Course Completion Work

GRADUATE IN DATA SCIENCE AND ARTIFICIAL INTELLIGENCE

Analysis of the relationship between training time and performance for professional hockey players

Paola Katherine Cardoso Pacheco
ADVISOR: Marta Ribeiro Hentschke

PUCRS online uol edtech

Content

1. SUMMARY

Data mining is extremely important when dealing with strategy, since it is possible to detect trends, predict some results, model, and collect information, and recognize patterns. [1]

In this work we will explore data from the National Hockey League (NHL), which provides data from previous years related to hockey metrics. The entire construction of the study is based on manipulating data in Google Collaboratory.

This research presents the study of metrics related to time difference gaps from the date of hiring a player to his early career playing in a season.

After cleaning data, the performance averages based on the time gaps that groups of players with the same gap have between draft, their hiring and their first season were constructed by regression lines.

In addition to data cleaning, data visualization was used to make it clearer and facilitate the graphical interpretation of the difference in performance.

2. INTRODUCTION

Data preprocessing is a fundamental pillar of what we now know as Big Data. For data analysis and visualization, the pre-processing stage is essential.

The study consists of steps ranging from data extraction, cleaning, standardization, and reorganization of data so that it is possible to perform analyses after having normalized data. [2]

Preprocessing consists of applying cleaning techniques, normalization, organization of data so that it can be possible to perform analyses and visualizations.

After data cleansing, it is necessary to mine, apply methods to search for patterns in the data and, with this, generate visualizations that bring responses to the case study. [3]

This work focuses on finding patterns by visualizing data for data from the NHL.

The case study consists of analyzing the relationship of the gap between draft and the first season, with the performance of the players. A brief summary of the draft process and first season: when hockey players are signed, the term in question used is draft since there is no contract offer, but a list of players ordered by performance, where each player is assigned to a team.

The draft process begins with the list of teams ordered according to the ranking obtained in the competition of the previous season, with the number one team being the champion and the number thirty-two team, last position.

To assign a player to the team, the two lists, players and teams are used, and a cross is made, where the best player is assigned to the last team and the player in last place is assigned to the current champion.

In other words, the team of position number thirty-two, last ranked team, will receive the number one player in the ranking of best players, the penultimate team, number thirty-one, will receive the second-best player on the draft list, ranked number two. This process occurs until the team that was in first place receives the player that is last on the draft list.

Once a player is "hired", assigned to the referring team, the team in question is responsible for defining when the athlete will begin his career in the NHL. Until that time happens, the player is assigned to a lower league so that he can receive training.

The following case study is intended to relate gap times in training to the first three years of a player's performance once he starts playing in the major leagues (NHL).

The questions that will be addressed are:

- Is there a relationship between the immediate transition from the junior league to the NHL, to the player's performance?

- Is there a relationship between the length of stay in the minor leagues, and the player's performance in the NHL?

3. PLATAFORM

This work was developed inside a notebook, on the Google Collaboratory platform, a platform similar to Jupyter Notebook, but through a browser, browser. In other words, Collaboratory is a hosted Jupyter Notebook service that does not require extra settings for use. [4]

The workflow was performed according to the process described below:

1. *Download* data from and available on the NHL's official platform [5]
2. Data ingestion using Google Collaboratory.
3. Data preprocessing
4. Grouping of data
5. Data pool
6. Data cleansing
7. Data exploration using libraries: pandas, matplotlib, seaborn and

NumPy.

8. Visualization of the data graphically.

4. DATASET

For this analysis, public *datasets available* in the original NHL source were used. All *datasets* were obtained through the NHL stats *section website*.

The datasets used were BioInfo and Summary, presented below.

Dataset Summary with 2173 lines and following structure:

- Player
- Season
- Team:
- S/C: Total unsuccessful shots on goal, cases in which the goalkeeper manages to grab the puck.
- Pos: player position
- GP: number of games
- G: Total accumulated scores
- P: Total accumulated points
- A: Total assists on goal
- PIM: Total minutes which the player was held for penalty.
- PPG: goals scored during penalties
- P/GP: points per game
- PPP: points made during penalty shootouts
- SHG: goals scored near goal
- SHP: points scored near goal
- OTG: goals scored in extra time
- S: shoots
- S%: percentage of shoots compared to all players since the first season of games which was held in the year 1960

- TOI/GP: ice time per game
- FOW%: winning percentage

The BioInfo 711 rows dataset, with the following structure:

- Player: player name
- S/C: Total unsuccessful shots on goal, cases in which the goalkeeper manages to grab the puck.
- Pos: player position
- DOB: date of birth
- Birth City
- S/W: state or province of birth
- Ctry: Country
- Ntnlty: nationality
- Ht: Height
- Wt: Weight
- Draft Yr: year it was drafted
- Round
- Overall: position that was drafted
- 1st season: first season playing
- HDF: is part of the hall of fame
- G: Total accumulated scores
- P: Total accumulated points
- A: Total assists on goal

The dataset has null and improperly typed values, all fields that were used to obtain data were treated.

