

George Mason University

CS 504

Spring 2025

**Indications of Recession**

**Report**

**Project 2 Team:**

**Iris Duan,  
Roger Graham,  
Brian Haggerty,  
An Huynh**

## Table of Contents

Table of Contents .....	2
Table of Figures .....	4
Table of Tables .....	4
Table of Appendices .....	4
Abstract .....	5
1 Introduction .....	6
1.1 Background and Rationale .....	6
1.2 Research .....	16
1.3 Project Objectives .....	16
1.4 Problem Space .....	16
1.5 User Stories .....	17
1.6 Solution Space: .....	18
1.7 Product Vision: .....	19
1.7.1 Scenario #1 .....	19
1.7.2 Scenario #2 .....	19
1.8 Glossary of Acronyms and Terms (GOAT): .....	20
2 Data Acquisition .....	21
2.1 Overview: .....	21
2.2 Field Descriptions: .....	22
2.2.1 Common Field: .....	22
2.2.2 Unique Fields by Data Set: .....	23
2.3 Data Context: .....	24
2.4 Data Conditioning .....	26
2.5 Data Risks: .....	29
2.6 Data Quality Assessment: .....	29
2.7 Other Data Sources .....	30
3 Analytics and Algorithms .....	31
3.1 Methodology .....	31
3.2 Tools and Infrastructure .....	32
3.3 Data Preparation for Modeling .....	33
4 Visualizations .....	34
4.1 Visualization Tools and Implementation Details .....	34

4.2	Visualization Applications and Descriptions .....	35
4.3	Visualization Risks and Mitigation.....	42
5	Findings.....	43
5.1	Logistic Regression and Hypothesis Testing .....	44
5.2	Ordinary Least Square (OLS) .....	45
5.3	Random Forest Classifier.....	47
5.4	XGBoost Classifier .....	49
5.4	Interpretation of Economic Shifts (Table 3) .....	51
6	Summary .....	52
7	Future Work .....	53
	Appendix A: Code References.....	54
	Appendix B: Risk Matrix.....	55
	Appendix C: Agile Development.....	58
	Appendix D: Course Schedule.....	59
	Appendix E: Glossary of Acronyms and Terms (GOAT) .....	60
	References .....	62

## Table of Figures

Figure 1: Federal funds effective rate (1955-2025).....	7
Figure 2: Real Gross Domestic Product (1947–2024).....	8
Figure 3: Housing Price Index (1975-2024).....	9
Figure 4: Housing Affordability Index (2024–2025).....	10
Figure 5: Monthly supply of new homes (1959–2024).....	11
Figure 6: Housing starts (1960-2025).....	12
Figure 7: NBER recession indicators (1900-2024).....	13
Figure 8: Yield Curve Spread (10-Year vs. 3-Month), (2006-2012).....	14
Figure 9: Historical yield curve spread and U.S. recessions (1982–2025).....	14
Figure 10: Real Personal Income (1960–2025).....	15
Figure 11: Yield Curve Spread (T10Y3M), Recession vs. Non-Recession Periods .....	35
Figure 12: GDP Pre-2008 and Post-2008 .....	36
Figure 13: House Price Index Pre-2008 and Post-2008.....	37
Figure 14: Income Pre-2008 vs. Post-2008.....	38
Figure 15: Home supply Pre-2008 vs. Post-2008 .....	39
Figure 16: Newly Owned Homes Pre-2008 vs. Post-2008.....	40
Figure 17: Consumption Price Index (CPI) Pre-2008 vs. Post-2008.....	41
Figure 18: Logistic Regression Output (Recession Indicator vs. Treasury Spread).....	44
Figure 19: Ordinary Least Square Output (House Price Index vs. Treasury Spread).....	45
Figure 20: Random Forest Model Performance Metrics .....	47
Figure 21: Random Forest Model Features Importance .....	48
Figure 22: XGBoost Model Performance Metrics.....	49
Figure 23: XGBoost Model Features Importance.....	50

## Table of Tables

Table 1: Summary of Indicators Used .....	25
Table 2: Summary Data .....	26
Table 3: Indicators Pre-2008 vs Post-2008.....	51

## Table of Appendices

Appendix B: Risk Matrix.....	55
Appendix C: Agile Development.....	58
Appendix D: Course Schedule.....	59
Appendix E: Glossary of Acronyms and Terms (GOAT) .....	60

## Abstract

This study examines early warning signals of the 2008 financial crisis by analyzing macroeconomic and housing market indicators, particularly the yield curve and housing affordability, using data from 2006 to 2012. After preprocessing and lagging variables to reflect real-world forecasting, four models were tested: OLS regression, logistic regression, Random Forest, and XGBoost. Results showed that the yield curve spread, federal funds rate, and real income significantly predicted housing prices, while yield curve inversion was linked to recession probability. Random Forest, selected for its interpretability, achieved perfect classification performance along with XGBoost. The study underscores the value of robust preprocessing, model selection, and further validation across other downturns like 2020.

# 1 Introduction

## 1.1 Background and Rationale

An inverted yield curve has historically served as a reliable signal of future economic downturns, and as of early 2025, the curve remains inverted. At the same time, housing affordability has declined significantly across the United States. This dual pressure on the economy raises questions about which indicators provide the earliest and most reliable warnings of recession.

This project analyzes key macroeconomic and housing indicators from 2006 to 2012 to evaluate their behavior before, during, and after the 2008 financial crisis. The objective is to identify patterns in historical data that may serve as early signals of economic downturns or recovery periods. Rather than forecasting future recessions, this analysis examines how established indicators—such as the yield curve spread, housing affordability, and real GDP—reflected and responded to the dynamics of the 2008 crisis.

Each of the following indicators is discussed to provide essential context for its role in the broader economic system. The shaded regions of the figures below represent periods of recession. The red boxes represent the focus period of this study, 2006 to 2012.

## Federal Funds Rate:

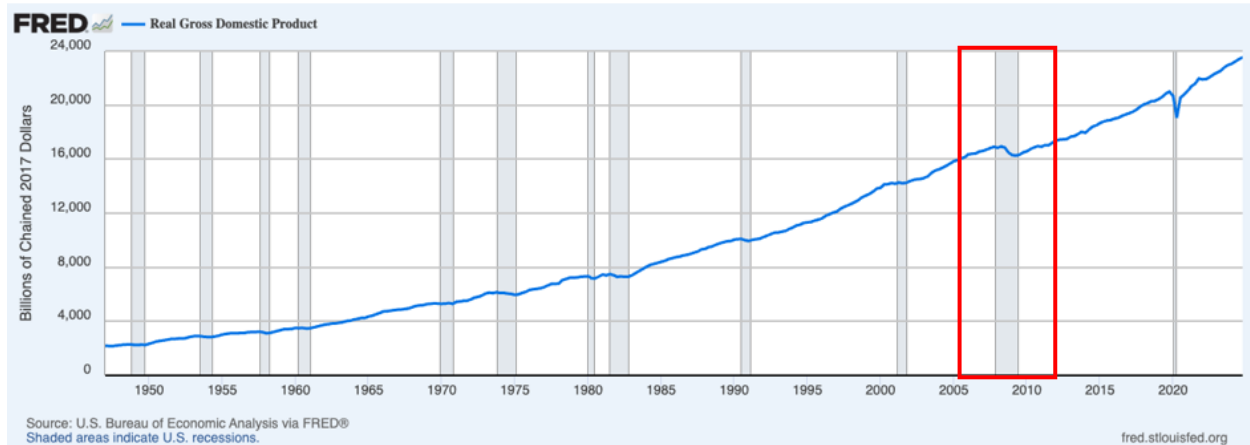


*Figure 1: Federal funds effective rate (1955-2025).*

*Source: Adapted from Federal Reserve Bank of St. Louis, by Board of Governors of the Federal Reserve System, 2024, <https://fred.stlouisfed.org/series/FEDFUNDS>. Public domain.*

Figure 1 shows the Federal Funds effective rate since 1955. The federal funds rate is the overnight interest rate at which banks lend reserves to one another. It is the primary monetary policy tool used by the Federal Reserve to influence liquidity and credit conditions. Lowering this rate reduces borrowing costs for banks, which typically translates into lower interest rates for consumers and businesses. During the 2008 financial crisis, the Federal Reserve cut this rate to near-zero levels to stimulate economic activity and avoid deflation. The trajectory of this rate is key to understanding the policy response to recessionary pressure (Board of Governors of the Federal Reserve System, 2024).

## Real Gross Domestic Product (GDP):



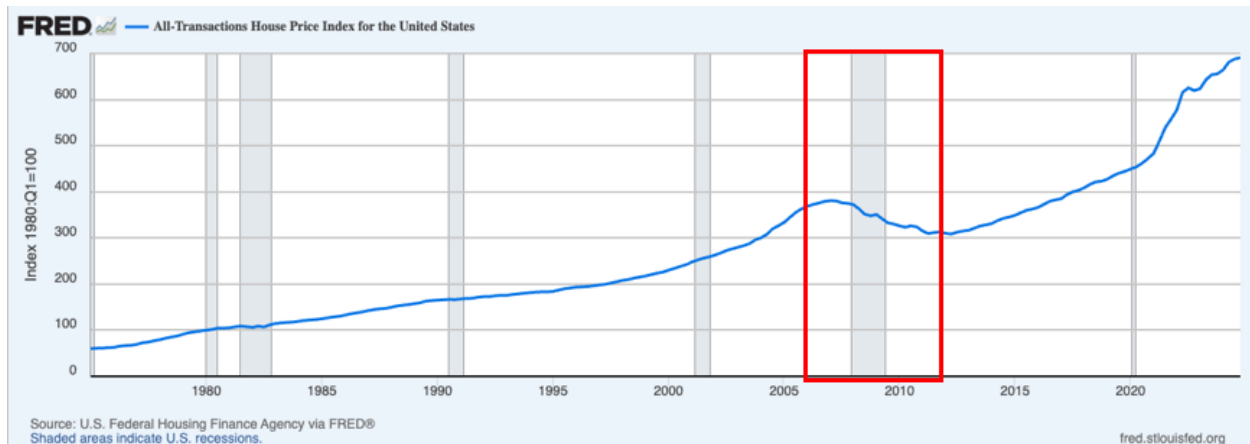
*Figure 2: Real Gross Domestic Product (1947–2024).*

*Source: Adapted from Federal Reserve Bank of St. Louis, by Bureau of Economic Analysis, 2024, <https://fred.stlouisfed.org/series/GDPC1>. Public domain.*

Figure 2 shows the Real Gross Domestic Product (GDP) since 1947. Real GDP measures the inflation-adjusted value of all U.S. goods and services. A sharp decline in GDP signals economic contraction, often accompanied by reduced output, lower demand, and rising unemployment. The steep decline in GDP during the 2008 recession reflects the severity of the downturn (Bureau of Economic Analysis, 2024).



## Housing Price Index (HPI):

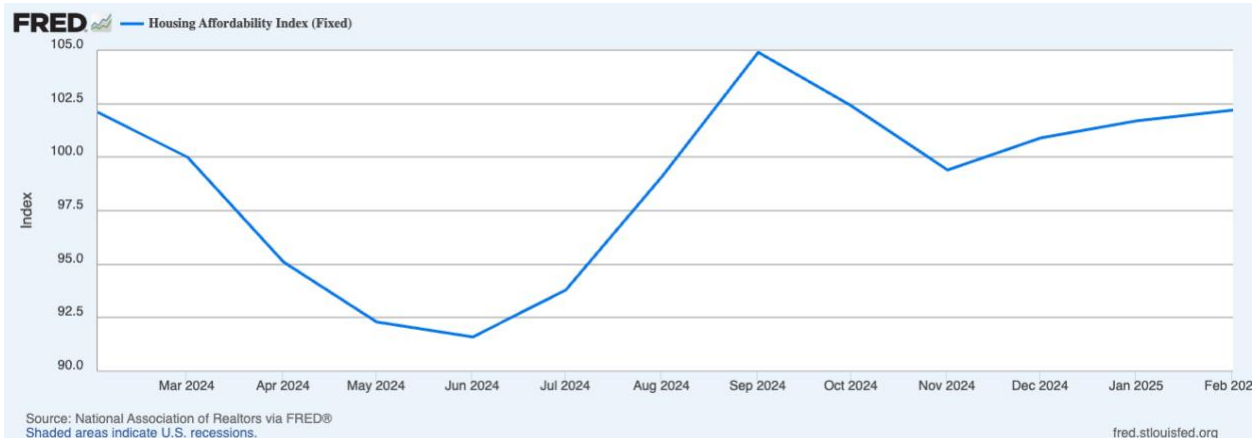


*Figure 3: Housing Price Index (1975-2024).*

*Source: Adapted from Federal Reserve Bank of St. Louis, by Federal Housing Finance Agency, 2024, <https://fred.stlouisfed.org/series/USSTHPI>. Public domain.*

Figure 3 shows the Housing Price Index (HPI) since 1975. The HPI, produced by the Federal Housing Finance Agency, tracks changes in single-family home prices over time. This project uses the Purchase-Only Index, which excludes refinancing activity and focuses solely on transaction-based price appreciation. The index is standardized to a base year and presented without units. For instance, a value of 150 with a base year of 100 implies a 50% increase in home prices since that year. Interpreting the rate of change, rather than absolute values, is central to understanding market volatility (Federal Housing Finance Agency, 2024).

