Trabalho de Conclusão de Curso

PÓS-GRADUAÇÃO EM CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL

Análise da relação entre tempo de treinamento e performance para jogadores profissionais de hockey

ALUNO: Paola Katherine Cardoso Pacheco
ORIENTADOR: Marta Ribeiro Hentschke

PUCRS online uol edtech

Sumário

1. RESUMO

A mineração de dados tem extrema importância quando tratamos de estratégia, uma vez que é possível detectar tendências, prever alguns resultados, modelar e coletar informações e reconhecer padrões. [1]

Nesse trabalho exploraremos dados da *National Hockey League (NHL)*, a qual disponibiliza dados de anos anteriores relacionados a métricas de hockey. Toda a construção do estude se baseia em manipular dados no Google Colaboratory.

Essa pesquisa apresenta o estudo de métricas relacionadas a *gaps* de diferença de tempo desde a data de contratação de um jogador até seu início de carreira jogando em uma temporada.

Após a limpeza dos dados, foi construído por linhas de regressão as médias de performance baseados nos *gaps* de tempo que grupos de jogadores com o mesmo *gap* tem entre *draft*, sua contratação e sua primeira temporada.

Além da limpeza de dados, foi usada visualização de dados para que ficasse mais claro e facilitasse a interpretação gráfica entre a diferença de performance.

2. INTRODUÇÃO

O pré-processamento de dados é um pilar fundamental do que hoje conhecemos como *Big Data*. Para análise e visualização de dados, é imprescindível a etapa de pré-processamento.

A estudo consiste em etapas desde extração de dados, limpeza, padronização e reorganização dos dados de forma que seja possível realizar análises após terem dados normalizados. [2]

Pré-processar consiste em aplicar técnicas de limpeza, normalização, organização dos dados para que possa ser possível realizar análises e visualizações.

Após realizada a limpeza dos dados, é necessário minerar, aplicar métodos para buscar padrões nos dados e, com isso, gerar visualizações que tragam repostas para o estudo de caso. [3]

Este trabalho foca na descoberta de padrões por meio de visualização de dados referentes a dados provenientes da NHL.

O estudo de caso consiste em analisar a relação do *gap* entre draft e a primeira temporada, com a performance dos jogadores. Um breve resumo do processo de *draft* e primeira temporada: quando jogadores de hockey são contratados, o termo em questão utilizado é *draft* uma vez que não existe oferta de contrato e sim uma lista com jogadores ordenados por rendimento, onde cada jogador é designado a um time.

O processo de *draft* começa com a lista de times ordenados de acordo com a classificação obtida na competição da temporada anterior, sendo o time de número um o campeão e o time de número trinta e dois, última posição.

Para assignar um jogador ao time, são utilizadas as duas listas, jogadores e times, e é realizado um cruzamento, onde o melhor jogador é assignado ao último time e o jogador em última colocação é assignado ao atual campeão.

Em outras palavras, o time da posição número trinta e dois, último time classificado, receberá o jogador de número um no ranking de melhores

jogadores, o penúltimo time, número trinta e um, receberá o segundo melhor jogador da lista de *draft*, classificado como número dois. Esse processo ocorre até chegar ao time que ficou em primeiro lugar receber o jogador que está como último na lista de *draft*.

Assim que um jogador é “contratado”, assignado ao time referente, o time em questão fica responsável por definir quando o atleta começará sua carreira na NHL. Até esse momento acontecer, o jogador é designado a uma liga inferior para que possa receber treinamento.

O estudo de caso a seguir tem a intenção de relacionar os tempos de *gap*, treinamento com os três primeiros anos de performance de um jogador, assim que ele começa a jogar na liga principal (NHL).

As perguntas que serão abordadas são:

- Existe relação entre a transição imediata da liga júnior para a NHL, com a performance do jogador?

- Existe relação entre o tempo de permanência na liga secundária, com a performance do jogador na NHL?

3. PLATAFORMA

Esse trabalho foi desenvolvido dentro de um *notebook*, na plataforma Google Colabolatory, plataforma similar ao Jupyter Notebook, porém por meio de navegador, *browser*. Em outras palavras, o Colaboratory é um serviço do Jupyter Notebook hospedado que não requer configurações extras para utilização. [4]

O fluxo de trabalho foi realizado conforme processo descrito abaixo:

1. *Download* de dados provenientes e disponíveis na plataforma oficial da NHL [5]

2. Ingestão dos dados utilizando o Google Colaboratory.
3. Pré-processamento dos dados
4. Agrupamento dos dados
5. Associação dos dados
6. Limpeza dos dados
7. Exploração dos dados utilizando-se das bibliotecas: pandas, matplotlib, seaborn e numpy.
8. Visualização dos dados de forma gráfica.

