**DATA MANAGEMENT PROJECT REPORT**

(Project Semester: August-December 2019)



***FIFA WORLD CUP ANALYSIS***

Submitted by

Joymalya Biswas

11704276

Programme and Section: B.Tech(CSE), KM060

Course Code: INT217

Under the Guidance of

**Savleen Kaur - 18306**

**Discipline of CSE/IT**

**Lovely School of Computer Science & Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that JOYMALYA BISWAS bearing Registration no. 11704276 has completed INT217 project titled, **“FIFA WORLD CUP 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.

**Savleen Kaur**

**School of Computer Science & Engineering**

**Lovely Professional University**

**Phagwara, Punjab.**

Date:

**DECLARATION**

I, Joymalya Biswas, student of B.Tech CSE 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.: 11704276 Joymalya Biswas

**ACKNOWLEDGEMENT**

I would like to express my special thanks of gratitude to my teacher Mrs. Savleen Kaur who gave me the golden opportunity to do this wonderful project of analysis of the data of a superstore namely “FIFA WORLD CUP ANALYSIS” which also helped me in doing a lot of research and I came to know about so many new things. I am thankful to them.Secondly, I would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame.

**Table of Content**

1. Introduction

2. Scope of the Analysis

3. Source of dataset

4. ETL process

5. Analysis on dataset (for each analysis)

1. Introduction
2. General Description
3. Specific Requirements, functions and formulas
4. Analysis results
5. Visualization

6. List of Analysis with results

7. References

8. Bibliography

**INTRODUCTION**

**Data Analysis** is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains.

The analytics team of the Football World Cup association and FIFA association anywhere in the world would love to check our through data analysis of each and every match leading to a well-organized and fruitful information. My analysis contains data on host teams, all stadiums, most supported team and teams which are most successful.

FIFA WORLD CUP ANALYSIS contains the following data fields: -

* Year – The year on which the following match is held.
* Date Time – Contains date and time on which the match is held.
* Stage – The level as Group A, B, C, D etc.
* Stadium – Stadium in which match is held.
* City – City in which match is held.
* Home Team Name – Host team name.
* Home team Goals – No of goals scored by the home team
* Away team goals – No of goals scored by the away team.
* Away team name – Away team name.
* Attendance – no. of people coming to enjoy the particular match.
* Half time home goals – goals scored by the home team by half time.
* Half time away goals – goals scored by the away team by half time.
* Referee – Name of referee
* Assistant 1 – Name of Assistant 1
* Assistant 2 – Name of assistant 2
* Home team initials – 3-word initials
* Away team initials - 3-word initials
* Winner – Winner team
* Total Goals – Total goals scored

**SCOPE OF ANALYSIS**

The s wants to see and analyze the sales trend month-wise and product-wise and work upon the lagging segments and outperforming employees accordingly. The Analytics team also wants to create analyze the database in depth to help the super store grow exponentially. The Analytics team wishes to answer the following objectives: -

1. Half time goals scored by the home and away team year wise.
2. No of goals scored in finals until now.
3. Analyzing the total attendance of crowd in Group matches, semifinals and Finals.
4. Overall Attendance for a particular city until now.
5. No. of people came to view match for particular team.
6. Country winning the maximum World Cups.
7. Goals scored by home teams and away team
8. Analyzing difference between total goals scored and total goals conceded by teams

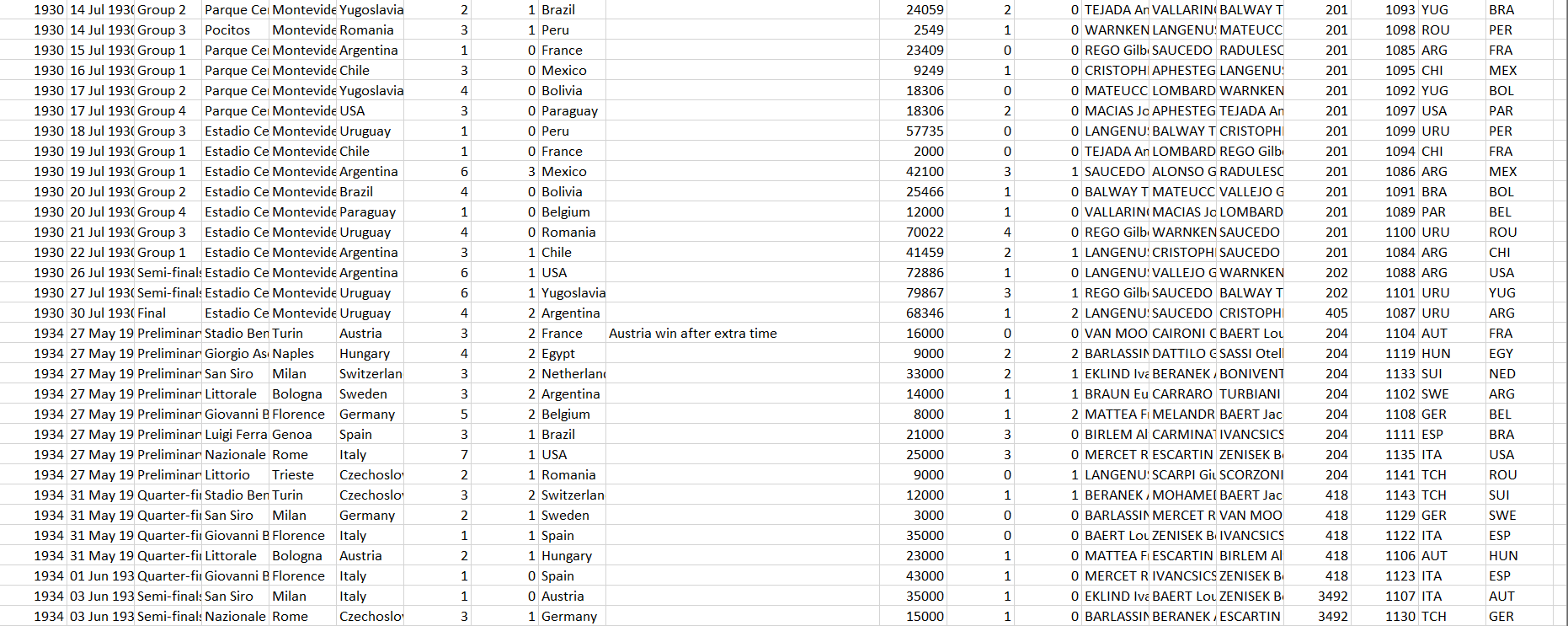
Aim of this project is to answer the above objectives in the form of visualization by creating a dashboard to convey the answers effectively and efficiently.

