Property Appreciation Estimation and Recommendation for Strategic Real Estate Investments

Introduction

Problem Statement and Motivation

The U.S. housing market is influenced by a multitude of factors—historical housing prices, macroeconomic indicators (e.g., mortgage rates, GDP), seasonal variations, and even political cycles. For real estate investors, identifying locations with the highest potential returns and understanding the trajectory of property values is critical. This project aims to forecast property value trends at the zip code level and determine the top areas to invest in over a one-year horizon.

Objectives

1.Examine Influential Factors: Investigate the effect of mortgage rates, seasonal patterns, and political/election cycles correlate with property value trends.

2.Forecast Housing Price Changes: Use time series analysis and regression models to predict short-term (1 month), medium-term (1 quarter), and longer-term (1 year) changes in property values.

3.Identify Optimal Investment Regions: Pinpoint top U.S. zip codes that are likely to yield the highest return on investment over a one-year period.

Data

Data sources

1. Zillow Housing Data (Zillow Home Value Index - ZHVI and Zillow Home Value Forecast - ZHVF)

- Description:
    - The ZHVI serves as a critical metric, representing the typical home value and tracking market changes for homes within the 35th to 65th percentile range. This data is available in both a smoothed, seasonally adjusted format and a raw measure, allowing for different types of analysis. Our dataset includes house prices on a monthly basis for every zip code in the United States, providing comprehensive coverage at various levels of geographical granularity—county, city, state, and zip code.

    - The ZHVF provides us with A month-ahead, quarter-ahead and year-ahead forecast of the Zillow Home Value Index (ZHVI). ZHVF is created using the all homes, mid-tier cut of ZHVI and is available both raw and smoothed, seasonally adjusted. We use these forecasts as the target variables to measure the performacne of our models.
- Size: Approximately 26,338 rows × 305 columns.
- Coverage: County, city, state, and zip code-level granularity spanning over two decades.
- Reference: Zillow's official website (obtained via downloadable CSV). https://www.zillow.com/research/data/:

RegionID	SizeRank	RegionName	RegionType	StateName	State	City	Metro	CountyName	2000-01-31	...	2023-11-30	2023-12-31	2024-01-31
84603	7958	60601	zip	IL	IL	Chicago	Chicago-Naperville-Elgin, IL-IN-WI	Cook County	232456.542741	...	343421.323487	344257.361587	344565.242611
84604	20351	60602	zip	IL	IL	Chicago	Chicago-Naperville-Elgin, IL-IN-WI	Cook County	271796.201836	...	292289.531820	289469.058508	285246.999747
84605	21722	60603	zip	IL	IL	Chicago	Chicago-Naperville-Elgin, IL-IN-WI	Cook County	NaN	...	332176.136008	329449.819693	324993.520639

2. Mortgage30US Dataset (FRED - Federal Reserve Economic Data)

- Description: 30 year monthly fixed mortgage rates in the US from 2000 to 2024.
- Size: Approximately 1,297 rows × 2 columns.
- Reference:Federal Reserve Bank Website (obtained via downloadable CSV) https://fred.stlouisfed.org/series/MORTGAGE30US

DATE	MORTGAGE30US
2000-01-07	8.15
2000-01-14	8.18
2000-01-21	8.26
2000-01-28	8.25

Data Cleaning & Preprocessing

1. Zillow Housing Data

In [ ]:
```
df2 = pd.read_csv("Zip_zhvi_uc_sfrcondo_tier_0.33_0.67_sm_sa_month.csv")
print(df2.shape)
df2 = df2.sort_values(by="RegionName")
df2

# Convert date columns to datetime if needed and sort by date
date_columns = df2.columns[9:] # the first 9 columns are metadata
df2[date_columns] = df2[date_columns].apply(pd.to_datetime, errors='coerc
```

```
# Handle missing values - for example, forward fill NaN values
df2.fillna(method='ffill', inplace=True)
df2.fillna(method='bfill', inplace=True)

# Convert datetime columns to numeric values (e.g., Unix timestamp) for p
df2[date_columns] = df2[date_columns].apply(lambda x: x.view(np.int64))

# Calculate monthly percent change for property values for each ZIP code
df2_pct_change = df2.groupby('State')[date_columns].pct_change(axis=1)

# Adding zipcode column to the df1_pct_change
df2_pct_change.insert(0, 'State', df2['State'])

# Display the trend over time for each ZIP code
print(df2_pct_change)

# Save the results to a CSV file
df2_pct_change.to_csv("df2_pct_change.csv", index=False)
```

2. Mortgage Data

- **Adjusting Mortgage Rate Data (Flipping Across the Y-Axis)**: Given that mortgage rates are inversely related to property appreciation (i.e., as mortgage rates increase, property prices generally decrease), we flipped the mortgage rate data across the Y-axis. This transformation allowed for a direct, more meaningful comparison of how fluctuations in mortgage rates correlated with changes in property values, making the analysis more intuitive for understanding their relationship.

