Kelvin-Rotich824 /
Phase_4_Project_Time_Series_Analysis

Code Issues Pull requests Actions Projects Wiki Security Insights

master Phase_4_Project_Time_Series_Analysis / Go to file t Add file ···

File successfully deleted.

Kelvin-Rotich824 Delete Phase 4 Project Notebook.ipynb now

Name	Name	Last commit date
.ipynb_checkpoints	Completed the project	2 hours ago
Images	images	yesterday
PHASE 4 GROUP PROJECT...	Finished project	4 minutes ago
Phase 4 Project Notebook ...	Finished project	4 minutes ago
Project presentation.pdf	Finished project	4 minutes ago
README.md	Update README.md	17 minutes ago
arima_model.pkl	Completed the project	2 hours ago
zillow_data.csv	Initialized repository	4 days ago

README.md

TIME SERIES MODELING FOR ZILLOW REAL ESTATE PRICES.

Problem Statement

To develop a time series model that can be used to predict and help determine the top five zip codes in which to invest.

Project Overview

The dataset encompasses details on a range of attributes, including RegionID, RegionName, City, State, Metro, SizeRank, CountyName, and the value representing real estate prices. This dataset, known as the Zillow Housing Dataset, has been obtained from the Zillow Research Page.

Business Understanding

Real estate investment stands as a profitable and ever-evolving industry, demanding meticulous analysis and strategic decision-making. A fictitious real estate investment firm is currently in search of insights to pinpoint the top five zip codes offering promising investment opportunities. To tackle this inquiry, we leverage historical data sourced from Zillow Research.

Components

- **Jupyter Notebook** The [Jupyter Notebook](https://github.com/Kelvin-Rotich824/Phase_4_Project_Time_Series_Analysis/blob/master/Phase 4 Project Notebook.ipynb) Our key deliverable contains details of our approach and methodology, data cleaning, exploratory data analysis, and model building and validation.

I recommend using nbviewer to view the Jupyter Notebook.

- **Presentation** The presentation gives a high-level overview of our approach, findings and recommendations for non-technical stakeholders. It is aimed to be between 5 and 10 minutes long.

- **Data**

The dataset can be found in the file "*zillow_data.csv*" in the Data folder, in this repository. It was originally provided in the following [repository](https://github.com/Kelvin-Rotich824/Phase_4_Project_Time_Series_Analysis/blob/master/Phase 4 Project Notebook.ipynb.

Technologies/ Packages

- Python version: 3.11.9
- Matplotlib version: 3.1.3
- Seaborn version: 0.9.0
- Pandas version: 0.25.1
- Numpy version: 1.16.5
- Statsmodels version: 0.10.1
- Scikit-learn version: 0.21.2
- TensorFlow version: v2.15.0

To get started

1. Clone this repository - guidance.
2. Dataset can be found in the file "zillow_data.csv".
3. Check the requirements in the Technologies section above and download libraries if necessary.

