Name - ARYAN BAJAJ

Task 5 - Exploratory Data Analysis - Sports (Level - Advanced)

In [1]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

In [2]:

```
# Reading data from the link
matches = pd.read_csv('C:/Users/HP/Desktop/New folder/matches.csv',encoding='latin1')
deliveries = pd.read_csv('C:/Users/HP/Desktop/New folder/deliveries.csv',encoding='latin1')
```

Understanding the Data

In [3]:
```
matches.head(5)
```

Out[3]:

	id	season	city	date	team1	team2	toss_winner	toss_decision	result	dl_applied	winner	win_by_runs	win_by_wickets
0	1	2017	Hyderabad	05-04-2017	Sunrisers Hyderabad	Royal Challengers Bangalore	Royal Challengers Bangalore	field	normal	0	Sunrisers Hyderabad	35	0
1	2	2017	Pune	06-04-2017	Mumbai Indians	Rising Pune Supergiant	Rising Pune Supergiant	field	normal	0	Rising Pune Supergiant	0	7
2	3	2017	Rajkot	07-04-2017	Gujarat Lions	Kolkata Knight Riders	Kolkata Knight Riders	field	normal	0	Kolkata Knight Riders	0	10
3	4	2017	Indore	08-04-2017	Rising Pune Supergiant	Kings XI Punjab	Kings XI Punjab	field	normal	0	Kings XI Punjab	0	6
4	5	2017	Bangalore	08-04-2017	Royal Challengers Bangalore	Delhi Daredevils	Royal Challengers Bangalore	bat	normal	0	Royal Challengers Bangalore	15	0

In [4]:
```
deliveries.tail(5)
```

Out[4]:

	match_id	inning	batting_team	bowling_team	over	ball	batsman	non_striker	bowler	is_super_over	...	bye_runs	legbye_runs	noball_
179073	11415	2	Chennai Super Kings	Mumbai Indians	20	2	RA Jadeja	SR Watson	SL Malinga	0	...	0	0	
179074	11415	2	Chennai Super Kings	Mumbai Indians	20	3	SR Watson	RA Jadeja	SL Malinga	0	...	0	0	
179075	11415	2	Chennai Super Kings	Mumbai Indians	20	4	SR Watson	RA Jadeja	SL Malinga	0	...	0	0	
179076	11415	2	Chennai Super Kings	Mumbai Indians	20	5	SN Thakur	RA Jadeja	SL Malinga	0	...	0	0	
179077	11415	2	Chennai Super Kings	Mumbai Indians	20	6	SN Thakur	RA Jadeja	SL Malinga	0	...	0	0	

5 rows × 21 columns

In [5]:
```
matches.describe()
```

Out[5]:

	id	season	dl_applied	win_by_runs	win_by_wickets
count	756.000000	756.000000	756.000000	756.000000	756.000000
mean	1792.178571	2013.444444	0.025132	13.283069	3.350529
std	3464.478148	3.366895	0.156630	23.471144	3.387963
min	1.000000	2008.000000	0.000000	0.000000	0.000000
25%	189.750000	2011.000000	0.000000	0.000000	0.000000
50%	378.500000	2013.000000	0.000000	0.000000	4.000000
75%	567.250000	2016.000000	0.000000	19.000000	6.000000
max	11415.000000	2019.000000	1.000000	146.000000	10.000000

In [6]:
```
deliveries.describe()
```

Out[6]:
```
             match_id          inning            over            ball   is_super_over      wide_runs        bye_runs    legbye_runs    noball_runs
count   179078.000000   179078.000000   179078.000000   179078.000000   179078.000000   179078.000000   179078.000000   179078.000000   179078.000000   17
mean      1802.252957        1.482952       10.162488        3.615587        0.000452        0.036721        0.004936        0.021136        0.004183
std       3472.322805        0.502074        5.677684        1.806966        0.021263        0.251161        0.116480        0.194908        0.070492
min          1.000000        1.000000        1.000000        1.000000        0.000000        0.000000        0.000000        0.000000        0.000000
25%        190.000000        1.000000        5.000000        2.000000        0.000000        0.000000        0.000000        0.000000        0.000000
50%        379.000000        1.000000       10.000000        4.000000        0.000000        0.000000        0.000000        0.000000        0.000000
75%        567.000000        2.000000       15.000000        5.000000        0.000000        0.000000        0.000000        0.000000        0.000000
max      11415.000000        5.000000       20.000000        9.000000        1.000000        5.000000        4.000000        5.000000        5.000000
```

