KARNATAK LAW SOCIETY'S

GOGTE INSTITUTE OF TECHNOLOGY

UDYAMBAG, BELAGAVI-590008

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

A COURSE ACTIVITY REPORT ON

"Data Visualization of the FIFA World Cup"

Submitted for the requirements of 6th semester B.E. in ISE for

"DATA MINING"

Submitted by

Name	USN
V Gopinath	2GI19IS056

Under the guidance of

Prof.Pandurang Upparamani

Prof., Dept. of ISE

Academic Year 2022 (Even semester)

Certificate

Karnatak Law Society's
GOGTE INSTITUTE OF TECHNOLOGY
Udyambag Belagavi -590008
Karnataka, India.

Department of Information Science and Engineering

This is to certify that the Project work titled **"Data Visualization of the FIFA World Cup"** carried out by **V Gopinath,** bearing **USN:2GI19IS056 submitted** in partial fulfilment of the requirements for 6th semester B.E. in INFORMATION SCIENCE AND ENGINEERING, Visvesvaraya Technological University, Belagavi. It is certified that all corrections/suggestions indicated have been incorporated in the report. The project report has been approved as it satisfies the academic requirements in respect of research work prescribed for the said degree.

Date:03/09/2022 **Signature of Guide**
Place: Belagavi

Prof. Pandurang Upparamani
Prof in Dept. of ISE
KLS Gogte Institute of Technology,
Belagavi

Team Members Details:

S. No.	USN	Student Name
1	**2GI19IS056**	**V GOPINATH**

Marks Allocation:

	Batch No. : 03		
1.	Project Title:Data Visualization of the FIFA Worldcup	Marks Range	USN 2GI19IS056
2.	Problem statement (PO2)	0-1	
3.	Objectives of Defined Problem statement(PO1,PO2)	0-2	
4.	Design / Algorithm/Flowchart/Methodology(PO3)	0-3	
5.	Implementation details/Function/Procedures/Classes and Objects (Language/Tools)(PO1,PO3,PO4,PO5)	0-4	
6.	Working model of the final solution (PO3,PO12)	0-5	
7.	Report and Oral presentation skill (PO9,PO10)	0-5	
	Total	20	

TABLE OF CONTENTS

INTRODUCTION

Data visualization is the graphical representation of information and data.By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.Data visualization is one of the steps of the data sceince process, which states that after data has been collected, processed and modeled, it must be visualized for conclusions to be made.

Data visualization provides a quick and effective way to communicate information in a universal manner using visual information. There are many types of data visualization. The most common are scatter plots, line graphs, pie charts, bar charts, heat maps, area charts, choropleth maps and histograms. Visualization is done by using several libraries in python like – matplotlib, seaborn etc.

Good data visualizations allow us to **reason and think effectively** about our data. By presenting information visually, it allows us offload internal cognition to the perceptual system. If we see numerical data in a table, we may be able to find a trend, but it will take a significant amount of work on our part to recognize and conceptualize that trend. By plotting that data visually, that trend becomes immediately clear to our mind through our perceptual system.

It's hard to think of a professional industry that doesn't benefit from making data more understandable. Every STEM field benefits from understanding data—and so do fields in government, finance, marketing, history, consumer goods, service industries, education, sports, and so on. While we'll always wax poetically about

data visualization (you're on the Tableau website, after all) there are practical, real-life applications that are undeniable. And, since visualization is so prolific, it's also one of the most useful professional skills to develop. The better you can convey your points visually, whether in a dashboard or a slide deck, the better you can leverage that information. The concept of the citizen data scientist is on the rise. Skill sets are changing to accommodate a data-driven world. It is increasingly valuable for professionals to be able to use data to make decisions and use visuals to tell stories of when data informs the who, what, when, where, and how. While traditional education typically draws a distinct line between creative storytelling and technical analysis, the modern professional world also values those who can cross between the two: data visualization sits right in the middle of analysis and visual storytelling.

ABSTRACT

Datasets are the foundation and starting point for visualizing your data. They are defined on the connections to your data and provide access to the specific tables in the data store.

A dataset is the logical representation of the data you want to use to build visuals. It is a logical pointer to a physical table or a defined structure in your data source. Datasets may represent the contents of a single data table or a data matrix from several tables that may be in different data stores on the same connection.

