REIn: Real Estate Investment)

Predict House Prices and future trends for Top 10 Global economies.

Executive Summary

Predict Future Prices

The primary goal of this project is to forecast house prices for Top 10 Global Economies.

Identify key attributes

Determine main attributes that influence real estate prices.

Insights

Real Estate Investment tool that offers insights for home-owners, government, financial institutions and investors.

Problem Statement

There are **multiple models** currently used by realtors, government and financial institutions to price a home at time of sale or appraise the value of a home.

What is missing is one **global model** that everyone can access not only to understand the value of their home but also **predict future values** to assist investors and homeowners make decisions on whether to sell now or in the future.

This would help investors to tap into **global real estate markets** as an alternative to stocks.

Related Work

Existing Solutions

- **Zillow:** Online real estate database.
- **MLS:** Pricing tool used by real estate agents.
- **Property appraisal Tools:** Used by banks, mortgage institutions and government.

Limitations of existing Solutions

- Zillow’s price estimator *zestimate* has a median error of **7.5%.**
- MLS database only has properties sold or listed for sale.
- Appraisal tools use data only for houses sold recently.
- Global database missing.

Project Phases and Timelines

Phase	Description	Timelines
Phase 1	Build a prediction model for houses in Dallas Fort Worth metroplex in the last 1 year. (Phase1 part of the project is what I plan on completing for this course).	9/30/2024
Phase 2	Expand the model to include all homes sold in US in the last 1 year.	10/31/2024
Phase 3	Historical Data (last 10 years)	11/30/2024
Phase 4	Expand to include top 10 countries based on real estate value.	12/31/2024

Top 10 Real Estate Markets

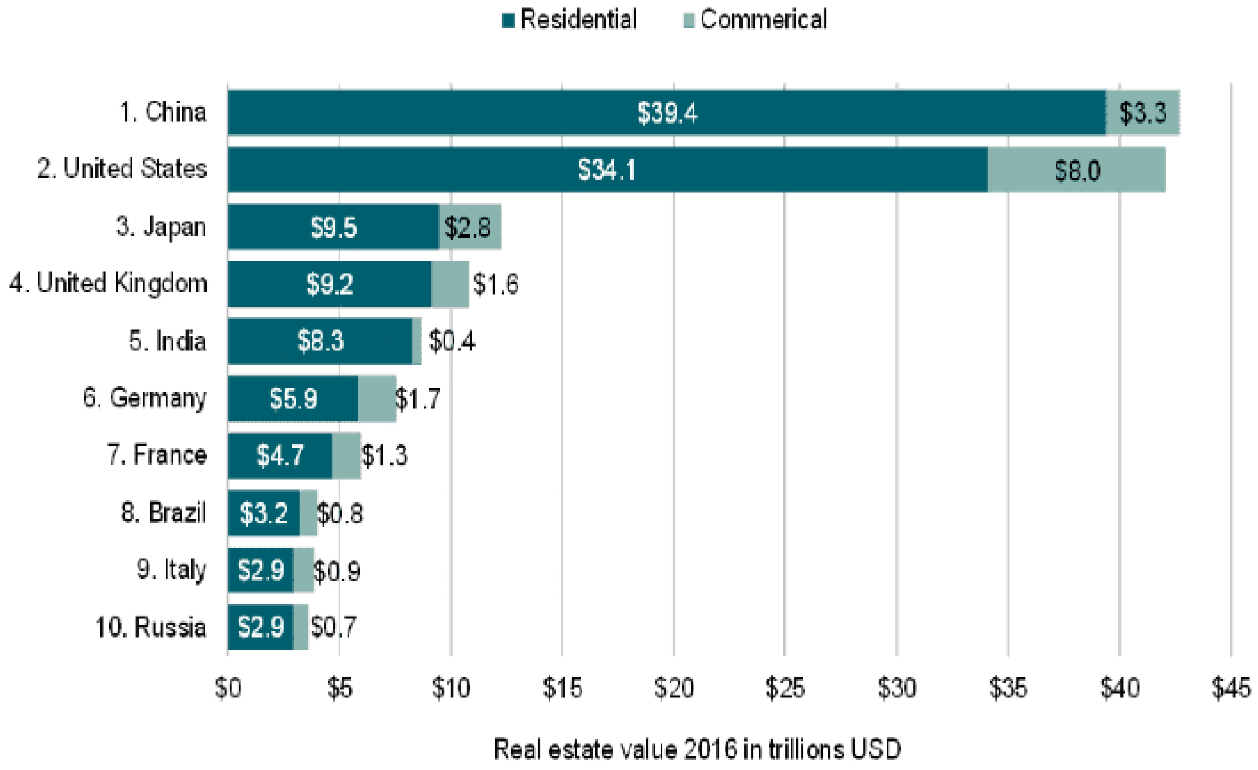


Image courtesy Saville.com

Project Phase 1 (Proposed Work)

Initial Submission

Dataset and approach

- Dataset of Houses sold in Dallas Fort Worth Metroplex from MLS website.
- Identify key attributes that impact house prices.
- Build a Model to predict house prices in Dallas Fort Worth.

