NSSA 220
Task Automation with Interpreted Languages

Pandas

Instructor: Dr. Fahed Jubair

RIT DUBAI

Pandas

- Pandas is a popular python library for analyzing and manipulating datasets, given in various formats such as CSV and Excel sheets

- To install Pandas, execute the command:

```
pip3 install pandas
```

Example Data Set

- For demonstration, we will use the below dataset, which is given as a CSV file, called TeddyBallgame.csv

Year	Age	GamesPlayed	HomeRuns	RunsBattedIn	BattingAverage
1940	21	144	23	113	0.344
1941	22	143	37	120	0.406
1942	23	150	36	137	0.356
1946	27	150	38	123	0.342
1947	28	156	32	114	0.343
1948	29	137	25	127	0.369
1949	30	155	43	159	0.343
1950	31	89	28	97	0.317
1951	32	148	30	126	0.318
1952	33	6	1	3	0.4
1953	34	37	13	34	0.407
1954	35	117	29	89	0.345
1955	36	98	28	83	0.356
1956	37	136	24	82	0.345
1957	38	132	38	87	0.388
1958	39	129	26	85	0.328
1959	40	103	10	43	0.254
1960	41	113	29	72	0.316

A Starting Example

- Run the following program and see the output

```
import pandas as pd

df = pd.read_csv('TeddyBallgame.csv')

print("shape is", df.shape)
print(df.head(5))
print(df.columns)
print(df.index)
```

Pandas is commonly imported using the ‘pd’ alias

Read the CSV file into a **dataframe**

Print the dimensionality

Print the top five rows

Print the name of the columns

Print index information

DataFrame Object

- The primary data structure for representing data as a 2D table with labeled columns and rows

- There is a rich support for attributes and methods to analyze, manipulate, and visualize data inside dataframes

- Each columns inside a dataframe is called a series, and represented as a Series object

Accessing Data in a DataFrame

- Run the following program and see the output

```
import pandas as pd

df = pd.read_csv('TeddyBallgame.csv')

print(df.loc[0])                    # print row 0
print(df.loc[0:4])                  # print rows 0-4
print(df['Year’])                   # print the column labeled 'Year'
print(df[['Age','HomeRuns']])       # print two columns
print(df['Age'][0])                 # print the first row in ‘Age'
print(df[['Age','HomeRuns']][0:9]) #rows 0-9 in the given columns
```

Printing Data Information

```python
import pandas as pd

df = pd.read_csv('TeddyBallgame.csv')

print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19 entries, 0 to 18
Data columns (total 6 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Year            19 non-null     int64
 1   Age             19 non-null     int64
 2   GamesPlayed     19 non-null     int64
 3   HomeRuns        19 non-null     int64
 4   RunsBattedIn    19 non-null     int64
 5   BattingAverage  19 non-null     float64
dtypes: float64(1), int64(5)
memory usage: 1.0 KB
```

Mathematical Functions

```python
import pandas as pd

df = pd.read_csv('TeddyBallgame.csv')

print(df['GamesPlayed'].sum())
print(df['GamesPlayed'].max())
print(df['GamesPlayed'].min())
print(df['HomeRuns'].mean())
print(df['HomeRuns'].median())
print(df['BattingAverage'].std())
```

Creating New Columns

```python
import pandas as pd

df = pd.read_csv('TeddyBallgame.csv')

df['GamesPlayedPercentage'] = df['GamesPlayed'] / df['GamesPlayed'].sum()
df['GamesPlayedPercentage'] = df['GamesPlayedPercentage'] * 100
df['GamesPlayedPercentage'] = round(df['GamesPlayedPercentage'], 2)

print(df.head(5))
print(df.columns)
```

Creating New Columns
Another Example

```
import pandas as pd
df = pd.read_csv('TeddyBallgame.csv')

def convert(a):
    if a < 25:
        return 'D'
    elif a < 50:
        return 'C'
    elif a < 100:
        return 'B'
    else:
        return 'A'

df['letter'] = df['GamesPlayed'].apply(convert)

df.to_csv('newData.csv')
```

Write dataframe to a CSV file

Data Filtering

```
import pandas as pd
df = pd.read_csv('TeddyBallgame.csv')

df1 = df[df['GamesPlayed'] < 100]
print(df1)

df2 = df[df['Age'].isin([27,30,35])]
print(df2)
```

