Housing Dashboard (DB Ver)

By Raymond Song, Alan Zhang

Why Housing Dashboard?

- You move to a new area and start looking for housing, you want to know how the listing compares to others
- For similar listings in your area, is the listing more expensive or less expensive
- Does it offer the amenities that you want?
- Is the surrounding area convenient for living (Groceries, Transportation)?

Databases

MongoDB:

- Best solution as we want to aggregate listings from multiple sources
  - Zillow, Apartment.com, Craigslist, listings from property websites
  - Listings gathered from the San Diego County
- Store housing listings for easier aggregates.
- Individual houses are not related through relationships

Databases (Cont.)

Neo4j:

- Spatial package provides geographic search.
- Houses are not stored in Neo4j, only passed along to neo4j as a tuple pair (lat, long) to find nearby businesses
- Easier querying between Businesses and Categories as well as bus stops and routes

Prepping the raw data

- Collected using a combination of web scraping and API
- Unify features have different column names
  - E.g. num_baths instead of baths
  - Convert all of them to standard key
- Ensure that data such as “$4,600+” gets stored as 4600 so it can be queried
- Try to remove anomalies from the data
- Tried to gather all amenities listed as its own column into one column named ‘amenities’

```
garden = pd.read_csv('../data/housing/raw/garden_community-2022-07-05.csv')
garden.head()
```

[60] ✓ 0.1s Python

...

ame	community_address	community_mobile	overview_text	floor_plan_name	bath_num	bed_num	price_info	area_info	...	Upgraded Residence Amenities	Appliances
ıxury ıents	9065 Sydney Court, San Diego, CA	(858) 358-7681	Up to First Month Free!* *All prices and spec...	Compass	1 Bath	Studio	$2,725 - $3,225	662 - 743 sq.ft.	...	NaN	NaN
ıxury ıents	9065 Sydney Court, San Diego, CA	(858) 358-7681	Up to First Month Free!* *All prices and spec...	Escape	1 Bath	1 Bedroom	$3,135 - $3,610	874 sq.ft.	...	NaN	NaN
ıxury ıents	9065 Sydney Court, San Diego, CA	(858) 358-7681	Up to First Month Free!* *All prices and spec...	Escape with Loft	1 Bath	1 Bedroom	$3,610 - $3,960	1016 - 1105 sq.ft.	...	NaN	NaN
ıxury ıents	9065 Sydney Court, San Diego, CA	(858) 358-7681	Up to First Month Free!* *All prices and spec...	Evolve	1 Bath	1 Bedroom	$3,135 - $3,410	880 - 909 sq.ft.	...	NaN	NaN
ıxury ıents	9065 Sydney Court, San Diego, CA	(858) 358-7681	Up to First Month Free!* *All prices and spec...	Evolve with Loft	1 Bath	1 Bedroom	$3,610	1028 sq.ft.	...	NaN	NaN

Size	AvailabilityRent	Monthly Fees	One Time Fees	Transportation	Education	Pet Policy	Distance	Duration	Parking	Amenities
['Studio, $2,320 – $2,415, Studio, ,, 1 bath, ...	$2320-3610	NaN	* Application Fee: $25\n* Cat Deposit: $250	Transit / Subway, Distance, Transit / Subway, ...	Colleges & Universities, Distance, Colleges & ...	Cats Allowed, Restrictions:, $250 deposit for ...	NaN	NaN	Restrictions:, $250 deposit for one cat. $400 ...	Community Amenities, Pool, Fitness Center, Lau...
['Florentine, $2,445 – $2,595, 1 bed, ,, 1 bat...	$2445-3655	NaN	* Application Fee: $25\n* Cat Deposit: $250	Transit / Subway, Distance, Transit / Subway, ...	Colleges & Universities, Distance, Colleges & ...	Cats Allowed, Restrictions:, $250 deposit for ...	NaN	NaN	Restrictions:, $250 deposit for 1 cat. $400 de...	Community Amenities, Pool, Fitness Center, Lau...
['Soledad, $2,475 – $2,725, Studio, ,, 1 bath,...	$2475-4950	NaN	* Application Fee: $46	Transit / Subway, Distance, Transit / Subway, ...	Colleges & Universities, Distance, Colleges & ...	Parking, Garage, 1 space, Assigned Parking, On...	NaN	NaN	Garage, 1 space, Assigned Parking, One Bedroom...	Community Amenities, Pool, Fitness Center, Ele...
['1C, $3,665 – $5,650, 1 bed, ,, 1 bath, ,, 90...	$3665-8710	NaN	* Application Fee: $47	Transit / Subway, Distance, Transit / Subway, ...	Colleges & Universities, Distance, Colleges & ...	Other Fees, Application Fee, $47	NaN	NaN	NaN	Community Amenities, Pool, Fitness Center, Lau...
['One Bedroom / One Bath, $2,875 – $3,325, 1 b...	$2875-4350	NaN	* Application Fee: $49\n* Cat Deposit: $500\n*	Transit / Subway, Distance, Transit / Subway, ...	Colleges & Universities, Distance, Colleges & ...	Dogs Allowed, Restrictions:, We welcome 2 pets...	NaN	NaN	Restrictions:, We welcome 2 pets per apartment...	Community Amenities, Pool, Fitness Center, Clu...

```
1  zillow = pd.read_csv('../data/housing/raw/zillow-2022-07-05.csv')
2  zillow.head()
```

