Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Análise, seleção e agrupamento de características relevantes das equipes da NFL durante os anos de 2013-2016 e suas relações - um estudo de caso.

Mariana Oliveira Araujo

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Análise, seleção e agrupamento de características relevantes das equipes da NFL durante os anos de 2013-2016 e suas relações - um estudo de caso.

Mariana Oliveira Araujo
Orientador(a): Maria Silvia de Assis Moura

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Outubro de 2022

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

Analysis, selection and clustering of relevant characteristics from NFL teams during the years of 2013-2016 and their correlations - a case study.

Mariana Oliveira Araujo
Advisor: Maria Silvia de Assis Moura

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
September 2022

Mariana Oliveira Araujo

Análise, seleção e agrupamento de características relevantes das equipes da NFL durante os anos de 2013-2016 e suas relações - um estudo de caso.

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Mariana Oliveira Araujo e aprovado pela banca examinadora.

Aprovado em dia de mês de ano

Banca Examinadora:

- Profa. Dra. Maria Sílvia de Assis Moura
- Prof. Dr. José Carlos Fogo
- Prof. Dr. Danilo Lourenço Lopes

Aos meus familiares e amigos por todo suporte.

Agradeço primeiro a Deus, pois foi Ele que me proporcionou a vida e sem Ele nada seria possível.

À minha família que me proporcionou todos os meios para eu conseguir chegar a este momento e me deu suporte em todos os momentos difíceis. Em especial ao meu irmão Vinicius por ter me compartilhado comigo anos incríveis.

À minha orientadora Professora Maria Silvia, por todos os conselhos acadêmicos e de vida. Além de me incentivar a realizar um trabalho sobre um assunto de muita importância para mim. E à banca avaliadora por aceitarem participar e darem contribuições para a melhora desse trabalho.

Aos meus amigos que ajudaram a deixar a vida universitária mais leve e equilibrada. Em especial à minha amiga Gabriele, minha estrelinha que ainda olha por mim e foi minha força para terminar o que ela não pode. À Ana Luiza por ter me escutado, me dado apoio e estado do meu lado nos momentos mais difíceis. À Daniele, a melhor veterana, por ter me ensinado muito sobre a vida e me mostrado que não estou sozinha. E à Lais Merllo, por ter me dado apoio principalmente no início dessa jornada.

De forma especial, à Adriana Azevedo, Cláudia Barbutti e Cláudia Penteado por terem me dado conhecimento e força espiritual durante minha evolução e conhecimento sobre a vida.

"Espere o melhor, prepare-se para o pior e aceite o que vier."

(CPM 22)

Resumo

Neste projeto de Trabalho de Conclusão de Curso propusemos o estudo de variáveis referentes a times da ***NFL*** (*National Football League*) com o objetivo de analisar, selecionar e agrupar as características, e também as relações entre estas. Dados disponibilizados pelo site oficial da ***NFL***, foram agrupados e disponibilizados em arquivos de dados no site *Kaggle*, de acordo com posições e características dos jogadores. Métodos de estatística multivariada são apresentados e então feita uma análise de dados provenientes de informações de um período de quatro anos dos quais serão analisados os resultados dos jogos de todos os times durante a temporada regular. Foram utilizados os métodos de Análise Componentes Principais e Análise de Agrupamentos, para selecionar variáveis pertinentes ao objetivo do estudo e agrupar o restante para uma melhor análise, respectivamente.

Palavras Chave: *Classificação; Análise de Componentes Principais; Análise de Agrupamentos;* ***NFL***..

Abstract

On this work, the study of different ***NFL*** *(National Football League) variables was proposed, with the main goal of analysing, selecting and clustering their characteristics and its correlations. Data on the official* ***NFL*** *website was grouped and made available on file datas on the website Kaggle, by the players positions and characteristics. Multivariate Statistics data methods are presented, and then an analysis of data from a four year period is presented, in order to analyze the results of all teams during a regular season. The Principal Components Analysis and Cluster Analysis methods were used to select and group variables for a better analysis.*

Keywords: *Classification; Principal component analysis; Cluster Analysis; NFL..*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A ***National Football League (NFL)*** é a principal liga de Futebol Americano no mundo, sendo umas das quatro maiores ligas de esporte dos **EUA** (**NBA, NFL, NHL, MLB**); deixando de ser apenas um esporte para virar um *show business*, o qual movimenta bilhões de dólares em *merchandisings*, salários, ingressos, etc (JovemPan (2017)). Ao contrário do que a maioria das pessoas pensa, o futebol americano não consiste apenas em um esporte de força bruta, mas também é um jogo tático complexo, oriundo de estratégias militares, com o objetivo de despistar o oponente, assim avançando no campo e marcando pontos.

Para a proteção dos atletas a ***NFL*** faz constantes análises e atualizações em suas regras para prevenir lesões graves, principalmente na região da cabeça, para tentar prolongar a vida do atleta no esporte e fora dele. De alguns anos pra cá, se fala muito sobre a parte mental do jogo, o que motivou a ***NFL*** a focar em alguns projetos e incentivar os jogadores a cuidarem de sua saúde mental, por se tratar de um jogo muito intenso, com altos valores monetários envolvidos e os jogadores terem em sua maioria carreiras curtas dentro do esporte (Dorfman (2021)).

A temporada regular, da qual foram obtidos os dados para esse estudo, consiste de 32 times separados em duas conferências, ***NFC*** e ***AFC*** (*National e American Football Conference*); cada uma das ligas conta com 16 times, alocados em quatro divisões igualitárias (*North, South, East* e $West$), como mostra na Figura 1.1. Dentro das divisões os times jogam duas vezes contra cada time, totalizando seis jogos de temporada regular cada, não contando os jogos de *playoffs*/finais, tendo também mais dez jogos contra outros times presentes nas conferências, estipulados pela ***NFL***.

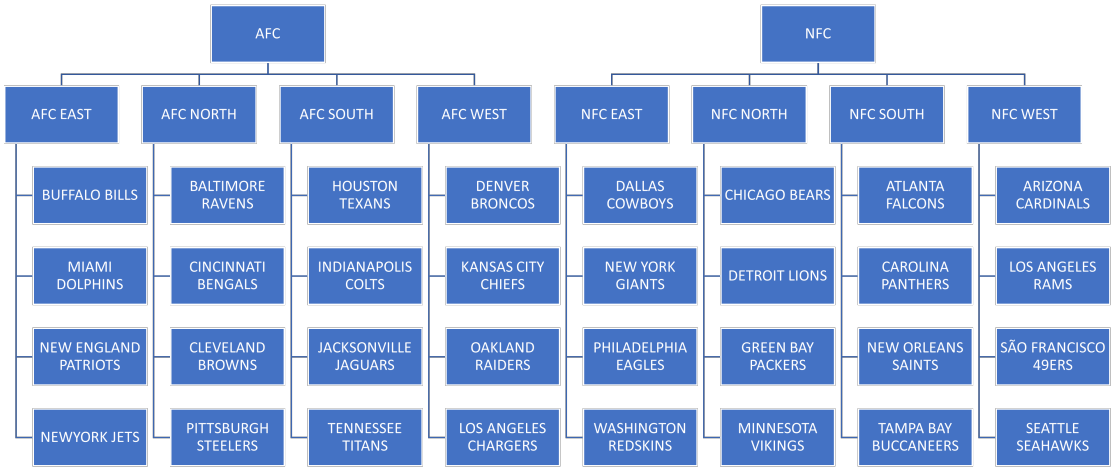

Figura 1.1: Fluxograma times da NFL.
Fonte: Produzida pela autora.

Como não é possível jogar contra todos os times da liga em apenas uma única temporada, a ***NFL*** estipulou um ciclo de quatro anos, no qual todos os times terão a oportunidade de competir contra todos os adversários da liga (Garcia (2017)).

O melhor time de cada divisão se classifica automaticamente para os *playoffs*, e para completar os seis times de cada conferência a avançar, o quinto e o sexto melhores times da classificação geral também garantem suas vagas. Como apenas 12 times dos 32 originais jogam as finais, e estas são jogos de eliminação, ou seja, o perdedor está fora do campeonato, decidimos analisar apenas os dados referentes aos jogos das temporadas regulares de um ciclo de quatro anos, para que assim exista a mesma quantidade de jogos e todas as comparações possíveis entre times.

O conjunto de dados foi obtidos no site *Kaggle* (Kendallgillies (2017)), no qual haviam conjuntos separados de jogadores, que foram agrupados por nós nos respectivos times em que jogaram no período. As variáveis são as características do jogo, isto é, as estatísticas coletadas em todas as partidas de uma temporada regular.

Com o interesse em estudar as características do jogo, em uma equipe da ***NFL***, iremos construir a análise time a time, agrupando as estatísticas dos mesmos até a temporada de 2019, terminada em fevereiro de 2020.

Por causa da pandemia do SARS-COV-19 algumas regras foram mudadas para me-

lhor se adequarem ao período social vigente, e por se tratarem de mudanças temporárias, os dados das temporadas de 2020 e 2021 não serão considerados para os fins deste estudo, já que podem trazer anomalias ao modelo.

No Capítulo 2 será feita uma breve explicação sobre o jogo, as formas de pontuações, as posições dos jogadores e as separações nas equipes. No Capítulo 3 será explicado o material utilizado para o estudo, tal como a explicação de como foram encontrados os bancos de dados, bem como foram compactados em um único conjunto de dados. No Capítulo 4 serão apresentados as metodologias para diminuir a quantidade de variáveis e agrupá-las. No Capítulo 5 serão apresentados as análises e os resultados encontrados. E no último Capítulo foi feita a conclusão do estudo de caso.

1.1 Objetivo

Este estudo tem como seu principal objetivo analisar, selecionar as características com maior contribuição para explicar a variabilidade dentro do conjunto de dados e, por fim agrupar as caraterísticas relevantes de acordo com a similaridade e a relações dessas características entre os três times presentes em cada equipe da ***NFL***.

O objetivo secundário do mesmo é a análise e classificação das variáveis que influenciam o rendimento de jogadores e times durante a temporada regular da liga, em um período de quatro anos.

Capítulo 2

Sobre o jogo

Para uma melhor compreensão do jogo, este capítulo será divido em partes, cada uma destas explicará uma parte desse complexo esporte.

2.1 Posições

Cada equipe é dividida em três principais times: ataque, defesa e especialistas, cada umas dessas possuem técnicos e coordenadores, que orientam esses jogadores.

2.1.1 Time de Ataque

O time de Ataque é o principal responsável por avançar em campo e marcar os pontos. Para desempenhar tais funções, estão relacionados os seguintes atletas:

- ***QuarterBack (QB)***: Responsável por ditar as jogadas (terrestre ou aérea).
- ***Running Back (RB)***: Jogadas terrestres.
- ***Fullback (FB)***: Jogadas terrestres e bloqueio.
- ***Wide Receiver (WR)***: Recebedor, principalmente jogadas aéreas.
- ***Tight End (TE)***: Recebedor e bloqueio/proteção ao QB.
- ***Linha Ofensiva (OL): Center, Guards, Tackle***: Responsáveis pela proteção do QB.

2.1.2 Time de Defesa

O time de Defesa é formado por jogadores responsáveis por não deixar o adversário marcar os pontos, que estão definidos abaixo:

- ***Safeties (S)***: Últimos jogadores para marcação.
- ***Cornerbacks (CB)***: Marcam os recebedores.
- ***Linebackers (LB)***: Ajudam a pressionar o QB e fazem marcação dos recebedores.
- ***Linha Defensiva (DL): Defensive End e Defensive Tackle.***: Responsáveis por pressionar a OL e tentar chegar ao QB, essa ação é denominada por *Rushing*.

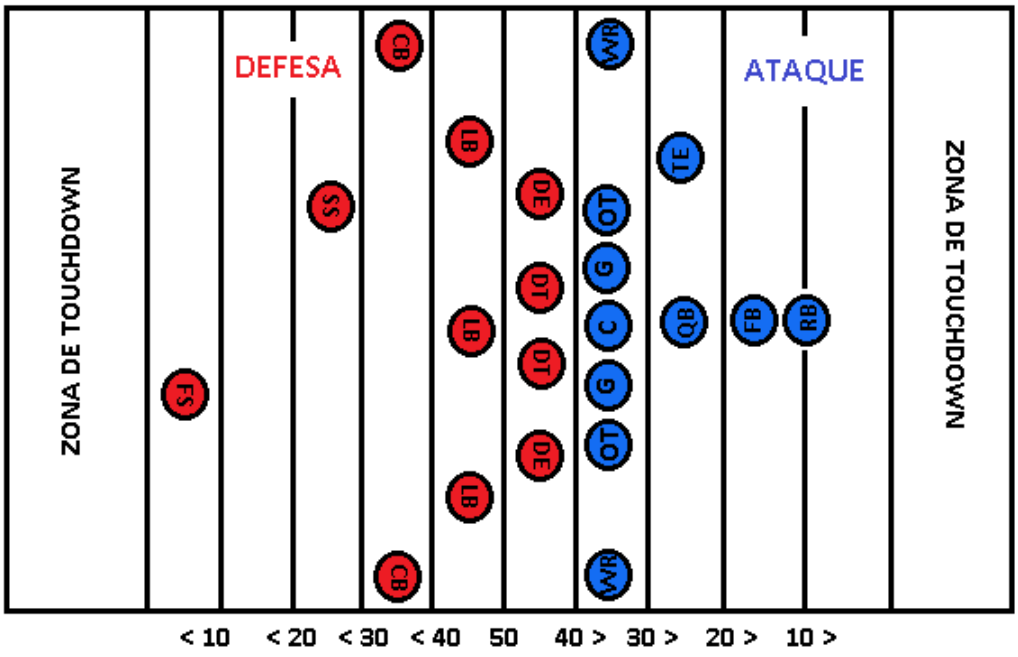

Figura 2.1: Esquema tático básico dos times de Defesa e Ataque.
Fonte: Produzida pela autora.

2.1.3 Time de Especialistas

O time de Especialistas: formado por jogadores especializados em jogadas em que ocorre um chute: *Punt, Kickoff, Field Goal*, que estão definidos abaixo:

- ***Kicker (K)***: Chuta para marcar pontos.
- ***Punter (P)***: Chutes de devolução da bola.

