Summary

Task: The objective was to extract data from Icons.com to determine which footballer's signature holds the highest value among current stars with surnames starting from A to C.

To find the most valuable signature, a seemingly straight-forward approach would be to calculate the average price of all products signed by each player. The average prices of each player's products would then be compared to yield the most valuable signature. However, product prices are determined by factors other than just the player's signature. They vary significantly based on product type, presentation style, and dimensions. Hence, this calls for a more comprehensive solution that takes into account the outlined complexities.

I explored two different approaches, each with its own advantages and limitations, depending on the use case.

Approach 1 (Most Accurate):

This approach finds the most common product across product type (photo, shirts, boots, etc), presentation type (framed, unframed, acrylic box, etc), and dimension of the frame/ photo. Then choosing that particular product, the prices are compared with each players' of the same selected item. This will result in an accurate representation of whose signature has more value.

Limitation: Not all players have the same product type, leading to data gaps. As a result, some players are excluded due to a lack of comparable data. A separate list of excluded players is provided in the final output.

Approach 2 (Simpler Alternative):

This method selects the highest-priced product for each player and compares it with others. It ensures that every player with at least one listed product is included in the comparison.

Limitation: The most expensive product might be costly due to additional factors such as framing or display costs rather than the signature itself.

Since the task was not specific on the use case, I have added both the methods and derived the results from that.

Tech Stack and choosing the appropriate scraping framework:

Initial thought after exploring the website was to go with selenium since lots of data were dynamically loaded. But upon further inspection most of the parts can be extracted using static webpage scrapers. So this solution uses selenium to get links to each products for all the players since those are loaded using JS; and then uses BeautifulSoup to gather product details to populate the DataFrames.

Though the whole project could have been achieved with selenium, using BeautifulSoup with requests where applicable is beneficial as it is much faster compared to selenium.

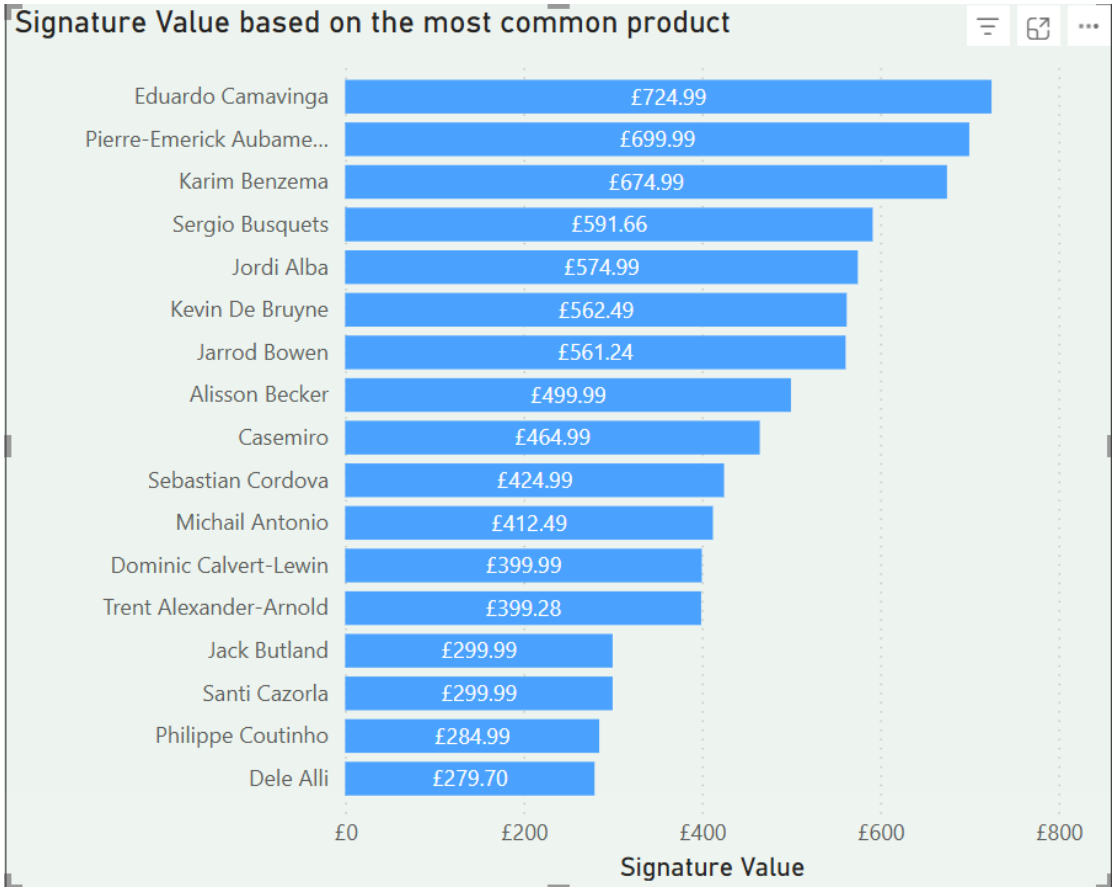
The final DataFrames are then converted to csv files as output.