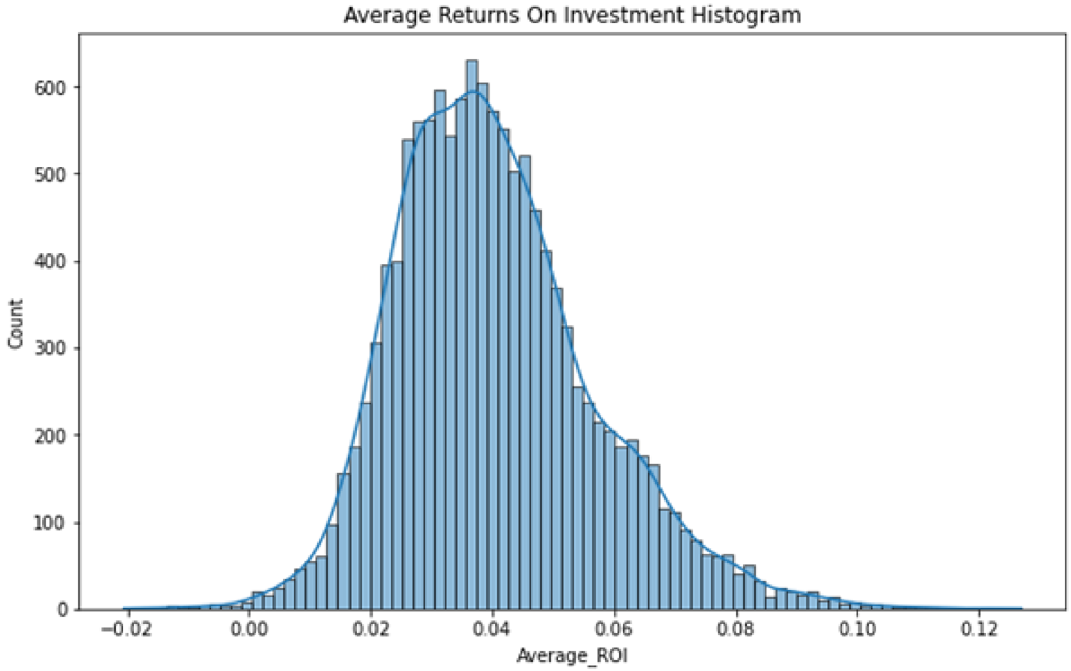
1. Data Wrangling

Here we will work on data cleaning, handling missing values, data transformation, handling duplicates, data reshaping, and other processes to ensure that we have a clean, structured, and suitable format for analysis and modeling

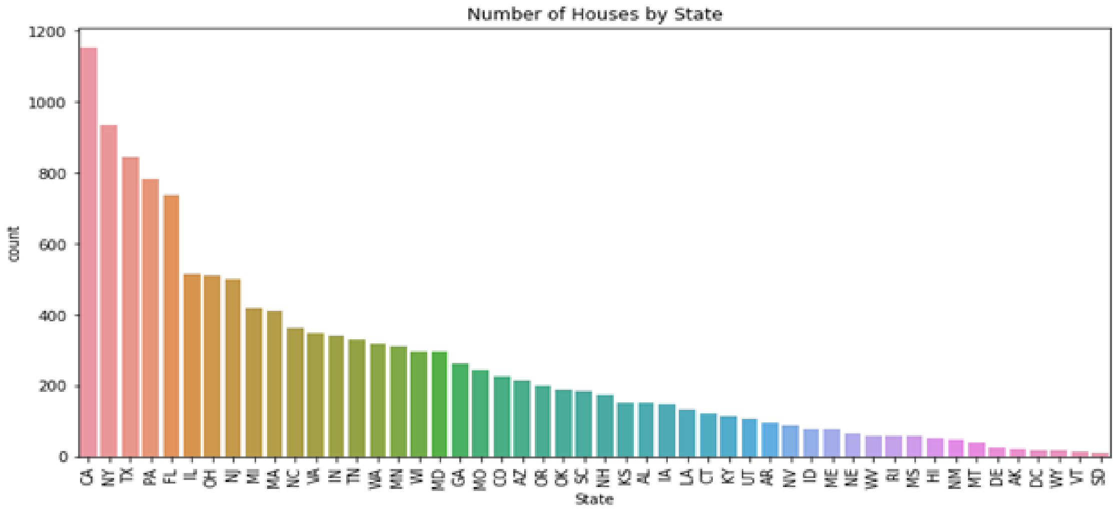
2. Exploratory Data Analysis (EDA)

Here we will explore the different features of the dataset to gain a better understanding of the data. We will use data visualization to uncover trends and patterns. We will use Feature Engineering to create new features from existing ones and perform One-Hot Encoding on categorical variables that we will require for analysis.

