Rutvik Gavaskar

BANA-680 Data Management for Business Analytics

Assignment A2

Pandas Data Management

In [36]:
```
import pandas as pd
import numpy as np
#read csv with death causes
deathcause_data=pd.read_csv('C:/Users/rutvi/Downloads/NCHS_-_Leading_Causes_of_Death__United_States.csv')
#removing United states from the state column and All causes from the Cause Name column
deathcause_data = deathcause_data[deathcause_data['State'] != 'United States']
deathcause_data =deathcause_data[deathcause_data['Cause Name'] != 'All causes']
deathcause_data.head(2)
```

Out[36]:

	Year	113 Cause Name	Cause Name	State	Deaths	Age-adjusted Death Rate
0	2012	Nephritis, nephrotic syndrome and nephrosis (N...	Kidney disease	Vermont	21	2.6
1	2016	Nephritis, nephrotic syndrome and nephrosis (N...	Kidney disease	Vermont	30	3.7

In [38]:
```
#read excel with census data; notice that the data available is from 2010 to 2018
population_data=pd.read_excel('C:/Users/rutvi/Downloads/nst-est2018-01.xlsx',header=3,index_col=0)
population_data.head(2)
```

Out[38]:

	Census	Estimates Base	2010	2011	2012	2013	
United States	308745538.0	308758105.0	309326085.0	311580009.0	313874218.0	316057727.0	31
Northeast	55317240.0	55318430.0	55380645.0	55600532.0	55776729.0	55907823.0	5

In [39]:
```
cols_year_deaths=[0,4]
yearly_deaths = deathcause_data[deathcause_data.columns[cols_year_deaths]]
#notice that data available is from 1999 to 2016
grouped_years = yearly_deaths.groupby('Year').sum().T
#Total deaths per year
grouped_years
```

Out[39]:

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007
Deaths	1905826	1902194	1899358	1918873	1912115	1864133	1889981	1854676	1846301

In [40]:
```
#common years from both tables
available_year_data = list(grouped_years.columns.intersection(population_dat
a.columns))
available_year_data
```

Out[40]:
```
[2010, 2011, 2012, 2013, 2014, 2015, 2016]
```

In [41]:
```
#In the generic form, mortality rates are calculated as: deaths/population*1
0^3  ,for per 1000 population
death_rate=(grouped_years[available_year_data]/population_data[available_yea
r_data].loc['United States'])*1000
print("Deaths per 1000 Americans has not drastically changed over the years
 2010-2016 and is as follows:")
death_rate
```

```
Deaths per 1000 Americans has not drastically changed over the years 2010-2
016 and is as follows:
```

Out[41]:

Year	2010	2011	2012	2013	2014	2015	2016
Deaths	5.988338	5.99949	5.97879	6.044184	6.088224	6.276112	6.296191

Analysis:

In the death trend above from year 2010 to 2016 death rate has remained steady. 5 to 6 people have died among every 1000 Americans. Hence Americans are facing a steady likelihood of death.

What are the four leading causes of death for Americans?

In [42]:
```
cols_cause_deaths=[2,4]
cause_deaths = deathcause_data[deathcause_data.columns[cols_cause_deaths]]
grouped_cause = cause_deaths.groupby('Cause Name').sum()
#Listing top for Cause Names with maximum death counts
four_leading_causes=grouped_cause.sort_values(by=['Deaths'],ascending=False)
.iloc[0:4]
print('Four leading causes of death for Americans: \n')
four_leading_causes.reset_index()
```

```
Four leading causes of death for Americans:
```

Out[42]:

	Cause Name	Deaths
0	Heart disease	11575183
1	Cancer	10244536
2	Stroke	2580140
3	CLRD	2434726

Analysis:

From the above result it is evident that the top 4 causes of death are:

1) Heart disease

2) Cancer

3) Stroke

4) CLRD

Do individual states show the same four leading causes of death?

