Scalable Data Engineering Pipeline for Local Health Insights Using CDC PLACES (2020–2023)

Dr. Reihaneh Samsami, Sapna Baniya, Suraj Thapa
M.S. in Data Science
University of New Haven

December 7, 2025

Abstract

This report presents a distributed and scalable data engineering pipeline for analyzing county-level health indicators using the CDC PLACES dataset for the years 2020–2023. The system is implemented on Amazon Web Services (AWS) using an S3-based data lake, AWS Glue for large-scale ETL, Amazon Athena for interactive SQL analytics, and Python-based tooling for visualization and machine learning. The pipeline ingests raw multi-year CSV files, standardizes and optimizes them into Parquet format, and exposes a curated analytics layer that supports both descriptive statistics and a K-Means clustering model for grouping U.S. counties into health risk profiles. The proposed architecture is modular, scalable, and reproducible, and can be extended to future PLACES releases with minimal changes.

1 Introduction

Population health analytics increasingly depends on the ability to process large-scale, multi-year, and multi-region datasets. The CDC PLACES "Local Data for Better Health" county data product provides model-based estimates of key health indicators—such as obesity, diabetes, frequent mental distress, and short sleep duration—for all U.S. counties and the District of Columbia. While the dataset is rich and comprehensive, it is provided as large flat CSV files, and repeated yearly releases quickly accumulate to hundreds of megabytes. Extracting meaningful insights from such data requires scalable systems capable of automated ingestion, cleaning, transformation, storage, and analysis.

In modern data ecosystems, **data engineering pipelines** serve as the backbone for analytical and machine learning workflows. Organizations across every industry increasingly rely on automated, distributed pipelines to continuously collect data, standardize formats, enforce quality, optimize storage, and expose curated datasets to analysts and applications. A well-designed pipeline ensures that data is *reliable, reproducible, scalable, and query-efficient*, enabling downstream teams to focus on insights rather than manual data wrangling. As datasets grow in size and heterogeneity, the ability to build such pipelines becomes essential for ensuring that analytics and machine learning remain accurate and cost-effective.

The importance of data engineering is amplified in public health, where policy decisions must be informed by up-to-date, geographically detailed metrics. A scalable cloud-native pipeline allows multi-year health data to be processed consistently, compared across regions, and prepared for models that can identify emerging risks or clustered health patterns. Such infrastructure is not only valuable for academic research but also mirrors the real-world systems used in healthcare technology companies, government agencies, insurance organizations, and public health institutions.

From a career perspective, this project offers hands-on experience with the tools and patterns most demanded in today's data industry, including cloud data lakes, distributed ETL (Extract–Transform–Load), schema design, Parquet optimization, SQL-based analytics, and integration with machine learning workflows. Building an end-to-end pipeline demonstrates mastery of both engineering and analytical thinking—a combination necessary for roles such as Data Engineer, Machine Learning Engineer, Analytics Engineer, and Cloud Engineer. Beyond technical skills, the project highlights the ability to design systems that scale, automate repetitive data processes, and support actionable insights in real-world domains such as population health.

The goal of this project is therefore to design and implement a complete, cloud-native data engineering pipeline that can ingest, clean, integrate, and analyze PLACES county-level data for multiple years (2020–2023). The system must support scalable processing, efficient querying, and downstream machine learning models, while remaining simple enough to be reproducible in an educational setting.

2 Data and Problem Description

The CDC PLACES county dataset provides model-based prevalence estimates for 36 measures across several domains:

- Health outcomes (e.g., obesity, diabetes, coronary heart disease, stroke),
- Health-related behaviors (e.g., physical inactivity, binge drinking, smoking),
- Health status and mental health (e.g., general health, frequent mental distress),
- Disabilities and social needs (e.g., mobility disability, housing insecurity).

Each record corresponds to a specific *county-year-measure* combination and includes:

- Temporal attributes: `Year`,
- Geographic attributes: `StateAbbr`, `StateDesc`, `LocationName`, `LocationID`,
- Measure metadata: `Category`, `CategoryID`, `Measure`, `MeasureId`, `Short_Question_Text`,
- Numerical fields: `Data_Value` (prevalence), `Low_Confidence_Limit`, `High_Confidence_Limit`, `TotalPopulation`,
- Geospatial fields: `Geolocation` (POINT(longitude latitude)).

After merging the yearly files (2020–2023) and removing unused columns (empty footnotes, symbols), the dataset contains approximately 850,000 rows. The central analytical questions are:

1. How do key health indicators (e.g., obesity, diabetes, mental distress, short sleep) vary over time and across states?
2. Can we group U.S. counties into meaningful health profiles using clustering based on multiple indicators?

3 System Architecture

3.1 High-Level Architecture

The proposed system follows a data lake pattern with three logical layers stored in Amazon S3:

- **Raw zone:** Original CDC PLACES CSV files, organized by year.
- **Processed zone:** Intermediate cleaned data with standardized schema, still relatively close to the source structure.
- **Curated zone:** Analytics-ready Parquet tables optimized for querying and modeling.

AWS Glue jobs (Spark) perform batch ETL from raw to processed to curated. Amazon Athena uses the AWS Glue Data Catalog to query curated data via SQL. For modeling and visualization, result subsets are exported (via Athena `UNLOAD`) back to S3 and consumed by local Python notebooks (pandas, matplotlib, scikit-learn).

3.2 Architecture Diagram

Figure 1 illustrates the end-to-end architecture of the pipeline.

