MSA 8200 Predictive Analytics

Final Project Report

Zillow Housing Data

Shreyashi Mukhopadhyay

Shalmali Joshi

Namit Srivastava

Nhi Nguyen

Introduction:

Zillow houses a portfolio of the largest and most vibrant real estate and home-related brands on the web and mobile.Zillow.com is the most popular online real-estate website and mobile app in the United States. Zillow provides updated home information to tens of millions of buyers and sellers, Real Estate Agents, and Financial Institutions among other affiliates every day.

Zillow focuses on all stages of the housing lifecycle including renting, buying, selling, financing and home improvement. Zillow and its affiliates offer customers an on-demand experience for selling, buying, renting, and financing with transparency and seamless end-to-end service.

As the most-visited real estate website in the United States, the Zillow Group empowers its customer base with unparalleled data around homes and connects them with the right local professionals to help. A primary feature of the Zillow website is the **Zestimate**—a home-valuation tool that provides buyers and sellers with the estimated market value for a specific home. Zillow currently offers Zestimate for more than one hundred million homes in the U.S., with hundreds of attributes for each property.

Project Description:

Our goal is to analyze the Zillow Economics Data and identify the patterns in the housing market especially highlighting the most expensive and the least expensive housing markets in the US.

This project encompasses a complete analysis of the Zillow Economic data through data visualization, data exploration, pattern discovery and sales forecasting through time series modelling in R as the programming language.

Data source:

The data has been sourced from the Zillow Economic dataset on Kaggle.com at https://www.kaggle.com/datasets/zillow/zecon?select=State_time_series.csv

Additional Research data: https://www.zillow.com/research/data/

We have used the **State_time_series.csv** dataset from the Zillow Economic dataset at Kaggle for our analysis.

This dataset has been constructed with housing data from 1996 up until the last update to this data in 2017. The dataset has been created by compiling and organizing public data and Zillow's own proprietary data.

Data description:

The dataset contains: 13212 obs. of 86 variables. However, since the data is available for only for 82 variables, our final dataset description is as follows:

```
[1] "The data set has 13212 rows and 82 columns"
```

Null "NA" Values:

There are **633838** null values in total across all the variables.

The missing values are attributed to the missing data between 1996 – 2010 for majority of the variables. The data from 2010 – 2017 is available for all the 82 variables.

.

```
## Counting Null Values
```{r}

print(sum(is.na(state)))

```
```

```
[1] 633838
```

The variables in the dataset are described as under:

[1] "Date"

[2] "Region Name"

[3] "DaysOnZillow_AllHomes"

[4] "HomesSoldAsForeclosuresRatio_AllHomes"

[5] "InventorySeasonallyAdjusted_AllHomes"

[6] "InventoryRaw_AllHomes"

[7] "MedianListingPricePerSqft_1Bedroom"

[8] “MedianListingPricePerSqft_2Bedroom”

[9] “MedianListingPricePerSqft_3Bedroom”

[10] “MedianListingPricePerSqft_4Bedroom”

[11] “MedianListingPricePerSqft_5BedroomOrMore”

[12] “MedianListingPricePerSqft_AllHomes”

[13] “MedianListingPricePerSqft_CondoCoop”

[14] “MedianListingPricePerSqft_DuplexTriplex”

[15] “MedianListingPricePerSqft_SingleFamilyResidence”

[16] “MedianListingPrice_1Bedroom”

[17] “MedianListingPrice_2Bedroom”

[18] “MedianListingPrice_3Bedroom”

[19] “MedianListingPrice_4Bedroom”

[20] “MedianListingPrice_5BedroomOrMore”

[21] “MedianListingPrice_AllHomes”

[22] “MedianListingPrice_CondoCoop”

[23] “MedianListingPrice_DuplexTriplex”

[24] “MedianListingPrice_SingleFamilyResidence”

[25] “MedianPctOfPriceReduction_AllHomes”

[26] “MedianPctOfPriceReduction_CondoCoop”

[27] “MedianPctOfPriceReduction_SingleFamilyResidence”

[28] “MedianPriceCutDollar_AllHomes”

[29] “MedianPriceCutDollar_CondoCoop”

[30] “MedianPriceCutDollar_SingleFamilyResidence”

[31] “MedianRentalPricePerSqft_1Bedroom”

[32] “MedianRentalPricePerSqft_2Bedroom”

[33] “MedianRentalPricePerSqft_3Bedroom”

[34] “MedianRentalPricePerSqft_4Bedroom”

[35] “MedianRentalPricePerSqft_5BedroomOrMore”

[36] “MedianRentalPricePerSqft_AllHomes”

[37] “MedianRentalPricePerSqft_CondoCoop”

[38] “MedianRentalPricePerSqft_DuplexTriplex”

[39] “MedianRentalPricePerSqft_MultiFamilyResidence5PlusUnits”

[40] “MedianRentalPricePerSqft_SingleFamilyResidence”

[41] “MedianRentalPricePerSqft_Studio”

[42] “MedianRentalPrice_1Bedroom”

[43] “MedianRentalPrice_2Bedroom”

[44] “MedianRentalPrice_3Bedroom”

[45] “MedianRentalPrice_4Bedroom”

[46] “MedianRentalPrice_5BedroomOrMore”

[47] “MedianRentalPrice_AllHomes”

[48] “MedianRentalPrice_CondoCoop”

[49] “MedianRentalPrice_DuplexTriplex”

[50] “MedianRentalPrice_MultiFamilyResidence5PlusUnits”

[51] “MedianRentalPrice_SingleFamilyResidence”

[52] “MedianRentalPrice_Studio”

[53] “MedianSoldPricePerSqft_AllHomes”

[54] “MedianSoldPricePerSqft_CondoCoop”

[55] “MedianSoldPricePerSqft_SingleFamilyResidence”

[56] “MedianSoldPrice_AllHomes”

[57] “ZHVIPerSqft_AllHomes”

[58] “PctOfHomesDecreasingInValues_AllHomes”

[59] “PctOfHomesIncreasingInValues_AllHomes”

[60] “PctOfHomesSellingForGain_AllHomes”

[61] “PctOfHomesSellingForLoss_AllHomes”

[62] “PctOfListingsWithPriceReductionsSeasAdj_AllHomes”

[63] “PctOfListingsWithPriceReductionsSeasAdj_CondoCoop”

[64] “PctOfListingsWithPriceReductionsSeasAdj_SingleFamilyResidence”

[65] “PctOfListingsWithPriceReductions_AllHomes”

[66] “PctOfListingsWithPriceReductions_CondoCoop”

[67] “PctOfListingsWithPriceReductions_SingleFamilyResidence”

[68] “PctTransactionsThatArePreviouslyForeclosuredHomes_AllHomes”

[69] “PriceToRentRatio_AllHomes”

[70] “Turnover_AllHomes”

[71] “ZHVI_1bedroom”

[72] “ZHVI_2bedroom”

[73] “ZHVI_3bedroom”

[74] “ZHVI_4bedroom”

[75] “ZHVI_5BedroomOrMore”

[76] “ZHVI_AllHomes”

[77] “ZHVI_BottomTier”

[78] “ZHVI_CondoCoop”

[79] “ZHVI_MiddleTier”

[80] “ZHVI_SingleFamilyResidence”

[81] “ZHVI_TopTier”

[82] “ZRI_AllHomes”

[83] “ZRI_AllHomesPlusMultifamily”

[84] “ZriPerSqft_AllHomes”

[85] “Zri_MultiFamilyResidenceRental”

[86] “Zri_SingleFamilyResidenceRental

Zillow Metrics:

Zillow Home Value Index ZHVI: Measures monthly changes in Zestimate and captures both the level and appreciation of home values across a wide variety of geographies and housing types.

Zillow Rent Index ZRI: Adjusted measure of the median rent across a region and house type.

Rent Listing Metrics: Median rents and prices for homes based on region and type and bedroom count.

Inventory Metrics: On-site data from Zillow which includes list prices, inventory, price cuts, and how long a post remained on Zillow before sale.

Homes Listing Metric: Median Listing prices for Data on the sell price of homes based on housing type.

Housing type definitions: Zillow classifies Homes with a county record as:

Studios

Condos

1-Bedroom

2-Bedroom

3- Bedroom

4- Bedroom

5 or more Bedroom

Single-family Home

Multi-family Home

Duplex/Triplex

Tiers: By Metro: Zillow determines the price tier cutoffs that divide all homes into thirds using the full distribution of estimated home values. The resultant categories defined by these cutoffs as Bottom, Middle, and Top Tiers.

Data Pre-processing:

→ We first read the data file state_time_series.csv and convert it into the dataframe state.

→ We then convert the date column into date format using:

```
## Convert the date column into date format

state$Date<-as.Date(state$Date)
```

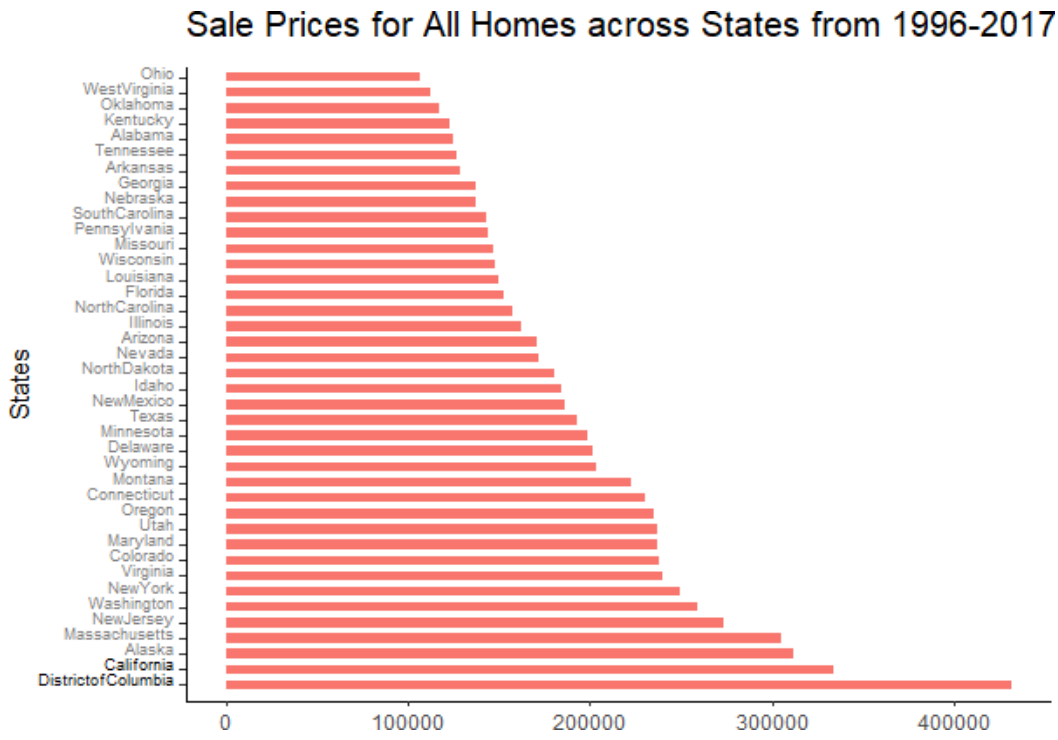
→ We then create a new data column called Year and store the value of Year from Date.

