ZILLOW

PROJECT

FORECASTING REAL ESTATE PRICES

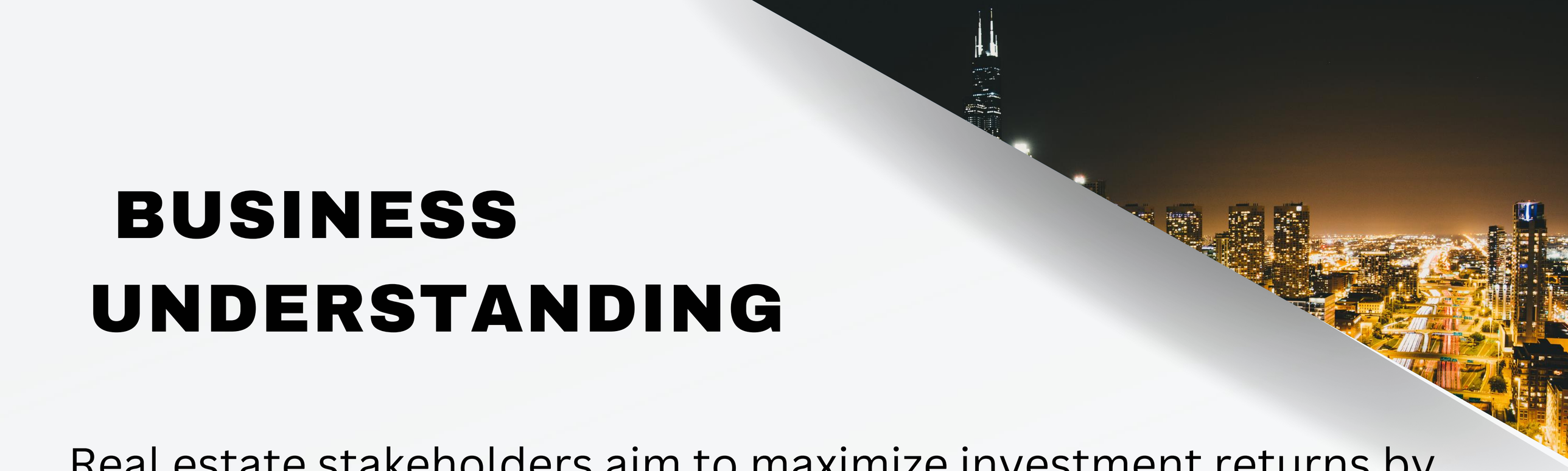
https://github.com/KirigoY/Group1_Phase4_Project/blob/master/starter_notebook.ipynb

CONTENT

INTRODUCTION

- The real estate market is vital to the global economy, impacting finance, construction, and urban development.

- Accurate price forecasting is crucial for investors, policymakers, homeowners, and developers.



BUSINESS UNDERSTANDING

Real estate stakeholders aim to maximize investment returns by understanding key factors affecting property values, leading to smarter investment decisions.

PROBLEM STATEMENT

What are the top 5 zip codes?

PROFIT

The market volatility and unforeseen factors can make this difficult to archieve

RISK

Estate price market downturns, economic fluctuations, and regional disparitie are prone to colapse.

TIME HORIZON

Being unable to know varying market cycles and regional differences may not be clear.

GOALS AND OBJECTIVES

01 Develop a model to forecast real estate prices and identify the top 5 ZIP codes for investment

02 Assess investment risks and recommend optimal time horizons for real estate investments

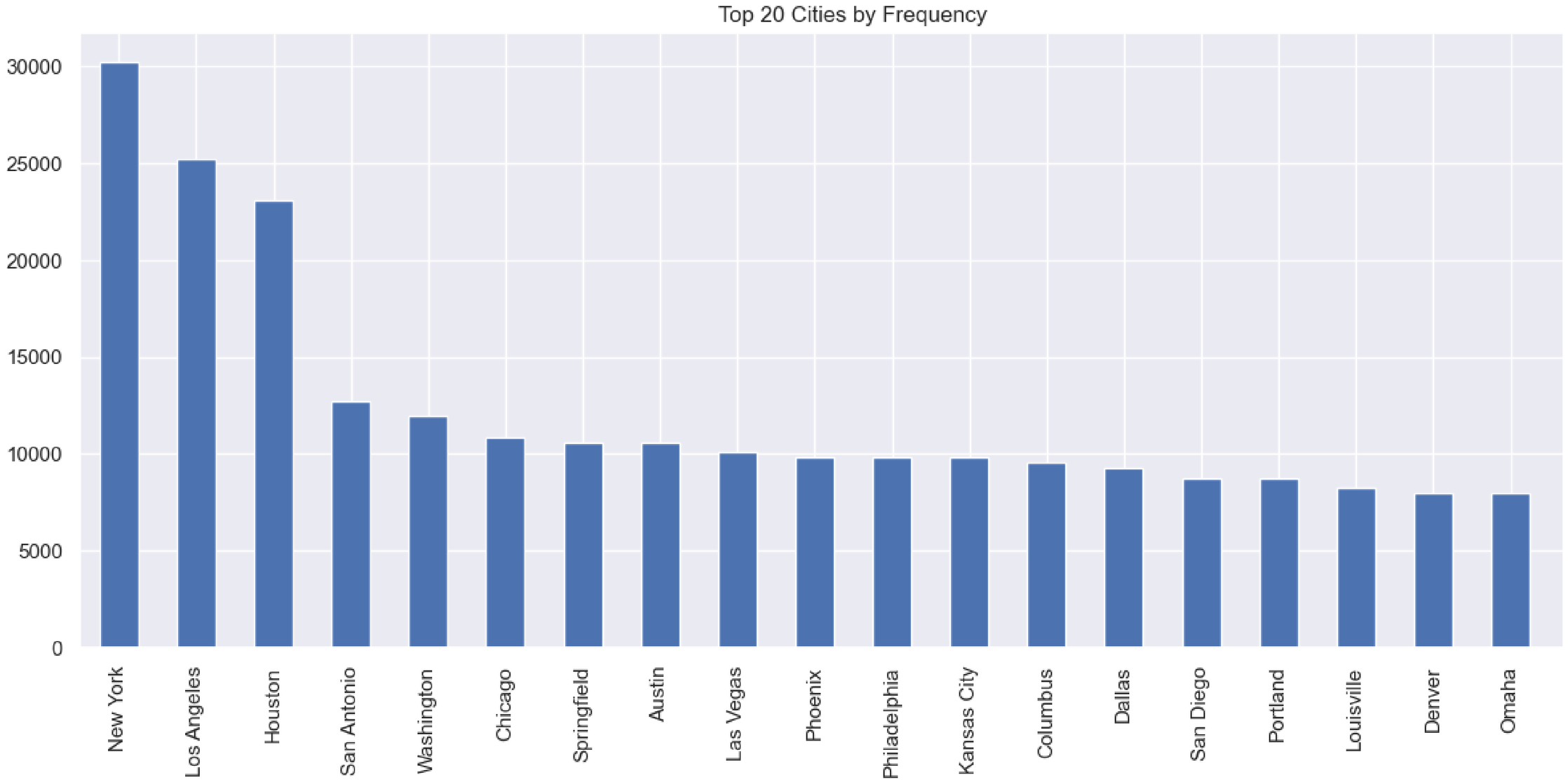
03 Evaluate profit margins to prioritize high-yield investment opportunities.

DATA UNDERSTANDING

- The dataset used for this project was downloaded from Zillow housing data. The dataset spans from April 1996 to April 2018 and provides detailed monthly home value data for a wide range of regions across the United States.
- The dataset has 14,723 rows and 272 columns. The information contained within the columns is as follows, as described by the data dictionary:
  - RegionID
  - RegionName
  - City
  - State
  - Metro
  - CountyName
  - SizeRank
  - Monthly Home Values

PRICE DISTRIBUTION

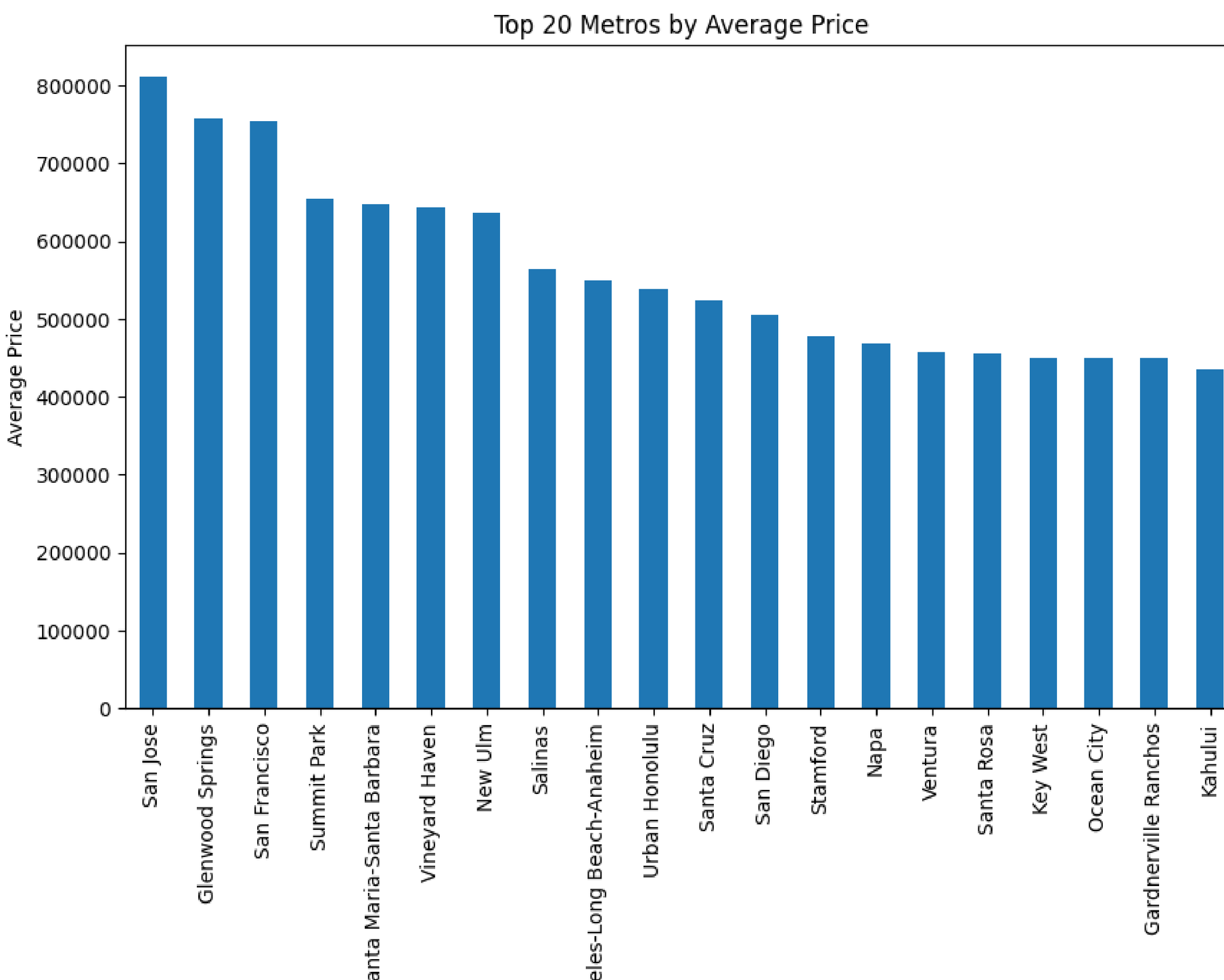
- Robust Models: All zip codes have strong predictive models.
- Best Fit: Zip codes 3 and 4 fit the data best.
- Normality Issues: Some zip codes may have inconsistencies in data.

