IT495 - Exploratory Data Analysis Project Presentation

May 17, 2023

IPL Data Analysis

...

Group-09

Rohan Shah

202218002

B.Tech (ECE)

Karan Parashar

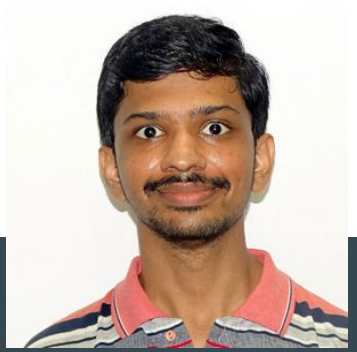
202218004

BSc Physics

Mohammed Nauman

202218010

Bsc Statistics



Nisarg Shah

202218034

BSc Statistics

Overview

- In this project we have done analysis of IPL matches played between year 2008 - 2023 using exploratory data analysis techniques. We have used python 3 language for analysis, where python libraries such as numpy & pandas used for mathematical computations and for visualization we used matplotlib and seaborn.
- We have also used the pytrend modules for google trend analysis,
- For predictive analysis we used machine learning techniques such as linear regression and K - Means clustering.

Topics to be covered in presentation

Topic 1

Dataset Description:

- An introduction about the data we have used here in analysis

Topic 2

Exploratory Data Analysis:

- An in depth analysis of the IPL matches played in year between 2008 - 2023

Topic 3

Machine learning:

- Machine learning techniques applied on the data.

Dataset Description

...

Attribute	Description
ID	Match id
INNINGS	Innings of the match that was being played
OVER	Over of the match
BALL	Ball in particular over (0 to 6)
BATSMAN	Batsman on strike
NON_STRIKER	Non striker batsman
BOWLER	Current bowler
BATSMAN_RUNS	Run scored by batsman on that ball
EXTRA_RUNS	Extra type runs (wide, no ball etc.)

Attribute	Description
TOTAL_RUNS	Total runs of batsman and extra
NON_BOUNDARY	Whether boundary was not scored
IS_WICKET	Whether wicket was taken on the ball
DISMISSAL_KIND	If wicket, type of dismissal
PLAYER_DISMISSED	If wicket, name of player
EXTRA_TYPE	Type of extra runs (wide, no ball etc.)
BATTING_TEAM	Batting Team
BOWLING_TEAM	Bowling Team

Exploratory Data Analysis

...

Numpy

It is a scientific computing library for Python, used for handling huge data sets. it provides advanced mathematical operations and a multi-dimensional structure known as ndarray.

Pandas

Numpy is the foundation for the open-source Python package known as Pandas. primarily used to perform data analysis tasks and related tabular data manipulation in DataFrames.

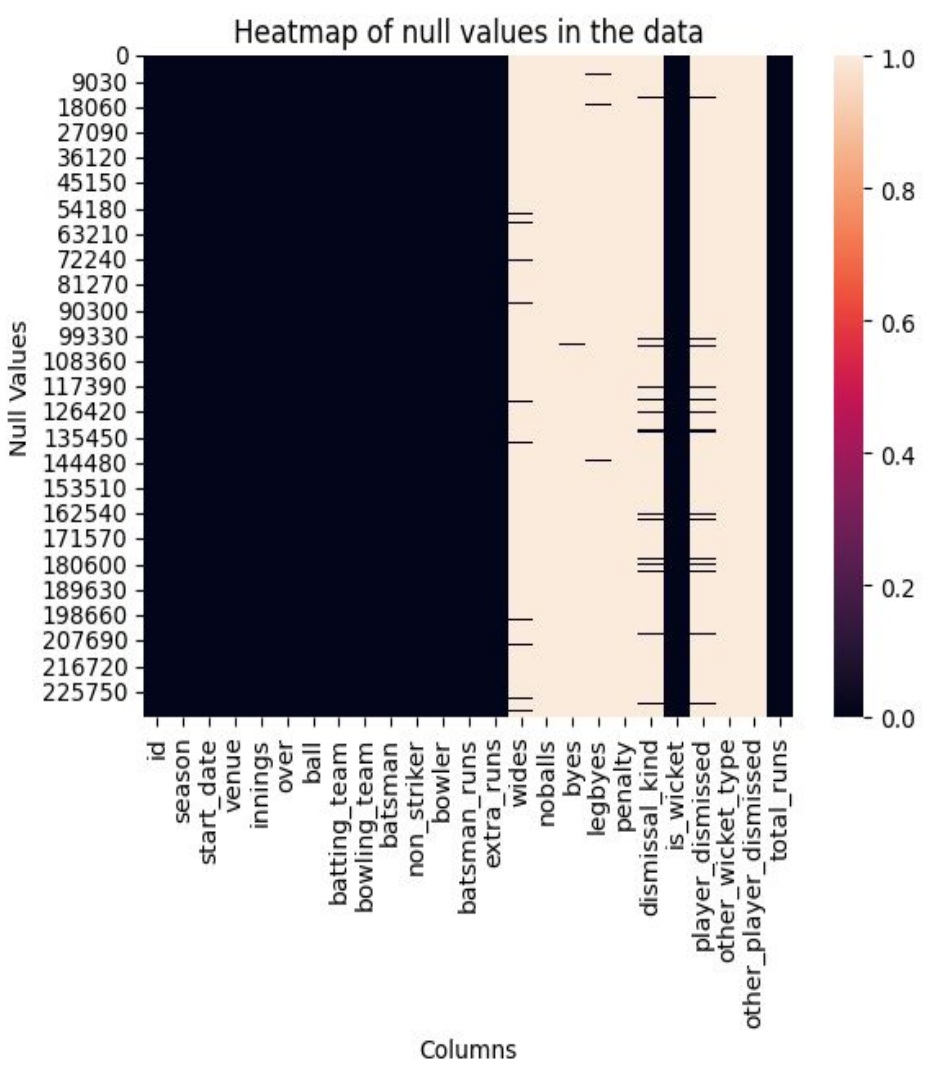
Matplotlib

Python's Matplotlib library is a complete tool for building animated, static, and interactive visualisations. It supports various plots like scatter-plot, bar charts, histograms, box plots, line charts, pie charts, etc.