**ETL PROCESS**

In computing, extract, transform, load (ETL) is a process in database usage to prepare data for analysis, especially in data warehousing. Data extraction involves extracting data from homogeneous or heterogeneous sources, while data transformation processes data by transforming them into a proper storage format/structure for the purposes of querying and analysis; finally, data loading describes the insertion of data into the final target database such as an operational data store, a data mart, or a data warehouse. A properly designed ETL system extracts data from the source systems, enforces data quality and consistency standards, conforms data so that separate sources can be used together, and finally delivers data in a presentation-ready format so that application developers can build applications and end users can make decisions.

Precisely, ETL is defined as a process that extracts the data from different RDBMS source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. ETL stands for Extract, Transform and Load.

Before ETL, the dataset looked like this. **This data is taken from Kaggle.**



Through the process of ETL, we are going to clean the dataset and bring all the entities to their proper data format.

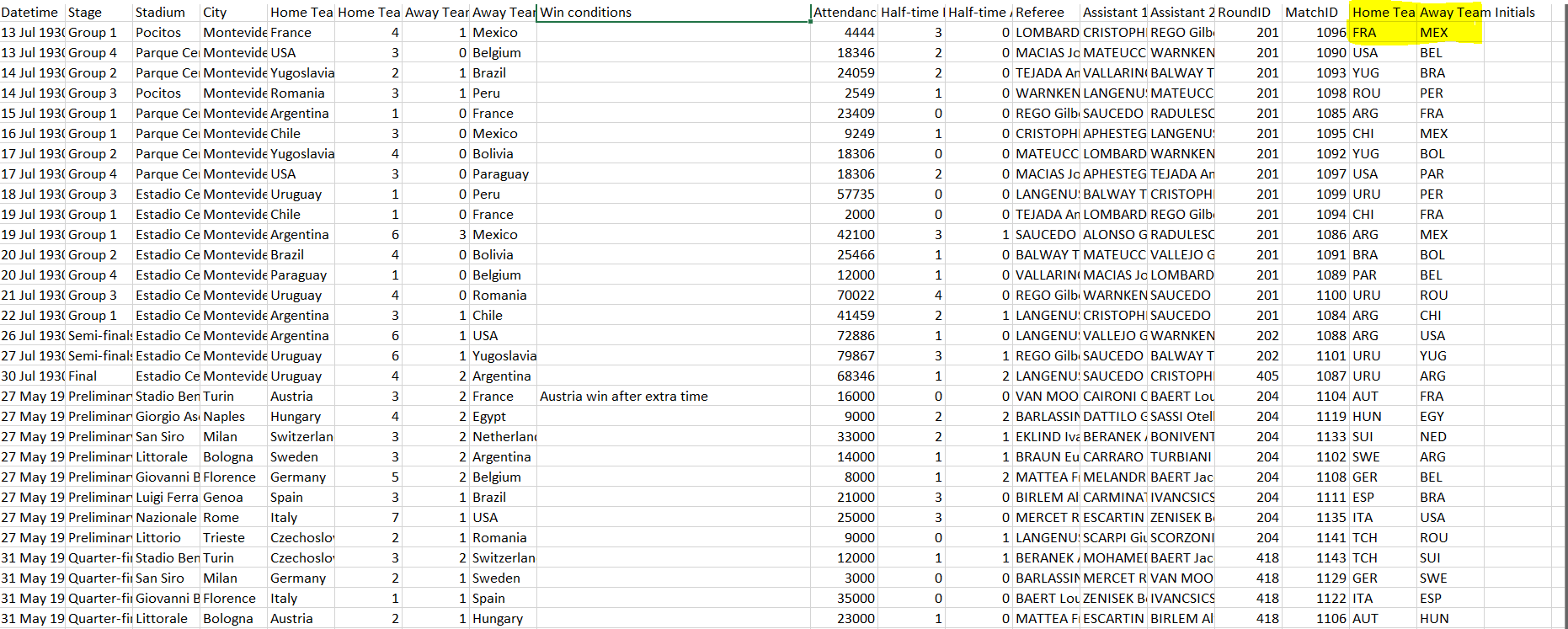
**Step 1: Removing the blank cells from the dataset.**

For this, select the whole dataset. Go to Find and Select in the Home tab of excel. Select Go to Special from the drop-down menu and then tick the blank option. All the blank cells will be selected. Then go to Delete option in the home tab again and select Delete Rows from the drop-down menu. This will remove any rows with blank cells.