5. METHODOLOGY USED

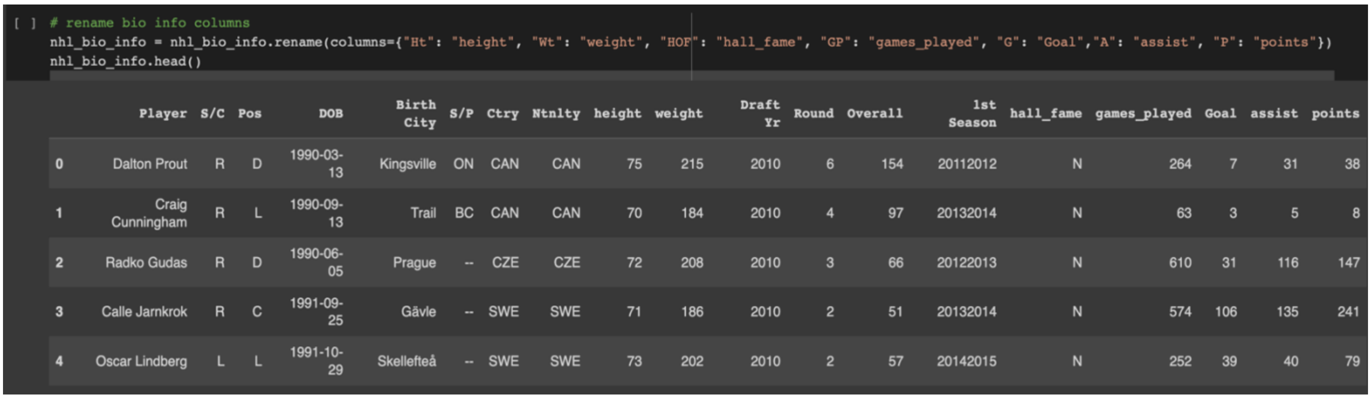
The data was extracted by downloading from the original NHL website in the form of a Microsoft Excel Open XML Spreadsheet (XLSX) file and modified to the form of CSV (comma separated value). After the download, the file versioning system was uploaded to the GitHub platform.

The whole process was carried out using the Google Collaboratory platform, where no configurations were required and/or installed additional libraries.

With the entire environment ready for analysis, the notebook was created and developed using the Python language for data cleaning and execution of the project.

To pre-process the data, validations were used, modifications such as:

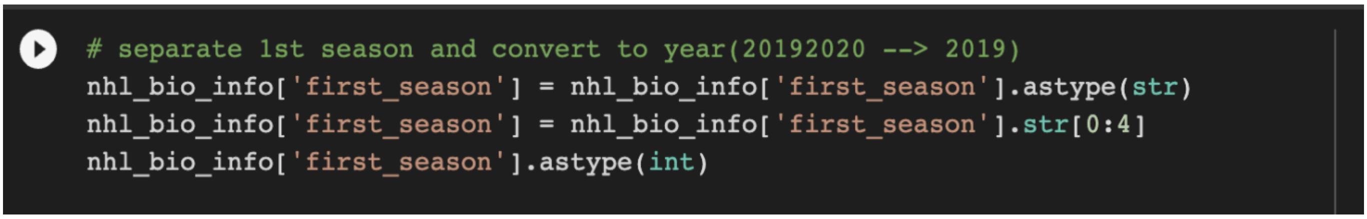
- Multiple fields have been renamed to facilitate the recognition of variables and standardize names.



```
# 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

- The '*First Season'* field has been changed to become a date, being considered only the element that represents the first year of the season.

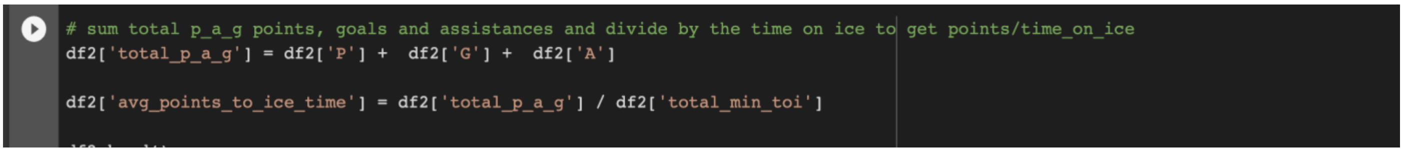


```
# separate 1st season and convert to year(20192020 --> 2019)
nhl_bio_info['first_season'] = nhl_bio_info['first_season'].astype(str)
nhl_bio_info['first_season'] = nhl_bio_info['first_season'].str[0:4]
nhl_bio_info['first_season'].astype(int)
```

- The 'AVG_points_per_game' field was created to generate metrics of average points per game for each player.
- The 'min_ice_toi' field was created and modified with the intention

of calculating the sum in minutes of ice time that the player remained playing.

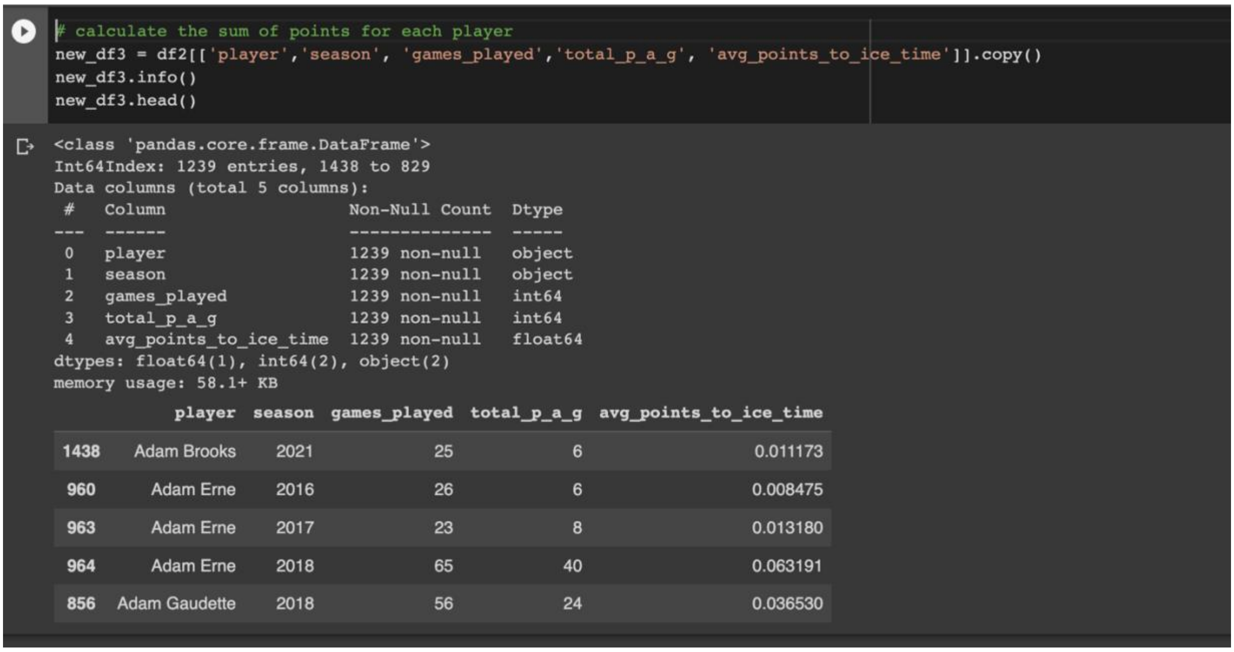
- The 'total_p_a_g' field was created to represent the sum of points, assists and goals made by each player being grouped by season (year played)



