Final Report

1. Team

Member Name: Gangyu Pan

Major: Master of Applied data science ; Bachelor of Information and computer science

2. Description

For the whole family, a suitable house is very important to them, because this is the core area of the family every day. There are some factors that affect house prices, which is very complicated. In this case, how to find the most suitable house for people to buy? I will design a recommendation system to help people choose houses. The system will show people more complexity than just price and location. I will provide more information, such as: public security in this area, whether it is a school district housing, whether the house is surrounded by scenic spots, climate, etc. These are all factors that affect housing prices. Through multiple levels of screening, this system will provide buyers with more suitable options for selection only.

3. Implementation process

When I want to deal with the above problems, I need to collect a lot of relevant data. Then, I need to process these raw data. For example, remove some abnormal data; fill in some vacant values; carry out data type conversion and so on. After processing, I use spark to process it again. Connect realtor data with university_info, use spark to calculate the distance between each house and its nearby school, and then determine whether it is a school district house; also calculate the average sqft for each postal house. Then, transfer the processed data to firebase real-time data. Finally, use flask+css+html+js to build a UI and connect it with firebase to achieve filtered query. Finally, the result is displayed on the front end of the html.

Raw Data: Handle Outliers

1).Realtor Data. From: real estate agent API:

Dataset attributes: estate address, price, property type, area, neighborhood_name, advertiser_id, lead_forms, photo, facility, update time, bed , bath , lat, lon, sqft , state etc.

Dataset process: This dataset has too much information for me, such as facility, neighborhood_name, advertiser_id, lead_forms etc. First, I need to pick some useful attributes. Then, I just keep these attributes in the dataset. For some attributes, there are some outliers, and I also need to deal with them so that they are easy to use in subsequent steps.

This is the csv file for this dataset:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	address	postal_cc	beds	baths	price	city	state	sqft	lat	lon			
2	200 Rectc	10280	1	1	########	New York	New York	547 sq ft	40.70847	-74.0164			
3	200 Rectc	10280	1	1	########	Manhattar	New York	640 sq ft	40.70847	-74.0164			
4	2782 Batc	11235	4	4	########	Brooklyn	New York	1,096 sq	40.5857	-73.9364			
5	2070 E 21	11229	3	3	########	Brooklyn	New York	1,512 sq	40.60222	-73.9512			
6	303 E Eas	10075	2	2	########	New York	New York	1,346 sq	40.77203	-73.9557			
7	140 W 22r	10011	1	2	########	New York	New York	1,063 sq	40.74211	-74.0007			
8	727 Ditma	11218	0	0	########	Brooklyn	New York	sq ft N/A	40.6366	-73.9705			
9	3207 Beve	11226	0	0	########	Brooklyn	New York	sq ft N/A	40.6455	-73.9467			
10	84-34 107	11418	3	2	########	Richmond	New York	1,376 sq	40.69951	-73.8427			
11	102-22 18	11423	3	2	########	Hollis	New York	1,120 sq	40.70727	-73.7707			
12	32 Gramer	10003	1	1	########	New York	New York	sq ft N/A	40.73687	-73.985			
13	200 Centr	10019	2	3	########	New York	New York	1,650 sq	40.76668	-73.9797			
14	1047 E 3r	11230	3	3	########	Brooklyn	New York	1,312 sq	40.62553	-73.9737			
15	51 5th Av	10003	2	2	########	New York	New York	sq ft N/A	40.73439	-73.9944			
16	2548 Matt	10467	4	2	########	Bronx	New York	1,914 sq	40.86464	-73.8623			
17	251 7th S	11215	1	1	########	Brooklyn	New York	671 sq ft	40.67198	-73.988			
18	1329 Hanc	11237	12	6	########	Brooklyn	New York	sq ft N/A	40.6944	-73.9094			
19	41-72 For	11373	6	3	########	Flushing	New York	sq ft N/A	40.74576	-73.877			
20	235 E 22r	10010	1	1	########	New York	New York	800 sq ft	40.73777	-73.9823			
21	247 W 46t	10036	1	2	########	New York	New York	999 sq ft	40.75985	-73.9874			
22	373 E 157	10451	4	2	########	Bronx	New York	sq ft N/A	40.82214	-73.916			
23	34-39 106	11368	0	0	########	Flushing	New York	sq ft N/A	40.75493	-73.862			
24	141-60 84	11435	1	1	########	Briarwood	New York	815 sq ft	40.71255	-73.816			

2).Crime rate dataset. From: scrape from website

Dataset attributes: chance_crime, city, crime_rate, state etc.

Dataset process: This dataset is fine for saving, because I can use all infomation. I just need to save it to the firebase database. I can use it to interface in the later step by matching the city with estate. For example, if this estate in the 'New York', I also can see the infomation about this city crime rate and chance crime in the interface.

