IPL Base Price Modelling and Visualization using Linear Regression

Rajat Chelani

Bachelor of Engineering
Institute of Engineering, Jiwaji University, Gwalior, India

Abstract

The Indian Premier League (IPL) is a professional Twenty20 cricket league in India contested during April and May of every year by teams representing Indian cities. The aim of the research was to predict the base price of players for IPL Auction. The performance of players from different leagues and world tournaments were collected and important features including Runs Scored, Strike Rate, Average, Wickets, Economy Rate were considered. The initial step of our action plan was to clean the data for which we used Python's Pandas. Once the data was cleansed and organized in a structured way, we calculated the Base Price Score, which was a summation of the extracted features, and each feature had its own weightage depending on the impact on IPL. The next step was to calculate the Average Score depending on the weightage given on the basis of different leagues and tournaments. The machine algorithm which was used here was Multiple Linear Regression as, in the above case, we had multiple dependent variables for an independent variable i.e. Base Price Score. For measuring the accuracy of the model, Root Mean Square Error method was incorporated.

Keywords: IPL, Linear Regression, Price Prediction, T-20

1. Introduction

The Indian Premier League (IPL) is a professional Twenty20 cricket league in India contested during April and May of every year by teams representing Indian cities. The league was founded by the Board of Control for Cricket in India (BCCI) in 2007, and is regarded as the brainchild of Lalit Modi, the founder and former commissioner of the league. The IPL is the most-attended cricket league in the world. In 2010, the IPL became the first sporting event in the world to be broadcast live on YouTube. Currently, with eight teams, each team plays each other twice in a home-and-away round-robin format in the league phase.

1.1 Auction Procedure

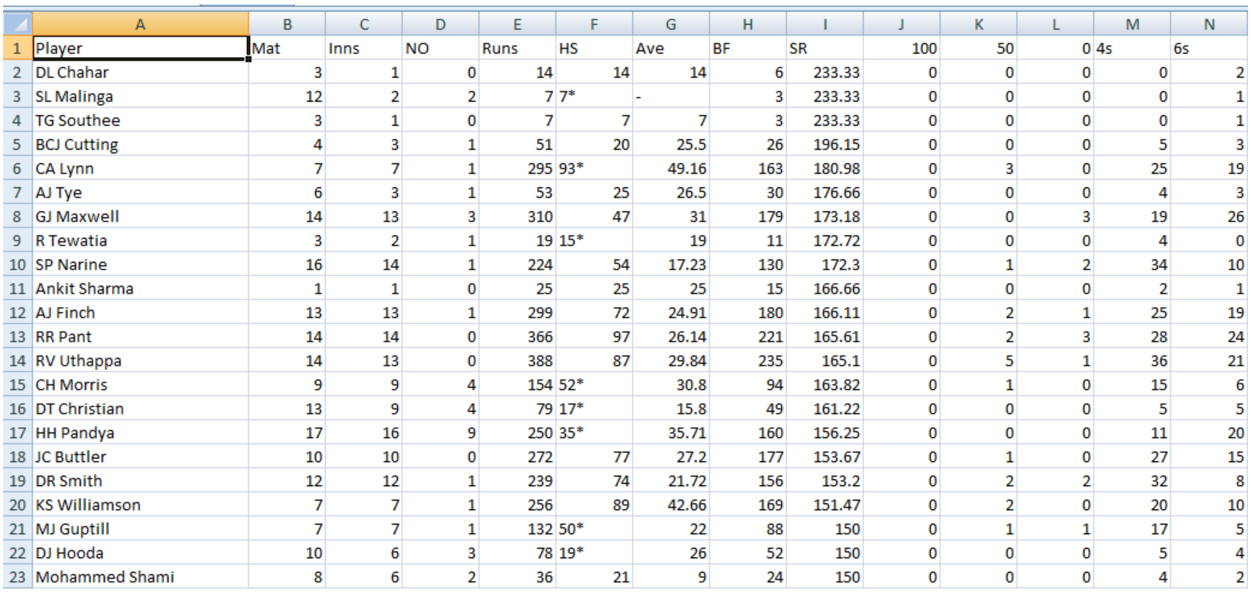
A team can acquire players through the annual player auction. The players who are going under the bidding process have to set a base price which will be the beginning price for the auctioneers. Players are bought by the franchise that bids the highest for them. A player will go unsold if no team bids for them. The auctioneer will give the franchises an option of listing the unsold players they are interested in and will start the bidding for those players for a second time with the base price of the player slashed to half of the original price. If the players remain unsold for the second time, they will be considered unsold in the auction.

1.2 Objective

The aim of this project is to predict the base price of players for the upcoming season of Indian Premier League considering the previous performances of players in various formats and leagues. The performance data of players are collected and important features are extracted. The objective of the work is to analyse the data and predict the base price of the players by using machine learning algorithm Linear Regression. Since input is about previous performances that are unstructured, we perform pre-processing, extract features on to which are important for a player, then calculate the base price score and generate the base price based on the range of score generated, and also plots graph for the result.

