

# Predicting San Jose Home Prices

*Aaron Lopez*

*6/25/2017*

## Project Description

Affordable housing is a major concern in the Bay Area of California. Many people are priced out of affordable housing options where the median home price has reached \$1 million in San Jose, California.<sup>1</sup> This trend impacts low and median income families looking to secure home ownership. Politicians and voters have voiced their concerns and willingness to address this issue by approving new policies at the ballot box. Last November voters passed Measure A allowing the city to issue \$950 million in new bonds for affordable housing.<sup>2</sup> The current mayor of San Jose, Sam Liccardo, has said, “San Jose is facing an affordable housing crises.”<sup>3</sup> To help address this issue I’d like to take a look at recent median home price trends and also predict how prices will change in the near term. This information will help to track of affordability and give insights into the effectiveness of government policy changes.

There are two major client groups. First, the city of San Jose who has made housing affordability a priority as the city continues to grow and city officials create urban planning policy. Local leaders have expressed their desire to introduce an implement policy to make housing more affordable for low and middle-income residents. Understanding the trend and where housing prices are headed will help politicians determine what policy may be needed or developed. The second group of clients would be low and median income individuals seeking to find affordable housing. This information can help them make an informed decision on when and if to purchase a home in the near future.

To predict near term median home prices I will use three different approaches, linear regression, time-series (ARIMA), and regression tree. I will then compare each model to actual median home prices from a test data set. Data will be used from the city of San Jose outlined below.

## Data Set

The main source of data will be from the City of San Jose. Economic and labor data including unemployment, total jobs, jobs by sector, median home prices, median rental prices will all be utilized along with broader economic data such as average mortgage rates from FRED. The data set is approximately 6 years of monthly data from 2009-2015 after merging and cleaning up any missing values. There are 6 csv files all under the “Data” file.

apt\_rents.csv  
home\_prices.csv  
jobs\_by\_sector.csv  
mortgagerates.csv  
total\_jobs.csv  
unemployment.csv

San Jose provides relatively clean uniform data however it can be a little outdated and incomplete. I would have liked to include crime data but this was missing at the time of my analysis and a time frame for requesting the data was uncertain. Including this information may be useful in further analysis.

## Data Wrangling

The main steps to cleaning up the data were simplifying the variable names, formatting strings, and converting objects. The city of San Jose has some consistency in their data sets but to combine all the data I had to do

some clean up of individual csv files before combining it all into one data frame for analysis.

```
# Load Data and Libraries

library(dplyr)
library(tidyr)
library(reshape2)
library(plyr)
jobs_by_sector <- read.csv("Data/jobs_by_sector.csv")
home_prices <- read.csv("Data/home_prices.csv")
total_jobs <- read.csv("Data/total_jobs.csv")
unemployment <- read.csv("Data/unemployment.csv")
apt_rents <- read.csv("Data/apt_rents.csv")
mortgagerates <- read.csv("Data/mortgagerates.csv")
jobs_by_sector <- tbl_df(jobs_by_sector)
total_jobs <- tbl_df(total_jobs)
unemployment <- tbl_df(unemployment)
apt_rents <- tbl_df(apt_rents)
mortgagerates <- tbl_df(mortgagerates)

# Bring column names (dates) into a unique column so table is in long format
jobs_by_sector <- melt(jobs_by_sector, variable.name = "Sector")

# Rename column to correct variable name
colnames(jobs_by_sector)[2] <- "Date"

# Move sector names into column names moving the table back to wide format
jobs_by_sector <- spread(jobs_by_sector, Sector, value)

# Rename column names to be short and concise
colnames(jobs_by_sector)[3] <- "Education"
colnames(jobs_by_sector)[4] <- "Finance"
colnames(jobs_by_sector)[6] <- "Hospitality"
colnames(jobs_by_sector)[8] <- "Mining"
colnames(jobs_by_sector)[9] <- "Other"
colnames(jobs_by_sector)[10] <- "Business"
colnames(jobs_by_sector)[11] <- "Public"
colnames(jobs_by_sector)[12] <- "Trade"

# Changing dates from string to date format as.Date returns NA for "Sept"
# abbreviation so first we'll change that to "Sep"
jobs_by_sector$Date <- gsub("Sept", "Sep", jobs_by_sector$Date)
# Adding a day
jobs_by_sector$Date <- paste0("01.", jobs_by_sector$Date)
# Format date
jobs_by_sector$Date <- as.Date(jobs_by_sector$Date, format = "%d.%b.%Y", "%b/%d/%Y")

# Setup column names for gather function to put back into long format
jobs_by_sector <- jobs_by_sector %>%
  gather(Construction, Education, Finance, Information, Hospitality,
        Manufacturing, Mining, Other, Business, Public, Trade, Unclassified,
        key = "JobSector", value = "NumJobs")

# Setting up dates "Sept" replaced with "Sep"
```

```

home_prices$Date <- gsub("Sept", "Sep", home_prices$Date)

# Adding a day then format to date
home_prices$Date <- home_prices$Date %>%
  paste("01", sep = " ") %>%
  as.Date(format = "%b %Y %d", "%b/%d/%Y")

# Same steps for total_jobs df
total_jobs$Date <- gsub("Sept", "Sep", total_jobs$Date)
total_jobs$Date <- total_jobs$Date %>%
  paste("01", sep = " ") %>%
  as.Date(format = "%b %Y %d", "%b/%d/%Y")

# Again for unemployment df
unemployment$Date <- gsub("Sept", "Sep", unemployment$Date)
unemployment$Date <- unemployment$Date %>%
  paste("01", sep = " ") %>%
  as.Date(format = "%b %Y %d", "%b/%d/%Y")

# Slightly different strategy for apt_rents df. Dates are in integer format.
# First I added the day then converted to date format.
apt_rents$Date <- paste0("01/", apt_rents$Date)
apt_rents$Date <- apt_rents$Date %>%
  as.Date(format = "%d/%m/%Y", "%b/%d/%Y")

# Cleaning up mortgage rates with more concise variable names then date format
colnames(mortgagerates)[1] <- "Date"
colnames(mortgagerates)[2] <- "Rates"
mortgagerates$Date <- mortgagerates$Date %>%
  as.Date(format = "%Y-%m-%d", "%b/%d/%Y")

# Combine all data frames by date and arranging in descending order
clean_df <- join_all(list(jobs_by_sector, apt_rents, home_prices, total_jobs,
                        unemployment, mortgagerates), by="Date", type="left")
clean_df <- arrange(clean_df, desc(Date))

# Cleaning up some formatting to convert data types to numeric by removing dollar
# signs and commas
clean_df$Condo.Townhome <- gsub("\\$|,", "", clean_df$Condo.Townhome)
clean_df$Single.Family.Home <- gsub("\\$|,", "", clean_df$Single.Family.Home)
clean_df$X2.bedroom <- gsub("*,", "", clean_df$X2.bedroom)
clean_df$X1.Bedroom <- gsub("*,", "", clean_df$X1.Bedroom)
clean_df$Average <- gsub("*,", "", clean_df$Average)

# Converting integers to numeric across variables
clean_df[3:9] <- lapply(clean_df[3:9], as.numeric)

# Converting job sectors from characters to factors
clean_df$JobSector <- as.factor(clean_df$JobSector)

# Renaming variables with more descriptive concise names
colnames(clean_df)[4] <- "AvgAptRent"
colnames(clean_df)[5] <- "Avg1bdAptRent"

