**A Project Report**

**On**

**EXPLORATORY DATA ANALYSIS AND PREDICTION OF HOUSING DATA**

Submitted by:

Pradeep Chhetri

SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR THE APPRENTICESHIP OF FINANCIAL DATA ANALYST APPRENTICE PROGRAM

UNDER CODERUSH

OCTOBER 2022

**Table of Contents**

[INTRODUCTION 3](#_Toc116230118)

[KEY ASPECTS 4](#_Toc116230119)

[Data 4](#_Toc116230120)

[Methods 4](#_Toc116230121)

[Data Visualization and Analysis 4](#_Toc116230122)

[Results: 15](#_Toc116230123)

[CONCLUSION AND RECOMMENDATIONS 17](#_Toc116230124)

[APPENDIX 18](#_Toc116230125)

[Bibliography 21](#_Toc116230126)

# INTRODUCTION

Housing market has been one of the popular markets. The pricing of houses and real estate is dependent upon various factors. Before the involvement of data, many of the decisions made in real estate were mainly based on experience and limited analysis of trends. Now, property professionals use real-time and detailed information to make informed decisions. (GetSmarter.com, 2022)

In this project, we will study the movement of house prices over time and various factors that have influenced the housing prices. We will perform various exploratory data analysis, visualizations and predictions to draw meaningful insights from available data. This project will help us to have in-depth knowledge about housing prices.

The major questions in the project are: How has the housing market performed over time across different states in the US? How would the investment in real estate of United States perform over the long run? How bad does real estate perform during market crashes? How will be the performance of models for predicting house prices based upon certain independent variables?

# KEY ASPECTS

## Data

For the dataset, we will be using Housing data provided by Zillow. Zillow Economic Dataset consists of housing price data of 20 years (1996-2017). Using the historic housing data, we can perform various analyses and predict the future trend of housing data. The dataset is a time-series dataset. Time series data, also referred to as time-stamped data, is a sequence of data points indexed in time order. Time-stamped is data collected at different points in time. These data points typically consist of successive measurements made from the same source over a time interval and are used to track change over time. (influxdata.com, 2022)

The original data set has 82 variables and 13212 observations. The descriptions about the variables are provided in the data dictionary. In the dataset, we have different features like Date, Median Listing Price, Median Rental Price, Sales Price, Zillow Rental Index, Zillow Home Value Index etc. of different types of houses.

## Methods

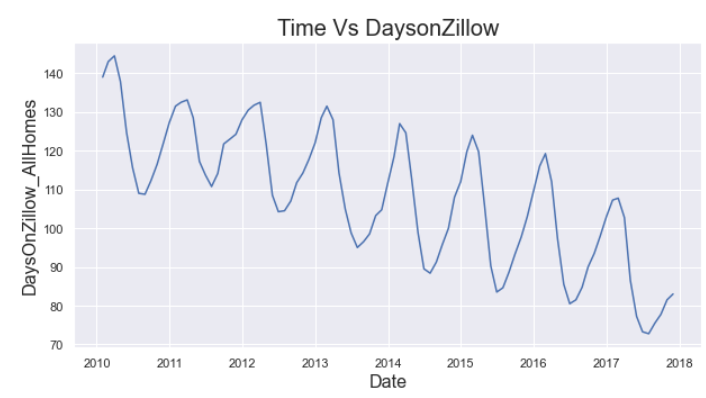
For data analysis, we will be using the Python Programming language. Firstly, the dataset will be loaded using the pandas library, and proper analysis of features will be done. We will clean the data as per our requirement, and deal with the missing values. Amongst the various features available, the suitable features will be selected for further analysis of data. At first, we will perform univariate analysis of various features, and draw insights from the data. Correlation between various features will be calculated in order to find the relationship between various data features. As the selected dataset is a time-series data, we will perform various time-series analysis in order to study the different trends and patterns over time. Similarly, we will also build a model to predict the price of houses as well as other features using different time series forecasting methods like ARIMA. Likewise, we will also use regression models to predict house prices.

## Data Visualization and Analysis

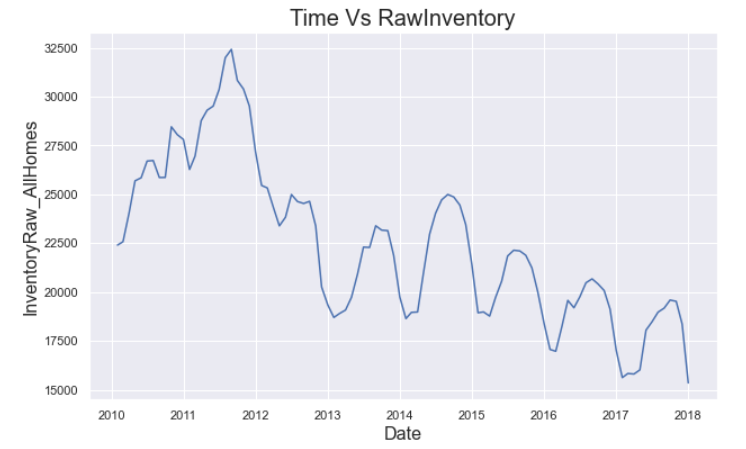
We intend to visualize and analyze the given housing dataset and identify how the different variables affects the price of houses in different regions of United States over time.

