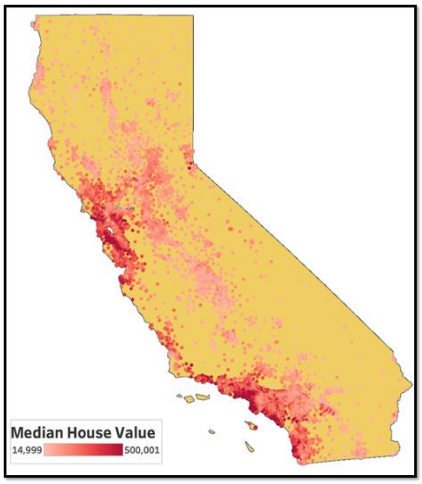
Executive Summary - Predicting California Housing Prices with Census Data

Kenneth Thomas, Jimmy Chen, RaShonda Jones, Robert Kaszubski, Wally Contreras

The California housing market has changed for various reasons over the course of the past two decades and there are many factors contributing to home value. In preparation for creating our own model of home value, we looked at several different articles that attempt to capture changes in home price. One article studied the effects of Covid-19 on supply and demand over the past two years and noted how there was a shift in demand from urban to suburban areas. A second article explained how search queries can be used to predict both present and future home sales. The final article studied the effects of labeling a single-family home as "green" and suggested that by doing so, a small markup in price may be possible.

In this paper we modelled a dataset obtained from Kaggle, containing California Housing information from a 1990 Census. The dataset possessed 20,640 observations and 14 variables. The variables were composed of our response variable, Median House Value, and other variables pertaining to age, median income, location and population density. Figure 1 shows the distribution of median house values in California.

Figure 1. California Median house values



Before creating our model, we first explored the data to learn about the relationships between variables comprising our dataset. We saw that the Median Income of a block and the block's distance to the coast were the strongest indicators of house value. Many of the other variables were found to have a lot in common with one another, suggesting that they could be reduced into a single indicator.

Using Principal Component Analysis (PCA), we were able to cluster the majority of our variables into two distinct categories: density and location. This suggests that individual variables such as Total Rooms, Total Bedrooms, Population and Household, are better off being considered as a single latent variable, we titled "Density". Likewise, the variables Distance to LA, Distance to San Diego, Distance to San Jose and Distance to San Francisco, were consolidated to the variable "Location". We were then able to create our final model using regression analysis. It was composed of five variables: density, distance, Median Income, Median Age and Distance to coast, all of which are significant predictors of Median House Value. The final model had an R-squared value of 59%, suggesting it has a moderate ability to explain how the Median House Price varies, given the other variables. Also, using the model we were able to gain insight as to how each independent variable affects the Median House Price (see Table 1). With this model we were able to make predictions about the Median House Price. A graph of predicted vs. actual Median House Prices is given in Figure 2. Several other methods were explored to model Median House Prices, such as: Canonical Correlation Analysis, Lasso Regression and Cluster Analysis. However, none of these methods garnered the insight obtained from our final model nor were they as appropriate for our research question.

Table 1. Variable effects on Median House Price

Variables	Unit Increase	Expected Effect on Median House Price
Median Income	$10,000	$38,424 increase
Median Age	1 year	$1,233 increase
Distance to coast	1 meter	$0.68 decrease
Density score	1 (dimensionless)	$7,444 increase
Location score	1 (dimensionless)	$1,340 decrease

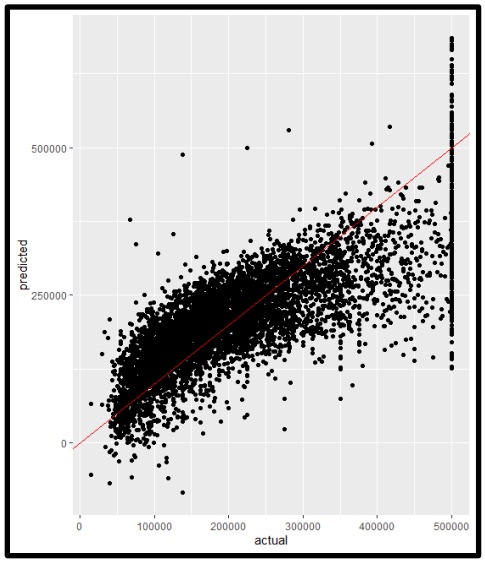


Figure 2. Actual values versus predicted values in test dataset

Several limitations exist for our model. The data used in constructing our model is from 1990, which makes drawing insight that might be applicable to the present day difficult. In the future we will consider using a newer dataset to construct our model. Also, both the Median Age of the homes and Median House Price were capped at 52 and 500,001 respectively. As a result, any value larger than the capped value was added to the maximum value count. For example, a home with a value of $700,000, was counted as $500,001. This makes the model inapplicable to values above the capped value. Also, this is not a model that can show causality between any of the predictor variables and the Median House Value - only correlations can be made.

In this study, we were able to successfully model the Median House Value using the five variables: density, distance, Median Income, Median Age and Distance to coast. The two most important variables affecting house value were median income and block density, with median age and location playing lesser roles. We see several potential future applications of our model in studies and in the real world. The model itself could be a key tool for realtors as well as potential buyers that want a quick valuation and perspective on the block they are moving into. Likewise, it can be used by construction companies and loan agencies (banks) particularly with the low interest rates of the present day to inform their decisions. We also see potential use from government agencies and officials especially with the devastating effects that the Covid-19 pandemic has on not only the housing market but those struggling to make ends meet. It can be a tool used to inform regulation and policy changes such as building new low income housing to accommodate those in need. Using the model you can see how housing value changes based on those demographic shifts. We also see application in future studies with not only our model but the latent variables created.

