Poorya behnamie

Task 6.1

Predicting Cancer Incidence Based on Environmental and Socioeconomic Factors

Project Overview

Objective: To analyze the impact of socioeconomic and environmental factors on cancer incidence and mortality rates across the United States from 2010 to 2020. The goal is to identify patterns and predictors that could inform healthcare policies and interventions to reduce cancer incidence and improve healthcare outcomes.

Given my background in healthcare, data analysis, and epidemiology, this project is highly relevant to my expertise. It allows me to combine my skills in data handling with a significant public health issue. Understanding the interplay between socioeconomic factors and cancer incidence is crucial, and I am excited about the potential to contribute to devising more targeted and effective health interventions.

Importance of the Project

This project has the potential to make a significant impact on public health by identifying and understanding the socioeconomic and environmental factors influencing cancer incidence and mortality. Through rigorous data analysis and predictive modeling, it aims to provide actionable insights that can guide effective interventions and policy decisions.

Datasets:

1: United States Cancer Statistics from the CDC

Data Source and Collection Method

The United States Cancer Statistics (USCS) dataset compiles comprehensive cancer registry data from the CDC's National Program of Cancer Registries (NPCR) and the NCI's SEER Program, along with mortality data from the CDC's National Center for Health Statistics (NCHS). Covering all 50 states, the District of Columbia, and Puerto Rico, it includes cancer cases diagnosed from 2001 to 2020. Data is collected from hospitals, physicians, and laboratories, reported to central cancer registries, and subjected to rigorous quality control to ensure accuracy.

Content and Relevance

The dataset includes data on leading cancer sites, incidence and death counts, and age-adjusted rates for incidence and mortality. This dataset is crucial for analyzing cancer incidence and mortality across various demographics and regions, allowing my project to explore the impact of socioeconomic factors on cancer outcomes and identify disparities.

Limitations and Ethical Considerations

While extensive, the dataset may have underreporting or misclassification issues and inconsistencies due to varying state reporting practices. The COVID-19 pandemic may have also affected data collection in 2020. Ethical considerations include ensuring data privacy and applying suppression rules to prevent individual identification, especially in small case numbers. These measures are vital for responsible data use and avoiding stigmatization of any population.

The USCS dataset is essential for understanding the relationship between socioeconomic factors and cancer, providing valuable insights for public health interventions.

Variables	Time dependency	Quantitative/Qualitative	Numeric/Nominal	Categorical/Continious
State	Invarient	Qualitative	Nominal	Categorical
Year	Invarient	Quantitative	Numerical	Discrete
State-Year	Invarient	Qualitative	Nominal	Categorical
Gender	Invarient	Qualitative	Nominal	Categorical
Leading Cancer Sites	Invarient	Qualitative	Nominal	Categorical
Incidence Counts	Varient	Quantitative	Numerical	Discrete
Incidence Age-Adjusted Rate	Varient	Quantitative	Numerical	Countunous
Death Counts	Varient	Quantitative	Numerical	Discrete
Mortality Age-Adjusted Rate	Varient	Quantitative	Numerical	Countunous
Mortality-Incidence Age-Adjusted Rate Ratio	Varient	Quantitative	Numerical	Countunous
Population	Varient	Quantitative	Numerical	Discrete

State	Year	State-Year	Gender	Leading Cancer Sites	Incidence Counts	Incidence Age-Adjusted Rate	Death Counts	Mortality Age-Adjusted Rate	Mortality-Incidence Age-Adjusted Rate Ratio	Population
Alabama	2010	Alabama2010	Male	Prostate	3872	151.078	544	27.666	0.183	2323013
Alabama	2011	Alabama2011	Male	Prostate	3817	146.447	544	26.612	0.182	2328518
Alabama	2012	Alabama2012	Male	Prostate	3412	127.557	461	22.065	0.173	2336196
Alabama	2013	Alabama2013	Male	Prostate	3318	120.949	470	22.134	0.183	2343135
Alabama	2014	Alabama2014	Male	Prostate	3103	109.17	467	21.298	0.195	2348012
Alabama	2015	Alabama2015	Male	Prostate	3499	121.769	495	21.767	0.179	2352806
Alabama	2016	Alabama2016	Male	Prostate	3679	124.881	512	21.428	0.172	2357211
Alabama	2017	Alabama2017	Male	Prostate	3955	131.483	464	19.24	0.146	2360503
Alabama	2018	Alabama2018	Male	Prostate	3783	123.567	525	21.053	0.17	2365445
Alabama	2019	Alabama2019	Male	Prostate	3765	119.415	507	19.447	0.163	2371832
Alabama	2020	Alabama2020	Male	Prostate	3283	103.354	534	20.03	0.194	2376966
Alaska	2010	Alaska2010	Male	Prostate	394	127.199	43	23.926	0.188	371634
Alaska	2011	Alaska2011	Male	Prostate	348	101.032	39	20.699	0.205	375929
Alaska	2012	Alaska2012	Male	Prostate	314	94.29	35	17.96	0.19	381032
Alaska	2013	Alaska2013	Male	Prostate	283	81.668	39	17.507	0.214	385939
Alaska	2014	Alaska2014	Male	Prostate	278	82.1	53	23.072	0.281	385994
Alaska	2015	Alaska2015	Male	Prostate	261	70.437	35	13.988	0.199	386716
Alaska	2016	Alaska2016	Male	Prostate	376	98.078	43	18.835	0.192	388632
Alaska	2017	Alaska2017	Male	Prostate	364	96.391	39	16.997	0.176	386986
Alaska	2018	Alaska2018	Male	Prostate	415	103.81	48	20.045	0.193	384304
Alaska	2019	Alaska2019	Male	Prostate	432	104.392	65	24.225	0.232	382813
Alaska	2020	Alaska2020	Male	Prostate	386	92.552	49	17.538	0.189	381537

2: County Health Rankings & Roadmaps (CHR&R)

State	Year	State-Year	Deaths	Years of Potential Life Lost Rate	% Fair or Poor Health	Average Number of Physically Unhealthy Days	Average Number of Mentally Unhealthy Days	% Low birthweight	% Adult smoking	% Adults with Obesity	Food Environment Index	% Physically Inactive	% With Access to Exercise Opportunities	% Excessive Drinking	% Drivin Deaths w Alcoho Involvem
Arkansas	2020	Arkansas2020	48,253.00	9,337.34	23.27	4.82	5.28	9.07	23.71	35.00	5.10	30.40	63.53	17.25	26
California	2020	California2020	350,612.00	5,253.06	17.61	3.86	3.73	6.88	11.47	24.30	8.80	17.70	93.07	18.13	28
Colorado	2020	Colorado2020	53,905.00	5,943.02	13.79	3.29	3.71	9.06	14.67	22.40	8.40	14.80	90.46	21.28	33
Connecticut	2020	Connecticut2020	35,724.00	5,748.12	12.99	3.30	3.79	7.80	12.54	26.30	8.20	19.90	94.02	20.47	31
Delaware	2020	Delaware2020	12,857.00	7,938.04	16.27	3.70	4.16	8.89	17.45	32.40	7.80	27.30	86.47	19.81	25
District of Columbia	2020	District of Columbia2020	8,280.00	7,955.66	14.69	3.28	4.17	9.96	15.38	24.00	8.50	17.40	100.00	23.45	28
Florida	2020	Florida2020	266,657.00	7,187.87	19.51	4.01	4.16	8.68	14.89	27.20	6.90	25.80	88.74	19.73	22
Georgia	2020	Georgia2020	133,085.00	7,616.11	18.42	3.95	4.18	9.74	16.34	32.30	6.50	26.40	75.49	16.81	20
Hawaii	2020	Hawaii2020	14,415.00	5,865.43	15.43	3.22	3.38	8.31	14.39	24.60	7.60	19.60	92.54	22.95	31
Idaho	2020	Idaho2020	18,269.00	6,169.43	15.11	3.74	3.97	6.85	14.96	29.00	7.70	20.40	78.90	17.10	31
Illinois	2020	Illinois2020	144,231.00	6,647.08	15.92	3.58	3.80	8.35	15.88	29.70	8.70	21.60	90.80	21.54	31
Indiana	2020	Indiana2020	92,231.00	8,251.60	18.19	3.95	4.65	8.08	21.72	33.90	7.00	26.70	75.24	18.65	18
Iowa	2020	Iowa2020	35,779.00	6,232.21	13.46	3.06	3.53	6.72	17.40	34.30	8.50	22.60	82.95	25.77	26
Kansas	2020	Kansas2020	35,107.00	7,078.55	16.28	3.62	4.05	7.16	17.94	33.00	6.70	23.90	80.12	18.21	19
Kentucky	2020	Kentucky2020	74,341.00	9,505.08	21.84	4.58	5.05	8.81	24.11	34.60	6.90	28.70	71.09	17.16	25
Louisiana	2020	Louisiana2020	72,099.00	9,516.02	21.42	4.32	5.01	10.70	21.09	36.30	5.20	28.00	75.03	19.72	32
Maine	2020	Maine2020	18,350.00	7,020.76	17.07	4.19	5.00	7.19	19.40	29.80	8.00	20.80	69.97	21.98	35
Maryland	2020	Maryland2020	70,898.00	7,197.71	15.17	3.37	3.71	8.66	12.61	31.60	8.70	21.90	92.57	15.41	28
Massachusetts	2020	Massachusetts2020	68,473.00	5,609.69	13.52	3.50	4.29	7.54	13.69	25.00	9.20	20.00	94.47	23.55	30
Michigan	2020	Michigan2020	132,221.00	7,535.14	18.34	4.31	4.71	8.51	20.12	32.40	7.00	23.10	85.50	20.92	29
Minnesota	2020	Minnesota2020	52,125.00	5,314.26	12.89	3.14	3.54	6.63	15.52	29.00	8.90	19.60	86.91	23.19	29