4. DATASET

Para essa análise, foram utilizados *dataset* públicos disponíveis na fonte original da NHL. Todos os *dataset* foram obtidos através do site NHL seção *stats*.

Os *dataset* utilizados foram o BioInfo e Summary, apresentados abaixo.

Dataset Summary com 2173 linhas e seguinte estrutura:

- Player: nome do jogador
- Season: temporada
- Team: time
- S/C: Total de chutes a gol sem sucesso, casos que o goleiro consegue agarrar o disco.
- Pos: posição do jogador
- GP: número de jogos
- G: Total de gols acumulados
- P: Total de pontos acumulados
- A: assistências a gol
- PIM: Total de minutos o qual o jogador ficou detido por penalidade.
- PPG: gols feitos durante pênaltis
- P/GP: pontos por jogo
- PPP: pontos feitos durante pênaltis
- SHG: gols feitos próximos ao gol
- SHP: pontos feitos próximos ao gol
- OTG: gols feitos em tempo extra
- S: chutes
- S%: porcentagem de chutes comparados a todos os jogadores desde a primeira temporada de jogos a qual foi realizada no ano de 1960

- TOI/GP: tempo no gelo por jogo
- FOW%: porcentagem de vitórias

O dataset BioInfo 711 linhas, com a seguinte estrutura:

- Player: nome do jogador
- S/C: Total de chutes a gol sem sucesso, casos que o goleiro consegue agarrar o disco.
- Pos: posição do jogador
- DOB: data de nascimento
- Birth City: Cidade de nascimento
- S/P: estado ou província de nascimento
- Ctry: pais de nascimento
- Ntnlty: nacionalidade
- Ht: altura
- Wt: peso
- Draft Yr: ano da contratação
- Round: rodada
- Overall: qual posição que foi contratado no ranking
- 1st season: primeira temporada jogando
- HDF: faz parte do hall da fama
- G: todos os gols
- P: pontos
- A: assistências a gol

O *dataset* possui valores nulos e tipados de forma inadequada, todos os campos que foram utilizados para obtenção de dados foram tratados.

5. METODOLOGIA UTILIZADA

Os dados foram extraídos por meio de *download* do site original da NHL, na forma de arquivo XLSX (*Microsoft Excel Open XML Spreadsheet*) e foram modificados para a forma de CSV (*comma separated value*). Após o *download* foi realizado o *upload* para o sistema de versionamento de arquivos na plataforma GitHub.

Todo o processo foi realizado utilizando a plataforma Google Colaboratory, onde não foram necessárias configurações e/ou instalar bibliotecas adicionais.

Com todo o ambiente pronto para análise, o *notebook* foi criado e desenvolvido utilizando-se a linguagem Python para limpeza e execução do

projeto.

Para pré-processar os dados, foram utilizadas validações, modificações como:

- Foram renomeados múltiplos campos para facilitar o reconhecimento das variáveis e padronizar os nomes.

```
[ ] # rename bio info columns
nhl_bio_info = nhl_bio_info.rename(columns={"Ht": "height", "Wt": "weight", "HOF": "hall_fame", "GP": "games_played", "G": "Goal","A": "assist", "P": "points"})
nhl_bio_info.head()
```

	Player	S/C	Pos	DOB	Birth City	S/P	Ctry	Ntnlty	height	weight	Draft Yr	Round	Overall	1st Season	hall_fame	games_played	Goal	assist	points
0	Dalton Prout	R	D	1990-03-13	Kingsville	ON	CAN	CAN	75	215	2010	6	154	20112012	N	264	7	31	38
1	Craig Cunningham	R	L	1990-09-13	Trail	BC	CAN	CAN	70	184	2010	4	97	20132014	N	63	3	5	8
2	Radko Gudas	R	D	1990-06-05	Prague	--	CZE	CZE	72	208	2010	3	66	20122013	N	610	31	116	147
3	Calle Jarnkrok	R	C	1991-09-25	Gävle	--	SWE	SWE	71	186	2010	2	51	20132014	N	574	106	135	241
4	Oscar Lindberg	L	L	1991-10-29	Skellefteå	--	SWE	SWE	73	202	2010	2	57	20142015	N	252	39	40	79

