Finding Details in the City

Data Engineering Track

Group 3: Steven Anthony, Kaleb Decker,
Maddie Haughton, Em Needles, Dena Wetmore

Searching API

- User Input
- How to use
- What you can expect to get out of this search

Geoapify Docs:
https://apidocs.geoapify.com/docs/places/#about

Dictionary of categories to search for & parameters

Build URL string

Get all categories of a specific place that includes the search parameters

URL Building

```python
# Create dictionary to house each category that will be captured
category_dict = {'activity' : activity, 'commercial' : commercial, 'commerical.catering' : commercial_catering, 'accomodation' : accomodations, \
                 'entertainment' : entertainment, 'leisure': leisure, 'parking' : parking, 'wheelchair.yes' : wheelchair}

# Set up string to input categories to add to base url
string_of_categories = ""

# Capture the key for each category we want to include in the search
for k,v in category_dict.items():
    if v:
        string_of_categories = string_of_categories + k + ','

# Remove last comma
string_of_categories = string_of_categories[:-1]

# API string built for flexibility
url2 = f"https://api.geoapify.com/v2/places?categories={string_of_categories}&filter={location_filter}&limit={limit}&offset={offset}&apiKey={apikey}"
```

Assigning Data to Dataframes

- What are the goals of the things we chose to pull from the json response?
- Why the separate tables?

Normalize json response

Different goals for each table

Troubleshooting

- Initial raw dataframe

	Property	Address2	Place_ID
0	Boyle Heights Senior Citizen Center	2839 East 3rd Street, Los Angeles, CA 90023, U...	51e64ee384248d5dc059b8917ebfff044140f00102f901...
1	Subliminal Projects	1331 Sunset Boulevard, Los Angeles, CA 90026, ...	516256f42609905dc059023122d13f094140f00102f901...
2	Mush	1617 Silver Lake Boulevard, Los Angeles, CA 90...	51cd2231343c915dc05987726c74540b4140f00102f901...
3	Associated of Los Angeles	2585 East Olympic Boulevard, Los Angeles, CA 9...	51edd3ca0b3c8e5dc059b4a1c92051034140f00102f901...
4	The Broad Museum	221 South Grand Avenue, Los Angeles, CA 90012,...	51da90e9b909905dc0590f09aa73f8064140f00102f901...

- Clean dataframe - allows the user to better sort and gather information needed for a query

	Place_ID	Property	Address	City	State	Zip	Country
0	51e64ee384248d5dc059b8917ebfff044140f00102f901...	Boyle Heights Senior Citizen Center	2839 East 3rd Street	Los Angeles	CA	90023	United States of America
1	516256f42609905dc059023122d13f094140f00102f901...	Subliminal Projects	1331 Sunset Boulevard	Los Angeles	CA	90026	United States of America
2	51cd2231343c915dc05987726c74540b4140f00102f901...	Mush	1617 Silver Lake Boulevard	Los Angeles	CA	90026	United States of America
3	51edd3ca0b3c8e5dc059b4a1c92051034140f00102f901...	Associated of Los Angeles	2585 East Olympic Boulevard	Los Angeles	CA	90021	United States of America
4	51da90e9b909905dc0590f09aa73f8064140f00102f901...	The Broad Museum	221 South Grand Avenue	Los Angeles	CA	90012	United States of America

- Second dataframe created linking place_id. This dataframe is necessary to hold all of the data from each unique place_id. Allows the user to pull from a destination and then use the unique place id to search through categories.

	Place_ID	Category	Index
0	51e64ee384248d5dc059b8917ebfff044140f00102f901...	activity	1
1	51e64ee384248d5dc059b8917ebfff044140f00102f901...	activity.community_center	2
2	51e64ee384248d5dc059b8917ebfff044140f00102f901...	building	3
3	51e64ee384248d5dc059b8917ebfff044140f00102f901...	building.public_and_civil	4
4	516256f42609905dc059023122d13f094140f00102f901...	building	5

Pre-Database Data

Transform & Export

Used Polars for speed & clarity

Wouldn't present data in this form to customer

Output CSV files based on city input

Preparing the Data for SQL

shape: (423, 7)

