Trabalho da A2

Gustavo Tironi e Luis Felipe Marciano

2023-06-03

Formulação e documentação da ideia

A escolha da base se deu a partir de uma exploração no **Kaggle**, na qual nos deparamos com a base de dados sobre os jogadores draftado na NBA. Definimos então, que essa seria a base utilizada. Analisando a base, vimos que, além de apresentar os jogadores, a posição no qual foram draftados e os times, desde 1989, também haviam as estatísticas de cada jogador ao longo de sua carreira. Conforme pode ser visualizado abaixo.

```
base_de_dados <- read.csv("nbaplayersdraft.csv", sep = ",")

kable(head(base_de_dados))  %>%
kable_styling(latex_options = "striped", stripe_index = c(1,2, 5:6))
```

Antes de partimos para a hipótese, cabe uma explicação sobre o processo de Draft da NBA.

O que é o Draft da NBA?

O Draft da NBA é o processo em que jogadores amadores são escolhidos por franquias e entram oficialmente na maior liga de basquetebol profissinal do mundo, a NBA. O evento é composto por duas rodadas de 30 escolhas cada, com uma ordem pré-definida. Os times se revezam e cada um seleciona um jogador por rodada. Ou seja, 60 atletas são recrutados. Portanto, há de se imaginar que os primeiros escolhidos do Draft são melhores jogadores e farão melhores performances em sua carreira na liga.

Com isso, surgiu o seguinte questionamento, que veio a se tornar a hipótese a ser respondida com esse trabalho.

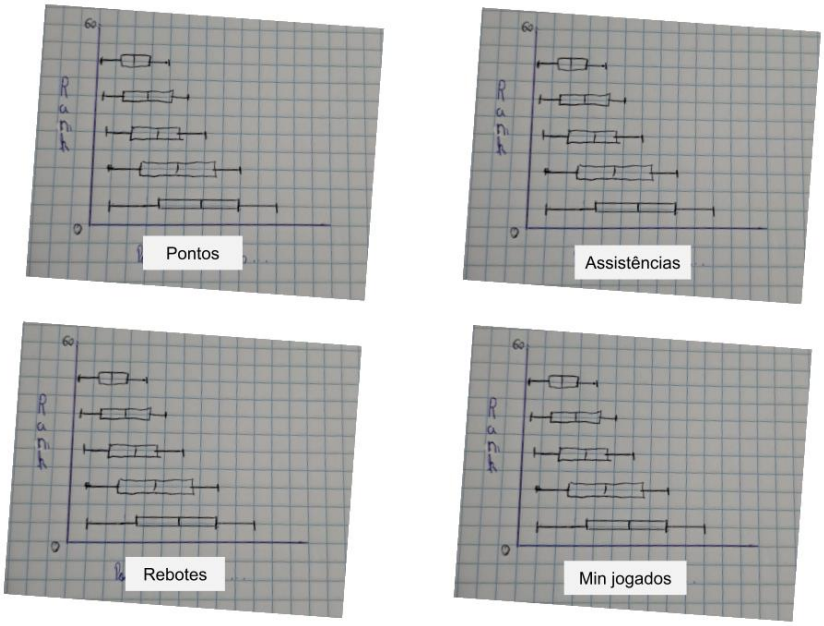
Hipótese

A posição em que o jogador foi draftado realmente tem correlação com o seu desempenho durante a carreira?

id	year	rank	overall_pick	team	player	college	years_active	games	minutes_played	points
1	1989	1	1	SAC	Pervis Ellison	Louisville	11	474	11593	4494
2	1989	2	2	LAC	Danny Ferry	Duke	13	917	18133	6439
3	1989	3	3	SAS	Sean Elliott	Arizona	12	742	24502	10544
4	1989	4	4	MIA	Glen Rice	Michigan	15	1000	34985	18336
5	1989	5	5	CHH	J.R. Reid	UNC	11	672	15370	5680
6	1989	6	6	CHI	Stacey King	Oklahoma	8	438	7406	2819

Dessa forma, pretendemos validar ou invalidar a hipótese acima, a partir da visualização dos dados da base escolhida.

Para tal, inicialmente, pensamos em criar quatro gráficos categórico vs quantitativo, onde caso exista alguma correlação os jogadores escolhidos nas primeiras posições, devem ter estatísticas melhores do que aqueles escolhidos nas últimas. Sendo assim, para validar a hipótese, usando como métrica as estatísticas **pontos**, **rebotes**, **assistências** e **minutos jogados**, esperamos que os gráficos se apresentem conforme o rascunho abaixo.



Análise Exploratória da base de dados

Como temos a hipótese a ser respondida em mente, trataremos a princípio apenas as variáveis pertinentes, sendo que, caso venha a se tornar necessário a inclusão de outras variáveis, iremos analisá-las posteriormente.

