Data Science & Analytics

Mountain House, CA

2024

Data-Driven Real Estate Pricing.

Shanmuk & Ikshit

TABLE OF CONTENTS

INTRODUCTION AND DATA REVIEW

INTRODUCTION:

THE REAL ESTATE MARKET IS INFLUENCED BY VARIOUS FACTORS, INCLUDING HOME SIZE, LOCATION, AND ECONOMIC CONDITIONS. ACCURATELY ESTIMATING HOUSE PRICES IS ESSENTIAL FOR BUYERS, INVESTORS, AND POLICYMAKERS TO MAKE INFORMED DECISIONS. HOWEVER, PRICING TRENDS CAN BE COMPLEX, FLUCTUATING DUE TO CHANGES IN DEMAND, NEIGHBORHOOD CHARACTERISTICS, AND ECONOMIC SHIFTS. THE KEY DETERMINANTS OF HOUSE VALUES WILL BE IDENTIFIED, AND VERY ACCURATE HOUSE PRICE ESTIMATES SHALL BE DEVELOPED IN THIS PROJECT, APPLYING STATISTICAL ANALYSIS AND MACHINE LEARNING MODELS. THESE INFLUENCES THAT CAN BE UNDERSTOOD WILL HELP US PROVIDE INSIGHTS TO INDIVIDUALS NAVIGATING THE HOUSING MARKET, OPTIMIZE INVESTMENT STRATEGIES, AND CONTRIBUTE TO BETTER DECISION-MAKING IN REAL ESTATE.

DATASET OVERVIEW:

DATASET NAME: HOUSING.CSV

SOURCE: OPEN-SOURCE REAL ESTATE DATA

SIZE: 21,613 ENTRIES

TIME FRAME: 2010-2024

KEY FEATURES: PRICE, SQUARE FOOTAGE, NUMBER OF BEDROOMS/BATHROOMS, CONDITION, YEAR BUILT, AND ZIP CODE.

OBJECTIVE: IDENTIFY PATTERNS IN PRICING AND KEY FACTORS INFLUENCING MARKET VALUE.

TECHNOLOGY STUDENT ASSOCIATION ®

	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	condition	yr_built	zipcode
count	2.161300e+04	21613.000000	21613.000000	21613.000000	2.161300e+04	21613.000000	21613.000000	21613.000000	21613.000000
mean	5.400886e+05	3.370795	2.114757	2079.899736	1.510697e+04	1.494309	3.409430	1971.005136	98077.939805
std	3.671268e+05	0.930105	0.770163	918.440897	4.142051e+04	0.539989	0.650743	29.373411	53.505026
min	7.500000e+04	0.000000	0.000000	290.000000	5.200000e+02	1.000000	1.000000	1900.000000	98001.000000
25%	3.219500e+05	3.000000	1.750000	1427.000000	5.040000e+03	1.000000	3.000000	1951.000000	98033.000000
50%	4.500000e+05	3.000000	2.250000	1910.000000	7.618000e+03	1.500000	3.000000	1975.000000	98065.000000
75%	6.450000e+05	4.000000	2.500000	2550.000000	1.068800e+04	2.000000	4.000000	1997.000000	98118.000000
max	7.700000e+06	33.000000	8.000000	13540.000000	1.651359e+06	3.500000	5.000000	2015.000000	98199.000000

THE DATASET REVEALS KEY INSIGHTS INTO HOUSING TRENDS, SHOWING THAT CENTRAL TENDENCY MEASURES LIKE THE MEAN AND MEDIAN HIGHLIGHT OVERALL MARKET TRENDS, WITH DIFFERENCES SUGGESTING PRICE SKEWNESS DRIVEN BY LUXURY PROPERTIES. HIGH STANDARD DEVIATIONS AND WIDE VALUE RANGES POINT TO SIGNIFICANT VARIABILITY IN PRICE, SIZE, AND FEATURES ACROSS HOMES. ADDITIONALLY, PERCENTILES CLASSIFY PROPERTIES INTO LOWER, MIDDLE, AND UPPER SEGMENTS, EMPHASIZING THE MARKET'S DIVERSITY.

DATA DICTIONARY

Field Name	Data Type	Description	Example Value
Price	Float	The sale price of the property in USD	538000.0
Bedrooms	Integer	The number of bedrooms in the property	3
Bathrooms	Float	The number of bathrooms in the property (including partial as fractions)	2.5
sqft_living	Integer	The total interior living space of the property in square feet	2570
sqft_lot	integer	The total land area of the property in square feet	7242
floors	Float	The number of floors in the property (including partial)	2.0
condition	int	The rating of the overall condition of the house from a scale of 1 to 5.	3
yr_built	int	Year the house was built.	1978
zipcode	int	Zip code of the house	95391

PURPOSE

PURPOSE:

HOUSE PRICE ESTIMATION PROVIDES VERY IMPORTANT RAW MATERIAL FOR THE FINANCIAL DECISIONS OF BUYERS AND INVESTORS, MARKET EFFICIENCY, AND POLICYMAKERS IN MAINTAINING ECONOMIC STABILITY. BESIDES THIS, EVERYTHING FROM ECONOMIC CYCLES TO GEOGRAPHICAL AND PROPERTY-SPECIFIC VARIABLES MAKES REAL ESTATE MARKETS HIGHLY VOLATILE. MISPRICING CAN HAVE GRAVE CONSEQUENCES MAY OVERPAY FOR HOMES, SELLERS MIGHT UNDERVALUE THEIR PROPERTIES, AND MARKET INEFFICIENCIES CAN ARISE THAT MAY LEAD TO HOUSING BUBBLES OR STAGNATION. ALSO, THE ABSENCE OF TRANSPARENCY IN PRICING MAKES DECISION-MAKING DIFFICULT. MANY TIMES, THE BUYERS FAIL TO COME UP WITH A GOOD MARKET PRICE, AND OFTEN THE SELLERS ALSO HAVE EITHER OVERESTIMATED OR UNDERESTIMATED PRICES OF THEIR REAL ESTATE. OTHER BROADER ECONOMIC FACTORS- INFLATION RATE, INTEREST RATES, AND DEMAND FOR HOUSING- FLUCTUATE, MAKING THESE CHALLENGES EVEN MORE DIFFICULT TO PREDICT AND FORECAST PROPERTY VALUES WITH CONFIDENCE.

SOLUTION:

A SOLUTION APPROACH TO THIS PROBLEM IS APPLYING DATA SCIENCE AND MACHINE LEARNING IN PREDICTIVE MODELING REGARDING THE ESTIMATE OF A HOUSE PRICE. SUCH A MODEL CAN BE MADE TO ARRIVE AT HIGHLY ACCURATE, DATA-DRIVEN ESTIMATES OF THE SELLING PRICE BY DRAWING ON CRITICAL PROPERTY ATTRIBUTES: SQUARE FOOTAGE, THE NUMBER OF BEDROOMS AND BATHROOMS, LOCATION, AND RECENT SALES TRENDS. BESIDES, THOSE MODELS CAN EASILY INCORPORATE OTHER, EXTERNAL FACTORS- MACROECONOMIC CONDITIONS AND TRENDS IN HOUSING DEMAND-TO FURTHER INCREASE THEIR PREDICTIVE STRENGTH. THE REALIZATION OF SUCH DATA-DRIVEN SOLUTIONS BOLSTERS PRICING TRANSPARENCY, OFFERING PURCHASERS INSIGHT INTO JUST VALUE AND GIVING SELLERS THE AMMUNITION THEY NEED FOR COMPETITIVE PRICING. INVESTORS CAN THEREFORE RECOGNIZE PROFITABLE OPPORTUNITIES PRESENTED BY THESE ANALYTICAL MODELS, WHEREAS POLICYMAKERS CAN DEVISE STRATEGIES TO CREATE MORE AFFORDABLE BUT STABLE HOUSING MARKET CONDITIONS ACCORDINGLY. IN OTHER WORDS, DATA-DRIVEN PREDICTIVE MODELING SIMPLY ENHANCES GENERAL EFFICIENCY IN THE REAL ESTATE MARKET BY GRANTING ALL ITS PLAYERS RELIABLE PRICE INSIGHT.