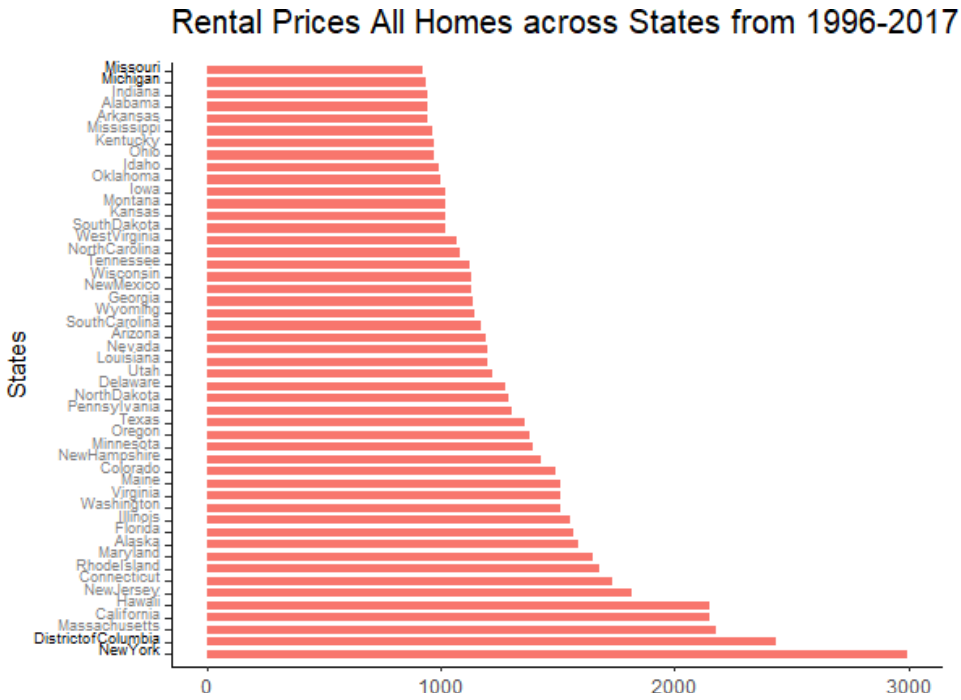
```
## Creating the column Year from the Date column

state['Year'] = year(state$Date)
```

→ We are not dropping the NA values from the data since it results in the loss of the data present in the other variables within the same row of the dataframe.

Visualizing the Data via Exploratory Data Analysis:



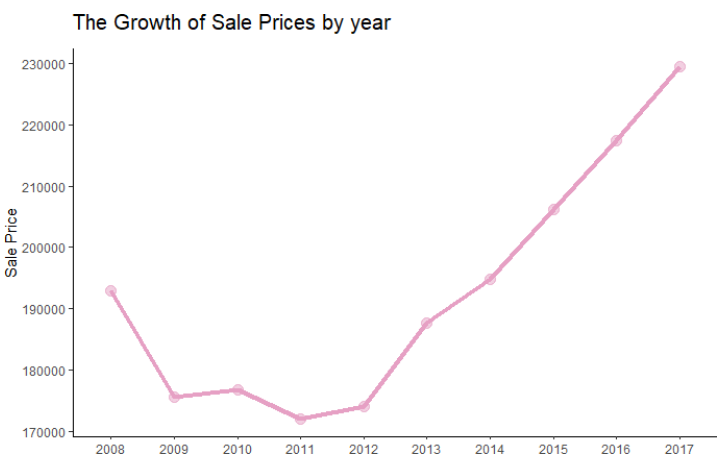


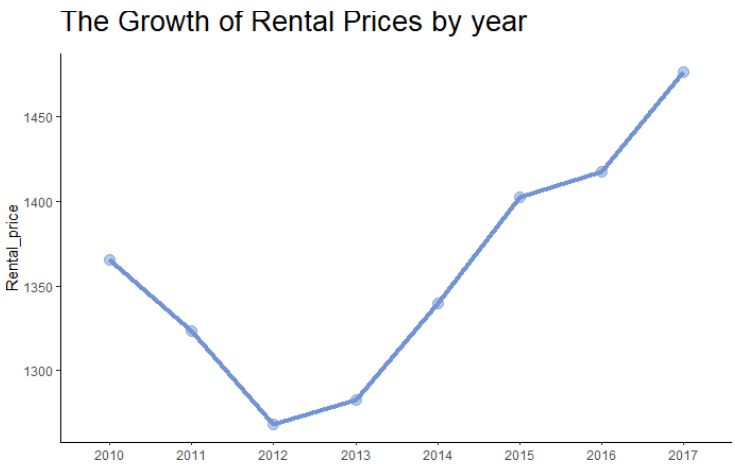
→ The plots of the overall Sale Prices and Rental Prices for All Homes across all states in US from 1996 – 2017 shows that:

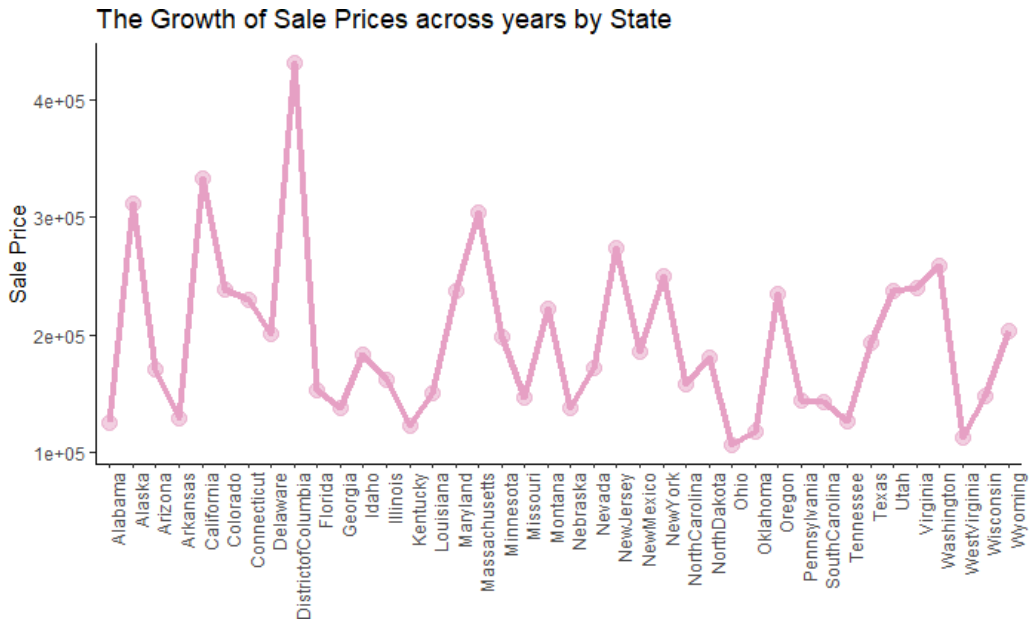
The District of Columbia, California, Alaska, New Jersey, Massachusetts are the most expensive states for buying a Home.

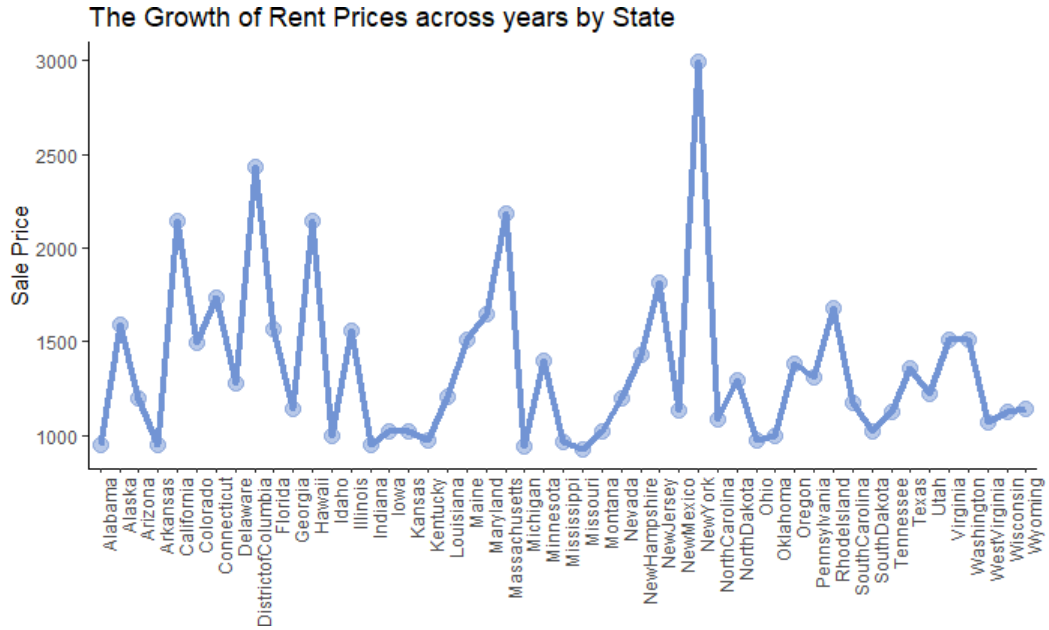
New York, District of Columbia, Massachusetts, California, and Hawaii are the most expensive states for renting a Home.

Growth in the Sale Prices and Rental Prices across Years and State:





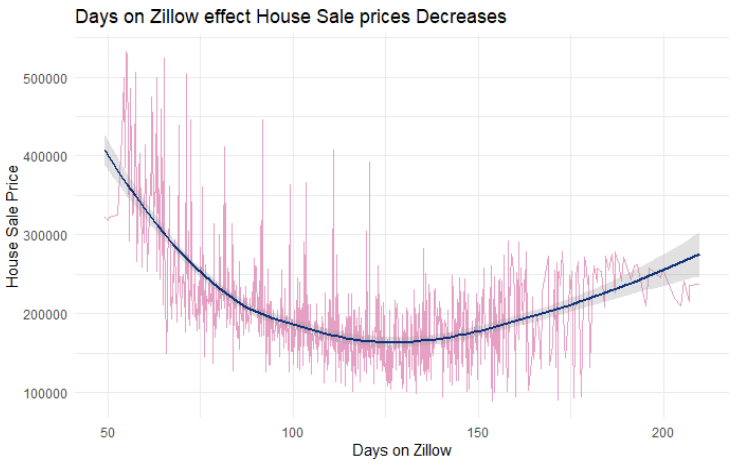


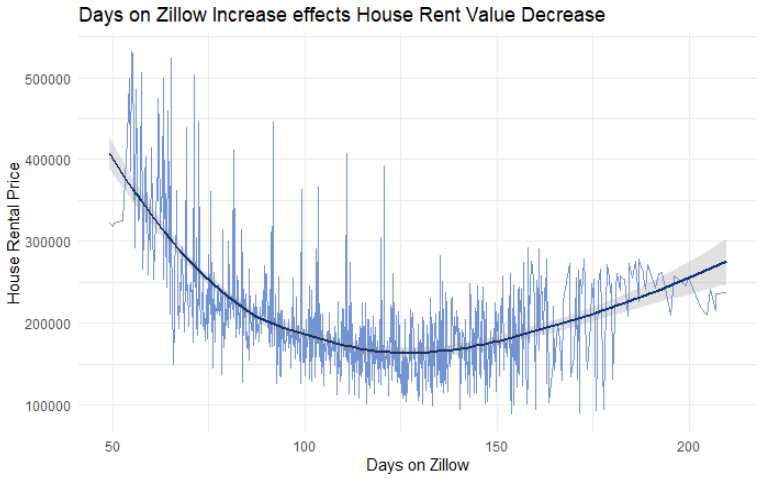


→ The plots of the Growth in the Sale Prices and Rental Prices across Years and State show that there is a strong upward trend in the growth of the Rental and Sale prices from 2012 onwards.

→ The plots of the Growth in the Sale Prices and Rental Prices by State show that the District of Columbia has shown the highest growth in the Buying market whereas New York has shown the highest growth in the Renters market.

