```
# cricket data

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

# 1 loading the dataset from the github link
cricketdata = pd.read_csv('https://raw.githubusercontent.com/Deepsphere-AI/LVA-Batch5-Assessment/main/Final%20Dataset%20-%20IPL.csv')

print(type(cricketdata))
print('\n')
cricketdata.shape # prints the no of rows and columns
#cricketdata
#cricketdata.head()
#cricketdata.tail()
```

```
<class 'pandas.core.frame.DataFrame'>


(74, 20)
```

```
#2 check and handle the missing values
df = pd.DataFrame(cricketdata)

#df_m = pd.DataFrame(cricketdata, na.value=['na','NaN'])
df.isnull()  # checks if there is any NULL values then gives TRUE
#missing values , there are no missing values in the given data

df.drop_duplicates() # dropping the duplicate values
# no duplicate values exist in the dataset so 0 rows are dropped
```

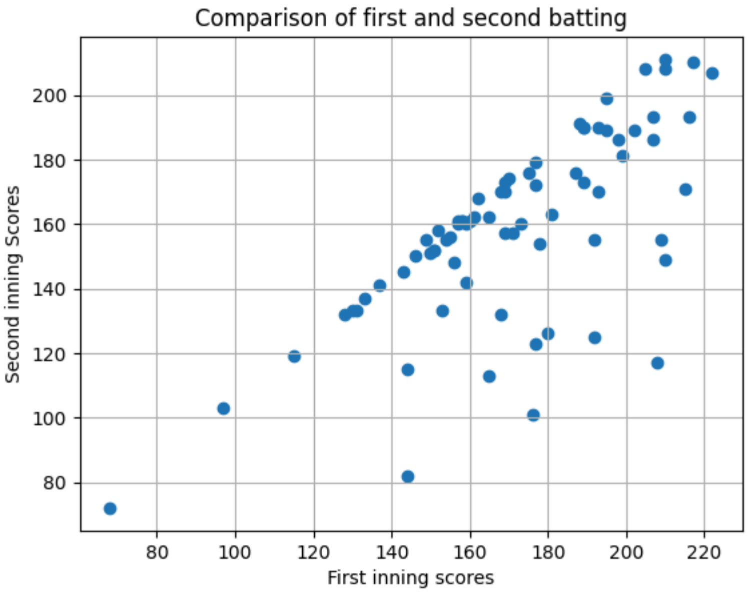
	match_id	date	venue	team1	team2	stage	toss_winner	toss_de
0	1	March 26,2022	Wankhede Stadium, Mumbai	Chennai	Kolkata	Group	Kolkata	
1	2	March 27,2022	Brabourne Stadium, Mumbai	Delhi	Mumbai	Group	Delhi	
2	3	March 27,2022	Dr DY Patil Sports Academy, Mumbai	Banglore	Punjab	Group	Punjab	
3	4	March 28,2022	Wankhede Stadium, Mumbai	Gujarat	Lucknow	Group	Gujarat	
4	5	March 29,2022	Maharashtra Cricket Association Stadium,Pune	Hyderabad	Rajasthan	Group	Hyderabad	
...	...	...	...	...	...	...	...	
69	70	May 22,2022	Wankhede Stadium, Mumbai	Hyderabad	Punjab	Group	Hyderabad	
70	71	May 24,2022	Eden Gardens, Kolkata	Gujarat	Rajasthan	Playoff	Gujarat	
71	72	May 25,2022	Eden Gardens, Kolkata	Banglore	Lucknow	Playoff	Lucknow	
72	73	May 27,2022	Narendra Modi Stadium, Ahmedabad	Banglore	Rajasthan	Playoff	Rajasthan	
73	74	May 29,2022	Narendra Modi Stadium, Ahmedabad	Gujarat	Rajasthan	Final	Rajasthan	

