Yash Mantri
The Shri Ram School, Moulsari,
Gurugram, Haryana, India

A statistical analysis of the Indian premiere league 2018-2021

Yash Mantri

Abstract
Founded in 2007, the Indian Premiere League (IPL) is one the most watched sports leagues in the world. The event is followed and enjoyed by millions of Indians and cricket lovers across the globe.
This study determines the quantitative factors that have played a role in winning matches in the last three seasons of the IPL. This is done using statistical tools such as Karl Pearson's and Multiple Correlation along with regression analysis and the t-test. The qualifiers and finalists of the IPL 2021 are predicted using the binomial distribution with the help of past winning percentages. Cricket is not only governed by numerical factors; therefore, the predicted qualifiers are further investigated for a qualitative factor analysis using the chi-square test and rank correlation. This overall analysis successfully predicts the team that has the highest chance of winning the IPL 2021 considering both quantitative and qualitative factors.

Introduction
Cricket, also known as the Gentleman's Game, is currently the world's second most popular sport and loved by people from all around the globe.
The first reference to cricket can be traced back to the early 1600s in England when it was played in grammar schools, villages, and farm communities. The first official international match was played in 1877 between England and Australia.
Test cricket is the traditional form of the game which has been played since then. It comprises of two innings each and is played over five days. It is known as the pinnacle form because it tests teams over a long duration.
Almost a decade later, in the 1980s, One Day Internationals (ODIs) gained popularity. This is a quicker format of the game with 50 overs played each side and comprises of one innings each. The well-known International Cricket Council (ICC) Cricket World Cup is contested every four years in this format.
Twenty20 cricket, the newest and modern format, revolutionized the game when it was introduced in 2003. It brought with it, rule changes and less overs that saw the beginning of power hitting and a whole new audience. It triggered the adoption of new skill sets and innovations in both batsmen and bowlers. A typical T20 match takes three hours to complete and includes creative batting, skillful bowling, and brilliant fielding. Other than the ICC Cricket Twenty20 championship, many T20 tournaments have emerged over the years including the Big Bash League, Caribbean Premiere League, Super Smash and of course the Indian Premiere League.
The Indian Premiere League (IPL) was founded by the Board of Control for Cricket in India in 2007. The league is contested by eight teams based out of eight different Indian cities- Each team is formed with the help of shuffling to arrive at four foreign players and seven Indian players. This shuffling process is achieved through an auction which takes place every year before the start of IPL. The IPL is the most-attended cricket league in the world and was ranked sixth by average attendance among all sports leagues.

Corresponding Author:
Yash Mantri
The Shri Ram School, Moulsari,
Gurugram, Haryana, India

Aim

To determine the most effective quantitative factors that have played a role in winning matches in the last three seasons of the Indian Premiere League and further analyze qualitative factors of teams.

To make predictions about the qualifiers and winner of the 2021 season using various statistical tools and probability distributions.

Data Collection (IPL 2020)

Table 1: IPL 2020 Data Collection

Team	Matches Won	Average of top 5 run scorers	Average of top 3 Strike Rates	Total number of 6s hit	Total number of 4s hit	Dot balls per match	Batsmen with strike rate greater than 140	Total 50s per team	Bowling Average of top 2 opening pacers from each team	Total number of wickets taken per match
Mumbai Indians	9	422.4	172.05	137	222	46	6	17	16.62	5.9
Delhi Capitals	8	412	144.66	88	236	40	4	14	20.01	5.7
Sunrisers Hyderabad	7	369.8	148.42	79	213	42	3	17	18.9	5.62
Royal Challengers Bangalore	7	358	135.46	43	176	36	2	14	18.7	5
Kolkata Knight Riders	7	321.8	141.79	86	197	37	2	11	26.3	5.2
Punjab Kings	6	373	156.13	98	179	43	2	14	24	5.07
Chennai Super Kings	6	308.6	148.17	75	188	38	2	12	28	5
Rajasthan Royals	6	310.8	160.913	105	171	42	3	11	30	4.3

Karl Pearson's Coefficient of Correlation

The Pearson Product Moment Correlation shows the linear relationship between an independent variable X and dependent variable Y. It is used to find the association between the two variables.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - \sum(x)^2 \right]\left[n\sum y^2 - \sum(y)^2 \right]}}$$

Here, (i) n is the sample size of the dataset and equals 8 (ii) x is the data set of the quantitative values of the factor being

analyzed (independent variables) (iii) y is the data-set of the number of matches won (dependent variables)

Note: The data from the IPL 2020 season has been analyzed.