3.3 Logical Data Flow

The pipeline proceeds through the following stages:

1. **Ingestion:** Raw PLACES CSV files for 2020–2023 are downloaded from the CDC data portal and uploaded to the S3 raw zone using a consistent folder layout, e.g.,

   ```
   s3://better-health-places/raw/places/year=2021/...
   s3://better-health-places/raw/places/year=2022/...
   ```

2. **Processing (Glue ETL):** An AWS Glue job written in PySpark:
   - Reads the raw CSV files with an explicit schema,
   - Removes empty or unused columns (e.g., footnote symbols),
   - Cleans numerical fields such as `TotalPopulation` by stripping commas and casting to `BIGINT`,
   - Parses the `Geolocation` field of the form `POINT(lon lat)` into separate `longitude` and `latitude` columns,
   - Standardizes column names to a consistent, lower-case format where appropriate.

   The cleaned intermediate outputs are stored in the S3 processed zone.

3. **Curation:** In the same Glue job, a fact-like table is created:
   - Columns include: `year`, `stateabbr`, `statedesc`, `locationname`, `category`, `measureid`, `short_question_text`, `data_value`, `low_confidence_limit`, `high_confidence_limit`, `totalpopulation`, `longitude`, `latitude`.
   - The table is written out as columnar Parquet files to the curated zone:

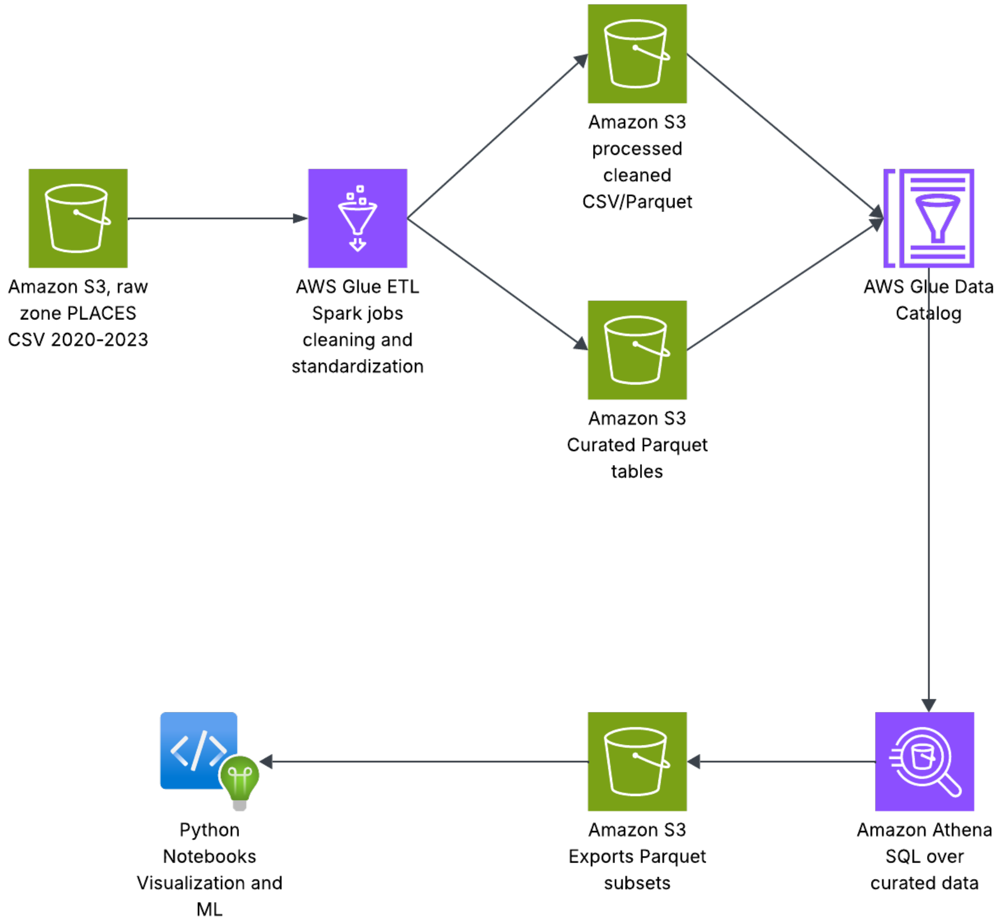
Amazon S3, raw zone PLACES CSV 2020-2023
AWS Glue ETL Spark jobs cleaning and standardization
Amazon S3 processed cleaned CSV/Parquet
Amazon S3 Curated Parquet tables
AWS Glue Data Catalog
Amazon Athena SQL over curated data
Amazon S3 Exports Parquet subsets
Python Notebooks Visualization and ML

Figure 1: High-level architecture of the CDC PLACES data engineering pipeline on AWS.

```
s3://better-health-places/curated/fact_health_nopart/
```

4. **Cataloging and Querying:** An external table `better_health_curated.curated_fact_health` is defined in Athena over the curated Parquet location via the Glue Data Catalog. Interactive queries (e.g., state-level summaries, year-on-year changes, measure filtering) are executed using standard SQL.

5. **Export for Analytics:** For heavier analytical tasks and visualization, selected Athena query results are exported using `UNLOAD` to an S3 exports prefix in Parquet format. These exports are downloaded locally and loaded into Python using `pandas.read_parquet`.

4 Methodology (CRISP–DM)

The project follows the CRISP–DM methodology, adapted to a data engineering and analytics setting.

4.1 Business Understanding

The business goal is to support public health analysis by:

- Monitoring changes in health indicators (e.g., obesity, diabetes, mental distress, sleep) over time,
- Comparing states and counties,
- Identifying clusters of counties with similar health risk profiles to guide targeted interventions.

