Exploring Relationships between Healthcare Outcomes and Local Industry

By Adrienne Martz, Gretchen Lam and Sophia Cheng

Motivation

Background

While the relationship between socioeconomic factors and health outcomes has been studied (Bhatt et al., 2022), we wanted to look at it through a slightly different lens by exploring the relationship between the local economy (dominant industries) and health outcomes.

There are a couple papers that have studied the connection between health and industry sector. For example, Arheart et al. (2011) examined the relationship between self-rated poor health and mortality rates with occupations and industry sectors.

However, there is no published research looking at the connection between population level health measures and locally dominant industries. This is the focus for our project.

Objectives

Using county-level health measures and county-level industry data, we would like to answer the following questions:

- Do we observe any relationships between the local industry and selected health measures?
- Are health outcomes better when local industry is one type over another (ex. do industries with more physical labor have less adult obesity)?

Change in Scope

In our original proposal, we wanted to look at Medicare claims per capita and see if counties with the same local industry had similar claim activity. Due to the volume of data and time constraints, we decided to scale back the scope and omit the claims data from our analysis.

Data Sources

CHR Primary: County Health Rankings Analytic Dataset

The County Health Rankings Analytic dataset contains county-level measures from a variety of national and state data sources that are used to rank counties by their overall health. The dataset contains data combined from prior years and was released in 2021. The dataset includes:

Health Outcomes such as Frequent Mental Distress
Health Factors such as Limited Access to Healthy Foods

https://www.countyhealthrankings.org/sites/default/files/media/document/analytic_data2021.csv

Size	3,196 rows x 690 columns
Format	CSV
Access Method	Download from website

CBP Secondary: County Business Patterns Dataset

County Business Patterns is an annual series that provides county-level economic data by industry as identified by the North American Industry Classification System (NAICS) code released by the U.S. Census Bureau. The 2021 dataset includes the number of establishments, employment during the week of March 12, and annual payroll.

https://www2.census.gov/programs-surveys/cbp/datasets/2021/cbp21co.zip

Size	1,090,164 rows x 25 columns
Format	TXT
Access Method	Download from website

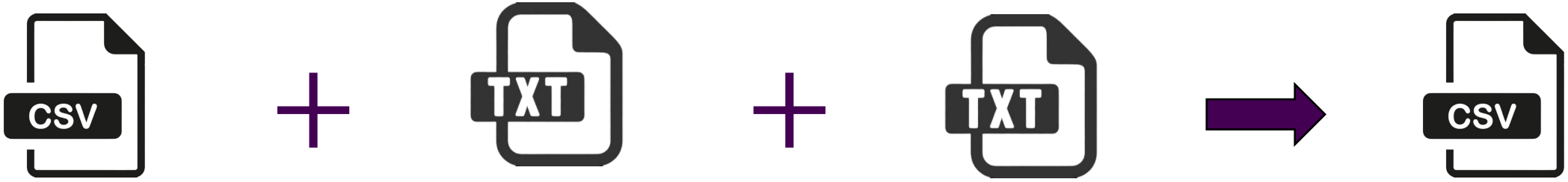
NAICS Secondary: NAICS Code to Industry Name

This dataset maps the NAICS code to the industry name.

https://www2.census.gov/programs-surveys/cbp/technical-documentation/reference/naics-descriptions/naics2017.txt

Size	2,003 rows x 2 columns
Format	TXT
Access Method	Download from website

Data Manipulation



CHP
Initial Shape: 3,196 x 690

Data Cleaning and Manipulation:
- Converted FIPS column from int to string (match CBP FIPS column type)
- Dropped unused columns (health metrics) and rows (non-county)
- Extract county names - remove “County”
- Subset for health metrics of interest using regex and manual list

Final Shape: 3,007 x 35

CBP
Initial Shape: 1,090,164 x 25

Data Cleaning and Manipulation:
- Combine state and county FIPS codes to get 5-digit FIPS code (string)
- Delete state and county codes when done
- Move FIPS column to front of dataframe

Final Shape: 1,090,164 x 22

NAICS
Initial Shape: 2,003 x 2

Data Cleaning and Manipulation:
- Set column names to lowercase for merging with CBP
- Rename “description” column to “Industry Description” so it’s more descriptive
- Filter for top-level NAICS codes

