CAUSES OF DISPARITIES IN LIFE EXPECTANCY AMONG COUNTIES IN USA

STATISTICAL DATA MINING PROJECT REPORT

TEAM NAME: DARK FOREST

TEAM MEMBERS: VENKATA SAI GAGAN DEEP ALUSURI, RAGHAV KHURANA, CHANDAN PATEL, MANOJ ARASADA

UNIVERSITY OF SOUTH FLORIDA
MUMA COLLEGE OF BUSINESS
TAMPA, FLORIDA, 33620

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY 3
1.1 PROBLEM SUMMARY 3
1.2 DATA 3
1.3 ANALYSIS 3
1.4 KEY FINDINGS 4
2. PROBLEM DEFINITON AND SIGNIFICANCE 4
2.1 TARGET CLIENTS 4
2.2 BUSINESS PROBLEM 4
2.3 PROBLEM SIGNIFICANCE 4
3. PRIOR LITERATURE 5
3.1 PREDICTORS USED IN THE PAPERS 5
3.2 PAPER 1: PREMATURE MORTALITY IN THE US: THE ROLES OF GEOGRAPHIC AREA, SOCIOECONOMIC STATUS, HOUSEHOLD STRUCTURE, AND HOUSEHOLD COMPOSITION [1] 6
3.3 PAPER 2: DETERMINANTS OF LIFE EXPECTANCY IN DEVELOPING COUNTRIES [2] 6
3.4 PAPER 3: HOW IMPORTANT ARE HEALTH CARE EXPENDITURES FOR LIFE EXPECTANCY? A COMPARITIVE, EUROPEAN ANALYSIS BY WIM. J.A VAN DEN HEUVEL PHD [3] 6
3.5 PAPER 4: SOCIAL DETERMINANTS OF HEALTH INEQUALITIES INTERNATIONAL CENTRE FOR HEALTH AND SOCIETY, UNIVERSITY COLLEGE LONDON, 1–19 TORRINGTON PLACE, LONDON WC1E 6BT, UK PROF MICHAEL MARMOT [4] 7
3.6 PAPER 5: IMPACT OF SOCIO-HEALTH FACTORS ON LIFE EXPECTANCY IN THE LOW AND LOWER MIDDLE-INCOME COUNTRIES [5] 7
3.7 PAPER 6: COUNTERVAILING EFFECTS OF INCOME, AIR POLLUTION, SMOKING, AND OBESITY ON AGING AND LIFE EXPECTANCY: POPULATION-BASED STUDY OF U.S. COUNTIES [6] 8
3.8 PAPER 7: COUNTY-LEVEL LIFE EXPECTANCY CHANGE: A NOVEL METRIC FOR MONITORING PUBLIC HEALTH ARUNA CHANDRAN 1,*,†, RITIKA PURBEY 1, KATHRYN M. LEIFHEIT 2, KIRSTEN MCGHIE EVANS 1, JOCELYN VELASQUEZ BAEZ 1 AND KERI N. ALTHOFF [7] 8
4. DATA SOURCE AND PREPARATION 9
4.1 SOURCES 9
4.2 DATA PREPARATION 9
5. PREDICTOR TABLE 9
5.1 VARIABLES, ESTIMATED EFFECT ON LIFE EXPECTANCY AND RATIONALE 9
6. DATA VISUALIZATIONS AND DESCRIPTIVE ANALYSIS 11
6.1 VISUALIZATIONS 11
6.2 CORRELATION ANALYSIS AND VARIABLE SELECTION 13
7. MODELING 13
7.1 MODELS AND THEIR RATIONALE 13
7.2 VARIABLE GLOSSARY 13
7.3 MODEL 1 – FIXED EFFECTS MODEL 14

7.4 MODEL 2 – MULTI LEVEL MODEL 14
7.5 MODEL SELECTION 15
7.6 QUALITY CHECKS 15
7.7 LMER MODEL INTERPRETATION: 15
8. RECOMMENDATIONS 15
9. REFERENCES 16
10. APPENDIX 17

1. EXECUTIVE SUMMARY

1.1 PROBLEM SUMMARY

The disparity in life expectancy between States in America is at its widest point in the past 40 years. Most Americans will live to be 78 years old, but if they were born in different areas, their life expectancy may be twenty years lesser compared to other regions. In this paper, we will examine the reasons for this disparity and understand how various social, economic and health factors effect life expectancy in the United States.

1.2 DATA

The data was collected for all the counties in the US for 5 years between 2019 to 2023. Here is a list of variables collected from various sources.

Variables:

Demographics - age, gender, race/ethnicity, socioeconomic status, education, and occupation.
Health behaviors - smoking, alcohol consumption, physical activity, and nutrition.
Environment - air/water quality, green spaces, and exposure to toxins.
Healthcare access - availability of providers, insurance coverage, and preventative care.
Social determinants - poverty, income inequality, and community safety.

Control variables:

Demographics - family size, marital status, immigration status, and location.
Health behaviors - access to healthy food, recreational facilities, and media exposure.
Health outcomes - healthcare access, insurance coverage, and preventative care.
Environment - weather patterns, location, and industrial activity.
Healthcare access - distance to facilities, transportation options, and cost of healthcare.
Social determinants - education level, employment status, and family structure.

1.3 ANALYSIS

Two models were used to analyze life expectancy data with various levels and time: a linear fixed effect model and an LMER model. The linear model accounted for individual-level factors, while the LMER model considered county-wide and state-wide variations. The analysis showed the importance of considering multiple levels and dimensions in understanding life expectancy variations, with implications for public health policies and interventions.

1.4 KEY FINDINGS

- Disparities in the income of individuals have a strong relation to how long they can expect to live. A rise of $10,000 in household income can increase life expectancy by 6 months.
- Even minimal levels of air pollution have the potential to significantly reduce life expectancy. A ten unit increase in PM2.5 (measured on a scale of 0-500) concentration can reduce LE by 1 year.
- Alcohol consumption does not seem to have any noticeable impact on reducing life expectancy. In fact, with an increase in percentage of individuals who reported excessive alcohol consumption appears to positively impact Life Expectancy. 1.2-year increase in LE for every 10% increase in percentage of individuals consuming excessive alcohol.