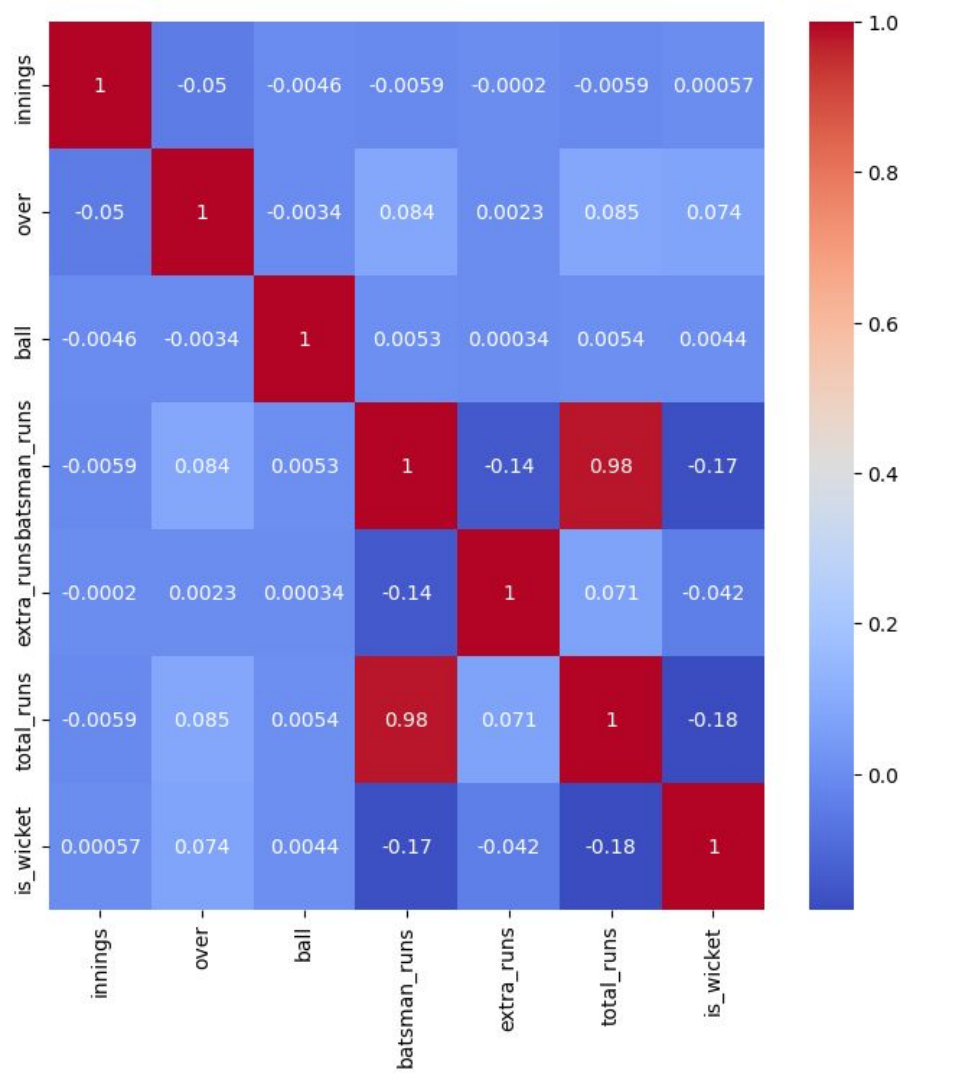
Scikit-learn

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It offers a variety of effective tools for statistical modelling and machine learning, including classification, regression and clustering.

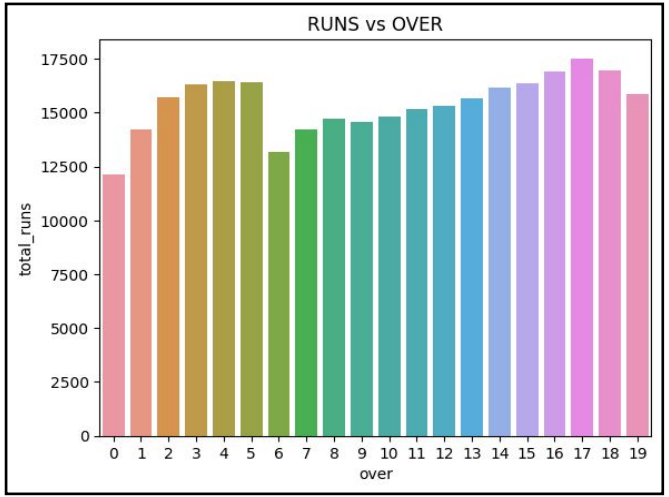
- Given graph shows the amount of missing values.
- Here white space shows the missing values.
- For features such as ‘id’,’ball’,’season’ etc there are no missing values while features such as ‘wide’, ‘no ball’ etc are sparse from which we can infer that frequency of wide and no balls are very less.