Other than providing access to data, datasets enhance data access and use in many ways, including (but not limited to):

- Table joins allow you to supplement the primary data with information from various other data sources. For more information, see *Data modeling*.

- Derived fields/attributes support flexible expressions, both for dimensions and for aggregates. For more information, see *Creating calculated fields*.

- Hiding fields enables you to eliminate the fields that are unnecessary to the business use case or to obscure sensitive data without affecting the base tables. For more information, see *Hiding dataset fields from applications*.

- Changing data types of the field attributes often helps you to deal with data types, or to ensure that numeric codes (like event ids) are processed correctly. For more information, see *Changing data type*.

- Changing the default aggregation of fields at the dataset level prevents common mistakes when building visuals. For more information, see *Changing field aggregation*.

- Providing user-friendly names for native columns or derived attributes often makes the visuals more accessible and saves some of the efforts of applying aliases to each field of the visual. For more information, see *Automatically renaming dataset fields* and *Custom renaming dataset fields*.

- Different datasets are created in different ways. In this post, you'll find links to sources with all kinds of datasets. Some of them will be machine-generated data. Some will be data that's been collected via surveys. Some may be data that's recorded from human observations. Some may be data that's been scraped from websites or pulled via APIs.

PROBLEM STATEMENT

The Data set given contains the data for FIFA world cup and its detailed information about the matches stats . Analyze the data set and its details , And represent the data set as organized form like bar charts , histogram , pie chart etc. Using python in arranged format with relevant attributes and names.

Thanks to visualization, most complex problems can be broken down into simpler elements for data scientists to figure out the optimal model architectures and solutions to complicated tasks. Hence, visualizations play a vital role in the successful completion of every major Data Science project. Without the use of visualization, it is nearly impossible to gauge the data patterns of a difficult task.

Here we will understand some of the basic features of data visualizations and try to interpret the benefits of exploratory data analysis while solving any kind of task. We will discuss a few essential libraries in python for visualization purposes . Then, we will have a detailed solution on the problem. Finally, we will conclude with a real-time example of interpreting these visualizations.

METHODOLOGY

Import the packages which are required to implement the Visualization of the Project like Matplotlib, Seaborn, Pandas, NumPy.

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

- Create publication quality plots.
- Make interactive figures that can zoom, pan, update.
- Customize visual style and layout.
- Export to many file formats.
- Embed in JupyterLab and Graphical User Interfaces.
- Use a rich array of third-party packages built on Matplotlib.

Seaborn is an open-source Python library built on top of matplotlib. It is used for data visualization and exploratory data analysis. Seaborn works easily with data frames and the Pandas library. The graphs created can also be customized easily.

Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.

NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It is open-source software. It contains various features including these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

After installing and importing these packages into the project try to implement them in order to visualize the given dataset.

At the beginning read the given dataset using read_csv, which is the built-in function in the Pandas library to read the CSV file.

After Drop the columns from the dataset which are not required for the Visualization of the Data.

Set the Index with respect to the data and time using set_index function which is included from the pandas library.

After this You can use the Seaborn Library to visualize the given data with respect to the required format

Experimental Results :

Implementation:

Data Visualization of the FIFA World Cup

In [1]:
```
#importing a bunch of packages
from matplotlib import pyplot as plt
import pandas as pd
import seaborn as sns
import numpy as np
%matplotlib inline
```

In [2]:
```
#Reading the Data
df = pd.read_csv('C:/Users/Admin/OneDrive/Desktop/Data Mining Project/Excel-to-Mysql-ETL-master/data/world.csv')
```

