Data Set Description

This data records many statistics from all regular season National Hockey League (NHL) games from the 2021 season. There are a total of 29 columns in my dataset, but key statistics I will be considering are each player's Age, Shots (S), Blocked Shots (BS), Time on Ice (TOI) and Average Time on Ice (ATOI).

Initial Plan for Exploration

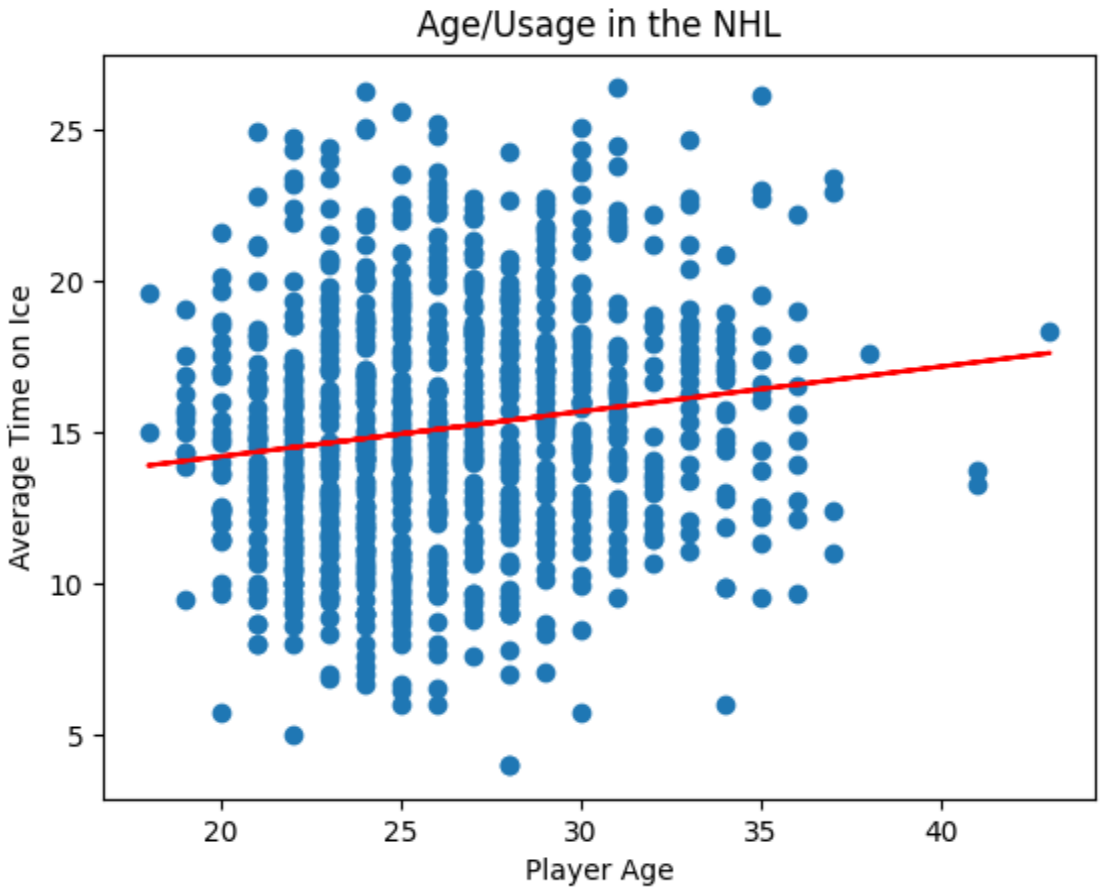
I will be using a combination of Python and CSV files downloaded from hockey-reference.com. I want to answer questions about hockey players and their game-by-game usage for their team. I further want to examine if there is a drop-off point between usage and productivity (i.e. Can players be given too much ice time?)

Data Cleaning Actions

I copied the table from https://www.hockey-reference.com/leagues/NHL_2021_skaters.html and found that due to in-season trades, some players show up as having multiple rows in the data set. Additionally, I know that from prior knowledge, there are two different players in the NHL named Sebastian Aho, so deleting duplicates correctly would be a challenge. I split the "Player" column (i.e. Vitaly Abramov\abramvi01) into two columns, delimited by the '\' character. This allowed me to access a unique identifier for each player, meaning I could use python to combine all of the "Michael Amadio\amadimi01" to account for his season totals with both the Kings and Senators, but NOT combine Hurricanes forward "Sebastian Aho\ahose01" with Islanders defenseman "Sebastian Aho\ahose02"

Key Findings and Insights

From the dataset, I found a positive correlation between an NHL player's age and the amount of time they spend on the ice in each game. This shows that coaches trust their more experienced players with more ice time than their younger players.



Other Hypotheses

My initial hypothesis regarding age versus ice time returned the positive correlation when examined on a linear regression model. I would further like to examine this data under polynomial regression to see if there is an age where average time on ice peaks and begins to decline, as could be assumed by seeing Joe Thornton, Patrick Marleau, and Zdeno Chara (the oldest 3 players, shown on the right) with ice times that are below the NHL average.

I would also like to examine the data when it is split by position, as NHL defensemen generally log more ice time than forwards. So while Zdeno Chara (oldest) looks to have a near-average ice time statistic, I have a hunch that it will be below average when only considering him with other defenders.

Finally, I would like to look at achievements of players (shots, goals, assists, blocked shots) compared to ice time and age. My hypothesis is that older players are used more, but younger players are more productive per minute on the ice.

Formal Significance Test

I realize that this significance test is a bit awkward since my data isn't a sample, but rather the entire population. I will thusly be taking a sample of 10 players and seeing if it estimates the true mean age. I used a random number generator to choose which players would comprise my sample.

Here is the description of my entire data set (913 players).

	Age	Average TOI
count	913.000000	913.000000
mean	26.260679	15.137980
std	4.095885	4.087153
min	18.000000	4.000000
25%	23.000000	12.129000
50%	26.000000	14.973000
75%	29.000000	17.918000
max	43.000000	26.375000

Here is the description of my sample set (10 players).

	Age	Average TOI
count	10.000000	10.000000
mean	29.500000	14.328300
std	4.648775	5.061403
min	21.000000	4.000000
25%	26.500000	12.482000
50%	30.500000	14.361500
75%	33.000000	18.250250
max	35.000000	20.667000

My sample mean was a whopping 29.5! When doing a one sided T-test to determine if this is significantly different from the population mean, we can see that, with a p-value of 2.17e-98, we can reject the null hypothesis and with great confidence say that these are different.

Input:

```
#Compute True mean
data3mean = data3["Age"].mean()

#Computer Sample Mean
data5mean = data5["Age"].mean()
print("Sample Mean: ", data5mean)

#One Sided T-Test to see if these are different:
tset, pVal = ttest_1samp(data3["Age"], data5mean)
print("P-Value: ", pVal)
```

Output:

```
Sample Mean:  29.5
P-Value:  2.1741943889216727e-98

Process finished with exit code 0
```

Suggestions for Next Steps

Next steps in the data analysis process, in addition to what has been outlined above, would be to take game-by-game ice time data and determine if there is an optimal amount of ice time to give each aged player to maximize their game impact.

Summary

This data was initially confusing to figure out due to the data cleaning process steps required (outlined above), but after cleaning up the issues, the data provided me with excellent insight into the ages and usages of NHL players. I would suggest that it would be very interesting to look at data from individual games rather than season averages and see if there are any major differences.