In [ ]:
```
fig = px.line(mort_df, x='DATE', y='MORTGAGE30US', title='24-Year Fixed M

# Update y-axis range
fig.update_yaxes(autorange="reversed")

# Show the plot
fig.show()
```

3. Preprocessing and feature engineering for ML analysis

- Melted table to transform the wide-format table into a long-format table

In [ ]:
```
df_melted = zhvi_dataset_clean.melt(id_vars=['RegionName'], var_name='Mon
```

- Created lag feautes and normalized the data

In [ ]:
```
#generate lag features
for lag in range(1, 13):  # Lags 1 to 12 months
    df_melted[f'Price_t-{lag}'] = df_melted.groupby('RegionName')['Price'
df_melted = df_melted.dropna()

#Set up scalers
def scale_data(X,Y):
    Xscaler = StandardScaler()
    Yscaler = StandardScaler()
```

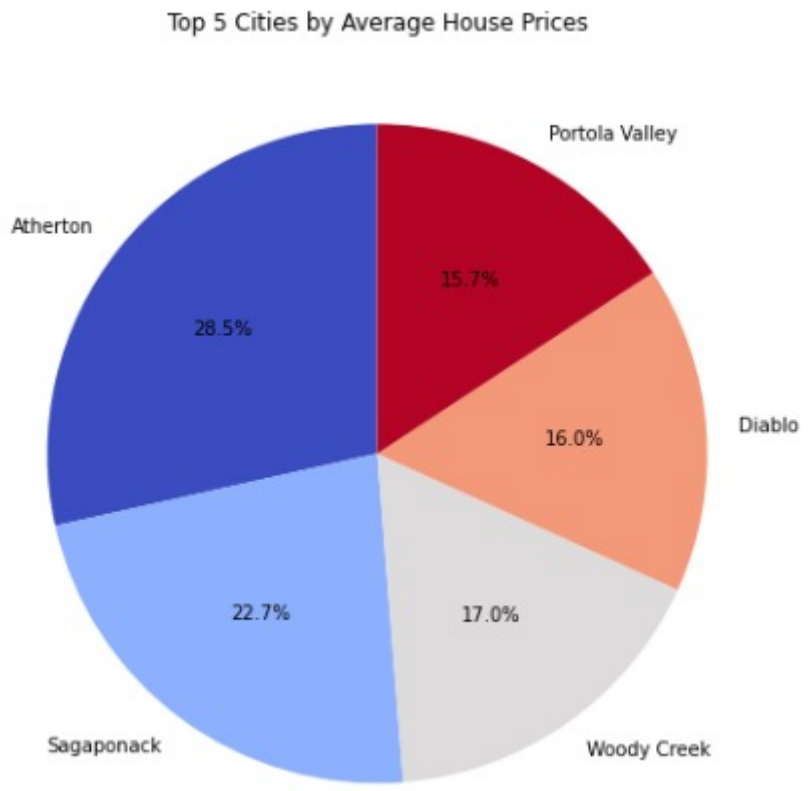
```
    #Fit the scaler on the feature data and transform it
    Xscaler = Xscaler.fit(X)
    Yscaler = Yscaler.fit(Y)
    return Xscaler, Yscaler
```

- Preprocessed the mortgage data by averaging the three monthly recordings to obtain a single value per month, preparing it for integration with the price dataset.

