

**House Price Prediction**

**About the dataset :-**

The USA property markets are a captivating playground for data analysts. By harnessing the power of data, we can unlock insights into future price trends, empowering informed decision-making for both buyers and sellers. Accurately predicting property prices is no longer a luxury, but a crucial tool for navigating these dynamic markets. After all, property values serve as a potent barometer of the overall market health and a nation's economic well-being.

This exceptional dataset, meticulously curated and made available on the renowned platform Kaggle, encompasses a vast collection of property sales records from Sydney and Melbourne the data has 4600 entries and 18 columns fortunately the data has no missing Values allowing us to dive straight into a rich exploration of market trends and price determinants.

**Columns :-**

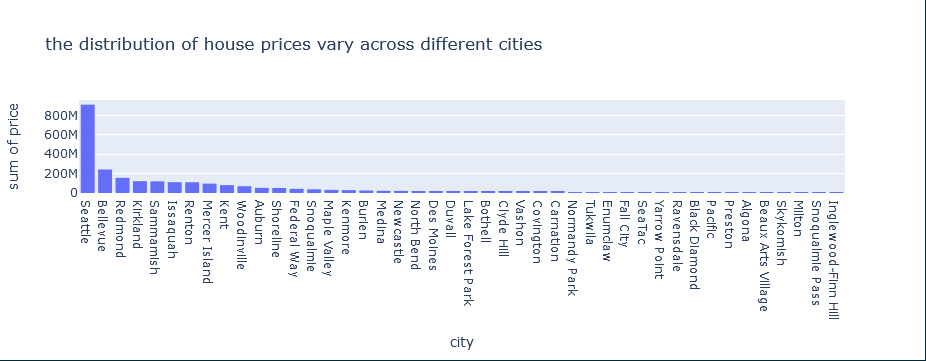
* **date (object):** This column captures the date of the property sale. It's important to note that this data type is stored as an object, so we might need to convert it to a datetime format for further analysis.
* **price (float64):** This crucial column holds the recorded sale price of the property. As it's a float data type, it can accommodate both whole numbers and decimals.
* **bedrooms (float64):** Here, we find the number of bedrooms each property boasts. Interestingly, this data type is also a float, suggesting the possibility of fractional bedrooms (e.g., studios with convertible spaces). We might need to investigate this further.
* **bathrooms (float64):** Similar to bedrooms, this column stores the number of bathrooms in each property. Again, the float data type might require clarification.
* **sqft\_living (int64):** This integer column holds the all-important square footage of the property's living area.
* **sqft\_lot (int64):** Here, we have the total square footage of the property's lot, including the living area and potentially additional outdoor space.
* **floors (float64):** This column captures the number of floors (including decimals for properties with basements or partial floors).
* **waterfront (int64):** A binary variable (coded as 1 or 0) indicating whether the property has waterfront access.
* **view (int64):** Similar to waterfront, this could be a binary value signifying the presence (1) or absence (0) of a desirable view.
* **condition (int64):** This numeric column likely represents the property's overall condition
* **sqft\_above (int64):** This integer column most likely captures the square footage of the finished living space above ground.
* **sqft\_basement (int64):** Here, we have square footage of the basement, providing insights into additional usable space.
* **yr\_built (int64):** This integer column stores the year the property was constructed.
* **yr\_renovated (int64):** This column likely holds the year (or absence of a value) when the property underwent renovations.
* **street (object):** This column likely contains the street address of the property. It's important to remember that due to privacy concerns, this data might be anonymized or partially obscured.
* **city (object):** Here, we find the city in which the property is located.
* **statezip (object):** This column combines state and zip code information. Depending on the data source, these might need to be separated for further analysis.
* **country (object):** A simple column confirming the country where the properties are located.

**Statistical Analysis:-**

* **Price:** The average property price is around $552,000, with a large standard deviation of over $563,000. This indicates a significant spread in prices, with some properties being much more expensive than others. The most expensive property in the dataset is listed for a whopping $26,590,000.
* **Bedrooms:** The average number of bedrooms is 3.4, with a minimum of 0 and a maximum of 9. The 25th percentile is 3 bedrooms, and the 75th percentile is 4 bedrooms. This suggests that most properties in the dataset have 3 or 4 bedrooms.
* **Bathrooms:** The average number of bathrooms is 2.16, with a minimum of 0 and a maximum of 8. The 25th percentile is 1.75 bathrooms, and the 75th percentile is 2.5 bathrooms. Similar to bedrooms, this indicates that most properties have 2 or 3 bathrooms.
* **Square Footage:** The average living area is around 2,140 square feet, with a standard deviation of over 960 square feet. The minimum living area is 370 square feet, and the maximum is 13,540 square feet. There is a large range in the sizes of the properties. The lot size also exhibits a large variation, with an average size of over 14,850 square feet and a standard deviation of over 35,880 square feet. The minimum lot size is 638 square feet, and the maximum is over 1,074,000 square feet.
* **Floors:** The average number of floors is 1.51, with a minimum of 1 and a maximum of 3.5. This suggests that most properties are single-story or two-story homes.
* **Waterfront:** Only a very small percentage (1%) of the properties have waterfront access.
* **View:** A similar proportion (24%) of properties have a view.
* **Condition:** The average condition rating is 3.45, on a scale of 1 (poor) to 5 (excellent). This suggests that most properties are in good condition.
* **Year Built:** The average year built is 1970, with a standard deviation of nearly 30 years. The minimum year built is 1900, and the maximum is 2014. This data suggests a mix of older and newer properties in the dataset.
* **Renovations:** The average year of renovation is 808, but there is a large standard deviation of nearly 980 years. The minimum year of renovation is 0, which likely indicates properties that have not been renovated. The data suggests that a significant number of properties have not been renovated

**Vizualization :-**

**1)**

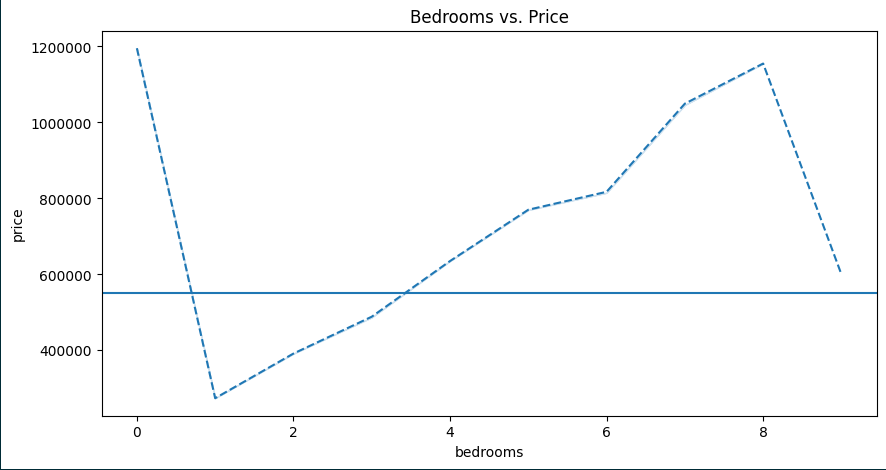
  
the plot above shows the distribution of prices across different cities we can clearly see that the Seattle tops the list compared to others this could be due to some reasons which are :-

* **High Demand:** Seattle is a major economic hub in the Pacific Northwest, attracting businesses from tech giants like Amazon and Microsoft. This strong job market fuels demand for housing, especially from professionals with higher salaries willing to pay more.
* **Low Unemployment:** Low unemployment rates in Seattle indicate a healthy economy with a high number of employed residents. This further strengthens buyer competition and potentially pushes prices higher.

**Nuances Compared to Neighboring Cities:**

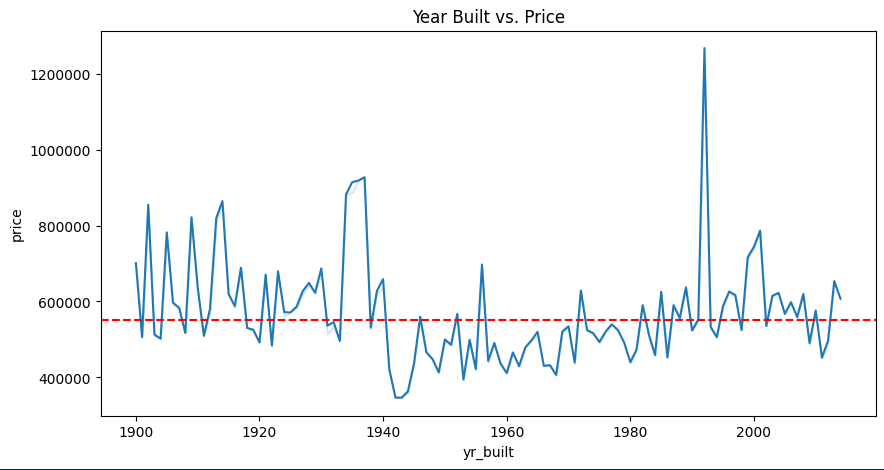
* **Bellevue, Redmond, Kirkland:** These Eastside suburbs offer a suburban lifestyle with good schools and amenities, attracting families who might be priced out of Seattle proper. However, they might still have high housing costs due to their proximity to major job centers and Seattle itself.
* **Sammamish, Issaquah, Renton:** These cities might offer a slightly more affordable alternative to Seattle, with potentially larger lots and a more suburban feel. However, their proximity to Seattle and major employers still contributes to their overall property values

