DATA ANALYSIS OF THE BAY AREA REAL ESTATE MARKET

PROJECT PRESENTATION: BAN 612

Contents

Meet the Team- Group 1

SHRIYA ARORA

AASTHA TANDON

AMRUTH REDDY

PALLAVI NAIR

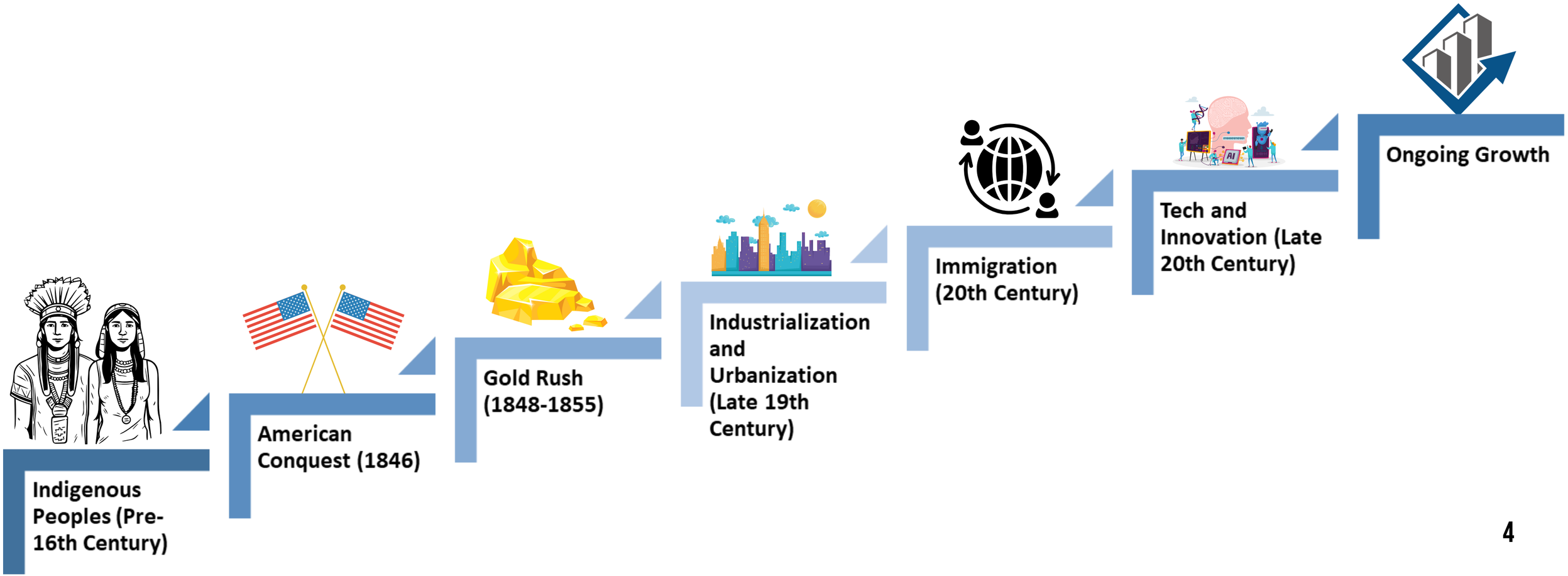
ELSIE YUAN

AISHWARYA RAVISHANKAR

WELCOME TO CALIFORNIA GOLD COUNTRY

1850 : 92,597 PEOPLE → **2023 : 39,0000000 PEOPLE***

*reported as of May 2023

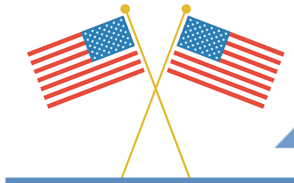


WELCOME TO CALIFORNIA GOLD COUNTRY

1850 : 92,597 PEOPLE → **2023 : 39 MILLION PEOPLE***

*reported as of May 2023

Indigenous Peoples (Pre-16th Century)



American Conquest (1846)

Gold Rush (1848-1855)

Industrialization and Urbanization (Late 19th Century)

Immigration (20th Century)

Tech and Innovation (Late 20th Century)

Ongoing Growth

“WE ARE LOOKING FOR AN INVESTMENT PROPERTY”

Alex and Sara

Source: iStock

Questions from Alex & Sara:

1. **Price variation by nearby places: Groceries, Restaurants, Parks, etc.**
2. **COVID-19 and Price Fluctuations: four-year period from 2019 to 2022**
3. **Property Price vs Inflation/Interest Rates & Correlation**
4. **Can we develop a predictive model to forecast future property prices?**
5. **Investor Perspective**

Interactive Map prepared for Alex & Sara

How we helped them

REDFIN

Acquire Data : Web Scrapping	Clean & Pre-Process Data	Perform Exploratory Data Analysis	Applying Appropriate Data Analytics Techniques	Develop Models	Interpret & Evaluate Results

Data Scrapping

1

Automating
Browser Actions

2

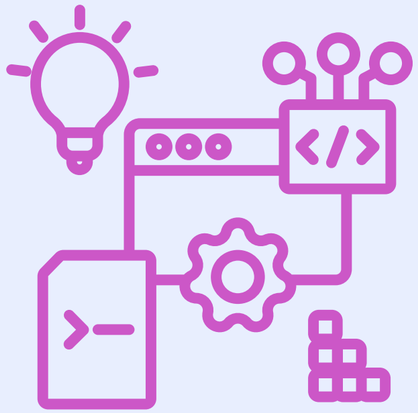
Internet
Protocol
Management

3

Step 1 - Initial Data Extraction

Step 2 - In-Depth
Data Parsing

4



Data
Integration

5

Data acquired !!

Data Collection : Scraping Redfin Website

- **Automating Browser Actions**: Leveraged Selenium to automate key browser actions such as button clicks, form filling, and page navigation.

- **IP management**: Used VPNs with frequent IP rotations and disabled automation flags in the Selenium driver to ensure uninterrupted data collection.

- **Repeatable Two-Step Data Collection for Each Bay Area Location:**

    - **Step 1 - Initial Data Extraction**: Interacted with the Redfin website, and utilized XPath, to gather essential data, including house URLs, for each location.

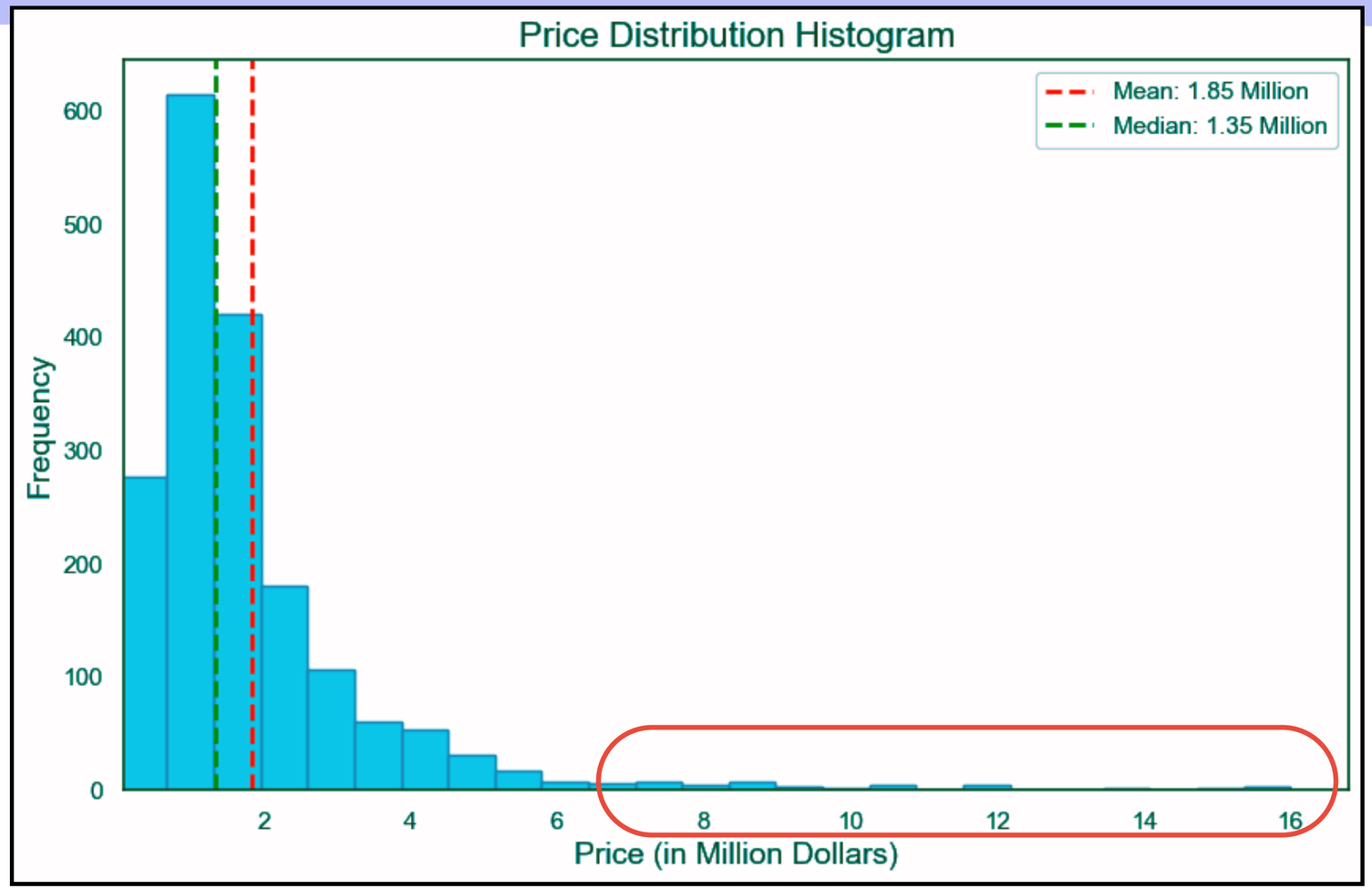
    - **Step 2 - In-Depth Data Parsing**: Processed each house URL using Beautiful Soup. Employed regex to extract various house-specific details, including **school ratings, interior specifications like flooring, heating, cooling, garage; walk, transit, and bike scores, nearby outlet counts, Redfin estimates, buyer compensation, estimated monthly costs, property sales history, and associated risk factors,** etc.