This is the csv file for this dataset:

	A	B	C	D	E	F
1	city	state	crime_rat	chance_crime		
2	Detroit	MI	19.5	1 in 51		
3	St. Louis	MO	19.2	1 in 51		
4	Memphis	TN	19	1 in 52		
5	Baltimore	MD	19	1 in 53		
6	Monroe	LA	17.9	1 in 55		
7	Danville	IL	17.5	1 in 56		
8	Wilmingtc	DE	15.8	1 in 62		
9	Alexandri	LA	15.8	1 in 63		
10	Camden	NJ	15.7	1 in 63		
11	Scranton	PA	15.7	1 in 63		
12	Pine Bluf	AR	15.5	1 in 64		
13	Springfie	MO	15.3	1 in 65		
14	Little Rc	AR	15.3	1 in 65		
15	Saginaw	MI	15.2	1 in 65		
16	San Berna	CA	15.2	1 in 65		
17	Clevelanc	OH	15.2	1 in 65		

3). School infomation dataset From: download from website

Dataset attributes: city, state, lat, lon, school name, street, unitid, zip etc.

Dataset process: Even the attributes of this data set are useful for my project, but some schools are useless in future processing. Therefore, I filter the data by all real estate cities to reduce the number of schools in this data set.

This is the csv file for this dataset:

	A	B	C	D	E	F	G	H	I	J
1	UNITID	NAME	STREET	CITY	STATE	ZIP	NMCNTY	LAT	LON	
2	100654	Alabama A	4900 Meri	Normal	AL	35762	Madison C	34.78337	-86.5685	
3	100663	Universit	Administr	Birmingha	AL	35294-011	Jefferson	33.5057	-86.7993	
4	100690	Amridge U	1200 Tayl	Montgomer	AL	36117-355	Montgomer	32.36261	-86.174	
5	100706	Universit	301 Spark	Huntsvill	AL	35899	Madison C	34.72456	-86.6404	
6	100724	Alabama S	915 S Jac	Montgomer	AL	36104-027	Montgomer	32.36432	-86.2957	
7	100733	Universit	500 Unive	Tuscaloos	AL	35401	Tuscaloos	33.20702	-87.5296	
8	100751	The Unive	739 Unive	Tuscaloos	AL	35487-01	Tuscaloos	33.21188	-87.546	
9	100760	Central A	1675 Cher	Alexander	AL	35010	Tallapoos	32.92478	-85.9453	
10	100812	Athens St	300 N Bea	Athens	AL	35611	Limestone	34.80679	-86.9647	
11	100830	Auburn Ur	7440 East	Montgomer	AL	36117-359	Montgomer	32.36736	-86.1775	
12	100858	Auburn Ur	M	Auburn	AL	36849	Lee Count	32.59938	-85.4883	
13	100937	Birmingha	900 Arkac	Birmingha	AL	35254	Jefferson	33.51377	-86.8506	
14	101028	Chattahoc	2602 Coll	Phenix Ci	AL	36869	Russell C	32.42391	-85.0315	
15	101116	South Uni	5355 Vaug	Montgomer	AL	36116	Montgomer	32.34268	-86.2165	
16	101143	Enterpris	600 Plaza	Enterpris	AL	36330-13	Coffee Co	31.2975	-85.837	
17	101161	Coastal A	1900 U S	Bay Minet	AL	36507-269	Baldwin C	30.85134	-87.7782	
18	101189	Faulkner	5345 Atla	Montgomer	AL	36109-339	Montgomer	32.38418	-86.2164	
19	101240	Gadsden S	1001 Geor	Gadsden	AL	35903	Etowah Co	33.994	-85.9914	
20	101277	New Begir	421 Martl	Albertvil	AL	35951	Marshall	34.27871	-86.1969	
21	101286	George C	1141 Wall	Dothan	AL	36303-923	Dale Cour	31.31527	-85.4658	
22	101295	George C	801 Main	Hancevill	AL	35077-20	Cullman C	34.07244	-86.7819	
23	101301	George C	3000 Earl	Selma	AL	36703-28	Dallas Co	32.44592	-87.0133	
24	101365	Herzing U	280 West	Birmingha	AL	35209	Jefferson	33.46847	-86.8325	

Spark: Process Datasets

1. Realtor Dataset

First, I read the realtor dataset from csv that we saved before. After that, I want to calculate the average sqft by postal code in realtor dataset.This attribute can provide users with a reference to whether the house is too large or too small in this area.

