# MODELS AND VISUALIZATIONS
# FOR HOUSING PRICE PREDICTION

A Thesis

Presented to the

Faculty of

California State Polytechnic University, Pomona

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science

In

Mathematics

By

Anh Komagome-Towne

2016

## SIGNATURE PAGE

**THESIS:** MODELS AND VISUALIZATIONS FOR HOUSING PRICE PREDICTION

**AUTHOR:** Anh Komagome-Towne

**DATE SUBMITTED:** Fall 2016

Department of Mathematics and Statistics

Dr. Adam King _____
Thesis Committee Chair
Mathematics & Statistics

Dr. Hoon Kim _____
Mathematics & Statistics

Dr. Jennifer Switkes _____
Mathematics & Statistics

## ACKNOWLEDGMENTS

I would like to thank my thesis advisor Dr. Adam King for his time and guidance. I also would like to thank Dr. Hoon Kim and Dr. Jennifer Switkes for agreeing to serve on my thesis committee. Many thanks go to my friends and family for their support.

## ABSTRACT

The main objective of this thesis is to compare house price models for single family properties in Pasadena, California, and to explore these results through visualizations. We collected data on 5,142 listings from Redfin.com for our analysis. The sample contains all single family homes, townhouses, and condominiums sold between October 2012 and October 2015 in several Pasadena areas. Using R, a programming language and software environment for statistical computing, we will process and analyze real estate data in order to make predictions regarding sale prices.

In order to select a prediction method, various regression models are explored and compared to find an appropriate fit. Methods discussed include multiple linear regression, k-nearest-neighbors, tree-based methods (including decision trees, bagging, and random forests), and nonlinear regression techniques (splines and generalized additive models). We compare and assess the performance of these methods by examining median percent prediction error.

Features used for prediction consist of commonly listed specifications including interior square footage, lot size, number of bedrooms, number of bathrooms, year the house was built, and date of sale. The relationship between location and housing values is investigated through heat maps superimposed onto maps of the Pasadena area.

Results indicate that generalized additive models perform best. Additionally, all models and methods agreed with respect to the direction and magnitude of the relationships between the predictors and housing prices. In particular, when holding all other features fixed, interior square footage, lot size, and the number of bathrooms are positively related to house prices, while the number of bedrooms surprisingly has a negative impact on the

price. Along with spatial coordinates, square footage is found to be the most important feature and accounts for a large amount of price variation. Our best median percent error is approximately 10%, which means our predictions are typically off from the actual price by 10%. These results offer insights into the features that drive home prices, and furthermore these results practically help sellers to offer their homes at fair prices and help buyers avoid paying too much for their homes.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Motivation and Goals

The last decade has witnessed recurrent inconsistency in the U.S. housing market due to economic fluctuations and market changes. In the United States, the annual sales of existing single-family homes from 2005 to 2008 fell by 30%, while their average price dropped about 26%, and new construction declined by a 64% (Lombra, 2012). Nationwide, home values decreased from an average price of $221,900 in 2006 to $177,700 in 2009, which is an approximately 20% reduction in only three years (NAR, 2009). The descent in home values raises questions among home owners and shoppers regarding how accurately the value of homes can be assessed and what attributes determine the desirability of certain homes compared to other houses on the market.

This instability in the housing market necessitates the need for better methods for assessing housing prices. Consequently, the predictive accuracy of housing models has gained much attention among scholars and has been extensively studied. In particular, we are interested in methods that can estimate the value of a home based on its attributes in

comparison to the current market price of similar houses. This ability to predict housing prices is important to anyone buying or selling a home, as well as to investors making asset allocation decisions.

The nature of real estate markets is inherently intricate. In addition, the diversity of housing attributes makes it challenging to construct a comprehensive model that can encompass all of the features. The value of a house is substantially affected by its own unique set of structural attributes (e.g. living area, number of rooms, age of the building, swimming pool, garage), locational attributes (proximity to business districts and accessibility to major highways) and neighborhood attributes (e.g. school district, income levels, unemployment rate, population density).

Given the complex nature of this problem, to attain a model that accurately predicts the value of a home is difficult for a variety of reasons. One reason is that data sets with the complete information on all mentioned attributes are not easy to access. Secondly, the effect of a specific attribute on housing price may not have the same amount of significance on price variation across different regions. For example, the effect of square footage on price variation may not be consistent for all neighborhoods. Also, certain attributes are valued differently in various areas. For instance, in colder regions, a fireplace may have a greater value while houses with swimming pools are more desirable in a warmer climate. For those reasons, housing price predictions are usually conducted on a specific location, and are difficult to generalize across different geographic regions. Since the housing price largely depends on location, it is necessary to account for spatial effects to improve prediction accuracy.

Despite the discussed challenges, past research has shown that it is possible, at least partially, to assess the future value of a house. We will discuss some of the past studies in the next section. Within the limited data available, our goals in this study are to assess

the prediction accuracy from various models and give clues of how certain features affect home prices. This information could prove useful to investors, buyers, and sellers of homes.

## 1.2 Review of Previous Work on House Price Prediction

A variety of research has been developed to model housing prices and property values. First developed in the 60's, hedonic regression has been the most common approach because it allows the total housing expenditure to be decomposed into the values of the individual components.

Hedonic modeling is a widespread method in which a property is assumed to be a heterogeneous good and can be broken down into characteristics such as structural features (e.g. year built, square footage, etc.) and locational factors (e.g. distance to commercial areas and major streets, etc.). The relationship between sale prices and housing characteristics along with neighborhood properties is measured by the hedonic regression (Rosen, 1974). In his works, Rosen defined "goods are valued for their utility-bearing attributes or characteristics. Hedonic prices are defined as the implicit prices of attributes and are revealed to economic agents from observed prices of differentiated products and the specific amounts of characteristics associated with them."

Malpezzi et al. (1987) compared housing to a bag of groceries. Bags with various sizes contain different items that contribute to the overall price. Therefore, regression analysis and the hedonic method can be used to determine the contributing effect each characteristic has on the overall price. Since housing is a heterogeneous commodity, making direct comparison of rents and values is difficult. Malpezzi et al. suggested that the hedonic approach is a convenient method to decompose expenditures on a heteroge-

nous commodity into price and quantity. The hedonic approach also provides a direct way to adjust for quality differences between old and new housing, making direct estimates of economic depreciation possible.

In Nguyen and Cripps (2001), the authors compared multiple regression analysis (MRA) to artificial neural networks (ANN) using three differently sized training sets of single family houses. General house attributes such as the square feet of living area, the number of bedrooms, the number of bathrooms, the number of years since the property was built, the quarter the property sold, and whether or not the property has a garage or carport were used for their predictions. The researchers obtained price fluctuation results and proposed that while MRA performs best on smaller sized training sets, ANN were found to outperform multiple regression models when the size of the dataset increases.

Sirmans et al. (2005) studied various hedonic models predicting housing prices that have been published over the previous decade. This study identified the most common housing attributes that were used in hedonic pricing models and measured the effect each attribute has accordingly on the overall price. Some of the attributes often found in hedonic modeling studies include, but are not limited to, lot size and its logarithmic transformation, square footage and its logarithmic transformation, age of the house, number of bathrooms, number of bedrooms, fireplace, air-conditioning, basement, pool, and time on market. Out all the mentioned attributes, lot size and square footage appeared most frequently and were found to have the most significant impacts on housing prices. Other studies find significant nonlinearities between home value and age (Do and Grudnitski, 1993; Grether and Mieszkowski, 1974) and home value and square footage (Goodman and Thibodeau, 1995).

Hedonic models that are done on housing prices typically have to account for spatial effects. Basu and Thibodeau (1998) explained that spatial dependence exists for a

few reasons. One is that houses in the same neighborhood are usually zoned and constructed during the same time with similar structural design. This implies that structural attributes such as living area, number of rooms, age of the house, etc. are correlated with geographic location. Another reason is that houses in the same area have the same proximity to major business centers and share locational amenities. The distance to work and accessibility to local services are contributing factors to determining the value of houses.

While the hedonic method is widely acceptable for accommodating real estate attributes to make predictions, one disadvantage is its complexity. Issues such as variable interactions, heteroscedasticity (non-constant variance), multicollinearity, non-linearity and outlier data points can hinder the performance of the hedonic price model in housing prices. The second disadvantage is that hedonic models are limited within a specific studied location. Thus, hedonic models are generally used to gain insight into one particular market and it is difficult to generalize across different geographic regions (Sirmans et al., 2005).

# Chapter 2

# Data Set

In this thesis, R is used extensively for data processing as well as mathematical and statistical computations. R is an open source programming language and environment that provides graphical techniques, predictive modeling and data analysis (R Development Core Team, 2008). R can be downloaded from the R Project website:
`http://www.r-project.org/`

## 2.1 Data Acquisition

The dataset used for this research consists of actual real estate transactions collected from Redfin, which is a real estate listing website that provides web-based brokerage services in the United States (Redfin, 2004). The downloaded files come in comma-separated value (CSV) format. Each cell inside each data files is separated by a special character, which usually is a comma. We initially gathered 26 CSV files, each containing between 300 and 500 properties listed for sale. The number of separate files is due to the fact that Redfin limits the number of observations that can be included in a single file. In this thesis, we define an observation as a listing of an individual property and its

corresponding attributes. Each observation in our data set includes 33 attributes (property characteristics) originally. We then merged the 26 separate files to create one complete CSV file consisting of 5,142 observations.

The data was read into R with the `read.csv()` function, which returned a data frame with 33 columns and 5,142 rows. However, this data set included observations not relevant to our prediction problem (for example, located outside of Pasadena or with an unknown location), and we removed all observations not meeting the following inclusion criteria:

1. Property was sold between October 18, 2012 and October 18, 2015 (inclusive).
2. Property is a single family home, town home, or condominium.
3. Property has a non-missing address.
4. Property has a non-missing last sale price.
5. Property is a house in Pasadena or South Pasadena.

## 2.2 Data Cleaning

Data in raw form are not ideal for analysis. To be useful for predictive modeling the data must first be cleaned. The purpose of data cleaning is to improve the quality of data by detecting and removing errors and inconsistencies from data. Data quality problems often are present due to misspellings during data entry, missing information, duplicates, redundant data in different representations (e.g. region listings of "South East Pasadena" versus "SE Pasadena"), or other general invalid data. The data cleaning process includes eliminating missing or duplicate information, filtering unnecessary data, and consolidating disparate data representations to provide access to consistent and accurate data.

### 2.2.1 Incomplete, Missing Data

In data preparation, problems arise when there are missing or empty values. A missing value in a variable occurs when an actual value exists but was omitted in the course of data entering. An empty value in a variable is one for which no real-world value exists or can be obtained. These values are expected to be corrected before we perform Exploratory Data Analysis.

It is essential to handle missing values properly because such values are generally hard to be digested and processed in R. Furthermore, we may assume inaccurate inference due to missing data. For instance, some modeling methods ignore missing and empty values, while others automatically substitute suitable values into the missing parts. The result obtained by one without missing values will differ from ones where the missing values are present.

One method to deal with missing data is to perform data imputation. Data imputation is the statistical technique for substituting missing or inconsistent data items with an estimated value based on other available information. However, when the percentage of observations with missing values is small (at most 10% of the sample) it is reasonable to omit those observations from the analysis. Since our data set consisted of roughly 90% complete cases, we simply omitted the observations with missing data, and hence no imputation was needed or performed.

### 2.2.2 Variable Selection and Filtering

This section describes the variable selection (column subsetting) and data filtering (observation subsetting) process. To prepare the dataset for analysis, the following changes were made:

1. We make a new data frame with nonessential information disregarded. Removed variables include: state, days on market, next open house date, next open house start time, next open house end time, recent reduction date, source, listing ID, original source, favorite, interest, sale type, and listing status. The variables retained include: home type (single family residence, condominium, or townhome), address, list price, number of bedrooms, number of bathrooms, location in Pasadena (Northeast, Northwest, Southeast, Southwest, Pasadena, and South Pasadena), square footage, lot size, year built, number of parking spots, parking type, original list price, last sale date, last sale price, latitude, longitude, and a categorical variable indicating whether the property is a short sale.

2. We omitted the 25 observations with home types of multi-family unit, vacant land, or unclassified type, and retained all single family residences (SFR), townhomes, and condominiums. We did this because our goal is to predict prices of homes consumers purchase with the intent to occupy as their primary residence.

3. We only included observations with sale dates between October 18, 2012 and October 18 2015. At the time of initial data collection, October 18 was the latest date for which data on all Pasadena home sales was available, and we selected the prior three year time period to comprise our sample. In this step, 276 transactions that were sold before or after that time period were omitted.

4. There are 12 rows that have "Undisclosed" address values, which we removed. The second issue is that some properties are listed more than once. These entries have the exact same housing features and the last sale price. We proceeded to eliminate these 30 duplicated observations from the data set.

5. We identified the following features as being essential for our prediction model:

number of bedrooms, number of bathrooms, square footage, year built, and lot size (for single family residences). In addition, last sale price was our target for prediction. Hence, we removed the 119 observations that had missing values for any of these variables.

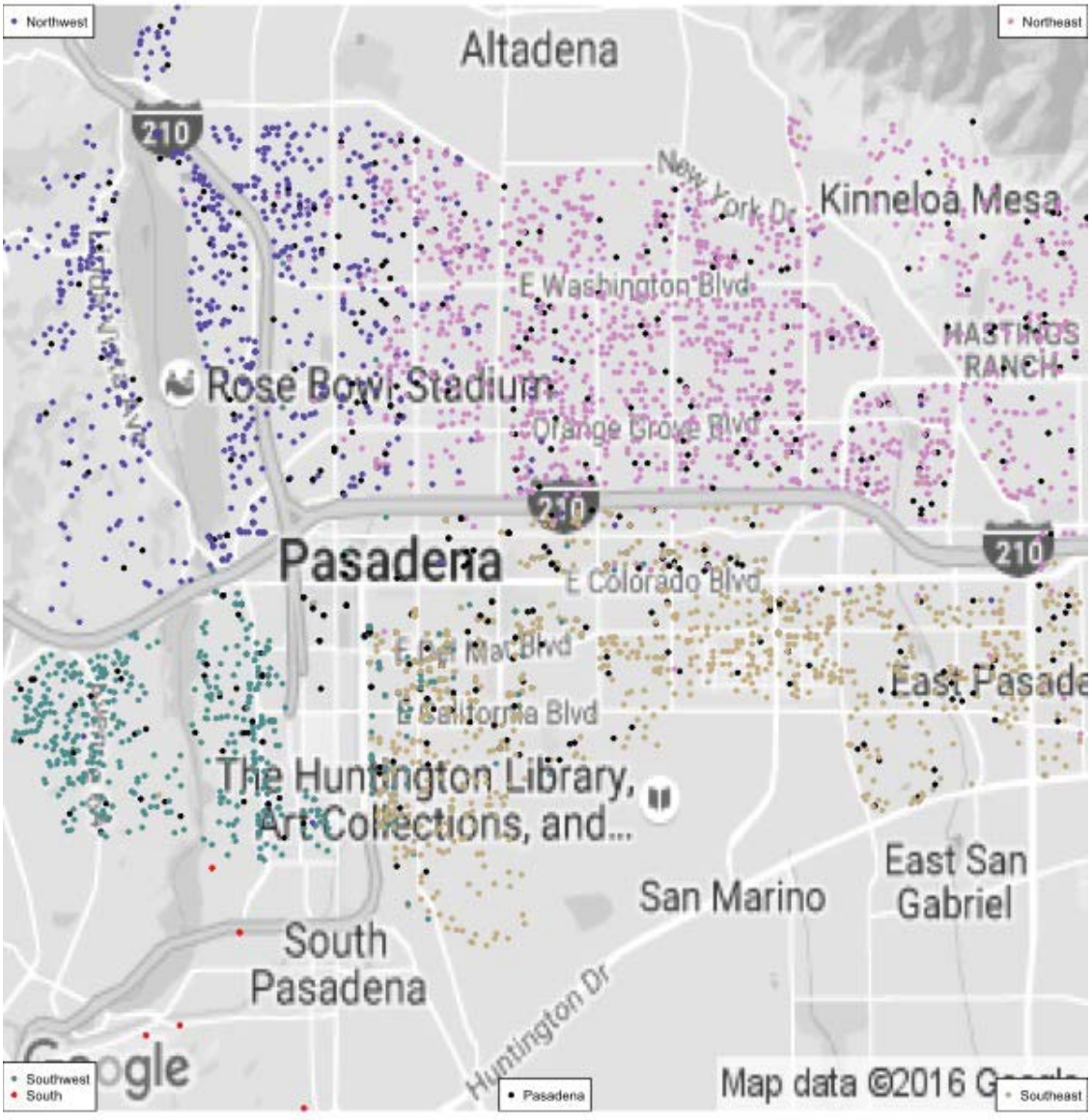
6. To limit geographic variation in housing prices within our data set, we removed any properties outside of Pasadena and South Pasadena or observations with unusual zip codes that do not appear to belong in this region. This step eliminated another 64 properties.

7. We removed 335 observations that we determined corresponded to currently active listings, since we instead wished to use only closed sales for our data set.

### 2.2.3 Inconsistent, Incorrect Data

One challenge we faced was to reconcile differing addresses in our data set. A big number of misspellings and incorrect information was presented during data entry. Location was recorded within Pasadena under inconsistent names such as Northeast, NE, Northwest, NW, Southeast, SE, Southwest, SW, Pasadena, South Pasadena, and many "Not Defined" values. We created a new variable called `cleanLocation` and stored the location parameters under five uniform values: Northeast, Northwest, Southeast, Southwest, and South Pasadena. Any unspecified observations were classified under "Pasadena" in general. We then used two variables `latitude` and `longitude` to project spatial coordinates of these observations onto Google Maps data. Figure 2.1 suggests that the location specifications which provided a quadrant (e.g., North-East) are approximately accurate. However, the general category of "Pasadena" appears to span the entire map. As shown in Figure 2.1, the black dots can be found in any of the four quadrants. We then conclude

Figure 2.1: Home locations within Pasadena with region variable color-coded.



that the `location` attribute does not contain reliable information and should be excluded from the analysis. This will entail no loss of information, since we later incorporate exact spatial coordinates (latitude and longitude) directly into our models.

## 2.3 Exploratory Data Analysis

The goal of this section is to provide in-depth preliminary investigations into our data set. In particular, we want to identify the outliers and examine the distributions as well as the descriptive statistics of all the existing variables.

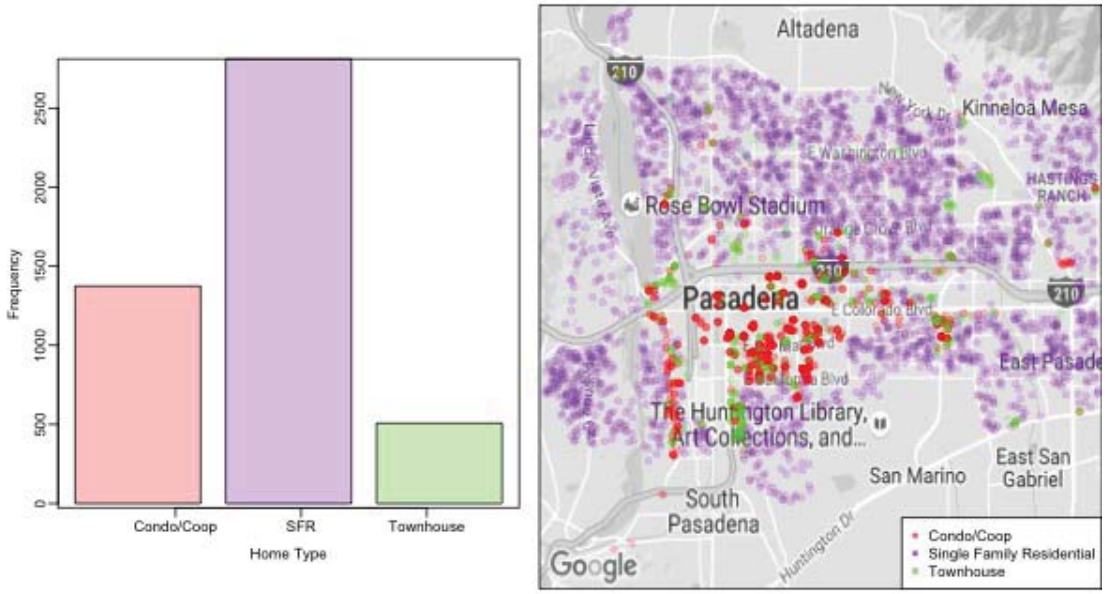
### 2.3.1 Data Introduction and Description

After data cleaning, our data set comprises 4,691 observations of 23 variables with complete information on spatial coordinates (no missing value on `latitude` and `longitude`).

In most regression analysis studies that involve real estate forecasting, the outcome variable of interest is almost always the property's sale price. The characteristic features or attributes of the property (e.g., square footage, lot size, bedrooms, etc.) that are used to predict the home values are called independent, explanatory, or predictor variables.

The aggregate data set contains three different home types: single family residences (SFR's), condominiums, and townhouses. Figure 2.2 gives relative proportions and the geographic distribution of these three types. In the map, the opacity of the color indicates the housing density. Out of 4,691 properties in the city of Pasadena, our aggregate data consists of 2,810 SFR's, 1,374 condominiums and 507 townhouses. SFR's are the typical real estate type for the outer Pasadena districts and the surrounding areas. Condominiums and townhouses are located with high density in the central area of the city. In addition, we can see a multitude of condos that are densely concentrated along the Long Beach

Figure 2.2: Counts and locations of condos, single-family residences, and townhomes within Pasadena.




freeway.

### 2.3.2 Univariate Exploration

We use univariate analysis to obtain the descriptive statistics for each single variable in the agregate data set. As discussed, our data set after cleaning comprises a total of 23 variables available for each observation. This list of variables contains detailed information such as home type, address, zip codes, sale type, etc. For analysis purposes, we only select certain essential variables that are relevant to our predictive modeling study.

The predictors are divided into two categories: spatial and non-spatial. Non-spatial predictors are housing structural features or attributes including square footage (`sqft`), lot size (`lotsize`), number of bedrooms (`beds`), number of bathrooms (`baths`), parking spots (`parkingspots`), and year of construction (`yearbuilt`). In addition to the structural

attributes, the geographic coordinates `latitude` and `longitude` of each property are also used as spatial predictors. This locational information allows the implementation of spatial methods introduced in the next chapter. Lastly, the dependent variable is the price the home sold for, `lastsaleprice`. We will also investigate the listing price (`listprice`) to draw a comparison between the actual sold price of a house and its listed value.

### Outlying Observations

Outliers are observations that are far from the large majority of the remainder of the data values. Outliers have an impact on the regression fits and may hinder a model's analyzability. Data visualization and univariate exploration are efficient ways to identify potential outliers. Some of the outliers are not incorrect data, but are far outside the range of data that we care about in our analysis. For example, any house with an extremely large lot size or square footage will have very high leverage if we include that variable linearly. We found and eliminated several obvious outliers from our cleaned dataset such as a property with 14 bedrooms and a property with 23 bedrooms. These observation are not relevant to our mainstream house price prediction but can drastically bias the fit estimates and predictions.

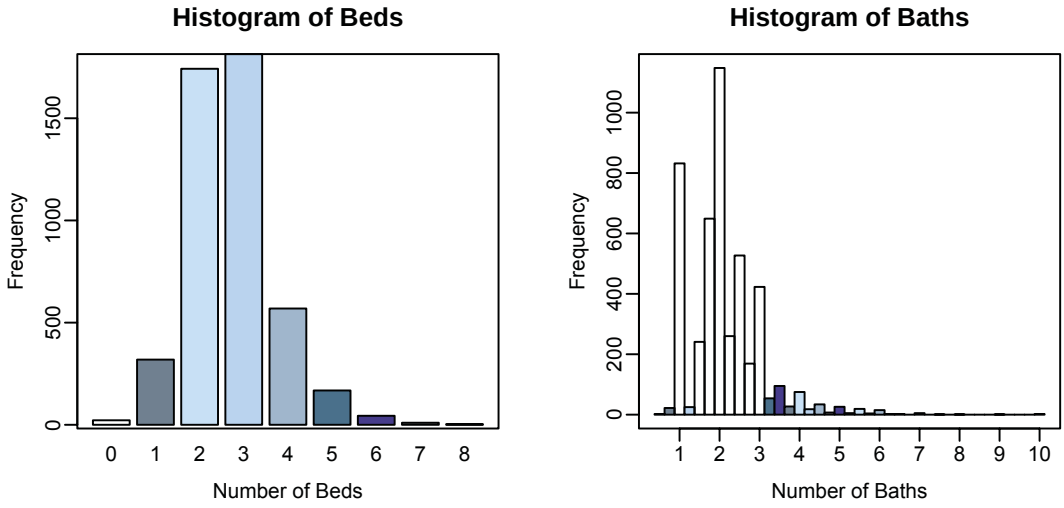
### Discrete Variables

There are three quantitative discrete variables which all assume a small number of possible values. They are `beds`, `baths` and `parkingspots` representing number of bedrooms, number of bathrooms, and number of parking spots respectively. However, `parkingspots` have not always been reliably reported, and a closer exploration from our dataset confirms this speculation. Hence, we are not including `parkingspots` for further analysis. The variable `yearbuilt` is quantitative discrete but with a large num-

Table 2.1: Univariate Summary Statistics For Number of Bedrooms and Bathrooms

| | Mean | S.D. | Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|---|---|---|
| Number of Beds | 2.71 | 1.000766 | 0 | 2 | 3 | 3 | 8 |
| Number of Baths | 2.12 | 0.9093696 | 0.5 | 1.75 | 2 | 2.5 | 10 |

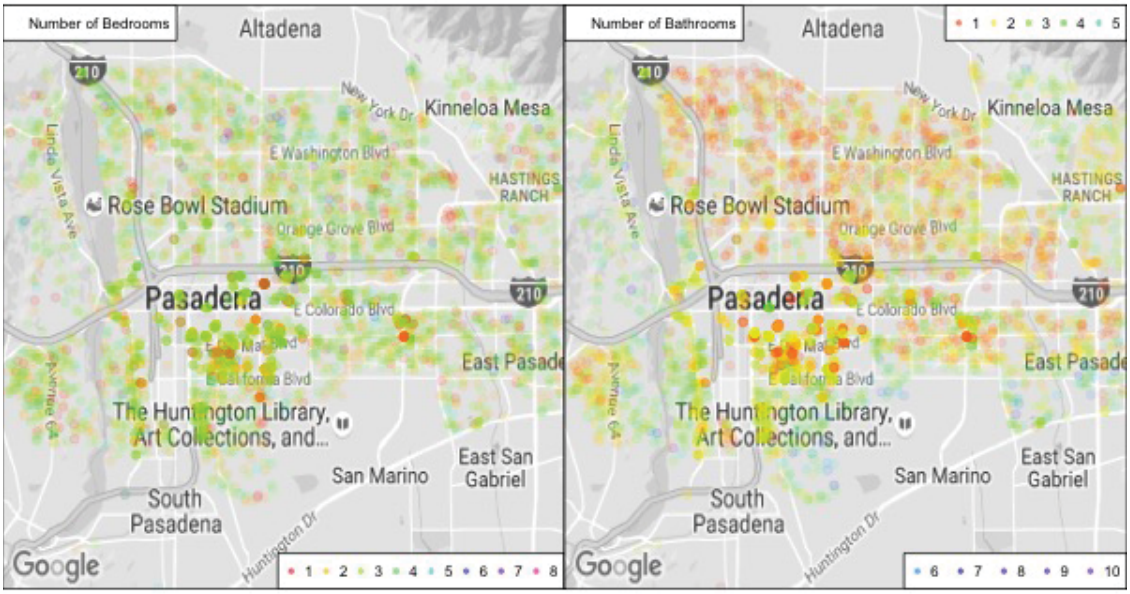
Figure 2.3: Histograms of discrete features.




ber of possible values. Treating `yearbuilt` as a discrete variable allows us to study the frequency of houses constructed each year in the histogram in Figure 2.5.

Table 2.1 summarizes the basic statistics of the two variables `beds` and `baths`, and Figure 2.3 shows their frequency distributions. Figure 2.4 displays the spatial heat maps of the number of bedrooms and bathrooms in the aggregate data set across the city of Pasadena. The `beds` variable is shown to have a roughly bell-shaped distribution in Figure 2.3. A large number of the data is concentrated at the values of one, two, three, and four bedrooms. A small proportion of houses possessing seven or eight bedrooms appear to be single family homes with relatively large square footage. As discussed, we already eliminated observations with aberrant values, e.g. houses with 14 and 23 bedrooms.

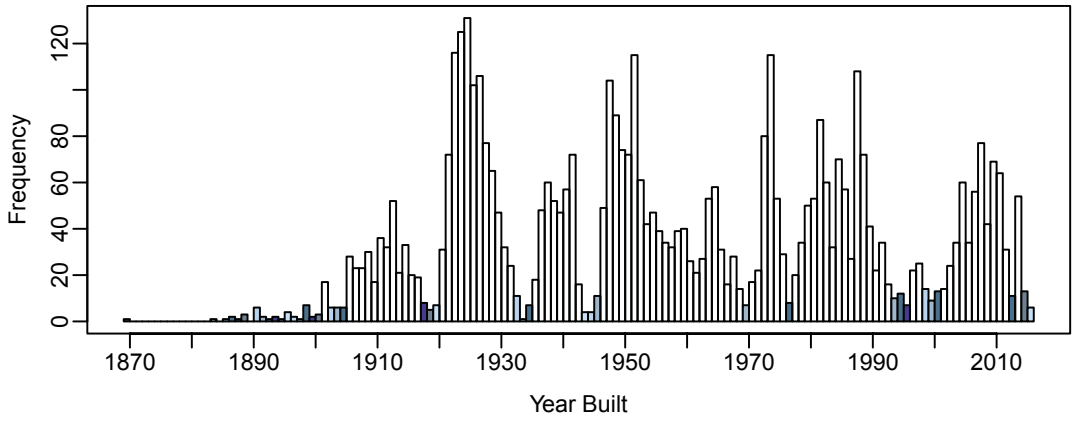
Figure 2.4: Heat maps displaying distributions of the number of bedrooms and bathrooms within Pasadena.




For the variable `baths`, an increment of a quarter unit represents an addition of a bath element such as a toilet, a sink, a shower stall, or a tub. For example, a 3/4 bath has three fixtures with no tub. It is reasonable to see in our data set, a majority of observations concentrate on values that are whole numbers such as one, two, and three. It is important to take into consideration the inaccuracy of the data, which often occurred due to rounding. People have the tendency to round up figures; for instance, a property with 1.75 baths is often recorded as a 2 bathroom house.

In the `yearbuilt` histogram (Figure 2.5), we see several clusters corresponding to booms in housing construction, which often relates to major world and economic events. For instance, the "roaring twenties" between WWI and the Great Depression was a period of intensive house-building activity. In 1922, citizens were encouraged to build, remodel, and improve their homes. The nationwide movement was known as the Better Homes Movement. The post WWII also witnessed evident booms in housing construction. A

Figure 2.5: Histogram of housing construction year.




home which was built in the late 1940s throughout the 1970s is considered as a post-war house. Based on the histogram and what we know about world events, we discretize the year built variable into a new categorical variable encoding these clusters of housing booms. Cutting points for discretization are reasonable at the years: 1865, 1920, 1933, 1945, 1975, and 2015.

We project the new categorical variable `yearbuilt.discretized` onto the map of Pasadena (Figure 2.6) to get a visual distribution of the construction time. The spatial plot exhibits apparent clusters of each construction period. Figure 2.6 implies a spreading out from the central city to outer areas over time. This expansion of housing construction is explained by the increase of housing needs over time as the population grew. Based on the previous investigation on home type distribution in Figure 2.2, we conclude that most condominiums and townhouses are newer constructions and are both located compactly in the center of Pasadena.

Figure 2.6: Spatial distribution of housing construction year.

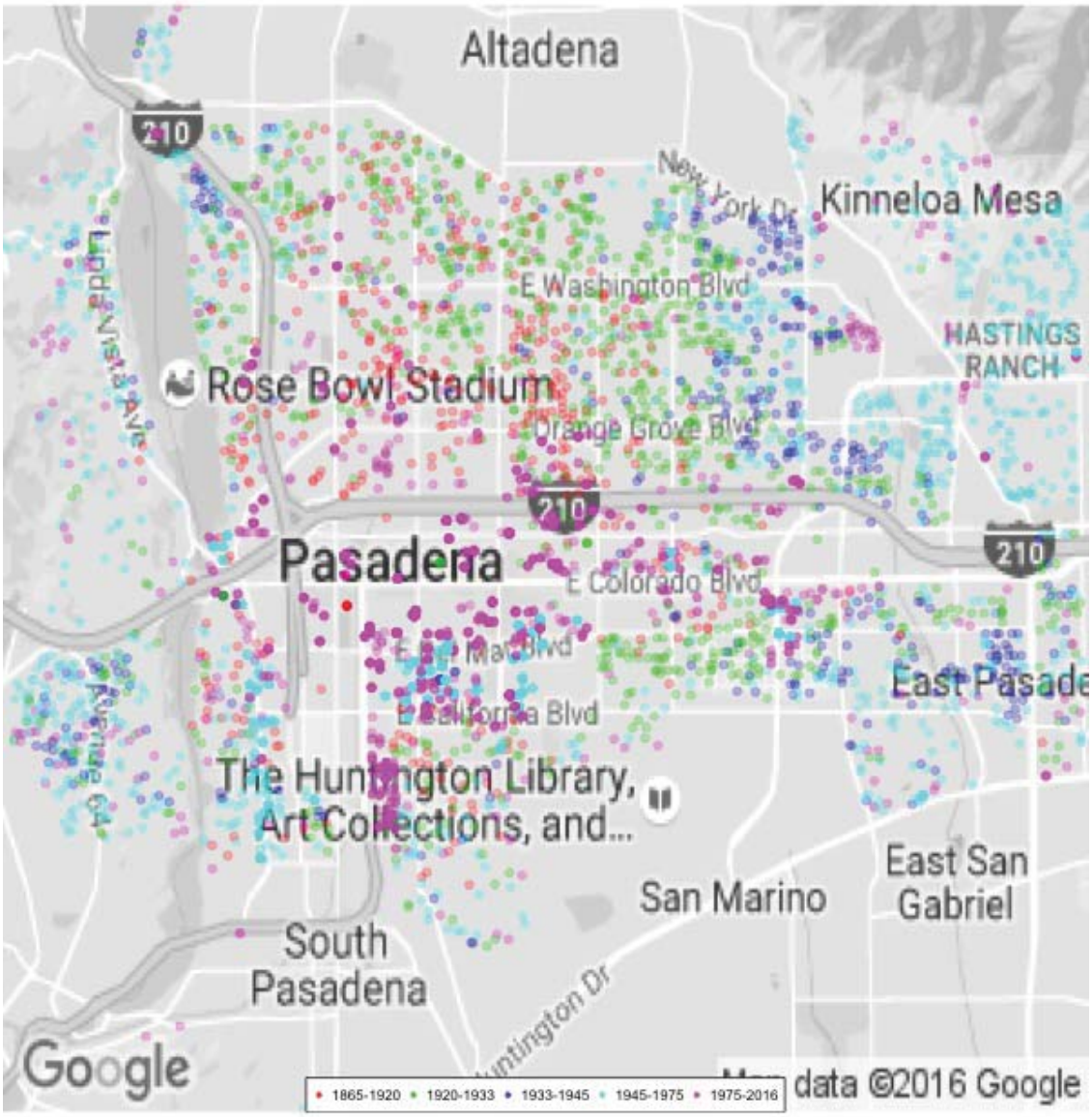


Table 2.2: Univariate summary statistics in the aggregate data set.

| | Mean | Median | S.D. | Min. | Max. |
|---|---|---|---|---|---|
| List Price (dollars) | 808,638 | 625,000 | 67,8740 | 129,000 | 10,800,000 |
| Sale Price (dollars) | 815,177 | 626,000 | 686,135 | 52,000 | 8,885,000 |
| Interior Square Footage ($ft^2$) | 1,740 | 1,494 | 1,001 | 384 | 12,711 |

**Continuous Variables**

Continuous variables include `sqft` (interior square footage of the property), `lotsize` (size of the lot measured in $ft^2$), `listprice` (the original listing price), and `lastsaleprice` (the actual selling price). Table 2.2 shows the univariate summary statistics of these continuous variables for the aggregate data (the aggregate data include SFR's, condos, and townhouses). It is important to note that we do not include `lotsize` in the aggregate data for the following reason: as a contributing factor, `lotsize` is useful for predicting the value of a single family residence, but is not meaningful for condominiums and townhouses, since the lot belongs to the commons. Hence, the inclusion of `lotsize` will be considered when analyzing the SFR subset alone, and not for the aggregate data set.

