DSC 465 – Data Visualization

Project Report

Group Name: Indian Premier League

Group Members:
Yashkumar Rajubhai Prajapati
Pradeep Shane
Kothari Ayush Sachin
Shiv Gandhi

Introduction

About Data

The data consists of two major datasets from the Indian Premier League, which is a professional Twenty20 cricket league in India. The datasets include detailed information regarding match deliveries and general match summaries. Such datasets provide a comprehensive view of each match and performances of every individual and overall team statistics. The time range of this dataset is 2008 to 2023.

❖ **deliveries.csv**

→ Numerical:

- over: Integer (Over number)
- ball: Integer (Ball number within the over)
- batsman_runs: Integer (Number of runs scored by the batsman on this delivery)
- extra_runs: Integer (Number of extra runs)
- total_runs: Integer (Total runs scored on this delivery)

→ Categorical:

- match_id: Integer (Unique identifier for each match)
- inning: Integer (Inning number, 1 or 2)
- batting_team: String (The team currently batting)
- bowling_team: String (The team currently bowling)
- batter: String (Name of the batsman facing the ball)
- bowler: String (Name of the bowler delivering the ball)
- non_striker: String (Name of the non-striking batsman)
- extras_type: String (Type of extra runs, if any)
- player_dismissed: String (Name of the player dismissed, if any)
- dismissal_kind: String (Type of dismissal, if any)
- fielder: String (Name of the fielder involved in the dismissal, if any)
- is_wicket: Integer (1 if a wicket fell on this delivery, 0 otherwise)

❖ **matches.csv**

Numerical:

- id: Integer (Unique identifier for each match)
- result_margin: Integer (Margin of victory, number of runs or wickets)
- target_runs: Integer (Target runs set for the chasing team, if applicable)
- target_overs: Integer (Number of overs allocated for the chasing team, if applicable)

Categorical:

- match_type: String (Type of the match, e.g., League, Playoff)
- player_of_match: String (Name of the player awarded as the player of the match)
- venue: String (The stadium where the match was played)
- team1: String (Name of the first team)
- team2: String (Name of the second team)
- toss_winner: String (Team that won the toss)
- toss_decision: String (Decision made by the toss winner, bat or field)
- winner: String (Team that won the match)

- result: String (The result of the match, e.g., runs, wickets)
- super_over: String (Indicates if a super over was played, Y for yes, N for no)
- method: String (Method of result determination, if applicable)
- umpire1: String (Name of the first umpire)
- umpire2: String (Name of the second umpire)

→ Time series:

- ❖ date: Date (The date of the match)
- ❖ season: Integer (Year of the season)

→ Geographical:

- ❖ city: String (The city where the match was played)

→ For creating visualizations in tableau, we have merged two datasets into one dataset called merged.csv using inner join with primary key of match_id (deliveries.csv) and Id (matches.csv).

Objective

Our work aims to provide a simple analysis of the Indian Premier League (IPL). These insights will help stakeholders understand the dynamics of the IPL teams and players.

→ Player Performance: We analyzed IPL player performance through metrics like runs and wickets. Visualizations highlight top performers, providing key insights into individual achievements.

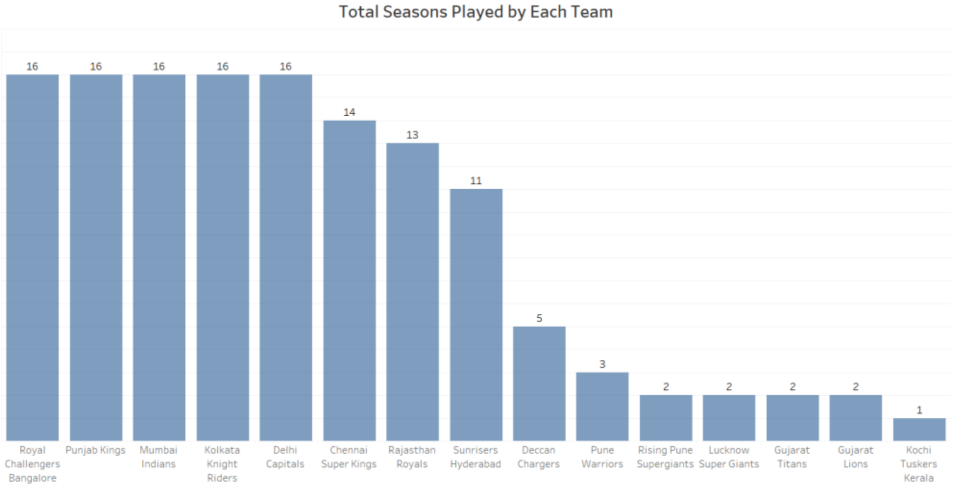
→ Examining Team Performances: We examined team performance by focusing on matches won, total runs, and wickets. Visualizations reveals successful IPL teams, batting and bowling performance and areas of improvement over time

Exploratory Analysis:

1. Bar Chart: Total Seasons Played by Each Team in IPL

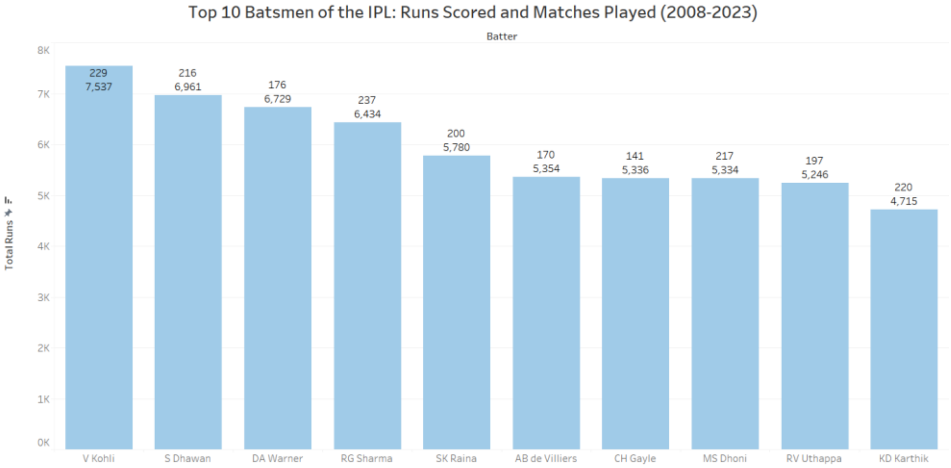
This bar chart is indicating the total number of seasons played by each team from the year 2008 to 2023. This chart is important for the whole analysis and visualization throughout this report because it helps in comparing team performances based on how many seasons they played.

From the chart we can see that Royal Challengers Bangalore, Punjab Kings, Mumbai Indians, Kolkata Knight Riders and Delhi Capitals are the team who have played every season. Other teams have not played every season so while comparing team performance we must take that into consideration.



2. Bar Chart: Top 10 Batsmen of the IPL (By Runs) and Matches Played

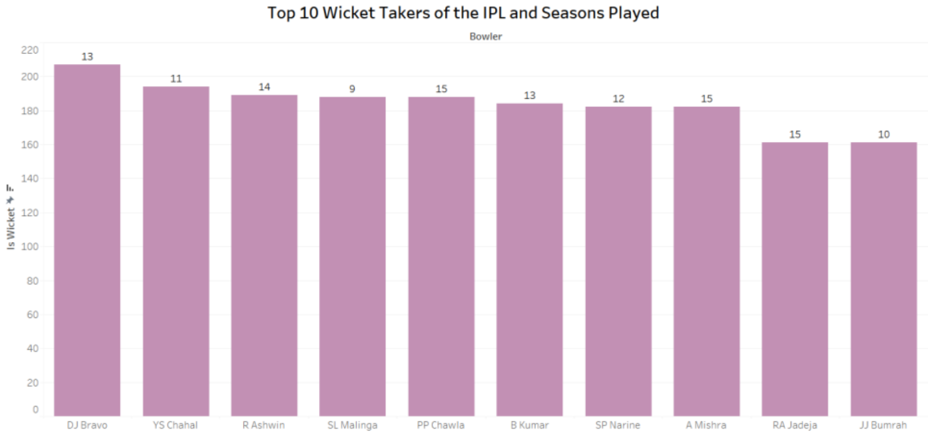
Below bar chart represents Total Runs of top 10 batsmen and how many matches they have played. Total matches played by each player is shown by labelling number on top of the bar. This graph explores performance of batsman across the years.



3. Bar chart: Top 10 Bowlers of the IPL and Seasons Played.

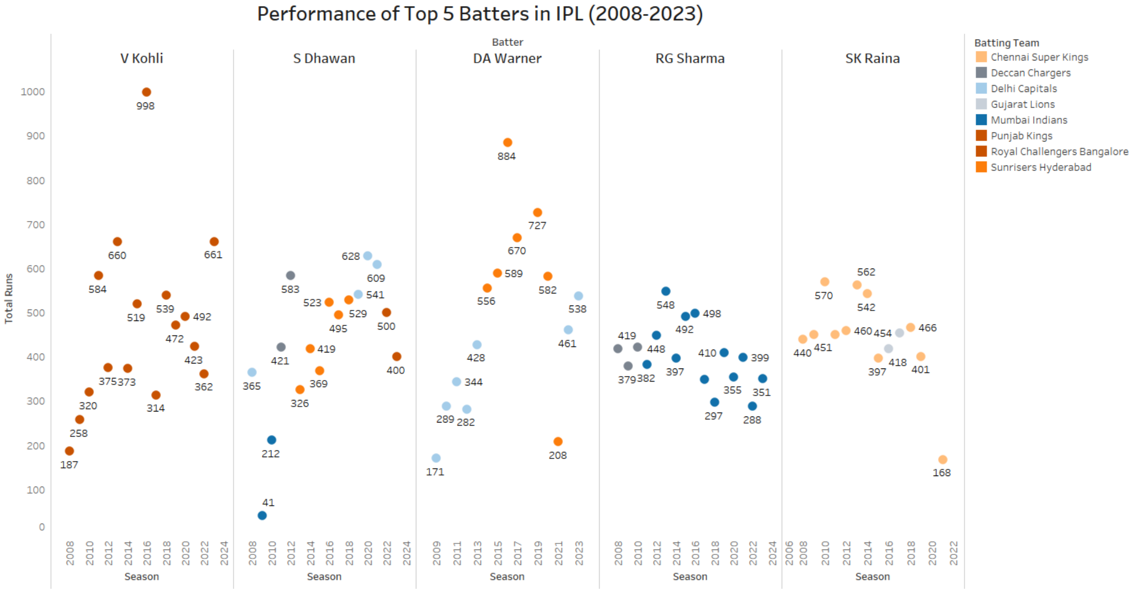
Below chart shows top 10 bowlers(In regards of wickets) and how many seasons they played.

This chart explores variables bowler, Is Wicket and distinct count of seasons.



