Real Estate Analysis for BAY AREA

GROUP MEMBERS

DEEPANSH MALVIYA

JASVITHA BUGGANA

KRUTIKA DESHPANDE

Contents

Introduction

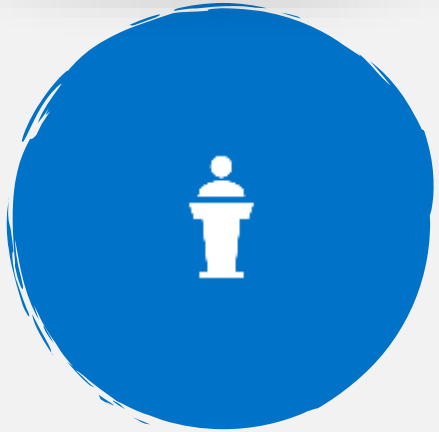
- The San Francisco Bay Area's real estate market has been a topic of intrigue and concern for both residents and investors alike.
- Over the past decade, the region has witnessed astonishing fluctuations in property values, from the meteoric rise pre-pandemic to recent shifts in response to the global economic landscape.
- Our goal is to provide you with a comprehensive analysis of the San Francisco Bay Area's real estate market, enabling you to make informed decisions, whether you're a potential buyer, seller, or investor.
- We'll dive into data-driven insights, leveraging advanced analytics to uncover the market's true dynamics.

Contoso

Web Scraping Process

Main URL

https://www.redfin.com/



Objective

To extract data from multiple real estate listings posted on the Redfin website in popular California cities.

Libraries

Beautiful Soup, Requests, Pandas, NumPy, Matplotlib, seaborn

- The code extracts a list of URLs from a main URL using the requests and beautifulsoup4 libraries in Python.
- We manually extracted the URL's for the 9 counties
- We wrote a code to extract the first 4 pages of each county which gave us information of 36 pages.
- The property extensions are concatenated with https://www.redfin.com/ to get the complete web address of each property.
- The code then visits each property URL and extracts the desired information using span, div tags along with their respective attributes.
- To handle any missing attributes or errors, a try-except block is used in the code.
- The output of the code is a list of dictionaries that contains the extracted information for each property.
- Finally, the code saves the output as a CSV file for easy analysis and sharing.

Contoso

Data Cleaning and wrangling

- **Data Anomalies observed***:* There were columns containing ‘MISSING’ values & Some data points contained special characters like $, % etc.

- **Handling Missing Data**: To deal with missing data in columns like ‘beds’ ‘Bath’, ‘SQFT’, ‘School_Rating’, ‘Bike_score’, ‘TransitScore’, we have imputed missing values with appropriate mean values. We also had a property with 56 bedrooms which had to be dropped as it a clear outlier.

- The missing values in county were imputed using ffill.

- While extracting data we have extracted columns like state, street address, but they were not needed for our analysis, So we have dropped them.

- SQFT & Price we have had a lot of outliers which we found through histogram, So we had to remove them.

- We have replaced the unusually high values in the "Bike_Score," "WalkScore”, “Beds”, “Baths”, “SQFT” and "TransitScore" columns with the respective means of those columns. This data cleaning step can help ensure that the data is more reasonable and representative for analysis.

- *Data Type Conversion*: Convert columns like Price, WalkScore, Bike_Score, and TransitScore to their appropriate data types for our analysis.