2. PROBLEM DEFINITON AND SIGNIFICANCE

2.1 TARGET CLIENTS

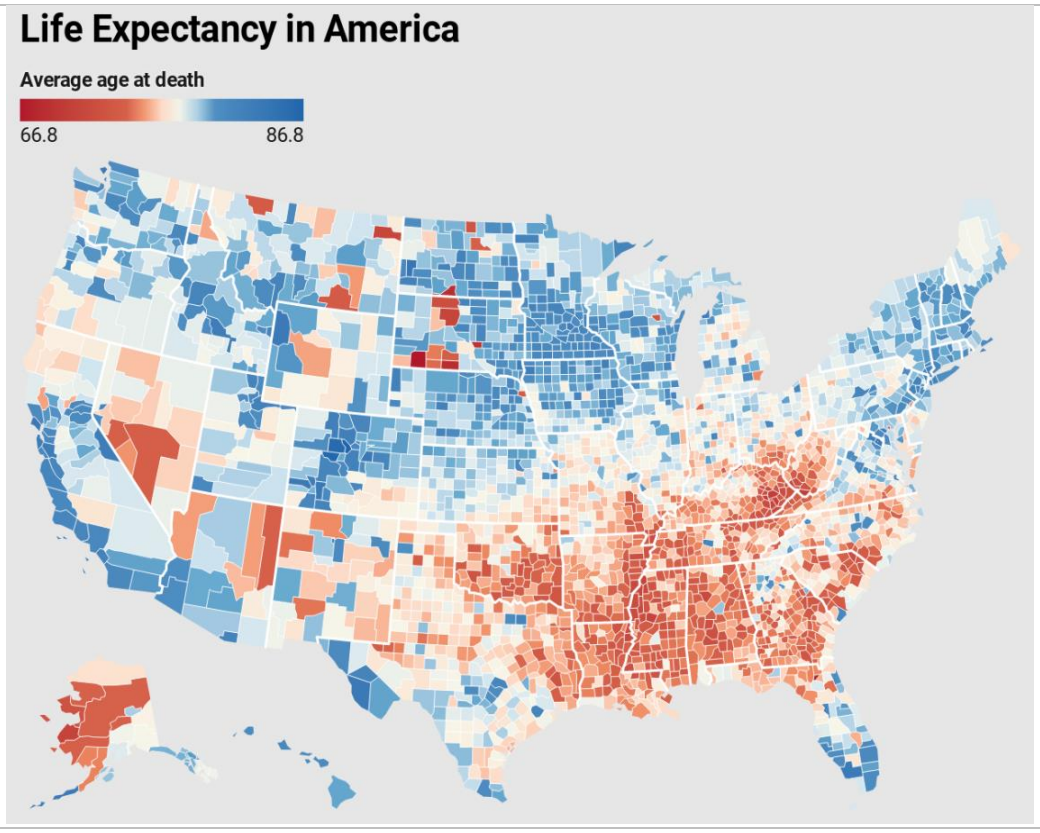
The target clients for the statistical analysis on the causes of disparities in life expectancy among US counties may include public health officials, researchers, government agencies, businesses, social security administration, non-profit organizations, and the general public.

2.2 BUSINESS PROBLEM

Estimate the impact of each factor on Life Expectancy in the United States and help policy makers, governments, businesses, and other sectors understand where to focus their attention and resources, and to try and make meaningful changes to help people live healthier and longer lives.

2.3 PROBLEM SIGNIFICANCE

Life expectancy disparities among US states have widened in the past 40 years, with some Americans dying over ten years earlier depending on where they live. Despite advances in medicine and quality of life, many Americans have not benefited equally. This disparity is illustrated in a graphic below.



3. PRIOR LITERATURE

3.1 PREDICTORS USED IN THE PAPERS

- Geographic area (e.g., region, urban-rural status), socioeconomic status (e.g., income, education), household structure (e.g., family type, household size), and household composition (e.g., age and sex of household members)
- Economic factors: Gross domestic product (GDP) per capita, income inequality, and poverty rates.
- Health factors: Access to health care services, vaccination coverage, and HIV prevalence.
- Demographic factors: Age structure, fertility rates, and literacy rates.
- Environmental factors: Water and sanitation facilities, air pollution, and access to clean energy.
- Political factors: Political stability, government effectiveness, and corruption levels.
- Income, Pollution, Obesity, Smoking, %Black, %Hispanic, Median age, %over 65 yrs. and indicator variables for the nine census division areas.
- Age group, Total County population, median income quartile indicators, population density quartile indicators, proportion of individuals with a 4-year degree, PRCSDA status and census region indicators.

3.2 PAPER 1: PREMATURE MORTALITY IN THE US: THE ROLES OF GEOGRAPHIC AREA, SOCIOECONOMIC STATUS, HOUSEHOLD STRUCTURE, AND HOUSEHOLD COMPOSITION [1]

PREDICTORS:

The predictors used in the study include geographic area (e.g., region, urban-rural status), socioeconomic status (e.g., income, education), household structure (e.g., family type, household size), and household composition (e.g., age and sex of household members)
Y – Life expectancy (LE)

MODELS:

logistic regression models, Cox proportional hazards models, and random-effects Poisson regression models

KEY FINDINGS:

Mortality rates vary geographically in the US, with certain regions having higher rates. Lower socioeconomic status is strongly associated with premature mortality. Household structure and composition also affect mortality rates, with non-traditional structures and living alone being linked to higher rates.
The effects of household structure on mortality are largely mediated by socioeconomic status.
Geographic area, socioeconomic status, and household structure together explain a substantial portion of the variation in premature mortality rates in the US.

3.3 PAPER 2: DETERMINANTS OF LIFE EXPECTANCY IN DEVELOPING COUNTRIES [2]

PREDICTORS:

Economic factors: Gross domestic product (GDP) per capita, income inequality, and poverty rates.
Health factors: Access to health care services, vaccination coverage, and HIV prevalence.
Demographic factors: Age structure, fertility rates, and literacy rates.
Environmental factors: Water and sanitation facilities, air pollution, and access to clean energy.
Political factors: Political stability, government effectiveness, and corruption levels.
Y – Life expectancy (LE)

MODELS:
It primarily uses descriptive statistics and correlation analysis to identify the determinants of life expectancy in developing countries.

KEY FINDINGS:
The results suggest that factors such as access to safe water, education, per capita income, and health expenditure are positively associated with life expectancy.
In contrast, infant mortality rates, prevalence of HIV/AIDS, and the proportion of the population living in urban areas are negatively associated with life expectancy.
The study highlights the importance of investing in public health infrastructure, education, and poverty reduction programs to improve life expectancy in developing countries.

3.4 PAPER 3: HOW IMPORTANT ARE HEALTH CARE EXPENDITURES FOR LIFE EXPECTANCY? A COMPARITIVE, EUROPEAN ANALYSIS BY WIM. J.A VAN DEN HEUVEL PHD [3]

PREDITORS:
Does not focus on specific predictors, but instead examines the relationship between health care expenditures and life expectancy in European countries. The paper compares life expectancy and health care expenditures across 27 European countries, and also looks at other factors that may affect life expectancy, such as income and education.
Y – Life expectancy (LE)

MODELS:
The paper uses multiple linear regression models and structural equation modeling to analyze the data.

KEY FINDINGS:
The study compares the relationship between health care expenditures and life expectancy across 27 European countries.
The results show a positive association between health care expenditures and life expectancy, but the relationship is weak and non-significant in some countries.
Other factors such as socioeconomic conditions, lifestyle, and environmental factors have a stronger impact on life expectancy.
The study suggests that investing in health care alone may not be enough to improve life expectancy, and a comprehensive approach that addresses social determinants of health is needed.