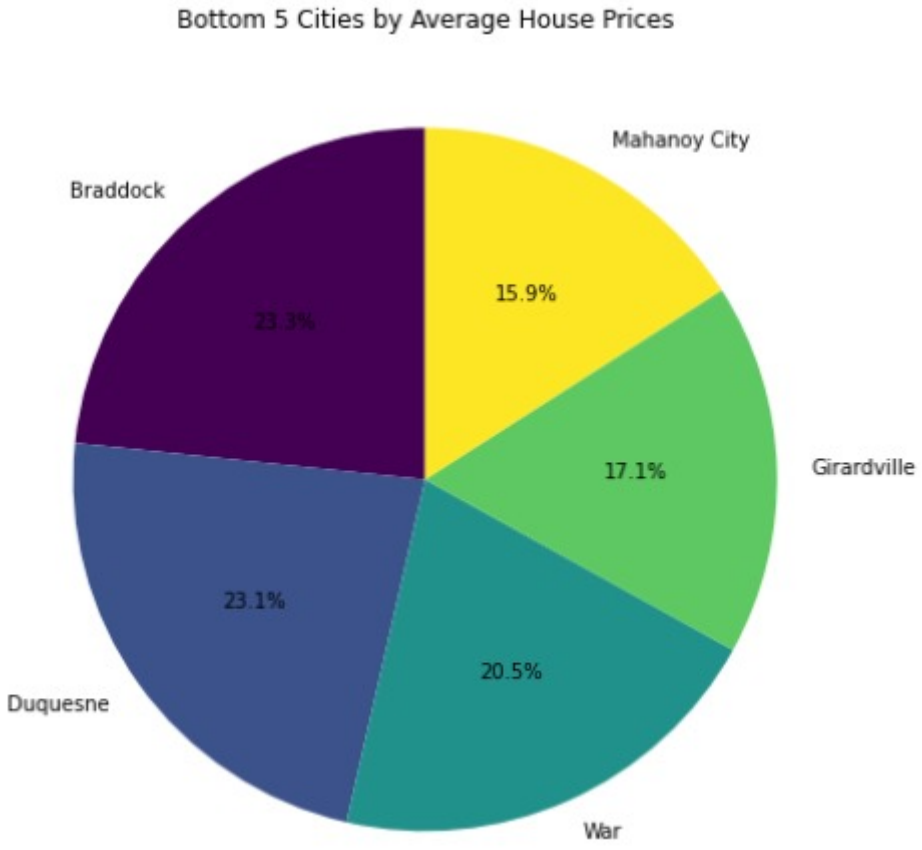
In [ ]:
```
mort_df['DATE'] = pd.to_datetime(mort_df['DATE'])

# Extract year-month from 'DATE' in b
mort_df['YearMonth'] = mort_df['DATE'].dt.to_period('M')

# Calculate the monthly average mortgage rate
monthly_avg_mortgage = mort_df.groupby('YearMonth')['MORTGAGE30US'].mean(
monthly_avg_mortgage.rename(columns={'MORTGAGE30US': 'Avg_Mortgage'}, inp
```

Exploratory Data Analysis and Visualization

Top 5 Cities with the Highest and Lowest Average House Prices in the United States



- Atherton: Represents the largest share of average house prices among the top cities, accounting for 28.5%, indicating it is the most expensive city in this group.
- Sagaponack: Comes second with 22.7%, followed by Woody Creek (17.0%), Diablo (16.0%), and Portola Valley (15.7%).
- The distribution shows that Atherton and Sagaponack dominate, while the remaining three cities contribute relatively smaller proportions.



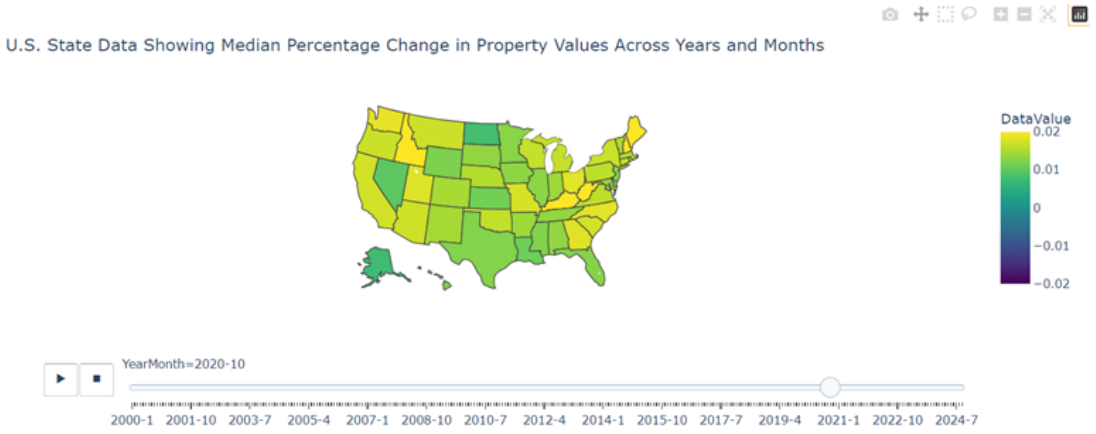
- Braddock and Duquesne: These two cities dominate the chart, with nearly equal shares of 23.3% and 23.1%, indicating their house prices are slightly higher than the other three.
- War: Represents 20.5% of the total, followed by Girardville (17.1%) and Mahanoy City (15.9%).