Checking the rows of the file

In [3]:
```
df.head()
```

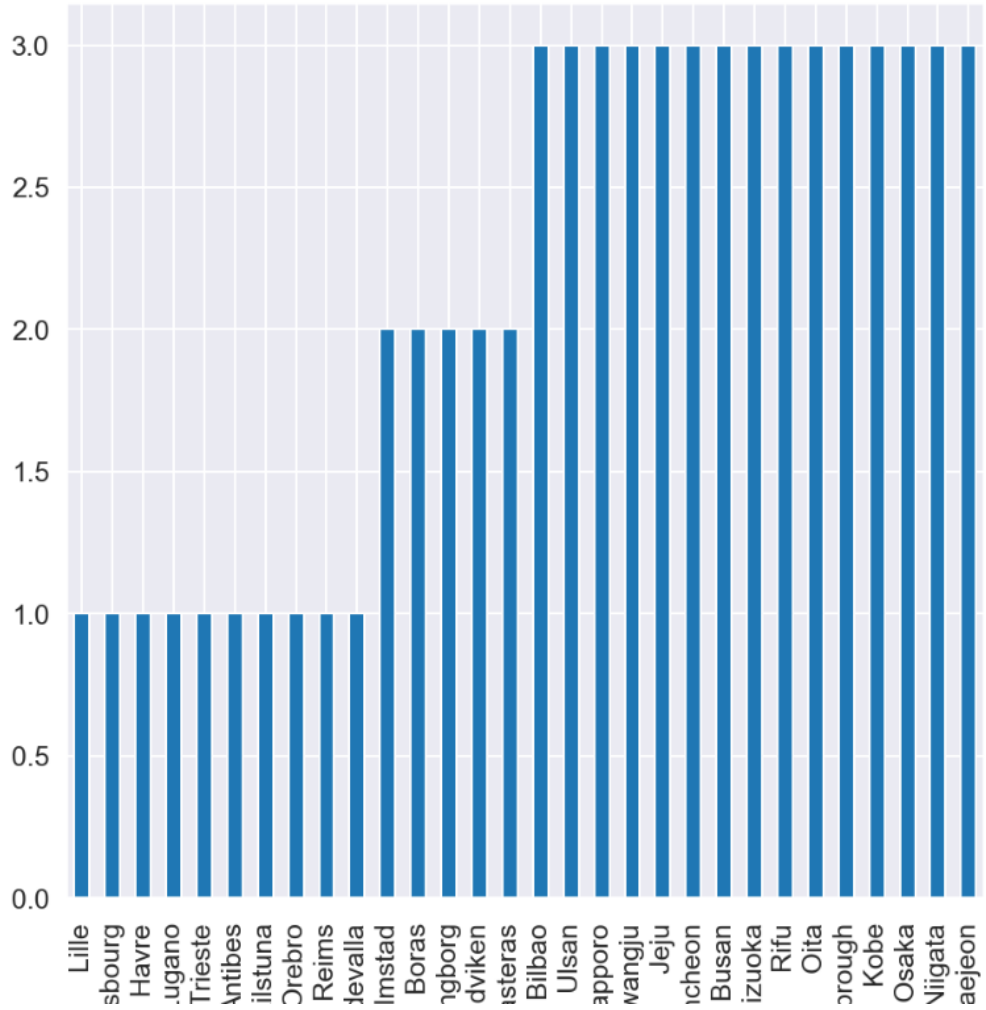
Out[3]:

	Year	Datetime	Stage	Stadium	City	Home Team Name	Home Team Goals	Away Team Goals	Away Team Name	Win conditions	Attendance	Half-time Home Goals	Half-time Away Goals	Referee	Assistant 1	Assistant 2
0	1930	13 Jul 1930 - 15:00	Group 1	Pocitos	Montevideo	France	4	1	Mexico		4444.0	3	0	LOMBARDI Domingo (URU)	CRISTOPHE Henry (BEL)	REGO Gilberto (BRA)
1	1930	13 Jul 1930 - 15:00	Group 4	Parque Central	Montevideo	USA	3	0	Belgium		18346.0	2	0	MACIAS Jose (ARG)	MATEUCCI Francisco (URU)	WARNKEN Alberto (CHI)
2	1930	14 Jul 1930 - 12:45	Group 2	Parque Central	Montevideo	Yugoslavia	2	1	Brazil		24059.0	2	0	TEJADA Anibal (URU)	VALLARINO Ricardo (URU)	BALWAY Thomas (FRA)
3	1930	14 Jul 1930 - 14:50	Group 3	Pocitos	Montevideo	Romania	3	1	Peru		2549.0	1	0	WARNKEN Alberto (CHI)	LANGENUS Jean (BEL)	MATEUCCI Francisco (URU)
4	1930	15 Jul 1930 - 16:00	Group 1	Parque Central	Montevideo	Argentina	1	0	France		23409.0	0	0	REGO Gilberto (BRA)	SAUCEDO Ulises (BOL)	RADULESCU Constantin (ROU)