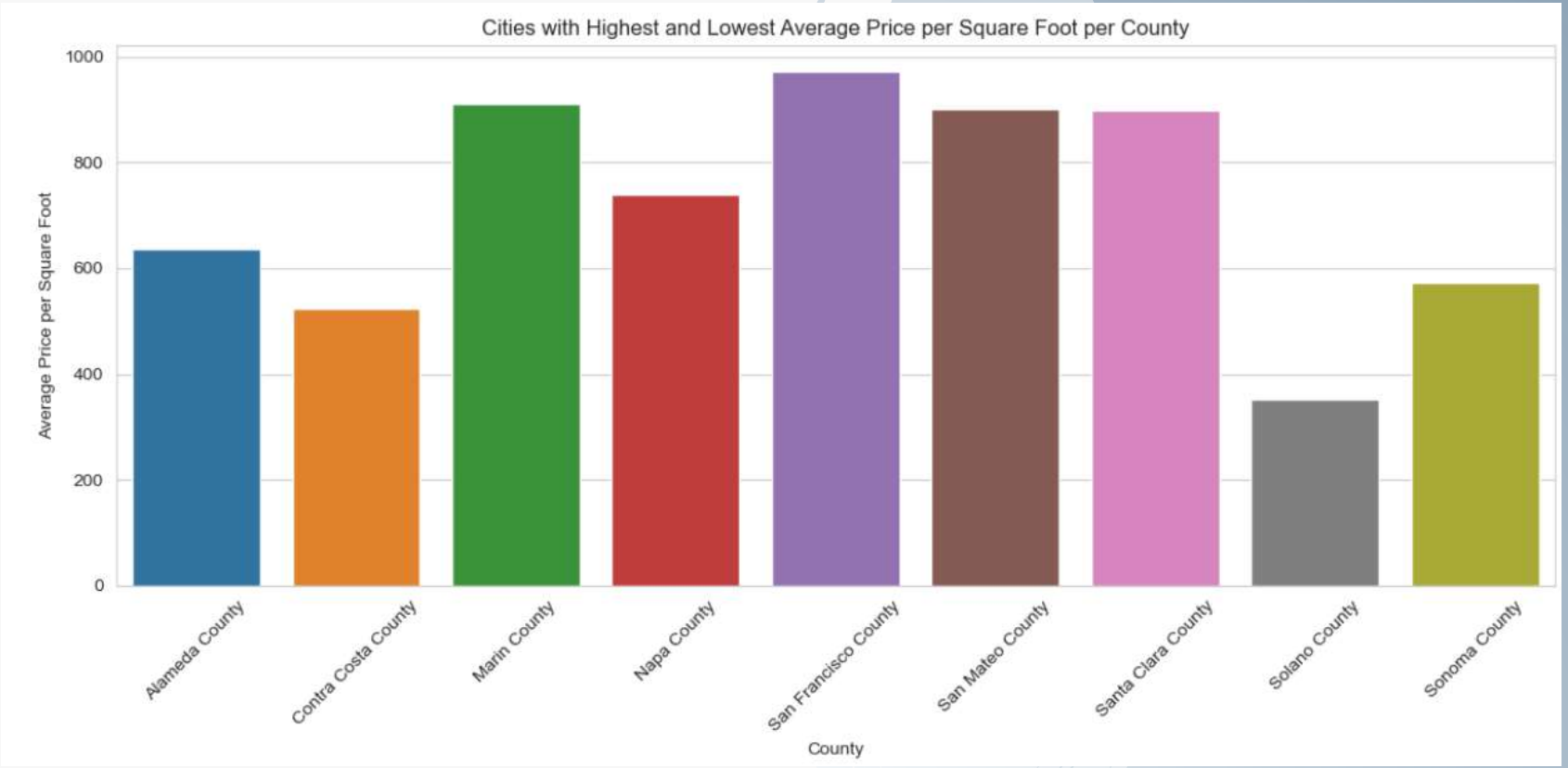
Contoso

Dataset Description

The data extracted consists of **1429 rows and 12 columns**	
Zip Code	Location of the property
City	City of the property
Garage Availability	Garage available with the property
Transit Score	How well location is served by public transit
List Price	Listed price/selling price/market vale for the property
Bedrooms	Number of rooms in the property
Bathrooms	Number of baths in the property
Bike Score	Measures how bike friendly neighborhood is
Walk Score	Based on walking distance from that address to various amenities
School Rating	Rating of school for a particular city
SQFT	Gives you the sqft of the property
County	Gives us information about where is the property located

What are the cities with the highest and lowest average price per square feet per county & outliners?

- In this question calculates the average price per square foot for different counties in the dataset. We identify the counties with the highest and lowest average price per square foot and visualizes the results using a bar graph



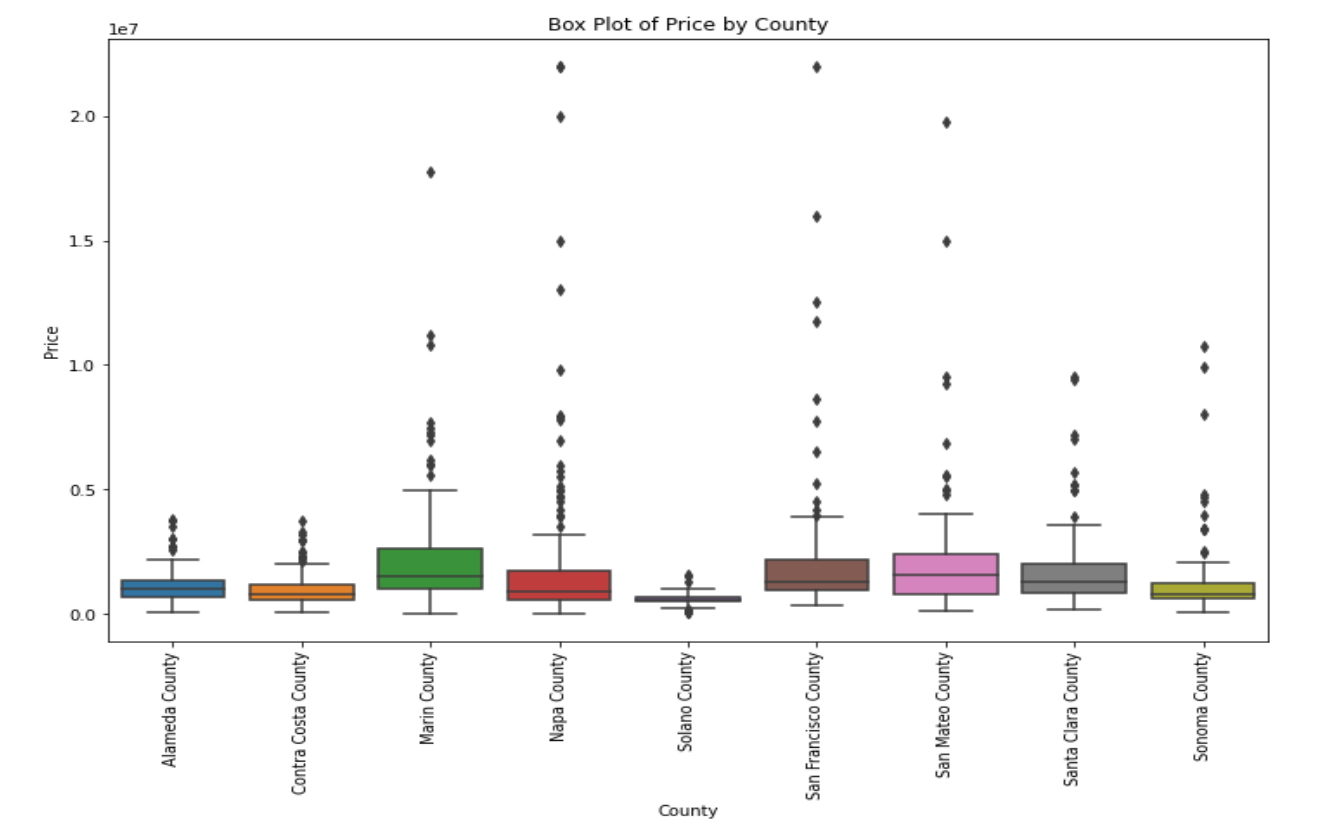
Observations:

- *San Francisco county has the highest per sqft price in the bay area which is followed closely by Santa Clara and San Mateo Counties.*
- *Solano county has the lowest per sqft price in the bay area which is less than half the price per sqft of Santa Clara or San-Mateo counties.*
- *Average Price per Sqft across all counties seem to be more than 600$*

Contoso

What are the cities with the highest and lowest average price per square feet for county & outliners?

- Using box plots to find outliers in data is a valuable technique for identifying extreme values or anomalies. We have created a box plot to visualize the distribution of property prices by county. Outliers, which are data points significantly different from the majority, can be spotted in this visualization.
- Boxes display the distribution of prices within each county. Outliers are displayed as individual data points beyond the whiskers of the boxes.



Observations:

- The median line inside the box represents the median price within each county.
- The height of the box indicates the interquartile range (IQR).
- The longer the box, the greater the spread of prices.
- From the picture we can state that the marin county, napa, sf and san mateo has many outliers and

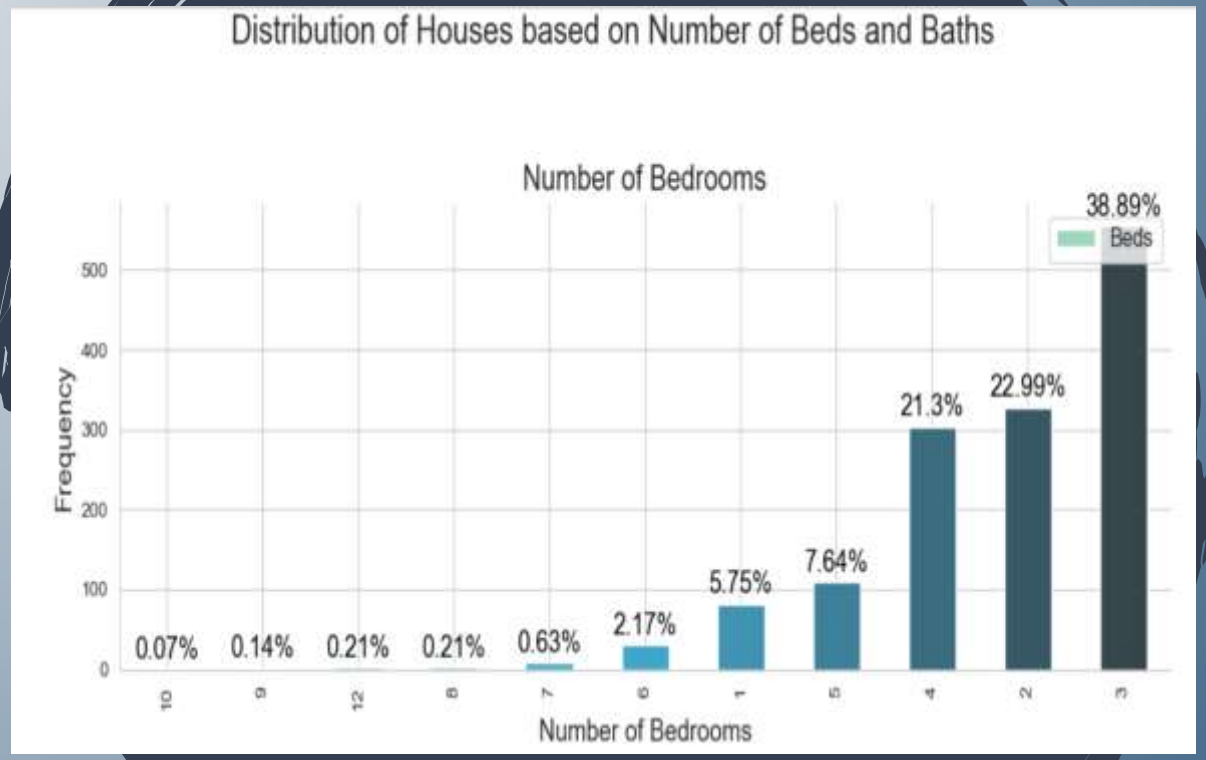
Does the presence of a garage significantly affect property prices?