✓ 0.0s Python

	statusText	detailUrl	latLong	units	...	isUserClaimingOwner	isUserConfirmedClaim	pgapt	sgapt	zestima
.	Solazzo Apartments Homes	/b/solazzo-apartments-homes-la-jolla-ca-5Xk5YK/	{'latitude': 32.86568, 'longitude': -117.235725}	[{'price': '$3,165+', 'beds': '1'}]	...	NaN	NaN	NaN	NaN	Na
.	Ocean House on Prospect Apartment Homes	/b/ocean-house-on-prospect-apartment-homes-la-...	{'latitude': 32.841976, 'longitude': -117.27955}	[{'price': '$5,340+', 'beds': '1'}, {'price': ...	...	NaN	NaN	NaN	NaN	Na
.	Apartment for rent	https://www.zillow.com/homedetails/7623-Eads-A...	{'latitude': 32.842457, 'longitude': -117.27549}	NaN	...	False	False	ForRent	For Rent	Na
.	For Rent	/b/La-Jolla-CA/32.841022,-117.275635_ll/	{'latitude': 32.841022, 'longitude': -117.275635}	[{'price': '$5,975+', 'beds': '3'}]	...	NaN	NaN	NaN	NaN	Na
.	Apartment for rent	https://www.zillow.com/homedetails/2205-Camini...	{'latitude': 32.81866, 'longitude': -117.234184}	NaN	...	False	False	ForRent	For Rent	1304800

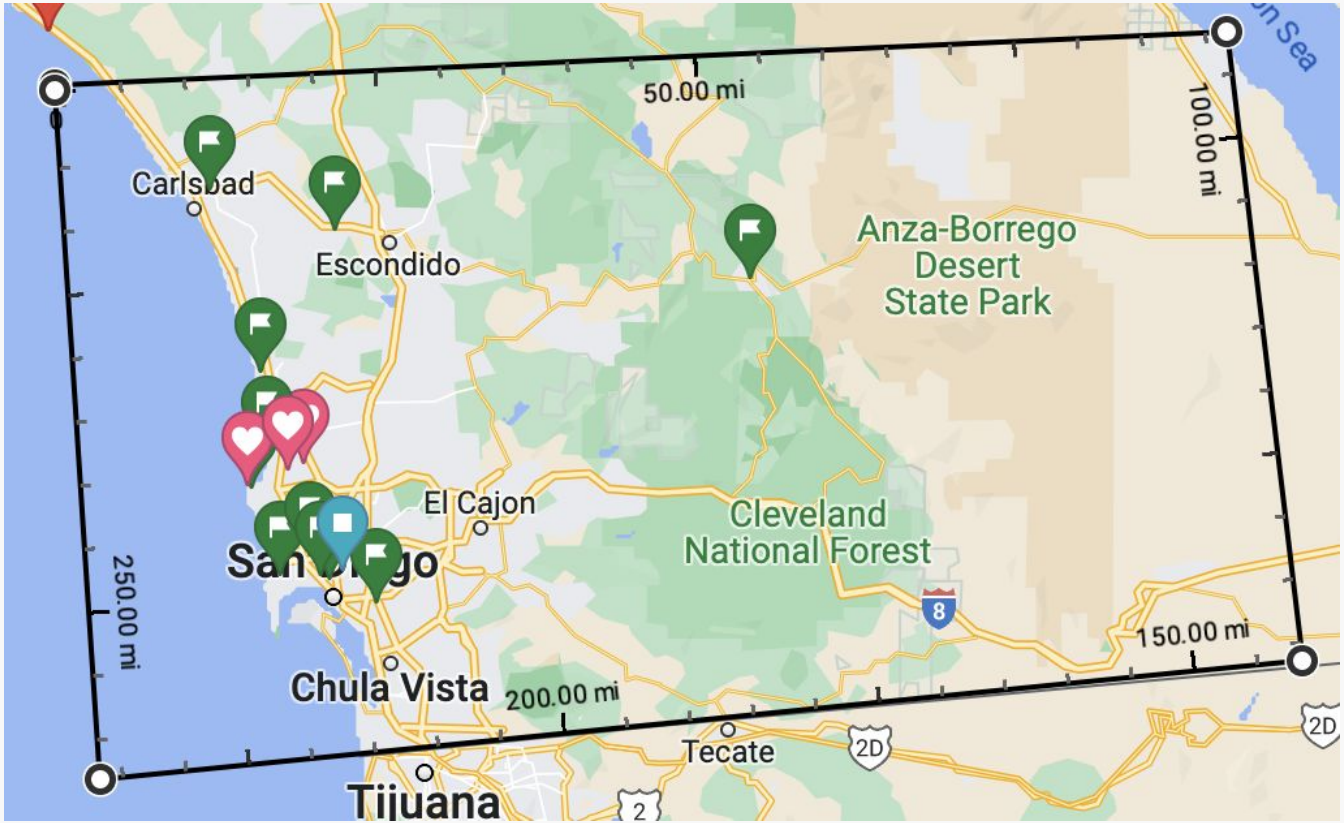
- Need to extract latitude and longitude from the dictionary and get the price and beds for properties that have multiple listings