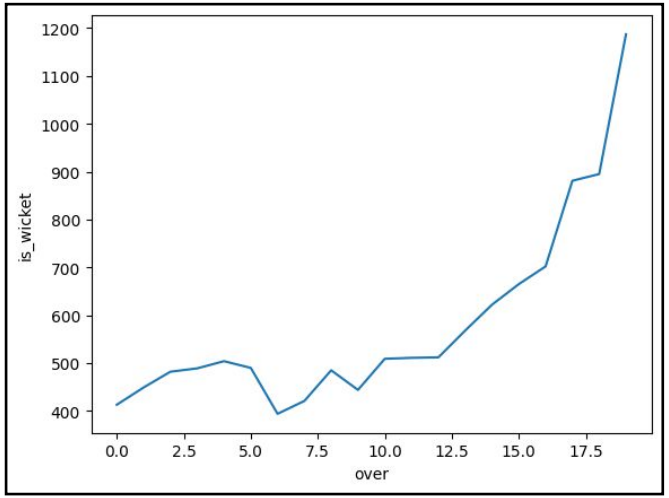
- Given heatmap shows the correlation between features.
- As we can see there is strong correlation between ‘total_runs’ and ‘batsman_run’ (+0.98) features, indicating that not all the run are contributed by batsman, some of them are extra.
- While other features are pairwise weakly dependant on each other.



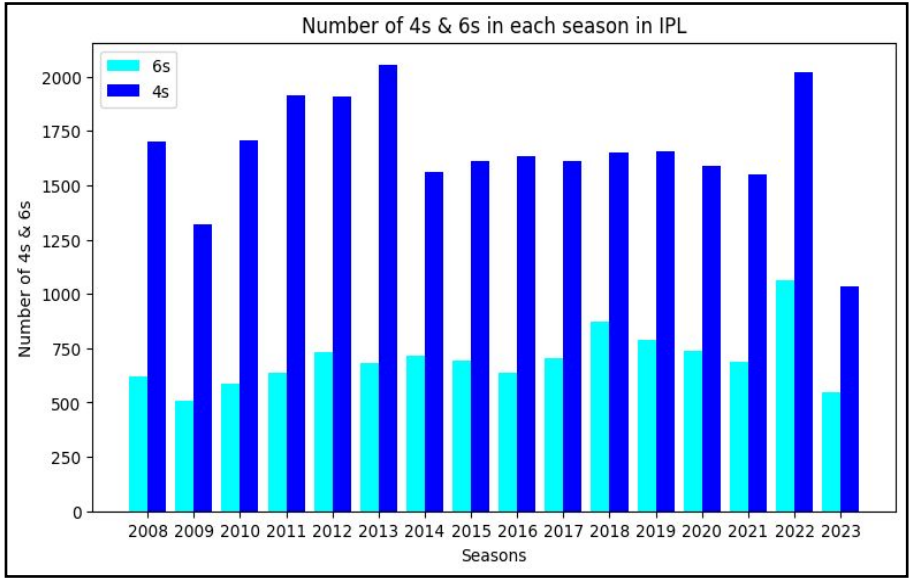
- Total runs scored in particular over
- As we can see from the graph in 18th over highest runs has been scored.
- Also karl pearson's correlation coefficient between over vs total_runs is: **0.5781**



- Total wickets taken particular over
- As we can see from the graph in death overs, the number of wickets spikes.
- Karl pearson's correlation coefficient between over vs wickets is: **0.7961**



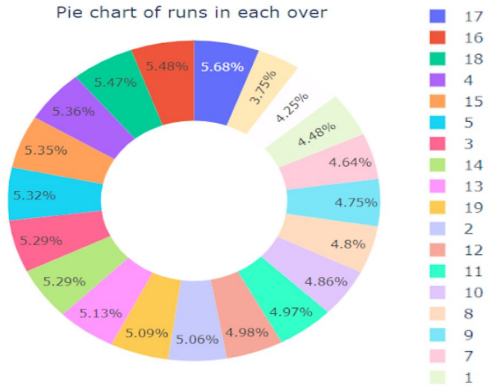
- Total sixes scored in particular season
- As we can see from the graph in year 2022 has highest sixes among all seasons.

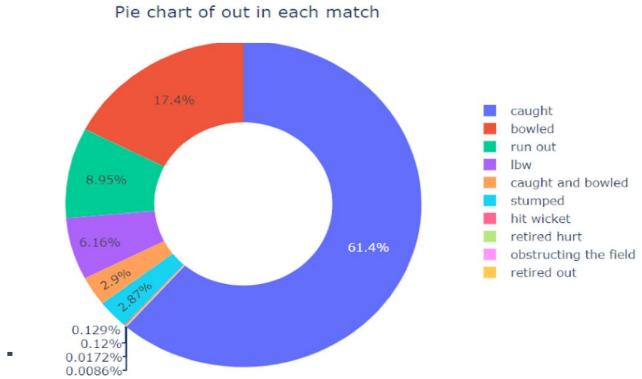


- Total fours scored in particular season
- As we can see from the graph in year 2013 most number of fours has been scored.

