Mental Health Demand & Supply Analysis Across U.S. States

Using Google Search Trends and Workforce Data

Capstone Final Report

Submitted by Group 4:

Erica Zhao, Jianjun Gao, Qinbin Huang, Sandhya Karki

Instructor: Professor: Yuxiao Huang

Course: DATS 6501_10 Data Science Capstone

Submission Date: April 22, 2025

MS in Data Science

The George Washington University

Columbian College of Arts and Sciences

1. Introduction

1.1 Project Background

Mental health issues are on the rise across the United States, with significant disparities in service accessibility and provider availability across states and regions. Especially in the post-pandemic era, demand for mental health support has surged—yet supply remains uneven and often inadequate. Traditional data sources such as patient intake numbers or hospital reports often suffer from time lags and limited granularity. In contrast, digital behavioral traces, such as Google search trends, may provide near real-time signals of population-level emotional needs and help identify underserved regions.

1.2 Problem Statement

This project addresses the mismatch between growing mental health service demand and the uneven supply of mental health providers across the U.S. The goal is to identify states with potential service gaps by modeling demand signals derived from online search behavior and comparing them to official provider statistics.

1.3 Objectives

- Quantify mental health-related search demand using Google Trends data at the state level.
- Aggregate and normalize provider supply data using HRSA datasets.
- Build predictive models to uncover spatial patterns of unmet need.
- Visualize the demand–supply gap to support policy planning and public health intervention.

2. Data Sources & Preprocessing

2.1 Data Sources

We integrated three primary datasets in this project, each serving a distinct role in capturing demand, supply, and target construction:

1. Google Trends Search Data

Using the Pytrends API, we retrieved monthly search interest scores for 400+ mental health-related keywords between 2020 and 2024. These keywords span diverse themes, including symptoms (e.g., “depression test”), emotional expressions (e.g., “crying for no reason”), therapy accessibility (e.g., “affordable therapy”), and insurance-related queries (e.g., “does insurance cover depression”). The resulting dataset, us_trends_monthly_cleaned.csv, contains normalized, time-aligned trend scores at the national level.

2. HRSA Mental Health Workforce Data (AHRF)

We used data from the Health Resources and Services Administration (HRSA) Area Health Resource Files to capture the supply side, namely, the number and distribution of licensed mental health professionals at the state level. The focus metric was FTE (Full-Time Equivalent) per 100,000 population.

3. Combined Depression/Anxiety Dataset

We constructed a unified target variable, Combined_Value, representing mental health demand at the state-month level. This was derived by aggregating selected anxiety and depression-related search terms per state, adjusted and normalized for consistency.

2.2 Keyword Strategy & Collection via Pytrends

Our keyword list was curated through domain research and iterative filtering. It includes:

- Diagnostic terms (e.g., “clinical depression symptoms”, “anxiety diagnosis”)
- Financial/insurance terms (e.g., “cheap counseling”, “insurance therapy coverage”)
- Therapy modalities (e.g., “CBT”, “group therapy”, “EMDR therapy”)
- Gender/demographic variations (e.g., “depression in teenagers”)

Given the Google Trends API constraint (max 5 keywords per call), we processed the keywords in batches of 3. The data extraction script included:

- Exponential backoff and retry logic
- 20-second pauses between API calls to prevent rate-limiting
- Error tracking with logging for failed batches
- Monthly resolution ("today 5-y" → reshaped to monthly average)

2.3 Data Merging & Cleaning Logic

To align Google Trends and HRSA data, we standardized all date columns into a unified Month_Year format. The merging occurred at the **state + Month_Year** granularity. We dropped rows with missing values, removed non-informative columns (e.g., raw timestamps, duplicated IDs), and retained only clean, aligned records for modeling.