3.5 PAPER 4: SOCIAL DETERMINANTS OF HEALTH INEQUALITIES INTERNATIONAL CENTRE FOR HEALTH AND SOCIETY, UNIVERSITY COLLEGE LONDON, 1–19 TORRINGTON PLACE, LONDON WC1E 6BT, UK PROF MICHAEL MARMOT [4]

PREDITORS:
Social and economic factors, such as income, education, employment, and housing. Political and social factors, including government policies and social norms. Structural factors, such as discrimination, racism, and gender inequality.
Y – Life expectancy (LE)

MODELS:
The paper is a literature review and discussion of the social determinants of health inequalities.

KEY FINDINGS:
There is a strong association between social determinants such as income, education, and occupation and health outcomes, including life expectancy and mortality rates.
Health inequalities are not just a result of individual choices and behaviors but are shaped by broader social, economic, and political factors.
Interventions and policies aimed at reducing health inequalities need to address the social determinants of health, such as poverty, social exclusion, and unequal access to education and employment opportunities.
The paper advocates for a broader approach to health policy that includes not just health care interventions but also upstream interventions that address the social determinants of health.

3.6 PAPER 5: IMPACT OF SOCIO-HEALTH FACTORS ON LIFE EXPECTANCY IN THE LOW AND LOWER MIDDLE-INCOME COUNTRIES [5]

PREDITORS:
Income, Education, Health Expenditure, Physicians Density, Sanitation, Clean Water Access, Malnutrition, HIV/AIDS, Tuberculosis, Malaria, Cardiovascular Diseases
Y – Life expectancy (LE)

MODELS:
Multiple Regression Models

KEY FINDINGS:
The study found that income, education, health expenditure, physician density, sanitation, and access to clean water were positively associated with life expectancy in low and lower-middle income countries. On the other hand, malnutrition, HIV/AIDS, tuberculosis, malaria, and cardiovascular diseases were negatively associated with life expectancy. The authors suggest that interventions aimed at improving education, healthcare access, and sanitation can have a positive impact on life expectancy in these countries. Additionally, addressing the burden of communicable diseases can also improve life expectancy.

3.7 PAPER 6: COUNTERVAILING EFFECTS OF INCOME, AIR POLLUTION, SMOKING, AND OBESITY ON AGING AND LIFE EXPECTANCY: POPULATION-BASED STUDY OF U.S. COUNTIES [6]

PREDITORS:
Y – Life expectancy (LE) / Exceptional Aging (EA)
X variables – Income, Pollution, Obesity, Smoking, %Black, %Hispanic, Median age, %over 65 yrs. and indicator variables for the nine census division areas.

MODELS:
Multiple Regression Models

KEY FINDINGS:

An interesting finding from the study was that policy on one factor alone like reducing air pollution is not a strong enough tool to increase life expectancy. When comparing the tradeoffs with the other variables, this study found that a ten µg/m3 reduction in PM2.5 and a $5,000 increase (adjusted for inflation, base year 2000) in real, per-capita income corresponded to the same increase in Life expectancy.
Additionally, they also found that while greater income does increase life expectancy, a bend in the relationship between income and longevity occurs at about 40,000$, indicating that beyond this level increases in income are associated with smaller increases in longevity.

3.8 PAPER 7: COUNTY-LEVEL LIFE EXPECTANCY CHANGE: A NOVEL METRIC FOR MONITORING PUBLIC HEALTH ARUNA CHANDRAN 1, *,†, RITIKA PURBEY 1, KATHRYN M. LEIFHEIT 2, KIRSTEN MCGHIE EVANS 1, JOCELYN VELASQUEZ BAEZ 1 AND KERI N. ALTHOFF [7]

PREDITORS:
Annual death counts were grouped into 25-34, 35-44 and similar age groups.
A Negative Binomial model was fitted with variables: Age group, Total County population, median income quartile indicators, population density quartile indicators, proportion of individuals with a 4-year degree, PRCSDA status and census region indicators to **calculate the estimated death counts (for counties with <5 3 year averaged observed deaths)**
A Linear Regression model was used to estimate the average change in life expectancy per year for each county and counties were categorized as having increased, decreasing or no change in LE.
Y – Life expectancy (LE)

MODELS:
Multinomial Regression Models and Poisson Models.

KEY FINDINGS:

Linking the County health rankings data was a good move by these researchers since it adds an additional layer to this analysis to help health officials in resource allocation.

It was found that counties with increasing LE between 2011-2016 had significantly lower COVID mortality compared to no-change counties. However, counties with decreasing LE had similar mortality rates to counties with no change for reasons unanswered.

Higher rates of smoking, obesity, unemployment, children in poverty and single parent households understandably led to a decrease in LE. However, binge drinking, motor vehicle crash mortality and preventable hospital stays did not impact LE.

4. DATA SOURCE AND PREPARATION

4.1 SOURCES

- The County Health Rankings, a program of the University of Wisconsin Population Health Institute, measures the health of nearly all counties in the nation and ranks them within states. The Rankings are compiled using county-level measures from a variety of national and state data sources. These measures are standardized and combined using scientifically informed weights.
- Counties in each of the fifty states are ranked according to summaries of a variety of health measures. Those ranking in the healthiest 75-100% of counties are the "healthiest." Counties are ranked relative to the health of other counties in the same state.
- County wise health ranking data https://www.countyhealthrankings.org/
- United States Mortality Rates and Life Expectancy by County, Race, and Ethnicity 2000-2019
- (https://ghdx.healthdata.org/record/ihme-data/united-states-life-expectancy-by-county-race-ethnicity-2000-2019)

4.2 DATA PREPARATION

- Population data was acquired from https://data.census.gov/
- Combined and merged the county wise data from 2019-2023.
- Sheets were merged using R and Tableau Prep builder to combine data from various sources and handle missing values.
- Missing values in certain columns (e.g. -food environment index) were filled in using the county-wise data from https://datausa.io/

