Housing Prices in San Diego, CA

Jacob A. Brubaker
San Diego State University
December 2024

Abstract

This project develops a predictive model for housing prices in San Diego using historical trends and price-per-square-foot data. By analyzing publicly available data and applying data science techniques, including regression models, this study identifies key patterns in property values over time. The analysis aims to assist future homebuyers and real estate professionals in making informed decisions. Key steps include data collection, cleaning, exploratory analysis, model development, and evaluation. The findings reveal trends that suggest a gradual increase in property values, providing valuable insights for navigating the dynamic San Diego housing market.

Approach

The main objective of this project is to develop a model that can predict future housing prices in San Diego based on historical and current data. This model will utilize price per square foot to ensure that some houses To achieve this, I followed a structured approach, divided into several stages:

Data Collection: The foundation of this project was a reliable dataset capturing housing trends in San Diego. After considering various sources, including Zillow, I selected a dataset from Kaggle that met key requirements: it included sale prices, dates, and property sizes. However, this dataset was published eight months ago and only contained data up to 2021, leaving out the years 2022–2024. This limitation introduced uncertainty into the model's ability to accurately forecast recent and future trends. Future iterations of this project could address this gap by integrating real-time data from platforms like Zillow or Redfin using APIs, or by sourcing updated datasets as they become available.

To ensure the data was usable for analysis, I focused on properties located specifically in San Diego, CA, filtering out entries from other locations like San Diego, TX. The dataset covered various property types, providing an overview of the housing market. This broader scope allowed the model to account for diverse housing trends beyond single-family homes, offering more generalizable insights.

The selected dataset provided a solid basis for exploratory analysis and modeling but highlighted the importance of keeping data current for predictive accuracy. Despite its limitations, the dataset served as an effective starting point for identifying historical trends in San Diego's real estate market.

Data Preparation: Preparing the data was a crucial step to ensure it was clean, consistent, and ready for analysis. Missing values, such as sale prices or property sizes, were removed altogether. I also identified and dealt with outliers by using methods like the interquartile range (IQR) and visualizing them with boxplots. Some of these were corrected, while others were removed entirely. To make the dataset more insightful, I created a new feature called price per square foot. For consistency, I scaled continuous variables like square footage and prices so they were on the same

scale, which helps some models perform better. These steps gave me a dataset that was both reliable and ready to use for building predictive models.

Model Selection and Development: After preparing the dataset, linear regression was selected as the primary predictive model for forecasting housing prices. While more advanced models like non-linear regression and random forests were considered for capturing potential nonlinear relationships, linear regression was ultimately used due to its simplicity, interpretability, and sufficient performance for the dataset at hand.

Evaluation and Iteration: The model was evaluated based on performance metrics like R-squared and Mean Squared Error. Continuous refinement, including feature optimization and hyperparameter tuning, was done to improve prediction accuracy. By following this approach, the goal is to develop a reliable and practical tool for predicting housing prices in the San Diego area, providing valuable insights for potential buyers and real estate professionals.

Data-Analysis

To start off my analysis, I needed a solid dataset with information on housing prices in San Diego. Initially, I thought about using platforms like Zillow, but I ended up finding what I needed on Kaggle. Kaggle offered a dataset that perfectly matched my needs, with detailed records of home sales, including price sold for, date sold, and size in square feet.

Unlike some analyses that focus exclusively on single-family homes, I worked with data covering various property types. This broader approach provided a more comprehensive view of the San Diego housing market. The dataset also spanned multiple years, making it ideal for identifying long-term trends.

After downloading the data, I focused on cleaning it. This included removing entries with missing or invalid values, particularly for sale dates and property sizes. By ensuring the dataset was complete and accurate, I set a solid foundation for exploratory data analysis (EDA) and built predictive models to forecast future housing trends.

Evaluation

Pandas & Data Frames

The Pandas library was a crucial tool throughout my analysis, providing an efficient way to handle and analyze the data. The DataFrame structure allowed me to easily organize the real estate data, where each row represented a property sale and each column contained key details like price, size, and year sold. This made it straightforward to filter and aggregate the data, helping me focus on the specific market I was interested in—San Diego, CA.

With Pandas, I could quickly compute new metrics, like price_per_sqft, which became a key factor in understanding trends in the market. Its groupby() function was especially helpful when looking at averages across different categories, like bedrooms or bathrooms, to identify patterns in property values. Overall, Pandas allowed

me to manipulate the data efficiently, which made the analysis smoother and faster. It was an essential tool in transforming raw data into actionable insights, helping me focus on the trends and relationships that were most important for the model.