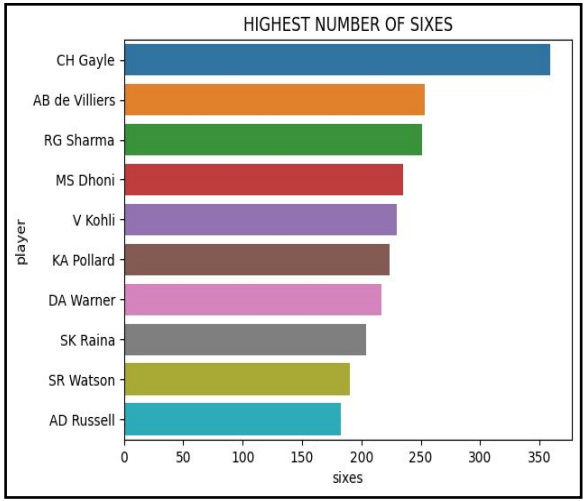
- Graph is showing the percentage of runs scored in particular overs. We can see that most runs have been scored in the 17th over.

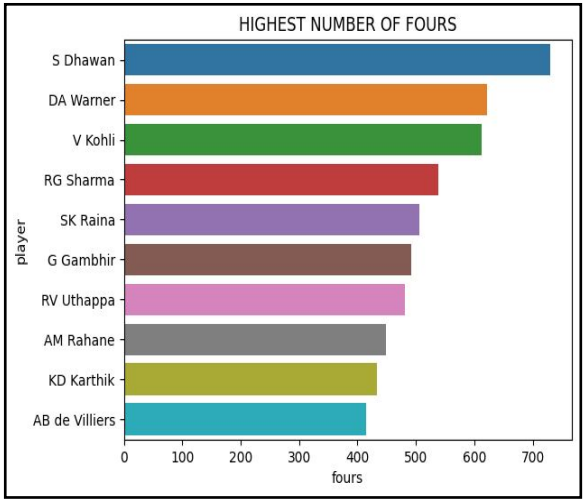
- Graph is showing percentage of wickets fallen and its cause. We conclude that the reason for the majority of wickets to fall is ‘Catch out’, followed by ‘Bowled’.

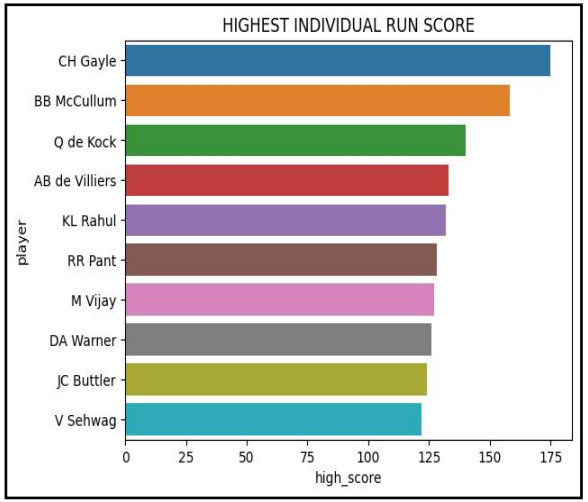




Batsmen Analysis

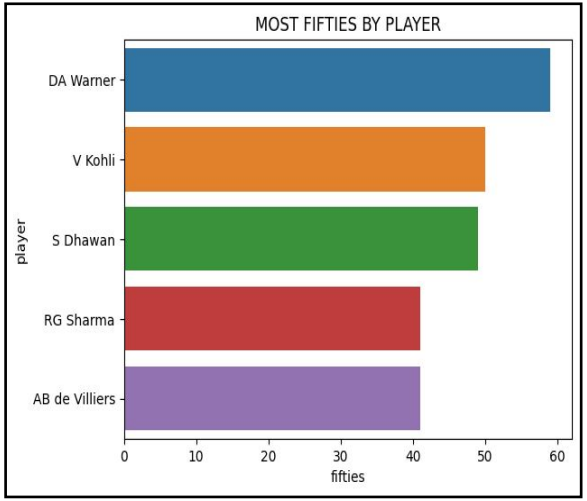


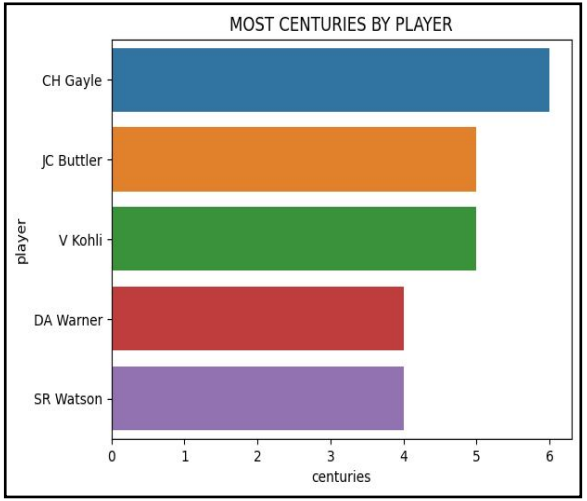


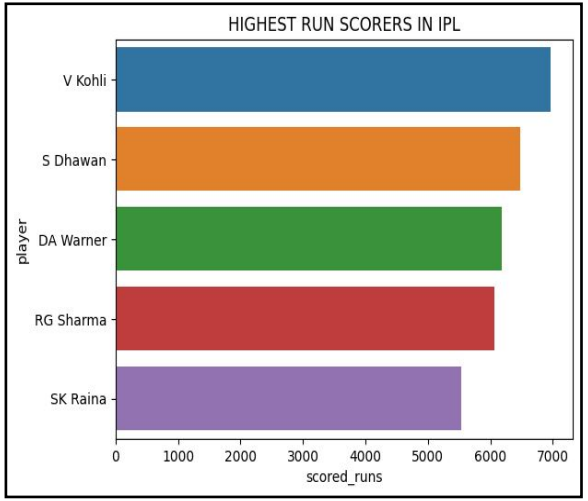


- Highest number of sixes scored by an individual batsman
- Highest number of fours scored by an individual batsman
- Highest runs scored by an individual batsman in on match