- O campo '*First Season'* foi alterado para se tornar uma data, sendo considerado somente o elemento que representa o primeiro ano da temporada.

```
# separate 1st season and convert to year(20192020 --> 2019)
nhl_bio_info['first_season'] = nhl_bio_info['first_season'].astype(str)
nhl_bio_info['first_season'] = nhl_bio_info['first_season'].str[0:4]
nhl_bio_info['first_season'].astype(int)
```

- O campo 'AVG_points_per_game' foi criado com o intuito de gerar métricas de média de pontos por jogo de cada jogador.
- O campo 'min_ice_toi' foi criado e modificado com a intenção de calcular a soma em minutos de tempo por gelo que o jogador permaneceu jogando.
- O campo 'total_p_a_g' foi criado para representar a soma de pontos, assistências e gols feitas por cada jogador sendo agrupadas por season (ano jogado)

```
# sum total p_a_g points, goals and assistances and divide by the time on ice to get points/time_on_ice
df2['total_p_a_g'] = df2['P'] +  df2['G'] +  df2['A']

df2['avg_points_to_ice_time'] = df2['total_p_a_g'] / df2['total_min_toi']
```

- O *dataset* ‘new_df3’ foi criado para calcular a soma de pontos por cada jogador com os demais campos:
    - player: nome do jogador
    - season: temporada
    - games_played: jogos que foram jogados
    - total_p_a_g: total de pontos, gols e assistências
    - avg_points_to_ice_time: média de pontos feitos pelo tempo no gelo

```
# calculate the sum of points for each player
new_df3 = df2[['player','season', 'games_played','total_p_a_g', 'avg_points_to_ice_time']].copy()
new_df3.info()
new_df3.head()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1239 entries, 1438 to 829
Data columns (total 5 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   player                 1239 non-null   object
 1   season                 1239 non-null   object
 2   games_played           1239 non-null   int64
 3   total_p_a_g            1239 non-null   int64
 4   avg_points_to_ice_time 1239 non-null   float64
dtypes: float64(1), int64(2), object(2)
memory usage: 58.1+ KB
```

	player	season	games_played	total_p_a_g	avg_points_to_ice_time
1438	Adam Brooks	2021	25	6	0.011173
960	Adam Erne	2016	26	6	0.008475
963	Adam Erne	2017	23	8	0.013180
964	Adam Erne	2018	65	40	0.063191
856	Adam Gaudette	2018	56	24	0.036530

- O *dataset* ‘new_df4’ foi criado para exibir os seguintes dados:
    - player: nome do jogador
    - first_season: primeira temporada
    - draft_year: ano que foi contratado
    - hall_fame: está no hall da fama

```
[ ] # create new df
    new_df4 = nhl_bio_info[['player', 'first_season', 'draft_year', 'hall_fame']].copy()
    new_df4.info()
    new_df4.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 710 entries, 0 to 709
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   player        710 non-null    object
 1   first_season  710 non-null    object
 2   draft_year    710 non-null    int64 
 3   hall_fame     710 non-null    object
dtypes: int64(1), object(3)
memory usage: 22.3+ KB
```

	player	first_season	draft_year	hall_fame
0	Dalton Prout	2011	2010	N
1	Craig Cunningham	2013	2010	N
2	Radko Gudas	2012	2010	N
3	Calle Jarnkrok	2013	2010	N
4	Oscar Lindberg	2014	2010	N

- Foi realizado um merge com ambos dataset (new_df3 e new_df4) que forma um novo dataset usando o nome do jogador como chave de união.

```
[ ]  # merge 2 df
     new_df_merged = pd.merge(new_df3, new_df4, on="player")
     new_df_merged.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 871 entries, 0 to 870
Data columns (total 8 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   player                  871 non-null    object
 1   season                  871 non-null    object
 2   games_played            871 non-null    int64
 3   total_p_a_g             871 non-null    int64
 4   avg_points_to_ice_time  871 non-null    float64
 5   first_season            871 non-null    object
 6   draft_year              871 non-null    int64
 7   hall_fame               871 non-null    object
dtypes: float64(1), int64(3), object(4)
memory usage: 61.2+ KB
```

```
[ ]  new_df_merged.head()
```

	player	season	games_played	total_p_a_g	avg_points_to_ice_time	first_season	draft_year	hall_fame
0	Adam Brooks	2021	25	6	0.011173	2019	2016	N
1	Adam Erne	2016	26	6	0.008475	2016	2013	N
2	Adam Erne	2017	23	8	0.013180	2016	2013	N
3	Adam Erne	2018	65	40	0.063191	2016	2013	N
4	Adam Gaudette	2018	56	24	0.036530	2017	2015	N

Após a etapa de pré-processamento foi criada a pergunta que dará origem ao tema da pesquisa em questão:

- Como podemos relacionar a performance de um jogador de hockey com o tempo que ele teve de gap entre ser contratado e começar a jogar pelo time.