Predicting California Housing Prices with Census Data

Kenneth Thomas, Jimmy Chen, RaShonda Jones, Robert Kaszubski, Wally Contreras

Abstract

Housing data from the 1990 California Census is explored to identify the primary predictors of median house prices, by city block. We utilized several different analysis techniques but focused on Principal Component Analysis (PCA) and Linear Regression for our final model. PCA revealed two components, Density and Location, with a Cronbach's alpha of 0.8. The linear regression model uses the two components, is statistically useful for predictions and explains 59% of variance in median housing prices. Median income, median age, distance to coast, location, and density are all significant predictors of median housing values. This analysis was completed using 70% of the dataset, the remaining 30% was used to validate the linear model; the adjusted R-squared value of predicted values was 0.589. We conclude that it is feasible to predict median housing prices by block from census data and further discuss primary contributors.

Introduction

The California housing market has changed drastically mainly due to the Covid-19 pandemic. Consumers' buying habits and motives have changed. People are looking for more space in their living area as they are working from home. Tight inventory and low mortgage rates, similar to national housing market trends, are fueling the rise in California home prices. Many cities across California are experiencing a shortage of homes due to soaring demand. While, Covid-19 and the shift to remote work and learning has played a large part in this, we believe that other factors are still strong influencers in median housing values across California. To best explore this we took to a dataset containing information from the 1990 California census, before Covid-19 tore ripples through the housing market to find the best predictors of housing prices.

Our dataset was sourced from Kaggle and although it may not help with predicting current housing prices, it will give the audience a prediction of trends of the California housing market over the years. The data pertains to the houses found in California districts and some summary statistics about them from the 1990 census data. Overall, there are 20640 observations and 14 variables. The variables are as follows: Longitude, Latitude, Median Age (of a house within a block; a lower number is a newer building), Total Rooms (within a block), Total Bedrooms (within a block), Population (within a block), Households (a group of people residing within a home unit, within a block), Median Income (within a block; measured in tens of thousands of US Dollars), Median House Value(within a block measured in US Dollars) and five "Distance-to" variables to key locations in California (measured in meters). The original dataset was found in the StatLib repository. It was cleaned and modified prior to being posted on Kaggle. The primary modification was the inclusion of distance variables representing distances from housing blocks to major California cities (Los Angeles, San Diego, San Jose, San Francisco), and to the Pacific coast. These distances were calculated using the Haversine formula with Longitude and Latitude.

Literature Review

Prior to diving into the analysis, we researched existing studies and articles to get a better grasp of housing prices in California. The articles discussed California housing trends and different methods useful for predicting housing prices. These three articles helped us develop research questions for our dataset.

The first article looked at the most recent trends within 2020-2021. The California real estate market in 2021 has to do with a shift in demand among home buyers. In April 2020, home sales quickly decreased as the state of California went into a pandemic-driven lockdown. Buyers stopped shopping, and many sellers pulled their homes off the market. This is due to the many unknowns and uncertainties of the COVID-19 pandemic. The housing market initially went into a halt. The real estate and mortgage industries adapted to the changed world by implementing digital workflows that eliminated the need for face-to-face contact. This caused a rebound in home prices with a shift in demand from urban to suburban areas. Small towns and rural areas became more popular. Homebuyer interest in these less-populous areas increased as many moved to working remotely. Currently, many cities across the state of California are experiencing a shortage of homes for sale as there are not enough properties on the market to meet the demand from buyers. The increased demand and low supply caused a spike in the prices of homes. The median price for a single-family home in California hit $818,260 in May 2021 (B.C 2021).

The second literary source looks at how data from search engines such as Google gives an accurate and simple prediction to future business activities. This is an exploratory study investigating whether online search behavior can predict underlying economic activity. Data was collected by looking at the volume of Internet search queries related to real estate from Google trends that provide weekly and monthly reports on various industries. It allows users to obtain a query index pertaining to a specific phrase, such as “housing price.” Google Trends has also systematically captured online queries and categorized them into several predefined categories such as computer and electronics, finance and business, and real estate. Data was collected on the volume of sales of existing single-family housing units from the National Association of Realtors for all fifty states in the United States and the District of Columbia from the first quarter of 2006 to the third quarter of 2011.

The article uses a simple seasonal autoregressive model to estimate the relationship between search indices and housing market indicators. The volume of housing sales and the house price index (HPI). A single class of explanatory variables is studied: search indices for housing related queries for each state in the United States. Then it examines whether housing-related search indices could forecast future home sales. It only uses the past housing statistics to predict the future housing trends because the present housing sales and HPI are not available, proving accurate predictions are possible using dated data. The research study concluded that using housing sales data, there is evidence that search terms are correlated with future sales and prices in the housing market. This evidence lends credibility to the hypotheses that web search can be used to predict future economic activities. The microdata collected using Google Trends may prove to be one of the most powerful tools for helping consumers, businesses, and government officials make accurate predictions about the future so that they can make effective and efficient decisions (Wu & Brynjolfsson, 2009).

The residential sector accounts for 33% of electricity consumption in the U.S. One promising trend is the rise of homes labeled by a third party as “green” or energy efficient. The third article conducts a hedonic pricing analysis of all single-family home sales in California

over the time period 2007 to 2012. It finds that homes labeled with a green label transact at a small premium relative to otherwise comparable, non-labeled homes. There is evidence of spatial variation in this capitalization such that both environmental ideology and local climatic conditions play a role in explaining the variation in the green premium across geographies. The disadvantage in housing markets and the subsequent decrease intransaction prices may also have an impact on the willingness to pay for more efficient, green homes. It has been documented that prices are more procyclical for durables and luxuries as compared to prices of necessities and nondurables (Kahn & Kok, 2014).