The variable `listprice` often reflects a subjective estimation of a home. We will not include `listprice` in our predictive analysis because it is not a prime indicator for the actual value of a property. Instead, one purpose of our predictions is to provide home sellers with information with which to set list prices. However, we will investigate the difference between the listed price and the actual sold price of a house from 2012 to 2015 in the Multivariate Exploratory section.

To gain an insight into the relationship between the value of a home and its location, we display `lastsaleprice` on the map of Pasadena. Figure 2.7 depicts the geographic distribution and the variability of sale prices across space. In this figure, warmer colors

Figure 2.7: Spatial distribution of sale price (for all home types).

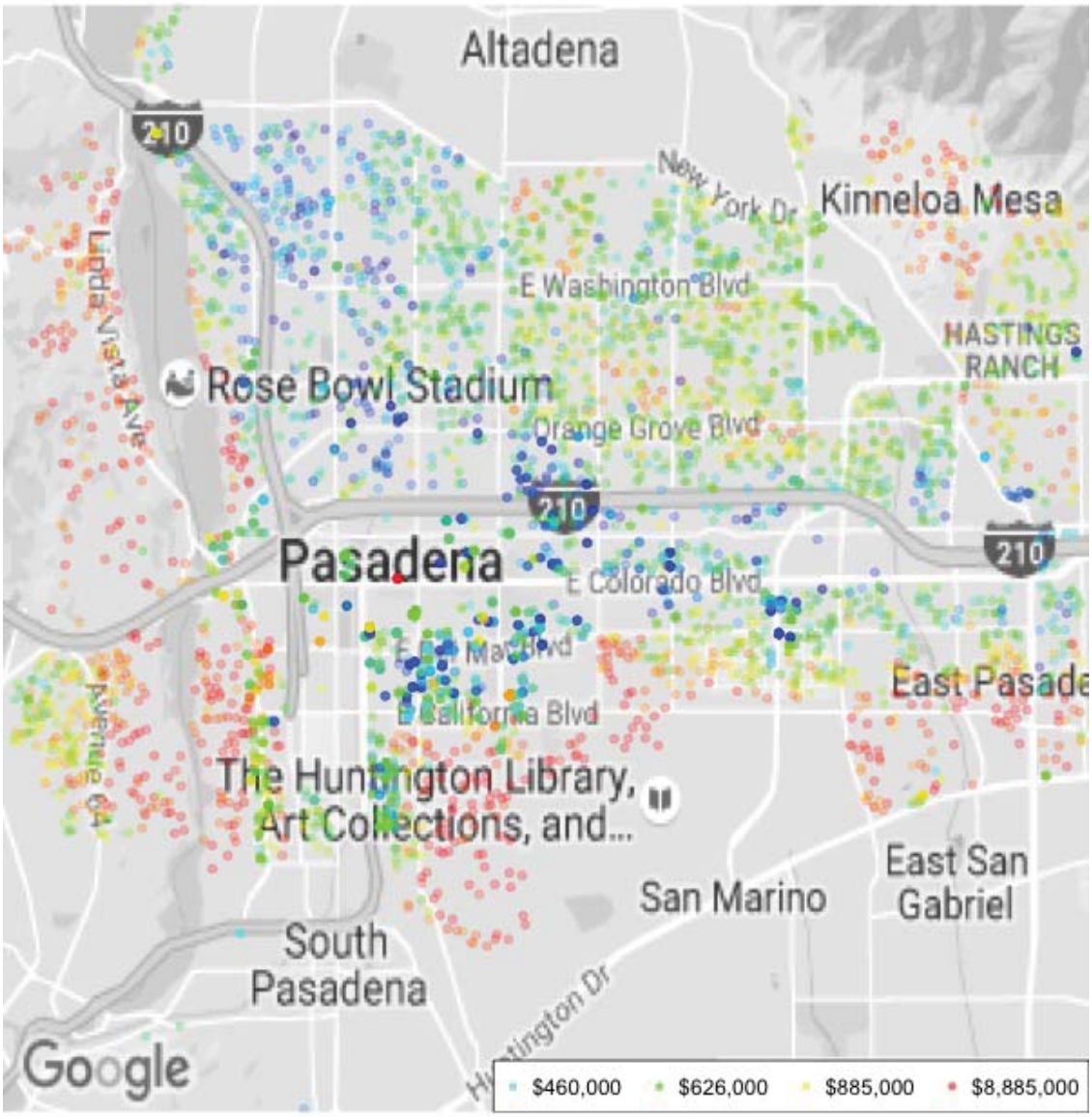


Table 2.3: Correlations between essential variables (aggregate data set)

| | Sale Price | Beds | Baths | Square Footage | Year Built |
|---|---|---|---|---|---|
| Sale Price | 1.00 | 0.56 | 0.69 | 0.88 | -0.07 |
| Number of Beds | 0.56 | 1.00 | 0.64 | 0.71 | -0.28 |
| Number of Baths | 0.69 | 0.64 | 1.00 | 0.80 | 0.15 |
| Square Footage | 0.88 | 0.71 | 0.80 | 1.00 | -0.08 |
| Year Built | -0.07 | -0.28 | 0.15 | -0.08 | 1.00 |

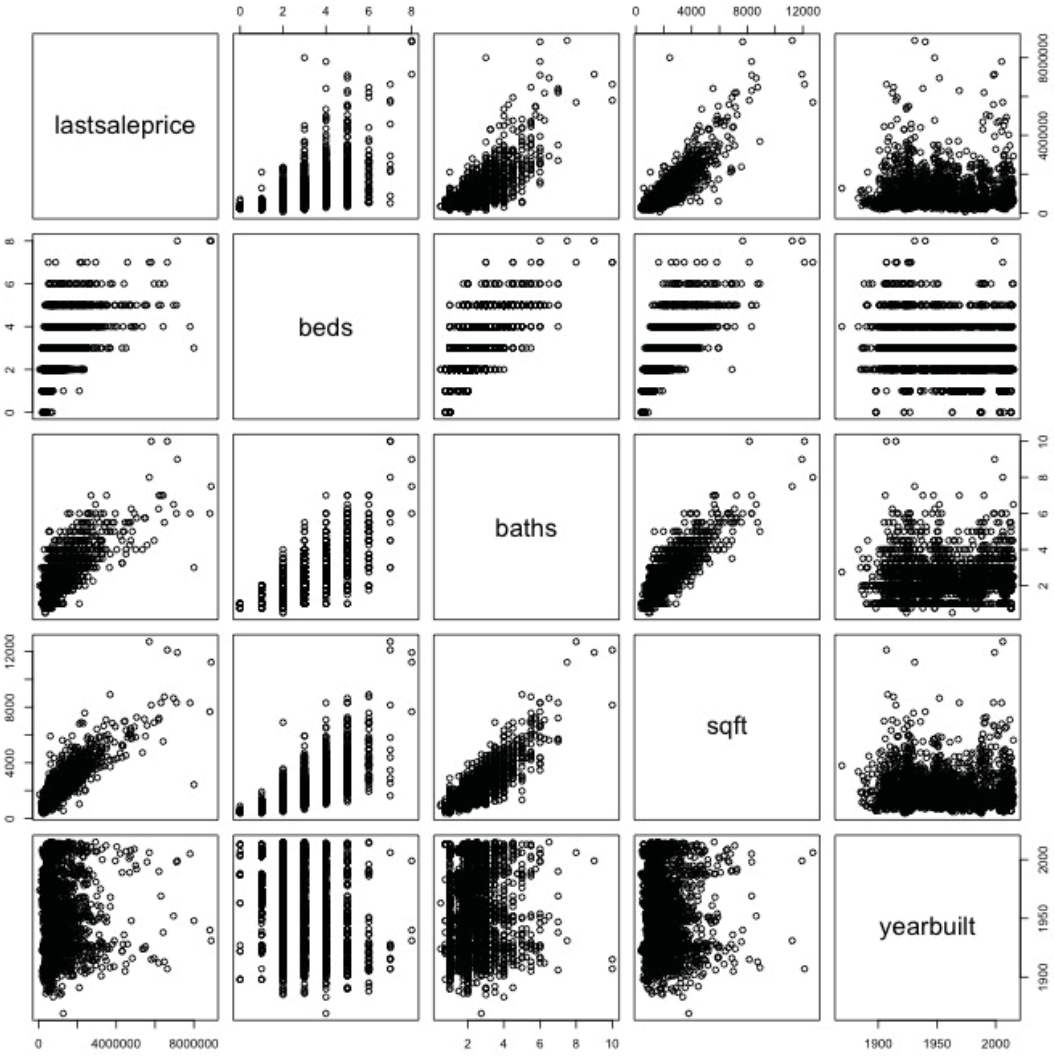
(orange and red) correspond to higher prices while cooler colors (blue and green) correspond to lower prices. The color's transparency, ranging from light to opaque, indicates housing density. Higher price properties are located in the outer districts of Pasadena and in the suburban areas. The lower price properties are found in the center of Pasadena, along the 210 highway and major traffic routes. Clearly, there is a large amount of spatial variation in housing prices.

### 2.3.3 Multivariate Exploration

This section describes the relationships between pairs of our variables. The correlation between variables is a measure that is commonly used to analyze how two variables are linearly related. In the context of housing attributes, it is not surprising to find certain variables are highly correlated with each other. For example, typically a house with more bedrooms also accommodates more bathrooms and apparently has a larger size. The relationship between the number of bedrooms and bathrooms is consistently correlated, and so is the relationship between the number of rooms and square footage.

Table 2.3 describes the correlation coefficients of our essential variables. The feature correlation plots (Figure 2.8) show a relatively high dependency of our response outcome

Figure 2.8: Feature correlations



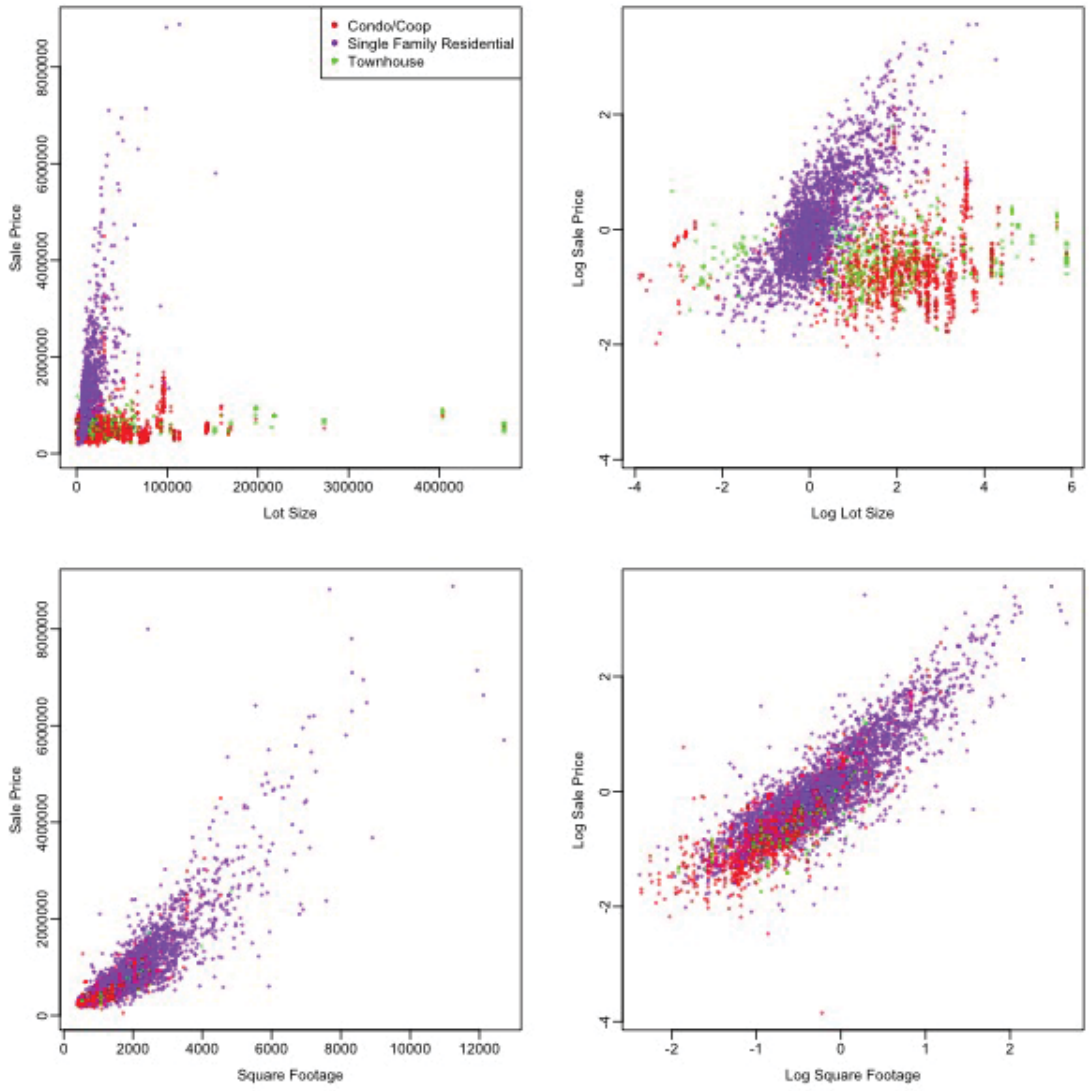
lastsaleprice on the housing attributes such as number of bedrooms beds, number of bathrooms baths, and square footage sqft. Variables that are highly correlated with the response variable are likely significant when the sample size is large, as it is for our dataset. We thus can consider the correlation coefficient as an indicator of variable significance in our regression analysis. Note that these coefficients were calculated for the aggregate data of all properties in the whole dataset. We have not included lotsize as a variable yet.

### 2.3.4 Data Subsetting

Due to the different nature of different home types, we divided the original data set into three separate subsets: Single Family Residence (SFR), condominium, and townhouse. We purposely excluded all the single family homes with missing values of lotsize since it is only meaningful and critical inside of SFR subset. This filtering process reduced the SFR subset down to 2,548 observations. The other subsets contain 1,374 condominiums and 507 townhouses.

Size generally appears to be positively correlated with price, and we can study their scatter plots to understand the distribution. As discussed earlier, lotsize is not a contributing factor to predicting the value of condominiums and townhouses as it is not meaningful and interpretable. Figure 2.9 illustrates the correspondence between the value of a property and its size (interior square footage and lot size). The top two scatter plots both depict an approximately linear relationship between lot size and sale price for single family homes (in purple). However, the log-log plot appears to have constant variance. As expected, the data for condominiums (in red) and townhouses (in green) are noisy and show no discernible pattern. The bottom two scatter plots suggest an overall increasing linear behavior for all properties, as square footage is a major factor that

Figure 2.9: Relationships between sale price and the size of the house (square footage and lot size) on original scale and log-log scale.



influences the cost and is often a primary concern of the home buyers. To increase our statistical precision, we are going to mainly analyze the SFR subset from this point on.

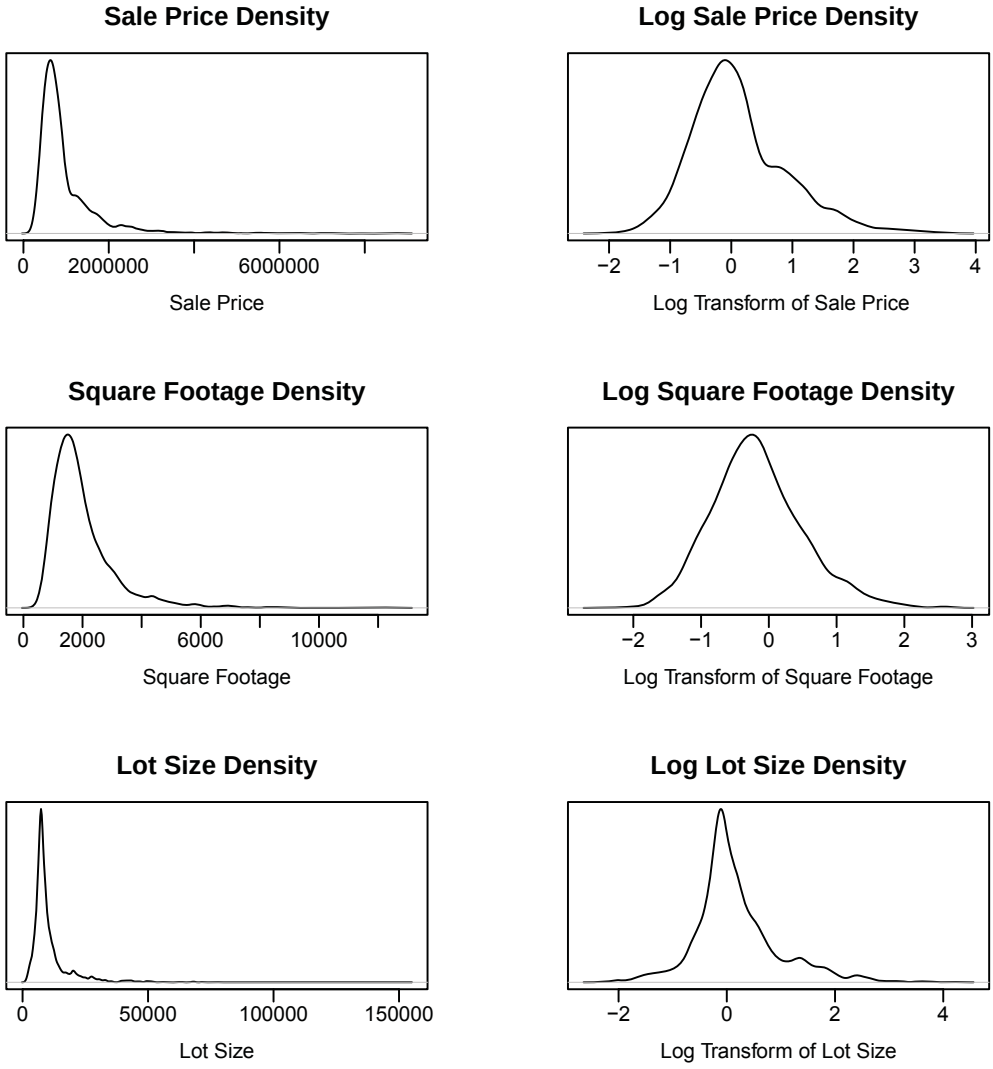
Figure 2.9 also illustrates a certain advantages when a variable is transformed to the logarithmic scale over its original scale. First, the relationship between our variables appears to be more linear in the log-log scale. Secondly, using logarithmic transformation helps reduce heteroscedasticity. As shown in the scatter plots, the data in the original scale form a fan shaped with an increased spread in the residuals. This problem is alleviated as we apply the log-log transformation to our data. The last advantage is that logarithmic transformation minimizes the leverage of outliers (e.g. houses with massive lot size and square footage). To obtain more normally distributed variables, we will perform a logarithm transformation on sale prices, square footage and lot size.