```

```

colnames(clean_df)[6] <- "Avg2bdAptRent"
colnames(clean_df)[7] <- "AvgPriceHome"
colnames(clean_df)[8] <- "AvgPriceCondo"
colnames(clean_df)[9] <- "TotalJobs"
colnames(clean_df)[10] <- "URateSJ"
colnames(clean_df)[11] <- "URateSJMetro"

# Removing missing data
clean_df <- na.omit(clean_df)

# Adding a month variable for exploratory analysis
month_prices <- c('Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep',
                  'Oct', 'Nov', 'Dec')
clean_df$Month <- factor(format(clean_df$Date, "%b"), levels = month_prices)

# Create final new csv file
write.csv(clean_df, file = "Data/clean_df.csv")

# Summary of the final data frame
str(clean_df)

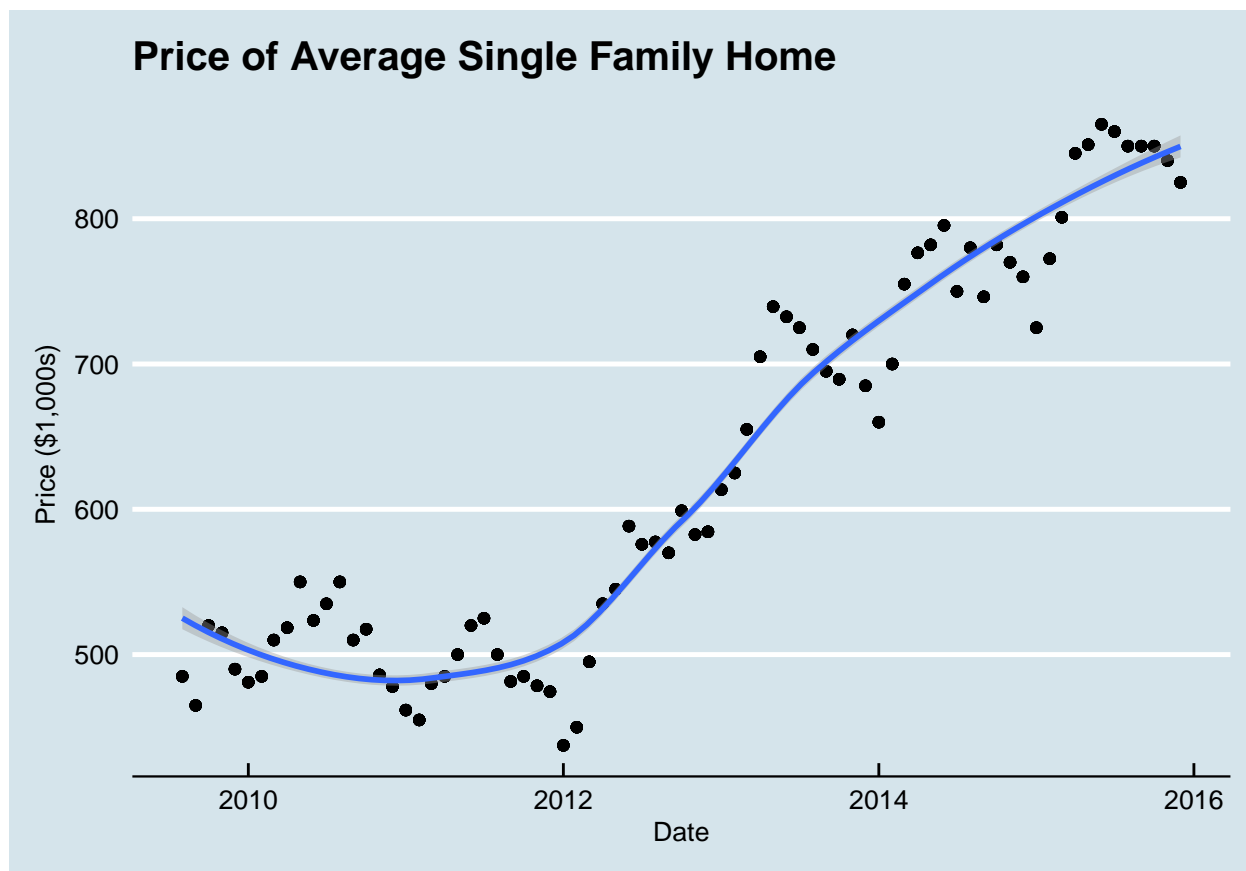
## 'data.frame': 924 obs. of 13 variables:
## $ Date : Date, format: "2015-12-01" "2015-12-01" ...
## $ JobSector : Factor w/ 12 levels "Business","Construction",...: 2 3 4 6 5 7 8 9 1 10 ...
## $ NumJobs : num 22886 82406 16574 11972 41518 ...
## $ AvgAptRent : num 2791 2791 2791 2791 2791 ...
## $ Avg1bdAptRent: num 2469 2469 2469 2469 2469 ...
## $ Avg2bdAptRent: num 3073 3073 3073 3073 3073 ...
## $ AvgPriceHome : num 825000 825000 825000 825000 825000 825000 825000 825000 825000 825000 ...
## $ AvgPriceCondo: num 465000 465000 465000 465000 465000 465000 465000 465000 465000 465000 ...
## $ TotalJobs : num 403194 403194 403194 403194 403194 ...
## $ URateSJ : num 4.2 4.2 4.2 4.2 4.2 4.2 4.2 4.2 4.2 4.2 ...
## $ URateSJMetro : num 3.9 3.9 3.9 3.9 3.9 3.9 3.9 3.9 3.9 3.9 ...
## $ Rates : num 3.96 3.96 3.96 3.96 3.96 3.96 3.96 3.96 3.96 3.96 ...
## $ Month : Factor w/ 12 levels "Jan","Feb","Mar",...: 12 12 12 12 12 12 12 12 12 12 ...
## - attr(*, "na.action")=Class 'omit' Named int [1:228] 925 926 927 928 929 930 931 932 933 934 ...
## .. ..- attr(*, "names")= chr [1:228] "925" "926" "927" "928" ...

```

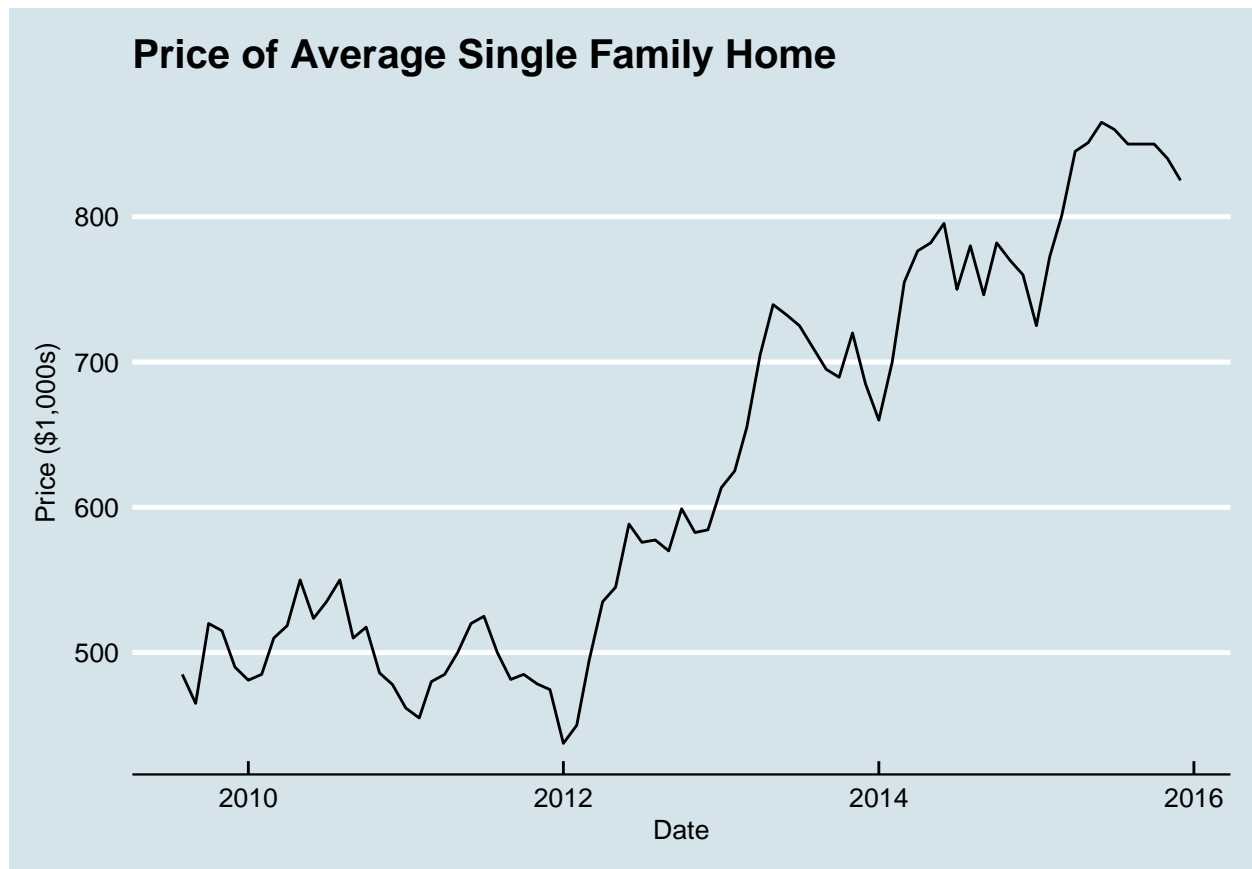
## Exploratory Analysis

To start exploring the data set let's look at correlations between the variables. Rental prices, house prices, condo prices, jobs, and unemployment all are highly correlated either positively or negatively, as one would expect. The two variables that are the most interesting are "Rates" and "NumJobs". "Rates" refers to the average 30-year mortgage interest rate and "NumJobs" is the number of jobs tied to each job sector i.e. construction, business, finance, etc. The positive or negative correlations are what you would expect but not as strong as some of the other variables.