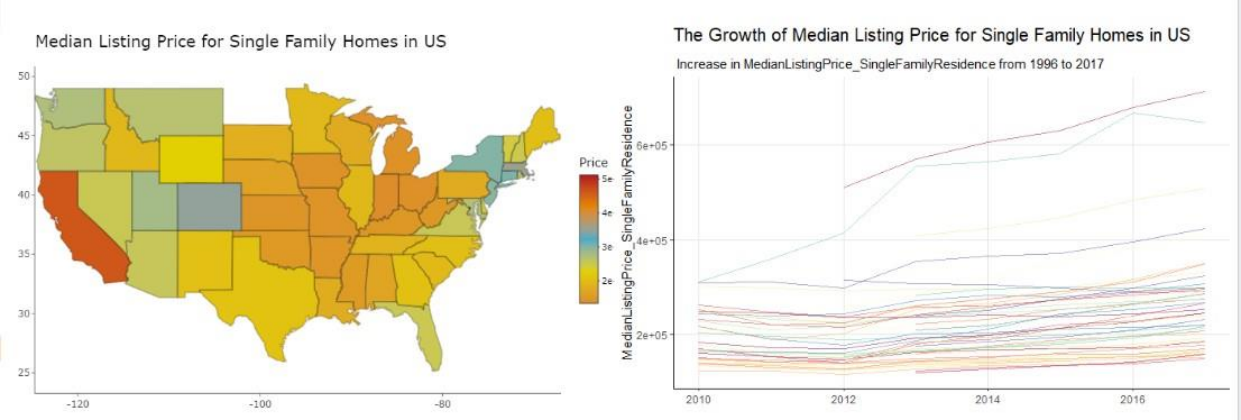
Days on Zillow effect on sale price and rent price





→ The plots of the Days on Zillow show that, the longer the listing stays on the market, the value of the property declines with time and after hitting a substantial low it again picks up momentum.

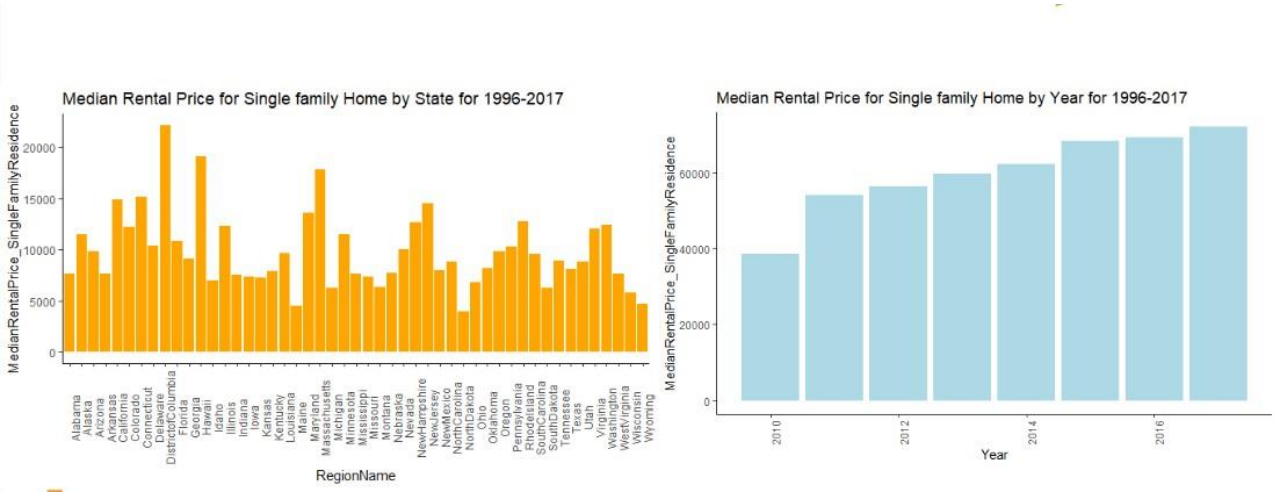
Median Listing price for Single Family Home across years in US

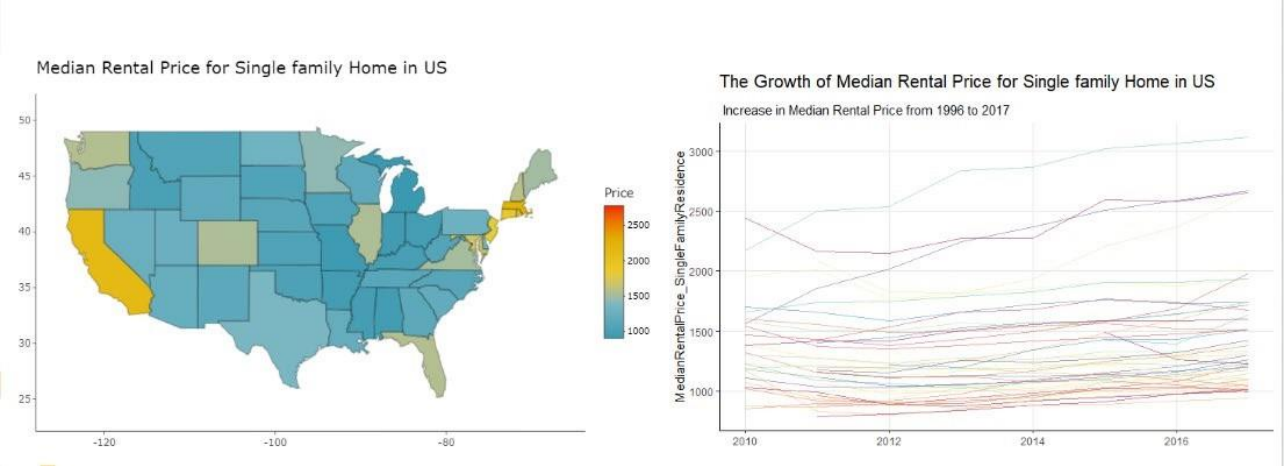


→ The Plots of the Median Listing price for single family homes in US shows us that:

- California, District of Columbia, Hawaii, and Massachusetts show the highest median listing price for single family homes which range from $500,000 and above.
- The median listing price for Single family homes has shown significant growth in Listing prices from 2012 onward.

Median Rental price for Single Family Home across states and years

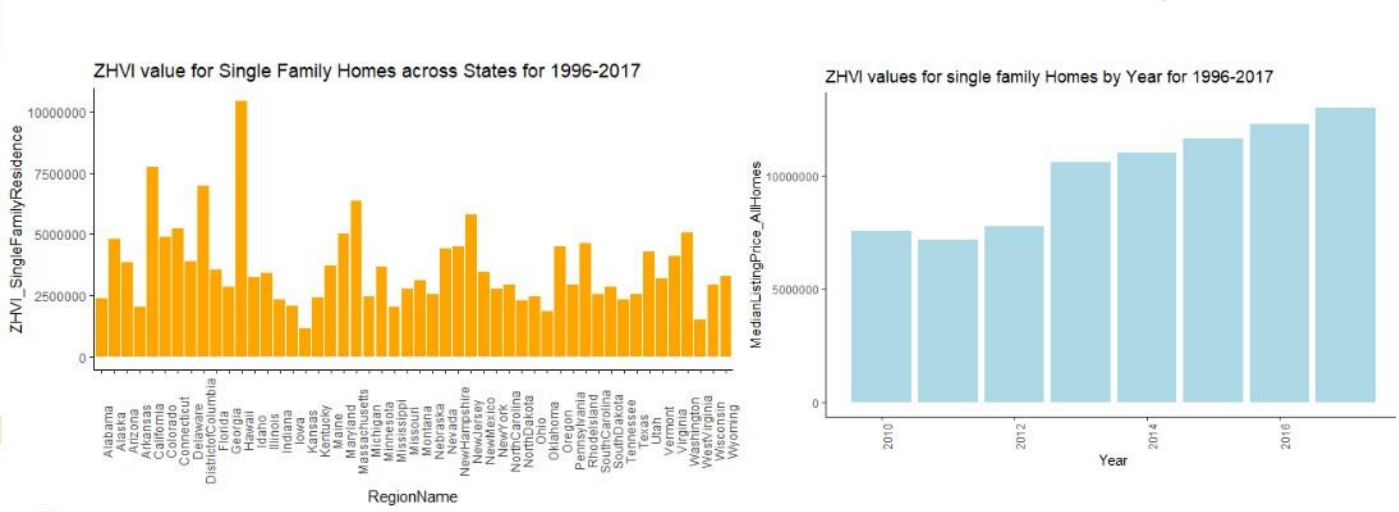


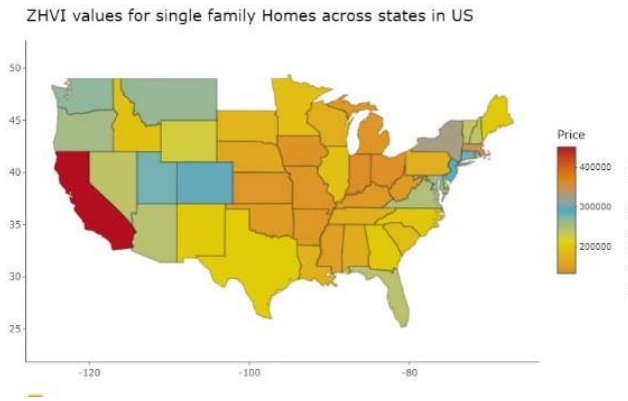


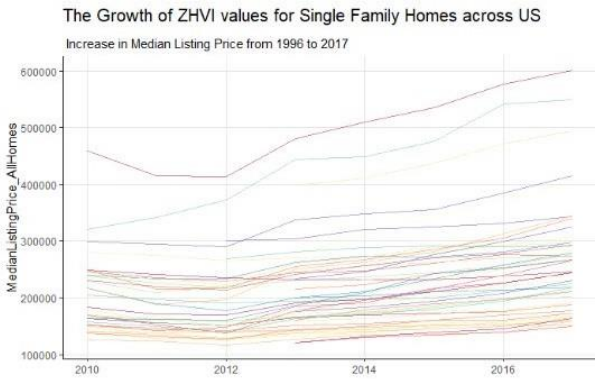
→ The Plots of the Median Rental price for single family homes in US shows us that:

- District of Columbia, Hawaii, Massachusetts, and California, show the highest median listing price for single family homes which range from $2000 and above.
- The median Rental price for Single family homes across US has shown steady growth in listing prices from 2010 onwards.

