SYED NAQI HASSAN NAQVI
20K-1347

Introduction To 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.

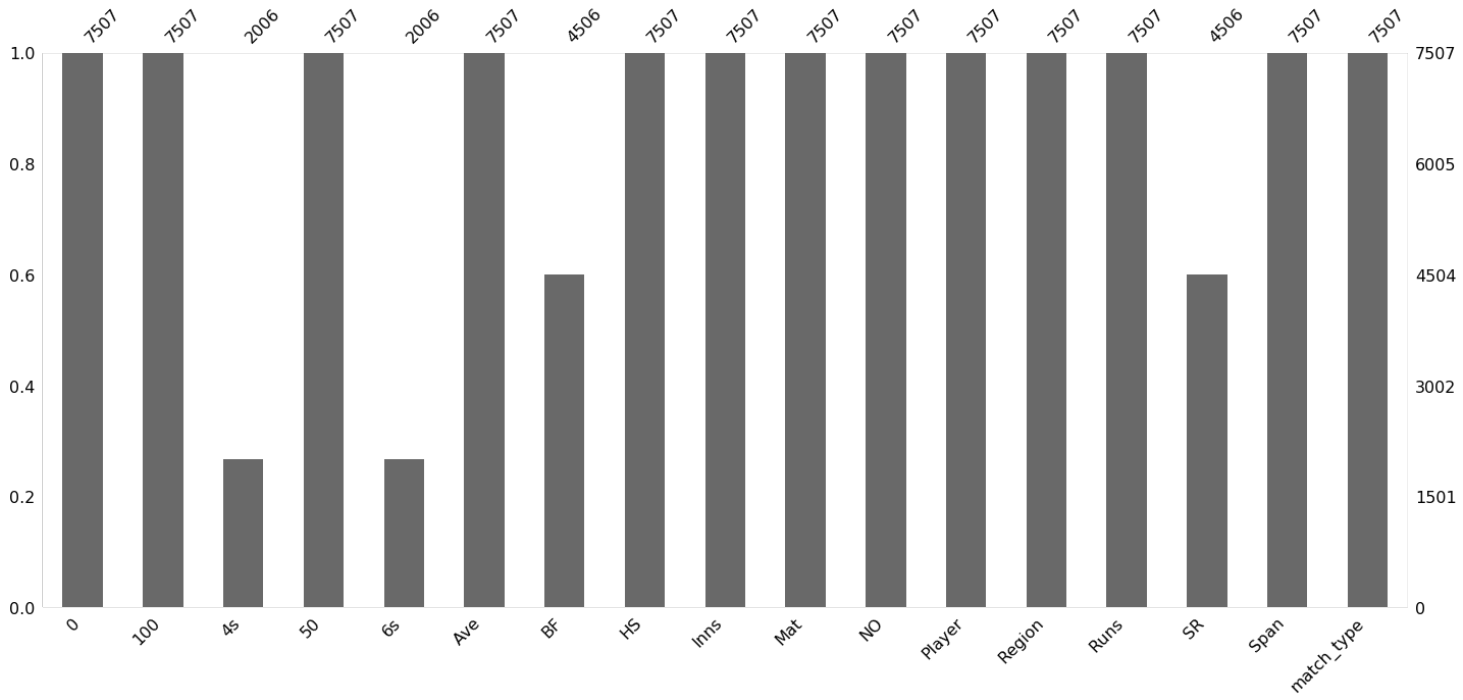
Types of Players

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.

SYED NAQI HASSAN NAQVI
20K-1347

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:

```
BATTERS's data type
************
 Player    object
Span      object
Mat        int64
Inns      object
NO        object
Runs      object
HS        object
Ave       object
BF        object
SR        object
100       object
50        object
0         object
dtype: object
```

```
BALLERS's data types:
************
 Player    object
Span      object
Mat        int64
Inns      object
Balls     object
Runs      object
Wkts      object
BBI       object
Ave       object
Econ      object
SR        object
4         object
5         object
dtype: object
```

```
FIELDERS's data types:
************
 Player     object
Span       object
Mat         int64
Inns        int64
Dis         int64
Ct          int64
St          int64
Ct Wk       int64
Ct Fi       int64
MD         object
D/I       float64
dtype: object
```

After:

```
BATTERS's data types:
************
 Player     object
Mat         int64
Inns        int64
NO          int64
Runs        int64
HS          int64
Ave       float64
BF          int64
SR        float64
100         int64
50          int64
0           int64
start       int64
end         int64
dtype: object
```

```
BALLERS's data types:
************
 Player     object
Mat         int64
Inns        int64
Balls       int64
Runs        int64
Wkts        int64
BBI        object
Ave       float64
Econ      float64
SR        float64
4           int64
5           int64
start       int64
end         int64
dtype: object
```

```
FIELDERS's data types:
************
 Player     object
Mat         int64
Inns        int64
Dis         int64
Ct          int64
St          int64
Ct Wk       int64
Ct Fi       int64
MD         object
D/I       float64
start       int64
end         int64
dtype: object
```

SYED NAQI HASSAN NAQVI
20K-1347

Exploratory Data Analysis

The analysis was done from two perspectives. First, data was explored for each play type by merging all match formats for that type. Second, data was explored only for ODI matches for each player type.