Module 2: Final Project – MILESTONE 1: EDA

By

Elvis Opoku, Kevin Tushar Pandya, Neha Ajinkya Nagurkar, Vidhi Manojkumar Patani,

Smit Pareshbhai Ranpariya

College of Professional Studies: Northeastern University

Prof. Shahram Sattar

Subject: ALY6040: DATA MINING APPLICATION

Date: April 28th, 2024

Final Project – MILESTONE 1: EDA

INTRODUCTION

The Disability and Health Data System (DHDS) data set is an extensive online collection of state-level data on persons with impairments. It covers six functional impairment types: cognition, hearing, mobility, vision, self-care, & independent living. DHDS allows users to access critical data on the physical and psychological well-being of people with disabilities, providing insights into their difficulties and needs across a variety of disciplines.

Initially, the dataset had 644,356 items and 32 columns that included a wide variety of disability and health-related characteristics. However, following extensive data pre-processing as well as cleaning methods, such as resolving missing values and deleting extraneous columns, the collection of data was reduced to 644,356 items and 17 columns. This guaranteed that the data was ready for future analysis and research, allowing for important insights into the features and patterns of disability among people.

OBJECTIVE

The primary goal of this research is to examine the DHDS dataset and uncover patterns, trends, and discoveries about disability and health in adults. Specifically, the analysis focuses at:

- Preprocess the dataset to verify its integrity and completeness.
- Conduct EDA to better understand the distribution of and interactions between variables.
- Identify any unusual values, missing numbers, or discrepancies in the dataset.
- Extract valuable information to help shape public health initiatives and laws.

SCOPE

The research focuses on the DHDS dataset, which contains data on disability kinds, health indicators, demographic variables, and geographic areas. The scope covers data preparation, visualization, & statistical analysis to acquire a thorough knowledge of adult disability and health patterns.

DATA DESCRIPTION

Final Project – MILESTONE 1: EDA

The DHDS dataset has 644,356 records and 32 columns. Before the cleaning process, the dataset comprised multiple columns, which include 'Year', 'LocationAbbr', 'LocationDesc', 'Category', 'Indicator', 'Response', 'Data_Value_Unit', 'Data_Value_Type', 'Data_Value', 'Low_Confidence_Limit', 'High_Confidence_Limit', 'Number', 'WeightedNumber', 'StratificationCategory1', 'Stratification1', 'CategoryID', as well as 'IndicatorID'.

Following cleaning of data, the dataset was improved to contain 644,356 items and 17 columns. Missing values were found in multiple columns, notably 'Data_Value', 'Low_Confidence_Limit', 'High_Confidence_Limit', 'Number', & 'WeightedNumber'. To solve the, a data imputation approach was used, in which missing values were supplied by indicator means.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Year	644356.0	NaN	NaN	NaN	2018.437783	1.69878	2016.0	2017.0	2018.0	2020.0	2021.0
LocationAbbr	644356	65	DC	10068	NaN	NaN	NaN	NaN	NaN	NaN	NaN
LocationDesc	644356	65	District of Columbia	10068	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Category	644356	8	Health Risks & Behaviors	147336	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Indicator	644356	42	Body mass index category among adults 18 years...	36864	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Response	635140	62	No	183218	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Data_Value_Unit	644356	1	%	644356	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Data_Value_Type	644356	2	Age-adjusted Prevalence	483384	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Data_Value	578578.0	NaN	NaN	NaN	42.25289	28.947211	0.1	17.5	35.9	66.2	100.0
Low_Confidence_Limit	578578.0	NaN	NaN	NaN	37.898096	28.456481	0.0	13.6	30.0	60.3	99.9
High_Confidence_Limit	578578.0	NaN	NaN	NaN	46.728883	29.148859	0.1	21.8	42.3	71.8	100.0
Number	578578.0	NaN	NaN	NaN	1518.155129	7131.018111	1.0	122.0	344.0	1004.0	327817.0
WeightedNumber	578578.0	NaN	NaN	NaN	856674.406697	4137684.208675	43.0	43178.0	147578.5	495804.25	181223676.0
StratificationCategory1	644356	3	Disability Status	479228	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Stratification1	644356	9	Any Disability	239614	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CategoryID	644356	8	HLTHRB	147336	NaN	NaN	NaN	NaN	NaN	NaN	NaN
IndicatorID	644356	42	BMI	36864	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Fig: 1: Summary Statistics: Pre & Post-Data Cleaning and Imputation

