GEOGRAPHIC PATTERNS OF HOME PRICE APPRECIATION IN CALIFORNIA COUNTIES

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

Mauricio Gutierrez

2023

SIGNATURE PAGE

THESIS:	GEOGRAPHIC PATTERNS OF HOME PRICE APPRECIATION IN CALIFORNIA COUNTIES
AUTHOR:	Mauricio Gutierrez
DATE SUBMITTED:	Summer 2023
	Department of Mathematics and Statistics

Dr. Adam King _____
Thesis Committee Chair
Mathematics & Statistics

Dr. Jennifer Switkes _____
Mathematics & Statistics

Dr. Alan Krinik _____
Mathematics & Statistics

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. Adam King, for his invaluable guidance, support, and encouragement throughout this process. His knowledge and advice played a huge role in shaping the trajectory of this thesis. I would also like to give a big thanks to my committee members Dr. Jennifer Switkes and Dr. Alan Krinik for sharing their endless knowledge and providing support in times where I believed myself as incapable.

I would also like to share my appreciation towards my colleagues and friends who have offered their endless support and kept pushing me to move forward in times where I felt I couldn't.

Special thanks goes to my family as well for their unconditional and infinite love. They have pushed me towards greatness and offered a source of motivation to follow my dreams.

Finally, I want to give acknowledgement to the Mathematics and Statistics community of Cal Poly Pomona for providing me with the tools to succeed in this endeavor. The work of the faculty and staff in this department has been nothing short of inspirational.

ABSTRACT

For our study, we utilized a comprehensive dataset from the state of California accessed from Redfin, containing various property types in California counties: all residential, single family residential, condominiums, multi-family units and townhouses. We performed pre-processing and exploratory analysis incorporating several different visualizations such as spaghetti plots, counting plots, geographical plot and scatter plots to analyze house sale price changes over time (by month) both by county and by property type. Our exploration investigates housing market trends between 2012 and 2021 obtaining a scope of counties that experienced significant increases, slight increase, no changes or decreases in home prices. Doing such analysis is a key to analyze spatio-temporal patterns, in relation to spatial (geographical) and temporal (time) dimensions and seeing evolution of our data in these two dimensions. Additionally, we fitted models to predict monthly median sale price and determine trends we see over time and examine which variables carry the largest influence for home sale prices. The predictive models were evaluated by employing inference plots, determining if the fitted regression line captured trends and seasonal patterns properly. This study will be beneficial in understanding what factors play a role in influencing changes that could occur in the following years to come and what other researchers can do in order to mitigate the risk of a market crash in the foreseeable future. By excavating these influential factors and trends, our research contributes to a firm grip of future market dynamics and aids in making informed decisions for all stakeholders involved in the housing market.

Contents

List of Tables

List of Figures

Chapter 1

Introduction of Data

In this chapter we will provide background of the dataset and a summary of variables contained in this dataset. This data contains monthly median sale prices for all counties in the United Sates and each county may contain five property types: All Residential, Condominiums, Multi-Family (2-4 Unit), Single Family Residential and Townhouse. Note that for some months in time, some counties may not contain observations for some property types, meaning we have no recording for some of these types. In other instances, some counties may have months or up to years of missing information. It is likely possible these counties don't contain enough population or enough homes available to sustain a fast paced housing market. In this chapter, we will also provide a brief literature review that provides similar works and that are parallel to this study on predicting home prices.

1.1 Study Objective

In the past few years the world encountered COVID-19, which itself has brought challenges to the US economy. In the end of the COVID-19 era, there has been

a surge in housing demand in response to lower mortgage interest rates, which inherently illustrates a structural break since March 2020. Since this structural break the housing market has still witnessed increasing demand [22]. Housing prices across the US have seen significant growth, with the Case-Shiller National Home Price Index claiming that real house prices have rose 45 percent from February 2012 through May 2020. It has been suggested that the US will soon face an economic recession and a financial crisis may soon have waves that could disrupt financial intermediation and then hit the US economy [2]. The uncertainty of the housing market raises questions about future housing costs, which directly affect buyers and sellers. This question requires the use of predictive models that could evaluate seasonal and yearly growth and predict home values.

Our objective of this study is to use methods to approximate the growth or reduction of median house prices and to assess the spatio-temporal patterns in our data set. Analyzing which counties were more likely to undergo price hikes or drops and which were most volatile to market changes. The housing market is influenced and driven by the economic adjustments faced throughout the years. The stubborn hoisted prices in today's market leave a lot to be desired in affordability, specifically for firs-time homebuyers. Elevated prices, especially in the California coastal markets have raised concerns over affordability for both lower and middle class families.

Modeling this objective will involve some intricate details that will require apprehension on geographical and temporal petterns in our dataset and the aggregate and distributional effects of the COVID-19 era. In our data, there are complications that were noted, such as counties with a larger population. Specifically, in counties with a larger population we expect a faster-paced market selling homes

at higher volumes. To observe changes over geographical space over different regions of California, it will be beneficial to observe these changes for coastal urban, coastal rural, inland urban and inland rural counties. This study will provide a better understanding of how circumstances affect home prices as housing is essential for family welfare, stability and wealth. Using the computation capacity of the R programming language to predict housing prices will help understand the impact on other vital economic factors [13].

1.2 Data Background

This study will use the R programming language for pre-processing and computations. This program is extensively utilized in the industry for exploratory analysis, modeling, cluster analysis, etc.

The data was downloaded from Kaggle and is a collection of median sale prices on a month to month basis throughout the United States by State, Metro, and Zip code from January 2012 to December 2021 [9]. This data was accessed from Redfin, a website for real estate brokerage, published on January 9th, 2022 [14]. For our purpose we filtered only the observations that were in California and will only focus on this state.

The file on Kaggle originally is in tab-separated values (TSV) format. Instead of using this format, we converted the file to comma-separated values (CSV) format, a common format used for statistical analysis in the R programming language. In the data's original form we have 563,123 rows of data accounting for all states in the US. However, we will be investigating our local housing market of California. Therefore, we have filtered all values to only be located in the state of California, resulting

in a reduction to 19,967 rows of data, containing 41 of 58 counties of California throughout the period 2012 to 2021. These are a conjunction of properties sold in a monthly period. There are 5 property types: All Residential, Condo/Co-op, Multi-Family (2-4 Unit), Single Family Residential and Townhomes. All Residential is a combination of all property types, meaning if an observation contains any one of the other four property types, it will contain an observation for the All Residential property type. In that regard, a county can contain up to 5 observations in one month period. For the most diverse example, we will consider Los Angeles county. In the month of January, 2012, Los Angeles contains all 5 property types. Table 1.1 demonstrates all 5 property types and the number of homes sold that month respective to each property type. The property type that sold the most in this month was Single Family Residential, selling $3,565$ homes and the property type that sold the least in this particular month was Townhouse with only 257 sold in the month of January. Lastly, All Residential is all other four property types added together.

Table 1.1: Properties sold in LA county for the month of January, 2012, including: all residential, condo/ co-op, Multi-family, single family residential, and townhouses

Property Type	Homes Sold
All Residential	5057
Condo/Co-op	880
Multi-Family (2-4 Unit)	355
Single Family Residential	3565
Townhouse	257

1.2.1 Variables in our Data

In the pre-processing section, we will exploit rigorous techniques in order to select the most significant variables that carry the most impact on median sale prices. Identifying these variables that may have the most effect on median sale prices is vital in achieving our objective of improving the accuracy and relevance of our analysis. To achieve this goal, we will perform comprehensive exploratory data analysis to gain insights into potential relationships between median sale prices and different variables. This thorough exploration will reveal of potential associations and concealed patterns, bringing the rise of influential factors for variations of median sale price.

Some important variables to make note of are:

- **period_begin**: Beginning of month period in date format: mm/dd/yyyy
- **period_end**: End of month period in date format: mm/dd/yyyy
- **Region**: county indicator
- **property_type**: 5 types of property a county could sale (mentioned in previous slide)
- **median_sale_price**: The media sale price of all properties that were sold one month period
- **homes_sold**: Number of homes/properties sold in month period for a specified property type

Furthermore, in the following chapter, we will address matters from missingness of observations, as well as outliers. These imperative steps in data refinement will

render strength and reliability in our data analysis, which in turn ensures superior accuracy and authenticity in our results.

For a comprehensive list and description of all variables, please see Appendix A.

1.3 Literature Review

As of late, the price of real estate has been an ongoing issue, where the market inhabits price fluctuations and is a general concern for the public. There have been many works over the decades that have assessed and modeled home prices where the most common approaches use linear regression and hedonic regression. Specifically, in our study we will be focusing on longitudinal data analysis.

Longitudinal analysis is a study on analyzing measurements of the same subject(s) taken repeatedly over time, which allows the direct study of change over time. In particular, those performing these studies are interested in the temporal patterns of characteristics in a data collection [5]. In this study, this will help analyze sale price changes over time per county and property type.

Nagaraja et al. (2011) used random effects models to predict home prices composed of a fixed time effect and a random ZIP code effect combined with an autoregressive component [10]. These researchers found that this outperformed a conventional mixed effects model. Another advantage their model had was that it remained easy to interpret at both the micro and macro levels, in spite of including several features inherent in the data.

Soltani et al. (2022) mentioned how conventional housing price prediction methods rarely consider the spatiotemporal non-stationary problem in large data vol-

umes. This study used machine learning models to understand the effects of housing characteristics such as property details and neighborhood quality. Decision trees in these scenarios have better performance than linear models [16]. Ensemble machine learning models, like Gradient-Boosting and Random Forest provide better predictions for home prices. In these models, the researchers added a spatiotemporal lag variable to improve prediction accuracy, which is a weighted sum or weighted average of the neighboring values for that variable. It was claimed in this paper that spatial and temporal variations are the most impactful factors of housing value changes (Yao & Stewart Fotheringham, 2016)[21]. Pace et al. (2000) also conducted a similar study using spatial-temporal autoregression and the substantial benefits when including spatial and temporal information [12]. In their study, they were able to generate 46.9% less for sum of square errors (SSE) using this particular autoregression with 14 variables compared to a 12-variable regression. Just as they mentioned, our data is an example of data that has been organized by unit of time (monthly) and by geographical location (county, state, etc.), therefore our data contains spatial and temporal characteristics.

Many methods may be used to estimate housing prices such as linear regression, Decision trees, Random Forests, boosting methods, Support Vector Machines, etc. Huang used all of these models and the main motivation was to find improved accuracy in predicting home prices using advanced methods. Concluding on this study, Huang found the most accurate models were tree based non-linear models with the lowest mean squared-errors [8].

Another common method used for predicting home prices is hedonic regression. According to the CFI team, hedonic regression is a regression technique used to estimate the value of a good, service or asset by splitting the product into its

different features and how much each feature contributes to the value of the product [17]. Dubin conducted research on predicting home prices using hedonic regression and noting of the importance of being accurate when predicting home prices. They estimated a hedonic regression using ordinary least squares, and then using these coefficients to cast predictions. The author stated one large issue using this method; it ignores correlation between neighboring homes. Dubin aimed to integrate these correlations when calculating the regression coefficients [4].

Chapter 2

Data Analysis

This chapter turns its focus to data analysis, which will provide comprehensive visualizations of median sale prices with other variables such as time (monthly), property type and more. These visualizations include plots investigating patterns of data missingness, spatio-temporal patterns and spaghetti plots.

This chapter will also dive into feature engineering, generating new variables based on features available in our dataset. In addition, we will pre-process our data, which will provide a clean slate with data that is a lot more usable for our purposes in median sale price prediction. This step is crucial in mitigating bias in our studies and removing outliers that may influence our predictions as these outliers are special and uncommon cases. Outliers prevent from a balanced representation of our data. With the pre-processed data we will include plots previously created for comparison purposes.

2.1 Exploratory Data Analysis

It is to be expected that data may harbor outliers and some anomalies in its original, unfiltered and unedited state. For this reason, data visualizations are created to see any patterns of missingness or outliers that are to be excluded when performing advanced procedures. Once abnormalities in our data are discovered we must perform pre-processing to remove impurities, duplicates or any observations deemed as irrelevant. Doing direct analysis with raw data can also lead to misunderstandings and raise concerns with unethical practices. It is an integral part of our study to perform this step for data reliability.

As mentioned before, in its original state, we have $19,967$ observations available. In pursuit of rigorous analysis and accuracy, our exploratory analysis plays a pivotal role in diagnosing observations that require removal from any further consideration, not doing so may skew our findings. We will later talk about this in further detail in the pre-processing section once we analyze visualizations made in this section.

Note that all plots in this section are done with the raw data; in the pre-processing section we re-plot with the pre-processed data for comparison purposes.

2.1.1 Missingness of Data

It is a very rare occasion where data is balanced and complete. This brings challenges for further analysis and may bring lackluster results when modeling our data. We will be checking this completeness by visualizations that depict the number of observations present for each county by property type. Our data only contains 41 out of 58 counties present. We do not have a concrete answer as to why 17 counties were not present in our dataset. However, some logical reason

may be the lack of urban centers and house availability. Figure 2.1 has calendar month on the x-axis, property type on the left vertical axis and county on the right vertical axis. The points drawn are colored by property type and they represent observations present by county for a specified calendar month. For example, if we look at the first point drawn in Los Angeles county for the single family residential property type, this signifies an observation present for January, 2012. On the contrary if we look at the townhouse property type for Glenn county, we do not see an observation drawn until at some point in 2017. That is, Glenn county has no townhouse observations present in the dataset until 2017. Furthermore, Figure 2.1 shows that Los Angeles county is a great example for a county that boasts data completeness, while Glenn county is not. It was reported that Los Angeles has almost 10 million residents while Glenn county has only about 28,000 residents [15]. This signifies a lack of population in Glenn county to require an abundant variety of property types.

In this section, we have only provided one example regarding missingness of data. Appendix B covers this information in more detail.

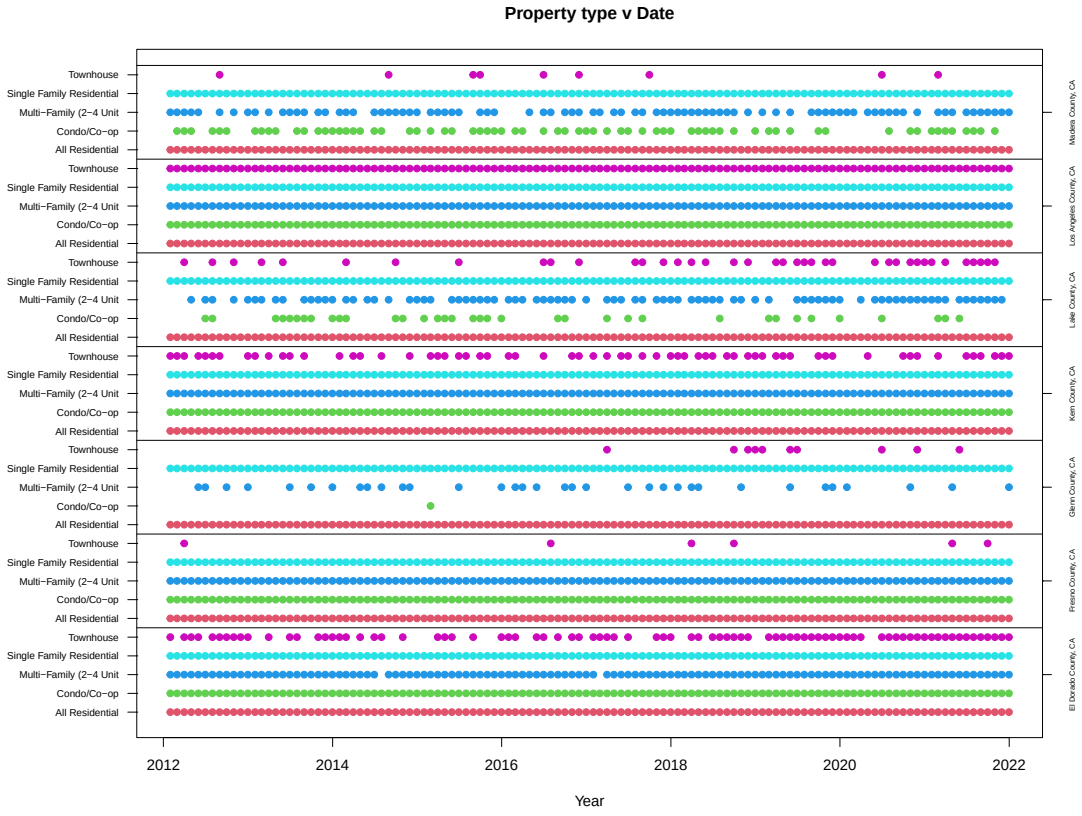


Figure 2.1: Property type vs Period End for Counties E-M

2.1.2 Spatio-Temporal Mapping with Shapefile Plots

Moving forward, we embark on an immersive expedition through the spatio-temporal dynamics of the housing market, enabling us to see links in demographic trends, urbanization, and housing price fluctuations.

To complete this task, a shapefile was used to produce a map of the state of California with county borders [11]. Our original data was filtered for February, 2012 and December, 2021 and the variables selected were region and median sale price. Furthermore, we filtered for each property type to create separate plots pertaining for each type of property, and this will be very helpful in analyzing data from a granular perspective. These columns of data were added onto the shapefile in order to produce coherent plots and be able to fine tune the features in our plots. The ggplot package was utilized to produce such plots filled by a continuous color range which ranges from blue, representing lower median sale price to red, invoking higher median sale prices [20]. Nearby shades of purple occupied the spectrum indicating median sale prices falling between red and blue. Therefore, a map plot was created for February, 2012 and a map plot was created for December, 2021, each county was filled depending on the median sale price observed for a selected property type.

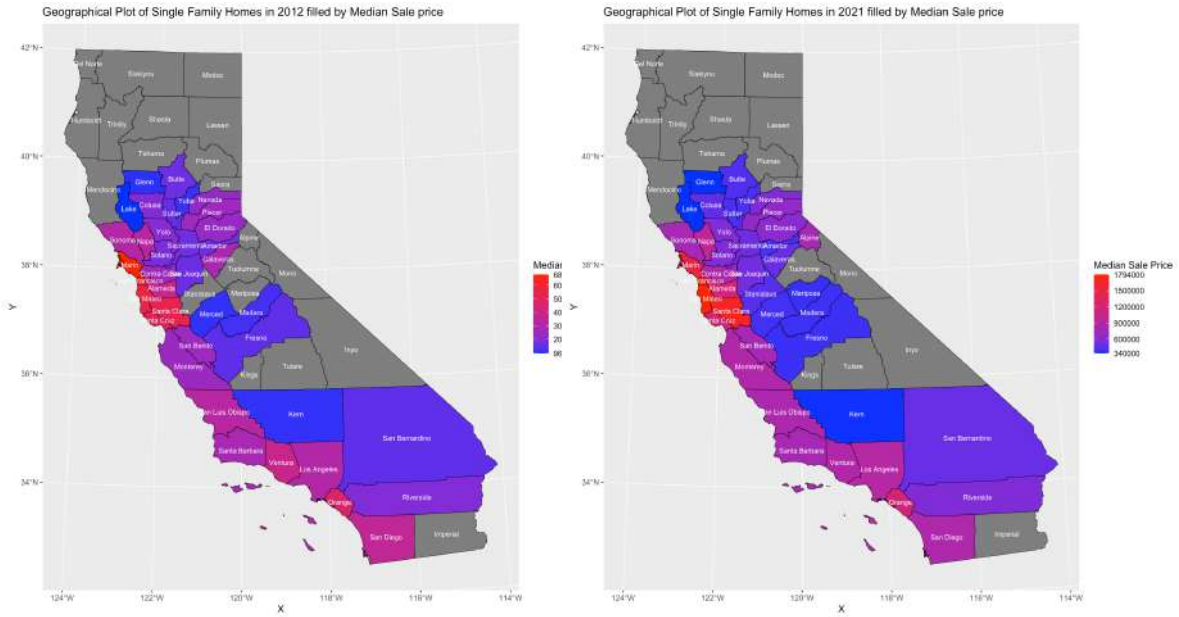
Furthermore, we calculated a percent change metric and was integrated into geographical plots like above, with similar implementation of the color range. These plots will be useful in identifying the magnitudes of price changes per county from 2012 to 2021.

Going forward, any work for multi-family (2-4 Unit) will be ceased since prediction on these properties will be inconsistent and unfair. This verdict was insinuated from acknowledgement that these properties have a single owner renting out multi-

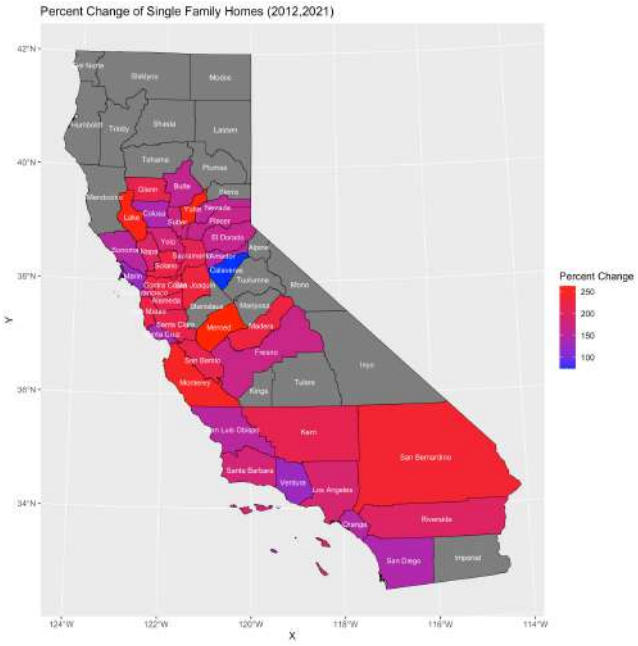
ple units, however it is noted that properties with 2 units might be priced differently compared to properties with 4 units. In a more practical sense we should consider these as separate observations. However, without more information we cannot separate these observations to entail multi-family with 2 units only, 3 units only and 4 units only.

Nonetheless, spatio-temporal visualizations is a very powerful tool in revealing patterns and trends over the dimensions of space and time. The integration of these two dimensions allow for a very powerful and robust analysis in the structure of our data. These visualizations garner the power of exploring evolution of our data across counties throughout time, illuminating potential correlations.

Figures 2.2(a), 2.2(b) and 2.2(c) depict plots filled by median sale prices for single family homes for 2012, 2021 and the percent change respectively. In Figure 2.2(a), the most expensive single family homes are found in the Bay Area and in Orange county. Note that the trend is that coastal counties tend to have higher median sale prices than inland counties. For instance, Glenn, Lake, Kern, Madera, Merced, and Yuba counties produced the lowest median sale prices. In 2021 (Figure 2.2(b)), results are similar, with the addition of Stanislauss and Mariposa which also saw lower prices. In addition, the legend in these two plots display a dramatically increased range of median sale prices. The percent changes in Figure 2.2(c) shows that all counties experienced a positive percent change in median sale prices. Counties in the Bay Area and many nearby show a 200% increase in prices. South coastal counties experience increased prices, though to a lower degree than counties in the Bay Area. San Bernardino, Merced and Madera counties were inland counties that saw percent change greater than 200%. The only county that seemed to have the least change was Calaveras county.

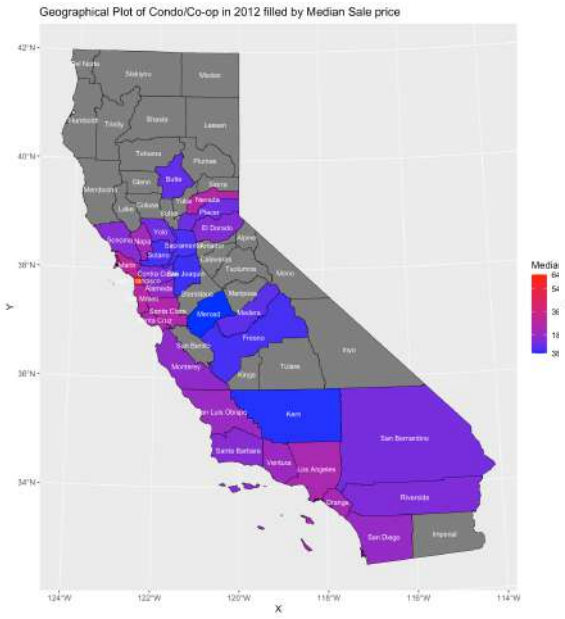


(a) Price for Single Family Homes in 2012 (b) Price for Single Family Homes in 2021

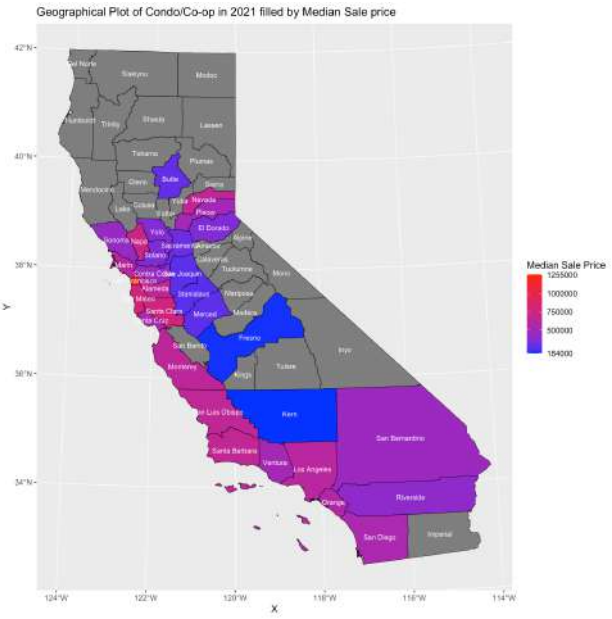


(c) Median Sale Price Percent Change

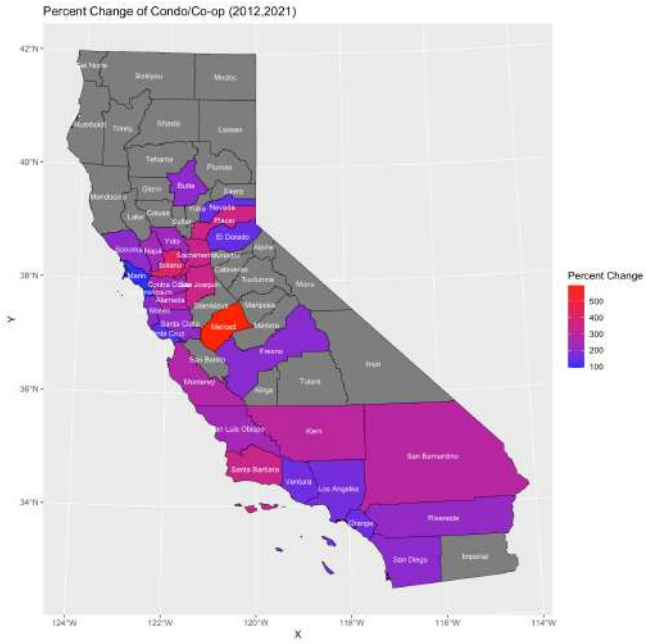
Figure 2.2: Single Family Residential Geographical Plots with Raw Data



(a) Price for Condos/Co-ops in 2012



(b) Price for Condos/Co-ops in 2021

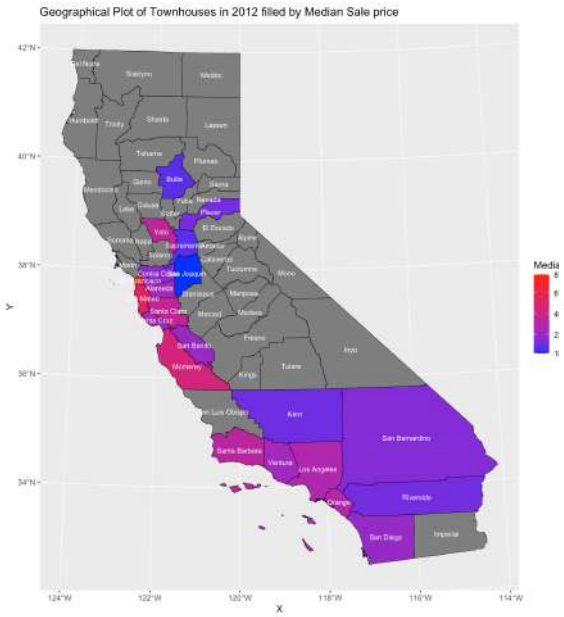


(c) Median Sale Price Percent Change

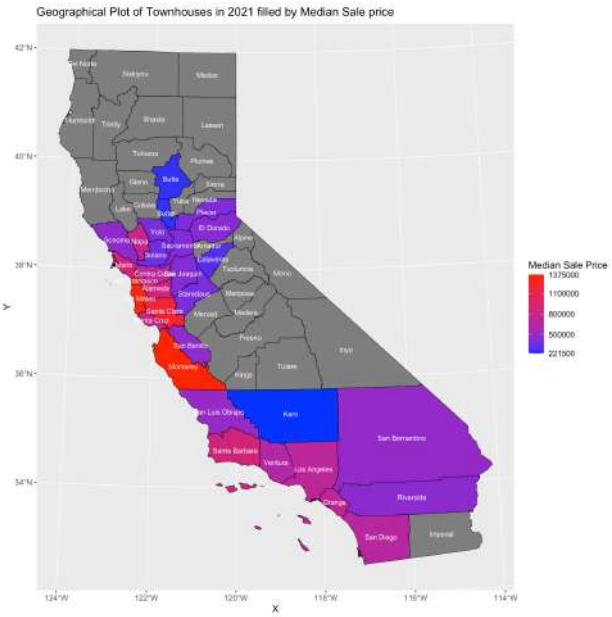
Figure 2.3: Condominiums Geographical Plots with Raw Data

Figure 2.3(a) depicts median sale prices for condo/co-op properties in 2012. The county with the highest price at this time was San Francisco, while the lowest priced counties were Fresno, Kern, Merced, Solano, Sacramento and San Joaquin. Similar to observations made in Figure 2.2(a), here prices of condos located in coastal counties come at a premium compared to inland counties, making condos in these areas more desirable. Figure 2.3(b) depicts prices in 2021, demonstrating a similar trend where again coastal counties are costly. Most counties in 2021 were shaded in hues closer to red, so cost of condominiums became exorbitant in the past decade. Further supporting these price hikes, in Figure 2.3(c), nearly all counties saw at least a 200% price increase. As expected, San Francisco county, along with Marin county, had the least percent change since these were higher priced counties. The county that witnessed the highest percent change is Merced county, which was anticipated due to its lower prices compared to neighbouring counties.

