PHASE 4 PROJECT :REAL ESTATE TIME SERIES ANALYSIS.

1). Business Understanding

Real estate investment is a lucrative and dynamic industry that requires careful analysis and decision-making. The fictional real estate investment firm is seeking guidance on identifying the top 5 zip codes for investment opportunities. To address this question, historical data from Zillow Research is utilized.

i) Background:

Real estate investment is a lucrative and dynamic industry that requires careful analysis and decision-making. The fictional real estate investment firm is seeking guidance on identifying the top 5 zip codes for investment opportunities. To address this question, historical data from Zillow Research is utilized. The dataset contains information on various attributes, including RegionID, RegionName, City, State, Metro, SizeRank, CountyName, and value (real estate prices).

ii). Main Objective:

The main objective of this project is to identify the top 5 zip codes that offer the best investment potential in terms of real estate prices. By analyzing historical trends and patterns, the project aims to provide actionable insights to the investment firm, enabling them to make informed decisions on where to allocate their resources.

Specific Objectives:

- Analyze Historical Data: The project involves analyzing the historical data of real estate prices across different zip codes. This includes understanding the trends, patterns, and fluctuations in property values over time.
- Identify Promising Zip Codes: Using the analysis of historical data, the project aims to identify the zip codes that have shown consistent growth, stability, or potential for future appreciation. These zip codes are considered the most favorable for investment.
- Consider Location Factors: In addition to the historical performance, the project also takes into account location-specific factors such as city, state, and metro. This information helps assess the overall desirability and attractiveness of the investment opportunities.
- Evaluate Market SizeRank: The SizeRank attribute provides insights into the relative size and competitiveness of the real estate market in each zip code. This factor helps gauge the potential opportunities and risks associated with investing in a particular area.

2). Data Understanding

The dataset contains information on various attributes, including RegionID, RegionName, City, State, Metro, SizeRank, CountyName, and value (real estate prices). Our dataset is the Zillow Housing Dataset which was sourced from Zillow Research Page.

In order to understand how our dataset looks like lets get a preview of this data by loading it. Below are the column names in our dataset.

- RegionID -This is unique Id for the Regions
- SizeRank -This is the ranking done based on the size of the region
- RegionName - This field contains the zip code of the region.
- RegionType- Type of region is Zip.
- StateName - State

- City - This column provide the specific City Name of Housing Data
- Metro - This provide the name of the metro city around that region
- County Name - This is the county name for that region
- Months Column - These columns contains the prices of region for every month

3). Data Preparation

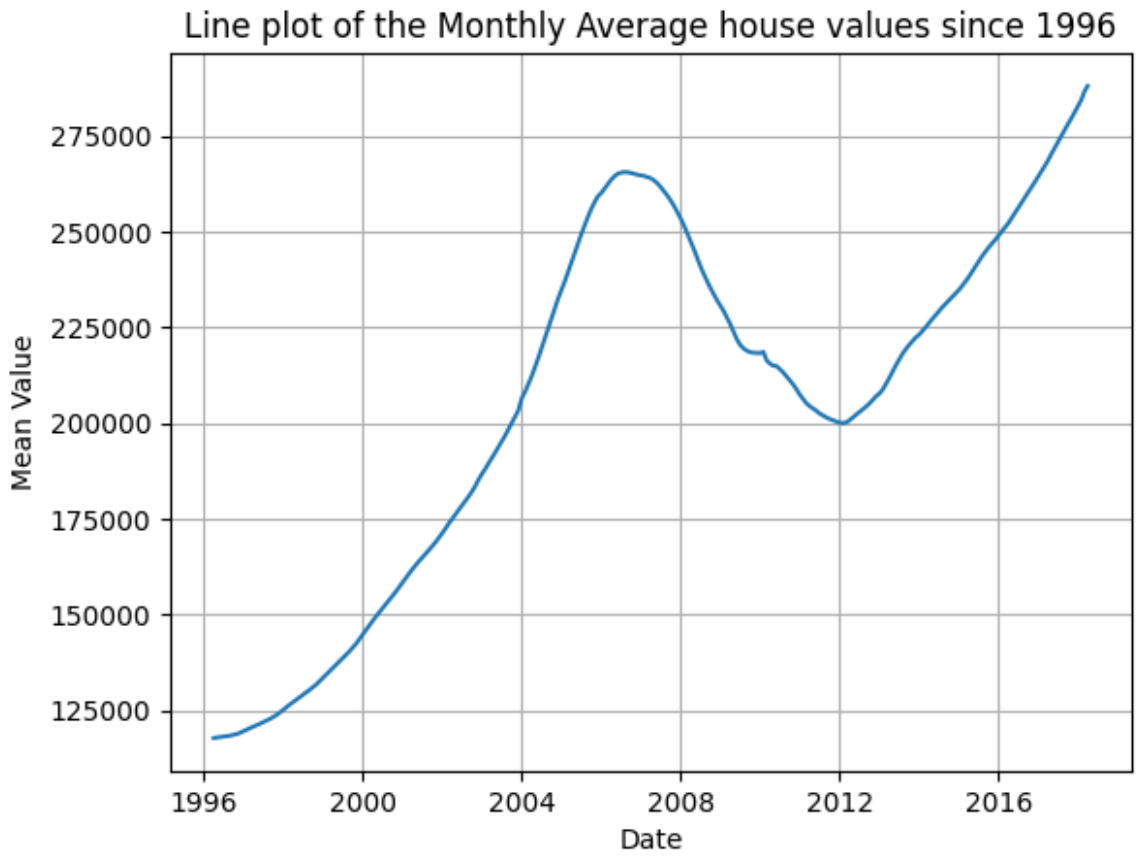
This is to make the data in a format that is good to feed to our model.It involves the following series of steps:

- Cleaning the data
- Checking for and dealing with missing values
- Reshaping our dataset from wide to long format

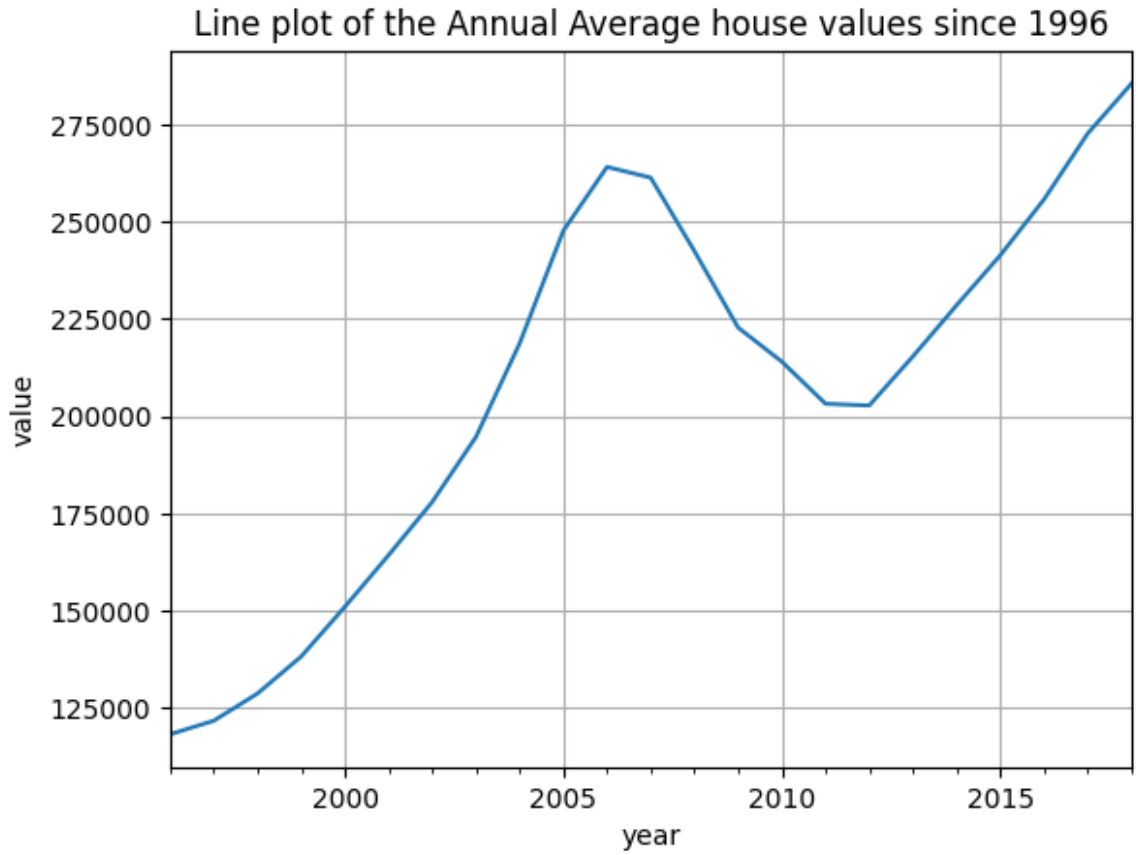
4). Exploratory Data Analysis

This is basically trying to figure out more about our data , its behaviours and patterns This involves the following :

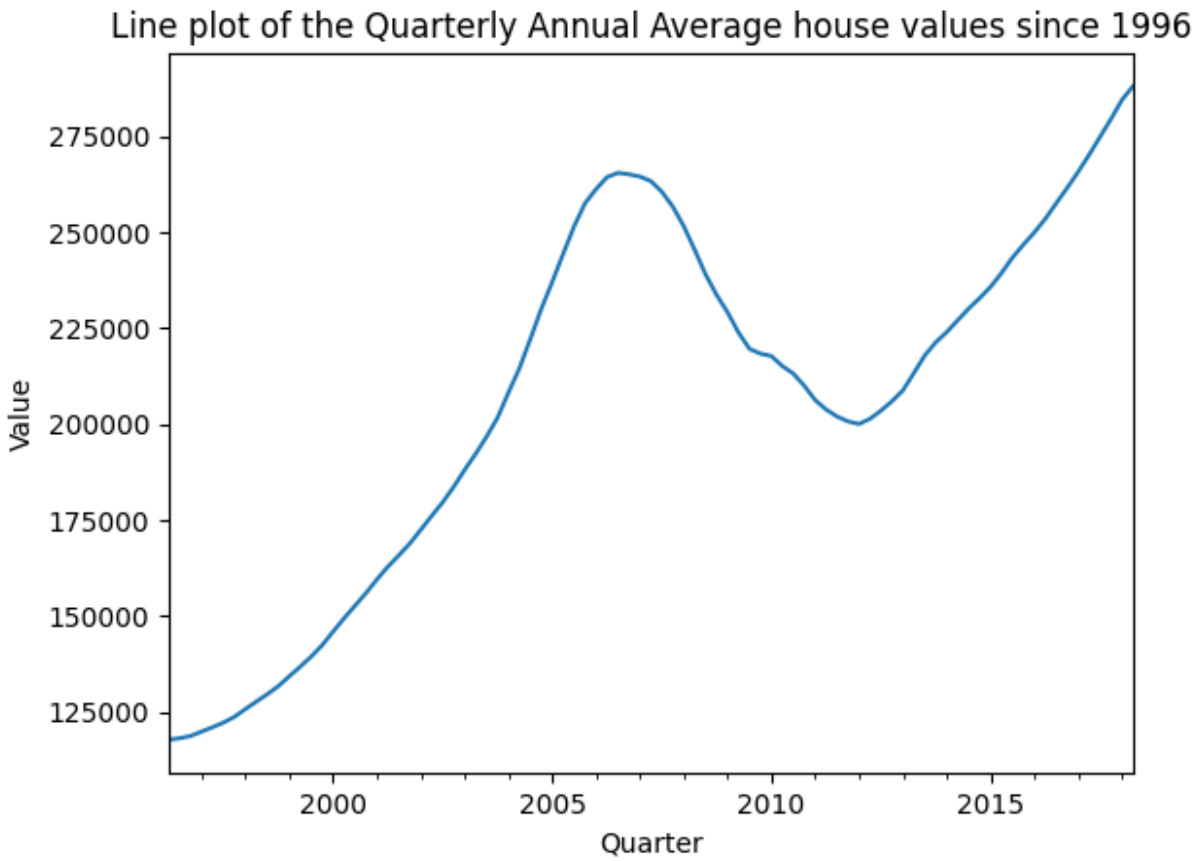
- Grouping the data by month.



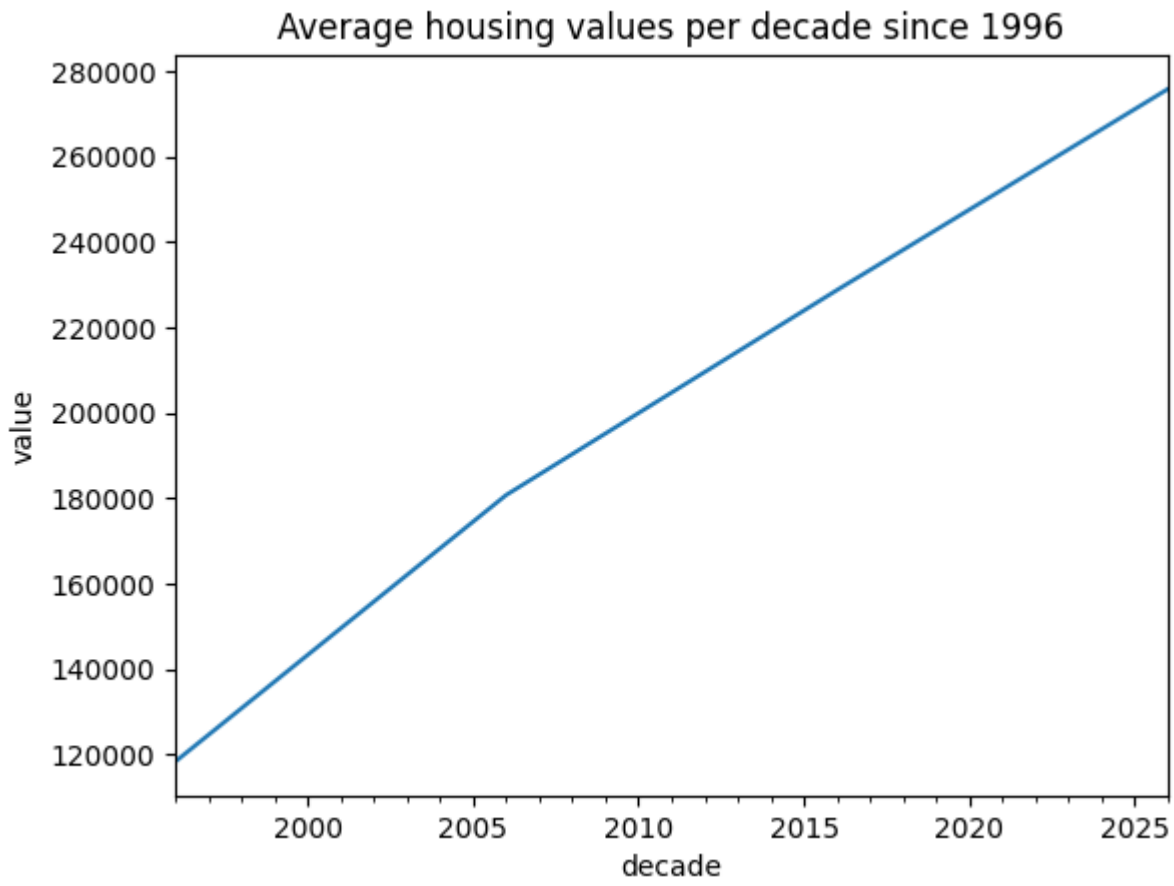
- Group the data yearly .



- Grouping per quarter and plotting .

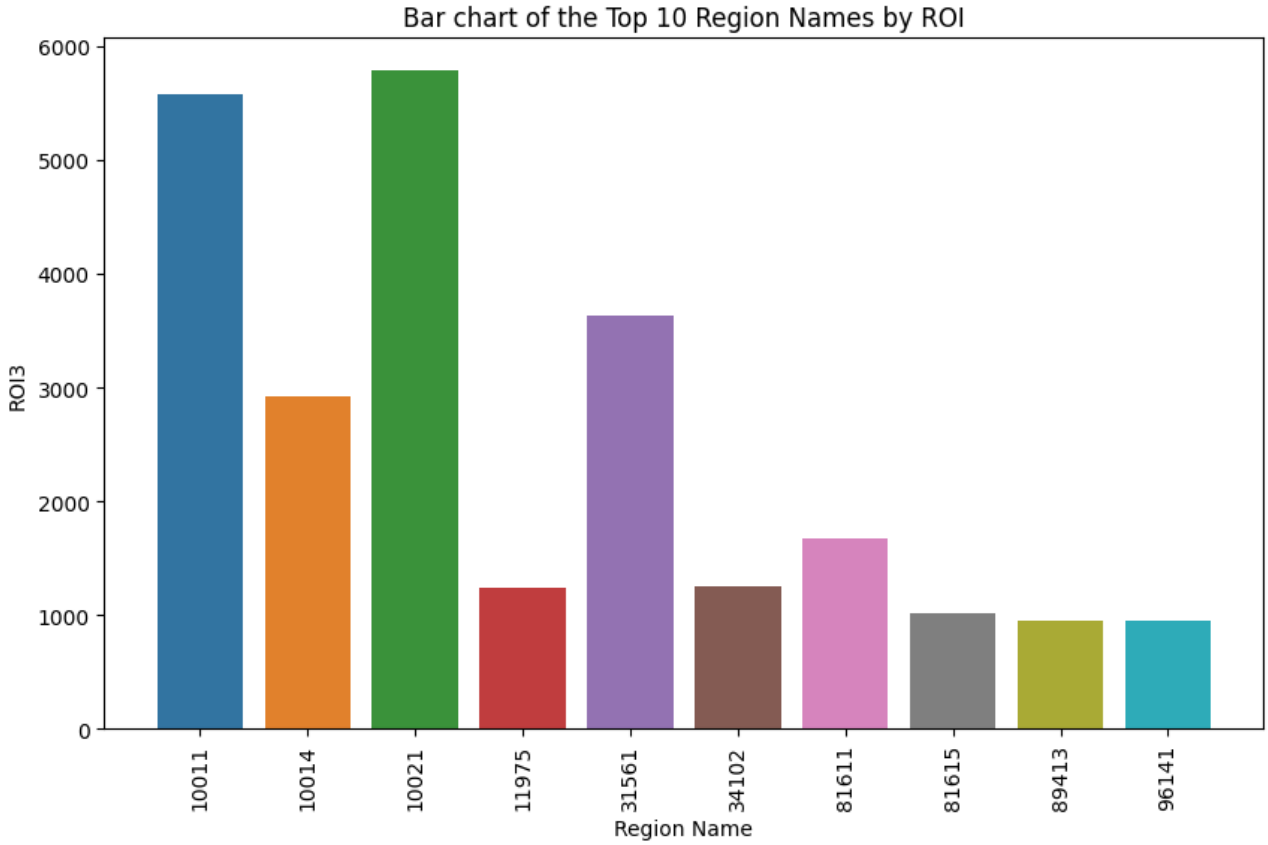


  - Grouping per decade

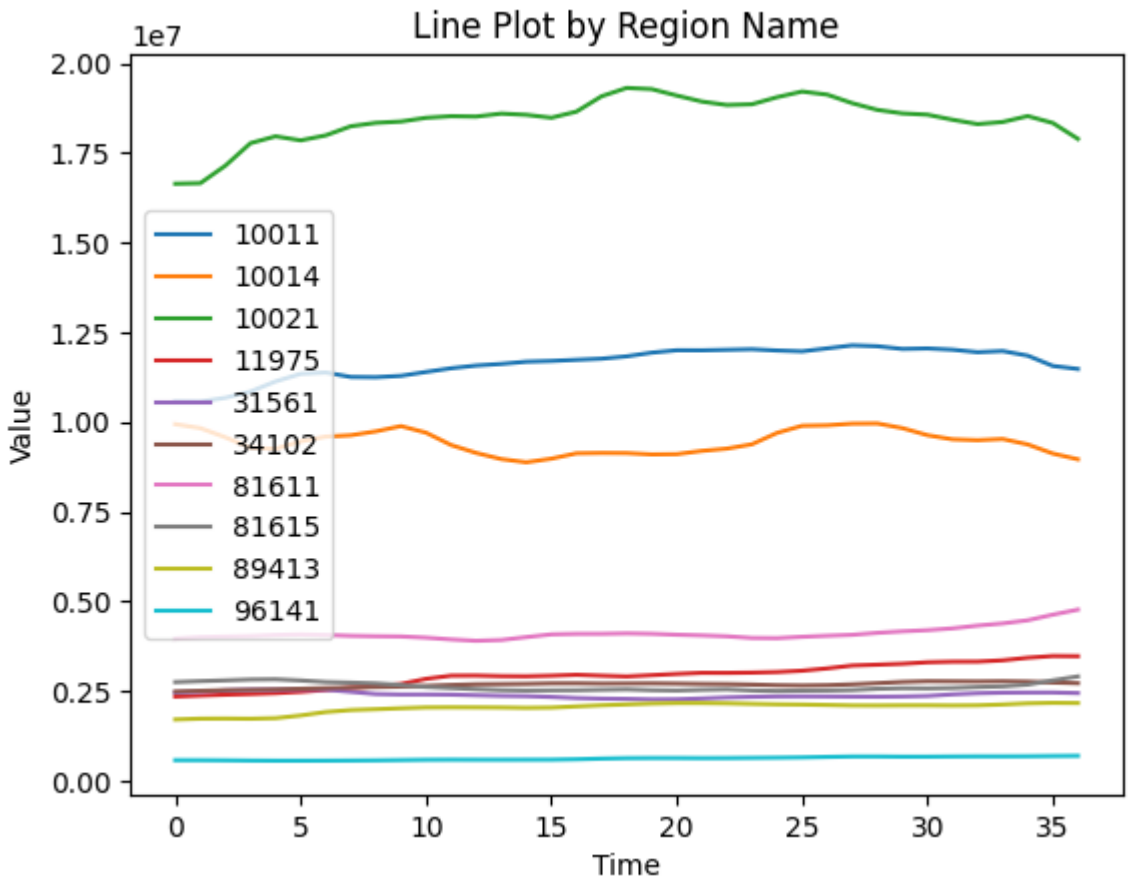


- Finding the top 5 Best regions

Here we seek to find the five best regions by uing the return of investment, where high returns show the best regions. Below is a plot of our findings



- Checking for trends and seasonality Here we seek to find the relevant trends in the dataset , based on the best regions, below is our plots



- Checking for the rolling statistics
- Performing Dullers test.
- Checking for stationarity
- Detrending our dataset
- Deseasonalizing our dataset
- Performing Seasonal decomposition

5). Modelling

This is now creating various models to forecast our data . we created the following models :

- ARIMA modelels
- SARIMAX models.
- PROPHET models.
  First weplot the auto-correlation plots , then we do the modelling.

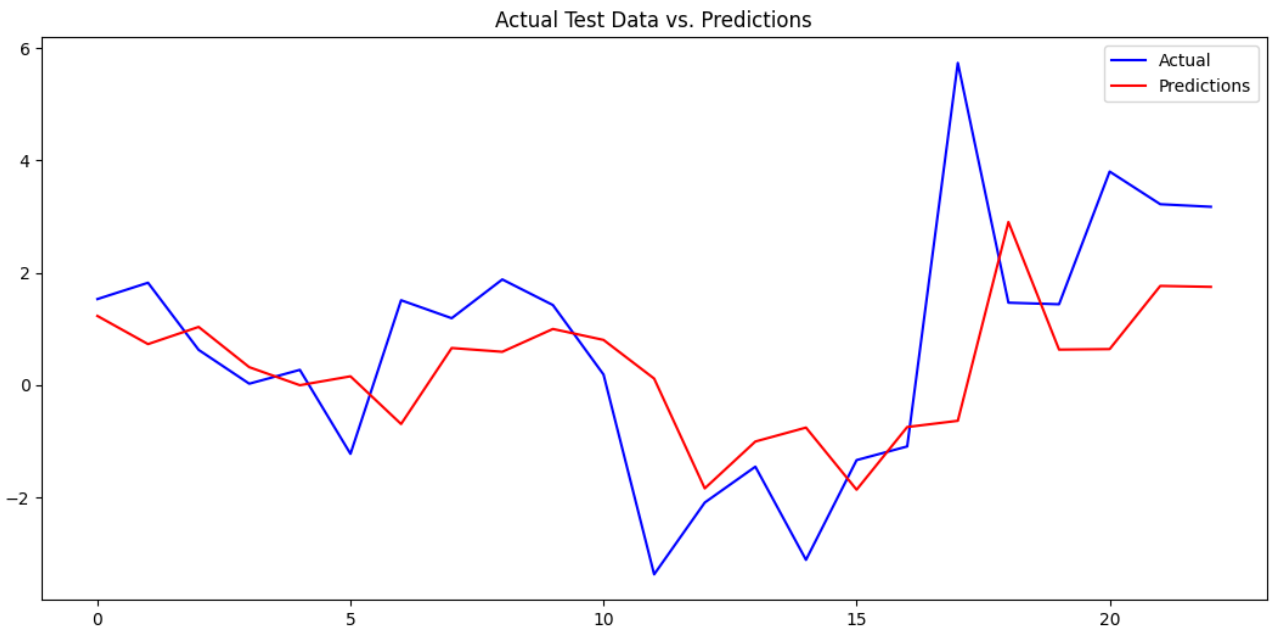
1. **Arima Models**
   We created an Arima model, below is the statistical results.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:               seasonal   No. Observations:                   36
Model:                 ARIMA(1, 1, 1)   Log Likelihood                -111.620
Date:                Tue, 20 Jun 2023   AIC                            229.239
Time:                        17:11:03   BIC                            233.905
Sample:                    05-01-2015   HQIC                           230.850
                         - 04-01-2018
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.1313      0.297     -0.441      0.659      -0.714       0.452
ma.L1         -0.6924      0.213     -3.245      0.001      -1.111      -0.274
sigma2        33.6579      8.259      4.075      0.000      17.471      49.845
===================================================================================
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):                 0.41
Prob(Q):                              0.82   Prob(JB):                         0.81
Heteroskedasticity (H):               1.58   Skew:                            -0.23
Prob(H) (two-sided):                  0.44   Kurtosis:                         3.27
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

2. **Sarimax Models**
   Below is the statistical result using the Sarimax models.

```
                               SARIMAX Results
==========================================================================
Dep. Variable:               seasonal   No. Observations:                   28
Model:               SARIMAX(1, 0, 1)   Log Likelihood                 -85.317
Date:                Tue, 20 Jun 2023   AIC                            176.635
Time:                        17:27:31   BIC                            180.631
Sample:                    05-01-2015   HQIC                           177.856
                         - 08-01-2017
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.7520      0.636      1.183      0.237      -0.494       1.998
ma.L1         -0.6145      0.726     -0.847      0.397      -2.037       0.808
sigma2        25.8880      6.345      4.080      0.000      13.452      38.324
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 6.55
Prob(Q):                              0.95   Prob(JB):                         0.04
Heteroskedasticity (H):               1.94   Skew:                            -1.12
Prob(H) (two-sided):                  0.34   Kurtosis:                         3.75
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Below is a plot of the actual and predicted values of our Sarimax model with a 1.94 RMSE



3. ** Prophet Models** We used this model to forecast and predict values, below are some of our predicted values (**Note: The yhat means the predicted values**)

	ds	yhat	identifier
0	2015-04-01	1.060583e+07	10011
1	2015-05-01	1.060216e+07	10011
2	2015-06-01	1.067983e+07	10011
3	2015-07-01	1.083111e+07	10011
4	2015-08-01	1.108970e+07	10011

7). Summary

After performing time series analysis on the 10 zip codes and forecasting total returns for up to three years, we reccomend the company to invest in the following 3 zipcodes:

- 81611 - Location: Aspen, CO (R.O.I - 132.378817)
- 10021 - Location: New York, NY (R.O.I - 111.795552)
- 34102 - Location: Naples, FL (R.O.I - 7.605307)

As for the other 6 zip codes, they are not fit for investement given the negative returns.