```
# 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']
```

- The ‘new_df3' dataset was created to calculate the sum of points for each player with the other fields:
    - player: player name
    - season
    - games_played
    - total_p_a_g: total points, goals and assists
    - avg_points_to_ice_time: average points made by time on ice



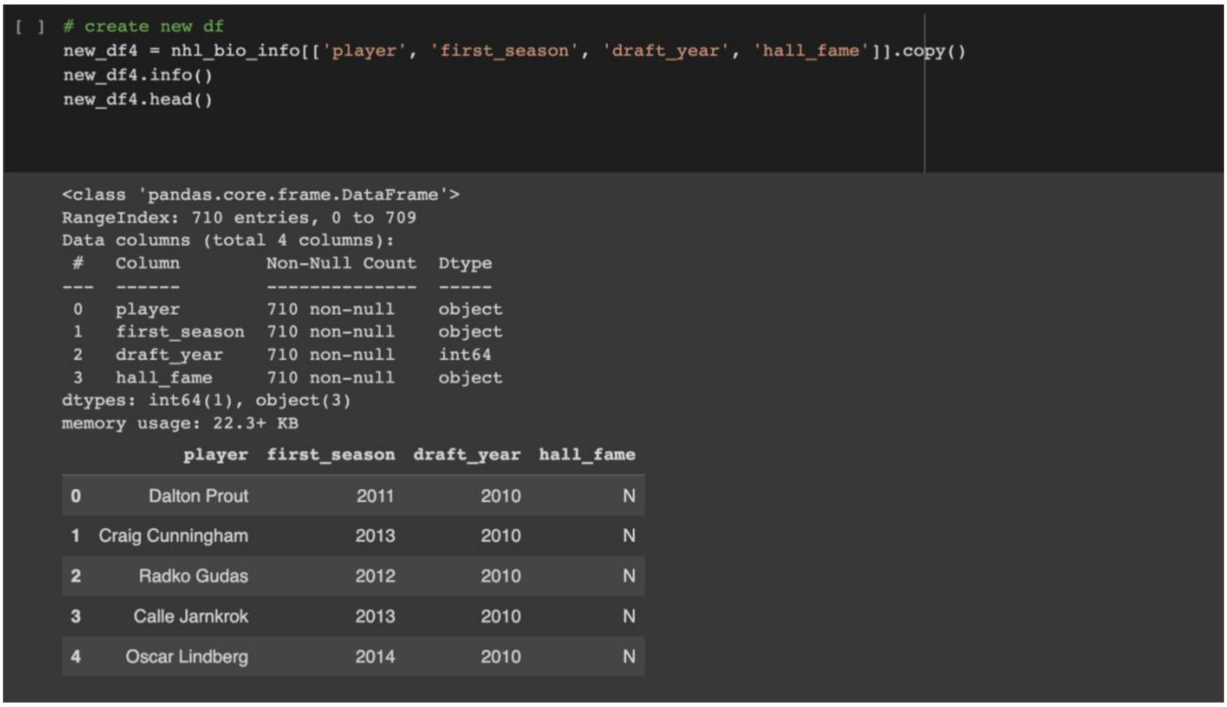
```
# calculate the sum of points for each player
new_df3 = df2[['player','season', 'games_played','total_p_a_g', 'avg_points_to_ice_time']].copy()
new_df3.info()
new_df3.head()
```