DATA IMPUTATION: HANDLING MISSING VALUES

Missing values are ubiquitous in datasets & can have a substantial influence on the accuracy and reliability of results. Missing values were found in several columns of the DHDS dataset, including 'Data_Value', 'Low_Confidence_Limit', 'High_Confidence_Limit', 'Number', and 'WeightedNumber'. To resolve the missing values and verify the dataset's comprehensiveness, a systematic technique towards data imputation was used.

Final Project – MILESTONE 1: EDA

Initially, the 'Data_Value' column was analyzed using the 'Indicator' to find missing data. It was thought that missing values in this field were inadvertent and needed imputation to ensure data integrity. The values that were missing for 'Data_Value', 'Low_Confidence_Limit', 'High_Confidence_Limit', 'Number', and 'WeightedNumber' had been then imputed using the average values computed for each indicator.

The data imputation procedure included the following steps:

- Indicator Means Calculation: The groupby function in pandas was used to generate mean values for 'Data_Value', 'Low_Confidence_Limit', 'High_Confidence_Limit', 'Number', and 'WeightedNumber'.
- Supplying Missing Value: The values that were missing in each of the columns were filled with the computed mean values for the related indicators. This was accomplished by mapping the average values to the relevant indicators and populating the values that were missing accordingly.

	Year	LocationAbbr	LocationDesc	Category	Indicator	Response	Data_Value_Unit	Data_Value_Type	Data_Value	Low_Confidence_Limit	...	Res ca
0	2019	DC	District of Columbia	Demographics	Employment status among adults 18 years of age...	Employed	%	Age-adjusted Prevalence	26.303808	21.790104	...	
1	2020	NJ	New Jersey	Demographics	Education level among adults 18 years of age o...	College Graduate	%	Age-adjusted Prevalence	15.000000	11.700000	...	
2	2020	WV	West Virginia	Demographics	Education level among adults 18 years of age o...	High School Graduate	%	Age-adjusted Prevalence	69.000000	63.300000	...	
3	2020	RI	Rhode Island	Demographics	Income level among adults 18 years of age or o...	$25,000 < 35,000$	%	Age-adjusted Prevalence	10.700000	7.200000	...	
4	2020	HHS8	HHS Region 8	Demographics	Income level among adults 18 years of age or o...	$25,000 < 35,000$	%	Age-adjusted Prevalence	13.500000	11.100000	...	

5 rows × 88 columns

Fig: 2: Summary Statistics Post-Cleaning and Imputation

Final Project – MILESTONE 1: EDA

EXPLORATORY DATA ANALYSIS (EDA)

DISTRIBUTION OF DATA VALUES

Our investigation began by examining a histogram, which provided a complete picture of the variation of data values. This visualization allows us to determine the dataset's central tendency, dispersion, and skewness. We learned about the fundamental distribution pattern and probable outliers by dividing the information into intervals and indicating the frequency on occurrences inside each one.