- ***Holder***: Jogador que segura a bola para o chute.
- ***Long Snapper***: Joga a bola para o *Holder*.
- ***Kick Returner***: Jogador que recebe a bola do chute e corre para avançar em campo.
- ***Gunner***: Corre pelo campo para tentar derrubar o *returner*.
- ***Jammers***: Tentam bloquear os *Gunners*.
- ***Blockers***: Bloqueiam para dar tempo de ocorrer o chute.

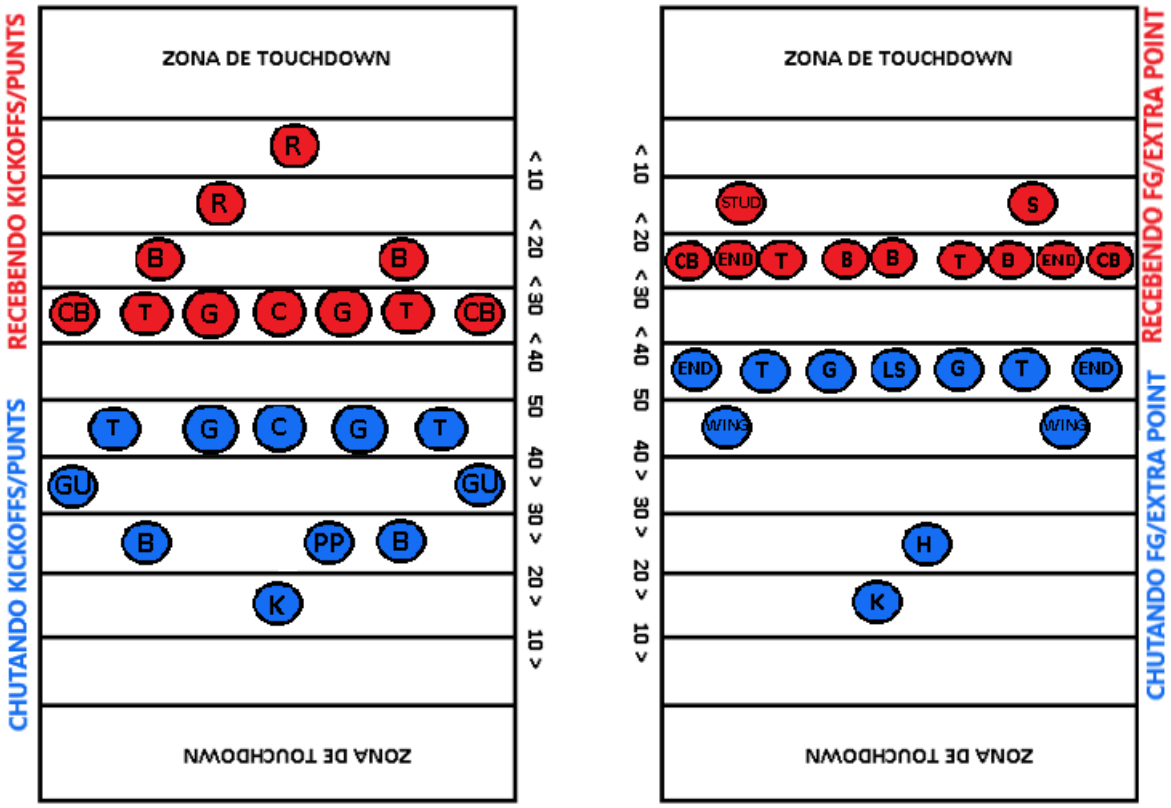

Figura 2.2: Esquema tático básico dos times de Especialistas.
Fonte: Produzida pela autora.

2.2 Pontuações

Existem algumas formas de pontuar no jogo, essas estão definidas na Tabela 2.1:

Tabela 2.1: Tabela da pontuações do Futebol Americano.

Pontuação	**Descrição**	**Pontos**
TouchDown	Jogador entra na zona de $TouchDown$.	6
Ponto extra	Chute extra, pós $TouchDown$ (sempre mesma distância).	1
Conversão de 2 pontos	Segundo $TouchDown$ em seguida.	2
Field Goal	Chute que entra no *Goal* (ocorre no último posicionamento da bola) .	3
Safety	Jogador com a bola, é derrubado na própria zona de $TouchDown$.	2
Pick six	$TouchDown$ do time de defesa.	6

2.3 Faltas e punições

A ***NFL*** possui muitas regras, que tem como função controlar as equipes, evitar sérias lesões, deixar a liga competitiva e manter o interesse do público. Podem sofrer penalidades tanto os jogadores, quanto as equipes, quanto organizações.

Durante as partidas, faltas de jogo são geralmente penalizados com jardas, isto é, o time que cometeu a falta retrocede jardas em campo e caso algum jogador tenha uma medida antidesportiva, pode ser ejetado do jogo. Para casos mais sérios este pode tomar punições e ser impedido de jogar em outras partidas. Caso a equipe e/ou organização infringir alguma regra, o time pode sofrer punições como perder escolhas de $Draft$, evento que permite que os times escolham os melhores atletas das universidades, seguindo uma ordem estabelecida pela ***NFL***, que é basicamente: o pior classificado da última temporada é o primeiro a escolher no $Draft$.

Alguns tipos de faltas:

- *Illegal Formation*: é qualquer formação de ataque que não é permitida pela regra, como ter menos de sete jogadores na linha de *scrimmage*, punição de cinco jardas.
- *False start*: Ocorre quando um jogador de ataque se movimenta antes de a bola ser colocada em jogo, induzindo o movimento de um defensor, punição de cinco jardas.
- *Offside*: é quando um jogador de defesa se movimenta antes de a bola ser colocada em jogo, induzindo o movimento de um defensor, punição de cinco jardas.

- *Too Many Men On The Field*: cada equipe pode ter no máximo 11 jogadores dentro de campo por jogada, caso tiver mais é punido com cinco jardas.
- *Unnecessary Roughness*: é a falta quando ocorre uma abordagem excessivamente violenta para parar o adversário. A infração é penalizada com a perda de 15 jardas. Se o árbitro entender como flagrante, o jogador pode ser desqualificado.
- *Face Mask*: é quando um jogador segura na grade do capacete de seu adversário, punido com 15 jardas.

2.4 O Jogo

O jogo tem quatro tempos de 15 minutos, caso termine empatado ao final do quarto tempo, existe a prorrogação de 15 minutos também, se ainda não tiver um ganhador joga-se mais um tempo de prorrogação, e assim continua até que um dos times tenha uma pontuação maior do que a do outro. O jogo tem como principal objetivo avançar em campo e marcar pontos.

Logo antes de começar o jogo, joga-se uma moeda, o ganhador escolhe se quer começar recebendo a bola no primeiro tempo ou no terceiro. O jogo é iniciado com as esquipes de especialistas em ambos as equipes em campo, com um $kickoff$, que é um chute do meio de campo, a equipe adversária tenta recepcionar essa bola e avançar em campo o máximo possível. Assim que essa jogada acaba, a equipe com a posse de bola entra com seu time de ataque e o outro com o time de defesa.

O time de ataque possui quatro tentativas para avançar no mínimo dez jardas em campo, *First Down* é quando o ataque consegue chegar na marca de dez jardas a partir do começo, e 45 segundos realizar cada uma dessas chances:

- enquanto o time conseguir o *First Down*, o ataque continua em campo para assim tentar marcar um *Touchdown*.
- caso não consiga:
  - pode tentar realizar um *Field Goal* para pontuar, isso geralmente ocorre se a bola estiver posicionada no meio campo pelo menos.
  - se não for tentar o chute para pontuação, o time é obrigado a realizar o chute de devolução da bola, para que assim a outra equipe tenha chance de marcar

pontos.

Toda essa campanha de um ataque é chamado de *Drive*.

Durante essas tentativas de pontuação, pode acontecer roubos da bola, bloqueios, isto é, o time de defesa consegue ter a posse da bola e desta forma, serão trocados os times em campo sem ser realizado um chute.

Capítulo 3

Material

Os dados usados para esse estudo, possuem tanto variáveis quantitativas, como peso, altura, quantidade de faltas cometidas, quanto variáveis qualitativas, como time do jogador, posição em que joga, local de nascimento de jogador.

Em cada linha do banco de dados é apresentado o nome do jogador, o time em que joga e variáveis como:

- *Características pessoais dos jogadores:* Número, posição, time atual, altura, peso, idade, aniversário, local de nascimento, faculdade frequentada, ensino médio frequentado, local onde estudou o ensino médio, tempo de experiência, em anos, na ***NFL***.

- *Características do jogo, estatísticas de*: defensivas, *Field Goals*, *Fumbles* (jogador perde a posse da bola), retorno de *Kickoff*, *Kickoff*, linha ofensiva, passes, retorno de *punt*, *punting*, recepções, *rushing*.

- *Estatísticas de posições específicas: Quarterback - QB*, *Running back*, *Wide Receiver* e *Tight End*, *Offensive Line*, *Defensive Lineman*, *Kickers*, *Punters*.

O conjunto de dados foi compilado por um usuário do *site Kaggle*, este adquiriu através dos dados estatísticos fornecidos pelo site oficial da ***NFL***. São 19 conjuntos de dados, com aproximadamente 81195 observações no total, entre eles se encontram dados sobre estatísticas básicas de jogadores e outros são divididos em posições ou características de jogos parecidos.

Estão disponíveis dados de jogadores desde 1970 a 2016, e por conta do ciclo de quatro anos adotado pela ***NFL***, escolhemos os anos 2013, 2014, 2015 e 2016 para realizar

essa análise, dado que quanto mais recente a escolha, mais informações foram coletadas e armazenadas. Os bancos de dados do *site*, estavam organizados com os jogadores nas linhas e as variáveis nas colunas. Em todos os bancos existiam pelo menos seis colunas iguais: identificação do jogador, nome do jogador, ano das informações daquela linha, o time que estava naquele ano, número utilizado pelo jogador e posição em que o jogador atua.

Alguns exemplos de como estavam divididos os bancos de dados:

- Nos dados de estatísticas básicas, que possui 16 variáveis com informações sobre características básicas por jogador, como peso, altura, escola e universidade que estudou, data de aniversário e local de nascimento.

- Nos dados sobre o time de defesa, encontra-se 17 variáveis de jogadores como time em que joga, ano em análise, posição que joga e aspectos defensivos, como quantidade de *tackle* (acabar com a jogada derrubando o jogador de posse da bola), *sacks* (*tackle* no QB) e interceptação.

- No banco de dados sobre a posição de $Kicker$, como ano de referência, quantidade de jogos que participou, *Field Goal* mais longo, chutes bloqueados, quantidade de pontos extras feitos e quantidade de $Kickoffs$ realizados.

- Os dados referentes à *rushing*, possui 18 variáveis sobre a ação de *Rushing* pelo ataque, que é apressar a defesa para avançar em campo, como posição do jogador, jardas conseguidas com *Rushing*, quantidade de *FirstDown* conseguido com *Rushing* e *Rushing* mais longo.

Como todos os conjuntos de dados possuem o time e ano em que foram coletadas as informações, iremos juntar as variáveis por time nos anos de interesse, citados anteriormente. O conjunto final de dados terá como observações os times por ano, as variáveis serão as variáveis dos outros bancos de dados, já retirando variáveis como nome do jogador e números de identificação destes, já os valores serão a soma, ou porcentagem, ou média, de acordo com a forma mais pertinente de apresentar os valores, para assim representar adequadamente o time e não os jogadores como indivíduos. Vale ressaltar que o cálculo das variáveis de porcentagem, foram realizadas através de média ponderada das observações.

3.1 Estrutura dos dados

A base de dados utilizada nesse estudo, foi criada a partir junção de outros conjuntos de dados disponibilizados pela ***NFL*** e agrupado por um usuário do *site Kaggle*. Os bancos estavam estruturados em bancos separados por algumas posições e características referentes ao jogo, nos quais as unidades de medida eram os jogadores nos anos de 1970 a 2016.

O primeiro passo para a criação do banco de dados, foi filtrar todos os dados pelo período de tempo desejado, 2013 a 2016. O segundo passo foi juntar os jogadores dos mesmos times, dessa forma as unidades de medida deixaram de ser jogadores e passaram a ser os times. Neste momento foi feita a junção das observações, analisando cada variável individualmente para saber qual a melhor forma de conseguir um valor total. Após isto, algumas variáveis foram tiradas do conjunto de dados, quando eram medidas repetidas. Só então, após cada conjunto arrumado individualmente, as observações foram colocadas em uma só planilha. Ao olhar as variáveis conjuntamente, foram identificadas mais variáveis iguais e assim, olhando atentamente para estas, foram somadas ou excluídas.

Com a planilha finalmente pronta, foi realizada uma análise sobre a variabilidade de cada variável, na qual as que não possuíam uma variabilidade maior que 70% foram retiradas do conjunto de dados final. Neste banco de dados, as unidades de observação são os times por cada ano em estudo, que totalizam 128 unidades observacionais e 31 variáveis a serem analisadas. Estas estão listadas abaixo, juntamente com a descrição de cada uma delas.

- ***Passing Yards***: Quantidade de jardas de passes por temporada.
- ***TD Passes***: Quantidade de passes para **TD** por temporada.
- ***Passer Rating***: Nota dada pela ***NFL*** ao ***QB***, varia de 0 à 158.3. Tal fórmula é calculada da seguinte forma e pode ser encontrada em *sites* oficiais da Liga:

$$PasserRating_{NFL} = \left(\frac{mm(a) + mm(b) + mm(c) + mm(d)}{6} \right) \times 100,$$

onde

$mm(x) = \mathbf{max}(\mathbf{0}, \mathbf{min}(\mathbf{x}, \mathbf{2}, \mathbf{375}))$

$a = \left(\frac{\mathbf{COMP}}{\mathbf{ATT}} - 0,3 \right) \times 5$

$b = \left(\frac{\mathbf{YARDS}}{\mathbf{ATT}} - 3\right) \times 0,25$

$c = \left(\frac{\mathbf{TD}}{\mathbf{ATT}}\right) \times 20$

$d = 2,375 - \left(\frac{\mathbf{INT}}{\mathbf{ATT}} \times 25\right)$

onde

ATT= Passes tentados,

COMP= Passes completos,

YARDS= Jardas aéreas,

TD=Passes para *Touchdown*

INT= Passes interceptados.