**2)**

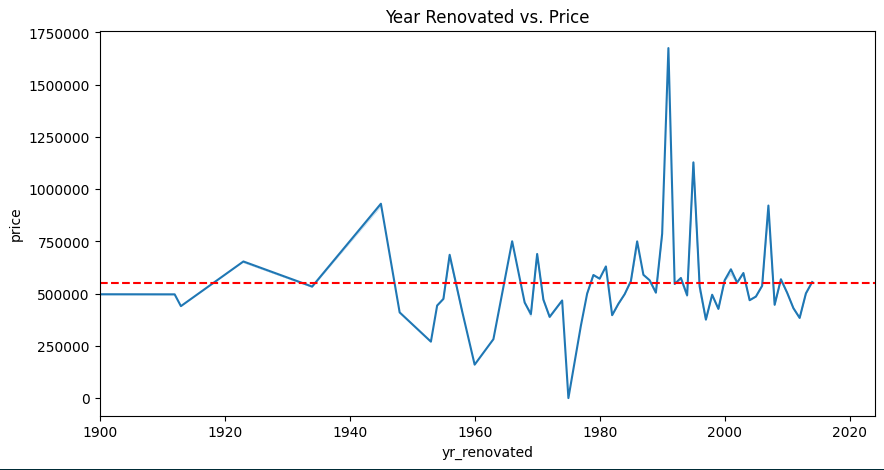


The above shows the relation between Bedrooms and price There's a generally positive correlation between the number of bedrooms and price. This indicates that properties with more bedrooms tend to cost more The horizontal line represents the average price across all the data points. This line serves as a benchmark to compare individual properties. Houses with data points above the line likely have more bedrooms and are presumably more expensive than the average, while those below the line have fewer bedrooms and might be more affordable

**3)**

here is a positive correlation between the year a house was built and its price. This means that newer houses tend to be more expensive than older houses. However, it is important to note that this is just a general trend and there may be many exceptions. For example, a luxury house built in the 1920s could still be more expensive than a small, starter home built in 2010.

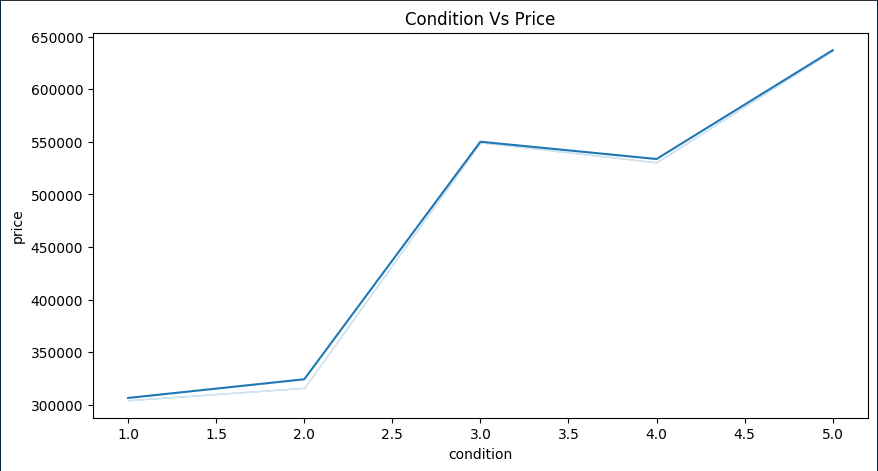
**4)**



There is a positive correlation between the year of renovation and the price. This means that houses renovated in more recent years tend to be more expensive than those renovated in earlier years. There could be a number of reasons for this, such as:

* The cost of renovation materials and labor has increased over time.
* More recent renovations may be more extensive or use higher quality materials.
* Renovated homes may be more energy-efficient, which can be a selling point for buyers.