Player Performance:

1. Scatterplot: Unveiling the IPL's Run-Scoring Titans: A Look at the Top Batsmen (2008-2023)



Message: This visual breakdown celebrates the highest run-getters in the IPL and how they have dominated at the crease in every season from the start of the league in 2008 till 2023. The colourful dots are colour-coded according to the team that batsman represents, so loyalties and rivalries can be identified.

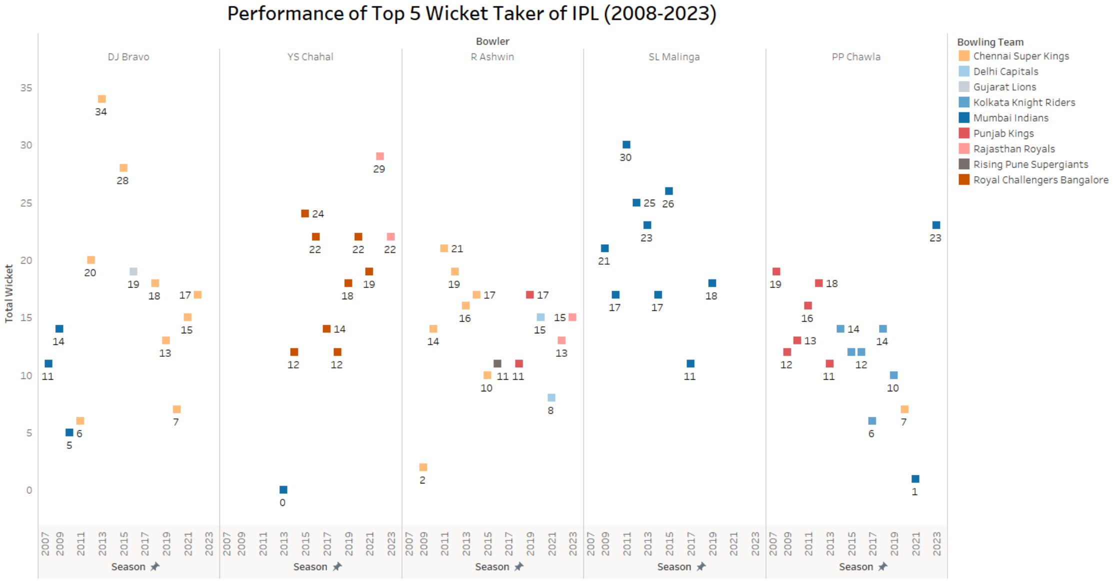
Variables mapping:

- x-axis: season
- y-axis: Total Runs
- label :sum of total runs
- colour: batting team
- sorting: the sum(total runs) by descending order
- Filter: Top 5 (Keep Only)

Refinement Process: Initially we considered the line chart just showing the path how each player has performed. After the feedback we got from the class presentation we improved it to the scatterplot showing the respective team of the players with colours. The colour palette chosen are colourblind friendly, so it remains safe to the colourblind, making it accessible to a broader audience.

Role in Analysis/Story: By tracking a particular data point along the X-axis, you can observe a batsman's run-scoring trajectory over the course of their IPL career. It allows for understanding team-wise representation by denoting colour to identify team affiliations and may point towards underlying trends in player movement across the teams.

2. Scatterplot: Performance of Top 5 Wicket Takers of IPL (2008-2023)



Message: The message conveyed by this scatter plot goes a long way to help the analysis as it enables viewers to see clearly and interestingly the bowling prowess of these greats of the IPL, not just see who the top wicket-takers are, but how this has unfolded over the years and the various teams they have represented.

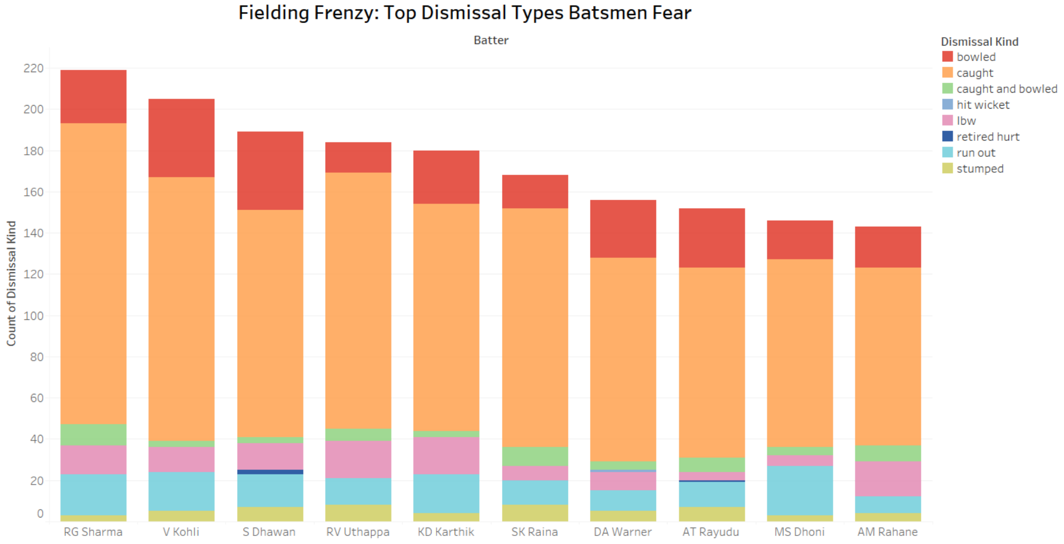
Variables Mapping:

- x-axis: season
- y-axis: is_wicket
- label :sum of is_wicket
- colour: bowling team
- sorting: the sum(is_wicket) by descending order
- Filter: Top 5 (Keep Only)

Refinement Process: Same as the previous char originally, we drafted line chart. However, a scatter plot was most suited for the application of this data set, as it is the most effective way to highlight individual bowler performance by the totality of their wickets and season. This will avoid any confusion that could potentially stem from the creation of overlapping lines if a line chart were used instead. Colours used are colourblind friendly.

