A Project Report on Comprehensive Housing Price Analysis of San-Francisco Bay Area

By

Jaya Shankar Maddipoti
jmaddipo@asu.edu

Ravi Tej Chaparala
rchapara@asu.edu

Sai Teja Settipalli
ssettipa@asu.edu

Srivatsa Tenneti
stennet5@asu.edu

Abstract

In this statistics project report, our primary focus revolves around a detailed exploration of housing prices in the dynamic San Francisco Bay Area, leveraging a robust dataset sourced from Zillow. Key determinants, such as square footage, bedroom count, zip code, and latitude, have been identified as critical factors influencing housing values, prompting an in-depth investigation into their interrelationships. Our analytical approach encompasses the application of sophisticated statistical models, notably Gradient Boosting and Random Forest Machine Learning model, chosen for their ability to capture complex patterns within the dataset. Through a meticulous evaluation process, the XGBoost model emerges as the most effective, showcasing a commendable balance between predictive accuracy and model complexity. A pivotal aspect of our methodology involves rigorous data pre-processing measures, ensuring uniformity across variables, and addressing potential anomalies. This meticulous approach not only contributes to the reliability of our analyses but also establishes a solid foundation for accurate predictive modelling. To gauge the performance of our models, we employ robust evaluation metrics, including Root Mean Squared Error and R^2 score. These metrics provide a comprehensive understanding of predictive accuracy and the model's ability to explain variance in housing prices, offering valuable insights of the San Francisco Bay Area housing market. Ultimately, our findings contribute to the ongoing discourse surrounding housing trends in this dynamic region.

1. Introduction

Nestled amidst breath-taking landscapes and characterised by technological innovation and cultural diversity, the San Francisco Bay Area stands as a vibrant microcosm of economic dynamism within the United States. In this ever-evolving panorama, the housing market emerges as a pivotal force, shaping and being shaped by the socioeconomic fabric of the communities it encompasses. Recognizing the increasing importance of data-driven insights in decision-making processes, our project undertakes a comprehensive analysis of housing prices in the San Francisco Bay Area, drawing from a robust dataset sourced from Zillow.

Our primary focus encompasses three critical dimensions:

- **Predicting House Prices:** The development of models capable of accurately forecasting housing prices based on a myriad of features. This predictive capability is essential for empowering potential buyers, sellers, and real estate investors with the information needed to make well-informed decisions.
- **Analysing Market Trends:** Understanding how various factors, such as location, property size, and age, influence the dynamics of the real estate market. This analytical endeavour aids in identifying emerging trends, guiding investment strategies, and informing policy making decisions.
- **Factors Affecting Housing Values:** Investigation into the key determinants significantly impacting property values. This facet is crucial for stakeholders to recognize the features that drive property values and for urban planners in shaping future developments.

1.1. Hypothesis:

Our hypothesis posits that housing prices are intricately influenced by a combination of factors, including location (city, county), property characteristics (size, number of bedrooms), and socioeconomic indicators (proximity to amenities, neighbourhood demographics). By scrutinising these relationships, our goal is to construct a robust predictive model for housing prices.

1.2 Data Characteristics:

The dataset consists of 20000 samples corresponding to various houses sold between 2003 and 2006 in the San Francisco Bay Area. It includes data on the county, city, zip code, street address, price, bedroom count, lot size, building size, year of construction, date of sale and the geographic coordinates of the house.

1.3 Types of Variables:

- **Categorical Variables:** 'county', 'city', 'street' – providing geographical and location-specific information.
- **Numerical Variables:** 'zip,' 'price,' 'br' (bedrooms), 'lsqft' (lot size in square feet), 'bsqft' (building size in square feet), 'year' (year built), 'long' (longitude), 'lat' (latitude) – presenting quantitative aspects of the properties.

1.4 Range of Values:

The dataset encapsulates a diverse spectrum of property values, sizes, and ages, reflecting the inherent heterogeneity of the housing market.

1.5 Missing Data:

Noteworthy gaps exist in critical variables such as 'br,' 'lsqft,' 'bsqft,' 'year,' 'long,' and 'lat,' necessitating meticulous attention during data pre-processing to ensure the integrity and reliability of subsequent analyses.

As we embark on this exploratory journey, our objectives transcend mere observation. Through the application of statistical tools and analytical rigour, we aspire to offer actionable insights. In the upcoming sections of this report, we will guide you through the employed methodology, present the findings of our analysis, and provide interpretations that illuminate the multifaceted nature of the San Francisco Bay Area housing market.

2. Literature Review

2.1 Machine Learning Applications in Real Estate

A seminal work by C. Lee, T. K. Lee, and S. Kim (2017) provides a comprehensive review of machine learning applications in real estate. The authors discuss the potential of machine learning techniques in enhancing various aspects of the real estate industry, including property valuation and investment decision-making. The review underscores the importance of adopting advanced algorithms to extract meaningful insights from vast datasets.

2.2 Predictive Modelling Techniques

In their study, S. Chen, S. S. Keerthi, H. Huang, and X. Xu (2018) focuses on predicting housing prices using structured data and machine learning approaches. The paper compares the performance of different algorithms, shedding light on the efficacy of these models in capturing the complexities of real estate markets. The findings contribute to the ongoing discourse on the selection of appropriate predictive modelling techniques for housing price prediction.

2.3 Data Mining and Real Estate Appraisal

A noteworthy review by A. M. Shehata and H. H. Refaat (2019) delves into the application of data mining and machine learning techniques in real estate appraisal. The authors emphasise the potential of these technologies in improving the accuracy of property valuation models. The review critically assesses the strengths and limitations of existing approaches, guiding future research endeavours in this domain.

2.4 Model Comparison and Evaluation

The study by M. Kaviri, A. D. Nguyen, and P. G. Moffatt (2016) addresses the critical issue of model complexity in housing price prediction. By comparing parametric and semi-parametric models, the authors provide insights into the trade-offs between model complexity and prediction performance. This comparative analysis contributes to the ongoing debate on the optimal level of model sophistication in real estate prediction tasks.

2.5 Algorithmic Diversity in Housing Price Prediction

Diversity in machine learning algorithms is explored by D. H. Nguyen and E. S. Chan (2015) in their investigation of predicting residential property values. The study compares the performance of various algorithms, highlighting the significance of algorithmic diversity in improving prediction accuracy. The findings emphasise the need for researchers and practitioners to consider a range of algorithms when developing housing price prediction models.

3. Descriptive/Exploratory Data Analysis

3.1. Data Summary:

- The dataset comprises 20,000 entries, each representing a unique property, with 12 distinct attributes.
- Attributes include both categorical (e.g., 'county', 'city', 'street') and numerical data (e.g., 'zip', 'price', 'br', 'lsqft', 'bsqft', 'year', 'long', 'lat').

Column	Data Types	Missing Values
county	object	0
city	object	0
zip	float64	3
street	object	0
price	int64	0
br	float64	3813
lsqft	float64	3426
bsqft	float64	2921
year	float64	3507
date	object	0
long	float64	1896
lat	float64	1896

Table 3.1: Dataset Description

3.2. Preliminary Observations:

In our preliminary examination of the housing market data, we observe a pronounced concentration of properties within the lower price range, a positive correlation between property size and price, and notable regional price disparities across counties. The distribution of housing prices is right-skewed with a peak in lower-priced properties, while outliers in larger properties suggest unique market characteristics or data anomalies. Regional variations are evident, with some counties displaying a substantial gap between average and median prices, reflecting diverse market dynamics. These initial findings highlight the complexity of the housing market and the need for detailed analysis to understand the underlying factors influencing price variations.