Tasks

- **Statistical Analysis:** Correlation and chi-sq tests to identify key attributes.
- **Normalization:** Normalize data to ensure consistency across different regions.
- **Models:** Linear Regression, Decision Trees, Neural networks.

Proposed Work: updates (Slides 9-15)

Data Source

- Data of houses sold in Denton and Collin county downloaded from MLS website. It includes 27914 records and 31 attributes.

Review Attributes

- 60 duplicate records (same MLS ID) deleted.
- Deleted 4 attributes: **S No**, **MLS Id**, **MLS Status** (closed for all records) and **Standard Status**(closed for all records except 2 which got recently sold so status still showing as pending).
- Property Type: 27487 records (98.7%) are Single Family Homes. Insufficient data for other property types such as Ranch, Condos, Townhomes so remove them.

Handling Missing Values

Attribute	Missing Values	% Missing	Solution
Subdivision Name	58	0.2%	Keep blank for now.
Pool YN	152	0.6%	Small number. Remove 152 records since Pool may be a key attribute.
Original List Price	1	0.0%	Keep List price same as Close price.
Waterfront YN	18091	65.7%	Delete attribute since most data is missing.
HOA Fee	6258	22.7%	Keep blank for now.
Fencing	5193	18.9%	Keep blank for now.
Flooring	2508	9.1%	Keep blank for now.
HOA Fee Includes	6429	23.3%	Keep blank for now.
Lot Size	17	0.1%	Can be back calculated from **Acres** field
# Parking Spaces	27547	100.0%	Delete attribute since 100% data is missing
High School Name	235	0.9%	Keep blank for now.
Middle School Name	1699	6.2%	Keep blank for now.
Elementary School Name	133	0.5%	Keep blank for now.
Acres	1	0.0%	Remove since acres may be a key attribute.

Handling 0 and Negative Values

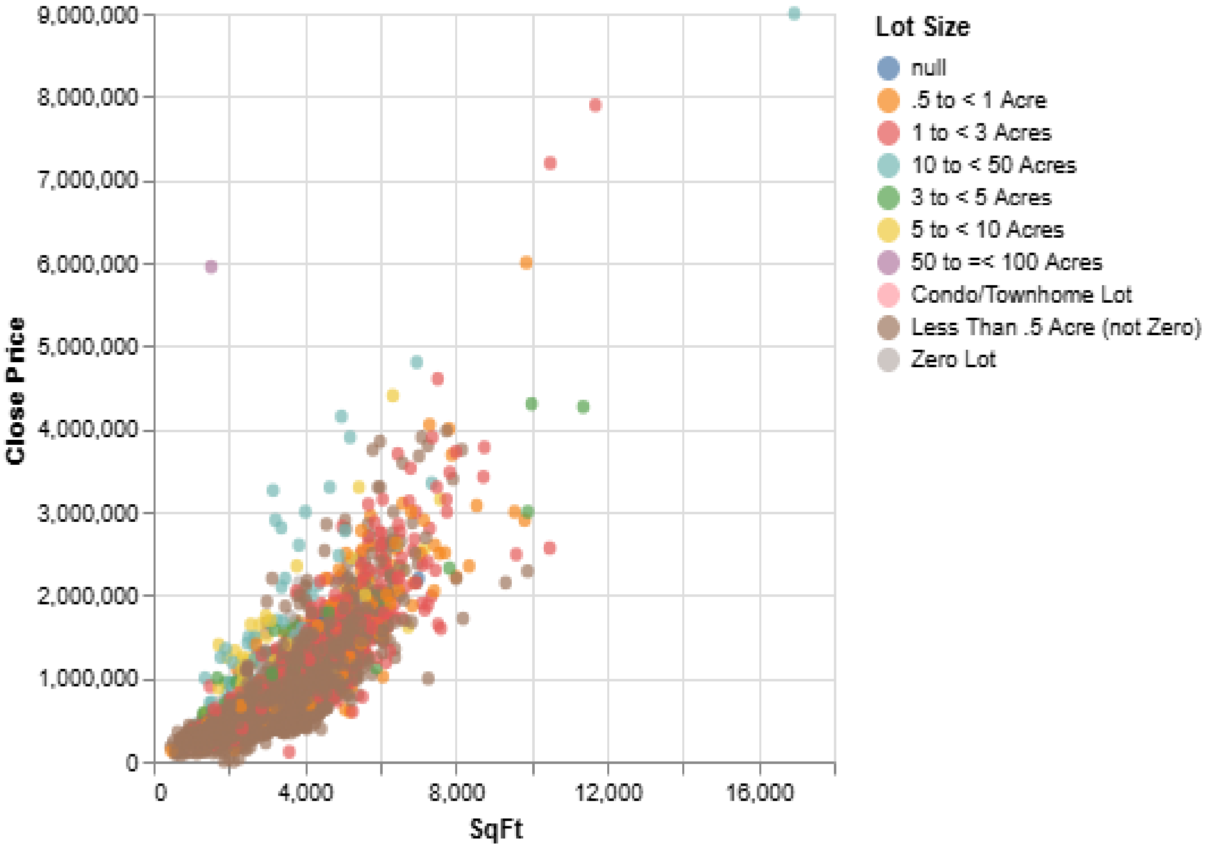
Attribute	Count of 0 values
Beds	6
Baths	0 except where Beds = 0
Acres	483
SqFt./Living Area	0