url	post_id	post_date	description	num_beds	num_baths	price	address	sqft	features	scrape_date
https://sandiego.craig	7495575791	2022-06-12T09:15:4	SAN DIEGO, MISSIC x 187OR Text 187 t to text our team dire	2	2	4125		-1	['air conditioning', 'c	2022-07-05 0:11:14
https://sandiego.craig	7498454629	2022-06-19T06:55:4	g 2 rooms one has a	2	shared	1400	6634 alcala knolls dr	-1	['house', 'laundry on	2022-07-05 0:11:17
https://sandiego.craig	7500408184	2022-06-23T16:22:2	all Now - show conta This is a 3 Bedroom	3	2	4850	5040 Camino San Fe	1205	['EV charging', 'air c	2022-07-05 0:11:22
https://sandiego.craig	7493144309	2022-06-06T13:29:2	Apartment Living Yo x 99OR Text 99 to to text our leasing te	2	2	3372	2185 Station Village	-1	['apartment', 'w/d in u	2022-07-05 0:11:24
https://sandiego.craig	7498345212	2022-06-18T17:14:4	Apex Mission Valley Residents are always Join us! Call today a	2	2	3500		-1	['EV charging', 'air c	2022-07-05 0:11:27
https://sandiego.craig	7503686488	2022-07-01T14:48:4	San Diego's most po -Large Closets -Pet-Friendly Commu -Centralized Locatio -Short Term Lease -Washer/Dryer -Select Homes Belov -Private Balcony -Wheelchair Accessi -Sundeck -Street Parking -After-Hours Courtesy	2	2	3570	5395 Napa St	-1	['EV charging', 'air c	2022-07-05 0:11:29
https://sandiego.craig	7492691615	2022-06-05T12:30:3	Society, managed b 610 Del Sol San Diego, CA 9210 show contact info The Society offers yo This two bedroom tw VIEW OUR WEBSITE http://www.hollandre [#9253505]	2	2	4000	610 Del Sol	-1	['EV charging', 'air c	2022-07-05 0:11:31

CA Businesses

- JSON data gathered from the SDSC database and cleaned
- Filtered down to SD businesses as businesses outside of the San Diego County will not have neighboring nodes
    - Save down on resources (for the scope of this project)

```python
with open('../data/ca_business.json', 'r') as f:
    data = json.load(f)

sd_businesses = []
for line in tqdm(data):
    if type(line) == dict:
        lat, lng = line['latitude'], line['longitude']
        if lat < 33.363442 and lat > 32.534317 and lng > -117.547484 and lng < -115.974969:
            if '_id' in line:
                del line['_id']
            sd_businesses.append(line)

with open('../data/sd_business.json', 'w') as f:
    json.dump(sd_businesses, f)
```


50.00 mi
100.00 mi
150.00 mi
200.00 mi
250.00 mi
Carlsbad
Escondido
Anza-Borrego
Desert
State Park
El Cajon
Cleveland
National Forest
Chula Vista
Tecate
Tijuana
8
2D
2

SD Business (Open)

Find out whether the business state is ‘permanently closed’ and remove it from the database

```python
# Find businesses with cafe or restaurant in the category
closed_pattern = re.compile(r"(?i)\b(?:permanently closed)\b")
closed_filer = {"state": closed_pattern}

# Define the update operation
update_closed = {"$set": {"closed": True}}

result_closed = collection.update_many(closed_filer, update_closed)
print("Businesses marked as closed:", result_closed.modified_count)

# Define the filter condition for the remaining businesses
remaining_filter = {"closed": {"$exists": False}}

# Define the update operation to set the remaining businesses as not closed
update_remaining = {"$set": {"closed": False}}

# Update the remaining businesses to set closed as false
result_remaining = collection.update_many(remaining_filter, update_remaining)
print("Remaining businesses marked as not closed:", result_remaining.modified_count)
```