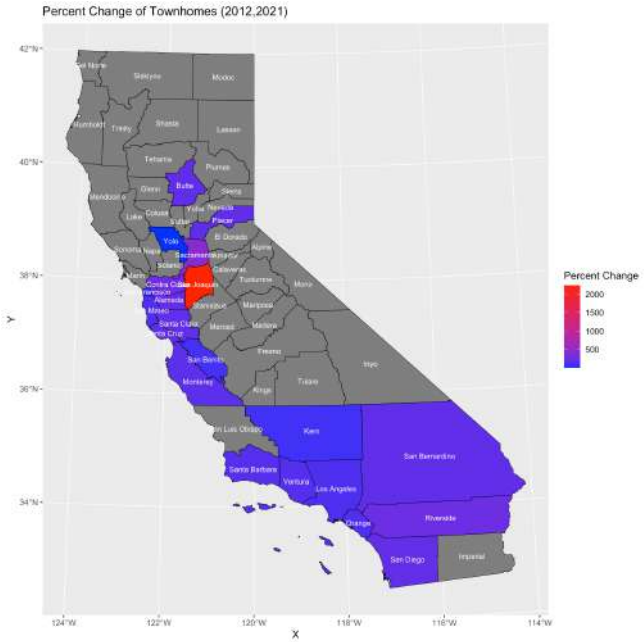
Continuing a similar trend, Figure 2.4 illustrates that San Francisco county holds the most costly homes in California. The Bay Area (surrounding counties of San Francisco county) are seen as more expensive than their inland counterparts. This remains true for southern coastal counties as well, although they are not as costly as townhouses in the Bay Area. In 2021, there were jarring median prices that reached and surpassed the million dollar mark. Examples of this are San Francisco, Santa Clara, San Mateo and Monterey counties. The cheapest counties to live in at the time were Kern, Butte and Sutter counties. Figure 2.4(c), illustrates that San Joaquin county had the most abrupt percent change out of all counties and Yolo county had the least percent change in median sale prices.



(a) Price for Townhomes in 2012



(b) Price for Townhomes in 2012



(c) Median Sale Price Percent Change

Figure 2.4: Townhome Geographical Plots with Raw Data

2.2 Pre-processing

This section entails detecting anomalies and outliers; where we ensure their removal to prevent skewed analysis. We will also filter any observations that are duplicates in order to enhance our dataset. Another step in this process involves taking transformations of variables in order to assume a normal or close to normal distribution of a variable. This will also prevent any negative median sale prices, which is not meaningful in our analysis (i.e., no homes are sold for for less than zero dollars). Lastly, we want to feature engineer, a very useful tool for improving the performance of our models.

Once we clean the data, we are interested in conducting applied longitudinal analysis. A prominent feature of longitudinal studies is that it partakes a study of individuals with repeated measurements over time [5]. In this study, we refer to these individuals as the counties.

2.2.1 Data Removal

As expected, we have determined some inconsistencies in our data, such as columns with missing values and outliers. For instance, Alpine county only sold one single family home during May of 2015 with a listed median sale price of $\$2,550,000$, a substantial increase from the month prior at $\$332,600$. Some of these abberations are present due to inadequate number of houses sold. In addition, about $3,000$ observations of the $19,967$ observations sold less than five homes. Observations that sold less than 5 were removed for those that were for all property types. Many of these observations were from less populous counties, which inherently cause slower housing markets in these areas (see appendix A for population sizes).

Due to the nature of how the all residential is collected, it may be a duplicate in some cases. For example, if for a month there was only single family residences sold for a county then, all residential will duplicate this information due to being an aggregate of all property types. Therefore, we will cease further analysis for the all residential property type. In addition, since multi-family homes vary from two to four units, it is an unfair comparison as two unit homes may be more expensive on average compared to four unit homes. So, we will not consider this property type for any further analysis. Therefore, we filtered observations for single family residential, condominium and townhouse property types.

For columns that have NA's we will be imputing and interpolating depending on the number of NA's visible. If one column has thousands of NA's these variables are to be removed and if an observation contains too many NA's then that observation will be removed. From the exploratory plots and data inspection we were able to determine anomalies in this dataset and tackled this shortcoming by removing unnecessary observations that may drive models towards inaccuracy.

2.2.2 Feature Engineering

Feature engineering is known as the creation of new variables from existing variables or information in our dataset. These new features will allow our models to learn more about the data and may improve performance. Feature engineering not only involves variable creation, but it also includes feature selection, transformation, one-hot encoding, dimensionality reduction and more [23]. For our study, we will be gravitating towards feature selection, transformation, one-hot encoding and feature creation.

To summarize the changes made to the original data we have provided some

bullet points below:

- **Feature selection**: we selected period end, median sale price, property type and region as important variables.
- **Feature creation 1**: Create a variable called area type which consisted of 4 categories regarding the geographical location of counties: coastal urban, coastal rural, inland urban, inland rural.
- **Feature creation 2**: We included a population variable from data gathered from US Census. US Census calculated estimates for each county on a yearly basis. That is, we were unable to find data of population on granular level covering month to month. Therefore, for all months for 2012, per county, we added 2012 estimates and there is no variation monthly [18] [19].
- **Feature creation 3**: We created another time variable (time_since) which began at 0 for the first month and ended at 119 for the last month of recording (this will be used for modeling purposes).
- **one-hot enconding**: Lastly, we created a binary version of area type only accounting for coastal or inland (no rural or urban categories included).

We created another area type variable to utilize in our model building chapter to create separate models since "rural" and "urban" categories already have an implication of population. Therefore, we wish to isolate these variable in the model building section.

2.2.3 Spatio-Temporal Mapping with Shapefile plots

Now, we wish to analyze which counties had a higher or lower median sale prices compared to other counties. In addition, we aim to determine the percent change for each county. Figure 2.5 achieves these objectives where Figure 2.5 (a) & (b) is median sale price for single family homes in 2012 and 2021 respectively. Figure 2.5(c) is the percent change for median sale price. These types of figures are known as shapefile plots, which are geographical plots that shade regions using a given color gradient. We have plotted these figures with the raw dataset in the previous chapter and the same color gradient applies here [11]. That is, a county with a lower median sale price is shaded in a hue of blue, a county with a higher median sale price is shaded in a hue of red and any counties in between are shaded in a hue of purple.

In Figure 2.5(a) the most expensive counties in 2012 were those located in the Bay Area, while lower priced counties were inland not populous counties such as Lake, Glenn and Yuba. Other inland counties that belonged in the lower price bracket were Merced, Madera and Kern. In Figure 2.5(b) similar results are seen where the Bay Area continues to have higher priced single family homes, and lower priced counties remained the same except for Glenn county moving up in the price bracket. Figure 2.5(c) displays that all counties except for Calaveras county sought gains from 2012 to 2021.

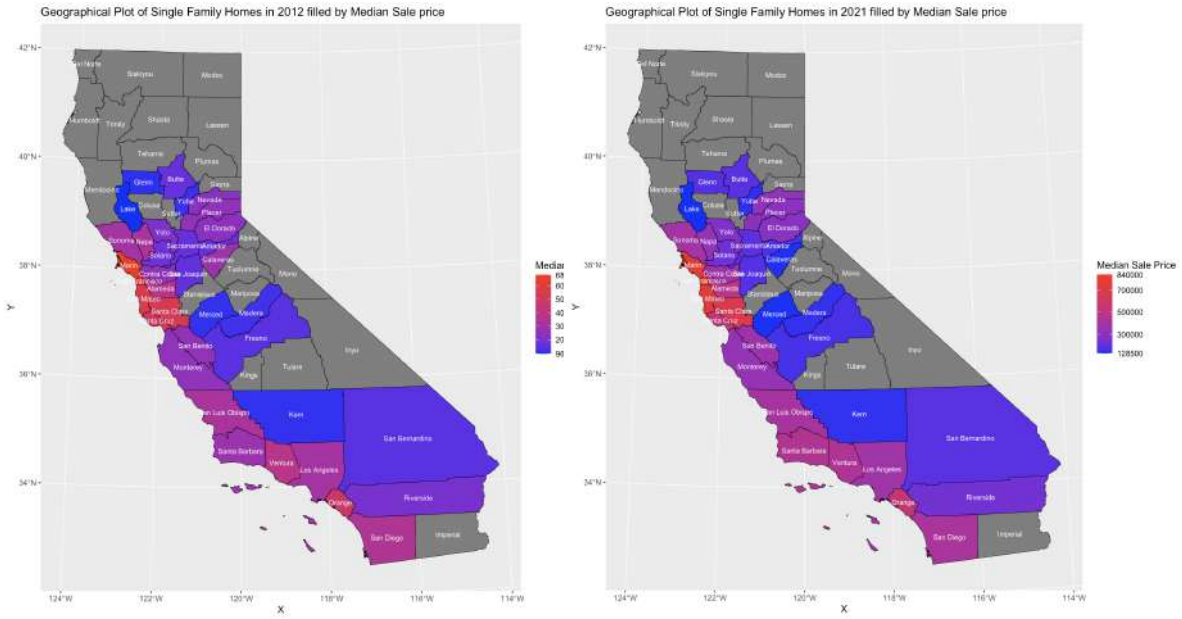
Comparing these results to those in Figure 2.2, there is a lot of resemblance and crossover, except that information was lost on some counties in 2021. For the percent change plots all counties received gains in Figure 2.2(c), however in Figure 2.5(c), Calaveras county experienced a loss in growth. Nevertheless, Calaveras county was still the only county shaded in a hue of blue for both versions of the

figure meaning it was at the lower end of the spectrum.

Figure 2.6, depicts pricing analysis for condominiums containing three plots just as shown in Figure 2.5. Figure 2.6(a) displays that the less expensive counties were Kern, Fresno, San Joaquin, Sacramento and Solano. The most expensive counties here were in the Bay Area and there is a recurring trend that coastal counties tend to be more expensive than inland counties. Figure 2.6(b) has similar results to that of Figure 2.6(a); however, one can bear to notice that the legend shows a price increase from 2012 to 2021. In 2012, the minimum price of condominiums were $\$46,500$ and in 2021, the minimum increased to $\$68,750$. Similarly, the maximum price of condominiums was $\$649,000$ in 2012, and increased to $\$720,000$ in 2021. These price changes are indicative of inflation in the housing market. This issue of increasing house prices results from the varied demands in each county and can also be attributed from the rich landscapes some counties have. Figure 2.6(c) illustrates how this demand has affected price change between 2012 and 2021. The counties that accrued the most change in price were Santa Barbara and Placer counties, while San Bernardino, Ventura and San Francisco had the least percent change.

Comparing the results from Figure 2.6 to those from Figure 2.3, the observations made for Figure 2.6(a) and 2.3(a) were the same regarding the most affordable counties and that the most expensive counties were located in the Bay Area. Now making a comparison with Figures 2.6(b) and 2.3(b), both shared in common that Fresno and Kern were affordable counties compared to surrounding counties. However, in Figure 2.6(b) demonstrated more counties shaded in blue than those found in figure 2.3(a).

Lastly, comparing Figure 2.6(c) and 2.3(c), percent change dropped, previously the minimum percent change for condos was around 100% and the maximum was

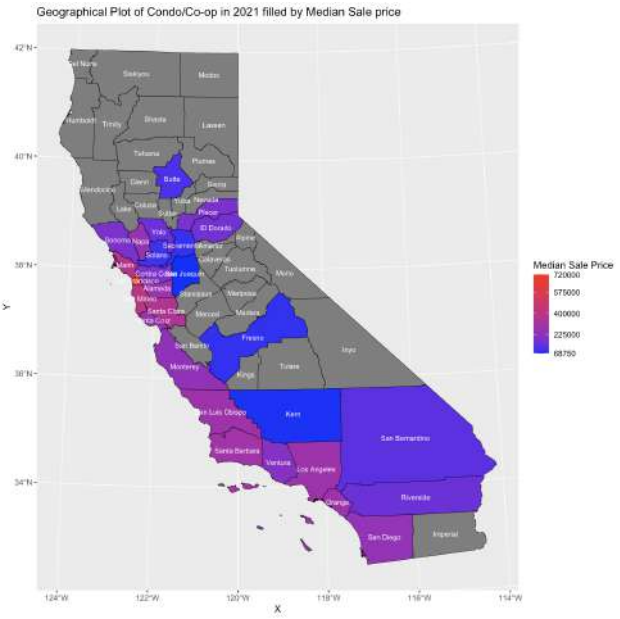


(a) Price for Single Family Homes in 2012 (b) Price for Single Family Homes in 2021

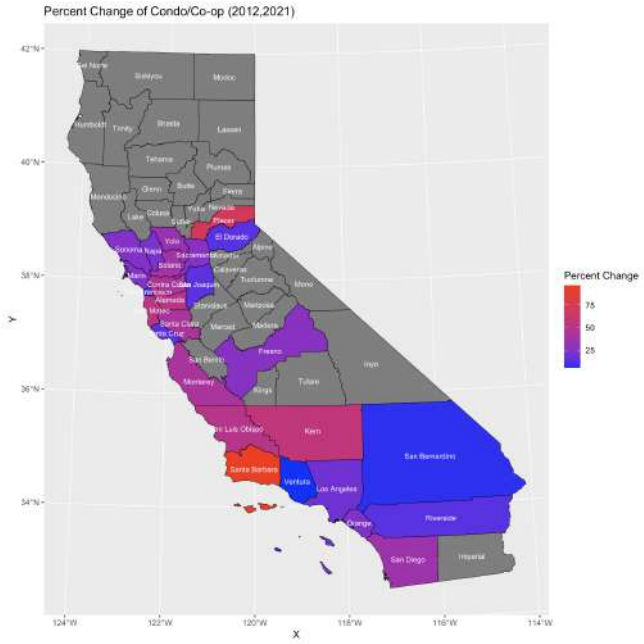
(c) Median Sale Price Percent Change

Figure 2.5: Single Family Residential Geographical Plots with clean data

(a) Price for Condos/Co-ops in 2012



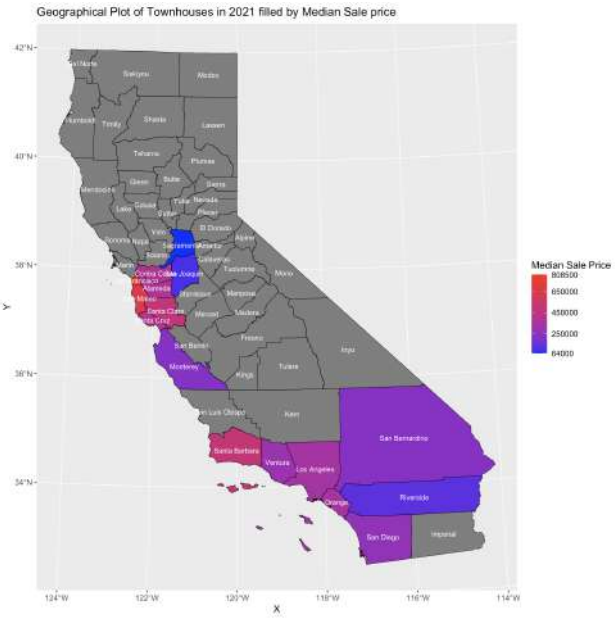
(b) Price for Condos/Co-ops in 2021



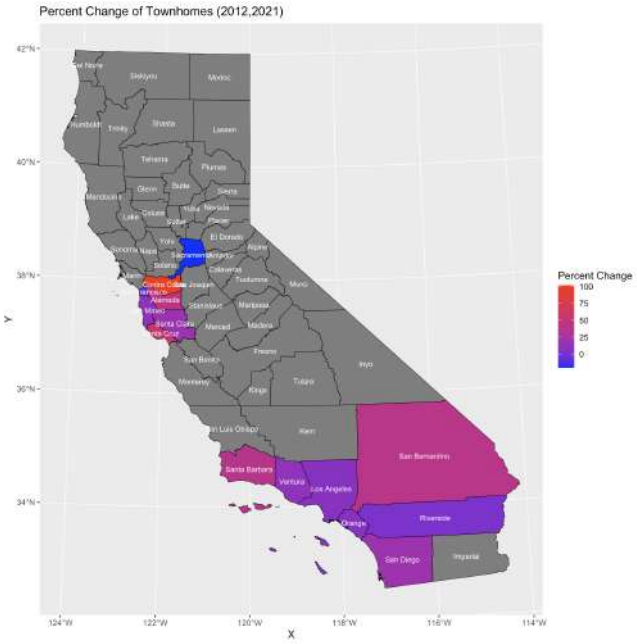
(c) Median Sale Price Percent Change

Figure 2.6: Condos/Co-ops Geographical plots with clean data

(a) Price for Townhomes in 2012



(b) Price for Townhomes in 2021



(c) Median Sale Price Percent Change

Figure 2.7: Townhomes Geographical plots with clean data

above 500%, whereas in the cleaned version of the dataset the minimum was around 10% and the max was around 96%. There seemed to be some differences between these two versions of the plot, in Figure 2.6(c) there was no data on December of 2021 for Merced county, which was not the case for Figure 2.3(c); Merced county experienced one of the highest percent change for condos. This was likely due to less availability of Condos for sale. That is, once we excluded months that sold less than 5 homes (or condos in this case) it removed this month's worth of data for Merced county of the condo property type. More than likely, the number of condos sold this month were lower than usual and those sold were overpriced, thus experiencing a sky-high percent change. In the new version of the percent change plot, San Bernardino and Ventura counties were now included in the lower percent change bracket, and Placer and Santa Barbara counties sought the highest percent change.

Figure 2.7 contains the same figures as the previous two shown except now we are analyzing behavior for townhomes. The first difference noticed between Figure 2.7 and 2.4 is that now our data is more limited in regards to how many counties we have data for. Figure 2.7(a) shows that in 2012 the most costly counties were San Francisco and San Mateo while the least expensive were Riverside and Sacramento counties. Figure 2.7(b) shows an additional county for December of 2021 which is Monterey county. Otherwise, we have similar analysis from Figure 2.7(a). In addition, the counties in the Bay Area shifted closer to a red hue, meaning that we saw more homes characterized in the upper price bracket. Figure 2.7(c) implies that Sacramento and San Francisco counties displayed the least change in median sale price, whereas Contra Costa county displayed the most percent change. Since San Bernardino county did not seem to shift in color dramatically, it was not expected

to have experienced the percent change reported. San Bernardino county shows a percent change of at least 50%, where we can conclude that this was the inland county that demonstrated the most change in median sale price. A comparison between Figure 2.7 and Figure 2.4, shows the outliers that were present in the data in raw form. For instance, in Figure 2.4(b) we saw that the maximum for median sale price in December of 2021 was about $1.3 million, and in Figure 2.7(b) this amount dropped to around $800,000.

Also, Figures 2.4(c) and 2.7(c) show there was a drastic difference on the shading. Most counties in Figure 2.4(c) were shaded between dark purple and blue and only San Joaquin county was shaded in red, meaning San Joaquin county contained an outlier that caused a heavy tailed distribution on median sale prices for townhomes, causing all other counties to have changed less over time in comparison. Otherwise, we can conclude that there is a similar trend for all these figures, that is the Bay Area tended to be more expensive than inland counties and southern coastal counties. Also, coastal counties in general tend to be in the higher price bracket compared to their inland counterparts. Therefore, these geographic analysis tend to be dependent and correlation. Nearby counties tend to have similar prices compared to more distance counties.

2.2.4 Scatter Plots investigating Percent Change

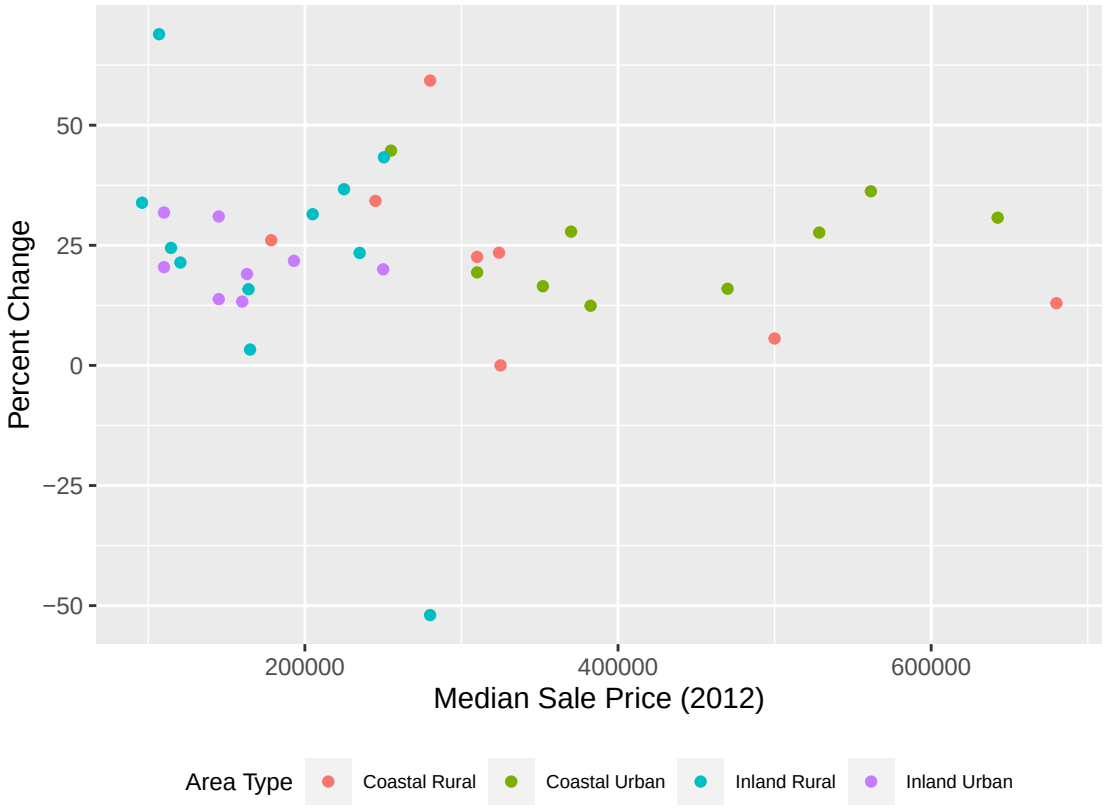


Figure 2.8: Scatter plot for Percent change vs Median Sale Price for Single Family Homes

The last set of figures are scatter plots for Percent Change vs Median Sale Price for February of 2012. Figure 2.8 is a scatter plot for Single Family Residences and it clearly displays how inland counties are in one cluster. Coastal counties are in shown in another cluster in terms of median sale price, where coastal counties are seen as experiencing higher prices compared to inland counties. The percent change tells a different narrative, that no one area had larger percent change when compared to other areas. However, when accounting for percent change by county, Figure 2.8 displays that Calaveras county resulted in a 51% loss with our preprocessed

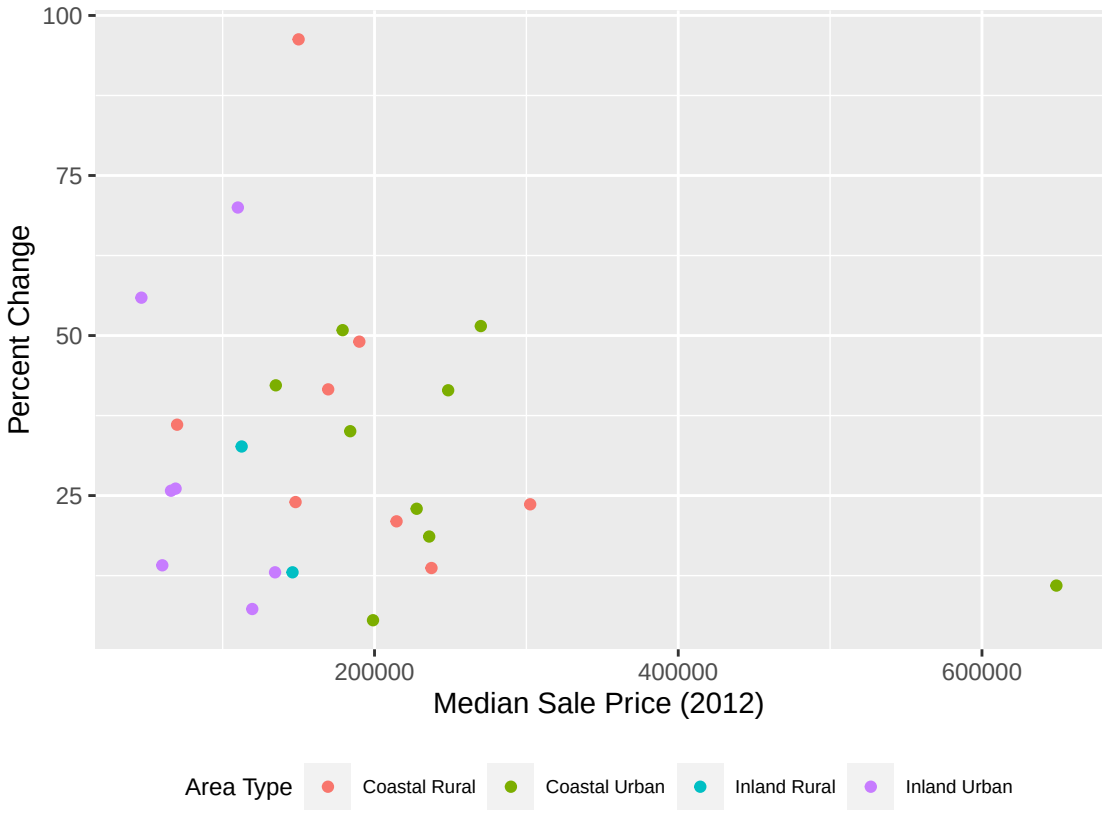


Figure 2.9: Scatter plot for Percent change vs Median Sale Price for Condos

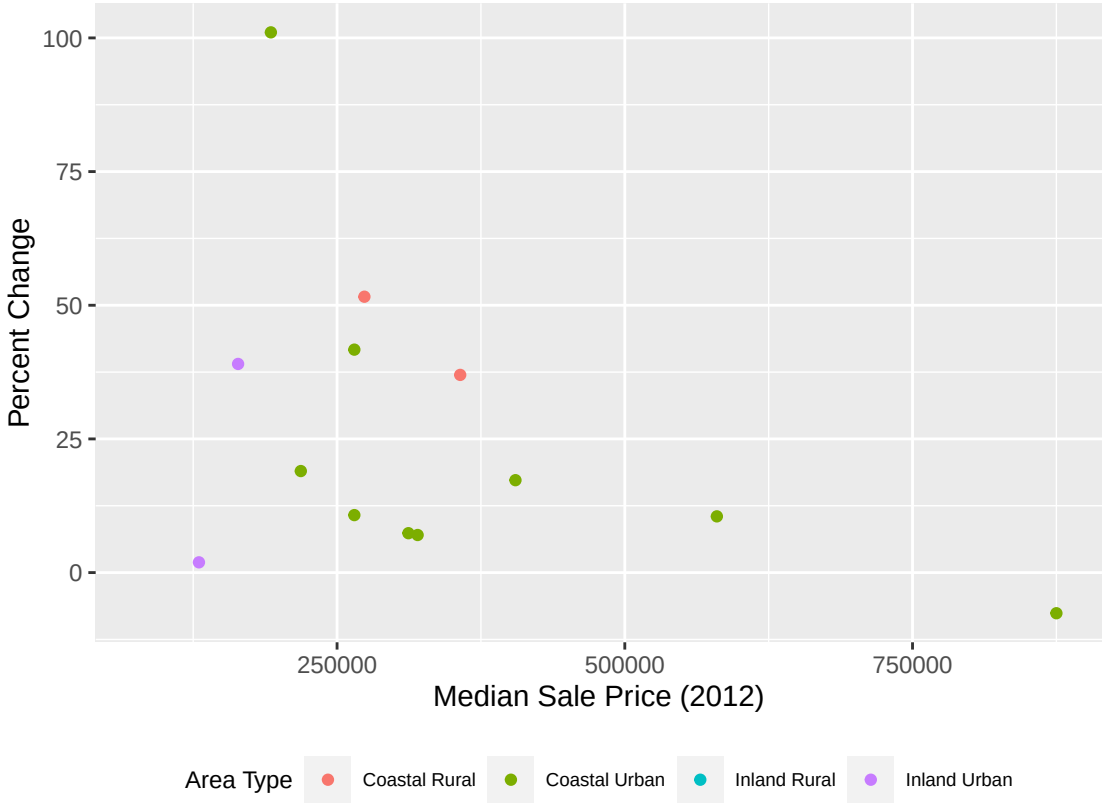


Figure 2.10: Scatter plot for Percent change vs Median Sale Price for Townhouses

data. Similarly in Figure 2.9 and Figure 2.10, coastal counties tended to be more expensive than their inland counterparts and here wasn't a clear distinction for percent change. In Figure 2.10, Sacramento and San Francisco resulted with a 20% and 7.6% percent losses respectively.

