**Predictive Analytics Final Project**

**Identifying Investment Opportunity (Zillow)**

****

**Georgia State University, Spring 2022**

**Predective Analytics - MSA 8200**

**(March 2022 – April 2022)**

**Table of Contents**

1. Introduction

1.1 General Introduction……………………………………………………..

1.2 Business Statement ………....

1. Exploratory Data Analysis (EDA)

2.1 Data Components……………………………………..

2.2 Data Cleeaning………………………………………..

2.3 Data Preprocessing………………………………

2.4 House Mean Value by Metro………………………………..

2.5 Home Value Analysis

2.6 Identifying the top 10 Zip codes based on highest ROI

1. Model Implementation

3.1 SARIMA………………………………………………

3.2 LSTM

1. Conclusion
2. Challenges and Future Scope
3. References

**CHAPTER 1 - INTRODUCTION**

**1.1 General Introduction**

In todays world people are tying to find variable sources to invest amnd make money , be it investing in stocks or investing in Cryptocurrency, there is also a viable source through which we can make money that is through investing in property or houses. We have a dataset for Zillow to do the trick, where the column ROI i.e return in investment in the particular area. Here, we'll be looking at how to apply predictive analytics techniques which we have learned in the class.

**1.2 Business Statement**

Which zip codes should one invest in California to maximize the predicted one-year return?​ Since California state has one of the most expensive houses in the US, so we want to analyze this state for the Return on Investment on houses in this state.​We used 5 years (2016-2021) Return On Investment rate to measure 'best'.

**CHAPTER 2 - EXPLORATORY DATA ANALYSIS**

**2.1 Data Components**

The Data consists of Zipcodes of CA state​, Zillow Home Value Index (ZHVI)**:** A smoothed, seasonally adjusted measure of the typical home value and market changes across a given region and housing type. It reflects the typical value for homes in the 35th to 65th percentile range.​We have gathered the data for the past 11 years (2011-2021) from Zillow's website.​

**2.2 Data Cleaning**

We dropped the NULL values for date columns​, and dropped the following columns - ​ 'RegionID' 'City', 'State’, ‘County Name’, ‘Region Type’, ‘State Name'. CA has 1575 zipcodes(after removing the ones which had missing values). We will narrow this to give us top 5 zipcodes one should consider investing based on forecasts of 1 year ROI



**2.3** **Data Preposessing**

We reshaped the Data format from Wide to long to make it time series compatible data​ and then used the melt function to implement the format change​. We also changed Zipcode dtype to 'str', 'Date' column is DateTime.

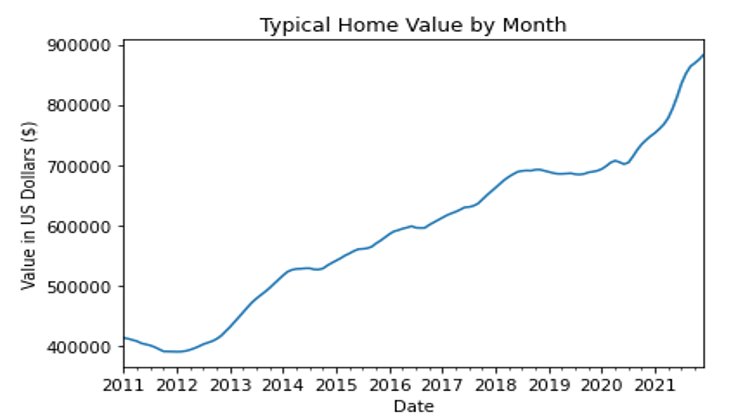
**2.4 House Mean Value by Metro​**

After grouping by metro, we calculated the mean value of the houses for analysis and found that San Jose-Sunnyvale-Santa Clara metro has the highest 11-year mean value



**2.5 Home Value Analysis**

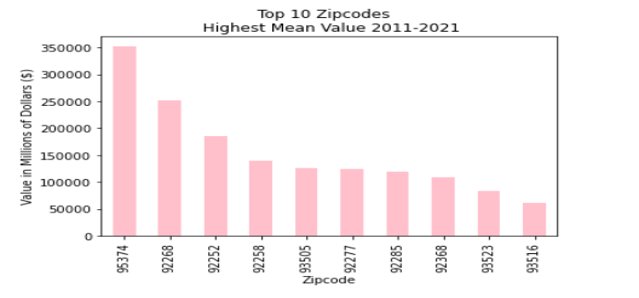
The 11 year average CA home price is $590375​ which is low but we looking at 11 yeras data and considering the prices before it tends to come out high, also this made us realise that taking more current data for the Analysis purpose.



**2.6 Identifying the top 10 Zip codes based on highest ROI​**

ROI was calculated using the last 5 year data​ and Top 10 Zip codes were identified using ROI, on which we implemented the timeseries for forecasting​. We analysed in the previous step that 5 year data would give us a better result.



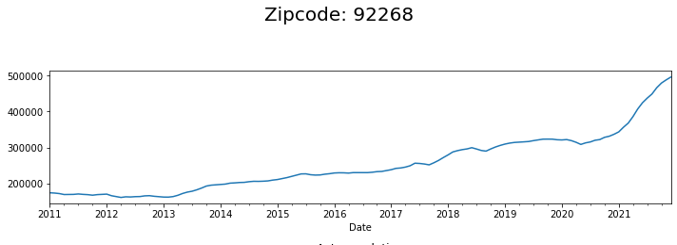


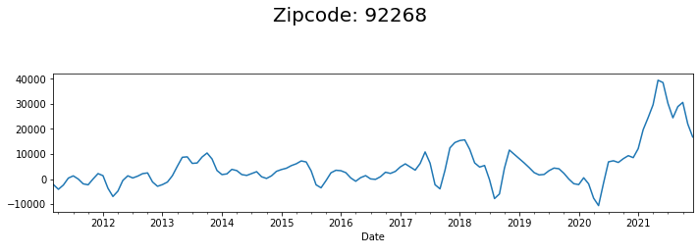
**CHAPTER 3- MODELS IMPLEMENTATION**

* **SARIMA**

A seasonal autoregressive integrated moving average (SARIMA) model is one step different from an ARIMA model based on the concept of seasonal trends. In many time-series data, frequent seasonal effects come into play. It adds three new hyperparameters to specify the autoregression (AR), differencing (I), and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

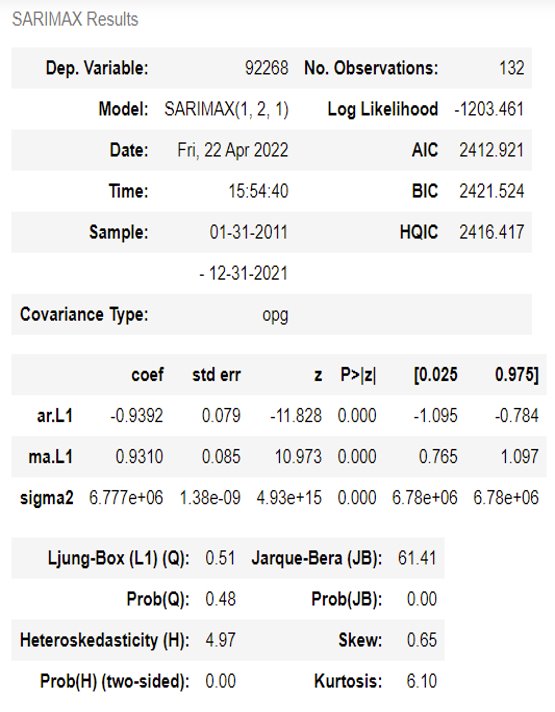
Through analysis we found out each zipcode is different in terms of its neighborhood varying in crime rates, pollution levels, public facilities, etc which is the reason we model each zip code separately​, we utilized the auto arima function to model each of the ten zip codes separately​, For instance, the time series plot in zipcode 92268 is depicted below. After taking the second difference, it is almost stationary.​

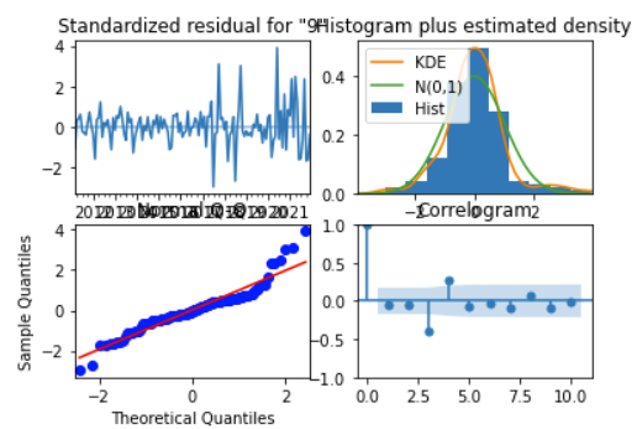




**​**

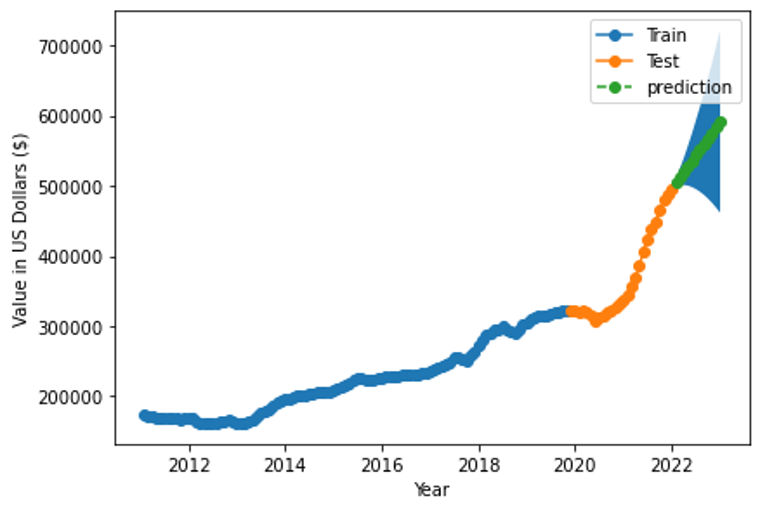
**Model Results for Zipcode 92268**

​ 



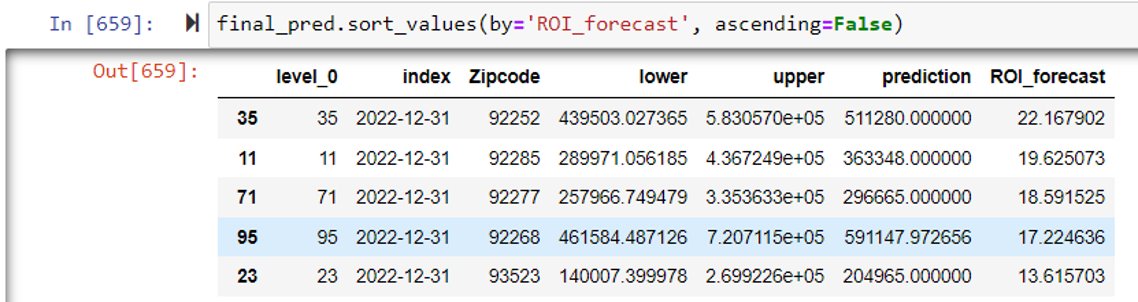
**Predictions for Zipcode 92268**

The predictions for the next 12 months – Jan 2022-Dec 2022 ​.Based on these predictions, we calculate the ROI forecast values for each zip code. For zipcode 92268, and get ROI forecast of 17.22%.



**Recommendation based on ROI calculated through 1-year forecast values**

As per our recommendation based on the time series analysis, the top 5 zipcodes in CA for investment purposes are 92252, 92285, 92277, 92268, 93523​.



* **LSTM**

Long short-term memory (LSTM) units or blocks are part of a recurrent neural network structure. Recurrent neural networks are made to utilize certain types of artificial memory processes that can help these artificial intelligence programs to more effectively imitate human thought.

—-----------------------------------



**CHAPTER 5- CONCLUSION**

**CHAPTER 6- CHALLENGES AND FUTURE SCOPE**

**CHAPTER 7- REFERENCES**