# Project Scope and Background

The project aimed to develop a predictive model for anticipating future black swan events in the Florida housing market using Freddie Mac Housing Data. Motivated by the inadequacy of traditional models revealed by the COVID-19 pandemic's unprecedented market fluctuations, the project sought to identify early warning signs of potential crises. The focus was on leveraging historical data to analyze the housing market's response to unforeseen events, with a specific emphasis on understanding the broader context of housing crises.

A graph of a rising price

Description automatically generated with medium confidence

# Data Sources

The project utilized diverse and robust data sources to ensure the accuracy and relevance of the analysis. Key sources included:

* Freddie Mac's house price index and mortgage rate data.
* Employment data from Florida's labor statistics.
* Housing units and active listing data from Federal Reserve Economic Data (FRED).
* Appraisal data from the Federal Housing Finance Agency (FHFA).
* Federal funds rate data from FRED.
* Work-from-home and airline travel data to capture pandemic-related trends.
* Social media data from Reddit (Florida-related subreddits) and Google Search results to understand public sentiment.

To ensure data reliability, the project focused on data from credible and official sources like Freddie Mac, FRED, and FHFA. The relevance was maintained by using recent data, specifically tailored to capture the pandemic's impact on the housing market.  
  
Methodology and Models

The methodology centered around integrating sentiment analysis with traditional economic data. Sentiment scores were derived using Hugging Face transformers library from social media and news sources. The integration process involved using sentiment scores as predictors in these models, alongside traditional economic indicators, to predict housing market trends. Machine Learning models used -

* ARIMA: Used for its strength in analyzing time series data, focusing on historical trends.
* Random Forest: Chosen for its ability to handle complex datasets and reduce overfitting.
* XGBoost: Selected for its efficiency with large datasets and versatility.

In addition to these models, Large Language Models (LLMs) such as OpenAI’s ChatGPT 4, ChatGPT 3.5 and Google PaLM were explored using advanced prompt engineering methods.

# Key Results and Interpretations

The project's significant findings include:

* Model Performance: The XGBoost model, incorporating public sentiment, showed a 14.83% reduction in error rate compared to the base model, indicating the effectiveness of integrating sentiment analysis in predictive models.
* Limitations of LLMs: Current Large Language Models (LLMs), as of December 2023, were found to be promising but not sufficiently reliable for the analysis, underscoring the need for further development and refinement in this area.
* Impact of Public Sentiment: The analysis revealed a direct correlation between public sentiment and changes in the house price index. This emphasizes the importance of including public perception and media coverage in predictive models for more accurate forecasts.

These findings highlight the evolving nature of real estate market analysis during unpredictable events like the pandemic, suggesting a shift towards more dynamic and comprehensive analytical approaches that encompass traditional data, AI, and public sentiment.

### LLM Model –

A graph with a line

Description automatically generated

### ARIMA model –

A graph with a line going up

Description automatically generated

### Random Forest Model –

A graph with a line

Description automatically generated

### XGBoost Model –

A graph with a line

Description automatically generated

# Challenges Encountered

* Exclusion of Post-2020 Data: This limited the ability to capture daily shifts and micro-market trends after the onset of COVID-19.
* Quality Data Acquisition Costs: High costs of accessing quality data sources were a constraint. We prioritized the most impactful data sources within budget limitations.
* API Restrictions: Challenges in obtaining sufficient social media data due to API limitations were addressed by diversifying data sources and maximizing the use of available data.

# Future Implications

The findings from our housing price analysis in Florida, particularly with the integration of sentiment analysis and advanced predictive models, had significant implications for the future:

* Enhanced Predictive Models: The success of integrating sentiment analysis with traditional data in predicting housing market trends indicates a shift towards more holistic and nuanced predictive models in real estate analytics. This approach can be replicated in other markets and for other types of economic forecasting.
* Broader Application of AI in Real Estate: The exploration of LLMs, despite their current limitations, opens the door for further research and development of AI tools specifically tailored to real estate market analysis. This could lead to more accurate and timely predictions, benefiting investors, policymakers, and consumers.
* Sentiment Analysis as a Standard Tool: The project underscores the value of sentiment analysis in economic forecasting. As social media and online platforms continue to influence public opinion and market trends, incorporating sentiment data could become a standard practice in various fields of analytics.