Observation

- The first graph demonstrates the disparity in house prices among the top cities, with Atherton being significantly more expensive.
- The second graph shows a more balanced distribution of house prices among the bottom cities, emphasizing affordability and lesser variability.

Influence of Macroeconomic Indicators

Assumption: Mortgage rates reached record lows during the latter half of 2020 and into early 2021. This is the same period during the Covid pandemic, when the demand for housing kept increasing. There are macroeconomic indicators that can explain this trend.



Worked out:

- Averaged out the percentage change in property values across states to get a single trend representing the whole of the U.S.
- Observed various macroeconomic indicators like mortgage rates, GDP, Federal interest rates during the period and tried to fit a factor that can tightly reason for the trend.

Observation: The median percentage change in property values (from ZHVI – Zillow Home Value Index) across different states considering the dataset's granularity on zip codes stays extremely positive (+) during the end of 2020 through 2021. This was observed from the interactive choropleth map.

Finding:

- The inverse graph of mortgage rates is fitting the average percentage increase trend in property values across the U.S.

Seasonal Trend Analysis

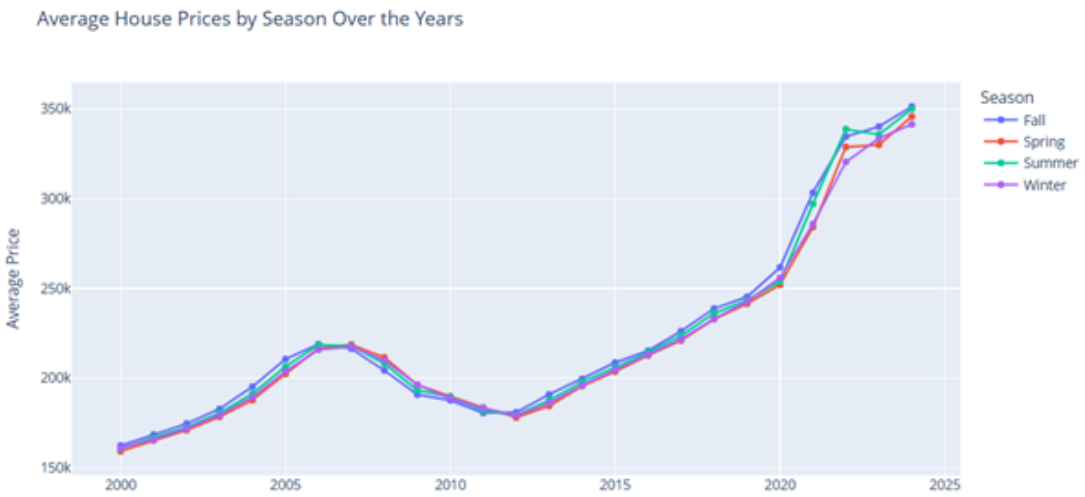
Introduction: In this analysis we investigate potential seasonal trends in average house prices over the years. Using raw data with a datetime column, we categorized each entry into seasons (Fall, Winter, Spring, Summer) based on the month. This approach allows us to analyze any fluctuations in house prices across seasons over a long period.

Methodology:

1. Data Preparation: The raw data contained a datetime column, which was converted into rows. We then added a new column, "Season," categorizing each month into Fall, Winter, Spring, or Summer.
2. Graph Analysis: After adding the seasonal column, we plotted a line graph to compare the average house prices for each season over the years. The graph below illustrates these seasonal trends.

Observations:

- The chart displays the average house prices over the years from around 2000 to 2025.
- Each line represents house prices for a different season (Fall, Spring, Summer, Winter).
- There's a general upward trend in average house prices over time, with notable dips and peaks.