- ***Completion Passes Percentage:*** Porcentagem de passes em que o jogador que recebe a bola não a perde.
- ***Number of Receptions***: Número de recepções de passes.
- ***Receiving TDs***: Quantidade de *Touchdowns* por passes.
- ***Receptions Longer than 20 Yards***: Recepções quando a bola percorre mais de 20 jardas no ar.
- ***Receptions Longer than 40 Yards***: Recepções quando a bola percorre mais de 40 jardas no ar.
- ***First Down Receptions***: Recepções de passe resultando em *First Down*.
- ***Yards Per Game***: Jardas totais por jogo.
- ***Rushing Attempts***: Tentativas de corridas.
- ***Rushing Yards***: Jardas percorridas por corridas.
- ***Rushing TDs***: Quantidade de *Touchdowns* por corridas.
- ***Rushing First Downs***: Corridas resultando em *First Down*.
- ***Ints***: Quantidade de interceptações.
- ***Total Tackles***: Quantidade de términos de jogada ao derrubar o jogador adversário com a posse da bola.

- ***Assisted Tackles***: Quantidade de ajuda para acabar com a jogada ao derrubar o jogador adversário com a posse da bola.
- ***Sacks***: Acabar com a jogada derrubando o ***QB***, esse tendo a posse da bola.
- ***Passes Defended***: Quantidade de passes defendidos, desviando, interceptando ou forçando o adversário a perder a posse da bola.
- ***Int Yards***: Jardas por interceptação.
- ***Fumbles***: Jogador que esta com a bola perde a posse e é recuperada pelo time adversário.
- ***Longest FG Made***: *Field Goal* concebido mais distante.
- ***FG Percentage***: Porcentagem de *Field Goal*
- ***Kickoffs***: Quantidade de chutes do meio de campo, de início dos tempos ou após ponto por chute.
- ***Touchbacks***: A bola sai pela linha de fundo após o chute, $Punt$ e $Kickoffs$.
- ***Kickoffs Returned***: Quantidade de $Kickoffs$ retornados.
- ***Punt Returns***: Quantidade de $Punts$ retornados.
- ***Punt Fair Catches***: Recurso de segurança utilizado pelo recebedor do $Punt$, se realizado a jogada para, não pode avançar em campo.
- ***Punts***: Quantidade de $Punts$ chutados.
- ***Kicks Yards Returned***: Quantidade de jardas retornadas após $Punt$ e $Kickoffs$.
- ***Percentage of Extra Points Made***: Porcentagem de *Extra Points* feitos.

Apesar de das unidades de observação serem os times em anos consecutivos, os time possuem uma grande rotação de jogadores, sendo comum o mesmo time troca de elenco e comissão técnica entre uma temporada regular e outra, não tendo algum problema com a suposição de independência, que foi aceita.

Capítulo 4

Métodos

Como o conjunto de dados trabalhado nesse projeto possui uma grande quantidade de variáveis, iremos utilizar o método de Análise de Componentes Principais (ACP), para averiguar quais são as variáveis de maior importância a serem analisadas. (dos Anjos, 2018). Após essa análise, iremos utilizar o método de Análise de Agrupamentos (*Clusters*), para agregar unidades experimentais, com base nas características que elas possuem (Pedro Ferreira, 2020).

Segundo da Costa Moraes (2016), a ACP é a decomposição das variáveis observadas em componentes independentes entre si, de maneira a maximizar a variância total explicada. Se variáveis forem altamente correlacionadas, poucos componentes explicariam uma maior quantidade de variância na amostra. Assim, usamos a ACP com o objetivo de reduzir os dados para obtenção do mínimo número de fatores necessários para assim explicar o máximo da variância representada pelas variáveis originais.

A ACP é um método normalmente utilizado para intermediar análises multivariadas, neste trabalho foi seguido por uma Análise de Agrupamentos. Este tem o intuito de reduzir a quantidade de variáveis, selecionando apenas os componentes com maior contribuição (autovalores altos) e como consequência do método, gerando um conjunto de dados sem multicolinearidade das variáveis. E como resposta, teremos as componentes de maior importância para a análise final.

A Análise de Agrupamentos é um método que agrupa unidades experimentais ou variáveis em grupos com características em comuns, é uma ferramenta que se adapta para qualquer situação, possibilitando um levantamento e análise de dados mais assertivo (Trecsson, 2016).

Para resumir, o objetivo básico de Análise de Agrupamentos é descobrir um bom

agrupamento natural das observações (ou variáveis). Porém é necessário inicialmente desenvolver uma escala qualitativa para medir a associação entre as observações, isto é, o quanto essas variáveis são similares (similaridade) (Pedro Ferreira, 2020).

Segundo Hair (2009), uma importante suposição é a representatividade da amostra, observações atípicas devem ser analisadas previamente para não introduzir um viés na estimação da estrutura de agrupamento dos dados. Portanto, todos os esforços devem ser feitos para garantir a representatividade da amostra e que os resultados possam ser generalizáveis para a população de interesse. (SAATE, 2007)

Para este trabalho iremos utilizar primeiramente o método ACP e após ao possuir só as variáveis com significância utilizar o método de agrupamentos. Realizaremos isso, pelo fato do nosso objetivo é diminuir a quantidade de variáveis para a análise e depois agrupar, e assim não correr o risco de perder informações ao descartar grupos inteiros.

4.0.1 Análise de Componentes Principais

Com o tamanho do conjunto de dados a ser utilizada, torna-se quase impossível a análise de todas as variáveis, no entanto para que não se perca informações simplesmente por eliminar variáveis, desta forma será utilizada primeiramente a Análise de Componentes Principais, com o intuito de reduzir variáveis (Rocha, 2020).

A análise de componentes principais é uma técnica estatística multivariada que transforma linearmente um conjunto original com $\boldsymbol{p}$ variáveis ($\mathbf{X_1}$,..., $\mathbf{X_p}$) em um conjunto com um número menor (k) de variáveis não correlacionadas ($\mathbf{Y_1}$,..., $\mathbf{Y_p}$) denominadas componentes principais, que explicam uma parcela substancial das informações do conjunto original e são combinações lineares das variáveis originais, na qual Y_1 explica a maior parcela da variabilidade total dos dados e assim por diante.

A análise de componentes principais depende somente da matriz de covariância ou da matriz de correlação de ($\mathbf{X_1}$,..., $\mathbf{X_p}$), de dimensão p, sendo p o número de variáveis, requer suposição de que as variáveis sejam independentes e identicamente distribuídas, não é necessário que tenham distribuição normal. A ACP tem como objetivos principais a redução da dimensionalidade dos dados, a obtenção de combinações interpretáveis das variáveis, e a descrição e entendimento da estrutura de correlação das variáveis (Barroso e Artes, 2003).

De uma maneira geral, os componentes principais são calculados através dos autovalores da matriz variância - covariância (var-cov) $\Sigma_{p\times p}$, uma matriz simétrica, isto é, os

elementos $\sigma_{ij} = \sigma_{ji}$, sendo $i, j = 1, 2, ..., p$.

$$\boldsymbol{\Sigma_{p \times p}} = \begin{cases} \sigma_{ij}^2, \text{ se } i = j \\ \sigma_{ij}, \text{ se } i \neq j. \end{cases} \tag{4.1}$$

Desta forma, temos a matriz var-cov a seguir:

$$\boldsymbol{\Sigma_{p \times p}} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \cdots & \sigma_{2p} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \cdots & \sigma_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \sigma_{3p} & \cdots & \sigma_p^2 \end{pmatrix} \tag{4.2}$$

Após pronta a matriz var-cov, é feita uma padronização e a matriz $\boldsymbol{\Sigma_{p \times p}}$ se transforma na matriz de correlação $\boldsymbol{R}$, quadrada simétrica de ordem p, em que os elementos $\rho_{ij} = \rho_{ji}$ e para $i \neq j$ $\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_i^2 \sigma_j^2}}$, com $i, j = 1, 2, ..., p$.

$$\boldsymbol{R_{p \times p}} = \begin{cases} 1, \text{ se } i = j \\ \rho_{ij}, \text{ se } i \neq j. \end{cases} \tag{4.3}$$

E assim, temos a matriz de correlação a seguir:

$$\boldsymbol{R_{pxp}} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \rho_{23} & \cdots & \rho_{2p} \\ \rho_{13} & \rho_{23} & 1 & \cdots & \rho_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \rho_{3p} & \cdots & 1 \end{pmatrix} \tag{4.4}$$

Através da matriz $\boldsymbol{R}$, são calculados os autovalores e os autovetores. Esse cálculo ocorre substituindo os valores dos autovalores no polinômio característico (P_A), que é calculado através da fórmula (2.5), na qual *det* é o determinante da matriz $\boldsymbol{R}$, I_p é a matriz identidade, da mesma dimensão da matriz $\boldsymbol{R}$ e λ são os autovalores associados a matriz $\boldsymbol{R}$.

$$P_A(\lambda) = \det(\lambda I_p - R) \tag{4.5}$$

Os valores de λ que igualam o P_A a 0 são chamados de autovalores. Por sua vez os autovetores são calculados através dos autovalores, a partir da subtração de um a um dos valores de λ na matriz $\boldsymbol{R}$, achando assim as coordenadas dos autovetores. Os valores das coordenadas dos autovetores, nos informarão a contribuição para cada componente principal e os valores dos autovalores será o quanto a componente é explicada através da variância.

4.0.2 Análise de Agrupamentos

Utilizamos as variáveis representativas dos Componentes Principais obtidos para dar seguimento ao estudo, queremos agrupar estas.

O método se resume a um algoritmo de classificação, o qual pode ser apresentado em quatro etapas:

- Definição dos **n** indivíduos, no qual cada indivíduo forma uma classe.
- Agrupamento dos dois indivíduos mais similares, formando uma única classe, o número de classes restantes passa a ser **n-1**.
- Repetição do passo anterior até que tenhamos uma só classe.
- Caracterização das classes obtidas a partir das variáveis utilizadas na classificação.

Existem diversos métodos de agrupamento, dentre eles, métodos hierárquicos e não hierárquicos. Os métodos hierárquicos podem ser usados de forma exploratória e flexível, captando observações discrepantes, similaridades e agrupamentos iniciais, isto é, antes da análise entre as observações de interesse, segmentando estas em diferentes grupos com características distintas. Após explorar os dados com os agrupamentos hierárquicos, podemos refinar as segmentações com agrupamentos não hierárquicos, que requerem um número de agrupamentos definido previamente. Essas duas técnicas são normalmente utilizadas de forma complementar (Souza, 2007).

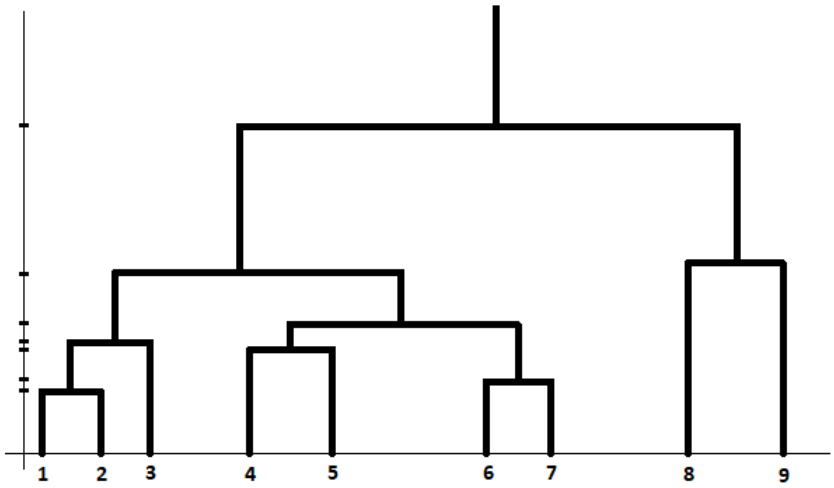

Figura 4.1: Demonstração da ordem de agrupamento através do Dendrograma.
Fonte: Produzida pela autora.

Neste trabalho utilizaremos métodos hierárquicos para agrupar as componentes principais, após descobrir a quantidade de grupos, será realizado o método não hierárquico para uma melhor análise. Dado que nosso objetivo é encontrar quais as características são importantes para a formação do time ideal, seria incorreto definirmos anteriormente as características, por isso iremos realizar os dois métodos.

Para que se defina o número ótimo de grupos, são realizados cálculos e gráficos específicos de cada método, porém a decisão de quantos grupos terão ao final é função do pesquisador ou analista.

Métodos Hierárquicos

No agrupamento hierárquico, existem algumas técnicas que podem ser utilizadas, essas são divididas em essencialmente em duas formas, o aglomerativo que vai agrupando através de pequenos grupos isolados, e o divisor que começa com um só grupo e vai se dividindo até que fiquem todos separados.

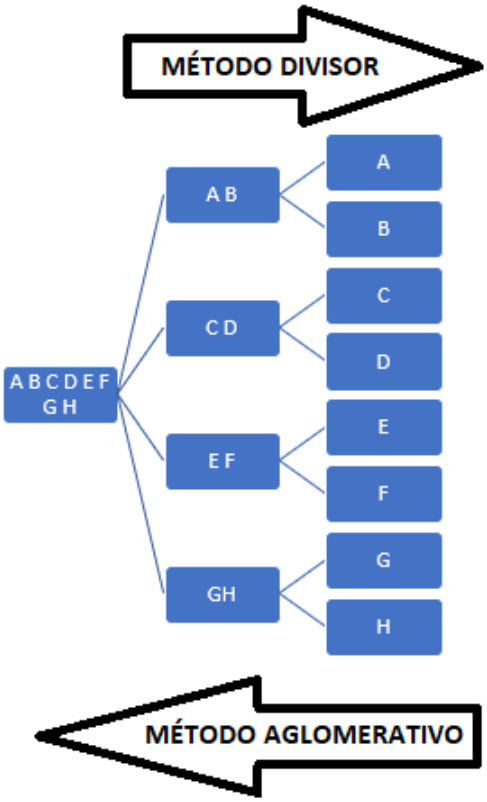

Figura 4.2: Ilustração agrupamento hierárquico.
Fonte: Produzida pela autora.

O método aglomerativo inicia com pequenos grupos de grandes características em comum, ao passo em que se vai adicionando mais observações a esses grupos a similaridade vai se perdendo. Já o método do divisor, começa com uma menor similaridade, mas ao passo em que os grupos vão diminuindo, aumenta a similaridade entre as observações em cada grupo. A partir do momento em que as distâncias entre os grupos for muito grande, isto é, não existir mais similaridade entre os grupo, é possível parar o método e assim terá o número de agrupamentos com a similaridade desejada, desta forma utilizaremos o método aglomerativo.

