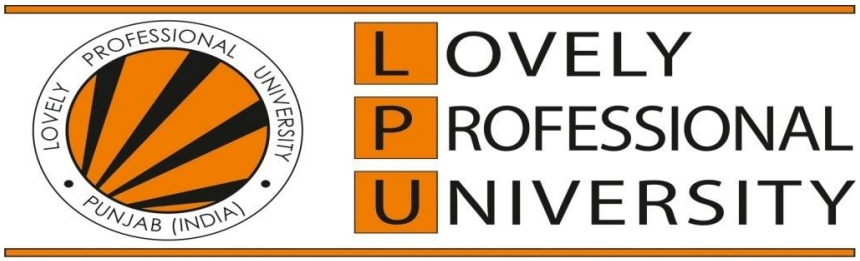
DATA MANAGEMENT PROJECT REPORT

(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

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.

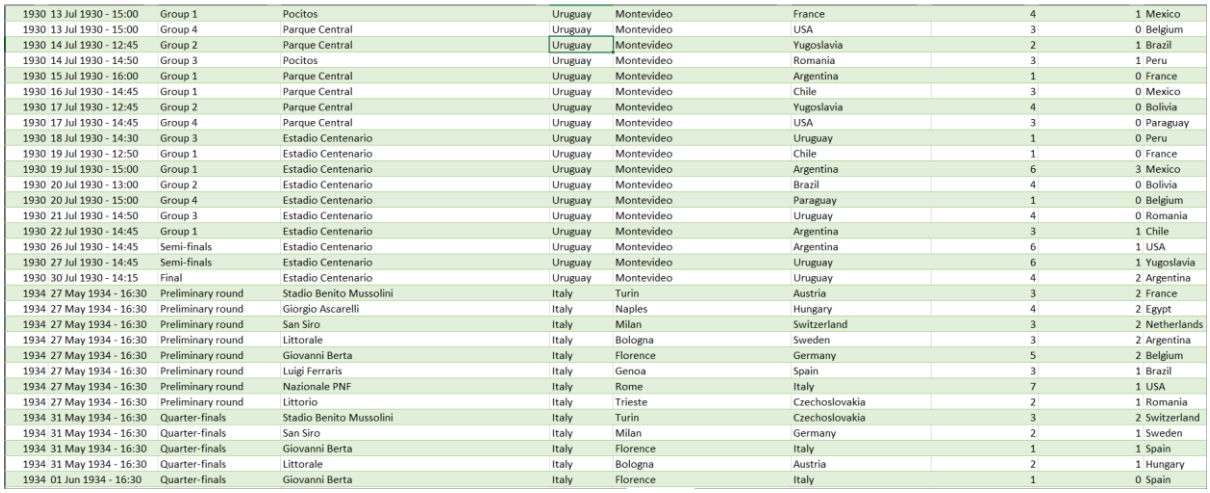
Before ETL, the dataset looked like this. **This data is taken from Kaggle.**

1930	14 Jul 193(	Group 2	Parque Ce(	Montevide	Yugoslavia	2	1	Brazil		24059	2	0	TEJADA An	VALLARIN(	BALWAY T	201	1093	YUG	BRA
1930	14 Jul 193(	Group 3	Pocitos	Montevide	Romania	3	1	Peru		2549	1	0	WARNKEN	LANGENU(	MATEUCC	201	1098	ROU	PER
1930	15 Jul 193(	Group 1	Parque Ce(	Montevide	Argentina	1	0	France		23409	0	0	REGO Gilb	SAUCEDO	RADULESC	201	1085	ARG	FRA
1930	16 Jul 193(	Group 1	Parque Ce(	Montevide	Chile	3	0	Mexico		9249	1	0	CRISTOPH(	APHESTEG	LANGENU(	201	1095	CHI	MEX
1930	17 Jul 193(	Group 2	Parque Ce(	Montevide	Yugoslavia	4	0	Bolivia		18306	0	0	MATEUCC	LOMBARD	WARNKEN	201	1092	YUG	BOL
1930	17 Jul 193(	Group 4	Parque Ce(	Montevide	USA	3	0	Paraguay		18306	2	0	MACIAS Jo	APHESTEG	TEJADA An	201	1097	USA	PAR
1930	18 Jul 193(	Group 3	Estadio Ce	Montevide	Uruguay	1	0	Peru		57735	0	0	LANGENU(	BALWAY T	CRISTOPH(	201	1099	URU	PER
1930	19 Jul 193(	Group 1	Estadio Ce	Montevide	Chile	1	0	France		2000	0	0	TEJADA An	LOMBARD	REGO Gilb	201	1094	CHI	FRA
1930	19 Jul 193(	Group 1	Estadio Ce	Montevide	Argentina	6	3	Mexico		42100	3	1	SAUCEDO	ALONSO G	RADULESC	201	1086	ARG	MEX
1930	20 Jul 193(	Group 2	Estadio Ce	Montevide	Brazil	4	0	Bolivia		25466	1	0	BALWAY T	MATEUCC	VALLEJO G	201	1091	BRA	BOL
1930	20 Jul 193(	Group 4	Estadio Ce	Montevide	Paraguay	1	0	Belgium		12000	1	0	VALLARIN(	MACIAS Jo	LOMBARD	201	1089	PAR	BEL
1930	21 Jul 193(	Group 3	Estadio Ce	Montevide	Uruguay	4	0	Romania		70022	4	0	REGO Gilb	WARNKEN	SAUCEDO	201	1100	URU	ROU
1930	22 Jul 193(	Group 1	Estadio Ce	Montevide	Argentina	3	1	Chile		41459	2	1	LANGENU(	CRISTOPH(	SAUCEDO	201	1084	ARG	CHI
1930	26 Jul 193(	Semi-final	Estadio Ce	Montevide	Argentina	6	1	USA		72886	1	0	LANGENU(	VALLEJO G	WARNKEN	202	1088	ARG	USA
1930	27 Jul 193(	Semi-final	Estadio Ce	Montevide	Uruguay	6	1	Yugoslavia		79867	3	1	REGO Gilb	SAUCEDO	BALWAY T	202	1101	URU	YUG
1930	30 Jul 193(	Final	Estadio Ce	Montevide	Uruguay	4	2	Argentina		68346	1	2	LANGENU(	SAUCEDO	CRISTOPH(	405	1087	URU	ARG
1934	27 May 19	Preliminar	Stadio Ben	Turin	Austria	3	2	France	Austria win after extra time	16000	0	0	VAN MOO	CAIRONI C	BAERT Lou	204	1104	AUT	FRA
1934	27 May 19	Preliminar	Giorgio Ast	Naples	Hungary	4	2	Egypt		9000	2	2	BARLASSIN	DATTILO C	SASSI Otel	204	1119	HUN	EGY
1934	27 May 19	Preliminar	San Siro	Milan	Switzerlan	3	2	Netherlan		33000	2	1	EKLIND Iva	BERANEK /	BONIVENT	204	1133	SUI	NED
1934	27 May 19	Preliminar	Littorale	Bologna	Sweden	3	2	Argentina		14000	1	1	BRAUN Eu	CARRARO	TURBIANI	204	1102	SWE	ARG
1934	27 May 19	Preliminar	Giovanni B	Florence	Germany	5	2	Belgium		8000	1	2	MATTEA Fi	MELANDR	BAERT Jac	204	1108	GER	BEL
1934	27 May 19	Preliminar	Luigi Ferra	Genoa	Spain	3	1	Brazil		21000	3	0	BIRLEM Al	CARMINA	IVANCSICS	204	1111	ESP	BRA
1934	27 May 19	Preliminar	Nazionale	Rome	Italy	7	1	USA		25000	3	0	MERCET R	ESCARTIN	ZENISEK Bo	204	1135	ITA	USA
1934	27 May 19	Preliminar	Littorio	Trieste	Czechoslo	2	1	Romania		9000	0	1	LANGENU(	SCARPI Gi	SCORZONI	204	1141	TCH	ROU
1934	31 May 19	Quarter-fi	Stadio Ben	Turin	Czechoslo	3	2	Switzerlan		12000	1	1	BERANEK /	MOHAMEI	BAERT Jac	418	1143	TCH	SUI
1934	31 May 19	Quarter-fi	San Siro	Milan	Germany	2	1	Sweden		3000	0	0	BARLASSIN	MERCET R	VAN MOO	418	1129	GER	SWE
1934	31 May 19	Quarter-fi	Giovanni B	Florence	Italy	1	1	Spain		35000	0	0	BAERT Lou	ZENISEK Bo	IVANCSICS	418	1122	ITA	ESP
1934	31 May 19	Quarter-fi	Littorale	Bologna	Austria	2	1	Hungary		23000	1	0	MATTEA Fi	ESCARTIN	BIRLEM Al	418	1106	AUT	HUN
1934	01 Jun 193	Quarter-fi	Giovanni B	Florence	Italy	1	0	Spain		43000	1	0	MERCET R	IVANCSICS	ZENISEK Bo	418	1123	ITA	ESP
1934	03 Jun 193	Semi-final	San Siro	Milan	Italy	1	0	Austria		35000	1	0	EKLIND Iva	BAERT Lou	ZENISEK Bo	3492	1107	ITA	AUT
1934	03 Jun 193	Semi-final	Nazionale	Rome	Czechoslo	3	1	Germany		15000	1	0	BARLASSIN	BERANEK /	ESCARTIN	3492	1130	TCH	GER

