Assignment 2 - Pandas

Atharva Ramgirkar 19BCE0114

In [ ]:

In [2]:
```python
import pandas as pd
import numpy as np
```

1.

In [3]:
```python
df1 =pd.DataFrame({'Day':[1,2,3,4,5,6],
                   'Vistors':[1200,700,5500,540,400,340],
                   'BounceRate':[20,30,22,15,10,35]})
df1
```

Out[3]:

	Day	Vistors	BounceRate
0	1	1200	20
1	2	700	30
2	3	5500	22
3	4	540	15
4	5	400	10
5	6	340	35

In [5]:
```python
df1=pd.DataFrame({'HPI':[90,70,60,80],
                  'Int_rate':[2,3,1,2],
                  'Ind_GDP':[50,35,40,45],
                  'Year':[2015,2016,2017,2018]})
df1
```

Out[5]:

	HPI	Int_rate	Ind_GDP	Year
0	90	2	50	2015
1	70	3	35	2016
2	60	1	40	2017
3	80	2	45	2018

In [15]:
```python
df2=pd.DataFrame({'HPI':[90,70,100,20],
                  'Int_rate':[2,5,3,9],
                  'Ind_GDP':[50,35,40,45],
                  'Month':['Jan','Feb','Mar','Apr']})
df2
```

Out[15]:

	HPI	Int_rate	Ind_GDP	Month
0	90	2	50	Jan
1	70	5	35	Feb
2	100	3	40	Mar
3	20	9	45	Apr

Slicing

In [8]:
```python
df1
```

Out[8]:

	HPI	Int_rate	Ind_GDP	Year
0	90	2	50	2015
1	70	3	35	2016
2	60	1	40	2017
3	80	2	45	2018

In [10]:
```python
df1.iloc[1,3]
```

Out[10]:
```
2016
```

In [11]:
```python
df1.iloc[1:,3]
```

Out[11]:
```
1    2016
2    2017
3    2018
Name: Year, dtype: int64
```

In [12]:
```python
df1.iloc[1:,2:]
```

Out[12]:

	Ind_GDP	Year
1	35	2016
2	40	2017
3	45	2018

In [13]:
```python
df1.iloc[2:,:3]
```

Out[13]:

	HPI	Int_rate	Ind_GDP
2	60	1	40
3	80	2	45

Merging

In [19]:
```python
df4=pd.merge(df1,df2,on="HPI")
df4
```

Out[19]:

	HPI	Int_rate_x	Ind_GDP_x	Year	Int_rate_y	Ind_GDP_y	Month
0	90	2	50	2015	2	50	Jan
1	70	3	35	2016	5	35	Feb

In [20]:
```python
df5=pd.merge(df1,df2,on="HPI",how="outer")
df5
```

Out[20]:

	HPI	Int_rate_x	Ind_GDP_x	Year	Int_rate_y	Ind_GDP_y	Month
0	90	2.0	50.0	2015.0	2.0	50.0	Jan
1	70	3.0	35.0	2016.0	5.0	35.0	Feb
2	60	1.0	40.0	2017.0	NaN	NaN	NaN
3	80	2.0	45.0	2018.0	NaN	NaN	NaN
4	100	NaN	NaN	NaN	3.0	40.0	Mar
5	20	NaN	NaN	NaN	9.0	45.0	Apr

In [22]:
```python
df6=pd.merge(df1,df2,on="HPI",how="right")
df6
```

Out[22]:

	HPI	Int_rate_x	Ind_GDP_x	Year	Int_rate_y	Ind_GDP_y	Month
0	90	2.0	50.0	2015.0	2	50	Jan
1	70	3.0	35.0	2016.0	5	35	Feb
2	100	NaN	NaN	NaN	3	40	Mar
3	20	NaN	NaN	NaN	9	45	Apr

Joining

In [28]:
```python
df7 = df5.join(df6,lsuffix="_l")
df7
```

Out[28]:

	HPI_l	Int_rate_x_l	Ind_GDP_x_l	Year_l	Int_rate_y_l	Ind_GDP_y_l	Month_l	HPI	Int_rate_x	Ind_GDP_x
0	90	2.0	50.0	2015.0	2.0	50.0	Jan	90.0	2.0	50.0
1	70	3.0	35.0	2016.0	5.0	35.0	Feb	70.0	3.0	35.0
2	60	1.0	40.0	2017.0	NaN	NaN	NaN	100.0	NaN	NaN
3	80	2.0	45.0	2018.0	NaN	NaN	NaN	20.0	NaN	NaN
4	100	NaN	NaN	NaN	3.0	40.0	Mar	NaN	NaN	NaN
5	20	NaN	NaN	NaN	9.0	45.0	Apr	NaN	NaN	NaN

In [29]:
```python
df8 = df6.join(df5,lsuffix="_l")
df8
```

Out[29]:

	HPI_l	Int_rate_x_l	Ind_GDP_x_l	Year_l	Int_rate_y_l	Ind_GDP_y_l	Month_l	HPI	Int_rate_x	Ind_GDP_x
0	90	2.0	50.0	2015.0	2	50	Jan	90	2.0	50.0
1	70	3.0	35.0	2016.0	5	35	Feb	70	3.0	35.0
2	100	NaN	NaN	NaN	3	40	Mar	60	1.0	40.0
3	20	NaN	NaN	NaN	9	45	Apr	80	2.0	45.0

Concatenation

In [31]:
```python
df9=pd.concat([df7,df8])
df9
```

Out[31]:

	HPI_l	Int_rate_x_l	Ind_GDP_x_l	Year_l	Int_rate_y_l	Ind_GDP_y_l	Month_l	HPI	Int_rate_x	Ind_GDP_x
0	90	2.0	50.0	2015.0	2.0	50.0	Jan	90.0	2.0	50.0
1	70	3.0	35.0	2016.0	5.0	35.0	Feb	70.0	3.0	35.0
2	60	1.0	40.0	2017.0	NaN	NaN	NaN	100.0	NaN	NaN
3	80	2.0	45.0	2018.0	NaN	NaN	NaN	20.0	NaN	NaN
4	100	NaN	NaN	NaN	3.0	40.0	Mar	NaN	NaN	NaN
5	20	NaN	NaN	NaN	9.0	45.0	Apr	NaN	NaN	NaN
0	90	2.0	50.0	2015.0	2.0	50.0	Jan	90.0	2.0	50.0
1	70	3.0	35.0	2016.0	5.0	35.0	Feb	70.0	3.0	35.0
2	100	NaN	NaN	NaN	3.0	40.0	Mar	60.0	1.0	40.0
3	20	NaN	NaN	NaN	9.0	45.0	Apr	80.0	2.0	45.0

In [32]:
```python
df10=pd.concat([df9,df8])
df10
```

Out[32]:

	HPI_l	Int_rate_x_l	Ind_GDP_x_l	Year_l	Int_rate_y_l	Ind_GDP_y_l	Month_l	HPI	Int_rate_x	Ind_GDP_x
0	90	2.0	50.0	2015.0	2.0	50.0	Jan	90.0	2.0	50.0
1	70	3.0	35.0	2016.0	5.0	35.0	Feb	70.0	3.0	35.0
2	60	1.0	40.0	2017.0	NaN	NaN	NaN	100.0	NaN	NaN
3	80	2.0	45.0	2018.0	NaN	NaN	NaN	20.0	NaN	NaN
4	100	NaN	NaN	NaN	3.0	40.0	Mar	NaN	NaN	NaN
5	20	NaN	NaN	NaN	9.0	45.0	Apr	NaN	NaN	NaN
0	90	2.0	50.0	2015.0	2.0	50.0	Jan	90.0	2.0	50.0
1	70	3.0	35.0	2016.0	5.0	35.0	Feb	70.0	3.0	35.0
2	100	NaN	NaN	NaN	3.0	40.0	Mar	60.0	1.0	40.0
3	20	NaN	NaN	NaN	9.0	45.0	Apr	80.0	2.0	45.0
0	90	2.0	50.0	2015.0	2.0	50.0	Jan	90.0	2.0	50.0
1	70	3.0	35.0	2016.0	5.0	35.0	Feb	70.0	3.0	35.0
2	100	NaN	NaN	NaN	3.0	40.0	Mar	60.0	1.0	40.0
3	20	NaN	NaN	NaN	9.0	45.0	Apr	80.0	2.0	45.0

In [34]:
```python
df11=pd.concat([df1,df2],axis=1)
df11
```

Out[34]:

	HPI	Int_rate	Ind_GDP	Year	HPI	Int_rate	Ind_GDP	Month
0	90	2	50	2015	90	2	50	Jan
1	70	3	35	2016	70	5	35	Feb
2	60	1	40	2017	100	3	40	Mar
3	80	2	45	2018	20	9	45	Apr

In [ ]:

2.

