Term Project Report on

Indian Premier League (IPL) – All Seasons overall stats (EDA Analysis)

Section-1: Overview The Indian Premier League (IPL) is a men's professional T20 franchise cricket league in India. In 2007, the Board of Control for Cricket in India (BCCI) established the league. The league began in 2008, with eight clubs named after various Indian cities and states. It is held once a year. Later in the league's evolution, more teams were added.

The Indian Premier League is a professional Twenty20 cricket league in India. The league features ten franchises representing various Indian cities and states. Every year, the IPL features several international and domestic players and takes place between March and May.

Our project is Mainly based on data for each ball for the entire tournament from **2008 to 2022.** The data will be updated every day until the season's final match. The match ID is given by ESPN Cric info and is also used by other datasheets and acting as the primary key. Along with that we are adding the data of each match result in detail held from the inauguration of IPL start date to till date. So this EDA will provide us to visualize the overall stats of each player and keep track of them in scoring number of runs per ball and wickets taking by the bower as well. The data which we have created for our project consists of information regarding all players which are of: no. of wickets taken by a bowler, no. of wide deliveries, no. of runs conceded -Scored, no. of sixes conceded-Scored, no. of fours conceded-scored, dot balls bowled, no balls bowled, economy rate of the bowler, no. of overs bowled, no. of maiden overs bowled, season of the match in which the bowler is playing, team for which the bowler is playing currently, name of the opposition team, bowling team, venue of the match, name of the country in which the match is being played, innings in which the bowler is bowling, whether the bowler is captain for that particular match or not, career span of the bowler and total no. of matches played by the bowler in his career, no.of matches played by a batsmen in his career.

NOTE: In above overview Consided represents – bowler activity where as Scored represent batsmen stats.

Section-2: Dataset

The dataset we developed for our research has 2,26,906 instances (rows) and 36 characteristics (columns). Instances are made up of numerous data types such as integer, float, Boolean (Yes /No), and string. All of these examples correspond to the matching values of 21 characteristics. The data type of the instance will be determined by the feature to which it corresponds.

The dataset has the following dimensions: **2,26,906 * 36**

As previously stated, there are 36 distinct sorts of features associated with overall IPL seasons ball by ball stats. The following is a full description of each feature:

ID: which is nothing but match ID (number)

Innings: Innings number (1st innings or 2nd innings) [i.e each match has 2-innings]

Overs: over number (each innings has 20 overs 0.1 to 19.6)

Ballnumber: Ball number

Batter: Name of the batter playing as a Striker (Position no #1)

~~Bowler~~ Name : Name of the bowler who is bowling

non-striker: Name of the batter playing as a Non-Striker (Position no #2)

extra_type: Extra type may be – wide, No-ball, Free-hit.. etc…

batsman_run: No.of runs scored by a batsmen

extras_run: Extra runs may include Byes, Leg-byes, over throw etc…

total_run: total runs scored per delivery and over

non_boundary: keep track of boundaries per delivery

isWicketDelivery: Out or Not-out

player_out: Name of batsmen who got dismissed.

dismissal_kind: It tells how the batsmen got out like caught-out, Run-out, Bowled, LBW etc…

fielders_involved: Name of the fielder involved in dismissal kind.

BattingTeam: Tells the name of team who is batting.

City: match held city-Name

Date: Date of the Match

Season: Season number

MatchNumber: Match Number

Team1:Home team

Team2: Away team

Venue: Venue (stadium name)

TossWinner: Name of the team that won toss

TossDecision: Elected to field/bat _Captain decision

SuperOver: Super over held or not (Yes/No)

WinningTeam: Team won -Name

WonBy: type of winning – runs/wickets/DLS

Margin: Related to runs and wickets that a team won by…

Method: Method -NA

Player_of_Match: Player of the match-Award winner

Team1Players: Team#1 line-ups after toss

Team2Players: Team#2 line-ups after toss

Umpire1: Name of ON-field umpire #1

Umpire2: Name of ON-field umpire 2

Name of the bowler/batsmen, no. of wickets taken by the bowler, no. of runs conceded by the bowler, no. of runs scored by the batsmen economy rate of the bowler, strike rate of batsmen dot balls bowled by the bowler, no balls bowled by the bowler, season of the current IPL match and bowling team, batting team are some of the important features of our dataset. Similarities and differences in instances can be found like, if the same player has played for two different teams then same name of the player can be identified.

