# PREDICTING HOUSE PRICES IN ARIZONA

**Sponsor: Freddie Mac**

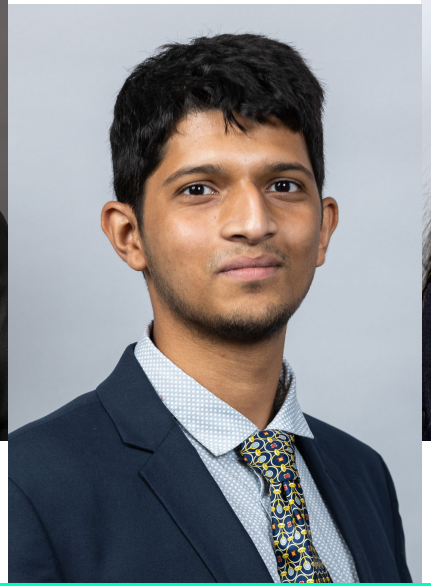
**Team: Anirudh Murali, Akshaya Krishnan, Neha Sharma, Rahul Murugkar, Shrushti Shah**

SHRUSHTI SHAH

ANIRUDH MURALI

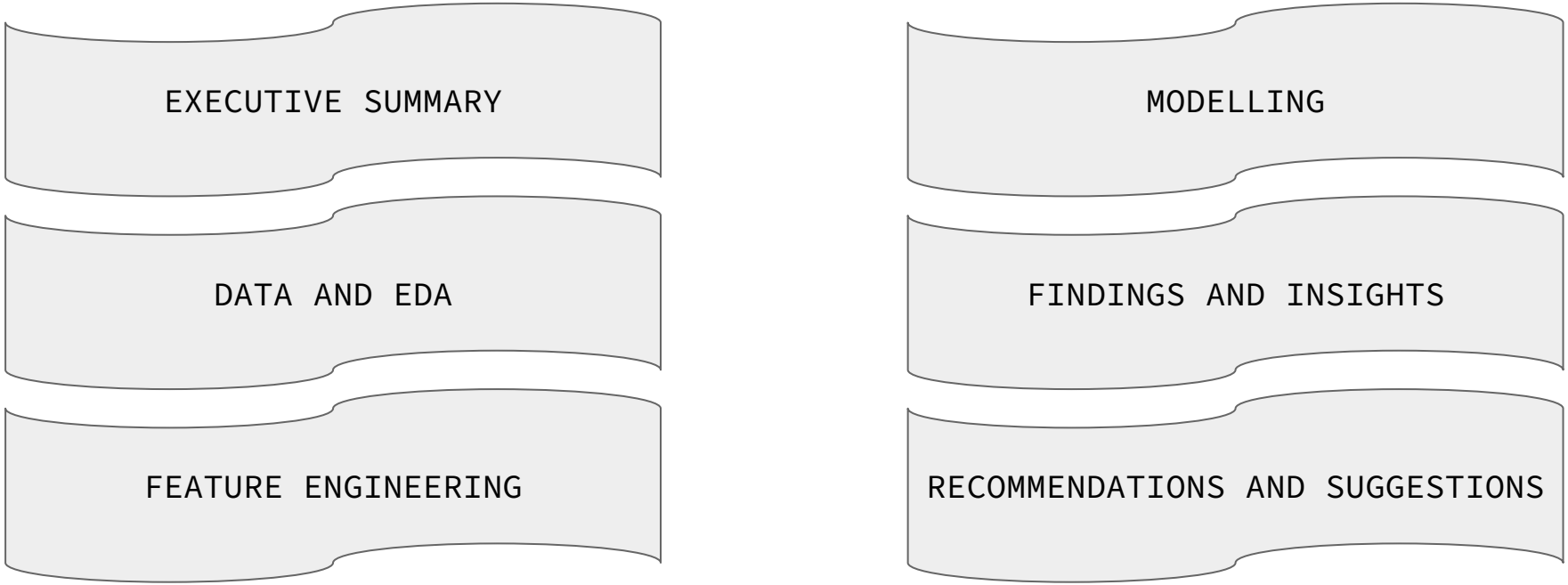
AKSHAYA KRISHNAN

RAHUL MURUGKAR



NEHA SHARMA

# TEAM


EXECUTIVE SUMMARY
DATA AND EDA
FEATURE ENGINEERING
MODELLING
FINDINGS AND INSIGHTS
RECOMMENDATIONS AND SUGGESTIONS

# AGENDA

# Executive summary

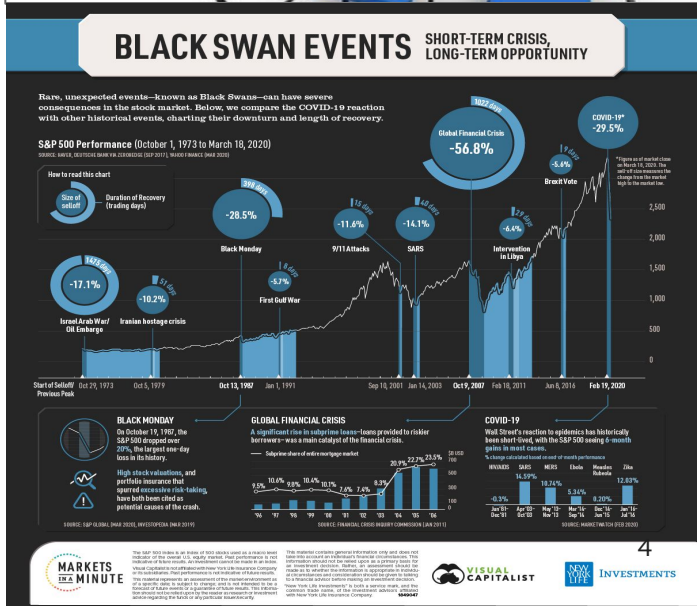
## PURPOSE

→ To minimize the risks associated with real estate investments

→ Anticipating market trends and black swan events

→ Developing financial planning or portfolio management, and having an accurate predictive model

→ Emphasizing on developing models that perform well during Black Swan events like COVID-19

# Executive summary

## GOALS

→ Develop Accurate Forecasting Models

→ Black Swan Event Resilience

→ Quick Recovery and Adjustment

→ Enhance Decision-Making Capabilities

# PROJECT OBJECTIVE AND BUSINESS PROBLEM OVERVIEW

**Business Problem:**

- The challenge at hand is creating a predictive model for Arizona's House Price Index (HPI) that can reliably detect and foresee black swan events.
- Financial crises, natural catastrophes, or any other unforeseen circumstances that have a major impact on the Arizona real estate market could be considered among these occurrences.

**Objective:**

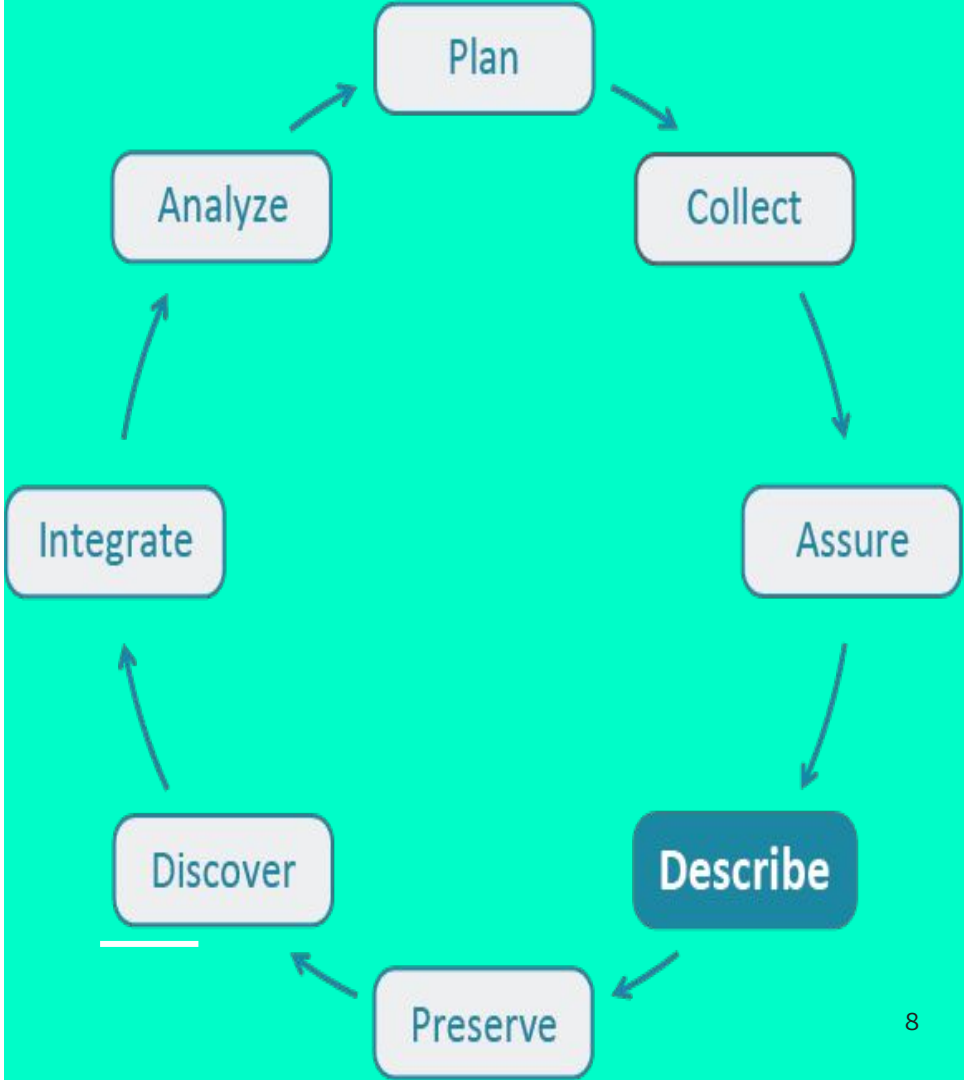
- The main goal of this project is to create, put into practice, and verify a predictive model for the Arizona House Price Index, with an emphasis on spotting and reacting to black swan events.

