**INTRODUCTION TO DATA MANAGEMENT PROJECT REPORT**

August-December 2020

***ODI CRICKET MATCH ANALYSIS***

Submitted by-

Akash Kushwah

Registration No- 11803007

Section- KM076

Course Code- INT217

Under the Guidance of

**Vasudha Ma’am 23636**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**



**CERTIFICATE**

This is to certify that Akash Kushwah bearing Registration no. 11803007 has completed INT 217 project titled, ODI cricket match analysis under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

Signature and Name of the Supervisor

Designation of the Supervisor

Lovely Professional University

Phagwara, Punjab.

Date: 27/11/2020

**DECLARATION**

I, Akash Kushwah student of ............................ (Program name) 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: 27/11/2020 Signature

Registration No.11803007 Name of the student Akash Kushwah

.

**ACKNOWLEDGEMENT**

I would like to express my deepest appreciation to all those who provided me the possibility to complete this report. A special gratitude I give to my **mentor** whose contribution in stimulating suggestions and encouragement helped me to coordinate my project especially in writing this report. I express my thanks to my institution **Lovely Professional University** for giving me an opportunity to learn this interesting topic. I also convey my regards to my faculty assistance all through this training named **“Data Science Project on Cricket match analysis”.** Once again, I would like to thank all my supporters from the core of my heart.

|  |  |  |
| --- | --- | --- |
| S.NO | TITLE | PAGE |
| 1 | INTRODUCTION | **6** |
| 2 | OBJECTIVE | **7** |
| 3 | SOURCE OF DATA | **7** |
| 4 | ETL PROCESS | **8** |
| 5 | ANALYSIS ON DATASET | **9** |
| 6 | VISAULIZATION | **21** |
| 7 | LIST OF ANALYSIS WITH RESULT | **25** |
| 8 | REFERENCE | **28** |
| 9 | BIBLOGRAPHY | **29** |

**INTRODUCTION**

A simple [database, such as Excel](https://www.lifewire.com/create-a-database-in-excel-3123446), holds all information about one subject in a single table. Relational databases, on the other hand, consist of many tables with each one containing information about different, but related topics.

**FIELDS**

Each item of information in a database record, such as a telephone number or street number, is referred to as a **field**. In Excel, the individual cells of a worksheet serve as fields, since each cell can contain a single piece of information about an object.

**FIELD NAMES**

It is vital to organize a database so that you can sort or filter the data to find specific information. Adding column headings, known as field names, makes it easier to enter data in the same order for each record.

Additionally, Microsoft has several data tools to make it easier to work with vast amounts of data stored in Excel tables and to help keep it in good condition.

**USING A FORM FOR RECORDS**

One of those tools is the data form. You can use it to find, edit, enter, or delete records, in tables containing up to 32 fields or columns.

The default form includes a list of field names in the order they are arranged in the table, to ensure that users enter records correctly. Next to each field name is a text box for entering or editing the individual fields of data.

While it is possible to create custom forms, creating and using the default form is often all that is needed.

**REMOVE DUPLICATE DATA RECORDS**

A common problem with all databases is data errors. In addition to simple spelling mistakes or missing fields of data, duplicate data records can be a concern as a data table grows in size.

Another of Excel's data tools can be used to [remove these duplicate records](https://www.lifewire.com/how-to-remove-duplicates-in-excel-4691045) — either exact or partial duplicates.

**RECORDS**

In database terminology, a **record** holds all the information or data about one specific object in the database. In Excel, each [cell](https://www.lifewire.com/what-is-a-spreadsheet-cell-4178208) in a worksheet contains one item of information or value.

**SORTING DATA IN EXCEL**

Sorting means to reorganize data according to a specific property, such as sorting a table alphabetically by last names or chronologically from oldest to youngest.

Excel's sort options include sorting by one or more fields, custom sorting, such as by date or time, and [sorting by rows](https://www.lifewire.com/sort-data-in-excel-4178206) which makes it possible to reorder the fields in a table.

**Objective:**

The objective the project to analyze the data set and find the solutions for different queries like:

1. The maximum, minimum and the average runs scored by the team in 1st inning.
2. The maximum, minimum and the average runs scored by the team in 2nd inning.
3. Which team has won how many matches in which year.
4. Average runs of teams including first and second inning.
5. The winning percentage of each team in each year.
6. To minimize the whole dataset in a simple and presentable way.
7. To find the winner of the match.

