Project Topic
Analysis of leading causes of death in the United States from 1999 to 2017

Project Definition:

- This project analyzes deaths across the United States and on a statewide level (mostly New Jersey because that is where our university is located) from the years 1999 to 2017 (18 year period). The project focuses on the ten leading causes of death (but mostly the top two leading causes) and analyzes differences in deaths for each cause.
- The problem we aim to solve is how we can use this data to hopefully address the reason behind some of the causes of deaths and how we can utilize this information to prevent, reduce, or mitigate deaths from such causes across the country.
- Some questions that the project answers include what is the leading cause of death for each state, what is the leading cause of death per year, the total number of deaths per year, if the number of deaths have increased or decreased per year, if there are any patterns in these trends over the 18 year period, and why some states experience an increase in deaths for a specific cause while others experience a decrease.
- This project is relevant/relates to this course because it includes data collection, data cleaning, data analysis, data visualization, and predictive modeling. The data collected will be used to predict future deaths in the country as a whole and within NJ.

Novelty and Importance:

- Analyzing leading causes of death is important because we can use these findings to come up with actions that could help lower or prevent the risk of certain causes of deaths, as well as coming up with interventions for high risk populations. If one state experiences a decrease in deaths for a specific cause over time but another state experiences an increase, we can conduct research to find out why, and a state could learn to adopt policies or practices from another state that has experienced a decrease in deaths.
- Public health is very important and when many people die of the same causes, we want to know why and if there are any personal, environmental, or other factors at play. By finding the reason behind some of the leading causes of death, we can better come up with strategies and measures to reduce risks and hopefully prevent deaths from increasing in the future.

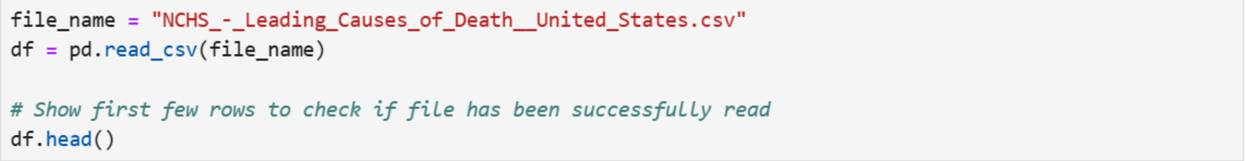
Progress and Contribution:

- Data was collected from the “NCHS - Leading Causes of Death: United States” csv file, which can be downloaded and retrieved at: https://healthdata.gov/dataset/NCHS-Leading-Causes-of-Death-United-States/nxxk-8p52/about_data
- Libraries used:
  - Data collection and cleaning: pandas
  - Data analysis: sqlite3, sqlalchemy
  - Data visualization: matplotlib (for bar and line graphs)
  - Predictive modeling: sklearn (for linear regression model)
- Some key findings/summary of results from the project:
  - The top two leading causes of death are heart disease and cancer, with heart disease being the leading cause of death every year from 1999 to 2017.
  - Deaths from heart disease have gone down from 1999 to 2012, but have increased since then.
  - Total number of deaths per year has increased.
  - Total number of deaths in NJ in 2017 is around the same as 1999

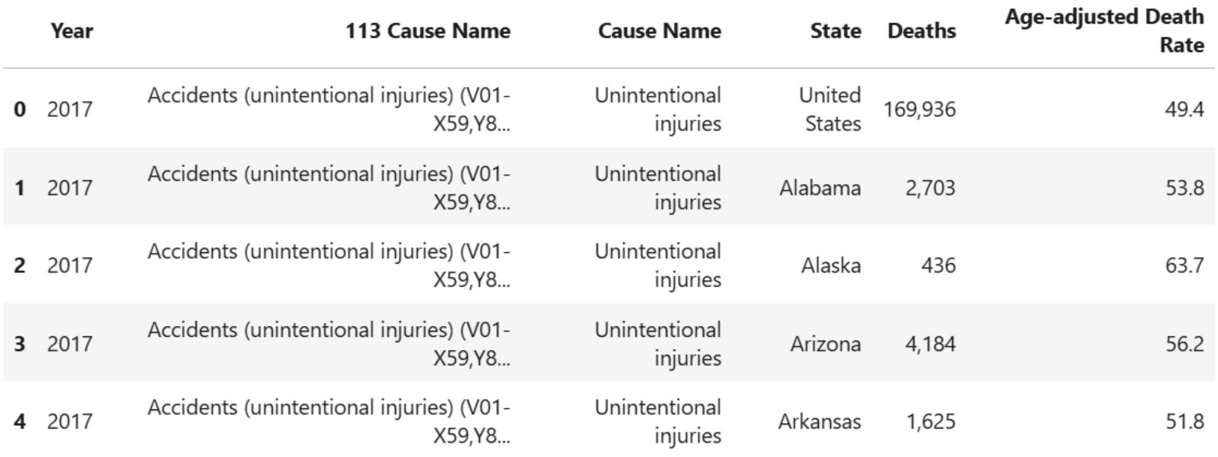
- Heart disease is the leading cause of death in NJ every year
- Deaths from heart disease and cancer in NJ have gone down over the years.

Implementation/Code

1. **Data Collection:** (Note: some code have been cropped out because they are too long, but full code can be seen on codebench/github)
    - Use pandas to load the csv file into a dataframe and show first few rows.

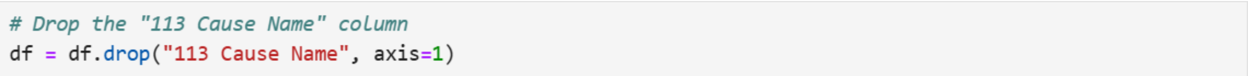


```
file_name = "NCHS_-_Leading_Causes_of_Death__United_States.csv"
df = pd.read_csv(file_name)

# Show first few rows to check if file has been successfully read
df.head()
```

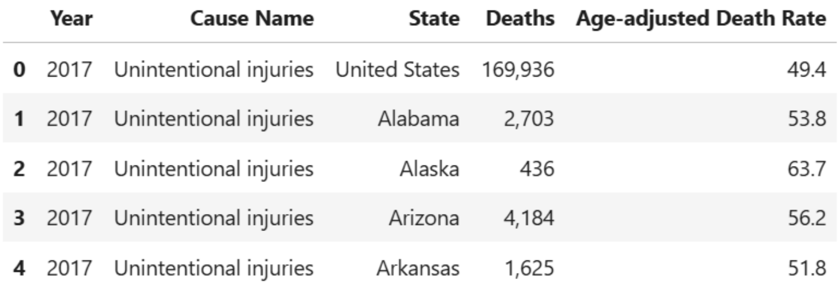


	Year	113 Cause Name	Cause Name	State	Deaths	Age-adjusted Death Rate
0	2017	Accidents (unintentional injuries) (V01-X59,Y8...	Unintentional injuries	United States	169,936	49.4
1	2017	Accidents (unintentional injuries) (V01-X59,Y8...	Unintentional injuries	Alabama	2,703	53.8
2	2017	Accidents (unintentional injuries) (V01-X59,Y8...	Unintentional injuries	Alaska	436	63.7
3	2017	Accidents (unintentional injuries) (V01-X59,Y8...	Unintentional injuries	Arizona	4,184	56.2
4	2017	Accidents (unintentional injuries) (V01-X59,Y8...	Unintentional injuries	Arkansas	1,625	51.8

2. **Data Cleaning:**
    - Remove the “113 Cause Names” column because it is essentially the same as the “Cause Names” column, making it redundant. Result shows that the column has been removed.

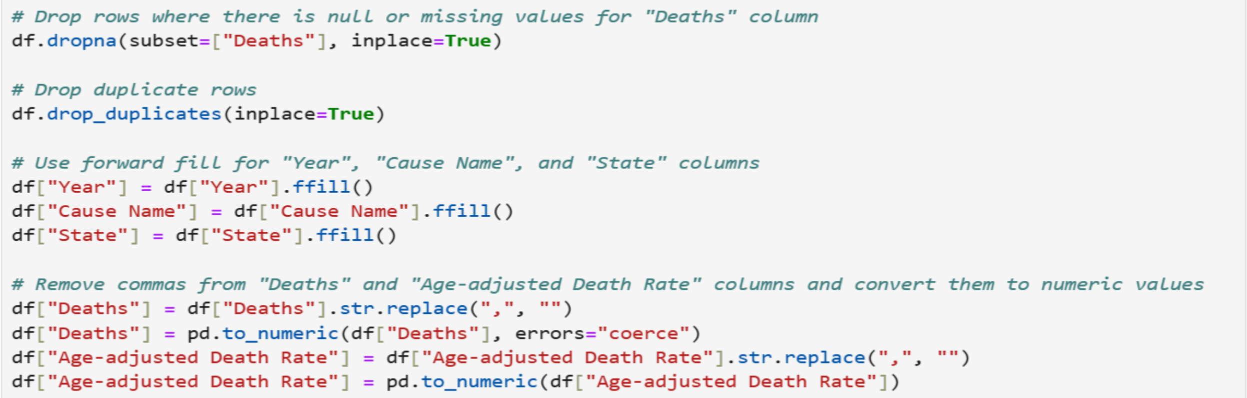


```
# Drop the "113 Cause Name" column
df = df.drop("113 Cause Name", axis=1)
```

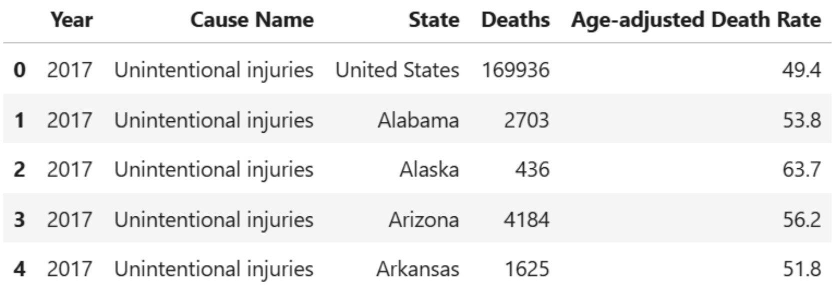


	Year	Cause Name	State	Deaths	Age-adjusted Death Rate
0	2017	Unintentional injuries	United States	169,936	49.4
1	2017	Unintentional injuries	Alabama	2,703	53.8
2	2017	Unintentional injuries	Alaska	436	63.7
3	2017	Unintentional injuries	Arizona	4,184	56.2
4	2017	Unintentional injuries	Arkansas	1,625	51.8

- Drop rows with missing values in “Deaths” column, drop duplicate rows, and convert values from “Deaths” and “Age-adjusted Death Rate” to numeric. Result shows that numeric values have been successfully converted