2.3 Longitudinal Analysis

The core foundation of longitudinal studies is that measurements of the same individual are taken at multiple points in time, allowing the direct study of change over time [5]. On the contrary, a cross-sectional study observes individuals who appear once [5]. This thesis study is a prime example of a longitudinal study, where the individuals are the counties, which are repeatedly collected monthly over 10 years. The primary focus of this type of study is to assess individual specific changes or formally known as within-individual changes. More specifically, we are interested in the median sale prices by county throughout the duration of the study. A longitudinal analysis becomes a requirement in order to assess differences amongst each county, and the changes within each county. We will handle this analysis by performing visualizations that will describe patterns and trends for median sale prices.

In some instances, the number of occasions recorded may be different for each individual in the study. It is rarely ever the case where data contains individuals that were all collected with the same number of occasions and at the same points of time. For example, in our data we have Los Angeles county recorded 360 number of observations throughout the span of 10 years (3 times a month due to 3 property types, thus 36 number of occasions for each year) and this is repeated on a monthly

basis. Therefore, a subset of our data that only contains Los Angeles county is considered as complete. However, this is not the case for all counties, since some counties were not recorded at the same number of occasions (missing data for some properties types in a month, or no data available for a month(s) for a specific county), meaning our dataset as a whole is not complete and inconsistent.

2.3.1 Spaghetti Plots with Raw Data

Before proceeding with the plots, we needed to break up counties into divisions in order in order to view the data from a granular perspective. Henceforth, we will use the area type variable we feature engineered containing the four categories: coastal rural, coastal urban, inland rural and inland urban.

The x-axis contains the month, the y-axis is median sale price and a line was created for each county. These lines were colored by the area_type variable and would indicate if one category experienced noteworthy changes compared to other categories. Note that these plots produce several lines, which are known as "spaghetti plots." When plotting these figures it made aware that there were some outliers that had considerably higher median sale prices. Upon further inspection, many of these observations in their respective counties and property types, have sold a fraction of the houses compared to previous months. This is where we will see the outlying observations as mentioned in the data removal section.

Figure 2.11 is a prime example of a spaghetti plot involving all counties for single family homes; however, do note that when involving all individuals in this type of plot makes it difficult to distinguish between-subject variability and within-subject variability [5]. Fitzmaurice et al. stated that between-subject variability articulates natural variation in individuals' tendency to respond [5]. For instance,

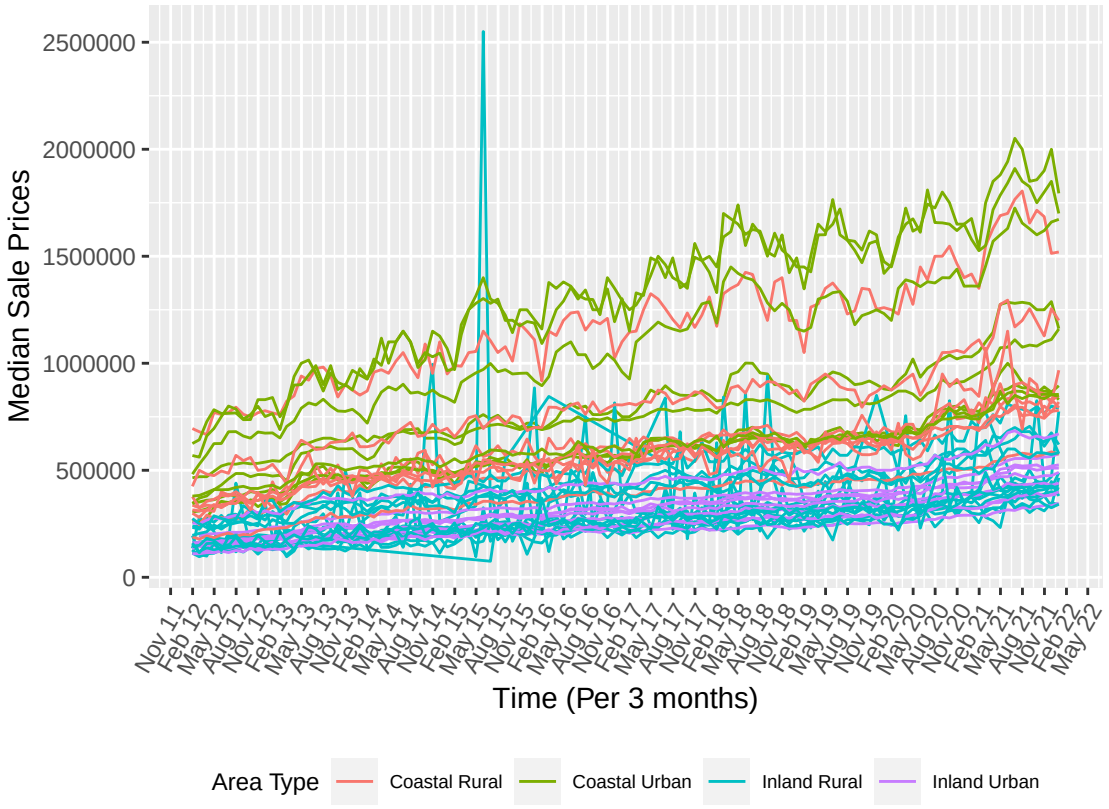


Figure 2.11: Spaghetti plot: Median Sale Price vs Time for Single Family Homes using Raw Data

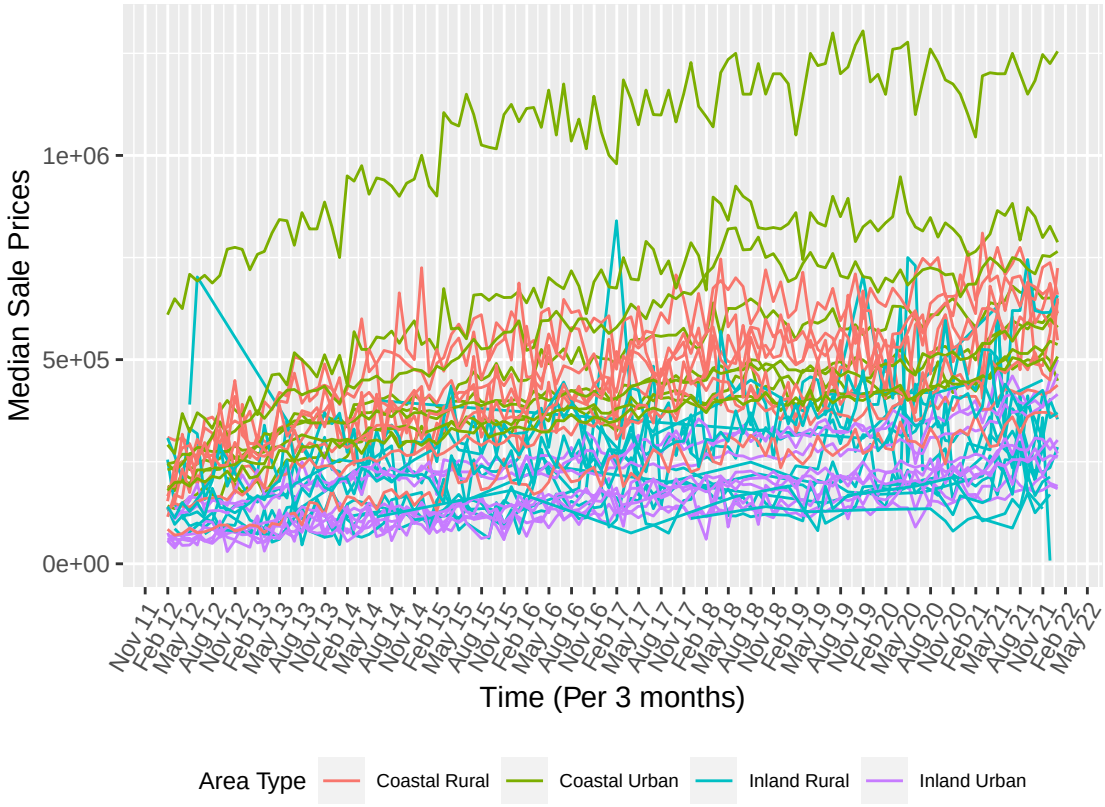


Figure 2.12: Spaghetti plot: Median Sale Price vs Time for Condo/Co-op using Raw Data

it may be seen that some counties consistently have higher median sale prices than the average, while other counties below the average. Therefore, one source of positive correlation among the repeated measures is the variability in the response variable between individuals in the population. In addition, we may also have diversity between counties meaning that some counties may have different trajectories over time. That is, some counties may have a higher increase/decrease or lower increase/decrease of median sale price over time. Another source of variation is within-individual variation which is individual-specific and unpredictable [5].

In Figure 2.11 it is inherent that one county that is inland and not populous has outlying observations, meaning that the median sale price is unusually high.

In these circumstances, we don't want our data to influence our prediction models and give inaccurate representations in the future. As mentioned earlier, in many occasions only less than 5 homes were for sale during this certain month periods for a particular county and property type. That is, in some particular months, only a few homes were for sale and these particular homes were expensive and outlandish homes that were not typically for sale often. The same observations follows for some inland rural counties for condiminiums in Figure 2.12. In Figure 2.13, it was noted that there were outliers for the townhouse property type across coastal rural, coastal urban and inland urban areas. As mentioned earlier, this was due to outrageously priced homes up for sale and less inventory for that property type in certain months.

Spaghetti plots are also introduced to determine any seasonality and trends of variables and are an extension of longitudinal data. The lines shown are colors by area type, which is a variable we featured engineered depending on the geographical location of the county in question. For Figure 2.11 there is an upward trend with some counties having a steeper slope than others. Other than the outliers, coastal urban counties (in green) tend to contain highest priced single family homes and tend to have higher gains than counties in other areas. Coastal not populous counties (red) had similar gains, however tended to have lower prices than coastal urban counties. Inland urban (purple) counties showed a small rate of change compared to other areas (i.e., less steep compared to coastal areas), but still had some gains. Inland not populous counties were hard to determine since there was an abundance of fluctuation. Figure 2.12 is another spaghetti plot now depicting median sale prices over time for condos. Visualizations from the previous figure apply here as well, where the most expensive counties were coastal urban, followed

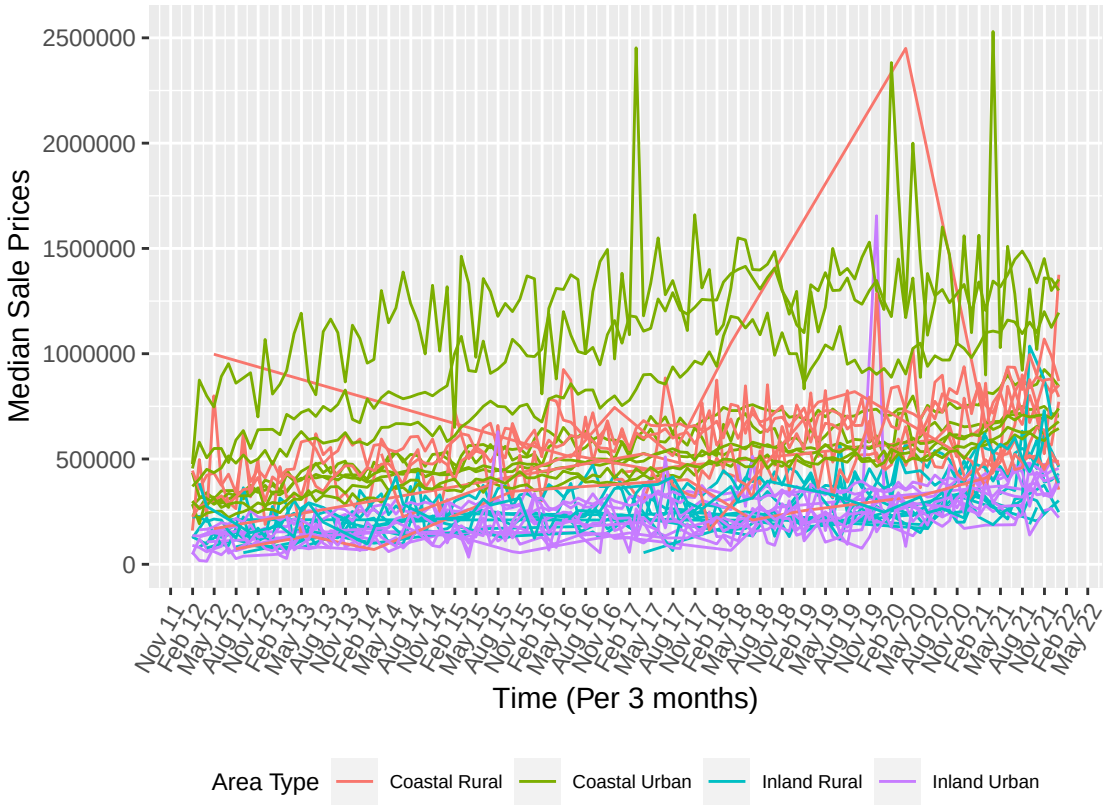


Figure 2.13: Spaghetti plot: Median Sale Price vs Time for Townhouses using Raw Data

by coastal not populous counties both with similar gains. Inland urban counties had small gains and inland not populous counties were still sporadic in nature. Figure 2.13, shows that townhouse properties did not have as steep of an increase compared to single family residences. Townhomes do follow the trend where coastal counties tend to have a higher median sale price compared to inland counties.

2.3.2 Spaghetti Plots with Pre-processed Data

Figures 2.14, 2.15 and 2.16 are all spaghetti plots as seen earlier in this study, which now plots the clean data. A quick comparison between with all plots shows outliers that were removed. The removal of these outliers are crucial since these could have a big impact on our analysis and skew results for prediction models, deeming them as less accurate for our studies.

For example, looking at Figure 2.11 shows two inland rural counties that stood out with an extreme jagged behavior, but now in Figure 2.14 we see those lines have been smoothed. Therefore, the removal of less than 5 homes sold has stripped the huge fluctuations of median sale price in these counties. Now comparing Figures 2.12 and 2.15 we can interpret the same differences for condominiums where the latter figure becomes smoother due to removal of outlying observations. However, when now comparing Figures 2.13 and 2.16, some lines for townhomes became relatively flat due to less observations present. There were a few observations from a coastal urban county that was exceeding 2 million dollars, thus we removed these. We also removed one observation from a coastal rural county that had a high median sale price in December, 2019 compared to neighboring months.

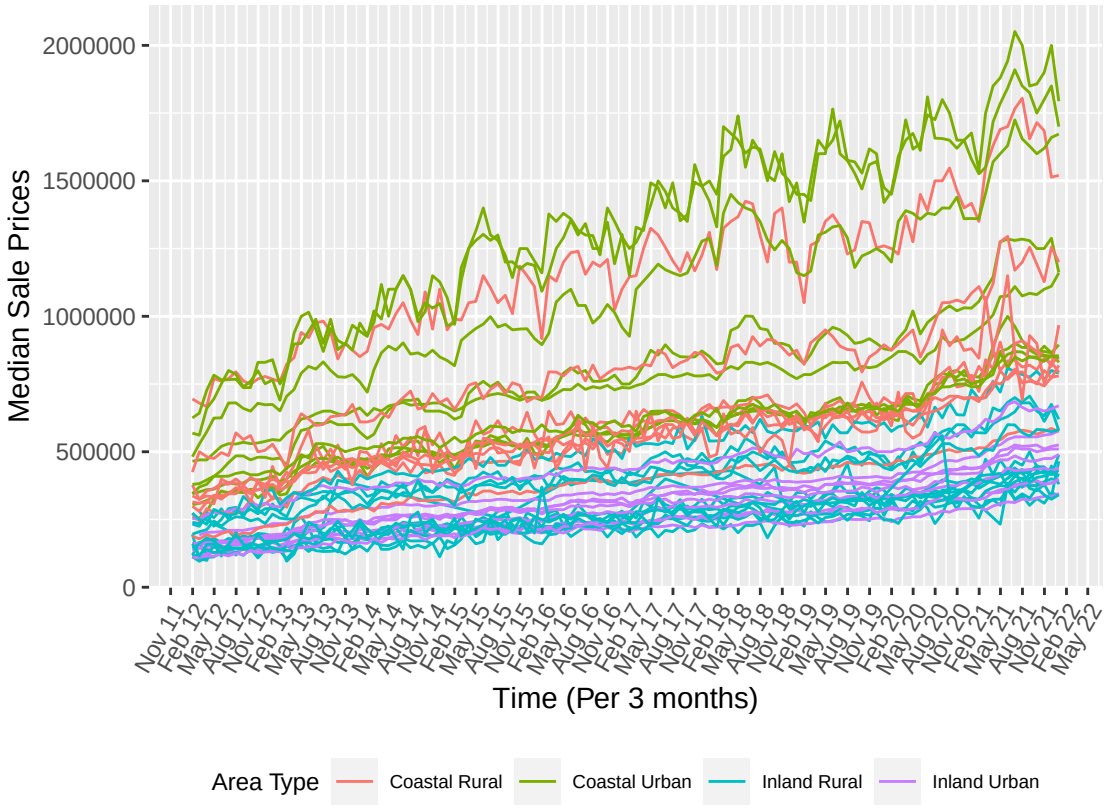


Figure 2.14: Spaghetti plot: Median Sale Price vs Time for Single Family Homes with clean data

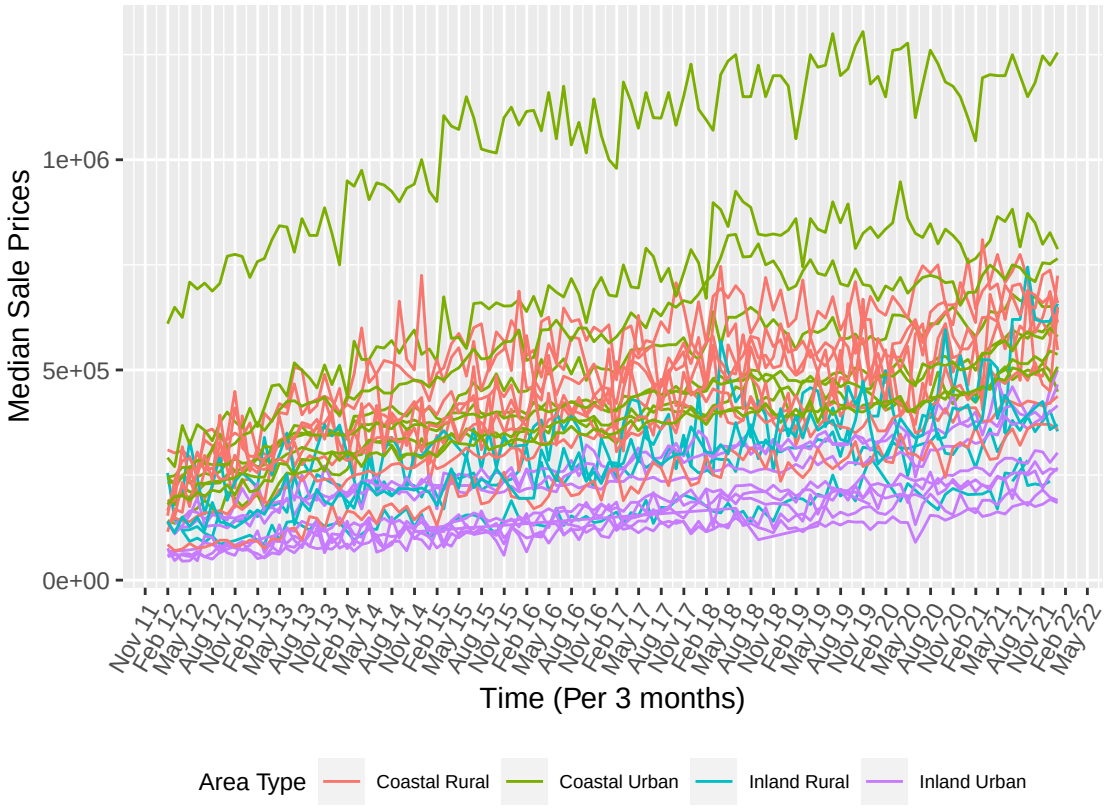


Figure 2.15: Spaghetti plot: Median Sale Price vs Time for Condo/Co-op with clean data

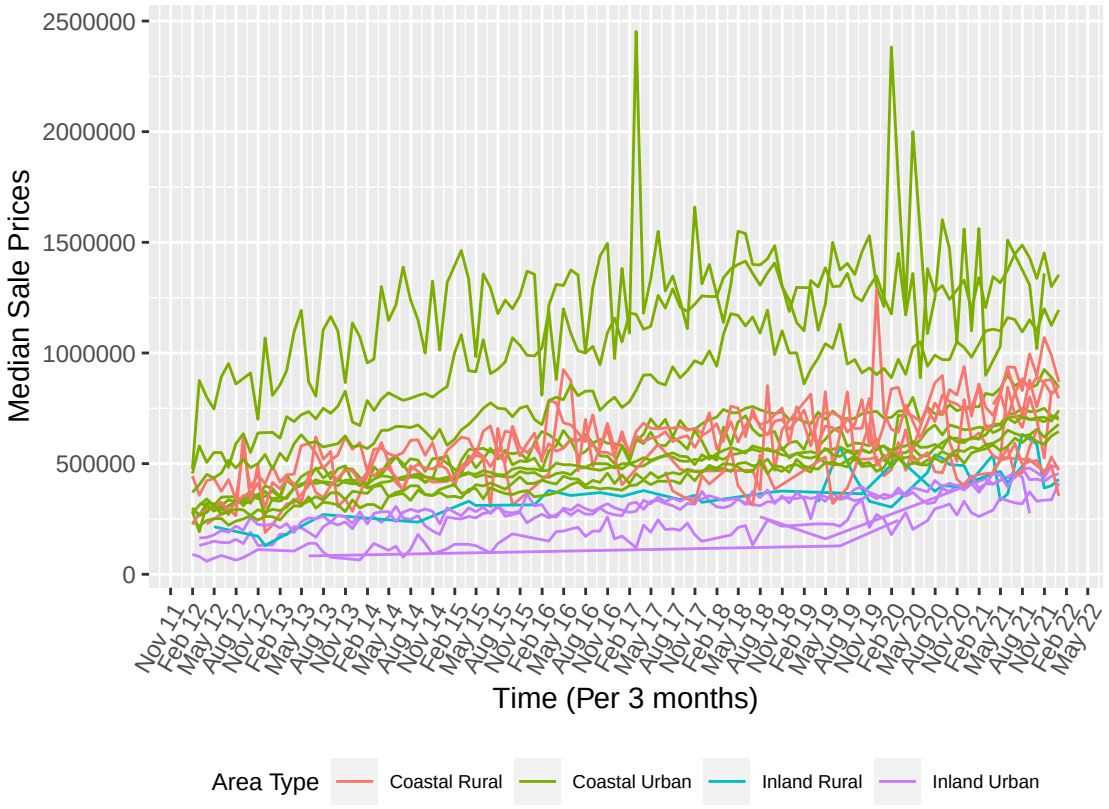


Figure 2.16: Spaghetti plot: Median Sale Price vs Time for Townhouses with clean data

Chapter 3

Model Building

Moving forward, we will now fit our data to a variety of models. In the following sections, we will discuss the different model types we used and give a concise background on how these models work and how they are formulated. The models that will be used are Linear Regression, Generalized Additive Models and Linear Mixed effect models. In each section, our goal is to fit, visualize and assess the performance of these models. The importance of modeling to predict home price values is to provide valuable insights and forecast values based on the available data in hand. Predictive modeling is a renown tool in the statistical realm that plays a vital role in various fields and industries. Taking advantage of this tool will help to help understand the changes the housing market could face over time.

3.1 Linear Regression

Linear regression is an elementary predictive method to create fitted values, in which case we can assess the performance of these models by visualizations and observing the output summary of the models. These models have a response

variable (dependent variable, the variable we wish to predict) and the predictors (independent variables). In R, we may use different libraries to code the formula for a linear regression model which involve a variety of expressions which may only include one predictor or a variety of combinations of predictors.

The most basic form is below:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{3.1}$$

where Y is the variable we wish to predict, X is the independent variable used to predict the value of Y, meaning there is some belief that X has some effect with the response. β_0 is the intercept term that is the value of Y when X is set to zero. β_1 is the slope of the line representing the change in Y for one unit change in X. Lastly, ϵ is known as the error term, which mathematically speaking is the difference between the observed value of Y and the mean value of Y.

In the case that we have a more complicated model, the formula now becomes:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \tag{3.2}$$

where X_1 and X_2 are the predictors and β_1 and β_2 are the slopes respectively (i.e., change in Y for one unit change in X_1 or X_2). The fourth term in this formulation is an interaction term which allows to inspect how the relationship between Y and the variable X_1 changes depending on the values of the variable X_2, which is the exact definition of our β_3 coefficient. That is, the above formulation can test on the differences of responses from one independent variable and at different levels of another independent variables. Note that this is just an example; we may have more complicated terms such as involving transformations and including

polynomials, which we will explain in detail in the upcoming subsections. The objective of regression analysis is to determine a line that best fits the observation points and allows us to understand the relationship between the response and the predictors.