We also engineered the following:

- **Temporal features**: year, month, quarter (derived from date)
- **Categorical encoding**: one-hot encoding of state and Indicated
- **Target alignment**: ensured Combined_Value matched the temporal index of trend predictors

2.4 Feature Engineering & Dataset Summary

After processing, the final modeling dataset included:

- ~5,990 records (state-month combinations from 2020–2024)
- 100+ keyword-derived features (normalized search volume per term)
- 3 temporal features and 50+ encoded categorical variables
- Log-transformed or scaled target for XGBoost (when applicable)

All datasets were saved in .csv format and version-controlled via GitHub for transparency and reproducibility.

3. Exploratory Data Analysis (EDA)

3.1 Temporal Trends: Yearly and Monthly Patterns

To understand population-level mental health fluctuations over time, we first visualized the aggregate search interest in depression and anxiety from 2020 to 2024. The data revealed two consistent patterns:

- **Pandemic Spike (2020–2021)**: Both anxiety and depression searches surged in early 2020, peaking around mid-2021—likely in response to the COVID-19 crisis and its prolonged social impact.
- **Seasonal Variation**: Anxiety-related searches tended to spike during the fourth quarter (October to December), while depression searches showed mild increases during winter months, aligning with known patterns like Seasonal Affective Disorder (SAD).

Monthly averages were plotted across all states using smoothed line graphs, enabling the identification of cyclic mental health patterns and abnormal peaks.

3.2 State-Level Comparisons

We computed the average Combined_Value per state, reflecting normalized mental health-related search intensity. States were ranked based on their overall scores:

- **Top 5 States (Highest Demand)**: California, New York, Texas, Florida, and Illinois consistently showed the highest average values, possibly due to population density, service shortages, or greater public awareness.
- **Bottom 5 States (Lowest Demand)**: North Dakota, South Dakota, Wyoming, Vermont, and Alaska had significantly lower scores—potentially reflecting either lower need or underreporting.

These comparisons were visualized using bar plots, which clearly depicted the inter-state disparities and emphasized the value of localized policy responses.

3.3 Regional Differences

To explore broader geographic patterns, we assigned each state to one of the four U.S. Census Bureau regions: Northeast, Midwest, South, and West.

- **Southern states** generally exhibited higher mental health-related search intensity.
- **Western states** showed more variability across months, possibly due to demographic diversity or resource differences.
- **Northeastern states** had more consistent scores, while **Midwestern states** showed lower average demand.

These regional differences were visualized via violin plots and regional overlays on choropleth maps, helping identify clusters of high or low psychological burden.

3.4 Keyword-Level Trend Observations (U.S. Aggregate)

We also examined keyword volatility across time using the nationwide dataset us_trends_monthly_cleaned.csv. High-variance keywords included:

- "seasonal depression"
- "am I depressed"
- "always tired"
- "crying for no reason"
- "chronic depression"

These keywords not only exhibited sharp peaks but also aligned with real-world seasonal and societal events. Density plots and box plots were used to assess their distribution and outlier behavior, helping guide feature selection for modeling.

4. Linear Regression Modeling – Part I

4.1 Modeling Pipeline & Feature Handling

We began our predictive modeling with multivariate linear regression as a baseline. Before modeling, we performed several data preparation steps:

- **Target Variable**: The outcome Combined_Value represents aggregated depression/anxiety search intensity at the state-month level.
- **Feature Set**: Predictors included Year, Month, state dummies, Indicated (Depression/Anxiety), and derived temporal features like Quarter.
- **Preprocessing**:
    - One-hot encoding was applied to categorical features such as State and Indicated.
    - Temporal features were normalized using StandardScaler.
    - Multicollinearity was evaluated via VIF (Variance Inflation Factor) and handled via feature selection when needed.

This setup allowed us to benchmark basic linear performance before moving into more complex modeling (e.g., XGBoost).

4.2 Depression vs Anxiety Submodels

To better understand condition-specific drivers, we split the dataset into two subsets: one containing depression-related rows, and another for anxiety.

For each:

- We built independent OLS (Ordinary Least Squares) models using statsmodels.
- Key predictors included state-level dummies, month, and year.
- Depression models showed lower variance and fewer significant predictors, whereas anxiety models demonstrated stronger seasonal and regional trends.

The coefficient tables helped reveal which states or months contributed most significantly to the demand variation per condition.

4.3 Regional Submodel Decomposition

To investigate spatial variability, we created submodels for each U.S. Census region (Northeast, Midwest, South, West). Each submodel used the same regression setup but filtered the data by region.

- Southern states had larger coefficients, indicating stronger fluctuation in demand.
- Western region models had more volatility in both predictors and residuals.
- Midwestern states exhibited weaker trends overall, suggesting either lower signal or more uniform demand.

This decomposition was critical to verify whether regional disparities were persistent across time and conditions.

4.4 Evaluation & Findings

Model performance was assessed using traditional metrics:

- **RMSE** values ranged between 7.9 and 9.8 depending on subset
- **R^2** values were modest (~0.32–0.49), indicating partial explanatory power

Key takeaways:

- Linear models were able to capture general trends and seasonality, especially in anxiety-related cases.
- Depression submodels had lower explanatory strength, possibly due to greater noise or subtler signals.
- Regional models exposed meaningful differences, justifying more granular policy or clinical interventions in certain states.

These findings laid the groundwork for the subsequent use of non-linear models like XGBoost.

5. XGBoost Regression Modeling – Part II

5.1 Data Preparation & Temporal Split

We reused the cleaned dataset created during the linear regression phase. To preserve time structure and prevent leakage, we applied a time-based train/test split:

- **Train Set**: First 80% of rows (chronologically sorted by Month_Year)
- **Test Set**: Last 20% of rows
- **Target Variable**: Either raw Combined_Value or log-transformed log1p(Combined_Value) depending on model sensitivity

The feature set included keyword-derived trend columns (normalized), as well as derived temporal fields such as year and month. Categorical columns like State and Indicated were excluded to simplify modeling and reduce feature sparsity.

5.2 Depression Model

We first trained an XGBoost Regressor to predict depression-related search intensity.

- **Input**: Search trend vectors for depression-related keywords
- **Model Setup**:
  - Objective: reg:squarederror
  - Estimators: 300
  - Learning rate and tree depth were tuned for convergence
- **Evaluation**:
  - RMSE: ~10.5 on test set
  - The model captured general seasonality and trend patterns, but underperformed on sudden spikes (e.g., early 2020 COVID onset)

This suggests that while search trend signals reflect depressive patterns, they may lack the granularity or sensitivity to track abrupt public health shocks.

5.3 Anxiety Model

A second XGBoost Regressor was trained on anxiety-specific trends.

- **Input**: Same structure as above, but filtered for anxiety keywords
- **Evaluation**:
  - RMSE: ~9.81
  - This model performed slightly better than the depression model, tracking both seasonal spikes and smoother trend transitions more effectively
  - However, it still showed limitations in predicting sharp, short-term surges

The higher performance may reflect more consistent search behavior in anxiety-related terms, or better alignment with real-world stressors (e.g., exam seasons, holidays, financial stress).

5.4 Summary & Comparison of XGBoost Models

Model Type	*RMSE(Test)*	*Strengths*	*Weakeness*
Depression	~10.5	Capture baseline trends	Misses high-variance peaks
Anxiety	~9.81	Captures seasonality and transitions	Still limited on sharp, sudden events

XGBoost models demonstrated superior performance compared to linear regressors in capturing non-linear patterns and seasonal variations in mental health-related search data. Among the two target categories, the anxiety-focused model yielded lower RMSE and better alignment with both trend transitions and cyclical peaks. In contrast, the depression model exhibited higher residual error, possibly due to noisier signals or weaker digital proxies.

Despite the improved predictive accuracy, both models exhibited limitations in capturing abrupt surges in demand, such as those driven by unexpected socio-political or public health events. These findings underscore the potential of integrating digital behavioral data with external variables to enhance modeling robustness. The results also informed our SHAP-based interpretability phase, which aimed to further analyze the relative impact of each feature on model output.