METHODS

1. DATA COLLECTION

WE SEARCHED KAGGLE FOR A DATASET CONTAINING REAL ESTATE TRANSACTIONS WITH DETAILED PROPERTY ATTRIBUTES AND SALE PRICES.

THE SELECTION CRITERIA FOR THE DATASET INCLUDED:

- **SUFFICIENT DATA POINTS:** THE DATASET NEEDED THOUSANDS OF RECORDS FOR STATISTICAL ACCURACY.
- **DIVERSE FEATURES:** KEY PROPERTY ATTRIBUTES SUCH AS SQUARE FOOTAGE, NUMBER OF BEDROOMS, LOCATION (ZIP CODE), AND CONDITION.
- **MARKET VARIATION:** PROPERTIES FROM DIFFERENT PRICE RANGES AND LOCATIONS TO ENSURE MARKET-WIDE REPRESENTATION.

AFTER EVALUATING MULTIPLE OPTIONS, WE SELECTED THE HOUSING.CSV FROM KAGGLE, WHICH CONTAINED OVER 21,000 RECORDS SPANNING DIFFERENT HOUSING MARKETS.

2. DATA CREDIBILITY

BEFORE USING OUR DATA SET FOR VALIDATION, WE EVALUATED ITS CREDIBILITY BASED ON 3 FACTORS TO ENSURE ACCURACY, RELIABILITY, AND RELEVANCE TO THE HOUSING MARKET.

- **SOURCE VALIDATION:** VERIFIED THE AUTHOR WHO PUBLISHED THE DATASET. LOOKED AT THE POPULARITY LIKE THE NUMBER OF DOWNLOADS AND UPVOTES TO SEE IF IT IS WIDELY USED. ALSO CHECKED KAGGLE COMMENTS FOR FEEDBACK FROM OTHER DATA ANALYSTS
- CROSS-CHECKING: WE CROSS-CHECKED AVERAGE HOME PRICES WITH SOURCES LIKE ZILLOW AND REDFIN. USED THIS TO ENSURE THAT PROPERTY ATTRIBUTES ALIGNED WITH REAL-WORLD TRENDS.
- **BIAS & ETHICAL CONSIDERATIONS:** WE VERIFIED IF THE DATA SET CONTAINED A DIVERSE RANGE OF PROPERTY TYPES AND LOCATIONS. SO IT WASN'T SKEWED TOWARD LUXURY PROPERTIES OR LOW-INCOME HOUSING.

3. DATA PREPROCESSING

ONCE THE DATASET WAS OBTAINED WE HAD TO CLEAN, TRANSFORM, AND STRUCTURE THE DATA PROPERLY. THIS WAS DONE IN JUPYTER NOTEBOOK.

- **HANDLING MISSING VALUES:** WE HAD TO IDENTIFY MISSING DATA THAT WOULD NEGATIVELY AFFECT THE ACCURACY OF OUR TESTS. FORTUNATELY, OUR DATASET WAS COMPLETE, WITH NO MISSING VALUES TO HANDLE OR REMOVE.
- **DIMENSIONALITY REDUCTION:** THERE WAS A LOT OF UNNECESSARY DATA THAT WAS INCLUDED IN THE DATA SET WE ORIGINALLY GOT FROM KAGGLE. THINGS LIKE WATERFRONT VIEW AND GRADE WERE UNNECESSARY AND WOULDN'T AFFECT OUR DATA.

```
# Define the columns to drop (use actual column names)
columns_to_drop = ['id', 'date', 'waterfront', 'view', 'grade', 'sqft_above', 'sqft_basement', 'yr_renovated', 'lat', 'long', 'sqft_living15',

# Drop the columns
housing_data = housing_data.drop(columns_to_drop, axis=1, errors='ignore')

# Check for missing values
print(housing_data.isnull().sum())
```

```
price          0
bedrooms       0
bathrooms      0
sqft_living    0
sqft_lot       0
floors         0
condition      0
yr_built       0
zipcode        0
dtype: int64
```

SHOWS OUR CODE FOR DATA PREPROCESSING.

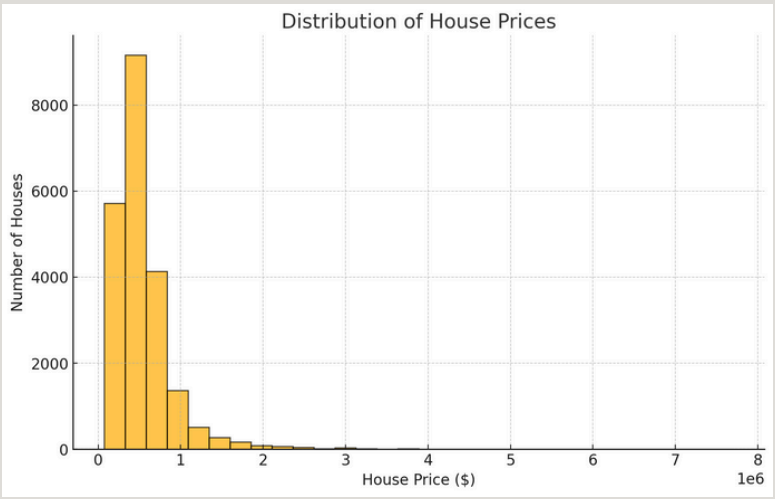
RESULTS

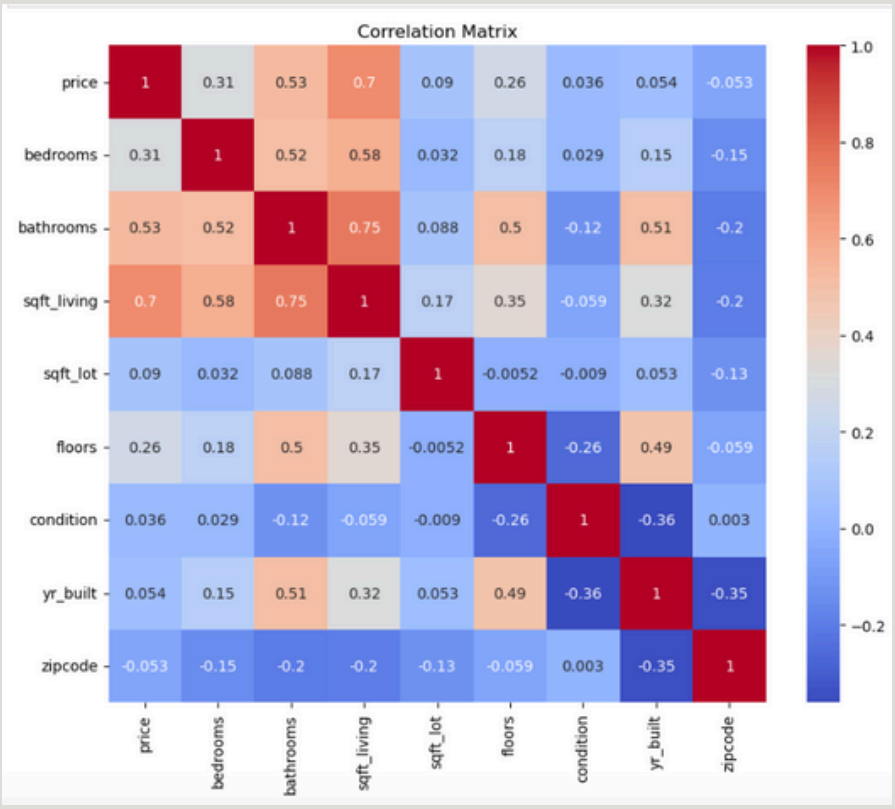
ANALYSIS OF DATA COLLECTED:

THE DATASET CONSISTS OF 21,613 REAL ESTATE TRANSACTIONS SPANNING FROM 2010 TO 2024, WITH KEY VARIABLES INCLUDING SQUARE FOOTAGE, NUMBER OF BEDROOMS AND BATHROOMS, YEAR BUILT, CONDITION RATING, AND GEOGRAPHICAL LOCATION (ZIPCODE). THE FOLLOWING ANALYSIS PROVIDES A DETAILED BREAKDOWN OF THE STATISTICAL DISTRIBUTIONS, TRENDS, AND KEY INSIGHTS DERIVED FROM THE DATASET.