✓ 0.8s

```
Businesses marked as closed: 3251
Remaining businesses marked as not closed: 43141
```

Cafes/Restaurants

```python
# Find businesses with cafe or restaurant in the category
food_pattern = re.compile(r"(?i)\b(?:cafe|restaurant)\b")
food_filter = {"category": food_pattern}

# Define the update operation
update_food = {"$set": {"food": True}}

# Update documents matching the filter condition
result = collection.update_many(food_filter, update_food)

# Print the number of documents updated
print("Food businesses labeled updated:", result.modified_count)

# Define the filter condition for the remaining businesses
remaining_filter = {"food": {"$exists": False}}

# Define the update operation to set the remaining businesses as not closed
update_remaining = {"$set": {"food": False}}

# Update documents matching the filter condition
result = collection.update_many(remaining_filter, update_remaining)

# Print the number of documents updated
print("Non Food businesses labeled updated:", result.modified_count)
```

✓ 1.3s

```
Food businesses labeled updated: 8550
Non Food businesses labeled updated: 37842
```

Grocery Stores

```python
# Find businesses with grocery in the category
grocery_pattern = re.compile(r"(?i)\b(?:grocery|food store)\b")
grocery_filter = {"category": grocery_pattern}

# Define the update operation
update_grocery = {"$set": {"grocery": True}}

# Update documents matching the filter condition
result = collection.update_many(grocery_filter, update_grocery)

# Print the number of documents updated
print("Grocery/Convenience updated:", result.modified_count)

remaining_filter = {"grocery": {"$exists": False}}

update_remaining = {"$set": {"grocery": False}}

res = collection.update_many(remaining_filter, update_remaining)

# Print the number of documents updated
print("Non Grocery/Convenience updated:", res.modified_count)
```

✓ 0.9s

```
Grocery/Convenience updated: 887
Non Grocery/Convenience updated: 45505
```