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.

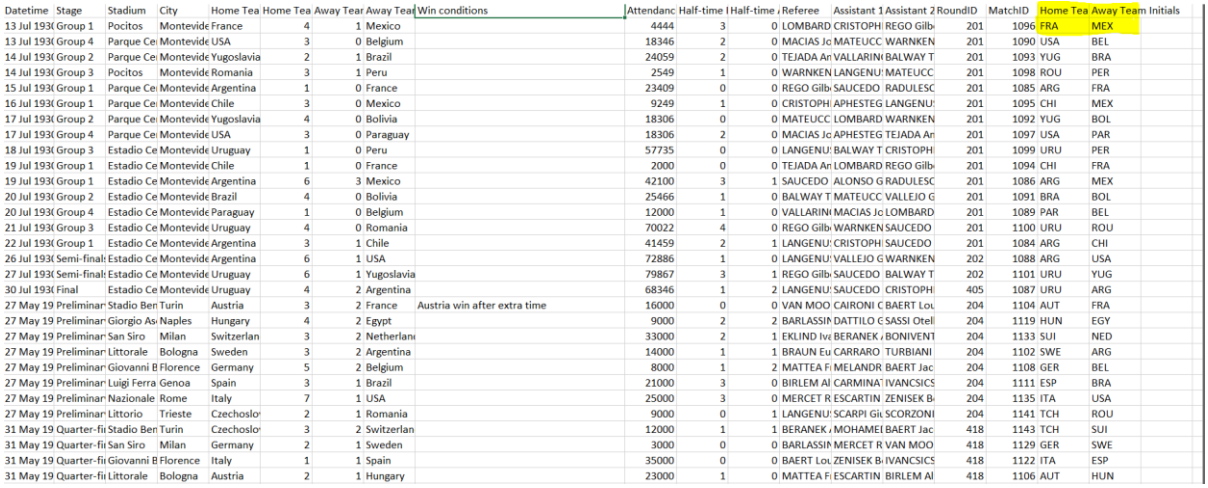
For this, select the whole dataset. Go to Find and Select in the Home tab of excel. Select Go to Special from the drop-down menu and then tick the blank option. All the blank cells will be selected. Then go to Delete option in the home tab again and select Delete Rows from the drop-down menu. This will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



1930	13 Jul 1930 - 15:00	Group 1	Pocitos	Uruguay	Montevideo	France	4	1	Mexico
1930	13 Jul 1930 - 15:00	Group 4	Parque Central	Uruguay	Montevideo	USA	3	0	Belgium
1930	14 Jul 1930 - 12:45	Group 2	Parque Central	Uruguay	Montevideo	Yugoslavia	2	1	Brazil
1930	14 Jul 1930 - 14:50	Group 3	Pocitos	Uruguay	Montevideo	Romania	3	1	Peru
1930	15 Jul 1930 - 16:00	Group 1	Parque Central	Uruguay	Montevideo	Argentina	1	0	France
1930	16 Jul 1930 - 14:45	Group 1	Parque Central	Uruguay	Montevideo	Chile	3	0	Mexico
1930	17 Jul 1930 - 12:45	Group 2	Parque Central	Uruguay	Montevideo	Yugoslavia	4	0	Bolivia
1930	17 Jul 1930 - 14:45	Group 4	Parque Central	Uruguay	Montevideo	USA	3	0	Paraguay
1930	18 Jul 1930 - 14:30	Group 3	Estadio Centenario	Uruguay	Montevideo	Uruguay	1	0	Peru
1930	19 Jul 1930 - 12:50	Group 1	Estadio Centenario	Uruguay	Montevideo	Chile	1	0	France
1930	19 Jul 1930 - 15:00	Group 1	Estadio Centenario	Uruguay	Montevideo	Argentina	6	3	Mexico
1930	20 Jul 1930 - 13:00	Group 2	Estadio Centenario	Uruguay	Montevideo	Brazil	4	0	Bolivia
1930	20 Jul 1930 - 15:00	Group 4	Estadio Centenario	Uruguay	Montevideo	Paraguay	1	0	Belgium
1930	21 Jul 1930 - 14:50	Group 3	Estadio Centenario	Uruguay	Montevideo	Uruguay	4	0	Romania
1930	22 Jul 1930 - 14:45	Group 1	Estadio Centenario	Uruguay	Montevideo	Argentina	3	1	Chile
1930	26 Jul 1930 - 14:45	Semi-finals	Estadio Centenario	Uruguay	Montevideo	Argentina	6	1	USA
1930	27 Jul 1930 - 14:45	Semi-finals	Estadio Centenario	Uruguay	Montevideo	Uruguay	6	1	Yugoslavia
1930	30 Jul 1930 - 14:15	Final	Estadio Centenario	Uruguay	Montevideo	Uruguay	4	2	Argentina
1934	27 May 1934 - 16:30	Preliminary round	Stadio Benito Mussolini	Italy	Turin	Austria	3	2	France
1934	27 May 1934 - 16:30	Preliminary round	Giorgio Ascarelli	Italy	Naples	Hungary	4	2	Egypt
1934	27 May 1934 - 16:30	Preliminary round	San Siro	Italy	Milan	Switzerland	3	2	Netherlands
1934	27 May 1934 - 16:30	Preliminary round	Littorale	Italy	Bologna	Sweden	3	2	Argentina
1934	27 May 1934 - 16:30	Preliminary round	Giovanni Berta	Italy	Florence	Germany	5	2	Belgium
1934	27 May 1934 - 16:30	Preliminary round	Luigi Ferraris	Italy	Genoa	Spain	3	1	Brazil
1934	27 May 1934 - 16:30	Preliminary round	Nazionale PNF	Italy	Rome	Italy	7	1	USA
1934	27 May 1934 - 16:30	Preliminary round	Littorio	Italy	Trieste	Czechoslovakia	2	1	Romania
1934	31 May 1934 - 16:30	Quarter-finals	Stadio Benito Mussolini	Italy	Turin	Czechoslovakia	3	2	Switzerland
1934	31 May 1934 - 16:30	Quarter-finals	San Siro	Italy	Milan	Germany	2	1	Sweden
1934	31 May 1934 - 16:30	Quarter-finals	Giovanni Berta	Italy	Florence	Italy	1	1	Spain
1934	31 May 1934 - 16:30	Quarter-finals	Littorale	Italy	Bologna	Austria	2	1	Hungary
1934	01 Jun 1934 - 16:30	Quarter-finals	Giovanni Berta	Italy	Florence	Italy	1	0	Spain

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.