## Housing Affordability Index (HAI):



*Figure 4: Housing Affordability Index (2024–2025).*

*Source: Adapted from Federal Reserve Bank of St. Louis, by National Association of Realtors, 2024, <https://fred.stlouisfed.org/series/FIXHAI>. Public domain.*

Figure 4 shows the Housing Affordability Index (HAI), but only since 2024. The HAI evaluates whether a median-income household can afford a mortgage on a median-priced home under typical lending conditions. A value of 100 indicates perfect affordability; values above 100 suggest higher affordability, while those below 100 indicate financial strain. Despite falling home prices during the 2008 recession, affordability dropped due to rising unemployment and tighter lending standards. As a result, this index captures both pricing and access-to-credit dynamics (National Association of Realtors, 2024).

However, the HAI was not used in this analysis due to limited historical data availability. The series only spans from 2024 onward and does not cover the core analysis window of 2006–2012. To approximate affordability during that period, this project used a combination of the Housing Price Index (to reflect cost trends), real income (to estimate household purchasing power), and housing supply (to evaluate market pressure). These variables were available across the full study window and served as an effective proxy for affordability conditions in the absence of HAI data.

While data substitution is certainly not unprecedented, this combination of data sets was selected by the project team based on data availability and economic factors considered in the HAI.

## Monthly Supply of New Homes:



*Figure 5: Monthly supply of new homes (1959–2024).*

*Source: Adapted from Federal Reserve Bank of St. Louis, by U.S. Census Bureau, 2024, <https://fred.stlouisfed.org/series/MSACSR>. Public domain.*

The monthly supply of new homes indicator measures how many months it would take to deplete the current inventory of new homes at the existing sales pace. Figure 5 shows the monthly supply of new homes since 1959. A six-month supply represents market equilibrium. Values above six months suggest excess supply and weakening demand (a buyer's market), while values below six months reflect a tight supply and elevated demand (a seller's market). The sharp increase in supply during the 2008 crisis highlights declining buyer activity and market saturation (U.S. Census Bureau, 2024a).

## Housing Starts:

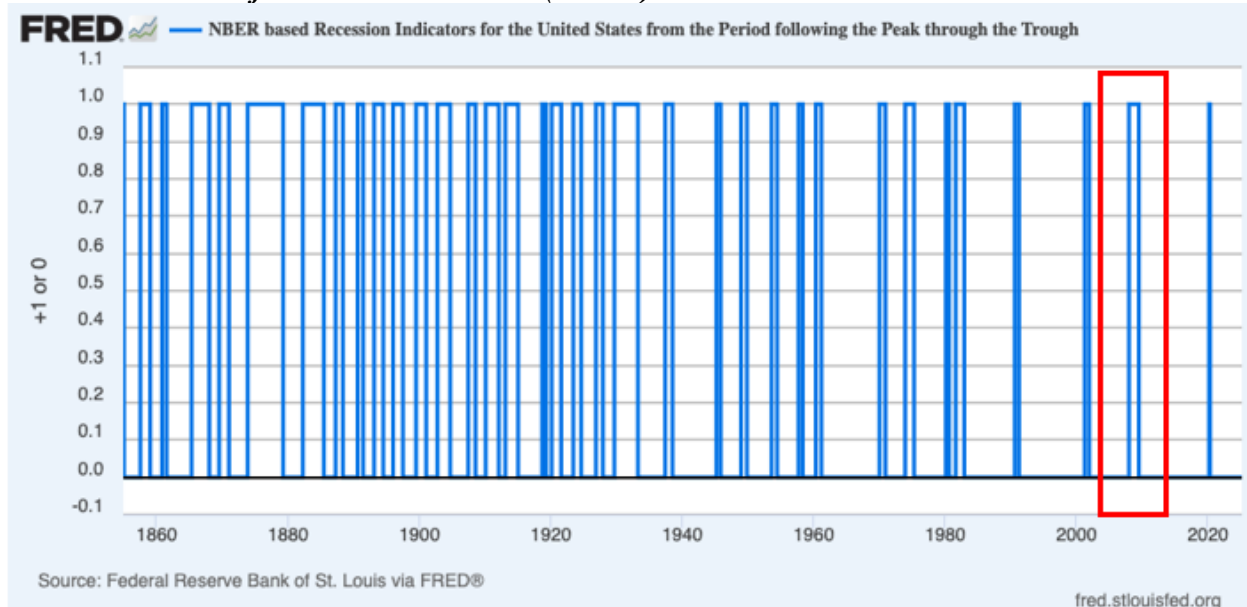


Figure 6: Housing starts (1960-2025).

Source: Adapted from Federal Reserve Bank of St. Louis, by U.S. Census Bureau, 2024, <https://fred.stlouisfed.org/series/HOUST>. Public domain.

Housing starts, as shown in Figure 6, quantifies new privately owned residential construction projects initiated each month, reported as a seasonally adjusted annual rate (SAAR). This includes both single-family and multi-family dwellings. A decline in housing starts typically signals reduced builder confidence and pessimism about future demand. The steep drop in starts between 2007 and 2009 was an early signal of housing market instability (U.S. Census Bureau, 2024b).

### ***National Bureau of Economic Research (NBER) Recession Indicators:***



*Figure 7: NBER recession indicators (1900-2024).*

*Source: Adapted from Federal Reserve Bank of St. Louis, by National Bureau of Economic Research, 2024, <https://fred.stlouisfed.org/series/USREC>. Public domain*

The National Bureau of Economic Research (NBER) provides official designations of recession periods, as shown in Figure 7. The NBER identifies months and quarters of turning points, where values of one represent recessionary periods and 0s represent expansion periods. For this time series, the recession begins the first day of the period following a peak and ends on the last date of the period of a trough. This project uses the quarterly NBER recession indicator series to anchor comparisons across macroeconomic and housing variables.

## Yield Curve Spread (10-Year vs. 3-Month):

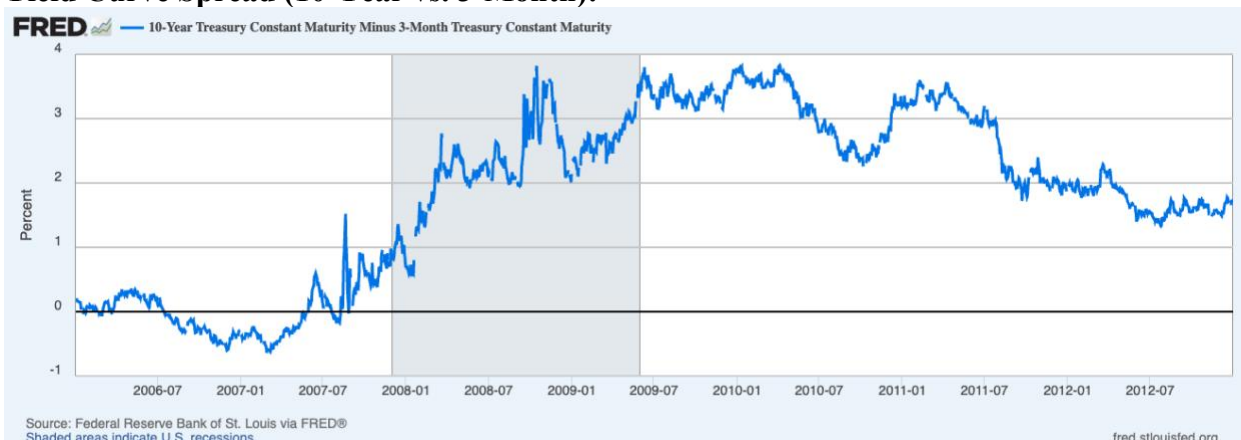


Figure 8: Yield Curve Spread (10-Year vs. 3-Month), (2006-2012).

Source: Adapted from Federal Reserve Bank of St. Louis, by Board of Governors of the Federal Reserve System, 2024. <https://fred.stlouisfed.org/series/T10Y3M>. Public domain.

The yield curve plots interest rates for Treasury securities with different maturities. The difference between the 10-year and 3-month Treasury yields—commonly referred to as the yield curve spread—is a widely accepted indicator of economic outlook. A negative spread (inversion) indicates that short-term interest rates exceed long-term rates, reflecting market expectations of an economic slowdown. Figure 8 shows this yield curve spread for the period of interest of 2006 to 2012.

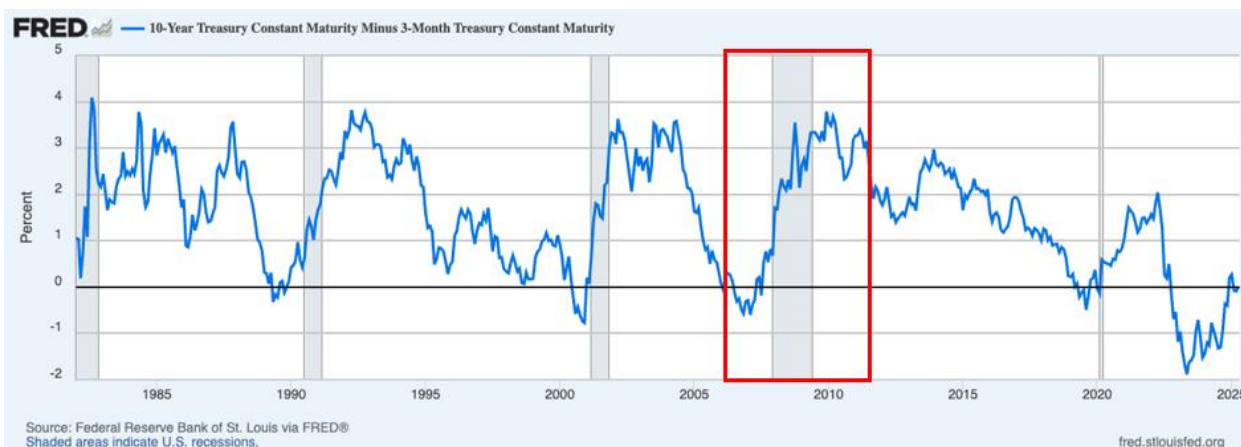
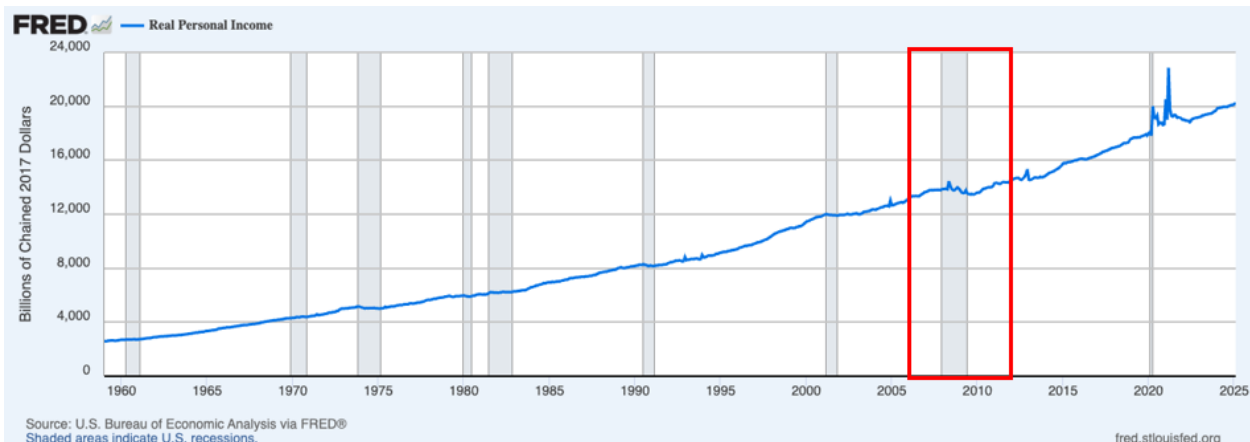


Figure 9: Historical yield curve spread and U.S. recessions (1982–2025).

Source: Adapted from Federal Reserve Bank of St. Louis, by Board of Governors of the Federal Reserve System, 2024. <https://fred.stlouisfed.org/series/T10Y3M>. Public domain.

Figure 9 shows the yield curve since 1982. This spread has inverted ahead of every U.S. recession since 1950, including the 2008 financial crisis, where the inversion preceded GDP contraction by several quarters. As such, it remains a critical tool for monitoring macroeconomic risk (Federal Reserve Bank of St. Louis, 2024).

## Real Personal Income:



*Figure 10: Real Personal Income (1960–2025)*

*Source: Adapted from Federal Reserve Bank of St. Louis, Real Personal Income (RPI), 2024.  
<https://fred.stlouisfed.org/series/RPI>. Public domain.*

Real Personal Income (RPI) since 1960 is shown in Figure 10. It reflects the total income earned by individuals in the United States, adjusted for inflation. Unlike nominal income, RPI accounts for changes in purchasing power over time and is expressed in chained 2012 dollars to control for inflation. Chained dollars refers to a way to adjust for inflation that also accounts for changes in what people buy over time. This measure is widely used to evaluate the real economic well-being of households.

In this project, RPI served as a substitute for the income component of the Housing Affordability Index (HAI), which lacked historical coverage for the 2006–2012 period. When analyzed alongside the Housing Price Index and housing supply, RPI enabled a practical approximation of affordability trends, revealing how economic downturns influenced household capacity to participate in the housing market.

## 1.2 Research

Some of the research team came to this project with a good understanding of the economic factors involved, which provided an excellent starting point for initial research. They were able to propose potential data sets, as well as where to find them. This led to iterative discussions about which ones were appropriate for our purposes, as well as additional research to better understand what each economic factor was and how it is determined.