Methods

CLARA Clustering was utilized as a data exploration tool and revealed three clusters. The clusters are characterized as: 1) a low-population density cluster in the North 2) a low-population density cluster in the South, and 3) A high-population density cluster across the state. These clusters are shown in figure 1, supporting charts are provided in the Appendix of this document showing relation to population and median house value.

Figure 1. CLARA Clustering

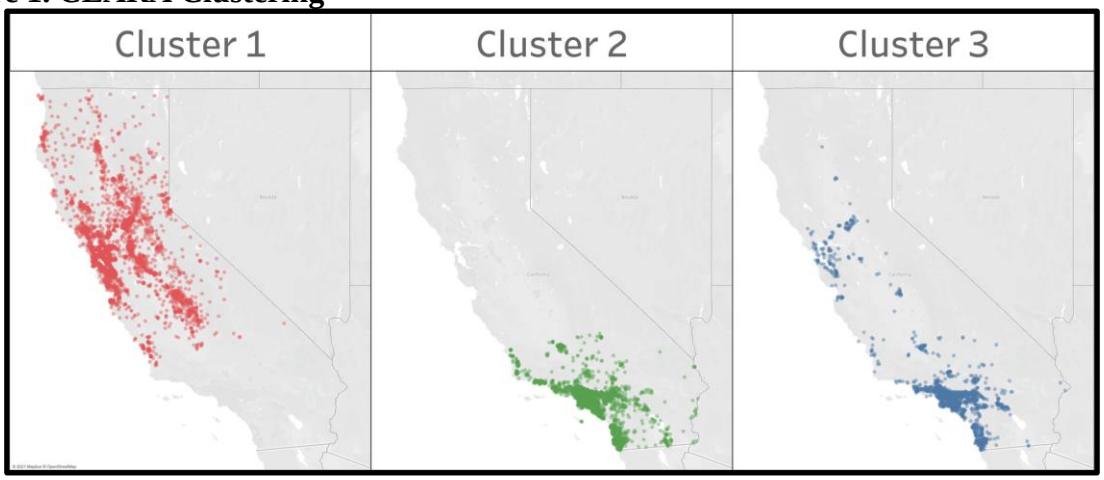
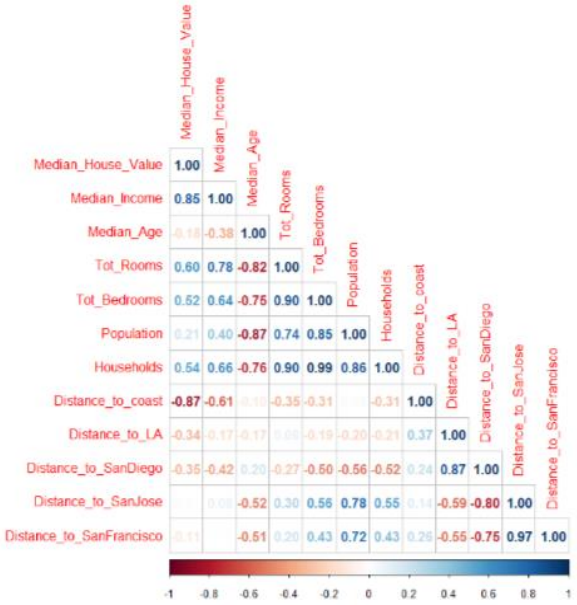


Figure 2. Correlations of all Variables

Latitude and Longitude were removed prior to any further analysis as we are not performing a geospatial analysis. The dataset was then split into a training dataset for analysis and a test dataset for model validation. Initial exploratory analysis involved observing the correlations of the variables in the dataset as seen in Figure 1. The highest correlations in relation to Median House Value were observed between Distance to Coast, and Median Income. Distance to Coast was strongly negatively correlated with a value of -0.87 while Median Income was strongly positively correlated with a value of 0.85. Several independent variables saw high correlations among each other - a possible sign of multicollinearity in the dataset. Los Angeles and San Diego are geographically near to one another as are San Jose and San Francisco. A strong positive correlation was observed among the distance variables between cities near each other, and a strong negative correlation was observed among the cities far from each other. High correlations were likewise observed among the population-based (per block) variables with Total



Rooms, Total Bedrooms, Population, and Households all seeing strong positive correlations among one another but Median Age showing strong negative correlation to each aforementioned variable.

As we are attempting to predict a value, we knew we would have to utilize some form of regression. An initial model was created utilizing all of the variables present in the dataset. This confirmed the presence of multicollinearity upon inspection of the Variance Inflation Factors (VIF). Distance to San Jose and Distance to San Francisco saw VIF scores above 100. The majority of the variables had VIF scores above 10 indicating that multicollinearity was present and had to be dealt with. The model itself saw all but one significant variable and an adjusted R-squared of 0.634. However, due to the presence of severe multicollinearity, this model had to be discarded.

Lasso regression was performed on the independent variables as a means of feature selection and to potentially reduce multicollinearity within the model. However, the analysis did not result in a reduction of many features - with only Distance to San Jose being removed - and the resulting model contained most of the same variables as the original model along with the same issue of multicollinearity. We also didn't have an issue of overfitting with our final model, so we did not require Lasso Regression later on.