### 2.3.5 Data Transformation

Another solution to outlier removal is data transformation. Data transforming is done to minimize the impact caused by outliers. The need for data transformation depends on our modeling methods. Since one of the main methods used in this thesis is multiple linear regression, we ideally want to meet three assumptions: the relationship between explanatory variables and response variables is approximately linear; the explanatory variables are not highly skewed (which would lead to high leverage); and prediction errors are all mutually independent.

Data can have skewed distributions. Figure 2.10 displays a side by side comparison of the distributions before and after the transformation. Before being log-transformed, the density of sale price, square footage, and lot size in the SRF subset all exhibit a positively skewed distribution (the means are to the right of the median values). Monetary amounts like sale prices are commonly encountered with a skewed distribution and the

Figure 2.10: Kernel density estimate before and after logarithmic transformation.



log transform of it is often normally distributed. Applying logarithmic transformation can resolve skewness, restore the symmetry, and improve the data normality.

Our predictive analysis will be conducted mostly on the SFR subset because it is important to include `lotsize` in the prediction. Preliminary data analysis shows that the median sale price for the single family homes is fortuitously a nice round value of $750,000. Therefore, we decide to set $750,000 as the center value (0 value) for our log-transformed price. Similarly, for the log transformation of square footage, we pick 0 to correspond to 2,000 square feet. For lot size, 0 corresponds to 8,000 square feet.

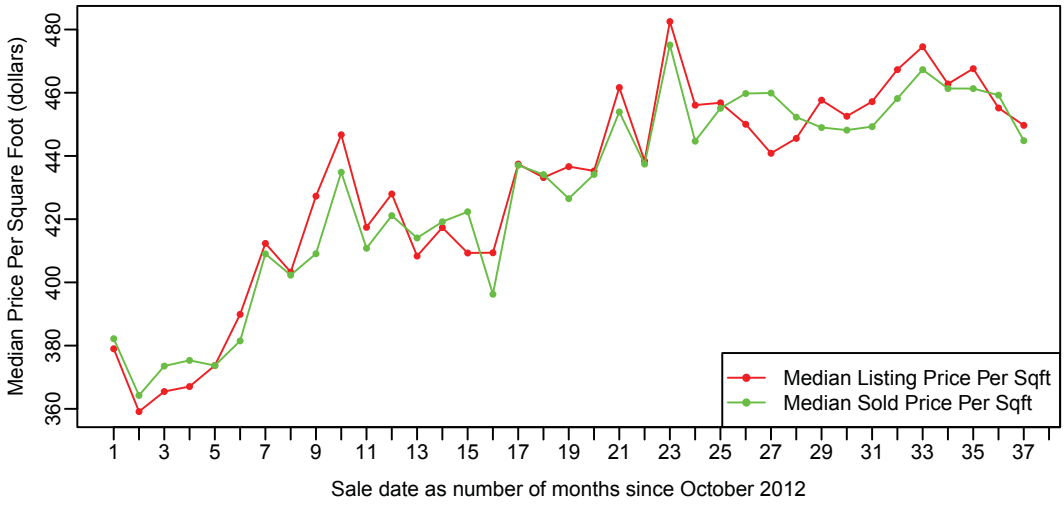
It is also notable to mention here that our logarithmic transformations are done on base 2. We prefer this particular base for simple interpretability purposes. On the $\log_2$ scale, a doubling or a reduction to 50% is translated to change of +1 or -1. These fold changes are relatively simple to interpret. Moreover, this scale produces more fine grained values than using higher bases (e.g. $\log_{10}$).

### 2.3.6 New Variables Introduction

As explained in the previous section, we have now created three new variables including: logarithmic transformation on the sale price (`log2price`), square footage (`log2sqft`) and lot size (`log2lotsize`).

One of the new variables is created based on a general understanding of the housing market. Our data are available for a 3-year period starting from October 18, 2012 to October 18, 2015. We use the existing variable `lastsaledate` to create a new variable called `duration`. This variable quantifies the number of days since October 18, 2012 to the day the property is officially sold. It is important to note that `duration` does not denote the concept of "days on market" in a traditional real estate listing context. The variable `duration` is merely a calendar time measurement starting from October 18, 2012

Figure 2.11: Monthly house price variation for the aggregate data set.




and is only meaningful in the scope of our data set.

We then divide `duration` into intervals with length of 30 days to represent month-long periods. This variable can account for the temporal fluctuations within the prices and is helpful to be included in the analysis. Figure 2.11 shows the trends in price variation in both listing and selling price of the aggregate data set. Each datapoint in the line chart marks the median price per square foot in every 30-day cycle of listing and selling prices. The chart suggests an overall increasing trend in housing transaction prices. This trend indicates the rise in prices in the 3-year period of October 2012 to October 2015. The median price per square foot for listing price is generally higher than that for sold price between 2012 and 2015. The minimal of median sold price occurred in November 2012. Since then, the median sold price per square foot gradually rises by about 25% over a three-year period.

# Chapter 3

# Data Modeling Methods

In this chapter, we discuss the concept of regression and related methods and the role they play in predicting future housing prices. Regression analysis is a statistical technique used to determine the contributing effect of a set of explanatory variables on the response variable. Within the scope of this thesis work, regression analysis is an appropriate approach because it allows us to assess the contribution of each variable individually. Specifically, one of our goals is to determine which attributes have a significant impact on the overall price of a home. Furthermore, we will implement regression methods to build predictive models on housing prices and evaluate their performance.

We introduce three different types of regression techniques in this chapter: linear regression, nonlinear regression, and regression trees. Some particular methods in each type are also presented, such as splines and Generalized Additive Models (GAM) in the nonlinear regression; bagging, random forest and K-Nearest Neighbors (KNN) in the regression trees.

## 3.1 Linear Regression

Linear regression is the simplest and earliest predictive method, which involves using a linear combination of predictors (independent variables) to estimate a continuous outcome (dependent variable). The objective of linear regression models is to estimate the regression coefficient vector $\beta$ in a way that the *mean squared error* (average squared discrepancy between the observed and predicted outcome values) is minimized.

Given a dataset of $n$ observations, a linear regression model with $p$ predictors is written as:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{3.1}$$

where $y_i$ represents the continuous response for the $i^{th}$ observation, $\beta_j$ is the regression coefficient for the $j^{th}$ variable, $x_{i,j}$ represents the $j^{th}$ variable value for the $i^{th}$ observation, and $\varepsilon_i$ is the random error term. We wish to estimate the $\beta_0$, $\beta_1$,...,$\beta_p$ by obtaining

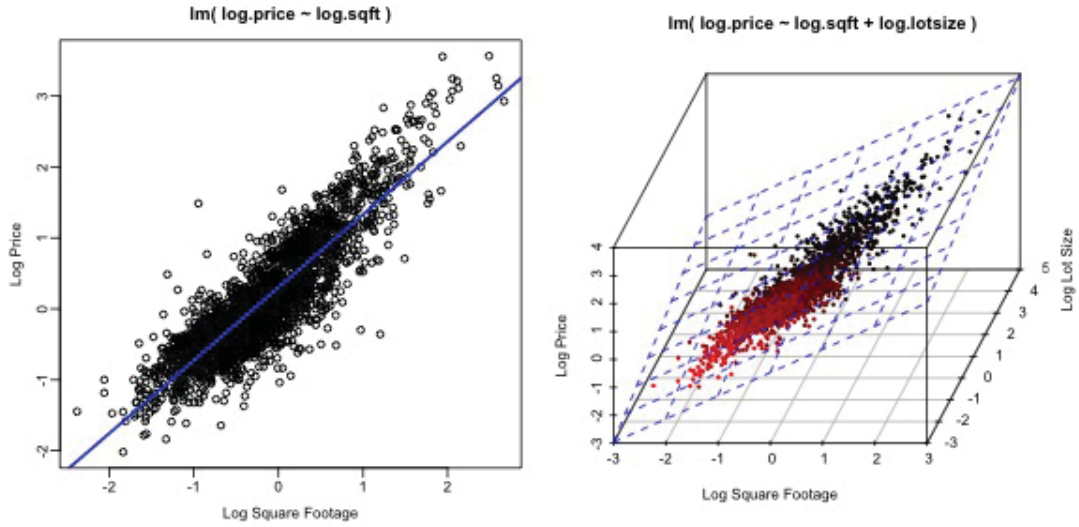
$$\hat{y_i} = b_0 + b_1 x_1 + \cdots + b_p x_p \tag{3.2}$$

in such a way that $\sum_i (y_i - \hat{y}_i)^2$ is minimized. The $b_j$, for $j = 0, 1, \ldots, p$, terms are called the *regression coefficients*. Figure 3.1 provides two examples of linear regression models generated by package `Scatterplot3d` (Ligges and Mächler, 2003). The first plot is a simple linear regression model, in which we fit a simple line onto the data based on one single predictor. The second plot fits a plane (for two predictors: log-transformed square footage and log-transformed lot size) to predict the value $\hat{y}_i$ of the outcome: log-transformed sale price.

In linear regression analysis, there are several assumptions: the error terms $\varepsilon_i$ are independent, uncorrelated and normally distributed with mean of zero and constant variance $\sigma^2$ (a.k.a. homoskedasticity).

The advantage of the linear regression model (3.1) is that it has high interpretability

Figure 3.1: Linear regression models.




of the coefficients and a good predictability in cases of small sized training data sets (Hastie et al., 2001). A weakness of linear regression is that it is sensitive to outliers, which is a common tendency of most datasets. The performance of the linear model is often influenced by the presence of a small number of outliers in the dataset.

While the linear regression method is a rather simplistic way to capture the complexity of housing predictions, there are fundamental concepts in linear regression that are used to develop other regression techniques. Many modern statistical learning approaches, such as splines and generalized additive models, can be considered as generalizations or extensions of linear regression.

## 3.2 Nonlinear Regression

As linear regression models have many advantages such as computational simplicity and high interpretability, linearity may not be an appropriate assumption for the relationship between predictors and the predicted outcome in many cases. There are a number

of regression models which we can use to determine the specific characteristics of the nonlinearity in the data set.

### 3.2.1 Splines

Spline methods use piecewise polynomial functions of degree $d$ in a variable $x$ to fit the model. The domain of the input variable is divided into continuous intervals with knots, which are the points where the piecewise functions connect. If we place $k$ knots, then we are fitting $k+1$ different piecewise segments to the data. Using more knots leads to a more flexible piecewise polynomial. A spline of degree $d$ is formed by connecting polynomial segments of degree $d$ so that: the function is continuous; the function has $d-1$ continuous derivatives at each knot; and the $d^{th}$ derivative is constant between knots. Spline functions possess a high degree of smoothness at the knots and are very useful for modeling arbitrary functions. The basic spline model is:

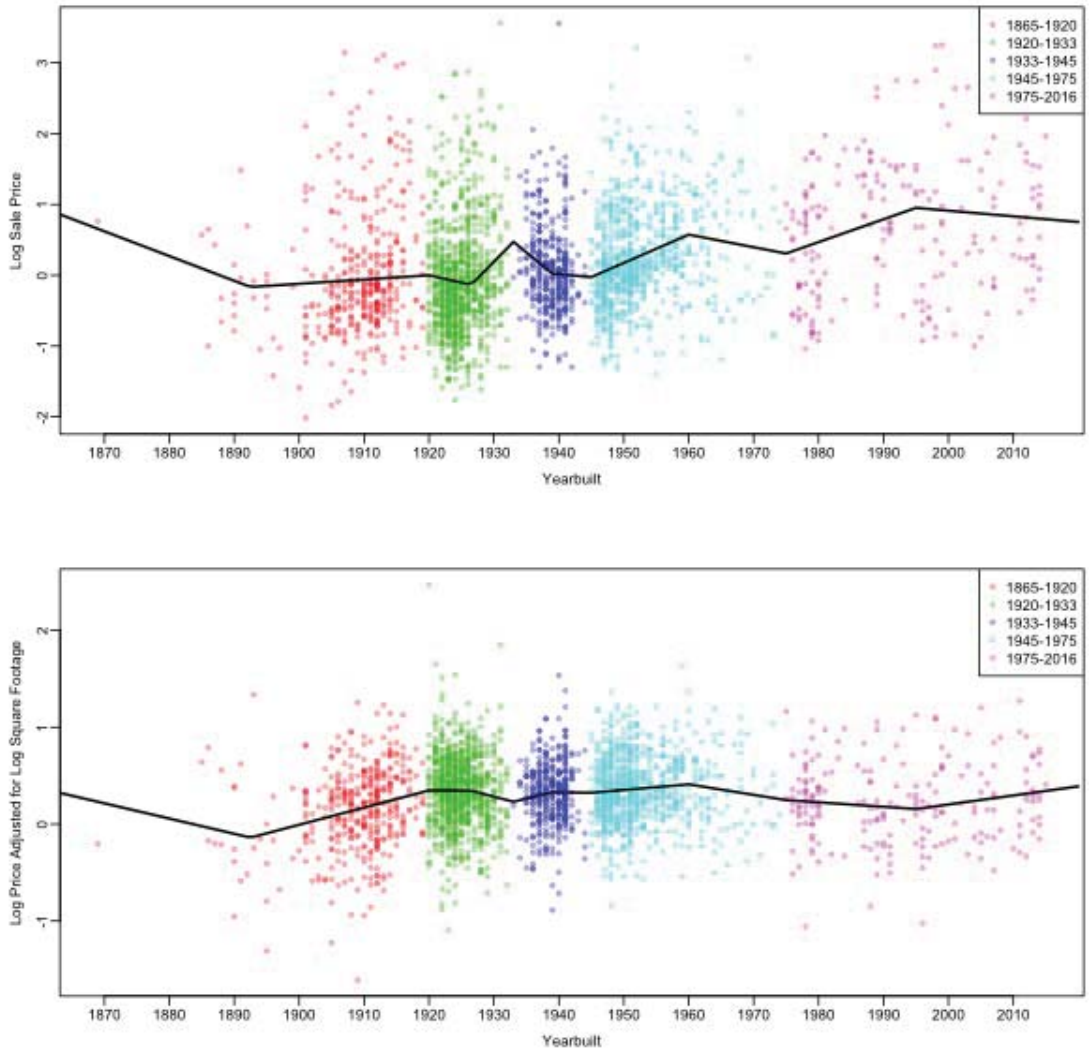
$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{k+3} b_{k+3}(x_i) + \varepsilon_i \tag{3.3}$$

where the basis functions $b_1(.), b_2(.), \ldots, b_k(.)$ are fixed and known. The basis functions are $b_j(x_i) = x_i^j$ for polynomial regression, and $b_j(x_i) = I(c_j \leq x_i < c_{j+1})$ for piecewise constant functions.

#### Linear Spline

A linear spline is a piecewise linear function which is essentially piecewise polynomials of degree 1. These spline functions are everywhere continuous and have slope changes at the knot locations. Figure 3.2 shows examples of piecewise linear functions fitted to the house price data. Both models use linear splines to model the relationship between year of construction and log price. The timeline was divided into different intervals cor-

Figure 3.2: Top panel shows the linear splines when using only `yearbuilt` to predict log price. Bottom panel is the partial residual scatter plot showing linear splines when using `yearbuilt` to predict log price while adjusting for `log2sqft`.



responding to social and economic events as mentioned in section 2.3.2. There are nine knots placed at the boundaries between periods and at the midpoints of the periods: 1865-1920, 1920-1933, 1933-1945, 1945-1975, and 1975-2015. The second model includes a (linear) adjustment for log square footage. In each plot, we can see how slope changes at the knots are chosen via the least squares criterion to match the scatter plot optimally.

The first panel seems to indicate that newer homes, especially those built in the 1990's or later, have higher average prices. However, from the second panel, which includes an adjustment for square footage, we see that this relationship appears to be explained by newer homes being larger on average. Indeed, in the second panel it appears that homes built between the early 1900's and 1960 command the highest average prices for a given square footage.

### Cubic Spline

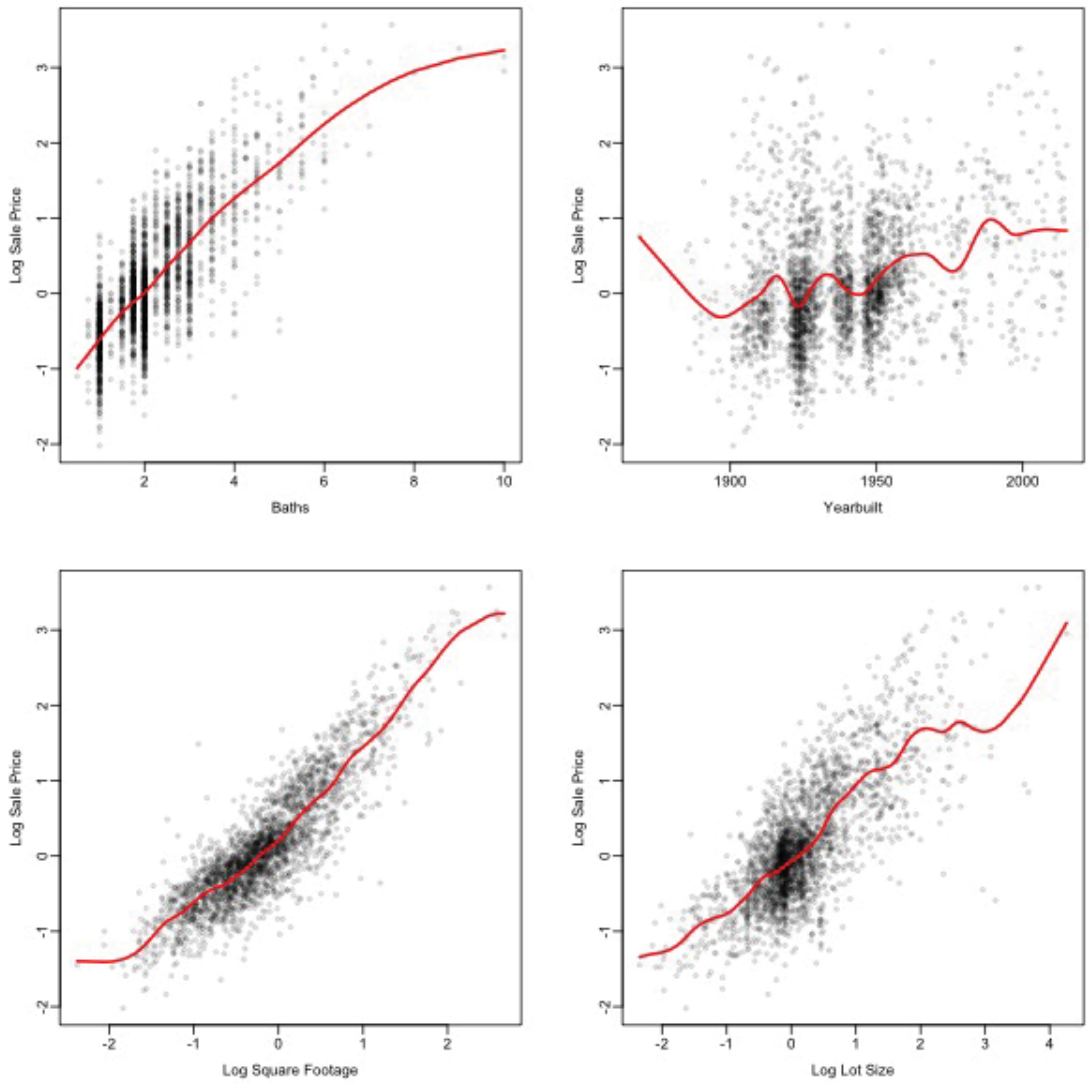
The smoothness of the piecewise interpolation can usually be improved by varying not only the slopes but also the concavities or higher derivatives at each knot to match the data values. Cubic splines are a good candidate for such cases. In fact, the most popular spline is the natural cubic spline, which fits cubic polynomials between knots but linear models outside of the first and last knots. The second derivative of each polynomial is commonly set to zero at the endpoints. This is a definition of a natural cubic spline as this provides a boundary condition that completes the system of equations.