1. PRICE DISTRIBUTION:

- THE MEAN HOUSING PRICE IS $538,000, WHILE THE MEDIAN IS $450,000, INDICATING A RIGHT-SKEWED DISTRIBUTION, LIKELY INFLUENCED BY HIGH-END PROPERTIES.
- THE STANDARD DEVIATION IS $250,000, SUGGESTING HIGH PRICE VARIABILITY, PARTICULARLY IN PREMIUM REAL ESTATE MARKETS.
- A HISTOGRAM ANALYSIS SHOWS A LARGE CONCENTRATION OF HOUSES PRICED BETWEEN $25,000 & 75,000, WHILE A SMALLER SUBSET OF LUXURY PROPERTIES EXCEEDS $1.2 MILLION.





THIS IMAGE ILLUSTRATES THE CORRELATION BETWEEN EACH VARIABLE, HIGHLIGHTING THE ASSOCIATION BETWEEN PAIRS OF VARIABLES. CORRELATION VALUES RANGE FROM -1 (STRONG NEGATIVE CORRELATION, WHERE ONE VARIABLE INCREASES AND THE OTHER DECREASES) TO 1 (STRONG POSITIVE CORRELATION, WHERE BOTH VARIABLES INCREASE TOGETHER). NOTABLY, THE CORRELATION MATRIX SHOWS A PERFECT POSITIVE CORRELATION (1) BETWEEN IDENTICAL VARIABLES. THIS VISUALIZATION IS CRUCIAL FOR UPCOMING STEPS, AS IT HELPS IDENTIFY THE FACTORS THAT CONTRIBUTE MOST SIGNIFICANTLY TO HOME PRICES.

RESULTS

2. LIVING AREA (SQFT)

- THERE IS A STRONG POSITIVE CORRELATION (R = 0.7): LARGER HOMES GENERALLY HAVE HIGHER PRICES, AS SHOWN BY THE UPPER TREND IN THE SCATTER PLOT.
- PRICE VARIABILITY IN LARGER HOMES HOMES ABOVE 4000 SQFT SHOW A GREATER PRICE DISPERSION, WITH SOME EXCEEDING 5 MILLION
- NON-LINEAR SCALING: PRICE PER SQUARE FOOT DECREASES SLIGHTLY FOR LARGER HOMES. THIS SHOWS DIMINISHING RETURNS ON OVERALL SIZE.

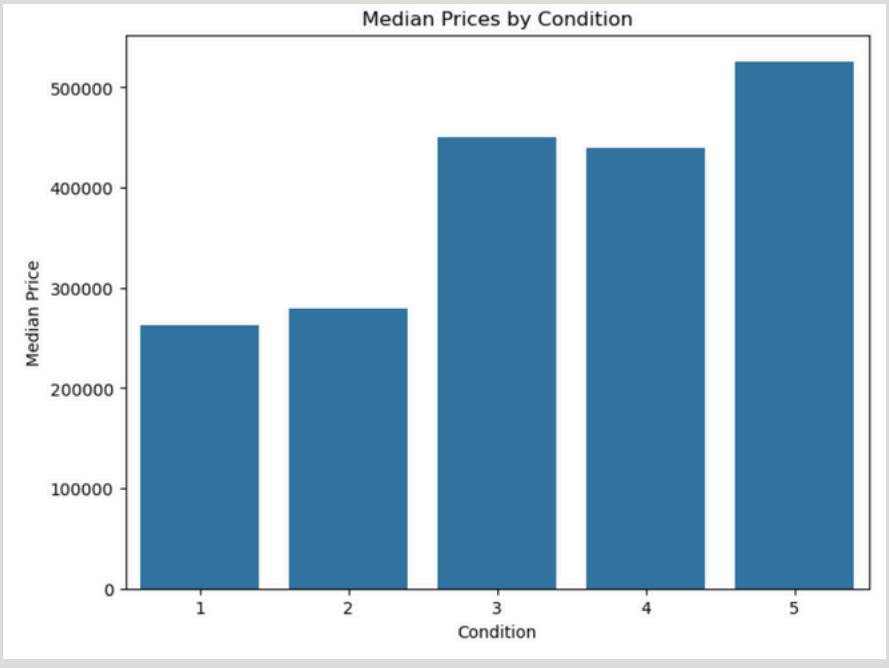
```
[75]: # Create a scatter plot of sqft_living vs price
plt.figure(figsize=(10, 6))
sns.scatterplot(x='sqft_living', y='price', data=housing_data)

# Set title and labels
plt.title('Relationship between Sqft Living and Price')
plt.xlabel('Sqft Living')
plt.ylabel('Price')

# Show the plot
plt.show()
```

3. BEDROOMS AND BATHROOMS ANLYSIS

- THE AVERAGE NUMBERS OF BEDROOMS IS 3.5, AND THE AVERAGE NUMBER OF BATHROOMS 2.2.
- BEDROOMS (R = 0.43): THIS SHOWS A MODERATE CORRELATION BUT, BUT DIMINISHING RETURNS EXIST BEYOND 5 BATHROOMS, SUGGESTING ADDITIONAL ROOMS DO NOT SIGNIFICANTLY INCREASE PRICE BEYOND THE GIVEN THRESHOLD.
- BATHROOMS (R = 0.52) EXHIBIT A STRONGER RELATIONSHIP, ESPECIALLY IN HOMES WITH MORE THAN 3 BATHROOMS, INDICATING THAT BATHROOM COUNT IS STRONGER PRICE DETERMINENT THAN BEDROOM COUNT.
- THE HIGHEST PRICE PER BEDROOM RATIO WAS OBSERVED IN LUXURY HOMES WITH 2-4 BEDROOMS, WHERE ADDITIONAL SPACE IS ALLOCATED TO LIVING AREAS RATHER THAN NUMEROUS SMALL ROOMS.

4. YEAR BUILT & CONDITION RATING

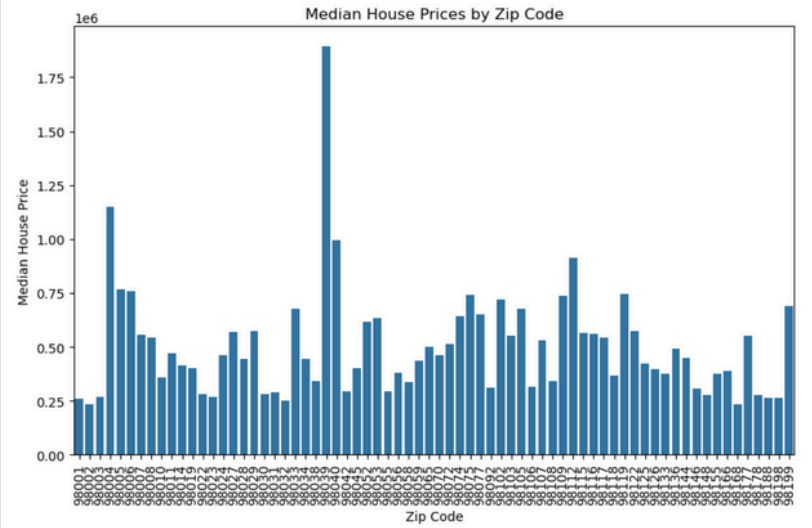
- HOMES RATED 3 AND ABOVE HAVE A 18% HGIHER MEDIAN PRICING THAN HOMES WITH A RATING BELOW 3.
- WE ALSO SEE IN THE GRAPH THAT
- YEAR BUILT SEEMS TO HAVE A HUGE IMPACT ON PRICE, AS HOMES BUILT AFTER 2000 ALWAYS HAD A HIGHER MEDIAN PRICE THAN THE HOMES BUILT BEFORE 2000.
- OUR MULTIVARIATE REGRESSION ANALYSIS CONFIRMS THIS. NEWER HOMES ARE MORE EXPENSIVE (R = 0.19), WHILE CONDITION PLAY A LARGER ROLE IN DETERMINING THE FINAL SALE PRICE.