Brief:

These are some of the significant characteristics that may be present in an IPL dataset. Other variables that may be included in the dataset include the names of the players who batted and bowled, the number of runs and wickets taken by specific players, and other match statistics. Each instance or row in the dataset represents a single IPL tournament match. The features for each match would be recorded in the dataset's columns. The similarity across cases is that they all represent an IPL tournament match. The distinctions across instances would be determined by the different teams involved, the various venues, the various players, and the various match outcomes.

Section-3: Exploratory Data Analysis (EDA) Graphical Insights

We have performed various operations on our dataset. At first we have tried all the basic operations like head, tail, describe, info, drop etc.. and then we started web scrapping the data form espncricinfo [https://www.espncricinfo.com/ci/engine/series/313494.html?view=records] we tried to find relation among various features like innings, overs, ballnumber, batsmen_run, extra_runs, total_run, non_boundary, iswicketdilevery . We found that there is a string relation between iswicketdelivery and no.ofballs/overs bowled. That tells us that if a bowler bowls more overs/balls then there is a higher probability of taking a wicket. Regarding this data we have generated the heatmap before merging the dataset as well as after merging the dataset.

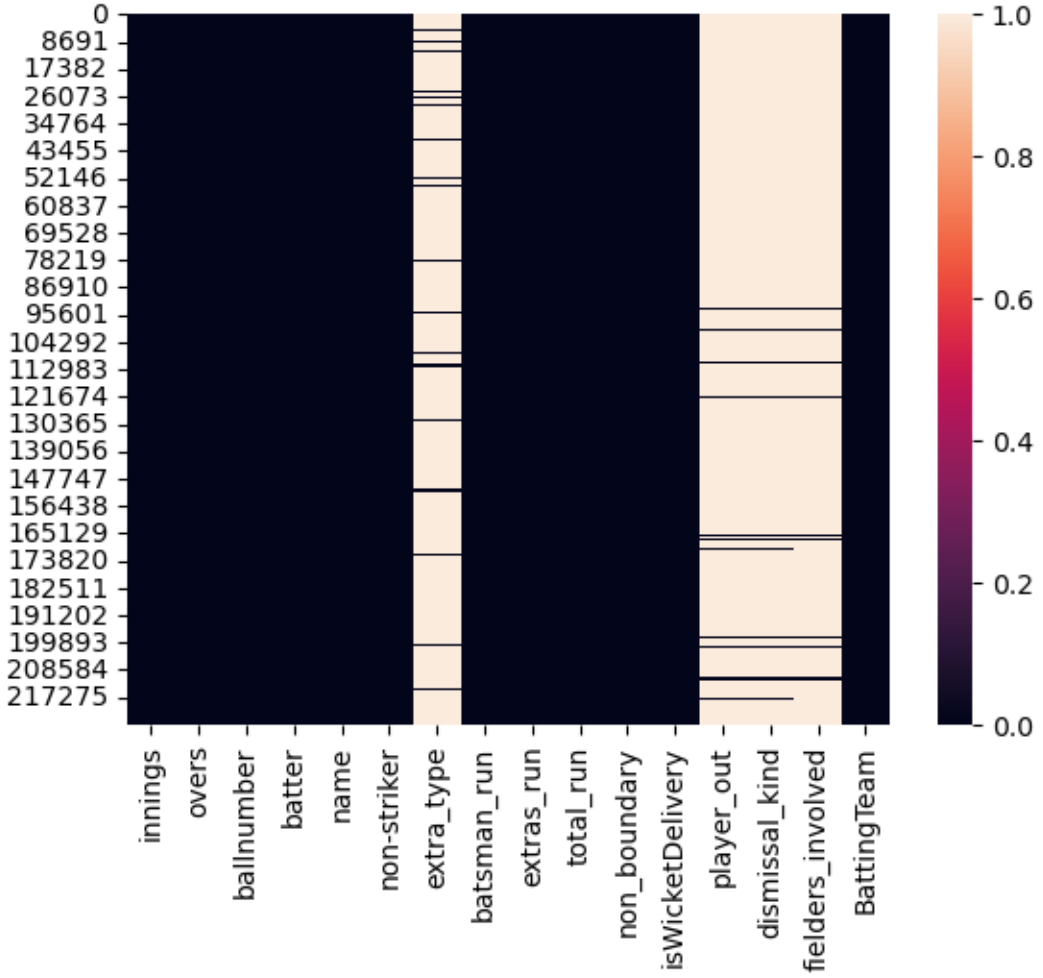
Brief :