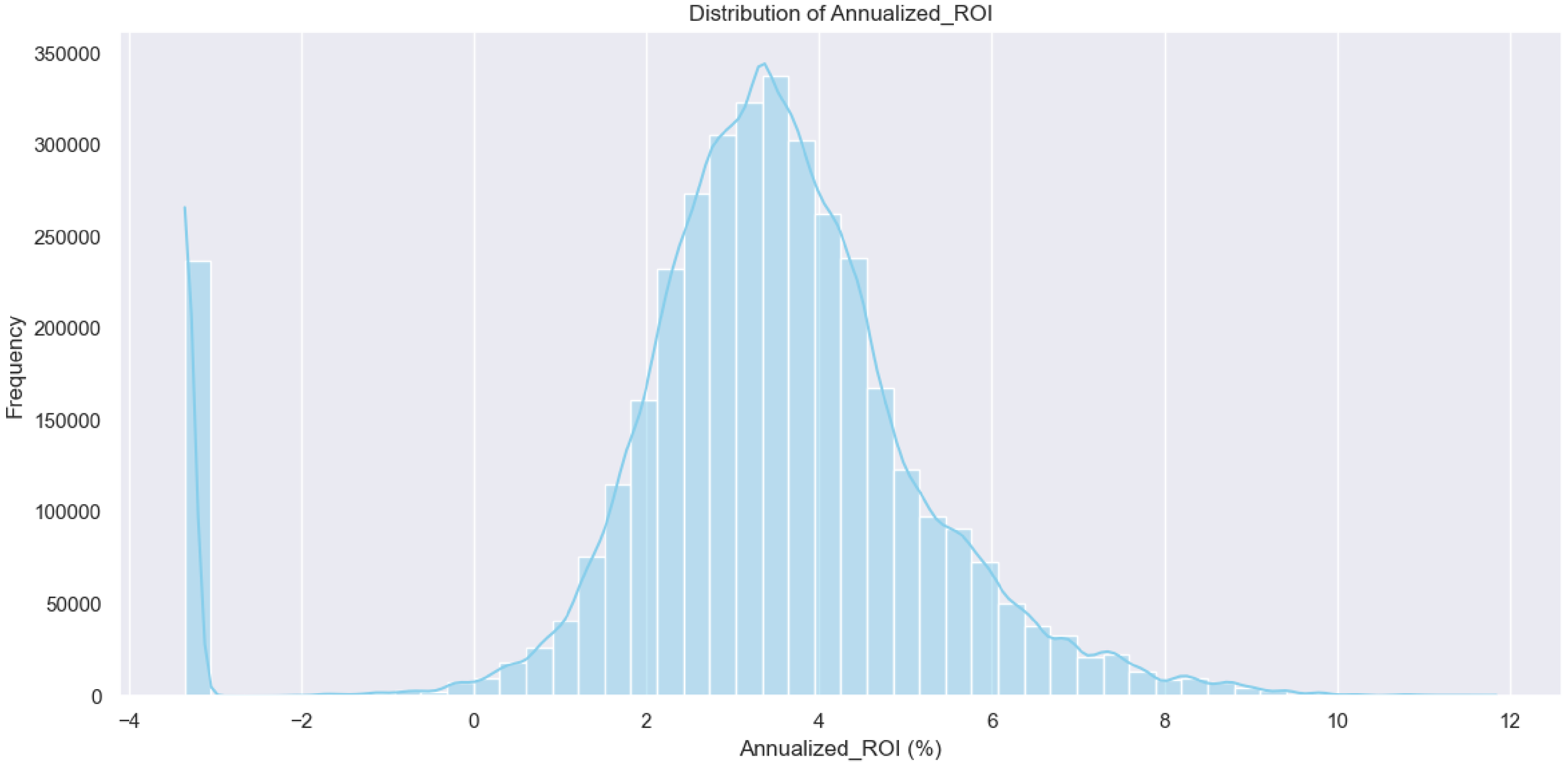


This plot shows the top 20 most frequent cities in the dataset.

Metr
o

These are hubs for technology, finance, and other high-paying industries, which can drive up housing prices





Most investments seem to yield an annualized ROI of around 4%, indicating a potentially good profit margin for the majority of real estate investments.

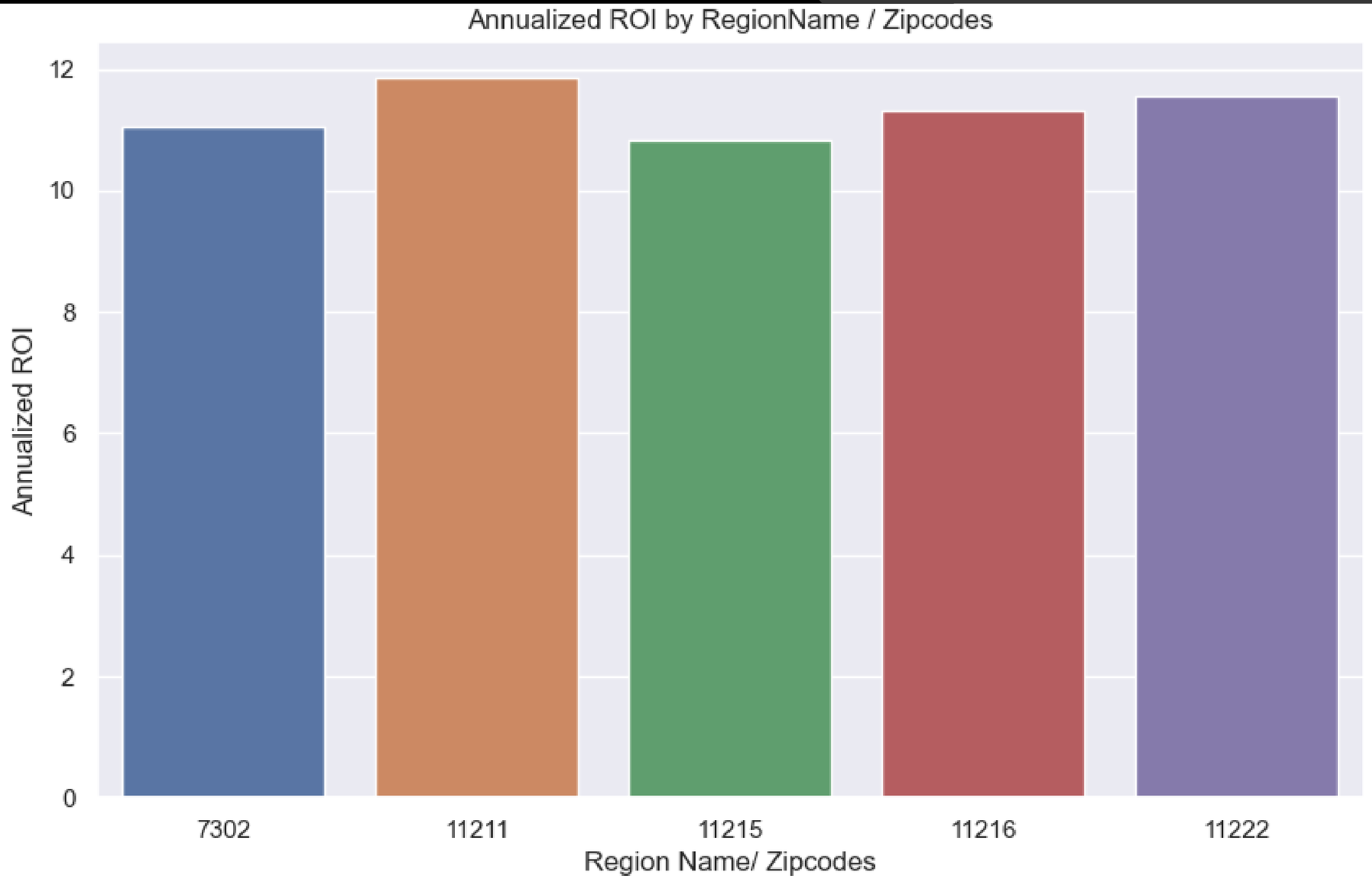
ANNUALIZED ROI

Bivariate Analysis

Annualized ROI: ~ This zip codes have the top 5 highest return on investment, making it the most attractive for real estate investment in terms of profitability.

KEY

11211 (New York, Kings)
11222 (New York, Kings)
11216 (New York, Kings)
07302 (Jersey City, Hudson)
11215 (New York, Kings)



AVERAGE PRICE BY COUNTY

The chart shows the top 20 counties by average property price in dollars:

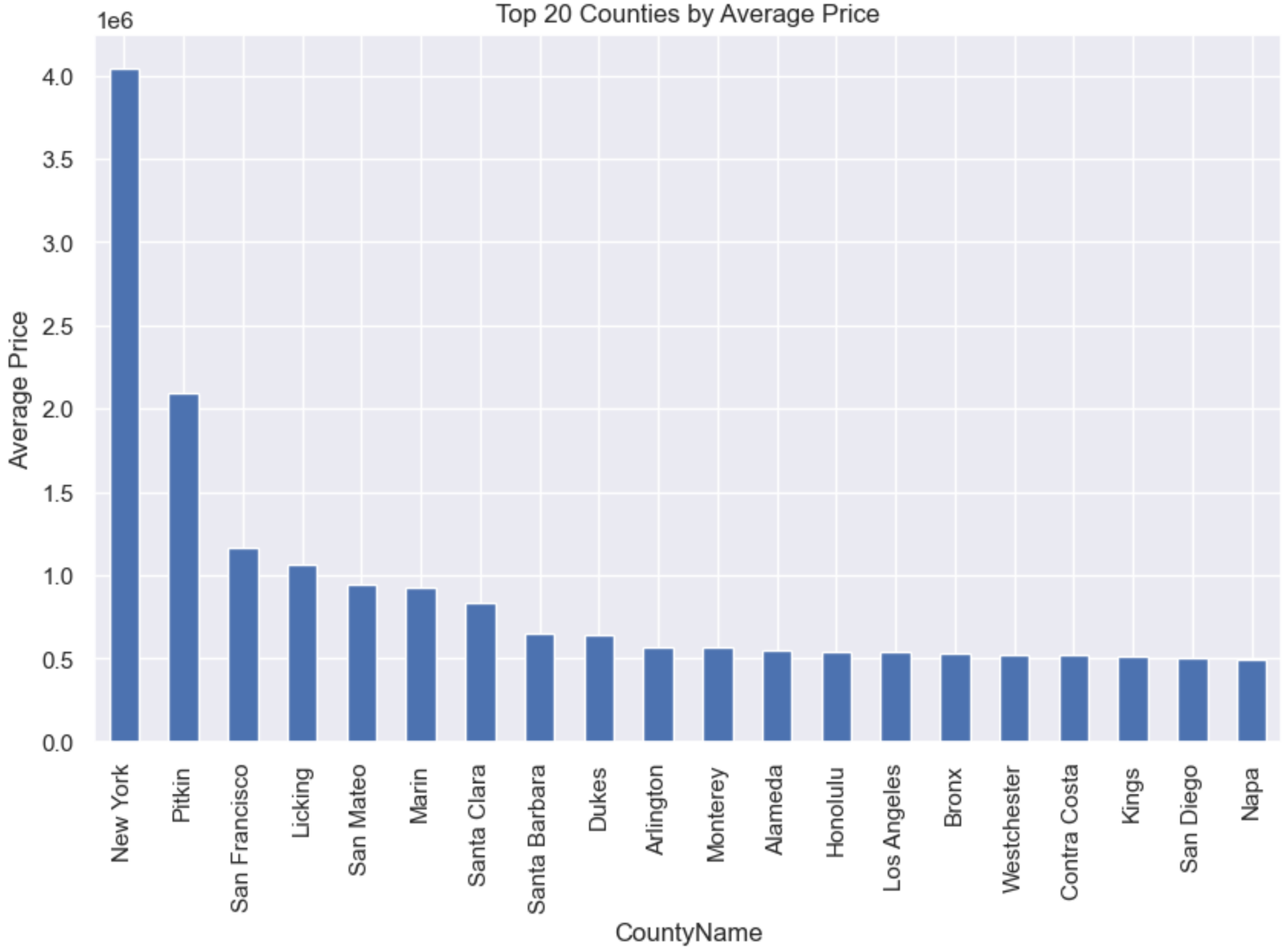
1. *New York*: Nearly $4,000,000.

2. *Pitkin*: Just over $2,000,000.

3. *San Francisco* and others (Licking, San Mateo, etc.): $1,000,000 to $1,500,000.

