**Executive Summary**

This Housing Market Trends project aims to analyze and compare current trends and issues within the regional and national housing markets. The housing market is experiencing fluctuations due to economic shifts, demographic changes, and policy interventions, affecting housing prices, availability, and market dynamics. Key findings from the project indicate a significant disparity between the growth rates of median housing prices and household incomes, particularly evident from 2000 to 2023. This growing gap suggests potential affordability challenges that may worsen, as projected trends up to 2027 indicate continued escalation in housing prices relative to income levels. Moreover, geospatial analysis within Washington State reveals substantial regional price variations, with higher prices concentrated around urban centers like Seattle due to strong economic opportunities and amenities. These insights underscore the critical impact of location-specific factors on housing markets. The project leverages advanced predictive modeling techniques, incorporating a wide array of variables to forecast future housing market trends. This approach enhances the predictive accuracy and reliability of market forecasts, enabling stakeholders to make more informed decisions regarding real estate investments and policy developments. Future research recommendations include the adoption of segmented geospatial analysis to further refine market understanding at a micro-level and the continued use of sophisticated predictive models to navigate future market dynamics effectively. Overall, the project provides stakeholders with a deep understanding of the housing market's complexities, equipping them with the knowledge to undertake better strategic planning and policy formulation. The integration of detailed analysis with advanced data analytics ensures that decision-making is well-informed, risks are mitigated, and investment strategies are optimized, contributing to a more stable and efficient housing market.

**Table of Contents**

[Project Scope 5](#_Toc172747817)

[Data Analytics Project/Project Scope 5](#_Toc172747818)

[Problem Importance 5](#_Toc172747819)

[Research/Problem Analysis 6](#_Toc172747820)

[Critical Success Factors (CSFs) 7](#_Toc172747821)

[Key Performance Indicators 8](#_Toc172747822)

[Project Insights of your Data Analysis 9](#_Toc172747823)

[Project Milestones 10](#_Toc172747824)

[Completion History 10](#_Toc172747825)

[Lessons Learned 11](#_Toc172747826)

[Data Set Description 12](#_Toc172747827)

[High-Level Data Diagram 13](#_Toc172747828)

[Data Definition/Data Profile 14](#_Toc172747829)

[Data Preparation/Cleansing/Transformation 23](#_Toc172747830)

[Data Preparation 23](#_Toc172747831)

[Data Cleansing 23](#_Toc172747832)

[Data Transformation 24](#_Toc172747833)

[Data Analysis 25](#_Toc172747834)

[Data Visualization 27](#_Toc172747835)

[Descriptive Statistics 27](#_Toc172747836)

[Data Visualization Definitions 27](#_Toc172747837)

[Data Visualization 1 28](#_Toc172747838)

[Data Visualization 2 30](#_Toc172747839)

[Data Visualization 3 32](#_Toc172747840)

[Proposed Visualizations 34](#_Toc172747841)

[Predictive Models 36](#_Toc172747842)

[Data Modeling Definitions 36](#_Toc172747843)

[Predictive Model 1 36](#_Toc172747844)

[Predictive Model 2 40](#_Toc172747845)

[Predictive Model 3 44](#_Toc172747846)

[Predictive Model Review 47](#_Toc172747847)

[Analysis Justification 50](#_Toc172747848)

[Findings 50](#_Toc172747849)

[Review of KPIs 51](#_Toc172747850)

[Review Significance 52](#_Toc172747851)

[Recommendations for Future Analysis 53](#_Toc172747852)

[References 55](#_Toc172747853)

# 

# Project Scope

# Data Analytics Project/Project Scope

The project I am proposing aims to analyze the current trends and issues within the regional housing market and compare them to the national housing market. The housing market is experiencing considerable fluctuations due to various factors, such as economic shifts, demographic changes, and policy interventions. These fluctuations impact housing prices, availability, and overall market dynamics, which in turn affect potential homeowners, real estate investors, and policy makers. By conducting a comparative analysis, this project seeks to identify whether regional trends mirror national patterns or if there are unique regional factors at play. Understanding these trends is crucial for making informed decisions in real estate investments and policy formulation. For instance, economic shifts such as inflation rates, employment levels, and GDP growth can significantly influence both regional and national housing market.

The primary problem I am addressing is the need to identify and understand these trends and issues within the regional housing market and how they align or diverge from national trends. This involves evaluating various indicators such as median housing prices, inventory levels, and demographic factors. Additionally, the project will explore the impact of economic policies, interest rates, and employment rates on the housing market. By leveraging data analytics, I aim to uncover key factors influencing price variations and market dynamics, providing actionable insights for stakeholders. This comparative analysis will help identify potential opportunities and risks within the regional housing market in the context of broader national trends. The data analytics problem that I am analyzing is to evaluate and compare the regional housing market trends with national trends, identifying key factors influencing price variations and market dynamics. By understanding these dynamics, we can provide valuable insights for decision-makers in the real estate sector and help anticipate future market movements.

# Problem Importance

This project was selected due to the ongoing volatility in the housing market, which has significant economic and social implications. According to Forbes Advisor Rothstein, home prices are at or near all-time highs. For 2024, demand continues to outpace housing supply as many homeowners remain locked in at ultra-low mortgage rates, and mortgage rates are not decreasing either (Rothstein, 2024). Recent fluctuations in housing prices, influenced by factors such as inflation, interest rates, and post-pandemic shifts in housing demand, have made it crucial to understand both regional and national market dynamics. Analyzing these trends will provide a comprehensive overview of the housing market, helping to address the uncertainties faced by various stakeholders. This project aims to fill the knowledge gap and offer valuable insights to navigate the current housing market landscape.

The importance of the project cannot be overstated, as housing market trends have far-reaching impacts on economic stability, community development, and individual financial security.By understanding these trends, stakeholders can engage in better strategic planning and policy formulation. Additionally, identifying discrepancies between regional and national markets can highlight unique local challenges and opportunities, leading to more targeted interventions. Effective data analytics in this domain can drive better decision-making, mitigate risks, and enhance investment strategies. Consequently, the project holds the potential to foster a more stable and efficient housing market, benefiting a wide range of stakeholders. In essence, the insights gained from this project can pave the way for more informed and strategic decisions within the housing sector.

This project is crucial for real estate investors, policymakers, and potential homeowners who rely on accurate and timely information to make informed decisions. Real estate investors can use the insights to optimize their investment strategies and identify high-growth potential areas. Policymakers can benefit from the detailed analysis to craft policies that address specific regional needs and promote housing affordability and stability. Potential homeowners will gain a better understanding of the market, aiding them in making informed purchasing decisions. Moreover, the project will contribute to sustainable development goals and improved risk management practices across the housing sector. Ultimately, it provides a valuable resource for all involved parties, fostering a more informed and resilient housing market.

# 

# Research/Problem Analysis

The Zillow Home Value Index (ZHVI) dataset has been used extensively to track home value trends over time across different regions. Researchers have applied time series analysis, machine learning models, and statistical methods to forecast future prices, identify influential factors, and assess market stability. For example, a study by Gupta and Walstrum (2023) utilized the ZHVI dataset to analyze the impact of economic indicators on housing prices. They implemented various machine learning algorithms to predict price trends. They found that from January 2020 to August 2023, typical U.S home increases by 41% and 19% even after adjusting for inflation. Additionally, their study highlighted a positive correlation between metro area population and home price within U.S regions. This analysis underscores the significance of demographic factors in understanding housing market dynamics.

The Housing Price dataset by Sukhmandeepsingh Brar has been used for predictive modeling and trend analysis. Ezzat (2024) analyzed this dataset to identify key determinants of housing prices, using Python notebook to quantify the influence of various attributes such as number of bedrooms, bathrooms, grade, and view. This study demonstrated how individual property characteristic impact overall housing prices, providing a granular understanding of price determinants. Both of these studies contribute valuable insights into housing market dynamics, leveraging data analytics to inform stakeholders. However, there remain gaps in the research that need addressing, particularly the integration of regional and national trends, the impact of external shocks, and the use of advanced predictive techniques.

As mentioned, while these studies have provided valuable insights into housing market dynamics, there are still gaps that need addressing. Firstly, most existing studies focus either on regional or national trends independently, with limited comprehensive analysis integrating both levels to compare and contrast their dynamics. Secondly, there is a need for more detailed analysis of how external shocks, such as pandemics or economic crises, differently impact regional versus national housing markets. Lastly, although machine learning models have been applied, there is scope for utilizing more advanced techniques, such as deep learning and ensemble models, to improve prediction accuracy. By combining the Housing Price Dataset, the ZHVI dataset, and economic indicators from the Federal Reserve Economic Data (FRED), this project aims to fill these gaps by providing an integrated analysis of regional and national housing market trends using advanced predictive analytics. By doing so, it will offer a more nuanced understanding of the factors driving housing prices and the differential impacts of external shocks.

### Critical Success Factors (CSFs)

Critical Success Factors (CSFs) are the essential elements that must be in place for a project to achieve its goals. They help identify what is necessary to reach desired outcomes and ensure project success. According to Rockart (1979), CSFs are "the limited number of areas in which results, if they are satisfactory, will ensure successful competitive performance for the organization." These factors focus on key areas that must go right for the project to succeed. For the housing market analysis project, the desired outcome is to provide accurate and insightful analysis of housing market trends. To achieve this, several critical success factors must be identified and met.

The first CSE is the availability of comprehensive and high-quality data. Access to reliable and up-to-date datasets covering both regional and national housing market is crucial. The data must include a wide range of variables, such as economic indicators, demographic data, and housing prices. This ensures that the analysis captures all relevant factors influencing the housing market. Additionally, using advance analytical techniques, such as machine learning and statistical methods, is essential to extract meaningful insights from the data. High-quality data and robust analytical method will enable the project to deliver accurate and valuable results.The second CSF is the availability of appropriate tools and technologies. Data acquisition tools, databases, and data cleaning software are necessary to ensure data quality and facilitate efficient data processing.These tools help in collecting, storing, and preparing data for analysis. Additionally, having access to advanced analytical software, such as Python's scikit-learn, SAS Enterprise Miner, and visualization tools like Tableau, will enable the team to perform complex analyses and present findings effectively. The right set of tools and technologies is crucial for achieving the project's goals and ensuring its success. The third CSF is the expertise and skills. One must possess strong skills in data collection, cleaning, and preprocessing to prepare the data for analysis. Expertise in advanced analytical techniques, such as machine learning, deep learning, and statistical modeling, is also essential. Additionally, skills in data visualization and interpretation are necessary to effectively communicate the findings to stakeholders.

# 

# Key Performance Indicators

Key Performance Indicators (KPIs) are measurable target used to evaluate the success of an organization, employee, or project in meeting objectives for performance. According to Hennigan (2023), KPIs are important tool to evaluate achievements, analyze issues and solve problems, thereby benefiting the overall health of an organization. For the housing market analysis project, identifying relevant KPIs is crucial to ensure the project is on track and its outcomes are beneficial. This section will discuss three indicators related to the completion of the project, the benefits of the analysis outcomes, and the overall impact on the stakeholders.

The first indicator is achieving 100% of project milestones within the planned timeframe. This is a crucial KPI for measuring the efficiency and effectiveness of project management. This indicator tracks whether all key tasks and phases of the project, such as data collection, data cleaning, model development, and analysis reporting, are completed as scheduled. Maintaining adherence to the timeline ensures that the project progresses smoothly, resources are utilized efficiently, and potential delays are minimized. Regular progress reviews and milestone checks can help identify any deviations early, allowing for timely corrective actions. Meeting 100% of project milestones on time demonstrates the project’s ability to stay on track and adhere to its planned schedule. Therefore, this KPI is essential for ensuring that the project is delivered as planned, contributing to its overall success.

The second indicator is the accuracy of the predictive models in generating reliable and actionable insights. The effectiveness of the data analysis in generating reliable and actionable insights is another critical KPI. This can be measured by comparing predicted housing prices to actual prices using metrics such as Mean Absolute Error (MAE) or Root Mean Square Error (RMSE). High accuracy in predictions of higher than 75% indicates that the models developed are robust and capable of providing valuable forecasts. This KPI ensures that the analytical techniques and models used in the project are effective in capturing the complexities of the housing market. By achieving low error rates such as less than 25%, the analysis demonstrates its reliability and usefulness in real-world applications. Therefore, maintaining high accuracy in predictive models is essential for the success of this project.

The third indicator is the stakeholder perception of the analysis’s value and impact on decision-making. Assessing the perceived value and impact of the analysis on stakeholders' decision-making processes is a vital KPI. This can be measured by evaluating how the insights and forecasts generated from the analysis influence the stakeholders’ decision-making processes and long-term strategies, including real estate investors, policymakers, and potential homeowners. Positive stakeholder feedback and adoption of new strategies based on the analysis could indicate that the analysis has effectively addressed their needs and contributed to more informed decisions. This KPI helps in understanding the practical benefits of the project outcomes and how well they align with stakeholder expectations. It also provides insights into areas for improvement, ensuring that future analyses can be even more targeted and effective. Thus, assessing the stakeholder perception and benefit is a critical measure of the project's overall effectiveness and relevance.

# 

# Project Insights of your Data Analysis

The analysis is expected to uncover significant trends in housing prices at both regional and national levels. This includes identifying periods of rapid price increases or decreases, seasonal variations, and long-term growth or decline patterns. By comparing regional and national housing markets, the project will highlight both similarities and differences, providing a comprehensive understanding of the market dynamics. This comparative analysis will shed light on how local factors might cause regional deviations from national trends, thereby providing a more nuanced view of the housing market. The predictive models developed are anticipated to demonstrate high accuracy in forecasting housing prices, offering reliable tools for stakeholders. Furthermore, the project aims to identify key factors that significantly influence housing prices, such as economic indicators of median household income. These findings will be integral in understanding the driving forces behind housing price fluctuations.

