Fortune 1000 Companies by HDI

Rafael Coelho, Gunay Azizova

Project Overview

Topic

Top 1000 American Company Performance vs. HDI(Human Development Index)

Why

We wanted to have a business case project related to the economy sphere, which also has solid data.

Datasets

D1: Fortune 1000, 2023 (Kaggle Dataset)
D2: HDI of US States, 2022
(Web Scraping Wikipedia)

Hypothesis

H0: Company Concentration vs. HDI
H1: Impact of Top Sectors on HDI
H2: Company Revenue vs. HDI

Project Overview: Quick Definitions

Fortunes 1000

An annual list of the 1000 largest public American companies maintained by Fortune.

- american public companies
- ranked by revenue

HDI (Human Development Index)

An index used by United Nations, composed of:

- life expectancy
- education (years of schooling)
- per capita income

Project Overview:
Data Cleaning & Analysis

- **Data Cleaning of Dataset 1 (Cleaning)**
  - Renaming, filling null values, lowercase values, converting data, dropping columns, etc.
- **Web Scraping of Dataset 2 (Cleaning & Wrangling)**
  - Making numerical values usable, dropping non-state values, renaming
- **Merging & Exporting of both Datasets to Excel**
  - Goal was to preserve the datasets both in their merged and individual formats, in order to avoid risk
- **Visualization through Tableau**
  - Connected to our data source (Excel DF)
  - Created Charts based on the data, in separate Worksheets
  - Combined primary Charts into a final Dashboard
  - Publish to Tableau Public & Exported the Charts

Data Wrangling and Cleaning

The Attempt ->

The Main Challenge: Table with alternating Rowspans

16	North Dakota	0.934
	Rhode Island	
18	Illinois	0.932
19	Alaska	0.931
	Nebraska	
	Utah	

Key Takeaway:

There is always a different path :)

The Solution: Pandas (read_html method)

```
url = 'https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_Human_D
tables = pd.read_html(url)
df=tables[0]
```

```
for i in soup.find_all("tr"):

    if i == 0:
        continue

    for j in i.find_all("td"):
        ### J BECOMES DEFINED
        temp_var = j.get_text().strip()
        ### J BECOMES DEFINED
        if "." in temp_var:
            temp_var = temp_var.split(" ")[0]
            if "~" in temp_var:
                temp_var = temp_var[1:]
        try:
            temp_var = float(temp_var)
        except:
            if len(j) >= 3:
                state_list.append(temp_var)
            else:
                if "—" in temp_var:
                    rank_list.append(0)
                else:
                    if temp_var[-1] == "]":
                        temp_var = temp_var.split(" [")[0]
                        state_list.append(temp_var)
                        last_state = temp_var
                    else:
                        state_list.append(temp_var)
                        last_state = temp_var
                        #print(last_state)
                        #print(temp_var)
        else:
            if temp_var < 0.999999 :
                hdi_list.append(temp_var)
                last_hdi = temp_var
            else:
                rank_list.append(temp_var)
                last_rank = temp_var
    temp_var = str(temp_var)
```

Teamwork & Project Management

Project Structure

- From the get-go, we defined how we want to approach the Project and we stuck to it

Data Selection -> **Data Processing** -> **Analysis** -> **Visualization**

- It was important to us to completely finish a Step, before moving on to the next