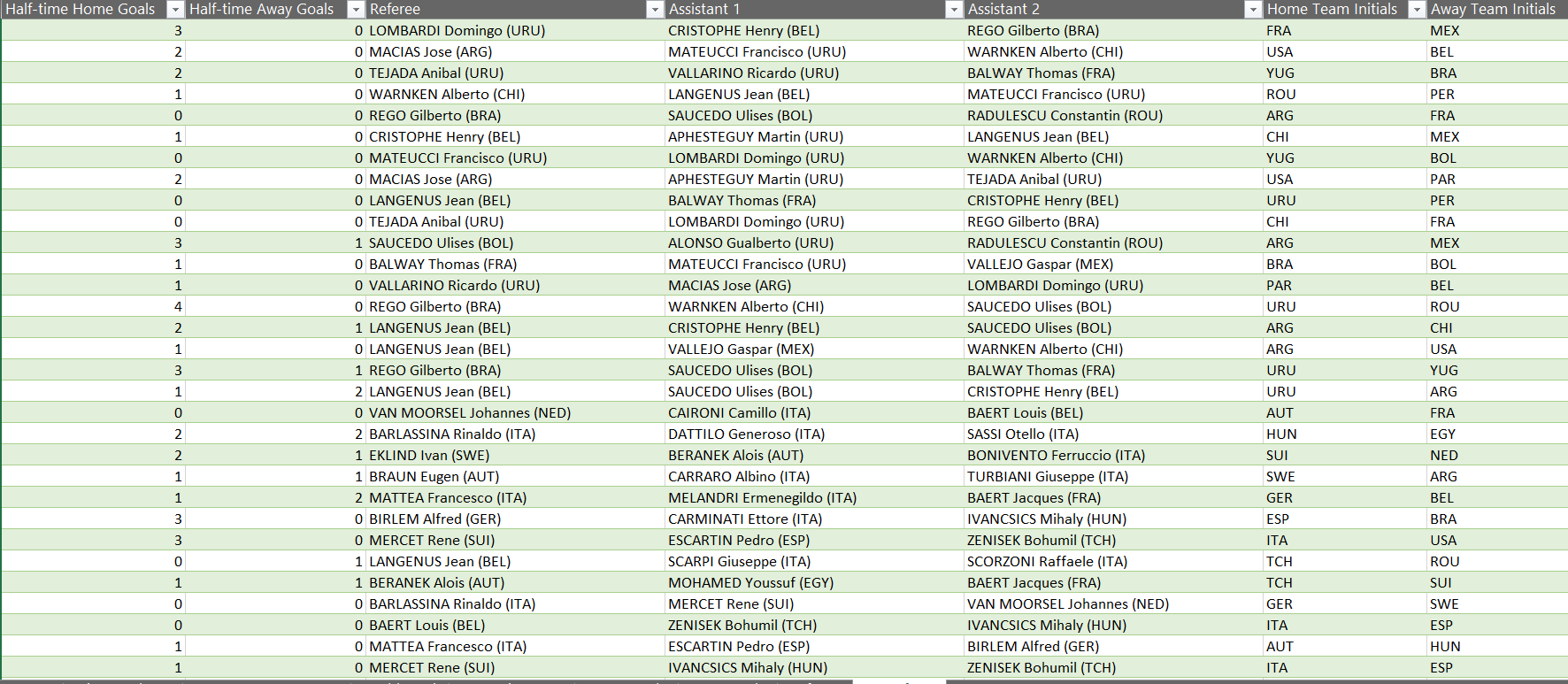
**Step 2: Removing columns which are not properly defined or not crucial to our analysis.**

For this we will columns which are redundant like the column with just the index numbers. For this we will select that particular column and then go to delete option in the home tag and then select Delete Columns from the drop-down menu.



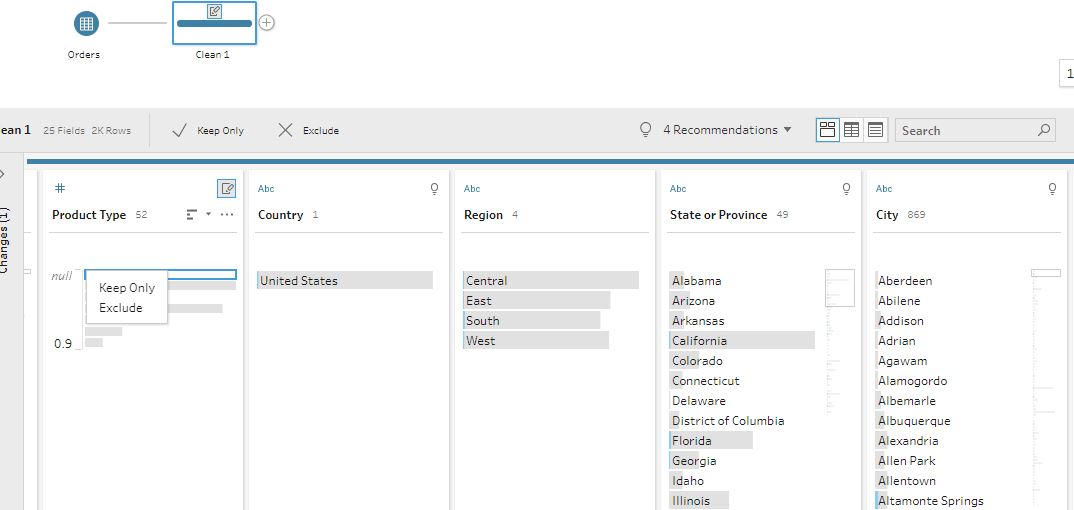
**Step 3: Giving proper and appropriate column names.**

The dataset does not have proper columns so our next step would be to giver proper column names to the columns wherever required.



