



# DATA SCIENCE CASE STUDY

Counties with Increase Mortality Rate Prediction

## Abstract

The objective of the document is present a noble way to predict the counties in United States which shall have a higher mortality rate compared to past

Gursewak Singh Sidhu  
Sidhus234@gmail.com

# DATA SCIENCE CASE STUDY

## Problem Statement:

### Description

The Institute for Health Metrics and Evaluation (IHME) has published estimates of annual county-level age-standardized mortality rates for 21 mutually exclusive causes of death from 1980 to 2014. This data is available from <http://ghdx.healthdata.org/record/united-states-mortality-rates-county-1980-2014> and you are free to supplement this dataset with any additional publicly available data you see fit.

## Summary:

The report covers below key topics:

- Understand the shift in Mortality rates in United States over a period of time (1980 – 2014)
- Pinpoint primary key causes of Mortality rate in 2014, based on gender
- Understand the causes that have resulted in increased mortality over a period of time
- Key causes, where mortality rates have fallen over the given time
- Understand the top causes of Mortality at State and gender level
- Analyse the top cause of increased mortality in each state
- Built a predictive model to predict counties which could have a higher mortality

# DATA SCIENCE CASE STUDY

## Index:

### Mortality Rates in United States:

- *Top 5 reasons for Mortality in 2014:*
- *Top 5 reasons which saw a major jump in mortalities:*
- *Top 5 reasons which saw a major reduction in mortalities:*

### Primary causes of increase in Mortality in 2014 (w.r.t 1980):

- *Males:*
- *Females:*

### Primary causes which saw a significant dip in Mortality (1980 to 2014):

- *Males:*
- *Females:*

### Predictive Model:

- *Objective:*
- *Design:*
- *Data:*
- *Performance:*
  - *Part 1: Random Forest Model:*
  - *Part 2: Gradient Boosting Model:*

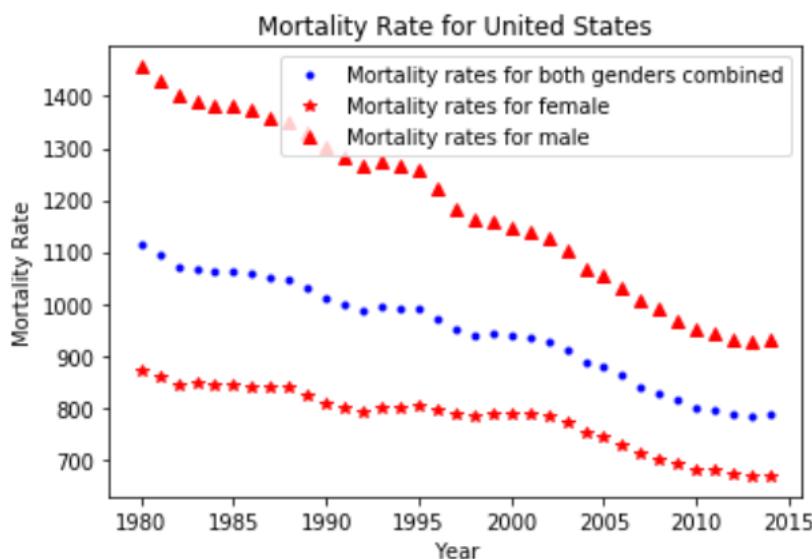
**Appendix 1: Table of top 2 reasons for mortalities in each state:**

**Appendix 2: Trend of Mortality Rates in each state**

# DATA SCIENCE CASE STUDY

## Mortality Rates in United States:

Mortality rates in United States have fallen from a high of 1120 (per 100,000) in 1980's to a low of 800 (per 100,000) in 2014. Over years, mortality rates have decreased at population level, as well as gender level. Breaking down the mortality by gender level, we observe that Males had a higher mortality rate than females by a margin of 600 in 1980, which has reduced to almost 200 in 2014. This means, that though the mortality rate for both male and female has come down with time, the effect of improved healthcare, reduction in death due to other causes has been higher for males.



**Fig:** Mortality rate for United States

From above graph, we can observe that:

- Mortality rates have decreased from 1,500 to almost 950 for males
- Mortality rates have decreased from 900 to almost 700 for females

The increased reduction of mortality rate in males by a significant margin (higher than women) could be attributed:

- Reduction in deaths in work-place
- Forces of Nature, War and legal intervention

# DATA SCIENCE CASE STUDY

## Top 5 reasons for Mortality in 2014:

- **Population Level:** Non-communicable disease is the primary cause of mortality, contributing about 44.2% at national level. Cardiovascular diseases, Neoplasms, Neurological disorders, and Diabetes, urogenital, blood, and endocrine diseases are other four causes. Combined, these five causes are responsible for almost ~82% of the mortalities.
- **Males:** Non-communicable diseases, Cardiovascular diseases, Neoplasms, Neurological disorders, Injuries are top five reasons for mortality in males. Together, these reasons, result in almost 81% of the mortalities in men. The major difference between the population (both genders combined) and men is the last 5<sup>th</sup> reason. Though, in overall population, Diabetes, urogenital, blood, and endocrine diseases is 5<sup>th</sup> reason, for males, it is Injuries. This could be because of multiple reasons. Men in general, are still employed in higher proportion in blue collar and risky jobs.
- **Females:** The top five reasons for mortalities is similar to the general population. Females have seen a lower improvement in reduction in mortalities over the period of 1980 to 2014.

## Top 5 causes for Mortality in US Population are:

Cause ID	Cause Name	% Contribution
409	Non-communicable diseases	44.233026
491	Cardiovascular diseases	15.923330
410	Neoplasms	12.100924
542	Neurological disorders	6.008843
586	Diabetes, urogenital, blood, and endocrine diseases	3.521814

## Top 5 causes for Mortality in US Males are:

Cause ID	Cause Name	% Contribution
409	Non-communicable diseases	43.572549
491	Cardiovascular diseases	15.989049
410	Neoplasms	12.495839

# DATA SCIENCE CASE STUDY

Cause ID	Cause Name	% Contribution
542	Neurological disorders	4.842448
687	Injuries	4.049257

**Top 5 causes for Mortality in US Females are:**

Cause ID	Cause Name	% Contribution
409	Non-communicable diseases	45.106112
491	Cardiovascular diseases	15.796139
410	Neoplasms	11.944118
542	Neurological disorders	7.191806
586	Diabetes, urogenital, blood, and endocrine diseases	3.630300

**Top 5 reasons which saw a major jump in mortalities:**

- **Population Level:** “Mental and substance use disorders” saw the highest jump of 188% in mortalities during the period, followed by “HIV/AIDS, and tuberculosis” (74% increase). “Chronic respiratory diseases”, “Maternal disorders”, “Neglected tropical diseases and malaria” are the other top three causes which saw a significant increase of 25 to 30%.
- **Males:** Compared to overall population, the jump in mortalities by “Mental and substance use disorders” is only 156%, 30% lower than the general combined population. “Neglected tropical diseases and malaria” which saw an overall jump of only 24%, increased by 46% for males. “Diabetes, urogenital, blood, and endocrine diseases”, and “Neurological disorders” are the last two causes, which are different from the general population.
- **Females:** “Chronic respiratory diseases” saw a jump of 95% for mortalities in females, whereas it saw an overall jump of only 20% in combined population. “Neglected tropical diseases and malaria” does not figure in top 5 causes, which saw a jump, though for an overall population, it saw a jump of 30%.

# DATA SCIENCE CASE STUDY

**Top 5 causes which increased for Mortality in US Population are:**

Cause ID	Cause Name	% Change1980-2014
558	Mental and substance use disorders	188.392306
296	HIV/AIDS and tuberculosis	74.348946
508	Chronic respiratory diseases	29.734155
366	Maternal disorders	29.247451
344	Neglected tropical diseases and malaria	24.973543

**Top 5 causes which increased for Mortality in US Males are:**

Cause ID	Cause Name	% Change1980-2014
558	Mental and substance use disorders	156.098135
296	HIV/AIDS and tuberculosis	62.445688
344	Neglected tropical diseases and malaria	46.190035
586	Diabetes, urogenital, blood, and endocrine diseases	19.498276
542	Neurological disorders	16.090767

**Top 5 causes which increased for Mortality in US Females are:**

Cause ID	Cause Name	% Change1980-2014
558	Mental and substance use disorders	261.142898
508	Chronic respiratory diseases	95.305125
296	HIV/AIDS and tuberculosis	64.166832
366	Maternal disorders	30.982418
542	Neurological disorders	21.020457

# DATA SCIENCE CASE STUDY

## Top 5 reasons which saw a major reduction in mortalities:

- **Population Level:** “Forces of nature, war, and legal intervention”, saw the most drop-in mortality rates at the population level, as well as at gender level. The second most drop was observed in “Neonatal disorders”, at almost the same rate for population, independent of gender. This was followed by “Cardiovascular diseases”, “Transport Injuries”, and “Other non-communicable diseases”. There was no difference observed at the gender level.

## Top 5 causes which decreased for Mortality in US Population are:

Cause ID	Cause Name	% Change1980-2014
728	Forces of nature, war, and legal intervention	-79.535567
380	Neonatal disorders	-63.853234
491	Cardiovascular diseases	-50.195349
688	Transport injuries	-45.449236
640	Other non-communicable diseases	-40.032023

## Top 5 causes which decreased for Mortality in US Males are:

Cause ID	Cause Name	% Change1980-2014
728	Forces of nature, war, and legal intervention	-82.420831
380	Neonatal disorders	-65.228181
491	Cardiovascular diseases	-55.347948
688	Transport injuries	-48.062508
640	Other non-communicable diseases	-40.774052

## Top 5 causes which decreased for Mortality in US Females are:

Cause ID	Cause Name	% Change1980-2014
728	Forces of nature, war, and legal intervention	-73.219381
380	Neonatal disorders	-61.951784
491	Cardiovascular diseases	-45.548424
688	Transport injuries	-40.019257
640	Other non-communicable diseases	-39.174121

# DATA SCIENCE CASE STUDY

## Top Causes of Mortality in each state:

The table ([Table](#)) gives the top contributors for mortality in 2014 in each state. For all the states, the top two reasons for mortality are:

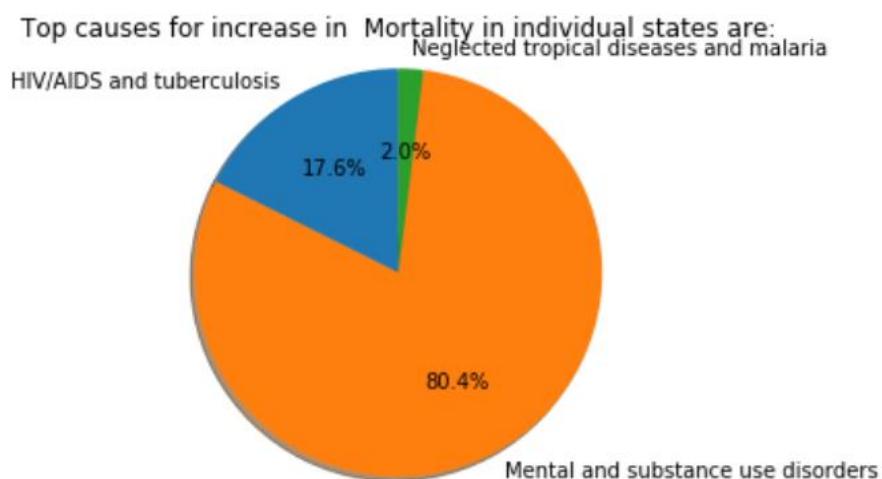
- Non-communicable diseases
- Cardiovascular diseases

Though the percentages vary for each state, but on average, around 42-44% of the mortalities in each state are caused by Non-Communicable diseases and approx. 15% are caused by Cardiovascular diseases. There is no major difference between males and females. The top 2 reasons for mortalities are same as in general population for each state.

## Primary causes of increase in Mortality in 2014 (w.r.t 1980):

When, we look at the top reasons that resulted in increased mortalities over the period (1980 to 2014):

- In 80.4% (41 states), the primary reason which saw highest jump was “Mental and substance use disorders”
- In 17.6% (9 states), “HIV/AIDS and tuberculosis”, saw the highest increase. (Delaware, District of Colombia, Florida, Georgia, Maryland, Mississippi, New York, North Carolina, and South Carolina)
- And in remaining 1 state (South Dakota), “Neglected tropical diseases and malaria” saw the highest jump.



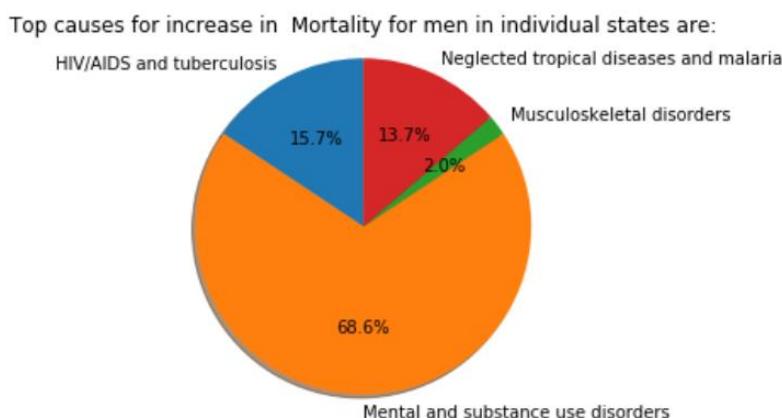
**Fig:** Causes with highest increase

# DATA SCIENCE CASE STUDY

## Males:

Compared to general population, the trends in males are bit different:

- In 35 states, “mental and substance use disorder” was the primary reason for increased male mortalities
- In 8 states, “HIV/AIDS and tuberculosis” was the primary reason, which saw the highest increase in mortalities (Delaware, District of Columbia, Georgia, Maryland, Mississippi, New York, North Carolina, and South Carolina)
- “Neglected tropical diseases and malaria” saw the highest increase in 7 states (Arizona, Colorado, Nebraska, New Mexico, North Dakota, South Dakota, and Wyoming)
- “Musculoskeletal disorders” saw the highest increase in Alaska

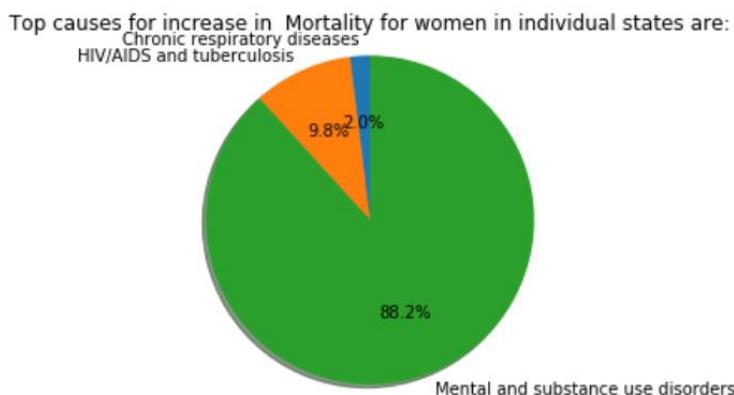


**Fig:** Major reasons for increased mortalities in males

## Females:

Females saw higher increase in mortalities by “mental and substance use disorders” in most states (45),

- “HIV/AIDS and tuberculosis” saw the highest jump in 5 states (District of Columbia, Florida, Georgia, Maryland, and New York)
- North Carolina saw the highest jump in mortalities for females in “Chronic respiratory disease”



**Fig:** Top reasons for increase in mortalities in females

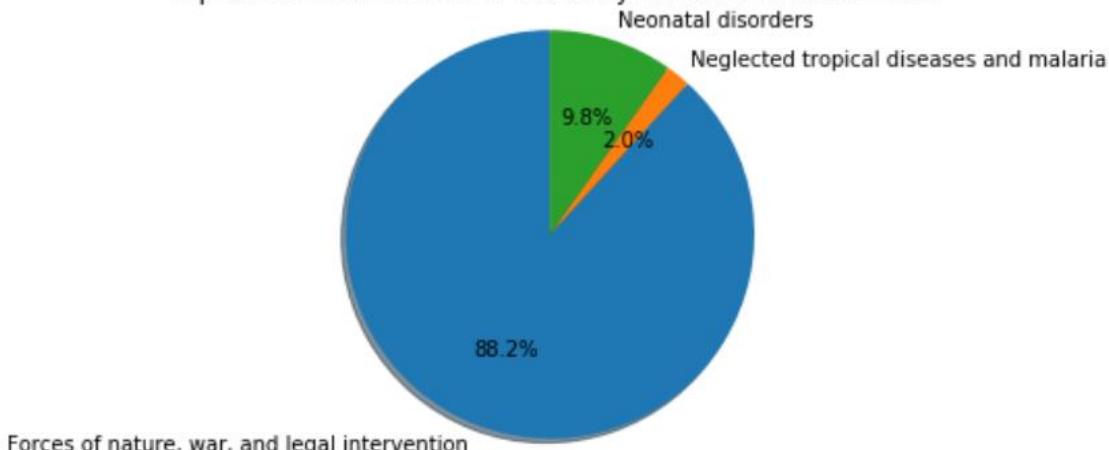
# DATA SCIENCE CASE STUDY

## Primary causes which saw a significant dip in Mortality (1980 to 2014):

### Overall Population:

- For overall population, the “Forces of nature, war, and legal intervention”, saw most reduction in mortality rates in 45 states.
- With improvement in healthcare, “Neonatal disorders”, saw the most drop in another 5 states (Maine, Montana, New Hampshire, Vermont, and Washington )
- In one state, Hawaii, “Neglected tropical diseases and malaria” saw the most drop.

