Final Project

Group 99 : Jason Li, Jiayun Huang, Huimin Guan,
Tao Chen, Yiyang Chen

AGENDA

01
Data Analysis

02
Challenge

03
Final Algorithm

04
Results

05
Summary

Data Analysis

Discover Dta

- **7 variables**
  Business_id, Name, Address, City, State, Zip-code, Size
- **Left data**
  94,586 data point
- **Right data**
  91,792 data point

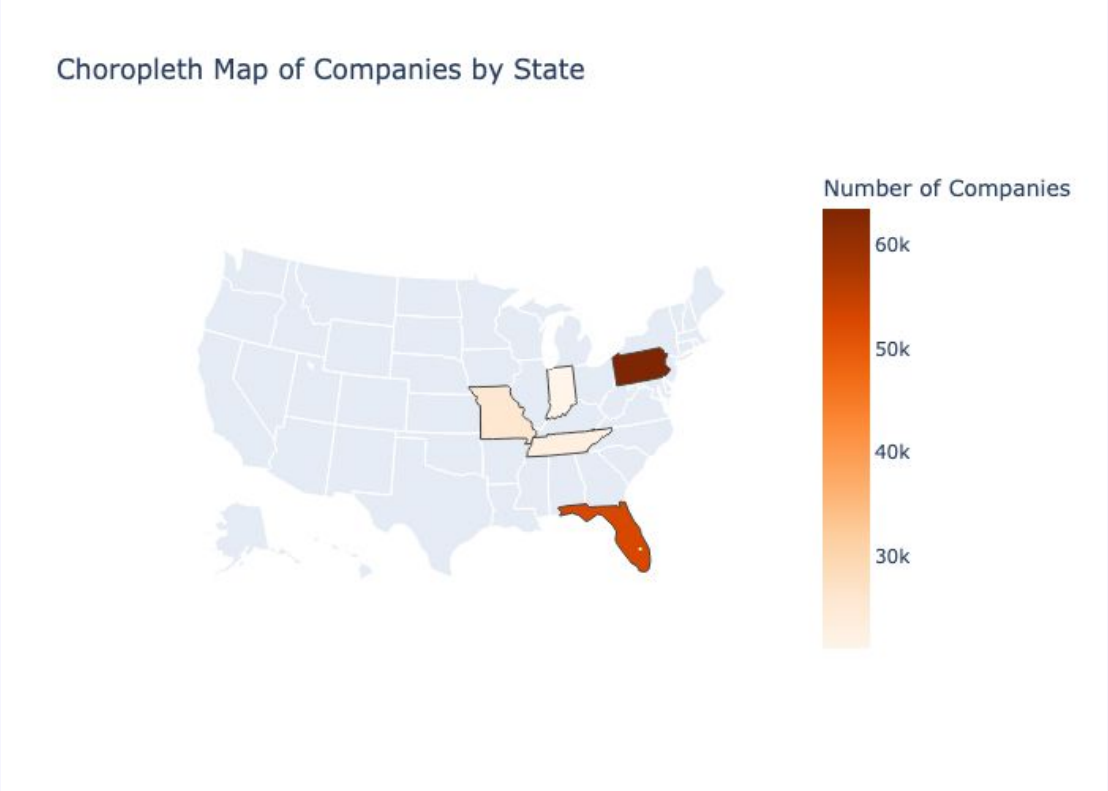
Prepare Data

- Select the columns we need
- Exact the first 5 number of zip

Clean Data

- Convert all the text to lowercase
- Remove all the special characters for name, address, city, state
- Remove all the “LLC”,”INC”,LTD”