In [4]:
```
df.tail()
```

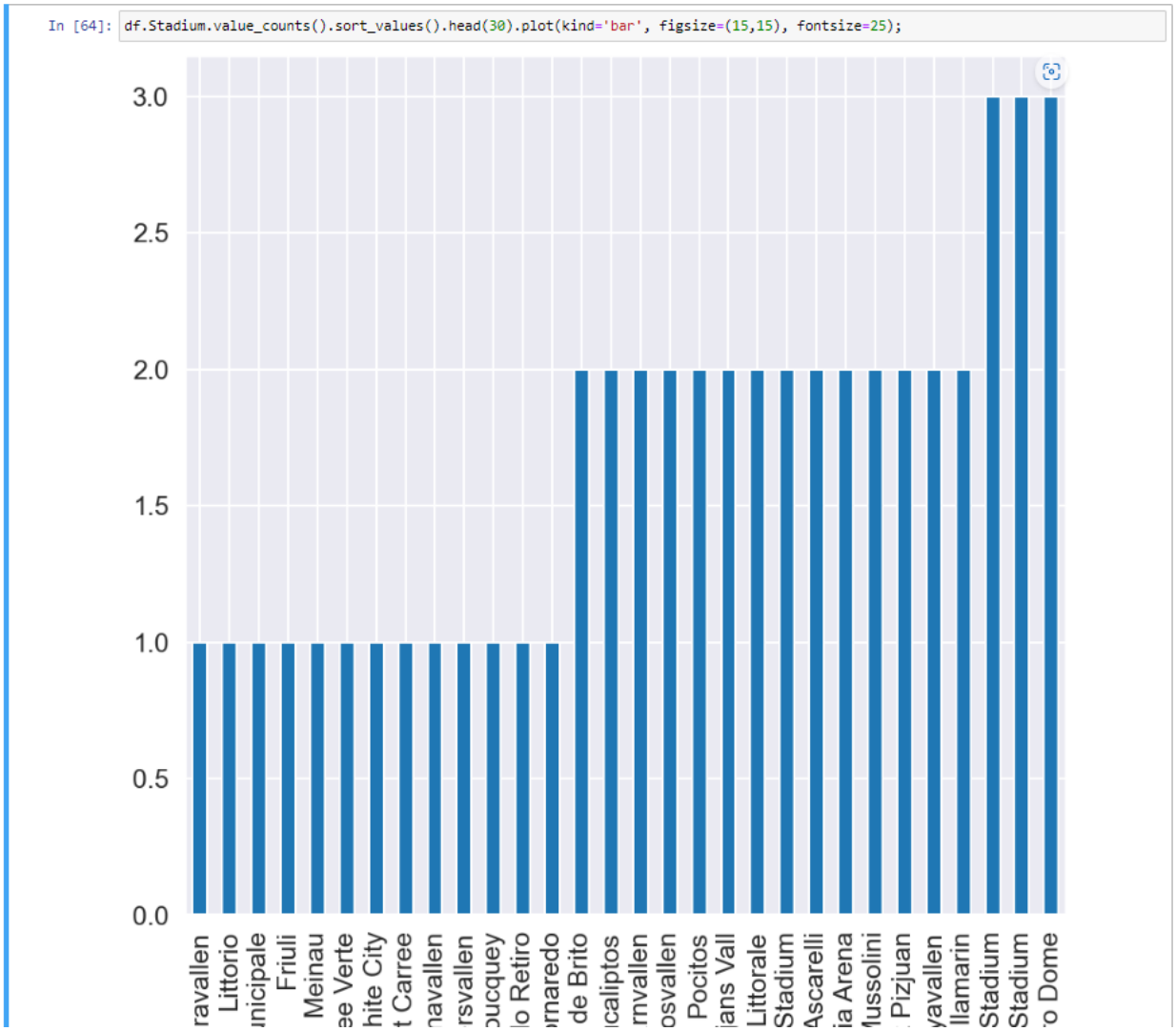
How many matches are been played in each city

```
In [63]: df.City.value_counts().sort_values().head(30).plot(kind='bar', figsize=(15,15), fontsize=25);
```



