**CENTRO UNIVERSITÁRIO DE BELO HORIZONTE**
**Graduação - Ciência da Computação**

JOSUÉ FILIPE PONTES NERY - 119122537
MARCELA CRISTYNE - 11827313
THAIS CORDEIRO PEREIRA - 11821085

# UC: ANÁLISE DE DADOS E BIG DATA

Prática 06 - Estatística Descritiva

**Belo Horizonte**
**2º Semestre de 2021**

**Objetivo:** obter estimativas pontuais e por intervalos de 95% de confiança. Vocês deverão obter os dados dos times **campeões que disputaram a série A do campeonato** Brasileiro de 2003 a 2020. Os dados de interesse são: pontuação final, número de gols marcados, número de gols sofridos e o total de Km 's viajados. Este último parâmetro pode ser um valor aproximado.
Usem o RStudio para os itens abaixo. Não esqueçam de **interpretar os resultados.**

1) Descrever a fonte para a coleta dos dados.
Foi usado como referência os dados encontrados nos sites Transfermarkt, RSSSF Brazil e Wikipédia.

---


TRANSFERMARKT GMBH & CO. KG. **TRANSFERMARKT**. [S. l.], entre [2000 e 2021]. Disponível em: https://www.transfermarkt.com.br/serie-a/startseite/wettbewerb/BRA1. Acesso em: 22 out. 2021.

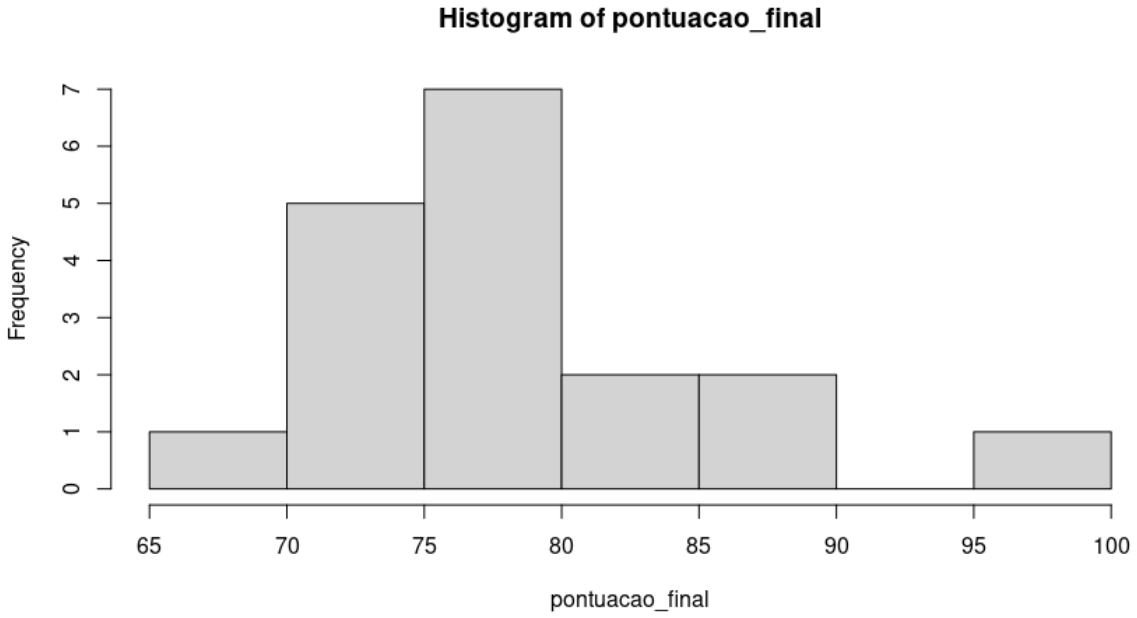
DIOGO, Julio Bovi; REC.SPORT.SOCCER STATISTICS FOUNDATION; RSSSF BRAZIL. **Brazil - List of Champions.** [S. l.], entre [1997 e 2021]. Disponível em: https://rsssfbrasil.com/tablesae/brcamp.htm. Acesso em: 22 out. 2021.