3.1.1 Cubic Polynomial on Time

We commence by fitting a model that included the variable time_since as a 3^{rd} degree polynomial, which as mentioned above is a variable that is the elapsed time set monthly. Since there are 120 months available in our data, the time variable starts at time 0 (first month of recording) and ends at time 119 (the last month of recording). Thus, the model formulation looks like:

$$\begin{aligned} Y = \beta_0 + \beta_1 X_c + \beta_2 X_{ir} + \beta_3 X_{iu} + \beta_4 X_t + \beta_5 X_t^2 + \beta_6 X_t^3 + \beta_7 (X_c : X_t) + \\ \beta_8 (X_{ir} : X_t) + \beta_9 (X_{iu} : X_t) + \beta_{10} (X_c : X_t^2) + \beta_{11} (X_{ir} : X_t^2) + \\ \beta_{12} (X_{iu} : X_t^2) + \beta_{13} (X_c : X_t^3) + \beta_{14} (X_{ir} : X_t^3) + \\ \beta_{15} (X_{iu} : X_t^3) + \epsilon \end{aligned} \tag{3.3}$$

where Y is the median_sale_price variable, and β_0 is the intercept or the median sale price value at time 0. X_t is the time_since variable, and X_c, X_{ir}, X_{iu} are the factor levels for the area_type variable. The base level for area_type is (level 1) is the coastal_not_populous, and level 2, 3, and 4 are coastal_urban, inland_not_populous and inland_urban respectively (from this point on the levels will referred to as coastal rural, coastal urban, inland rural and inland urban respectively). The ":" symbol represents interaction terms and β_1 through β_{15} are all slopes. Note that, if an observation falls within one of the area type categories, it will be replace with an

Table 3.1: Table of Coefficient estimates for cubic polynomial model stratified by Single Family Residential

Coefficient	Estimate	**Coefficient**	Estimate
β_0	341490.31738	β_8	-6024.97378
β_1	91367.64489	β_9	-5215.89058
β_2	-170550.77294	β_{10}	-10.50925
β_3	-193981.59777	β_{11}	87.29674
β_4	10492.13750	β_{12}	67.52180
β_5	-138.06122	β_{13}	0.02431
β_6	0.79449	β_{14}	-0.49564
β_7	2492.70469	β_{15}	-0.37546

"1" and if not "0". That is, if we are fitting an observation that is coastal urban, X_c will be a "1" while X_{in} and X_{iu} will be "0", meaning we are left with a regression line that looks like:

$$\begin{aligned} Y = \beta_0 + \beta_1 X_c + \beta_4 X_t + \beta_5 X_t^2 + \beta_6 X_t^3 + \\ \beta_7 (X_c : X_t) + \beta_{10} (X_c : X_t^2) + \\ \beta_{13} (X_c : X_t^3) + \epsilon \end{aligned} \tag{3.4}$$

The benefits of having a polynomial variable allows us to capture nonlinear relationships between median sale prices (response) and the elapsed time (predictor). In addition, this will allow for greater fit to our data, meaning our model becomes more flexible and results in less bias and lower residuals. However, some drawbacks may be that we may be overfitting observations and can lead to less interpretability.

Figures 3.1, 3.2 and 3.3 displays the regression lines plotted by area type and

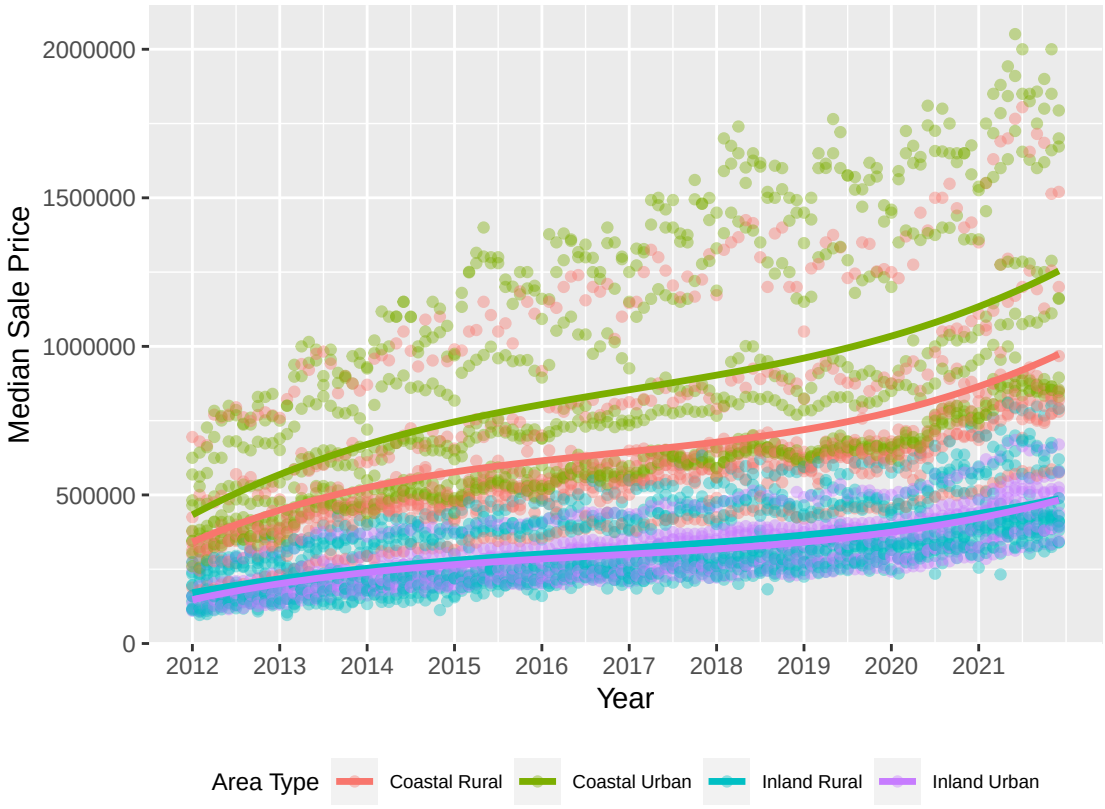


Figure 3.1: Regression lines for Cubic Polynomial of Time for Single Family Residential

stratified by property type. In other words, this first plot is a result from a model fitted using a subset of our preprocessed data that contains only Single Family Residential observations. The general trend of Figure 3.1 is that the regression lines are increasing over time, with a slight uptick towards the last year. Now observing the positions of these curves, the coastal urban curve is the highest and inland urban is the lowest. Overall, coastal counties tend to have higher median sale prices than inland counties; while inland rural and inland urban are only marginally different.

Note that from Equation 3.4, $\beta_0 + \beta_1$ is the coefficient which is the intercept for the coastal urban area type and is the highest intercept out of the four area types in our data as shown in Table 3.1. The lowest intercept is inland urban counties, the purple line, which was not expected to its counterpart, inland rural areas, as exhibited from Table 3.1 the intercept will become $\beta_0 + \beta_3 = \$341,490.32 - \$193,981.60 = \$147,508.72$, a substantial drop. Respectively, the intercepts for inland rural and coastal rural are $\$341,490.32$ (the base level) and $\$170,939.55$.

We initially thought that including the interaction term between time and the area type would delineate a difference in trend lines between each area type. However, the results in Figure 3.1 say otherwise, since the results are relatively parallel to one another, demonstrating little to no interaction between area type and time.

Figure 3.2 now fits regression lines for the condo property type. Reiterating the comments made from Figure 3.1, we have an increasing trend. Meaning our models predicted that median sale prices increased over time on average. In contrast to what we have seen with single family residential, the trends here seem to be flatter which slower increase perceived during the economic recovery in mid 2020. However, it still holds true that the coastal urban area is the highest line and the median

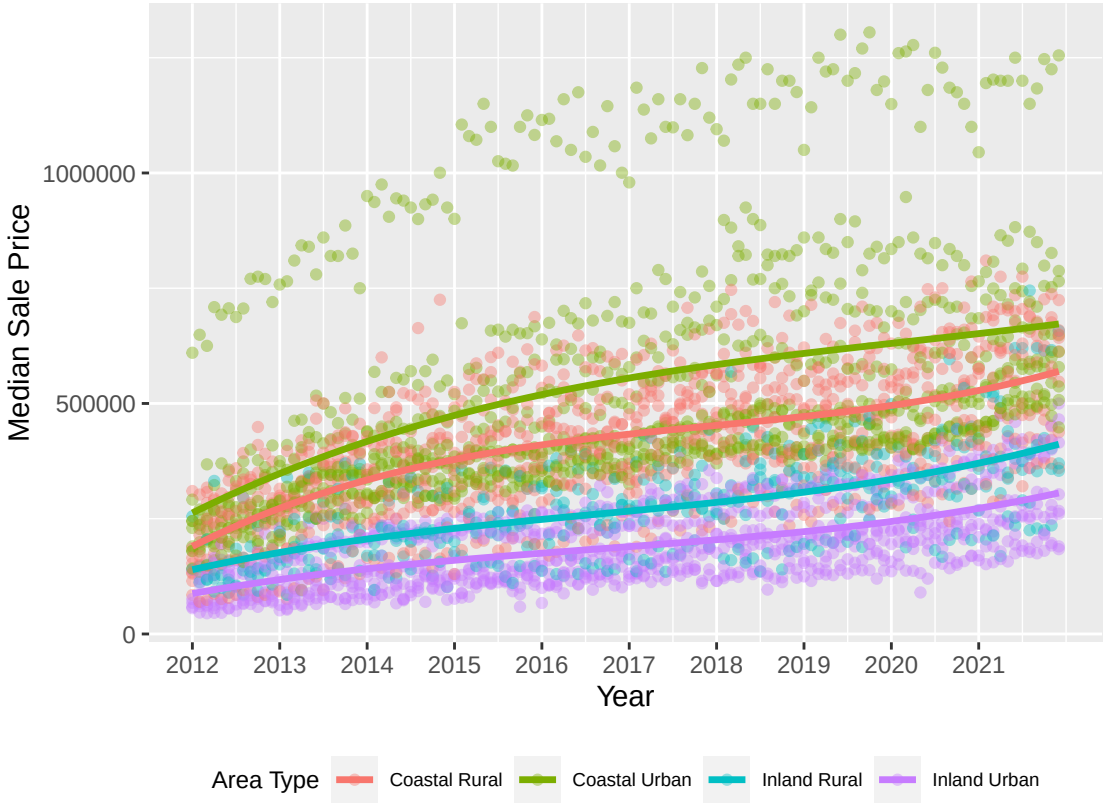


Figure 3.2: Regression lines for Cubic Polynomial of Time for Condos

Table 3.2: Table of Coefficient estimates for cubic polynomial model stratified by Condominiums

Coefficient	Estimate	**Coefficient**	Estimate
β_0	189820.9489	β_8	-4402.3731
β_1	72721.3053	β_9	-5050.7963
β_2	-50260.4995	β_{10}	27.6452
β_3	-102194.3623	β_{11}	52.7920
β_4	7917.1214	β_{12}	58.8044
β_5	-89.2663	β_{13}	-0.2079
β_6	0.4165	β_{14}	-0.1967
β_7	-92.3064	β_{15}	-0.2334

sale prices tend to be higher for coastal counties than inland counties.

Table 3.2 tells us that if we are in the coastal urban area the intercept is $\beta_0+\beta_1 = \$189,820 + \$72,721 = \$262,541$ which is the highest price at time 0, while the other areas will subtract from the β_0 intercept, suggesting lower intercepts at time 0. Figure 3.2 delineates that an interaction here was required.

Now, turning our attention to Figure 3.3, this now plots fitted regression lines for the townhouse property type. Inspecting the results from this figure, exhibits similar behavior to that seen in Figure 3.2, demonstrating a slight increase over time. That is, the predicted median sale prices exude a slight increase over time on average. Furthermore, these predictions imply that coastal counties have higher predicted values on average compared to inland counties. In contrast to Figure 3.1, this plot shows a flatter increase and does not have a noticeable uptick in summer of 2020.

Figure 3.3: Regression lines for Cubic Polynomial of Time for Townhomes

Table 3.3: Table of Coefficient estimates for cubic polynomial model stratified by Townhomes

Coefficient	Estimate	**Coefficient**	Estimate
β_0	306733.2190	β_8	-4032.1495
β_1	61942.9755	β_9	-5108.6652
β_2	-163116.0272	β_{10}	54.6172
β_3	-184034.4211	β_{11}	61.3483
β_4	10226.1992	β_{12}	57.5568
β_5	-124.8100	β_{13}	-0.3191
β_6	0.5712	β_{14}	-0.2721
β_7	-1316.9193	β_{15}	-0.1861

Table 3.3 has similar findings to the previous tables, however at a slight different scale and by far single family homes have the highest base intercept compared to other property types. Nevertheless, over time (linearly) we observe similar growth on average townhomes to single family residences.

Overall, in this section we have found that single family residences tend to be the most expensive property type on average. Coastal urban was consistently the area with the highest median sale price and coastal rural always followed in second place. It is interesting to note that inland princes don't vary for single family residences, but do for other property types.

3.1.2 Quintic Polynomial on Time

Now we turn to a more complicated model which is a 5th degree polynomial, which specifically entails the factored area type variable having an interaction with

5th degree polynomial of time. These more complex models will now have more coefficient estimates than the cubic polynomial model. Models that include higher degree terms may assist in implying a more complex relationship between time and median sale prices. That is, we may need a curve that captures the seasonality of our data. Equation 3.5 is a formulation of the model we used to fit our data with a 5th degree polynomial on time and interaction with the area type variable.

$$\begin{aligned} Y = \beta_0 + \beta_1 X_c + \beta_2 X_{ir} + \beta_3 X_{iu} + \beta_4 X_t + \\ \beta_5 X_t^2 + \beta_6 X_t^3 + \beta_7 X_t^4 + \beta_8 X_t^5 + \\ \beta_9 (X_c : X_t) + \beta_{10} (X_{ir} : X_t) + \beta_{11} (X_{iu} : X_t) + \\ \beta_{12} (X_c : X_t^2) + \beta_{13} (X_{ir} : X_t^2) + \beta_{14} (X_{iu} : X_t^2) + \\ \beta_{15} (X_c : X_t^3) + \beta_{16} (X_{ir} : X_t^3) + \beta_{17} (X_{iu} : X_t^3) + \\ \beta_{18} (X_c : X_t^4) + \beta_{19} (X_{ir} : X_t^4) + \beta_{20} (X_{iu} : X_t^4) + \\ \beta_{21} (X_c : X_t^5) + \beta_{22} (X_{ir} : X_t^5) + \beta_{23} (X_{iu} : X_t^5) + \epsilon \end{aligned} \tag{3.5}$$

Since this is now a 5th degree polynomial, we have 23 coefficient estimates, where the first two lines of the equation are the factor variable terms only and the time variable terms from degree 1 to degree 5. The 5 lines after are the interaction terms with the respective factor level and the polynomial terms. Note that we should proceed with caution with models with high degree polynomials like this one, since the risk of overfitting is high. Also, as the degree fit is higher, it will becomes more challenging to interpret models. In the case with the model for Equation 3.5, if for an observation for a given month in time, for example in the

inland urban area, the equation becomes:

$$\begin{aligned} Y = \beta_0 + \beta_3 X_{iu} + \beta_4 X_t + \beta_5 X_t^2 + \beta_6 X_t^3 + \beta_7 X_t^4 + \beta_8 X_t^5 + \\ \beta_{11}(X_{iu} : X_t) + \beta_{14}(X_{iu} : X_t^2) + \beta_{17}(X_{iu} : X_t^3) + \\ \beta_{20}(X_{iu} : X_t^4) + \beta_{23}(X_{iu} : X_t^5) + \epsilon \end{aligned} \tag{3.6}$$

More specifically, if we are looking at San Bernardino county, which is considered an inland urban area, from the month of December of 2021 (the last month of our data collection), then we have equation 3.7.

$$\begin{aligned} Y = \beta_0 + \beta_3 \cdot 1 + \beta_4 \cdot 119 + \beta_5 \cdot 119^2 + \beta_6 \cdot 119^3 + \beta_7 \cdot 119^4 + \beta_8 \cdot 119^5 + \\ \beta_{11} \cdot 1 \cdot 119 + \beta_{14} \cdot 1 \cdot 119^2 + \beta_{17} \cdot 1 \cdot 119^3 + \\ \beta_{20} \cdot 1 \cdot 119^4 + \beta_{23} \cdot 1 \cdot 119^5 + \epsilon \end{aligned} \tag{3.7}$$

We used equations for each area type to derive the optimal line of fit, which resulted in four distinct lines. Figure 3.4 brings forth these results where the lines and points are colored by the area type. Notice how in this plot we were able to properly fit data in the last two years worth of data due to the global pandemic, which lead a systemic rise in median sale prices. Consequently, affordability concerns for first time home-buyers arises due to this sharp increase. In Hermann's study, he found that the percent change in home prices was in the double-digits at the end of 2020, which was the first since the beginning of 2014. The global pandemic introduced in late 2019, reached its peak in spring and summer of 2020. During the peak of the pandemic, home prices remained stationary, but as the economy began recovery in mid 2020, prices sought a steep increase for coastal areas. Inland counties only had a slow and steady increase over time. Table 3.4 is a list of all coefficient estimates

Table 3.4: Table of Coefficient estimates for quintic polynomial model stratified by Single Family Residential

Coefficient	Estimate	**Coefficient**	Estimate
β_0	358186.03017241	β_{12}	-418.08900224
β_1	69563.12305430	β_{13}	17.51257618
β_2	-182735.29181455	β_{14}	-61.12853583
β_3	-207316.67459800	β_{15}	9.60763399
β_4	7897.14555487	β_{16}	0.22903286
β_5	-58.926957862	β_{17}	1.91616123
β_6	0.27820883	β_{18}	-0.09366366
β_7	-0.00439126	β_{19}	-0.00061702
β_8	0.00003963	β_{20}	-0.01721626
β_9	8912.74447756	β_{21}	0.00032255
β_{10}	-4010.93150123	β_{22}	-0.00001477
β_{11}	-2465.01163275	β_{23}	0.00004579

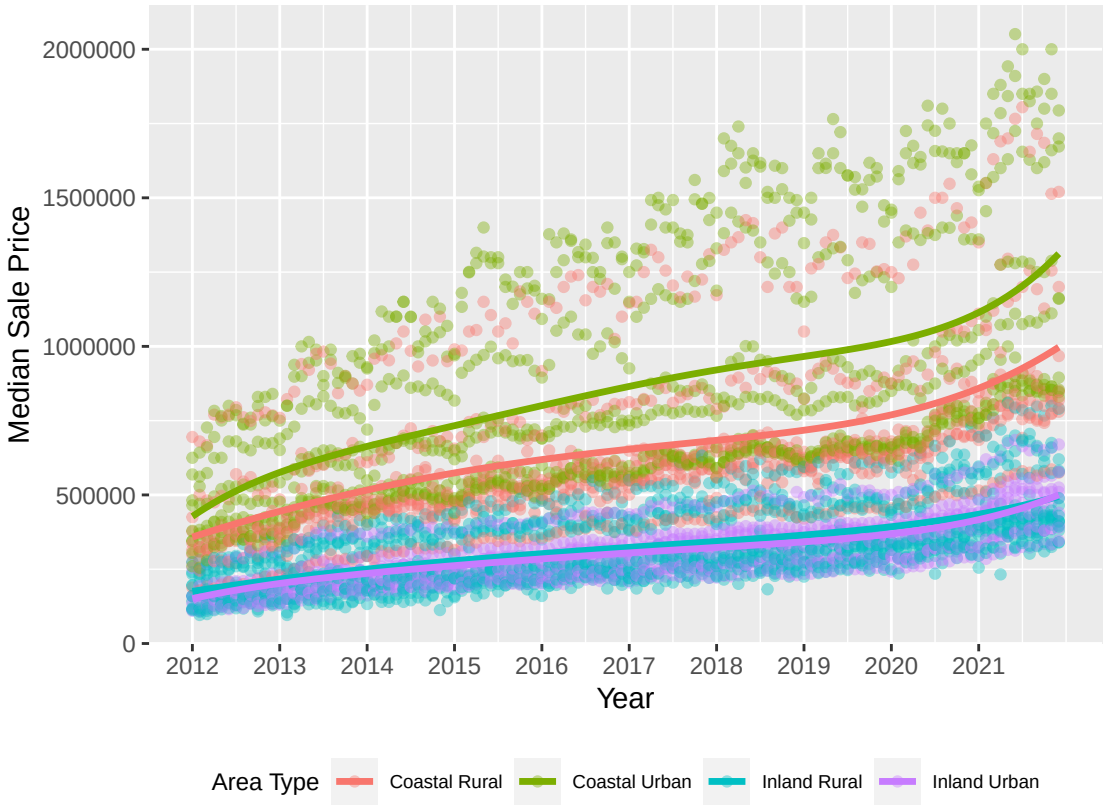


Figure 3.4: Regression line for Quintic Polynomial of Time for Single Family Residential

Table 3.5: Table of Coefficient estimates for quintic polynomial model stratified by Condominiums

Coefficient	Estimate	**Coefficient**	Estimate
β_0	182558.291429080	β_{12}	-139.968460767
β_1	62325.419630012	β_{13}	-46.743289353
β_2	-60607.778269931	β_{14}	57.571408444
β_3	-102983.349906842	β_{15}	3.579054852
β_4	10197.022485267	β_{16}	1.718866242
β_5	-239.226323398	β_{17}	-0.252507443
β_6	4.024319740	β_{18}	-0.035931818
β_7	-0.035835831	β_{19}	-0.015900117
β_8	0.000124896	β_{20}	0.000483490
β_9	2702.494864553	β_{21}	0.000120934
β_{10}	-2368.493333138	β_{22}	0.000047704
β_{11}	-4972.362717316	β_{23}	-0.000002365

for the model fitted with the single family property type.

Now we stratified for condominiums and fit our model which resulted in the regression lines fit in Figure 3.5. In repetition, we see a slight increase over time, where the right-end tails have a sharper increase due to price hikes seen in those times. Just before these steep increases, the rate of change slowed, experienced during the economy shutdown from the COVID-19 pandemic. The green data points that are widely segregated from the rest are points observed from San Francisco county, which easily surpasses as the most expensive county in California. It is shocking that the inland urban areas are cheaper than rural areas. These lower

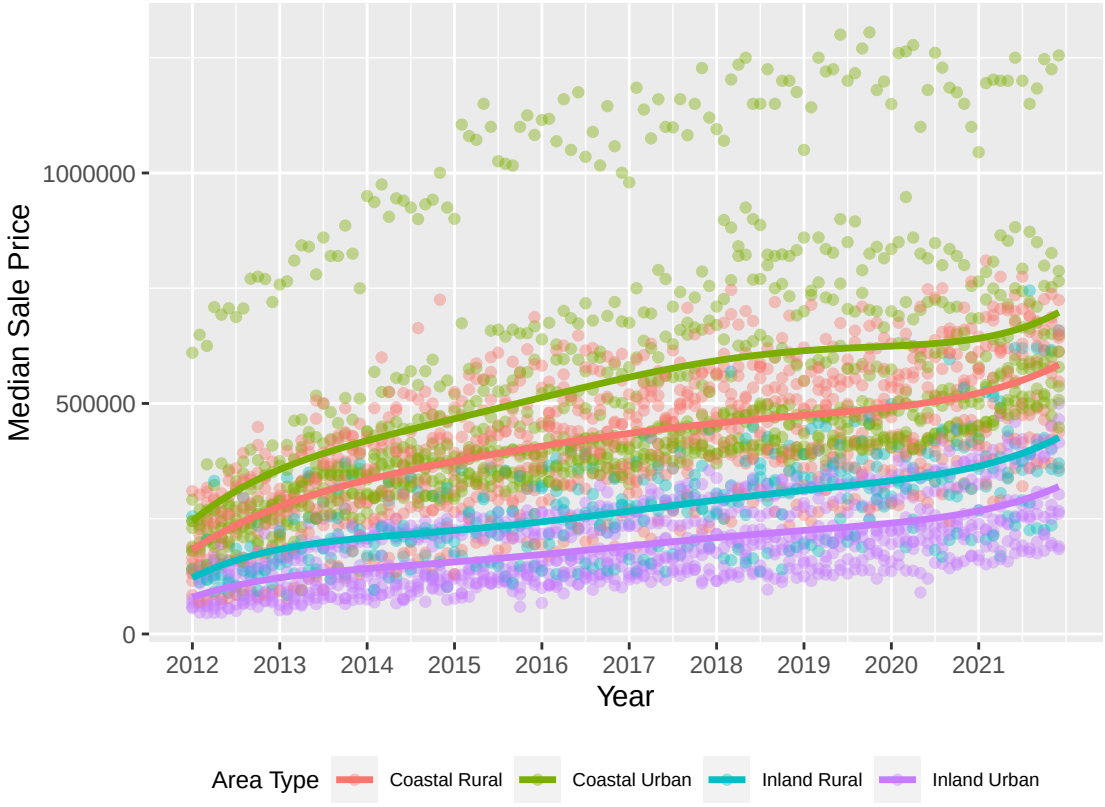


Figure 3.5: Regression lines for Quintic Polynomial of Time for Condos

Table 3.6: Table of Coefficient estimates for quintic polynomial model stratified by Townhomes

Coefficient	Estimate	**Coefficient**	Estimate
β_0	326901.8757765	β_{12}	-652.8588239
β_1	17527.8732105	β_{13}	939.5293327
β_2	-89688.2911643	β_{14}	-318.1115391
β_3	-212506.4190106	β_{15}	15.7717746
β_4	6083.8054366	β_{16}	-18.7324600
β_5	80.4277312	β_{17}	7.8760472
β_6	-3.3869689	β_{18}	-0.1537139
β_7	0.0327742	β_{19}	0.1661734
β_8	-0.0000979	β_{20}	-0.0735533
β_9	10428.4011588	β_{21}	0.0005207
β_{10}	-20406.8289379	β_{22}	-0.0005367
β_{11}	1622.6125608	β_{23}	0.0002400

prices could be attributed to other local effects not collected in this data such as crime rates, incomes and more.

Figure 3.6 is stratified by townhomes and colored by area type, mirroring the results in the past plots. For urban areas, we discern similar tendencies for townhomes with perceptible inflation towards the last two years of data. However, a diverging narrative materializes for rural townhomes, especially in inland rural areas where an identifiable decline surfaces in the last months of the data collection, which presents a different outcome from urban areas. On the contrary, coastal rural areas experienced a less pronounced increase.

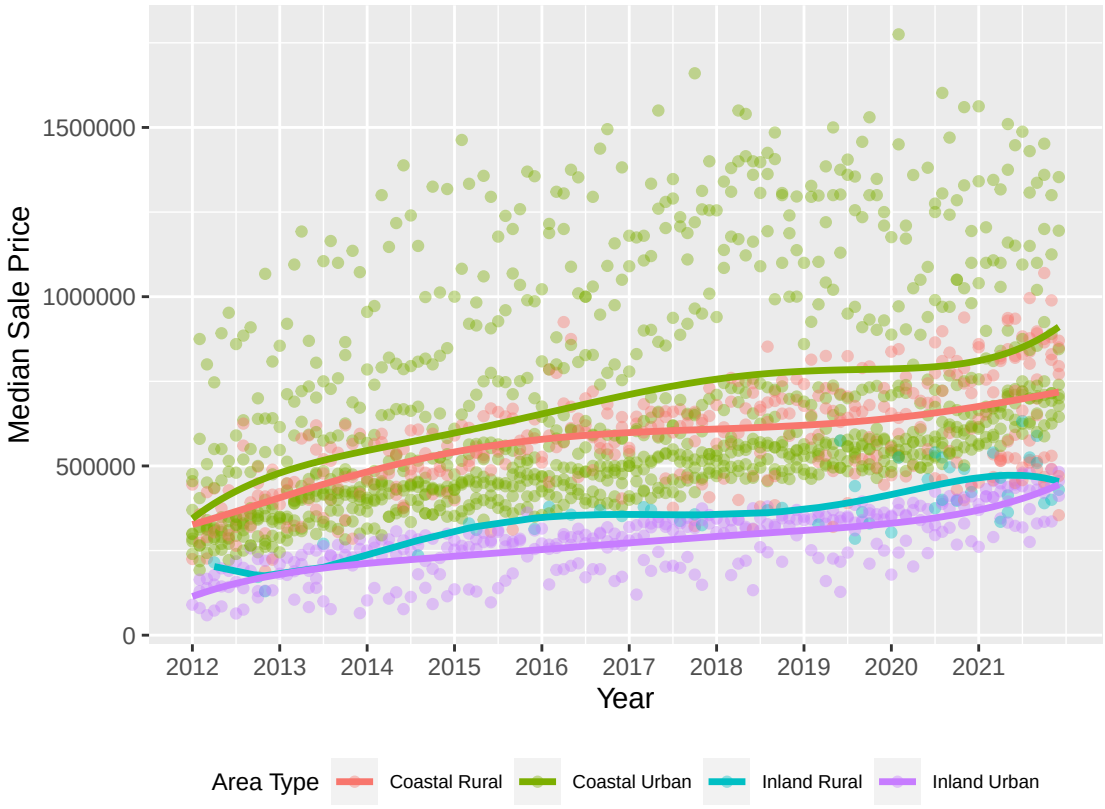


Figure 3.6: Regression lines for Quintic Polynomial of Time for Townhomes

Out of all three property types, single family residences tend to go through the most change during COVID-19 recovery. That is, in mid 2020 median sale prices experienced a steep increase. This trajectory was less pronounced for condominium and townhouses property types. Similar to the results for the cubic polynomial, only single family residences contained median sale prices that were not much different for inland rural and inland urban counties. However, for the condominium and townhouse properties, these values were more different, where inland rural had higher median sale prices. More specifically, the inland rural line for the townhouse properties experienced some oscillation which may be due to the lack of observations, therefore this prediction is unreliable. Overall, coastal counties were more expensive than inland counties for all property types.