Antes de qualquer análise mais aprofundada, devemos entender nossa base de dados. Batendo o olho no arquivo **.cvs**, podemos ver que cada linha representa um jogador e cada coluna uma variável a respeito dele. Assim, podemos começar a tratar as variáveis. Para responder à hipótese, precisamos, acima de tudo, conseguir identificar a posição em que o jogador foi escolhido no draft. Para isso, temos as variáveis **rank** e **overall_pick**, que são o mesmo, e tratam exatamente da posição em que o jogador foi draftado. Essas, são variáveis qualitativas ordinais que vão de 1 a 60.

Tendo a posição definida, precisamos olhar agora para as variáveis que nos ajudarão a definir o desempenho do jogador. Previamente, definimos que as estatísticas **pontos**, **rebotes**, **assistências** e **minutos jogados** seriam as responsáveis por determinar o desempenho. Essa escolha vai conforme a comunidade de basquete, que frequentemente usa essas estatísticas para definir a grandeza de um jogador, principalmente os **pontos**. Contudo, entendendo a diversidade de posições no basquete, não é justo analisar apenas pelos pontos, e sim pelo conjunto da obra, ou seja, **pontos**, **rebotes** e **assistências**. Junto a isso, os **minutos jogados** são de extrema importância, já que um jogador importante para o time, jogará mais minutos em cada partida.

Anaálise Unidimensional

Analisando o banco de dados, vemos que a coluna **minutes_played** traz os minutos jogados, a coluna **points** traz os pontos totais, a coluna **total_rebounds** traz os rebotes e a coluna **assists** traz as assistências. Todas essas, são variáveis quantitativas discretas.

Com as variáveis devidamente, definidas, podemos começar a análise exploratória. Inicialmente, iremos aplicar a função **summary** em todas as variáveis, para observar seu comportamento.

```
base_de_dados_resumida <- select(base_de_dados, points, minutes_played,total_rebounds, assists)

summary(base_de_dados_resumida)
```

```
##      points      minutes_played  total_rebounds     assists
##  Min.   :    0   Min.   :    0   Min.   :    0   Min.   :    0.0
##  1st Qu.:  265   1st Qu.:  838   1st Qu.:  128   1st Qu.:   46.0
##  Median : 1552   Median : 4204   Median :  656   Median :  257.0
##  Mean   : 3580   Mean   : 8399   Mean   : 1497   Mean   :  774.3
##  3rd Qu.: 5150   3rd Qu.:13246   3rd Qu.: 2139   3rd Qu.:  910.0
##  Max.   :37062   Max.   :52139   Max.   :15091   Max.   :12091.0
##  NA's   :253     NA's   :253     NA's   :253     NA's   :253
```

Já pudemos identificar a ocorrência de valores nulos na base. Analisando-a mais profundamente, podemos determinar que os valores nulos correspondem a jogadores que foram draftados, mas que nunca jogaram na NBA. Por isso, resolvemos desconsiderar esses valores, sem perda de dados significantes.

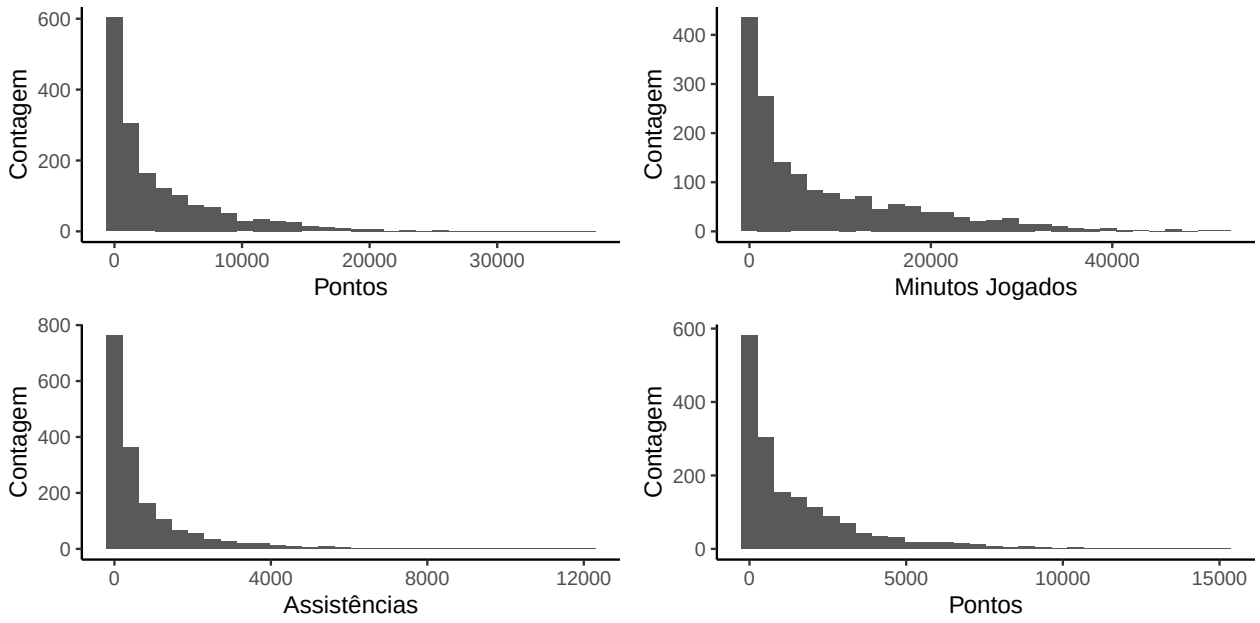
```
base_de_dados_resumida <- na.omit(select(base_de_dados, points, minutes_played, total_rebounds, assists)

summary(base_de_dados_resumida)
```

```
##      points      minutes_played  total_rebounds     assists
##  Min.   :    0   Min.   :    0   Min.   :    0   Min.   :    0.0
##  1st Qu.:  265   1st Qu.:  838   1st Qu.:  128   1st Qu.:   46.0
##  Median : 1552   Median : 4204   Median :  656   Median :  257.0
##  Mean   : 3580   Mean   : 8399   Mean   : 1497   Mean   :  774.3
##  3rd Qu.: 5150   3rd Qu.:13246   3rd Qu.: 2139   3rd Qu.:  910.0
##  Max.   :37062   Max.   :52139   Max.   :15091   Max.   :12091.0
```