Batsmen Analysis

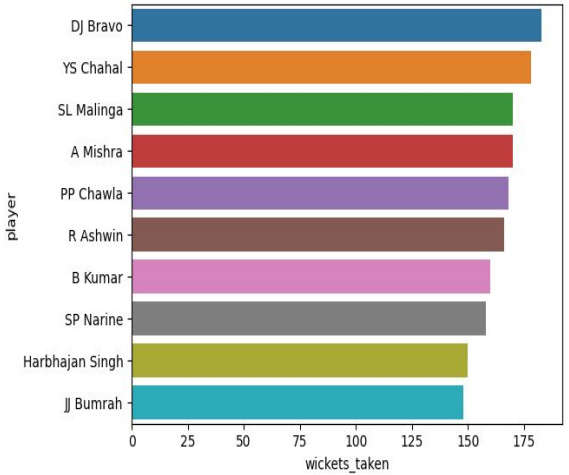




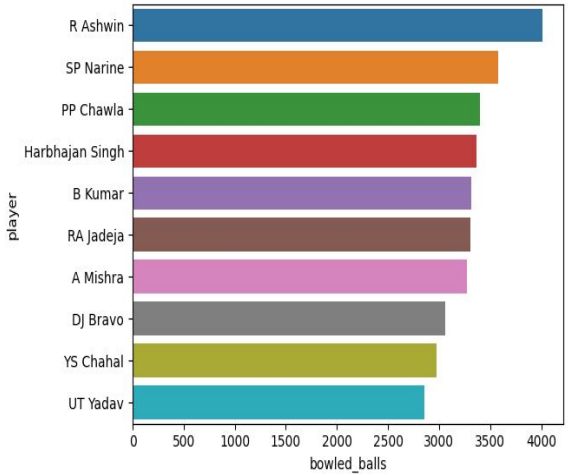


- Highest number of half - centuries scored by an individual batsman
- Highest number of centuries scored by an individual batsman
- Highest runs scored by an individual batsman in IPL

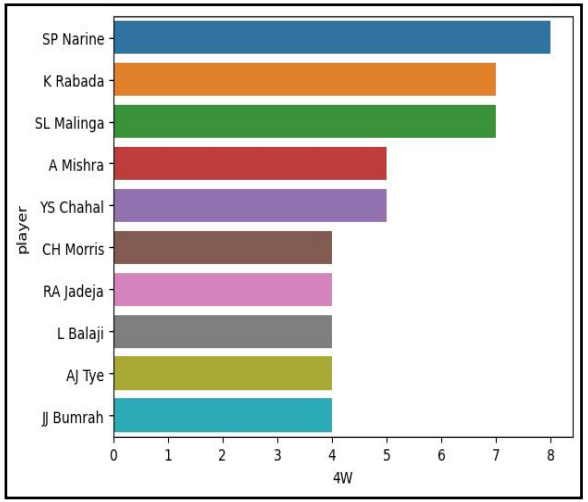
Bowler Analysis



- Highest number of wicket taken by an individual player in IPL



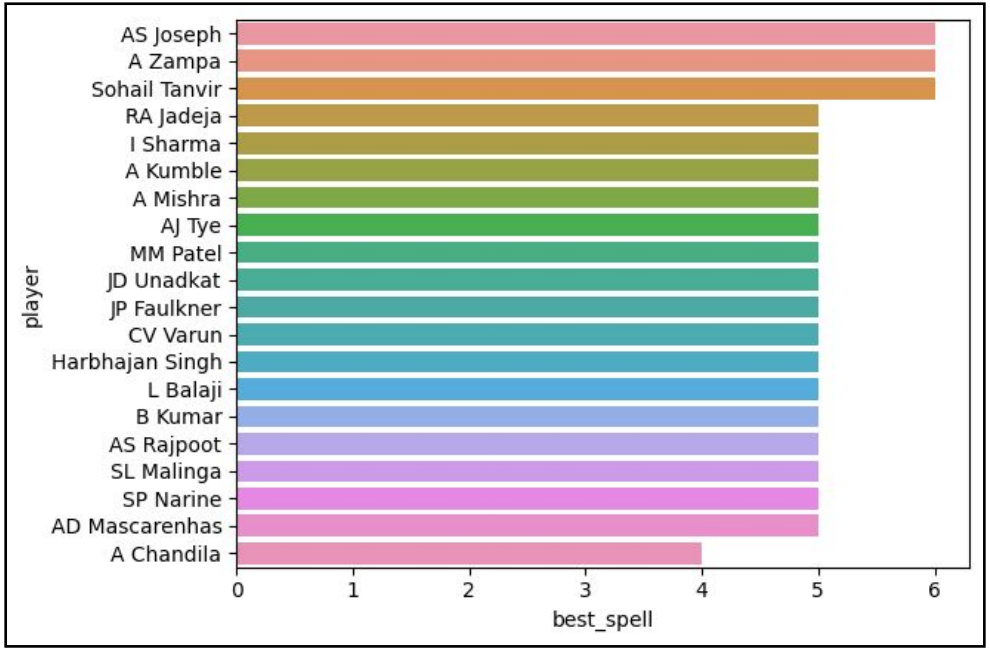
- Highest number of bowl bowled by an individual player in IPL



- Highest number of 4 wicket howl by a bowler in IPL

Bowler Analysis

- This plot shows us the best performance of bowlers in each match sorted by number of wickets in the spell. The best performing bowlers in this case are AS Joseph, A Zampa, and Sohail Tanvir.