3.1.3 Findings for Linear Regression

These nuanced findings paint a comprehensive picture of the paths unveiled by each property type, offering valuable insights into the heterogeneous dynamics impacting the real estate market in the varied regions within California.

On average, we observe a constant rise of median sale price over time; however the linear regression models may have difficulty in capturing the diverse growth behavior by county. Despite this limitation, the results denote an upward trend with a steeper increase in the last two years of data, which coincides with the shutdown and resurgence of the economy from the COVID-19 era. Single family residential data emerges as the most reliable for our use case due to its availability and provides the steepest increase in mid 2020. This was the only property type to have inland rural and inland urban seem to correspond to little difference in median sale prices for the decade. The gaps were more noticeable for the other property

types and overall coastal urban areas were always shown as containing the highest median sale prices, followed by coastal rural.

3.2 Generalized Additive Model

A Generalized Additive Model is a powerful model utilized for more flexible methods on fitting data and making predictions. This model allows to add smooth functions for nonlinear relationships between the dependent and independent variables. Generalized Additive Models (GAM) is a combination of generalized linear models and additive models and was originally developed by Trevor Hastie and Robert Tibshirani [7].

Instead of fitting simple linear relationships in linear regression, GAMs transforms the independent variables into smoothing functions (smoothing splines) to garner nonlinear trends and heterogeneous patterns in our data. The smoothing intensity of the splines depend on the smoothing parameter chosen, for our studies we kept it as default, meaning it will choose the parameter for our models. The term "additive" is derived from the configuration of the independent smoothing functions for each predictor.

The model presumes that the mean value of Y is an additive function of nonlinear functions of each predictor [7]. In very general form a GAM takes the appearance of:

$$Y = \beta_0 + S(X_1) + S(X_2) + ... + S(X_p) + \epsilon \tag{3.8}$$

where β_0 is the intercept and $S(X_i)$ represents the smoothing function for the i^{th} predictor variable, i=1,...,p.

3.2.1 Simple GAM with one smooth function

First, we explore the most simple case, involving one smoothing function for the time variable. Incorporating one smooth function between time and median sale price into a generalized additive model assists in the attainability of a nonlinear relationship. This is necessary since prices change on a seasonal basis and as seen earlier change when global issues arise, such as pandemics which inherently cause slow downs in the economy. Equation

$$log(Y) = \beta_0 + S(X_t) + \epsilon \tag{3.9}$$

3.9 shows that log(Y) is the logarithmic of the response for median sale price, β_0 is the intercept when $X_t = 0$, X_t being the time variable and lastly, ϵ is the error term. The objective here is to estimate the optimal shape with the data fed to the model. We used the logarithm to prevent as these models underestimated housing price values; thus, this prevents negative home price values which are not meaningful. For interpretation purposes we exponentiated these fitted values.

As mentioned in the introduction to this section, using smooth functions on predictors effectively use regression splines, where these splines are necessary when a relationship is not best explained when fitting a straight line. What a regression spline does is that it breaks the independent variable into chunks or segments, more formally known as "knots," and fits a unique polynomial function within each segment. These polynomial functions are of a lower degree, such as quadratic or cubic, allowing for gradual changes in the dependent variable as the independent variable changes [7]. For our purposes, we set these to defaults, meaning our model chooses the number of knots for the smoothing function. After the number of

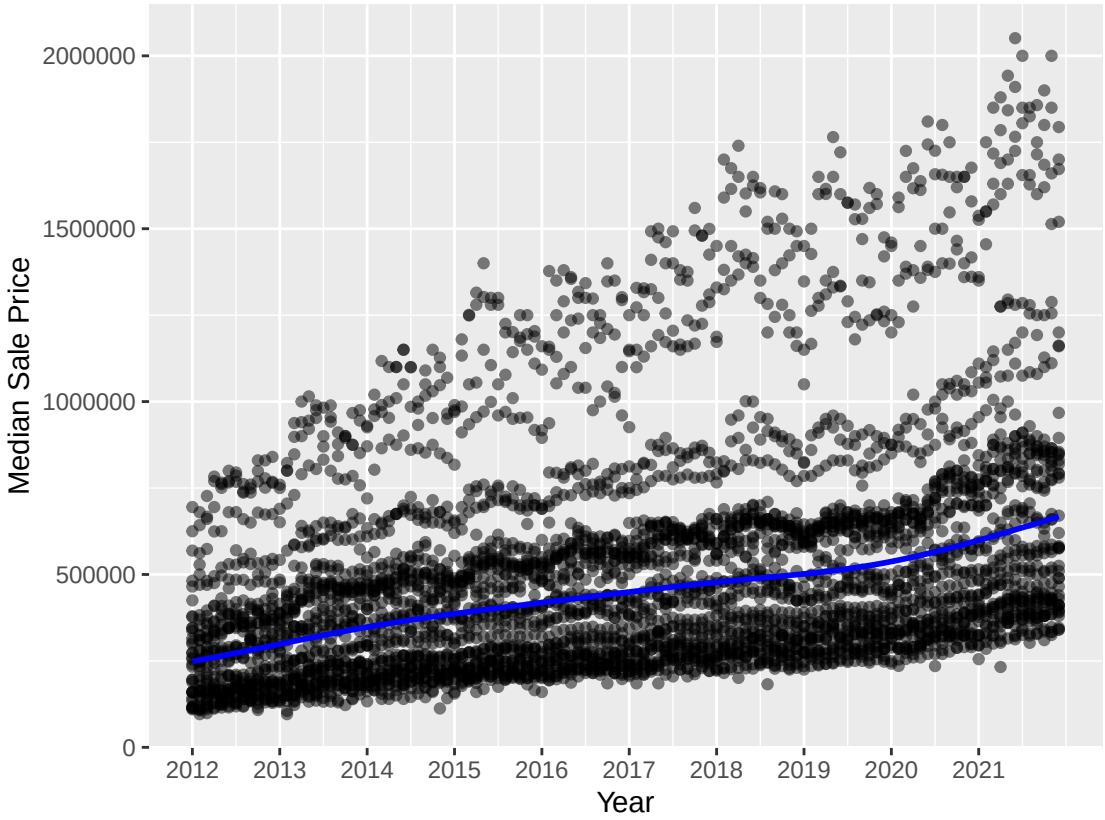


Figure 3.7: Regression line for GAM with fixed effect on time for Single Family Residential

knots is set (automatically in our case), basis functions are chosen per knot. Since this equation is non-parametric, this means we cannot extract coefficient estimates, except for the intercept of course. One may derive code to extract coefficient estimates from each knot, however this is not practical and may be hard to interpret, as interpretations from linear regression does not apply. Instead of using coefficient estimates, smooth functions in GAMs involve estimating parameters (these are not fixed parameters like those in linear regression), which does not necessarily mean the smooth function is parametric, it is still non-parametric since parameters are not fixed, stemming from basis functions that calculate the shape of the curve [7].

These smooth functions are assembled representing polynomial functions in each

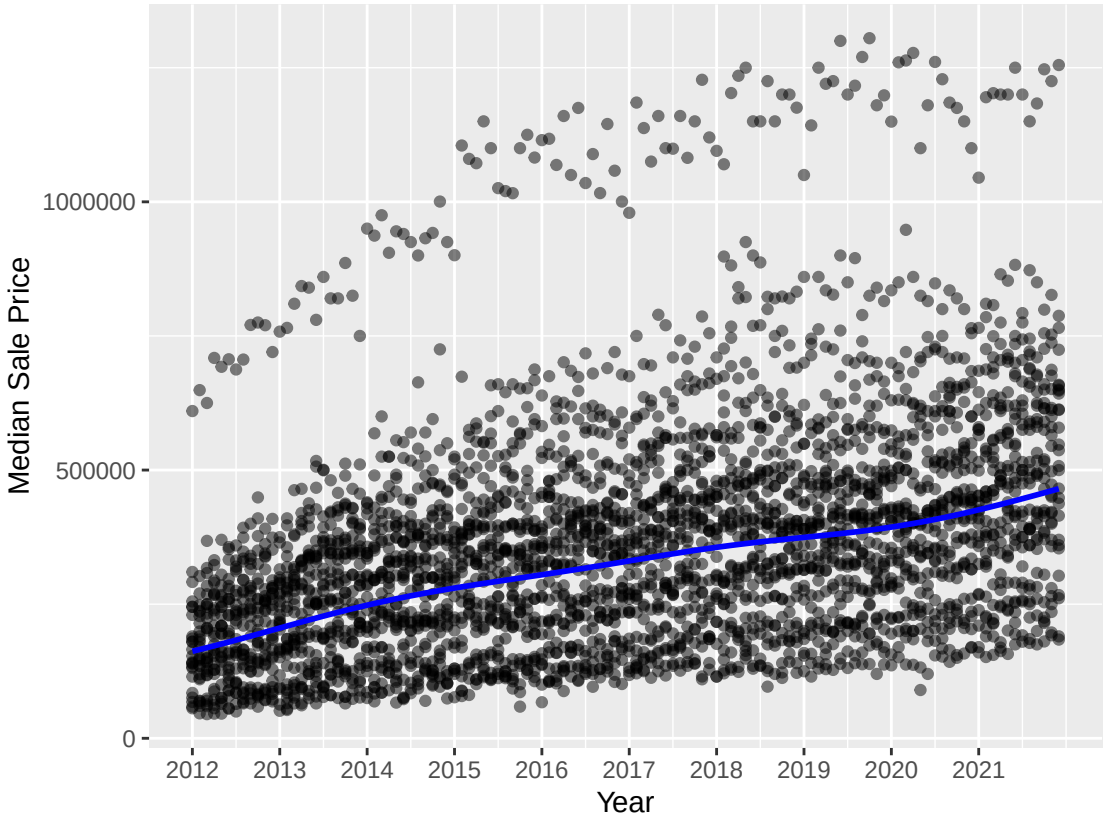


Figure 3.8: Regression line for GAM with fixed effect on time for Condos

partition. The coefficients for each basis function use ordinary least squares or maximum likelihood estimation to be estimated. These coefficient estimates are what give the flexibility and steepness of the curve for each partition.

Figure 3.7 demonstrates the fitted regression line using the smooth function of time to determine the relationship with median sale prices for Single Family properties. The blue regression line showcases a smooth curve, expressing a resemblance to the plots in prior sections. In particular, there is a propelled increase rate in the last two years of data, a consistent conformity of prominent growth.

Note that the models for this property type will be more robust due to the density of our data. Figure 3.8 shows the model fit for the condominium property

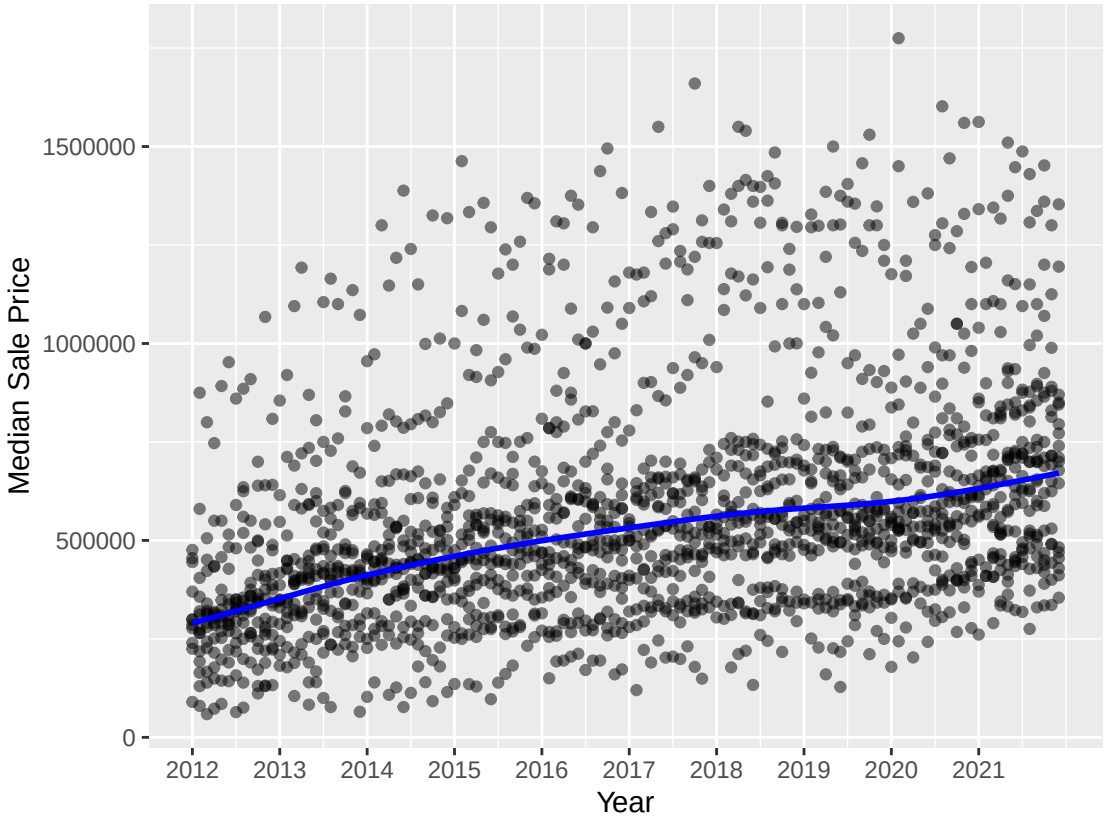


Figure 3.9: Regression line for GAM with fixed effect on time for Townhomes

type. A comparison with single family residential shows that these curves experienced a flatter increase. This distinction can be attributed to additional homes sold on average for single family residences compared to condominiums. This disparity stems from the nature of more single family homes being available to the public, directly impacting sale volumes.

Lastly, Figure 3.9 shows the same plot for townhomes which is the data with the least availability, deeming these models as less reliable. As mentioned in the pre-processing section, we had removed homes that sold less than 5 in any given month to remove outliers to remove any influence they may have in the model estimates. The increase for median sale price was gradual over time. This model

seems to underfit our data, therefore we wish to include more variables to have more variance explained.

3.2.2 Random Intercept GAM

We proceed now with random intercepts. A random intercept is utilized in the context of mixed-effects models, which refers to individual specific or group specific intercept that changes randomly depending on different levels of a categorical variable. These type of random effects are best used with longitudinal studies. Indeed we wish to model counties as we may expect heterogeneity in the responses among counties due characteristics of homes such as number of rooms, county, population and more. Assessing within-individual changes may give insight on response trajectories, focusing on temporal order of events.

Therefore, our objective in this section is to model a random intercept which will effectively model a different mean sale price for each county at time zero. This leads to counties being parallel to one another when regression lines are plotted, as visualized in the upcoming plots. The upside of utilizing these models is the ability to have subjects with varying number of observations taken at different points in time. Hence, this is very useful since the counties in our data have different number of observations due to many counties not having a complete set of observations or having months of data not present. Once we incorporate time with a smooth function and the counties variable with random effects, this yields the model:

$$log(Y_{ij}) = \beta_0 + b_i + S(T_{ij}) + \epsilon_{ij} \tag{3.10}$$

where β_0 is the intercept, $S(T_{ij})$ is smooth function of time and ϵ_{ij} is the error

for the i^{th} county, at the j^{th} occasion, where $i = 1, ..., N$, N being the number of counties we have in our data and $j = 1, ..., n_i$, n_i being the number of occasions for the i^{th} county. The second term is applying the random intercept, being county specific, where b_i is the i^{th} county. Where intercept for the i^{th} county is is $(\beta_0 + b_i)$. Lastly, $log(Y_{ij})$ is the logarithmic predicted median sale price for the i^{th} county, at the j^{th} occasion. As seen before, we predicted logarithmic values of median sale price Y_{ij} to prevent any negative values being predicted and for interpretability purposes we exponentiated these values after. We fitted 3 models to inspect each property type at a more granular level. That is, we will have a model for each the three property types; thus n will vary as each property type will contain a different number of counties represented.

Now onto the visualizations, Figure 3.10 plots the regression lines with random intercepts for single family residences. It is apparent that all these lines are parallel with one another except for one line in the bottom left panel. This is caused by no data being available within that time frame so a straight line is drawn between the two time points. Therefore, the model fitted regression lines for data that was available in our dataset, which is the reason for some lines being drawn late or being cut off early. This plot is faceted by area type and demonstrates gradual growth over time, with higher inflation seen in the right tails of the lines. There is a detectable drop in median sale prices in the end of 2019 for all area types. The differences in counties here are the median sale prices seen at time zero (intercept), accounting for price differences in predominantly wealthy and predominantly poor counties.

The condominium property types shows similarities to results in Figure 3.10, gravitating to a steady increase over time until a drop occurred in the end of 2019

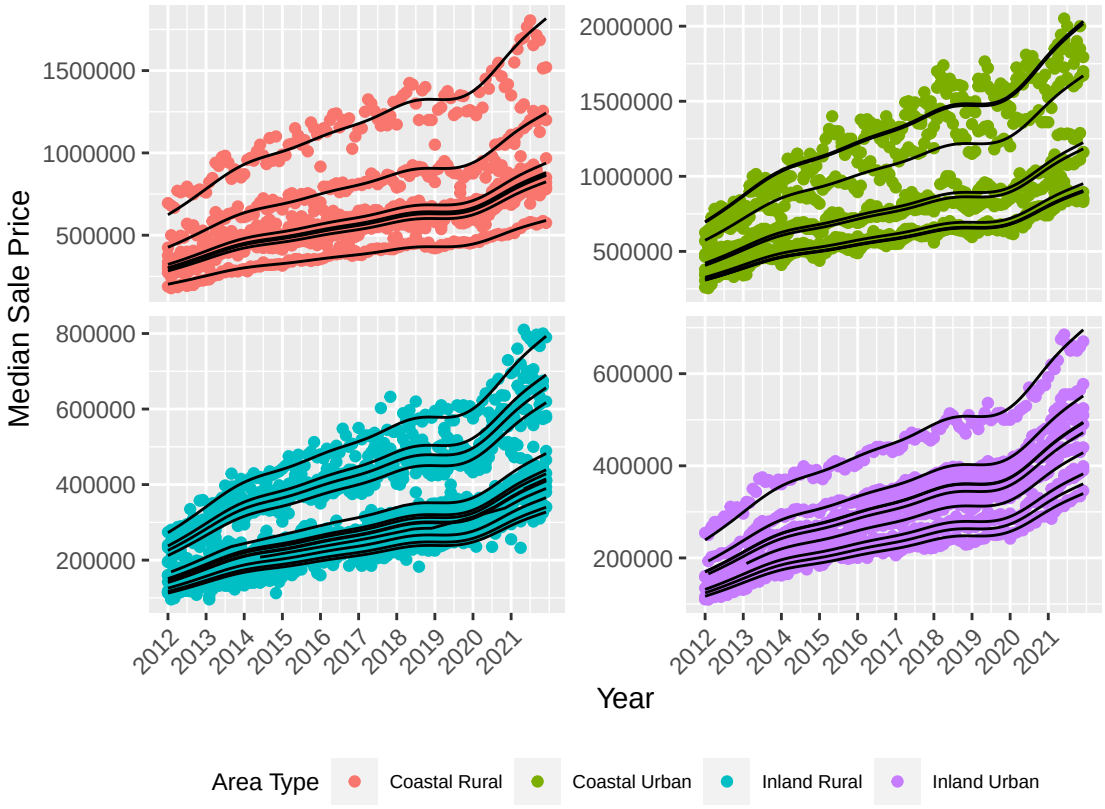


Figure 3.10: Regression lines for GAM with Random Intercept for Single Family Residential

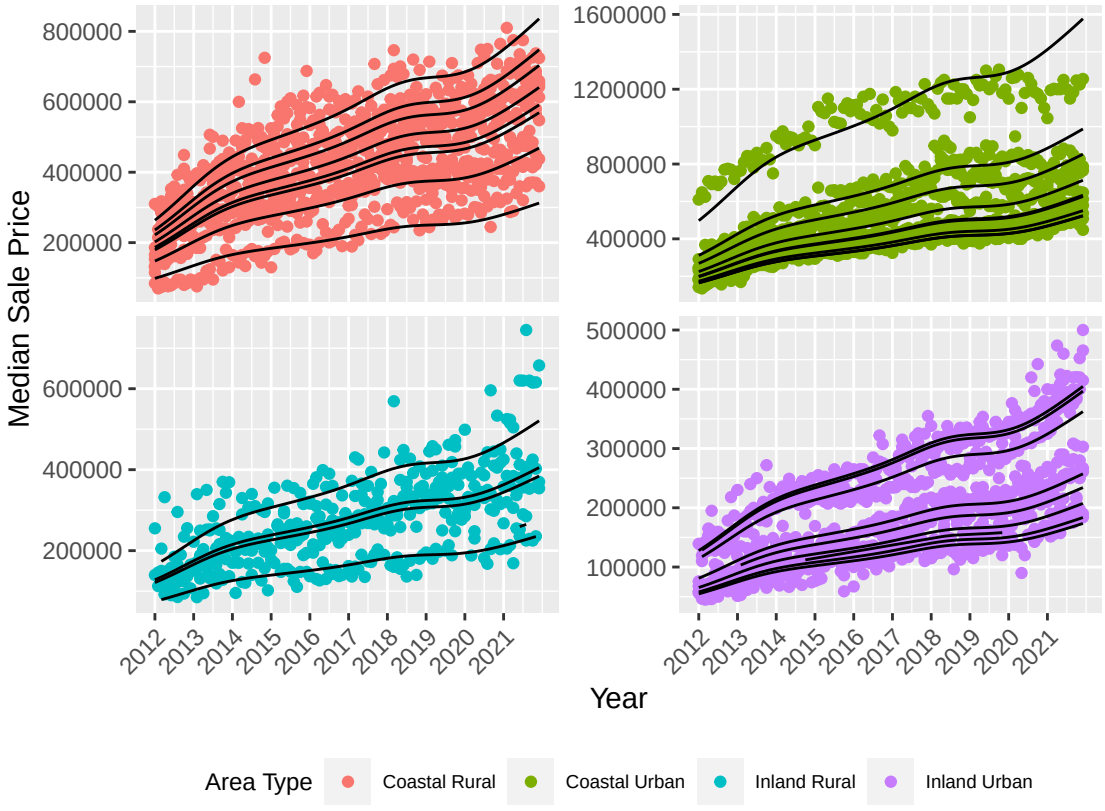


Figure 3.11: Regression lines for GAM with Random Intercept for Condos

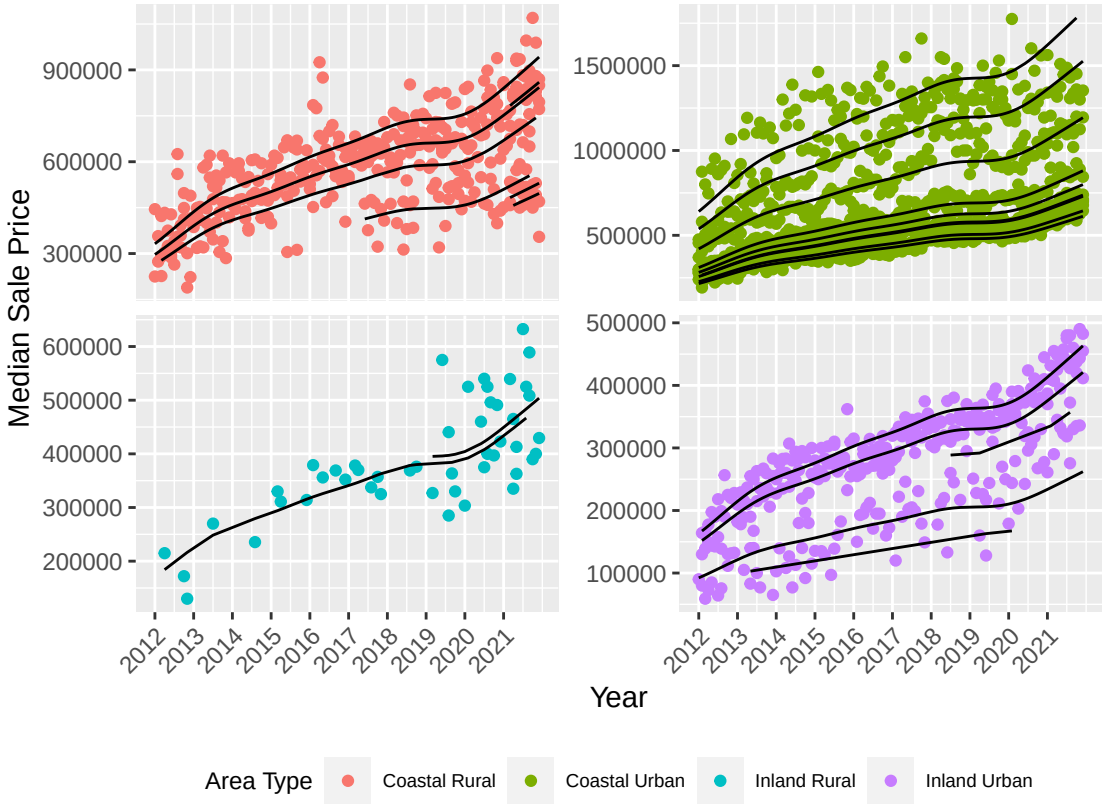


Figure 3.12: Regression lines for GAM with Random Intercept for Townhomes

(start of COVID-19). Figure 3.11 saw another steep increase in the beginning of 2020 due to the economy starting its course of recovery.

Lastly, as shown in Figure 3.12, townhomes experienced a similar trend to the previous plots. Nevertheless, the Townhouse property type contains the least number of observations out of all property types. This scarcity of observations leads to several incomplete lines, where the blame lies on the unbalanced data. As seen in our data, townhouse properties sell less on average than single family residences. Since many lines were incomplete, we could not fully encompass the seasonal changes in those counties and changes seen pre and post pandemic. Even though some lines were limited in the number of observations, we were able to

distinguish upward trends seeing the same drop in previous plots.

3.2.3 Random Intercept and Random Slope GAM

An extension of the previous section, now we will explore having both a random intercept and random slope in our models. Adding a random slope will now allow to capture changes over time for different counties, diversifying the rate of increase or decrease per county. In other words, regression coefficients vary randomly among counties, allowing for extra flexibility. The following is representation of the formulation containing both random intercept and random slope:

$$log(Y_{ij}) = \beta_0 + b_{1i} + b_{2i} \cdot T_{ij} + S(T_{ij}) + \epsilon_{ij} \tag{3.11}$$

allowing to capture different rates of change and intercepts of median sale prices for each county. The new term here b_{1i} accounts for the random intercept change per county, that is our intercept for the i^{th} county is $(\beta_0 + b_{1i})$. The term $S(T_{ij})$ is a fixed slope where i and j are identical to the previous section. The new term, $b_{2i} \cdot T_{ij}$ is a random slope, meaning the slope changes based on the county.

Figure 3.13 plots our objective of including both random intercept and random slopes for single family residences. Even though we opted for counties to have distinct slopes, they look similar across the board and look alike to plots that were random intercepts only. The random slope is visible for some counties, but do not vary too much from the rest of the counties. For these figures we only predicted available observations on data that was available. Thus, for some counties we see only partial lines drawn.