- **Data Integration**: Concatenated the data obtained from both Step 1 and Step 2, providing a comprehensive dataset for each location.

- **Final Dataset**: Once this is done for each of the Bay Area locations, datasets from entire locations are merged into a single, cohesive file in the Data Preprocessing phase.

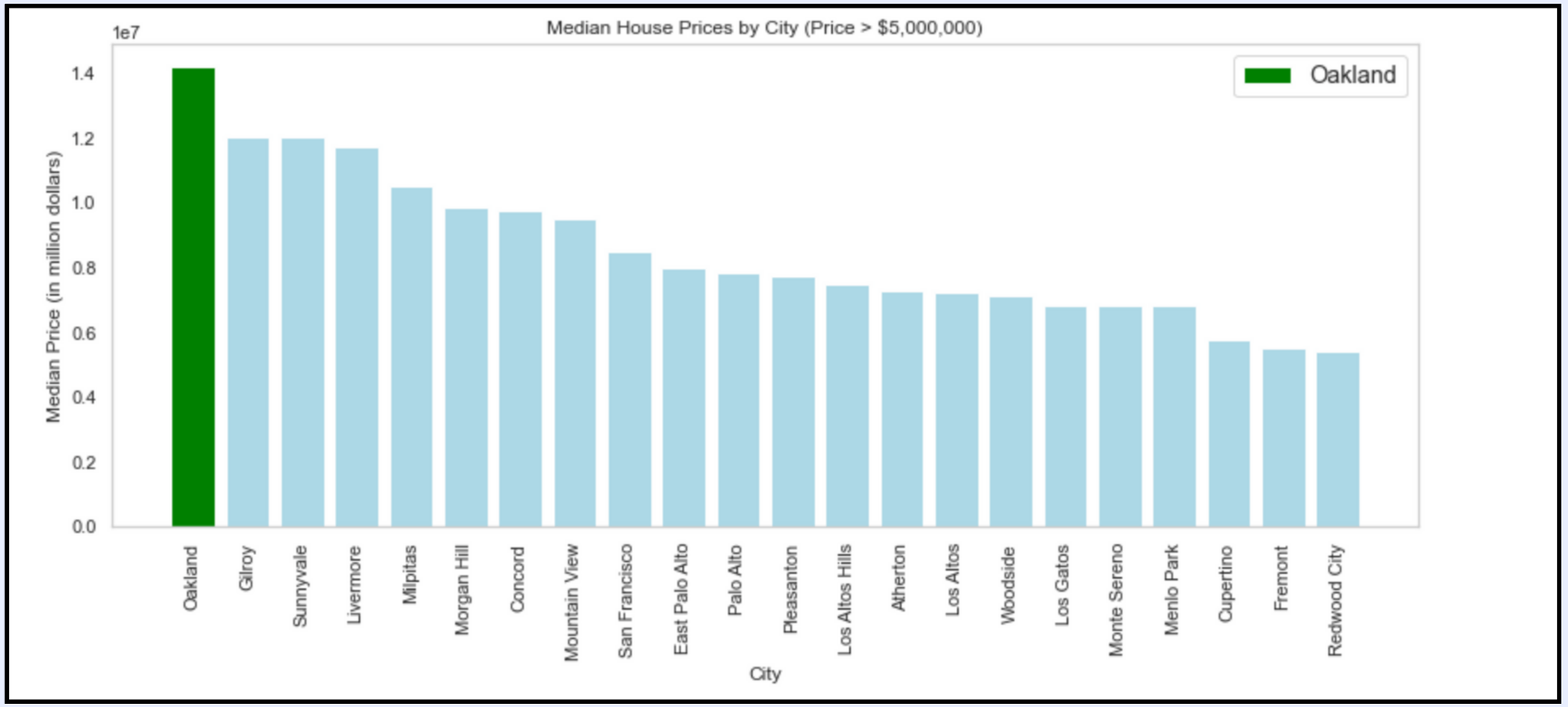
Exploratory Data Analysis

Dataset Summary: 1946 observations, 35 features



Based on the entire dataset: Houses sold between 2000 to 2023

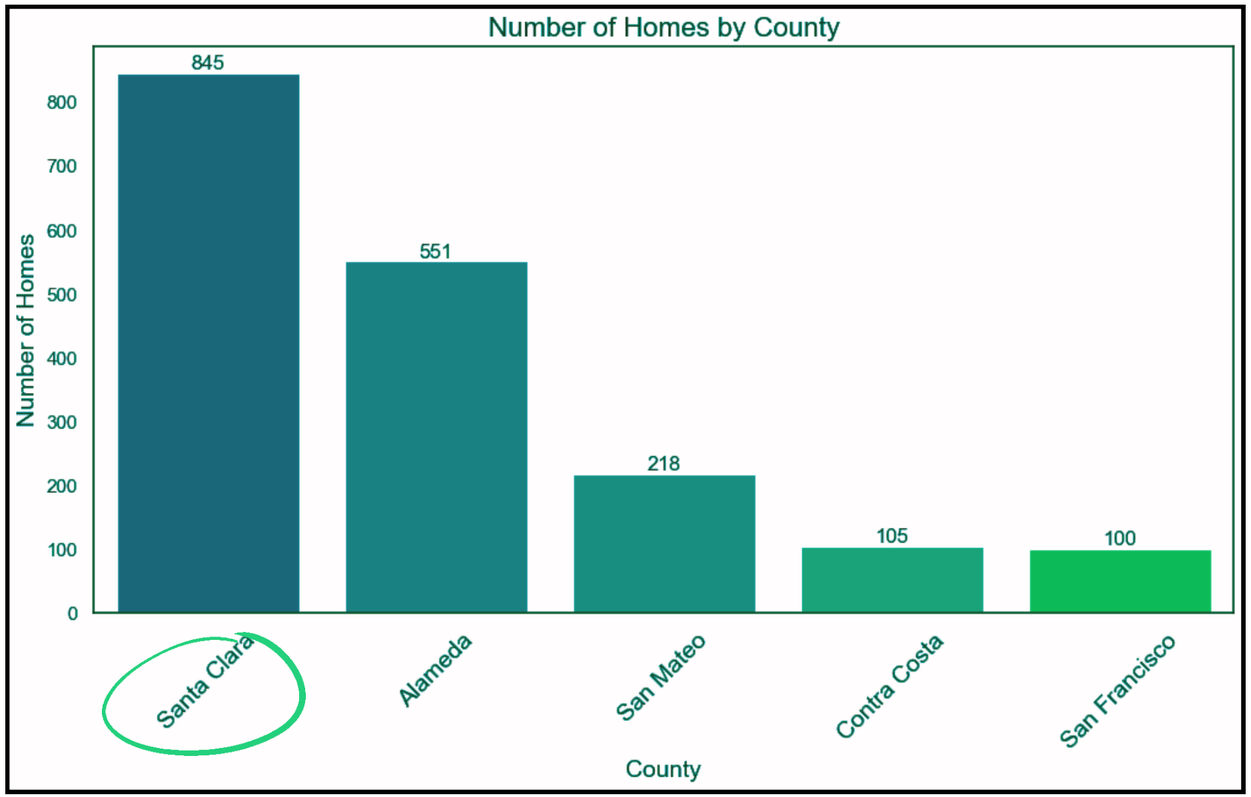
Median House Prices across Cities



- Cities where median prices are greater than 5 million dollars
- Oakland has the highest median price : 14 million dollars