6. SHAP Interpretability Analysis

6.1 SHAP Method Overview

To understand how individual features (i.e., search terms) influenced the predictions of our XGBoost models, we applied SHAP (SHapley Additive exPlanations). SHAP assigns a contribution score to each feature for every individual prediction, enabling both global and local interpretability.

We used shap.Explainer to compute SHAP values and generated summary plots that rank features by their average absolute impact across the test set. This analysis provides insights into which search terms most strongly affect model output and whether their effects are positively or negatively correlated.

6.2 Most Influential Keywords

SHAP analysis revealed several high-impact features in both depression and anxiety models:

- **Top-ranked keywords** included:
  - “seasonal depression”
  - “chronic depression”
  - “depression fatigue”
  - “depression weight loss”
  - “always tired”

These terms consistently contributed large positive SHAP values, indicating strong alignment with elevated mental health search trends. Their medical or clinical phrasing may signal more serious underlying concerns, making them predictive of higher Combined_Value.

6.3 Gendered & Clinical Term Patterns

We observed that gender- and population-specific terms had asymmetrical influence on model outputs:

- **“depression in women”** had a notably stronger impact than **“depression in men”**, suggesting potential gender differences in digital mental health expression or search behavior.
- Terms related to diagnosis (e.g., “clinical depression symptoms”, “anxiety screening”) had greater predictive weight than vague emotional expressions like “feeling sad” or “unmotivated”.

This differentiation supports the idea that medically-framed, structured search phrases are more informative than open-ended or casual queries in detecting population-level mental health signals.

6.4 Interpretability Summary

The SHAP analysis confirmed that:

- Feature importance was consistent with real-world health semantics
- Structured and clinically relevant keywords had stronger model impact
- Demographic-specific search terms revealed disparities in behavior and signal strength

These insights not only validate model behavior but also offer guidance for future work on personalized or demographically segmented mental health forecasting systems. In particular, medically framed and chronicity-related search terms emerged as the most predictive, suggesting that the model is especially responsive to structured digital health expressions.