Figure 3.14 gives the same plot for the condominium property type, showing

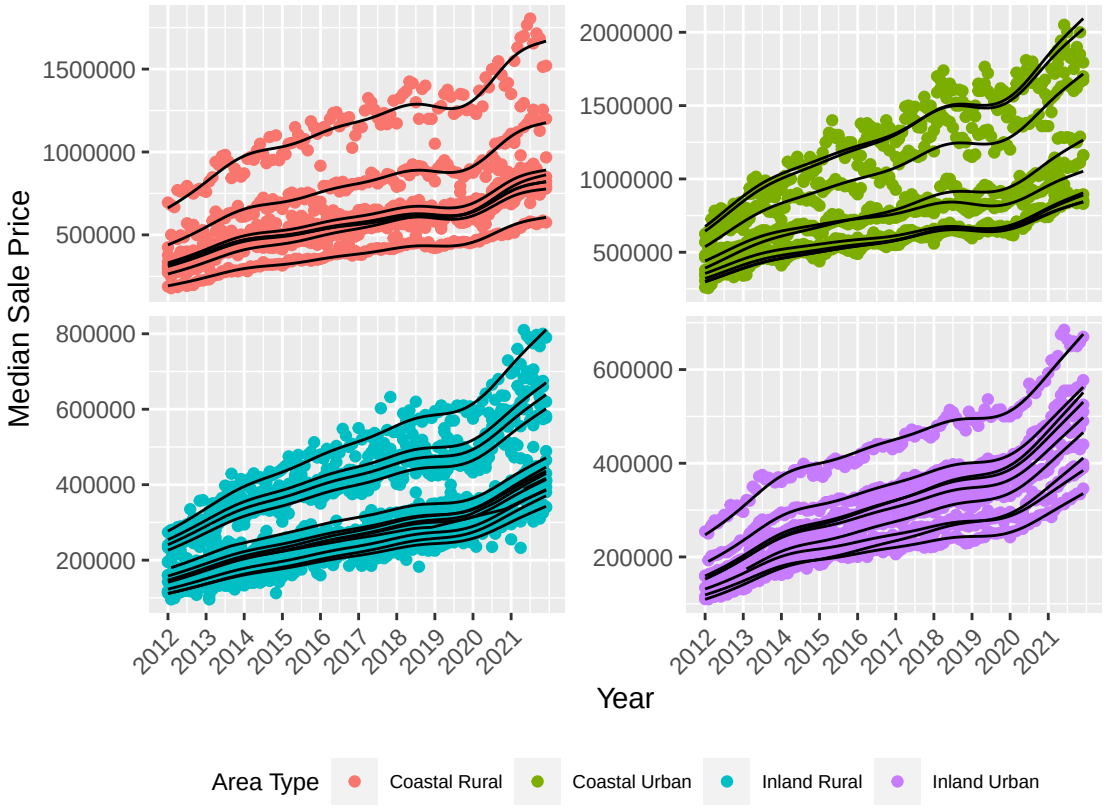


Figure 3.13: Regression line for GAM with Random Intercept and Slope for Single Family Residential

similar patterns with some counties showing less growth than others on average. Alpine county, an inland rural county, only has predicted values for time points, shown as the smallest line drawn in the inland rural subplot.

The last plot in this subsection is Figure 3.15, plotting fitted regression lines for the townhouse property type. There are similar patterns to before except there are more lines that do not span the whole width of the independent variable, time and cannot capture proper trends seen for a full year. For instance, in the coastal rural subplot there are two insignificant lines seen at the end having downward trajectories, meaning those counties had a decrease in price over that period. However, these results may be misleading for these specific counties since we do not have enough information to capture yearly pattern, or an overall pattern for a decade.

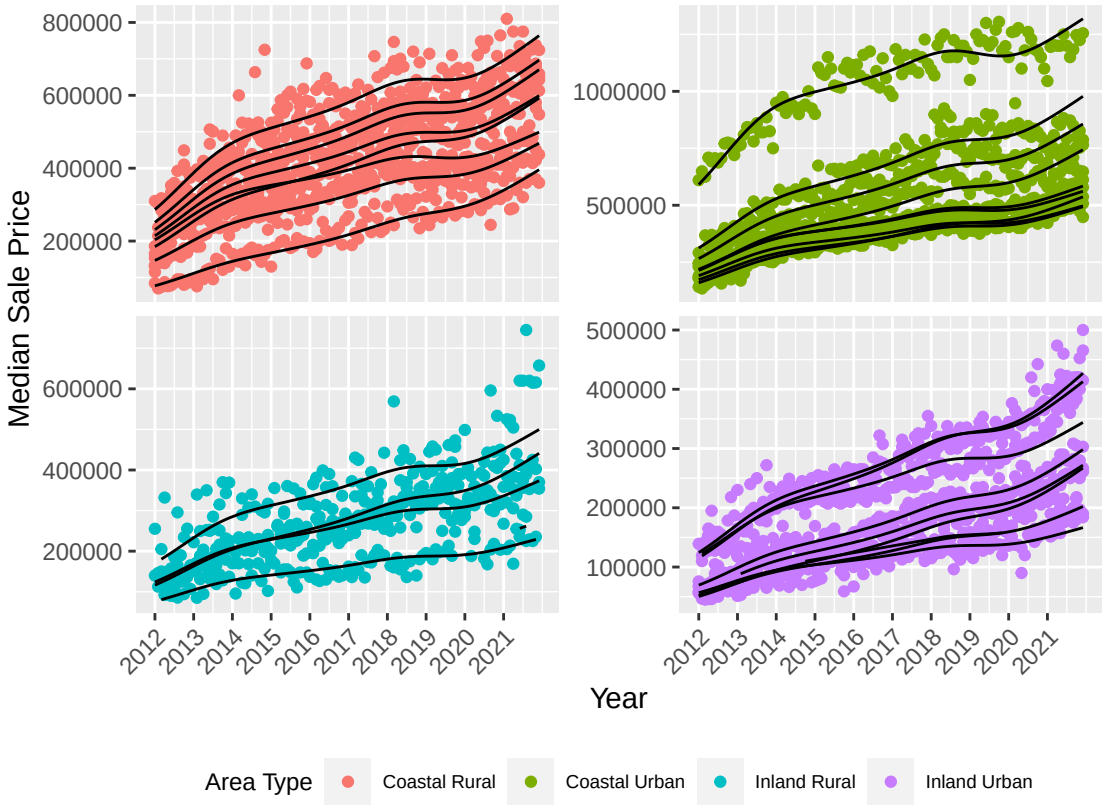


Figure 3.14: Regression line for GAM with Random Intercept and Slope for Condos

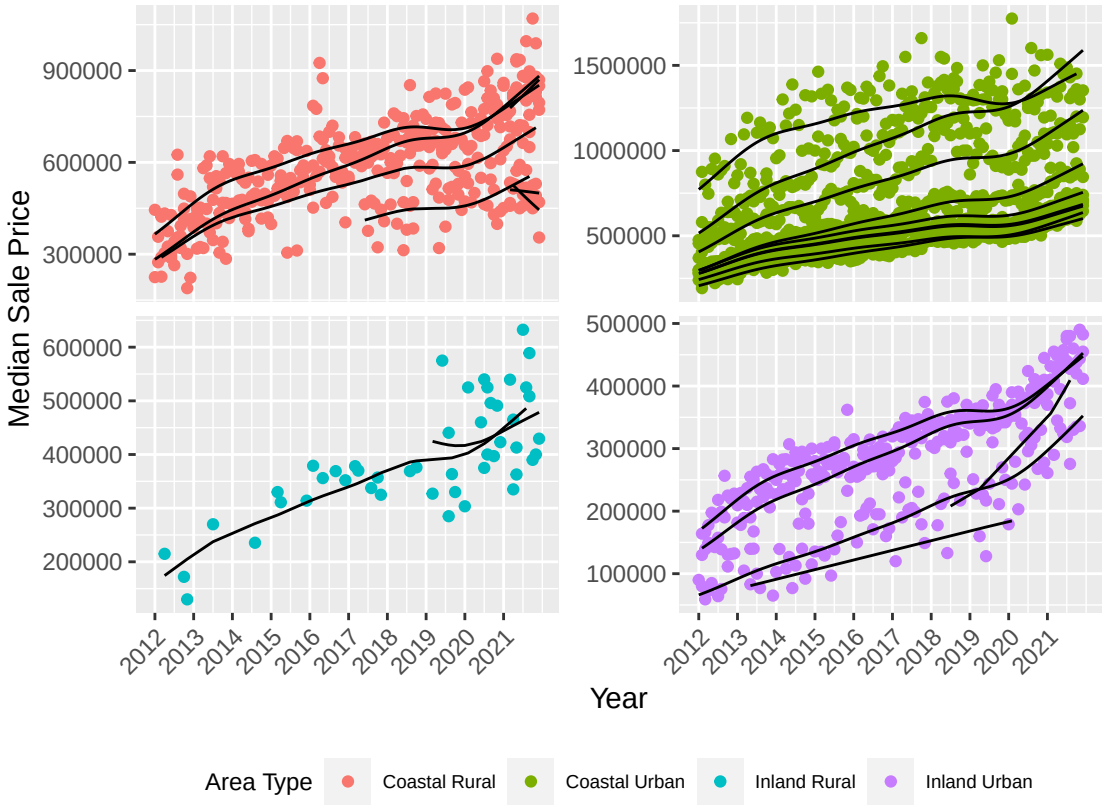


Figure 3.15: Regression line for GAM with Random Intercept and Slope for Townhomes

3.2.4 GAMs including population

Now we will turn our focus to GAM models including populations. As mentioned in chapter 2, we included two variables regarding the area. The area type variable took into account population size, with categories coastal rural, coastal urban, inland rural and inland urban.

One variable was inclusive on population size: coastal rural, coastal urban, inland rural and inland urban. These were decided based on the location of each county and the population size as well as metropolitan areas. The second variable regarding area was a binary variable, only including coastal and inland as categories. Thus, this second variable did not incorporate information on population size. The last variable needed for these models is the population sizes themselves. The models have a similar fitting to those studied above. The first model we will consider is:

$$log(Y) = \beta_0 + S(X_t, K = 7) + X_a + S(log(X_p), K = 7) + \epsilon \tag{3.12}$$

where each smoothing term takes the form of

$$S(X_t, k = 7) = \sum_{k=1}^{7} \beta_k f_k(X_t) \tag{3.13}$$

and similarly

$$S(X_p, k = 7) = \sum_{k=1}^{7} \beta_k f_k(X_p) \tag{3.14}$$

The new variables here are the area type variable including the aforementioned two categories X_a and the population variable, X_p with a smoothing function for nonlinearity. X_t represents the time value. We took the logarithm of the popu-

lation variable to have values normally distributed. The parameter k within the smoothing functions denotes the number of knots or number of basis functions, f_k, for the smooth terms and controls the flexibility of the model, where a lower k chosen leads to a less flexible model and a higher k value leads to a more flexible model. Furthermore, the placement of these knots refers to the specific points along the interval of the independent variable where the smoothing function can differ in curvature or slope [7]. In addition, by definition since we placed 7 knots in each of the smoothing terms, they partition the range of each variable into $k + 1 = 8$ subintervals. We chose $k = 7$ for both variables to prevent overfitting but it is a big enough value to capture nonlinearity. Going further in detail, the number of basis functions stands for the number for components utilized to construct the smooth curve and builds these curves from a spline which are continuous piecewise polynomials (function made up of simpler functions). Then, the relationship between median sale prices and each independent variables is modeled as a weighted sum of these basis functions, where the weight is denoted by β_k [7]. Therefore, our objective is to determine a balance for model complexity and goodness-of-fit. This value was kept fixed for all subsequent models fitted.

We fitted three models, one per property type, with the same variables and formulation shown in Equation 3.12. The reason to fit these models separately by property type is to capture different trends for each property type, as each property may have median sale prices differ in trends over time. We casted predictions on two new data frames; one for coastal and one for inland, which were created from each property type, rendering a total of six new data frames. These data frames were created to understand the relationship with median sale price and population at a more granular level for each property type and specified area. As in either

Table 3.7: Table of Minimum and Maximum of Population Sequences

Area Type	Property type	Min Population	Max Population
Coastal	Single Family Residential	130,000	10,000,000
Coastal	Condominium	130,000	10,000,000
Coastal	Townhouse	130,000	10,000,000
Inland	Single Family Residential	15,000	2,500,000
Inland	Condominium	100,000	2,500,000
Inland	Townhouse	190,000	2,500,000

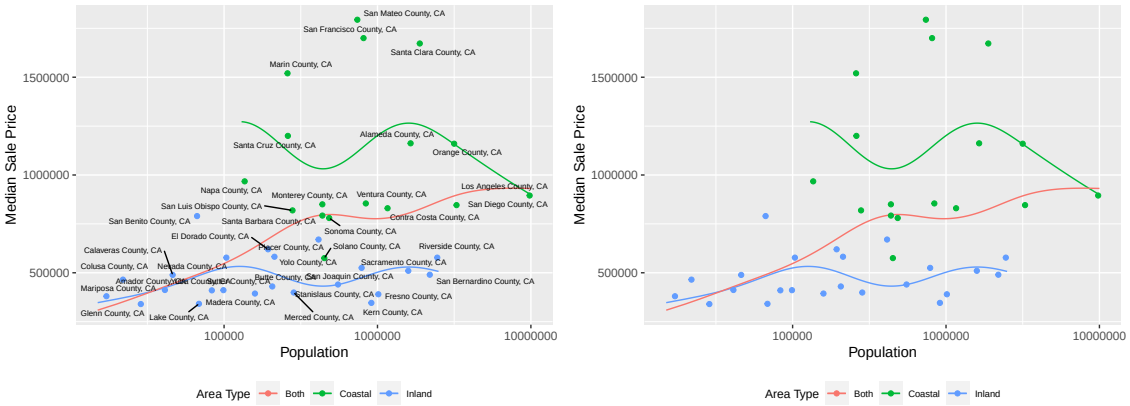
coastal or inland may see more growth than the other depending on population sizes and the property type. These two new data frames included three columns each: population sequence, time (fixed to one value) and area type (coastal and inland, only one category will be present for each data frame). The population sequences were unique and were in increments of $5,000$ for each respective property type and area type as shown in Table 3.7, so each of six data frames had a unique sequence. Furthermore, we wish to see patterns for the relationship between median sale price and population for the last month of recording. Therefore, the time column will only contain the fixed value, 119 (corresponding to December 2021) for all population entries and since each population type is tied to an area, each data frame will only contain one category or level. Predictions for median sale price values were made for each respective data set to account for the effect population and each sole area type had on median sale prices (blue and green lines for Figure 3.16, 3.18 and 3.20). These predictions were added as new columns for each respective data frame. Then, we row bind these two data sets in order to be able to take advantage of the ggplot package and create comprehensive visualizations.

Now we wish to see the effect of population itself on median sale prices. Thus, in this next formulation we excluded area type:

$$log(Y) = \beta_0 + S(X_t, k = 7) + S(log(X_p), k = 7) + \epsilon \tag{3.15}$$

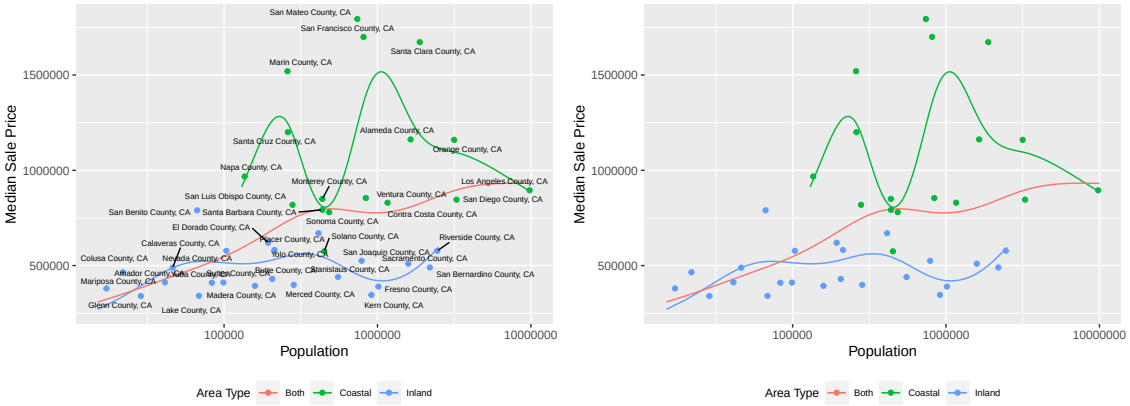
where the formulations for the smoothing terms are the same as shown in Equations 3.13 and 3.14. This model was fitted nine times: one model was fitted for the original data set of each property type (3 models total, one for single family, one for condos and one for townhouse data) and 2 separate models were fitted for the coastal and inland subsets of each property type (single family inland and coastal, condos inland and coastal and townhouse inland and coastal, total of 6 models). For each model fitted with the original datasets, we casted predictions (red lines for all figures) on the combined data set mentioned above and for the models segregated with property and area types, we casted predictions on the data frames created (the unmerged version of the combined data set) since the models were created from each area type. Thus, we wanted to make predictions (green and blue lines on Figures 3.17, 3.19 and 3.21) on the inland data frame from inland subsetted models and predictions on the coastal data frame from the coastal subsetted models. We created these separate models and predictions to determine if we could achieve a more accurate prediction.

Figures 3.16 and 3.17 displayed the lines of best fit for median sale price vs population for single family homes. We have also added the observation points for the last month of recording of our data and labeled the county for each respective point. We have made plots without county labels as a tool of visualization aid. Figures 3.16 and 3.17 show a trend of larger population sizes for coastal counties



(a) Price vs Pop Reg Lines - sfh (1) (b) Price vs Pop Reg Lines - sfh No Labs (1)

Figure 3.16: Price vs Population Regression Lines - Single Family Homes (1)

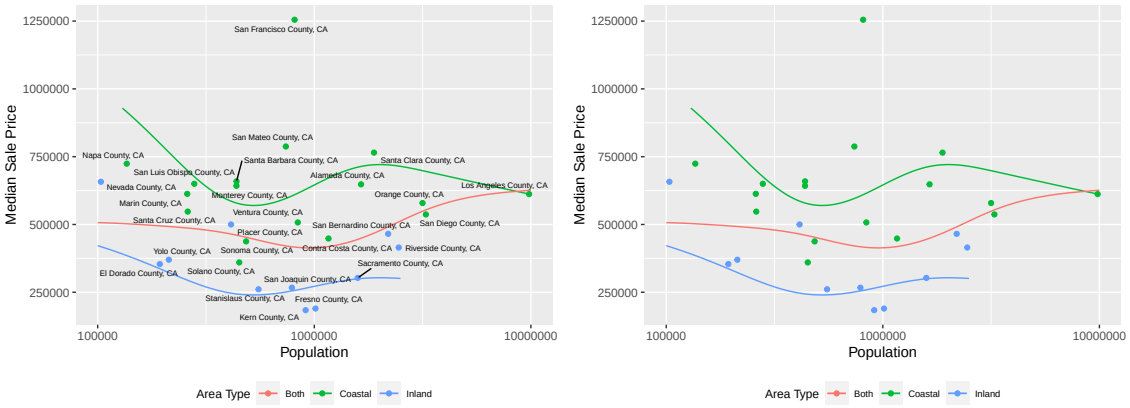


(a) Price vs Pop Reg Lines - sfh (2) (b) Price vs Pop Reg Lines - sfh No Labs (2)

Figure 3.17: Price vs Population Regression Lines - Single Family Homes (2)

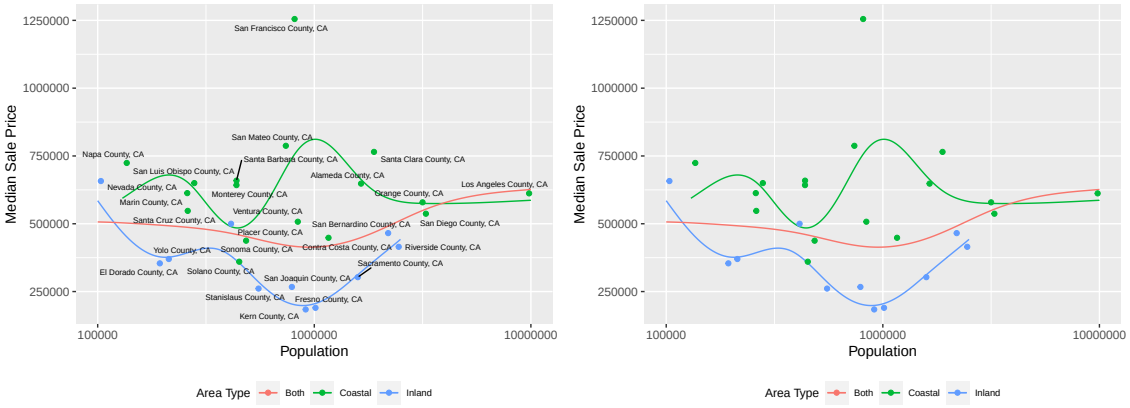
on average, compared to inland counties. Coastal counties also demonstrate higher median sale prices and both these inferences align with the points, where blue points tend to be closer to the blue curve and green points with the green curve. For Figure 3.16, it is interesting to note that the curve for inland counties has slight variation regardless of population size, meaning we expect prices to hover over $\$500,000$ for single family homes, for populations greater than 100,000. Inland counties with less than $100,000$ residents will find homes less than $\$500,000$ on average. For the green curve, we also noticed oscillation, however as population increases to 10,000,000 residents, prices decreased. This could be from external factors as mentioned earlier in this study, such as NOX levels, crime rates and more. Buyers seeking homes in coastal counties can expect to pay above $\$1,000,000$ on average regardless on population size for a single family home. The red fitted curve is drawn from the model excluding area type, but including the full data set for single family homes, and we can detect that this curve sits in the middle between coastal and inland median sale prices. However, an interesting trend seen in this red curve is that median sale prices gradually increase as population increases. Figure 3.17 contains the same red curve, thus we will refrain from any further analysis. The coastal and inland curves for Figure 3.17 demonstrate similar trends to Figure 3.16, however we have a higher magnitude of oscillation for the coastal counties.

Figures 3.18 and 3.19 exhibit the lines of best fit for median sale price vs population for condominiums. It is apparent that the observation for San Francisco county is an outlier for December of 2021. For coastal lines in Figure 3.18 and 3.19, it is inconclusive whether we can decipher an increasing or decreasing trend on median sale prices for increasing population size. It is also difficult to draw any conclusions for inland counties. However, we may infer that the housing does vary when across



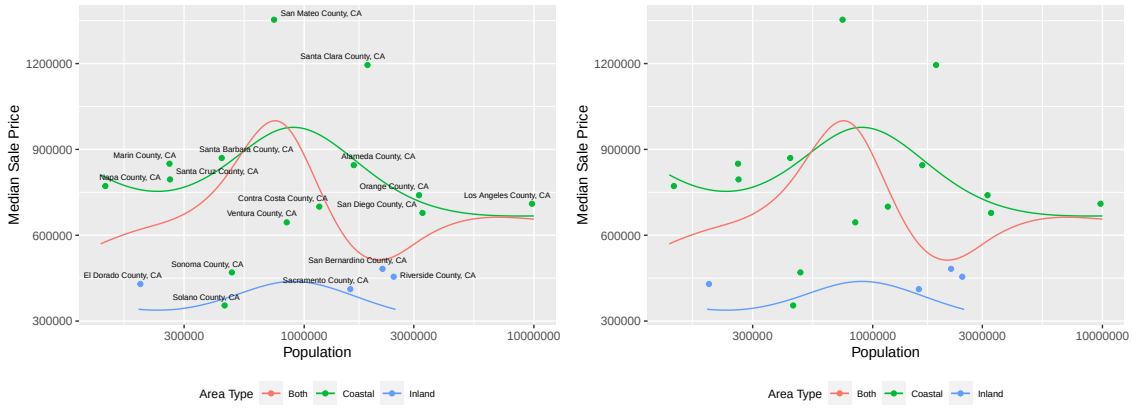
(a) Price vs Pop Reg Lines - con (1) (b) Price vs Pop Reg Lines - con No Labs (1)

Figure 3.18: Price vs Population Regression Lines - Condominiums (1)



(a) Price vs Pop Reg Lines - con (2) (b) Price vs Pop Reg Lines - con No Labs (2)

Figure 3.19: Price vs Population Regression Lines - Condominiums (2)

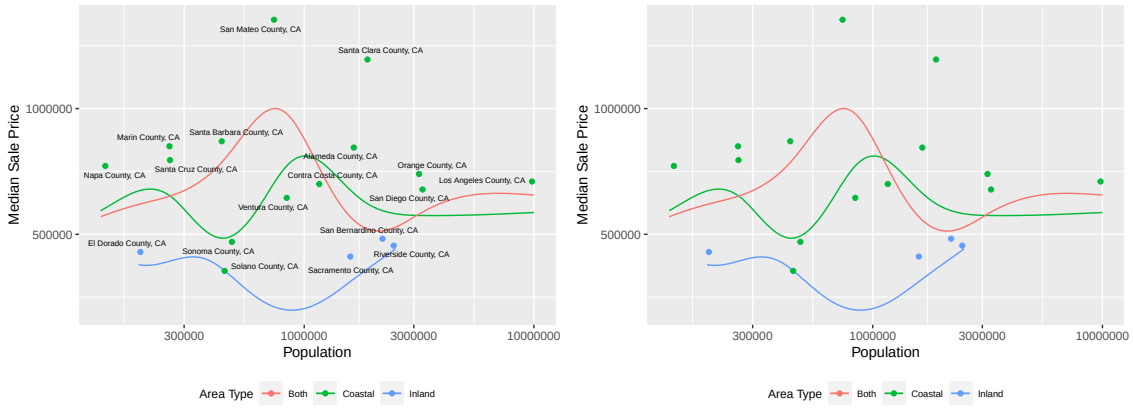


(a) Price vs Pop Reg Lines - Th (1) (b) Price vs Pop Reg Lines - Th No Labs (1)

Figure 3.20: Price vs Population Regression Lines - Townhomes (1)

different regions across the state. Coastal counties typically houses more expensive condominiums with an average price hovering over $\$600,000$. Inland counties have relatively cheaper condominiums, with an average around $\$300,000$. Accounting for both areas (state as a whole), prices sit steady with an average approximately $\$500,000$.

Figures 3.20 and 3.21 demonstrates the same figure for townhomes. The results obtained from these figures are largely inconclusive and implies population does not have a strong effect on median sale prices for townhomes. However, we can deduce that coastal counties tend to have more expensive townhomes than inland counterparts for population held fixed.



(a) Price vs Pop Reg Lines - Th (2) (b) Price vs Pop Reg Lines - Th No Labs (2)

Figure 3.21: Price vs Population Regression Lines - Townhomes (2)

3.2.5 Findings for GAMs/Comparison to LMER

In this section we found similarities across all types of models fitted. There was an upward trend throughout the decade, with one drop seen in the last two years of the data, during the time of the pandemic shutdown, causing a slowdown in the economy and a period of less homes sold. The single family residences contained the most number of observations for modeling and visualization purposes, while townhouses fared worse in this regard, causing anomalies and incompleteness in regression lines for some counties. Due to the sheer size of the number of single family residences compared to townhomes, single family residences are available for the masses. In addition, townhomes sell less properties a month, and in some cases our data showed many observations sold less than 5 homes in many months, causing the median sale price to be influenced by only a few homes for that month. In particular, some of these observations stem from luxurious and expensive homes, causing these to be outliers. This creates an issue for model accuracy for this property type.

Regression lines fitted with population did not capture any significant trends

for condominiums and townhomes and therefore these lines may be implying that population does not play a huge role in median sale prices for these property types. When not controlling for inland and coastal (including both) the regression line for single family residences show a strong trend where median sale prices increased as population sizes increased. So, there is correlation between single family residential prices and population. When looking at the four areas in California, we have seen continued evidence that coastal areas exhibit higher prices and boast larger populations. Nonetheless, when looking within a geographical region (i.e., coastal) there is no association between population and prices.

Chapter 4

Conclusion

In the following sections we will provide a summary and discussion of our results. We will draw conclusions from these results by observing overall patterns and trends displayed in the results. We will also mention any limitations that lead to less than favorable results. Last of all, we will give a brief discussion on future pathways we can explore to improve this study and enhance our understanding of the housing market.