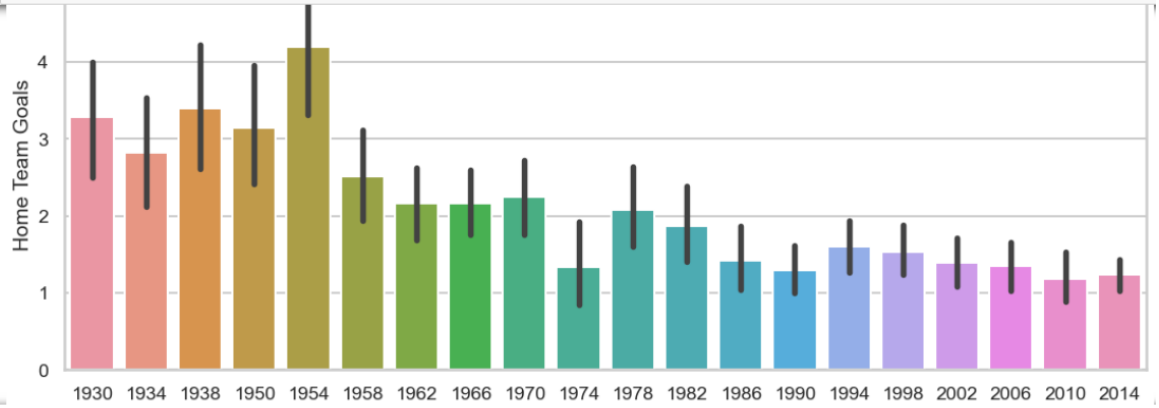
How many matches are been played in each Stadium

```
In [64]: df.Stadium.value_counts().sort_values().head(30).plot(kind='bar', figsize=(15,15), fontsize=25);
```



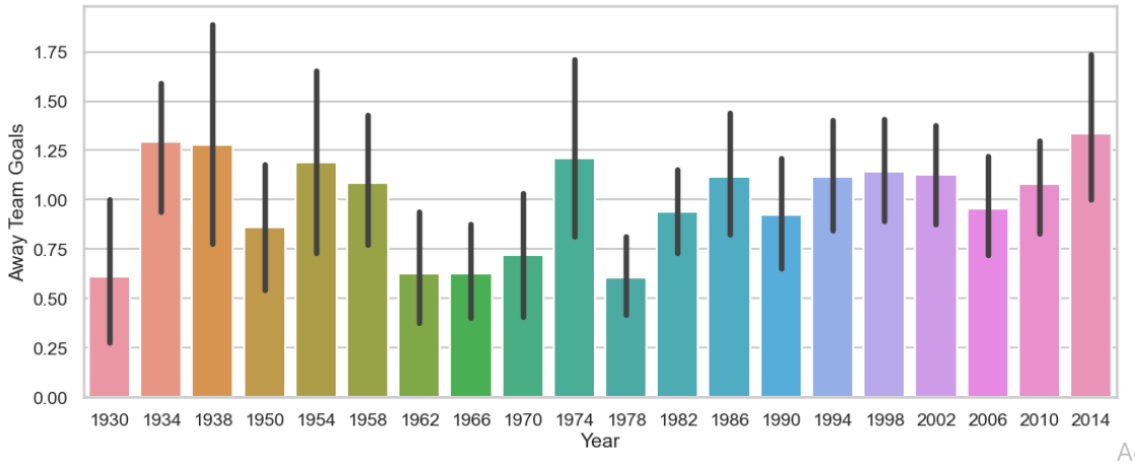
Home team goals per year

In [16]:
```
sns.set_style('whitegrid')
sns.set_context("poster", font_scale=0.8)

f, ax = plt.subplots(figsize=(18, 7))
ax = sns.barplot(data=df, x='Year', y='Home Team Goals')
plt.show()
```