In [7]:
```
matches.shape
```

Out[7]:
```
(756, 18)
```

In [8]:
```
deliveries.shape
```

Out[8]:
```
(179078, 21)
```

In [9]:

```
matches.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 756 entries, 0 to 755
Data columns (total 18 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   id               756 non-null    int64 
 1   season           756 non-null    int64 
 2   city             749 non-null    object
 3   date             756 non-null    object
 4   team1            756 non-null    object
 5   team2            756 non-null    object
 6   toss_winner      756 non-null    object
 7   toss_decision    756 non-null    object
 8   result           756 non-null    object
 9   dl_applied       756 non-null    int64 
 10  winner           752 non-null    object
 11  win_by_runs      756 non-null    int64 
 12  win_by_wickets   756 non-null    int64 
 13  player_of_match  752 non-null    object
 14  venue            756 non-null    object
 15  umpire1          754 non-null    object
 16  umpire2          754 non-null    object
 17  umpire3          119 non-null    object
dtypes: int64(5), object(13)
memory usage: 106.4+ KB
```

In [10]:

```
deliveries.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 179078 entries, 0 to 179077
Data columns (total 21 columns):
 #   Column            Non-Null Count   Dtype 
---  ------            --------------   ----- 
 0   match_id          179078 non-null  int64 
 1   inning            179078 non-null  int64 
 2   batting_team      179078 non-null  object
 3   bowling_team      179078 non-null  object
 4   over              179078 non-null  int64 
 5   ball              179078 non-null  int64 
 6   batsman           179078 non-null  object
 7   non_striker       179078 non-null  object
 8   bowler            179078 non-null  object
 9   is_super_over     179078 non-null  int64 
 10  wide_runs         179078 non-null  int64 
 11  bye_runs          179078 non-null  int64 
 12  legbye_runs       179078 non-null  int64 
 13  noball_runs       179078 non-null  int64 
 14  penalty_runs      179078 non-null  int64 
 15  batsman_runs      179078 non-null  int64 
 16  extra_runs        179078 non-null  int64 
 17  total_runs        179078 non-null  int64 
 18  player_dismissed  8834 non-null    object
 19  dismissal_kind    8834 non-null    object
 20  fielder           6448 non-null    object
dtypes: int64(13), object(8)
memory usage: 28.7+ MB
```

In [11]:

```
matches.dtypes
```

Out[11]:

```
id                  int64
season              int64
city               object
date               object
team1              object
team2              object
toss_winner        object
toss_decision      object
result             object
dl_applied          int64
winner             object
win_by_runs         int64
win_by_wickets      int64
player_of_match    object
venue              object
umpire1            object
umpire2            object
umpire3            object
dtype: object
```

In [12]:
```
deliveries.dtypes
```

Out[12]:
```
match_id             int64
inning               int64
batting_team        object
bowling_team        object
over                 int64
ball                 int64
batsman             object
non_striker         object
bowler              object
is_super_over        int64
wide_runs            int64
bye_runs             int64
legbye_runs          int64
noball_runs          int64
penalty_runs         int64
batsman_runs         int64
extra_runs           int64
total_runs           int64
player_dismissed    object
dismissal_kind      object
fielder             object
dtype: object
```

In [13]:

```
matches.nunique()
```

Out[13]:

```
id                 756
season              12
city                32
date               546
team1               15
team2               15
toss_winner         15
toss_decision        2
result               3
dl_applied           2
winner              15
win_by_runs         89
win_by_wickets      11
player_of_match    226
venue               41
umpire1             61
umpire2             65
umpire3             25
dtype: int64
```

In [14]:

```
deliveries.nunique()
```

Out[14]:

```
match_id            756
inning                5
batting_team         15
bowling_team         15
over                 20
ball                  9
batsman             516
non_striker         511
bowler              405
is_super_over         2
wide_runs             6
bye_runs              5
legbye_runs           6
noball_runs           5
penalty_runs          2
batsman_runs          8
extra_runs            7
total_runs           10
player_dismissed    487
dismissal_kind        9
fielder             499
dtype: int64
```

In [15]:

```
matches.columns
```

Out[15]:

```
Index(['id', 'season', 'city', 'date', 'team1', 'team2', 'toss_winner',
       'toss_decision', 'result', 'dl_applied', 'winner', 'win_by_runs',
       'win_by_wickets', 'player_of_match', 'venue', 'umpire1', 'umpire2',
       'umpire3'],
      dtype='object')
```

In [16]:
```
deliveries.columns
```

Out[16]:
```
Index(['match_id', 'inning', 'batting_team', 'bowling_team', 'over', 'ball',
       'batsman', 'non_striker', 'bowler', 'is_super_over', 'wide_runs',
       'bye_runs', 'legbye_runs', 'noball_runs', 'penalty_runs',
       'batsman_runs', 'extra_runs', 'total_runs', 'player_dismissed',
       'dismissal_kind', 'fielder'],
      dtype='object')
```