**Step 4: Excluding the NULL values from the data.**

We’ll be using Tableau prep for this work as it’ll make the work simple and faster because we might not know how many null values could be there in this huge data set. Tableau helps us doing one step cleaning with ease.



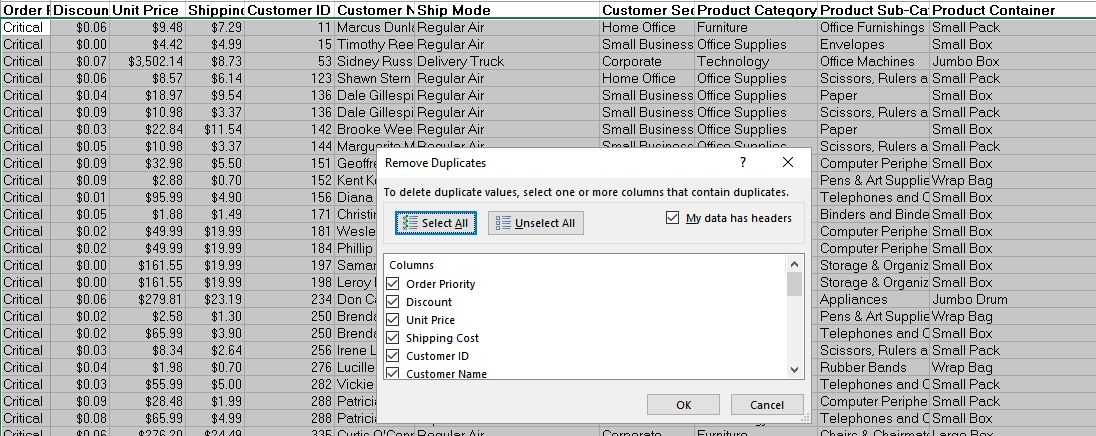
**Step 5: Improvising Proper Data Formatting**

Without proper Data Formatting, proper analysis will not take place. So, we will bring down certain columns to their proper format. For example, the dates should be in the date format and price and sales should be in currency format for better results.



**Step 6: Removing Duplicate Values**

It might be possible that our data may be containing duplicate values which may hinder in precise analysis. So, our last task in ETL will be removing duplicate values and making our data perfect for analysis.



**ANALYSIS OF DATASET**

1. **Half Time goals scored by the home and away teams:**

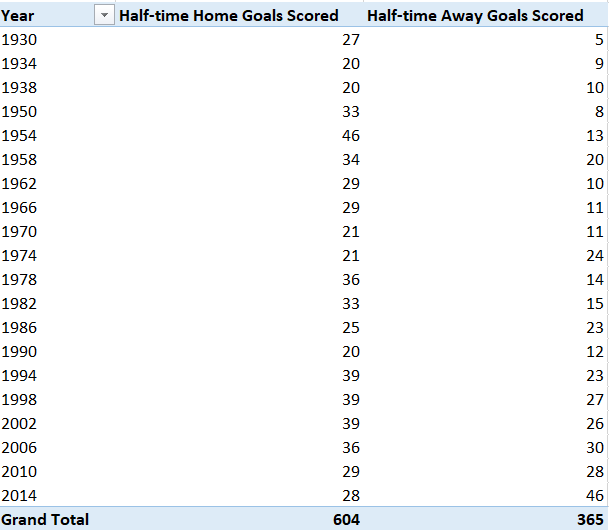
**Description:**

Total number of goals scored by home and away teams year by year from the year 1930 to 2014 to depict the mental confidence and usual trend of goals scored. From the trend given below we can see home team scores 604 goals in foist half and away team scores 365 goals. There is a massive goal difference leading to home team advantage.

**Specific function and requirements**

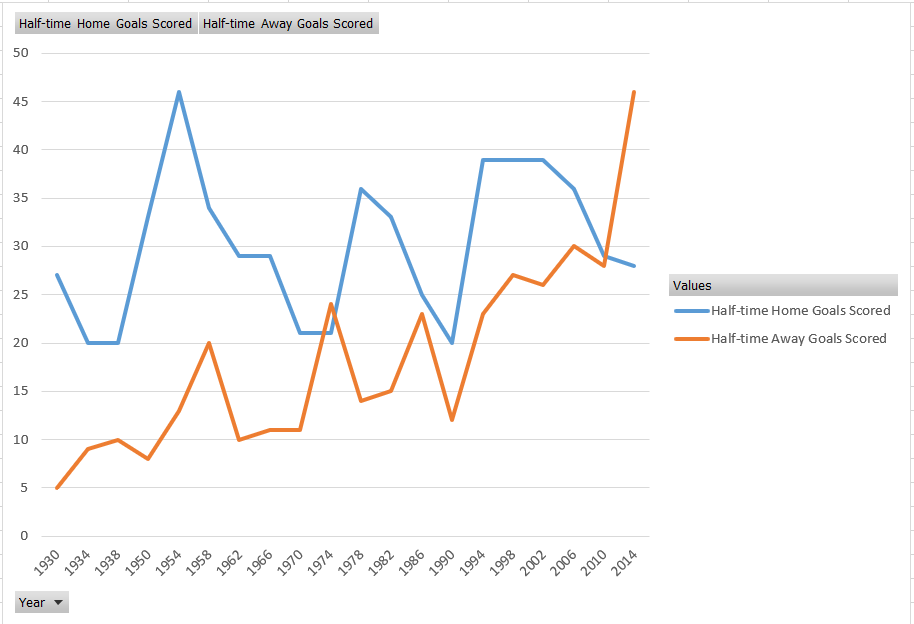
We have to create a pivot table to determine the difference in goals and then visualize it on graph.