De uma forma simplificada, o passo a passo para realizar o método aglomerativo é:

- Passo 1: Colocar todas as variáveis como grupos individuais.
- Passo 2: Combinar os grupos com mais similaridade, isto é, maior similaridade, pela matriz de similaridade.
- Passo 3: Calcular os novos valores da matriz de similaridade, após o agrupamento do passo anterior.
- Passo 4: Repetir os passos 2 e 3 até que todas as variáveis formem um único grupo.

A matriz de similaridades (D) é uma matriz simétrica de ordem $\boldsymbol{p}$, na qual os valores da matriz (d_{ij}), que são valores reais, não negativos, pois são a distância entre i e j. Abaixo encontra-se uma exemplificação da matriz D:

$$D_{p\times p} = \begin{pmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1p} \\ d_{12} & 0 & d_{23} & \cdots & d_{2p} \\ d_{13} & d_{23} & 0 & \cdots & d_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{1p} & d_{2p} & d_{3p} & \cdots & 0 \end{pmatrix} \tag{4.6}$$

Para se agrupar as observações, existem alguns métodos, estes tem as distâncias calculadas por diferentes maneiras.

- **Vizinho mais próximo:** são agrupados os grupos que tem a menor medida de similaridade, no caso a distância.
- **Vizinho mais distante:** são agrupados os grupos que tem a maior medida de similaridade, no caso a distância.
- **Ward:** usa a soma de quadrados entre grupos como medida de similaridade, é um método que minimiza a variância interna, tendo como resultado grupos aproximadamente iguais.
- **Centróide:** a medida de similaridade utilizada nesse método é a distância entre centroides calculada pelo ponto médio de cada grupo.

Métodos Não Hierárquicos

No agrupamento não hierárquico, a técnica mais utilizada e conhecida é o método *K-médias*. O método do *K-médias* é um algoritmo que agrupa em centroides observações de acordo com a similaridade, a medida de dissimilaridade usada é a distância Euclidiana, no entanto, para utilizar esse método é necessário definir antes de tudo a quantidade final de agrupamentos que deseja-se ter. Para tal, usaremos o número de agrupamentos definido pelo método hierárquico (Izbicki e dos Santos, 2020).

De uma forma simplificada, o passo a passo para realizar o método *K-médias* é:

- Passo 1: Definidos K centros provisórios.
- Início do processo de agrupamento.

- Passo 2: Combina cada observação com o centro na qual possuí maior similaridade, formando K grupos.
- Passo 3: Calcula novos centros, para os novos agrupamentos.
- Passo 4: Repete os passos 2 e 3, até estabilizar os agrupamentos.

Esse método necessita a definição do K antes de começar o processo, mas ao utilizar o método hierárquico para definir K, essa desvantagem é minimizada. De modo positivo, o método tem uma eficiente aplicabilidade para grandes conjuntos de dados, além de que um indivíduo pode ser realocado durante o processo.

Capítulo 5

Resultados

5.1 Análise descritiva e exploratória dos dados

Com o intuito de verificar o comportamento das variáveis do banco de dados estudado, foram realizadas algumas análises descritivas de características muito analisadas em jogos da ***NFL***. Através das medidas resumo das variáveis será realizada uma análise para ter uma visão inicial geral do comportamento dos dados.

Tabela 5.1: Tabela das medidas resumo da variáveis dos Times de Ataque em estudo.

Variáveis	**Min.**	1^o **Qu.**	**Mediana**	**Média**	3^o **Qu.**	**Max.**
Passing Yards	0,00	3358,00	3874,00	3524,00	4343,00	5572,00
TD Passes	0,00	19,00	23,50	24,12	30,25	55,00
Passer Rating	21,33	68,05	82,00	80,42	94,05	130,00
Completion Passes Percentage	0,00	46,81	59,97	56,99	67,00	89,03
Number of Receptions	30,00	2720,00	325,50	313,60	359,50	472,00
Receiving TDs	9,00	17,00	22,00	23,15	28,00	48,00
Receptions Longer than 20 Yards	20,00	39,75	46,00	46,29	53,00	72,00
Receptions Longer than 40 Yards	1,00	7,00	9,00	8,81	11,00	21,00
First Down Receptions	12,00	149,80	181,00	176,00	202,20	265,00
Yards Per Game	10,98	16,38	18,52	20,76	21,16	253,70
Rushing Attempts	36,00	351,00	402,50	382,60	438,20	545,00
Rushing Yards	55,06	739,78	1205,00	1170,27	1616,00	2266,00
Rushing TDs	1,00	7,00	11,00	11,32	15,00	29,00
Rushing First Downs	6,00	66,75	87,00	81,30	100,25	146,00

Pela Tabela 5.1, é interessante ressaltar que nenhum time obteve a nota máxima de *Passador*, também que todas as variáveis possuem um melhor coeficiente de variação. Ao olhar para os dois tipos de *Touchdown*, os aéreos (*Receiving*) acontecem em média duas vezes a mais do que os terrestres (*Rushing*), no entanto, ao analisar as jardas percorridas por cada uma dessas variáveis (*Passing Yards*e *Rushing Yards*), a relação é a contrária da anterior. Por fim, podemos observar que as recepções de mais de 40 jardas, consideradas longas acontecem poucas vezes durante a temporada. Vale ressaltar que valores mínimos iguais a zero, mostram uma falta de sintonia no time de ataque, isso pode acontecer por condições climáticas não favoráveis, lesões, ou até a estratégia do próprio time de priorizar outro tipo de jogada.

Tabela 5.2: Tabela das medidas resumo da variáveis dos Times de Defesa em estudo.

Variáveis	**Min.**	1^o **Qu.**	**Mediana**	**Média**	3^o **Qu.**	**Max.**
Ints	0,00	10,00	13,00	13,58	17,00	42,00
Total Tackles	573,00	838,80	941,00	929,30	1014,20	1378,00
Assisted Tackles	103,00	196,50	246,50	255,50	310,00	490,00
Sacks	15,00	30,50	36,00	36,11	42,50	62,00
Passes Defended	27,00	59,00	70,00	69,52	78,25	120,00
Int Yards	11,00	109,00	178,50	190,60	253,80	516,00
Fumbles	10,00	23,00	28,00	28,67	34,00	91,00

Ao analisar a Tabela 5.2, apenas uma variáveis possui o mínimo em zero, mostrando que de modo geral os times de defesa apresentam estatísticas positivas, isto é, os times de defesa são mais consistentes e possuem boas estatísticas. Dentre os passes defendidos, que pode ser interceptação, $fumbles$ e dificultar com boas marcações, a que mais ocorre aproximadamente são os $fumbles$. Ao ser realizada uma interceptação, esses jogadores conseguem realizar pelo menos um pequeno avanço em campo.

Tabela 5.3: Tabela das medidas resumo da variáveis dos Times Especiais em estudo.

Variáveis	**Min.**	1^o **Qu.**	**Mediana**	**Média**	3^o **Qu.**	**Max.**
Longest FG Made	0,00	51,00	53,00	48,79	55,00	64,00
FG Percentage	0,00	79,12	84,55	77,33	88,90	100,00
Kickoffs	0,00	67,50	79,50	70,09	87,00	115,00
Touchbacks	0,00	30,75	40,50	37,69	52,00	81,00
Kickoffs Returned	0,00	23,00	32,00	30,34	40,25	73,00
Punt Returns	0,00	24,00	30,00	30,23	38,00	60,00
Punt Fair Catches	0,00	13,00	18,00	18,12	22,00	111,00
Punts	0,00	61,00	73,00	66,43	81,25	109,00
Kicks Yards Returned	0,00	506,00	702,00	702,30	895,20	1640,00
Percentage of Extra Points Made	0,00	92,30	97,55	88,08	100,00	100,00

Com as medidas resumo do time de especialistas da Tabela 5.3, chama a atenção as duas medidas de porcentagem, as quais possuem altas chances de pontuação, isto mostra

que existem poucos erros nessas jogadas. Nas estatísticas de chutes retornados, mostra grande semelhança entre as jogadas. E por fim, o ***FG*** mais longo concebido foi de 64 jardas, o que justifica o motivo de não ser comum as tentativas de chute para pontuar antes do meio de campo.

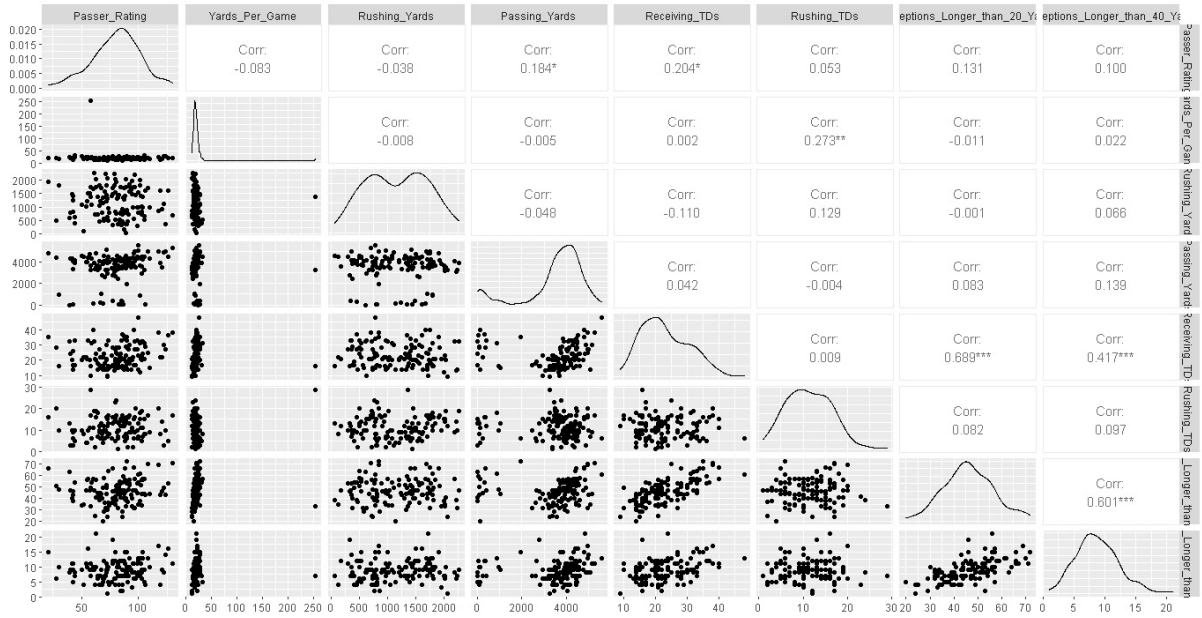

Figura 5.1: Matrix Plot do time de Ataque

Analisando a Figura 5.1, a correlação entre as variáveis de ataque, as maiores relações são as de *Recepções maiores de 20 jardas* e *TD por recepção*, *recepções maiores de 20 jardas* e *recepções maiores de 40 jardas*, com aproximadamente 0,7 e 0,6 respectivamente. A variável de *jardas por jogo* tem um ponto em destaque, separado do restante, esse ponto é o *Buffalo Bills - 2016*, essa temporada foi a que o time quebrou mais recordes. Já as outras variáveis possuem uma nuvem de pontos aparentemente dispersa.

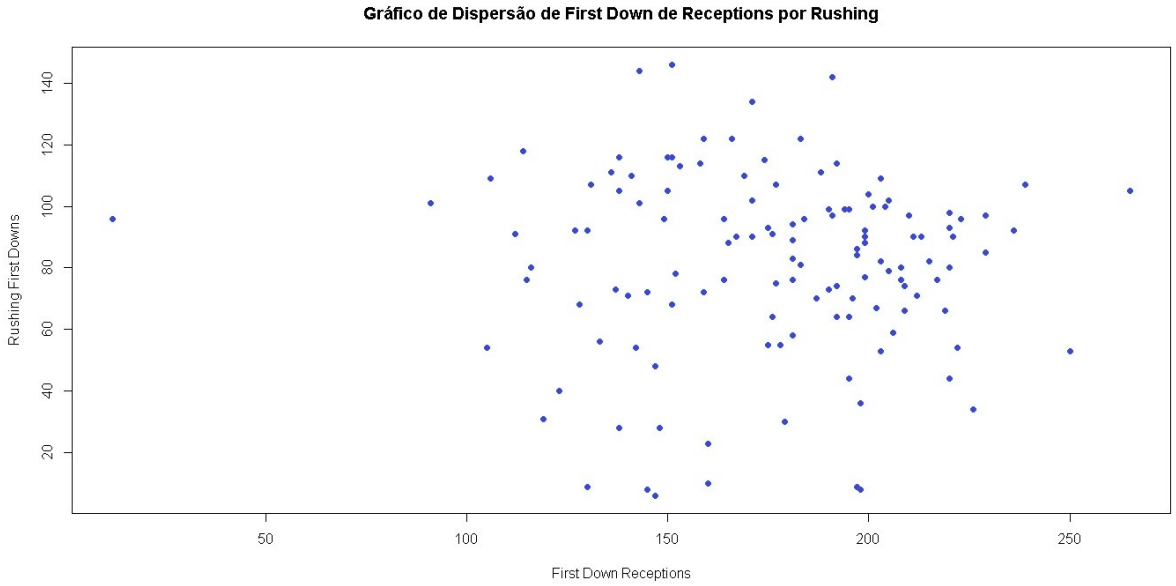

Figura 5.2: Gráfico de Dispersão de First Down de Receptions por Rushing.