Data Source and Collection Method

The second dataset for my project comes from the County Health Rankings & Roadmaps (CHR&R) program, a collaboration between the University of Wisconsin Population Health Institute and the Robert Wood Johnson Foundation. This dataset provides detailed socioeconomic data for U.S. counties from 2010 to 2020, including factors like education, income, and housing. I combined ten individual datasets from this program to create a comprehensive view of each state's socioeconomic features over the decade.

The data is collected from external, administrative sources such as the National Center for Health Statistics and the Census Bureau's Population Estimates Program. These sources provide official records on births, deaths, and population estimates. The data is age-adjusted to ensure fair comparisons across counties with different age structures and considers various health determinants.

Limitations and Ethical Considerations

However, the dataset has some limitations. Data reliability can vary, especially for smaller counties, and methodological changes over time can affect comparability. Additionally, life expectancy calculations can be complex and sensitive to age structure and infant mortality.

Content and Relevance

This dataset is crucial for my project as it highlights socioeconomic disparities across regions, providing insights needed for analyzing health outcomes and developing targeted interventions. Its comprehensive nature and focus on health determinants make it an essential resource for addressing health disparities in the United States.

Data Cleaninng:

1: United States Cancer Statistics from the CDC

Primary Column	Column Type		Column Rename	Missing Value	Duplicates	Mixed-Type Data
State	object	object	State	-	-	-
Year	Int64	Int64	Year	-	-	-
State-Year	object	object	State_Year	-	-	-
Gender	object	category	Gender	-	-	-
Leading Cancer Sites	object	category	Leading_Cancer_Sites	-	-	-
Incidence Counts	Float64	Int64	Incidence_Counts	22	-	-
Incidence Age-Adjusted Rate	Float64	Float64	Incidence_Age_Adjusted_Rate	22	-	-
Death Counts	Float64	Int64	Death_Counts	22	-	-
Mortality Age-Adjusted Rate	Float64	Float64	Mortality_Age_Adjusted_Rate	22	-	-
Mortality-Incidence Age-Adjusted Rate Ratio	Float64	Float64	Mortality_Incidence_Age_Adjusted_Rate_Ratio	24	-	-
Population	Float64	Int64	Population	-	-	-

I have encountered a problem with the cancer data set for Nevada and Indiana in 2020. All numeric values were missing, and deleting these rows was not an option as they are important for the analysis. However, using the mean or median for imputation was not suitable due to the differences between the first year and last year's values. Incidence and mortality rates for different types of cancer are not the same. One potential solution could be to use the 2019 data to fill in missing values to keep up with trends. Nevertheless, it's important to acknowledge that this method has its limitations.

Also, mortality incidence rate ratio was missing in the District of Columbia in 2010 for two cancers. I manually calculated and then imputed it.

```
RangeIndex: 6171 entries, 0 to 6170
Data columns (total 11 columns):
 #   Column                                    Non-Null Count  Dtype
---  ------                                    --------------  -----
 0   State                                     6171 non-null   object
 1   Year                                      6171 non-null   int64
 2   State_Year                                6171 non-null   object
 3   Gender                                    6171 non-null   category
 4   Leading_Cancer_Sites                      6171 non-null   category
 5   Incidence_Counts                          6171 non-null   int64
 6   Incidence_Age_Adjusted_Rate               6171 non-null   float64
 7   Death_Counts                              6171 non-null   int64
 8   Mortality_Age_Adjusted_Rate               6171 non-null   float64
 9   Mortality_Incidence_Age_Adjusted_Rate_Ratio 6171 non-null   float64
 10  Population                                6171 non-null   int64
```

2: County Health Rankings & Roadmaps (CHR&R)

#	Column	Non-Null Count	Dtype
0	State	561 non-null	object
1	Year	561 non-null	float64
2	State-Year	561 non-null	object
3	Deaths	459 non-null	object
4	Years of Potential Life Lost Rate	561 non-null	object
5	% Fair or Poor Health	561 non-null	float64
6	Average Number of Physically Unhealthy Days	561 non-null	float64
7	Average Number of Mentally Unhealthy Days	561 non-null	float64
8	% Low birthweight	561 non-null	float64
9	% Adult smoking	561 non-null	float64
10	% Adults with Obesity	561 non-null	float64
11	Food Environment Index	408 non-null	float64
12	% Physically Inactive	561 non-null	float64
13	% With Access to Exercise Opportunities	408 non-null	float64
14	% Excessive Drinking	561 non-null	float64
15	% Driving Deaths with Alcohol Involvement	408 non-null	float64
16	% Uninsured	561 non-null	float64
17	Primary Care Physicians Rate	561 non-null	float64
18	Mental Health Provider Rate	561 non-null	float64
19	% With Annual Mammogram	560 non-null	float64
20	% Flu Vaccinated	153 non-null	float64
21	% Unemployed	561 non-null	float64
22	% Children in Poverty	559 non-null	float64
23	Average Daily PM2.5	543 non-null	float64
24	% Severe Housing Problems	561 non-null	float64
25	Life Expectancy	153 non-null	float64
26	% Adults with Diabetes	561 non-null	float64
27	% Limited Access to Healthy Foods	508 non-null	float64
28	Drug Overdose Mortality Rate	408 non-null	float64
29	% Insufficient Sleep	306 non-null	float64
30	Median Household Income	561 non-null	float64
31	% Less Than 18 Years of Age	561 non-null	float64
32	% 65 and Over	561 non-null	float64
33	% Black	561 non-null	float64
34	% American Indian & Alaska Native	561 non-null	float64
35	% Asian	561 non-null	float64
36	% Native Hawaiian/Other Pacific Islander	561 non-null	float64
37	% Hispanic	561 non-null	float64
38	% Non-Hispanic White	561 non-null	float64
39	% Female	561 non-null	float64
40	% Rural	561 non-null	float64

I dealt with missing values by first removing columns with a missing value proportion greater than 20%. I also deleted a row where all values were NaN. After that, I addressed the remaining columns with missing values by using the mean of each column or the mean of that state to handle the missing values.

Figure 1. Columns with missing values

%_With_Annual_Mammogram	1
%_Unemployed	0
%_Children_in_Poverty	2
Average_Daily_PM2.5	18
%_Severe_Housing_Problems	0

I encountered missing values in some columns where the proportion of missing data was over 20%, so I decided to remove those columns from my analysis. Additionally, I found a single row with NaN values in all columns, which I also removed. Even after dropping these columns, I still had three columns with missing values. To address this, I calculated the mean for each column or used the mean for that specific state to handle the missing values.

Following the handling of missing values, I identified and managed any duplicate entries and checked for mixed data types in columns. Once both databases were cleaned, I merged them to create a single, clean database ready for analysis. Finally, I added a new column called "region" based on the location of each state.

Region	
South	2057
West	1573
Midwest	1452
Northeast	1089

Final Cleaned Data set

#	Column	Non-Null Count	Dtype
0	State	6171 non-null	object
1	Year	6171 non-null	int64
2	State_Year	6171 non-null	object
3	Gender	6171 non-null	object
4	Leading_Cancer_Sites	6171 non-null	object
5	Incidence_Counts	6171 non-null	int64
6	Incidence_Age_Adjusted_Rate	6171 non-null	float64
7	Death_Counts	6171 non-null	int64
8	Mortality_Age_Adjusted_Rate	6171 non-null	float64
9	Mortality_Incidence_Age_Adjusted_Rate_Ratio	6171 non-null	float64
10	Population	6171 non-null	int64
11	%_Fair_or_Poor_Health	6171 non-null	float64
12	Average_Number_of_Physically_Unhealthy_Days	6171 non-null	float64
13	Average_Number_of_Mentally_Unhealthy_Days	6171 non-null	float64
14	%_Low_birthweight	6171 non-null	float64
15	%_Adult_smoking	6171 non-null	float64
16	%_Adults_with_Obesity	6171 non-null	float64
17	%_Physically_Inactive	6171 non-null	float64
18	%_Excessive_Drinking	6171 non-null	float64
19	%_Uninsured	6171 non-null	float64
20	Primary_Care_Physicians_Rate	6171 non-null	float64
21	Mental_Health_Provider_Rate	6171 non-null	float64
22	%_With_Annual_Mammogram	6171 non-null	float64
23	%_Unemployed	6171 non-null	float64
24	%_Children_in_Poverty	6171 non-null	float64
25	Average_Daily_PM2.5	6171 non-null	float64
26	%_Severe_Housing_Problems	6171 non-null	float64
27	%_Adults_with_Diabetes	6171 non-null	float64
28	Median_Household_Income	6171 non-null	int64
29	%_Less_Than_18_Years_of_Age	6171 non-null	float64
30	%_65_and_Over	6171 non-null	float64
31	%_Black	6171 non-null	float64
32	%_American_Indian_&_Alaska_Native	6171 non-null	float64
33	%_Asian	6171 non-null	float64
34	%_Native_Hawaiian/Other_Pacific_Islander	6171 non-null	float64
35	%_Hispanic	6171 non-null	float64
36	%_Non_Hispanic_White	6171 non-null	float64
37	%_Female	6171 non-null	float64
38	%_Rural	6171 non-null	float64
39	Region	6171 non-null	object

Final Data Profile

Column	Description
State	The U.S. state or territory where the cancer cases or deaths were recorded.
Year	The calendar year in which the cancer cases or deaths occurred.
Gender	The sex of the individuals (male or female) whose cancer data is being reported.
Leading Cancer Sites	The primary organs or tissues where the highest incidence of cancer is found.
Incidence Count	The number of new cancer cases diagnosed within the specified population and time period.
Incidence Age-Adjusted Rate	The rate of new cancer cases per 100,000 population, adjusted to the age distribution of a standard population (year 2000 U.S. standard population).
Death Count	The number of deaths attributed to cancer within the specified population and time period.
Mortality Age-Adjusted Rate	The rate of cancer deaths per 100,000 population, adjusted to the age distribution of a standard population (year 2000 U.S. standard population).
Mortality Incidence Age-Adjusted Rate Ratio (MIR)	The ratio of the age-adjusted mortality rate to the age-adjusted incidence rate, indicating the proportion of diagnosed cases that result in death.
Population	The total population of the specified area and demographic group, used as the denominator in calculating incidence and mortality rates.
% Fair or Poor Health	Percentage of adults that report fair or poor health
Average Number of Physically Unhealthy Days	Average number of reported physically unhealthy days per month
Average Number of Mentally Unhealthy Days	Average number of reported mentally unhealthy days per month
% Low birthweight	Percentage of births with low birthweight (<2500g)
% Smokers	Percentage of adults that reported currently smoking
% Adults with Obesity	Percentage of adults that report BMI >= 30

% Physically Inactive	Percentage of adults that report no leisure-time physical activity
% Excessive Drinking	Percentage of adults that report excessive drinking
% Uninsured	Percentage of people under age 65 without insurance
Primary Care Physicians Rate	Primary Care Physicians per 100,000 population
Mental Health Provider Rate	Mental Health Providers per 100,000 population
% With Annual Mammogram	Percentage of female Medicare enrollees having an annual mammogram (age 65-74)
% Unemployed	Percentage of population ages 16+ unemployed and looking for work
% Children in Poverty	Percentage of children (under age 18) living in poverty
Average Daily PM2.5	Average daily amount of fine particulate matter in micrograms per cubic meter
% Severe Housing Problems	Percentage of households with at least 1 of 4 housing problems: overcrowding, high housing costs, or lack of kitchen or plumbing facilities
Diabetes prevalence	Percentage of adults aged 20 and above with diagnosed diabetes.
Median household income*	The income where half of households in a county earn more and half of households earn less.
% below 18 years of age	Percentage of population below 18 years of age.
% 65 and older	Percentage of population ages 65 and older.
% Non-Hispanic Black	Percentage of population that is non-Hispanic Black or African American.
% American Indian & Alaska Native	Percentage of population that is American Indian or Alaska Native.
% Asian	Percentage of population that is Asian.
% Native Hawaiian/Other Pacific Islander	Percentage of population that is Native Hawaiian or Other Pacific Islander.
% Hispanic	Percentage of population that is Hispanic.
% Non-Hispanic White	Percentage of population that is non-Hispanic White.
% Females	Percentage of population that is female.
% Rural	Percentage of population living in a rural area.

Limitations and Biases:

Both the United States Cancer Statistics (USCS) and County Health Rankings & Roadmaps (CHR&R) datasets offer comprehensive and valuable insights into public health, yet they come with inherent biases. The USCS data, sourced from the CDC's NPCR, NCI's SEER Program, and the NCHS, may face underreporting and misclassification issues due to varying diagnostic criteria and reporting standards across regions. Regional variability in healthcare infrastructure and the impact of temporal changes, such as improved diagnostic technology and the COVID-19 pandemic, can further skew the data. Socioeconomic factors affecting healthcare access and reporting might also lead to underrepresentation of certain populations, thereby introducing biases in the analysis of cancer incidence and mortality rates.

Similarly, the CHR&R dataset, which aggregates health measures across nearly every U.S. County from various national and state sources, is subject to biases stemming from inconsistent data collection methods and the use of national averages to fill gaps for smaller populations. Measurement errors, particularly in small counties, and the effects of age adjustment can obscure true health burdens. Changes in data collection methods over time and regional differences in reporting practices also affect the comparability of health outcomes. Additionally, socioeconomic and policy factors influencing health measures can introduce biases, as regions with better resources and effective policies may show improved health outcomes not solely due to underlying health conditions but due to superior infrastructure and services.

Both datasets are invaluable for public health research, but users must be aware of their potential biases. Recognizing these biases is crucial for accurately interpreting the data and making informed decisions.

Questions

- What are the trends in cancer incidence and mortality rates across different states from 2010 to 2020?
- How do socioeconomic factors influence cancer incidence and mortality rates across different counties?
- Which socioeconomic factors (e.g., income, education, employment) have the most significant impact on cancer outcomes?
- Are there significant disparities in cancer outcomes between urban and rural areas?
- How do health behaviors, clinical care access, and environmental factors contribute to these disparities?
- How have changes in healthcare policies or public health interventions impacted cancer outcomes over the past decade?

Hypotheses

Hypothesis 1:

Null Hypothesis (H0): There is no significant trend in cancer incidence and mortality rates across different states from 2010 to 2020.

Alternative Hypothesis (H1): There is a significant trend in cancer incidence and mortality rates across different states from 2010 to 2020, varying by demographic factors.

Hypothesis 2:

Null Hypothesis (H0): Socioeconomic factors do not significantly influence cancer incidence and mortality rates across different counties.

Alternative Hypothesis (H1): Socioeconomic factors significantly influence cancer incidence and mortality rates across different counties, with certain factors having a more pronounced impact.

Hypothesis 3:

Null Hypothesis (H0): There are no significant disparities in cancer outcomes between urban and rural areas.

Alternative Hypothesis (H1): There are significant disparities in cancer outcomes between urban and rural areas, influenced by health behaviors, clinical care access, and environmental factors.