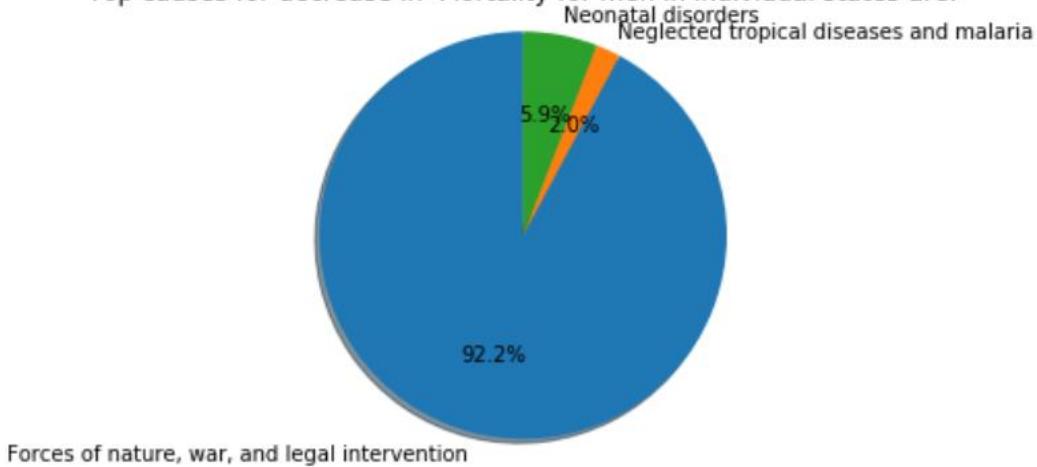
Top causes for decrease in Mortality in individual states are:



### Males:

- In males, in 47 states, “Forces of nature, war, and legal intervention” saw the most reduction in mortality rates.
- In Another 3 states, “Neonatal disorder” was the cause, which saw highest reduction (Montana, New Hampshire, Washington)
- “Neglected tropical diseases and malaria” in Hawaii.

Top causes for decrease in Mortality for men in individual states are:

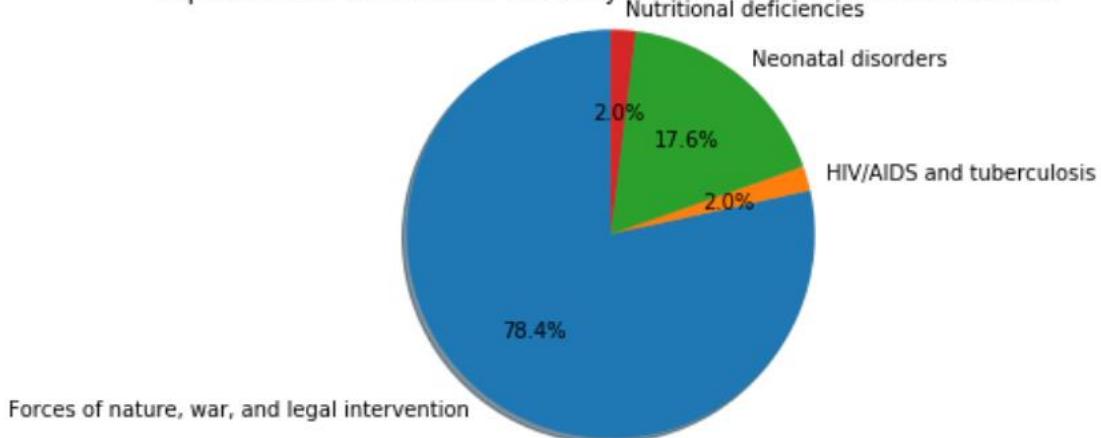


# DATA SCIENCE CASE STUDY

## Females:

- For females, in 39 states, “Forces of nature, war, and legal intervention” saw highest drop.
- “Neonatal disorder” saw the most drop in 9 states (Alaska, Maine, Massachusetts, Mississippi, New Hampshire, New Jersey, New York, Vermont, and Washington)
- “HIV/AIDS and tuberculosis” and “Nutritional deficiencies” saw the most drop in 1 state each (Hawaii, and District of Colombia respectively)

Top causes for decrease in Mortality for women in individual states are:



## Predictive Model:

This part covers the details on the predictive model. Every year mortalities, result in increased outgo for an insurance firm. If, we can predict the counties with higher expected mortality rate in future, the firm could then take a call on provisions, capital adequacy and investment decisions. This could also help in deciding the premium rate at a county level.

### Objective:

The objective of the model is to predict the counties (in United States), which will see the mortality rate increase by >1.5% in future.

### Design:

The dependent variable is defined as below:

2010	2011	2012	2013	2014
Observation Period		Lag Period		Prediction period

The data from 2010, 2011 and 2012 will be used to predict the counties, which will see increased mortalities by >1.5%, with respect to 2012 base.

### Data:

Below three data sources were used in this

# DATA SCIENCE CASE STUDY

1. Institute for Health Metrics and Evaluation (IHME) has published estimates of annual county-level age-standardized mortality rates
2. County level population migration, birth, death data from US census (<https://www2.census.gov/programs-surveys/popest/datasets/2010-2013/counties/>)
3. County level population data by race (<https://www2.census.gov/programs-surveys/popest/datasets/2010-2013/counties/totals/>)

## Performance:

### Part 1: Random Forest Model:

- Random forest is a machine learning model, which used majority voting to classify the observation. Random Forest Classifier is an ensemble algorithm, which creates a set of decision trees from a randomly selected subset of the training set, which then aggregates the votes from different decision trees to decide the final class of the test object.
- It performs better than the logistic model in terms of predicting counties which will see a significant increase in mortalities in future.
- The model has a KS of 44.6%, and 42.6% in train and test sample.

### Part 2: Gradient Boosting Model:

- A gradient boosting model is a CHAIN of decision trees that each make a vote. But instead of each learning in isolation, when you add a new one to the chain, it tries to improve a bit on what the rest of the chain already thinks. So, a new tree's decision is influenced by all the trees that have already voiced an opinion.
- It performs better than the logistic and Random Forest model in terms of performance.
- It has a KS of 52.4% and 44.6% for train and test sample respectively.

# DATA SCIENCE CASE STUDY

## Appendix 1: Table of top 2 reasons for mortalities in each state:

state	Cause Name	Cause ID	% Contribution
ALABAMA	Non-communicable diseases	409	43.64
	Cardiovascular diseases	491	15.60
ALASKA	Non-communicable diseases	409	43.63
	Cardiovascular diseases	491	13.89
ARIZONA	Non-communicable diseases	409	43.63
	Cardiovascular diseases	491	14.47
ARKANSAS	Non-communicable diseases	409	43.45
	Cardiovascular diseases	491	16.79
CALIFORNIA	Non-communicable diseases	409	44.87
	Cardiovascular diseases	491	16.52
COLORADO	Non-communicable diseases	409	43.50
	Cardiovascular diseases	491	14.41
CONNECTICUT	Non-communicable diseases	409	44.22
	Cardiovascular diseases	491	15.31
DELAWARE	Non-communicable diseases	409	44.03
	Cardiovascular diseases	491	15.77
DISTRICT_OF_COLUMBIA	Non-communicable diseases	409	43.51
	Cardiovascular diseases	491	17.80
FLORIDA	Non-communicable diseases	409	44.05
	Cardiovascular diseases	491	15.89
GEORGIA	Non-communicable diseases	409	43.80
	Cardiovascular diseases	491	15.59
HAWAII	Non-communicable diseases	409	43.57
	Cardiovascular diseases	491	15.74
IDAHO	Non-communicable diseases	409	44.11
	Cardiovascular diseases	491	14.90
ILLINOIS	Non-communicable diseases	409	44.53
	Cardiovascular diseases	491	16.25
INDIANA	Non-communicable diseases	409	44.45
	Cardiovascular diseases	491	15.62
IOWA	Non-communicable diseases	409	44.55
	Cardiovascular diseases	491	16.08
KANSAS	Non-communicable diseases	409	43.98
	Cardiovascular diseases	491	15.50
KENTUCKY	Non-communicable diseases	409	44.11
	Cardiovascular diseases	491	15.19
LOUISIANA	Non-communicable diseases	409	43.52
	Cardiovascular diseases	491	15.91
MAINE	Non-communicable diseases	409	44.65
	Cardiovascular diseases	491	13.66

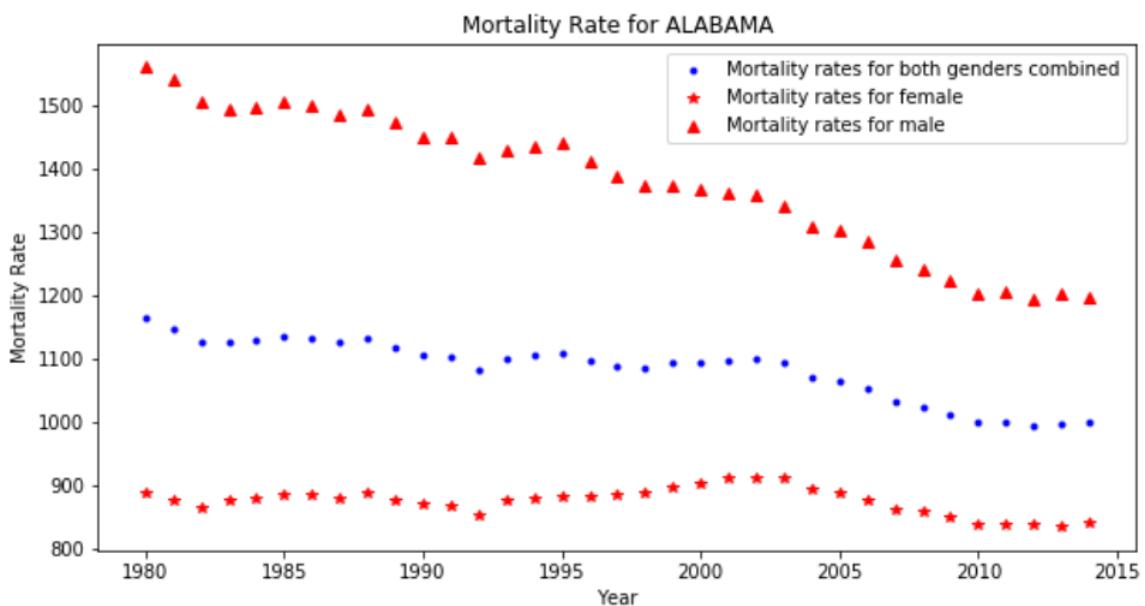
# DATA SCIENCE CASE STUDY

MARYLAND	Non-communicable diseases	409	43.47
	Cardiovascular diseases	491	16.42
MASSACHUSETTS	Non-communicable diseases	409	44.80
	Cardiovascular diseases	491	14.51
MICHIGAN	Non-communicable diseases	409	44.71
	Cardiovascular diseases	491	17.05
MINNESOTA	Non-communicable diseases	409	44.60
	Cardiovascular diseases	491	13.32
MISSISSIPPI	Non-communicable diseases	409	43.20
	Cardiovascular diseases	491	16.82
MISSOURI	Non-communicable diseases	409	43.92
	Cardiovascular diseases	491	16.48
MONTANA	Non-communicable diseases	409	43.22
	Cardiovascular diseases	491	14.19
NEBRASKA	Non-communicable diseases	409	44.47
	Cardiovascular diseases	491	14.59
NEVADA	Non-communicable diseases	409	43.74
	Cardiovascular diseases	491	15.74
NEW_HAMPSHIRE	Non-communicable diseases	409	44.91
	Cardiovascular diseases	491	14.66
NEW_JERSEY	Non-communicable diseases	409	44.70
	Cardiovascular diseases	491	16.94
NEW_MEXICO	Non-communicable diseases	409	42.89
	Cardiovascular diseases	491	13.86
NEW_YORK	Non-communicable diseases	409	44.44
	Cardiovascular diseases	491	18.42
NORTH_CAROLINA	Non-communicable diseases	409	43.83
	Cardiovascular diseases	491	14.94
NORTH_DAKOTA	Non-communicable diseases	409	44.13
	Cardiovascular diseases	491	15.54
OHIO	Non-communicable diseases	409	44.72
	Cardiovascular diseases	491	15.71
OKLAHOMA	Non-communicable diseases	409	44.04
	Cardiovascular diseases	491	16.95
OREGON	Non-communicable diseases	409	44.76
	Cardiovascular diseases	491	13.86
PENNSYLVANIA	Non-communicable diseases	409	44.32
	Cardiovascular diseases	491	16.26
RHODE_ISLAND	Non-communicable diseases	409	44.75
	Cardiovascular diseases	491	15.42
SOUTH_CAROLINA	Non-communicable diseases	409	43.81
	Cardiovascular diseases	491	14.79
SOUTH_DAKOTA	Non-communicable diseases	409	43.62
	Cardiovascular diseases	491	15.28
TENNESSEE	Non-communicable diseases	409	43.80
	Cardiovascular diseases	491	16.01

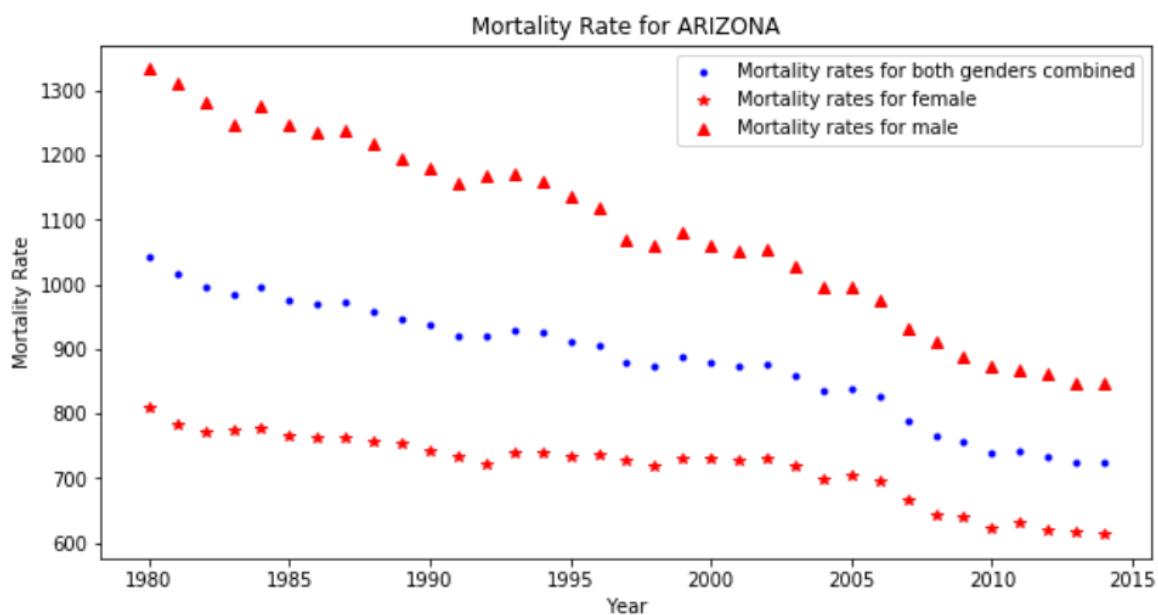
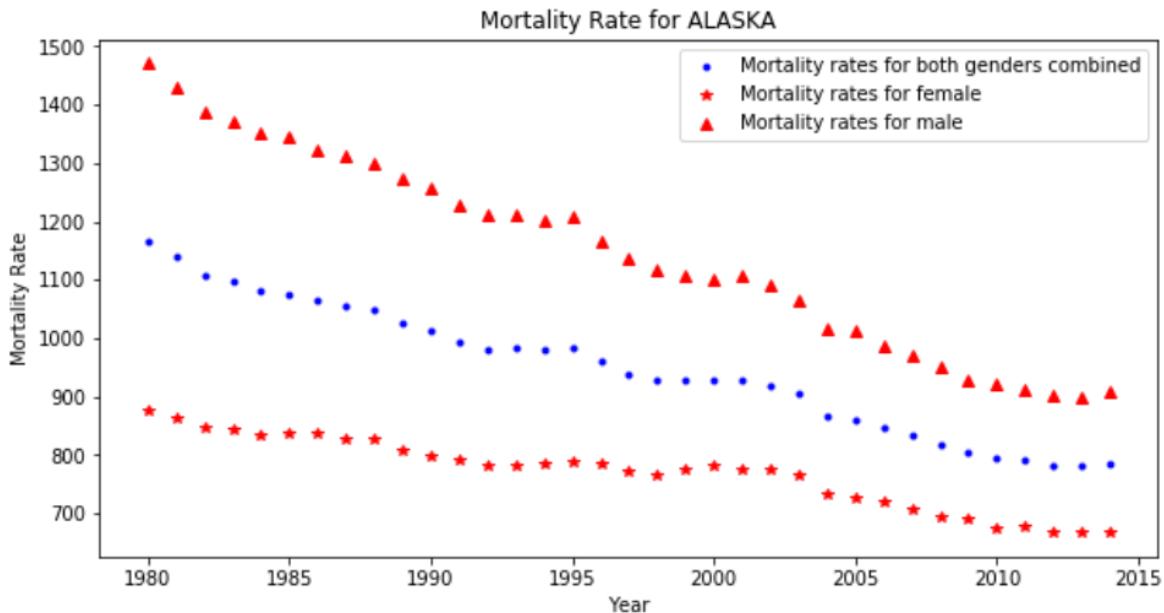
# DATA SCIENCE CASE STUDY