Final Shape: 20 x 2

C_H_B

Data Cleaning and Manipulation:
- Merge CBP with NAICS using NAICS codes
- Merge CBP/NAICS with CHP using FIPS codes
- Move identifying columns to front (FIPS, state, county)

Final Shape: 47,560 x 57

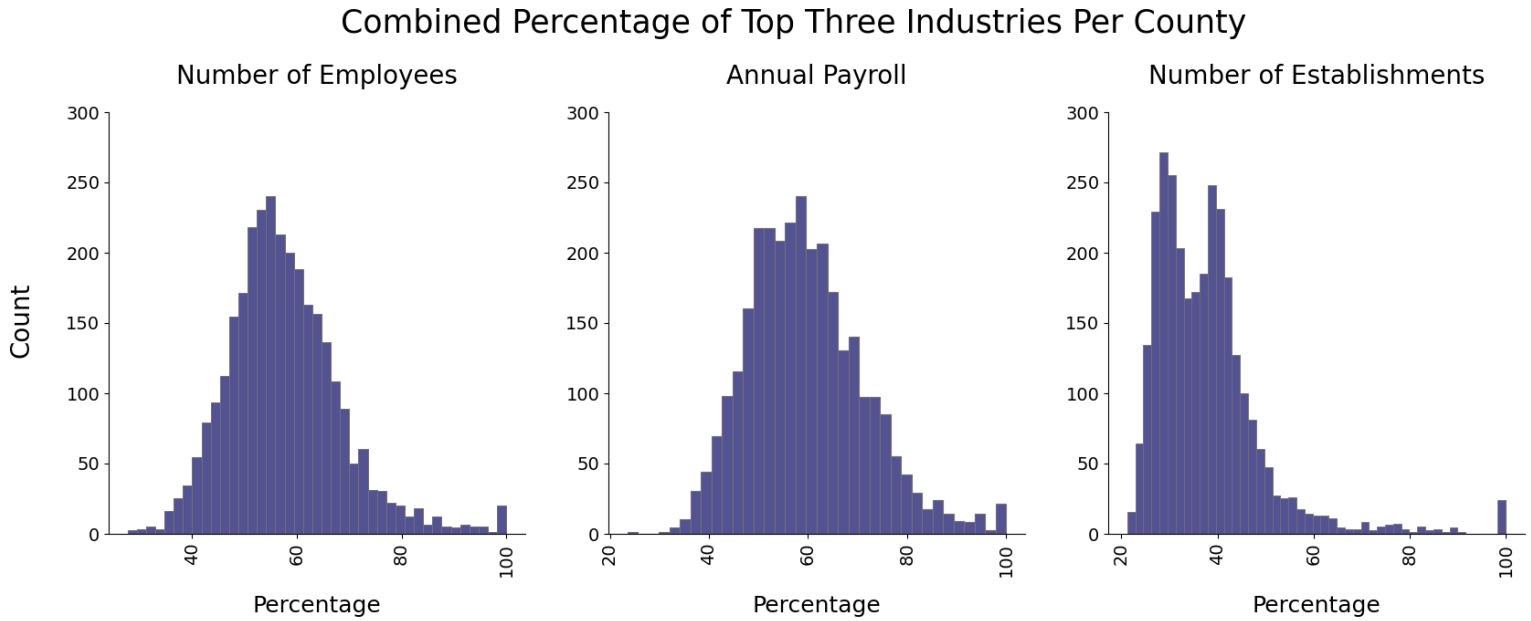
Analysis & Visualizations

Selection of Dominant Industries

The CBP dataset had several levels of granularity to identify the industry in a given county. We chose to look at the top-level category to narrow our analysis. We investigated **three different ways** to identify the dominant industry in a county:

number of employees *annual payroll* *number of establishments*

For each method, we looked at what percentage of each county is represented by the number one ranked industry. The distributions reveal that the top industry generally represents **less than 40%** of the county. Since distributions were right skewed and below 50%, we decided to look at **the top three industries** in each county to capture more of the workforce. For the number of employees and annual payroll, the top three industries combined mostly represent **40 – 80%** of the counties while the top three industries only account for **20 – 60%** of the counties using number of establishments.



Analysis & Visualizations

Selection of Dominant Industries

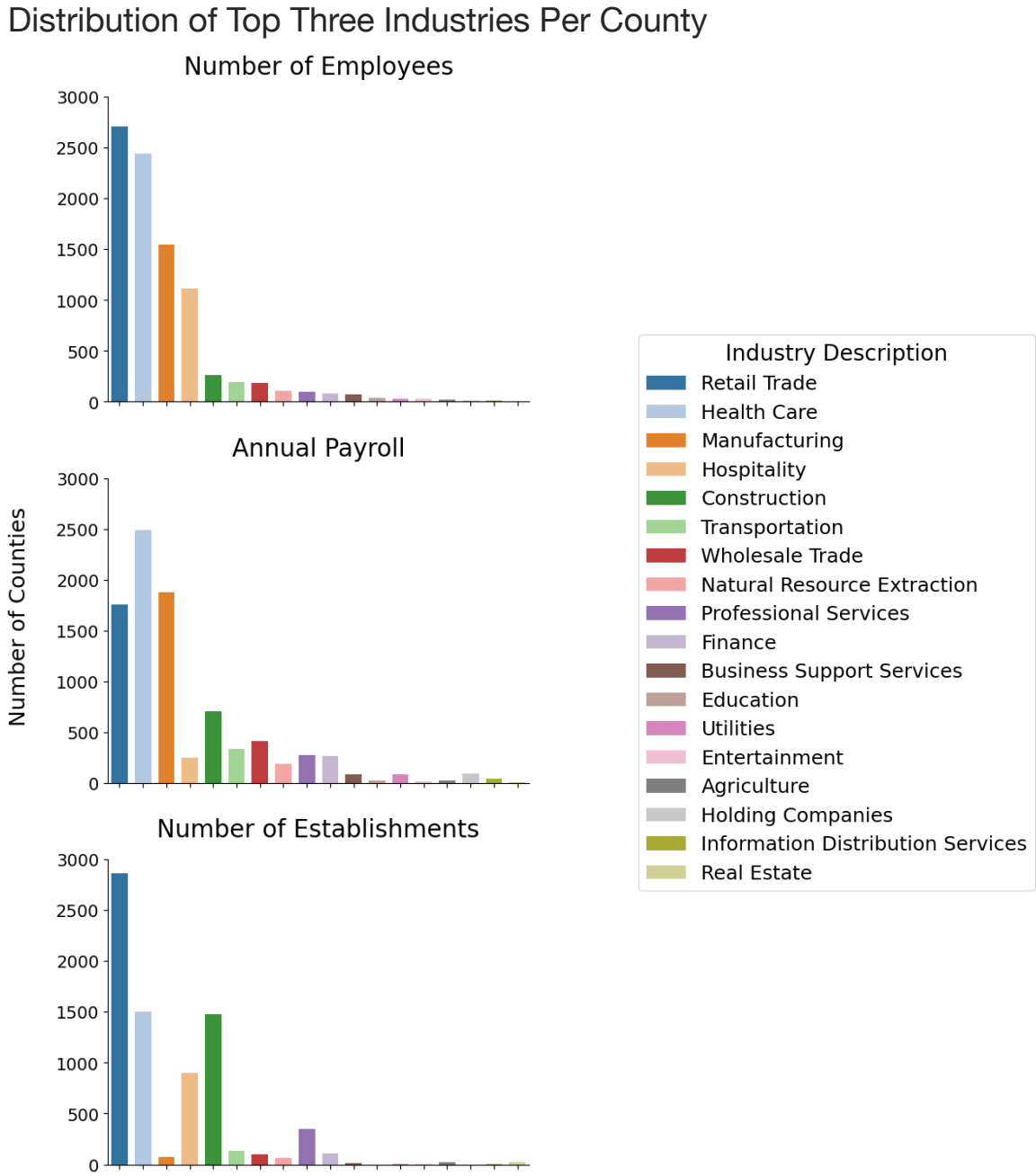
To determine which method to use for selecting the top industries, we looked at the distributions of industry type for the top three industries in a county as identified by each method.

The number of employees and annual payroll methods show the **same top three industries** while number of establishments shows the same top two.

Given that the prior slide showed the top three industries as identified by number of establishments does not cover the majority of the workforce in a county and shows different results than the other two methods for determining the top industries, we eliminated the number of establishments as a method.

After some consideration, we felt using annual payroll would be skewed by industries that are more lucrative and chose the **number of employees** as our metric.

Industries that were not present in 20 or more counties were dropped from our analysis as we did not feel there was enough data to support any relationships found with the health measures.



Analysis & Visualizations

Selection of Health Measures

The CHR dataset contains information on Health Outcomes, which measure how healthy residents are today, and Health Factors, which represent factors that will impact their health in the future. The dataset is used to calculate the County Health Rankings and incorporates 35 measures into their rankings model.

We chose **18 outcomes and factors** to look at. The table to the right lists the categories for the measures and lists one measure from that category that we looked at when creating our health profiles for the different industries.

We included *Quality of Life* measures to get a sense of the physical and mental difficulties residents face daily. *Access to Care* and *Quality of Care* measures highlight if there is a lack of access to adequate care in a county. *Air & Water Quality* measures capture the direct impact of local industry in the area. *Alcohol & Drug Use* measures signal if there are severe addiction issues in a county. *Diet & Exercise* measures indicate how healthy the lifestyles of residents are. By looking at *Unemployment*, we can get a sense if the top three industries provide enough jobs in the county. *Housing & Transit* can show if wages allow for suitable housing.

Selected Measures of Interest
(One Example per Category)

Health Outcomes	
Length of Life	*Life Expectancy*
Quality of Life	*Frequent Mental Distress*

Health Factors	
Access to Care	*Primary Health Providers*
Air & Water Quality	*Air Pollution**
Alcohol & Drug Use	*Excessive Drinking*
Diet & Exercise	*Adult Obesity*
Employment	*Unemployment**
Housing & Transit	*Severe Housing Problems**
Quality of Care	*Preventable Hospital Stays**