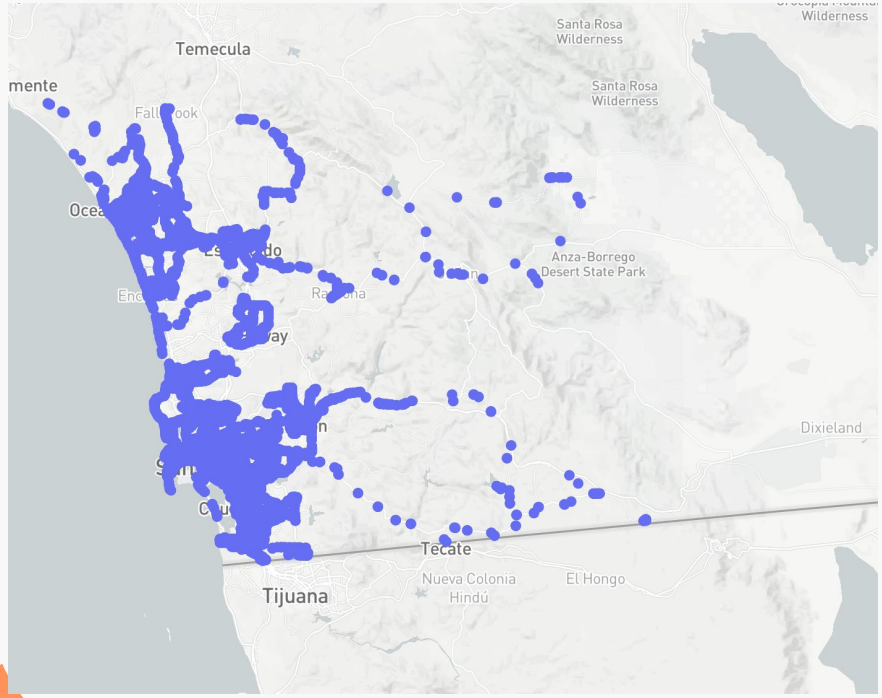
Schools

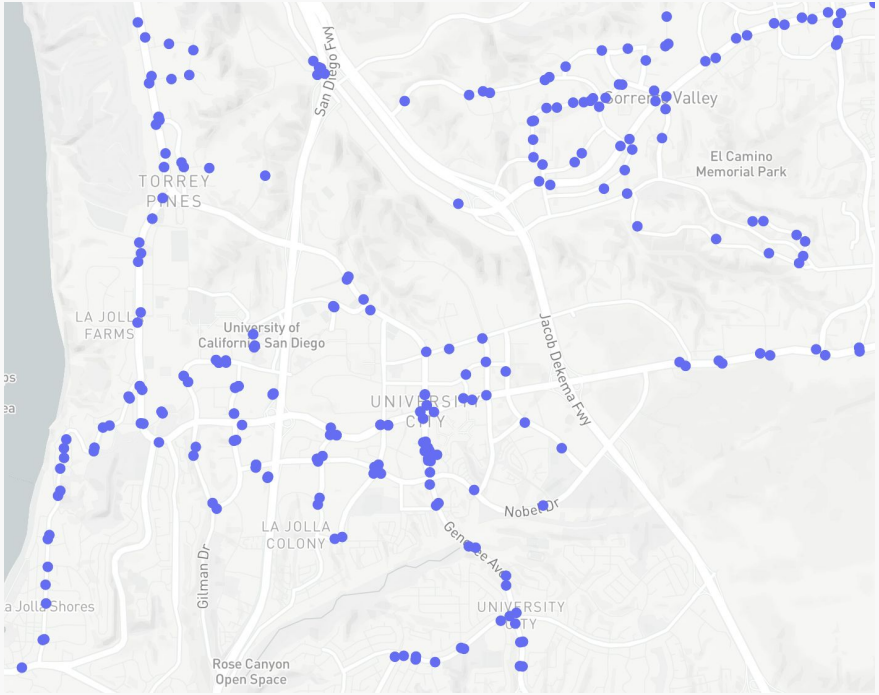
```python
# Find businesses with grocery in the category
school_pattern = re.compile(
    r"(?i)^(?:after school|charter school|education|elementary school|middle school|private university|primary school|preschool|university)$"
)
school_filter = {"category": school_pattern}

# Define the update operation
update_school = {"$set": {"school": True}}

# Update documents matching the filter condition
result = collection.update_many(school_filter, update_school)

# Print the number of documents updated
print("Schools updated:", result.modified_count)

remaining_filter = {"school": {"$exists": False}}

update_remaining = {"$set": {"school": False}}

res = collection.update_many(remaining_filter, update_remaining)

# Print the number of documents updated
print("Non School updated:", res.modified_count)
```

✓ 1.4s

```
Schools updated: 192
Non School updated: 46200
```

Bus stops


Temecula
Santa Rosa Wilderness
Santa Rosa Wilderness
Anza-Borrego Desert State Park
Dixieland
Tecate
Nueva Colonia Hindú
El Hongo
Tijuana


San Diego Fwy
Correa Valley
El Camino Memorial Park
TORREY PINES
LA JOLLA FARMS
University of California San Diego
Jacob Dekema Fwy
LA JOLLA COLONY
Gilman Dr
Genesee Ave
UNIVERSITY
Rose Canyon Open Space

Filtering the listings

- Price : 2400 - 4000, Sqft >= 850, In-Unit Washer Dryer, Allows pet

```python
# Find listing that has a price between 2400 - 4000 a month, In-Unit Washer Dryer and Allows pet

pipeline = [
    {"$match": {
        "price": {"$gte": 2400, "$lte": 4000},
        "sqft": {"$gte": 850},
        "$or": [
            {"amenities": {"$regex": ".*In-Unit Washer & Dryer.*"}},
            {'laundry': {'$regex': '.*w/d in unit*'}}
        ],
        "$or": [
            {'pets': {'$regex': '.*cats ok*'}},
            {'Pet Policy': {'$regex': '.*Cats Allowed*'}}
        ]
    }},
]

filtered_res = housing_collection.aggregate(pipeline)
for res in filtered_res:
    print(res)
```