TEXAS	Non-communicable diseases	409	44.09
	Cardiovascular diseases	491	16.05
UTAH	Non-communicable diseases	409	43.51
	Cardiovascular diseases	491	15.12
VERMONT	Non-communicable diseases	409	44.45
	Cardiovascular diseases	491	14.61
VIRGINIA	Non-communicable diseases	409	43.95
	Cardiovascular diseases	491	15.29
WASHINGTON	Non-communicable diseases	409	44.81
	Cardiovascular diseases	491	14.76
WEST_VIRGINIA	Non-communicable diseases	409	44.09
	Cardiovascular diseases	491	15.13
WISCONSIN	Non-communicable diseases	409	44.35
	Cardiovascular diseases	491	15.14
WYOMING	Non-communicable diseases	409	43.06
	Cardiovascular diseases	491	14.69

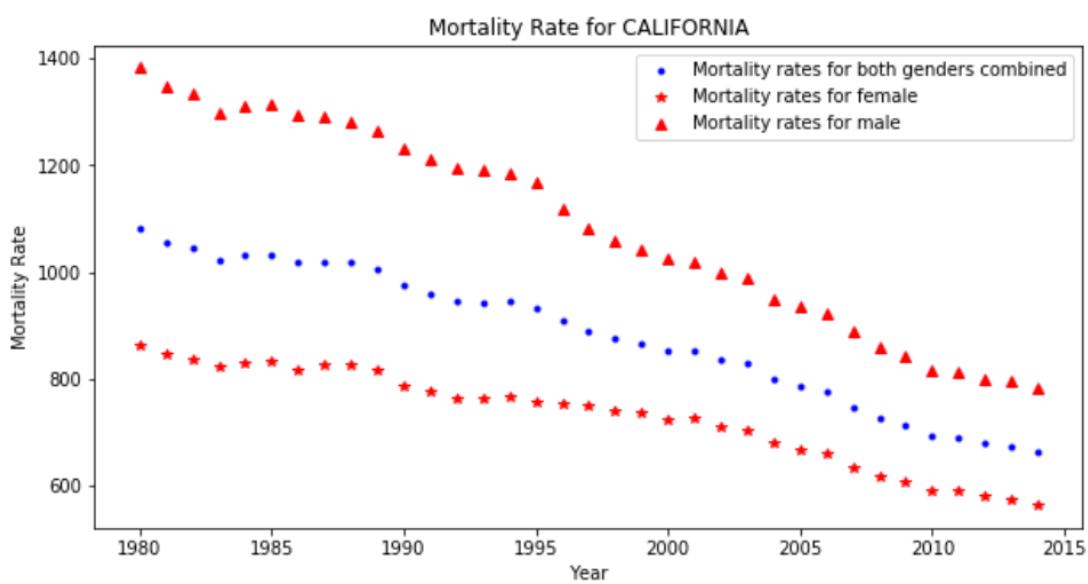
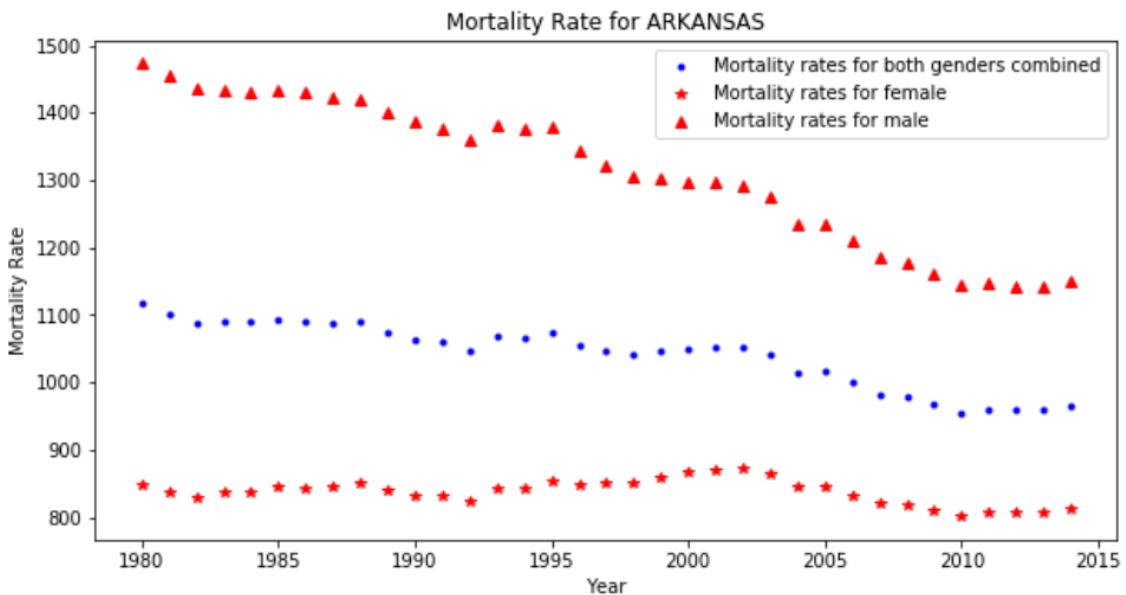
## Appendix 2: Trend of Mortality Rates in each state



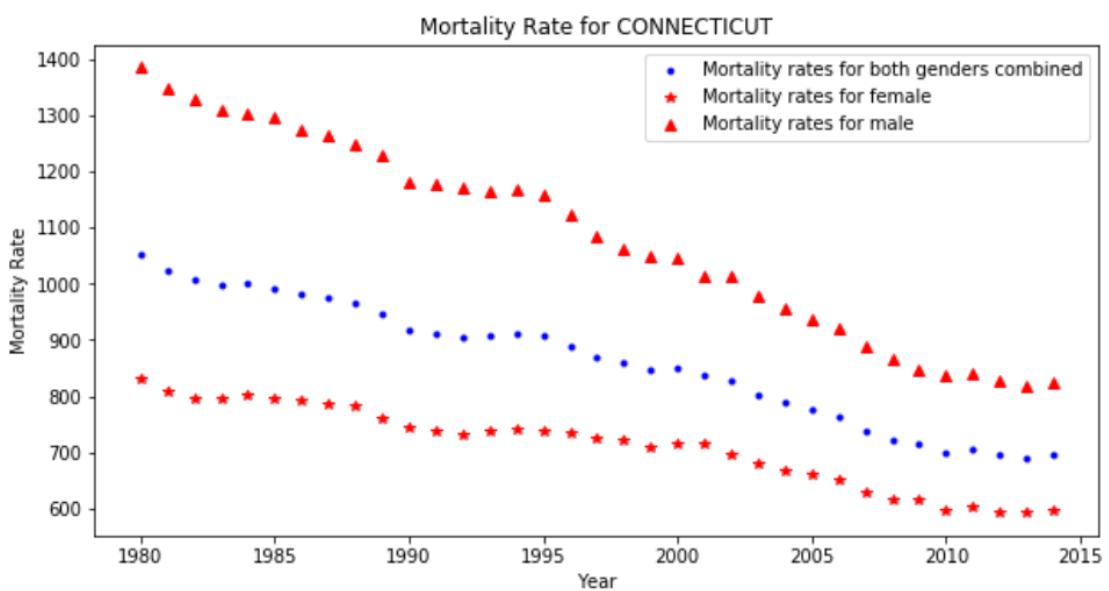
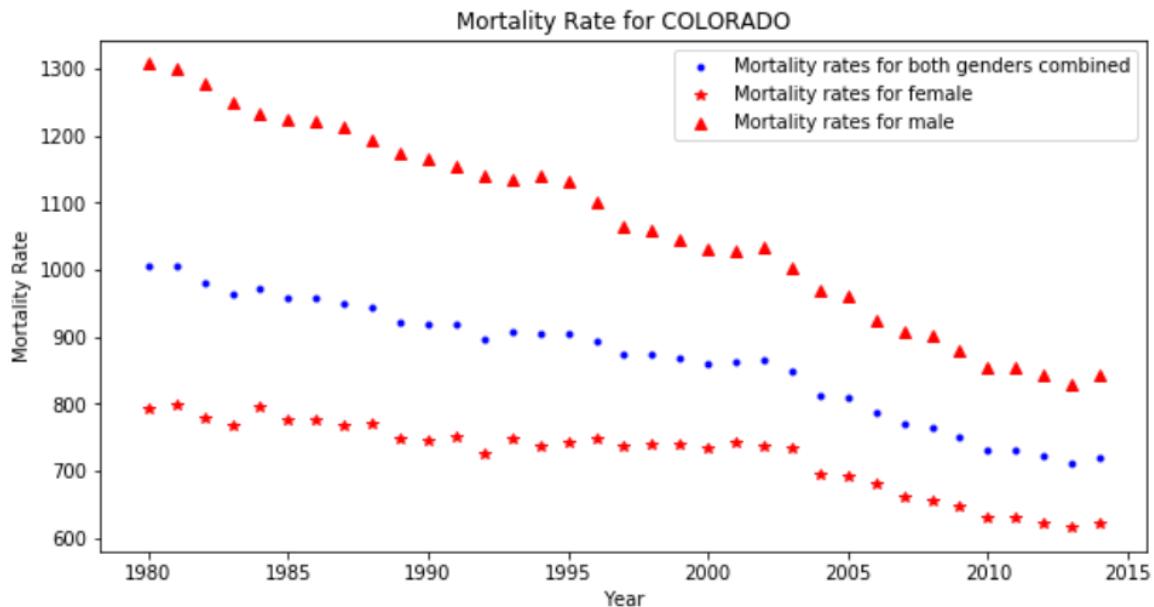
# DATA SCIENCE CASE STUDY



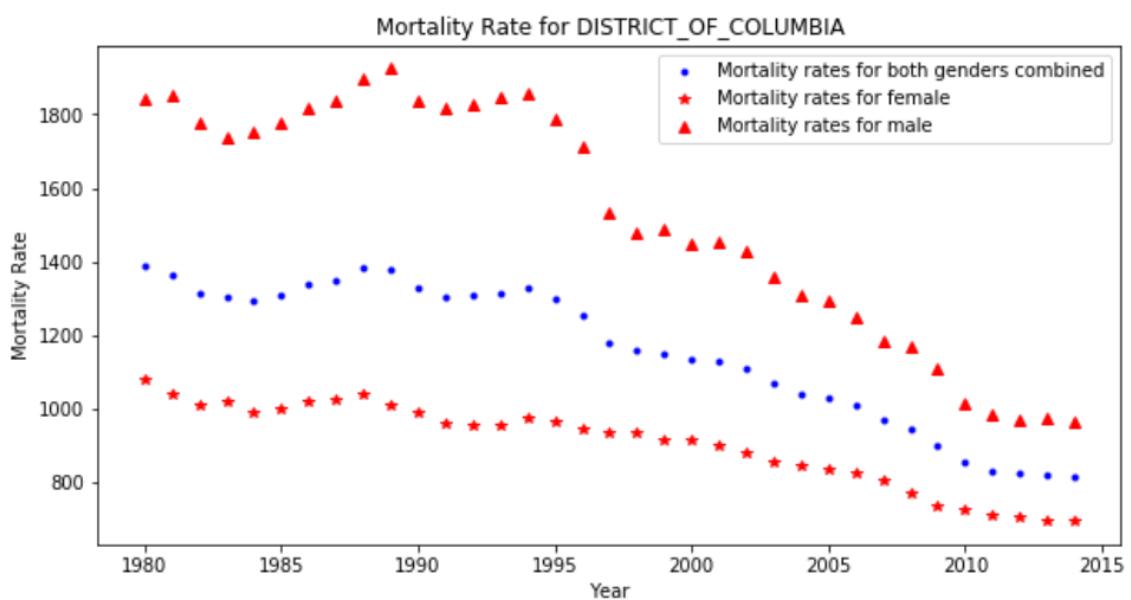
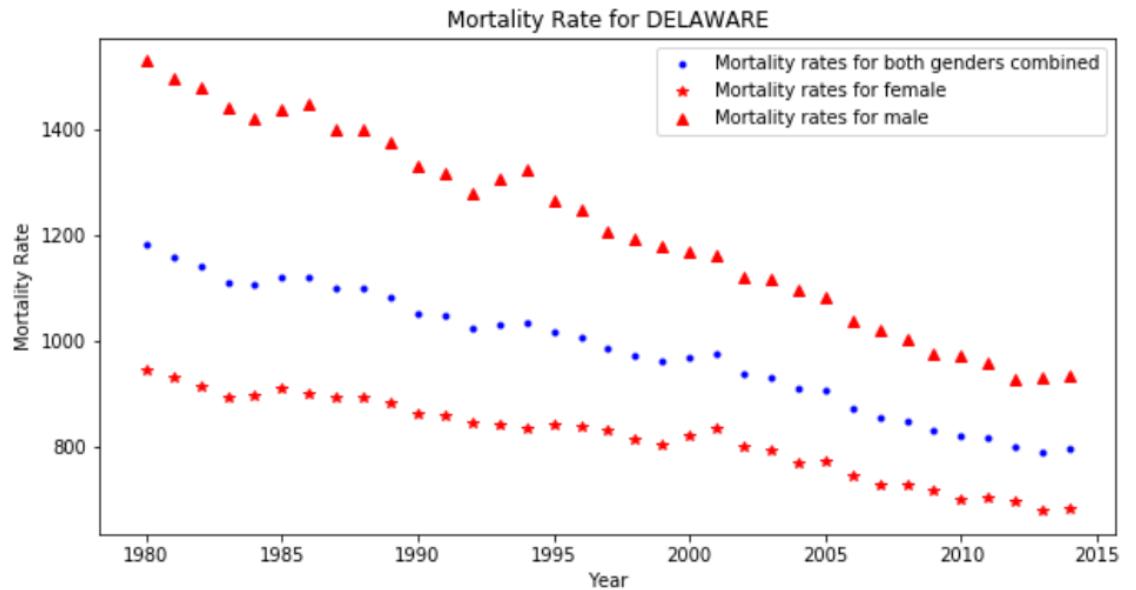
# DATA SCIENCE CASE STUDY



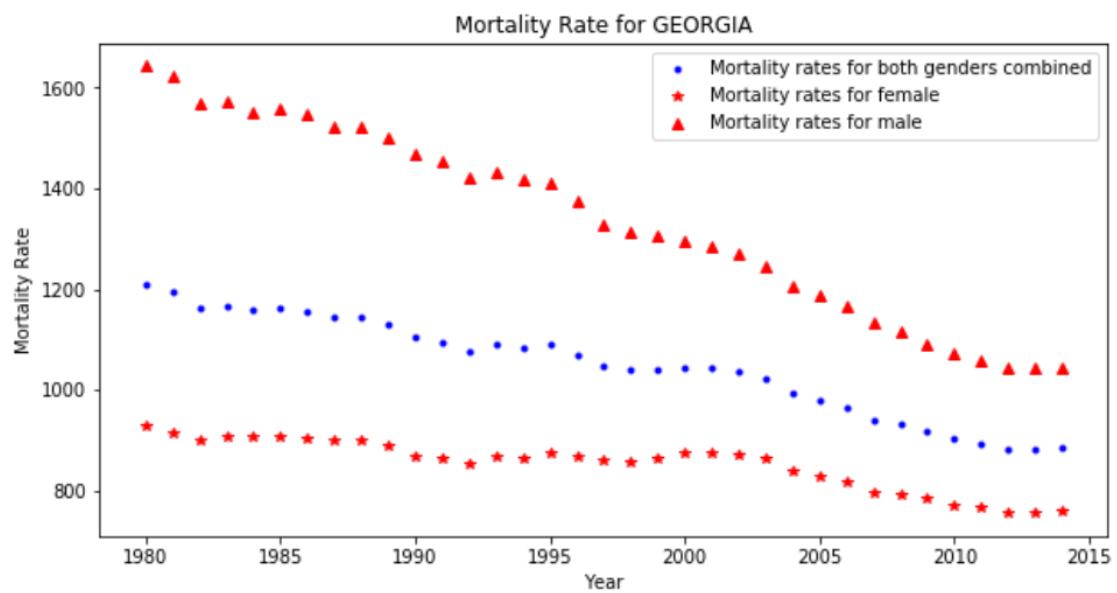
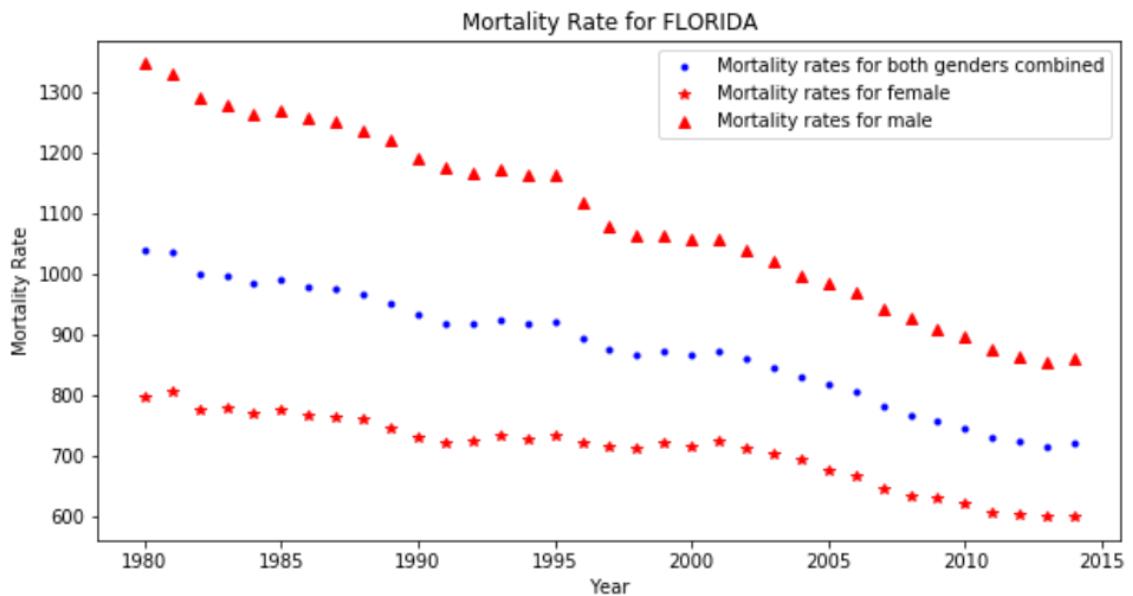
# DATA SCIENCE CASE STUDY