# The DATA

# DATA DESCRIPTION

- Timeframe for Analysis:

- Initial dataset was from 1975-2023 with 1849 rows
- Utilized data from 2000 to 2022.
- Trained models on 2000-2019 data; 2020-2022 used for validation.



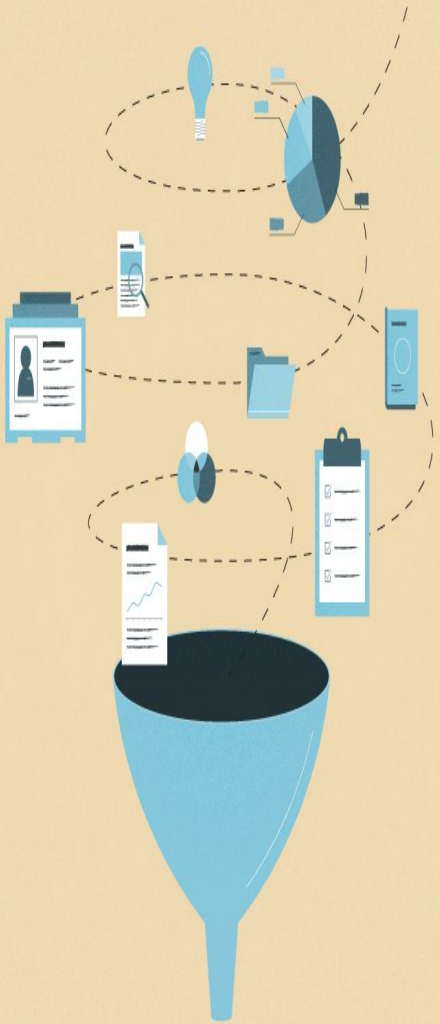
# FEATURES

- Consumer Price Index(CPI)
- Gross Domestic Product(GDP)
- Mortgage
- Unemployment Rate
- Median Income
- Population
- Property Tax
- Construction Permits
- 90-Day Delinquency
- Median Price  Per Sq. Ft.

# External sources

- Economic Data
- Building Permits Survey Data
- Consumer Price Index for All Urban Consumers
- Mortgages 30–89 days delinquent
- Mortgage Rates - Freddie Mac
- Yuma, Arizona, had second highest unemployment rate but largest over-the-year decrease
- Despite high unemployment, Yuma’s agribusiness continues to thrive
- Sierra Vista area ranked 3rd-most affordable housing market

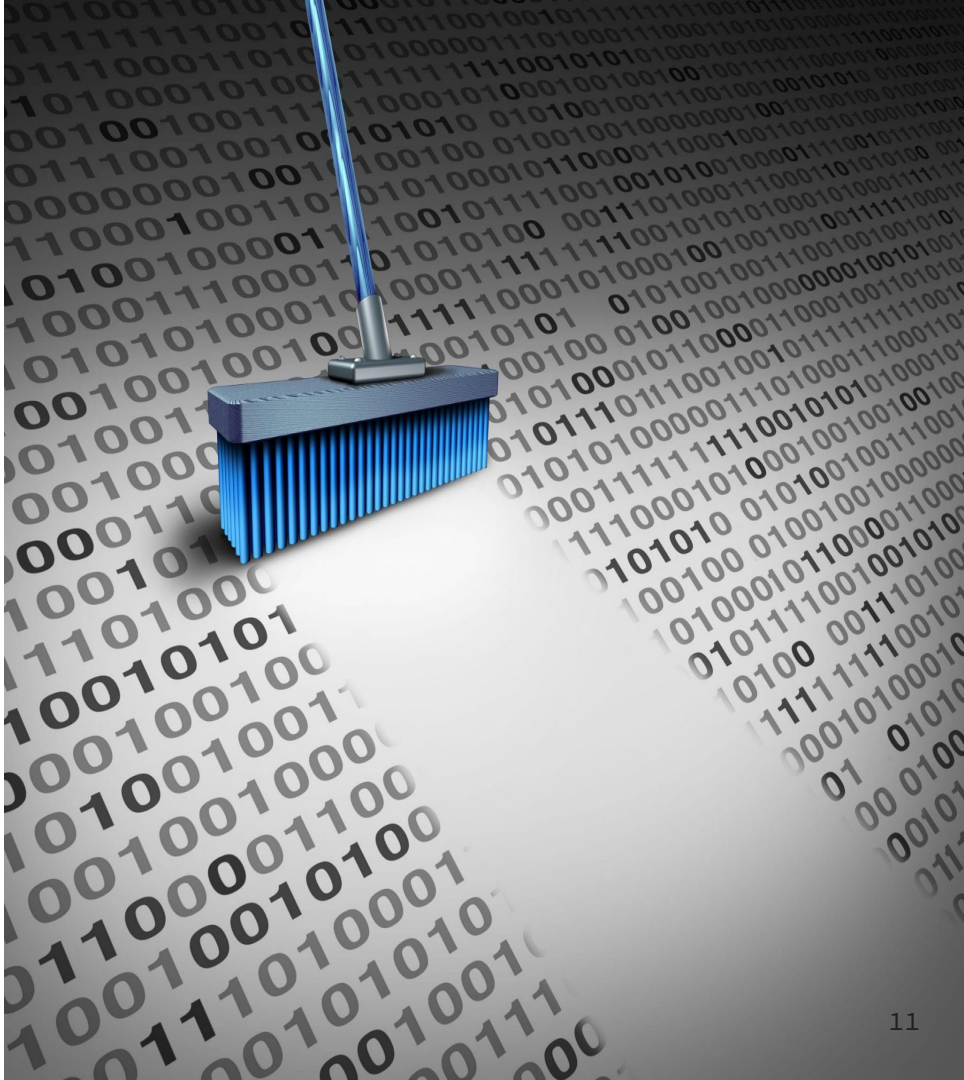
*Click on the above bullets to be directed to the source*



# DATA CLEANING AND PREPARATION

In preparing our dataset for analysis, we encountered variations in the availability of certain features across different geographical levels and timeframes. To maintain consistency and facilitate meaningful comparisons, we implemented a thoughtful approach to handle missing or incomplete data. The following outlines our methodology:

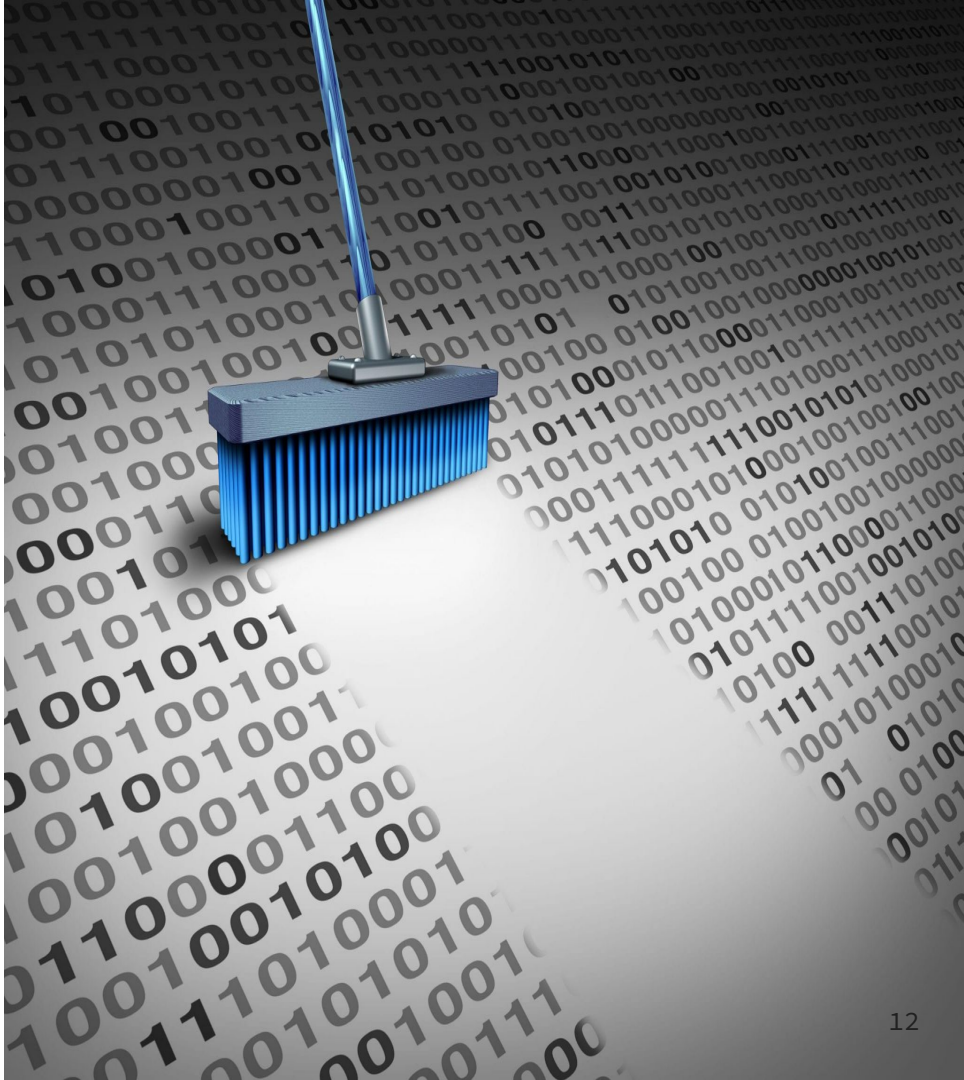
**State-Level Features:**
For features available only at the state level (e.g., Mortgage rates, CPI, Median Income, Property Tax), we adopted a pragmatic approach. Given the absence of specific data at the Metropolitan Statistical Area (MSA) level, we assumed uniform values for these features across all MSAs within Arizona for the corresponding time frames.



