Real Estate Data Analysis and Visualization

Team Members: Matthew Adent, Cecilia Rocha, Asia Byrne, Arya Maredia, Josh Ehlke

Date: 2/20/2024

Tools Used: SQL, Python (Pandas, Plotly, Dash/Streamlit), Tableau/Power BI

1. Introduction

The real estate market is constantly evolving. Property listings are affected by a range of factors including economic shifts, interest rates, geographical location, local amenities, and seasonal demand fluctuations. This project aims to analyze a comprehensive dataset of real estate listings to extract meaningful insights into **property pricing, types of homes, geographic distribution, and market trends over time**.

By leveraging **SQL for data extraction, Python for data analysis, and interactive visualization tools like Python Anywhere**, we provide an in-depth view of property listings and trends. Our goal is to make real estate data **more accessible, interactive, and insightful**.

---

2. Purpose of the Project

The primary objective of this project is to **understand real estate market trends**, identify **patterns in property pricing**, and listings. By analyzing price distributions, and listing trends over time, we can answer key real estate questions such as:

- How do listing prices vary across different counties and cities? We wanted to know the geographical distribution of real estate prices. By exploring how prices differ from county to county, and from city to city, we can pinpoint which areas are seeing higher demand or price growth.
- How has the real estate market's offerings (distribution) changed from the 80s compared to the 2010s? By identifying the most frequent price ranges for properties, we can better understand affordability, price trends and locations that are thriving.
- What are the common price ranges for listed properties? By identifying the most frequent price ranges for properties, we can better understand affordability, price trends and locations that are thriving.

---

3. Why Data Applications?

Real estate data is **large, structured, and constantly changing**. Using **data applications** provides:

- **Efficient storage & retrieval:** SQL databases allow fast querying and filtering of large datasets.
- **Automated analysis:** Python scripts can clean, transform, and summarize real estate trends.
- **Scalability:** The application can handle new real estate listings dynamically.

---

4. Why Interactive Visualizations?

Traditional spreadsheets and reports limit data exploration. **Interactive dashboards provide**:

- **Dynamic filtering:** Users can adjust time ranges, property types, and locations.
- **Geospatial analysis:** Maps can highlight price variations across cities/counties.
- **Historical trends:** Time-series charts show how the market has changed.
- **User engagement:** Investors can explore pricing patterns based on specific filters.

Using **\Python Anywhere**, we create a **visually engaging experience**.

---

5. Dataset Overview

The dataset contains real estate listings with the following attributes:

- **Property Information:**
  - `id`: Unique identifier for each property
  - `stateId, countyId, cityId`: Location details
  - `country`: Country of listing
- **Listing Details:**
  - `city:` City listing is located
  - `price:` Price of property
  - `bedrooms:` Amount of bedrooms in the property
  - `bathrooms:` Amount of bathrooms in the property
  - `datePostedString`: Date when the property was listed
  - `homeType`: Type of home listed like, apartment, condo, lot, multi family, single family, townhouse
- **Pricing Data:**
  - Price distributions and value ranges across multiple properties.

Motivation for Choosing the Dataset

This dataset provides **rich insights into the real estate market**, allowing for:

- **Price analysis** across different counties or cities.
-

Data Cleaning

We went through the following steps to ensure that our data was cleaned sufficiently for the purpose of this project:

- Loaded the data into a DataFrame within a Jupyter notebook for easier handling
- Dropped rows where the “hasBadGeocode” attribute was true
- Dropped rows where the price was 0, since price is arguably THE main factor to consider when buying a house
- Dropped columns that we deemed irrelevant for the scope of our project. For example, the house description column was irrelevant because we wanted to create a dashboard that gives quick, mostly numerical summaries of thousands of properties, and having a text description on each point on the map would clutter the pop-up
- Convert the columns into appropriate types. As an example, “price” into int (since all amounts were in whole dollars), “homeType” into string

---

6. Example SQL Queries

1. Get the Total Number of Properties Listed per State

sql
CopyEdit

```
SELECT streetAddress, city, zipcode, price, bedrooms, bathrooms,
livingArea, yearBuilt, homeType
    FROM real_estate
    WHERE city LIKE '%{city}%'
```

2. Find the Average Price of Properties by County

sql
CopyEdit

```
SELECT REPLACE(county, ' County', '') AS county, ROUND(AVG(price), 2)
AS avgPrice
```

```
    FROM real_estate
    WHERE city LIKE '%{city}%' AND county IS NOT NULL
    GROUP BY county
```

3. Property Types

sql
CopyEdit

```
SELECT homeType, COUNT(*) AS count
    FROM real_estate
    WHERE city LIKE '%{city}%'
    GROUP BY homeType
;
```

---

7. Walkthrough of the Data Application

The application allows users to:

1. **Search properties** by location, price range, and listing type.
2. **Filter data dynamically** based on cities, and prices
3. **View trends over time** with interactive graphs.
4. **Compare property prices** between locations.