For that, we first performed univariate analysis of the different variables available in the dataset. The dataset contains housing price data of 20 years ((1996-2017). It was also observed that there were a lot of missing values in different variables for the time period between 1996-2010, and analysis was done only using the data after 2010 in case of those variables.

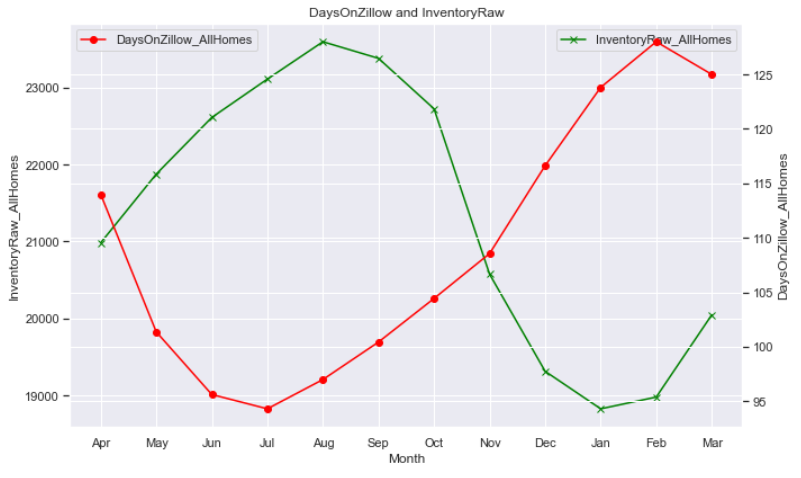
The dataset contains the features of 52 different regions in United States. Later, we will also discuss other variables grouped on basis of these 52 different regions and discuss the relationship of the features based on regions.



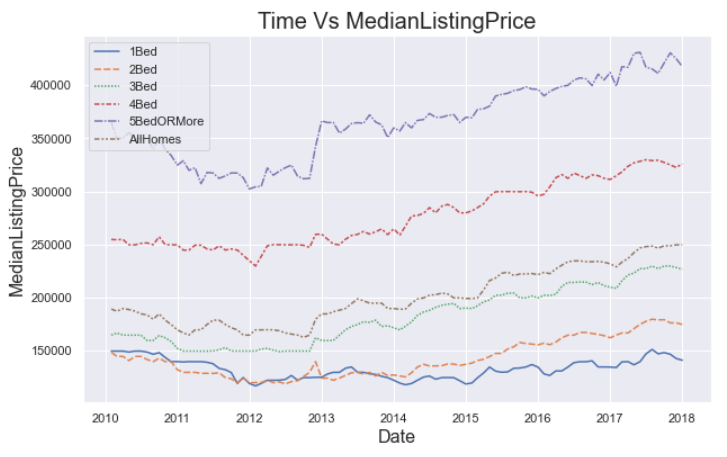
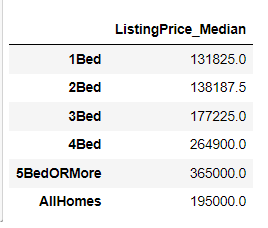
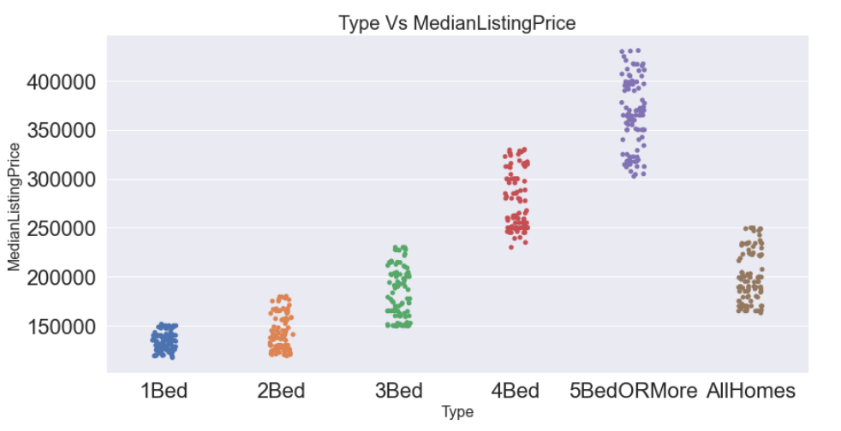
DaysOnZillow\_AllHomes is the median days on market of homes sold within a given month, including foreclosure re-sales. The feature seems to be recorded only since 2010. In addition, the DaysOnZillow seems to fall from 140 in 2010 to around 90 in 2017.



