Relatório final projeto de Computação para Análise de Dados.

Aluno: Paulo Henrique Calado Aoun

1. Introdução

A National Football League (NFL) é a liga mais poderosa do mundo, com uma renda anual de 13 bilhões de dólares americanos, e com suas 32 franquias figurando entre os 50 times esportivos mais ricos do mundo. A liga é dividida em 2 conferências, cada conferência com 16 times, e essas 2 conferências são divididas em 4 divisões com 4 times cada.

Neste trabalho iremos focar no desempenho dos 4 times (Tampa Bay Buccaneers, New Orleans Saints, Atlanta Falcons e Carolina Panthers) da NFC South (uma das 4 divisões da conferência NFC) que é disparadamente a divisão mais disputada de toda NFL, sendo a única que teve 3 times indo para a pós-temporada da liga.

Futebol Americano é um jogo extremamente estratégico, com padrões e tendências valiosas escondidas por trás dos números de cada time, esse trabalho tem o objetivo de analisar algumas das principais tendências desses times, e o seu comportamento durante os jogos através da criação de gráficos e análises estatísticas.

Um time de futebol americano é formado basicamente por três diferentes times, que são chamados de Ataque, Defesa e Times Especiais (Offense, Defense e Special Teams respectivamente). Quando o Ataque de um time está em campo, a Defesa do time adversário entra e os Times Especiais entram em situações específicas como na hora do chute de Kickoff ou Field Goal por exemplo.

Este trabalho visa entender algumas características dos times de Ataque dos times da NFC South citados acima. Entender a movimentação e os padrões do time adversário, assim como tendência dos principais jogadores, é um fator crucial que pode garantir uma importante vitória em um campeonato tão curto e disputado como é a NFL.

2. Pacotes requeridos

Para o projeto foram utilizados os seguintes pacotes:

- library(dplyr) : Utilizada para auxiliar na manipulação dos dados.
- library(tidyr) : Também utilizada para auxiliar na manipulação dos dados.
- library(ggplot2): Utilizada para auxiliar na criação de gráficos.

3. Preparação dos dados

Os dados para este trabalho foram obtidos a partir deste repositório no GitHub. Estes dados são referentes a todas as jogadas de todos os jogos da temporada regular de 2017 da NFL, e foram obtidos diretamente da API da NFL. O dataset possui 100 variáveis, de diferentes tipos, numéricas, textuais e até mesmo binárias (0 ou 1), os valores ausentes no dataset já são definidos como NA.

O primeiro para a análise dos dados é importar o arquivo csv para o ambiente do RStudio.

As principais variáveis utilizadas no trabalho foram:

- Posteam:
  - Tipo: Character
  - Indica qual o time que está no ataque.
- PlayType:
  - Tipo: Character
  - Indica qual o tipo de jogada que ocorreu como por exemplo, Pass, Run, Kickoff e etc.
- Yards.Gained:
  - Tipo: integer
  - Indica o número de jardas ganha na jogada.
- RunLocation:
  - Tipo: Character
  - Indica para qual lado do campo a corrida aconteceu, left, middle ou right.
- Rusher:
  - Tipo: Character
  - Indica o nome do corredor.
- Passer:
  - Tipo: Character
  - Indica o nome do lançador.
- Receiver:
  - Tipo: Character
  - Indica o nome do recebedor.
- PassOutcome:
  - Tipo: Character
  - Indica se o passe foi completo ou incompleto
- PassLocation:
  - Tipo: Character

    - Indica para qual lado do campo o passe foi tentado, left, middle ou right.
- Qtr:
    - Tipo: Integer
    - Indica qual o quarto do jogo.
- RushAttempt:
    - Tipo: Binária
    - Indica se uma corrida foi tentada na jogada.
- RunGap:
    - Tipo: Character
    - Indica qual o Gap da linha ofensiva que a corrida ocorreu.
- Time:
    - Tipo: Character
    - Indica o tempo de jogo.
- Win_Prob:
    - Tipo: numérico
    - Indica a probabilidade de vitória do time com a posse de bola.
- ScoreDiff:
    - Tipo: numérico
    - Indica a diferença no placar do time com a posse de bola em relação ao adversário.
- Yrdline100:
    - Tipo: numérico
    - Indica a distância para a Endzone do adversário.
- Opp_Touchdown_Prob:
    - Tipo: numérico
    - Indica a probabilidade do time adversário marcar um Touchdown
- Field_Goal_Prob:
    - Tipo: numérico
    - Indica a probabilidade de o time com a posse marcar um field goal
- FieldGoalDistance:
    - Tipo: numérico
    - Indica a distância do chute de Field Goal.

A primeira abordagem é transformar o campo Time em duas outras variáveis, minutos e segundos, para uma melhor e mais fácil análise, assim como transformar o tipo dele para numérico.

```
#importando dados
nflData <- read.csv("/Users/paulocalado/Documents/UFRPE/analiseDados/pbp_2017_huge.csv",
                    header = T,stringsAsFactors = F)
#separando a coluna time em minutes e seconds
nflData<-nflData%>%
          separate(time,
                   into = c("minutes","seconds"),
                   sep=':')
#como só a coluna minutes me interessa, só transformo ela para numérico
nflData$minutes<- as.numeric(nflData$minutes)
```