ZHVI values for Single Family Home across states and years



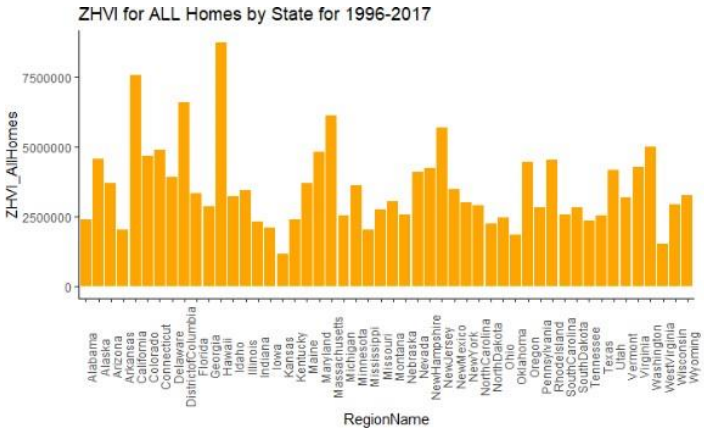


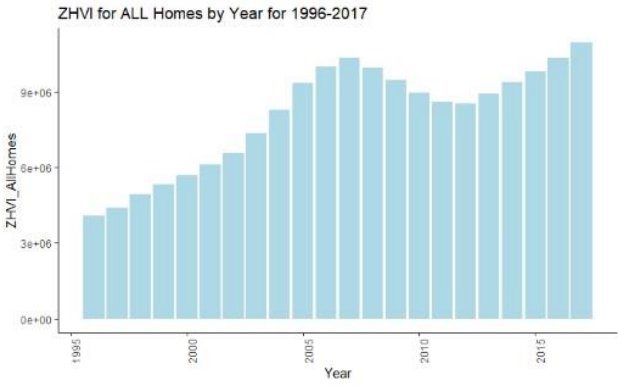


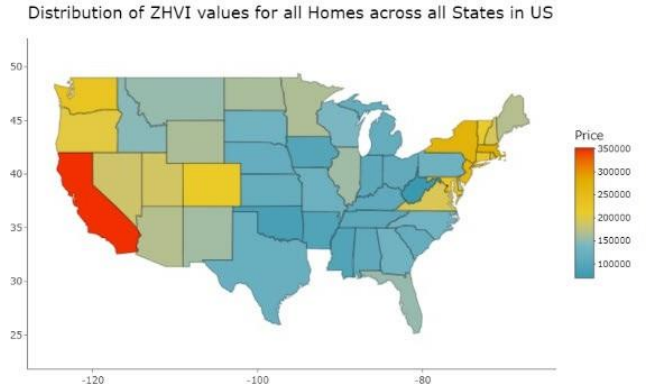
→ The Plots of the ZHVI values for listing price of single-family homes in US show us that:

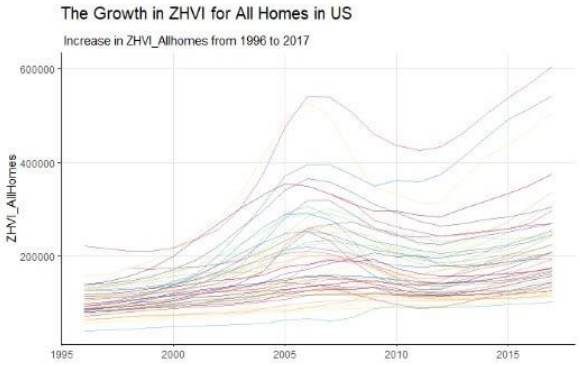
- District of Columbia, Hawaii, Massachusetts, and California, show the highest ZHVI values of listing price which range from $400,000 and above.
- The ZHVI values for listing price of Single-family Homes across US has shown steady increase in the growth of Listing prices from 2010 onwards.

ZHVI values for All Homes across states and years





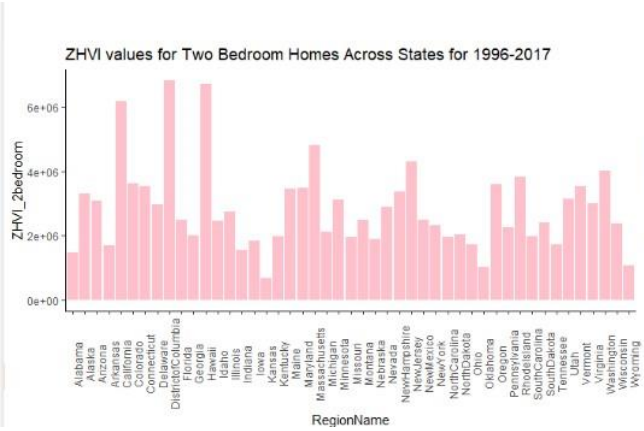


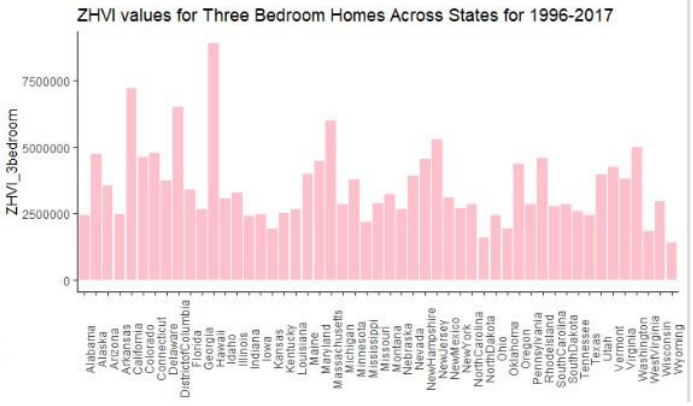


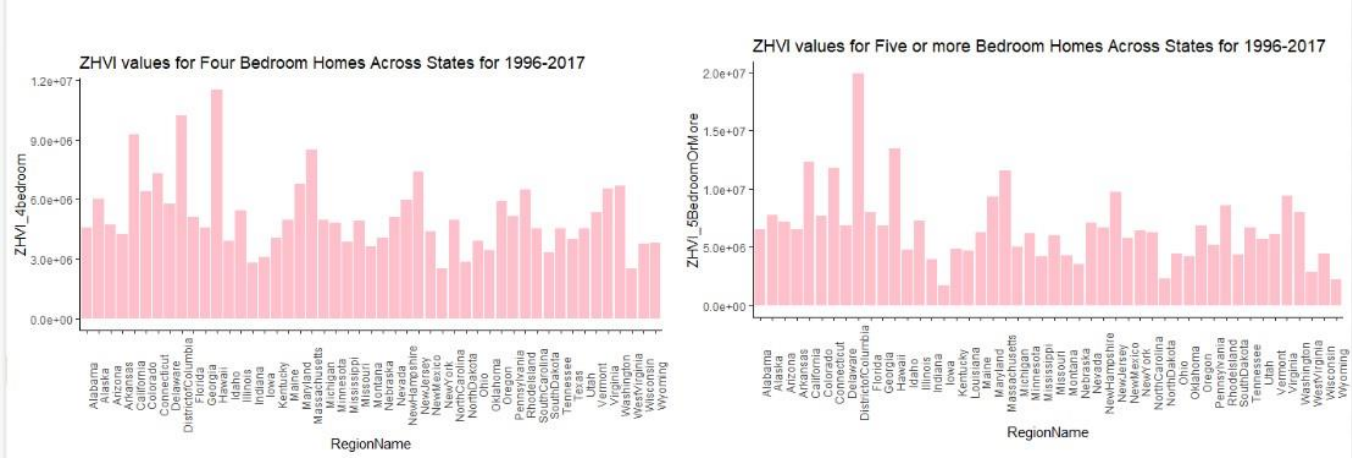
→ The Plots of the ZHVI values for listing price of All homes in US shows us that:

- District of Columbia, Hawaii, Massachusetts, and California, show the highest ZHVI listing prices for All homes which range from $350,000 and above.
- The ZHVI for listing price for All Homes across US has shown significant increase in the growth of Listing prices from 2005 onwards.

ZHVI values of 2 Bedroom and 3 Bedroom Homes across years in US



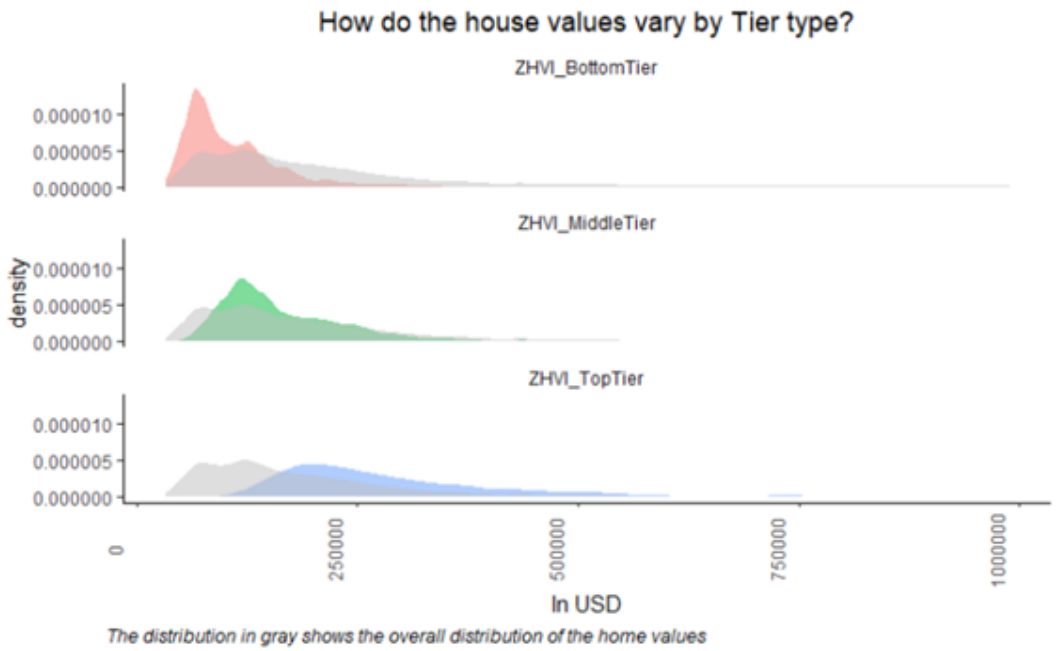


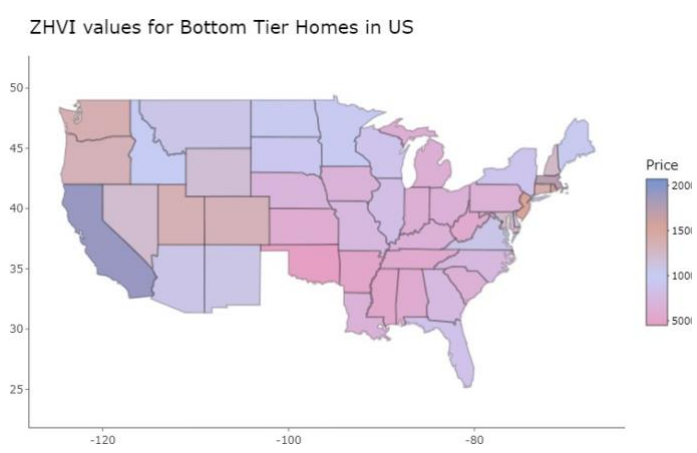


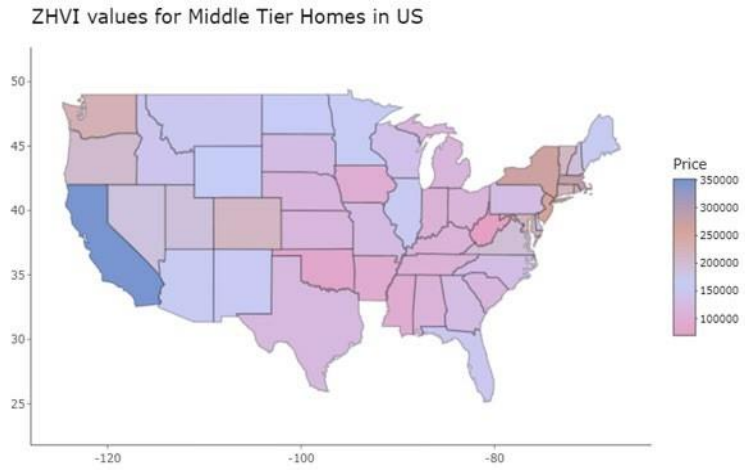
The Plots of the ZHVI values for listing price of 2 Bedroom, 3 Bedroom, 4 Bedroom and 5 Bedrooms in US shows us that:

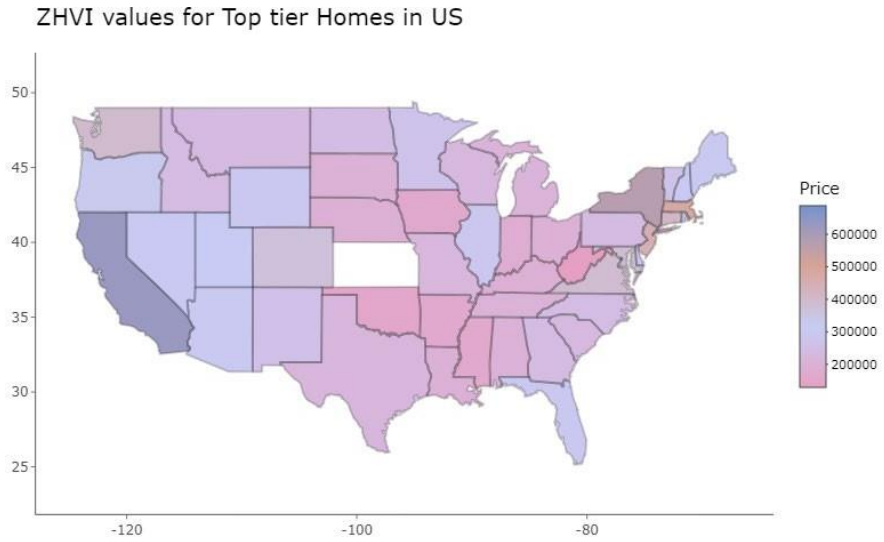
- The states of Connecticut, District of Columbia, Hawaii, Massachusetts, and California, show the highest ZHVI listing prices from 1996-2017.