The entire project is containerized using docker and the steps to run it locally is explained in the instructions.txt

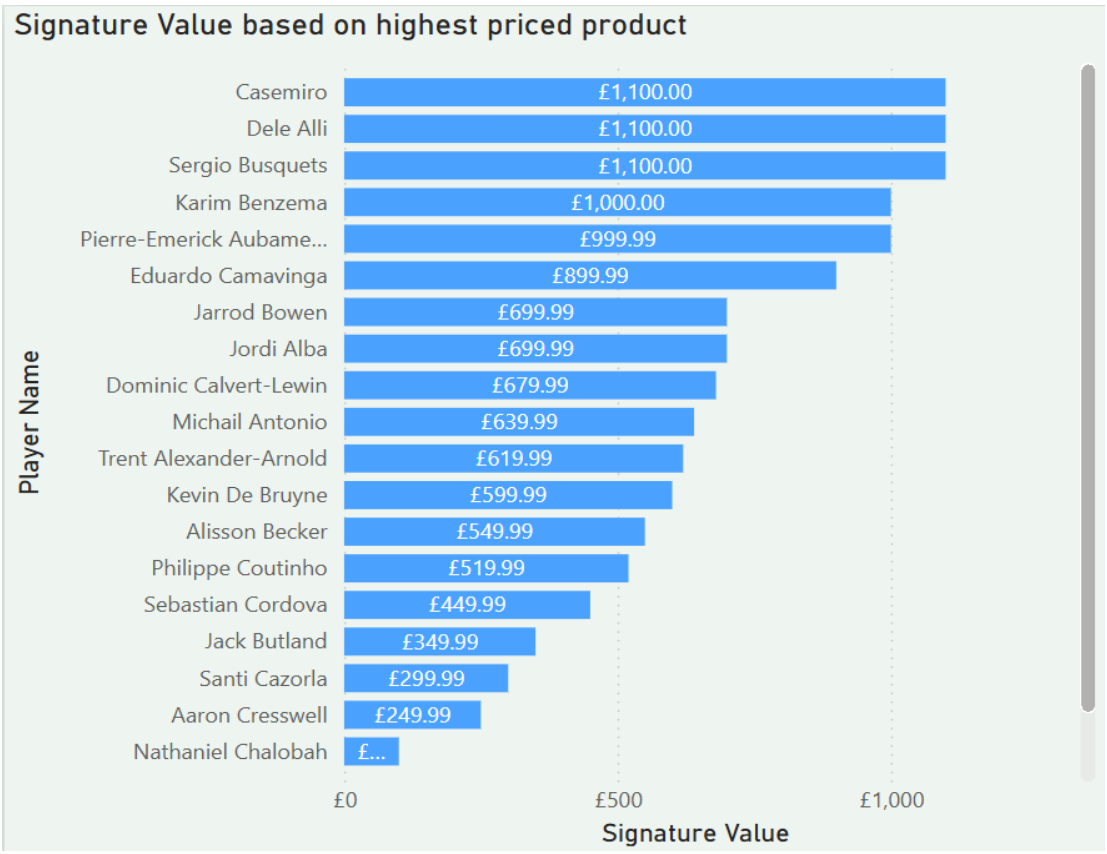
Results:

To determine the highest-valued signature fairly, Approach 1 identified **Eduardo Camavinga** as the player with the most valuable signature, based on a specific product (**framed signed shirt of dimension 885mm (H) x 780mm (W) x 30mm (D).**

The following visualization shows the list of players sorted by signature values according to approach 1.



For approach 2, which considers the highest priced product for each player, the results are displayed in the visualization below:



Additionally, because the project was super fun, I have also created an interactive dashboard on Power Bi for player portfolio analysis. The Power BI file (.pbix) is included in the project folder.