Pandas Library for beginner

Qusay AL-Btoush

github.com/qusaybtoush
https://www.linkedin.com/in/qusayal-btoush
https://www.kaggle.com/qusaybtoush1990

pandas

For beginner

Pandas is a popular Python library for data manipulation and analysis. It provides data structures and functions that make working with structured data (e.g., spreadsheets, SQL tables, and CSV files) more accessible and efficient. Pandas is commonly used in data science, data analysis, and machine learning tasks.

Qusay AL-Btoush

- https://github.com/qusaybtoush
- https://www.linkedin.com/in/qusayal-btoush/
- https://www.kaggle.com/qusaybtoush1990

In [17]:
```
# import the library pandas
import pandas as pd
```

In [16]:
```
#defined the data using dictionary

di= {"Name":["Ali","Sahra","Ahmad","Mouhammed","Ali","Sahra","Ali"],
    "Age":[ 7,8,14,20,7 ,6,11],
     "Mark": ["A","A","B","A","A","D","A"]
}
```

In [19]:
```
# First, using the function Data Frame from pandas to defined data frame

df = pd.DataFrame(data=di)  # save the data in variable df or any name
df # print
```

Out[19]:

	Name	Age	Mark
0	Ali	7	A
1	Sahra	8	A
2	Ahmad	14	B
3	Mouhammed	20	A
4	Ali	7	A
5	Sahra	6	D
6	Ali	11	A

In [30]:
```
# select item in the data using iloc or loc
#iloc = I can put the number of rows or columns #loc = I can put the name of columns
# df.iloc[the rows from : to , than the number of column from : to]

df.iloc[0:3 , 0:2] # the first 3 rows and 2 columns
```

Out[30]:

	Name	Age
0	Ali	7
1	Sahra	8
2	Ahmad	14

In [34]:
```
# here in loc I can put the name of columns
df.loc[0:3,["Name","Age"]]
```

Out[34]:

	Name	Age
0	Ali	7
1	Sahra	8
2	Ahmad	14
3	Mouhammed	20

In [38]:
```
# I can change any value
df.iloc[3,1]
```

Out[38]: 20

In [39]:
```
# I can change the value based on the index
df.iloc[3,1] = 30
df #print Data Frame
```

Out[39]:

	Name	Age	Mark
0	Ali	7	A
1	Sahra	8	A
2	Ahmad	14	B
3	Mouhammed	30	A
4	Ali	7	A
5	Sahra	6	D
6	Ali	11	A

In [40]:
```
# show the first 5 rows or you can choice the number head(number) but the default 5 and
df.head()
```

Out[40]:

	Name	Age	Mark
0	Ali	7	A
1	Sahra	8	A
2	Ahmad	14	B

	Name	Age	Mark
3	Mouhammed	30	A
4	Ali	7	A

In [42]:
```
# show the shape the data
df.shape # 7 rows and 3 columns
```

Out[42]:
```
(7, 3)
```

In [43]:
```
# show the data type and the information about the columns
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7 entries, 0 to 6
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   Name    7 non-null      object
 1   Age     7 non-null      int64 
 2   Mark    7 non-null      object
dtypes: int64(1), object(2)
memory usage: 296.0+ bytes
```

In [65]:
```
#check the unique value
df["Mark"].unique()
```

Out[65]:
```
array(['A', 'B', 'D'], dtype=object)
```

In [47]:
```
#check if there any missing value
df.isnull().sum()  # you can choice the name of column ex df["Name"].isnull()
```

Out[47]:
```
Name    0
Age     0
Mark    0
dtype: int64
```

In [48]:
```
#check if there any duplication in the data
df.duplicated().sum()
```

Out[48]:
```
1
```

In [50]:
```
#show the names of columns
df.columns
```

Out[50]:
```
Index(['Name', 'Age', 'Mark'], dtype='object')
```

In [52]:
```
# you can desribe all the data set
df.describe() # show the statistics
```

Out[52]:

	Age
count	7.000000
mean	11.857143
std	8.474050
min	6.000000
25%	7.000000
50%	8.000000
75%	12.500000
max	30.000000

In [53]:

```
# make filter the name Ahmad
df[df["Name"]== "Ahmad"]
```

Out[53]:

	Name	Age	Mark
2	Ahmad	14	B

In [59]:

```
#show the statistics
print ("The Max Age : ", df["Age"].max())
print ("The Min Age : ", df["Age"].min())
print ("The Avg Age : ", df["Age"].mean())
print ("The Total Age : ", df["Age"].sum())
```

```
The Max Age :  30
The Min Age :  6
The Avg Age :  11.857142857142858
The Total Age :  83
```

Drop

In [60]:

```
# Drop
#you can drop any columns or rows and the missing value and duplicate value

#drop column
df.drop(columns= "Name") # i can writing inside drop 'inplace =True' than the change wi
```

Out[60]:

	Age	Mark
0	7	A
1	8	A
2	14	B
3	30	A
4	7	A
5	6	D
6	11	A

In [61]:
```
# drop duplicate value
df.drop_duplicates() # i can writing inside drop 'inplace =True' than the change will b
```

Out[61]:

	Name	Age	Mark
0	Ali	7	A
1	Sahra	8	A
2	Ahmad	14	B
3	Mouhammed	30	A
5	Sahra	6	D
6	Ali	11	A

In [62]:
```
#drop null / missing value
df.dropna()
```

Out[62]:

	Name	Age	Mark
0	Ali	7	A
1	Sahra	8	A
2	Ahmad	14	B
3	Mouhammed	30	A
4	Ali	7	A
5	Sahra	6	D
6	Ali	11	A

In [68]:
```
# I can fill null value
df["Age"].fillna(5) # I don`t have missing value but if have this function will fill ba
```

Out[68]:
```
0     7
1     8
2    14
3    30
4     7
5     6
6    11
Name: Age, dtype: int64
```

Group by

In [64]:
```
# I can make group by using pandas
# I will make group in the column Make with the avg Age
df.groupby("Mark")["Age"].mean()
```

Out[64]:
```
Mark
A    12.6
```

```
B    14.0
D     6.0
Name: Age, dtype: float64
```

Import and Export the data set

- using pandas you can import data set from a lot of sources like , excel, web, csv, sql...

In [ ]:
```
# import data from CSV , defined the variable liked df

# df = pd.read_csv("link or the name file or the path.csv")

#once you have the data you can do a lot of things like analysis or check every thing i

# export the data set

# if you did change in the data set or you import the data from web or any sources and

#df.to_csv()  the name like df.to_csv("new name")
```

*If you have any questions feel free to contact me : Qusay AL-Btoush

- https://github.com/qusaybtoush
- https://www.linkedin.com/in/qusayal-btoush/
- https://www.kaggle.com/qusaybtoush1990

In [ ]:
```

```