Ice Around the World

Executive Summary

In this experiment, we will look into the National Hockey League (NHL) to find out if a certain body type is associated with a country, a position, or both. To do this, we will start by collecting our dataset which will include six attributes from each player in the NHL. In total, our hockey dataset will contain 741 instances and have six attributes. The attributes that will be collected include a player's name, position, shooting hand, height, weight, and country (Teams). Once all of our data has been collected, we can start our experiment. For the experiment, we will use the k-means clustering algorithm to cluster the data into clusters, based on the player's heights and weights, and then overlay our results with the countries, followed by the positions, and then finally the countries and positions. K-means clustering partitions the dataset into a k number of clusters that can be user controlled. To find these clusters, the k-means algorithm selects centroids (cluster means) from the dataset and then partitions like data into the cluster with the closest centroid. The clustering steps continue until no points change clusters. In our experiment, we will start by clustering the data into one to eight clusters to figure out how many clusters are within our dataset since we do not know the true number of clusters. In order to figure out the best number of clusters, we plot the convergence measurement from one to eight clusters and then use the elbow method to make our decision (Cluster Analysis). Along with the elbow method we will look at a 3D plot of the clusters to see if it will help us determine the correct number of clusters. When using the elbow method, we look at our convergence graph to figure out where it begins to plateau. The plateau in our convergence graph means the percentage of variance of our convergence begins to level off. Although, the elbow method is not the only method we should study when analyzing the convergence measurements graph, since it is based on human analysis and does not give a definitive answer (Gove). It gives a good representation of the approximate number of clusters that should be used, but it is recommended to verify the findings with another method of choice--we used a 3D plot. The 3D plot can visually show if the correct amount of clusters are there or if there should be more or less clusters in the dataset. Instead of just looking at one graph to determine the number of clusters, we will

go through the k-means algorithm and plotting the convergence measurements for all eight clusters a number of times and see what number of clusters is appropriate through multiple experiments. After going through the k-means clustering algorithm, we will use tables to overlay our data and see if there are relationships between body types, countries, and positions.

After conducting the experiment, we found that there are six clusters in the hockey dataset. We found that each cluster was based on the weight of the players and not their heights. The tables that we made showed no distinct relationships between body types, countries, and/or positions within the NHL. The two countries that the most players come from are Canada and the United States, but neither country dominated a certain body type or position. In the beginning, our assumptions were that there was not going to be a relationship between body types and countries, but that their might be one between body types and positions, and body types, country, and position.

Problem Description

The experiment we are conducting will try to find a pattern in our hockey dataset made up of attributes from every hockey player in the NHL. We will use the k-means algorithm to cluster our dataset into a number of clusters that will be determined from the experiment itself. To determine the number of clusters in our dataset, we will look at the plot of cluster convergence measurements for one to eight clusters. The elbow method and a 3D plot are the methods that will be used in selecting the correct number of clusters for our dataset. From there, we will overlay our clusters with the countries and/or positions to try to find relationships among body types, countries, and positions.

Analysis Technique

In this experiment, we will use the k-means clustering algorithm to cluster our hockey dataset into clusters based on body type. We collected our own data based on information from the official website of the NHL to obtain the attributes of all of the current NHL

players. Our dataset is called the hockey dataset and has 741 occurrences with six attributes. The attributes that we obtained include a player's name, position, shooting or catching hand, height, weight, and country. When obtaining the data, we kept track of the names of each player for ease of keeping track of which player we were on, but this information will not be used within the experiment itself. The only attributes that will be used to figure out a player's body type is his height and weight. We will conduct the k-means clustering algorithm for one to eight clusters, which will allows us to analyze a large range of clusters in order to figure out the appropriate amount of clusters for our hockey data. Our technique for narrowing down our clusters is called the elbow method. Once we have narrowed down the selection of clusters to two clusters, we will construct 3D plots for both clusters to determine the correct number of clusters within the hockey dataset. From here, we will overlay our clusters with the countries of origin to find a relationship between body types and countries. Once we have determined if there is a relationship between the body types and countries, we will then overlay the data with the positions. This will show us if there is a connection between body types, countries, and position. From this information we can analyze if there are certain players with certain body types that come from common countries to play a position. We expect that there will be an association between body types and positions, but not by country. Although, there is a possibility that there will be a relationship between a player's body type, position, and country.

Results

When looking at our convergence measurement graph and 3D plots (Figure 1 and 2, respectively), we were able to determine that there are six clusters within our dataset. When looking into each cluster, we found that they were mainly clustered based on weight rather than height.

Cluster	Range of Heights	Average Weight
1	5’10-6’6	207.8494
2	5’9-6’4	191.4331
3	5’10-6’6	199.3939
4	6’1-6’9	233.0222
5	5’11-6’7	217.8462
6	5’7-6’4	179.4054

Our results were mostly inconclusive when overlaying the six clusters with the countries or the positions. For a few clusters, we found there were some relationships between body types, positions, and countries, but nothing significant. Due to the clusters being partitioned based solely on weight, it makes it difficult to put one body type to one cluster. In the end, we can conclude that there is no true relationship between body types, countries, and positions in the NHL because no certain body type is needed to play any position in the NHL.