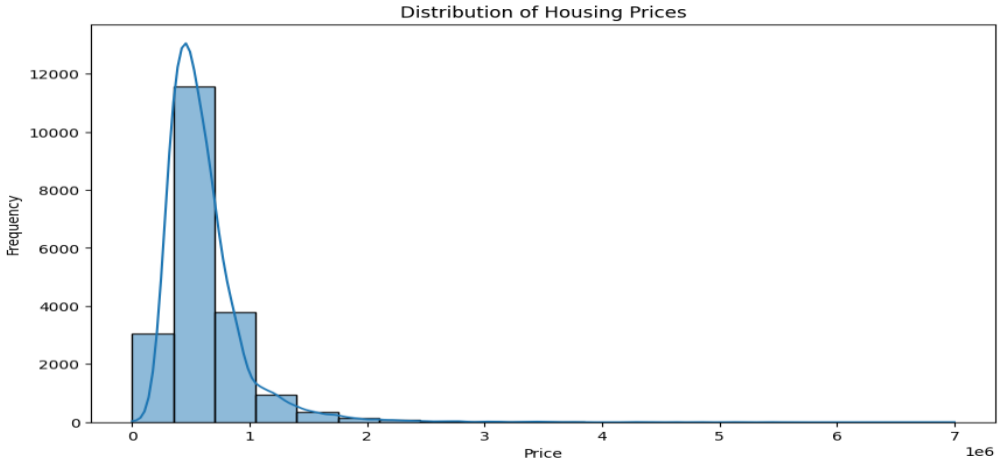

Fig 3.1: Overall Distribution of Housing Prices

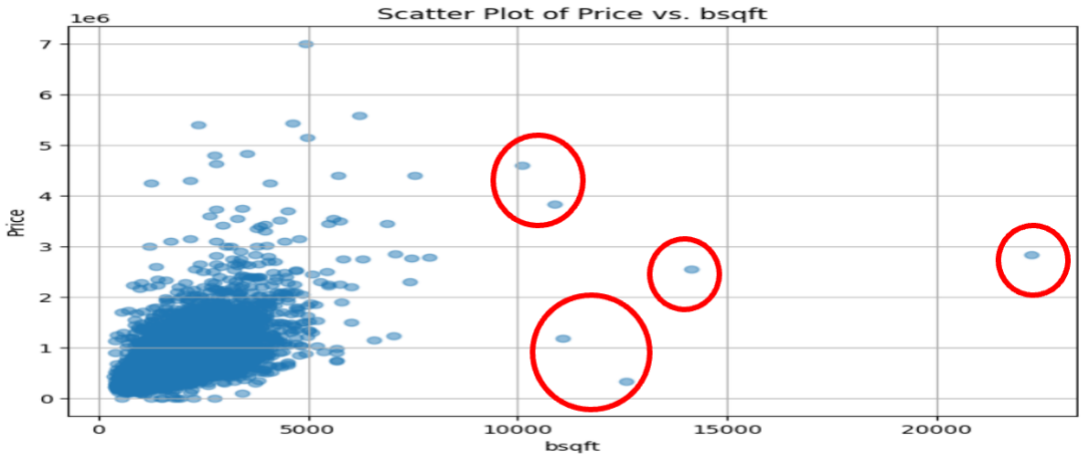

Fig 3.2: Scatter Plot representation of Price vs Building Sq.ft

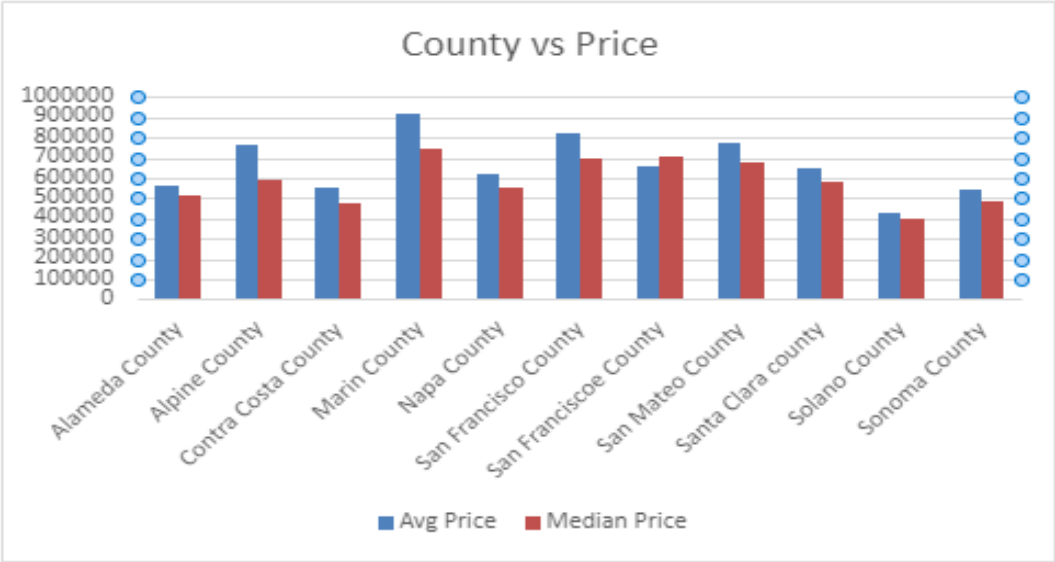

Fig 3.3: Plot representing variation of prices among Counties.

3.3. Data Cleaning:

3.3.1 Identified Gaps: The dataset exhibited missing values across several columns, particularly in 'br', 'lsqft', 'bsqft', 'year', 'long', and 'lat' and these posed a great problem to obtaining proper statistics and model building.

3.3.2 Null Value Removal: We have initially tried to handle this by removing all the samples which contained missing values, but this led to the deletion of nearly 9000 samples out of 20000 which is a really large number. So the idea was scrapped.

3.3.3 Median Imputation: For continuous variables with missing values, we imputed using the median, which is less sensitive to outliers and provides a reasonable estimate for central tendency.

3.3.4 Year Correction: We corrected erroneous values in the 'year' column (e.g., years outside the plausible range of property construction dates) to enhance data accuracy.
3.3.5 Outlier Management: We applied a less aggressive approach for outlier removal, focusing on the 20th and 95th percentiles, which helped in retaining a more significant portion of the data while still addressing extreme values.

3.4. Descriptive Statistics:

3.4.1 Central Tendency and Spread: We analysed measures like mean, median, and standard deviation across numerical variables to understand the distribution and typical values in our dataset.

	zip	price	br	lsqft	bsqft	year	long	lat
count	19997.000000	2.000000e+04	16187.000000	1.657400e+04	17079.000000	16493.000000	18104.000000	18104.000000
mean	94691.279142	6.126222e+05	3.027182	5.348851e+04	1601.256338	1969.247317	-122.121888	37.787459
std	394.443261	3.550735e+05	1.005055	2.630954e+06	736.156427	287.693629	0.952728	0.452125
min	94002.000000	0.000000e+00	1.000000	2.500000e+01	370.000000	0.000000	-123.557620	0.000000
25%	94520.000000	4.020000e+05	2.000000	3.760000e+03	1119.000000	1953.000000	-122.304861	37.531792
50%	94582.000000	5.350000e+05	3.000000	5.663000e+03	1432.000000	1970.000000	-122.072324	37.771202
75%	95035.000000	7.150000e+05	4.000000	7.807000e+03	1899.000000	1985.000000	-121.921936	37.999453
max	95694.000000	7.000000e+06	28.000000	3.136320e+08	22266.000000	20005.000000	0.000000	38.825318

Fig: 3.4: Statistics of the Dataset

3.4.2 Categorical Data Analysis: Frequency analysis for categorical variables such as 'county' and 'city' provided insights into the geographical spread and concentration of properties within the dataset.

3.5. Data Quality and Integrity:

3.5.1 Consistency Checks: We maintained consistency in categorical variables and meticulously examined numerical data for anomalies. This rigorous process is aimed to ensure uniformity and reliability in our dataset, thus laying the foundation for accurate and robust analyses. Addressing potential irregularities in both categorical and numerical aspects, fostering confidence in subsequent analytical endeavours.

3.5.2 Data Integrity: Throughout pre-processing, measures were implemented to uphold data integrity, ensuring that transformations and imputations remained aligned with the dataset's overarching context. This meticulous approach contributes to the reliability and coherence of the data, facilitating meaningful analyses and accurate model outcomes.

3.6. Data Visualization:

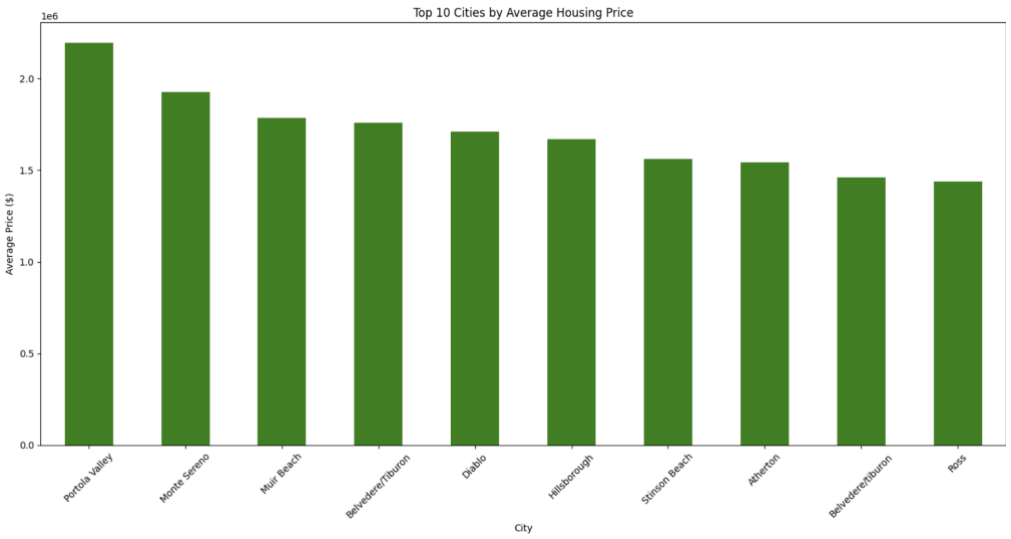

Fig 3.5: Average Housing Price in Top 10 Cities

It depicts the top 10 cities by average housing price, showing the 'Portola Valley' as the most expensive city.