```
<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

- The ‘new_df4' dataset was created to display the following data:
    - player

- first_season
- draft_year
- hall_fame

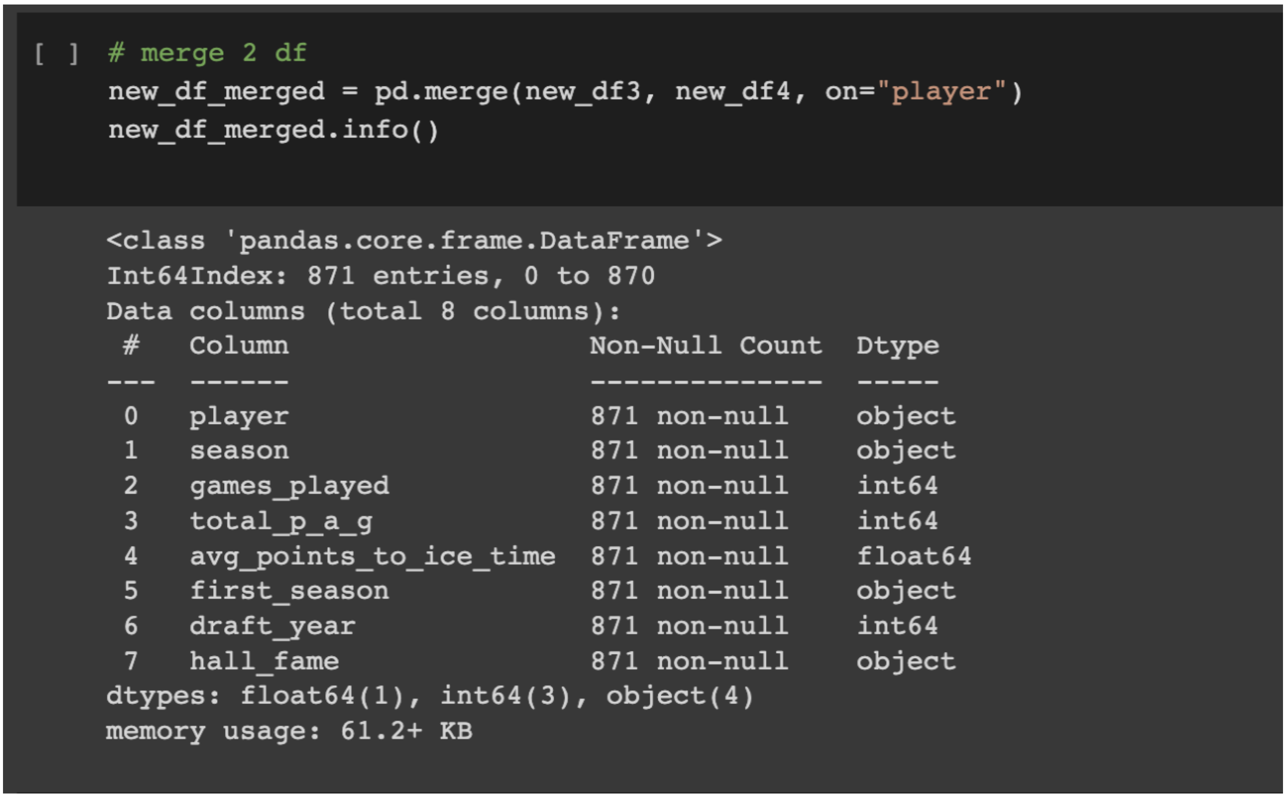


```
[ ] # create new df
    new_df4 = nhl_bio_info[['player', 'first_season', 'draft_year', 'hall_fame']].copy()
    new_df4.info()
    new_df4.head()
```

```
<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
```

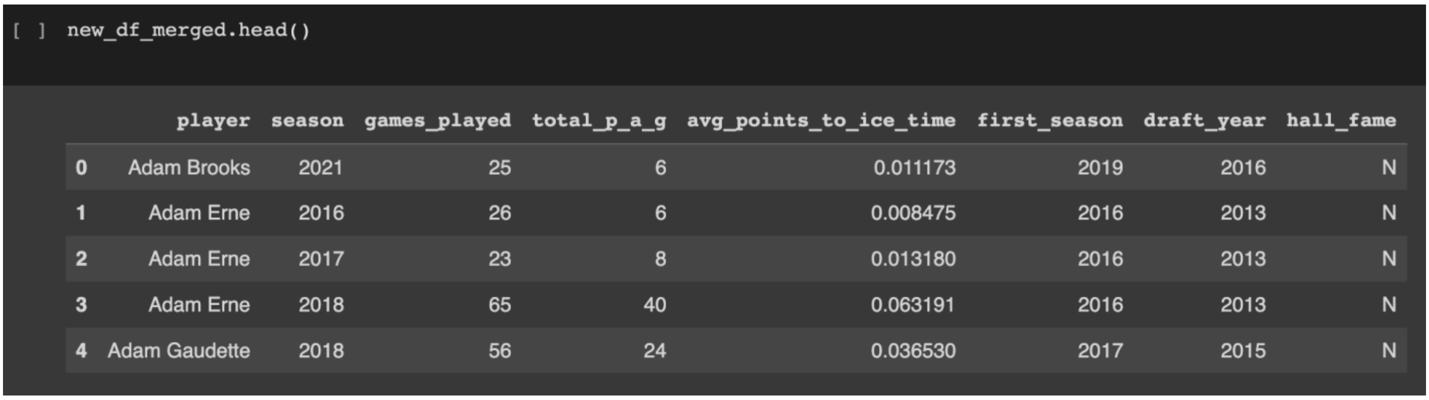
	player	first_season	draft_year	hall_fame
0	Dalton Prout	2011	2010	N
1	Craig Cunningham	2013	2010	N
2	Radko Gudas	2012	2010	N
3	Calle Jarnkrok	2013	2010	N
4	Oscar Lindberg	2014	2010	N

- A merge was performed with both dataset (new_df3 and new_df4) that forms a new dataset using the player's name as the join key.



```
[ ]  # 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()
```

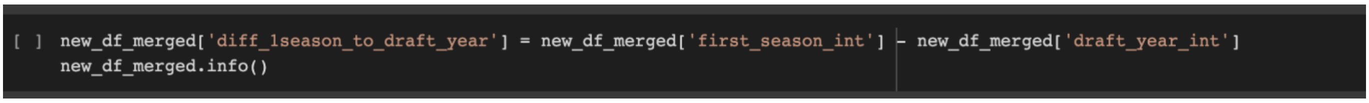
	player	season	games_played	total_p_a_g	avg_points_to_ice_time	first_season	draft_year	hall_fame
0	Adam Brooks	2021	25	6	0.011173	2019	2016	N
1	Adam Erne	2016	26	6	0.008475	2016	2013	N
2	Adam Erne	2017	23	8	0.013180	2016	2013	N
3	Adam Erne	2018	65	40	0.063191	2016	2013	N
4	Adam Gaudette	2018	56	24	0.036530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

After the pre-processing step, the question that will give rise to the research topic in question was created:

- How can we relate the performance of a hockey player to the time interval he had between being signed and starting to play for the team?