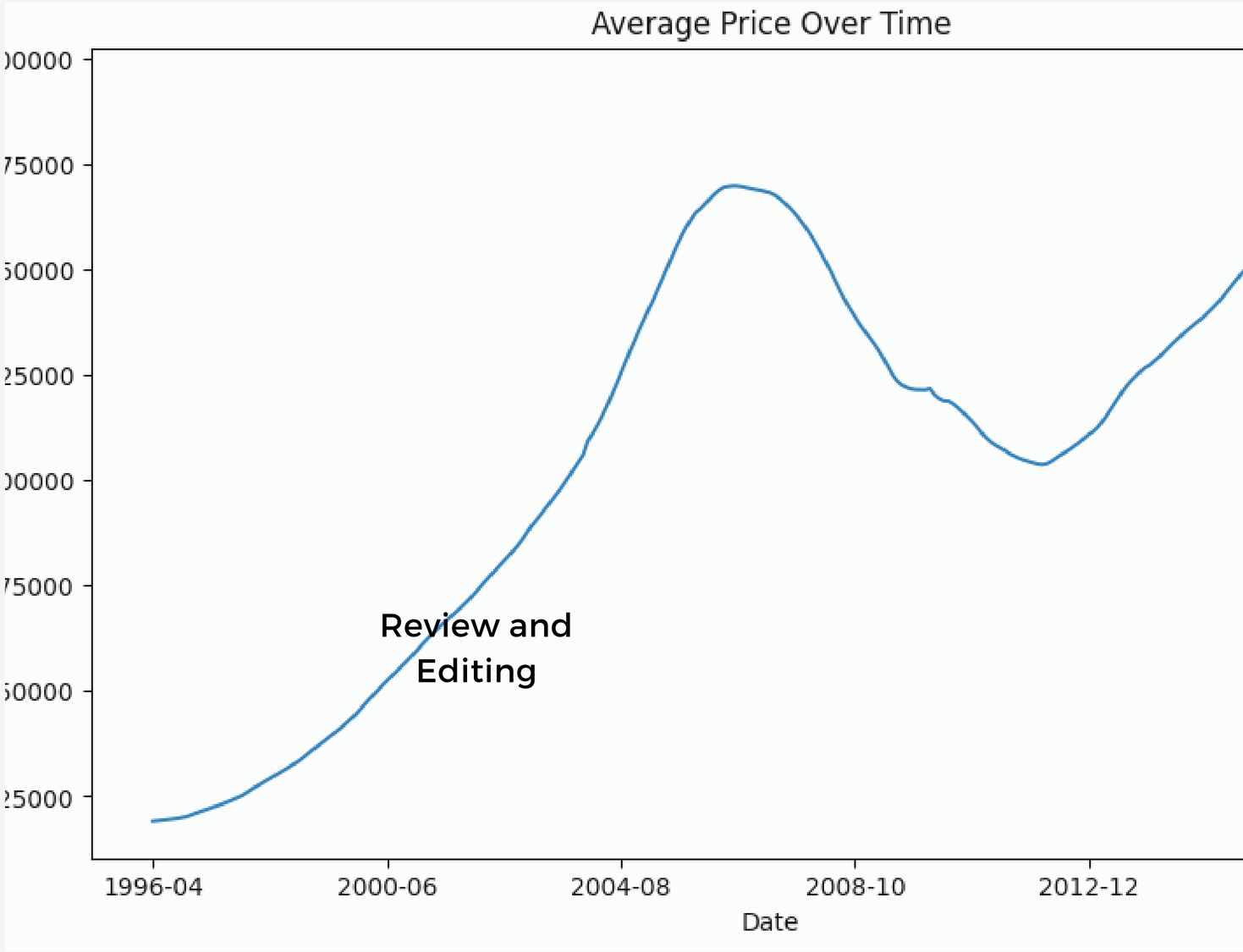
4. *Remaining counties*: Below $1,000,000.

New York County is the most expensive by a large margin.



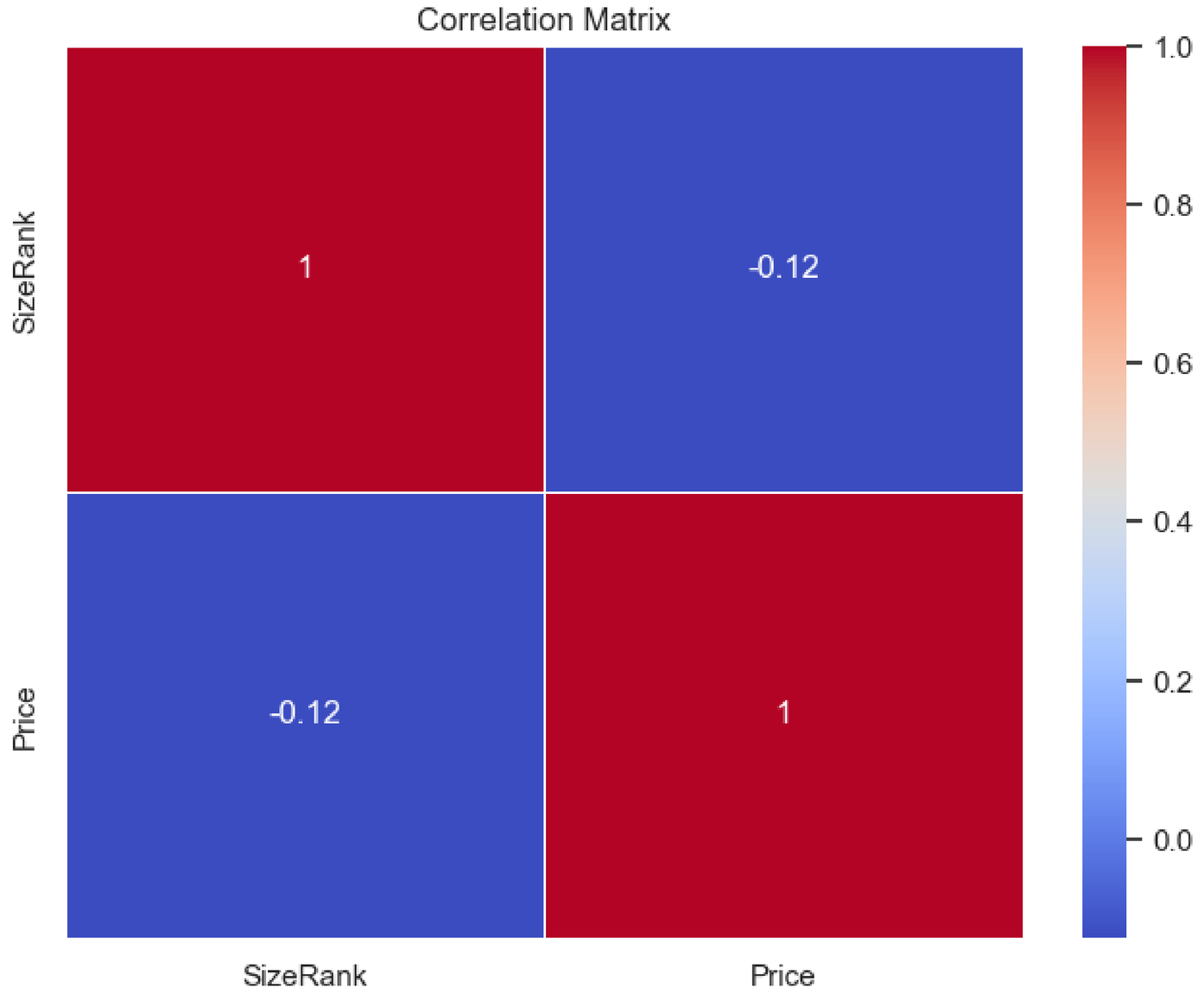
AVERAGE PRICE OVER TIME

- **Rising Prices (1996-2006)**: Prices increased from $100K to $275K.**
- **Peak and Drop (2006-2012)**: Prices peaked in 2008 at $275K, then dropped below $225K by 2012.**
- **Recovery (2012-2017)**: Prices rose again, reaching $300K by early 2017.**
- **This shows a cycle of growth, peak, decline, and recovery.**



CORRELATION BETWEEN PRICE AND SIZE RANK

- The correlation between SizeRank and Price is weakly negative (-0.12).
- This suggests that as the size rank of a region increases, the housing prices tend to decrease slightly, but the relationship is not strong.



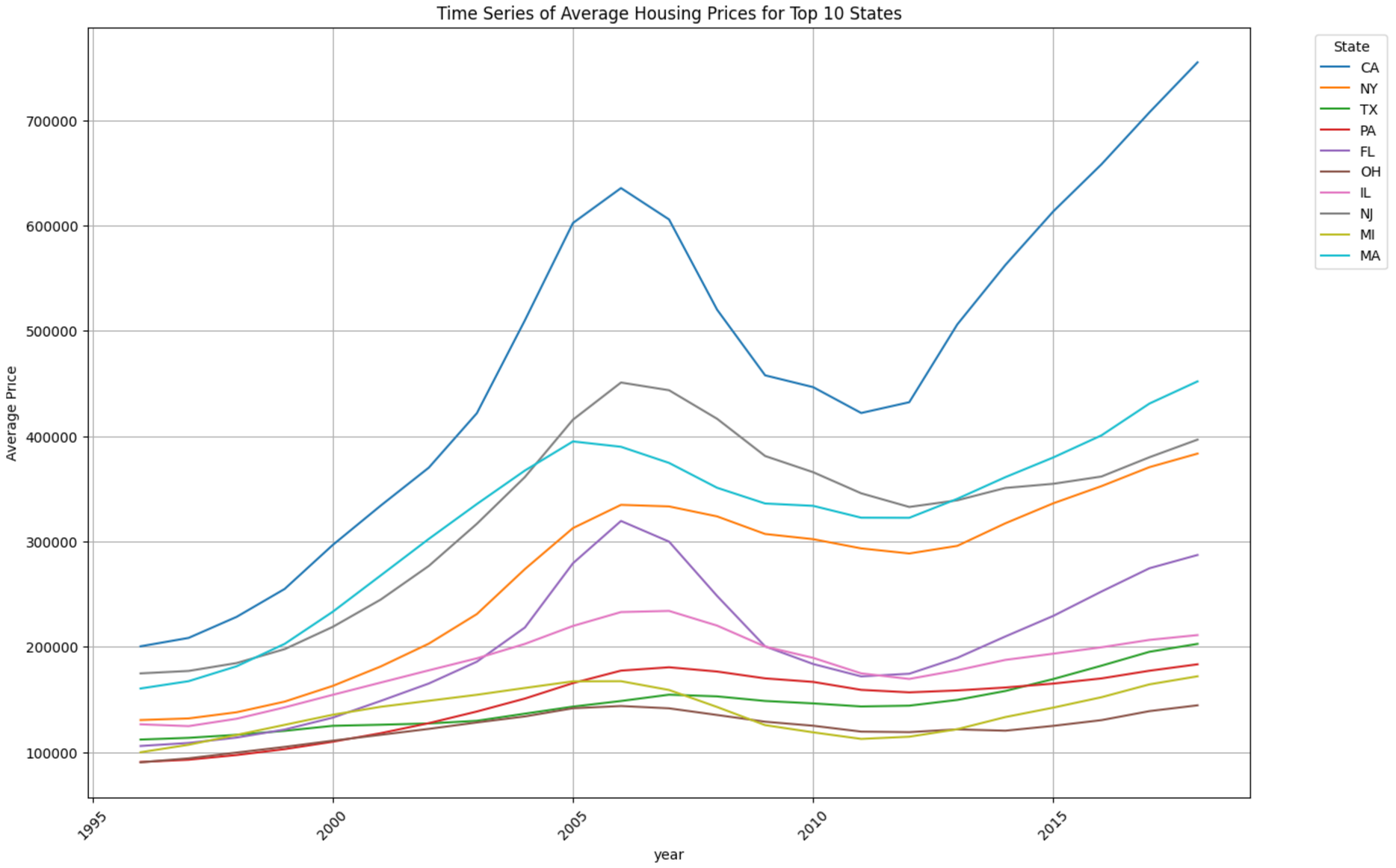
MARKET ANALYSIS

Multivariate Analysis

- **Market Sensitivity**: Fluctuations in average housing prices suggest the market is influenced by economic and seasonal factors..