Principal Component Analysis (PCA) was attempted as a means of interpretation but also to reduce multicollinearity by combining heavily correlated data into uncorrelated components. Looking at the eigenvalues, 4 components were recommended, however using the knee method on the scree plot, 2 components were recommended. The sampling adequacy and consistency of the data was tested, returning a KMO of 0.70, a Cronbach Alpha of 0.80, and a significant Bartlett's test of Sphericity. Initial PCA was conducted using four components and all variables excluding the dependent variable. This netted mixed results with PC3 and PC4 containing only 2 and 1 variable respectively. However, PC1 and PC2 showed clear and strong loadings with Total Rooms, Total Bedrooms, Population, Households, in PC1 and all distance variables sans Distance to Coast in PC2. We then utilized three components, netting similar results with Median Income placed in PC1 now but with a low loading, and PC3 still having only 2 variables. PCA was conducted again with only two components, once again reaching similar results but now excluding Median Income and Distance to Coast due to very low loadings. Those were two of the most heavily correlated variables to Median House Value, and we figured we should keep them in. These results reinforce the findings of the clustering analysis, where clusters are differentiated by population per block and geographic location.

Principal Component Analysis was re-attempted, now utilizing only the eight variables with the highest loadings that were placed in PC1 and PC2 earlier and two components. These were once again placed in PC1 and PC2 in the same fashion as described above when utilizing four factors and with similarly high loadings. PC1 was then renamed to Density, as in Population Density, while PC2 was renamed to Location. The principal component loadings are summarized

in Table 1. To gain a better understanding, the two new latent variables were then explored by attempting a canonical correlation analysis to see if any more information or insight could be extracted. Four variates were created and all four were significant. Variate 1 explained about 18% new information and Variate 2 explained about 17% new information. We could see that high loadings in Distance to Los Angeles and San Diego saw fairly low loadings across the board in the density variables. Contrastingly, high loadings in Distance to San Francisco and San Jose, saw high loading in the density variables. However, shared variance was low so this effort was not wholly successful but was helpful to see the behaviour of our Location variable in relation to Density and vice versa while confirming that we have two fairly distinct latent variables.

The scores from the two principal components were extracted from the new PCA and combined with the three variables excluded: Median Income, Median Age, and Distance to Coast. A new regression model was created using these three variables and the two components. The new model had low VIF scores nearing 1, indicating no multicollinearity, and contained all significant variables. The model's adjusted R-squared was 0.584, a minor change from our earlier model.

Table 1. Principal Component Loadings

Variables	**PC1 (Density)**	**PC2 (Location)**
Total Rooms	**0.959**	-0.024
Total Bedrooms	**0.980**	0.007
Population	**0.941**	0.051
Households	**0.985**	-0.004
Distance to LA	-0.064	**-0.932**
Distance to San Diego	-0.071	**-0.977**
Distance to San Jose	0.053	**0.954**
Distance to San Francisco	0.057	**0.978**
SS loadings	3.75	3.69
Proportion Var	0.469	0.461
Cumulative Var	0.469	0.930

*Rotation: **Varimax**
*Kaiser-Meyer-Olkin Overall MSA: 0.70
*Cronbach's Alpha: 0.8
*Bart's Test of Sphericity: p-value <0.001

The linear model was used to predict Median House Values for the test dataset (30% of the original dataset). The adjusted R-squared value for these predictions was 0.589. The RMSE was 73121 which is very close to the RMSE of 74745 using the training dataset indicating no overfitting. Actual values versus predicted values are shown in Figure 3.

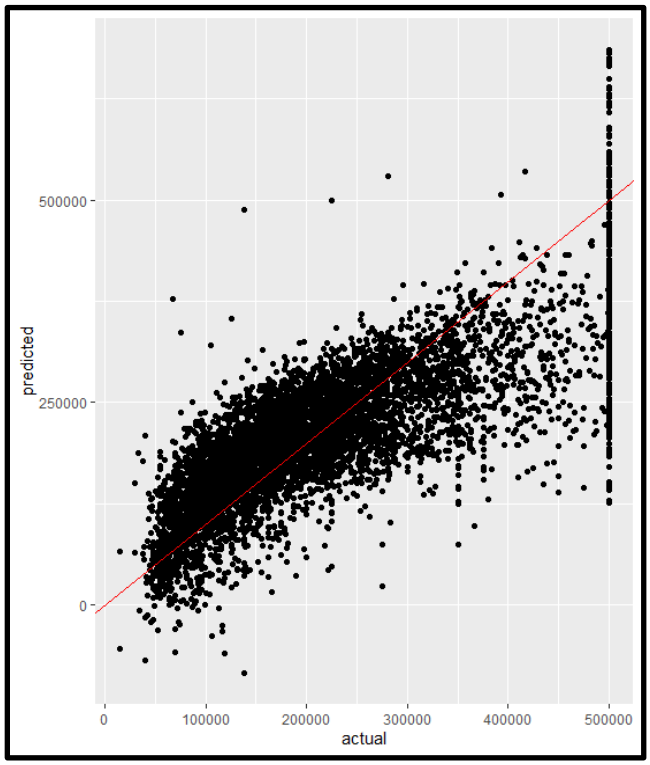


Figure 3. Actual values versus predicted values in test dataset

Discussion and Results

Principal components derived from the principal component analysis were interpreted as measures of population density and location desirability.

The density score contains variables which measure the total rooms per block, total bedrooms per block, population per block, and the number of households per block. The density score will increase as any of these density measures increase.

The location score includes variables which measure the distance from LA, San Diego, San Jose, and San Francisco. This score will increase with distance from San Jose and San Francisco and decrease with distance from LA and San Diego. The definitions of these principal components are reinforced by results from the clustering analysis where the clusters are differentiated by the population per block as well as by their geographic location.

