Capstone Project 2 Final Report

Predicting NHL player performance using non-game metrics with machine learning

Kevin Hill

Introduction

The National Hockey League (NHL) is often considered the smallest of the "Big 4" major North American sports, also consisting of the National League Football (NFL), Major League Baseball (MLB) and National Basketball Association (NBA) [1]. The NHL claims the lowest revenue and smallest fan base of those sports, but that has never deterred me from being a huge fan. Perhaps the smallest status has kept sports analytics from becoming a larger part of the NHL. On the other hand, it may be the speed and dynamic action that make it difficult to apply analytics to the sport. In any case, analytics is a relatively small aspect of NHL hockey, but is definitely becoming more popular and important.

For those who are not fans of the sport, I shall present a very brief summary of the sport. Hockey is played (in the NHL, on ice) by teams, who are allowed 5 skaters and 1 goaltender per team to be active at any given time. Each player uses a stick with a short bend at the end, shaped like the letter "L", and the object is to put the puck, a small rubber disk, into the opposing team's goal. The team scoring more goals when regulation time expires is the winner. A game is 60 minutes long, divided into 3 periods of 20 minutes each, with a short overtime if the teams are tied at the end of the game. When a player is penalized, that player is not allowed to participate for a short time, and that player's team is short a player as a result. This situation is called a power play for the team with the larger number of skaters, and a penalty kill for the penalized team. I enjoy the fast pace and fluid action of a hockey game and attempting to apply data science to the game is partly what inspired me to this capstone project. Another influence for this particular project is my brother-in-law, who played minor league hockey but was told by an NHL scout that he was 2 inches too short for consideration. The NHL has had a tendency to emphasize height for players, so I wanted to see if any evidence supported that belief.

The goal of my project is to predict the performance of an NHL player, specifically by working mainly with non-game metrics, such as height and weight. Non-game metrics are those things a player has little influence over, although height can change with age and weight with strength training, but most players tend to remain the same height and within a certain weight. A secondary goal is to determine which metrics have the most influence over a player's performance. I approached this project with the intention of determining how much attention should be paid to a player's non-game measurements. Thus, the primary clients I envisioned when planning this project include any NHL team, as well as other hockey teams outside the NHL, such as minor league teams linked to the NHL, for example the American Hockey League (AHL), and national hockey teams outside North America.

Data Wrangling

The data set I am working with was downloaded from the website NaturalStatTrick.com. The website has tabs for data organized by game, player and team, all collected from the 2007-2008 season through the 2017-2018 season. For this project, I am only interested in player data. In that tab, I used the data labelled “On-Ice”, “Individual” and “Bios”. Each of those collections of data contains different information relevant to studying player. The Bios data contains the information I’m immediately interested in, including physical characteristics such as height and weight, and background information that I plan to use as well, such as date of birth, nationality, and draft position. The Individual data contains data collected and attributed to an individual player, such as number of games played, points and shots. This data is traditionally important in predicting future player performance, although the focus of my study is on non-game metrics. Finally, the On-Ice data refers to events that happened while a player was on the ice, although not necessarily directly involved. This data is also important in determining a player’s performance, but requires more investigation, since the player may not have had an immediate influence on the event.

Much of my data wrangling involved removing features I was less interested in for this project. For instance, many data collected for hockey analytics involves an event for and an event against, as well as the ratio of the two. For this study, I decided I was only interested in the ratios, not the base numbers, so I only kept the ratios of that data. Also, I wanted all the data collected into a single data set, rather than spread across three separate sets, so I removed duplicate information that was present in all three data sets, such as a player’s name, team and position. One thing that is interesting to measure for player’s is what percent of their play begins in different “zones” of the ice, offensive, neutral or defensive. This data contained only offensive and defensive zone data, so I also calculated the neutral zone data.

My original plan was to build a baseline model using regularized linear regression with hyper-parameter tuning. Linear regression attempts to construct an equation of sorts, in which each input variable is assigned a weight, or importance, toward the output. In this study, the output is a player’s performance metric. The metric I was initially planning on predicting was points, because that is one of the primary measurements of the talent of a player. I had read an article suggesting there was no connection between a player’s height and their point production [2], so part of the goal of this project was to verify that finding. However, a player’s points is not the only metric teams are interested in, so I was also planning on predicting another useful statistic, puck possession, also known as Corsi. The data is also available through various websites, including NHL.com and Corsica.hockey.

Exploratory Data Analysis

The first thing I wanted to look at was how player height and weight have changed over time. I have been under the impression that the NHL valued player height, sometimes to very high degree, so I wondered if that were true and if that was any different now versus in previous seasons. Figure 1 shows the mean player height over the past 10 seasons and figure 2 shows the mean player weight over the same timeframe.

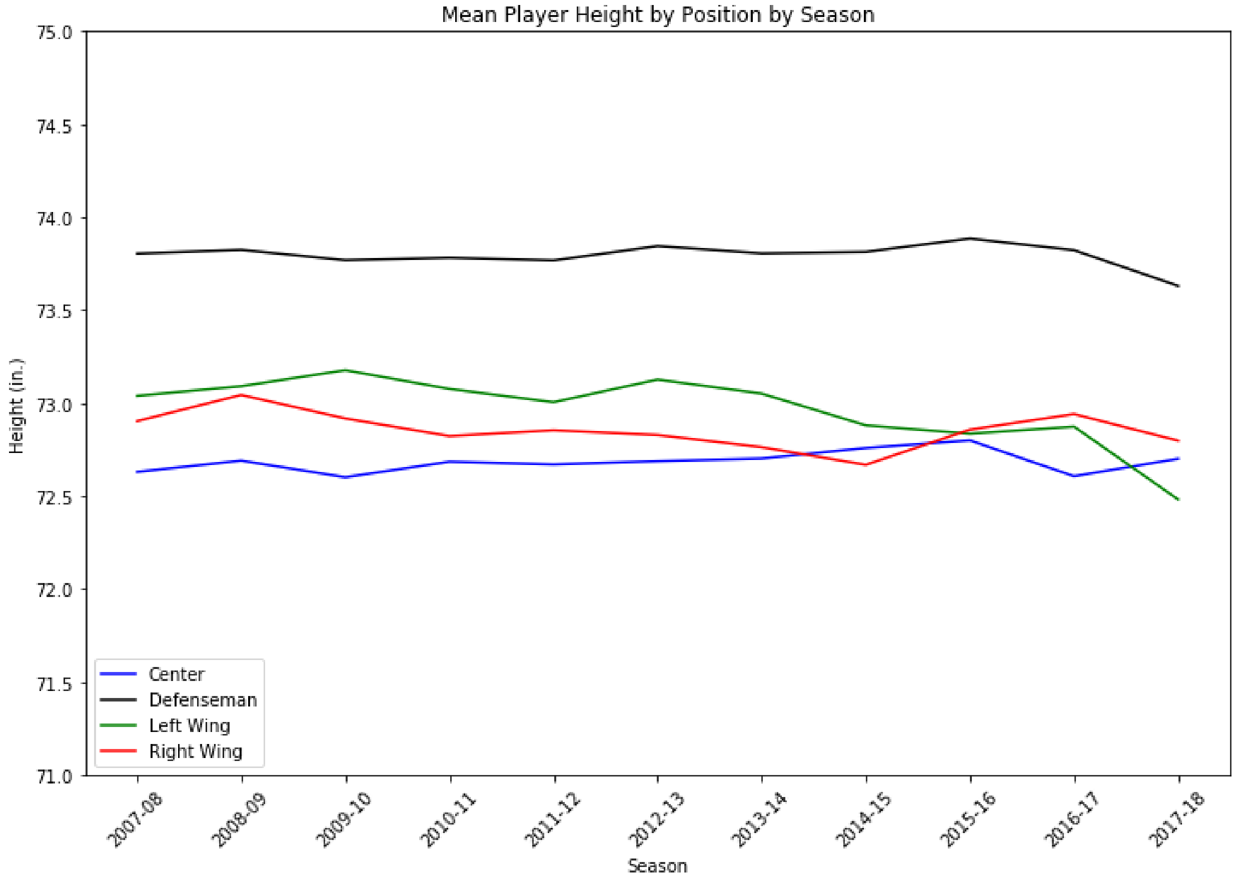


Figure 1. Mean Player Height over time

One thing to notice in both figures is the scale. Both figures may look more dramatic than they really are, but at zoomed out scale, everything would appear as a mostly flat line, so this at least shows what little change there has been. Figure 1 shows defensemen have a height advantage over forwards of about one inch. It also shows the mean height of all players haven't really changed much, except for left wings who have gotten roughly a half inch shorter. Figure 2 shows defensemen tend to weigh the most, and centers tend to weigh the least, with wings in between the two. It also shows most players have lost roughly 5 pounds over the last 10 seasons, but centers have generally maintained the mean weight over that time.

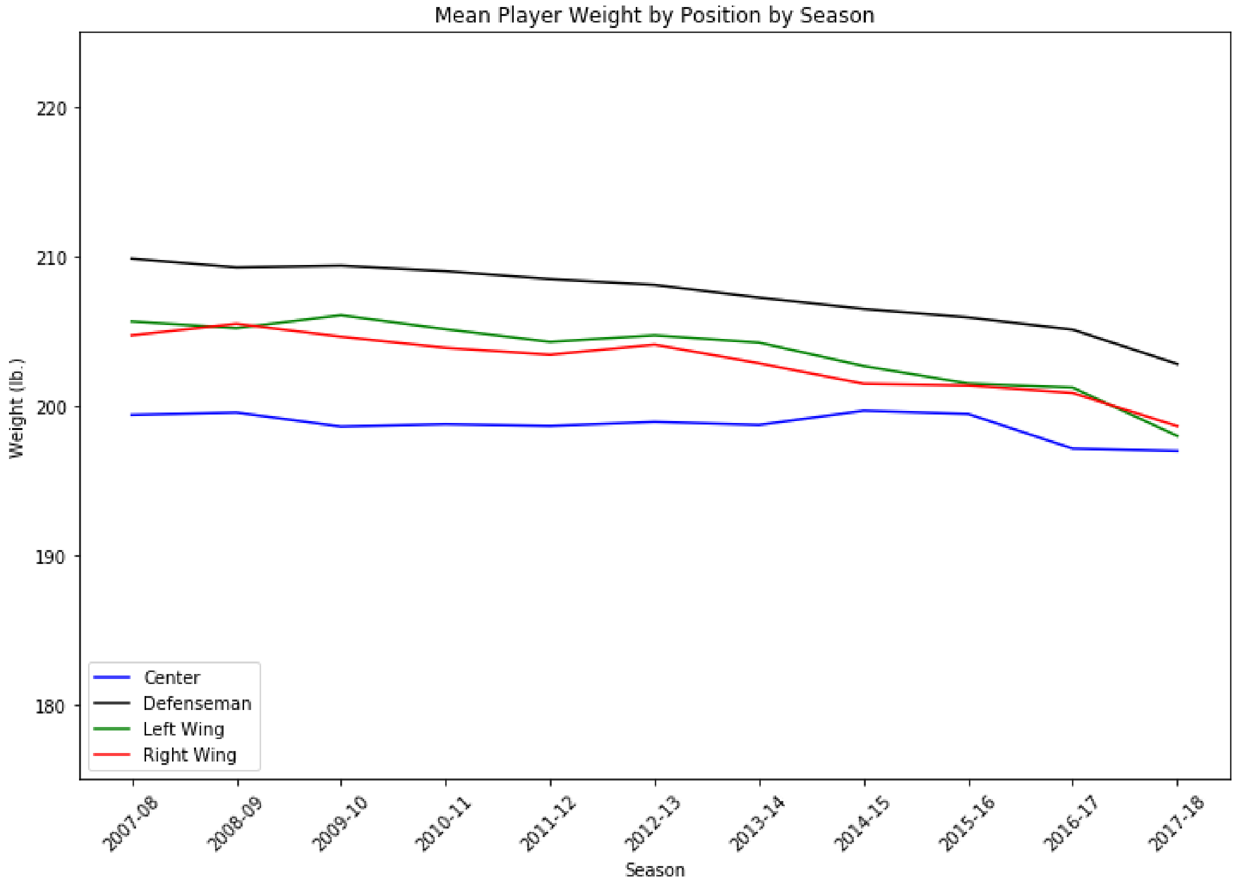


Figure 2. Mean Player Weight over time

Between both figures, it is clear that defensemen tend to be taller and heavier than forwards, and that hasn't really changed over time, but otherwise there are no real differences between players now and 10 seasons ago. I also plotted the mean height and weight of players by draft round but found nothing interesting there. The figures can be found in Appendix A of this report.

Another possible category to use as a non-game metric is a player's nationality. I was curious if perhaps one country's hockey program encourages their players to excel in different areas. Since my primary goal involves height and weight, I created a graph to show those measurements for the most common nationalities among NHL players, shown in figure 3. All players not from one of the major NHL origin countries are aggregated in the "Other" dot in the middle of the Canadian dot. Again, note the close scale of the graph. The size of the dots in the graph represent the number of players from that country. The most interesting dot is Slovakia, showing their players with a mean height around 74 inches and a mean weight around 210 pounds, compared to the mean Canadian player's height of just over 73 inches and weight around 204 pounds. There are two players from Slovakia who are at least 80 inches tall, so to see how much those two affected the means, I made a second graph of the data without them.

That pulled the Slovakian dot closer to the rest, but it still appeared like an outlier. That figure is in Appendix B.

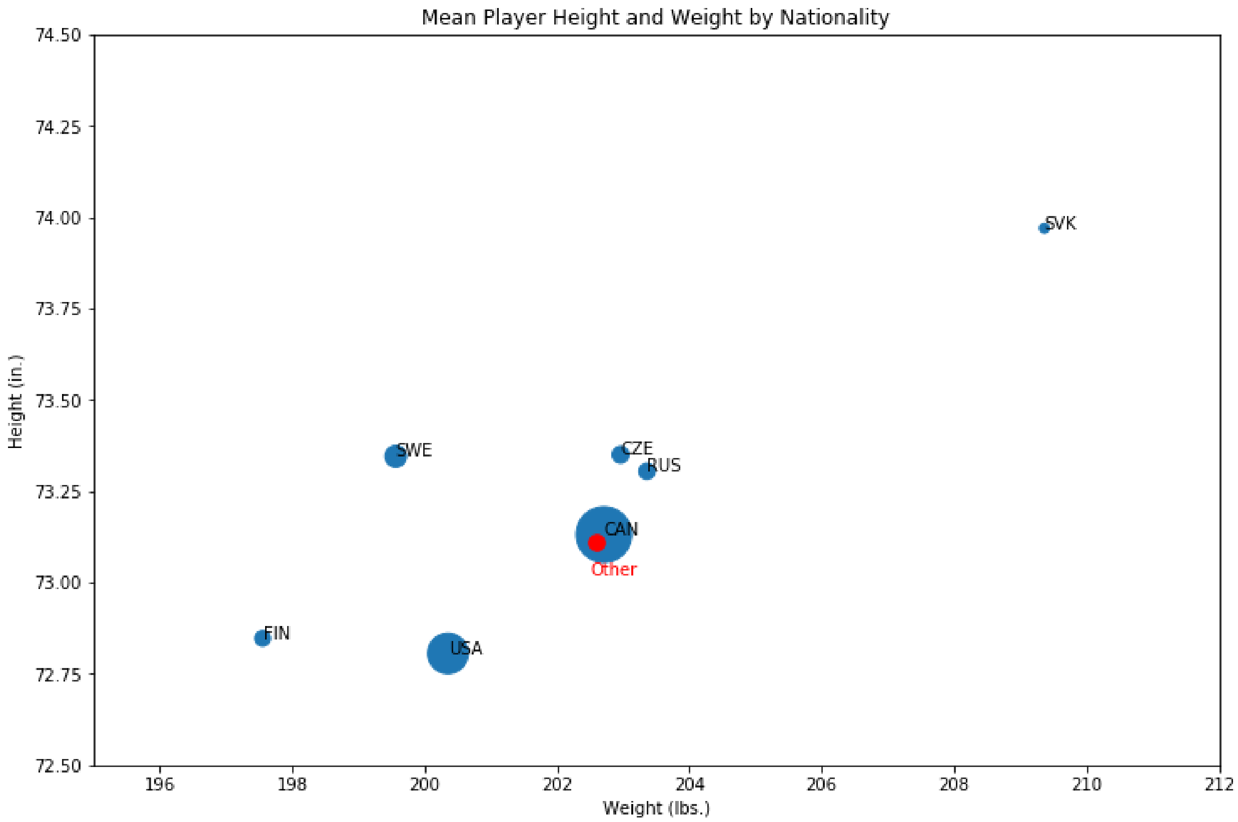


Figure 3. Height and Weight by Nationality

I was also curious about the difference in player count by country. Historically, and confirmed by the size of the dots in figure 3, the NHL has been dominated by players from Canada and, to a lesser extent, the USA. I created a graph of the total player counts which clearly shows that, which is available in Appendix C. It also shows a decline in the number of Canadian players over time and an increase in the number of American players, but players from other countries are difficult to interpret. I created a second graph without North American players in figure 4. That figure shows a tremendous growth in Swedish players, and a slight decrease in central European players, specifically Czech and Slovakian. Once again, all players not from one of the major NHL origin countries are aggregated in the “Other” line, which does appear to grow slightly over time.

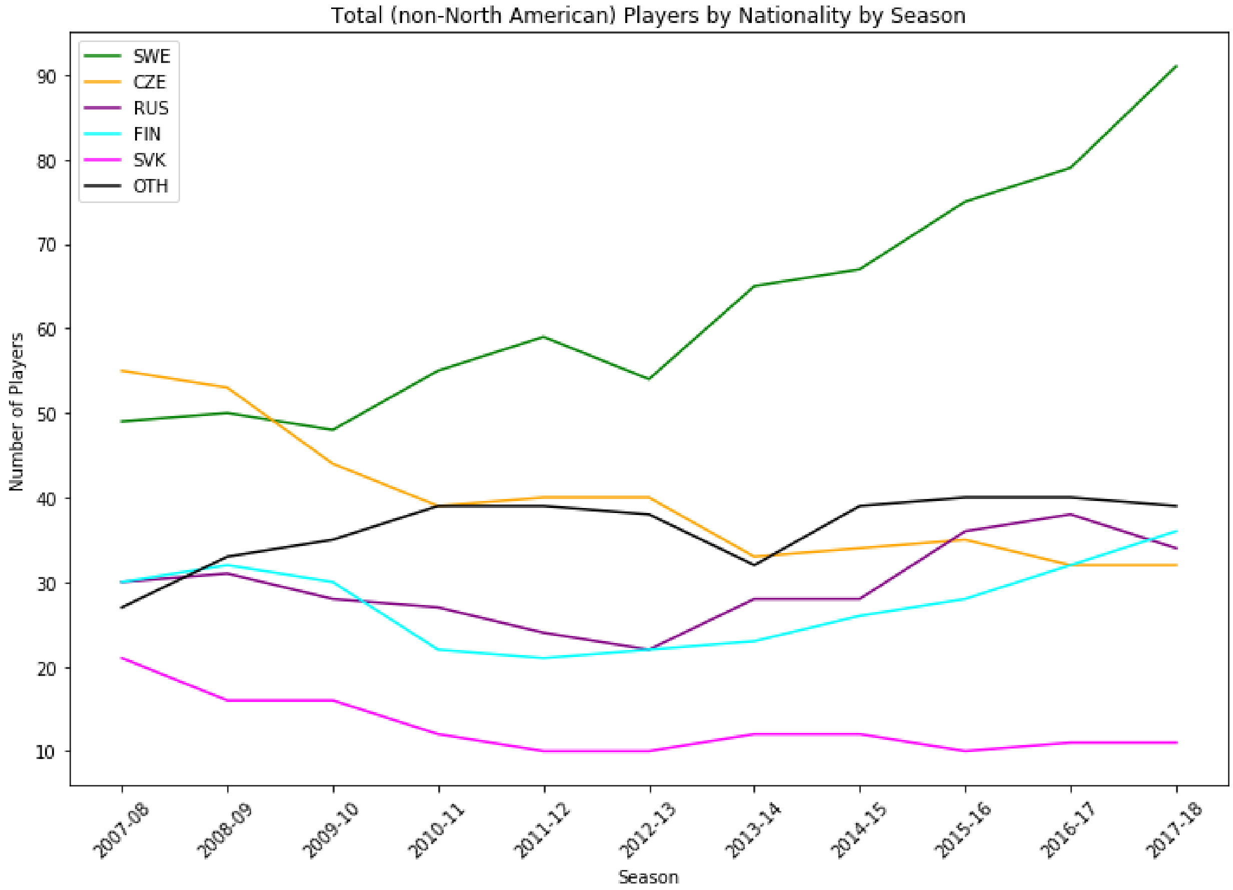


Figure 4. Non-North American player counts

My next step was to create plots showing some of the primary metrics to measure player performance by, based on height and weight and colored by country of origin, to again look for trends. That proved to be too messy to interpret, so I instead created the same plots but without the coloring and by country instead. There are too many graphs to show in the body of this report, so they are all available in Appendix D. However, I did spot a few interesting patterns, which I've shown in figure 5. As shown in figure 3, players from the USA tend to be shorter than players from all other countries, so I found it interesting to see the high number of blocked shots among taller American players. In the next section, I will find some estimated correlation coefficients, which is why the graph of Russian players Is unusual. The coefficients will show a negative relation between height and rush attempts, so it seems strange to see the spike in rushes by taller Russians. The Swedish player graph shows nothing by itself, but compared to players from other countries, lighter Swedes seem to giveaway the puck less. And finally, somewhat opposite the US player graph, Czech players tend to be heavier than players from other countries, so again I found it surprising to see the high number of goals scored by lighter Czech players.

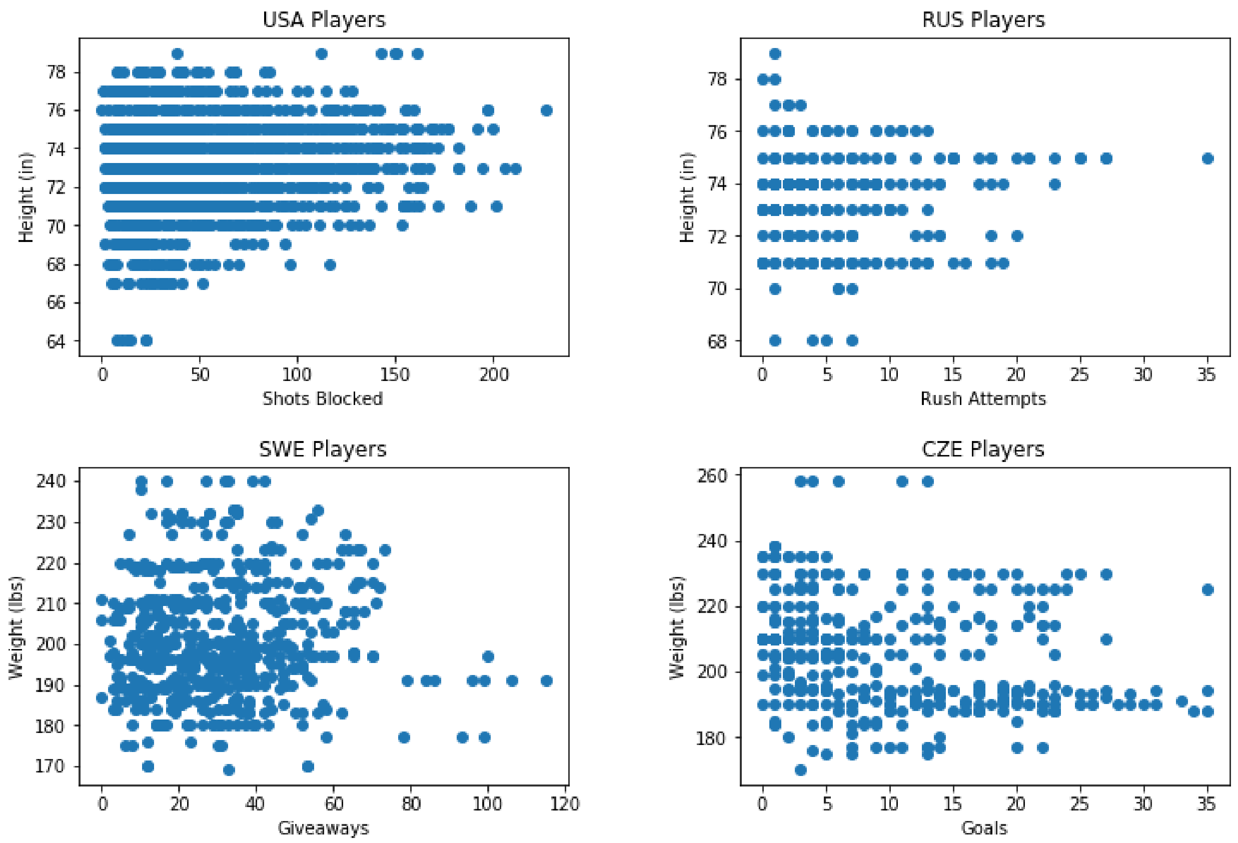


Figure 5. Sample scatter plots

My final exploratory step was to look at the affect of a player's age on their performance, creating some very interesting graphs in the process. Since the first metric that seems to be important is points, figure 6 shows the mean player points by age. Some things I found interesting in this graph are the initial spike in both points and goals when a player is very young. I wonder if this is due to the extreme talent of players who start playing at that age. Otherwise, players seem to peak around age 29 and decline after that. The other points I noticed were player points at 34 and 37, which I would guess are due to survivor bias, again due to the talent of players who last that long in the NHL. Since again, points aren't the only metric NHL teams are interested in, I also created a graph of puck possession, or Corsi, shown in figure 7. In this graph, the dotted line represents the point at which a player's team puck possession or team goal scoring is even, neither greater than or less than the opposing team's. The most interesting thing I saw in this graph was the higher percent of both puck possession and goal scoring when a player is under 25, which appears to drop under 50% until a player reaches their mid to late 30's. The final age graph I created was due to the estimated correlation coefficients I'll show in the next section again. The statistic that seems to have the highest correlation to non-game metrics ends up being hits, so I created that graph as well in figure 8. According to that graph, players tend to make the most hits between the ages of 25 and 30. However, their shot blocking appears to steadily increase until around 34, after which it drops off quickly.