CAMPEONATO BRASILEIRO DE FUTEBOL DE 2003 - SÉRIE A. In: WIKIPÉDIA, a enciclopédia livre. Flórida: Wikimedia Foundation, 2021. Disponível em: <https://pt.wikipedia.org/w/index.php?title=Campeonato_Brasileiro_de_Futebol_de_2003_-_S%C3%A9rie_A&oldid=60563887>. Acesso em: 22 out. 2021

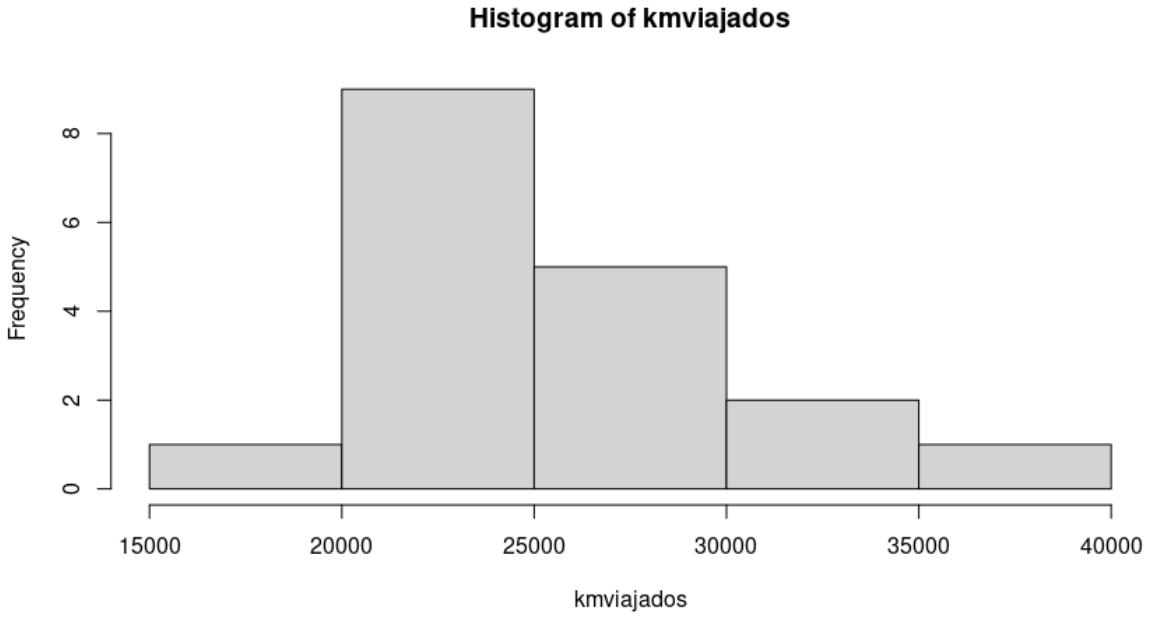

---

* Para fazer o cálculo dos Km's viajados foi utilizado o endereço eletrônico a seguir, https://pt.distance.to/ .
** Consideramos a distância em quilômetro (km), usando a menor distância encontrada entre os dois pontos (linha aérea), **valores aproximados**.
*** Identificamos a cidade e estado do time da casa, fazendo a comparação com o time visitante e considerando ida e volta. (Ex.:. Entre Belo Horizonte e São Paulo temos 489,56 km (linha aérea), considerando ida e volta fica 979,12 km)

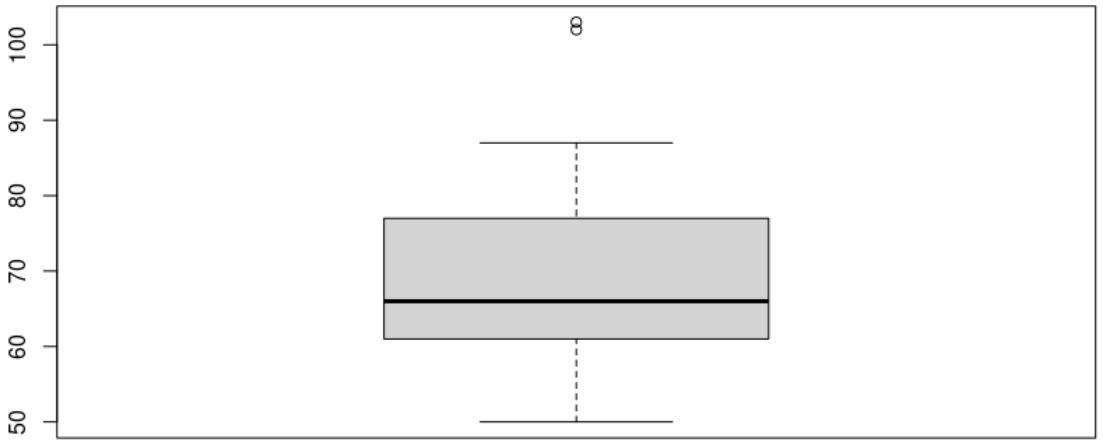
2) Construir histogramas para: ’pontuação’ e ’total de Km' s viajados’.
Entre os campeões do Brasileirão Série A, de 2003 a 2020, as pontuações mais frequentes estão entre 75 e 80, com um total de 7 componentes. Não foram encontrados elementos no intervalo entre 90 e 95, temos um outlier entre 95 e 100.