Role in Analysis/Story: This scatter plot is pivotal in understanding who the top wicket-takers are in the IPL. By looking at the Y-axis (total wickets), viewers can identify the bowlers with the highest wicket tallies. By following a specific data point across the X-axis, you can see a bowler's wicket-taking performance throughout their IPL career. The colour coding helps identify team affiliations and potentially uncover trends in player movement across teams.

3. Stacked Bar: Dismissal Kind across the IPL Batsmen



Message: This is a stacked bar chart that breaks down the top 10 dismissed batsmen in the IPL across different methods of being dismissed. It allows easy comparison between batsmen and can reveal league-level trends in dismissals.

Variables Mapping:

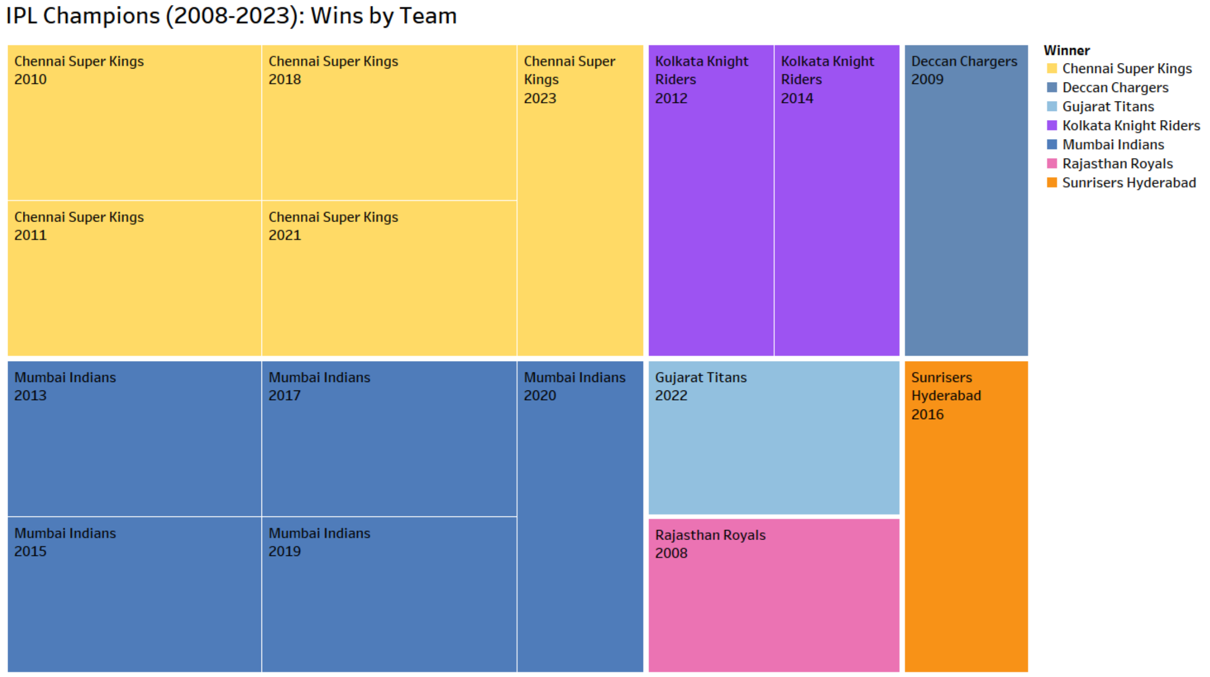
- x-axis: Batter
- y-axis: distinct count of dismissal_kind
- colour: dismissal_kind
- Sorting: distinct count of dismissal_kind with descending order
- Filter: Top10 (Keep Only)

Refinement: Stacked bar graphs are quite good for showing the breakdown of categorical data into its parts. The reasons this chart is a good choice in this case are that the chart shows various types of dismissal and a breakdown for each. Colours used are colourblind friendly.

Role in Analysis/Story: By inspecting the height and composition of each bar, it would allow viewers to understand which batsmen are more prone to specific dismissal types, for example, being caught out or being bowled. This could expose technical flaws in a batsman or areas bowlers might target.

Team Performance:

1. Tree Map: IPL Champions Wins by Team



Description: The image sent is a Tree Map, where each rectangular block represents a franchise that has won an IPL title. The area of each block relates to the number of IPL titles a particular franchise has won. We tried to give the colours of the blocks are of the team's franchise colours.

Message: The Tree Map shows very well and is easy to interpret how the IPL championships are distributed among franchises. The size and colour of each block allow viewers to quickly grasp which teams have been the most successful in terms of title wins.