Through linear regression, median income, median age, distance to coast, density score, and location score were identified as meaningful predictors of housing prices. Using these variables, the selected model is able to account for 59% of variance in median house prices. Furthermore, the model gives us an estimate of how much each variable contributes to median house prices; a summary is provided in table 2. Table 2 should be interpreted as: all else being equal,

- A $10,000 increase in median income is associated with a $38,425 increase median house value
- A one-year increase in median age is associated with a $1,233 increase in median house value
- A one-meter increase in distance to the coast is associated with a $0.68 decrease in median house value
- Each unit increase in the Density score is associated with a $7,444 increase median house value
- Each unit increase in the Location score is associated with a $1,341 decrease in median house value

Table 2. Linear regression beta coefficients

Variables	Unit Increase	Expected Effect on Median House Price
Median Income	$10,000	$38,425 increase
Median Age	1 year	$1,233 increase
Distance to coast	1 meter	$0.68 decrease
Density score	1 (dimensionless)	$7,444 increase
Location score	1 (dimensionless)	$1,341 decrease

The linear regression reveals a counterintuitive interpretation of the "location" principal component derived from the principal component analysis. We discover that increases in location scores are associated with a decrease in median house prices; revealing that median house prices in Los Angeles and San Diego are higher than median house prices in San Jose and San Francisco.

Limitations

Two types of limitations were encountered in this analysis, those associated with our dataset and those associated with the methods used.

A lot has changed in California since 1990. In a recent report, Cornett (2021) notes that as of January 2021, the median home price in Los Angeles County was $700,973 and the median house price in San Diego County was $670,649. This is far outside the range of our dataset, where the Median of Median house values is $179,700 and the maximum values are capped at $500,0001.

The use of the model discussed in the paper would not be advised for the current housing market. Because the methods used in our analysis rely on correlation methods, we're not able to make any causal claims and we should expect the model's predictive capability to change over time.

It's also important to note that the variables Median House Value and Median Age both capped their values at the upper end. For Median House Value, any observations of $500,001 and over were simply recorded as $500,001. Likewise for the Median Age there was a cap at 52 years meaning that any house 52 years or older, was always listed as just 52. Looking at the

distribution charts for both variables, we saw a significant chunk of the data falling in this bracket. This may have led to slight inaccuracies with our model. Essentially, a Median House Value in the millions is treated that same way as a value of half a million. Likewise, if there originally was a Median Age of, for example, 100, it would be treated the same way as an age of 52. We should be careful when the model predicts housing values less than $1,500 or greater than $500,000 as these are outside of the original dataset's range.

Future Work

Since the data from this project is from the 1990 census, it would be a good idea to analyze the same data using the 2000, 2010, and 2020 California census to see the trend over time. It would likewise be valuable to test the performance of our current model using up to date data to see how well it holds up. We could then see which locations have gotten more expensive and why? What is the median price of these same houses now? Adjusting them for inflation, is the housing market in California currently unrealistic for minimum-wage americans? The census data we used only contained information per block of people. More specific information and different variables such as square footage of the house, crime rates in the neighborhood, business diversity, and education levels could all contribute to the value of a house. Housing prices from other states such as Illinois, Texas, or New York could be compared to California. Is California overpriced? Why is the cost of housing in some states more than others?

Our model and findings can also be applied to outside studies as well as to real world applications. Our latent Density and Location variables created using PCA could be useful measures in other analysis. The model can be useful in computing Median House Values that can then be used in further studies. The model can be used by people in various professions such as those in real estate as a tool for quick valuation. It can also be used by those looking to purchase a home and see how that neighborhood fares in terms of housing value and even demographics. There is usage by those that operate construction companies, to inform their decision on whether to build on that block and how the value would change. There is likewise usage by those issuing loans for purchasing houses. Government officials and agencies can benefit from utilizing the model to inform policy decisions and introducing new regulations. They can see how demographic shifts would affect the property value of a block. This may be particularly important now with the Covid-19 pandemic displacing so many people in need. There are some limitations as it is on a per block, not per house basis, but we still see a lot of value in future applications of this model.

Conclusion

We used a 1990 California census dataset to try to determine the median home value per block in California using variables such as age, income, rooms, households, populations, and distances to various cities. We used Principal Component Analysis that ended with two components (renamed density and location) that combined with regression allowed us to find that the most important variables for predicting median home value are median income, followed by block density, and then median age. Living closer to the pacific coast and southern california by Los Angeles or San Diego also increases your home value slightly while being closer north to San Jose or San francisco will lower the home value slightly. Areas with higher income residents tend to have higher housing values. Likewise older houses tend to have higher value, a slightly surprising result pointing to a possible preference for older homes rather than modern ones.

Proximity to the coast is also a significant predictor of housing values with homes further away from the Pacific Ocean showing rapid decline in value. An older home along the coast of Los Angeles would likely be among the most expensive in the state. It would be interesting to check the housing market again using the most recent census and different types of variables to see the differences today.

References

5, B. C. | A. (2021, January 14). *5 predictions for the California housing market in 2021*. HBI News. http://www.homebuyinginstitute.com/news/california-housing-predictions-for-2021/.

Cornett, B. (2021, January 15). *Southern California Housing Forecast 2021: San Diego, Riverside, Los Angeles*. HBI News. http://www.homebuyinginstitute.com/news/southern-california-forecast-san-diego-riverside

Federsoriano. (2021, July 3). *California Housing Prices Data - Extra features.* Kaggle. https://www.kaggle.com/fedesoriano/california-housing-prices-data-extra-features

Khan, M. E., & Kok, N. (2014). The capitalization of green labels in the California housing market. *Regional Science and Urban Economics*, *47*, 25–34. https://doi.org/10.1016/j.regsciurbeco.2013.07.001

Wu, L., & Brynjolfsson, E. (2009). The future of prediction: How google searches foreshadow housing prices and sales. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2022293

Appendix

Additional Figures:

Figure 1. Canonical Correlation Helio Plot:

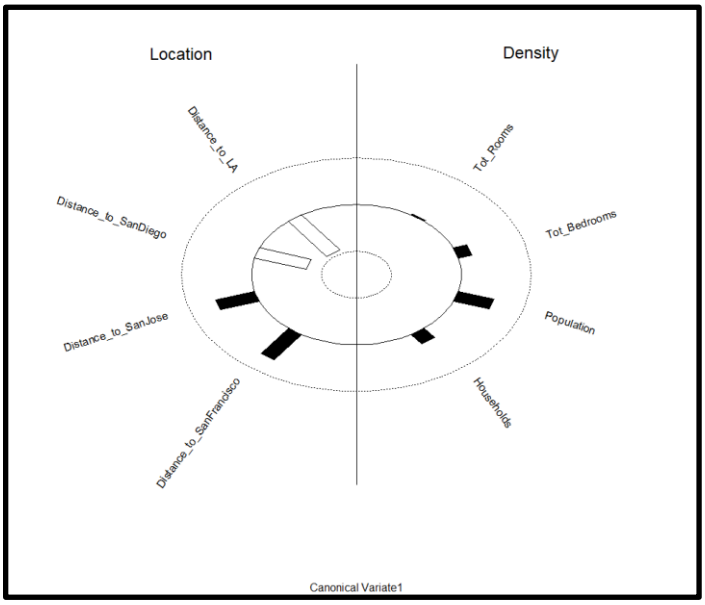


Figure 2. Cluster Analysis output

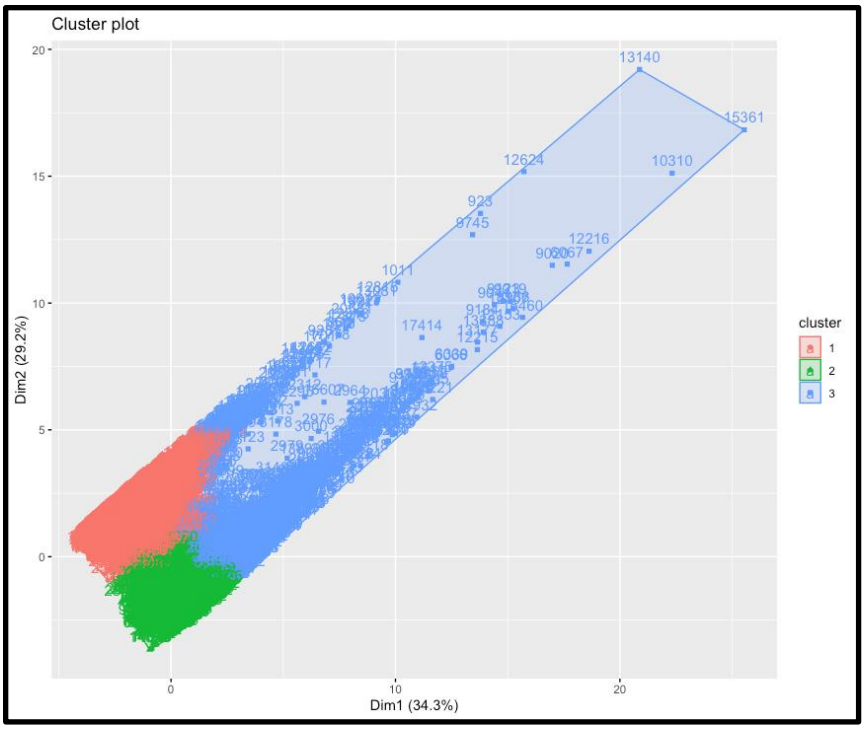


Figure 3. Cluster Analysis results: Cluster vs Population per block

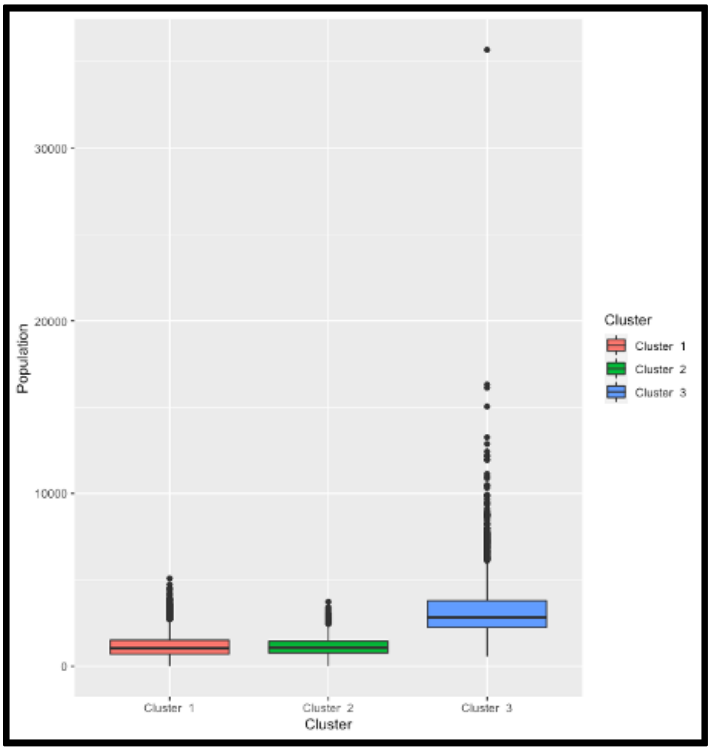
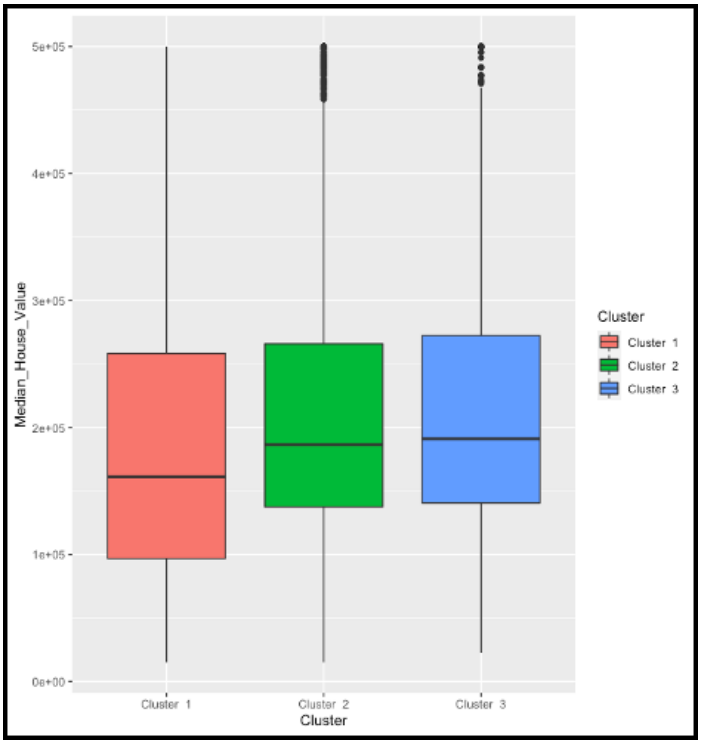