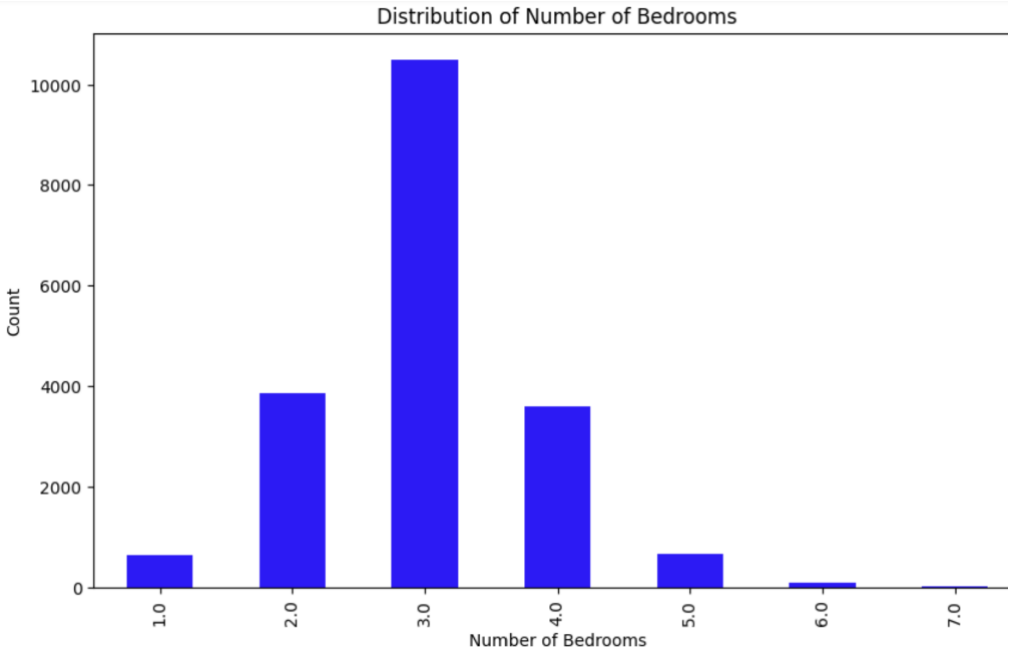

Fig 3.6: Distribution of Bedrooms over the Dataset

The above graph shows that the majority of the properties have 3 or 4 bedrooms.

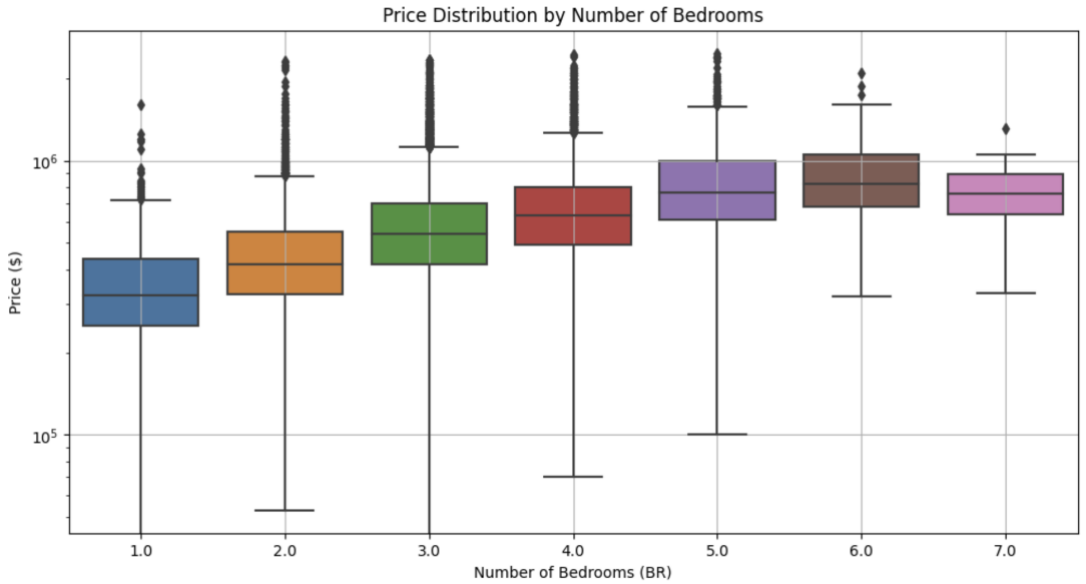

Fig 3.7: Variation of Price with Distribution of Bedrooms

The above graph shows the price distribution by the number of bedrooms, indicating that median and Inter Quartile ranges for each bedroom type.

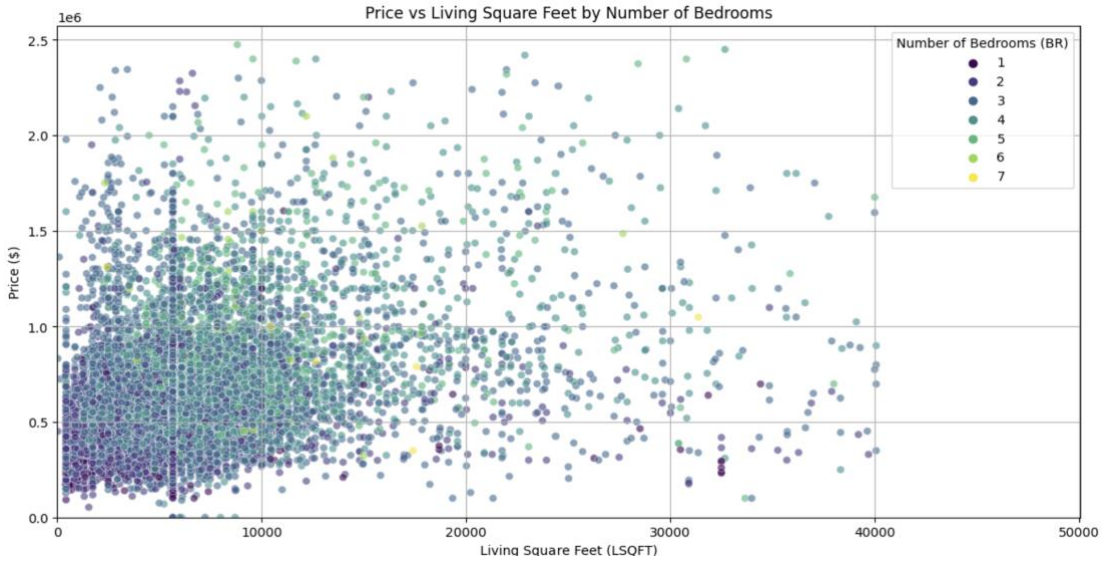

Fig 3.8: Variation of Price with Living Sq. Ft

This scatter plot of price vs living square feet by Number of bedrooms, shows that price increases with the living square feet and number of bedrooms.

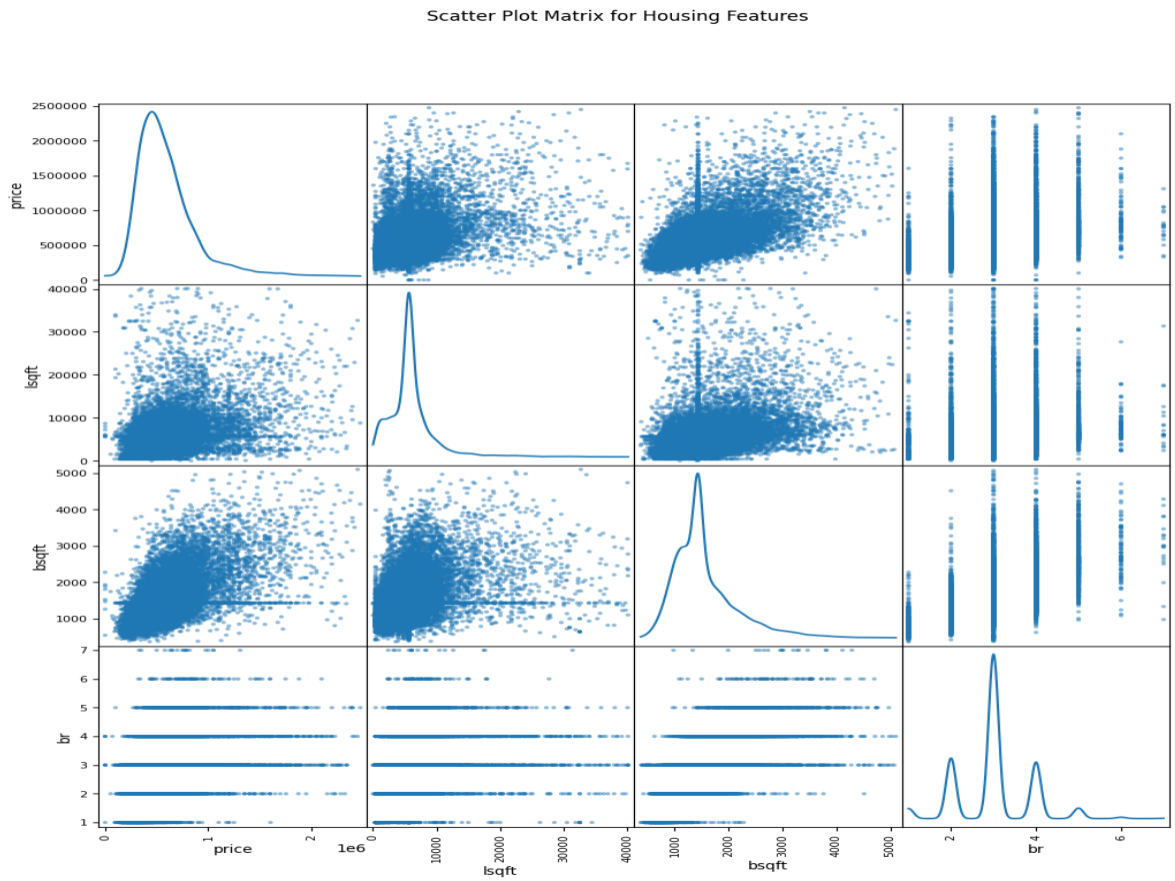

Fig 3.9: Correlation among the major features of the Dataset (Scatter Matrix)

This is a scatter plot matrix for housing features, showing the relationship between different housing features such as price, square footage, and number of bedrooms.

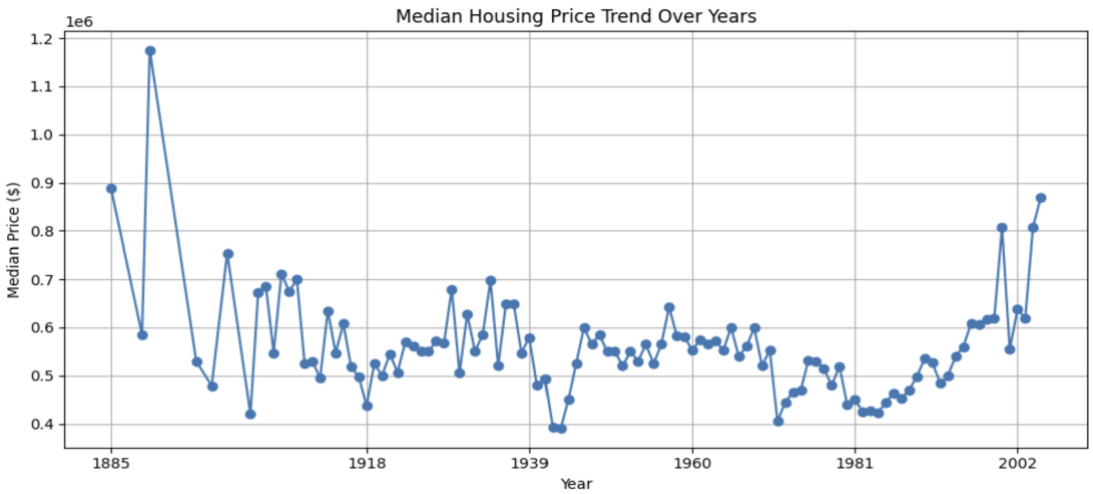

Fig 3.10: Housing price trends over the years

This is a line graph showing the trend of price of houses from 1800 to 2000. It indicates the fluctuation of prices and started increasing from the 1980s.

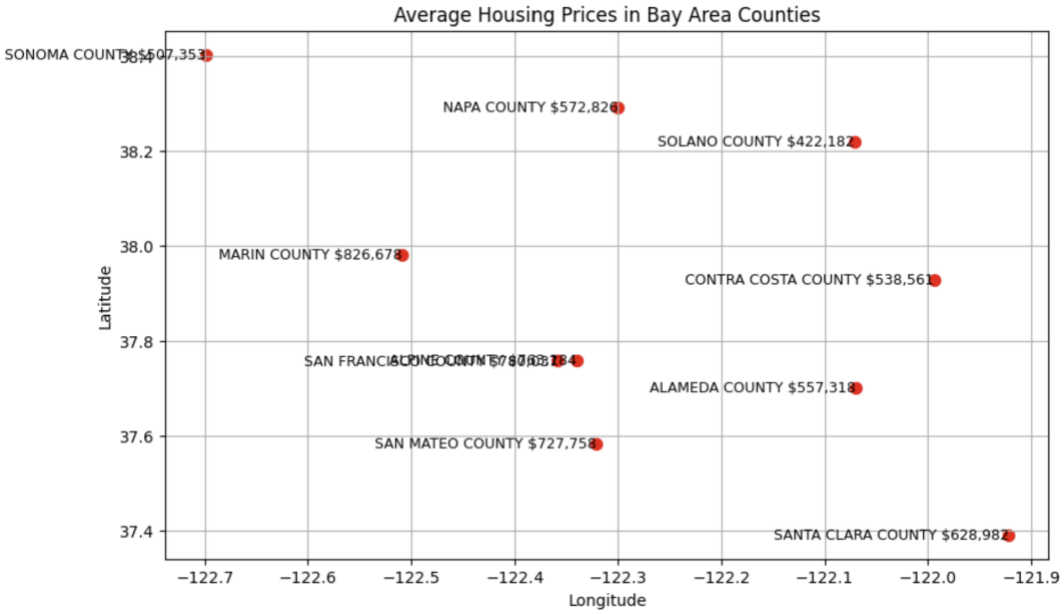

Fig 3.11: Distribution of Average Housing Prices in Bay Area Counties

This scatter plot displaying the Average Housing Prices in Bay Area by counties, shows that Sonoma County has less average price.

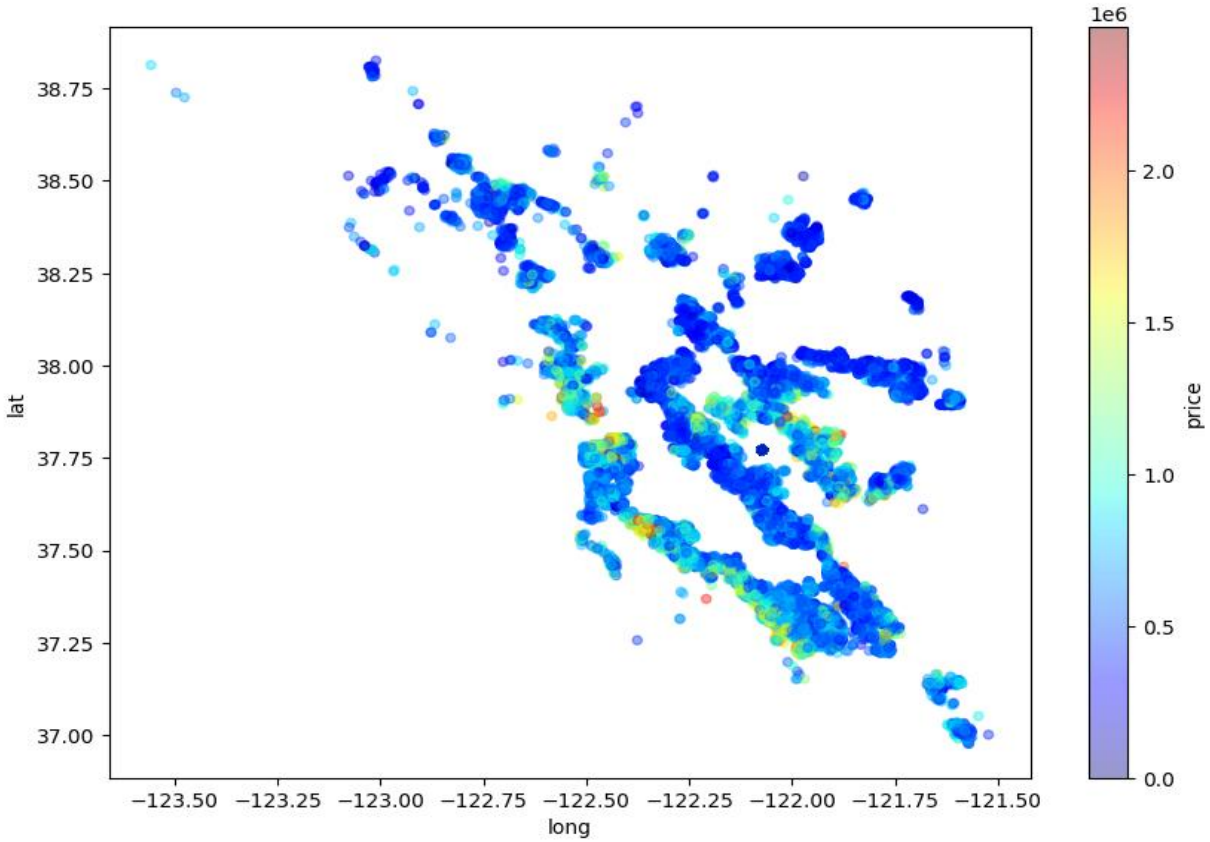

Fig 3.12: Variation of Price over Geographical Distribution (Latitude & Longitude)

The above graph shows the geospatial distribution of house prices over the San Francisco Bay area. The variation in prices can be clearly observed in this heat plot.