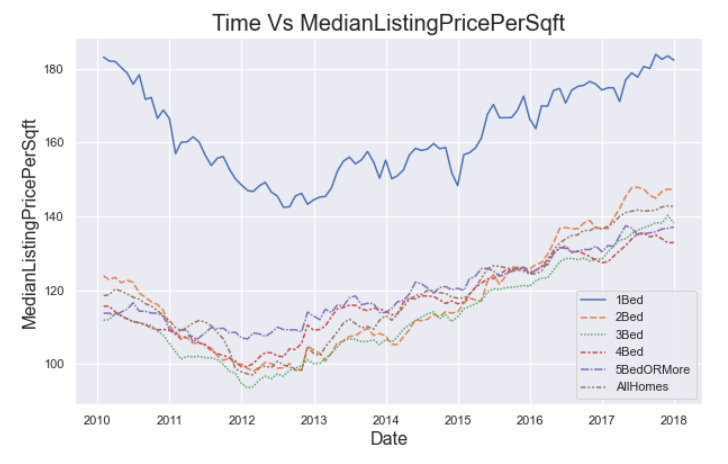
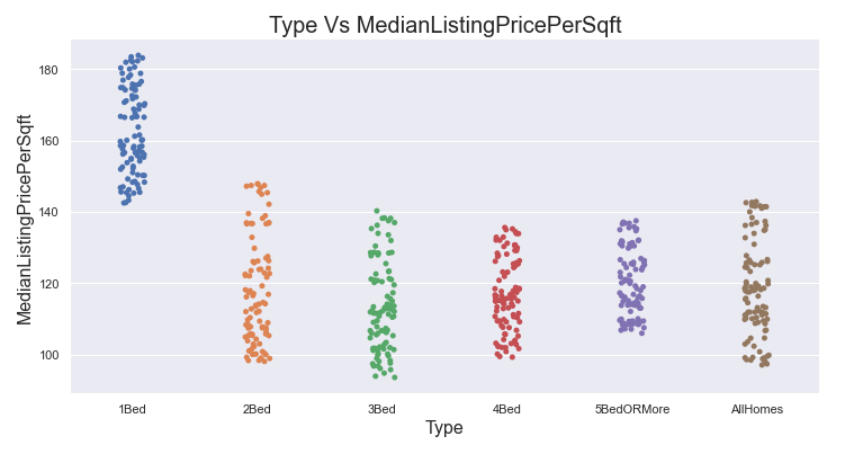
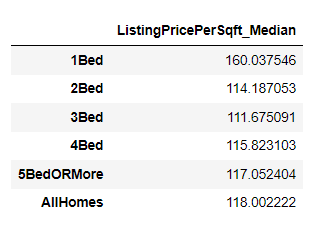
InventoryRaw\_AllHomes is the median of weekly snapshot of for-sale homes within a region for a given month. The inventory seems to be recorded only since 2010. In addition, the DaysOnZillow seems to rise from 22500 in 2010, reaching the peak of 32500 in 2011and finally dropping below 17500 in 2018.



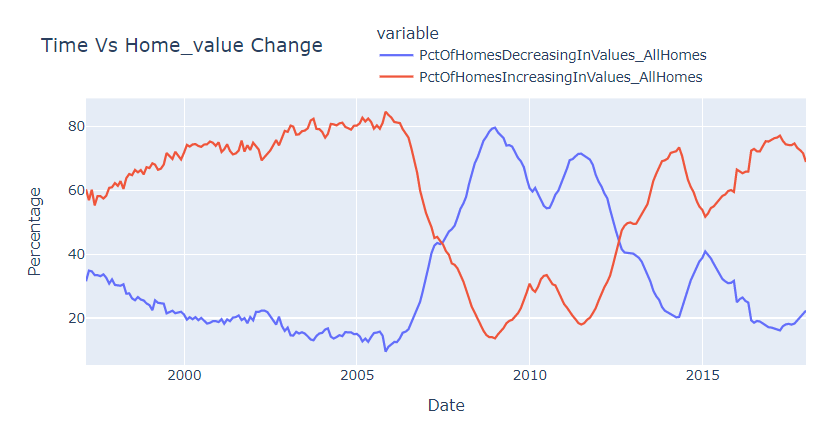
We can also see seasonal trend of data for DaysOnZillow which dipped on the month of July and peaked on the month of February. Similarly, the InventoryRaw is also seasonal which peaked on the month of August and peaked on the month of January. The two variables DaysOnZillow and InventoryRaw have an inverse relationship.



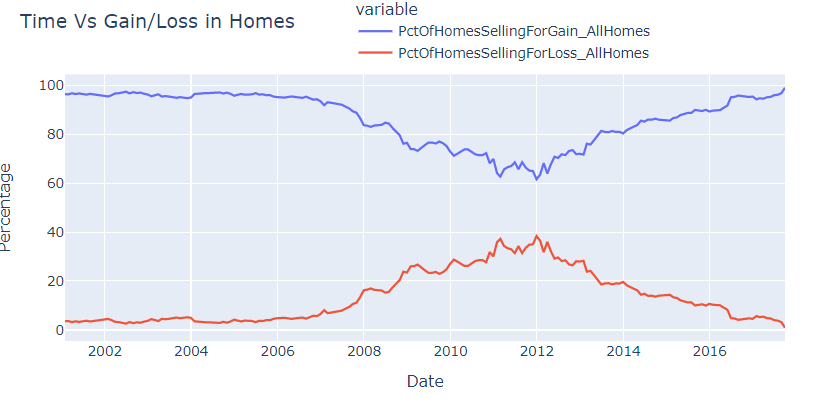
MedianListingPrice is the median of the list price (or asking price) for homes listed on Zillow. The above visualization shows the median listing price of houses is in long-term uptrend during the review period. Similarly, the median listing price is found to increase with increase in number of rooms in the house. While analyzing the data, we can see similar behavior in other variables like RentalPrice, ZHVI and few others.