According to Fidelity, the yield curve is the "graphical representation of the yields available for bonds of equal credit quality and different maturity dates." The slope of the yield curve is an indicator of the direction of interest rates and economic status.

The normal yield curve with an upward slope signifies economic growth without significant change in inflation rate. An inverted yield curve with downward slope is the predictor of economic recession. A flat yield curve can be a sign of economic uncertainty or transition.

Housing market stress refers to a condition where factors like high mortgage rates, high home prices, and shortage of inventory impact affordability and ability to buy a house. This condition can be caused or worsened by the economic downturn.

## 1.3 Project Objectives

To understand the correlation between yield curves and the housing market stress, along with other economic indicators such as inflation, interest rates, and GDP.

## 1.4 Problem Space

The problem domain for this project revolves around the impact of the inverted yield curve on the housing market. Specifically, it focuses on how the economic downturn affects housing affordability.

The problem spaces include analyzing the economic factors that are related in the U.S. recession of 2008, and the meaning of the inverted yield curve and how it correlates to the housing market stress.



## 1.5 User Stories

The following user stories outline the objectives and expected outcomes for each project sprint. Each story captures the perspective of a realistic user, providing clear guidance and transparency for stakeholders and project participants.

### **Primary User Story (Problem Statement):**

- As a Client Analyst, I want to access accurate and reliable historical data on housing markets and yield curves from the 2008 recession period to inform actionable business and investment decisions.

### **Sprint 1 User Stories:**

- As a Business Stakeholder, I want the project team to deliver a comprehensive project schedule, clear risk mitigation strategies, and multiple scenario analyses so that I can confidently oversee our analysis of economic indicators relevant to recession events.

### **Sprint 2 User Stories:**

- As a Data Analyst, I want to acquire, standardize, and validate data spanning the pre-, during-, and post-recession periods of 2006–2012, ensuring data quality and consistency for meaningful recession analysis.
- As a Database Developer, I want to identify and define entities, attributes, and relationships within our datasets to create an efficient and functional relational database that enables Client Analysts to perform effective data queries and analysis.

### **Sprint 3 User Stories:**

- As a Data Scientist, I want to define and implement algorithms and analytic methods on the prepared dataset to identify statistically significant trends related to the recession periods.
- As a Risk Analyst, I want to promptly identify and address potential risks or challenges that arise during initial analytics so the project can stay aligned with its intended scope.
- As a Business Stakeholder, I want ongoing updates and interim analytics results communicated clearly, so I can assess whether project adjustments or scope revisions are necessary.

### **Sprint 4 User Stories:**

- As a Data Visualization Specialist, I want to create clear, effective visualizations that accurately reflect our analytic findings, adhering to industry best practices for clarity and interpretability.
- As a Client Analyst, I want to review and validate visualization concepts and deliverables to ensure they effectively communicate insights needed for strategic decision making
- As a Business Stakeholder, I want visualizations to be reviewed and approved collaboratively so the final deliverables match project goals and stakeholder expectations.

### **Sprint 5 User Stories:**

- As a Data Analyst, I want to integrate all analytics, database components, and visualizations into a cohesive final product to ensure project completeness and consistency.
- As a Client Analyst, I want the final integrated deliverable clearly aligned with our original problem statement and business objectives to effectively support strategic decisions.
- As a Business Stakeholder, I want a comprehensive final presentation and detailed project report submitted, clearly communicating the project's findings, methodology, and recommendations.

### **1.6 Solution Space:**

This analysis delivers value to policymakers and mortgage lenders by providing insights into the yield curve and housing market conditions surrounding the 2008 recession.

In addition to these primary indicators, this analysis also considers other key economic factors such as Real Gross Domestic Product (GDP), inflation, and interest rates. Users derive value from this analysis by obtaining an understanding of how the inverted yield curve may signal an economic downturn and how the housing market may act as early signals of economic health.

Understanding these relationships and their impacts enables policymakers, financial institutions, and related interest groups to develop informed strategies, policies, and decisions.

## 1.7 Product Vision:

There are 2 different visions for how the deliverables of this project can be used. The first is that monetary policy makers can use this to provide insights to inform their decisions. The second is for Mortgage Lenders to help them understand the indicators for their own decision making.

### 1.7.1 Scenario #1

- **For:** Monetary policy makers who need a big picture of economic status and the factors that impact the economy.
- **Who:** Require comprehensive and accurate analysis to make better decisions regarding monetary policy.
- **The:** Correlation analysis to explore the relationship between yield curve, housing market, and other economic indicators.
- **Is a:** Useful method that provides insights into the intricate relationship between key economic variables.
- **That:** Uncover the relationship between these factors can provide more insights for investors and policy makers.
- **Unlike:** Individual datasets that stand alone.
- **Our product:** Uses statistical analysis to deep dive into multiple datasets and discover the relationship between them.
- **Caveats:** Finding data from valid and reliable sources. Successful analysis depends on combining data from various sources to have a complete data set, and database setup within an associated GitHub account for the team to consolidate efforts.

### 1.7.2 Scenario #2

- **For:** Mortgage lenders analytical staff who need to understand where the market is going so they can adjust their product mix accordingly.
- **Who:** Require good information to enable decision making to potentially make asks of their secondary mortgage partner on product that changes weekly with the mortgage market.
- **The:** Data-driven technology solution that uses leading economic indicators.
- **Is a:** Visual system that is easy to understand and be used to make decisions.
- **That:** Reduces the amount of work a mortgage lender team does to understand affordability and where the market is going.
- **Unlike:** Going to different web sites to gather information and figure out how and what to do.
- **Our product:** Combines all the economic indicators and provides an easy-to-understand interpretation of what they may mean.
- **Caveats:** Performance depends on data accuracy and availability. Not all economic indices may be useful.

### 1.8 Glossary of Acronyms and Terms (GOAT):

A complete Glossary of Acronyms and Terms can be found in Appendix E. It is important to note that not all the abbreviations or terms encountered in this research or included in this project are acronyms. Many are simply industry nomenclatures used by organizations such as the Fed to concisely represent longer titles.

## 2 Data Acquisition

### 2.1 Overview:

For this project the team worked with a plan to use data from reliable sources that were all publicly available. Initially we had gathered nine datasets, all sourced from the Federal Reserve Bank of St. Louis. These datasets included:

- Federal Funds Effective Rate
- Gross Domestic Product (GDP)
- House Price Index
- Housing Affordability Index
- Monthly Supply of New Homes
- New Privately Owned Housing
- Personal Consumption Expenditures Price Index (PCEPI)
- Personal Real Income (RPI)
- NBER Recession Indicators
- Yield Curve

Each dataset consists of two columns: date and value. The datasets are available on the FRED website and can be downloaded in CSV, Excel, Image, or PowerPoint formats. We primarily utilized the CSV format.

Most datasets cover the primary period of interest from 2006 to 2012. However, 1 data set, the Housing Affordability Index only contains data for 2024 and 2025, and some others have not yet been updated for 2025. The datasets are well-formatted and ready to use. They can be concatenated based on dates, though some minor data manipulation may be required.

Due to the limited range of Housing Affordability Index data, we substituted other data, including Housing Price Index, Real Income and housing supply, as discussed in the section following Figure 4.

## 2.2 Field Descriptions:

As mentioned previously, each data set included two columns: date and value. With only minor exceptions, the date field was similar in each set and is described in paragraph 2.2.1 below. All the second columns are unique to each dataset, so they are addressed in paragraph 2.2.2 below. These sections will include the names of each field, its associated data type, format, and any additional notes like exceptions or abnormalities.

### 2.2.1 Common Field:

Observation Date is common to each data set we are evaluating.

- `observation_date` (Type: string)- The 1<sup>st</sup> column of each data set is the date. Except for the NBER Recession Indicators, all dates are in the mm/dd/year format. Recession Indicators are in the year-mm-dd format. None of the dates are null.

### 2.2.2 Unique Fields by Data Set:

Each data set contains 2 columns. The 2nd column for each set is described below.

- **T10Y3M** (Type: real number)- The 2<sup>nd</sup> column of the Yield Curve data set is a percent. It is calculated as the spread between 10-Year Treasury Constant Maturity (BC\_10YEAR) and 3-Month Treasury Constant Maturity (BC\_3MONTH). 4.2% of the rows are nulls, which corresponds to holidays when markets were closed.
- **PCEPI** (Type: real number)- The 2<sup>nd</sup> column of the Personal Consumption Expenditures data set is an index. None of the indexes are null.
- **FEDFUNDS** (Type: real number)- The 2<sup>nd</sup> column of the Federal Funds Effective Rate data set is a percent. None of the rates are nulls.
- **GDPC1** (Type: real number)- The 2<sup>nd</sup> column of the Real GDP data set is a quantity in billions of Chained 2017 Dollars. None of the quantities are nulls.
- **FIXHAI** (Type: real number)- The 2<sup>nd</sup> column of the Housing Affordability Index data set is an index. None of the indexes are null. This data set is assessed to be valuable for analysis of future recessions but was not used for this project due to insufficient date range of coverage.
- **RPI Type:** real number)- The 2<sup>nd</sup> column of the Real Personal Income is personal income in chained 2009 dollars (RPI) is personal income in current dollars (PI) deflated by the PCE chained price index (PCEPI). None of the quantities are nulls.
- **USSTHPI** (Type: real number)- The 2<sup>nd</sup> column of the All-Transactions House Price Index for the United States data set is an index. None of the indexes are null.
- **MSACSRNSA** (Type: real number)- The 2<sup>nd</sup> column of the Monthly Supply of New Houses in the United States data set is a quantity representing months of supply. None of the quantities are nulls.
- **HOUST** (Type: real number)- The 2<sup>nd</sup> column of the New Privately-Owned Housing Units Started data set is a quantity in thousands of units. None of the quantities are nulls.
- **USRECQ** (Type: integer)- The 2<sup>nd</sup> column of the NBER based Recession Indicators data set is a binary set with 0 = no recession indicated, 1 = recession indicated. None of the integers are nulls.

## 2.3 Data Context:

This project analyzes how selected macroeconomic and housing indicators behaved during the period surrounding the 2008 financial crisis, officially dated from December 2007 to June 2009 by the National Bureau of Economic Research (NBER, 2024). The data range for most indicators begins in 2006 and extends through 2012 to capture pre-, during-, and post-recession trends.

Rather than building a predictive model, the objective is to identify recurring patterns in key variables and assess their potential as early warning indicators. The chosen timeframe enables comparison between leading signals—such as the yield curve and housing affordability—and outcome-based measures like real GDP contraction or recession markers.

All data were retrieved from publicly available, government-backed sources such as the Federal Reserve Bank of St. Louis (FRED), the Federal Housing Finance Agency (FHFA), the U.S. Census Bureau, and the Bureau of Economic Analysis. Data were downloaded in CSV format and cleaned and standardized using Excel and Python, with a focus on aligning timeframes across datasets. Resolution of nulls was not required due to completeness of each dataset.

Indicators that are reported at different frequencies (monthly vs. quarterly) were resampled or aggregated as needed to ensure comparability. For example, real GDP (quarterly) was aligned with monthly indicators using midpoint matching for interpretation. When required, transformation logic was applied, such as calculating average rates per month or deriving spreads (e.g., 10-year minus 3-month Treasury yields).

Variables were classified by their analytical purpose:

- **Leading indicators** (e.g., yield curve, housing starts, affordability)
- **Concurrent indicators** (e.g., housing price index, monthly supply of homes)
- **Outcome indicators** (e.g., real GDP)

These categories help frame the indicators' roles in understanding recession onset, depth, and recovery. This classification also informed the segmentation of the data into three analytical periods: pre-recession buildup, recession onset and peak, and post-recession recovery.

To minimize bias and allow replication, all original datasets were documented with source metadata and preserved in their original structure before transformation. A list of variables, units, frequency, and original source URLs is provided in Table 1.



<i>Table 1: Summary of Indicators Used</i>			
Indicator	Frequency	Source	Role
Federal Funds Rate	Daily	FRED	Monetary Policy Context
Real Gross Domestic Product (GDP)	Quarterly	BEA/FRED	Outcome Indicator
Housing Affordability Index (HAI) *	Monthly	NAR/FRED	Leading Indicator
* Housing Price Index (HPI)	Monthly	FHFA/FRED	Concurrent Indicator, Affordability Proxy – Cost Component
* Monthly Supply of New Homes	Monthly	Census/FRED	Concurrent Indicator, Affordability Proxy – Supply Component
* Real Personal Income (RPI)	Monthly	BEA/FRED	Affordability Proxy – Income Component
Housing Starts	Monthly	Census/FRED	Leading Indicator
NBER Recession Indicator	Quarterly	NBER/FRED	Benchmark
Yield Curve Spread (10Y–3M)	Daily	Fed/FRED	Leading Indicator

*\* Because the Housing Affordability Index lacked sufficient historical coverage, affordability trends were approximated using three components available from 2006–2012: the Housing Price Index (reflecting cost trends), Real Personal Income (purchasing power), and Monthly Supply of New Homes (market pressure). These variables served as a composite affordability proxy in place of the excluded HAI series. Real Personal Income was sourced from the Federal Reserve Bank of St. Louis under the Real Personal Income (RPI) series, expressed in billions of dollars.*

## 2.4 Data Conditioning

Data was sourced from the Federal Reserve Economic Data (FRED), U.S. Census Bureau, and Bureau of Labor Statistics. Indicators were selected based on economic theory and prior research. The response variable was a binary indicator of recession periods, defined using NBER's classification. Data transformations included creation of yield curve spreads (10Y-2Y and 10Y-3M), normalization, and categorical recession labeling.