Feito isso, é necessário então separar os dados em que os times da NFC South estão no ataque, já que este é o foco do trabalho como dito anteriormente. Abaixo temos como ficou o código para essa tarefa.

```
#separando apenas os times que focarei no trabalho
nfcSouthData <- nflData[nflData$posteam == "TB"|
                        nflData$posteam == "NO"|
                        nflData$posteam == "ATL"|
                        nflData$posteam == "CAR", c("posteam","qtr","minutes","PlayType","Yards.Gained",
                                                    "RunLocation","Rusher","Passer","Receiver",
                                                    "PassOutcome","PassLocation",
                                                    "RushAttempt","RunGap")]
```

Ao rodar esse código, temos a tabela a seguir com os dados ofensivos de todos os times da NFC South.

	posteam	qtr	minutes	PlayType	Yards.Gained	RunLocation	Rusher	Passer	Receiver	PassOutcome	PassLocation	RushAttempt	RunGap
1254	ATL	1	15	Kickoff	63	NA	NA	NA	NA	NA	NA	0	NA
1255	ATL	1	14	Run	6	right	D.Freeman	NA	NA	NA	NA	1	end
1256	ATL	1	14	Pass	4	NA	NA	M.Ryan	M.Sanu	Complete	left	0	NA
1257	ATL	1	13	Pass	19	NA	NA	M.Ryan	Ju.Jones	Complete	middle	0	NA
1258	ATL	1	12	Pass	14	NA	NA	M.Ryan	T.Gabriel	Complete	right	0	NA
1259	ATL	1	11	Run	3	left	T.Coleman	NA	NA	NA	NA	1	end
1260	ATL	1	11	Pass	3	NA	NA	M.Ryan	T.Coleman	Complete	middle	0	NA
1261	ATL	1	10	Run	5	left	D.Freeman	NA	NA	NA	NA	1	tackle
1262	ATL	1	10	Pass	7	NA	NA	M.Ryan	M.Sanu	Complete	right	0	NA
1263	ATL	1	9	Run	2	right	D.Freeman	NA	NA	NA	NA	1	tackle
1264	ATL	1	8	Run	-1	right	T.Coleman	NA	NA	NA	NA	1	tackle
1265	ATL	1	8	Field Goal	0	NA	NA	NA	NA	NA	NA	0	NA
1272	ATL	1	5	Pass	0	NA	NA	M.Ryan	T.Coleman	Incomplete Pass	left	0	NA
1273	ATL	1	5	Run	4	right	D.Freeman	NA	NA	NA	NA	1	tackle
1274	ATL	1	4	Pass	0	NA	NA	M.Ryan	M.Sanu	Incomplete Pass	right	0	NA
1275	ATL	1	4	Punt	14	NA	NA	NA	NA	NA	NA	0	NA
1287	ATL	2	14	Kickoff	0	NA	NA	NA	NA	NA	NA	0	NA
1288	ATL	2	14	Run	5	left	T.Coleman	NA	NA	NA	NA	1	guard
1289	ATL	2	14	Run	0	right	T.Coleman	NA	NA	NA	NA	1	tackle
1290	ATL	2	13	Pass	0	NA	NA	M.Ryan	M.Sanu	Incomplete Pass	middle	0	NA

4. Análise exploratória dos dados.

Esta seção tem o objetivo de mostrar o que pode ser descoberto a partir da análise dos dados do dataset da temporada regular de 2017 da NFL.

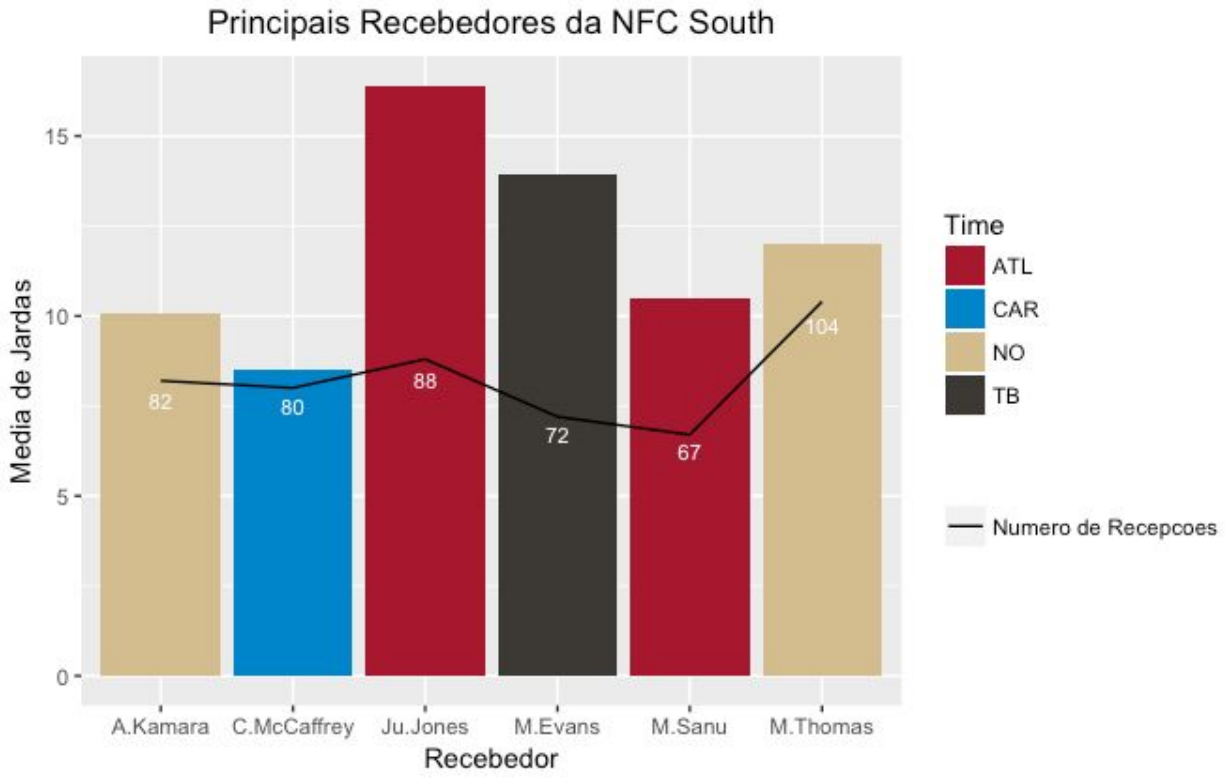
Nossa primeira análise visa procurar quais as principais “armas aéreas” da divisão(NFC South), ou seja, quais atletas que mais receberam passe na conferência, e seus respectivos números. Para chegar nesse atleta obviamente é preciso fazer uma manipulação dos dados, que é conseguida através do seguinte código:

```
#verificando média de jardas ganhas de cada recebedor da NFC South
nfcSouthReceivers <- nfcSouthData %>%
  group_by(Receiver,posteam) %>%
  filter(PlayType == "Pass" & PassOutcome=="Complete") %>%
  summarise(
    meanYardsGained = mean(Yards.Gained),
    numReception = sum(PlayType=="Pass"&PassOutcome=="Complete")
  )

#Ordenando para os recebedores com maior número de passes recebidos
nfcSouthReceivers<- arrange(nfcSouthReceivers,desc(numReception))
```