**Source of data:**

The dataset has take from [www.kaggle.com](http://www.kaggle.com).

<https://www.kaggle.com/nowke9/odi-matches-first-innings-scores>.

**ETL PROCESS:**

Extraction: Data extraction is a process that involves retrieval of data from various sources. Frequently, companies extract data in order to process it further, migrate the data to a data repository (such as a data warehouse or a data lake) or to further analyze it. It’s common to [transform the data](https://www.alooma.com/blog/what-is-data-transformation) as a part of this process. For example, you might want to perform calculations on the data — such as aggregating sales data — and store those results in the data warehouse. If you are extracting the data to store it in a data warehouse, you might want to add additional metadata or enrich the data with timestamps or geolocation data. Finally, you likely want to combine the data with other data in the target data store.

The dataset on which I am working it is the structured data.

**Structured data:**

If the data is structured, the data extraction process is generally performed within the source system. It’s common to perform data extraction using one of the following methods

**Full extraction:**

Data is completely extracted from the source, and there is no need to track changes. The logic is simpler, but the system load is greater.

**Transformation:**

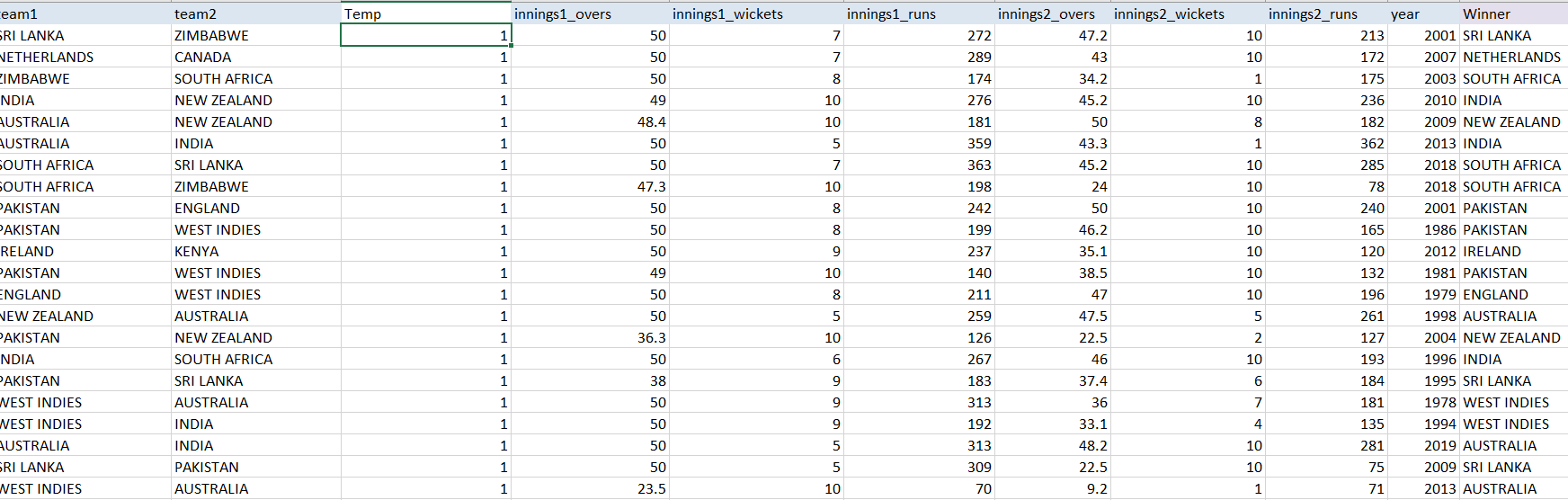
Transform is the process of converting the extracted data from its previous form into the form it needs to be in so that it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining the data with other data.

In my dataset I have add the winner column which is used to find the winner of the matches. In this I have used a **if function** to find the winner.

If(1st innings\_runs-2nd innings\_runs>0, 1st team winner,2nd team winner).