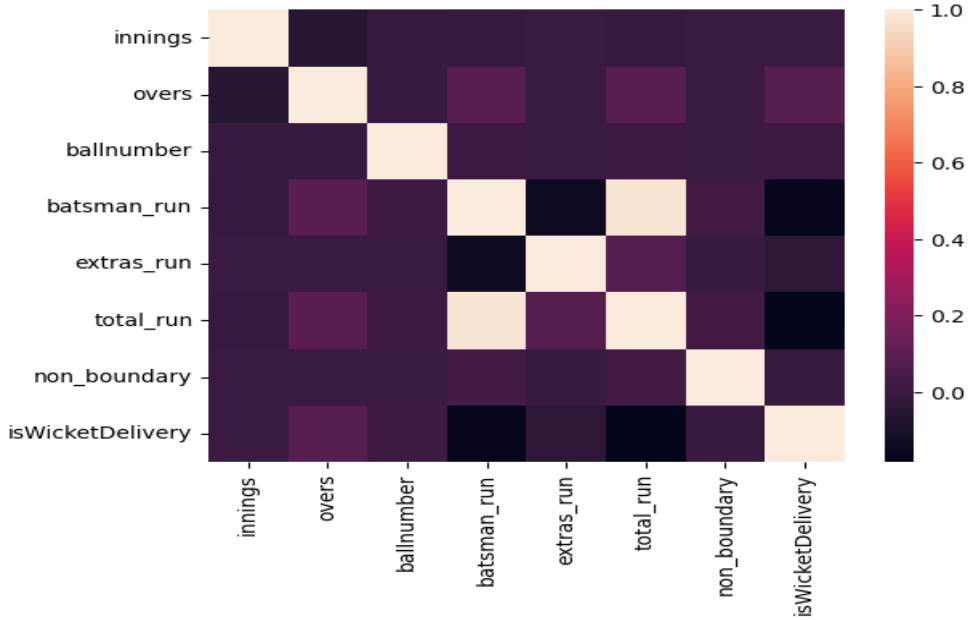
The combination of ball-by-ball IPL datasets with ESPN Cricinfo bowler analytics can provide important insights into bowler performance in IPL matches. By merging these two datasets, one can examine bowlers' performance against different batsmen, in different venues, and at different periods of the game on a ball-by-ball basis.

The ESPN Cricinfo bowler stats provide information about a bowler's career statistics such as their economy rate, average, and strike rate. By merging this data with the ball-by-ball IPL dataset, we can analyze how a bowler has performed in IPL matches, how they have adapted to different conditions, and how they have fared against different teams and batsmen.

This analysis can be used by team coaches and analysts to make data-driven decisions about team selection, game strategy, and player development. It can also be used by fans and media to gain insights into the performance of their favorite bowlers and teams. Overall, merging ball-by-ball IPL data with ESPN Cricinfo bowler stats can provide a more comprehensive understanding of the performance of bowlers in IPL matches.