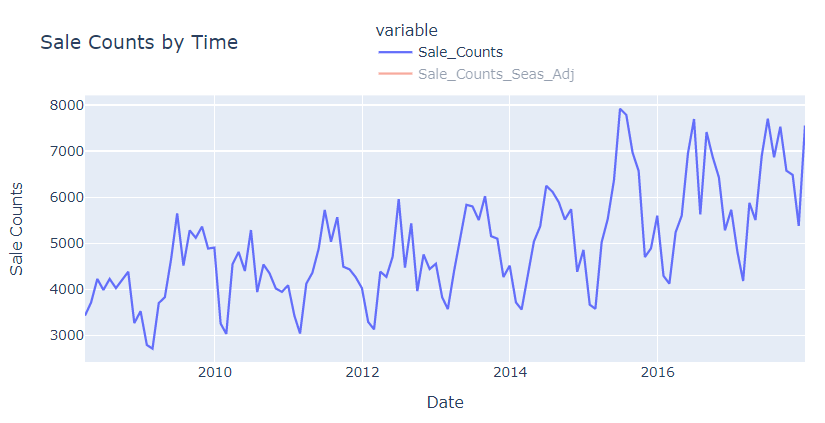
MedianListingPricePerSqft is the median of list prices divided by the square footage of a home. The above visualization shows the median listing price per square feet of houses is in long-term uptrend during the review period. Similarly, the median listing price per square feet is found to decrease with increase in number of rooms in the house. One bedroom houses had the highest per square feet listing price at around $160 per square feet.



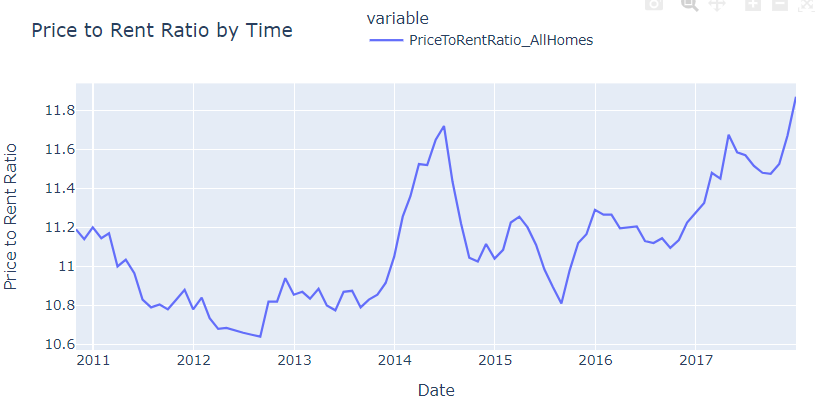
As a general trend, house prices are found to be up-trending in the long run, and the value of homes are found to be increasing in value. However, we observed high percentage of houses decreasing in value due to the housing market bubble during 2007-2008. During November 2008, 79.5 % of homes were found to be decreasing in value. After the economy recovered after the crisis, the value of homes seems to be restored.

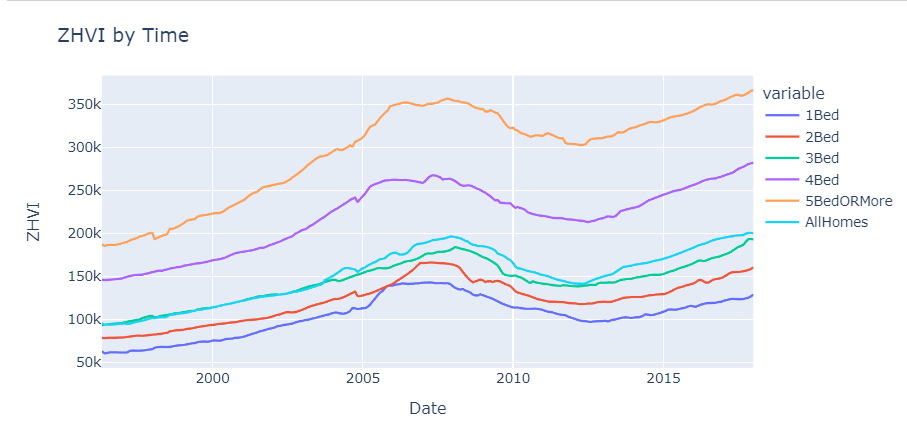


From the graph, we can observe that around 40% of houses were sold for loss during the peak crisis of 2008 as a result of housing market crash.

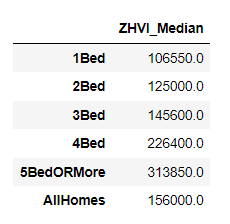
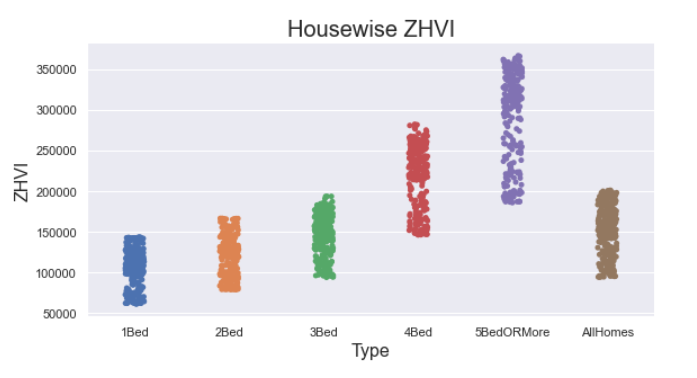


Sale Counts is the number of houses sold in a given time period. As like the house prices, the sale count is also observed to be in uptrend with a seasonal behavior during the review period, starting above 3000 in 2009 to reaching above 7000 in 2017.