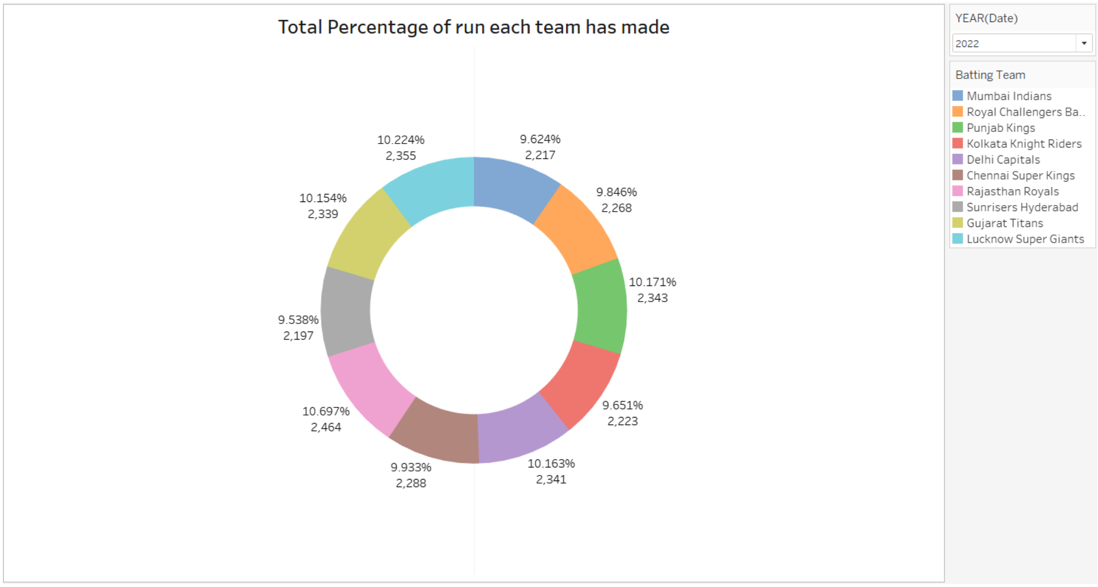
Variables Mapping:

- colour: Winner
- Size: distinct count of match_id
- Label: Winner and Season
- Filter: match_type by final

Refinement: Originally, we used packed bubble chart showing just the total amount of wins by each team. After the presentation we improved it to the tree map showing the year they won. With the actual colours of the team, the visualization becomes more readable and memorable.

Role in Analysis: This Tree Map can be very instrumental in the understanding of the distribution of IPL championships among the teams. It allows viewing people to identify dominant teams. The size of each block allows you to identify those teams that have the most IPL titles.

2. Donut Chart: Total Percentage of run each team has made



Message: The donut chart clearly and intuitively breaks down how the total runs scored in an IPL season are distributed across the teams. This really highlights how each team contributes to the total run count across that season. It shows the batting power of the team.

Variable Mapping:

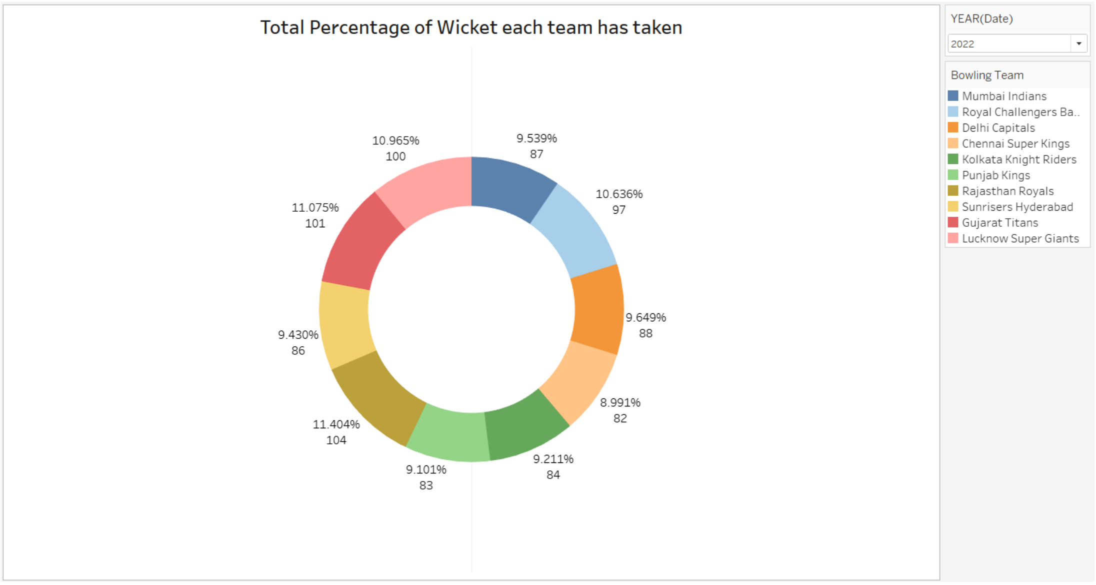
- Colour: Batting team
- Angle: Percentage total of sum of Total_Runs
- Label: Sum of Total Runs and Percentage total of Percentage total of sum of Total_Runs
- Filter: season

Here we have used aggregate of average zero to create inner and outer circle.

Refinement: In the first draft we tried to create small multiples showing each donut for each season however, after looking at the small multiples we found that viewer can not compare size directly because of the compressed view. That is why we tried the filter approach which gives viewer an interactivity to swap between seasons. And gave labels to show total percentages and total runs.

Role in Analysis/Story: This donut chart plays an important role in analysing the contributions of different teams in the context of one IPL season, and it allows the viewer to identify which teams scored the greatest number of runs. From the size of the slice, it can be clearly noticed.

3. Donut Chart: Total Percentage of Wicket each team has taken



YEAR(Date)
2022

Message: This donut chart would provide a clear and detailed description of the distribution of the total number of wickets taken in a season of the IPL among the playing teams. It would indicate the share of each team in the aggregate bowling performance for the concerned season. It shows the bowling power of the team.