Reviews of each data set revealed some of the key summary data, as well as areas that required data conditioning. Summary data is included in Table 2. It includes the date range covered for each data set, and the range of values for each second column, as described in the data fields of section 2.2 above.

<b>Table 2: Summary Data</b>				
Name	Earliest Date	Latest Date	Min	Max
Fed Funds Effective Rate	1/1/1955	9/1/2024	0.05	19.1
Gross Domestic Product	1/1/1947	7/1/2024	2,172.432	23,542.349
House Price Index	1/1/1975	7/1/2024	60.01	690.9
Housing Affordability Index	1/1/2024	1/1/2025	91.6	104.9
Monthly Supply of New Homes	1/1/1959	9/1/2024	478	2,494
Personal Consumption Expenditures Price Index	1/1/1959	2/1/2025	125.6	15.164
NBER Recession Indicators	1/1/1900	7/1/2024	This is a binary indicator, where: 1 = recession, 0 = not recession.	
Yield Curve	6/1/1976	3/27/2025	-2.41	2.91

Assessment of the data revealed the following about required conditioning.

The original source for Housing Affordability Index does not meet the required needs, as data only goes back to 2024, so it includes no 2008 recession indicators. As a result, this dataset was excluded. To retain affordability analysis, it was replaced with proxy data composed of the Housing Price Index, Real Personal Income, and Monthly Supply of New Homes—all of which were available across the full-time window.

Most other data sets do not have nulls or repeats. The only exception is the Yield Curve data, which contains approximately 4.2% nulls. The nulls in this set correspond to dates that the market was closed, so these nulls will not need to be cleaned up when analyzing this data set independently. When aligning data with other datasets, matching dates may become problematic, so the team decided on a convention for aligning the dates.

Each data set contains a different range of dates, with starts as early as 1900, and as current as March 2025 (with most ending in 2024). For conditioning, all the data sets were set to the specific ranges of the analysis. For pre-2008 recession indicators the team used data starting with 2006, and post-recession data ending with 2012.

All but one data set is in a common date format. The one exception is NBER Recession Indicators, so that date format was transitioned to match the others.

Four of the eight usable original data sets contain monthly increments of data (PCEPI, FEDFUNDS, MSACSRNSA, and HOUST), three are quarterly (GDPC1, USSTHPI, and USRECQ), and one (Yield Curve) is daily. When analyzed independently, the periodicity remained unchanged. When compared with other data, common periodicity was assigned.

- **Fed Funds Effective Rate** has the date range that meets our standards. Also, the values are in correct format and no null data is found. Date format is in YYYY-DD-MM and all rates are as of the first day of the month. Rates are expressed in decimal format to the second decimal place.
- **Gross Domestic Product (GDP)** record data at quarterly frequency. Date format is in YYYY-DD-MM and all rates are as of the first date of the month. Values are in decimal form and are dollar amounts measured in billions. This dataset is clean with correct data format and no missing values.
- **House Price Index** records the price index in the U.S. quarterly. Like other datasets found on FRED websites, the data is well-formatted and complete. There is a source on the FHFA for Housing Price Index on a monthly cadence for the Purchase-Only Index (excludes refinancing mortgages). This data goes all the way back to 1991 and is formatted well. <https://www.fhfa.gov/data/hpi/datasets?tab=monthly-data>. It contains the Datetime in MM/DD/YYYY format and decimals for the index. The data is split up into regions but also contains the whole US.
- **New Privately Owned Housing** reports the units of homes owned by individuals or families. The data was recorded at monthly level with a date range that satisfies the scope of our analysis. The unit value is recorded as integer type, and there is no missing data.
- **Monthly Supply of New Homes** dataset records the ratio of new houses for sale and new houses sold monthly. According to FRED, the statistics indicate how long the current for-sale would last in months, given the current sales rate if no new houses were built.
- **Personal Consumption Expenditures** – Chained-type Price Index (PCEPI): According to FRED, PCEPI “is a measure of the prices that people living in the United States, or those buying on their behalf, pay for goods and services.” PCEPI is known for capturing

inflation/deflation and changes in consumer behaviors. Like other datasets recorded by Fred, this dataset contains 2 columns – observation\_date and value. Date is formatted as YYYY-MM-DD and value is stored as float type. Data is recorded monthly with no null values.

- **Real Personal Income** - Calculated by the Federal Reserve Bank of St. Louis: Personal income in chained 2009 dollars (RPI) is personal income in current dollars (PI) deflated by the PCE chained price index (PCEPI).

## 2.5 Data Risks:

Appendix B contains 2 Risk Matrixes. The first summarizes Project Risks, such as revision control of the deliverables, loss of data or content, schedule slippage, project complexity, and security breaches. All these risks could contribute to not meeting customer expectations. Because the project is dependent on data, a separate risk matrix is also included to address specific Data Risks, including data availability and quality, correlation and causation analysis, external influences, and modeling challenges.

Each Risk Matrix lists and describes the applicable risks, assigns them each a probability of occurrence and a potential impact level, and concludes with planned mitigations (or acceptance of the risk, where warranted). Both the probability and impact level were assessed as Low, Medium or High.

## 2.6 Data Quality Assessment:

The data quality was assessed by evaluating the completeness, uniqueness, accuracy, atomicity, conformity and overall quality. Atomicity does not apply to selected data sets, so that is not addressed below.

- **Completeness:** Most of the datasets selected as candidates early in the project start well before the 2008 recession. Only 3 of the sets cover any data in 2025, but the rest go well into 2024. If limiting exploration to the years leading up to the 2008 recession, and stopping the analysis short of 2024, 8 of the 9 can be considered complete. Each data set was also checked for nulls. 8 of 9 were free of nulls, and the Yield Curve had only 4.2% of the record with nulls. Only Housing Affordability Index (FIXHAI) is incomplete for the purpose of the analysis (range of dates is 1/1/24-1/1/25).
- **Uniqueness:** Each data set has been checked to ensure that each data record is distinct, so duplicates will not be a problem.
- **Accuracy:** The Fed has established data governance processes to ensure information consistency and transparency. They assess the accuracy of their data through various methods, including comparing data from different sources, conducting data audits, and using data quality frameworks, as well as relying on timely and accurate filing of reports by financial institutions. (GAO)
- **Conformity:** The Fed has established data governance processes to ensure information consistency and transparency. They assess the accuracy of their data through various methods, including comparing data from different sources, conducting data audits, and using data quality frameworks, as well as relying on timely and accurate filing of reports by financial institutions (GAO). Their processes help ensure consistency and usability.
- **Overall Quality:** Overall, this is high quality data. Except for 1 incomplete data set that is not usable for this project's purposes, the remainder are complete, unique, accurate, and they demonstrate conformity.

## 2.7 Other Data Sources

Of the initial 9 data sets selected as candidates, the only one excluded is the Housing Affordability Index because it only contains data starting in 2024. Instead, this project is using Housing Price Index and Real Income data as a proxy. In addition to the indicated sources of data, the team have referred to other sources of information and analysis to help assess the data. A complete list of references in the Works Cited page.

### 3 Analytics and Algorithms

This section outlines the analytical framework and modeling techniques used to evaluate relationships between macroeconomic indicators, housing market behavior, and recessionary patterns. A combination of statistical and machine learning approaches was employed to analyze both continuous outcomes (e.g., house price index) and binary outcomes (e.g., recession classification).

#### 3.1 Methodology

The analysis employed multiple modeling techniques to evaluate the relationship between macroeconomic indicators and both housing market outcomes and recession probability. Each method was selected based on its suitability for the task. These techniques were used to analyze both continuous housing market outcomes (e.g., house price index) and binary economic indicators (e.g., recession classification via yield curve inversion). The four modeling approaches employed were:

- **Ordinary Least Squares (OLS) Regression:** OLS regression was applied to estimate the relationship between the housing price index and lagged macroeconomic variables, including the yield curve spread, federal funds rate, and inflation. This technique enabled interpretability of coefficient direction and magnitude.
- **Logistic Regression:** A logistic regression model was used to predict the likelihood of a yield curve inversion—a binary outcome—based on macroeconomic variables. The model estimated the probability that observed economic conditions corresponded to an inverted curve.
- **Random Forest Classifier:** A random forest model was employed to capture non-linear relationships and interaction effects. This ensemble method leveraged lagged features to incorporate temporal structure, offering strong performance in ranking variable importance. However, the tradeoff was reduced interpretability compared to linear models, warranting further validation.
- **XGBoost Classifier:** XGBoost was deployed to validate and enhance classification performance. By leveraging sequential tree boosting and regularization techniques, XGBoost reduced the risk of overfitting while maintaining strong predictive accuracy. Its ability to model complex feature interactions made it particularly well-suited to the classification of recession likelihood.

Collectively, these models offered complementary strengths: OLS enabled continuous value estimation with interpretable coefficients, logistic regression supported binary classification with statistical rigor, Random Forest uncovered non-linear patterns and feature importance, and XGBoost provided advanced ensemble-based performance optimization.

### 3.2 Tools and Infrastructure

All data preprocessing, modeling, and visualization tasks were conducted using the Python programming language. The analytical workflow was supported by the following libraries:

- **Pandas and NumPy** for data manipulation and transformation.
- **Matplotlib and Seaborn** for exploratory and final visualizations.
- **Statsmodels** for statistical analysis and linear modeling.
- **Scikit-learn** for classification models and performance evaluation.

Version control and team collaboration were managed using GitHub and Microsoft Teams.

These platforms enabled asynchronous contributions, branch-based development, and centralized documentation, ensuring consistency across development environments and facilitating effective project coordination.



### 3.3 Data Preparation for Modeling

The dataset spans the period from 2006 to 2012 and includes monthly and quarterly macroeconomic data sourced from public databases including the Federal Reserve and U.S. Census Bureau. To replicate a real-world predictive environment, key features were lagged by up to six months. These lagged variables included the yield curve spread, federal funds rate, and new privately owned housing units.

Preprocessing steps included z-score normalization of continuous variables to ensure model compatibility, as well as imputation of missing values using forward-fill techniques to preserve temporal integrity. A composite housing affordability index was developed using a weighted combination of real personal income, housing starts, and home price data, allowing for a more consistent proxy over the historical time frame.

For classification tasks such as predicting recession periods, the target variable was the NBER-defined recession indicator (binary). For some classification models, features such as the Treasury spread were binarized to emphasize inversion behavior.

## 4 Visualizations

This section presents key visualizations developed to support the analysis, clarify trends, and communicate model results. Visualizations were designed to highlight macroeconomic shifts surrounding the 2008 financial crisis and to validate key statistical and machine learning findings.

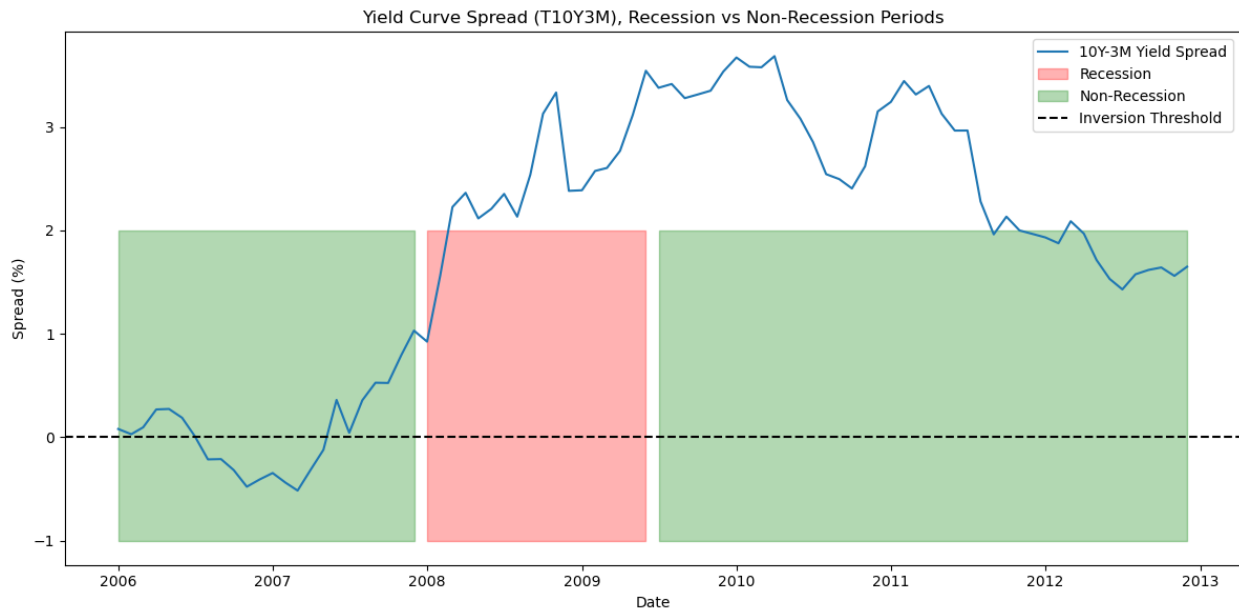
### 4.1 Visualization Tools and Implementation Details

This subsection outlines the tools, platforms, and methodologies used to create the project's visualizations, including implementation details, programming environment, testing procedures, and version control strategy.