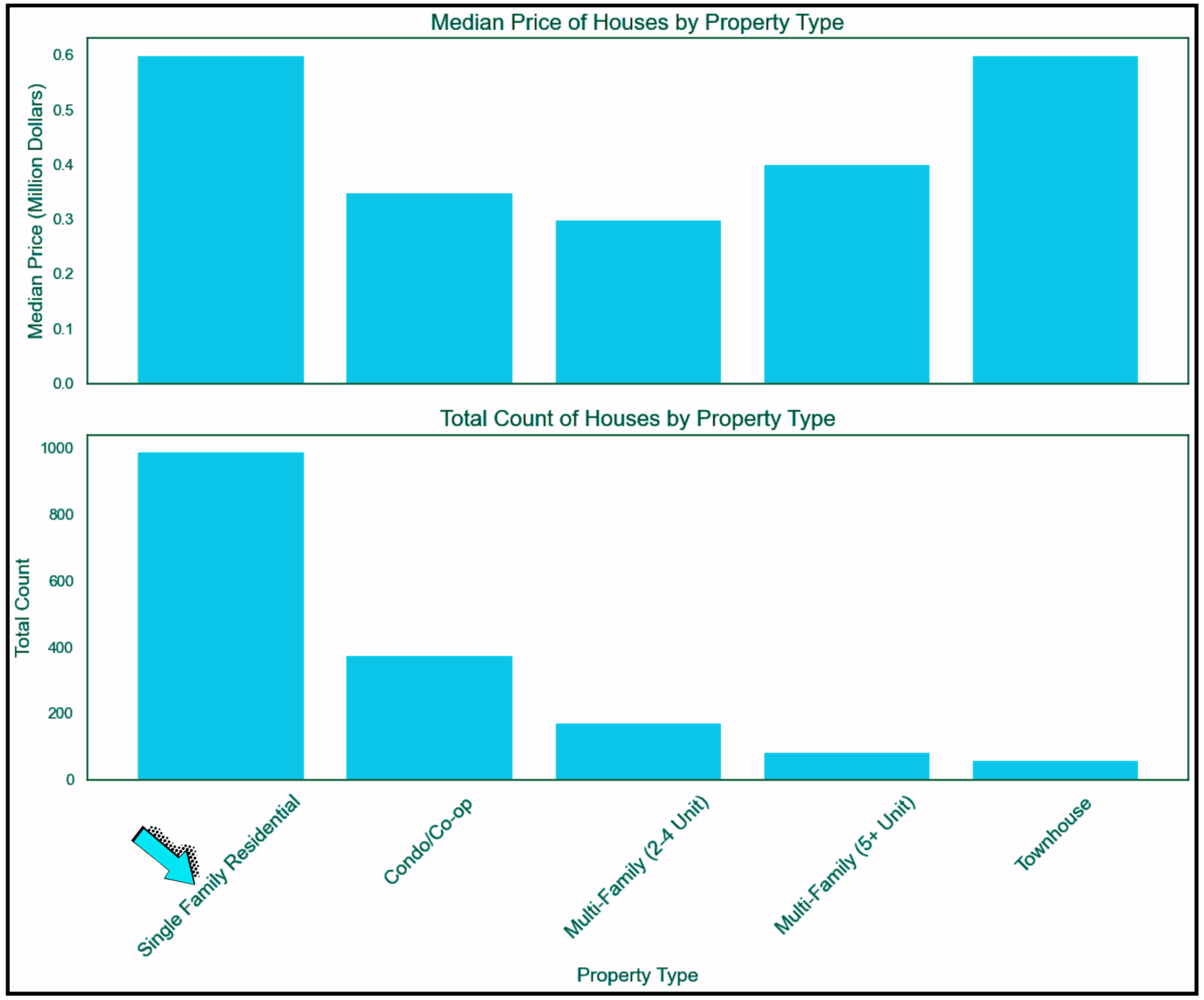
Exploratory Data Analysis

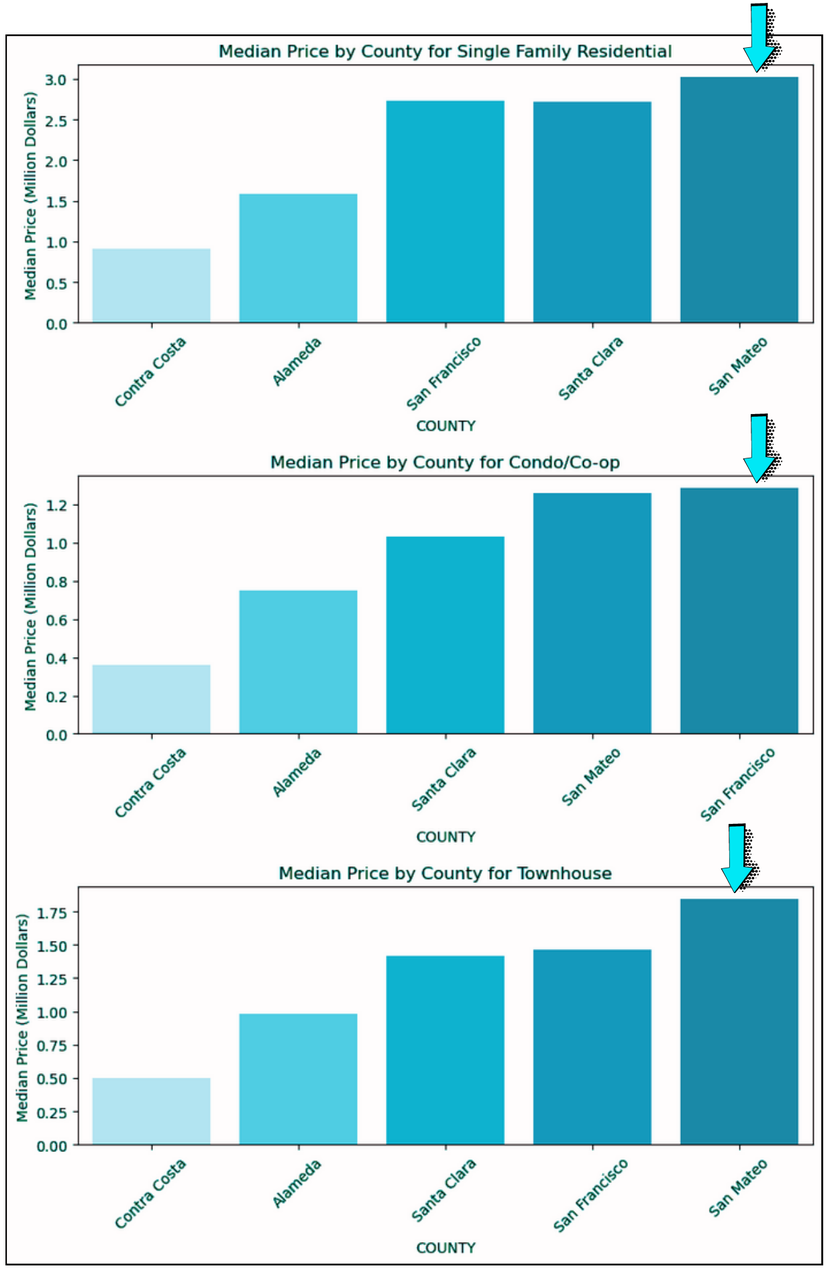
The cities extracted were mapped along the counties for further analysis



- Santa Clara county has the highest number of homes
- San Mateo county has the highest average price