**5)**

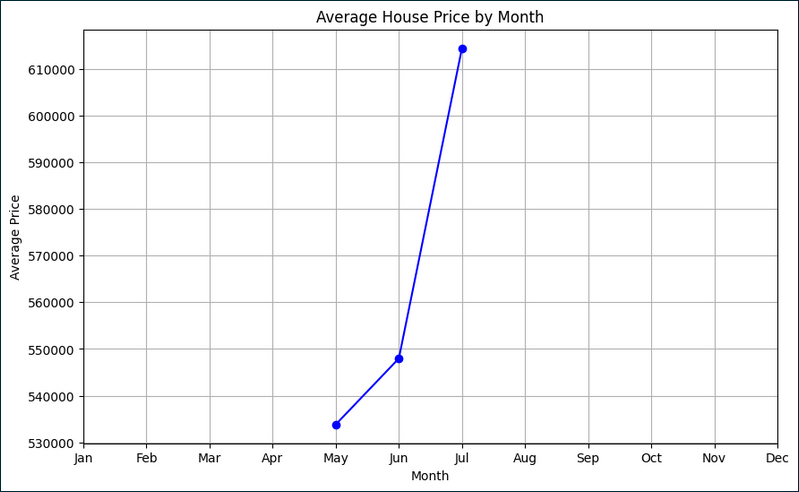


there is a positive correlation between the condition of the house and its price. This means that houses in better condition tend to be more expensive than houses in poorer condition. There are a few reasons for this. First, houses in better condition are likely to require less maintenance and repair, which can save buyers money in the long run. Second, houses in better condition may be more energy-efficient, which can also save buyers money on their utility bills. Finally, houses in better condition are more likely to be move-in ready, which means that buyers can avoid the hassle and expense of making repairs before they can move in..

Here are some additional insights I can glean from the graph:

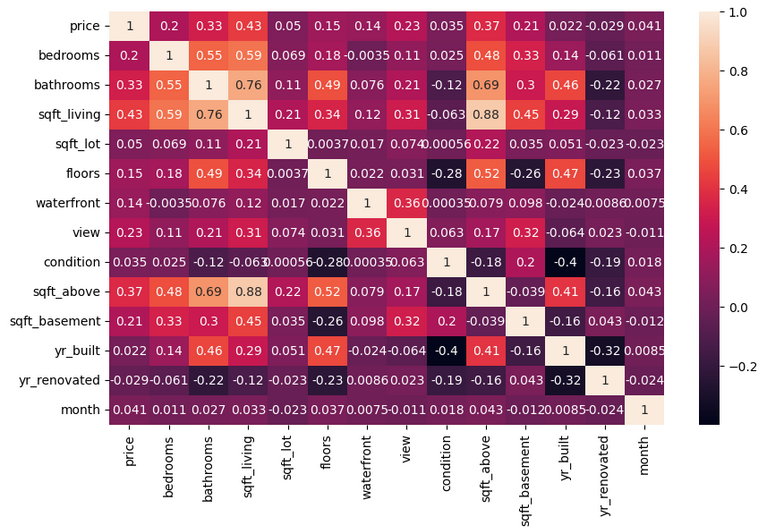
* The data points are spread out, which means there is a variation in price across different condition ratings.
* There isn't a perfectly straight line, so it's likely that the price isn't solely determined by condition. Other factors, such as location and size, also likely influence the price. For instance, a house in a desirable neighborhood may be more expensive than a house in a less desirable neighborhood, even if they are in similar condition. Similarly, a larger house, even in need of some repairs, may be more expensive than a smaller house in excellent condition.

**6)**



The image shows average house price by month however the data only contains May,June and July it shows upward trend with massive increase from june to July

**Statistical analysis :-**



* **Space Matters**: The correlation values reveal that the size of the living space (sqft\_living) has a strong positive relationship with the number of bedrooms (bedrooms), bathrooms (bathrooms), and the area above ground (sqft\_above). This suggests that larger houses tend to have more bedrooms, bathrooms, and overall living space. Homebuyers looking for spacious properties might find this correlation useful in their search.
* **Age vs. Condition**: The negative correlation between the year built (yr\_built) and the condition (condition) of the house indicates that older houses tend to be in poorer condition. This insight highlights the importance of considering the age of a property when assessing its overall condition. Renovation or maintenance may be necessary for older homes to maintain or improve their condition.
* **Renovation Impact**: Interestingly, there's a slight negative correlation between the year renovated (yr\_renovated) and the condition of the house (condition). This suggests that renovated houses may not always be in better condition than non-renovated ones. It's possible that renovations may focus more on aesthetics rather than structural improvements, impacting the overall condition differently.
* **Location Preference**: The correlation values don't directly include geographical features, but they indirectly suggest that certain features like waterfront (waterfront) and (Views) For instance, the positive correlation between waterfront, views and house prices (price) suggests that homes with waterfront location and scenic views tend to command higher prices, reflecting the preference for scenic locations.