Datetime	Stage	Stadium	City	Home Tea	Home Tea	Away Tear	Away Tea	Win conditions	Attendanc	Half-time I	Half-time /	Referee	Assistant 1	Assistant 2	RoundID	MatchID	Home Tea	Away Team Initials
13 Jul 193(	Group 1	Pocitos	Montevide	France	4	1	Mexico		4444	3	0	LOMBARD	CRISTOPH	REGO Gilb	201	1096	FRA	MEX
13 Jul 193(	Group 4	Parque Ce	Montevide	USA	3	0	Belgium		18346	2	0	MACIAS Jc	MATEUCC	WARNKEN	201	1090	USA	BEL
14 Jul 193(	Group 2	Parque Ce	Montevide	Yugoslavia	2	1	Brazil		24059	2	0	TEJADA An	VALLARIN	BALWAY T	201	1093	YUG	BRA
14 Jul 193(	Group 3	Pocitos	Montevide	Romania	3	1	Peru		2549	1	0	WARNKEN	LANGENU	MATEUCC	201	1098	ROU	PER
15 Jul 193(	Group 1	Parque Ce	Montevide	Argentina	1	0	France		23409	0	0	REGO Gilb	SAUCEDO	RADULESC	201	1085	ARG	FRA
16 Jul 193(	Group 1	Parque Ce	Montevide	Chile	3	0	Mexico		9249	1	0	CRISTOPH	APHESTEG	LANGENU	201	1095	CHI	MEX
17 Jul 193(	Group 2	Parque Ce	Montevide	Yugoslavia	4	0	Bolivia		18306	0	0	MATEUCC	LOMBARD	WARNKEN	201	1092	YUG	BOL
17 Jul 193(	Group 4	Parque Ce	Montevide	USA	3	0	Paraguay		18306	2	0	MACIAS Jc	APHESTEG	TEJADA An	201	1097	USA	PAR
18 Jul 193(	Group 3	Estadio Ce	Montevide	Uruguay	1	0	Peru		57735	0	0	LANGENU	BALWAY T	CRISTOPH	201	1099	URU	PER
19 Jul 193(	Group 1	Estadio Ce	Montevide	Chile	1	0	France		2000	0	0	TEJADA An	LOMBARD	REGO Gilb	201	1094	CHI	FRA
19 Jul 193(	Group 1	Estadio Ce	Montevide	Argentina	6	3	Mexico		42100	3	1	SAUCEDO	ALONSO G	RADULESC	201	1086	ARG	MEX
20 Jul 193(	Group 2	Estadio Ce	Montevide	Brazil	4	0	Bolivia		25466	1	0	BALWAY T	MATEUCC	VALLEJO G	201	1091	BRA	BOL
20 Jul 193(	Group 4	Estadio Ce	Montevide	Paraguay	1	0	Belgium		12000	1	0	VALLARIN	MACIAS Jc	LOMBARD	201	1089	PAR	BEL
21 Jul 193(	Group 3	Estadio Ce	Montevide	Uruguay	4	0	Romania		70022	4	0	REGO Gilb	WARNKEN	SAUCEDO	201	1100	URU	ROU
22 Jul 193(	Group 1	Estadio Ce	Montevide	Argentina	3	1	Chile		41459	2	1	LANGENU	CRISTOPH	SAUCEDO	201	1084	ARG	CHI
26 Jul 193(	Semi-final	Estadio Ce	Montevide	Argentina	6	1	USA		72886	1	0	LANGENU	VALLEJO G	WARNKEN	202	1088	ARG	USA
27 Jul 193(	Semi-final	Estadio Ce	Montevide	Uruguay	6	1	Yugoslavia		79867	3	1	REGO Gilb	SAUCEDO	BALWAY T	202	1101	URU	YUG
30 Jul 193(	Final	Estadio Ce	Montevide	Uruguay	4	2	Argentina		68346	1	2	LANGENU	SAUCEDO	CRISTOPH	405	1087	URU	ARG
27 May 19	Preliminar	Stadio Ben	Turin	Austria	3	2	France	Austria win after extra time	16000	0	0	VAN MOO	CAIRONI C	BAERT Lou	204	1104	AUT	FRA
27 May 19	Preliminar	Giorgio As	Naples	Hungary	4	2	Egypt		9000	2	2	BARLASSIN	DATTILO C	SASSI Otel	204	1119	HUN	EGY
27 May 19	Preliminar	San Siro	Milan	Switzerlan	3	2	Netherlan		33000	2	1	EKLIND Iva	BERANEK /	BONIVENT	204	1133	SUI	NED
27 May 19	Preliminar	Littorale	Bologna	Sweden	3	2	Argentina		14000	1	1	BRAUN Eu	CARRARO	TURBIANI	204	1102	SWE	ARG
27 May 19	Preliminar	Giovanni B	Florence	Germany	5	2	Belgium		8000	1	2	MATTEA Fi	MELANDR	BAERT Jac	204	1108	GER	BEL
27 May 19	Preliminar	Luigi Ferra	Genoa	Spain	3	1	Brazil		21000	3	0	BIRLEM Al	CARMINA	IVANCSICS	204	1111	ESP	BRA
27 May 19	Preliminar	Nazionale	Rome	Italy	7	1	USA		25000	3	0	MERCET R	ESCARTIN	ZENISEK B	204	1135	ITA	USA
27 May 19	Preliminar	Littorio	Trieste	Czechoslo	2	1	Romania		9000	0	1	LANGENU	SCARPI Gi	SCORZONI	204	1141	TCH	ROU
31 May 19	Quarter-fi	Stadio Ben	Turin	Czechoslo	3	2	Switzerlan		12000	1	1	BERANEK /	MOHAMEI	BAERT Jac	418	1143	TCH	SUI
31 May 19	Quarter-fi	San Siro	Milan	Germany	2	1	Sweden		3000	0	0	BARLASSIN	MERCET R	VAN MOO	418	1129	GER	SWE
31 May 19	Quarter-fi	Giovanni B	Florence	Italy	1	1	Spain		35000	0	0	BAERT Lou	ZENISEK B	IVANCSICS	418	1122	ITA	ESP
31 May 19	Quarter-fi	Littorale	Bologna	Austria	2	1	Hungary		23000	1	0	MATTEA Fi	ESCARTIN	BIRLEM Al	418	1106	AUT	HUN