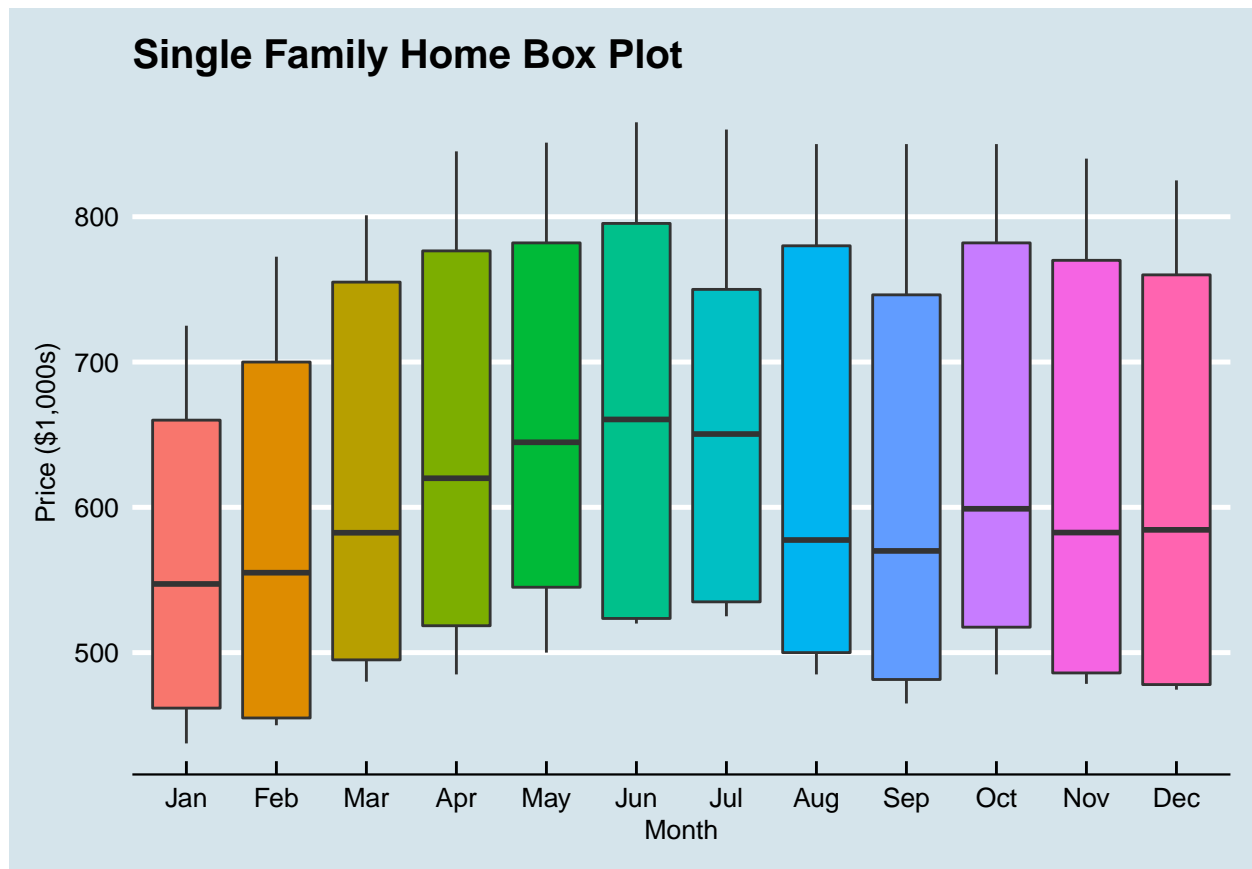
##		NumJobs	AvgAptRent	Avg1bdAptRent	Avg2bdAptRent
##	NumJobs	1.00000000	0.06235937	0.06197694	0.06256243
##	AvgAptRent	0.06235937	1.00000000	0.99648134	0.99476460
##	Avg1bdAptRent	0.06197694	0.99648134	1.00000000	0.99040318
##	Avg2bdAptRent	0.06256243	0.99476460	0.99040318	1.00000000
##	AvgPriceHome	0.06188430	0.92368942	0.90848338	0.92613877
##	AvgPriceCondo	0.06126985	0.90730429	0.88731460	0.90631008
##	TotalJobs	0.06526213	0.95552156	0.94966168	0.95863311
##	URateSJ	-0.06362893	-0.96168471	-0.96505948	-0.95925571
##	URateSJMetro	-0.06365025	-0.96249251	-0.96531654	-0.95975949
##	Rates	-0.03349193	-0.56699496	-0.60168275	-0.53628273
##		AvgPriceHome	AvgPriceCondo	TotalJobs	URateSJ
##	NumJobs	0.0618843	0.06126985	0.06526213	-0.06362893
##	AvgAptRent	0.9236894	0.90730429	0.95552156	-0.96168471
##	Avg1bdAptRent	0.9084834	0.88731460	0.94966168	-0.96505948
##	Avg2bdAptRent	0.9261388	0.90631008	0.95863311	-0.95925571
##	AvgPriceHome	1.00000000	0.97569997	0.94824225	-0.92856164
##	AvgPriceCondo	0.9757000	1.00000000	0.93882699	-0.90922165
##	TotalJobs	0.9482422	0.93882699	1.00000000	-0.97497483
##	URateSJ	-0.9285616	-0.90922165	-0.97497483	1.00000000
##	URateSJMetro	-0.9315962	-0.91185750	-0.97530145	0.99977450
##	Rates	-0.4376577	-0.41007855	-0.51319093	0.61154583
##		URateSJMetro	Rates		
##	NumJobs	-0.06365025	-0.03349193		
##	AvgAptRent	-0.96249251	-0.56699496		
##	Avg1bdAptRent	-0.96531654	-0.60168275		
##	Avg2bdAptRent	-0.95975949	-0.53628273		
##	AvgPriceHome	-0.93159617	-0.43765766		
##	AvgPriceCondo	-0.91185750	-0.41007855		
##	TotalJobs	-0.97530145	-0.51319093		
##	URateSJ	0.99977450	0.61154583		
##	URateSJMetro	1.00000000	0.61115102		
##	Rates	0.61115102	1.00000000		



Next let's start visualizing some of the data with a few plots. First, let's look at a time series of home prices. The data starts after the bottom of the financial crises in 2009 and then there is a second drop in prices a few years after with a subsequent rebound in more recent years to pre-recession levels. There seems to be some short-term trends or seasonality in housing prices that may be easier to visualize with other plots but overall the data set reflects a time of home price appreciation in San Jose.

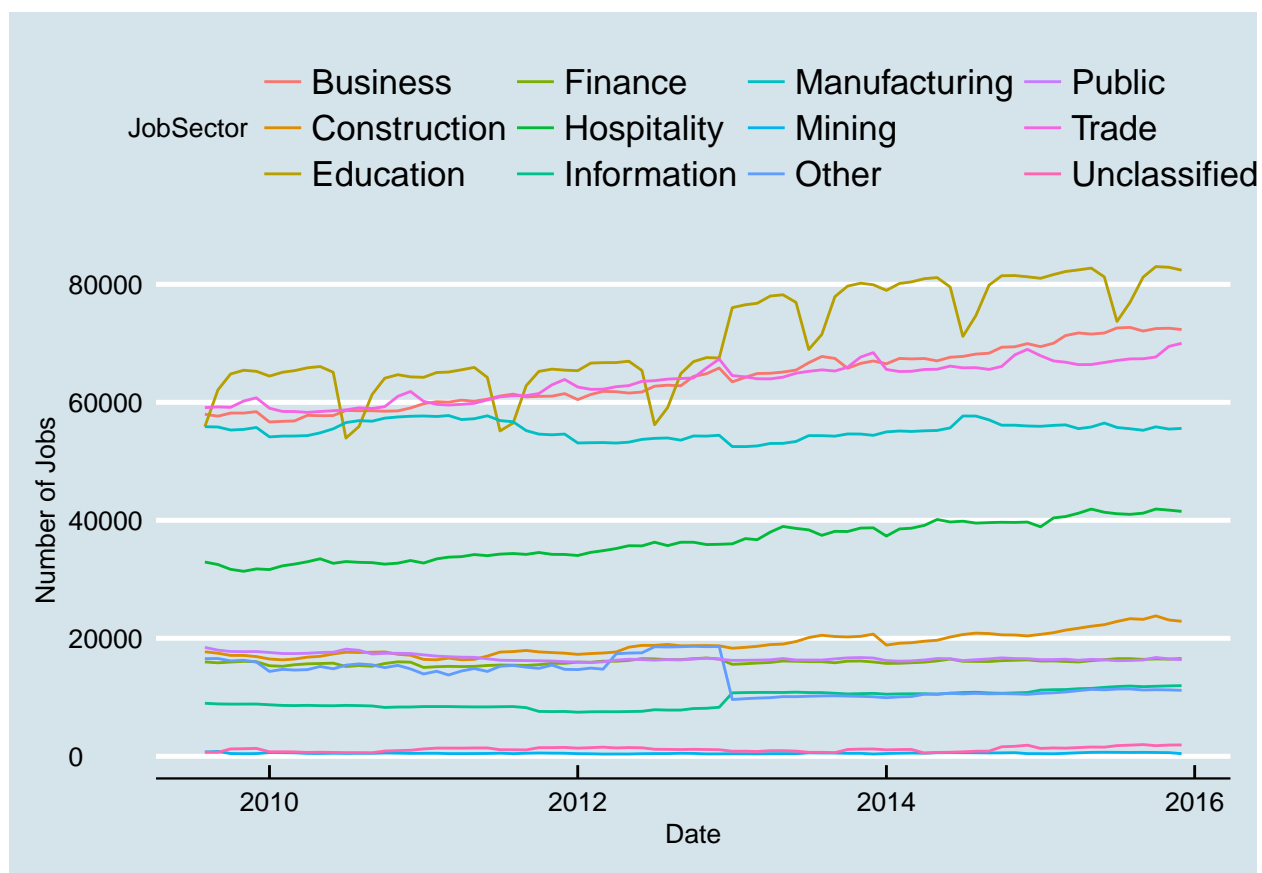


With a time series in a simple line graph its easier to see the up and down movement of home prices that perhaps reveals a seasonal trend. I would expect home prices to rise during the summer buying months when people are perhaps more inclined to go out and look for homes. During the winter months, people tend to stay in more although winters are not that harsh in San Jose. A box plot would be useful to analyze this on a monthly basis.