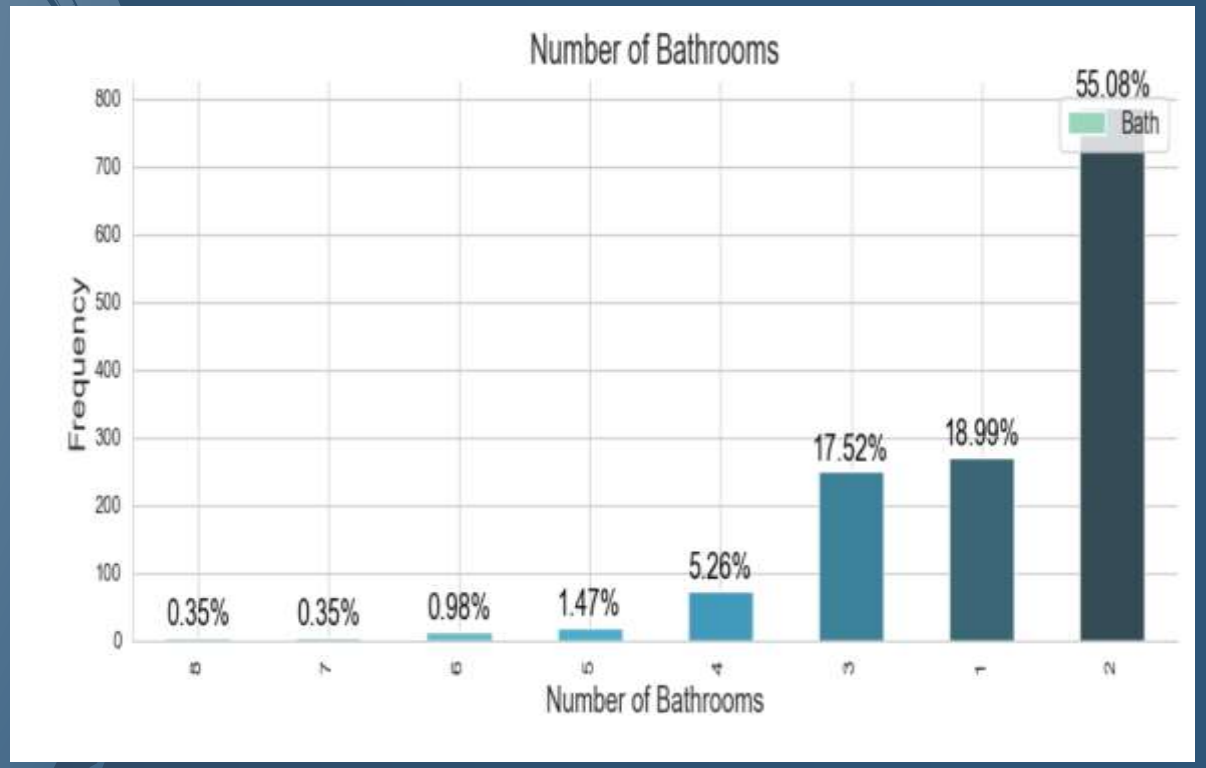
- **Statistical Test:** A two-sample t-test (independent t-test) is performed .This test compares the property prices between the two groups (with and without garages) to determine if there's a statistically significant difference in means. The results include the t-statistic and the p-value.

- **Price Difference Calculation:** Calculating the difference in mean property prices between the two groups.

- In our analysis, we examined the impact of garage availability on property prices. We divided the dataset into two categories: properties with garages and properties without garages. A two-sample t-test was performed, with a significance level (alpha) of 0.05.

Distribution of houses based on number of bedrooms and bathrooms





The bar plot depicts percentage the number of houses with different number of bedrooms and bathrooms. Houses with 3 bedrooms and Houses with 2 bathrooms are common among the listings.