Data Cleaning

After downloading the original dataset from Kaggle, I first reviewed the columns to identify those most relevant for my analysis and modeling. The dates were in a full format (e.g., YYYY-MM-DD), so I extracted just the year from the prev_sold_date column and created a new column called year_sold. This simplification helped focus the analysis on trends over time, rather than specific dates. I made sure the year_sold column was complete by removing any rows with empty values, ensuring the dataset was accurate and reliable. I also cleaned the price and size columns by eliminating rows with missing values to maintain data integrity.

Next, I calculated a new column, price_per_sqft, by dividing the price by the house size for each row, which became a key metric for understanding property values. Lastly, I filtered the data to include only properties located in San Diego, CA. This focused the analysis on the real estate market in San Diego and removed any properties from other cities, such as San Diego, TX, ensuring the analysis was geographically relevant and precise.

Exploratory Data Analysis

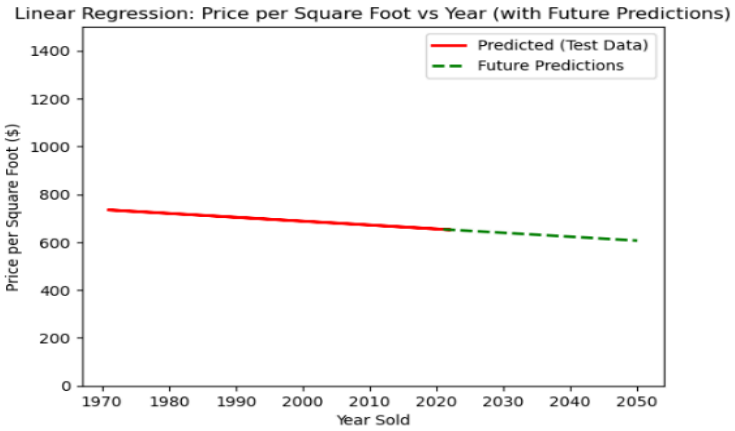
For the exploratory analysis, I began by examining the range of dates in the dataset, which spanned from 1971 to 2021. Given the wide range, I decided to simplify the analysis by focusing on the year sold rather than the full date. This allowed for a clearer view of the trends over time without getting restricted by minor specifics of each sale. The price_per_sqft column I created became central to uncovering patterns in property values, as it standardized the prices and allowed for easier comparison between properties of different sizes.

I also filtered the dataset to focus exclusively on properties located in San Diego, CA, ensuring that my analysis remained relevant to the area of interest. By narrowing the data in this way, I could better understand how the San Diego market had evolved over the years. This step helped me uncover important trends, such as changes in the average price per square foot and provided a solid foundation for future modeling and predictions specific to San Diego's real estate market.

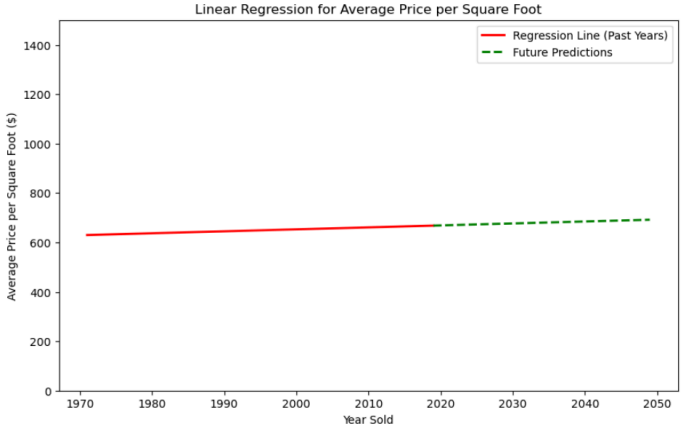
Linear Regression Model

To make a linear regression model, I chose to use the Scikit Learn library and their linear regression model to create a model that would estimate the future price per square foot of houses in San Diego. I fed the model “year_sold” as a target “X” and “price_per_sqft” as a target “y.” Using the linear regression model, I found that the average price per square foot has actually been on a slight decline over the years. I found this to be extremely surprising, yet after further thought, I realized that this trend could be influenced by several factors, such as a shift in demand for larger homes in the suburbs over smaller urban properties,

changes in the overall real estate market conditions, or even economic factors like inflation and interest rates that affect buying power.