**Results:**



**Visualization:**

The results are then visualized in the form of a stacked bar graph for both profit and sales



1. **No of goals scored in finals until now**

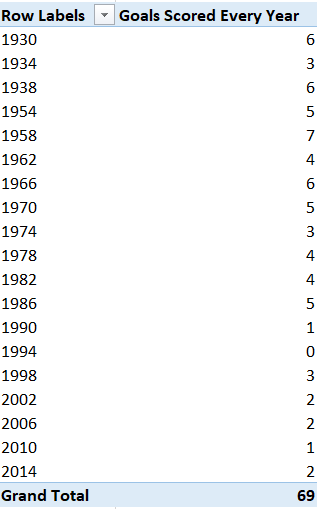
**Description:**

By calculating the current trend of the number of goals scored by home and away teams we can check who have strategical advantage and who is going under more pressure.

**Specific function and requirements**

As we can see total of 69 goals scored in finals out of which 1958 having the most of 7 goals and 1994 with least of 0 goals.

**Results:**



**Visualization:**

We will use a 3D clustered column to visualize the distribution.



1. **Analyzing the total attendance of crowd in group matches, Semi-Finals and Finals**

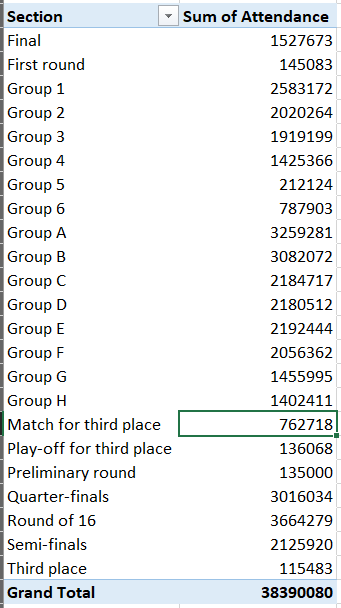
**Description:**

Describes the total attendance based on the group A,B,C,D and semifinals and finals and determining which stadium to be chosen, approx. amount of people expected security required etc.

**Specific function and requirements:**

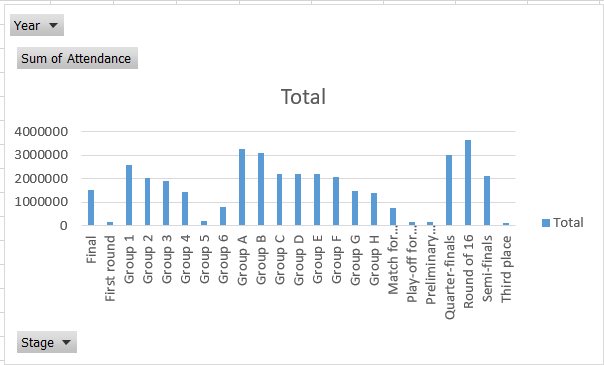
We have to create a pivot table. No specific functions are used. We then put the type of match and sum of sales in the columns.

**Results:**



**Visualization:**

The results are visualized with the help of line graph with a trend line displaying the trend of sales over months.



1. **Overall attendance for a particular city until now**

**Description:**

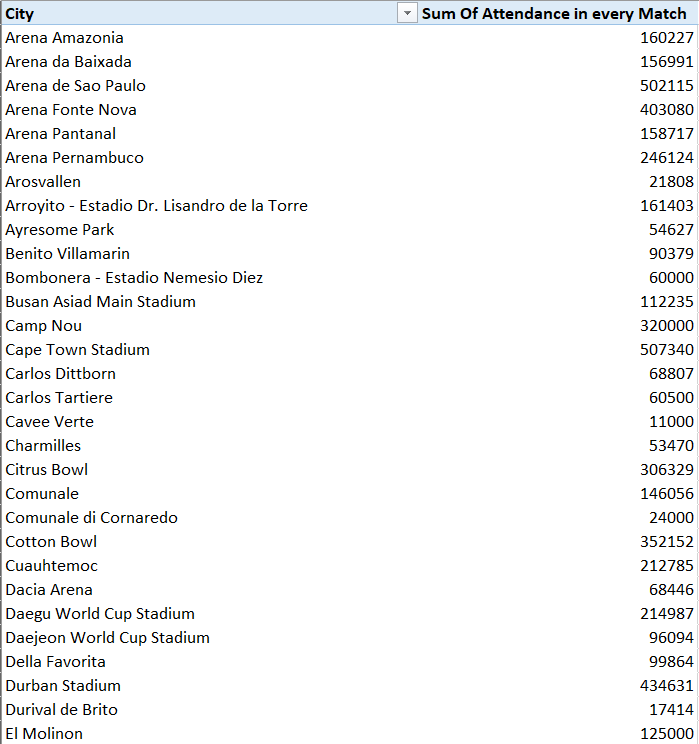
Determine the sum of attendance of all the people when they are home team.