Value Counts of Beds Grouped by County:

Beds	1	2	3	4	5	6	7	8	9	10	12	56
County												
Alameda County	16	27	66	36	11	4	0	0	0	0	1	0
Contra Costa County	4	32	64	35	20	1	0	0	1	0	2	0
Marin County	6	31	57	44	14	7	0	1	0	0	0	0
Napa County	0	34	77	25	12	5	4	0	0	0	0	0
San Francisco County	21	49	50	25	9	4	2	0	0	0	0	0
San Mateo County	18	53	32	30	7	5	1	1	1	0	0	1
Santa Clara County	9	33	68	32	14	1	1	0	0	1	0	0
Solano County	2	28	67	51	14	1	0	0	0	0	0	0
Sonoma County	6	41	74	26	8	3	1	1	0	0	0	0

Maximum Values by County:

County	
Alameda County	66
Contra Costa County	64
Marin County	57
Napa County	77
San Francisco County	50
San Mateo County	53
Santa Clara County	68
Solano County	67
Sonoma County	74

dtype: int64

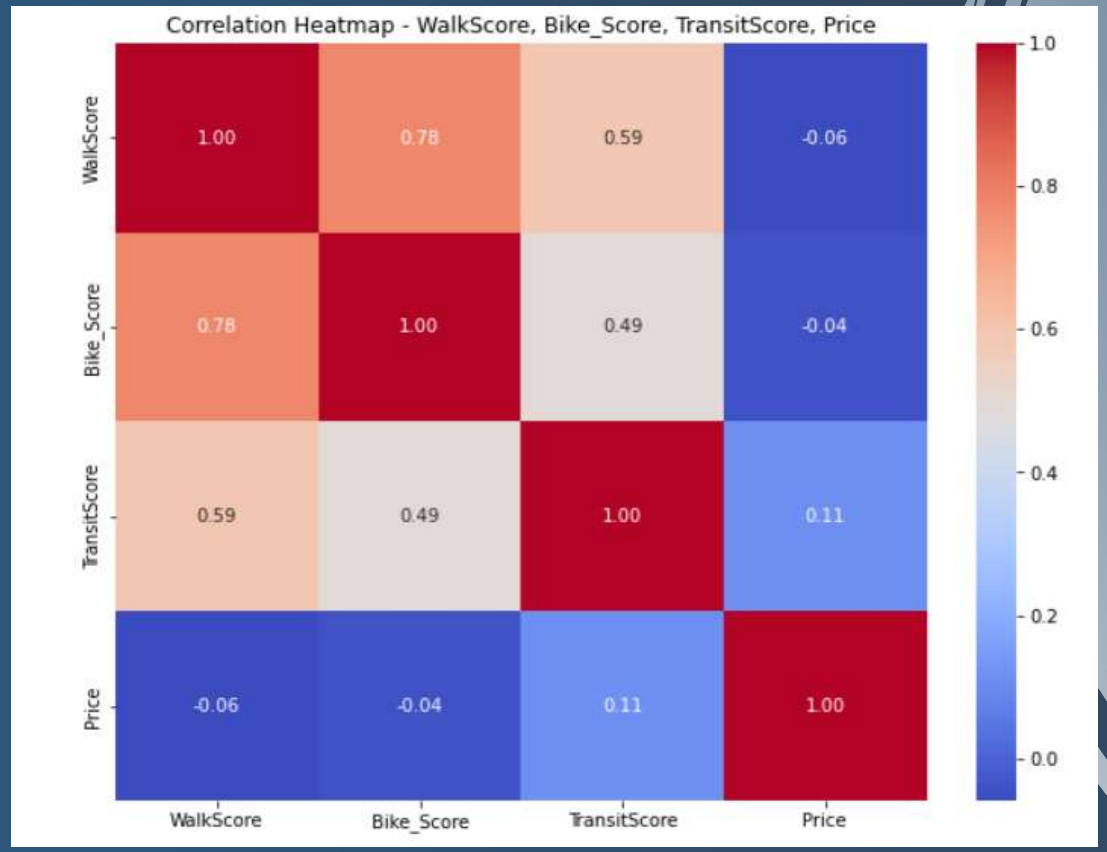
- We can group houses in counties by number of bedrooms
- We can notice that 3 bedrooms houses are high in all the counties.
- Napa county has the highest number of houses with 3 bedrooms

Column Name with Maximum Count by County:

County	
Alameda County	3
Contra Costa County	3
Marin County	3
Napa County	3
San Francisco County	3
San Mateo County	2
Santa Clara County	3
Solano County	3
Sonoma County	3