# DATA CLEANING AND PREPARATION

**Yearly Features:**
In cases where certain features were available on a yearly basis (e.g., CPI, Median Income, Population, GDP), we extrapolated these values uniformly across all twelve months of the respective year. This assumption allows us to maintain a monthly granularity while working with annually reported data.

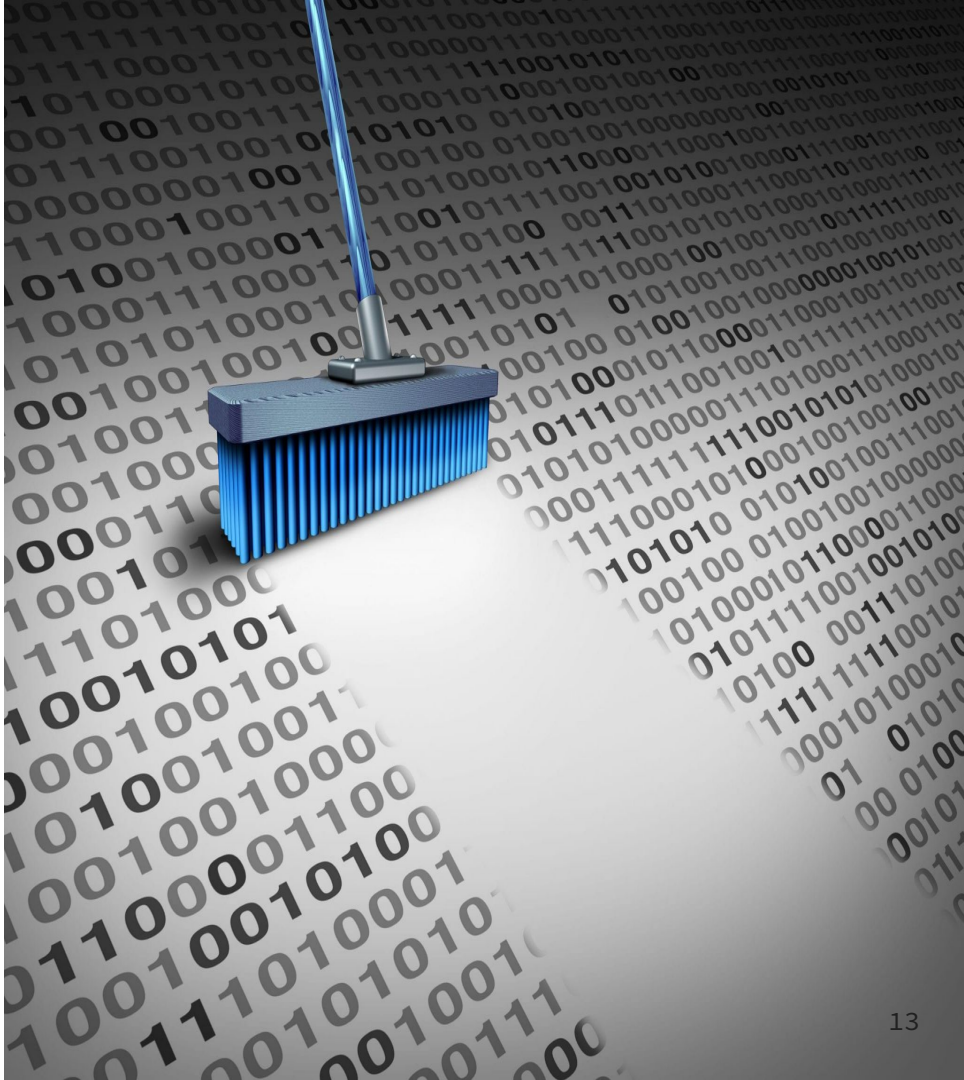
**Monthly Features:**
Features reported on a monthly basis (e.g., Mortgage rates, Unemployment Rate, Construction Permits) presented a straightforward integration. We utilized the available monthly data as reported without any imputations or assumptions, ensuring the preservation of temporal dynamics.



# DATA CLEANING AND PREPARATION

**Quarterly Features:**
For features reported quarterly (e.g., Property Tax), we uniformly distributed the reported values across the respective three months within the quarter. This approach aligns with our objective to maintain a consistent level of detail throughout our dataset.

**Rationale:**
Our approach balances the necessity for granularity and uniformity, acknowledging the limitations in data availability and reporting practices. While assumptions were made, they were applied judiciously to ensure a robust and coherent dataset for subsequent analysis.



# EXPLORATORY DATA ANALYSIS

## Will Arizona's Housing Market Crash?

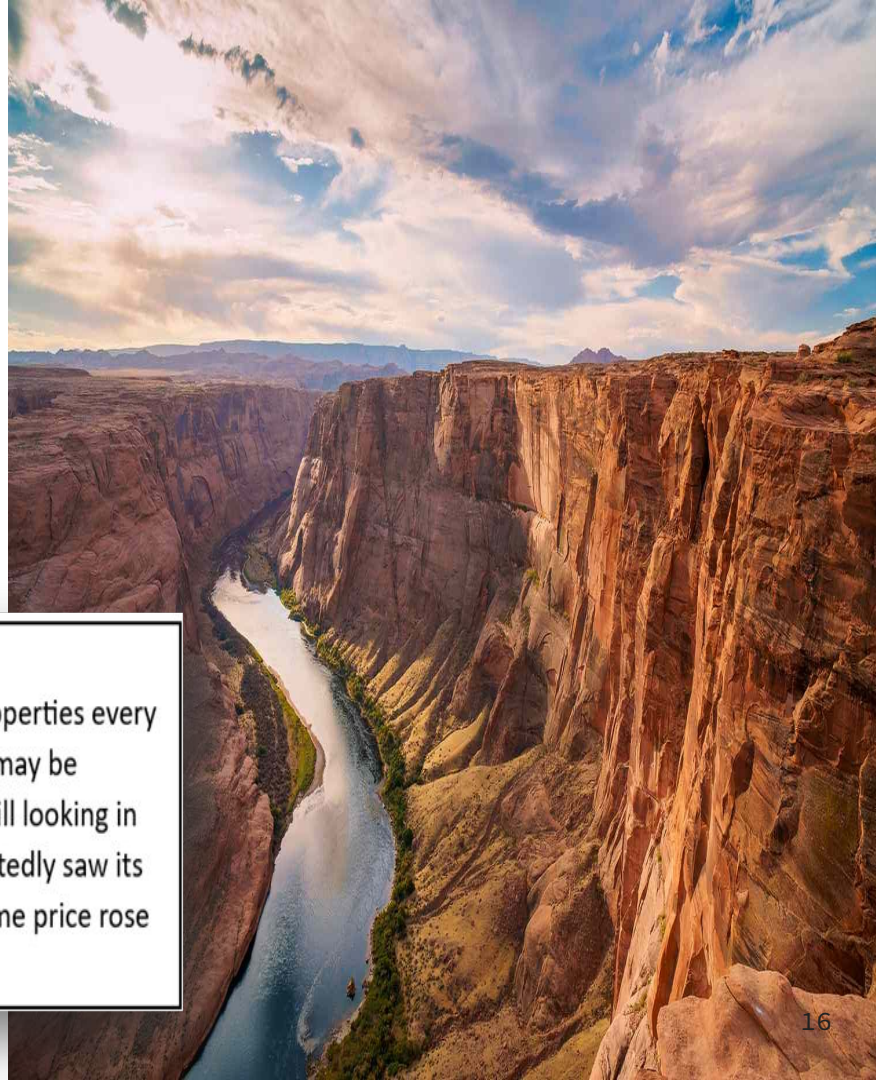
Some are pessimistic about the future of the Arizona housing market, especially in Phoenix.

At the end of 2022, the National Association of Realtors predicted a 15.8% drop in combined sales and prices for the Phoenix-Mesa-Scottsdale market in 2023. Tucson was expected to fare only slightly better with a 10.2% decline.

Despite the decline in sales and available listings, home prices are on the rise. This price increase is common amongst states with lowing housing inventories. However, potential homebuyers in Arizona and elsewhere may be waiting for prices to come down and to see where mortgage rates, which have jumped considerably in the past year, go next.