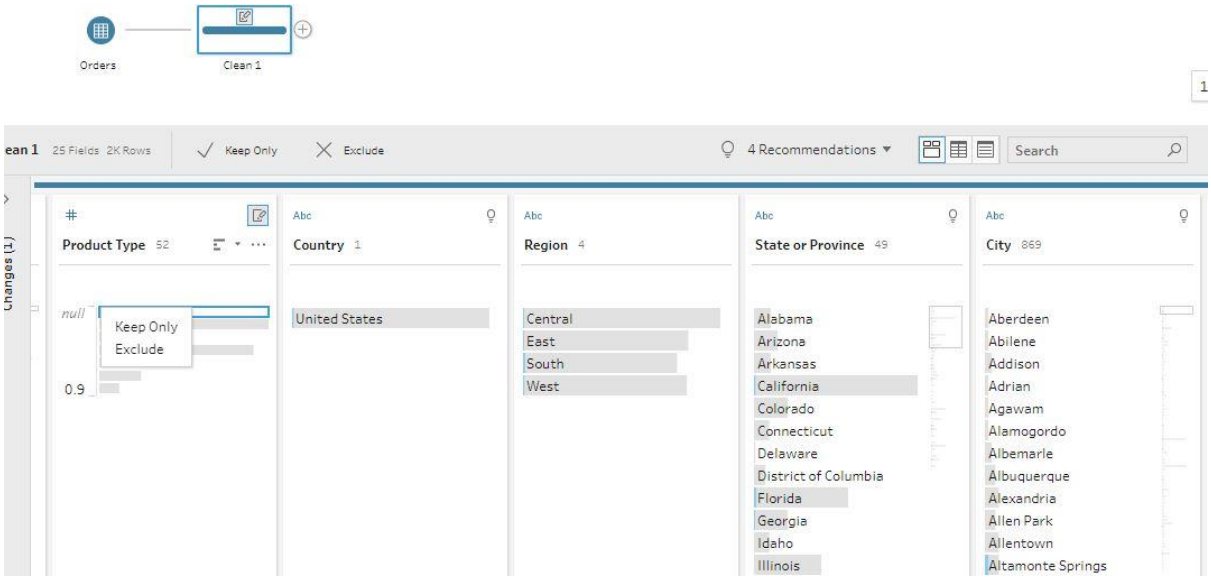
Step 3: Giving proper and appropriate column names.

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

Half-time Home Goals	Half-time Away Goals	Referee	Assistant 1	Assistant 2	Home Team Initials	Away Team Initials
3	0	LOMBARDI Domingo (URU)	CRISTOPHE Henry (BEL)	REGO Gilberto (BRA)	FRA	MEX
2	0	MACIAS Jose (ARG)	MATEUCCI Francisco (URU)	WARNKEN Alberto (CHI)	USA	BEL
2	0	TEJADA Anibal (URU)	VALLARINO Ricardo (URU)	BALWAY Thomas (FRA)	YUG	BRA
1	0	WARNKEN Alberto (CHI)	LANGENUS Jean (BEL)	MATEUCCI Francisco (URU)	ROU	PER
0	0	REGO Gilberto (BRA)	SAUCEDO Ulises (BOL)	RADULESCU Constantin (ROU)	ARG	FRA
1	0	CRISTOPHE Henry (BEL)	APHESTEGUY Martin (URU)	LANGENUS Jean (BEL)	CHI	MEX
0	0	MATEUCCI Francisco (URU)	LOMBARDI Domingo (URU)	WARNKEN Alberto (CHI)	YUG	BOL
2	0	MACIAS Jose (ARG)	APHESTEGUY Martin (URU)	TEJADA Anibal (URU)	USA	PAR
0	0	LANGENUS Jean (BEL)	BALWAY Thomas (FRA)	CRISTOPHE Henry (BEL)	URU	PER
0	0	TEJADA Anibal (URU)	LOMBARDI Domingo (URU)	REGO Gilberto (BRA)	CHI	FRA
3	1	SAUCEDO Ulises (BOL)	ALONSO Gualberto (URU)	RADULESCU Constantin (ROU)	ARG	MEX
1	0	BALWAY Thomas (FRA)	MATEUCCI Francisco (URU)	VALLEJO Gaspar (MEX)	BRA	BOL
1	0	VALLARINO Ricardo (URU)	MACIAS Jose (ARG)	LOMBARDI Domingo (URU)	PAR	BEL
4	0	REGO Gilberto (BRA)	WARNKEN Alberto (CHI)	SAUCEDO Ulises (BOL)	URU	ROU
2	1	LANGENUS Jean (BEL)	CRISTOPHE Henry (BEL)	SAUCEDO Ulises (BOL)	ARG	CHI
1	0	LANGENUS Jean (BEL)	VALLEJO Gaspar (MEX)	WARNKEN Alberto (CHI)	ARG	USA
3	1	REGO Gilberto (BRA)	SAUCEDO Ulises (BOL)	BALWAY Thomas (FRA)	URU	YUG
1	2	LANGENUS Jean (BEL)	SAUCEDO Ulises (BOL)	CRISTOPHE Henry (BEL)	URU	ARG
0	0	VAN MOORSEL Johannes (NED)	CAIRONI Camillo (ITA)	BAERT Louis (BEL)	AUT	FRA
2	2	BARLASSINA Rinaldo (ITA)	DATTILO Generoso (ITA)	SASSI Otello (ITA)	HUN	EGY
2	1	EKLIND Ivan (SWE)	BERANEK Alois (AUT)	BONIVENTO Ferruccio (ITA)	SUI	NED
1	1	BRAUN Eugen (AUT)	CARRARO Albino (ITA)	TURBIANI Giuseppe (ITA)	SWE	ARG
1	2	MATTEA Francesco (ITA)	MELANDRI Ermenegildo (ITA)	BAERT Jacques (FRA)	GER	BEL
3	0	BIRLEM Alfred (GER)	CARMINATI Ettore (ITA)	IVANCSICS Mihaly (HUN)	ESP	BRA
3	0	MERCET Rene (SUI)	ESCARTIN Pedro (ESP)	ZENISEK Bohumil (TCH)	ITA	USA
0	1	LANGENUS Jean (BEL)	SCARPI Giuseppe (ITA)	SCORZONI Raffaele (ITA)	TCH	ROU
1	1	BERANEK Alois (AUT)	MOHAMED Youssuf (EGY)	BAERT Jacques (FRA)	TCH	SUI
0	0	BARLASSINA Rinaldo (ITA)	MERCET Rene (SUI)	VAN MOORSEL Johannes (NED)	GER	SWE
0	0	BAERT Louis (BEL)	ZENISEK Bohumil (TCH)	IVANCSICS Mihaly (HUN)	ITA	ESP
1	0	MATTEA Francesco (ITA)	ESCARTIN Pedro (ESP)	BIRLEM Alfred (GER)	AUT	HUN
1	0	MERCET Rene (SUI)	IVANCSICS Mihaly (HUN)	ZENISEK Bohumil (TCH)	ITA	ESP

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.