dtype: int64

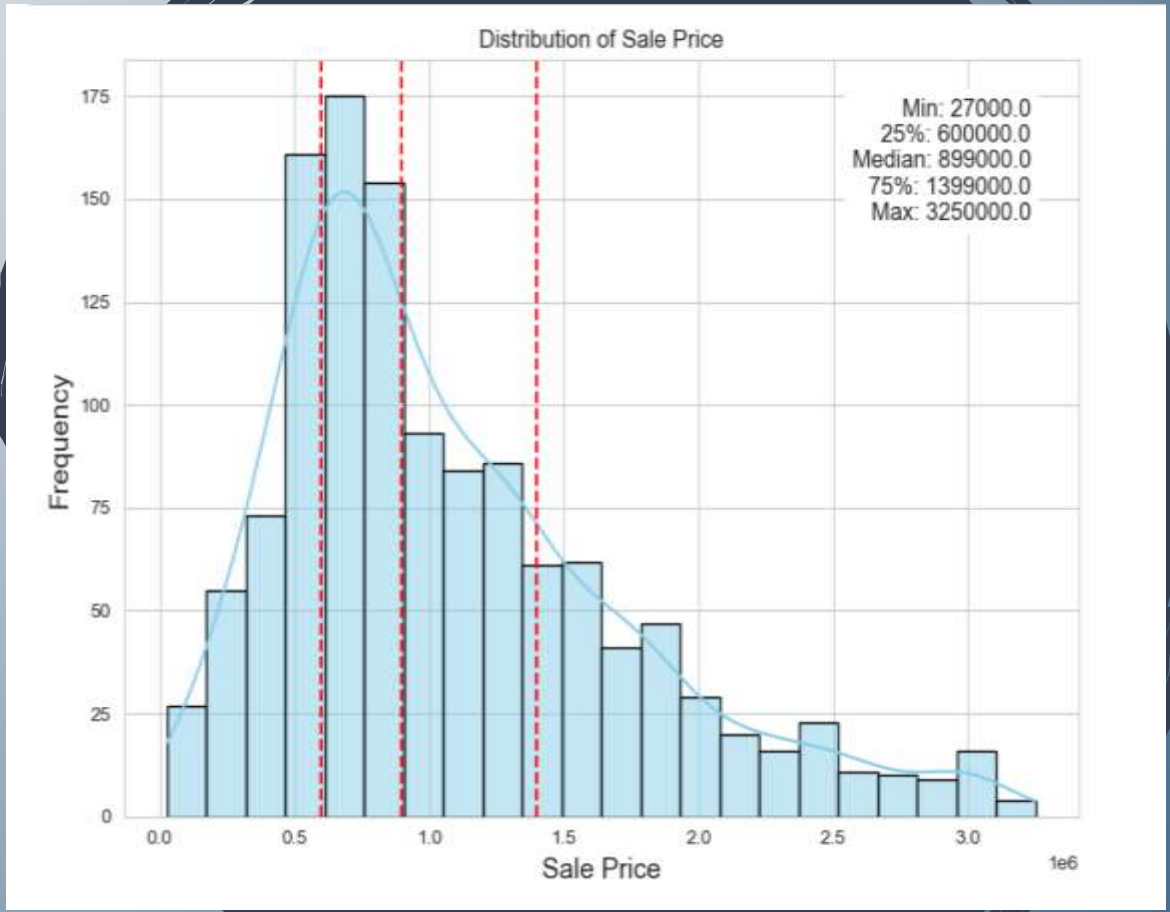
EFFECT OF WALK SCORE, BIKE SCORE, AND TRANSIT SCORE ON HOUSE PRICES



- To analyze the effect of all the scores ('walkscore,' 'bike_score,' 'transitscore') on the 'price' column, we have performed a multiple linear regression analysis. This analysis helped us to determine how well the scores collectively impact the 'price' column.
- The r-squared value is 0.032, indicating that only approximately 3.2% of the variability in 'price' can be explained by the three independent variables. This suggests that 'walkscore,' 'bike_score,' and 'transitscore' have a limited ability to impact property prices in the dataset.

- The regression analysis suggests that 'transitscore' have statistically significant effects on property prices. However, 'bike_score' & Walkscore does not appear to have a significant impact.
- The model, while statistically significant, explains only a small portion of the variation in property prices. This implies that other unaccounted factors likely play a more substantial role in determining property prices in this dataset.

DISTRIBUTION OF PRICE

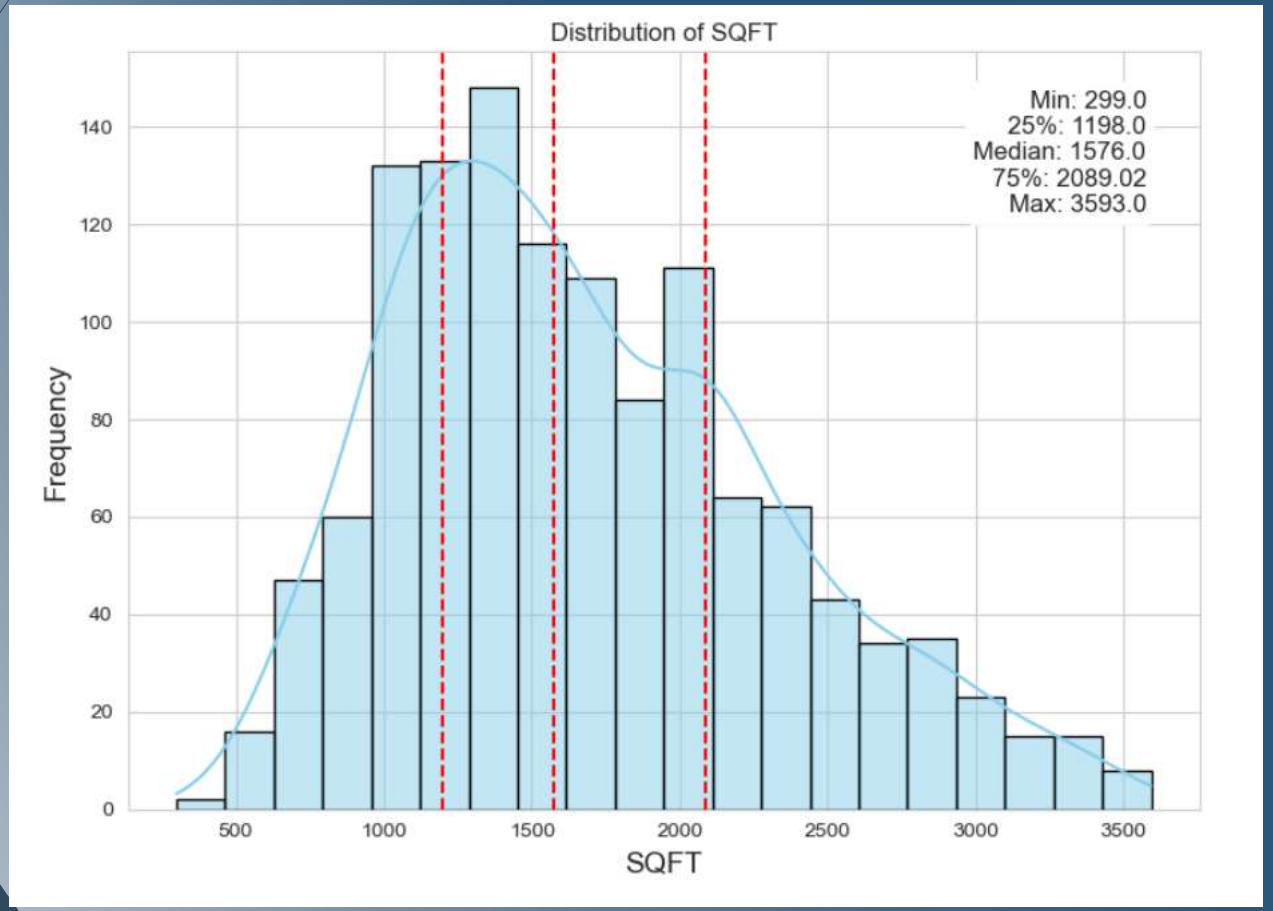


- The distribution of sale prices appears to be right-skewed, with a longer tail on the right side. This skewness suggests that while many properties have sale prices clustered around the median, a smaller number of properties have substantially higher sale prices.

DISTRIBUTION OF SQFT

- This histogram and kernel density estimate visually represent the distribution of SQFT.

- The distribution of SQFT appears to be unimodal and somewhat symmetric, with a peak near the median. This suggests that many properties have SQFT values clustered around the central value, resulting in a bell-shaped distribution.



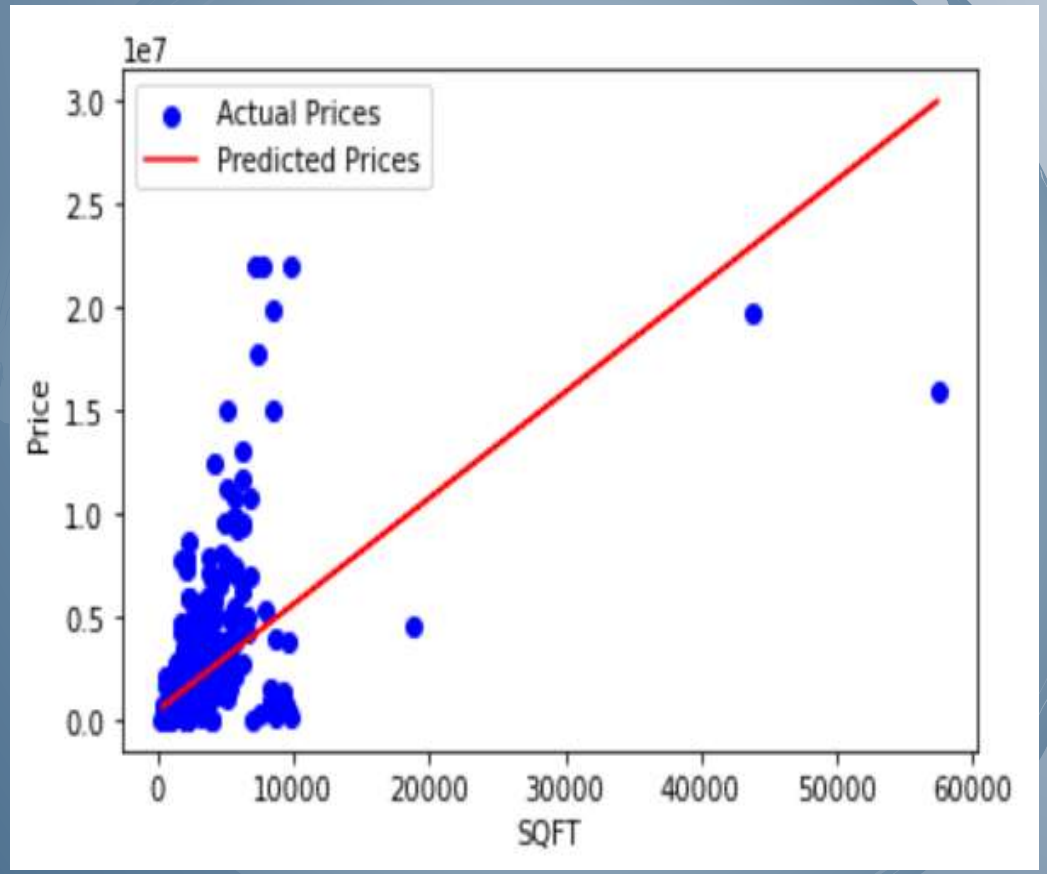
How does the size of property effect on price?