5. PREDICTOR TABLE

5.1 VARIABLES, ESTIMATED EFFECT ON LIFE EXPECTANCY AND RATIONALE

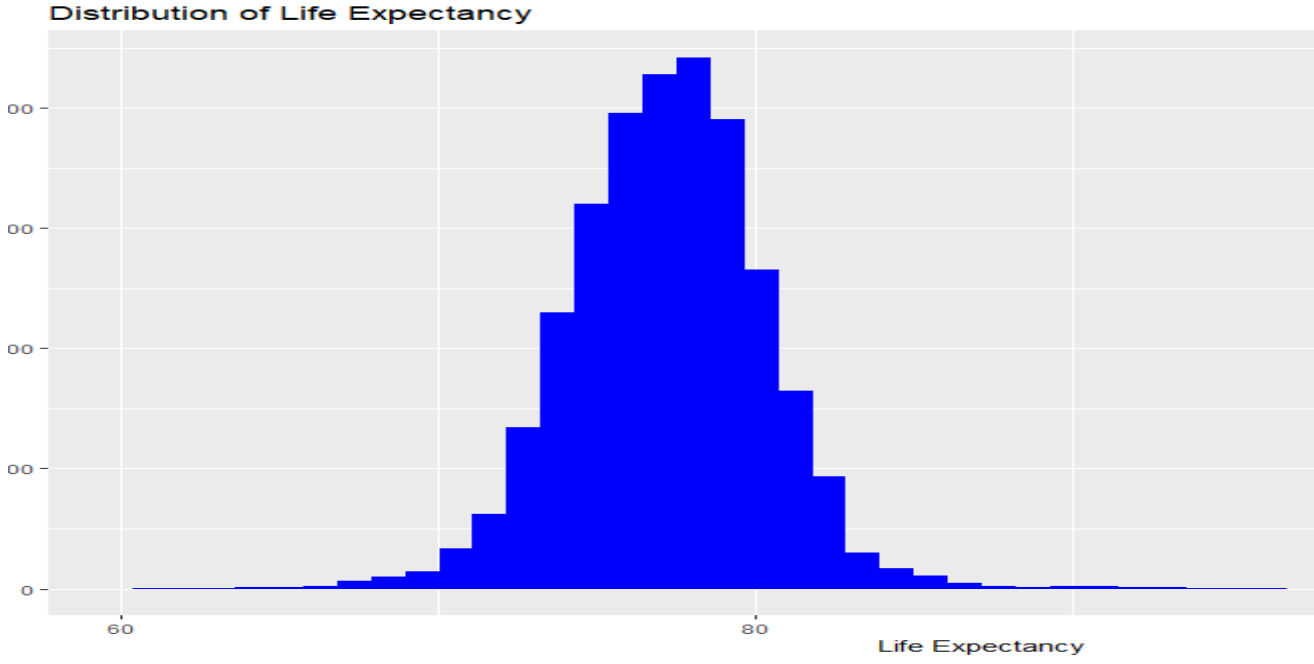
Predictor	Expected Effect	Rationale
Year	+	With increasing years, causes of decreasing LE should be identified and action is likely to be taken to reduce them.
State	?	Policy and laws by each state can influence LE depending on how strict/lax their regulations are.
% of people under frequent physical distress	-	Greater physical distress/ a greater number of distressed individuals can affect LE Negatively.
% of people under frequent mental distress	-	More people under mental distress can lower the LE for an area
% of diabetic individuals	-	Although diabetes is not a fatal disease by itself, it can be hypothesized that those with diabetes might be relatively unhealthier than healthy individuals and can lower the LE of an area

# of food insecure individuals	-	Food insecurity can lead to malnutrition and lower LE
% of people getting insufficient sleep	-	Insufficient sleep can impair the body's normal functions and can affect LE in the long term
No. of uninsured adults **No. of uninsured children**	-	Uninsured people are likely to visit the doctor less frequently due to fear of hefty medical bills and might miss obvious signs of illness. However, it can also be said that they might take better care of themselves knowing that they cannot afford hefty fees. Regardless, the expected effect is likely a negative impact to LE.
Household income	+	Greater income leads to a better quality of life which can improve LE
No. of households with severe cost burden	-	Severe money issues can cause mental stresses, inadequate nutrition, and little to no preventative medical care, likely leading to reduced LE
Population	+/-	Greater population likely indicates a more urban city and can provide better access to food, more varied housing and better healthcare facilities leading to improve LE however can also contribute negatively due to increased pollution, greater mental stress, and other factors.
% of population under 18 + **% of population over sixty-five**	+/-	The skew in population distribution can affect LE positively or negatively as a larger youth population can lead to increase LE due to better access to healthcare, environment and vaccinations compared to individuals over sixty-five who would not have had the same amenities growing up.
Years of Potential Life Lost Rate (YPLL)	Excluded	This variable does not help in analysis of life expectancy and is highly correlated. Additionally, since we cannot control for the process which was used to calculate this rate, using it in analysis can cause problems.
Deaths	Excluded	This variable does not add any value to our analysis and deaths are not a factor which contributes to life expectancy. Since we are trying to focus on predicting the factors affecting life expectancy, this variable is excluded.
Low Birthweight	Excluded	Low birthweight by itself is more of an individual factor and does not provide any information for the general area wide LE calculation. Additionally, this is not an actionable variable.
Alcohol-Impaired Driving Deaths	Excluded	Driving deaths are not a factor which affects LE. Since it is just a cause of death and an individual level choice, not something that affects the population.
Teen Births	Excluded	Number of births given by teenagers is irrelevant to our analysis of LE.
# Dentists	Excluded	The number of dentists in a county is irrelevant to our analysis of LE.
Voter Turnout	Excluded	Voter turnout is not a gauge of health and has no direct link to health outcome/life expectancy.
Mammography Screening	Excluded	Number of individuals undergoing mammography screening is irrelevant to our analysis since it affects a subset of the population and does not have any direct influence on a person's LE.
High School Completion	Excluded	Completing versus not completing high school does not have any link with Life expectancy.
% Some College	Excluded	LE has no relation with if you went to college or not.
Children in Single-Parent Households	Excluded	While this may affect social standing of an individual, as well as disposable income (the effect for which is accounted for in other variables), this variable has no discernable connection to LE.
Injury Deaths	Excluded	The number of deaths due to injury cannot help in analyzing LE.