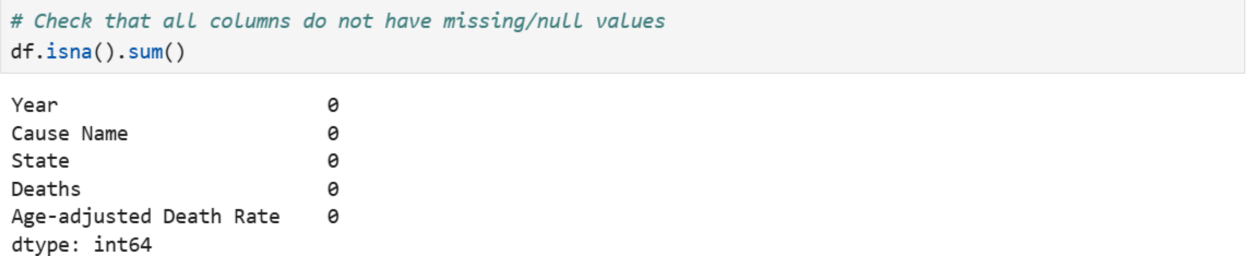


```
# Drop rows where there is null or missing values for "Deaths" column
df.dropna(subset=["Deaths"], inplace=True)

# Drop duplicate rows
df.drop_duplicates(inplace=True)

# Use forward fill for "Year", "Cause Name", and "State" columns
df["Year"] = df["Year"].ffill()
df["Cause Name"] = df["Cause Name"].ffill()
df["State"] = df["State"].ffill()

# Remove commas from "Deaths" and "Age-adjusted Death Rate" columns and convert them to numeric values
df["Deaths"] = df["Deaths"].str.replace(",", "")
df["Deaths"] = pd.to_numeric(df["Deaths"], errors="coerce")
df["Age-adjusted Death Rate"] = df["Age-adjusted Death Rate"].str.replace(",", "")
df["Age-adjusted Death Rate"] = pd.to_numeric(df["Age-adjusted Death Rate"])
```



	Year	Cause Name	State	Deaths	Age-adjusted Death Rate
0	2017	Unintentional injuries	United States	169936	49.4
1	2017	Unintentional injuries	Alabama	2703	53.8
2	2017	Unintentional injuries	Alaska	436	63.7
3	2017	Unintentional injuries	Arizona	4184	56.2
4	2017	Unintentional injuries	Arkansas	1625	51.8

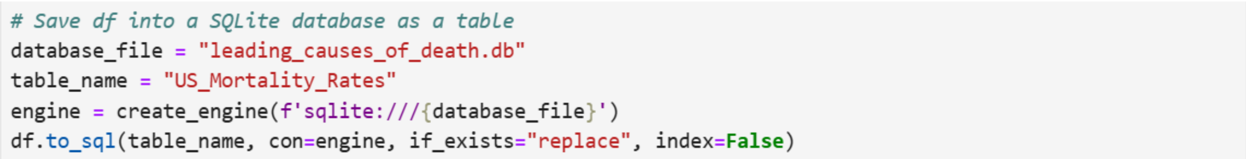
- Check for any null values in dataframe.



```
# Check that all columns do not have missing/null values
df.isna().sum()
```

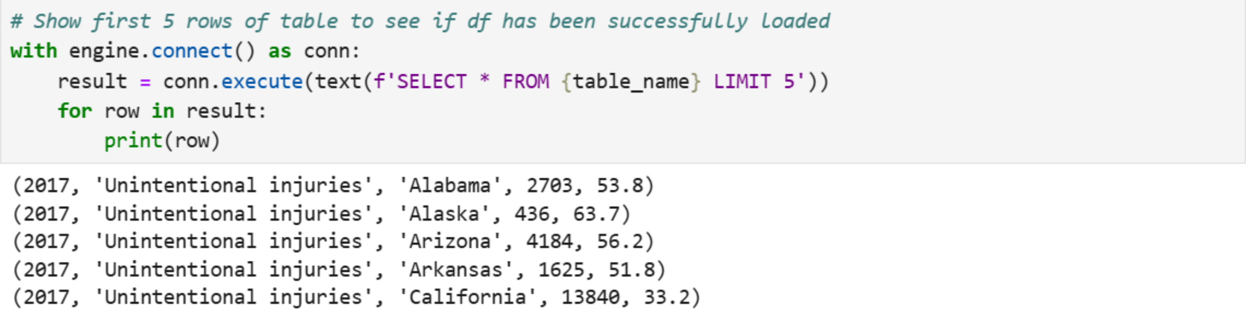
```
Year                       0
Cause Name                 0
State                      0
Deaths                     0
Age-adjusted Death Rate    0
dtype: int64
```

- Save results into a SQLite database as a table



```
# Save df into a SQLite database as a table
database_file = "leading_causes_of_death.db"
table_name = "US_Mortality_Rates"
engine = create_engine(f'sqlite:///{database_file}')
df.to_sql(table_name, con=engine, if_exists="replace", index=False)
```

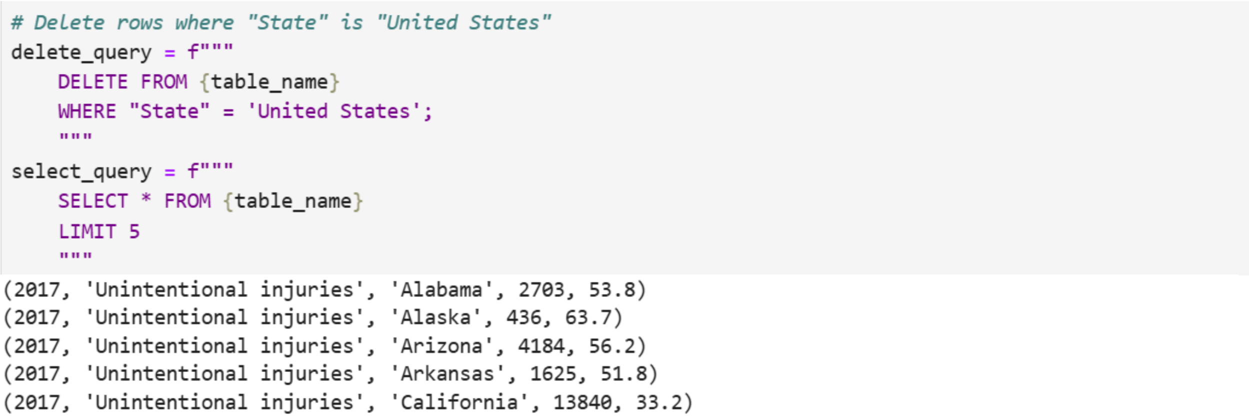
- Show first few rows to check if loading was successful



```
# Show first 5 rows of table to see if df has been successfully loaded
with engine.connect() as conn:
    result = conn.execute(text(f'SELECT * FROM {table_name} LIMIT 5'))
    for row in result:
        print(row)
```

```
(2017, 'Unintentional injuries', 'Alabama', 2703, 53.8)
(2017, 'Unintentional injuries', 'Alaska', 436, 63.7)
(2017, 'Unintentional injuries', 'Arizona', 4184, 56.2)
(2017, 'Unintentional injuries', 'Arkansas', 1625, 51.8)
(2017, 'Unintentional injuries', 'California', 13840, 33.2)
```

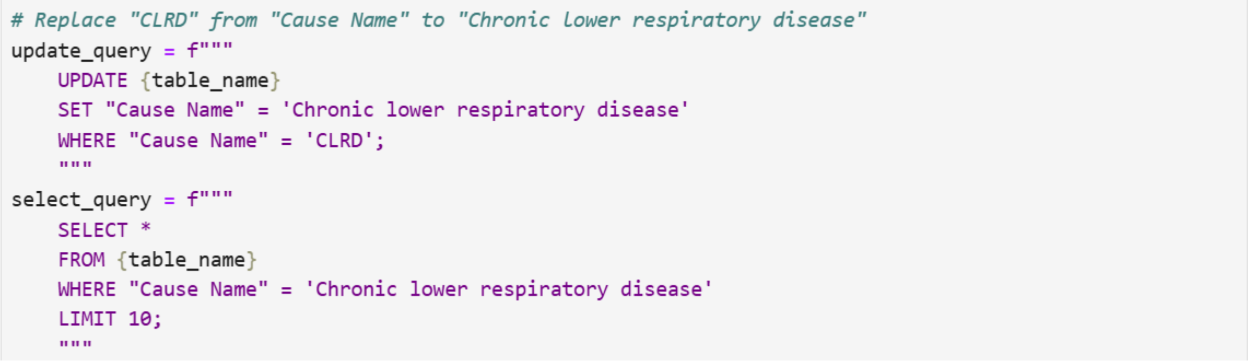
- Delete rows where the state name is United States because it will mess with calculations when we try to add up deaths from states. Result shows that US has been removed.



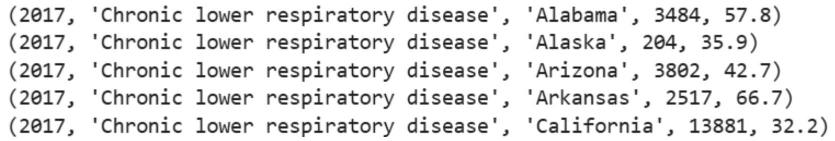
```
# Delete rows where "State" is "United States"
delete_query = f"""
    DELETE FROM {table_name}
    WHERE "State" = 'United States';
    """

select_query = f"""
    SELECT * FROM {table_name}
    LIMIT 5
    """
```

```
(2017, 'Unintentional injuries', 'Alabama', 2703, 53.8)
(2017, 'Unintentional injuries', 'Alaska', 436, 63.7)
(2017, 'Unintentional injuries', 'Arizona', 4184, 56.2)
(2017, 'Unintentional injuries', 'Arkansas', 1625, 51.8)
(2017, 'Unintentional injuries', 'California', 13840, 33.2)
```

- Replace “CLRD” to “Chronic lower respiratory disease” to make it easier to read. Result shows that the name has successfully been replaced.