1.3 Dataset

This dataset contains performance data of players of different leagues such as Indian Premier League, Big Bash League etc., from ESPNCRICINFO, which includes data from the year 2015 to 2017. This dataset includes performance attributes (Runs scored, Number of Matches, Strike Rate, Number of Wickets, Economy Rate, Number of 0s, 4s, 6s, 50s, 100s, Innings).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Player	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	0	4s	6s
2	DL Chahar	3	1	0	14	14	14	6	233.33	0	0	0	0	2
3	SL Malinga	12	2	2	7	7*	-	3	233.33	0	0	0	0	1
4	TG Southee	3	1	0	7	7	7	3	233.33	0	0	0	0	1
5	BCJ Cutting	4	3	1	51	20	25.5	26	196.15	0	0	0	5	3
6	CA Lynn	7	7	1	295	93*	49.16	163	180.98	0	3	0	25	19
7	AJ Tye	6	3	1	53	25	26.5	30	176.66	0	0	0	4	3
8	GJ Maxwell	14	13	3	310	47	31	179	173.18	0	0	3	19	26
9	R Tewatia	3	2	1	19	15*	19	11	172.72	0	0	0	4	0
10	SP Narine	16	14	1	224	54	17.23	130	172.3	0	1	2	34	10
11	Ankit Sharma	1	1	0	25	25	25	15	166.66	0	0	0	2	1
12	AJ Finch	13	13	1	299	72	24.91	180	166.11	0	2	1	25	19
13	RR Pant	14	14	0	366	97	26.14	221	165.61	0	2	3	28	24
14	RV Uthappa	14	13	0	388	87	29.84	235	165.1	0	5	1	36	21
15	CH Morris	9	9	4	154	52*	30.8	94	163.82	0	1	0	15	6
16	DT Christian	13	9	4	79	17*	15.8	49	161.22	0	0	0	5	5
17	HH Pandya	17	16	9	250	35*	35.71	160	156.25	0	0	0	11	20
18	JC Buttler	10	10	0	272	77	27.2	177	153.67	0	1	0	27	15
19	DR Smith	12	12	1	239	74	21.72	156	153.2	0	2	2	32	8
20	KS Williamson	7	7	1	256	89	42.66	169	151.47	0	2	0	20	10
21	MJ Guptill	7	7	1	132	50*	22	88	150	0	1	1	17	5
22	DJ Hooda	10	6	3	78	19*	26	52	150	0	0	0	5	4
23	Mohammed Shami	8	6	2	36	21	9	24	150	0	0	0	4	2

Figure 1.3.1: Sample Dataset

1.4 Problem Definition

Based on the data described in the previous section, the following are the problems to be solved:

1. To compute a Base Price Score (BP Score) for a given player based on his previous performances.
2. To find how many players fall in category of various base price ranges set by BCCI.
3. To find players with higher base prices in the category of All-Rounder, Batsmen and Bowlers.

2. Methodology

Programming Language used: **Python 3.4**

Machine Learning and Data Analysis Library: **SciKit Learn**

SciKit Learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Reading the Dataset

The dataset contains a total of 120 Comma Separated Value (CSV) Files. Hence to read the files Pandas is used.

Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

2.1 Pre-Processing and Feature Extraction

The pre processing of the data begins by removing unwanted rows from the file which contained information like date of match, venue, and team name of player. Since there were 9 leagues of which data were collected of different years, each league's data was merged into one. As the dataset contained 13 features, we decided to keep only few parameters for evaluation which were important.

The following features are considered for analysis:
For Batsman: {Runs Scored, Strike Rate, Average}
For Bowler: {Economy Rate, Number of Wickets, Average, Strike Rate}

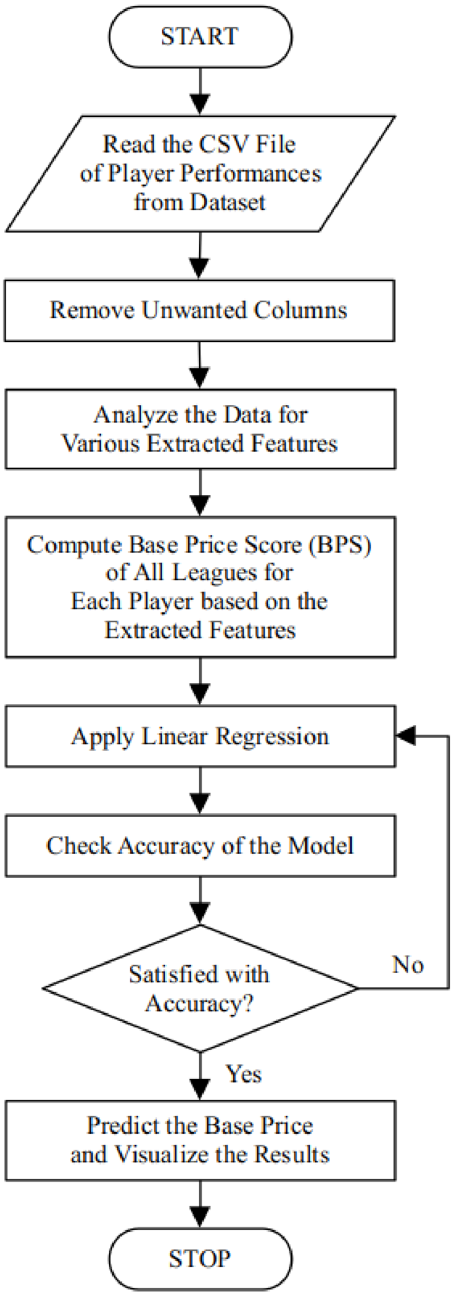
The Base Price Score of each player for each league is generated by giving weights to the extracted features.
For Batsmen weights of {0.4, 0.4, 0.2} were given to Runs Scored, Strike Rate and Average respectively.
For Bowlers weights of {0.4, 0.3, 0.15, 0.15} were given to Economy, Wickets Taken, Average and Strike Rate respectively.

Since the Base Price Score (BPS) calculated were very high so we normalised by taking log base 10.
Then a column called as "Avg L_Score" is made which contains the various values of mean of base price scores of all the leagues for particular player, which are later ranged/split according to the Base Price Range given by the BCCI and IPL committee.