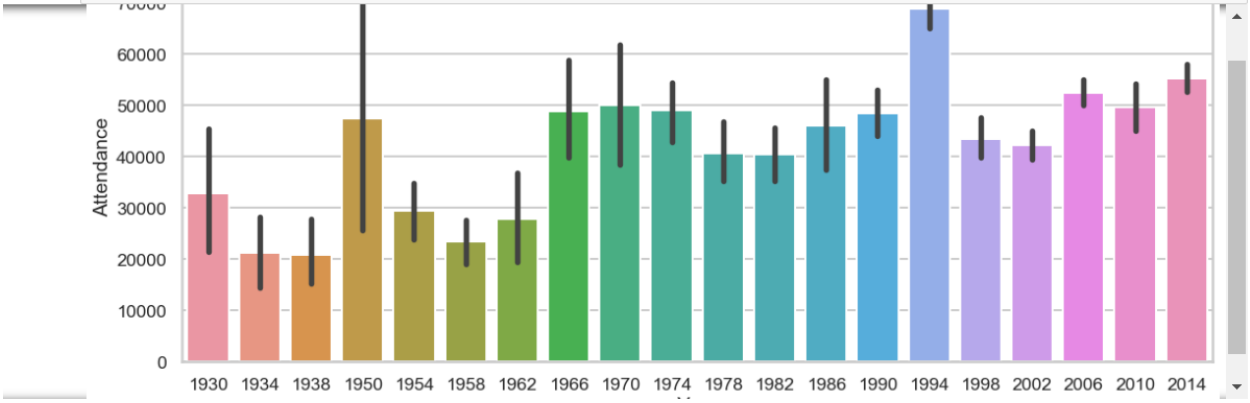
Away team goals per year

In [17]:
```
sns.set_style('whitegrid')
sns.set_context("poster", font_scale=0.8)

f, ax = plt.subplots(figsize=(18, 7))
ax = sns.barplot(data=df, x='Year', y='Away Team Goals')
plt.show()
```



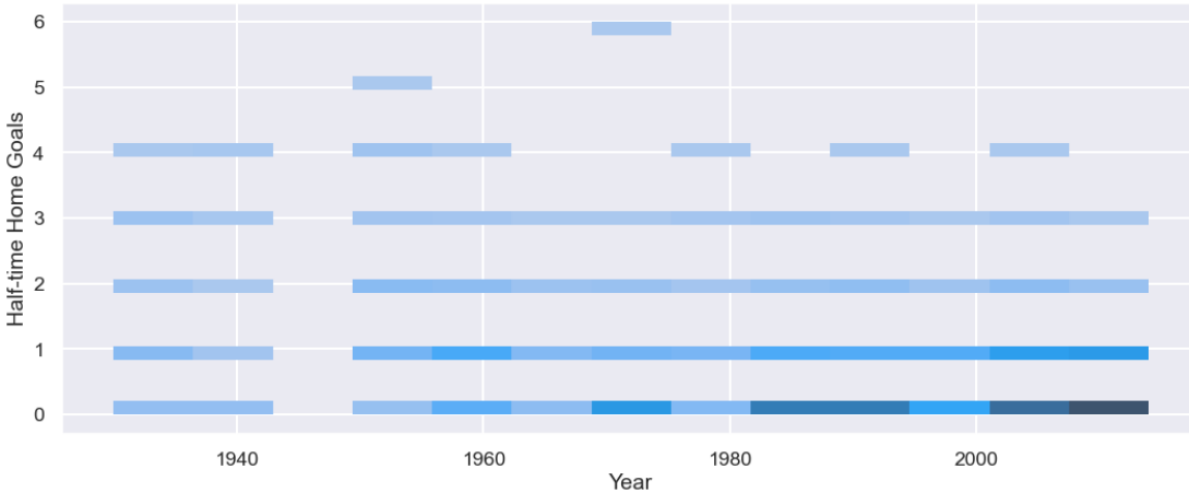
Attendence per year

In [18]:
```
sns.set_style('whitegrid')
sns.set_context("poster", font_scale=0.8)

f, ax = plt.subplots(figsize=(18, 7))
ax = sns.barplot(data=df, x='Year', y='Attendance')
plt.show()
```



