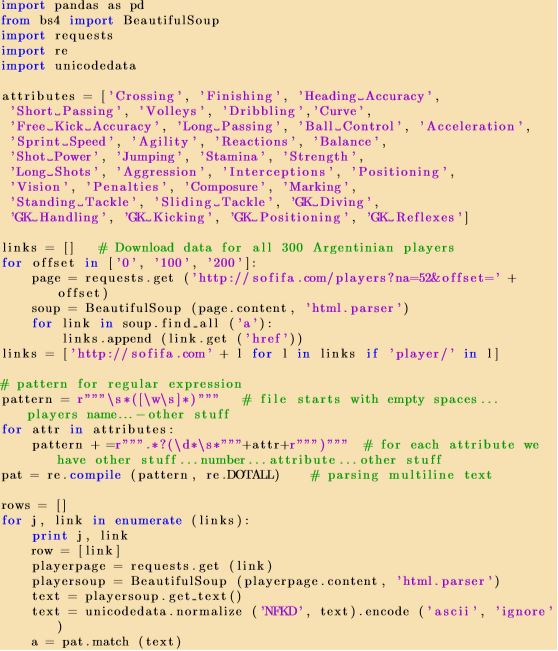
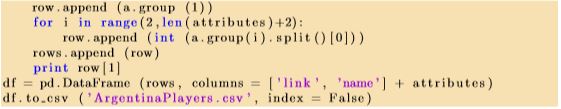
**Machine Learning- Clustering**

For this assignment, I am going to ‘scrape’ data on Argentinian soccer players from the EA sports FIFA games website.



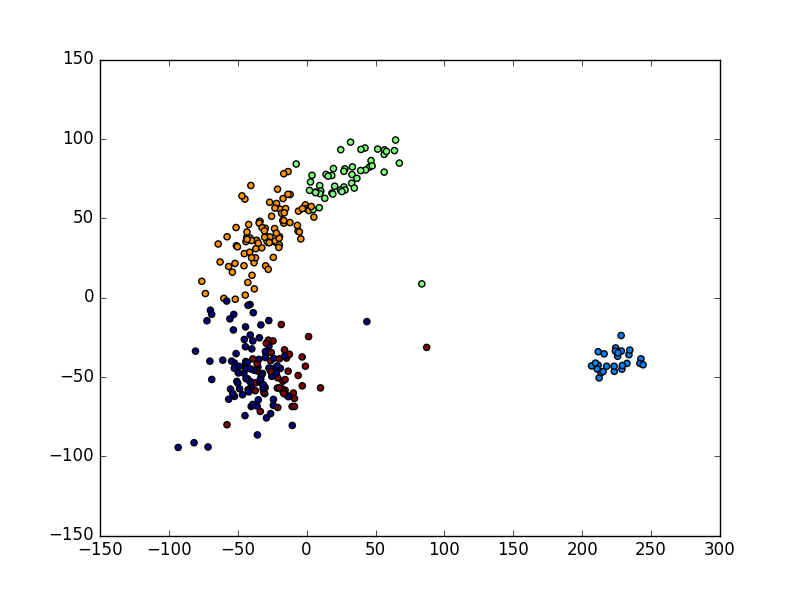


The code above does the following:

1. Creates a list of the attributes that each of our players will have later on (this will be populated from the website.
2. There is a number of steps followed in order to obtain the links for the 300 Argentinian players. In particular, we first get the basic website http://sofifa.com/players?na=52&oﬀset and decode it, by using the ‘requests.get’ command and by setting offset equal to 0, 100 and 200 respectively. When offset equals 0 we get access to the first web page which contains the first 100 players’ information, whereas when the offset equals 100 and 200 we get access to the next 100 players in the second and third web page respectively.
3. The HTML files are parsed by using the BeautifulSoup command (“html parser”) in order to be able to extract the players’ information. Looping through the player’s information, we extract all href attributes from player’s pages and then add them to ‘http://sofifa.com' address in order to obtain the final links of the 300 players and saved them in our ‘links’ list as long as ‘player/’ is included in our initial ‘links’ list (obtained through the loop).
4. Set pattern as the player’s name and then we add the objects of our initial ‘attributes’ list to the pattern. This procedure uses regular expression.
5. We parse the multiline text by using the ‘re.compile’ command. The last loop of our code focuses on obtaining the information - text we need for the attributes of each of our players. We again decode the contents of the web pages, parse the html file and obtain the text from it.
6. Finally, we create a data frame containing the link, name and all attributes of each of the 300 Argentinian players. More detailed comments can be found in the ‘footballscrape\_q2.py’ file along with the code.