Screenshots of the App: 📷 *(Include screenshots of your interactive dashboard and data visualizations.)*

Real Estate Map Dashboard

City:

Enter city name

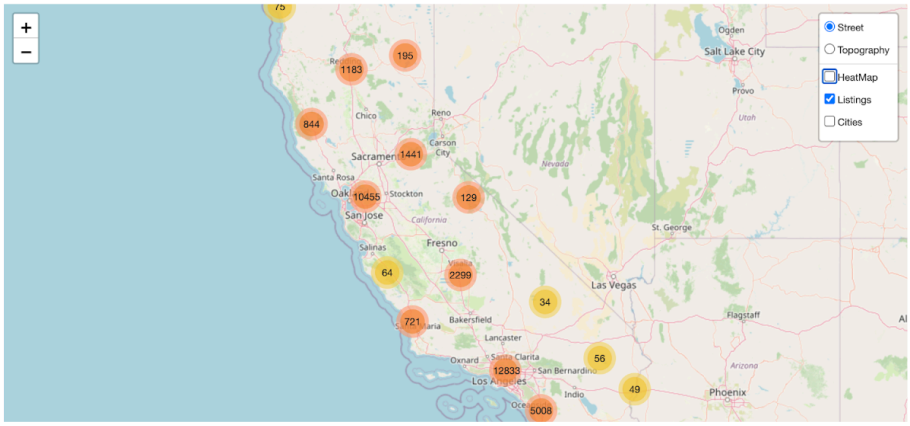
Min Price:

Min price

Max Price:

Max price

APPLY FILTERS



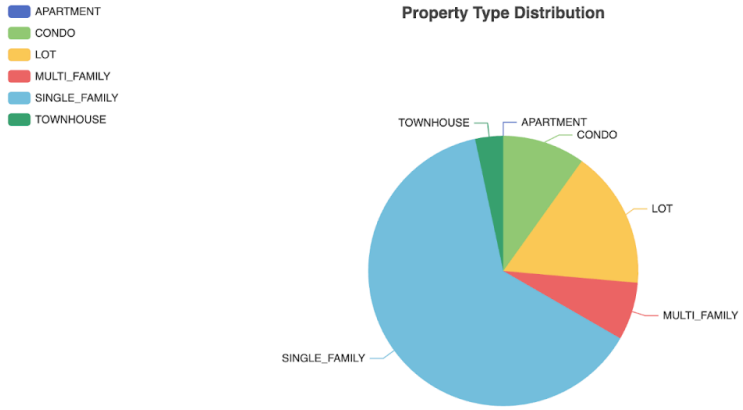
Real Estate General Visualization Dashboard

City:

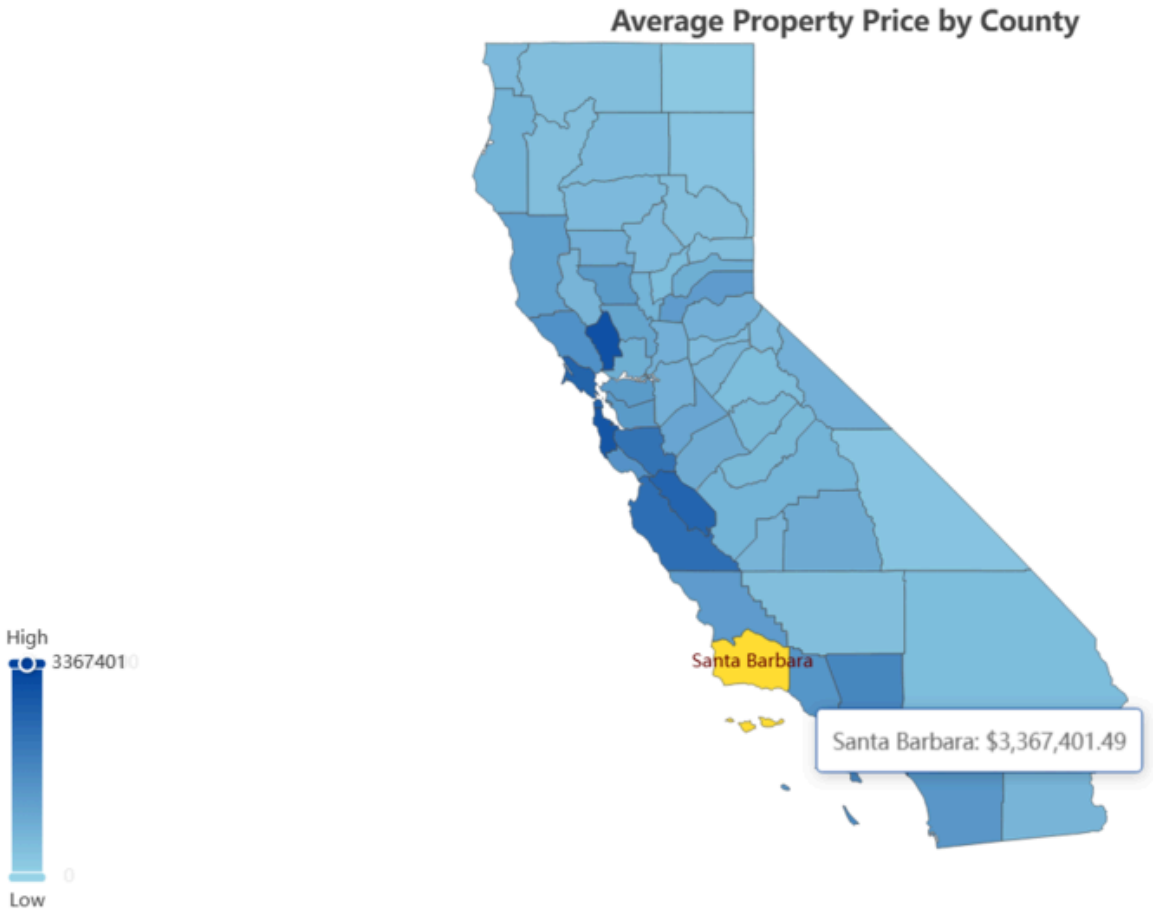
Enter city name

APPLY FILTERS

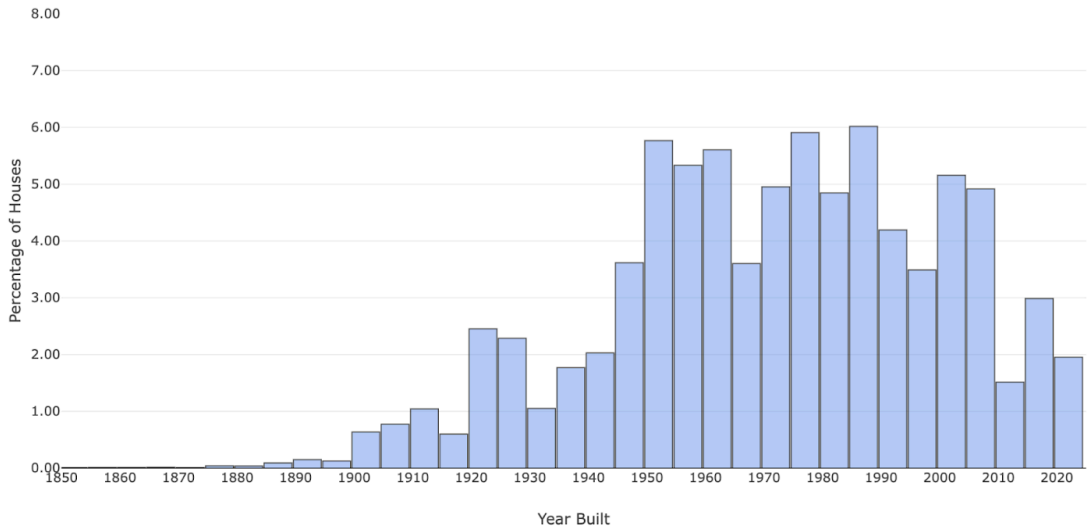
Property Type Distribution



Average Property Price by County



Property Age Distribution



Property Data Table

Show 10 entries Search:

Street Address	City	Zipcode	Price ($)	Bedrooms	Bathrooms	Living Area (sqft)	Year Built	Home Type
"0 OFerrall Dr"	Perris	92570.0	400000	0	0	0	0	LOT
"9810 ONeil Ct"	Jamestown	95327.0	539000	4	3	2148	1974	SINGLE_FAMILY
0	Earlimart	93219.0	112000	0	0	0	0	LOT
0	Brawley	92227.0	15000	0	0	0	0	LOT
0	Friant	93626.0	512000	0	0	0	0	LOT
0	San Joaquin	93660.0	69000	0	0	0	0	LOT
0	Pixley	93256.0	0	0	0	0	0	LOT
0	Kerman	93630.0	97500	2	0	0	1960	LOT
0	Exeter	93221.0	19500000	0	0	0	0	SINGLE_FAMILY
0 0 Alvord Mountain Rd	Newberry Springs	92309.0	60000	0	0	0	0	LOT

Showing 1 to 10 of 35,386 entries Previous 1 2 3 4 5 … 3,539 Next

---

8. Dashboard Design

The dashboard consists of:

1. **Geospatial Analysis (Map)**
    - Interactive **heatmap of property prices by location**.

2. **Time-Series Trends**
    - Line chart showing the **number of listings over time**.
3. **Property Distribution**
    - Histogram to **analyze property price ranges**.
4. **Filter Controls**
    - Dropdowns for cities, and price ranges.

Designed in **Tableau / Power BI / Dash**.

---

9. Questions Answered by the Dashboard

1. What are the most expensive and affordable cities in California for real estate?
2. How do property prices compare between urban and rural counties? (Ex. Los Angeles County for urban, Alpine County for rural)
3. What are the most popular types of homes?
4. What are the most common price ranges for properties?
5. Are there seasonal patterns in property listings?

---

10. Limitations & Bias in Data

- **Incomplete Data Collection**: The dataset is only from 2021-2024 and was only collected for the first six months of each year. This restricted our ability to account for seasonal or annual fluctuations in the real estate market. A more complete dataset that spans the entire year or several years would have provided a clearer picture of market dynamics.
- **Incomplete Data Records**: The dataset did not include complete records of sale prices, like if properties have changed ownership multiple times. This creates gaps in understanding property value trends, especially in high-demand markets. Without this information, we may not have captured the full picture of property values.
- **Missing Population Demographics:** Information such as household income, age distribution, and employment information, are crucial factors in determining housing demand and price fluctuations. Without understanding the demographics of the populations in various regions, it's difficult to correlate pricing trends to potential shifts in demand.
- **Property Conditions:** There was no data provided regarding the conditions of the homes or any renovations that had been made. Properties that have undergone significant renovations or upgrades may command higher prices, but without this information, it was difficult to assess the true value of a property based on its physical condition.

- **Impact of External Factors:** There is no accounting for external influences like natural disasters, economic crises, or shifts in local policies, which can drastically affect real estate prices.
- **Data Collection Bias**: The dataset may favor certain cities with more listings.
- **Market Changes**: The dataset reflects historical data, but real estate trends fluctuate.
- **Outliers in Prices**: Some properties may have extreme prices that skew averages.z
- **Price Bias in Modern vs. older Homes:** The dataset may introduce a bias in property prices, particularly in the valuation of modern homes versus older ones. Modern homes could be overvalued compared to historic older homes.
- **Potential Sampling Bias:** The data collected might have been biased due to how properties were sampled. If certain types of properties or markets were disproportionately represented—such as luxury homes or specific geographic areas—the results may not reflect the broader trends in the real estate market.

---

11. Conclusion & Future Work

Conclusion

This project successfully analyzed **real estate listings** through **data applications and interactive visualizations**, uncovering key insights about:

- Property price distributions
- Geographic trends in real estate
- Changes in listings over time

Future Work

- **Expand data sources**: Include Zillow, Redfin, or other property databases.
- **Enhance predictive modeling**: Use machine learning to **forecast property prices**.
- **Improve real-time updates**: Implement an automated **ETL pipeline** for live data.
- **Add user interactions**: Allow users to **input budget constraints and get recommendations**.