Social Health Index (SHI) Predictor Prototype: In-Depth Technical Report

Developed by: Ruchir Patel

Date: May 22, 2024

Table of Contents

1. Executive Summary

This report provides a comprehensive technical account of the development of a prototype Social Health Index (SHI) predictor. The primary objective was to construct a quantifiable measure of Social Determinants of Health (SDoH) applicable at both U.S. county and ZIP code granularities. This initiative leveraged publicly accessible datasets, primarily from the Centers for Disease Control and Prevention (CDC PLACES data) for core SDoH-related health outcomes and behaviors, and a consolidated SDoH dataset (largely based on the American Community Survey - ACS) for broader contextual indicators.

The methodology followed a systematic data processing pipeline, beginning with **data ingestion and preparation**, which involved loading raw CSV data and standardizing geographic identifiers, specifically ensuring County FIPS codes were in a 5-digit string format. This was followed by **targeted variable selection**. From CDCPlaces.csv, seven direct measures related to food insecurity, housing insecurity, transportation access, and social isolation/mental distress were chosen. Concurrently, from the SDoH2020Data_cleaned.csv, an initial set of approximately 46 variables pertinent to food, housing, and transportation was identified through keyword analysis, with a subsequent filter applied to retain only those columns exhibiting substantive data. A crucial **data transformation** step involved pivoting the CDCPlaces.csv data from a long to a wide format, aligning each county with its respective SDoH measures as distinct features.

Data integration was achieved by merging the processed CDC PLACES data with the selected SDoH2020 variables at the county level, employing a left merge strategy on CountyFIPS. The core of the project was the **county-level index construction**. This multi-stage process included handling missing data (NaNs) for all SHI constituent variables via median imputation; developing four domain-specific sub-indices (Food Insecurity, Housing Insecurity, Transportation Barriers, and Social Isolation); normalizing all constituent variables within each sub-index to a [0, 1] scale using Min-Max scaling to ensure equitable contribution (with variables oriented such that higher scores indicate greater adversity); calculating sub-index scores through unweighted averaging; and finally, computing an Overall_SHI by averaging the four sub-index scores.

Subsequently, **geographic mapping to the ZIP code level** was performed. This required acquiring a ZIP Code-to-County FIPS crosswalk via the HUD USPS API, processing this crosswalk data (including standardization of ZIP and CountyFIPS codes), and disambiguating ZIP codes that span multiple counties by prioritizing the county with the highest residential ratio. The county-level SHI scores were then merged with this processed crosswalk to assign an SHI to each ZIP code. The process concluded with **results exploration and contextualization**, which involved preliminary validation through internal consistency checks (correlation of Overall SHI with sub-indices) and interpretation of ZIP-level scores via percentile ranking and categorization.

The final output is a Pandas DataFrame (df_zip_shi) providing multi-dimensional SHI scores for approximately 39,490 U.S. ZIP codes. A conceptual discussion on leveraging Large Language

Models (LLMs) for enhancing SDoH data integration is also included, addressing its potential for future development. This report details the rationale behind each methodological choice, illustrates key data transformations, and discusses the structure of the resulting SHI predictor, aiming to provide a clear understanding for a technical audience.

2. Introduction and Objective

Understanding Social Determinants of Health (SDoH)

The conditions in which individuals are born, grow, live, work, learn, and age are collectively known as **Social Determinants of Health** (**SDoH**). These determinants encompass a wide array of economic and social factors, such as socioeconomic status, education level, neighborhood characteristics, employment opportunities, social support networks, and access to healthcare. Unlike clinical factors that relate directly to medical care, SDoH represent the broader environmental and societal influences that shape an individual's and a community's health.

The impact of SDoH on health outcomes is profound and well-documented. They are recognized as primary drivers of health inequities, contributing to disparities in life expectancy, chronic disease prevalence, and overall well-being across different population groups. For instance, limited access to nutritious food (food insecurity), unstable or unsafe housing (housing insecurity), lack of reliable transportation, and social isolation can all lead to poorer health outcomes and exacerbate existing health conditions.

Project Rationale and Objectives

Given the significant influence of SDoH, the ability to quantify and map these factors at granular geographic levels (such as counties and ZIP codes) is crucial. Such quantification allows for the identification of vulnerable populations, enabling more targeted interventions, informed policy development aimed at mitigating SDoH-related risks, and more equitable resource allocation to areas with the greatest need.

The primary objective of this project was to develop a robust prototype of a **Social Health Index** (**SHI**) **predictor**. This initiative aimed to:

1. Effectively combine and utilize SDoH indicators from disparate, publicly available datasets.
2. Construct a multi-dimensional SHI at the county-level, comprising several sub-indices that reflect key SDoH domains: Food Insecurity, Housing Insecurity, Transportation Barriers, and Social Isolation (including aspects of mental distress and lack of social support).
3. Translate these county-level SHI insights to the ZIP code level, providing a more localized assessment tool.
4. Employ clear, understandable, and replicable data processing and index construction techniques, primarily using the Pandas library for data manipulation and Scikit-learn for scaling.

