19CSE304 Data Science Final Assignment

Dheepthi Priyangha S J

CB.EN.U4CSE20217

A - Q1

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
%matplotlib inline
import seaborn; seaborn.set()
```

```
population=pd.read_csv("state-population.csv")
area=pd.read_csv("state-areas.csv")
abbrevs=pd.read_csv("state-abbrevs.csv")
df= pd.merge(population, abbrevs, how='outer',left_on='state/region', right_on='abbreviation')
df.head()
```

	state/region	ages	year	population	state	abbreviation
0	AL	under18	2012	1117489.0	Alabama	AL
2	AL	under18	2010	1130966.0	Alabama	AL
3	AL	total	2010	4785570.0	Alabama	AL
4	AL	under18	2011	1125763.0	Alabama	AL


```
df.shape
```

```
(2544, 6)
```

Indexing

```
df=df[df['state']=='Alabama']
df.head()
```

	state/region	ages	year	population	state	abbreviation
0	AL	under18	2012	1117489.0	Alabama	AL
1	AL	total	2012	4817528.0	Alabama	AL
2	AL	under18	2010	1130966.0	Alabama	AL
3	AL	total	2010	4785570.0	Alabama	AL
4	AL	under18	2011	1125763.0	Alabama	AL

```
df.loc[(df['ages']=='total')&(df['state']=='Alaska')].head(4)
```

	state/region	ages	year	population	state	abbreviation
48	AK	total	1990	553290.0	Alaska	AK
50	AK	total	1992	588736.0	Alaska	AK
53	AK	total	1994	603308.0	Alaska	AK
55	AK	total	1991	570193.0	Alaska	AK

```
df.iloc[:,[3,4]]
```

	population	state
0	1117489.0	Alabama
1	4817528.0	Alabama
2	1130966.0	Alabama
3	4785570.0	Alabama
4	1125763.0	Alabama
...	...	...
2539	309326295.0	NaN
2540	73902222.0	NaN

Drop Duplicate

2543	313873685.0	NaN

```
df=df.drop('abbreviation',1)
df.head()
```

```
<ipython-input-23-b5b0efca35ea>:1: FutureWarning: In a future version of pandas all arguments of DataFr
  df=df.drop('abbreviation',1)
```

	state/region	ages	year	population	state
0	AL	under18	2012	1117489.0	Alabama
1	AL	total	2012	4817528.0	Alabama
2	AL	under18	2010	1130966.0	Alabama
3	AL	total	2010	4785570.0	Alabama
4	AL	under18	2011	1125763.0	Alabama

Check null

```
df.isna().any()
```

```
state/region    False
ages            False
year            False
population       True
state            True
dtype: bool
```

```
df[df['state'].isna()].head()
```

	state/region	ages	year	population	state
2448	PR	under18	1990	NaN	NaN
2449	PR	total	1990	NaN	NaN
2450	PR	total	1991	NaN	NaN
2451	PR	under18	1991	NaN	NaN
2452	PR	total	1993	NaN	NaN

```
df.loc[df['state/region'] == 'AL', 'state'] = 'Alabama'
df.loc[df['state/region'] == 'USA', 'state'] = 'United States'
df.isnull().any()
```

```
state/region    False
ages            False
year            False
population       True
state           False
dtype: bool
```

```
df = pd.merge(df, area, on='state', how='left')
df.head()
```

	state/region	ages	year	population	state	area (sq. mi)
0	AL	under18	2012	1117489.0	Alabama	52423.0

```
df.isna().any()
```

```
state/region     False
ages             False
year             False
population        True
state            False
area (sq. mi)     True
dtype: bool
```

```
df['state'][df['area (sq. mi)'].isnull()].unique()
```

```
array(['United States'], dtype=object)
```

```
df.dropna(inplace=True)
df.head()
```

	state/region	ages	year	population	state	area (sq. mi)
0	AL	under18	2012	1117489.0	Alabama	52423.0
1	AL	total	2012	4817528.0	Alabama	52423.0
2	AL	under18	2010	1130966.0	Alabama	52423.0
3	AL	total	2010	4785570.0	Alabama	52423.0
4	AL	under18	2011	1125763.0	Alabama	52423.0