Place_ID	Property	Address	City	State	Zip	Country
str	str	str	str	str	str	str
"51512b54fd976c...	"Great Lakes Sc...	"601 Erieside A...	" Cleveland"	"OH"	"44114"	" United States...
"51b2dc88dd2d6c...	"International ...	"1501 North Mar...	" Cleveland"	"OH"	"44114"	" United States...
"5164ffa097066c...	"A Christmas St...	"1103 Rowley Av...	" Cleveland"	"OH"	"44109"	" United States...
"5118963fdf966d...	"Frederich Bicy...	"3800 Lorain Av...	" Cleveland"	"OH"	"44113"	" United States...
"5151fb08579b6a...	"Graffiti Caps"	"3200 Carnegie ...	" Cleveland"	"OH"	"44115"	" United States...
"51fa4800ed7d6c...	"Leavened"	"1633 Auburn Av...	" Cleveland"	"OH"	"44113"	" United States...
"514459f8fada6f...	"Christophier C...	"9308 Clifton B...	" Cleveland"	"OH"	"44102"	" United States...
"5166b418a7826c...	"Rock and Roll ...	"East 9th Stree...	" Cleveland"	"OH"	"44114"	" United States...
"516249b9fb1c6c...	"Cleveland Hung...	"Selzer Way"	" Cleveland"	"OH"	"44114"	" United States...
"51b6ea31aa8a6f...	"Family Dollar"	"8484 Madison A...	" Cleveland"	"OH"	"44102"	" United States...
"51f50803567a6a...	"Mawi Wireless"	"3507 Payne Ave...	" Cleveland"	"OH"	"44114"	" United States...
"51044d7ae63c6a...	"Children's Mus...	"3813 Euclid Av...	" Cleveland"	"OH"	"44115"	" United States...
...	...	...	...	...	...	...
"51d65647e92d6d...	"Riverview Welc...	"1701 West 25th...	" Cleveland"	"OH"	"44113"	" United States...
"51a01fa1c1016a...	"Rebuilders Exc...	"East 53rd Stre...	" Cleveland"	"OH"	"44114"	" United States...
"519d6bf35a646c...	"Darlafoxx"	"50 South Roadw...	" Cleveland"	"OH"	"44113"	" United States...

shape: (1_619, 3)

Index	Place_ID	Category
i64	str	str
1	"51512b54fd976c...	"building"
2	"51512b54fd976c...	"building.touri...
3	"51512b54fd976c...	"entertainment"
4	"51512b54fd976c...	"entertainment....
5	"51512b54fd976c...	"internet_acces...
6	"51512b54fd976c...	"wheelchair"
7	"51512b54fd976c...	"wheelchair.yes...
8	"51b2dc88dd2d6c...	"entertainment"
9	"51b2dc88dd2d6c...	"entertainment....
10	"51b2dc88dd2d6c...	"no_fee"
11	"51b2dc88dd2d6c...	"no_fee.no"
12	"5164ffa097066c...	"building"
...	...	...
1608	"5186b7bd93056c...	"service"
1609	"5186b7bd93056c...	"service.beauty...
1610	"5186b7bd93056c...	"service.beauty...

Importing Data to PostgreSQL

- Why this database?
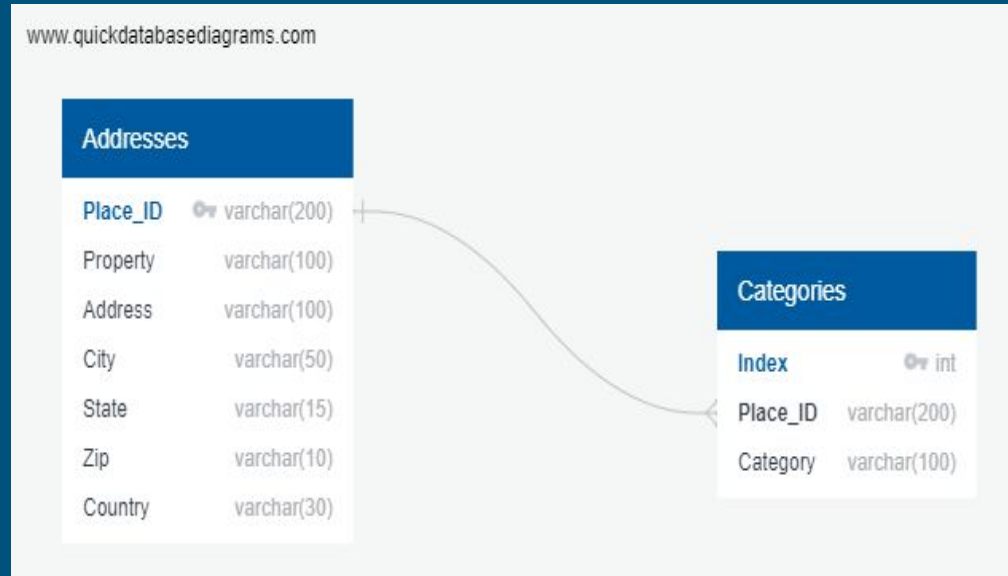- How can we use this moving forward?

ERD and schema to create tables

Import data to store in db and later query

How Data is Connected in the Database

- Place ID is distinct for each property (primary key) and exists in the Categories table as a foreign key
- In the categories table, each category is separated so Place ID is repeated (one to many relationship)
- The categories table was given an index to serve as the primary key
- Almost all columns in both tables contained data that was a mix between letter characters and numbers or just letter characters on their own - all datatypes are varchar types of various lengths except for the index of the categories table, which is an integer type

www.quickdatabasediagrams.com

Addresses

Place_ID	Oᵥ varchar(200)
Property	varchar(100)
Address	varchar(100)
City	varchar(50)
State	varchar(15)
Zip	varchar(10)
Country	varchar(30)

Categories

Index	Oᵥ int
Place_ID	varchar(200)
Category	varchar(100)

Conclusions

- What will be the outcome of our tool?
- Who could use it and why?

Purpose of dataset and user input options.

Is the use of this data ethical?

What are some other data sets that could be used in a similar application?

The End

Thanks for Watching