- **Visualization Tools:** Python (matplotlib, seaborn), Jupyter Notebooks
- **Programming Language:** Python 3.11
- **Code Size Estimate:** 100–150 lines per figure
- **Code Complexity:** Medium
- **Implementation Required:** Yes
- **Version Control:** GitHub (private repository, team-managed)
- **Testing Responsibility:** Team members (Huynh, Graham, Haggerty, Duan) conducted visual validation against processed data tables and model outputs

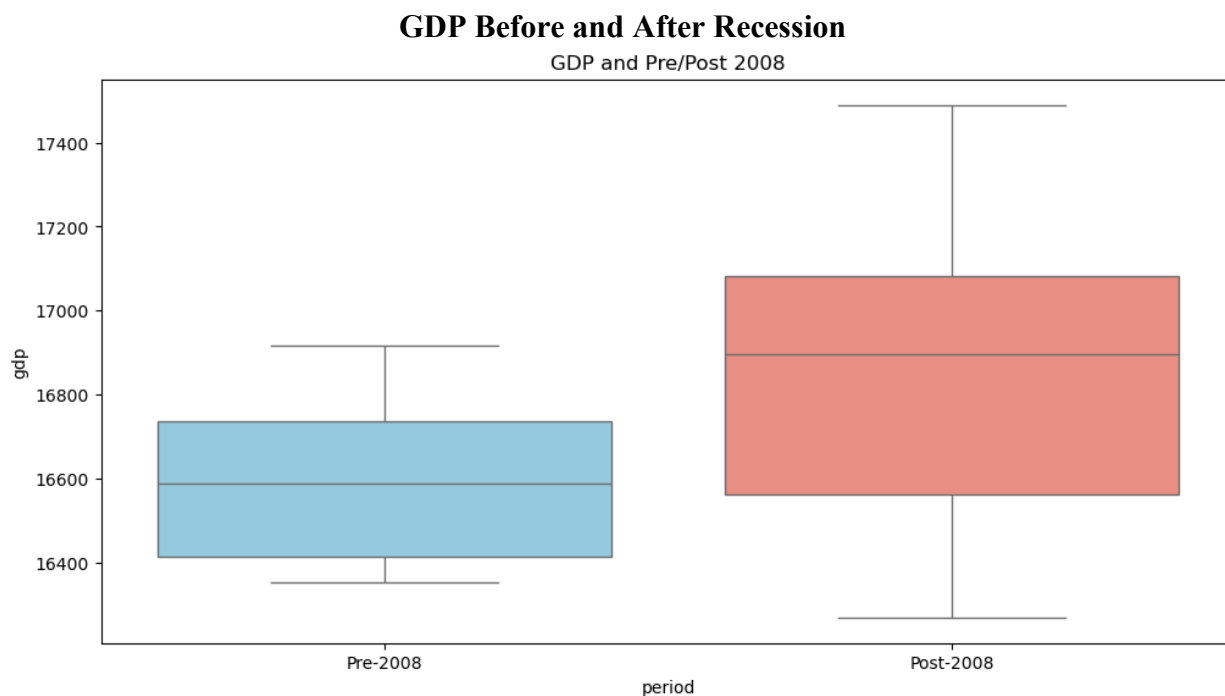
## 4.2 Visualization Applications and Descriptions

This subsection outlines the purpose, structure, and implementation details of the visualizations used in the study. It includes visualization types, tools applied, and contextual considerations such as programming language, version control, and associated risks.



*Figure 11: Yield Curve Spread (T10Y3M), Recession vs. Non-Recession Periods*  
*Source: Federal Reserve Bank of St. Louis; team analysis. Nguyen et al. (2024)*

To assess the relationship between the yield curve and economic downturns, the 10-Year minus 3-Month Treasury spread was analyzed in conjunction with official U.S. recession periods (as defined by the NBER). As shown in Figure 11, the curve inverted prior to the 2008 recession, consistent with patterns discussed in Section 1.

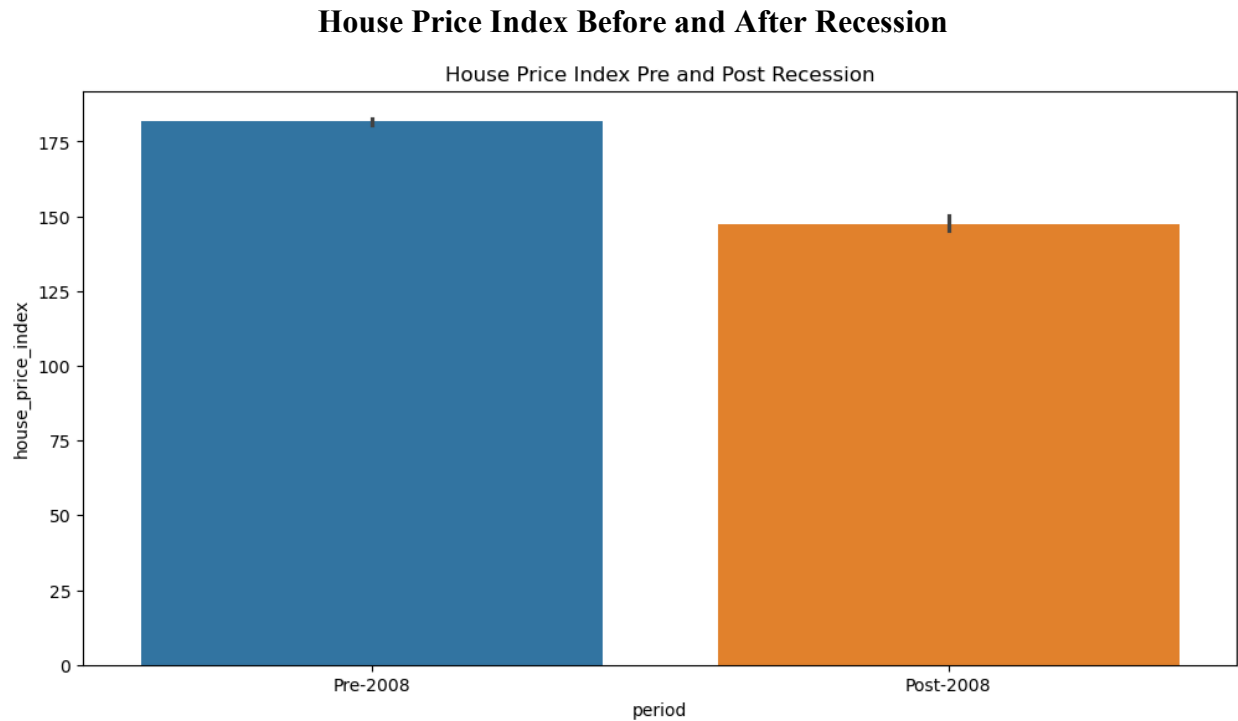


*Figure 12: GDP Pre-2008 and Post-2008*

*Source: Federal Reserve Bank of St. Louis; team analysis.*

Figure 12 visualizes the distribution of GDP before and after the recession.

Each box represents the minimum and maximum values (bottom and top whisker bars, respectively), the 25th percentile, the median, and the 75th percentiles of the data. This clearly shows that after the 2008 recession, median Gross Domestic Product is greater, indicating economic growth. It also demonstrates an increase in spread of GDP after the recession, which can be indicative of changes in economic dynamics. It is significant that the 1st quartile of the post-recession GDP is nearly the median of the pre-recession period.



*Figure 13: House Price Index Pre-2008 and Post-2008*

*Source: Federal Reserve Bank of St. Louis; team analysis.*

Figure 13 displays the house price index before and after the 2008 recession. The average house price index was significantly higher before the recession, which aligns with the housing bubble that preceded the financial crisis. The subsequent drop in the house price index indicates a crash of the real estate market following the recession.

## Income Before and After 2008 Recession

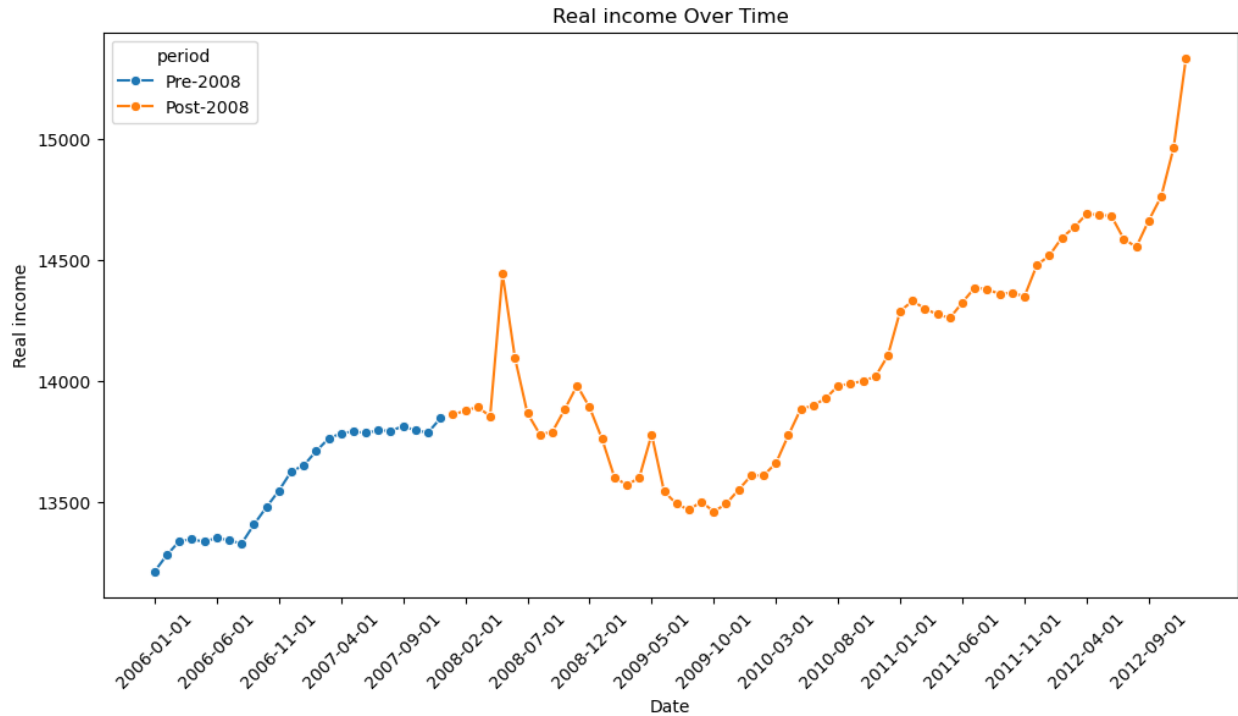
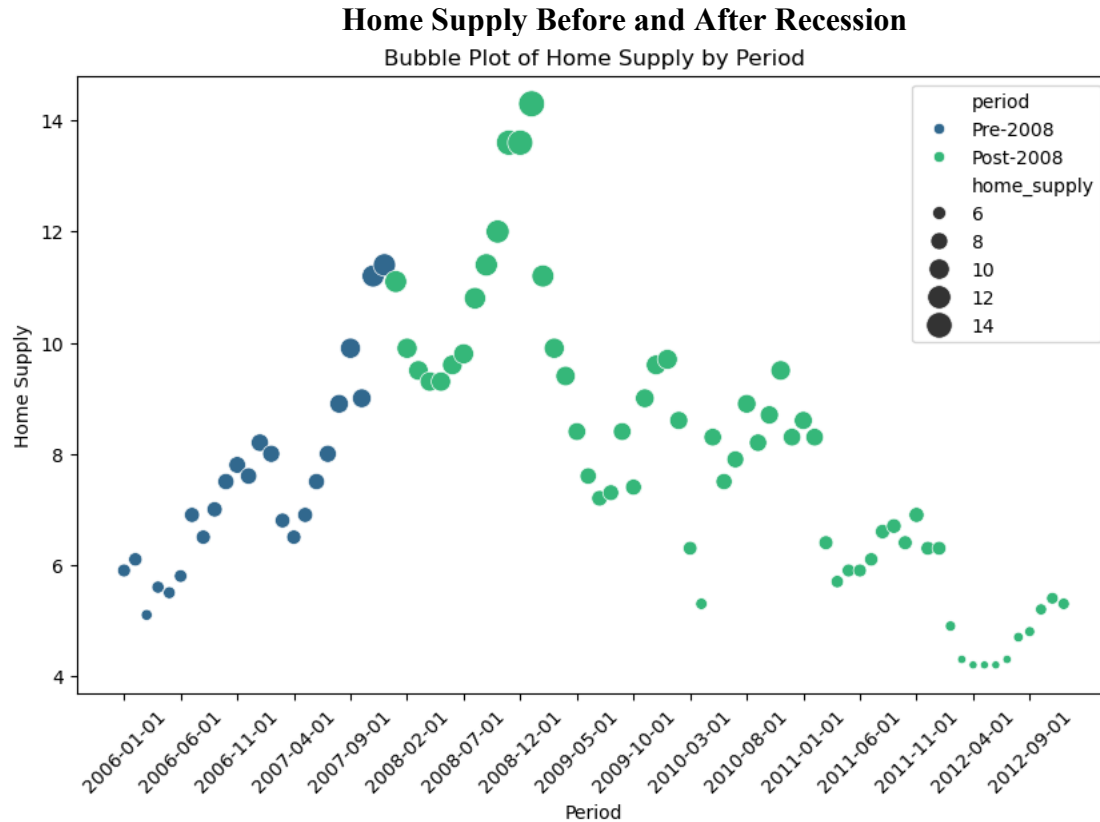


Figure 14: Income Pre-2008 vs. Post-2008

Source: Federal Reserve Bank of St. Louis; team analysis.

Figure 14 shows the increase in income after 2008 recession, which may reflect stimulus measures or sectorial shifts, where some sectors of the economy grow more strongly after the recession. Income and GDP move from one sector to another.