# Phoenix real estate entered 2020 on a high, which should ease the pandemic's market impact

*Rebuilding America: Metro Phoenix's housing market slowed in the early days of COVID-19 but is showing signs of bouncing back.*



**THE GAME OF SUPPLY AND DEMAND**

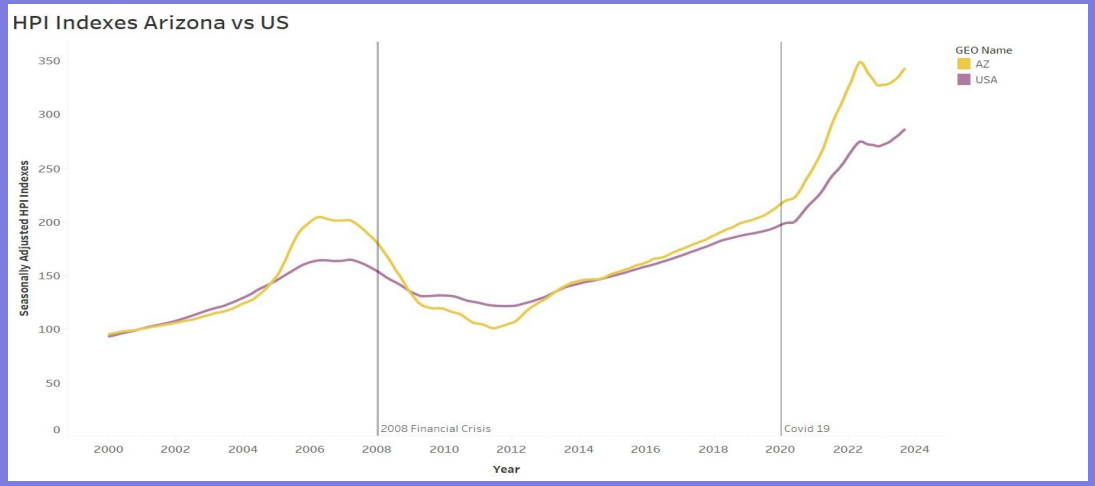
Before the outbreak, in some Arizona communities, sellers were actively listing properties every day. Now, homeowners who may have planned to put their properties up for sale may be choosing to wait until conditions improve. With reduced supply, buyers who are still looking in these markets have fewer options. For instance, in March, Flagstaff, Arizona, reportedly saw its already low inventory of homes for sale decline. At the same time, the median home price rose by six percent.

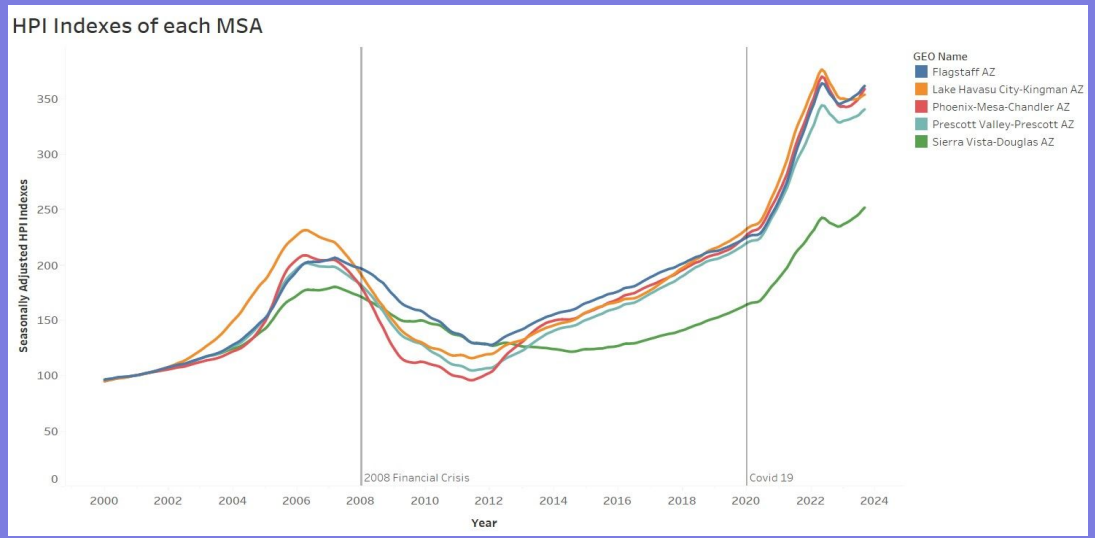
# House Pricing Index

According to the National Association of Home Builders, 92.9 percent of the homes sold in the third quarter of 2021 in the Sierra Vista-Douglas metropolitan statistical area were affordable to families earning the area's median income of $66,900.

That ranked Sierra Vista area the third-most affordable housing market in the U.S., outpacing all major markets.

The index ranks the Sierra Vista area as the most affordable housing market in its region and as the only area in Arizona in the top 50 percent of the list, which includes 238 areas nationwide. The next most affordable markets in Arizona are Yuma and Tucson at 125 and 126 respectively.



# PHOENIX REAL ESTATE

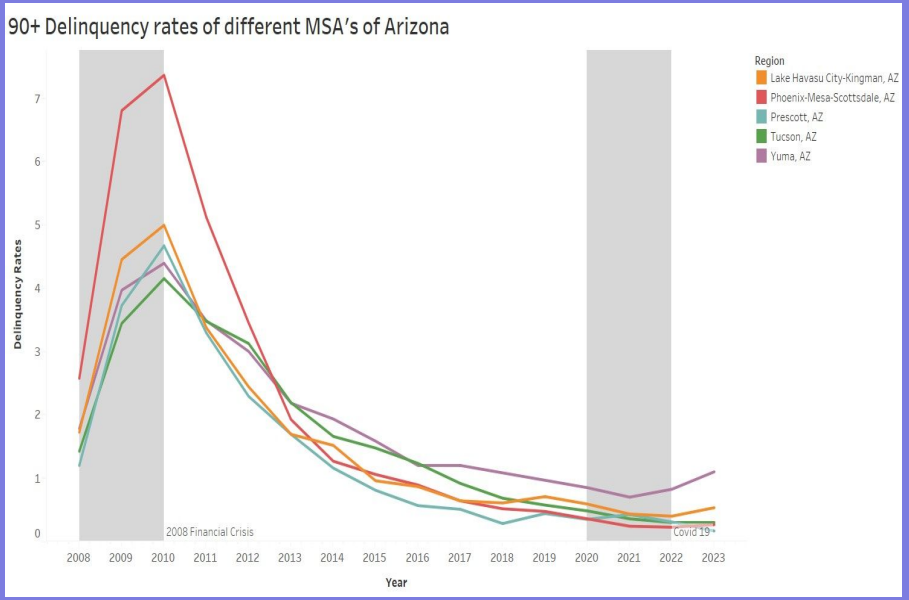
## HOW HAS COVID-19 AFFECTED THE MARKET

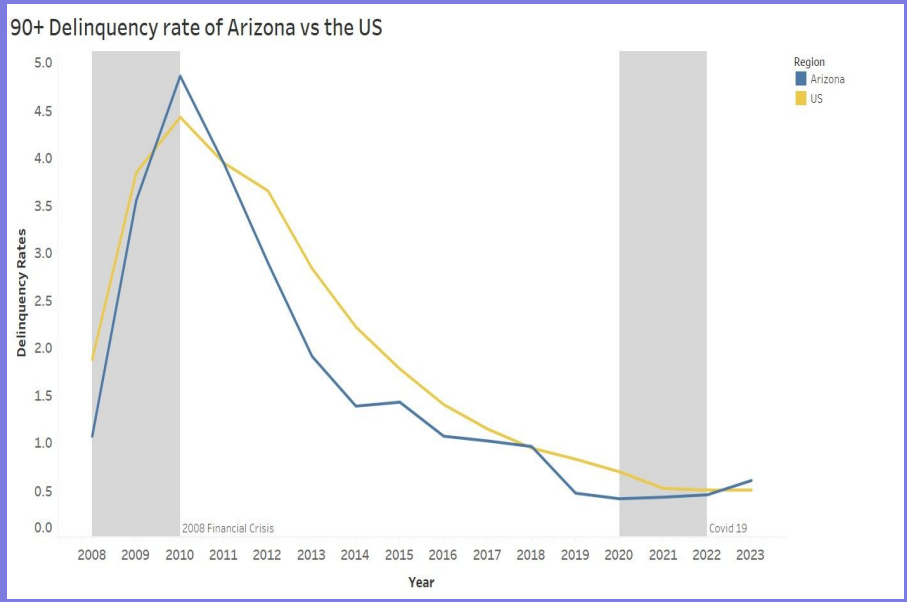
Interest Rates Are At A Record-Breaking Low Due To COVID

COVID-19 Has Pushed the Real Estate Industry To Be A Seller's Market

Tenants Unable to Pay Rent Due To Job Loss Caused by COVID-19

# Mortgage Delinquency Rates

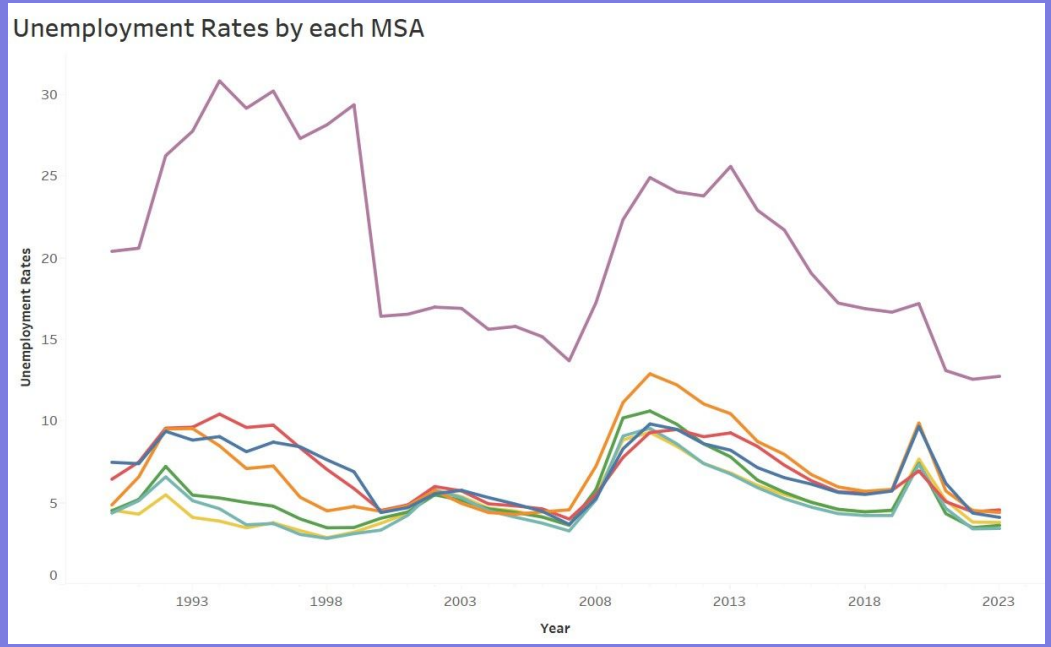


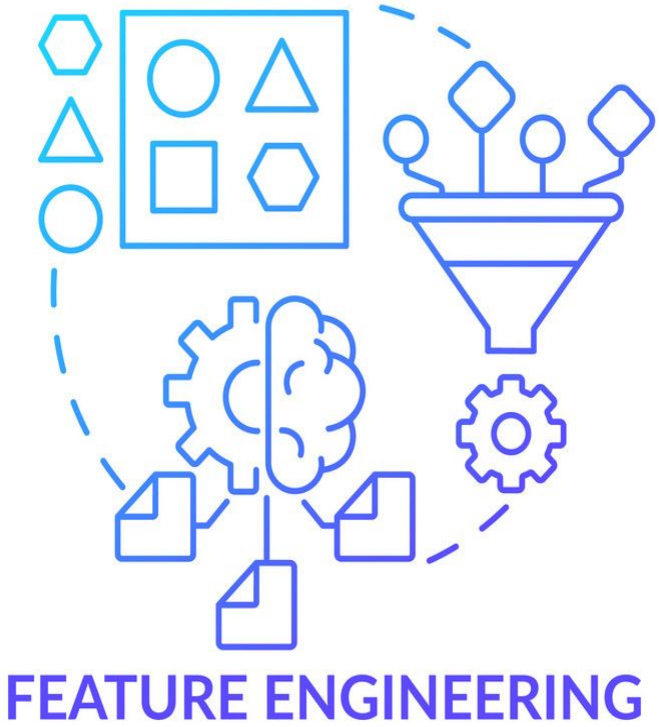

The 90+ mortgage delinquency rate is a measure of early stage delinquencies and can be an early indicator of the mortgage market's overall health. It captures borrowers that have missed one or two payments.

# Unemployment Rates

- Yuma, is an extreme-weather area, an agricultural area.
- In October 2019, 153 metro areas had unemployment rates of less than 3.0 percent and 2 areas had rates of at least 10.0 percent. El Centro, California (21.2 percent), and Yuma, Arizona (16.1 percent), had the highest unemployment rates.
- “We’re really tired of talking about the unemployment rate because it’s not reflective of this community,” said Julie Engel, CEO of the Greater Yuma Economic Development Corp.




**Despite high unemployment, Yuma’s agribusiness continues to thrive**




# FEATURE ENGINEERING

# Feature Engineering

## TREE BASED MODELS

Tree based models like Random Forests can capture nonlinear relationships. Their robustness and ability to handle diverse feature types, make them effective techniques for HPI prediction.

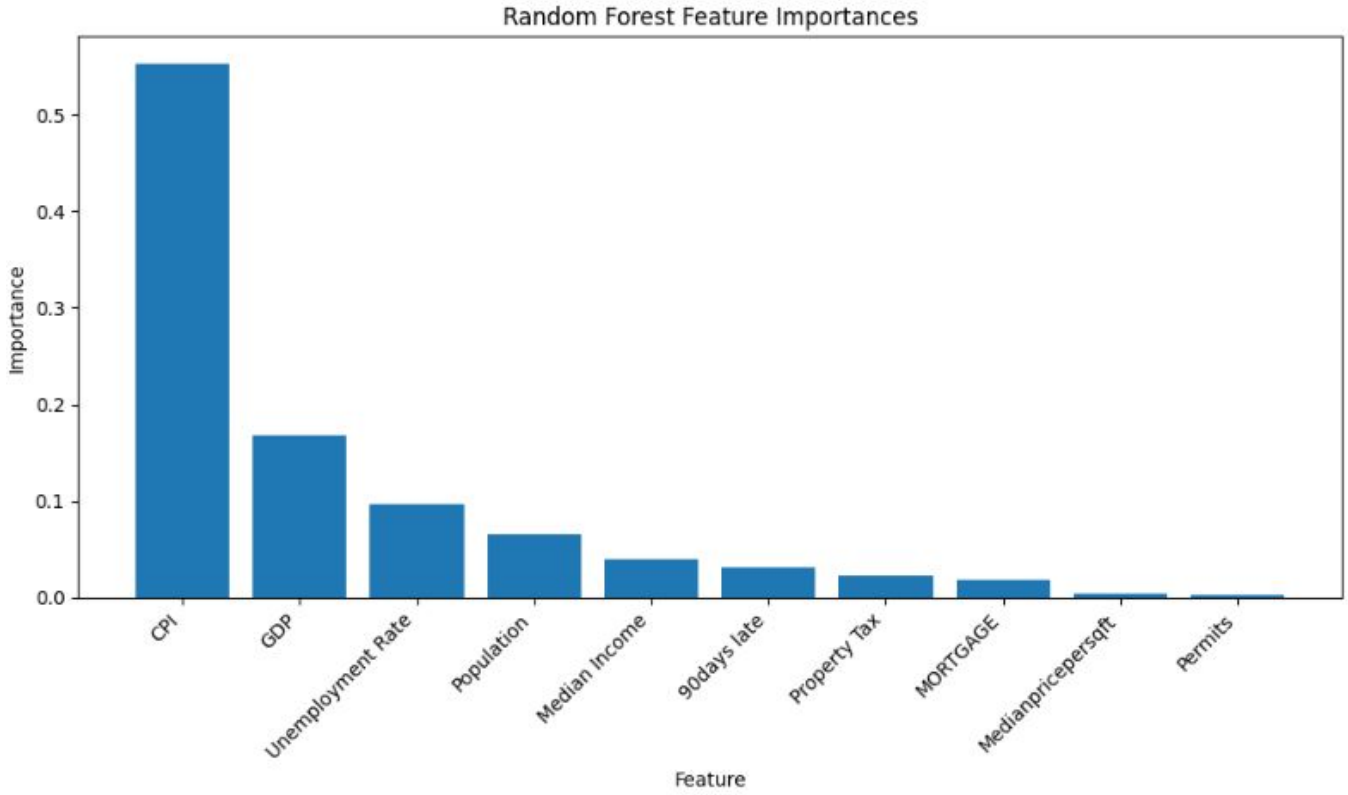
## LASSO REGRESSION

Lasso automatically selects features by driving insignificant variables to 0. In the context of HPI, Lasso helps create a “sparse” model  by emphasizing the most impactful features.

## RFECV

RFECV combines feature ranking and cross validation, ensuring robust performance. Its iterative process adapts to data characteristics, making it suitable for HPI prediction where importance of features may vary.