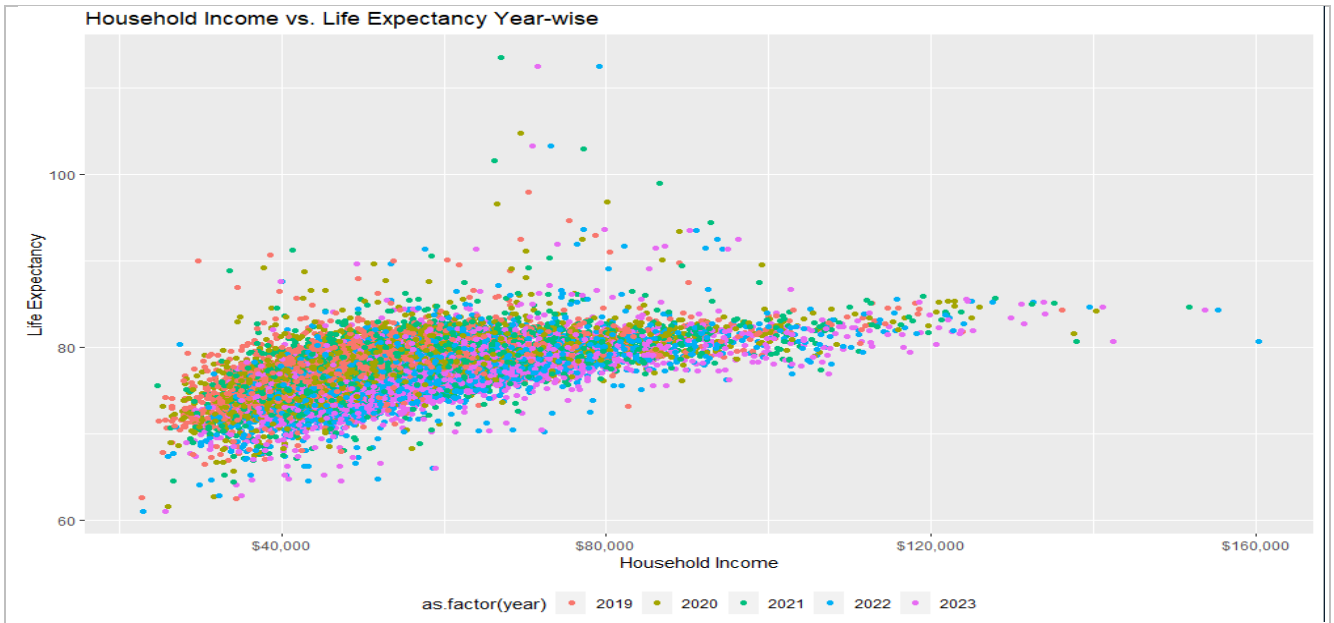
Driving Alone to Work	Excluded	Driving alone versus driving with others has no connection to Life Expectancy.
Long Commute - Driving Alone	Excluded	Neither long commute, nor driving alone has any connection with analyzing LE.
Broadband Access	Excluded	Having or not having broadband does not impact LE.
Reading Scores	Excluded	Scores/grades do not impact LE.
Traffic Volume	Excluded	The volume of traffic is irrelevant to LE and likely has a correlation with population count.
% Not Proficient in English	Excluded	A person's language skills are irrelevant to our analysis of LE.

6. DATA VISUALIZATIONS AND DESCRIPTIVE ANALYSIS

6.1 VISUALIZATIONS

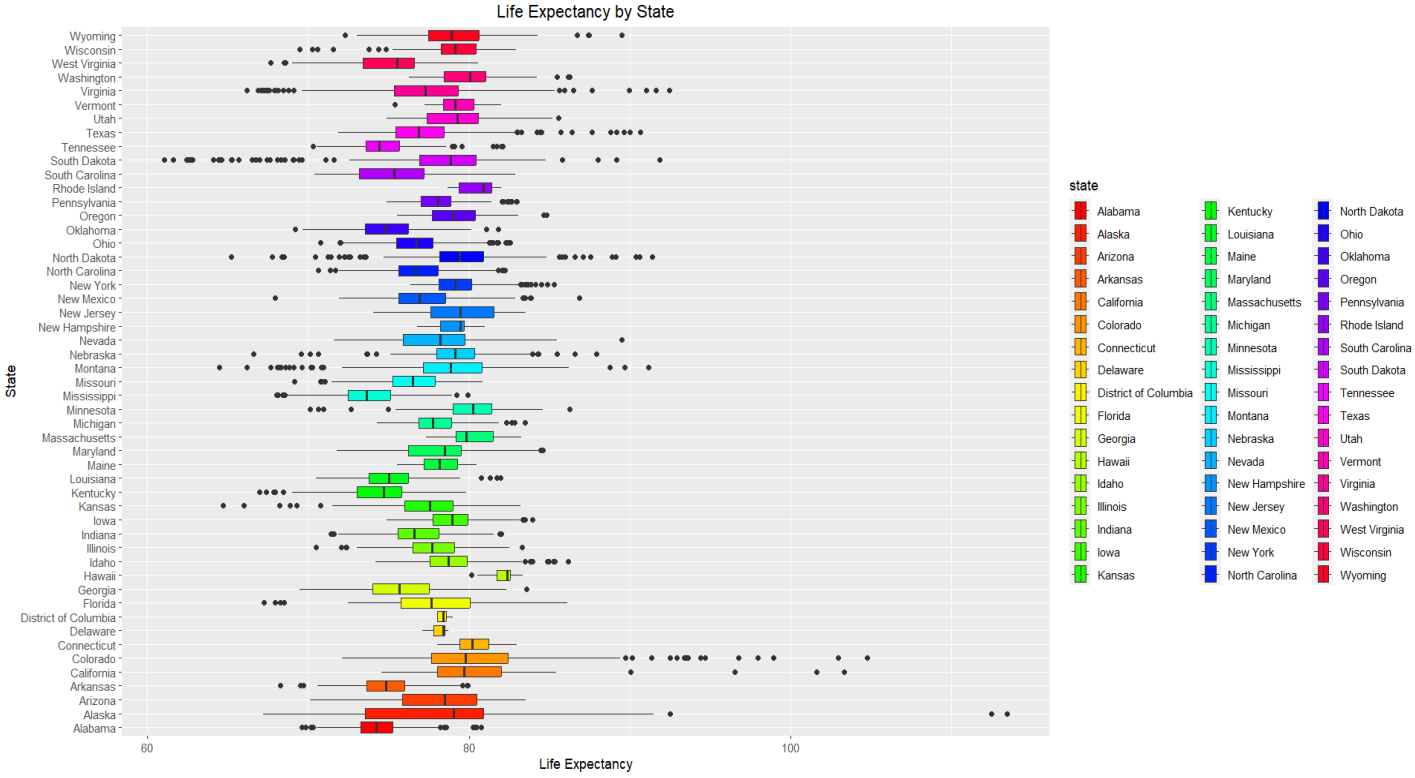


Life Expectancy for all the counties from 2019-2023
The histogram shows us that the distribution of life expectancy is roughly normal, with a peak around 75-80 years.



Household Income Vs Life Expectancy (Year-Wise)

The scatterplot shows us that there is a positive correlation between household income and life expectancy. Counties with higher household incomes tend to have higher life expectancies.



6.2 CORRELATION ANALYSIS AND VARIABLE SELECTION

Due to high correlation between percent of frequent physical distress and average number of physically unhealthy days, we dropped average number of physically unhealthy days as its effect was captured by the other variable.

Due to high correlation between number of households with severe cost burden and population. Population was dropped from the analysis because it was not a strong predictor of life expectancy compared to the former.

High correlation was observed between the number of primary care physicians and mental health providers. Removed the number of mental health providers from the analysis as primary care physicians is more considerable for the analysis.

7. MODELING

7.1 MODELS AND THEIR RATIONALE

Since Life expectancy is approximately normal, we can run linear models. However, the data has various levels (County and State) along with a time dimension.

Hence, we ran 2 models starting with a Linear Model, which is a fixed effect model, followed by LMER model to account for county-wide and state-wide variations in life expectancy.