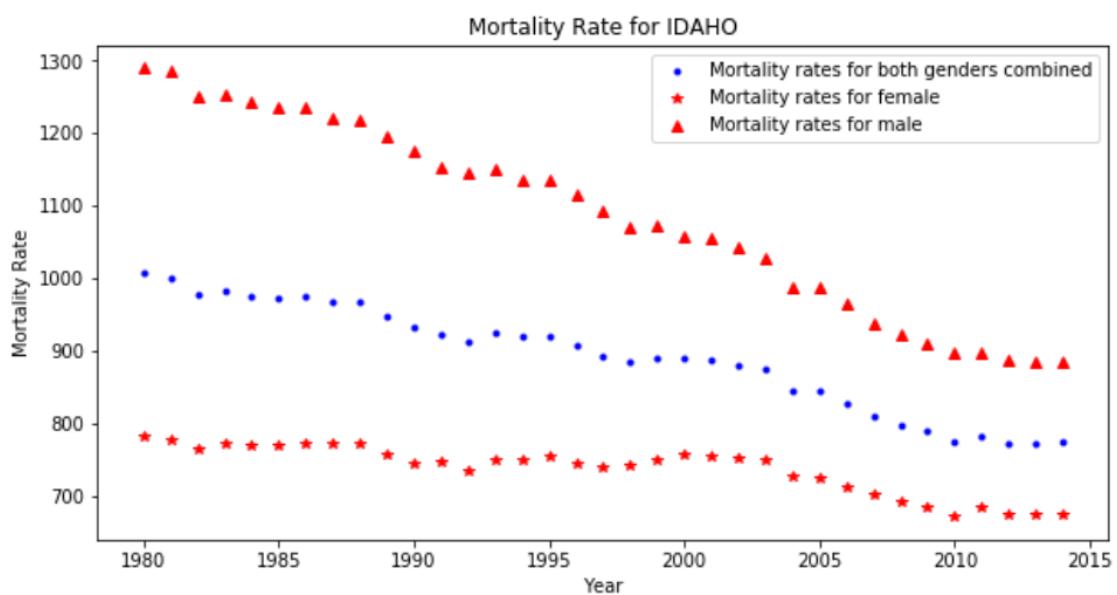
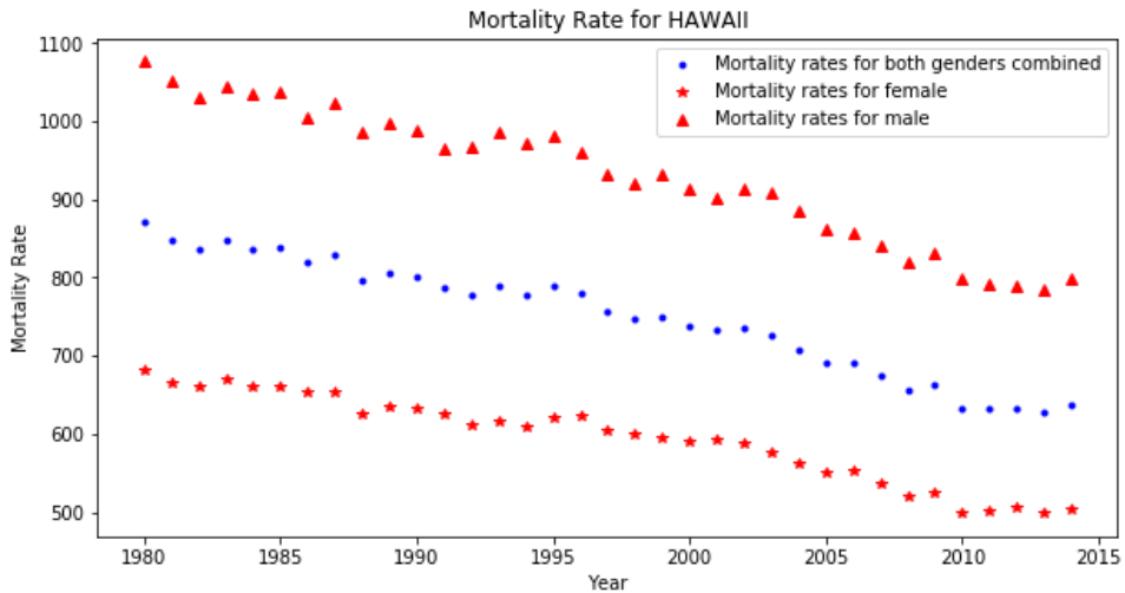
# DATA SCIENCE CASE STUDY



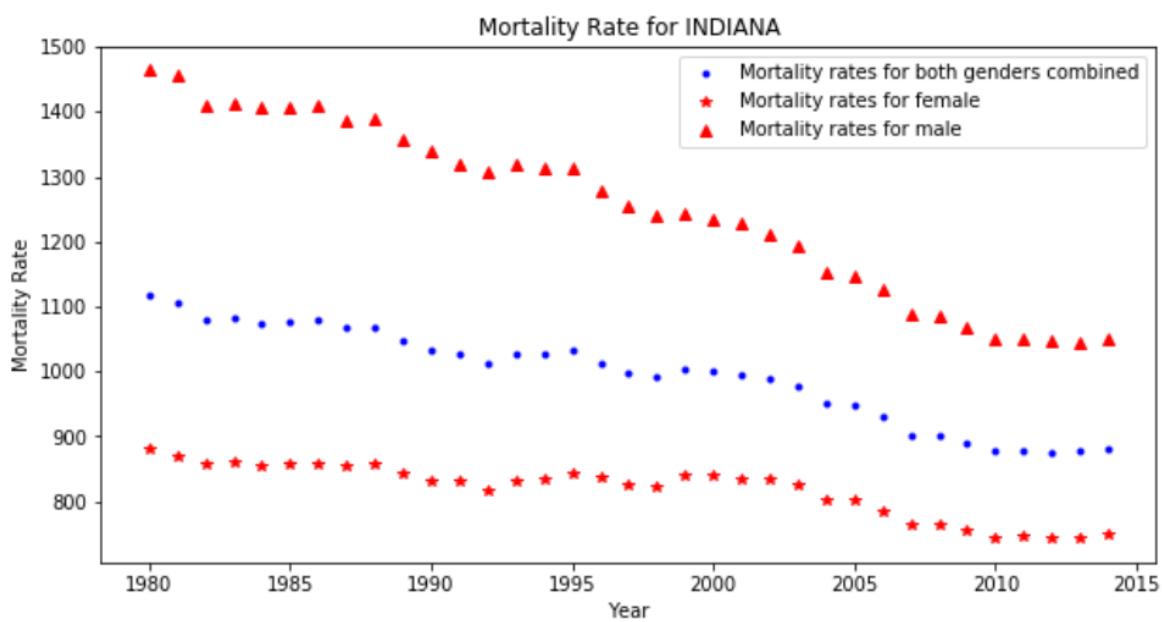
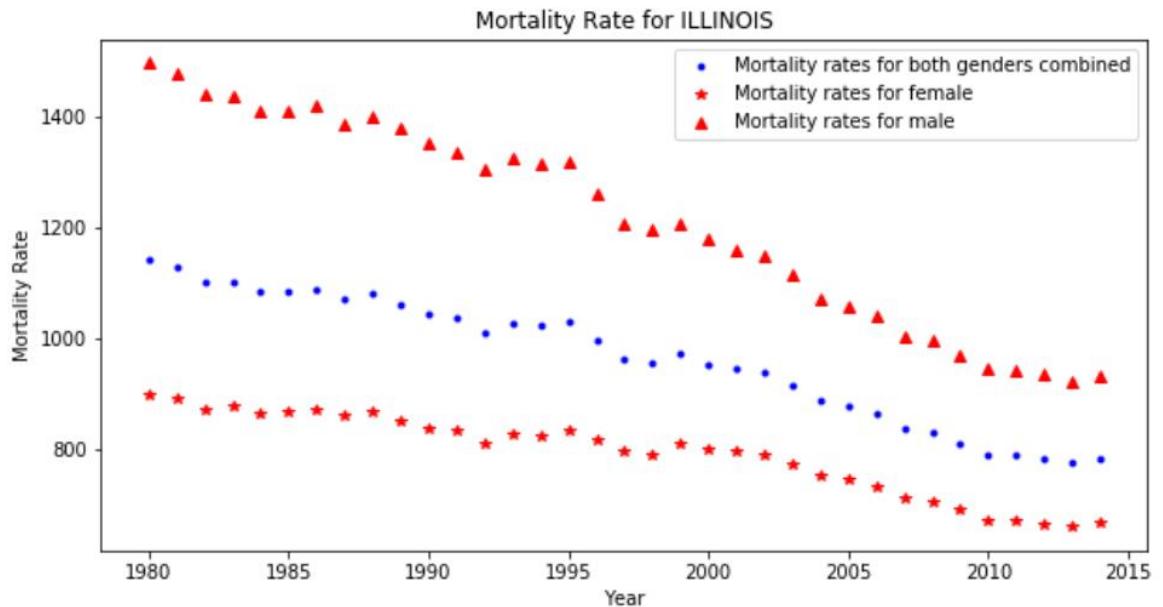
# DATA SCIENCE CASE STUDY



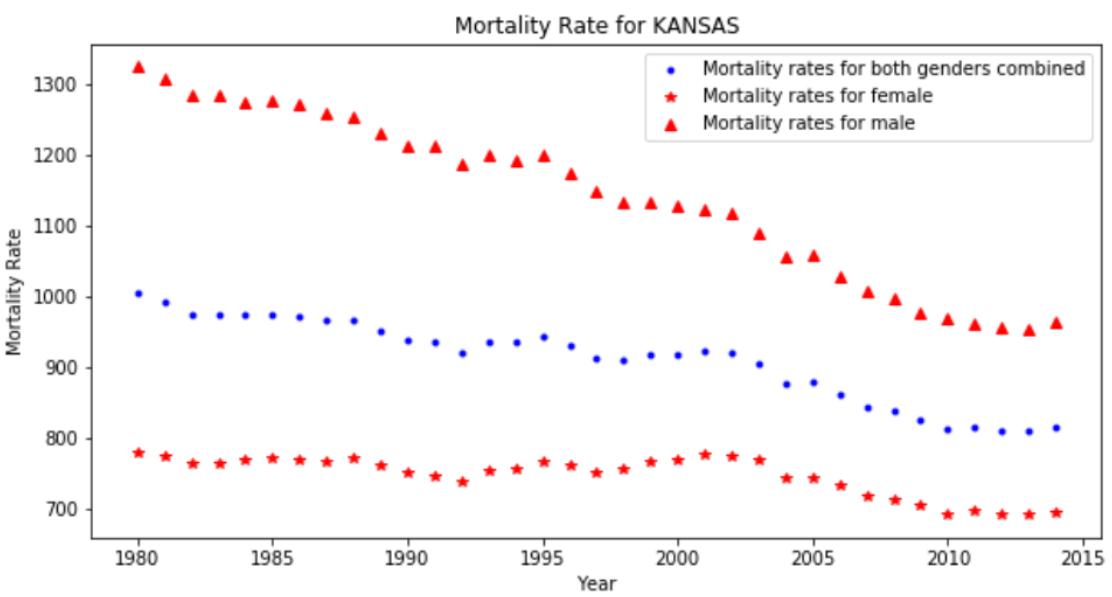
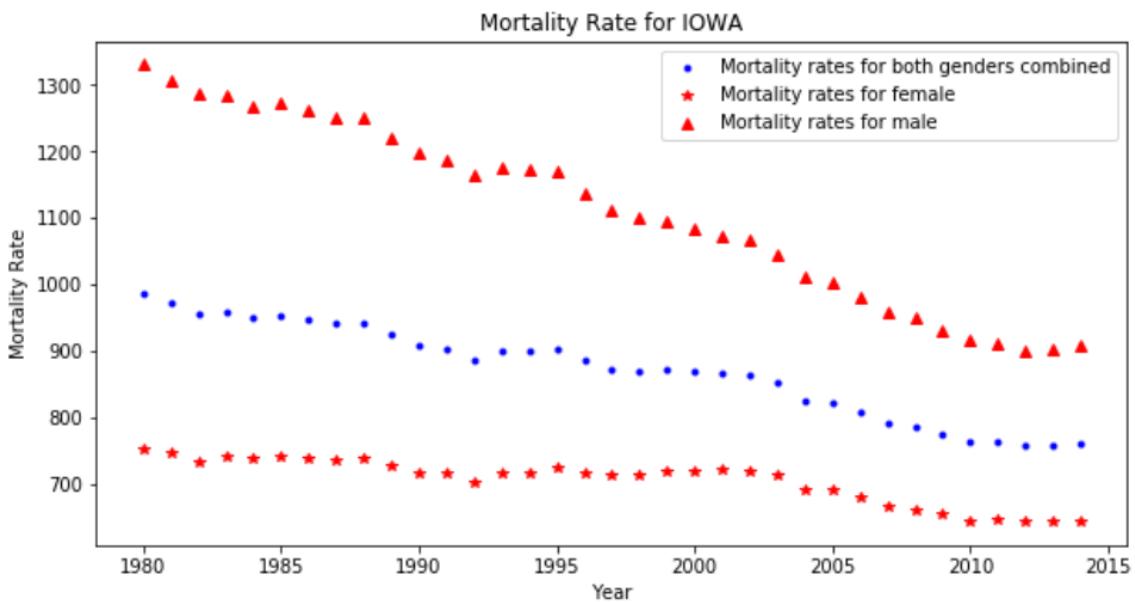
# DATA SCIENCE CASE STUDY



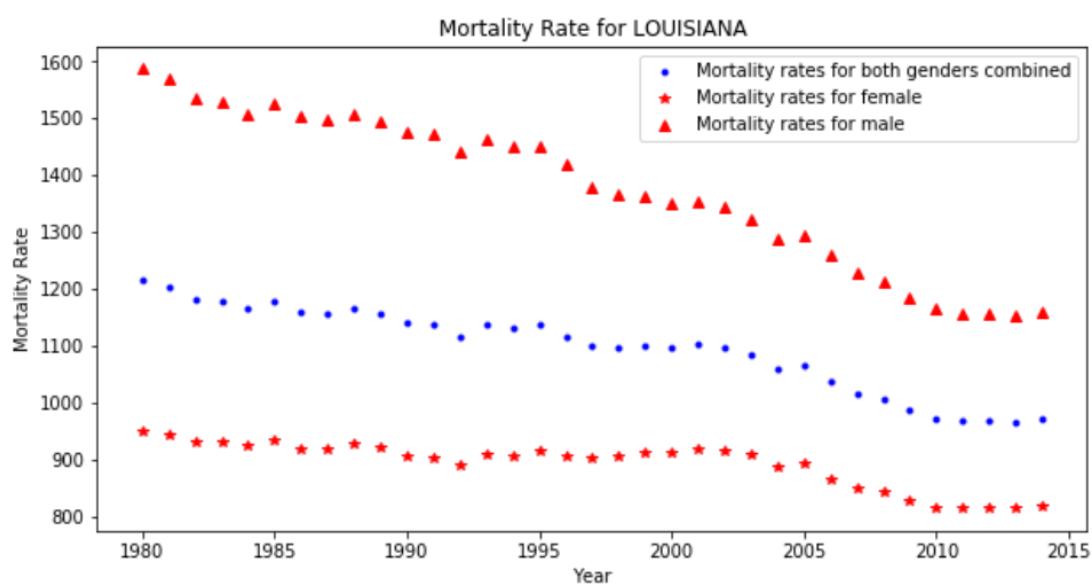
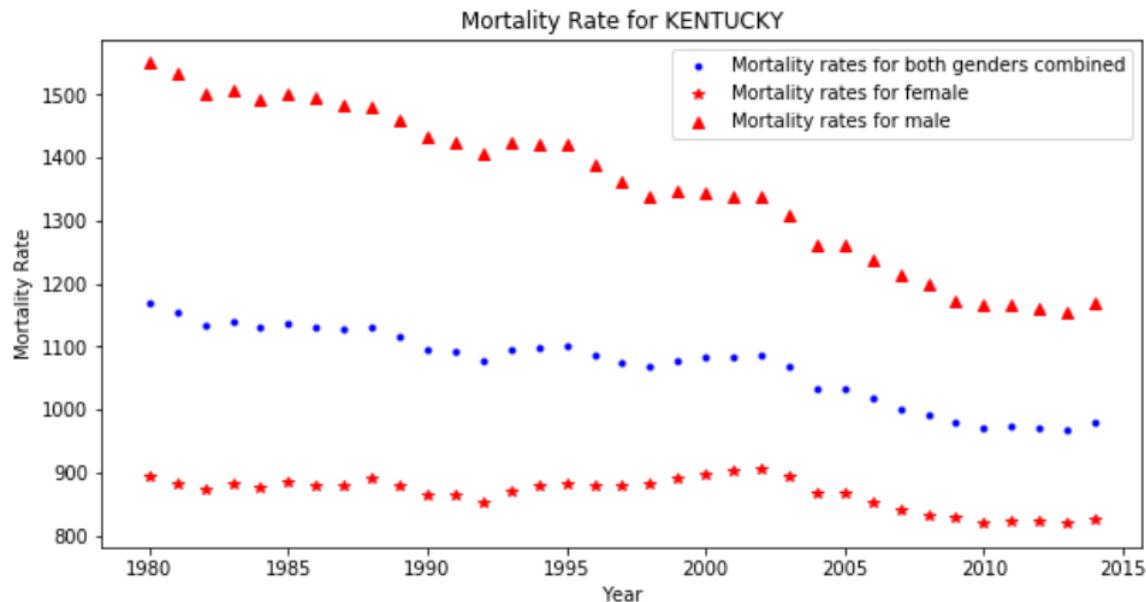
# DATA SCIENCE CASE STUDY