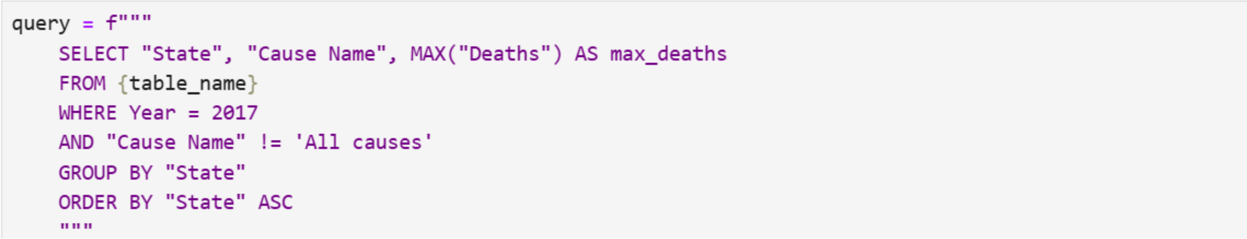


```
# Replace "CLRD" from "Cause Name" to "Chronic lower respiratory disease"
update_query = f"""
    UPDATE {table_name}
    SET "Cause Name" = 'Chronic lower respiratory disease'
    WHERE "Cause Name" = 'CLRD';
    """
select_query = f"""
    SELECT *
    FROM {table_name}
    WHERE "Cause Name" = 'Chronic lower respiratory disease'
    LIMIT 10;
    """
```

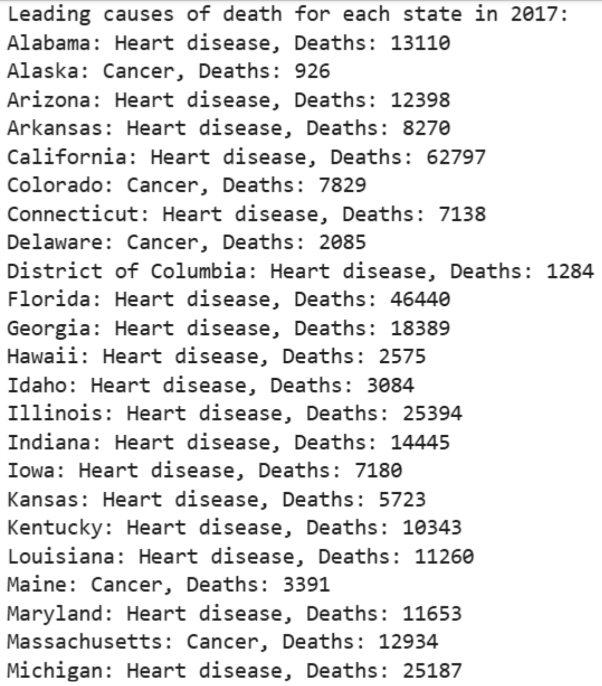


```
(2017, 'Chronic lower respiratory disease', 'Alabama', 3484, 57.8)
(2017, 'Chronic lower respiratory disease', 'Alaska', 204, 35.9)
(2017, 'Chronic lower respiratory disease', 'Arizona', 3802, 42.7)
(2017, 'Chronic lower respiratory disease', 'Arkansas', 2517, 66.7)
(2017, 'Chronic lower respiratory disease', 'California', 13881, 32.2)
```

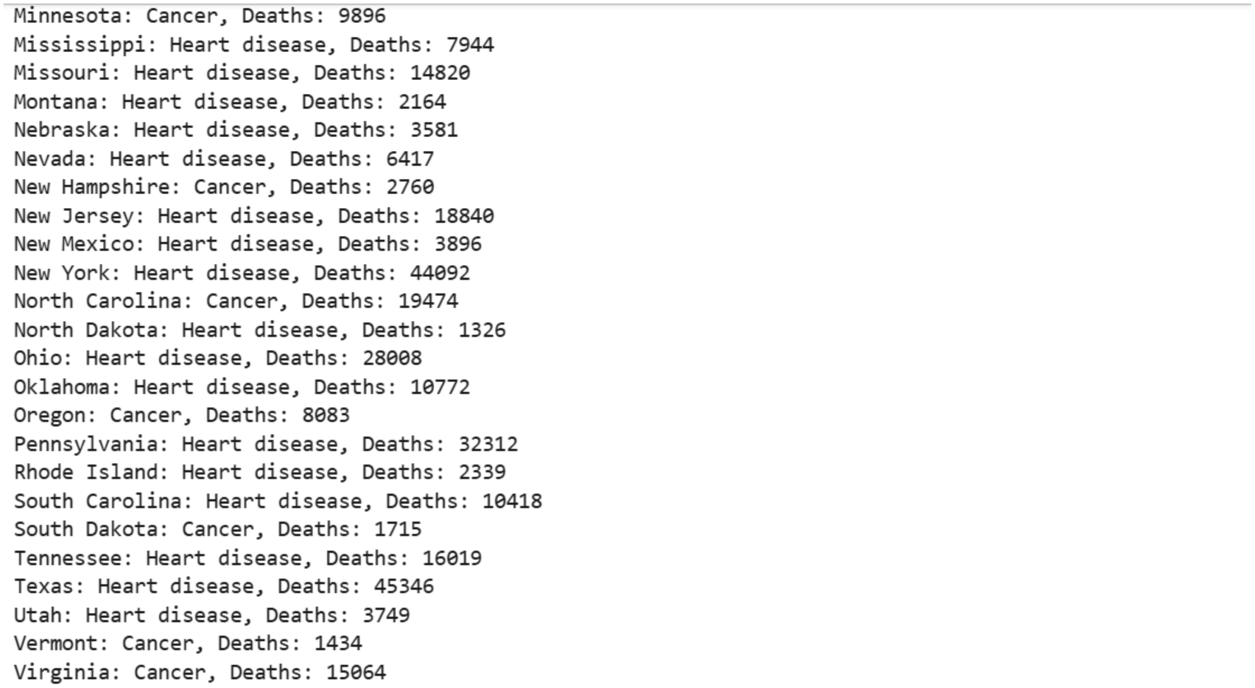
3. **Database/Analysis:** (again, the execution step of the code has been cropped)
   Starting with countrywide analysis:
   a. Query to find the leading cause of death for each state in 2017



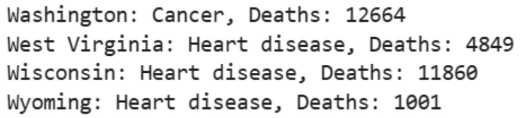
```
query = f"""
    SELECT "State", "Cause Name", MAX("Deaths") AS max_deaths
    FROM {table_name}
    WHERE Year = 2017
    AND "Cause Name" != 'All causes'
    GROUP BY "State"
    ORDER BY "State" ASC
    """
```



```
Leading causes of death for each state in 2017:
Alabama: Heart disease, Deaths: 13110
Alaska: Cancer, Deaths: 926
Arizona: Heart disease, Deaths: 12398
Arkansas: Heart disease, Deaths: 8270
California: Heart disease, Deaths: 62797
Colorado: Cancer, Deaths: 7829
Connecticut: Heart disease, Deaths: 7138
Delaware: Cancer, Deaths: 2085
District of Columbia: Heart disease, Deaths: 1284
Florida: Heart disease, Deaths: 46440
Georgia: Heart disease, Deaths: 18389
Hawaii: Heart disease, Deaths: 2575
Idaho: Heart disease, Deaths: 3084
Illinois: Heart disease, Deaths: 25394
Indiana: Heart disease, Deaths: 14445
Iowa: Heart disease, Deaths: 7180
Kansas: Heart disease, Deaths: 5723
Kentucky: Heart disease, Deaths: 10343
Louisiana: Heart disease, Deaths: 11260
Maine: Cancer, Deaths: 3391
Maryland: Heart disease, Deaths: 11653
Massachusetts: Cancer, Deaths: 12934
Michigan: Heart disease, Deaths: 25187
```

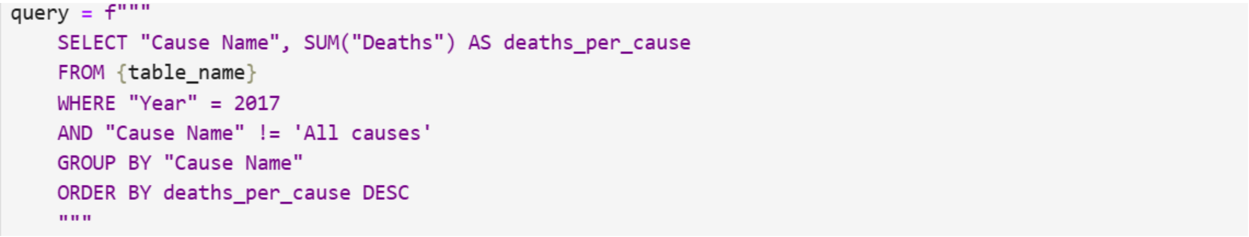


```
Minnesota: Cancer, Deaths: 9896
Mississippi: Heart disease, Deaths: 7944
Missouri: Heart disease, Deaths: 14820
Montana: Heart disease, Deaths: 2164
Nebraska: Heart disease, Deaths: 3581
Nevada: Heart disease, Deaths: 6417
New Hampshire: Cancer, Deaths: 2760
New Jersey: Heart disease, Deaths: 18840
New Mexico: Heart disease, Deaths: 3896
New York: Heart disease, Deaths: 44092
North Carolina: Cancer, Deaths: 19474
North Dakota: Heart disease, Deaths: 1326
Ohio: Heart disease, Deaths: 28008
Oklahoma: Heart disease, Deaths: 10772
Oregon: Cancer, Deaths: 8083
Pennsylvania: Heart disease, Deaths: 32312
Rhode Island: Heart disease, Deaths: 2339
South Carolina: Heart disease, Deaths: 10418
South Dakota: Cancer, Deaths: 1715
Tennessee: Heart disease, Deaths: 16019
Texas: Heart disease, Deaths: 45346
Utah: Heart disease, Deaths: 3749
Vermont: Cancer, Deaths: 1434
Virginia: Cancer, Deaths: 15064
```

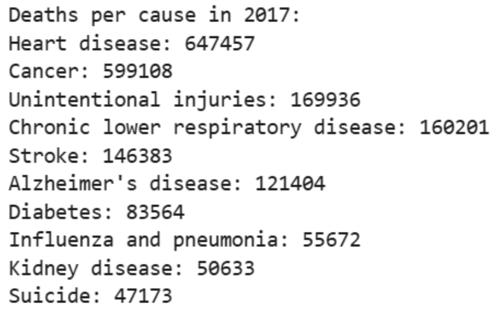


```
Washington: Cancer, Deaths: 12664
West Virginia: Heart disease, Deaths: 4849
Wisconsin: Heart disease, Deaths: 11860
Wyoming: Heart disease, Deaths: 1001
```

b. Query to find total number of deaths per cause in 2017

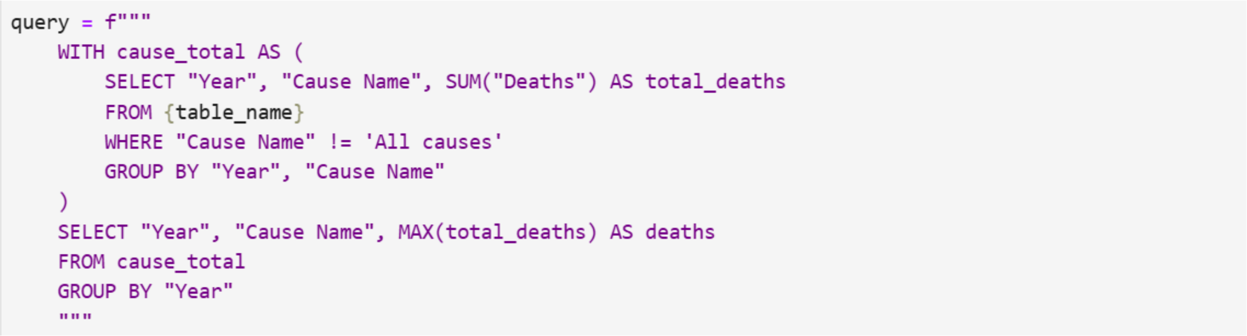


```
query = f"""
    SELECT "Cause Name", SUM("Deaths") AS deaths_per_cause
    FROM {table_name}
    WHERE "Year" = 2017
    AND "Cause Name" != 'All causes'
    GROUP BY "Cause Name"
    ORDER BY deaths_per_cause DESC
    """
```