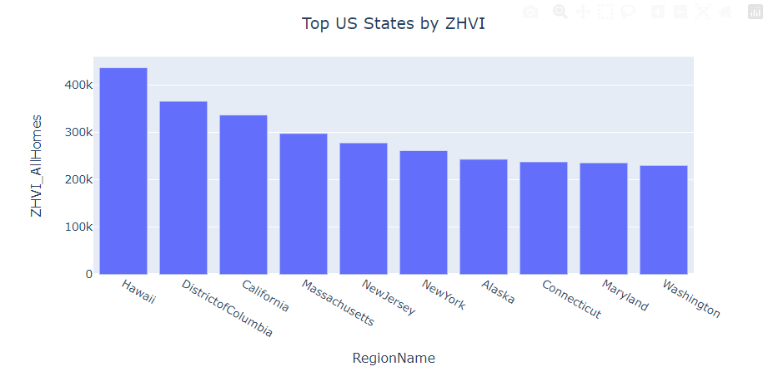
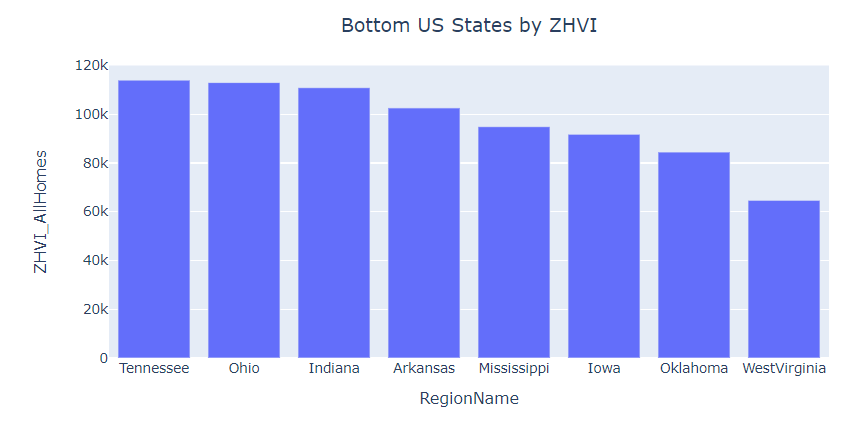
PriceToRentRatio is first calculated at the individual home level, where the estimated home value is divided by 12 times its estimated monthly rent price. Then, the median of all home-level price-to-rent ratios for a given region is then calculated. The Price to Rent Ratio has reached above 11.8 in 2017, which was 11.19 back in 2010.



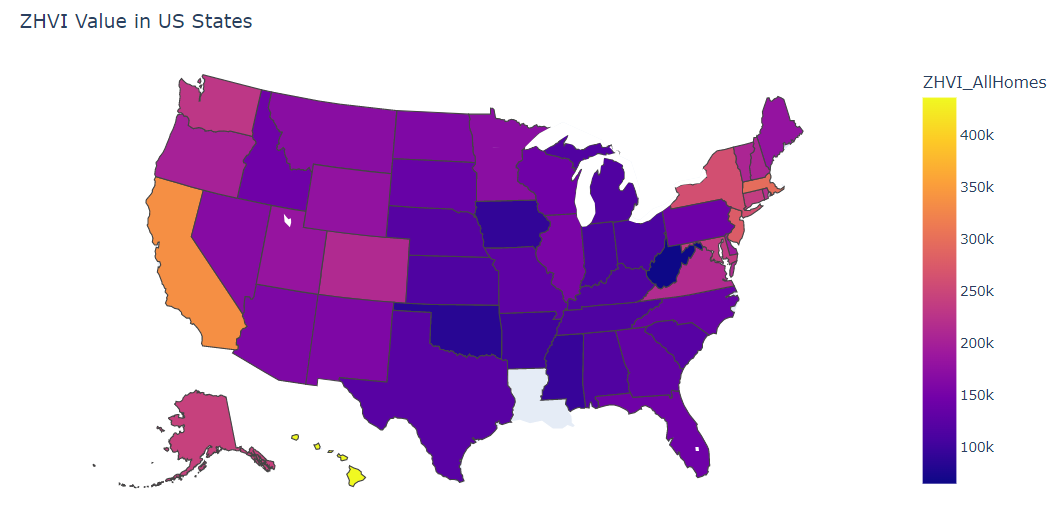
ZHVI is a smoothed seasonally adjusted measure of the median estimated home value across a given region and housing type.



We can observe that houses with more rooms tend to have higher prices and higher ZHVI. Houses with five or more bedrooms have a median ZHVI of $313k, while houses with one bedroom have a median ZHVI of $106k.