```
df.isna().any()
```



```
year             False
population       False
state            False
area (sq. mi)    False
dtype: bool
```

Aggregation

```
df.describe()
```

	year	population	area (sq. mi)
count	2476.000000	2.476000e+03	2476.000000
mean	2001.556543	3.482132e+06	73452.686591
std	6.917905	4.986552e+06	94687.159589
min	1990.000000	1.013090e+05	68.000000
25%	1996.000000	7.306692e+05	35387.000000
50%	2002.000000	1.557804e+06	56276.000000
75%	2008.000000	4.373440e+06	84904.000000
max	2013.000000	3.833252e+07	656425.000000

```
df['population'].sum()
```

```
8621759560.0
```

```
df['area (sq. mi)'].sum()
```

```
181868852.0
```

```
df.mad()
```

```
year             5.991466e+00
population       3.126783e+06
area (sq. mi)    4.740079e+04
dtype: float64
```

```
df.var()
```

```
<ipython-input-37-28ded241fd7c>:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated;
  df.var()
year             4.785741e+01
population       2.486570e+13
area (sq. mi)    8.965658e+09
dtype: float64
```

Merge And join

```
a=pd.DataFrame(population)
b=pd.DataFrame(area)
c=pd.DataFrame(abbrevs)
merged_final=pd.concat([a,b,c])
merged_final
```

	state/region	ages	year	population	state	area (sq. mi)	abbreviation
0	AL	under18	2012.0	1117489.0	NaN	NaN	NaN
1	AL	total	2012.0	4817528.0	NaN	NaN	NaN
2	AL	under18	2010.0	1130966.0	NaN	NaN	NaN
3	AL	total	2010.0	4785570.0	NaN	NaN	NaN
4	AL	under18	2011.0	1125763.0	NaN	NaN	NaN
...	...	...	...	...	...	...	...
46	NaN	NaN	NaN	NaN	Virginia	NaN	VA
47	NaN	NaN	NaN	NaN	Washington	NaN	WA
48	NaN	NaN	NaN	NaN	West Virginia	NaN	WV
50	NaN	NaN	NaN	NaN	Wyoming	NaN	WY

2647 rows × 7 columns

Grouping

```
df.head()
```

	state/region	ages	year	population	state	area (sq. mi)
0	AL	under18	2012	1117489.0	Alabama	52423.0
1	AL	total	2012	4817528.0	Alabama	52423.0
2	AL	under18	2010	1130966.0	Alabama	52423.0
3	AL	total	2010	4785570.0	Alabama	52423.0
4	AL	under18	2011	1125763.0	Alabama	52423.0

```
df.groupby('state').sum()
```

state	year	population	area (sq. mi)
Alabama	96072	1.342974e+08	2516304.0
Alaska	96072	1.998907e+07	31508400.0
Arizona	96072	1.602037e+08	5472288.0
Arkansas	96072	8.095872e+07	2552736.0
California	96072	1.042036e+09	7857936.0
Colorado	96072	1.309345e+08	4996800.0
Connecticut	96072	1.021652e+08	266112.0
Delaware	96072	2.390016e+07	93792.0
District of Columbia	96072	1.678336e+07	3264.0
Florida	96072	4.839901e+08	3156384.0
Georgia	96072	2.537002e+08	2853168.0
Hawaii	96072	3.734995e+07	524736.0
Idaho	96072	4.116600e+07	4011552.0
Illinois	96072	3.723697e+08	2780064.0
Indiana	96072	1.842422e+08	1748160.0
Iowa	96072	8.792585e+07	2701248.0
Kansas	96072	8.177131e+07	3949536.0
Kentucky	96072	1.217471e+08	1939728.0
Louisiana	96072	1.349554e+08	2488464.0
Maryland	96072	1.611601e+08	595536.0
Massachusetts	96072	1.867770e+08	506640.0
Michigan	96072	2.955700e+08	4646880.0
Minnesota	96072	1.492092e+08	4173264.0
Mississippi	96072	8.612558e+07	2324832.0
Missouri	96072	1.690720e+08	3346032.0
Montana	96072	2.743707e+07	7058208.0
Nebraska	96072	5.222119e+07	3713184.0
Nevada	96072	6.289339e+07	5307216.0
New Hampshire	96072	3.684437e+07	448848.0
New Jersey	96072	2.508986e+08	418656.0
New Mexico	96072	5.641301e+07	5836464.0
New York	96072	5.622154e+08	2614800.0
North Carolina	96072	2.467548e+08	2583408.0
North Dakota	96072	1.952047e+07	3393792.0
Ohio	96072	3.398054e+08	2151744.0