- Foi criada uma coluna para representar o tempo de *gap* entre a data

de contratação e quando o jogador efetivamente começou a jogar pelo time. ‘diff_1_season_to_draft_year’

```
[ ]  new_df_merged['diff_1season_to_draft_year'] = new_df_merged['first_season_int'] - new_df_merged['draft_year_int']
     new_df_merged.info()
```

Após todas as etapas de pré-processamento terem sido concluídas, foi dado início a visualização dos dados.

Todas as visualizações e desenvolvimento foram realizados com a utilização das bibliotecas: Pandas, Matplotlib, Seaborn e Numpy.

6. RESULTADO

Na fase inicial da análise foram realizadas análises e visualizações para visualizar quais seriam os tempos *gaps* que iriamos utilizar.

Verificou-se que o *gap*, diferença entre ano de *draft* o qual o jogador foi contratado e o ano que ele começou a jogar, variam entre 0 e 7 anos.

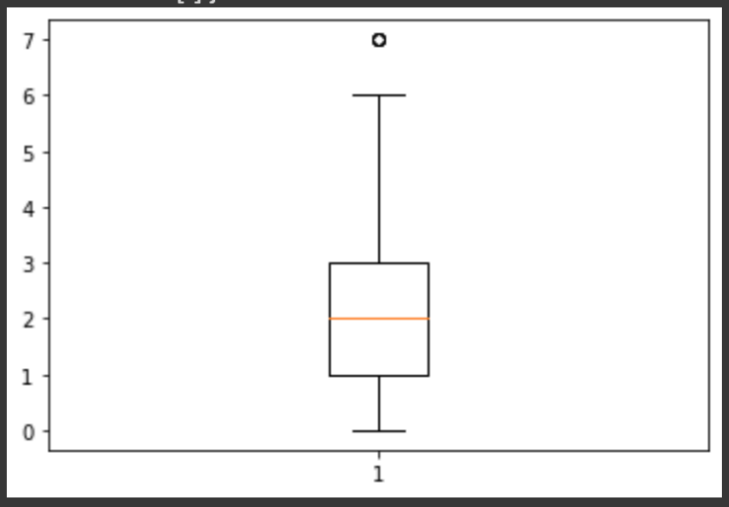
Em ambas as visualizações gráficas, podemos ver a diferença da primeira temporada de um jogador com a data de contratação.

Observamos que a diferença tem um range que varia de 1 até 7.

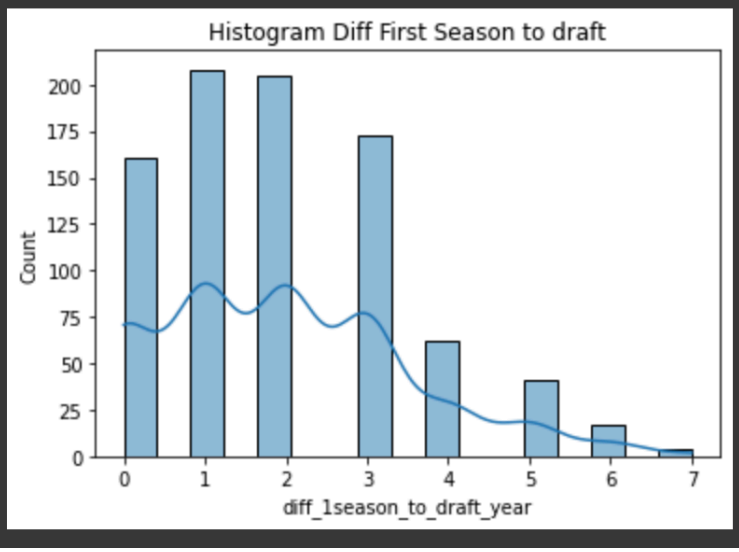
```
'''understanding the spread of the difference in years
from a player being drafted to having his first season playing in the nhl '''

plt.boxplot(new_df_merged['diff_1season_to_draft_year'])
```

```
{'whiskers': [<matplotlib.lines.Line2D at 0x7f9de6133b50>,
  <matplotlib.lines.Line2D at 0x7f9de60be0d0>],
 'caps': [<matplotlib.lines.Line2D at 0x7f9de60be310>,
  <matplotlib.lines.Line2D at 0x7f9de60be890>],
 'boxes': [<matplotlib.lines.Line2D at 0x7f9de6133550>],
 'medians': [<matplotlib.lines.Line2D at 0x7f9de60bee10>],
 'fliers': [<matplotlib.lines.Line2D at 0x7f9de6121ad0>],
 'means': []}
```