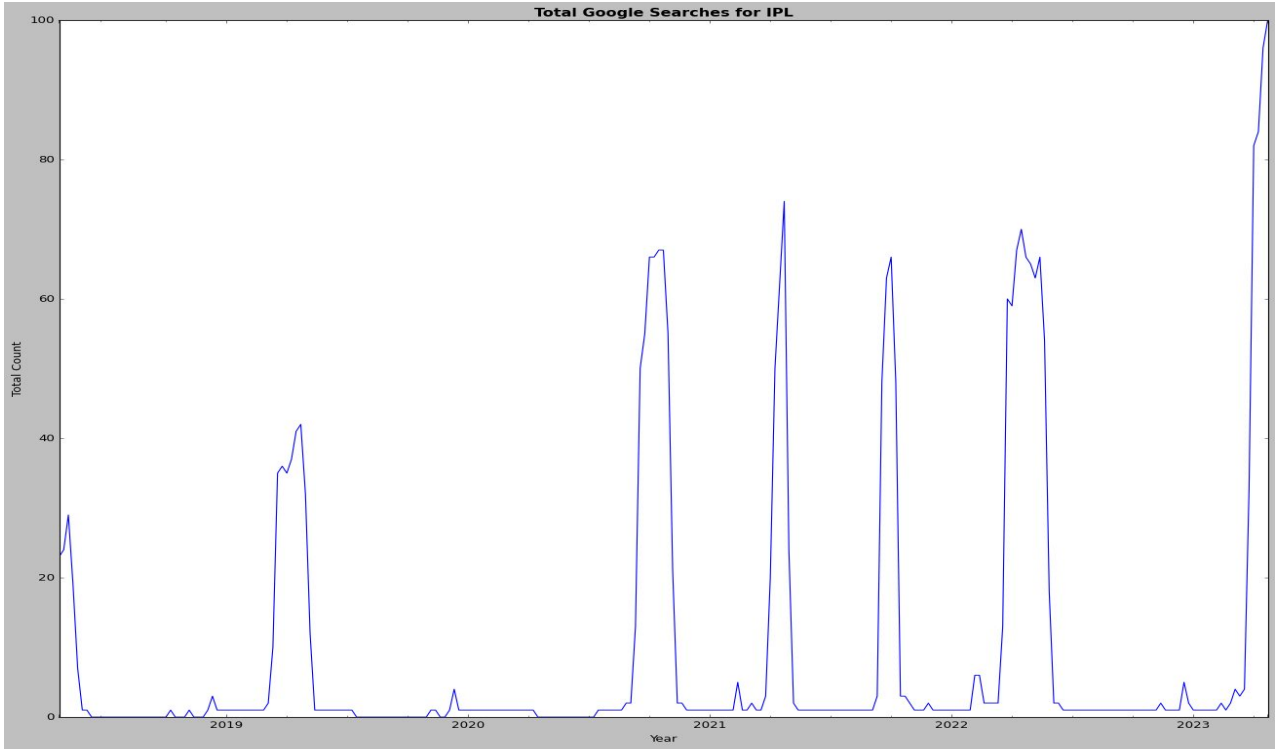


Trend Analysis

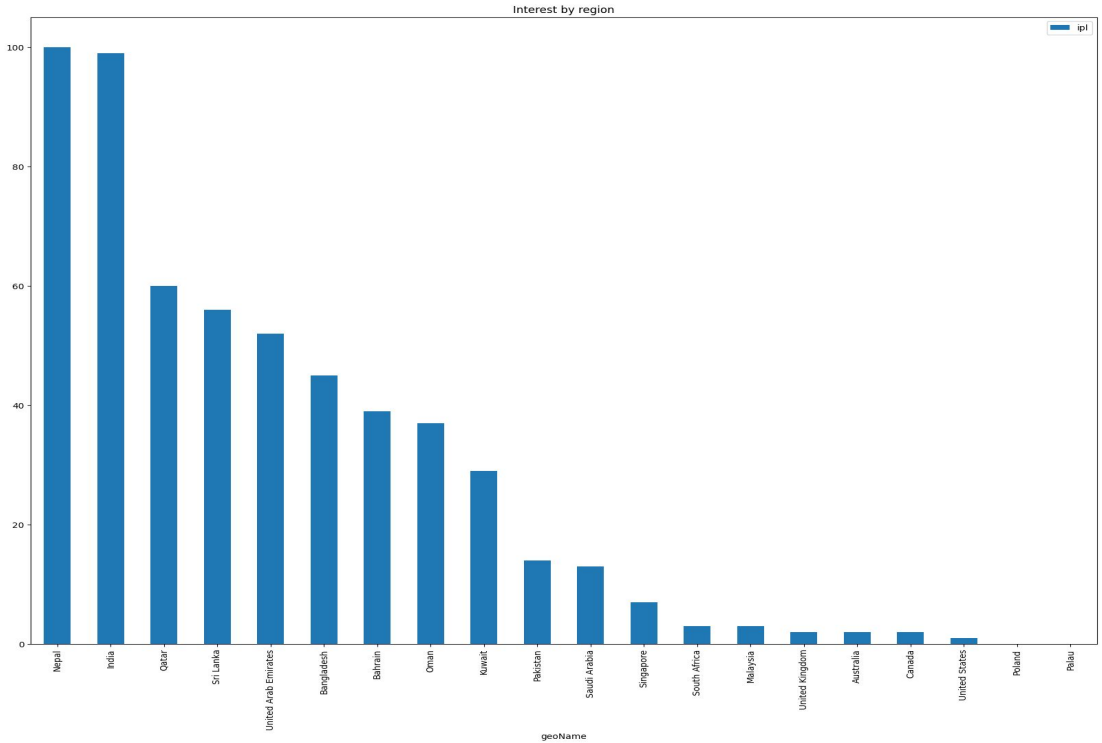
...

