IST652 - Scripting for Data Analysis **4/26/2023**

REPORT – MINI PROJECT 2: **Semi-structured Data Processing**

Anvitha Koojugodu Shashidhar

Data And Source

The dataset for the project has been downloaded from Kaggle. The data set is in JSON format and consists of data which is related to football (soccer) player transfers in the summer window of 2022. The data set contains information about 2,344 transfers, including player name, age, position, nationality, new club, and cost.

Data Exploration and Data Cleaning

The first step is to import the required libraries and load the JSON file using the pandas library. 'player_valuje' column is renamed to 'player_value.' Defined a function to convert the player's value to a float, removing any euro or million/ thousand denomination.

The next step involves calculating several fundamental data, such as the most expensive player, the costliest club, and the number of players by their new club's and country of origin. Additionally, two tables are created to list the top 10 nations based on the number of players from those nations and the top 10 nations based on the number of players who arrive at each nation's club.

Two bar graphs are plotted that show the top 10 highest and lowest transfers in terms of their cost.

The last step involves computing and plotting the total transfer fees spent by the top 10 clubs.

Comparison Questions with Unit of Analysis

The following questions were selected to analyze the dataset in a view different from what is available in the dataset:

1. What are the total transfer fees spent by each league, i.e., Premier League, LaLiga, and Ligue 1, and how do they compare?
   The unit of analysis is League, comparison value is Total Transfer Fees Spent. It is computed using groupby 'league_new_club' and computing the sum of 'player_value'.

2. Which are the top 10 clubs by total transfer fees spent, and how do they compare?
   The unit of analysis is Club, comparison value is Total Transfer Fees Spent. It is computed using groupby 'new_club' and computing the sum of 'player_value'.
3. What are the top 10 countries by the count of players from their origin and the top 10 countries by the count of players arriving at their club, and how do they compare?
   The unit of analysis is Country and the comparison value is Count of Players. It is computed using groupby 'country_origin_club' and 'country_new_club' and computing the count of players.

Description of the Program

The JSON file is loaded into a pandas data frame once the program imports the relevant libraries. The dataframe is then cleaned up by renaming a misspelled column, player_valuje, to player_value. The player_value column is then converted from a string containing a currency value (in Euros) to a float using a custom function convert_value().
The program calculates certain fundamental statistics pertaining to player transfers, generates tables listing the number of players by their new club and country of origin, and shows bar graphs to contrast the top 10 highest and lowest transfers as well as the top 10 clubs in terms of total transfer fees paid.
Overall, the program reads in data on football transfers, cleans and processes it, computes summary statistics, and creates various visualizations to help explore and understand the data.

Description of the result of the analysis

The club with the greatest overall transfer expenditure is determined to be the club having spent the most on transfers during this time. Fans and experts who are interested in following the financial activity of different football clubs may find this information valuable.
The most expensive player transfer during this time is displayed which can provide light on the best players' current market value in the football industry. Football fans and business stakeholders who are interested in following player valuation and the

transfer market may find this information to be helpful.

The program generated tables that list the top 10 nations that players traveled to and from during this time. These tables break down where players are coming from and going to, which can offer insight into how football players migrate across borders and their migratory patterns.

Two bar plots are plotted to display the top 10 highest and lowest transfer fees for certain players. These stories can provide light on the transfer market worth of players and the tendency to overpay for top talent.

The top 10 clubs during this time period's total transfer fees are displayed in a bar plot. This plot sheds light on the financial dealings and spending habits of these elite clubs.