

## DATA PREPARATION

**1. Merging data sets and/or records:-** We have only two csv files one is for match details and its result and the other is for ball by ball details. We need not merge data sets according to our target goals.

**2. Selecting a sample subset of data:-** We are selecting all rows from our data set. We are removing attribute "umpire1" "umpire2" "umpire3" from our data set as it is not relevant to our target goal and target user.

## 3. DATA CLEANING

### 3.1. Removing or replacing blank or missing values :-

We have missing values in city column. We are replacing missing values with city, having occurrence(frequency) count nearest to mean number of occurrences of matches in a stadium. Below tableau report proves it.

Null\_Values\_In\_City\_Column

City	Id	
Null	403	■
	404	■
	408	■
	410	■
	411	■
	416	■
	418	■
Abu Dhabi	399	■
	401	■
	402	■
	406	■
	412	■
	413	■
	417	■
Ahmedabad	121	■
	128	■
	136	■
	139	■
	423	■
	428	■
	439	■
	442	■
	467	■
	474	■
Bangalore	1	■
	11	■
	15	■

We also have large number of empty values in columns such as "dismissal type", "fielder", "player dismissed". We have replaced empty values in such columns with "None" string value. As it makes sense that in a particular delivery no player was dismissed, none was dismissal kind of that delivery.

Null\_Values\_In\_Dismissed\_Kind\_Column

Dismissal Kind	Match Id	
Null	1	Abc
	2	Abc
	3	Abc
	4	Abc
	5	Abc
	6	Abc
	7	Abc
	8	Abc
	9	Abc
	10	Abc
	11	Abc
	12	Abc
	13	Abc
	14	Abc
	15	Abc
	16	Abc
	17	Abc
	18	Abc
	19	Abc
	20	Abc
	21	Abc
	22	Abc

**3.2.Renaming columns to remove spaces:-** Our column names had spaces in between so we removed spaces using R script.

#### 4. DATA QUALITY

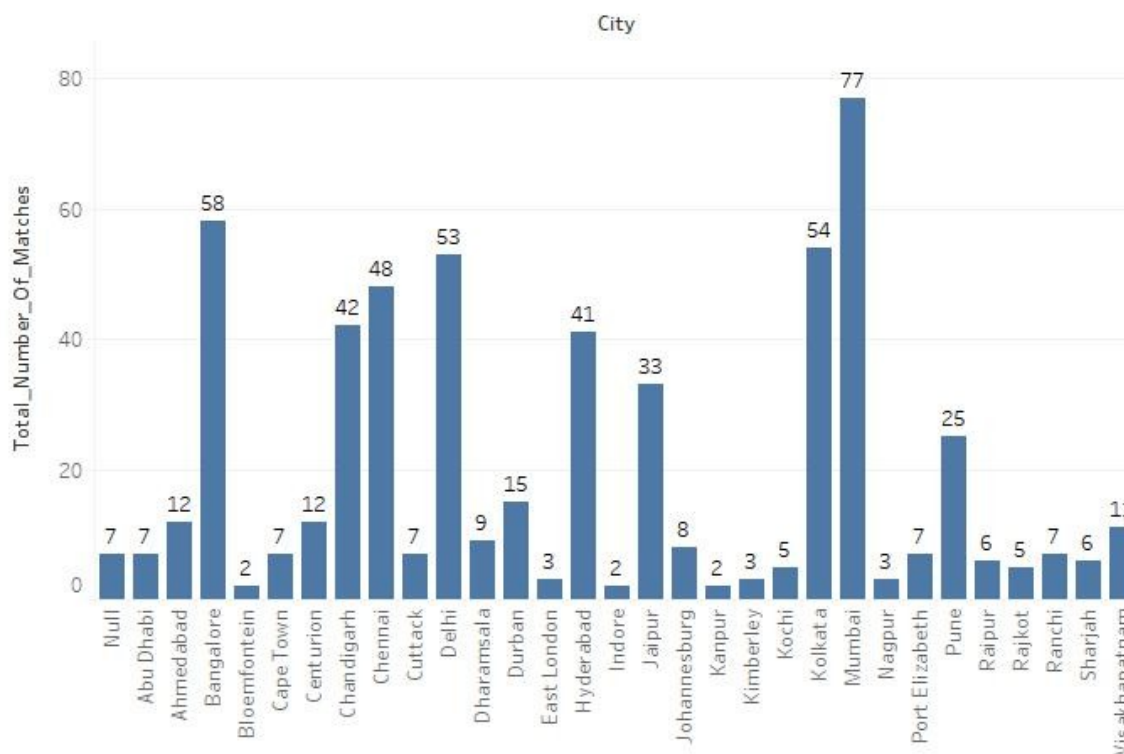
**4.1.**We have **18** attributes in **matches.csv** and **21** attributes in **deliveries.csv** which is sufficient for analysis.

**4.2.**We have different attributes with proper data types except attribute "di\_applied" which is of integer type(but it should be logical type) so we converted it to logical type which returns true or false.

**4.3."****Team****"** is an important attribute with respect to our analysis and we have **13** distinct teams participating in tournament which is good for analysis purpose.

**4.4."****Venue****"** is an important attribute and it has **35** distinct values and every venue has hosted some matches.Although some stadiums gets more chances in hosting matchesBelow tableau report proves that:-

City\_VS\_Number\_of\_matches



Count of Id for each City. The marks are labeled by count of Id.