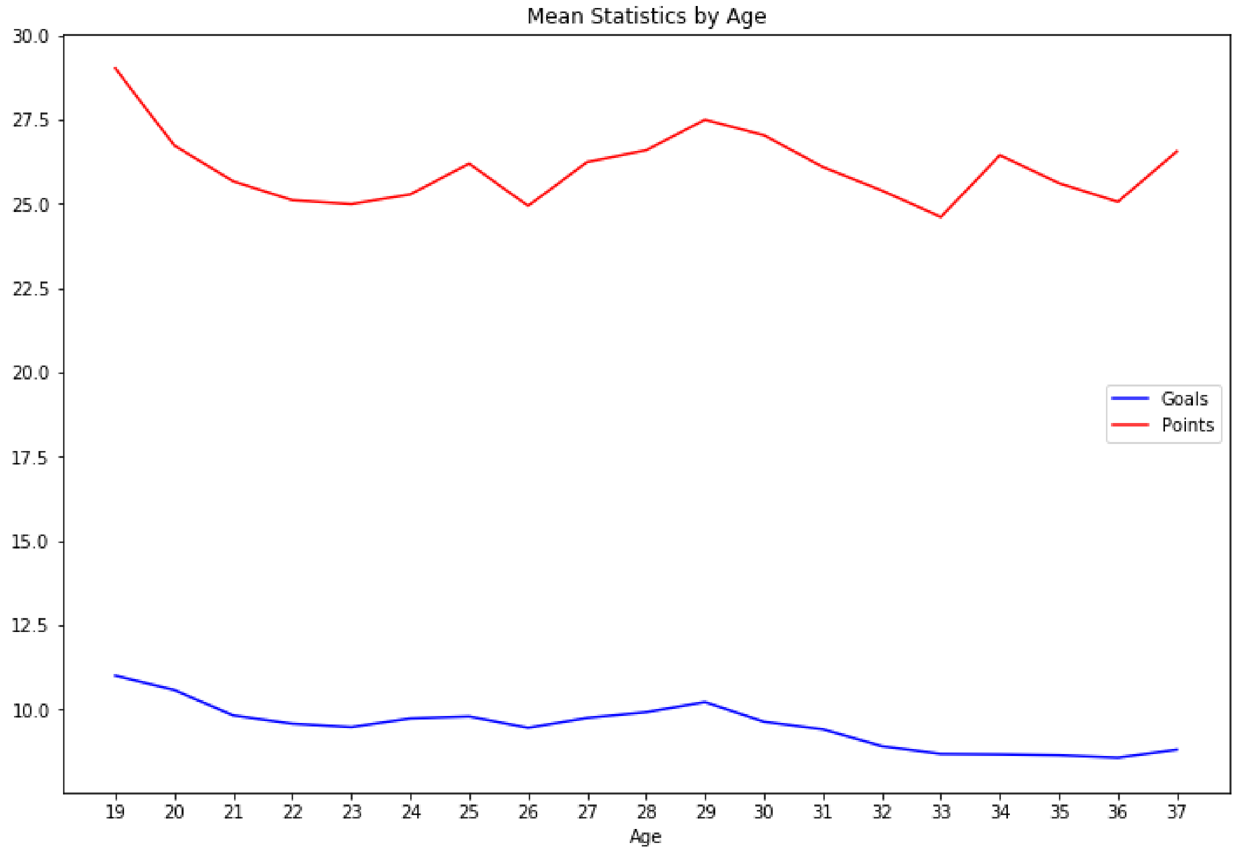


Figure 6. Points and Goals by Age

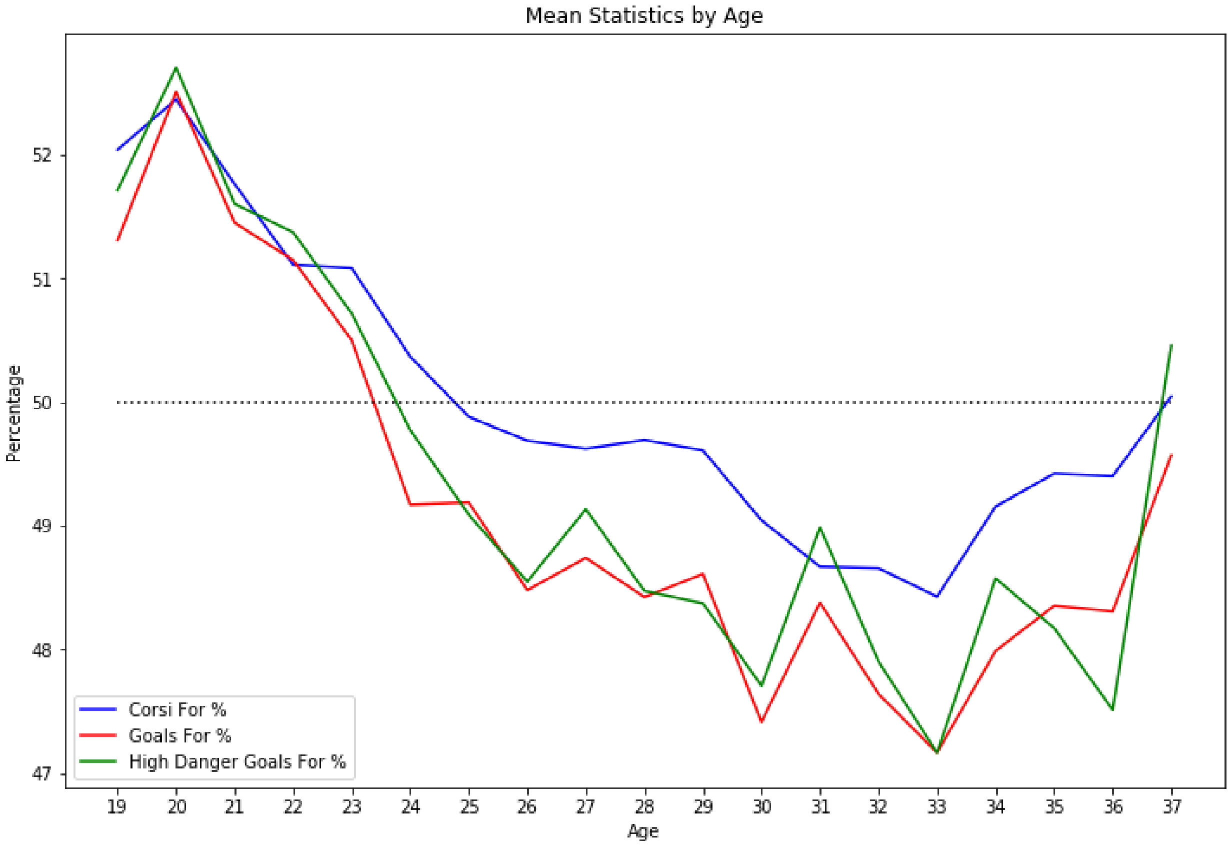


Figure 7. Corsi % and Goal % by Age

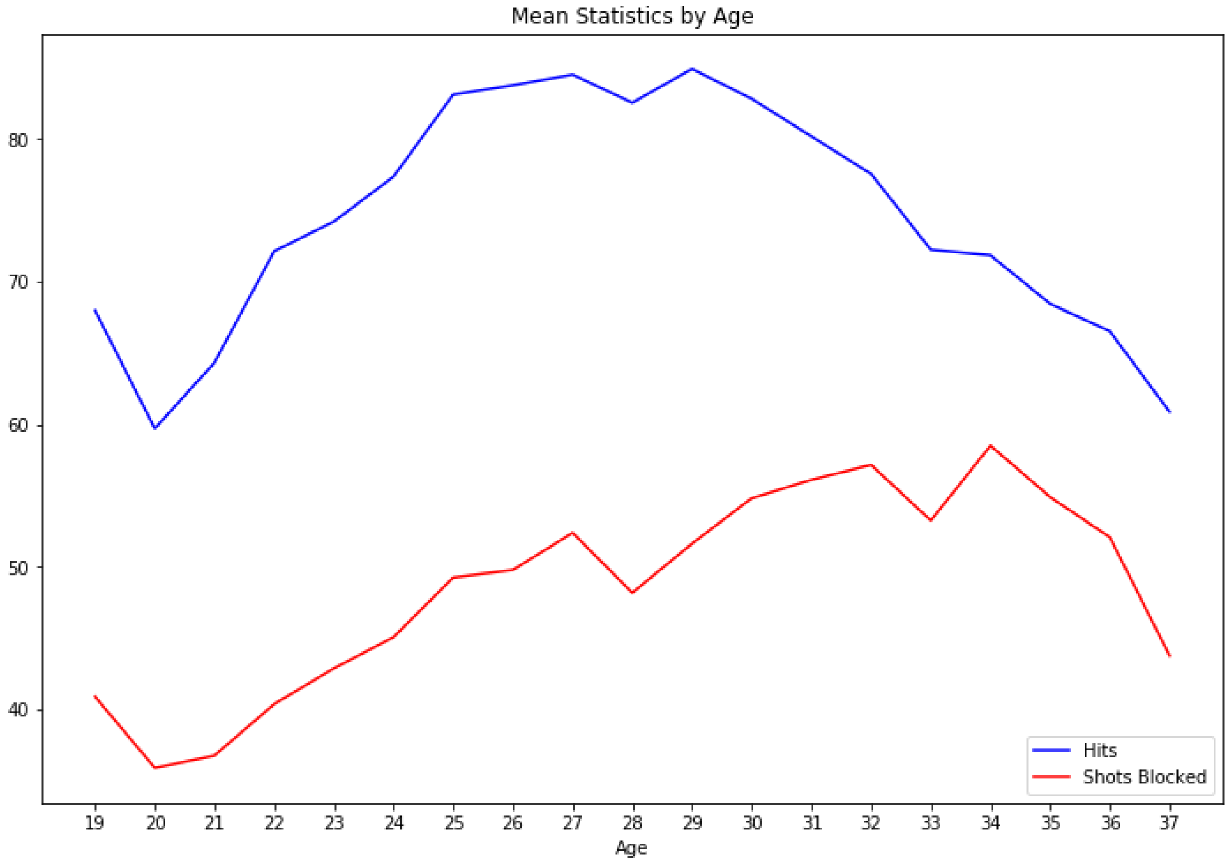
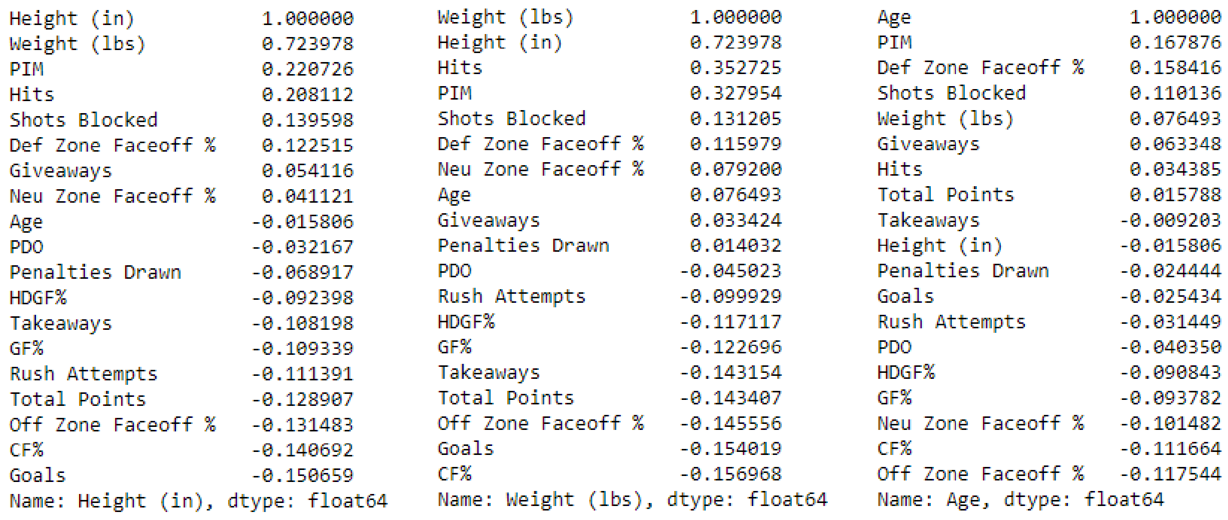


Figure 8. Hits and Shots Blocked by Age

Inferential Statistics

The goal of this project was to predict player performance based on non-game metrics, so the first thing I wanted to know was how much impact those metrics might have on measured statistics, so I ran some correlation tests for height, weight and age. I didn't want the outcomes skewed by players who didn't play very much, so I arbitrarily chose to calculate the correlations with only players who played more than 20 games in a given season, out of the total 82. I may revisit this in the future and calculate the correlation without eliminating that data, but I wanted only the regularly contributing players to affect the numbers right now. Figure 9 shows the estimated correlation coefficients I found.



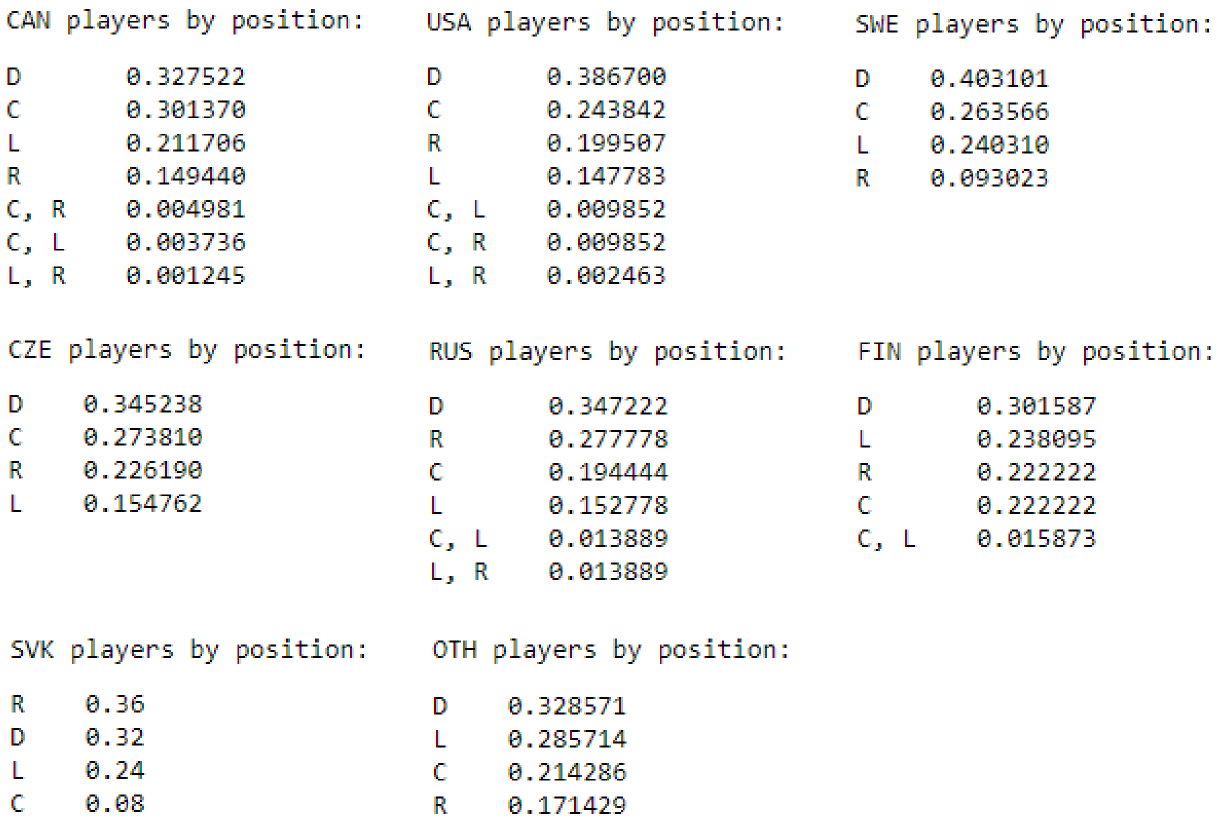
```
Height (in)             1.000000    Weight (lbs)            1.000000    Age                     1.000000
Weight (lbs)            0.723978    Height (in)             0.723978    PIM                     0.167876
PIM                     0.220726    Hits                    0.352725    Def Zone Faceoff %      0.158416
Hits                    0.208112    PIM                     0.327954    Shots Blocked           0.110136
Shots Blocked           0.139598    Shots Blocked           0.131205    Weight (lbs)            0.076493
Def Zone Faceoff %      0.122515    Def Zone Faceoff %      0.115979    Giveaways               0.063348
Giveaways               0.054116    Neu Zone Faceoff %      0.079200    Hits                    0.034385
Neu Zone Faceoff %      0.041121    Age                     0.076493    Total Points            0.015788
Age                    -0.015806    Giveaways               0.033424    Takeaways              -0.009203
PDO                    -0.032167    Penalties Drawn         0.014032    Height (in)            -0.015806
Penalties Drawn        -0.068917    PDO                    -0.045023    Penalties Drawn        -0.024444
HDGF%                  -0.092398    Rush Attempts          -0.099929    Goals                  -0.025434
Takeaways              -0.108198    HDGF%                  -0.117117    Rush Attempts          -0.031449
GF%                    -0.109339    GF%                    -0.122696    PDO                    -0.040350
Rush Attempts          -0.111391    Takeaways              -0.143154    HDGF%                  -0.090843
Total Points           -0.128907    Total Points           -0.143407    GF%                    -0.093782
Off Zone Faceoff %     -0.131483    Off Zone Faceoff %     -0.145556    Neu Zone Faceoff %     -0.101482
CF%                    -0.140692    Goals                  -0.154019    CF%                    -0.111664
Goals                  -0.150659    CF%                    -0.156968    Off Zone Faceoff %     -0.117544
Name: Height (in), dtype: float64   Name: Weight (lbs), dtype: float64  Name: Age, dtype: float64
```

Figure 9. Estimated correlation coefficients

The first thing that stands out are the top two positively correlated stats for height and weight being penalty minutes and hits, while shots blocked and defensive zone starts also show some positive correlation. Similarly, the most negatively correlated stats for height and weight are goals scored, corsi percentage, offensive zone starts and total points. The estimated coefficients for age are not as pronounced as the other two metrics, but penalty minutes and defensive zone starts appear the most positively correlated, with corsi percentage and offensive and neutral zone starts showing as the most negatively correlated.

Additionally, to see any differences between players of different nationalities, I calculated the correlation coefficients by player nationality, all of which are in Appendix E. Since the bulk of NHL players are Canadian, their correlations are very similar to the total numbers, except the most negative correlations tend to be slightly exaggerated for heavier Canadians. US players appear to have more positive correlations to penalty minutes and hits, and age seems to make their correlations both more positive and more negative. Swedish players tend to have less positive correlation and more negative correlation, other than hits, which may be more common for heavier Swedes, and older Swedish players appear to have very little negative correlations. In addition to the other statistics, Czech players have positive correlations in both defensive zone starts and shot blocking, but their negative correlations are more so and include scoring on the rush and age may affect their puck control. I'll explore their potential defensive tendencies shortly. Taller Russians do not show much positive correlation to anything, but also little negative correlation except puck possession, while heavier Russians appear to block more shots, and age seems to give Russians more positive correlations to point production and takeaways, but also giveaways and penalty minutes. Finns show the most positive correlation to shot blocking only, and oddly the most negative correlation to drawing penalties, and older Finns appear to mainly commit more penalties. The estimated coefficients for Slovakians show much more variety, but the sample size of those players is quite small, so I won't detail those.