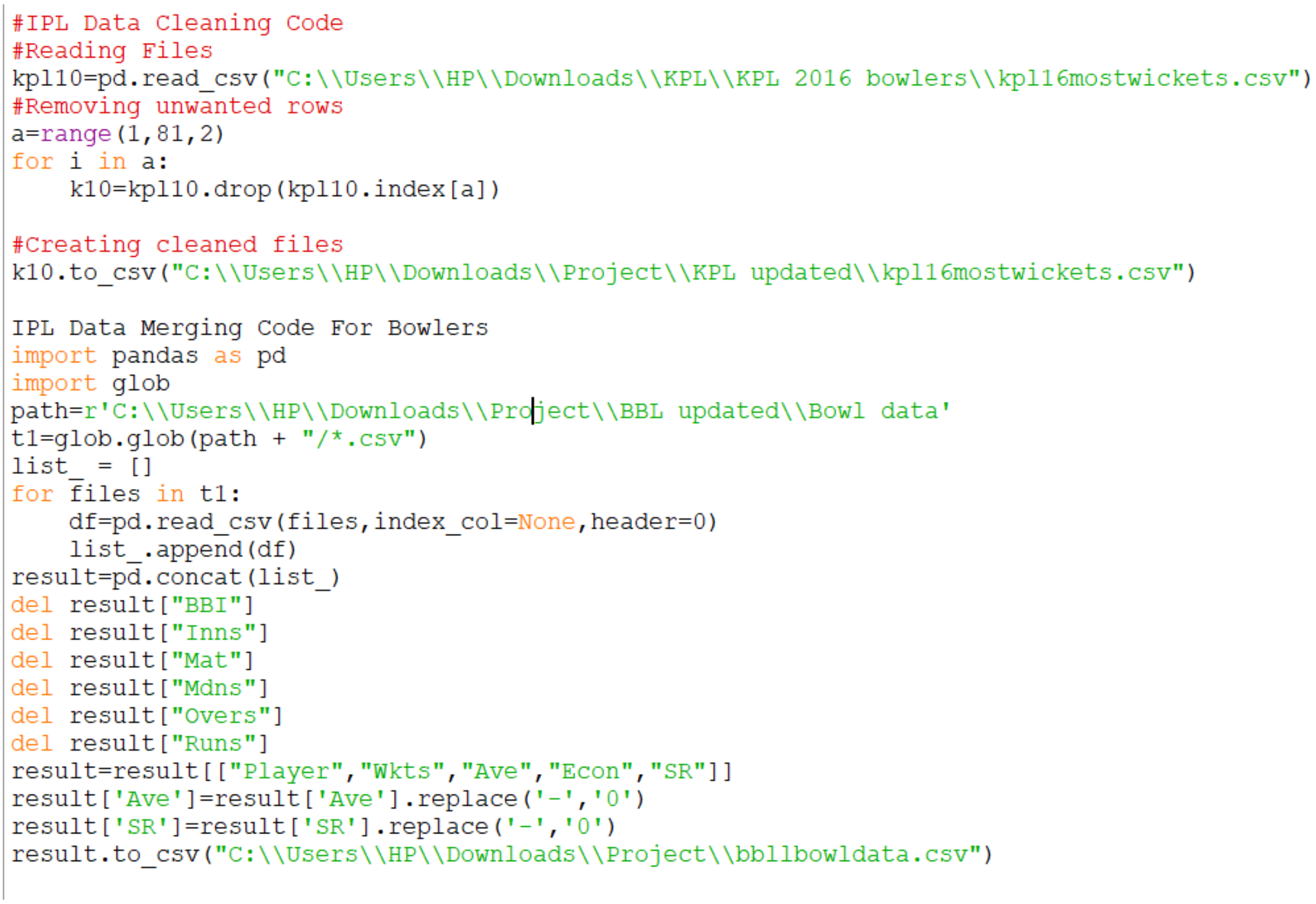
Then data is split into train and test data. The training data contains 80% of data and 20% data is kept for testing and scholastic sampling is used for splitting the data.

2.2 Training using Machine Learning Algorithm

- Machine Learning Model Multiple Linear Regression from SciKit Learn is applied on the dataset.
- Multiple Linear Regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. The independent variables can be continuous or categorical.
- The regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.
- The Avg_LScore Column is the target variable.

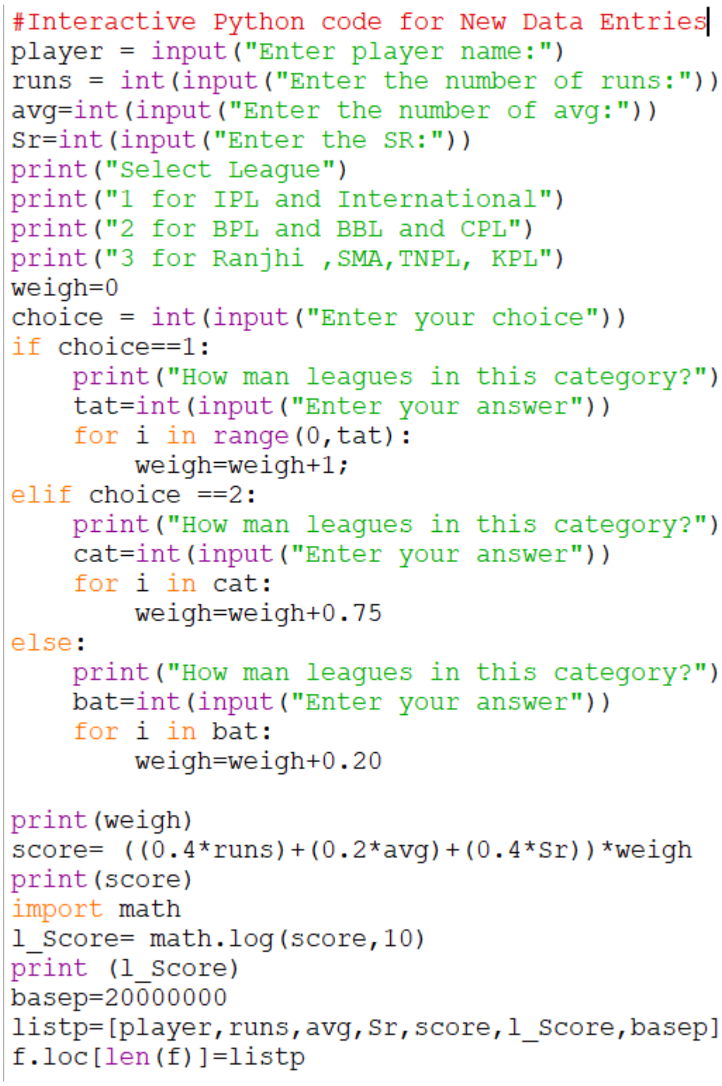

START
Read the CSV File of Player Performances from Dataset
Remove Unwanted Columns
Analyze the Data for Various Extracted Features
Compute Base Price Score (BPS) of All Leagues for Each Player based on the Extracted Features
Apply Linear Regression
Check Accuracy of the Model
Satisfied with Accuracy?
No
Yes
Predict the Base Price and Visualize the Results
STOP