Figure 4. Cluster Analysis results: Cluster vs Median House Value



R Code:

#Final Project

#libraries

library(glmnet)

library(ggplot2)

setwd('C:/Users/rober/Desktop/Grad/DSC424/Final Project')

housing_raw <- read.csv(file="California_Houses.csv", header=TRUE, sep=",")

summary(housing_raw)

head(housing_raw)

```
names(housing_raw)


sum(is.na(housing_raw))
#no missing values


#remove longitude and latitude
housing = housing_raw[,c(1:7,10:14)]


#Split into training/testing dataset
set.seed(987648)
s = sample(nrow(housing),as.integer(nrow(housing)*.7))
housing_Train <- housing[s,]
housing_Test <- housing[-s,]


#correlation / multicollinearity
M<-cor(housing_Train, method="spearman")
M
library(corrplot)
corrplot(cor(M,method="spearman"), method = "number", type = "lower")


MTest = ifelse(M < .01, T, F)
MTest
colSums(MTest) - 1
```

```
#distribution of dependent variable (Median_House_Value)
options(scipen=10000)
library(vioplot)

vioplot(housing_Train$Median_House_Value,
        lineCol = "white",
        border = "black",
        col=3)
title("housing_Train Value ($) - Distribution")
ylab("$")

library(ggplot2)
ggplot(housing_Train, aes(x=Median_House_Value)) + geom_histogram()

#Limitation in the Data: Median House Values 500,001 and over are all listed as 500,001

ggplot(housing_Train, aes(x=Median_Age)) + geom_histogram()

#Limitation in the Data: Median Age 52 and over are all listed as 52

################################################################################
##
#initial model
```

```
################################################################################
##


model1 <- lm(Median_House_Value ~ ., data=housing_Train)

model1

summary(model1)


#confirm multicollinearity

library(DescTools)

VIF(model1)

#variables in the hundreds, several over 10.


################################################################################
##

# Perform lasso regression

################################################################################
##


x <- dplyr::select(housing_Train, -c(Median_House_Value))

y <- dplyr::select(housing_Train, c(Median_House_Value))


x<- as.matrix(x)

y<- as.matrix(y)


model_lasso <- cv.glmnet(x, y, alpha = 1)
```

```
summary(model_lasso)


model_lasso$lambda.min

coef(model_lasso, s= model_lasso$lambda.min)


################################################################################
##

#Cluster Analysis

################################################################################
##

library(cluster)

library(factoextra)


h2 <-scale(housing_Train)


#determine # of clusters - this is a bit slow ~15 seconds

fviz_nbclust(h2, kmeans)

#Suggested 3 clusters


h2_clara <- clara(h2, 3, samplesize=100, pamLike = TRUE)


# Visualize - loads but slow

#fviz_cluster(h2_clara)
```

```
#Output for spacial visualization

clusteredHousing <-
data.frame(cbind(housing_Train,"Latitude"=housing_raw$Latitude[s],"Longitude"=housing_raw
$Longitude[s],"Cluster"=paste("Cluster ",h2_clara$clustering)))

#write.csv(clusteredHousing, file="ClusteredHousing.csv")

head(clusteredHousing)

#Chart by cluster

ggplot(clusteredHousing, aes(x=Cluster, y = Median_House_Value,group=Cluster, fill =
Cluster)) + geom_boxplot()

ggplot(clusteredHousing, aes(x=Cluster, y = Population,group=Cluster, fill = Cluster)) +
geom_boxplot()

ggplot(clusteredHousing, aes(x=Population, y=Median_House_Value, color=Cluster)) +
geom_point(size=3)



##############################################################################
##

#Principal Component Analysis

##############################################################################
##

#factorability of entire dataset:

library(psych)

KMO(housing_Train)

#.71



library(REdaS)

bart_spher(housing_Train)
```

```
#chi-square:263160 p-value <0.001

library(psych)
alpha(housing_Train,check.keys=TRUE)
#.79
#remove the dependent variable prior to PCA
housing_Trainy = housing_Train[,c(1)]
housing_Testy = housing_Test[,c(1)]

#all independent variables:
housing_Trainx = housing_Train[,c(2:12)]
housing_Testx = housing_Test[,c(2:12)]

#We can now run PCA
p = prcomp(housing_Trainx, center=T, scale=T)
p

#4 components using eigenvalues

plot(p)
abline(1, 0)

#4 using line
```

```
#3 using elbow


summary(p)

#4 is a good number to start with and explains 90%


p2 = psych::principal(housing_Trainx, rotate="varimax", nfactors=4, scores=TRUE)

p2

print(p2$loadings, cutoff=.45, sort=T)


#possible solution to multicollinearity - using PCA to turn all distance variables into one component

#PC3 and PC4 have only 2 and 1 loading respectively


#PCA with 3 components

p3 = psych::principal(housing_Trainx, rotate="varimax", nfactors=3, scores=TRUE)

p3

print(p3$loadings, cutoff=.4, sort=T)


#still only 2 variables in PC3


#PCA with 2 components

p4 = psych::principal(housing_Trainx, rotate="varimax", nfactors=2, scores=TRUE)

p4

print(p4$loadings, cutoff=.4, sort=T)
```

