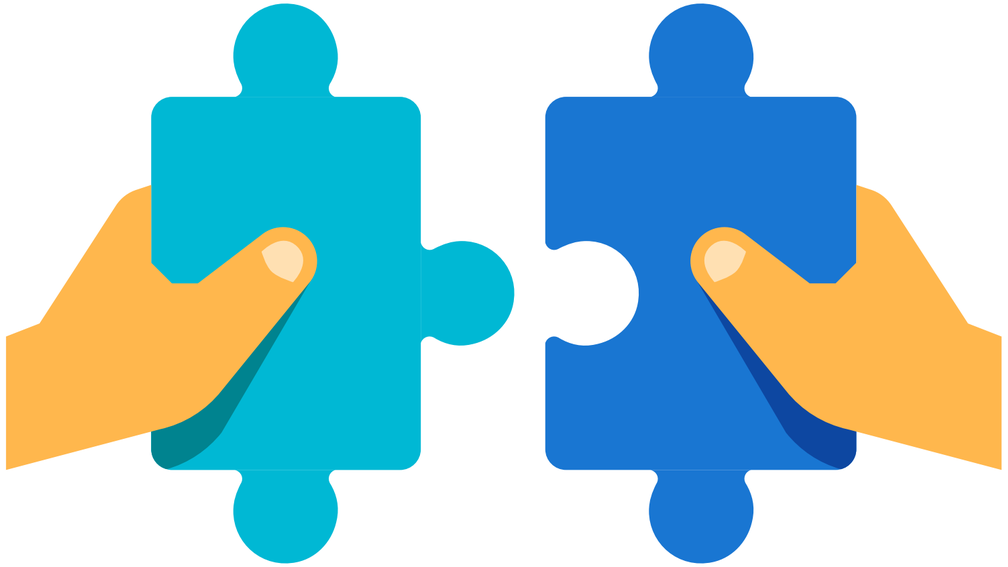

MATCHING ALGORITHM

PYTHON WARRIORS

DISCUSSION FLOW

1. EXPLORATORY DATA ANALYSIS
2. JOURNEY, ITERATIONS
3. FINAL ALGORITHM
4. DISCUSSION OF RESULTS
5. OVERALL LEARNING

NAS - COLUMNS - ADDRESSES

```
(98509, 7)
business_id     0
name            0
address         0
city            0
state           0
zip_code        0
size            0
dtype: int64
(94585, 7)
entity_id       0
name            0
address      2798
city            0
state           0
postal_code    37
categories     62
dtype: int64
```

```
business_id      int64
name            object
address         object
city            object
state           object
zip_code        object
size           float64
dtype: object
entity_id        int64
name            object
address         object
city            object
state           object
postal_code    float64
categories      object
dtype: object
```

address
2720 Broadway Center Blvd
2720 Broadway Center Blvd
8176 Woodland Center Blvd
2720 Broadway Center Blvd
8270 WOODLAND CENTER BLVD
...
501 Corporate Centre Drive
1107 BAPTIST WORLD CENTER DR.
501 Corporate Centre Drive
501 CORPORATE CENTRE DR Ste. 200
501 CORPORATE CENTRE DR STE 200

```
PA    32335
FL    28373
MO    15613
TN    11613
IN    10575
Name: state, dtype: int64
PA    34039
FL    26330
TN    12056
IN    11247
MO    10913
Name: state, dtype: int64
```

LIBRARIES USED: PANDAS, RE, FUZZY WUZZY

Left

	business_id	name	address	city	state	zip_code	size
0	1	SOURINI PAINTING INC.	12800 44th St N	Clearwater	FL	33762-4726	11.0
1	2	WOLFF DOLLA BILL LLC	1905 E 19th Ave	Tampa	FL	33605-2700	8.0
2	3	COMPREHENSIVE SURGERY CENTER, LLC	1988 GULF TO BAY BLVD, Ste 1	CLEARWATER	FL	33765-3550	8.0
3	4	FRANK & ADAM APPAREL LLC	13640 Wright Cir	Tampa	FL	33626-3030	12.0
4	5	MORENO PLUS TRANSPORT INC	8608 Huron Court unite 58	Tampa	FL	33614	8.0