A partir disso, é possível montar o seguinte gráfico:

Gráfico 1



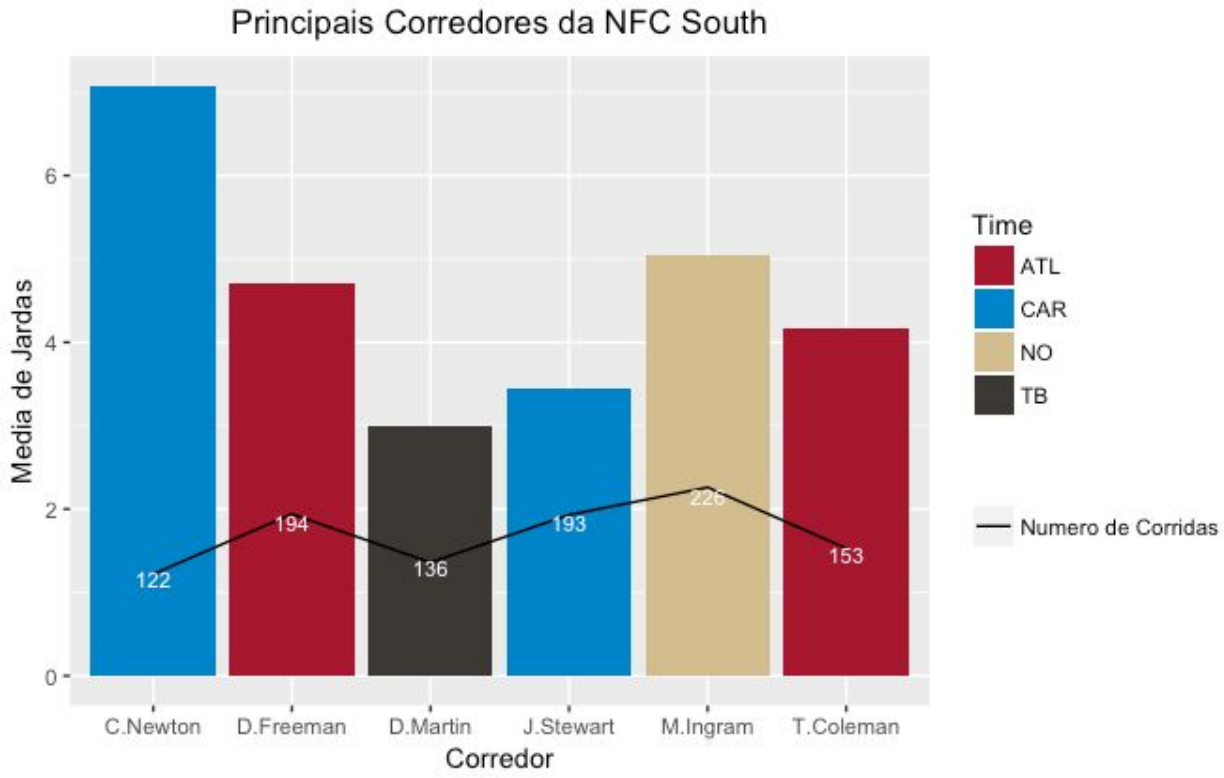
É possível tirar observações bastante relevantes deste gráfico, a primeira observação importante é que temos dois nomes que não são Wide Receivers (WR), ou seja, jogadores onde sua principal função é de receber passes, Alvin Kamara e Christian McCaffrey são os dois Runnin Backs (corredores do time, conhecido como RB) que estão na lista. Esse gráfico também nos mostra que o New Orleans Saints tem uma ameaça dupla (quando o time é bom tanto no jogo corrido quanto aéreo) muito forte, pelo fato de ter dois jogadores entre os 6 com mais passes recebidos na conferência e um deles ser o RB do time. Outra conclusão importante que aparece no gráfico é que o Carolina Panthers só possui um jogador na lista, e esse jogador é um RB, o que mostra uma certa deficiência do time no corpo de WRs, ou seja, o time que conseguir anular o McCaffrey consegue anular a principal arma aérea do Panthers. Julio Jones, um dos 3 melhores recebedores da liga encabeça a lista com a maior média de jardas por recepção (16.4 jardas) e sendo o segundo com mais passes recebidos.