Agora, podemos a analisar unidimensional das variáveis que serão utilizadas.

```
p_points <- ggplot(data = base_de_dados_resumida, mapping = aes(x = points)) +
  geom_histogram() +
  labs(x = "Pontos", y = "Contagem")

p_minutes <- ggplot(data = base_de_dados_resumida, mapping = aes(x = minutes_played)) +
  geom_histogram() +
  labs(x = "Minutos Jogados", y = "Contagem")

p_assists <- ggplot(data = base_de_dados_resumida, mapping = aes(x = assists)) +
  geom_histogram() +
  labs(x = "Assistências", y = "Contagem")
```

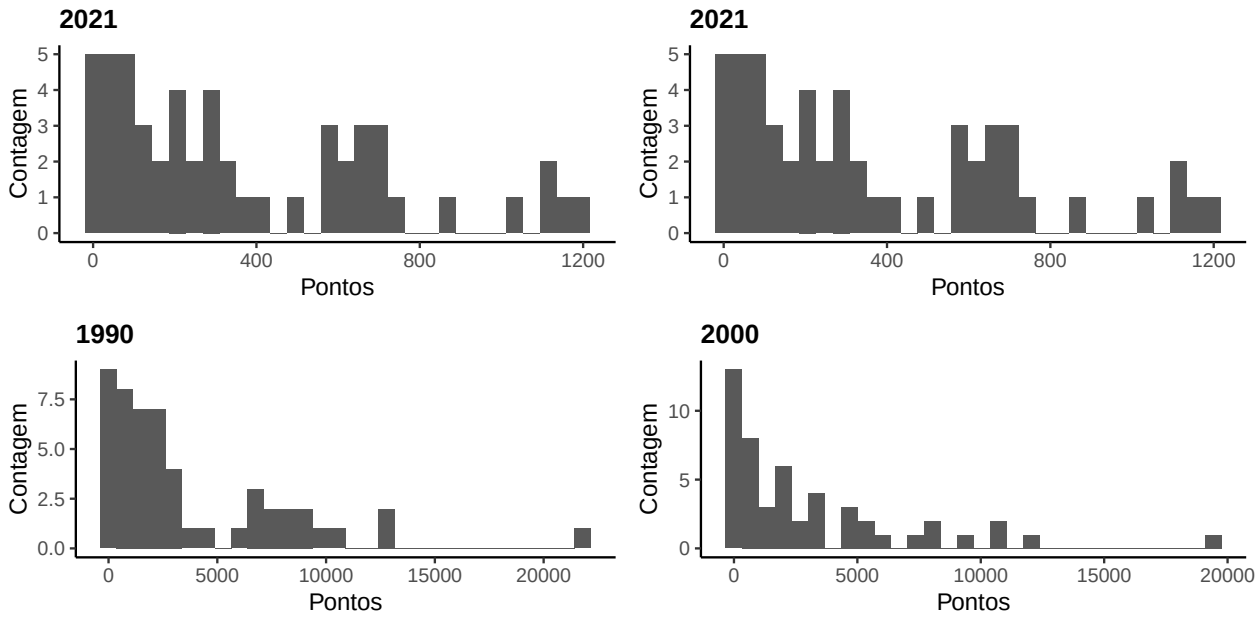
```
p_rebounds <- ggplot(data = base_de_dados_resumida, mapping = aes(x = total_rebounds)) +
  geom_histogram() +
  labs(x = "Pontos", y = "Contagem")

grid.arrange(p_points, p_minutes,
             p_assists, p_rebounds,
             ncol=2, nrow=2)
```



