**INTRODUCTION TO DATA MANAGEMENT**

**PROJECT REPORT**

(Project Semester August-December 2020)

***IPL DATA DASHBOARD***

Submitted by

**Pankaj Singh**

**11808783**

Programme and Section – Bachelor of Technology(CSE) – KM078

Course Code – INT217

Under the Guidance of

**Mrs Savleen Kaur – 18306**

**Asst. Professor, Database Systems**

**Discipline of CSE/IT**

**School of Computer Science & Engineering**

**Lovely Professional University, Phagwara**



**DECLARATION**

I, Pankaj Singh, student of B.Tech under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date:

Registration No.: 11808783 Pankaj Singh

**ACKNOWLEDGEMENT**

The success and outcome of this project required a lot of guidance and assistance from many people. All that I have done is only due to such supervision and assistance and I would not forget to thank them.

I respect and thank *Mrs. Savleen Kaur*, for providing me all support and guidance, which made me complete the project duly. I am extremely thankful to her for providing such a nice support and guidance.

I am thankful to and fortunate enough to get constant encouragement, support, and guidance from “School of Computer Science” which helped me in successfully completing my project work.

**TABLE OF CONTENT**

1. Introduction
2. Objectives/Scope of the Analysis
3. Source of dataset
4. ETL process
5. Analysis on dataset (for each analysis)
6. Introduction
7. General Description
8. Specific Requirements, functions and formulas
9. Analysis results
10. Visualization
11. References
12. Bibliography

**INTRODUCTION**

What is MS Excel?

Microsoft Excel is a helpful and powerful program for data analysis and documentation. It is a spreadsheet program, which contains the number of columns and rows, where each intersection of a column and a row is a “cell.” Each cell contains one point of data or one piece of information. By organizing the information in this way, you can make information easier to find, and automatically draw information from changing data.

The Excel Interface

**Home Tab**: The Home tab contains the most common text editing tools used in Excel.

Graphical user interface, application, Word

Description automatically generated

**Shortcut Menu**: The most common Excel commands and functions can be accessed quickly by using the shortcut menu. To access this feature, simply right-click the element you wish to edit if you are using a PC, or control-click the element you wish to edit if you are using a Mac. The options displayed will vary depending on the element you have selected.

**Formula Bar**: This will be one of the most useful tools as you use Excel. The formula bar allows you to see all the details and methods used to return what is seen in a cell. Whenever you input any information into a cell, the output, or result is what is shown once you move away from the cell. This is most prominent when using functions, as you do not see the whole equation in the cell in the worksheet, only the result. The formula bar is located below the ribbon and spans most of the window.

Graphical user interface, application

Description automatically generated

It is possible to hide the formula bar, though is not recommended. In order to hide or show the formula bar if it has mistakenly been hidden, go to **Excel Options** at the bottom right of the menu that opens when you click the Office button. Go to the **Advanced** option, and under **Display**, you can click the check box for **Show formula bar**. When done, simply click **Ok**.

Graphical user interface, application

Description automatically generated

**Adding a Worksheet**: By default, three worksheets are included in each Excel workbook. You can access the different worksheets by clicking the worksheet tabs just above the status bar.

Graphical user interface, application

Description automatically generated

To add a new worksheet, click the **Insert Worksheet** tab, to the right of the existing worksheet tabs.

**Modifying a Worksheet**

Adding a Cell: A row runs horizontally across a worksheet. To add a row to a worksheet:

|  |
| --- |
| Steps Actions   1. Click the worksheet where you would like the row to appear. 2. Click the Home tab. 3. Click the arrow on the Insert button in the Cells group. 4. Click Insert Sheet Rows |

Graphical user interface, application, Word

Description automatically generated

**Selecting a cell**: Before you can modify or format a cell, you must first select it. The table below will show you the quickest way to select cells.

|  |  |
| --- | --- |
| Cells to Select | Mouse Action |
| One cell | Click once in the cell |
| Entire row | Click the row heading |
| Entire column | Click the column heading |
| Entire worksheet | Click the **Select All** button located above the row headings and to the left of the column headings or press **Ctrl + A** on your keyboard. |
| Cluster of cells | Click and drag the mouse over the cell cluster. |

Formatting cells

The options for formatting cells can be accessed from the Format Cells dialog box. To open the Format Cells dialog box, do one of the following.