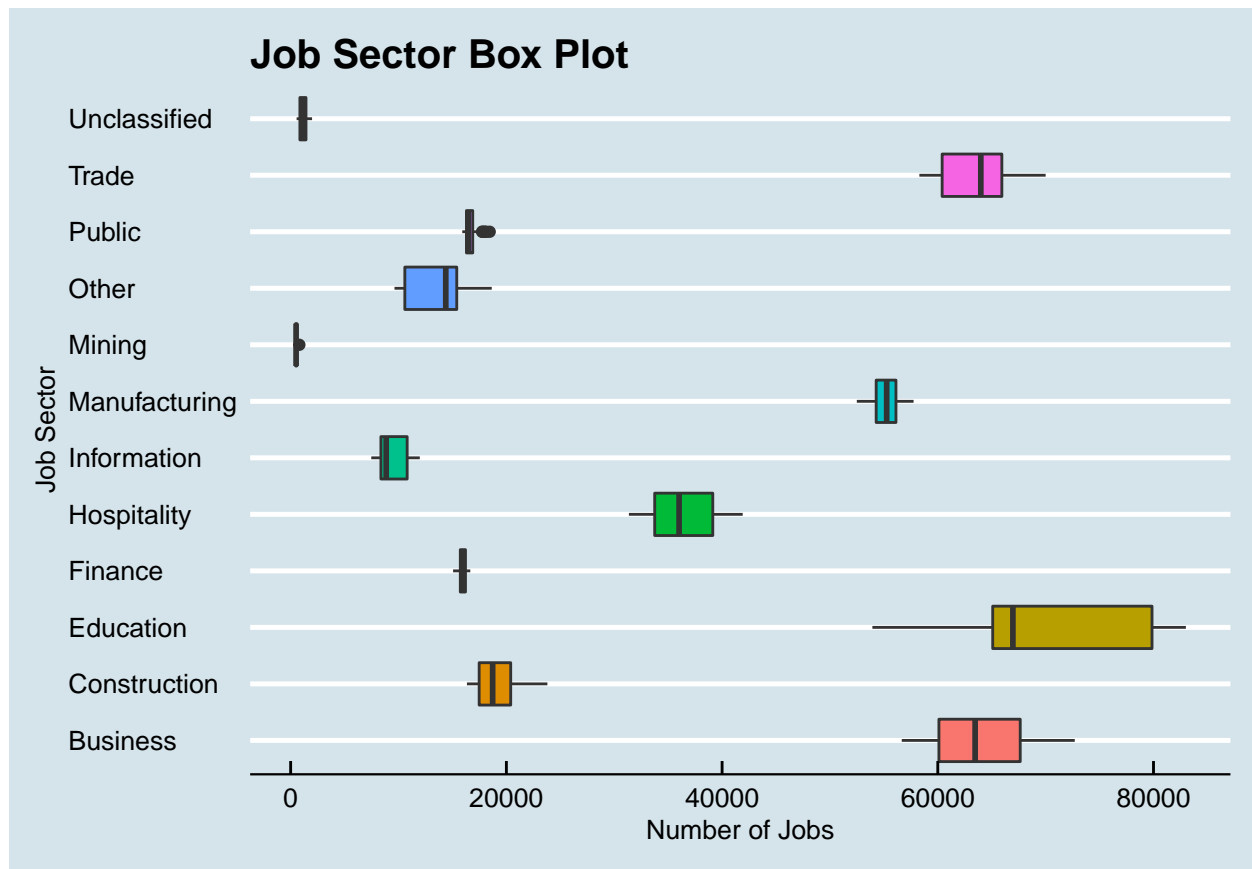


From the box plot above we can see home prices tend to be higher during the summer months of April, May, June, and July then start to come down as the winter months begin. From this data January and February seems to be the best time to buy if you're looking for a new home.

