Dana DiVincenzo BANA.680.01

Assignment 2: Pandas Data Management

Introduction

In this assignment, I was given a two data files - one on the leading causes of death in the United States, and one on the population of the United States. I demonstrate my Pandas data management skills by answering 4 questions which require manipulation and clensing off the supplied data files.

```
In [88]: import pandas as pd
```

```
In [89]: # before reading in files, I slightly edited their headers in Excel
         # read in files
         file = 'C:/Users/Owner/Downloads/USPop.csv'
         popdf = pd.read_csv(file)

         file = 'C:/Users/Owner/Downloads/Deaths.csv'
         deathsdf = pd.read_csv(file)
```

Question 1: Are Americans facing increasing, decreasing, or steady likelihood of death?

```
In [83]: # get data frame where state is "United States", cause name is "All causes",
         # then sort by "Year"
         # this data frame only displays the year and deaths columns
         totaldeathsdf = deathsdf.loc[(deathsdf["State"] == "United States") & \
                                      (deathsdf["Cause Name"] == "All causes"), \
                                     ["Year", "Deaths"]]
         totaldeathsdf = totaldeathsdf.sort_values("Year")

         # slice the data frame to be only from 2010 to 2016, then reset the indices
         totaldeathsinrangedf = totaldeathsdf.loc[totaldeathsdf["Year"].isin(range(2010
         , 2017))]
         totaldeathsinrangedf.reset_index(drop=True, inplace=True)

         # get populations from where the "Geographic Area" is "United States" from the
         years 2010 to 2016
         totalpopdf = popdf.loc[popdf["Geographic Area"] == "United States", "2010":"20
         16"]
```

In [84]:

```
# remove commas from population values
totalpopdf = totalpopdf.replace(",", "", regex=True)

# transpose the population data frame so it can be added as a columnn to total
deathsinrangedf
transpopdf = totalpopdf.transpose()

# reset the indices transpopdf
transpopdf.reset_index(drop=True, inplace=True)

# create a combined data frame that shows population and death data from all c
auses for
# the United States in the years 2010 to 2016
combined = totaldeathsinrangedf.join(transpopdf)
combined.columns.values[2]="United States Population"
# create an empty data frame to fill later on with the United States death rat
es from 2010 to 2016
question1df = pd.DataFrame(columns=["Year", "Deaths", "United States Populatio
n", "Death Rate"])

# calculates the death rate for each year from 2010 to 2016
for index, row in combined.iterrows():
    deathrate = (row["Deaths"] / int(row["United States Population"]))
    year = row["Year"]
    deaths = row["Deaths"]
    population = row["United States Population"]
    # populates a row in the dataframe "question1"
    question1df.loc[index] = [str(year), str(deaths), str(population), deathra
te]

# displays the death rate in the United States from 2010 to 2016
question1dfstyler = question1.style.set_properties(**{"text-align":"left"})
question1dfstyler.set_table_styles([dict(selector='th', props=[('text-align',
'left')])])
display(question1dfstyler.hide_index())
```

Year	Deaths	United States Population	Death Rate
2010	2468435	309326085	0.007980
2011	2515458	311580009	0.008073
2012	2543279	313874218	0.008103
2013	2596993	316057727	0.008217
2014	2626418	318386421	0.008249
2015	2712630	320742673	0.008457
2016	2744248	323071342	0.008494

The table above shows the death rates in the United States from the years 2010 to 2016. This was calculated by taking the total number of deaths for each year and dividing that by the United State's population of that year. Looking at the Death Rate column, it is obvious that the death rate is increasing. Therefore, Americans are facing increasing likelyhood of death.

Question 2: What are the four leading causes of death for Americans?

In [90]:

```
# get data frame where state is "United States", cause name is not "All cause
s, and
# the year is 2016. This data frame only displays the cause name and deaths co
lumn
USdeathsdf = deathsdf.loc[(deathsdf["State"] == "United States") &\
                        (deathsdf["Cause Name"] != "All causes") &\
                        (deathsdf["Year"] == int("2016")), ["Cause Name", "Dea
ths"]]

# sorts the data frame by deaths and resets the indices
USdeathsdf = USdeathsdf.sort_values("Deaths", ascending = False)
USdeathsdf.reset_index(drop=True, inplace=True)

# displays the top four causes of deaths in the United States in 2016
USdeathsdf.index+=1
USdeathsdf = USdeathsdf.head(4)
USdeathsdfstyler = USdeathsdf.style.set_properties(**{"text-align":"left"})
USdeathsdfstyler.set_table_styles([dict(selector='th', props=[('text-align',
'left')])])
display(USdeathsdfstyler)
```

	Cause Name	Deaths
1	Heart disease	635260
2	Cancer	598038
3	Unintentional injuries	161374
4	CLRD	154596

The table above shows the top four causes of death in the United States in 2016. This was found after sorting the leading causes of deaths data frame to find only the data on deaths for the United States from individual causes. It was then sorted from highest to lowest and the top 4 were displayed. Therefore, in 2016, the top 4 causes of death for Americans are Heart Disease, Cancer, Unintentional Injuries, and CLRD, in that order.