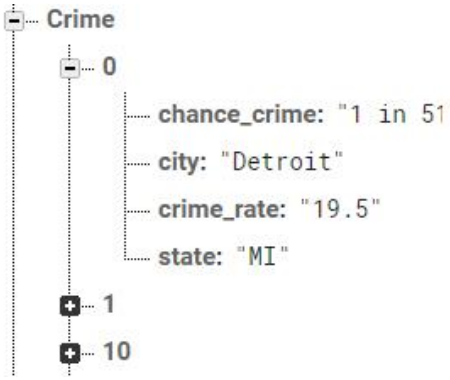
2. Realtor Dataset & Unviersity Dataset

I read these 2 datasets from csv files. Then, I calculate the distances for every house and every university. I just keep the minimum distance for every house. After that, I set a threshold to determine whether this house is a school district. If the distance is less than 10, it is a school district room; otherwise, it is not. At same time, I also store the nearly university in the dataset.

Firebase:

This is the whole picture for firebase realtime datasets.

1. Crime dataset



2. Realtor dataset


```
Realtor
  9099
    address: "5618 E 94th St in Southside Tulsa, Tulsa, 74137"
    Avg_price(PCode): 794220
    Avg_sqft(PCode): 4344
    baths: "2"
    beds: "4"
    city: "Tulsa"
    lat: "36.02707"
    lon: "-95.914525"
    Nearly_school: "Spartan College of Aeronautics and Technology"
    postal_code: "74137"
    price: "$220,000"
    School_district: "No"
    sqft: "2,227 sq ft"
    state: "Oklahoma"
```

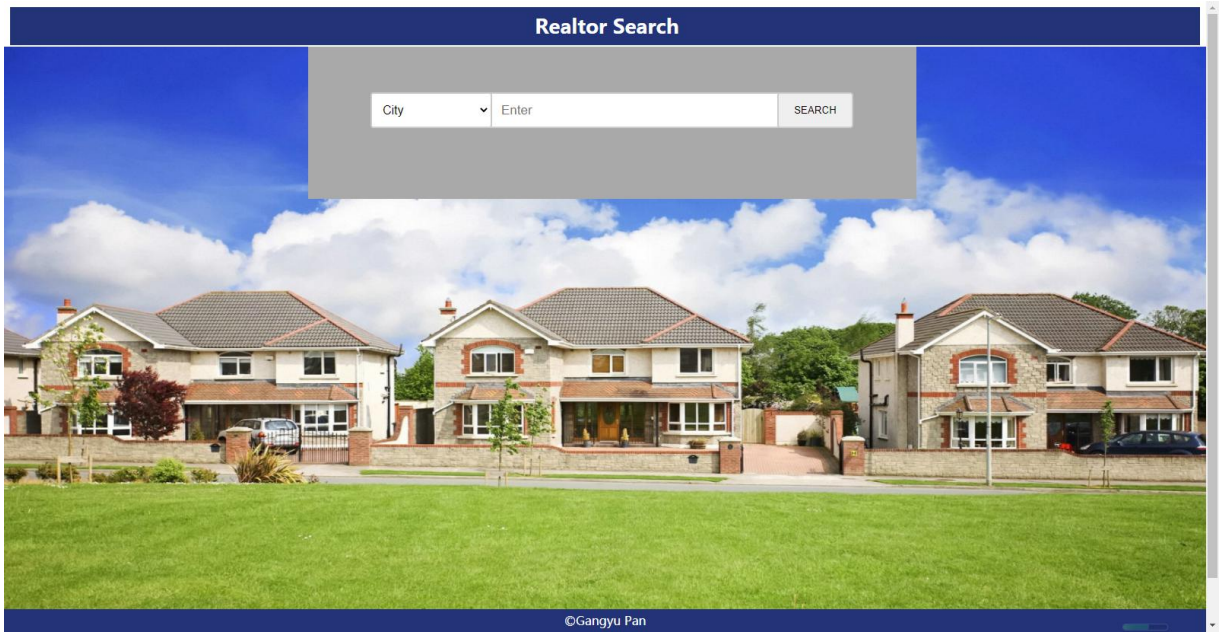

3. University dataset


```
School
  1813
    CITY: "Memphis"
    LAT: "35.134871"
    LON: "-89.992526"
    NAME: "Massage Institute of Memphis"
    NMCNTY: "Shelby County"
    STATE: "TN"
    STREET: "2076 Union Avenue, Suite 202"
    UNITID: "490470"
    ZIP
  1814
```