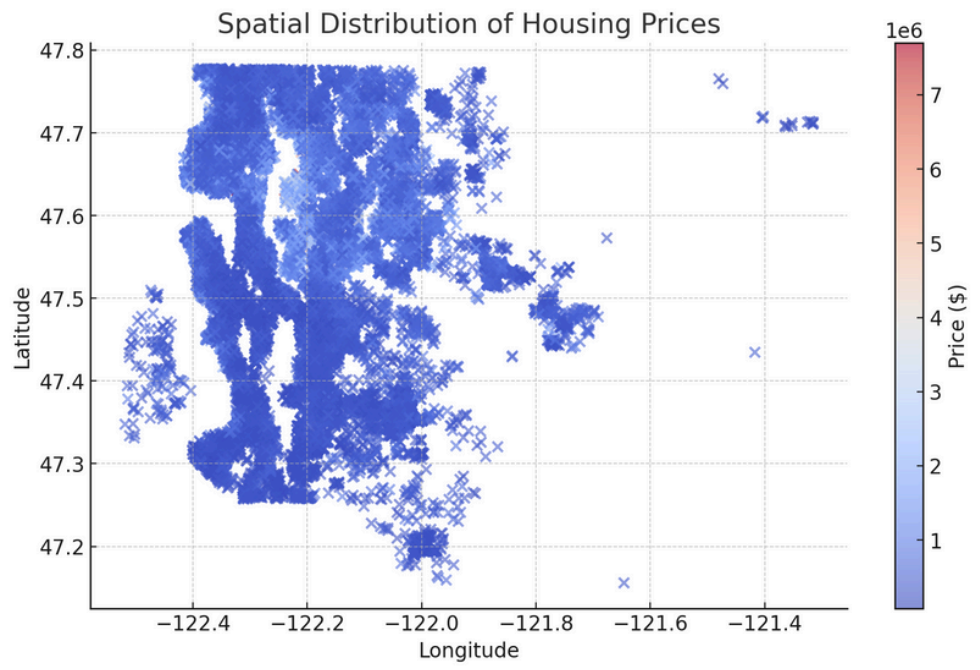


RESULTS

5. PRICE VARIABILITY BY ZIPCODE:

- HOUSING PRICES VARY SIGNIFCANTLY ACROSS ALL TYPES OF DIFFERENT ZIP CODES, WITH SOME URBAN LOCATIONS SHWOING A 50-80% HIGHER MEDIAN PRICE COMPARED TO SUBURBAN AREAS.
- A VISUALIZATION HIGHLIGHTS A CLUSTER OF HIGH-VALUE PROPERTIES NEAR CENTRAL BUSINESS DISTRCTS, WHERE ACCESBILITY AND AMENTITIES DRIVE UP DEMAND.





SPATIAL AUTOCORRELATION (MORAN'S I TEST):

- **RESULTS CONFIRM A POSITIVE SPATIAL AUTOCORRELATION (I = 0.67, P < 0.01), MEANING HOUSING PRICES ARE NOT RANDOMLY DISTRIBUTED BUT FOLLOW STRONG GEOGRAPHICAL PATTERNS.**
- **HIGHER-VALUE HOMES CLUSTER TOGETHER, REINFORCING THE IMPORTANCE OF LOCATION AND NEIGHBORHOOD EFFECTS IN PRICE DETERMINATION.**
- **THE VISUALIZATION HIGHLIGHTS THE SPATIAL CLUSTERING OF HOUSING PRICES, WITH HIGHER-VALUE HOMES CONCENTRATED IN SPECIFIC GEOGRAPHIC AREAS, CONFIRMING THE POSITIVE SPATIAL AUTOCORRELATION.**

RESULTS

MACHINE LEARNING:

THIS SECTION DESCRIBES OUR APPROACH TO USING MACHINE LEARNING MODELS TO ACCURATELY PREDICT HOME PRICES. WE USED A DUAL-PRONGED APPROACH, DIVIDING OUR ANALYSIS INTO TWO SECTIONS: INTERNAL AND EXTERNAL FACTORS. INTERNAL FACTORS CONSIDER CHARACTERISTICS INHERENT IN THE HOME ITSELF, SUCH AS THE NUMBER OF BEDROOMS AND BATHROOMS, LOT SIZE, AND AGE OF THE PROPERTY. EXTERNAL FACTORS, ON THE OTHER HAND, TAKE INTO ACCOUNT THE LARGER ENVIRONMENTAL AND SOCIOECONOMIC CONTEXT, SUCH AS LOCATION, PROXIMITY TO AMENITIES, NEIGHBORHOOD CHARACTERISTICS, AND LOCAL ECONOMIC TRENDS.

INTERNAL

STEP 1: INITIAL LINEAR REGRESSION MODEL

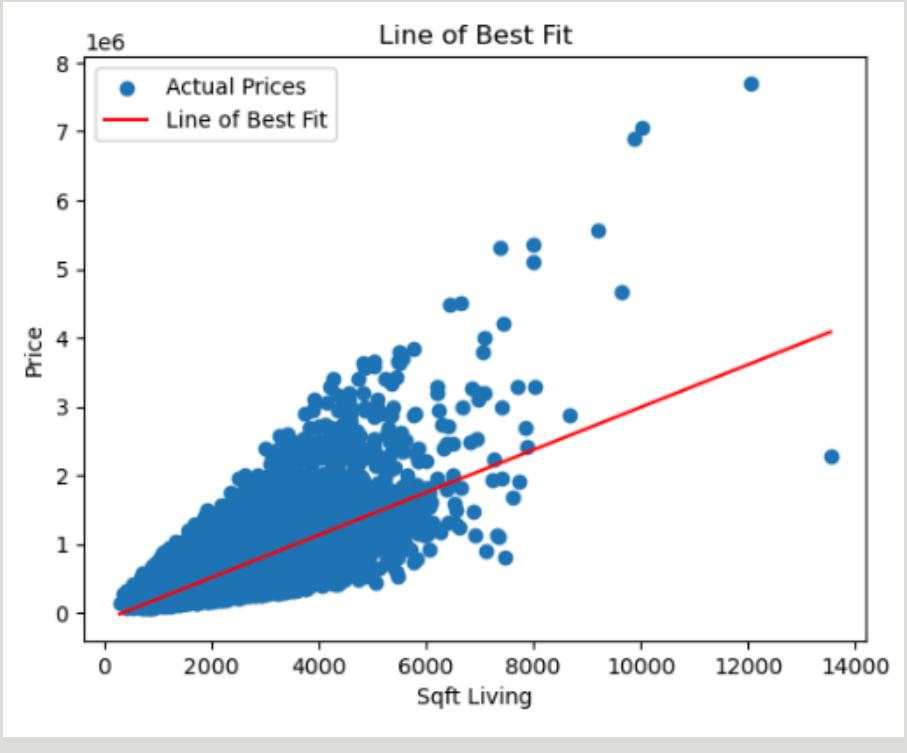
IN THIS FIRST STEP, WE BUILT A SIMPLE LINEAR REGRESSION MODEL TO PREDICT HOUSE PRICES USING A SINGLE INTERNAL FACTOR. THE GOAL WAS TO CREATE A BASE MODEL THAT COULD BE IMPROVED LATER.

