Trends in Heart Disease and Mortality: 1999-2019 Data Analysis

Group 2: Sincere Estrada, Sean Guzman, Lauren Golden, Cory Scarnecchia

Proposal

Until the COVID-19 pandemic, heart disease was the leading cause of death in the United States. In fact, in a CDC article published in August, 1999 states: “Despite remarkable progress, heart disease and stroke remain leading causes of disability and death.”

Substantial research and resources have been allocated over the last 20 years to study and disseminate information to reduce the occurrence of cardiovascular disease. These efforts are often covered in the media, and educational resources are embedded in health curricula for schools around the country.

The results have been clear: mortality due to cardiovascular disease has dropped. In looking at a two CDC datasets, and integrated census data, we will draw conclusions about how and why the change has occurred, and if the story is the same nationally as it is for New Jersey.

Research Questions

- How did mortality due to heart disease and stroke change from 1999 to 2019 (by different demographic features)?

- How did mortality due to heart disease and stroke change in New Jersey as compared to the nation?

- What contributing factors led to this change (i.e. smoking, diabetes, obesity)? Why?

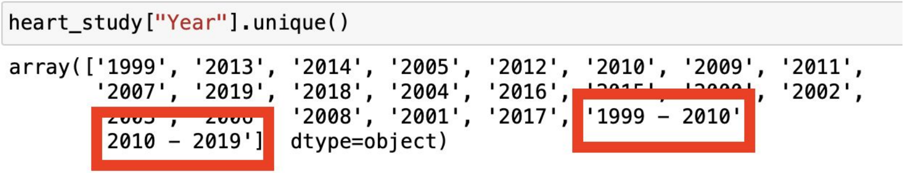
Initial Import and Cleaning

Data Sources

- CDC: Rates and Trends in Heart Disease and Stroke Mortality Among US Adults (35+) by County, Age Group, Race/Ethnicity, and Sex – 2000-2019
  - Original shape: 5,770,240 rows x 21 columns
  - Re-shaped to: 3,108,720 rows × 10 columns
- CDC: U.S. Chronic Disease Indicators (CDI)
  - Original shape: 1,185,676 rows x 34 columns down
  - Re-shaped to: 1,133,018 rows × 12 columns

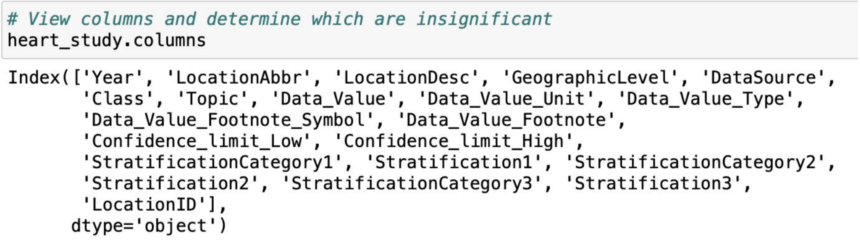
Data Cleaning Process

- Use .columns and .unique to determine which columns and values make sense for our study before dropping
- Rename columns
- Drop missing values / NaN



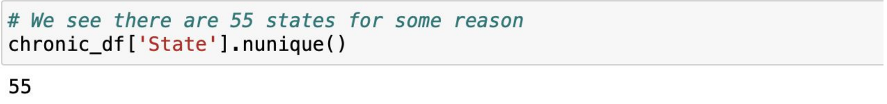
```
heart_study["Year"].unique()
```

```
array(['1999', '2013', '2014', '2005', '2012', '2010', '2009', '2011',
       '2007', '2019', '2018', '2004', '2016', '2015', '2000', '2002',
       '2003', '2006', '2008', '2001', '2017', '1999 - 2010',
       '2010 - 2019'], dtype=object)
```



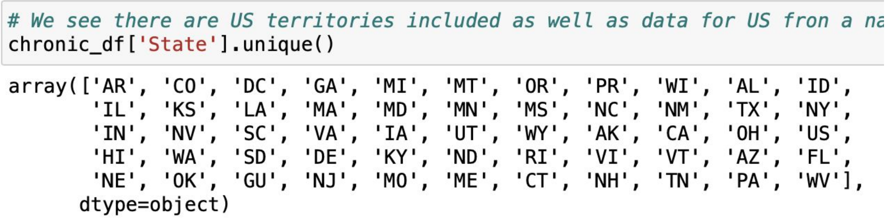
```
# View columns and determine which are insignificant
heart_study.columns
```

```
Index(['Year', 'LocationAbbr', 'LocationDesc', 'GeographicLevel', 'DataSource',
       'Class', 'Topic', 'Data_Value', 'Data_Value_Unit', 'Data_Value_Type',
       'Data_Value_Footnote_Symbol', 'Data_Value_Footnote',
       'Confidence_limit_Low', 'Confidence_limit_High',
       'StratificationCategory1', 'Stratification1', 'StratificationCategory2',
       'Stratification2', 'StratificationCategory3', 'Stratification3',
       'LocationID'],
      dtype='object')
```



```
# We see there are 55 states for some reason
chronic_df['State'].nunique()
```