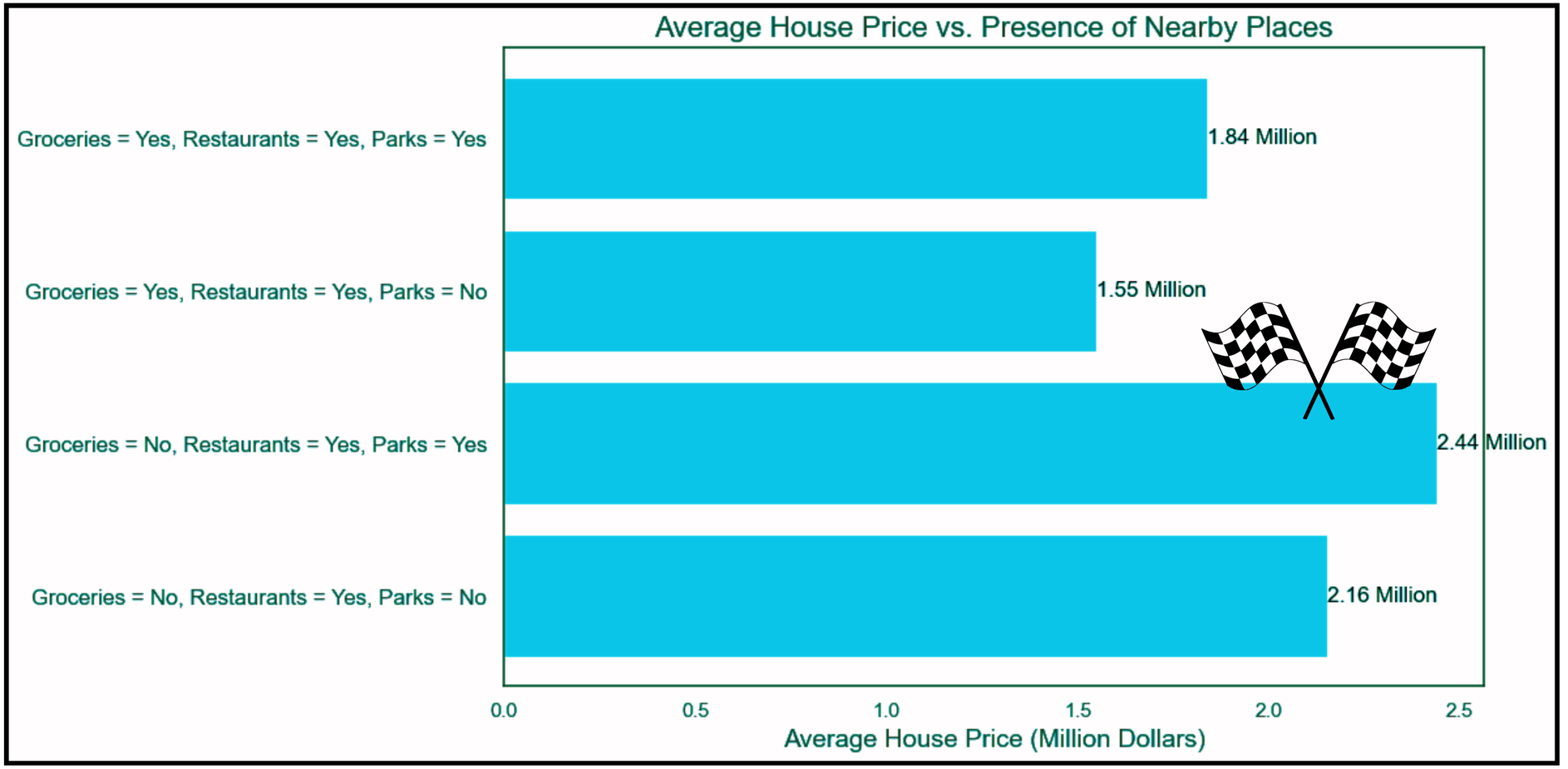
Exploratory Data Analysis





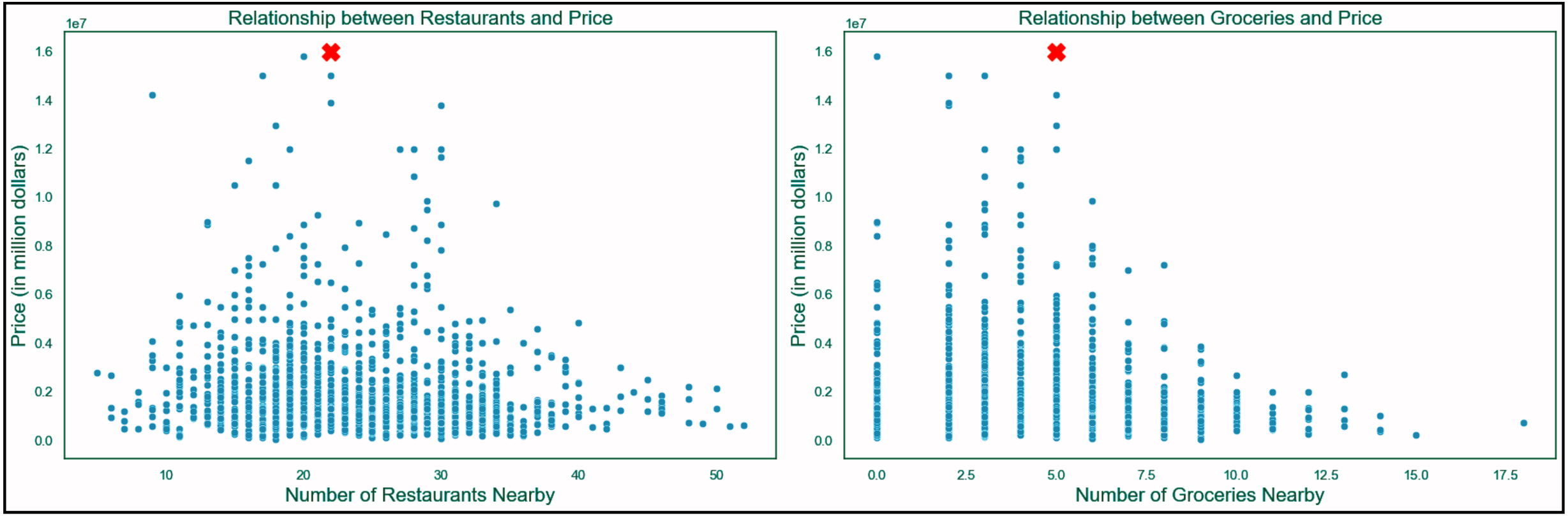
- **Single Family homes are highest in number across Bay Area with median price almost equivalent to Townhouses**
- **San Mateo county has the highest median price for Single Family & Townhouse, SF has highest median price price for Condo**

Price variation by nearby places (1/2)



- **Houses near both restaurants and parks have the highest average price, standing at approximately $2.44 Mn**
- **Properties located near restaurants also command premium prices ($2.16 Mn)**
- **Combinations of nearby amenities have a significant influence on housing prices.**

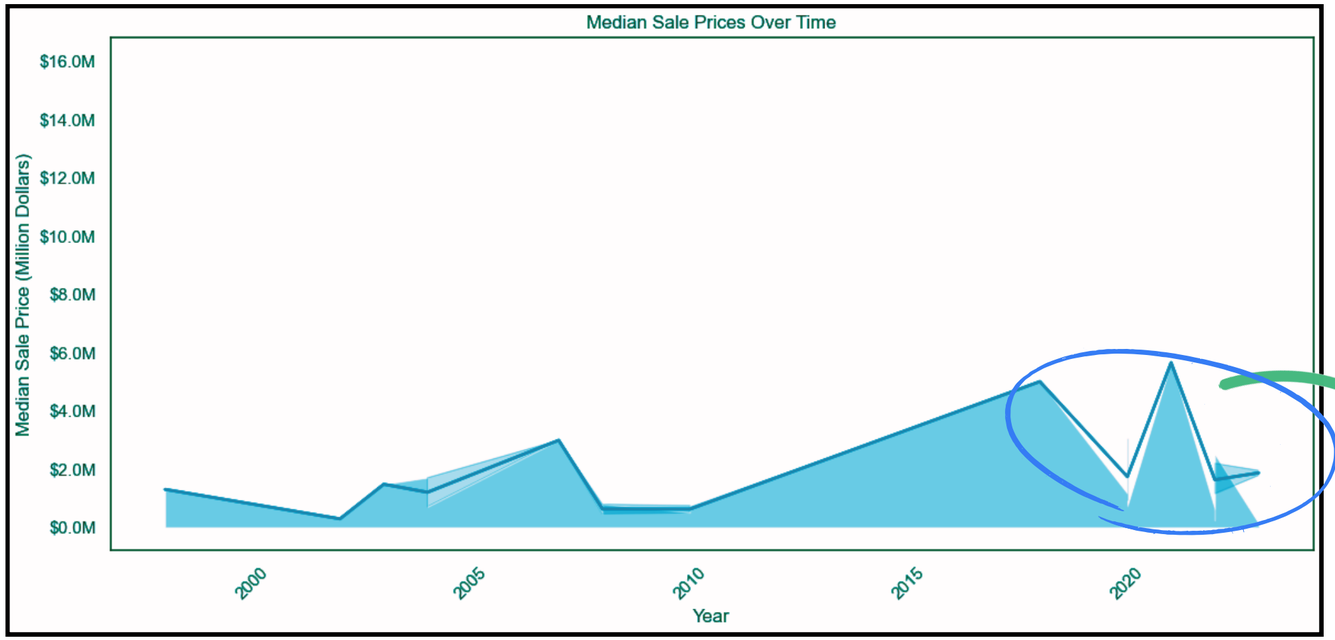
Price variation by nearby places (2/2)