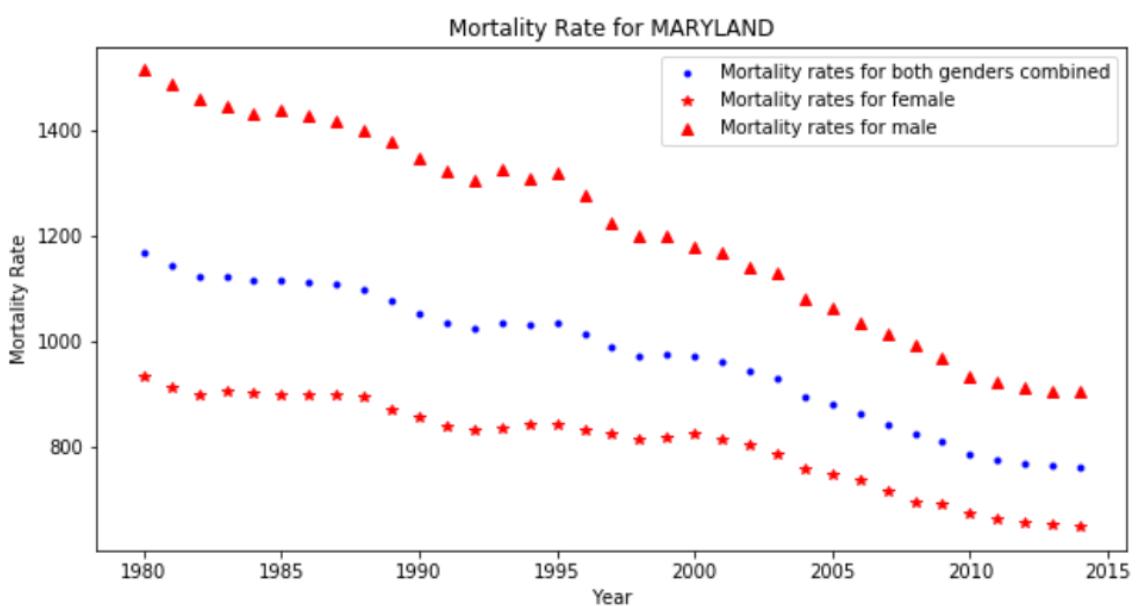
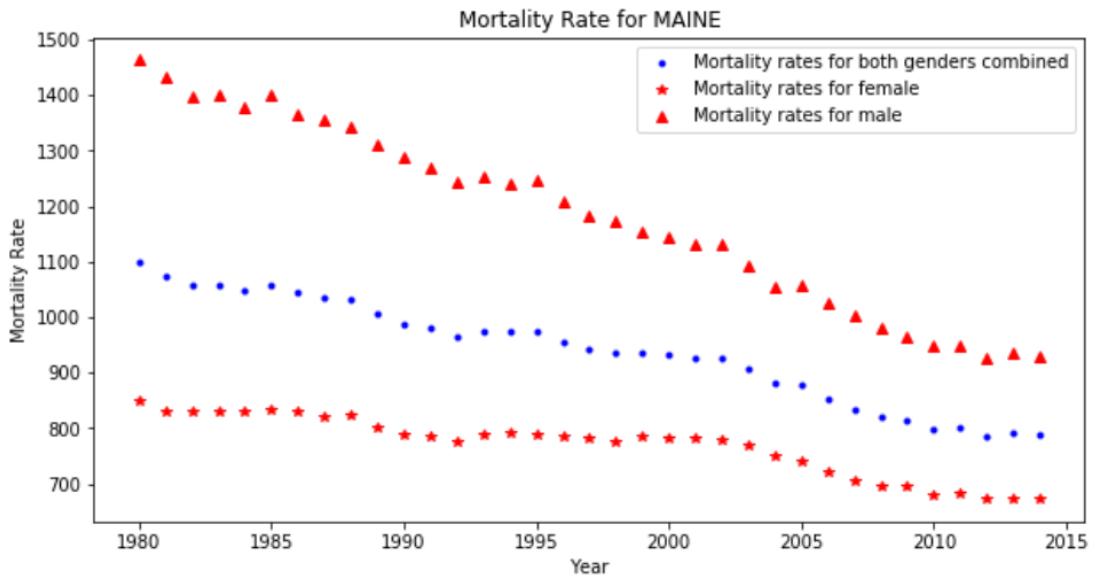
# DATA SCIENCE CASE STUDY



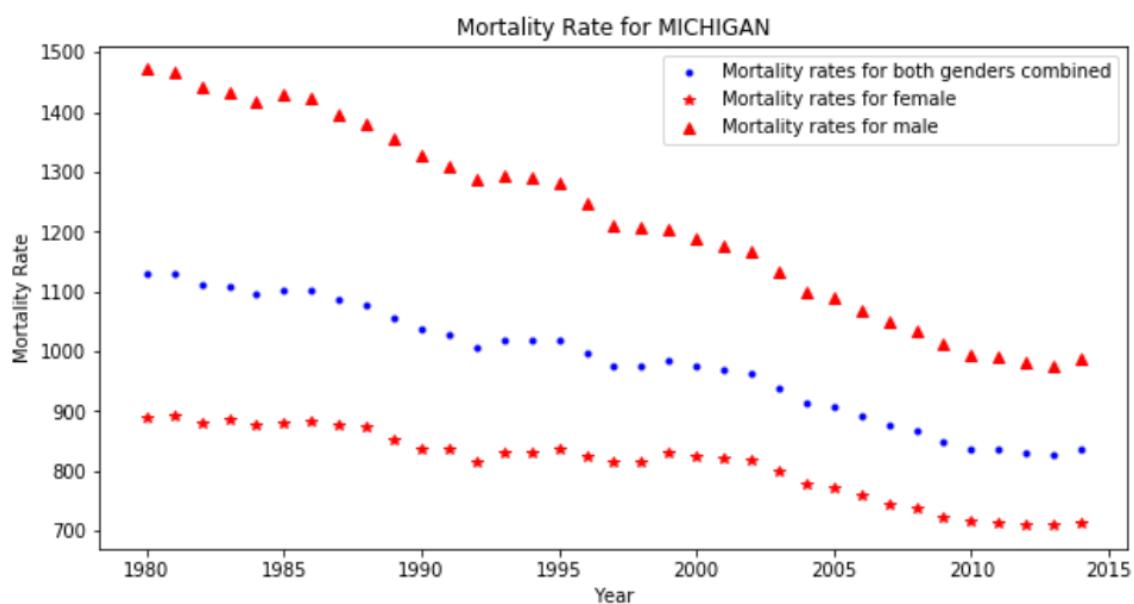
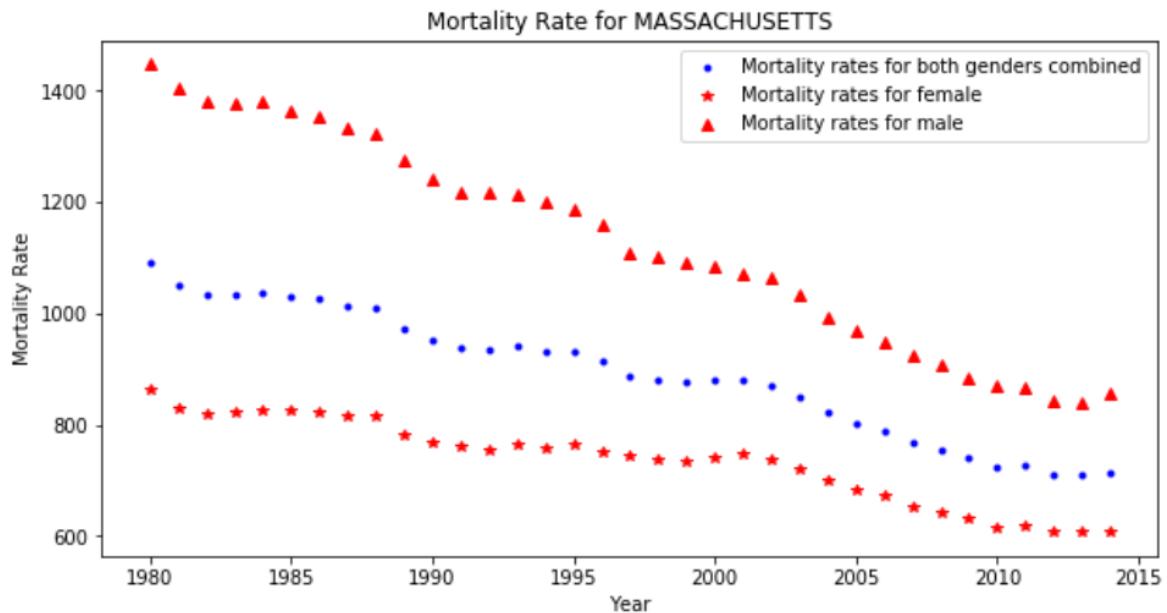
# DATA SCIENCE CASE STUDY



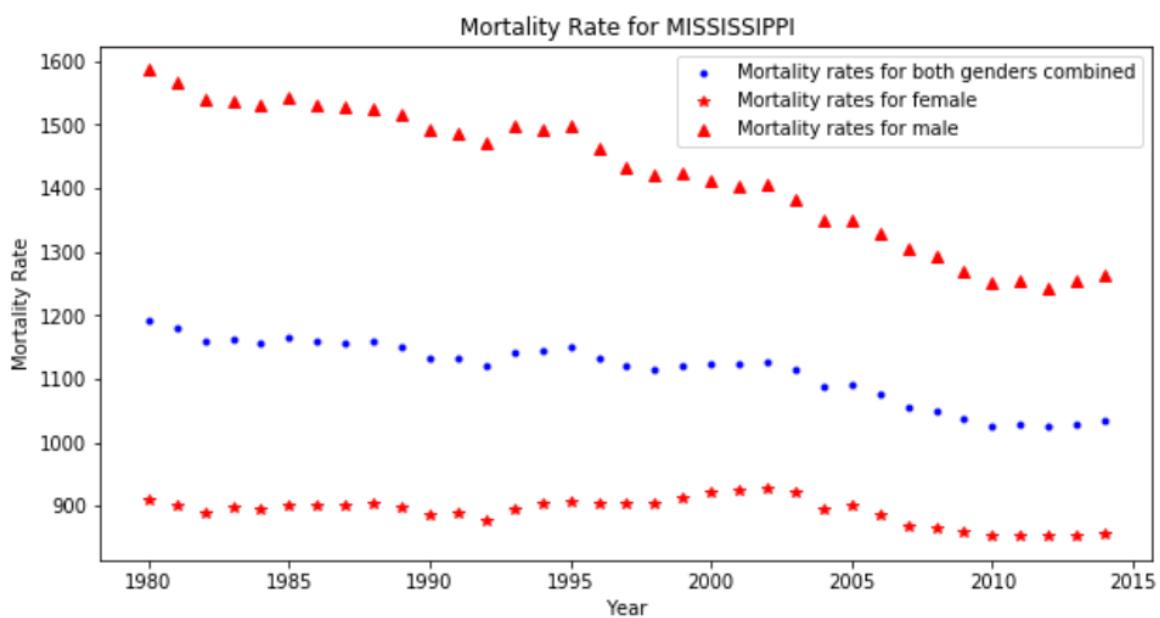
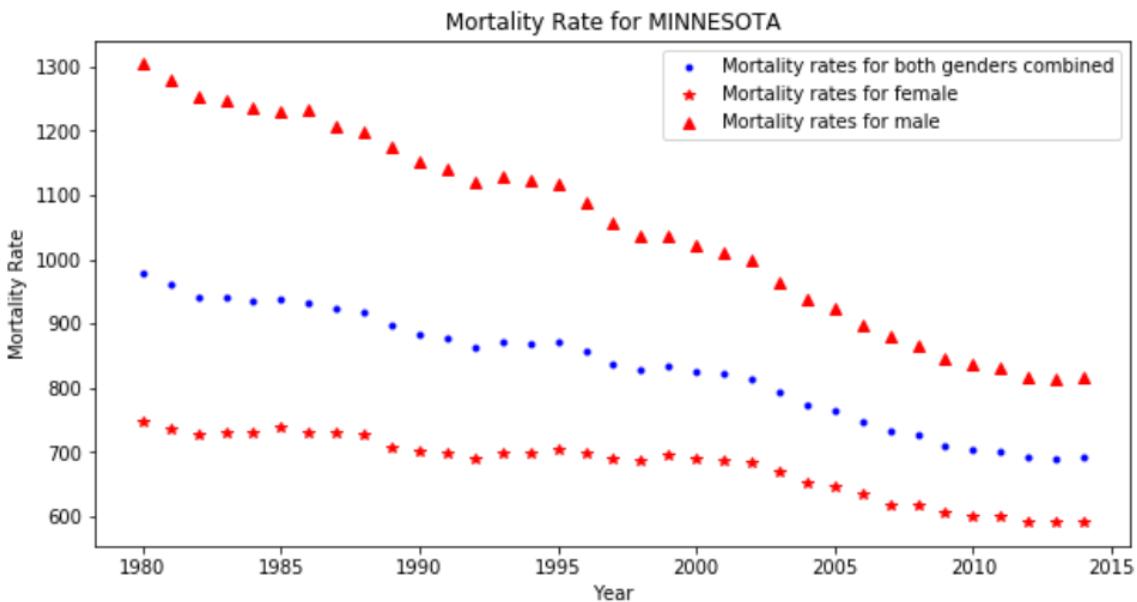
# DATA SCIENCE CASE STUDY



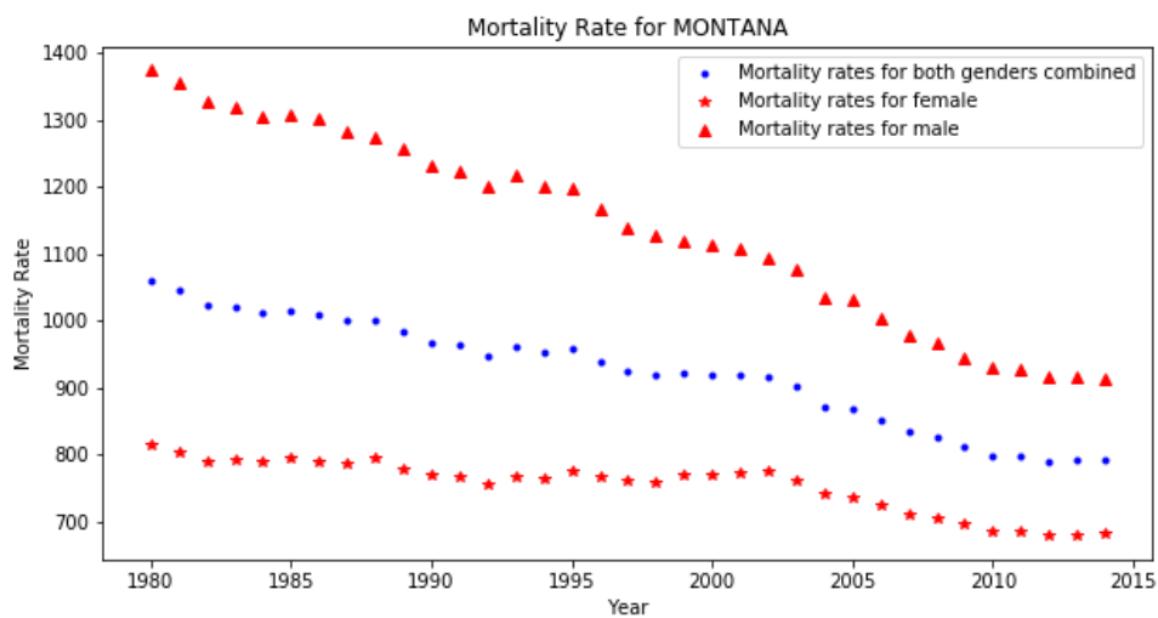
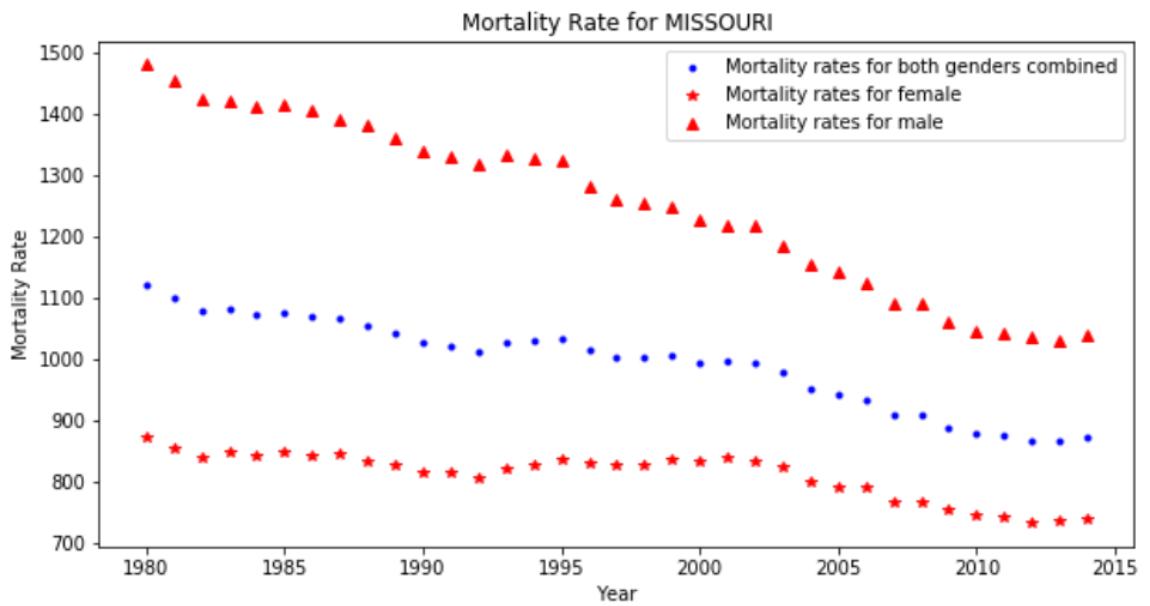
# DATA SCIENCE CASE STUDY