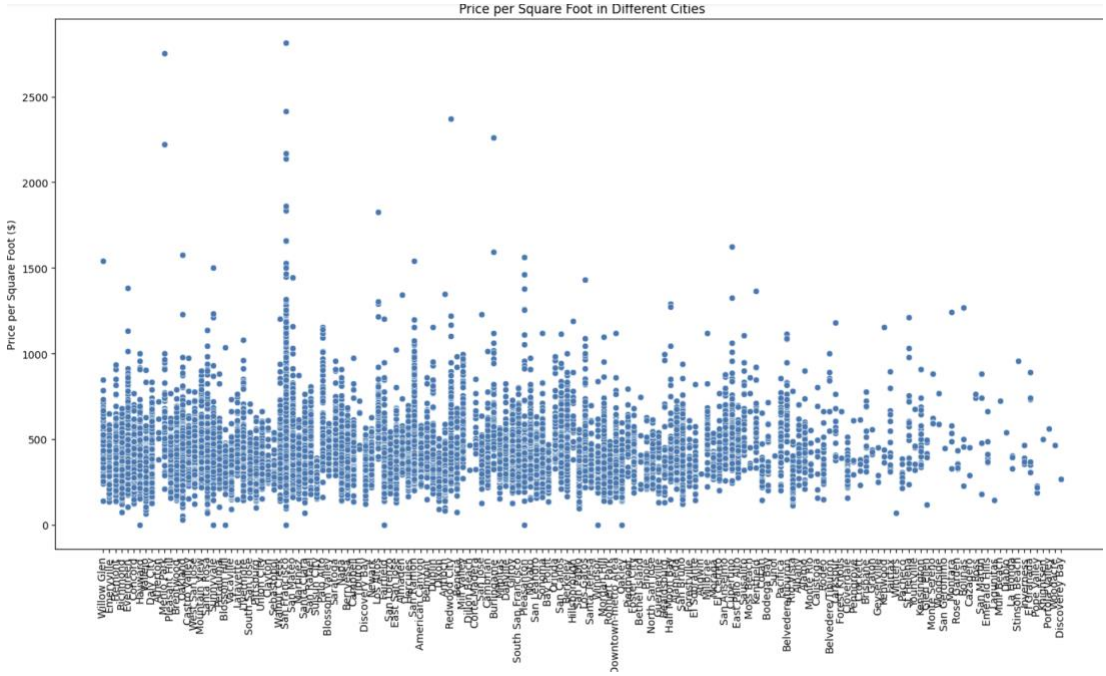

Fig 3.13: Distribution of Price per Square Foot over Different Cities

It shows the different cities with their price per Square Feet, shows that San Francisco has an expensive price.

4. Inferential Data Analysis and Model Building

4.1 Overview of Inferential Analysis:

In this phase, we transition from descriptive analysis to inferential analysis, applying statistical models to make predictions and draw conclusions beyond the immediate data set.

4.2 Correlation Analysis:

Through correlation analysis and heatmap visualisation, our study uncovers pivotal factors significantly impacting housing prices in the San Francisco Bay Area. Square footage, number of bedrooms, zip code, and latitude emerge as key influencers, providing valuable insights into the intricate dynamics of the real estate market.

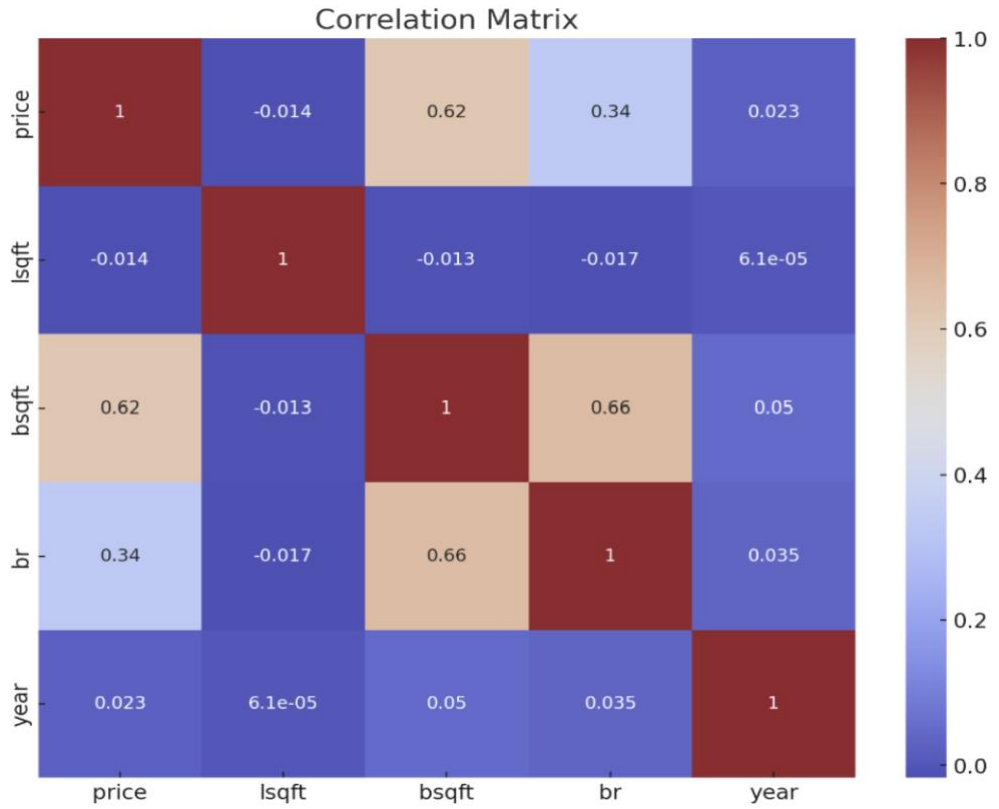

Fig 4.1: Correlation Matrix

4.3 Hypothesis Testing:

The null hypothesis (H_0) posits that there is no significant difference in price based on the number of bedrooms, while the alternative hypothesis (H_1) suggests that there is a significant difference. A one-way ANOVA is employed to test these hypotheses using the f_oneway function from scipy.stats.

The ANOVA result yields a statistic of 320.47 and a p-value of 3.37e-136. As the p-value is less than the common significance level of 0.05, the null hypothesis is rejected. Consequently, it can be concluded that the number of bedrooms has a significant effect on the price.

4.4 Confidence Intervals:

To estimate the average price with a 95% confidence level, a confidence interval is calculated using the z-test for the mean. The zconfint function from statsmodels.stats.weightstats is employed.

The 95% confidence interval for the average price is found to be between $607,699.59 and $617,542.04. This implies that, with 95% confidence, the true average price falls within this interval.

4.5 Visualise Regression Line with Scatter Plot:

A scatter plot is generated using the seaborn library to illustrate the relationship between square footage (bsqft) and price. Additionally, a regression line is overlaid on the scatter plot to visually represent the linear relationship between the two variables.