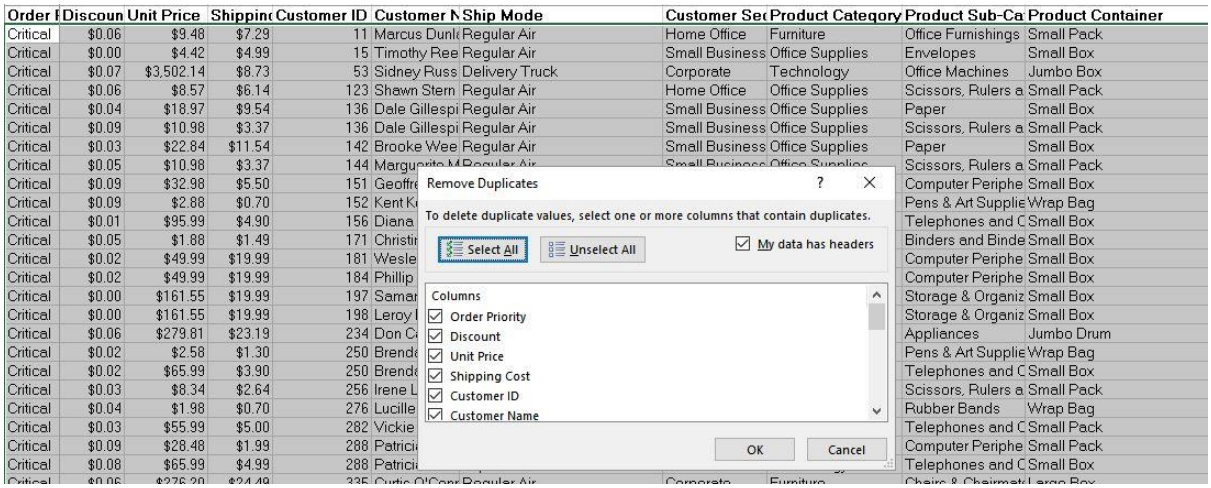
Step 5: Improvising Proper Data Formatting

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.

26 Jul 1930 - 14:45	Semi-finals	Estadio Centenario	Uruguay	Montevideo
27 Jul 1930 - 14:45	Semi-finals	Estadio Centenario	Uruguay	Montevideo
30 Jul 1930 - 14:15	Final	Estadio Centenario	Uruguay	Montevideo
27 May 1934 - 16:30	Preliminary round	Stadio Benito Mussolini	Italy	Turin
27 May 1934 - 16:30	Preliminary round	Giorgio Ascarelli	Italy	Naples
27 May 1934 - 16:30	Preliminary round	San Siro	Italy	Milan
27 May 1934 - 16:30	Preliminary round	Littorale	Italy	Bologna
27 May 1934 - 16:30	Preliminary round	Giovanni Berta	Italy	Florence
27 May 1934 - 16:30	Preliminary round	Luigi Ferraris	Italy	Genoa
27 May 1934 - 16:30	Preliminary round	Nazionale PNF	Italy	Rome
27 May 1934 - 16:30	Preliminary round	Littorio	Italy	Trieste
31 May 1934 - 16:30	Quarter-finals	Stadio Benito Mussolini	Italy	Turin
31 May 1934 - 16:30	Quarter-finals	San Siro	Italy	Milan
31 May 1934 - 16:30	Quarter-finals	Giovanni Berta	Italy	Florence
31 May 1934 - 16:30	Quarter-finals	Littorale	Italy	Bologna
01 Jun 1934 - 16:30	Quarter-finals	Giovanni Berta	Italy	Florence
03 Jun 1934 - 16:30	Semi-finals	San Siro	Italy	Milan
03 Jun 1934 - 16:30	Semi-finals	Nazionale PNF	Italy	Rome
07 Jun 1934 - 18:00	Match for third place	Giorgio Ascarelli	Italy	Naples
10 Jun 1934 - 17:30	Final	Nazionale PNF	Italy	Rome
04 Jun 1938 - 17:00	First round	Parc des Princes	France	Paris
05 Jun 1938 - 17:00	First round	Velodrome Municipale	France	Reims
05 Jun 1938 - 17:00	First round	Stade Olympique	France	Colombes
05 Jun 1938 - 17:00	First round	Stade Municipal	France	Toulouse
05 Jun 1938 - 17:00	First round	Stade Vï¿½lodrome	France	Marseilles
05 Jun 1938 - 17:30	First round	Stade de la Meinau	France	Strasbourg
05 Jun 1938 - 18:30	First round	Cavee Verte	France	Le Havre
09 Jun 1938 - 18:00	First round	Stade Municipal	France	Toulouse
09 Jun 1938 - 18:00	First round	Parc des Princes	France	Paris
12 Jun 1938 - 17:00	Quarter-finals	Stade du Parc Lescure	France	Bordeaux
12 Jun 1938 - 17:00	Quarter-finals	Victor Boucquey	France	Lille

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:

Year	Half-time Home Goals Scored	Half-time Away Goals Scored
1930	27	5
1934	20	9
1938	20	10
1950	33	8
1954	46	13
1958	34	20
1962	29	10
1966	29	11
1970	21	11
1974	21	24
1978	36	14
1982	33	15
1986	25	23
1990	20	12
1994	39	23
1998	39	27
2002	39	26
2006	36	30
2010	29	28
2014	28	46
Grand Total	**604**	**365**

Visualization:

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

2. **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:

Row Labels	Goals Scored Every Year
1930	6
1934	3
1938	6
1954	5
1958	7
1962	4
1966	6
1970	5
1974	3
1978	4
1982	4
1986	5
1990	1
1994	0
1998	3
2002	2
2006	2
2010	1
2014	2
Grand Total	**69**

Visualization:

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

3. **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:

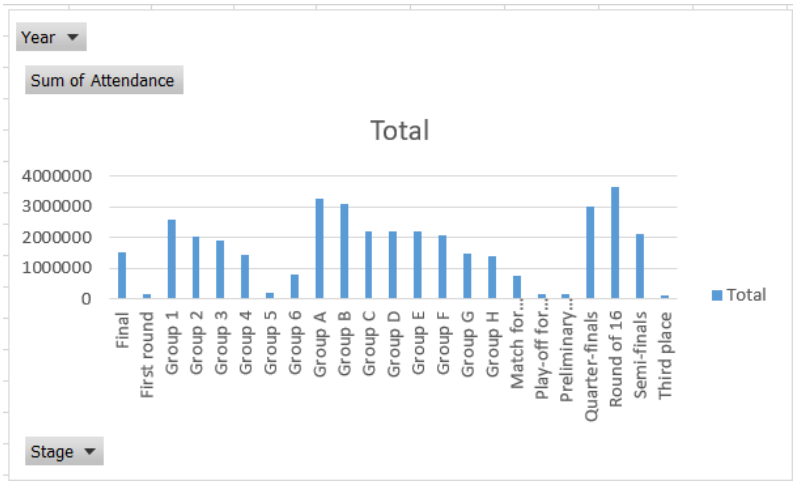
We have to create a pivot table. No specific functions are used. We then put the type *of match and sum of sales in the columns.*

Results:

Section	Sum of Attendance
Final	1527673
First round	145083
Group 1	2583172
Group 2	2020264
Group 3	1919199
Group 4	1425366
Group 5	212124
Group 6	787903
Group A	3259281
Group B	3082072
Group C	2184717
Group D	2180512
Group E	2192444
Group F	2056362
Group G	1455995
Group H	1402411
Match for third place	762718
Play-off for third place	136068
Preliminary round	135000
Quarter-finals	3016034
Round of 16	3664279
Semi-finals	2125920
Third place	115483
Grand Total	**38390080**

Visualization:

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



4. **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

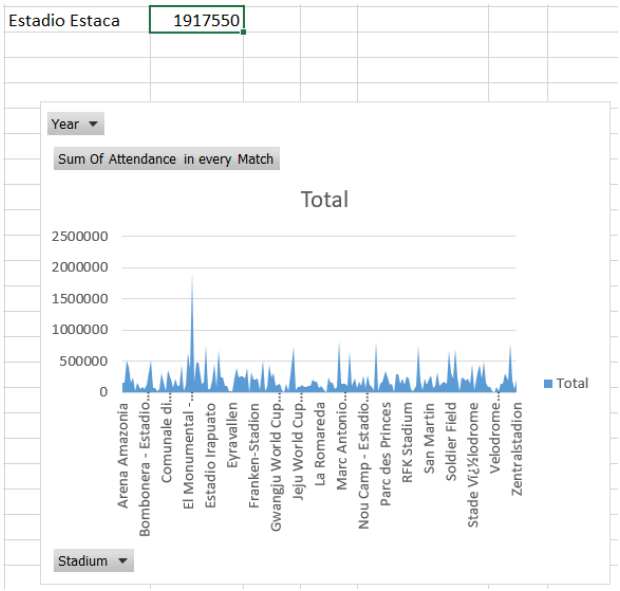
Results:

Here is a little chuck of data

City	Sum Of Attendance in every Match
Arena Amazonia	160227
Arena da Baixada	156991
Arena de Sao Paulo	502115
Arena Fonte Nova	403080
Arena Pantanal	158717
Arena Pernambuco	246124
Arosvallen	21808
Arroyito - Estadio Dr. Lisandro de la Torre	161403
Ayresome Park	54627
Benito Villamarin	90379
Bombonera - Estadio Nemesio Diez	60000
Busan Asiad Main Stadium	112235
Camp Nou	320000
Cape Town Stadium	507340
Carlos Dittborn	68807
Carlos Tartiere	60500
Cavee Verte	11000
Charmilles	53470
Citrus Bowl	306329
Comunale	146056
Comunale di Cornaredo	24000
Cotton Bowl	352152
Cuauhtemoc	212785
Dacia Arena	68446
Daegu World Cup Stadium	214987
Daejeon World Cup Stadium	96094
Della Favorita	99864
Durban Stadium	434631
Durival de Brito	17414
El Molinon	125000

Visualization:

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



5. **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:

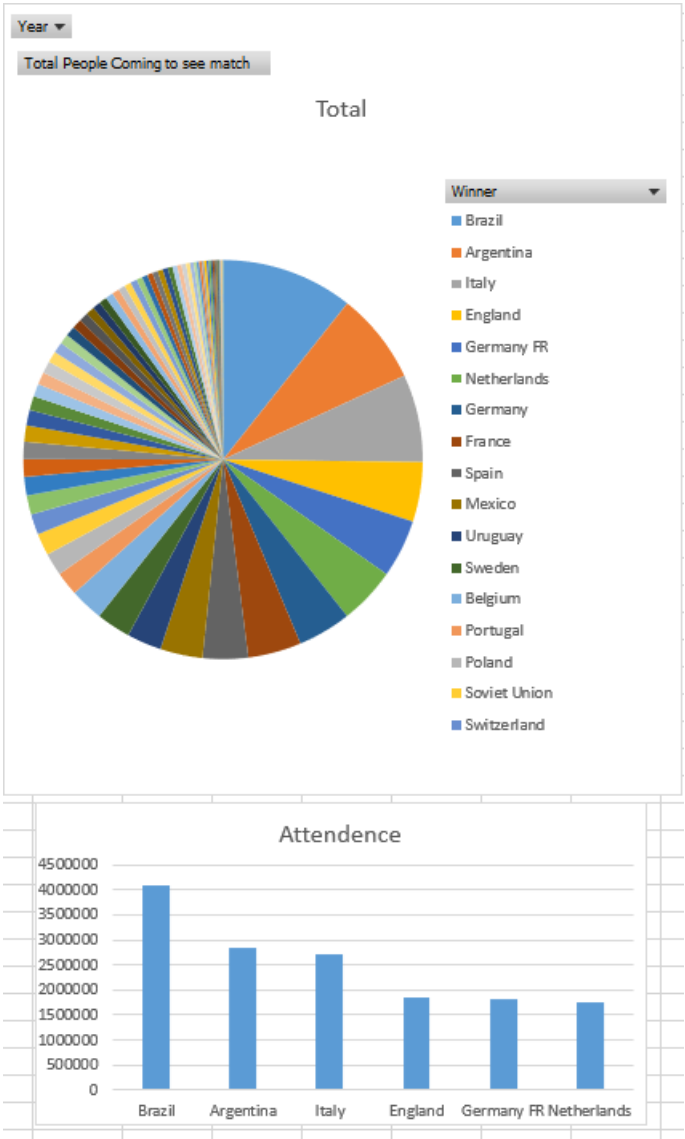
We get cities with highest fans.here we potray top 5 countries with hoighest attendence and snips of data.

Results:

Country	Total People Coming to see match
Brazil	4104110
Argentina	2844002
Italy	2727117
England	1855457
Germany FR	1804613
Netherlands	1768126
Germany	1659795
France	1655871
Spain	1409775
Mexico	1316594
Uruguay	1076418
Sweden	1065610
Belgium	1032703
Portugal	742088
Poland	703594
Soviet Union	683633
Switzerland	637616
Chile	595110
Yugoslavia	580546
Romania	542619
Colombia	538412
Korea Republic	503438
USA	480942
Austria	436912
Paraguay	399956
Hungary	384996
Bulgaria	374488
Denmark	345431
Cameroon	334000
Croatia	307454
Czechoslovakia	303523
Ghana	303447
Costa Rica	289629
Nigeria	257713
rn">Republic of Ireland	250676
Turkey	243286
Norway	241399
Saudi Arabia	227081
Greece	212657

Visualization:

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



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

Description:

Determines the country with maximum FIFA world cup trophy.

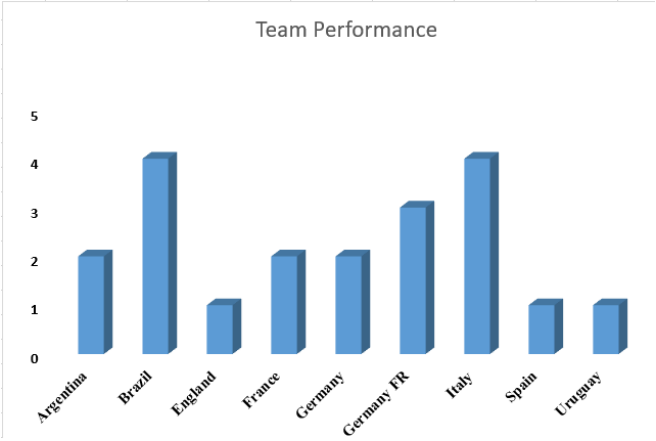
Specific function and requirements:

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

Results:

Row Labels	Number of time Won
Argentina	2
Brazil	4
England	1
France	2
Germany	2
Germany FR	3
Italy	4
Spain	1
Uruguay	1
Grand Total	**20**

Visualization:



7. **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.

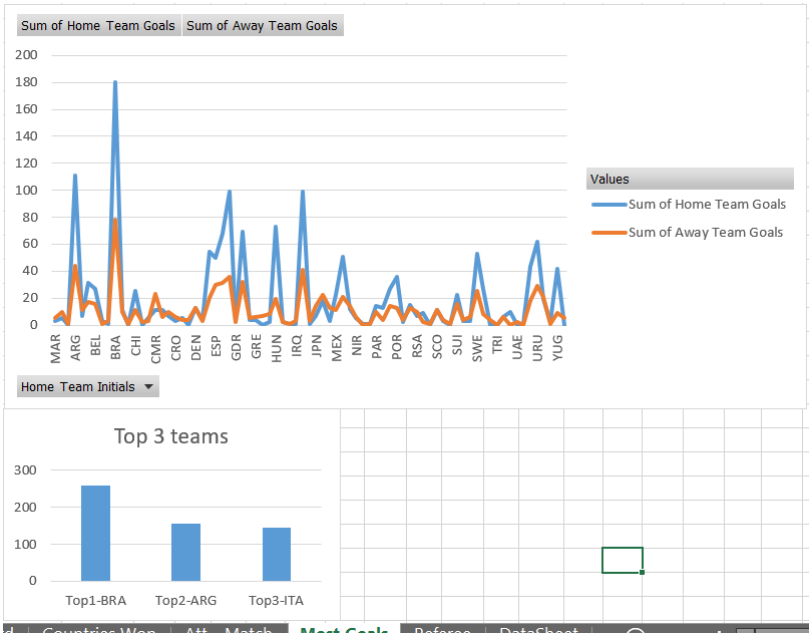
Specific function and requirements:

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

Results:

MAR	3	5
ALG	5	10
ANG	0	1
ARG	111	44
AUS	7	11
AUT	31	17
BEL	27	16
BIH	3	1
BOL	1	3
BRA	180	78
BUL	11	10
CAN	0	1
CHI	25	11
CHN	0	2
CIV	5	3

Visualization:



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

Description:

Difference between goals scored and conceded by teams.