The project will provide detailed insights into the housing market, helping stakeholders understand current conditions and future prospects. By comprehending the factors driving price changes, stakeholders can better anticipate and respond to market fluctuations. These insights will be crucial for making informed decisions regarding investments, policy-making, and market strategies. For real estate investors, the analysis will help optimize investment strategies by identifying high-growth potential areas. Policymakers will benefit from detailed data to craft policies that address specific regional needs and promote housing affordability and stability. Potential homeowners will gain a better understanding of the market, aiding them in making informed purchasing decisions. Overall, the project aims to contribute to sustainable development goals and improved risk management practices across the housing sector, providing a valuable resource for all involved parties.

# Data Set Description

For this project, I will utilize three comprehensive data sets that provide extensive information on housing prices, home value trends, and economic indicators. The primary dataset is the Housing Price Dataset from Kaggle, which offers a rich collection of property listings spanning over 21600 rows and encompassing 21 attributes (Brar, 2024). It includes crucial information such as the number of bedrooms, bathrooms, living area size, lot size, and zip codes, providing a comprehensive insight into sold housing in Washington State during 2014 and 2015. This dataset is essential for our project as it serves as the primary source for predicting regional housing prices. The detailed attributes allow for thorough exploratory data analysis (EDA) and the development of robust predictive models. Furthermore, the quantitative dataset's diversity enables effective feature engineering and understanding of the intricate relationships between various variables and the target variable, which is the price (Calzon, 2024).

Complementing this dataset is the Zillow Home Value Index (ZHVI) dataset, which is collected and aggregated by Zillow and sourced from Kaggle, containing over 292 rows and 52 columns (Mulla, 2024). This quantitative dataset tracks the monthly median home value in each state from 2000 to the present, which is particularly useful for time-series analysis. Its extensive temporal coverage and granularity make it invaluable for analyzing real estate trends over time and across different locations. Utilizing this dataset alongside the Housing Price Dataset enriches our analysis by providing insights into broader market dynamics, facilitating the forecasting of future prices, and aiding in making informed investment decisions. By merging the ZHVI dataset with the Housing Price Dataset, we can analyze how historical home value trends influence current housing prices in specific regions, enhancing our predictive models’ robustness.

The third dataset comes from the Federal Reserve Economic Data (FRED) system, which provides median household income data across various states. This dataset includes 24 rows and 13 columns with multiple series such as MEHOINUSA672N (National), MEHOIUSCAA672N (California), and others, represents 12 states covering data points from 2000 to present (Federal Reserve Bank of St. Louis, 2024). Each series provides annual median household income, offering a vital economic context for our housing price analysis. The integration of this quantitative dataset with the Housing Price Dataset and ZHVI will allow us to examine the relationship between household income and housing prices, providing a more comprehensive understanding of the economic factors driving the housing market. This holistic approach ensures that our analysis is not only predictive but also explanatory, providing valuable insights into the factors influencing housing prices (Calzon, 2024).

The process of combining these datasets involves a multi-step process to ensure seamless integration and comprehensive analysis. For instance, we will align the datasets based on common attributes such as geographical identifiers (e.g., zip codes, state names) and time periods. This alignment will enable us to merge the datasets accurately, ensuring that each row in the combined dataset represents a coherent data point with information from all three sources. Next, we will handle any missing data through imputation techniques or exclusion, depending on the analysis requirements. Ensuring data consistency and completeness is crucial for developing accurate predictive models. Finally, we will perform feature engineering to create new variables that capture the relationships between housing prices, home value trends, and economic indicators. This process will enhance our models’ explanatory power and predictive accuracy, providing a comprehensive understanding of the factors influencing housing prices.

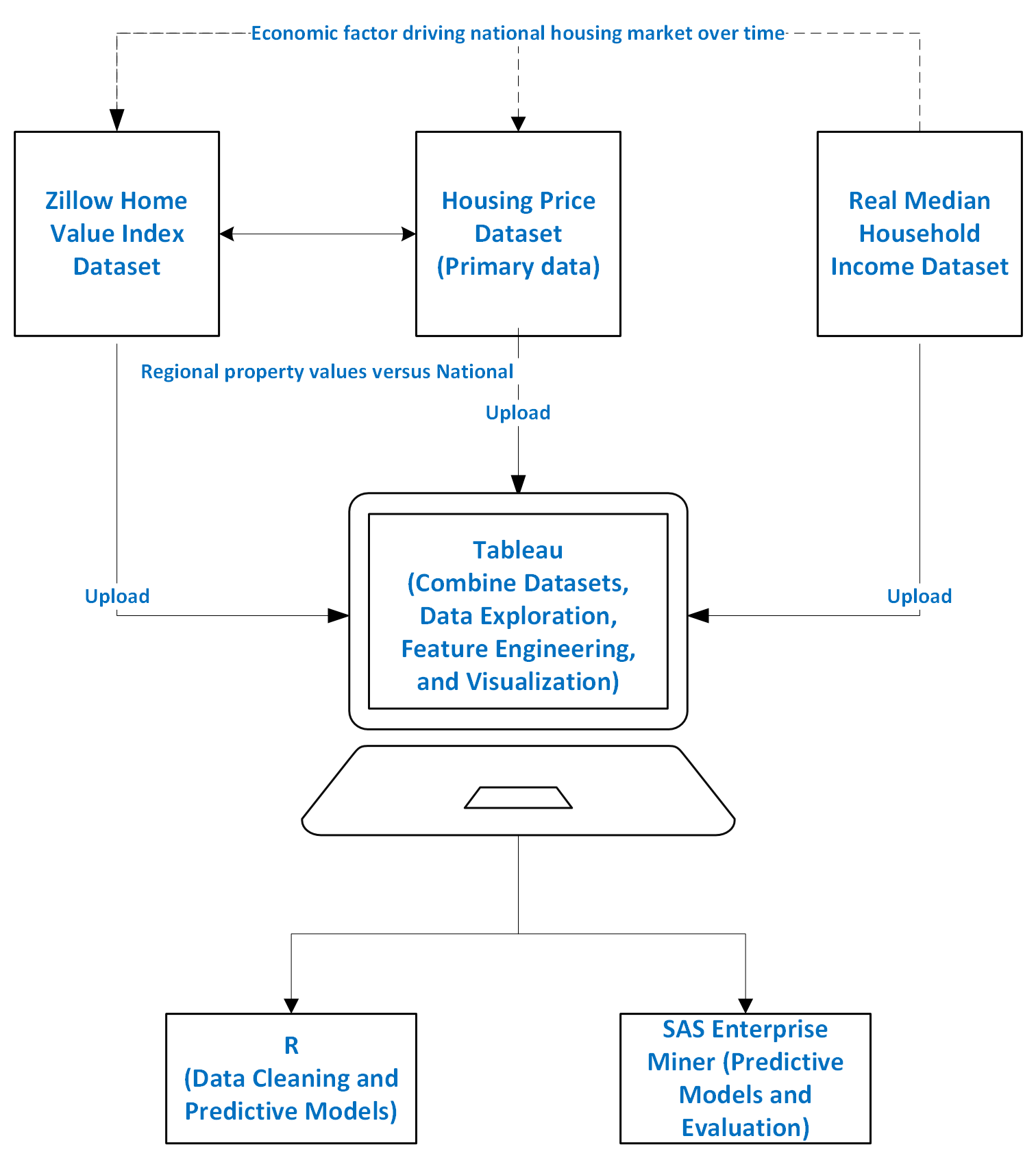
### High-Level Data Diagram

The high-level data diagram illustrates the relationships between three primary datasets used in the analysis: the Housing Price Dataset from Kaggle, the Zillow Home Value Index (ZHVI) dataset, and the Median Household Income Data from the Federal Reserve Bank of St. Louis (FRED). The diagram shows the integration points between these datasets. The housing price data can be directly correlated with the ZHVI data to assess how individual property values compare to broader market trends. Similarly, the median household income data provides an economic backdrop that can help explain variations in housing prices and values across different regions and time periods.

By combining these datasets, we can conduct a comprehensive analysis that includes:

predictive modeling of housing prices using property attributes, time-series analysis of home value trends and their economic drivers, and comparative studies of regional housing markets and the impact of income levels on housing affordability. This integrated approach addresses gaps in existing research by providing a more holistic view of the housing market, incorporating both micro-level property data and macro-level economic indicators.

Analyzing multiple datasets involves a comprehensive approach that combines data preprocessing, exploratory data analysis (EDA), feature engineering, predictive modeling, and visualization. Step 1 involves data preprocessing, which includes data combining and cleaning to check for missing values and handle them appropriately, identifying and rectifying any inconsistencies or errors, and standardizing categorical variables. Data integration ensures the temporal and geographic dimensions align properly when merging datasets. Data transformation converts date columns to datetime objects and normalizes numerical features. Step 2 involves EDA, which includes calculating descriptive statistics, visualizing data distributions, performing geospatial and temporal analyses, and identifying relationships between features. Step 3 involves feature engineering to create new features and generate interaction terms. These three steps can be done using Tableau. Step 4 involves predictive modeling, where appropriate models are selected, trained, and evaluated using metrics like MAE, MSE, RMSE, and R-squared, with steps for hyperparameter tuning and cross-validation to ensure model robustness and generalization. Step 4 can be done using both R and SAS Enterprise Miner to create multiple predictive and machine learning models. Below is a graphical representation of the relationships between the datasets and approach to comprehensive analysis:

**

**Figure 1. High Level Diagram with Multiple Datasets**

### 

### Data Definition/Data Profile

**Housing Price Dataset**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Definition** | **Data Type** | **Outliers** | **Frequency of Nulls** | **Potential Quality Issues** |
| ID | Unique identifier for each house | Integer | No | None | None |
| Date | Date when the house was sold | Date | No | None | None |
| Price | Sale price of the house | Integer | Yes | None | Potential outliers in extreme values |
| Bedrooms | Number of bedrooms | Integer | Yes | None | Possible misentries (e.g., extremely high values) |
| Bathrooms | Number of bathrooms | Float | Yes | None | Possible misentries or potential outliers in extreme values or unrealistic value of 0 |
| Sqft\_living | Square footage of the living space | Integer | Yes | None | Outliers in very high or low square footage |
| Sqft\_lot | Square footage of the lot | Integer | Yes | None | Outliers in very high or low square footage |
| Floors | Number of floors | Float | None | None | None |
| Waterfront | Whether the house is waterfront | Integer | No | None | Binary value should be checked for validity |
| View | Quality of the view from the house | Integer | Yes | None | Range check needed (should be 0-4) |
| Condition | Condition of the house | Integer | Yes | None | Range check needed (should be 1-5) |
| Grade | Construction grade | Integer | Yes | None | Range check needed (should be 1-13) |
| Sqft\_above | Square footage above ground | Integer | Yes | None | Should not exceed sqft\_living |
| Sqft\_basement | Square footage of the basement | Integer | Yes | None | Check consistency with sqft\_living and sqft\_above |
| Yr\_built | Year the house was built | Integer | Yes | None | Unrealistic years (e.g., in the future) |
| Yr\_renovated | Year the house was renovated | Integer | Yes | Many | Missing or 0 values for non-renovated house |
| Zipcode | Zip code where the house is located | Integer | No | None | None |
| Lat | Latitude of the house | Float | No | None | Geographic validation needed |
| Long | Longitude of the house | Float | No | None | Geographic validation needed |
| Sqft\_living15 | Living area size of 15 nearest properties in square feet | Integer | Yes | None | Consistency check with sqft\_living |
| Sqft\_lot15 | Lot size of 15 nearest properties in square feet | Integer | Yes | None | Consistency check with sqft\_lot |

**Zillow Home Index Value Dataset**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Definition** | **Data Type** | **Outliers** | **Frequency of Nulls** | **Potential Quality Issues** |
| The monthn | Date each month | Date | None | None | None |
| Virginia | Mean home value for the region | Float | None | None | None |
| California | Mean home value for the region | Float | None | None | None |
| Florida | Mean home value for the region | Float | None | None | None |
| New York | Mean home value for the region | Float | None | None | None |
| New Jersey | Mean home value for the region | Float | None | None | None |
| Texas | Mean home value for the region | Float | None | None | None |
| Michigan | Mean home value for the region | Float | None | None | None |
| Massachusetts | Mean home value for the region | Float | None | None | None |
| Arizona | Mean home value for the region | Float | None | Yes | One missing value that needs to be addressed |
| Washington | Mean home value for the region | Float | None | None | None |
| Colorado | Mean home value for the region | Float | None | None | None |
| Illinois | Mean home value for the region | Float | None | None | None |
| District of Columbia | Mean home value for the region | Float | None | None | None |
| Nevada | Mean home value for the region | Float | None | None | None |
| Hawaii | Mean home value for the region | Float | None | None | None |
| New Hampshire | Mean home value for the region | Float | None | None | None |
| Utah | Mean home value for the region | Float | None | None | None |
| Georgia | Mean home value for the region | Float | None | None | None |
| Montana | Mean home value for the region | Float | None | Yes | Many missing value that needs to be addressed |
| Minnesota | Mean home value for the region | Float | None | None | None |
| Louisiana | Mean home value for the region | Float | None | None | None |
| Maryland | Mean home value for the region | Float | None | None | None |
| Pennsylvania | Mean home value for the region | Float | None | None | None |
| South Carolina | Mean home value for the region | Float | None | None | None |
| North Carolina | Mean home value for the region | Float | None | None | None |
| Vermont | Mean home value for the region | Float | None | None | None |
| Tennessee | Mean home value for the region | Float | None | None | None |
| Oregon | Mean home value for the region | Float | None | None | None |
| New Mexico | Mean home value for the region | Float | None | Yes | Many missing value that needs to be addressed |
| Rhode Island | Mean home value for the region | Float | None | None | None |
| Alaska | Mean home value for the region | Float | None | Yes | One missing value that needs to be addressed |
| Maine | Mean home value for the region | Float | None | None | None |
| Alabama | Mean home value for the region | Float | None | None | None |
| Wisconsin | Mean home value for the region | Float | None | None | None |
| Arkansas | Mean home value for the region | Float | None | None | None |
| Mississippi | Mean home value for the region | Float | None | None | None |
| Indiana | Mean home value for the region | Float | None | None | None |
| West Virginia | Mean home value for the region | Float | None | Yes | One missing value that needs to be addressed |
| Idaho | Mean home value for the region | Float | None | Yes | One missing value that needs to be addressed |
| North Dakota | Mean home value for the region | Float | None | Yes | Many missing value that needs to be addressed |
| Connecticut | Mean home value for the region | Float | None | None | None |
| Kentucky | Mean home value for the region | Float | None | None | None |
| Missouri | Mean home value for the region | Float | None | None | None |
| Kansas | Mean home value for the region | Float | None | None | None |
| Delaware | Mean home value for the region | Float | None | None | None |
| Wyoming | Mean home value for the region | Float | None | Yes | Many missing value that needs to be addressed |
| Oklahoma | Mean home value for the region | Float | None | None | None |
| South Dakota | Mean home value for the region | Float | None | Yes | One missing value that needs to be addressed |
| Nebraska | Mean home value for the region | Float | None | None | None |
| Iowa | Mean home value for the region | Float | None | None | None |
| Ohio | Mean home value for the region | Float | None | None | None |

**Real Median Household Income Dataset**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Definition** | **Data Type** | **Outliers** | **Frequency of Nulls** | **Potential Quality Issues** |
| Date | The date when the data was recorded | Date | None | None | None |
| MEHOINUSA672N | Median household income in the USA | Integer | None | None | None |
| MEHOINUSCAA672N | Median household income in California | Integer | None | None | None |
| MEHOINUSCOA672N | Median household income in Colorado | Integer | None | None | None |
| MEHOINUSDCA672N | Median household income in District of Columbia | Integer | None | None | None |
| MEHOINUSFLA672N | Median household income in Florida | Integer | None | None | None |
| MEHOINUSGAA672N | Median household income in Georgia | Integer | None | None | None |
| MEHOINUSHIA672N | Median household income in Hawaii | Integer | None | None | None |
| MEHOINUSMDA672N | Median household income in Maryland | Integer | None | None | None |
| MEHOINUSMAA672N | Median household income in Massachusetts | Integer | None | None | None |
| MEHOINUSNYA672N | Median household income in New York | Integer | None | None | None |
| MEHOINUSVAA672N | Median household income in Virginia | Integer | None | None | None |
| MEHOINUSWAA672N | Median household income in Washington State | Integer | None | None | None |

# Data Preparation/Cleansing/Transformation

### Data Preparation

The initial data preparation process involves using R for its powerful data manipulation and cleaning capabilities. This begins with loading the datasets into R and performing an initial inspection to understand the structure and content of the data. By utilizing functions like ‘read.csv()’ and ‘str()’, we can gain insights into the data types and identify any immediate issues such as missing values or inconsistencies. Handling missing values is a critical step, which can be managed using functions like ‘na.omit()’ to remove rows with missing data or ‘impute()’ to fill in missing values based on statistical methods like mean or median imputation. Additionally, ensuring the correct data types for each field is essential; this can be done by converting columns using functions like ‘as.Date()’ for dates or ‘as.numeric()’ for numerical data. This initial phase lays the groundwork for a robust dataset, free from glaring issues that could skew further analysis.

To handle inconsistencies, the next step involves standardizing data formats across the dataset. This includes ensuring that all data fields follow a consistent format, such as ‘YYYY-MM-DD’, which can be done using the ‘as.Date()’ function with the appropriate format parameter. However, before using R to prepare the data, Excel is utilized to label the variable such as state names instead of state code for the median income data, and also the date data in housing data set is altered by removing ‘T000000’ from the date values, such as from ‘20141013T000000’ to ‘20141013’. Similarly, categorical variables need to be standardized to ensure consistency in naming conventions and values, which can be achieved using functions like ‘factor()’ or ‘gsub()’ for string replacements. Filters are applied to remove outliers or incorrect values, ensuring that only valid and relevant data points are included in the analysis. This step is crucial for maintaining data integrity and reliability, allowing for accurate and meaningful analysis.

The final step in data preparation involves integrating multiple datasets and performing necessary data transformations using Tableau. This can include merging datasets through intuitive joins and calculated fields ensures seamless alignment of temporal and geographical dimensions. Data transformations such as creating new calculated fields (e.g., price per square foot) can be performed by entering the formula in the dialog box. Additionally, normalizing numerical features to a common scale ensures that all variables contribute equally to the analysis, which can also be achieved using calculated fields. By combining these steps, we ensure that the datasets are comprehensive, consistent, and ready for in-depth analysis.

### Data Cleansing

Data cleansing starts with addressing missing values and outliers, which can significantly impact the quality of the analysis. In R, missing values are identified using the ‘is.na()’ function, and can be handled through deletion ‘na.omit()’ or imputation (‘mean()’, ‘median()’, or advanced methods like KNN imputation). Among three datasets, only the ZHVI dataset has missing values. Thus, the mean imputation was used to fill in missing values so no rows would be omitted. Outliers are detected using statistical techniques such as the IQR method or z-scores and can be removed or adjusted accordingly. For instance, the ‘boxplot.stats()’ function can be used to identify and remove extreme values. This step ensures that the dataset is free from anomalies that could distort the analysis.

Ensuring data consistency involves correcting any discrepancies in the dataset. This includes standardizing categorical variables to maintain uniformity across the data. In R, this can be done using the ‘factor() function to convert variables into factors with consistent levels. In Tableau, this can be done using create group feature. Additionally, any inconsistencies in data formats or numerical scales are corrected with using appropriate functions ‘as.Date()’ for dates and ‘scale()’ for numerical data. This step is crucial for maintaining the accuracy and reliability of the dataset, ensuring that all values are consistent and correctly formatted.

Further data cleansing is performed in Tableau through visual inspection and refinement of the data. Tableau’s visualization tools help in spotting trends and outliers, allowing for interactive filtering and adjustment of the data. For example, using Tableau’s group feature mentioned above to combine categories to standardize categorical variables. Calculated fields can also be created for normalization, ensuring that all numerical values fall within a specific range. This visual and interactive approach to data cleansing ensures that the datasets are not only clean but also optimized for analysis and visualization. Overall, combining the strengths of R and Tableau provides a comprehensive approach to data preparation and cleansing, ensuring that the datasets are robust, reliable, and well-prepared for the subsequent stages of analysis and modeling, ultimately enhancing the overall quality and accuracy of the project.

### Data Transformation

In the context of our analysis, creating additional features can significantly enhance the predictive power and insights derived from the dataset. One important feature we can create is the price per square foot. This feature is derived by dividing the property price by the living area (square footage). This metric is crucial as it provides a standardized way to compare property value across different sizes and types of homes. It can help identify overvalued or undervalued properties and highlight trends in real estate pricing within different regions. By incorporating this feature, we can gain deeper insights into market dynamics and price variations, which are essential for accurate predictions and meaningful analysis.

Another valuable feature to add is the seasonality index, which captures the impact of seasonal variations on property prices and sales volumes. This feature can be derived from the date of property listing and can categorize data into seasons (e.g., winter, spring, summer, and fall) or specific months. Seasonal trends can significantly influence real estate markets, with certain times of the year typically seeing higher activity and price fluctuations. By including a seasonality index, we can better understand these patterns and adjust our models accordingly to improve prediction accuracy. This feature is particularly important for forecasting purposes, as it allows the models to account for temporal variations that could impact future market behavior.

The price per square foot feature is particularly important as it standardizes property values, allowing for a more nuanced comparison across different properties. This is critical in real estate analysis where property sizes and types vary widely. By having a standardized metric, we can identify trends and outliers more effectively. It also aids in the valuation of properties, providing a clearer picture of market conditions. This feature supports both exploratory data analysis and predictive modeling, making it a foundational element for understanding property values within the dataset. The seasonality index is essential for capturing the temporal dynamics of the real estate market. Seasonal trends can greatly impact market activity, with certain periods experiencing higher buying and selling activity. By incorporating this feature, we can better model and predict these fluctuations, leading to more accurate and reliable forecasts. This is particularly important for stakeholders who rely on timing to make informed decisions about buying, selling, or investing in real estate. The seasonality index thus adds a layer of temporal granularity to the analysis, enhancing the overall depth and accuracy of our insights.

### Data Analysis

For the analysis, I will be using Tableau, R, and SAS Enterprise Miner. Each of these tools has distinct strengths that make them suitable for different aspects of data analysis, from visualization to predictive modeling. Tableau’s user-friendly interface allows for creating interactive and visually appealing dashboards, while R and SAS Enterprise Miner are powerful for statistical analysis and predictive modeling. This combination leverages the strengths of all tools, ensuring thorough data analysis, accurate modeling, and effective communication of insights. These tools collectively enable comprehensive data exploration, analysis, and presentation. By leveraging the strengths of each, I can ensure a robust approach to data analysis, from initial exploration to final reporting.

Tableau is an excellent tool for creating interactive visualizations and dashboard without requiring extensive programming skills. Its drag-and-drop functionality facilitates quick visual exploration and analysis of data. According to Katerova (2023), Tableau’s ability to create a wide range of visualizations, including bar charts, line graphs, scatter plots, heat maps, and geographic maps, makes it a versatile tool for data visualization. Moreover, the creation of interactive dashboards that can be shared with stakeholders enhances data-driven decision-making. The user-friendly interface and powerful analytics capabilities of Tableau make it an ideal tool for presenting complex data in an accessible and engaging manner. These combinations provide a robust and comprehensive approach to explore and interpreting data.

RStudio is a powerful tool for statistical analysis, offering a wide range of packages for performing complex data manipulation, statistical modeling, and visualization. The large and active community around R provides extensive resources and support for troubleshooting and learning, making it a reliable choice for data scientists (Hair Jr. et. al, 2021). Additionally, according to Jain (2020), R is particularly well-suited for data science due to its flexibility and the breadth of statistical techniques it supports. Additionally, R can easily integrate with databases, web services, and other software tools, making it versatile for various data analysis workflows. This flexibility and integration capability ensure that R can handle diverse data analysis tasks effectively.

On the other hand, SAS Enterprise Miner is a user-friendly tool that allows for the creation of complex models without extensive programming knowledge. The software offers huge array of statistical functions with great GUI interface for people to learn quickly (Jain, 2020). This makes it accessible to users with varying levels of technical expertise. The software is particularly strong in predictive modeling and machine learning, providing a wide range of algorithms and techniques. According to Hall et. al. (2014), SAS Enterprise Miner excels in model accuracy, scalability, and evaluation, making it suitable for large datasets and complex analytical tasks. The comprehensive features of SAS Enterprise Miner make it a valuable tool for predictive modeling in this project.

# Data Visualization

### Descriptive Statistics

The datasets include several fields/variables that are practical for analysis in terms of their descriptive statistics. Key variables include median housing prices, inventory levels, and median household income. Median housing prices measure the central tendency of housing prices within a region. The mean and standard deviation provide insights into the average price and the variability around this average, respectively. Quantiles help understand the distribution of prices and identify any skewness. Inventory levels indicate the number of houses available for sale. The mean, standard deviation, and quantiles describe the average availability and its dispersion. Median housing income indicates central value of household income in a region across nation. The standard deviation and quantiles provide insights into the income distribution and variation.

### 

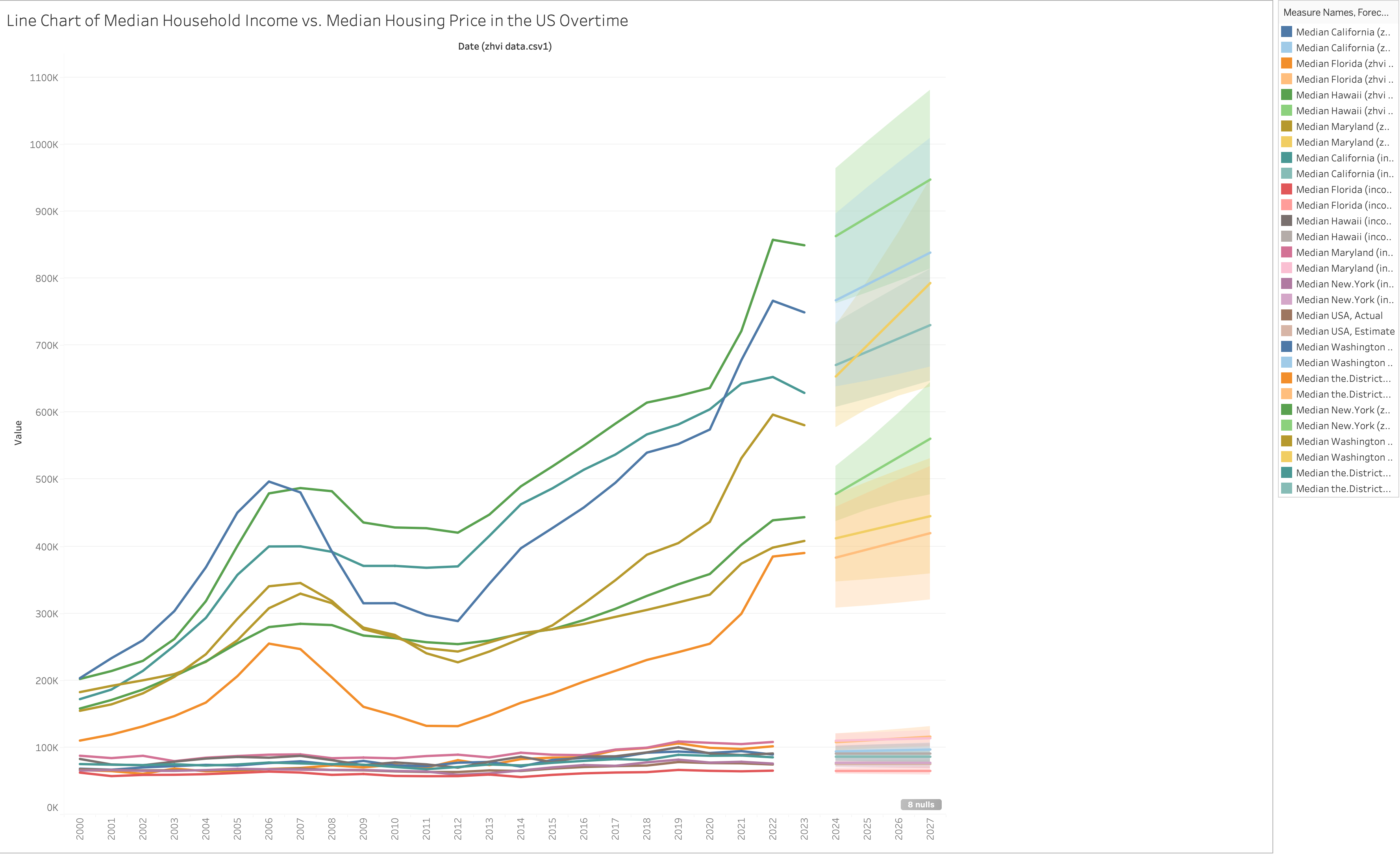
### Data Visualization Definitions

For the visualization analysis, line charts and geospatial analysis map were used. Line charts are an essential visualization technique for showing trends over time. They consist of a series of data points called 'markers' connected by straight line segments. This type of chart is particularly useful for identifying patterns and changes in data over periods (Reimers & Harvey, 2024). For example, in the context of housing market analysis, line charts can effectively illustrate the fluctuations in median housing prices over several years, providing a clear visual representation of trends. Line charts can also compare multiple datasets simultaneously by using different colored lines for each dataset. This makes it easy to see how different regions' housing markets are performing relative to each other. The simplicity and clarity of line charts make them a preferred choice for presenting time series data in a comprehensible manner. In addition to their straightforwardness, line charts can incorporate various statistical annotations, such as trend lines or forecast, to enhance their informational and predicting value. These additional features help in understanding the underlying trends and predicting future movements. Thus, line charts are not only versatile but also highly effective for a broad range of applications, from simple data exploration to detailed statistical analysis.

Geospatial analysis maps are powerful visualization tools that provide spatial context to data, enabling the identification of geographical patterns and trends (Ahasan et. al, 2022). These maps are created by overlaying data onto geographical maps, which can highlight variations across different regions. By visualizing data geographically, stakeholders can gain insights into location-specific factors that may influence market dynamics. This type of visualization is particularly effective in identifying clusters or hotspots where certain trends are more pronounced. For instance, a geospatial map can reveal areas with rapid price increases or significant inventory shortages, guiding strategic decisions for real estate investments or policy interventions. By leveraging geospatial analysis maps, researchers and analysts can uncover hidden patterns and relationships that are not immediately apparent in tabular data, making them an invaluable tool for both exploratory and explanatory data analysis.

### 

### Data Visualization 1



The first visualization is a line chart presents the median household income versus the median housing price in the US over time. Different colored lines represent different regions showcasing how both metrics have evolved from the year 2000 to 2023. The x-axis indicates the years, while the y-axis represents the value, which spans from 0 to 900K. The lower color lines represent the median household income while the higher color lines represent the median housing price. The chart includes different colored lines to differentiate between various regions, making it easier to compare trends. This visualization is crucial for understanding the relationship between household income and housing prices, offering a clear visual comparison over the selected time frame.

The inclusion of future projections up to 2027 adds an additional layer of analysis, providing insights into expected trends based on current data. By observing the chart, one can quickly identify periods of rapid growth, stability, or decline in both household income and housing prices. The shaded areas around each line indicate potential confidence intervals or variation within the forecasting data from 2024 to 2027. The color-coded lines help distinguish between the different regions, ensuring clarity in the presented information. It also highlights the disparities between states, with some states experiencing more significant increases in housing prices compared to others. Overall, this line chart is a powerful tool for visualizing complex time series data, making it easier to grasp significant patterns and trends over a lengthy period.

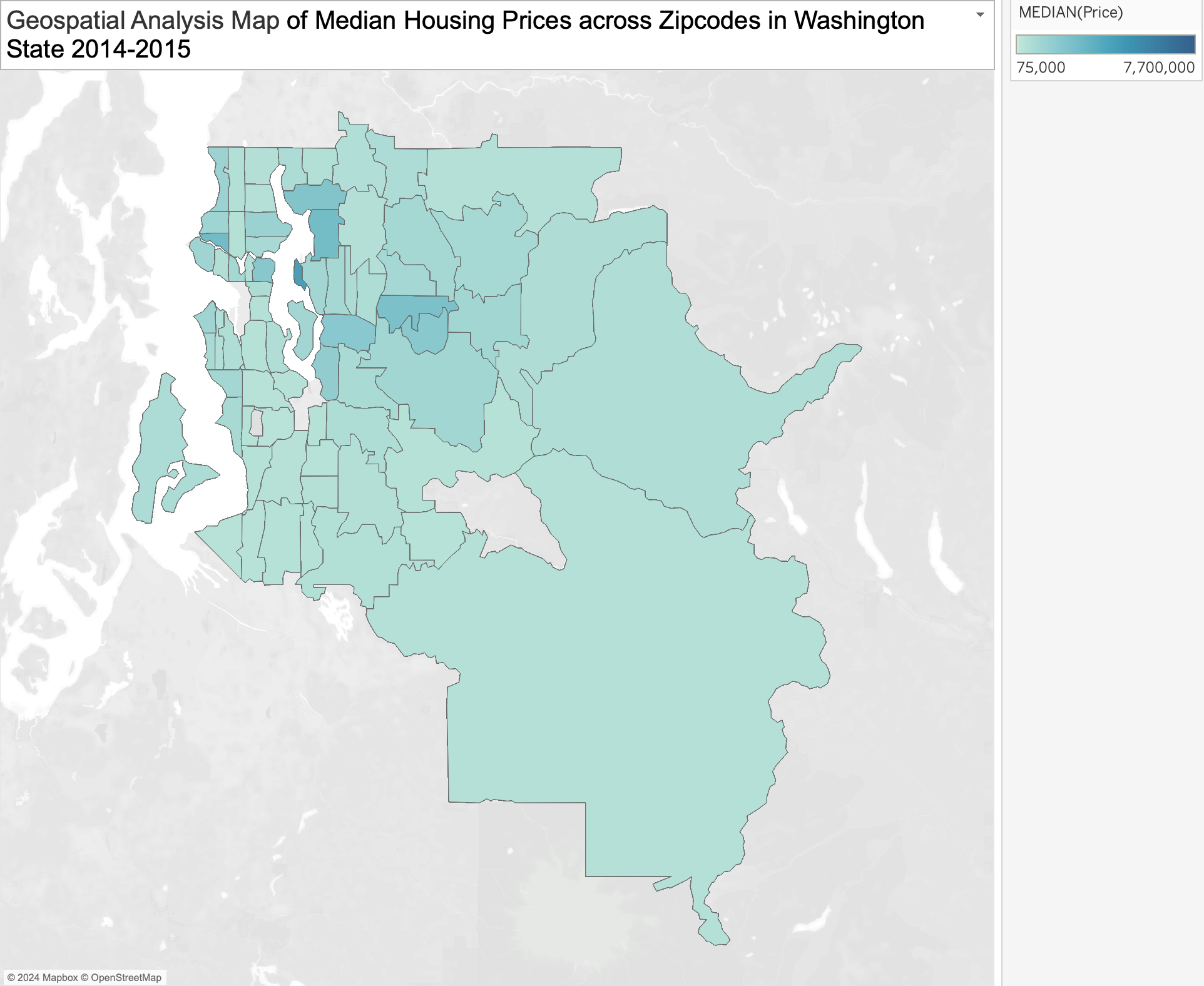
From this visualization, several key insights emerge. Firstly, it is evident that median housing prices have generally been on an upward trend, especially since the mid-2010s after recovering from a downward trend in 2008 crisis. This trend indicates a significant increase in housing market values across various regions, suggesting a robust growth phase. Some of the fastest growing regions are Hawaii, California, and Washington. In contrast, the median household income lines appear flatter, indicating a more modest growth rate over the same period. This disparity suggests that housing prices have been rising faster than household incomes, which could imply affordability issues for potential homebuyers.

Additionally, the shaded confidence intervals reveal the degree of uncertainty or variability within each dataset. Regions with broader shaded areas indicate higher volatility in housing prices or incomes, whereas narrower shaded areas suggest more stability. The visualization also highlights distinct periods where housing prices either surged or dipped, potentially corresponding with economic events such as the 2008 financial crisis or recent economic policies during and after the pandemic. The forecasting trends also help identify potential regions with higher median household income increase and lower median housing price increase, such as Maryland or District of Columbia. These insights can help policymakers, real estate investors, and potential homeowners make more informed decisions based on historical and projected data trends.

This visualization alters the scope of expected results by highlighting the pronounced disparity between the growth rates of housing prices and household incomes. Initially, the expectation might have been that both metrics would grow at relatively comparable rates, maintaining a stable affordability ratio. However, the chart clearly shows that housing prices have escalated much faster than incomes, suggesting a growing affordability gap. Especially, in some of the fastest growth regions such as California and Washington, it is shocking that the median household income is lower than some other regions such as Maryland. This finding necessitates a deeper analysis into the factors driving this divergence, such as economic policies, supply and demand dynamics, or regional economic conditions.

Moreover, the future projections up to 2027 indicate that this trend is expected to continue, potentially exacerbating affordability issues. This discovery shifts the focus of the project towards understanding the implications of this trend on different socioeconomic groups and identifying potential interventions to address housing affordability. It also underscores the importance of considering regional variations, as some areas may experience more severe disparities than others. It is also important to not look over the history of economic events to project future data trends. Overall, these insights prompt a re-evaluation of the project's objectives, emphasizing the need to explore targeted strategies to balance housing prices with household incomes for sustainable market growth.

### Data Visualization 2



The second visualization is a Geospatial analysis map. In housing market analysis, geospatial maps can illustrate regional disparities in median housing prices across zip codes in Washington state in 2014 and 2015. Geospatial maps are effective for visualizing data with a geographical component, enabling the identification of spatial patterns and trends. This map leverages GIS technology to provide a detailed view of housing prices at the zip code level, offering insights into local variations. The use of a color gradient helps to intuitively convey the range of housing prices, making it easy to identify areas with higher or lower prices.

As discussed, the map uses different shares of color to represent varying housing price levels ranging from $75,000 to $7,700,000, with darker shades indicating higher prices. It shows that there are specific locations that has higher median housing prices compared to other locations. Each zip code is outlined and filled with a color that corresponds to the median price within that area. By looking closely at different zip codes or areas, stakeholders can identify housing trends and make informed decisions. For example, it is clear that areas around Seattle have significantly higher price than other surrounding areas. This visualization allows for a spatial comparison of housing prices within the state, highlighting regional differences.

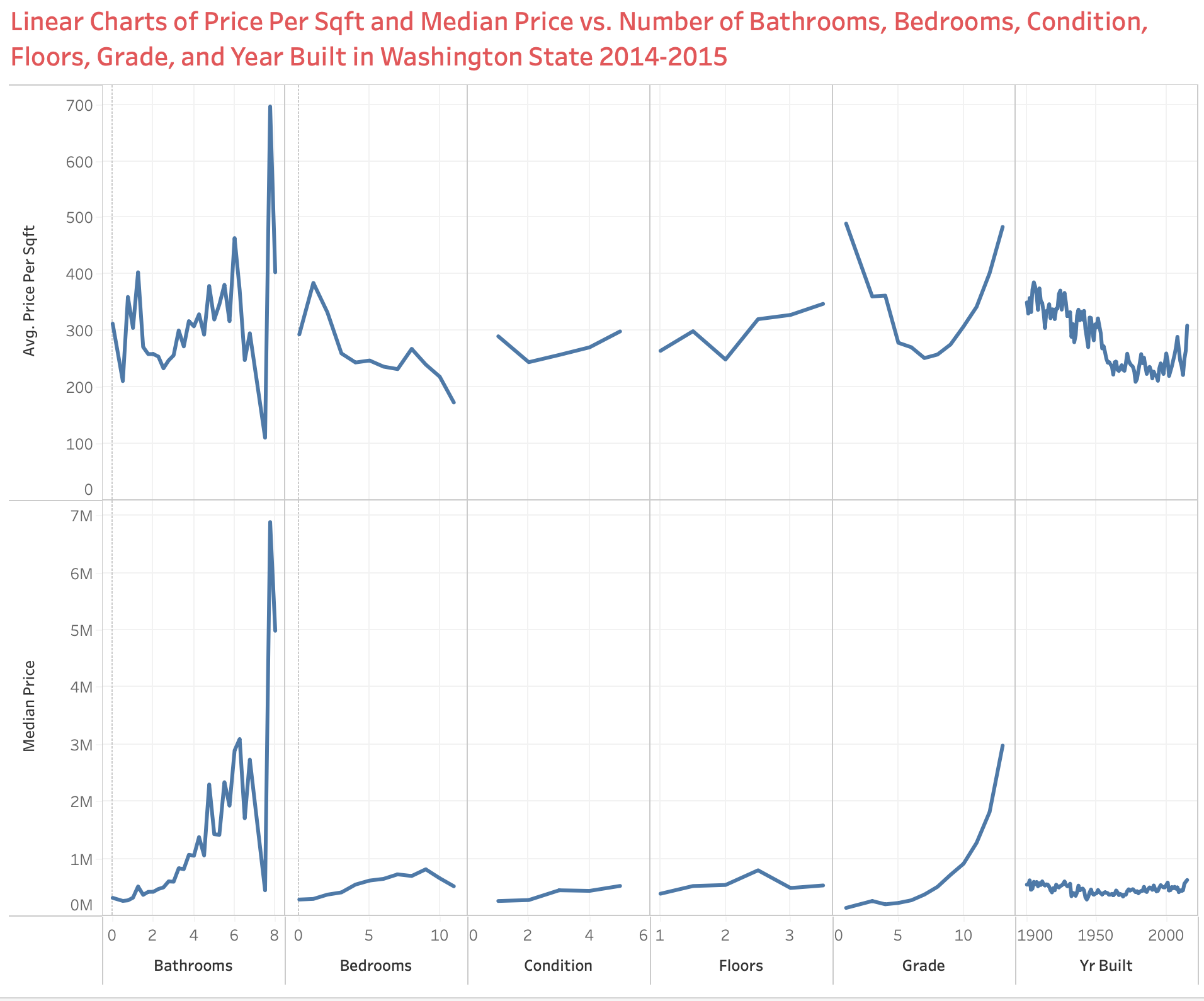
The geospatial analysis map reveals significant regional variations in housing prices within Washington State. Some areas, particularly those in the central and western parts of the state near Seattle, have higher median housing prices compared to other regions. This suggests that housing demand and market conditions vary significantly across the state, likely influenced by factors such as proximity to urban centers, economic opportunities, and local amenities. This aligns with the well-known economic activity and population density in these regions, driving up real estate values. On the other hand, the central and southern parts of the state exhibit much lower median housing prices, indicating a more affordable housing market. This could be due to various factors such as lower population density, less economic activity, and a higher availability of land.