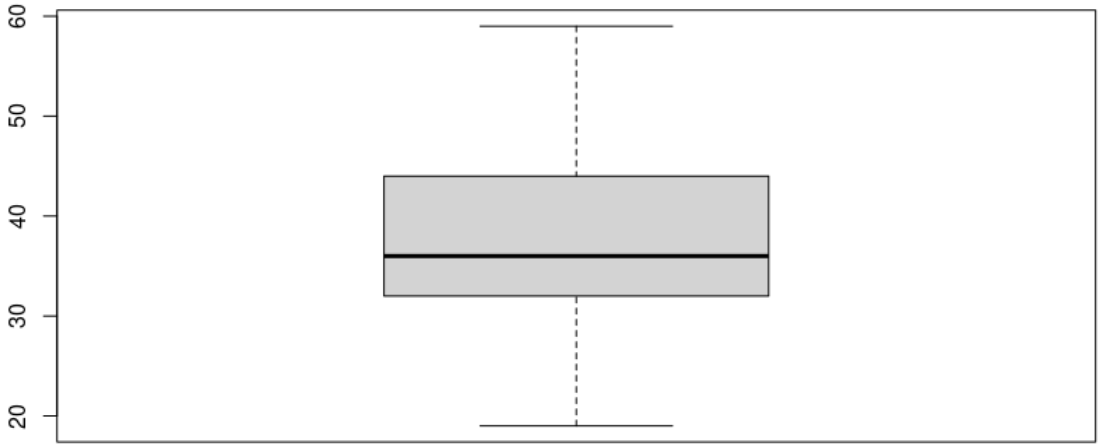

Já em relação aos quilômetros viajados temos uma maior frequência entre elementos que viajaram de 20000 km a 25000 km, não apresentando outlier.



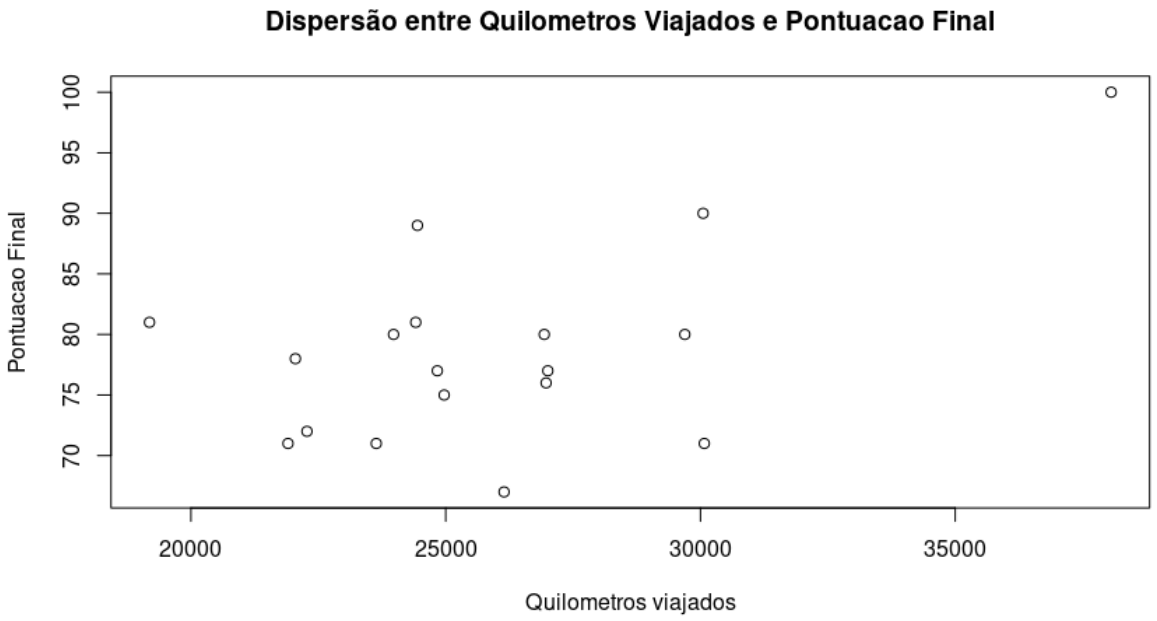
3) Construir box-plots para 'número de gols marcados' e 'número de gols sofridos'.
Número de gols marcados -
Valor mínimo aproximadamente cinquenta e valor máximo entre 80 e 90, podemos observar outlier próximo de 100 gols.