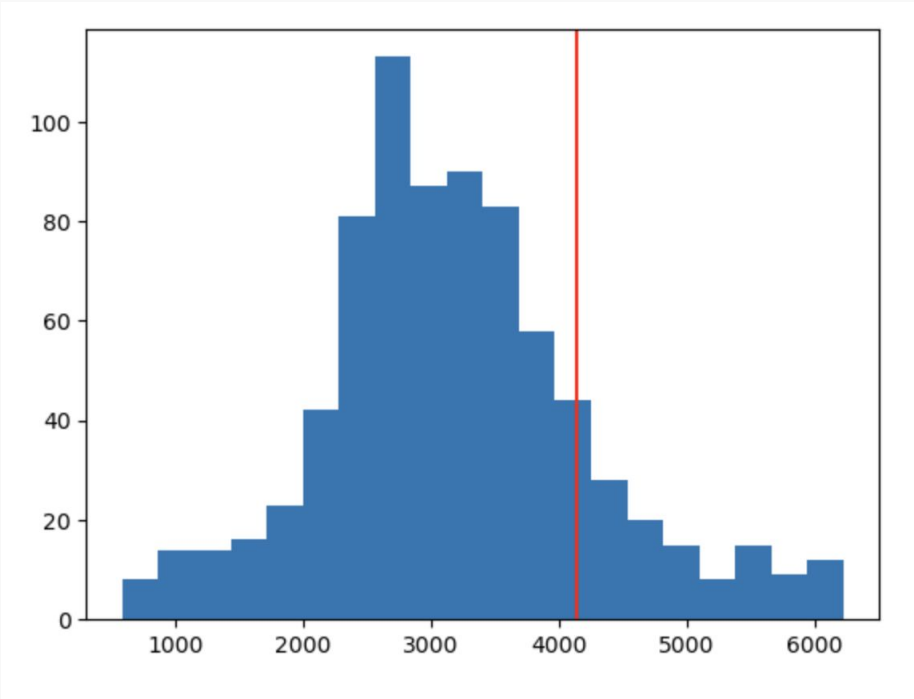
[12] ✓ 0.1s Python

```
{'_id': ObjectId('648baffb41b4b5c44bc0c309'), 'beds': 2.0, 'baths': 2.0, 'price': 3135.0, 'sqft': 1004.0, 'name': 'La Jolla International G
{'_id': ObjectId('648baffb41b4b5c44bc0c30a'), 'beds': 2.0, 'baths': 2.0, 'price': 3357.5, 'sqft': 1215.0, 'name': 'La Jolla International G
{'_id': ObjectId('648baffb41b4b5c44bc0c30c'), 'beds': 2.0, 'baths': 2.0, 'price': 3135.0, 'sqft': 1004.0, 'name': 'La Jolla International G
{'_id': ObjectId('648baffb41b4b5c44bc0c30d'), 'beds': 2.0, 'baths': 2.0, 'price': 3357.5, 'sqft': 1215.0, 'name': 'La Jolla International G
{'_id': ObjectId('648baffb41b4b5c44bc0c340'), 'beds': 2.0, 'baths': 2.0, 'price': 3260.0, 'sqft': 1010.0, 'name': 'La Scala Apartments', 'C
{'_id': ObjectId('648baffb41b4b5c44bc0c341'), 'beds': 2.0, 'baths': 2.0, 'price': 3290.0, 'sqft': 969.0, 'name': 'La Scala Apartments', 'Cc
{'_id': ObjectId('648baffb41b4b5c44bc0c344'), 'beds': 2.0, 'baths': 2.0, 'price': 3260.0, 'sqft': 1010.0, 'name': 'La Scala Apartments', 'C
{'_id': ObjectId('648baffb41b4b5c44bc0c345'), 'beds': 2.0, 'baths': 2.0, 'price': 3290.0, 'sqft': 969.0, 'name': 'La Scala Apartments', 'Cc
```

Pricing Distribution Query + Graph

```python
# Perform the aggregation pipeline
pipeline = [
    {"$group": {"_id": None, "prices": {"$push": "$price"}}},
    {"$project": {"_id": 0, "prices": 1}},
    {"$unwind": "$prices"},
    {"$group": {"_id": "$prices", "count": {"$sum": 1}}},
    {"$sort": {"_id": 1}}
]

result = list(housing_collection.aggregate(pipeline))

prices = [item['_id'] for item in result if item['_id'] is not None]
prices = [
    int(re.match(r"\$(\d[\d,]*)", price).group(1).replace(',', ''))
    if isinstance(price, str) else int(price) for price in prices
]
```

✓ 0.0s

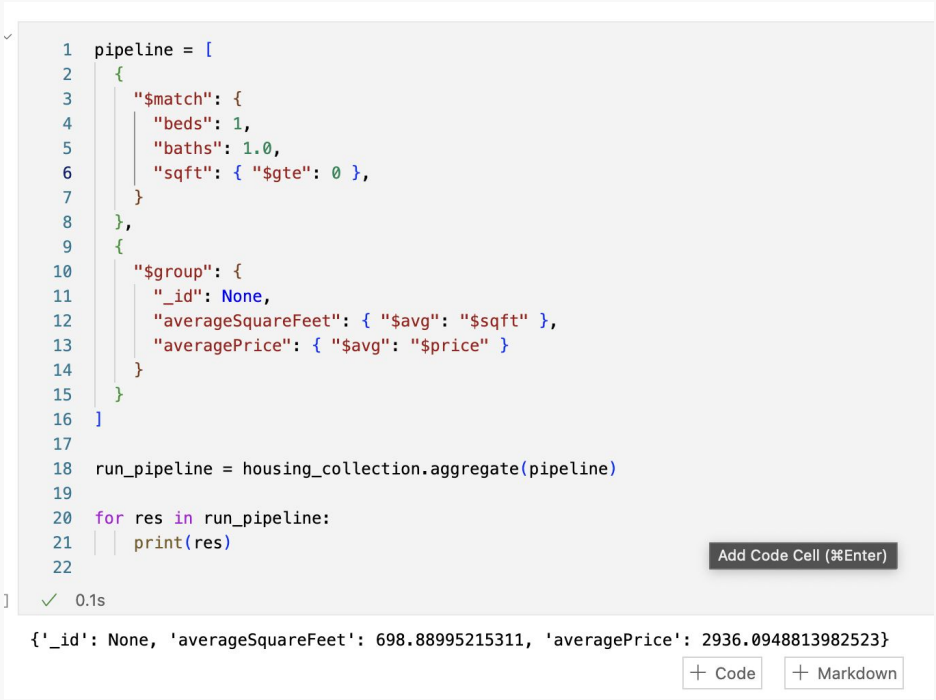


Filter out outliers using the interquartile range method

Speedup improvement

- Save the median, median, and standard deviation in a Redis database so that we can quickly draw up the gaussian distribution without querying all of the data and save one for each postal code.