7.2 VARIABLE GLOSSARY

Percent.Smokers	Percentage of adults that reported currently smoking
Percent.Adults.with.Obesity	Percentage of adults that report BMI >= 30
Food.Environment.Index	Indicator of access to healthy foods - 0 is worst, 10 is best.
Percent.Excessive.Drinking	Percentage of adults that report excessive drinking.
Average.Daily.PM2.5	Average daily amount of fine particulate matter in micrograms per cubic meter
Presence.of.Water.Violation	County affected by a water violation: 1-Yes, 0-No
No.of,households.with.severe.cost.burden	Number of households facing severe cost burden due to high housing costs.
Percent.adult.uninsured	Percentage of people above age 18 without insurance
Percent.food.insecure	Percentage of people facing food insecurity
Percent.diabetic	Percentage of people who are diabetic
Percent.frequent.physical.distress	Percentage of individuals who reported physical distress in the last 30 days.
Percent.65.and.over	Percentage of individuals who are 65 and above.

7.3 MODEL 1 – FIXED EFFECTS MODEL

```
FE_Model <- lm(life.expectancy ~  state + year + percent.frequent.physical.distress + percent.diabetic  +
percent.food.insecure + percent.adult.uninsured+ household.income + no.households.with.severe.cost.burden
+ percent.65.and.over  + percent.female +  percent.smokers + percent.adults.with.obesity +
food.environment.index + percent.excessive.drinking +.primary.care.physicians  + percent.rural +
            + average.daily.pm2.5 + presence.of.water.violation, data=d)
```

Variable	β Coefficient	Interpretation Value (in terms of years)
Year 2020	-0.203	Reduced LE by 2.4 months (compared to 2019)
Year 2021	0.861	Increased LE by 10 months (compared to 2019)
Year 2022	-0.230	Reduced LE by 3 months (compared to 2019)
Year 2023	0.063	No noticeable impact (compared to 2019)
Percent.frequent.physical.distress	0.141	Increases LE by 1.7 months.
Percent.diabetic	-0.072	Reduces LE by <1 month.
Percent.food.insecure	0.001	No Noticeable Impact to LE.
Percent.adult.uninsured	0.014	No Noticeable Impact to LE.
Household.income	0.00005	Increases LE by 6 months for every $10,000.
no.households.with.severe.cost.burden	0.00001	No Noticeable Impact to LE.
percent.65.and.over	0.016	No Noticeable Impact to LE.
percent.smokers	-0.428	Reduces LE by 5 months.
percent.adults.with.obesity	-0.052	Reduces LE by 0.5 month.
food.environment.index	0.002	No Noticeable Impact to LE.
percent.excessive.drinking	0.116	Increases LE by 1.4 months.
no.primary.care.physicians	-0.001	No Noticeable Impact to LE.
percent.rural	0.004	No Noticeable Impact to LE.
average.daily.pm2.5	-0.106	Reduces LE by 1.3 months.
presence.of.water.violationYes	0.037	No Noticeable Impact to LE.

7.4 MODEL 2 – MULTI LEVEL MODEL

```
re_model <- lmer(life.expectancy ~ year + percent.frequent.physical.distress  + percent.diabetic +
percent.food.insecure+ percent.adult.uninsured +household.income + no.households.with.severe.cost.burden
+ percent.65.and.over + percent.female +  percent.smokers + percent.adults.with.obesity +
food.environment.index + percent.excessive.drinking + no.primary.care.physicians+ percent.rural +
average.daily.pm2.5 + presence.of.water.violation + (1 | state), data=d, REML=FALSE)
```

Variable	β Coefficient	Interpretation Value (in terms of LE)
Year 2020	-0.200	Reduced LE by 2.4 months (compared to 2019)
Year 2021	0.840	Increased LE by 10 months (compared to 2019)
Year 2022	-0.245	Reduced LE by 3 months (compared to 2019)
Year 2023	0.029	No noticeable impact (compared to 2019)
Percent.frequent.physical.distress	0.132	Increases LE by 1.5 months.
Percent.diabetic	-0.073	Reduces LE by <1 month.
Percent.food.insecure	0.0005	No Noticeable Impact to LE.
Percent.adult.uninsured	0.014	No Noticeable Impact to LE.
Household.income	0.00005	Increases LE by 6 months for every $10,000.
no.households.with.severe.cost.burden	0.00001	No Noticeable Impact to LE.
percent.65.and.over	0.017	No Noticeable Impact to LE.
percent.smokers	-0.421	Reduces LE by 5 months.
percent.adults.with.obesity	-0.051	Reduces LE by 0.5 month.
food.environment.index	0.007	No Noticeable Impact to LE.
percent.excessive.drinking	0.121	Increases LE by 1.5 months.
no.primary.care.physicians	-0.001	No Noticeable Impact to LE.
percent.rural	0.004	No Noticeable Impact to LE.
average.daily.pm2.5	-0.105	Reduces LE by 1.2 months.
presence.of.water.violationYes	0.033	No Noticeable Impact to LE.

7.5 MODEL SELECTION

Even though the outputs of the models don't differ widely in terms of the estimated effect of each predictor. since the linear model does not account for the statewide differences and the multi-level model takes those effects into account, therefore, we will choose to make recommendations based on the multi-level model.

7.6 QUALITY CHECKS

Although LMER models are robust to Linearity, Normality and Equality of variances assumptions. But we still need to perform tests on multi-collinearity and Independence.

7.7 LMER MODEL INTERPRETATION:

- In 2021, even during the height of the pandemic, the average life expectancy increased by about 10 months compared to 2019.
- We can see that a 10% increase in the percent of diabetic individuals equates to a roughly 9-month reduction in LE for a given area.
- A $10,000 increase in the household income has a positive impact equating to 6 months increased LE for a given area.
- To no surprise, smoking has a negative impact to Life expectancy, with a 10% increase in the percentage of population that smokes equating to a 4+ year reduction in LE for any given area.
- Obesity in adults also has a negative impact to LE, with every 10% increase in percentage of adults with obesity equating to a 6-month reduction in LE for any given area.
- Strangely, excessive alcohol consumption seems to have a positive impact to LE, with a 10% increase in the percentage of adults consuming excessive alcohol equating to a 1.2-year increase in life expectancy for any given area.
- Air pollution (measured as daily average PM2.5 concentration) has a negative impact on LE, with every 10-point increase in the average daily concentration of particles, reducing LE by 1 year.

8. RECOMMENDATIONS

- Obesity and some form of diabetes go hand in hand, we can see that both diabetes and obesity negatively impact LE and effects to mitigate their impact can help the public lead healthier and longer lives. By scrutinizing the amount of processed ingredients, added sugars and nutritional content in our food, policymakers and government agencies can attempt to fix this problem directly at the source and helping citizens by limiting access to highly processed and unhealthy food choices.

