**Project Background**

**Motivation**: A couple would like to invest in real estate in Massachusetts, USA. They are specifically looking for a multifamily property and have a maximum budget of $800,000. It’s no news that house prices are affected by certain factors; hence, the prices fluctuate. Knowing the best time to purchase a property can give a good Return on Investment (ROI). They have asked for my help as a data analyst so they can make data-driven investment decisions.

**Scope**: This project will look at real estate listings within Massachusetts, USA.

**Objective**: To understand the housing market in Massachusetts and identify the factors that influence house prices.

**Data source:** The data was sourced from Kaggle- [USA Real Estate Dataset](https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset)

**Data collection method**: Data was collected from realtor.com via web scraping. Realtor.com is a real estate listing website operated by the News Corp subsidiary Move, Inc. and based in Santa Clara, California. It is the second most visited real estate listing website in the United States as of 2021, with over 100 million monthly active users.

**Data contents**: the data contains real estate listings in the US, categorized by states and cities. It has 12 variables; status, price, bed, bath, acre\_lot, full\_address, street, city, state, zip\_code, house\_size, and sold\_date.

**Data relevance:** The data shows real estate listings across the US. It contains relevant information that meets the project objectives.

**Data profile**

*Variables and Data Types:*

Status: qualitative, time-invariant, nominal

Price: quantitative, time-variant, discrete

Bed: quantitative, time-invariant, discrete. Though quantitative, at some point in the analysis it will be treated as qualitative (i.e categorical)

Bath: quantitative, time-invariant, discrete. Though quantitative, at some point in the analysis it will be treated as qualitative (i.e categorical)

Acre\_lot: quantitative, time-invariant, continuous

Full\_address: qualitative, time-invariant, nominal

Street: qualitative, time-invariant, nominal

City: qualitative, time-invariant, nominal

State: qualitative, time-invariant, nominal

Zip\_code: qualitative, time-invariant, nominal

House\_size: quantitative, time-invariant, discrete

Sold\_date: quantitative, time-invariant, continuous

**Summary (before data selection and cleaning):** 12 variables, 923,159 records

**Data Integrity Issues (Changed/Fixed Records)**

A subset of the original data was used for analysis. Below is the data flow as changes were made. Total- is the number of records after each action.

**Summary (after data selection and cleaning):** 8 variables, 3,048 records

**Quantitative (summary statistics)**

Table

Description automatically generated

**Qualitative**

City

* Value count-346
* Mode: Boston

Zip\_code

* Value count-433
* Mode-1201

**Data limitations/ethics**

* There are some controversies about web scraping, however, it is considered ethical if the rules of the website are respected.
* Other factors which could potentially influence home prices like age of the property, type of property (condo, single family, multi-family, townhome etc), interest rates, location in terms of quality of schools and proximity to key areas were not captured in the available data.
* There were a lot of duplicate records and missing records that had to be removed. I chose to remove them and not impute data so that I can use the actual data for prediction.
* The data doesn’t specify the type of house. This would have been useful since I am more interested in multi-families.
* There was large variation in the number of datapoints across the years. This could have impacted the time series analysis.

**Key Questions**

Which location have the highest number of houses sold?

Which locations have the highest and lowest prices?

Is there a trend in house prices?

What factors affect house prices?

What is the most popular number of bedrooms and bathrooms?

What is the most common price range? What is the minimum, maximum, average and median price of a house?

What is the price forecast for the next two years?

What features should I consider in terms of number of beds, baths and house size?

When is the best time to purchase a property?

**Hypothesis:** The price of the house increases with increase in number of baths and house size.