2.1 Univariate Analysis



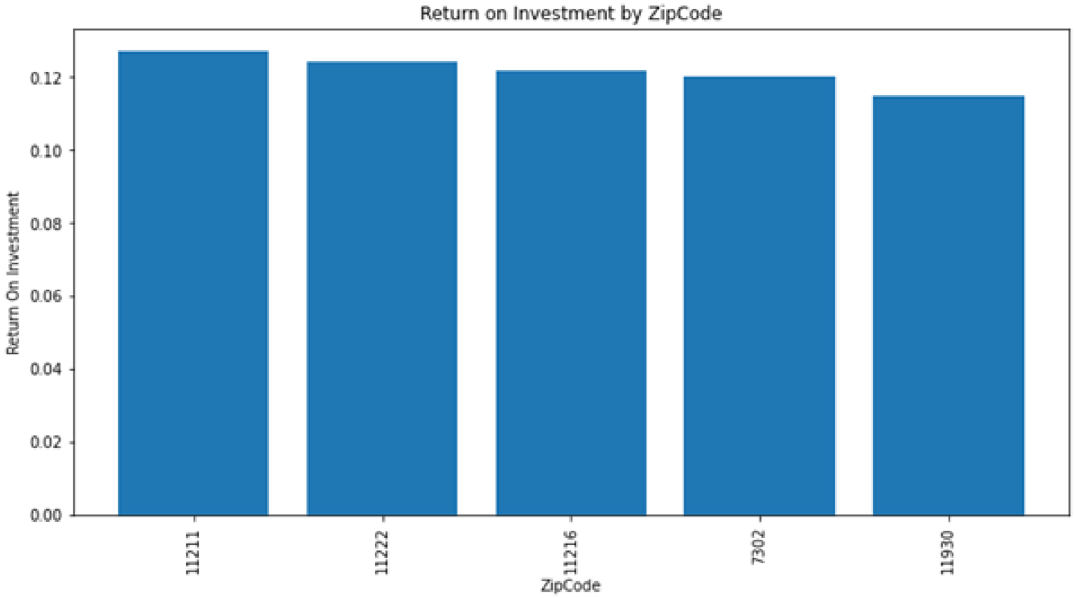
From the histogram above, we can conclude that the distribution of the average returns on investment has a normal distribution.



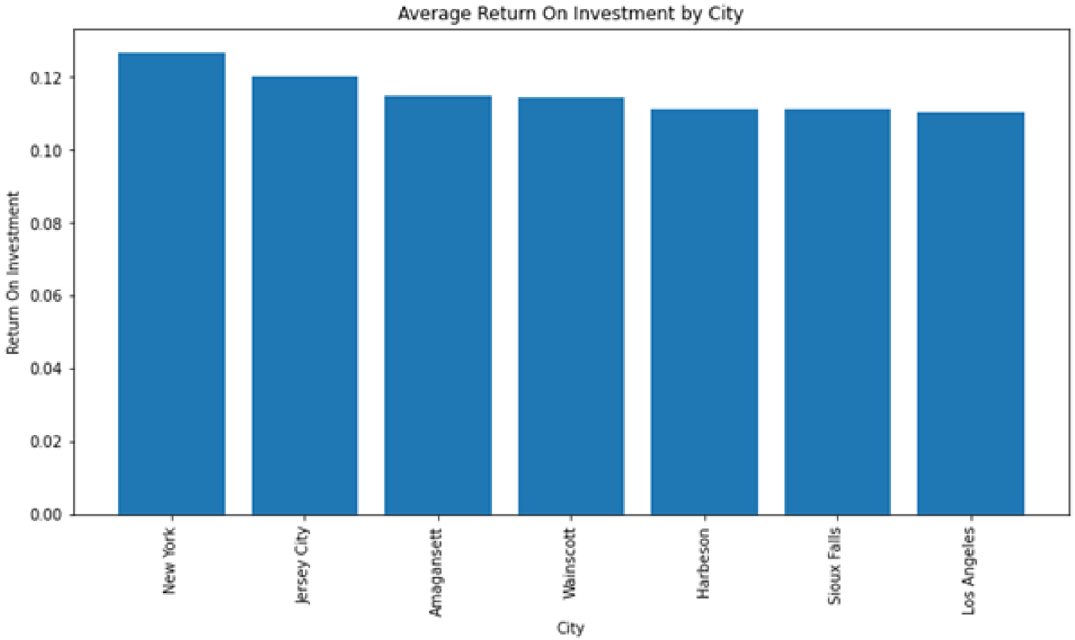
From the visualization above, California, New York, Texas, Pennsylvania, and Florida emerged as the states with the highest count of houses. At the same time, South Dakota, Vermont, Washington DC, Wyoming, and Arkansas registered the lowest number of houses, respectively.