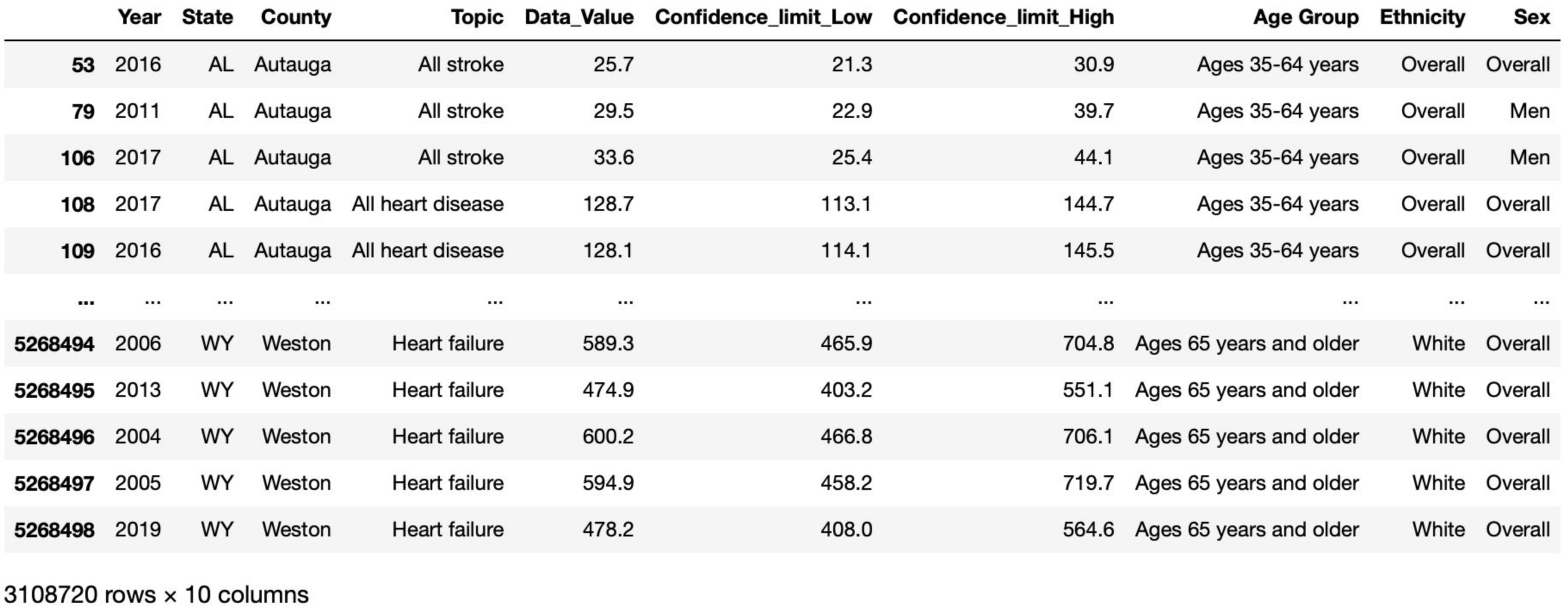
```
55
```



```
# We see there are US territories included as well as data for US fron a na
chronic_df['State'].unique()
```

```
array(['AR', 'CO', 'DC', 'GA', 'MI', 'MT', 'OR', 'PR', 'WI', 'AL', 'ID',
       'IL', 'KS', 'LA', 'MA', 'MD', 'MN', 'MS', 'NC', 'NM', 'TX', 'NY',
       'IN', 'NV', 'SC', 'VA', 'IA', 'UT', 'WY', 'AK', 'CA', 'OH', 'US',
       'HI', 'WA', 'SD', 'DE', 'KY', 'ND', 'RI', 'VI', 'VT', 'AZ', 'FL',
       'NE', 'OK', 'GU', 'NJ', 'MO', 'ME', 'CT', 'NH', 'TN', 'PA', 'WV'],
      dtype=object)
```

Cleaned Data Frame



	Year	State	County	Topic	Data_Value	Confidence_limit_Low	Confidence_limit_High	Age Group	Ethnicity	Sex
53	2016	AL	Autauga	All stroke	25.7	21.3	30.9	Ages 35-64 years	Overall	Overall
79	2011	AL	Autauga	All stroke	29.5	22.9	39.7	Ages 35-64 years	Overall	Men
106	2017	AL	Autauga	All stroke	33.6	25.4	44.1	Ages 35-64 years	Overall	Men
108	2017	AL	Autauga	All heart disease	128.7	113.1	144.7	Ages 35-64 years	Overall	Overall
109	2016	AL	Autauga	All heart disease	128.1	114.1	145.5	Ages 35-64 years	Overall	Overall
...	...	...	...	...	...	...	...	...	...	...
5268494	2006	WY	Weston	Heart failure	589.3	465.9	704.8	Ages 65 years and older	White	Overall
5268495	2013	WY	Weston	Heart failure	474.9	403.2	551.1	Ages 65 years and older	White	Overall
5268496	2004	WY	Weston	Heart failure	600.2	466.8	706.1	Ages 65 years and older	White	Overall
5268497	2005	WY	Weston	Heart failure	594.9	458.2	719.7	Ages 65 years and older	White	Overall
5268498	2019	WY	Weston	Heart failure	478.2	408.0	564.6	Ages 65 years and older	White	Overall

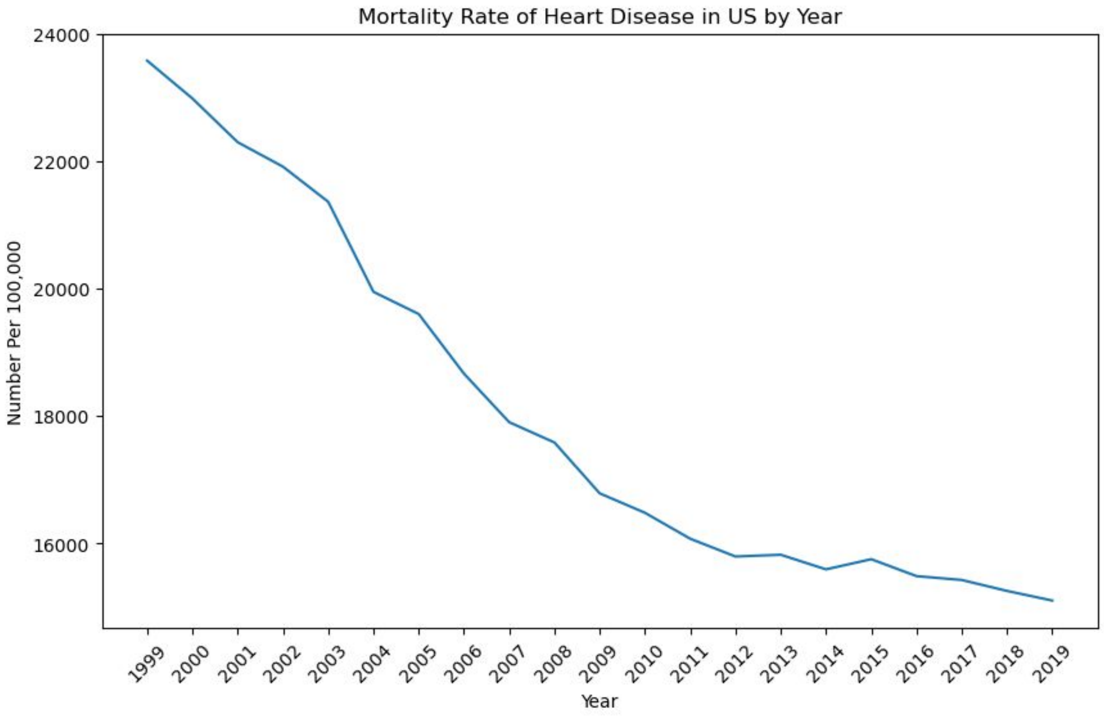
3108720 rows × 10 columns

National Trends including Supplemental Chronic Illness Data

The Overall Data

- The overall average in the United States for mortality of heart Disease.

- This showed us it was going down but didn't tell us enough as to why.



Key Questions

- How did mortality due to heart disease and stroke change from 1999 to 2019 (by different demographic features)?