- A column was created to represent the *gap time* between the signing date and when the player effectively started playing for the team.

'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()
```

After all the preprocessing steps have been completed, the data was previewed.

All visualizations and development were performed using the libraries: Pandas, Matplotlib, Seaborn and Numpy.

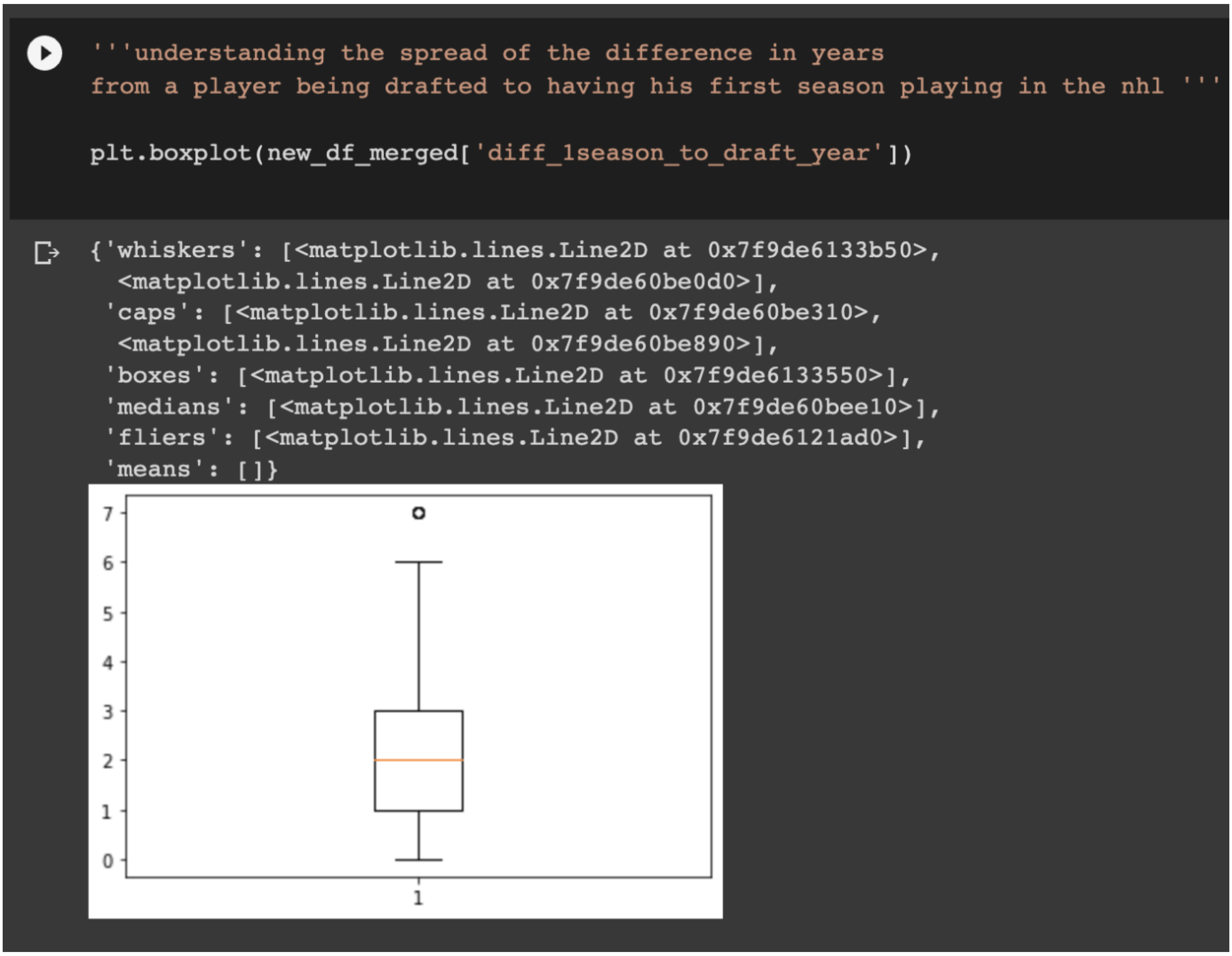
6. RESULT

In the initial phase of the analysis, analyses and visualizations were performed to visualize what gaps *times* we would use.

It was found that *the gap*, difference between *the year of the draft* that the player was hired and the year he started playing, vary between 0 and 7 years.

In both graphical views, we can see the difference of a player's first season with the signing date.

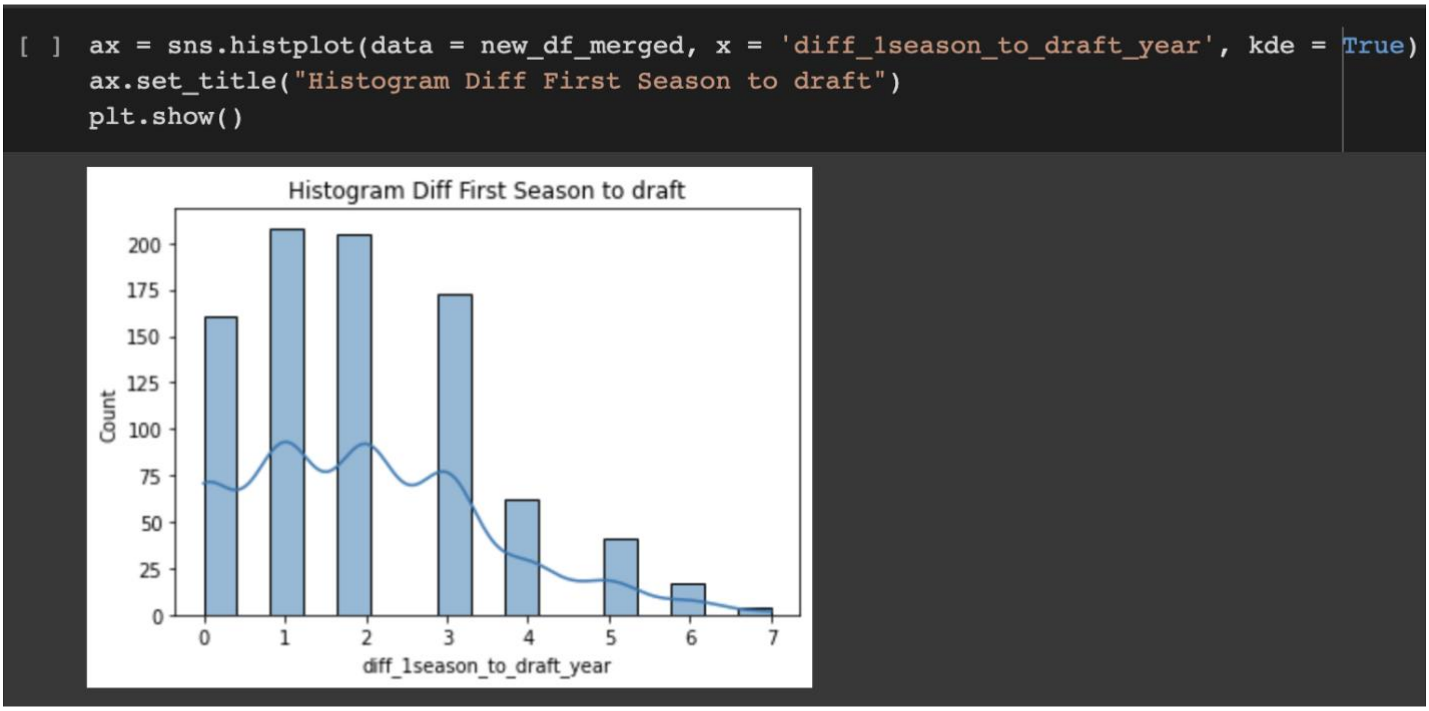
We observed that the difference has a range ranging from 1 to 7.