The higher correlations to defensive statistics for Czech players got me curious about the pool of players from there. I thought perhaps more Czech players play defense, so I calculated the number of players by position from each country, but found the percent of Czech defensemen to be in the bottom third, so the correlations aren't due to just position, and this project wasn't concerned with following up that research. Figure 10 shows the breakdown of position by country.



```
CAN players by position:

D       0.327522
C       0.301370
L       0.211706
R       0.149440
C, R    0.004981
C, L    0.003736
L, R    0.001245
```

```
USA players by position:

D       0.386700
C       0.243842
R       0.199507
L       0.147783
C, L    0.009852
C, R    0.009852
L, R    0.002463
```

```
SWE players by position:

D    0.403101
C    0.263566
L    0.240310
R    0.093023
```

```
CZE players by position:

D    0.345238
C    0.273810
R    0.226190
L    0.154762
```

```
RUS players by position:

D       0.347222
R       0.277778
C       0.194444
L       0.152778
C, L    0.013889
L, R    0.013889
```

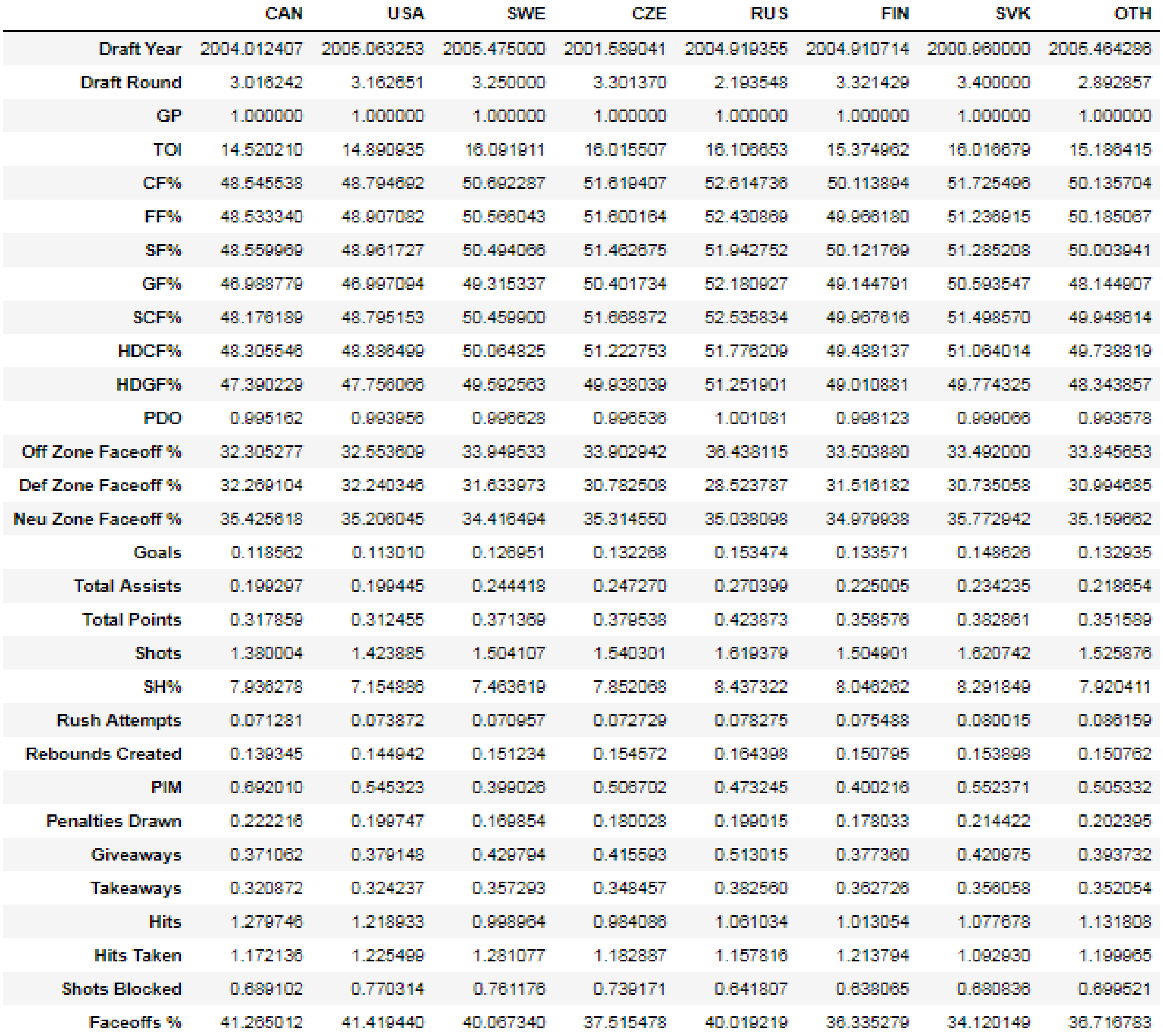
```
FIN players by position:

D       0.301587
L       0.238095
R       0.222222
C       0.222222
C, L    0.015873
```

```
SVK players by position:

R    0.36
D    0.32
L    0.24
C    0.08
```

```
OTH players by position:

D    0.328571
L    0.285714
C    0.214286
R    0.171429
```

Figure 10. Player positions by country

Lastly, I wanted to discover if there were any patterns to any measured statistics when the players are broken down by country. First, I "normalized" the stats to be per game and computed the mean stats of the players by country. Figure 11 shows those stats and what stood out to me was the higher mean stats for Russian players in the majority of the categories. That prompted me to do some hypothesis testing, focusing mainly on the popular stats of points and corsi percent. The first null hypothesis I tested was, the difference between the mean total points for Russian players and the mean total points for non-Russian players is zero, with the alternate hypothesis being the difference between mean total points is more or less than zero. I ran a SciPy 2-tailed t-test and found a p-value of 0.004454, which is less than 0.01, so that null hypothesis can be rejected. The second null hypothesis I tested was the difference between the mean Corsi % for Russian players and non-Russian players is zero, with the alternate hypothesis being the difference is not zero. Running the same SciPy test, I found a p-value of 0.000001, which is much less than 0.01, so that null hypothesis can also be rejected. Not being satisfied with just those two tests, I tested two more null hypotheses. A null hypothesis test of the difference between the mean goals of Russian and non-Russian players produced a p-value of 0.035507, which is slightly indeterminate, being less than 0.05 but greater than 0.01, so that null hypothesis cannot necessarily be rejected, and perhaps needs



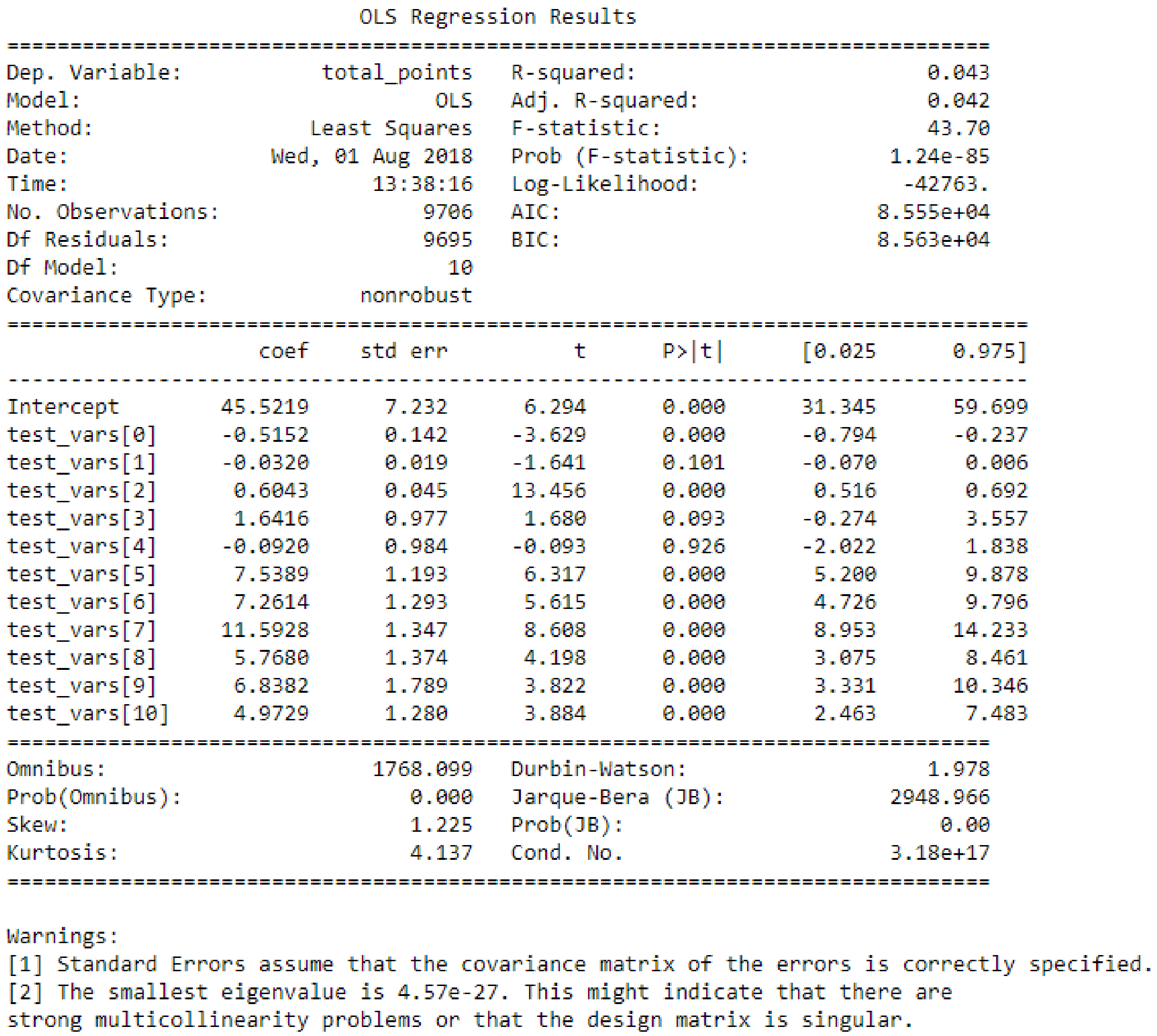
	CAN	USA	SWE	CZE	RUS	FIN	SVK	OTH
Draft Year	2004.012407	2005.063253	2005.475000	2001.589041	2004.919355	2004.910714	2000.960000	2005.464286
Draft Round	3.016242	3.162651	3.250000	3.301370	2.193548	3.321429	3.400000	2.892857
GP	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
TOI	14.520210	14.890935	16.091911	16.015507	16.106653	15.374962	16.016679	15.186415
CF%	48.545538	48.794692	50.692287	51.619407	52.614736	50.113894	51.725496	50.135704
FF%	48.533340	48.907082	50.566043	51.600164	52.430869	49.966180	51.236915	50.185067
SF%	48.559969	48.961727	50.494066	51.462675	51.942752	50.121769	51.285208	50.003941
GF%	46.988779	46.997094	49.315337	50.401734	52.180927	49.144791	50.593547	48.144907
SCF%	48.176189	48.795153	50.459900	51.668872	52.535834	49.967616	51.498570	49.948614
HDCF%	48.305546	48.886499	50.064825	51.222753	51.776209	49.488137	51.064014	49.738819
HDGF%	47.390229	47.756066	49.592563	49.938039	51.251901	49.010881	49.774325	48.343857
PDO	0.995162	0.993956	0.996628	0.996536	1.001081	0.998123	0.999066	0.993578
Off Zone Faceoff %	32.305277	32.553609	33.949533	33.902942	36.438115	33.503880	33.492000	33.845653
Def Zone Faceoff %	32.269104	32.240346	31.633973	30.782508	28.523787	31.516182	30.735058	30.994685
Neu Zone Faceoff %	35.425618	35.206045	34.416494	35.314550	35.038098	34.979938	35.772942	35.159662
Goals	0.118562	0.113010	0.126951	0.132268	0.153474	0.133571	0.148626	0.132935
Total Assists	0.199297	0.199445	0.244418	0.247270	0.270399	0.225005	0.234235	0.218654
Total Points	0.317859	0.312455	0.371369	0.379538	0.423873	0.358576	0.382861	0.351589
Shots	1.380004	1.423885	1.504107	1.540301	1.619379	1.504901	1.620742	1.525876
SH%	7.936278	7.154886	7.463619	7.852068	8.437322	8.046262	8.291849	7.920411
Rush Attempts	0.071281	0.073872	0.070957	0.072729	0.078275	0.075488	0.080015	0.086159
Rebounds Created	0.139345	0.144942	0.151234	0.154572	0.164398	0.150795	0.153898	0.150762
PIM	0.692010	0.545323	0.399026	0.506702	0.473245	0.400216	0.552371	0.505332
Penalties Drawn	0.222216	0.199747	0.169854	0.180028	0.199015	0.178033	0.214422	0.202395
Giveaways	0.371062	0.379148	0.429794	0.415593	0.513015	0.377360	0.420975	0.393732
Takeaways	0.320872	0.324237	0.357293	0.348457	0.382560	0.362726	0.356058	0.352054
Hits	1.279746	1.218933	0.998964	0.984086	1.061034	1.013054	1.077678	1.131808
Hits Taken	1.172136	1.225499	1.281077	1.182887	1.157816	1.213794	1.092930	1.199965
Shots Blocked	0.689102	0.770314	0.761176	0.739171	0.641807	0.638065	0.680836	0.699521
Faceoffs %	41.265012	41.419440	40.067340	37.515478	40.019219	36.335279	34.120149	36.716783

Figure 11. Stats per game by player country

more study. Finally, the null hypothesis of the difference between the mean shooting percentage of Russian and non-Russian players found a p-value of 0.169475, which is greater than 0.05, so that null hypothesis cannot be rejected. In the end, there may be statistical evidence to support Russian players scoring more goals and contributing more to puck possession than non-Russian players.

Baseline and Advanced Model Construction

I began my baseline model by building a linear regressor using StatsModels, because I have read that that is better for initial testing, and SciKitLearn's model is better for optimizing. I first fit the model to find points from just player height, since that was the intent of this project. Fitting the model refers to attempting to create the linear equation to find the desired output from the inputs, in this case the output is points and the input is only player height. The summary showed an R-squared score of 0.004, which shows the model had extremely poor accuracy. The R-squared score is a measure of how well the equation finds the correct output and ranges from zero to one, where zero never finds the correct output and one suggests the model finds the correct output exactly every time. Next, I fit the model from just player weight and found an R-squared score of 0.002, which was even worse. Moving on, I fit the model from just player age with the R-squared score climbing to 0.018, which means the model was still not helpful. Then, I fit the model from player nationality and got an R-squared score of 0.02, so clearly each individual metric was not scoring well. Finally, I fit the model to all the previous metrics and found the best R-squared score so far, at 0.043, but even that was so low it meant the model was still performing very poorly. The first four summaries are in Appendix F and the last one is shown here in figure 12.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:           total_points   R-squared:                       0.043
Model:                            OLS   Adj. R-squared:                  0.042
Method:                 Least Squares   F-statistic:                     43.70
Date:                Wed, 01 Aug 2018   Prob (F-statistic):           1.24e-85
Time:                        13:38:16   Log-Likelihood:                -42763.
No. Observations:                9706   AIC:                         8.555e+04
Df Residuals:                    9695   BIC:                         8.563e+04
Df Model:                          10
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept        45.5219      7.232      6.294      0.000      31.345      59.699
test_vars[0]     -0.5152      0.142     -3.629      0.000      -0.794      -0.237
test_vars[1]     -0.0320      0.019     -1.641      0.101      -0.070       0.006
test_vars[2]      0.6043      0.045     13.456      0.000       0.516       0.692
test_vars[3]      1.6416      0.977      1.680      0.093      -0.274       3.557
test_vars[4]     -0.0920      0.984     -0.093      0.926      -2.022       1.838
test_vars[5]      7.5389      1.193      6.317      0.000       5.200       9.878
test_vars[6]      7.2614      1.293      5.615      0.000       4.726       9.796
test_vars[7]     11.5928      1.347      8.608      0.000       8.953      14.233
test_vars[8]      5.7680      1.374      4.198      0.000       3.075       8.461
test_vars[9]      6.8382      1.789      3.822      0.000       3.331      10.346
test_vars[10]     4.9729      1.280      3.884      0.000       2.463       7.483
==============================================================================
Omnibus:                     1768.099   Durbin-Watson:                   1.978
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2948.966
Skew:                           1.225   Prob(JB):                         0.00
Kurtosis:                       4.137   Cond. No.                     3.18e+17
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 4.57e-27. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Figure 12. Summary of linear regression results using only non-game metrics

A different method for testing the accuracy of a model is to plot what are called the residuals, which are the difference between the predicted outputs of the model and the actual statistics. The plot of the residuals is shown in figure 13. The obvious thing to notice is the large proportion of data in the negative. This suggests the model tends to mostly predict points that are less than the actual value, with the peak value near -15.

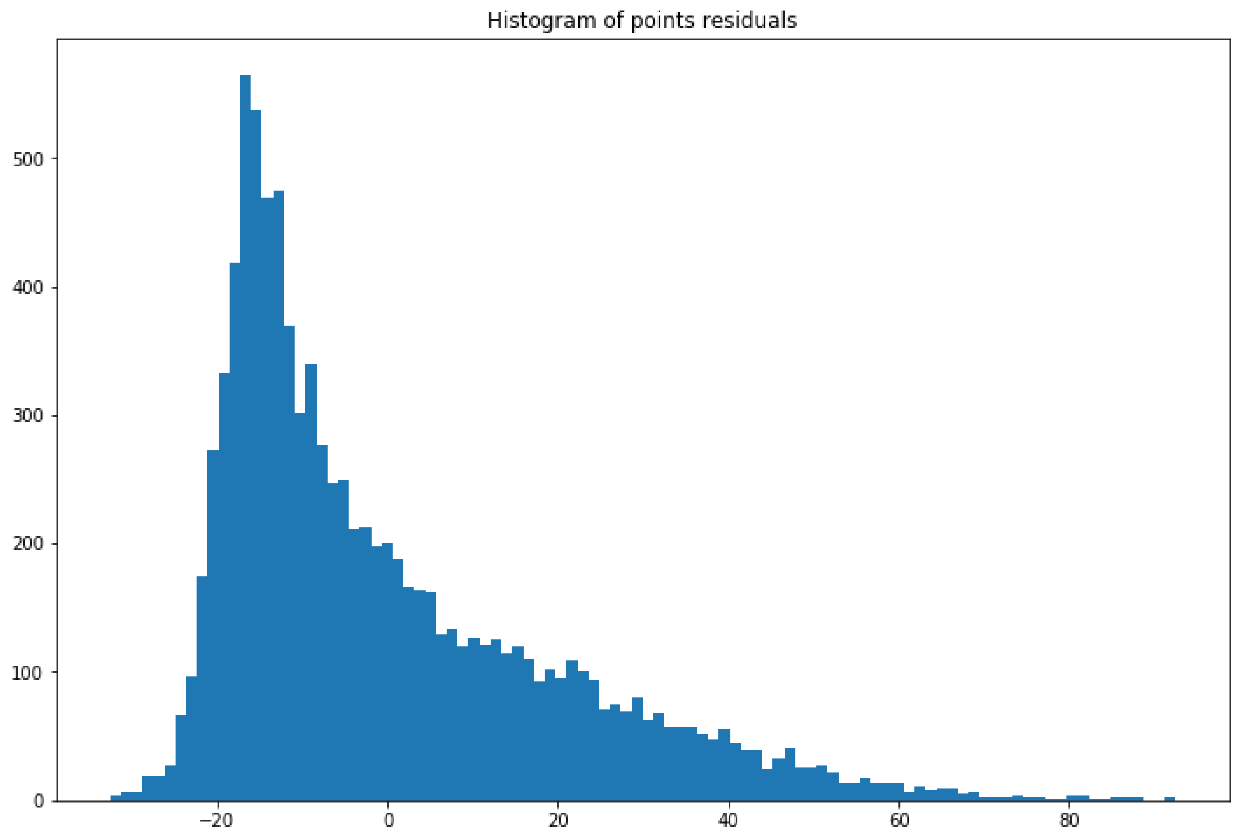


Figure 13. Plot of residuals from points linear regressor

In an attempt to improve those outcomes, I created a linear regressor using SciKitLearn as well. To fit this model, I created what is known as a train-test-split of the data, using the default setting of training at 75%. This method divides, or splits, the data into two different sets, one composed of 75% of the data which is used to train the model, and the other composed of the remaining 25% which is used to test the accuracy of the model on data it hasn't been exposed to. Part of training the model is called hyperparameter tuning, which involves adjusting several parameters of the model to try to maximize the accuracy. I calculated the best accuracy using L1 regularization, also known as Lasso, and L2 regularization, also known as Ridge, but didn't find any real improvement, suggesting the original model was not overfitting much. Overfitting happens when the model learns to predict well with the training data it knows, but doesn't predict well with unknown data.

Since linear regression seemed extremely inaccurate, I decided to try a different model, a random forest regressor, setting values for square root, log, and no maximum of features, as well as number of trees, for hyperparameter tuning. Random forest regressors begin with what is known as a decision tree. The tree is created by that model examining each variable and branching at different values toward the output, in this case, a player's points. A completely optimized tree yields the same results every time, so tends to work extremely well during training, but relatively poorly on unknown data. To compensate, decisions can be given in random order, forcing the tree to reach different outputs. The forest refers to the creation and training of many of these decision trees and taking the most common output as the predicted one. This produced better scores than the linear regressors but were again still low, with a best R-squared score of 0.134112. The method I used in computing the R-squared score for random forests is known as the OOB score, which stands for Out of Bag, and refers to using bootstrap aggregating, or bagging, on sub-samples of the data for training.

Since points prediction was not the only goal of this project, I repeated all the previous work with the output of corsi % instead of points, but all of the results were worse than the results for points. The R-squared score for the linear regressor was 0.04 and for the random forest regressor was 0.014711, suggesting puck possession is even harder to predict with non-game metrics than points. However, the metric I found earlier that seemed most positively correlated to non-game metrics was hits, so I did one final round of calculations to predict hits. This time I calculated an R-squared score of 0.106 for the linear regressor and 0.140201 for the random forest regressor. So, predicting hits showed the best performance of any of regressors, but still was not even close to good enough to suggest a helpful model. Shown in figures 14, 15 and 16 are the plots for predicted versus actual points, corsi % and hits, all generated from the random forest regressors. None of the figures show any real shape to suggest a successful model.