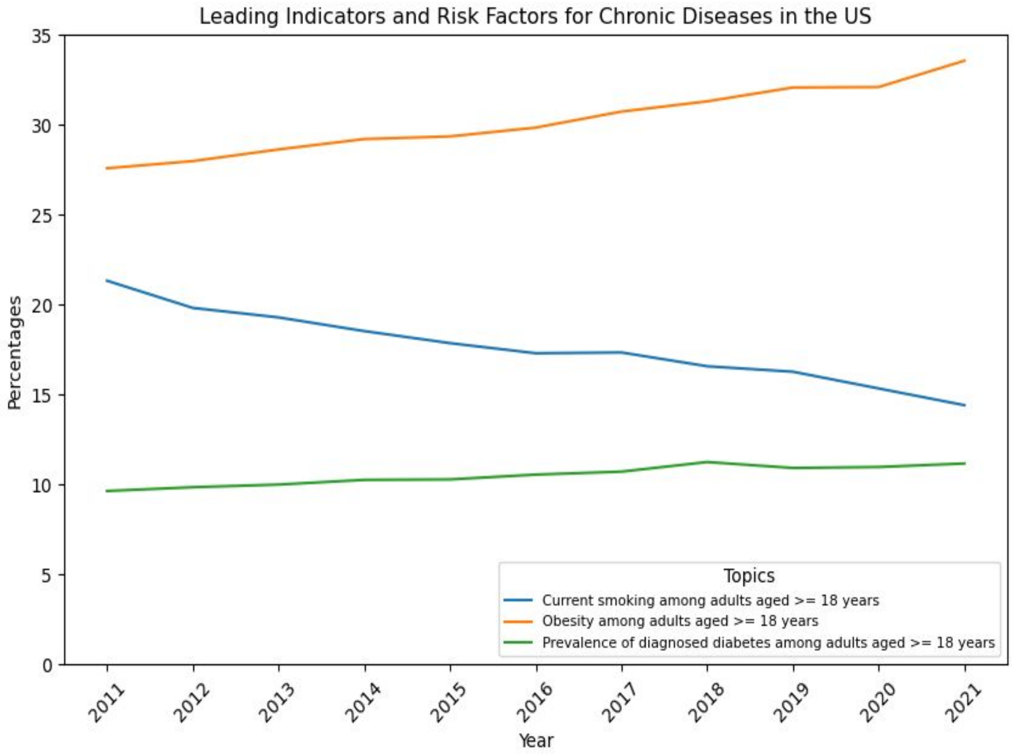
- What contributing factors led to this change (i.e. smoking, diabetes, obesity)? Why?

Finding Contributing Factors

- We found data for common things that could be a contributing factor to our original data.

- Two of these Indicators were not following the same trend.

- While smoking among adults seemed to be trending the same as our data

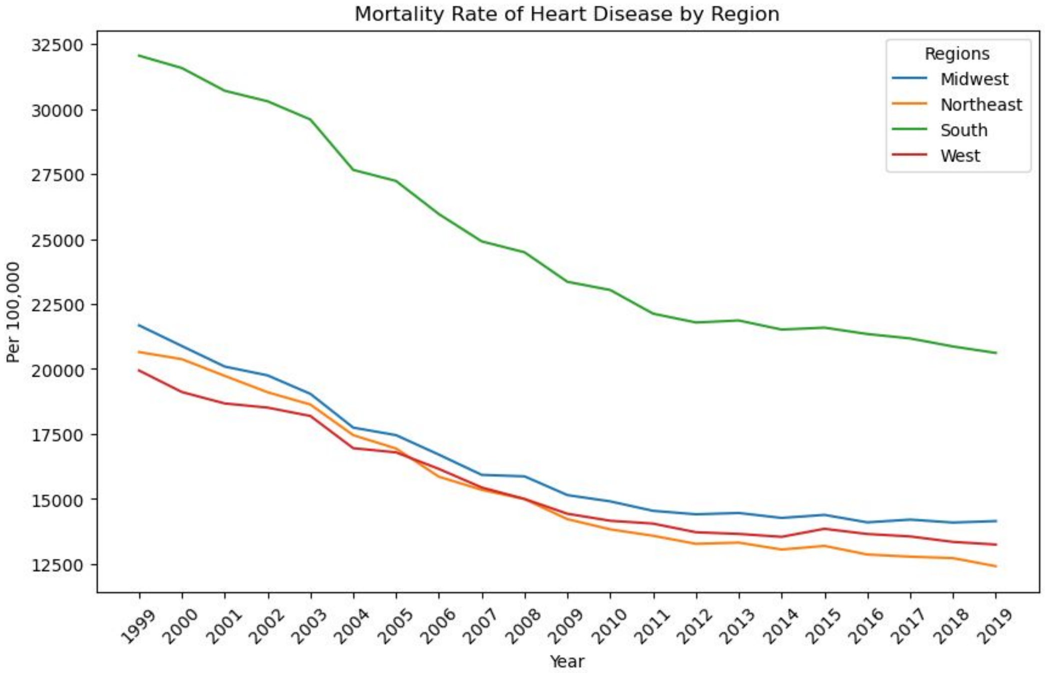


The US by Region

- We split the states into regions to see if there were any trends there.

- By region the overall is still trending down though the South has the highest overall and the other 3 are relatively about the same.

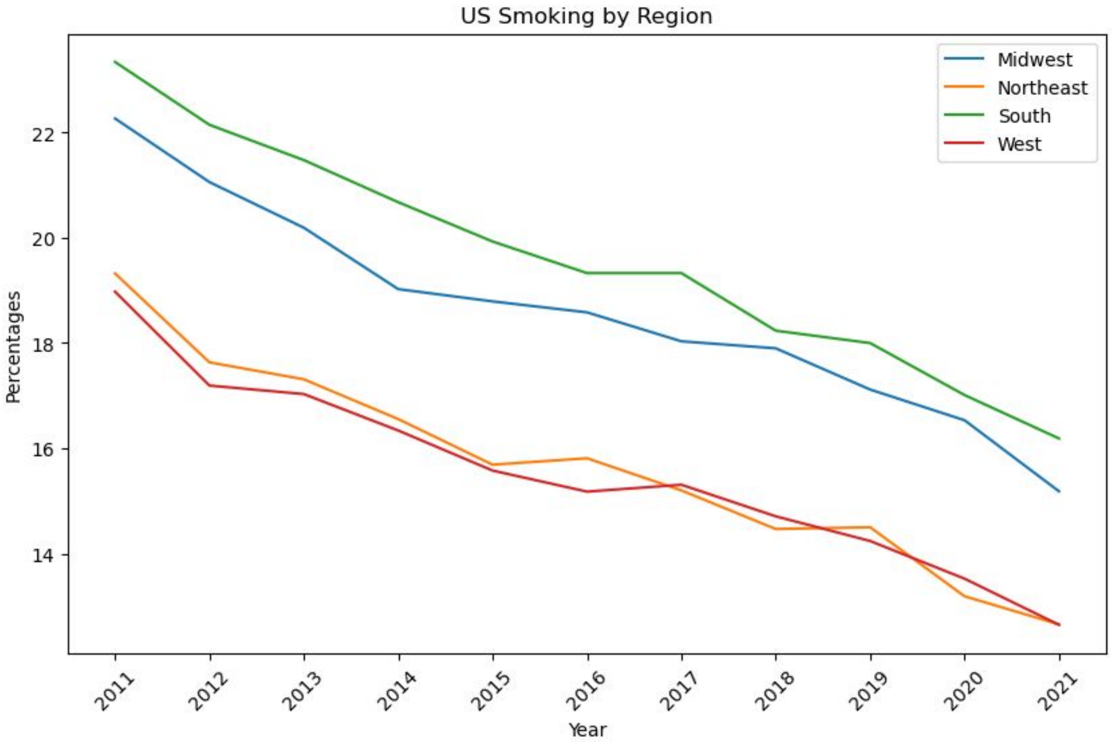
- We wanted to see if our contributing factors followed this trend