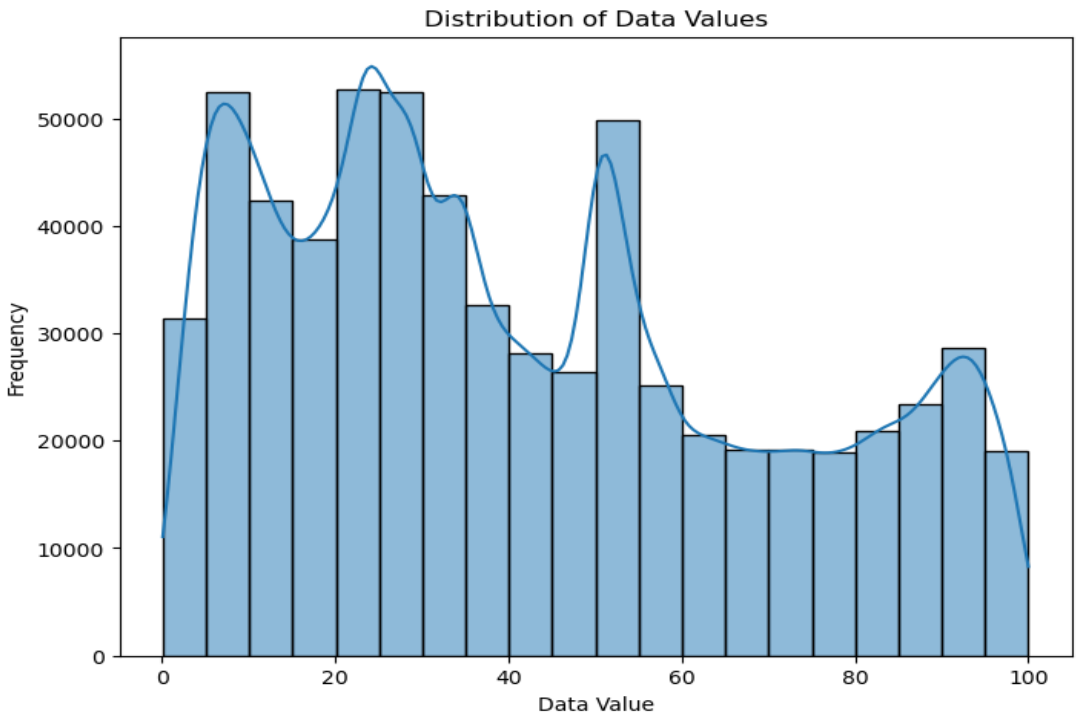

Fig: 3: Histogram of Data Values

TOP 10 LOCATIONS BY RESPONSE COUNT

Following on to the bar plots, we looked at the top ten locations according to response counts. This image allowed us to identify the geographic locations with the greatest levels of data gathering activities. By categorizing the locations in order of decreasing of response count, we identified areas of interest for additional investigation and resource allocation.