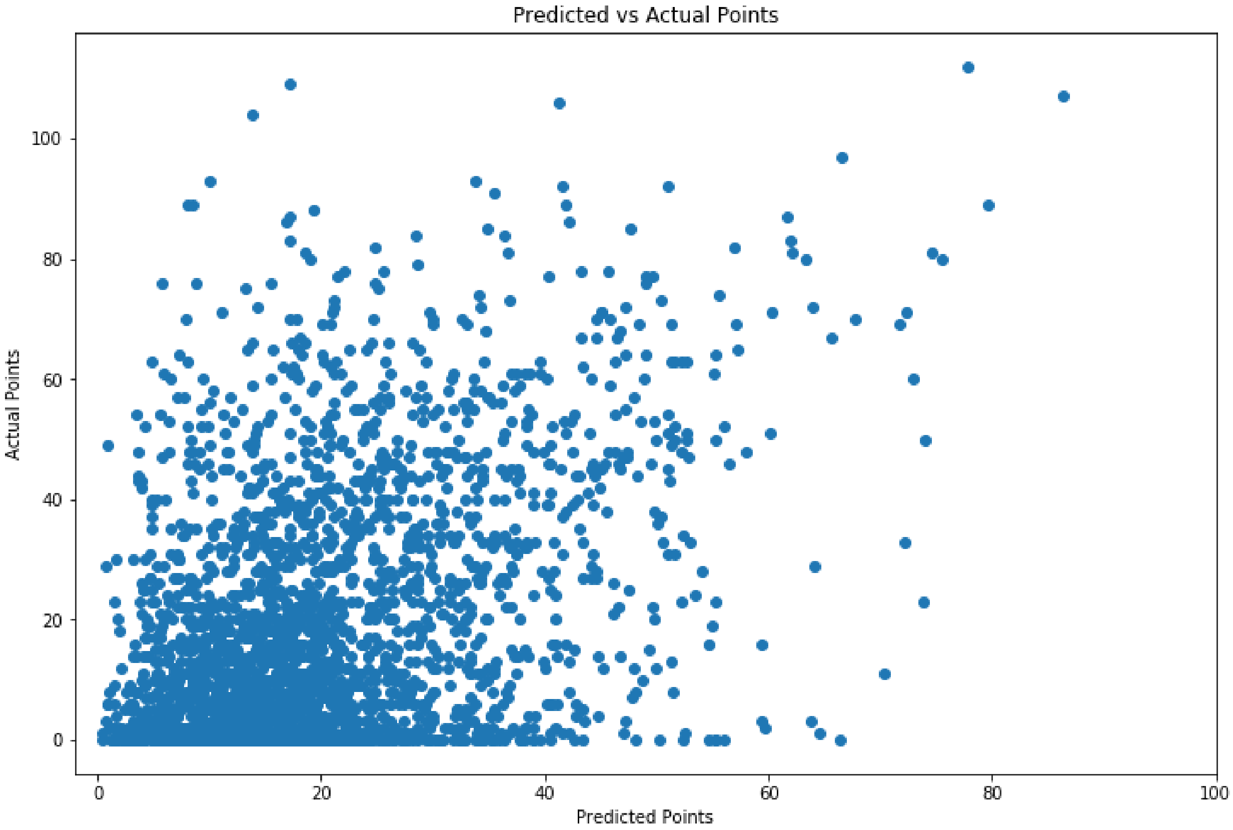


Figure 14. Predicted versus actual points from the random forest regressor

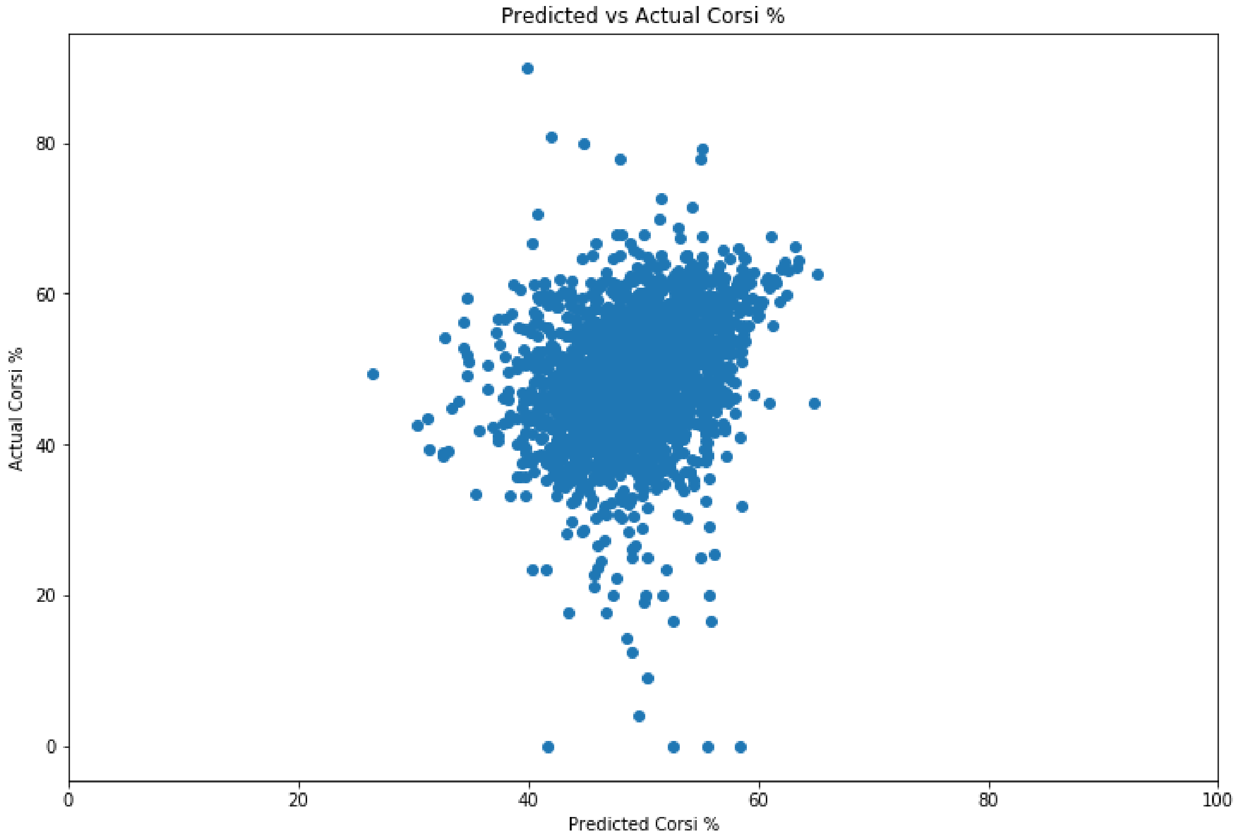


Figure 15. Predicted versus actual corsi % from the random forest regressor

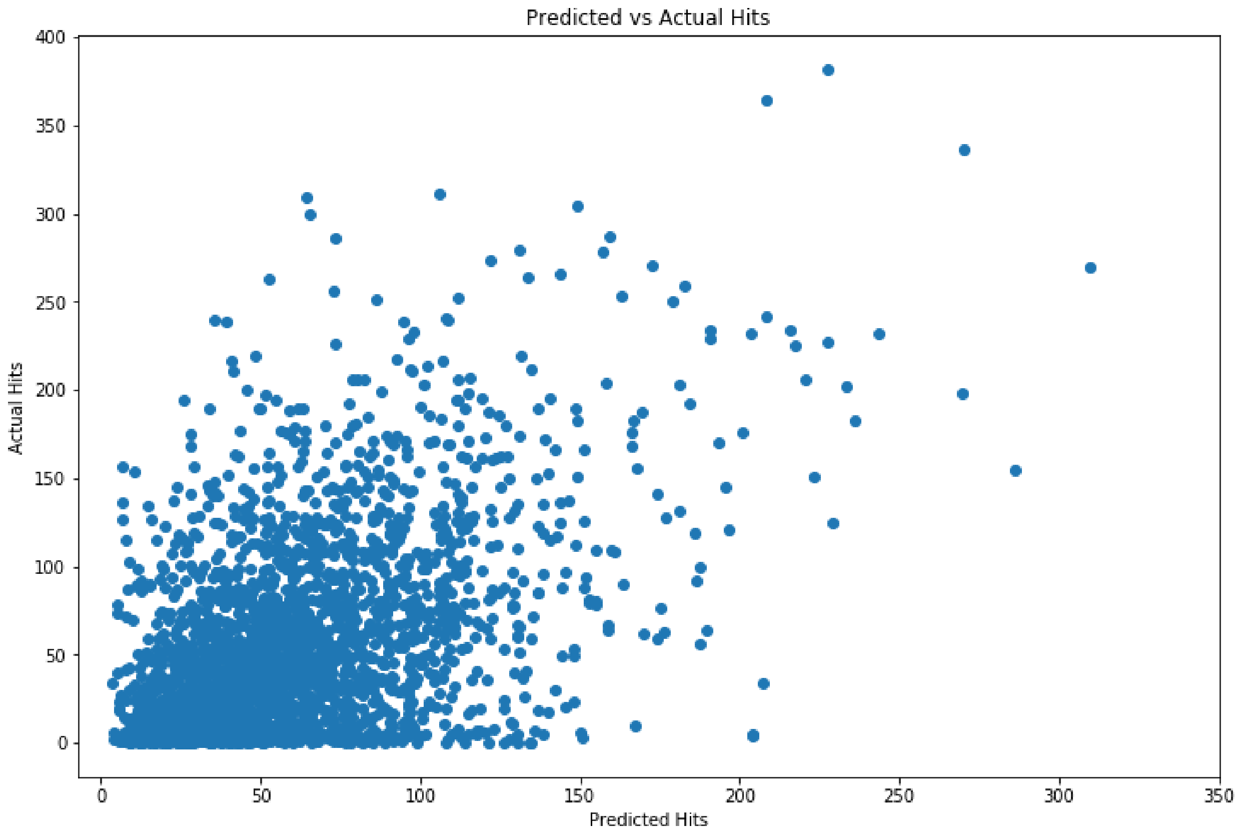
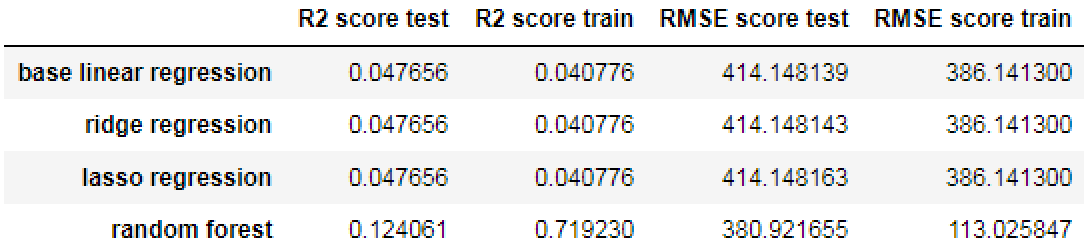


Figure 16. Predicted versus actual hits from the random forest regressor

Finally, figure 17 shows a table of the R-squared and RMSE scores for the various models. The random forest clearly shows the best performance, with the highest R-squared scores and the lowest RMSE, which stands for root mean square error, and refers to the difference between the correct output and the models' predictions, with smaller being better and zero being the best possible score.



	R2 score test	R2 score train	RMSE score test	RMSE score train
base linear regression	0.047656	0.040776	414.148139	386.141300
ridge regression	0.047656	0.040776	414.148143	386.141300
lasso regression	0.047656	0.040776	414.148163	386.141300
random forest	0.124061	0.719230	380.921655	113.025847

Figure 17. Table of R-squared and RMSE scores

Conclusions

My final results suggest predicting player performance using only non-game metrics is not reliable. Also, the article referenced in [2] has been verified in that player height doesn't seem to have much if any correlation to point production. Of further note, non-game metrics alone don't seem to have much influence on points, corsi percentage or really any statistic other than perhaps hits or shots blocked. I didn't find the project without merit however, finding many interesting notes. For instance, there appears to be some measurable difference between players of different nationalities, at least in some specific stats. Also, the figure showing the mean height and weight of players of different countries was a little surprising. Finally, seeing how player performance is affected by age was also very revealing.

Recommendations to Clients

According to my findings, any NHL teams, and likely any other hockey teams, still placing heavy emphasis on player height are probably looking at the wrong measurements. That isn't to say player height has no importance, but teams should not necessarily look to tall players to be scoring contributors. A team looking to boost their physical play or needing help defensively can certainly try to build their advantage in player height. Speaking of defensive play, teams trying to improve that area might apparently target Czech players, whereas teams wanting an improvement in scoring and puck possession may want to investigate Russian players. Also, the NHL recently has seen the value in younger, faster players, but apparently coaches still trust older players to begin play when starting in their own defensive zone. Age seems to have a progression as well, beginning with better puck possession from young players, to prime age players showing the most scoring and hitting, to older players being the best at blocking shots. Hopefully, this gives NHL teams some information to improve their weaknesses or build on their strengths.

Future Work

The most obvious omission from this project is goalie statistics. However, the website I collected my data from did not have that information available. Thus, the next step for this research would be to find those statistics from other sources and perform the same calculations for goalies as well, who also seem to suffer or benefit from height discrimination. Further study can be made to determine the mean height, weight and/or age of individual NHL teams to see if any of the results of this project can help explain a team's defensive or scoring statistics based on those metrics. Also, player data can be collected again after future seasons and added to this existing data to track changes in the results.

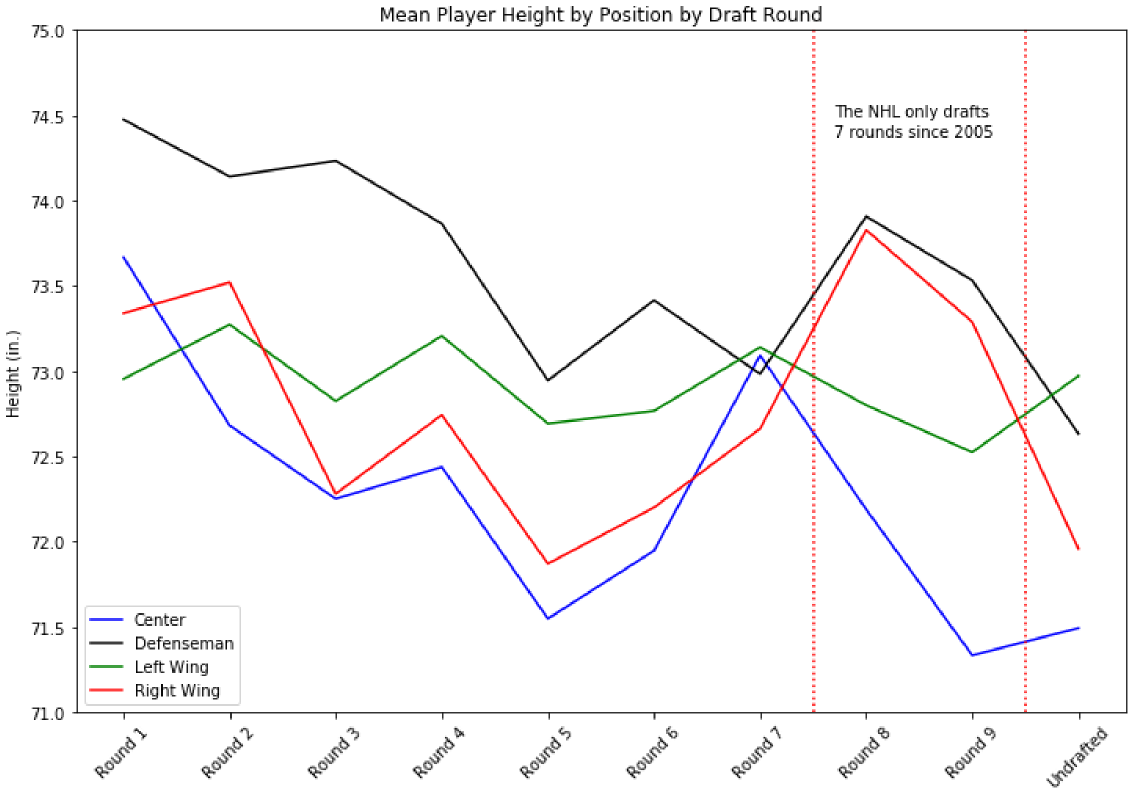
References

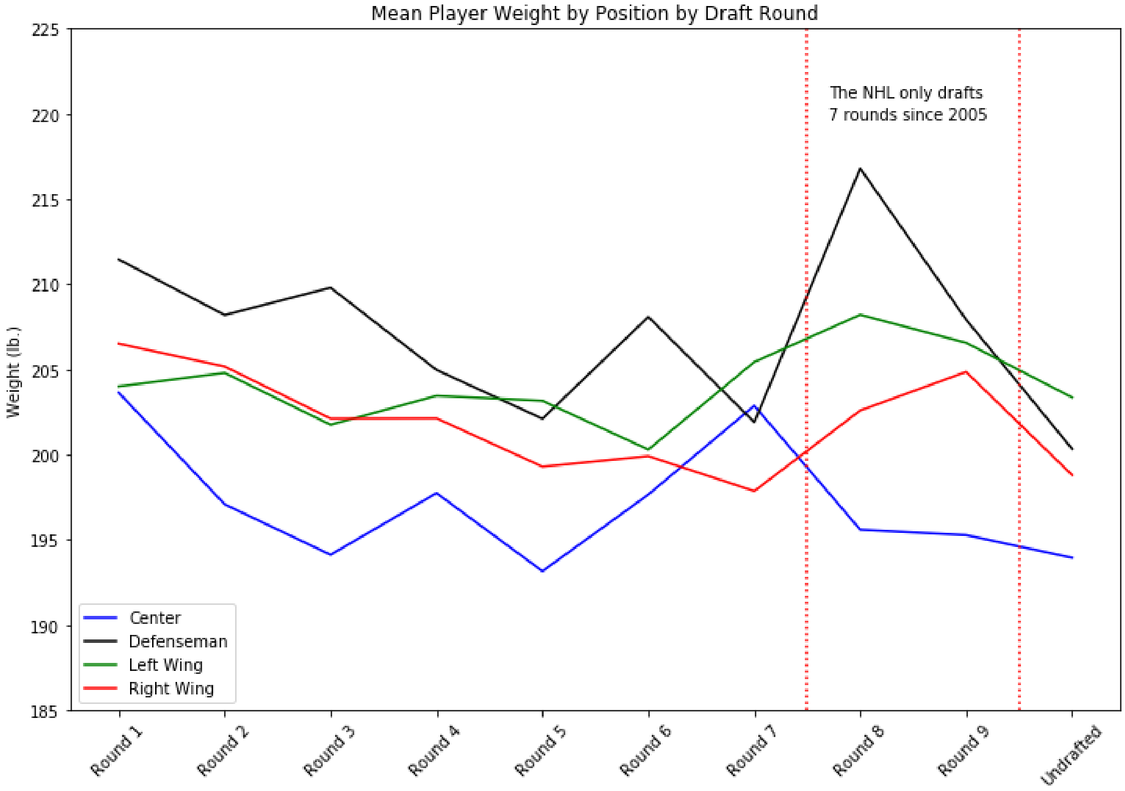
1. Denson, Ben: The NHL is dead. The Cornell Daily Sun. http://cornellsun.com/2016/02/09/denson-the-nhl-is-dead/ (2016)

2. Curry, P.: Hockey analytics: Does size really matter in the NHL? The Star https://www.thestar.com/sports/hockey/2014/03/13/hockey_analytics_does_size_really_matter_in_the_nhl.html (2014)

Appendix A

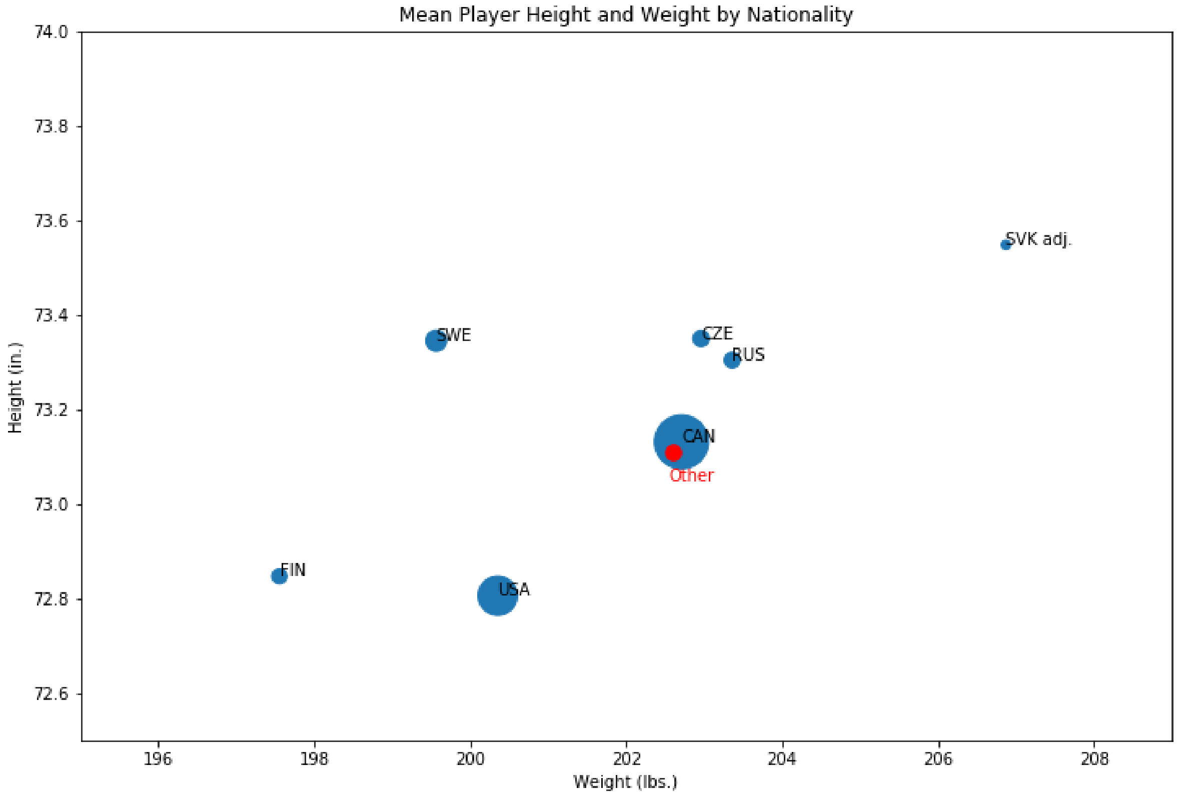
Mean Player Height and Weight by Draft Round




Mean Player Weight by Position by Draft Round
Weight (lb.)
225
220
215
210
205
200
195
190
185
The NHL only drafts
7 rounds since 2005
Center
Defenseman
Left Wing
Right Wing
Round 1
Round 2
Round 3
Round 4
Round 5
Round 6
Round 7
Round 8
Round 9
Undrafted