```
'''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': []}
```

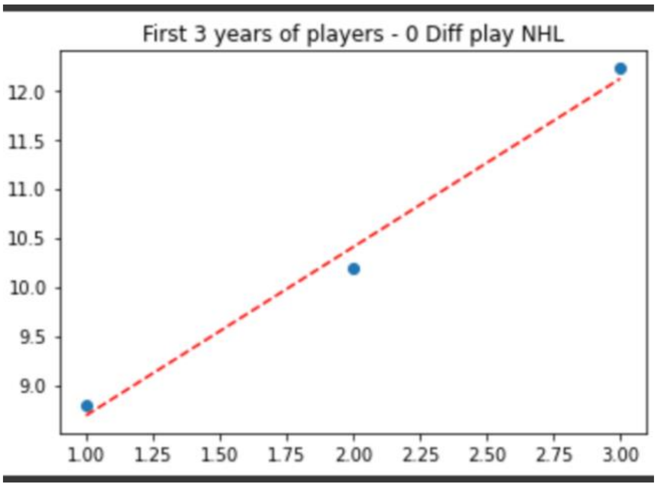


```
ax = sns.histplot(data = new_df_merged, x = 'diff_1season_to_draft_year', kde = True)
ax.set_title("Histogram Diff First Season to draft")
plt.show()
```

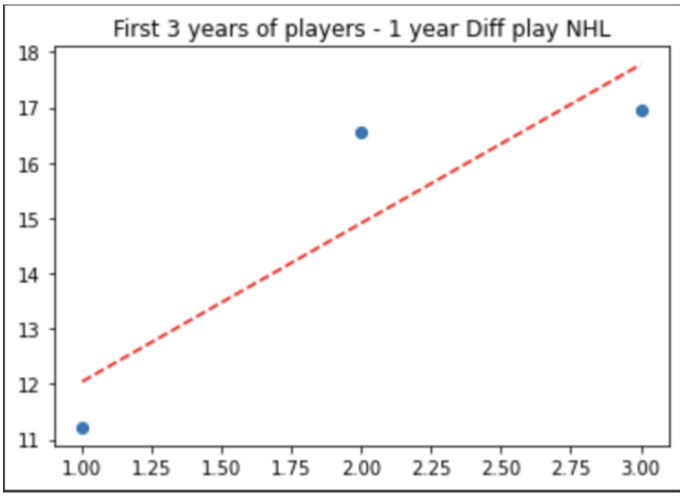
Histogram Diff First Season to draft
200
175
150
125
100
75
50
25
0
Count
0
1
2
3
4
5
6
7
diff_1season_to_draft_year

As a step following the analysis, visualizations were generated for each group of players, in their first 3 years playing, after starting to play for the NHL.

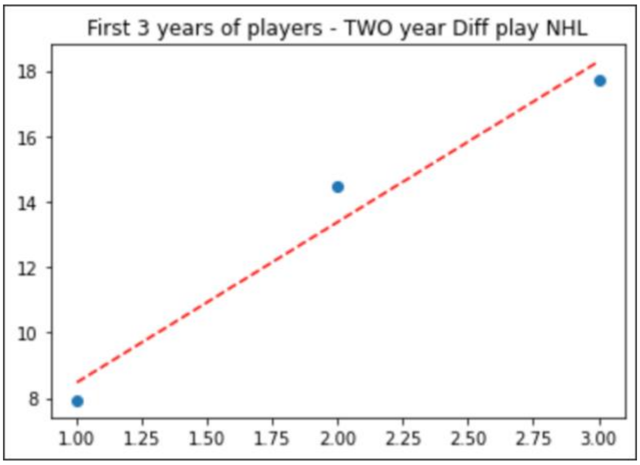
- (***gap* = 0**) – Players who started playing the year they were hired.



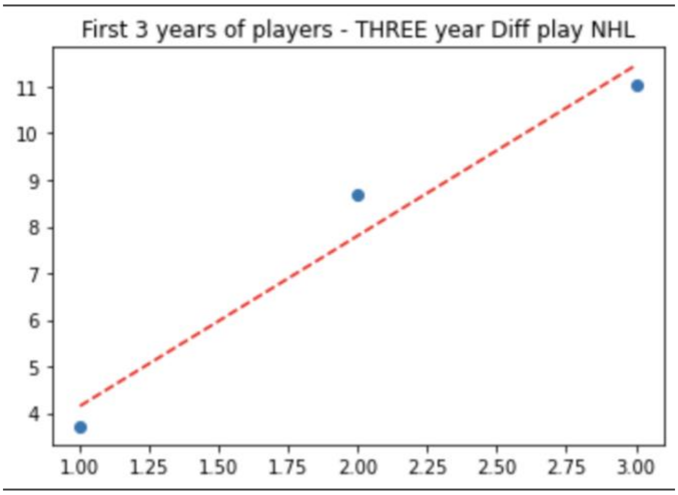
- (***gap* = 1**) – Players who started playing 1 year after the year they were hired.



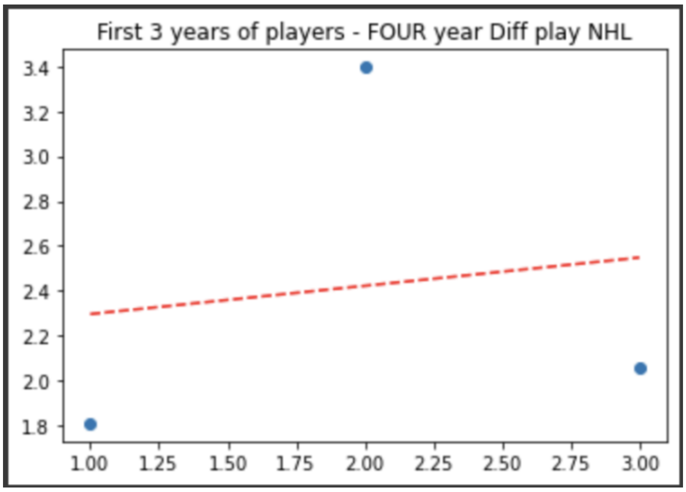
- (***gap* = 2**) – Players who started playing 2 years after the year they were hired.



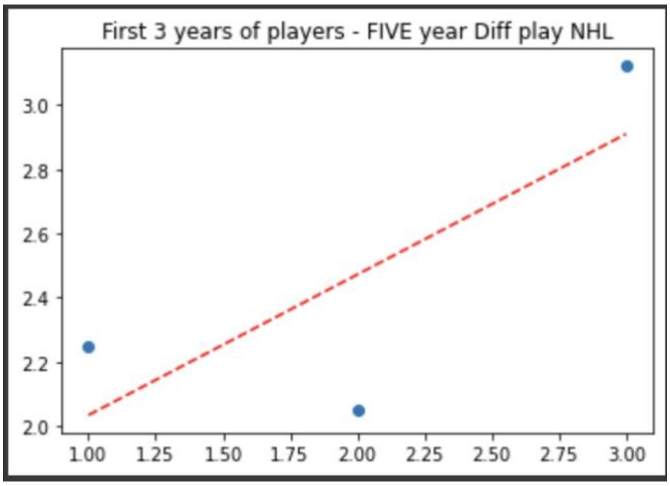
- (***gap* = 3**) – Players who started playing 3 years after the year they were hired.



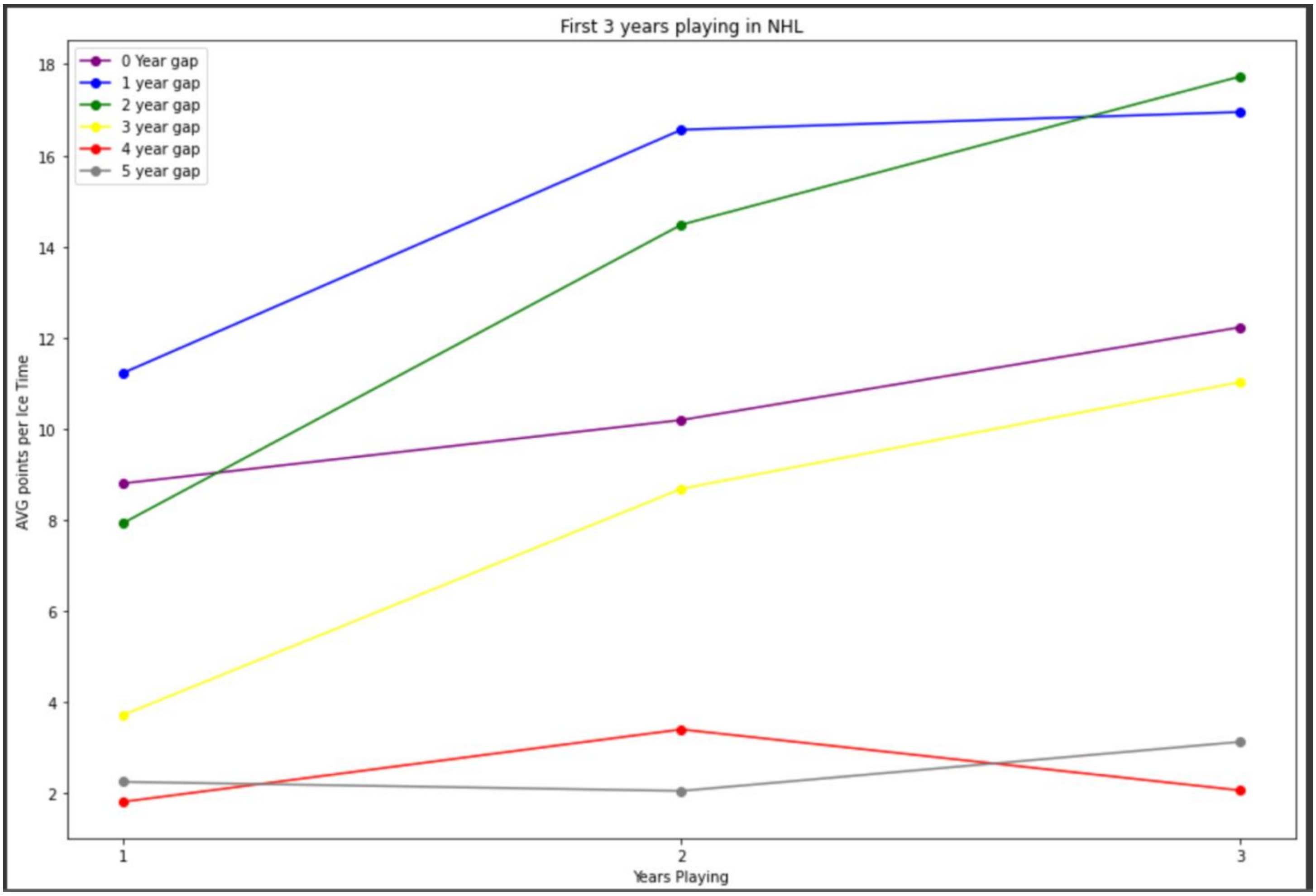
- (***gap* = 4**) – Players who started playing 4 years after the year they were hired.



- (***gap* = 5**) – Players who started playing 5 years after the year they were hired.



As a last step of analysis and visualization, a *plot was created* with all *the gaps* so that they could be compared, facilitating the visualization because they have the same *yaxis*.



7. CONCLUSION AND FUTURE WORK

Data visualization and processing are considered increasingly important for decision making. The stages of collection, organization, analysis, and synthesis are essential for decisions and impact. Data processing and visualization techniques for decision making are used in numerous branches and areas. [7]

This work had as main objective the realization of a case study using real data and publicly available for data analysis of the relationship between performance of players with the gap time that are between being drafted and beginning to effectively play for the team.