### Smoothing Spline

Another approach to fitting a smooth curve to a set of noisy observations is to use the smoothing spline. The goal is to find the function $g(x)$ that minimizes:

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt \tag{3.4}$$

Figure 3.3: Smoothing splines models of log price using the `smooth.spline()` R function.



The first term is the usual least squares criterion, and the second term is a regularization term that penalizes variation or lack of smoothness in the curve. The parameter $\lambda$ is a nonnegative tuning parameter that controls the amount of smoothness. The function $g$ that minimizes the quantity in equation (3.4) is known as a smoothing spline.

The smoothing splines are developed to control the complexity of the fit by establishing a trade-off between the closeness of the data and the smoothness in the function. The integral is evaluated over the range of the $x_i$. A smoothing spline has a knot at each data point, but introduces a penalty for lack of smoothness. As $\lambda \to 0$ (no smoothing), the smoothing spline converges to a function that interpolates the data, and as $\lambda \to \infty$ (infinite smoothing), we get a straight line fitted by ordinary least squares. Figure 3.3 shows the results from fitting smoothing splines to the log transformation of price by using `smooth.spline()` function in R.

### 3.2.2 Generalized Additive Models

A generalized additive model (GAM) written as:

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \varepsilon_i \qquad (3.5)$$

is an extension of the multiple linear regression model given in equation (3.1), in which each linear component $\beta_j x_{ij}$ is replaced by a sum of smooth functions, $\sum_{j=1}^{p} f_j(x_{ij})$, to allow the non-linear relationship between each predictor and the response variable. It is called an additive model because each separate $f_j$ for each $x_j$ is calculated and then all of their contributions are added together .

GAMs were originally introduced and popularized by Trevor Hastie and Robert Tibshirani in 1990 to integrate properties of generalized linear models with additive models (Hastie and Tibshirani, 1990). It is applicable in many areas of prediction. Generalized

Figure 3.4: The estimated partial response functions for the additive model.

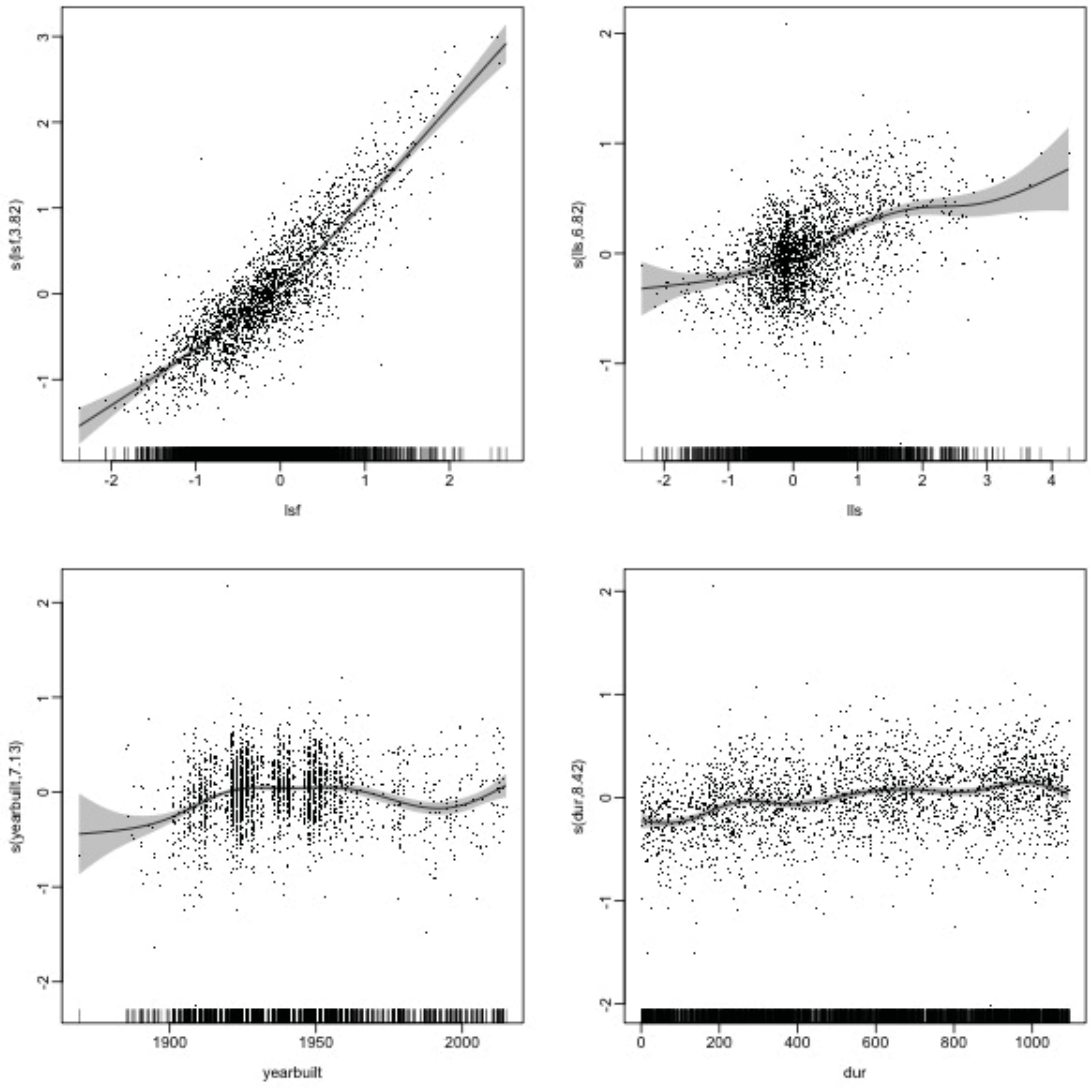


Figure 3.5: The joint smoothing of longitude and latitude (perspective map).

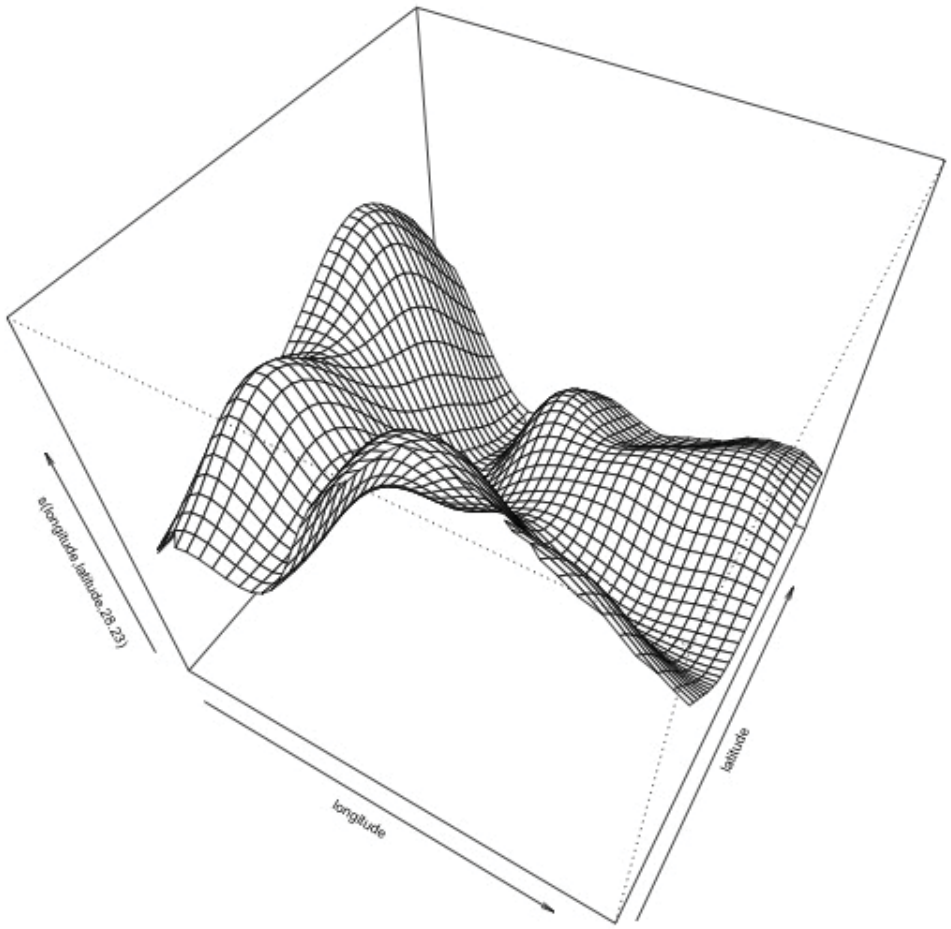
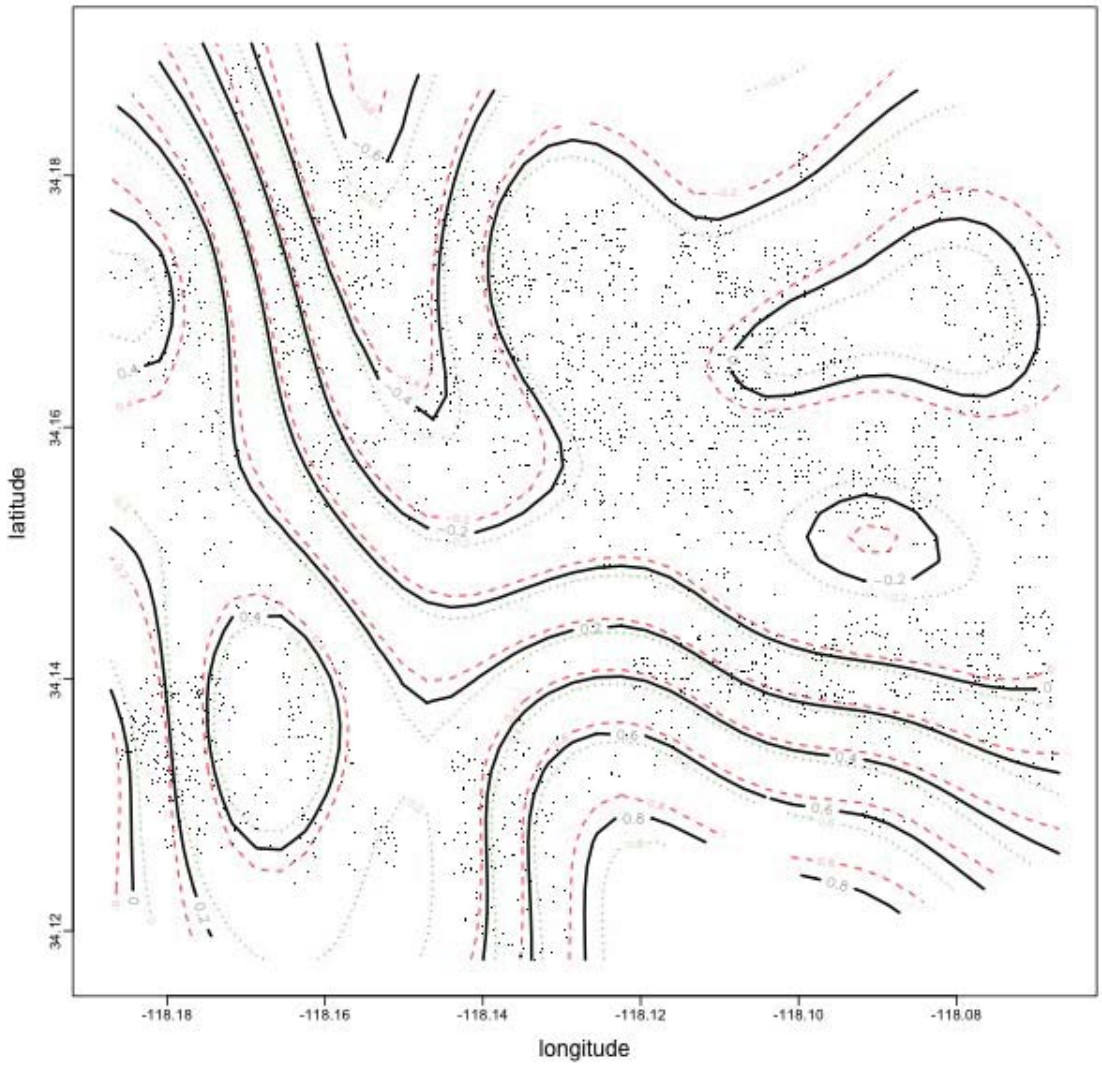


Figure 3.6: The joint smoothing of longitude and latitude (contour Map).



additive models are suitable for exploring the data set and visualizing the relationship between a univariate dependent variable, $y$, to some independent variables, $x_i$. In an additive regression model, the marginal impact of a single variable is independent of the values of the other variables in the model. The additive property allows us to study the contribution of each $x_j$ on $y$ individually while holding all of the other variables fixed. Hence, GAMs provide a useful representation in inference because we can simply look at the output of the model and make simple statements about the effects of the predictive variables.

In Figure 3.4 we fit an additive model with four variables: the log of square footage `log2sqft`, the log of lot size `log2lotsize`, the year of construction `yearbuilt`, and `duration` (the number of days past the date October 18, 2012 on which the house was sold), using the `gam` function from the `mgcv` package (Wood, 2006). Each plot represents the variable's response shape, independent of the other variables, in relation to the estimates of the log price in the multivariate model. The degree of smoothing is indicated in the y-axis label. The tick marks along the horizontal axes are rug plots that show the distribution of the observed input values. The shaded region represents $\pm 2$ standard errors around the response curve and is wider where there are fewer observations. The vertical scale is logarithmic.

Figure 3.5 and 3.6 are the corresponding perspective and contour map generated by the `gam` package (Hastie, 2015) to represent our inferences for the two spatial variables. The joint smoothing of the spatial coordinates `latitude` and `longitude` is incorporated to display the predictions of house prices. However, the plots lack some degree of precision when locating the peaks and troughs within Pasadena. This motivates superimposing estimates onto maps using colors to represent prediction values. Hence, heat maps superimposed on geographic maps are demonstrated later to be more effective for relat-

ing inferences to specific locations, and are thus used primarily in the remainder of this thesis.

## 3.3 Regression Trees and Other Predictive Models

Tree-structured classification and regression are nonparametric computationally intensive methods that have greatly increased in popularity during the past dozen years. They can be applied to data sets with large number of observations and variables. Tree-structured methods are highly resistant to outliers.

### 3.3.1 Basic Regression Trees

As a special form of non-linear regression, a regression tree is recursively constructed through a binary partitioning process. In the regression tree method, the target variable is predicted through an iterative process of bifurcating the data into branches with leaves and nodes. Each branch represents an attribute value that either leads to a decision node or splits into a different branch. At each split point, the error between the predicted value and the actual values is squared and then added up to get a Residual Sum of Squares (RSS). The algorithm selects the branch that minimizes the RSS.

Given our dataset with four features included as potential predictors (number of bedrooms, the number of bathrooms, the year built of a property, and lot size), Figure 3.7 is a crude example of a regression tree which predicts the price of houses based on the input variables. We note that not all features are used by the tree, e.g., the number of beds does not contribute to the decision nodes. Figure 3.8 is another example in which the response variable (log price) is determined by the geographic coordinates of the property, i.e. longitude and latitude. This regression tree can also be demonstrated on a spatial map

Figure 3.7: Regression tree on sale price (dollars) with four inputs: `beds`, `baths`, `yearbuilt` and `lotsize`.

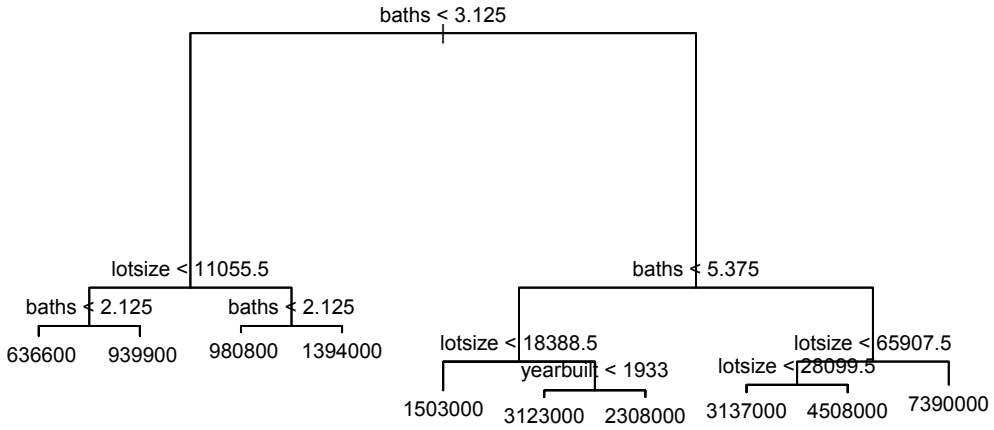



Figure 3.8: Regression tree on sale price (log-scale) based on spatial coordinates: `latitude` and `longitude`.