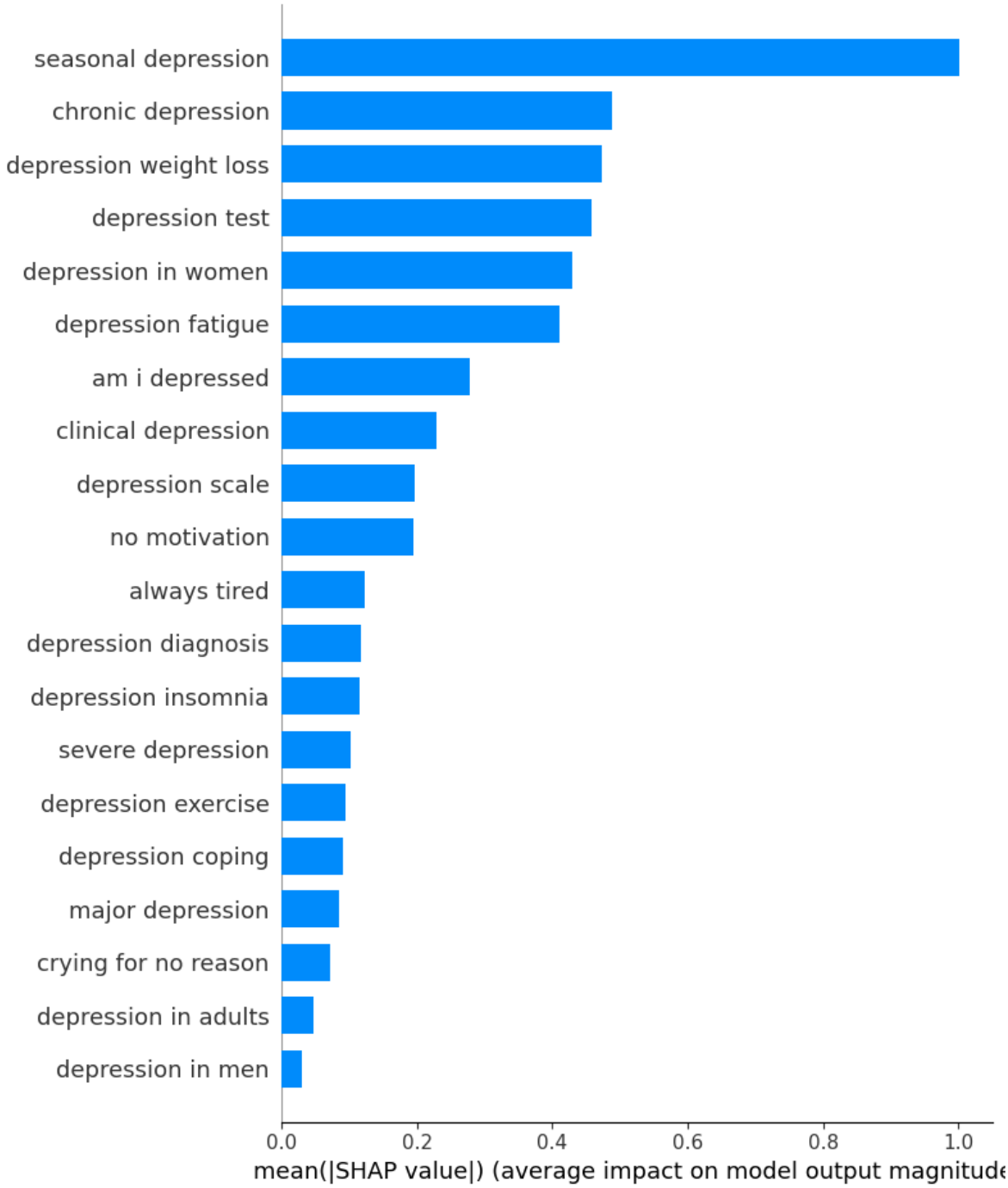

Figure 6.1 Summary bar plot (top features)

Figure 6.1 shows a summary bar plot of the top-ranked features based on their mean SHAP value. The highest-impact term was *"seasonal depression"*, followed by *"chronic depression"*, *"depression weight loss"*, *"depression in women"*, and *"depression fatigue"*. These terms all reflect either medical framing or long-term physical/psychological burden, reinforcing the model's sensitivity to clinical and chronic mental health concerns.

In contrast, emotionally expressive but vague terms such as *"crying for no reason"* and *"depression in men"* had much lower SHAP values, indicating weaker influence on model predictions. This highlights the challenge of capturing unstructured or low-intensity digital signals, especially when they lack medical specificity.

7. Project Deployment & Tech Stack

7.1 Tools and Libraries

The project was implemented in Python, using the following tools and libraries across different stages:

- **Data Collection**:
  - pytrends for Google Trends API access
  - pandas, os, time for batch processing and storage
- **Preprocessing & Analysis**:
  - pandas, numpy, scikit-learn, statsmodels for data cleaning, feature engineering, and linear modeling
  - matplotlib, seaborn, plotly for visualization
- **Machine Learning Models**:
  - xgboost for gradient boosting regression
  - scikit-learn for pipeline design and evaluation
- **Interpretability**:
  - shap for SHAP value calculation and summary plotting
- **App Deployment**:
  - streamlit used to develop an optional demo interface for showcasing model outputs and interactive charts

All environments were maintained in Jupyter Notebooks and version-controlled via GitHub.

7.2 GitHub Repository Structure

The entire codebase is organized and available in the following GitHub repository:
https://github.com/sandhya-mgs2/Capstone-4A/

Directory overview:

- data/ — cleaned CSV datasets and keyword files
- notebooks/ — exploratory data analysis, modeling, SHAP scripts
- app/ or streamlit_app.py — optional Streamlit interface
- README.md — project summary and instructions
- requirements.txt — dependency list for reproducibility

All contributions were version-controlled, with branches used to manage development stages and collaboration among team members.

