MIS 3335 – Data Analysis Using Python

Homework 5 – Data types, strings, and GroupBy

Your assigned task is to complete the data operations described below using pandas, Python, and the Jupyter Notebook. The data file for this assignment is named "`fifa19_ver1.csv`" and is available on Blackboard with the other course data. ***Submit your notebook at the appropriate link in Blackboard***. Do not submit the data file.

This dataset is a subset (17 columns) of individual player data from the video game FIFA 19. The complete dataset (89 columns) is used to generate the performance of video game characters to approximate the play of the real players. This dataset was created by someone who used Python to scrape the data from the game's website. More about the origins of the dataset is found at https://www.kaggle.com/karangadiya/fifa19.

The only change made to the original dataset from Kaggle.com is to remove columns. None of the data itself has been changed.[1]

Part 1

1. Read the data from the csv file with the ID column set to be the index value (we did this in class in the "Data Types" notebook when we loaded the Taiwan data). Store the data in a data frame object named "fifa."
2. Several of the data types will not be what we want when the fifa data frame is first created. Fix the data types and data as needed to produce the following results for `fifa.info()`.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 18207 entries, 158023 to 246269
Data columns (total 16 columns):
Name                        18207 non-null object
Age                         18207 non-null int64
Nationality                 18207 non-null object
Overall                     18207 non-null int64
Potential                   18207 non-null int64
Club                        17966 non-null object
Value                       18207 non-null object
Preferred Foot              18159 non-null category
International Reputation    18159 non-null category
Weak Foot                   18159 non-null category
Skill Moves                 18159 non-null category
Body Type                   18159 non-null object
Position                    18147 non-null object
Jersey Number               18207 non-null object
player_photo                18207 non-null object
wage_euros                  18207 non-null float64
dtypes: category(4), float64(1), int64(3), object(8)
memory usage: 2.5+ MB
```

[1] Okay, technically one cell value out of the 309,536 total cells was changed. But that's it.

To do this, you will need to do several things. These include (but are not limited to) the following:

- Change data types of some of the existing columns. Compare the data read from the file to the example to determine which ones.
- Even though it will be an object, make sure the jersey number looks like an integer. So, the value should look like "7," not "7.0." Hint: you will need to replace missing values with 0 to make it work.
- The "Photo" column contains a URL to the player's picture. Save just the name of the picture file itself. For example, the value for L. Messi in the original data is "https://cdn.sofifa.org/players/4/19/158023.png" but we want to trim it down to just "158023.png" when you are done *because photo file names are different lengths but every value has the same path characters*. Store the picture file name in a column named "player_photo" and drop the original "Photo" column.
- The "Wage" column currently contains problematic data: "€", a number, and a "K." Retain only the number as a float in a column named "wage_euros" and drop the original "Wage" column. Feel free to create an intermediate column if you need to as long as you drop it when you no longer need it.

Make sure you don't overlook anything that was not detailed above. Double-check your results against the example `fifa.info()` output on page 1.

Part 2

In this part of the assignment, you will do some aggregation using GroupBy on your cleaned data frame. Show the code and the results to the following questions. You don't have to isolate every answer. It is acceptable to show more results than necessary. [Except using 'describe()' as your answer is not acceptable.]

1. How many players are left footed?
2. Which clubs have the three highest weekly payroll totals?
3. What are the five most common jersey numbers?
4. Players from which country (i.e., "Nationality") have the lowest average age?

TIP: Remember that when you are looking at a single column of a data frame, it is a pandas Series. Any methods are applied to the column the same way they would be applied to a stand-alone series. Methods named nlargest() and nsmallest() will be very helpful here.

As always, figure out what needs to be done before you start doing it.

A description of the fields in this dataset is on the last page.

This is a brief description of the fields in the “fifa19_ver1.csv” data.

Field Name	Description
ID	A unique number identifying each player. Should be used as the index value when reading the data.
Name	The player’s name
Age	The player’s age in years
Photo	A URL to the player’s photo.
Nationality	The country the player is from.
Overall	A value from 1 to 100 representing the player’s overall skill level.
Potential	A value from 1 to 100 representing the player’s potential overall skill level.
Club	The name of the professional club currently employing the player.
Value	The overall market value of the player.
Wage	The weekly wage paid to the player.
Preferred Foot	The player’s dominant kicking foot.
International Reputation	A value from 1 to 5 (best) representing the player’s reputation.
Weak Foot	A value from 1 to 5 (best) representing the player’s skill level with their non-dominant foot.
Skill Moves	A value from 1 to 5 (best) representing the player’s proficiency with moves that require a high level of skill.
Body Type	A category of body type. Elite or unique players are their own body type.
Position	A text abbreviation of the player’s preferred position.
Jersey Number	The number the player wears on their jersey when playing for the club.