We can observe that the states with highest house prices are Hawaii, DistrictofColumbia, California etc. with price range above $ 300k. Similarly, the states with lowest house prices are West Virginia, Oklahoma, Iowa etc. with price range below $ 100k.



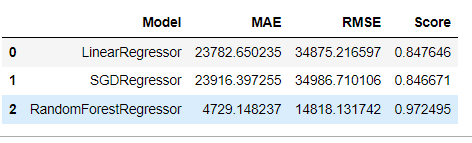
Similarly, we also plotted the same information in Chloropleth plot to depict the value of ZHVI across different states in United States.

**Regression:**

We built different regression models which can predict the value of ZHVI based on other independent variables. At first, we selected the features within the AllHomes category for uniformity, where we had 22 features initially. Variables having high number of null values were dropped and multi-collinearity was checked among the remaining variables.

We converted the date column into datetime datatypes, and categorically encoded the ‘RegionName’ column. We added an extra column ‘Year’ from ‘Date’ column. And, ZHVI\_AllHomes was added as a target variable. In the dataset, most of the columns had null values before 2010, so we only considered the data after 2009 in the regression analysis. Even in the remaining features, two columns (MedianRentalPrice\_AllHomes and PctOfHomesIncreasingInValues\_AllHomes) had few missing values which was replaced with the year-wise mean value.

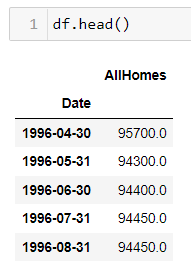
The performance of these models is shown in the table below:



Among the three models, Random Forest Regression performed the best with lowest Root Mean Squared Error and highest model score.

**ARIMA:**

Similarly, we also performed time series analysis prediction using ARIMA model on the variable ZHVI\_AllHomes. First, the daily median across different regions was used for ARIMA model.



We performed the seasonal decomposition where the time series was decomposed into trend component, seasonal component and residual component.

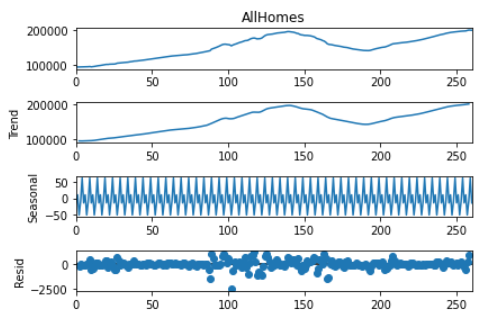
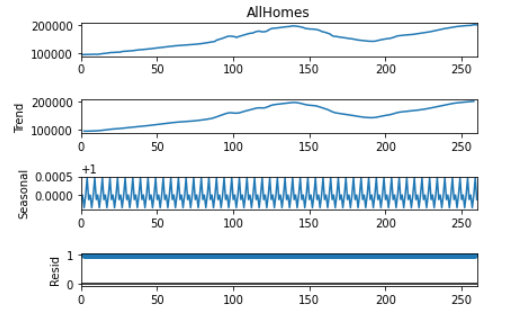
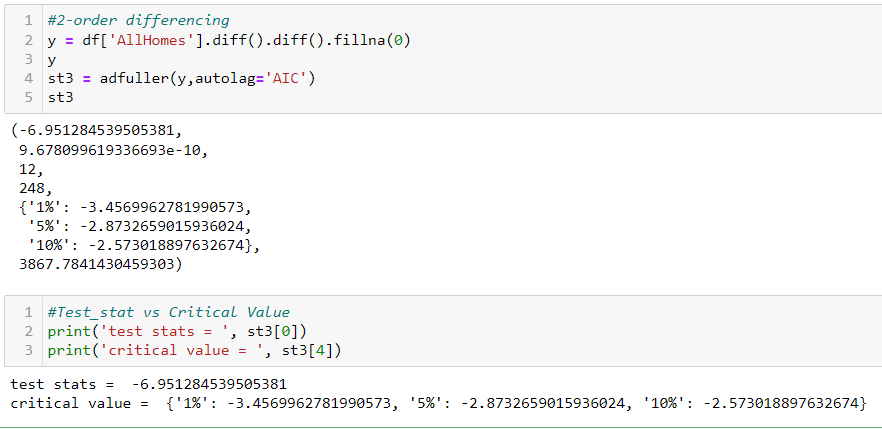


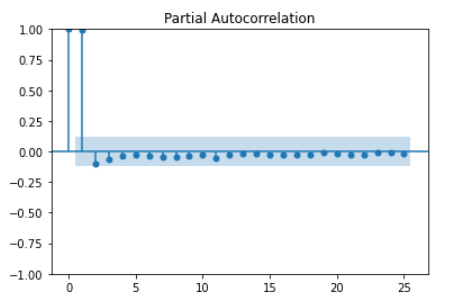
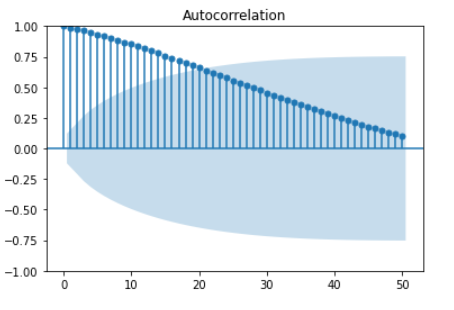
Figure: Multiplicative Decomposition

Figure: Additive Decomposition

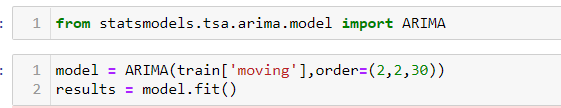
We also performed the ADF test to check the stationarity of the dataset. The data was found to be non-stationary only after 2nd-order differencing.