* Indicates measure was used to calculate the ranking in the CHR data

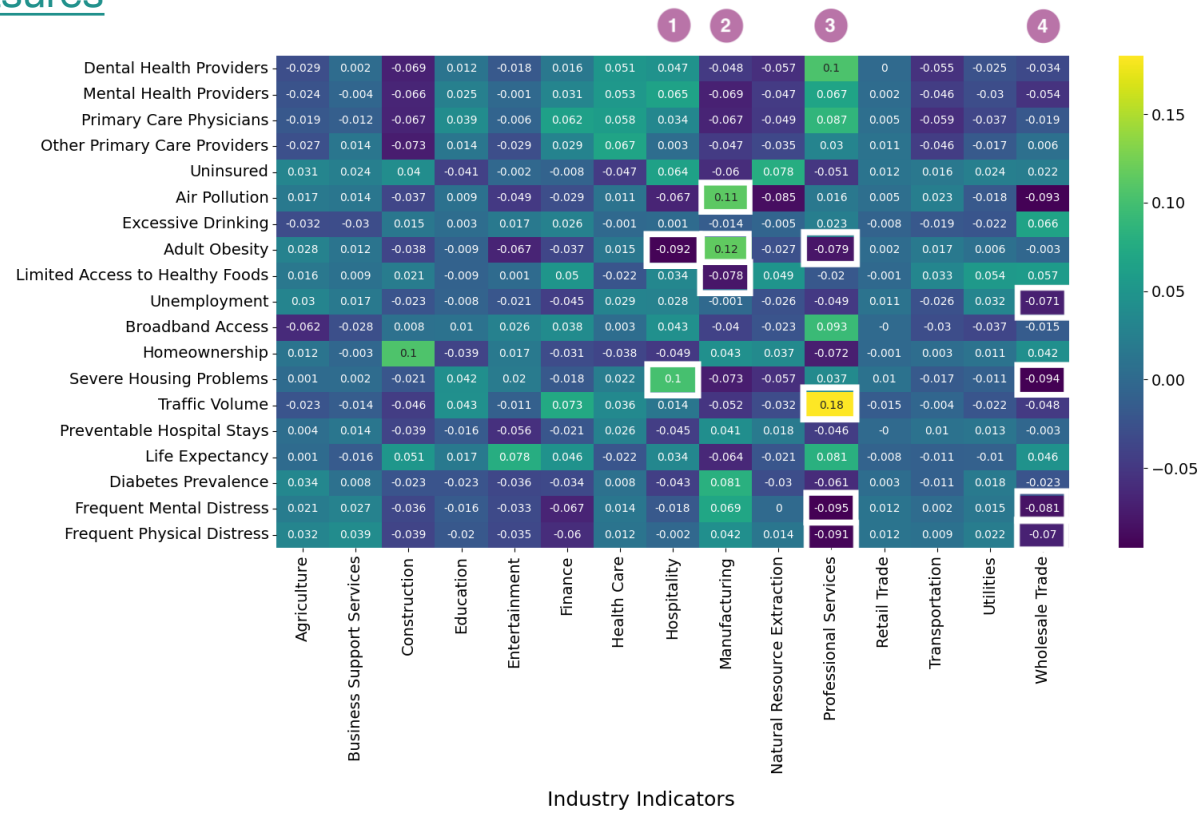
Analysis & Visualizations

Correlations Between Dominant Industries and Health Measures

To look at the relationships between our selected health measures and the local dominant industries in a county, we created **dummy** (indicator) variables for each industry type to signify their presence in a county.

Although the correlations to the right do not have large coefficients, we can see some positive and negative relationships between the variables.

Highlighting some of the industries and correlations of interest:



1 **Hospitality** – There is a positive relationship with severe housing problems, which may be due to scarce housing. There is also a negative relationship with adult obesity.

2 **Manufacturing** – There is a positive relationship with air pollution which seems likely since factories produce pollution. Despite a negative relationship with limited access to healthy foods, there is a positive relationship with adult obesity.

3 **Professional Services** – There is a positive relationship with traffic volume, which means more people are commuting. There is a negative relationship with frequent physical and mental distress, which may play a part in having a negative relationship with adult obesity.

4 **Wholesale Trade** – There is a negative relationship with unemployment, which may explain the negative relationship with housing problems. This may also contribute to the negative relationship with frequent physical and mental distress.

Analysis & Visualizations

Health Profiles

Using the correlations, we were able to select health measures that had stronger relationships with the different industries. The table below describes the measures we included in our health profile.

Measure	Definition
Primary Care Physicians	Originally rate of number of providers/100K population, but was modified to be per person
Mental Health Providers	Originally rate of number of providers/100K population, but was modified to be per person
Uninsured	Percentage of population under 65 without health insurance
Frequent Physical Distress	Percentage of adults reporting 14 or more days of poor physical health per month (age-adjusted)
Frequent Mental Distress	Percentage of adults reporting 14 or more days of poor mental health per month (age-adjusted)
Limited Access to Healthy Foods	Percentage of population who are low-income and do not live close to a grocery store
Adult Obesity	Percentage of adult population (age 18 and older) that reports a BMI greater than or equal to 30 kg/m2 (age-adjusted)
Severe Housing Problems	Percentage of households with at least 1 of 4 housing problems: overcrowding, high housing costs, lack of kitchen facilities, or lack of plumbing facilities

We included the provider measures to see how access to care differs. Uninsured would indicate if there is a financial burden with receiving care. Frequent distress can give a sense of day-to-day health. Lack of access to healthy foods would indicate food deserts and would most likely play a role in adult obesity.

To create the parallel plots, we calculated the median value for each of our measures by industry type and plotted the largest five industries and the smallest five industries as they showed differing health outcomes.