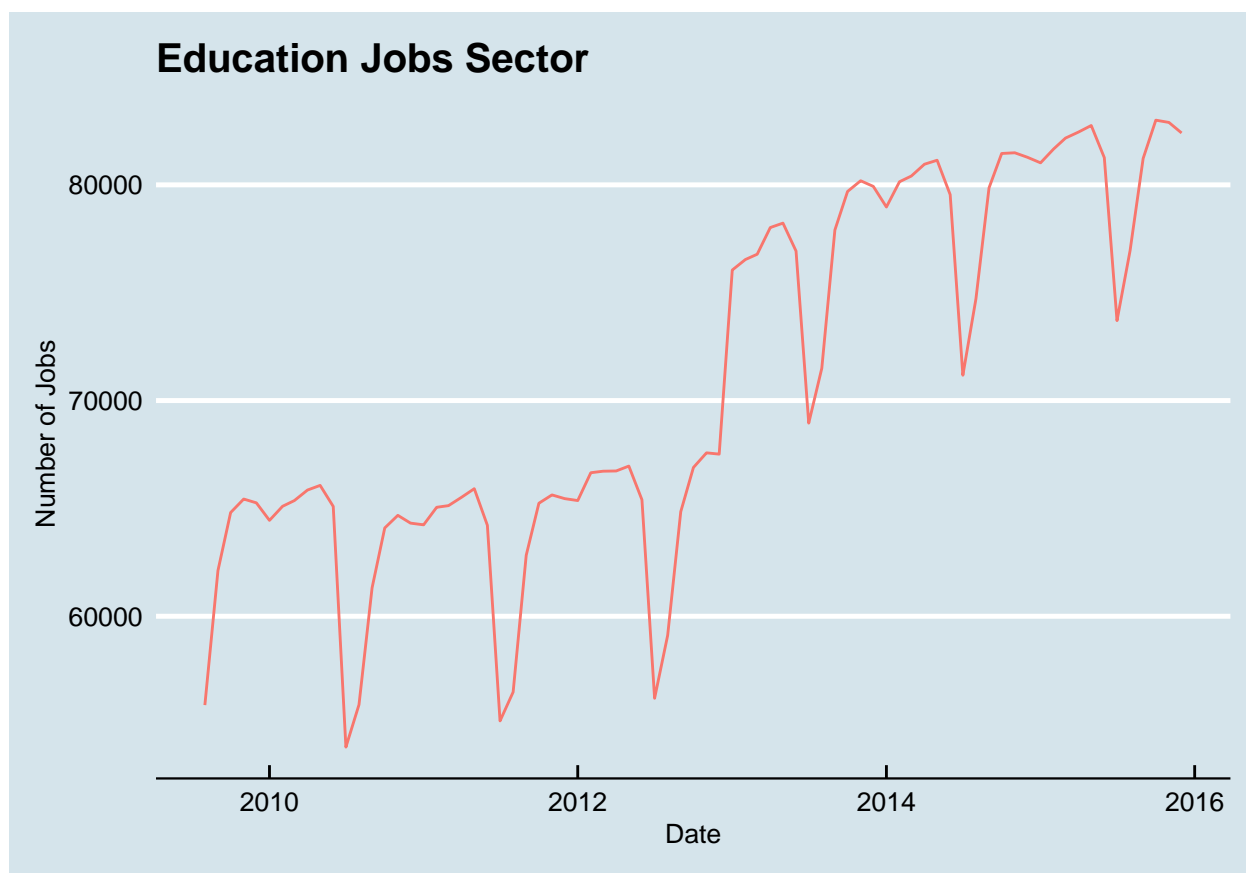




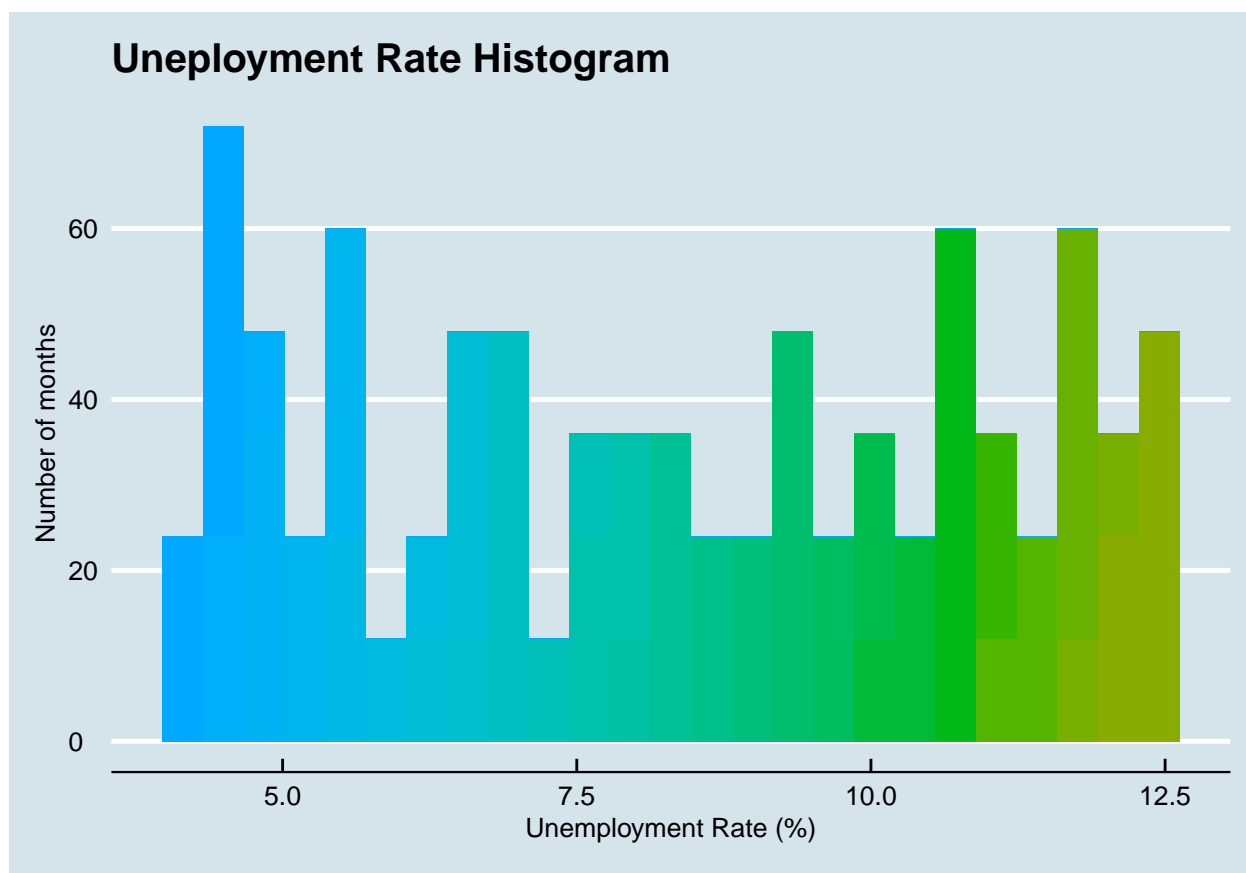
The second set of time series plots shows the total number of jobs by job sector (see README file for long form job sector names). Education, Business, and Trade are the top sectors in terms of number of jobs. The education sector shows high yearly variation most likely due to teachers getting let go for the summer months when students aren't in school.



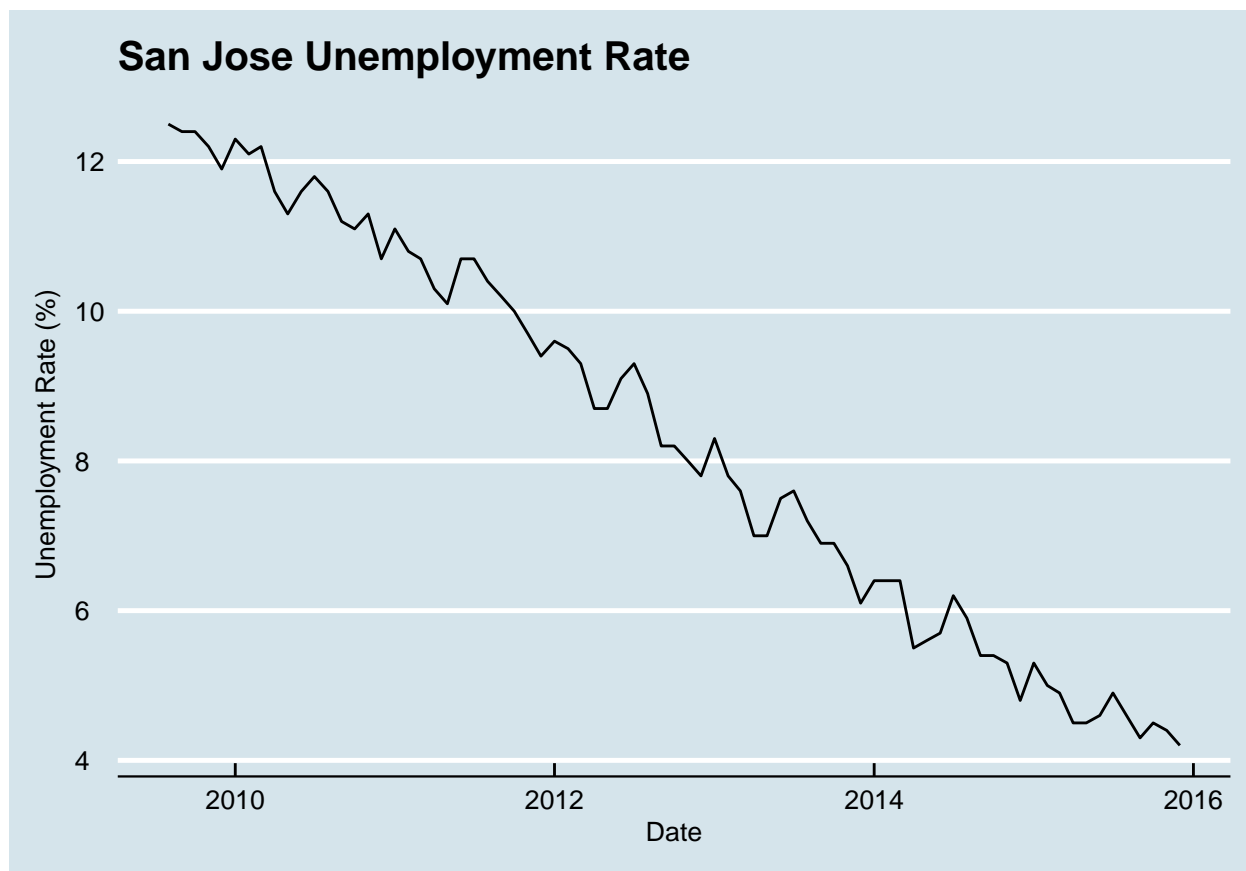
If we look at a box plot the top jobs sectors also have the most variability in the number of jobs. Manufacturing is the fourth place sector but shows a lot less variability when compared to the other top sectors. It's important to consider the data set is mostly in a time of strong economic growth so manufacturing may or may not be more variable through a full business cycle that includes both weak and strong economic growth.