4.2 Data Understanding

After ingestion and initial exploration, the following characteristics were noted:

- Multi-year coverage (2020–2023) with roughly similar schema,
- Multiple measures per county-year, identified by `MeasureId` and `Short_Question_Text`,
- Modeled prevalence values (`Data_Value`) with confidence intervals and population estimates,
- Some measures are reported biannually (e.g., certain screening behaviors).

4.3 Data Preparation

Within Glue ETL and Athena, the main preparation steps are:

- Type casting for `year`, `totalpopulation`, and numeric health measures,
- Trimming and standardizing state codes and location names,
- Extracting and separating latitude and longitude from the `Geolocation` field,
- Filtering to a subset of key measures for modeling: obesity (`OBESITY`), diagnosed diabetes (`DIABETES`), frequent mental distress (`MHLTH`), and short sleep duration (`SLEEP`).

For modeling, a feature view is built in Athena for a target year (e.g., 2023):

```
CREATE OR REPLACE VIEW better_health_curated.health_features_2023 AS
SELECT
    year,
    stateabbr,
    statedesc,
    locationname,
    locationid,
    totalpopulation,
    MAX(CASE WHEN measureid = 'OBESITY'  THEN data_value END) AS obesity_rate,
    MAX(CASE WHEN measureid = 'DIABETES' THEN data_value END) AS diabetes_rate,
    MAX(CASE WHEN measureid = 'MHLTH'    THEN data_value END) AS mental_distress_rate,
    MAX(CASE WHEN measureid = 'SLEEP'    THEN data_value END) AS short_sleep_rate
FROM better_health_curated.curated_fact_health
WHERE year = 2023
GROUP BY
    year, stateabbr, statedesc, locationname, locationid, totalpopulation;
```

4.4 Modeling

The modeling objective is to cluster counties into groups with similar health indicator profiles. In Python, the steps are:

1. Load the feature view for 2023 from S3 exports using pandas.
2. Select the four measures as features:

$$X = [\text{obesity\_rate}, \text{diabetes\_rate}, \text{mental\_distress\_rate}, \text{depression\_rate}].$$

3. Drop rows with missing feature values.
4. Standardize features using `StandardScaler` to zero mean and unit variance.
5. Train a K-Means model with $k = 4$ clusters:

$$\text{cluster} = \text{KMeans}(n\_clusters = 4).fit\_predict(X_{\text{scaled}}).$$

4.5 Evaluation

Model evaluation is primarily based on:

- **Cluster interpretability:** Each cluster's mean values for obesity, diabetes, mental distress, and short sleep are computed and compared.
- **Visualization:** Scatter plots (e.g., obesity vs. diabetes) and PCA projections colored by cluster are used to assess separation and interpret the groupings.
- **Epidemiological plausibility:** Clusters with consistently high values across indicators are interpreted as high-risk health profiles, while clusters with low prevalence across indicators are interpreted as relatively healthier profiles.

5 Results

This section presents the empirical outcomes of the end-to-end data engineering pipeline and the analytical insights generated from the curated PLACES dataset. Results include pipeline validation, descriptive analytics, geographic visualizations, and K-Means clustering of U.S. counties based on selected health indicators.

5.1 Pipeline Validation and Summary

The implemented cloud-native pipeline successfully processed all multi-year PLACES datasets (2020–2023) and produced a clean, analytics-ready fact table in Parquet format. The system satisfies all Distributed and Scalable Data Engineering requirements:

- **Data Ingestion:** Batch ingestion of multi-year CSV files into Amazon S3 (raw zone).
- **Data Storage:** A structured S3 data lake with raw, processed, and curated layers; optimized Parquet storage for efficient Athena querying.

- **Data Processing:** AWS Glue (Spark) jobs performing schema normalization, removal of comma separators in numeric fields, geolocation parsing, data type enforcement, and model-ready dataset preparation.
- **Data Consumption:** Amazon Athena for SQL analytics and Python-based visualization for geographic, temporal, and statistical analysis.
- **Model Deployment Concept:** A K-Means clustering model for county-level health segmentation, saved to S3 for reproducibility and potential deployment in scheduled Glue/EMR batch inference workflows.

The curated output dataset contains over one million records across four years, demonstrating the scalability of the distributed ETL pipeline.

5.2 Geographic Health Indicators: Connecticut and National Overview

Figures 3, 5, 7, and 9 display county-level health indicators for both Connecticut and the United States in 2023. These maps and bar charts reveal clear geographic patterns:

- Connecticut counties show moderate variation in obesity, diabetes, and short sleep prevalence, with no extreme outliers.
- Nationally, southern and Appalachian regions exhibit the highest obesity and diabetes burdens.
- Mental health distress is elevated in several rural midwestern and southern counties.
- Short sleep duration clusters strongly in the southeastern states, aligning with known disparities in sleep-related risk factors.

5.3 Year-over-Year Trends (2020–2023)

State-level year-on-year (YoY) comparisons show a consistent national trend:

- Obesity increased in most states between 2020 and 2021, coinciding with pandemic-related behavior changes.
- Diabetes rates show smaller YoY variation but remain persistently high in southern states.
- Mental distress sharply increased during the early pandemic and stabilized in 2022–2023.
- Short sleep duration shows gradual increases in most states, especially in southeastern regions.

These temporal insights illustrate how chronic disease burden shifted over the pandemic period.