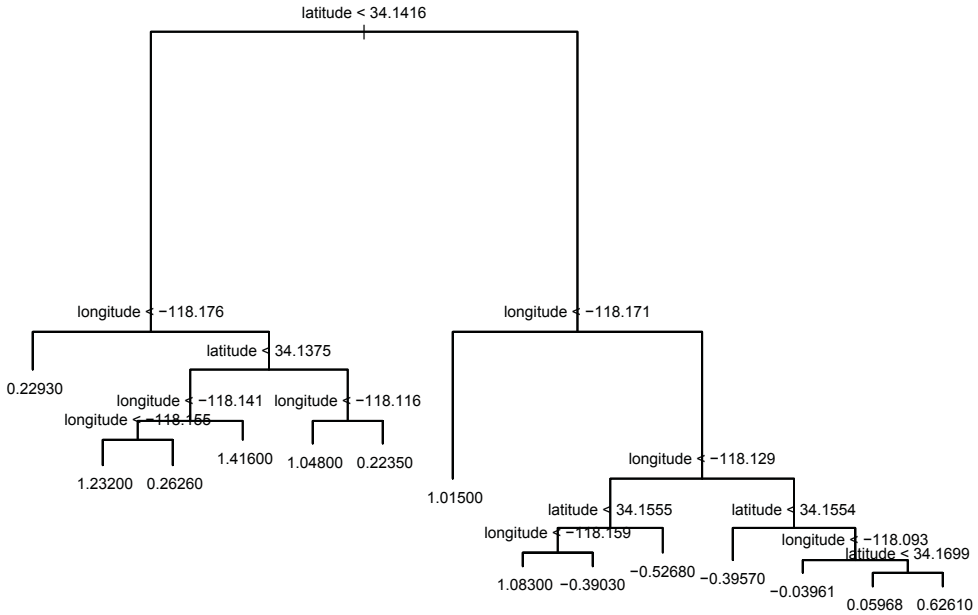
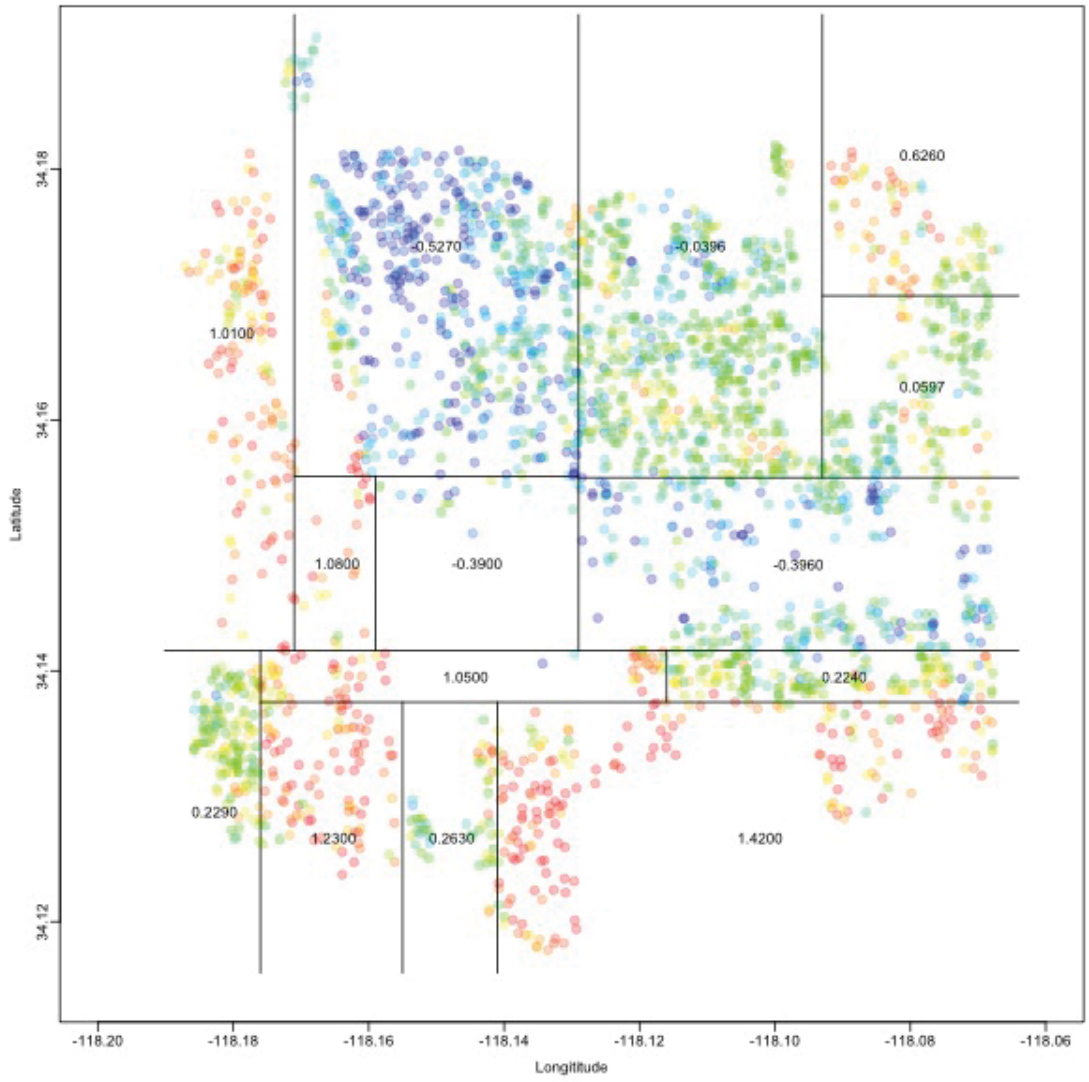


Figure 3.9: Map of log prices on single family residences (color-coded by decile, red indicating most expensive, blue indicating least expensive), and their partitions by regression tree.



(Figure 3.9) to illustrate the same decision splits as shown in Figure 3.8. It is more visual to see the partition of log price is divided into clusters of different colors corresponding to different price quantiles.

The process of building a regression tree is described in these following steps. We divide the set of possible values for $X_1, X_2, \ldots, X_p$ into $J$ distinct and non-overlapping regions, $R_1, R_2, \ldots, R_J$. For every observation that falls into the region $R_j$, we make the same prediction, which is simply the mean of the response values for the training observations in $R_j$. Our goal is to find regions $R_1, \ldots, R_J$ that minimize the RSS given by:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \tag{3.6}$$

The procedure then recursively continues splitting each partition into smaller groups as the method moves down each branch in such a way that at each stage, the split that minimizes the RSS over all possible splits is the one that is chosen. A node is called a terminal node if the sum of squared deviations from the mean in that node is below some threshold.

Regression trees possess a sequential structure that is not only accurate but also simple to compute and interpret. Another advantage is that they are non-parametric and could be applied for large datasets without prior knowledge of the relationship between the predictor variables and the predicted outcome.

### 3.3.2 Bagging and Random Forest

Bootstrap aggregation, commonly known as *bagging* for short, was invented by Leo Breiman to be a simple yet effective ensemble algorithm for improving the accuracy and reducing the instability of single decision trees (Breiman, 1996). The idea is the same as the following result from elementary statistics: Given a set of $n$ independent observations

$Z_1, \ldots, Z_n$, each with variance $\sigma^2$, by taking the average of the set of $n$ observations, we reduce the variance to $\frac{\sigma^2}{n}$.

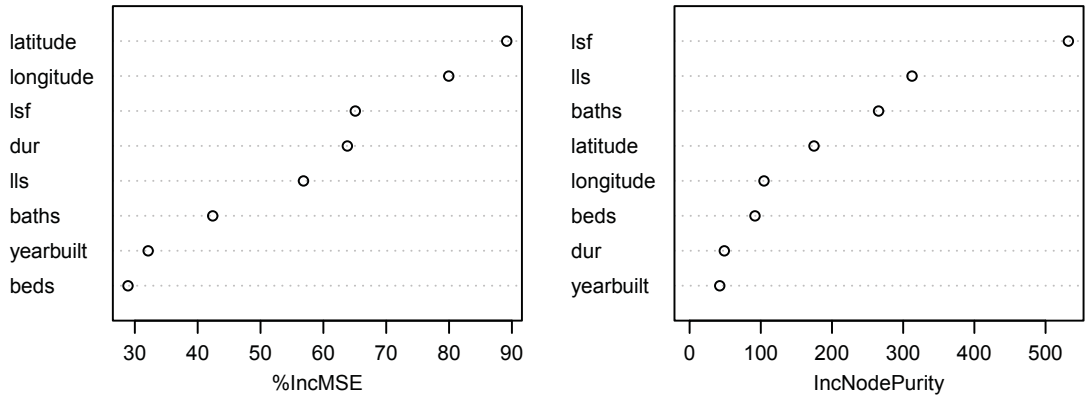
To increase the prediction accuracy of a high-variance statistical learning model, we build separate prediction models $\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x)$ on B separate training sets, and then average the resulting predictions. Since it is not practical to always obtain multiple training sets, bagging generates several new training sets by using random bootstrap sampling from the original dataset with replacement. Then, a set of tree models can be trained independently by applying the regression tree algorithm on the new training sets. The predicted responses are finally calculated by averaging all the models $\hat{f}^{*b}(x)$:

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}(x) \tag{3.7}$$

Random forest is a closely related method to bagging in which we also construct a large number of decision trees on bootstrapped training samples. To construct a random forest, every time a split in a decision tree is considered, we take a random sample of $m$ predictors from the full set of $p$ predictors, and only splits on this subset of predictors are allowed at this stage. This process is repeated for each node until the tree is large enough. The main difference between bagging and random forests is the choice of predictor subset size $m$. Random forests can work with a very large number of predictors and reduce bias of fitted values and estimated splits. Additionally, random forests provide an improvement over bagged trees by decorrelating the trees since each predictor will have several opportunities to be the predictor defining a split. This process improves the reliability and reduces variablility in the average of the resulting trees.

We can assess the importance of predictor variables in a random forest model by random permuting a variable value vector, and observing the resulting increase in MSE. The first graph of Figure 3.10 shows the average increase in percentage of the MSE when a variable is assigned values by random permutation. The second graph indicates

Figure 3.10: Dot chart of variable importance as measured by a random forest.




the node purity measured by Gini Index, which is the the difference on RSS every time a split takes place at a certain variable. Note that we obtain different rankings for different predictors in each graph.

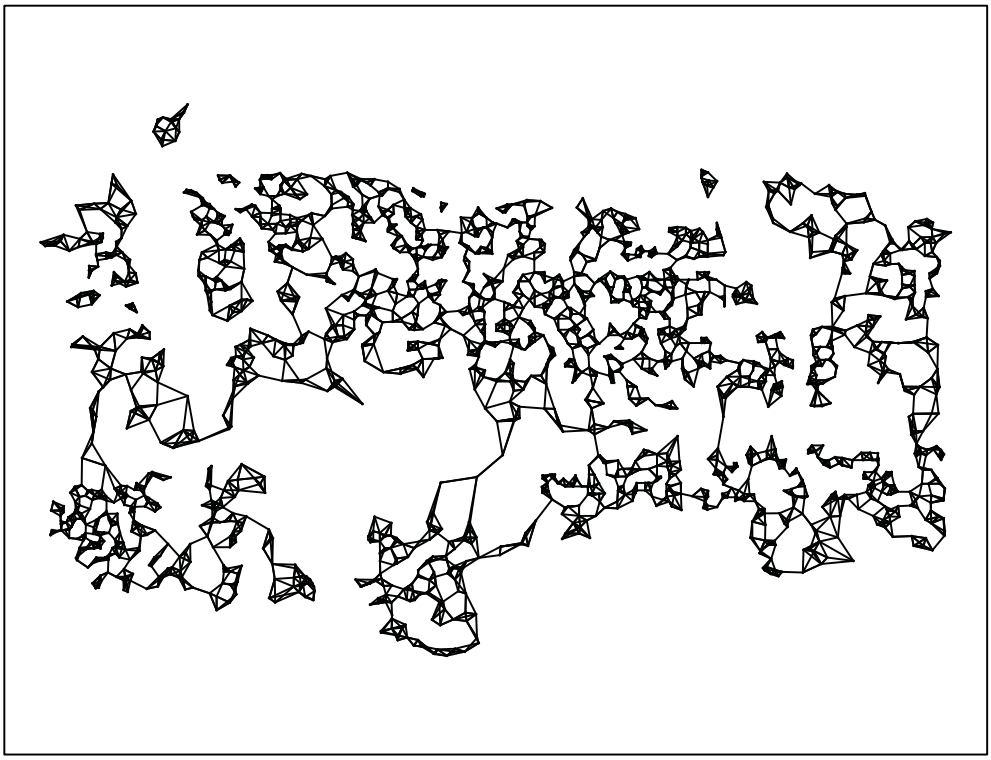
### 3.3.3 K-Nearest Neighbors Regression

The K-Nearest Neighbors (KNN) methods is a non-parametric method used for classification and regression, whereby predictions at each point are determined by the values of the $K$ closest observations in the data set, where "closeness" is measured by a distance function on the feature space. For example, Figure 3.11 shows a map connecting each home in our dataset to its $K = 4$ nearest neighbors determined by physical location (latitude and longitude) within Pasadena. Typically, the Euclidean distance is used as the distance measure:

$$d = \sqrt{\sum_{j=1}^{p} (x_{aj} - x_{bj})^2} \tag{3.8}$$

Given a new input $x_0$, the prediction $\hat{f}(x_0)$ is created by identifying the set $N_0$ con-

Figure 3.11: $K$ nearest neighbor neighborhood structure for the Pasadena housing price data, using $K = 4$.



sisting of the $K$ nearest neighbors and averaging their response values:

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i \tag{3.9}$$

In KNN regression, the KNN algorithm is used for estimating continuous variables. The approach discussed in this thesis uses a weighted average of the $K$ nearest neighbors, weighted by the inverse of their distance. The algorithm follows three steps. First, the Euclidean or Mahalanobis distance of the observed value is computed. Then the observations are ordered by increasing distance. After that, an optimal number $k$ of nearest

neighbors is calculated based on RMSE. Finally, the weighted average is determined by the inverse distance with the $k$-nearest neighbors. The latitude and longitude information allows the calculation of distances between properties, which is an essential step for the two KNN models considered in this study.

# Chapter 4

# Results

In this chapter, we discuss the application of the methods from Chapter 3 to the single-family residence subset of the Pasadena home price data set from Chapter 2. In particular, the dataset consists of 2,548 sale records for single-family homes in Pasadena from October 18, 2012 to October 18, 2015. A parsimonious set of six property attributes (lot size, square footage, number of bedrooms, number of bathrooms, year built, duration) along with spatial coordinates (latitude and longitude) of the properties are used.

We have three aims. First, we compare the accuracies of the models discussed in the previous chapter for predicting the sale price of single family homes. Next, we discuss which variables appear most strongly associated with housing prices, and comment on the magnitude and direction of these associations. Finally, we discuss how spatial location within Pasadena appears to be related to housing prices, and discuss how spatial variation in housing prices is reduced as we include additional predictors.

## 4.1 Accuracy of Housing Price Prediction

In this section, we compare various methods by testing different models on the single-family residential data set. To assess the quality of our prediction methods, the performance of each model is evaluated by two criteria: the median absolute error and the median percent error. Both values are measures of the predictive power of the model.

If $\hat{y}_i$ is the predicted value and $y_i$ is the corresponding true value of the log sale price of the $i^{\text{th}}$ home, then the *median absolute error (MAE)* estimate is defined as:

$$\text{MAE}(y, \hat{y}) = \text{median}(\mid y_1 - \hat{y}_1 \mid, \ldots, \mid y_n - \hat{y}_n \mid) \tag{4.1}$$

Choosing a model that minimizes the median absolute residuals is intuitive because median absolute error is robust to outliers and can be interpreted as the distance around true values from our predictions.

It is also important to mention that our data for single family homes are using the $\log_2$ version of sale price, lot size and square footage. The *median percent error (MPE)* is obtained by taking the median absolute error of the $\log_2$ of sale price, raising 2 to that power, and then subtracting 1:

$$\text{MPE}(y, \hat{y}) = 2^{\text{MAE}(y, \hat{y})} - 1 \tag{4.2}$$

### 4.1.1 Non-spatial Models

Non-spatial models estimate the value of a home solely based on the characteristic features of the house. We use `log2lotsize`, `log2sqft`, `beds`, `baths`, `yearbuilt`, and `duration` (which is an absolute calendar time measurement equal to the number of days past the date October 18, 2012 on which the house was sold) as the predictor variables to predict `log2price`. Different non-spatial modeling approaches are tested and evaluated

to find the best performing model as judged by MAE or equivalently MPE. The following methods are used for our non-spatial models:

1. Simple Linear Regression (SLR) with each predictor separately
2. Multiple Linear Regression (MLR) with all predictors
3. Linear splines with each predictor separately
4. Linear splines for all predictors included all at once
5. Decision trees using all predictors at once
6. Random forest using all predictors at once
7. GAM model for all predictors included all at once

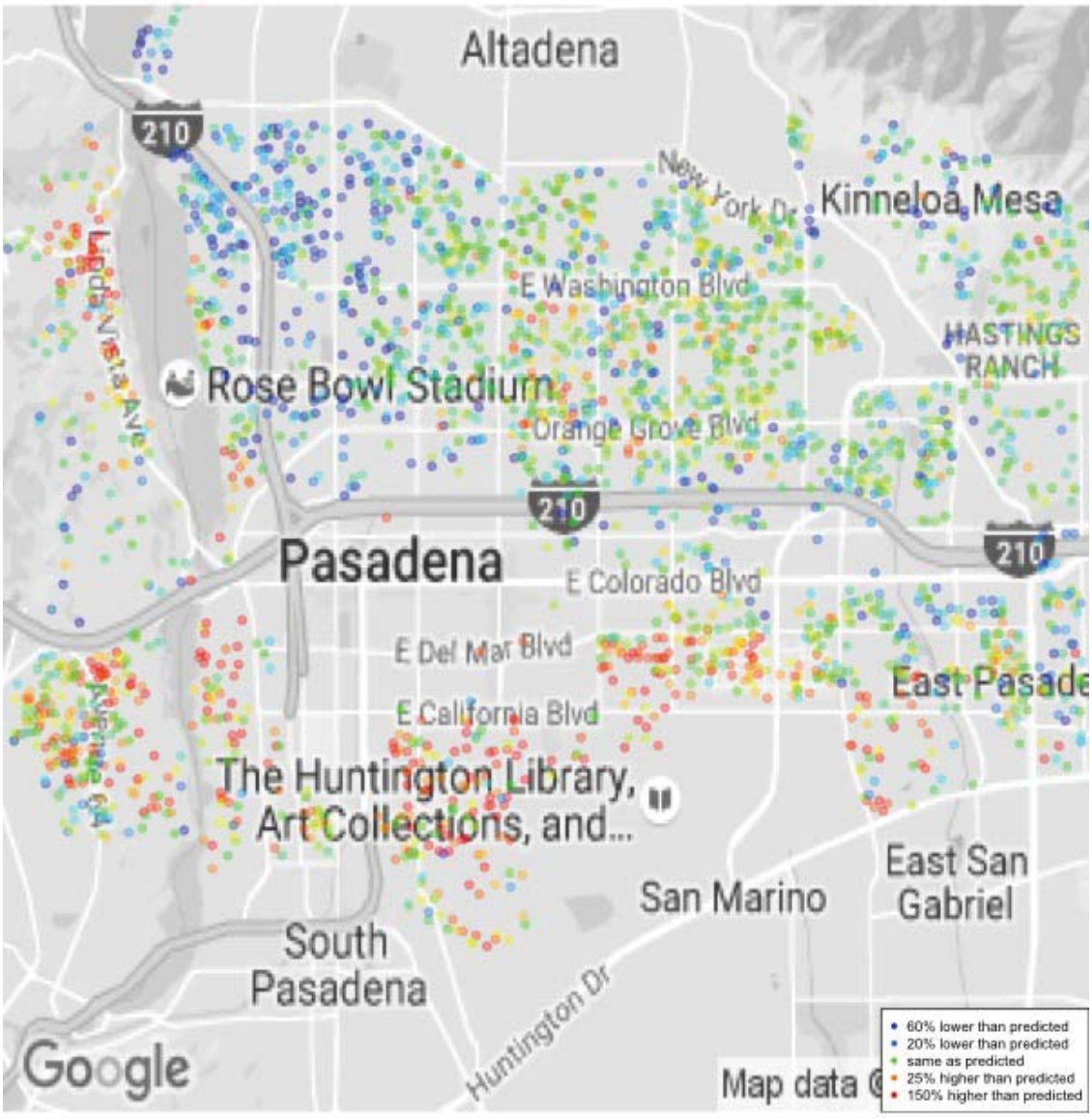
Table 4.1 shows the comparison of the discussed methods. Among all the models with respect to median percent error, the linear splines using all non-spatial predictors achieved the lowest median percent error, followed by GAM, and MLR all employing six non-spatial variables combined as predictors. The models that perform best are linear splines and GAM with nearly equivalent results with MPEs of 13.98% and 14.18% respectively. While the GAM model is constructed by using the `gam` function in the `mgcv` package in R, we implemented the linear splines models using custom R code.