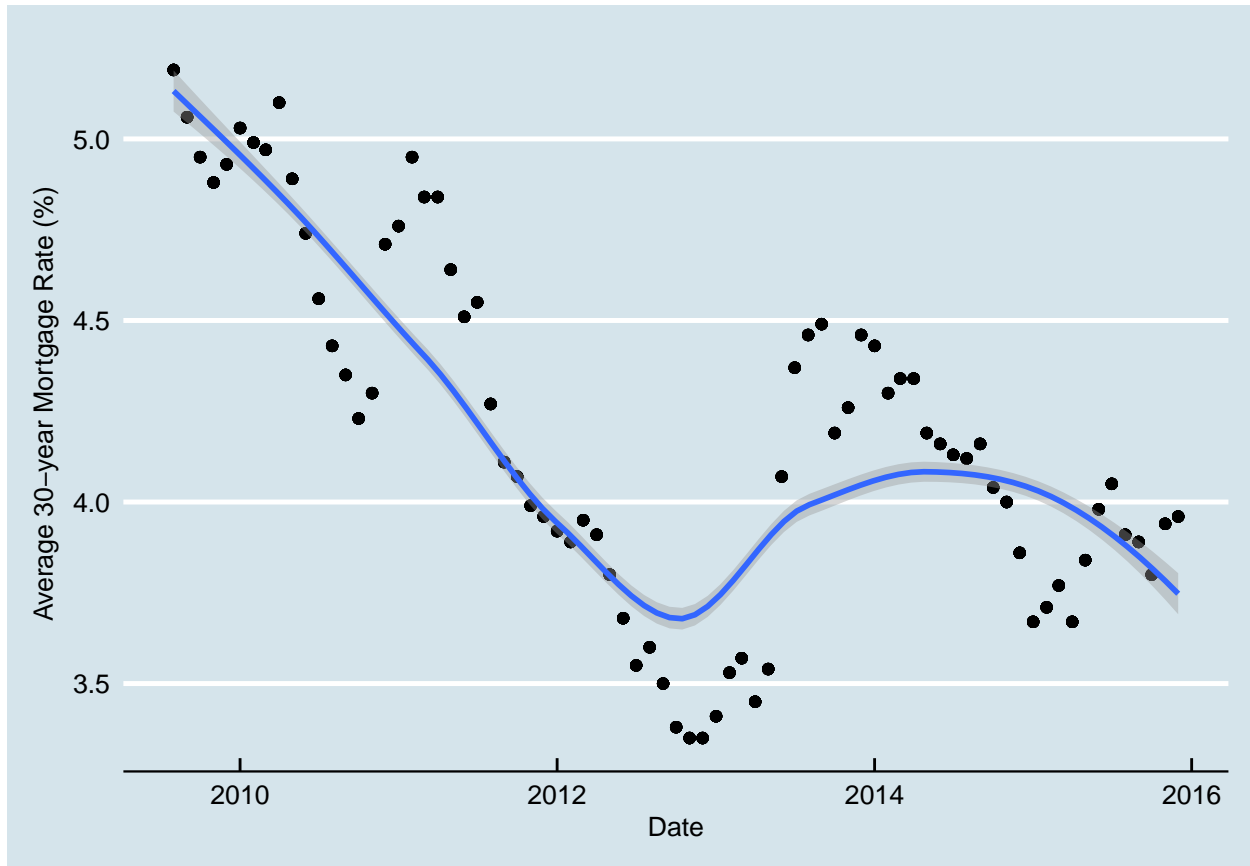
Lastly, I've pulled out the Education sector to show how the number of education jobs drops significantly each year during the summer months as you might expect. It's also interesting how the variation seems to be in the same range from 2009 - 2012 then spikes to a new level in 2013. Perhaps more teachers were hired overall due to a budget increase or the methodology changed for counting education jobs. This sector also includes health services jobs so perhaps that is the reason for the overall increase.



Taking a look at the unemployment rate in our data set I thought a histogram might be interesting to look at. When counting the number of observations for each unemployment rate in a histogram we can see the rate stood the most months at 5.4 and 5.5% and ranged from 4.6 to 12.6%.



The unemployment rate has trended downward over the range of our data set. This would indicate we are looking only at a time when economic conditions were improving and not a full market cycle that would reflect both improving and declining economic conditions.



Lastly, a final time series plotting average 30-year mortgage rates shows a historical fall in rates since 2008 that bottomed in late 2012 and have remained range bound between 3.5 and 4.5% since. This may have helped the rise in average home prices recently as lower rates create a more favorable lending environment for borrowers to purchase homes. Rates started to trend downward due to the financial crises and have remained historically low ever since.

## Model Building

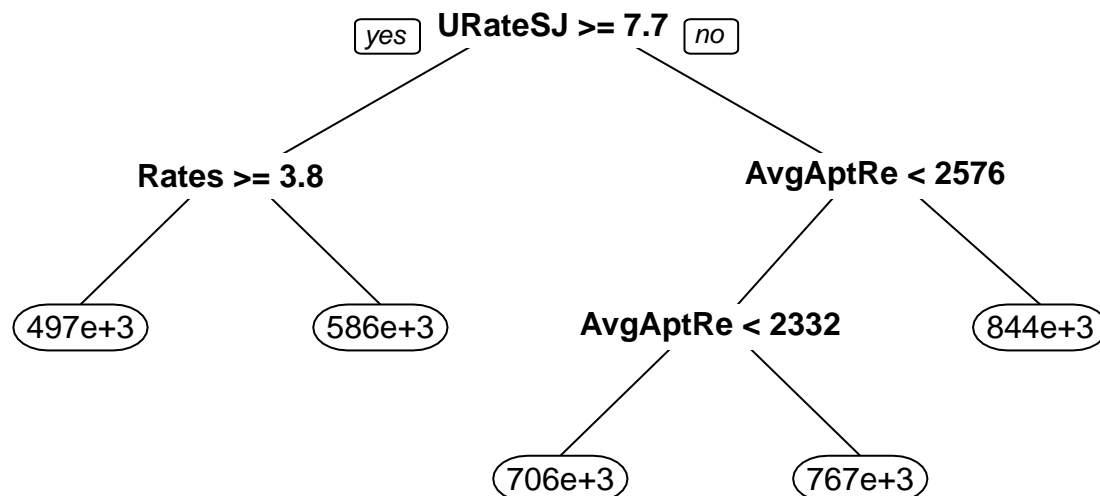
Now that we've taken some time to do some analysis I'm going to look at three different models and determine from test data which one does the best job at predicting average home prices. To compare each model I'll use root-mean-square error (RMSE) results. RMSE measures the square root mean differences between the values our models predict and the actual values from the test data set. The model with the lowest RMSE corresponds to the model with the best predictive accuracy. To compare each model fairly we will use the test data set from the time-series model since we cannot use randomly selected test data for a time-series model.

### Linear Regression

First, I started with a linear regression model. From analyzing our data set and looking at all the variables I chose average apartment rents, unemployment rate, and mortgage rates as the best variables for the model. All variables were highly significant with a high adjusted R-squared of 0.899. To test the model I created a random train and test data set from the original data set. I then made predictions using the test data set to calculate the RMSE using the predicted and real average home prices. For the linear regression model the RMSE was 42673.19. I will compare this result with the tree regression model next.

## Tree Regression

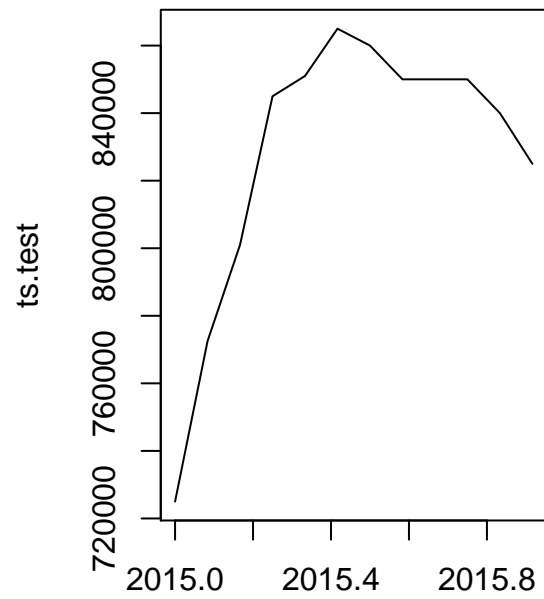
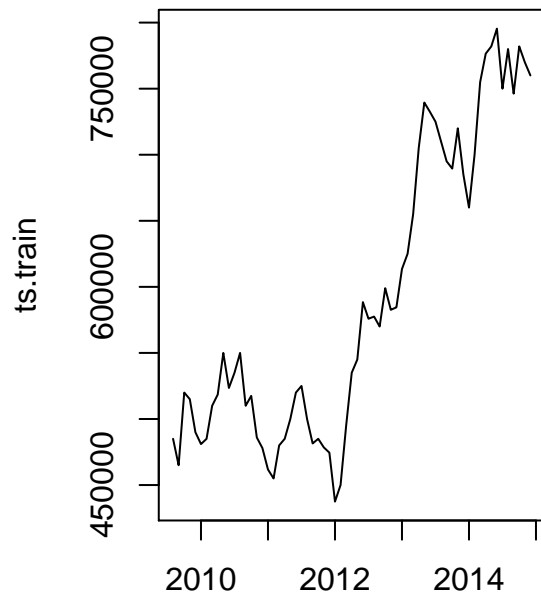
In the first step of the tree regression model I used the same variables as the linear regression model. This got an RMSE of 24843.1, which outperformed the linear regression model. To see if I could make improvements on the tree model I then used cross validation using all the variables in the data set. This improved the RMSE dramatically over both the linear model and the first tree model to 3591.5. The complexity is much greater than the original tree model but the accuracy is significantly higher, this is the best model so far. Finally, I will compare these results with the time series model.



## Time Series

For the time series model the train and test data sets cannot be random since we have to structure our time series in a logical order from past to present. The train data will be the oldest data and the most recent data will be used for the test data set. The time series analysis will be an autoregressive integrated moving average (ARIMA) model. To build the best model we had to be sure to incorporate seasonal trends. As we saw in the exploratory analysis, average home prices in San Jose would rise in the summer and fall in the winter. This seasonal trend had to be taken into consideration when building the time series model and to try and improve the forecasting we assumed the model to be multiplicative. To make it additive we took the log of the time series data sets and built our model from that. This gave us an RMSE of 0.02252769. To compare this result to the linear and tree regression models I also had to compare the predictions of those models to the log of the time series test data set. The results were calculated manually and outlined in the table below. The model with the highest accuracy according to RMSE is the regression tree model followed by the linear regression model and finally the time series model which is the least accurate.