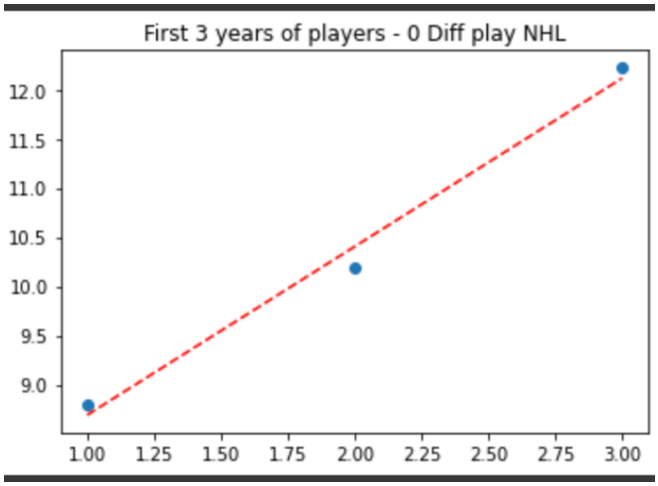



```
[ ] ax = sns.histplot(data = new_df_merged, x = 'diff_1season_to_draft_year', kde = True)
    ax.set_title("Histogram Diff First Season to draft")
    plt.show()
```

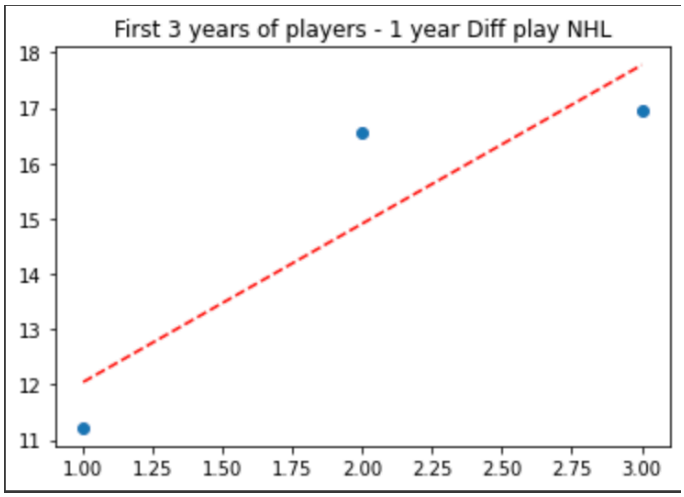


Como passo seguinte a análise, foram geradas visualizações para cada grupo de jogadores, nos seus 3 primeiros anos jogando, após começarem a jogar pela NHL.

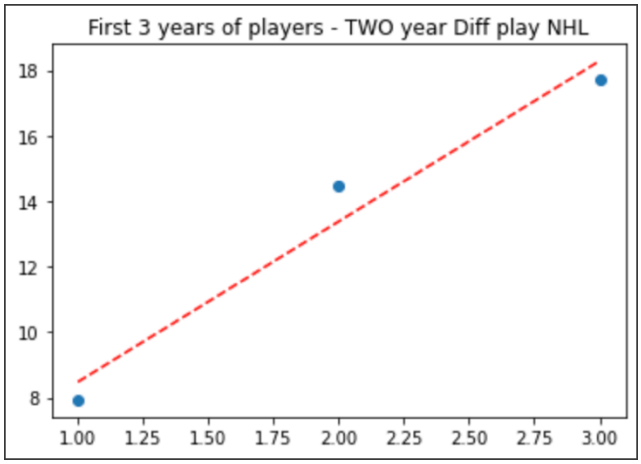
- (***gap* = 0**) – Jogadores que começaram a jogar no ano em que foram contratados.