**Feature selection :-**

1. **Data Preparation**:
   * I start by preparing the dataset, hp, by dropping columns that are not needed for the feature selection process. These columns include 'date', 'street', 'city', 'statezip', 'country', and 'price'.
   * The remaining columns are assigned to the variable X, which represents the features used for prediction.
   * The target variable, 'price', is assigned to the variable y.
2. **Feature Selection**:
   * I import the RFE (Recursive Feature Elimination) class from sklearn.feature\_selection and the SVR (Support Vector Regression) class from sklearn.svm.
   * i instantiate the SVR model with a linear kernel as the estimator for feature selection.
   * I specify n\_features\_to\_select=10, indicating that i want to select the top 10 features.
   * The step=1 parameter in RFE indicates that one feature is removed at each iteration.
3. **Feature Selection Process**:
   * I fit the RFE selector to ir feature matrix X and target vector y using the fit() method. This process evaluates the importance of each feature by training the SVR model and recursively eliminating the least important features until the desired number of features is reached.
4. **Output Processing**:
   * After fitting the RFE selector, I obtain the selected features using the get\_support() method, which returns a boolean mask indicating the selected features.
   * I use this boolean mask to filter the column names of X, resulting in a list of selected feature column names stored in selected\_cols.

### Output Explanation:

The selected\_cols variable contains the names of the top 10 features selected by the RFE algorithm. These features are considered to be the most important for predicting house prices based on the SVR model with a linear kernel. Here's a detailed explanation of each selected feature:

* **bedrooms**: The number of bedrooms in the house.
* **bathrooms**: The number of bathrooms in the house.
* **sqft\_living**: The total square footage of the living space.
* **floors**: The number of floors in the house.
* **waterfront**: A binary indicator (0 or 1) representing whether the property has a waterfront view.
* **view**: An index from 0 to 4 representing the quality of the view from the property.
* **condition**: An index from 1 to 5 representing the overall condition of the house.
* **sqft\_above**: The square footage of the house above ground level.
* **sqft\_basement**: The square footage of the basement.
* **yr\_built**: The year the house was built.

These selected features provide valuable insights into the factors that significantly influence house prices in the dataset. By focusing on these features, i can potentially improve the predictive performance of my model while reducing computational complexity and overfitting

**Linear Regression :-**

* **Data Preparation**:
  + I select the features ('bedrooms', 'bathrooms', 'sqft\_living', 'floors', 'waterfront', 'view', 'condition', 'sqft\_above', 'sqft\_basement', 'yr\_built') from the dataset hp and assign them to the feature matrix X.
  + The target variable 'price' is assigned to the target vector y.
* **Train-Test Split**:
  + I split the dataset into training and testing sets using train\_test\_split from sklearn.model\_selection.
  + The split ratio is set to 55% for testing (test\_size=0.45), and the random state is fixed for reproducibility (random\_state=111).
* **Model Training**:
  + I instantiate a Linear Regression model using LinearRegression from sklearn.linear\_model.
  + The model is trained on the training data using the fit() method.
* **Model Evaluation**:
  + The training and testing R-squared scores are calculated using the score() method on both the training and testing data.
  + Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are calculated using mean\_absolute\_error(), mean\_squared\_error(), and mean\_squared\_error(squared=False) respectively.

### Output Explanation:

* **Training R-squared score**: The R-squared value measures the proportion of variance in the target variable that is explained by the independent variables in the model. In this case, the training R-squared score is 0.1627, indicating that the model explains approximately 16.27% of the variance in house prices on the training data.
* **Testing R-squared score**: The testing R-squared score is 0.4530, which means that the model performs better on the testing data compared to the training data, explaining approximately 45.30% of the variance in house prices.
* **Mean Absolute Error (MAE)**: MAE represents the average absolute difference between the predicted and actual values. In this case, the MAE is approximately 166,455, indicating that, on average, the model's predictions are off by around $166,455.
* **Mean Squared Error (MSE)**: MSE measures the average of the squares of the errors. The MSE value is approximately 66,502,739,125, indicating the average squared difference between the predicted and actual values.
* **Root Mean Squared Error (RMSE)**: RMSE is the square root of the MSE, representing the average magnitude of the error. Here, the RMSE is approximately 257,881, indicating that, on average, the model's predictions are off by around $257,881.