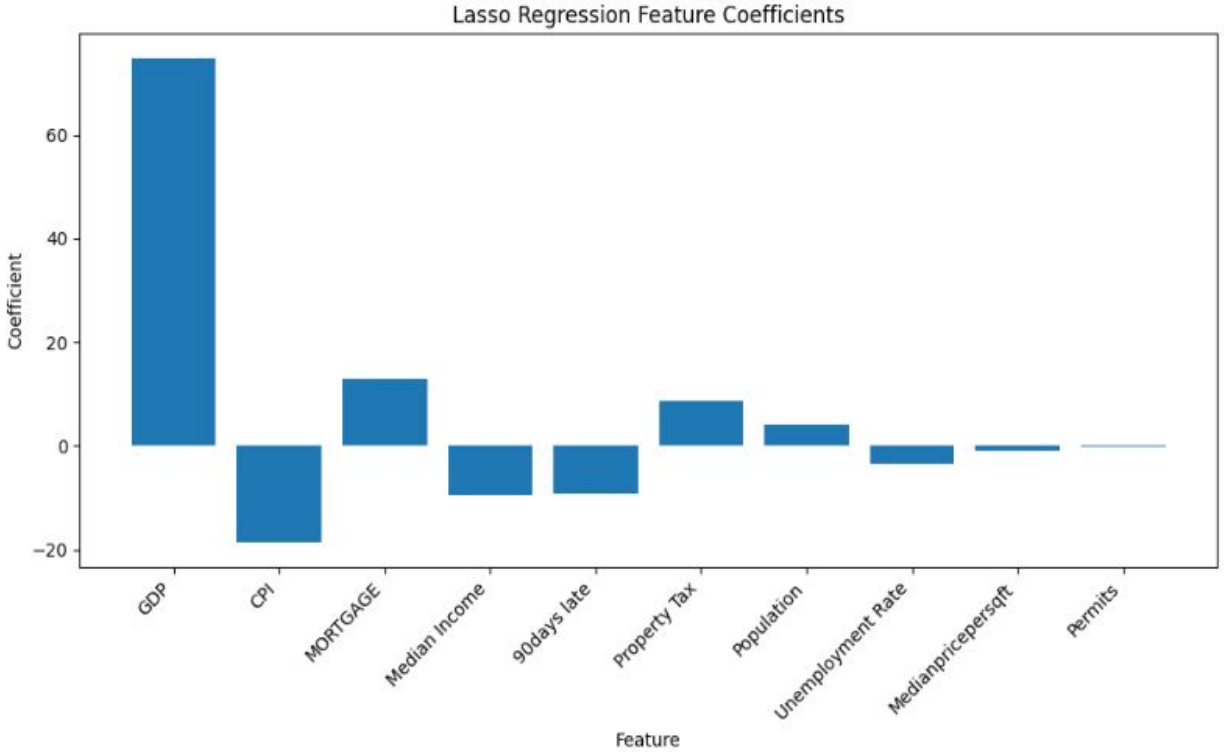
# Random forest results




Mean Squared Error: 14.33717013239702

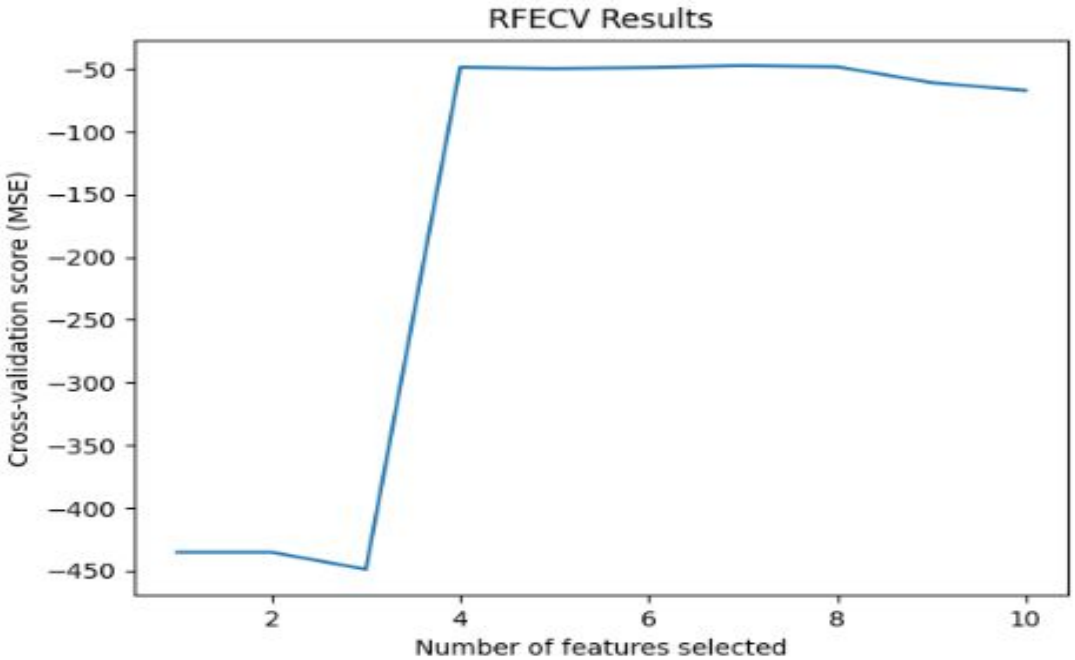
| | Feature | Importance |
|---|---|---|
| 9 | GDP | 0.420719 |
| 4 | Unemployment Rate | 0.190798 |
| 0 | CPI | 0.122777 |
| 5 | 90days late | 0.072691 |
| 3 | Population | 0.067915 |
| 8 | Property Tax | 0.054964 |
| 2 | MORTGAGE | 0.035596 |
| 1 | Median Income | 0.027118 |
| 7 | Permits | 0.003823 |
| 6 | Medianpricepersqft | 0.003599 |

# Lasso regression results




| | Feature | Coefficient |
|---|---|---|
| 9 | GDP | 74.828718 |
| 0 | CPI | -18.657764 |
| 2 | MORTGAGE | 12.870407 |
| 1 | Median Income | -9.516373 |
| 5 | 90days late | -9.266904 |
| 8 | Property Tax | 8.593666 |
| 3 | Population | 4.073547 |
| 4 | Unemployment Rate | -3.478556 |
| 6 | Medianpricepersqft | -0.848440 |
| 7 | Permits | -0.143497 |

# Recursive feature elimination cross validation results

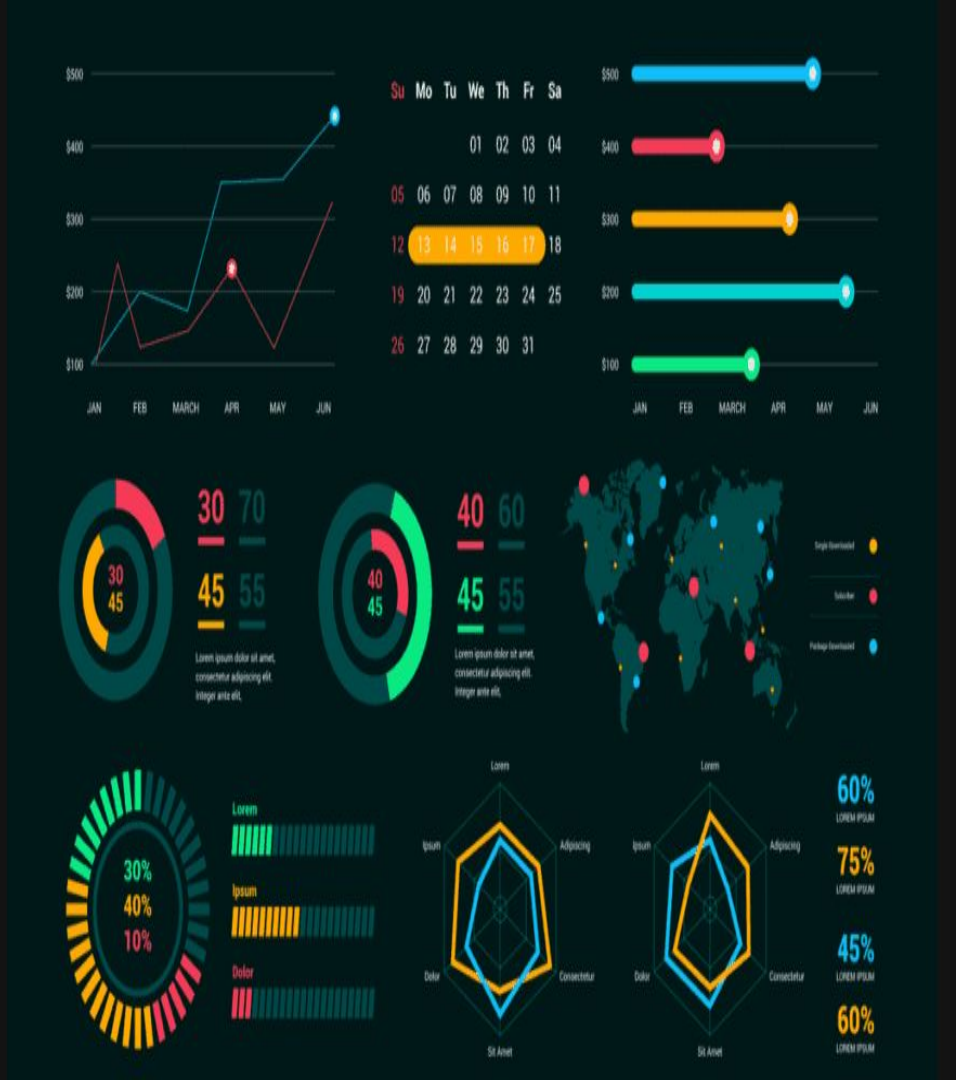



```
Mean Squared Error: 25.693694036394586
Selected Features:
Index(['CPI', 'Median Income', 'Population', 'Unemployment Rate',
       '90days late', 'Property Tax', 'GDP', 'MORTGAGE'],
      dtype='object')
```

# Results

After performing all the feature selection steps, we have narrowed it down to the following:

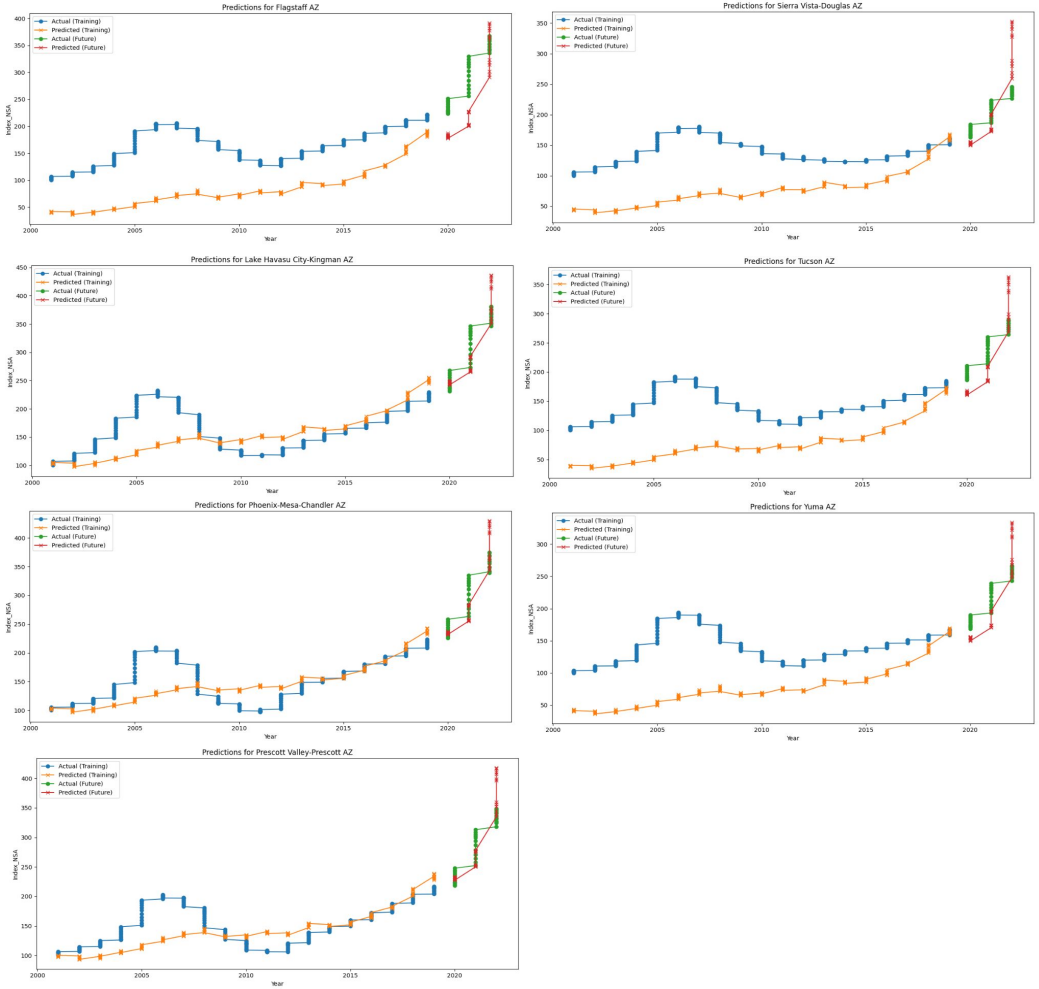
- Consumer Price Index(CPI)
- Gross Domestic Product(GDP)
- Mortgage
- Unemployment Rate
- Median Income
- Population
- Property Tax
- 90-Day Delinquency


$500
$400
$300
$200
$100
JAN FEB MARCH APR MAY JUN
Su Mo Tu We Th Fr Sa
01 02 03 04
05 06 07 08 09 10 11
12 13 14 15 16 17 18
19 20 21 22 23 24 25
26 27 28 29 30 31
30 70
45 55
40 60
45 55
30%
40%
10%
Lorem
Ipsum
Dolor
Adipiscing
Consectetur
Sit Amet
60%
75%
45%
60%

# MODELLING

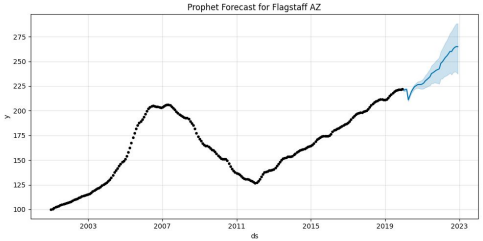
# LSTM

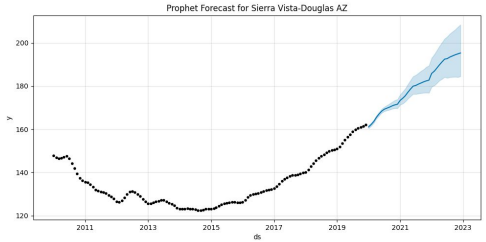
LSTM stands for long short-term memory networks, used in the field of Deep Learning. It is a variety of recurrent neural networks (RNNs) that are capable of learning long-term dependencies, especially in sequence prediction problems.

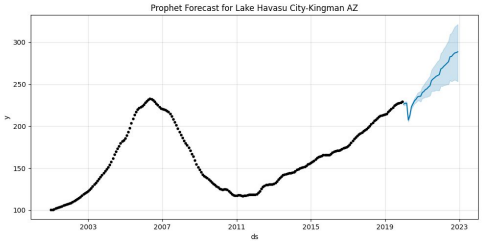


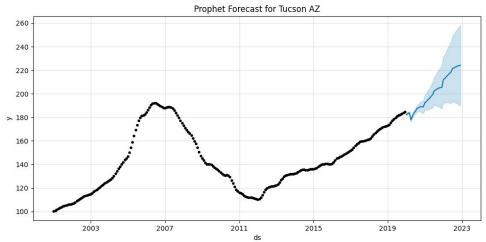
# Prophet Model

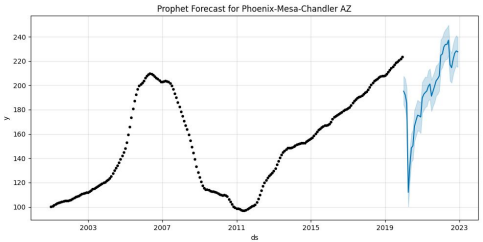
Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

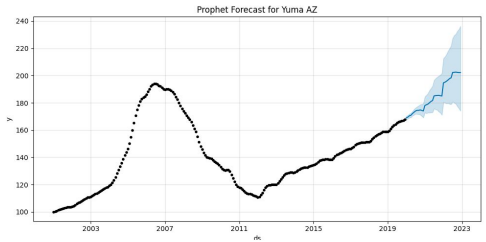


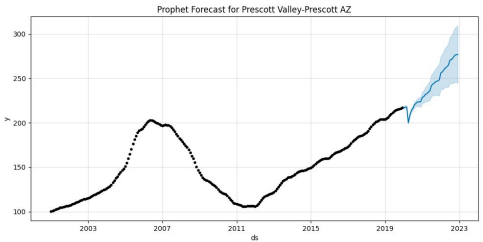



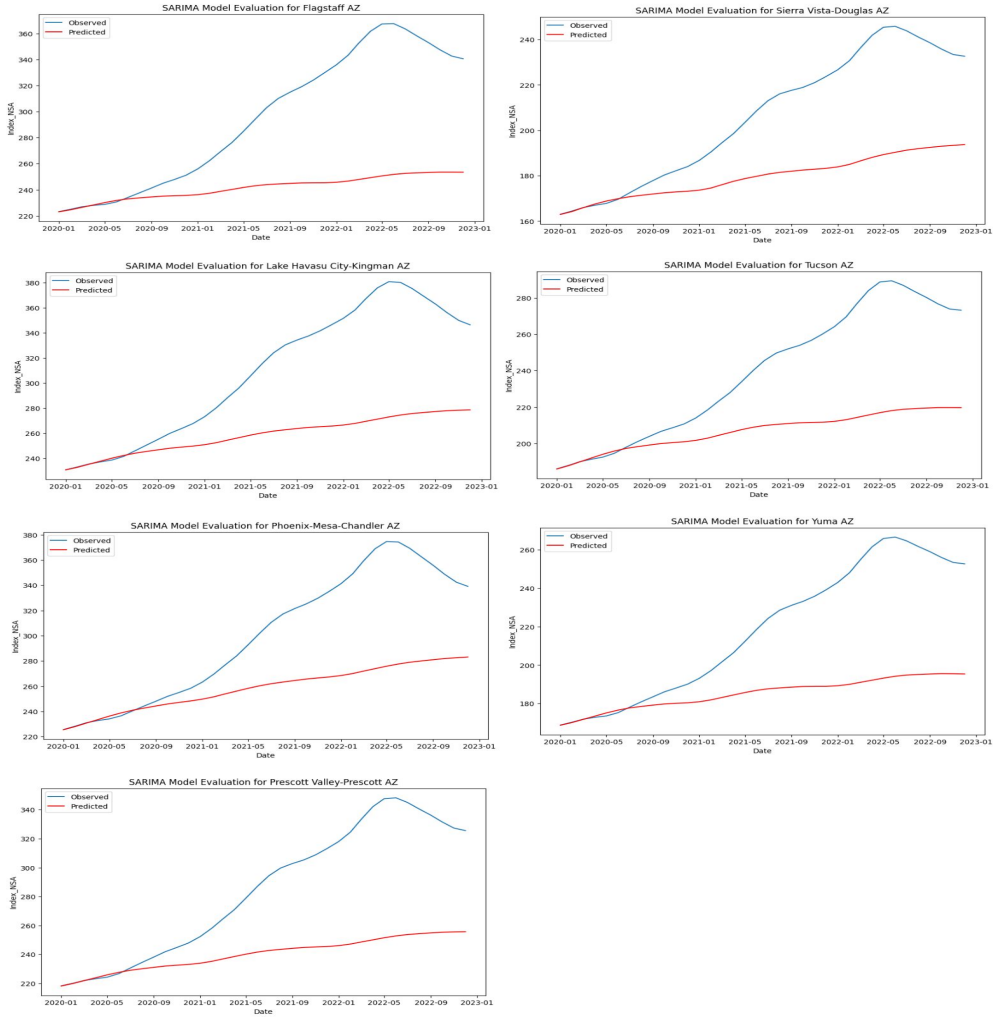



# SARIMA

SARIMA (Seasonal Autoregressive Integrated Moving Average) is a statistical technique used for forecasting time series data, a series of observations recorded at regular intervals over time. SARIMA models are a combination of autoregressive (AR) models, moving average (MA) models, and differencing.