Average SQFT and Price



```
pipeline = [
  {
    "$match": {
      "beds": 1,
      "baths": 1.0,
      "sqft": { "$gte": 0 },
    }
  },
  {
    "$group": {
      "_id": None,
      "averageSquareFeet": { "$avg": "$sqft" },
      "averagePrice": { "$avg": "$price" }
    }
  }
]

run_pipeline = housing_collection.aggregate(pipeline)

for res in run_pipeline:
    print(res)
```

Add Code Cell (⌘Enter)

0.1s

```
{'_id': None, 'averageSquareFeet': 698.88995215311, 'averagePrice': 2936.0948813982523}
```

+ Code + Markdown

- About 700 square feet on average and $2,937 per month
  (Price accurate as to July 2022)

Average SQFT and Price

```
pipeline = [
  {
    "$match": {
      "beds": 2,
      "baths": 2.0,
      "sqft": { "$gte": 0 },
    }
  },
  {
    "$group": {
      "_id": None,
      "averageSquareFeet": { "$avg": "$sqft" },
      "averagePrice": { "$avg": "$price" }
    }
  }
]

run_pipeline = housing_collection.aggregate(pipeline)

for res in run_pipeline:
    print(res)

```

✓ 0.0s

```
{'_id': None, 'averageSquareFeet': 1234.6165413533834, 'averagePrice': 4136.829201101928}
```

- About 1230 square feet on average and $4,136 per month
  (Price accurate as to July 2022)

Neo4j

Businesses and Transit

- Create Business and Category nodes
- Add Businesses to Spatial Index
- Similar to creating transit stops

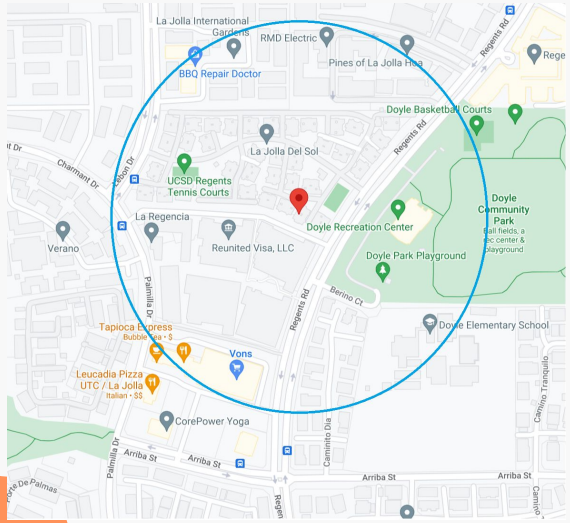
```
load_business_query = """CALL apoc.load.json("file:///active_business.json")
YIELD value
MERGE (b:Business {name: value.gmap_id})
SET b.name = value.name
SET b.latitude = value.latitude
SET b.longitude = value.longitude
WITH b, value
UNWIND value.category AS cat
MERGE (c:Category {name: toLower(cat)})
MERGE (b)-[:IS_IN]->(c);
"""
```

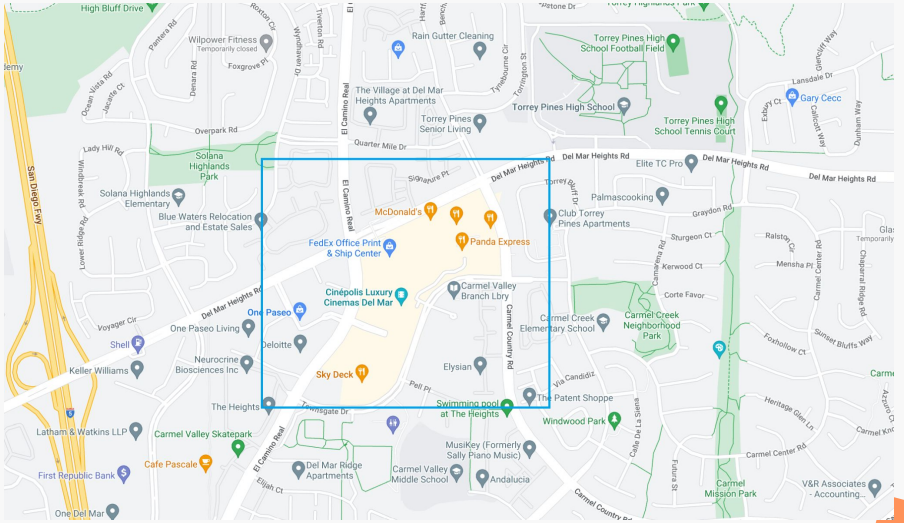
```
# Add Businesses to spatial index
run_query("""MERGE (n:Business) with n CALL spatial.addNode('geom', n) YIELD node RETURN node;""")
```