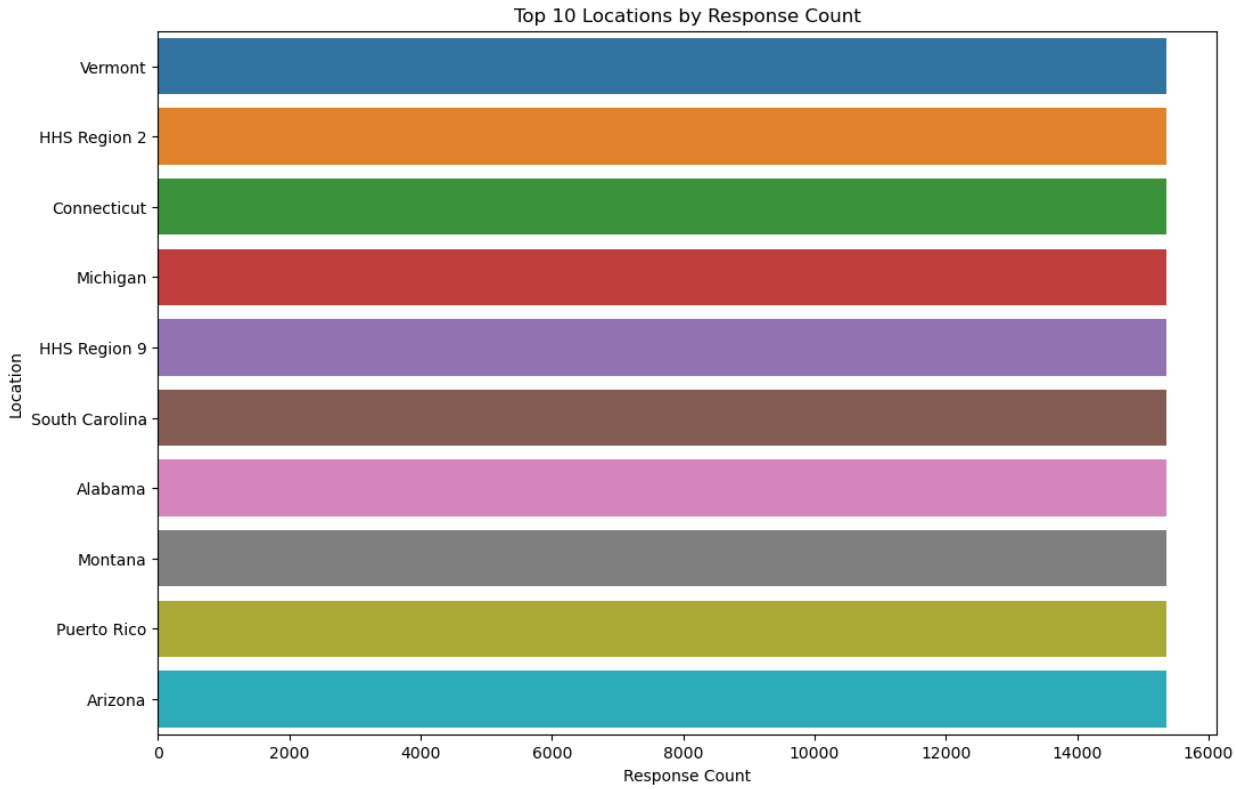

Fig: 4: Bar plot showing top 10 locations by response count

COUNT OF STRATIFICATION1 CATEGORIES

Using a count plot, we examined the frequency distribution of categories inside the 'Stratification1' variable. This visualization gave a thorough breakdown of each category's frequency, revealing the dataset's demographic mix. We obtained a better picture of this dataset's demographic variety by examining how numbers were distributed across different categories.

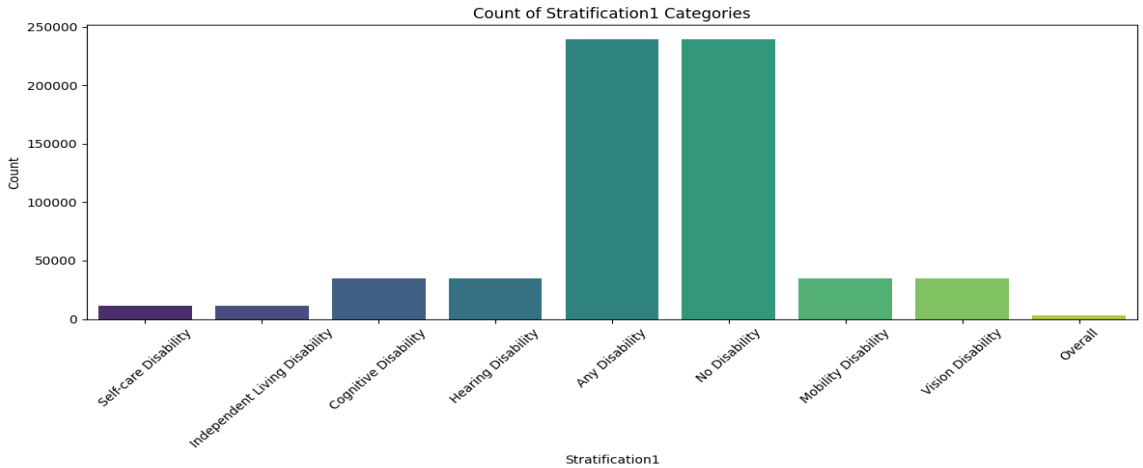

Fig: 5: Count of Stratification1 Categories

Final Project – MILESTONE 1: EDA

RESPONSE PERCENTAGE DISTRIBUTION BY STRATIFICATION1

Using violin plots, we investigated the pattern of distribution of response proportions among the 'Stratification1' categories. This depiction represented the central tendency, dispersion, and shape of answer percentage distributions across each demographic group. By displaying a density of data points throughout the violin plot, we were able to identify trends and variances in response percentages among demographic groupings.