punctuation

Right

redundant column

	entity_id	name	address	city	state	postal_code	categories
0	1	The UPS Store	87 Grasso Plaza Shopping Center	Affton	MO	63123.0	Shipping Centers, Local Services, Notaries, Ma...
1	2	St Honore Pastries	935 Race St	Philadelphia	PA	19107.0	Restaurants, Food, Bubble Tea, Coffee & Tea, B...
2	3	Perkiomen Valley Brewery	101 Walnut St	Green Lane	PA	18054.0	Brewpubs, Breweries, Food
3	4	Sonic Drive-In	615 S Main St	Ashland City	TN	37015.0	Burgers, Fast Food, Sandwiches, Food, Ice Crea...
4	5	Famous Footwear	8522 Eager Road, Dierbergs Brentwood Point	Brentwood	MO	63144.0	Sporting Goods, Fashion, Shoe Stores, Shopping...

uppercase

abbreviations

punctuation

capitalization

'.0' after each postal code

		Left Dataset	Right Dataset
Differences	**Columns**	The ‘size’ column is redundant	The ‘categories’ column is redundant
	Column Names	business_id zip_code	entity_id postal_code
	Data Types	The ‘zip_code’ column is object	The ‘postal_code’ column is float64
	Null Values	None	2798 missing values in the ‘address’ column 37 missing values in the ‘postal_code’ column
	Zip Codes	Contains 9-digit zip codes	‘.0’ after postal codes
Commonalities	**States**	Both have 5 states: PA, FL, MO, TN, IN	
	Abbreviations	Both contain various abbreviations in ‘name’, ‘address’, and ‘city’ columns	
	Letter Case	A mix of lowercase, uppercase, and capitalization	
	Spelling	Different spellings of the same words	
	Punctuations	Punctuations in ‘name’, ‘address’, and ‘city’ columns	

JOURNEY: DATA CLEANING

Left Dataset
Drop "size" column
Change "zip_code" column to 5-digit
Remove punctuation
Change all letters to lowercase

	business_id	name	address	city	state	zip_code
0	1	sourini painting inc	12800 44th st n	clearwater	FL	33762
1	2	wolff dolla bill llc	1905 e 19th ave	tampa	FL	33605
2	3	comprehensive surgery center llc	1988 gulf to bay blvd ste 1	clearwater	FL	33765
3	4	frank adam apparel llc	13640 wright cir	tampa	FL	33626
4	5	moreno plus transport inc	8608 huron court unite 58	tampa	FL	33614

Right Dataset
Drop "categories" column,
Convert "zip_code" column to string
Remove punctuation
Change all letters to lowercase

	entity_id	name	address	city	state	postal_code
0	1	the ups store	87 grasso plaza shopping center	affton	MO	63123
1	2	st honore pastries	935 race st	philadelphia	PA	19107
2	3	perkiomen valley brewery	101 walnut st	green lane	PA	18054
3	4	sonic drivein	615 s main st	ashland city	TN	37015
4	5	famous footwear	8522 eager road dierbergs brentwood point	brentwood	MO	63144

JOURNEY: USPS DICTIONARY

Left & Right Dataset

Iterate USPS Dictionary over "address" columns

address
12800 44th st n
1905 e 19th ave
1988 gulf to bay blvd ste 1
13640 wright cir
8608 huron court unite 58

address
12800 44th street north
1905 east 19th avenue
1988 gulf to bay boulevard suite 1
13640 wright circle
8608 huron court unite 58

```
abbrev_dict = {
    'aly': 'alley',
    'ave': 'avenue',
    'blvd': 'boulevard',
    'byp': 'bypass',
    'cir': 'circle',
    'ct': 'court',
    'dr': 'drive',
    'expy': 'expressway',
    'hwy': 'highway',
    'ln': 'lane',
    'pkwy': 'parkway',
    'pl': 'place',
    'pt': 'point',
    'rd': 'road',
    'sq': 'square',
    'st': 'street',
    'ter': 'terrace',
    'trl': 'trail',
    'ste':'suite',
    'e':'east',
    'w':'west',
    's':'south',
    'n':'north',
    'bldg':'building',
    'mlk':'martin luther king',
    'jfk':'john f kennedy',
    '1st':'first',
    '2nd':'second',
    '3rd':'third',
    '4th':'fourth',
    '5th':'fifth',
    '6th':'sixth',
    '7th':'seventh',
    '8th':'eighth',
    '9th':'ninth',
    '10th':'tenth'
}
```