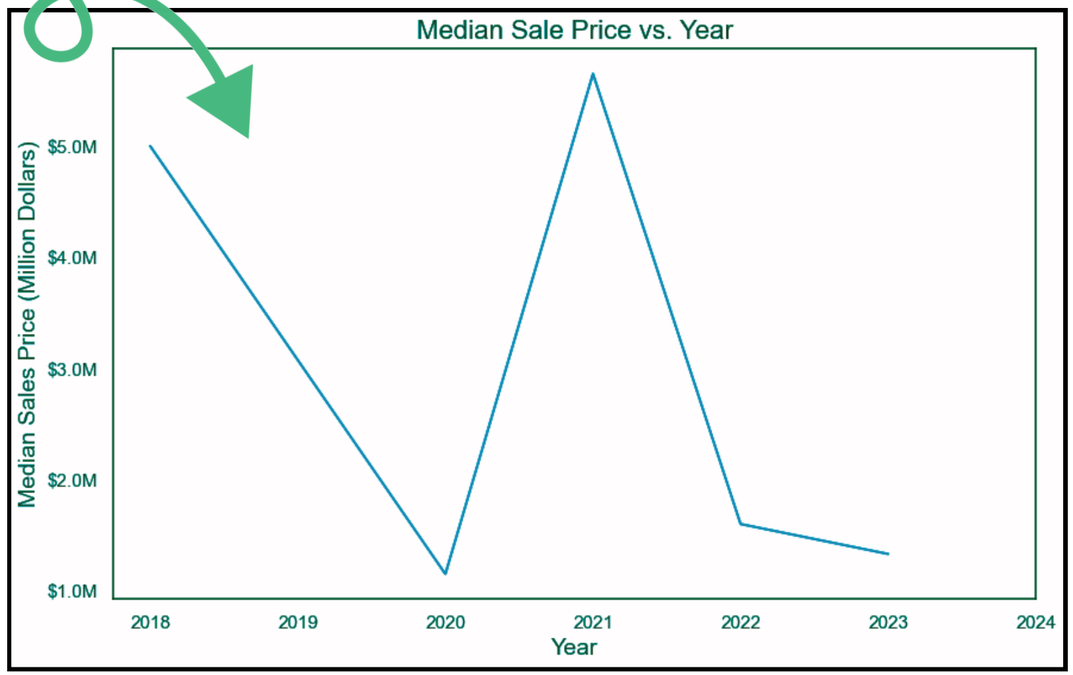
- **Notably, the scatter plots identifies an interesting data point – the property with the highest price.**
- **Property stands out due to its exceptional price, and it's association with 5 nearby grocery stores and 23 nearby restaurants.That said, the charts show variability in pricing and the relationship isn't strictly linear.**

What do these findings mean for Alex & Sara -"Consider these areas as attractive opportunities."

COVID-19 and Price Fluctuations

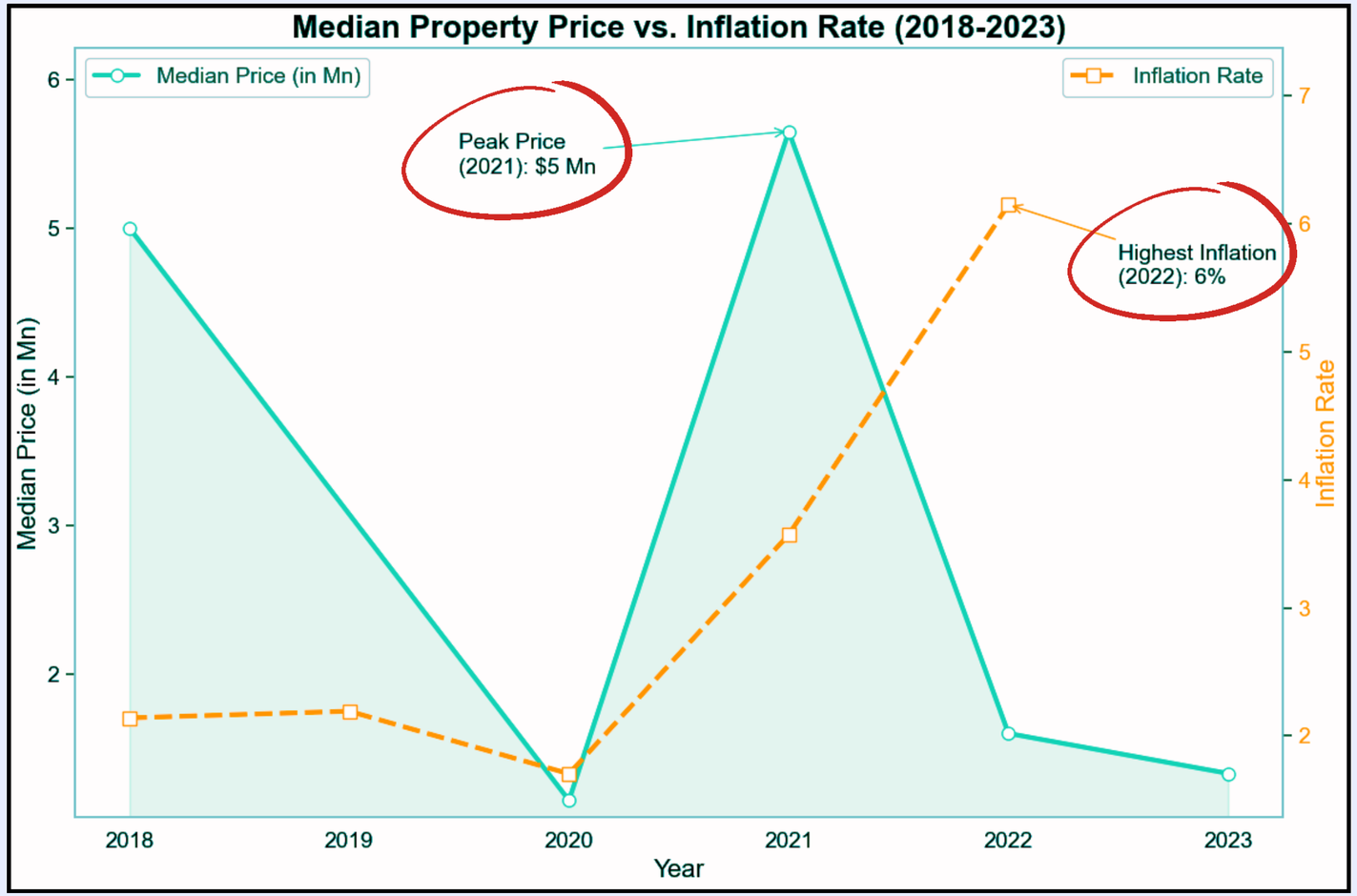


- Left chart presents the median sales price of properties from 2000 to 2023
- Includes outliers to analyse the extremes for the Covid -19 analysis
- Let's zoom in at how the COVID-19 pandemic influenced real estate prices over the years-



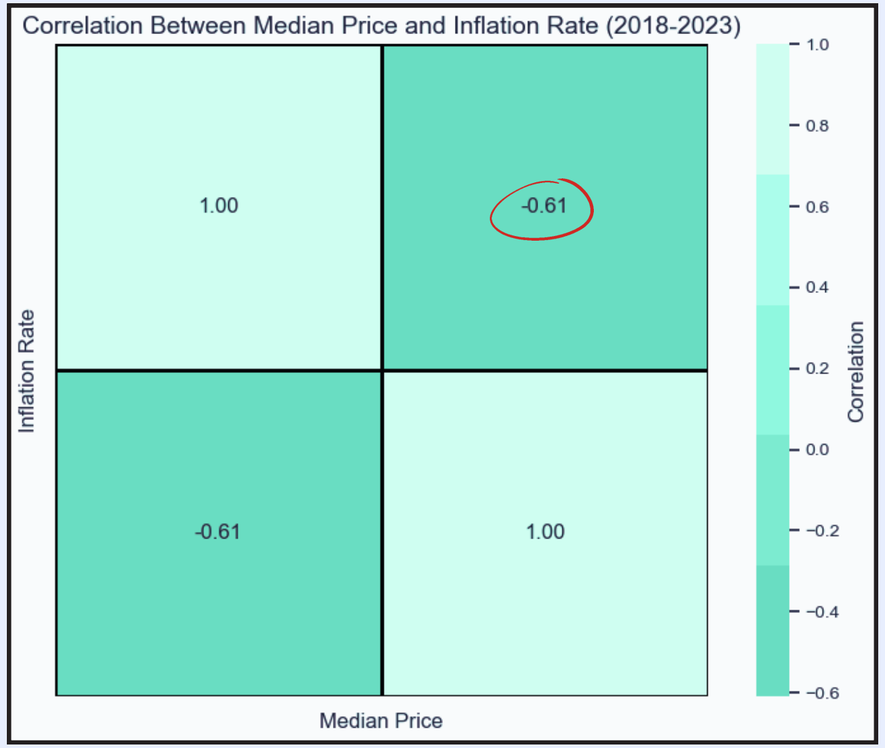
- In 2018, we observe a gradual upward trend in median prices
- In 2020, a notable disruption occurred—COVID-19 hit. Impacting the real estate landscape, where housing prices dropped by 45%
- Move forward to 2021, shows a remarkable recovery in the median prices of 217%
- In 2023 the median prices continue their upward trajectory, suggesting a strong and resilient real estate market