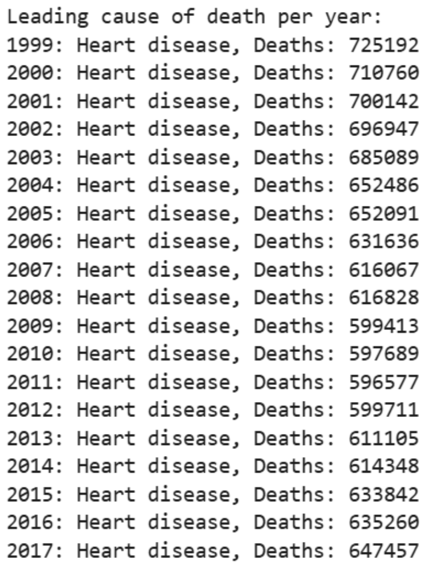


```
Deaths per cause in 2017:
Heart disease: 647457
Cancer: 599108
Unintentional injuries: 169936
Chronic lower respiratory disease: 160201
Stroke: 146383
Alzheimer's disease: 121404
Diabetes: 83564
Influenza and pneumonia: 55672
Kidney disease: 50633
Suicide: 47173
```

c. Query to find leading causes of death per year

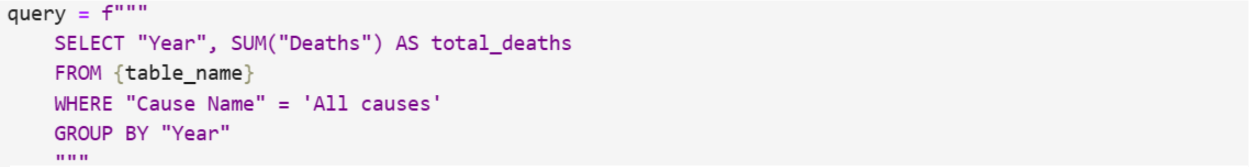


```
query = f"""
    WITH cause_total AS (
        SELECT "Year", "Cause Name", SUM("Deaths") AS total_deaths
        FROM {table_name}
        WHERE "Cause Name" != 'All causes'
        GROUP BY "Year", "Cause Name"
    )
    SELECT "Year", "Cause Name", MAX(total_deaths) AS deaths
    FROM cause_total
    GROUP BY "Year"
    """
```

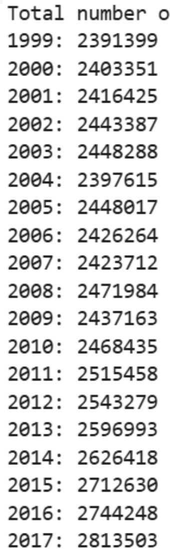


```
Leading cause of death per year:
1999: Heart disease, Deaths: 725192
2000: Heart disease, Deaths: 710760
2001: Heart disease, Deaths: 700142
2002: Heart disease, Deaths: 696947
2003: Heart disease, Deaths: 685089
2004: Heart disease, Deaths: 652486
2005: Heart disease, Deaths: 652091
2006: Heart disease, Deaths: 631636
2007: Heart disease, Deaths: 616067
2008: Heart disease, Deaths: 616828
2009: Heart disease, Deaths: 599413
2010: Heart disease, Deaths: 597689
2011: Heart disease, Deaths: 596577
2012: Heart disease, Deaths: 599711
2013: Heart disease, Deaths: 611105
2014: Heart disease, Deaths: 614348
2015: Heart disease, Deaths: 633842
2016: Heart disease, Deaths: 635260
2017: Heart disease, Deaths: 647457
```

d. Query to find total number of deaths per year



```
query = f"""
    SELECT "Year", SUM("Deaths") AS total_deaths
    FROM {table_name}
    WHERE "Cause Name" = 'All causes'
    GROUP BY "Year"
    """
```



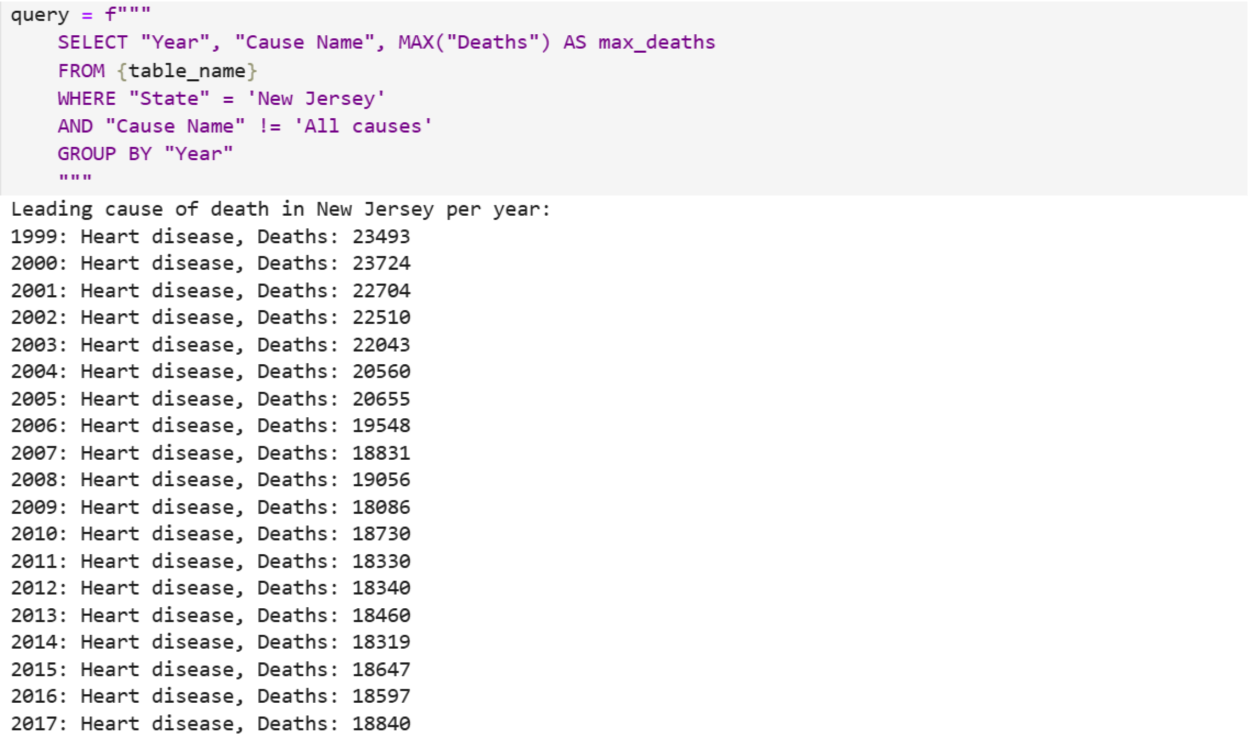
```
Total number of deaths per year:
1999: 2391399
2000: 2403351
2001: 2416425
2002: 2443387
2003: 2448288
2004: 2397615
2005: 2448017
2006: 2426264
2007: 2423712
2008: 2471984
2009: 2437163
2010: 2468435
2011: 2515458
2012: 2543279
2013: 2596993
2014: 2626418
2015: 2712630
2016: 2744248
2017: 2813503
```

- Based on data gathered from parts a, b, c, and d, the total number of deaths have increased over the years and the two top leading causes of death overall are heart disease and cancer. Research shows that heart disease is very common due to a combination of multiple factors, such as an unhealthy diet, being physically inactive, use of tobacco, consumption of alcohol, air pollution, high blood pressure and sugar, and being overweight and obese. Cancer in the US is also very common due to many of the same factors, along with other factors such as age, eating a western diet, having diabetes, and getting too little sleep. In addition to sharing many of the same risk factors, heart disease and cancer may be linked, with cardiovascular diseases increasing the likelihood of cancer and cancer treatments increasing the likelihood of heart problems.

- Sources:
    - https://www.who.int/health-topics/cardiovascular-diseases
    - https://www.cancercenter.com/community/blog/2023/01/why-are-cancer-rates-rising-in-adults-under-50
    - https://www.hackensackmeridianhealth.org/en/healthu/2021/03/22/how-heart-disease-may-be-linked-to-cancer
    - https://www.cancercenter.com/community/blog/2024/02/cancer-and-cardiovascular-disease

Moving on to state analysis (NJ):

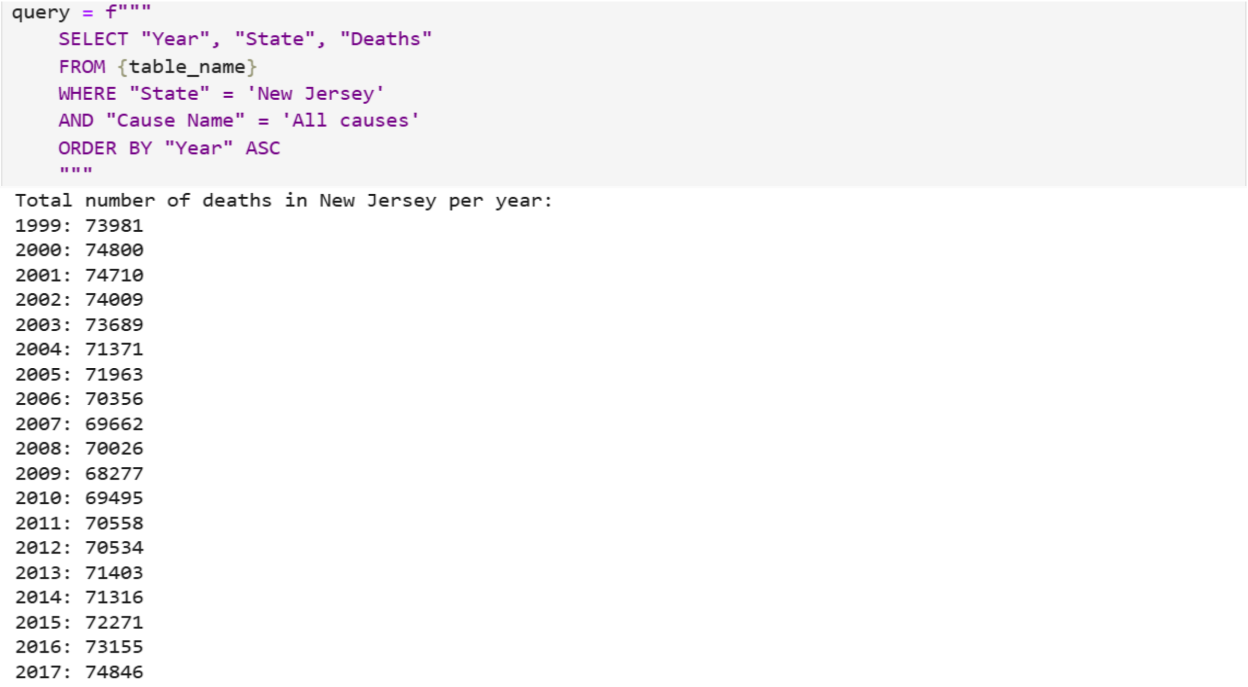
a. Query to find leading cause of death per year in NJ



```
query = f"""
    SELECT "Year", "Cause Name", MAX("Deaths") AS max_deaths
    FROM {table_name}
    WHERE "State" = 'New Jersey'
    AND "Cause Name" != 'All causes'
    GROUP BY "Year"
    """
```

```
Leading cause of death in New Jersey per year:
1999: Heart disease, Deaths: 23493
2000: Heart disease, Deaths: 23724
2001: Heart disease, Deaths: 22704
2002: Heart disease, Deaths: 22510
2003: Heart disease, Deaths: 22043
2004: Heart disease, Deaths: 20560
2005: Heart disease, Deaths: 20655
2006: Heart disease, Deaths: 19548
2007: Heart disease, Deaths: 18831
2008: Heart disease, Deaths: 19056
2009: Heart disease, Deaths: 18086
2010: Heart disease, Deaths: 18730
2011: Heart disease, Deaths: 18330
2012: Heart disease, Deaths: 18340
2013: Heart disease, Deaths: 18460
2014: Heart disease, Deaths: 18319
2015: Heart disease, Deaths: 18647
2016: Heart disease, Deaths: 18597
2017: Heart disease, Deaths: 18840
```