**Loading:** I had made a little bit change in my dataset. I have Added a winner column and made a change in my dataset.

**Result:**



**Analysis of dataset**

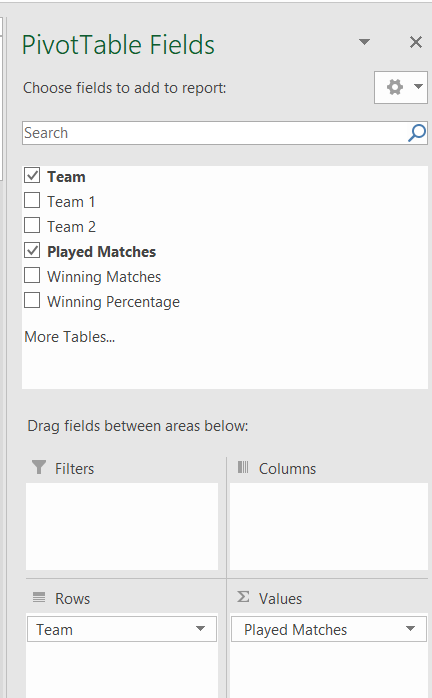
**1. Winning percentage of teams:**

In the given dataset I have used the given data to find the winning percentage of each team, In which I have used some basic mathematics calculations.

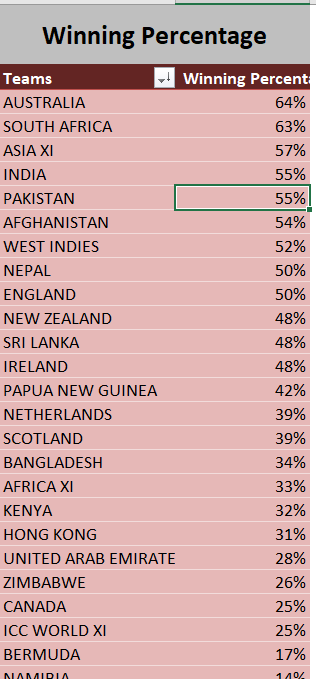
Here we are the showing the winning percentage of the team from year 1975 to 2017. We find the winning percentage of the team by using:

Winning percent = Matches wins/Total matches played\*100.

From here we get the knowledge which has won how many matches out of they had played.



**Result:**



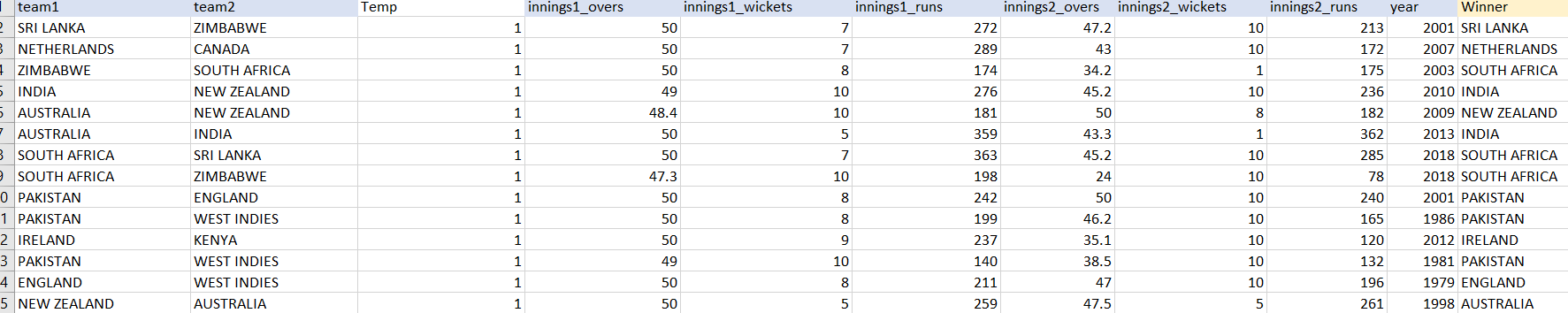
**2.Winner of the match:**

The given dataset is off cricket match in which we are finding that which team has won the match, To find the winner of the match first we have to make a new column of **winner,** and then we have to apply the if function to find the winner.

Here we are showing which team has won the match by using the if function

If(1st innings\_runs-2nd innings\_runs >0, 1st team winner,2nd team winner)

**Result:**



**3. Teams wins in each year:**

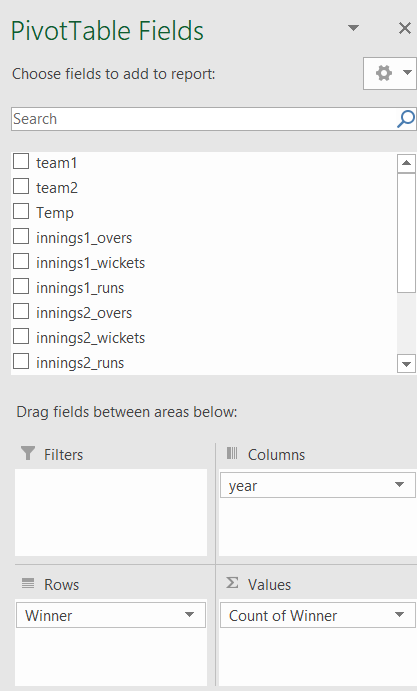
In the given data there are many teams and had played many matches from the year 1971 to 2019 so to keep the wins record of each team I have used a function to find the solution.