Cleaning the Data

In [17]:
```
# Finding all the NULL Values

matches.isnull().sum()
```

Out[17]:
```
id                   0
season               0
city                 7
date                 0
team1                0
team2                0
toss_winner          0
toss_decision        0
result               0
dl_applied           0
winner               4
win_by_runs          0
win_by_wickets       0
player_of_match      4
venue                0
umpire1              2
umpire2              2
umpire3            637
dtype: int64
```

In [18]:
```
matches.dropna(inplace=True)
```

In [19]:

```
matches.isnull().sum()
```

Out[19]:

```
id                 0
season             0
city               0
date               0
team1              0
team2              0
toss_winner        0
toss_decision      0
result             0
dl_applied         0
winner             0
win_by_runs        0
win_by_wickets     0
player_of_match    0
venue              0
umpire1            0
umpire2            0
umpire3            0
dtype: int64
```

In [20]:
```
deliveries.isnull().sum()
```

Out[20]:
```
match_id                 0
inning                   0
batting_team             0
bowling_team             0
over                     0
ball                     0
batsman                  0
non_striker              0
bowler                   0
is_super_over            0
wide_runs                0
bye_runs                 0
legbye_runs              0
noball_runs              0
penalty_runs             0
batsman_runs             0
extra_runs               0
total_runs               0
player_dismissed    170244
dismissal_kind      170244
fielder             172630
dtype: int64
```

In [21]:
```
deliveries.drop(['player_dismissed', 'dismissal_kind','fielder'], axis=1,inplace = True)
```

In [22]:
```
deliveries.isnull().sum()
```

Out[22]:
```
match_id            0
inning              0
batting_team        0
bowling_team        0
over                0
ball                0
batsman             0
non_striker         0
bowler              0
is_super_over       0
wide_runs           0
bye_runs            0
legbye_runs         0
noball_runs         0
penalty_runs        0
batsman_runs        0
extra_runs          0
total_runs          0
dtype: int64
```

Relationship Analysis

In [23]:
```
matches.corr()
```

Out[23]:

	id	season	dl_applied	win_by_runs	win_by_wickets
id	1.000000	0.999288	-0.159826	0.001860	0.056888
season	0.999288	1.000000	-0.158800	0.005130	0.054076
dl_applied	-0.159826	-0.158800	1.000000	-0.051445	-0.013603
win_by_runs	0.001860	0.005130	-0.051445	1.000000	-0.549351
win_by_wickets	0.056888	0.054076	-0.013603	-0.549351	1.000000

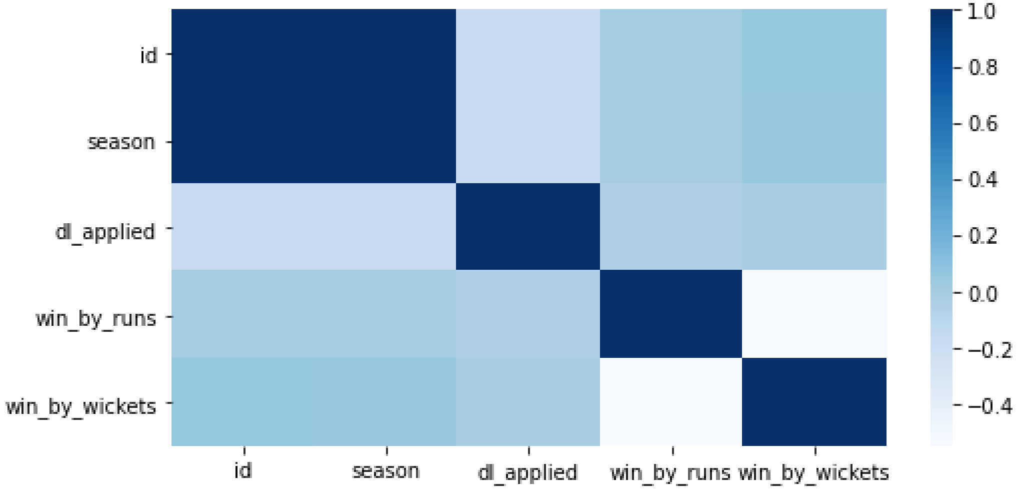
In [24]:
```
deliveries.corr()
```

Out[24]:

	match_id	inning	over	ball	is_super_over	wide_runs	bye_runs	legbye_runs	noball_runs	penalty_runs	batsman_rur
match_id	1.000000	0.003958	0.008268	-0.001349	-0.009150	-0.007549	0.000905	-0.012429	-0.004623	-0.001475	0.03351
inning	0.003958	1.000000	-0.050076	-0.003943	0.084154	0.001201	-0.000757	-0.001996	-0.000904	0.003442	-0.00536
over	0.008268	-0.050076	1.000000	-0.007424	-0.034329	-0.010003	0.012111	-0.004764	0.016984	-0.000979	0.08670
ball	-0.001349	-0.003943	-0.007424	1.000000	-0.001143	-0.004665	0.006602	-0.002727	0.000567	0.000711	0.00795
is_super_over	-0.009150	0.084154	-0.034329	-0.001143	1.000000	-0.001019	0.001353	0.001735	0.013640	-0.000071	0.01012
wide_runs	-0.007549	0.001201	-0.010003	-0.004665	-0.001019	1.000000	-0.006196	-0.015855	-0.008675	0.012817	-0.09457
bye_runs	0.000905	-0.000757	0.012111	0.006602	0.001353	-0.006196	1.000000	-0.004596	-0.002515	-0.000142	-0.01893
legbye_runs	-0.012429	-0.001996	-0.004764	-0.002727	0.001735	-0.015855	-0.004596	1.000000	-0.006434	-0.000362	-0.07010
noball_runs	-0.004623	-0.000904	0.016984	0.000567	0.013640	-0.008675	-0.002515	-0.006434	1.000000	-0.000198	0.00483
penalty_runs	-0.001475	0.003442	-0.000979	0.000711	-0.000071	0.012817	-0.000142	-0.000362	-0.000198	1.000000	-0.00259
batsman_runs	0.033510	-0.005362	0.086701	0.007950	0.010125	-0.094579	-0.018936	-0.070106	0.004832	-0.002591	1.00000
extra_runs	-0.013323	-0.000531	-0.002479	-0.002576	0.003504	0.720916	0.332352	0.554458	0.194899	0.057882	-0.11480
total_runs	0.030727	-0.005485	0.086326	0.007414	0.010891	0.059077	0.051946	0.048075	0.046427	0.009755	0.97727

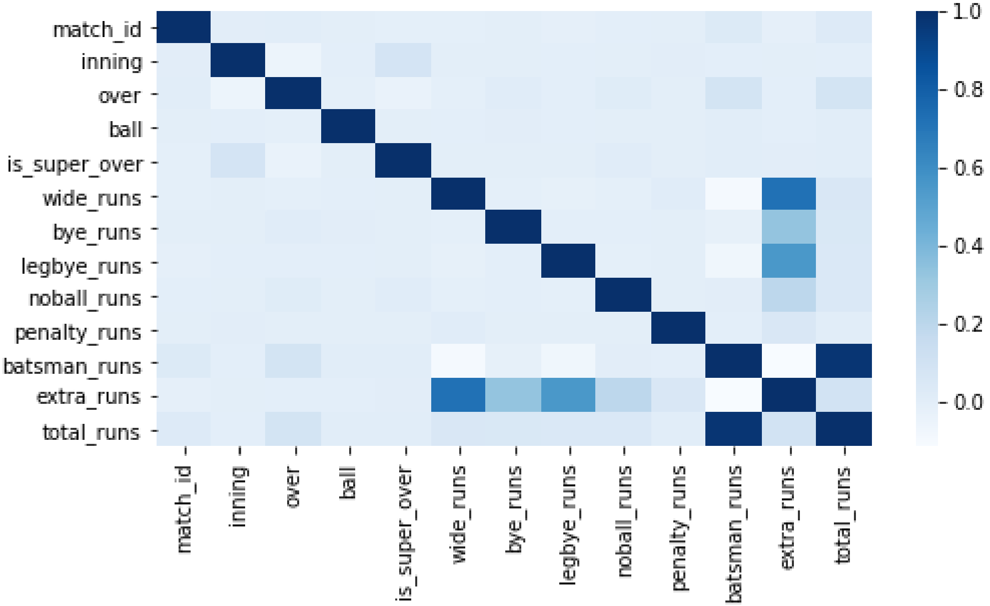
In [25]:
```
plt.figure(figsize=(8,4))
sns.heatmap(matches.corr(),cmap='Blues',annot=False)
```

Out[25]:
```
<AxesSubplot:>
```



```
In [26]: plt.figure(figsize=(8,4))
         sns.heatmap(deliveries.corr(),cmap='Blues',annot=False)
```

```
Out[26]: <AxesSubplot:>
```


match_id
inning
over
ball
is_super_over
wide_runs
bye_runs
legbye_runs
noball_runs
penalty_runs
batsman_runs
extra_runs
total_runs