74 rows × 20 columns

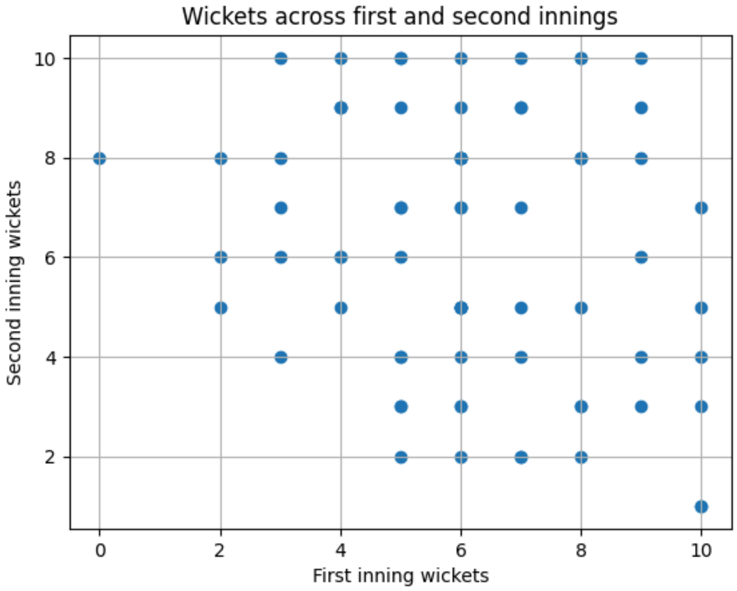
```
#3 mean,median,mode,range,variance,standard deviation

# here we are considering for few of the numerical data rows such as FIRST INNINGS SCORE, SECOND INNINGS SCORE AND by what MARGIN they

#FIRST INNINGS SCORE
print("The mean of first innings score is : ", df['first_ings_score'].mean())
print("The median of first innings score is : ", df['first_ings_score'].median())
print("The variance of first innings score is : ", df['first_ings_score'].var())
print("The standard deviation of first innings score is : ", df['first_ings_score'].std())
print("The mode of first innings score is : ", df['first_ings_score'].mode())


# SECOND INNINGS SCORE

print("The mean of second innings score : ", df['second_ings_score'].mean())
print("The median of second innings score : ", df['second_ings_score'].median())
print("The variance of second innings score : ", df['second_ings_score'].var())
print("The standard deviation of second innings score : ", df['second_ings_score'].std())
print("The mode of second innings score : ", df['second_ings_score'].mode())
```

```
The mean of first innings score is :  171.1216216216216
The median of first innings score is :  169.5
The variance of first innings score is :  843.806923361718
The standard deviation of first innings score is :  29.0483549166165
The mode of first innings score is :  0    169
1    177
2    189
3    210
Name: first_ings_score, dtype: int64
The mean of second innings score :  158.54054054054055
The median of second innings score :  160.0
The variance of second innings score :  858.4435394298408
The standard deviation of second innings score :  29.299207146778578
The mode of second innings score :  0    155
1    161
Name: second_ings_score, dtype: int64
```

```
#4 Data Visualisation (histogram, scatter plot, boxplot, bar charts, pie charts)

#scatterplot b/w first innings and second innings

f_innings = df['first_ings_score']
s_innings = df['second_ings_score']
righttoss = df[['toss_winner','match_winner']]

#plt.bar(f_innings, s_innings)
plt.scatter(f_innings, s_innings) # the scores are plotted agains the scatter plot

plt.xlabel('First inning scores')
plt.ylabel('Second inning Scores')
plt.title('Comparison of first and second batting')
plt.grid()
# toss decision of toss winner and match result
```



```
f_wickets = df['first_ings_wkts']
s_wickets = df['second_ings_wkts']

plt.scatter(f_wickets,s_wickets)
plt.xlabel('First inning wickets')
plt.ylabel('Second inning wickets')
plt.title('Wickets across first and second innings')
plt.grid()
```



```
# 5 correlation btwn the data
# different match conditions like toss decisions, innings score or venue impact match outcomes

df_2 = df.groupby('won_by').value_counts()
print(df.groupby('won_by')['won_by'].value_counts()) # denotes how many matches won by (how many) runs and by (how many) wickets


df_3 = df['margin'].mean()
print("The mean of the margin is : ", df_3)
```

```
won_by
Runs       37
Wickets    37
Name: count, dtype: int64
The mean of the margin is :  16.972972972972972
```

```
# 6 outlier detection

#outlier is some abnormal data point in the data set. It is not useful when it comes to statistical analysis and it can be discarded in
# the outliers can be determined by the box plot.

sns.boxplot(df['first_ings_score'])
```

```
# 7 compare team and individual performances across different matches and venues

df_2 = df[['match_winner','venue']]
df_2.groupby('match_winner').value_counts() # the team performances are analysed across different teams in different venues.
```

```
match_winner  venue
Banglore      Wankhede Stadium, Mumbai                          4
              Dr DY Patil Sports Academy, Mumbai                2
              Maharashtra Cricket Association Stadium,Pune      2