Visualization



Challenges

```
# Calculate the similarity scores for each pair of entities
matches = []
for i, row1 in left_data.iterrows():
    for j, row2 in right_data.iterrows():
        # Calculate similarity of name and address columns
        name_similarity = jaccard_similarity(row1["name_clean"], row2["name_clean"])
        address_similarity = jaccard_similarity(row1["address_clean"], row2["address_clean"])
        # Calculate similarity of zip code columns
        zip_similarity = zip_code_similarity(row1["postal_code"], row2["zip_code"])
        # Combine the similarities to get an overall confidence score
        confidence_score = (name_similarity + address_similarity + zip_similarity) / 3
        # Add the match to the list if the confidence score is above a certain threshold
        if confidence_score >= 0.9:
            matches.append((row1["entity_id"], row2["business_id"], confidence_score))

# Output the matches as a CSV file
matches_df = pd.DataFrame(matches, columns=["entity_id", "business_id", "confidence_score"])
matches_df.to_csv("matches.csv", index=False)
```

63m 45.9s

ALGORITHM

01

Data cleaning

Lower case
Special Character
Zip code
Company name suffix (inc/llc/ltd)

02

Sorting

Increase the likelihood of matching

03

Merging

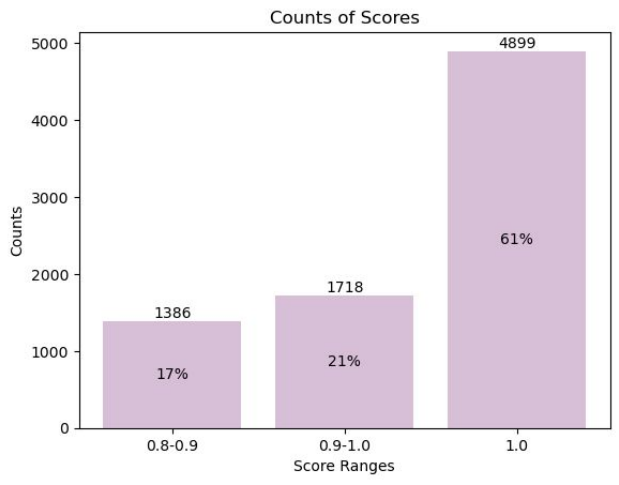
Zip code
City
State
First word of names and addresses

04

Fuzzy Wuzzy

Compare names and addresses
Levenshtein distance

Result



	left_id	right_id	score	name_left	name_right	address_left	address_right
0	46770	80468	1.0	iOptik	IOPTIK, INC	8354 Bustleton Ave	8354 Bustleton Avenue
1	76725	37627	1.0	Mertz Auto Body	MERTZ AUTO BODY INC.	989 Gravois Rd	989 Gravois Rd
2	76725	38413	1.0	Mertz Auto Body	MERTZ AUTO BODY, INC.	989 Gravois Rd	989 GRAVOIS RD
3	43026	71911	1.0	Newtown Nails	NEWTOWN NAILS	29 Swamp Rd	29 Swamp Rd
4	43026	71912	1.0	Newtown Nails	NEWTOWN NAILS	29 Swamp Rd	29 Swamp RD
...	...	...	...	...	...	...	...
7998	63062	20386	0.8	Khonsari Law Group	KHONSARI LAW GROUP PLLC	150 2nd Ave N, 970,	150 Second Avenue 970
7999	87806	12753	0.8	SB Health and Beauty Spa	SB SPA LLC	116 S Oregon Ave	116 S Oregon Ave,
8000	8082	72910	0.8	Coco Blue Nail & Spa	COCO BLUE OLD CITY LLC	108 N 2nd St, Ste 102	108 N 2nd Street #102
8001	34501	69043	0.8	Vince's Gulf	VINCES SERVICE STATION	5430 Ridge Ave	5430 Ridge Ave
8002	34501	69037	0.8	Vince's Gulf	VINCE'S SERVICE STATION INC	5430 Ridge Ave	5430 RIDGE AVENUE

8003 rows × 7 columns

Summary

What we have learned so far:

1. Data cleaning and preparation
2. Data volume
3. Use Internet resources
4. New library and calculation methods
5. Validation and testing

Github

https://github.com/rockcyy/apan5210_team99_group_project