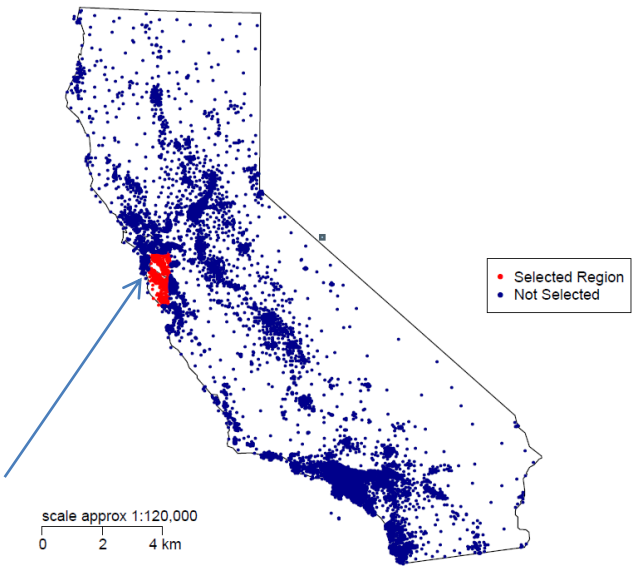
Daniel Prusinski
Predict 412

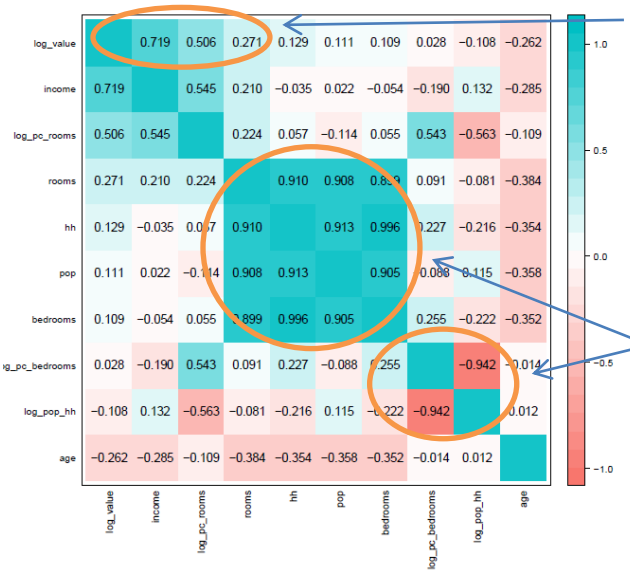
Programming Assignment 6: Fitting a Spatial Model

Within real-estate, there have been three rules that dominate the price of a home "location, location, and location." While location is an important factor when defining the price of a home, there are additional factors that affect the overall price of a home.

The following exploratory data analysis (EDA) employs spatial data and various modeling methods to predict housing prices within a defined area. California has been home to housing booms and busts similar to that of the Gold rushes, and this EDA focuses on California 1990 US Census data. The dataset for this EDA includes 20,640 unique observations and 9 variables. All outputs, graphs, and visualizations are found within Appendix 1. The data is divided into training (2/3) and testing (1/3).

Data transformations are a necessary component of a thorough EDA. In order to manipulate the data such that it followed a more normal distribution, the following variables had a log transformation: Value, Age, Log(Rooms/Population), Log(Bedrooms/Population), Log(Population/Households), and Households (Output 1 & 2). After the transformations, including squaring and cubing income, the variables better interacted with one another such that the modeling was more insightful (Output 3). California is a vast state, and the EDA focused on San Francisco and the neighboring area (Output 4). As seen to the right, the area selected is small geographically, but densely populated. The selected area had 1,962 observations and 1,307 were training and 655 were testing data.





The correlation heat matrix to the left shows a couple variables have favorable correlation with house values. An issue to consider when modeling is the correlation between the variables, which violates an assumption of linear regression, and is known as multi-collinearity (Output 6). Before delving into the spatial modeling, analyzing a linear model without the spatial coordinates was helpful for establishing a baseline. The linear regression model had an adjusted R-squared (ARS) of .643 and root mean-squared error (RMSE) of .283 on the training data and .702 ARS as well as a .259 RMSE on the testing data (Output 7). Full tree regression produced a RMSE of .274 and an ARS of .67 on the training data and a RMSE of .293 and an ARS of .62 on the testing data output 8 shows the tree and technical node output. The full random forest produced the best metrics out of the three models tested. With an R-squared (RS) of .950 on the training data and a .730 on the testing data, over fitting might have been issue. A similar trend also occurred with the RMSE, the training data produced a .116 and a .245 on the testing data. Output 9 shows the technical output as well as the variable importance, and income, log_pc_rooms, and log_pop_hh were the most important variables. Geographically weighted regression produced the best testing and training models. The RS for the training model was .884 and .770 on the testing data. For the RMSE, the training data produced a .163 and .232 on the testing data. The hybrid prediction accounted for .935 of the variance (training) and .813 variance on testing data. Overall, the geographically weighted regression model was appropriate and best fit the data for the EDA.