Com essa análise, já podemos observar algo interessante. Podemos ver que há uma maior ocorrência de todas as variáveis nos valores mais baixos. Analisando a base de dados de forma visual, identificamos como uma possível causa disso, que jogadores que foram draftados em anos mais recentes, como 2020, têm pouco tempo de carreira e, consequentemente, têm menores estatísticas. Para confirmar isso, resolvemos replicar a analise, para a variavel pontos, separando por pelo ano em que o jogador foi draftado.

```
dados_year_points <- na.omit(select(base_de_dados, points, year))

dados_2021 <- dados_year_points %>% filter(year == 2021)
dados_2020 <- dados_year_points %>% filter(year == 2020)
dados_1990 <- dados_year_points %>% filter(year == 1990)
dados_2000 <- dados_year_points %>% filter(year == 2000)

p_2021 <- ggplot(data = dados_2021, mapping = aes(x = points)) +
  geom_histogram() +
  labs(x = "Pontos", y = "Contagem", title = "2021")

p_2020 <- ggplot(data = dados_2020, mapping = aes(x = points)) +
  geom_histogram() +
  labs(x = "Pontos", y = "Contagem", title = "2020")

p_1990 <- ggplot(data = dados_1990, mapping = aes(x = points)) +
  geom_histogram() +
  labs(x = "Pontos", y = "Contagem", title = "1990")
```

```
p_2000 <- ggplot(data = dados_2000, mapping = aes(x = points)) +
  geom_histogram() +
  labs(x = "Pontos", y = "Contagem", title = "2000")

grid.arrange(p_2021, p_2021,
             p_1990, p_2000,
             ncol=2, nrow=2)
```



Aqui, vale destacar que a escala dos eixos estão extremamente diferentes, para melhor visualização e pela dificuldade em programar outra maneira onde as escalas estejam iguais e seja visível. Contudo, para o nosso propósito, esses gráficos irão servir. Como podemos observar, nossa suspeita inicial se confirmou, pois nos anos mais recentes (2021 e 2020), a maioria dos dados se concentram abaixo dos **1000** pontos, enquanto os dados de mais antigos (1990 e 2020) se concentram até a região dos **5000** pontos. Isso já é um empecilho para o uso desses dados, pois dessa forma, não poderemos analisar bem a performance dos jogadores, já que jogadores mais antigos serão favorecidos. Para contornar esse problema, precisamos ponderar essas estatísticas pelo tempo jogado. Por sorte, temos na base de dados, todas estatísticas ponderadas por partidas. Então mudaremos o foco, e começaremos a analisar essas estatísticas, recomeçando a análise. Para tal, serão usadas as variáveis ponderadas **points_per_game**, **average_assists**, **average_total_rebounds** e **average_minutes_played**.

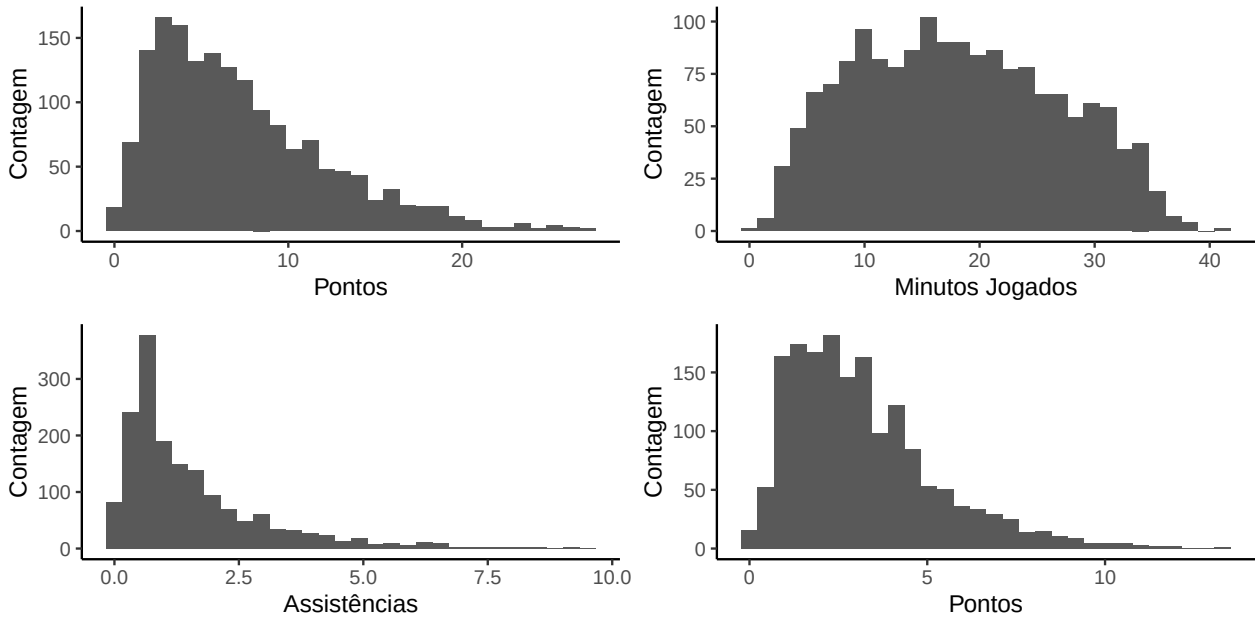
```
base_de_dados_resumida <- na.omit(select(base_de_dados, points_per_game, average_assists, average_total_

summary(base_de_dados_resumida)
```

```
##  points_per_game  average_assists average_total_rebounds average_minutes_played
##  Min.   : 0.000   Min.   :0.000   Min.   : 0.000         Min.   : 0.00
##  1st Qu.: 3.400   1st Qu.:0.500   1st Qu.: 1.700         1st Qu.:11.00
##  Median : 6.200   Median :1.100   Median : 2.800         Median :17.70
##  Mean   : 7.276   Mean   :1.551   Mean   : 3.194         Mean   :18.13
##  3rd Qu.:10.000   3rd Qu.:2.100   3rd Qu.: 4.200         3rd Qu.:24.80
##  Max.   :27.200   Max.   :9.500   Max.   :13.300         Max.   :41.10
```

```
##       rank
##  Min.   : 1.0
##  1st Qu.:13.0
##  Median :26.0
##  Mean   :26.9
##  3rd Qu.:40.0
##  Max.   :60.0
```

