**Data Science Tools & Techniques**

Sir Farrukh Hassan – Fall 2020

Term Project Report

# **Dataset Description**

The chosen dataset has 9 CSV files in total containing statistics about various cricketers. The data recorded is from 1877 till 2019. There are 3 different files for different match formats and 3 files for each for different types of players.

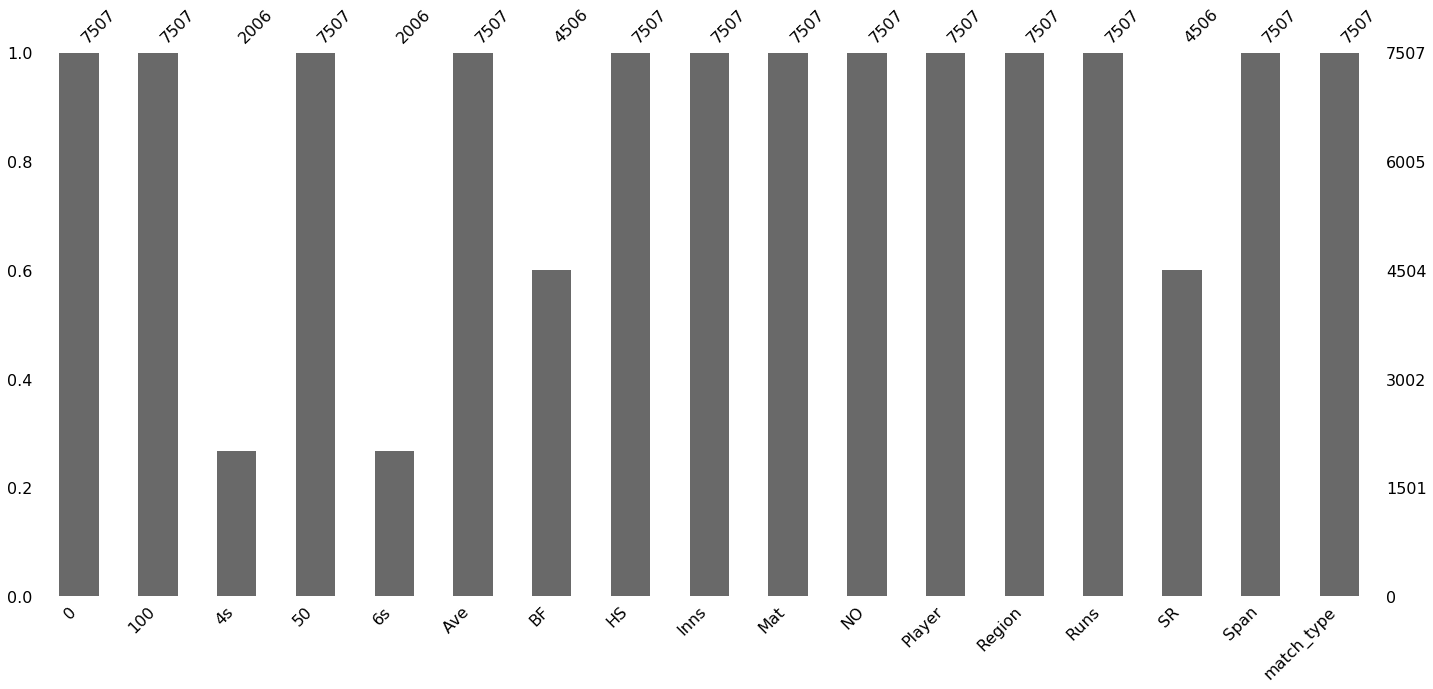
## Types of Players

1. Batter
2. Baller
3. Fielder

## Types of Players

1. ODI
2. Test
3. T20

# **Data Preprocessing**



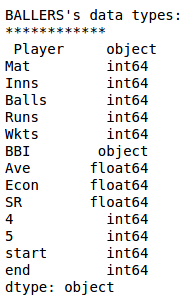
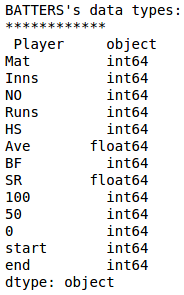
There are common attributes between all files, however, after merging all match formats into a single file for each player type, missing values were encountered. These missing values were result of a player not participating in a different match format.

As shown visually, in the above diagram for ballers dataset, after merging ODI, T20 and Test matches datasets, missing values were encountered.

Next, data-types for each dataset were analyzed, it was found that some numeric columns had numbers in quotations. This was corrected using custom code for each incorrect column. Also Span was spitted into two new columns that represented start and end of a cricketer’s career.