5. Conceptually evaluate the potential role of advanced technologies, specifically Large Language Models (LLMs), in streamlining and enhancing SDoH data integration processes for future iterations or more complex models.

This technical report provides a detailed walkthrough of the project's life cycle, from data acquisition and preprocessing to index construction, geographic mapping, and initial results interpretation, thereby offering a comprehensive understanding of the SHI predictor prototype.

3. Data Sources and Acquisition

The development of the SHI predictor relied on three key data components, each acquired and processed to contribute to the final index.

The first primary source was the **CDC PLACES Data** (**CDCPlaces.csv**), obtained from the Centers for Disease Control and Prevention (CDC). This dataset served as a foundational element, providing model-based, small-area estimates for a multitude of health outcomes, risk behaviors, prevention practices, and health status indicators across various geographic levels, including counties. For this project, its primary utility was as a source for direct measures related to food insecurity, housing insecurity, transportation access, and indicators of social isolation and mental distress. The LocationID column within this dataset was identified and subsequently processed as the County FIPS code.

The second key dataset was a **Consolidated SDoH Indicators dataset** (**SDoH2020Data_cleaned.csv**). This comprehensive file (referenced via the path /Users/ruchirpatel/Downloads/SDOH2020.csv) appears to be a pre-compiled aggregation of SDoH variables, largely derived from the U.S. Census Bureau's American Community Survey (ACS) for the year 2020, potentially supplemented by indicators from other public sources such as the Area Health Resources Files (AHRF). Containing nearly 700 variables, it offered a rich source of contextual SDoH data related to demographics, socioeconomic status (income, poverty, employment), housing characteristics (type, cost, quality), and education, primarily at the county level. The COUNTYFIPS column in this dataset served as the county-level identifier.

Finally, to map county-level indices to ZIP codes, a **HUD USPS Crosswalk** was acquired. This data, facilitating the linkage between ZIP codes and County FIPS codes, was fetched programmatically from the U.S. Department of Housing and Urban Development (HUD) User Portal API, specifically its USPS Validation Service. The API endpoint https://www.huduser.gov/hudapi/public/usps was queried with parameters type=2 (specifying a ZIP-to-County crosswalk) and query=all. This attempt to retrieve a comprehensive national mapping required an authenticated API token. The successful API response provided a JSON object, which was then parsed into a pandas DataFrame (df_hud_crosswalk_raw). Key fields from this crosswalk included zip (the 5-digit ZIP code), geoid (identified as the 5-digit County FIPS code for type=2 responses), and res_ratio (the proportion of residential addresses in that ZIP code falling within the specified county), a crucial field for disambiguating ZIP codes that may span multiple counties.

4. Data Preparation and Preprocessing

A rigorous data preparation phase was undertaken to ensure the quality, consistency, and compatibility of the data for subsequent merging and index construction. All operations were performed using the Pandas library in Python unless otherwise specified.

4.1. Processing CDCPlaces.csv

The CDCPlaces.csv dataset underwent several transformations to extract and structure the core SDoH measures at the county level. The process commenced with loading the CSV file into a pandas DataFrame, df_cdc_places. Initial inspection confirmed its dimensions (240,886 rows, 22 columns) and general data structure. A notable observation was that the LocationID column, intended to represent County FIPS codes, was read as an int64 data type by pandas. This implied that any leading zeros, which are significant components of standard FIPS codes (e.g., '01001' for Autauga County, AL), would have been omitted during the initial CSV parsing.

To rectify this and ensure accurate geographic merging, the LocationID column was standardized. A new column, CountyFIPS, was created by first converting LocationID to a string type using the .astype(str) method. Subsequently, the .str.zfill(5) method was applied to pad these strings with leading zeros, guaranteeing a uniform 5-character length for all FIPS codes. Following this standardization, the DataFrame was filtered to retain only rows where the newly created CountyFIPS conformed to a 5-digit numeric regular expression pattern (^\d{5}$). This filtering step, resulting in the DataFrame df_cdc_places_filtered, served as a validation, confirming that all original rows contained LocationID values that could be successfully converted to standard 5-digit FIPS formats. This process identified 3,145 unique county FIPS codes within the dataset.

The subsequent step involved selecting the specific SDoH measures relevant to the SHI's predefined domains. A list of seven measure descriptions was used to filter df_cdc_places_filtered based on its Measure column. The selected measures were:

- 'Food insecurity in the past 12 months among adults'
- 'Received food stamps in the past 12 months among adults'
- 'Housing insecurity in the past 12 months among adults'
- 'Lack of reliable transportation in the past 12 months among adults'
- 'Feeling socially isolated among adults'
- 'Lack of social and emotional support among adults'
- 'Frequent mental distress among adults'
  This selection process yielded df_cdc_selected_measures, a DataFrame containing 35,306 rows, reflecting the multiple measures available per county.