- We actively used GitHub and avoided any merge conflicts

Teamwork

- Strong Collaboration
- Aligned on Project Goals

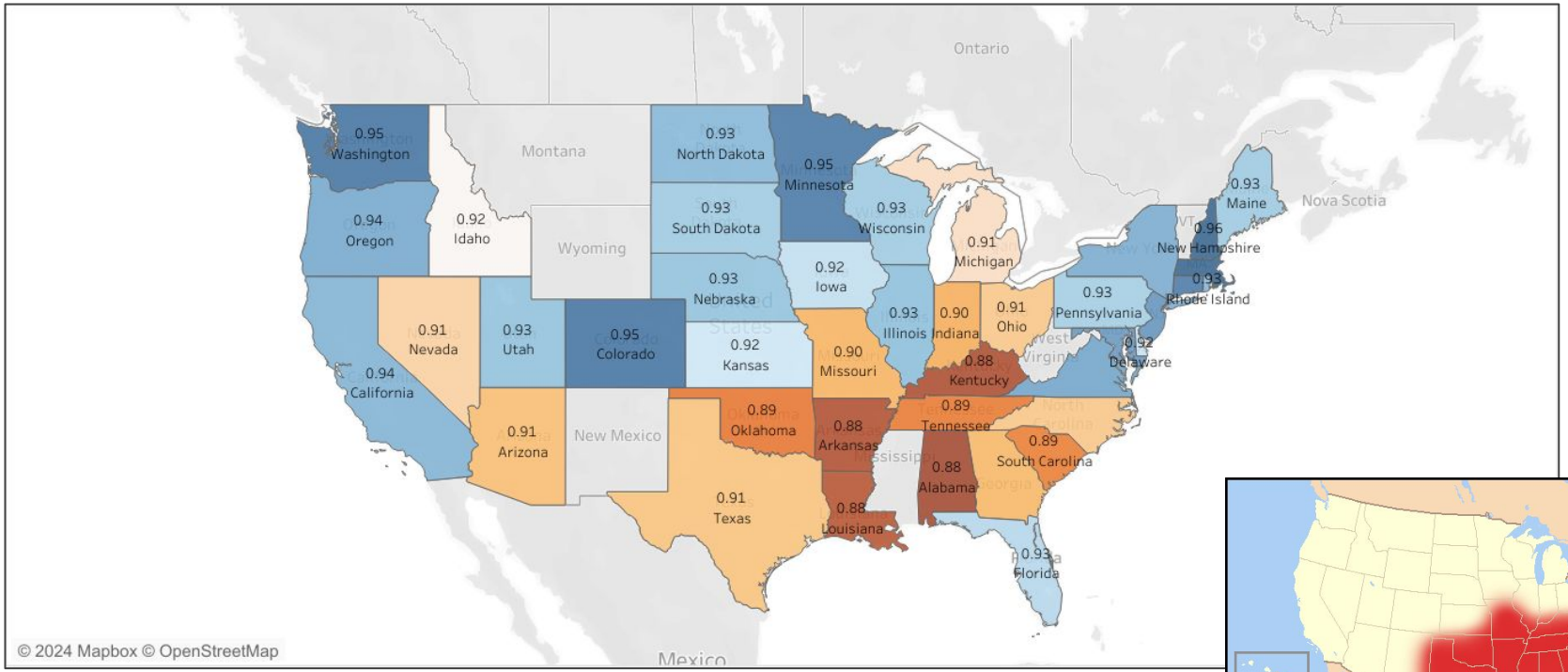
Creative Approach to Blockers

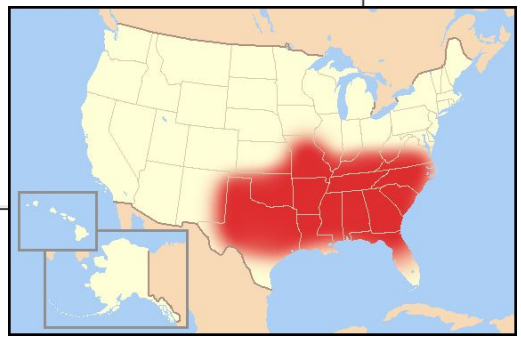
- Being open to different approaches kept us from blocking project progress.

Exchange of Knowledge

- GitHub: risk management
- Tableau: friendly visualisation

Conclusion & Insights:
Human Health Development by State[1]

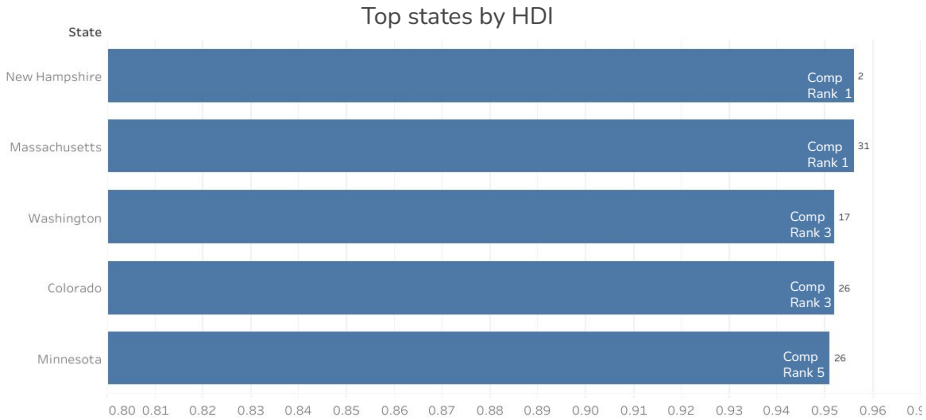




1: Montana[924], Wyoming[936], New Mexico[884], Mississippi[858], West Virginia[870] & Vermont[945], do not have any 1000 Fortune Headquarters and have been excluded from this analysis.