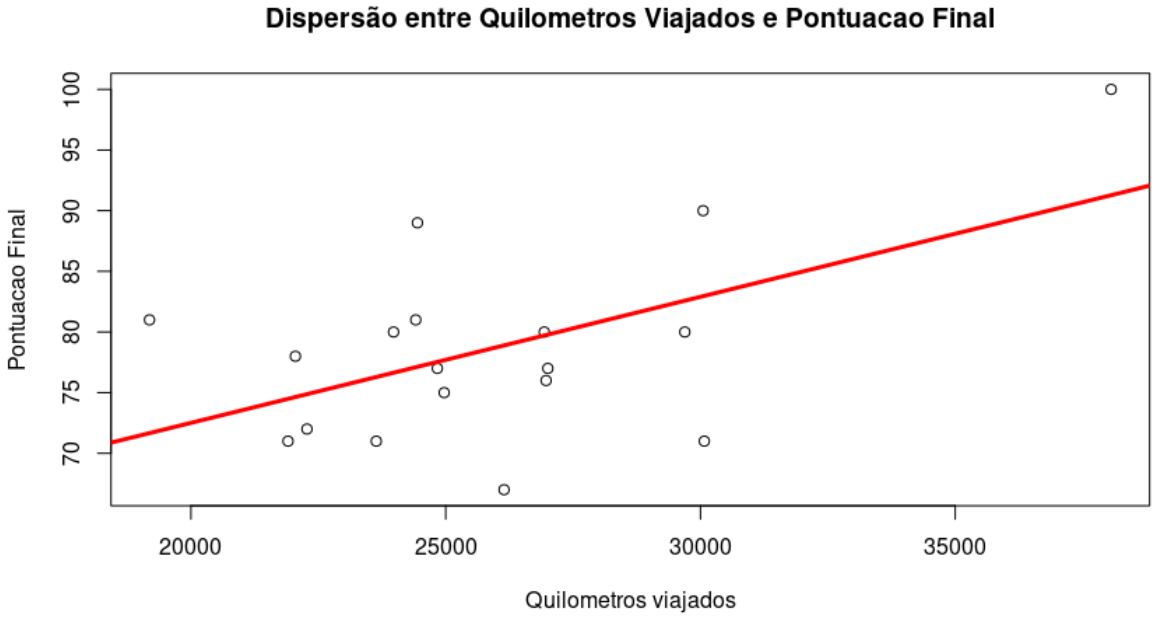


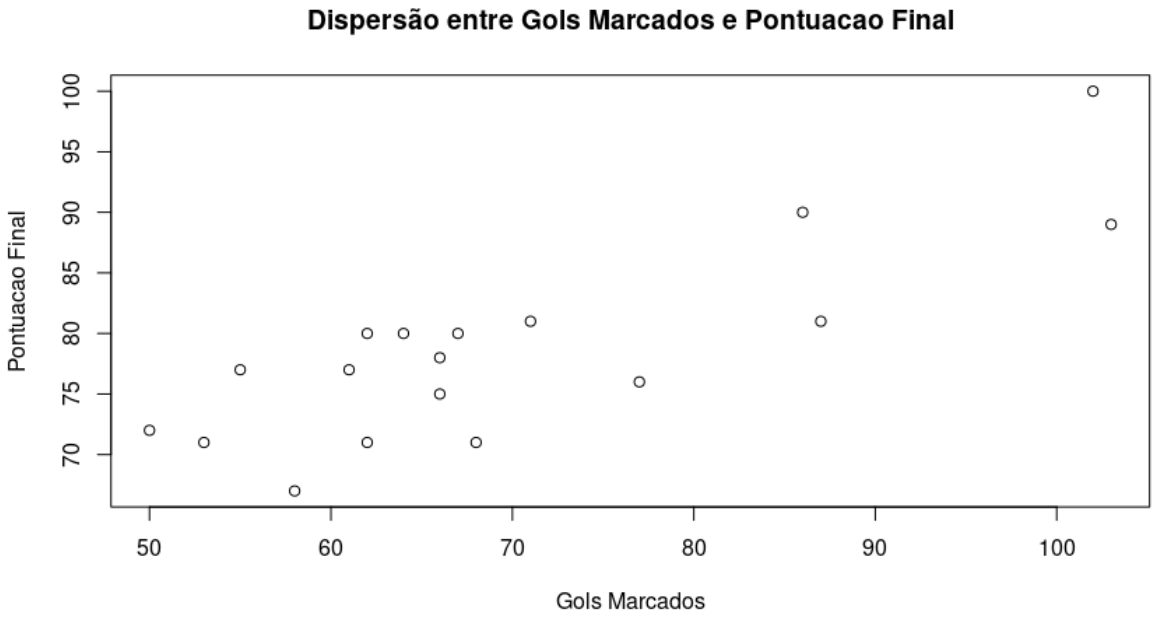

Número de gols sofridos-
Valor mínimo aproximadamente 20 e valor máximo próximo de 60, sem a presença de outlier.