4.1 Results

Observing regression lines fit we found the most reliable data was single family residential observations for our modeling purposes; they boasted the highest base intercept out of all property types, while condominiums being the lowest. Despite these differences intercepts, all our of inference plots suggest that median sale price is increasing over time on average. However, since the quintic polynomial models provided little improvement over the cubic polynomial models, they were overfitting our data. Models that included population as a variable only showed a significant

increasing trend for single family residences when we don't account for geography. That is, coastal county prices will hold similar prices and inland counties will hold similar price with small variation, for time held fixed. Overall, coastal counties were determined to be more expensive than their inland counterparts for time held fixed, where coastal urban counties were consistently higher than coastal rural.

We visualized disparities in the townhome GAM models, but this was due to the lack of information being assimilated into models. Some of these disparities showed some counties decreasing, and the length of some of these lines of fit did not carry enough observations to provide in depth analysis on seasonal and yearly patterns. These challenges came to fruition from not enough homes sold for particular months in the data. This is also likely from removal of observations; however including these removed observations introduces outliers which may impede on model performance as they have disproportionate impact leading to biased estimates and overfitting. Our goal in this study was to be as accurate as possible, and the first vital step we took before modeling the data is by cleaning our data to prevent any distortions in our data visualizations as it brings challenges to proper interpretations.

It was also interesting to see that population was not an important factor for changes in median sale prices for the condiminium and townhouse property types whether looking at inland, coastal, or both. On the other hand, the trend for including both in the mix for single family residences displayed a median sale price increase as population sizes increased.

From the trends seen, we expect all counties (except for those with lack of observations) to have an increasing trend over time in the following months to come. However, the market can be very sensitive to adjustments, such as inflation, crime rates and interest rates. Which may turn away stakeholders looking to buy or sell

properties. The ever-changing dynamic of the housing market may be contributed from policies such as affordable housing programs, rent control, first time buyer assistance, zoning and land use regulations, property taxes, homeless prevention programs, new developments ,etc., which are all factors not included in our data set.

4.2 Limitations and Future explorations

As mentioned in the conclusion, there were a lot of factors not included in our data set such as policies, crime rates, interest rates, all of which would be beneficial to include for future model fitting. If we included such variety in our data, we would include more visualizations to seek relationships between these new predictors and median sale price. In addition, including more information may provide profound insights on factors that affect median sale prices as well as improve model accuracy.

Another avenue we could have journeyed in was provide spatial autocorrelation plots also known as Moran scatter plots, which are powerful tools used to visualize the tendency of neighboring locations or observations displaying similar values of a variable. Such plots are created from fitting spatial models. This type of plot contains actual values on the x-axis, the average of neighboring locations on the y-axis, each location being plotted as a point [6]. Then, a line of best fit is drawn; if the scatter points hug the line, it suggests positive spatial autocorrelation, which denotes that similar observations tend to be closer in space. If the points have more spread along the line of fit, it implies dissimilarity [3].

One more further exploration we could have included was to provide a map of residual plots. Residuals are the differences between the observed value and fitted

value, which represent the unexplained variation in the data that the model could not apprehend. That is we would perform a similar task with the Geographical plots seen in this study, however instead of analyzing median sale prices directly, we would plot residuals where we would apply a color spectrum indicating residual values that are negative, positive, or if they sit close to zero.

Bibliography

[1] United States Census Bureau. "gct-ph1 – population, housing units, area, and density: 2010 – county – census tract", 2020. 2010 United States Census Summary File 1.

[2] Philipp Carlsson-Szlezak, Martin Reeves, and Paul Swartz. What coronavirus could mean for the global economy. *Harvard business review*, 3(10):1–10, 2020.

[3] Harold D. Clarke and Jim Granato. Autocorrelation. In Kimberly Kempf-Leonard, editor, *Encyclopedia of Social Measurement*, pages 111–119. Elsevier, New York, 2005. ISBN 978-0-12-369398-3. doi: https://doi.org/10.1016/B0-12-369398-5/00157-2. URL `https://www.sciencedirect.com/science/article/pii/B0123693985001572`.

[4] Robin Dubin. Predicting house prices using multiple listings data. *The Journal of Real Estate Finance and Economics*, 17:35–59, 02 1998. doi: 10.1023/A:1007751112669.

[5] Garret M. Fitzmaurice, Nan M. Laird, and James H. Ware. *Applied Longitudinal Analysis*. John Wiley Sons, Inc., 2011. ISBN 9781119513469. doi: 10.1002/9781119513469.

[6] R.P. Haining. Spatial autocorrelation. In Neil J. Smelser and Paul B.

Baltes, editors, *International Encyclopedia of the Social Behavioral Sciences*, pages 14763–14768. Pergamon, Oxford, 2001. ISBN 978-0-08-043076-8. doi: https://doi.org/10.1016/B0-08-043076-7/02511-0. URL https://www.sciencedirect.com/science/article/pii/B0080430767025110.

[7] Trevor Hastie and R. Tibshirani. *Generalized Additive Models*. John Wiley Sons, Ltd, 2006. ISBN 9780471667193. doi: https://doi.org/10.1002/0471667196.ess0297.pub2. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/0471667196.ess0297.pub2.

[8] Yitong Huang. Predicting home value in california, united states via machine learning modeling. *Statistics, Optimization amp; Information Computing*, 7 (1):66–74, Jan. 2019. doi: 10.19139/soic.v7i1.435. URL http://www.iapress.org/index.php/soic/article/view/2019-M-5.

[9] Thuy Le. Redfin housing market data 2012-2021, Feb 2022. URL https://www.kaggle.com/datasets/thuynyle/redfin-housing-market-data?resource=download.

[10] Chaitra H. Nagaraja, Lawrence D. Brown, and Linda H. Zhao. An autoregressive approach to house price modeling. *The Annals of Applied Statistics*, 5(1):124 – 149, 2011. doi: 10.1214/10-AOAS380. URL https://doi.org/10.1214/10-AOAS380.

[11] California Department of Forestry, DFG Fire Protection (using data from BOR, and DOC FMMP). Boundaries and administrative and political divisions, 2009. Held by Berkeley.

[12] R.Kelley Pace, Ronald Barry, Otis W. Gilley, and C.F. Sirmans. A method

for spatial–temporal forecasting with an application to real estate prices. *International Journal of Forecasting*, 16(2):229–246, 2000. ISSN 0169-2070. doi: https://doi.org/10.1016/S0169-2070(99)00047-3. URL https://www.sciencedirect.com/science/article/pii/S0169207099000473.

[13] R Core Team. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

[14] Inc. Redfin. Redfin: Real estate, homes for sale, mls listings, agents, 2004.

[15] 2023 World Population Review. Population of counties in california (2023), 2023. URL https://worldpopulationreview.com/states/california/counties.

[16] Ali Soltani, Mohammad Heydari, Fatemeh Aghaei, and Christopher James Pettit. Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms. *Cities*, 131:103941, 2022. ISSN 0264-2751. doi: https://doi.org/10.1016/j.cities.2022.103941. URL https://www.sciencedirect.com/science/article/pii/S0264275122003808.

[17] CFI Team. Hedonic regression method, 2020. URL https://corporatefinanceinstitute.com/resources/data-science/hedonic-regression-method/#.

[18] Population Division U.S. Census Bureau. Annual estimates of the resident population for counties in california: April 1, 2010 to july 1, 2019 (co-est2019-annres-06) [data file], 2020. URL https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-total.html.

[19] Population Division U.S. Census Bureau. Annual estimates of the resident population for counties in california: April 1, 2020 to july 1, 2022 (co-est2022-pop-06) [data file], 2023. URL https://www.census.gov/data/tables/time-series/demo/popest/2020s-counties-total.html.

[20] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

[21] Jing Yao and A. Fotheringham. Local spatiotemporal modeling of house prices: A mixed model approach. *The Professional Geographer*, 68:1–13, 06 2015. doi: 10.1080/00330124.2015.1033671.

[22] Yunhui Zhao. Us housing market during covid-19: Aggregate and distributional evidence. In *IMF Working Paper No. 2020/212*. SSRN, 2020. URL https://ssrn.com/abstract=3744679.

[23] A. Zheng and A. Casari. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly, 2018. ISBN 9781491953242. URL https://books.google.com/books?id=Ho0UvgAACAAJ.

Appendices

Appendix A

Variable Descriptions

period begin/period end: The beginning of the month or end of the month listed as mm/dd/yyyy.

region: The county listed for a particular period.

Average sale to list: The mean ratio of each home's sale price divided by their list price. Excludes properties with a sale price of 50%.

Home sales: Total number of homes with a sale date during a given time period.

Inventory: Total number of active listings on the last day of a given time period.

Median active list ppsf: The median list price per square foot of all active listings.

Median active list price: The median list price of all active listings.

Median active listings: The median of how many listings were active on each day within a given time period.

Median days on market: The number of median days between the date the home was listed for sale and when the home went off-market. Excludes homes that sat on the market for more than 1 year.

Median days to close: The median number of days a home takes to go from pending to sold.

Median list price: The most recent median listing price covering all homes with a listing date during a given time period.

Median list price per square foot: The median of the most recent listing price divided by the total square feet of the property (not the lot) covering all homes with a listing date during a given time period.

Median listing with price drops: The median of how many listings were active on each day and whose current list price is less than the original list price within a given time period.

Median sale price: The median of final home sale prices covering all homes with a sale date during a given time period.

Median sale price per square foot: The median of final home sale price divided by the total square feet of the property (not the lot) covering all homes with a sale date during a given time period.

Months of supply: When data are monthly, it is inventory divided by home sales. This tells you how long it would take supply to be bought up if no new homes came on the market.

New listings: Total number of homes with a listing added date during a given time period.

Off market in two weeks: The total number of homes that went under contract within two weeks of their listing date.

Pending home sales: Total homes that went under contract during the period. Excludes homes that were on the market longer than 90 days.

Percent of median active listings with price drops: The share of median active listings that dropped their price in a given time period

Percent of total active listings with price drops: The share of total active listings that dropped their price in a given time period.

Percent off market in two weeks: The share of pending sales that went under contract within two weeks of their listing date.

Percent total homes sold with price drops: The share of homes with a sale date during a given time period where the sale price is less than the latest listing price.

Region: County of the specified state.

Sold above list: The median percent of homes sales with a sale price greater than their latest list price covering all homes with a sale date during a given time period.

Total active listings: The total number of listings that were active at any point during a given time period.

Total homes sold with price drops: The total number of homes with a sale date during a given time period and where the sale price is less than the latest listing price.

Appendix B

Missing Patterns in Data

B.1 Missingness of Raw Data

When handling and analyzing observations in a dataset, the presence of missing data, read as NA's in the R data programming language, poses as a challenge when performing analysis of our dataset. These missing values are usually due to incomplete data entry, lost files etc. Depending on the severity of missing data, it may not always be convenient to remove those observations since it is less information that we are feeding into our models. If there are thousands of observations missing a value, the more practical application is to interpolate the data. Data interpolation is an estimation method used to fill those missing values, which can be performed by the calculating a mean or mode from neighboring observations. If the number of observations with missing data is small, then removing those observations may be a better option.

It was noted that our data contained only 41 counties of California out of the 58 counties. Therefore, we are missing observations for 17 counties. These absent

Table B.1: Population size per county A-C

County	Population size
Alameda	1,733,977
Alpine	1,212
Amador	41,188
Butte	209,121
Calaveras	45,205
Colusa	21,965
Contra Costa	1,200,997

counties contained smaller populations on average, thus the market in these counties were slow in comparison to urbanized counties. Therefore, we will continue our research with only the counties that are present in our dataset. With further data analysis, the median variables that were updated on a month over month (variable(s) ending in "mom") or on a year over year (variable(s) ending in "yoy") contained thousands of missing data. Therefore, these variables will be disregarded in our continuation of our study.

Below are six plots of property type over time broken down by county. Our goal is to assess if certain counties show significant data missing for the property type variable. To reiterate, the property type variable consists of 5 different categories: All Residential (all types of property), Condominiums, Multi-Family Residential, Single Family Residential and Townhouses.

Figure B.1 furnishes insights into the missing data observed in Alpine county pertaining to Condominiums, Multi-Family and Townhouse property types. Evidently, Amador, Calaveras and Colusa counties show similar patterns. To contex-

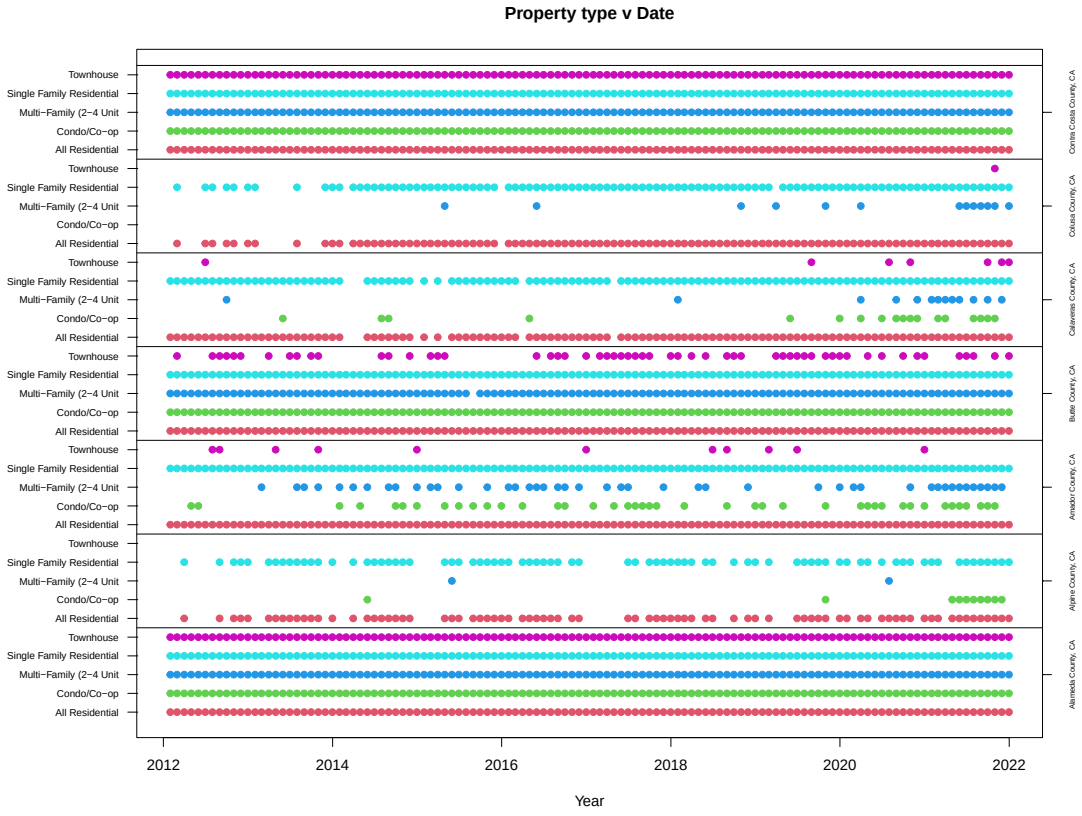


Figure B.1: Property type vs Period End for Counties A-C

tualize these results, Table B.1 provides population statistics of California counties as of 2023. This shows that Alameda boasts a population size of 1,733,977 and is large in comparison to other counties [15].

The implications of these results emphasizes the lack of urban centers in previously mentioned counties, which are less populous and designated by a shortage of property diversity. Another prominent example bolstering these findings is Contra Costa County containing a population size of 1,200,997 (note that there may be a small disparity since the recorded population sizes were recorded in 2023 compared to the last observations being December 2021) [15]. Such significance of population accentuates the presence of a more diverse urban topography.

Due to these findings, it is presumed that lack of data for certain property types for some counties can be linked to their demographic traits. Scarcity of property type diverseness can be attributed to the lack of population and local urban centers. These elaborate analyses further enhance our understanding housing market dynamics.

Analogous to the results in Figure B.1, Figure B.2 presents compelling realizations regarding Fresno, Glenn, Lake, and Madera counties, representing an absence of observations for townhouses. Glenn county does not contain enough observations for almost all property types. On the contrary, Los Angeles county surfaces as embracing all property types throughout the decade of collected data. Los Angeles county grants us the possibility of gaining insights of the housing market at a more granular level. Table B.2, further enhances these results, offering a different viewpoint. Although Fresno county contains one million residents, it presents a lack of Townhouse observations unlike El Dorado, which is a fraction of Fresno's population [15].

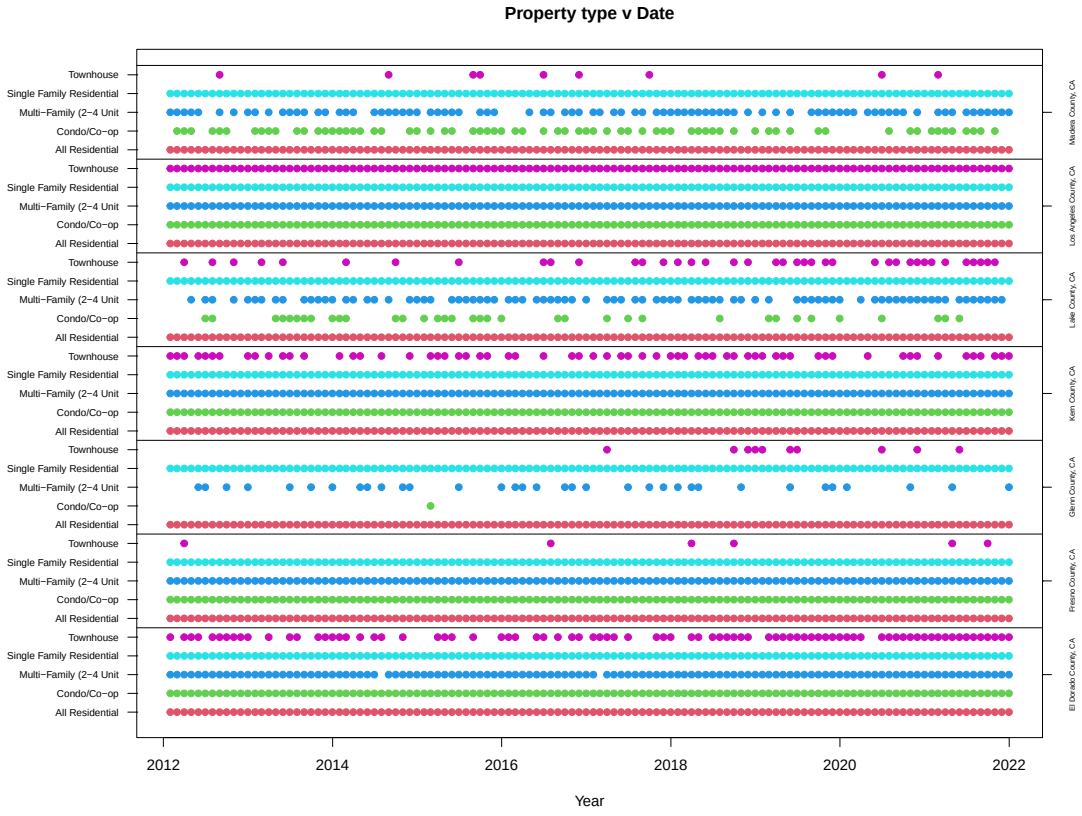


Figure B.2: Property type vs Period End for Counties E-M

Table B.2: Population size per county E-M

County	Population size
El Dorado County	194,224
Fresno County	1,032,114
Glenn County	29,157
Kern County	930,115
Lake County	69,213
Los Angeles County	10,072,629
Madera County	157,872

Decoding these analyses allows to draw informed conclusions and uncover complications of the housing market landscape. These housing dynamic is further evidence in enhancing the relationship between demographic patterns and property types.

As shown in Figure B.3, Marin, Monterey and Napa counties had observations missing in the townhouse category. Merced county had some missing observations for townhouse, multi-Family and condominiums, while Nevada county had some missing observations for townhouse and multi-family properties. Mariposa county had a considerable amount of missing observations throughout all categories.

Figure B.3 further reinforces empirical evidence on presence of missing observations have some relationship to the demographic patterns. This figure unveils the prominent absence of townhouse observations for Marin, Monterey and Napa counties. Napa county's bulk of townhomes observations are after 2020, which raises the question if some of these were from new construction and expansions. In addition, Merced county displays partial missing observations for all property types

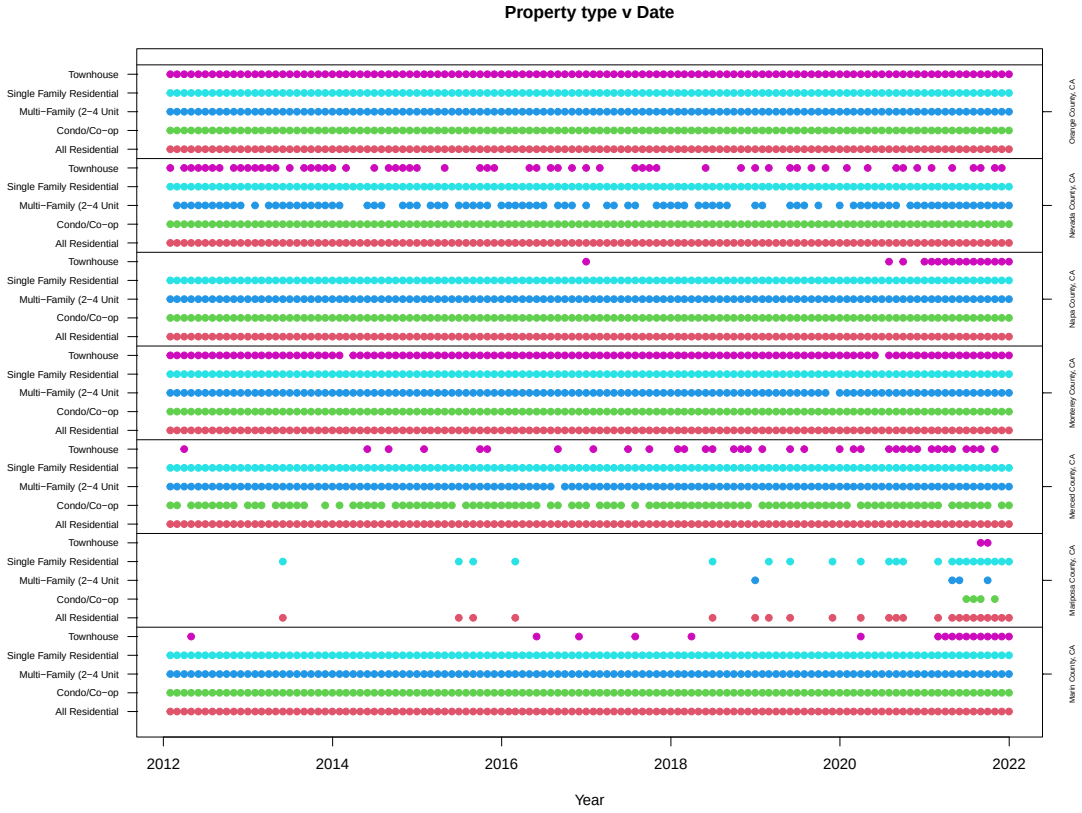


Figure B.3: Property type vs Period End for Counties M-O

Table B.3: Population size per county M-O

County	Population size
Marin County	265,294
Mariposa County	16,795
Merced County	288,825
Monterey County	446,229
Napa County	138,481
Nevada County	103,285
Orange County	3,240,017

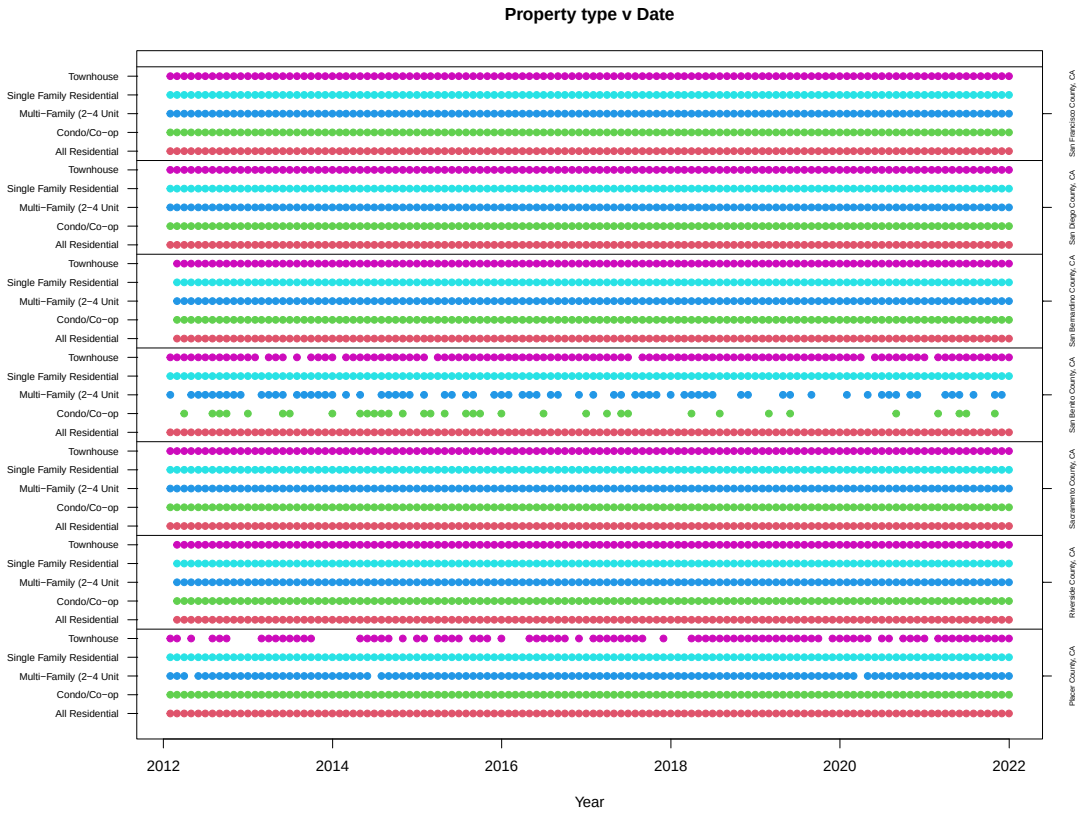


Figure B.4: Property type vs Period End for Counties P-S

except for single family residential, while Nevada county contains missing data for townhomes and multi-family properties. On the other hand, Mariposa county, one of the least populous in our data as seen in Table B.3, is inadequate in providing diverse data over time, plaguing an issue for accuracy in predictive modeling. [15] Nonetheless, the counties that provide the fullest data are Orange and Monterey counties, which are the most populous in Figure B.3.

Figure B.4 reports San Benito, while reporting a complete set of observations for single family residential and almost complete for Townhouses, it suffers from missing observations for condo/co-op and multi-family. Although it may not be absolute, it could be due to less availability of these properties leading to some

Table B.4: Population size per county P-S

County	Population size
Placer County	421,632
Riverside County	2,486,747
Sacramento County	1,634,936
San Benito County	66,891
San Bernardino County	2,225,586
San Diego County	3,359,630
San Francisco County	894,584

months where no sales occurred. Placer county to some extent displayed missing observations for townhouse and multi-family property types. Both Riverside and San Bernardino counties lack records for any property types in the first month of the data collection, leading to questions about data acquisition.

Urban metropolitan areas such as Sacramento, San Diego and San Francisco, all contain complete data for all property types in the decade of 2012 through 2021 and we see that their population sizes mirror these results [15]. Although San Francisco's population is considerably smaller to other counties, it is densely populated and the smallest area at 46.9 square miles [1].

Figures B.5 shows that San Mateo, Santa Barbara and Santa Clara counties did not have any missing observations. San Joaquin and Solano counties displayed townhouse observations missing. Santa Clara county exhibited only three missing observations for multi-family homes and lastly, San Luis Obispo county displayed missing observations for townhouses and only 2 missing for multi-family homes.