Another key insight is the visualization of economic and social stratification within the state. The high concentration of wealth and higher median prices in specific zip codes highlights areas of affluence and potentially higher socioeconomic status. These areas may be experiencing higher demand due to their attractiveness for living, investment, or development. Conversely, the lighter shaded regions could indicate areas that might be more economically challenged or have less market demand. This geographic disparity can influence various socio-economic policies, urban planning, and resource allocation by state and local governments. Understanding these regional differences is crucial for stakeholders in the real estate market, as it helps identify opportunities and challenges in different parts of the state.

The discoveries from the map alter some expected results regarding the uniformity of real estate price across Washington State. The initial expectation was that housing prices would show some regional variation but would be relatively uniform across the state. However, the geospatial analysis map reveals substantial disparities in housing prices, with certain areas significantly outpacing others. This unexpected finding suggests that the housing market is highly localized, with different regions experiencing unique market dynamics. Another unexpected discovery is the extent of high-value areas even beyond the immediate urban centers. While it was anticipated that urban areas like Seattle would have higher housing prices, the map shows that some surrounding zip codes also exhibit high median prices.

This discovery necessitates a deeper exploration of the factors driving these regional differences. This indicates a spillover effect where the high demand and limited supply in central urban areas drive up prices in neighboring regions. Additionally, the lower-than-expected prices in some central and southern regions suggest potential opportunities for development and investment that might not have been considered otherwise. The scope of the project may now include analyzing the impact of local economic conditions, demographic trends, and policy interventions on housing prices. Additionally, there may be a need to investigate the implications of these disparities for housing affordability and access in different regions, providing a more nuanced understanding of the housing market.

### Data Visualization 3

**

The third visualization consists of linear charts depicting the relationship between various property attributes and real estate prices in Washington State from 2014 to 2015. The attributes examined include the number of bathrooms, bedrooms, condition, floors, grade, and year built. Each subplot contains two-line graphs: one showing the average price per square foot and another showing the median price, both plotted against the respective attribute on the x-axis. The y-axes on the above left of each subplot measure the average price per square foot, while the y-axes on the left below measure the median price. The charts reveal significant variations and patterns across different property attributes, with some exhibiting clear trends and others showing more scattered data points. The x-axis for bathrooms, bedrooms, condition, floors, and grade are numerical, while the year-built ranges from 1900 to 2000s. The title succinctly summarizes the content of the charts, emphasizing the temporal and geographical scope of the data.

The first chart, for bathrooms, shows an interesting trend where properties with eight bathrooms have an exceptionally high average price per square foot and median price while properties with nine bathrooms have slightly lower value and seven bathrooms have extremely low value. The second chart, for bedrooms, demonstrates a relatively stable trend with a slight increase in median price as the number of bedrooms increases, up to ten while the average price-per-sqft has a downward trend. The condition chart shows that properties in the best condition (grade 5) have the highest prices per square foot and median prices, while those in poorer conditions exhibit lower values. The floors chart suggests that properties with two or three floors tend to have higher prices than those with only one floor. The grade chart reveals a significant increase in median prices for properties with higher grades, particularly those with a grade of ten and above while the average price-per-sqft does not show the same pattern. Lastly, the year-built chart shows a general trend of increasing median prices for more recently built properties, with some fluctuations but the average price-per-sqft chart indicates that properties from the 1900s tend to have higher value.

The insights from these visualizations provide a detailed understanding of how different attributes affect property prices in Washington State during 2014-2015. For instance, properties with a higher number of bathrooms, particularly eight, are associated with significantly higher prices, indicating that such homes may be luxury properties. The relatively stable increase in median price with the number of bedrooms suggests a steady market demand for larger homes. In terms of condition, the data indicates that homes in excellent condition command higher median prices, emphasizing the importance of property maintenance and upgrades in enhancing property value. However, in term of average price-per-sqft, increasing in number of bedrooms, condition, or grade does not guarantee an increase in value.

The analysis of floors shows a preference and higher valuation for multi-story homes over single-story homes, which could reflect buyer preferences for larger living spaces without occupying larger plots of land. The grade attribute clearly shows a steep increase in median prices with higher grades, underscoring the value of premium-quality construction and finishes. In term of average price-per-sqft, we do not see the same pattern. The year-built analysis indicates a median price premium for newer homes, which is consistent with expectations as newer homes typically offer more modern amenities and construction standards. Surprisingly, older homes have significantly higher average price-per-sqft compared to newer constructions. The reason can be attribute to the neighborhood and geographic of older homes. These insights are critical for real estate investors, developers, and homeowners aiming to maximize their property value by focusing on these key attributes.

While the general trends align with common expectations, some specific discoveries significantly alter the anticipated results. For example, the exceptionally high prices associated with properties having eight bathrooms but extremely low with seven bathrooms were unexpected and suggest the presence of ultra-luxury and wrongly input properties that drastically skew the data. This outlier indicates a niche market that significantly affects the overall analysis and should be considered separately to avoid misinterpretation of the general market trends. Additionally, the relatively flat and downward trend for the number of bedrooms beyond seven contradicts the expectation that larger homes with more bedrooms would consistently command higher prices. This may indicate a market saturation point or diminishing returns for additional bedrooms beyond a certain number.

Another surprising discovery is the relatively low impact of the condition attribute on the price compared to other attributes like grade and number of bathrooms. This suggests that while maintenance is important, other factors such as construction quality and modern amenities play a more significant role in determining property value. Furthermore, the fluctuations in prices for properties built between 1900 and 2000 indicate that historical factors and specific periods of architectural significance might influence property values more than the general trend of newer being better. These findings challenge the conventional wisdom and highlight the need for a more nuanced approach when analyzing real estate market data. Overall, the insights gained from these visualizations underscore the complexity of the real estate market and the importance of considering multiple factors in property valuation. The unexpected results highlight the need for detailed, attribute-specific analysis to accurately understand market dynamics and inform strategic decisions in real estate investments and development.

### Proposed Visualizations

A valuable addition to the existing visualizations would be a time-series analysis chart that tracks the changes in median housing prices across different zip codes over the years. This line chart could feature multiple lines, each representing a different zip code, with the x-axis representing the months and year, and the y-axis representing the median housing prices. This visualization would allow us to observe trends over time, identify periods of rapid growth or decline, and correlate these changes with broader economic events or policy changes. The time-series analysis would be beneficial because provides a dynamic view of the housing market, unlike the static snapshot currently available. By examining how prices have evolved over time, stakeholders can better understand the long-term trends and potential future directions of the market. This could help real estate investors identify which areas are consistently appreciating in value and which might be stagnating or declining. Additionally, policymakers could use this data to evaluate the impact of housing policies over time and adjust strategies accordingly to promote balanced growth across the state.

Moreover, this visualization can reveal cyclical patterns in the real estate market, such as seasonal variations or the impact of economic cycles. For instance, it can help identify whether certain areas experience price drops during economic recessions and recover quickly afterward. Understanding these patterns can be crucial for timing investments and making informed decisions about buying or selling properties. This longitudinal perspective is essential for anyone looking to make data-driven decisions in the housing market. Lastly, a time-series analysis can highlight the resilience of certain markets compared to others. Some areas might show a steady increase in prices despite economic downturns, indicating strong market fundamentals and a robust economy. Conversely, areas that exhibit significant volatility might be riskier investments but could also offer higher returns for those willing to take on more risk. This detailed temporal analysis provides a comprehensive view that complements the geographical and attribute-based insights from the current visualizations.

Another valuable addition would be a correlation matrix heatmap showing the relationships between various housing attributes and their impact on median housing prices. This heatmap would display the correlation coefficients between pairs of variables, such as the number of bedrooms, bathrooms, property condition, grade, year built, and price per square foot. Each cell in the matrix would be color-coded to represent the strength and direction of the correlation, with darker colors indicating stronger correlations. This visualization would be valuable because it provides a clear and concise summary of how different attributes influence housing prices. By identifying strong correlations, stakeholders can prioritize certain features when making decisions about buying, selling, or renovating properties. For instance, if the number of bathrooms has a strong positive correlation with housing prices, homeowners might consider adding more bathrooms to increase their property’s value. Conversely, attributes with weak or negative correlations can be deprioritized in investment decisions.

Additionally, a correlation matrix heatmap can help uncover hidden relationships between variables that are not immediately obvious from the linear charts or the geospatial map. For example, it might reveal that certain combinations of attributes, like having a high-grade home with a large number of bedrooms, are particularly valuable. This kind of insight can inform more nuanced investment strategies and development plans, targeting specific combinations of features that maximize property value. Furthermore, this visualization can aid in risk assessment and portfolio diversification for real estate investors. By understanding which attributes are most strongly correlated with price fluctuations, investors can better anticipate how changes in one attribute might affect their overall portfolio value.

# Predictive Models

### Data Modeling Definitions

To predict housing prices and market trends, I employed several machine learning models, including linear regression, decision tree, and random forest. Linear regression is a fundamental and widely used statistical method for predictive modeling, served as a starting point due to its simplicity and interpretability (Mali, 2024). It aims to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. It provided a baseline for understanding the linear relationships between housing prices and various predictors. However, the linear regression model's performance was limited by its inability to capture complex, non-linear interactions between variables.

**Decision trees** are a versatile and intuitive machine learning technique used for both regression and classification tasks. They work by recursively splitting the data into subsets based on the value of input features, creating a tree-like model of decisions (Song & Lu, 2015). Each node in the tree represents a feature, each branch represents a decision rule, and each leaf node represents an outcome or a prediction. Decision trees are appreciated for their simplicity and interpretability, as they mimic human decision-making processes. They can handle both numerical and categorical data and require little data preprocessing. However, decision trees are prone to overfitting, especially when they are deep and complex. Despite their potential for overfitting, decision trees are a powerful tool and form the basis for more advanced ensemble methods such as random forests and gradient boosting machines.

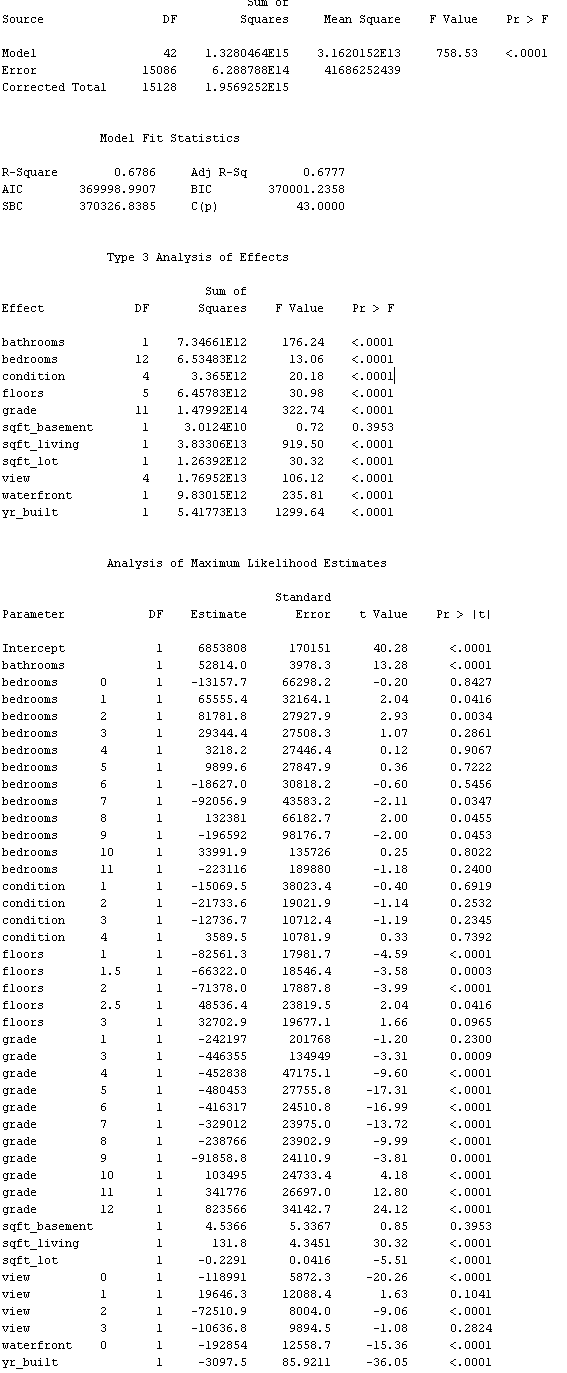
The random forest model, an ensemble learning method, was particularly robust in predicting housing prices. A random forest constructs a multitude of decision trees during training time and outputs the mean prediction of the individual trees (Knode, 2016). By combining multiple decision trees, random forest improved prediction accuracy and handled non-linear relationships and interactions between variables effectively. They can also handle large datasets with higher dimensionality and are relatively efficient. The model's ability to manage high-dimensional data and its robustness against overfitting made it an excellent choice for this analysis. Overall, the combination of these techniques demonstrates their complementary strengths in analyzing and predicting housing market price, which linear regression provides a baseline, decision tree classifies market segments, while random forest delivers high accuracy in predictions.

### Predictive Model 1

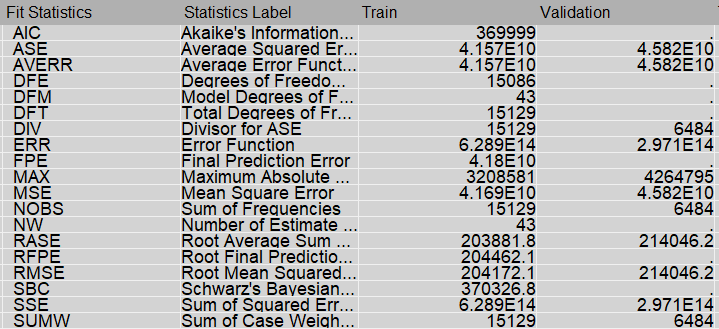
The linear regression model aims to predict the target variable, which is the housing price based on multiple input features including number of bathrooms, bedrooms, floors, grade, condition, square foot living, view, etc. The dataset comprises various types of variables, including binary, interval, and nominal variables. The model uses these features to construct a linear relationship that can predict the target variable as accurately as possible. The training dataset contains 15,129 observations and the validation dataset contains 6484 observations. The model involves 43 parameters, which likely include the coefficients for each predictor variable, the intercept, and possibly other model-specific parameters. The error distribution is assumed to be normal, and the link function used is the identity function, which is typical for linear regression models.

The analysis of variance (ANOVA) indicates that the model is statistically significant with an F-value of 758.53 and a p-value less than 0.0001, demonstrating that the model as a whole has predictive power. The model fit statistics further reveal the efficacy of the regression model. The R-squared value is 0.6786, and the adjusted R-squared value is 0.6777. These values indicate that approximately 67.86% of the variance in the target variable (price) is explained by the model, which suggests a relatively strong model fit. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are also provided, with values of 369998.9907 and 370001.2358, respectively. These metrics help in comparing models and selecting the one with the best trade-off between fit and complexity.

The type 3 analysis of effects indicates the significance of each predictor variable in the model. For instance, the number of bathrooms has an F-value of 176.24 with a p-value less than 0.0001, indicating a highly significant effect on the price. Similarly, other variables such as the number of bedrooms and various categorical predictors also show significant F-values, suggesting their importance in the model. The analysis provides a comprehensive understanding of how different variables contribute to predicting the target variable. The regression parameter estimates give detailed insights into the relationship between each predictor and the target variable. For example, the parameter estimate for bathrooms is 7.34661E12, reflecting a positive relationship with the price.



**Table 1. Linear Regression Model Results**

****

**Table 2. Linear Regression Model Fit Statistics**

The linear regression model provides significant insights into the factors influencing the price. The high R-squared value indicates that the model captures a substantial portion of the variability in price, suggesting that the selected predictors are relevant and influential. The standard errors, t-values, and p-values for each parameter estimate are also reported, allowing for hypothesis testing on the significance of each predictor. The residuals and predicted values are examined to assess the model's predictive accuracy and identify any potential outliers or anomalies. The significance of variables such as bathrooms and bedrooms highlight the importance of these features in determining price, which aligns with intuitive expectations in real estate valuation.

Moreover, the detailed parameter estimates and their associated statistics offer a deeper understanding of the magnitude and direction of each predictor's effect on price. These estimates reveal that the number of bathrooms, grade, square foot living area, year built, view, and waterfront status are among the leading indicators of housing price. For instance, an increase in the number of bathrooms and higher grades of construction and finish quality are strongly associated with higher property prices. Similarly, larger living areas, newer construction years, properties with views, and those situated on waterfronts significantly elevate housing prices. This detailed information can be invaluable for stakeholders, such as real estate agents, property developers, and investors, in making informed decisions based on empirical data.

One unexpected finding from the model results is the significant effect of certain nominal variables, which might have been initially considered less important. For instance, variables related to the quality of the property, such as the condition and quality ratings, or specific location attributes, such as neighborhood characteristics and proximity to amenities, showed high significance levels. These findings alter the initial scope of expected results by highlighting the substantial impact of qualitative and locational factors on housing prices. This underscores the necessity of considering a wide range of factors beyond just the physical attributes of the property when predicting price. The significance of these nominal variables suggests that subjective and contextual factors play a crucial role in real estate valuation. This realization prompts a more holistic approach to property valuation, integrating qualitative assessments and neighborhood-level data to capture the true value of a property accurately.

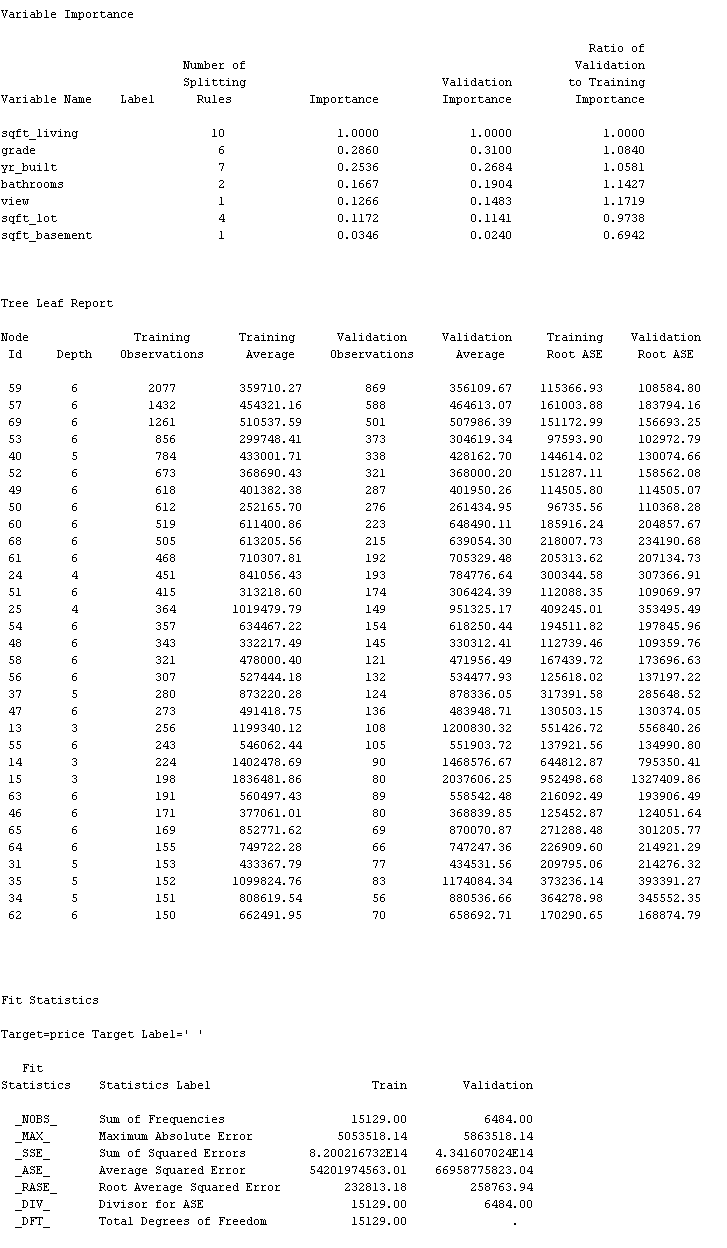
Another notable discovery is the presence of significant residuals in certain cases, indicating that the model may not fully capture some aspects of the variability in price. These residuals suggest the presence of additional factors not included in the model, such as macroeconomic conditions, market trends, or unobserved property characteristics. For instance, fluctuations in interest rates, changes in local economic conditions, and shifts in buyer preferences can all influence housing prices but may not be directly accounted for in the model. This realization calls for a broader approach in future modeling efforts, incorporating more diverse data sources to enhance predictive accuracy. By broadening the scope of data collection and incorporating these additional factors, future models can achieve higher accuracy and provide more reliable predictions, ultimately benefiting all stakeholders involved in the real estate industry. This holistic perspective can lead to better decision-making and optimized outcomes in the relevant domain, ensuring that all influential factors are considered in property valuation.

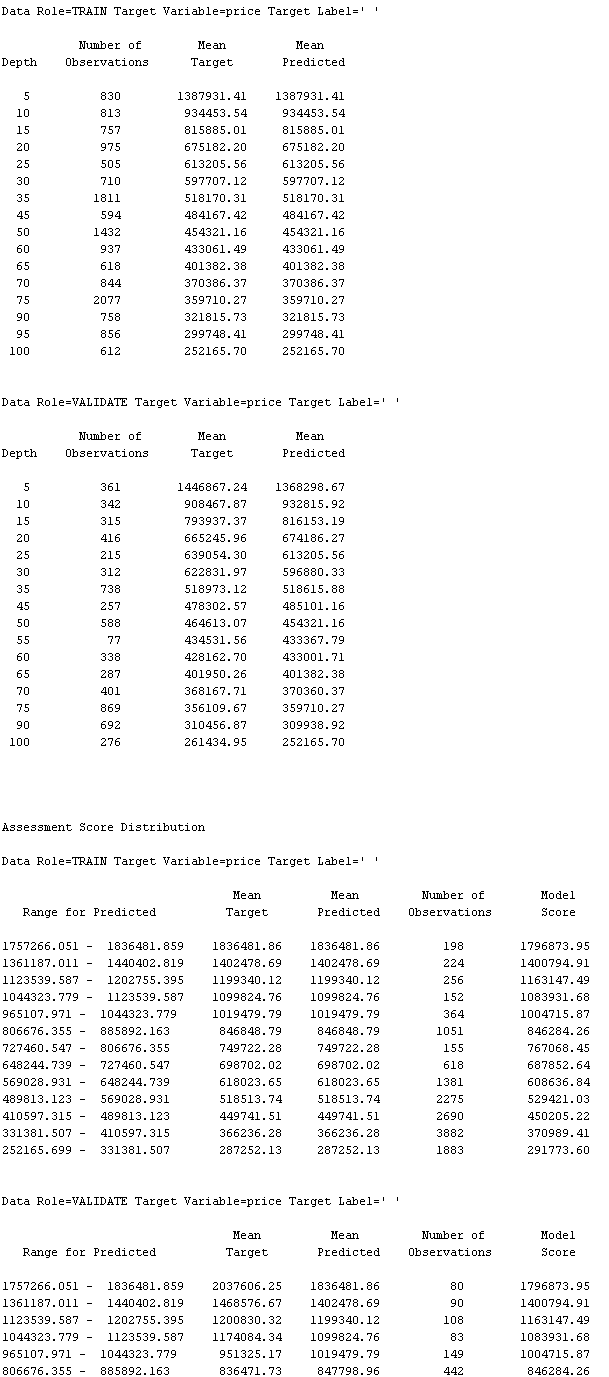
### Predictive Model 2

The decision tree model is aimed at predicting the target variable, which is the price, by recursively splitting the dataset based on the input features to create branches and leaves that represent different value ranges and categories. In this specific model, the primary variables influencing the splits are likely to be those with the most significant impact on the target variable. The model's construction involves calculating measures such as Gini impurity or entropy to determine the best splits. These measures assess the homogeneity of the target variable within the branches. A low value of impurity indicates that the nodes contain predominantly one class, which is ideal for accurate predictions. The tree continues to grow until it meets stopping criteria, such as a maximum depth of 6 or minimum number of samples per leaf of 150, to prevent overfitting.

Similar to linear regression model, the training dataset contains 15,129 observations and the validate dataset contains 6484 observations. There are 11 independent variables including binary, interval, and nominal variables. The “Variable Importance” section provides critical insights into which variables had the most significant impact on the model's decisions. For instance, sqft\_living was identified as the most influential variable with an importance score of 1.0000, followed by 'grade' and yr\_built with lower but still significant importance scores. These metrics highlight the key drivers of the model's predictions and help prioritize areas for further data collection and analysis. The ratio of validation to training importance for each variable also indicates how well the model generalizes to new data, with most ratios being close to 1.0, suggesting good generalization.

The “Tree Leaf Report**”** presents detailed information about the performance of individual nodes within the tree. For example, node 59 at depth 6 has 2,077 training observations with an average price prediction of $359,710.27 and a root ASE (Average Squared Error) of $115,366.93. The validation set for this node includes 869 observations with an average price prediction of $356,109.67 and a root ASE of $108,584.80. These metrics provide a granular view of the model's performance across different segments of the data, helping to identify areas where the model performs well and where it may need improvement.





**Table 3. Decision Tree Model Results**

The decision tree model offers valuable insights into the relationships between various predictors and the target variable, price. One key insight is the importance of sqft\_living as the primary predictor, which significantly influences the model's decisions. This finding aligns with real estate market trends, where the living area of a property is a crucial factor in determining its value. Understanding this relationship helps stakeholders focus on critical factors when analyzing property prices. Another important insight is the varying importance of other predictors such as grade, yr\_built, and bathrooms. These variables, while not as influential as sqft\_living, still play significant roles in the model's predictions.

Furthermore, the importance scores suggest that properties with higher grades, newer construction, and more bathrooms tend to have higher prices. These insights can guide future data collection efforts, ensuring that these key variables are accurately measured and included in the analysis. The Fit Statistics section shows that the model has a root average squared error (RASE) of $232,813.18 for the training set and $258,763.94 for the validation set. The closeness of these values indicates that the model performs consistently across both sets, suggesting good generalization. Additionally, the maximum absolute error for the validation set is slightly higher than the training set, which is typical and indicates areas where the model could be further refined.

One discovery that altered the scope of the expected results is the varying performance of the model across different nodes and segments of the data. The Tree Leaf Report shows that some nodes have significantly higher errors than others, highlighting areas where the model may struggle to make accurate predictions. For instance, node 59 at depth 6, with a training average of $359,710.27, has a root ASE of $115,366.93, while node 25 at depth 4 has a much higher training average of $1,019,479.79 and a root ASE of $409,245.01. This substantial difference in error metrics suggests that the model's accuracy varies significantly depending on the data segment. It also indicates that certain nodes may be influenced by outliers or less representative data. This insight suggests that the model could benefit from additional tuning or the incorporation of more sophisticated algorithms to handle complex patterns in the data. It also emphasizes the need for continuous monitoring and evaluation of the model's performance to identify and address potential weaknesses.

