Pricing Analysis of Northern California Housing

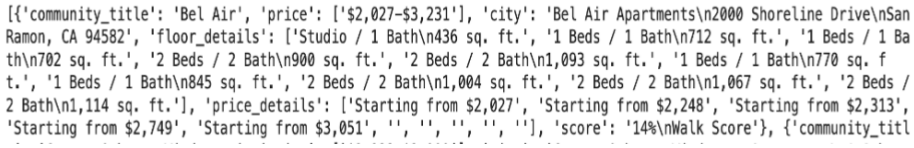
Problem Statement:
The project aims to scrape inclusive and up-to-date data on property listings, including pricing details, to gain insights into the current state of the housing market in Northern California by Essex**.**

Data Source: The data is collected through web scraping techniques, targeting Essex website.

Key Attributes: Property Price, City name, No of rooms, No of baths, Square ft details, Accessibility score.

Tools & Libraries: Python (Scripting), Selenium (Web scraping), CSV File (Data storage), Pandas (Data exploration and analysis), Seaborn, Matplot lib (Data Visualizations), Scikit learn (ML Models) & Jupyter Notebook (IDE).

Web Scraping: The data was extracted through a web scraping technique called selenium. This involves importing required libraries, navigating to the website, interacting with the elements to get the required details like Price (Dependent variable), geographic details, floor details(Independent variables). The extracted data is as shown below:

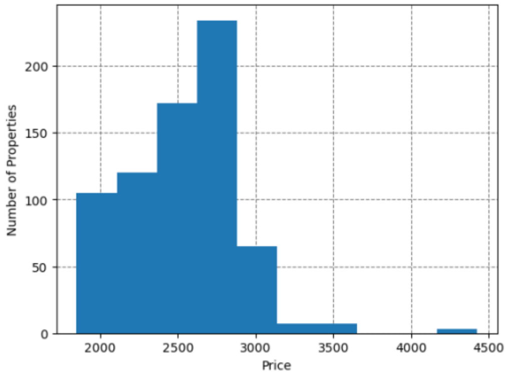


```
[{'community_title': 'Bel Air', 'price': ['$2,027-$3,231'], 'city': 'Bel Air Apartments\n2000 Shoreline Drive\nSan
Ramon, CA 94582', 'floor_details': ['Studio / 1 Bath\n436 sq. ft.', '1 Beds / 1 Bath\n712 sq. ft.', '1 Beds / 1 Ba
th\n702 sq. ft.', '2 Beds / 2 Bath\n900 sq. ft.', '2 Beds / 2 Bath\n1,093 sq. ft.', '1 Beds / 1 Bath\n770 sq. f
t.', '1 Beds / 1 Bath\n845 sq. ft.', '2 Beds / 2 Bath\n1,004 sq. ft.', '2 Beds / 2 Bath\n1,067 sq. ft.', '2 Beds /
2 Bath\n1,114 sq. ft.'], 'price_details': ['Starting from $2,027', 'Starting from $2,248', 'Starting from $2,313',
'Starting from $2,749', 'Starting from $3,051', '', '', '', '', ''], 'score': '14%\nWalk Score'}, {'community_titl
```

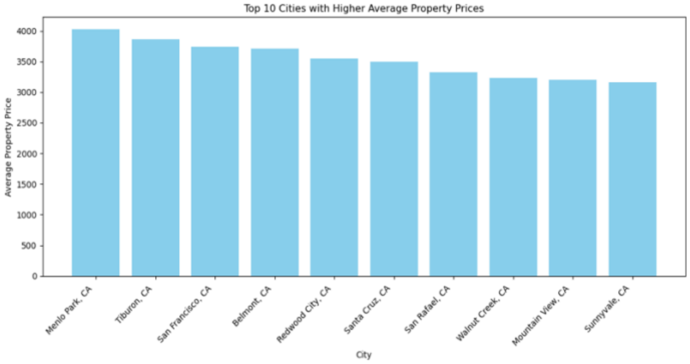
The data of each variable was stored in a dictionary, and then into a list. Then list has been exported to a CSV file for further analysis.