The correlation between above mentioned features can be viewed below:

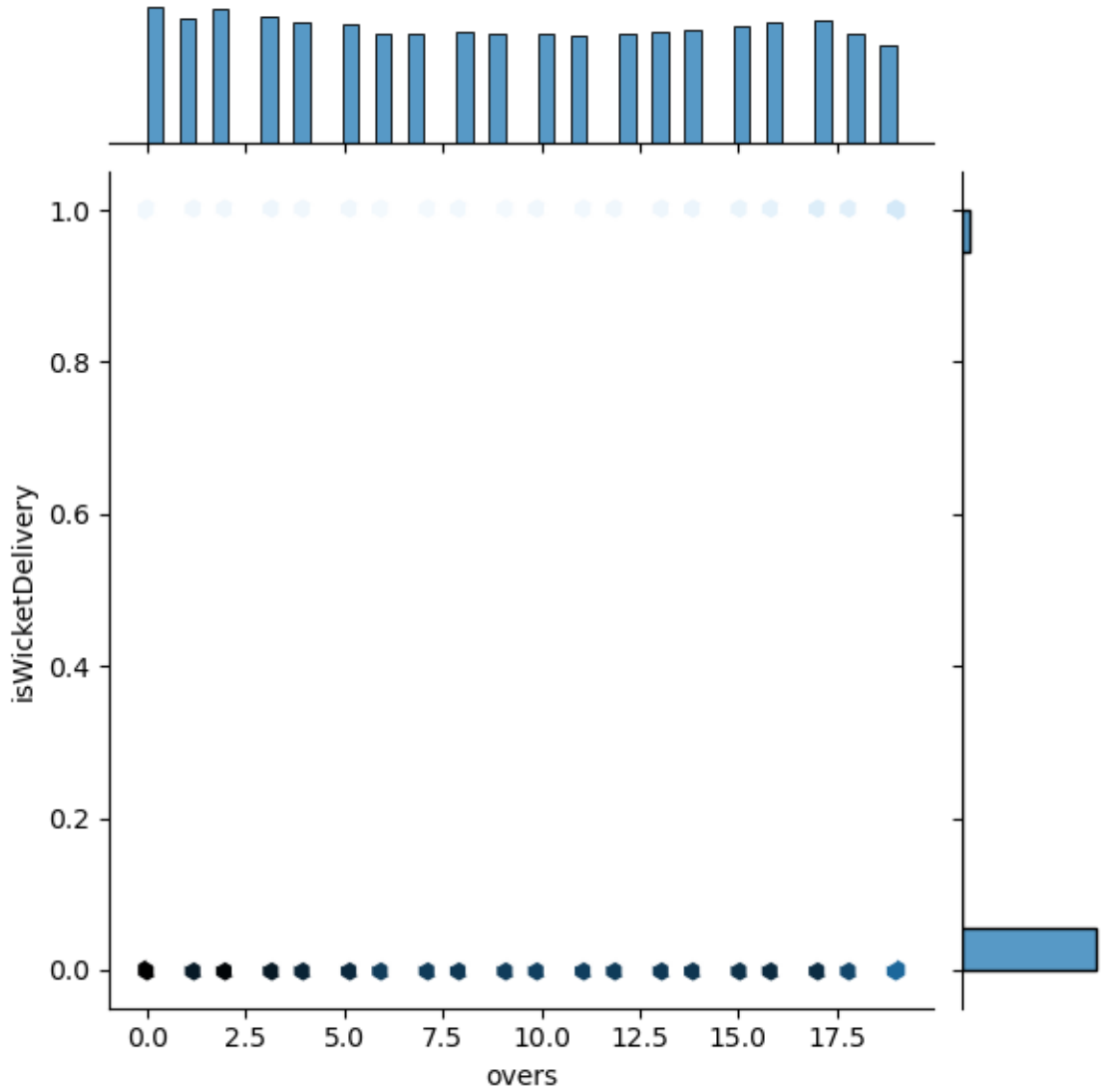




We also observed that if a bowler bowls more Overs then the probability for him of taking wickets increases. This is because they are bowling more deliveries, and each delivery presents a chance for a wicket.

However, it is crucial to note that there are other other elements that can influence a bowler's ability to take wickets. These considerations can include the quality of the opposing batting lineup, the pitch and weather conditions, the captain's fielding placements and tactics, and the form and talent of the bowler themselves.

As a result, while the number of overs bowled is an essential component in a bowler's ability to take wickets, it is not the only determinant. When examining a bowler's performance and forecasting their future success, it's critical to take into account all essential elements.

