**Data Profile:**

**6)Data Cleaning Process**

-No duplicates found

-124 missing values on Shooting\_Percentage column (replaced missing values with zeros since the players have scored 0 goals and therefore will have 0 shooting percentage)

-311 missing values on Faceoff\_Percentage column (replaced missing values with zeros since players are not centermen therefore they haven’t taken any faceoffs)

-1 missing value in Blocks column (replaced with a zero since the player only played 1 game/All of these single missing values are from the same player)

-1 missing value in Faceoff\_Wins column (replaced with a zero since the player played 1 game and was defenseman (position doesn’t take faceoffs)

-1 missing value in Faceoff\_Losses column (replaced with a zero since the player played 1 game and was defenseman (position doesn’t take faceoffs)

-Changed column names to be more understandable/ I included a key to what each column means and what the names were changed to (its below)

-removed EV.1,PP.1,SH.1 columns (duplicated column/unnecessary)

-I changed the ‘Hart\_Winner?’column into a new column ‘Hart\_Trophy?’. I changed 0 to No and 1 to Yes with yes meaning the player won the Hart Trophy that year and no meaning the player didn’t. I then deleted the old ‘Hart\_Winner?’ column.

-I reordered the columns to where the season/year at the front of the dataframe

-I removed the 2004 player data because I decided to analyze the salary cap era (2005 season to now)

-Started with a data shape of (13,186, 31) and ended with data shape of (12,270, 28)

-Added a new variable called ‘Playtime\_Status’ that divides players time\_on\_ice into categories of playtime in a season. (ex. Little playtime, top minutes, etc.)

-Added a new variable called ‘State\_Province’ that includes the state/provine a team plays in

**7) Understanding the Data (Added Columns are in RED)**

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Column Description | Data Type | Time Variant |
| Season | Which NHL season/year the player played in | Quantitative | Yes |
| Rank | Players ranked by total points earned that season | Quantitative | No |
| Player | Player name | Qualitative | No |
| Age | Age of the player that season | Quantitative | Yes |
| Position | Position the player plays | Qualitative | No |
| Team | Team the player plays on that season | Qualitative | No |
| State\_Province | Includes the state/province a team plays in | Qualitative | No |
| Games\_Played | Total games played that season for a player | Quantitative | No |
| Goals | Total goals scored by a player that season | Quantitative | No |
| Assists | Total assists by a player that season | Quantitative | No |
| Points | Total points (goals and assists combined) by a player that season | Quantitative | No |
| Plus\_Minus | Calculated by subtracting the total number of goals allowed by a player's team while the player is on the ice from the total number of goals scored by the player's team while the player is on the ice | Quantitative | No |
| Penalty\_ Minutes | Minutes of penalties a player has been assessed in that season | Quantitative | No |
| Point\_Shares | An estimate of the number of points contributed by a player. [Calculating Points Shares](https://www.hockey-reference.com/about/point_shares.html) (better explanation) | Quantitative | No |
| Even\_Strength\_Goals | Total goals scored while neither your team or opposing team is on a powerplay | Quantitative | No |
| Powerplay\_Goals | Goals scored by player when his team is on a powerplay | Quantitative | No |
| Shorthanded\_Goals | Goals scored by a player when the opposing team is on a powerplay | Quantitative | No |
| Game\_Winning\_Goals | Game winning goals scored by a player in a season. | Quantitative | No |
| Shots | Total shots on goal in a season by a player | Quantitative | No |
| Shooting\_Percentage | Calculated by dividing a player’s goals by shots on goal | Quantitative | No |
| Time\_On\_Ice | Time (in minutes) spent in all games in a season | Quantitative | No |
| Playtime\_Status | Creating a new variable that divides players time\_on\_ice into categories of playtime in a season. (ex. Little playtime, top minutes, etc.) | Qualitative | No |
| Avg\_Time\_On\_Ice | The average amount of time the player spent on the ice in the games he played (total time on ice divided by games played) | Quantitative | No |
| Blocks | Total shot blocks in a season by a player | Quantitative | No |
| Hits | Total hits in a season by a player | Quantitative | No |
| Faceoff\_Wins | Total faceoff wins in a season by a player | Quantitative | No |
| Faceoff\_Losses | Total faceoff losses in a season by a player | Quantitative | No |
| Faceoff\_Percentage | The number of faceoffs won by a player divided by total faceoffs | Quantitative | No |
| Hart\_Trophy? | Did a player win the hart trophy (voted most valuable player in a season) that season? | Qualitative | No |
| Votes | The amount of hart trophy votes that player received that year | Quantitative | No |
|  |  |  |  |
|  |  |  |  |

**8)Consider limitations and ethics**

-The limitations with this dataset are that the data only contains player data from 2004-2018. Goalies are excluded in this dataset which have been useful to compare their data with players.

