ABOUT THE PROJECT

T20 World Cup 2022: Dataset

The dataset contains data about all the matches from the super 12 stage to the final of the ICC Men's T20 World Cup 2022. Below are all the features in the dataset:

- venue: The venue where the match was played
- team1: the team that batted first
- team2: the team that batted second
- stage: stage of the match (super 12, semi-final, or final)
- toss winner: the team that won the toss
- toss decision: the decision of the captain after winning the toss
- first innings score: runs scored in the first innings
- first innings wickets: the number of wickets lost in the first innings
- second innings score: runs scored in the second innings
- second innings wickets: the number of wickets lost in the second innings
- winner: the team that won the match
- won by: how the team won the match (wickets or runs)
- player of the match: the player of the match
- top scorer: the player who scored highest in the match
- highest score: the highest runs scored in the match by the player
- best bowler: the player who took the most wickets in the match
- best bowling figure: the number of wickets taken and runs given by the best bowler in the mat

MY goal and question to be answered during the analysis are:

Overall Tournament Performance:

- What is the average first innings score in the super 12 stage, semi-finals, and the final?
- Is there any significant difference in the average second innings score between the stages?
- Which team had the highest average first innings score throughout the tournament?

Top Performers:

- Who were the top run-scorer and top wicket-taker of the tournament?
- What was the highest individual score by a player in a match?
- What was the best bowling figure by a player in a match?

Team Performance:

- Which team had the most wins in each stage of the tournament?
- Which team had the highest and lowest first innings scores in a match?
- Which team had the highest and lowest second innings scores in a match?

- Which team had the highest and lowest first innings scores in a match?
- Which team had the highest and lowest second innings scores in a match?

Player of the Match:

- Which top 5 player receive the "Player of the Match" award?

Toss Analysis:

- What percentage of matches were won by the team that won the toss?
- Did the team winning the toss have a significant advantage in any particular stage?

Trends and Patterns:

- The average first innings score over time as the tournament progressed.
- Comparing the distribution of second innings scores across different stages of the tournament.
- Calculating the average first innings score at each venue, to help me identify venues where teams tend to score higher or lower.
- Calculating the trend in toss decisions (batting or bowling) across different stages of the tournament

The following are the steps that would be followed during this project

Step 1: Gathering Data

Step 2: Assessing data

Step 3: Cleaning data

Step 4: Analyzing

Step 5: Visualization

Step 6: Reporting

Step 1: Gathering Data

In [1]:

```
#Loading important packages
library("tidyverse")
library("ggplot2")
```

```
Registered S3 methods overwritten by 'ggplot2':
  method         from
  [.quosures     rlang
  c.quosures     rlang
  print.quosures rlang
Registered S3 method overwritten by 'rvest':
  method            from
  read_xml.response xml2
-- Attaching packages --------------------------------------- tidyverse
1.2.1 --
v ggplot2 3.1.1       v purrr   0.3.2
v tibble  2.1.1       v dplyr   0.8.0.1
v tidyr   0.8.3       v stringr 1.4.0
v readr   1.3.1       v forcats 0.4.0
-- Conflicts ------------------------------------------ tidyverse_conflic
ts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

In [2]:
```
#Reading the csv file into a variable t20_worldcup
t20_worldcup <- read_csv("t20-world-cup-22.csv")
```

```
Parsed with column specification:
cols(
  venue = col_character(),
  team1 = col_character(),
  team2 = col_character(),
  stage = col_character(),
  `toss winner` = col_character(),
  `toss decision` = col_character(),
  `first innings score` = col_double(),
  `first innings wickets` = col_double(),
  `second innings score` = col_double(),
  `second innings wickets` = col_double(),
  winner = col_character(),
  `won by` = col_character(),
  `player of the match` = col_character(),
  `top scorer` = col_character(),
  `highest score` = col_double(),
  `best bowler` = col_character(),
  `best bowling figure` = col_character()
)
```

Step 2: Assessing data

In [3]:
```
#Checking the dataset visually
t20_worldcup
```

venue	team1	team2	stage	toss winner	toss decision	first innings score	first innings wickets	second innings score
SCG	New Zealand	Australia	Super 12	Australia	Field	200	3	111
Optus Stadium	Afghanistan	England	Super 12	England	Field	112	10	113
Blundstone Arena	Ireland	Sri lanka	Super 12	Ireland	Bat	128	8	133
MCG	Pakistan	India	Super 12	India	Field	159	8	160
Blundstone Arena	Bangladesh	Netherlands	Super 12	Netherlands	Field	144	8	135
Blundstone Arena	Zimbabwe	South Africa	Super 12	Zimbabwe	Bat	79	5	51
Optus Stadium	Sri lanka	Australia	Super 12	Australia	Field	157	6	158
MCG	Ireland	England	Super 12	England	Field	157	10	105
MCG	New Zealand	Afghanistan	Super 12	NA	NA	NA	NA	NA
SCG	South Africa	Bangladesh	Super 12	South Africa	Bat	205	5	101
SCG	India	Netherlands	Super 12	India	Bat	179	2	123
Optus Stadium	Zimbabwe	Pakistan	Super 12	Zimbabwe	Bat	130	8	129
MCG	Afghanistan	Ireland	Super 12	NA	NA	NA	NA	NA
MCG	Australia	England	Super 12	NA	NA	NA	NA	NA
SCG	New Zealand	Sri lanka	Super 12	New Zealand	Bat	167	7	102
The Gabba	Bangladesh	Zimbabwe	Super 12	Bangladesh	Bat	150	7	147
Optus Stadium	Netherlands	Pakistan	Super 12	Netherlands	Bat	91	9	95
Optus Stadium	India	South Africa	Super 12	India	Bat	133	9	137
The Gabba	Australia	Ireland	Super 12	Ireland	Field	179	5	137
The Gabba	Afghanistan	Sri lanka	Super 12	Afghanistan	Bat	144	8	148
The Gabba	England	New Zealand	Super 12	England	Bat	179	6	159
Adelaide Oval	Zimbabwe	Netherlands	Super 12	Zimbabwe	Bat	117	10	120
Adelaide Oval	India	Bangladesh	Super 12	Bangladesh	Field	184	6	145
SCG	Pakistan	South Africa	Super 12	Pakistan	Bat	185	9	108
Adelaide Oval	New Zealand	Ireland	Super 12	Ireland	Field	185	6	150

venue	team1	team2	stage	toss winner	toss decision	first innings score	first innings wickets	second innings score
Adelaide Oval	Australia	Afghanistan	Super 12	Afghanistan	Field	168	8	164
SCG	Sri lanka	England	Super 12	Sri lanka	Bat	141	8	144
Adelaide Oval	Netherlands	South Africa	Super 12	South Africa	Field	158	4	145
Adelaide Oval	Bangladesh	Pakistan	Super 12	Bangladesh	Bat	127	8	128
MCG	India	Zimbabwe	Super 12	India	Bat	186	5	115
SCG	New Zealand	Pakistan	Semi-final	New Zealand	Bat	152	4	153
Adelaide Oval	India	England	Semi-final	England	Field	168	6	170
MCG	Pakistan	England	Final	England	Field	137	8	138

In [4]:
```
#Checking the dataset to know the number of observation and variable
glimpse(t20_worldcup)
```

```
Observations: 33
Variables: 17
$ venue                    <chr> "SCG", "Optus Stadium", "Blundstone Aren
a"...
$ team1                    <chr> "New Zealand", "Afghanistan", "Ireland",
"...
$ team2                    <chr> "Australia", "England", "Sri lanka", "In
di...
$ stage                    <chr> "Super 12", "Super 12", "Super 12", "Sup
er...
$ `toss winner`            <chr> "Australia", "England", "Ireland", "Indi
a"...
$ `toss decision`          <chr> "Field", "Field", "Bat", "Field", "Fiel
d",...
$ `first innings score`    <dbl> 200, 112, 128, 159, 144, 79, 157, 157, N
A,...
$ `first innings wickets`  <dbl> 3, 10, 8, 8, 8, 5, 6, 10, NA, 5, 2, 8, N
A,...
$ `second innings score`   <dbl> 111, 113, 133, 160, 135, 51, 158, 105, N
A,...
$ `second innings wickets` <dbl> 10, 5, 1, 6, 10, 0, 3, 5, NA, 10, 9, 8,
NA...
$ winner                   <chr> "New Zealand", "England", "Sri lanka",
"In...
$ `won by`                 <chr> "Runs", "Wickets", "Wickets", "Wickets",
"...
$ `player of the match`    <chr> "Devon Conway", "Sam Curran", "Kusal Men
di...
$ `top scorer`             <chr> "Devon Conway", "Ibrahim Zadran", "Kusal
M...
$ `highest score`          <dbl> 92, 32, 68, 82, 62, 47, 59, 62, NA, 109,
6...
$ `best bowler`            <chr> "Tim Southee", "Sam Curran", "Maheesh Th
ee...
$ `best bowling figure`    <chr> "03-Jun", "05-Oct", "Feb-19", "Mar-30",
"A...
```

In [5]:
```
#Checking for the summary statistics of  the dataset
summary(t20_worldcup)
```

```
    venue              team1              team2              stage
 Length:33          Length:33          Length:33          Length:33
 Class :character   Class :character   Class :character   Class :characte
r
 Mode  :character   Mode  :character   Mode  :character   Mode  :characte
r




 toss winner        toss decision      first innings score
 Length:33          Length:33          Min.   : 79.0
 Class :character   Class :character   1st Qu.:134.0
 Mode  :character   Mode  :character   Median :157.0
                                       Mean   :153.4
                                       3rd Qu.:179.0
                                       Max.   :205.0
                                       NA's   :3
 first innings wickets second innings score second innings wickets
 Min.   : 2.000        Min.   : 51.0        Min.   : 0.000
 1st Qu.: 5.250        1st Qu.:113.5        1st Qu.: 5.000
 Median : 7.500        Median :136.0        Median : 6.000
 Mean   : 6.867        Mean   :130.8        Mean   : 6.233
 3rd Qu.: 8.000        3rd Qu.:147.8        3rd Qu.: 9.000
 Max.   :10.000        Max.   :170.0        Max.   :10.000
 NA's   :3             NA's   :3            NA's   :3
    winner             won by             player of the match  top scorer
 Length:33          Length:33          Length:33           Length:33
 Class :character   Class :character   Class :character    Class :charact
er
 Mode  :character   Mode  :character   Mode  :character    Mode  :charact
er




 highest score    best bowler        best bowling figure
 Min.   : 32.00   Length:33          Length:33
 1st Qu.: 52.50   Class :character   Class :character
 Median : 62.00   Mode  :character   Mode  :character
 Mean   : 64.07
 3rd Qu.: 70.25
 Max.   :109.00
 NA's   :3
```

In [6]:
```
#Checking for any missing Values
any(is.na(t20_worldcup))
```

TRUE

In [7]:
```
# finding the location of missing values
which(is.na(t20_worldcup))
```

141 145 146 174 178 179 207 211 212 240 244 245 273 277 278
306 310 311 336 339 343 344 369 372 376 377 402 405 409 410
438 442 443 471 475 476 504 508 509 537 541 542

In [8]:
```
#counting the total missing values
sum(is.na(t20_worldcup))
```

42

In [9]:
```
#finding the location of missing values column wise
sapply(t20_worldcup, function(x) which(is.na(x)))
```

$venue

$team1

$team2

$stage

$`toss winner`

9 13 14

$`toss decision`

9 13 14

$`first innings score`

9 13 14

$`first innings wickets`

9 13 14

$`second innings score`

9 13 14

$`second innings wickets`

9 13 14

$winner

6 9 13 14

$`won by`

6 9 13 14

$`player of the match`

6 9 13 14

$`top scorer`

9 13 14

$`highest score`

9 13 14

$`best bowler`

9 13 14

$`best bowling figure`

9 13 14

In [10]:
```
#counting the missing values by column wise
sapply(t20_worldcup, function(x) sum(is.na(x)))
```

venue	0
team1	0
team2	0
stage	0
toss winner	3
toss decision	3
first innings score	3
first innings wickets	3
second innings score	3
second innings wickets	3
winner	4
won by	4
player of the match	4
top scorer	3
highest score	3
best bowler	3
best bowling figure	3

In [11]:
```
#Checking for duplicates
sum(duplicated(t20_worldcup))
```

0

After assessing our tables based on quality and tidiness it was observed that :

- We have some missing data in dataset which can be dropped due to the observation from the assessment
- Also the column names should be properly renamed with underscore instead of having spaces in between the names
- Then the last column named best_bowling_figure should be dropped, because it contains a date instead of figures
- Lastly a new column will be created to calculate the average_best_bowling_figure(More like saying the best_bowling_figure)

Step 3: Cleaning data

In [12]:
```
#First we make the copy of the dataset before cleaning
t20_worldcup_copy <- t20_worldcup
```

In [13]:
```
#Removing the rows with NA
t20_worldcup_copy <- t20_worldcup_copy %>% drop_na()

#Checking if row where properly removed
sum(is.na(t20_worldcup_copy))
```

0

In [14]:

```
#Another check visually
t20_worldcup_copy
```

venue	team1	team2	stage	toss winner	toss decision	first innings score	first innings wickets	second innings score
SCG	New Zealand	Australia	Super 12	Australia	Field	200	3	111
Optus Stadium	Afghanistan	England	Super 12	England	Field	112	10	113
Blundstone Arena	Ireland	Sri lanka	Super 12	Ireland	Bat	128	8	133
MCG	Pakistan	India	Super 12	India	Field	159	8	160
Blundstone Arena	Bangladesh	Netherlands	Super 12	Netherlands	Field	144	8	135
Optus Stadium	Sri lanka	Australia	Super 12	Australia	Field	157	6	158
MCG	Ireland	England	Super 12	England	Field	157	10	105
SCG	South Africa	Bangladesh	Super 12	South Africa	Bat	205	5	101
SCG	India	Netherlands	Super 12	India	Bat	179	2	123
Optus Stadium	Zimbabwe	Pakistan	Super 12	Zimbabwe	Bat	130	8	129
SCG	New Zealand	Sri lanka	Super 12	New Zealand	Bat	167	7	102
The Gabba	Bangladesh	Zimbabwe	Super 12	Bangladesh	Bat	150	7	147
Optus Stadium	Netherlands	Pakistan	Super 12	Netherlands	Bat	91	9	95
Optus Stadium	India	South Africa	Super 12	India	Bat	133	9	137
The Gabba	Australia	Ireland	Super 12	Ireland	Field	179	5	137
The Gabba	Afghanistan	Sri lanka	Super 12	Afghanistan	Bat	144	8	148
The Gabba	England	New Zealand	Super 12	England	Bat	179	6	159
Adelaide Oval	Zimbabwe	Netherlands	Super 12	Zimbabwe	Bat	117	10	120
Adelaide Oval	India	Bangladesh	Super 12	Bangladesh	Field	184	6	145
SCG	Pakistan	South Africa	Super 12	Pakistan	Bat	185	9	108
Adelaide Oval	New Zealand	Ireland	Super 12	Ireland	Field	185	6	150
Adelaide Oval	Australia	Afghanistan	Super 12	Afghanistan	Field	168	8	164
SCG	Sri lanka	England	Super 12	Sri lanka	Bat	141	8	144
Adelaide Oval	Netherlands	South Africa	Super 12	South Africa	Field	158	4	145
Adelaide Oval	Bangladesh	Pakistan	Super 12	Bangladesh	Bat	127	8	128
						first	first	second

venue	team1	team2	stage	toss winner	toss decision	first innings score	first innings wickets	second innings score
MCG	India	Zimbabwe	Super 12	India	Bat	186	5	115
SCG	New Zealand	Pakistan	Semi-final	New Zealand	Bat	152	4	153
Adelaide Oval	India	England	Semi-final	England	Field	168	6	170
MCG	Pakistan	England	Final	England	Field	137	8	138

In [15]:

```
#Renaming Columns
t20_worldcup_copy <- t20_worldcup_copy%>%
rename(toss_winner = "toss winner",
       toss_decision = "toss decision",
       first_innings_score = "first innings score",
       first_innings_wickets = "first innings wickets",
       second_innings_score = "second innings score",
       second_innings_wickets = "second innings wickets",
       won_by = "won by",
       player_of_the_match = "player of the match",
       top_scorer = "top scorer",
       highest_score = "highest score",
       best_bowler = "best bowler",
       best_bowling_figure = "best bowling figure")

#Checking the column names
colnames(t20_worldcup_copy)
```

'venue' 'team1' 'team2' 'stage' 'toss_winner' 'toss_decision' 'first_innings_score' 'first_innings_wickets' 'second_innings_score' 'second_innings_wickets' 'winner' 'won_by' 'player_of_the_match' 'top_scorer' 'highest_score' 'best_bowler' 'best_bowling_figure'

In [16]:

```
# Dropping best_bowling_figure column by using the subset function which h
t20_worldcup_copy <- subset(t20_worldcup_copy, select = c(venue, team1, te
                                toss_winner, toss_decision, first_innings_
                                second_innings_score, second_innings_wicke
                                player_of_the_match, top_scorer, highest_s
#Checking our dataset
t20_worldcup_copy
```

venue	team1	team2	stage	toss_winner	toss_decision	first_innings_score
SCG	New Zealand	Australia	Super 12	Australia	Field	200
Optus Stadium	Afghanistan	England	Super 12	England	Field	112
Blundstone Arena	Ireland	Sri lanka	Super 12	Ireland	Bat	128
MCG	Pakistan	India	Super 12	India	Field	159
Blundstone Arena	Bangladesh	Netherlands	Super 12	Netherlands	Field	144
Optus Stadium	Sri lanka	Australia	Super 12	Australia	Field	157
MCG	Ireland	England	Super 12	England	Field	157
SCG	South Africa	Bangladesh	Super 12	South Africa	Bat	205
SCG	India	Netherlands	Super 12	India	Bat	179
Optus Stadium	Zimbabwe	Pakistan	Super 12	Zimbabwe	Bat	130
SCG	New Zealand	Sri lanka	Super 12	New Zealand	Bat	167
The Gabba	Bangladesh	Zimbabwe	Super 12	Bangladesh	Bat	150
Optus Stadium	Netherlands	Pakistan	Super 12	Netherlands	Bat	91
Optus Stadium	India	South Africa	Super 12	India	Bat	133
The Gabba	Australia	Ireland	Super 12	Ireland	Field	179
The Gabba	Afghanistan	Sri lanka	Super 12	Afghanistan	Bat	144
The Gabba	England	New Zealand	Super 12	England	Bat	179
Adelaide Oval	Zimbabwe	Netherlands	Super 12	Zimbabwe	Bat	117
Adelaide Oval	India	Bangladesh	Super 12	Bangladesh	Field	184
SCG	Pakistan	South Africa	Super 12	Pakistan	Bat	185
Adelaide Oval	New Zealand	Ireland	Super 12	Ireland	Field	185
Adelaide Oval	Australia	Afghanistan	Super 12	Afghanistan	Field	168
SCG	Sri lanka	England	Super 12	Sri lanka	Bat	141
Adelaide Oval	Netherlands	South Africa	Super 12	South Africa	Field	158
Adelaide Oval	Bangladesh	Pakistan	Super 12	Bangladesh	Bat	127

venue team1 team2 stage toss_winner toss_decision first_innings_score

venue	team1	team2	stage	toss_winner	toss_decision	first_innings_score
MCG	India	Zimbabwe	Super 12	India	Bat	186
SCG	New Zealand	Pakistan	Semi-final	New Zealand	Bat	152
Adelaide Oval	India	England	Semi-final	England	Field	168
MCG	Pakistan	England	Final	England	Field	137

In [17]:

```
# Calculating the total_innings_score by adding first_innings_score and se
t20_worldcup_copy$total_innings_score <- t20_worldcup_copy$first_innings_s

# Calculating the total_innings_wickets by adding the first_innings_wicket
t20_worldcup_copy$total_innings_wickets <- t20_worldcup_copy$first_innings

#Adding the averages together to get our best_bowling_figure
t20_worldcup_copy$average_best_bowling_figure = t20_worldcup_copy$total_in


#Checking Our result
t20_worldcup_copy
```

venue	team1	team2	stage	toss_winner	toss_decision	first_innings_score
SCG	New Zealand	Australia	Super 12	Australia	Field	200
Optus Stadium	Afghanistan	England	Super 12	England	Field	112
Blundstone Arena	Ireland	Sri lanka	Super 12	Ireland	Bat	128
MCG	Pakistan	India	Super 12	India	Field	159
Blundstone Arena	Bangladesh	Netherlands	Super 12	Netherlands	Field	144
Optus Stadium	Sri lanka	Australia	Super 12	Australia	Field	157
MCG	Ireland	England	Super 12	England	Field	157
SCG	South Africa	Bangladesh	Super 12	South Africa	Bat	205
SCG	India	Netherlands	Super 12	India	Bat	179
Optus Stadium	Zimbabwe	Pakistan	Super 12	Zimbabwe	Bat	130
SCG	New Zealand	Sri lanka	Super 12	New Zealand	Bat	167
The Gabba	Bangladesh	Zimbabwe	Super 12	Bangladesh	Bat	150
Optus Stadium	Netherlands	Pakistan	Super 12	Netherlands	Bat	91
Optus Stadium	India	South Africa	Super 12	India	Bat	133
The Gabba	Australia	Ireland	Super 12	Ireland	Field	179
The Gabba	Afghanistan	Sri lanka	Super 12	Afghanistan	Bat	144
The Gabba	England	New Zealand	Super 12	England	Bat	179
Adelaide Oval	Zimbabwe	Netherlands	Super 12	Zimbabwe	Bat	117
Adelaide Oval	India	Bangladesh	Super 12	Bangladesh	Field	184
SCG	Pakistan	South Africa	Super 12	Pakistan	Bat	185
Adelaide Oval	New Zealand	Ireland	Super 12	Ireland	Field	185
Adelaide Oval	Australia	Afghanistan	Super 12	Afghanistan	Field	168
SCG	Sri lanka	England	Super 12	Sri lanka	Bat	141
Adelaide Oval	Netherlands	South Africa	Super 12	South Africa	Field	158
Adelaide Oval	Bangladesh	Pakistan	Super 12	Bangladesh	Bat	127

venue team1 team2 stage toss_winner toss_decision first_innings_score

venue	team1	team2	stage	toss_winner	toss_decision	first_innings_score
MCG	India	Zimbabwe	Super 12	India	Bat	186
SCG	New Zealand	Pakistan	Semi-final	New Zealand	Bat	152
Adelaide Oval	India	England	Semi-final	England	Field	168
MCG	Pakistan	England	Final	England	Field	137

Step 4: Analyzing

Overall Tournament Performance

- Calculating average first innings score by stage

In [18]:

```
#Calculating the average first innings score by stage
average_first_innings_score <- t20_worldcup_copy %>%
group_by(stage) %>%
summarise(average_score = mean(first_innings_score))%>%
arrange(desc(average_score))

average_first_innings_score
```

stage	average_score
Semi-final	160.0000
Super 12	156.3462
Final	137.0000

- Is there any significant difference in the average second innings score between the stages?

In [19]:

```
# Testing for significant difference in second innings scores between stag
anova_result <- aov(second_innings_score ~ stage, data = t20_worldcup_copy
summary(anova_result)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
stage        2   1722   861.0   2.119   0.14
Residuals   26  10565   406.4
```

Result

The ANOVA results suggest that there are no statistically significant differences in the second innings scores across the different stages of the T20 World Cup matches as there is no strong evidence to reject the null hypothesis which is also indicated by the relatively large p-value (0.14) for the stage variable.

- Which team had the highest average first innings score throughout the tournament?

In [20]:

```
# Calculating the average first innings score for team1
team1_average_scores <- t20_worldcup_copy %>%
  group_by(team1) %>%
  summarize(average_first_innings_score = mean(first_innings_score))

# Calculating the average first innings score for team2
team2_average_scores <- t20_worldcup_copy %>%
  group_by(team2) %>%
  summarize(average_first_innings_score = mean(first_innings_score))

# Merging the results for team1 and team2
average_scores_combined <- bind_rows(
  team1 = team1_average_scores,
  team2 = team2_average_scores
)

# Finding the team  with the highest average first innings score
team_with_highest_average <- average_scores_combined %>%
  arrange(desc(average_first_innings_score)) %>%
  head(1)

# Print the result
cat("Team with the highest average first innings score:", team_with_highes
cat("Average first innings score:", team_with_highest_average$average_firs
```

```
Team with the highest average first innings score: South Africa
Average first innings score: 205
```

Top Performers:

- Who were the top run-scorer and top wicket-taker of the tournament?

In [21]:

```
# Top run_scorer of the tournament
top_scorer <- t20_worldcup_copy %>%
group_by(player_of_the_match) %>%
summarise(total_runs = sum(highest_score))%>%
arrange(desc(total_runs))%>%
head(1)

#Top wicket_taker of the tournament
top_bowler <- t20_worldcup_copy %>%
group_by(best_bowler) %>%
summarise(total_wickets = sum(first_innings_wickets,second_innings_wickets
arrange(desc(total_wickets))%>%
head(1)


# Print the result
cat("Top run-scorer of the tournament is:", top_scorer$player_of_the_match
cat("With a score of :", top_scorer$total_runs, "\n")
cat("Top wicket-taker of the tournament is :", top_bowler$best_bowler, "\n
cat("With a score of :", top_bowler$total_wickets, "\n")
```

```
Top run-scorer of the tournament is: Virat Kohli
With a score of : 146
Top wicket-taker of the tournament is : Sam Curran
With a score of : 40
```

- What was the highest individual score by a player in a match?

- What was the highest individual score by a player in a match?

In [22]:
```
# Calculating the highest individual score
highest_individual_score <- t20_worldcup_copy%>%
group_by(top_scorer)%>%
arrange(desc(highest_score)) %>%
head(1)

# Print the result
cat("Highest individual score by a player in a match:", highest_individual
cat("And the player is :", highest_individual_score$top_scorer, "\n")
```

```
Highest individual score by a player in a match: 109
And the player is : Rilee Rossouw
```

- What was the best bowling figure by a player in a match?

In [23]:
```
#Calculating the best average bowling figure by a player in a match
best_bowling_figure <- t20_worldcup_copy %>%
group_by(best_bowler)%>%
arrange(average_best_bowling_figure) %>%
head(1)

cat("Best average bowling figure by a player in a match:", best_bowling_fi
cat("And the player is :", best_bowling_figure$best_bowler, "\n")
```

```
Best average bowling figure by a player in a match: 14.30769
And the player is : Shadab Khan
```

Team Performance:

- Which team had the most wins in each stage of the tournament?

In [24]:
```
#Calculating the team with the most wins in each stage
team_with_most_wins <- t20_worldcup_copy %>%
  group_by(stage, winner) %>%
  summarise(matches_won = n()) %>%
  arrange(stage, desc(matches_won)) %>%
  slice(1)

team_with_most_wins
```

stage	winner	matches_won
Final	England	1
Semi-final	England	1
Super 12	India	4

- Which team had the highest and lowest first innings scores in a match?

In [25]:

```
#Calculate the maximum and minimum first innings scores for team1
team1_max_and_min_scores <- t20_worldcup_copy %>%
  group_by(team1) %>%
  summarize(max_first_innings_score = max(first_innings_score),
            min_first_innings_score = min(first_innings_score))

# Calculate the maximum and minimum first innings scores for team2
team2_max_and_min_scores <- t20_worldcup_copy %>%
  group_by(team2) %>%
  summarize(max_first_innings_score = max(first_innings_score),
            min_first_innings_score = min(first_innings_score))

# Merge the results for team1 and team2
max_and_min_scores_combined <- bind_rows(
  team1 = team1_max_and_min_scores,
  team2 = team2_max_and_min_scores
)

# Find the team (either team1 or team2) with the highest and lowest first
team_with_highest_score <- max_and_min_scores_combined %>%
  arrange(desc(max_first_innings_score)) %>%
  head(1)

team_with_lowest_score <- max_and_min_scores_combined %>%
  arrange(min_first_innings_score) %>%
  head(1)

# Print the results
cat("Team with the highest first innings score:", team_with_highest_score$
cat("Highest first innings score:", team_with_highest_score$max_first_inni

cat("Team with the lowest first innings score:", team_with_lowest_score$te
cat("Lowest first innings score:", team_with_lowest_score$min_first_inning
```

```
Team with the highest first innings score: South Africa
Highest first innings score: 205
Team with the lowest first innings score: Netherlands
Lowest first innings score: 91
```

- Which team had the highest and lowest second innings scores in a match?

In [26]:

```
#Calculating the maximum and minimum first innings scores for team1
team1_max_and_min_scores <- t20_worldcup_copy %>%
  group_by(team1) %>%
  summarize(max_first_innings_score = max(second_innings_score),
            min_first_innings_score = min(second_innings_score))

# Calculating the maximum and minimum first innings scores for team2
team2_max_and_min_scores <- t20_worldcup_copy %>%
  group_by(team2) %>%
  summarize(max_first_innings_score = max(second_innings_score),
            min_first_innings_score = min(second_innings_score))

# Merging the results for team1 and team2
max_and_min_scores_combined <- bind_rows(
  team1 = team1_max_and_min_scores,
  team2 = team2_max_and_min_scores
)

# Finding the team with the highest and lowest second innings scores
team_with_highest_score <- max_and_min_scores_combined %>%
  arrange(desc(max_first_innings_score)) %>%
  head(1)

team_with_lowest_score <- max_and_min_scores_combined %>%
  arrange(min_first_innings_score) %>%
  head(1)

# Print the results
cat("Team with the highest second innings score:", team_with_highest_score
cat("Highest first innings score:", team_with_highest_score$max_first_inni

cat("Team with the lowest second innings score:", team_with_lowest_score$t
cat("Lowest first innings score:", team_with_lowest_score$min_first_inning
```

```
Team with the highest second innings score: India
Highest first innings score: 170
Team with the lowest second innings score: Netherlands
Lowest first innings score: 95
```

Player of the Match:

- Which top 5 player receive the "Player of the Match" award?

In [27]:
```
# Counting the number of times each player received "Player of the Match"
player_of_the_match <- t20_worldcup_copy %>%
group_by(player_of_the_match) %>%
summarise(count = n())%>%
arrange(desc(count))%>%


head(5)
player_of_the_match
```

player_of_the_match	count
Sam Curran	2
Shadab Khan	2
Suryakumar Yadav	2
Taskin Ahmed	2
Virat Kohli	2

Toss Analysis:

- What percentage of matches were won by the team that won the toss?
- Did the team winning the toss have a significant advantage in any particular stage?

In [28]:
```
# Calculating the percentage of matches won by the team winning the toss
percentage_matches_won_by_toss_winner <- t20_worldcup_copy %>%
  filter(winner == toss_winner) %>%
  summarize(percentage = (n() / nrow(t20_worldcup_copy)) * 100)

# Print the percentage of the matches won by toss winner
cat("Percentage of matches won by the team winning the toss:", percentage_

# Performing a chi-squared test for independence while i create a continge
contingency_table <- table(t20_worldcup_copy$toss_winner, t20_worldcup_cop

# Perform the chi-squared test
chisq_test_result <- chisq.test(contingency_table)

# Print the chi-squared test result
print(chisq_test_result)
```

```
Percentage of matches won by the team winning the toss: 44.82759 %

Warning message in chisq.test(contingency_table):
"Chi-squared approximation may be incorrect"


        Pearson's Chi-squared test

data:  contingency_table
X-squared = 126.63, df = 110, p-value = 0.1328
```

Result

- The percentage of matches won by the team winning the toss is 45% approximately which shows that winning the toss does not guarantee that you will win matches.
- Also, Based on the results of Pearson's Chi-squared test, with a p-value of 0.1328, there is no significant association or relationship between which team won the toss (toss_winner) and which team won the match (winner). In other words, the outcome of the match does not appear to be dependent on which team won the toss.

Visualization and Reporting

Trends and Patterns:

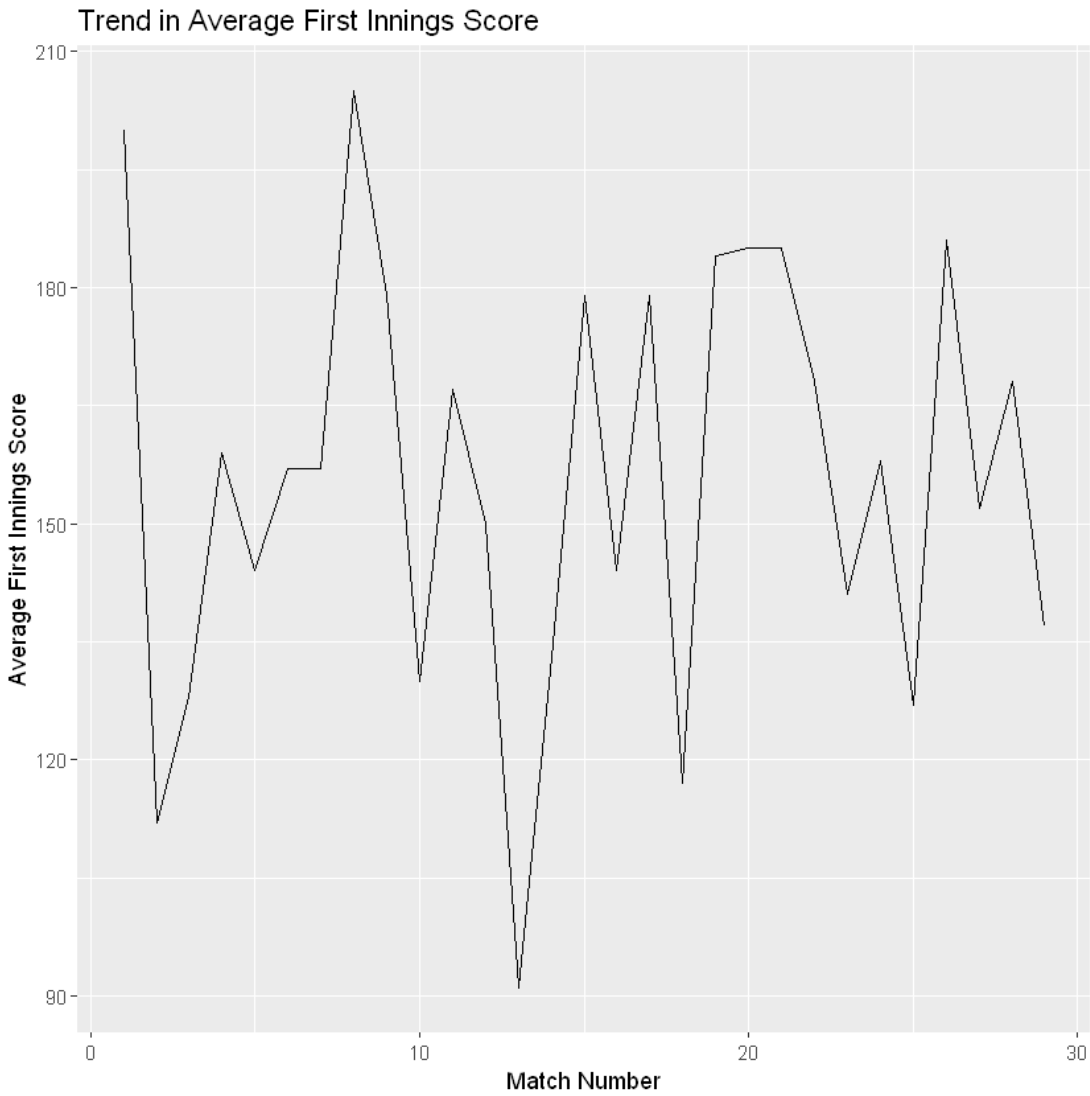
- The average first innings score over time as the tournament progressed.

In [29]:

```
# Creating a sequence of numbers to represent match order
t20_worldcup_copy$match_number <- seq(1, nrow(t20_worldcup_copy))

# Calculating the average first innings score over time (matches)
avg_first_innings_score <- t20_worldcup_copy %>%
arrange(match_number) %>%
group_by(match_number) %>%
summarise(avg_score = mean(first_innings_score))

# Creating a line plot to visualize the trend
ggplot(avg_first_innings_score, aes(x = match_number, y = avg_score)) +
  geom_line() +
  labs(x = "Match Number", y = "Average First Innings Score") +
  ggtitle("Trend in Average First Innings Score")
```

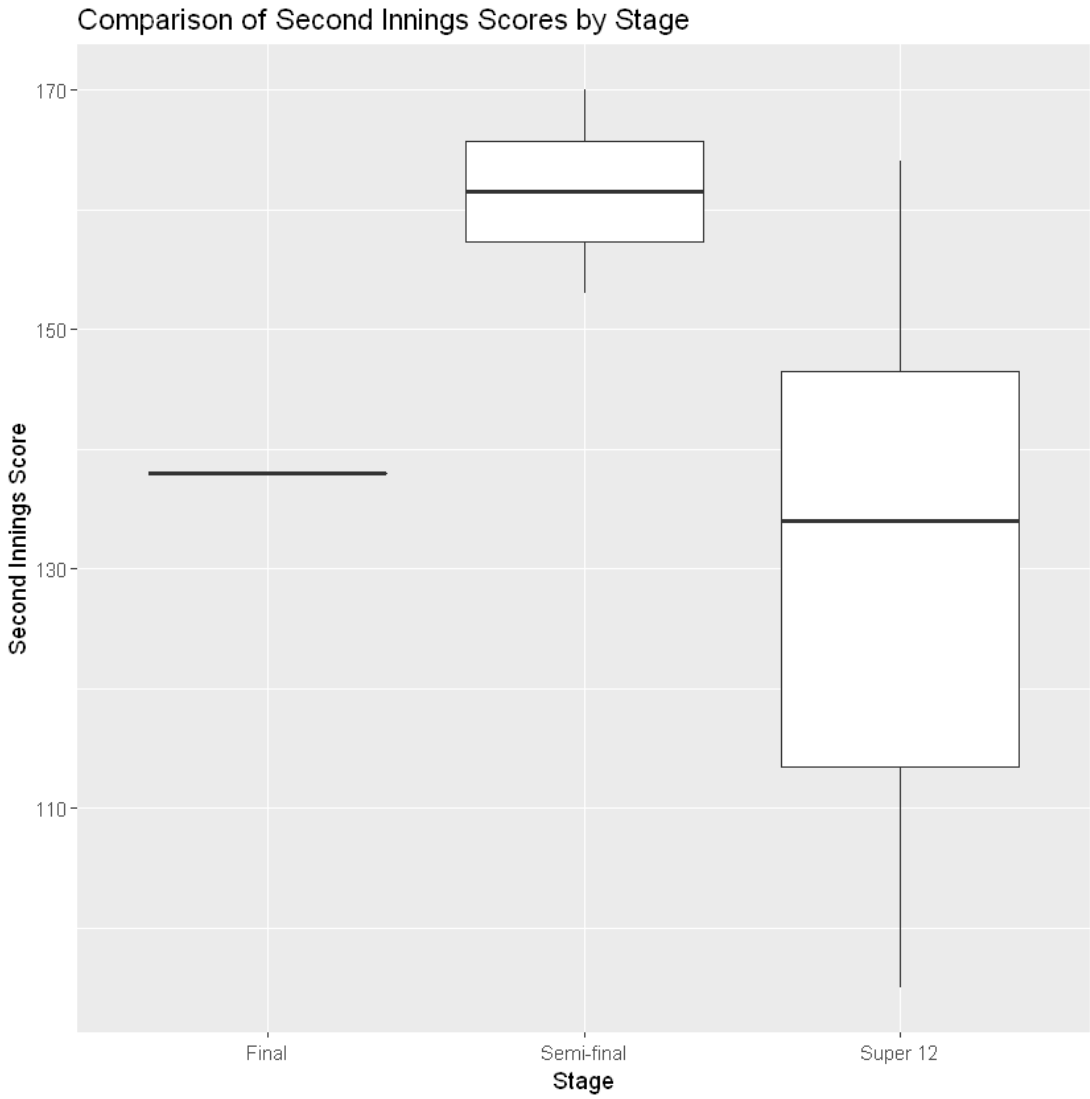


Result

- I create a new column match_number in the dataset using the seq() function to generate a sequence of numbers from 1 to the number of rows in the dataset. This sequence represents the order of the matches.
- I then proceeded to calculate the average first innings score over time (based on the order of matches) and create a line plot to visualize the trend. The x-axis now represents the "Match Number," which is essentially the order of the matches in the dataset.
- This way, you can analyze the trend in the average first innings score over time

Comparing the distribution of second innings scores across different stages of the tournament.

In [30]:
```
# Creating a box plot to compare second innings scores by stage
ggplot(t20_worldcup_copy, aes(x = stage, y = second_innings_score)) +
  geom_boxplot() +
  labs(x = "Stage", y = "Second Innings Score") +
  ggtitle("Comparison of Second Innings Scores by Stage")
```



Result

The graph also shows that the average second innings scores decrease from the Super 12 stage to the Final stage of the tournament. This suggests that the quality of bowling and fielding improves as the tournament progresses.

This is likely due to the following factors:

- The teams that progress to the later stages of the tournament are typically the stronger teams.
- The teams have more time to prepare for the later stages of the tournament.
- The pressure is higher in the later stages of the tournament, which can motivate players to perform at their best.

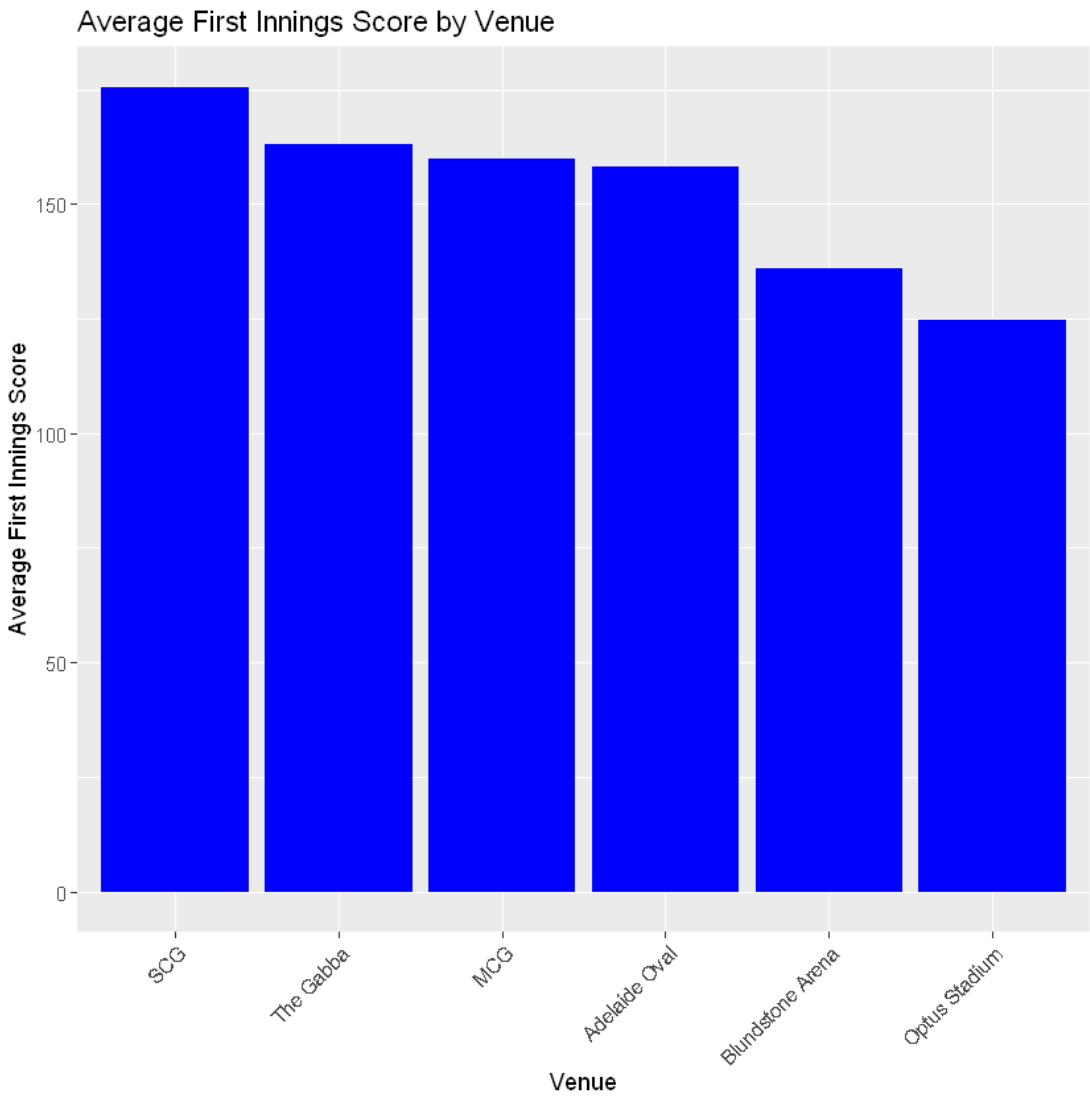
Venue Analysis:

- Calculating the average first innings score at each venue, to help me identify venues where teams tend to score higher or lower.

In [31]:

```
# Calculate the average first innings score by venue
avg_first_innings_score_venue <- t20_worldcup_copy %>%
  group_by(venue) %>%
  summarise(avg_score = mean(first_innings_score))

# Create a bar plot to show average first innings scores at each venue
ggplot(avg_first_innings_score_venue, aes(x = reorder(venue, -avg_score),
  geom_bar(stat = "identity", fill = "blue") +
  labs(x = "Venue", y = "Average First Innings Score") +
  ggtitle("Average First Innings Score by Venue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Result

- SCG has the highest average first innings score
- While Optus Stadium has the lowest average first innings score.

This suggests that this venue is generally less favorable to batsmen and we can also say some of the factors below can affect the outcome of a match Such as:

- The weather conditions

say some of the factors below can affect the outcome of a match such as:

- The weather conditions
- The pitch conditions
- And the quality of the opposition.

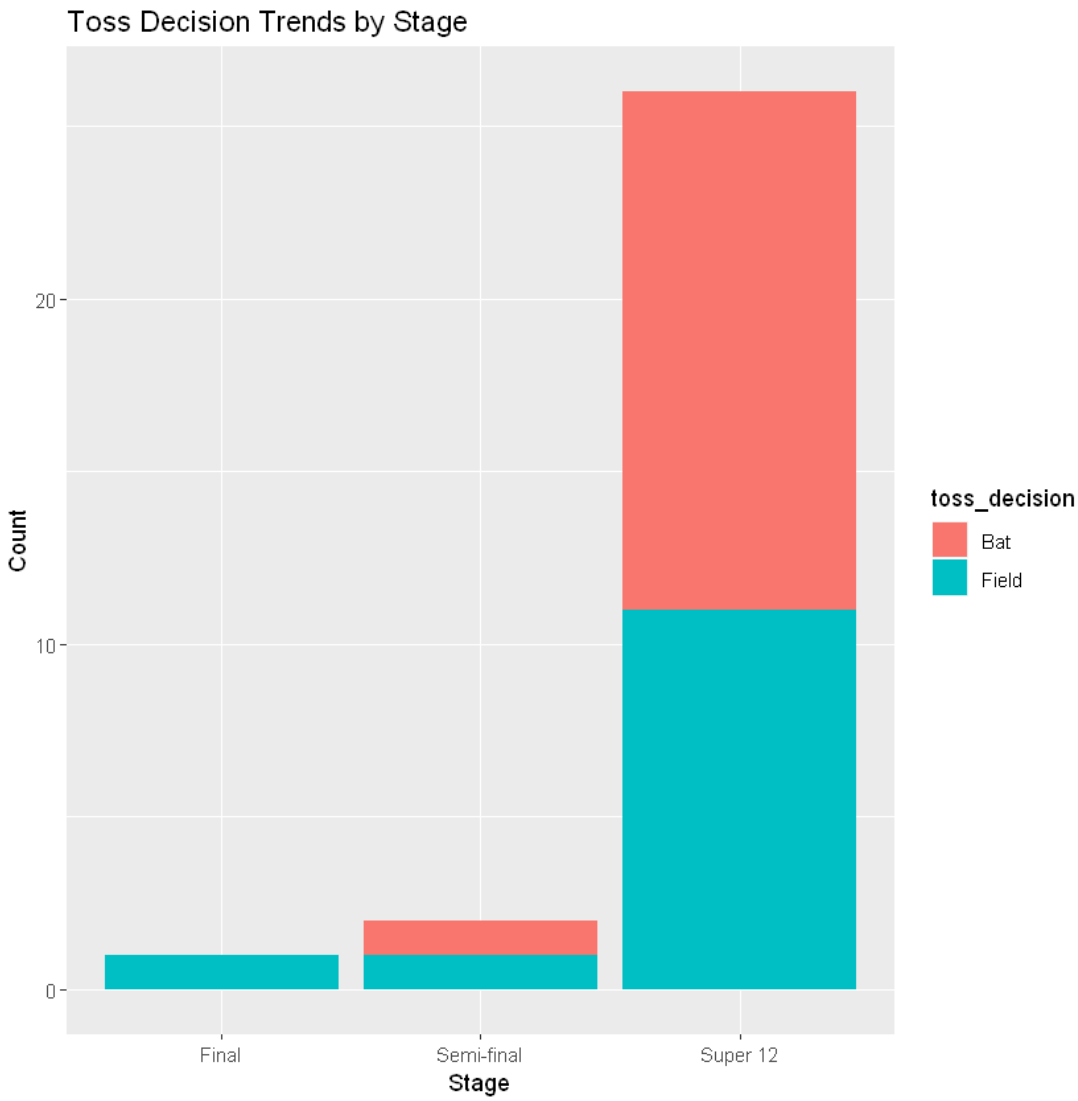
Toss Decision Trends

- Calculating the trend in toss decisions (batting or bowling) across different stages of the tournament

In [32]:

```
# Creating a stacked bar plot to visualize the trend in toss decisions
toss_decision_trend <- t20_worldcup_copy %>%
  group_by(stage, toss_decision) %>%
  summarise(count = n())

ggplot(toss_decision_trend, aes(x = stage, y = count, fill = toss_decision
  geom_bar(stat = "identity") +
  labs(x = "Stage", y = "Count") +
  ggtitle("Toss Decision Trends by Stage")
```



Result

The graph shows that teams are more likely to choose to bat first in the Super 12 stage than in the Semi-Final and Final stages.

- This suggests that teams value the advantage of batting first in the early stages of the

than in the Semi-Final and Final stages.

- This suggests that teams value the advantage of batting first in the early stages of the tournament.
- Which makes teams want to get off to a good start in the tournament and build momentum. Another possibility is that teams feel that they have a better chance of winning if they bat first, as this gives them more time to set a target for the opposition.

The graph also shows that teams are more likely to choose to field first in the Semi-Final and Final stages of the tournament. Which suggests that teams value the advantage of chasing in the later stages of the tournament