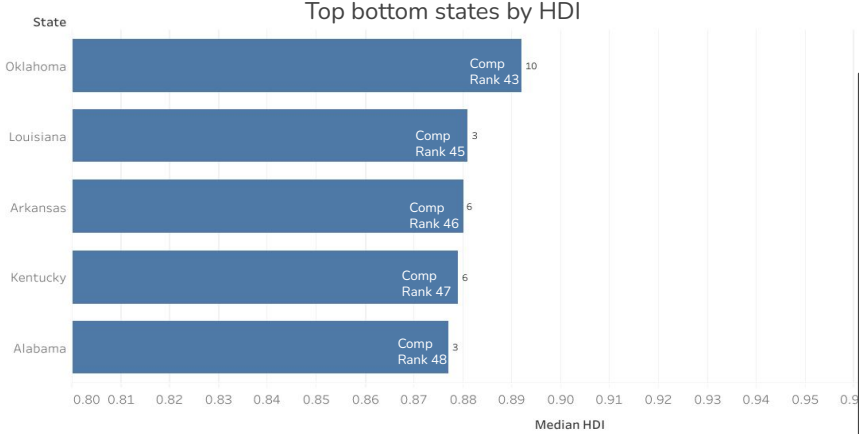
Fun Fact: Bible Belt has the worst HDI

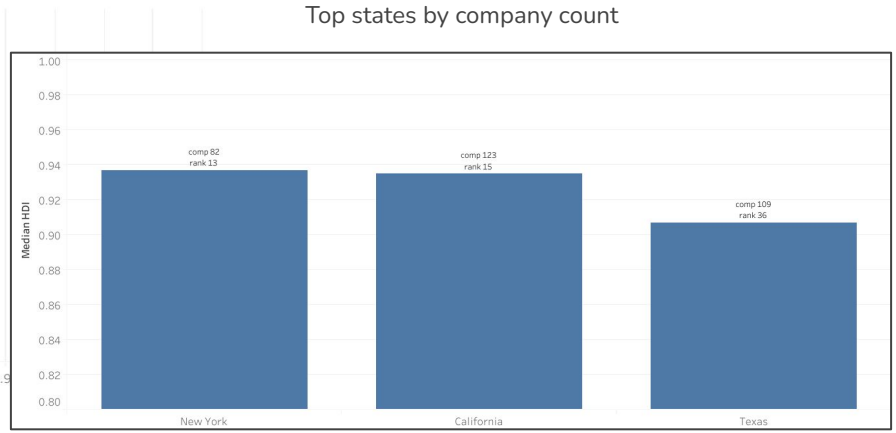
H0: Company concentration is negative on HDI



H0 is FALSE

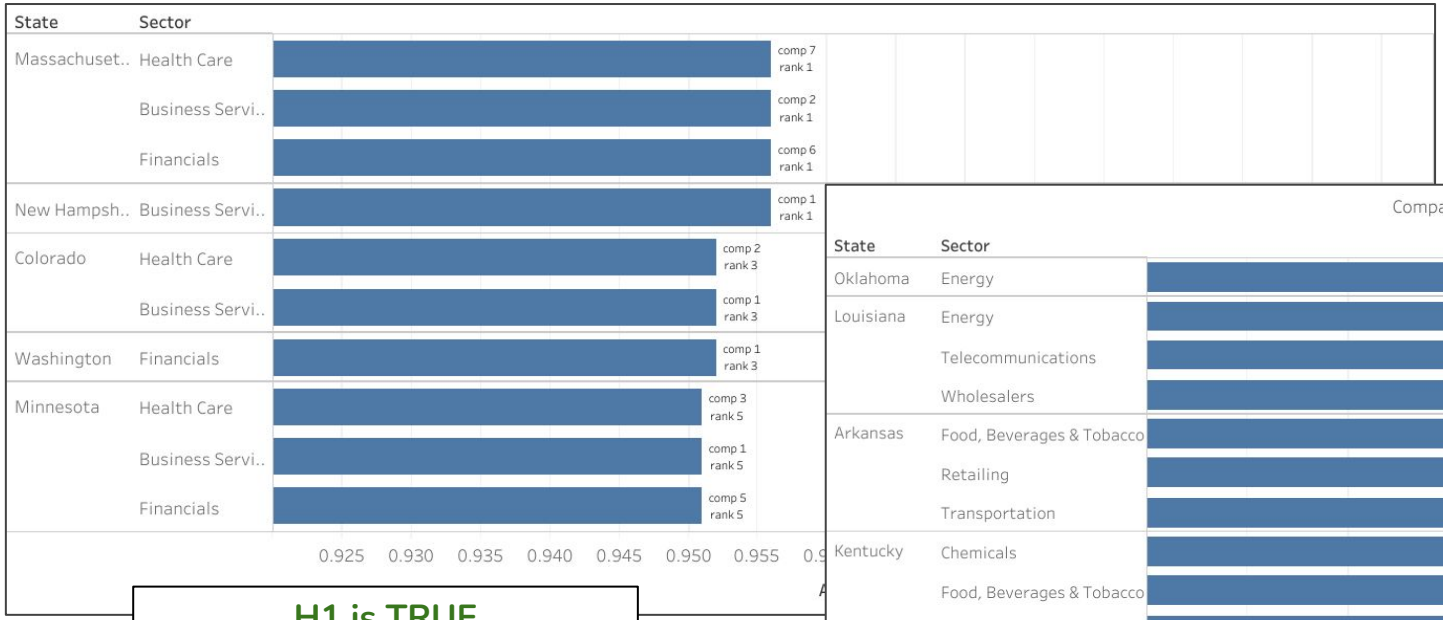
- No clear correlation between **Top 5/Bottom 5** states by Fortune 1000 companies and HDI.
- **High-concentration** states (California, Texas, New York) **rank neutrally**, outside the top/bottom 20% in HDI.