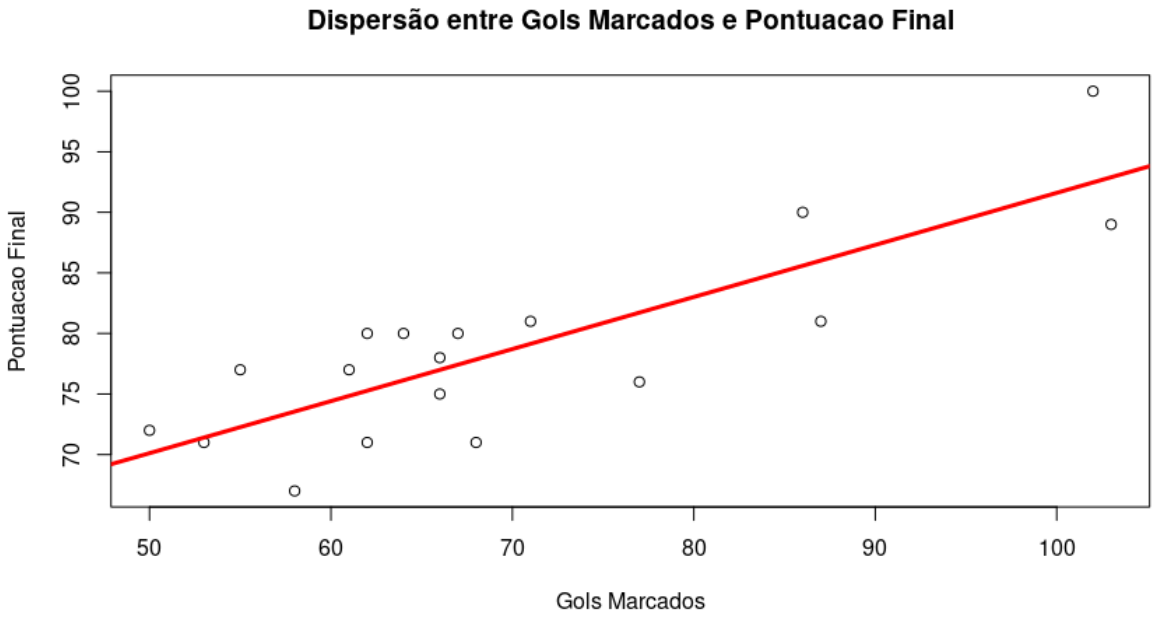


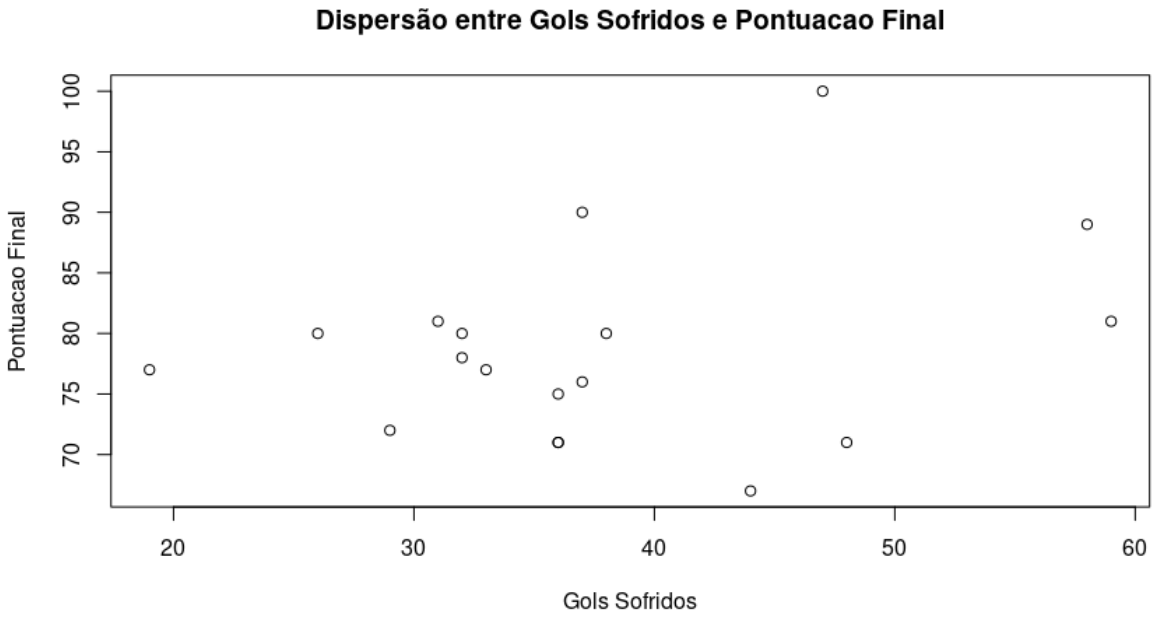
4) Construir gráficos de dispersão para analisar o ’total de Km’s viajados’ vs a ’pontuação’, ’número de gols marcados’ vs a ’pontuação’, e ’número de gols sofridos’ vs a ’pontuação’.