In this we are showing which has won how many matches in which year.

We use this formula to find the winner:

If(1st innings\_runs-2nd innings\_runs>0, 1st team winner,2nd team winner)

And then we use them in pivot table as count of winner.



**Result:**



**4.Batting first average score:**

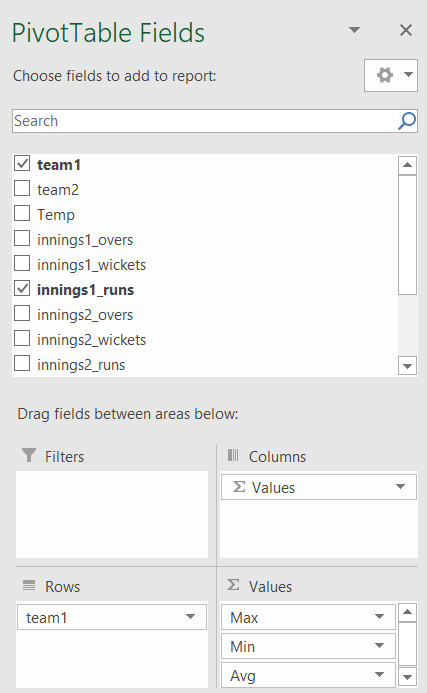
The average batting first score is the runs that score in the first inning of the match, here we have to find that how many runs had been scored In the first inning from 1971 to 2019.

In this analysis we will find the maximum, minimum and the average runs scored by the teams while batting first.

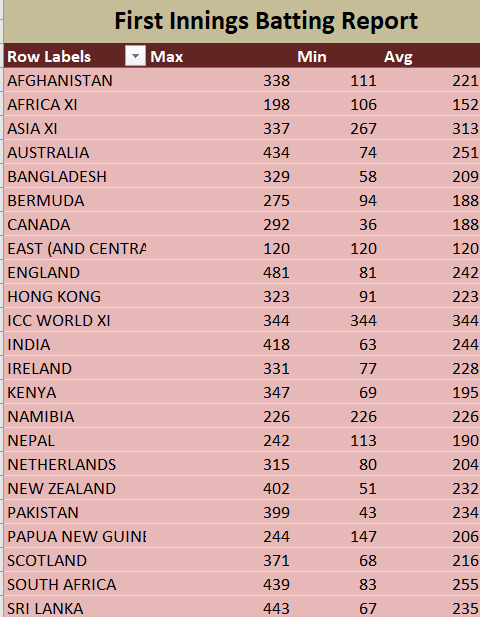
To find the average we simply use the formula:

Average runs scored in 1st inning = Total number of runs scored in 1st inning/Total number of 1st inning played.

And to specify the data more we have to use the pivot table.



**Result:**



**5.Batting second average score:**

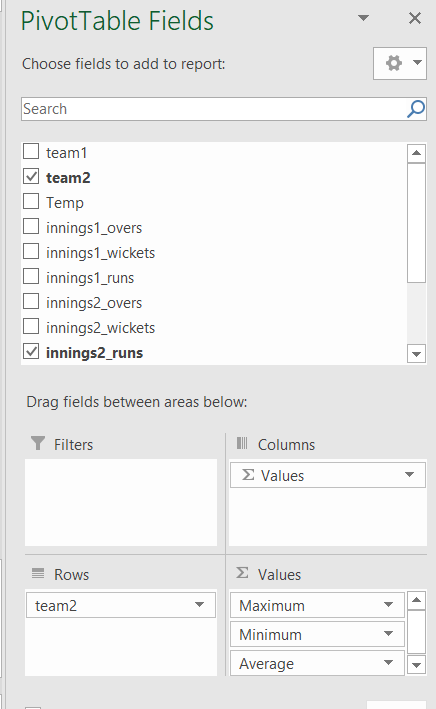
The average batting second score is the runs that score in the second inning of the match, here we have to find that how many runs had been scored In the second inning from 1971 to 2019.