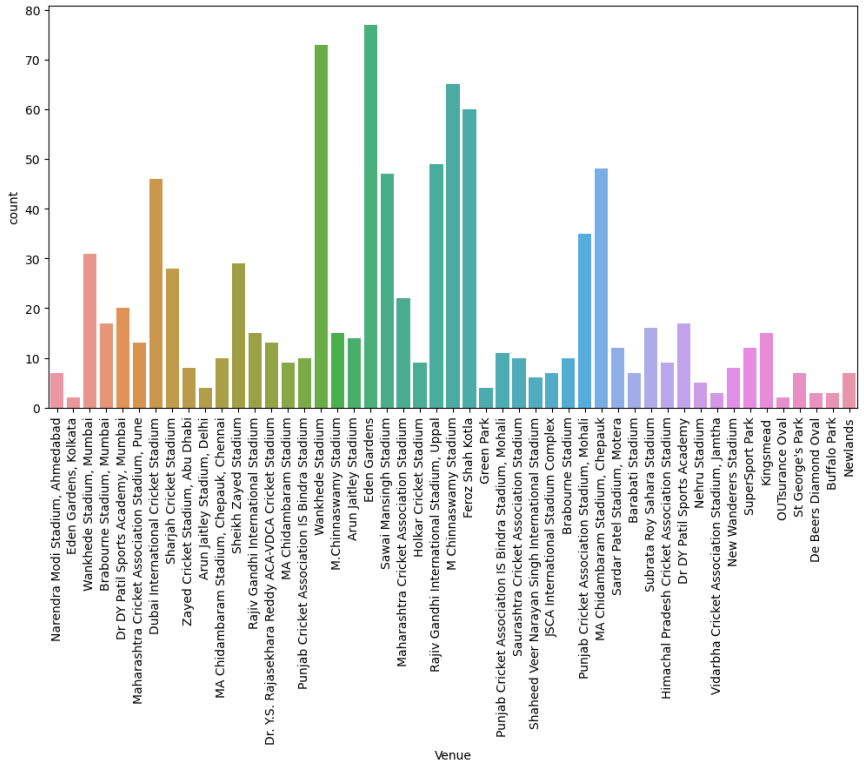


Analyzing the number of matches held in different venues can provide valuable insights into various aspects of the IPL tournament. Some of the benefits of analyzing the number of matches held in different venues are:

 Identifying popular venues: By analyzing the number of matches held in different venues, we can identify which venues are more popular among the teams and the IPL audience. This information can be useful for scheduling future matches and deciding which venues to prioritize.

 Evaluating team performance: The number of matches played by a team in a particular venue can provide insights into their performance in that venue. Teams that perform well in a particular venue are likely to have a better chance of winning matches played in that venue.

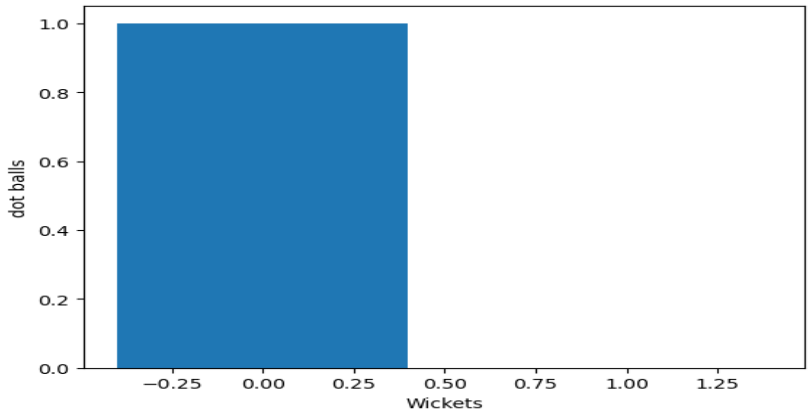
 Understanding home advantage: Teams that play their home matches in a particular venue may have a home advantage due to factors such as familiarity with the pitch and the support of the local crowd. Analyzing the number of matches played by teams in different venues can help us understand the extent of this home advantage. Etc….



We have also tried to check the relation between all the bowlers in our dataset with all possible features. So that we will get an idea of how much strong relation exists between bowlers and other features.

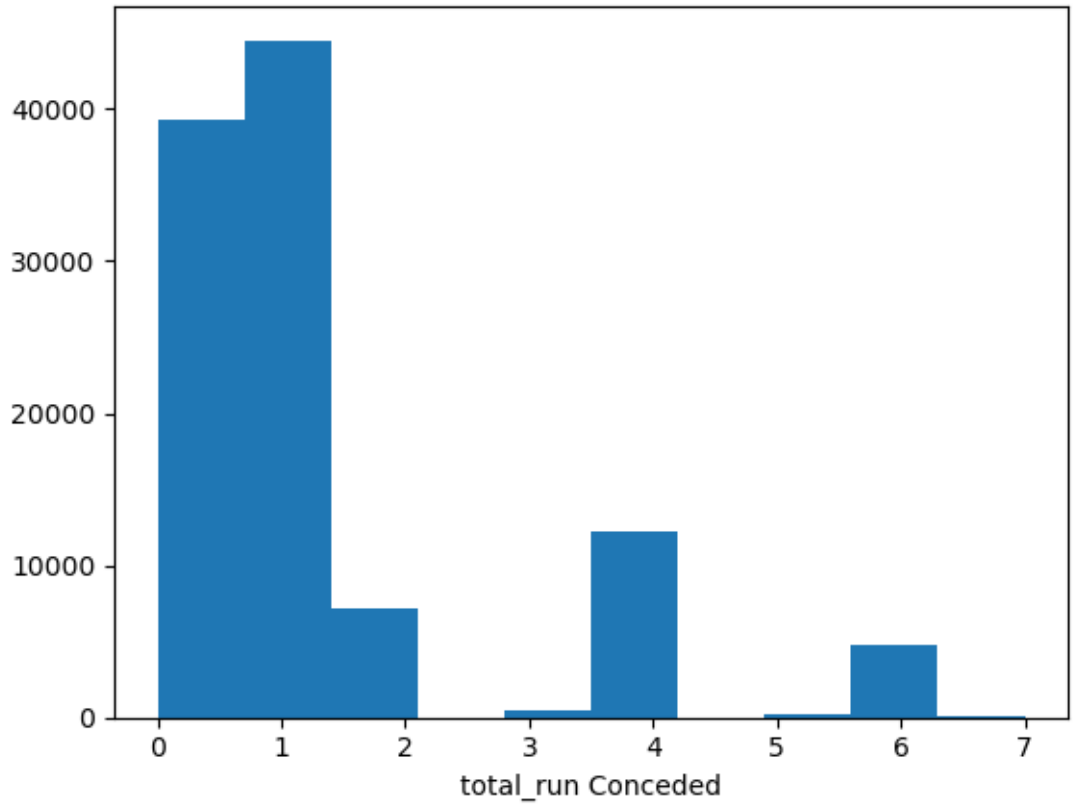
Analyzing the relation between all the bowlers in the dataset with all possible features can provide insights into the factors that influence a bowler's performance. By examining the strength of the relationships between bowlers and other features, we can identify which factors have the greatest impact on a bowler's performance and develop strategies to improve their performance.

We've observed the influence of no. of wickets taken by a bowler on bowling dot balls. There is a higher chance of bowling dot balls when a bowler picks up less no.of wickets and the probability of bowling dot balls decreases slightly when a bowler takes more wickets. Surprisingly, if a bowler has taken no wicket then there's less chance of him bowling dot balls. This observation can be viewed below:



For a bowler conceding More no. of runs (Includes 6s ,4s) is the worst thing to happen. We have observed that large no. of bowlers didn't concede more than a single six. As the count of conceding sixes increases, the no. of bowlers in that category decreases.

Furthermore, we can discover any outliers or extreme results by evaluating the distribution of total runs conceded. These outliers may signal outstanding bowler performances or faults with their bowling strategy that must be addressed. We may make informed selections regarding which bowlers to select for upcoming matches and which methods to utilize if we understand the distribution of total runs conceded.