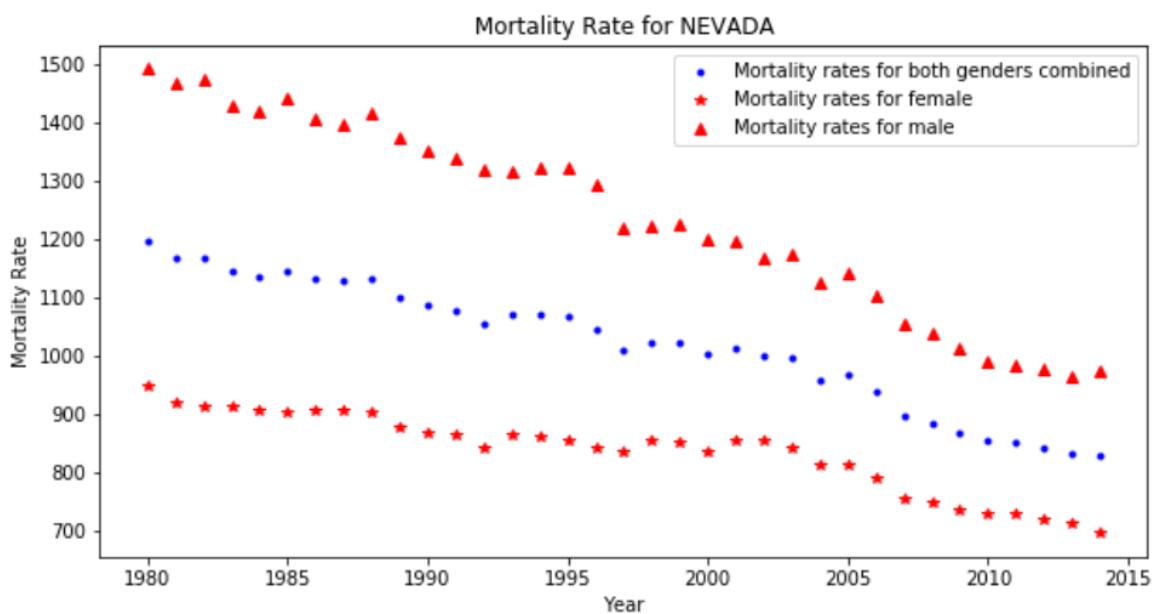
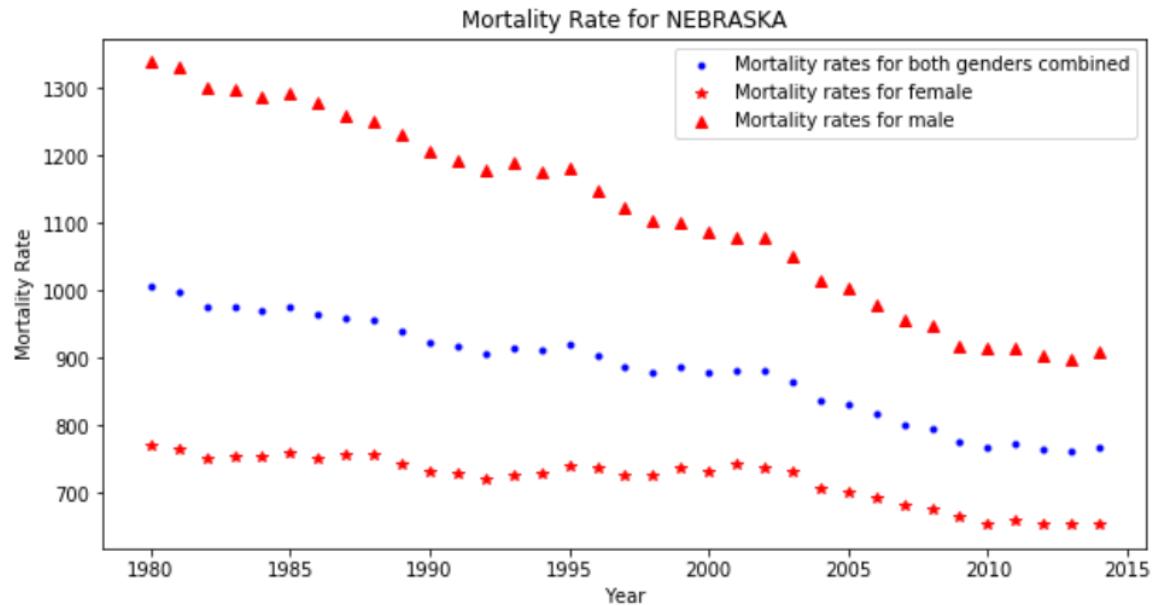
# DATA SCIENCE CASE STUDY



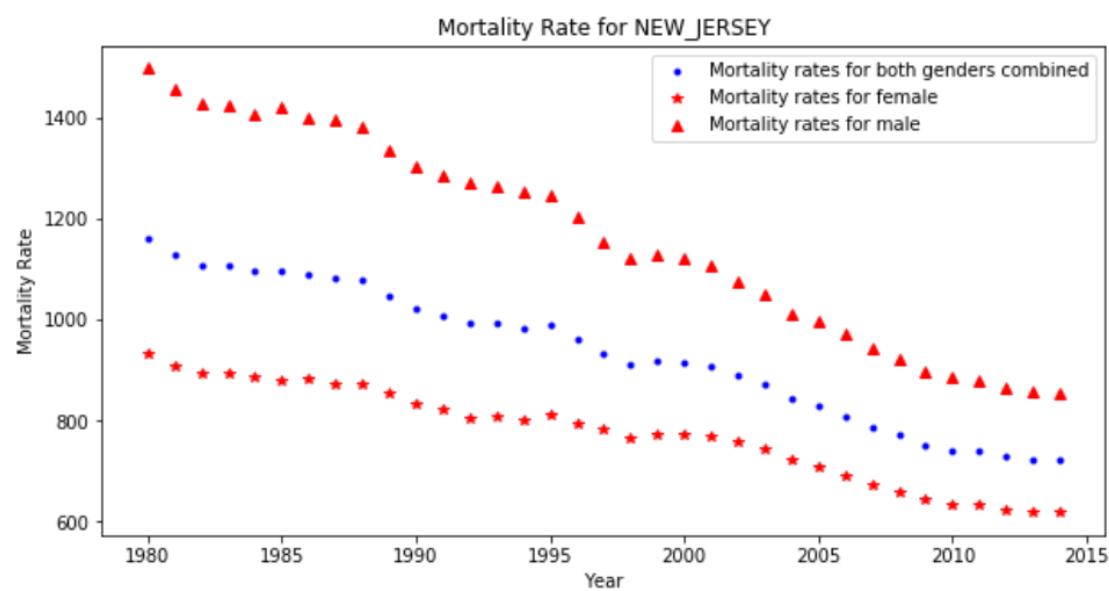
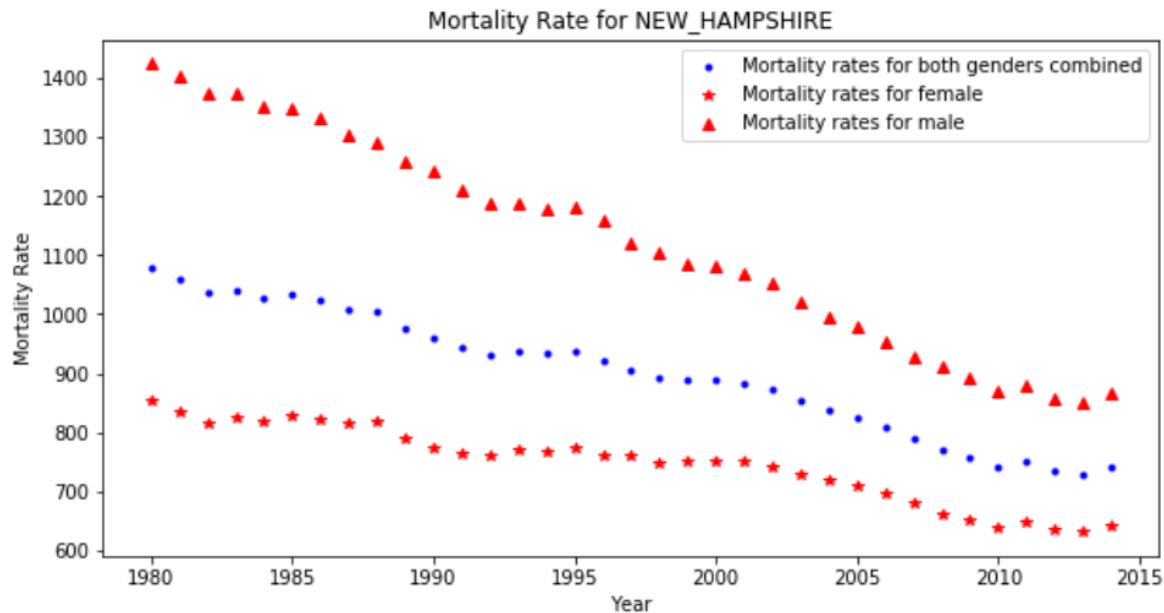
# DATA SCIENCE CASE STUDY



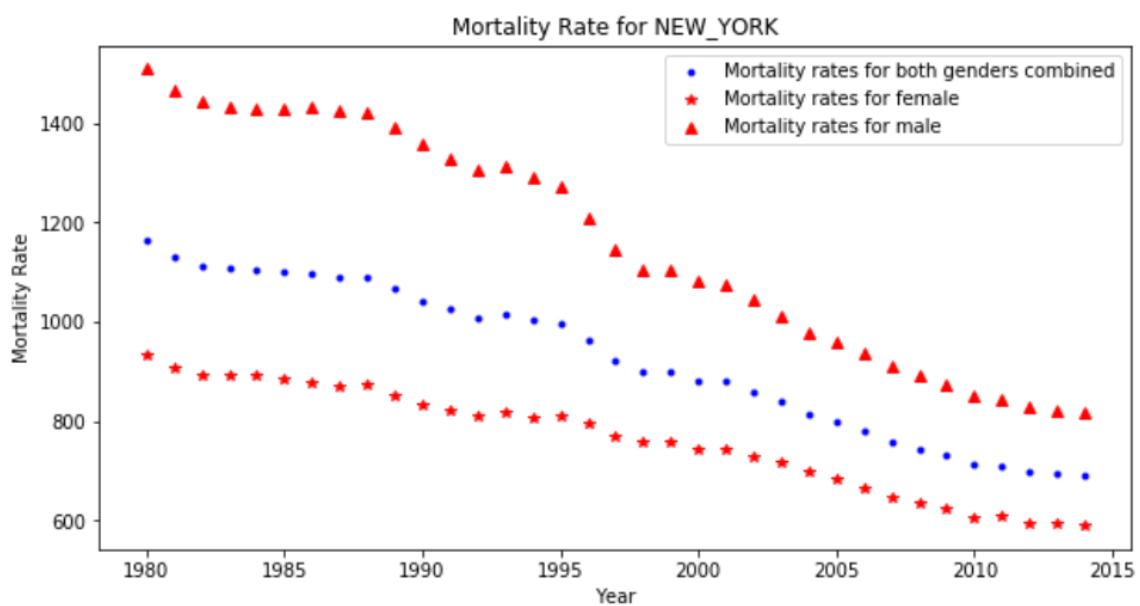
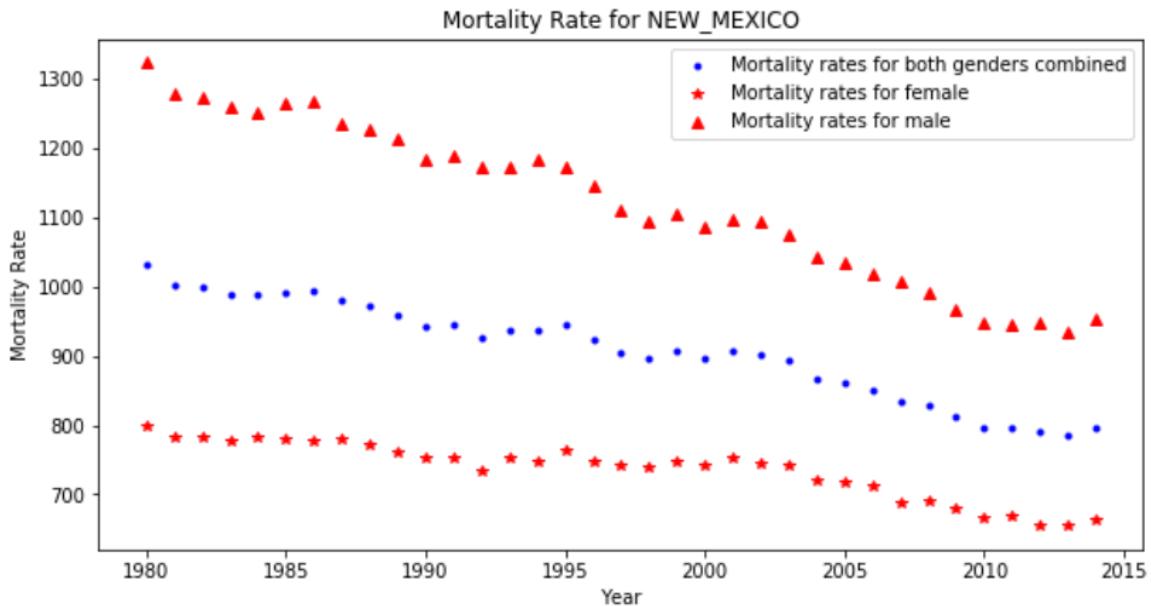
# DATA SCIENCE CASE STUDY



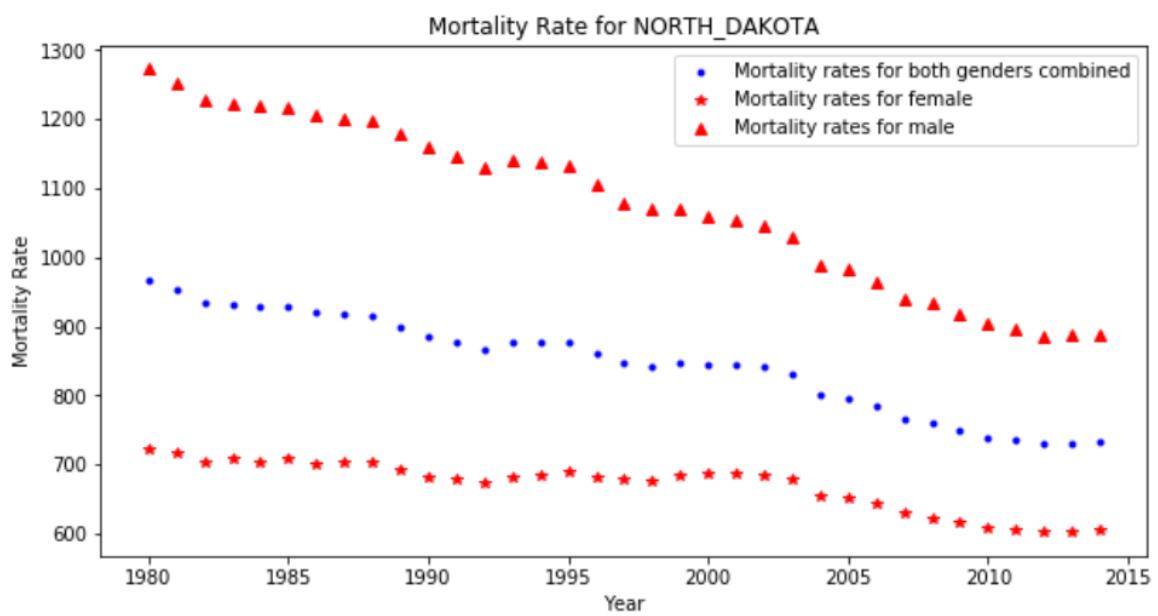
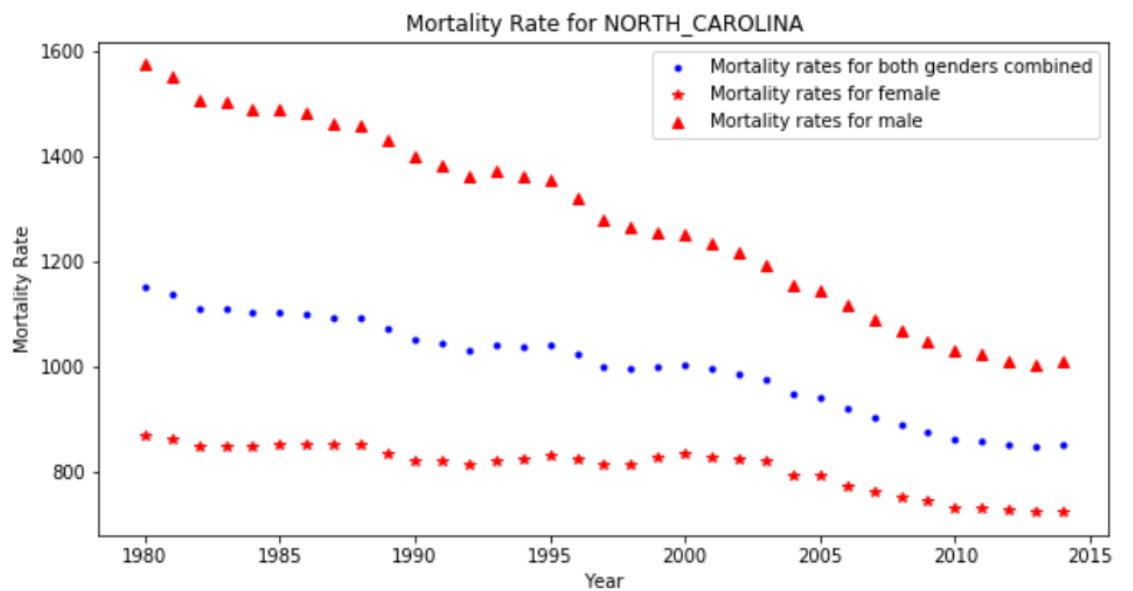
# DATA SCIENCE CASE STUDY