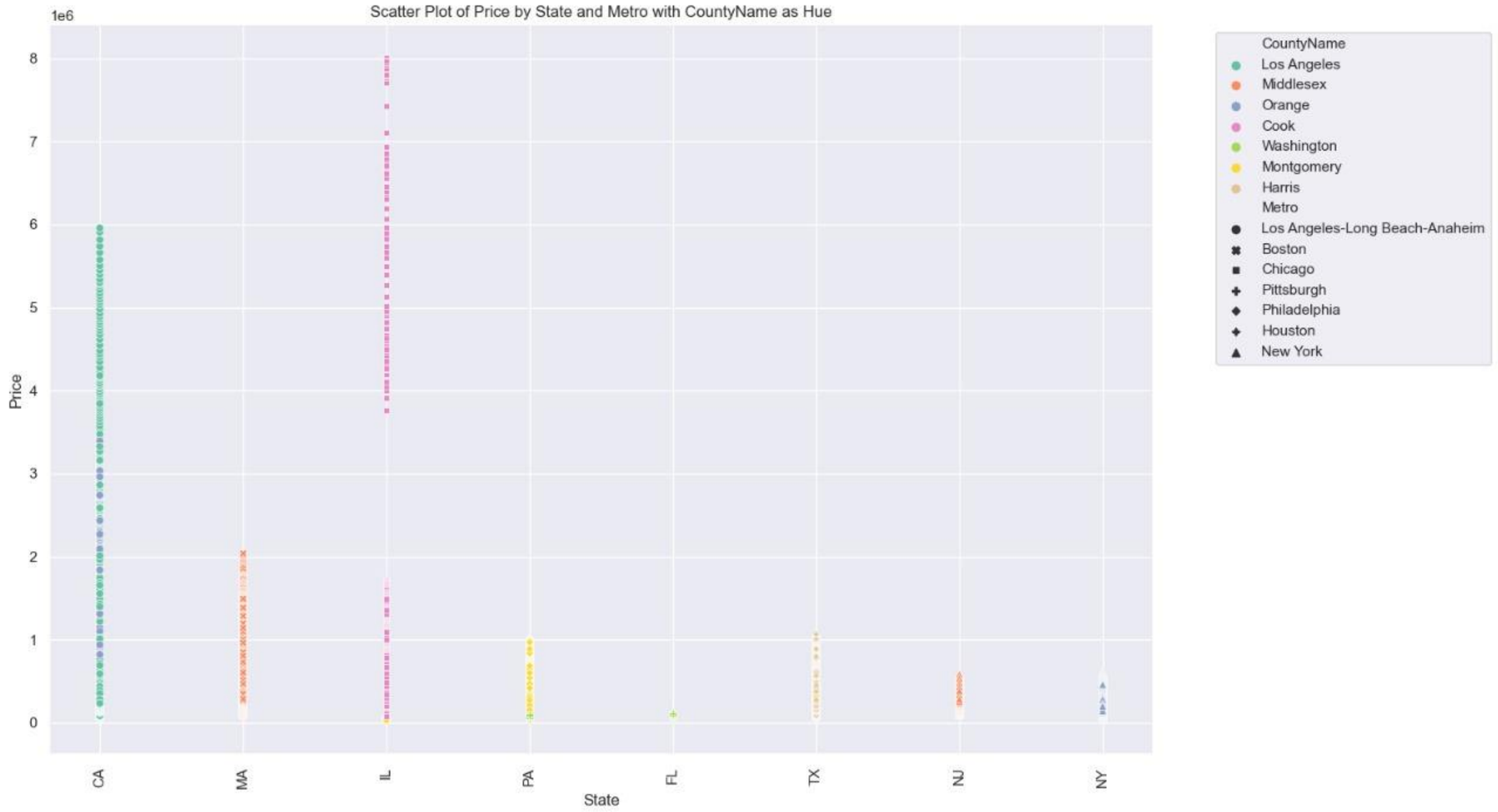
- **Competitive Market:** Housing prices in these states are significantly higher than the national average.

- **Strong Demand**: A steady increase in average housing prices across all states indicates strong demand



PRICE BY STATE AND METRO WITH COUNTYNAME AS HUE

Multivariate Analysis



Overall Trend:

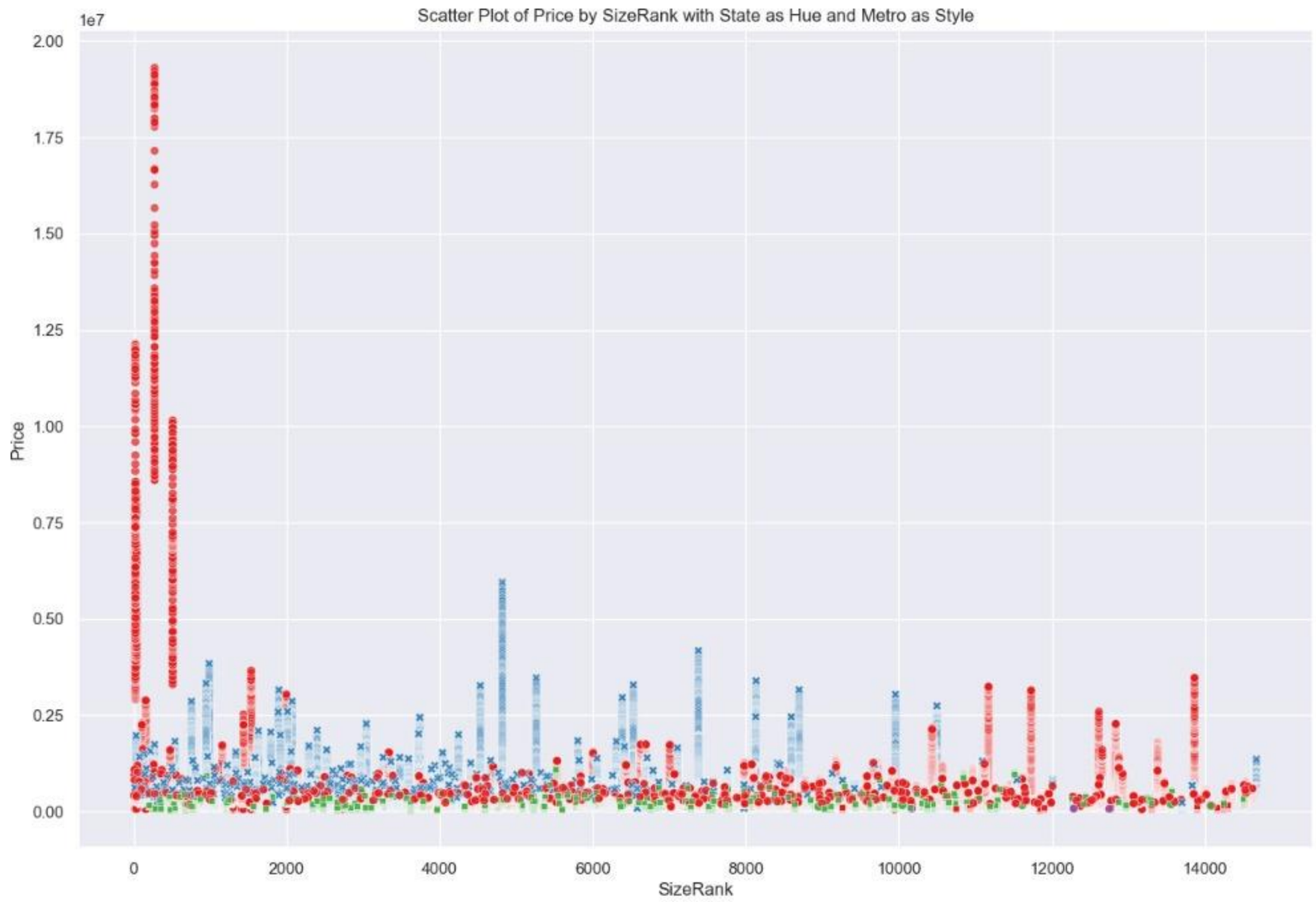
- Significant variation in prices across states and counties.
- California and New York display the highest price points.

County Highlights:

- *Highest Price Points:* Los Angeles (CA) and New York.
- *Other Counties:* Middlesex (MA), Orange (CA), Cook (IL), Washington (PA), Montgomery (PA), and Harris (TX) show lower price ranges.

PRICE BY SIZERANK WITH STATE AS HUE AND METRO AS STYLE

Multivariate Analysis



Metro Highlights:

New York: High prices.

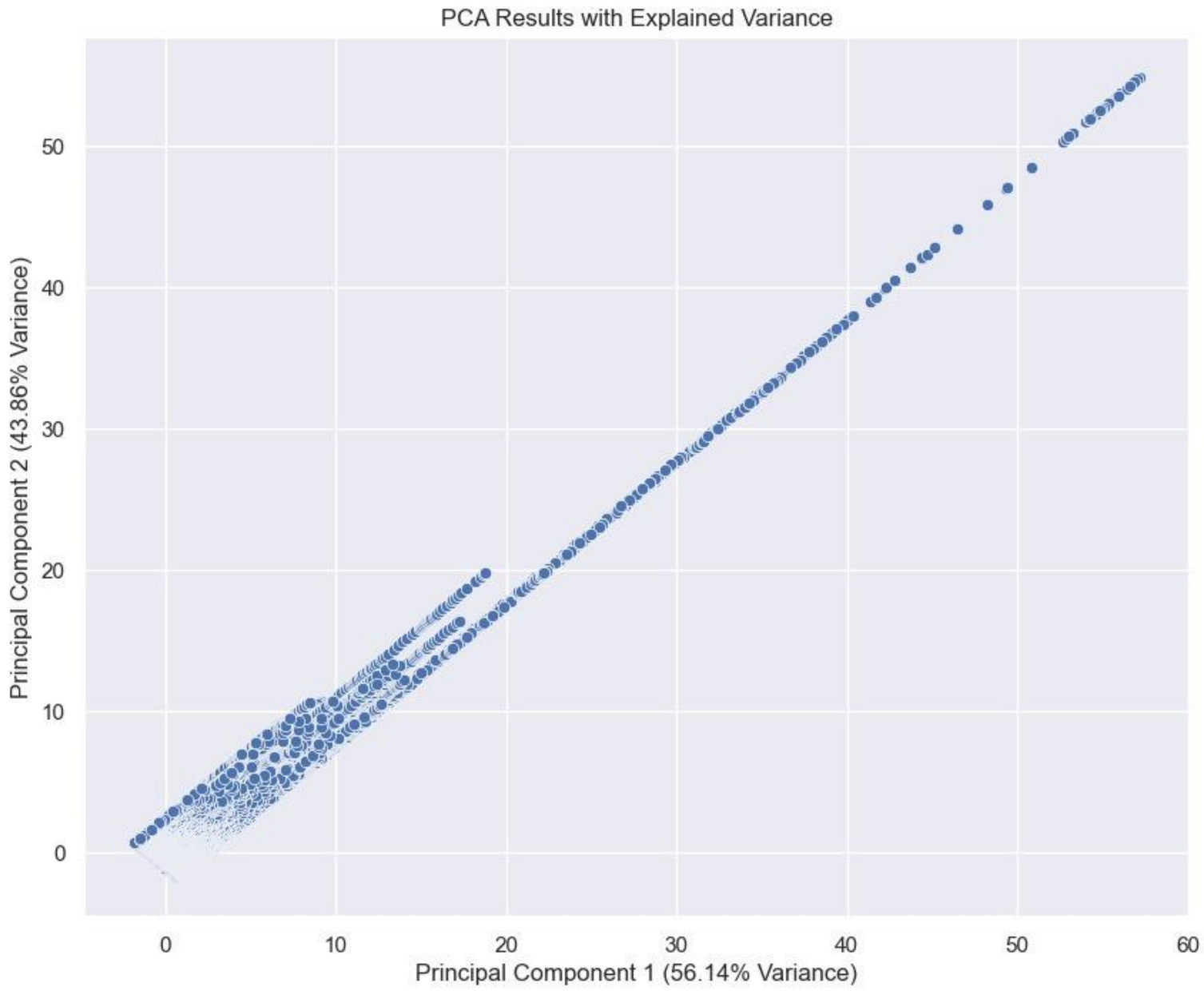
Los Angeles: Wide variance, several high prices.

Philadelphia: Lower prices.

Summary:

- Big price differences across states and metros.
- New York leads in high prices, followed by California.
- Pennsylvania and Texas have lower prices.