Although, there are a few relationships within the clusters. The few relationships that we found within the data are listed below:

- 30.72% of Americans are from Cluster 1
- 46.67% of players in Cluster 4 are defensemen
- 34.39% of players in Cluster 2 are centers
- 40.96% of players in Cluster 1 are defensemen
- 27.84% of centers are in Cluster 2
- 28.81% of defensemen are in Cluster 1
- 33.66% of Canadian centers are from Cluster 2

- 28.83% of Canadian defensemen are from Cluster 1
- 33.33% of American defensemen are from Cluster 1
- 52.06% of centers come from Canada
- 47.03% of defensemen come from Canada
- 41.79% of goalies come from Canada
- 51.16% of left wing forwards come from Canada
- 42.61% of right wing forwards come from Canada

When broken down, nothing above is significant enough to find any relationships between body types, countries, and positions.

Body Type by Country

	1	2	3	4	5	6
AUT	1	0	1	0	1	2
CAN	80	79	61	19	68	48
CHE	1	2	2	0	0	1
CZE	6	7	3	3	3	4
DEU	0	0	2	1	2	2
DNK	2	2	0	1	0	2
EST	1	0	0	0	0	0
FIN	5	8	4	0	7	8
FRA	1	1	2	0	0	0
ITA	1	0	0	0	0	0
KAZ	0	0	1	0	0	0
LVA	0	0	1	0	0	0
NOR	0	0	0	0	1	1
RUS	8	8	5	3	4	5
SVK	4	1	2	1	0	2
SVN	0	0	0	0	1	0
SWE	9	19	18	3	10	7
USA	47	30	30	14	33	29

Body Type by Position

	1	2	3	4	5	6
C	35	54	43	4	27	31
D	68	40	41	21	51	15
G	12	12	10	7	13	13
LW	30	27	23	9	15	25
RW	21	24	15	4	24	27

Body Type by Country and Position

Centers (C)

	1	2	3	4	5	6
AUT	0	0	0	0	0	1
CAN	19	34	19	2	13	14
CHE	0	0	0	0	0	1
CZE	2	1	3	1	1	0
DEU	0	0	0	0	1	0
DNK	1	1	0	0	0	0
EST	1	0	0	0	0	0
FIN	1	1	1	0	2	2
FRA	0	0	0	0	0	0
ITA	0	0	0	0	0	0
KAZ	0	0	0	0	0	0
LVA	0	0	1	0	0	0
NOR	0	0	0	0	0	0
RUS	1	2	2	0	0	2
SVK	0	0	0	0	0	0
SVN	0	0	0	0	1	0
SWE	2	6	6	0	3	2
USA	8	9	11	1	6	9

Defensemen (D)

	1	2	3	4	5	6
AUT	0	0	0	0	0	0
CAN	32	20	17	11	25	6
CHE	0	1	2	0	0	0
CZE	3	1	0	1	1	0
DEU	0	0	1	0	1	0
DNK	0	0	0	0	0	1
EST	0	0	0	0	0	0
FIN	1	2	1	0	2	1
FRA	1	1	0	0	0	0
ITA	1	0	0	0	0	0
KAZ	0	0	0	0	0	0
LVA	0	0	0	0	0	0
NOR	0	0	0	0	0	0
RUS	3	2	2	2	2	0
SVK	1	0	1	1	0	0
SVN	0	0	0	0	0	0
SWE	4	6	5	0	6	2
USA	22	7	12	6	14	5

Goalies (G)

	1	2	3	4	5	6
AUT	0	0	0	0	0	0
CAN	4	5	5	1	8	5
CHE	0	0	0	0	0	0
CZE	1	0	0	0	0	1
DEU	0	0	0	1	0	1
DNK	0	0	0	1	0	0
EST	0	0	0	0	0	0
FIN	0	2	0	0	2	2
FRA	0	0	0	0	0	0
ITA	0	0	0	0	0	0
KAZ	0	0	1	0	0	0
LVA	0	0	0	0	0	0
NOR	0	0	0	0	0	0
RUS	2	0	0	0	0	1
SVK	0	0	1	0	0	1
SVN	0	0	0	0	0	0
SWE	0	2	2	2	0	1
USA	5	3	1	2	3	1

Left Wing Forwards (LW)

	1	2	3	4	5	6
AUT	1	0	1	0	1	0
CAN	15	13	13	3	8	14
CHE	0	1	0	0	0	0
CZE	0	2	0	0	0	0
DEU	0	0	0	0	0	0
DNK	1	0	0	0	0	0
EST	0	0	0	0	0	0
FIN	0	2	1	0	0	2
FRA	0	0	2	0	0	0
ITA	0	0	0	0	0	0
KAZ	0	0	0	0	0	0
LVA	0	0	0	0	0	0
NOR	0	0	0	0	1	0
RUS	1	1	0	1	1	1
SVK	1	0	0	0	0	1
SVN	0	0	0	0	0	0
SWE	3	3	3	0	1	1
USA	8	5	3	5	3	6

Right Wing Forwards (RW)