2.3.1 IPL Data Cleaning Code



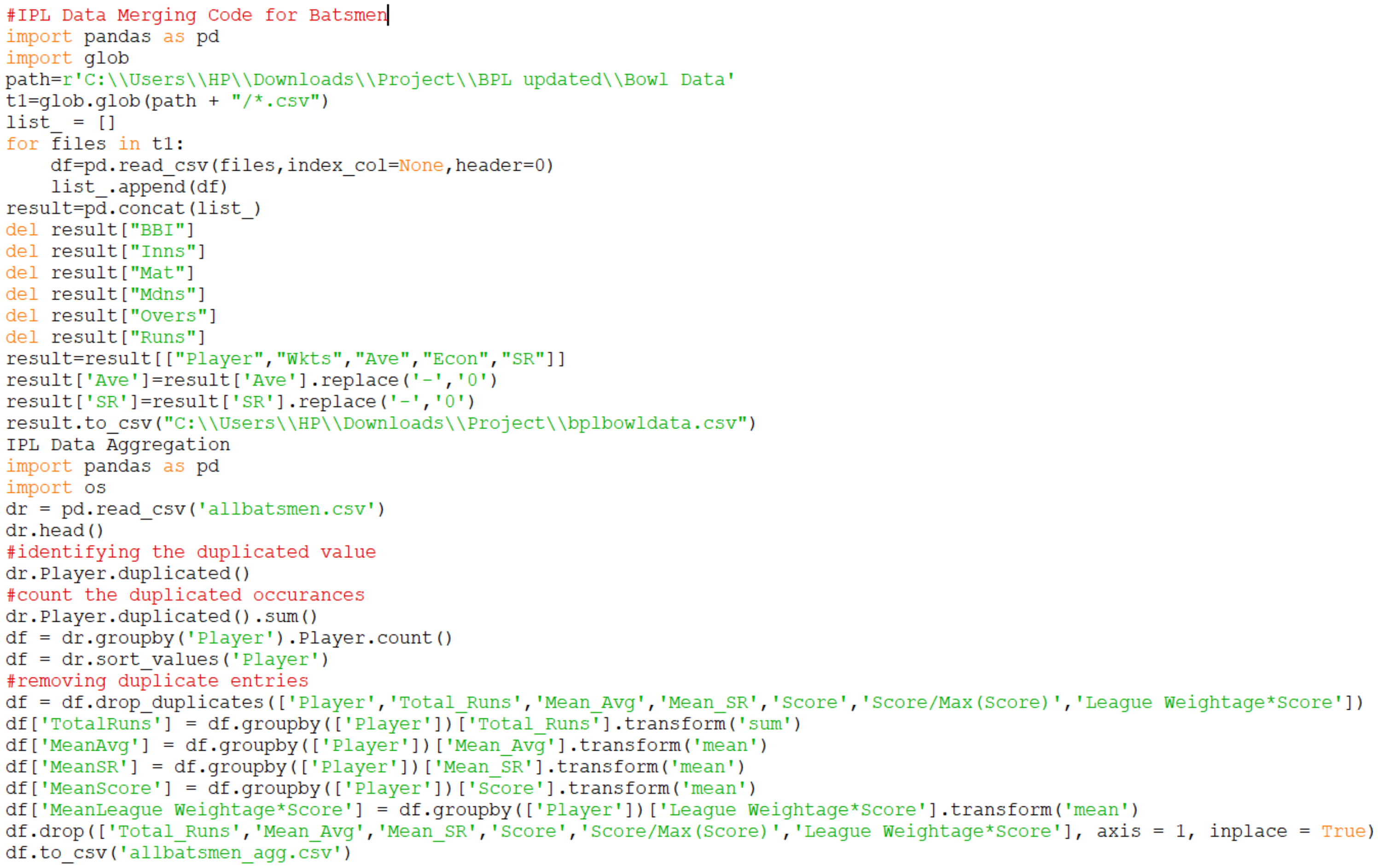
```
#IPL Data Cleaning Code
#Reading Files
kpl10=pd.read_csv("C:\\Users\\HP\\Downloads\\KPL\\KPL 2016 bowlers\\kpl16mostwickets.csv")
#Removing unwanted rows
a=range(1,81,2)
for i in a:
    k10=kpl10.drop(kpl10.index[a])

#Creating cleaned files
k10.to_csv("C:\\Users\\HP\\Downloads\\Project\\KPL updated\\kpl16mostwickets.csv")

IPL Data Merging Code For Bowlers
import pandas as pd
import glob
path=r'C:\\Users\\HP\\Downloads\\Project\\BBL updated\\Bowl data'
t1=glob.glob(path + "/*.csv")
list_ = []
for files in t1:
    df=pd.read_csv(files,index_col=None,header=0)
    list_.append(df)
result=pd.concat(list_)
del result["BBI"]
del result["Inns"]
del result["Mat"]
del result["Mdns"]
del result["Overs"]
del result["Runs"]
result=result[["Player","Wkts","Ave","Econ","SR"]]
result['Ave']=result['Ave'].replace('-','0')
result['SR']=result['SR'].replace('-','0')
result.to_csv("C:\\Users\\HP\\Downloads\\Project\\bbllbowldata.csv")
```