Smoking By Region

- Follows the same trend as our overall data for heart disease in the US.

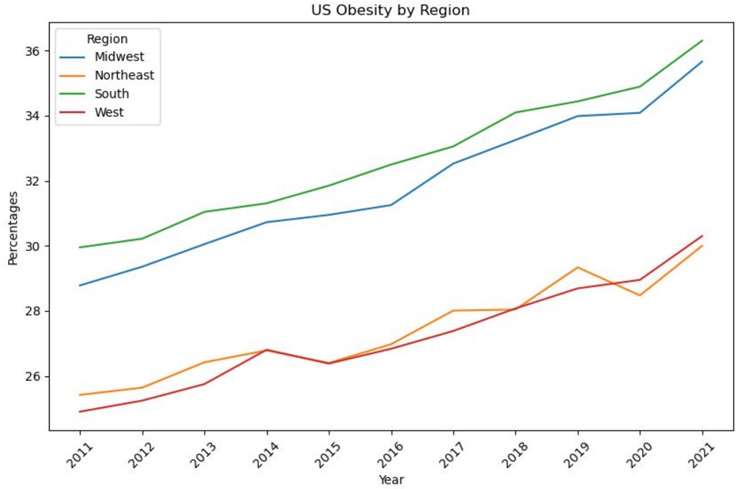
- The south is overall higher but the west and northeast are around the same overall

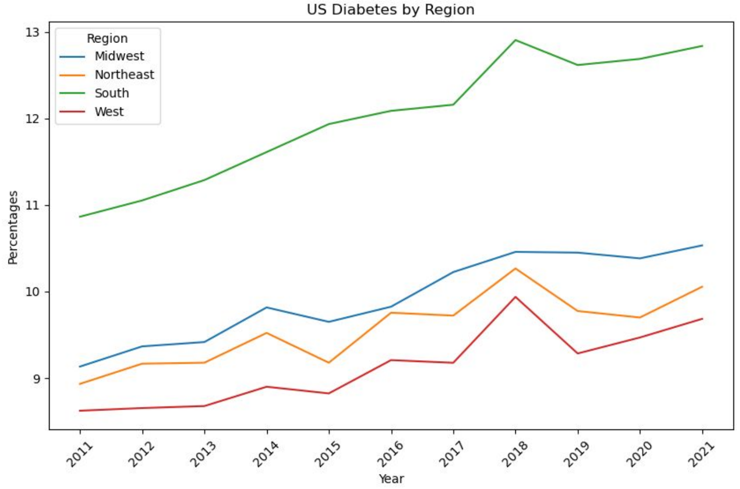


No Connection

- Based on what our overall data is doing these these don't seem to be a contributing factor as why it would be going down.

- The only thing this data shows is that the south seems to to be the highest overall in all the categories while the West and Northeast seem to around the same values.



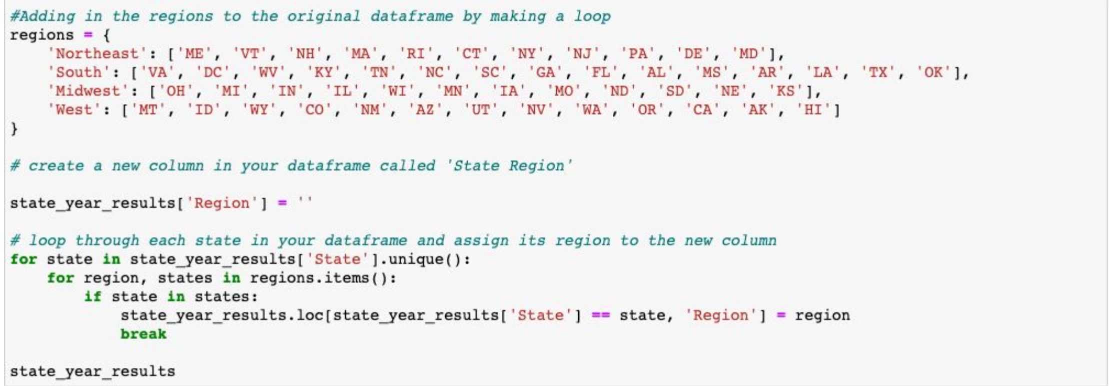


Prepping the Data for Overall and by Region

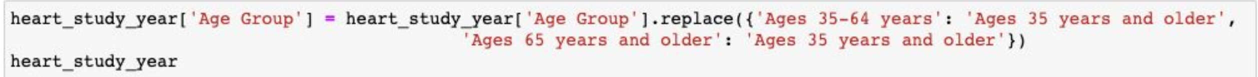
- Made a region dictionary and a loop to assign each state a region it fell into.