The objectives to be contemplated were related to the analysis in question and the use of different libraries, as well as the use of a notebook to manipulate data in different datasets and interacting with all data, libraries, and visualizations in the same environment.

From this case study, it was possible to individually visualize trendlines referring to each time gap in which we can have a positive correlation. In addition, it was found that when we performed the plots, we noticed that the greater the gaps, the lower the performance of a group of players. Players who have a difference of 4 and 5 years starting to play after being signed, are players who have lower averages than the others.

We can also see that players who have between one and two years of gap, have better performances compared to the others.

We can conclude that between one and two years is the best gap a player can have, being hired, and being trained for the period between 1 and 2 years present better performance, since they are players who start with a good average of points in the season and have considerable positive linear growth, the performance gain in three seasons is notorious.

When we analyze players who come straight from the junior league, despite being players who start with a good average of points in the season, we identify very little performance increase, we can identify a positive linear growth, but almost nil, since the average points follows almost with the same value after three seasons.

Regarding players with between three and five years of gap, it is possible to identify a low average of points in the first season.

For players with a three-year gap, we noticed a performance gain quite similar to players with a 2-year gap, but at the end of three seasons they remain below the average points.

For four-year-old players, we identified a drop in performance after the second season. Players 5-year gap, on the other hand, present a drop in performance in the first year and after that there is an improvement, albeit very discreet.