PCA COMPONENTS AND VARIANCE



KEY POINTS

1. Axes:
X-axis: Principal Component 1 (56.14% variance)
Y-axis: Principal Component 2 (43.86% variance)

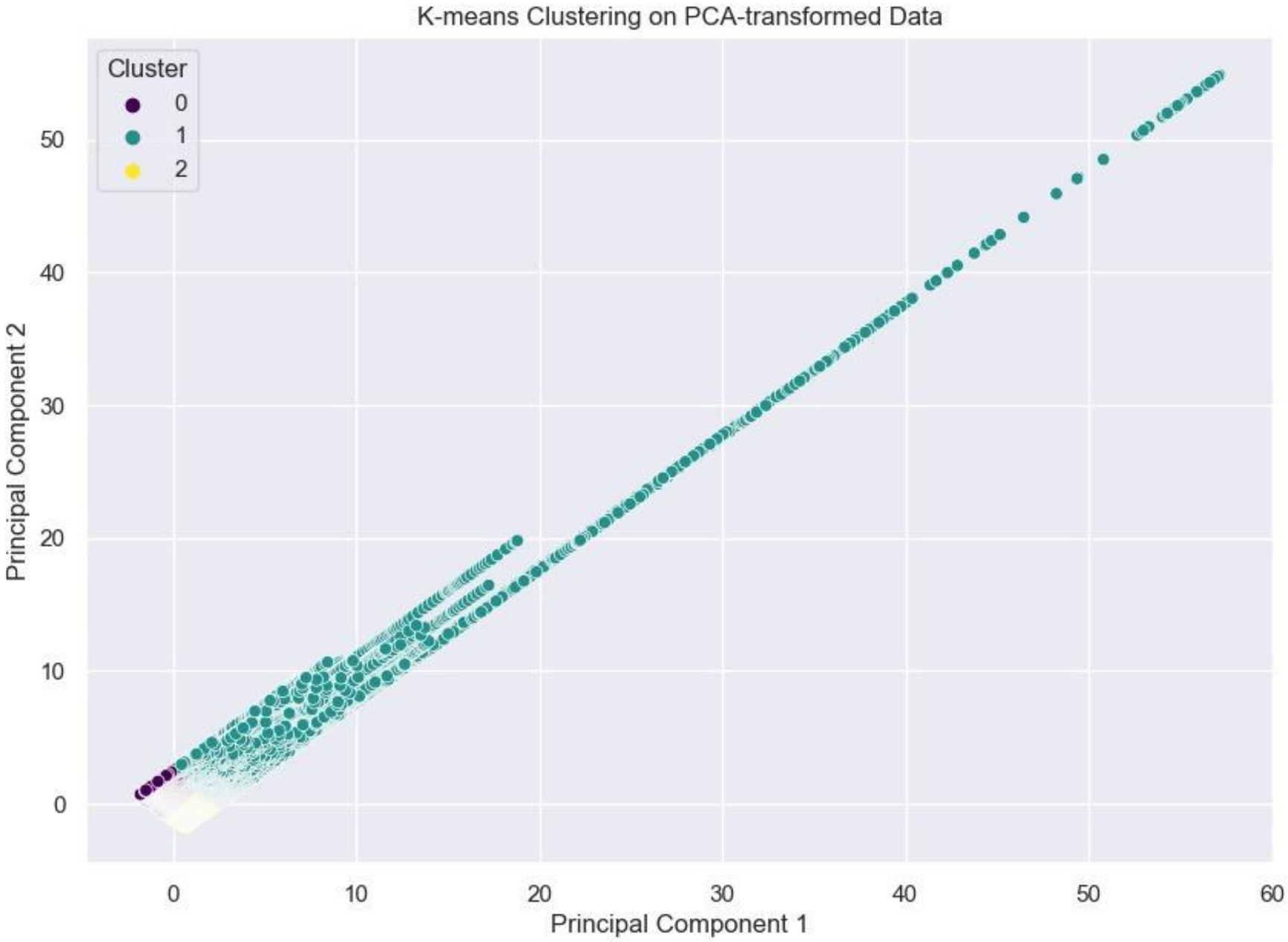
2. Trend:
Data points form a clear diagonal line.
Indicates a strong linear relationship between components.

3. Variance Explained:
Together, these components explain 100% of the variance.
PC1 (56.14%) + PC2 (43.86%).

Summary:
- First two components capture all variability.
- Strong linear relationship between them.

IDENTIFYING GROUPS WITH K-MEANS CLUSTERING

Multivariate Analysis



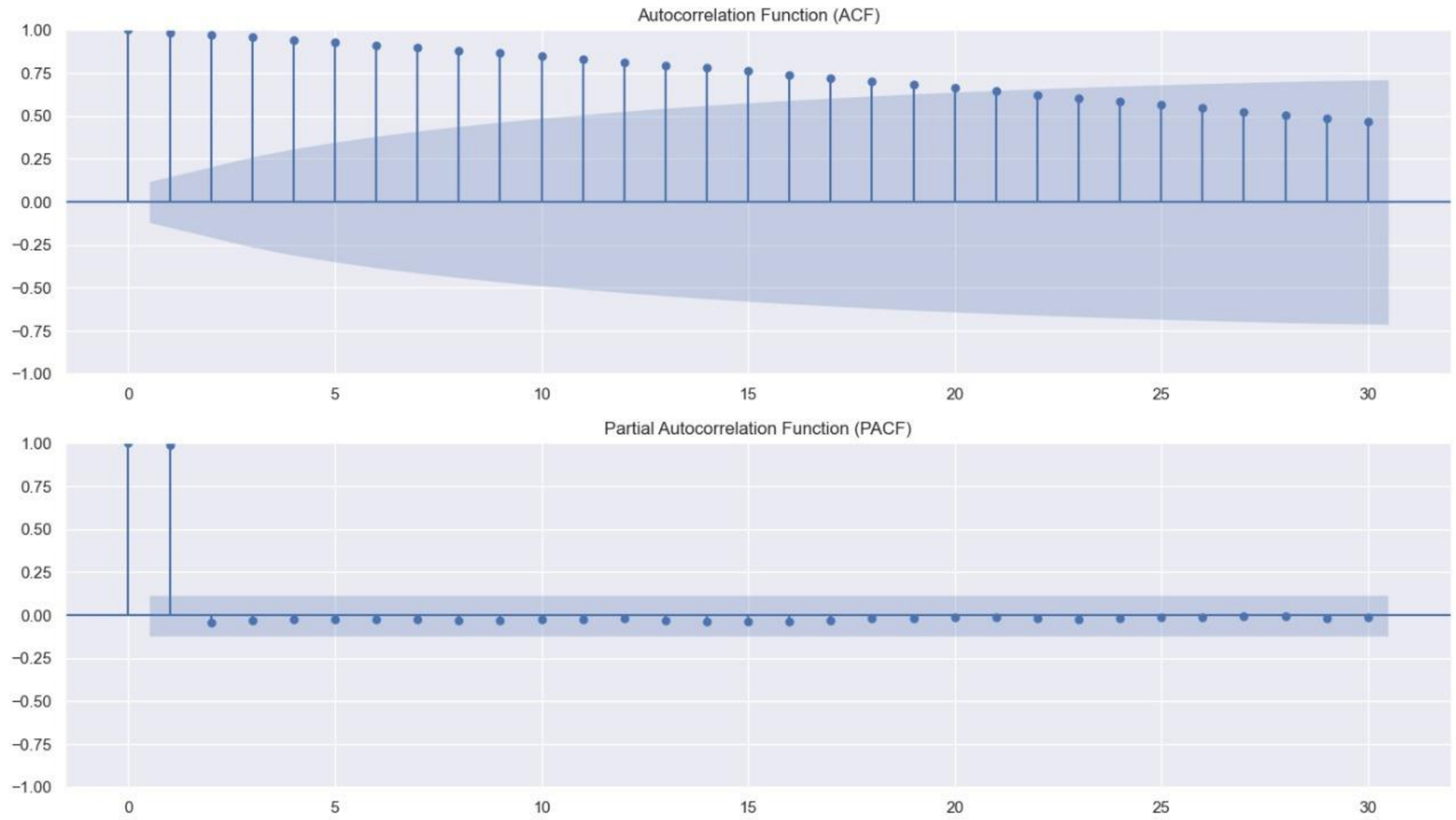
Distribution

Cluster 0 (Purple):
Data points with lower values.
Cluster 1 (Green):
Most prevalent, wide range of values.
Cluster 2 (Yellow):
Less frequent, localized in higher values.

Interpretation

Three distinct groups identified.
Cluster 1:
Most widespread across values.
Cluster 0 and 2:
Smaller, more localized.

ACF AND PACF PLOTS ANALYSIS



PACF Plot:
Significant correlation at lag 1 implies immediate past values are most predictive.
Suggests an AR(1) model (autoregressive model with one lag) might effectively capture the data's behavior.

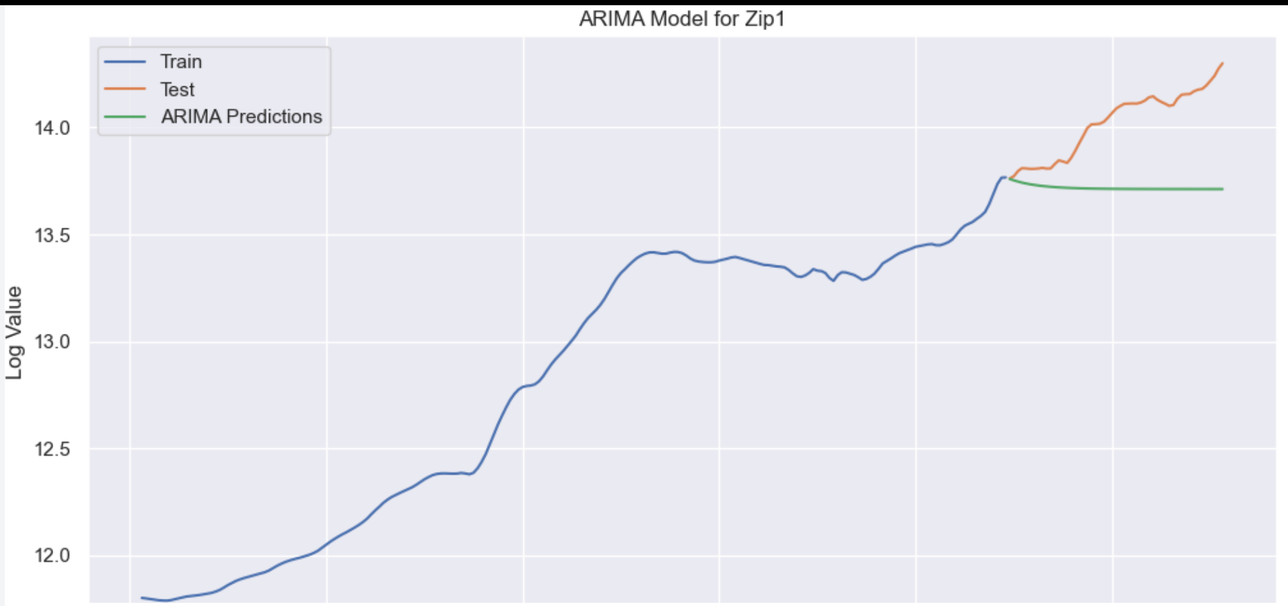
ACF Plot:
Indicates strong correlation with recent past values.
Suggests the presence of trends or seasonal patterns in the data.

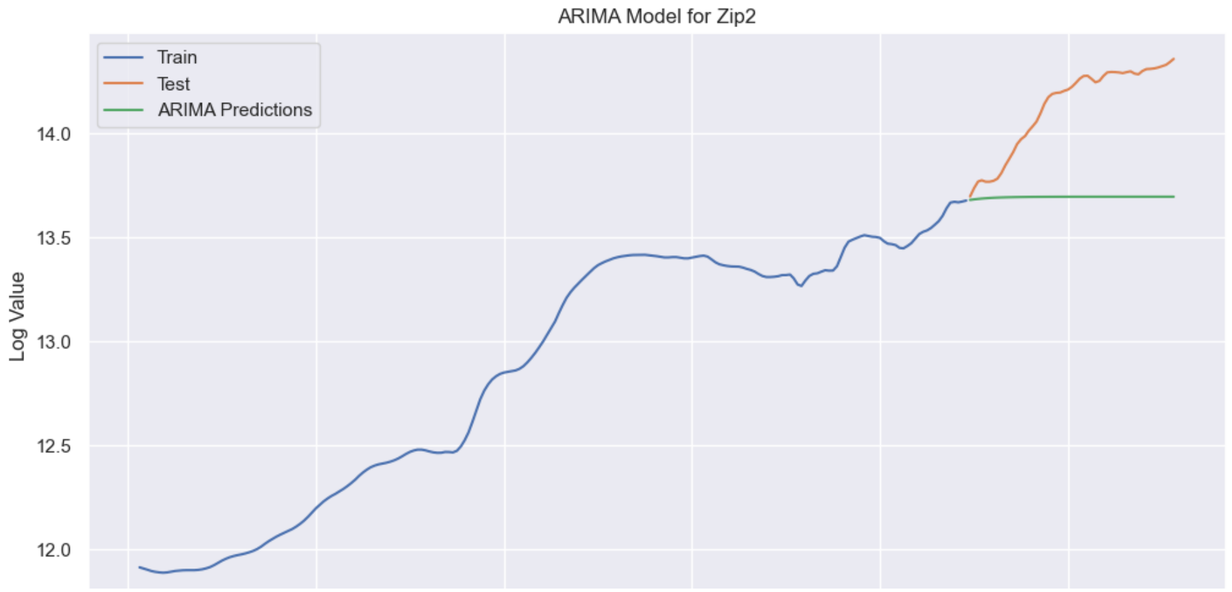
Overall:

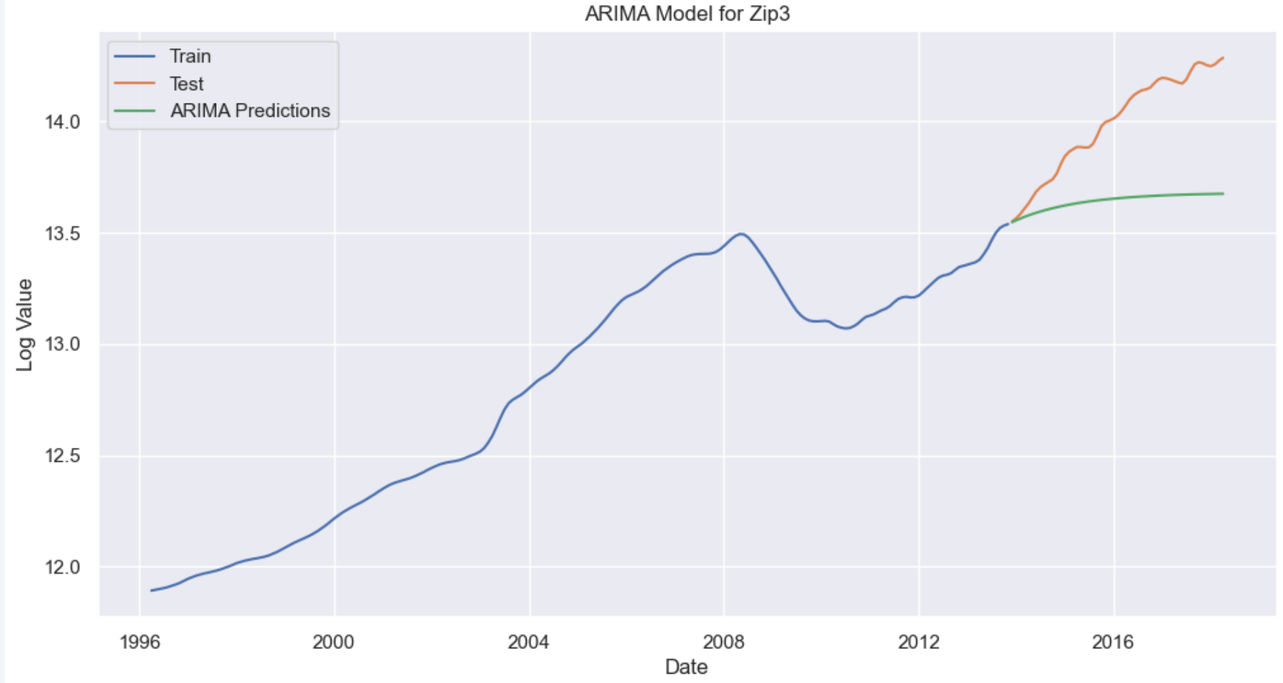
- Together, these plots highlight the potential need for differencing to stabilize the data.
- Useful for identifying suitable time series models to improve forecasting accuracy.