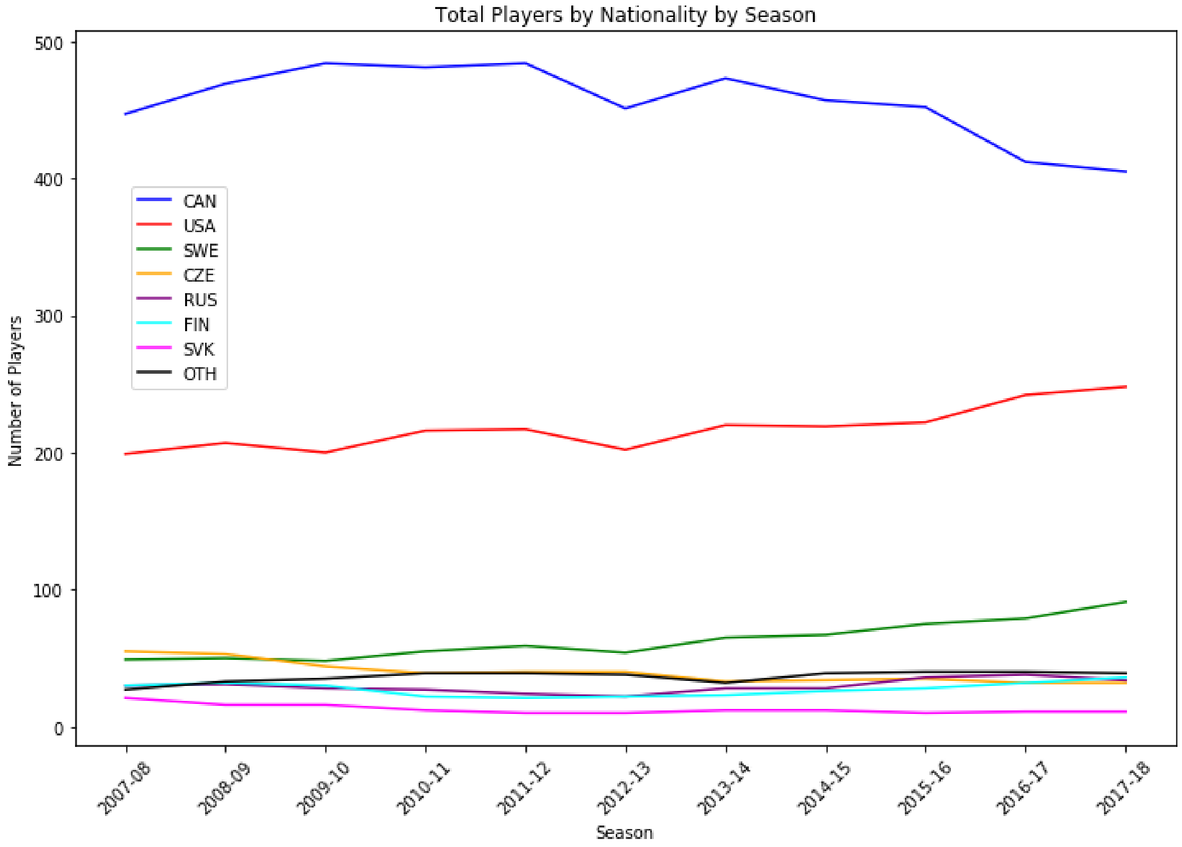
Appendix B

Adjusted Mean Player Height and Weight by Nationality



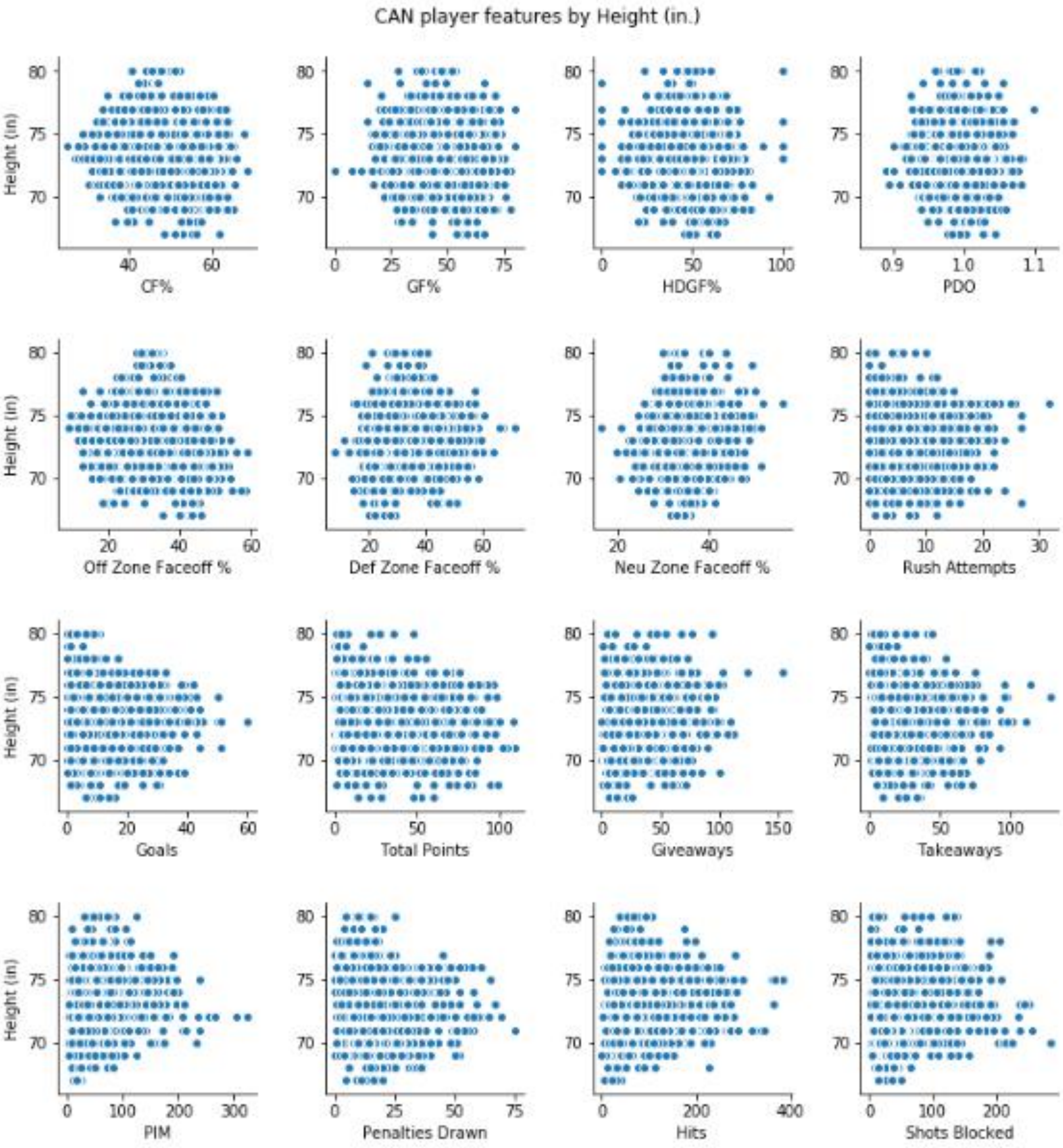
Appendix C

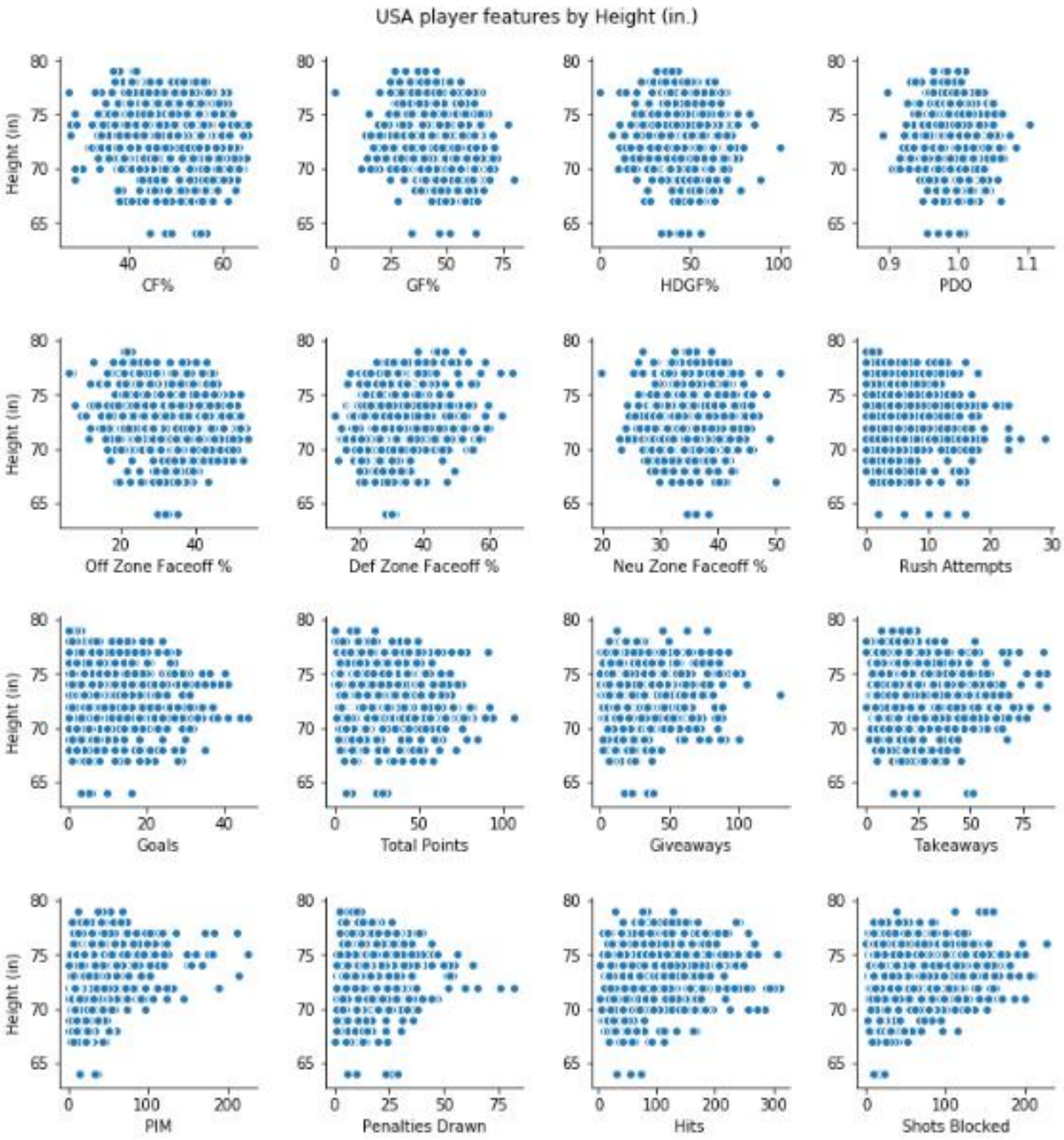
Total Player Count by Nationality

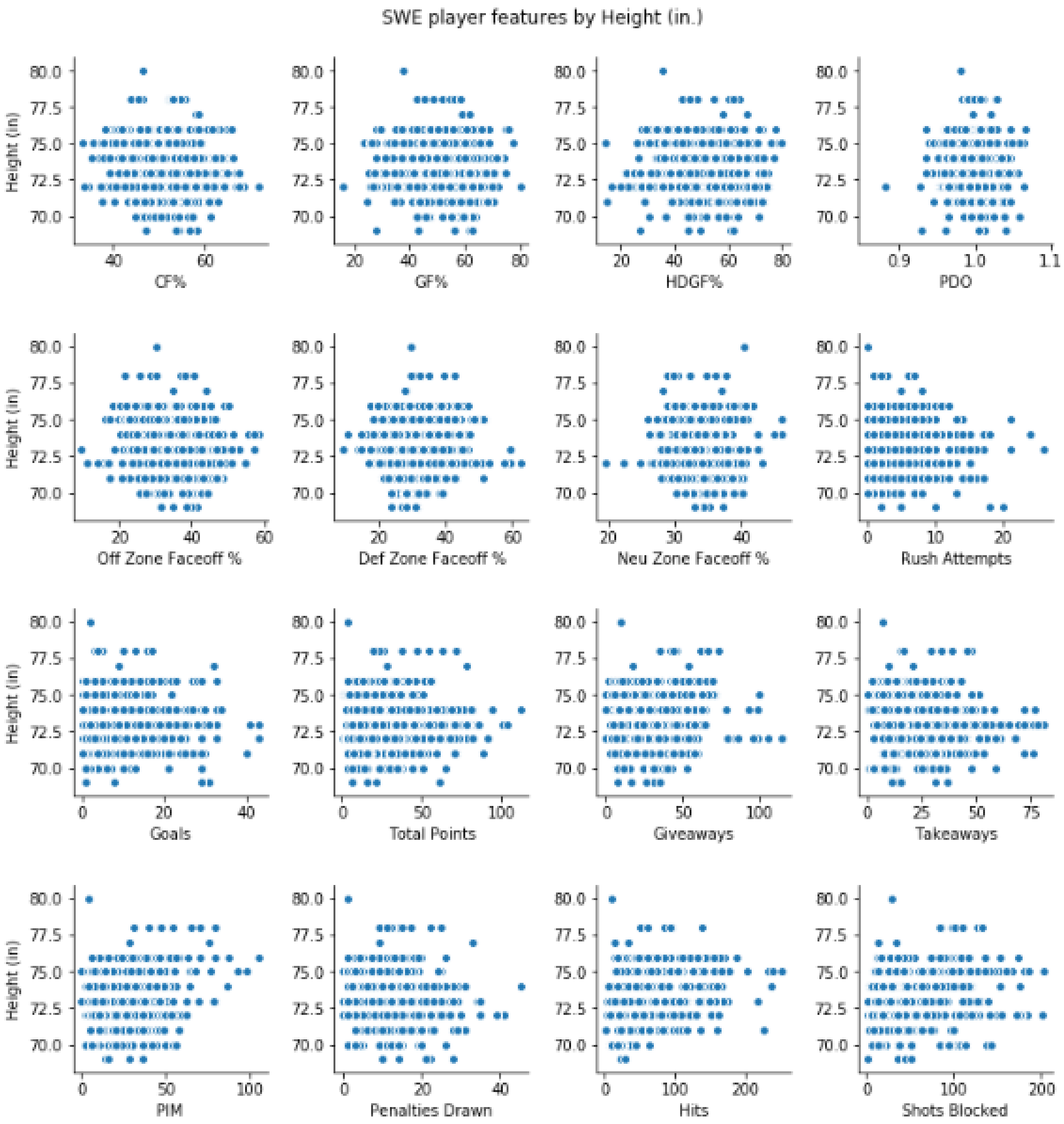


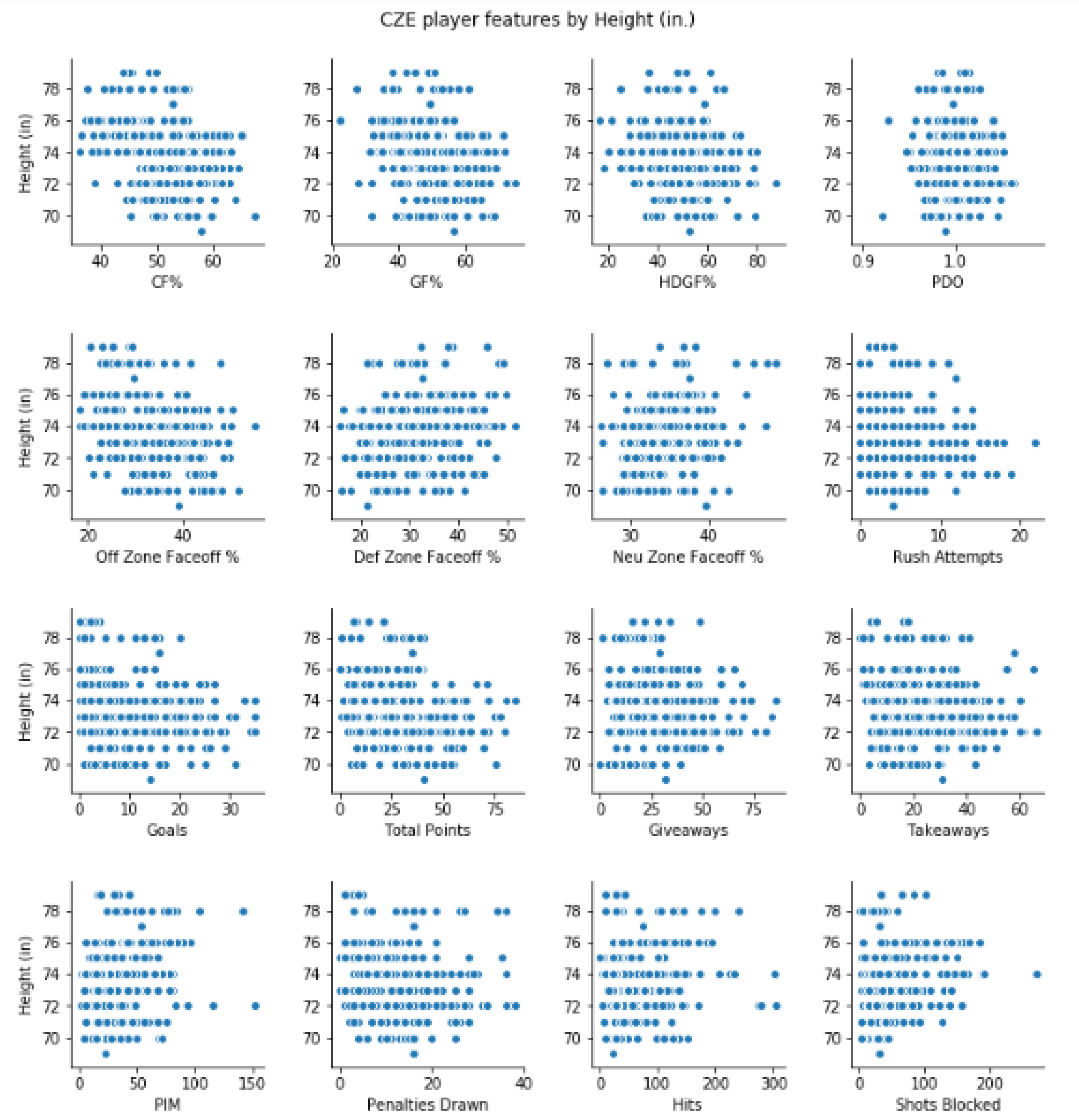
Appendix D

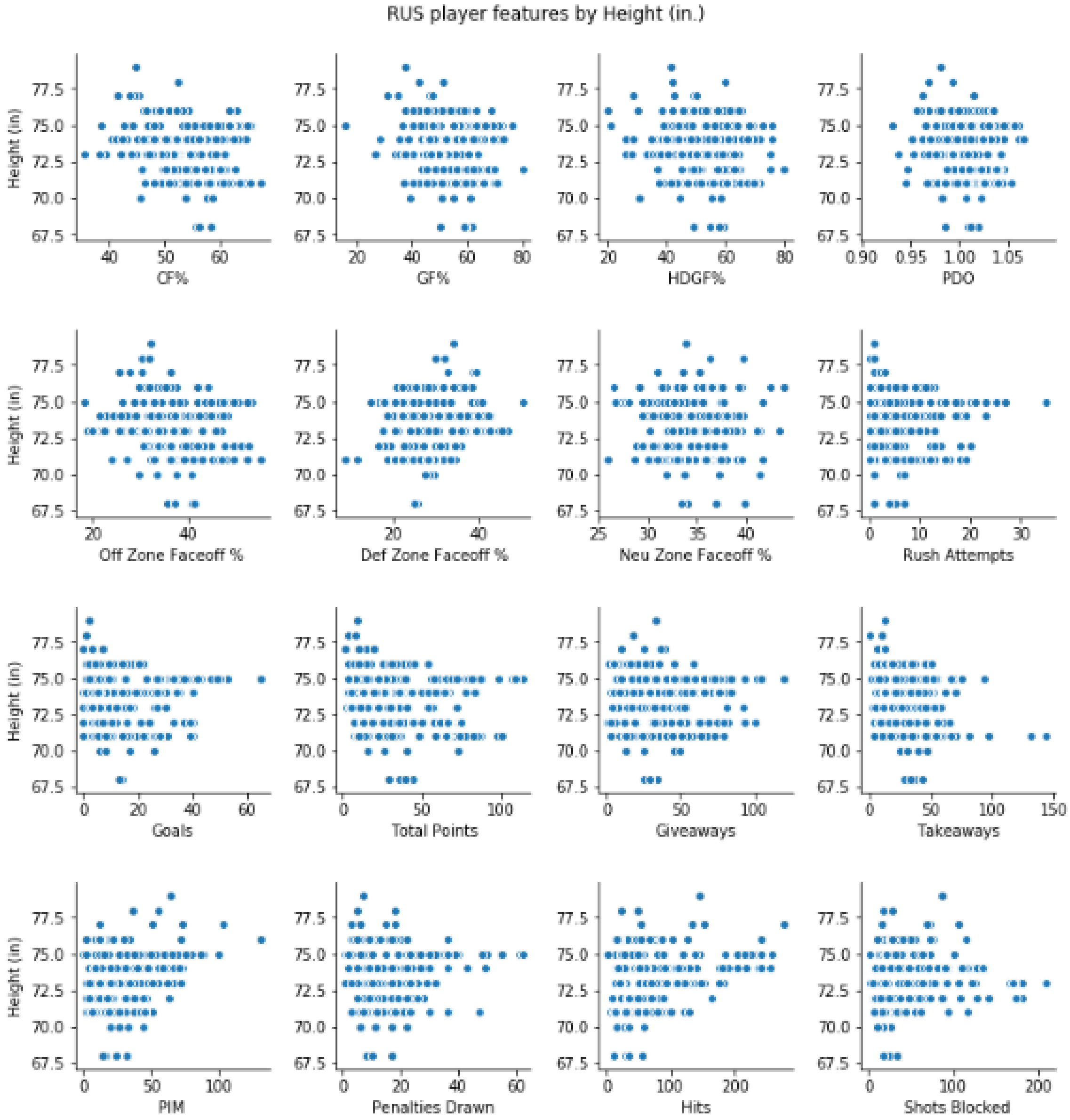
Scatter Plots of Various Metrics by Player Height and Weight and Nationality

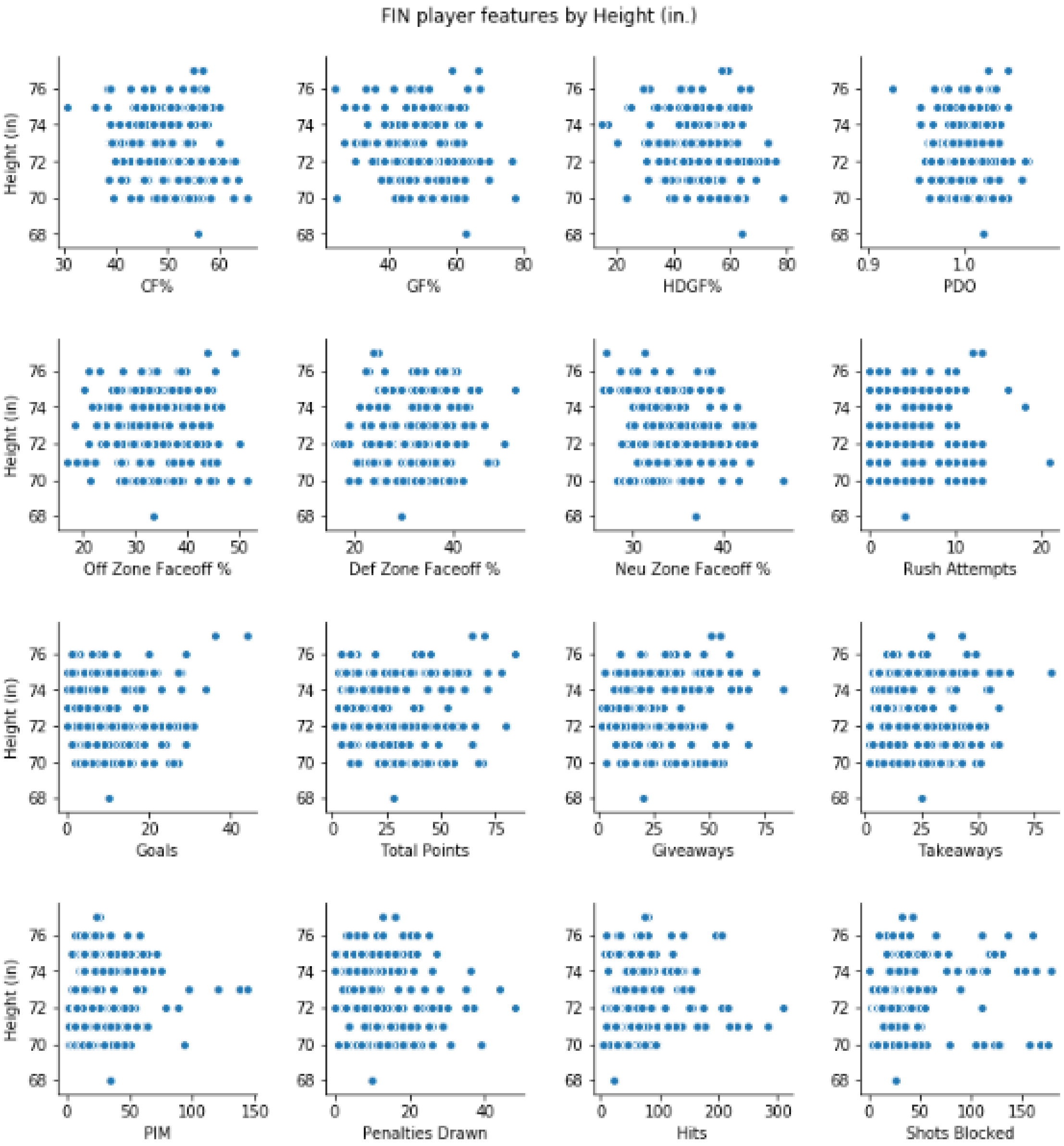


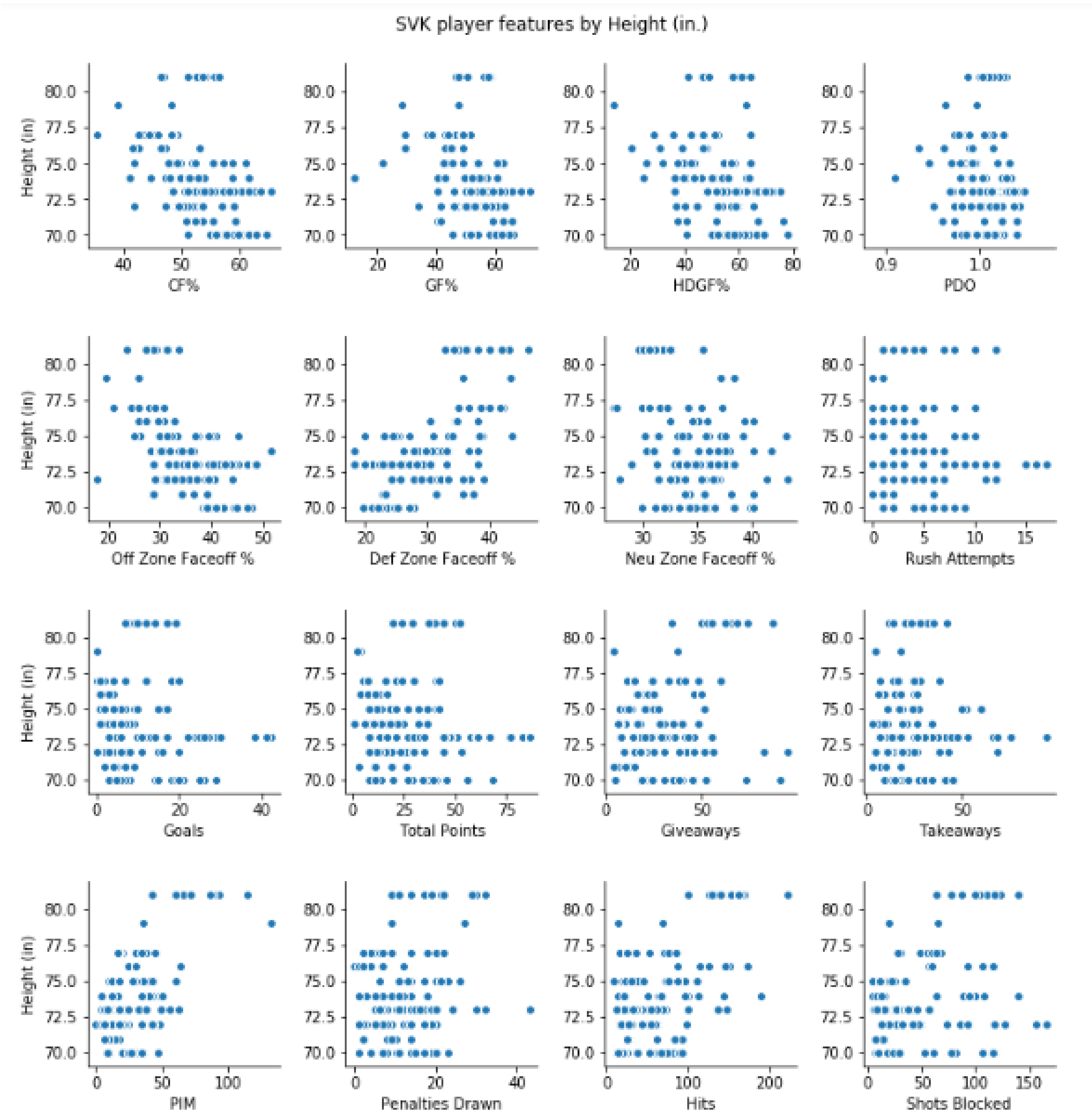

USA player features by Height (in.)
Height (in)
80
75
70
65
CF%
40
60
GF%
0
25
50
75
HDGF%
0
50
100
PDO
0.9
1.0
1.1
Off Zone Faceoff %
20
40
Def Zone Faceoff %
20
40
60
Neu Zone Faceoff %
20
30
40
50
Rush Attempts
0
10
20
30
Goals
0
20
40
Total Points
0
50
100
Giveaways
0
50
100
Takeaways
0
25
50
75
PIM
0
100
200
Penalties Drawn
0
25
50
75
Hits
0
100
200
300
Shots Blocked
0
100
200

