Project Name: Baseball Game Logs Analysis Project

Objective

The objective of this project is to analyse baseball game logs to gain insights into game trends, attendance patterns, and scheduling. The specific goals include:

1. **Loading and Preprocessing Data**: Import the dataset and handle data types and formatting.
2. **Creating and Analysing Pivot Tables**: Summarize and visualize game attendance data by day of the week and year.
3. **Visualizing Trends**: Use scatter plots and other visualizations to explore trends in game attendance over time.

Summary

The dataset used for this analysis consists of baseball game logs, including details such as game dates, home team names, and attendance figures. The project involves several key steps:

1. **Data Loading**:
    - Imported the dataset from a CSV file.
    - Handled potential warnings related to data types using low_memory=False.
2. **Data Preprocessing**:
    - Converted the date column to a datetime format and extracted the year for further analysis.
3. **Pivot Table Analysis**:
    - Created a pivot table to analyze the distribution of games by day of the week and year.
    - Normalized the data to show the proportion of games played each day of the week.
4. **Visualizations**:
    - **Pivot Table Visualization**: Created an area plot to show the proportion of games played on each day of the week over the years.
    - **Scatter Matrix Plot**: Visualized relationships between numeric variables such as year and attendance using scatter matrices.
    - **Yearly Attendance**: Aggregated attendance data by year and plotted a scatter plot to explore trends over time.

Results

1. **Pivot Table Analysis**:
    - The area plot indicates the distribution of games across different days of the week and how this distribution has changed over the years. Each day's proportion of games is represented, allowing for comparison of game scheduling patterns.

2. **Scatter Matrix Plot**:
    - The scatter matrix revealed potential correlations between attendance and year. Diagonal plots show the density of attendance values, while off-diagonal plots reveal relationships between different variables.
3. **Yearly Attendance Scatter Plot**:
    - The scatter plot of total annual attendance showed trends in game attendance over time, highlighting fluctuations and potential growth or decline in attendance.

Conclusion

The Baseball Game Logs Analysis Project provides valuable insights into game scheduling and attendance trends:

- **Game Scheduling**: The area plot analysis shows how game distribution varies by day of the week and can inform scheduling strategies.
- **Attendance Trends**: The scatter matrix and yearly scatter plot reveal trends and relationships in attendance, which can be useful for future planning and analysis.

Code :

```
import pandas as pd
#Read the csv file
#q1 = pd.read_csv("https://raw.githubusercontent.com/swapnilsaurav/Dataset/master/baseball_game_logs.csv")
#Display top 5 dataset
#print(q1.head())


#DtypeWarning: for remove this warning we have to use following code
g1 = pd.read_csv("https://github.com/adityapathak0007/Project_Portfolio/blob/8274188e799b1615269ad2d041760683d27a0db5/Baseball%20Game%20Logs%20Analysis%20Project/baseball_game_logs.csv", low_memory=False)
print(g1.head())


#Create a Multiindex
'''
A Multiindex is a hierarchical index that groups the data so that you can easily summarize it. For e.g. by grouping by home_team
and then by the game attendance, you can plot the average attendance per home team
'''
#df2 = g1.set_index(['h_name','attendance'])
#print(df2[0:10])


#check the data with the 'describe()' method
#print(df2.describe())
```

```
'''
Analyze the data:
We will be creating a spreadsheet-style Pivot table using Pandas. Syntax is:
(pandas.pivot_table(data, values=None, index=None, columns=None, aggfunc='mean',
fill_value=None, margins=False, dropna=True, margins_name='All', observed=Flase)
The levels in the pivot table will be stored in MultiIndex objects(hierarchical indexes)
on the index and columns of the result DataFrame.
'''
#Analyzing baseball games
import matplotlib.pyplot as plt

#convert q1 date type to date
g1['date'] = pd.to_datetime(g1['date'], format='%Y%m%d')
#print(q1['date'])
#Take Year
g1['year'] = g1['date'].dt.year



#Creating PIVOT

games_per_day = g1.pivot_table(index='year',columns='day_of_week',values='date',aggfunc=len)
games_per_day = games_per_day.divide(games_per_day.sum(axis=1),axis=0)

ax = games_per_day.plot(kind='area',stacked='true')
ax.legend(loc='upper right')
ax.set_ylim(0,1)

plt.show()
```

```
'''
Scatter Plot:
We will use Pandas scatter_matrix method to explore trends in data. We will use Panda's Scatter
matrices (pair plots). A scatter matrix(pairs plot) compactly plots all the numeric variables we
have in a dataset against each other. In Python, this data visualization technique can be carried
out with many libraries but if we are using Pandas to load the data, we can use the base
scatter_matrix method to visualize the dataset. Scatter_Matrix takes a dataframe as input.
'''


#Plot Scatter plot
from pandas.plotting import scatter_matrix
#scatter_matrix takes dataframe as input, we need to create our own dataframe before we call scatter_matrix
#We will use year and attendance
#Plotting attendance v year


dfgroup = pd.DataFrame(zip(g1['year'],g1['attendance']),index=g1["day_of_week"],columns=["year",'attendance'])
scatter_matrix(dfgroup,figsize=(12,12), diagonal='kde')
plt.show()



#Scatter Plot by aggregating the data at Year level:


import seaborn as sns
att_yr = g1.groupby('year').agg({'attendance': ['sum']}).reset_index()
att_yr.plot.scatter(x='year',y='attendance',c='red')
plt.show()
```