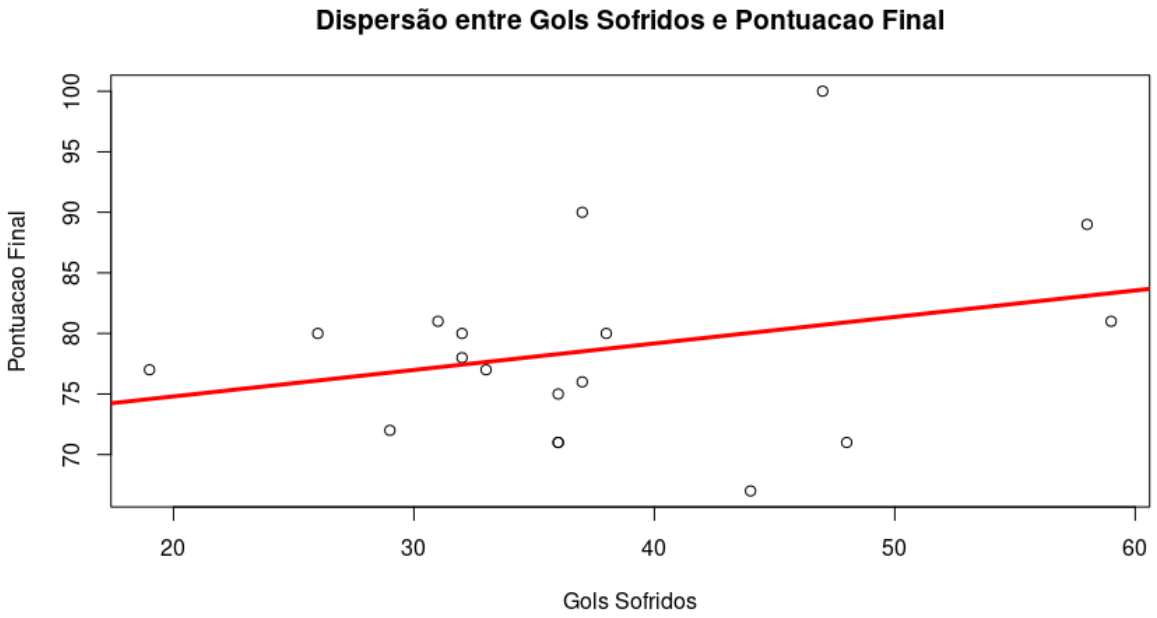



Dispersão entre Gols Marcados e Pontuacao Final
Pontuacao Final
70
75
80
85
90
95
100
50
60
70
80
90
100
Gols Marcados


Dispersão entre Gols Marcados e Pontuacao Final
Pontuacao Final
70
75
80
85
90
95
100
50
60
70
80
90
100
Gols Marcados


Dispersão entre Gols Sofridos e Pontuacao Final
Pontuacao Final
70
75
80
85
90
95
100
20
30
40
50
60
Gols Sofridos


Dispersão entre Gols Sofridos e Pontuacao Final
Pontuacao Final
70
75
80
85
90
95
100
20
30
40
50
60
Gols Sofridos

5) Obter estimativas por intervalos de 95% de confiança para cada um dos parâmetros analisados.

| | **limite inferior** | **limite superior** |
|---|---|---|
| pontuacao_final: | 74.68471 | 82.64863 |
| total_golsM: | 62.20783 | 77.56994 |
| total_golsS: | 32.54677 | 42.78656 |
| kmviajados: | 23819.15 | 28028.68 |

6) O que é possível concluir com toda a análise realizada dos itens anteriores?

É importante manter em mente que a base de dados usada é bem pequena, com apenas 18 elementos e cinco variáveis, sendo elas:

**[clube] -** variável categórica tipo character, consta o nome do clube.
**[ano] -** variável contínua, ano do campeonato, tipo inteiro (Ex.: 2003)
**[pontuacao_final] -** variável discreta tipo numeric, pontuação final do respectivo campeão.
**[total_golsM] -** variável discreta tipo numeric, total de gols marcados;
**[total_golsS] -** variável discreta tipo numeric, total de gols sofridos;
**[kmviajados] -** variável discreta tipo numeric, total de quilômetros viajados.

O modelo calculou a possibilidade, com 95% de confiança, dos próximos campeões pontuarem entre aproximadamente 74.68 e 82.64 (pontuação final), terminarem o campeonato entre aproximadamente 62.20 e 77.56 gols marcados, 32.54 e 43.78 gols sofridos e 23819.15 e 28028.68 quilômetros viajados.