In this analysis we will find the maximum, minimum and the average runs scored by the teams while batting second.

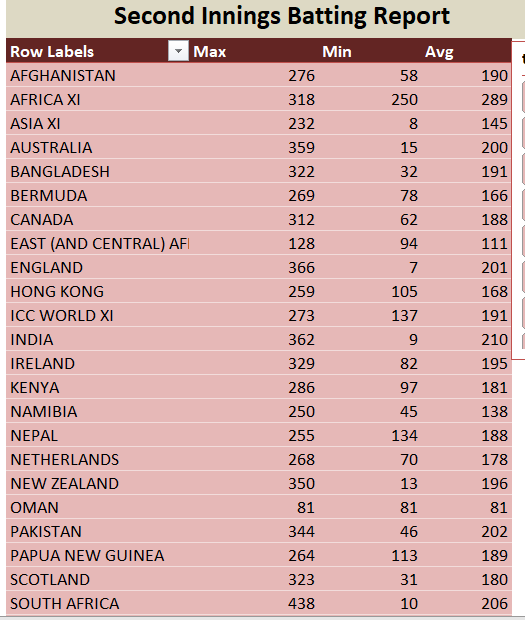
To find the average we simply use the formula:

Average runs scored in 2nd inning = Total number of runs scored in 2 inning/Total number of 2nd inning played.

And to specify the data more we have to use the pivot table.



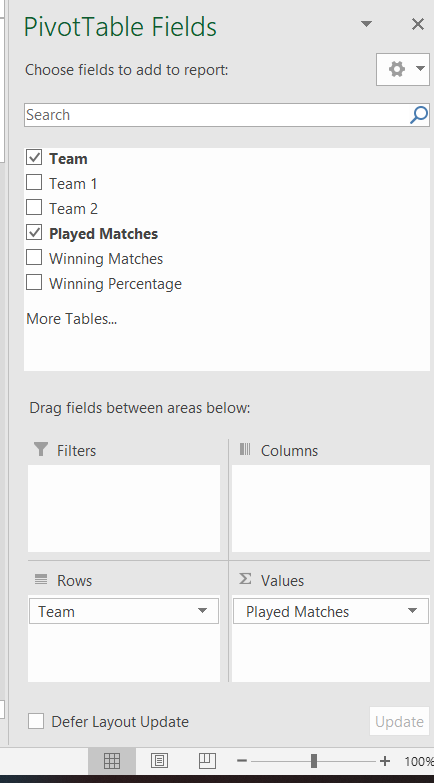
**Result:**



**6. Matches Played:**

From the dataset we have to find that which has played how many matches from the year 1971 to 2019.

To find we have to use the pivot table scenario.

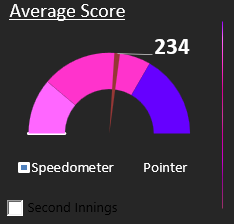
**Pivot table:** 

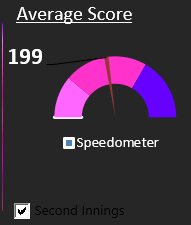
**Result:**



1. **To find the average score of both innings on one click.**

Here I have use the combination of two charts(pie chart and drought chart ) and a checkbox to find the average score of both team on one click.





1. **Use of the hyperlink to connect the page:**

In the dashboard I have to use the hyperlink to connect the different pages.