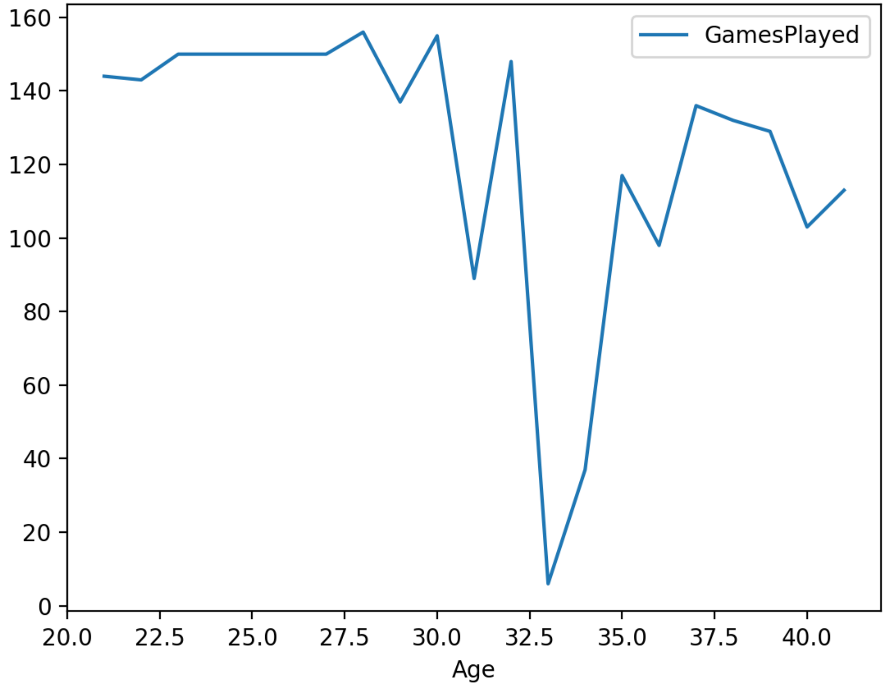
Data Plotting

```
import matplotlib.pyplot as plt
import pandas as pd

df = pd.read_csv('TeddyBallgame.csv')

df.plot(kind = 'line', x = 'Age', y = 'GamesPlayed')

plt.show()
```



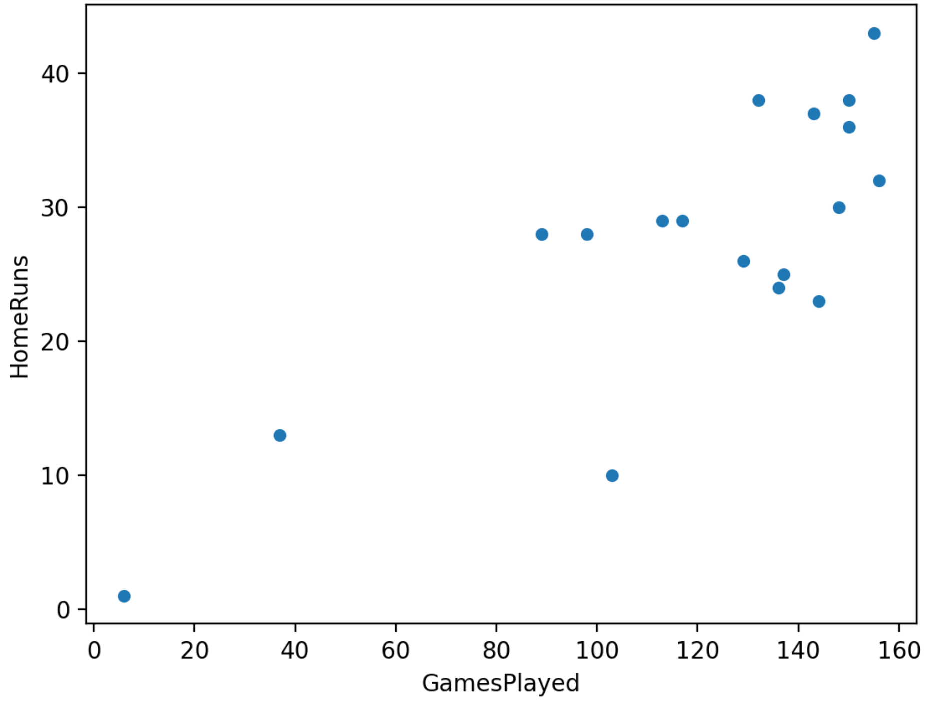
Data Plotting
Another Example

```
import matplotlib.pyplot as plt
import pandas as pd

df = pd.read_csv('TeddyBallgame.csv’)

df.plot(kind='scatter', x='GamesPlayed', y='HomeRuns')

plt.show()
```



Exercise

- Write a python script that generates a bar plot with proper labels and legends that show three bars:
  - The first bar represents the total count of games Teddy played during the age 20-29
  - The second bar represents the total count of games Teddy played during the age 30-39
  - The third bar represents the total count of games Teddy played during the age 40-49