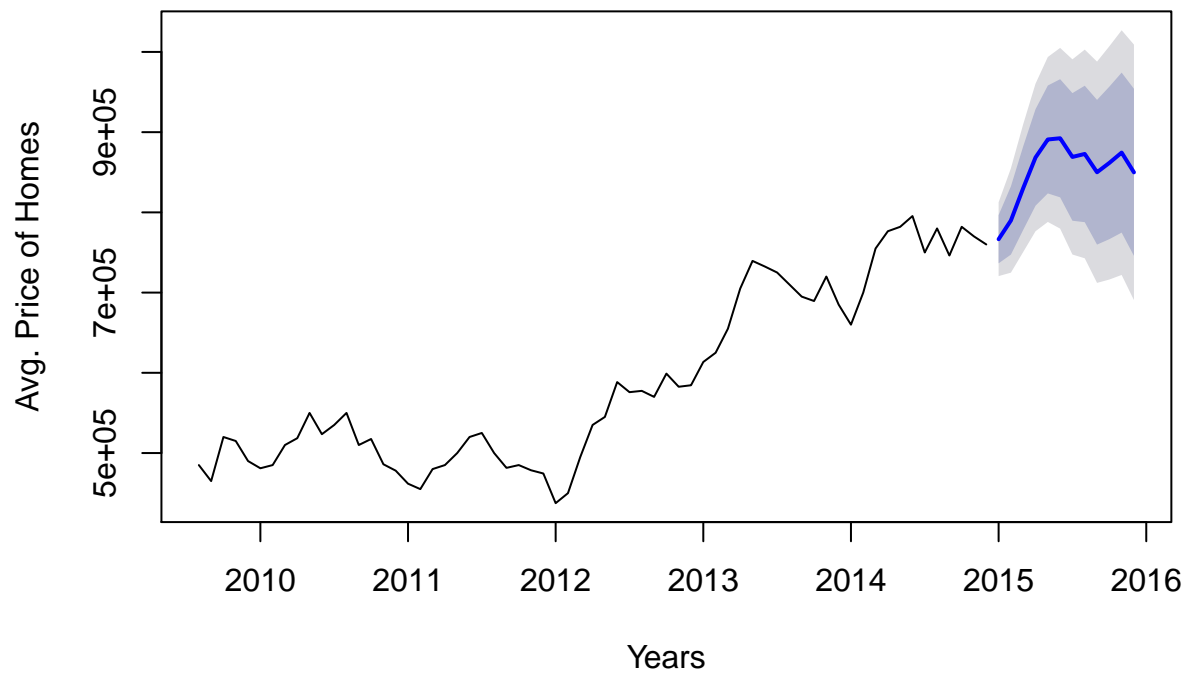
Model	Linear	Tree	Time Series
RMSE	0.012206	0.001140	0.022528