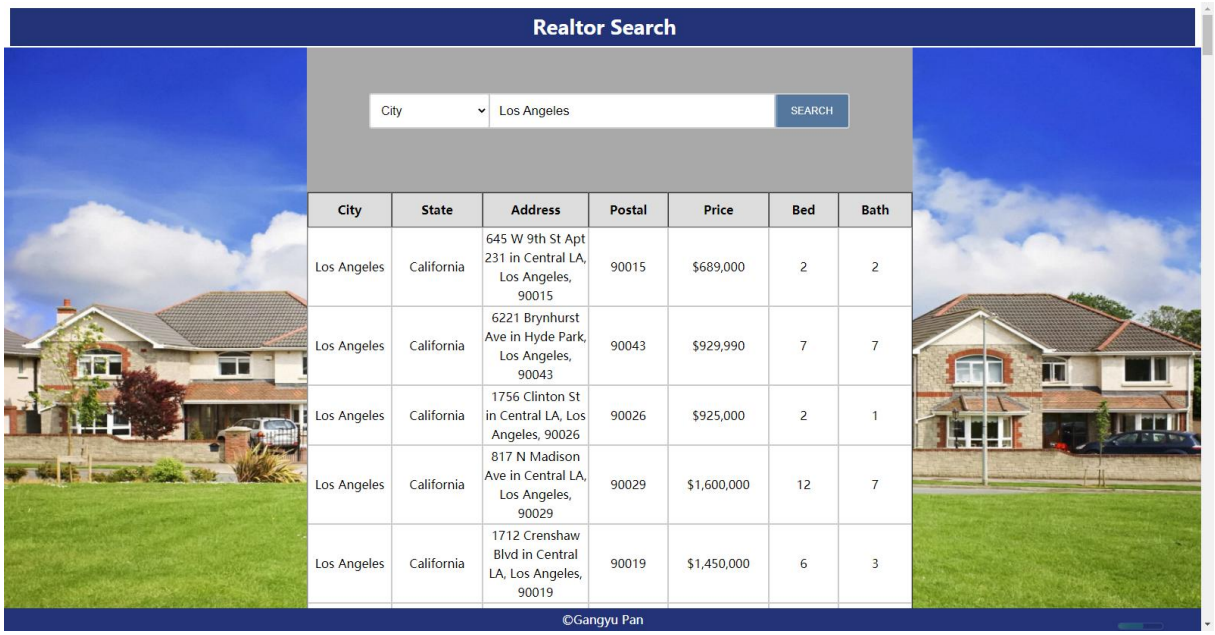

UI

I use flask to create an app file and get datasets from firebase. Then, I transfer the json file to html and js code. During this processing, I use html to design the different display on the front end. I also designed the drop-down box filter search function, after this operation, the content displayed on the page is the filtered data.

First, I will show you the home page:



This is a test case in my project, I choose the city option and give a city “Los Angeles”. Then, this table shows us the relational contents. You also can choose other options, like price, postal, bed and so on. This is result in the web:



City	State	Address	Postal	Price	Bed	Bath
Los Angeles	California	645 W 9th St Apt 231 in Central LA, Los Angeles, 90015	90015	$689,000	2	2
Los Angeles	California	6221 Brynhurst Ave in Hyde Park, Los Angeles, 90043	90043	$929,990	7	7
Los Angeles	California	1756 Clinton St in Central LA, Los Angeles, 90026	90026	$925,000	2	1
Los Angeles	California	817 N Madison Ave in Central LA, Los Angeles, 90029	90029	$1,600,000	12	7
Los Angeles	California	1712 Crenshaw Blvd in Central LA, Los Angeles, 90019	90019	$1,450,000	6	3

4. Experiences and Lessons Learned

For this project, during the process of completion, I encountered many problems and learned a lot. At the same time, some of the knowledge learned in class was also applied to achieve very good results.

First of all, at the beginning, I used the crawler process of Python basic learning last

semester to collect a lot of data. I also used the API to access the database to get the massive amount of data that I wanted more conveniently. After getting the data, there are a lot of data that give me a very headache, and I can't deal with some outliers or missing values. I spent a lot of time processing this part. Then after this part, I used the spark that I learned in class to process the massive data to read in the data, and then perform operations to achieve the results I want. It is also passed into the firebase database learned in class to make the subsequent page interaction more convenient. What made me spend a lot of time was the process of building the UI, because I had never touched the process of building web pages and connecting to databases.I searched the Internet for a method flask can be operated, and I learned from many videos on YouTube. Then, imitate and add a lot of your own designs to deal with the project. In the process of processing the front-end display, a lot of data did not achieve the results I wanted. For example, the data read in firebase is a string in js, and it is difficult for me to store the value after obtaining the key value. I tried and searched for information, forced to convert it to Object first, and then performed operations to get the value successfully. In this series of attempts, I learned a lot of new knowledge, which benefited me a lot. Although the final project did not have many functions, this was also the first step I tried.

I used the spark processing data that I learned in class and the method of storing data in firebase in my project.

5. Recording and Github Link

Github: https://github.com/Lucas0717/house-search
Recording:https://drive.google.com/file/d/1xx5spVc1jgPggDbH0iP3Ww-5CJ-3J9XX/view?usp=sharing