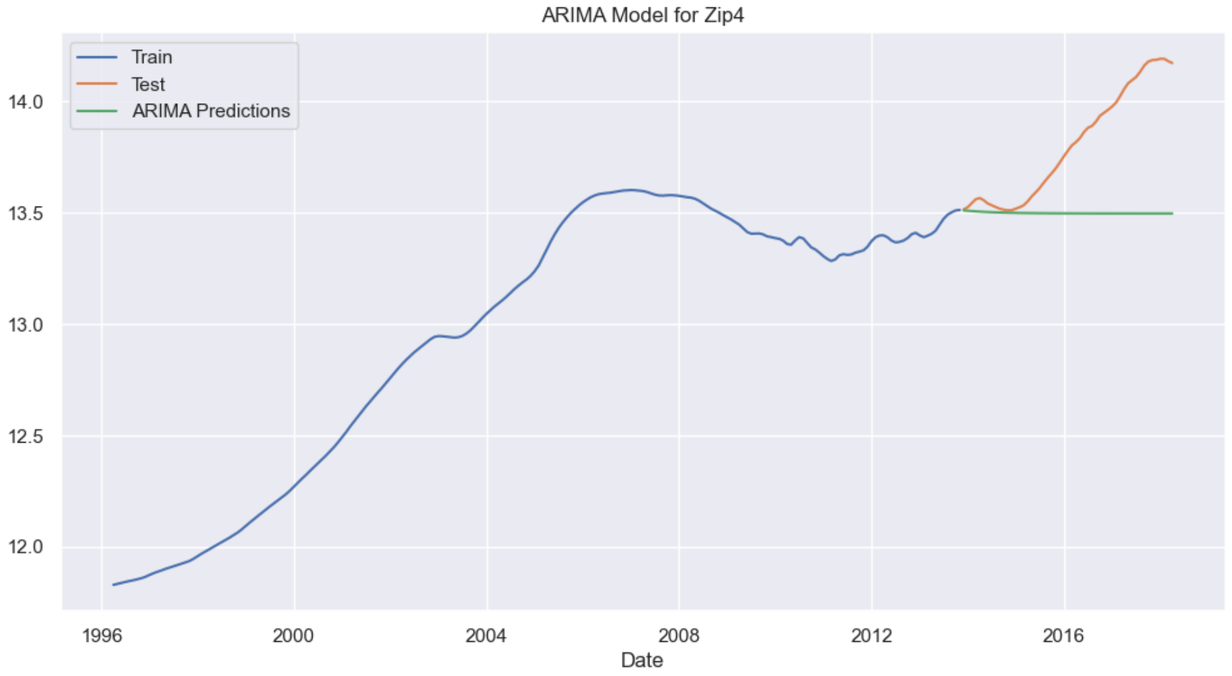
ARIMA

Modeling



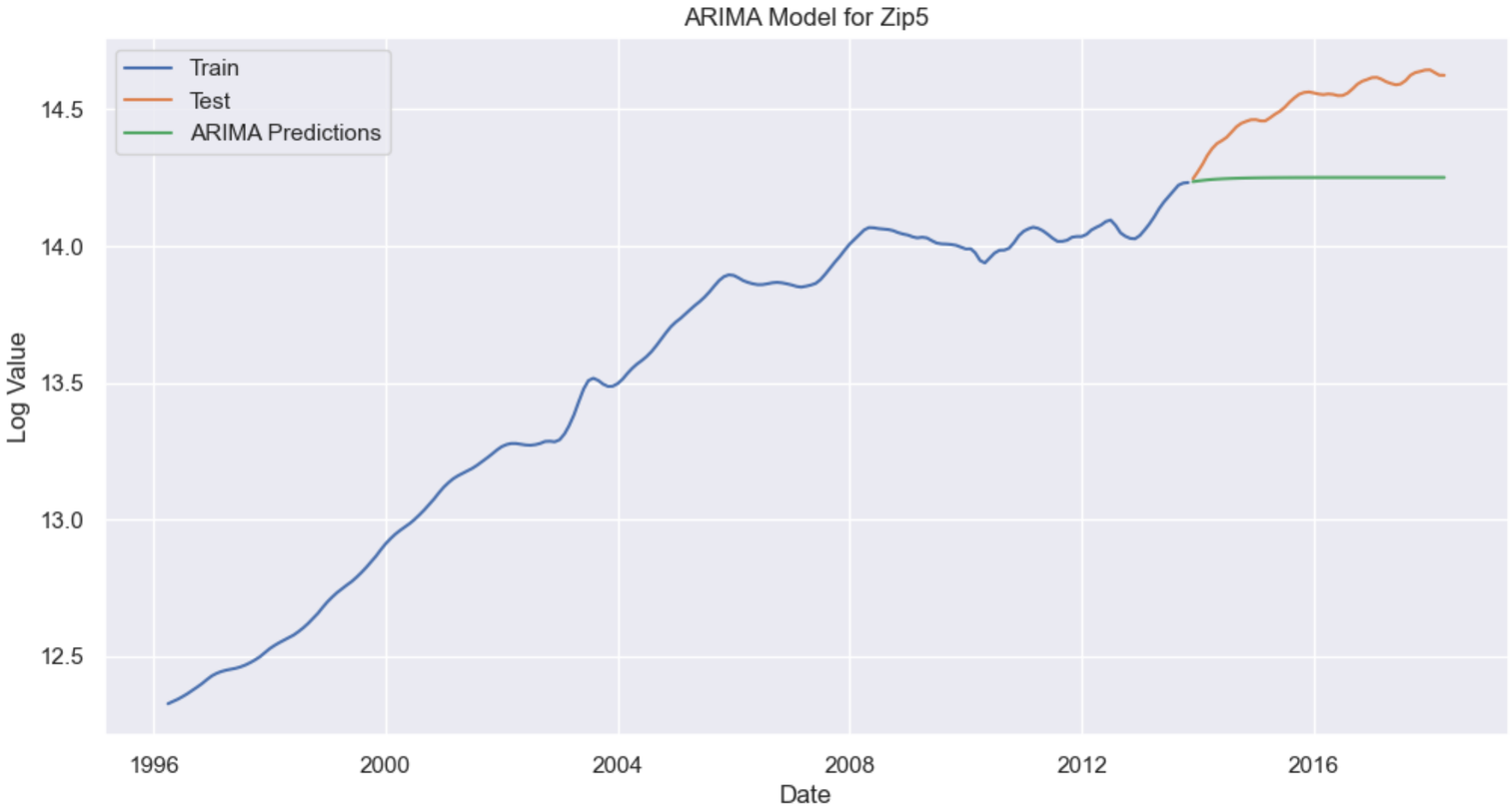






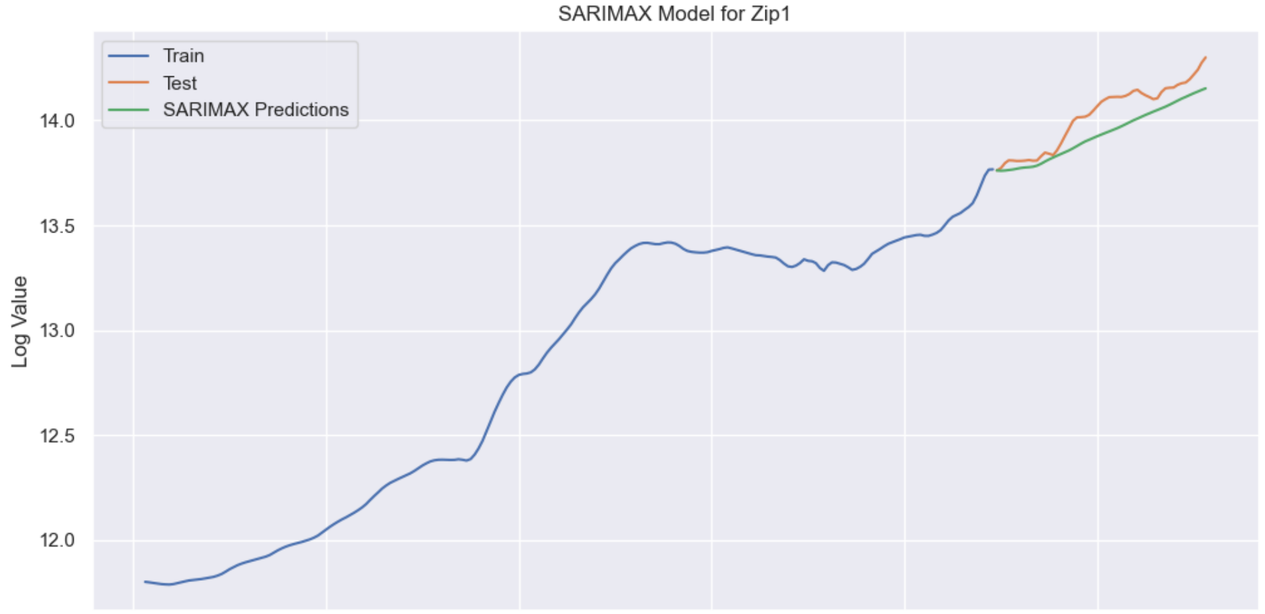
MODEL PERFORMANCE

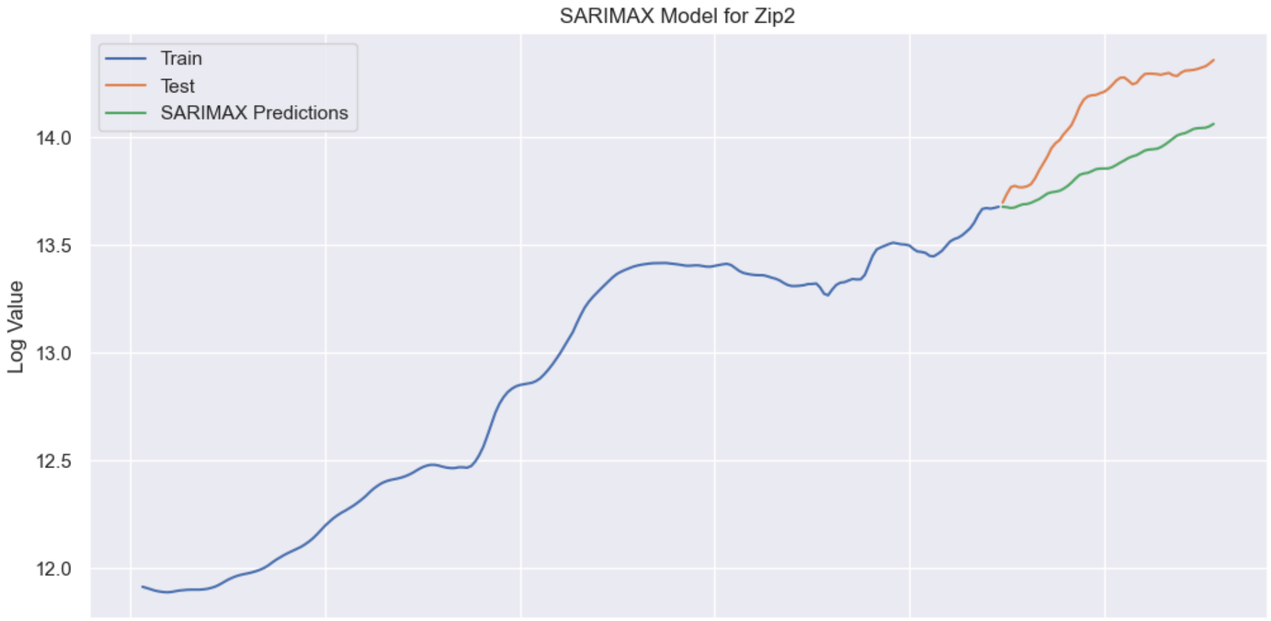
- Key Contributions: The ARIMA models show significant contributions from key parameters.
- Good Data Fit: High log likelihood and low AIC/BIC values suggest the models fit the data well.
- Predictive Accuracy: The models provide reasonable predictions, with lower error metrics indicating better performance.