Neo4J Spatial

- Allows for finding all nodes within a bounding box
- Allow to find nodes within distance of point

Library Repo

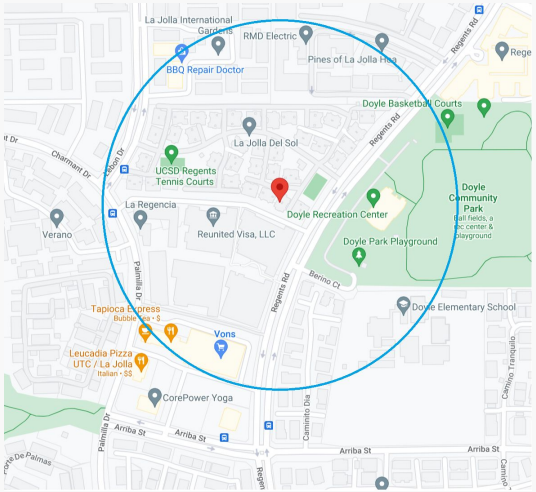




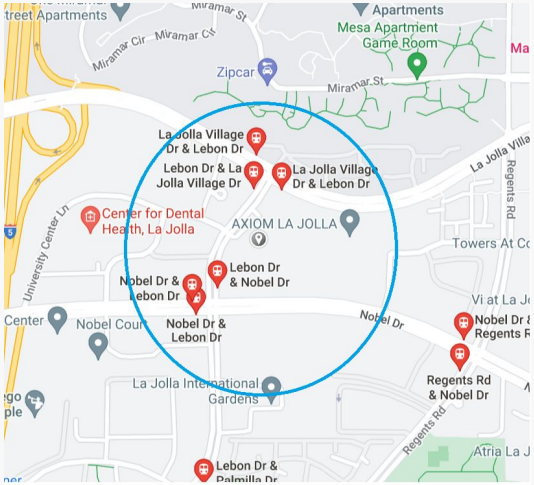
Queries

Business with optional Category

within radius of point

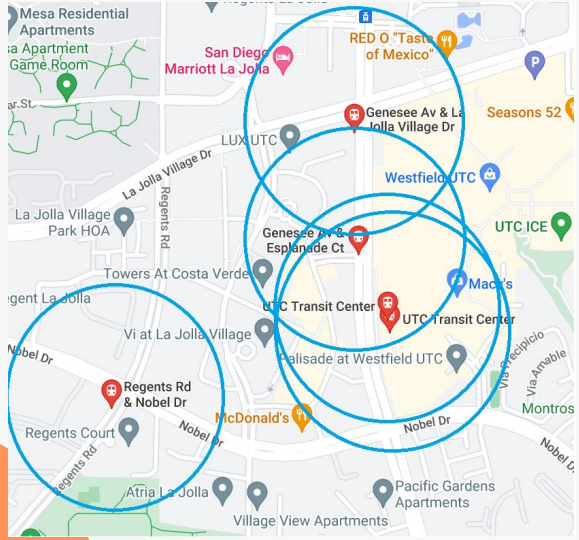


Bus stops within radius of point

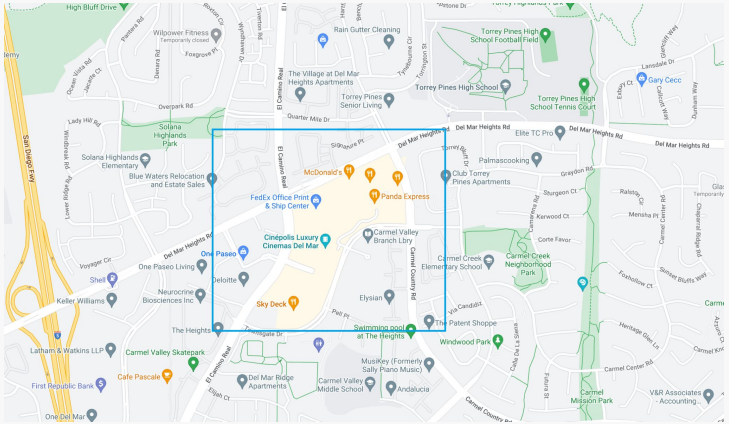


Queries (Cont)

Businesses with radius of all bus stops on a Route



Businesses and/or Bus Stops within bounding box



Neo4j Demo