Time

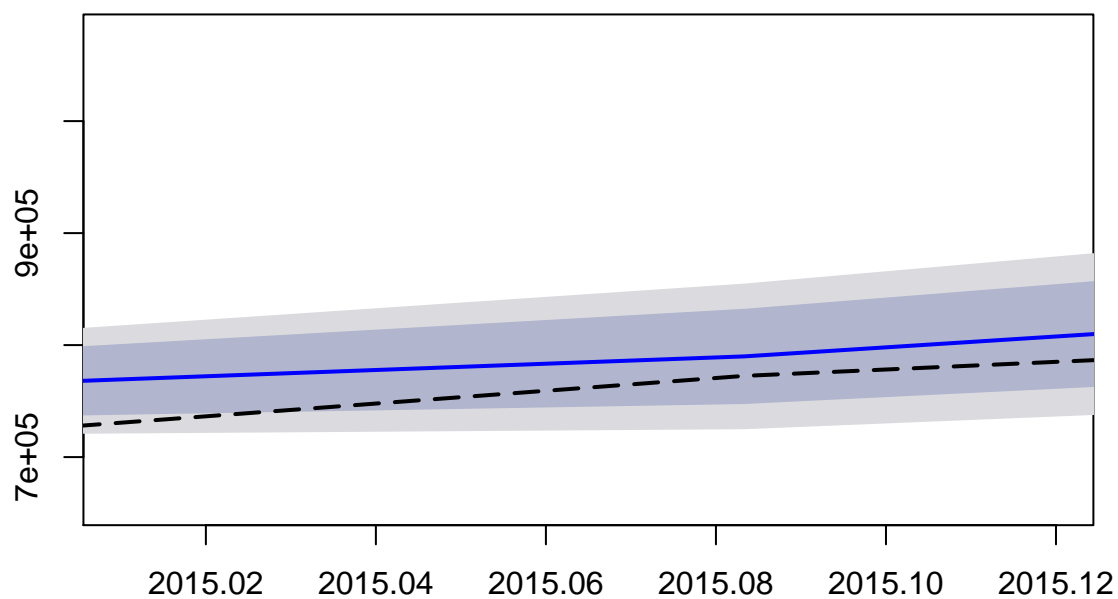
Time

**Forecasts from ARIMA(0,1,0)(1,1,0)[12]**



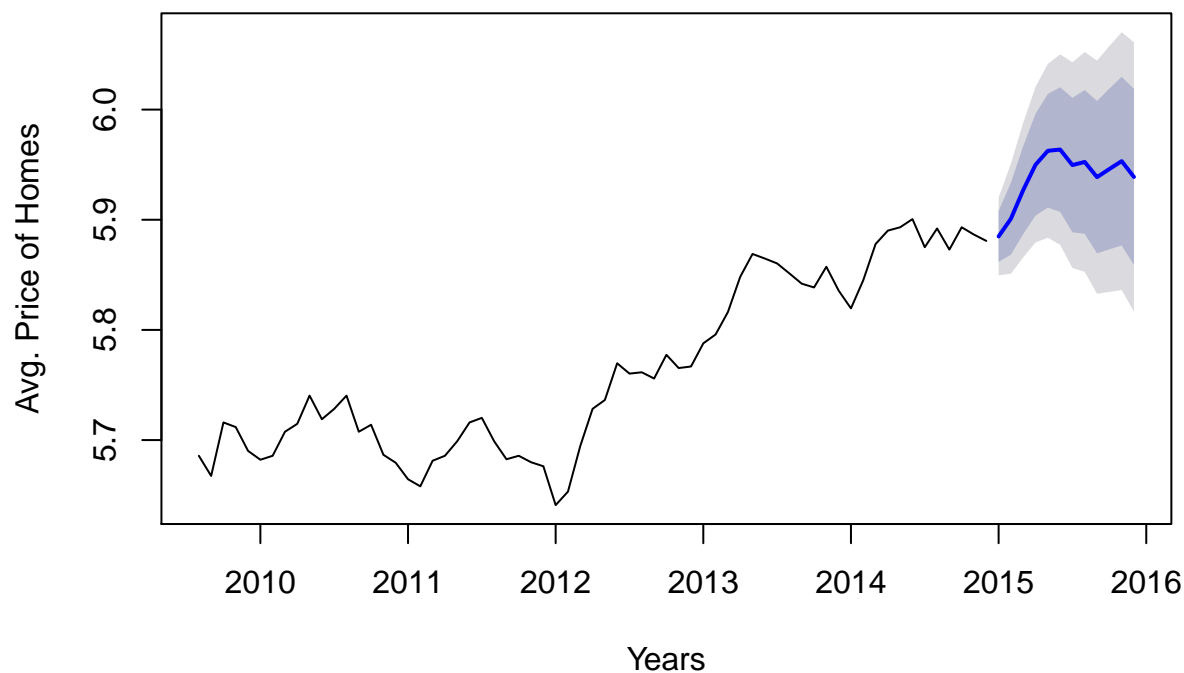


### Forecasts from ARIMA(0,1,0)(1,1,0)[12]

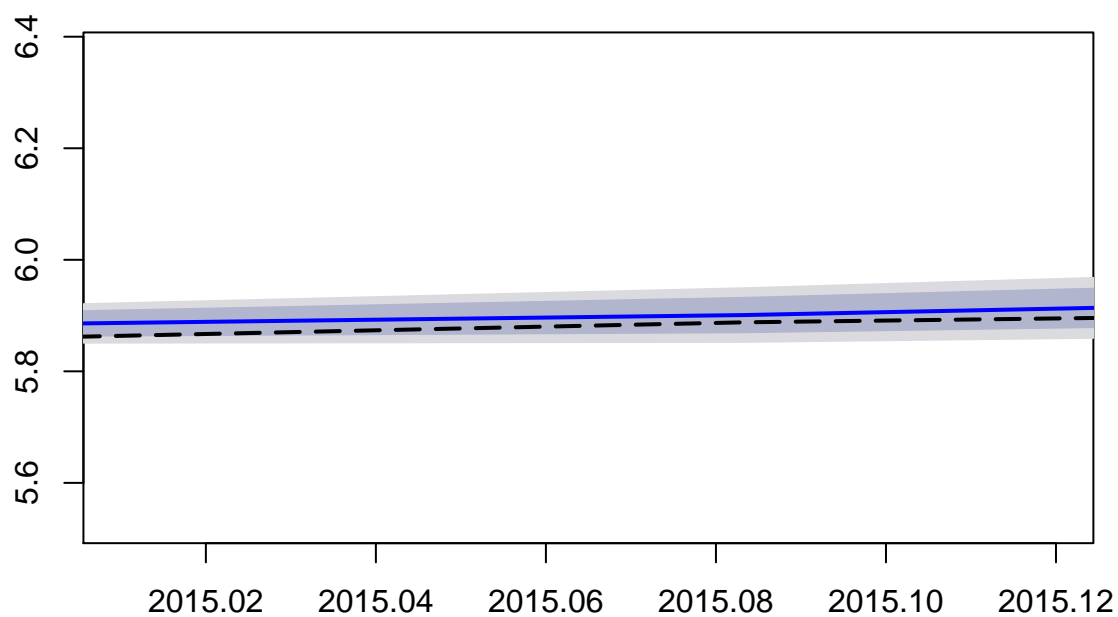


##	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
## Training set	793	20779	15403	0.103	2.57	0.219	-0.1687	NA
## Test set	-23502	26347	23502	-2.877	2.88	0.335	0.0291	1

### Forecasts from ARIMA(0,1,0)(1,1,0)[12]

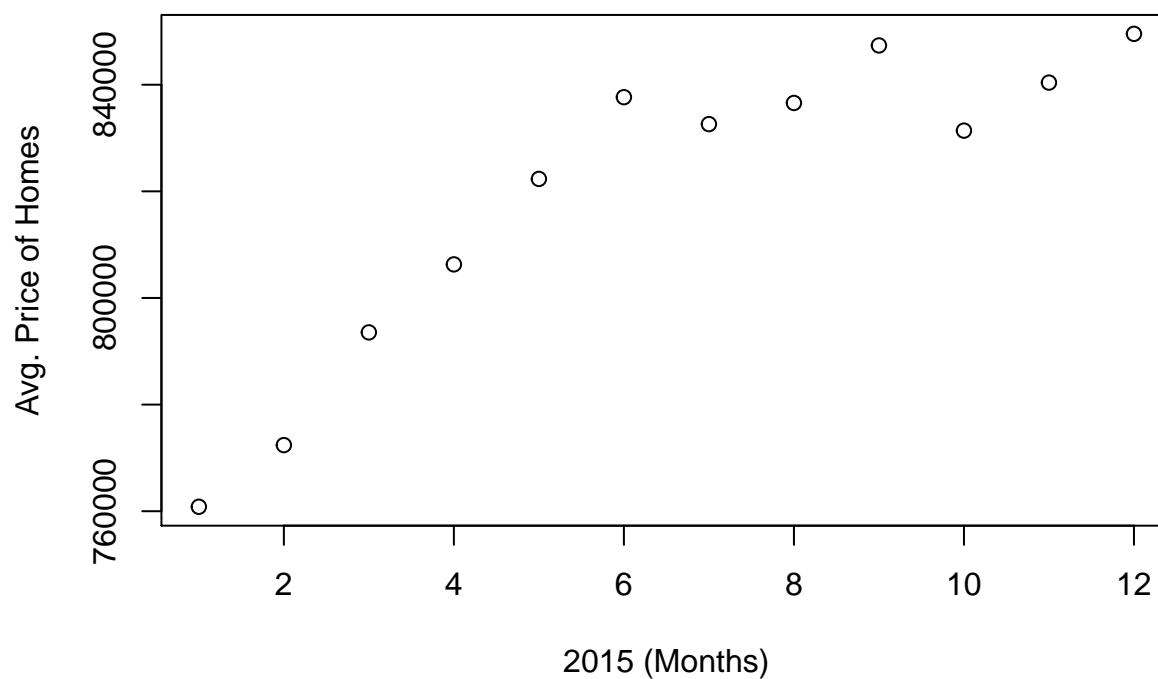


## Forecasts from ARIMA(0,1,0)(1,1,0)[12]

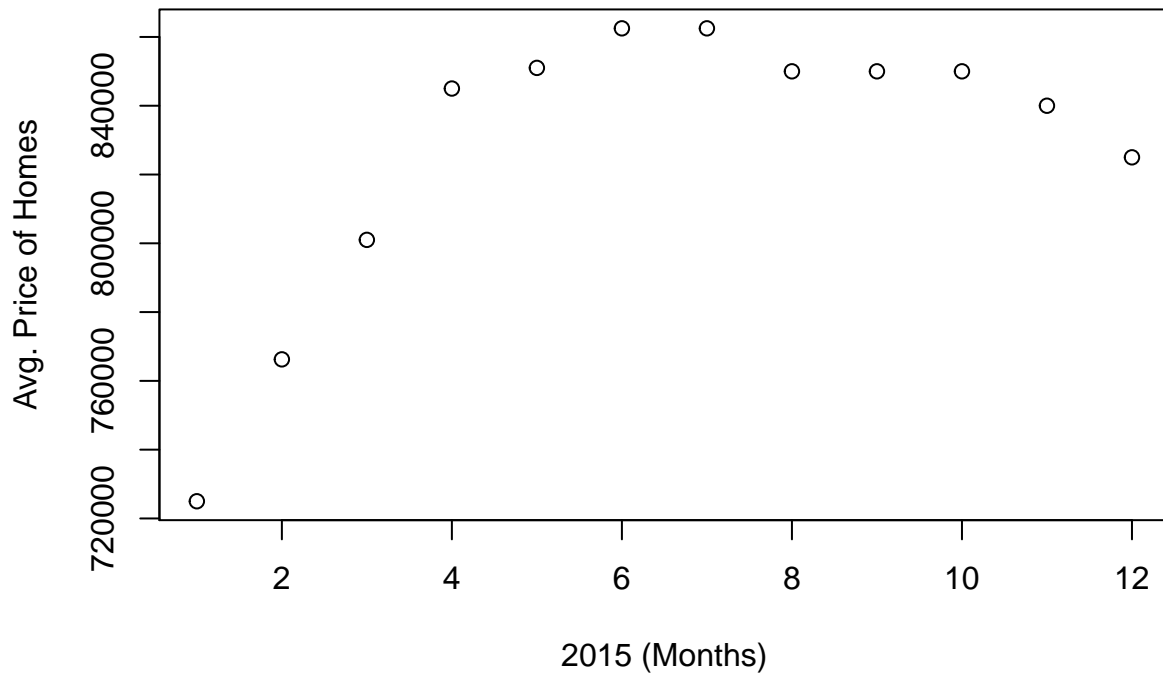


```
##              ME    RMSE    MAE      MPE  MAPE  MASE   ACF1
## Training set -2.09e-06 0.0159 0.0120 -8.42e-05 0.209 0.234 -0.0945
## Test set    -2.16e-02 0.0225 0.0216 -3.64e-01 0.364 0.419  0.0100
##              Theil's U
## Training set      NA
## Test set          1.77
```

## Linear Model Predictions



## Tree Model Predictions



## Results and Discussion

After analyzing the data set and working through building our models to predict average home prices for San Jose there are a few takeaways for our client groups. First, in the near term housing prices will continue to rise. For someone looking to buy a home it may be advantages to wait until prices fall if buying a home isn't a short term necessity. Secondly, when the time comes to purchase a new home it may be more beneficial to buy during the winter months when seasonal trends show prices to be the lowest. This could save buyers a substantial amount of money compared to purchasing a home during the summer months. Lastly, from a government policy perspective, it seems as though current initiatives to create affordable housing for more people aren't having a significant impact on average home prices. However, it is hard to say definitively since this data set is limited and more variables would need to be considered. Further analysis should be done by the city of San Jose to measure the impact of policy initiatives to verify they are working as intended. The subject of affordable housing will continue to be of interest in cities with limited space and growing populations. San Jose is at the heart of Silicon Valley where innovation and wealth creation thrive, making the area a very popular and expensive place to live. Affordable housing will continue to be an issue for the businesses and people who live in San Jose.

## Appendix

### References

1. <https://www.nar.realtor/sites/default/files/reports/2017/embargoes/2017-q1-metro-home-prices/metro-home-prices-q1-2017-single-family-2017-05-15.pdf>
2. <http://yesonaffordablehousing.org/>
3. <http://www.sanjoseca.gov/index.aspx?NID=5256>