Trend Analysis

- This plot shows the trend of Google searches for IPL over time using interest_over_time( ) method.
- As we can see, the searches were low in 2019, but have increased significantly nearing 2023
- Also there is a delay in 2020 as IPL was postponed.



Trend Analysis



- This bar graph represents the relative popularity of search for the word “IPL” on the Google.
- For plotting the above graph, we have extracted the information using trends.interest_by_region( ) method which searches by region

Machine Learning

...

Here we have applied two machine learning algorithms.

1. K - mean clustering algorithm

2. Linear Regression

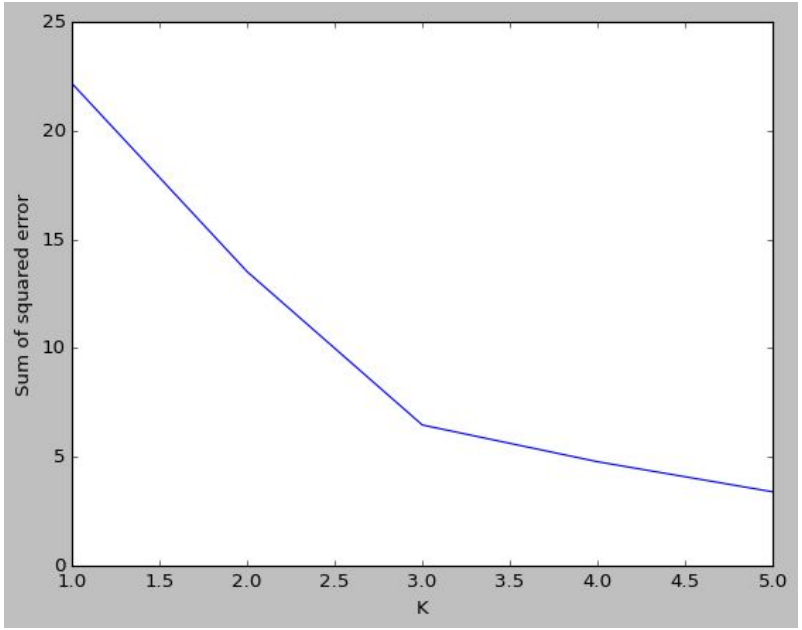
3. Kth nearest neighbor

- K - Means clustering is an unsupervised machine learning algorithm where the data is divided into k clusters or groups having same characteristics. Aim is to minimize the sum squared distance between object and it's assigned cluster centroid.

- Linear Regression is a supervised machine learning algorithm where the linear relationship between independent variables and dependent variables is taken into consideration to predict the outcomes.
- A best fit line which has the minimum sum squared error between original and predicted point is calculated and makes predictions based on that line

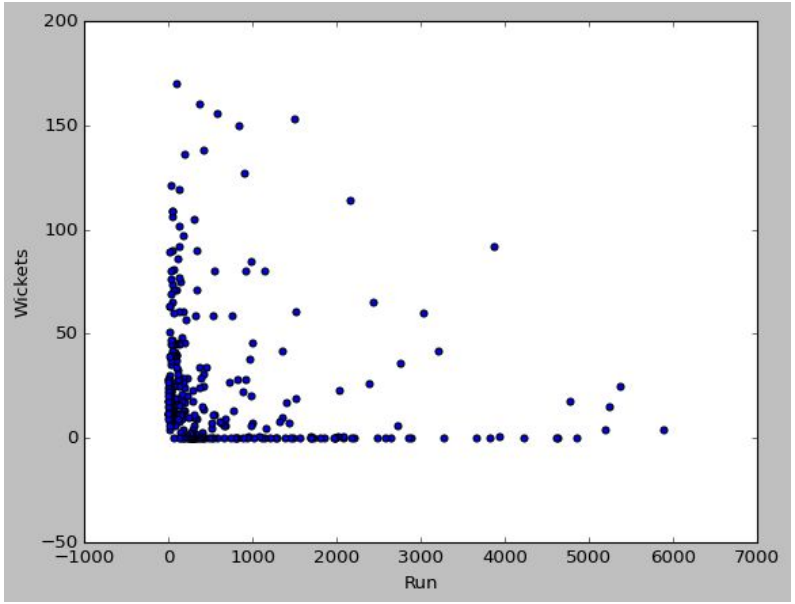
- KNN is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.
- K in KNN defines how many neighbors will be checked to determine the classification of a specific query point

