Proposal

Topic: Find the most suitable house

1. Team Information:

Team member: Gangyu Pan

Undergraduate major: Information and Computer Sciences

Experiences: Python (DSCI 510, DSCI 552), SQL(Database application), Matlab(Undergraduate)

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. Datasets:

I collected a set of sold house information data from the real estate agent

api, including many useful basic house information. For example, address, price, property type, area, facility, update time etc., which are important for people who want to buy a house. For some useless information, I delete it to make storage more convenient.

I plan to find a website that contains crime scores for every postal code in the United States. I want to obtain this data set through api as an influencing factor for selecting houses. By matching the first data set to stitch the entire dataset, this will provide buyers with more accurate recommendations.

For some school district rooms, I want to collect some information myself. Like calculating the distance between these houses and setting a range to determine whether it is a school district house or not, this is also an important factor affecting the price and buyers, because these types of houses will reduce college expenses. I will add these data to make more suitable recommendations to buyers.

For other factors, depending on the situation, if there is extra time, more time will be added to make the data set richer.

For these datasets, I will save as a json file to upload to firebase, which will be a realtime database. I will use this dataset to connect with my system, the data will be displayed on the interface.

4. Sample Data:

This is the raw house sales dataset. I only provided part of the house

information, but it has not been processed yet. Later, I will filter it to keep only important information. You can check the following pictures:


```
{'address': {'city': 'Jackson Heights',
             'county': 'Queens',
             'fips_code': '36081',
             'lat': 40.765143,
             'line': '23-55 79th St',
             'lon': -73.889712,
             'neighborhood_name': 'Northwestern Queens',
             'postal_code': '11370',
             'state': 'New York',
             'state_code': 'NY'},
 'agents': [{...}],
 'baths': 3,
 'baths_full': 3,
 'beds': 5,
```

```
'lot_size': {'size': 1875, 'units': 'sqft'},
'mls': {'abbreviation': 'LINY',
        'id': '3287298',
        'name': 'OneKeyMLS',
        'plan_id': None,
        'type': 'mls'},
'office': {'id': 'b6b3878fe3a998d0472be3be68939e80',
           'name': 'Keller Williams Rlty Landmark'},
'page_no': 1,
'photo_count': 1,
'price': 1589000,
'products': ['core.agent', 'core.broker', 'co_broke'],
'prop_status': 'for_sale',
'prop_type': 'multi_family',
'property_id': 'M4422858845',
'rank': 9,
```


For crime data, I just get one location related information.

```
{
  "total_incidents": 2,
  "total_pages": 1,
  "incidents": [
    {
      "city_key": "AUS",
      "incident_code": "20191010131",
      "incident_date": "2019-04-11T01:40:00.000Z",
      "incident_type": "Alcohol-related offense",
      "incident_official_type": "DWI",
      "incident_source_name": "Austin_Police_Department_Crime_Reports",
      "incident_description": "DWI at 816 LAVACA ST",
      "incident_latitude": 30.27146322,
      "incident_longitude": -97.74426176,
      "incident_address": "816 LAVACA ST"
    },
    {
      "city_key": "AUS",
      "incident_code": "20183640046",
      "incident_date": "2018-12-30T02:48:00.000Z",
      "incident_type": "Alcohol-related offense",
      "incident_official_type": "DWI",
      "incident_source_name": "Austin_Police_Department_Crime_Reports",
      "incident_description": "DWI at W 9TH ST / LAVACA ST",
      "incident_latitude": 30.27154076,
      "incident_longitude": -97.74415195,
      "incident_address": "W 9TH ST / LAVACA ST"
    }
  ]
}
```


Website:

https://www.realtor.com/ (get sale house basic information)

https://www.crimeometer.com/crime-data-api (get crime api)

5. Plan to approach the problem:

For this question, I will first get the original data set. From the real estate

agent api, I can select each state and country, which will make my data set more comprehensive. Then, I need to limit the number of houses in each location, otherwise, the entire US data set will be too much. In order to get the crime rate in a specific location, I will use the Crimeometer API to get it. As for determining the school district house, I will calculate the house data set of the distance to the nearby university and set a range to determine whether it is a school district house.

I will use Spark in my data processing project. It can help me process data efficiently and perform data aggregation, which is very important for my project. I have never used spark before, and I don’t know how to use it. It takes time to explore and research. By processing the dataset, I can get more useful datasets to upload my Firebase dataset.

Finally, I want to design an interface to display this data set. Users can filter some houses according to their requirements. Then, this system will give the most suitable houses for house buyers.