Remove records with 0 values for these attributes

Attribute	Negative Values
Days on Market	46

Change these values to 0 days on Market

Correlation Analysis and Outlier detection: SqFt and Sale Price

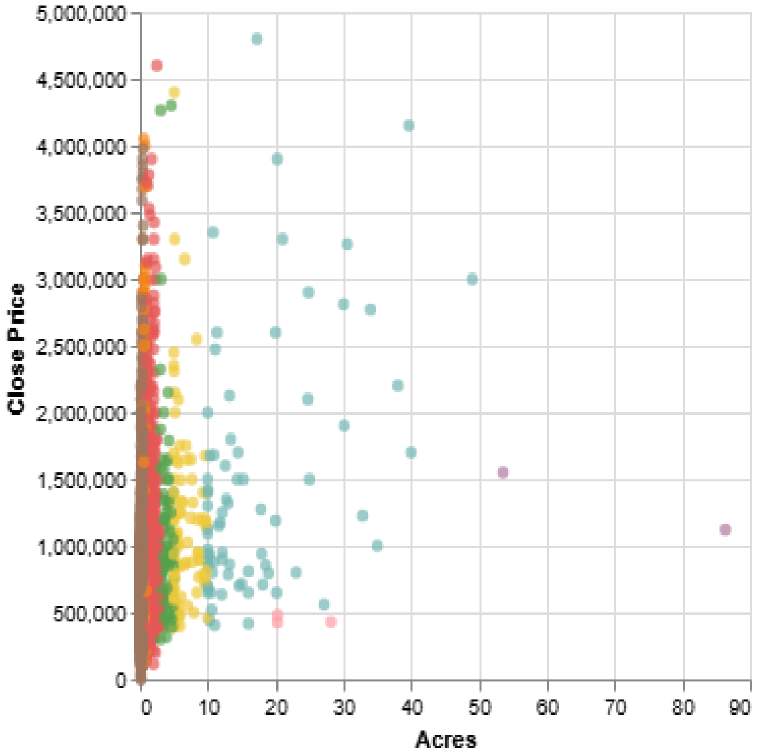


Strong Correlation between Living Area and Sale Price - 0.83

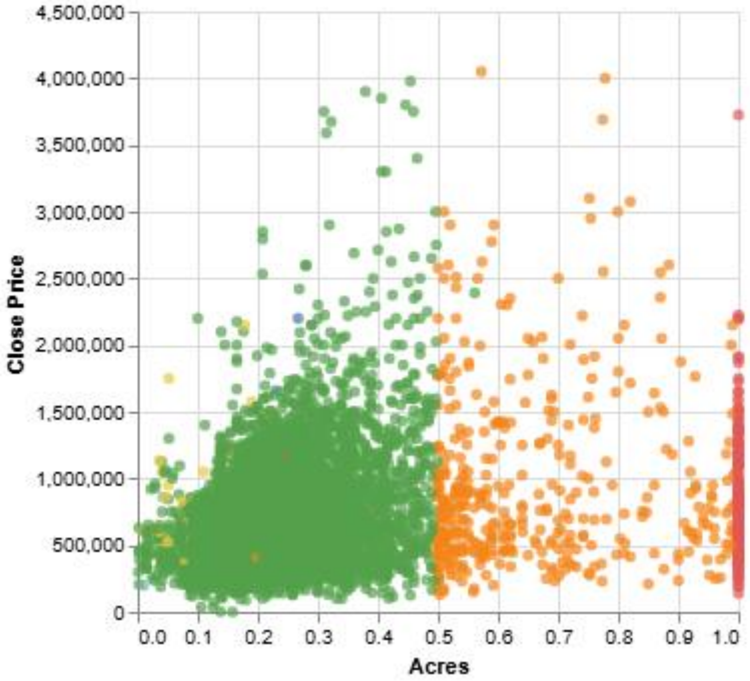
Outliers: 5 points above 5M which can be treated as outliers and removed since they are skewing the data.

7 points below 100,000 and they seem to be incorrect entries.