b. Query to find total number of deaths per year in NJ

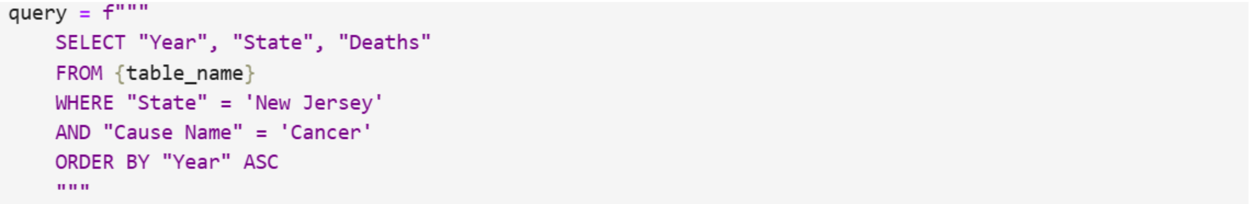


```
query = f"""
    SELECT "Year", "State", "Deaths"
    FROM {table_name}
    WHERE "State" = 'New Jersey'
    AND "Cause Name" = 'All causes'
    ORDER BY "Year" ASC
    """
```

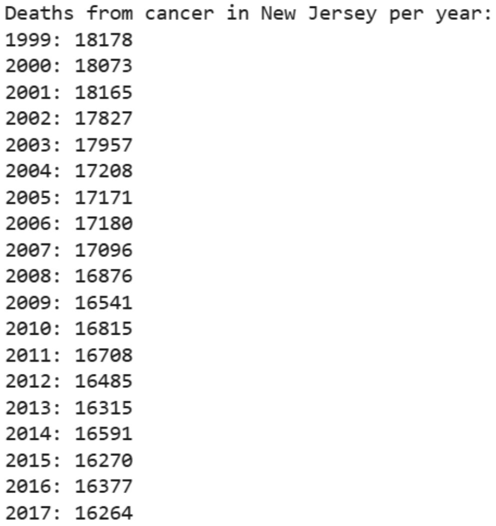
```
Total number of deaths in New Jersey per year:
1999: 73981
2000: 74800
2001: 74710
2002: 74009
2003: 73689
2004: 71371
2005: 71963
2006: 70356
2007: 69662
2008: 70026
2009: 68277
2010: 69495
2011: 70558
2012: 70534
2013: 71403
2014: 71316
2015: 72271
2016: 73155
2017: 74846
```

- From parts a, we have gathered that the leading cause of death in NJ per year is heart disease, but the deaths have decreased over the years. From part b, we can see that the total deaths in NJ have been relatively stable, with 2017 and 2019 being almost the same (difference of 865). So one question is then why did the total number of deaths in the US increase significantly from 1999 to 2017? Some guesses include more deaths from other causes in NJ, more deaths from other states, or more deaths due to an increasing population, which we will explore in parts c and d.

c. Query to find total deaths from cancer per year in NJ



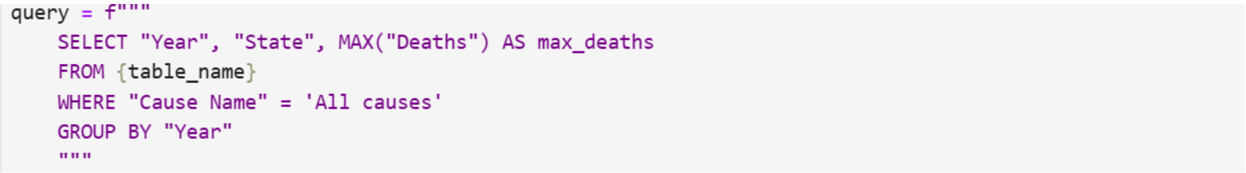
```
query = f"""
    SELECT "Year", "State", "Deaths"
    FROM {table_name}
    WHERE "State" = 'New Jersey'
    AND "Cause Name" = 'Cancer'
    ORDER BY "Year" ASC
    """
```



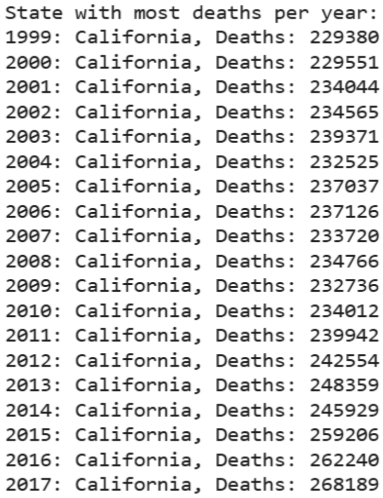
```
Deaths from cancer in New Jersey per year:
1999: 18178
2000: 18073
2001: 18165
2002: 17827
2003: 17957
2004: 17208
2005: 17171
2006: 17180
2007: 17096
2008: 16876
2009: 16541
2010: 16815
2011: 16708
2012: 16485
2013: 16315
2014: 16591
2015: 16270
2016: 16377
2017: 16264
```

- From part c, we can see that the total number of deaths from cancer, the next leading cause, has decreased over the years in NJ. So more deaths from other causes in NJ is unlikely to have caused an increase in deaths in the US.

d. Query to find the state with most deaths per year



```
query = f"""
    SELECT "Year", "State", MAX("Deaths") AS max_deaths
    FROM {table_name}
    WHERE "Cause Name" = 'All causes'
    GROUP BY "Year"
    """
```

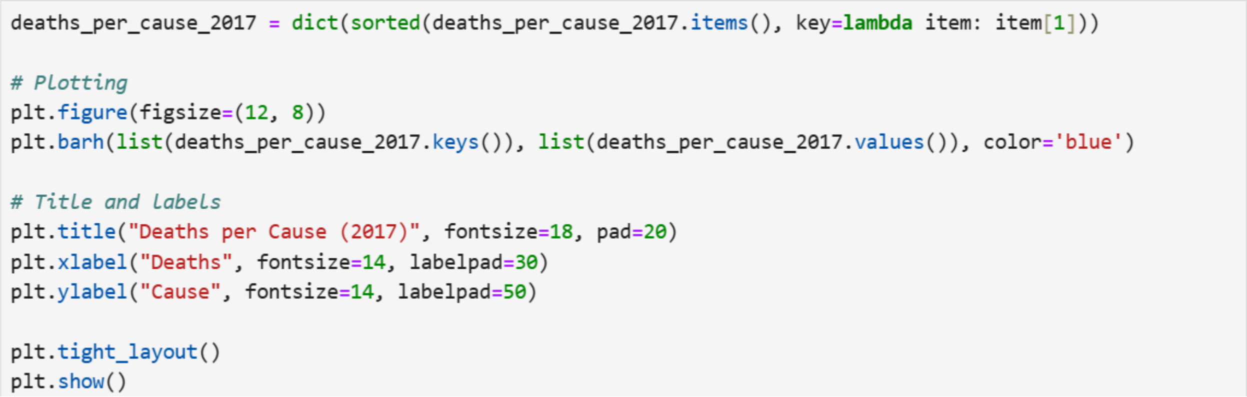


```
State with most deaths per year:
1999: California, Deaths: 229380
2000: California, Deaths: 229551
2001: California, Deaths: 234044
2002: California, Deaths: 234565
2003: California, Deaths: 239371
2004: California, Deaths: 232525
2005: California, Deaths: 237037
2006: California, Deaths: 237126
2007: California, Deaths: 233720
2008: California, Deaths: 234766
2009: California, Deaths: 232736
2010: California, Deaths: 234012
2011: California, Deaths: 239942
2012: California, Deaths: 242554
2013: California, Deaths: 248359
2014: California, Deaths: 245929
2015: California, Deaths: 259206
2016: California, Deaths: 262240
2017: California, Deaths: 268189
```