Section 4: CONCLUSION

In conclusion, merging datasets can be a useful technique in IPL EDA to gain more insights and analyze the data more effectively. By combining two or more datasets, we can identify patterns and relationships that may not be evident when analysing the data separately. In the case of IPL datasets, merging different datasets such as player performance data, team data, match data, and toss data can provide a better understanding of the game and help us identify trends and patterns in the data.

By analysing the merged data, we can draw meaningful insights that can help teams make more informed decisions on team selection, strategy, and game planning. For instance, we can analyze the performance of players in different teams, identify the impact of toss on team performance, and evaluate the impact of weather conditions on match outcomes.

Therefore, merging datasets is an essential technique for IPL EDA, as it helps us to better understand the data, identify insights, and make more informed decisions.

Brief:

Overall, using our dataset we have performed Exploratory Data Analysis (EDA) and pulled some useful insights and made it ready for an ML algorithm. Some visualizing insights from graphs and plots are very useful to estimate the predicted value. Few data insights were very precise and gave us few observations which we didn't even think of. The dimensions of the dataset was descent to know significant information about bowlers. Using this dataset it can be predicted that what is the performance of a particular bowler is going to look like in future, and few other predictions which are related to bowling statistics.

Appendix A Data Collection and analysis of original dataset

We have gathered the information from many sources, like Kaggle and cricket websites. The initial (original) dataset, which we obtained from Kaggle, contains information about overall metrics(stats) of ball by ball in all IPL seasons from 2008 through 2022- also combined with matches going on current season (2023)

Source of the original dataset: https://www.kaggle.com/datasets/vora1011/ipl-2008-to-2021-all-match-dataset?select=IPL_Ball_by_Ball_2008_2022.csv

Current season auto updated through espncricinfo.in.

The original dataset consists of large no. of instances (rows) and features (columns). First just to know the initial data we've performed '.head()' which resulted in top 5 rows of the dataset. To know more information in depth of the original dataset, we've performed '.info()' operation

on it which resulted in total entries (rows) and column names. The total entries were 2,26,906(rows) and 36 columns.

Shape of the dataset is: **2,26,906 * 36**

There were a total of 36 columns in it which all are related to all batting, bowling, venue, scores etc... It consists of few features like home team, away team, match id, season, bowling team, venue, city, state, innings id, runs conceded and few other columns which are statistically used for bowling analysis. Apart from that, to view the statistical analysis of the original dataset, we've utilized '.describe()' operation which resulted in min, max, standard deviation, mean, 25th percentile, 50th percentile and 75th percentiles of specific features.

We just want to concentrate on the parameters (columns) which are closely related to and have an impact on bowling. So, we want to get rid of few columns.

Data Cleaning on original dataset

We've dropped few columns which were we thought not useful for our main category 'name' which represents name of the bowler. '.drop(columns names, axis=1,iplace=True)' was used to drop specific columns which are ['Inns', 'Overs', 'Mdns', 'Runs', 'Wkts', 'BBI', 'Ave', 'Econ', 'SR', '4', '5']. Here, 'axis=1' specifies dropping columns and 'inplace =True' indicates to reflect this operation on original dataset. After dropping five columns, the original dataset consisted of 29 columns.

To check whether this data is either cleaned or not, we've used 'heatmap' to visualize null values in the dataset. It resulted in zero null values.

Web Scrapping

We have found another 3rd dataset in 'espncricinfo' website. The dataset is about 'bowlers who took most wickets of all IPL seasons'.

URL of the second dataset:

https://stats.espncricinfo.com/ci/engine/records/bowling/most_wickets_career.html?id=117;type=trophy

To web scrape the dataset, we copied and pasted the 'URL' of the webpage link corresponding to the dataset into a variable name. We read the URL in the same way as a CSV file, but to display it as a CSV file, we added '[0]' to the URL read procedure. Because the dataset was in the form of a list when it was shown, '[0]' was used here. So, we simply utilized an indexing operation to obtain the dataset (in the form of a list).