Assumptions:

- The dataset used for this chart includes sufficient historical data on seasonal average house prices, possibly sourced from a reliable real estate database.
- The chart aims to show whether there's a seasonal effect on house prices, looking for distinct price differences between seasons over the years.

Findings:

- Similar Trends Across Seasons: All four seasons display similar patterns in price increases and decreases over time, with minimal differences between them,

especially after 2015.
- Peaks and Troughs: There's a peak in prices around 2005, a dip after 2007, a steady low around 2010, and then significant growth from around 2015 onward.
- Minor Seasonal Differences: Seasonal variation appears minor, with all seasonal lines staying close to each other across the years.

Conclusion:

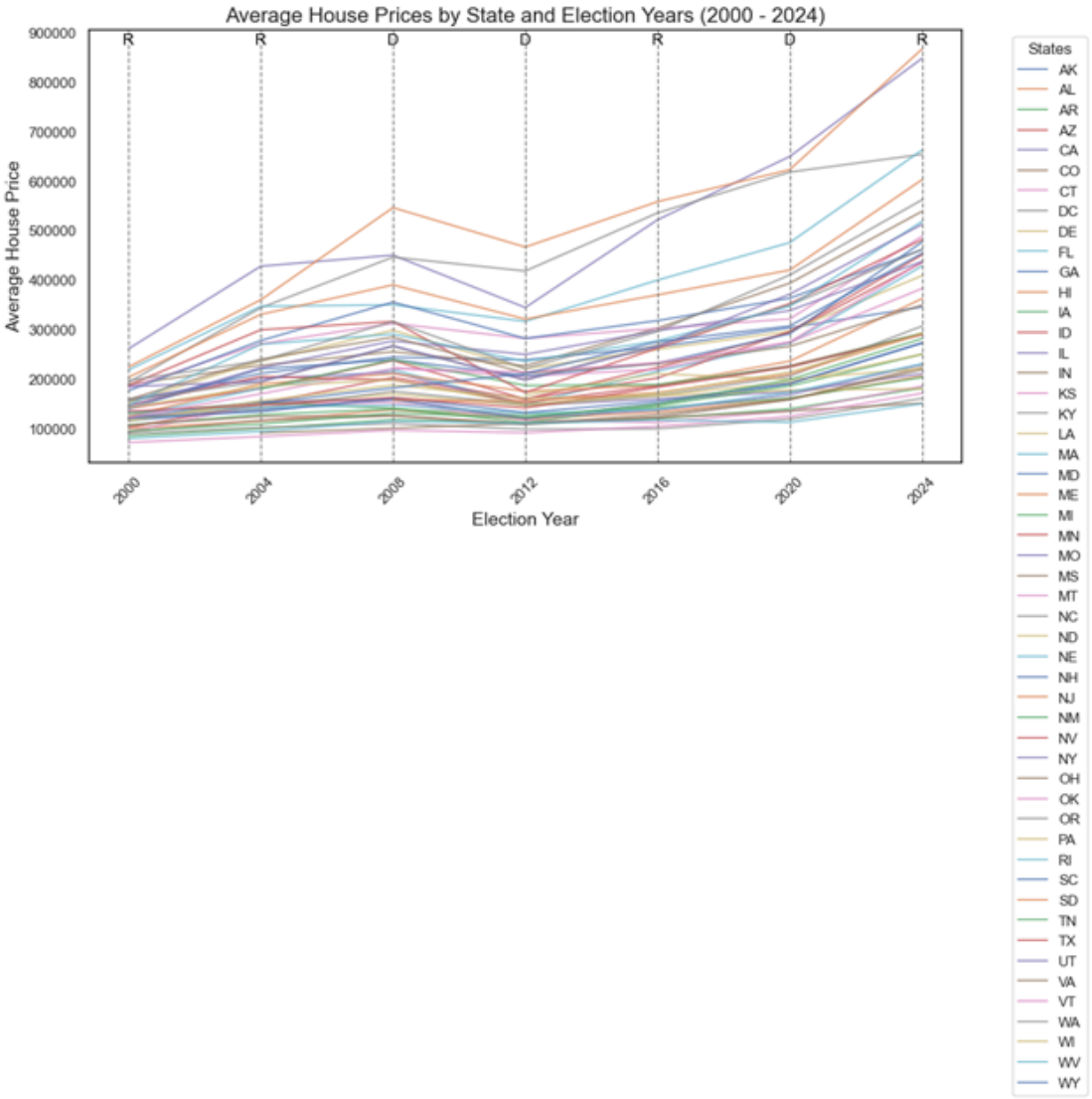
- The average house prices appear to follow a consistent pattern across seasons, with no significant seasonal effect observed in recent years. Prices largely vary based on broader economic cycles rather than seasonal factors, suggesting that, at least in this dataset, seasonality doesn't have a strong impact on average house prices over the years.