K - Means Clustering

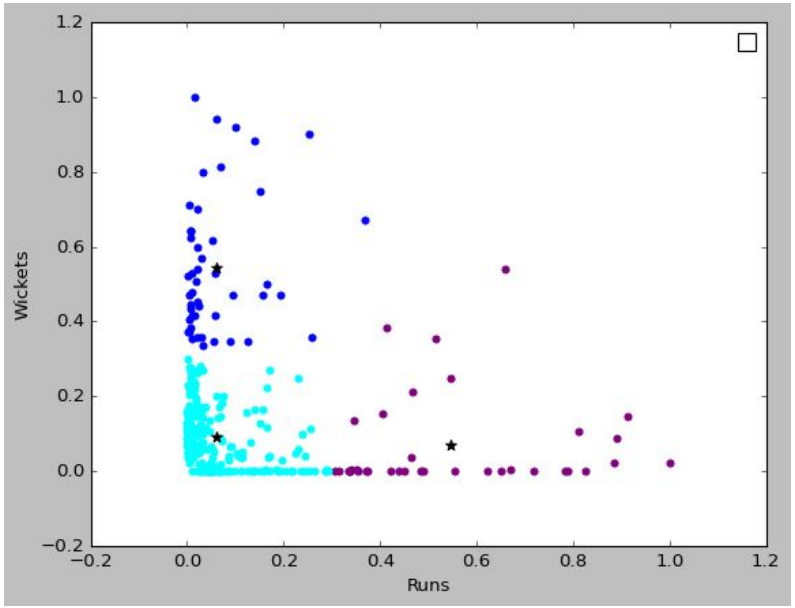


- ➔ To find the optimal number of clusters, we use a method named as ‘Elbow method’.
- ➔ In elbow method we calculate SSE (within cluster sum of square error). As we can see in graph, as we increase number of clusters, there is a decrease in SSE.
- ➔ When there is no major change in SSE value or we can say when line is becoming almost parallel to x axis, we take that k value as optimal number of clusters.

K - Means Clustering



➔ Scatter Plot between Run scored vs Wickets taken by player



➔ Scatter Plot between Run scored vs Wickets taken by player after applying K - Means Clustering.

Kth Nearest Neighbours

- KNN is a model used for classification as well as regression
- We have assumed the 3 clusters formed by the k-means algorithm as 3 classes
- Number of neighbours used = 5
- Number of input features = 8, including scored_runs, wickets_taken, batting_innings, etc.
- Testing set size = 20%
- Distance measure : Euclidean (L2 Norm)
- Following slide shows performance measures:

Kth Nearest Neighbours

- Confusion Matrix

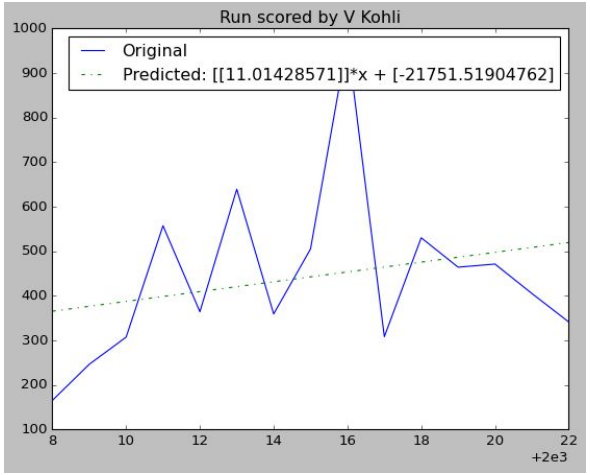
Actual / Predicted	Class 1	Class 2	Class 3
Class 1	9	0	0
Class 2	0	23	3
Class 3	0	1	31

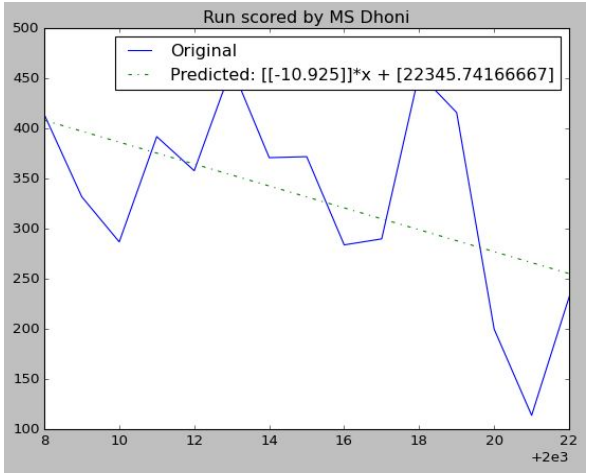
- Result of Classifier

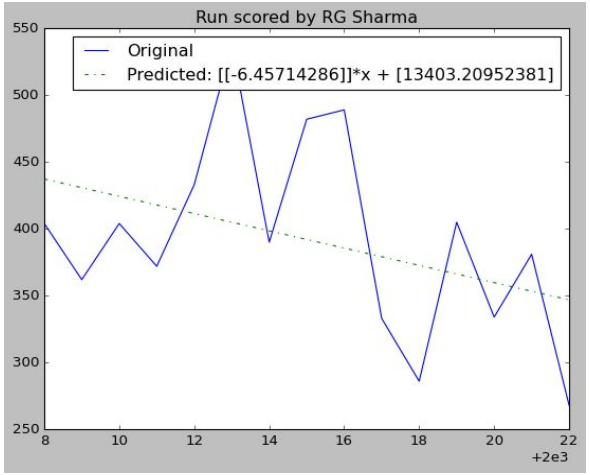
Class	Accuracy	Precision	Recall	F1-Score
1	0.9718	1	0.8889	0.9412
2	0.9718	0.9737	0.9737	0.9737
3	0.9718	0.9167	1	0.9565

- We can see 4 erroneous classifications, 3 for class 3 and 1 for class 2

Linear Regression







- Here we are predicting the run score by any player, given past performance in IPL. We are taking ‘year’ as an independent variable and run_scored as a dependent or target variable and trying to predict the score of the player using linear regression