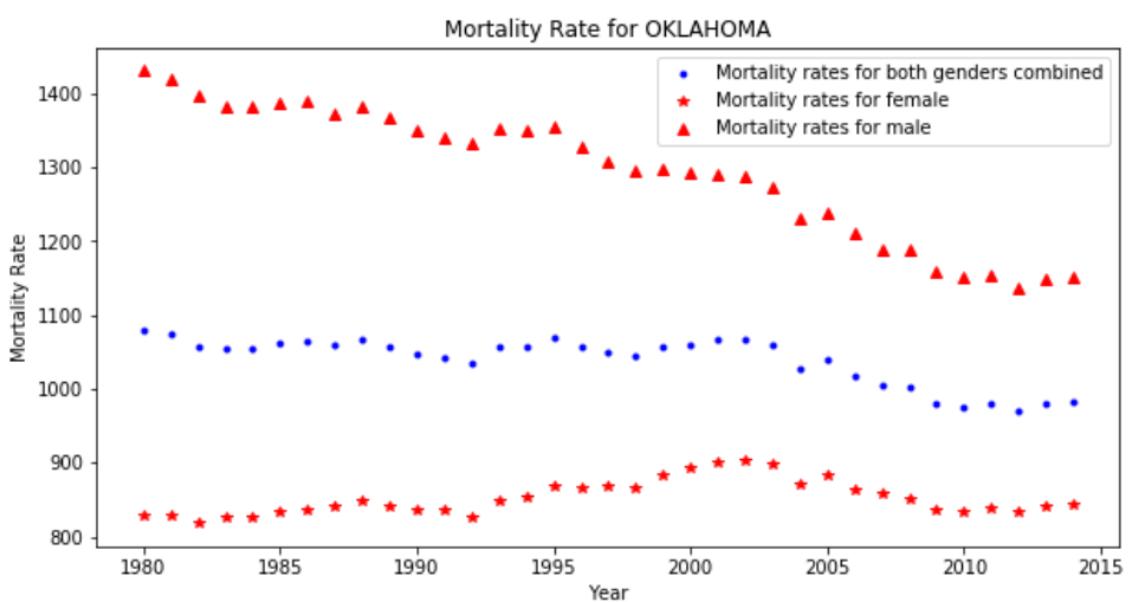
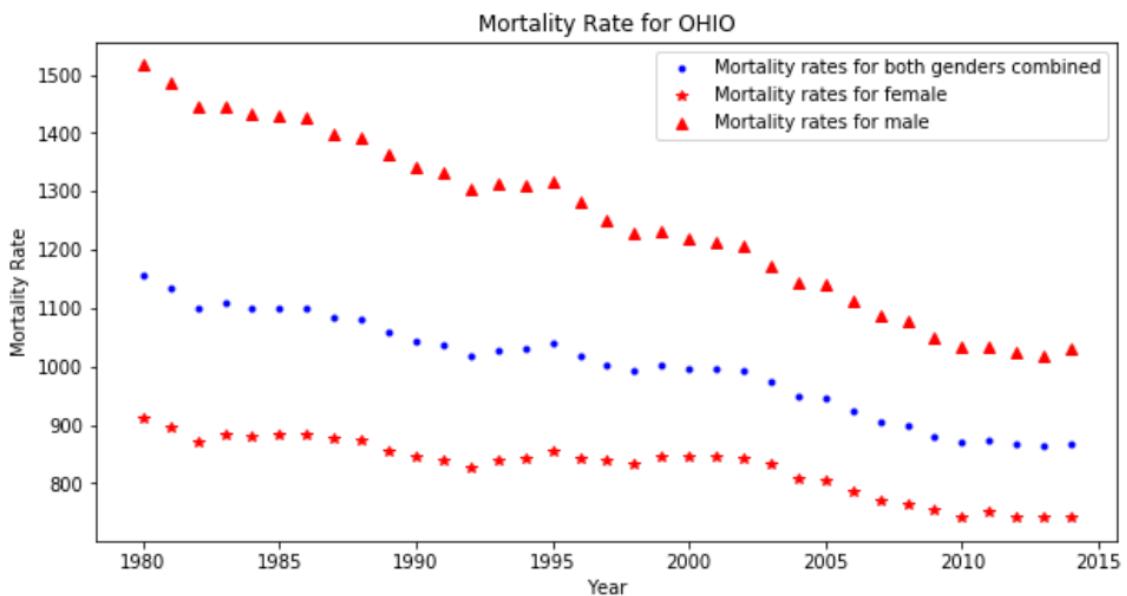
# DATA SCIENCE CASE STUDY



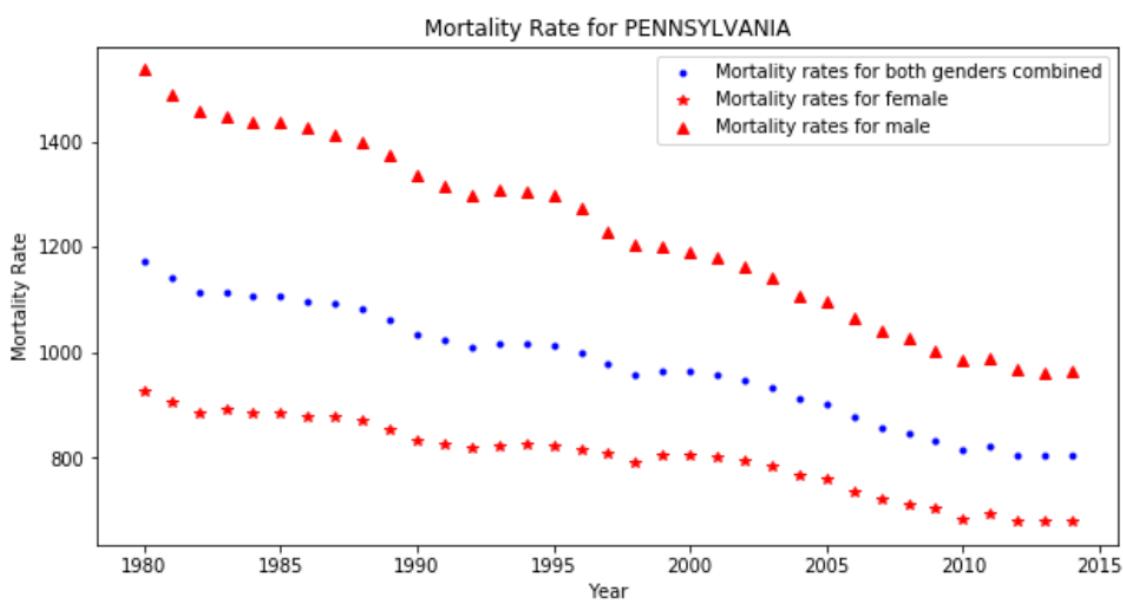
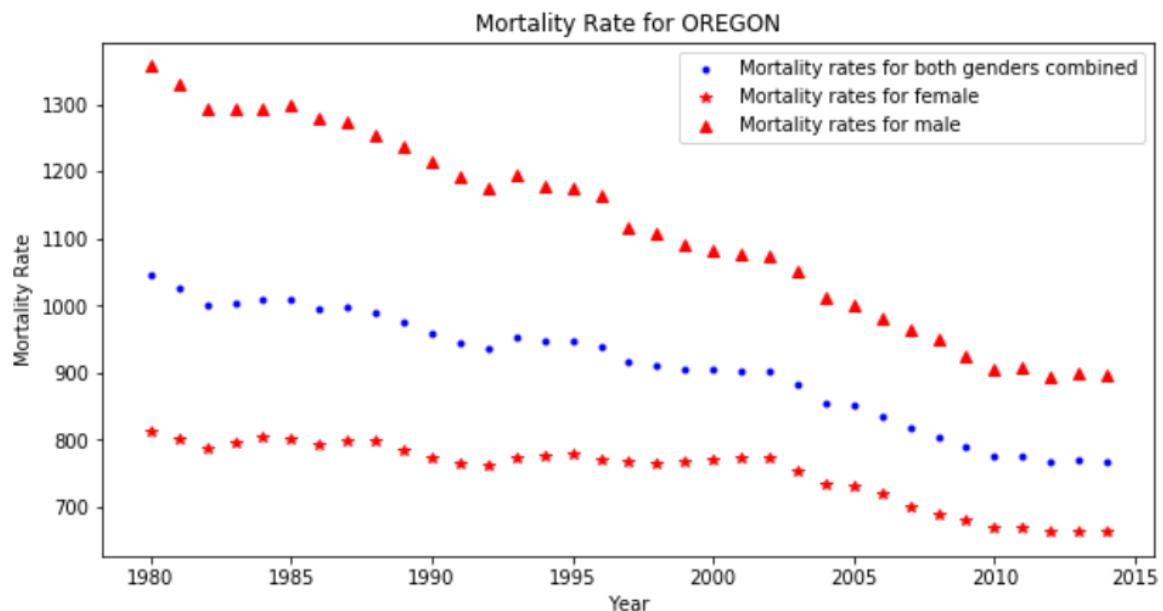
# DATA SCIENCE CASE STUDY



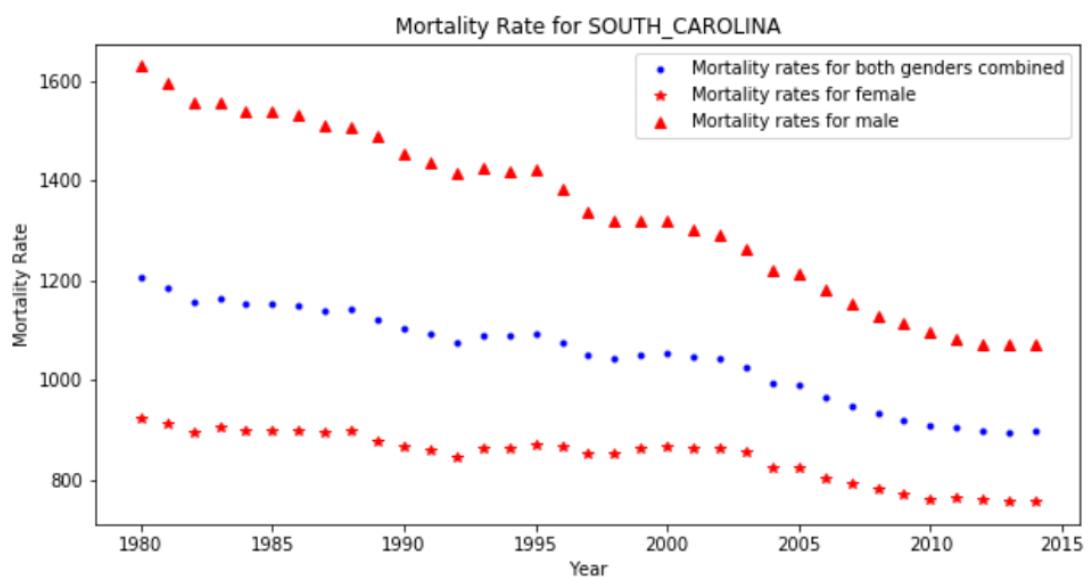
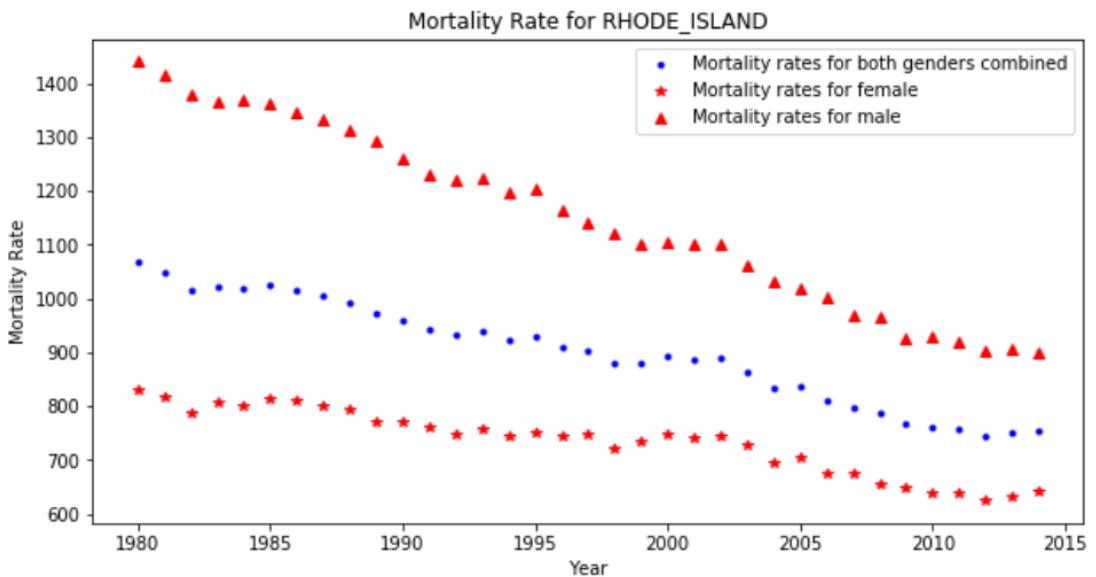
# DATA SCIENCE CASE STUDY



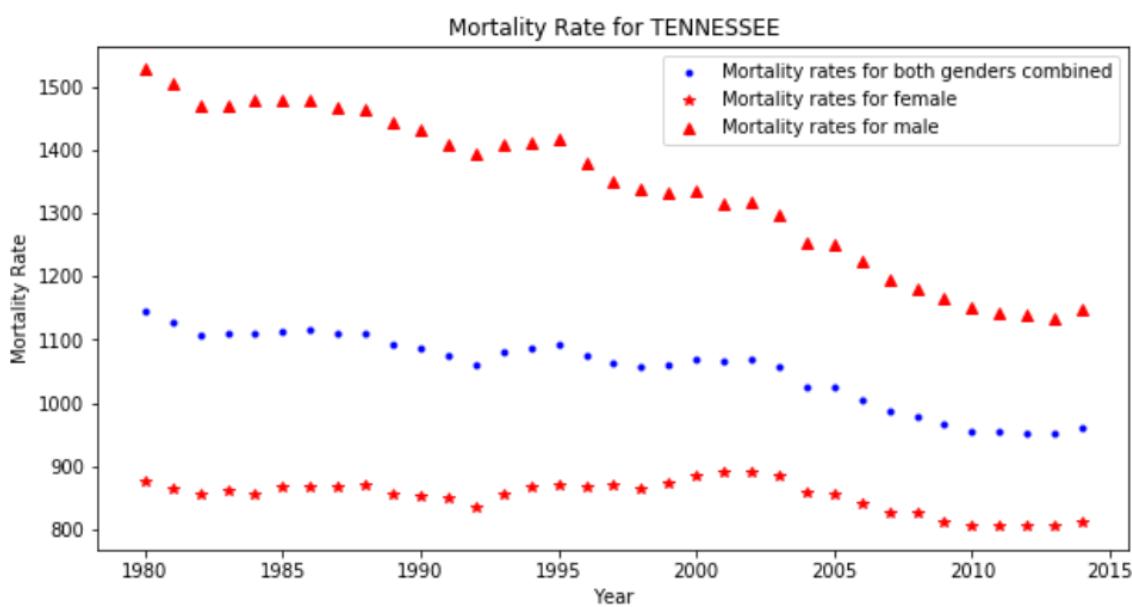
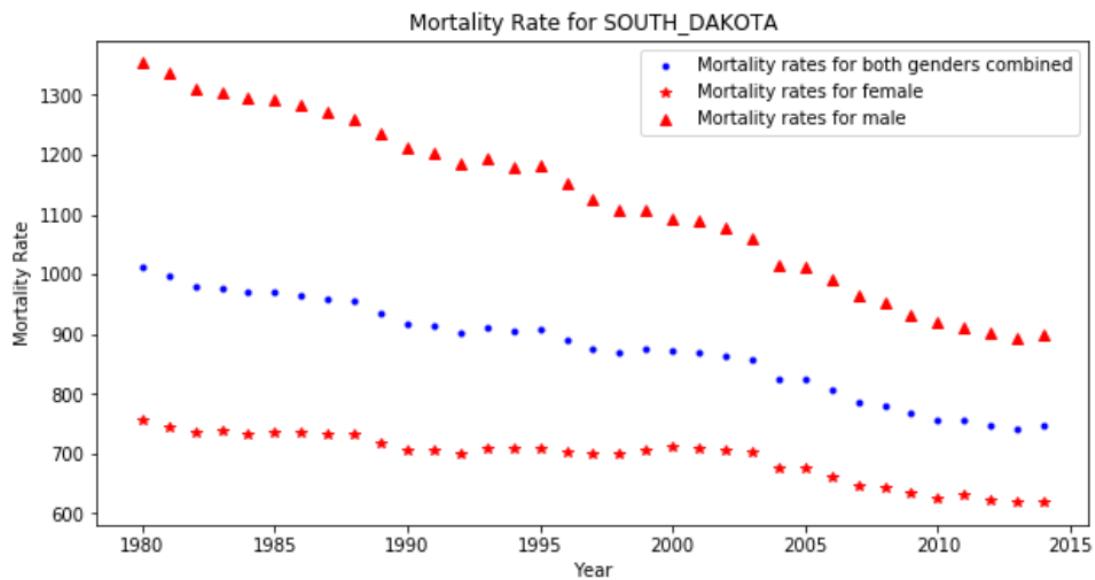
# DATA SCIENCE CASE STUDY