Variable Mapping:

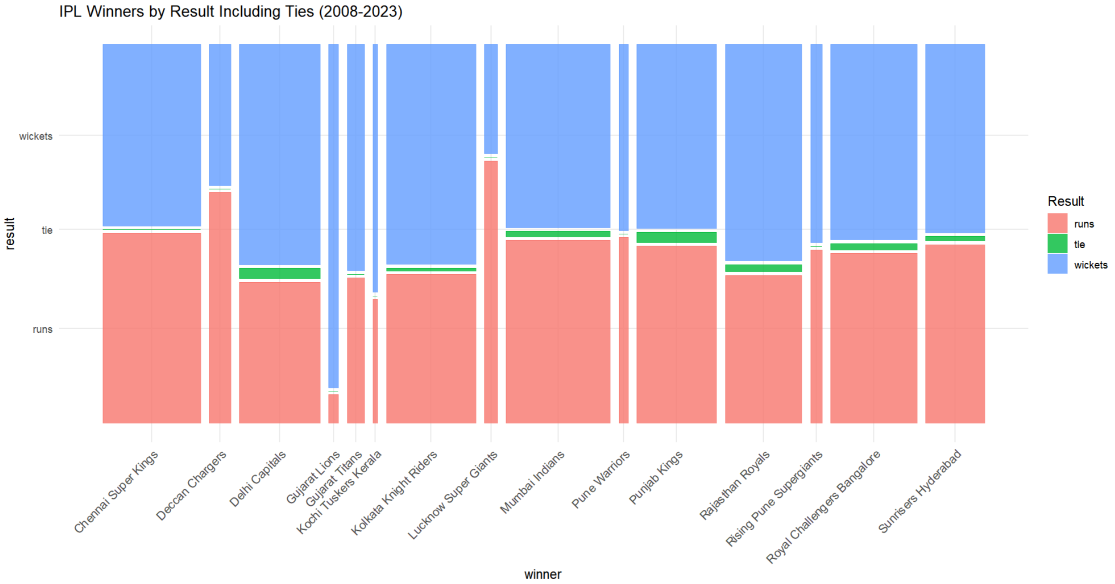
- Colour: Bowling team
- Angle: Percentage total of sum of is _wicket
- Label: Sum of is_wicket and Percentage total of sum of is _wicket
- Filter: season

Here we have used aggregate of average zero to create inner and outer circle.

Refinement: Same as the above.

Role in Analysis/Story: This donut chart would serve the purpose of understanding bowling performances across teams in a single season of the IPL. The size of the slices indicates the high takers of wickets for those specific seasons. Viewers can easily deduce how the various teams have performed in terms of taking wickets during that season of the IPL. By Comparing the donut charts of multiple seasons stake holders can identify if there is a trend in the way wickets are distributed among the teams over IPL seasons.

4. Mosaic Plot: IPL Winners by Result (Created in RStudio)



Message: This mosaic plot communicates quite well the distribution of match wins by IPL teams across different result types and gives one quite a different perspective on how teams fare within the IPL. To see how the IPL teams have won their matches against wins with some result type, viz. runs, wickets, and ties for the seasons. Each tile's area is equivalent to its team's matches won under the stated specifications.

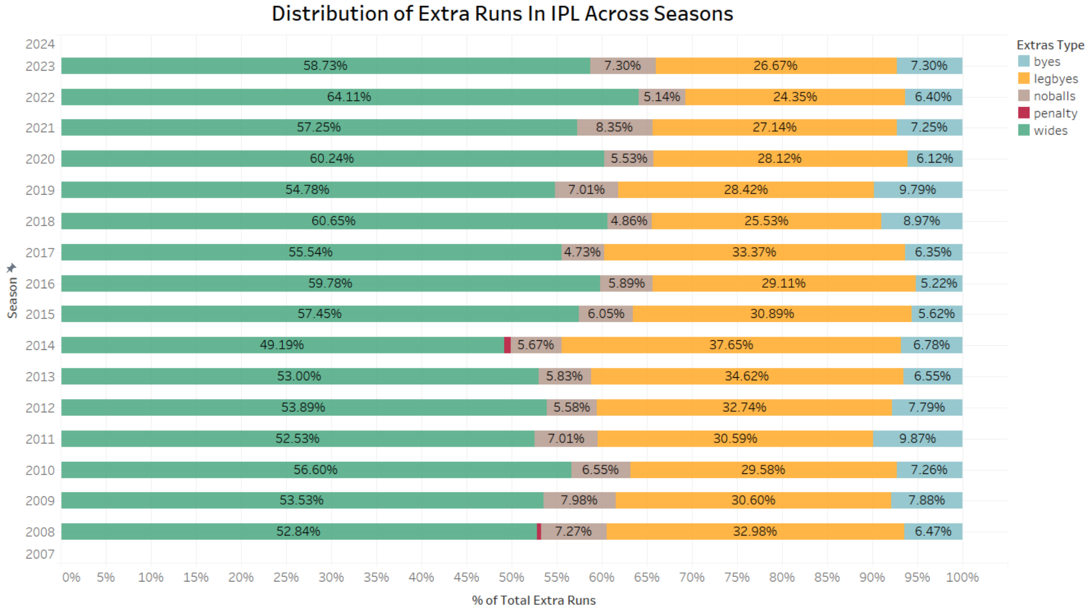
Variables Mapping:

- x-axis: winner
- y-axis: result type
- colour: result type
- size: distinct count of match_id for each combination of winner and result

Refinement: Before creating the plot, we removed null values from the result type while keeping the ties. To create this plot, we calculated distinct count of match_id for each combination of winner and result.

Role in Analysis/Story: This helps audience in identifying which teams have fared better in IPL on winning large numbers of the same when compared to others. This allows you to compare the team's performance across different win types very easily. For example, you might find that a team has huge tiles under the "Runs" section but very small tiles under "Wickets" which suggests that they win more by scoring big runs than by making it difficult for their opponents to lose their wickets.