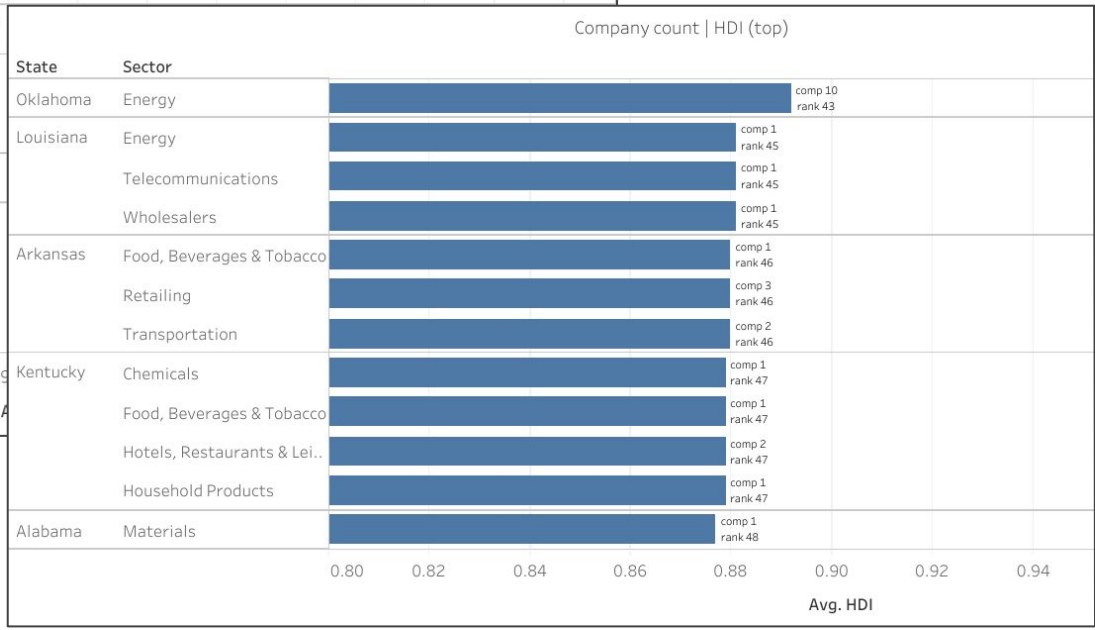


H1: Company Concentration in Top Sectors Positively Correlates with State HDI

Top states by HDI & top sectors



Bottom States without top sectors



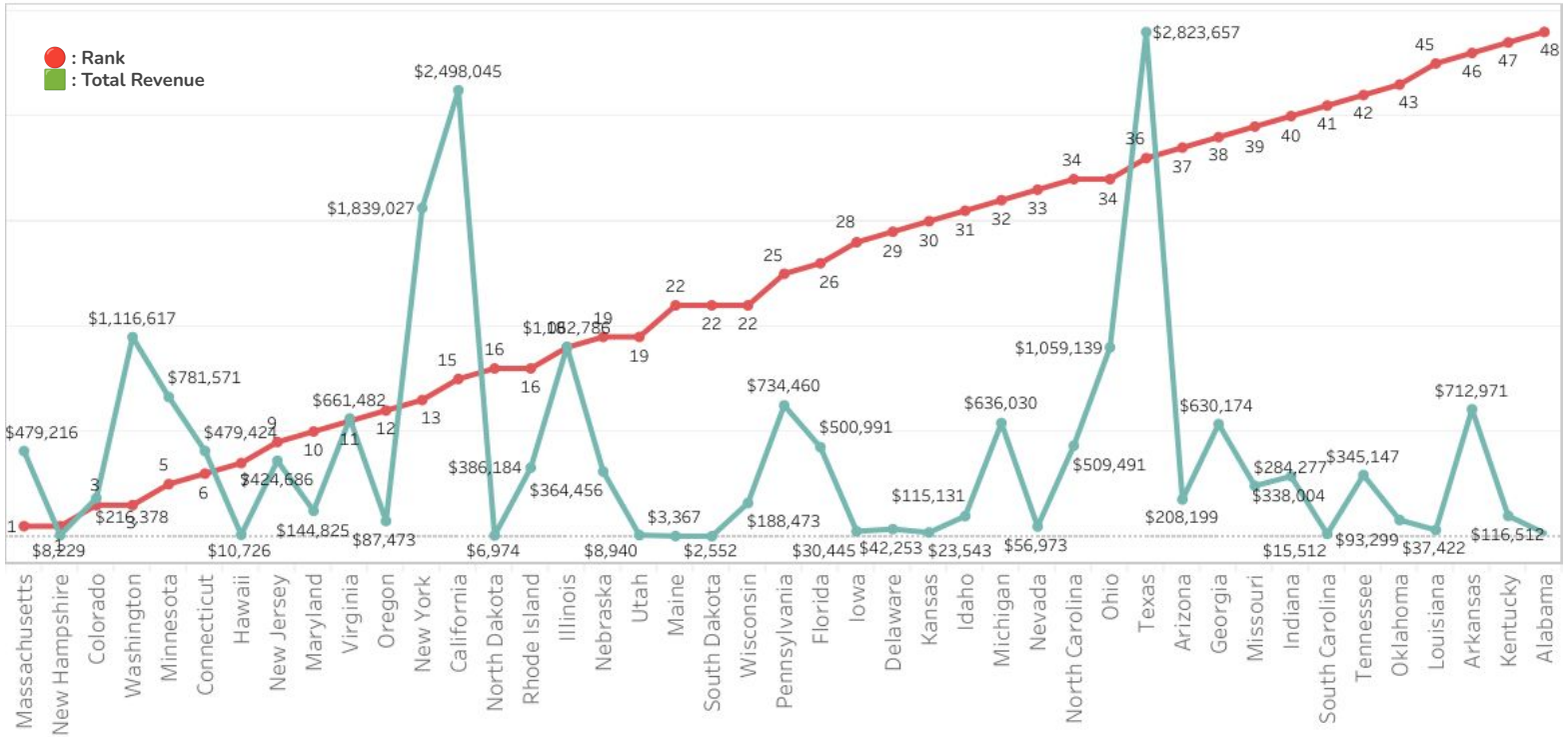
H1 is TRUE

- **Top 5 states** have companies in top sectors (**Health, Services & Finance**).
- States **without** companies in these **sectors** rank in the **bottom 20%**.

H2: States with Top Revenue-Generating Companies Show Positive HDI Correlation

H2 is FALSE

States by Revenue & Ranking



Top 3 by Revenue

#1: **TX** with $2.82 trillion, ranked 36th by HDI

#2: **CA** with $2.50 trillion, ranked 15th by HDI

#3: **NY** with $1.84 trillion, ranked 13th by HDI

Thank you!

ARCHIVE

Project Overview

Topic: Top 1000 American Company Performance vs. HDI(Human Development Index)

Why: We wanted to have a business case project related to the economy sphere, which also has solid data.

Dataset 1 -> Kaggle Dataset, contains 2023 performance of Top 1000 American Companies
Dataset 2 -> Webscraping Wikipedia, contains 2022* HDI performance of American States

H0: States with a higher concentration of top companies exhibit a negative correlation with Human Development Index (HDI) scores, suggesting that economic concentration in corporate hubs may not directly translate into broader human development outcomes.

H1: States with a high concentration of the highest contributing sectors to the US Economy (Health, Services & Finance), have a positive correlation to the their HDI.

H2: States with highest revenue-generating Fortunes 1000 companies, have a positive correlation with their HDI.

Exploratory Data Analysis:

- Initial structuring and accessing of the Data & Data Frames on Python

- In sequence, Dataframes were explored on Tableau:
  - To better visualize and understand our results
  - To come to conclusions regarding our key hypothesis

Project Overview: Quick Definitions

Fortunes 1000

An annual list of the 1000 largest public American companies maintained by Fortune.

- american public companies
- ranked by revenue

HDI (Human Development Index)

An index used by United Nations, composed of:

- life expectancy
- education (years of schooling)
- per capita income