Elections vs House Prices

- In the years 2000-2008, during the Republican Administration under Geoge W Bush, it was considered as the period of economic expansion. Interest rates were low, the availability of credit was easy, and minimal regulations around mortgages led to a housing boom which encouraged the investments in real estate. The American Dream Downpayment Act made it easier for people to own homes.
- But during the global financial crisis in 2008, this steep decline in home prices across the US as foreclosures surged and the housing dema d dropped, that created the economic downturn.
- In 2008-2016, during the Democratic Administration under Obama, the ARRA (American Recovery and Reinvestment Act) in 2009 aimed to stimulate the economic growth and support recovery, ideally it tried to stabilize the situation which can be seen from the graph. The recovery started to restore confidence in the economy. Priority was also given to affordable housing and assistance, though they had a limited effect on curbing the overall rise in house prices due to high ongoing demand and slow supply growth
- During the Republican administration in 2016 - 2020 under Trump underwent through a major policy, Tax Cuts and Jobs Act, that had provisions for lowering taxes for some individuals, which boosted disposable income for many and eventually made the housing market bit more welcoming, which made the house prices rise and that can be seen from the upward trend.
- During 2020-2021, mortgage rates were at their all time low, also covid brought demand in housing which caused a rise in demand, subsequently rise in property rates.

Prediction based on recent elections:

- In the 2024 fall, the feds have decreased the interest rates for the first time after 2020, also the elections held in November 2024, the republican party has come into power, substituting the democrats. Predicting the previous trends, it can be projected to be increasing ahead, since fed rates are low, this leads to lower mortgage rates increasing the housing demand and higher house rates.



Modeling and Statistical Analysis

Models

1. AutoRegressive Integrated Moving Average (ARIMA)
2. Linear Regression

Evaluvation Metrics

- Average Absolute Difference: This metric measures the absolute difference between the model's predicted values and Zillow's ZHVF estimates.
- Directional Accuracy: This metric evaluates the percentage of predictions that correctly capture the direction (positive or negative) of change in comparison to Zillow's ZHVF values.

Rationale for Including Directional Accuracy

Directional accuracy is included to assess whether the model can at least capture the correct trend of appreciation or depreciation, even if the absolute difference is high.

Since Zillow uses a highly complex model to generate ZHVF values, the absolute difference may naturally be larger. By incorporating directional accuracy, we aim to determine whether the model successfully captures the overall direction of the trend, providing additional insights into its predictive reliability.

1. ARIMA (AutoRegressive Integrated Moving Average) Model:

- ARIMA was used to model the data based solely on historical prices, with an individual model fitted for each zip code.
- The models were trained and evaluated using four different time windows: the past 5 years, 10 years, 15 years, and the entire dataset.
- The goal was to determine how much of the historical price data is relevant for accurate future predictions.

[Insert image of graphs here]

Results

- Based on our metrics, both the absolute difference and directional accuracy scores improve when the entire dataset is used for future forecasting. Given these results, we decided to train our linear regression model using the entire dataset.

2. Linear Regression

Setup

- Similar to ARIMA, we fit a linear regression model to each zip code to predict its future prices.
- To train the linear regression model, we perform feature engineering to transform the data into a supervised learning problem. Specifically, we create lag features by using the 12 prior months' prices for each zip code, with the target variable being the price for the next month.
- To predict future prices, we use the most recent 12 months of prices for each zip code. We predict the next price, append it to the input data, remove the oldest price, and repeat this process until we have predictions for the desired time range.
- We experimented with different feature degrees to evaluate model performance, specifically testing degree = 1 and degree = 2.
- Our exploratory data analysis (EDA) revealed an inverse relationship between mortgage rates and housing prices, prompting us to include mortgage rates as a feature. We conducted hypothesis testing to evaluate whether their inclusion significantly impacts model performance by comparing metrics of models trained with and without mortgage rates.