Proportional Stacked Bar Chart: Distribution of Extra Runs in IPL



Message: This proportional stacked bar chart gives a detailed breakdown of how the various types of extra runs, such as byes, leg byes, wides, etc., contribute to the total extra runs tally over different IPL seasons.

Variable Mapping:

- x-axis: total percent of sum of extra runs
- y-axis: season
- colour: extras type
- label: total percent of sum of extra runs
- Tooltip: Sum of extra runs

Refinement: A proportional stacked bar chart was chosen as the final way to go instead of line or pie to represent the data, and it was used due to the following reasons: It clearly shows both the total extra runs across seasons (represented by entire bar height) and the relative contribution of each extra run type (proportional height of each segment within the bar).

Role in Analysis/Story: This proportional stacked bar chart is quite important for understanding the different roles that extra runs play in the IPL. This helps to track trends over time. By examining the bars across the various seasons, one can find out whether there are trends in the proportions of extra runs scored by each type over time. For each season (bar), it is straightforward for viewers to compare the relative scale of various extra run types (for example, a season may have a higher percentage of wides than it does leg byes).

Analysis

Unveiling the Powerhouses and Performance Nuances of IPL

This data-driven exploration took us at the heart of the IPL, not only to see the most dominant teams across the seasons but also to visualize finer details of player performance and scoring trends. Here are some takeaways:

Powerhouse Teams: The visual analysis of the data consistently sets up a great story. the Mumbai Indians and the Chennai Super Kings are perennial IPL powerhouses. Donut charts likely highlighted these two franchises at the forefront of both most total runs and wickets taken, respectively solidifying their top-contender status.

Beyond the Big Two: Though the Mumbai Indians and the Chennai Super Kings are the big two, scatter plots probably teased out any additional high-performing teams and individual players who were consistently challenging the leading dominance with at least a few teams.

Performance Nuances: Stacked bar charts that would show dismissals by type, this will reveal player vulnerabilities to different types of bowling. Specific teams may have a higher percentage of batsmen dismissed caught behind the wicket, whereas others could be more susceptible to swing bowling. This will show opportunities for strategic enhancement.

Shifting Tides: Mosaic plots highlighting areas that were within the scope of seasonal patterns would allow us to compare win patterns over the seasons. This may reveal something like teams are winning by narrower margins or an increase in high-scoring contests.

Beyond the Scoreboard: The analysis stays beyond just simple runs and wickets. One can certainly look further into trends in extra runs, like byes and wides, using proportional stacked bar charts. This can let one look deeper into umpire tendencies or perhaps specific wicket-keeping techniques that impact scoring.

Conclusion:

To Conclude, this data visualization journey provided a truly insightful exploration of the IPL. It has helped us identify not only the league's elite teams and players but also many strategic nuances, dismissal patterns, and an ever-evolving landscape of scoring. This multifaceted analysis empowers us to appreciate the intricate details that contribute to the enthralling matches in the IPL and its huge global cricketing phenomenon.

Our Learnings from this project:

This IPL data visualization project served as a valuable learning experience. It reinforced the importance of selecting appropriate chart types for the data being presented. Scatter plots proved effective for identifying top performers, while donut charts clearly showcased dominant teams. Stacked bar charts provided insights into dismissal patterns, revealing areas for strategic improvement. Overall, the project solidified the power of data visualization in transforming complex data into clear and insightful narratives.

Code for Mosaic Plot (RStudio):

```
# Importing necessary libraries
library(ggplot2)
library(ggmosaic)
library(dplyr)

# Remove rows with NA values but keep 'Tie' results
cleaned_data <- merged %>%
  filter(!is.na(winner) & !is.na(result))

# Calculate distinct count of match_id for each combination of winner and result
distinct_count_data <- cleaned_data %>%
  group_by(winner, result) %>%
  summarise(distinct_match_id_count = n_distinct(match_id)) %>%
  ungroup()

# Plotting with ggplot2 and geom_mosaic
ggplot(data = distinct_count_data) +
  geom_mosaic(aes(weight = distinct_match_id_count, x = product(result, winner), fill = result)) +
  labs(fill = "Result") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 10)) +
  ggtitle("IPL Winners by Result Including Ties (2008-2023)")
```

Individual Report – Yashkumar Prajapati

My key contributions include:

Data Preprocessing: I done the preprocessing part of the dataset before starting. I removed inconsistencies, missing observations, or bad formatting in the dataset. After the data was successfully pre-processed, I proceeded to undertake the creation of the following visualizations to scrutinize different aspects of the IPL.

Mosaic Plot: I designed and developed a Mosaic plot focusing on the distribution of IPL match wins across teams and modes of results, either wins by runs, wickets, or tie, across a specific timeframe.

Proportional Stacked Bar Chart: I developed a proportional stacked bar chart focused on the changing landscape of extra runs (byes, leg byes, wides) in the IPL over seasons. More specifically, uncover the trends associated with how the different extra run types of proportions towards the total runs have changed over seasons.

Stacked bar chart: I developed a stacked bar chart to investigate the partition of the getting out of the batsmen due to the different dismissal methods in the IPL. This visualization can be useful for identifying potential weaknesses of the batsmen in certain types of dismissals and areas which the bowlers should target to gain higher strategic leverage.

Tree map: I plotted a tree map to analyse the distribution of IPL championship titles among all franchises, providing an effective way to show briefly which teams have the most championship titles in comparison to others.