2.3.2 Interactive Python code for New Data Entries



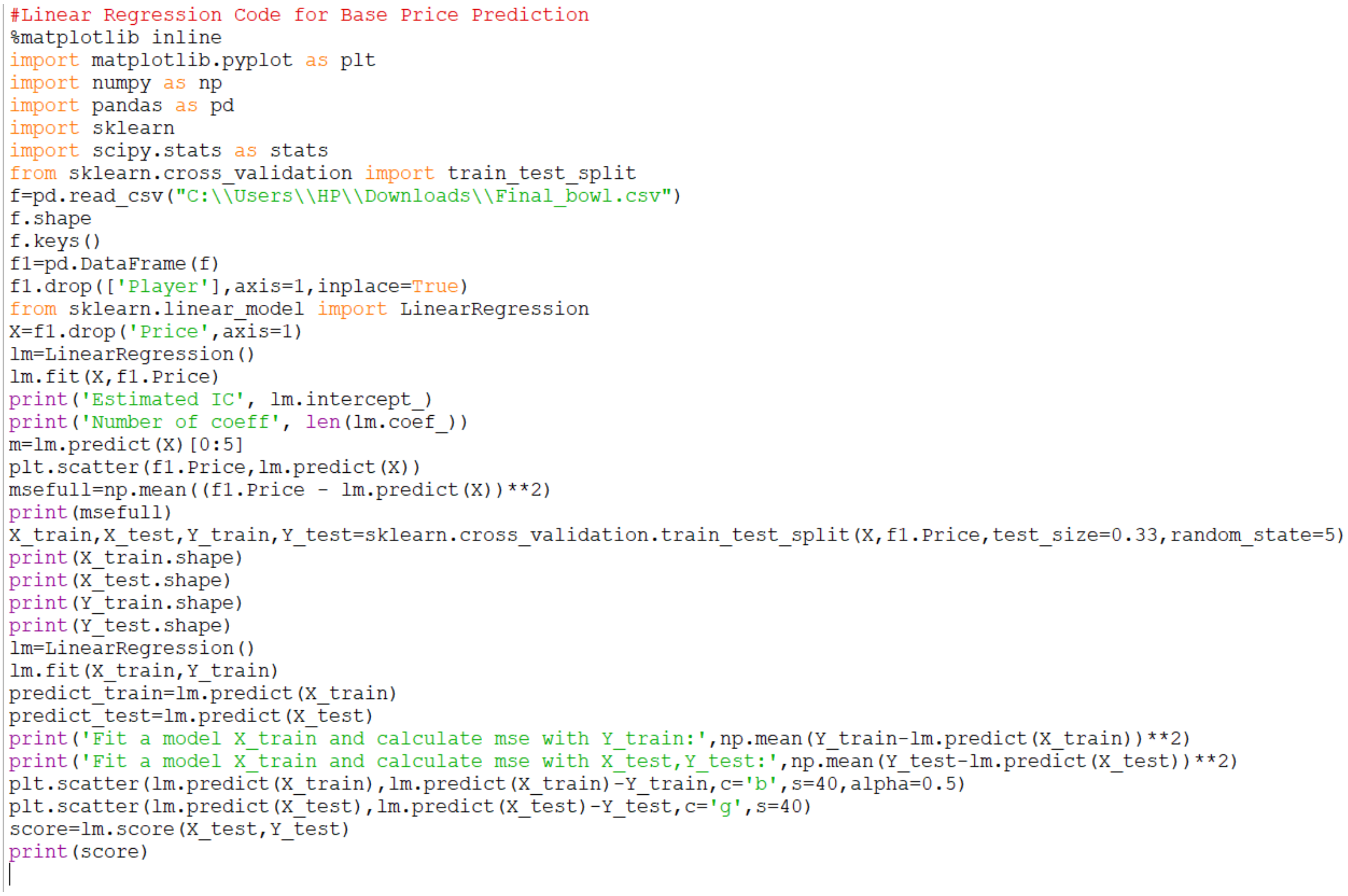
```
#Interactive Python code for New Data Entries
player = input("Enter player name:")
runs = int(input("Enter the number of runs:"))
avg=int(input("Enter the number of avg:"))
Sr=int(input("Enter the SR:"))
print("Select League")
print("1 for IPL and International")
print("2 for BPL and BBL and CPL")
print("3 for Ranjhi ,SMA,TNPL, KPL")
weigh=0
choice = int(input("Enter your choice"))
if choice==1:
    print("How man leagues in this category?")
    tat=int(input("Enter your answer"))
    for i in range(0,tat):
        weigh=weigh+1;
elif choice ==2:
    print("How man leagues in this category?")
    cat=int(input("Enter your answer"))
    for i in cat:
        weigh=weigh+0.75
else:
    print("How man leagues in this category?")
    bat=int(input("Enter your answer"))
    for i in bat:
        weigh=weigh+0.20

print(weigh)
score= ((0.4*runs)+(0.2*avg)+(0.4*Sr))*weigh
print(score)
import math
l_Score= math.log(score,10)
print (l_Score)
basep=20000000
listp=[player,runs,avg,Sr,score,l_Score,basep]
f.loc[len(f)]=listp
```

2.3.3 IPL Data Merging Code for Batsmen

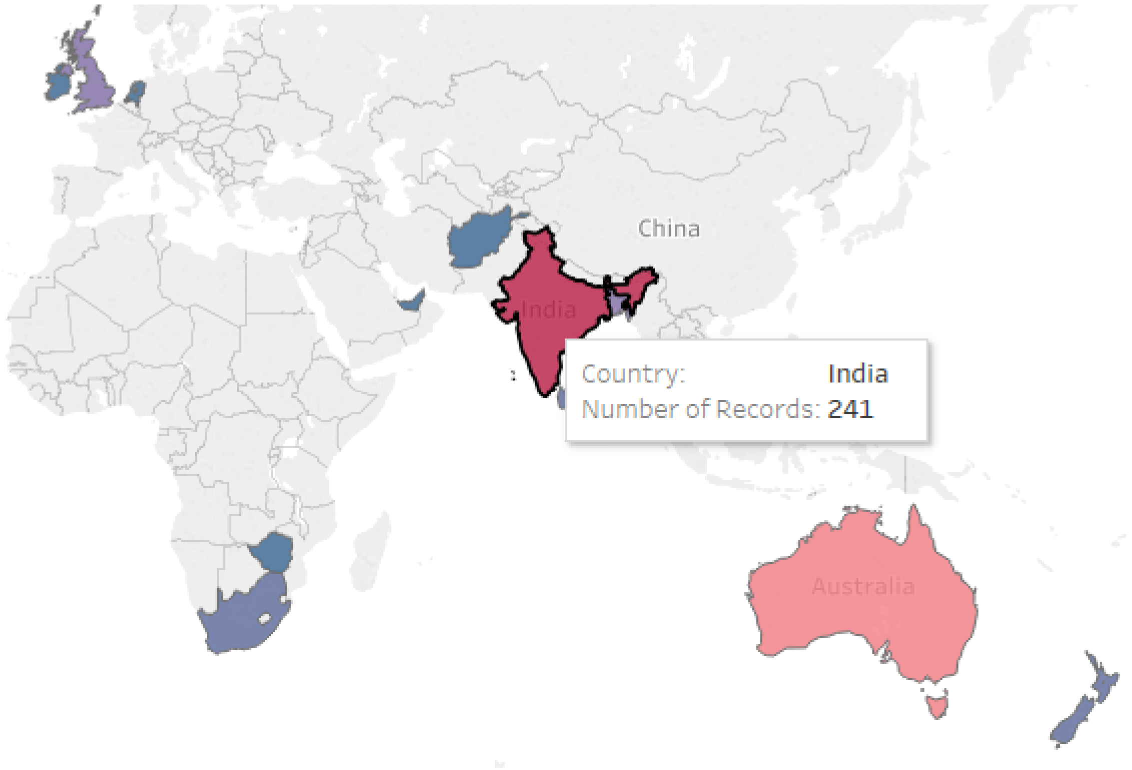


```
#IPL Data Merging Code for Batsmen
import pandas as pd
import glob
path=r'C:\\Users\\HP\\Downloads\\Project\\BPL updated\\Bowl Data'
t1=glob.glob(path + "/*.csv")
list_ = []
for files in t1:
    df=pd.read_csv(files,index_col=None,header=0)
    list_.append(df)
result=pd.concat(list_)
del result["BBI"]
del result["Inns"]
del result["Mat"]
del result["Mdns"]
del result["Overs"]
del result["Runs"]
result=result[["Player","Wkts","Ave","Econ","SR"]]
result['Ave']=result['Ave'].replace('-','0')
result['SR']=result['SR'].replace('-','0')
result.to_csv("C:\\Users\\HP\\Downloads\\Project\\bplbowldata.csv")
IPL Data Aggregation
import pandas as pd
import os
dr = pd.read_csv('allbatsmen.csv')
dr.head()
#identifying the duplicated value
dr.Player.duplicated()
#count the duplicated occurances
dr.Player.duplicated().sum()
df = dr.groupby('Player').Player.count()
df = dr.sort_values('Player')
#removing duplicate entries
df = df.drop_duplicates(['Player','Total_Runs','Mean_Avg','Mean_SR','Score','Score/Max(Score)','League Weightage*Score'])
df['TotalRuns'] = df.groupby(['Player'])['Total_Runs'].transform('sum')
df['MeanAvg'] = df.groupby(['Player'])['Mean_Avg'].transform('mean')
df['MeanSR'] = df.groupby(['Player'])['Mean_SR'].transform('mean')
df['MeanScore'] = df.groupby(['Player'])['Score'].transform('mean')
df['MeanLeague Weightage*Score'] = df.groupby(['Player'])['League Weightage*Score'].transform('mean')
df.drop(['Total_Runs','Mean_Avg','Mean_SR','Score','Score/Max(Score)','League Weightage*Score'], axis = 1, inplace = True)
df.to_csv('allbatsmen_agg.csv')
```