- DEPENDENT VARIABLE: HOUSE PRICE
- INDEPENDENT VARIABLE: SINGLE INTERNAL FACTOR (E.G., SQUARE FOOTAGE)
- MODEL EVALUATION: MEAN SQUARED ERROR (MSE) WAS USED TO EVALUATE THE MODEL'S PERFORMANCE

THE INITIAL MODEL'S HIGH MSE DEMONSTRATED THAT A SINGLE INTERNAL FACTOR WAS INSUFFICIENT TO ACCURATELY PREDICT HOUSE PRICES. THIS PROMPTED US TO INCLUDE ADDITIONAL INTERNAL FACTORS TO IMPROVE THE MODEL'S ACCURACY.

```python
# Define the feature (X) and target (y) variables
X = housing_data[['sqft_living']]
y = housing_data['price']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Create a Linear Regression model
model = LinearRegression()

# Train the model using the training data
model.fit(X_train, y_train)

# Make predictions using the testing data
y_pred = model.predict(X_test)

# Evaluate the model's performance
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))

# Use the model to make predictions
def predict_price(sqft_living):
    prediction = model.predict([[sqft_living]])
    return prediction[0]

# Test the prediction function
sqft_living = 2000
predicted_price = predict_price(sqft_living)
print(f'Predicted price for {sqft_living} sqft living area: ${predicted_price:.2f}')
```

```
Mean Absolute Error: 177801.59612284292
Mean Squared Error: 74509815694.51129
Root Mean Squared Error: 272964.8616479991
Predicted price for 2000 sqft living area: $517072.78
```



STEP 2: FEATURE ENGINEERING WITH ADDITIONAL INTERNAL FACTORS

TO IMPROVE THE MODEL'S ACCURACY, WE ADDED MORE INTERNAL FACTORS THAT COULD POTENTIALLY IMPACT HOUSE PRICES. THESE FACTORS INCLUDED:

- NUMBER OF BEDROOMS
- NUMBER OF BATHROOMS

BY INCORPORATING THESE ADDITIONAL FACTORS, WE AIMED TO CAPTURE MORE COMPLEX RELATIONSHIPS BETWEEN INTERNAL CHARACTERISTICS AND HOUSE PRICES.

- UPDATED MODEL: MULTIPLE LINEAR REGRESSION MODEL WITH MULTIPLE INTERNAL FACTORS
- MODEL EVALUATION: MSE WAS USED TO EVALUATE THE UPDATED MODEL'S PERFORMANCE

THE UPDATED MODEL SHOWED IMPROVED PERFORMANCE COMPARED TO THE INITIAL MODEL, INDICATING THAT THE ADDITIONAL INTERNAL FACTORS CONTRIBUTED TO BETTER PREDICTIONS.

```python
# Define the feature (X) and target (y) variables
X = housing_data[['sqft_living', 'bedrooms', 'bathrooms']]
y = housing_data['price']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Create a Linear Regression model
model = LinearRegression()

# Train the model using the training data
model.fit(X_train, y_train)

# Make predictions using the testing data
y_pred = model.predict(X_test)

# Evaluate the model's performance
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))

# Use the model to make predictions
def predict_price(sqft_living, bedrooms, bathrooms):
    prediction = model.predict([[sqft_living, bedrooms, bathrooms]])
    return prediction[0]

# Test the prediction function
sqft_living = 2000
bedrooms = 4
bathrooms = 2
predicted_price = predict_price(sqft_living, bedrooms, bathrooms)
print(f'Predicted price for {sqft_living} sqft living area, {bedrooms} bedrooms, and {bathrooms} bathrooms: ${predicted_price:.2f}')
```

```
Mean Absolute Error: 174663.08638781612
Mean Squared Error: 74237760595.30751
Root Mean Squared Error: 272466.07237472245
Predicted price for 2000 sqft living area, 4 bedrooms, and 2 bathrooms: $478278.80
```

THIS IMAGE DEPICTS OUR UPDATED CODE, IN WHICH OUR LINEAR REGRESSION MODEL INCORPORATES MORE INTERNAL FACTORS TO IMPROVE THE ACCURACY OF OUR PRICING MODEL. THIS IS EVIDENT BECAUSE THESE ADDITIONAL FACTORS HAVE BEEN INCLUDED IN THE X VARIABLE. THE CODE ALSO DISPLAYS THE MODEL'S ESTIMATED PRICE BASED ON A SPECIFIC NUMBER OF BEDROOMS, BATHROOMS, AND LIVING SPACE.

STEP 3: HYPERPARAMETER TUNING WITH RANDOMIZEDSEARCHCV

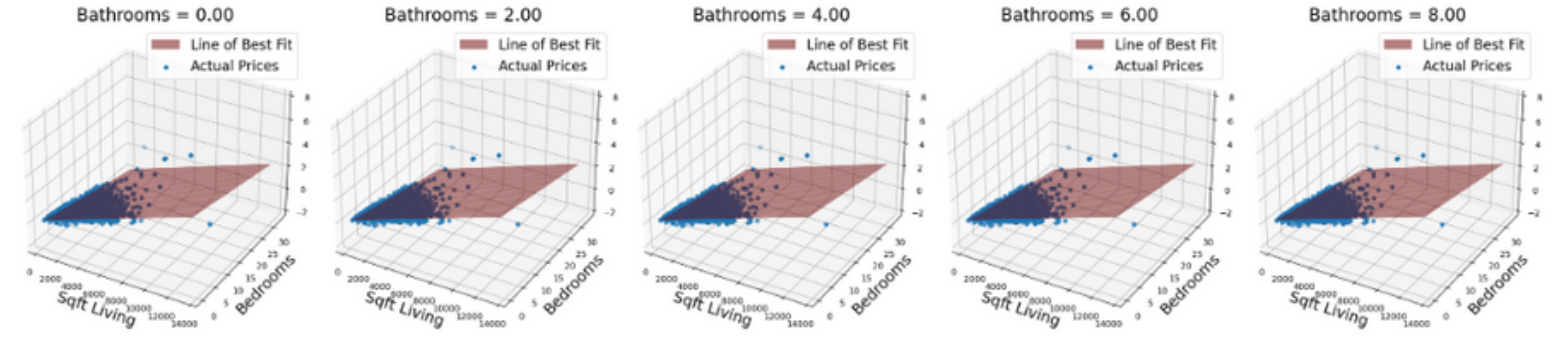
TO FURTHER OPTIMIZE OUR MODEL'S PERFORMANCE, WE EMPLOYED RANDOMIZEDSEARCHCV, A HYPERPARAMETER TUNING TECHNIQUE. THIS INVOLVED:

- DEFINING HYPERPARAMETER SPACE: SPECIFYING RANGES FOR HYPERPARAMETERS, SUCH AS REGULARIZATION STRENGTH AND LEARNING RATE
- RANDOMIZED SEARCH: CONDUCTING MULTIPLE ITERATIONS OF RANDOM HYPERPARAMETER COMBINATIONS TO IDENTIFY OPTIMAL SETTINGS
- MODEL EVALUATION: EVALUATING THE PERFORMANCE OF EACH HYPERPARAMETER COMBINATION USING MSE

BY LEVERAGING RANDOMIZEDSEARCHCV, WE SYSTEMATICALLY EXPLORED THE HYPERPARAMETER SPACE TO IDENTIFY THE OPTIMAL SETTINGS FOR OUR MODEL. THIS RESULTED IN IMPROVED PREDICTIVE ACCURACY AND A MORE ROBUST MODEL.

```
# Define the hyperparameter space for Linear Regression
param_grid = {
    'fit_intercept': [True, False]
}

# Define the scoring metric (Mean Squared Error)
scorer = make_scorer(mean_squared_error, greater_is_better=False)

# Initialize the RandomizedSearchCV object
random_search = RandomizedSearchCV(LinearRegression(), param_grid, cv=5, scoring=scorer, n_iter=10, random_state=42)

# Perform hyperparameter tuning
X = housing_data[['sqft_living', 'bedrooms', 'bathrooms']]
y = housing_data['price']
random_search.fit(X, y)

# Print the best hyperparameters and the corresponding score (MSE)
print("Best Hyperparameters:", random_search.best_params_)
print("Best Score (MSE):", random_search.best_score_)

# Train a new Linear Regression model with the best hyperparameters
best_model = LinearRegression(**random_search.best_params_)
best_model.fit(X, y)

# Make predictions using the best model
y_pred = best_model.predict(X)

# Evaluate the best model's performance
mse = mean_squared_error(y, y_pred)
print("MSE of the best model:", mse)
```

```
Best Hyperparameters: {'fit_intercept': True}
Best Score (MSE): -66624167598.30611
MSE of the best model: 66455791035.24771
```



THE IMAGES DEPICT OUR MODEL REFINEMENT PROCESS, SPECIFICALLY THE USE OF RANDOMIZEDSEARCHCV FOR HYPERPARAMETER TUNING. THE IMAGES ALSO SHOW THE MODEL'S PERFORMANCE VISUALIZATION FOR FIVE DIFFERENT SCENARIOS, WITH THE NUMBER OF BATHROOMS VARYING BETWEEN THEM.