In [55]:
```
#notice that data available is from 1999 to 2016
grouped_state = deathcause_data[['State','Cause Name','Deaths']].groupby(['State','Cause Name']).aggregate(sum).sort_values(by=['Deaths'],ascending=False).reset_index('State')
state_four_leading_causes=grouped_state.groupby('State').apply(lambda x:x.iloc[0:4]).reset_index('Cause Name')
#The Frequency counts of contents in the Cause Name column display the same 4 Causes
print('===> States having a cause in the top 4\n\n',state_four_leading_causes['Cause Name'].value_counts(),'\n\n===>Individual states :\n')
#Get names of individual states with causes in top 4
state_names_four_lead = state_four_leading_causes.reset_index(drop='State')
state_names_four_lead = state_names_four_lead.groupby('Cause Name').apply(lambda x: x['State'].unique())
print(state_names_four_lead)
```

```
===> States having a cause in the top 4

 Heart disease             51
Cancer                     51
Stroke                     42
CLRD                       39
Unintentional injuries     20
Alzheimer's disease         1
Name: Cause Name, dtype: int64

===>Individual states :

Cause Name
Alzheimer's disease                                          [North Dakota]
CLRD                       [Alabama, Arizona, Arkansas, California, Color...
Cancer                     [Alabama, Alaska, Arizona, Arkansas, Californi...
Heart disease              [Alabama, Alaska, Arizona, Arkansas, Californi...
Stroke                     [Alabama, Alaska, Arkansas, California, Connec...
Unintentional injuries     [Alaska, Arizona, Colorado, District of Columb...
dtype: object
```

Analysis:

===>

1) 51 states have Cancer and Heart Disease in the top four causes.

2) 42 states have Stroke in the top four causes.

3) 39 states have CLRD in the top four causes.

4) 20 states have Unintentional injuries in the top four causes.

5) 1 state has Alzheimer's disease in the top four causes.

Hence individual states do not show the same four leading causes of death as compared to all states together.

===>The second part displays the list of the top four causes in the individual states and confirms the reason presented in the first part of the result. The causes are as follows:

1) Alzheimer's disease

2) CLRD

3) Cancer

4) Heart disease

5) Stroke

6) Unintentional injuries

Hence individual states do not show the same four leading causes of death as compared to all states together.

Are there year-by-year changes in the four leading causes of death nationwide?

In [57]:
```
grouped_year = deathcause_data[['Year','Cause Name','Deaths']].groupby(['Yea
r','Cause Name']).aggregate(sum).sort_values(by=['Deaths'],ascending=False).
reset_index('Year')
year_four_leading_causes = grouped_year.groupby('Year').apply(lambda x:x.ilo
c[0:4]).reset_index('Cause Name')
lead_causes_year = year_four_leading_causes['Cause Name'].value_counts()
year_count = grouped_year['Year'].unique()
print('Total number of years for which data is available:',len(year_count))
print('Number of years in which Causes are present in top four:\n', lead_cau
ses_year)
#Calculating Year by Year Change
group_lead_cause = year_four_leading_causes.reset_index(drop='Year')
group_lead_cause = group_lead_cause.groupby('Cause Name').apply(lambda x: x[
'Year'].unique())
print('\n\nCauses in Years:\n',group_lead_cause)
```

```
Total number of years for which data is available: 18
Number of years in which Causes are present in top four:
 CLRD                    18
Cancer                   18
Heart disease            18
Stroke                   14
Unintentional injuries    4
Name: Cause Name, dtype: int64


Causes in Years:
 Cause Name
CLRD                    [1999, 2000, 2001, 2002, 2003, 2004, 2005, 200...
Cancer                  [1999, 2000, 2001, 2002, 2003, 2004, 2005, 200...
Heart disease           [1999, 2000, 2001, 2002, 2003, 2004, 2005, 200...
Stroke                  [1999, 2000, 2001, 2002, 2003, 2004, 2005, 200...
Unintentional injuries                          [2013, 2014, 2015, 2016]
dtype: object
```

Analysis:

In the past 18 year trend from 1999 to 2016, Heart Disease, Cancer and CLRD have remained in the top 4 spots, however Stroke rates have gone down and unintentional injury rates have risen in the years 2013 to 2016. Which means that Stroke was a part of the top four causes until 2012 and year 2013 onwards, unintentional injuries have increased and replaced Stroke in the top four casues of deaths. Until 2012, there were no changes in the top four causes of death and 2013 onwards Stroke was replaced in the top four by Unintentonal injuries. Hence yes, there are year by year changes in the four leading causes of death nationwide.