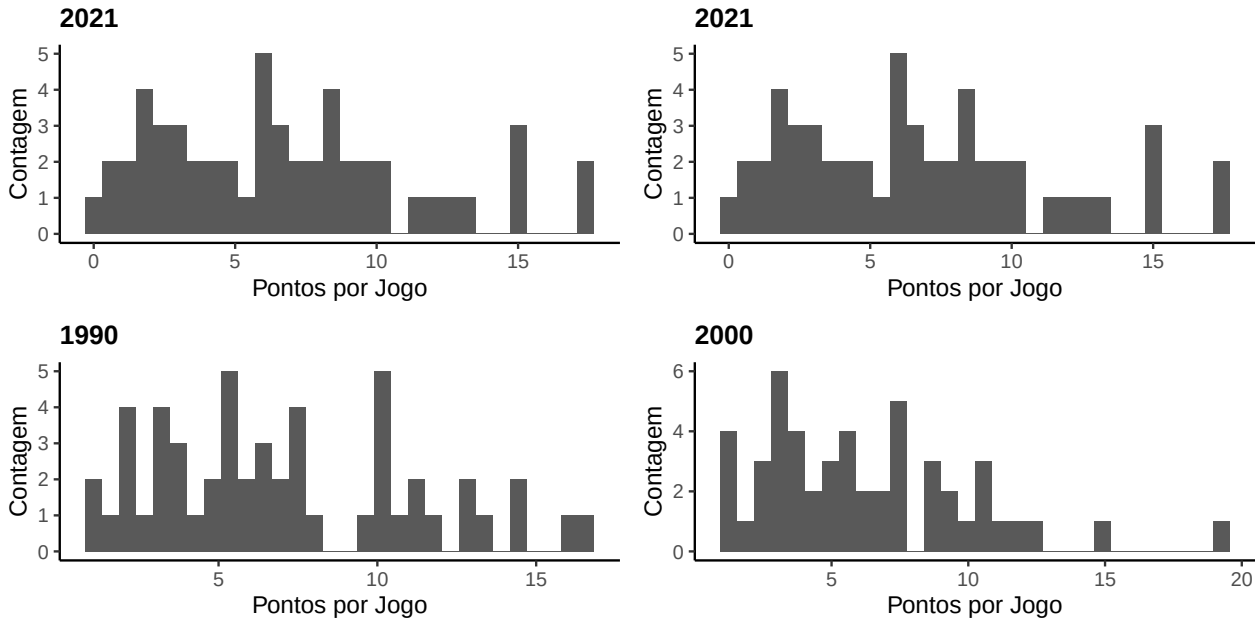
Então são plotados os gráficos.

```
p_points <- ggplot(data = base_de_dados_resumida, mapping = aes(x = points_per_game)) +
  geom_histogram() +
  labs(x = "Pontos", y = "Contagem")

p_minutes <- ggplot(data = base_de_dados_resumida, mapping = aes(x = average_minutes_played)) +
  geom_histogram() +
  labs(x = "Minutos Jogados", y = "Contagem")

p_assists <- ggplot(data = base_de_dados_resumida, mapping = aes(x = average_assists)) +
  geom_histogram() +
  labs(x = "Assistências", y = "Contagem")

p_rebounds <- ggplot(data = base_de_dados_resumida, mapping = aes(x = average_total_rebounds)) +
  geom_histogram() +
  labs(x = "Pontos", y = "Contagem")

grid.arrange(p_points, p_minutes,
             p_assists, p_rebounds,
             ncol=2, nrow=2)
```



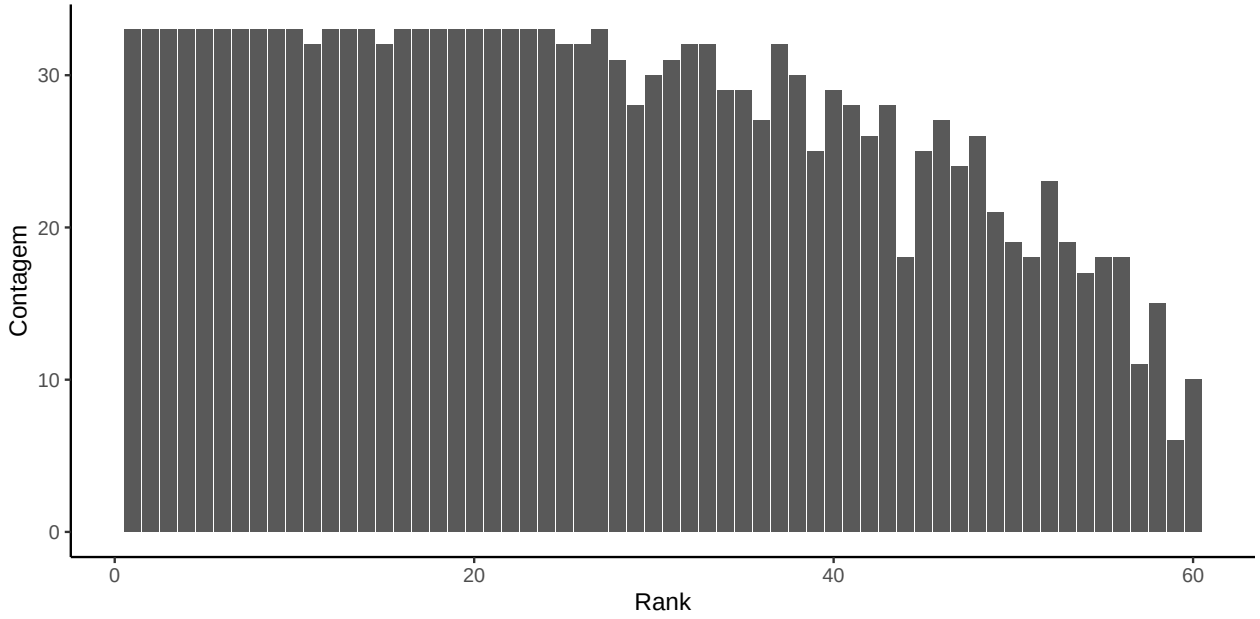
Agora, podemos notar uma melhor distribuição dos dados. Contudo, para confirmar isso, repetiremos a analise dos anos mais recentes e de anos mais antigos, para comparação e confirmação.

```r
dados_year_points <- na.omit(select(base_de_dados, points_per_game, year))

dados_2021 <- dados_year_points %>% filter(year == 2021)
dados_2020 <- dados_year_points %>% filter(year == 2020)
dados_1990 <- dados_year_points %>% filter(year == 1990)
dados_2000 <- dados_year_points %>% filter(year == 2000)

p_2021 <- ggplot(data = dados_2021, mapping = aes(x = points_per_game)) +
  geom_histogram() +
  labs(x = "Pontos por Jogo", y = "Contagem", title = "2021")

p_2020 <- ggplot(data = dados_2020, mapping = aes(x = points_per_game)) +
  geom_histogram() +
  labs(x = "Pontos por Jogo", y = "Contagem", title = "2020")

p_1990 <- ggplot(data = dados_1990, mapping = aes(x = points_per_game)) +
  geom_histogram() +
  labs(x = "Pontos por Jogo", y = "Contagem", title = "1990")

p_2000 <- ggplot(data = dados_2000, mapping = aes(x = points_per_game)) +
  geom_histogram() +
  labs(x = "Pontos por Jogo", y = "Contagem", title = "2000")

grid.arrange(p_2021, p_2021,
             p_1990, p_2000,
             ncol=2, nrow=2)
```



Novamente se confirma a distribuição mais igualitária dos dados, indicando que essas variáveis são mais eficazes para determinação do desempenho dos jogadores. Por fim, resolvemos analisar a variável **Rank**, que também será utilizada.

```
ggplot(data = base_de_dados_resumida, mapping = aes(x = rank)) +
  geom_histogram(stat = "count") +
  labs(x = "Rank", y = "Contagem")
```



Com o gráfico, fica evidente que há menos dados de jogadores nas últimas posições do rank, contudo, isso era esperado, já que algumas posições do draft só começaram em anos mais recentes. Além disso, muito dos NA's retirados estavam nessas posições.

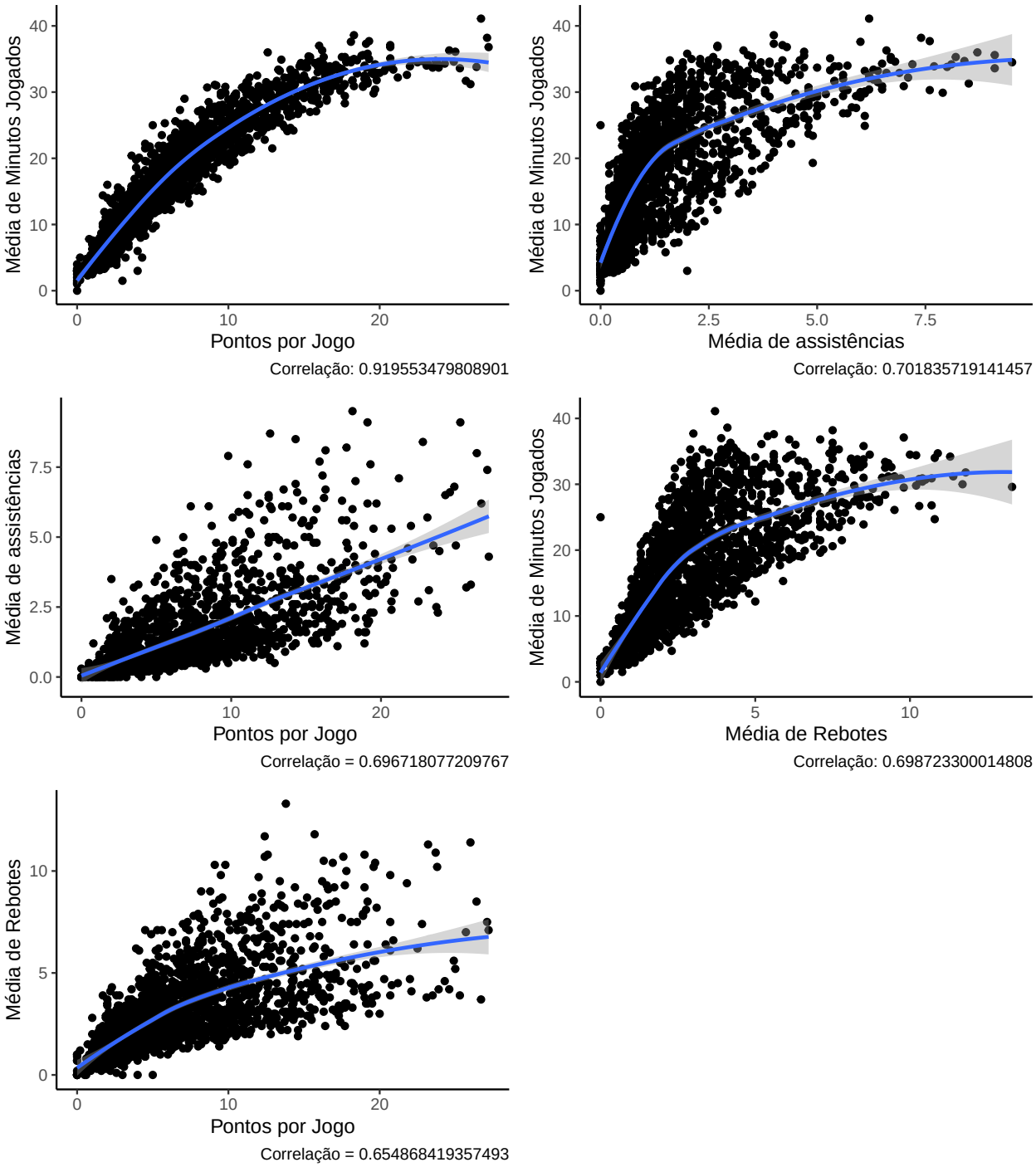
Com isso, finalizamos nossa análise unidimensional, e podemos partir para a análise bidimensional.

Análises Bidimensionais

Para a análise bidimensional, analisaremos a correlação entre as variáveis escolhidas e plotaremos os gráficos bidimensionais. Por se tratarem de variáveis quantitativas, o gráfico utilizado será o de dispersão.

```
g1 <- ggplot(base_de_dados_resumida, mapping = aes(x = points_per_game, y = average_minutes_played)) +
  geom_point() + geom_smooth(method = "loess") +
  labs(x = "Pontos por Jogo",
       y = "Média de Minutos Jogados",
       caption =sprintf("Correlação: %s", cor(base_de_dados_resumida$points_per_game, base_de_dados_resu

g2 <- ggplot(base_de_dados_resumida, mapping = aes(x = average_assists, y = average_minutes_played)) +
  geom_point() + geom_smooth(method = "loess") +
  labs(x = "Média de assistências",
       y = "Média de Minutos Jogados",
       caption =sprintf("Correlação: %s", cor(base_de_dados_resumida$average_assists, base_de_dados_resu

g3 <- ggplot(base_de_dados_resumida, mapping = aes(x = average_total_rebounds, y = average_minutes_playe
  geom_point() + geom_smooth(method = "loess") +
  labs(x = "Média de Rebotes",
       y = "Média de Minutos Jogados",
```

```
       caption =sprintf("Correlação: %s", cor(base_de_dados_resumida$average_total_rebounds, base_de_da

g4 <- ggplot(base_de_dados_resumida, mapping = aes(x = points_per_game, y = average_assists)) +
  geom_point() + geom_smooth(method = "loess") +
  labs(x = "Pontos por Jogo",
       y = "Média de assistências",
       caption = sprintf("Correlação = %s", cor(base_de_dados_resumida$points_per_game, base_de_dados_r


g5 <- ggplot(base_de_dados_resumida, mapping = aes(x = points_per_game, y = average_total_rebounds)) +
  geom_point() + geom_smooth(method = "loess") +
  labs(x = "Pontos por Jogo",
       y = "Média de Rebotes",
       caption =sprintf("Correlação = %s", cor(base_de_dados_resumida$points_per_game, base_de_dados_re

grid.arrange(g1, g2, g4, g3, g5, ncol = 2, nrow = 3)
```



Com isso, podemos ver que há uma grande correlação entre **média de pontos** e **minutos jogados**. As outras variáveis também mostraram certo grau de correlação, mas não como as já citadas. Com essa análise, pode-se identificar que há uma relação entre a **média de pontos** com as outras variáveis e entre os **minutos jogados** e as outras variáveis.