5.4 K-Means Clustering of U.S. Counties

To group similar counties, a K-Means model (k = 4) was trained on standardized 2023 indicators: obesity, diabetes, short sleep, and mental distress.

Clusters can be summarized as follows:

- **Cluster 0 (Healthiest):** Lowest obesity/diabetes, relatively low short sleep and distress.
- **Cluster 1 (High-Risk):** Highest obesity and diabetes prevalence; widespread short sleep and elevated mental distress.
- **Cluster 2 (Behavioral-Risk Dominant):** Moderate metabolic disease but high short sleep and distress.
- **Cluster 3 (Intermediate):** Middle-range values across indicators, representing transitional regions.

Figures 10 and 11 illustrate cluster separation on two feature pairs.

These clusters highlight structural geographic disparities in health outcomes and risk behaviors, demonstrating the analytical value of a scalable data pipeline integrated with machine learning.

6 Discussion

The architecture demonstrates how cloud-native data engineering can transform raw, messy public health CSV files into a structured, analytics-ready platform. By leveraging S3, Glue, and Athena, the pipeline efficiently handles multi-year data, separates concerns across layers, and provides a flexible interface for both SQL-based and Python-based analyses.

From a public health perspective, clustering counties by multiple indicators uncovers meaningful profiles of health burden that may not be apparent when focusing on a single metric. These profiles can support targeted interventions, funding decisions, and further research, especially when combined with demographic and socioeconomic variables in future work.

7 Conclusion

This project delivers a complete, scalable data engineering pipeline for the CDC PLACES county dataset, covering ingestion, storage, processing, consumption, visualization, and initial machine learning. The modular architecture can easily accommodate new yearly releases and additional measures. Future extensions include integrating external covariates, implementing time-series forecasting, and deploying an interactive dashboard for health planners.

References

[1] Centers for Disease Control and Prevention (CDC). PLACES: Local Data for Better Health. https://www.cdc.gov/places.

[2] Amazon Web Services. AWS Glue, Amazon Athena, and Amazon S3 Documentation. https://docs.aws.amazon.com/.

[3] Pedregosa et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2011.

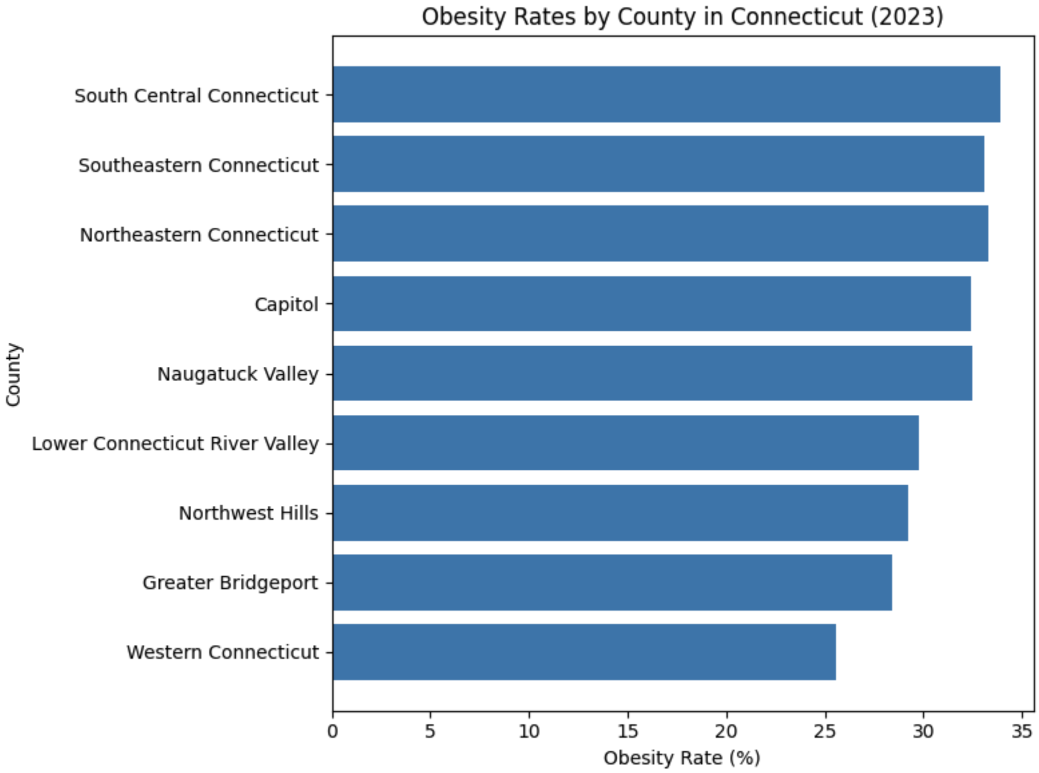

Figure 2: County-level obesity prevalence in Connecticut, 2023.