2.3 Bivariate Analysis

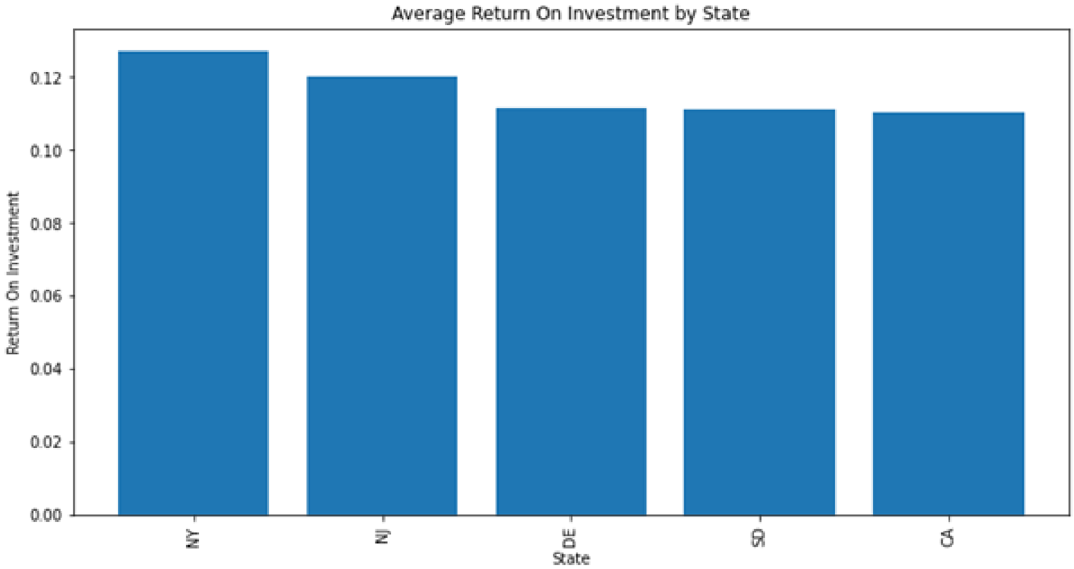
From the bar graph we can conclude the Region with Zip code 11211 located in New York State had the highest Return on Investment.



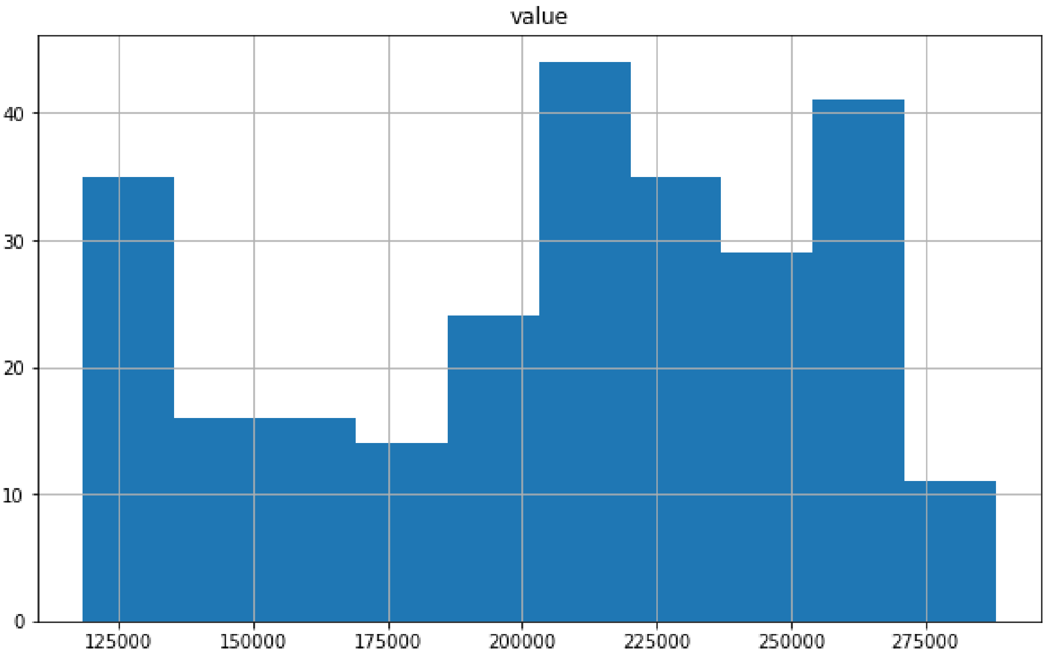
The graph below shows the relationship between the Return on Investment and the Different cities provided. We can therefore conclude that New York City has the highest Return on Investment followed by Jersey City, Wainscott, Amagansett, Hartsel, Los Angeles, and Washington.



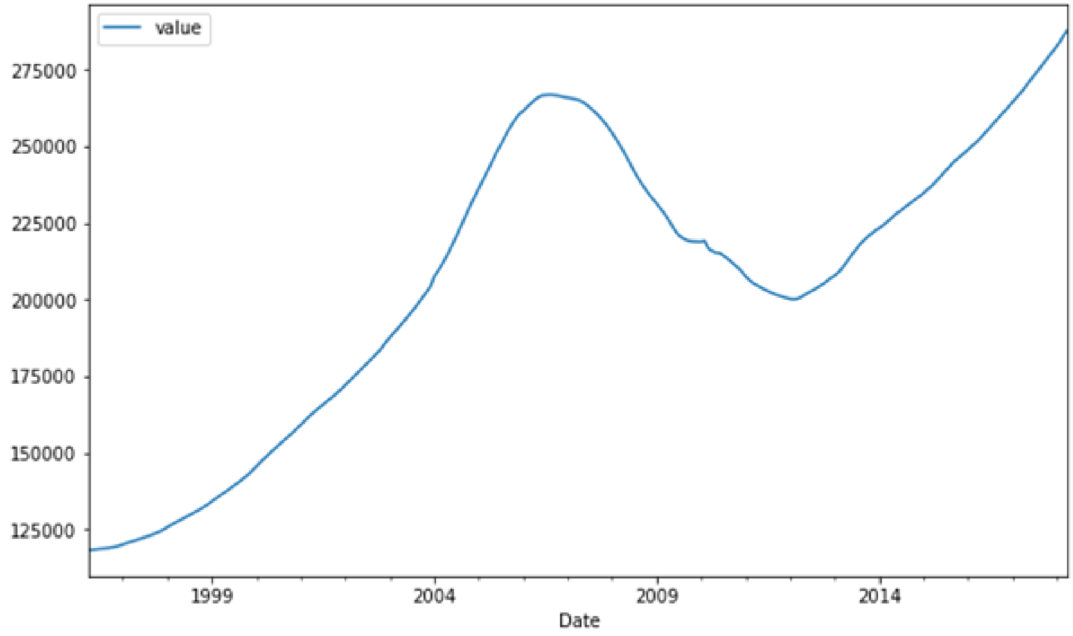
Furthermore, the states with the highest average returns on investment are New York, New Jersey, Delaware, South Dakota, and California.



The data did not follow a normal distribution as indicated by the histogram below.