**Specific function and requirements**

It helps us in determing

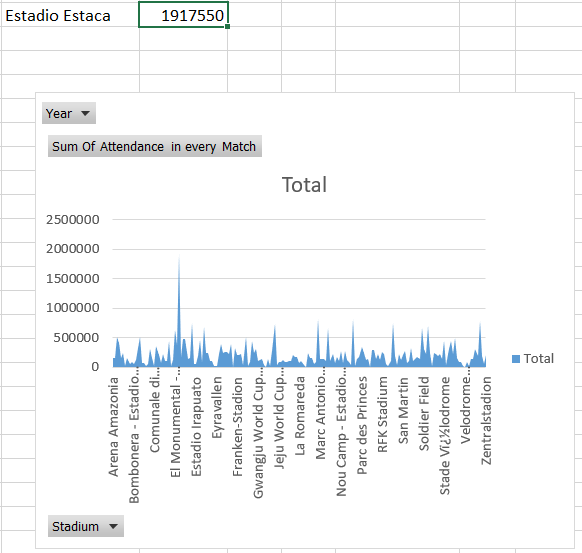
**Results:**

Here is a little chuck of data



**Visualization:**

We visualize the above results with the help of area chart created using area chart.



1. **No. of people came to view match for particular team.**

**Description:**

By comparing the attendance of all cities we get top 5 cities with highest audience.

**Specific function and requirements:**

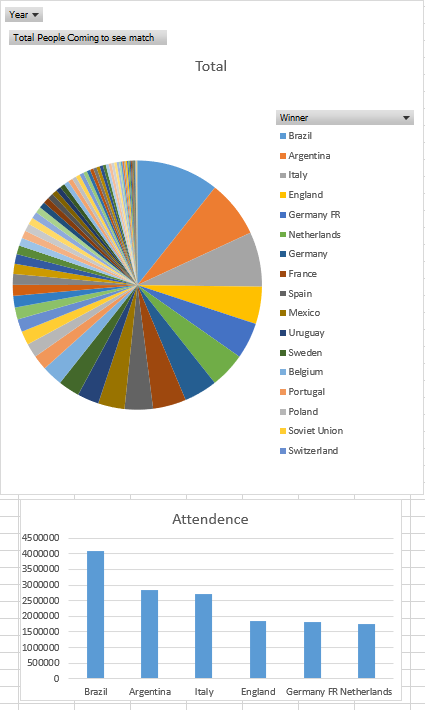
We get cities with highest fans.here we potray top 5 countries with hoighest attendence and snips of data.

**Results:**

****

**Visualization:**

The results are visualized in the form of pie and bar graphs.

****

1. **Country winning the maximum world cups**

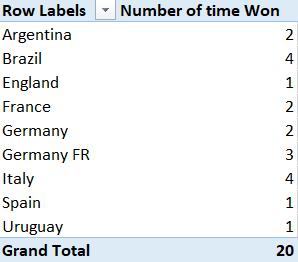
**Description:**

Determines the country with maximum FIFA world cup trophy.

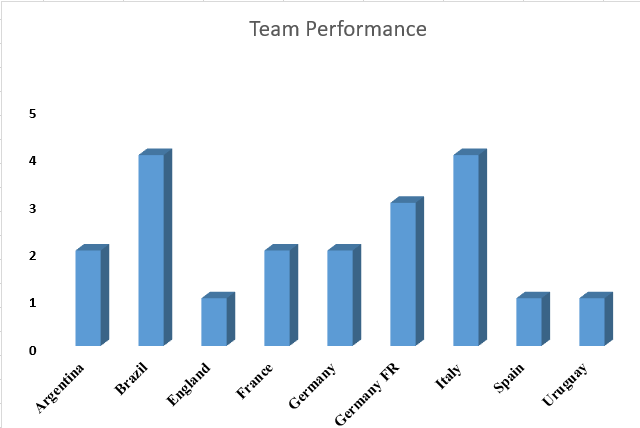
**Specific function and requirements:**

We have to create a pivot table. No specific functions are used.

**Results:**

****

**Visualization:**



1. **Goals Scored by home teams and away teams**

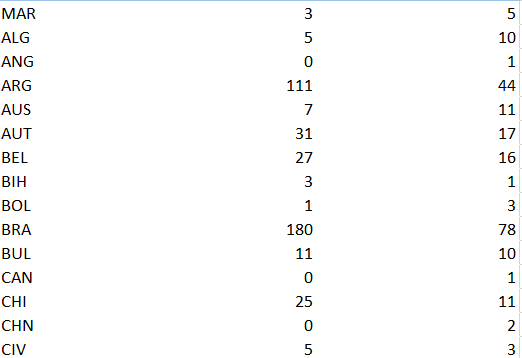
**Description:**

Describes the total goals scored by home and away teams and also top 3 teams scoing most goal difference.