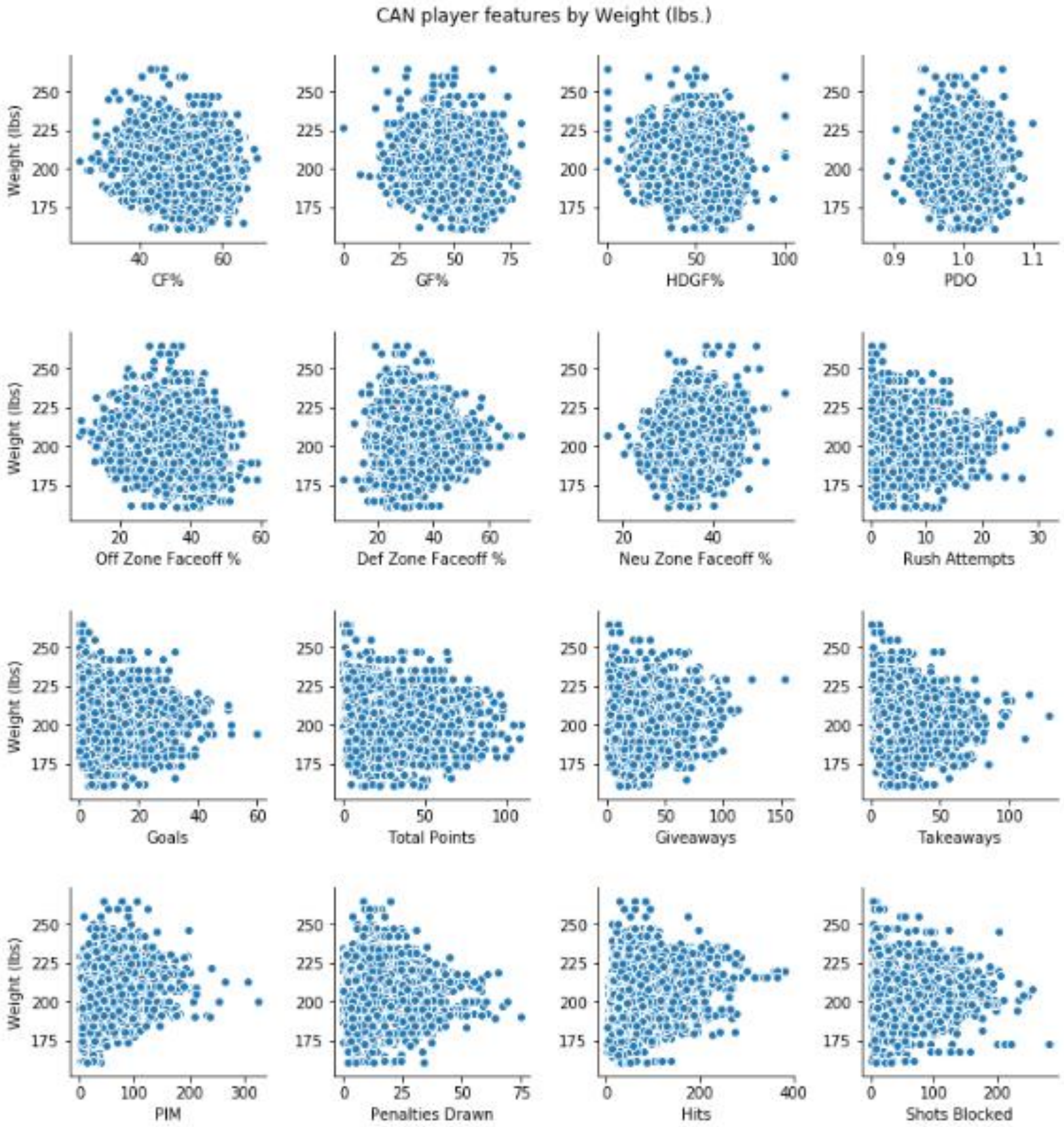

SWE player features by Height (in.)
Height (in)
80.0
77.5
75.0
72.5
70.0
CF%
40
60
GF%
20
40
60
80
HDGF%
20
40
60
80
PDO
0.9
1.0
1.1
Off Zone Faceoff %
20
40
60
Def Zone Faceoff %
20
40
60
Neu Zone Faceoff %
20
30
40
Rush Attempts
0
10
20
Goals
0
20
40
Total Points
0
50
100
Giveaways
0
50
100
Takeaways
0
25
50
75
PIM
0
50
100
Penalties Drawn
0
20
40
Hits
0
100
200
Shots Blocked
0
100
200

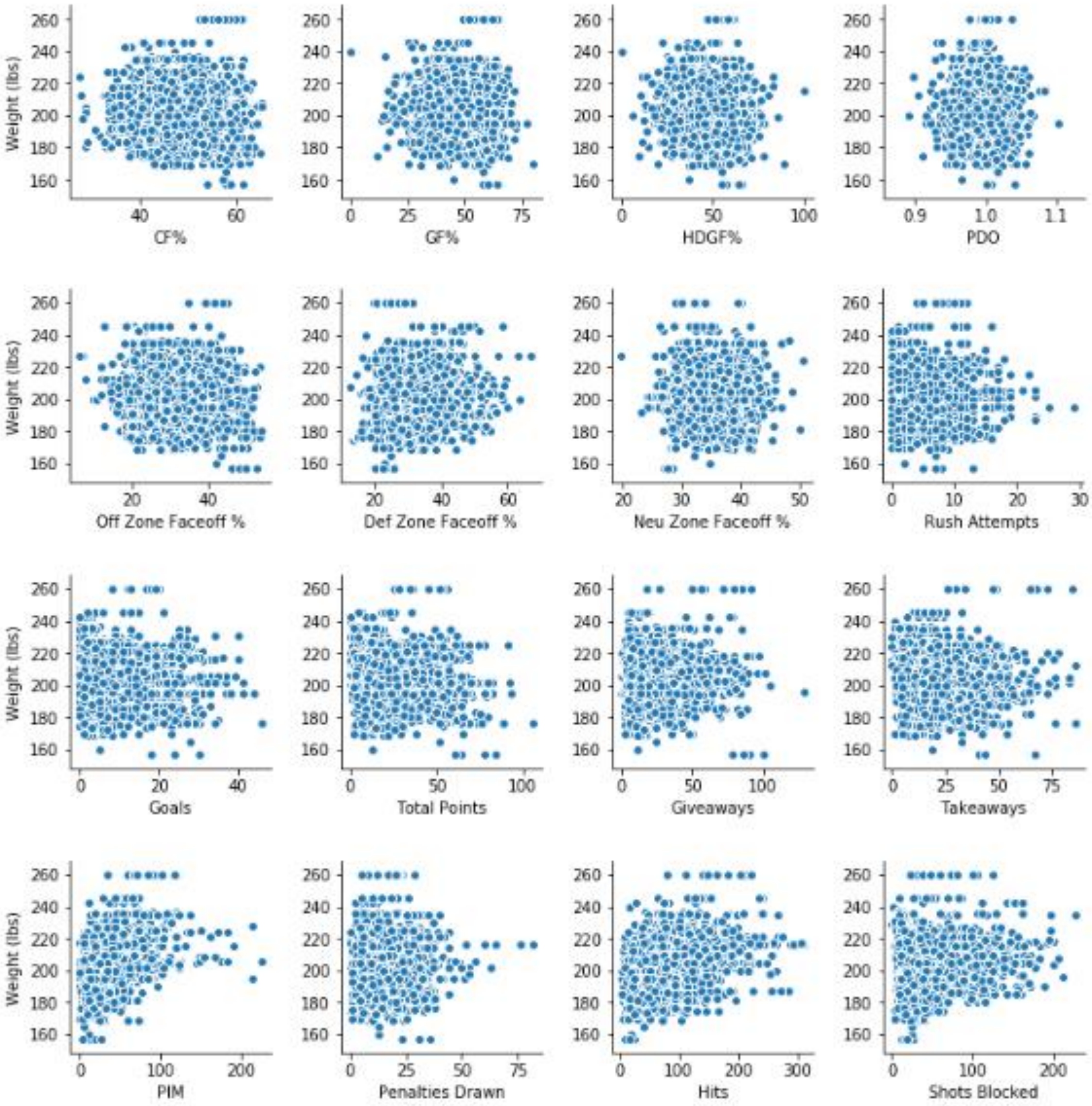

CZE player features by Height (in.)
Height (in)
78
76
74
72
70
CF%
40
50
60
GF%
20
40
60
HDGF%
20
40
60
80
PDO
0.9
1.0
Off Zone Faceoff %
20
40
Def Zone Faceoff %
20
30
40
50
Neu Zone Faceoff %
30
40
Rush Attempts
0
10
20
Goals
0
10
20
30
Total Points
0
25
50
75
Giveaways
0
25
50
75
Takeaways
0
20
40
60
PIM
0
50
100
150
Penalties Drawn
0
20
40
Hits
0
100
200
300
Shots Blocked
0
100
200

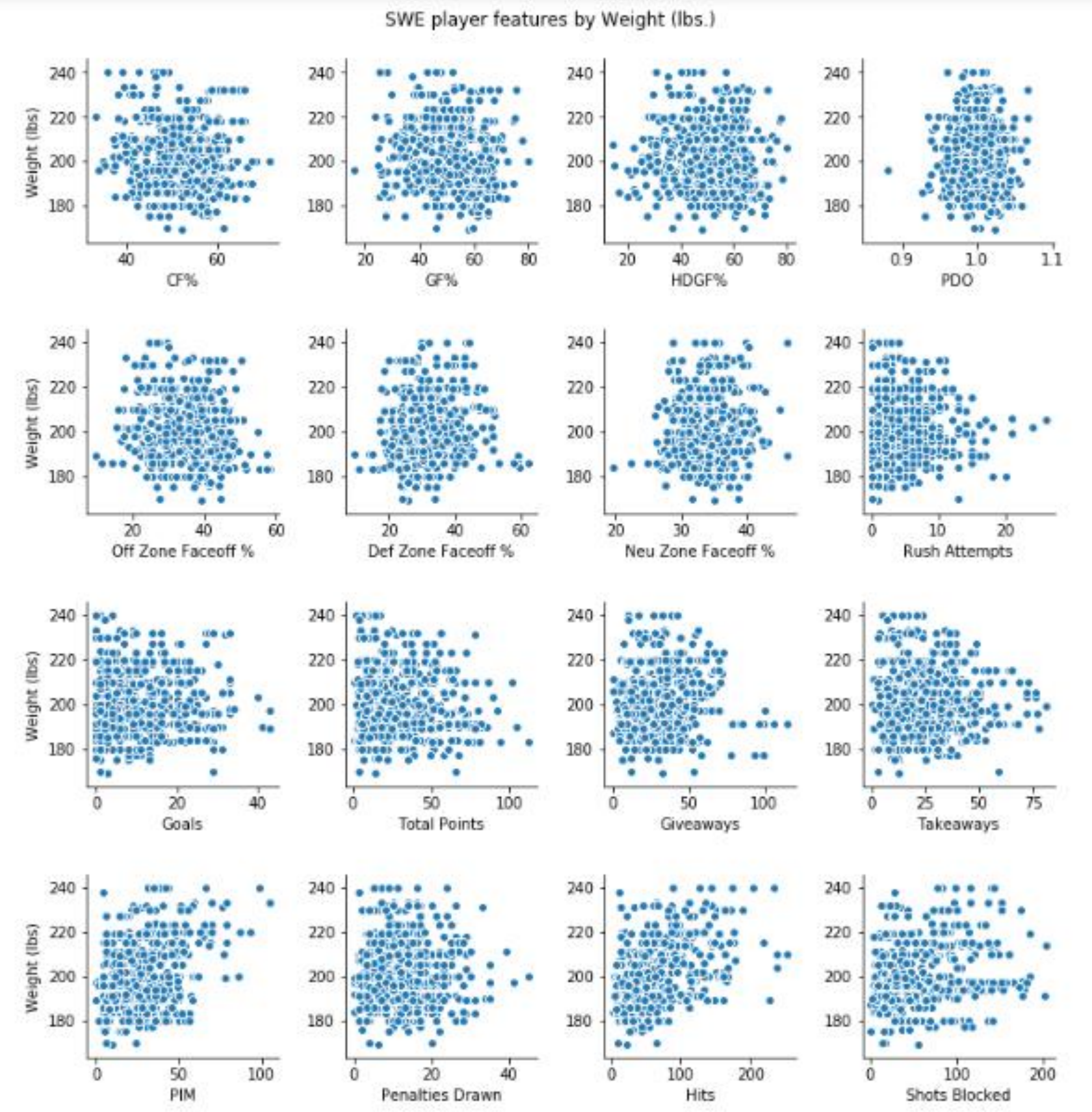

RUS player features by Height (in.)
Height (in)
77.5
75.0
72.5
70.0
67.5
CF%
GF%
HDGF%
PDO
Off Zone Faceoff %
Def Zone Faceoff %
Neu Zone Faceoff %
Rush Attempts
Goals
Total Points
Giveaways
Takeaways
PIM
Penalties Drawn
Hits
Shots Blocked


FIN player features by Height (in.)
Height (in)
76
74
72
70
68
CF%
30
40
50
60
GF%
40
60
80
HDGF%
20
40
60
80
PDO
0.9
1.0
Off Zone Faceoff %
20
30
40
50
Def Zone Faceoff %
20
40
Neu Zone Faceoff %
30
40
Rush Attempts
0
10
20
Goals
0
20
40
Total Points
0
25
50
75
Giveaways
0
25
50
75
Takeaways
0
25
50
75
PIM
0
50
100
150
Penalties Drawn
0
20
40
Hits
0
100
200
300
Shots Blocked
0
50
100
150

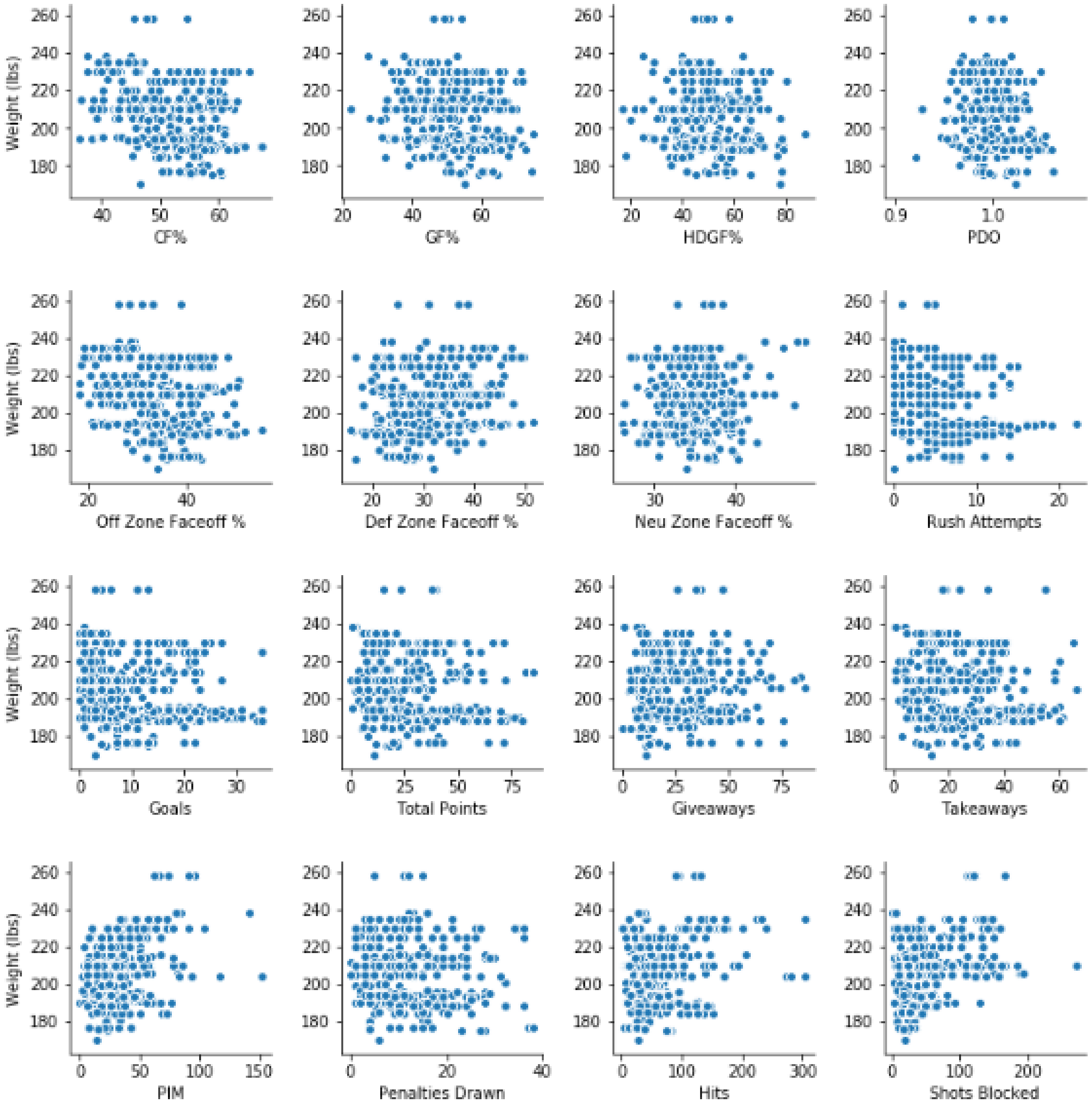

SVK player features by Height (in.)
Height (in)
80.0
77.5
75.0
72.5
70.0
CF%
40
50
60
GF%
20
40
60
HDGF%
20
40
60
80
PDO
0.9
1.0
Off Zone Faceoff %
20
30
40
50
Def Zone Faceoff %
20
30
40
Neu Zone Faceoff %
30
35
40
Rush Attempts
0
5
10
15
Goals
0
20
40
Total Points
0
25
50
75
Giveaways
0
50
Takeaways
0
50
PIM
0
50
100
Penalties Drawn
0
20
40
Hits
0
100
200
Shots Blocked
0
50
100
150

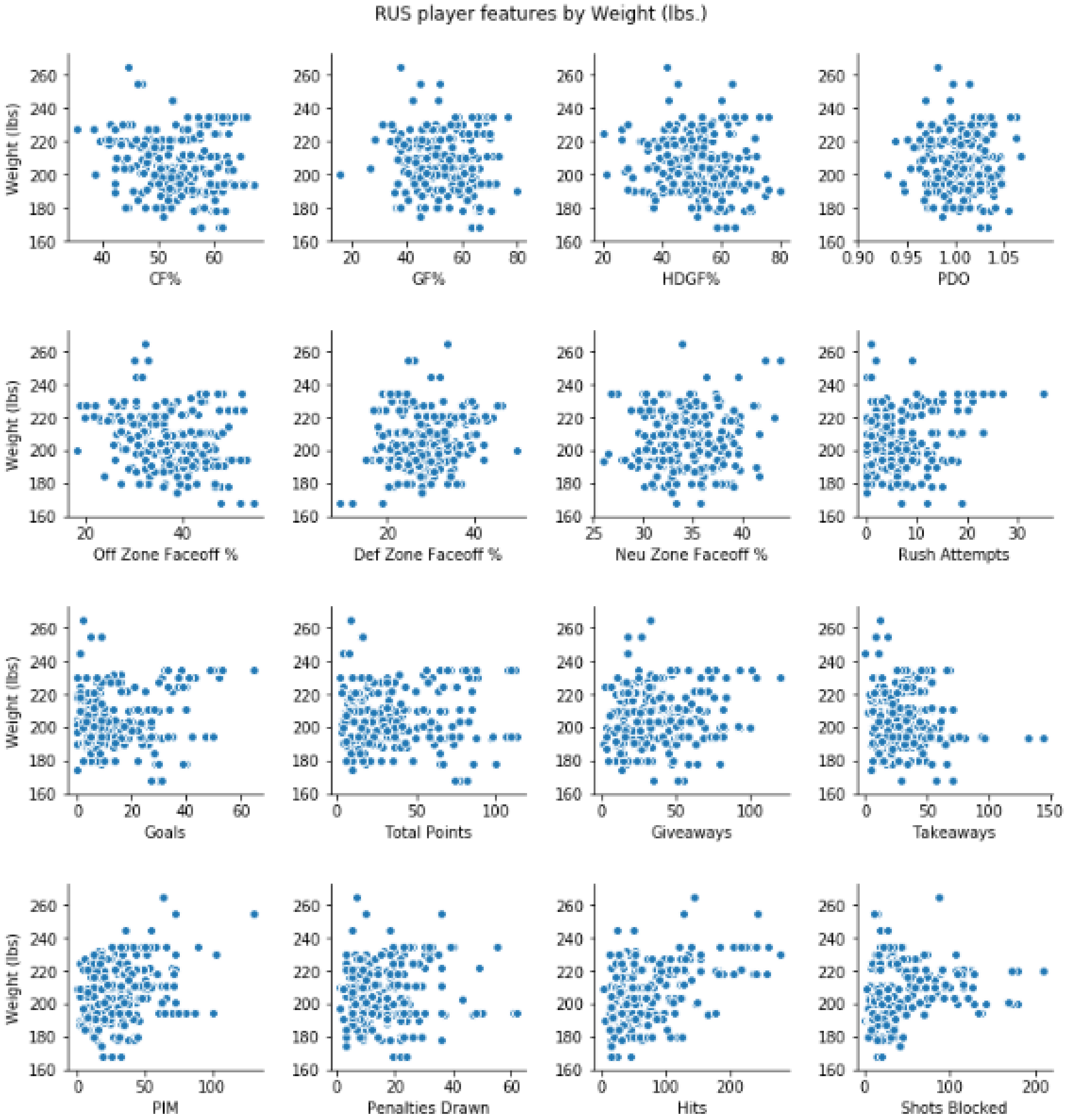

CAN player features by Weight (lbs.)
Weight (lbs)
175
200
225
250
CF%
40
60
GF%
0
25
50
75
HDGF%
0
50
100
PDO
0.9
1.0
1.1
Off Zone Faceoff %
20
40
60
Def Zone Faceoff %
20
40
60
Neu Zone Faceoff %
20
40
Rush Attempts
0
10
20
30
Goals
0
20
40
60
Total Points
0
50
100
Giveaways
0
50
100
150
Takeaways
0
50
100
PIM
0
100
200
300
Penalties Drawn
0
25
50
75
Hits
0
200
400
Shots Blocked
0
100
200


USA player features by Weight (lbs.)

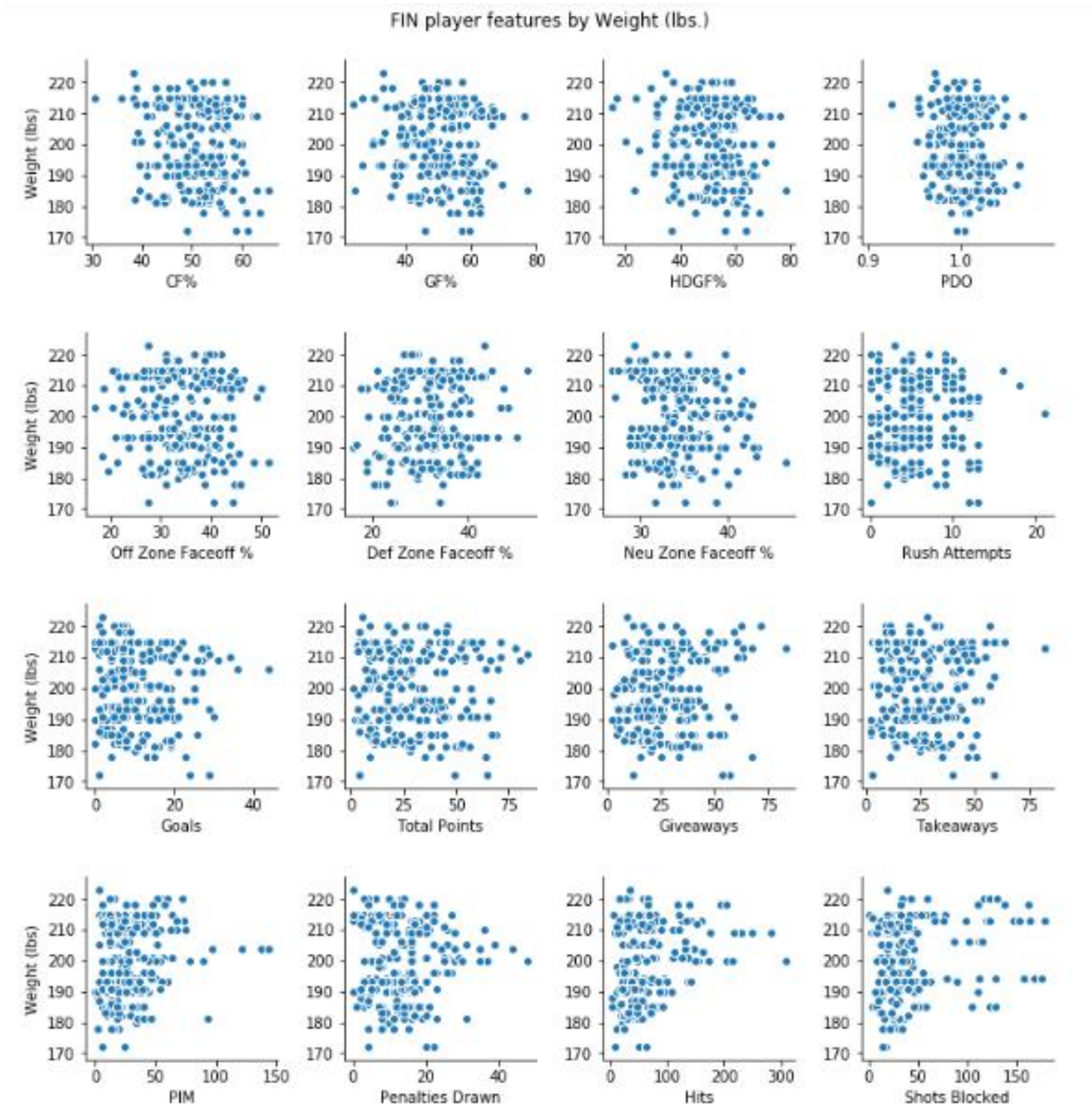

SWE player features by Weight (lbs.)
Weight (lbs)
240
220
200
180
CF%
GF%
HDGF%
PDO
Off Zone Faceoff %
Def Zone Faceoff %
Neu Zone Faceoff %
Rush Attempts
Goals
Total Points
Giveaways
Takeaways
PIM
Penalties Drawn
Hits
Shots Blocked