	1	2	3	4	5	6
AUT	0	0	0	0	0	1
CAN	10	7	7	2	14	9
CHE	1	0	0	0	0	0
CZE	0	3	0	1	1	3
DEU	0	0	1	0	0	1
DNK	0	1	0	0	0	1
EST	0	0	0	0	0	0
FIN	3	1	1	0	1	1
FRA	0	0	0	0	0	0
ITA	0	0	0	0	0	0
KAZ	0	0	0	0	0	0
LVA	0	0	0	0	0	0
NOR	0	0	0	0	0	1
RUS	1	3	1	0	1	1
SVK	2	1	0	0	0	0
SVN	0	0	0	0	0	0
SWE	0	2	2	1	0	1
USA	4	6	3	0	7	8

Issues

Since all of the players within the NHL, no matter the position, are all within a small range of body types, it made it difficult to separate the data into significant clusters. We believe that is why the clusters were based mainly on the weight of the player instead of his height. After reviewing the data, the NHL is probably one of the least diverse leagues in all of professional sports.

Appendices

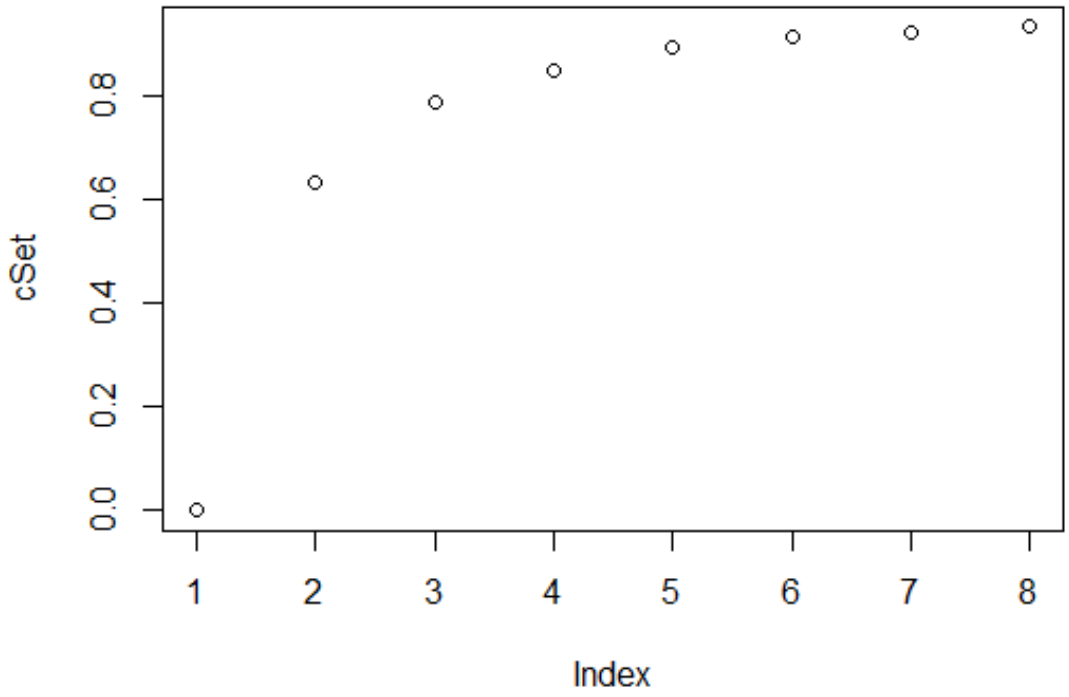

Figure 1-Convergence Measurements (index is the number of clusters used in the k-means algorithm)

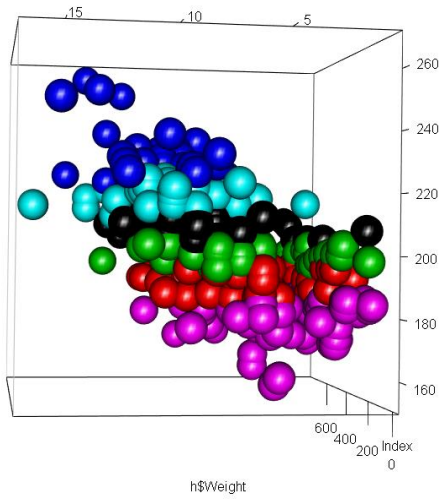

Figure 2-3D Plot of Clusters

References

Cluster Analysis: Basic Concepts and Algorithms. (n.d.). Retrieved from https://www.bing.com/cr?IG=50515960EDD742D48D7B9F02F77A927A&CID=35969EA9A9256AC0052F9741A8146BD7&rd=1&h=X75Mhp4mTdZCWPW00ywIrIWKxZtfgbssdyqBc5BjRao&v=1&r=https://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf&p=DevEx,5082.1

Gove, R. (2015, December 3). *Using the Elbow Method to Determine the Number of Clusters for K-Means Clsutering*. Retrieved from Robert Gove's Block: https://bl.ocks.org/rpgove/0060ff3b656618e9136b

Teams. (n.d). Retrieved from https://www.nhl.com/info/teams