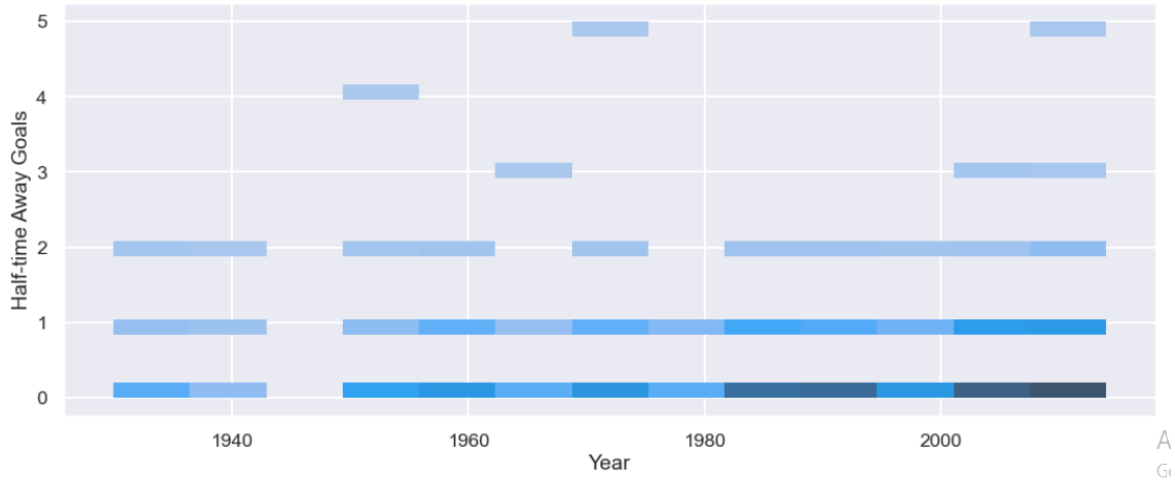
Half time away goals

[54]:
```
sns.set_style('darkgrid')
sns.set_context("poster", font_scale=0.8)

f, ax = plt.subplots(figsize=(18, 7))
ax = sns.histplot(data=df, x='Year', y='Half-time Home Goals')
plt.show()
```



Half time away goals

[n [55]:
```
sns.set_style('darkgrid')
sns.set_context("poster", font_scale=0.8)

f, ax = plt.subplots(figsize=(18, 7))
ax = sns.histplot(data=df, x='Year', y='Half-time Away Goals')
plt.show()
```



Data Set (FIFA World cup Stats):

Year	Datetime	Stage	Stadium	City	Home Tea	Home Tea	Away Tear	Away Tear	Win condit	Attendanc	Half-time I	Half-time ,	Referee	Assistant 1	Assistant 2	RoundID	MatchID	Home Tea	Away Team Initials
1930	13 Jul 193(	Group 1	Pocitos	Montevide	France	4	1	Mexico		4444	3	0	LOMBARD	CRISTOPH	REGO Gilb	201	1096	FRA	MEX
1930	13 Jul 193(	Group 4	Parque Ce	Montevide	USA	3	0	Belgium		18346	2	0	MACIAS Jo	MATEUCC	WARNKEN	201	1090	USA	BEL
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 wir	16000	0	0	VAN MOO	CAIRONI C	BAERT Lou	204	1104	AUT	FRA
1934	27 May 19	Preliminar	Giorgio As	Naples	Hungary	4	2	Egypt		9000	2	2	BARLASSIN	DATTILO G	SASSI Otell	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 B	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

The dataset used for experimentation was retrieved from “kaggle.com”. The measures used for evaluating the performance are precision, recall, and accuracy. The results are narrowed down to the analysis eliminating the noises and the errors in the data set.

Conclusion

We need data visualization because a visual summary of information makes it easier to identify patterns and trends than looking through thousands of rows on a spreadsheet. It's the way the human brain works. Charts and graphs make communicating data findings easier even if you can identify the patterns without them.

As we conclude our brief study on data visualization, it is clear that the field is rich in potential applications in diverse disciplines, at the same time we need to be aware of its practical and ethical complexities. Previously we implemented the data set in many forms which are easy to comprehend and interpret, using python and important libraries.

References:

1.https://www.tableau.com/learn/articles/data-visualization.

2. https://www.adamenfroy.com/data-visualization-tools.

3. https://seaborn.pydata.org/

4. https://pandas.pydata.org/

5. https://matplotlib.org/