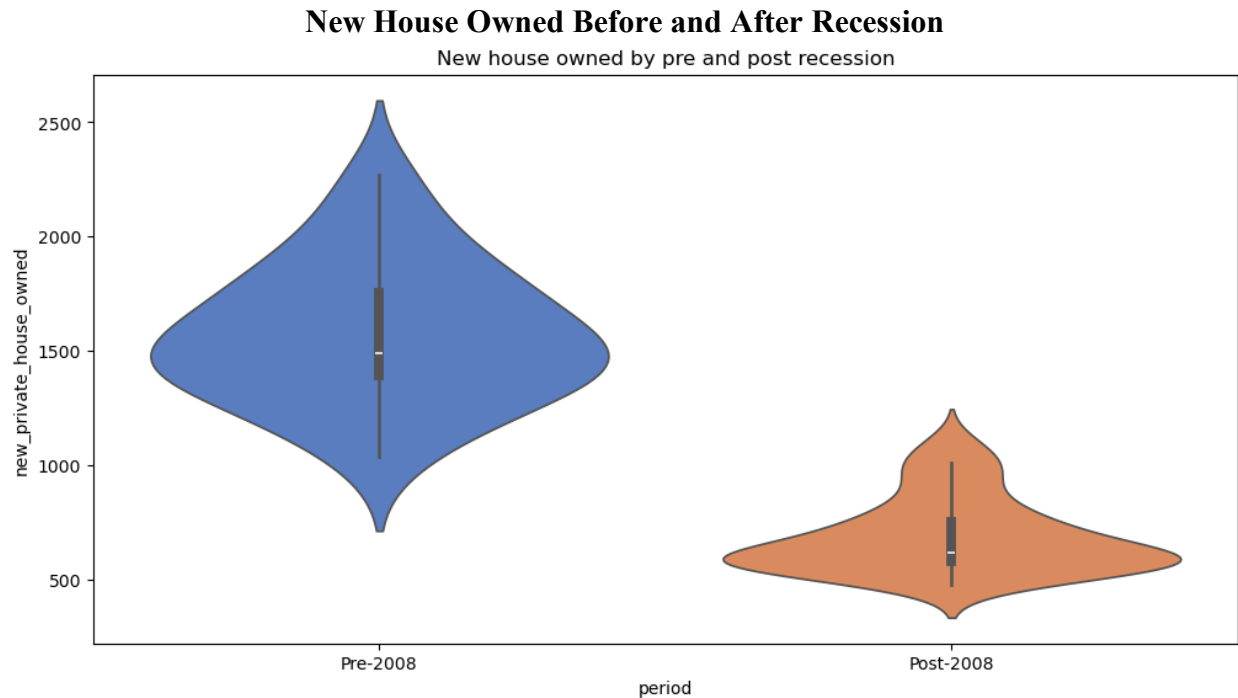
After the 2008 recession, tech and healthcare recovered much faster and grew stronger compared to other sectors in the economy. More workers are allocated to these growing sectors, and higher incomes are assigned to these sectors.



*Figure 15: Home supply Pre-2008 vs. Post-2008*

*Source: Federal Reserve Bank of St. Louis; team analysis*

In figure 15, the bubble plot is used to represent the inventory of housing before and after recession. The size of the bubbles clearly demonstrates that there was a housing boom before and during recession and a sharp decline after recession.



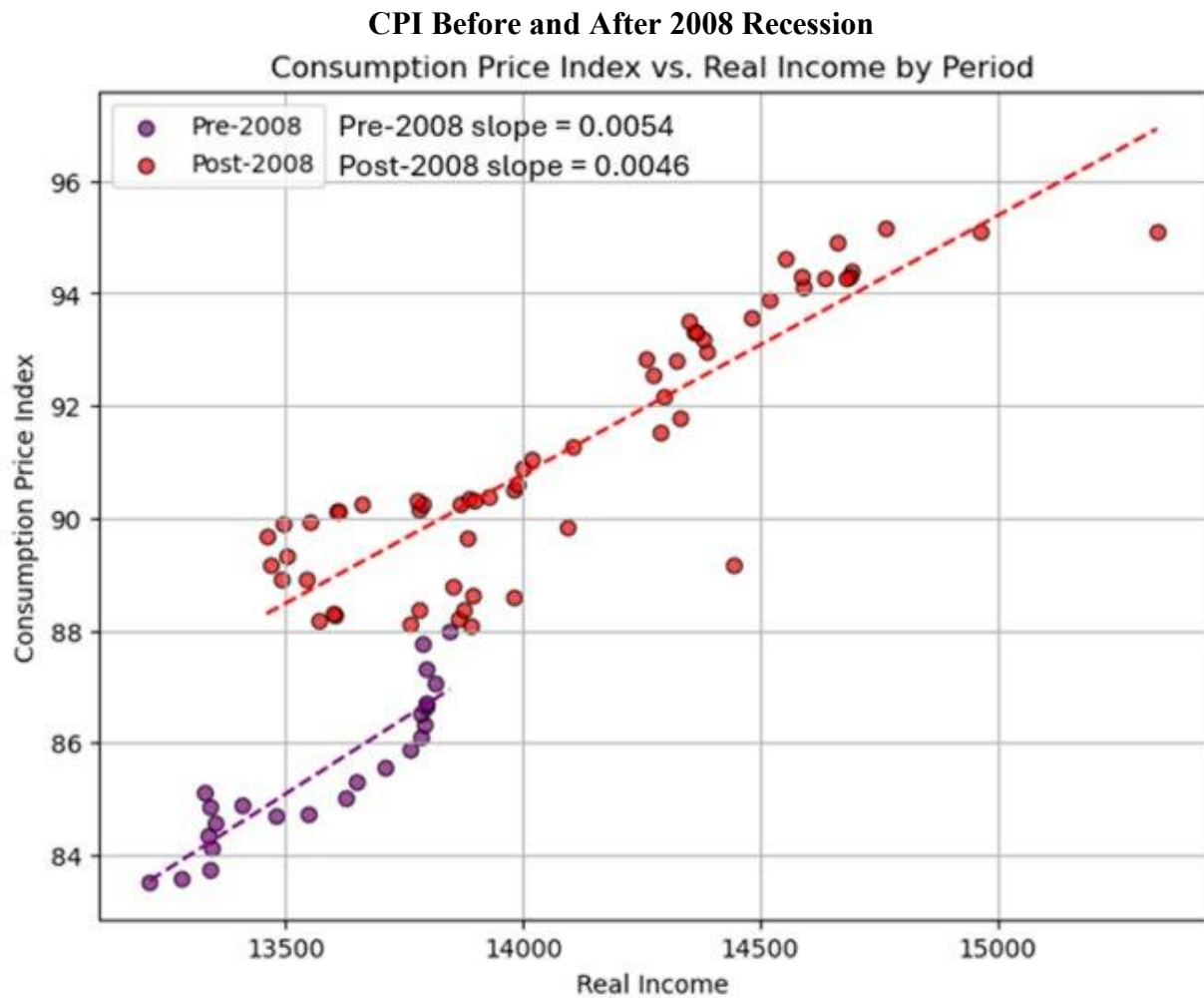
*Figure 16: Newly Owned Homes Pre-2008 vs. Post-2008*

*Source: Federal Reserve Bank of St. Louis; team analysis.*

Figure 16 is a Violin Plot of newly owned homes before and after the 2008 recession. It shows the distribution of new private homes owned in the U.S. before and after the 2008 recession, with the scale in thousands of units. The pre-recession period mostly covers 1,000,000-2,300,000 units, with a median around 1,500,000 units. The post-recession period spans 500,000-1,000,000 units, with a median of about 750,000 units.

This figure shows that before the recession, the number of newly owned homes was much higher, and more widely spread out. After the recession, the number of new homes owned dropped significantly and had generally less variation. The decrease in newly private home owned indicates declining demands that further leads to construction plummeting.





*Figure 17: Consumption Price Index (CPI) Pre-2008 vs. Post-2008*

*Source: Federal Reserve Bank of St. Louis; team analysis.*

Figure 17 is a Scatter Plot showing the relationship between Real Income and the Consumption Price Index (CPI). The pre-2008 recession data is shown in purple and illustrates that both real incomes and CPI were generally lower in the pre-recession period, and there was a moderately increasing trend. This shows that even as people earned more, prices didn't skyrocket as much, indicating a healthy economic balance.

In the post-recession period, both the real income and CPI shifted higher, and with an even greater positive trend. This means that even as incomes recovered after the recession, prices increased even faster, meaning real purchasing power was reduced.

### 4.3 Visualization Risks and Mitigation

Visualization risks are summarized in a table in Appendix B. Risks include misleading visuals, loss of context, misuse of color, and excess clutter. In most cases, the mitigations include team review of content produced by others to ensure the right messages are being communicated.

## 5 Findings

This section presents and interprets the key outcomes from the regression and classification models used in this study. It synthesizes the statistical significance of recession predictors, evaluates model performance, and explains feature importance.

## 5.1 Logistic Regression and Hypothesis Testing

A logistic regression model was used to assess the relationship between the yield curve spread and the probability of recession. The binary dependent variable was defined by the NBER recession indicator, while the independent variable was the treasury maturity spread (10-year minus 3-month).

Logit Regression Results						
Dep. Variable:	recession_indicator	No. Observations:	84			
Model:	Logit	Df Residuals:	82			
Method:	MLE	Df Model:	1			
Date:	Sat, 19 Apr 2025	Pseudo R-squ.:	0.06044			
Time:	12:52:40	Log-Likelihood:	-41.007			
converged:	True	LL-Null:	-43.645			
Covariance Type:	nonrobust	LLR p-value:	0.02162			
	coef	std err	z	P> z	[0.025	0.975]
const	-2.3940	0.636	-3.763	0.000	-3.641	-1.147
treasury_maturity	0.5204	0.246	2.114	0.034	0.038	1.003

Figure 18: Logistic Regression Output (Recession Indicator vs. Treasury Spread)

Source: Team analysis based on FRED T10Y3M series and NBER data. Nguyen et al. (2024)

The regression yielded a p-value of 0.02162, indicating a statistically significant relationship between curve inversion and recession likelihood ( $p < 0.05$ ), as shown in Figure 18.

The hypotheses test below summarizes the significance of the assessment.

### Hypotheses:

- $H_0$ : No relationship exists between the yield curve and the probability of a recession.
- $H_1$ : A statistically significant relationship exists between the yield curve and the probability of a recession.

### Testing Metric:

- $p\text{-value} < 0.05$ : Indicates statistical significance.
- $p\text{-value} \geq 0.05$ : Indicates no statistically significant relationship.

### Result:

- Logistic regression yielded a p-value of 0.02162, confirming a statistically significant relationship between yield curve inversion and recession probability.

## 5.2 Ordinary Least Square (OLS)

To assess which macroeconomic indicators predict housing price changes, an OLS regression was applied with the House Price Index as the dependent variable. The model achieved an adjusted R-squared of 0.971, indicating that 97.1% of the variance in housing prices could be explained by the model.

The regression included multiple macroeconomic features, including a lagged version of the yield curve spread—representing the difference between 10-year and 3-month treasury maturities at prior time points. Specifically, yield curve spreads were tested at 1-month, 3-month, and 6-month lags.

OLS Regression Results						
Dep. Variable:	house_price_index	R-squared:	0.974			
Model:	OLS	Adj. R-squared:	0.971			
Method:	Least Squares	F-statistic:	285.0			
Date:	Tue, 29 Apr 2025	Prob (F-statistic):	2.31e-50			
Time:	18:50:44	Log-Likelihood:	-187.04			
No. Observations:	78	AIC:	394.1			
Df Residuals:	68	BIC:	417.6			
Df Model:	9					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	233.4371	27.501	8.488	0.000	178.561	288.314
federal_rate	2.6844	1.036	2.591	0.012	0.617	4.752
real_income	-0.0029	0.002	-1.158	0.251	-0.008	0.002
consumption_price_index	-1.2536	0.714	-1.755	0.084	-2.679	0.172
home_supply	0.8165	0.219	3.721	0.000	0.379	1.254
gdp	0.0030	0.004	0.759	0.450	-0.005	0.011
new_private_house_owned	0.0167	0.004	4.354	0.000	0.009	0.024
yieldcurve_lag1	0.7627	0.963	0.792	0.431	-1.159	2.684
yieldcurve_lag3	1.7280	1.106	1.563	0.123	-0.479	3.935
yieldcurve_lag6	-3.3672	0.798	-4.220	0.000	-4.960	-1.775
Omnibus:	4.113	Durbin-Watson:	0.476			
Prob(Omnibus):	0.128	Jarque-Bera (JB):	3.819			
Skew:	-0.473	Prob(JB):	0.148			
Kurtosis:	2.469	Cond. No.	1.87e+06			

Figure 19: Ordinary Least Square Output (House Prince Index vs. Treasury Spread)

Source: Team analysis based on FRED data. Nguyen et al. (2024)

According to Figure 19, significant predictors ( $p < 0.05$ ) included:

- Federal Funds Rate
- Home Supply
- Newly Privately Owned Housing Units
- Yield Curve (lagged by 6 months)

Notably, the coefficient for the 6-month lagged yield curve was negative and statistically significant, indicating that an inverted curve six months earlier was associated with a decline in housing prices. This suggests that yield curve behavior may serve as a leading indicator of housing market stress.

These results guided feature selection for subsequent machine learning models by emphasizing variables with meaningful economic influence.