Results

- The model achieves a directional accuracy of 47% and an average absolute difference of 55 when predicting house prices one year ahead.

- Increasing the model degree to 2 resulted in a slight reduction in the error term but increased both bias and variance:
  - Degree 1: Average RMSE: 0.0030 Bias: 8.33×10−78.33×10−7 Variance: 0.1664
  - Degree 2: Average RMSE: 0.0103 Bias: 3.10×10−53.10×10−5 Variance: 0.1778
  - Given these results, the simpler degree-1 model was chosen.
- Hypothesis testing revealed that including the mortgage rate does not yield a statistically significant improvement in model performance, with p-values for directional accuracy as follows:
  - One Month: 0.3441
  - One Quarter: 0.4093
  - One Year: 0.4626
  - Despite this, the mortgage rate was included in the final model due to its slight improvement in directional accuracy (boosting it from 46% to 47%).

Inference

Using our linear regression model, we identified the top 5 zip codes for investment that are projected to yield the highest profits by September 2025.

```
x =['RegionName', 'RegionType', 'StateName', 'State', 'City', 'Metro',
    'CountyName']
res[x]
```

✓ 0.0s

	RegionName	RegionType	StateName	State	City	Metro	CountyName
0	48505	zip	MI	MI	Flint	Flint, MI	Genesee County
1	61605	zip	IL	IL	Peoria	Peoria, IL	Peoria County
2	36610	zip	AL	AL	Prichard	Mobile, AL	Mobile County
3	71103	zip	LA	LA	Shreveport	Shreveport-Bossier City, LA	Caddo Parish
4	62914	zip	IL	IL	Cairo	Cape Girardeau, MO-IL	Alexander County

Discussion and Next Steps

Model Limitations

- Model Complexity: The linear regression model struggles to accurately capture the non-linear relationships that are often present in housing market trends, especially since we limited exploration to models with a maximum degree of 2.
- Feature Engineering: Our model was trained using only raw price data and mortgage rates.

Future Enhancements

- Advanced Modeling Techniques:
  - Implement machine learning models like XGBoost, Random Forests, or neural

networks to capture non-linear relationships and interactions between variables.

- Explore ensemble methods to combine the strengths of multiple models.

- Enhanced Feature Engineering:
  - Introduce time-series-specific features such as moving averages, seasonality adjustments, and more lagged variables.
  - Include other indicators like population growth, crime data, umemployment rate and median income of each zipcode.

Conclusion

This project leveraged a comprehensive dataset of U.S. housing prices, enriched with mortgage rates and contextualized by economic and political factors, to forecast property value changes and identify high-potential investment opportunities. Through ARIMA analysis, we found that using the entire dataset yielded optimal results for forecasting future prices. Our linear regression model with degree=1 provided the best balance of simplicity and performance, with the inclusion of mortgage rates marginally improving accuracy. Additionally, our analysis confirmed a significant inverse relationship between mortgage rates and housing prices, aligning with economic theory. With these insights, we identified several promising zip codes for future investments.

While our approach successfully identified promising zip codes for future investments, we acknowledge the limitations of simpler models. Advanced machine learning techniques and the integration of additional features—such as crime rates, unemployment levels, etc—are necessary to further enhance predictive accuracy. Zillow's estimates, for instance, leverage neural network models and richer data inputs.

Peer Evaluation

- Armaan: Conducted ML analysis using ARIMA and Linear Regression models, implemented feature engineering (e.g., lag features, macroeconomic variables), and evaluated model performance metrics.
- Abhiram: Explored regional and temporal trends, cleaned datasets for consistency, and produced visualizations (e.g., maps, trend lines) to support analysis.
- Vamsi: Data collection from Zillow and FRED, explored dataset structure and quality, and identified key variables (e.g., mortgage rates, property values) for analysis and the EDA for macro-economic trends vs property appreciation.
- Sushanth: Preprocessed data by calculating percentage changes, aligning mortgage rates with property values, and generated visualizations showcasing inverse

correlations and seasonal trends(EDA)

- Niyati: Bridged EDA and ML phases by uncovering trends (seasonal, macroeconomic), implementing ARIMA models, and validating their performance against observed data.
- Purva: Focused on cleaning datasets (handling NaN, formatting columns) and conducted EDA to reveal state-level and election-cycle trends through visualizations.