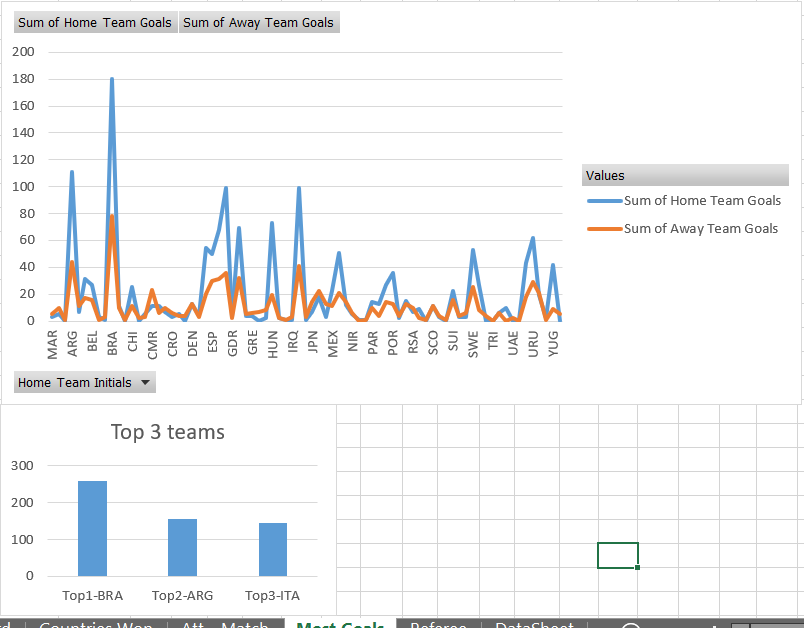
**Specific function and requirements:**

We have to create a pivot table. No specific functions are used. Only a few snips of the data shown

**Results:**



**Visualization:**

****

1. **Analyzing difference between total goals scored and total goals conceded**

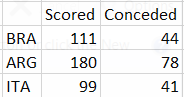
**Description:**

Difference between goals scored and conceded by teams.

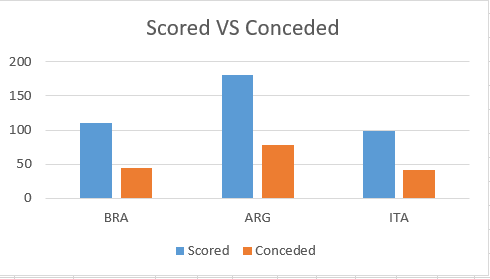
**Specific function and requirements:**

Using Excel data and formulas:

**Results:**

****

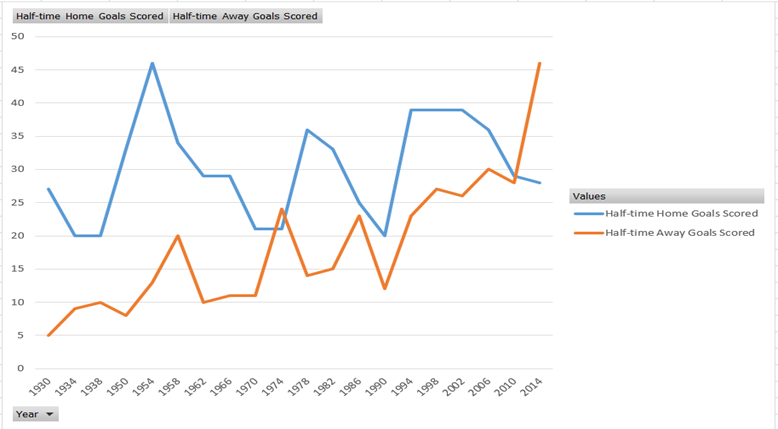
**Visualization:**

****

**ANALYSIS RESULTS**

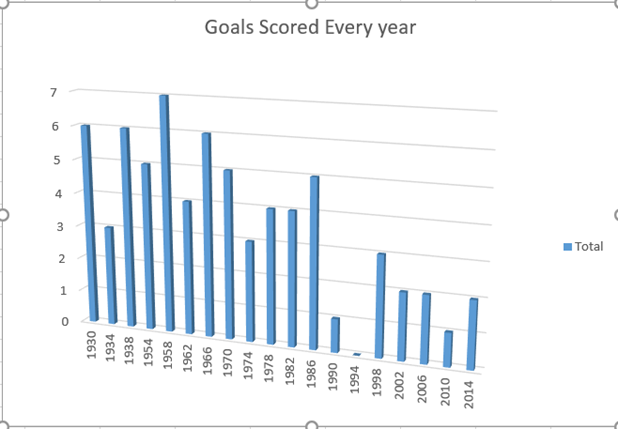
1. **Half Time goals scored by the home and away teams:**

The half time goals scored by the home teams were a lot higher then that of away team as they have a better understanding of ground.



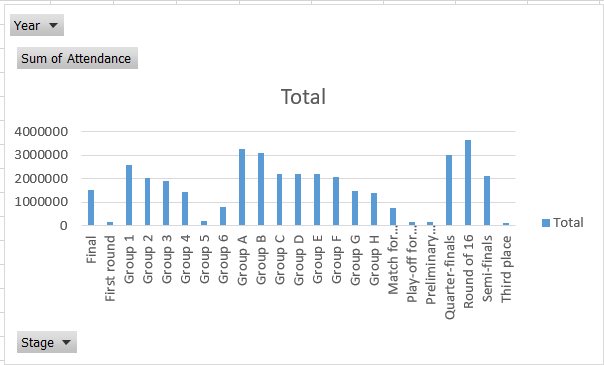
1. **No of goals scored in finals until now**

Until now 69 goals are scored in finals and most of the time host team have won just 1-time draw have happened in finals and once a 0 scored match.