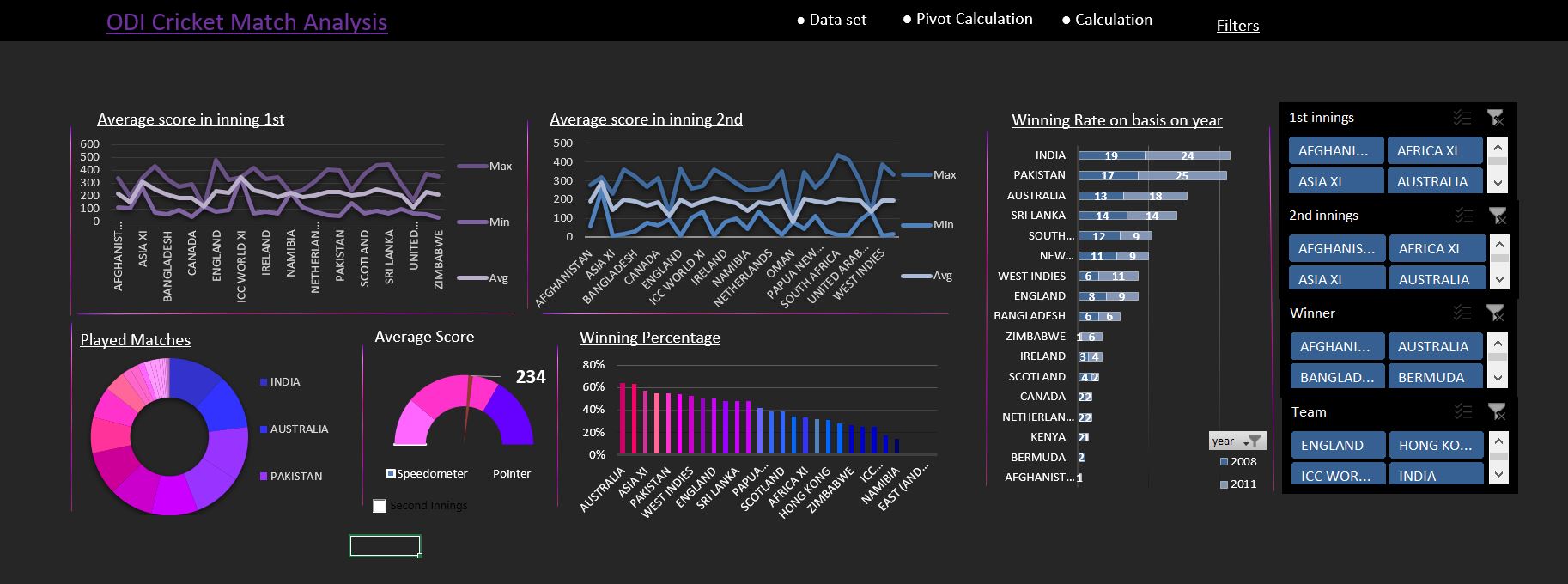


**Visualization:**

This is the front page of my project which gives the knownledge abiout the content contains in the sheet.

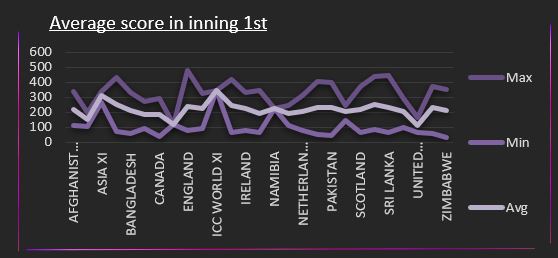


This dashboard contains of multiple graph which tells us about different scenarios going in the dataset.

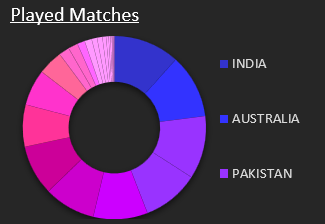


What different charts are showing in the above dash board.

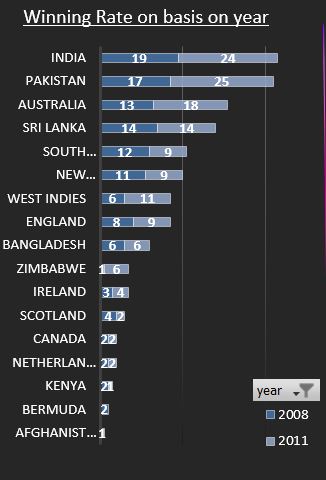
**Line Chart:** Data that's arranged in columns or rows on a worksheet can be plotted in a line chart. In a line chart, category data is distributed evenly along the horizontal axis, and all value data is distributed evenly along the vertical axis. In this dashboard line chart is showing the average 1st inning and 2nd inning runs.



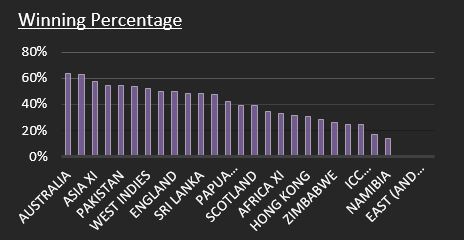
**Pie Chart:** Data that's arranged in one column or row on a worksheet can be plotted in a pie chart. Pie charts show the size of items in one data series, proportional to the sum of the items. The data points in a pie chart are shown as a percentage of the whole pie. In dashboard the matches played by different countries is shown by Pie Chart.