* Select the cell you want to format. PC users: right click the cell. Mac users: press Control and click the cell. From the shortcut menu, click **Format Cells**.
* On the **Home** tab, in the **Number** group, click the arrow.

Graphical user interface, application

Description automatically generated

The Format Cells dialog box has the following options for formatting cells:

* **Number tab** – allows you to specify the numerical data type, for example currency, date, percentage.
* **Alignment tab** – allows you to change the position and alignment of the data within a cell.
* **Font tab** – allows you to change the attributes of cell font, including font face, size, style, and color.
* **Border** – allows you to select a border style for your cell from a variety of different options.
* **Fill** – allows you to add shading and background colour to a cell.
* **Protection** – allows you to lock or hide a cell.

Functions and Formulas

An important feature of the Excel spreadsheet program is that it allows you to create formulas that will automatically calculate results. Without formulas, a spreadsheet is not much more than a large table for displaying text.

**Formulas**: A formula is an equation that makes calculations based on the data in your spreadsheet. Formulas are entered in a cell in your worksheet. They must begin with an equal sign, followed by the addresses of the cells that will be calculated upon, with an appropriate operand placed in between. Once the formula is typed into the cell, the calculation executes immediately. The formula appears in the formula bar.



In the example below, a formula has been created for calculating the subtotal of number of textbooks. This formula multiplies the quantity and price of each textbook, and then adds the totals to give the combined cost of all books.

Graphical user interface, application, table, Excel

Description automatically generated

**Linking Worksheets**: You can create a formula that uses data from two different worksheets. This can be done within the same workbook or across different workbooks. The base formula is written as “sheet name! cell address” when linking cells from worksheets within the same workbook. The base formula is written as “[workbook name. xlsx] sheet name! cell address” when linking cells from different workbooks. For example, the value of cell A1 in Worksheet 1 and cell A2 in Worksheet 2 can be added using the formula “=A1+Sheet2! A2”. Similarly, suppose Worksheet 1 was in a workbook named Book1.xlsx, and Worksheet 2 was in a workbook called Book2.xlsx, the same cells could be added using the formula “= [Book1. xlsx] Sheet1! $ A$1+A2”. This formula would of course be entered inside Sheet2 of Book2.xlsx.

**Relative, Absolute and Mixed Referencing**: Relative referencing is the practice of calling cells by just their column and row labels (such as “A1”). When a formula contains relative referencing and it is copied from one cell to another, Excel does not create an exact copy of the formula. It will change cell addresses relative to the row and column they are moved to. For example, if a simple addition formula in cell C1 “=(A1+B1)” is copied to cell C2, the formula would change to “=(A2+B2)” to reflect the new row.

To prevent this from happening, cells must be called by absolute referencing. This is accomplished by placing dollar signs “$” within the cell addresses in the formula. Continuing the previous example, if the formula in cell C1 reads “=($A$1+$B$1)”, the value of cell C2 will be the sum of cells A1 and B1. Both the column and row of both cells are absolute and will not change when copied. Mixed referencing can also be used where the row OR column is fixed, but not both. For example, in the formula “=(A$1+$B2)”, the row of cell A1 is fixed and the column of cell B2 is fixed.

**Basic Functions**: Functions can be a more efficient way of performing mathematical operations than formulas. For example, if you wanted to add the values of cells D1 through D10, you would type the formula “=D1+D2+D3+D4+D5+D6+D7+D8+D9+D10”. A shorter way would be to use the SUM function and simply type “=SUM (D1:D10)”. Several other function commands and examples of functions are given in the table below:

|  |  |  |
| --- | --- | --- |
| Function | Example | Description |
| SUM | =SUM(A1:A100) | Finds the sum of cells A1 through A100 |
| AVERAGE | =AVERAGE(B1:B10) | Finds the average of cells B1 through B100 |
| MAX | =MAX(C1:C100) | Returns the highest number from C1 through C100 |
| MIN | =MIN(D1:D100) | Returns the lowest number from D1 through D100 |
| SQRT | =SQRT(D10) | Finds the square root of the value in cell D10 |
| TODAY | =TODAY()\_ | Returns the current date. |

**Charts**

Charts allow you to present data entered into a worksheet in a visual format, using a variety of graph types. Before you can make a chart you must first enter data into a worksheet.

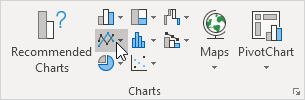
Create a chart: To create a line chart, execute the following steps.

1. Select the range A1:D7.

Table

Description automatically generated

2. On the Insert tab, in the Charts group, click the Line symbol.



3. Click Line with Markers.

Graphical user interface, application, Word

Description automatically generated

Result:

Chart, line chart

Description automatically generated

**Change Chart Type**

You can easily change to a different type of chart at any time.

1. Select the chart.

2. On the Design tab, in the Type group, click Change Chart Type.

Graphical user interface, application

Description automatically generated

Graphical user interface, chart, bar chart

Description automatically generated

3. On the left side, click Column.

4. Click OK.

Result:

Chart, bar chart

Description automatically generated

### Switch Row/Column:

If you want to display the animals (instead of the months) on the horizontal axis, execute the following steps.

1. Select the chart.

2. On the Design tab, in the Data group, click Switch Row/Column.

Graphical user interface, application

Description automatically generated

Result:

Chart, bar chart

Description automatically generated

### Legend Position:

To move the legend to the right side of the chart, execute the following steps.

1. Select the chart.

2. Click the + button on the right side of the chart, click the arrow next to Legend and click Right.

Diagram

Description automatically generated

**Result:**

Chart, bar chart

Description automatically generated

### Data Labels:

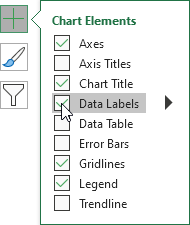
You can use data labels to focus your readers' attention on a single data series or data point.

1. Select the chart.

2. Click a green bar to select the Jun data series.

3. Hold down CTRL and use your arrow keys to select the population of Dolphins in June (tiny green bar).

4. Click the + button on the right side of the chart and click the check box next to Data Labels.



**Result:**

Chart, bar chart

Description automatically generated

**Pivot Table**

Pivot tables are one of Excel's most powerful features. A pivot table allows you to extract the significance from a large, detailed data set.

**OBJECTIVE/SCOPE OF ANALYSIS**

As already mentioned, IPL is a popular cricket series played annually in India from 2007. So, we have generated a lot of data since then which helps in prediction of the future performances of teams and players.

I have analysed the data as per my knowledge and covered the various objectives which are mentioned below-

1. To analysis the player’s batting and bowling skills.
2. To compare the teams by their count number of match wins.
3. To find out the popular playgrounds.
4. To show the toss distribution of team’s session-wise.
5. To find the number of boundaries by each team session-wise.
6. To show the players distribution in IPL from different countries.
7. To find the average run rate in IPL.

**SOURCE OF DATASET**

There are various sources for different types of datasets on the internet world in which Kaggle, GitHub, DataQuest are some of the popular opensource websites which provides the data. For my project I have explored on various platforms and found a interesting dataset at [Kaggle.com](http://Kaggle.com) which contains all the details of last 10 years of IPL data. As IPL is a popular cricket series organise by BCCI every year since 2007, so there is a lot of things we can explore in data.

Link to the dataset - <https://www.kaggle.com/ramjidoolla/ipl-data-set>

**ETL PROCESS**

In most organizations, data goes through an ETL (extract, transform and load) process before it is available for reporting. During the ETL process, data is extracted from a data source, then transformed, validated, standardized, corrected, quality checked and ultimately loaded for reporting.

Typically, the ETL process is carried out by robust enterprise-grade ETL applications such as SQL Server Integration Services (SSIS) or other third-party tools. But in some organizations, business users do some of this ETL work right in Excel, often referred to as data shaping and transformation.

The challenge with this method? The ETL process in Excel is usually a time-consuming, manual process that is not easy to automate.

There is various transformation we include in ETL are –

* Remove unnecessary columns.
* Rename the column headers.
* Remove the null values.
* Data validation.
* Standardization of Data.
* Removing any redundancy and other data cleaning.

I have used Tableau Prep application for ETL process and cleaning of data. As the source dataset contains six different csv file namely *deliveries.csv, matches.csv, most\_runs\_average\_strikerate.csv, Players.xlsx, teams.csv, teamwise\_home\_and\_away.csv;* below are the process used on each of the data file –

First of all the csv files were loaded in Tableau Prep and following are the operation performed-

1. *matches.csv*

* In columns *team1, team2, toss\_winner and winner*, *Rising Pune Supergiants* is groups with *Rising Pune Supergiant*.
* Null values have been excluded from *winner* column.
* In *venues* columns, same playground with misspellings and abbreviated form grouped together.

2. *most\_runs\_average\_strikerate.csv* is renamed as batsman and *id* column is added for which contains the id of each player which can be further used for linking with other relations.

3. *Player\_id* column is added to *Player.xlsx* and file is renamed as players.

4. teamwise\_home\_and\_away.csv is renamed as teamwise\_home\_away.

5. In *deliveries.csv* following are transformation that have been made –

* In columns *batting\_team and bowling\_team*, *Rising Pune Supergiants* is groups with *Rising Pune Supergiant*.
* For a few match ids, match data is not there, so we remove those deliveries by matching the *match\_id* from *matches.csv*

**ANALYSIS ON DATASET**

***Analysis 1***

**Introduction –** Team comparison based on number of matches wins against each other teams

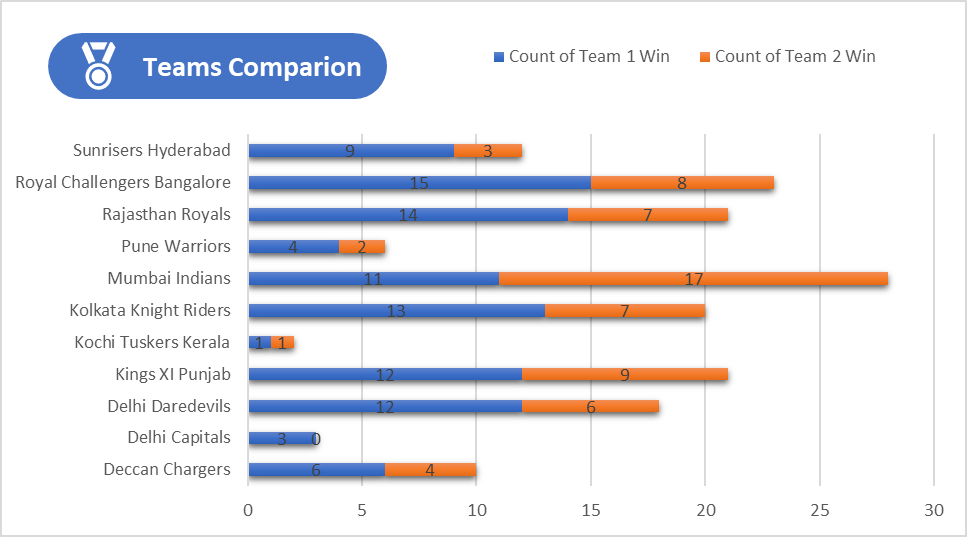
**General Description –** A team would be selected and compared with all other teams and we will able to see that which teams won how many matches against the selected team.

**Specific Requirements, functions, and formulas –** Data Pivoting and COUNTIFS

**Analysis results –** We can easily find out that that which teams dominate against each other and can their numbers of wins and loss.

**Visualization –**





***Analysis 2***

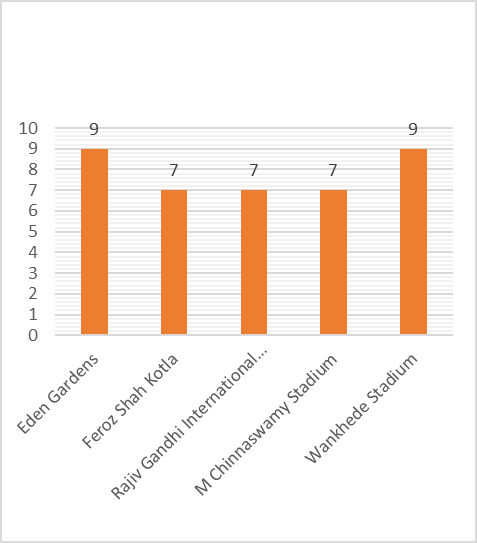
**Introduction –** Find out the top playground/venues each season

**General Description –** Top 5 playground in the selected season from slicer will be displayed on chart

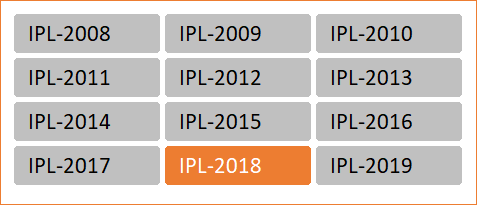
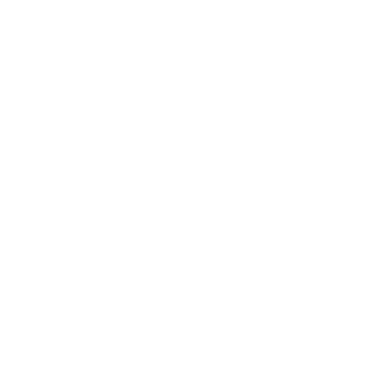
**Specific Requirements, functions, and formulas –** Table Pivoting and filter data only for top 5 venues

**Analysis results –** We can find out the popular playground year wise and check the number of matches played at that venue

**Visualization –**



**Popular Playground**



***Analysis 3***

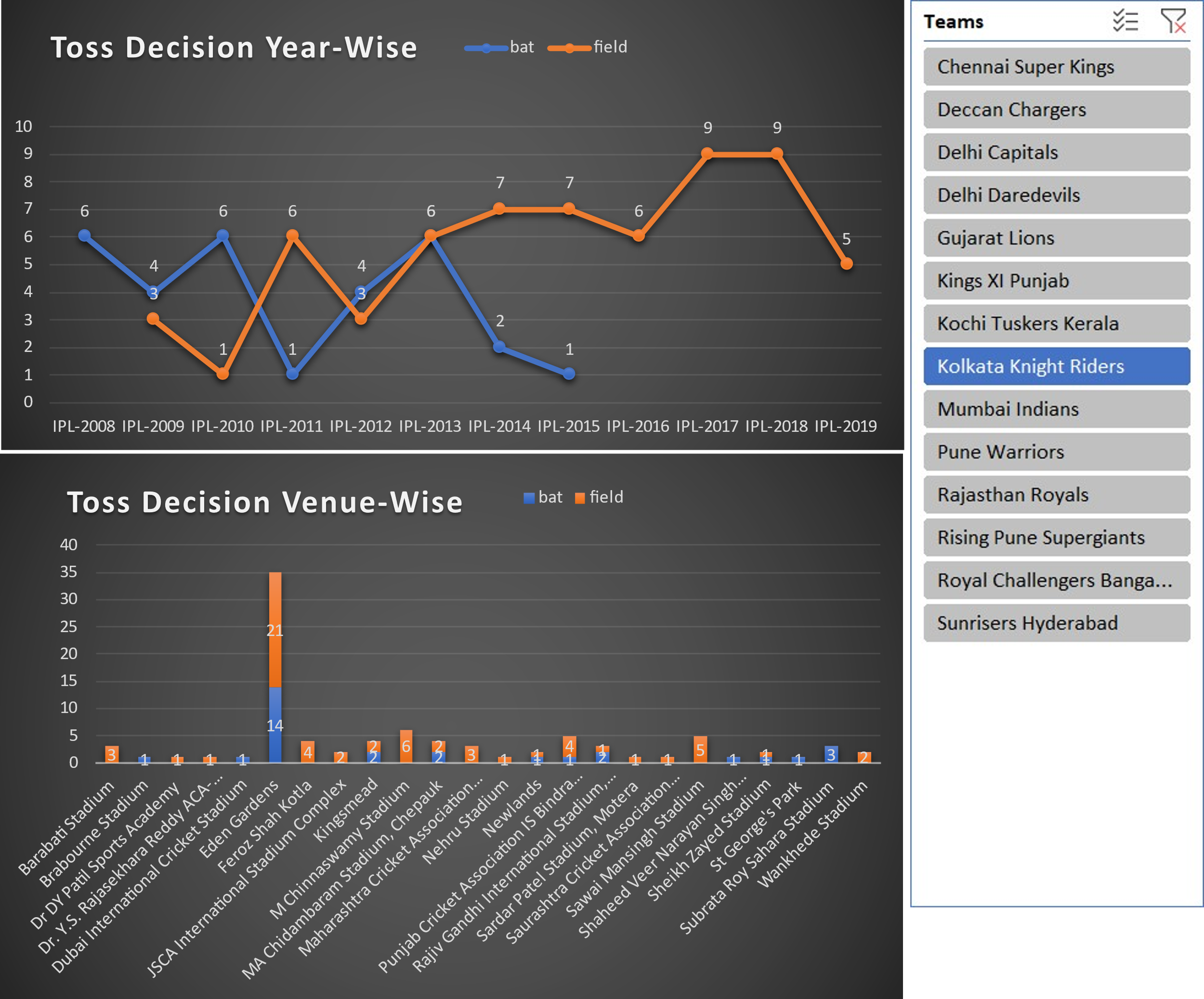
**Introduction –** Analysis of toss decision taken by teams’ year-wise and according to venues.

**General Description –** We can analyse the toss decision team-wise according to seasons and venues. We can see the decision of batting or bowling overs years and according to venues.

**Specific Requirements, functions, and formulas –** Data Pivoting and Slicer

**Analysis results –** We can see that at Wankhede, Punjab Cricket Association, M Chinnaswamy and at Eden Gardens Stadium teams used to choose bowling more than the batting.

**Visualization –**



***Analysis 4***

**Introduction –** Count of Boundaries Teamwise

**General Description –** The chart generated will show the count of boundaries based on selected season/year from a slicer as stacked chart of 4 and 6 runs.

**Specific Requirements, functions, and formulas –** Data Pivoting and Charts

**Analysis results –** Stacked chart of count of 4 and 6 team-wise.

**Visualization –**

***Analysis 5***

**Introduction –** Players distribution

**General Description –** The chart generated will show the count of boundaries based on selected season/year from a slicer as stacked chart of 4 and 6 runs.

**Specific Requirements, functions, and formulas –** Data Pivoting and Charts

**Analysis results –** Stacked chart of count of 4 and 6 team-wise.

**Visualization –**

***Analysis 6***

**Introduction –** To find out the average run over the data over-wise. We can analyse teams strategies according the this run rate.

**Specific Requirements, functions, and formulas –** Data Pivoting, Charts, Average()

**Analysis results –** We find out that in first five overs run rate gradually increases and then fall gradually in next 3-4 overs and then increase slowly till 20th over.

**Visualization –**

**REFERENCES**

* + - [**https://support.microsoft.com/en-us/excel?ui=en-us&rs=en-us&ad=us**](https://support.microsoft.com/en-us/excel?ui=en-us&rs=en-us&ad=us)
    - [**https://edu.gcfglobal.org/en/excel2016/**](https://edu.gcfglobal.org/en/excel2016/)
    - [**https://support.microsoft.com/en-us/excel?ui=en-us&rs=en-us&ad=us**](https://support.microsoft.com/en-us/excel?ui=en-us&rs=en-us&ad=us)
    - [**https://excelexposure.com**](https://excelexposure.com/)
    - [**https://spreadsheetpage.com**](https://spreadsheetpage.com/)
    - [**https://excel-bytes.com**](https://excel-bytes.com/)
    - [**https://www.excelfunctions.net**](https://www.excelfunctions.net/)
    - [**https://best-excel-tutorial.com**](https://best-excel-tutorial.com/)
    - [**http://www.homeandlearn.org**](http://www.homeandlearn.org/)
    - [**https://www.contextures.com/tipteh.html**](https://www.contextures.com/tipteh.html)
    - **https://www.kaggle.com**

**BIBLIOGRAPHY**

* Dataset from Kaggle.com

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

* Information from [www.Google.com](file:///C:\Users\shash\Desktop\www.Google.com)
* Learnt some things from YouTube channels