CS 6140 Project Data collection report

Harshitha Manduva U1080236
Sai Varun Addanki U1210409
Sachin Boban U1210407

Data details:

How did you obtain your data ?
We obtained our data from Kaggle. It has been scraped from a website https://sofifa.com/. It details attributes for every player registered in the latest edition of FIFA 19 database.

How large is your data
Our data in tabular format consists of 18000 rows each with 89 columns. The size of the csv file is ~9 MB.

Details about the data
Each row of the dataset represents a player and a thorough description of his abilities. Each player has a unique ID and each column represents attributes such as Age, Nationality, Club, Wages etc. We are even more interested in the non-categorical/numerical attributes of a player such as his speed, strength, passing accuracy etc.
Below is the sample cleaned data frame in the dataset

https://github.com/harshi-M/Mining-Project/blob/master/doc/SampleDataFrame.txt

Details about the data source
The data that we require, is the different aspects of various soccer players.
We obtained FIFA 19 soccer player ratings on Kaggle. Though FIFA 19 is a soccer simulation video game developed by EA, ratings within the game are considered as the best, even among professional soccer players. These ratings showcase the players in the game based on their performances from the past year of world soccer.

Data Cleaning Status:

In what format are you storing your data. Describe the abstract data type, not just the file format
The dataset (a csv file) was quite easy to process. We did have to work on some missing values or missing data which were replaced by a value -1.

Implementation wise we storing in the form of pandas dataframe. We could've simply kept it as a 2 dimensional array, but dataframe will be very useful and fast down the line for any kind of manipulation.

Prior to storing the data, we did some data cleaning to make the data uniform and following are the major changes:

- Columns such as the Joined Date, which represent the date when a player joined a club format was changed.
- The field height had units in feet which were converted to inches for uniformity
- Each field was given its own data type e.g. Age float, name string etc.
- The form of a player added a +2 value to the attributes, we separated it out to a different column
- The market Value of a player was mentioned in various currencies and denominations which was changed to a uniform integer format.

Did you need to process the original data to get it into an easier, more compressed format (e.g., convert from one format to another one)?
As the data in hand was appropriate, we did not compress the data.

How would you simulate similar data?
To simulate similar data, i.e. to add a new player, we could simply interpolate values of players of similar ability playing in similar clubs. This would be un-necessary for the most part as there are already 18000 data points in the dataset to work with.

- What would happen if our dataset was larger
  Currently we stored our datasets in the form of dataframes. Had our dataset been much larger, we would've had to sample our data to obtain representative values and used them instead of the whole data. A larger dataset would also mean changes to our implementation and we would've used a better suited data structure than dataframes.

- Underlying model?
  FIFA is a game that has been in business since the 90s. We believe EA sports, the author of the game and these attribute values has curated this data over 10 years, which each players score being set based on his score from last year. We believe there isn't a strong underlying mathematical model to the data. We would have to interpolate existing values to create new data points. That being said, currently the data is large enough. We are vary of changing the columns/attributes as each player is strong/weak on certain attributes that can set him apart from the rest of the players.