dados nos gerou uma construção de um banco Data *Warehouse*.

Mais antes disso o mais importante foi estudar uma metodologia que nos levou a coletar os dados, saber usá-los, como usar analise exploratória, diferencial os, e por fim lê-los e analisá-los de forma correta. Com esse estudo dessa metodologia, podemos afirmar que o trabalho ficou mais dinâmico e fácil de fazer, pois seguindo as regras conseguimos manusear os dados para ter uma ótima análise e leitura dos mesmos.

Após a construção do DW, foram analisados todos tipos de dados para extração tendo em vista uma relação futura dos mesmos para uma leitura mais concisa e bem desenvolvida.

As ferramentas que foram utilizadas como SQL Server para construção do DW e RStudio para análise desses dados foram uteis e favorável para o projeto, SQL Server nos ajudou a ter uma visão de diagrama, fácil de leitura, códigos mais simples e fáceis de utilizar. Já o RStudio também foi muito importante para análise exploratória dos dados, tendo uma visão de gráficos, matrizes e formulas de média, mediana e moda dos dados, nessas análises foram feitos diversos gráficos diferentes para chegar em resultados diferentes, porém, sendo valiosos para o projeto! Sem essas diversas formas de resultados não iriamos conseguir algo conciso nos dados.

Concluímos que a extração foi importante para análise futuras, foi extraído os seguintes dados para análise dos times: times, jogos, vitorias e derrotas e para análise dos jogadores foram extraídos os dados das ações como: gols, assistências, cartões, desarmes, passes, faltas e defesas. Com essas informações extraídas foram feitas as análises dentro da ferramenta RStudio relacionando-as e fazendo análise exploratória das mesmas.

Em nossa primeira análise, analisamos a média de gols em uma competição,

nossa informação gerou uma média de 0.1323 que reflete a média de gols de uma competição na atualidade, feito isso fizemos análise exploratória para poder enxugar os dados e acharmos pontos importantes como números mais exatos e médias mais concisas para visualização, esse tipo de análise foi feito para outras ações: faltas, assistência, passes e cartões, junto a isso foi feito os quantis de cada um deles sendo média, mediana, mínimo e máximo, pois resultam na importância da análise. Já para análise de times foi feito análise em matrizes onde tivemos uma melhor avalição de vitorias, derrotas e empates. Podendo identificar também, alguns times que jogaram duas vezes ou mais fora de casa e obtiveram dois ou mais jogos dentro de casa, onde isso não é comum em uma competição.

Foram feita uma análise de grupos utilizando a metodologia de algoritmos K-means, nessa análise juntamos duas ações que se completam para um time chegar ao um gol em uma partida, sendo elas: gols e passes, com essa junção podemos identificar os grupos de jogadores que mais podem chegar a fazer um gol numa partida que estão nas posições de atacantes e meio-campo,

também foram identificados jogadores que dificilmente fazem gols que estão nas posições: zagueiros, laterais e goleiros.

Para trabalhos futuros podem ser coletados mais dados de ações como: Tempo do jogador dentro de campo, posse de bola de cada time por partida, como o time fazem seus gols sendo bola parada, bola aérea, sendo no primeiro tempo ou segundo tempo entre outras! Com esses dados podemos fazer análises de quais jogadores correm mais dentro de campo, quais times tem mais posse de bola e contra quais times e podendo ver o poder que o time tem em fazer gols sendo por bola aérea ou parada. Podendo utilizar jogadores reservas e fazer análises de quais jogadores reservas se descatam nas partidas, para quando um jogador se contundir ou estiver suspenso o técnico saber qual jogador vai colocar no lugar dele.

Podemos construir um banco com mais campeonatos para tornar o banco mais robusto e fazer comparações de como cada time joga em campeonatos diferentes.

Utilizando o K-means podem ser criados grupos de como os times chegam a uma vitória cruzando duas ou mais ações, pois podem ser feitas análises em 4D, utilizando 4 ações diferentes e criando varios tipos de grupos para o projeto.

5.REFERÊNCIAS BIBLIOGRÁFICAS

DEWSON, Robin. **SQL SERVER 2005** Para Desenvolvedores. Rio de Janeiro ED: Alta Books 2005

PORTER, Michael E. **Vantagem Competitiva**. 21. ed. Rio de Janeiro: Campus, 1989.

EGEA, Miguel. SQL SERVER 2005 Fundamentos de Banco de Dados. São Paulo, ED: Bookman 2005.

BRAUN, W.J. & Murdoch, D.J. (2007) A first course in statistical programming with R. Cambridge University Press, Cambridge.

WU, Junjie Wu. ***Advances in K-means Clustering***: A Data Mining Thinking. Ed: Pai2014

ZIKOPOULOS, Paul. "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data" ED: Osbome 2011

RAMOS, Raniere Ramos, **Qual a importância da Análise Exploratória de Dados?,** 08/09/2015, Disponível em: <https://oestatistico.com.br/analise-exploratoria-de-dados/>

DEL NERO, Marco Polo Del Nero, 2017, **Regras de Futebol**, 06º Ed: Rio de Janeiro.

http://www.bigdatabusiness.com.br/se-voce-se-interessa-por-big-data-precisa-entender-o-crisp-dm/ Acesso em:

LEUDENSCHLAGER, Douglas Leudenschlager, **Iniciar o assistente de importação e exportação do SQL Server**, 20/06/2017, Disponível em: <https://docs.microsoft.com/pt-br/sql/integration-services/import-export-data/start-the-sql-server-import-and-export-wizard?view=sql-server-2017> Acesso em:

LEWIS, Michael Lewis. Moneyball **O Homem que Mudou o Jogo**. São Paulo, 1º ED: Bookman 2015.

FERNANDES SR. Trabalho informatizado e distúrbios psicoemocionais: estudo seccional em três empresas de processamento de dados em Salvador, Bahia [**Dissertação de Mestrado**] Salvador: Faculdade de Medicina da Universidade Federal da Bahia; 1992.

AMO, Sandra de. **Técnicas de Mineração de Dados**. Disponível em. Última visita em 24 de novembro de 2010.

GRECO, Cesar Greco, **Média de gols é a maior das últimas 10 aberturas**, CBF, 14/05/2017, Disponível em: <https://www.cbf.com.br/noticias/campeonato-brasileiro-serie-a/media-de-gols-e-a-maior-das-ultimas-10-aberturas-1#.WvWeIqinHIU> Acesso em:

GACIBA, Leonardo Gaciba, **Largada Fair Play**, 11/05/2014, Disponível em: < http://sportv.globo.com/platb/blog-do-gaciba/tag/faltas/> Acesso em:

OLIVEIRA, Marco Oliveira, **O jogador mais advertido no brasileirão em cartões toma 1 a cada 2 jogos**, 13/07/2017, Disponível em: https://www.foxsports.com.br/news/313145-esta-na-degola-o-jogador-mais-advertido-no-brasileirao-em-cartoes-toma-1-a-cada-2-jogos. Acesso em:

RODRIGUES, Rodolfo Rodrigues, **Brasileirão tem maior média de amarelos do mundo,** 26/05/2015, Disponível em: <https://futebolemnumeros.blogosfera.uol.com.br/2015/05/26/brasileirao-tem-maior-media-de-amarelos-do-mundo/> Acesso em:

REIS, Rafael Reis, **Neymar sofre recorde de faltas e é o jogador que mais apanha na Europa**, 30/03/2017, Disponível em: <https://blogdorafaelreis.blogosfera.uol.com.br/2017/03/30/neymar-sofre-recorde-de-faltas-e-e-o-jogador-que-mais-apanha-na-europa/> Acesso em: