Time Series Modeling for Real Estate Investments: A Consultant's Guide to Identifying the Best Zip Codes using Zillow Research Data

1. BUSINESS UNDERSTANDING

Most people's wealth is mostly comprised of real estate, and this is particularly true for many American homeowners. The real estate market is influenced by a variety of variables, including governmental regulations, demographics of prospective buyers, affordability, disparities in housing access, location, cash flows, liquidity, and the overall status of the economy. For the buyers, the process may be laborious due to the numerous variables. In order to help investors choose which real estate possibilities to pursue, Naruto Consultants are working to develop a predictive time series model.

Main Objective

To develop a time series model that would predict the future prices of houses for Naruto Investments to invest in

Specific Objectives

- To act as a consultant for Naruto investment firm and provide a solid recommendation for the top 5 best zip codes for investment.
- To find change in house prices over time.

2. DATA UNDERSTANDING

This Dataset was obtained from Zillow Website and comprises of 14723 rows and 272 columns in Wide Format.

Dataset columns	**Column Description**	**Data type**
RegionID	Represents a unique ID for each region	integer (int64)
RegionName	Represents the name of the region/ also the zip code	integer (int64)
City	Represents the city where the region is located	string (object)
State	Represents the state where the region is located	string (object)
Metro	Represents the metropolitan area where the region is located (if applicable)	string (object)
CountyName	Represents the name of the county where the region is located	string (object)

SizeRank	Represents the relative size of the region compared to other regions in the dataset	integer (int64)
1996 upto 2018	Represents the median home price for the region in months and years	float (float64)

3. DATA PREPARATION

The data was stored in a CSV file and loaded into Python using the Pandas library as shown below:

	RegionID	RegionName	City	State	Metro	CountyName	SizeRank	1996-04	1996-05	1996-06	1996-07	1996-08	1996-09	1996-10	1996-11
0	84654	60657	Chicago	IL	Chicago	Cook	1	334200.0	335400.0	336500.0	337600.0	338500.0	339500.0	340400.0	341300.0
1	90668	75070	McKinney	TX	Dallas-Fort Worth	Collin	2	235700.0	236900.0	236700.0	235400.0	233300.0	230600.0	227300.0	223400.0
2	91982	77494	Katy	TX	Houston	Harris	3	210400.0	212200.0	212200.0	210700.0	208300.0	205500.0	202500.0	199800.0
3	84616	60614	Chicago	IL	Chicago	Cook	4	498100.0	500900.0	503100.0	504600.0	505500.0	505700.0	505300.0	504200.0
4	93144	79936	El Paso	TX	El Paso	El Paso	5	77300.0	77300.0	77300.0	77300.0	77400.0	77500.0	77600.0	77700.0

Tidying the data

We applied a pd.melt function to make the dataset easily readable and hence it can be used for modeling and exploration. The dataset was transformed from a wide format to a long format making it change to 3901595 rows and 11 columns. Our data cleaning process involved the following steps:

1. **Handling missing values**

We identified missing values in the Metro, %ROI, and ROI price columns, filling the null values on the Metro column with “Missing”. The remaining two were handled in the preprocessing stage.

2. **Checking for duplicates**

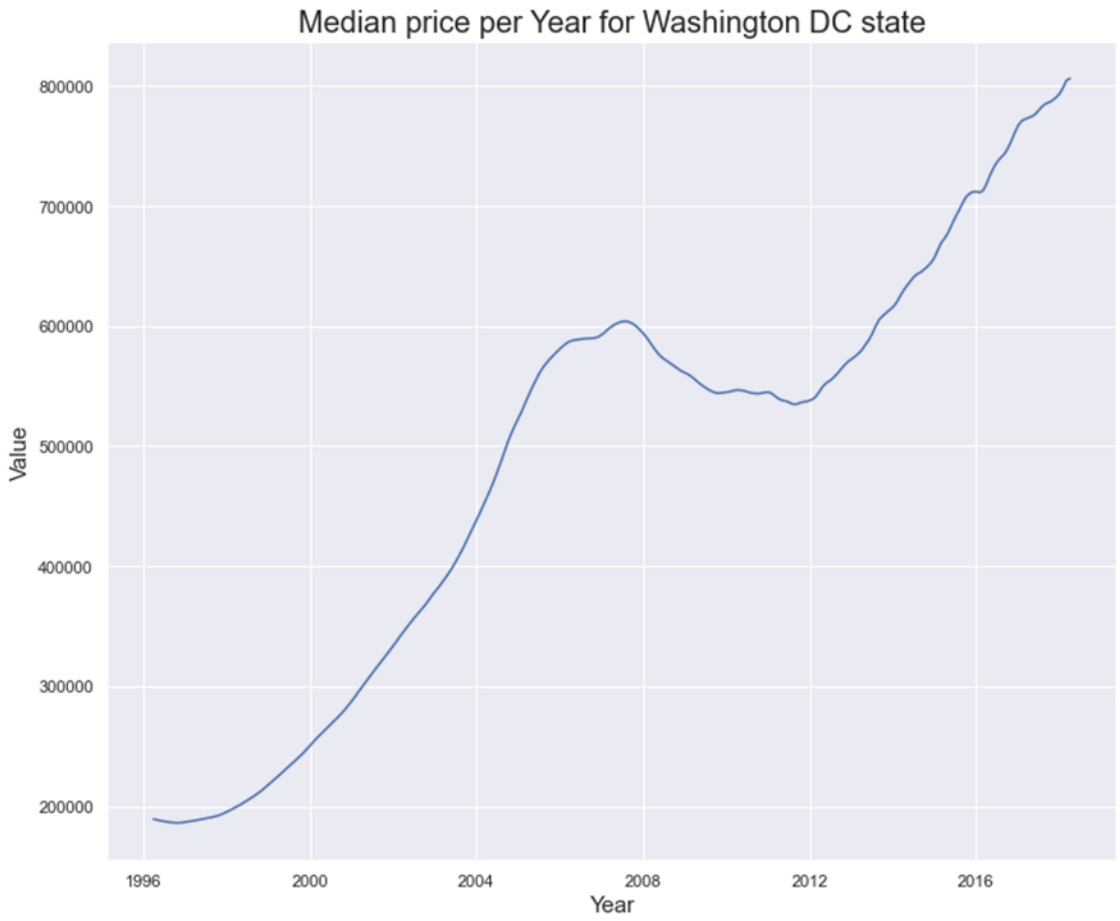
The dataset had no duplicates.

External Dataset Validation

The Zillow dataset was compared with external datasets from the National Association of Realtors (NAR), the US Census Bureau, and the Bureau of Economic Analysis (BEA). The NAR provides home sales and price data, while the US Census Bureau provides population data, and the BEA provides GDP data. Analysis was done on the trend of home prices in Washington from 2002 to 2018 using the Zillow dataset and compared it to the trend in the NAR dataset.

Also the impact of the 2008 housing market bubble burst on home prices in Washington was examined using the NAR dataset and compared it to Zillow's trend data. Additionally, analysis on the trend of population and GDP growth was done using the US Census Bureau and BEA data, respectively, to examine their impact on the real estate market in Washington and compared them to Zillow's data.

Based on the results, the Zillow dataset seems to be good to proceed with modeling. The findings can be used to make informed decisions and provide recommendations for the real estate industry in Washington, while keeping in mind the limitations of the data analysis and the need for further research to gain a complete picture of the real estate market in the US. The full article can be found here.



Exploratory Data Analysis

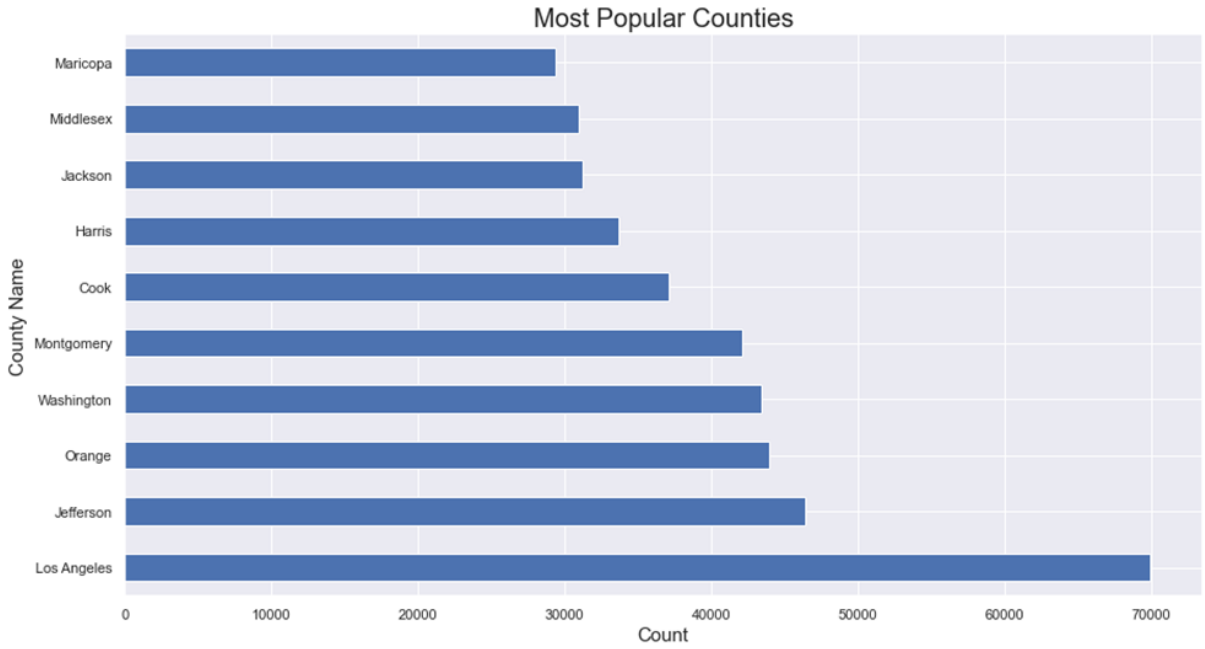
House prices have been trending upwards from 1996-2008 until the house market crash where the house prices drastically went down and stabilized around 2012.

Through the univariate analysis, we discovered:

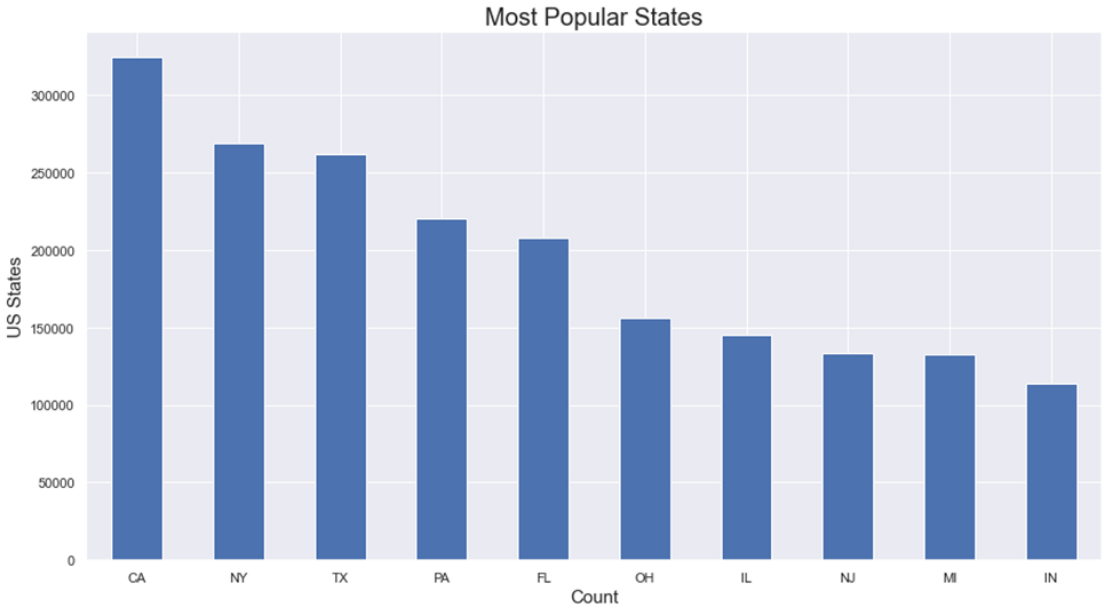
- Los Angeles is the most popular County in the dataset.
- California is the most popular state.
- New York is the most popular city followed by Los Angeles.

We answered the following questions:

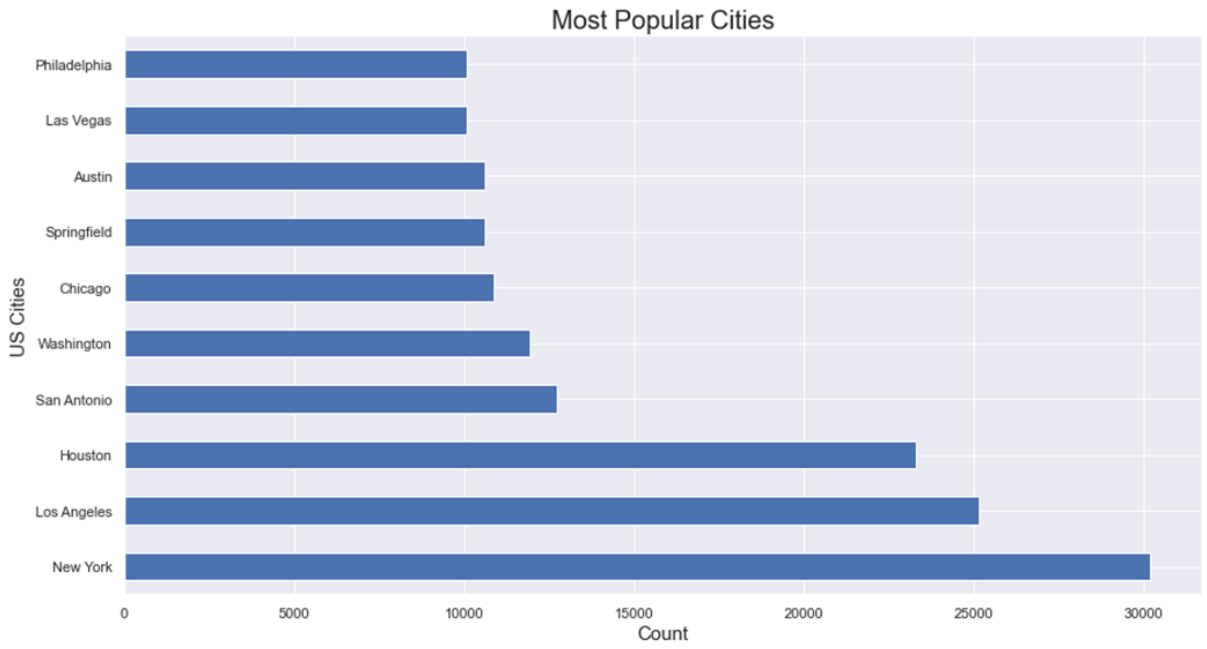
a. What are the top 10 most popular Counties?



b. What are the top 10 most popular States?

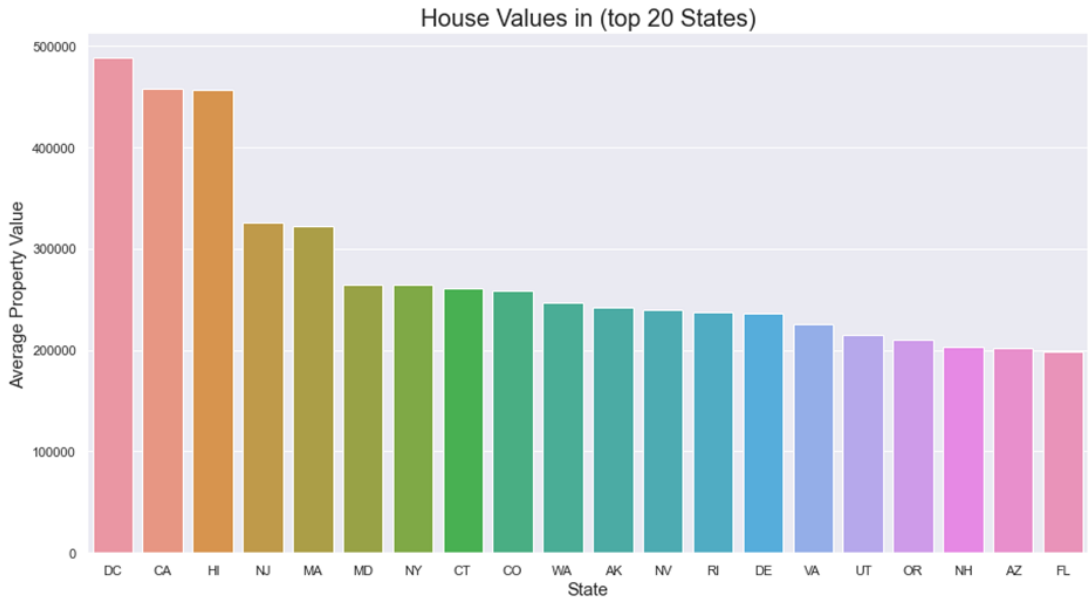


c. What are the top 10 most popular Cities?

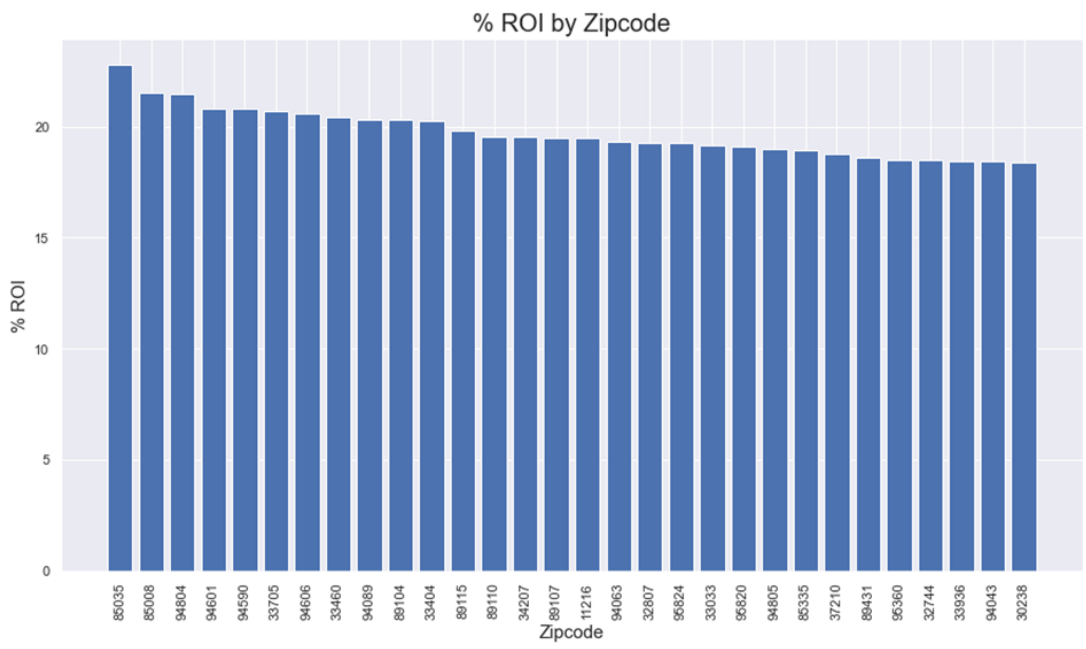


Through the bivariate analysis, we discovered California, Washington DC and New Jersey states have the highest average property value.

d. What is the average profit margin per state?

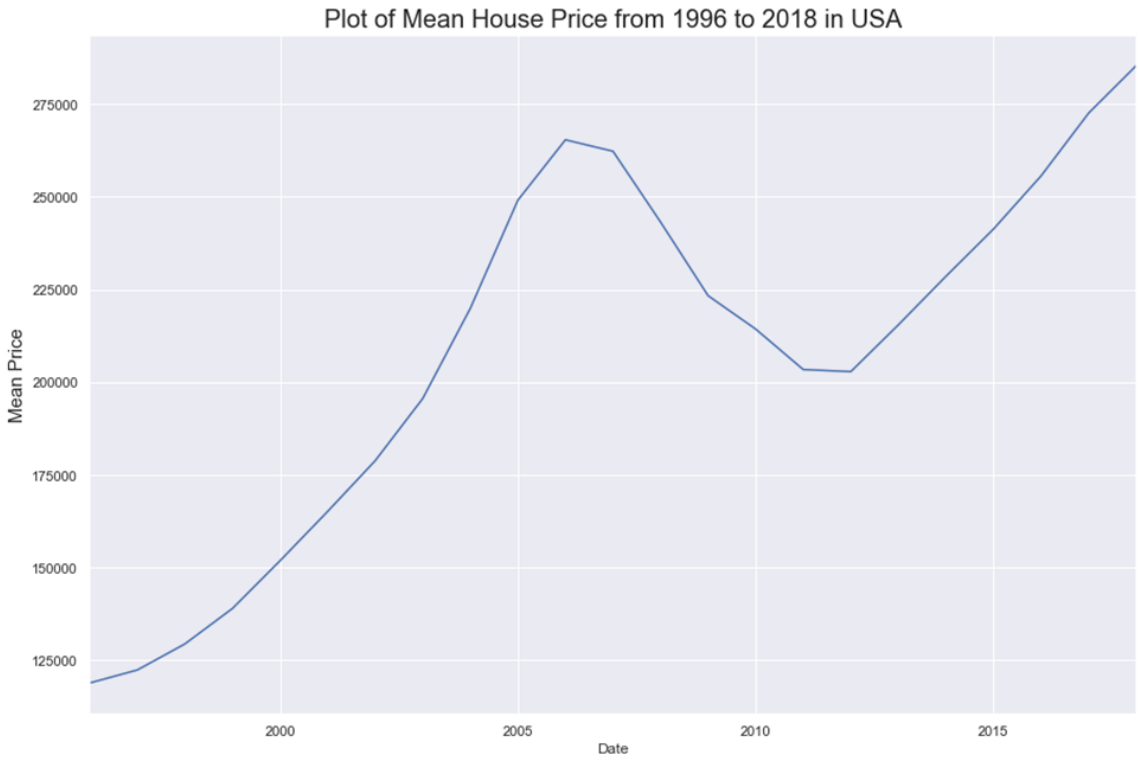


e. What is the mean percentage return on investment by zip code?



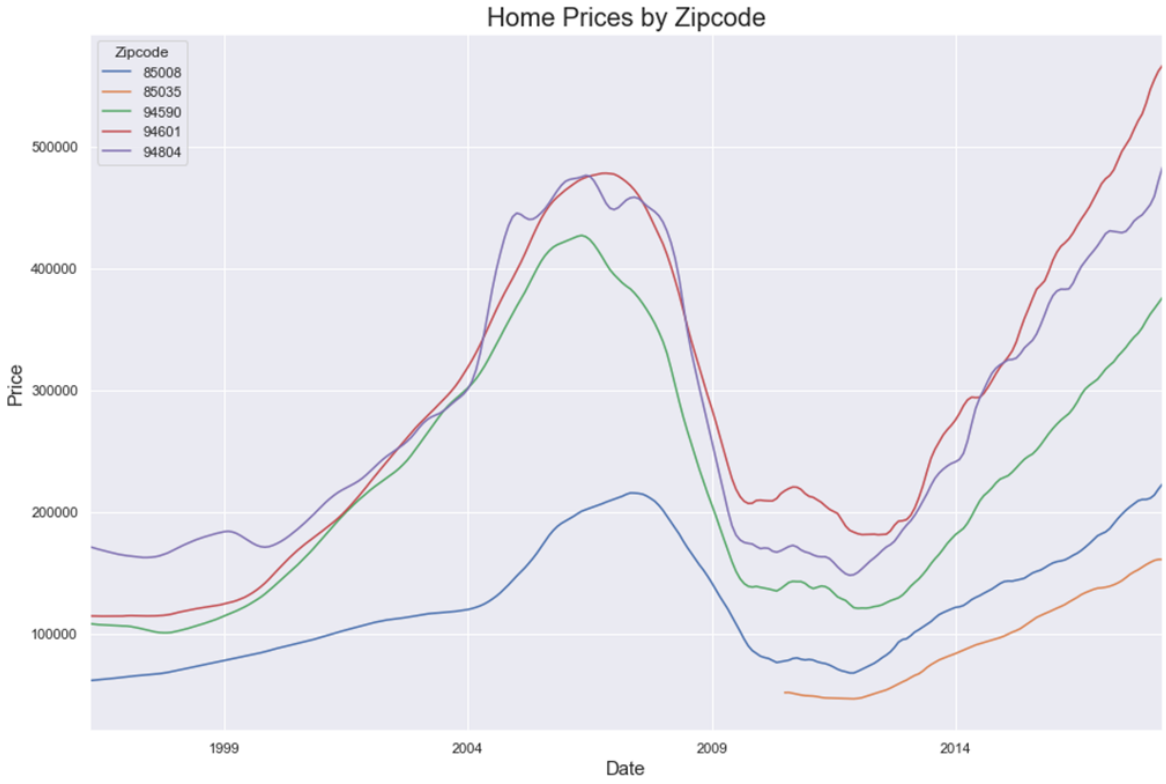
From this graph, zip code 85035 seems to be the most profitable zip code at 22.8% ROI from 2012 to 2018.

f. What is the change in house prices from 1996 to 2018?



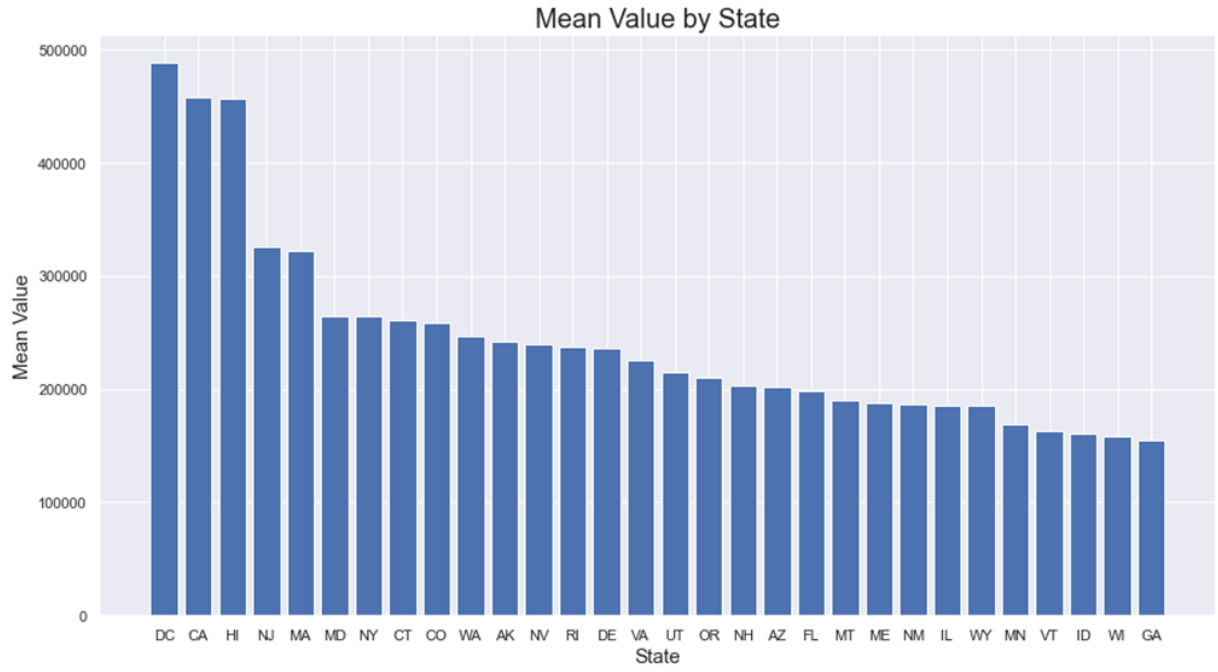
This shows that the house price had been trending upwards from 1996-2008 until the house market crash where the house prices drastically went down and stabilized around 2012. After this the house price has been trending upwards once again till 2018.

g. What is the time series data for the top 5 zip codes ranked by percentage return on investment?



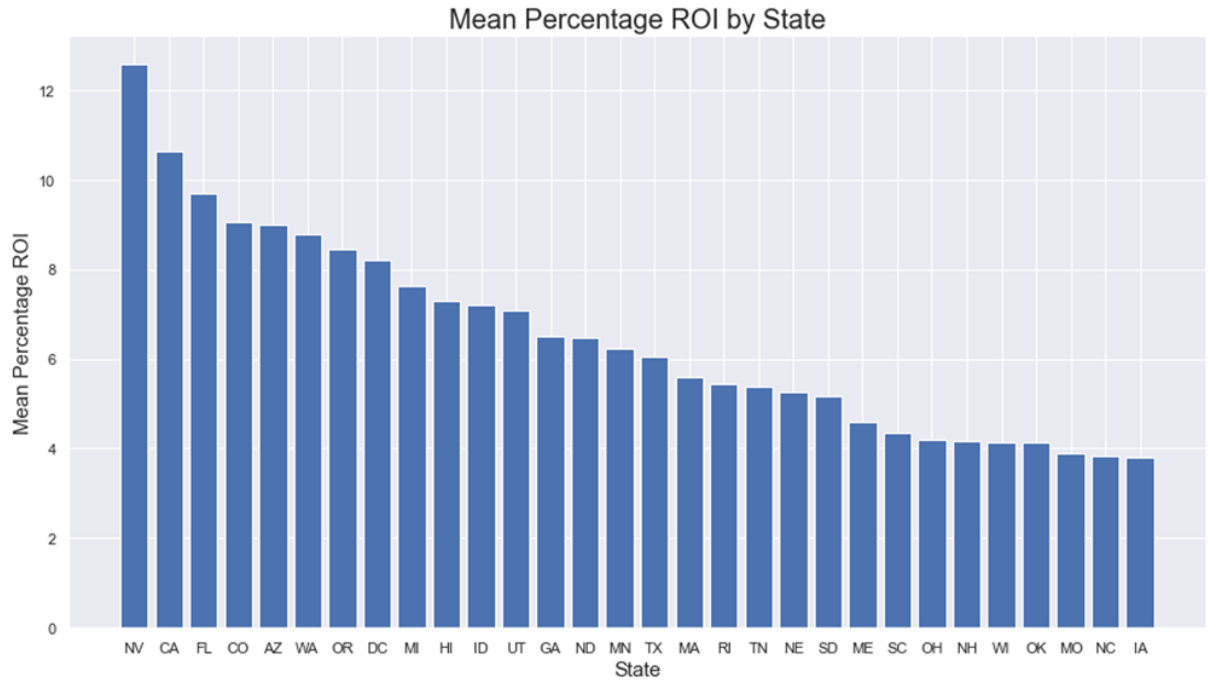
As observed, the percentage of return on investents for the top five zip codes had a crash in 2008 to 2012 and then from there they all have an upward linear trend.

h. What is the mean price of houses by state?



From the above graph, New Jersey has the highest mean value by State

i. What is the mean percentage return on investment by state?



Nevada has the highest mean percentage return on investment

Data Preprocessing

An assumption that the data is stationary is made when dealing with time series models. However, we verified that the data is stationary by performing the Dickey Fuller test and also using the Rolling mean.

4. MODELING

Our aim was to identify the top five zip codes to invest in. Hence, we used five different models for each of the zip codes to forecast their prices thus giving the investors an informed decision. We arrived to ARIMA(2,0,1)(0,0,0)[0] intercept as the best model with a total fit time of 2.481 seconds.

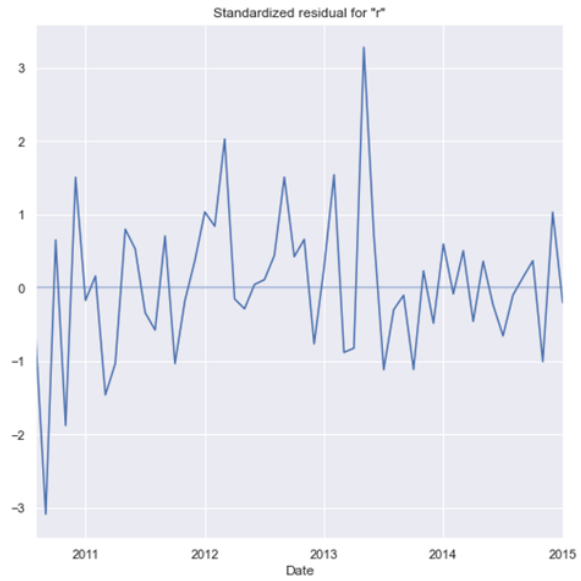
We then fit an ARIMA Model on the training series using parameters retrieved from the AUTO ARIMA model. The results were as shown below:

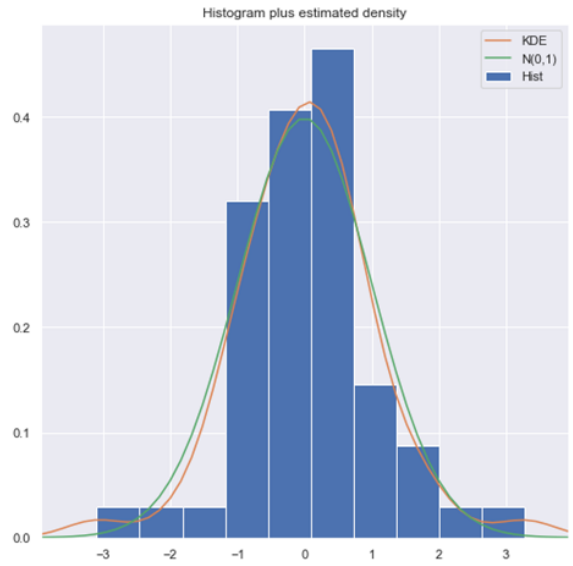
SARIMAX Results

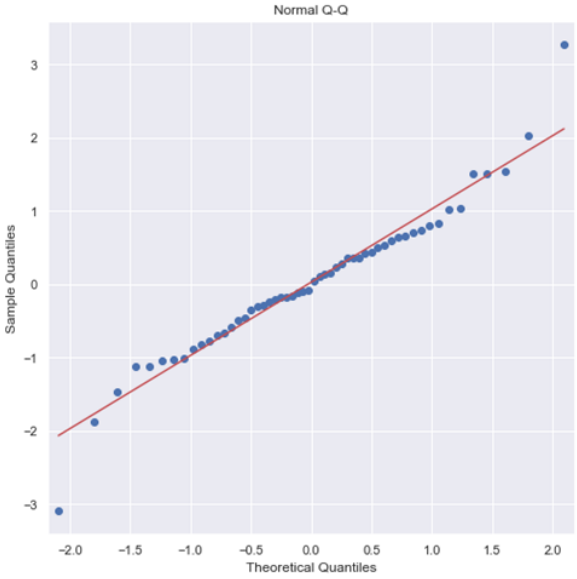
Dep. Variable:	ret	No. Observations:	54
Model:	ARIMA(2, 0, 1)	Log Likelihood	210.348
Date:	Thu, 16 Mar 2023	AIC	-410.696
Time:	20:32:05	BIC	-400.751
Sample:	08-01-2010	HQIC	-406.860
	- 01-01-2015		
Covariance Type:	opg		

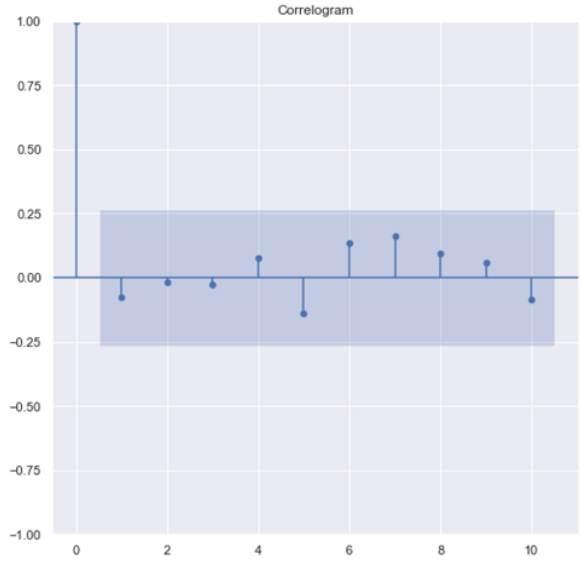
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0115	0.008	1.456	0.145	-0.004	0.027
ar.L1	0.4476	0.245	1.827	0.068	-0.033	0.928
ar.L2	0.4027	0.242	1.661	0.097	-0.072	0.878
ma.L1	0.8430	0.186	4.536	0.000	0.479	1.207
sigma2	2.304e-05	3.5e-06	6.572	0.000	1.62e-05	2.99e-05

Ljung-Box (L1) (Q):	0.35	Jarque-Bera (JB):	8.82
Prob(Q):	0.55	Prob(JB):	0.01
Heteroskedasticity (H):	0.23	Skew:	0.12
Prob(H) (two-sided):	0.00	Kurtosis:	4.97









The residuals should not be correlated and they should have a normal distribution to satisfy the normality assumptions.

- The qq- plot on the bottom left shows that the residuals follow a linear trend line hence they are normally distributed.
- The correlogram plot on the bottom left show there are low correlations with their lagged version. This tells us that there isn't any obvious seasonality in our series.

- The histogram has a bell curve showing that the residuals are normally distributed which is a good thing.

5. EVALUATION

We recorded the mean squared error as 0.00508602223386915. This tells us that our monthly returns would be off by 0.005% if this model is used, which is a good thing since it is almost negligible.

Conclusion and Recommendations

We concluded that all the zip codes have an encouraging predicted price seeing as they are in the positive, apart from the 85035 zip code. The investor can therefore decide to invest in zip codes 94804, 94590, 94601, and 85008 as they have a positive return on investment.

Follow-up Questions

- Did we have the right data? Yes, the data was correct and verified
- Do we need other data to answer our question? Yes, it would be effective if we had additional information apart from the price to determine whether other socio-economic factors would affect decisions on investments in certain areas or not. Facilities like schools and hospitals, and level of security can affect whether one would invest in a given area or not.
- Did we have the right question? The question chosen was correct.