**INT 217 PROJECT REPORT**

DATA SCIENCE

***IPL DATA Analysis 2008 - 2019***

Submitted by

**Animesh Shukla**

Registration No: 11907984

Computer Science and Engineering

Section: KM004

Course Code INT217

Logo

Description automatically generated

Under the Guidance of

**Sandeep Kaur**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**DECLARATION**

I, Animesh Shukla , student of Lovely Professional University of CSE Department of Lovely Professional University, Punjab, hereby declare that all the analysis done and information given on this report have been performed and completed by my own.

Name of the student: Animesh Shukla

Date:12 December 2021

Signature:

Text, letter

Description automatically generated

**ACKNOWLEDGEMENT**

A **Project** is a temporary, unique and progressive attempt or endeavor made to produce some kind of a tangible or intangible result (a unique product, service, benefit, competitive advantage, etc.). It usually includes a series of interrelated tasks that are planned for execution over a fixed period of time and within certain requirements and limitations such as cost, quality, performance, others.

In successfully completing this project, many people have helped me. I would like to thank all those who are related to this project . Primarily, I would thank God for being able to complete this project with success. Then I will thank my subject teacher Ms Sandeep Kaur, under whose guidance I learned a lot about this project. Her suggestions and directions have helped in the completion of this project.

At last, I would like to extend my heartfelt thanks to my parents because without their help this project would not have been successful. Finally, I would like to thank my dear friends who have been with me all the time.

## Contents

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Sr.No. | Particulars | | 1. | Introduction | | 2. | Scope of The Analysis | |  |  | | 3. | ETL Process | | 4. | Source of the dataset | | 5. | Analysis and Results | | 6. | Dashboards | | 7. | References | |  | |

Introduction

The **Indian Premier League** (**IPL**) is a professional men's [Twenty20 cricket](https://en.wikipedia.org/wiki/Twenty20_cricket) league, contested by ten teams based out of ten Indian cities .The league was founded by the [Board of Control for Cricket in India](https://en.wikipedia.org/wiki/Board_of_Control_for_Cricket_in_India) (BCCI) in 2007. It is usually held between March and May of every year and has an exclusive window in the [ICC Future Tours Programme](https://en.wikipedia.org/wiki/ICC_Future_Tours_Programme).

The IPL is the most-attended cricket league in the world and in 2014 was ranked sixth by average attendance among all sports leagues. In 2010, the IPL became the first sporting event in the world to be broadcast live on [YouTube](https://en.wikipedia.org/wiki/YouTube). The brand value of the IPL in 2019 was ₹47,500 crore (US$6.3 billion), according to [Duff & Phelps](https://en.wikipedia.org/wiki/Duff_%26_Phelps). According to BCCI, the 2015 IPL season contributed ₹1,150 crore (US$150 million) to the [GDP](https://en.wikipedia.org/wiki/GDP) of the [Indian economy](https://en.wikipedia.org/wiki/Economy_of_India). The 2020 IPL season set a massive viewership record with 31.57 million average impressions and with an overall consumption increase of 23 per cent from the 2019 season.

There have been [fourteen seasons](https://en.wikipedia.org/wiki/List_of_Indian_Premier_League_seasons_and_results) of the IPL tournament. The current IPL title holders are the [Chennai Super Kings](https://en.wikipedia.org/wiki/Chennai_Super_Kings), winning the [2021 season](https://en.wikipedia.org/wiki/2021_Indian_Premier_League). The venue for the [2020 season](https://en.wikipedia.org/wiki/2020_Indian_Premier_League) was moved due to the [COVID-19 pandemic](https://en.wikipedia.org/wiki/COVID-19_pandemic_in_India) and games were played in the [United Arab Emirates](https://en.wikipedia.org/wiki/United_Arab_Emirates).