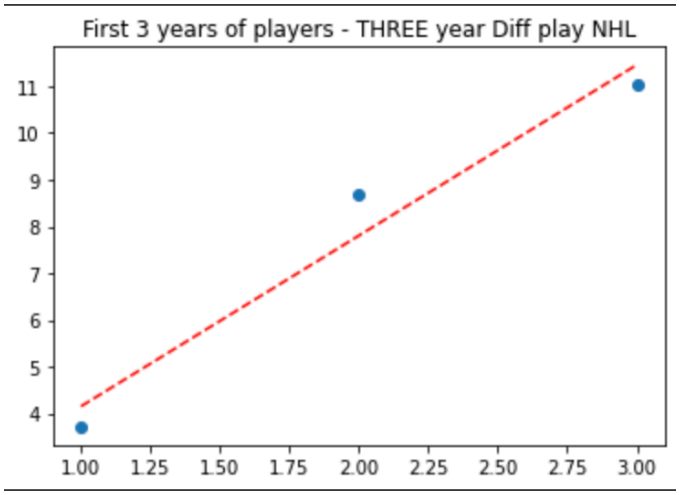
- (***gap* = 1**) – Jogadores que começaram a jogar 1 ano após o ano em que foram contratados.



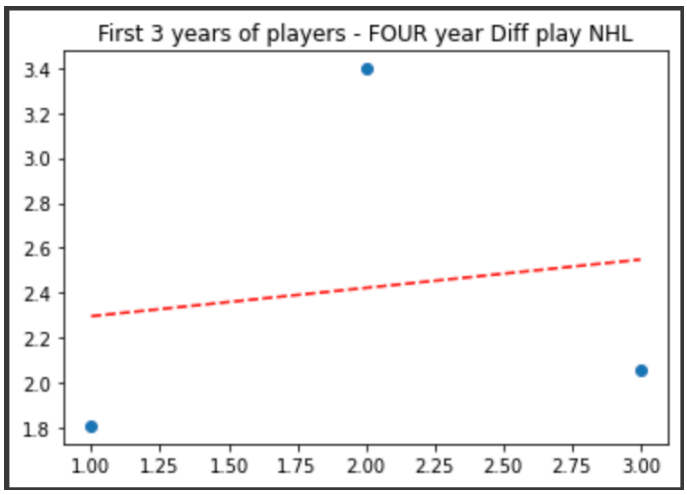
- (***gap* = 2**) – Jogadores que começaram a jogar 2 anos após o ano em que foram contratados.



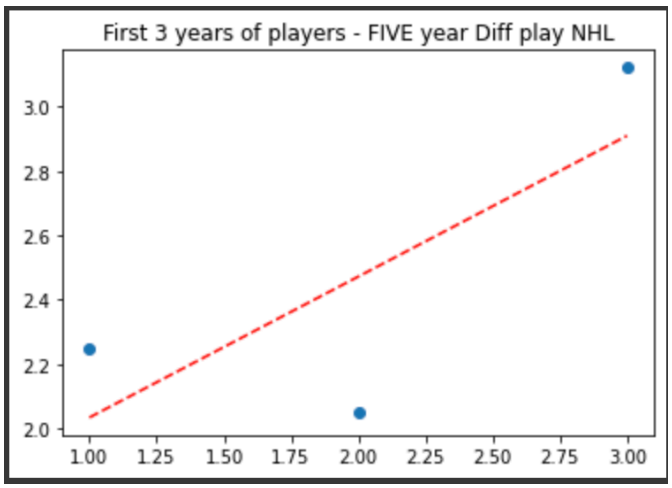
- (***gap* = 3**) – Jogadores que começaram a jogar 3 anos após o ano em que foram contratados.