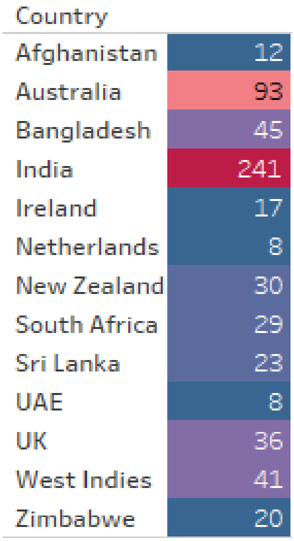
2.3.4 Linear Regression Code for Base Price Prediction



```
#Linear Regression Code for Base Price Prediction
%matplotlib inline
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import sklearn
import scipy.stats as stats
from sklearn.cross_validation import train_test_split
f=pd.read_csv("C:\\Users\\HP\\Downloads\\Final_bowl.csv")
f.shape
f.keys()
f1=pd.DataFrame(f)
f1.drop(['Player'],axis=1,inplace=True)
from sklearn.linear_model import LinearRegression
X=f1.drop('Price',axis=1)
lm=LinearRegression()
lm.fit(X,f1.Price)
print('Estimated IC', lm.intercept_)
print('Number of coeff', len(lm.coef_))
m=lm.predict(X)[0:5]
plt.scatter(f1.Price,lm.predict(X))
msefull=np.mean((f1.Price - lm.predict(X))**2)
print(msefull)
X_train,X_test,Y_train,Y_test=sklearn.cross_validation.train_test_split(X,f1.Price,test_size=0.33,random_state=5)
print(X_train.shape)
print(X_test.shape)
print(Y_train.shape)
print(Y_test.shape)
lm=LinearRegression()
lm.fit(X_train,Y_train)
predict_train=lm.predict(X_train)
predict_test=lm.predict(X_test)
print('Fit a model X_train and calculate mse with Y_train:',np.mean(Y_train-lm.predict(X_train))**2)
print('Fit a model X_train and calculate mse with X_test,Y_test:',np.mean(Y_test-lm.predict(X_test))**2)
plt.scatter(lm.predict(X_train),lm.predict(X_train)-Y_train,c='b',s=40,alpha=0.5)
plt.scatter(lm.predict(X_test),lm.predict(X_test)-Y_test,c='g',s=40)
score=lm.score(X_test,Y_test)
print(score)
```



3. Visualization



Country	
Afghanistan	12
Australia	93
Bangladesh	45
India	241
Ireland	17
Netherlands	8
New Zealand	30
South Africa	29
Sri Lanka	23
UAE	8
UK	36
West Indies	41
Zimbabwe	20

Figure 3.1: Distribution of Players from Different Countries Represented by a Heat Map