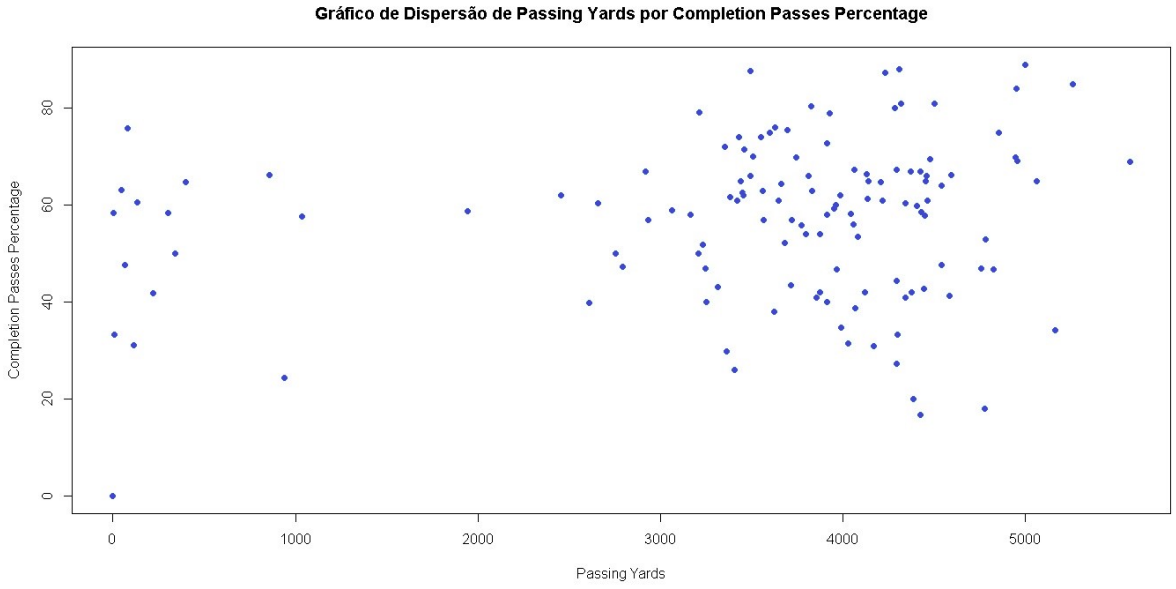

Figura 5.3: Gráfico de Dispersão de Passing Yards por Completion Passes Percentage

A Figura 5.2 possui uma nuvem de pontos dispersa, na qual aparentemente possuí *outlier* e a Figura 5.3 parece possuir duas nuvens de pontos, uma com valores pequenos de jardas aéreas e outra com valores alto de passes aéreos.

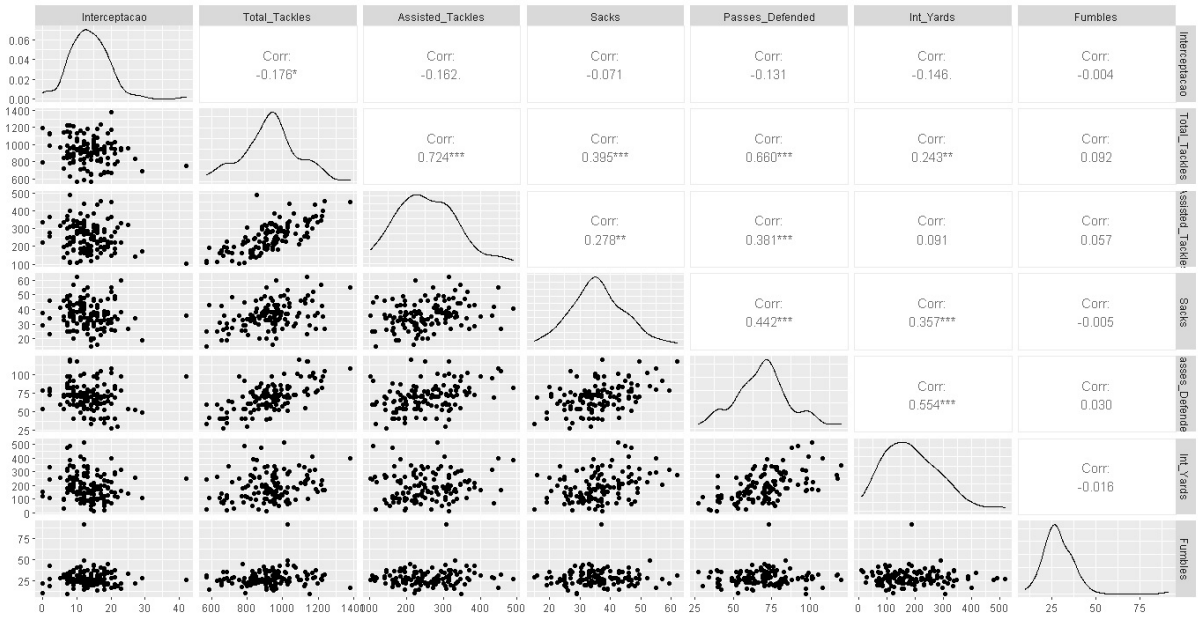

Figura 5.4: Matrix Plot do time de Defesa

A correlação das variáveis de defesa observadas na Figura 5.4, possuem um valor de 0,73 entre quantidade total de *tackles* e ajuda para realizar *tackles*, tal como uma correlação positiva de 0,66 entre quantidade *total de tackles* e *passes defendidos*. Com exceção da variável *Fumble*, as nuvens de pontos das variáveis parecem estar bem distribuídas. Quanto a existência de *outliers*, apenas duas variáveis aparentam ter, *Fumble* e *interceptação*.

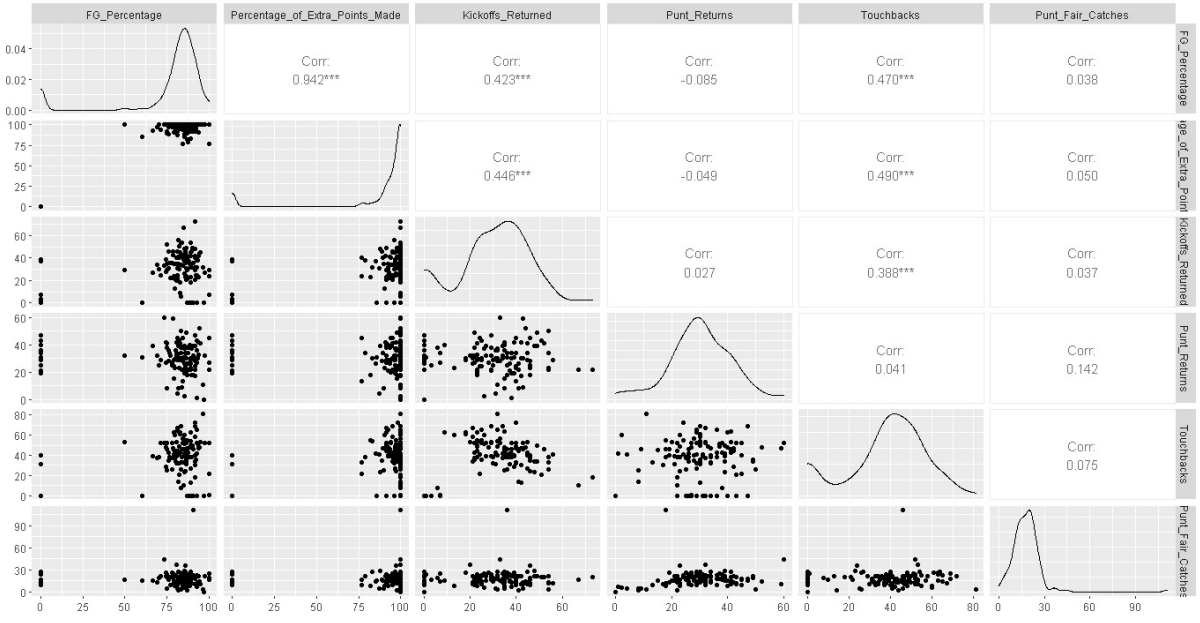

Figura 5.5: Matrix Plot do time de Especialistas

Na Figura 5.5, encontramos a maior correlação entre variáveis, que é uma correlação

positiva de 0,94 entre as porcentagens de *FG* e *Pontos Extras convertidos*, essa correlação alta obtida é devido aos pontos influentes na posição (0,0). Ao analisar as nuvens de pontos, estas possuem formas variadas entre as variáveis dos times de especialistas. Quanto a existência de *outliers*, as duas variáveis que medem porcentagens e *Punt Fair Catches* aparentam possuir pelo menos um ponto discrepante.

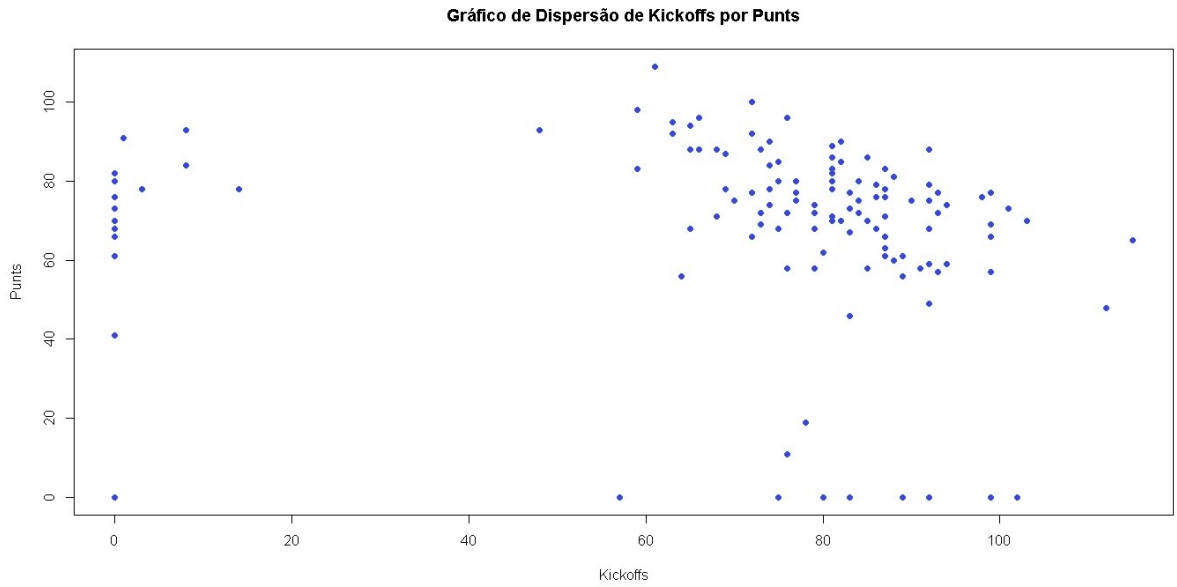

Figura 5.6: Gráfico de Dispersão de Kickoffs por Punts

Através do Gráfico de dispersão da Figura 5.6, podemos identificar quatro grupos de pontos. O primeiro quando existe valores próximos de zero da variável $Kickoffs$ e altos valores da variável$Punts$, o segundo é o oposto do anterior, grandes valores para a variável $Kickoffs$ e valores próximos a zero de $Punts$. O terceiro grupo possui valores altos para ambas as variáveis, e o último apresenta valores zerados para ambas as variáveis.

5.2 Análise de componentes Principais

Para uma análise inicial, foi realizada uma matriz de correlação das variáveis com as primeiras cinco dimensões:

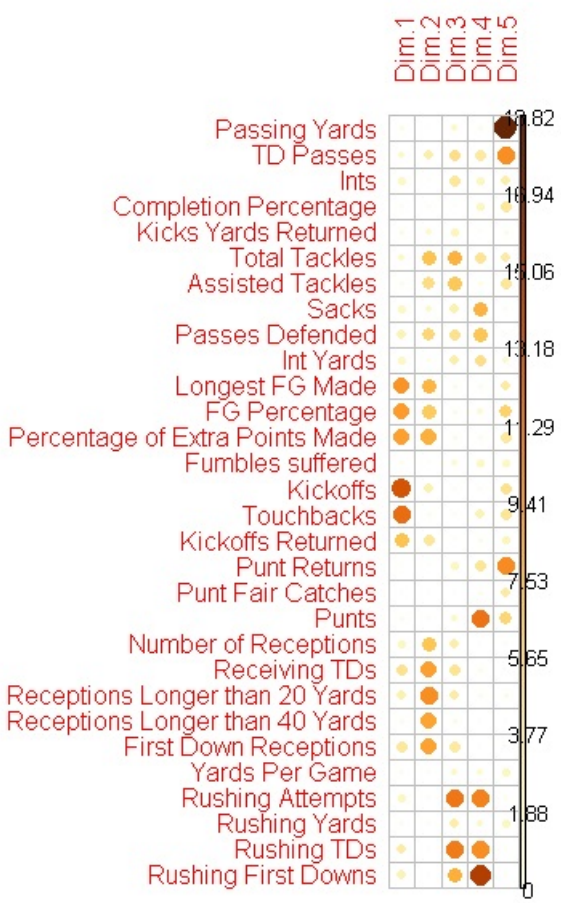

Figura 5.7: Contribuição das Variáveis para os Componentes Principais.

Pela Figura 5.7, nota-se que as medidas associadas a *Longest FG Made*, *FG percentage* e *Percentage of Extra Points made* são bem representadas pelas dimensões 1 e 2. As medidas associadas a *Rushing Attempts*, *Rushing TDs* e *Rushing First Downs* são bem representados pelas dimensões 3 e 4. Em particular, nota-se que o primeiro componente representa melhor as características relacionadas aos times de especialistas, já o terceiro componente em particular é bem representado por características de jogadas terrestres. Em especial, a variável *Passing Yards* tem um grande representatividade apenas na dimensão 5.

Para uma melhor análise, pode-se checar o *Scree Plot* (gráfico de cotovelo) dos componentes, com este conseguimos checar a representatividade dos dados por dimensão e assim definir o número final de dimensões que será escolhida:

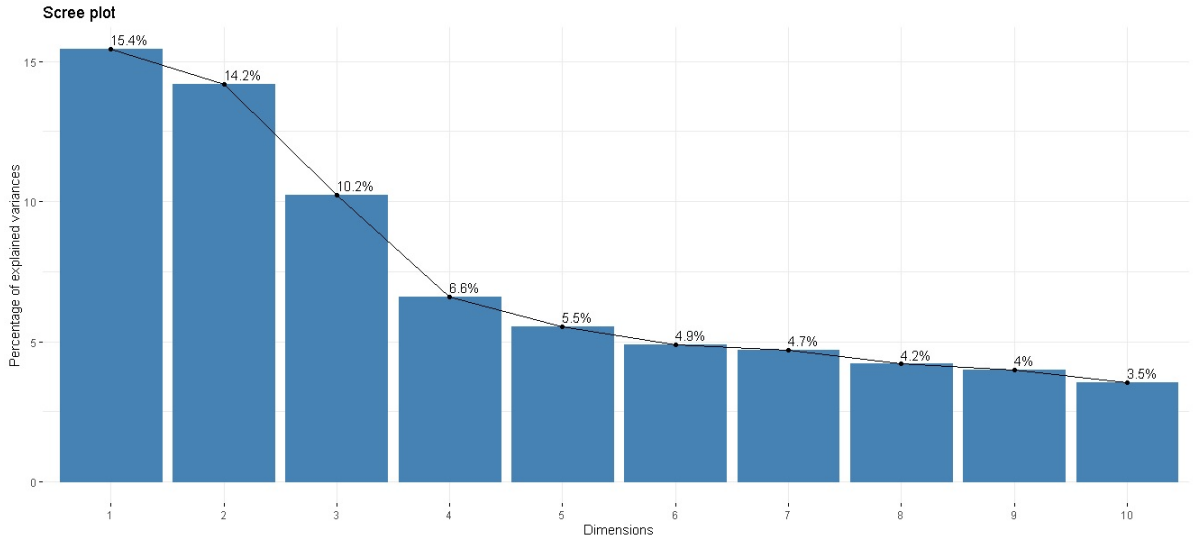

Figura 5.8: Scree Plot

Observa-se pelo *Scree Plot* a formação de um cotovelo no terceiro componente, isto é, a partir da quarta dimensão a variação entre as dimensões são pequenas, indicando que a escolha de três componentes pode ser adequada, tendo uma porcentagem variância total acumulada de aproximadamente 40%. Além disso, os primeiros três componentes possuem as maiores significâncias, principalmente os dois primeiros com porcentagem acentuadamente maior que os demais componentes posteriores. Apesar disso pode-se analisar como se dá a representatividade de cada variável com relação a cada componente escolhido, buscando entender de que forma cada variável é explicada pelo conjunto de componentes com maior variabilidade total explicada.

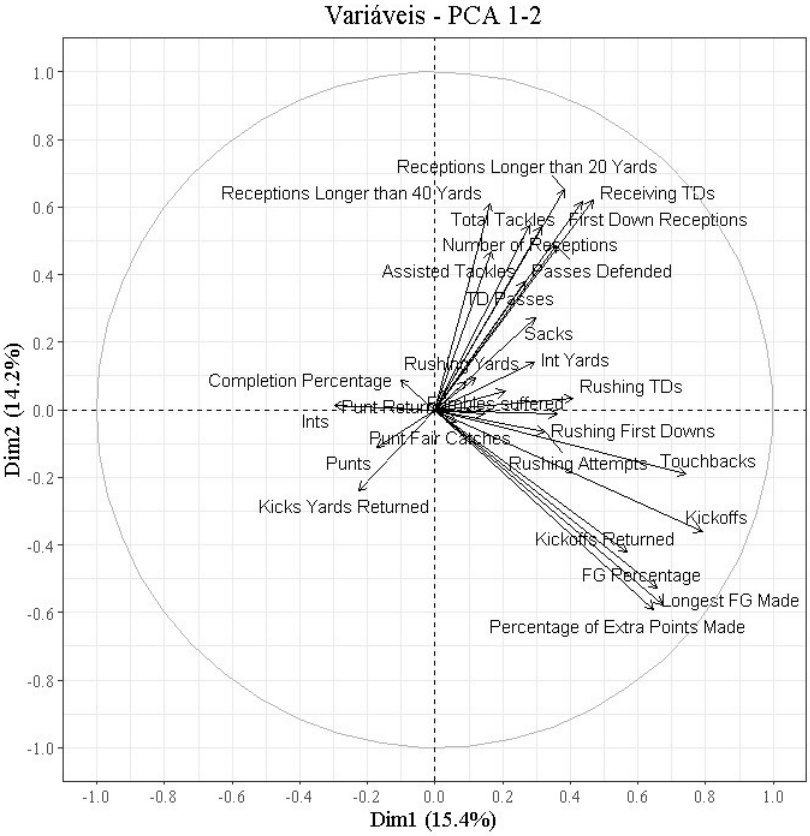

Figura 5.9: Círculo de correlação planos (Dim1 x Dim2).

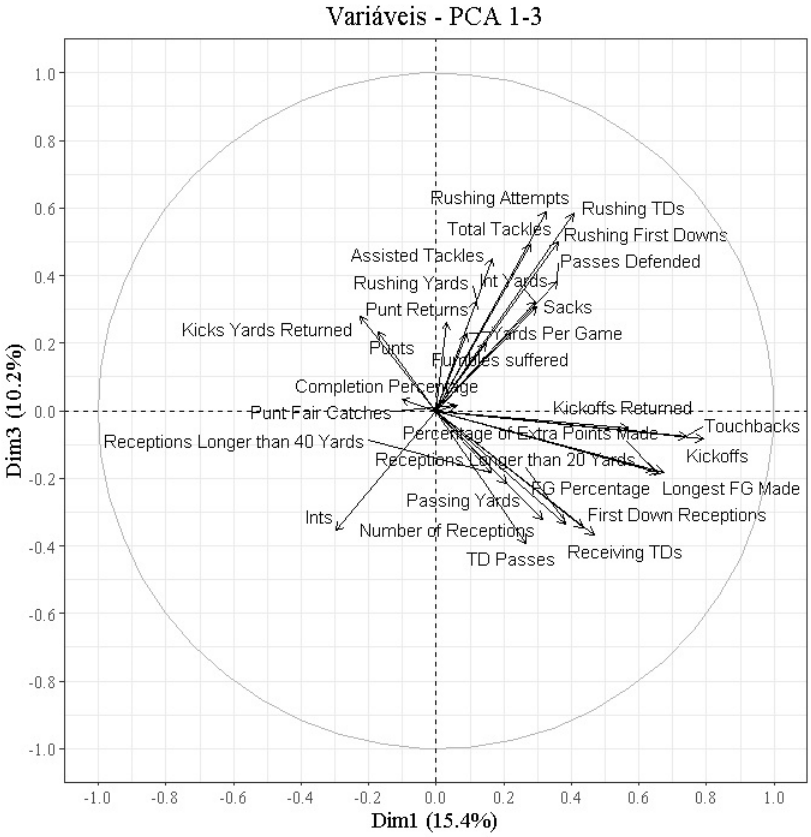

Figura 5.10: Círculo de correlação planos (Dim1 x Dim3).

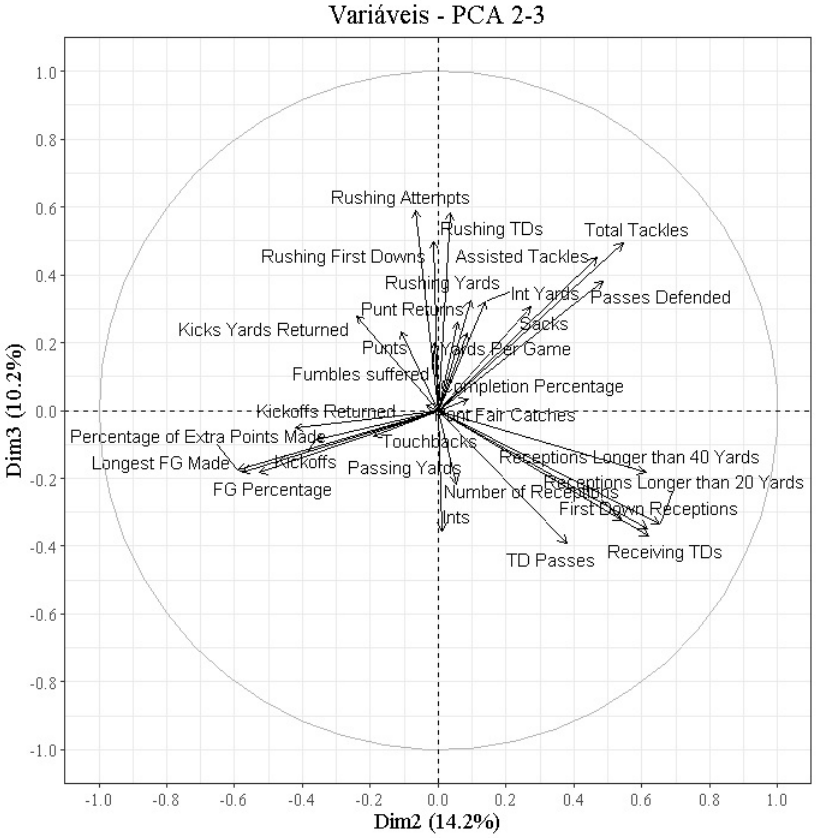

Figura 5.11: Círculo de correlação planos (Dim2 x Dim3).

Pelo Círculo de Correlação do Primeiro Plano (Dim1 x Dim2) é possível distinguir que as variáveis referentes time de especialistas não são bem explicadas por esse plano, ficando em grande maioria separadas das características de ataque e defesa. As variáveis de ataque como *Longas Recepções*, *TD recebidos* e *Primeiras decidas recebidas* são melhores representadas, em comparação ao primeiro plano, assim como as variáveis de defesa como *Passes defendidos*, *Tackles (assistência e total).*

Podemos observar pelo Segundo Plano (Dim1 x Dim3) que a variável *Punt retornado* é bem representada pelo segundo plano, pois está bem próxima ao círculo unitário. As variáveis relacionadas à *jogadas terrestres* estão sendo melhor representadas em comparação ao primeiro plano, porém *Interceptações* passam a ser mal representadas nesse plano.

Em relação ao Círculo unitário do Terceiro Plano (Dim2 x Dim3), temos que as variáveis relacionadas à *Jogadas terrestres*, próximas do círculo unitário, desta forma, bem representadas pelo Primeiro Plano, em contraponto, as *Interceptações* são representadas negativamente por esse plano. As variáveis relacionadas ao time de defesa *Tackles, Passes defendidos, Sacks* e *Yards por Interceptações* estão sendo melhor representadas em

comparação ao Segundo Plano.

Pelo fato das características do time de especialistas estar representado separadamente das características dos outros times, será feita uma análise das características desse time separadamente.

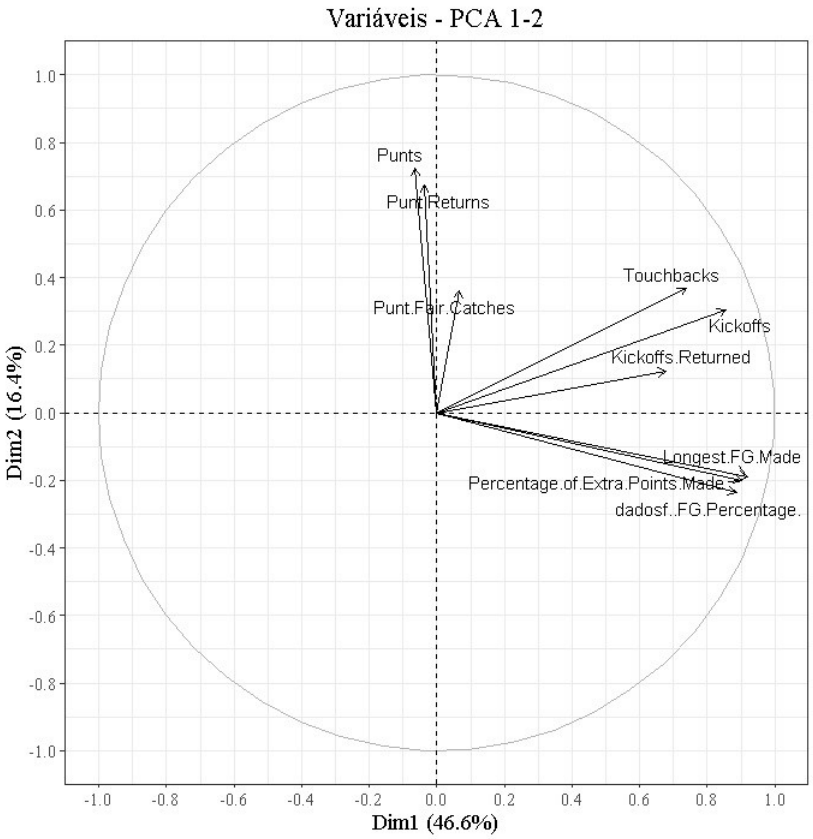

Figura 5.12: Círculo de correlação planos do time de especialistas (Dim1 x Dim2).

Ao observar o Primeiro Plano (Dim1 x Dim2), temos as variáveis relacionas à $Punts$ bem representadas, pois são próximas ao círculo unitário. Já as variáveis relacionas à $Kickoffs$ são melhores representadas em comparação ao Primeiro Plano.

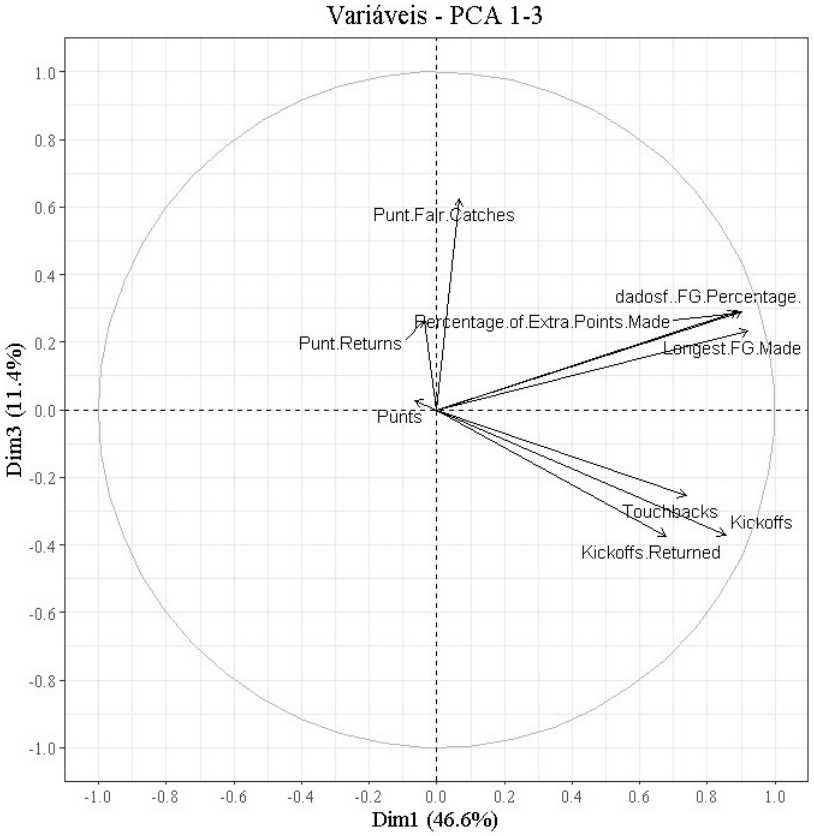

Figura 5.13: Círculo de correlação planos do time de especialistas (Dim1 x Dim3).