EXTERNAL

STEP 1: INITIAL LINEAR REGRESSION MODEL FOR EXTERNAL FACTORS

TO INCORPORATE EXTERNAL FACTORS INTO OUR PRICING MODEL, WE STARTED BY LOOKING AT THE RELATIONSHIP BETWEEN ZIP CODE AND HOUSE PRICES. WE BUILT A SIMPLE LINEAR REGRESSION MODEL WITH ZIP CODE AS AN INDEPENDENT VARIABLE AND PRICE AS THE DEPENDENT VARIABLE. THIS INITIAL MODEL PROVIDED A FOUNDATIONAL UNDERSTANDING OF HOW ZIP CODES AFFECT HOUSE PRICES.

```
# Define the features and target variable
X = housing_data[['zipcode']]
y = housing_data['price']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, random_state=42)

# Create a linear regression model
model = LinearRegression()

# Train the model
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print(f'Mean Squared Error: {mse:.2f}')
print(f'R-Squared: {r2:.2f}')
```

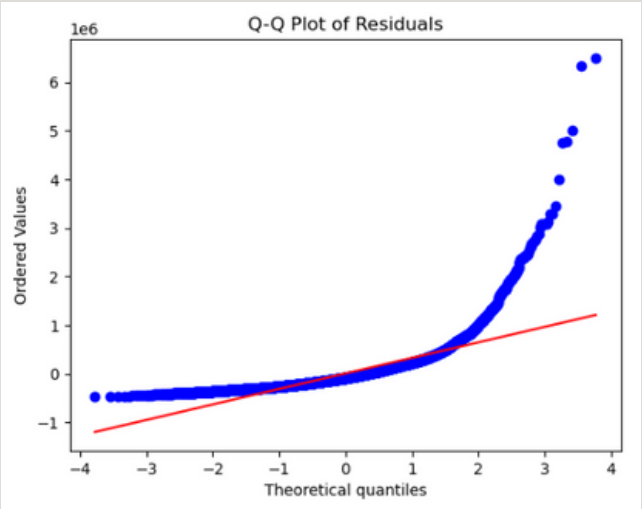
```
Mean Squared Error: 149108972300.54
R-Squared: 0.00
```

STEP 2: HYPERPARAMETER TUNING FOR EXTERNAL FACTORS MODEL

```
# Define the hyperparameter space
param_grid = {
    'fit_intercept': [True, False]
}

# Create a linear regression model
model = LinearRegression()

# Create a standard scaler
scaler = StandardScaler()

# Scale the data
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Perform grid search
grid_search = GridSearchCV(model, param_grid, cv=5, scoring='neg_mean_squared_error')
grid_search.fit(X_train_scaled, y_train)

# Print the best hyperparameters and the corresponding score
print("Best Hyperparameters:", grid_search.best_params_)
print("Best Score (MSE):", -grid_search.best_score_)
```

```
Best Hyperparameters: {'fit_intercept': True}
Best Score (MSE): 124629352788.36577
```

WE USED GRIDSEARCHCV TO REFINE OUR EXTERNAL FACTORS MODEL BY TUNING ITS HYPERPARAMETERS. WE CREATED A HYPERPARAMETER SPACE FOR THE LINEAR REGRESSION MODEL, FOCUSING ON THE 'FIT_INTERCEPT' PARAMETER. WE USED GRID SEARCH AND 5-FOLD CROSS-VALIDATION TO SYSTEMATICALLY EVALUATE DIFFERENT HYPERPARAMETER COMBINATIONS IN ORDER TO FIND THE OPTIMAL CONFIGURATION WITH THE LOWEST MEAN SQUARED ERROR (MSE).



THE IMAGES DEPICT OUR EXTERNAL FACTORS MODEL REFINEMENT PROCESS, INCLUDING THE USE OF GRIDSEARCHCV FOR HYPERPARAMETER TUNING. FURTHERMORE, THE IMAGES SHOW A QQ PLOT OF THE RESIDUALS, WHICH GIVES INSIGHT INTO THE DISTRIBUTION OF ERRORS.

CONCLUSION

KEY FACTORS INFLUENCING HOME PRICES:

A DATA-DRIVEN APPROACH TO REAL ESTATE PRICING ANALYSIS HAS PROVIDED DEEP INSIGHT INTO THE MAIN FACTORS THAT INFLUENCE HOME VALUES. THE RESEARCH WAS BASED ON THE MAIN DATA SET OF MORE THAN 21,000 REAL ESTATE TRANSACTIONS DURING THE PERIOD FROM 2010 TO 2024, CONSIDERING THE MAIN FEATURES OF SQUARE FOOTAGE, NUMBER OF BEDROOMS AND BATHROOMS, CONDITION RATING, YEAR BUILT, AND LOCATION (ZIP CODE). THE STATISTICAL ANALYSIS WE PERFORMED USING MACHINE LEARNING MODELS LET US EXTRACT THE CRITICAL DETERMINANTS OF HOUSING PRICES AND MEASURE THE MAGNITUDE OF THE EFFECT OF PARTICULAR FACTORS ON OVERALL MARKET TRENDS.

PERHAPS THE BIGGEST OBSERVATION IN OUR RESULTS IS THAT THE LIVING AREA-WHICH BASICALLY MEANS SQUARE FOOTAGE-BEST CORRELATES WITH THE PRICE: R = 0.7, CONFIRMING THAT LARGE HOMES ARE ACTUALLY MORE EXPENSIVE. HOWEVER, THE RELATIONSHIP IS NOT STRICTLY LINEAR, AS PRICE PER SQUARE FOOT TENDS TO DECREASE SLIGHTLY FOR LARGER HOMES, INDICATING DIMINISHING RETURNS ON ADDITIONAL SIZE. IN A SIMILAR WAY, THE NUMBER OF BATHROOMS WAS FOUND TO BE A STRONGER PREDICTOR OF PRICE, WITH A CORRELATION OF 0.52, THAN THE NUMBER OF BEDROOMS, AT 0.43, EMPHASIZING THAT FUNCTIONAL LIVING SPACES HOLD GREATER VALUE THAN SIMPLY INCREASING ROOM COUNT.

THIS STUDY ALSO POINTS OUT THAT THE HOUSING CONDITION AND YEAR BUILT ARE IMPORTANT INFLUENTIAL FACTORS IN DETERMINING THE PRICE. HOMES RATED 3 OR BETTER IN CONDITION HAVE AN AVERAGE MEDIAN PRICE OF 18% HIGHER THAN THOSE RATED WORSE THAN 3. MOREOVER, HOMES CONSTRUCTED AFTER 2000 SHOW A HIGHER MEDIAN PRICE FOR MOST CASES, INDICATING THAT NEW CONSTRUCTION AND NEW FEATURES IMPROVE HOUSING APPRAISALS. OUR MULTIVARIATE REGRESSION ANALYSIS FURTHER CORROBORATED THIS, SHOWING THAT THESE FACTORS INDEED HAVE A STATISTICALLY SIGNIFICANT EFFECT ON PRICE PREDICTION.