- The anti-smoking campaign has worked very effectively in the past, with smoking rates in US adults at an all-time low. While commendable, more action can be taken to reduce smoking even further by taking inspiration from other countries (like New Zealand) or imposing more tax on cigarettes to discourage their use even further. Additionally, taking inspiration from India and adding more striking visuals to each box, might also help in reducing smoking.

- Increased household income has a positive impact on LE. Local governments can play close attention to this metric and focus on working hard to increase the economic standing of their constituents through creating local jobs, incentivizing businesses to move into their area, considering tax rebates and focusing on improving the local and national economy.

- Air pollution is a very serious issue, passing stringent local/national laws to enforce stricter guidelines on air pollutants discharged by industries, cars and incentivizing the push toward a greener future through electric vehicles, solar/wind/nuclear power generation and making cities more walkable to reduce

reliance on vehicles as well as enabling more physical exercise in the process would be a great move to help tackle both the issue of pollution as well as physical and mental well-

9. REFERENCES

[1]: https://ajph.aphapublications.org/doi/abs/10.2105/AJPH.89.6.893

[2]: Determinants of Life Expectancy in Developing Countries https://muse.jhu.edu/article/231979/pdf

[3]: HOW IMPORTANT ARE HEALTH CARE EXPENDITURES FOR LIFE EXPECTANCY? A COMPARITIVE, EUROPEAN ANALYSIS BY WIM. J.A VAN DEN HEUVEL PHD https://www.jamda.com/article/S1525-8610(16)30559-X/fulltext

[4]: SOCIAL DETERMINANTS OF HEALTH INEQUALITIES INTERNATIONAL CENTRE FOR HEALTH AND SOCIETY, UNIVERSITY COLLEGE LONDON, 1–19 TORRINGTON PLACE, LONDON WC1E 6BT, UK PROF MICHAEL MARMOT https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(05)71146-6/fulltext

[5] IMPACT OF SOCIO-HEALTH FACTORS ON LIFE EXPECTANCY IN THE LOW AND LOWER MIDDLE-INCOME COUNTRIES https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4441932/

[6] COUNTERVAILING EFFECTS OF INCOME, AIR POLLUTION, SMOKING, AND OBESITY ON AGING AND LIFE EXPECTANCY: POPULATION-BASED STUDY OF U.S. COUNTIES https://ehjournal.biomedcentral.com/articles/10.1186/s12940-016-0168-2

[7] COUNTY-LEVEL LIFE EXPECTANCY CHANGE: A NOVEL METRIC FOR MONITORING PUBLIC HEALTH ARUNA CHANDRAN 1,*,†, RITIKA PURBEY 1, KATHRYN M. LEIFHEIT 2, KIRSTEN MCGHIE EVANS 1, JOCELYN VELASQUEZ BAEZ 1 AND KERI N. ALTHOFF https://www.mdpi.com/1660-4601/19/17/10672

Data Sources:

- https://data.census.gov/
- (https://ghdx.healthdata.org/record/ihme-data/united-states-life-expectancy-by-county-race-ethnicity-2000-2019)
- https://www.countyhealthrankings.org/
- https://datausa.io/
- https://wonder.cdc.gov/

10. APPENDIX

10.1 LIMITATIONS

Given that the data used for our analysis is individual level health data, it does come with certain limitations. Although data was sourced from reputable and government websites, we do not know or have any control over how the data was collected. Since we have no control over the underlying data generation process, it does affect the validity of the recommendations that we are making as well as the statistical output of our models. The results may not represent reality and must not be taken at face value.

The data was inconsistently populated, and the statistical output sometimes went against common sense, although we did spend considerable time on the data cleaning and feature engineering, it is possible that some inconsistencies may have slipped into the final dataset.

10.2 STARGAZER OUTPUT

	Dependent variable:	
	life.expectancy	
	OLS	*linear mixed-effects*
	(1)	(2)
stateAlabama	-1.927***	
	(0.159)	
stateAlaska	-1.224**	
	(0.536)	
stateArizona	-1.142***	
	(0.242)	
stateArkansas	-1.386***	
	(0.153)	
stateCalifornia	-1.744***	
	(0.166)	
stateColorado	-0.706***	
	(0.147)	
stateConnecticut	-1.845***	
	(0.306)	
stateDelaware	-1.266***	
	(0.481)	
stateDistrict of Columbia	-3.776***	
	(0.821)	
stateFlorida	-0.371**	
	(0.144)	
stateGeorgia	-1.205***	
	(0.133)	
stateIdaho	-0.906***	
	(0.165)	

statelllinois	-1.254***
	(0.129)
stateIndiana	-0.500***
	(0.139)
statelowa	-0.597***
	(0.122)
stateKansas	-1.378***
	(0.127)
stateKentucky	-1.204***
	(0.142)
stateLouisiana	-1.261***
	(0.146)
stateMaine	-1.095***
	(0.224)
stateMaryland	-2.230***
	(0.199)
stateMassachusetts	-1.989***
	(0.241)
stateMichigan	-0.352***
	(0.130)
stateMississippi	-2.186***
	(0.154)
stateMissouri	-0.355***
	(0.129)
stateMontana	-1.121***
	(0.144)
stateNebraska	-0.929***
	(0.125)
stateNevada	-1.557***
	(0.222)
stateNew Hampshire	-1.938***
	(0.275)
stateNew Jersey	-2.234***
	(0.211)
stateNew Mexico	-2.270***
	(0.188)
stateNew York	-0.392***
	(0.139)
stateNorth Carolina	-0.901***
	(0.133)
stateNorth Dakota	-0.544***
	(0.144)
stateOhio	-0.265*

	(0.138)	
stateOklahoma	-1.861***	
	(0.154)	
stateOregon	-0.688***	
	(0.170)	
statePennsylvania	-0.498***	
	(0.140)	
stateRhode Island	-1.533***	
	(0.378)	
stateSouth Carolina	-2.213***	
	(0.160)	
stateSouth Dakota	-0.925***	
	(0.138)	
stateTennessee	-1.146***	
	(0.146)	
stateTexas	-2.226***	
	(0.123)	
stateUtah	-3.191***	
	(0.219)	
stateVermont	-1.643***	
	(0.237)	
stateVirginia	-1.202***	
	(0.132)	
stateWashington	-0.808***	
	(0.171)	
stateWest Virginia	-0.280*	
	(0.168)	
stateWisconsin	-1.035***	
	(0.138)	
stateWyoming	-1.227***	
	(0.198)	
year2020	-0.203***	-0.200***
	(0.048)	(0.047)
year2021	0.861***	0.840***
	(0.065)	(0.065)
year2022	-0.230***	-0.245***
	(0.068)	(0.068)
year2023	0.063	0.029
	(0.117)	(0.116)
percent.frequent.physical.distress	0.141***	0.132***
	(0.021)	(0.021)
percent.diabetic	-0.072***	-0.073***
	(0.007)	(0.007)