Tier wise analysis of All Homes across years in US





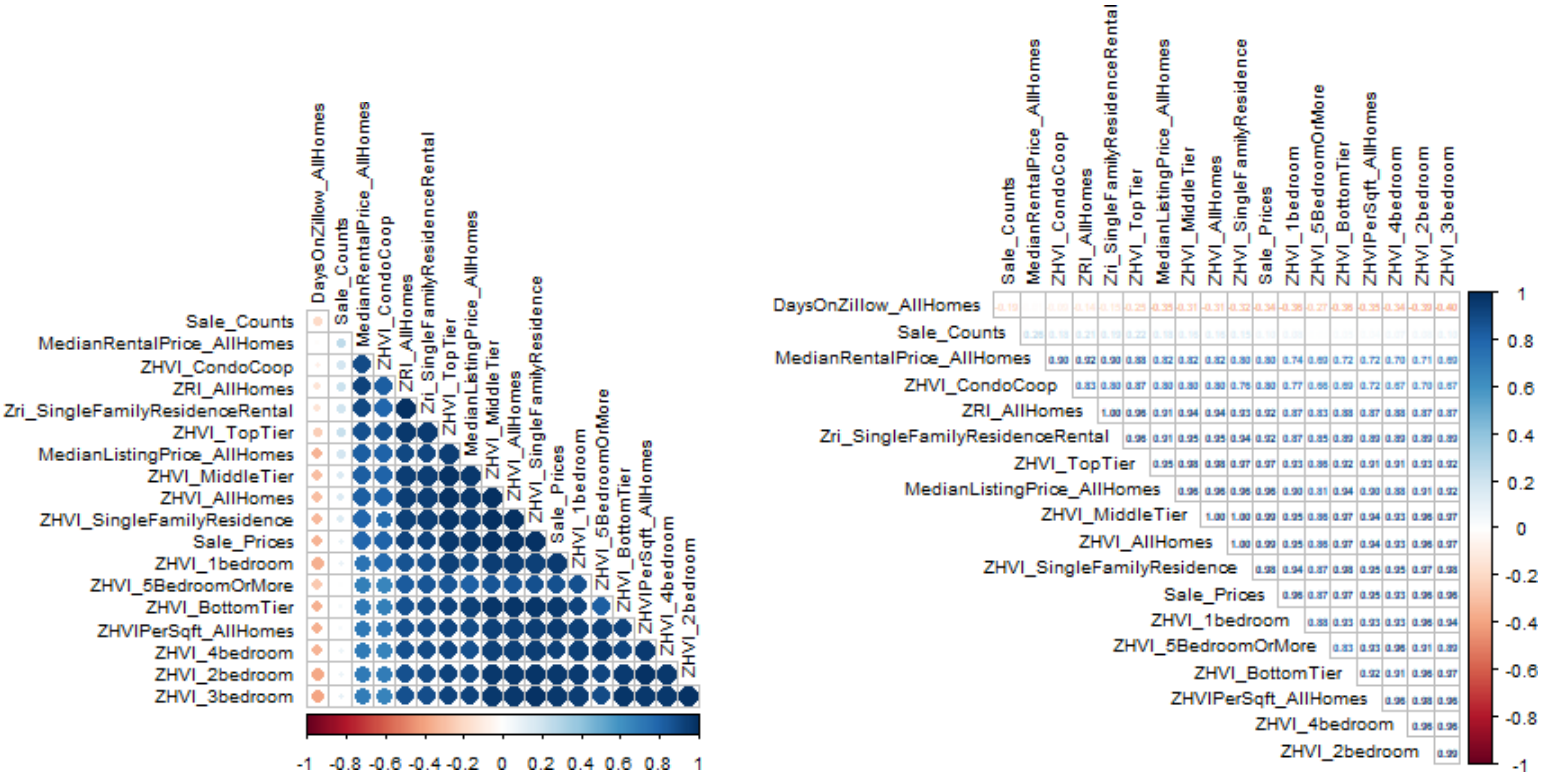




The Tier-wise analysis All homes across US Metros shows us that:

- Bottom Tier Homes are priced between $50,000 and $200,000.
- Middle Tier Homes are priced between $100,000 and $350,000
- Top Tier Homes are priced between $200,000 and $600,000.
- The states of District of Columbia, Hawaii, Massachusetts, and California, show the highest ZHVI values across all the three Tiers across US from 1996-2017.

Correlation matrix between metrics



→ The correlation plot of the variables shows that:

- There is a strong negative correlation between the DaysonZillow variable versus all the other variables which indicates that as the Days on Zillow period of the listing increases, the sale prices of the property decreases.
- All the other variable except for Sale_Counts are highly positively correlated with each other which indicates that as the property value of a certain house type increases the price of the other House type also increases.

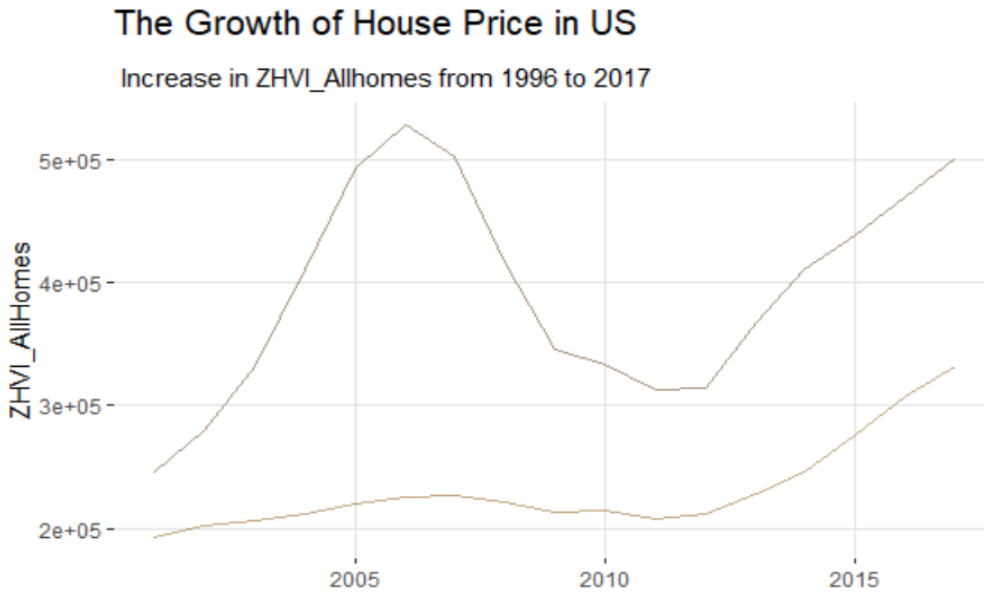
Data cleaning and EDA:

- Removed all unnecessary columns to focus on research for ZHVI All Homes, 1-5 Bedroom Homes, Condos and Top, Middle Bottom Tier Homes
- Removed all NA Values and took subset of dataset to focus on following attributes for my further analysis:

```
> colnames(train)
 [1] "Date"                          "RegionName"
 [3] "ZHVIPerSqft_AllHomes"          "PctOfHomesSellingForGain_AllHomes"
 [5] "ZHVI_1bedroom"                 "ZHVI_2bedroom"
 [7] "ZHVI_3bedroom"                 "ZHVI_4bedroom"
 [9] "ZHVI_5BedroomOrMore"           "ZHVI_AllHomes"
[11] "ZHVI_BottomTier"               "ZHVI_CondoCoop"
[13] "ZHVI_MiddleTier"               "ZHVI_SingleFamilyResidence"
[15] "ZHVI_TopTier"                  "year"
```

Data Visualization:

House Price changes in the US (1996-2017) based on region:



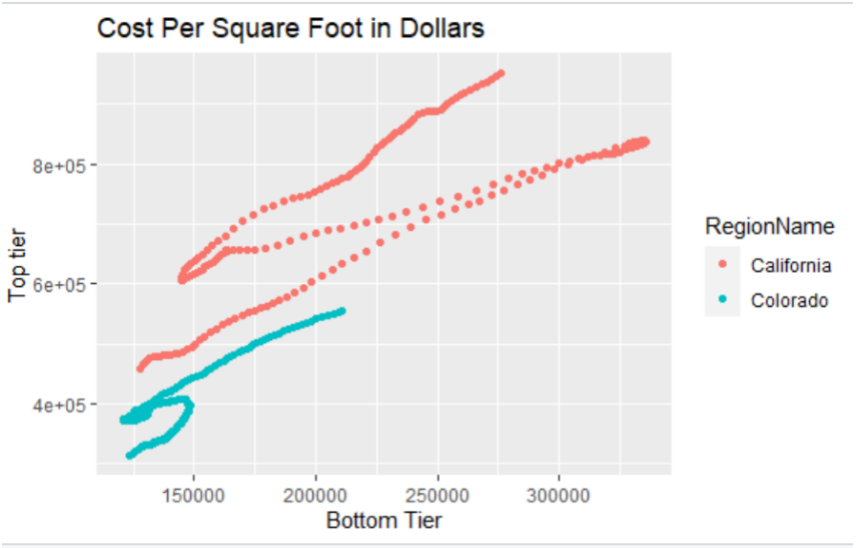
These were the salient observations gained from the above plot:

- Real estate market was at its peak in 2005: found multiple reasons like reinstated IT stability, global trade, and political stability.
- It crashed miserably in 2008 and continued to remain low till 2011 due to world-wide recession, monetary crisis due to oil and bankruptcy of major banks like Lehman Brothers
- There was a revival in real estate market post 2012 when it again started going upwards onto 2017.

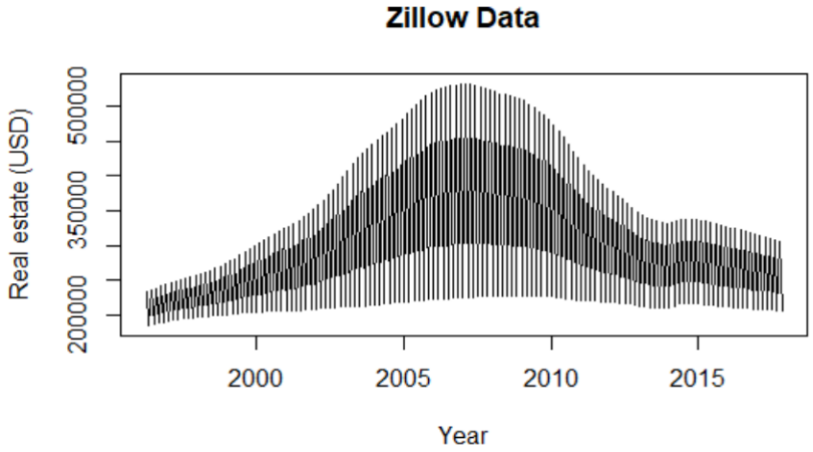
Bottom Tier vs Top Tier Cost Per Square Feet Distribution:

These were the salient observations gained from the above plot:

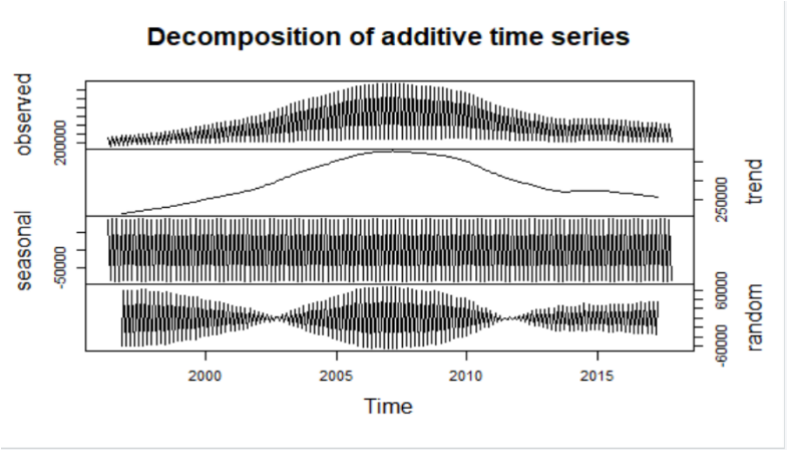
- Across all years, Cost Per Square Feet has been highest in **California** while **Colorado** has the lowest cost per square feet (bottom tier).