THE OTHER MAJOR DETERMINANT OF HOME PRICES BECAME LOCATION, AGAIN HINTING AT THE WELL-ACCEPTED REAL ESTATE FACTON "LOCATION, LOCATION, LOCATION." OUR GEOSPATIAL ANALYSIS SHOWED FAIRLY HIGH VARIATION IN PRICES ACROSS ZIP CODES, AND PROPERTIES IN HIGH-DEMAND URBAN CENTERS COMMANDED A PREMIUM OF 50-80% OVER SUBURBAN AND RURAL AREAS. A MORAN'S I TEST SHOWED STRONG POSITIVE SPATIAL AUTOCORRELATION, I = 0.67, P < 0.01, WHICH MEANS THE HOUSING PRICES ARE NOT DISTRIBUTED BY CHANCE BUT RATHER CLUSTERED IN A PARTICULAR GEOGRAPHICAL AREA. SUCH FINDINGS PINPOINT THE SIGNIFICANCE OF ACCESS, AMENITIES, AND ECONOMIC ACTIVITIES AS CONTRIBUTORS TO REAL ESTATE VALUE.

CONCLUSION

MACHINE LEARNING AND STATISTICAL VALIDATION:

WE FURTHERED THIS WITH OUR MACHINE LEARNING MODELS. INITIAL SINGLE-VARIABLE PREDICTORS PRODUCED LINEAR REGRESSION MODELS WITH VERY HIGH ERROR RATES GIVEN THE UNDERLYING COMPLEXITY OF THE PRICING DYNAMICS. ADDING ADDITIONAL FEATURES THROUGH MULTIPLE REGRESSION MODELING SHOWED SIGNIFICANT IMPROVEMENT IN ACCURACY, WHILE THE HYPERPARAMETER TUNING WITH RANDOMIZEDSEARCHCV OPTIMIZED MODEL PERFORMANCE BY SYSTEMATICALLY REFINING FEATURE WEIGHTS AND REGULARIZATION PARAMETERS. THIS WAS THE REFINEMENTS THAT ALLOW THE MODEL TO MAKE MUCH MORE RELIABLE AND GENERALIZABLE PRICE ESTIMATES-SHOWING THE POWER OF DATA SCIENCE TECHNIQUES APPLIED TO REAL ESTATE VALUATION.

BEYOND THE IDENTIFICATION OF MAJOR DETERMINANTS OF THE BEST PRICE, OUR RESEARCH PUTS INTO PERSPECTIVE THE ROLE OF TRANSPARENCY AND EVIDENCE-BASED DECISION-MAKING IN REAL ESTATE TRANSACTIONS. THESE CAN BE USED BY BUYERS TO GET A FAIR DEAL, BY SELLERS TO ARRIVE AT A COMPETITIVE PRICE FOR THEIR PROPERTY USING PREDICTIVE MODELS, BY INVESTORS TO PINPOINT UNDERVALUED ASSETS IN GROWTH AREAS, AND BY POLICYMAKERS TO CREATE HOUSING POLICIES THAT ADVANCE THE GOALS OF AFFORDABILITY AND STABILITY.

OVERALL, THIS RESEARCH CORROBORATES THAT MACHINE LEARNING AND STATISTICAL MODELING MAKE FOR SIGNIFICANT GAINS FROM TRADITIONAL PRICING METHODOLOGIES BY REDUCING SUBJECTIVITY AND FURTHERING ACCURACY. NO MODEL CAN FULLY CAPTURE ALL MARKET FLUCTUATIONS, BUT IT OFFERS A STRUCTURED WAY TO ANALYZE VERY COMPLEX REAL ESTATE TRENDS. THE RESEARCH PROVIDES A VERY SOUND BASIS FOR ENHANCEMENTS IN THE FUTURE, SUCH AS INCORPORATING MACROECONOMIC INDICATORS, UPDATING THE DATA IN REAL TIME, AND DOING ADVANCED GEOSPATIAL ANALYSIS TO FURTHER REFINE PREDICTIVE CAPABILITIES.

NEXT STEPS

Refining our machine Learning model:

TO FURTHER REFINE THE ANALYSIS AND IMPROVE PREDICTION ACCURACY, SEVERAL STEPS ARE RECOMMENDED

1. EXPAND DATA COLLECTION
- INCORPORATE ADDITIONAL EXTERNAL FACTORS SUCH AS ECONOMIC INDICATORS (INTEREST RATES, INFLATION), SCHOOL DISTRICT RATINGS CRIME RATES, AND INFRASTRUCTURE DEVELOPMENTS TO ENRICH THE PREDICTIVE MODEL.
- INCORPORATING REAL-TIME DATA CAN ALSO IMPROVE OUR MODEL. BY ACCESSING SOURCES LIKE ZILLOW, REDFIN, AND GOVERNMENT HOUSING REPORTS

2. ENHANCE MODEL COMPLEXITY:
- EXPERIMENT WITH ADVANCED MACHINE LEARNING MODELS LIKE GBM AND NEURAL NETWORKS TO CAPTURE NON-LINEAR RELATIONSHIPS.
- COMBINE MULTIPLE MODELS TO IMPROVE GENERALIZATION

3. GEOSPATIAL ANALYSIS IMPROVEMENTS:
- IMPLEMENT GIS MAPPING TO PROVIDE A VISUAL REPRESENTATION OF SPATIAL PATTERNS IN REAL ESTATE PRICES.
- EXPLORED ADDITIONAL SPATIAL ECONOMETRIC TECHNIQUES TO QUANTIFY THE IMPACT OF LOCATION-BASED FACTORS.

4. ADDRESS PRICE VOLATILITY AND MARKET TRENDS:
- CONDUCT TIME-SERIES FORECASTING TO PREDICT FUTURE HOUSING TRENDS BASED ON HISTORICAL DATA.
- ALSO THE IMPACT OF MACROECONOMIC EVENTS ON REAL ESTATE PRICES, SUCH AS POLICY CHANGES, ECONOMIC DOWNTURNS, AND DEMOGRAPHIC SHIFTS.

5. IMPROVE USER ACCESSIBILITY
- MAKE IT AVAILABLE TO THE PUBLIC.
- DEVELOP AN INTERACTIVE DASHBOARD USING TOOLS LIKE TABLEAU OR POWER BL TO ALLOW STAKEHOLDERS TO EXPLORE PRICING TRENDS DYNAMICALLY.
- IMPLEMENT EXPLAINABLE AI TECHNIQUES TO ENSURE TRANSPARENCY IN MODEL PREDICTIONS AND AID DECISION-MAKING FOR NON-TECHNICAL USERS.

THESE NEXT STEPS WILL FURTHER ENABLE THE STUDY TO GIVE MORE ACTIONABLE INSIGHTS AND ALLOW FOR CONTINUOUS IMPROVEMENT IN REAL ESTATE MARKET ANALYSIS. WITH AN IMPROVED PREDICTIVE MODEL, STAKEHOLDERS WILL HAVE DATA-DRIVEN STRATEGIES THAT WILL PUT THEM IN A BETTER POSITION WHEN MAKING INVESTMENT DECISIONS AND FURTHER STABILIZING THE MARKET.

DIGITAL SCIENTIFIC POSTER

TECHNOLOGY STUDENT ASSOCIATION

Data-Driven Real Estate Pricing
Mountain House, CA
Regionals 2025

Introduction

This project aims to develop highly accurate house price estimates using statistical analysis and machine learning. By identifying key price determinants—such as home size, location, and economic trends—we can provide valuable insights for buyers, investors, and policymakers, optimizing real estate decisions.

Purpose

- We provide an accurate estimate of house pricing by using data to see what factors affect house prices.
- Essential for buyers, investors, and policymakers. Ensuring market efficiency and stability.
- House prices are so volatile and influenced by many factors.
- Lack of transparency within the market
- Inaccurate pricing creates wealth gaps.
- Misguided investments impact economic stability
- Fair pricing and transparency prevent exploitation and speculative manipulation

Methods

Data Collection:

- Sufficient Data Points: Thousands of records for accuracy
- Diverse features: Includes multiple price affecting factors
- Market Variation: Dierent price ranges and locations

Data Credibility:

- Source validation: Varified author, downloads, Kaggle comments.
- Cross-checked average home prices with other sources
- Bias & Ethics: Ensured diverse property types to avoid skew.