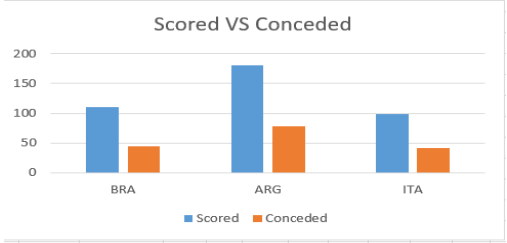
Specific function and requirements:

Using Excel data and formulas:

Results:

	Scored	Conceded
BRA	111	44
ARG	180	78
ITA	99	41

Visualization:



ANALYSIS RESULTS

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

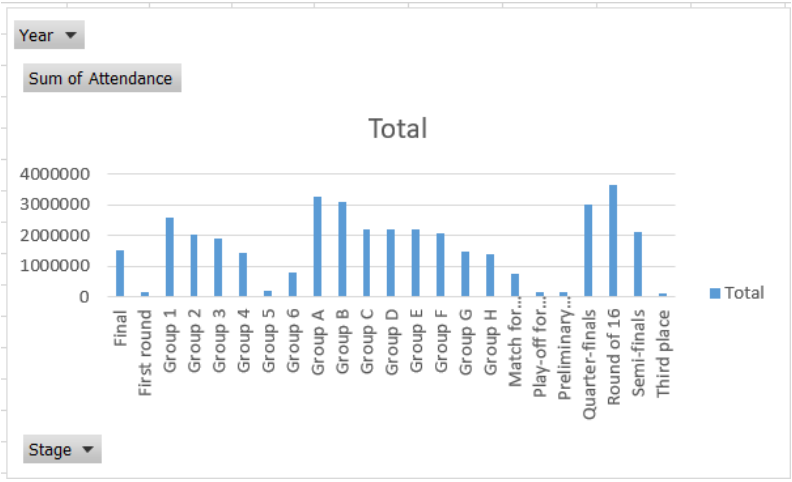
   The half time goals scored by the home teams were a lot higher then that of away team as they have a better understanding of ground.

2. **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.

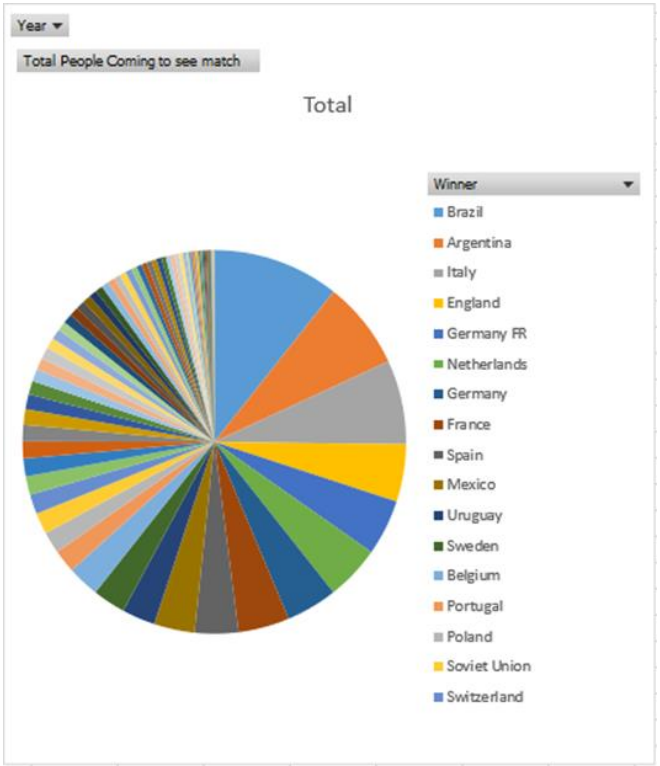
3. **Analyzing the total attendance of crowd in group matches, Semi-Finals and Finals**

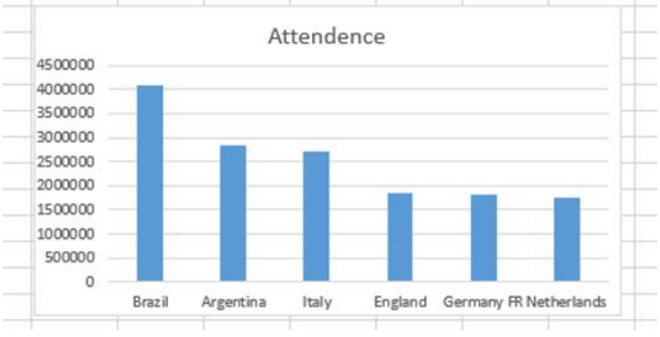
   Most Crowded stadiums are present in the semifinals then comes finals after that group stages match so book a stadium acc. to that.



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

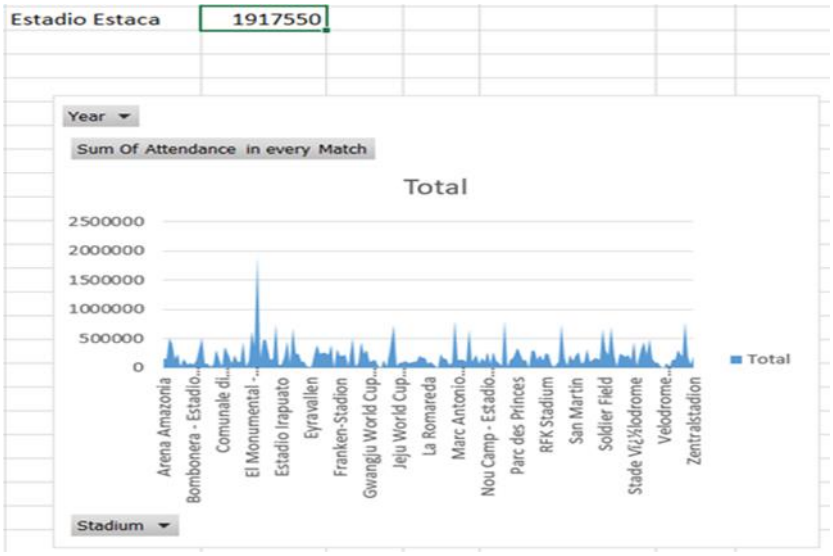
Most people came to see match for brazil then argentina,Italy England and so on..