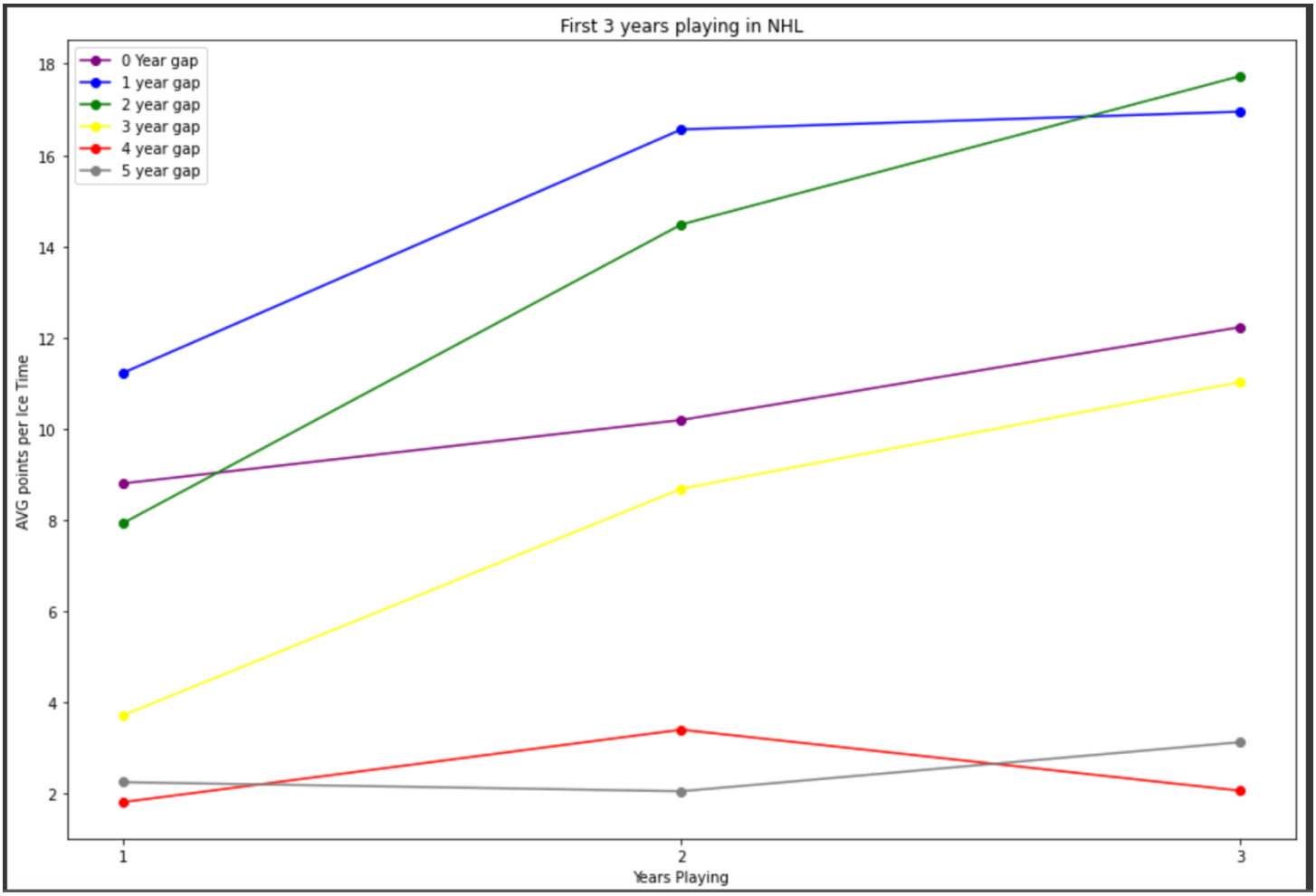
- (***gap* = 4**) – Jogadores que começaram a jogar 4 anos após o ano em que foram contratados.



- (***gap* = 5**) – Jogadores que começaram a jogar 5 anos após o ano em que foram contratados.



Como última etapa de análise e visualização, foi criado um *plot* com todos os *gaps* para que pudessem ser comparados, facilitando a visualização por possuírem o mesmo *yaxis*.



7. CONCLUSÃO E TRABALHOS FUTUROS

Visualização e processamento de dados são considerados cada vez mais importantes para a tomada de decisão. As etapas de coleta, organização, análise e síntese são essenciais para decisões e impacto. Técnicas de processamento e visualização de dados para tomada de decisão são utilizadas em inúmeros ramos e áreas.[7]

Este trabalho teve como principal objetivo a realização de um estudo de caso utilizando dados reais e disponíveis publicamente para análise de dados da relação entre performance de jogadores com o tempo de *gap* que ficam entre serem contratados e começarem a efetivamente jogar pelo time.

Os objetivos a serem contemplados eram relacionados a análise em questão e ao uso de diferentes bibliotecas, bem como o uso de *notebook* para manipular dados em diferentes *dataset* e realizando interação de todos os dados, bibliotecas e visualizações no mesmo ambiente.