To transform this "long" format data (where each measure for a county occupied a separate row) into a "wide" format more suitable for analysis, the pivot_table() function was utilized. Prior to pivoting, a subset of necessary columns (CountyFIPS, StateAbbr, Year, Measure, Data_Value) was selected into df_cdc_to_pivot. The Data_Value column, containing the actual measure percentages or rates, was converted to a numeric type using pd.to_numeric(errors='coerce'). To handle any potential duplicate entries for the same CountyFIPS and Measure (particularly if data for multiple years or minor variations existed), the data was sorted by CountyFIPS, Measure, and then Year (in descending order for Year). Subsequently, drop_duplicates(subset=['CountyFIPS', 'Measure'], keep='first') was applied, effectively retaining the data for the most recent year available for each measure within each county. The pivot operation was then performed with CountyFIPS, StateAbbr, and Year set as the index, Measure used to define the new column headers, and Data_Value populating the cells. The resulting df_cdc_county_pivot DataFrame, after resetting the index, contained 3,145 rows and 10 columns. For enhanced usability, column names derived from the Measure values were programmatically cleaned by replacing spaces and special characters with underscores and shortening common descriptive suffixes (e.g., _in_the_past_12_months_among_adults was changed to _adults_12mo).

4.2. Processing SDoH2020Data_cleaned.csv

The SDoH2020Data_cleaned.csv dataset, characterized by its extensive width (approximately 700 columns), served as a source for supplementary SDoH indicators. The preparation of this dataset began with loading the file into the df_sdoh_full DataFrame. A critical first step was the identification and standardization of its county-level geographic identifier. The COUNTYFIPS column was located (allowing for case variations in the column name) and subsequently converted to a 5-digit string format using .astype(str).str.zfill(5), ensuring consistency with the CountyFIPS column in the already processed df_cdc_county_pivot.

Given the dataset's breadth, a targeted approach to column selection was imperative to maintain focus and manageability. An initial list of 46 SDoH-related column name templates (e.g., ACS_PCT_TRANSPORT, ACS_PCT_HH_FOOD_STMP), which had been previously identified through keyword searches relevant to the SHI domains of food, housing, and transportation, was used as a starting point. To ensure that only variables with practical utility were incorporated into the analysis, a data quality check was performed on these potential columns. For each of the 46 column templates, its corresponding actual column name in df_sdoh_full was found (using a case-insensitive match). Then, each identified column was assessed for data presence by checking if it contained at least one non-missing (non-NaN) value, using the condition df_sdoh_full[actual_col_casing].notna().any(). Only those columns that satisfied this condition, along with the mandatory COUNTYFIPS column, were included in the final selection. This filtering process resulted in the df_sdoh_selected DataFrame, which comprised 3,229 rows and 47 columns (this count includes CountyFIPS and the 46 SDoH variables that successfully passed the non-empty data check).

4.3. Merging Prepared DataFrames

The final step in the data preparation phase was to consolidate the processed information from both primary sources into a single, comprehensive county-level dataset. This was achieved by merging df_cdc_county_pivot (which contained the core SDoH measures derived from CDC PLACES data) with df_sdoh_selected (which held the supplementary indicators from the

SDoH2020 dataset).

The merge operation was performed using the pandas pd.merge() function. The common key for this join was the standardized CountyFIPS column present in both DataFrames. A how='left' merge strategy was employed, designating df_cdc_county_pivot as the "left" DataFrame. This approach is crucial as it ensures that all 3,145 counties present in the CDC PLACES data (our anchor dataset for the core SDoH measures) were retained in the final merged output. If a particular county from df_cdc_county_pivot did not have a corresponding CountyFIPS entry in df_sdoh_selected, the columns originating from df_sdoh_selected for that county would be populated with NaN (Not a Number) values. The resulting DataFrame, named df_merged_county, contained 3,145 rows and 56 columns. A subsequent analysis of NaN values in this merged dataset revealed that 10 of the counties from the original CDC PLACES data did not find a corresponding match in the SDoH2020Data_cleaned.csv data, leading to NaNs in the columns derived from the latter for these specific counties. This df_merged_county served as the complete county-level dataset for the subsequent SHI construction.

5. Social Health Index (SHI) Construction (County Level)

The construction of the Social Health Index (SHI) from the integrated df_merged_county DataFrame involved a multi-stage analytical process, designed to be transparent and methodologically sound, transforming raw data points into meaningful composite indicators.