The [Indian Cricket League](https://en.wikipedia.org/wiki/Indian_Cricket_League) (ICL) was founded in 2007, with funding provided by [Zee Entertainment Enterprises](https://en.wikipedia.org/wiki/Zee_Entertainment_Enterprises). The ICL was not recognized by the [Board of Control for Cricket in India](https://en.wikipedia.org/wiki/Board_of_Control_for_Cricket_in_India) (BCCI) or the [International Cricket Council](https://en.wikipedia.org/wiki/International_Cricket_Council) (ICC) and the BCCI were not pleased with its committee members joining the ICL executive board. To prevent players from joining the ICL, the BCCI increased the prize money in their own domestic tournaments and also imposed lifetime bans on players joining the ICL, which was considered a rebel league by the board.

On 13 September 2007, on the back of India's victory at the [2007 T20 World Cup](https://en.wikipedia.org/wiki/2007_T20_World_Cup), BCCI announced a franchise-based [Twenty20 cricket](https://en.wikipedia.org/wiki/Twenty20_cricket) competition called Indian Premier League. The first season was slated to start in April 2008, in a "high-profile ceremony" in [New Delhi](https://en.wikipedia.org/wiki/New_Delhi). BCCI vice-president [Lalit Modi](https://en.wikipedia.org/wiki/Lalit_Modi), who spearheaded the IPL effort, spelled out the details of the tournament including its format, the prize money, franchise revenue system and squad composition rules. It was also revealed that the IPL would be run by a seven-man governing council composed of former India players and BCCI officials and that the top two teams of the IPL would qualify for that year's [Champions League Twenty20](https://en.wikipedia.org/wiki/Champions_League_Twenty20). Modi also clarified that they had been working on the idea for two years and that the IPL was not started as a "knee-jerk reaction" to the ICL. The league's format was similar to that of the [Premier League](https://en.wikipedia.org/wiki/Premier_League) of England and the [NBA](https://en.wikipedia.org/wiki/National_Basketball_Association) in the United States.

In order to decide the owners for the new league, an [auction](https://en.wikipedia.org/wiki/Auction) was held on 24 January 2008 with the total base prices of the franchises costing around $400 million. At the end of the auction, the winning bidders were announced, as well as the cities the teams would be based in: [Bangalore](https://en.wikipedia.org/wiki/Bangalore), [Chennai](https://en.wikipedia.org/wiki/Chennai), [Delhi](https://en.wikipedia.org/wiki/Delhi), [Hyderabad](https://en.wikipedia.org/wiki/Hyderabad), [Jaipur](https://en.wikipedia.org/wiki/Jaipur), [Kolkata](https://en.wikipedia.org/wiki/Kolkata), [Mohali](https://en.wikipedia.org/wiki/Mohali), and [Mumbai](https://en.wikipedia.org/wiki/Mumbai). In the end, the franchises were all sold for a total of $723.59 million. The Indian Cricket League soon folded in 2008.

Some IPL Teams



Scope of The Analysis

The Indian Premier League, also known as IPL and branded as Dream11 Indian Premier League 2020, was the thirteenth season of the IPL, a professional Twenty20 cricket (T20) league established by the Board of Control for Cricket in India (BCCI) in 2007.  
I have used Microsoft Excel for the analysis.

**Objectives:**

# Objective 1: Performance of each team in every season of IPL.

# Objective 2: Total number of matches held in each stadium in every Season.

* Objective 3: Team winning matches by runs in every season.
* Objective 4: Team winning matches by wickets in every season.

# Objective 5: Most Player of the Match win by any team in each season.

* Objective 6: Percentage of toss decisions after winning toss.

ETL(Extract-Transfer-Load) Process

ETL stands for Extraction, Transformation, and loading. ETL is a process which is used to Extract data, Transform the data and loading of the data to the final source. ETL follows a process of loading the data from the source system to the Data Warehouse.

### **Extraction**

Extraction is the first process where data from different sources like text file, XML file, Excel file, or various other sources are collected.

