🏡 USA Real Estate Market Analysis

---

1. Project Overview

Objective

The objective of this project is to analyze residential real estate listings across the United States to uncover **pricing patterns, geographic trends, property characteristics, and market structure**. The dashboard is designed to support **data-driven decision-making** for stakeholders such as investors, analysts, and real estate professionals.

Key Business Questions

- How are listings distributed geographically across the U.S.?
- What cities command the highest median prices and price per square foot?
- How does house size relate to listing price?
- Which property configurations (beds/baths) are most valuable?
- What proportion of the market consists of standard vs luxury listings?

---

2. Dataset Description

Data Source

Public real estate listings dataset containing residential properties across the United States.

Key Columns Used

- price, price_per_sqft
- house_size (sq ft), bed, bath
- city, state, zip_code
- status (for_sale, ready_to_build)
- listing_type (standard vs luxury/extreme)

Data Volume

- **Total listings:** ~683,000
- **Standard listings:** 672,636

- **Luxury / extreme listings:** 10,350

---

3. Data Cleaning & Preparation in Jupyter Notebook (python)

Steps Performed

- **Loading data:** Imported the dataset using pandas
- **Initial exploration:** Used df.shape to check the size of dataset and df.head() to look at the first rows (the sample subset of dataset)

	brokered_by	status	price	bed	bath	acre_lot	street	city
0	103378.0	for_sale	105000.0	3.0	2.0	0.12	1962661.0	Adjuntas
1	52707.0	for_sale	80000.0	4.0	2.0	0.08	1902874.0	Adjuntas
2	103379.0	for_sale	67000.0	2.0	1.0	0.15	1404990.0	Juana Diaz
3	31239.0	for_sale	145000.0	4.0	2.0	0.10	1947675.0	Ponce
4	34632.0	for_sale	65000.0	6.0	2.0	0.05	331151.0	Mayaguez

- **Column Standardization:** Standardize column names first using

  df.columns.str.strip().str.lower() .str.replace(' ', '_').str.replace('-', '_') for readability and documentaion
- Drop the invalid columns with df.drop(['unnamed:_12', 'unnamed:_13', 'unnamed:_14'], axis=1, inplace=True)
- Check the duplicates and data types using df.duplicated().sum() and df.dtypes
- Strip the data to avoid leading and tailing space issue; and standardize the datatype to consolidate data accuracy

- **Missing Data Handling:** check missing values with df.info(). Drop null values in critical columns (price, state, status) and impute nulls in (bed, bath) with median and in city column with unknown.
- **Handling Invalid Values:** Check the invalid values in key numerical columns and assigned as missing values pd.NA. Drop again the nulls in price and house size. Valiate null values with df.isna().mean().sort_values(ascending=False) for missing ratio and df.isna().sum() for number of missing values for each column

- **Feature Engineering:**

  Created price_per_sqft by dividing price by house_size

  Created prev_sold_year from prev_sold_date

- **Validate Data:** Used df.info() to check structure and df.describe(include="all") for summary statistics. Checking distinct states. Sorted the data according to "prev_sold_year" and "prev_sold_date".

```
<class 'pandas.core.frame.DataFrame'>
Index: 682986 entries, 0 to 1048573
Data columns (total 14 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   brokered_by     682367 non-null  Int64
 1   status          682986 non-null  string
 2   price           682986 non-null  Int64
 3   bed             682986 non-null  Int64
 4   bath            682986 non-null  Int64
 5   acre_lot        533643 non-null  float64
 6   street          678801 non-null  Int64
 7   city            682986 non-null  string
 8   state           682986 non-null  string
 9   zip_code        682937 non-null  Int64
 10  house_size      682986 non-null  Int64
 11  prev_sold_date  398244 non-null  datetime64[ns]
 12  price_per_sqft  682986 non-null  float64
 13  prev_sold_year  398244 non-null  Int64
dtypes: Int64(8), datetime64[ns](1), float64(2), string(3)
memory usage: 83.4 MB
```