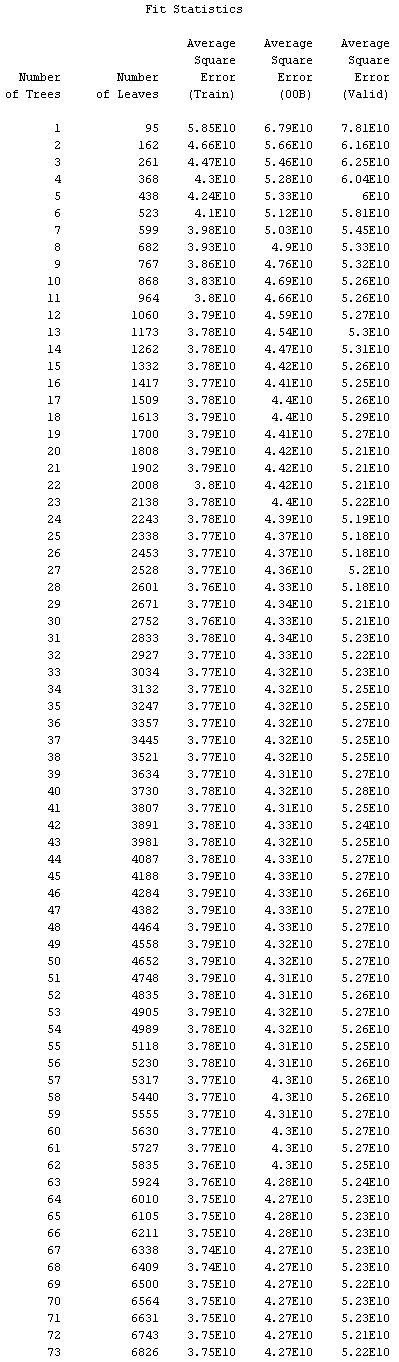
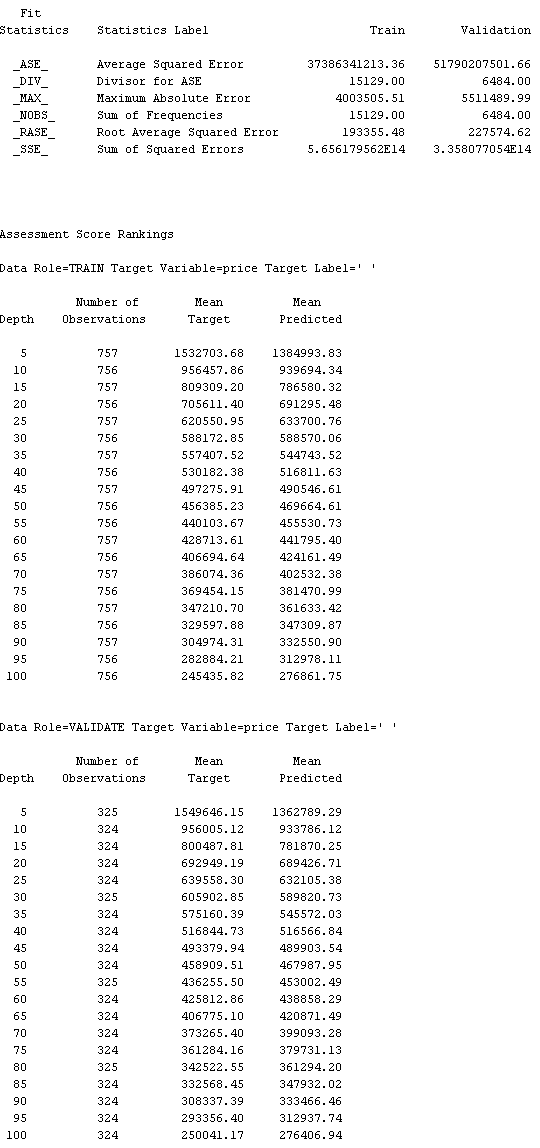
Another discovery is the assessment score rankings and distribution tables provide a detailed view of how the model's predictions vary across different ranges of the target variable. These tables reveal that the model tends to perform better in certain price ranges than others, indicating potential biases or limitations in the training data. For example, the model predicts high-end properties (with prices over $1,000,000) with less accuracy, as evidenced by the higher validation root ASE for these segments. This inconsistency suggests that the training data might not be adequately representative of all price ranges, leading to biased predictions. Addressing these issues may require gathering more representative data or applying advanced techniques such as ensemble learning to improve overall accuracy and robustness. This realization shifts the project's scope towards enhancing data diversity and model complexity to achieve more reliable predictions.

### Predictive Model 3

The random forest model uses 100 actual trees with an in-bag fraction of 0.6. It has been configured with several default parameters such as a prune threshold of 0.1, leaf fraction of 0.00001, and a maximum depth of 50. Category bins are set to 30, with interval bins at 100. The model was also trained and validated using a dataset with 21,613 observations, split into 15,129 for training and 6,484 for validation. The baseline fit statistics show an average square error of 129,349,274,164 for training and 148,011,916,961 for validation. The fit statistics provide detailed error metrics across different numbers of trees, showing a trend of decreasing error with more trees, indicating the model's improvement as more trees are added.

The results show that as the number of trees increases, the average square error decreases for both training and validation sets, suggesting that the model is learning effectively. For instance, with one tree, the average square error for training is 5.85E10, and for validation, it is 7.81E10. As the number of trees increases to seven, the average square error for training decreases to 3.98E10, and for validation, it decreases to 5.45E10. These results indicate that the model's accuracy improves with more trees, which is a characteristic of Random Forest models due to their ensemble learning approach.

The detailed fit statistics also highlight the number of leaves used in each tree, which increases as more trees are added. For example, with three trees, there are 261 leaves, and with seven trees, there are 599 leaves. This increase in complexity helps the model capture more intricate patterns in the data. The complexity added by more leaves helps the model generalize better on unseen data by reducing bias. The mean predicted target values and the number of observations across different ranges of predicted values are provided, giving insights into the distribution and accuracy of predictions.



**Table 4. Random Forest Model Results**

From the model's performance metrics, it is clear that the Random Forest model benefits significantly from an increased number of trees. The consistent decrease in average square error with more trees suggests that the model is effectively capturing the underlying data patterns and reducing overfitting through ensemble averaging. This ensemble effect helps in balancing the variance and bias, making the model more robust and stable. The error reduction from the initial trees to the later stages demonstrates the model's robustness and stability. The stability of the model across different numbers of trees shows that it is not overly sensitive to a specific number of trees, providing flexibility in its deployment.

The insights also highlight the importance of appropriate parameter settings, such as the in-bag fraction and category bins. The chosen parameters seem to provide a good balance between complexity and generalization, allowing the model to perform well on both training and validation datasets. The in-bag fraction, set at 0.6, controls the proportion of data used to train each tree, ensuring that each tree is trained on a diverse subset of data, which aids in reducing overfitting. Category bins set at 30 and interval bins at 100 ensure that the model can handle different types of data distributions effectively. The model's ability to handle missing values as valid entries without a significant drop in performance is another notable insight, indicating its robustness to data imperfections.

One key discovery is the significant impact of the number of trees on the model's accuracy. This finding suggests that future models should explore even higher numbers of trees to potentially improve performance further. The model's performance metrics indicate that more trees lead to lower error rates, but there is a point where the gains may start to diminish, and computational costs may outweigh the benefits. Additionally, the model's sensitivity to parameters like in-bag fraction and category bins indicates that parameter tuning could be crucial for optimizing model performance. Fine-tuning these parameters can lead to significant improvements in model accuracy and generalization capabilities.

Another discovery is the model's handling of missing values without substantial performance degradation. This capability could expand the project's scope to include datasets with missing values without requiring extensive data preprocessing. The model's robustness in this aspect can save time and resources, allowing for a broader application of the Random Forest model in various real-world scenarios where data completeness is often a challenge. This flexibility is particularly valuable in industries where data collection is ongoing, and the dataset may evolve over time, ensuring that the model remains effective without frequent retraining.

### Predictive Model Review

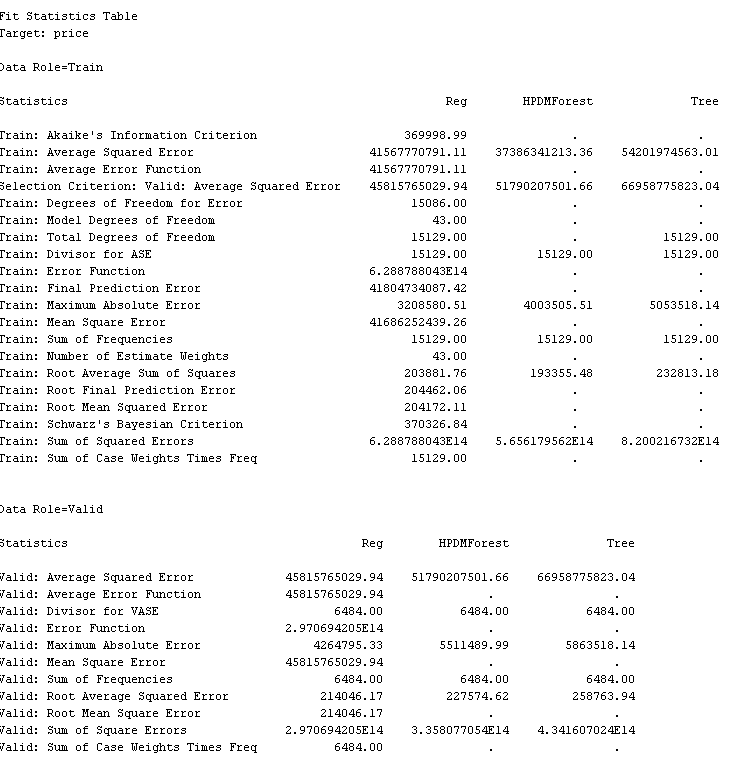
Based on the model comparison result, the linear regression model shows the lowest Average Squared Error (ASE) on the validation datasets, indicating its superior predictive accuracy. ASE is a key metric that evaluates the average of the squares of the errors, which means the lower the ASE, the closer the model's predictions are to the actual values. The linear regression model’s lower ASE indicates that it consistently makes smaller errors in its predictions compared to the other models. In contrast, the random forest model has a lower ASE than the decision tree but higher than regression on validation datasets but has the lowest on the training dataset, placing it in the middle. The decision tree exhibits the highest ASE, making it the least accurate among the three models.

The linear regression model also demonstrates the lowest Maximum Absolute Error on both training and validation data, suggesting it has fewer large deviations in its predictions compared to the other models. Maximum Absolute Error measures the largest single error in the model’s predictions, and a lower value indicates that the model is less prone to making significant errors. The random forest model has a higher Maximum Absolute Error than regression but lower than the decision tree, indicating that while it has some significant errors, they are not as severe as those in the decision tree model. This is critical for applications where large prediction errors can be particularly detrimental. The decision tree’s higher Maximum Absolute Error highlights its tendency to occasionally make very inaccurate predictions, reducing its reliability.

Similarly, the Root Mean Squared Error (RMSE) follows the same pattern, with the linear regression model having the lowest RMSE on the validation data, followed by random forest and then the decision tree. This further corroborates the consistency and reliability of the linear regression model over the other two. RMSE is another measure of the average magnitude of the errors, taking the square root of the average of squared differences between predicted and actual values. The lower RMSE of the linear regression model suggests it provides more consistent and reliable predictions, reducing the impact of larger errors more effectively than the other models. The random forest model, while better than the decision tree, still has a higher RMSE than the regression model, indicating it is less precise. The decision tree’s highest RMSE further corroborates its status as the least reliable model, as it indicates larger average errors in its predictions.

Based on the comparative analysis, the linear regression model is recommended as the champion model. It consistently shows the lowest errors across various metrics, including ASE, Maximum Absolute Error, and RMSE, for validation datasets. This consistency in performance demonstrates that the linear regression model is highly effective in capturing the underlying patterns in the data. The model’s ability to minimize both average and maximum errors is crucial for applications where prediction accuracy and reliability are paramount. This indicates that the linear regression model provides more accurate and reliable predictions, making it the most robust choice among the models compared.

The consistent performance of the linear regression model across different error metrics highlights its effectiveness in capturing the underlying patterns in the data without overfitting. Overfitting occurs when a model is too closely fitted to the training data, making it less effective on new, unseen data. For instance, the random forest model indicates overfitting of data with lower ASE and RMSE on training data but higher on validation data compared to linear regression model. In conclusion, the linear regression model’s lower Maximum Absolute Error indicates fewer large prediction errors, which is crucial for applications requiring high precision. Furthermore, the lower RMSE confirms its ability to minimize prediction errors more effectively than random forest and decision tree models. This solidifies the linear regression model’s position as the best choice for deployment, ensuring more accurate and dependable performance in real-world applications.



**Table 5. Model Comparison Result**

**Final Results**

### Analysis Justification

The analysis of housing prices and market trends using machine learning models, including linear regression, decision trees, and random forests, was strategically designed to capture both simple and complex patterns in the data. Linear regression was chosen as the baseline model due to its straightforward approach to modeling linear relationships, allowing for easy interpretation of how individual predictors, such as the number of bedrooms, bathrooms, and square footage, influence housing prices. Despite its limitations in handling non-linear interactions, linear regression offers a clear understanding of which factors are most directly correlated with price changes, providing a foundational perspective that can guide further analysis. This model is particularly valuable in identifying key variables and setting a benchmark for performance against more sophisticated models. By highlighting the strengths and weaknesses of linear regression, the analysis sets the stage for more advanced techniques that can address its limitations, offering a comprehensive view of the data's underlying structure.

Building on the insights gained from linear regression, decision trees were employed to explore non-linear relationships and interactions between variables. Decision trees offer a more nuanced analysis by recursively partitioning the data into subsets based on feature values, thus capturing complex patterns that linear regression might miss. Their interpretability is also beneficial, as the tree structure visually represents decision paths and the importance of different features in predicting housing prices. However, decision trees are susceptible to overfitting, especially when they grow too deep. This analysis underscores the need to balance model complexity with generalization, highlighting the importance of setting appropriate stopping criteria, such as maximum tree depth and minimum samples per leaf, to maintain model robustness. The decision tree analysis reveals how factors interact in diverse ways, providing insights into market segmentation and the heterogeneous nature of housing prices across different market conditions.