Presentation and Communication: As a team member, I also presented our findings to the class. I explained the design choices of the above charts, the key insights, and what kind of importance can be derived from those charts in understanding the IPL. This helped me create a really strong base for communication, as it allowed me to represent in a graphical manner many complex sets of data with clear interpretations.

Learnings:

Selection of charts: Realized why it is important to select the appropriate chart based on the data and the story we want to convey. Each chart type has its strengths and limitations, and we need to learn these limitations to be able to communicate correctly.

Data storytelling: The prime importance of this course was data storytelling. Making clear charts is one task but making them and their inferences communicate and involve viewers in the story of the data is completely another.

Collaboration: This project developed my collaboration skills. Working effectively in a team to achieve one single goal was a must if we were to learn from this project.

This IPL Data Visualization project has indeed been a great learning process. Going from pre-processing data to creating and presenting visualizations, I developed my technical skills and grasped a better understanding of data storytelling through visuals. This project has further set down a foundation for further data visualization techniques and methodologies to be explored.

Individual Report – Pradeep Shane

Data Acquisition: The main development of the IPL involved datasets. I was successfully able to find some useful ones on Kaggle. These two datasets form the base of further exploration and visualization. Building on top of the acquired data, I was also active in the development of the visualizations to analyse the various perspectives surrounding the IPL—given below is one of my specific contributions.

Donut Charts: I contributed to the donut charts, which is a great visualization of the percentage category for the dimensions. It would help to analyse distribution of total tuns/wickets by team. By using donut charts, one could very well depict the percentage of total runs scored or wickets taken across a season or even a larger timeframe; hence, it gives the trend of dominant teams across the league in their batting or bowling domain.

Learnings:
Data Source Identification: I have realized the need for relevant data identification and access for a project. Kaggle turned out to be a vital source of data on IPL, developed practice for browsing different genres of its huge repository of data.

Data Understanding: Studying the data for the IPL, I was able to practice understanding the data structure and pattern to find out the relevant variables that will help in achieving the targeted goal of the project. This would be of upmost importance; otherwise, changing the form of data is necessary for effective manipulation and further operation.

Visualization Choice: The class spent time learning the importance of choosing the right kind of visualization to represent the data and the story that is to be told using the data. Working with donut charts gave me the opportunity to investigate how to effectively use this visualization to tell the story of the data using percentages.

Wrap-Up: I helped secure the foundational data for this project, thus playing a significant role in improving the exploratory analysis of IPL. Learning and finding where to get the data made me realize how easy it is to start with data visualization. Also, I was able to learn donut charts to increase my visualization ideas.

Individual Report – Ayush Kothari

Contribution:

This was a visualization-heavy project that had an analysis focus on visualizing the performance of players in the IPL. I was responsible for the development and delivery of scatterplots, which are the significant visualization for this project.

Scatterplots and Player Performance Analysis: I chose scatterplots since they were used to represent individual player performance across time in quite a visually effective way. By analysing the scatterplots, below conclusions can be drawn:

Top Performers: We can easily identify the greatest number of runs scored by a player (batsmen) or most wickets taken by a player (bowler) in scatter plots.

Performance Trends: A trace of a particular data point on X-axis gives us the information related to the performance trajectory of a player across his/her IPL career.

Class Presentation: I played a vital role in the development and delivery of the introduction, data and objective part of the class presentation.

Learnings:

Choice of Visualization: Working with scatter plots, gave a lot of experience to understand how it projected individual player performance and trends over time. Scatter plots do not create confusion, as sometimes happens with line overlap in line charts.

Data Storytelling through Visualization: Developed the scatter plots and presentation to refine skills in telling stories through data visualization. We use colour coding the teams to see if player movements across teams could be highlighted.

Presentation Skills: Prepared and presented the introduction and objective of the class presentation to enhance my presentation and communication skills.

By preparing and presenting scatterplots, a key visualization tool, I was able to refine my skills in data storytelling via visualization. Additionally, I prepared and presented the introduction and objective section of the class presentation to augment presentation and communication skills.

Individual Report – Shiv Gandhi

The Three Bar Charts

I was actively participating in exploring these three key visualizations to attain insights from the IPL landscape.

Team Participation Patterns: This was likely in the form of a bar chart to show the number of seasons each IPL team had taken part in across some time let's say 2008-2023. Team participation history is an important aspect of the later analysis, because from it, one can make better distinctions between different teams' performance over the various seasons.

Unveiling Run-Scoring Titans: This will be a scatter plot outlining who the key beast run scorers have been in the history of the IPL and their run-scoring performances across the different seasons. Through analysis of this chart, one can identify the highest total run-scorers and subsequently follow their performance trends throughout their IPL careers.

Unveiling Top Wicket-Takers' Longevity: Following up on the batting, the next visualization is made on the highest wicket-takers of the IPL. However, I have included another unit of measure, the number of seasons played, for each bowler. This allows it to avoid a common pitfall and enables viewers to understand not only who the bowlers with the highest wickets are but also consider their experience and longevity in the IPL.

Key Learnings: The process of initial exploration thus helped in learning some aspects crucial for analysis:

Power of Visualization: Developing these visualizations helped understand the power of data storytelling through data visualization. Good charts significantly improve the ability to understand and derive hidden insights from the chart created.

Importance of Context: In this exploration phase, it was nothing but the importance of context to understand the data. Knowing the participation history of a team or the length of a career of a participant serves crucial added context to plain performance metrics.

This initial data visualization and exploration in addition to the above learnings have worked well in going a step further with understanding the IPL player/team performance trends, eventually leading to engaging more stories from the IPL world.