A critical initial step within this phase was **handling missing values (NaNs)**. The df_merged_county contained NaNs that could have arisen from several sources, including non-matching FIPS codes during the merge operation, inherent missing data in the original source files, or as a result of coercing non-numeric entries to numeric during earlier data type conversions. For the specific variables selected to constitute the SHI sub-indices, median imputation was chosen as the primary strategy. This technique is generally preferred over mean imputation for SDoH data because the median is a more robust measure of central tendency, less susceptible to the influence of outliers and skewed distributions, which are common characteristics of socio-economic and health indicators. The implementation involved first ensuring all SHI constituent variables were converted to a numeric type using pd.to_numeric(errors='coerce'). Then, for each of these numeric columns, if any NaN values were present, they were filled with that column's calculated median using the fillna() method (e.g., df_merged_county[col].fillna(df_merged_county[col].median(), inplace=True)). A fallback mechanism was implemented: if a column's median was itself NaN (a rare scenario, possibly if a column became entirely NaNs after coercion), any remaining NaNs in that column were filled with 0, accompanied by a printed warning, to allow subsequent calculations to proceed without error.

The SHI was architected around **four domain-specific sub-indices**, reflecting key areas of social health: Food Insecurity, Housing Insecurity, Transportation Barriers, and Social Isolation. Specific variables from df_merged_county, originating from both CDCPlaces.csv and SDoH2020Data_cleaned.csv, were carefully assigned to these sub-indices. The variable lists for each were as follows:

- **Food Insecurity Index Variables:** Food_insecurity__adults_12mo, Received_food_stamps__adults_12mo, ACS_PCT_HH_1FAM_FOOD_STMP, ACS_PCT_HH_FOOD_STMP, ACS_PCT_HH_FOOD_STMP_BLW_POV, ACS_PCT_HH_NO_FD_STMP_BLW_POV.
- **Housing Insecurity Index Variables:** Housing_insecurity__adults_12mo, ACS_PCT_RENTER_HU_COST_30PCT, ACS_PCT_RENTER_HU_COST_50PCT, ACS_PCT_OWNER_HU_COST_30PCT, ACS_PCT_OWNER_HU_COST_50PCT, ACS_PCT_VACANT_HU, ACS_PCT_HU_NO_FUEL, ACS_PCT_HU_PLUMBING, ACS_PCT_HU_KITCHEN.
- **Transportation Barriers Index Variables:** Lack_of_reliable_transportation__adults_12mo, ACS_PCT_HU_NO_VEH, ACS_PCT_WORK_NO_CAR, CDCW_TRANSPORT_DTH_RATE, ACS_PCT_PUBL_TRANSIT.
- **Social Isolation Index Variables:** Feeling_socially_isolated_among_adults, Lack_of_social_and_emotional_support_among_adults, Frequent_mental_distress_among_adults, ACS_PCT_CHILDREN_GRANDPARENT.

A crucial step in constructing these sub-indices was the **normalization of their constituent variables**. SDoH indicators often originate with disparate units and scales (e.g., percentages ranging from 0-100, rates per 100,000 population, or raw counts). To ensure that each variable contributed equitably to its respective sub-index score, and to prevent variables with intrinsically larger absolute values from disproportionately influencing the index, Min-Max scaling was applied. The MinMaxScaler from the sklearn.preprocessing library was utilized to transform the values of all variables within each sub-index group to a common [0, 1] range. For this prototype, a critical assumption was made: for all selected variables, a higher value inherently indicates a higher level of risk or a more adverse SDoH condition (e.g., a higher percentage of food insecurity is considered worse). If variables with an inverse relationship were to be included in future iterations (where lower values indicate worse conditions), they would typically require transformation (e.g., 1 - normalized_value) before being averaged into an index.

Once the relevant variables were numerically processed and normalized, each of the **four sub-index scores** for a given county was calculated. This was achieved by taking the simple arithmetic mean of the normalized (0-1 scaled) values of its assigned constituent variables. This equal-weighting approach was chosen for its simplicity, transparency, and ease of interpretation in this prototype stage. Finally, an **Overall_SHI** was computed for each county by averaging its four sub-index scores (Food_Insecurity_Index, Housing_Insecurity_Index, Transportation_Barriers_Index, and Social_Isolation_Index). This provided a single, composite measure intended to reflect aggregate social health challenges. These five new index columns were then added to the df_merged_county DataFrame, making it the primary repository of county-level SDoH insights.

6. Mapping County-Level SHI to ZIP Code Level

The final stage in the development of the SHI predictor was to translate the county-level SHI scores to the ZIP code level, thereby fulfilling a key project objective for a more localized assessment tool. This required a reliable method to associate ZIP codes with their corresponding counties.