# FINDINGS AND INSIGHTS

# LSTM OUTPERFORMS SARIMA AND PROPHET MODELS

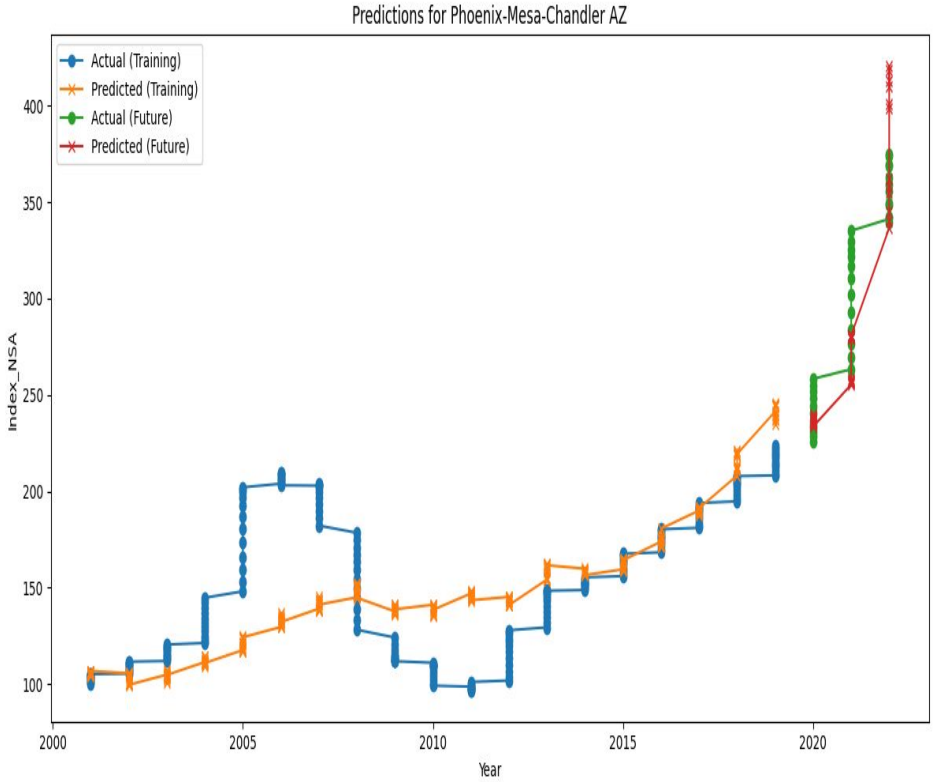
- LSTM, as an RNN, excels in capturing intricate patterns and dependencies within time series data.
- LSTMs adeptly handle non-linear patterns, a challenge for linear models like SARIMA.
- LSTMs automatically extract relevant features from input data, simplifying complex time series analysis.
- LSTM allows end-to-end learning, eliminating the need for explicit feature engineering.
- LSTM's capabilities make it a potent choice, potentially outperforming traditional models and facilitating efficient time series forecasting.

```
LSTM RMSE: 32.43309079486738
LSTM MSE: 1037.4974637792368
LSTM MAE: 25.505099984160967
LSTM R2: 0.6450042903878351
LSTM Accuracy: 0.20912247474747475
```

```
SARIMA RMSE: 42.11929759481705
SARIMA MSE for: 1774.035229880761
SARIMA MAE for: 32.98231501501993
```

```
Prophet RMSE: 88.72826621799615
Prophet MSE: 7872.705226051596
Prophet MAE: 69.87116842932862
```

# LSTM Model Outshines in Phoenix-Mesa-Chandler MSA




Mean Squared Error for Phoenix-Mesa-Chandler AZ: 1119.8180840393281

Root Mean Squared Error for Phoenix-Mesa-Chandler AZ: 33.463683061482165

Mean Absolute Error for Phoenix-Mesa-Chandler AZ: 26.018230944818793

R-squared for Phoenix-Mesa-Chandler AZ: 0.5590790610902517

Training Accuracy: 0.20891043397968606

```
Predictions for Phoenix-Mesa-Chandler AZ:
[[236.98006]
 [235.37195]
 [235.77498]
 [239.73581]
 [237.69592]
 [236.37253]
 [238.815  ]
 [236.9146 ]
 [236.09471]
 [235.10797]
 [234.27365]
 [233.59912]
 [255.31653]
 [255.83235]
 [258.1996 ]
 [257.83926]
 [256.6795 ]
 [257.153  ]
 [279.9168 ]
 [279.36243]
 [279.58652]
 [280.94147]
 [280.64777]
 [280.81473]
```

# Comparing Model Performance: Seasonally Adjusted vs. Non-Seasonally Adjusted HPI Index

- Both Seasonally Adjusted and Non-Seasonally Adjusted HPI Index demonstrate comparable forecasting performance.
- Seasonally Adjusted Index exhibits a slightly better forecasting performance.
- Seasonal adjustments contribute to a more stable and reliable forecasting model.
- The choice between the two may depend on the inherent seasonality within the HPI data.
- Consider the specific requirements of the forecasting task to determine the most suitable index.

```
LSTM RMSE for Non-Sesonally Ajusted Index: 32.7386423429818
LSTM MSE for Non-Sesonally Ajusted Index: 1079.7368298617803
LSTM MAE for Non-Sesonally Ajusted Index:: 26.05525273702047
LSTM R2 for Non-Sesonally Ajusted Index: 0.633082859882456
LSTM Accuracy for Non-Sesonally Ajusted Index:: 0.2063960570454077
```

```
LSTM RMSE for Sesonally Ajusted Index: 32.27892950342557
LSTM MSE for Sesonally Ajusted Index: 1051.9053785081114
LSTM MAE for Sesonally Ajusted Index:: 25.587424601739475
LSTM R2 for Sesonally Ajusted Index: 0.6400744008297661
LSTM Accuracy for Non-Sesonally Ajusted Index:: 0.20787741749592398
```

# CHALLENGES AND OBSTRUCTIONS

# DATA CONSTRAINTS

- **Incomplete MSA-Level Data:** Assumed values at the MSA level based on statewide data due to unavailability, potentially introducing inaccuracies in regions with distinct real estate dynamics.
- **Temporal Inconsistencies:** Diiverse temporal resolutions (annual, monthly, quarterly), posed challenges in achieving uniform analysis during specific events.
- **Historical Data Discrepancies:** Excluded parameters with insufficient historical data (e.g., data from 2016 onwards), impacting the model's comprehensiveness.
- **Limited Coverage of Events:** Relied on historical data, and the absence of information on certain events may limit the model's accuracy during similar future occurrences.
- **Data Imputation Techniques:** Utilization of data imputation techniques to handle missing values or gaps in the dataset may introduce uncertainties.

# External constraints

- **Heterogeneous Market Conditions:** Difficulty capturing diverse regional influences on housing prices.
- **Dynamic Economic Factors:** Challenged by rapid, unpredictable economic shifts during black swan events.
- **Cyclic Market Patterns:** Difficulty accommodating non-linear disruptions in cyclic market trends.
- **Data Privacy and Ethics:** Ensuring ethical handling of sensitive data and addressing potential biases are critical considerations.
- **Lag in Data Reporting**:Delays in official data reporting may hinder timely and accurate predictions during rapidly changing market conditions.
- **Changing Demographics:**Incomplete consideration of demographic shifts like migration, age groups,  limits accurate predictions on housing demand and market dynamics.

# RECOMMENDATIONS AND SUGGESTIONS

1. **Leverage LSTM for Optimal Forecasting:**

The LSTM model excels in capturing complex temporal patterns, making it an effective choice, for enhanced predictive accuracy.It provides valuable insights of housing market dynamics during unforeseen events, contributing to more informed decision-making.

2. **Spatial and Temporal Analysis:**

Integrate spatial analysis to account for location-specific factors that impact house prices and use temporal analysis to capture time-related trends and cyclic patterns in the housing market.

# RECOMMENDATIONS AND SUGGESTIONS

**3. Market Segmentation:**
Explore the potential of creating distinct models for different market segments (e.g., luxury homes, affordable housing) to accommodate the diverse market conditions.

**4. Incorporate External Data Sources:**
Incorporate external data, such as social media sentiment analysis, crime rates, or data on infrastructure development projects, to get a more comprehensive perspective of factors influencing house prices.

# RECOMMENDATIONS AND SUGGESTIONS

**5. Scalability and Deployment:**
Verify that the model is scalable to effectively manage a growing volume of data, and investigate potential options for deploying the model in the real world.

**6. Continuous Model Monitoring and Maintenance:** Establish a framework for continuous monitoring of model performance and incorporate regular updates to adapt to evolving market conditions.

# Thank You!