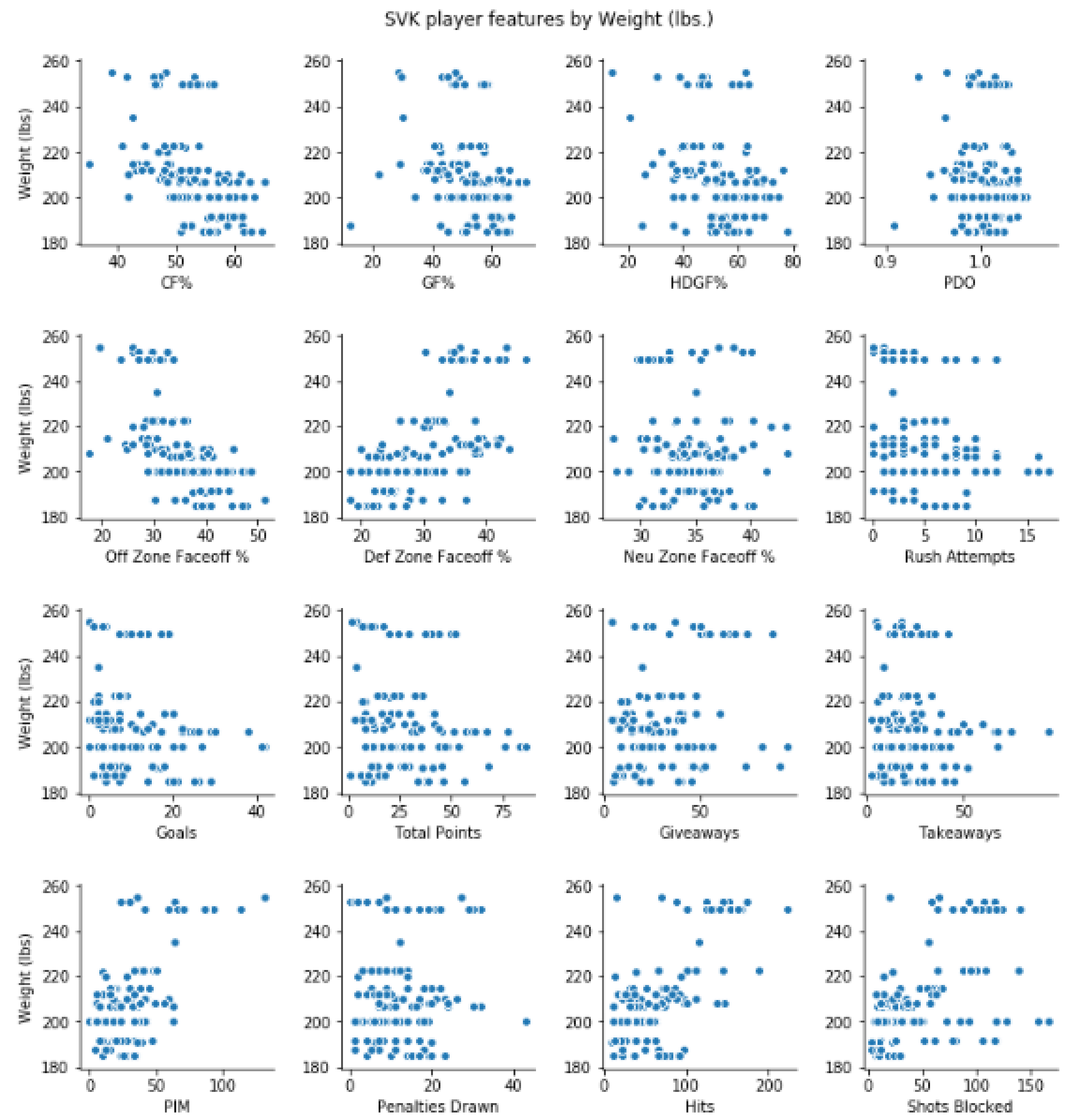
CZE player features by Weight (lbs.)





RUS player features by Weight (lbs.)

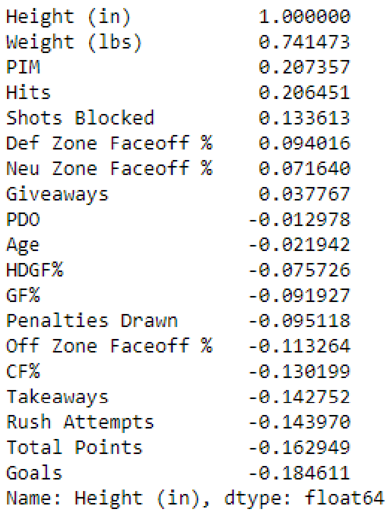

FIN player features by Weight (lbs.)
Weight (lbs)
220
210
200
190
180
170
CF%
30
40
50
60
GF%
40
60
80
HDGF%
20
40
60
80
PDO
0.9
1.0
Off Zone Faceoff %
20
30
40
50
Def Zone Faceoff %
20
40
Neu Zone Faceoff %
30
40
Rush Attempts
0
10
20
Goals
0
20
40
Total Points
0
25
50
75
Giveaways
0
25
50
75
Takeaways
0
25
50
75
PIM
0
50
100
150
Penalties Drawn
0
20
40
Hits
0
100
200
300
Shots Blocked
0
50
100
150


SVK player features by Weight (lbs.)
Weight (lbs)
260
240
220
200
180
CF%
40
50
60
GF%
20
40
60
HDGF%
20
40
60
80
PDO
0.9
1.0
Off Zone Faceoff %
20
30
40
50
Def Zone Faceoff %
20
30
40
Neu Zone Faceoff %
30
35
40
Rush Attempts
0
5
10
15
Goals
0
20
40
Total Points
0
25
50
75
Giveaways
0
50
Takeaways
0
50
PIM
0
50
100
Penalties Drawn
0
20
40
Hits
0
100
200
Shots Blocked
0
50
100
150

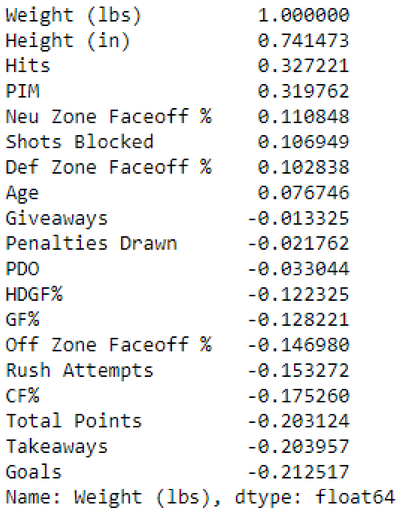
Appendix E

Estimated Correlation Coefficients by Nationality

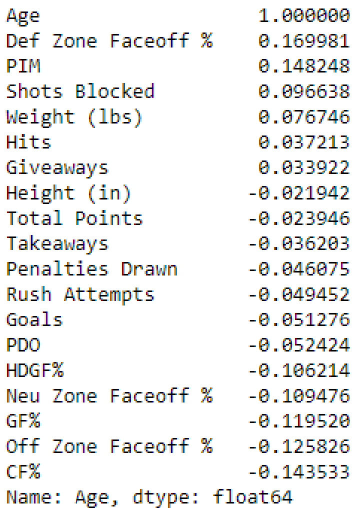
Canadian player estimated correlation coefficients



```
Height (in)             1.000000
Weight (lbs)            0.741473
PIM                     0.207357
Hits                    0.206451
Shots Blocked           0.133613
Def Zone Faceoff %      0.094016
Neu Zone Faceoff %      0.071640
Giveaways               0.037767
PDO                    -0.012978
Age                    -0.021942
HDGF%                  -0.075726
GF%                    -0.091927
Penalties Drawn        -0.095118
Off Zone Faceoff %     -0.113264
CF%                    -0.130199
Takeaways              -0.142752
Rush Attempts          -0.143970
Total Points           -0.162949
Goals                  -0.184611
Name: Height (in), dtype: float64
```

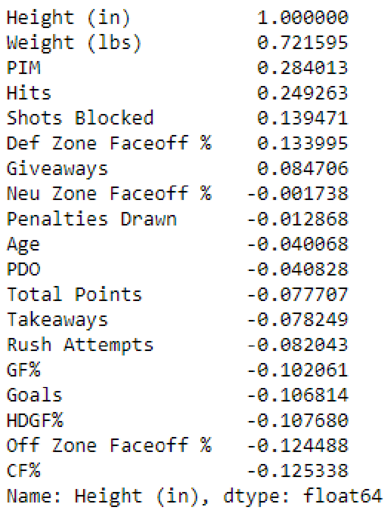


```
Weight (lbs)            1.000000
Height (in)             0.741473
Hits                    0.327221
PIM                     0.319762
Neu Zone Faceoff %      0.110848
Shots Blocked           0.106949
Def Zone Faceoff %      0.102838
Age                     0.076746
Giveaways              -0.013325
Penalties Drawn        -0.021762
PDO                    -0.033044
HDGF%                  -0.122325
GF%                    -0.128221
Off Zone Faceoff %     -0.146980
Rush Attempts          -0.153272
CF%                    -0.175260
Total Points           -0.203124
Takeaways              -0.203957
Goals                  -0.212517
Name: Weight (lbs), dtype: float64
```

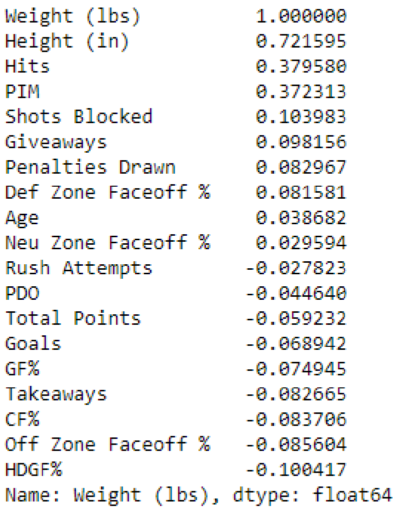


```
Age                     1.000000
Def Zone Faceoff %      0.169981
PIM                     0.148248
Shots Blocked           0.096638
Weight (lbs)            0.076746
Hits                    0.037213
Giveaways               0.033922
Height (in)            -0.021942
Total Points           -0.023946
Takeaways              -0.036203
Penalties Drawn        -0.046075
Rush Attempts          -0.049452
Goals                  -0.051276
PDO                    -0.052424
HDGF%                  -0.106214
Neu Zone Faceoff %     -0.109476
GF%                    -0.119520
Off Zone Faceoff %     -0.125826
CF%                    -0.143533
Name: Age, dtype: float64
```

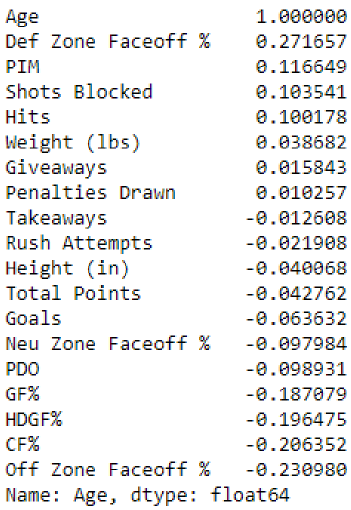
US player estimated correlation coefficients



```
Height (in)             1.000000
Weight (lbs)            0.721595
PIM                     0.284013
Hits                    0.249263
Shots Blocked           0.139471
Def Zone Faceoff %      0.133995
Giveaways               0.084706
Neu Zone Faceoff %     -0.001738
Penalties Drawn        -0.012868
Age                    -0.040068
PDO                    -0.040828
Total Points           -0.077707
Takeaways              -0.078249
Rush Attempts          -0.082043
GF%                    -0.102061
Goals                  -0.106814
HDGF%                  -0.107680
Off Zone Faceoff %     -0.124488
CF%                    -0.125338
Name: Height (in), dtype: float64
```

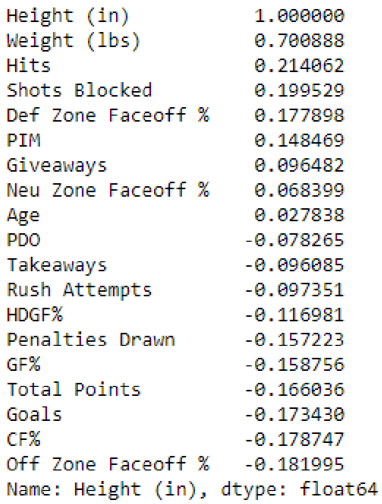


```
Weight (lbs)            1.000000
Height (in)             0.721595
Hits                    0.379580
PIM                     0.372313
Shots Blocked           0.103983
Giveaways               0.098156
Penalties Drawn         0.082967
Def Zone Faceoff %      0.081581
Age                     0.038682
Neu Zone Faceoff %      0.029594
Rush Attempts          -0.027823
PDO                    -0.044640
Total Points           -0.059232
Goals                  -0.068942
GF%                    -0.074945
Takeaways              -0.082665
CF%                    -0.083706
Off Zone Faceoff %     -0.085604
HDGF%                  -0.100417
Name: Weight (lbs), dtype: float64
```

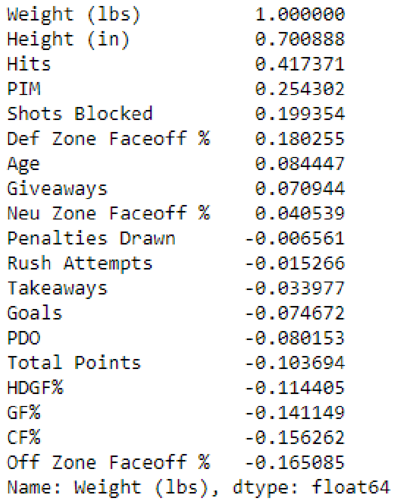


```
Age                     1.000000
Def Zone Faceoff %      0.271657
PIM                     0.116649
Shots Blocked           0.103541
Hits                    0.100178
Weight (lbs)            0.038682
Giveaways               0.015843
Penalties Drawn         0.010257
Takeaways              -0.012608
Rush Attempts          -0.021908
Height (in)            -0.040068
Total Points           -0.042762
Goals                  -0.063632
Neu Zone Faceoff %     -0.097984
PDO                    -0.098931
GF%                    -0.187079
HDGF%                  -0.196475
CF%                    -0.206352
Off Zone Faceoff %     -0.230980
Name: Age, dtype: float64
```

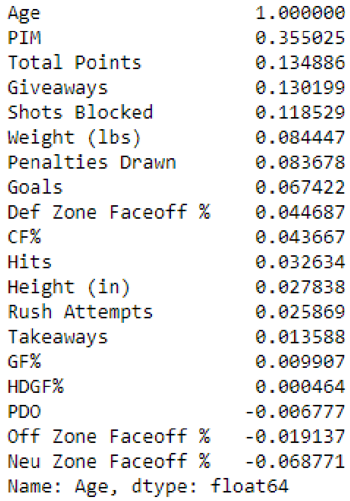
Swedish player estimated correlation coefficients



```
Height (in)            1.000000
Weight (lbs)           0.700888
Hits                   0.214062
Shots Blocked          0.199529
Def Zone Faceoff %     0.177898
PIM                    0.148469
Giveaways              0.096482
Neu Zone Faceoff %     0.068399
Age                    0.027838
PDO                   -0.078265
Takeaways             -0.096085
Rush Attempts         -0.097351
HDGF%                 -0.116981
Penalties Drawn       -0.157223
GF%                   -0.158756
Total Points          -0.166036
Goals                 -0.173430
CF%                   -0.178747
Off Zone Faceoff %    -0.181995
Name: Height (in), dtype: float64
```

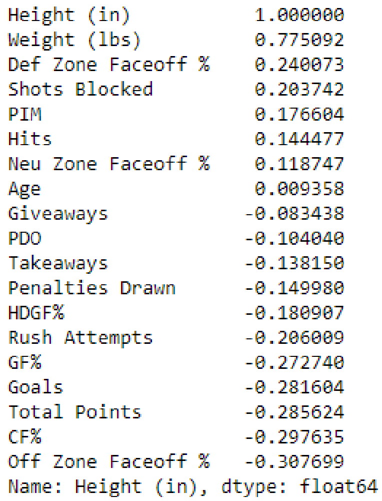


```
Weight (lbs)           1.000000
Height (in)            0.700888
Hits                   0.417371
PIM                    0.254302
Shots Blocked          0.199354
Def Zone Faceoff %     0.180255
Age                    0.084447
Giveaways              0.070944
Neu Zone Faceoff %     0.040539
Penalties Drawn       -0.006561
Rush Attempts         -0.015266
Takeaways             -0.033977
Goals                 -0.074672
PDO                   -0.080153
Total Points          -0.103694
HDGF%                 -0.114405
GF%                   -0.141149
CF%                   -0.156262
Off Zone Faceoff %    -0.165085
Name: Weight (lbs), dtype: float64
```

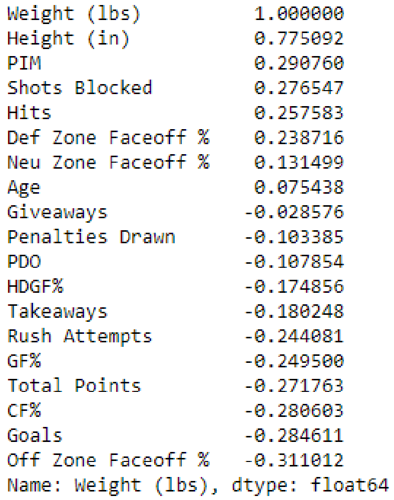


```
Age                    1.000000
PIM                    0.355025
Total Points           0.134886
Giveaways              0.130199
Shots Blocked          0.118529
Weight (lbs)           0.084447
Penalties Drawn        0.083678
Goals                  0.067422
Def Zone Faceoff %     0.044687
CF%                    0.043667
Hits                   0.032634
Height (in)            0.027838
Rush Attempts          0.025869
Takeaways              0.013588
GF%                    0.009907
HDGF%                  0.000464
PDO                   -0.006777
Off Zone Faceoff %    -0.019137
Neu Zone Faceoff %    -0.068771
Name: Age, dtype: float64
```

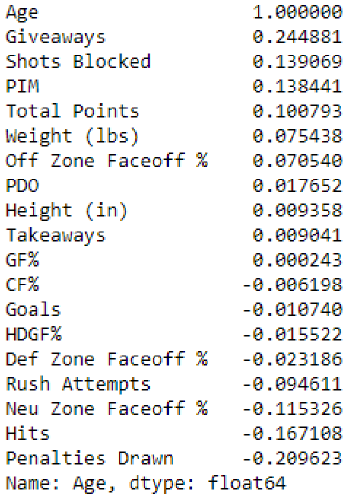
Czech player estimated correlation coefficients



```
Height (in)            1.000000
Weight (lbs)           0.775092
Def Zone Faceoff %     0.240073
Shots Blocked          0.203742
PIM                    0.176604
Hits                   0.144477
Neu Zone Faceoff %     0.118747
Age                    0.009358
Giveaways             -0.083438
PDO                   -0.104040
Takeaways             -0.138150
Penalties Drawn       -0.149980
HDGF%                 -0.180907
Rush Attempts         -0.206009
GF%                   -0.272740
Goals                 -0.281604
Total Points          -0.285624
CF%                   -0.297635
Off Zone Faceoff %    -0.307699
Name: Height (in), dtype: float64
```

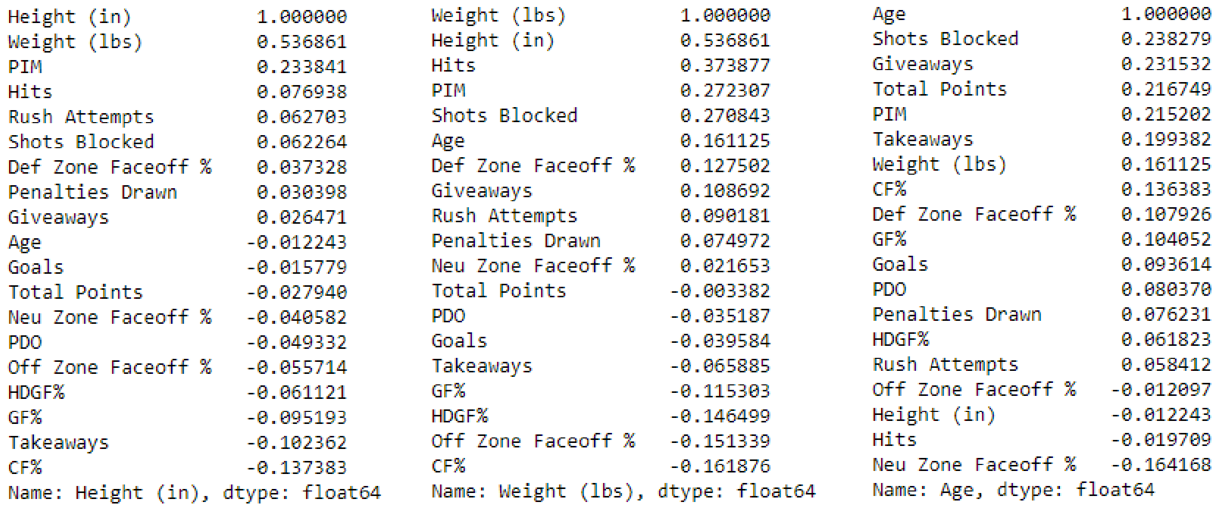


```
Weight (lbs)           1.000000
Height (in)            0.775092
PIM                    0.290760
Shots Blocked          0.276547
Hits                   0.257583
Def Zone Faceoff %     0.238716
Neu Zone Faceoff %     0.131499
Age                    0.075438
Giveaways             -0.028576
Penalties Drawn       -0.103385
PDO                   -0.107854
HDGF%                 -0.174856
Takeaways             -0.180248
Rush Attempts         -0.244081
GF%                   -0.249500
Total Points          -0.271763
CF%                   -0.280603
Goals                 -0.284611
Off Zone Faceoff %    -0.311012
Name: Weight (lbs), dtype: float64
```



```
Age                    1.000000
Giveaways              0.244881
Shots Blocked          0.139069
PIM                    0.138441
Total Points           0.100793
Weight (lbs)           0.075438
Off Zone Faceoff %     0.070540
PDO                    0.017652
Height (in)            0.009358
Takeaways              0.009041
GF%                    0.000243
CF%                   -0.006198
Goals                 -0.010740
HDGF%                 -0.015522
Def Zone Faceoff %    -0.023186
Rush Attempts         -0.094611
Neu Zone Faceoff %    -0.115326
Hits                  -0.167108
Penalties Drawn       -0.209623
Name: Age, dtype: float64
```

Russian player estimated correlation coefficients

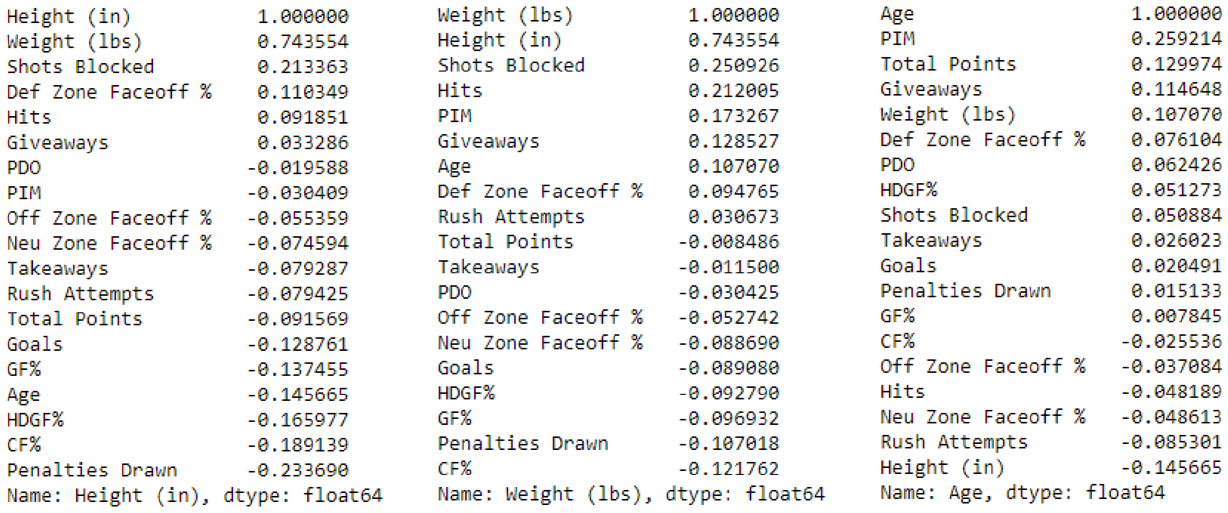


```
Height (in)              1.000000
Weight (lbs)             0.536861
PIM                      0.233841
Hits                     0.076938
Rush Attempts            0.062703
Shots Blocked            0.062264
Def Zone Faceoff %       0.037328
Penalties Drawn          0.030398
Giveaways                0.026471
Age                     -0.012243
Goals                   -0.015779
Total Points            -0.027940
Neu Zone Faceoff %      -0.040582
PDO                     -0.049332
Off Zone Faceoff %      -0.055714
HDGF%                   -0.061121
GF%                     -0.095193
Takeaways               -0.102362
CF%                     -0.137383
Name: Height (in), dtype: float64
```

```
Weight (lbs)             1.000000
Height (in)              0.536861
Hits                     0.373877
PIM                      0.272307
Shots Blocked            0.270843
Age                      0.161125
Def Zone Faceoff %       0.127502
Giveaways                0.108692
Rush Attempts            0.090181
Penalties Drawn          0.074972
Neu Zone Faceoff %       0.021653
Total Points            -0.003382
PDO                     -0.035187
Goals                   -0.039584
Takeaways               -0.065885
GF%                     -0.115303
HDGF%                   -0.146499
Off Zone Faceoff %      -0.151339
CF%                     -0.161876
Name: Weight (lbs), dtype: float64
```

```
Age                      1.000000
Shots Blocked            0.238279
Giveaways                0.231532
Total Points             0.216749
PIM                      0.215202
Takeaways                0.199382
Weight (lbs)             0.161125
CF%                      0.136383
Def Zone Faceoff %       0.107926
GF%                      0.104052
Goals                    0.093614
PDO                      0.080370
Penalties Drawn          0.076231
HDGF%                    0.061823
Rush Attempts            0.058412
Off Zone Faceoff %      -0.012097
Height (in)             -0.012243
Hits                    -0.019709
Neu Zone Faceoff %      -0.164168
Name: Age, dtype: float64
```