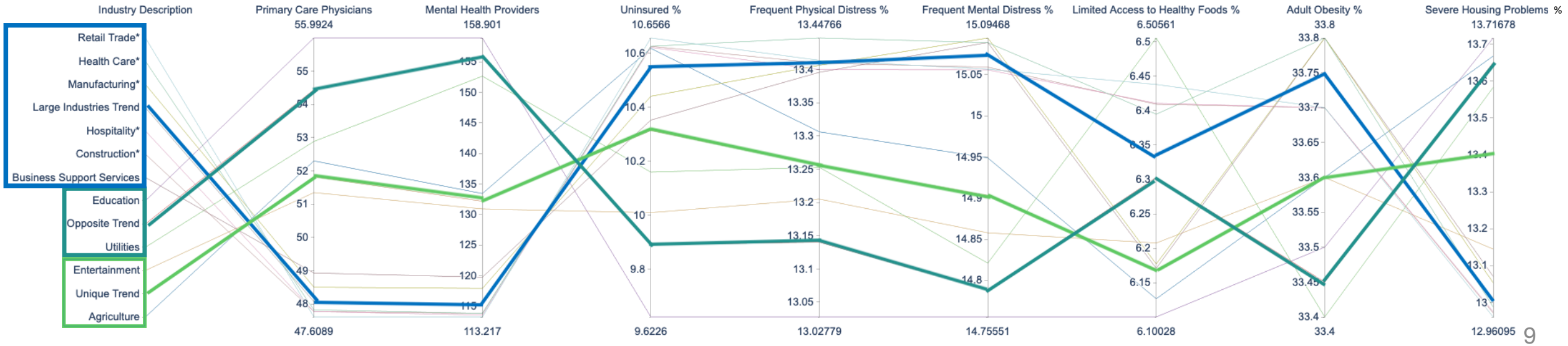
Analysis & Visualizations

Health Profiles

While there are some measures that have small ranges in values, there are a few trends to note. The lines in the visualization below represent the average values of the grouped industries:

- All large industries and Business Support Services **have very similar health profiles** with lower access to providers, higher uninsured, higher levels of frequent distress and lower severe housing problems.
- Education and Utilities stand out as they **have opposite trends** from all other industries with better access to health care providers, lower levels of uninsured, less frequent distress and lower adult obesity. They have higher levels of severe housing problems. They diverge on access to healthy foods.
- Entertainment and Agriculture **have unique health profiles**. They have moderate access to providers, lower frequent distress, more access to healthy foods and moderate adult obesity. They diverge on uninsured and severe housing problems.

* Indicates in visualization that industry is one of the top five largest



Analysis - Findings

Conclusions

Our main goals for this project were to determine if there are any relationships between health measures and dominant industries at the county level, and if health outcomes are better for one industry over another.

Using a correlation matrix, we were able to **identify small positive and negative correlations** between some industries and health measures. For example, counties where manufacturing was a dominant industry showed a positive correlation with adult obesity.

We also identified **industries with similar health metrics**. For example, the five largest industries all have lower access to medical providers and higher levels of frequent physical and mental distress.

We identified another trend that was the **opposite** of this for Education and Utilities, which are two very different industries. They have better access to health care providers, lower levels of uninsured, less frequent distress and lower adult obesity.

While the industries are not similar in type, there are industries that **have better outcomes** than others.

Limitations and Future Directions

Due to short timeframe of this project, we have several limitations that we were not able to address:

- We used the default top-level NAICS industry categories, which were sometimes too broad and included dissimilar occupations. We may see stronger correlations with health outcomes if we picked more **narrowly-defined categories**.
- We compared health measures and industries by county, which can have very small or very large populations. It would be ideal to **compare data for smaller regions** (ex. zip code or census tracts) or select similarly-sized counties to compare.
- Counties with the same dominant industries may differ in other characteristics such as race and gender, which could impact the correlations with health measures. We did not look at the impact of these **confounding variables** in this analysis. A future direction could include breakdowns of health measures by these variables.

Statement of Work

Data Cleaning & Manipulation

Gretchen and Adrienne wrote code to merge the three datasets together. This included limiting the datasets to the measures and industries we were interested in. Sophia assisted on code quality and best practices.

Analysis & Visualizations

Sophia created visualizations to help with our initial EDA. All three developed visualizations to support and display our narrative. Gretchen and Adrienne further developed the visualizations, including fine-tuning to make the visualizations expressive and effective.

Report

Adrienne outlined the slide docs and wrote initial drafts and bullet points that were edited and completed by Gretchen and Sophia.

Collaboration

Throughout the project, we met via Zoom three days a week and collaborated on coding using notebooks in Deepnote. Versioning was a small issue, but we worked out how to handle it. With future work we would explore using GIT as well as setting up an environment that could handle larger datasets.

References

Arheart, K. L., Fleming, L. E., Lee, D. J., LeBlanc, W. G., Caban-Martinez, A. J., Ocasio, M. A., McCollister, K. E., Christ, S. L., Clarke, T., Kachan, D., Davila, E. P., & Fernandez, C. A. (2011). Occupational vs. Industry Sector Classification of the US Workforce: Which approach is more strongly associated with worker health outcomes?. *Am. J. Ind. Med.*, 54(10), 748-757. https://doi-org.proxy.lib.umich.edu/10.1002/ajim.20973

Bhatt, J., Batra, N., Davis, A., Rush, B., & Gerhardt, W. (2022, June 22). *US health care can't afford health inequities*. Deloitte Insights. https://www2.deloitte.com/us/en/insights/industry/health-care/economic-cost-of-health-disparities.html