Assignment

SKILLS

Q1. List any five functions of the pandas library with execution.

Answer.

1. read_csv()

read_csv() function helps read a comma-separated values (csv) file into a Pandas DataFrame.

2. head()

head(n) is used to return the first n rows of a dataset. By default, df.head() will return the first 5 rows of the DataFrame. If you want more/less number of rows, you can specify n as an integer.

3. describe()

describe() is used to generate descriptive statistics of the data in a Pandas DataFrame or Series. It summarizes central tendency and dispersion of the dataset.

4. memory_usage()

memory_usage() returns a Pandas Series having the memory usage of each column (in bytes) in a Pandas DataFrame.

6. loc[:]

loc[:] helps to access a group of rows and columns in a dataset, a slice of the dataset, as per our requirement. For instance, if we only want the last 2 rows and the first 3 columns of a dataset, we can access them with the help of loc[:].

Q2. Given a Pandas DataFrame df with columns 'A', 'B', and 'C', write a Python function to re-index the DataFrame with a new index that starts from 1 and increments by 2 for each row.

Answer:

```
def reindex_df(df):
    new_index = range(1, 2*len(df)+1, 2)
    new_df = df.set_index(pd.Index(new_index))
    return new_df
```

Q3. You have a Pandas DataFrame df with a column named 'Values'. Write a Python function that iterates over the DataFrame and calculates the sum of the first three values in the 'Values' column. The function should print the sum to the console.

For example, if the 'Values' column of df contains the values [10, 20, 30, 40, 50], your function should calculate and print the sum of the first three values, which is 60.

Answer:

```
import pandas as pd
data = {
        "Column":[10, 20, 30, 40, 50]
        }
df = pd.DataFrame(data)
print(df)

Sum = df.Column[0:3].sum()
print(Sum)
```

OUTPUT:

```
    Column
0       10
1       20
2       30
3       40
4       50
60
```

Q4. Given a Pandas DataFrame df with a column 'Text', write a Python function to create a new column 'Word_Count' that contains the number of words in each row of the 'Text' column.

Answer:

```
df = pd.DataFrame({'Text': ['Data Science', 'Python', 'Pandas DataFrame', 'C','Java','Machine Learning']})

df['Word_Count'] = df['Text'].apply(len)

print(df)
```

OUTPUT:

```
              Text  Word_Count
0     Data Science          12
1           Python           6
2  Pandas DataFrame          16
3                C           1
4             Java           4
5  Machine Learning          16
```

Q5. How are DataFrame.size() and DataFrame.shape() different?

Answer: * DataFrame size()

The size property is used to get an int representing the number of elements in this object and Return the number of rows if Series. Otherwise, return the number of rows times the number of columns if DataFrame.

*DataFrame shape()

The shape property is used to get a tuple representing the dimensionality of the Pandas DataFrame.

Q6. Which function of pandas do we use to read an excel file?

Answer: Pandas read_excel() function is used for reading the Excel files.

Q7. You have a Pandas DataFrame df that contains a column named 'Email' that contains email addresses in the format 'username@domain.com'. Write a Python function that creates a new column 'Username' in df that contains only the username part of each email address.

The username is the part of the email address that appears before the '@' symbol. For example, if the email address is 'john.doe@example.com', the 'Username' column should contain 'john.doe'. Your function should extract the username from each email address and store it in the new 'Username' column.

Answer:

```
import pandas as pd

df=pd.DataFrame({"Email":["joe@gmail.com","roger@nasa.com","detas@uncouncil.com","robert@pandas.com
","Jenny@uk.org","michelle.clark@nypd.net"]})

df['Username'] = df['Email'].str.split('@', expand=True)[0]

df
```

OUTPUT:

	Email	Username
0	joe@gmail.com	joe
1	roger@nasa.com	roger
2	detas@uncouncil.com	detas
3	robert@pandas.com	robert
4	Jenny@uk.org	Jenny
5	michelle.clark@nypd.net	michelle.clark

Q8. You have a Pandas DataFrame df with columns 'A', 'B', and 'C'. Write a Python function that selects all rows where the value in column 'A' is greater than 5 and the value in column 'B' is less than 10. The function should return a new DataFrame that contains only the selected rows.