Finnish player estimated correlation coefficients

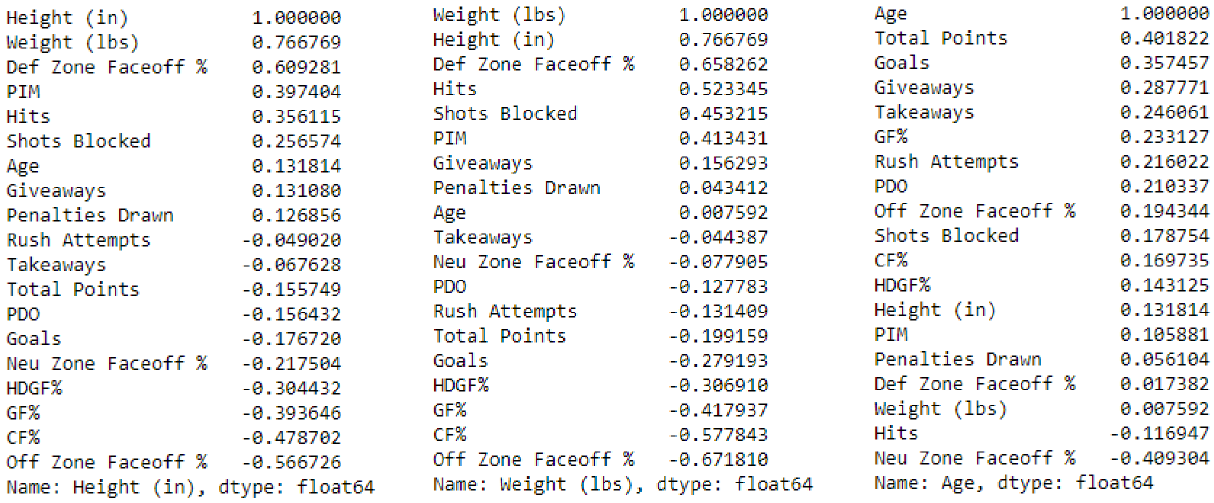


```
Height (in)              1.000000
Weight (lbs)             0.743554
Shots Blocked            0.213363
Def Zone Faceoff %       0.110349
Hits                     0.091851
Giveaways                0.033286
PDO                     -0.019588
PIM                     -0.030409
Off Zone Faceoff %      -0.055359
Neu Zone Faceoff %      -0.074594
Takeaways               -0.079287
Rush Attempts           -0.079425
Total Points            -0.091569
Goals                   -0.128761
GF%                     -0.137455
Age                     -0.145665
HDGF%                   -0.165977
CF%                     -0.189139
Penalties Drawn         -0.233690
Name: Height (in), dtype: float64
```

```
Weight (lbs)             1.000000
Height (in)              0.743554
Shots Blocked            0.250926
Hits                     0.212005
PIM                      0.173267
Giveaways                0.128527
Age                      0.107070
Def Zone Faceoff %       0.094765
Rush Attempts            0.030673
Total Points            -0.008486
Takeaways               -0.011500
PDO                     -0.030425
Off Zone Faceoff %      -0.052742
Neu Zone Faceoff %      -0.088690
Goals                   -0.089080
HDGF%                   -0.092790
GF%                     -0.096932
Penalties Drawn         -0.107018
CF%                     -0.121762
Name: Weight (lbs), dtype: float64
```

```
Age                      1.000000
PIM                      0.259214
Total Points             0.129974
Giveaways                0.114648
Weight (lbs)             0.107070
Def Zone Faceoff %       0.076104
PDO                      0.062426
HDGF%                    0.051273
Shots Blocked            0.050884
Takeaways                0.026023
Goals                    0.020491
Penalties Drawn          0.015133
GF%                      0.007845
CF%                     -0.025536
Off Zone Faceoff %      -0.037084
Hits                    -0.048189
Neu Zone Faceoff %      -0.048613
Rush Attempts           -0.085301
Height (in)             -0.145665
Name: Age, dtype: float64
```

Slovakian player estimated correlation coefficients



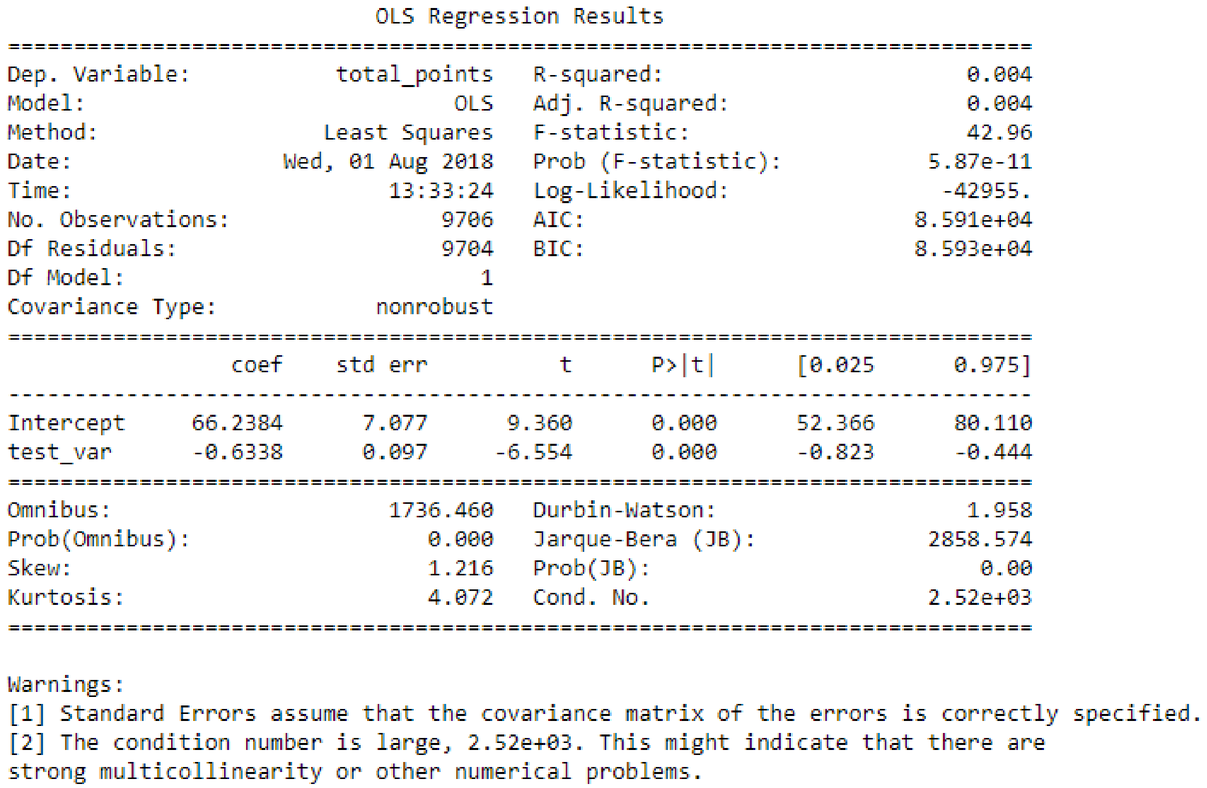
```
Height (in)            1.000000
Weight (lbs)           0.766769
Def Zone Faceoff %     0.609281
PIM                    0.397404
Hits                   0.356115
Shots Blocked          0.256574
Age                    0.131814
Giveaways              0.131080
Penalties Drawn        0.126856
Rush Attempts         -0.049020
Takeaways             -0.067628
Total Points          -0.155749
PDO                   -0.156432
Goals                 -0.176720
Neu Zone Faceoff %    -0.217504
HDGF%                 -0.304432
GF%                   -0.393646
CF%                   -0.478702
Off Zone Faceoff %    -0.566726
Name: Height (in), dtype: float64
```

```
Weight (lbs)           1.000000
Height (in)            0.766769
Def Zone Faceoff %     0.658262
Hits                   0.523345
Shots Blocked          0.453215
PIM                    0.413431
Giveaways              0.156293
Penalties Drawn        0.043412
Age                    0.007592
Takeaways             -0.044387
Neu Zone Faceoff %    -0.077905
PDO                   -0.127783
Rush Attempts         -0.131409
Total Points          -0.199159
Goals                 -0.279193
HDGF%                 -0.306910
GF%                   -0.417937
CF%                   -0.577843
Off Zone Faceoff %    -0.671810
Name: Weight (lbs), dtype: float64
```

```
Age                    1.000000
Total Points           0.401822
Goals                  0.357457
Giveaways              0.287771
Takeaways              0.246061
GF%                    0.233127
Rush Attempts          0.216022
PDO                    0.210337
Off Zone Faceoff %     0.194344
Shots Blocked          0.178754
CF%                    0.169735
HDGF%                  0.143125
Height (in)            0.131814
PIM                    0.105881
Penalties Drawn        0.056104
Def Zone Faceoff %     0.017382
Weight (lbs)           0.007592
Hits                  -0.116947
Neu Zone Faceoff %    -0.409304
Name: Age, dtype: float64
```

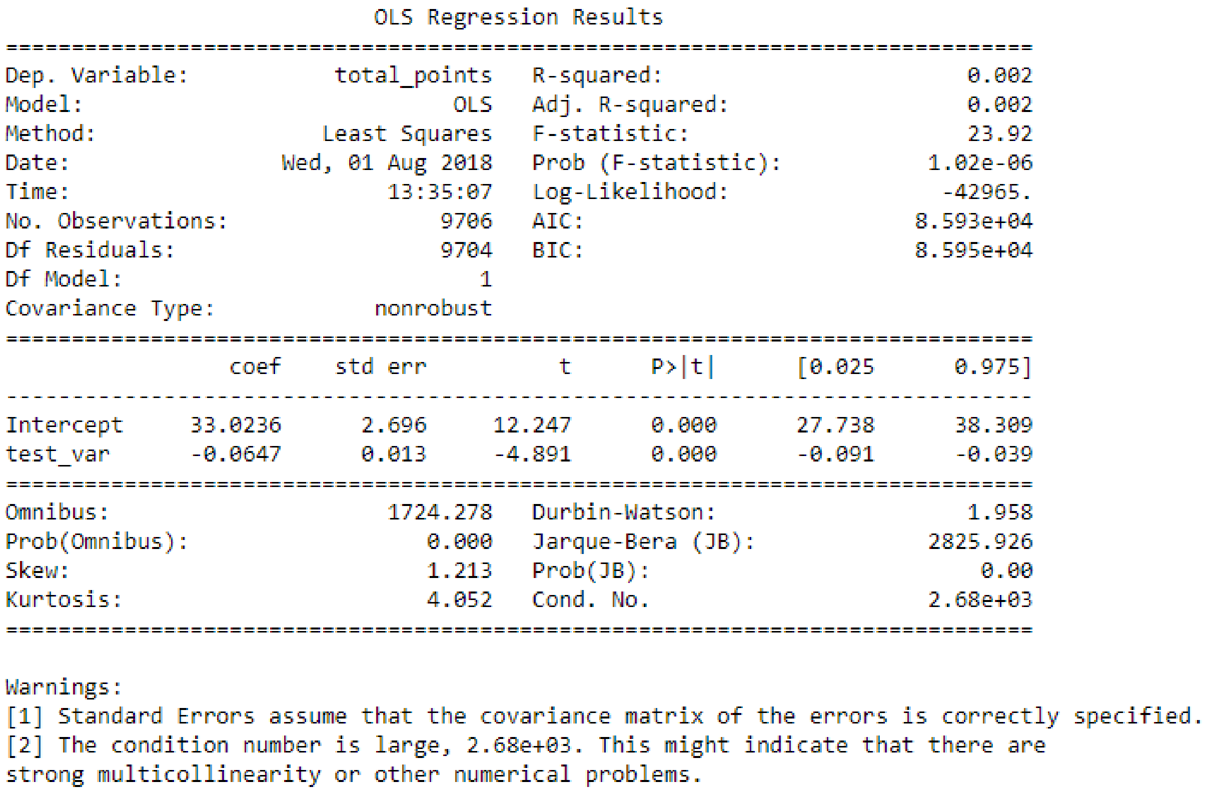
Appendix F

Linear Regression Summary Reports



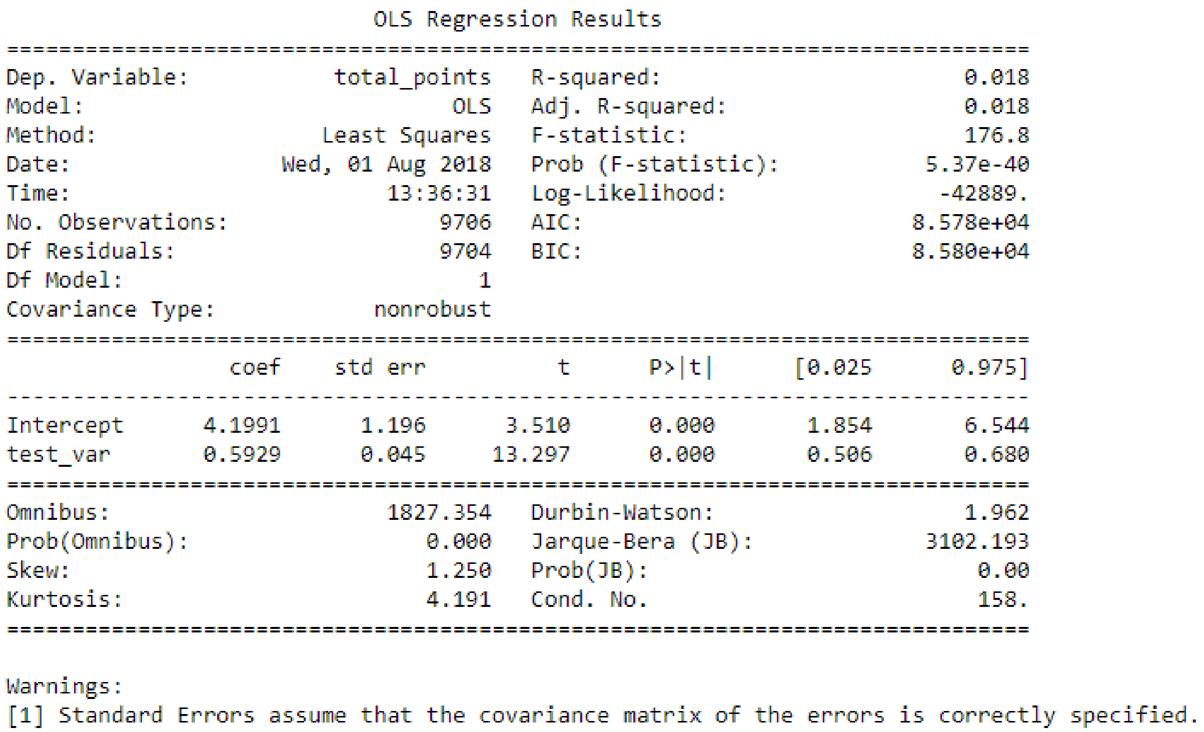
```
                            OLS Regression Results
==============================================================================
Dep. Variable:           total_points   R-squared:                       0.004
Model:                            OLS   Adj. R-squared:                  0.004
Method:                 Least Squares   F-statistic:                     42.96
Date:                Wed, 01 Aug 2018   Prob (F-statistic):           5.87e-11
Time:                        13:33:24   Log-Likelihood:                -42955.
No. Observations:                9706   AIC:                         8.591e+04
Df Residuals:                    9704   BIC:                         8.593e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     66.2384      7.077      9.360      0.000      52.366      80.110
test_var      -0.6338      0.097     -6.554      0.000      -0.823      -0.444
==============================================================================
Omnibus:                     1736.460   Durbin-Watson:                   1.958
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2858.574
Skew:                           1.216   Prob(JB):                         0.00
Kurtosis:                       4.072   Cond. No.                     2.52e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.52e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Summary of linear regression using only height



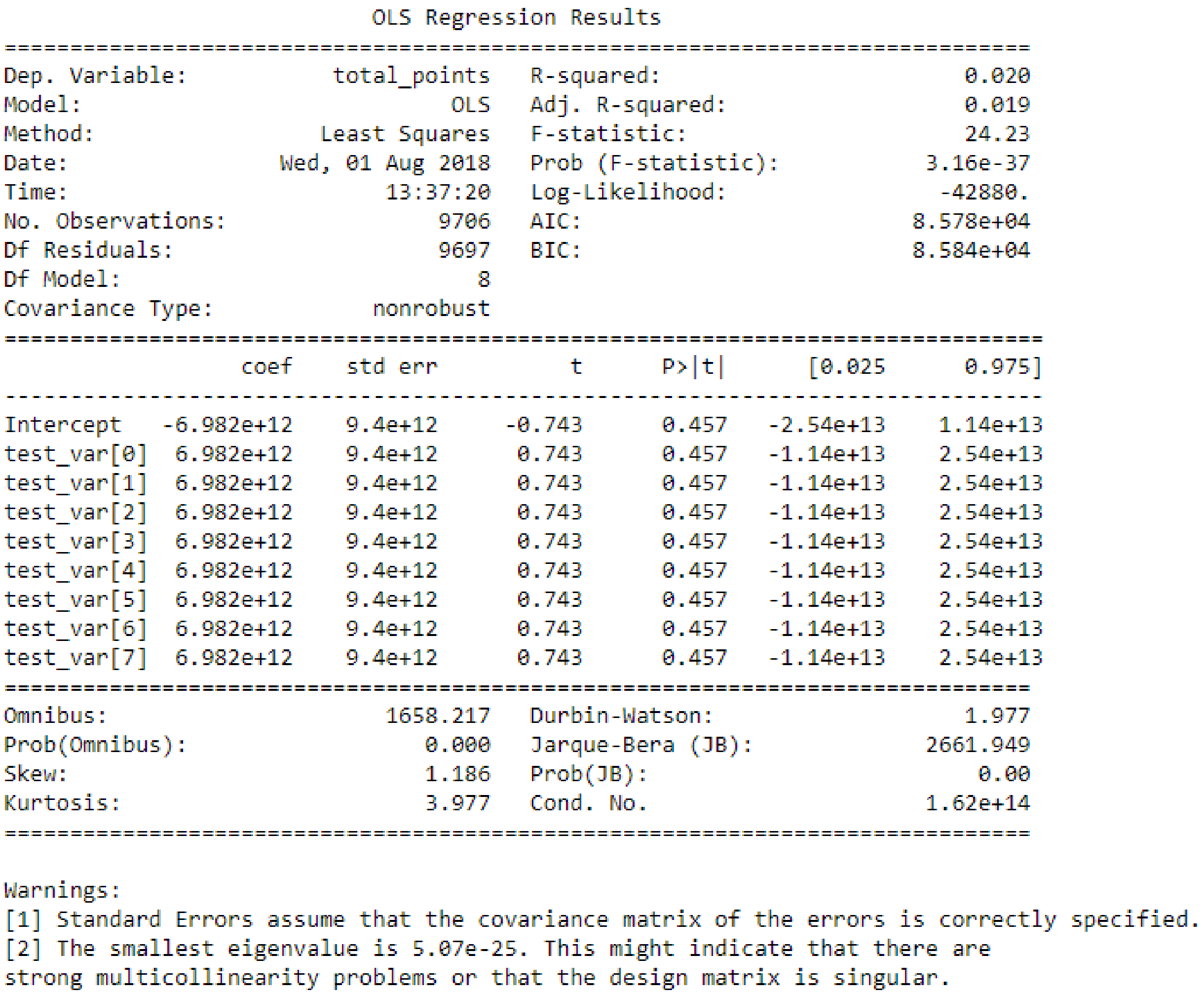
```
                            OLS Regression Results
==============================================================================
Dep. Variable:           total_points   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     23.92
Date:                Wed, 01 Aug 2018   Prob (F-statistic):           1.02e-06
Time:                        13:35:07   Log-Likelihood:                -42965.
No. Observations:                9706   AIC:                         8.593e+04
Df Residuals:                    9704   BIC:                         8.595e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     33.0236      2.696     12.247      0.000      27.738      38.309
test_var      -0.0647      0.013     -4.891      0.000      -0.091      -0.039
==============================================================================
Omnibus:                     1724.278   Durbin-Watson:                   1.958
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2825.926
Skew:                           1.213   Prob(JB):                         0.00
Kurtosis:                       4.052   Cond. No.                     2.68e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.68e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Summary of linear regression using only weight



```
                            OLS Regression Results
==============================================================================
Dep. Variable:           total_points   R-squared:                       0.018
Model:                            OLS   Adj. R-squared:                  0.018
Method:                 Least Squares   F-statistic:                     176.8
Date:                Wed, 01 Aug 2018   Prob (F-statistic):           5.37e-40
Time:                        13:36:31   Log-Likelihood:                -42889.
No. Observations:                9706   AIC:                         8.578e+04
Df Residuals:                    9704   BIC:                         8.580e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      4.1991      1.196      3.510      0.000       1.854       6.544
test_var       0.5929      0.045     13.297      0.000       0.506       0.680
==============================================================================
Omnibus:                     1827.354   Durbin-Watson:                   1.962
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3102.193
Skew:                           1.250   Prob(JB):                         0.00
Kurtosis:                       4.191   Cond. No.                         158.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Summary of linear regression using only age



```
                            OLS Regression Results
==============================================================================
Dep. Variable:           total_points   R-squared:                       0.020
Model:                            OLS   Adj. R-squared:                  0.019
Method:                 Least Squares   F-statistic:                     24.23
Date:                Wed, 01 Aug 2018   Prob (F-statistic):           3.16e-37
Time:                        13:37:20   Log-Likelihood:                -42880.
No. Observations:                9706   AIC:                         8.578e+04
Df Residuals:                    9697   BIC:                         8.584e+04
Df Model:                           8
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept   -6.982e+12    9.4e+12     -0.743      0.457   -2.54e+13    1.14e+13
test_var[0]  6.982e+12    9.4e+12      0.743      0.457   -1.14e+13    2.54e+13
test_var[1]  6.982e+12    9.4e+12      0.743      0.457   -1.14e+13    2.54e+13
test_var[2]  6.982e+12    9.4e+12      0.743      0.457   -1.14e+13    2.54e+13
test_var[3]  6.982e+12    9.4e+12      0.743      0.457   -1.14e+13    2.54e+13
test_var[4]  6.982e+12    9.4e+12      0.743      0.457   -1.14e+13    2.54e+13
test_var[5]  6.982e+12    9.4e+12      0.743      0.457   -1.14e+13    2.54e+13
test_var[6]  6.982e+12    9.4e+12      0.743      0.457   -1.14e+13    2.54e+13
test_var[7]  6.982e+12    9.4e+12      0.743      0.457   -1.14e+13    2.54e+13
==============================================================================
Omnibus:                     1658.217   Durbin-Watson:                   1.977
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2661.949
Skew:                           1.186   Prob(JB):                         0.00
Kurtosis:                       3.977   Cond. No.                     1.62e+14
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 5.07e-25. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Summary of linear regression using only nationality