5.3 Random Forest Classifier

The Random Forest Classifier model was trained to classify whether the treasury maturity yield curve spread is above or below its median value, using six-month lagged economic indicators.

The model incorporates the following features: House Price Index, Newly Privately Owned Housing Units, Housing Supply, Federal Funds Rate, Real Income, Consumer Price Index (CPI), and Gross Domestic Product (GDP).

To enable binary classification, the treasury maturity yield curve spread was transformed into a binary target variable by thresholding at its median value. Observations above the median were labeled as “high” and those below as “low,” allowing the model to distinguish between elevated and suppressed yield curve environments.

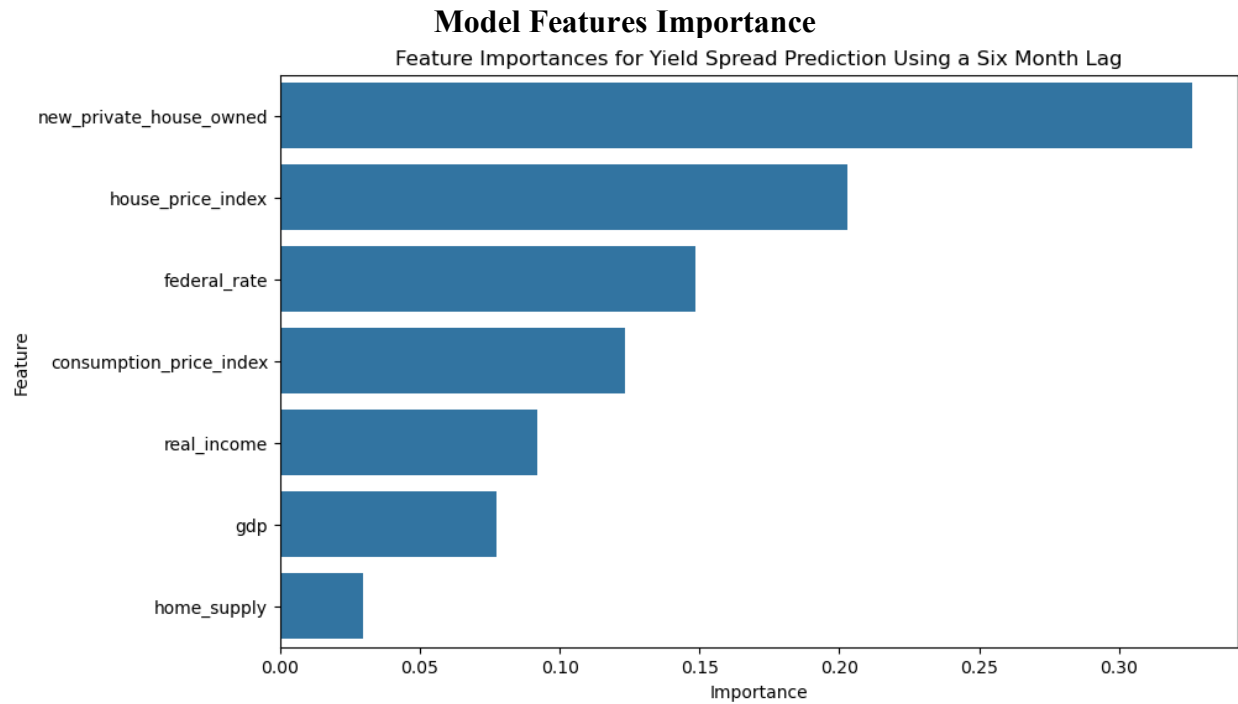
The dataset was divided into training and testing sets using an 80/20 split, with 80% allocated for model training and 20% reserved for evaluation.

Model Performance Metrics:				
AUC Score: 1.0				
Classification Report:				
	precision	recall	f1-score	support
0	0.89	1.00	0.94	8
1	1.00	0.88	0.93	8
accuracy			0.94	16
macro avg	0.94	0.94	0.94	16
weighted avg	0.94	0.94	0.94	16

Figure 20: Random Forest Model Performance Metrics

Source: Team analysis based on FRED data. Nguyen et al. (2024)

According to the Model Performance Metrics in Figure 20, the model’s AUC score indicates that the model can perfectly distinguish between high and low treasury maturity yield curve spreads.



*Figure 21: Random Forest Model Features Importance*

*Source: Team analysis based on FRED data. Nguyen et al. (2024)*

A model importance features analysis was performed to give insight into which economic factors most influenced the treasury maturity yield curve spread.

According to Figure 21, the Random Forest Classifier model identified the top three economic indicators influencing the treasury maturity yield curve spread: Newly Private Home Owned, House Price Index, Federal Funds Rate.



5.4 XGBoost Classifier

To validate Random Forest results and assess potential performance gains, an XGBoost classifier was implemented on the same lagged dataset.

The model used the same features, target, and split in the Random Forest Classifier.

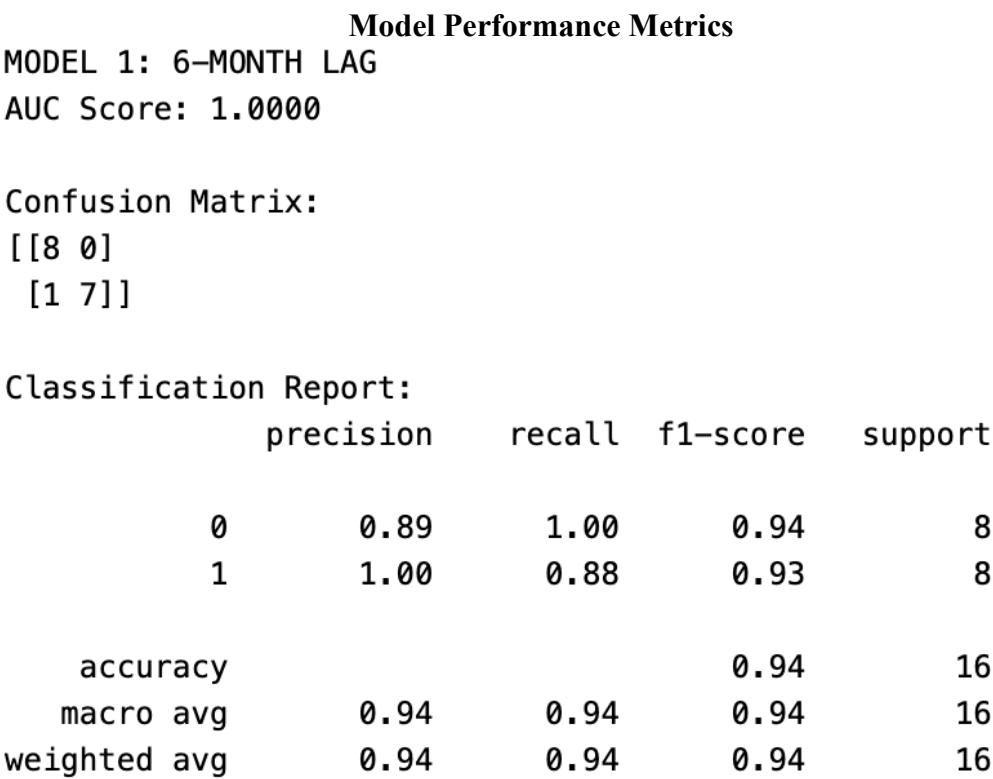
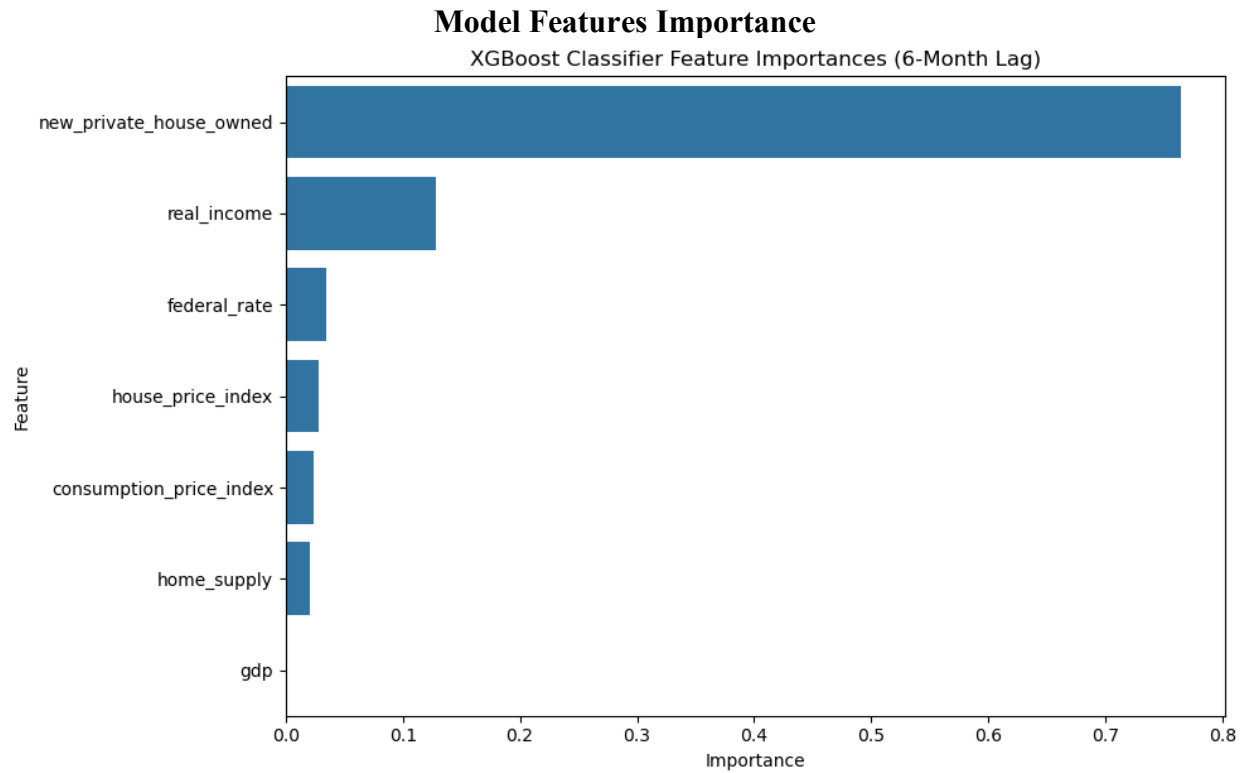


Figure 22: XGBoost Model Performance Metrics

Source: Team analysis based on FRED data. Nguyen et al. (2024)

According to Figure 22, the model’s AUC score demonstrates that this model is a strong candidate for predicting treasury maturity yield curve spread based on key macroeconomic indicators.



*Figure 23: XGBoost Model Features Importance*

*Source: Team analysis based on FRED data. Nguyen et al. (2024)*

According to Figure 23, the XGBoost's model features importance analysis revealed the top three economic indicators influencing the treasury maturity yield curve spread are: Newly Private Home Owned, Real Income, and Federal Funds Rate

## 5.4 Interpretation of Economic Shifts (Table 3)

To evaluate broader economic shifts, summary statistics were calculated for key indicators before and after the 2008 recession.








Table 3: Summary of Economic Indicators- Pre-2008 vs. Post-2008							
Period	House Price Index	New Private House Owned	Home Supply	Federal Rate	Real Income	Consumption Price Index	GDP
Pre-2008	181.71	1,576.88	7.48	4.99	13,580.61	85.52	16,597.80
Post-2008	147.50	687.03	7.91	0.50	14,087.98	91.19	16,883.10
Trend							

Table 3: Indicators Pre-2008 vs Post-2008

Source: Team analysis using FRED datasets. Nguyen et al. (2024)

Table 3 presents key economic indicators before and after the 2008 crisis. Notable changes were observed in housing activity, income levels, interest rates, and GDP. Interpretations are summarized below:

- **House Price Index:** The marked decline reflects real estate deflation following the housing bubble.
- **New Privately Owned Homes:** The sharp decrease signals reduced construction activity and weakened demand.
- **Home Supply:** A modest post-recession increase suggests slower turnover and potential inventory buildup.
- **Federal Funds Rate:** Post-crisis levels near zero reflect standard monetary responses to economic contraction.
- **Real Income:** A rise in income, despite economic disruption, may be attributed to stimulus measures or sectoral shifts.
- **Consumption Price Index:** Persistent increases indicate ongoing inflationary pressures.
- **GDP:** The recovery and eventual surpassing of pre-recession GDP confirms long-term macroeconomic rebound.

These patterns confirm that the macroeconomic environment experienced significant restructuring during and after the crisis, supporting the statistical findings and validating the selected predictors.

## 6 Summary

In conclusion, after training and evaluating both the Random Forest Classifier and the XGBoost Classifier on a small, well-formatted dataset of 85 data points, both models achieved a perfect AUC-ROC score of 1. While this indicates strong performance, it also suggests a potential overfitting issue due to the limited sample size.

Ultimately, the Random Forest model was selected because of its consistency with the OLS (Ordinary Least Squares) statistical model. The OLS model demonstrates that variance in the House Price Index can be explained by the federal funds rate, housing supply, newly privately owned housing units, and the treasury maturity lagged by six months. Both models in the project aim to predict the treasury maturity lagged by six months. The top three indicators in the Random Forest model are newly privately owned homes, the federal funds rate, and the House Price Index. Additionally, the Random Forest model is less sensitive to small datasets compared to XGBoost.