```
df.groupby('ages').sum()
```

ages	year	population	area (sq. mi)
total	2477927	6.892642e+09	90934426.0
under18	2477927	1.729118e+09	90934426.0
South Dakota	96072	2.323334e+07	3701808.0

```
df.groupby('ages').mean()
```

ages	year	population	area (sq. mi)
total	2001.556543	5.567562e+06	73452.686591
under18	2001.556543	1.396702e+06	73452.686591

```
df.groupby('year').sum()
```

year	population	area (sq. mi)
1990	313841326.0	7573768.0
1991	318293960.0	7573768.0
1992	323023408.0	7573768.0
1993	327513533.0	7573768.0
1994	331766762.0	7573768.0
1995	335751543.0	7573768.0
1996	339627803.0	7573768.0
1997	343567670.0	7573768.0
1998	347285522.0	7573768.0
1999	350986232.0	7573768.0
2000	359438268.0	7580798.0
2001	362536470.0	7580798.0
2002	365450402.0	7580798.0
2003	368085401.0	7580798.0
2004	370965830.0	7580798.0
2005	373881077.0	7580798.0
2006	376941383.0	7580798.0
2007	380007220.0	7580798.0
2008	382905139.0	7580798.0
2010	388064004.0	7580798.0
2011	390040693.0	7580798.0
2012	392075149.0	7580798.0
2013	394143865.0	7580798.0


```
df.groupby('year').mean()
```

Transformation

```
1990  3.076876e+06    74252.627451
```

```
df.groupby('state').transform(lambda x:x-x.mean())
```

```
<ipython-input-46-54974673b5fa>:1: FutureWarning: Dropping invalid columns in DataFrameGroupBy.transfor
  df.groupby('state').transform(lambda x:x-x.mean())
```

	year	population	area (sq. mi)
0	10.5	-1.680373e+06	0.0
1	10.5	2.019666e+06	0.0
2	8.5	-1.666896e+06	0.0
3	8.5	1.987708e+06	0.0
4	9.5	-1.672099e+06	0.0
...	...	...	...
2491	3.5	-1.470545e+06	0.0
2492	4.5	-1.498163e+06	0.0
2493	4.5	1.319090e+06	0.0
2494	5.5	-1.525750e+06	0.0
2495	5.5	1.284055e+06	0.0

2476 rows × 3 columns

```
2007  3.653916e+06    72892.288462
```

Apply


```
total=df['population'].sum()
proportion=[]
def Percentage(df):
  proportion.append(df['population']/total)
  return proportion
a=Percentage(df)
print(a)
```

```
[0       0.000130
1       0.000559
2       0.000131
3       0.000555
4       0.000131
          ...
2491    0.000104
2492    0.000101
2493    0.000428
2494    0.000098
2495    0.000424
Name: population, Length: 2476, dtype: float64]
```

Pivot

```
df.pivot_table(index='ages',columns='year')
```

	area (sq. mi)						
year	1990	1991	1992	1993	1994	1995	1996
ages							
total	74252.627451	74252.627451	74252.627451	74252.627451	74252.627451	74252.627451	74252.627451
under18	74252.627451	74252.627451	74252.627451	74252.627451	74252.627451	74252.627451	74252.627451

2 rows × 48 columns

0s completed at 10:34 PM