- Also had to group the data to only include the overalls

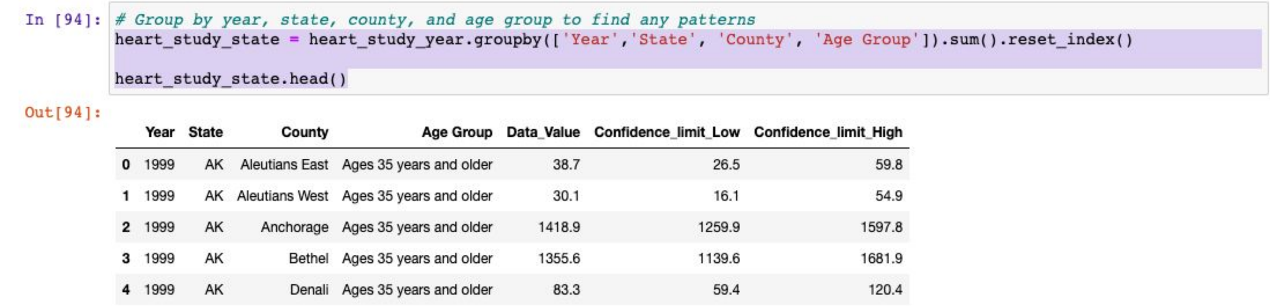
- Had to rename values in a column that was broken out by age to the same name a do a groupby.sum to get the total value



```
#Adding in the regions to the original dataframe by making a loop
regions = {
    'Northeast': ['ME', 'VT', 'NH', 'MA', 'RI', 'CT', 'NY', 'NJ', 'PA', 'DE', 'MD'],
    'South': ['VA', 'DC', 'WV', 'KY', 'TN', 'NC', 'SC', 'GA', 'FL', 'AL', 'MS', 'AR', 'LA', 'TX', 'OK'],
    'Midwest': ['OH', 'MI', 'IN', 'IL', 'WI', 'MN', 'IA', 'MO', 'ND', 'SD', 'NE', 'KS'],
    'West': ['MT', 'ID', 'WY', 'CO', 'NM', 'AZ', 'UT', 'NV', 'WA', 'OR', 'CA', 'AK', 'HI']
}

# create a new column in your dataframe called 'State Region'

state_year_results['Region'] = ''

# loop through each state in your dataframe and assign its region to the new column
for state in state_year_results['State'].unique():
    for region, states in regions.items():
        if state in states:
            state_year_results.loc[state_year_results['State'] == state, 'Region'] = region
            break

state_year_results
```



```
heart_study_year['Age Group'] = heart_study_year['Age Group'].replace({'Ages 35-64 years': 'Ages 35 years and older',
                                                                 'Ages 65 years and older': 'Ages 35 years and older'})
heart_study_year
```



```
In [94]: # Group by year, state, county, and age group to find any patterns
         heart_study_state = heart_study_year.groupby(['Year','State', 'County', 'Age Group']).sum().reset_index()

         heart_study_state.head()
```

Out[94]:

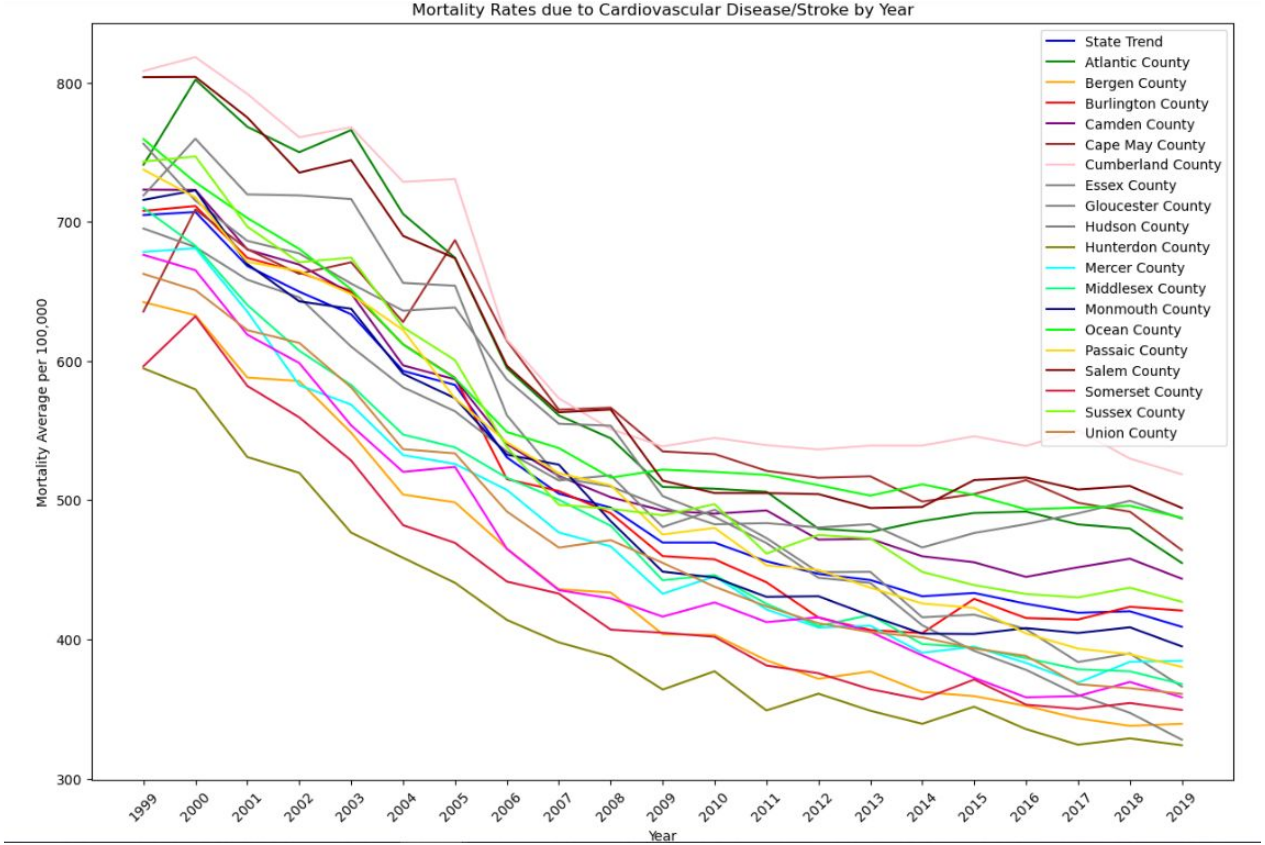
	Year	State	County	Age Group	Data_Value	Confidence_limit_Low	Confidence_limit_High
0	1999	AK	Aleutians East	Ages 35 years and older	38.7	26.5	59.8
1	1999	AK	Aleutians West	Ages 35 years and older	30.1	16.1	54.9
2	1999	AK	Anchorage	Ages 35 years and older	1418.9	1259.9	1597.8
3	1999	AK	Bethel	Ages 35 years and older	1355.6	1139.6	1681.9
4	1999	AK	Denali	Ages 35 years and older	83.3	59.4	120.4