We then calculate the residuals from our best model (linear splines), and plot these residuals versus spatial location. Figure 4.1 has revealed several non-random spatial clusters in the city of Pasadena. Certain areas are clustered by higher-value properties and others are clustered by lower-value ones relative to what we would expect given the specifications of the houses. By excluding the spatial coordinates, we assume a consistent relationship between house prices and house features only. However, the non-spatial

Table 4.1: Non-spatial Models

| Method | Predictor Variables | MAE | MPE |
|---|---|---|---|
| 1. SLR | beds | 0.381 | 0.3020 |
| 1. SLR | baths | 0.317 | 0.2457 |
| 1. SLR | log2sqft | 0.2346 | 0.1765 |
| 1. SLR | log2lotsize | 0.319 | 0.2475 |
| 1. SLR | year built | 0.443 | 0.359 |
| 1. SLR | duration | 0.489 | 0.4034 |
| 2. MLR | all non-spatial predictors | 0.2053 | 0.153 |
| 3. Linear splines | beds | 0.38 | 0.3015 |
| 3. Linear splines | baths | 0.303 | 0.2337 |
| 3. Linear splines | log2sqft | 0.222 | 0.1665 |
| 3. Linear splines | log2lotsize | 0.307 | 0.2375 |
| 3. Linear splines | year built | 0.433 | 0.34998 |
| 3. Linear splines | duration | 0.4834 | 0.398 |
| 4. Linear splines | all non-spatial predictors | 0.1888 | 0.1398 |
| 5. Decision trees | all non-spatial predictors | 0.238 | 0.1796 |
| 6. Random forest | all non-spatial predictors | 0.1962 | 0.1457 |
| 7. GAM | all non-spatial predictors | 0.1913 | 0.1418 |

Figure 4.1: Heat map of residuals from the linear splines model with all non-spatial predictors.



approach may not be the best fit because it only captures the direct effects of physical characteristics but fails to incorporate the proximity of homes to desirable amenities like shopping, transportation, schools and other high-end homes. The existence of spatial variations in the plot motivates the inclusion of geographic coordinates into the model.

### 4.1.2 Spatial Models

Regression models that predict the value of a home based on its relationship with non-spatial features only are often conducted out of convenience. When a fair number of spatial clusters are presented in the dataset, it is reasonable to suggest that location has a substantial relationship with property values. Spatial models can improve the prediction accuracy by taking into account the implications of geography. We employ the following spatial methods and compare their performance in Table 4.2.

1. Decision trees using only spatial coordinates (latitude and longitude) as predictor variables

2. GAM using only spatial coordinates (latitude and longitude) as predictors

3. KNN only using spatial coordinates (latitude and longitude) as predictors where $k = 4$. This $k$ was determined by empirically choosing $k$ to minimize MPE.

4. Decision trees using spatial coordinates plus all other characteristic attributes as predictors.

5. GAM using spatial coordinates plus all other characteristic attributes as predictors.

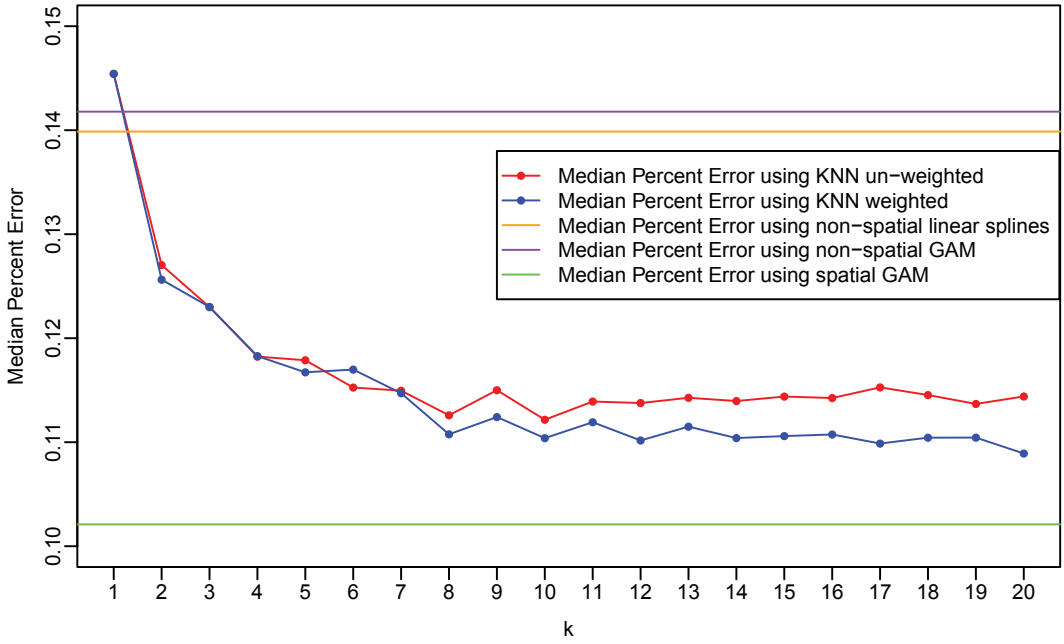
6. KNN using spatial coordinates plus linear splines used for all non-spatial predictors.

Table 4.2: Spatial Table

| Method | Predictor Variables | MAR | MPE |
|---|---|---|---|
| 1. Decision trees | latitude, longitude | 0.27 | 0.20581 |
| 2. GAM | latitude, longitude | 0.339 | 0.2647 |
| 3. KNN | latitude, longitude | 0.2478 | 0.1874 |
| 4. Decision trees | latitude, longitude and all other predictors | 0.237 | 0.178 |
| 5. GAM | latitude, longitude and all other predictors | 0.1402 | 0.1021 |
| 6. KNN & linear splines | latitude, longitude (KNN) all other predictors (splines) | 0.149 | 0.1089 |

The construction of the 6$^{\text{th}}$ method, the KNN method using spatial coordinates along with linear splines using all non-spatial predictors, is implemented as follows: In order to pick up more fine-grained spatial features, we fit all characteristic attributes (excluding latitude and longitude) in a non-spatial model using the linear spline method, since linear splines are shown to have the best performance out of the non-spatial techniques. The residuals extracted from this non-spatial model are then treated as outputs for a KNN model that uses latitude and longitude as predictors. The predicted residuals from this KNN, $E_{\text{KNN}}$, are added to the predicted values from the previous non-spatial linear splines model, $\hat{Y}_{\text{non-spatial}}$ to obtain the final predicted values $\hat{Y}_{\text{KNN/spline}} = \hat{Y}_{\text{non-spatial}} + E_{\text{KNN}}$. We refer to this approach as the *KNN/spline* approach for short. We repeat this approach multiple times with various values of $k$ (the number of neighbors for the KNN) to select the optimal $k$ value. The best possible value using the weighted version is between $k = 8$ and $k = 20$ based upon the comparison plot in Figure 4.2.

Figure 4.2: Median Percent Errors of the best non-spatial and spatial methods.



Results from Table 4.2 imply no remarkable improvement when latitude and longitude are added to the model using decision trees. The median percent error of this model with spatial coordinates slightly goes down to 17.8% compared to 17.96% without spatial coordinates. On the other hand, the inclusion of spatial location substantially improved the GAM performance, with a reduction in the median percent error from 14.18% to 10.21%. This reduction suggests that spatial coordinates are effective at removing spatial variation from the data. The GAM method that fits spatial coordinates along with all other characteristic attributes appears to have the best performance. The median percent error is about 10% which means our predictions are typically off from the true price by 10%. This is a good outcome considering our price data spans a wide range between $185,000 to $8,885,000. Results also show that, with the optimal value of k, our KNN/spline method performs almost as well as the spatial GAM model, with an MPE of 10.89%.

## 4.2 Variable Associations and Importance

We now describe the multivariate relationships between our non-spatial predictors and housing prices. Figures 4.3 and 4.4 present partial residual plots with fitted curves for each of the six predictors for the GAM and for linear spline models, respectively. First, we describe how each variable appears to be associated with prices. Following that, we discuss which variables appear most important for accurately predicting housing prices.

Assuming all other predictors are fixed, as lot size increases from 2,000 square feet to 32,000 square feet, prices increase steadily, approximately doubling over this range. In other words, a 10% increase in lot size yields a 1.3% increase in sale price. However, single family homes with lot sizes over 32,000 square feet do not appear to have higher

Figure 4.3: Curves representing each predictor's relationship with log price using GAM package.

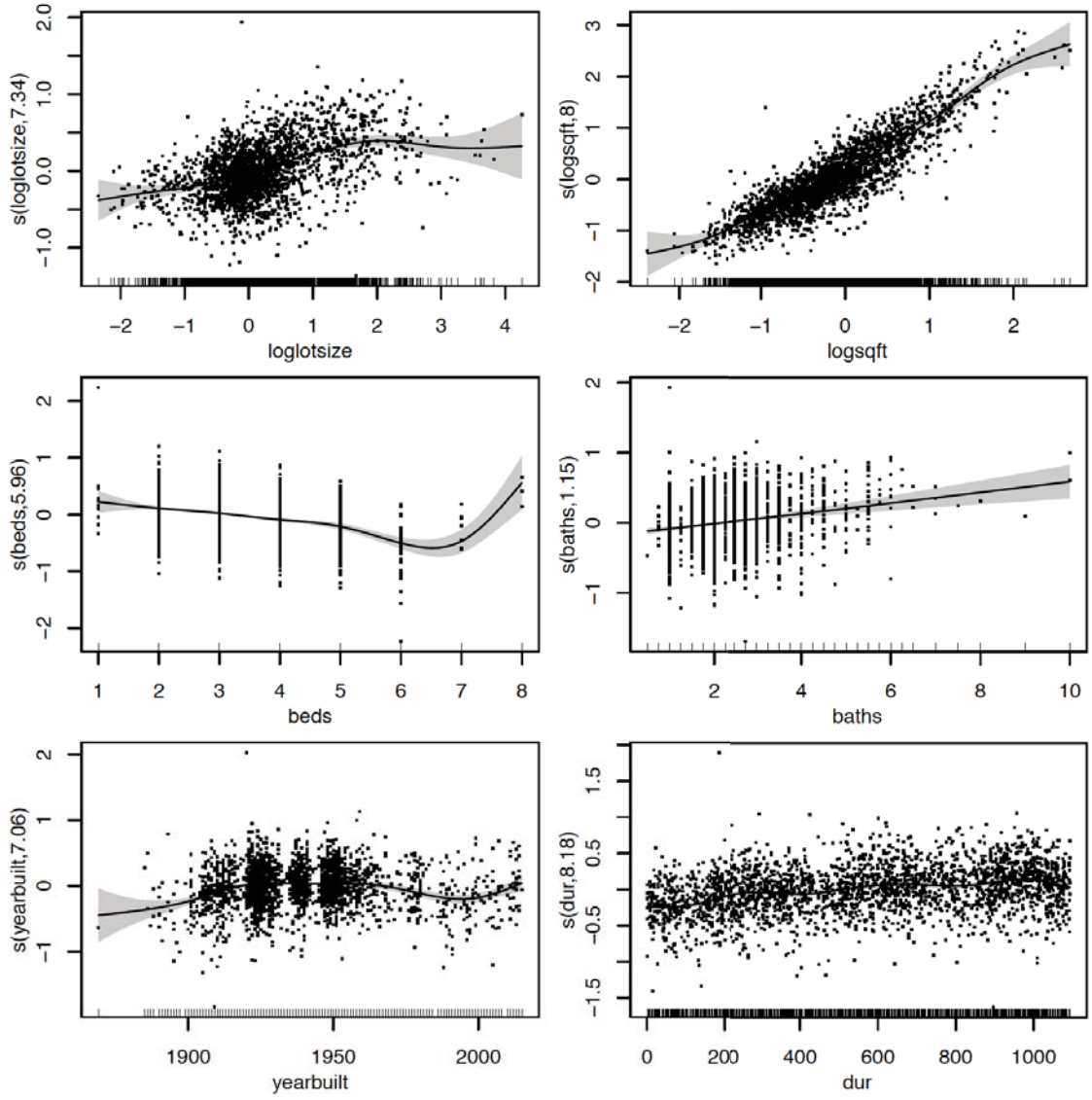
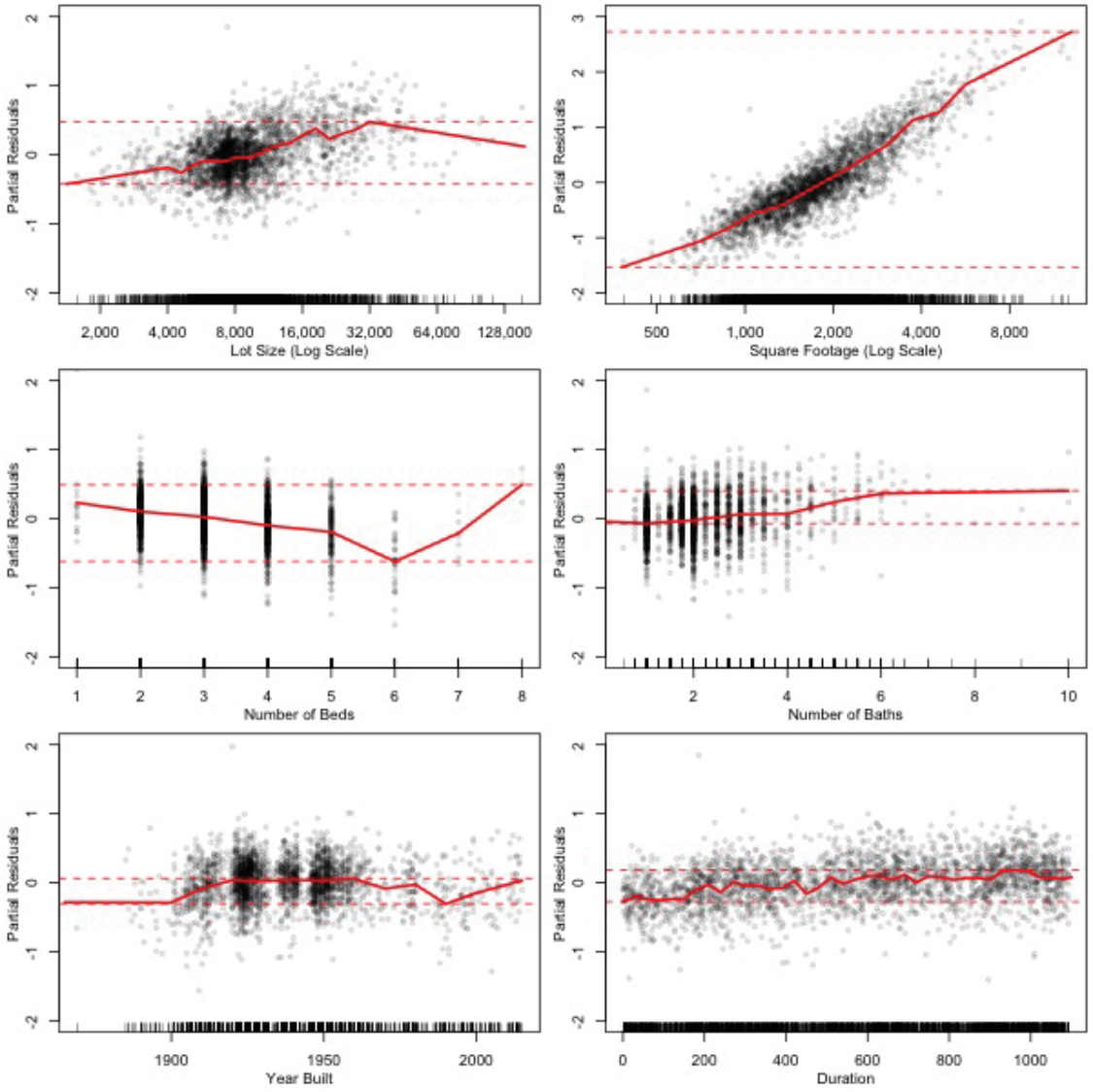


Figure 4.4: Curves representing each predictor's relationship with log price using linear splines.



prices and in fact may be less expensive, although we have limited data available for houses with such large lot sizes.