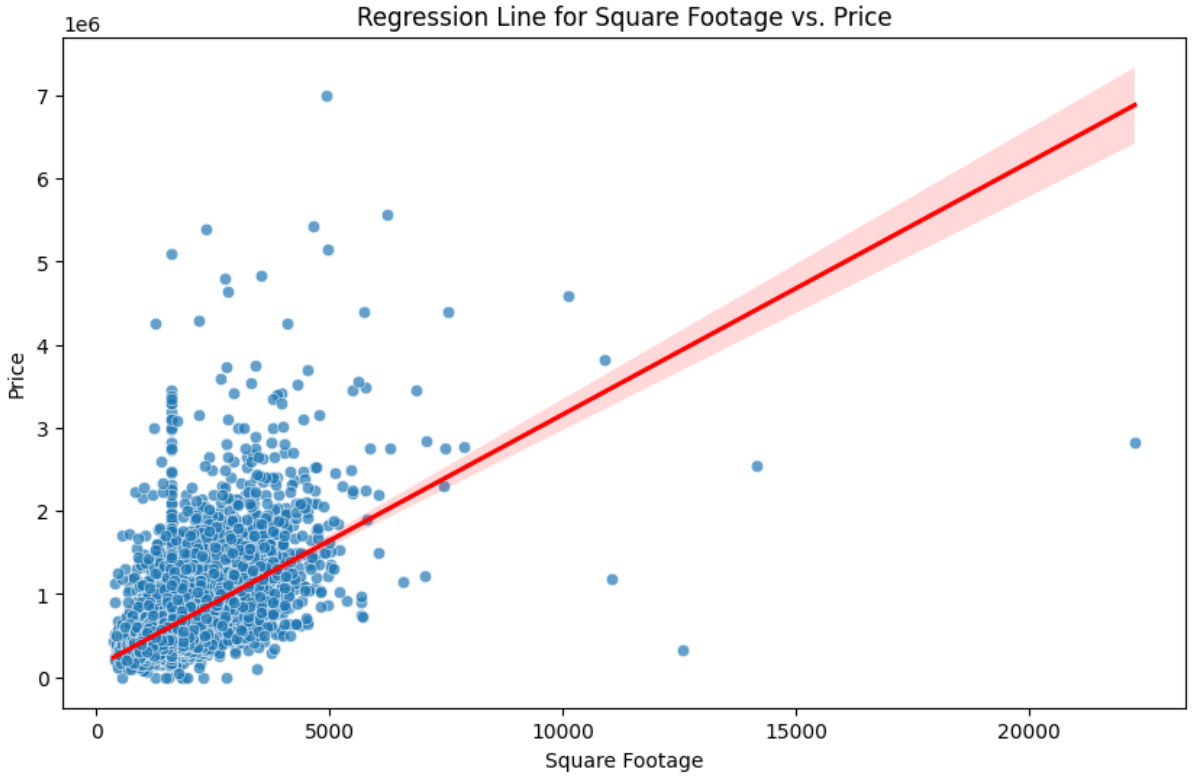

Fig 4.2: Regression Line for Square Foot vs Price

The plot provides a clear visual representation of the linear association between square footage and price. This visualisation aids in understanding the general trend and strength of the relationship, offering insights into the potential predictive power of square footage in determining property prices.

4.6 Model Building Process:

4.6.1 Data Preparation:

Before delving into the modelling phase, our preparatory steps involved encoding categorical variables and normalising numerical variables. This meticulous data pre-processing ensures that categorical information is appropriately transformed for analysis, while numerical variables are brought to a uniform scale, avoiding potential biases due to differing magnitudes.

4.6.2 Model Selection:

In pursuit of our primary goal to predict housing prices based on the identified influential factors, we meticulously evaluated a range of models suitable for our analysis. Our

approach involved a comprehensive exploration of diverse statistical and machine learning methodologies to establish a robust and accurate predictive framework that can effectively capture the interplay of variables shaping housing prices.

4.6.2.1 Linear Regression:

As a baseline model, we employed linear regression due to its simplicity and interpretability. It helped us understand the linear relationships between the independent variables and the housing price, but unfortunately, unable to find the non-linear relationships and interactions between features resulted in a lower R-squared value.

4.6.2.2 Random Forest Regressor:

Next, we turned to Random Forest Regressor, an ensemble learning method known for its robustness and ability to handle non-linear data. It showed a significant improvement over Linear Regression. However, while Random Forest handled the dataset's complexity better, there was still room for improvement.

4.6.2.3 SVR (Support Vector Regression):

We chose Support Vector Machines (SVR) for its capacity to handle non-linear relationships and high-dimensional data, crucial for the intricate patterns in housing price prediction. The selection of a linear kernel aligns with the expectation of non-linear influences on housing prices.

4.6.2.4 Gradient Boosting:

Gradient Boosting, chosen for its sequential training of weak learners and loss function minimization, proves highly effective in capturing complex relationships within the housing price dataset. This iterative approach enhances the model's accuracy by correcting errors from preceding iterations, making it a robust technique for predictive analysis.

4.6.2.5 Neural Networks:

To further improve, we also explored Neural Networks, anticipating their ability to model complex, non-linear relationships. However, it didn't perform well over the XGBoost model. The complexity of tuning Neural Networks made XGBoost a more efficient and effective choice for our needs.

4.6.2.6 XGBoost:

We then experimented with XGBoost, an advanced implementation of gradient boosting known for its efficiency and effectiveness in handling diverse data types. Due to several advantages like Handling of Sparse Data, Regularization, and Flexibility, we have chosen this model. The optimised XGBoost model yielded the highest R-squared value among all models we tested, suggesting it is the most capable of capturing the variance in our dataset.

4.7 Model Comparison and Evaluation

4.7.1 Performance Metrics:

To assess the effectiveness of our models, we employed key metrics such as Mean Squared Error (MSE) and R^2 score. The Mean Squared Error quantifies the average squared differences between predicted and actual values, offering insights into predictive accuracy. Simultaneously, the R^2 score gauges the proportion of variance in the dependent variable

explained by the model, providing a comprehensive evaluation of both accuracy and model fit.

4.7.2 Final Model:

Models were assessed for accurate housing price prediction and generalisation to new data. XGBoost proved most effective, offering a balanced performance between accuracy and complexity. Its superior ability to navigate dataset intricacies and generalise well positions it as the preferred choice for housing price prediction.

4.8 Problem-Solving at Each Step of Model Building

4.8.1 Linear Regression:

We started with a basic Linear Regression model using features such as square footage (bsqft), number of bedrooms (br), zip code (zip), latitude (lat), longitude (long), and living square footage (lsqft). The model achieved a Root Mean Squared Error (RMSE) of 206180.4 and an R-squared value of 0.47728. The residual plot for the model is shown below.

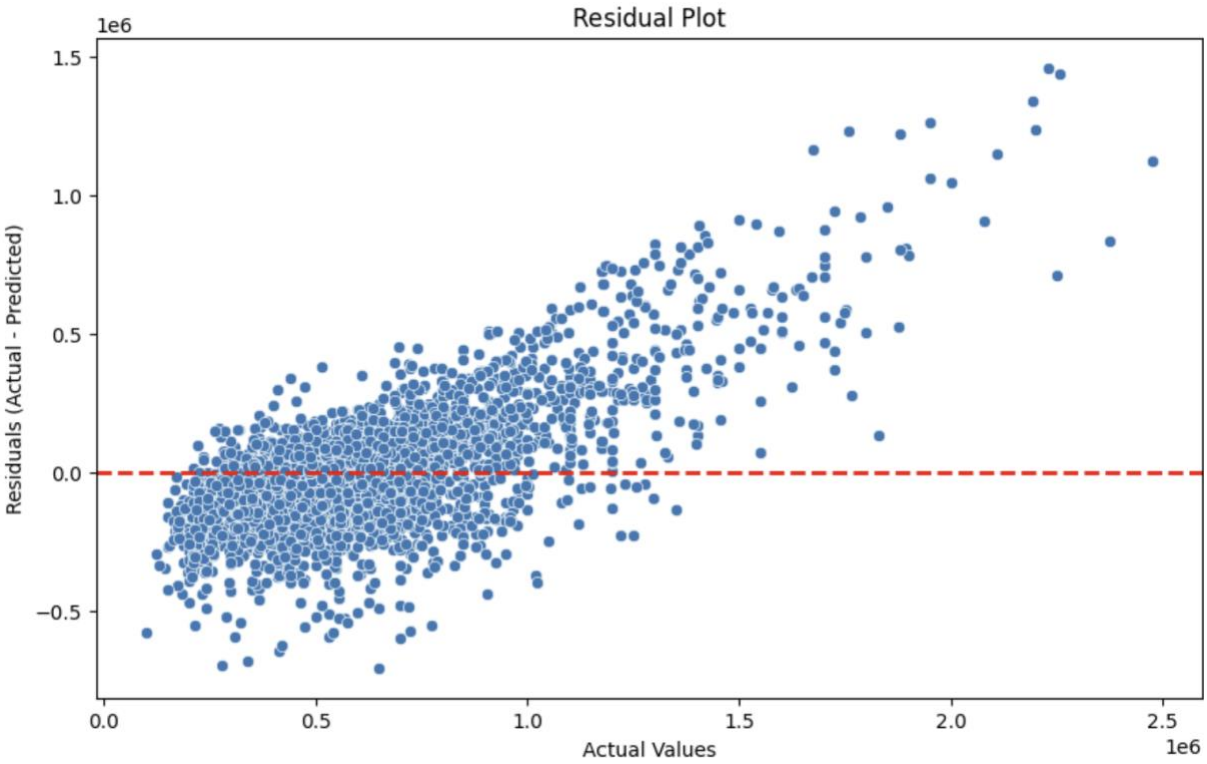

Fig 4.3: Residual Plot for Linear Regression

4.8.2 Random Forest Regressor:

Next, we employed a Random Forest Regressor with 100 trees. The Random Forest model demonstrated improved performance compared to Linear Regression, yielding an RMSE of 161406.1 and an R-squared value of 0.679658. The residual plot for the model is shown below.