NJ Up Close: Regional Data and Integrated Census API

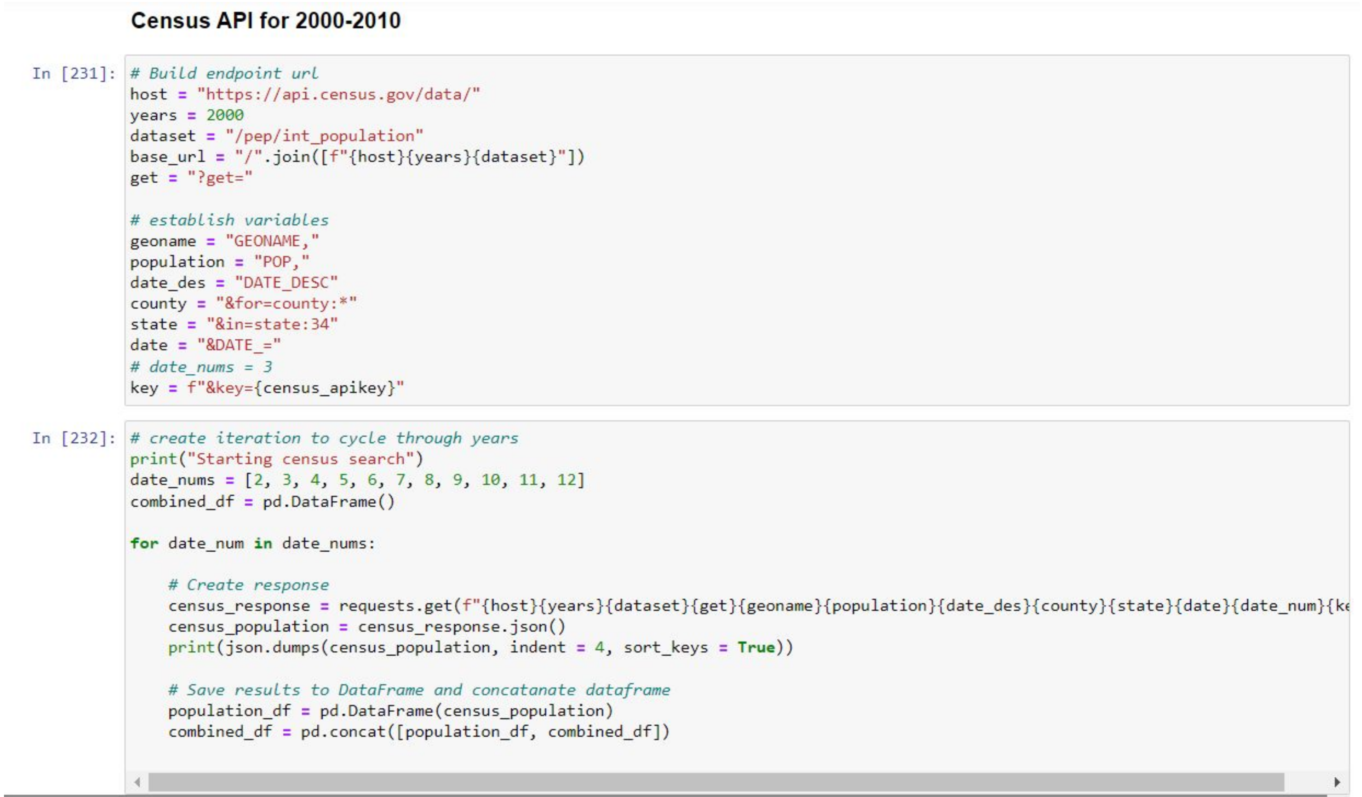
Key Question

- **How did mortality due to heart disease and stroke change in New Jersey?**

New Jersey Observations: Regional Graph (line)



Census API



Census API for 2000-2010

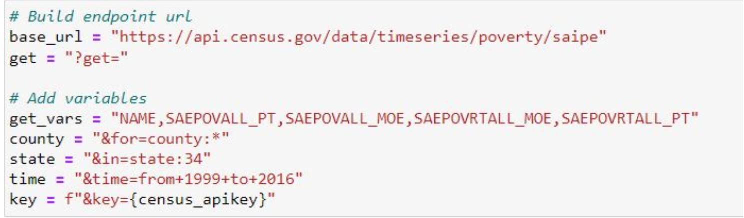
In [231]:

```
# Build endpoint url
host = "https://api.census.gov/data/"
years = 2000
dataset = "/pep/int_population"
base_url = "/".join([f"{host}{years}{dataset}"])
get = "?get="

# establish variables
geoname = "GEONAME,"
population = "POP,"
date_des = "DATE_DESC"
county = "&for=county:*"
state = "&in=state:34"
date = "&DATE_="
# date_nums = 3
key = f"&key={census_apikey}"
```

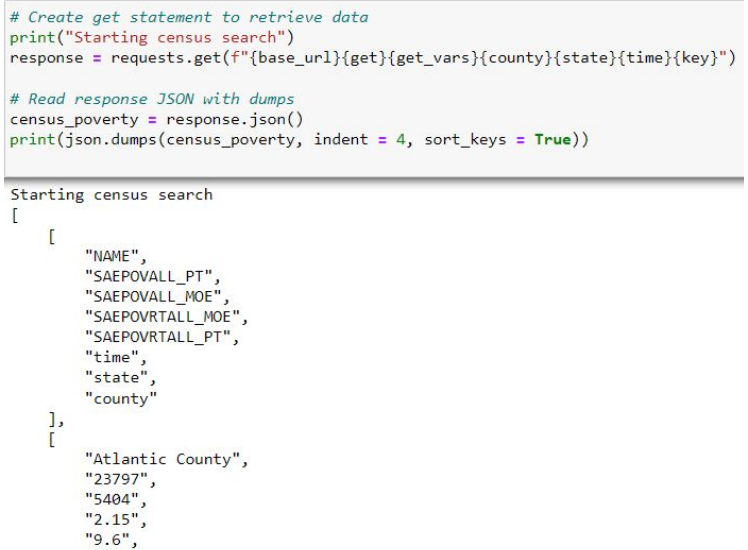
In [232]:

```
# create iteration to cycle through years
print("Starting census search")
date_nums = [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
combined_df = pd.DataFrame()

for date_num in date_nums:

    # Create response
    census_response = requests.get(f"{host}{years}{dataset}{get}{geoname}{population}{date_des}{county}{state}{date}{date_num}{ke
    census_population = census_response.json()
    print(json.dumps(census_population, indent = 4, sort_keys = True))

    # Save results to DataFrame and concatanate dataframe
    population_df = pd.DataFrame(census_population)
    combined_df = pd.concat([population_df, combined_df])
```

Census API

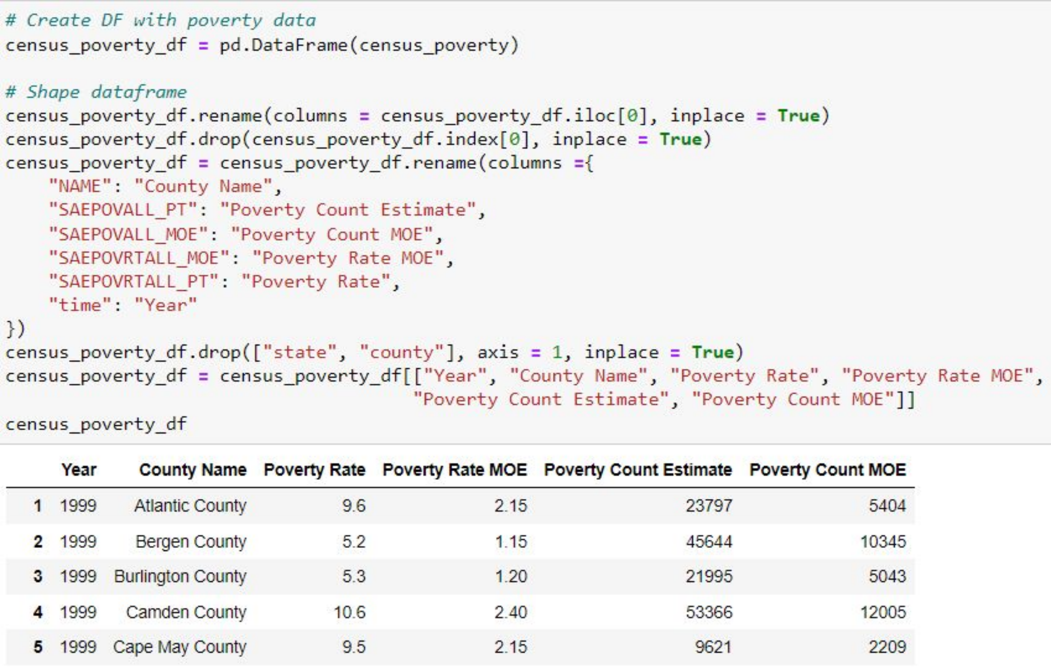


```python
# Build endpoint url
base_url = "https://api.census.gov/data/timeseries/poverty/saipe"
get = "?get="

# Add variables
get_vars = "NAME,SAEPOVALL_PT,SAEPOVALL_MOE,SAEPOVRTALL_MOE,SAEPOVRTALL_PT"
county = "&for=county:*"
state = "&in=state:34"
time = "&time=from+1999+to+2016"
key = f"&key={census_apikey}"
```



```python
# Create get statement to retrieve data
print("Starting census search")
response = requests.get(f"{base_url}{get}{get_vars}{county}{state}{time}{key}")

# Read response JSON with dumps
census_poverty = response.json()
print(json.dumps(census_poverty, indent = 4, sort_keys = True))
```

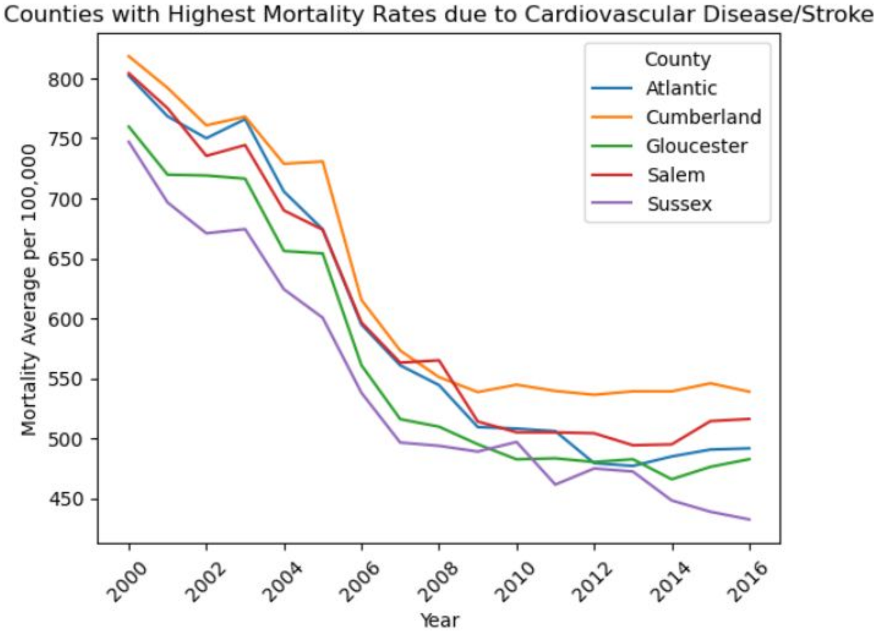
```
Starting census search
[
    [
        "NAME",
        "SAEPOVALL_PT",
        "SAEPOVALL_MOE",
        "SAEPOVRTALL_MOE",
        "SAEPOVRTALL_PT",
        "time",
        "state",
        "county"
    ],
    [
        "Atlantic County",
        "23797",
        "5404",
        "2.15",
        "9.6",
```

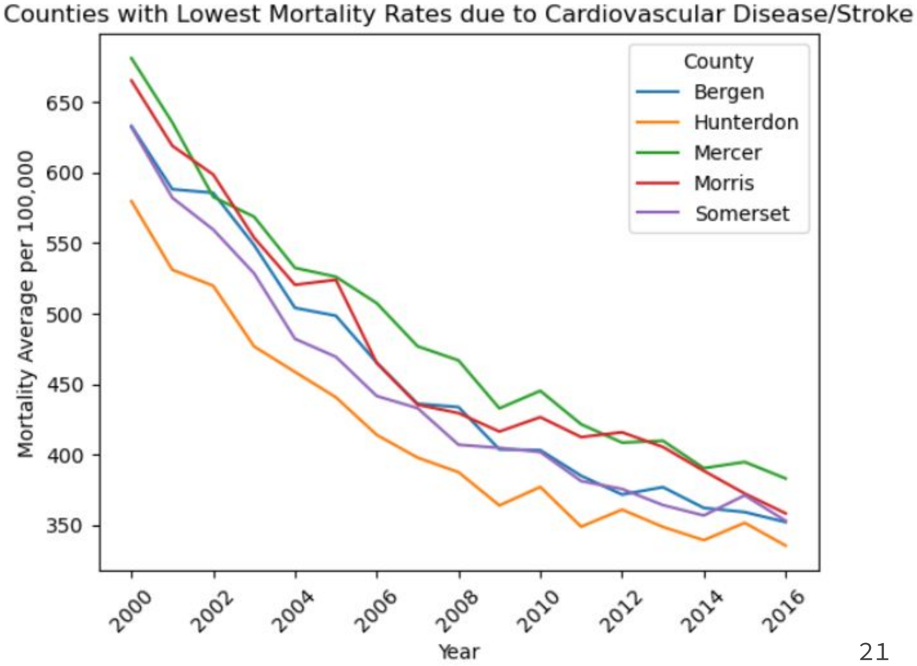


```python
# Create DF with poverty data
census_poverty_df = pd.DataFrame(census_poverty)

# Shape dataframe
census_poverty_df.rename(columns = census_poverty_df.iloc[0], inplace = True)
census_poverty_df.drop(census_poverty_df.index[0], inplace = True)
census_poverty_df = census_poverty_df.rename(columns ={
    "NAME": "County Name",
    "SAEPOVALL_PT": "Poverty Count Estimate",
    "SAEPOVALL_MOE": "Poverty Count MOE",
    "SAEPOVRTALL_MOE": "Poverty Rate MOE",
    "SAEPOVRTALL_PT": "Poverty Rate",
    "time": "Year"
})
census_poverty_df.drop(["state", "county"], axis = 1, inplace = True)
census_poverty_df = census_poverty_df[["Year", "County Name", "Poverty Rate", "Poverty Rate MOE",
                                       "Poverty Count Estimate", "Poverty Count MOE"]]
census_poverty_df
```

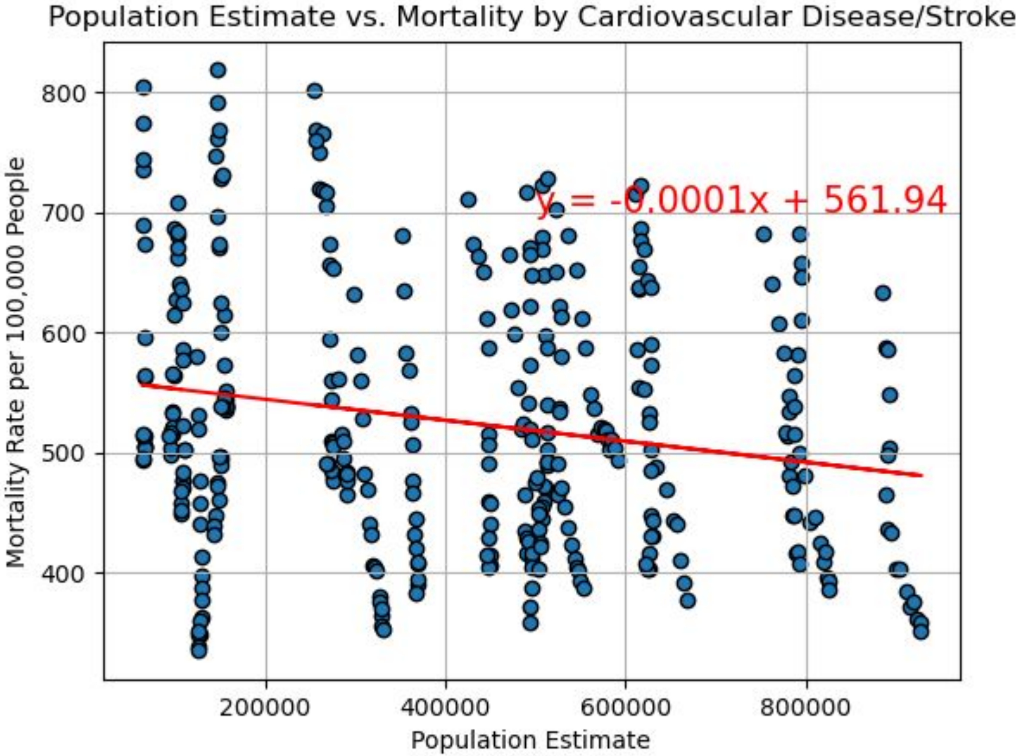
	Year	County Name	Poverty Rate	Poverty Rate MOE	Poverty Count Estimate	Poverty Count MOE
1	1999	Atlantic County	9.6	2.15	23797	5404
2	1999	Bergen County	5.2	1.15	45644	10345
3	1999	Burlington County	5.3	1.20	21995	5043
4	1999	Camden County	10.6	2.40	53366	12005
5	1999	Cape May County	9.5	2.15	9621	2209

New Jersey Observations: Regional Graph (with census data)

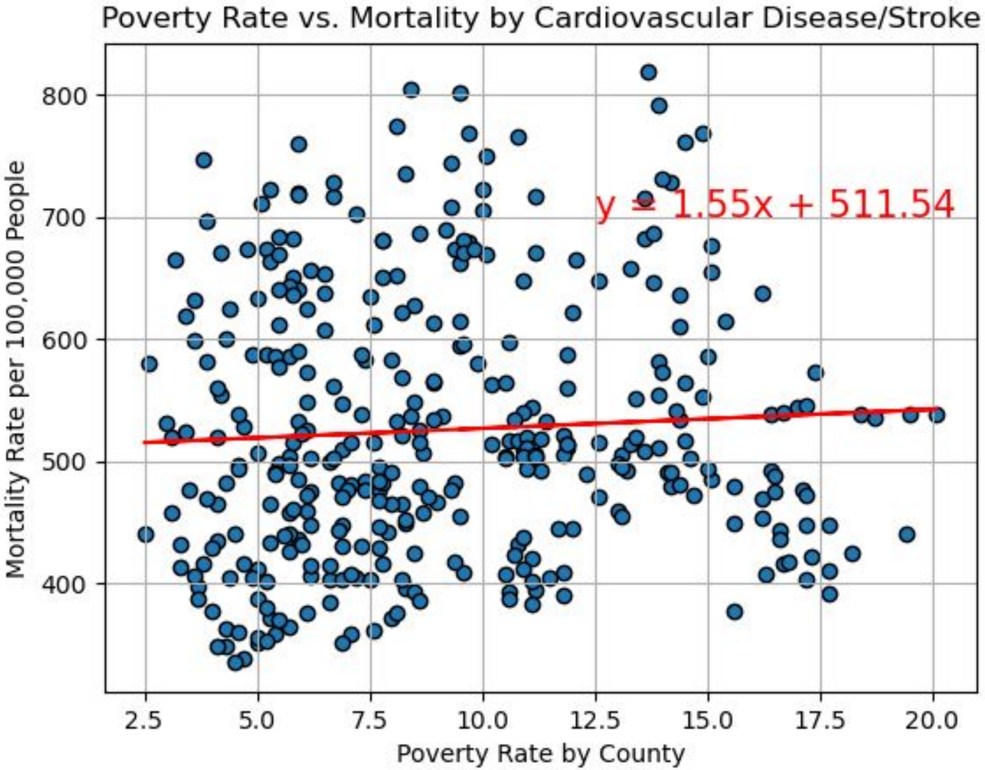




New Jersey Observations



New Jersey Observations



Conclusion

Final Conclusions

- **Mortality due to cardiovascular disease and stroke was on the decline as to 2019, along with some factors including smoking, both nationally and within NJ.**

- **Conversely, obesity and diabetes rates are on the rise and are considered risk factors for heart disease**

- **The regions of NJ who seem to be most affected are more remote than others. Further study is needed.**

- **Poverty does show correlation with mortality outcomes due to heart disease.**

- ***All data is pre-COVID-19, and further study is needed to see how these rates have been changed due to the pandemic.**