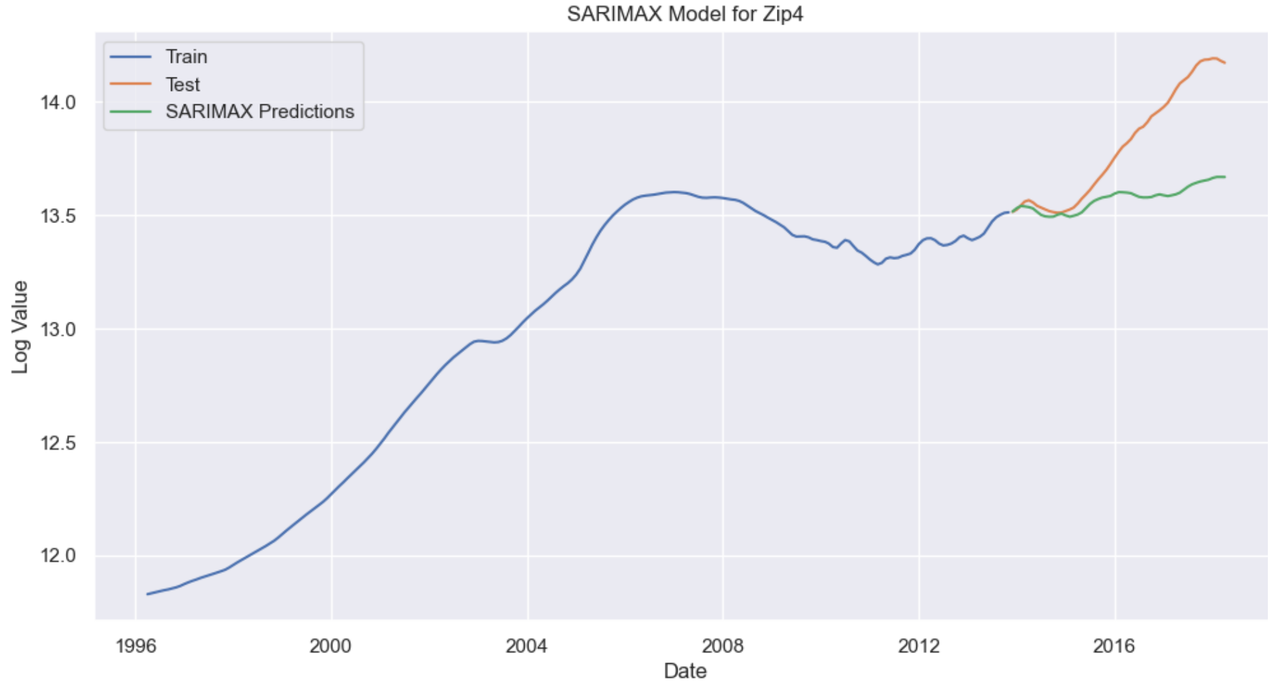


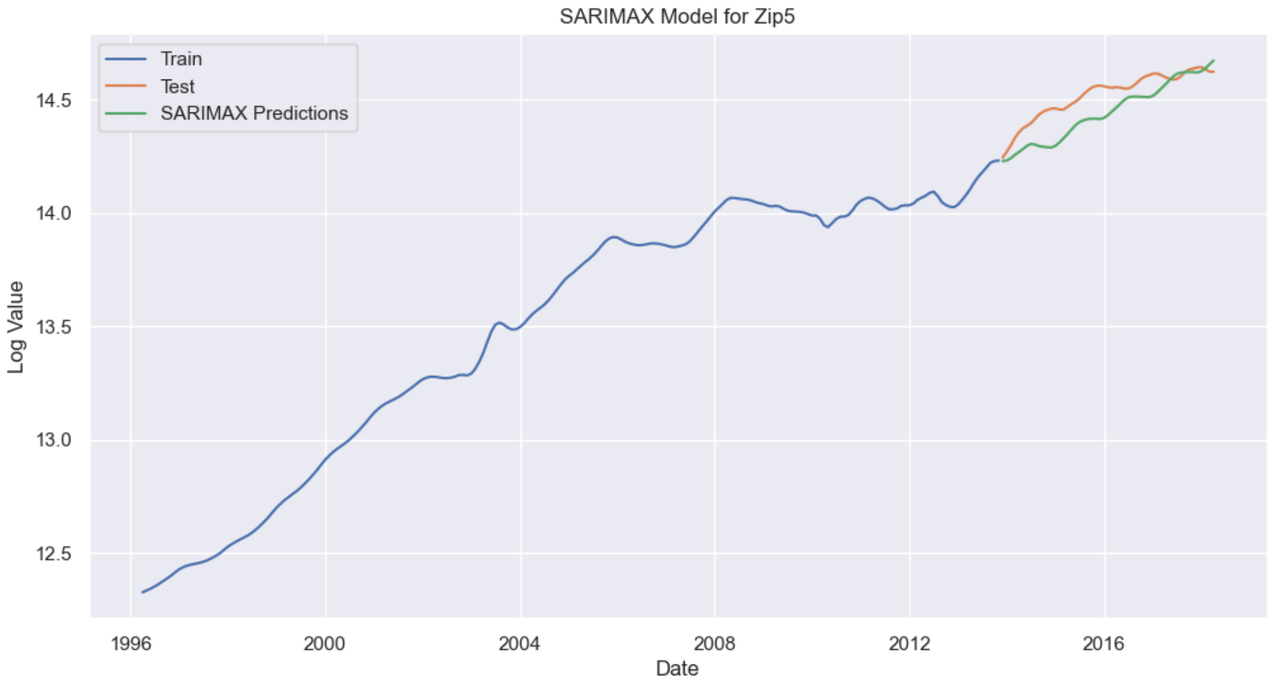
SARIMAX

Modeling

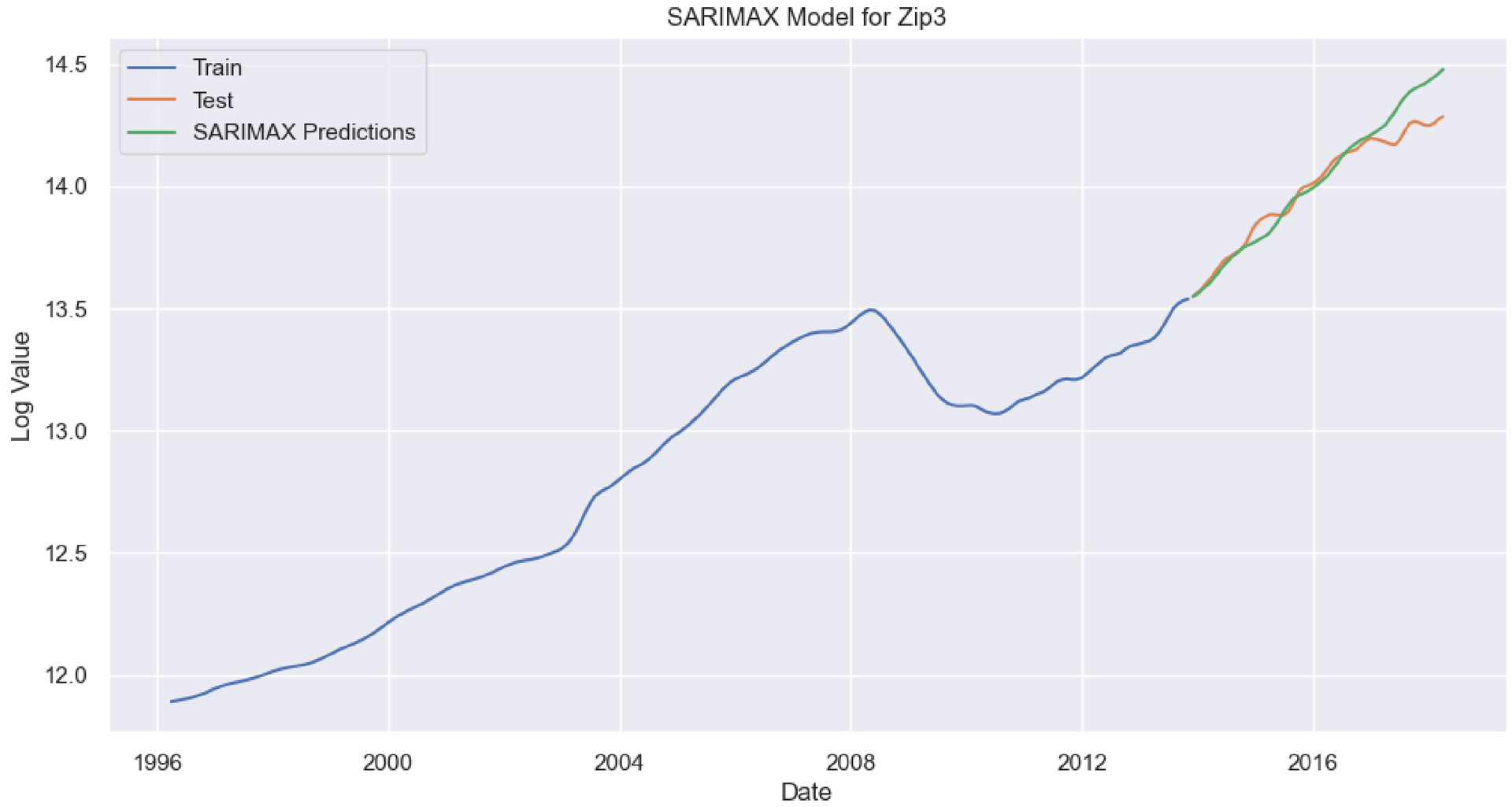




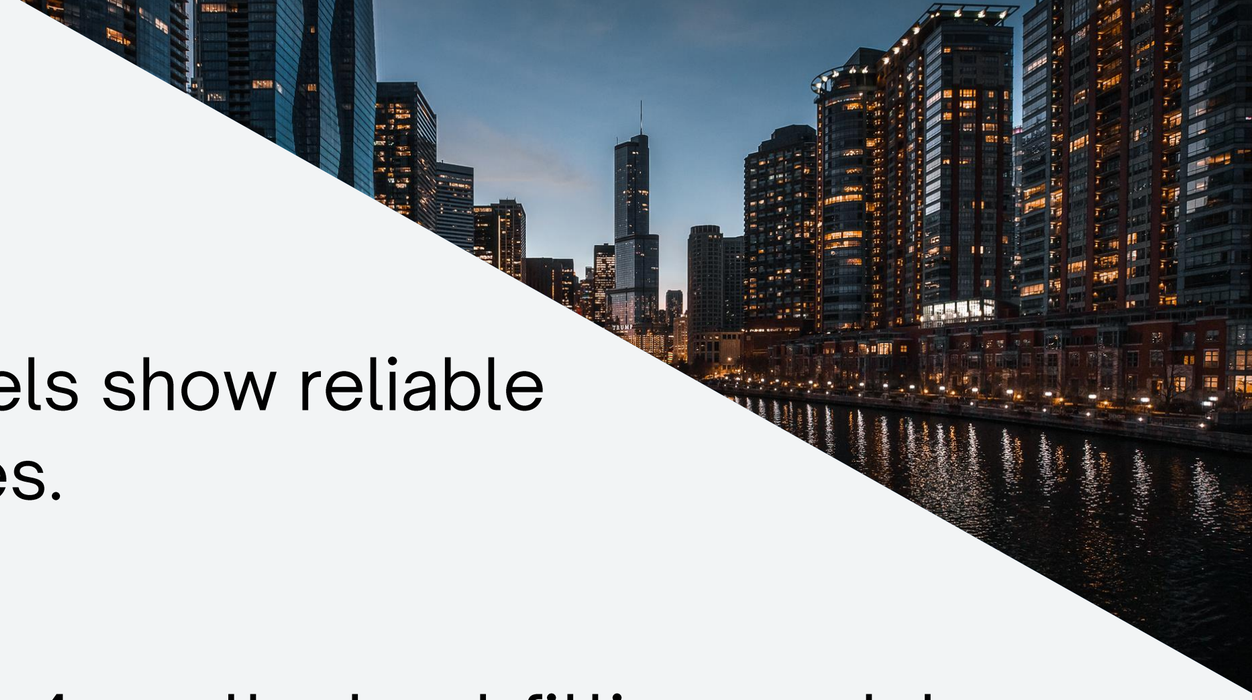




STATISTICS : SARIMAX



Zip3 stands out with the highest likelihood and lowest AIC, suggesting it has the best model among the five ZIP codes.



RESULT ANALYSIS

- **Strong Model Reliability**: All models show reliable parameter estimates.

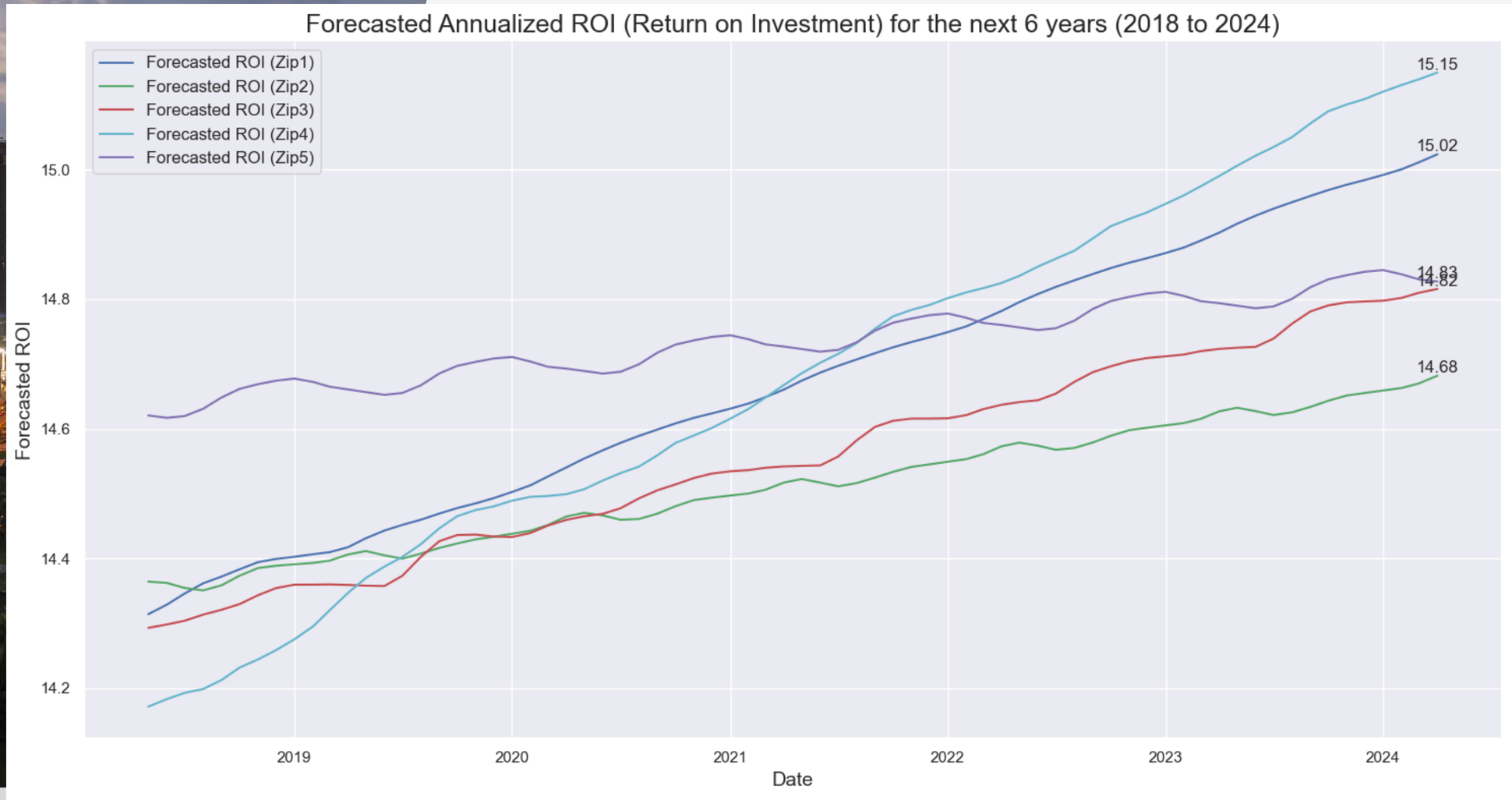
- **Top Performers:** Zip3 and Zip4 are the best-fitting models overall.

- **Residual Concerns:** Zip1, Zip2, Zip4, and Zip5 may have issues with data consistency.

MODEL DEPLOYMENT (SARIMAX)

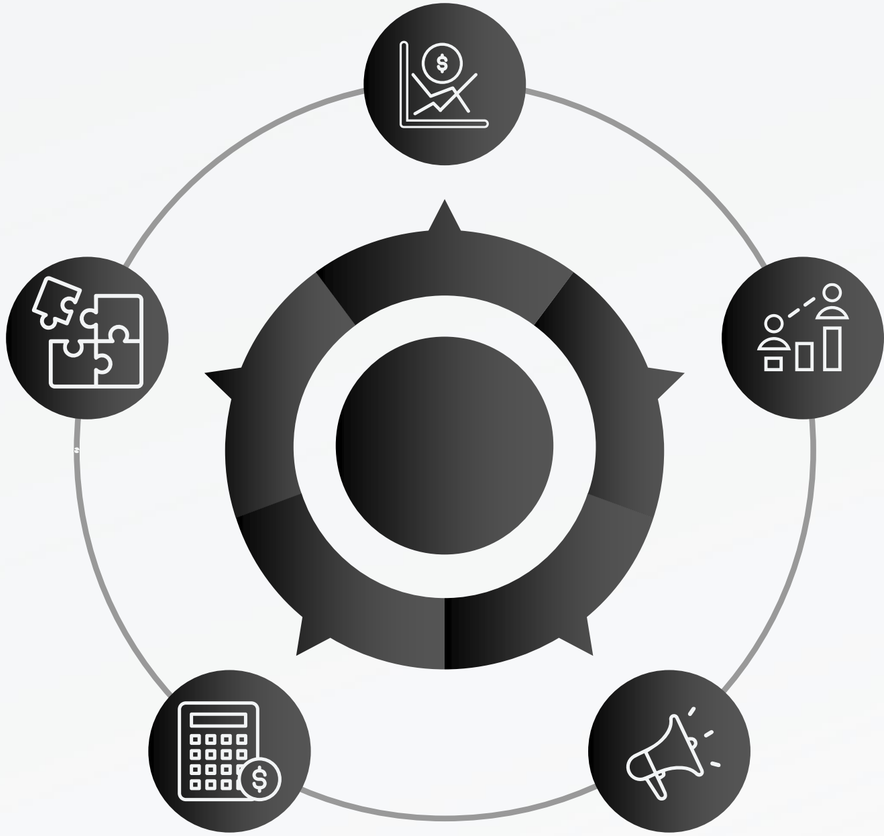
- SARIMAX avoids underestimating or overestimating future values
- Capability to handle seasonality inherent data
- SARIMAX lies in its ability to incorporate external factors, denoted as exogenous variables.

The forecasted ROI (2018-2024), Zip 3 is expected to perform the best at 14.8%, while Zip 5 is expected to perform low at 7.4%.



CONCLUSION

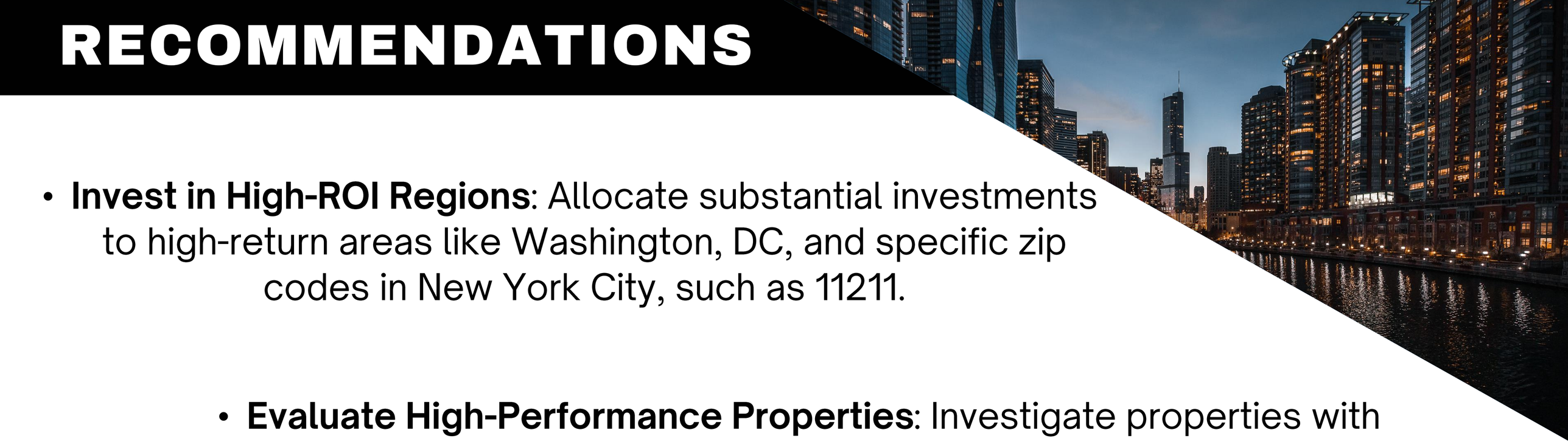
- Specific zip codes such as 11211 in Brooklyn consistently show high annualized returns. Investing in these areas could yield significant returns over time.



- While high-ROI areas offer potential for growth, balancing the portfolio with stable markets can safeguard against market fluctuations.

- To mitigate risk, it's advisable to diversify investments across multiple regions and property types

RECOMMENDATIONS



- **Invest in High-ROI Regions**: Allocate substantial investments to high-return areas like Washington, DC, and specific zip codes in New York City, such as 11211.

- **Evaluate High-Performance Properties**: Investigate properties with exceptional metrics to understand their potential for high returns or associated risks. Consider including select outliers in your portfolio if they match your risk tolerance and goals.

- **Diversify Investments**: Create a diversified portfolio by spreading investments across different regions and property types. This strategy will help manage risk while maximizing potential returns from both high-growth and stable markets.


THANK
YOU

OUR TEAM

1. Caleb Kenyatta
2. Sonia Ojay
3. Shuru Ebale
4. Magdalene Ondimu
5. Yvonne Mwangi
6. Joseph Karumba

Group 1: Phase 4

https://github.com/KirigoY/Group1_Phase4_Project/blob/master/starter_notebook.ipynb