Time series modelling:



On executing the ADF test, we find that the p-Value is high and need to decompose the time series:



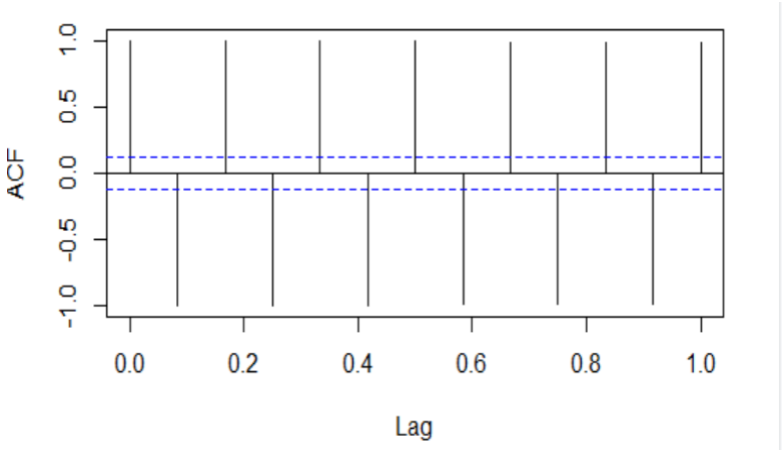
- There is trend but no definite seasonality: Peaks in the end of the year and small peaks in the summer

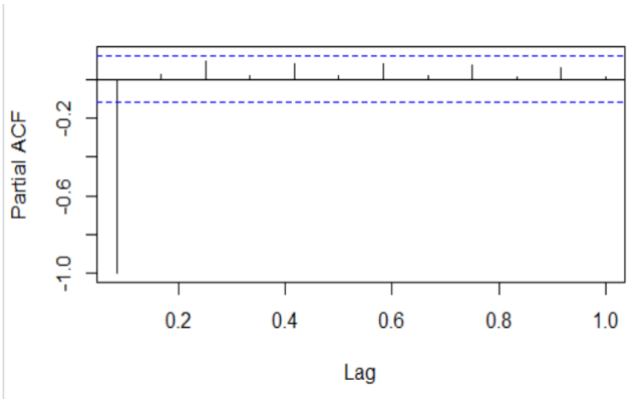
months.

- The trend in home prices was at highest in 2005, decreased rapidly from 2008 to 2011 and increased again from 2014 to 2017.
- With log and difference transformations, we are successful in making time series stationary.

ACF and PACF:

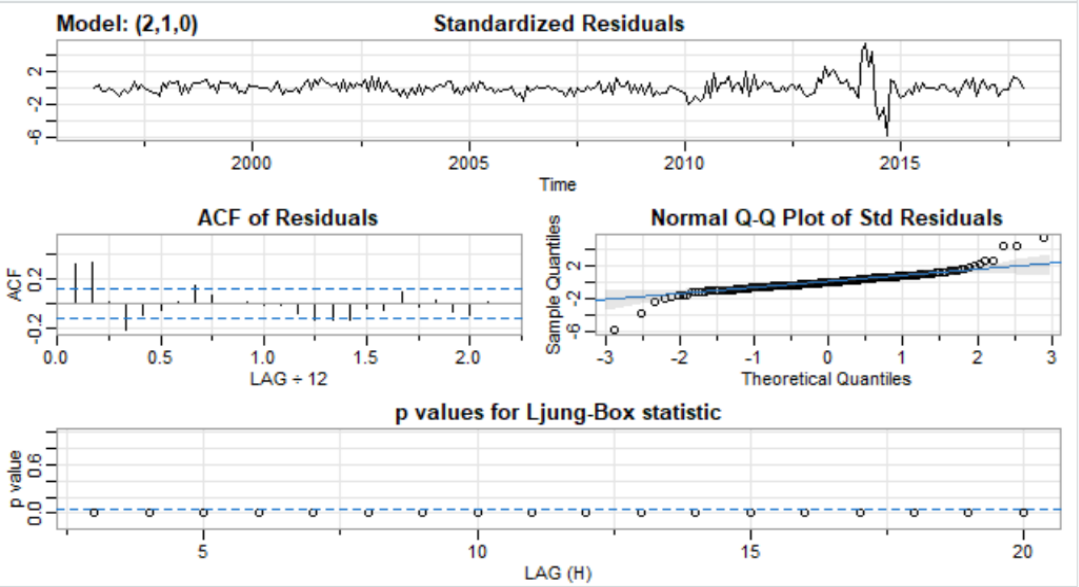
- As per the ACF plot, it is an AR (1) model. PACF spikes at 0 position, indicating an MA (0) model
- Further the seasonal pattern is 12 months.





ARIMA and SARIMA:

The ARIMA model with Seasonality component does not work well in our case due to indefinite seasonality. Model does not give good statistics and proper convergence. Therefore, tried the SARIMA (2,1,0) with additional variables which converged but did not give particularly satisfactory results.

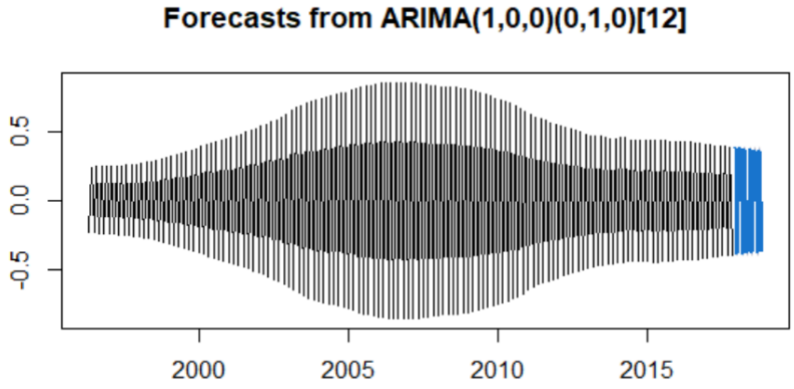


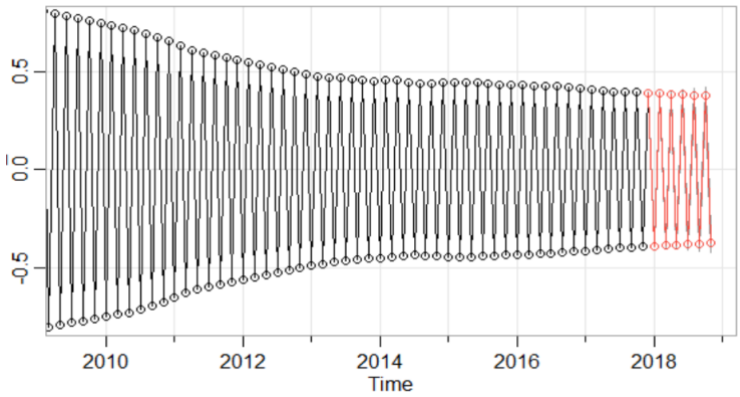
Observations:

- The Standardized residuals show some pattern until end of 2005 where the real estate sales are highest. They again decrease at start of 2011 and then remain constant. Since there is no continuous pattern, it resembles white noise.
- Similarly, ACF also resembles the Standardized Residuals in terms of white noise.
- QQ plot has few outliers in the beginning but otherwise has a good fit
- The P-values are all below the significance level which again suggests white noise.

Sales Forecasting:

Suggests continuous pattern in Real Estate Home Sales after 2017. It means demand for houses will remain constantly high over next few years. We have similar observations with 1. ARIMA with seasonal component and 2. SARIMA





Next, we will explore a few advanced model build options for this dataset.

Section 3: Advanced Modeling with Additional variables

In this section, we attempted to predict the pattern based on regression-based models and other variables. This data from Zillow on housing prices were affected by several external factors for which data was not provided (as the analysis showed). There are two significant changes in trends (Bends) as highlighted in the charts below which were difficult to Model. However, apart from these two significant changes if we build the model on other cross-sectional view of the data, the model does well and is heavily correlated on preceding values.

Before starting on the modeling process, summarizing key takeaways from the initial modeling process

Key Takeaways from EDA for Building the Model

- The Sale Price of the homes vary significant by region for obvious reasons. So, need to build separate model for each region
- For each tier prediction, the model must be built separately. We can predict the direction of the tiers from each other but not the actual price by tier
- The sales price is bimodal, with no seasonality and driven most likely by external factors
- To explore opportunity to find other data to predict sales, especially the upward and downward trend, employment rates (from Labor statistics) and Interest rates were also compared
- Based on the above Linear Regression was used to fit and predict the trend
- The time series heavily depended on preceding values
- The available variables were not able to predict the change in trends
- Different Regions have change in trends(bends) around same time (as we would expect) but the magnitude was different

The trends seen across different regions selected (Georgia, New York and California) and target variables (Low Tier, Medium Tier and Top Tier) is same, but the magnitude of change is different

Chart 3.1

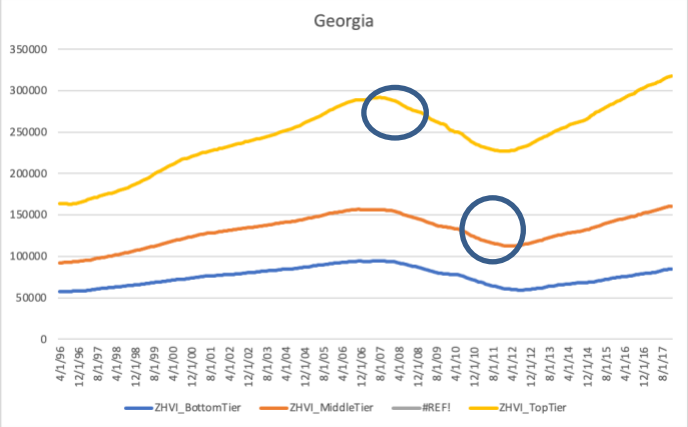
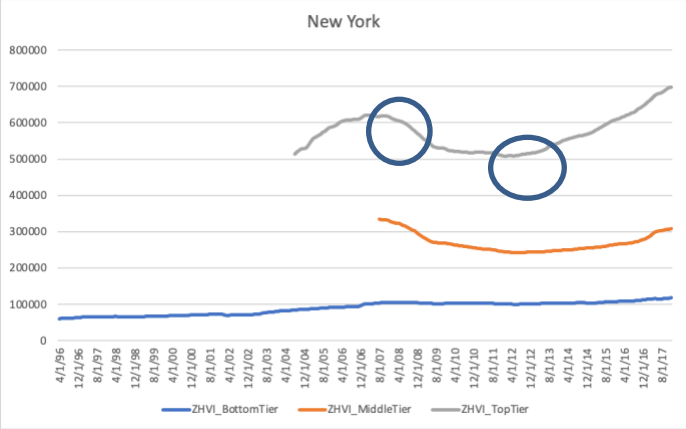
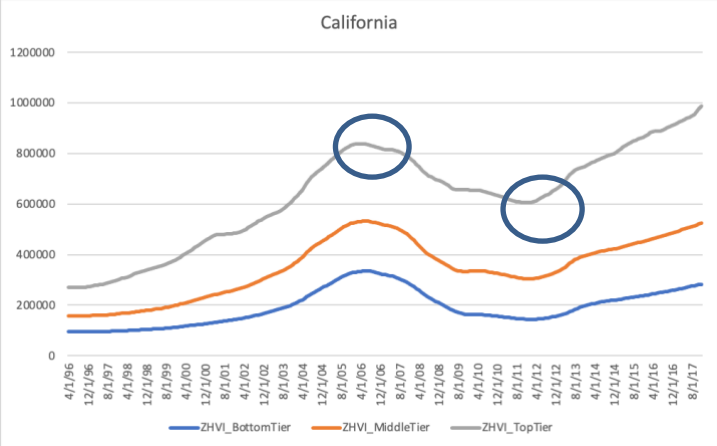