The random forest model, an ensemble learning technique, was incorporated to further enhance prediction accuracy by averaging the outcomes of multiple decision trees. This approach mitigates the overfitting tendency of individual decision trees by introducing randomness in feature selection and data sampling, resulting in more stable and reliable predictions. Random forests excel in capturing intricate relationships within high-dimensional datasets, offering robustness to variations in the data and the ability to model complex interactions. The analysis of random forests demonstrates their superior performance in handling large datasets with numerous variables, making them particularly suited for predicting housing prices with greater precision. Additionally, the model's feature importance scores provide valuable insights into which variables have the most significant impact on price predictions, guiding strategic decisions in real estate investment and policy-making. By combining the strengths of individual trees into a cohesive model, random forests deliver a comprehensive solution that balances accuracy and generalization, ensuring reliable insights into housing market dynamics.

### Findings

The housing market project has yielded several critical insights through detailed visualizations. One major finding is the divergence between the growth rates of median housing prices and household incomes, as seen in the line chart spanning from 2000 to 2023. This chart shows that while housing prices have seen significant increases, especially post-2008 financial crisis, median household incomes have grown at a much slower pace. The visualization with future projections up to 2027 suggests this trend will continue, potentially exacerbating the affordability issues for homebuyers. This disparity raises concerns about the sustainability of the housing market growth and highlights the need for policies aimed at bridging the gap between income and housing costs.

Geospatial analysis of housing prices across Washington state in 2014 and 2015 reveals significant regional disparities. Areas around urban centers like Seattle display considerably higher prices than other regions, illustrating a strong market demand linked to economic opportunities and amenities. Conversely, central and southern regions of the state showed lower housing prices, which might attract development and investments overlooked due to the high costs in more urbanized areas. This analysis underlines the importance of geographic factors in real estate pricing and the potential for targeted economic development to enhance market accessibility.

Further, the examination of property attributes in Washington State underscores how certain features like the number of bathrooms, condition, and grade of homes influence housing prices. The charts show a clear preference for homes with higher grades and more bathrooms, indicating a market inclination towards luxury properties or those with better amenities. Meanwhile, older homes command a higher price per square foot, suggesting a valuation based on location or historical significance rather than mere age. These findings provide valuable insights for stakeholders in the real estate market, highlighting the need for a nuanced understanding of how various attributes impact property values. These findings from the housing market project offer valuable perspectives for stakeholders, including policymakers, investors, and potential homeowners, providing them with a data-driven foundation to make informed decisions and develop strategies to address the challenges within the housing market.

**Table 6: Growth Rate Disparities between Housing Prices and Household Incomes**

|  |  |  |  |
| --- | --- | --- | --- |
| Year Range | Median Household Income Growth | Median Housing Price Growth | Future Projection Impact |
| 2000-2023 | Slow growth | Significant increases | Affordability issues exacerbated up to 2027 |
| Post-2008 | Minimal increase | Rapid recovery and growth | Continued disparity |

**Table 7: Regional Housing Price Differences in Washington State (2014-2015)**

|  |  |  |  |
| --- | --- | --- | --- |
| Region | Housing Price Level | Economic Opportunities | Potential for Development |
| Urban Centers (e.g., Seattle) | High | Strong | Low due to high costs |
| Central and Southern Regions | Low | Moderate to Low | High, overlooked opportunities |

**Table 8: Impact of Property Attributes on Housing Prices in Washington State**

|  |  |  |
| --- | --- | --- |
| Property Attribute | Impact on Housing Prices | Market Preference |
| Number of Bathrooms | Higher prices with more bathrooms | Preference for luxury |
| Grade | |  | | --- | |  |  |  | | --- | | Higher prices with higher grades | | High-quality finishes and conditions favored |
| Age of Home | Older homes have higher prices per square foot | Valued for location or historical significance |

### Review of KPIs

First, achieving 100% of project milestones within the planned timeframe stands out as a major accomplishment. This indicates effective project management and adherence to the schedule, which is crucial in ensuring timely data analysis and reporting. The ability to meet deadlines without compromising the depth and quality of analysis exemplifies a well-organized project framework. This accomplishment likely contributed positively to the project's overall credibility, underpinning its findings and recommendations.

However, the project fell short of the KPI for accuracy in its predictive models, achieving an R-squared value of 0.6786 and an adjusted R-squared value of 0.6777. While these values suggest that approximately 67.8% of the variance in the target variable (housing prices) is explained by the model, the goal was to surpass the 75% threshold. Despite this, the model's performance is still commendable and indicates a substantial capture of the variability in housing prices without overfitting, suggesting a reliable predictive framework. This shortfall highlights an opportunity for further refinement of the model, possibly by exploring additional variables, employing more sophisticated modeling techniques, or increasing the dataset's size and diversity.

The impact of the analysis on stakeholders' decision-making processes has been notably positive and transformative. For real estate investors, the project's findings have enabled the optimization of investment strategies and the identification of high-growth areas, which can lead to more targeted and profitable investments. Policy-makers have utilized the data to address specific regional needs and promote housing affordability, which is crucial in mitigating the effects of rising housing prices on accessibility and societal equity. Finally, potential homebuyers are better equipped to make informed purchasing decisions, armed with knowledge about how different property attributes and regional characteristics influence housing prices. This empowerment helps them navigate the complex housing market more effectively, potentially leading to better financial and lifestyle outcomes.

In summary, while the project excelled in execution and impact, the accuracy of predictive modeling did not fully meet the predefined KPI. Future efforts might focus on enhancing the predictive power of the models while maintaining the practical relevance and applicative value of the analysis to continue supporting stakeholders effectively. This balanced approach will ensure that the project not only meets but exceeds all performance expectations in subsequent iterations.

### 

### Review Significance

The housing market project has proven to be significant in several respects, notably in enhancing understanding of the complexities within the housing market across different regions. By analyzing historical data and current trends in housing prices and household incomes, the project has illuminated the growing disparity between these two metrics, which is a central concern for economic sustainability and social equity. This analysis has provided stakeholders with a clear visualization of how housing affordability issues are evolving, highlighting the urgent need for interventions that can bridge the gap between rising housing prices and lagging income growth. Additionally, the project has offered insights into the recovery trends post-financial crisis and projected future developments, making it an invaluable resource for planning and policy formulation. The findings have underscored the necessity for targeted economic policies and have facilitated a deeper discussion among policymakers, investors, and the public about strategic solutions to housing challenges.

Furthermore, the project's geospatial analysis of housing prices in Washington State has brought to light the significant regional variations within a single state, offering a detailed picture of how location-specific factors such as economic opportunities, public amenities, and demographic characteristics influence housing markets. This segmentation has provided a granular level of detail that is often overlooked in broader market analyses, allowing for more precise and effective policy and investment decisions. The ability to pinpoint which areas are experiencing price inflation and which are undervalued presents opportunities for strategic development and investment that could balance regional economic growth and improve housing accessibility. This aspect of the project is particularly significant for local governments and urban planners who are tasked with promoting equitable development and managing urban sprawl. The insights from this analysis are crucial for tailoring housing policies and development initiatives to the unique needs and potentials of different communities.

Finally, the application of advanced predictive modeling techniques in this project has set a new standard for forecasting in the real estate sector. By incorporating a wide range of variables, including economic indicators, demographic data, and even potential impacts of climate change, the project has demonstrated the value of using sophisticated models to predict future housing market trends. These models enable stakeholders to anticipate changes and make informed decisions well in advance, thereby minimizing risks and maximizing returns on investment. The predictive insights also help in preparing for demographic shifts and economic changes that could affect housing demand and prices. Moreover, the ability to model future scenarios allows policymakers and developers to test the potential impacts of their policies or projects in a simulated environment before implementing them, thereby refining strategies and ensuring more effective outcomes. Overall, the project's findings significantly contribute to the strategic planning and proactive management of housing markets, ensuring that growth is both sustainable and inclusive.

### Recommendations for Future Analysis

For a more nuanced understanding and effective targeting of housing market interventions, a detailed approach to future analysis is recommended. Beginning with segmented geospatial analysis, this method delves into the granular differences in housing prices at the neighborhood or community level. This detailed examination can reveal how local factors such as school district quality, crime rates, demographics, and public transportation accessibility influence housing prices. For instance, neighborhoods with better school districts or lower crime rates may exhibit higher property values. Conversely, areas with poor public transit options might show lower housing prices. Such an analysis would provide stakeholders, including urban planners and real estate developers, with crucial information to tailor their strategies according to specific community needs and characteristics.

Advancing towards predictive modeling, employing more sophisticated techniques such as machine learning and deep learning can significantly enhance the forecasting accuracy of housing market trends. These models can process and analyze large datasets to uncover complex patterns that traditional statistical methods may miss. By incorporating a broader set of variables, including economic indicators like unemployment rates, demographic shifts, and potentially even climate change impacts, predictive models can offer more robust forecasts. This approach would not only improve the precision of price predictions but also help in understanding how macroeconomic and environmental changes could shape future housing markets. Such models could be crucial for long-term strategic planning by government bodies and real estate investors aiming to anticipate and mitigate future market disruptions.

Lastly, integrating these advanced predictive models with segmented geospatial analysis could provide a comprehensive toolkit for analyzing and forecasting housing market dynamics. This integrated approach allows for the simultaneous consideration of macro-level economic trends and micro-level local conditions. By doing so, it becomes possible to create highly detailed and localized housing market forecasts that are directly applicable to specific areas. This could be particularly beneficial for policy-makers and investors looking to implement targeted interventions that promote housing affordability and community development. For example, this analysis could identify potential areas for development incentives or affordable housing projects, aligning investment efforts with community needs and market demand trends. Overall, these advanced analytical techniques would greatly enhance the decision-making capabilities of all housing market stakeholders by providing deeper insights and more accurate predictions.

# References

Ahasan, R., Alam, M. S., Chakraborty, T., & Hossain, M. M. (2022). *Applications of GIS and geospatial analyses in COVID-19 research: A systematic review*. Retrieved from: https://doi.org/10.12688/f1000research.27544.2

Brar, S. (March, 2024). *Housing Price Dataset*. Retrieved from: <https://www.kaggle.com/datasets/sukhmandeepsinghbrar/housing-price-dataset>

Calzon, Bernardita (Jan 2024). *A Guide To The Methods, Benefits & Problems of The Interpretation of Data.* Retrieved from: <https://www.datapine.com/blog/data-interpretation-methods-benefits-problems/#data-interpretation-definition>

Ezzat, M. (March, 2024). *Housing Price Dataset*. Retrieved from: <https://www.kaggle.com/code/mariamezzathammam/housing-price-dataset/notebook>

Federal Reserve Bank of St. Louis. (2024). *Median Household Income Data*. Retrieved from: <https://fred.stlouisfed.org/graph/?id=MEHOINUSA672N,MEHOINUSCAA672N,MEHOINUSCOA672N,MEHOINUSDCA672N,MEHOINUSFLA672N,MEHOINUSGAA672N,MEHOINUSHIA672N,MEHOINUSMDA672N,MEHOINUSMAA672N,MEHOINUSNYA672N,MEHOINUSVAA672N,MEHOINUSWAA672N>

Gupta, E., & Walstrum, T. (September, 2023). *Regional Differences in Home Price Growth Since the Start of Covid-19 Are a Continuation of Long-Run Trends*. Retrieved from: [https://www.chicagofed.org/publications/blogs/midwest-economy/2023/regional-differences-home-price-growth#ftn3](https://www.chicagofed.org/publications/blogs/midwest-economy/2023/regional-differences-home-price-growth)

Hair, J.F., Hult, G.T.M., Ringle, C.M., Sarstedt, M., Danks, N.P., Ray, S. (Nov 2021). Overview of R and RStudio. *Springer Link.* Retrieved from: https://doi.org/10.1007/978-3-030-80519-7\_2

Hall, P., Dean, J., Kabul, I.K., Silva, J. (2014). *An Overview of Machine Learning with SAS Enterprise Miner.* Retrieved from: https://support.sas.com/resources/papers/proceedings14/SAS313-2014.pdf

Hennigan, Laura (Apr 2023). *What Is A KPI? Definition & Examples*. Retrieved from: <https://www.forbes.com/advisor/business/what-is-a-kpi-definition-examples/>

Jain, Kunal (Jun 2020). *Python vs R vs SAS - Which Tools Should I Learn for Data Science?*Retrieved from: <https://www.analyticsvidhya.com/blog/2017/09/sas-vs-vs-python-tool-learn/>

Katerova, Lia (Dec 2023). *Why Choose Tableau for Your Big Data Visualization?*Retrieved from: <https://b-eye.com/blog/why-choose-tableau-for-your-data-visualizations/>

Knode, Steve (Nov 2016). *Ensemble Models – Bagging\_Boosting\_Random\_Forest.* [Video].Retrieved from: <https://vimeo.com/189862546>

Mali, Kavita (Jul 2024). *What Is Linear Regression?* Retrieved from: <https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/>

Mulla, Rob (May 2024). *Zillow Home Value Index*. Retrieved from: https://www.kaggle.com/datasets/robikscube/zillow-home-value-index?resource=download&select=ZHVI.csv

Reimers, S., & Harvey, N. (March 2024). Bars, Lines, and Points: The Effect of Graph Format on Judgemental Forecasting. *International Journal of Forecasting.* Retrieved from: <https://doi.org/10.1016/j.ijforecast.2022.11.003>

Rockart, John (Mar 1979). *Chief Executives Define Their Own Data Needs.* Retrieved from: <https://hbr.org/1979/03/chief-executives-define-their-own-data-needs>

Rothstein, Robin (May 2024). *Housing Market Predictions For 2024: When Will Home Prices Be Affordable Again?* Retrieved from: [https://www.forbes.com/advisor/mortgages/real-estate/housing-market-predictions/#:~:text=Home%20prices%20are%20at%20or,Hepp%2C%20chief%20economist%20at%20CoreLogic](https://www.forbes.com/advisor/mortgages/real-estate/housing-market-predictions/)

Song, Y. Y., & Lu, Y. (2015). Decision Tree Methods: Applications For Classification and Prediction. *Shanghai Archives of Psychiatry.* Retrieved from: <https://doi.org/10.11919/j.issn.1002-0829.215044>