Question 3: Do individual states show the same four leading causes of death?

In [86]:

```
# get data frame where state is not "United States", cause name is not "All ca
uses", the year is 2016
# this data frame only displays the state, cause name, and deaths columns
statedeathsdf = deathsdf.loc[(deathsdf["State"] != "United States") &\
                        (deathsdf["Cause Name"] != "All causes") &\
                        (deathsdf["Year"] == int("2016")), ["State", "Cause Na
me", "Deaths"]]

# sorts the data frame by state name and resets the indices
statedeathsdf = statedeathsdf.sort_values("State", ascending = True)
statedeathsdf.reset_index(drop=True, inplace=True)

# create an empty data frame to fill later on with the top 4 causes of death f
rom every individual state
question3df = pd.DataFrame(columns=["State", "Cause 1", "Cause 2", "Cause 3",
"Cause 4"])

# creates an index to track the row in question3df
idx = 0

# for each state, get the 4 leading causes of death and add them to question3d
f
for state in statedeathsdf["State"].unique():
    # splices out just the individual state
    statedf = statedeathsdf.loc[statedeathsdf["State"] == state]
    # gets the four highest deaths for this state
    statedf = statedf.nlargest(4, "Deaths")
    # creates an empty list to fill with four leading causes of death for this
state
    listofcauses = []
    # iterates over each cause for the state
    for index, row in statedf.iterrows():
        # gets the cause name and append it to the list of causes of death
        causename = row["Cause Name"]
        listofcauses.append(causename)
    # adds a new row to question3df filled with the state and top four causes
 of death
    question3df.loc[idx] = [state, *listofcauses]
    idx += 1

# displays the top four causes of death for every state in 2016
question3styler = question3df.style.set_properties(**{"text-align":"left"})
question3styler.set_table_styles([dict(selector='th', props=[('text-align', 'l
eft')])])
display(question3styler.hide_index())
```

State	Cause 1	Cause 2	Cause 3	Cause 4
Alabama	Heart disease	Cancer	CLRD	Stroke
Alaska	Cancer	Heart disease	Unintentional injuries	CLRD
Arizona	Heart disease	Cancer	Unintentional injuries	CLRD
Arkansas	Heart disease	Cancer	CLRD	Stroke
California	Heart disease	Cancer	Stroke	Alzheimer's disease
Colorado	Cancer	Heart disease	Unintentional injuries	CLRD
Connecticut	Heart disease	Cancer	Unintentional injuries	CLRD
Delaware	Cancer	Heart disease	CLRD	Unintentional injuries
District of Columbia	Heart disease	Cancer	Unintentional injuries	Stroke
Florida	Heart disease	Cancer	Unintentional injuries	CLRD
Georgia	Heart disease	Cancer	CLRD	Unintentional injuries
Hawaii	Heart disease	Cancer	Stroke	Unintentional injuries
Idaho	Heart disease	Cancer	CLRD	Unintentional injuries
Illinois	Heart disease	Cancer	Stroke	CLRD
Indiana	Heart disease	Cancer	CLRD	Unintentional injuries
Iowa	Heart disease	Cancer	CLRD	Unintentional injuries
Kansas	Heart disease	Cancer	CLRD	Unintentional injuries
Kentucky	Heart disease	Cancer	CLRD	Unintentional injuries
Louisiana	Heart disease	Cancer	Unintentional injuries	Stroke
Maine	Cancer	Heart disease	CLRD	Unintentional injuries
Maryland	Heart disease	Cancer	Stroke	Unintentional injuries
Massachusetts	Cancer	Heart disease	Unintentional injuries	CLRD
Michigan	Heart disease	Cancer	CLRD	Unintentional injuries
Minnesota	Cancer	Heart disease	Unintentional injuries	CLRD
Mississippi	Heart disease	Cancer	CLRD	Unintentional injuries
Missouri	Heart disease	Cancer	CLRD	Unintentional injuries
Montana	Heart disease	Cancer	CLRD	Unintentional injuries
Nebraska	Cancer	Heart disease	CLRD	Stroke
Nevada	Heart disease	Cancer	CLRD	Unintentional injuries
New Hampshire	Cancer	Heart disease	Unintentional injuries	CLRD
New Jersey	Heart disease	Cancer	Unintentional injuries	Stroke
New Mexico	Heart disease	Cancer	Unintentional injuries	CLRD
New York	Heart disease	Cancer	Unintentional injuries	CLRD
North Carolina	Cancer	Heart disease	Unintentional injuries	CLRD
North Dakota	Heart disease	Cancer	Unintentional injuries	Alzheimer's disease