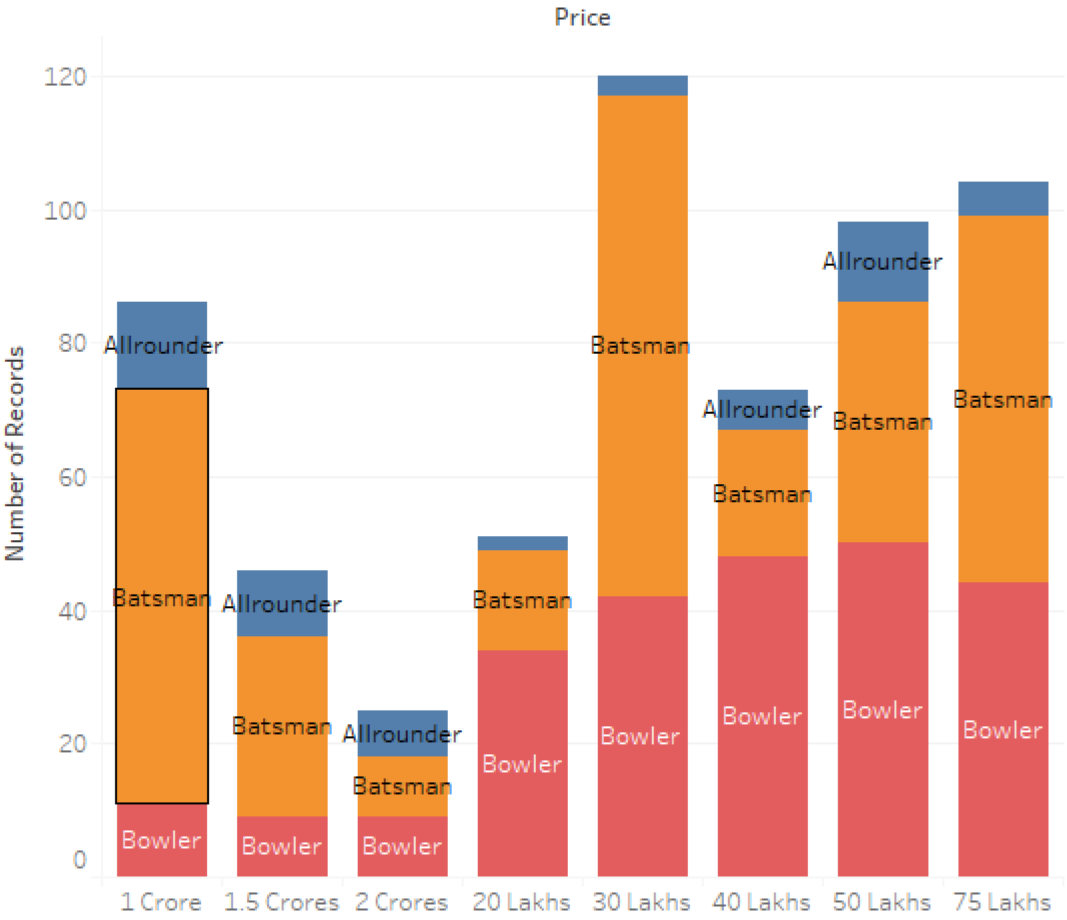
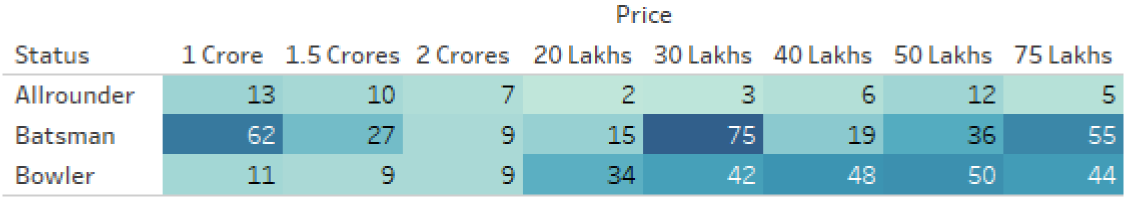


Figure 3.2: Distribution of Players from Different Categories in Various Base Prices Set by IPL Governing Committee



	Price							
Status	1 Crore	1.5 Crores	2 Crores	20 Lakhs	30 Lakhs	40 Lakhs	50 Lakhs	75 Lakhs
Allrounder	13	10	7	2	3	6	12	5
Batsman	62	27	9	15	75	19	36	55
Bowler	11	9	9	34	42	48	50	44