### **Transformation**

Transformation is the second step of the ETL process, where all the collected data has been transformed into the same format. The format can be anything as per our requirement. In this step, a set of rules of functions are applied to the extracted data to convert it into a single standard format. It may involve the following tasks:

* **Filtering:** Only specific attributes are loading into the data warehouse.
* **Cleaning:** Filling up the null values with specific default values.
* **Joining:** Join the multiple attributes into the one.
* **Splitting:** Splitting the single attribute into multiple attributes.
* **Sorting:** Sort the tuples based on the attributes.

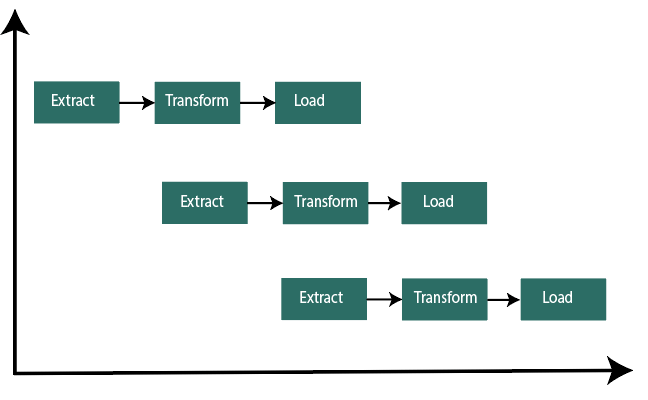
### **Loading**

Loading is the final step of the ETL process. The big chunk of data is collected from various sources, transformed them, and finally loaded to the data warehouse.

ETL is a process to extract the data from different source systems, transform the data, and load the data into the data warehouse. ETL process requires active input from various stakeholders including, developers, analysts, testers, top executive.

ETL (Extract, Transform and Load) is an automated process of extracting the information from the raw data which is required for analysis and transforms it into a format that can serve business needs and loads it into a data warehouse. ETL typically summarize data to reduce its size and improve performance for a specific type of analysis.

**ETL process uses the pipelining concept. In this concept, as soon as the data is extracted, it can be transformed, and during the period of transformation, new data can be obtained. And when the modified data is being loaded into the data warehouse, the already extracted data can be transformed.**



When we build an ETL infrastructure, we must integrate data sources, carefully plan, and test to ensure that we transform the source data correctly.

## Source of The Dataset

# The Dataset used in this project is taken from Kaggle named “IPL\_Data\_Set”

# Li

Link to Dataset:

https://www.kaggle.com/ramjidoolla/ipl-data-set

KAGGLE

**Kaggle**, a subsidiary of [Google LLC](https://en.wikipedia.org/wiki/Google_LLC), is an online community of [data scientists](https://en.wikipedia.org/wiki/Data_science) and [machine learning](https://en.wikipedia.org/wiki/Machine_learning) practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

Kaggle got its start in 2010 by offering machine learning competitions and now also offers a public data platform, a cloud-based workbench for data science, and Artificial Intelligence education. Its key personnel were Anthony Goldbloom and [Jeremy Howard](https://en.wikipedia.org/wiki/Jeremy_Howard_(entrepreneur)). [Nicholas Gruen](https://en.wikipedia.org/wiki/Nicholas_Gruen) was founding chair succeeded by [Max Levchin](https://en.wikipedia.org/wiki/Max_Levchin) . Equity was raised in 2011 valuing the company at $25 million. On 8 March 2017, Google announced that they were acquiring Kaggle.

## **Kaggle community**

In June 2017, Kaggle announced that it passed 1 million registered users, or Kagglers, and as of 2021 has over 8 million registered users. The community spans 194 countries. It is a diverse community, ranging from those just starting out to many of the world's best known researchers.

Kaggle [competitions](https://www.kaggle.com/competitions) regularly attract over a thousand teams and individuals. Kaggle's community has thousands of [public datasets](https://www.kaggle.com/datasets) and [code snippets](https://www.kaggle.com/kernels) (called "kernels" on Kaggle). Many of these researchers publish papers in peer-reviewed journals based on their performance in Kaggle competitions.

By March 2017, the [Two Sigma Investments](https://en.wikipedia.org/wiki/Two_Sigma_Investments) fund was running a competition on Kaggle to code a trading algorithm.

## **Kaggle's services**

* [Machine learning competitions](https://www.kaggle.com/competitions): this was Kaggle's first product. Companies post problems and machine learners compete to build the best algorithm, typically with cash prizes.
* [Kaggle Kernels](https://www.kaggle.com/kernels): a cloud-based workbench for data science and machine learning. Allows data scientists to share code and analysis in Python, R and R Markdown. Over 150K "kernels" (code snippets) have been shared on Kaggle covering everything from sentiment analysis to object detection.
* [Public datasets platform](https://www.kaggle.com/datasets): community members share datasets with each other. Has datasets on everything from bone x-rays to results from boxing bouts.
* [Kaggle Learn](https://www.kaggle.com/learn/overview): a platform for AI education in manageable chunks.

**How Kaggle’s competitions work**

1. The competition host prepares the data and a description of the problem.
2. Participants experiment with different techniques and compete against each other to produce the best models. Work is shared publicly through Kaggle Kernels to achieve a better benchmark and to inspire new ideas. Submissions can be made through Kaggle Kernels, through manual upload or using the Kaggle [API](https://en.wikipedia.org/wiki/Application_programming_interface). For most competitions, submissions are scored immediately (based on their predictive accuracy relative to a hidden solution file) and summarized on a live leaderboard.
3. After the deadline passes, the competition host pays the prize money in exchange for "a worldwide, perpetual, irrevocable and royalty-free license [...] to use the winning Entry", i.e. the algorithm, software and related [intellectual property](https://en.wikipedia.org/wiki/Intellectual_property) developed, which is "non-exclusive unless otherwise specified".

Alongside its public competitions, Kaggle also offers private competitions limited to Kaggle's top participants. Kaggle offers a free tool for data science teachers to run academic machine learning competitions, [Kaggle In Class](https://inclass.kaggle.com/). Kaggle also hosts recruiting competitions in which data scientists compete for a chance to interview at leading data science companies like [Facebook](https://en.wikipedia.org/wiki/Facebook), Winton Capital, and [Walmart](https://en.wikipedia.org/wiki/Walmart).

## **Impact of Kaggle competitions**

Kaggle has run hundreds of machine learning competitions since the company was founded. Competitions have ranged from improving gesture recognition for [Microsoft Kinect](https://en.wikipedia.org/wiki/Microsoft_Kinect)[[10]](https://en.wikipedia.org/wiki/Kaggle#cite_note-10) to making a [football](https://en.wikipedia.org/wiki/Soccer) [AI](https://en.wikipedia.org/wiki/Artificial_intelligence) for [Manchester City](https://en.wikipedia.org/wiki/Manchester_City) to improving the search for the [Higgs boson](https://en.wikipedia.org/wiki/Higgs_boson) at [CERN](https://en.wikipedia.org/wiki/CERN).

Competitions have resulted in many successful projects including furthering the state of the art in HIV research, chess ratings and traffic forecasting.[[14]](https://en.wikipedia.org/wiki/Kaggle#cite_note-14) Most famously, [Geoffrey Hinton](https://en.wikipedia.org/wiki/Geoffrey_Hinton) and George Dahl used deep [neural networks](https://en.wikipedia.org/wiki/Neural_networks) to win a competition hosted by [Merck](https://en.wikipedia.org/wiki/Merck_%26_Co.). And Vlad Mnih (one of Hinton's students) used deep neural networks to win a competition hosted by Adzuna. This helped show the power of deep neural networks and resulted in the technique being taken up by others in the Kaggle community. Tianqi Chen from the University of Washington also used Kaggle to show the power of [XGBoost](https://en.wikipedia.org/wiki/Xgboost), which has since taken over from [Random Forest](https://en.wikipedia.org/wiki/Random_forest) as one of the main methods used to win Kaggle competitions.

Several academic papers have been published on the basis of findings made in Kaggle competitions.[[15]](https://en.wikipedia.org/wiki/Kaggle#cite_note-15) A key to this is the effect of the live leaderboard, which encourages participants to continue innovating beyond existing best practice.[[16]](https://en.wikipedia.org/wiki/Kaggle#cite_note-16) The winning methods are frequently written up on the Kaggle blog, [*Kaggle Winner's Blog*](https://medium.com/kaggle-blog).

## Analysis of Dataset

#### Performance of each team in every season of IPL :

The analysis in this shows the performance of each team in different seasons of the IPL.

* 1. Specific Requirements/Functions :

i)Pivot table of Import Data

ii)Pivot table of Export Data iii)SLICERS AND HYPER LINKS ARE USED

iv)LINE Chart is Used

* 1. Analysis Results:

# we have observed that in over the years the performance of Chennai super kings has gone down from 2008 to 2019.

# Chart, line chart Description automatically generated

#### Total number of matches held in each stadium in every Season:

* 1. Introduction: The analysis shows that how many times matches have been held in a stadium in each season.

.

* 1. Specific Requirements/Functions and Formulas: i) Pivot table of the data of Import

ii)With the help of data Clustered Bar chart is plotted. iii)Slicers are used to change the season of the ipl.

* 1. Analysis Results:

By this we came to know that the stadiums MA Chidambaram Stadium and Rajiv Gandhi International Cricket stadium held the most no. of matches in 2019.

* 1. Visualization:

Graphical user interface

Description automatically generated

#### Team winning matches by runs in every season:

* 1. Introduction: The analysis shows the teams matches wins by runs.
  2. Specific Requirements/Functions and Formulas:
     + Pivot table of the imported data
     + With the help of data Clustered Bar chart is plotted.
     + With the help of Slicers,we can change season. c)Analysis Results:
     + By bar graph we can observe that Mumbai Indians have won most

matches by runs in 2019.

Chart

Description automatically generated

#### Team winning matches by wickets in every season:

* 1. Introduction: The analysis shows the teams matches wins by runs.
  2. Specific Requirements/Functions and Formulas:

i)Pivot table of the data of wickets and year.

ii)With the help of data Clustered Bar chart is plotted. c)Analysis Results:

* + - Kolkata Knight Riders is the teams which wins the most matches by wicket taking in 2019.

Chart, bar chart

Description automatically generated

#### Most Player of the Match win by any team in each season:

* 1. Introduction: The analysis shows the total player of the match award win by any team in each season.

b)Specific Requirements/Functions and Formulas:

1. Pivot table of the data total Player of the match

ii)With the help of data clustered column chart is plotted.

iii) Slicers is used.

c)Analysis Results:

we can observe that most Player of the match win by any team in each season.

Chart, bar chart

Description automatically generated

#### PERCENTAGE OF TOSS DECISION:

* 1. Introduction: The analysis shows that what a team chooses after winning a toss, batting or fielding.

. b)Specific Requirements/Functions and Formulas:

1. Pivot table of the data of toss decision ii)With the help of data pie chart is plotted.

c)Analysis Results:

From the pie chart, we can observe that most of teams are choosing field option after winning the toss

Chart

Description automatically generated

# DASHBOARDS OF THE PROJECT

Chart, line chart

Description automatically generated

Graphical user interface

Description automatically generated

## References

1. [www.kaggle.com](http://www.kaggle.com/)
2. [www.youtube.com](http://www.youtube.com/)
3. [www.google.com](http://www.google.com/)