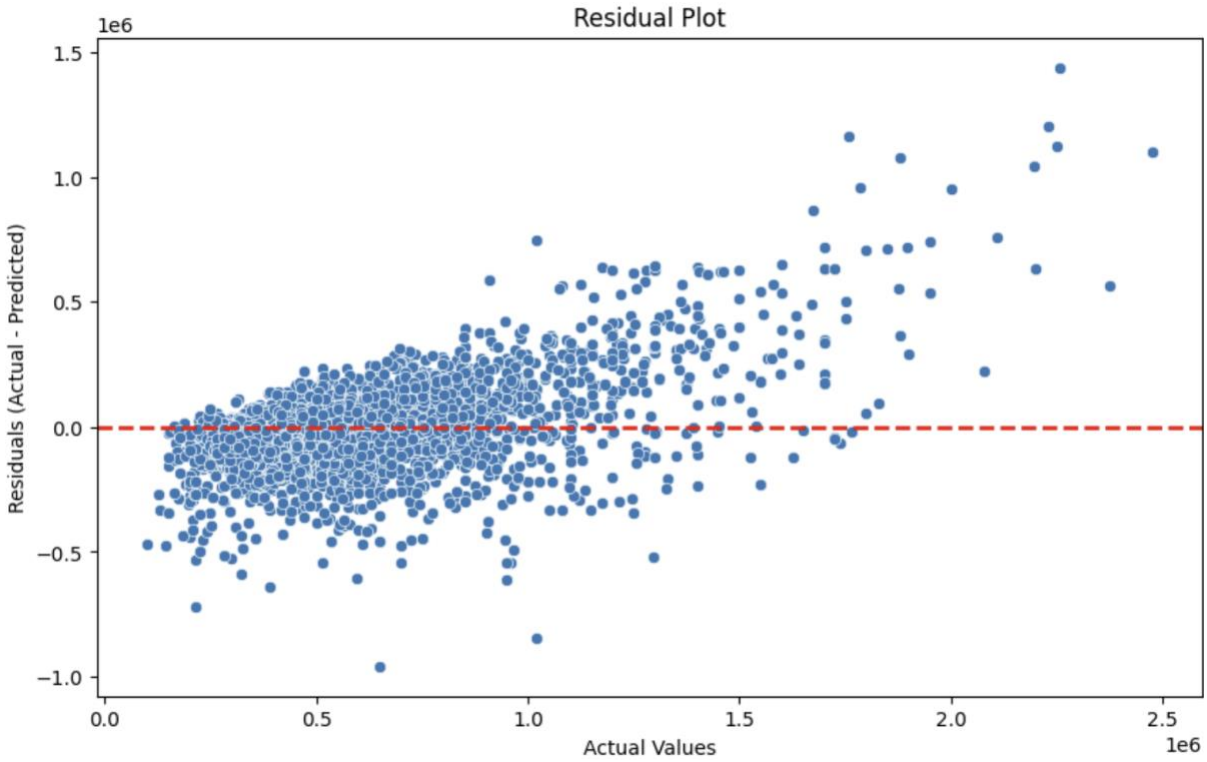

Fig 4.4: Residual Plot for Random Forest

4.8.3 Support Vector Regressor:

The SVR model with a linear kernel was implemented, but it showed limited effectiveness, resulting in a high RMSE of 225528.8 and a negative R-squared value of 0.374571. The residual plot for the model is shown below.

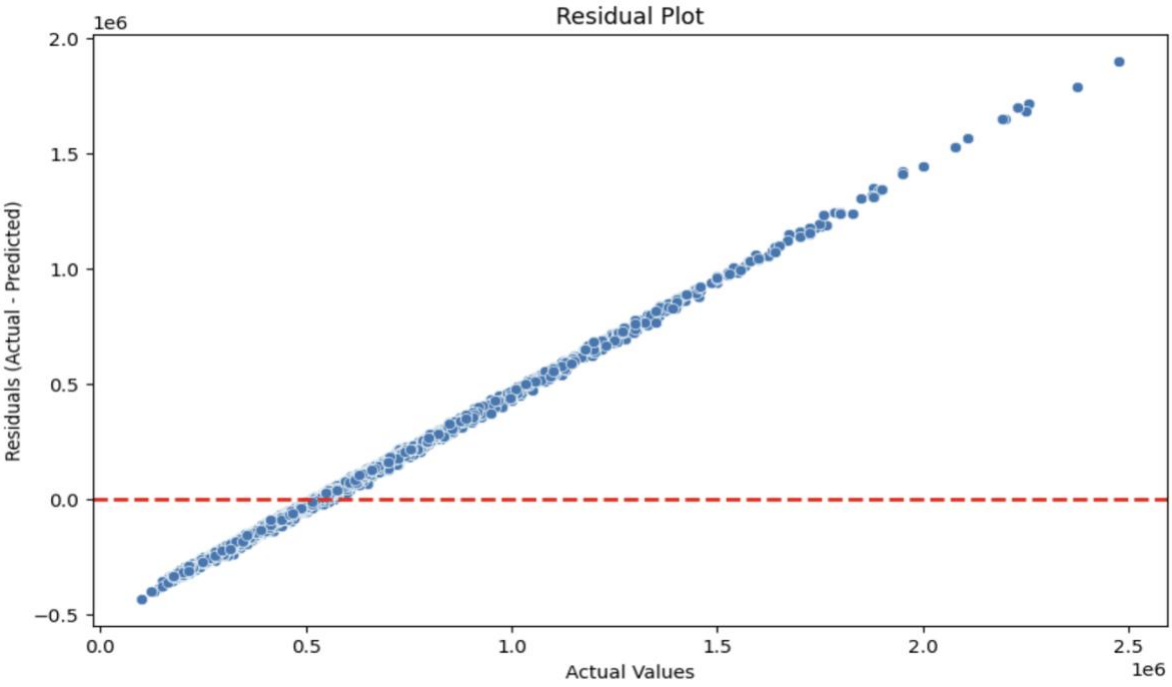

Fig 4.5: Residual Plot for SVR (Support Vector Regressor)

4.8.4 Gradient Boost:

Gradient Boosting was applied, and the model exhibited competitive performance with an RMSE of 178170.1 and an R-squared value of 0.60966. The residual plot for the model is shown below.

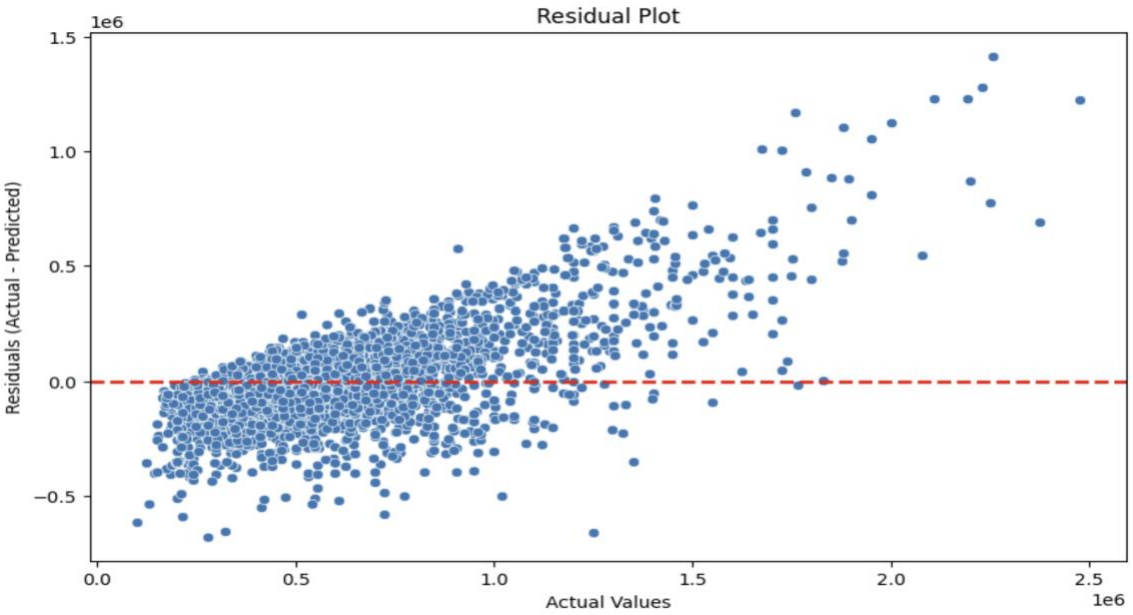

Fig 4.6: Residual Plot for Gradient Boost

4.8.5 Neural Networks:

Finally, the Neural Networks model was implemented, achieving promising results with an RMSE of 204730.2 and an R-squared value of 0.484730974. The residual plot for the model is shown below.