#now Median_Income and Distance_to_coast have disappeared - distance_to_coast was highly correlated

#to Median_House_Value so we should probably keep it anyways

#Lets try PCA using only the Density and Location variables

#Exclude Median_Income, Distance_to_coast, and Median_Age from PCA

housing_Trainx2 = housing_Trainx[,c(3:6,8:11)]

housing_Testx2 = housing_Testx[,c(3:6,8:11)]

#factorability of the data used in Final PCA

#KMO

KMO(housing_Trainx)

#.70

#Bart's Sphericity Test

bart_spher(housing_Trainx)

#df = 55, Chi-squared=248509, p-value <0.001

#Chronbach's Alpha

alpha(housing_Trainx,check.keys=TRUE)

#.8

```
#the variables excluded:

housing_Trainxother = housing_Trainx[,c(1:2,7)]

housing_Testxother = housing_Testx[,c(1:2,7)]


#new PCA

p5 = psych::principal(housing_Trainx2, rotate="varimax", nfactors=2, scores=TRUE)

p5

print(p5$loadings, cutoff=.4, sort=T)


#2 components with strong loadings - removes multicollinearity and still interpretable


#Get the scores

scores<-p5$scores

summary(scores)

##############################################################################
#

#Canonical  Correlation

##############################################################################
#

#looking at the canonical correlation between the 2 components:

library(yacca)

density <- housing_Trainx[,c(3:6)]

location <- housing_Trainx[,c(8:11)]
```

```
c2 = cca(location,density)
summary(c2)

#CV1
helio.plot(c2, cv=1, x.name="Location",
        y.name="Density")

#CV2
helio.plot(c2, cv=2, x.name="Location",
        y.name="Density")

#CV3
helio.plot(c2, cv=3, x.name="Location",
        y.name="Density")
#CV4
helio.plot(c2, cv=4, x.name="Location",
        y.name="Density")

##############################################################################
##
#Linear Model with PCA and model validation with test set
##############################################################################
##
#combine the dependent variables, excluded x variables, and the 2 new components:
```

```
newhousing_Train <- cbind(housing_Trainy,housing_Trainxother,scores)

colnames(newhousing_Train) <- c("Median_House_Value", "Median_Income", "Median_Age",
"Distance_to_coast", "Density", "Location")

#get PC scores for test set

pcScores_test <- predict(p5, data=housing_Testx2)

newhousing_Test <- cbind(housing_Testy,housing_Testxother,pcScores_test)

colnames(newhousing_Test) <- c("Median_House_Value", "Median_Income", "Median_Age",
"Distance_to_coast", "Density", "Location")


#create new data frame

a = data.frame(newhousing_Train)


#new regression

model2 <- lm(Median_House_Value ~ ., data=a)

model2


#check new VIFs

VIF(model2)

#No multicollinearity!


summary(model2)

#R2 of .584 Adj-R2:0.5842

#all variables significant, location has p-value of 0.03, others <0.001
```

```
cperformance <- function(coefficients, predicted, actual) {
  n = length(predicted)
  p = nnzero(coefficients)-1
  df = n-p-1
  SSE <- sum((predicted - actual)^2)
  SST <- sum((actual - mean(actual))^2)
  R2 <- 1 - SSE / SST
  Adj_R2 <- R2 -(1-R2)*(p/df)
  RMSE = sqrt(SSE/n)
  # Model performance metrics
  data.frame(
    RMSE = RMSE,
    R2 = R2,
    Adj_R2  = Adj_R2
  )
}

pred_Test <- predict(model2,newdata = data.frame(newhousing_Test))

cperformance(coefficients(model2), pred_Test, housing_Testy)
#RMSE: 73121   R2: 0.5893   Adj-R2:0.5890
#Residual plots:
res_lmfit_test <-pred_Test- newhousing_Test$Median_House_Value
```

```
plot(res_lmfit_test)

plot(newhousing_Test$Median_House_Value,res_lmfit_test)

plot(newhousing_Test$Median_Income,res_lmfit_test)

plot(newhousing_Test$Median_Age,res_lmfit_test)

plot(newhousing_Test$Distance_to_coast,res_lmfit_test)

plot(newhousing_Test$Density,res_lmfit_test)

plot(newhousing_Test$Location,res_lmfit_test)


# Plot of the actual vs. predicted.

# First assign the x and y variables.

actual <- newhousing_Test$Median_House_Value

predicted <- pred_Test


# Create a dataframe from the x and y variables.

pred_dataframe <- data.frame(actual, predicted)


library(tidyverse) # used for plotting

pred_v_actual2 <- ggplot2::ggplot(data=pred_dataframe, aes(x=actual, y=predicted)) +
geom_point() + geom_abline(color="red")

pred_v_actual2


#comparing to the training data

pred_Train <- predict(model2,newdata = data.frame(newhousing_Train))

cperformance(coefficients(model2), pred_Train, housing_Trainy)
```

#RMSE: 74745 R2: 0.5843 Adj_R2:0.5842