              Eden Gardens, Kolkata                             1
Chennai       Dr DY Patil Sports Academy, Mumbai                3
              Maharashtra Cricket Association Stadium,Pune      1
Delhi         Brabourne Stadium, Mumbai                         4
              Dr DY Patil Sports Academy, Mumbai                2
              Wankhede Stadium, Mumbai                          1
Gujarat       Maharashtra Cricket Association Stadium,Pune      3
              Wankhede Stadium, Mumbai                          3
              Brabourne Stadium, Mumbai                         2
              Dr DY Patil Sports Academy, Mumbai                2
              Eden Gardens, Kolkata                             1
              Narendra Modi Stadium, Ahmedabad                  1
Hyderabad     Dr DY Patil Sports Academy, Mumbai                3
              Brabourne Stadium, Mumbai                         2
```

```
               Wankhede Stadium, Mumbai                       1
Kolkata        Wankhede Stadium, Mumbai                       3
               Maharashtra Cricket Association Stadium,Pune   2
               Dr DY Patil Sports Academy, Mumbai             1
Lucknow        Dr DY Patil Sports Academy, Mumbai             3
               Wankhede Stadium, Mumbai                       2
               Maharashtra Cricket Association Stadium,Pune   2
               Brabourne Stadium, Mumbai                      2
Mumbai         Wankhede Stadium, Mumbai                       2
               Dr DY Patil Sports Academy, Mumbai             1
               Brabourne Stadium, Mumbai                      1
Punjab         Dr DY Patil Sports Academy, Mumbai             2
               Wankhede Stadium, Mumbai                       2
               Brabourne Stadium, Mumbai                      2
               Maharashtra Cricket Association Stadium,Pune   1
Rajasthan      Wankhede Stadium, Mumbai                       3
               Brabourne Stadium, Mumbai                      3
               Maharashtra Cricket Association Stadium,Pune   2
               Narendra Modi Stadium, Ahmedabad               1
               Dr DY Patil Sports Academy, Mumbai             1
Name: count, dtype: int64
```

```
# 8 focus on key player performances  'Player of the match' , assess the impact of top scorers and best bowlers on their team success

key_player = df['player_of_the_match']
top_scorers = df['top_scorer']
best_bowlers = df['best_bowling']
print(key_player.value_counts().head(5)) # shows the most player with player of matches\
print('\n')
print(top_scorers.value_counts().head(5))  # shows the top scorers in a match
print('\n')
best_bowlers.value_counts().head(5) # shows the best bowlers in a match
```

```
player_of_the_match
Kuldeep Yadav      4
Jos Buttler        3
Umesh Yadav        2
Quinton de Kock    2
David Miller       2
Name: count, dtype: int64


top_scorer
Jos Buttler         7
Quinton de Kock     5
Liam Livingstone    4
Shubman Gill        4
KL Rahul            4
Name: count, dtype: int64


best_bowling
Yuzvendra Chahal    5
Rashid Khan         4
T Natarajan         3
Kagiso Rabada       3
Jasprit Bumrah      3
Name: count, dtype: int64
```

9 Summary of the dataset

From the given IPL Dataset, we can analyse multiple trends from it.

1. The match outcomes of winning and losing are equal in both the first innings and second innings as both the half of the teams that batted first won and other half of the matches are won by teams with second batting.
2. From the batting trends in the matches we can say that the scores each of the innings are above the 160. (From the scatter plot)
3. Based on the analysis, the most valuable player of the season is 'Kuldeep Yadav' with 4 player of the match awards. The top scorer is "Jos Butler" and the top bowler is "Yuzvendra Chahal"