7.3 Demo Page

We developed a lightweight prototype using streamlit to visualize model predictions and keyword trends. Key features include:

- Interactive keyword trend explorer (monthly line plots)
- Depression/anxiety prediction interface (based on search inputs)
- Summary SHAP explanation page

The prototype is locally deployable via streamlit run and demonstrates how search-driven mental health insights could be integrated into a public-facing analytics tool.

8. Conclusion & Future Work

8.1 Key Findings

This project demonstrated the feasibility and value of leveraging Google search trends as a proxy for analyzing population-level mental health patterns across the United States. By applying exploratory data analysis, linear regression, and XGBoost modeling techniques, we uncovered several meaningful insights.

Anxiety-related search behavior exhibited more pronounced seasonal patterns and led to better-performing models compared to depression-related trends. Moreover, states with higher search intensity frequently aligned with regions already known for limited mental health service availability, indicating potentially unmet needs. The most influential model features were structured, medically framed keywords such as “seasonal depression” and “clinical depression symptoms,” which consistently contributed to higher predicted values. Lastly, the SHAP analysis further validated the interpretability of our model and clearly identified the search terms with the greatest predictive impact.

8.2 Limitations & Challenges

Despite the promising results, several limitations should be acknowledged. One key challenge is the lack of clinical ground truth labels. Because the model relies on online search behavior as a proxy for mental health need, there is no direct link to actual prevalence data or diagnostic outcomes. This introduces a degree of uncertainty in interpreting the model’s predictions.

In addition, search behavior is inherently influenced by demographic, socioeconomic, and regional disparities in internet access and digital literacy, which may introduce systematic bias. Another notable limitation is the model’s limited responsiveness to abrupt, event-driven changes such as pandemic-related spikes. Without integrating real-time external data sources, the model struggles to capture these temporal disruptions. Finally, this study focused solely on state-level patterns, leaving finer-grained spatial and demographic insights unexplored.

8.3 Future Directions

Building on this work, several enhancements can be pursued to increase the accuracy, robustness, and real-world utility of the model. Incorporating external signals such as news coverage, social media activity, or public policy shifts could improve the model's ability to respond to sudden events and reflect changing psychological environments.

Furthermore, integrating demographic and socioeconomic variables would allow for a more nuanced analysis of disparities in mental health expression and access to care. Scaling the model to a city or county level would also enable more localized and targeted interventions. Developing a real-time dashboard interface could help translate the model into a functional public health monitoring tool, offering accessible visual insights for policy stakeholders and researchers.

Lastly, validating the model against clinical datasets, such as prevalence estimates from the CDC or provider shortage reports from HRSA, would provide essential calibration and enhance its credibility. Together, these directions can help transform the current framework into a comprehensive, interpretable, and actionable digital health surveillance system.

References & Data Sources

Data Sources

1. **Google Trends Search Data**
   Monthly mental health-related keyword trends (2020–2024)
   Accessed via Pytrends API
   GitHub script: Crawl data code from google.ipynb
2. **HRSA Area Health Resources Files (AHRF)**
   State-level data on licensed mental health professionals
   Source: https://data.hrsa.gov/data/download
3. **Self-Engineered Target Variable: Combined_Value**
   Aggregated keyword trend scores by state and month
   Dataset: combined_depression_anxiety.csv
4. **Mental Health Keyword List**
   Curated and categorized search terms
   File: cleaned_mental_health_keywords.txt (see GitHub repo)

Tools & Libraries

- Python 3.9
- pandas, numpy, scikit-learn, statsmodels
- xgboost, shap
- matplotlib, seaborn, plotly
- streamlit (for prototype demo)

GitHub Repository: Capstone-4A GitHub