5. Overall attendance for a particular city until now

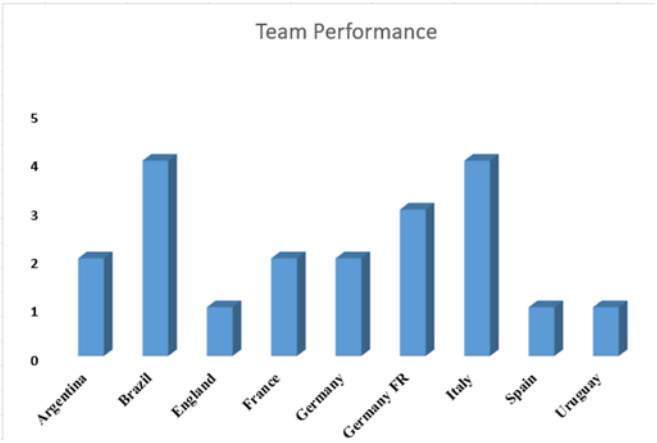
Most attendance is there for Estadio Estaca and so on



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.

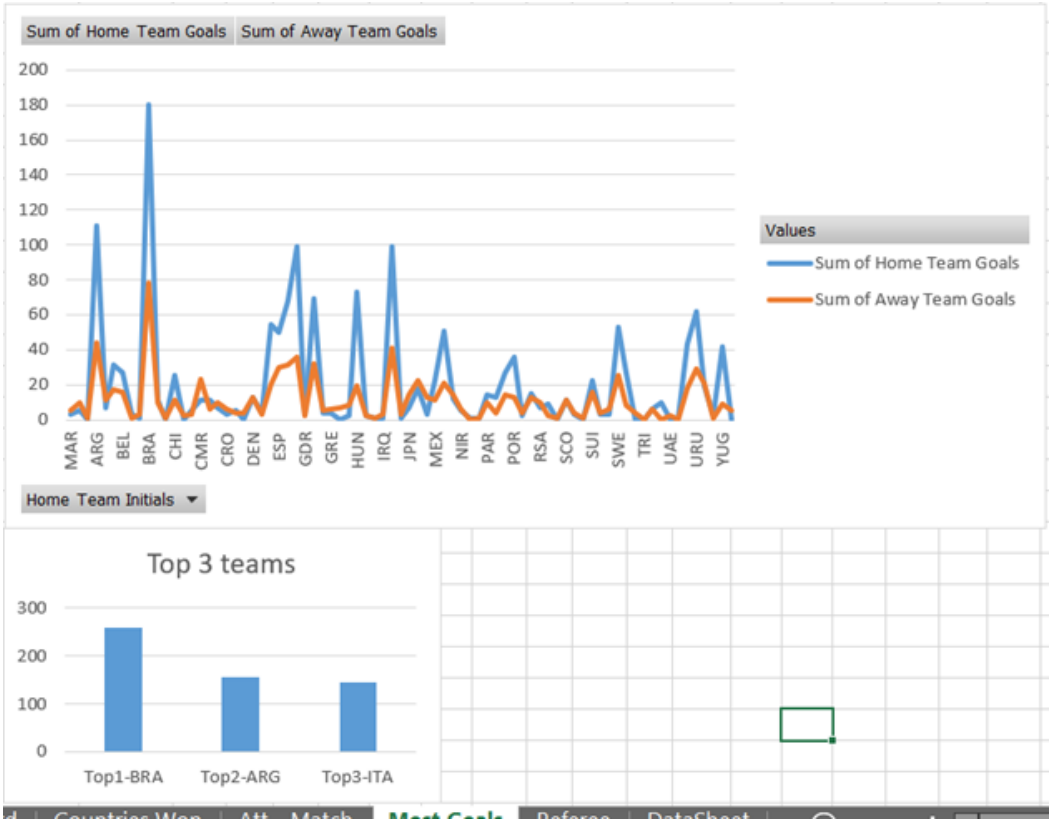
6. Country winning the maximum world cups

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



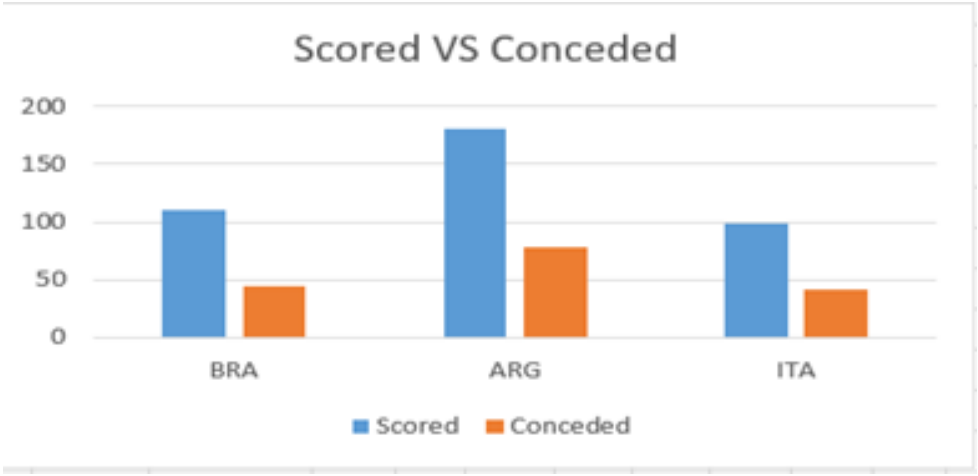
7. Goals Scored by home teams and away teams

Home team scores goals way larger then away team teamwise.

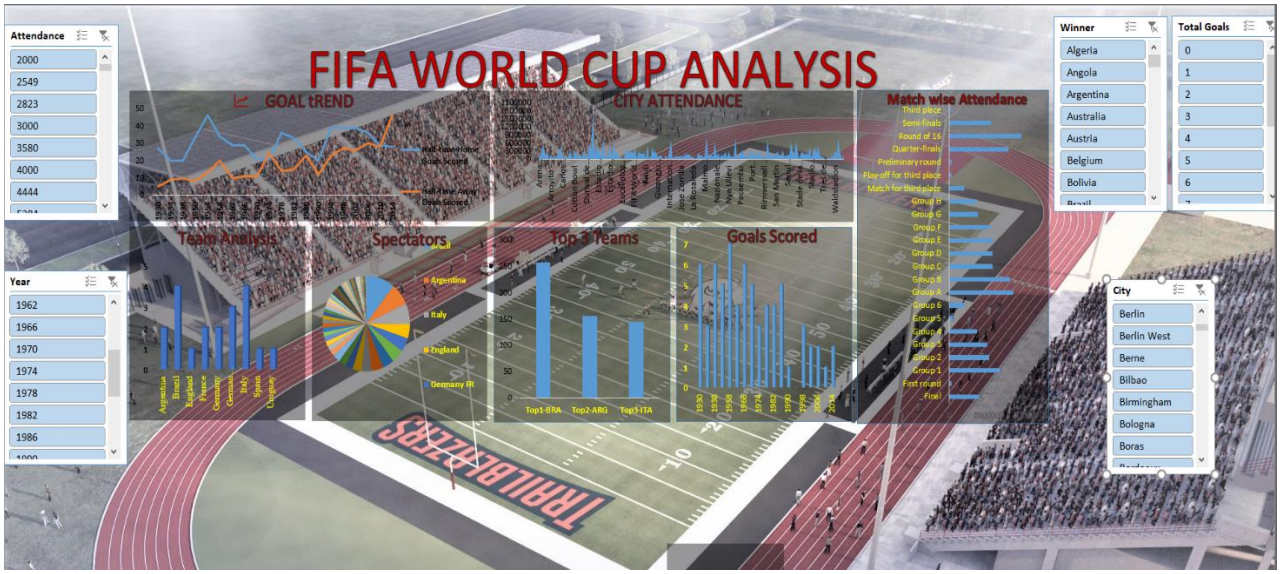


8. Analyzing difference between total goals scored and total goals conceded

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