Correlation Analysis and Outlier detection: Acres and Sale Price

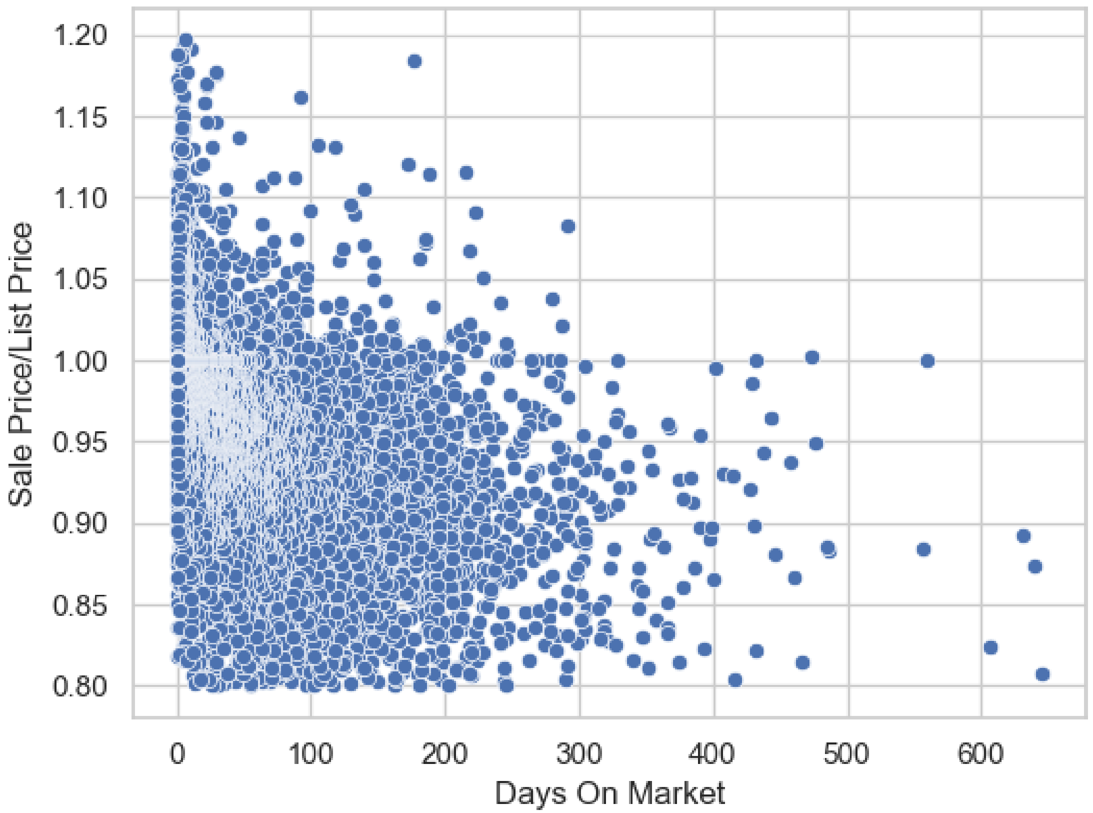


Filter to only homes <1 acre



Weak Correlation between Acres (Lot Size) and Sale Price - 0.27

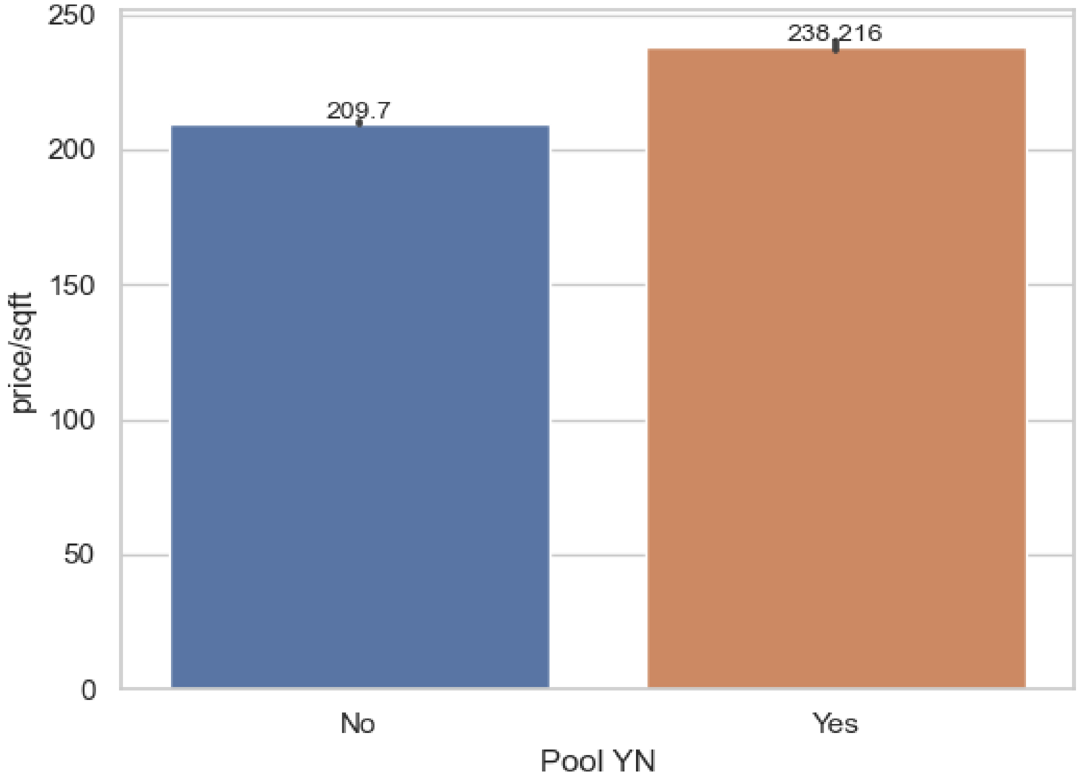
Correlation Analysis: Days On Market and Sale Price as % of List Price



Correlation score = -0.49 so more days the house is in the market, lesser would be the Sale Price.

Days On Market	Sale Price/List Price
0-15	99.5%
15-30	96.7%
30-60	95.1%
60-90	93.6%
>90	91.6%

Categorical Attributes: Pool



Chi-Sq value of 2.87 so statistically the difference is not significant.

We may still consider it if we decide to go for Decision Tree or Neural Network.

Input Features based on initial analysis

Numerical

1. Living Area
2. Acres (Lot Size)
3. Days On Market

Categorical

1. Pool
2. School District
3. Beds
4. Bath

Model Training and Evaluation

Model	Attributes	Model Type	Accuracy (R-sq.)
Model 1	Living Area	Regression	70.5%
Model 2	Add Lot Size	Regression	72.8%
Model 3	Add Bedrooms, Bath, Pool	Regression	79.1%
Model 4	Add School District	Regression	82.7%
Model 5	Same as Model 4	Neural Network	79.4%

Based on the current set of attributes the best result is 82.7% from a Linear Model. Need to research if more attributes can be identified and added to get to 95% accuracy. Examples of new attributes to be researched: age of home, flooring type, garage features, neighborhood rating (scale of 1-10), location rating (scale of 1 -10) etc.

Conclusion

Project Summary

- Built a linear model to predict house prices for 2 counties (Denton and Collin) in Dallas Fort Worth metroplex in Phase 1 with 82.7% accuracy. Improve the model to identify new data sources and gather additional attributes. Replicate this methodology to include all of US in Phase 2 and top global markets in Phase 4.

- Planned uses of the tool :
  - Predictor of current and future home values.
  - Investment tool for global real estate investments
  - A global standard for property valuation.

Key findings & Future Work

Phase 1: *Identify a data source that gives us new attributes that can be used to determine quality of house, age, neighborhood rating/safety, location benefits etc. to come up with a model that gives us at least 95% accuracy.*

Expand to other phases detailed in Slide 5 and countries identified in Slide 6.