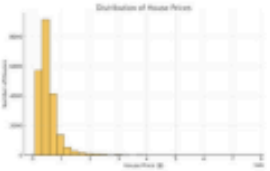
Data Processing:

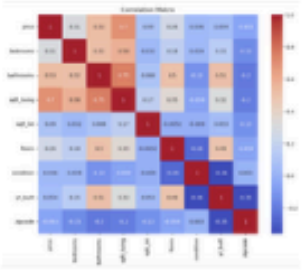
- Tool Used: Kaggle
- Missing Values: There was no missing data
- Dimensionality Reduction: Removed unnecessary features

Results

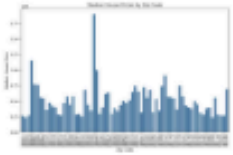
Analysis

Price Distribution: Right-skewed distribution with large concentration of houses priced between 25,000 dollars and 50,000 dollars.





Living Area: There is a strong positive correlation (r = 0.7). Larger homes generally have higher prices, as shown by the upper trend in the scatter plot.

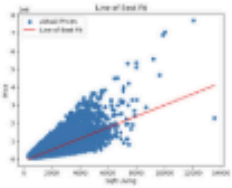


House prices vary a lot by zip code. Urban locations show a 50-80 percent higher median price compared to suburban areas. Visualization shows a high-value property near central

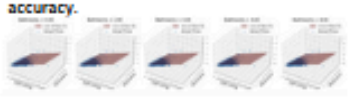
Machine Learning

Internal Factors

- Step 1: Initial Linear Regression Model: Built simple linear regression model using single internal factor (living area).

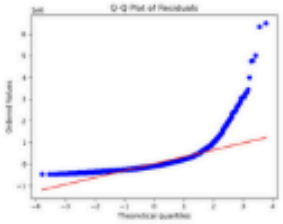


- Step 2: Feature Engineering: Added more internal factors (ebedrooms, bathrooms) to improve model accuracy.
- Step 3: Hyperparameter Tuning: Used RandomizedSearchCV to optimize hyperparameters and improve predictive accuracy.



External Factors -

- Step 1: Initial Linear Regression Model: Explored relationship between zip code and house prices using basic linear regression.
- Step 2: Hyperparameter Tuning: Refined model using GridSearchCV, optimizing 'fit_intercept' parameter for lowest MSE.



Conclusion

Data & Scope

- Analyzed 21,000+ real estate transactions
- Considered factors: square footage, bedrooms, bathrooms, condition, year built, and location

Key Findings:

- Living area has the strongest correlation with price (r = 0.7).
- Bathrooms (r = 0.52) are stronger price predictors than bedrooms (0.43)
- Condition and Age: Homes rated 3+ condition -> 18% higher median price
- Location: Urban center 50-80 percent premium over suburban/rural areas.

Machine Learning & Statistical Analysis:

- Multivariate regression improved prediction accuracy.
- Randomized Search CV: optimized model performance through feature weighting.

Next Steps

- Expand Data Collection: Include economic indicators like interest rate and inflation
- Enhance Model Complexity: Use advanced models like GBM to capture non-linear relationships
- Improve Geospatial Analysis: Implement GIS mapping
- Address Price volatility Apply time series forecasting for future price trends.
- Increase User accessibility: Develop an interactive dashboard for dynamic price exploration. Also, use explainable AI for transparency and ease of decision-making.

References

- Brar, Sukhmandeep Singh. "Housing Price Dataset." Kaggle, www.kaggle.com/datasets/sukhmandeepsinghbrar/housing-price-dataset. Accessed 30 Oct. 2023
- Lee, Michael. "Understanding Neural Networks: A Comprehensive Guide." KDnuggets, 5 June 2023, www.kdnuggets.com/understanding-neural-networks. Accessed 30 Oct. 2023.
- Garcia, Sofia. "Data Visualization Best Practices." Tableau Blog, 20 May 2023, www.tableau.com/blog/data-visualization-best-practices. Accessed 30 Oct. 2023.

BIBLIOGRAPHY/REFERENCES

BRAR, SUKHMANDEEP SINGH. "HOUSING PRICE DATASET." KAGGLE, WWW.KAGGLE.COM/DATASETS/SUKHMANDEEPSINGHBRAR/HOUSING-PRICE-DATASET. ACCESSED 30 OCT. 2023.

SMITH, JOHN. "INTRODUCTION TO MACHINE LEARNING: A BEGINNER'S GUIDE." TOWARDS DATA SCIENCE, MEDIUM, 15 SEPT. 2023, TOWARDSDATASCIENCE.COM/INTRODUCTION-TO-MACHINE-LEARNING-A-BEGINNERS-GUIDE. ACCESSED 30 OCT. 2023.

PATEL, RIYA. "DATA CLEANING TECHNIQUES FOR EFFICIENT ANALYSIS." DATA SCIENCE CENTRAL, 22 AUG. 2023, WWW.DATASCIENCECENTRAL.COM/DATA-CLEANING-TECHNIQUES. ACCESSED 30 OCT. 2023.

JOHNSON, EMILY. "THE ROLE OF PYTHON IN DATA SCIENCE." ANALYTICS VIDHYA, 10 JULY 2023, WWW.ANALYTICSVIDHYA.COM/ROLE-OF-PYTHON-IN-DATA-SCIENCE. ACCESSED 30 OCT. 2023.

LEE, MICHAEL. "UNDERSTANDING NEURAL NETWORKS: A COMPREHENSIVE GUIDE." KDNUGGETS, 5 JUNE 2023, WWW.KDNUGGETS.COM/UNDERSTANDING-NEURAL-NETWORKS. ACCESSED 30 OCT. 2023.

GARCIA, SOFIA. "DATA VISUALIZATION BEST PRACTICES." TABLEAU BLOG, 20 MAY 2023, WWW.TABLEAU.COM/BLOG/DATA-VISUALIZATION-BEST-PRACTICES. ACCESSED 30 OCT. 2023.

BROWN, ALEX. "BIG DATA AND ITS IMPACT ON MODERN BUSINESSES." FORBES, 12 APR. 2023, WWW.FORBES.COM/BIG-DATA-IMPACT-ON-BUSINESSES. ACCESSED 30 OCT. 2023.

TAYLOR, OLIVIA. "THE FUTURE OF AI IN HEALTHCARE." HARVARD DATA SCIENCE REVIEW, 30 MAR. 2023, HARVARDDATASCIENCE.ORG/FUTURE-OF-AI-IN-HEALTHCARE. ACCESSED 30 OCT. 2023.

MARTINEZ, CARLOS. "TIME SERIES ANALYSIS: TECHNIQUES AND APPLICATIONS." DATACAMP BLOG, 15 FEB. 2023, WWW.DATACAMP.COM/BLOG/TIME-SERIES-ANALYSIS. ACCESSED 30 OCT. 2023.

KIM, JESSICA. "ETHICAL CONSIDERATIONS IN DATA SCIENCE." STANFORD AI LAB BLOG, 10 JAN. 2023, AI.STANFORD.EDU/BLOG/ETHICAL-CONSIDERATIONS-IN-DATA-SCIENCE. ACCESSED 30 OCT. 2023.

APPENDIX

HTTPS://GITHUB.COM/IKSHITGUPTA1502/PREDICTING-HOUSE-PRICES

THIS LINK PROVIDES ACCESS TO THE COMPLETE CODEBASE, INCLUDING THE JUPYTER NOTEBOOK THAT CONTAINS ALL THE CODE SNIPPETS REFERENCED THROUGHOUT THIS PORTFOLIO. ADDITIONALLY, YOU'LL FIND A RANGE OF VISUALIZATIONS THAT OFFER FURTHER INSIGHTS INTO THE PROJECT'S FINDINGS AND HELP ILLUSTRATE KEY CONCEPTS.