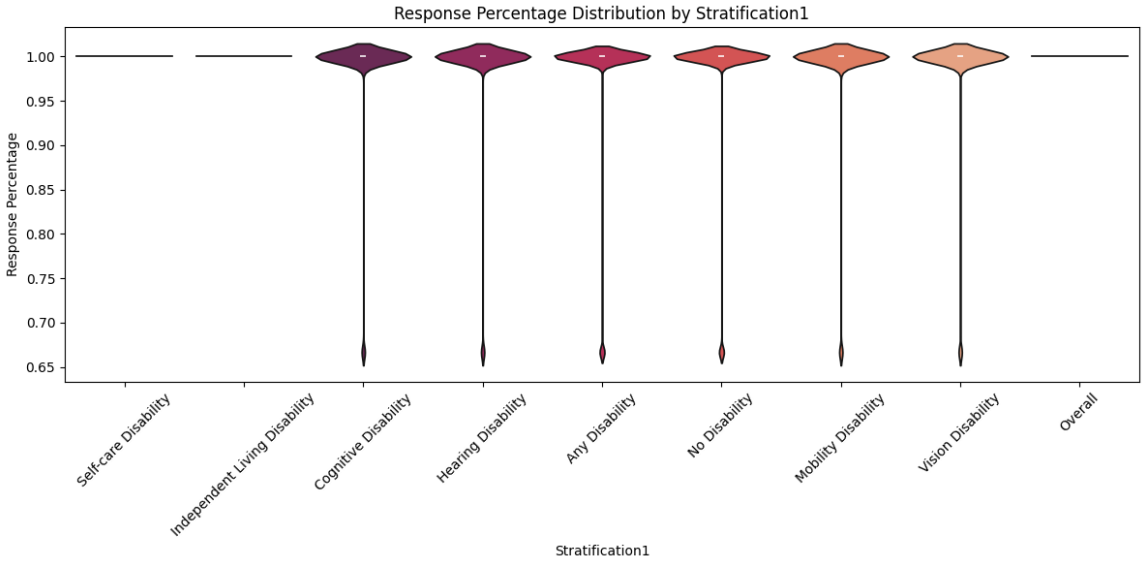

Fig: 6: Violin plot of Response Percentage by Stratification1

DISTRIBUTION OF CATEGORIES BY STRATIFICATION1

We used a count plot to evaluate the distribution among categories inside 'Stratification1' among various demographic groupings. This visualization allowed for the comparison during category frequencies inside each demographic category, revealing information on the relative frequency of distinct categories across various demographic groups.