Figure 3.3: Tabular View of Players in Different Categories of Base Price

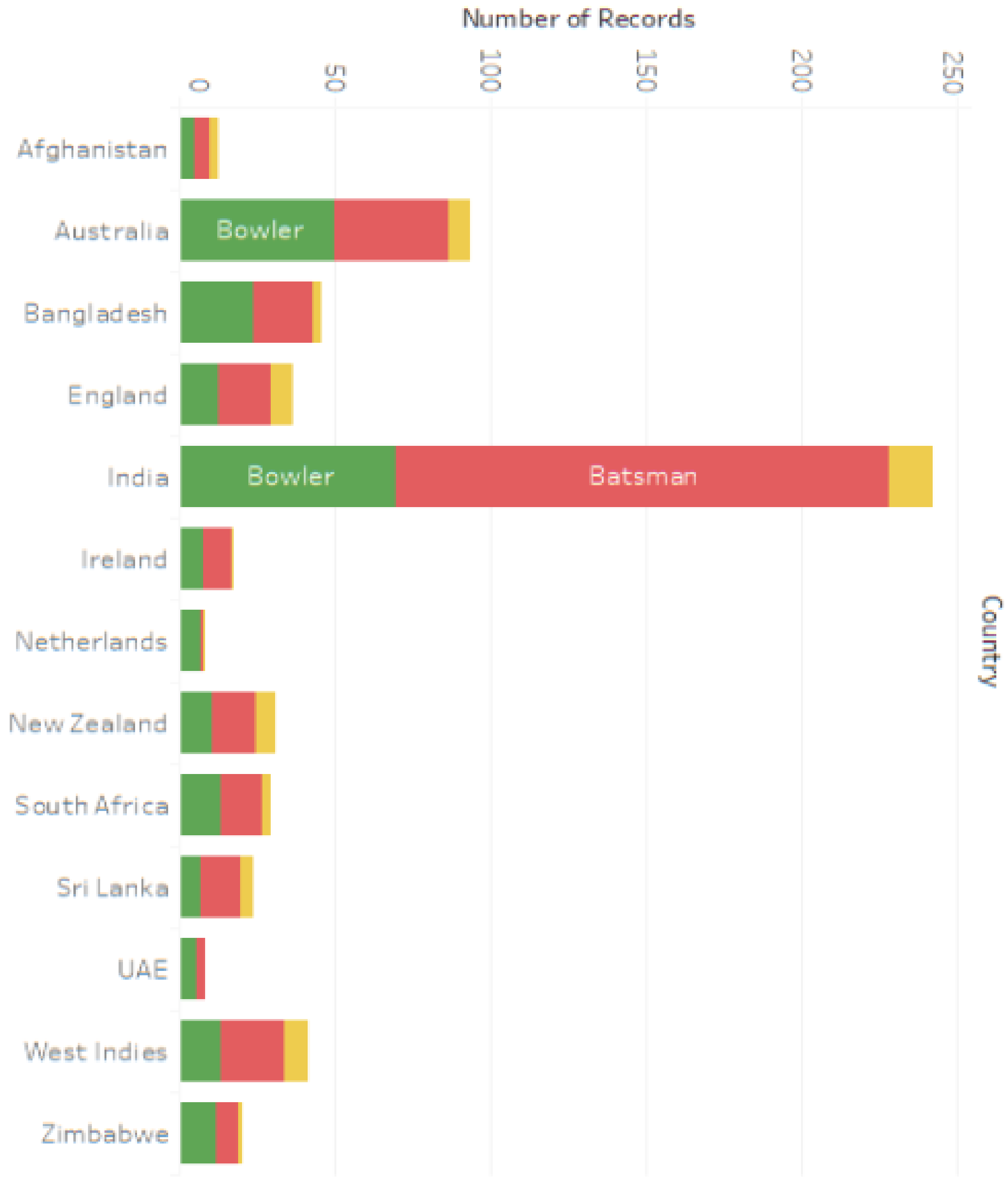


Figure 3.4: Number of Players from Each Country Categorized by Bowlers, Batsmen and All-Rounders

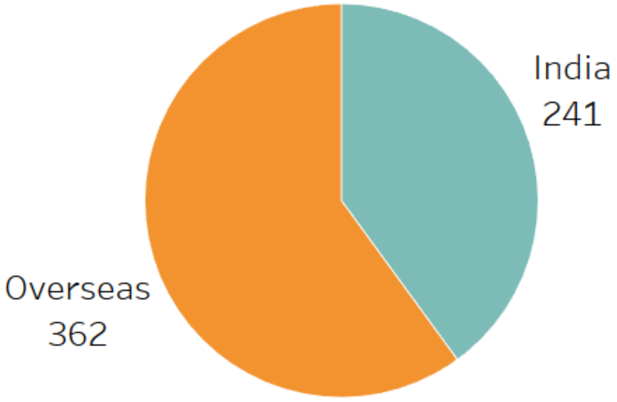


Figure 3.5: Number of Indian Players vs Overseas Players

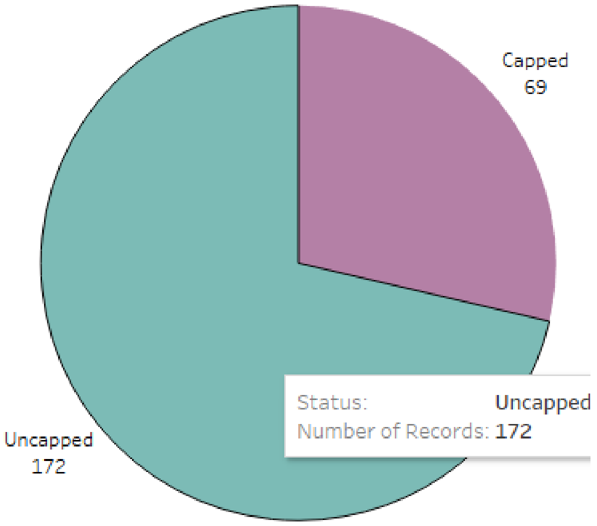


Figure 3.6: Number of Capped vs Uncapped from India

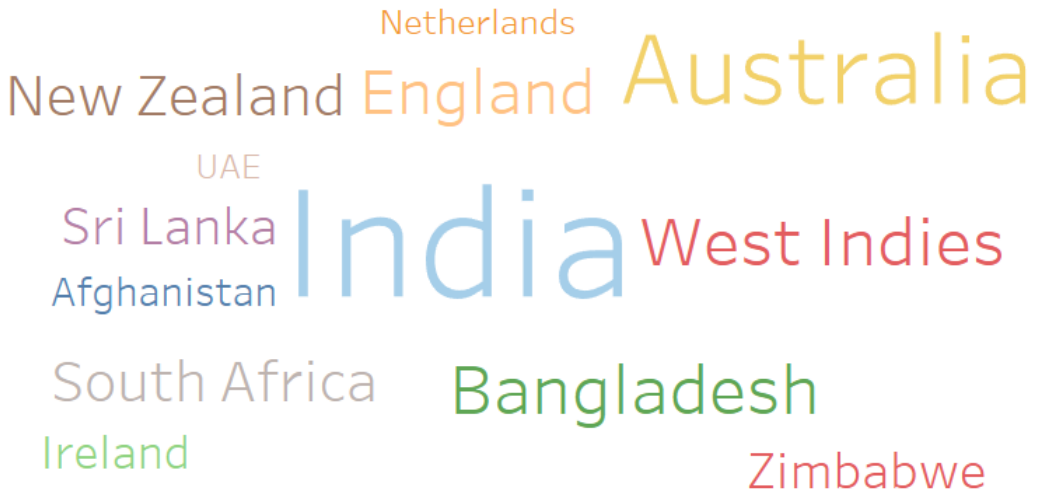


Figure 3.7: Word Cloud Representing Number of Participations in Auction from Each Country

References

1. Data source: https://www.espncricinfo.com/
2. Linear Regression https://www.geeksforgeeks.org/ml-linear-regression/
3. General information from https://en.wikipedia.org/wiki/Indian_Premier_League
4. Hyun-Il Lim (2019). IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC) https://ieeexplore.ieee.org/xpl/conhome/8746989/proceeding
5. Sainathan Ganesh Iyer, Anurag Dipakkumar Pawar (2019). International Conference on Smart Systems and Inventive Technology (ICSSIT) https://ieeexplore.ieee.org/xpl/conhome/8966524/proceeding
6. B. Pavlyshenko (2016). IEEE International Conference on Big Data https://ieeexplore.ieee.org/xpl/conhome/7818133/proceeding

7. Mengyu Huang (2020). International Conference on Computer Vision, Image and Deep Learning (CVIDL) https://ieeexplore.ieee.org/xpl/conhome/9270288/proceeding
8. Lyn Bartram, Michael Correll, Melanie Tory (2021). Untidy Data: The Unreasonable Effectiveness of Tables https://research.tableau.com/paper/untidy-data-unreasonable-effectiveness-tables

Acknowledgement

This research paper was written under the mentorship of Professor Alpana Sharma. She is the Head of Department of Computer Science in Institute of Engineering, Jiwaji University, Gwalior.