REAL ESTATE Data Analysis Project

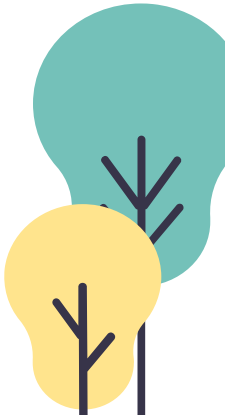
Overview

This project analyzes a comprehensive housing dataset containing over **2.2 million** property listings across the U.S., each described by **12** key attributes. These include **pricing** information, number of **bedrooms** and **bathrooms**, and **location** details. Notable features are **total land area** and **interior living space**. Additional fields capture the property's status **whether** ready for sale or sold, **broker** information, and **previous sale date**.

The dataset enables in-depth exploration of housing market trends and property characteristics.

U.S.
REAL ESTATE
MARKET

TABLE OF CONTENTS



DATA CLEANING: DEALING WITH

Unnecessary Columns

Excluded the broker agency & street names as both had so many categories, making them almost meaningless.

Unresonable Timeframe

Excluded incorrectly written dates.

Incorrect Location

Dropped the properties located outside the US, to maintain geographical consistency.

DATA CLEANING: DEALING WITH

Inconsistent City Names

Mapped the city names to data from the US Postal Service based on the zip code to ensure its consistency.

Incorrect Values

Excluded properties that had any incorrect entries.

All clean to go!

FEATURE ENGINEERING

Computed the price per square foot for every property

Standardizes pricing by property size to allow fair comparisons and reveal pricing inconsistencies across listings.

Flagged if the property was recently sold or not

Adds a binary filter to highlight recent sales (before vs after 2020), enables trend tracking against historic data.

Grouped properties by seasons it was sold in

Categorizes each sale by season to uncover demand cycles and support seasonal pricing analysis.

OUTLIERS

Price

Dropped **1%:**
Moderate house size side sold for over $2 billion

of Bedrooms

Dropped **0.5%:**
Propeties with over 200 bedrooms

of Bathrooms

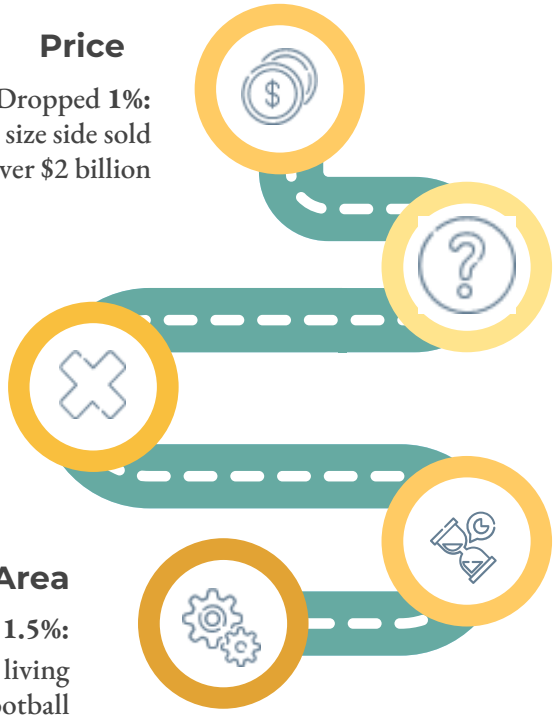
Dropped **0.9%:**
Properties with over 400 bathrooms, meanwhile having 1-5 bedrooms

Land Area

Dropped **1.7%:**
Properties that had land ares of 10,000 acres, almost the size of cities.

Living Area

Dropped **1.5%:**
Properties with living areas as big as 17 football fields



DATA VISUALIZATION

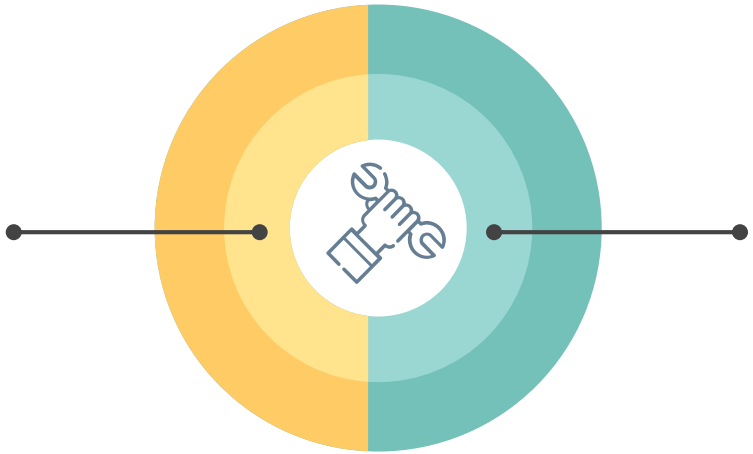
Streamlit Application

MISSINGS

Minimal Percentage

Dropped

Properties with missing features were excluded during data cleaning, as long as the proportion of missing data remained within the acceptable 5% treshold.



Significant Percentage

Imputed

During the data preprocessing, the mssing values below the 40% threshold were imputed numerically and categorically.

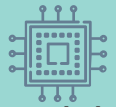
DATA PREPROCESSING

Separated features as input and target variables for modelling.

Divided the dataset into training and testing datasets

Imputed missing values in numerical columns

Applied scaling to numerical features for normalization



Encoded categorical variables into numerical format

THANK YOU!