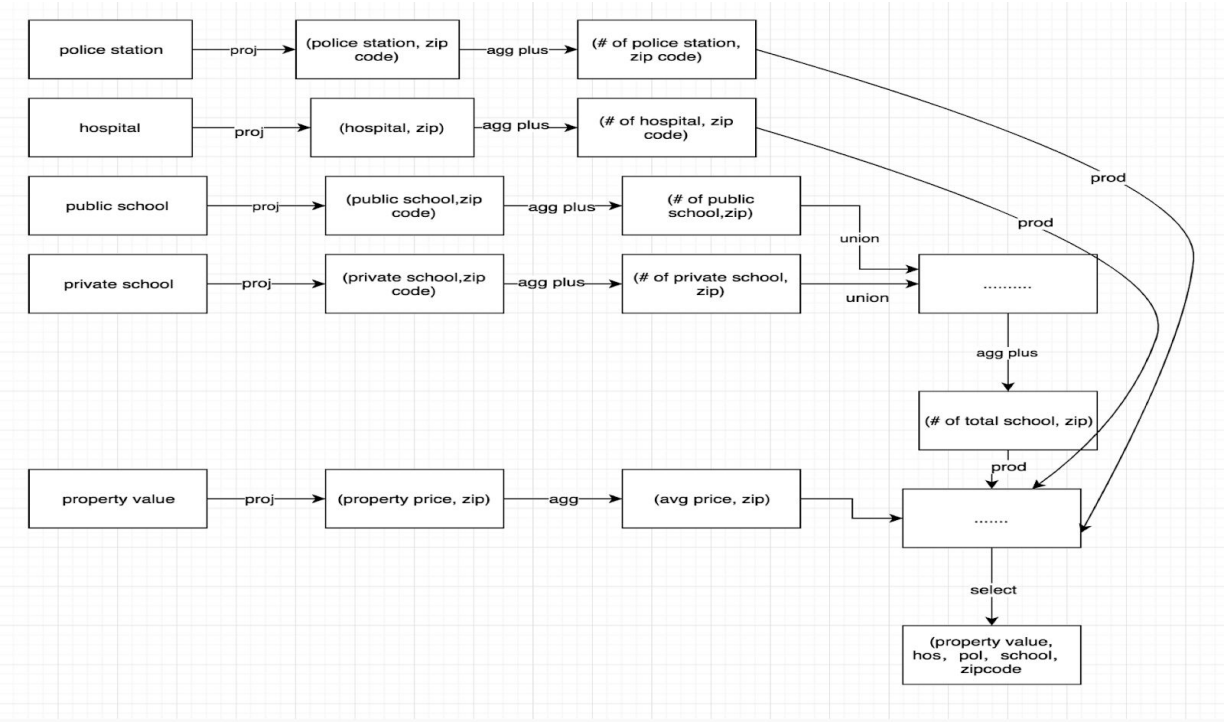
Yiyan Zhou
Rui Pang

CS 504 Final Report

Boston is one of the biggest metros in the US, the house price varies among different districts. It's hard for people to consider all the factors when they make a decision on house location. In our following projects, we want to find the factors which can influence the property value in Boston. Among the factors we try to find which factor has more influence on property value compared with other factors. Finally, based on our analysis and work, we build a prediction model for Boston property value. If a user inputs the information about the facilities of a specific zone, our model will output a prediction average property value.

To achieve our goal, in Project 1, we obtain the information about the number of hospitals, police stations, and schools in different areas of Boston by extracting data sets: Boston Hospital, Boston Police Station, Boston School, and Boston Property Value from Analyze Boston. In order to find the correlation between each factor and the property value, we calculate the number of each facility (school, hospital, police station) in each area and project each facility in the form {Number of Facilities, Zip Code}. Then, we merge all collections into one collection with zip code as key. In this case, we obtain a table that combines all our retrieved information together to perform further research on. The graph below shows that how we convert the initial data set to the processed dataset using basic building blocks.