Similarly, the increase in interior square footage, between 500 square feet and 8,000 square feet, results in roughly an eightfold rise in price when holding all other predictors fixed. The practical interpretation of this finding is that for each doubling of square footage, price increases by 86%. We can also say when an interior size of a house is 10% bigger, buyers should expect to pay roughly 6.4% more assuming all other features equal.

Intuitively, houses with more bedrooms should be more expensive. However, this is primarily because these houses tend to be larger on average, whereas our inferences are regarding how prices change with number of bedrooms while holding other predictors constant. The relationship between number of bedrooms and price in our data indicates that for each additional bedroom the value of the house decreases by 7% with all other features held constant. The most practical explanation is that, for a fixed amount of living area, an extra bedroom would reduce the available space for other desirable features such as a larger kitchen, larger living room, and larger bathrooms. Another speculation could be that a property with many bedrooms is less likely a luxurious home. The response curve then shows a rapid turn in direction to a positive slope starting at 6. Although the relationship is less evident for houses with more than five bedrooms due to insufficient data, it is reasonable to expect large mansions with seven or eight bedrooms will sell for a much higher price.

Contrary to popular belief, for a fixed amount of square footage, a single family home is estimated to be less expensive with more bedrooms. However, interestingly, the value of the house increases with more bathrooms. The fitted response curve of baths implies a gradual increase in price as each extra unit of bath is added (recall that

one-quarter of a bath corresponds to one bathroom fixture, either a sink, toilet, shower, or tub). Specifically, on average, a two-bathroom house is expected to be 4.7% more expensive than a one-bathroom house assuming all other features of the house are fixed. Again, we do not have enough data to obtain accurate predictions for houses with more than six bathrooms.

Year built is another factor influencing a buyer's decision. The response curve for year built reveals certain tendencies of buyer preferences. For instance, given the same specifications such as lot size, square footage, and number of rooms, houses that were built in the 80's and 90's appear to be less popular among buyers as compared to houses that were built from the 20's to the 70's and new houses that were built in the 2000's. These preferences can be attributed to the exterior appearances and the layouts/designs that are associated with homes in these periods of construction.

The partial residuals for duration appear to be more noisy. Nonetheless, the overall trend of the response curve suggests that home prices overall increased by 38% over the course of the three year period October 2012 to October 2015, which equates to an average growth rate of 11% per year. This information could, for example, help prospective buyers and sellers adjust comparable transactions (called *comps* in the real estate industry, which are houses similar in location and features to the house under consideration that sold recently) occurring in the recent past to comparable values if these comps had been sold today.

Encouragingly, the GAM and linear spline models both predict a similar impact of property attributes on sale price. In summary, when setting all other features fixed, square footage and lot size have a large positive effect on sale price. The number of bedrooms has a negative effect. A positive effect is estimated for the number of bathrooms. Newer houses are estimated to be more expensive than houses that were built in the 80's and

90's, and houses that were built from the 1920's to the 1970's are more expensive than very old houses. We found an increasing trend between sale price and how many days past October 18, 2012 the house was sold. Hence the housing market improved substantially over the span of time under consideration.

In general, all discussed predictors are important components that account for differences in selling price. We can use the magnitude of these relationships to assess the relative importance of each feature. Our results suggest that out of the six predictors, square footage appears to be the most major factor as it accounts for a large portion of the variability in price. Lot size, number of bedrooms, and number of bathrooms tend to have less of an impact on price variation but still are significant explanatory variables for housing price prediction. Lastly, year built and duration, despite being not as important as the size of the house, are often features that motivate the home buying decision.

## 4.3 Spatial Variation in Housing Prices within Pasadena

Next we turn to identifying and describing regions within Pasadena corresponding to more or less affluent neighborhoods. Figure 4.5 provides a visualization of the results we obtained in Section 4.2. We introduce three heat maps which illustrate the improvements in our prediction accuracy when we include additional features into our models. Specifically, the top-left panel of Figure 4.5 provides a heat map of raw price values, with warmer colors corresponding to prices more expensive than the average in Pasadena, and cooler colors corresponding to less expensive homes. The top-right panel shows prices adjusted for the non-spatial features of our model. Here, warmer colors correspond to homes more expensive than we would expect given their non-spatial attributes such as square footage, number of bathrooms, etc. Finally, the bottom-left panel shows prices

adjusted for both non-spatial features and spatial location. Warmer colors here mean the homes are more expensive than the predictions made by our best model (including both spatial and non-spatial attributes), and cooler colors correspond to cases where the predictions from our best model were too high. In all three panels, we use the same color gradient scale, with blue corresponding to homes being 75% less expensive than predicted and red corresponding to homes 150% more expensive than predicted as extreme cases. Moreover, the opacity of the color represents the density of housing in that area.

From the top-left plot, nicer neighborhoods, which are colored closer to red, are located in the outer regions of Pasadena. Less affluent neighborhoods, which are colored closer to blue, are located near the 210 highway. Since the cluster separation appears to be closely related to high school boundaries, it is possible that a substantial portion of this spatial variation is due to school districts. Mapping of local high school district boundaries in the city of Pasadena and South Pasadena is depicted in Figure 4.6. Note that the blue cluster is within the approximate boundary of the John Muir High School district. The red cluster near The Huntington Library representing a more affluent neighborhood falls into the boundary of San Marino High School district.

The top-right plot displays the prediction error percentages when using only non-spatial features in the predictive models. It is apparent that the inclusion of non-spatial predictions accounts for a significant reduction in prediction errors. The map is predominantly green with the absence of red. However, there is still residual spatial correlation evident. It is depicted by the presence of several yellow clusters and a number of scattered blue spots across the board. This suggests that our non-spatial model is capable of predicting errors within an accuracy of 50% from the actual prices.

Finally, the bottom-left plot exhibits a transition in error percentages as we include spatial coordinates in addition to non-spatial predictors. The inclusion of geographic

Figure 4.5: Heat maps depicting size of prediction errors. Top left panel depicts errors when using no predictors (only the overall mean price) to predict prices. The top right panel gives errors when using all the non-spatial predictors. The bottom left panel gives errors when using both spatial and non-spatial predictors.

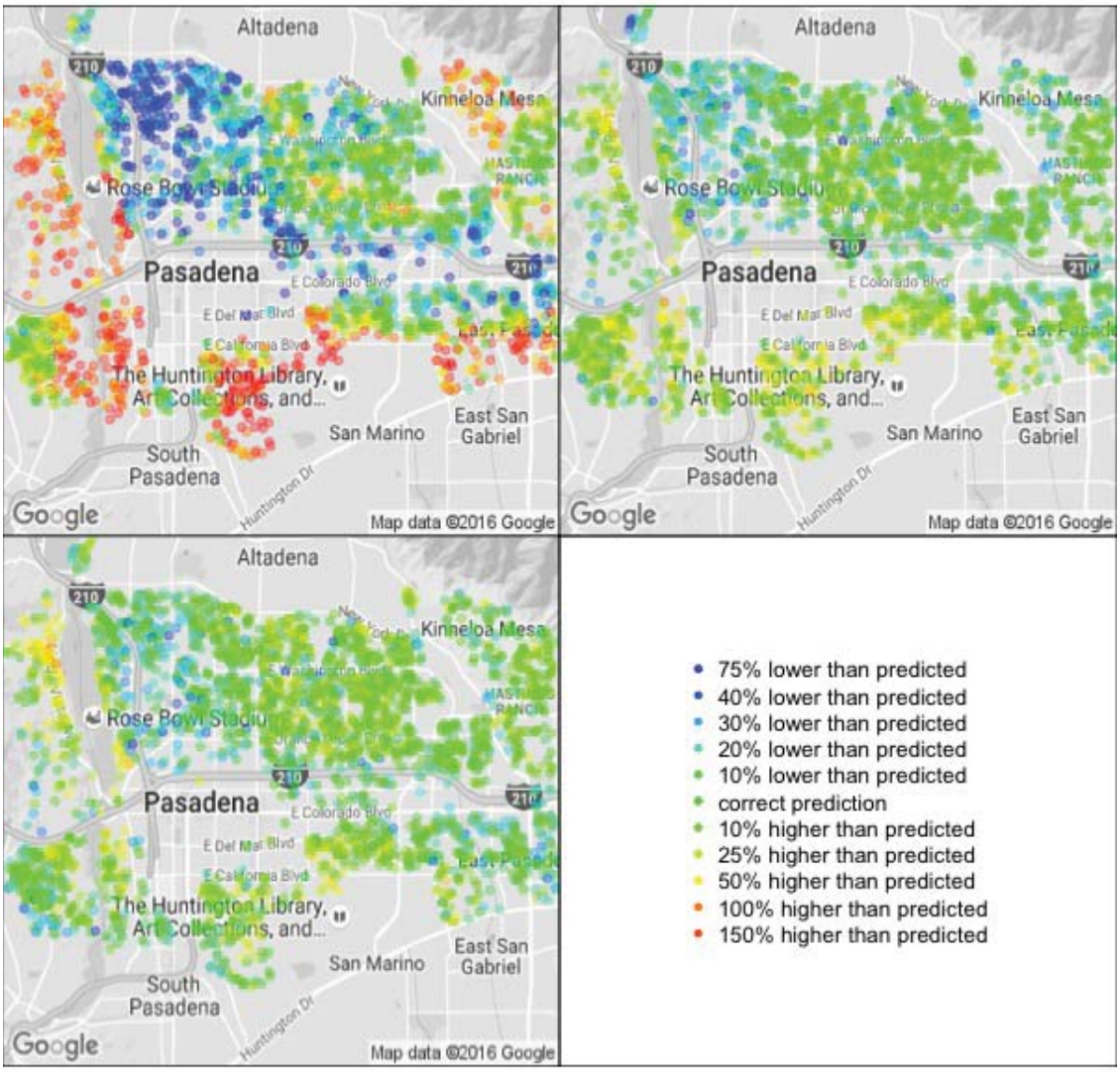
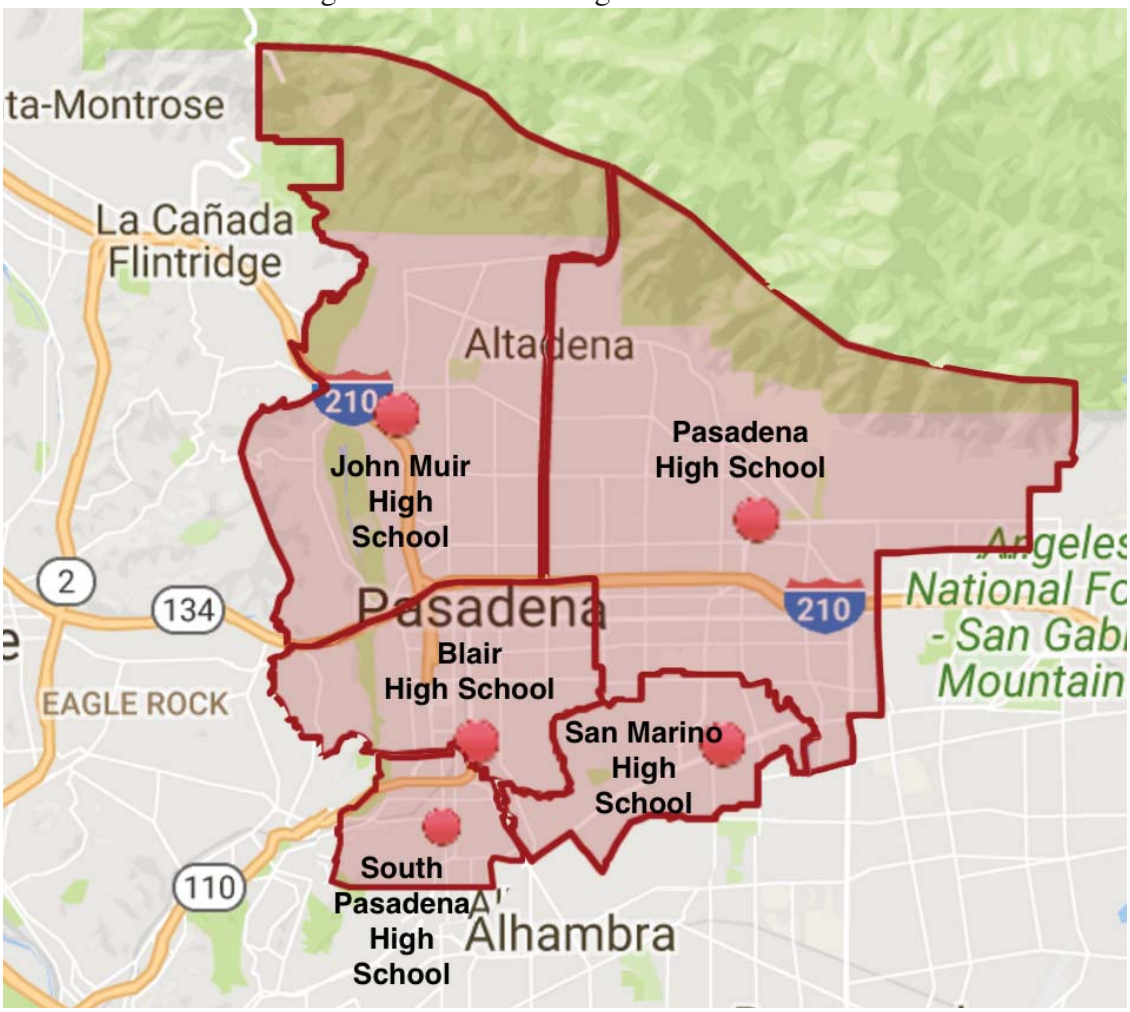


Figure 4.6: Pasadena high school districts.



location improves the accuracy of our predictions to a typical error 10% away from the actual prices. There is almost no spatial correlation left on the heat map. The plot shows virtually no red or blue cluster remain except for a small number of exceptional yellow and blue dots. This enhancement in prediction errors indicates a high level of spatial dependency in our data. This means we have succeeded in modeling the spatial component and explaining most of the variation in housing prices.

# Chapter 5

# Conclusion

With the use of a variety of analytical and graphical tools, we were able to evaluate the predictive performance of various housing price models applied to real data on single family homes in the city of Pasadena. In addition, our models also helped identify which characteristics of housing were most strongly associated with price and could explain most of the price variation. Furthermore, we were able to improve our models' prediction accuracy by accounting for the impact of spatial location.

The methods used in this study consisted of simple and multiple linear regression, linear splines, decision trees, random forests, and GAMs for non-spatial predictors. When spatial factors are incorporated, we employed decision trees, GAMs, KNN, and our KNN/splines combination method to predict housing prices. The models were compared and assessed using median absolute error and median percent error as performance metric criteria. While linear splines appeared to perform best with the non-spatial predictors (MPE $= 0.1398$), GAM achieved the lowest value of median percent error when accounting for both spatial and non-spatial predictors (MPE $= 0.1021$).

Another main goal of this thesis was to examine the importance of each predictor in

explaining price variation for a given set of housing attributes. Overall, our results provided practical information regarding the effect of various characteristics on house prices and their corresponding interpretations. In particular, when holding all other features fixed, square footage and lot size have a large positive effect on sale price. The number of bathrooms also appears to be positively related to price. One interesting finding from our work is that, contrary to popular belief, given the same housing specifications, the number of bedrooms has a negative relationship with house prices.

Moreover, the construction year of the house also has some influence on housing prices. Specifically, newer houses and houses built from the 1920's to the 1970's were estimated to be more expensive than houses that were built in the 1980's and 1990's, assuming other features fixed. The cause of this phenomenon is speculated to be the association of a house's exterior appearances or layout and its time of construction. Additionally, results obtained from sale dates implied an overall growth of 38% in sale price over a three-year span (October 18, 2012 to October 18, 2015). That is equivalent to an average annual growth rate of 11% in the Pasadena housing market.

The last finding in this thesis was that geographic coordinates, which are closely related to the quality of the school district, proved to be an important factor to building a reliable predictive model. Our results indicated that the inclusion of spatial coordinates in addition to all other non-spatial variables accounted for a large amount of prediction accuracy. We were able to improve the accuracy of our predictions to a typical error of 10% away from the actual prices and succeeded in modeling the spatial component and explaining most of the variation in housing prices.

These findings are useful to potential home buyers and real estate investors. The results from our study can help provide answers to home owners and shoppers when making decisions such as what housing attributes to consider in an effort to generate the

highest value of a home, how accurately the value of homes can be assessed, and finally how strongly home location is related to the value of a home.

Future research could investigate further issues. For instance, some desirable housing attributes were not included in our models, such as the presence of a garage, pool, or basement. With more predictor variables and more observations, better predictions can be made. Furthermore, it is established that the real estate price largely depends on neighborhood. The data for future exploration could include neighborhood characteristics such as school district, distance to major business centers, or accessibility to highways. If these data sets could be obtained, it would give clues to how these predictors affect housing prices. This information could prove useful for future investments and exploration.

# Bibliography

Basu, A. and Thibodeau, T. (1998). Analysis of spatial autocorrelation in house prices. *Journal of Real Estate Finance and Economics*, 17:61–85.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Do, A. Q. and Grudnitski, G. (1993). A neural network analysis of the effect of age on housing values. *The Real Estate Research*, 8(2):253–264.

Goodman, A. C. and Thibodeau, T. G. (1995). Age-related heteroskedasticity in hedonic house price equations. *Journal of Housing Research*, 6(1):25–42.

Grether, D. and Mieszkowski, P. (1974). Determinants of real values. *Journal of Urban Economics*, 1(2):127–145.

Hastie, T. (2015). *gam: Generalized Additive Models*. R package version 1.12.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning*. Springer.

Ligges, U. and Mächler, M. (2003). Scatterplot3d - an r package for visualizing multivariate data. *Journal of Statistical Software*, 8(11):1–20.

Lombra, R. (2012). The rise and fall of the u.s. housing market: Past, present, and future. *Journal Achievement USA*.

Malpezzi, S., Ozanne, L., and Thibodeau, T. G. (1987). Microeconomic estimates of housing depreciation. *Land Economics*, 63(4):375–377.

NAR (2009). *National Association of Realtors Existing Home Sales*. `http://www.realtor.org`.

Nguyen, N. and Cripps, A. (2001). Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. *Journal of Real Estate Research*, 22(3):333–335.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Redfin (2004). *Redfin: Real Estate, Homes for Sale, MLS Listings, Agents*. Seattle, WA.

Rosen, S. (1974). Hedonic prices and implicit markets:product differentiation in pure competition. *Journal of Political Economy*, 82(1):34–55.

Sirmans, S., Macpherson, D. A., and Zietz, E. N. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature*.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.