## 7 Future Work

This project focused on 2006 to 2012 to cover the periods leading up to and following the 2008 economic recession. The findings specifically apply to this period and event. A logical next step would be to evaluate if these same factors and conclusions apply to other recessions in U.S. history.

The project team noticed that there are certain patterns amongst the analyzed economic factors that may be predictive in nature. These patterns helped shape the hypotheses the team applied for this project. The team also noticed that the unique nature of the 2020 recession related to the COVID 19 pandemic caused significant variations in those same patterns. If there are further studies that explore the applicability of this project's findings to other recessions, interesting case studies could explore societal events like the pandemic to see how the events influence the economic factors.

Of course, any future study could also evaluate various machine learning models to compare economic factors and model effectiveness. Other models to consider could include Naive Bayes or Support Vector Machine (SVM).

Finally, this study focused on some very specific economic factors related to yield curve and housing market stresses. There are, however, a nearly infinite number of other economic factors that could be considered. Future work could explore various combinations of other factors, including macroeconomic indicators. These could include interest rates, consumer confidence, unemployment or inflation.

## Appendix A: Code References

The team utilized Python for this project. Python is a versatile programming language with numerous built-in libraries that can be applied to a variety of project purposes. To ingest, cleanse, and aggregate data, we used the pandas library. To create aesthetically pleasing and informative charts and graphs, we employed matplotlib and seaborn packages. Finally, the team leveraged the capabilities of the statsmodels and scikit-learn libraries to generate statistical data and build machine learning models.

To enhance team collaboration, we created a GitHub repository where each member could share data and code, ensuring that everyone remained updated on one another's work. This improved overall efficiency and minimized the likelihood of miscommunication.

- CS504 – Project Team 2 GitHub repository: <https://github.com/annguyenhuynh/CS504-006-Team2>
- Random Forest Classification Model: [https://github.com/annguyenhuynh/CS504-006-Team2/blob/main/random\\_forest\\_regression.ipynb](https://github.com/annguyenhuynh/CS504-006-Team2/blob/main/random_forest_regression.ipynb)
- XGBoost Classification Model: <https://github.com/annguyenhuynh/CS504-006-Team2/blob/main/NEWXGBOOST.ipynb>
- Hypothesis testing: [https://github.com/annguyenhuynh/CS504-006-Team2/blob/main/data\\_ingestion.ipynb](https://github.com/annguyenhuynh/CS504-006-Team2/blob/main/data_ingestion.ipynb) - Section 9

## Appendix B: Risk Matrix

Project Risk	Description	Probability	Impact Level	Mitigation
Loss of revision control for report and brief	With multiple Project Team members working on deliverables, it is easy to have mismatched versions.	High	Low	Work on local computers, and upload products to share folder (Teams/Data Analytics Project 2). Indicate dates on work.
Loss of data or content	May be especially challenging for large data sets.	Medium	Medium	Work on local computers, and upload products to share folder (Teams/Data Analytics Project 2).
Schedule slippage	Many factors can contribute to delays in schedule. End date is firm.	Medium	Medium	Publish team schedule. Periodic submissions of progress with weekly sprints. Address shortfalls in team meetings. Be attentive to team member-scheduling conflicts, and attempt to cover down on each other's efforts, when needed.
Project complexity inhibits determining conclusions	Project has taken on assessing a wide number of variables. There is some risk that it becomes unmanageable within schedule constraints.	Medium	Medium	Reserve opportunity to adjust scope of project as complexity is better understood. Remain engaged with customer, Dr. Baldo, especially before major scope changes.
Deliverables don't meet customer expectations	Meeting customer expectations is a top priority.	Medium	High	Regular customer engagement, including Sprint submissions and meeting invitations. Be especially attentive to his feedback.
Risk of security breach.	Using publicly accessible data. Nothing is classified. There is no reason to think other project groups will try to access or take advantage of work.	Low	Low	Accept- Risk is low. Mason Honor code.

Data Risk	Description	Probability	Impact Level	Mitigation
Data availability and quality	Missing historical data or unverified outdated data. Data granularity. Data inconsistencies: Different data sources may have different data standards, or definition of the same variables.	Low	High	Choose data from reliable sources.  Review data prior to analysis.  Standardize datasets before concatenating and storing.
Correlation versus Causation	The yield curve may be correlated with housing stress but does not cause it. Other economic factors may impact both variables. Hence, it's challenging to establish the causal link between the two variables.	Medium	High	Examine the historical case studies to understand the relationship between these 2 variables.
External influences	Federal Reserve decisions. Policy makers. Global economic factors.	Low	Low	Update data regularly as new policies are introduced or old policies are modified.
Statistical and Modeling challenges	Challenge of choosing the best model for the project. Overfitting or underfitting risks. Bad modeling results in wrong interpretation.	Medium	High	Use different metrics to measure the accuracy of the model.  Consider using regularization to minimize the noise, or cross-validation on train data.



<b>Visualization Risk</b>	<b>Description</b>	<b>Impact</b>	<b>Likelihood</b>	<b>Mitigation</b>
Misleading visuals	Wrong charts selection, charts distort scale or omit context– Leads to wrong conclusions.	<b>High</b>	<b>Medium</b>	Team members review visualizations developed by others.
Loss of context	Visuals don't have labels, legends, timeframes– cause confusion.	<b>Medium</b>	<b>Medium</b>	Ensure all visualizations have a complete summary.
Color misuse	Poor color choice– cause confusion.	<b>Low</b>	<b>High</b>	Use consistent color scheme. Team members review visualizations developed by others.
Clutter	Too much text/data in a visualization, or too many visualizations in a dashboard– key messages are lost in the noise.	<b>Medium</b>	<b>High</b>	Team members review visualizations developed by others.

## Appendix C: Agile Development

The Project 2 Team consisted of Iris Duan as Scrum Master, Brian Haggerty as Product Owner, and two Developers, Roger Graham and An Huynh.

The team utilized Agile Scrum Sprints for the development of this project. 4 sprints spanned 1-2 weeks each over an 8-week period.

YouTrack was used to record stories and track sprint progress. GitHub and Microsoft Teams was used to facilitate collaborative work on common products. GitHub was the principal repository for data and code, and Teams was used for communications and retaining files directly related to the report and the brief.

## Appendix D: Course Schedule

Dates	Milestones
Week1 (17-23 Mar)	Sprint 0 Submission due Sun 3/23 Team meeting Tue 3/18 @6:30 PM Team meeting Thu 3/20 @6:30 PM Team meeting Sat 3/22 @10:00 AM
Week 2 (24-30 Mar)	Sprint 1 Submission due Sun 3/30 Problem Definition & Project Plans Team meeting Tue 3/25 @6:30 PM Team meeting Sat 3/29 @10:00 AM
Week 3 (31 Mar-6 Apr)	Sprint 2 Submission due Sun 4/13 (wk4) Data Sets Team meeting Tue 4/1 @6:30 PM Team meeting Sat 4/5 @10:00 AM
Week 4 (7-13 Apr)	Sprint 2 Submission due Sun 4/13 Data Sets Team meeting Tue 4/8 @6:30 PM Team meeting Sat 4/12 @10:00 AM
Week 5 (14-20 Apr)	Sprint 3 Submission due Sun 4/27 (wk6) Algorithms & Analytics Team meeting Tue 4/15 @6:30 PM Team meeting Sat 4/19 @10:00 AM
Week 6 (21-27 Apr)	Sprint 3 Submission due Sun 4/27 Algorithms & Analytics Team meeting Tue 4/22 @6:30 PM Team meeting Sat 4/26 @10:00 AM
Week 7 (28 Apr-4 May)	Sprint 4 Submission due Sun 3/23 Visualizations Team meeting Tue 4/29 @6:30 PM Team meeting Sat 5/3 @10:00 AM
Week 8 (5-11 May)	Final Presentation Presentation: 5/7 @7:00 PM Course Complete 5/11

All times Eastern Time.

Standing Team Meetings Tuesdays at 6:30 PM. That meeting was used to set additional meetings in given week, as needed. Standing Team Meetings with Principal Customer representative were on Saturday mornings.

## Appendix E: Glossary of Acronyms and Terms (GOAT)

Note that not all the abbreviations or terms below are acronyms. Many are simply the nomenclature used by organizations such as the Fed to concisely represent longer titles.

- **BEA:** Bureau of Economic Analysis
- **Chained dollars:** A way to adjust for inflation that also accounts for changes in what people buy over time. It gives a more accurate picture of the economy by expressing values in the prices of a specific year, like 2017.
- **FEDFUNDS:** Federal Funds Rate. This is the official rate set by the Fed and is used to control interest rates and inflation. It is the rate for overnight interbank lending. It acts as the basis for other interest rates. Tracked and published by the Fed.
- **FHFA:** Federal Housing Finance Agency.
- **FRED:** Federal Reserve Economic Data
- **GDPC1:** Real U.S. GDP measured in billions of chained dollars, adjusted for inflation using chain-type quantity indexes. The “C” denotes chained dollars; the “1” is a FRED series identifier. The base year (e.g., 2017) is specified in the metadata and may change over time.
- **HOUST:** Housing Starts. Shows the number of new residential projects that began in a monthly period. A decline in builder activity can indicate an economic slowdown. Data from the US Census and UHUD.
- **MSACSRNSA:** Monthly Supply of Houses. Measures how many months it would take to sell all homes currently on the market at the current sales pace. Rising supply signals weakening demand and can signal price drops or recession. Data from FHFA.
- **PCEPI:** Personal Consumption Expenditure Price Index. This is the Fed's preferred measure of inflation, and it is used to influence monetary policy.
- **Real Personal Income:** Personal income in chained 2009 dollars (RPI) is personal income in current dollars (PI) deflated by the PCE chained price index (PCEPI).
- **Sectorial shift:** indications that some sectors of the economy grow more strongly after the recession. Income and GDP move from one sector to another.
- **UHUD:** U.S. Housing and Urban Development
- **USRECQ:** United States Recession Indicator. A binary indicator that shows whether the US was officially in a recession or not. The BEA provides this data.

- **USSTHPI:** U.S. All-Transactions House Price Index. Tracks changes in U.S. residential property prices. Can help identify sharp price increases which indicate unstable growth that may precede a recession. Data comes from the FHFA.
- **Yield Curve:** Measures the spread between long term and short-term interest rates on government issued bonds (10 years and 3 months for this project). It is derived from market forces and predictions. It has historically acted as a strong recession predictor.

## References

Board of Governors of the Federal Reserve System. (2024). Federal funds effective rate [FEDFUNDS]. Federal Reserve Economic Data (FRED), Federal Reserve Bank of St. Louis. <https://fred.stlouisfed.org/series/FEDFUNDS>

Bureau of Economic Analysis. (2024). Gross domestic product (GDP), real, chained dollars. U.S. Department of Commerce. <https://www.bea.gov/data/gdp/gross-domestic-product>

Bureau of Economic Analysis. (2024). Personal consumption expenditures price index [PCEPI]. Federal Reserve Economic Data (FRED), Federal Reserve Bank of St. Louis. <https://fred.stlouisfed.org/series/PCEPI>

Bureau of Economic Analysis. (2024). Real Personal Income [RPI]. Federal Reserve Economic Data (FRED), Federal Reserve Bank of St. Louis. <https://fred.stlouisfed.org/series/RPI>

Federal Housing Finance Agency. (2024a). FHFA house price index (purchase-only). <https://www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index.aspx>

Federal Housing Finance Agency. (2024b). HPI frequently asked questions. [https://www.fhfa.gov/DataTools/Downloads/Documents/HPI/HPI\\_FAQs.pdf](https://www.fhfa.gov/DataTools/Downloads/Documents/HPI/HPI_FAQs.pdf)

Federal Reserve Bank of St. Louis, 10-Year Treasury Constant Maturity Minus 2-Year Treasury Constant Maturity [T10Y2Y], retrieved from FRED, Federal Reserve Bank of St. Louis; <https://fred.stlouisfed.org/series/T10Y2Y>, April 2, 2025.

Federal Reserve Bank of St. Louis. (2024). 10-year Treasury constant maturity minus 3-month Treasury constant maturity [T10Y3M]. Federal Reserve Economic Data (FRED). <https://fred.stlouisfed.org/series/T10Y3M>

National Association of Realtors. (2024). Housing affordability index. <https://www.nar.realtor/research-and-statistics/housing-statistics/housing-affordability-index>

National Bureau of Economic Research. (2024). U.S. business cycle expansions and contractions. <https://www.nber.org/research/data/us-business-cycle-expansions-and-contractions>

Nguyen, A., Duan, I., Haggerty, B., & Graham, R. (2024). CS504-006 Team 2 project repository: Data folder [Data set]. GitHub. <https://github.com/annguyenhuynh/CS504-006-Team2/tree/main/data>

U.S. Census Bureau. (2024a). Monthly supply of new houses for sale [MSACSR]. Federal Reserve Economic Data (FRED), Federal Reserve Bank of St. Louis. <https://fred.stlouisfed.org/series/MSACSR>

U.S. Census Bureau. (2024b). New privately-owned housing units started [HOUST]. Federal Reserve Economic Data (FRED), Federal Reserve Bank of St. Louis.  
<https://fred.stlouisfed.org/series/HOUST>

U.S. Government Accountability Office. Federal Information Transparency.  
<https://www.gao.gov/federal-information-transparency#:~:text=Federal%20spending%20data.,these%20data%20remain%20top%20priorities>. Accessed April 2, 2025.

*What is a yield curve?* (n.d). Fidelity. <https://www.fidelity.com/learning-center/investment-products/fixed-income-bonds/bond-yield-curve>.