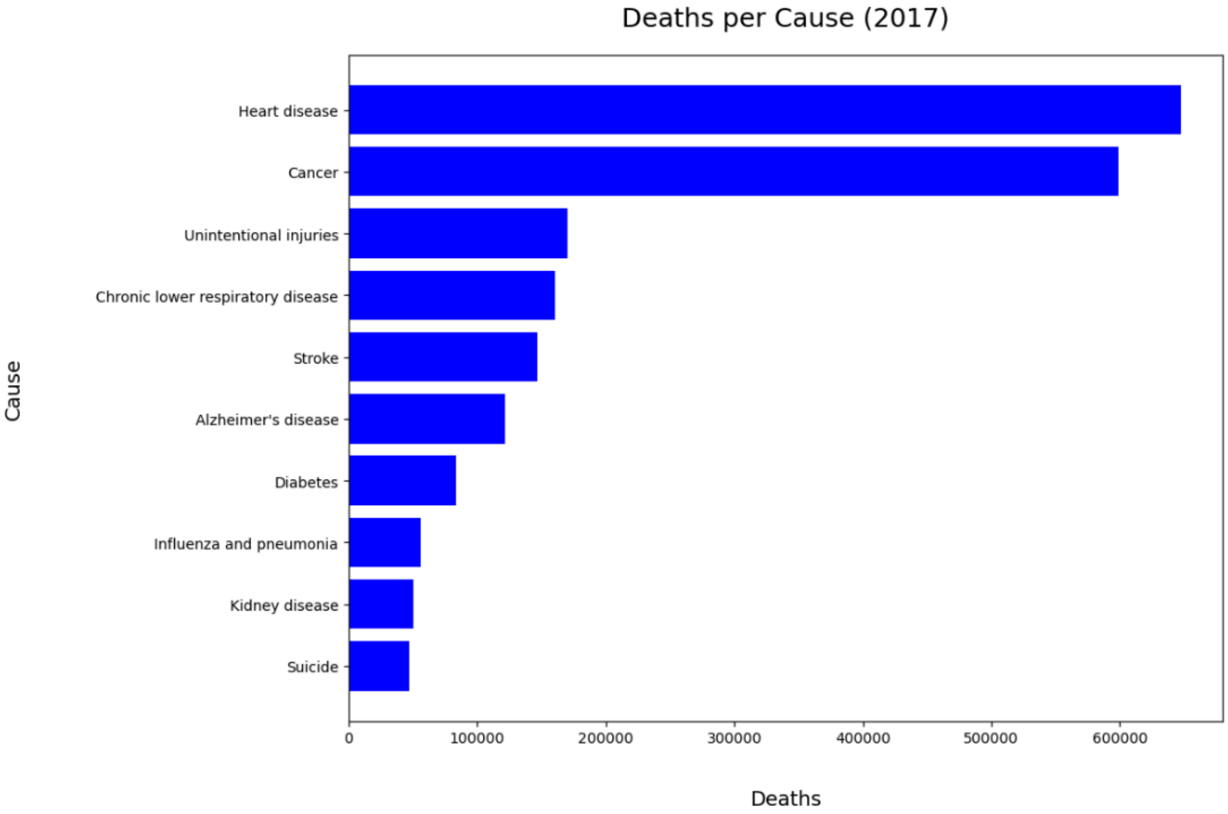
- From part d, we find out that the state with most deaths per year is California, and deaths have been increasing over the years. So even though deaths in NJ haven't increased much since 1999 and deaths from heart disease and cancer in NJ have actually gone down, deaths across the country are going up most likely due to an increase in deaths from some other states, like Cali, and an increase in the human population across the country as a whole.

4. Data Visualization

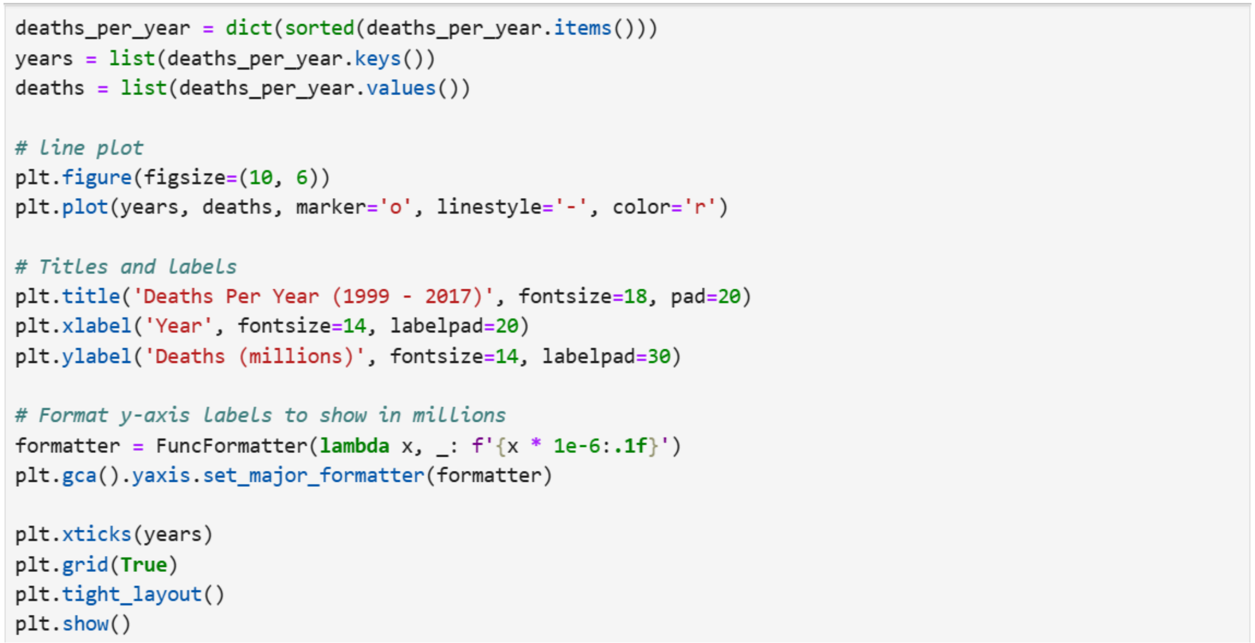
a. Bar graph for deaths per cause in 2017



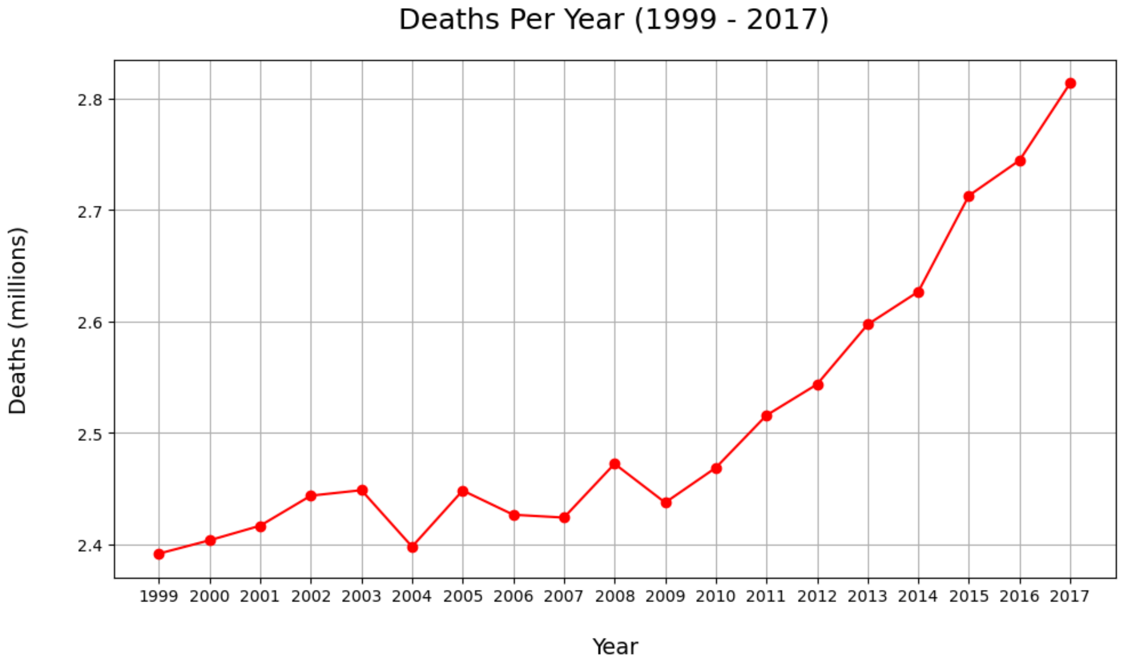
```
deaths_per_cause_2017 = dict(sorted(deaths_per_cause_2017.items(), key=lambda item: item[1]))

# Plotting
plt.figure(figsize=(12, 8))
plt.barh(list(deaths_per_cause_2017.keys()), list(deaths_per_cause_2017.values()), color='blue')

# Title and labels
plt.title("Deaths per Cause (2017)", fontsize=18, pad=20)
plt.xlabel("Deaths", fontsize=14, labelpad=30)
plt.ylabel("Cause", fontsize=14, labelpad=50)

plt.tight_layout()
plt.show()
```



b. Line graph for deaths per year (US)

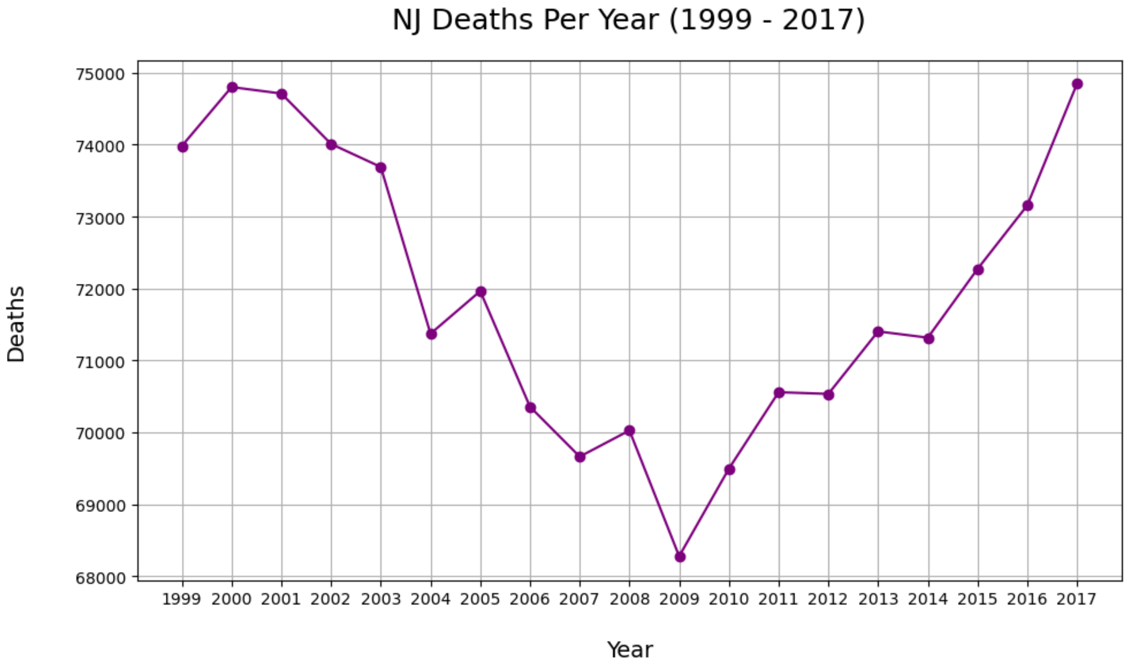


```
deaths_per_year = dict(sorted(deaths_per_year.items()))
years = list(deaths_per_year.keys())
deaths = list(deaths_per_year.values())

# Line plot
plt.figure(figsize=(10, 6))
plt.plot(years, deaths, marker='o', linestyle='-', color='r')

# Titles and labels
plt.title('Deaths Per Year (1999 - 2017)', fontsize=18, pad=20)
plt.xlabel('Year', fontsize=14, labelpad=20)
plt.ylabel('Deaths (millions)', fontsize=14, labelpad=30)

# Format y-axis labels to show in millions
formatter = FuncFormatter(lambda x, _: f'{x * 1e-6:.1f}')
plt.gca().yaxis.set_major_formatter(formatter)

plt.xticks(years)
plt.grid(True)
plt.tight_layout()
plt.show()
```