JOURNEY: GROUPING

- Divide into five groups by "State"
- Merge left and right dataset on "address" by "inner"

	business_id	name_x	address	city_x	state_x	zip_code	entity_id	name_y	city_y	state_y	postal_code
0	54565	marvin e kanze inc	1395 lawrence road	havertown	PA	19083	90040	marvin e kanze	havertown	PA	19083
1	80058	wire to wire llc	1395 lawrence road	havertown	PA	19083	90040	marvin e kanze	havertown	PA	19083
2	54567	commerce dujour corporation	2001 market street	philadelphia	PA	19103	11524	fedex office print ship center	philadelphia	PA	19103
3	54567	commerce dujour corporation	2001 market street	philadelphia	PA	19103	65559	panache hair design	philadelphia	PA	19103
4	54567	commerce dujour corporation	2001 market street	philadelphia	PA	19103	72850	paganos market and bar	philadelphia	PA	19103

FINAL ALGORITHM

Partial Ratio from FuzzyWuzzy python package

- **Fuzzywuzzy is a popular library to perform string comparison and matching**
- **The partial ratio feature calculates the ratio of the length of the longest common substring to the length of the shorter string. Useful when dealing with data contain:**
  - **typos**
  - **misspellings**
  - **some strings only appear in the longer piece of text**

LAYERS OF MATCHING

1. Split each dataset based on “state”
2. Merged left and right dataset based on “address”
3. Compute the confidence score based on “name” → The output of the score gives 92

left_df:

business_id	name	address	city	state	zip_code
54613	mayfair diner restaurant inc	7373 frankford avenue	philadelphia	PA	19136

right_df:

entity_id	name	address	city	state	postal_code
50438	mayfair diner	7373 frankford avenue	philadelphia	PA	19136

RESULTS

- There are 7164 matching results with confidence score greater than 80 out of 100
- There are totally 3576 pairs of matching records that are duplicates. This is manifested as one entity_id with multiple business_id high matches or vice versa.

	business_id	entity_id	confidence_score	name_x	address_x	name_y	address_y
0	54613	50438	100	mayfair diner restaurant inc	7373 frankford avenue	mayfair diner	7373 frankford avenue
1	57615	50438	100	mayfair diner restaurant inc	7373 frankford avenue	mayfair diner	7373 frankford avenue
2	54639	62088	100	vk collegeville diner	290 east main street	collegeville diner	290 east main street
3	55133	62088	100	vk collegeville diner	290 east main street	collegeville diner	290 east main street
4	54663	66731	100	reanimator coffee roasters llc	310 west master street	reanimator coffee	310 west master street
...	...	...	...	...	...	...	...
3571	37928	50745	100	saltire games inc	11723 pendleton pike	saltire games	11723 pendleton pike
3572	38292	2041	100	mimosa and a masterpiece llc	614 massachusetts avenue suite b	mimosa and a masterpiece	614 massachusetts avenue suite b
3573	38646	2041	100	mimosa and a masterpiece llc	614 massachusetts avenue suite b	mimosa and a masterpiece	614 massachusetts avenue suite b
3574	38808	55812	100	coco nails spa llc	5868 east 71st street suite k	coco nails	5868 east 71st street suite k
3575	38837	55812	100	coco nails and spa inc	5868 east 71st street suite k	coco nails	5868 east 71st street suite k

different business_id with same information

	business_id	entity_id	confidence_score
0	54565	90040	100
1	54597	72850	95
2	54573	62113	100
3	54576	21969	100
4	54613	50438	100
...	...	...	...
7159	38842	13456	100
7160	38862	48598	100
7161	38871	84891	100
7162	38882	73946	100
7163	38884	47838	100

GitHub

https://github.com/ccccliu919/5210Project.git
---Liying Liu

https://github.com/avporciuncula/APAN5210-Python/tree/main

SUMMARY

We use FuzzyWuzzy Partial Ratio from python package to match the two datasets and get 7164 matching results

Improved accuracy: Fuzzy matching algorithms can identify and correct spelling mistakes, typos, and other errors that can lead to inaccurate matches.

Efficiency: Fuzzy matching can be used to quickly and efficiently match large datasets, without the need for manual review and correction

Flexibility: Fuzzy matching algorithms can be configured to allow for a certain degree of variation in the data being matched, which can be useful for matching addresses that are entered in different formats or with different levels of detail.