The **crosswalk data facilitating this mapping was acquired programmatically** from the U.S. Department of Housing and Urban Development (HUD) User Portal API, specifically its USPS Validation Service. A GET request was constructed and sent to the API endpoint https://www.huduser.gov/hudapi/public/usps, utilizing parameters type=2 (which specifies a ZIP-to-County mapping) and query=all (an attempt to retrieve a comprehensive national dataset). This API call, which necessitated an authenticated token for access, successfully returned a dataset of 54,553 records. The JSON response was parsed into a pandas DataFrame named df_hud_crosswalk_raw. The key fields extracted from this response for the purpose of this project included zip (the 5-digit ZIP code), geoid (which, for type=2 API responses, directly represents the 5-digit County FIPS code), and res_ratio (the proportion of residential addresses within that ZIP code that are estimated to fall into the specified county).

This raw crosswalk data then underwent **processing and standardization** to prepare it for merging. The zip column was converted to a 5-digit string format and renamed to ZIP for clarity and consistency. Similarly, the geoid column, identified as the County FIPS code, was standardized to a 5-digit string and named CountyFIPS. A critical consideration in ZIP-to-county mapping is that a single ZIP code's geographic area can, in some instances, span multiple counties or parts of counties. The res_ratio field from the HUD data provides a quantitative basis to address this ambiguity by indicating the proportion of a ZIP code's residential units within a specific county. To assign a single, primary county to each ZIP code, the crosswalk data was sorted first by ZIP (ascending) and then by res_ratio (descending). Following this sort, the drop_duplicates(subset=['ZIP'], keep='first') method was applied. This operation effectively retained, for each unique ZIP code, only the county record that accounted for the highest proportion of its residential addresses, a standard and robust approach for disambiguating such many-to-many or one-to-many potential mappings. This processing yielded a cleaned DataFrame, df_zip_to_county, which contained 39,490 unique ZIP-to-County FIPS mappings.

The **final step in creating the ZIP-level SHI** was to merge this df_zip_to_county DataFrame with the df_merged_county DataFrame (the latter holding all county-level data including the calculated SHI sub-indices and the Overall_SHI). This merge was performed using CountyFIPS as the common join key, employing a how='left' strategy. This ensured that all 39,490 ZIP codes from the processed crosswalk were retained in the final output DataFrame, df_zip_shi. Consequently, each row in df_zip_shi represents a unique ZIP code, now augmented with the Overall_SHI and the four sub-indices corresponding to its primary associated county. The final df_zip_shi contained 62 columns. An analysis of NaN values within this ZIP-level DataFrame indicated that approximately 188 ZIP codes had null SHI scores. This was an expected outcome, attributable to their primary associated counties not being present in the df_merged_county data, which in turn was likely due to those counties either lacking initial data in the CDCPlaces.csv source or not finding a match during the earlier merge with the SDoH2020Data_cleaned.csv. This df_zip_shi DataFrame constitutes the core output of the SHI predictor prototype at the ZIP code level.

7. Results Exploration and Contextualization (County and ZIP Level)

Following the construction of the SHI and its mapping to ZIP codes, an initial exploration and contextualization of the results were performed to understand the characteristics of the generated indices and to provide methods for their interpretation.

At the **county level**, using the df_merged_county DataFrame, a conceptual validation of the Overall_SHI was attempted. The initial plan was to correlate the Overall_SHI with external socio-economic indicators not directly used in its construction, such as ACS_PCT_LT_HS (percentage of population with less than a high school diploma), ACS_PER_CAPITA_INC (per capita income), and ACS_PCT_UNEMPLOY (percentage unemployed), which were expected to be present in the SDoH2020Data_cleaned.csv. However, this specific external validation could not be fully executed because these particular ACS column templates were not found among the 47 columns that were ultimately selected from SDoH2020Data_cleaned.csv for inclusion in df_merged_county. The earlier keyword-based column selection for df_sdoh_selected (the subset of SDoH2020Data_cleaned.csv) had focused more narrowly on direct indicators of food, housing, and transportation needs. Consequently, these broader socio-economic metrics, which would have been valuable for such validation, were inadvertently omitted. This observation highlights an area for refinement in future iterations of this project: ensuring the inclusion of a wider array of pertinent socio-economic variables if they are available in the source data and deemed relevant for validation purposes. Despite this limitation, an **internal consistency check** of the Overall_SHI was performed and proved successful. The Overall_SHI demonstrated strong, positive Pearson correlations with its four constituent sub-indices: Food Insecurity Index (0.94), Social Isolation Index (0.90), Transportation Barriers Index (0.81), and Housing Insecurity Index (0.71). This alignment indicates that the sub-indices are cohesively and meaningfully contributing to the composite SHI, which is an expected and desirable characteristic of a well-constructed aggregate index.

At the **ZIP code level**, using the df_zip_shi DataFrame, several methods were employed to contextualize the SHI scores and make them more interpretable:

- **Descriptive Statistics:** The .describe() method was applied to the Overall_SHI and each sub-index column within df_zip_shi (after appropriate handling of NaNs for these index columns). This provided a statistical summary including count, mean, standard deviation, minimum, maximum, and quartile values (25th, 50th/median, 75th percentiles). These statistics offer crucial insights into the distribution and range of SDoH challenges across the approximately 39,000 ZIP codes for which scores were available. Given the Min-Max normalization applied to their components, all index scores are inherently scaled to fall primarily within a 0 to 1 range, where higher values consistently signify greater SDoH challenges.

- **Identification of High-Risk ZIP Codes:** To pinpoint areas of potentially high need, df_zip_shi was sorted in descending order based on the Overall_SHI. The top 10 ZIP codes exhibiting the most significant aggregate SDoH challenges were then identified and displayed, along with their respective sub-index scores. This provides a direct method for prioritizing areas for further investigation or intervention.
- **Percentile Ranking:** To understand the relative standing of any specific ZIP code within the dataset, its percentile rank for the Overall_SHI was calculated. This was demonstrated by taking a sample ZIP code from the high-SHI list (e.g., '99632', with an Overall_SHI of 0.6571) which was found to be in approximately the 99.97th percentile. This means it faces more severe SDoH challenges than 99.97% of other ZIP codes for which scores were available in the dataset, providing a clear comparative measure.
- **Categorization of SHI Scores:** To offer a more qualitative interpretation of the numerical index values, ZIP codes were categorized into four levels of concern: 'Low Concern', 'Medium-Low Concern', 'Medium-High Concern', and 'High Concern'. This categorization was achieved using the pd.qcut() function, which divides the Overall_SHI scores into quartiles (four equally sized groups based on rank). The distribution of ZIP codes across these categories was then examined (e.g., ZIP '60062' was categorized as 'Medium-High Concern').
- **Specific ZIP Code Lookup Functionality:** To facilitate easy querying of the SHI predictor for any given ZIP code, an interactive lookup capability was implemented. This involved utilizing the df_zip_shi_cleaned_with_category DataFrame (which is df_zip_shi with NaNs in Overall_SHI dropped and the SHI_Category column added). A Python function was created that prompts the user to enter a 5-digit ZIP code via the input() function. Upon entry, the system retrieves and displays the Overall_SHI, all four sub-index scores, and the derived SHI_Category for the specified ZIP code, if it exists in the dataset and has valid scores. This provides an immediate and practical application for users to assess the SDoH profile of a particular ZIP code.

These exploratory and contextualization steps confirm that the SHI scores are being generated in a structured manner and can be effectively used to differentiate and characterize SDoH levels across various geographic areas, providing actionable insights.

8. Conceptual Role of LLMs in SDoH Data Integration: A Technical Perspective

The integration of diverse Social Determinants of Health (SDoH) data, as highlighted by the initial project considerations, presents a significant data science challenge. This project, while manually curating variables for the prototype, acknowledges the transformative potential of Large Language Models (LLMs) in this domain. This section provides a more detailed technical perspective on this conceptual integration, drawing insights from contemporary research (e.g., Fensore et al., "Large Language Models for Integrating Social Determinant of Health Data").

8.1 The Core Challenge: Semantic Heterogeneity and Scale

Publicly available SDoH datasets (like the 9 sources initially mentioned for this project, including ACS, AHRQ SDoH Database, CDC PLACES, etc.) exhibit vast semantic heterogeneity. Variables representing similar concepts may have different names, units, or granularities. Manually mapping these variables to standardized SDoH domains (e.g., Economic Stability, Education Access and Quality, Health Care Access and Quality, Neighborhood and Built Environment, Social and Community Context) or specific analytical constructs (like the sub-indices in this SHI) is labor-intensive, requires extensive domain expertise, and is prone to inconsistencies, especially when dealing with hundreds or thousands of potential variables (e.g., the ~700 columns in SDoH2020Data_cleaned.csv).

8.2 LLM-Powered Semantic Annotation and Harmonization

LLMs offer a powerful solution for automating and enhancing this SDoH variable annotation and harmonization process. Their capabilities extend to several key areas:

- **Automated Variable Classification:** LLMs can perform zero-shot or few-shot classification of SDoH variables into predefined ontologies or domains. This typically involves providing the LLM with the variable name, its textual description (from a data dictionary, if available), and potentially sample data values. The LLM then predicts the most appropriate SDoH domain(s). The effectiveness of this classification relies on carefully engineered prompts. For instance, a prompt might be structured as: "Given the SDoH variable named '[VARIABLE_NAME]' with the description '[VARIABLE_DESCRIPTION]', classify it into one of the following five domains: [Domain1, Domain2, Domain3, Domain4, Domain5]. Provide the domain and a confidence score." For this project, an LLM could have been tasked to classify each of the ~700 columns in SDoH2020Data_cleaned.csv. A variable like ACS_PCT_HH_NO_FD_STMP_BLW_POV would ideally be mapped by an LLM to "Economic Stability" and "Food Insecurity."
- **Enhanced Semantic Understanding:** LLMs are not limited to simple keyword matching. Their training on vast text corpora allows them to discern semantic relationships and contextual nuances. For example, an LLM might correctly associate a variable related to "access to parks" with "Neighborhood and Built Environment" and potentially "Health Behaviors" (due to implications for physical activity), even if these exact terms are not present in the variable name or its immediate description.
- **Data Dictionary Generation/Enrichment:** In scenarios where datasets suffer from poor or missing metadata, LLMs could potentially assist in generating or enriching variable descriptions. This could be based on analyzing variable names, observed data patterns, and relationships with other, better-described variables, although this represents a more advanced application requiring careful validation.

8.3 Accuracy, Efficiency, and Scalability Gains

The application of LLMs to SDoH data integration promises significant improvements in several dimensions:

- **Benchmarking Accuracy:** The accuracy of LLM-based annotation can be rigorously benchmarked against manual annotations performed by domain experts or by comparing against established SDoH frameworks. Standard metrics such as precision, recall, F1-score, and inter-rater reliability (e.g., Cohen's Kappa) can be employed for this evaluation. Published research indicates that larger, instruction-tuned LLMs (e.g., Llama-2 70B-chat, GPT-4) can achieve high levels of agreement with human experts, particularly when provided with rich and descriptive variable metadata.

- **Computational Efficiency:** LLMs can annotate variables at a considerably faster rate than manual human efforts. For example, some studies estimate 1-5 seconds per variable for large models, compared to potentially minutes for a human annotator performing detailed research and classification. This efficiency translates to substantial time savings, especially when dealing with large-scale data integration projects involving numerous datasets and thousands of variables.
- **Scalability and Generalizability:** An LLM-based annotation pipeline offers excellent scalability and generalizability. Once developed and validated, such a pipeline can be readily applied to new SDoH datasets or to updated versions of existing ones with minimal, if any, retraining. This is crucial as the landscape of available SDoH data is continuously evolving and expanding. For instance, if additional data sources were to be incorporated into this SHI predictor, an LLM could rapidly process and categorize their variables, facilitating a much quicker expansion of the index.

8.4 Conceptual Application in This SHI Prototype

Had LLMs been directly implemented in this prototype's workflow, several stages could have been enhanced:

1. **Initial Data Exploration & Understanding (SDoH2020Data_cleaned.csv):** Instead of relying solely on manual review or simple keyword searches across the nearly 700 columns, an LLM could have provided an initial thematic breakdown of all variables. This would involve flagging columns potentially relevant to the core SHI domains of food, housing, transportation, and social context, thereby streamlining the initial screening process.
2. **Systematic Feature Selection:** Following the initial screening, the LLM could have been prompted to identify and rank the top *k* most relevant variables from SDoH2020Data_cleaned.csv for each of the four SHI sub-indices. This could potentially uncover relevant variables that might not be immediately obvious through keyword matching alone and would make the selection of supplementary variables more systematic and potentially more comprehensive. This could also have aided in identifying broader validation variables (like income or education levels) that were missed in the manual keyword approach.
3. **Harmonization Check:** If the project were to expand to include multiple diverse datasets, an LLM could assist in identifying conceptually similar but differently named variables across these sources. Recognizing such semantic equivalencies is critical for effective data harmonization, deduplication, or deciding on appropriate methods for combining related indicators.

8.5 Technical Considerations and Limitations for Implementation

While the potential of LLMs is significant, their practical deployment for SDoH data integration requires careful consideration of several technical aspects and inherent limitations:

- **Quality of Input Metadata:** The performance of LLMs is highly contingent on the quality and completeness of the input data, particularly variable names and their associated descriptions. Cryptic or poorly documented variable names (e.g., VARX001, X12B) will likely yield suboptimal classification results unless substantial contextual information can be inferred or provided.
- **Prompt Engineering:** The design of effective prompts is a critical and often iterative process. The structure of the prompt, the clarity of the instructions, the inclusion of relevant examples (for few-shot learning scenarios), and the specification of the desired output format all significantly influence the LLM's performance and the utility of its responses.
- **Computational Resources and Cost:** Accessing and utilizing powerful LLMs can have considerable resource implications. Commercial models often involve API usage costs that scale with the volume of data processed. Hosting large open-source models locally requires significant computational infrastructure, including high-end GPUs and substantial memory. Latency for processing large batches of variables also needs to be factored into workflow design.
- **Potential for Bias:** LLMs are trained on vast quantities of text data, much of which is sourced from the internet. As such, they can inadvertently learn and perpetuate societal biases present in their training data. This is a particularly critical concern in SDoH research, where biased interpretations or classifications of variables could lead to unfair or inequitable outcomes for vulnerable populations. All LLM outputs must be carefully scrutinized for potential biases.
- **Interpretability and Explainability:** The decision-making processes of complex LLMs can be opaque, often referred to as the "black box" problem. Understanding *why* an LLM classified a particular variable in a certain way can be challenging, though techniques for improving model explainability (e.g., attention mechanisms, feature attribution) are an active area of research.
- **Validation by Domain Experts:** LLM-generated annotations, classifications, or insights should always be treated as a preliminary or assistive step. Rigorous validation, review, and refinement by human domain experts (e.g., public health researchers, epidemiologists, sociologists) are indispensable to ensure the accuracy, contextual relevance, and ethical application of the LLM's outputs.
- **Reproducibility:** The outputs of some LLMs can exhibit a degree of stochasticity, meaning that identical inputs might not always produce identical outputs across multiple runs. Ensuring reproducible results for research or operational deployment may require setting specific model parameters (e.g., temperature=0 for more deterministic output) and careful versioning of the models, prompts, and datasets used.
- **Handling of Numerical Data:** While LLMs excel at processing and understanding natural language text, their ability to directly interpret the statistical properties, distributions, or

complex interrelationships of purely numerical SDoH variables to infer their semantic meaning is still an emerging capability. This often requires specialized prompting techniques or the integration of LLMs with traditional statistical methods.

8.6 Conclusion on LLM Integration

LLMs possess a strong and demonstrable potential to significantly streamline the integration of SDoH data for clinical studies and predictive modeling projects by performing accurate, automatic, and generalizable annotation of SDoH variables. They act as powerful assistive technologies that can manage the scale and complexity of SDoH data, allowing researchers to dedicate more time to higher-order analytical tasks, interpretation, and intervention design. However, their application must be approached with a clear understanding of their current limitations and a steadfast commitment to rigorous validation by domain experts to ensure responsible, ethical, and impactful use. For a future iteration of this SHI predictor, incorporating an LLM-assisted feature engineering and variable categorization pipeline would represent a valuable and significant enhancement, particularly if the scope were to expand to include a larger number of diverse data sources.

9. Project Conclusion and Future Directions

This prototype successfully demonstrates a robust methodology for constructing a multi-dimensional Social Health Index at the U.S. county level and subsequently mapping these insights to the ZIP code level. Through systematic data preparation—including standardization of geographic identifiers, targeted variable selection, and appropriate data transformations (pivoting)—and a transparent index construction process—involving median imputation for missing values, Min-Max normalization, and equal-weight averaging for sub-indices and the overall SHI—a valuable analytical tool was developed. The integration of data from CDCPlaces.csv and a comprehensive SDoH2020Data_cleaned.csv allowed for the creation of nuanced sub-indices for Food Insecurity, Housing Insecurity, Transportation Barriers, and Social Isolation, culminating in an Overall_SHI. The final df_zip_shi DataFrame, which assigns these county-derived SHI scores to 39,490 ZIP codes based on a HUD USPS crosswalk, serves as the core output of this predictor model prototype.

Future Directions could include:

- **Enhanced Feature Selection:** Incorporating a broader range of SDoH variables from SDoH2020Data_cleaned.csv (e.g., detailed income, education, employment metrics) to enable more comprehensive external validation and potentially richer indices. This could involve LLM-assisted variable screening.
- **Advanced Index Construction:** Exploring alternative normalization techniques, statistical methods for variable weighting (e.g., Principal Component Analysis - PCA, expert-defined weights), or machine learning-based approaches to constructing the SHI, rather than simple averaging.
- **Refined NaN Imputation:** Utilizing more sophisticated imputation techniques (e.g., k-Nearest Neighbors imputation, regression imputation) for missing data, particularly for

variables with higher rates of missingness if they are deemed critical.

- **Validation Against Health Outcomes:** Rigorously validating the SHI and sub-indices by correlating them with actual health outcome data at the county or ZIP level (e.g., prevalence of specific diseases, life expectancy, hospital readmission rates, mortality rates).
- **Temporal Analysis:** If multi-year data is consistently processed for all sources, analyzing trends in SHI scores over time to identify areas with improving or worsening social health conditions.
- **Geospatial Visualization:** Implementing mapping capabilities (e.g., using libraries like GeoPandas, Folium, or Kepler.gl) to visually represent SHI scores across geographic areas, which can be highly impactful for communication and identifying hotspots.
- **Direct LLM Implementation:** Moving from conceptual discussion to practical implementation of LLMs for tasks like SDoH variable description generation from metadata, semantic mapping of variables to SDoH domains across diverse datasets, or even generating natural language summaries of SHI findings for specific geographic areas.
- **Sensitivity Analysis:** Assessing how sensitive the SHI scores are to different methodological choices (e.g., different imputation methods, normalization techniques, or variable weighting schemes).

This prototype provides a solid foundation for further development and application in understanding and addressing the complex interplay of social determinants on community health.