Sobre os dados utilizado, foi observado através do histograma de pontuação final e quilômetros viajados, uma certa dispersão entre os dados, porém o total de elementos entre 70 e 80 pontos é maior que a soma de todos os outros, assim como a quantidade de elementos entre 20000 e 30000 quilômetros viajados é superior que a soma de todos os outros.

Os outliers foram observados tanto no histograma de pontuação final quanto no `boxplot de gols marcados e gráficos de dispersão. Podemos observar uma boa diferença entre os boxplot de gols marcados e gols sofridos, de forma que enquanto o maior valor para gols sofridos é aproximadamente 60 (gols) e sua caixa da amplitude interquartílica se localiza entre 30 e 50, o menor valor para gols marcados

é aproximadamente 50 (gols) e sua caixa da amplitude interquartílica se localiza entre 60 e 80.

Dentre os gráficos de dispersão vamos começar com os de quilômetros viajados vs pontuação final. Temos elementos um pouco dispersos com a existência de um outlier, e junto com a linha de tendência temos uma possível correlação, linear positiva, entre as variáveis analisadas.

Já na relação entre gols marcados vs pontuação final, encontramos elementos menos dispersos e rentes à linha de tendência. Também temos uma possível correlação, onde, quanto maior a quantidade de gols marcados, maior a pontuação final.

O mesmo pode ser observado no gráfico de gols sofridos vs pontuação final, porém, com a presença de outliers e elementos mais dispersos.  A linha de tendência também é crescente, mas a sua posição, quase que horizontal, demonstra a existência de uma possível relação entre as variáveis, porém uma relação mais fraca.