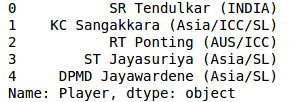
## Before correcting data types

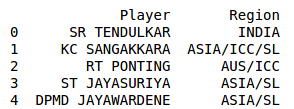
## After correcting data types



The fielders dataset had duplicates and were removed:



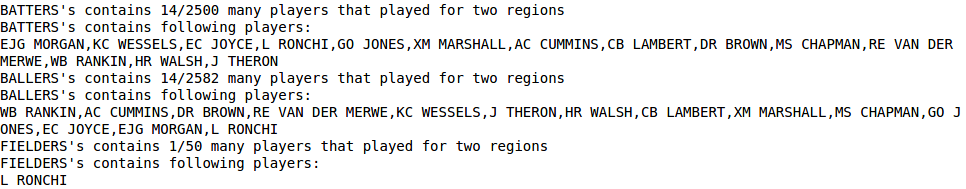
The player names column had regions in it as well, so they were moved to a separate column, following is their before and after comparison:



Apparently, the regions were not only country names, but some other identifiers as well, so they were corrected:



Even after correction, some regions had two countries listed. This was due to the fact that a player had played major matches for both countries. Such regions were retained as is. Following is a snapshot of such players as well as their counts:



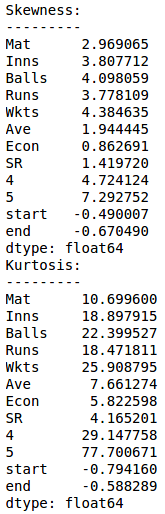
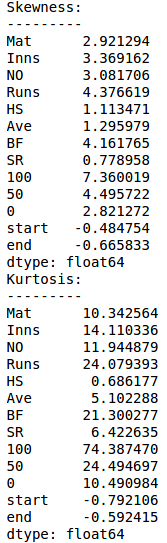
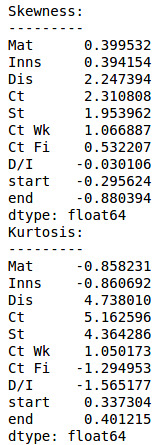
***\* The data was considered cleaned after all these steps and in the next section exploratory data analysis will be performed.***

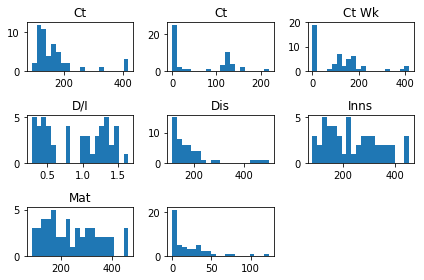
# **Exploratory Data Analysis**

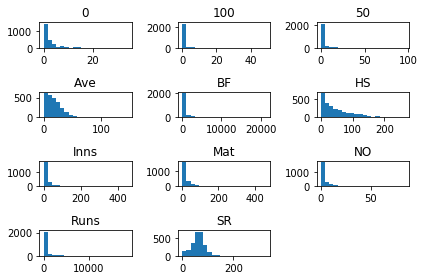
The analysis was done from two perspectives initially. First, data was explored for each player type by merging all match formats for that type. Second, data was explored only for ODI matches for each player type. The latter was suited better for analysis because players appearing in ODI matches ought to be professional cricketers unlike those who appeared once for a test match and then quit.

### Dataset Distribution

To mathematically explore the distribution. Skewness and kurtosis were computed for each columns. Following are the results for batters, ballers and fielders:



To better understand the distributions, histograms were plotted for each column for each dataset:

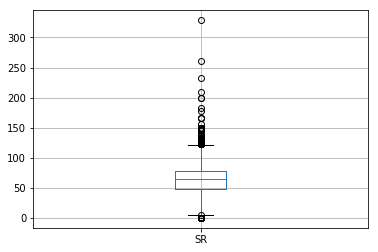
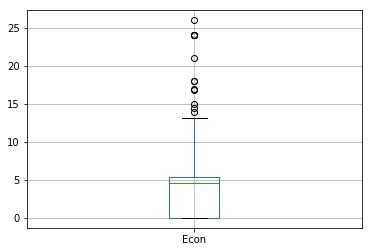


Batters Dataset

Ballers Dataset

Fielders Dataset

To get an overview of the performance of the cricketers, boxplots were created. For batsmen, strike-rate was used, for ballers economy and for fielders number of catches:



Ballers Data

Ballers Data