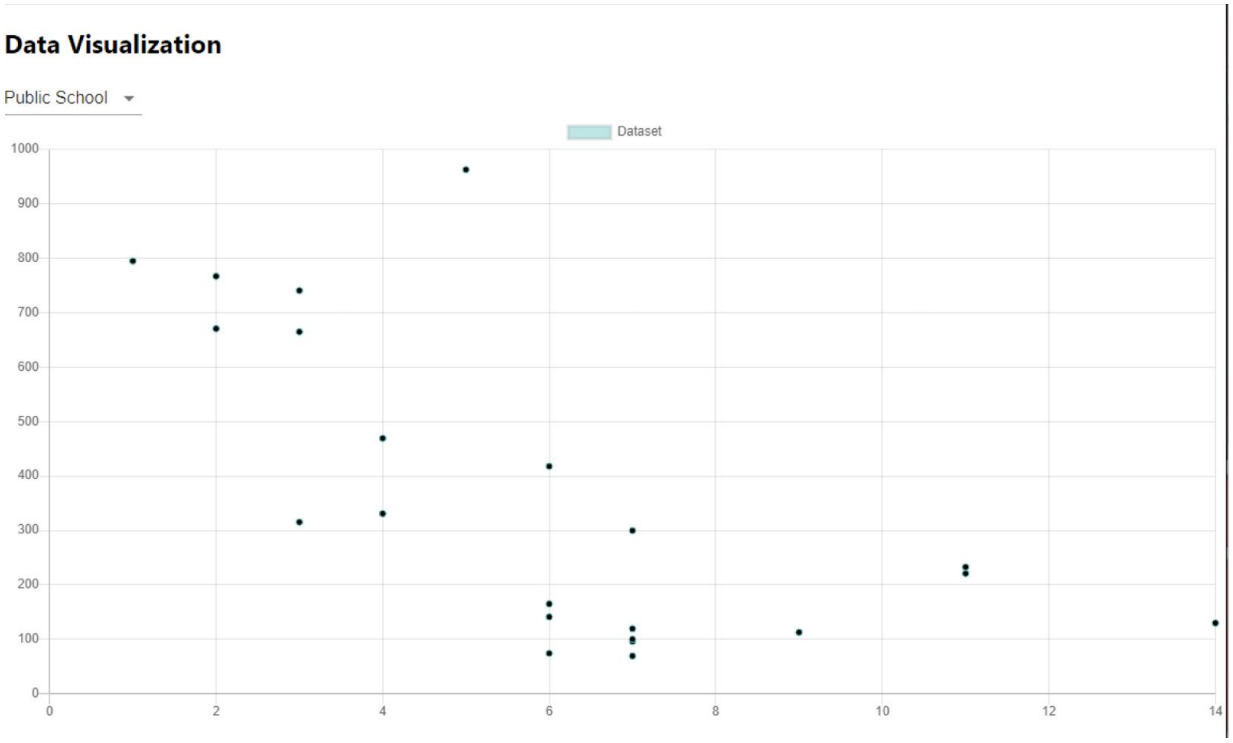


In project 2, we did more analysis on the data sets we obtained in project 1. Firstly, we did a statistic analysis on these data sets. We use Numpy package in python to calculate the correlation coefficient between each factor (police station, schools, and hospital) and property value. After calculation, we found that correlation between number of police stations and housing price is close to zero. Thus, we come to a conclusion that there is no correlation between number of police stations and housing value. This is due to that each zip code area has only one police station. Number of hospital shows a strong positive correlation with property value which follows our expectation. However, the correlation between number of schools and property value is opposite to what we expected. The correlation coefficient is negative, which means that when number of school increases, the property value drops. We think the reason is that there is some limitation of our data sets. We only count for facilities in each zip code, However, we also need to account for the size of each zip code area. Number of facilities in each square mile will be a better measure. Besides statistical analysis, we also used our data to build a house price prediction model. We use Scikit-Learn linear model and feed the model

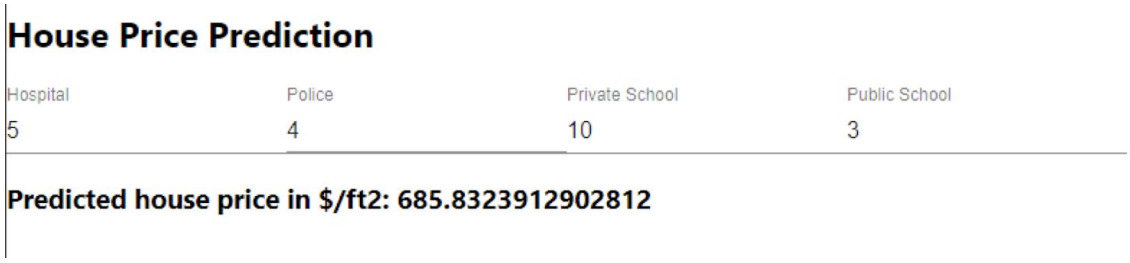
with our retrieved data as training data. User can just input several arguments like number of police station, hospitals and number of schools to get an estimated house price of certain area.

We implemented a web-based application in Project 3. The front end for the application is built in React.js, the back end is written in node.js, and we use MongoDB as our database. The project is located in folder “project3” of the group folder (ruipang_zhou482). To run the application, in the project folder, first run “execute.py” which is implemented in Project 1 and 2. Then, run “npm install” then “npm start” to start the web server. Also in “api” folder, run “npm install” then “node app.js” to start the backend.

We implemented two new features in Project 3. The first feature is a visualization on the data we processed in Project 1. In Project 1, we constructed a dataset that stores the house price, number of police stations, hospitals, public schools and private schools for each zip code area in Boston. A screenshot of the visualization is shown below. A user can select between police stations, hospitals, public schools and private schools to show a scatter plot of the number of selected facilities against average house price.



The second new feature is an interactive client-server application. In project two, we implemented a linear regression on average house price, against number of police stations, hospitals, public schools and private schools. The result of the regression is stored in MongoDB. The application uses this data to predict the house price given the number of hospitals, police stations, private schools and public school that a user typed in. A screenshot of the application is shown below.



In summarize, in these projects, we retrieve the data we need from authorized data websites. After preprocessing the data, we did statistical analysis on the data to get deeper view of the factors that may influence property value at Boston. Also, we build a prediction model based the analysis results and retrieved data. Finally, we build a web service to provide users friendly interface and visualization of results. In the future, we may change the way we divide the city, because we had some problem when doing analysis between number of school and house price. We plan to divide the city by blocks, so we need to find more specific data and do more preprocess on the retrieved data. Moreover, we have a more challenging idea. We think about building a model which can automatically choose the factors that may have influence on the data instead of we initially choose several factors by our experience, because there may be some hidden and surprising information among the data sets.