Chart 3.2



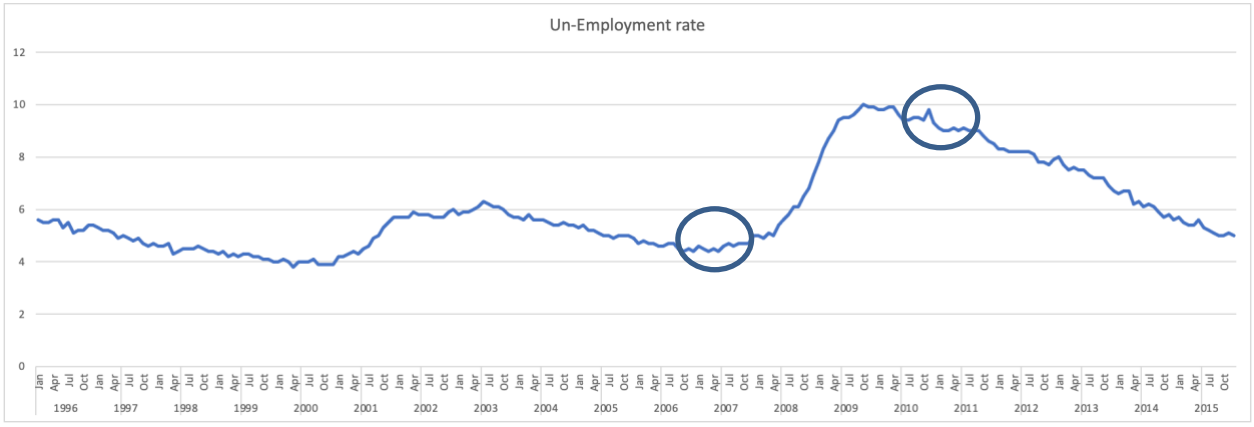


As shown in the above charts, there are two significant bends (change in trends). This is not seasonality and not cyclical. It's driven by external factors related to economy. To model this trend we looked at socio economic data available like unemployment rates from Bureau of Labor

(https://data.bls.gov/timeseries/LNS14000000) and interest rates from readily available data across websites

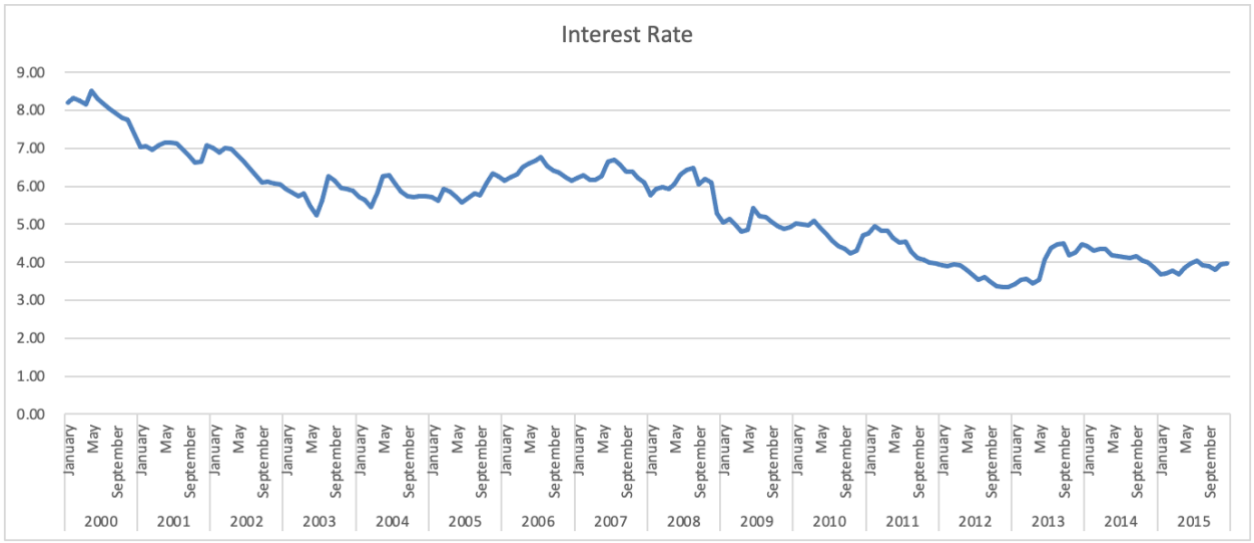
External Factors (Variables)

A) Unemployment Rate



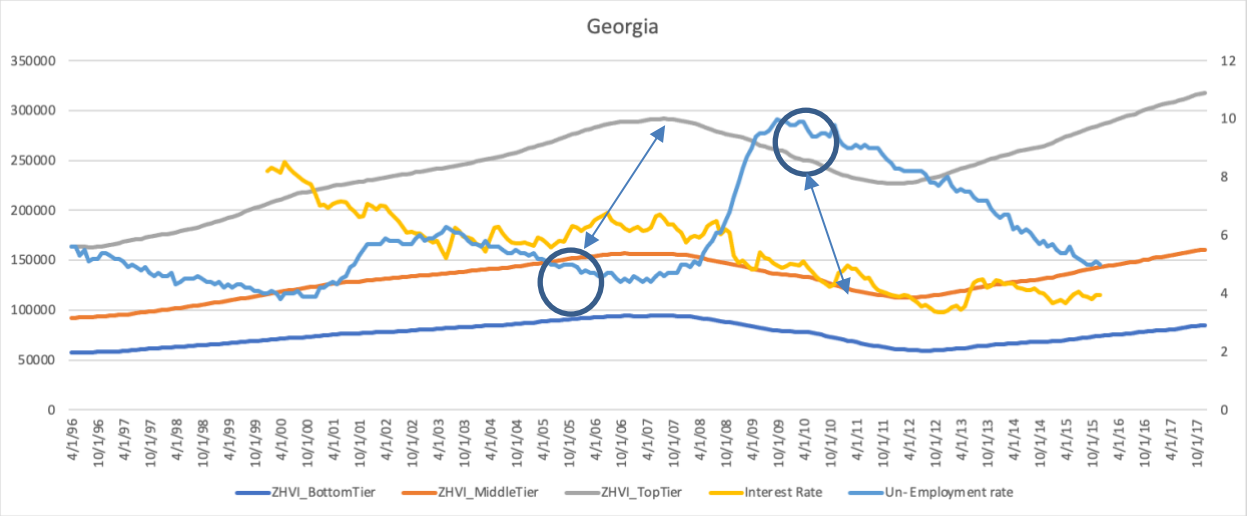
On analyzing the unemployment trends carefully, we see that the **unemployment rate can be used as proxy to economic trends** and it can help predict the bends (trends) in the sale prices, it seems to have negative correlation to the sales price (for obvious reasons)

B) Interest Rates



On analyzing the interest rates time series carefully, we see that the interest rates trends cannot help determining the change in trend (predict the bends) in the sale prices, however, they can help determine the slope of the curve

The chart below shows, the trends of Georgia along with the unemployment rates and interest rate trends for easy comparison



The bends can be predicted with the bends in Un-employment rates few months earlier

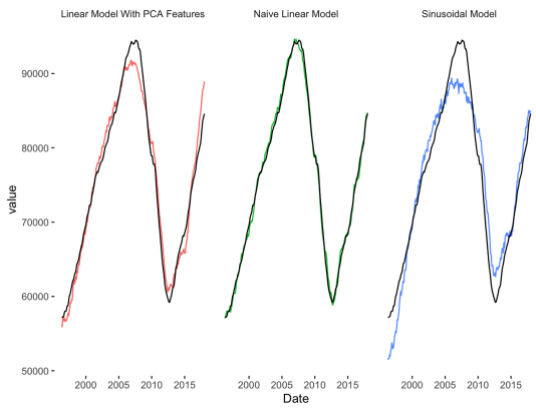
As shown above these external variables can be used to predict the change in trends. However, post incorporating these variables in the model, these variables were not as strong predictors as the preceding (lag months) itself

After trying multiple models, the models seems to overfit and the linear models seems to not provide incremental benefit in using them

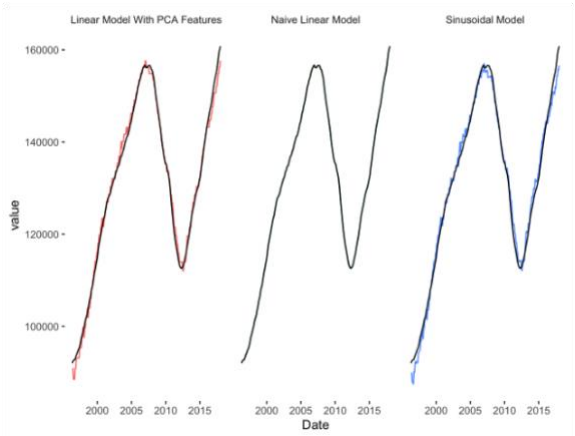
Model Building

The linear models seems to overfit

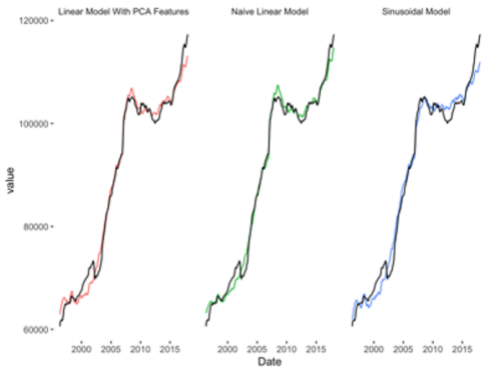
Georgia (ZHVI_BottomTier)



Georgia (ZHVI_Middle Tier)



New York (ZHVI_Top Tier)



Variables Used: "Date" "RegionName" "ZHVIPerSqft_AllHomes" "ZHVI_1bedroom" "ZHVI_2bedroom" "ZHVI_3bedroom" "ZHVI_4bedroom" "ZHVI_5BedroomOrMore" "ZHVI_AllHomes" "ZHVI_CondoCoop" "ZHVI_MiddleTier" "ZHVI_SingleFamilyResidence" "ZHVI_TopTier"