In [36]:
```python
df = pd.read_csv("Data.csv")
```

In [37]:
```python
df.head()
```

Out[37]:

	Country	Age	Salary	Purchased
0	France	44.0	72000.0	No
1	Spain	27.0	48000.0	Yes
2	Germany	30.0	54000.0	No
3	Spain	38.0	61000.0	No
4	Germany	40.0	NaN	Yes

In [38]:
```python
df.tail()
```

Out[38]:

	Country	Age	Salary	Purchased
5	France	35.0	58000.0	Yes
6	Spain	NaN	52000.0	No
7	France	48.0	79000.0	Yes
8	Germany	50.0	83000.0	No
9	France	37.0	67000.0	Yes

In [40]:
```python
df.dtypes
```

Out[40]:
```
Country       object
Age          float64
Salary       float64
Purchased     object
dtype: object
```

Null Values

In [51]:
```python
df_nulls = pd.DataFrame(df.isnull().sum(),
                        columns=["Null Values"])
df_nulls
```

Out[51]:

	Null Values
Country	0
Age	1
Salary	1
Purchased	0

In [ ]:

3.

Unique Values

In [54]:
```python
df.columns
```

Out[54]:
```
Index(['Country', 'Age', 'Salary', 'Purchased'], dtype='object')
```

In [55]:
```python
df.Country.unique()
```

Out[55]:
```
array(['France', 'Spain', 'Germany'], dtype=object)
```

In [56]:
```python
df.Purchased.unique()
```

Out[56]:
```
array(['No', 'Yes'], dtype=object)
```

Change Column Name

In [57]:
```python
df.columns
```

Out[57]:
```
Index(['Country', 'Age', 'Salary', 'Purchased'], dtype='object')
```

In [59]:
```python
df12 = df.rename(columns={"Country": "Region"})
df12
```

Out[59]:

	Region	Age	Salary	Purchased
0	France	44.0	72000.0	No
1	Spain	27.0	48000.0	Yes
2	Germany	30.0	54000.0	No
3	Spain	38.0	61000.0	No
4	Germany	40.0	NaN	Yes
5	France	35.0	58000.0	Yes
6	Spain	NaN	52000.0	No
7	France	48.0	79000.0	Yes
8	Germany	50.0	83000.0	No
9	France	37.0	67000.0	Yes

In [61]:
```python
df12.shape
```

Out[61]:
```
(10, 4)
```

Changing Index Values

In [74]:
```python
df13 = df12.set_index("Region")
df13
```

Out[74]:

	Age	Salary	Purchased
Region			
France	44.0	72000.0	No
Spain	27.0	48000.0	Yes
Germany	30.0	54000.0	No
Spain	38.0	61000.0	No
Germany	40.0	NaN	Yes
France	35.0	58000.0	Yes
Spain	NaN	52000.0	No
France	48.0	79000.0	Yes
Germany	50.0	83000.0	No
France	37.0	67000.0	Yes

Mean

In [76]:
```python
df13.Age.mean()
```

Out[76]:
```
38.77777777777778
```

In [77]:
```python
df13.Salary.mean()
```

Out[77]:
```
63777.77777777778
```

Median

In [78]:
```python
df13.Age.median()
```

Out[78]:
```
38.0
```

In [79]:
```python
df13.Salary.median()
```

Out[79]:
```
61000.0
```

Mode

In [86]:
```python
df13.Purchased.value_counts()
```

Out[86]:
```
Yes    5
No     5
Name: Purchased, dtype: int64
```

In [85]:
```python
df13.index.value_counts()
```

Out[85]:
```
France     4
Spain      3
Germany    3
Name: Region, dtype: int64
```

In [ ]: