





Chronic Disease Indicators

DSE 241 - Tarek Tarif



Motivation

<u>Disease Patterns</u>	The increasing prevalence of chronic diseases worldwide necessitates a comprehensive understanding of these conditions.
<u>Health Interventions</u>	Help identify high-risk populations, enabling the implementation of preventive measures and early interventions.
<u>Resource Allocation</u>	Guide public health officials and policymakers in prioritizing resources for the most prevalent or severe conditions.
<u>Monitoring/Evaluation</u>	Assess whether interventions are leading to desired health outcomes and inform necessary adjustments.



Dataset

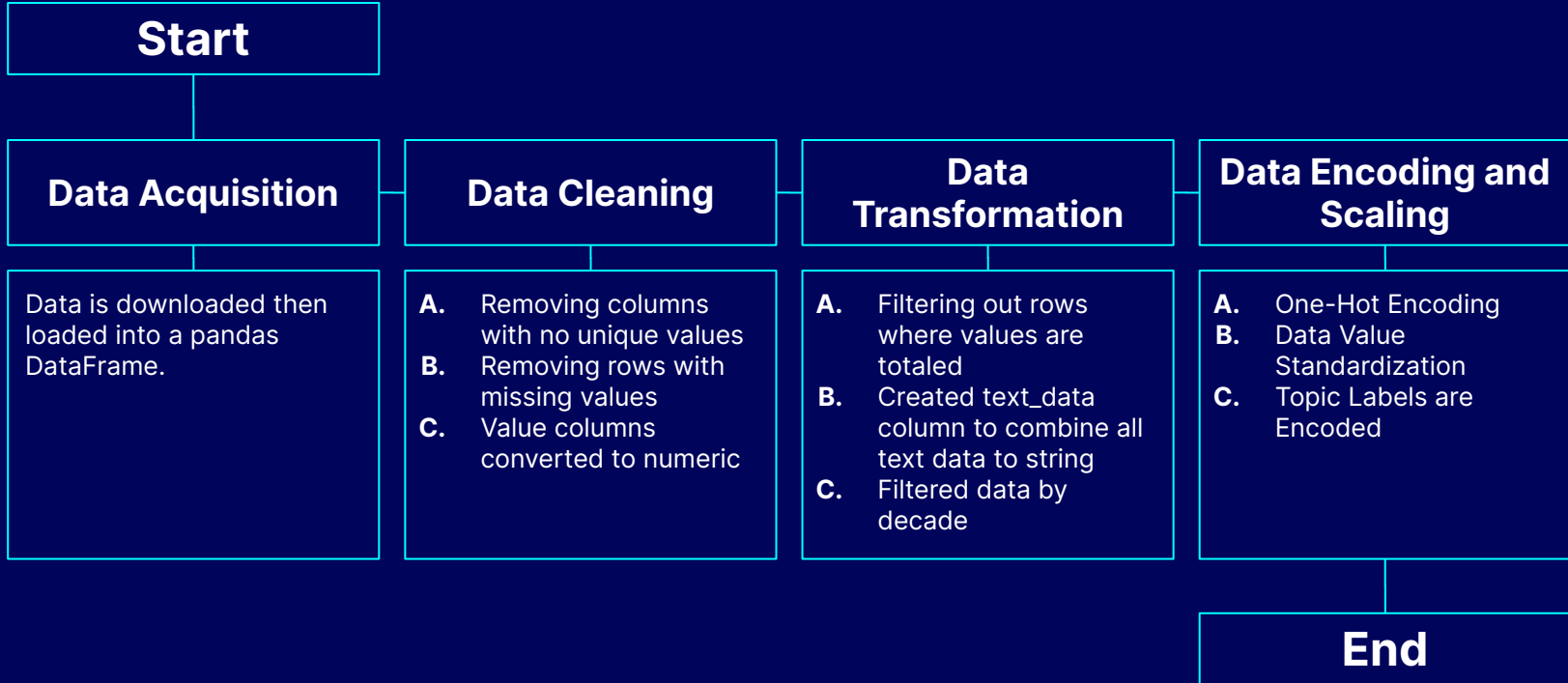
Dataset Source: [US Chronic Disease Indicators](#)

- The dataset is a collection of chronic disease indicators in the U.S., with data spanning from the 2001 to 2021. It includes information about the location, data source, topic, and question related to each data point.
- The dataset provides detailed data values, including the unit and type of the data value, and confidence limits to indicate the reliability of the data.
- The dataset incorporates stratification to allow for more specific analysis. This includes stratification categories and identifiers, allowing for detailed breakdowns such as gender and ethnicity.. It also includes geolocation coordinates for mapping and various identifiers for the location, topic, question, response, and data value type.

#	Column	Non-Null Count	Dtype
0	YearStart	794819 non-null	int64
1	YearEnd	794819 non-null	int64
2	LocationAbbr	794819 non-null	object
3	LocationDesc	794819 non-null	object
4	DataSource	794819 non-null	object
5	Topic	794819 non-null	object
6	Question	794819 non-null	object
7	DataValueUnit	691045 non-null	object
8	DataValueType	794819 non-null	object
9	DataValue	794819 non-null	float64
10	DataValueAlt	794819 non-null	float64
11	DataValueFootnoteSymbol	10723 non-null	object
12	DataValueFootnote	10723 non-null	object
13	LowConfidenceLimit	674338 non-null	float64
14	HighConfidenceLimit	674338 non-null	float64
15	StratificationCategory1	794819 non-null	object
16	Stratification1	794819 non-null	object
17	GeoLocation	794819 non-null	object
18	LocationID	794819 non-null	int64
19	TopicID	794819 non-null	object
20	QuestionID	794819 non-null	object
21	DataValueTypeID	794819 non-null	object
22	StratificationCategoryID1	794819 non-null	object
23	StratificationID1	794819 non-null	object
24	TextData	794819 non-null	object



Data Wrangling

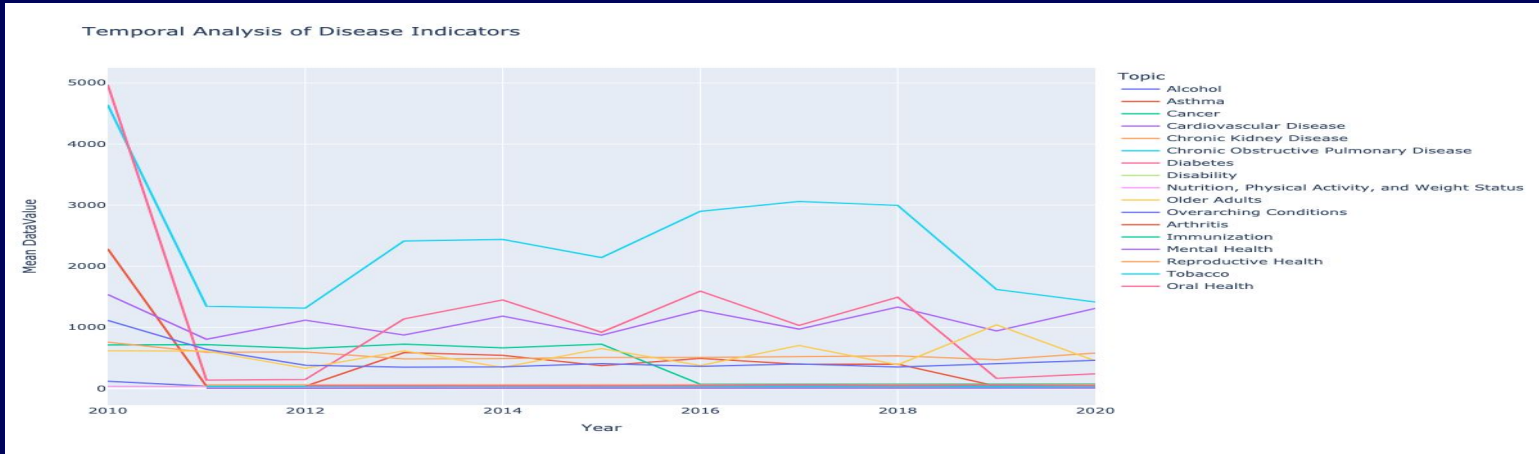


Tasks Outlined

Task	Problem	Visualization Purpose	Data Transformation	Use Case Scenarios
Temporal Analysis	Understand how various disease indicators have changed over time.	Data exploration and hypothesis confirmation.	Group the data by year and indicator, then plot the trends over time.	Public health officials can use this visualization to track the progression of different diseases and prioritize interventions accordingly.
Geographic Comparison	Identify geographic disparities in disease prevalence or incidence rates.	Data exploration and hypothesis confirmation.	Group the data by location and calculate summary statistics for each disease indicator.	Policymakers can use this visualization to allocate resources to regions with higher disease burdens.
Demographic Analysis	Explore how disease indicators vary across different demographic groups.	Data exploration and hypothesis confirmation.	Group the data by demographic variables (e.g., gender, age, race/ethnicity) and calculate summary statistics for each disease indicator.	Healthcare providers can use this visualization to identify demographic groups with higher disease burdens and tailor interventions accordingly.
Word Cloud	Understand the frequency of words used in disease reports.	Data exploration and hypothesis confirmation.	Combine all text data into a single string, count the frequency of each word, and generate a word cloud.	Researchers can use this to identify common words or themes in disease reports.
Predictive Modeling	Predict the disease indicator based on various features.	Model evaluation and performance assessment.	Preprocess the data, split the data into training and testing sets, train a model, make predictions, and evaluate the model's performance.	Healthcare providers and policymakers can use this model to anticipate disease indicators based on specific features

1 - Temporal Analysis

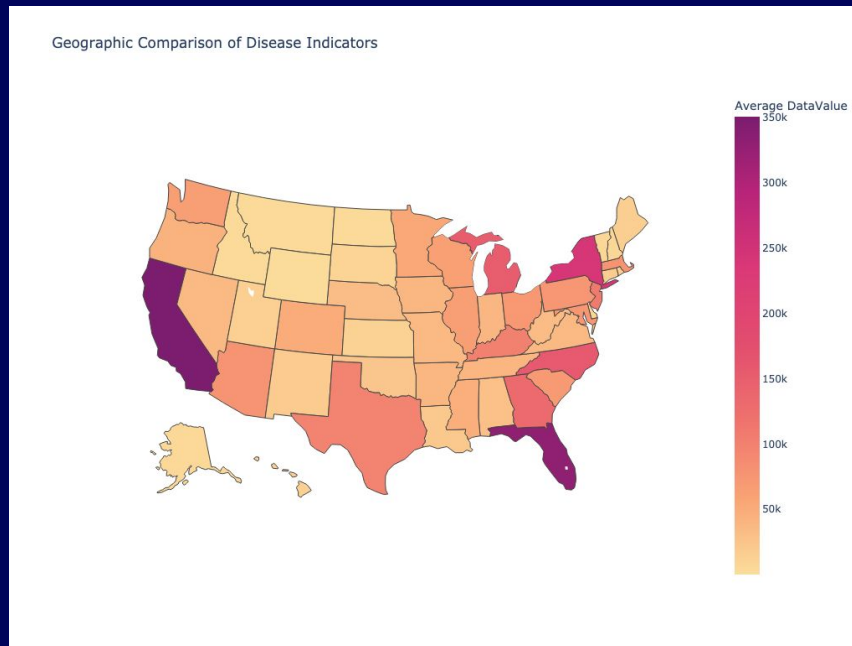
- This task involves analyzing the progression of disease indicators over time, providing a perspective on disease prevalence.
- Purpose: Identify trends and patterns to predict future disease prevalence and plan preventive measures.
- The line chart visualization allows for interactability by allowing the user to modify time periods and by the mean data value.



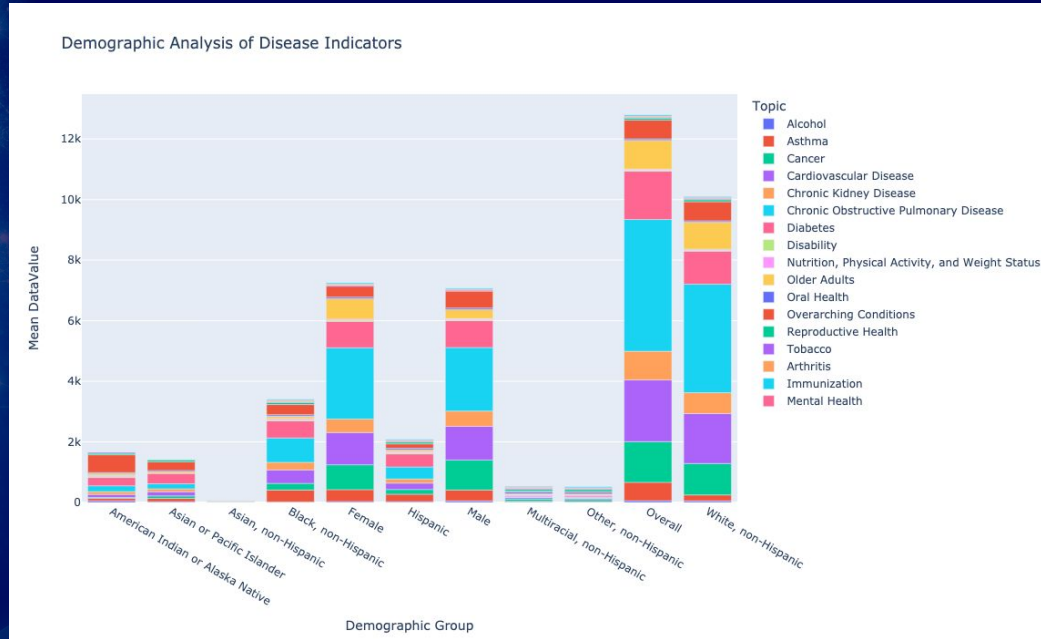
- Large amount (> 250k responses)

Identify hotspots for specific diseases.

Implementation allows for filtering by year and topic.



3 - Demographic Analysis



Visualization helps identify demographic groups with higher disease burdens, enabling healthcare providers to tailor interventions to these specific populations.

The stacked bar charts used in this task effectively illustrates these disparities, making it easy to compare disease prevalence across different demographic categories.



-

5 - Predictive Modeling

XGB Classifier

- This task involves using machine learning techniques to predict disease indicators based on various features.
- The performance of the predictive model is evaluated using metrics such as accuracy, confusion matrices, and classification reports, providing a measure of the model's predictive accuracy.

Region	Accuracy Score
Logistic Regression	0.13
Random Forest	0.27
XGBoost	0.50

Confusion Matrix

0	4110	469	444	12	173	1132	761	366	5	1	309	833	292	40	1179	14	508
1	80	7764	32	11	1	152	36	1754	6	23	0	3007	0	128	358	0	287
2	332	2069	1480	144	26	1173	1313	1654	0	3	1	325	352	32	730	1	160
3	0	18	24	30606	0	9	29	703	2	1	0	140	2	218	110	2	0
4	78	0	8	0	1826	577	2056	19	0	0	0	0	0	0	1	0	67
5	685	2155	488	49	262	6061	8104	3695	9	5	2	521	412	80	594	0	67
6	103	2728	226	176	81	1750	17868	3063	2	6	2	1506	102	71	478	13	86
7	380	2671	687	327	199	1856	3628	9005	4	14	0	917	107	264	602	1	295
8	1	409	1	0	0	0	0	1	151	1	0	208	3	5	6	0	0
9	2	1257	0	0	0	33	3	172	1	23	0	677	0	7	20	0	6
10	494	37	19	0	0	7	0	20	0	0	683	101	0	1	1191	0	144
11	71	3982	57	92	1	81	63	1954	7	43	0	8108	7	188	549	4	316
12	124	688	220	13	6	310	853	709	3	4	0	228	1220	64	164	7	119
13	43	686	60	144	0	94	34	1371	5	11	0	355	18	710	439	23	121
14	467	476	237	225	13	405	1180	856	16	13	482	1424	5	268	8738	33	389
15	12	147	0	131	0	1	33	338	0	3	0	261	9	53	252	151	7
16	374	1126	208	57	351	453	36	2329	2	13	4	1618	3	112	860	3	1493
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Predicted

Solutions

Task	Objective	Visualization	Results	Implications
Temporal Analysis	Understand how various disease indicators have changed over time.	Interactable line charts to display trends over time.	Identification of diseases showing itrends.	Helps in predicting future disease prevalence and planning preventive measures.
Geographic Comparison	Identify geographic disparities.	Choropleth maps to visualize disease prevalence across states.	Identification of regions with higher disease burdens.	Enables policymakers to prioritize interventions in high-prevalence areas.
Demographic Analysis	Analyze disease indicators across demographics.	Stacked bar chart to visualize disease prevalence across demographic categories.	Identification of demographic groups with higher disease burdens.	Helps healthcare providers tailor interventions to specific demographics.
Word Cloud	Understand the frequency of words used in disease reports.	Word cloud to visualize the frequency of words used in disease reports.	Identification of prevalent trends within responses	Informs emerging trends based on textual data
Predictive Modeling	Predict the disease indicator based on select features.	Visualizations of model performance metrics with a confusion matrix and classification report.	50% Accuracy Score	Enables healthcare providers, policymakers, and researchers to plan preventive measures and allocate resources effectively.

Questions?