For example, if df contains the following values:

A B C

0 3 5 1

1 8 2 7

2 6 9 4

3 2 3 5

4 9 1 2

Your function should select the following rows: A B C

1 8 2 7

4 9 1 2

The function should return a new DataFrame that contains only the selected rows.

Answer:

```
import pandas as pd
data={'A':['3','8','6','2','9'],
      'B':['5','2','9','3','1'],
      'C':['1','7','4','5','2']
     }
df=pd.DataFrame(data)
print(df)

SelectRow=df[df['A'] >'7']
SelectRow
```

Out Put :

```
A  B  C
0  3  5  1
1  8  2  7
2  6  9  4
3  2  3  5
4  9  1  2

   A  B  C
1  8  2  7
4  9  1  2
```

Q9. Given a Pandas DataFrame df with a column 'Values', write a Python function to calculate the mean, median, and standard deviation of the values in the 'Values' column.

Answer:

```
import pandas as pd
data={'Values':[10, 20, 30, 40, 50, 60, 70, 80, 90, 100],
      }
df=pd.DataFrame(data)
print(df)

Mean=df.mean()
Median=df.median()
Std=df.std()

print("Mean=", Mean)
print("Median=", Median)
print("Standard deviation =", Std)
```

OUTPUT:

```
   Values
0      10
1      20
2      30
3      40
4      50
5      60
6      70
7      80
8      90
9     100
Mean= Values    55.0
dtype: float64
Median= Values     55.0
dtype: float64
Standard deviation = Values     30.276504
```

dtype: float64

Q10. Given a Pandas DataFrame df with a column 'Sales' and a column 'Date', write a Python function to create a new column 'MovingAverage' that contains the moving average of the sales for the past 7 days for each row in the DataFrame. The moving average should be calculated using a window of size 7 and should include the current day.

Answer:

```
import pandas as pd
data={'Date':["1st July","2nd July","3rd July","4th July","5th July","6th July","7th July"],
      "Sales":[10,20,50,30,55,65,25]
      }
df=pd.DataFrame(data)
df['MovingAverage'] = df['Sales'].rolling(window=7, min_periods=1).mean()
print(df)
```

OUTPUT:

	Date	Sales	MovingAverage
0	1st July	10	10.000000
1	2nd July	20	15.000000
2	3rd July	50	26.666667
3	4th July	30	27.500000
4	5th July	55	33.000000
5	6th July	65	38.333333
6	7th July	25	36.428571

Q11. You have a Pandas DataFrame df with a column 'Date'. Write a Python function that creates a new column 'Weekday' in the DataFrame. The 'Weekday' column should contain the weekday name (e.g. Monday, Tuesday) corresponding to each date in the 'Date' column.

For example, if df contains the following values:

	Date
0	2023-01-01
1	2023-01-02
2	2023-01-03
3	2023-01-04
4	2023-01-05

Your function should create the following DataFrame:

	Date	Weekday
0	2023-01-01	Sunday
1	2023-01-02	Monday
2	2023-01-03	Tuesday
3	2023-01-04	Wednesday
4	2023-01-05	Thursday

The function should return the modified DataFrame.

Answer:

```
import pandas as pd
data={'Date':['2023-01-01','2023-01-02','2023-01-03','2023-01-04','2023-01-05']
    }
df=pd.DataFrame(data)
print(df)

df['Date'] = pd.to_datetime(df['Date'])

df['WeekDay'] = df['Date'].dt.day_name()

print(df)
```

OUTPUT:

```
         Date
0  2023-01-01
1  2023-01-02
2  2023-01-03
3  2023-01-04
4  2023-01-05
        Date    WeekDay
0 2023-01-01     Sunday
1 2023-01-02     Monday
2 2023-01-03    Tuesday
3 2023-01-04  Wednesday
4 2023-01-05   Thursday
```

Q12. Given a Pandas DataFrame df with a column 'Date' that contains timestamps, write a Python function to select all rows where the date is between '2023-01-01' and '2023-01-31'.

Answer:

```
def select_dates_between(df):
    start_date = '2023-01-01'
    end_date = '2023-01-31'
    mask = (df['Date'] >= start_date) & (df['Date'] <= end_date)
    selected_df = df.loc[mask]
    return selected_df

select_dates_between(data)
```

Q13. To use the basic functions of pandas, what is the first and foremost necessary library that needs to be imported?

Answer:

```
import pandas as pd
```