-I believe the data and data collection methods are ethical since the data belongs to the NHL which is open to the public and no privacy concerns are being broken. The website Hockey-reference is allowed to collect NHL statistics which was utilized in this dataset. Professional athletes understand that their name and statistics are open to the public.

**10)Define questions to explore.** In a third section of your project document, define a list of questions to explore with your analysis. As mentioned in the Exercise, you may want to revisit [Exercise 1.2: Starting with Requirements](https://careerfoundry.com/en/course/data-immersion/exercise/starting-with-requirements) for a recap on writing good questions.

**10)Define questions to explore**

1)Is the Hart trophy voted on by which player simply earned the most points in a given season?

2)Is there a correlation between which players are most likely to score short-handed goals?

3)Does age or position impact which players have the highest/lowest time on ice?

4)Is the number of shots a player takes in a year directly correlated to a higher goal total?

5)Which seasons or teams have seen the highest goals, assists, points and the lowest in the dataset’s timeframe?

6) Is there a correlation between higher hit totals and penalty minutes?

**6.3**

7)Which teams/states have the highest and lowest point shares

8) Do the players with the highest point shares in a season result in winning the hart trophy?

**6.2**

9)How much do powerplay goals make up of the total points for top minute/lots of playtime players?

10)What are the main factors to causing a very high or low plus minus?

11)Does time impact shooting percentage? (to see if goalies have an impact?)

**Hypothesis**

1)Most players with the highest and lowest plus minuses will be defenseman.

2)Teams in the western conference will have higher hit and penalty minute totals than the eastern conference.

3)The teams with the most blocked shots will have the most shorthanded goals.

4)The majority of hart trophy votes will go to centermen.

**EXTRA:**

**Data Set Notes:**

-TOT in team name means a player got traded mid-season

- 2005 season is not included due to lockout

-2013 season is shortened due to lockout

-No goalies are included in the dataset, purely player statistics

-Link to all hockey statistic abbreviation meanings:

<https://www.hockey-reference.com/about/glossary.html#:~:text=opportunities%20against%20PS-,Point%20shares%3B%20an%20estimate%20of%20the%20number%20of%20points%20contributed,ties%20plus%20overtime%2Fshootout%20losses>.

**Column Abbreviation Changes/Meanings:**

* 'Rk':'Rank' / Players ranked by total points earned that season.
* 'Pos':'Position' / Position the player plays.
* 'Tm':'Team' / Team the player plays on that season.
* 'GP':'Games\_Played' / Total games played that season for a player.
* 'G':'Goals' / Total goals scored by a player that season.
* 'A':'Assists' / Total assists by a player that season.
* 'PTS':'Points' / Total points (goals and assists combined) by a player that season.
* 'plusminus':'Plus\_Minus / Calculated by subtracting the total number of goals allowed by a player's team while the player is on the ice from the total number of goals scored by the player's team while the player is on the ice.
* ‘PIM: Penalty\_ Minutes’ / Minutes of penalties a player has been assessed in that season
* ‘PS:Point\_Shares’ / An estimate of the number of points contributed by a player. [Calculating Points Shares](https://www.hockey-reference.com/about/point_shares.html) (better explanation)
* 'EV':'Even\_Strength\_Goals' / Total goals scored while neither your team or opposing team is on a powerplay.
* 'PP':'Powerplay\_Goals' / Goals scored by player when his team is on a powerplay.
* 'SH':'Shorthanded\_Goals' / Goals scored by a player when the opposing team is on a powerplay.
* 'GW':'Game\_Winning\_Goals' / Game winning goals scored by a player in a season.
* 'S':'Shots' / Total shots on goal in a season by a player.
* 'S\_percent':'Shooting\_Percentage' / Calculated by dividing a player’s goals by shots on goal.
* 'TOI':'Time\_On\_Ice' / Time (in minutes) spent in all games in a season.
* 'ATOI':'Avg\_Time\_On\_Ice' / The average amount of time the player spent on the ice in the games he played (total time on ice divided by games played)
* 'BLK':'Blocks'/ Total shot blocks in a season by a player.
* 'HIT':'Hits'/ Total hits in a season by a player.
* 'FOW':'Faceoff\_Wins' / Total faceoff wins in a season by a player.
* 'FOL':'Faceoff\_Losses' / Total faceoff losses in a season by a player.
* 'FO\_percent':'Faceoff\_Percentage' / The number of faceoffs won by a player divided by total faceoffs.
* 'HART':'Hart\_Trophy?' / Did a player win the hart trophy (voted most valuable player in a season) that season?

**Other Non-Changed Column Meanings:**

* Player / Player name
* Age / Age of player that season
* Votes / The amount of hart trophy votes that player received that year
* Season / Which NHL season year