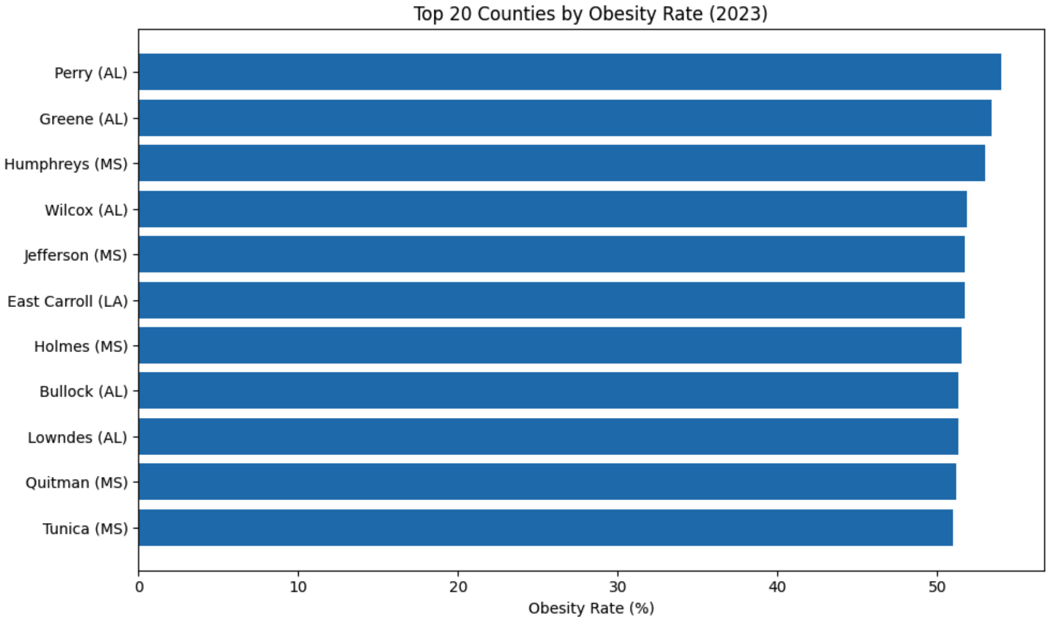

Figure 3: Top 20 U.S. counties with highest obesity prevalence, 2023.

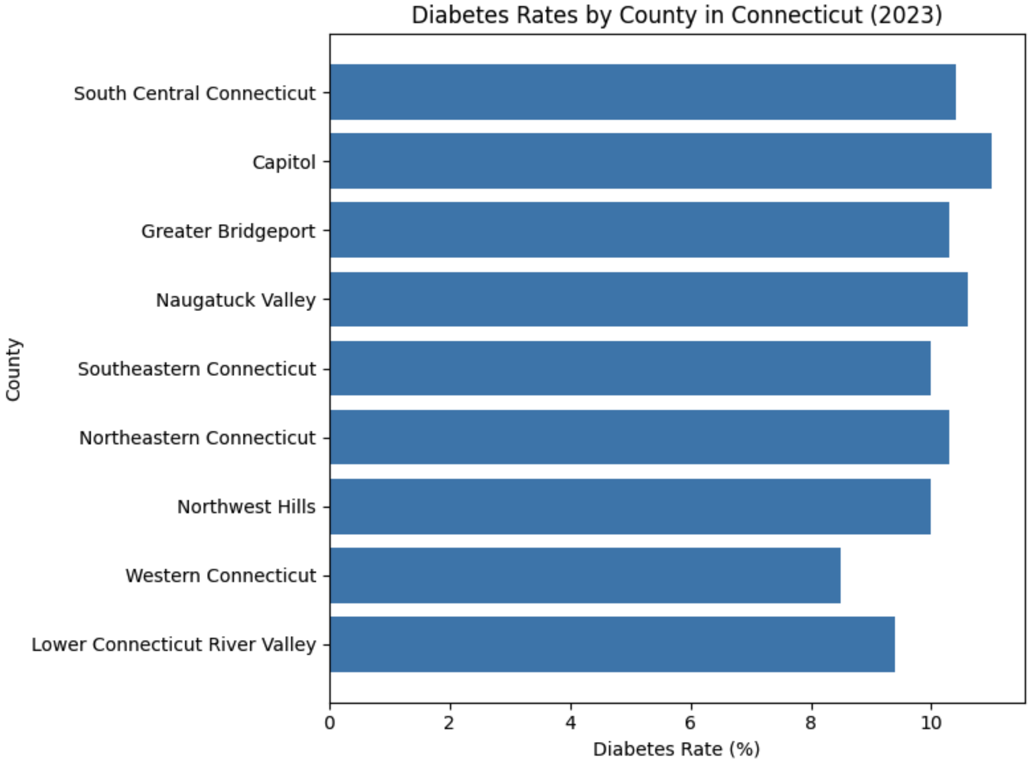

Figure 4: County-level diabetes prevalence in Connecticut, 2023.