Median Property Price vs. Inflation Rate



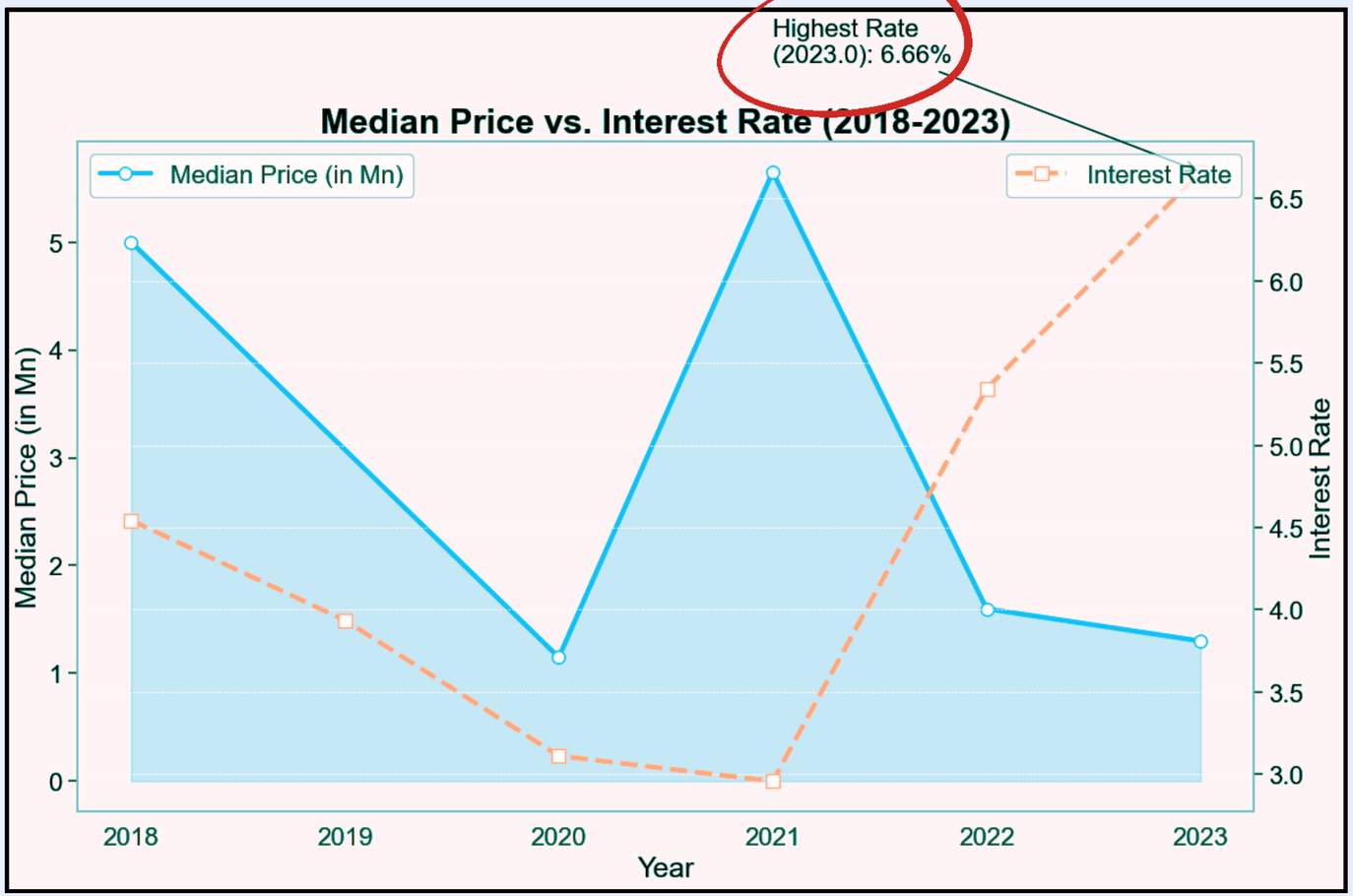
- On the top, in the line plot the pastel green represents the median property prices from 2018 to 2023
- The dashed orange line represents the inflation rate, highest rate occurred in 2022 - 6%

- Heatmap below, exhibits a moderate negative correlation of -0.61
- When inflation is high in 2022, it can put downward pressure on property prices.
- In 2021, when inflation is low or negative, makes real estate an attractive investment



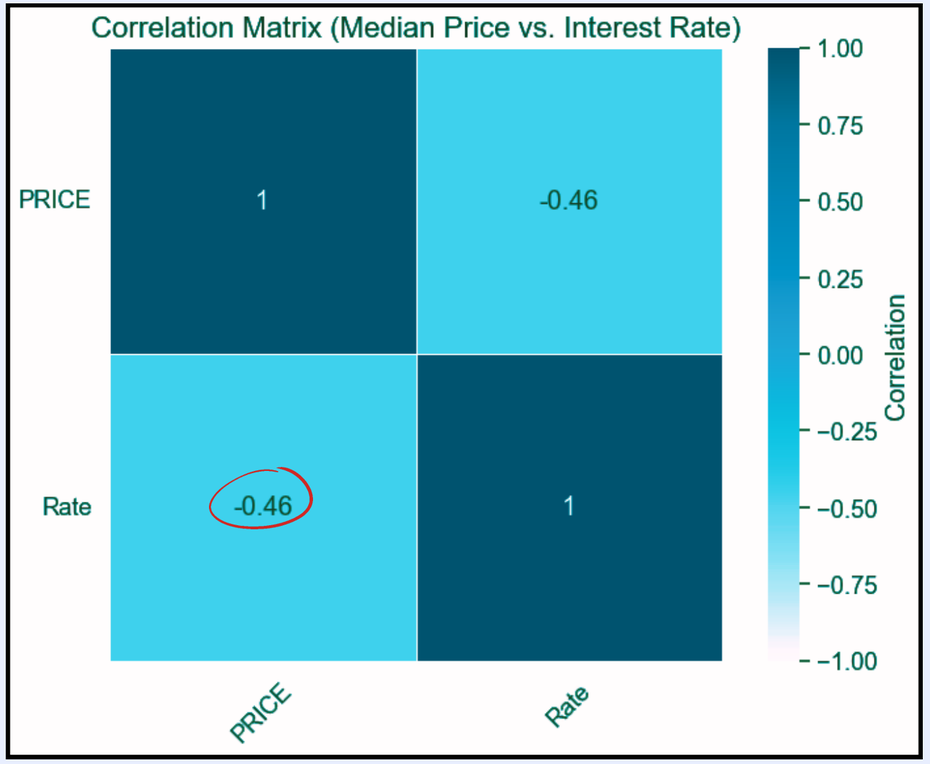
❓ For Alex & Sara- understanding this correlation is essential to consider how property prices may respond

Median Property Price vs. Interest Rate



- On the top, the dashed line represents 'Interest Rate' spanning from 2018 to 2023, highest interest rate of 6.66% in 2023
- The heatmap, shows a moderate negative correlation of -0.46 between interest rates and property prices

- During 2023 with rising interest rates, property prices may become more attractive
- In 2021 with low rates, property prices could be on an upward trajectory
- Off course, for the year 2020 where the relationship was completely impacted by covid a completely different picture can be seen



? For Alex & Sara- this correlation underscores the importance of monitoring interest rates

Pre-modeling

STEP 1:
Selection of Key Factors Related to House Price

'PROPERTY TYPE', 'CITY', 'ZIP OR POSTAL CODE'
'PRICE', 'BEDS', 'BATHS', 'SQUARE FEET',
'ELEMENTARY_SCHOOL_RATING',
'MIDDLE_SCHOOL_RATING', 'HIGH_SCHOOL_RATING',
'PRICE_PR_SQFT', 'REDFIN_ESTIMATE',
'EST_MONTHLY_PAY', 'PRICE_HISTORY1_PRICE',
'PRICE_HISTORY2_PRICE', 'PRICE_HISTORY3_PRICE'.