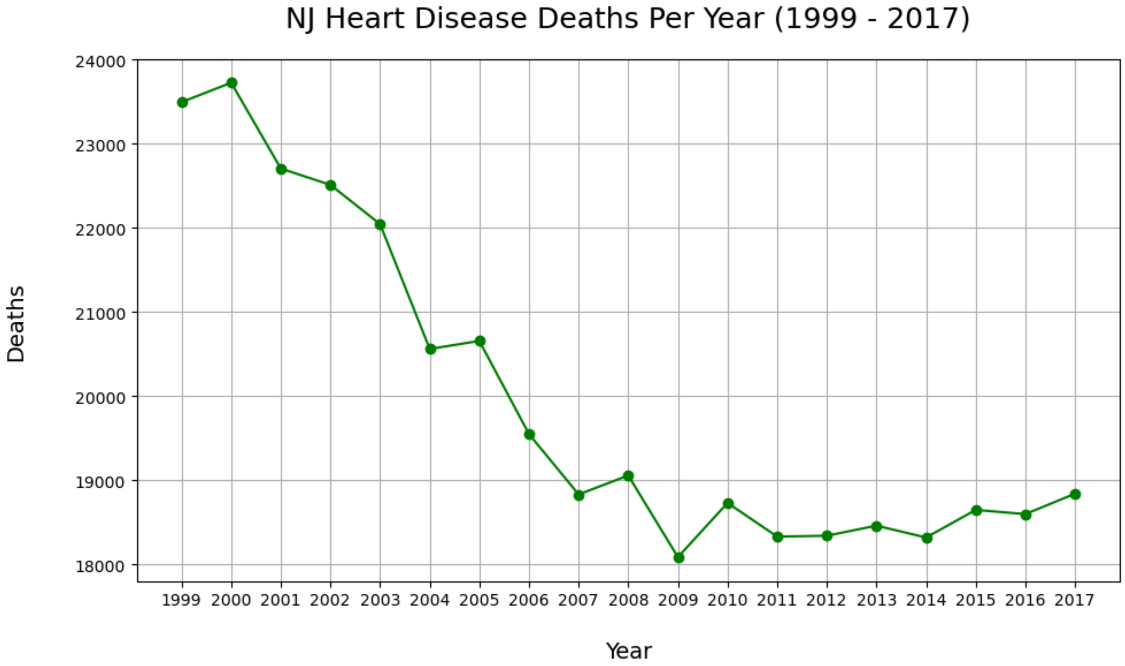


- Using the same method that was used to construct a line graph for part b, a line graph is also constructed for total deaths per year in NJ, deaths per year from heart disease in NJ, deaths per year from cancer in NJ, and deaths per year in California.

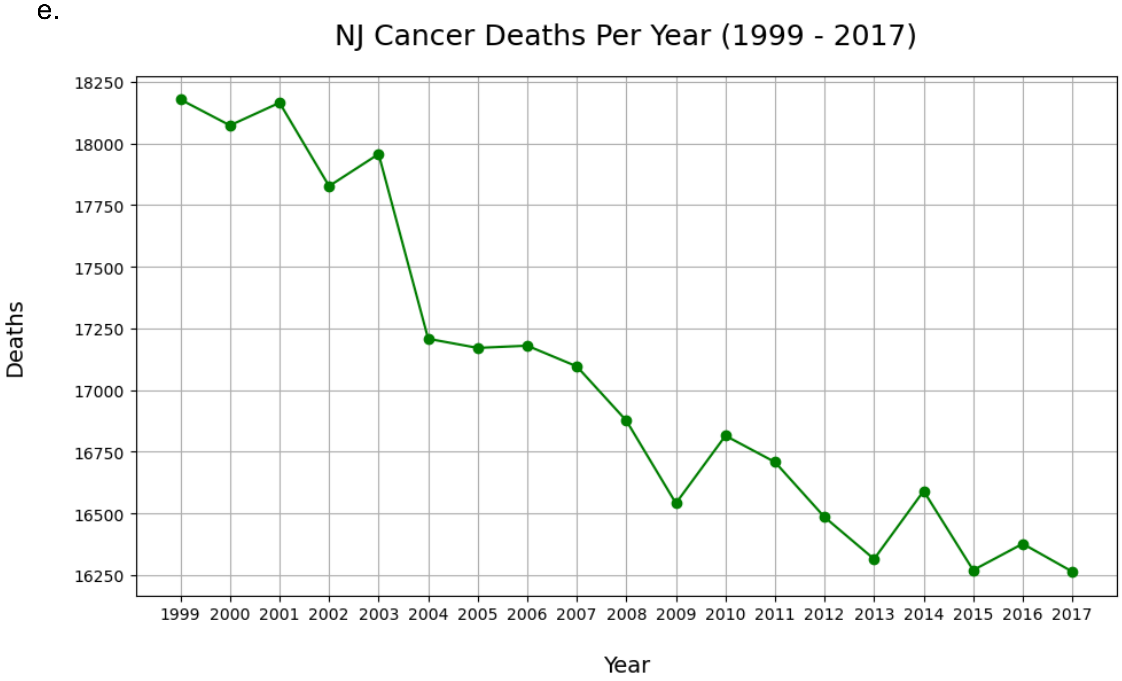
c.


NJ Deaths Per Year (1999 - 2017)
Deaths
75000
74000
73000
72000
71000
70000
69000
68000
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Year

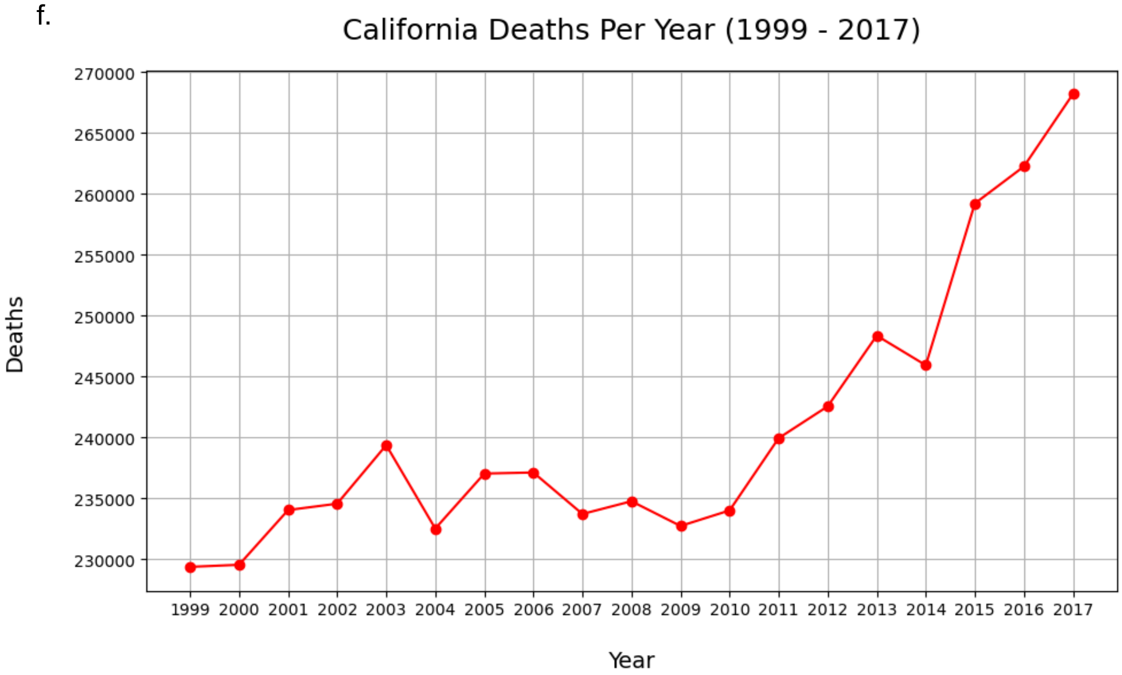
d.


NJ Heart Disease Deaths Per Year (1999 - 2017)
Deaths
24000
23000
22000
21000
20000
19000
18000
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Year

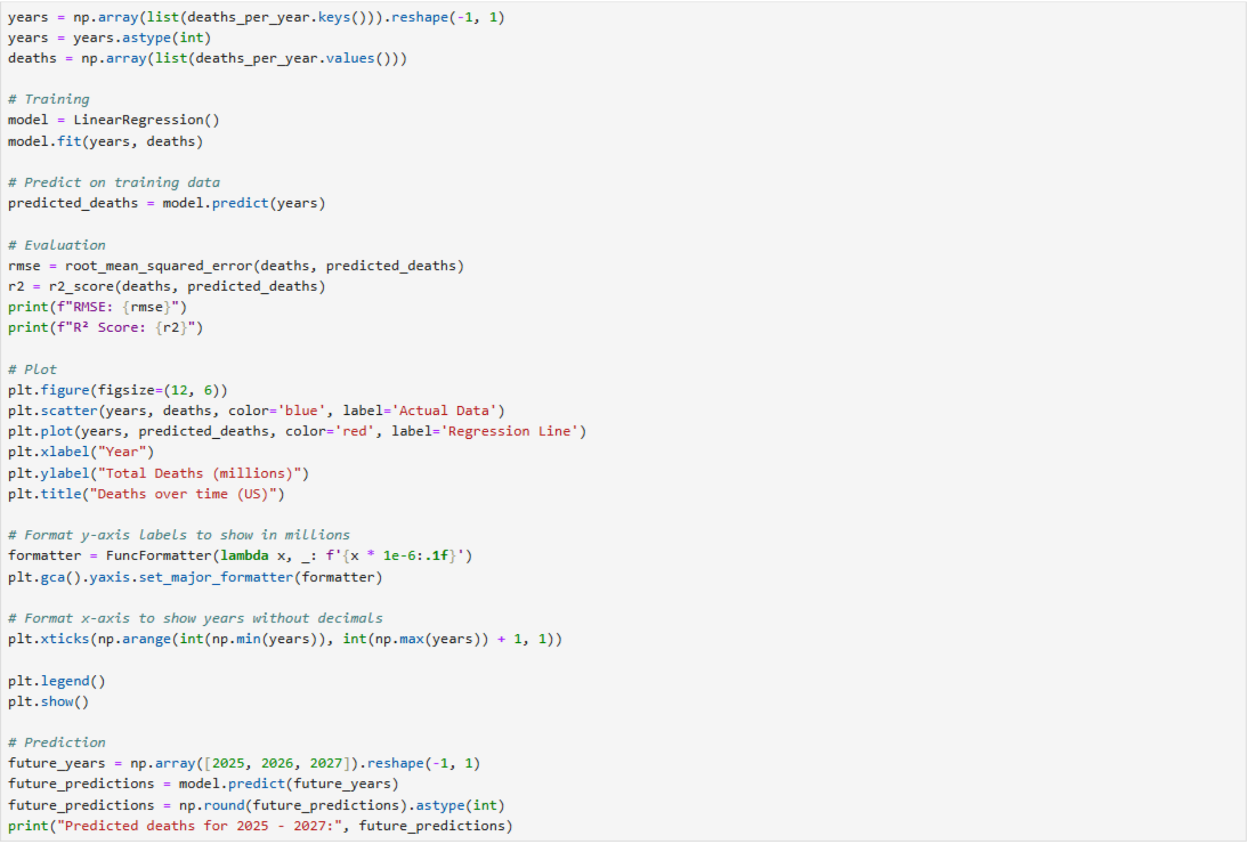
e.