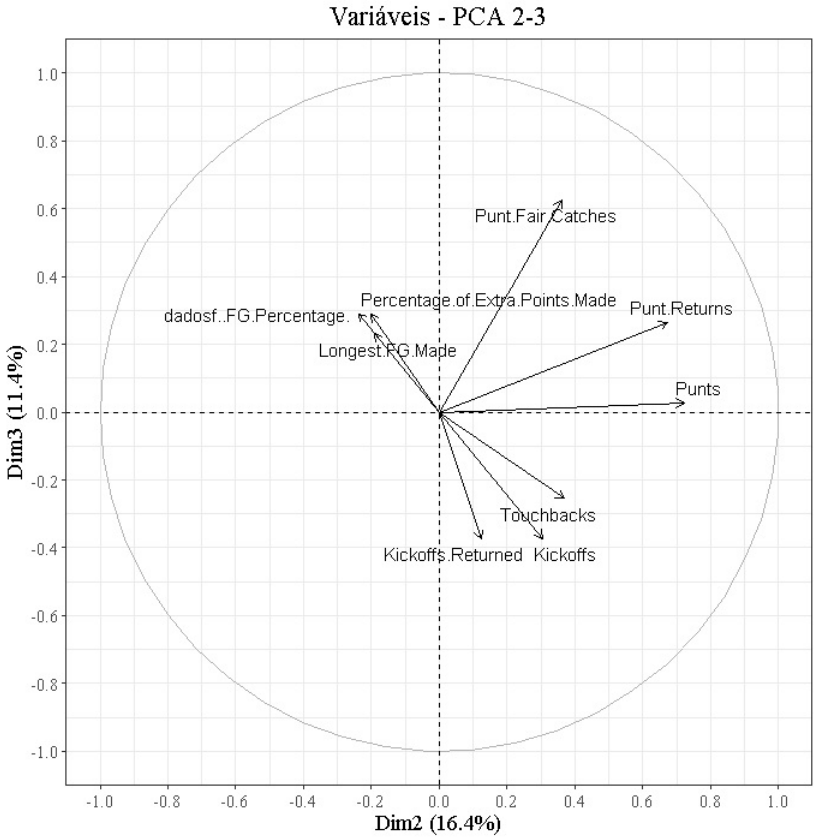

Figura 5.14: Círculo de correlação planos do time de especialistas (Dim2 x Dim3).

No Segundo Plano (Dim1 x Dim3) vemos as variáveis relacionadas a *Punts* sendo bem representadas, já que estão próximas ao círculo unitário. Já as variáveis relacionadas a *FG* e *Extra Points* melhores representadas em relação ao Primeiro Plano.

Pelo Terceiro Plano (Dim2 x Dim3), temos as variáveis relacionadas a *Punts* são melhores representados em relação ao Segundo Plano.

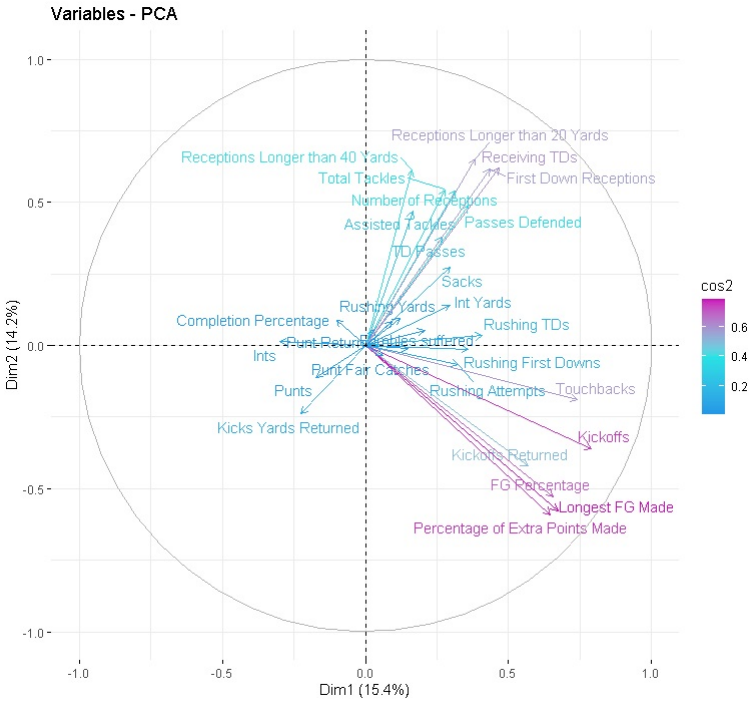

Figura 5.15: Escores fatoriais no plano (Dim 1, Dim 2) com a qualidade de representação de cada observação.

Pala Figura 5.15, nota-se que as observações mais próximas das médias possuem menores representações, enquanto observações mais acima ou abaixo da média para cada componente possuem melhor representatividade. Ou seja, observações com características de recepções (*Receptions Longer than 20 Yards*, *Receiving TDs* e *First Down Receptions*) tendem a ser melhor representadas que observações mais próximas da média como *Sacks*, *Interceped Yards* e *Rushing TDs*.

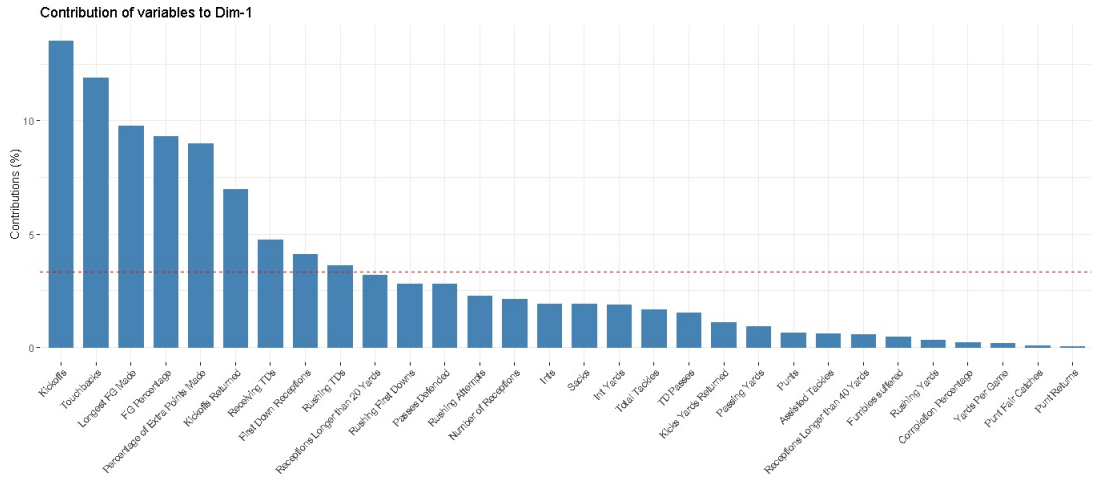

Figura 5.16: Contribuição das variáveis pela primeira dimensão.

Nota-se pela Figura 5.16 primeiramente o fato do primeira dimensão possuir nove variáveis com alta qualidade, estas são *Kickoffs*, *Touchbacks*, *Longest FG Made*, *FG percentage*, *Percentage of Extra Points made*, *Kickoffs returned*, *Receiving TDs*, *First Down Receptions*, *Rushing TDs* de representação. As demais variáveis possuem qualidade de representação pequena. Logo, conclui-se que o primeiro plano em geral, possui alta qualidade de representação para um grande número de variáveis. Chama atenção que seis das nove variáveis são relacionadas ao time de especialistas.

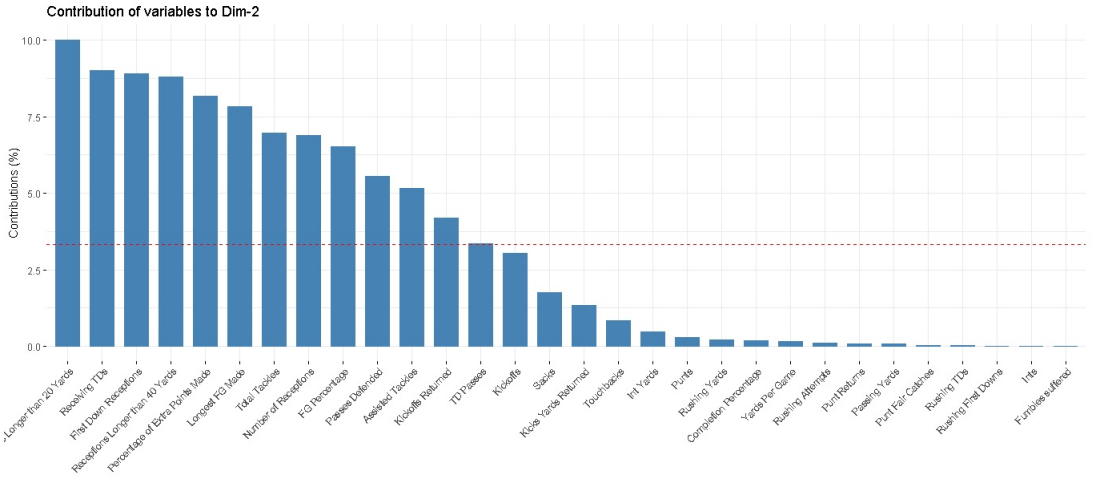

Figura 5.17: Contribuição das variáveis pela segunda dimensão.

Pela Figura 5.17 é possível ver que o segunda dimensão possui 13 variáveis com alta qualidade de representação. As demais variáveis possuem qualidade de representação pequena. Desta forma, conclui-se que o segundo plano em geral, possui alta qualidade de representação para um grande número de variáveis. Diferentemente do primeiro plano, neste já possuí variáveis referentes aos times de ataque, defesa e de especialistas com alta

qualidade representativa.

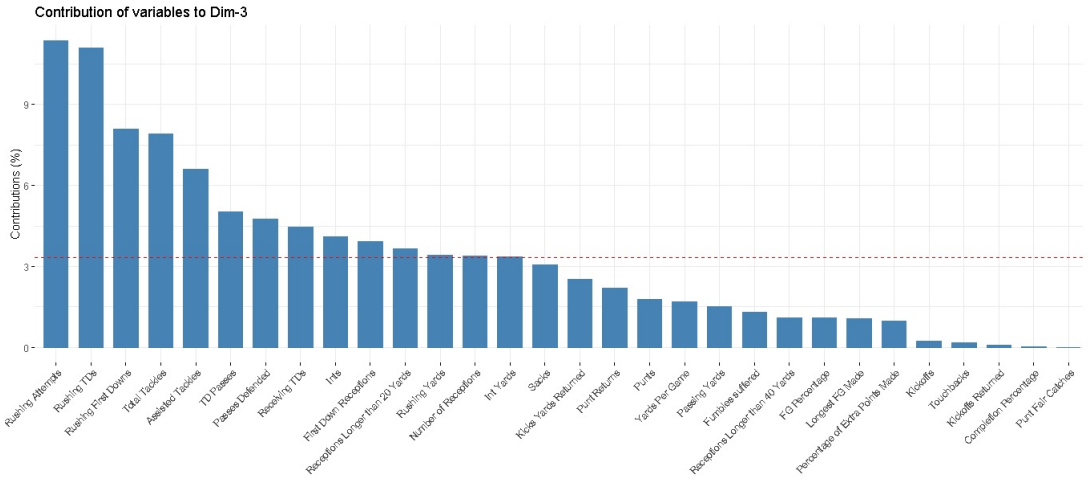

Figura 5.18: Contribuição das variáveis pela terceira dimensão.

Através da Figura 5.18, nota-se que que o terceira dimensão possui uma variáveis a mais com qualidade representativa do que o plano anterior, isto é, o terceiro plano possuí 14 variáveis com alta qualidade de representação. E diferentemente dos dois primeiros planos, neste não possuí nenhuma variável relacionadas ao time de especialistas.

Desta forma, chegamos aos três componentes principais:

- Componente principal 1 é dado por: $Kickoffs$, *Touchbacks*, *Longest FG Made*, *FG percentage*, *Percentage of Extra Points made*, *Kickoffs returned*, *Receiving TDs*, *First Down Receptions*, *Rushing TDs*.

- Componente principal 2 é dado por: *Recptions Longer than 20 Yards*, *Receiving TDs*, *First Down Receptions*, *Recptions Longer than 40 Yards*, *Porcentage of Extra Points Made*, *Longest FG Made*, *FG Porcentage*, *Number of Receptions*, *Total Tackles*, *Passes Defended*, *Assisted Tackles*, *Kickoffs Returned*, *TD Passes*.

- Componente principal 3 é dado por: *Rushing Attempts*, *Rushing TDs*, *Total Tackles*, *Rushing First Downs*, *Assisted Tackles*, *TD Passes*, *Passes Defended*, *Receiving TDs*, $Ints$, *First Down Receptions*, *Recptions Longer than 20 Yards*, *Rushing Yards*, *Int Yards*, *Number of Receptions*, *Sacks*.

De todas as 31 variáveis do conjunto de dados, foram usadas 22 variáveis para a definição dos três Componentes Principais resultantes, são elas:

(1) *TD Passes*
(2) *Ints*
(3) *Total Tackles*
(4) *Assisted Tackles*
(5) *Sacks*
(6) *Passes Defended*
(7) *Int Yards*
(8) *Longest FG Made*
(9) *FG percentage*
(10) *Percentage of Extra Points made*
(11) $Kickoffs$
(12) *Touchbacks*
(13) *Kickoffs returned*
(14) *Number of Receptions*
(15) *Receiving TDs*
(16) *Recptions Longer than 20 Yards*
(17) *Recptions Longer than 40 Yards*
(18) *First Down Receptions*
(19) *Rushing Attempts*
(20) *Rushing Yards*
(21) *Rushing TDs*
(22) *Rushing First Downs*

5.3 Análise de Agrupamentos

Para descobrirmos quais variáveis encontradas pelo método de Componentes Principais estão relacionadas, iremos utilizar primeiro o método hierárquico aglomerativo de $Ward$ e após decidirmos um número de agrupamentos finais, utilizaremos o método não-hierárquico para agrupar as variáveis.

5.3.1 Método Hierárquico

Para identificar a quantidade de agrupamentos finais que serão utilizados posteriormente no método não-hierárquico, utilizaremos para tal o método hierárquico aglomerativo da distância de $Ward$.