A partir desse estudo de caso, foi possível visualizar individualmente *trendlines* referentes a cada *gap* de tempo no qual podemos a existência de uma correlação positiva. Além disso verificou-se que quando realizamos os *plots*, notamos que quanto maior os *gap*s, menor a performance de um grupo de jogadores. Jogadores que possuem diferença de 4 e 5 anos a começarem a jogar depois de terem sido contratados, são jogadores que tem medias inferiores aos demais.

Podemos visualizar também que jogadores que possuem entre um e dois anos de *gap*, tem melhores performances comparados aos demais.

Podemos concluir que entre um e dois anos é o melhor *gap* que um jogador pode ter, ser contratado e ser treinado pelo período entre 1 e 2 anos apresentam melhores rendimentos, já que são jogadores que começam com uma boa média de pontos na temporada e tem um crescimento linear positivo considerável, é notório o ganho de performance em três temporadas.

Quando analisamos jogadores que vem direto da liga júnior, apesar de serem jogadores que começam com uma boa média de pontos na temporada, identificamos muito pouco aumento de performance, conseguimos identificar um crescimento linear positivo, porém quase nulo, já que a média de pontos segue quase que com o mesmo valor após três temporadas.

Em relação aos jogadores que tem entre três e cinco anos de *gap,* é possível identificar baixa média de pontos na primeira temporada.

Para os jogadores com três anos de *gap* notamos um ganho de performance bastante similar aos jogadores de 2 anos de *gap,* porém ao final de três temporadas ainda permanecem abaixo da média de pontos.

Para os jogadores com quatro anos, identificamos uma queda de performance passando a segunda temporada. Já os jogadores com 5 anos, apresentam uma queda de performance no primeiro ano e após existe uma melhora, ainda que muito discreta.

Durante o processo de análise e desenvolvimento foram notadas

oportunidades de melhorias e iterações futuras. Podemos citar algumas como melhorar a predição de dados uma vez que possamos utilizar um *range* maior de dados com mais temporadas, logo mais jogadores. Gerar mais visualizações com dados disponíveis no *dataset* como exemplo, podemos gerar visualizações baseadas em posições de jogadores para entender quais seriam as performances de cada tipo de jogador relacionado a sua posição de jogo.

Para trabalhos futuros seria interessante avaliar e entender a relação de performance com tipos de treinos realizado no período de *gap* de um jogador. Outra implementação possível seria agregar dados médicos para avaliar qualquer tipo de lesão de jogadores que podem afetar seu rendimento, entre outros dados que podemos agregar para gerar mais visualizações.

O *notebook* completo está disponível em um repositório no GitHub [8].

8. REFERENCIAS

1.Minerando dados: o que é, importância e ferramentas.
Disponível em:
https://www.totvs.com/blog/negocios/mineracao-de-dados/ Acesso em: 28 de outubro de 2022.

2.MALLEY, B.; RAMAZZOTTI, D.; WU, J. T. Data Pre-processing. **Secondary Analysis of Electronic Health Records**, p. 115–141, 2016.
Disponível em:
https://link.springer.com/chapter/10.1007/978-3-319-43742-2_12 Acesso em 28 de outubro de 2022.

3. CHEN, C.; HÄRDLE, W.; UNWIN, A. **Handbook of Data Visualization**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.
Disponível em:
https://link.springer.com/book/10.1007/978-3-540-33037-0 Acesso em 28 de outubro de 2022.

4.Frequently Asked Questions.

Disponível em:
https://research.google.com/colaboratory/faq.html Acesso em 28 de outubro de 2022.

5.NHL Stats.
Disponível em:
https://www.nhl.com/stats/ Acesso em 28 de outubro de 2022.

6.SHEN-HSIEH, A.; SCHINDL, M. Data visualization for strategic decision making. **Case studies of the CHI2002|AIGA Experience Design FORUM on - CHI '02**, 2002.
Disponível em:
https://dl.acm.org/doi/10.1145/507752.507756 Acesso em 28 de outubro de 2022.

7.MOORE, J. Data Visualization in Support of Executive Decision Making. **Interdisciplinary Journal of Information, Knowledge, and Management**, v. 12, p. 125–138, 2017.
Disponível em:
http://www.ijikm.org/Volume12/IJIKMv12p125-138Moore2889.pdf Acesso em 28 de outubro de 2022.

8. Repositório GitHub com *notebook* do projeto.
Disponível em:
https://github.com/PKpacheco/TCC_PUC_DataScience/blob/main/NHL_Data.ipynb