Data Wrangling of second dataset

To gain a general idea of the latest dataset, we just displayed it and noticed that it has 14 columns and 96 rows. The columns correspond to the player's name (bowler's name), the bowler's IPL career span, the number of IPL matches played by the bowler, and a few other statistically connected factors.

We want to combine the first and second datasets depending on the bowler's name. When compared to the original dataset, the column name of the bowler's name in the second dataset was changed. So we renamed it to 'name' by using '.rename(axis=1,inplace=True)', which correlates to the name to which the column name is being replaced and permanently reflects the operation.

Merging of two datasets

We want to merge these datasets based on Players name so that the resulting dataset will have meaning to it. To perform merge, we've utilized '.**merge(two datasets name, on='name')'.** It performs inner merge on two datasets based on one common column 'name' which refers to Player's name.

APPENDIX B: WHAT I DID & LEARNED

Sumanth Reddy Busireddy: In this project, I have learned a lot i.e creating a comprehensive dataset from the ESPN website can present various challenges that require careful consideration and problem-solving skills. One potential difficulty I have faced is the dynamic nature of websites, which may necessitate careful selection of appropriate web scraping tools and techniques to extract data accurately. Another challenge I have faced is data cleaning, which involves addressing missing, inconsistent, or duplicated data, as well as formatting inconsistencies.

Despite these challenges, developing a dataset from the ESPN website can provide us valuable learning opportunities. It also enhanced me in developing web scraping skills, improve data cleaning and pre-processing techniques, increase awareness of legal and ethical considerations in data collection, and foster good data documentation practices. Overcoming these challenges can result in a robust and reliable dataset that can be utilized for various data analysis and research purposes. Additionally, navigating the complexities of web scraping and data collection can build expertise in dealing with real-world data challenges and provide valuable experience for future data projects. I learned a lot from the 510 class which helped me in better understanding of data visualization. I am so happy that I took this course which challenged me but taught me a lot which can improve my career further.

Sri Swetha Kanduri: In this project, I worked in collaboration with my teammates in finding the datasets required and adding a few columns to the datasets manually for our EDA. Apart from this,

I've performed programming operations on data cleaning, data merging, and data wrangling. I took part in performing EDA and data manipulation. I also participated in project report creation.

While doing the project I have put my theoretical knowledge taught in class to practice. Deeply understood concepts of data visualization. Understood how to analyze the different visual representations. Learned different ways of indexing columns while data cleaning. Understood Data wrangling techniques. Learned how to deal with large datasets as our dataset is huge. Also learned the art of data compression by removing unwanted columns and discarding all null values in our dataset. Learned few logical operations on columns using pandas. This project helped me to understand how to translate user requirements into high-level software and data models In real-time.

Naveen Varma Pandeti :

I helped with the project by looking for the two datasets (which included web scraping) and creating the report. Aside from that, I've worked on data cleaning, merging, and wrangling. I assisted in EDA and data manipulation. Some of my discoveries include:

- ✓ I have learned how to loop through a column to acquire insights.
- ✓ I have learned a few logical operations on columns with pandas.
- ✓ Learnt how to merge two different datasets based on various types of merging operations.
- ✓ Learnt new visualization techniques in seaborn.
- ✓ Also I've learned how to webscrap the url data into heat maps
- ✓ Got to know what different operations we can perform while scrapping the data.
- ✓ Learnt removing warning messages while performing data visualization

Finally, I have gained a range of technical and soft skills by involving in this project. These skills can be valuable in a variety of settings, from data analysis and research to business and management. So, I Keep learning and applying these skills in future projects.

APPENDIX C: REFERENCES:

https://www.kaggle.com/datasets/deepcontractor/ipl-2021-ball-by-ball-dataset

https://www.kaggle.com/datasets/vora1011/ipl-2008-to-2021-all-match-dataset?select=IPL_Matches_2008_2022.csv

https://stats.espncricinfo.com/ci/engine/records/bowling/most_wickets_career.html?id=117;type=trophy

https://www.espncricinfo.com/ci/engine/series/313494.html?view=records

https://youtu.be/oRCAL14w1zs

https://youtu.be/-o3AxdVcUtQ