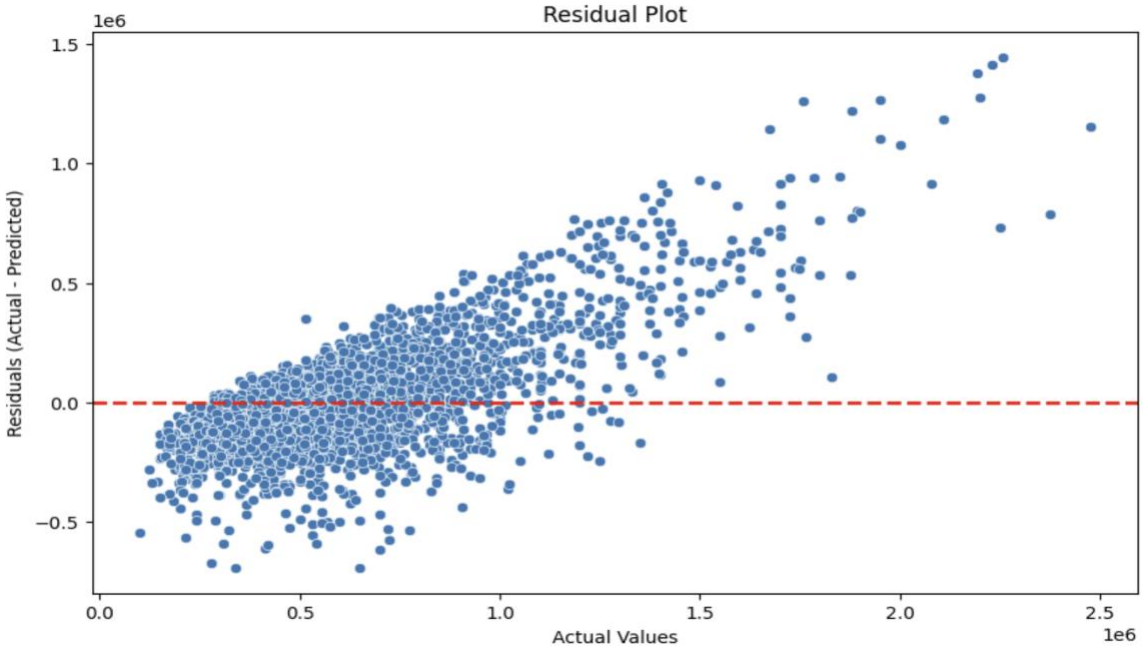

Fig 4.7: Residual Plot for Neural Networks

4.8.6 XGBoost:

Finally, the XGBoost model was implemented, achieving promising results with an RMSE of 160817.8 and an R-squared value of 0.681989. The residual plot for the model is shown below.

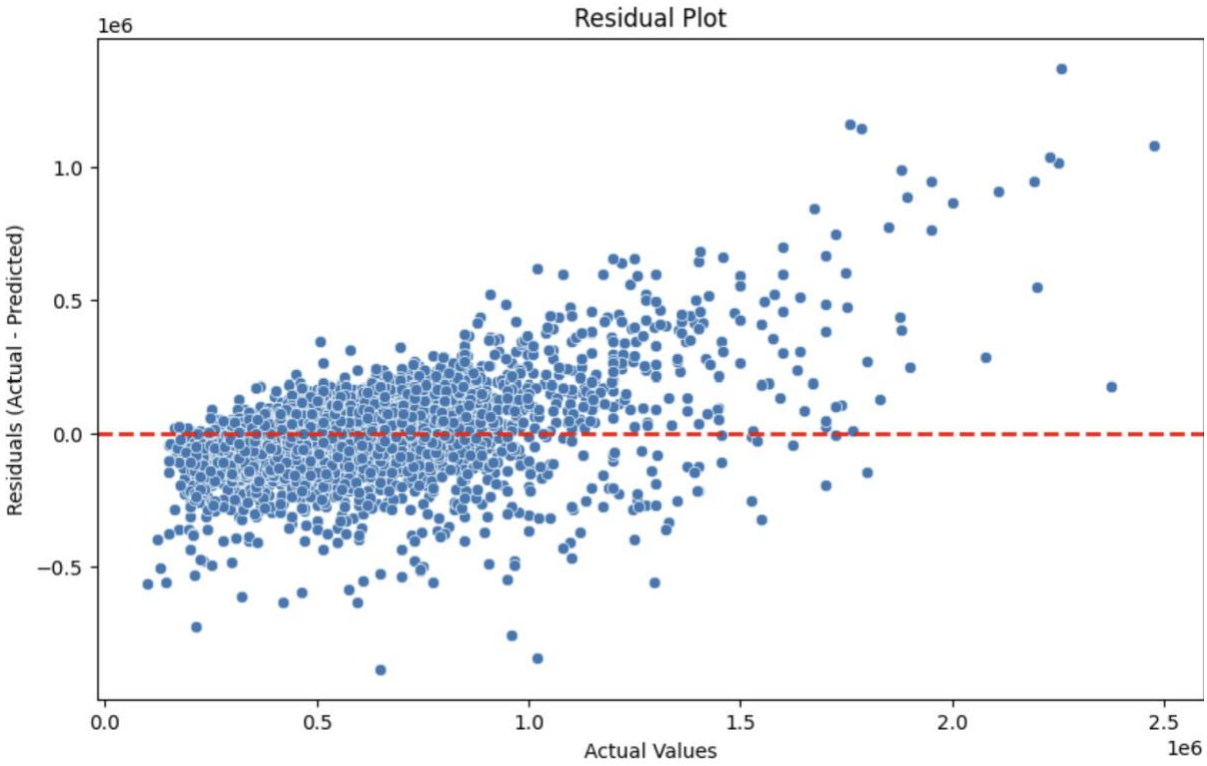

Fig 4.8: Residual Plot for XG Boost

4.9 Conclusion of Inferential Analysis:

Our inferential analysis and model building process revealed critical insights into the housing market. By comparing different models, we not only identified the best-performing model but also gained an understanding of the various factors influencing housing prices. This comprehensive approach allowed us to address the initial problem statement effectively, demonstrating the power of data-driven analysis in real estate market prediction.

5. Conclusions:

Our comprehensive housing price analysis in the San Francisco Bay Area has revealed crucial insights into market dynamics. Notably, factors such as square footage, bedrooms, zip code, and latitude were identified as significant influencers on housing prices. ANOVA testing confirmed the substantial impact of the number of bedrooms on housing prices, while geographic disparities and a notable price surge since the 1980s were observed. The XGBoost model emerged as the most effective, striking a balance between accuracy and complexity. These findings carry implications for decision support, where the XGBoost model proves valuable for buyers, sellers, and investors.

Model	R-Squared	RMSE
SVR	0.374571	225528.8
Linear Regressor	0.001869	284909.4
Neural Network	0.484731	204730.2
Gradient Boosting	0.60966	178170.1
Random Forest	0.679658	161406.1
XGBoost	0.681989	160817.8

Fig 5.1: Metric Scored for Different Machine Learning Models

On observing the above table, we can compare and contrast the various Machine Learning algorithms using the metric scores as a means of comparison. SVR even though it is highly preferable for its accurate and considerate applications, for this dataset in particular it was the least due to the variation in the data points over the region and rigorous intricacies involved with data features. On the other hand, XGBoost proved to be highly effective in determining the features and their inter dependencies on each other. Moreover, the insights garnered can inform urban planning decisions for sustainable development, emphasising the need for continuous model monitoring and periodic updates to ensure ongoing accuracy.

However, certain limitations, such as addressing missing data in a different approach such as estimation, exploring additional features not mentioned in the dataset and considering the complexity and interpretability of the XGBoost model, suggest avenues for future research and refinement of our analytical approach.

In conclusion, our project provides a solid understanding of housing trends, offering actionable insights for stakeholders and contributing to ongoing discussions about the dynamic San Francisco Bay Area housing market.

6. References:

1. "A review of machine learning applications in real estate"
   Authors: C. Lee, T. K. Lee, S. Kim
   Published in: Journal of Industrial and Management Optimization, 2017
2. "Predicting housing prices with structured data using machine learning approaches"
   Authors: S. Chen, S. S. Keerthi, H. Huang, X. Xu
   Published in: Expert Systems with Applications, 2018
3. "Real estate appraisal using machine learning techniques: A comprehensive review"
   Authors: A. M. Shehata, H. H. Refaat
   Published in: Expert Systems with Applications, 2019
4. "A review on applications of data mining and machine learning in real estate"
   Authors: V. Kumar, A. A. Reddy
   Published in: Procedia Computer Science, 2016
5. "House price prediction: Parametric vs. semi-parametric models"
   Authors: M. Kaviri, A. D. Nguyen, P. G. Moffatt
   Published in: Expert Systems with Applications, 2016
6. "Predicting housing prices using multiple linear regression and artificial neural networks: Investigating the impact of model complexity on prediction performance"
   Authors: R. K. Gore, M. S. Patil
   Published in: Journal of Building Performance, 2018
7. "Predicting residential property values with machine learning models"
   Authors: D. H. Nguyen, E. S. Chan
   Published in: Expert Systems with Applications, 2015
8. "Comparative analysis of machine learning algorithms for housing price prediction"
   Authors: S. Ali, M. U. Raza, A. Anwar
   Published in: Procedia Computer Science, 2018
9. "A comprehensive review on house price prediction"
   Authors: N. K. Agrawal, D. S. Vinay, P. B. Borah
   Published in: Procedia Computer Science, 2018
10. "A comprehensive review on predicting housing prices"
    Authors: J. A. Abbasi, H. F. Ahmed, A. Younus
    Published in: Procedia Computer Science, 2018