Agora vamos analisar o jogo terrestre da divisão, quais os principais corredores dela e seus números, que podem ser conseguidos através do código:

```
#verificando média de jardas corridas dos jogadores de cada time
nfcSouthRb<- nfcSouthData %>%
  group_by(Rusher,posteam) %>%
  filter(PlayType == "Run") %>%
  summarise(
    mediaJardas = mean(Yards.Gained),
    numCorridas = n()
  )
#Ordenando os corredores com maior número de corridas
nfcSouthRb<- arrange(nfcSouthRb,desc(numCorridas))
```

A partir destes dados podemos obter o seguinte gráfico:

Gráfico 2



Já é possível logo de cara observar que o Panthers possui dois jogadores na lista, e que o Tampa Bay Buccaneers só possui um jogador e que ele é o que possui menor média de jardas(2.99 jardas) de todos os seis. O péssimo jogo terrestre foi um fator crucial para o Buccaneers ser o único time dessa divisão de fora da pós-temporada, ou seja, é de extrema importância que a franquia invista num melhor jogo terrestre se quiser passar de fase na liga em 2018. Outra observação importante é que o corredor com maior média de jardas, Cam Newton (7.08 jardas) é na verdade o Quarterback (Lançador) da equipe, mas ele também é o com menos número de corridas tentadas, apenas 122. Esse gráfico também confirma a afirmação feita anteriormente da ameaça do Saints tanto no jogo aéreo quanto terrestre, com Mark Ingram sendo a segunda melhor média de jardas por corrida e tendo o maior número de tentativas.

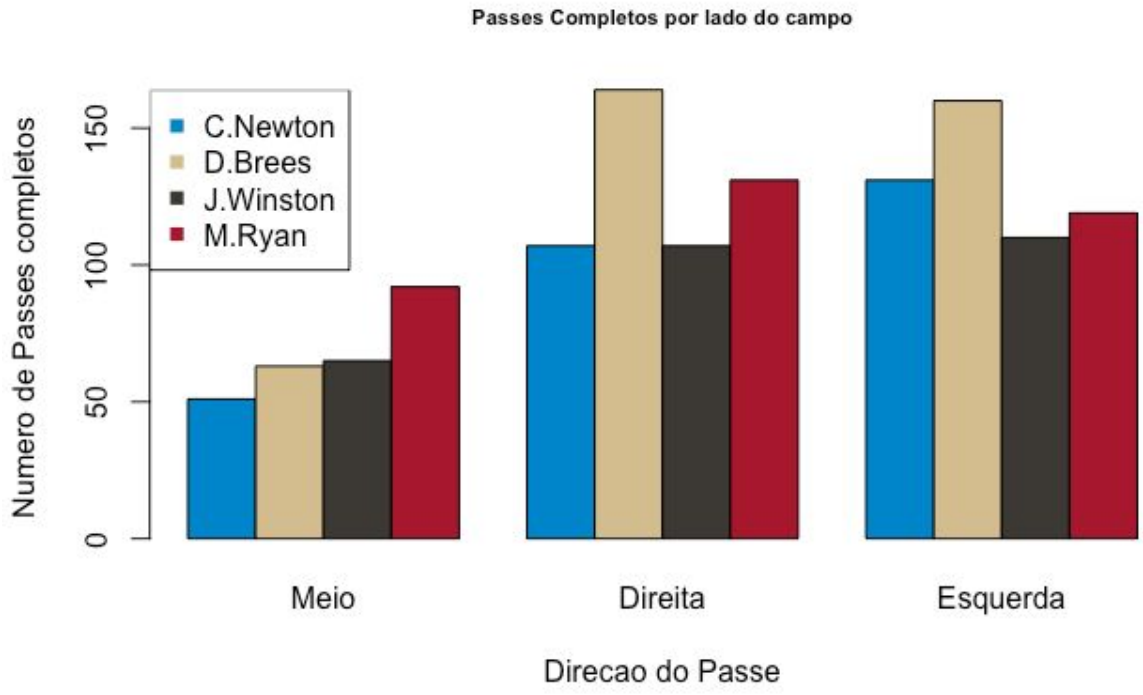
Nossa próxima análise é sobre tendências de passes que os principais Quarterbacks (QB) de cada franquia possa ter, para isso, vamos analisar a quantidade de passes para cada lado do campo. Os QBs titulares de cada time são Cam Newton no

Panthers, Drew Brees no Saints, Jameis Winston no Buccaneers e Matt Ryan no Atlanta Falcons. Primeiro é necessário separar os dados dos QBs, isso é feito a partir do código a seguir:

```
#separando para análisar tendências de passes para cada lado
#tive que usar uma matriz por conta da função para plotar os gráficos que não aceita dataframe
QBcompletePassesBySide<- as.matrix(nfcSouthData%>%
      group_by(Passer)%>%
      filter((PlayType == "Pass")&(Passer=="D.Brees"|Passer=="J.Winston"|Passer=="M.Ryan"|Passer=="C.Newton"))%>%
      summarise(
        MiddleCompleted = sum(PassOutcome=="Complete"&PassLocation=="middle",na.rm = T),
        RightCompleted = sum(PassOutcome=="Complete"&PassLocation=="right",na.rm = T),
        LeftCompleted = sum(PassOutcome=="Complete"&PassLocation=="left",na.rm = T)
      ))
#ajustando a matriz para conseguir plota-la da maneira correta no gráfico
completePassesBySide<- matrix(as.numeric(QBcompletePassesBySide[,2:4]),nrow=4,ncol=3)
```