	brokered_by	status	price	bed	bath	acre_lot	
count	682367.0	682986	682986.0	682986.0	682986.0	533643.000000	
unique	<NA>	2	<NA>	<NA>	<NA>	NaN	
top	<NA>	for_sale	<NA>	<NA>	<NA>	NaN	
freq	<NA>	666318	<NA>	<NA>	<NA>	NaN	
mean	51838.497184	NaN	556004.376639	3.298642	2.518656	13.810222	9353
min	2.0	NaN	1.0	1.0	1.0	0.010000	
25%	24866.0	NaN	210000.0	3.0	2.0	0.150000	
50%	52648.0	NaN	340000.0	3.0	2.0	0.250000	
75%	78552.0	NaN	539000.0	4.0	3.0	0.610000	
max	110142.0	NaN	2147483600.0	210.0	212.0	100000.000000	
std	29985.545813	NaN	3091643.80072	1.431049	1.365941	844.027916	5415

- **Exporting data:** Used the df.to_csv().

---

4. Data Analysis and Visualization in Power BI

Outlier Handling Strategy

- Identified and flagged extreme listings for transparency
- Extreme values were **not removed**

- Instead, they were **flagged and separated** to preserve market integrity and avoid misleading medians

DAX

- New Measure:

  Luxury or Extreme Listings

  Standard Listings

  Total Listing

- New Column:

  Acre Lot Outlier Flag

  Bed Bath Outlier Flag

  Clean Record Flag

  House Size Outlier Flag

  Price Outlier Flag

  Price per Sqft Outlier Flag

  Bed Category

Key Metrics (KPIs)

The dashboard highlights the following KPIs at the top level:

- **Total Listings:** 683K
- **Median Price:** $340K
- **Median Price per Sq Ft:** $182.64
- **Median House Size:** 1,800 sq ft
- **Median Bedrooms:** 3
- **Median Bathrooms:** 2
- **Data Coverage and Transparency (Standard Listings: 672636, Luxury or Extreme Listings: 10350)**

These KPIs provide a **quick snapshot of the U.S. residential market**.

---

5. Dashboard Design & Visuals

Visual Components Used

Visual	Purpose
KPI Cards	High-level market summary
Map (Bubble)	Geographic distribution of listings
Bar Charts	Top 10 cities by median price and price per sq ft
Scatter Plot + Trendline	Relationship between house size and price
Column Chart	Price comparison by bed category
Donut Chart	Listing status distribution
Slicers	Interactive filtering by city, price, beds, baths and house size

---

6. Key Insights & Findings

Market Overview

- The U.S. housing market shows a **median listing price of $340K**, with significant variation by city and property size.
- Listings are **densely concentrated in major metropolitan regions**, indicating strong urban housing supply.

Pricing Dynamics

- **Price per square foot** reveals value differences not visible in total prices alone, highlighting premium and value-driven markets.
- Larger homes generally command higher prices, but the **trend line shows diminishing returns** as size increases.

Property Configuration

- Homes with **4+ bedrooms** achieve the highest median prices, indicating strong demand for larger residential properties.

- Smaller homes remain competitive in price-per-square-foot terms.

Market Structure

- **Standard residential listings account for over 97%** of total inventory.
- Luxury and extreme listings form a small but influential segment, justifying separate treatment in analysis.

---

7. Stakeholder Value

For Investors

- Identify cities with high price-per-square-foot potential
- Understand demand patterns by house size and configuration

For Analysts

- Clear segmentation between standard and extreme listings
- Transparent handling of outliers

For Business Users

- One-page executive dashboard for fast decision-making
- Interactive filters to explore specific market segments

---

8. Tools & Technologies

- **Power BI** – Data modeling, DAX, and visualization
- **DAX** – KPI calculations and filtering logic
- **Excel / CSV** – Data preprocessing
- **Bing Maps** – Geographic visualization

---

9. Limitations & Assumptions

- Dataset represents listings, not final transaction prices
- Some regions may be overrepresented due to data availability

- Price trends do not account for temporal changes (time-series analysis not included)

---

10. Conclusion

This project demonstrates how **data visualization and analytics** can uncover meaningful insights from large-scale real estate data.
By combining KPIs, geographic analysis, and price-size relationships, the dashboard provides a **clear, actionable view of the U.S. housing market** while maintaining transparency and analytical rigor.

---

11. Future Enhancements

- Add time-based trend analysis
- Include mortgage rate or economic indicators
- Predictive modeling for price estimation
- Separate dashboards for luxury vs standard markets