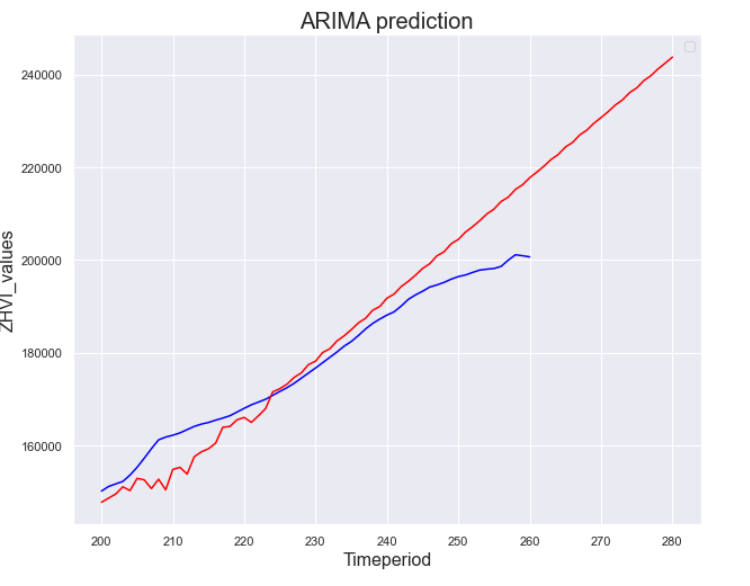
Then, we performed partial autocorrelation plot and auto correlation plot to find out the autoregressive term (p) and number of lagged forecasts (q) respectively.



From the plots, we determined the optimal value of p and q to be 2 and 30 respectively. Since, the dataset was converted into non-stationary data after 2nd-order differencing, the value of d was determined to be 2. So, the ARIMA model was fitted as following:



Finally, the dataset was predicted using ARIMA model as follows:



## Results

* DaysOnZillow seems to fall from 140 in 2010 to around 90 in 2017.
* The two variables DaysOnZillow and InventoryRaw have an inverse relationship.
* The median listing price is found to increase with increase in number of rooms in the house. While analyzing the data, we can see similar behavior in other variables like RentalPrice, ZHVI and few others.
* The median listing price per square feet is found to decrease with increase in number of rooms in the house.
* High percentage of houses decreasing in value due to the housing market bubble during 2007-2008.
* Around 40% of houses were sold for loss during the peak crisis of 2008 as a result of housing market crash.
* The sale count is also observed to be in uptrend with a seasonal behavior during the review period, starting above 3000 in 2009 to reaching above 7000 in 2017.
* Houses with more rooms tend to have higher prices and higher ZHVI.
* States with highest house prices are Hawaii, DistrictofColumbia, California etc. with price range above $ 300k. Similarly, the states with lowest house prices are West Virginia, Oklahoma, Iowa etc. with price range below $ 100k.
* Random Forest Regression performed the best with lowest Root Mean Squared Error and highest model score.
* The ZHVI dataset was predicted using ARIMA model after 2nd-order differencing.

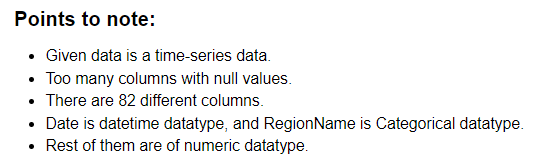
# CONCLUSION AND RECOMMENDATIONS

From the descriptive analysis and predictive modeling, we were able to draw various insights about the housing market in United States. From the analysis, we discovered that the number of rooms in a house is one of the major factors for determining the house prices, and the houses with higher number of rooms tend to have high listing price, rental price and Zillow Home Value Index. However, the houses with lower number of rooms tend to have higher listing price and rental price per square feet. We also observed that the housing market is up trending in the United States in the long run, except for few volatilities due to bubble formations and market crashes. During the market crashes, we observed around 40% of houses were sold for loss during the peak crisis of 2008 . The prices have been able to beat the previous peaks after the recovery from the crash, and the investment in real estate and housing seems attractive.

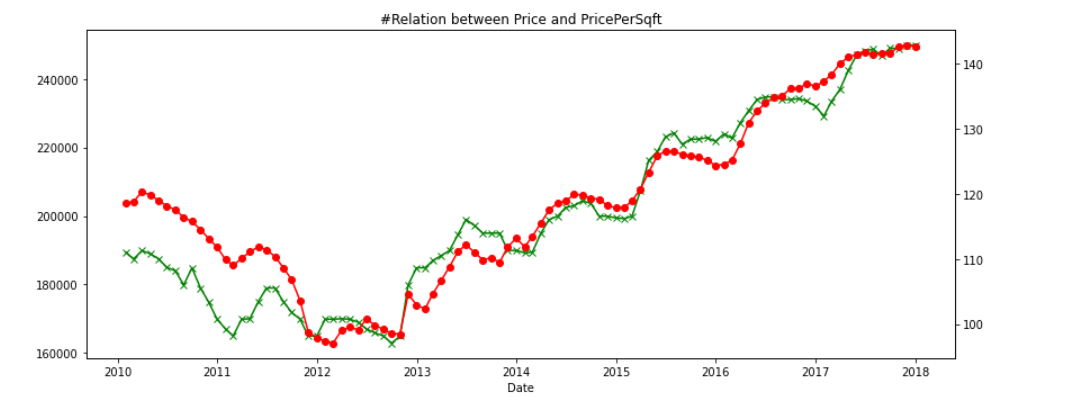
Similarly, Random Forest Regression performed the best with lowest Root Mean Squared Error and highest model score. The ZHVI dataset was predicted using ARIMA model after 2nd-order differencing. The use of machine learning and modeling for the predictions in house prices seems useful for helping the investors to buy and sale decisions based upon data. Models with further improvements can be utilized for making better decisions in the real estate market.

# APPENDIX

Information about Datatypes:



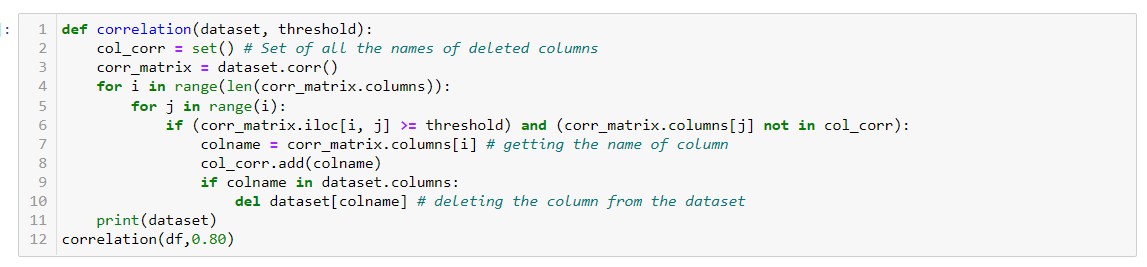
Relation between Price and PriceperSqft:



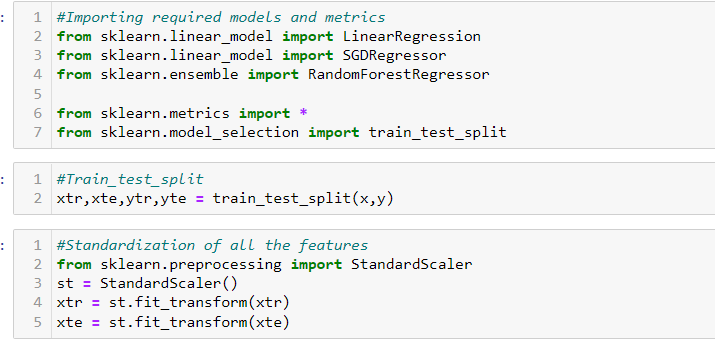
Chloropleth Code:



Code for Multicollinearity Test:



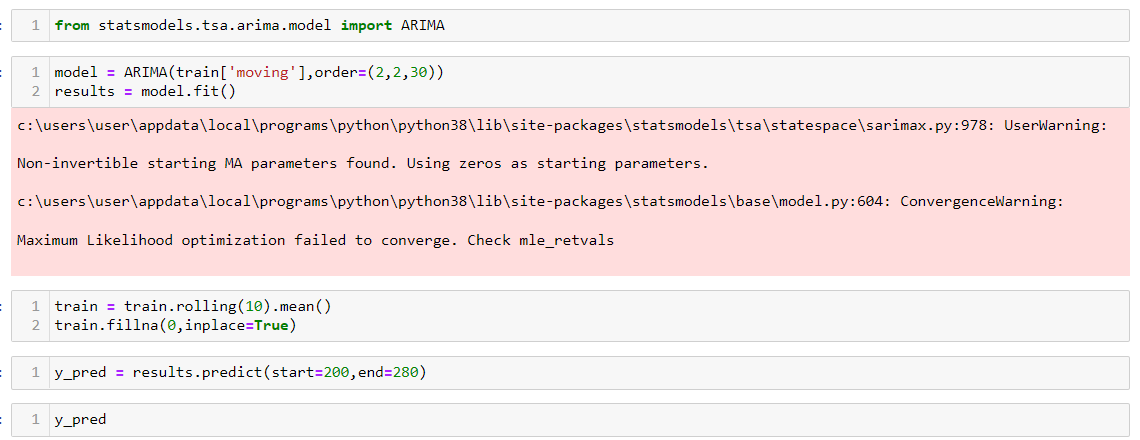
Importing models and Standardization:



Stationarity Test:



ARIMA Training and Prediction:



# Bibliography

GetSmarter.com. (2022). The Big Data Advantage in Real Estate Analysis. Retrieved from https://www.getsmarter.com/blog/career-advice/the-big-data-advantage-in-real-estate-analysis/

influxdata.com. (2022). Time-series Data.

Ozancan Özdemir, Ozancan. “House Price Prediction Using Machine Learning: A Case in Iowa.” 2022.

Yang, Julei. “A Hybrid Regression Technique for House Prices Prediction.” 2017.