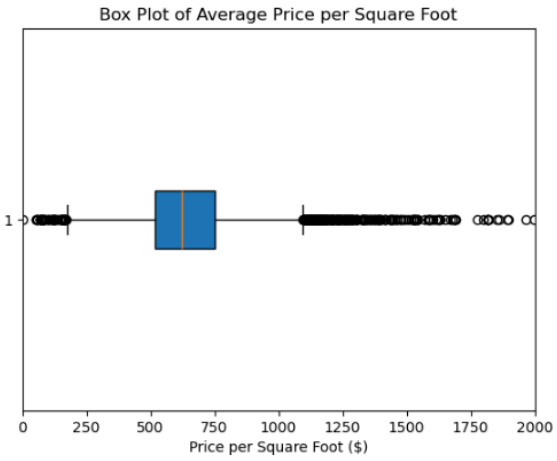
After seeing the model's prediction and the line going in a downward trend, I was slightly confused as I thought that overall, the average price per square foot should be on an upward trend. One possible explanation for this could be outliers that exist in the data, and after further investigation, I found that some years had some negative average price per square foot averages, causing the graph to have a downward trend. Because of this, I created a new model that checked for these negative values. Here is the new linear regression model that more accurately predicts the future average price per square foot in years to come:



This newer linear regression model reveals how overall, prices per square foot will continue to trend slightly upwards, while also revealing the importance of searching for outliers that may lead to unexpected results.

Visualization of the Data

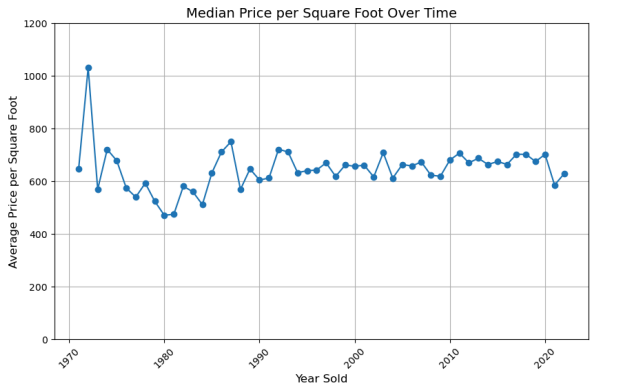
After doing some cleaning and EDA on the data, there were some interesting trends that I found. First off, I wanted to get an overall visualization of the data in boxplot form, just to be able to understand where the data should be centered around and visualize any outliers that may exist, here is the boxplot and some of the statistics:



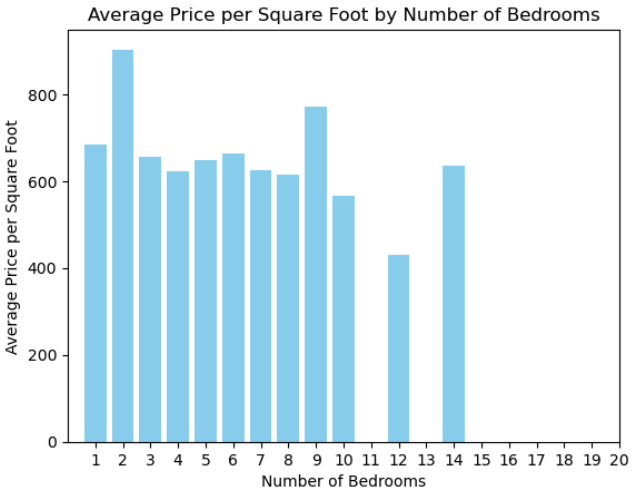
count	6494.000000
mean	657.132327
std	232.433604
min	1.166667
25%	519.500780
50%	623.326692
75%	749.687109
max	4794.520548

In this graph, we can see a lot of the important details around the most important feature of this research, the average price per square foot. I had to limit the x-axis to only display up to 2000 as the max was at 4797, which stretched out the visualization so

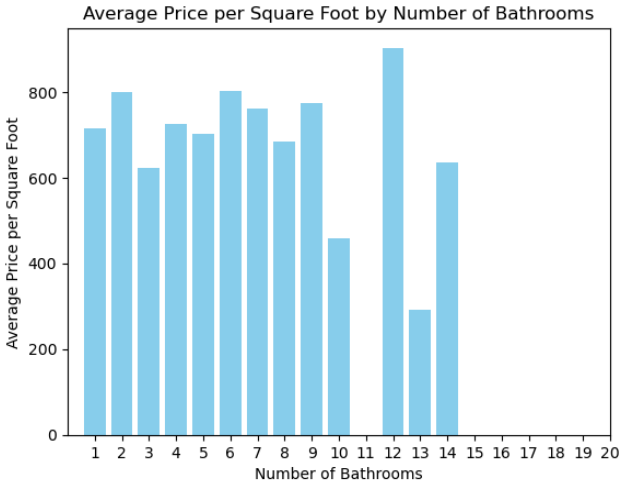
much that the graph was hard to make sense of. Some important things to note in this graph are that the overall mean for all years is 657.13, and 50% of the price_per_sqft is between 519.5 and 749.69. Moving forward, here is the graph of the Median Price per Square Foot over Time:



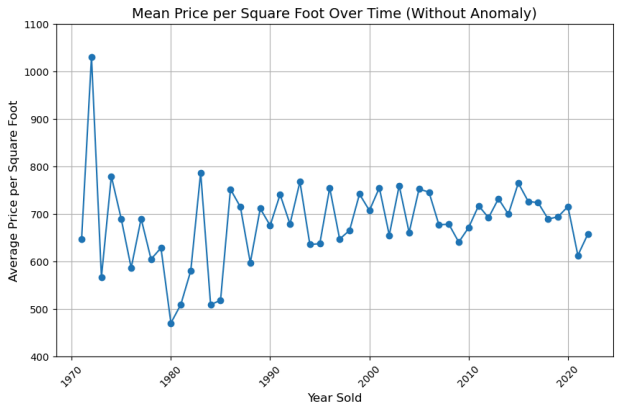
This line graph reveals the trends of how the median price of housing prices in San Diego has shifted over the years. There are a few visual spikes in terms of large jumps in average price per square foot over the years, however, an interesting visualization to see is the dropoff from 2020 to 2021, likely due to COVID-19.



This Bar graph reveals the average price per square foot by Number of Bedrooms. Surprisingly, the number of bedrooms in each house doesn't have that much of an effect on the price of the house after the second bedroom. The addition of a second bedroom reveals a large jump to around 900 dollars per square foot, but after, the value jumps back down to just over the 600 mark.



This bar graph reveals the average price per square foot by Number of Bathrooms. The trends reveal that having more bathrooms doesn't have a direct correlation between price per square foot and the number of bathrooms. There are slight jumps from one to two bathrooms and three to four but overall, there is no major trend to be observed. One important thing to note though is that more bathrooms, particularly 9, has a generally higher price compared to the others.



This line graph illustrates how average housing prices in San Diego have changed over the years, highlighting clear trends such as growth and declines. These shifts may be linked to major events like economic recessions, interest rate changes, or fluctuations in the housing market. Analyzing the data in this way helps identify patterns and understand how factors like population growth, job availability, or housing policies influence the market. While the overall trend is similar to the median graph, the mean graph exhibits more pronounced fluctuations. This makes it effective for uncovering significant trends and shifts over time.

References

Kaggle. (n.d.). *San Diego Real Estate Data*. Retrieved from https://www.kaggle.com/datasets

Pandas Development Team. (2020). *Pandas Documentation*. Retrieved from https://pandas.pydata.org/pandas-docs/stable/

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., et al. (2020). *Array programming with NumPy*. *Nature*, 585(7825), 357-362. https://doi.org/10.1038/s41586-020-2649-2

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). *Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research*, 12, 2825-2830. Retrieved from https://scikit-learn.org/

Conclusion

Based on the data analysis, visualizations, and the linear regression model, my prediction is that the average price per square foot in San Diego will continue to rise slowly over the next few years. This upward trend could be influenced by factors such as population growth, housing demand, inflation, and the overall economic climate in California. The model suggests a steady increase in property values, which may guide buyers and investors to expect gradual appreciation in the real estate market. However, it's important to note that these predictions are based on historical data, and unforeseen changes in the economy or market could affect the actual outcome. While the analysis provides valuable insights, potential buyers, sellers, and investors should remain cautious and consider external variables, such as interest rates or local policy changes, which could alter these predictions.

This project originated from my curiosity about San Diego's real estate market and how property values evolve over time. The goal was to create a predictive model for house prices in San Diego to help homebuyers make more informed decisions, and on a personal level, to give me a better understanding of the housing market as I plan to be a future homeowner. Given the volatility of the housing market and its dependence on various factors, I aimed to analyze current housing data to predict future price movements specific to San Diego. By using data-driven methods like linear regression and visualizations, I sought to uncover trends that could influence property investment decisions. The dataset

for this analysis was sourced from Kaggle, providing a comprehensive collection of real estate sales data, which enabled me to build a meaningful model and gain insights into the factors driving property values in the area.

In conclusion, this analysis serves as a tool for anticipating future shifts in the San Diego housing market. As the market continues to evolve, staying informed and leveraging data will be crucial for anyone navigating its complexities. I would also like to emphasize that all work presented in this project is entirely my own, with no external major projects used as a reference, other than the Kaggle dataset that provided the data for my analysis. All data cleaning and exploratory data analysis (EDA) were conducted specifically for this research, and the original dataset was created for a different purpose than the one I applied it to in this project.