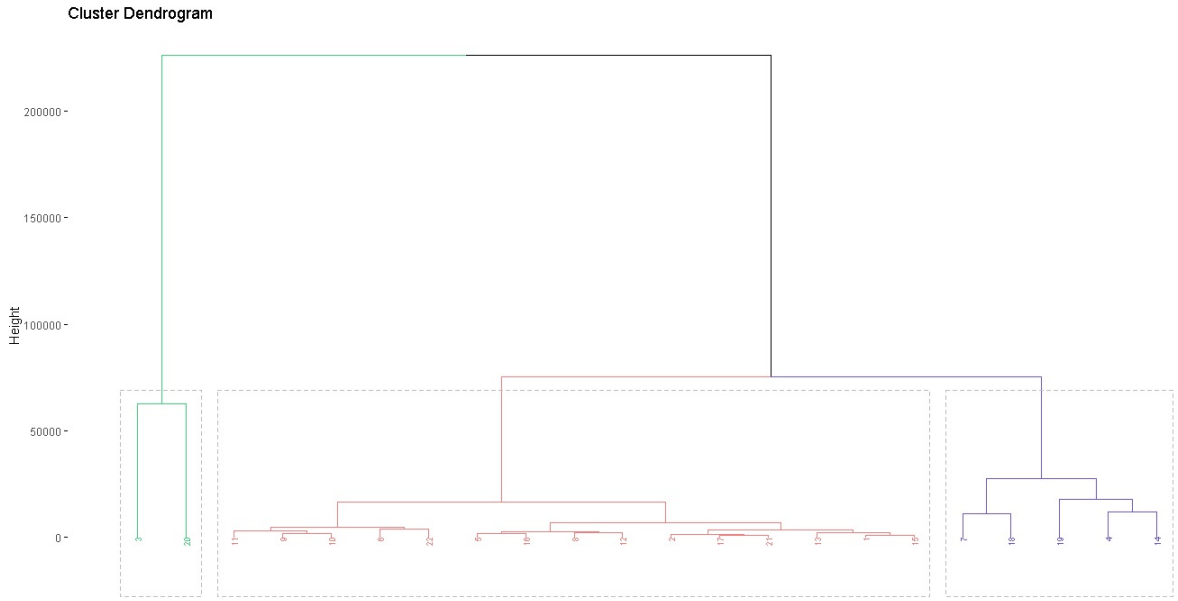

Figura 5.19: Agrupamentos de acordo com Método de $Ward$.

Pela Figura Figura 5.19, temos um indicativo de que um número final $k = 3$ seria adequado, no entanto para averiguar, utilizaremos alguns gráficos do método de $Ward$ que indicam a melhor quantidade de grupos finais.

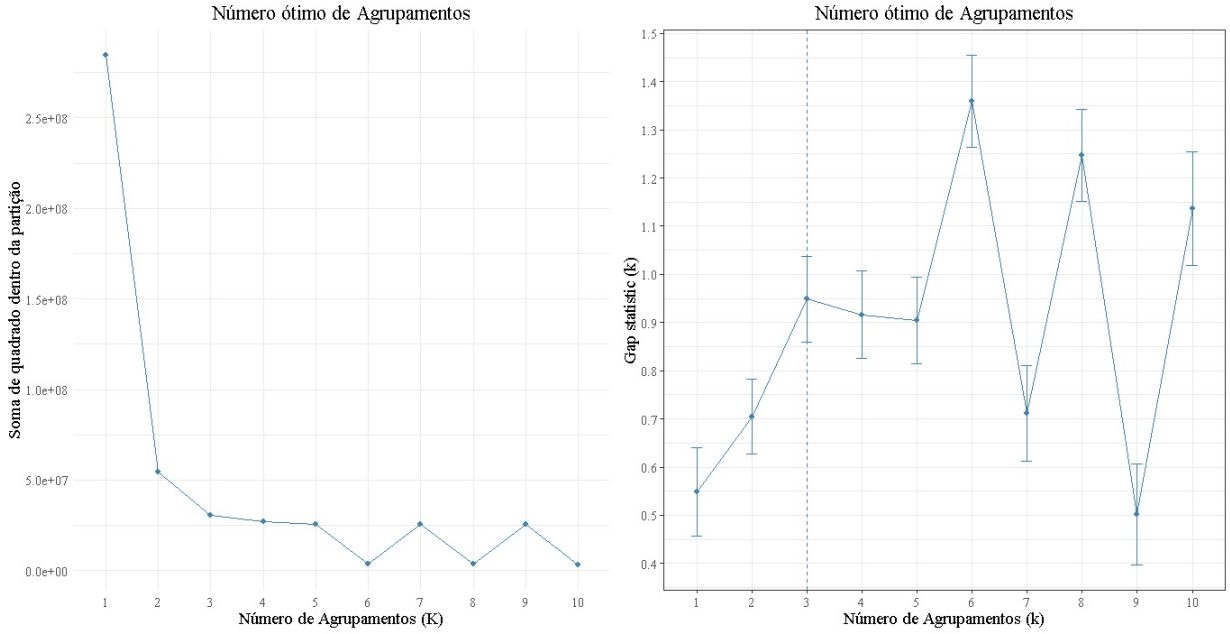

Figura 5.20: Número ótimo de agrupamentos.

Pela Figura Figura 5.20, podemos ver que o número ótimo de agrupamentos finais seria entre $k = 2$ e $k = 3$, ao escolher um número final de três grupos, os resultados se tornam mais palatáveis e a obtenção de informações sobre os resultados mais rica. Desta forma, pelo método hierárquico de $Ward$, escolhemos $k = 3$.

5.3.2 Método Não-hierárquico

A partir do número ótimo de agrupamentos encontrados anteriormente $k = 3$, agora podemos aplicar o método não-hierárquico de *K-médias* para agrupar as variáveis dos Componentes Principais.

Ao utilizar o método do *K-médias* obtivemos os seguinte grupos:

Tabela 5.4: Quantidade e quais variáveis estão presentes em cada agrupamento.

Grupos	**Variáveis**	**Quantidade de variáveis**
1	1,2,5,6,8,9,10,11,12,13,15,16,17,21,22	15
2	3,20	2
3	4,7,14,18,19	5

Em seguida a descrição de quais são essas variáveis, de cada agrupamento:

- Grupo 1: *TD Passes*, *Ints*, *Sacks*, *Passes Defended*, *Longest FG Made*, *FG percentage*, *Percentage of Extra Points made*, $Kickoffs$, *Touchbacks*, *Kickoffs returned*, *Receiving TDs*, *Recptions Longer than 20 Yards*, *Recptions Longer than 40 Yards*, *Rushing TDs*, *Rushing First Downs*.

- Grupo 2: *Total Tackles*, *Rushing Yards*.

- Grupo 3: *Assisted Tackles*, *Int Yards*, *Number of Receptions*, *First Down Receptions*, *Rushing Attempts*.

Podemos perceber pela Tabela 5.4 que os grupos dois e três possuem quantidade de variáveis muito diferente do primeiro grupo, este engloba a maior parte das variáveis, é importante ressaltar que os grupos finais foram os mesmos tanto no método hierárquico de *Ward* quanto ao método não-hierárquico de *K-médias*.

Ao analisar os grupos finais, temos o segundo com duas variáveis, elas se assemelham no aspecto que uma *jogada terrestre* só acaba com um *tackle*.

Com relação ao terceiro grupo temos variáveis que tem significância secundária, isto é, geralmente em um jogo o que é valorizado é o *tackle*, a *interceptação* em si, e quantidade de jardas percorridas que resultam em pontuações, ou seja, é valorizado o resultado final, e não os motivos que foram possível chegar a esse resultado.

No primeiro grupo, temos variáveis primárias em sua maioria, isto é, variáveis que são importantes sozinhas, que mostram a força e constância da equipe. São variáveis que definem o resultado de um jogo, e por isso valorizadas.

As análises anteriores são importantes para entendermos quais variáveis são significantes através do método estatístico de Componentes Principais, dentre todas as variáveis do banco de dados. Além, das composições de cada grupo final para entendermos características em comum através do método de Agrupamentos não-hierárquico de *K-médias*.

Capítulo 6

Conclusão e Trabalhos Futuros

Com esse Trabalho de Conclusão de Curso, analisamos, selecionamos as características com maior contribuição para explicar a variabilidade dentro do conjunto de dados e, agrupamos as caraterísticas relevantes de acordo com a similaridade e a relações dessas características entre os três times presentes em cada equipe da ***NFL*** obter bons resultados, através de métodos estatísticos.

Para começar a realizar o estudo, foi necessário a montagem do banco de dados de acordo com os anos definidos, 2013 a 2016. Além da incorporação dos dados nas equipes, antes divididos por jogadores. Fazendo todos os tratamentos dos dados, analisando individualmente cada variável e excluindo, quando duplicadas ou somando quando complementares. Resultando em 31 variáveis com 128 linhas de observações.

Com o conjunto de dados finalizada, foram feitas as análises sobre este. Pela análise descritiva, foi possível verificar o comportamento das variáveis de acordo com cada um dos três times (ataque, defesa e especialistas).

Já na análise descritiva, foi possível identificar que o time de especialistas pareciam ter valores discrepantes dos times de ataque e defesa, dando indicativos de que essas características seriam diferenciadas na análise.

Ao realizar o método de Componentes Principais, tivemos um resultado que mostrou que as variáveis relacionadas ao time de especialistas era realmente diferenciadas dos outros dois times, sendo interessante a análise destes separadamente.

Ao final da utilização do ACP, obtivemos três componentes principais, estes utilizavam 22 variáveis das 31 originais. Em que dois deles, características do time de especialistas eram de maior significância.

Através da ACP, foram utilizadas as variáveis encontradas para aplicar o método

de agrupamentos. Realizamos primeiro o método hierárquico aglomerativo pelo método da distância de $Ward$ para determinar o número de grupos finais e com esse valor $k = 3$, foi aplicado o método de agrupamentos não-hierárquico pelo método *K-médias* .

Os agrupamentos finais, mostraram um grupo com características de primeira importância, outro com as secundárias e no último duas variáveis relacionadas entre elas desse grupo.

Essas foram os resultados que foram possíveis de chegar para o final desse Trabalho de Conclusão de Curso. Através desse trabalho pode-se analisar a importância de cada time e o quanto cada um influenciam, como por exemplo a importância do time de especialistas.

Pelos componentes principais, tem-se uma ideia das principais características que os técnicos deveriam dar atenção principal. E olhando para os agrupamentos finais, como podem ser trabalhadas essas características conjuntamente, para que assim as equipes tenham melhores resultados.

Para trabalhos futuros, seria interessante realizar a mesma análise com mais dados e comparar com outros ciclos de quatro anos, analisando se as características mais interessantes variam de acordo com os ciclos estudados. Poderia estudar também a diferença entre temporadas por alteração de alguma regra, como o aumento de uma semana no calendário de jogos.

Referências Bibliográficas

Barroso, L. P. e Artes, R. (2003). Análise multivariada. *Universidade Federal de Lavras*, **1st ed**.

da Costa Moraes, M. B. (2016). Análise multivariada aplicada à contabilidade. <https://edisciplinas.usp.br/pluginfile.php/2204134/mod_resource/content/1/An%C3%A1liseMultivariada-Aula11.pdf>. Acesso em: 30 jan. 2022.

Dorfman, S. (2021). Nfl players tackle mental health in league's latest campaign. <https://www.palmbeachpost.com/story/lifestyle/2021/05/23/nfl-debuts-mental-health-initiative/5115679001/>. Acesso em: 28 jan. 2022.

dos Anjos, A. (2018). Análise de componentes principais. <https://docs.ufpr.br/~aanjos/SENSOMETRIA/slides/ACP.pdf>. Acesso em: 30 jan. 2022.

Garcia, M. (2017). Como as divisões da nfl foram montadas? <https://ligados32.lance.com.br/como-as-divisoes-da-nfl-foram-montadas/>. Acesso em: 28 jan. 2022.

Hair, J. F. e. a. (2009). *Multivariate Data Analysis: A Global Perspective.*, volume 7th ed. Upper Saddle River: Prentice Hall.

Izbicki, R. e dos Santos, T. M. (2020). *Aprendizado de máquina: uma abordagem estatística*. Rafael Izbicki.

JovemPan (2017). Além da final: shows e experiências com fãs garantem lucro bilionário à nfl. <https://jovempan.com.br/esportes/alem-da-final-shows-e-experiencias-com-fas-garantem-lucro-bilionario-nfl.html>. Acesso em: 28 jan. 2022.

Kendallgillies (2017). Nfl statistics. <https://www.kaggle.com/datasets/kendallgillies/nflstatistics/metadata?select=Basic_Stats.csv>. Acesso em: 10 dez. 2021.

Pedro Ferreira, F. (2020). Estatística multivariada 2. **Análise de Componentes Principais; Análise de Agrupamentos (Cluster Analysis)**.

Rocha, E. (2020). Entendendo de vez o que é pca - principal component analysis. <https://www.youtube.com/watch?v=p4bvCFygfW0>. Acesso em: 28 fev. 2022.

SAATE, S. d. A. a. A. d. T. E. (2007). Análise de cluster. <http://www5.eesc.usp.br/saate/index.php/saate/Indicar-a-T%C3%A9cnica/Associar/2.-%C3%81rvore-de-decis%C3%A3o/Gloss%C3%A1rio/An%C3%A1lise-de-Cluster>. Acesso em: 30 jan. 2022.

Souza, E. F. d. (2007). *Comparação e escolha de agrupamentos: uma proposta utilizando a entropia*. Ph.D. thesis, Universidade de São Paulo.

Trecsson, B. S. (2016). Análise de cluster: O que é e como aplicar? <https://www.trecsson.com.br/blog/planejamen/analise-de-cluster>. Acesso em: 30 jan. 2022.

Apêndice A

Apêndice A

Tabela A.1: Tabela com descrição dos símbolos utilizados.

Símbolos	**Descrição**
$\Sigma_{p\times p}$	matriz var-cov
σ_{ij}	elementos da matriz $\Sigma_{p\times p}$
$\boldsymbol{p}$	quantidade de variáveis de um conjunto ACP
$R_{p\times p}$	matriz de correlação
ρ_{ij}	fórmula de cálculo de $\rho =$ é $\frac{\sigma_{ij}}{\sqrt{\sigma_i^2\sigma_j^2}}$
k	quantidade de variáveis não correlacionadas
P_A	polinômio característico
λ	autovalores
I_p	matriz identidade de ordem p
(X_1,X_2, ..., X_p)	variáveis iniciais
(Y_1,Y_2, ..., Y_p)	variáveis transformadas
det	determinante de matriz
K	quantidade de agrupamentos
D	matriz de similariedade
d_{ij}	valores das distâncias entre observações