Com isso é possível gerarmos o seguinte gráfico:

Gráfico 3



Logo de cara podemos observar a baixa tendência que os Quarterbacks possuem em lançar no meio do campo, isso muito pelo fato desta ser a área mais “congestionada”, o que faz com que os quarterbacks procurem as laterais do campo. Outro fator importante

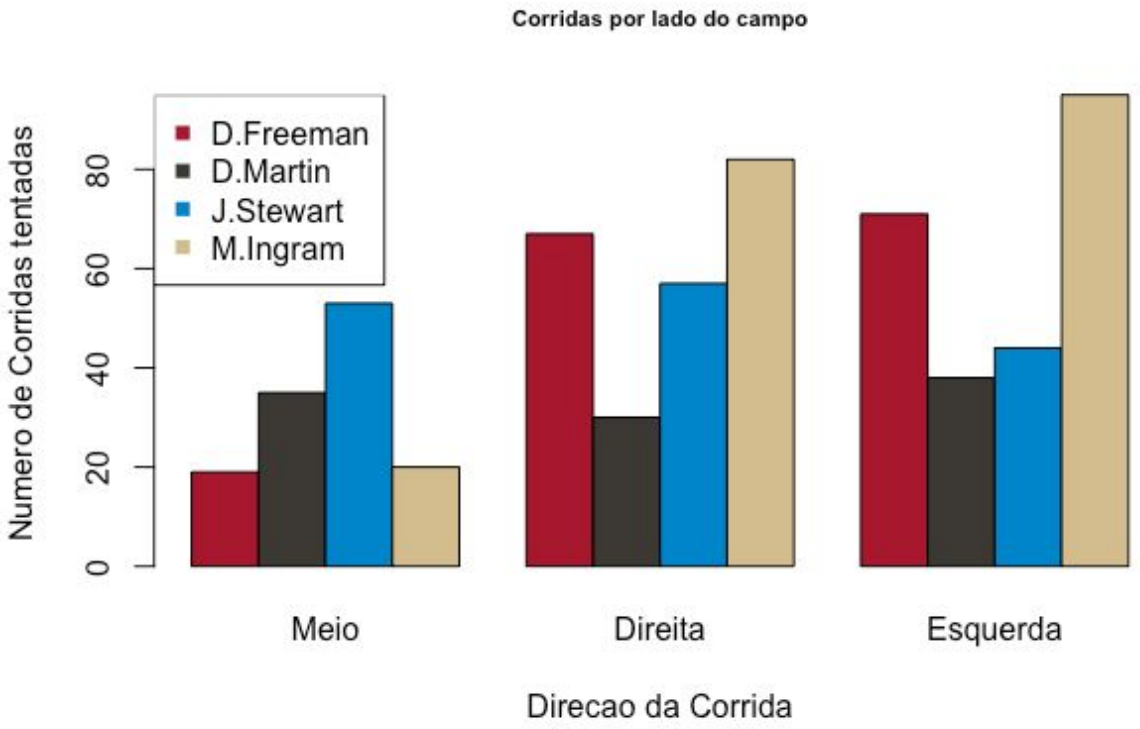
que podemos observar é a performance do Drew Brees, e a enorme diferença entre o número de passes completos para as laterais do campo em relação ao número de passes completos no meio do campo. Isso ocorre bastante pelo fato de seu RB Alvin Kamara ser um dos seus principais alvos como mostrado anteriormente, os RB geralmente fazem rota visando a lateral do campo. Outro importante a ser destacado é a clara preferência do Cam Newton para o lado esquerdo do campo, ou seja, uma defesa que consiga anular passes para o Christian McCaffrey no lado esquerdo do campo vai deixar o time do Panthers numa situação bem desconfortável.

Agora vamos verificar qual a tendência dos principais RBs desses times, ou seja, para qual lado do campo eles tendem a correr mais? Primeiro precisamos preparar os dados para criar o gráfico, isso é feito da seguinte maneira.

```
#Verificando agora o número de corridas para cada lado do campo dos principais RB de cada time
RBrunBySide<- as.matrix(nfcSouthData%>%
      group_by(Rusher)%>%
      filter((PlayType == "Run")&(Rusher=="D.Freeman"|Rusher=="D.Martin"|Rusher=="J.Stewart"|Rusher=="M.Ingram"))%>%
      summarise(
        MiddleRun = sum(RunLocation=="middle" & Yards.Gained>0,na.rm = T),
        RightRun = sum(RunLocation=="right" & Yards.Gained>0,na.rm = T),
        LeftRun = sum(RunLocation=="left" & Yards.Gained>0,na.rm = T)
      ))

#ajustando a matriz para deixa-la numérica
runsBySide<- matrix(as.numeric(RBrunBySide[,2:4]),nrow=4,ncol=3)
```