During the analysis and development process, opportunities for improvement and future iterations were noted. We can cite somehow to improve data prediction once we can use a wider range of data with more seasons, then more players. Generate more views with data available in the dataset as an example, we can generate views based on player positions to understand what would be the performances of each type of player related to their playing position.

For future work it would be interesting to evaluate and understand the performance relationship with types of training performed in the gap period of a player. Another possible implementation would be to aggregate medical data to assess any type of player injury that could affect their performance, among other data that we can aggregate to generate more views.

The full notebook is available in a repository on GitHub [8].

8. REFERENCES

1.Mining data: what it is, importance and tools.
Available in:
https://www.totvs.com/blog/negocios/mineracao-de-dados/ Theend of: October 28, 2022.

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

3. CHEN, C.; HÄRDLE, W.; UNWIN, A. **Handbook of Data Visualization**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.
Available in:
https://link.springer.com/book/10.1007/978-3-540-33037-0 Access On October 28, 2022.

4.Frequently Asked Questions.

Available in:
https://research.google.com/colaboratory/faq.html Access On October 28, 2022.

5.NHL Stats.
Available in:
https://www.nhl.com/stats/ Access on October 28, 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.
Available in:
https://dl.acm.org/doi/10.1145/507752.507756 Access On October 28, 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.
Available in:
http://www.ijikm.org/Volume12/IJIKMv12p125-138Moore2889.pdf Access on October 28, 2022.

8. GitHub repository *with* project notebook.
Available in:
https://github.com/PKpacheco/TCC_PUC_DataScience/blob/main/NHL_Data.ipynb