State	Cause 1	Cause 2	Cause 3	Cause 4
Ohio	Heart disease	Cancer	Unintentional injuries	CLRD
Oklahoma	Heart disease	Cancer	CLRD	Unintentional injuries
Oregon	Cancer	Heart disease	Unintentional injuries	CLRD
Pennsylvania	Heart disease	Cancer	Unintentional injuries	Stroke
Rhode Island	Heart disease	Cancer	Unintentional injuries	CLRD
South Carolina	Cancer	Heart disease	Unintentional injuries	CLRD
South Dakota	Heart disease	Cancer	Unintentional injuries	Alzheimer's disease
Tennessee	Heart disease	Cancer	CLRD	Unintentional injuries
Texas	Heart disease	Cancer	Stroke	Unintentional injuries
Utah	Heart disease	Cancer	Unintentional injuries	Stroke
Vermont	Heart disease	Cancer	Unintentional injuries	CLRD
Virginia	Cancer	Heart disease	Unintentional injuries	Stroke
Washington	Cancer	Heart disease	Alzheimer's disease	Unintentional injuries
West Virginia	Heart disease	Cancer	Unintentional injuries	CLRD
Wisconsin	Heart disease	Cancer	Unintentional injuries	CLRD
Wyoming	Heart disease	Cancer	Unintentional injuries	CLRD

The table above shows the leading 4 causes of death in 2016 for each state. As shown, the states do not all have the same top 4 causes of death in the same order, but are similar. Heart disease or cancer is always the number 1 cause of death in every state, and whichever cause (heart disease or cancer) is not the number 1 cause is the number 2 cause. The number 3 and 4 causes of death in every state are either unintentional injuries, CLRD, stroke, or Alzheimer's disease.

Question 4: Are there year-by-year changes in the four leading causes of death nationwide?

In [87]:

```
# creates an empty data frame to fill later on with the top 4 causes of death in the US for every year
question4df = pd.DataFrame(columns=["Year", "Cause 1", "Cause 2", "Cause 3", "Cause 4"])

# creates an index to track the row in question4df
idx = 0

# for each year, get the 4 leading causes of death nationwide and add them to question4df
for year in range(2010, 2017):
    # get data frame where state is "United States", cause name is not "All causes"
    # this data frame only displays the cause name, and deaths columns, for every year
    leadingdeathsdf = deathsdf.loc[(deathsdf["State"] == "United States") &\
                            (deathsdf["Cause Name"] != "All causes") &\
                            (deathsdf["Year"] == year), ["Cause Name", "Deaths"]]
    # sorts data frame by deaths, highest to lowest, and only shows top 4 causes
    leadingdeathsdf = leadingdeathsdf.sort_values("Deaths", ascending = False)
    leadingdeathsdf = leadingdeathsdf.head(4)
    # resets the indices
    leadingdeathsdf.reset_index(drop=True, inplace=True)
    # creates an empty list to fill with four leading causes of death
    listofcauses = []
    # get the cause name and append it to the list of causes
    for index, row in leadingdeathsdf.iterrows():
        causename = row["Cause Name"]
        listofcauses.append(causename)
    # add a new row to question4df filled with the year and top four causes
    question4df.loc[idx] = [year, *listofcauses]
    idx += 1

# displays the top four causes of death nationwide from 2010 to 2016
question4styler = question4df.style.set_properties(**{"text-align":"left"})
question4styler.set_table_styles([dict(selector='th', props=[('text-align', 'left')])])
display(question4styler.hide_index())
```

Year	Cause 1	Cause 2	Cause 3	Cause 4
2010	Heart disease	Cancer	CLRD	Stroke
2011	Heart disease	Cancer	CLRD	Stroke
2012	Heart disease	Cancer	CLRD	Stroke
2013	Heart disease	Cancer	CLRD	Unintentional injuries
2014	Heart disease	Cancer	CLRD	Unintentional injuries
2015	Heart disease	Cancer	CLRD	Unintentional injuries
2016	Heart disease	Cancer	Unintentional injuries	CLRD

This table shows the top 4 causes of death from 2010 to 2016 nationwide, and demonstrates the year-by-year changes. As shown, the number 1 cause of death in America is always heart disease, and the number 2 cause of death is cancer. The number 3 cause of death was consistently CLRD, until 2016 when it became unintentional injuries. The number 4 cause of death is either stroke, unintentional injuries, or CLRD.

Conclusion

To summarize this assignment, I was given two data files about US populations and causes of deaths, and was asked specific questions about the data. Using Pandas data manipulation functions such as .loc, .join, pd.DataFrame, and others, I was able to correctly clean and analyze the data to find the answers to the questions. I also improved my reporting skills using Jupyter Notebook by finding information online about formatting tables, which made them easier to read and more aesthetically pleasing.