

**UNIVERSITY OF HERTFORDSHIRE**

School of Computer Science

MSc Data Science and Analytics

7COM1079-Team Research and Development Project

**IPL TROPHY OVER YEARS (DATASET)**

**Group\_178**   
**Group members**  
1.Akhil Akuthota - 19020480  
2.Lingeshwar Pinnam - 19020478  
3.Manikanta Akuthota - 19022732  
4.Narendra Pandrinki - 19020479  
5.Nikhil Neelath - 19020481

**Supervisor: DR. John Noll**

**Table of contents**

**ABSTRACT 1**

1. **INTRODUCTION 1**
   1. RESEARCH QUESTIONS 1
   2. HYPOTHESIS 1
2. **DATA VISUALIZATION 2**
   1. BOXPLOT 2
   2. HISTOGRAM 3
   3. SCATTERPLOT 4
3. **ANALYSIS 6**
   1. STATISTICAL ANALYSIS 6
   2. PEARSON’S CORRELATION TEST 7

**4.CONCLUSION** ……………………………………………………………………………………………….9

**ABSTRACT:**

This data was collected from different franchises from IPL played over the years. This data includes specifics like wins and losses of each team, win percentage , and trophies won by each team. The objective of the project is to understand the relationship between the number of trophies won and the win percentage of a team.

**1. INTRODUCTION:**

The Indian Premier League(IPL) is a T20 cricket league in India in which eight teams contest representing eight different states or cities in India. Every team has different strategies to win the trophy every year but the team with more consistency wins more trophies.  
Do the teams with more wins really win the trophy or the team which performs well in the final wins the trophy?

The data used in this analysis is the dataset IPL trophy over years (2008 to 2020) which is taken from the Kaggle website. This dataset contains 6 variables and 12 observations.   
<https://www.kaggle.com/vasudevavivek/ipl-trophy-over-years-2008-2020>

Cricket is already a popular game and the introduction of IPL made it popular especially among the youth as IPL brings a new opportunity to the younger generations and is a good stage for a player to showcase their talent and get a chance to play for their national side. The main reason for taking this dataset is whether the win percentage of a team is the main factor for winning the trophy or not.

**1.1. Research Question:**

Is there a correlation between the number of trophies won and the percentages of Matches won by an IPl Team?

**1.2 Hypothesis:**

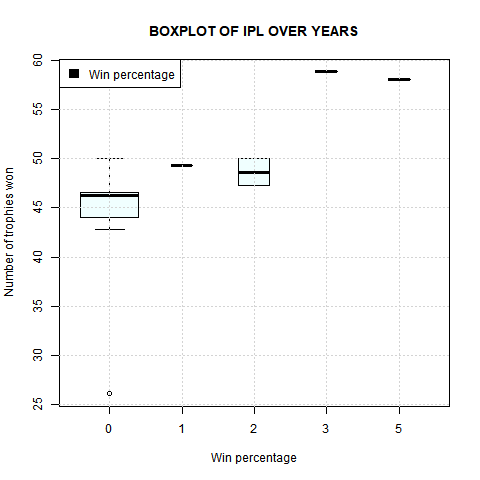
H0: There is no correlation between the number of trophies won and the percentage of matches won by an IPL team.

H1: There is a correlation between the number of trophies won and the percentage of matches won by an IPL team.

**2. Data Visualization**

We used data visualization to represent the data graphically by using visual elements like graphs, plots, etc. in R programming where we can clearly see and understand the data. The Importance of data visualization is to analyze the data and show the relationship between variables and observations. We performed a few representations of the dataset using the boxplot, histogram, and scatterplot.

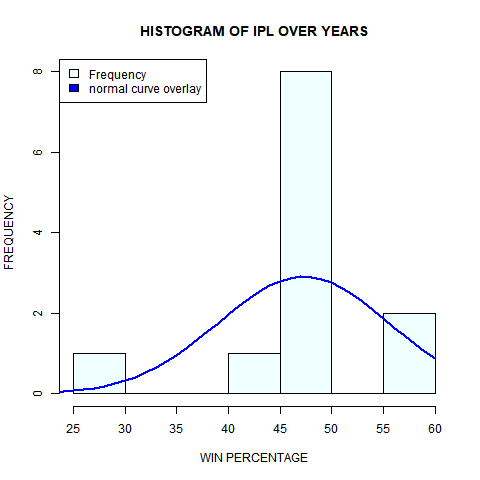
**2.1. BOXPLOT:**

Boxplot is used in explaining data analysis. Its visualization contains the minimum score, maximum score, first quartile, third quartile, and median. It gives a good indication of how the values in the data are spread out.

**Fig1: Boxplot**

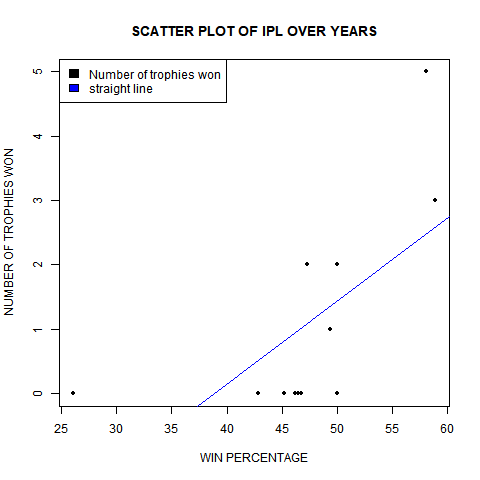
**2.2 HISTOGRAM:**

We used histogram for performing visualization of the dataset. The blue curve in the graph indicates the normal distribution curve in the histogram.



**Fig 2: Histogram**

**2.3 SCATTERPLOT:**

A scatterplot is used to show the relation between the two quantitative variables. The scatterplot is plotted between the win percentage and the number of trophies won. Here each dot represents data in a scatterplot. Using the scatterplot, we can see the correlation between the variables. By observing the scatterplot, we have seen that there are many points near to the correlation. 

**Fig 3: Scatterplot**

Code for visualization of boxplot, histogram and scatterplot:

**library(tidyverse)  
df <-read\_csv("IPL\_trophy\_over\_years.csv")  
pdf("visualization.pdf")**

**#boxplot**

**boxplot(df$`Win percentage` ~ df$`Number of trophies won`,data=df,varwidth = TRUE, main = "BOXPLOT OF IPL OVER THE YEARS",xlab="Win percentage",ylab = "Number of trophies won",col="azure")  
grid(nx=NULL,ny=NULL)  
legend("topleft", c("Win percentage"), fill=c("black"))**

**#SCATTER PLOT**

**plot(df$`Win percentage`,df$`Number of trophies won`, main = "SCATTER PLOT OF IPL OVER YEARS", xlab="WIN PERCENTAGE",ylab="NUMBER OF TROPHIES WON",pch = 20)  
x <- df$`Win percentage`  
y <- df$`Number of trophies won`  
model <- lm(y ~ x, data = df)  
abline(model, col = "blue")  
legend("topleft", c("Number of trophies won","straight line"), fill=c("black","blue"))**

**#HISTOGRAM**

**d<-read.csv("IPL\_trophy\_over\_years.csv")  
y<- d$Win.percentage  
h<- hist(y,10,breaks = 5,main = "HISTOGRAM OF IPL OVER YEARS",xlab ="WIN PERCENTAGE",ylab ="FREQUENCY",col = "azure" )  
legend("topleft", c("Frequency","normal curve overlay"), fill=c("azure","blue"))  
mn<- mean(y)  
stDt<- sd(y)  
x<- seq(20,60,1)  
y1<- dnorm(x,mean = mn,sd=stDt)  
y1<- y1\*diff(h$mids[1:2])\*length(y)  
lines(x,y1,col="blue",lwd=2)  
dev.off()**

**3.ANALYSIS:**

Code for the analysis:

**library(tidyverse)**

**options(warn=-1)**

**dataset <- read.csv("IPL\_trophy\_over\_years.csv")**

**x <- dataset$Win.percentage**

**y <- dataset$Number.of.trophies.won**

**new\_dataset <- spread(dataset,Win.percentage,Number.of.trophies.won)**

**summary(new\_dataset[,c(10:13)])**

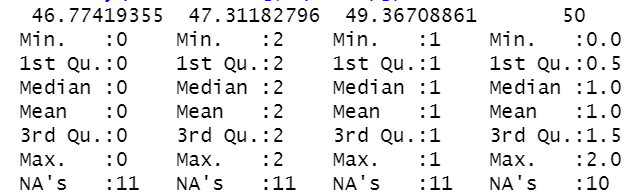
**summary(dataset)**

**cor(dataset[,2:6])**

**cor.test(dataset$Win.percentage,dataset$Number.of.trophies.won,method=c("pearson"),alt="greater",conf.level=0.99)**

**pairs(dataset[,2:6])**

**3.1. STASTISTICAL ANALYSIS:**

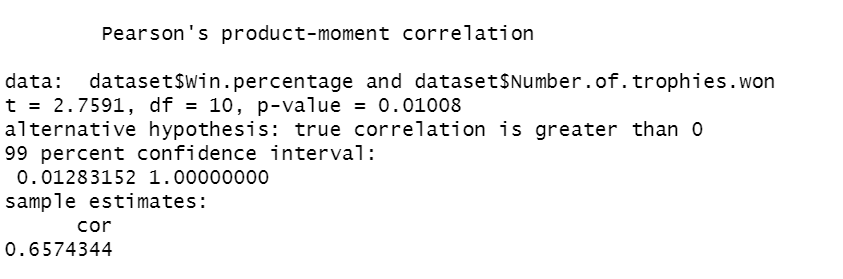
Upon performing statistical analysis of the IPL trophy dataset it results mean, median, minimum and maximum values and also shows the 1st and 3rd quartiles. For KXIP being minimum,1st quartile, median, mean,3rd quartile, and maximum values are zero and NA’s as 11. For SRH being minimum, 1st quartile, median, mean,3rd quartile, and maximum values are two and NA’s as 11. For RR being minimum, 1st quartile, median, mean,3rd quartile, and maximum values are one and NA’s as 11. Win percentage is 50 for both the teams which are KKR and RPS being minimum as 0.0, 1st quartile as 0.5, median as 1.0, mean as 1.0, 3rd quartile as 1.5 and maximum value as 2.0 and NA’s as 10.

**3.2. PEARSON’S CORRELATION TEST:**

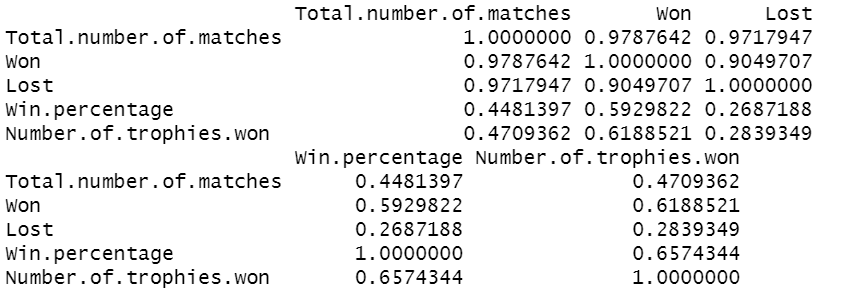
Performing Pearson’s method for both X and Y axis which are win percentage and number of trophies won resulted as t=2.7591, df=10, p-value=0.2016. True correlation is greater than 0, we found a 95% confidence interval. The correlation between win percentage and the number of trophies won is 0.66.

H0: There is no correlation between the number of trophies won and the percentage of matches won by an IPL team.

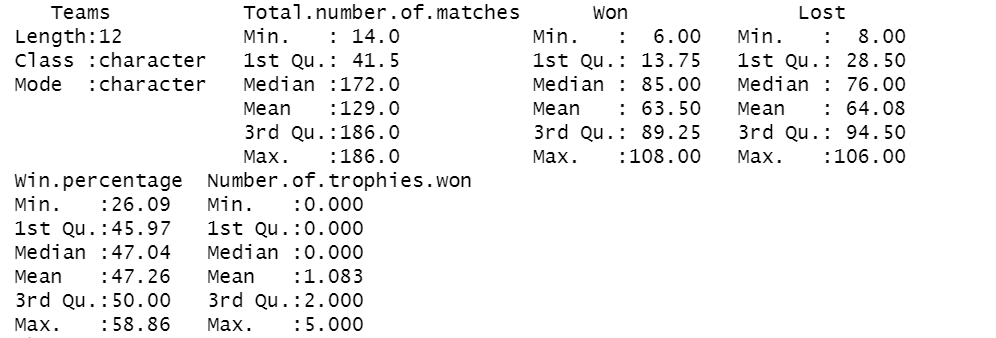
H1: There is a correlation between the number of trophies won and the percentage of matches won by an IPL team.



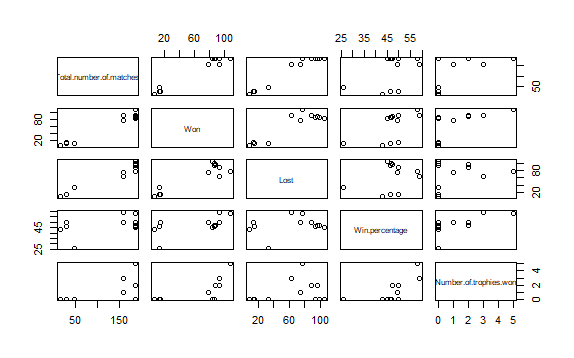
Then performed correlation in the dataset for the numerical data. The result is displayed below.



Summary of the dataset is displayed below:



The Pairs correlation in the dataset for the numerical data. The result is displayed below:



**4. CONCLUSION:**

From all the above tests and visualization, we observed that alternative hypothesis (H1): There is a correlation between the number of trophies won and the percentage of matches won by an IPL team is TRUE compared to the null hypothesis (H0). When there is an increase in win percentage there is an increase in the number of trophies won by the teams except for the teams CSK and RCB. Observation from the visualization and analysis we found out the graph is more to the linear side so there is a positive correlation which is greater than zero.