- In general , if the size increases the price of property increases, but majority of population might have budget constraints.

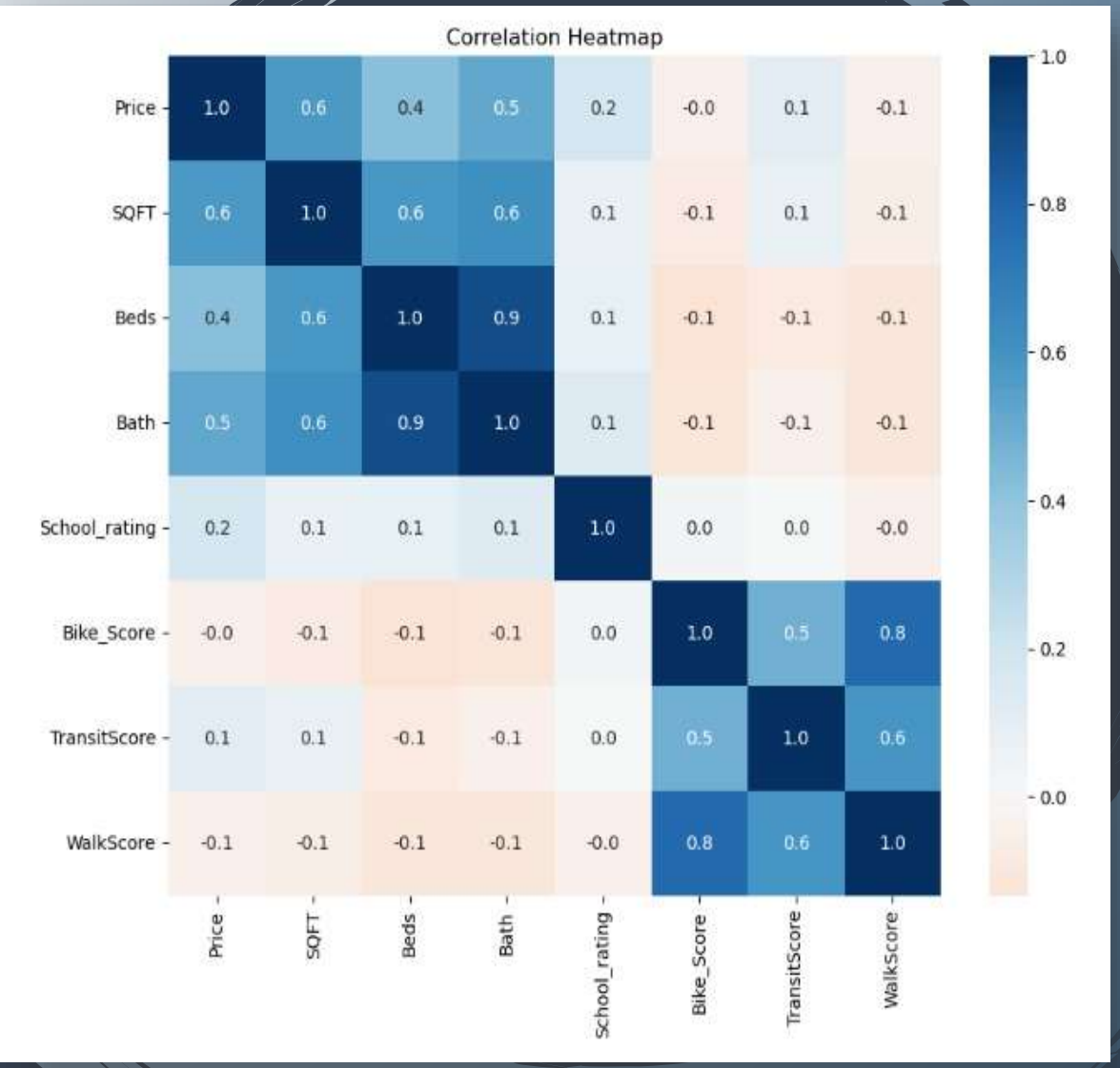
- Majority houses bought by the people are below 10000sft and the price for them is below 1million.

Contoso

- By performing the hypothesis on p-val<0.5. So the hypothesis is failed to reject. concludes significant difference b/w price and SQFT
- The correlation coeff=0.5753, concludes that there is a positive linear relation ship between price and size(SQFT)
- R square= 0.33, represents 33% sure that there is significance of Size on price.



CONCLUSION



- San Francisco county is the most expensive in term of price per SQFT
- Variables affecting Property Price:
  - Property Size (SQFT) – most significant
  - Garage availability – significantly
  - TransitScore - slightly
- Majorly, there are properties with 3 bedrooms or 2 bathrooms
- Napa County has the highest number of such properties

THANK YOU