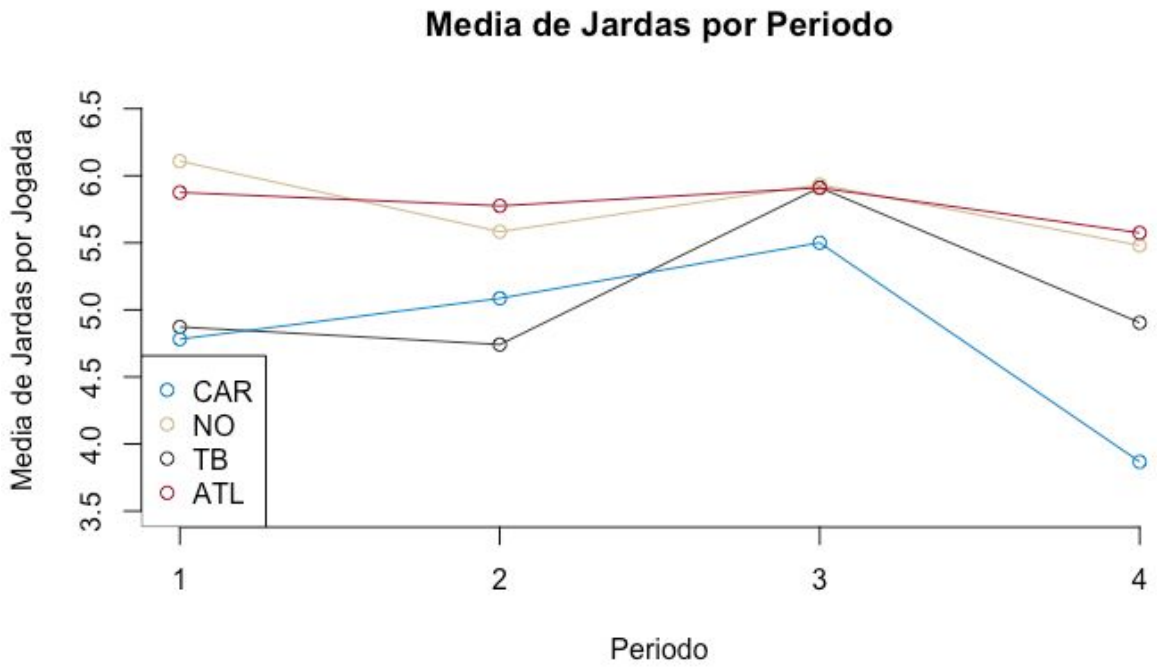
É importante perceber que só contei corridas que tiveram número de jardas positivos, dessa forma é possível gerar o seguinte gráfico



Podemos tirar algumas análises bem relevantes desse gráfico, se lembrarmos, destes quatro Running Backs, os dois com maiores médias de jardas foram Mark Ingram e Devonta Freeman, sendo Doug Martin o RB com a menor média. Quando analisamos o gráfico, é possível perceber que Mark Ingram quase não teve corridas pelo meio do campo, seus maiores ganhos foram sempre para as laterais, sendo o lado esquerdo do campo o com mais corridas positivas para o RB de New Orleans. O mesmo vale para o Freeman, onde seu foco também foi mais para as laterais do campo, tendo apenas 19 corridas no meio dele. Essa situação já não se repete com os RB com menores médias de jardas, Doug Martin e Jonathan Stewart (Buccaneers e Panthers respectivamente), o RB dos Buccaneers teve 103 corridas positivas na temporada, sendo 35 dessas para o meio do campo, ou seja, aproximadamente 34% de suas corridas. O mesmo vale para o corredor dos Panthers, que teve aproximadamente 34,4% das suas corridas para o meio do campo, diferentemente de Ingram e Freeman que tiveram respectivamente apenas 10,5% e 12,1% de suas corridas positivas em direção ao meio do campo. Como falado anteriormente, o meio do campo é a parte que possui uma maior concentração de jogadores, logo é natural que se tenha um menor ganho de jardas quando as corridas vão por essa direção, podemos concluir a partir disso que tentar mais corridas pelo meio do campo impacta diretamente e negativamente em uma menor média de jardas para o atleta.

Vamos agora analisar a quantidade média de jardas por jogada para cada período do jogo de cada time, a partir disso podemos gerar o gráfico