percent.food.insecure	0.001***	0.0005**
	(0.0002)	(0.0002)
percent.adult.uninsured	0.014***	0.014***
	(0.003)	(0.003)
household.income	0.00005***	0.00005***
	(0.00000)	(0.00000)
no.households.with.severe.cost.burden	0.00001***	0.00001***
	(0.00000)	(0.00000)
percent.65.and.over	0.016***	0.017***
	(0.004)	(0.004)
percent.female	0.001	0.001
	(0.002)	(0.002)
percent.smokers	-0.428***	-0.421***
	(0.012)	(0.011)
percent.adults.with.obesity	-0.052***	-0.051***
	(0.005)	(0.005)
food.environment.index	0.002	0.007
	(0.020)	(0.020)
percent.excessive.drinking	0.116***	0.121***
	(0.011)	(0.011)
no.primary.care.physicians	-0.001***	-0.001***
	(0.0001)	(0.0001)
percent.rural	0.004***	0.004***
	(0.001)	(0.001)
average.daily.pm2.5	-0.106***	-0.105***
	(0.013)	(0.013)
presence.of.water.violationYes	0.037	0.033
	(0.034)	(0.034)
Constant	82.753***	81.320***
	(0.458)	(0.448)
Observations	15,356	15,356
R^2	0.647	
Adjusted R^2	0.646	
Log Likelihood		-31,053.400
Akaike Inf. Crit.		62,152.800
Bayesian Inf. Crit.		62,328.500
Residual Std. Error	1.819 (df = 15286)	
F Statistic	406.721*** (df = 69; 15286)	
Note:		*p<0.1; **p<0.05; ***p<0.01

Output of random effects -

> ranef(re)

$state (Intercept)

Minnesota	1.198059131
Alabama	-0.646195231
Alaska	0.023238577
Arizona	0.130643926
Arkansas	-0.121255873
California	-0.443291224
Colorado	0.548429155
Connecticut	-0.483314934
Delaware	-0.007957165
District of Columbia	-1.020208431
Florida	0.860752371
Georgia	0.054248576
Idaho	0.359229863
Illinois	-0.018823562
Indiana	0.719552269
Iowa	0.609230683
Kansas	-0.130248650
Kentucky	0.040045090
Louisiana	-0.024970519
Maine	0.128925744
Maryland	-0.907250022
Massachusetts	-0.664548082
Michigan	0.860736005
Mississippi	-0.900964806
Missouri	0.869040444
Montana	0.105524315
Nebraska	0.297878743
Nevada	-0.282117117
New Hampshire	-0.597220202
New Jersey	-0.895020202
New Mexico	-0.932884151
New York	0.827545251
North Carolina	0.347047777
North Dakota	0.644903767
Ohio	0.942342936
Oklahoma	-0.578442336
Oregon	0.548167345
Pennsylvania	0.714674848
Rhode Island	-0.209545914
South Carolina	-0.930899086
South Dakota	0.294522306
Tennessee	0.096022371

Texas	-0.948011108
Utah	-1.756024571
Vermont	-0.360557582
Virginia	0.045819004
Washington	0.452783055
West Virginia	0.946902816
Wisconsin	0.170632767
Wyoming	0.022851633

with conditional variances for "state"

10.3 R Code:

```
rm(list=ls())
library(rio)
library(dplyr)
d = import("Final2019-2023.xlsx")
d <- d[complete.cases(d), ]

d$StateAndCounty  = paste(d$county, d$state,sep = ", ")
colnames(d)
#Converted # to %
d$percent.food.insecure                    =     (d$no.food.insecure / d$population) * 100.0
d$percent.adult.uninsured                  =     (d$no.adult.uninsured / d$population) * 100.0
d$percent.child.uninsured                  =     (d$no.children.uninsured / d$population) * 100.0
#d$percent.households.with.severe.cost.burden        =     (d$no.households.with.severe.cost.burden / d$population) * 100.0
d$percent.rural                            =     (d$no.rural / d$population) * 100.0
d$average.number.of.physically.unhealthy.days        =     d$average.number.of.physically.unhealthy.days * 12
d$average.number.of.mentally.unhealthy.days          =     d$average.number.of.mentally.unhealthy.days * 12



#Checking correlation
str(d)
boxplot(life.expectancy~ factor(presence.of.water.violation), data = d)
boxplot(life.expectancy ~ state, horizontal = TRUE, data = d)
```

```
str(d)
numeric_cols <- d %>% select(c(4,5,6,7,9,12,13,14,15,16,18,20,22,24,26,27,29,30,31,32,33,34,35,36,37,39,41,42,43,44))
library(openxlsx)
# Compute the correlation matrix using Pearson correlation
cor_matrix <- cor(numeric_cols, method = "pearson")
# Create an Excel workbook
wb <- createWorkbook()

# Add a sheet to the workbook
addWorksheet(wb, "Correlation Matrix")

# Write the column names to the sheet
writeData(wb, "Correlation Matrix", colnames(cor_matrix), startCol = 2, startRow = 1)

# Write the row names and correlation matrix to the sheet
writeData(wb, "Correlation Matrix", cbind(rownames(cor_matrix), cor_matrix), startCol = 1, startRow = 2)

# Save the Excel workbook
saveWorkbook(wb, "correlation_matrix.xlsx", overwrite = TRUE)

#High correlation between---
#Removing percent.frequent.mental.distress, percent.insufficient.sleep, population, percent,
#average.number.of.physically.unhealthy.days, average.number.of.mentally.unhealthy.days,
#no.mental.health.providers, percent.food.insecure, percent.child.uninsured

#Converting to factors.
col <- c("year", "county", "state", "presence.of.water.violation")
d[col] <- lapply(d[col], factor)

attach(d)
library(lme4)
re <- lmer(life.expectancy ~ year + percent.frequent.physical.distress  +percent.diabetic  + percent.food.insecure+ percent.adult.uninsured+
            + household.income + no.households.with.severe.cost.burden + percent.65.and.over
          + percent.female +  percent.smokers + percent.adults.with.obesity + food.environment.index + percent.excessive.drinking +
```

```
          + no.primary.care.physicians  + percent.rural +
          + average.daily.pm2.5 + presence.of.water.violation
         + (1 | state), data=d, REML=FALSE)

ranef(re)

d$state = relevel(d$state, 'Minnesota')
linear <- lm(life.expectancy ~  state + year + percent.frequent.physical.distress + percent.diabetic  + percent.food.insecure+
percent.adult.uninsured+
            household.income + no.households.with.severe.cost.burden + percent.65.and.over
          + percent.female +  percent.smokers + percent.adults.with.obesity + food.environment.index + percent.excessive.drinking +
           + no.primary.care.physicians  + percent.rural +
           + average.daily.pm2.5 + presence.of.water.violation, data=d)

library(stargazer)
stargazer(linear, re, type="html",out="OutputOfModels.htm")

library('car')
vif(re)
library(gvlma)
residuals <- resid(re)
durbinWatsonTest(residuals)
```