Note that the Bay Area consists of nine counties: Alameda, Contra Costa,

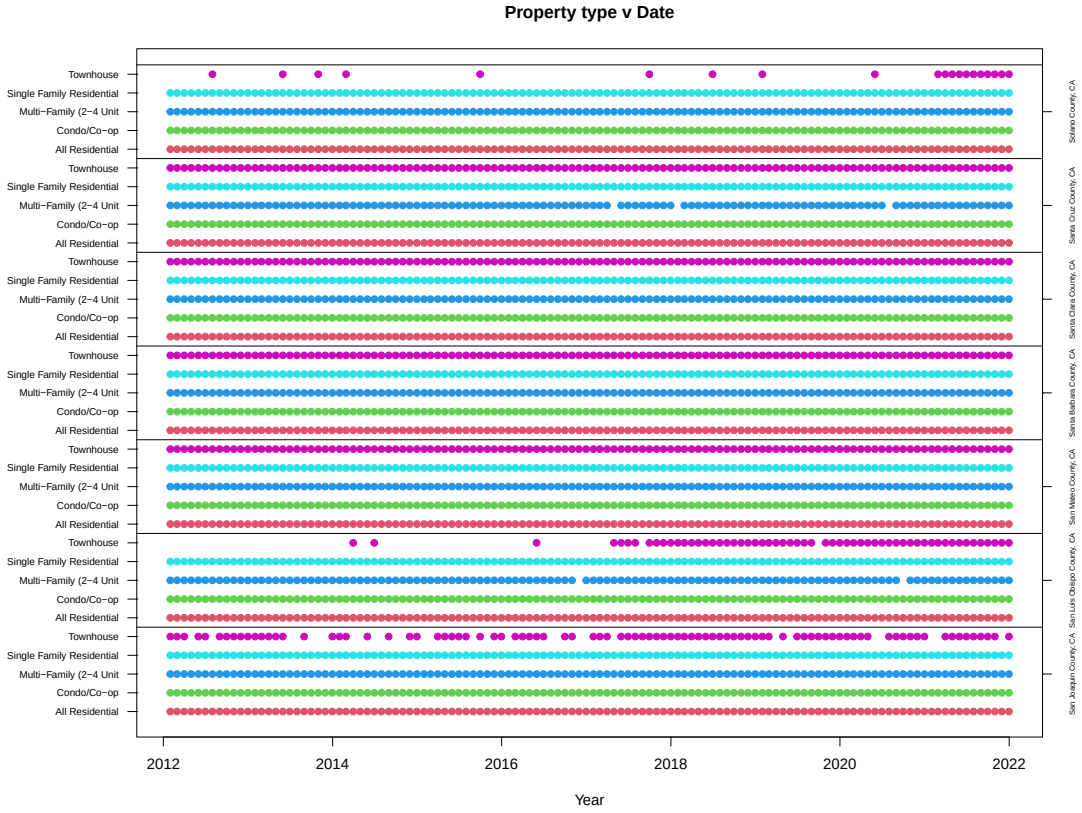


Figure B.5: Property type vs Period End for Counties S

Table B.5: Population size per county S

County	Population size
San Joaquin County	807,412
San Luis Obispo County	286,261
San Mateo County	778,239
Santa Barbara County	455,528
Santa Clara County	1,982,645
Santa Cruz County	273,405
Solano County	465,536

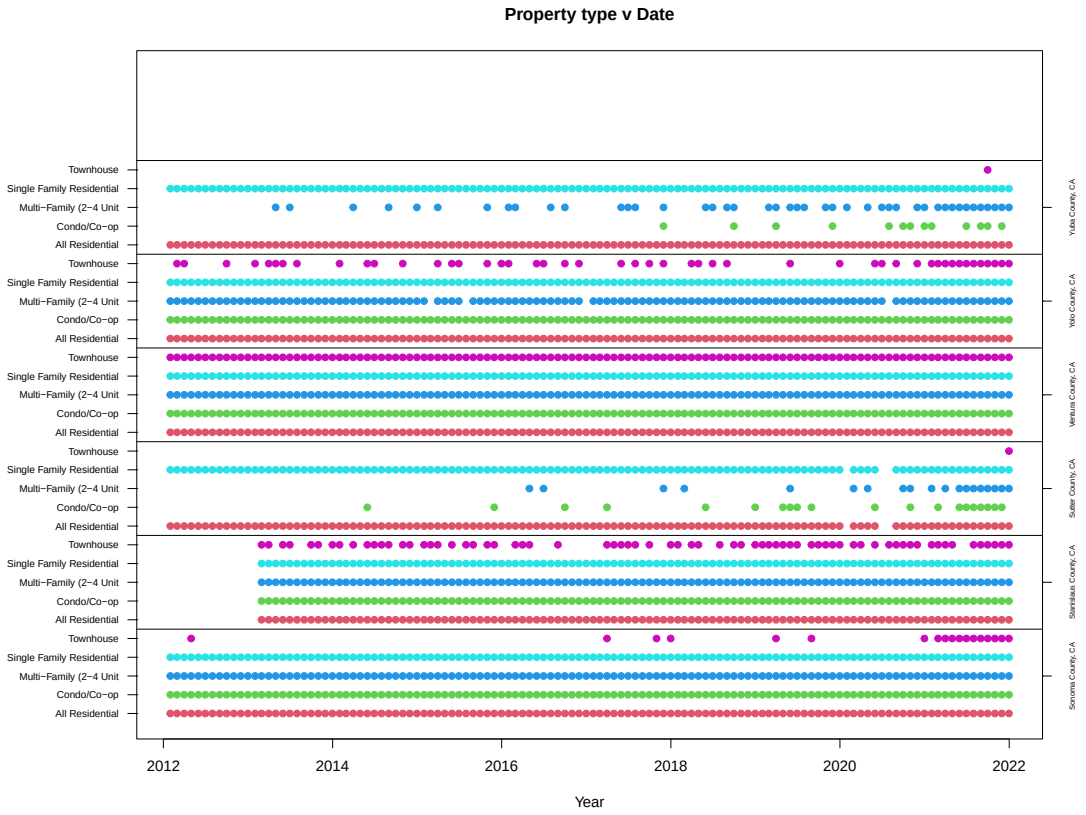


Figure B.6: Property type vs Period End for Counties S-Y

Marin, Napa, San Francisco, San Mateo, Santa Clara, Solano, and Sonoma.

Many of the counties in the Bay Area exhibit exemplary patterns in data completeness as uncovered in Figure B.5, and further examination shows parallel results with population sizes in Table B.5 [15]. Amid most of the Bay Area's exemplary patterns, Solano county illustrates a shortcoming on townhome observations, communicating a divergence from the tendency we perceived within the Bay Area.

Sutter and Yuba counties displayed missing observations to a considerable extent for townhouses, multi-family and condos. Yolo and Sonoma counties had copious townhouse observations missing. Stanislaus county did not have any records for the year of 2012, demonstrating challenges faced on capturing data over long

Table B.6: Population size per county S-Y

County	Population size
Sonoma County	490,357
Stanislaus County	564,404
Sutter County	101,103
Ventura County	849,999
Yolo County	221,068
Yuba County	84,401

timeframes. Lastly, Ventura county was the sole county to bear all observations throughout the decade.

Reflected in the insights from Figure B.6, Sutter and Yuba were recorded as the least populous counties demonstrating limited growth prospect in these areas. Yolo and Sonoma counties on the other hand, have larger populations in comparison, however, since these counties are not urban, townhomes are scarce [15].

Along with visualizing the scope of missing data, or lack thereof, it was also crucial to visualize the growth of median sale prices over the course of the decade. We wish to see these fluctuations over time by including prices in a heat map, indicating different colors for variations in price per county.

B.2 Missingness of Clean Data

Many of the plots in this section were plotted previously with the raw data in the pre-processing section. We now re-plot with the cleaned version of the data. Below are six plots of property type over time broken down by county. We want to

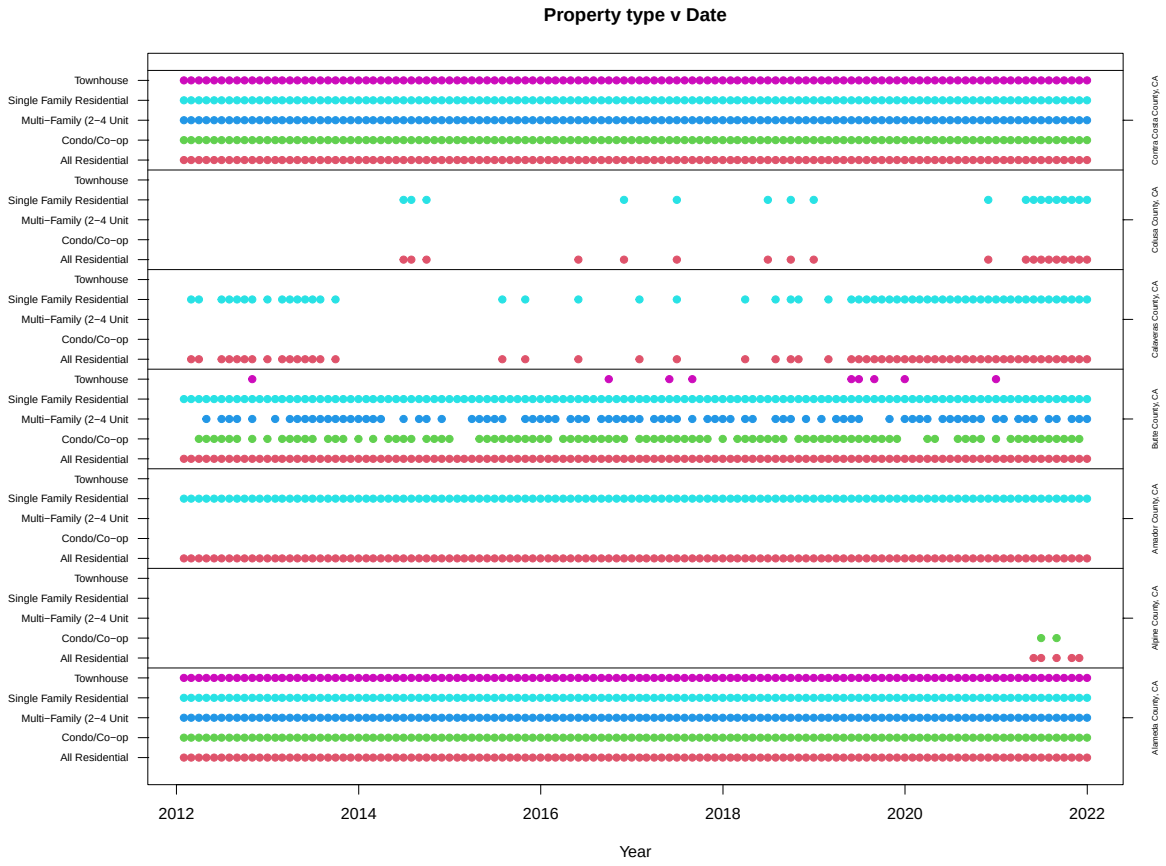


Figure B.7: Property type vs Period End Counties A-C with clean data

see if any counties show significant amounts of data missing for the property type variable. Figure B.7 visualizes property type over time, however this is now using the cleaned version of the data. Comparing this result to what was shown in Figure B.1 we see a lot more observations were removed. Note that before cleaning we had $19,967$ observations, after performing our filtering and

removals, we now have 14,428 observations. In figure B.7, Colusa county now only has 18 observations for the Single Family property type and nothing else. Previously Colusa county had over 100 observations for Colusa county for Single Family Homes. The more populated

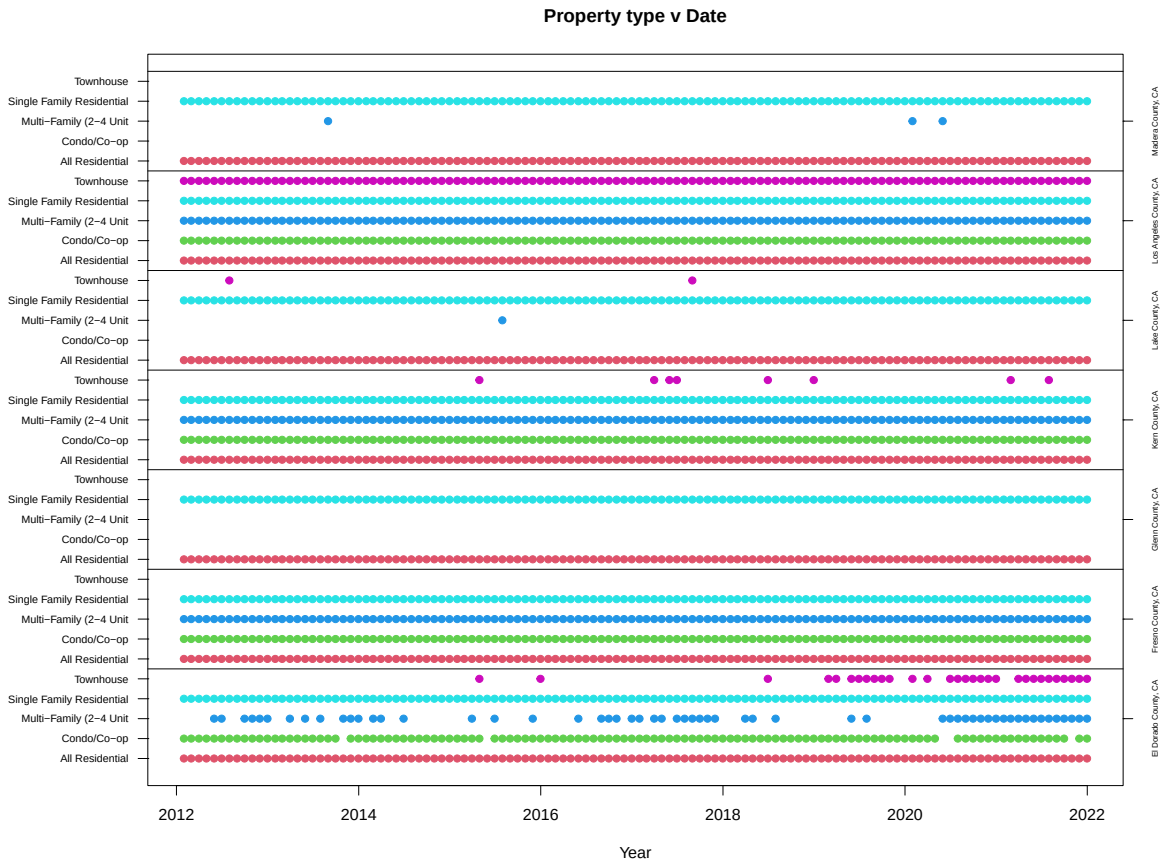


Figure B.8: Property type vs Period End Counties E-M with clean data

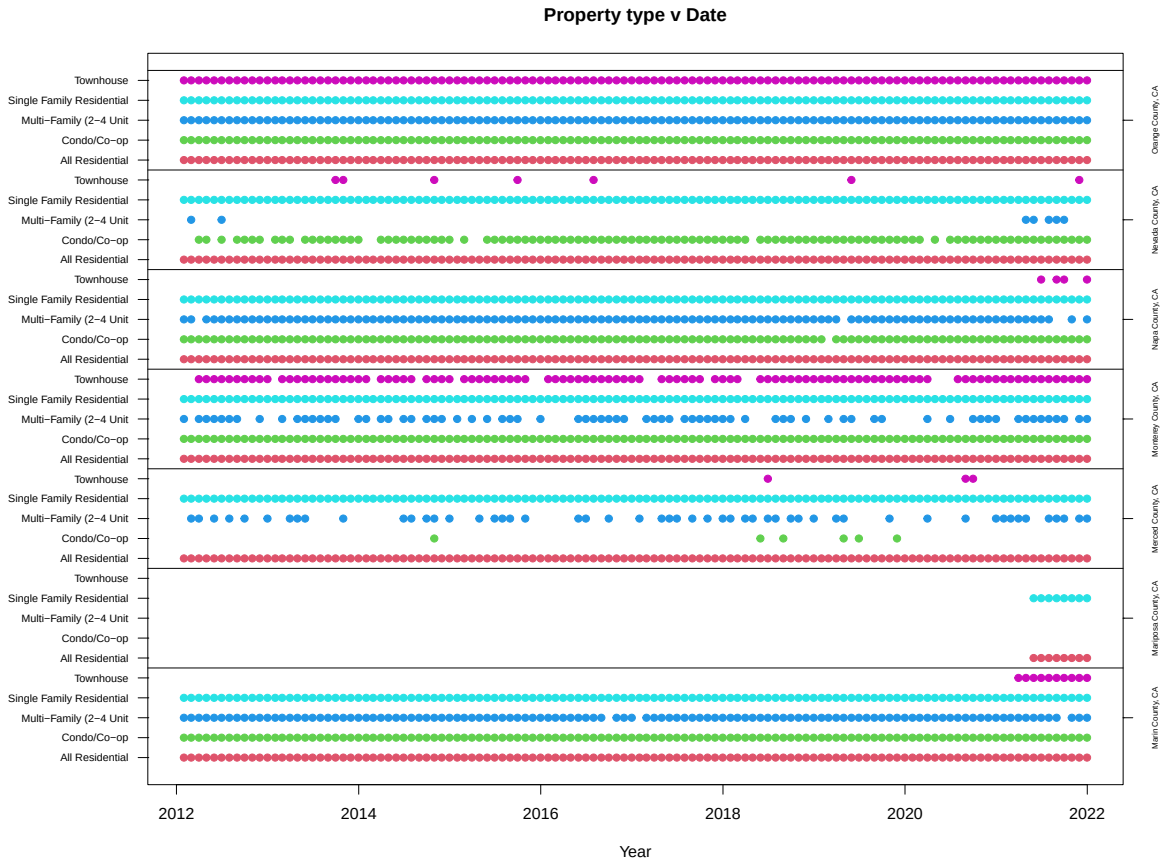


Figure B.9: Property type vs Period End Counties M-O with clean data

counties like Los Angeles county tended to not have any observations removed as more than 5 homes were sold each month. Less populated counties tended to have more observations removed as the housing market was typically smaller. Figure B.8 is comparable to figure B.2 except we have less points due to removals. Therefore, some counties have scarce observations with respect to some property types deeming them as unreliable. For instance, Fresno, Glenn, Kern, Lake and Madera counties have very few or no observations for Townhouses, deeming them as unreliable for prediction purposes for this property type, however they do have a plethora of entries for Single Family homes, so they may be used to prediction in this case.

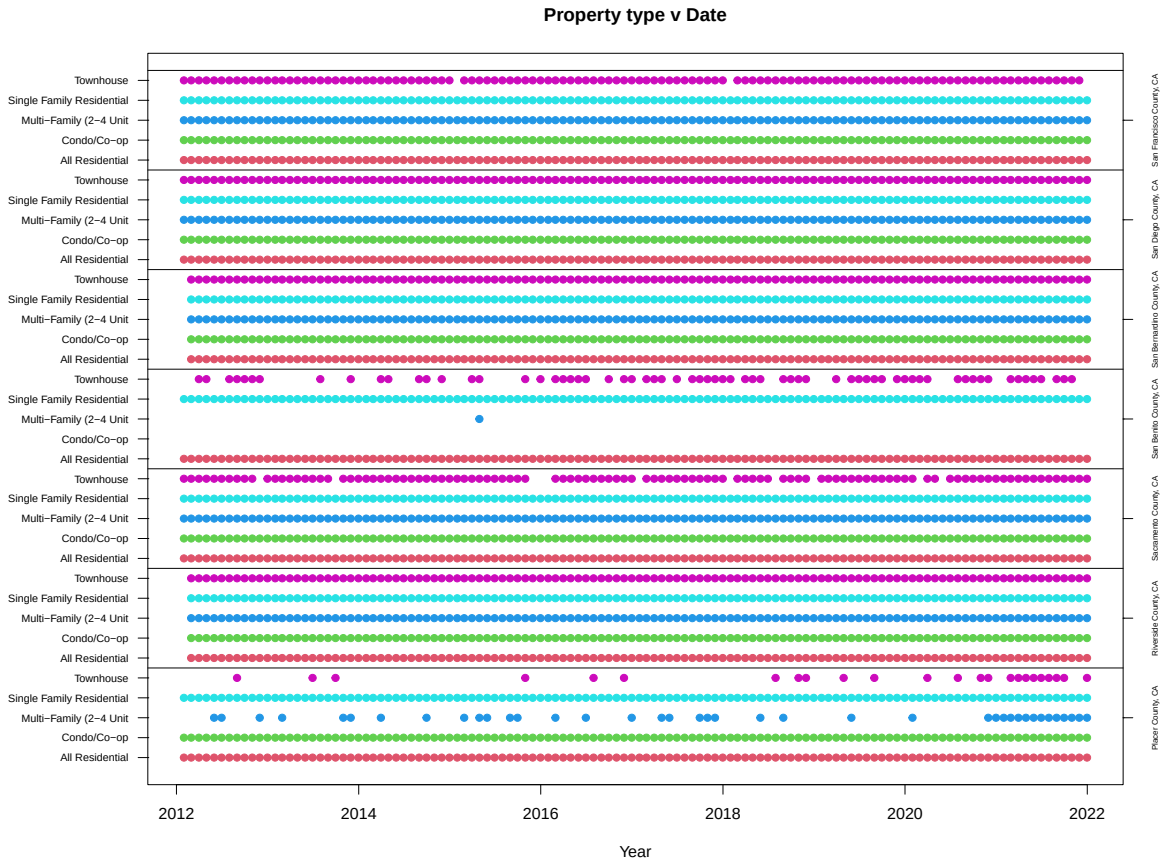


Figure B.10: Property type vs Period End Counties P-S with clean data

Figures B.9 continue this trend of missing observations, where Orange county was the only county that contained an entry for every month of every property type. Mariposa county now does not capture a full year worth of data and we do not have enough information on how prices have changed over time and if there is any seasonality.

Figures B.10 and B.11 had many counties that were complete or were shy from being complete. That is many counties had a complete set of observations for all properties over time.

Note that throughout these figures, there have been inconsistency and incom-

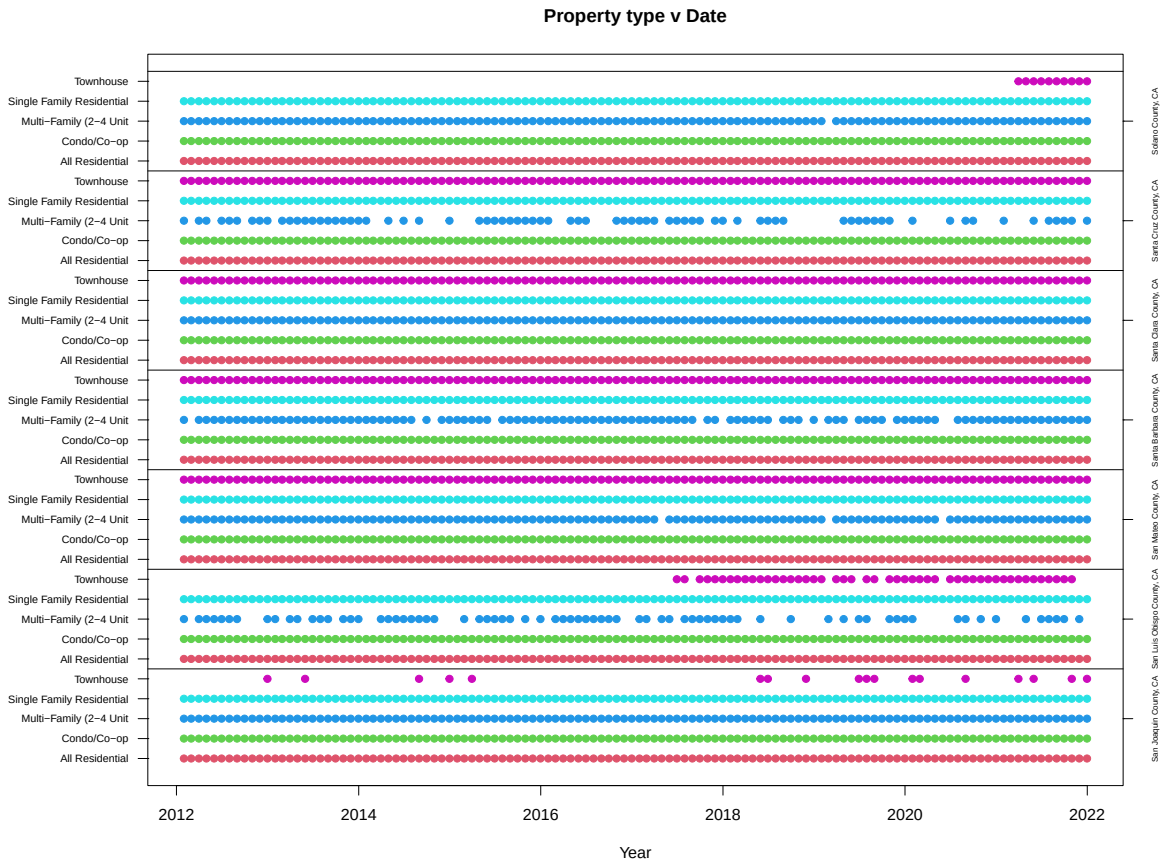


Figure B.11: Property type vs Period End Counties S with clean data

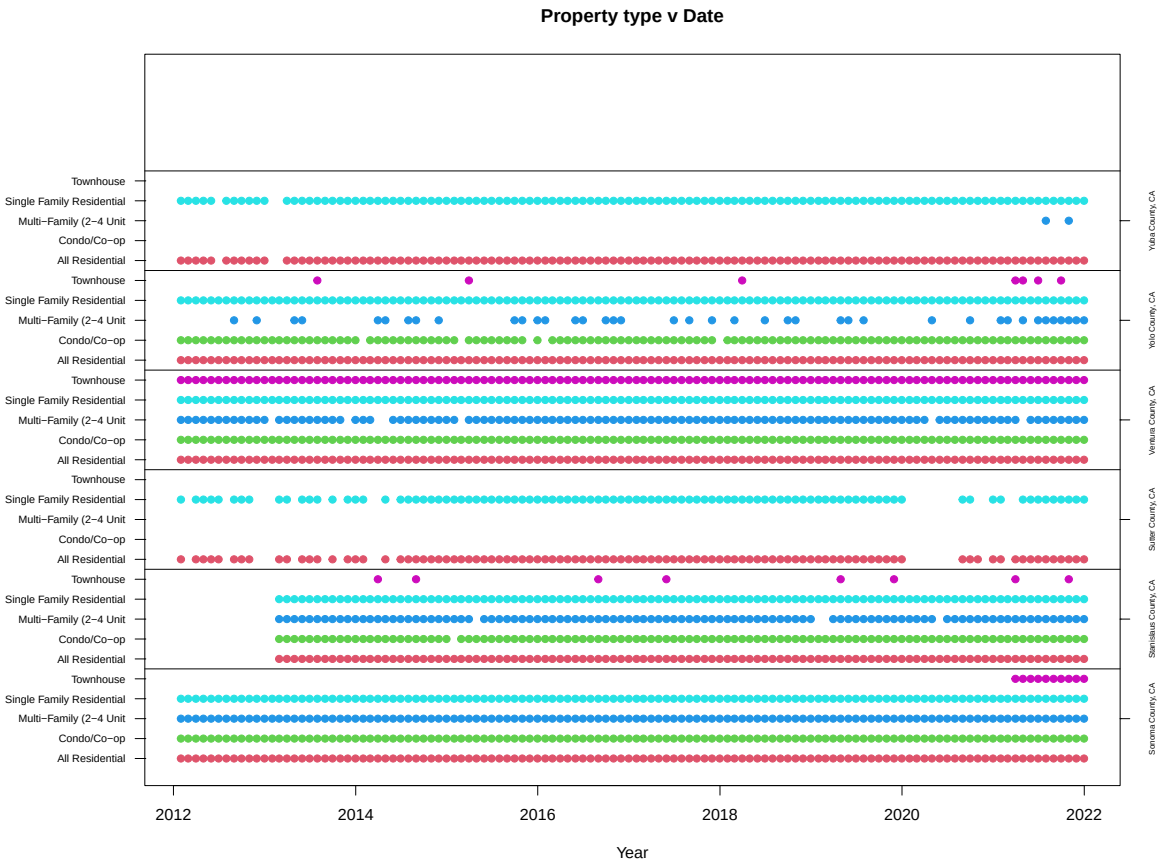


Figure B.12: Property type vs Period End Counties S-Y with clean data

pleteness. Meaning that many of counties (considered as an individual in this study) have a different number of observations. Figure B.12 is also an example of this recurring issue.

Beyond this point we have removed observations for Multi-Family observations since 2 unit, 3 unit or 4 unit should be investigated separately. However, due to the nature of how the data was collected, we don't have enough information to split these observations. We will also be removing All Residential since it is a combination of all property types, deeming it as the possibility of being a duplicate.