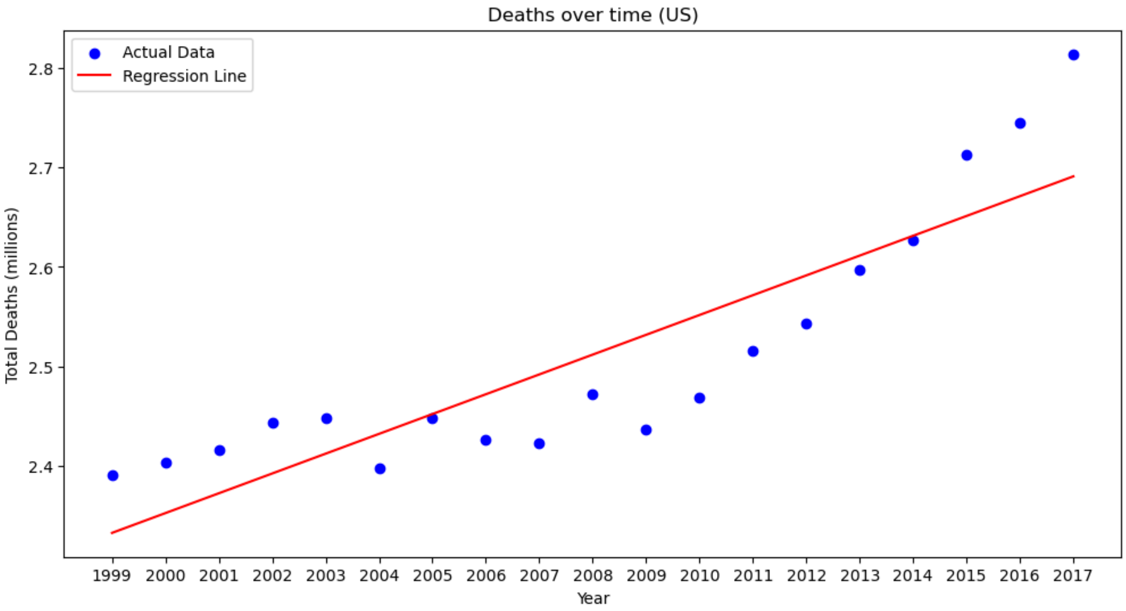
f.

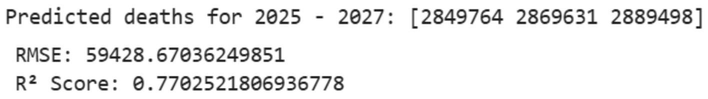


5. **Predictive Modeling**
   a. Prediction for deaths in US for 2025 - 2027



```python
years = np.array(list(deaths_per_year.keys())).reshape(-1, 1)
years = years.astype(int)
deaths = np.array(list(deaths_per_year.values()))

# Training
model = LinearRegression()
model.fit(years, deaths)

# Predict on training data
predicted_deaths = model.predict(years)

# Evaluation
rmse = root_mean_squared_error(deaths, predicted_deaths)
r2 = r2_score(deaths, predicted_deaths)
print(f"RMSE: {rmse}")
print(f"R² Score: {r2}")

# Plot
plt.figure(figsize=(12, 6))
plt.scatter(years, deaths, color='blue', label='Actual Data')
plt.plot(years, predicted_deaths, color='red', label='Regression Line')
plt.xlabel("Year")
plt.ylabel("Total Deaths (millions)")
plt.title("Deaths over time (US)")

# Format y-axis labels to show in millions
formatter = FuncFormatter(lambda x, _: f'{x * 1e-6:.1f}')
plt.gca().yaxis.set_major_formatter(formatter)

# Format x-axis to show years without decimals
plt.xticks(np.arange(int(np.min(years)), int(np.max(years)) + 1, 1))

plt.legend()
plt.show()

# Prediction
future_years = np.array([2025, 2026, 2027]).reshape(-1, 1)
future_predictions = model.predict(future_years)
future_predictions = np.round(future_predictions).astype(int)
print("Predicted deaths for 2025 - 2027:", future_predictions)
```

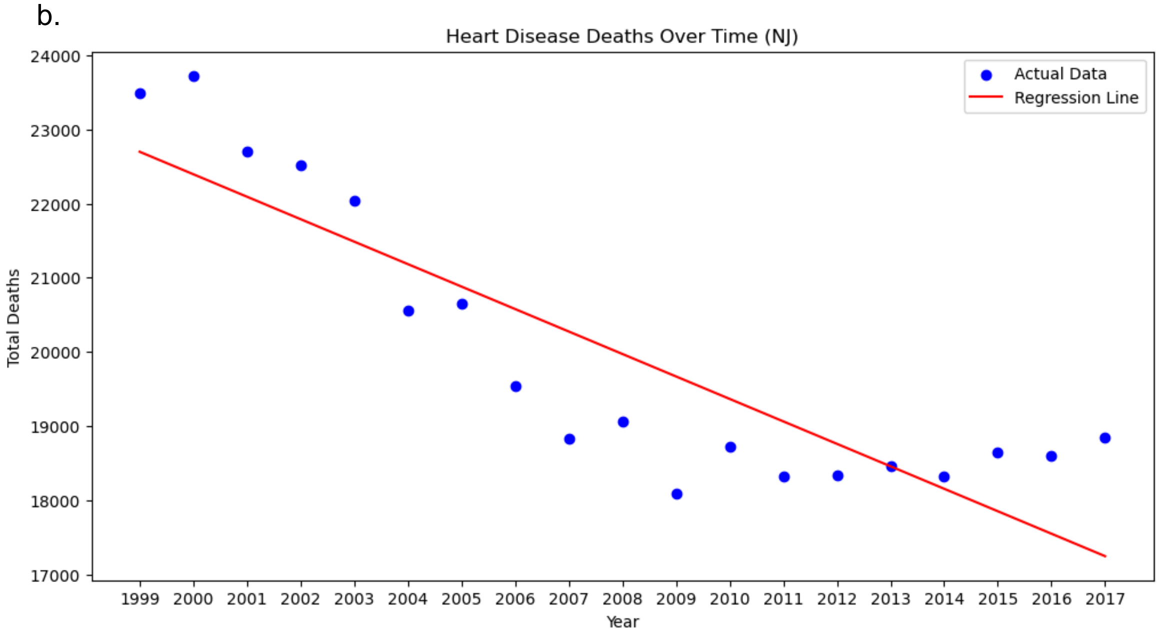




```
Predicted deaths for 2025 - 2027: [2849764 2869631 2889498]

RMSE: 59428.67036249851
R² Score: 0.7702521806936778
```

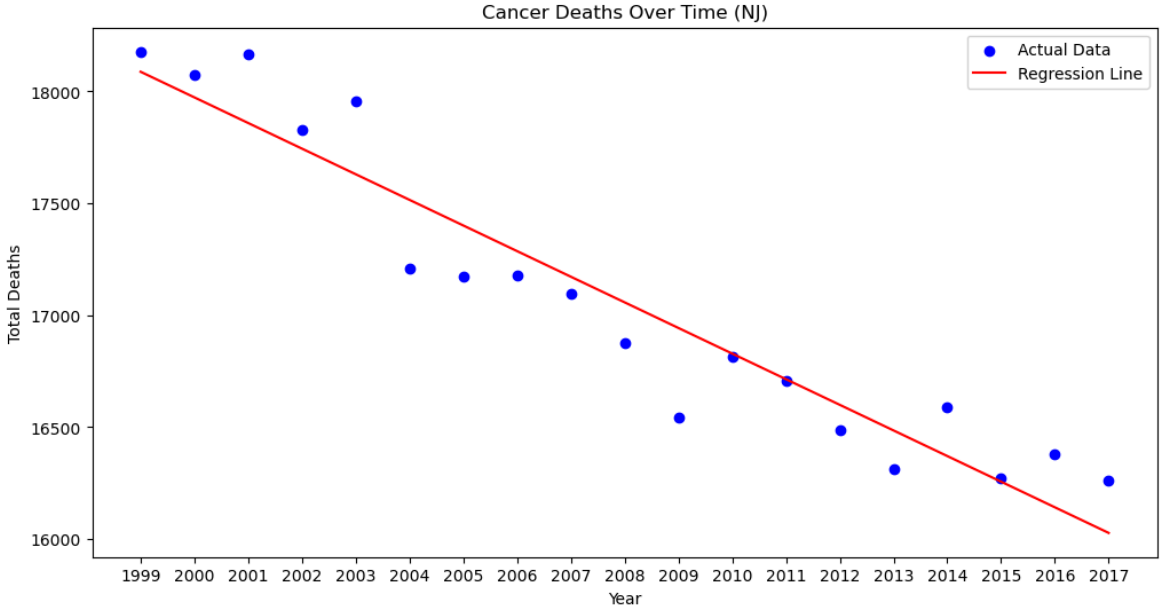
- Using the same method that was used to create the model and prediction for part a, models and predictions are also constructed for heart disease deaths over time in NJ and cancer deaths over time in NJ.

b.



```
Predicted deaths for 2025-2027: [14825 14522 14219]
RMSE: 916.2206123166964
R² Score: 0.7661587542742649
```

c.



```
Predicted deaths for 2025-2027: [15112 14997 14883]
RMSE: 203.54294003850956
R² Score: 0.9046646788360946
```

- Based on the predictions in parts b and c, it seems that cancer may surpass heart disease as the leading cause of death in New Jersey in the future.
- The prediction models seem to have pretty good fits, as R2 scores range from ~0.77 to ~0.90.

6. **Conclusion:**
   It seems that New Jersey is on a pretty good track, as its deaths have not increased much since 1999. Heart disease and cancer deaths have been going down as well and are predicted to decrease in the next few years. For states that are experiencing an increase in deaths, such as California, measures should be taken to prevent or reduce the risk of heart disease and cancer in individuals. A nutrient dense diet rich in fruits and vegetables and low in fat, salt, and sugar is recommended. Exercise and physical activity should be encouraged. Those who are overweight or obese should try to lose weight in order to lower risks. Rules and policies should be put in place to limit the use of tobacco and consumption of alcohol, and regulations should be put in place to reduce pollution. Individuals should also make sure to get enough sleep. All of these practices could help reduce heart disease and cancer in states experiencing an increase in deaths and ultimately reduce the total number of deaths in the US in the future.
7. **Github link:** https://github.com/jp1853/leading-causes-of-death-analysis