First, the factor of the mean of the bowling average of the top 2 opening pacers from each team has been taken. A player's bowling average is the number of runs they have conceded per wicket taken. Hence, the lower this value is for a bowler, the better it is considered. The correlation coefficient between the bowling average of top 2 opening pacers from each team and the matches won by that team in that season has been calculated:

Table 2: Karl Pearson's Coefficient of Correlation

Team	Bowling Average of top 2 opening pacers from each team (X)	Matches won(Y)	XY	X²	Y²
Mumbai Indians	16.62	9	149.58	276.2	81
Delhi Capitals	20	8	160	400	64
Sunrisers Hyderabad	19	7	133	361	49
Royal Challengers Bangalore	19	7	133	361	49
Kolkata Knight Riders	26.3	7	184.1	691.7	49
Punjab Kings	24	6	144	576	36
Chennai Super Kings	28	6	168	784	36
Rajasthan Royals	30	6	180	900	36
	$\sum x$ =182.92	$\sum y$ =56	$\sum xy$ =1251.68	$\sum x^2$ =4349.9	$\sum y^2$ =400

Calculating r,

$$r = \frac{8(1251.68) - (182.92)(56)}{\sqrt{[(8)(4349.9) - (33459.7)][(8)(400) - (3136)]}} = -0.786$$

(High Inverse Correlation)

Thus, the correlation between the two variables is a high and inverse relationship. Since a lower bowling average is better

for a bowler, that is why a negative association can be seen. As the coefficient of correlation is high (lower than -0.75), it can be assumed that it is a contributing factor in winning matches in the Indian Premiere League. The same formula has been applied to a number of possible factors and the coefficient of correlation between them has been obtained.

Table 3: Correlation between Factors and Winning matches

S. No.	Factors	Coefficient of Correlation	Remark
1.	Average runs scored by top 5 run scorers from each team	+0.8073	Strong Correlation
2.	Average of Top 3 Strike Rates from each team	+0.2887	Weak Correlation
3.	Average number of 6s hit per match	+0.4174	Weak Correlation
4.	Average number of 4s hit per match	+0.6405	Moderate Correlation
5.	Total number of 50s scored by each team	+0.6188	Moderate Correlation
6.	Batsmen in each team with a strike rate greater than 140	+0.8504	Strong Correlation

7.	Average number of dots per match	+0.3558	Weak Correlation
8.	Bowling Average of top 2 opening pacers from each team	-0.786	Strong Inverse Correlation
9.	Bowling economy of top two spinners	-0.092	Weak Inverse Correlation
10.	Total number of wickets taken per match	+0.826	Strong Correlation

It can be observed from the table that the factors which mostly contribute to winning matches are:
1. Average runs scored by top 5 run scorers from each team
2. Batsmen in each team with a strike rate greater than 140
3. Bowling Average of top 2 opening pacers from each team
4. Total number of wickets taken per match.

These four factors affect the winning of the teams by 61%-72.5% (Coefficient of determination)

T-test
The widely used t-test is be used next to establish whether this correlation is true for the population data which includes all T20 cricket matches and other Twenty20 Leagues. The t-test value is calculated using the following formula:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

After putting the respective for n and r, a t-test value of 3.959 for the relationship between matches won and batsmen in each team with strike rate greater than 140 was obtained. Our significance level has been taken as $a = 0.05$.
Null Hypothesis (Ho): The correlation between matches won and batsmen with strike rate greater than 140 is significant in the population data as well.
Alternate Hypothesis (Ha): The correlation between matches won and batsmen with strike rate greater than 140 is not significant in the population data.
Degree of Freedom- The formula for obtaining the value of the degrees of freedom is $DOF = n - 2$. A value of 6 as n=8 is obtained and finally a critical t-test value of 2.447 is obtained. Hence, as the value of 3.959 exceeds the critical value, the null hypothesis is accepted, establishing the correlation observed in the sample data to be statistically significant for the population data as well. The same was observed for the rest of the factors as well after repeating the same procedure. Thus, it can be proven that for all other twenty20 leagues, the same factors will contribute in winning. This can even be generalised to everyday gully cricket T-20 tournaments!

Regression Analysis
A linear regression model is one that assesses the relationship between a dependent variable and an independent variable. Next, the regression analysis is used to predict the matches won by a team with respect to the independent variable such as the total number of wickets taken by a team per match. The following equation is used in order to calculate the regression coefficient for the linear regression graph of y against x:

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

A value of 1.789 for b_{yx} is used using which the linear regression model expression using the following equation is calculated:

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 7 = 1.789(x - 5.23)$$

Finally, the equation can be re-arranged to form a simplified equation for y:

$$y = 1.789x - 2.36$$

Over here: (i) $\bar{y}$, having a value of 7, represents the mean value of y (number of matches won) and (ii) $\bar{x}$, having a value of 5.23, represents the mean value of x (number of wickets taken per match)

Similarly, the regression equations were found for the other three factors as well:

Average runs scored by top 5 run scorers from each team: $y = 0.0198x - 0.13$

Batsmen in each team with a strike rate greater than 140: $y = 0.643x + 5.07$

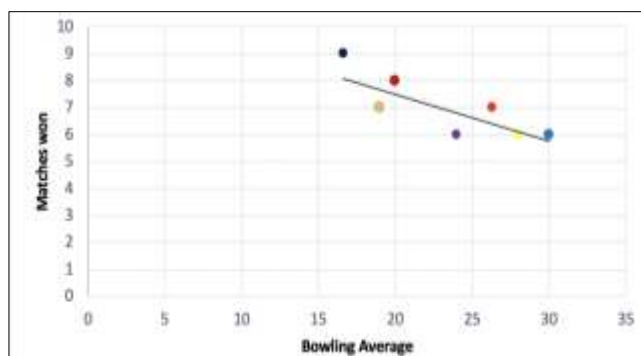Bowling Average of top 2 opening pacers from each team: $y = -0.1717x + 10.93$
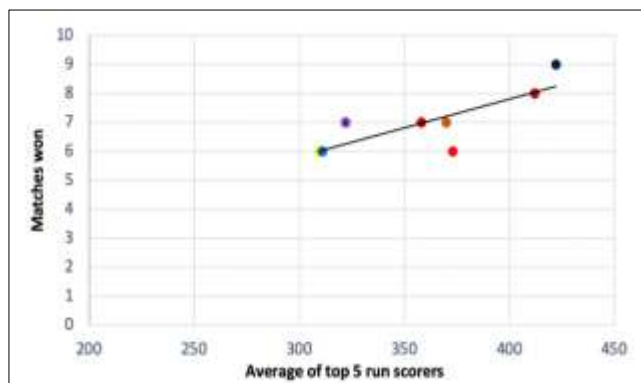

Fig 1: Bowling Average vs. Matches Won

Fig 2: Average of top 5 run scorers vs. Matches won

Multi-Variable Correlation
The relationship between two variables has been determined using correlation and used regression as a prediction tool to predict matches won in a season depending on certain factors.

Multi-variable correlation is a measure of how well a given variable (in this case it is the matches won by a team in a season of the IPL) can be predicted using a linear function of a set of other variables. It was decided that multi variable statistics will be used and the multi-variable correlation of total number of wickets taken per match will be measured. In addition, the bowling average of the top two opening pacers with matches won by a team will be calculated, which is given by the equation:

$$R_{y12} = \sqrt{\frac{r^2_{y1} + r^2_{y2} - 2r_{y1}r_{y2}r_{12}}{1 - r^2_{12}}}$$

Where

- r_{y1} = Correlation between matches won and total wickets taken per match.
- r_{12} = Correlation between total wickets taken per match and bowling average of top 2 opening pacers from each team.
- r_{y2} = Correlation between matches won and bowling average of top 2 opening pacers from team.

The values of the Pearson's Co-efficient of Correlation have been tabulated in the table below.

Table 4: Multiple Correlation

Factors	Pearson's Co-efficient of Correlation
Matches won and Total wickets taken per match (r_{y1})	+0.826
Matches won and Bowling average of top 2 opening pacers from team (r_{y2})	-0.786
Total wickets taken per match and Bowling average of top 2 opening pacers from each team (r_{12})	-0.802

$$R_{y12} = \frac{\sqrt{(0.826)^2 + (-0.786)^2 - [2(0.826)(-0.786)(-0.802)]}}{\sqrt{1 - (-0.802)^2}}$$

$$R_{y12} = 0.8515$$

Therefore, in the Indian Premiere League, Total wickets taken per match and Bowling average of top 2 opening pacers together have a positive impact of 72.51% on the Total number of matches won by a team in a season.
Similarly, we can take multiple factors into account and calculate how much they affect winning.

Net Run Rate
Teams that have a higher overall net run rate in the season tend to be the teams that end up qualifying for the playoffs and the winning teams usually have one of the highest net run rates in the season. So, what is net run rate? One hears of this terminology many times when it comes down to choosing between two teams that have won the same number of games. The team with a higher net run rate is always ranked higher than a team that has a lower net run rate given that the matches won by both teams are same. Net run rate is a statistic used in cricket used to put runs scored and conceded in comparison with the number of overs faced and bowled.

$$\text{Net Run Rate} = \frac{\text{Total runs scored}}{\text{Total Overs Faced}} - \frac{\text{Total runs conceded}}{\text{Total Overs Bowled}}$$

Now, a table with the average net run rate over 3 years has been calculated and shown below:

Table 5:

Team	Average Net Run Rate between 2018-20
Mumbai Indians	0.615
Delhi Capitals	-0.083
Sunrisers Hyderabad	0.490
Royal Challengers Bangalore	-0.216
Kolkata Knight Riders	-0.085
Punjab Kings	-0.305
Chennai Super Kings	-0.008
Rajasthan Royals	-0.427

Thus, it can be observed that the top four teams with the best average net run rate over the last 3 seasons are:
1. Mumbai Indians
2. Sunrisers Hyderabad
3. Chennai Super Kings
4. Delhi Capitals
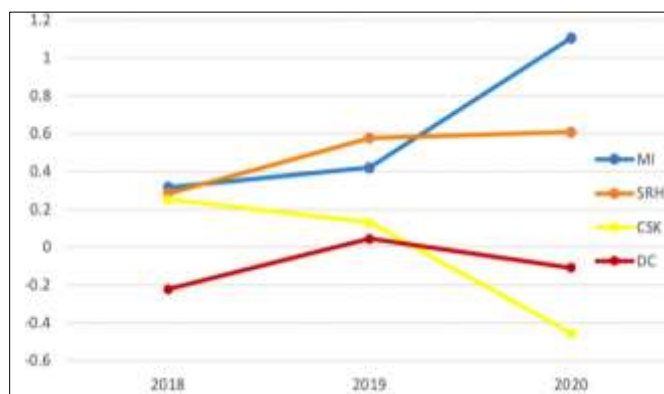
Fig 3: Net Run Rate (2018-2020)
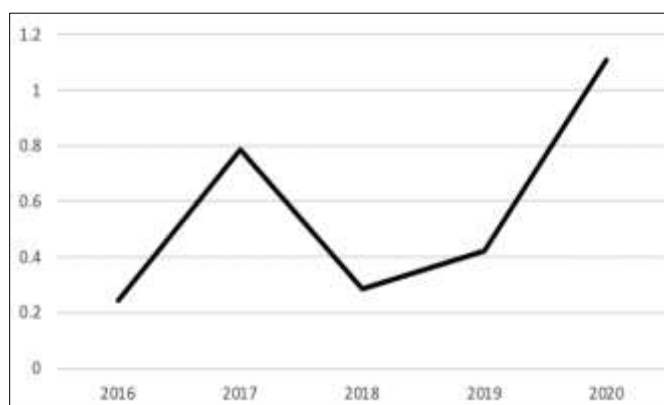
Fig 4: Net Run Rate of the winners of the IPL (2018-2020)

IPL 2021 was unfortunately postponed midway due to COVID-19. Most teams have already played 7 out of 14 matches. Looking at the teams above, three out of the four teams with highest net run rates are in the top four teams of the current points table so far.

Thus, the teams that we will further investigate are:
1. Mumbai Indians
2. Chennai Super Kings
3. Delhi Capitals

The aim is to make predictions about the qualifiers and winner of the IPL 2021 once it resumes.

Binomial Distribution
Cricket is a game of glorious uncertainties. Anything can change at any point in time and that is what makes it so exciting. The IPL has witnessed the breaking of numerous T20 records and many nail-biting matches. An example of this is the match played between Mumbai Indians and Rajasthan Royals in IPL 2014. This is considered to be one of the most dramatic thrillers in IPL history. All the hard work of the season came down to one ball. Mumbai Indians had to score a boundary off the last ball and Aditya Tare helped Mumbai clinch the game with a six, advancing Mumbai Indians to the playoffs leaving Wankhede Stadium in wild celebrations.
In mathematics, uncertainty is measured by probability. The binomial distribution is a common discrete probability distribution used in statistics. Using the winning probabilities of our three shortlisted teams, it is possible to use binomial distribution to calculate the probability of these teams making it to the playoffs and winning. The binomial distribution only counts two states, typically representing success or failure (in this case win or loss) given a fixed number of trials in the data.

The first step is to calculate the winning probabilities of the three teams based on the last three years.

$$\text{Win Probability\%} = \frac{\text{Total Matches Won}}{\text{Total Matches Played}} \times 100$$

Using this we get,
Mumbai Indians= 59.1 %
Chennai Super Kings= 57.2 %
Delhi Capitals= 52.38 %

Before we use binomial distribution, we need to take into account that before the IPL was postponed, Delhi Capitals had played 8 games and Mumbai Indians and Chennai Super Kings had played 7 games each.
Observing the last three seasons, a team that wins 7 or more matches (given a high net run rate), makes it to the playoffs. In most cases, a team that wins more than 8 matches, goes on to make it to the finals.
Now we can calculate the following probabilities for each team using the binomial distribution:

$$P_x = \binom{n}{x} p^x q^{n-x}$$

Mumbai Indians
Games left=7
Games required to win to reach playoffs= 3 or more
Games required to win to play the final= more than 4

Calculating,
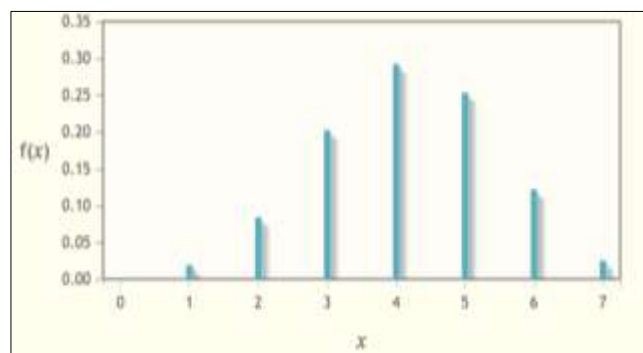Probability (Mumbai Indians will reach the playoffs):

P (Winning 3 matches) + P (Winning 4 matches) + P (Winning 5 matches) + P (Winning 6 matches) + P (Winning 7 matches) =

$^{7}C_{3} \times (0.591)^{3} \times (0.409)^{4} + {}^{7}C_{4} \times (0.591)^{4} \times (0.409)^{3} + {}^{7}C_{5} \times (0.591)^{5} \times (0.409)^{2} + {}^{7}C_{6} \times (0.591)^{6} \times (0.409)^{1} + {}^{7}C_{7} \times (0.591)^{7} \times (0.409)^{0}$

= 0.89477 = 89.477%

Using the same procedure,
Probability (Mumbai Indians will reach the final) =0.4004= 40.04%

Fig 5: Winning binomial Distribution for MI

This is the graphical representation of the binomial probability distribution of Mumbai Indian's next seven games.

Chennai Super Kings
Games left=7
Games required to win to reach playoffs= 2 or more
Games required to win to play the final= more than 3

Calculating,
Probability (Chennai Super Kings will reach the playoffs) =0.97275= 97.275%
Probability (Chennai Super Kings reach the final) =0.65427= 65.427%

Delhi Capitals
Games left=6
Games required to win to reach playoffs= 1 or more
Games required to win to play the final= more than 2

Calculating,
Probability (Delhi Capitals will reach the playoffs) =0.98833=98.83%

Probability (Delhi Capitals make the final) =0.6997=69.97%
It can be observed that it is very likely that all three of these teams will make the playoffs. The teams that are most likely to win 9 or more matches and reach the finals are Delhi Capitals with a chance of 69.97% and Chennai Super Kings with 65.427%. Thus, it can be predicted that Delhi Capitals or Chennai Super Kings have the highest chance of reaching the final in IPL 2021 currently.

Qualitative Factor Analysis using Chi-Square Test
Cricket can certainly not only be controlled by quantitative factors. There are numerous other qualitative factors that contribute to the overall performance of a team. These include

fielding, fitness, team motivation, home ground advantage and many more.

To investigate whether teams rank differently based on the qualitative factors, it was decided to use a chi-square test. A questionnaire was sent out to a random sample of 100 cricket enthusiasts who had been following the IPL 2021. They were asked to rate teams on a scale of 1 to 5(5 being best) keeping in mind their overall fielding, fitness, and team spirit.

The three teams we will apply the test on are Mumbai Indians, Delhi Capitals and

Chennai Super Kings.

Below shows the responses to the fielding rating of the three teams.

Table 6: Data from questionnaire

	Rating of 1,2 or 3 (Low)	Rating of 4 (Medium)	Rating of 5 (High)
Mumbai Indians	26	63	11
Chennai Super Kings	28	36	36
Delhi Capitals	23	56	21

Ho (Null Hypothesis): The rating of IPL teams is independent of fielding performance.

Ha (Alternative Hypothesis): The rating of IPL teams is dependent on fielding performance.

Degrees of Freedom= (Number of rows – 1) (Number of columns – 1) = 4

Level of Significance (a) = 5%

Critical value of c^2 for 4 degrees of freedom at 5% level of significance is 9.488

$$\text{Chi-Square Statistic } (c^2) = \sum \left(\frac{\text{Observed Value - Expected Value}}{\text{Expected Value}} \right)^2$$

Calculating the Chi-square statistic, the value we get is 21.5554.
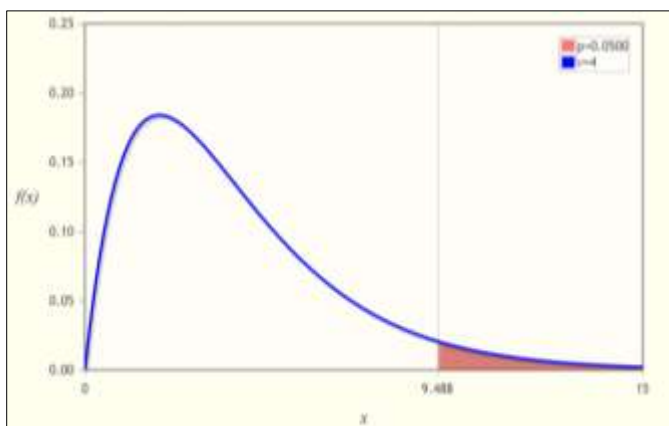
Fig 6: Graphical display of the chi-square distribution at 4 degrees of freedom.

Conclusion

Since the calculated c^2 value is more than the critical value; we can reject the null hypothesis proving that the rating of teams is dependent on fielding performance.

The same procedure was carried out for overall fitness and team spirit. The results obtained were the same and there was statistically significant evidence that the rating of teams is dependent on the qualitative factors.

Spearman's Rank Correlation Coefficient

It was decided to finally use Spearman's Rank correlation to investigate 6 qualitative factors for the two predicted finalists-Delhi Capitals and Chennai Super Kings. In order to carry out the most effective research, two senior cricket experts and analysts *Mr. Gaurav Kalra* (Senior Editor at ESPN) and *S. Ganesh* (Sports consultant and ex-manager of Punjab Kings) graciously provided us with their opinionated scores based on these factors out of 10. The six factors included Fielding, Fitness, Popularity, Team Energy, Team Balance, Hitting Power, and Death-Over Bowling.

The formula for Spearman's Rank Correlation is

$$r = 1 - \frac{6 \sum D^2}{n^3 - n}$$

As ranks are repeated, the rank correlation formula with the repeated ranks is as follows

$$r = 1 - \frac{6 \left[\sum D^2 + \sum (\frac{m^3 - m}{12}) \right]}{n^3 - n}$$

Where D the rank difference and n is is the total number of samples.

Table of Ranks for Both Teams

Table 7: Rank Correlation

Factor	Delhi Capitals		Chennai Super Kings	
	Expert 1	Expert 2	Expert 1	Expert 2
Fielding	6	7	7	6.5
Fitness	4	5	6	7
Popularity	6	6	8	8
Team Energy	7	6	8	8.5
Team Balance	7	8	6	5.5
Hitting Power	8	6.5	7	8.5
Death Over Bowling	7	6	7	6.5

After the total calculation using the formula, the rank correlation between ratings given by both experts is 0.712 for Chennai Super Kings and 0.411 for Delhi Capitals. These values indicate that both the senior analysts have moderately similar views regarding the qualitative aspects of both the teams, justifying their ratings. It was also observed that Chennai Super Kings was rated higher than Delhi Capitals by the two experts for most of the qualitative factors for the year 2021.

Based on the overall analysis, it can be predicted that Chennai Super Kings has the highest chance of winning the IPL 2021, having a high probability of winning more than 8 matches along with high ratings in their qualitative aspects-fielding, fitness, death-over bowling etc.

Conclusion

In accordance with the aim, the quantitative factors that have played a role in winning matches in the last three seasons have been obtained. These include total number of wickets taken per match, average of the top two opening bowlers (bowling average), top five run scorers average and number of batsmen with a strike rate of greater than 140 in each team. The qualifiers and finalists of the IPL 2021 have been predicted using the binomial distribution. The study predicts

that three teams- Mumbai Indians, Chennai Super Kings, and Delhi Capitals have the highest probability (more than 90%) of making it to the qualifiers. Delhi Capitals and Chennai Super Kings are the two predicted finalists. Using the Chi-square test and Rank Correlation, a qualitative factor analysis was successfully carried out. This gave us the result that the ratings of IPL teams are dependent on qualitative factors. The Rank Correlation showed a moderate positive correlation between the ratings given to both the teams by both the experts. Since Chennai Super Kings was rated higher than Delhi Capitals by both the experts in their qualitative aspect, they have been predicted as the winners of the IPL 2021.

It must be noted that cricket has numerous limitations. Probability and data analysis can only make predictions about the game. The fact that we cannot fully predict anything in the game of cricket till the match is over is what makes it so exciting.

References

1. Britannica. (n.d.). Indian Premier. Retrieved April 03, 2021. From https://www.britannica.com/topic/Indian-Premier-League
2. International Cricket Council. (n.d.) History of cricket. Retrieved April 03, 2021. From https://www.icc-cricket.com/about/cricket/history-of-cricket/early-cricket
3. Indian Premier League. (n.d.). By season stats. Retrieved April 03, 2021. From https://www.iplt20.com/stats/
4. Miller N. A very brief history of cricket. The Telegraph. Retrieved April 2016-2021. From https://www.telegraph.co.uk/only-in-britain/odi-cricket/very-brief-history-of-cricket/
5. ESPN. (n.d). Net run rate explained. Retrieved April 03, 2021. From https://www.espncricinfo.com/ci/content/page/429305.html
6. Douglas CC. Some Interpretations of the Multiple Correlation Coefficient. The American Statistician 1972;26(2):31-33. Taylor and Francis Ltd. Retrieved April 08, 2021, from https://www.jstor.org/stable/2683460
7. Huhn C. Correlation Revelation: The Search for Meaning in Pearson's Coefficient. The Mathematics Teacher 2016;110(2):98-103. National Council of Teachers of Mathematics. Retrieved April 09, 2021, from https://www.jstor.org/stable/10.5951/mathteacher.110.2.0098
8. Indo-Asian News Service (IANS). IPL has changed the dynamics of cricket, says BCCI boss Anurag Thakur: Cricket, News. India Today. Retrieved 17 April 2021. https://www.indiatoday.in/sports/cricket/story/bcci-president-anurag-thakur-on-indian-premier-league-345331-2016-10-06
9. Illowsky B, Dean S. Linear Regression and Correlation. Introductory Statistics (Open Stax). The Libre Texts Libraries. The UC Davis Library, the California State University Affordable Learning Solutions Program. Retrieved 07 April, 2021. From https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(OpenStax)/12%3A_Linear_Regression_and_Correlation/12.05%3A_Testing_the_Significance_of_the_Correlation_Coefficient
10. Indian Premier League. (n.d). IPL. Retrieved April 06, 2021. From http://iplfixture.com