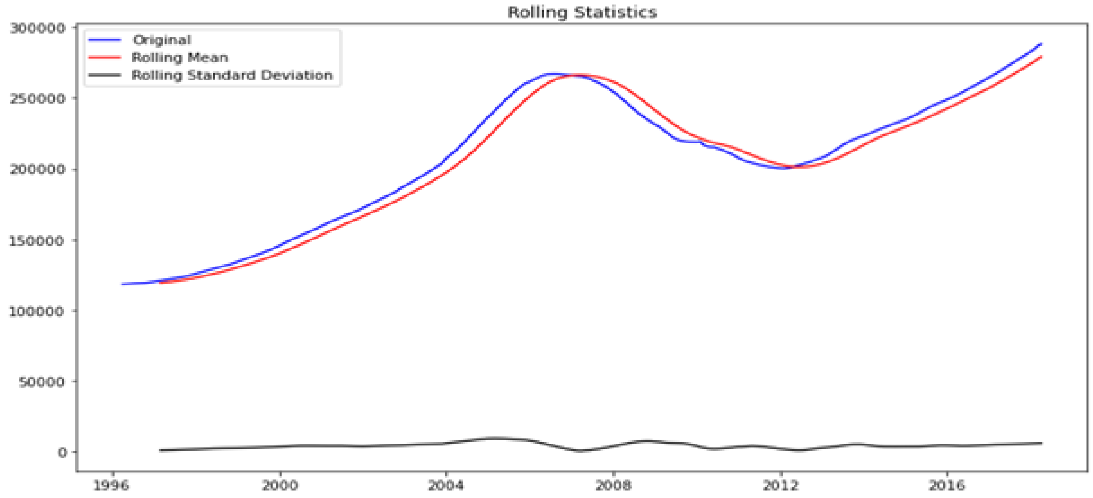


After monthly, quarterly, semi-annual, and annual resampling, the data exhibited an upward trajectory while lacking a distinct seasonality. Below is a plot after monthly resampling.



Beyond Notably, there was a conspicuous surge in house prices around 2008. This notable increase can be attributed to the global recession, which profoundly impacted the housing market in the United States during that period. The observed spike in house prices during 2008 aligns with the broader economic downturn, providing context for the temporary deviation from the overall trend in the real estate market.

We also conducted rolling statistics which is performed to check for any stationarity in our data. This will help us identify trends or patterns in the data, and the stationarity check involves examining whether these trends are constant over time. The line graph below shows that our data is not stationary from the rolling mean. We also performed a Dickey-Fuller Test, confirming our data was not stationary.



3. Modeling

The data preprocessing steps we took were that we subtracted the data with its weighted exponential rolling mean and differenced the data twice in order to achieve stationarity.

We then created 7 models, 4 ARIMA and 3 Prophet models, and determined that the best-performing model was the ARIMA(7, 1, 8) model. The table below shows the performance of each model.

Model	ARIMA(1,0,1)	ARIMA (1,0,3)	ARIMA (3,0,1)	ARIMA (7,1,8)	Prophet-1	Prophet-2	Final-Prophet
RMSE	121.94	122.23	122.24	118.11	142.35	122.51	121.87
MAE	85.97	86.27	86.18	86.26	95.67	87.13	86.22

4. Conclusion

From the above project, we met our objectives as follows:

1. The best zipcodes to invest based on ROI were: i. 11211 – Brooklyn, New York ii. 11222 – Brooklyn, New York iii. 11216 – Brooklyn, New York iv. 7302 – Jersey City, New Jersey v. 11215 – Brooklyn, New York. The best cities to invest in were: i. New York ii. New Jersey iii. Winscott iv. Amagansett v. Hartsel. The best states to invest were: i. New York ii. New Jersey iii. Colorado iv. California v. Washington DC.

2. We also noted that the real estate prices had an upward trend, meaning that the value increased over time, although there was no clear way to determine the best time to enter the market as the data was non-stationary.

3. Finally, the ARIMA(7,1,8) model was the best predictive model to forecast future real estate values.

Contributors:

Name	GitHub
Kelvin Rotich	https://github.com/Kelvin-Rotich824
Crystal Wanjiru	https://github.com/CrystalW123
Miriam Nguru	https://github.com/miriamnguru
Celiajoy Omiah	https://github.com/celiahjoyomiah
Paul Mbugua	https://github.com/Paulwaweru
Stephen Butiya	https://github.com/obystephen