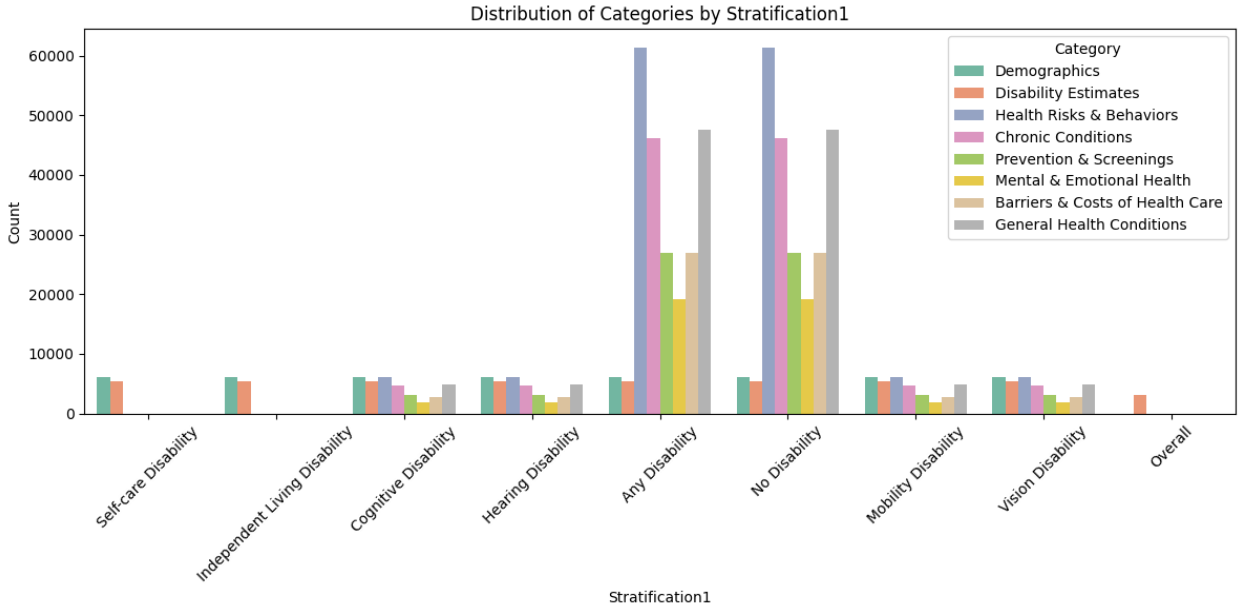

Fig: 7: Categories distributed by Stratification1

PROPORTION OF STRATIFICATION1 CATEGORIES

Finally, we used a pie chart to determine the fraction of 'Stratification1' groups. This visualization gave a comprehensive perspective of the demographic makeup by displaying each category's proportionate contribution to the overall population. We learned about the distribution of different demographic groupings in the dataset by presenting proportions as pie slices.

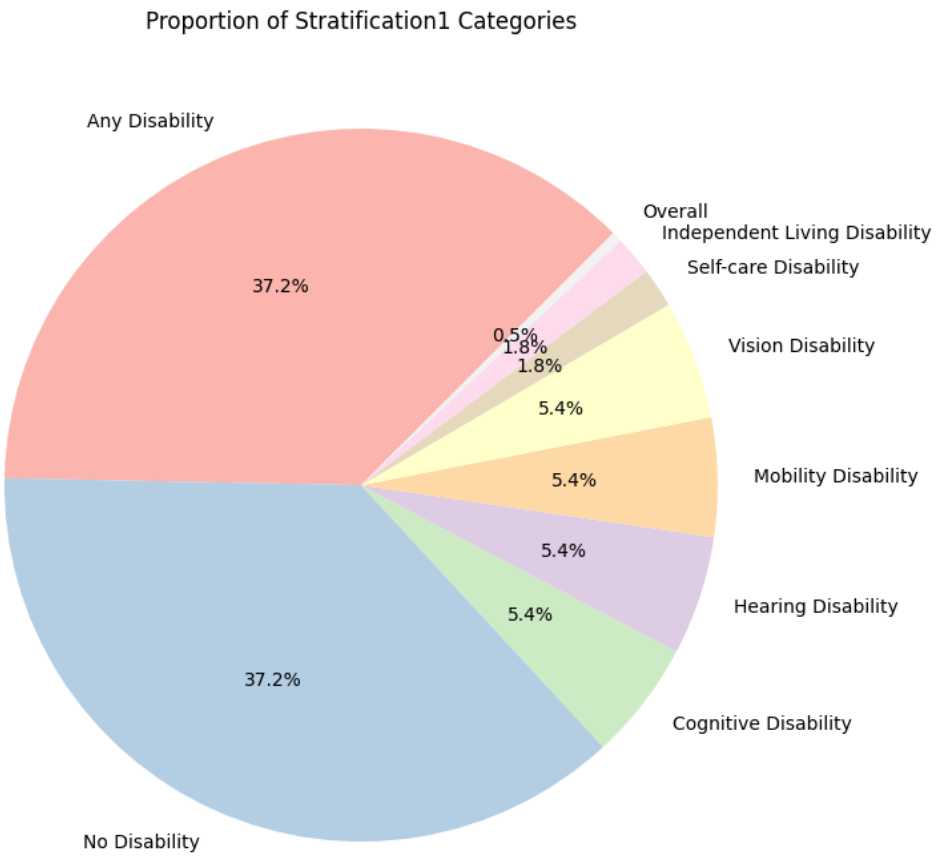

Fig: 8: Pie chart followed by Stratification1 Proportion

FINDINGS AND INSIGHTS

Several major discoveries resulted from the exploratory data analysis (EDA) conducted on the Disability and Health Data System (DHDS) data set, providing insight into adult disability and health trends.