Exploratory Data Analysis:

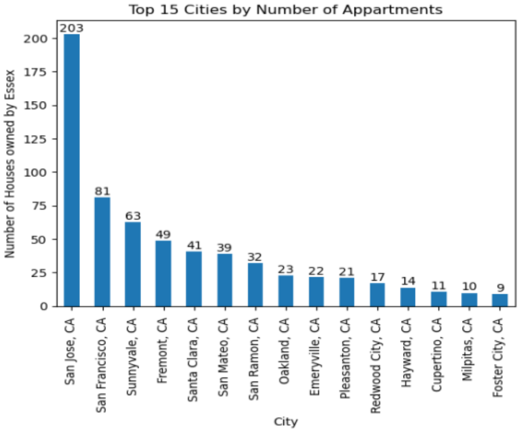
1. *Data Exploration:* Examined the data types, null values, duplicates, and missing values of the dataset.
2. *Data Transformation:*
   - Converted the floor_details column into NRooms, NBath, sqft.
   - Price into minPrice and maxPrice columns.
   - In score column, extracted the integer only and renamed it to Accesibility score.
   - City column contains complete address, hence extracted the city name only.
3. *Handling Null Values:* Checked for null values, particularly in the ‘ Property Price' column. Instead of dropping those records, those null values are replaced with the average rent of the respective city in which the property is located.
4. *Data Quality Check:* Verified data completeness and accuracy after transformations. Ensured that the dataset is now suitable for in-depth analysis.
5. *Data Analysis:* We can derive the following insights
   - The avg price increases in primary locations because of proximity to economic centers like Silicon Valley, IT Hub, good schools, attractive amenities, low crime rates etc.
   - From the below graphs, we can say that the avg price of the listed properties in Essex is from 2500 to 3100 USD.
   - The Essex apartments have 203 apartments in San Jose, mainly due to their proximity to places like university, other amenities etc.



Distribution of prices across properties



Highest avg prices by city

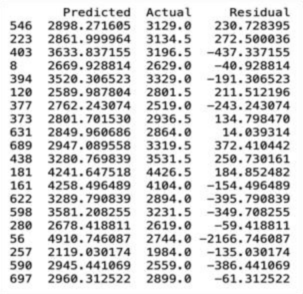


Distribution of Houses across City

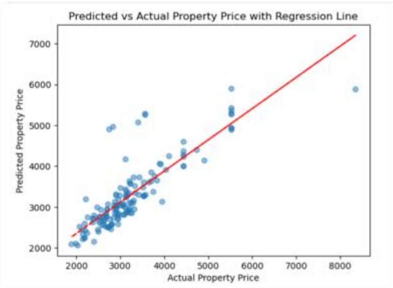
Machine Learning Models:

1. *Multiple Linear Regression:*
    - Trained and Tested the Multiple Linear Regression Model.
    - Splitted the data into training and testing sets, 80% of the data for training and 20% of the data for testing.
    - Considered multiple independent variables like sqft, NRooms, NBath, City, Accessibility score as predictors.
    - Target variable as Property Price.
    - Used R-squared and Adjusted R-square as model evaluation metrics.

    R-squared: 0.7032342500860779



```
          Predicted  Actual     Residual
546  2898.271605  3129.0   230.728395
223  2861.999964  3134.5   272.500036
403  3633.837155  3196.5  -437.337155
8    2669.928814  2629.0   -40.928814
394  3520.306523  3329.0  -191.306523
120  2589.987804  2801.5   211.512196
377  2762.243074  2519.0  -243.243074
373  2801.701530  2936.5   134.798470
631  2849.960686  2864.0    14.039314
689  2947.089558  3319.5   372.410442
438  3280.769839  3531.5   250.730161
181  4241.647518  4426.5   184.852482
161  4258.496489  4104.0  -154.496489
622  3289.790839  2894.0  -395.790839
598  3581.208255  3231.5  -349.708255
280  2678.418811  2619.0   -59.418811
56   4910.746087  2744.0 -2166.746087
257  2119.030174  1984.0  -135.030174
590  2945.441069  2559.0  -386.441069
697  2960.312522  2899.0   -61.312522
```



2. *Random Forest Regression:*
   - Like Multiple Linear Regression, we considered multiple independent variables like sqft, NRooms, NBath, City, Accessibility score as predictors for Random Forest Model.
   - Target variable as Property Price. Random Forest is defined with number of estimators = 100.
   - There is only very slight increase in the performance of the model compared to MLR

     R-squared (R2): 0.7139185026779504

3. *Decision Tree Regressor:*
   - Considered multiple independent variables like sqft, NRooms, NBath, City, Accessibility score as predictors. Target variable as Property Price.

     R-squared: 0.609038885620987

Overview of Exploratory Data Analysis:
Average price increases in primary locations due to proximity to economic centers, good schools, etc.
Average property price range in Essex is $2,500 to $3,100.

Machine Learning Model Performances:
Linear Regression: R² of 70%.
Random Forest Regression: R² of 71%.

Conclusion:
Random Forest Regression is most effective for predicting housing prices. Through this project, we are predicting the key attributes that are affecting the housing prices in Northern California.

This pricing analysis provides information on the economic factors which effects the property values, current market patterns and community development. This project aims to provide insights to stakeholders like investors, buyers, and sellers to make data driven decisions.