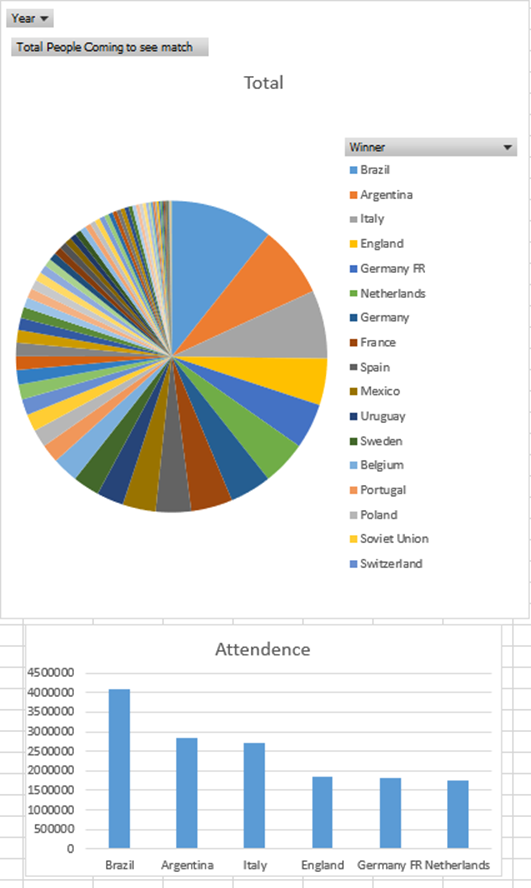
1. **Analyzing the total attendance of crowd in group matches, Semi-Finals and Finals**

Most Crowded stadiums are present in the semifinals then comes finals after that group stages match so book a stadium acc. to that.



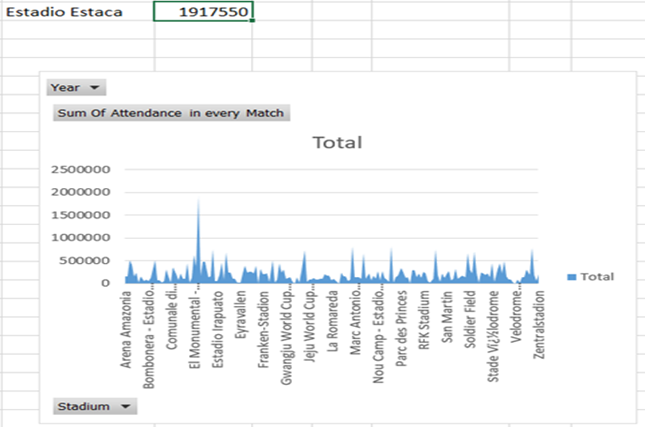
1. **Number of people came to view match for a particular team**

Most people came to see match for brazil then argentina,Italy England and so on..



1. **Overall attendance for a particular city until now**

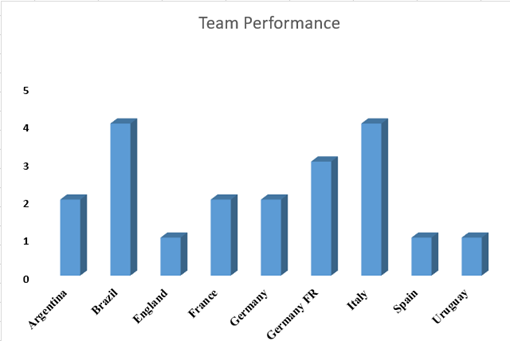
Most attendance is there for Estadio Estaca and so on



It is clear that tables are our best selling products followed by chairs and chair mats. We can work upon the one’s not performing well to increase their sales also.

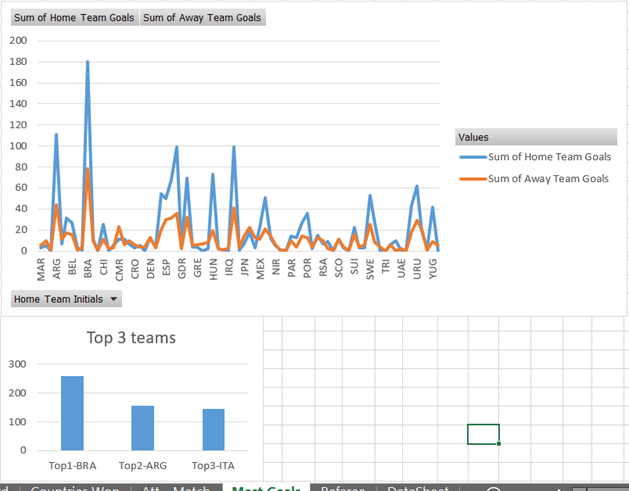
1. **Country winning the maximum world cups**

Maximum world cup is won by brazil and Italy together as 4 then comes Germany at 3 and then comes Argentina at 2.



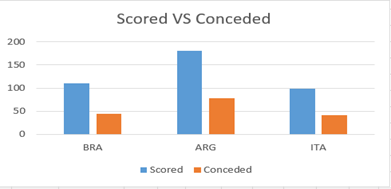
1. **Goals Scored by home teams and away teams**

Home team scores goals way larger then away team teamwise.



1. **Analyzing difference between total goals scored and total goals conceded**

Top 3 teams have large difference between scored and conceded that show their superiority over other teams.



**Final Dashboard**



**REFERENCES AND BIBLIOGRAPHY**

* Youtube
* Analytics Vidhiya
* Kaggle