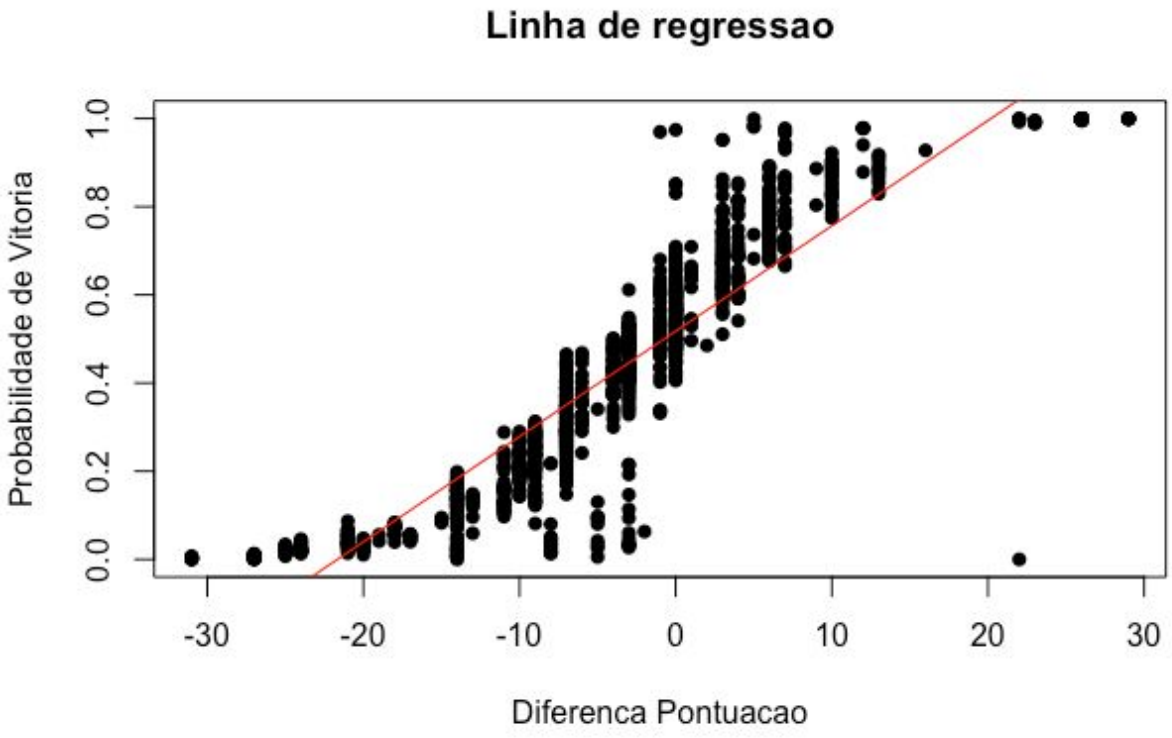
Gráfico 5



É possível observar a consistência dos dois melhores times da divisão (NO e ATL) durante todos os períodos do jogo, e a forma como a performance principalmente do Buccaneers varia, tendo seu pico apenas no terceiro período de jogo. Já o Carolina tem uma média horrível durante o último período, tendo em média apenas 3.87 jardas por jogada, o que mostra a dependência do time com o seu time de defesa. Esse gráfico nos mostra também que para ser ter um time competitivo dentro da liga, é necessário consistência durante todos os períodos do jogo.

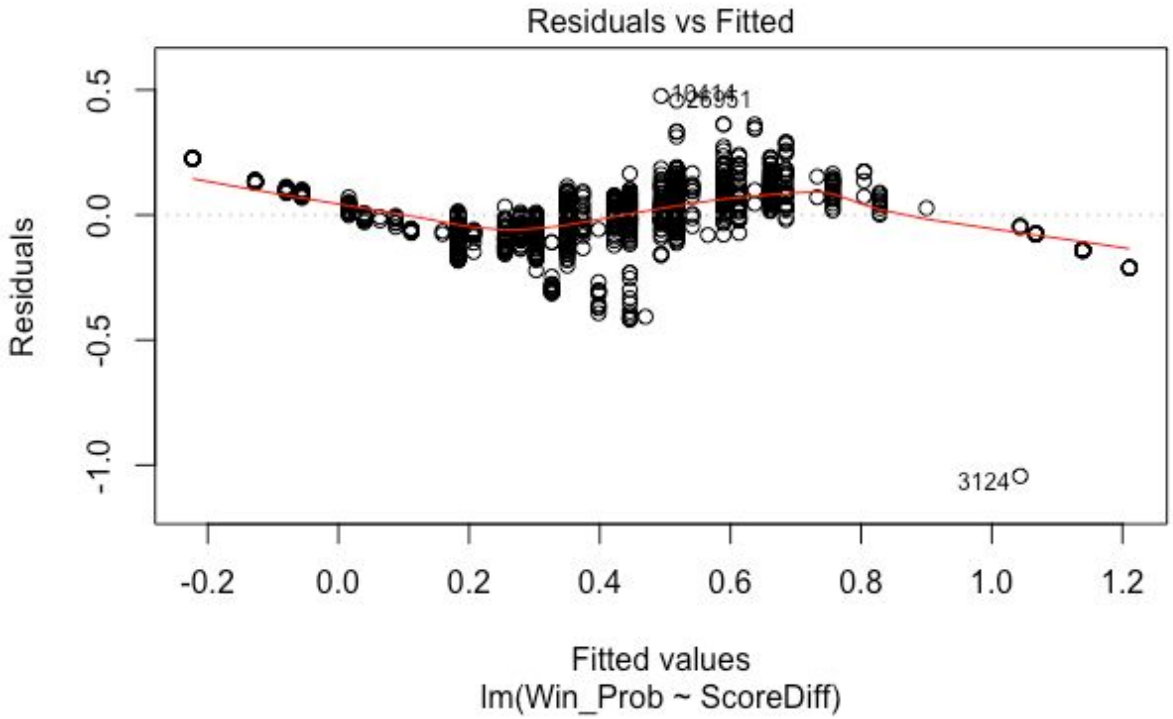
Agora vamos tentar fazer previsões a partir dos dados disponíveis para nós, primeiro vamos verificar a correlação entre duas variáveis, Win_Prob e ScoreDiff. Quando fazemos o teste de correlação linear entre as duas variáveis, obtemos o resultado de 0.9110598, ou seja, isso mostra uma correlação positiva muito forte entre essas duas variáveis. Isso faz muito sentido, afinal, quanto maior a diferença positiva no placar, maior a probabilidade do time vencer, isso fica bem visível quando observamos o gráfico da linha de regressão:

Gráfico 6



Mas será que dada uma diferença do placar é possível prever a probabilidade de vitória de um time? Para isso é necessário fazermos mais alguns testes, o primeiro é verificar se existe alguma relação estatística entre as duas variáveis, para isso o valor de Pr(>|t|) precisa ser menor do que 0.05, e no caso dessas duas variáveis, isso foi verdadeiro, obtivemos o resultado de 2e-16 para elas. O próximo teste é verificar se elas são provenientes de uma distribuição normal, isso é possível fazer através do teste de Shapiro, caso o valor de *p* seja menor que 0.05, então os valores não são provenientes de uma distribuição normal. No teste realizado obtivemos o valor de 2.2e-16, ou seja, eles não são provenientes de uma distribuição normal, logo não é possível fazer nenhuma previsão em cima dessas variáveis. O último teste seria verificar a variância homogênea entre as duas variáveis, e isso é facilmente observado através do seguinte gráfico:

Gráfico 7:



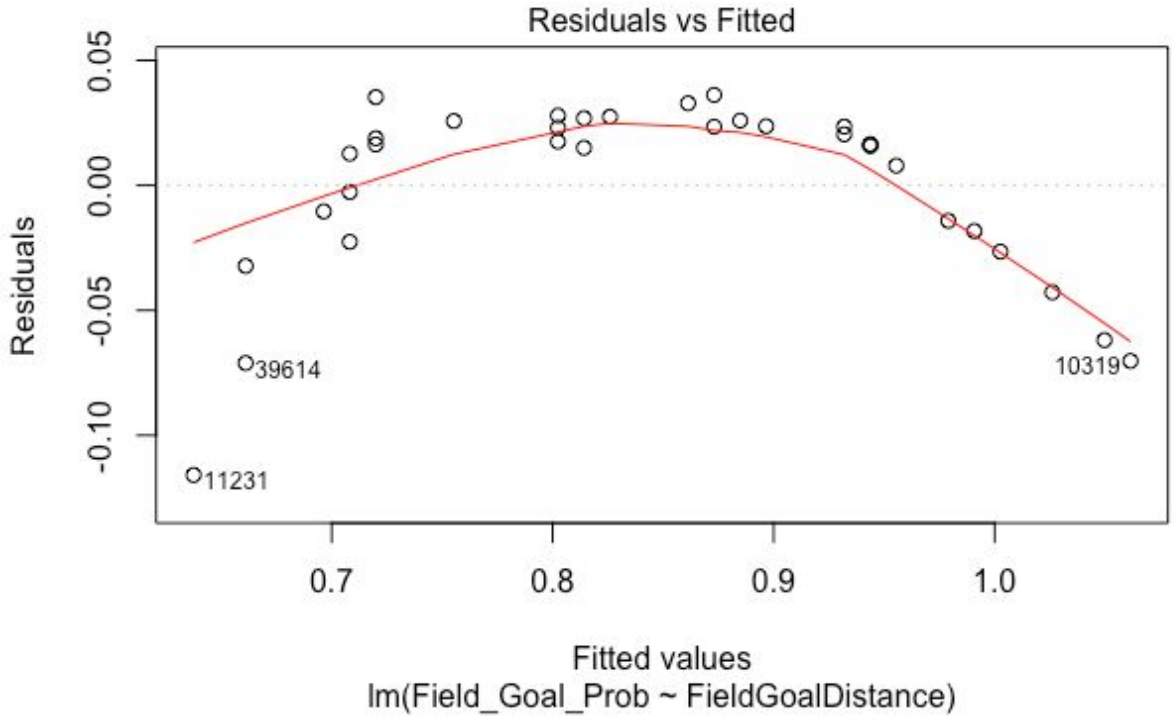
Nele podemos observar uma distribuição não tanto homogênea, o que também não tornaria possível fazer a previsão entre as duas variáveis.

A próxima análise é entre as variáveis yrdline100 e Opp_Touchdown_Prob, ou seja, a probabilidade do adversário marcar um Touchdown a medida que o time no ataque vai avançando em campo. Mais uma vez, primeiramente é necessário observar a correlação entre elas, obtemos 0.7043097 que é considerado um bom grau de associação, mas será que é possível fazer alguma previsão? Mais uma vez, através do modelo linear, obtemos que as duas variáveis possuem uma relação estatística entre si, pois o valor de Pr(>|t|) foi 2e-16. O próximo passo é verificar se essas observações são provenientes de uma distribuição normal, mais uma vez aplicamos o teste de Shapiro e obtivemos o valor de *p* 2.572e-15, ou seja, menor que 0.05, logo não são distribuidos normalmente, o que mostra que não é possível fazer previsões em cima dessas duas variáveis.

Por fim, vamos tentar analisar mais uma dupla de variáveis, Field_Goal_Prob e FieldGoalDistance, ou seja, será que a partir de uma determinada distância do Field Goal é possível prever qual a probabilidade de acerto? O teste de correlação nos traz um

resultado importante, -0.9587256, o que mostra uma forte correlação negativa, ou seja, a medida que uma variável cresce a outra diminui. Essa afirmação faz total sentido, pois quanto maior a distância do Field Goal, mais difícil é de acertar. É também possível verificar a relação estatística entre as duas variáveis por conta do valor de Pr(>|t|) = 2e-16. O próximo passo é verificar se são provenientes de uma distribuição normal, através do teste de Shapiro, e mais uma vez obtemos um valor menor que 0.05, nesse caso 4.074e-05. A última etapa é analisar se há uma variância homogênea, e isso mais uma vez é observado através do gráfico:

Gráfico 8:



É possível perceber que não há uma variância homogênea, logo, tanto por conta da distribuição normal quanto por conta da variância homogênea, não é possível fazer uma previsão em cima dessas duas variáveis.

5. Conclusão

Foi possível observar que a análise de dados pode se tornar bastante importante para verificar padrões e tendências de times e jogadores da NFL, numa escala profissional isso pode trazer vários benefícios. Entender como um adversário joga e quais seus principais jogadores, pode aumentar bastante as chances de vitória.

Outro ponto interessante é que foi possível observar a dificuldade de tentar prever algo no jogo, e isso que torna o esporte um entretenimento tão grande, pois no final, a incerteza do futebol americano prevalece.

Como trabalhos futuros, deseja-se a análise também de padrões dos times de defesa e até mesmo dos times de especialistas de outras divisões e conferências da NFL, assim como analisar os números de outras temporadas.