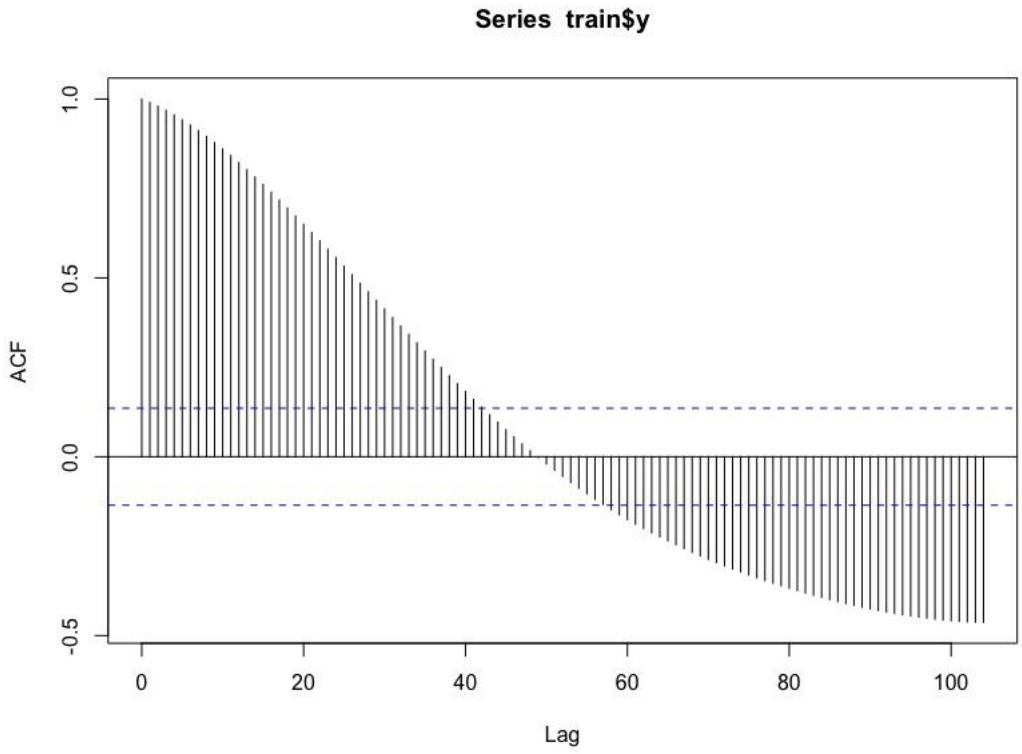
Summary of this section

- In this project where trends are not seasonal, the standard methodologies and model do a good prediction in the scenarios both pre and post a significant event (events which changes the trend) but cannot model the significant events
- The variables available are not sufficient to model the significant events
- Need to look at alternate data sources or fields to model the change in price. We can also look at Yield curve, availability, and price trends of raw materials etc. to predict the overall trends
- The linear model seems to be overfitting. BSTS model does a better prediction along with Time series analysis (post significant event trends)
- The preceding values are strong indicators of future predictions

Part 4: Prophet Model and BSTS Model

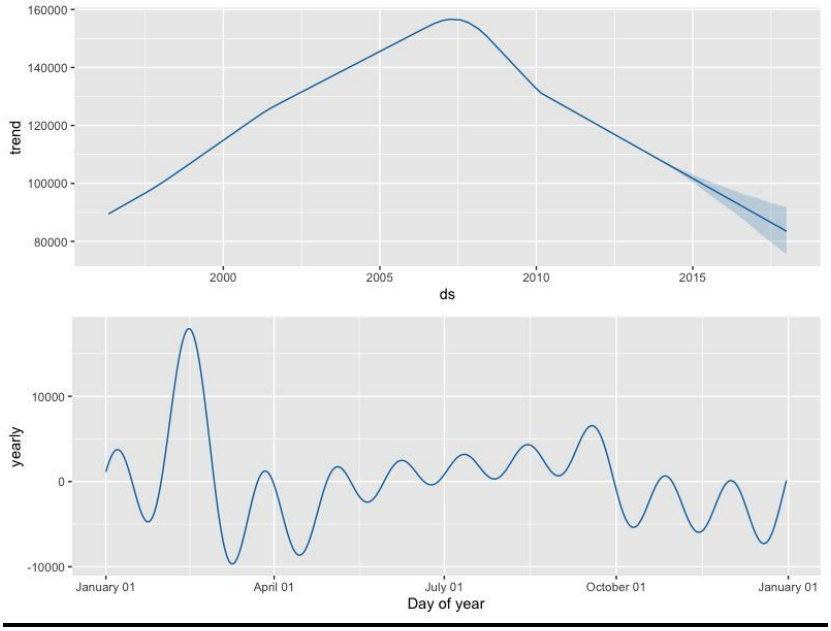
The next models were applied to the data are Prophet model and BSTS model. For simplification, only column ZHVI_AllHomes (in Georgia) was used to build models. There are several reasons for it. First, this column represents the index price of all types of home which would have more general information than other indexes. Second, there is no missing value from this column since 1996 which would provide enough continuous data to build the model.

Determine the original component of ZHVI_AllHomes:

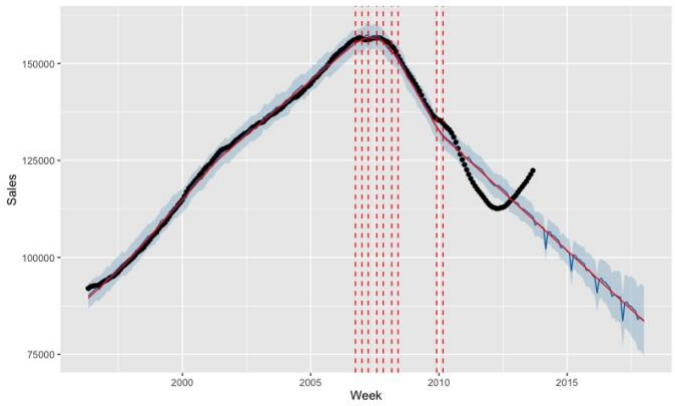


As shown in the ACF plot, the crucial component of the data is trend. Also, noticed that, there is no seasonality in this data column.

Prophet Model:



The above image shows the components of the Prophet model when the prediction is applied. A clear linear trend could be seen from the data. The yearly residual trends also show no definite pattern, which confirms the linearity of the time series.

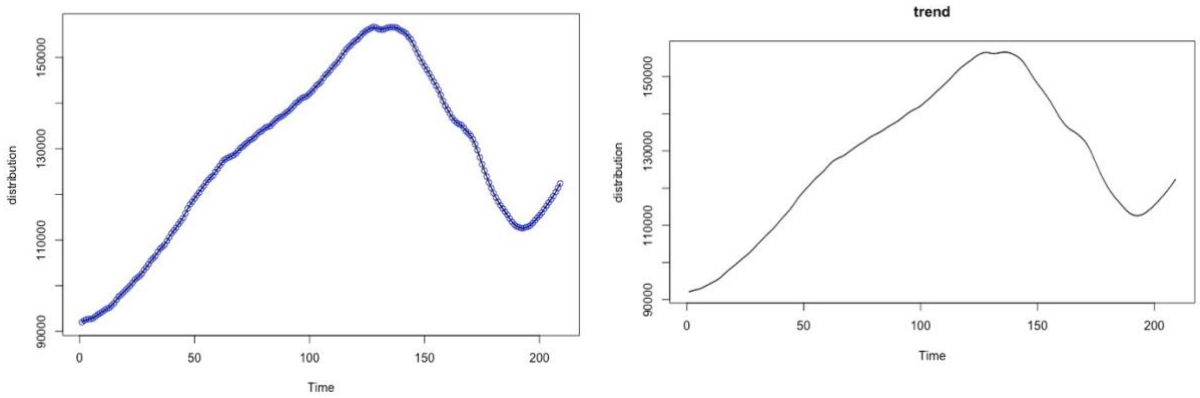


As shown in the above image, Prophet model does not work well with data that has no seasonality. Thus, it cannot accurately forecast the future outcomes from the data. The MSPE for this Prophet model is 31.05545%, which is very high for a forecasting model.

BSTS Model with different parameters:

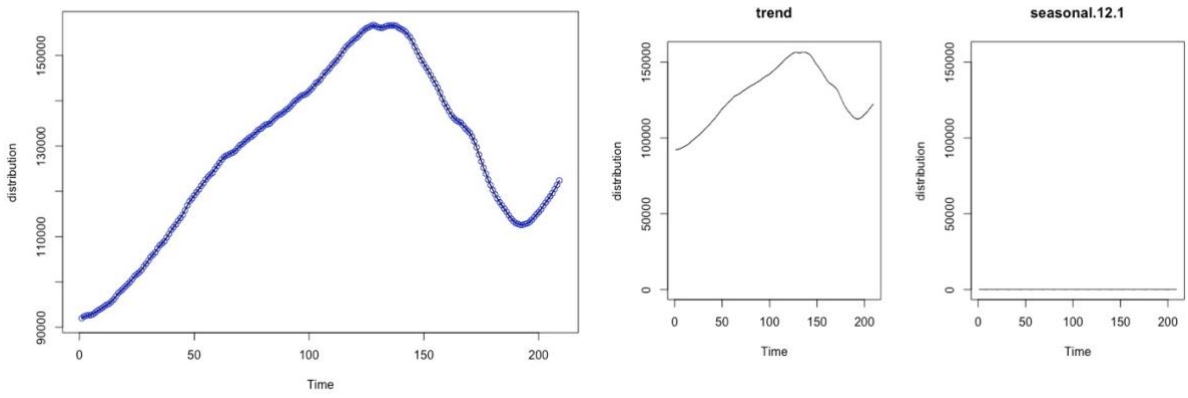
BSTS model is a good choice for ZHVI_AllHomes data because it can handle nonstationary time series. Also, it can forecast with few historical data points and handle high variability in the data.

BSTS model with Trend



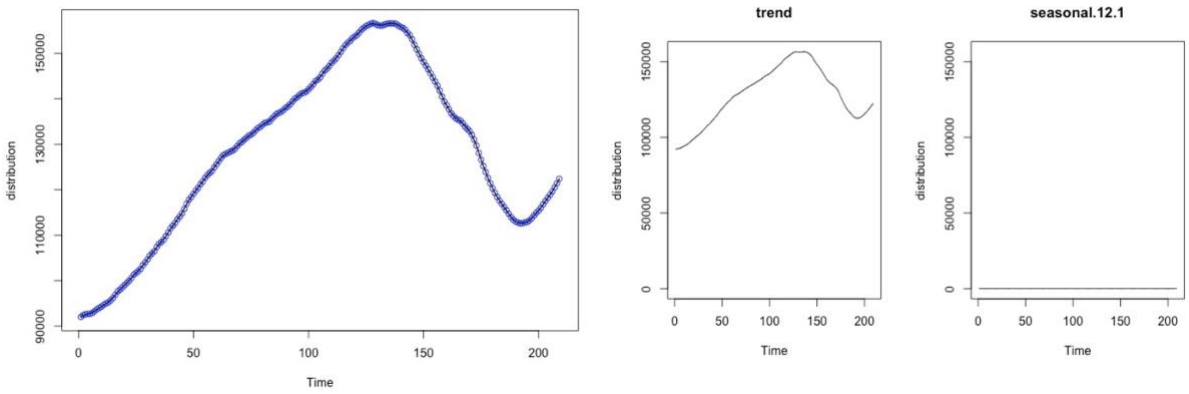
- Left image is the plot of BSTS model applying linear trend, and right image shows the component of the model
- Experience much better MSPE than Prophet model. For example, the model MSPE is 3.212157%, which slightly increased from the original model.

BSTS Model with Seasonality



- Introduced seasonality for the training model. Although trend was not introduced in the model, it still can capture the trend from the data.
- This model MSPE is 3.436987%, which increased from the original model.
- The increasing of MSPE suggested that the BSTS model should not include seasonality as a parameter.

BSTS Model with Trend and Seasonality



- Introduced both seasonality and trend for the training model, but model can only capture the trend from the data. The component graph shows no seasonality in the data.
- This model MSPE is 3.693665%, which increased from the original model (trend only).
- The increasing of MSPE suggested that the BSTS model should not include seasonality as a parameter.