**1.Used the sklearn.cluster.KMeans Python class to cluster the players into 5 clusters.**

We load our data obtained from question two and apply the k-means clustering method by using the sklearn.cluster.KMeans and sklearn.cluster.KMeans.fit commands. Here we should note that before applying the k-means method, we first remove all columns that contain text data. Finally, we get the labels of the clusters – 0, 1, 2, 3, 4 – and create a new data frame, called ‘clusters\_table’, which contains all of our initial data but including also a column with the cluster of each player. The plot obtained from our clustering analysis can be seen below:



**2. By inspecting the clusters and looking up individual players online, I assigned meaningful labels to the clusters.**

I came up with each cluster definition as follows:

*Cluster 0*

Players in this cluster have high average value for Agility, Acceleration, Balance, Dribbling, Ball Control, and Sprint Speed (more than 75). These types of attributes are necessary to penetrate enemy’s defense using dribbling and acceleration. And by looking at the players’ name - Lionel Messi, Sergio Aguero, Angel Di Maria and Paulo Dybala - we can assign this cluster as an ‘attacking midfielder/ forward’.

*Cluster 1*

Players in this cluster have high average value for GK Reflexes, GK Diving, GK Positioning, GK Handling, Reactions, and GK Kicking (more than 70). Those attributes are crucial for a goalie to anticipate a shot from an attacker or a dangerous crossing in his penalty area. And by looking at players’ names - Sergio Romero, Willy Cabalero and Oscar Ustari - we can assign this cluster as Goalkeeper.

*Cluster 2*

Players in this cluster have high average value for Strength, Aggression, Standing Tackle, Interceptions, Marking, Heading Accuracy, and Sliding Tackle (more than 72). These types of attributes are critical to not only win one-on-one battle but also reading attacker movement to steal his ball possession. These players also occasionally helped to score goal in set-piece situation since they have good heading accuracy. And by looking at the players’ name - Ezequiel Garay, Martin Demichelis, and Gonzalo Rodriguez - we can assign this cluster as ‘Central Defender or Defensive Midfilder’.

*Cluster 3*

Players in this cluster have high average value for Stamina, Aggression, Standing Tackle, Reactions, Short Passing and Interceptions (more than 72). These attributes are important since the players act as first line of defense and involve in ball distributions to other players. And by looking at the names - Javier Mascherano, Esteban Cambiasso, Pablo Zabaleta and Marcos Rojo - we can assign this cluster as ‘Fullback/Wingback’. J Mascherano is somewhat an anomaly here as he played as a central defender for Barcelona on some occasions, but there is not much defending to do when you are Barcelona, hence he was able to fit in.

*Cluster 4*

Players in this cluster have high average value for Positioning, Shot Power, Heading Accuracy, Finishing, and Ball Control (more than 72). These attributes are decisive criteria to create a goal either from live situation or set-piece like corner kick or free kick. They also can make chances for other team mates to score. And looking from the names - Gonzalo Higuain, Mauro Icardi, and Nicolas Biandi - we can assign this cluster as ‘Striker’.

In addition, by looking at these clustering results, we can suggest that Argentine National Team lacks player who has good dribbling combined with short passing which usually found in central midfield and acts as a playmaker (similar to Xavi, Fabregas, etc.). That is why Argentina usually relies on Lionel Messi to play in the middle since they have lots of winger players and he usually assigned as attacking midfielder and playmaker for his national team, which is somewhat different to his position for Barcelona.

**3. For a new and unknown player, the following attributes are available:**

|  |  |
| --- | --- |
| **Attribute** | **Value** |
| Crossing | 45 |
| Sprint Speed | 40 |
| Long Shots | 35 |
| Aggression | 45 |
| Marking | 60 |
| Finishing | 40 |
| GK Handling | 15 |

**For each of the 5 clusters from Step 4, computed the cluster centroid. Assigned the new player to the nearest cluster based on the distance to the cluster centroids, using only the available attributes.**

Since we have 34 attributes from the dataset and we have 5 different clusters, therefore each cluster centroid is given by a vector comprises of 34 values representing each of the attribute. However, we only have 7 attributes for the new player, therefore the centroids for the clusters are calculated based on just the 7 attributes.

Next step is to calculate Euclidian distances between centroids and the new player- we do iteration to look for every attribute in the vector containing new player’s attributes and take power of 2 from the difference between this value with respective attribute in the cluster centroid. Then we sum this result. Finally, the distance of new player from a cluster is given by the square root of this sum. We iterate this steps for all clusters and assign the new players to the minimum distance. In this case, the shortest distance is to cluster 2.