STEP 2:
Preprocessing of Missing Values

CATEGORICAL

- **Fill with 'Mode'**

```
categorical_cols = df.select_dtypes(include=['object']).columns
df[categorical_cols] = df[categorical_cols].fillna(df.mode().iloc[0])
```

NUMERICAL

- **Interpolation**

```
df = df.interpolate()
```

- **Fill with Mean**

```
numeric_cols = df.select_dtypes(include=['number']).columns
df[numeric_cols] = df[numeric_cols].fillna(df.mean())
```

```
NaN Counts for Each Column:
PROPERTY TYPE                0
CITY                         0
ZIP OR POSTAL CODE           0
PRICE                        0
BEDS                         0
BATHS                        0
SQUARE FEET                  0
ELEMENTARY_SCHOOL_RATING     0
MIDDLE_SCHOOL_RATING         0
HIGH_SCHOOL_RATING           0
PRICE_PR_SQFT                0
REDFIN_ESTIMATE              0
EST_MONTHLY_PAY              0
PRICE_HISTORY1_PRICE         0
PRICE_HISTORY2_PRICE         0
PRICE_HISTORY3_PRICE         0
dtype: int64
Total Number of NaN Values: 0
```

Linear Regression Model

Preparation & Splitting

```
features = df[['PROPERTY TYPE', 'CITY', 'ZIP OR POSTAL CODE', 'BEDS', 'BATHS', 'SQUARE FEET',
               'ELEMENTARY_SCHOOL_RATING', 'MIDDLE_SCHOOL_RATING', 'HIGH_SCHOOL_RATING',
               'PRICE_PR_SQFT', 'REDFIN_ESTIMATE', 'EST_MONTHLY_PAY',
               'PRICE_HISTORY1_PRICE', 'PRICE_HISTORY2_PRICE', 'PRICE_HISTORY3_PRICE']]

target = df['PRICE']
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.2, random_state=42)
```

Model Training

```
numeric_features = ['BEDS', 'BATHS', 'SQUARE FEET', 'ELEMENTARY_SCHOOL_RATING', 'MIDDLE_SCHOOL_RATING',
                    'HIGH_SCHOOL_RATING', 'PRICE_PR_SQFT', 'REDFIN_ESTIMATE', 'EST_MONTHLY_PAY',
                    'PRICE_HISTORY1_PRICE', 'PRICE_HISTORY2_PRICE', 'PRICE_HISTORY3_PRICE']
categorical_features = ['PROPERTY TYPE', 'CITY', 'ZIP OR POSTAL CODE']
numeric_transformer = StandardScaler()
categorical_transformer = Pipeline(steps=[
    ('onehot', OneHotEncoder(handle_unknown='ignore'))  # Handle unknown categories
])
preprocessor = ColumnTransformer(
    transformers=[
        ('num', numeric_transformer, numeric_features),
        ('cat', categorical_transformer, categorical_features)
    ])
pipeline = Pipeline(steps=[('preprocessor', preprocessor),
                           ('regressor', LinearRegression())])
pipeline.fit(X_train, y_train)
```

Model Evaluation

```
predictions = pipeline.predict(X_test)
mse = mean_squared_error(y_test, predictions)
r_squared = r2_score(y_test, predictions)

print('Mean Squared Error:', mse)
print('R-squared:', r_squared)
```

Mean Squared Error: 93093449491.52747

R-squared: 0.9752638953309994

Future Predictions

```
future_data_processed = pipeline.named_steps['preprocessor'].transform(future_data)
future_predictions = pipeline.named_steps['regressor'].predict(future_data_processed)
print('Future Property Price Predictions:', future_predictions)
```

Assume We have 3 properties

```
future_data = pd.DataFrame({
    'PROPERTY TYPE': ['Single Family', 'Condo', 'Multi-Family'],
    'CITY': ['Fremont', 'Hayward', 'San Francisco'],
    'ZIP OR POSTAL CODE': ['94536', '94542', '94122'],
    'BEDS': [2, 3, 4],
    'BATHS': [1, 2, 3],
    'SQUARE FEET': [3000, 4000, 2500],
    'ELEMENTARY_SCHOOL_RATING': [8, 7, 6],
    'MIDDLE_SCHOOL_RATING': [7, 6, 5],
    'HIGH_SCHOOL_RATING': [9, 8, 7],
    'PRICE_PR_SQFT': [400, 350, 300],
    'REDFIN_ESTIMATE': [600000, 400000, 550000],
    'EST_MONTHLY_PAY': [2500, 1800, 3000],
    'PRICE_HISTORY1_PRICE': [550000, 380000, 520000],
    'PRICE_HISTORY2_PRICE': [530000, 360000, 510000],
    'PRICE_HISTORY3_PRICE': [510000, 340000, 500000]
})
```

Future Property Price Predictions:
[362408.85163255
269637.12276663
458369.78452527]

Linear Regression Model:

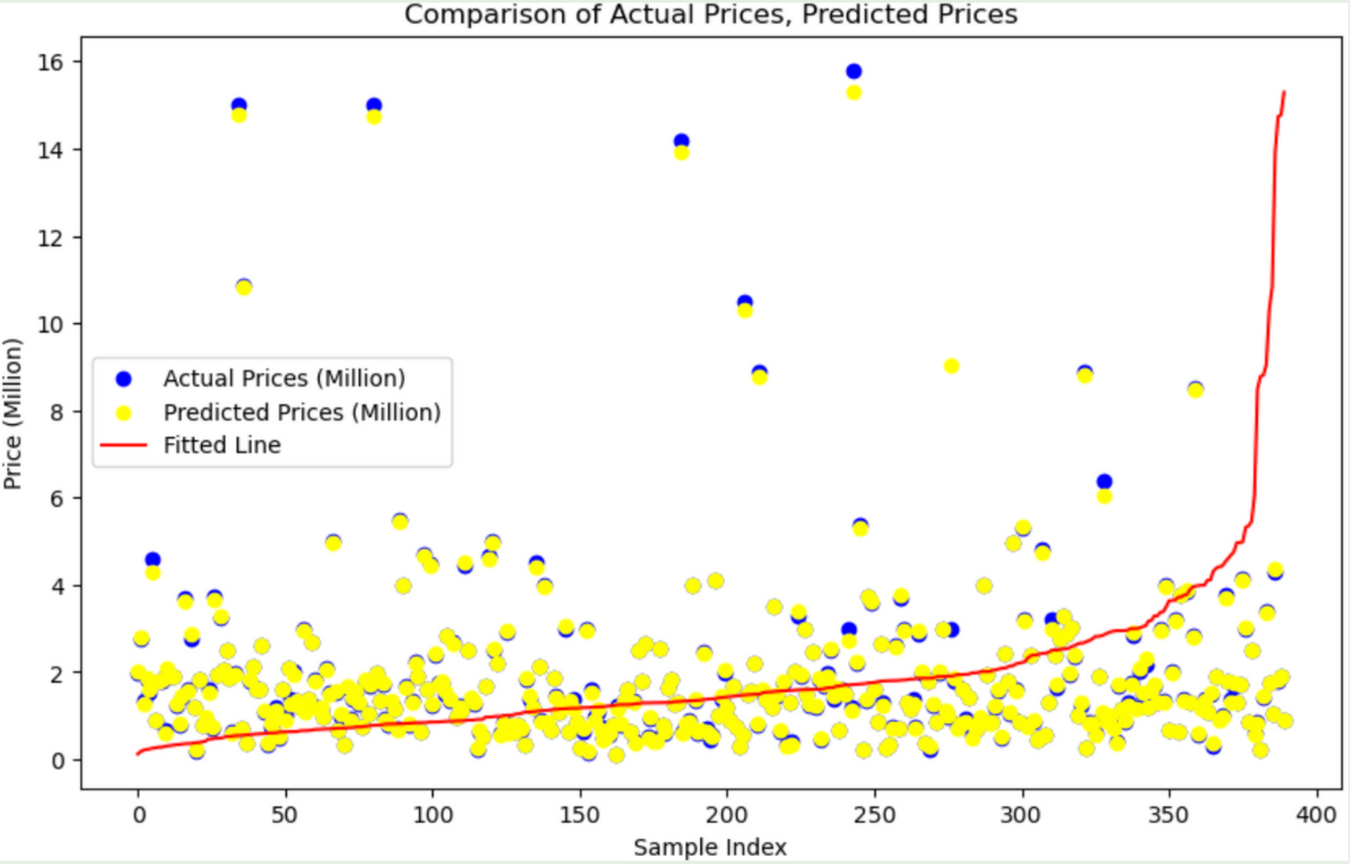
Coefficients DataFrame:

	Feature	Coefficient
0	BEDS	6474.969639
1	BATHS	10030.642208
2	SQUARE FEET	1253.267911
3	ELEMENTARY_SCHOOL_RATING	-5206.578778
4	MIDDLE_SCHOOL_RATING	-925.089921
...	...	...
266	ZIP OR POSTAL CODE_95139	10982.545780
267	ZIP OR POSTAL CODE_95140	-8865.301186
268	ZIP OR POSTAL CODE_95141	-13050.928411
269	ZIP OR POSTAL CODE_95148	-14213.614952
270	ZIP OR POSTAL CODE_95406	39541.075501

271 rows × 2 columns

Math Function:

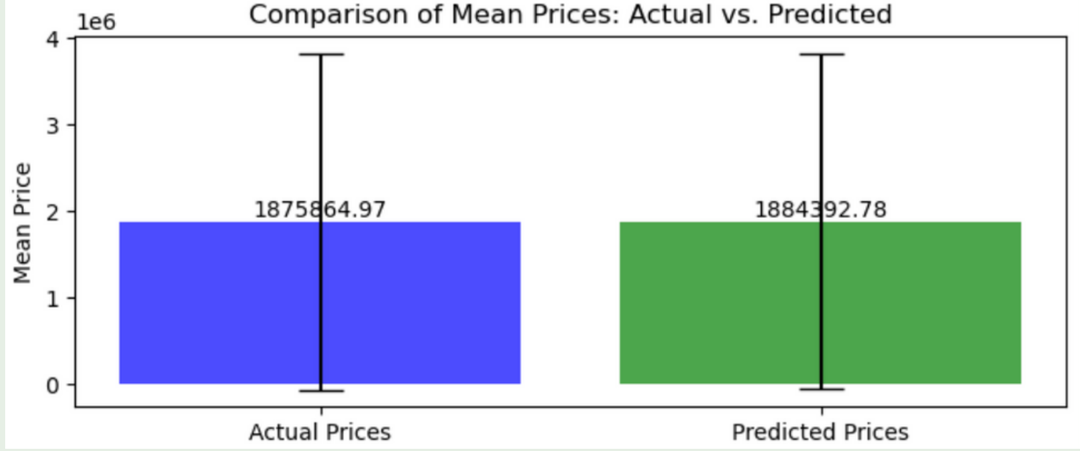
PRICE = 1851385.11 + (6474.97) * BEDS + (10030.64) * BATHS + (1253.27) * SQUARE FEET + (-5206.58) * ELEMENTARY_SCHOOL_RATING + (-925.09) * MIDDLE_SCHOOL_RATING + (2198.19) * HIGH_SCHOOL_RATING + (1968.22) * PRICE_PR_SQFT + ... + (1683843.21) * EST_MONTHLY_PAY + ... + (-114774.20) * PROPERTY TYPE_Condo/Co-op + (2676.69) * PROPERTY TYPE_Mobile/Manufactured Home + (-25659.33) * PROPERTY TYPE_Multi-Family (2-4 Unit) + (106891.35) * PROPERTY TYPE_Multi-Family (5+ Unit) + (70570.80) * PROPERTY TYPE_Other + (-1359.73) * PROPERTY TYPE_Single Family Residential + (-54168.97) * PROPERTY TYPE_Townhouse + (15823.39) * PROPERTY TYPE_Vacant Land + (1315.16) * CITY_Atherton ...



Distribution of actual prices, predicted prices, and future predictions:

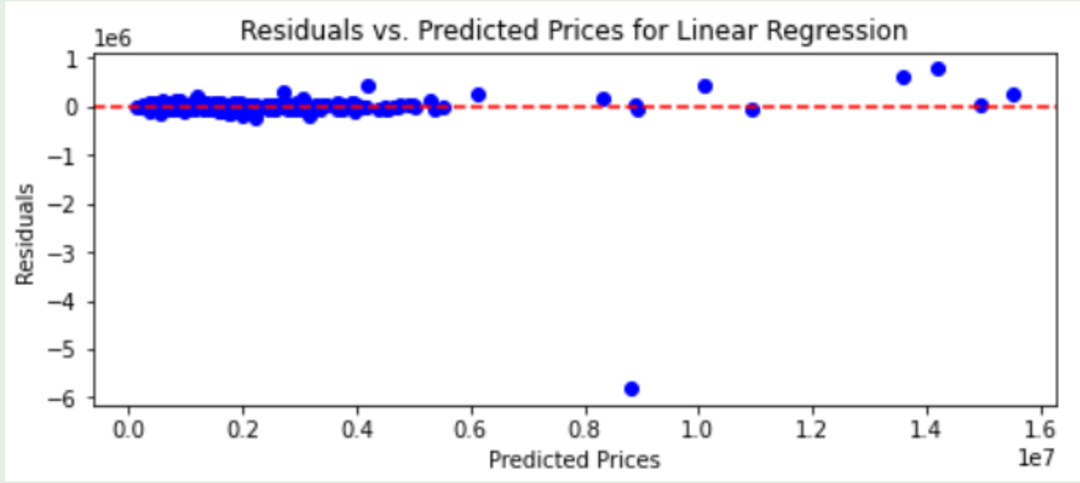
· scatter plot to show the distribution of these prices

· red fitted line to double check the overall distribution of the predictions: we have around 400 records in our test set, and most of them are between 0-3 million in price.



Comparison of Mean Prices

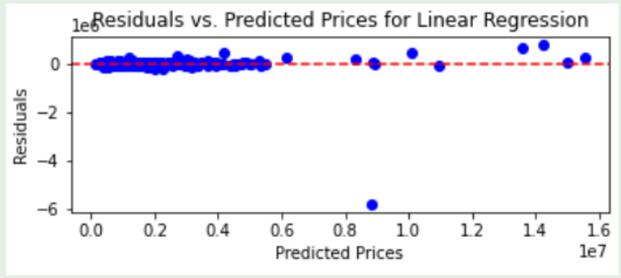
· bar graph (error bars reflect the uncertainty in these mean prices)

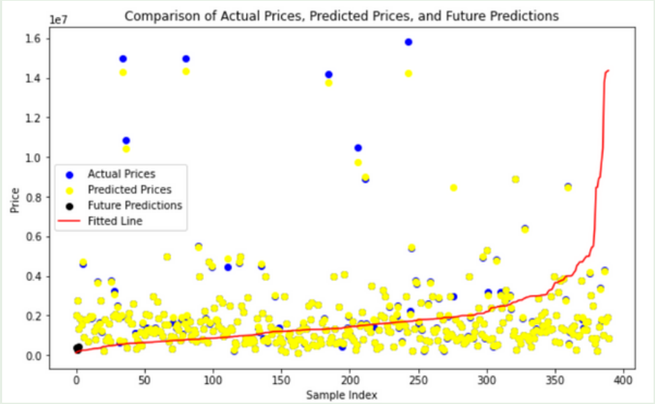


Residuals and Predicted Prices Relationship:

· scatter plot - dashed red line = the differences between actual and predicted prices

Random Forest Model





Preparation & Splitting...

Model Training

```
pipeline = Pipeline(steps=[('preprocessor', preprocessor),
                           ('regressor', RandomForestRegressor(n_estimators=100, random_state=42))])
```

Model Evaluation

Mean Squared Error: 89735009755.08913
R-squared: 0.97615627516329 51

Future Predictions

Future Property Price Predictions:
[360614.44
293895.98
436461.64]

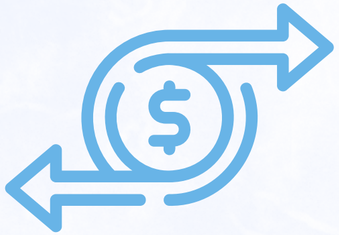
Mean Squared Error: 93093449491.52747
R-squared: 0.97526389533099 94

Future Property Price Predictions:
[362408.85163255
269637.12276663
458369.78452527]

Linear Regression Model

Investor Perspective

Investment Purpose



Expected Cashflow
&
Profit Opportunities

New Construction
Vs
Existing Property

Valuation of
the
Property

Leverage

Overall Real Estate Market

Property Location

Your Credit Score

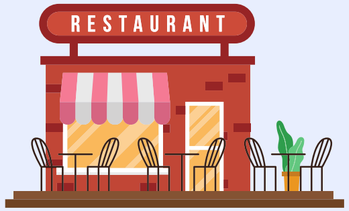
Source: Investopedia

Learning + Outcomes for Alex & Sara

Based on the initial analysis, it is evident that Santa Clara, Contra Costa and Alameda counties boast the highest number of affordably priced houses

Single-family and condo home types appear to have both a higher count and reasonable prices in the aforementioned regions

An attractive price range for a quality property falls between 1.55M - 1.83M, considering the convenient accessibility to groceries, parks, and restaurants



Property price drop post covid with high interest rates and rising inflation investment decision will depend on market conditions, cash flow goals and purchase power


THANK YOU
Q&A