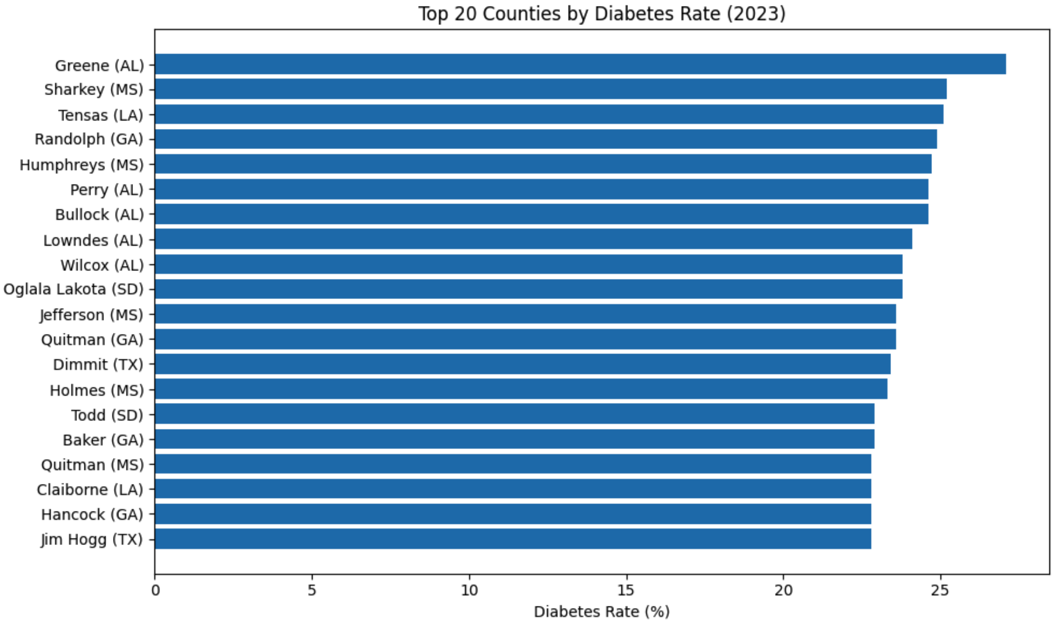

Figure 5: Top 20 U.S. counties with highest diabetes prevalence, 2023.

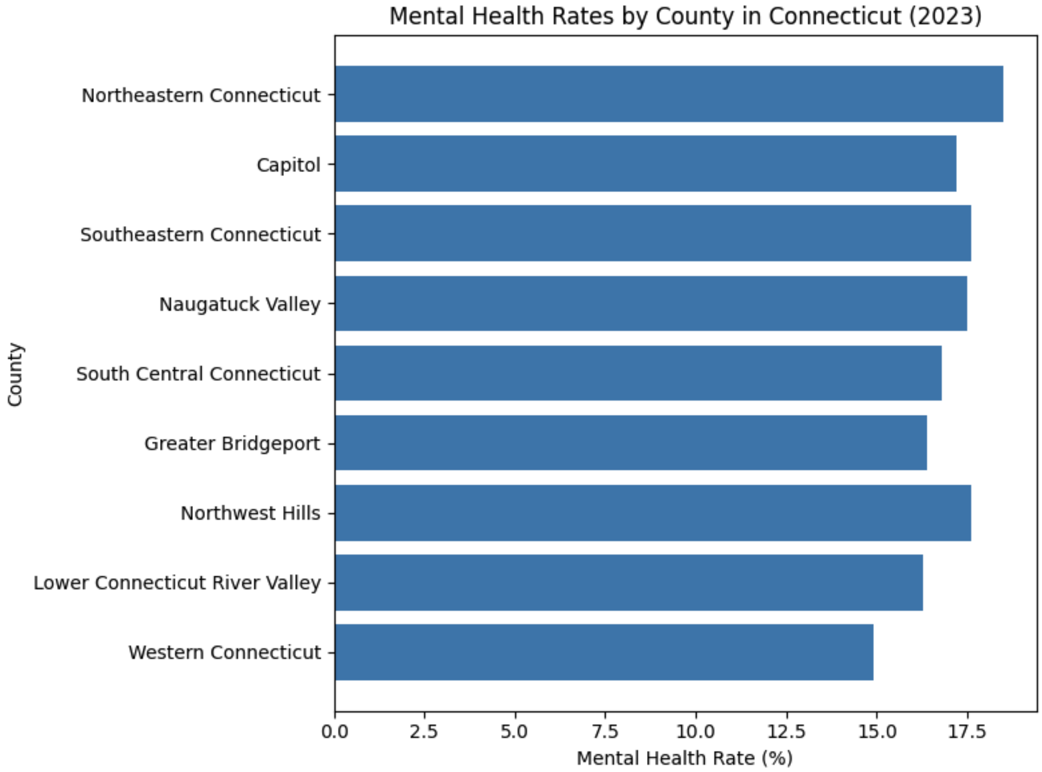

Figure 6: Frequent mental distress among adults in Connecticut, 2023.

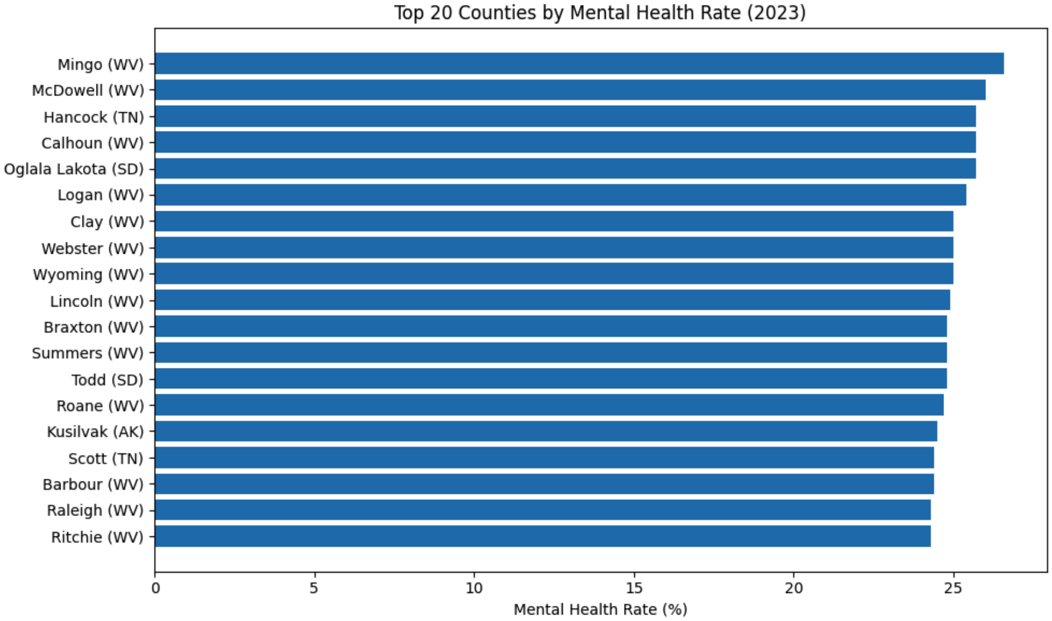

Figure 7: Top 20 U.S. counties with highest mental distress prevalence, 2023.

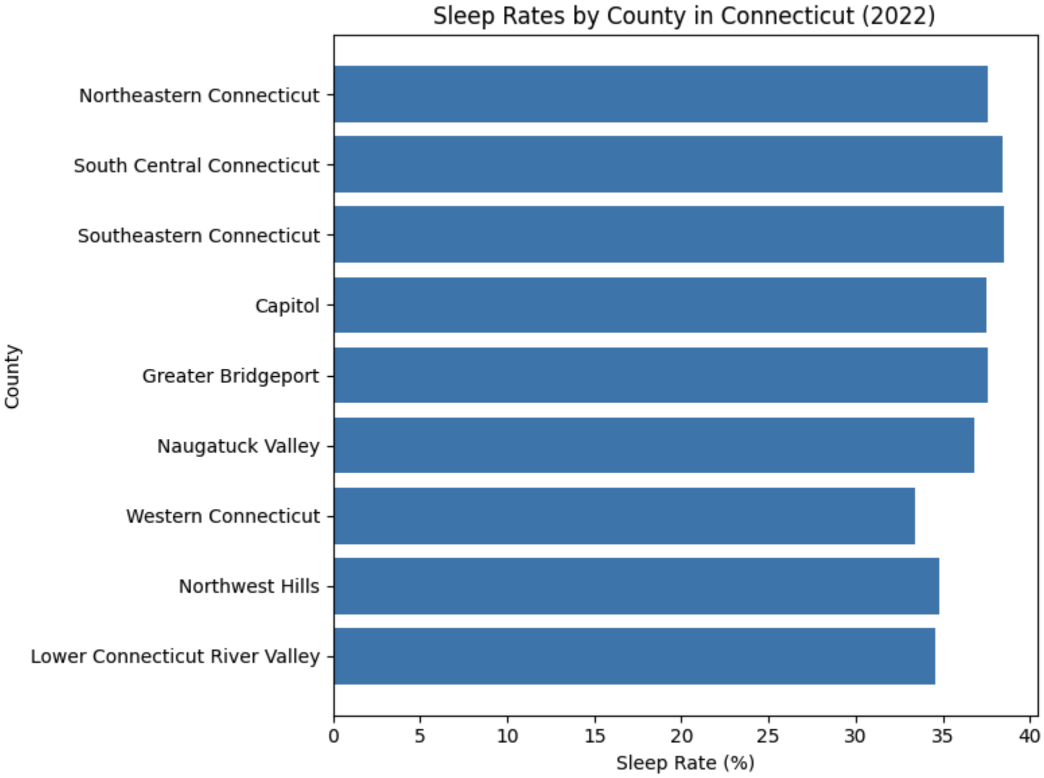

Figure 8: Short sleep duration in Connecticut, 2023.

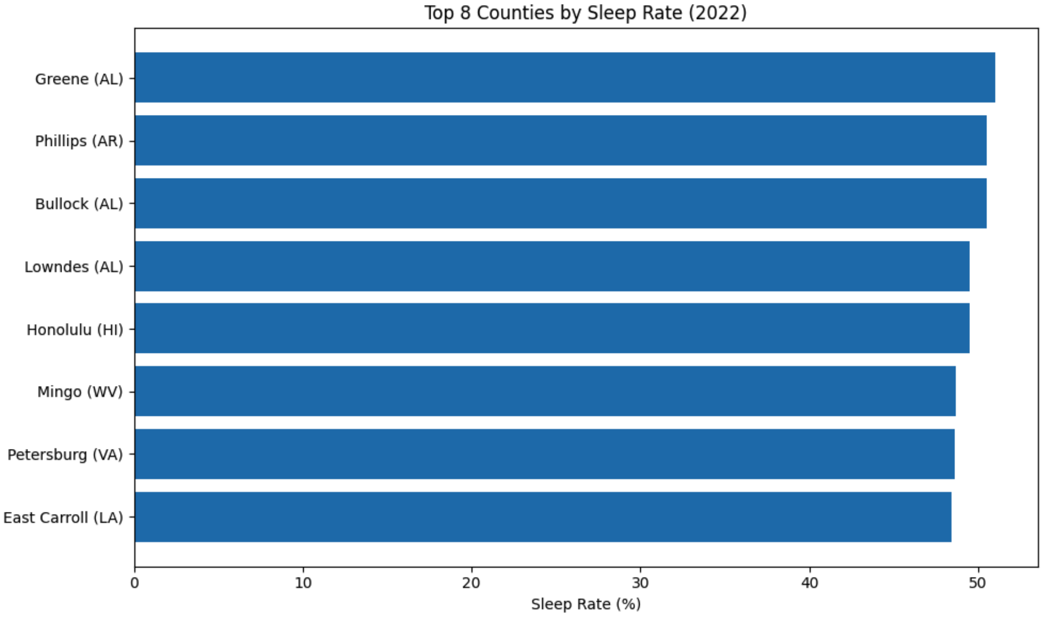

Figure 9: Top 20 U.S. counties with highest short sleep prevalence, 2023.

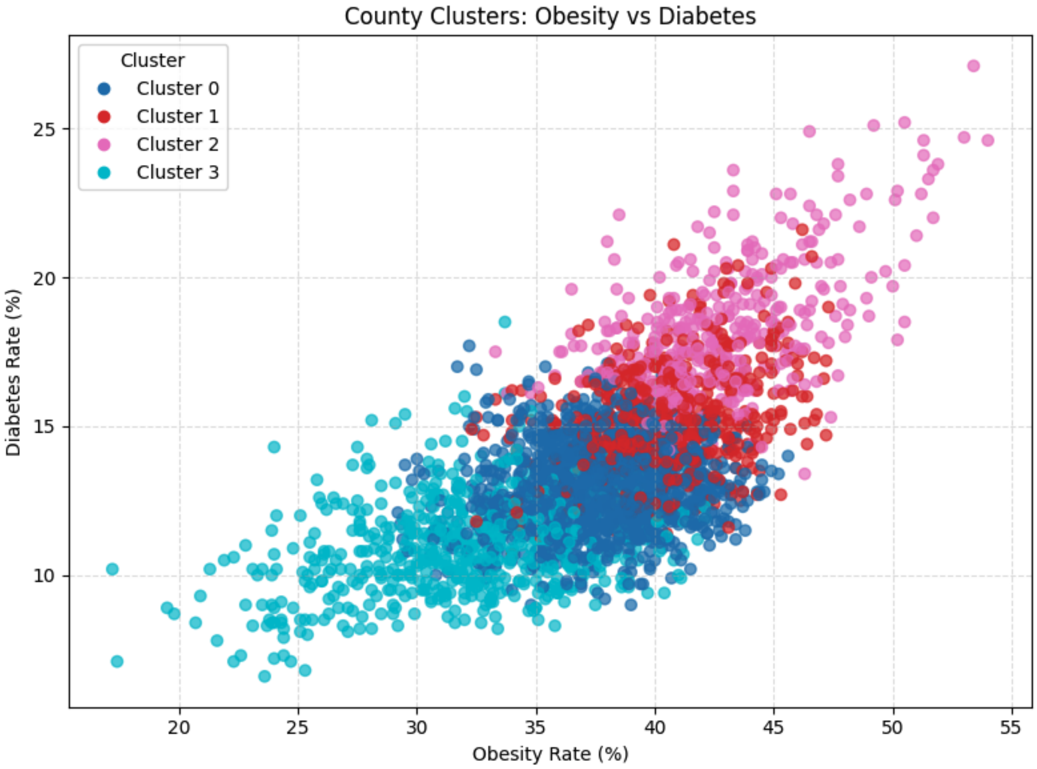

Figure 10: K-Means clustering on Obesity vs. Diabetes (2023).

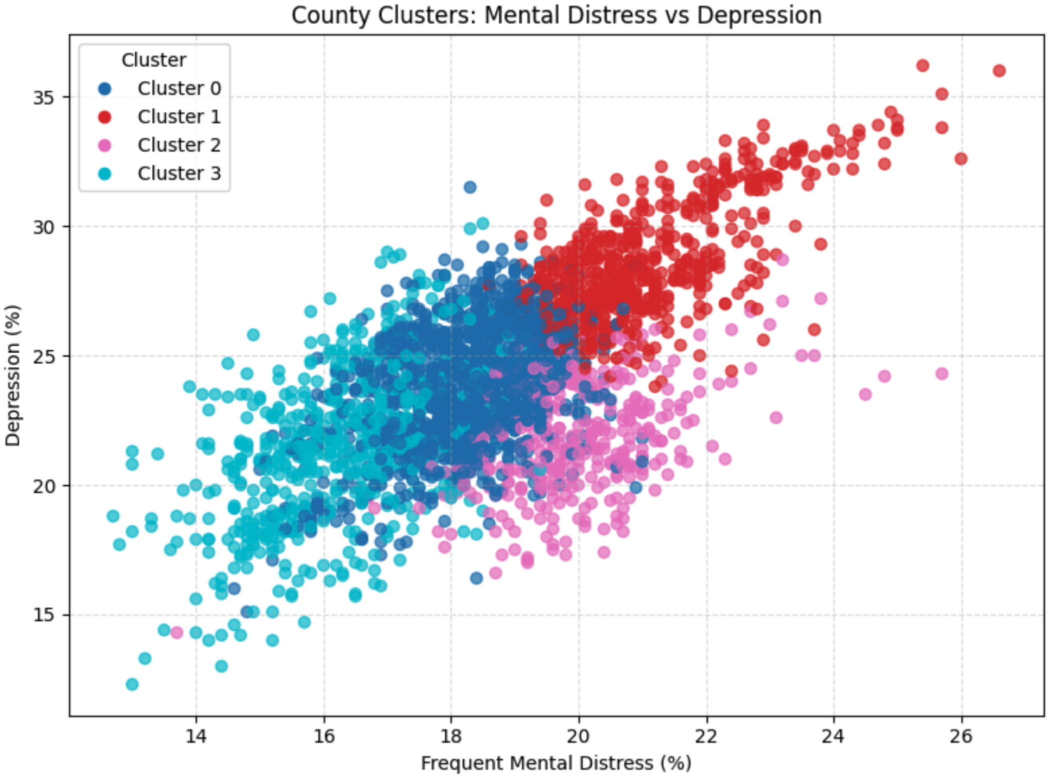

Figure 11: K-Means clustering on Mental Distress vs. Short Sleep (2023).