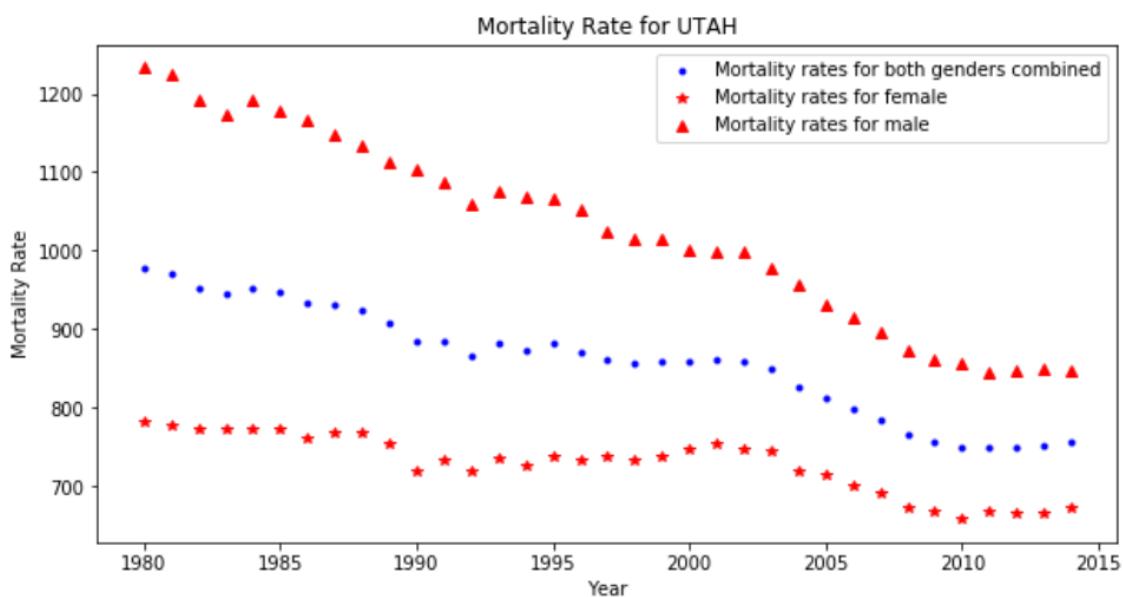
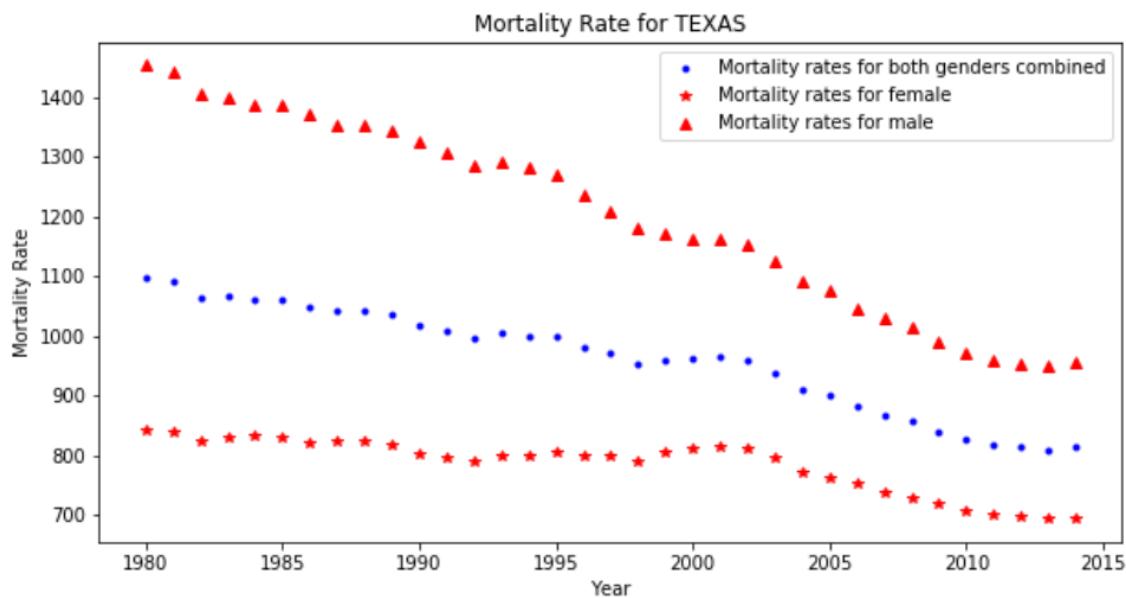
# DATA SCIENCE CASE STUDY



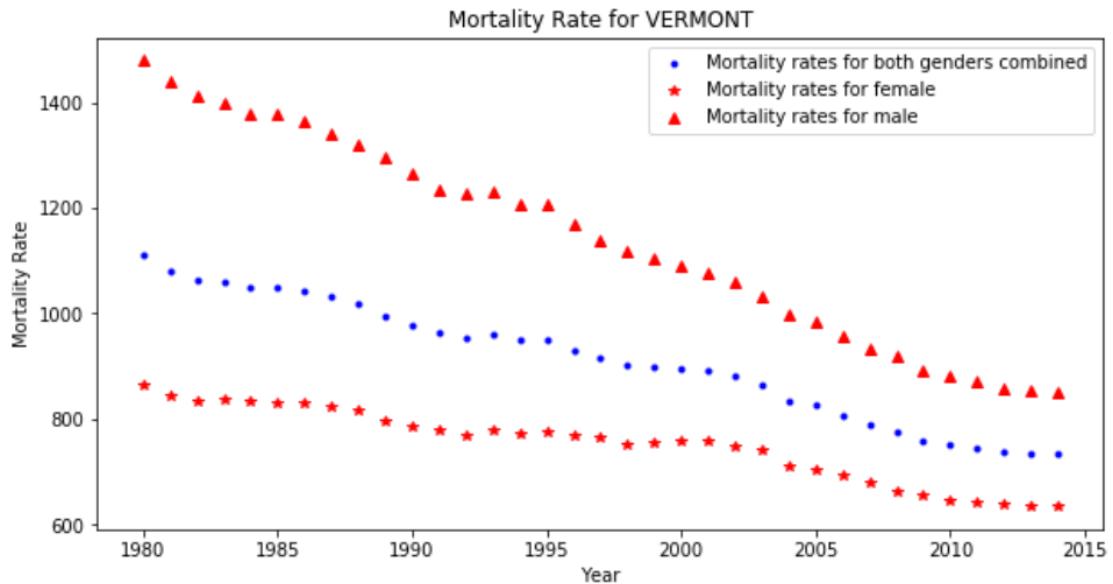
# DATA SCIENCE CASE STUDY



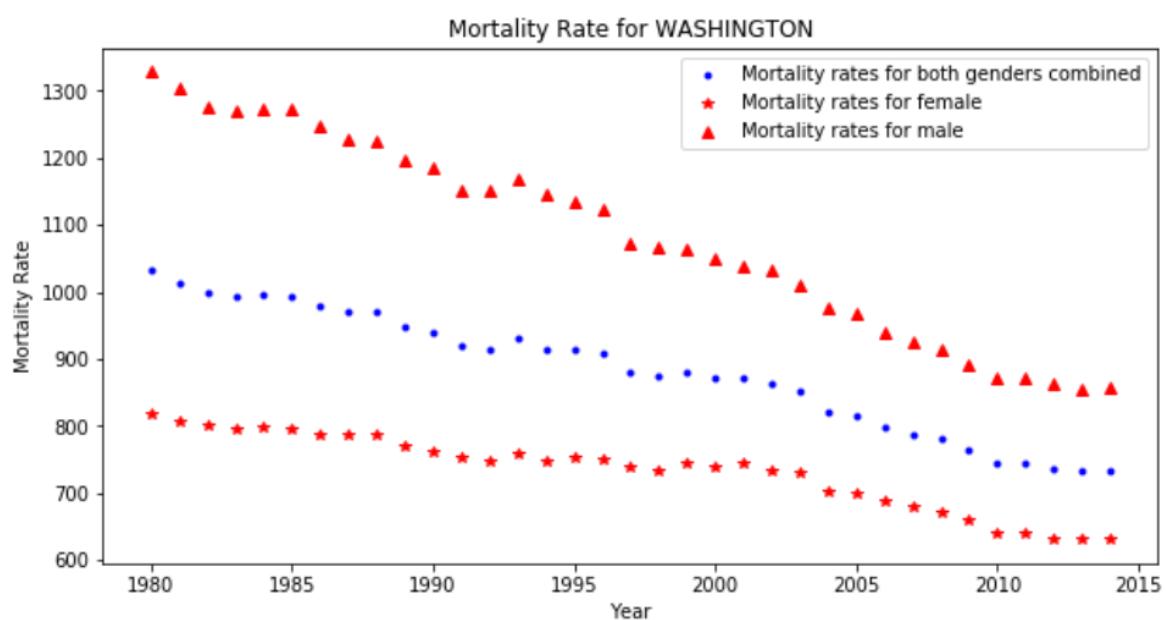
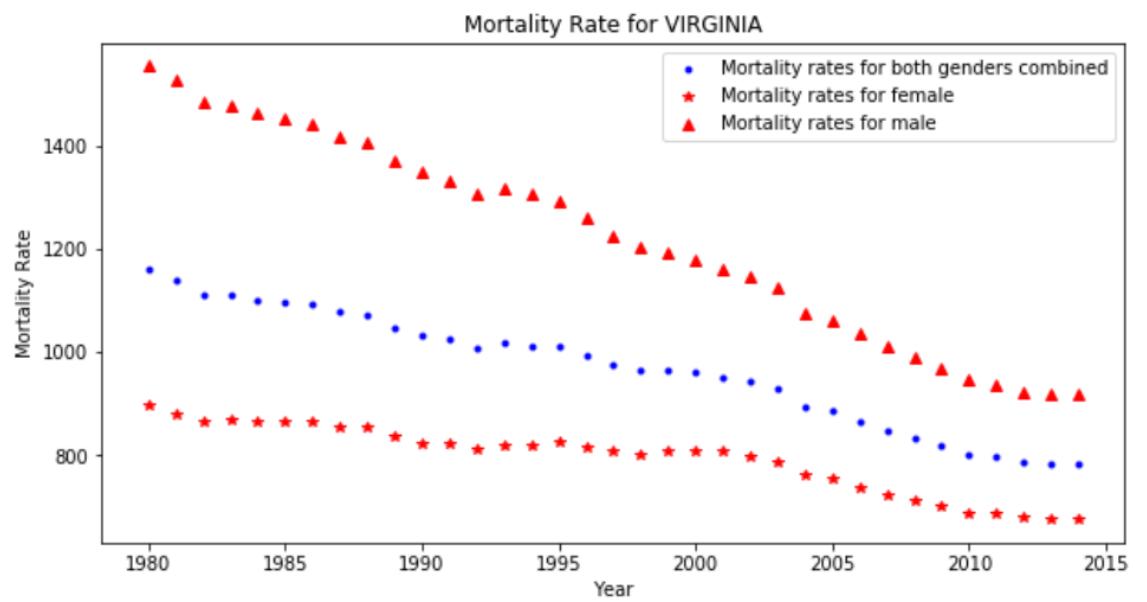
# DATA SCIENCE CASE STUDY



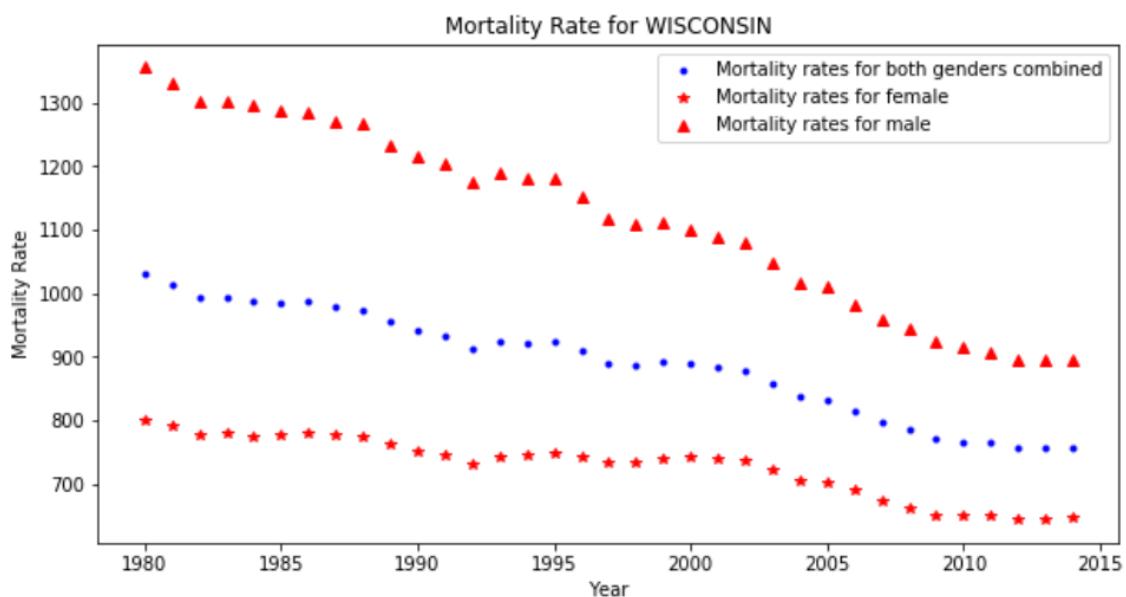
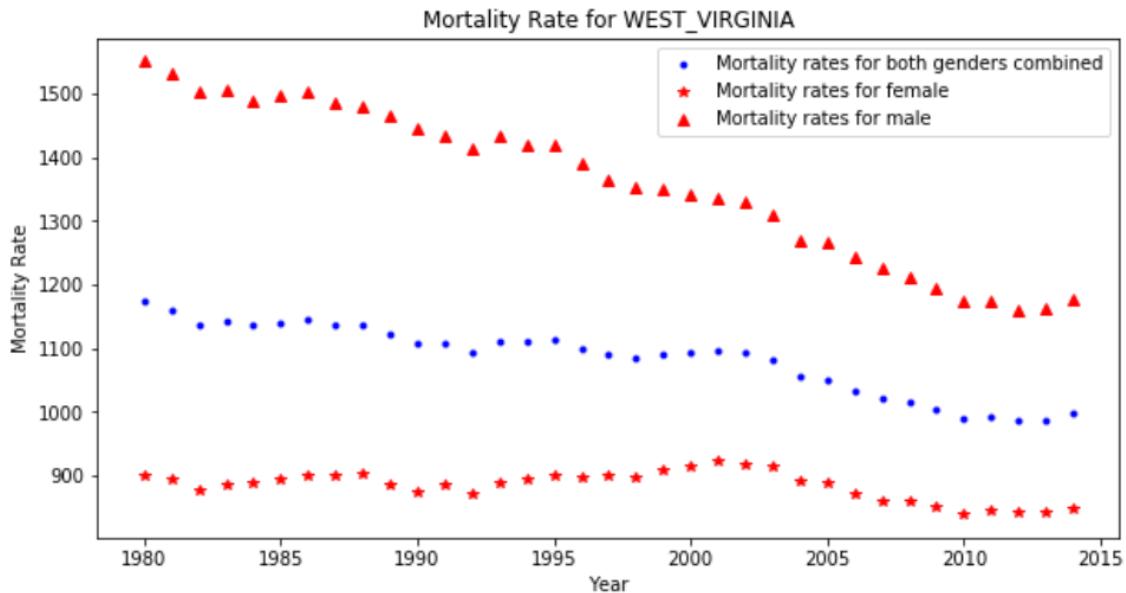
# DATA SCIENCE CASE STUDY



# DATA SCIENCE CASE STUDY



# DATA SCIENCE CASE STUDY



# DATA SCIENCE CASE STUDY

