**Data Science Challenge**

**1.Project title:** Find the most suitable house

**2.Dataset Description**

These datasets come from real estate agent api, crime rating information in the U.S. and school location in the U.S. First, 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. I also get a crime rating fors some cities in the United States, which is a very important factor for people to choose the nearby real estate.For some school district rooms, I collect some information for school location. Then, I can calculate the distance between these houses and set 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. These three datasets are related by latitude and longitude. Then, I can get some new attributes, like whether it is school district rooms.

**3.Data Science Challenges**

**1). Data Cleaning**

For the school database, there are too many data. How can we keep the schools that can be used in the realtor data and clear other schools? For example, if a state or city does not appear in the realtor, the school information in that area can be removed.

**2). Data Analysis**

Through the three databases I gave, how to recommend suitable houses through the search of buyers? Can the above three databases be integrated to set an indicator to determine the purchaser's substitution?