- The violin plot demonstrated the difference in response percentages between demographic groupings, which was an unexpected observation. This visualization showed differences in health outcomes as well as accessibility to healthcare services across demographic groups, emphasizing the significance of tackling health injustices.

- Furthermore, the geographic distribution of categories amongst demographic groupings revealed information on the frequency of various health disorders or socioeconomic

variables among different segments of the population. Variations in work position and earnings levels highlighted the link between socioeconomic determinants and health consequences.

These findings emphasize the need for extensive gathering and analyzing information in studies on public health. Policymakers, healthcare providers, and academics may use insights obtained from the DHDS dataset to create evidence-based policies for improving health equality and reducing inequities among individuals with disabilities.

PROPOSED NEXT STEPS

The following phases will include more in-depth statistical studies to investigate correlations and connections between variables. This includes:

- CORRELATION ANALYSIS: Analyzing the association between disability kinds and health indicators to uncover potential risk factors & comorbidities.
- REGRESSION ANALYSIS entails developing prediction models to better understand the influence of socioeconomic determinants on health outcomes in persons with disabilities.
- SPATIAL ANALYSIS: Investigating spatial variations in disability prevalence & healthcare access to help guide resource allocation along with service delivery.

Integrating qualitative data, which might include interviews or surveys, will assist contextualize quantitative findings and highlight unmet needs and priorities in the community.

BUSINESS QUESTIONS ADDRESSED

From a commercial standpoint, the data mining initiatives that utilize the DHDS dataset attempt to answer fundamental issues about healthcare service, resource allocation, especially policy development:

- Healthcare Service Tailoring: How should healthcare organizations adapt their services to better suit the requirements of individuals with disabilities, taking into account

differences in health outcomes as well as access to treatment among demographic groups?

- Addressing Healthcare inequalities: What are the root causes of healthcare inequalities among adults alongside disabilities, and in what ways can policymakers overcome structural impediments to fair healthcare access?
- Creating Inclusive Workplaces: How can companies and hiring managers develop inclusive workplaces that meet the requirements of individuals with disabilities while also encouraging diversity and inclusion in the workforce?
- Economic Implications: What exactly are the economic consequences of disability-related healthcare spending, and how can stakeholders improve resource allocation & affordability in healthcare delivery?
- Technology & Accessibility: How can technologies be used to improve access and communication between healthcare practitioners and people with disabilities, therefore enhancing medical results and quality of life?

CONCLUSION

The examination of the Disability and Health Data System (DHDS) dataset revealed important insights into the intricacies of disability and health in people. By rigorously preparing and examining the data, we discovered large discrepancies in health outcomes as well as accessibility to healthcare services throughout demographic groups. These findings highlight the necessity of tailored interventions in addressing current inequities and promoting health equity.

Moving ahead, we will focus on sophisticated statistical analysis with qualitative information integration to gain a better understanding of the links between factors and actual experiences of people with disabilities. From a commercial standpoint, our data mining initiatives are intended to drive evidence-driven choices in the delivery of healthcare services, resource allocation, workforce inclusion, and economic policy.

In short, the DHDS dataset serves as a solid platform for developing comprehensive healthcare policies and procedures that address the different needs of people with disabilities. We can work

together and do thorough analysis to create a more equal and readily available healthcare ecosystem across all.

REFERENCE

The CDC (2024) is the Centers for Disease Control and Prevention. Regarding the Health and Disability Data System (DHDS). taken from the announcement page at https://www.cdc.gov/dhds.

The NCBDDD stands for the National Center on Birth Defects and Developmental Disabilities. (2024). Information & Figures Regarding Developmental Disability and Birth Defects. taken from data.html at https://www.cdc.gov/ncbddd