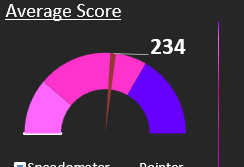
**Bar Chart:** Data that's arranged in columns or rows on a worksheet can be plotted in a bar chart. Bar charts illustrate comparisons among individual items. In a bar chart, the categories are typically organized along the vertical axis, and the values along the horizontal axis. Here which team has won how many matches in which year is shown by bar chart.



**Column Chart**: Data that’s arranged in columns or rows on a worksheet can be plotted in a column chart. A column chart typically displays categories along the horizontal (category) axis and values along the vertical (value) axis. In dashboard the winning percentage of each team has shown.



**Half pie Chart**: A half pie chart (also, half moon or semicircle chart) is a 180 degrees graph that represents the composition of a whole. Its total arc area corresponds to 100%, and its slices illustrate the part-to-whole relationships. The half pie chart works like a regular pie chart.



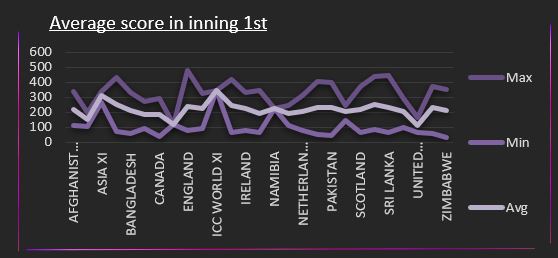
**Slicer**: Slicers provide buttons that you can click to filter tables, or PivotTables. In addition to quick filtering, slicers also indicate the current filtering state, which makes it easy to understand what exactly is currently displayed.



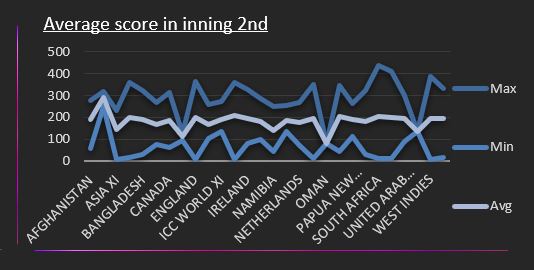
**Analysis with result:**

This is the final result of the Analysis which gives the knowledge about the specific things:

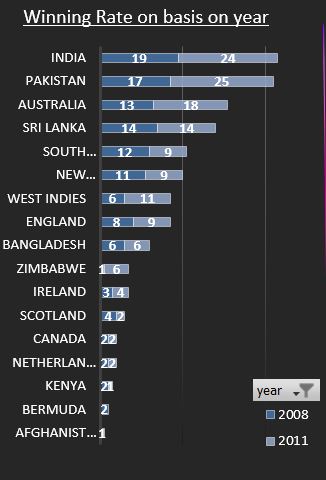
* The maximum, minimum and the average 1st inning score.



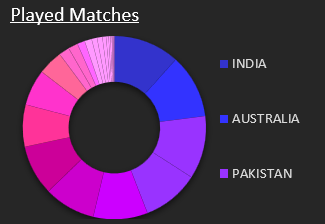
* The maximum, minimum and the average 2nd inning score.



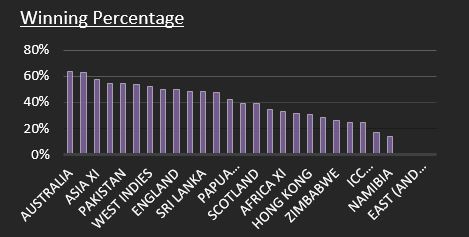
* Winning Rate of every team each year from 1971 to 2019.



* Total number of matches played by each team.



* Winning Percentage of each team.



**Reference:**

<https://www.kaggle.com/nowke9/odi-matches-first-innings-scores>.

<https://www.geeksforgeeks.org/find-excel-column-number-column-title/?ref=lbp>

<https://www.youtube.com/watch?v=RM8T1eYBjQY&t=106s>

Microsoft Excel Data Analysis and Business Modeling **Book by Wayne L. Winston.**

**Bibliography:**

1. C. J. Date, A. Kannan and S. Swamynathan, An Introduction to Database Systems, Pearson Education, Eighth Edition, 2009.
2. Abraham Silberschatz, Henry F. Korth and S. Sudarshan, Database System Concepts, McGraw-Hill Education (Asia), Fifth Edition, 2006.