Pensamento Editorial

Antes de partirmos para a produção peça gráfica e das visualizações, devemos planejar alguns tópicos do pensamento editorial a respeito da peça gráfica, tais como: a definição de um público alvo, definição da qualidade expressiva da peça e da solução de representação do dado.

A respeito do público alvo, a peça não é direcionada para o público geral, pois não há um compromisso em fazer gráficos de amplo conhecimento pela maioria das pessoas, mas sim fazer gráficos que melhor representem os dados. Porém, não é somente esse fator que limita o público, mas também o interesse pelo esporte. Portanto, o público alvo pode ser definido como apreciadores de basquete com algum conhecimento de estatística.

A qualidade expressiva do gráfico é neutra, pois não é nosso propósito impactar o leitor com uma representação dramática.

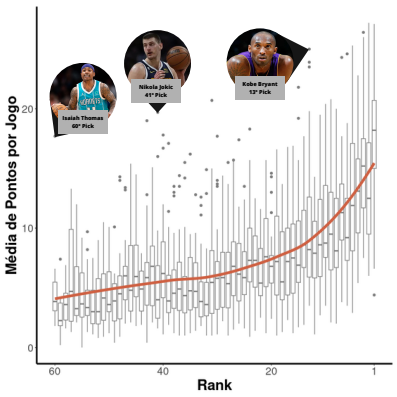
A escolha de representação do gráfico é a que é melhor indicada para gráficos categóricos X quantitativo, o boxplot. Porém, como o objetivo do gráfico é mostrar um comportamento, a informação mais importante não são os boxplots, mas sim a linha de tendência que passa por eles. Portanto, deve-se colorir com cores temática de basquete o que é mais importante, como a linha de tendência, e deixar em uma cor mais opaca outras informações que servem como contexto, a fim de não confundir a compreensão do leitor.

Finalmente, as ferramentas escolhidas para a realização da peça gráfica foram: ggplot, para execução dos gráficos; dplyr, para manipulação de dados; canvas, para design e algumas outras bibliotecas do R para execução do markdown e auxílio na criação dos gráficos.

Produção da peça gráfica

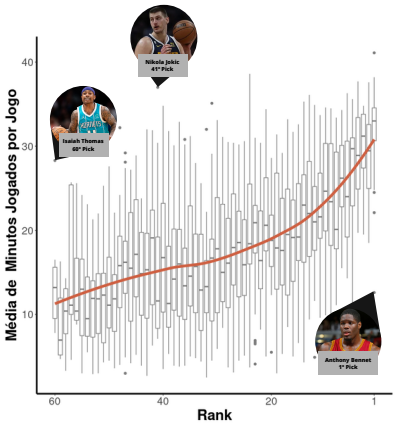
Com tudo apresentado até agora, foi possível, então, executar a peça gráfica desejada. O resultado, pode ser visualizado abaixo.

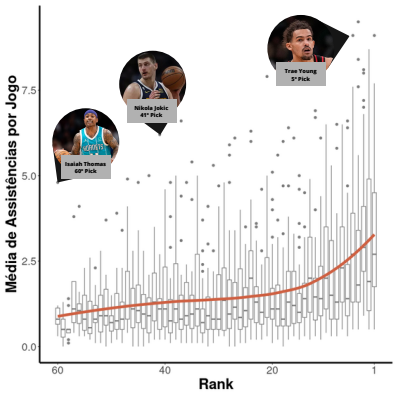
```
knitr::include_graphics("Posição no Draft e desempenho.jpg")
```

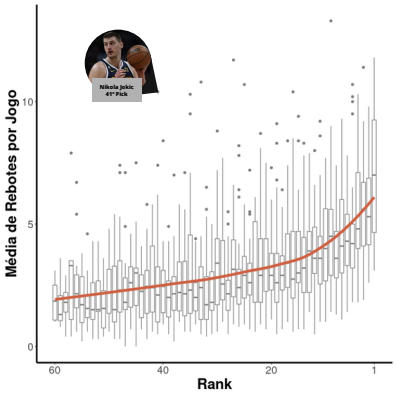


Posição no Draft e Desempenho

Gustavo Tironi e Luís Felipe







Com o trabalho finalizado, gostaríamos de acrescentar que a realização dessa atividade foi de extrema importância para aprimorar nossas habilidades com ggplot e R. Muitas funcionalidades usadas tiveram que ser aprendidas para que o resultado fosse como desejado. Com isso, pudemos aprender mais a pesquisar diretamente na documentação, além de ser engenhoso em alguns momentos para fazer dar certo rs.