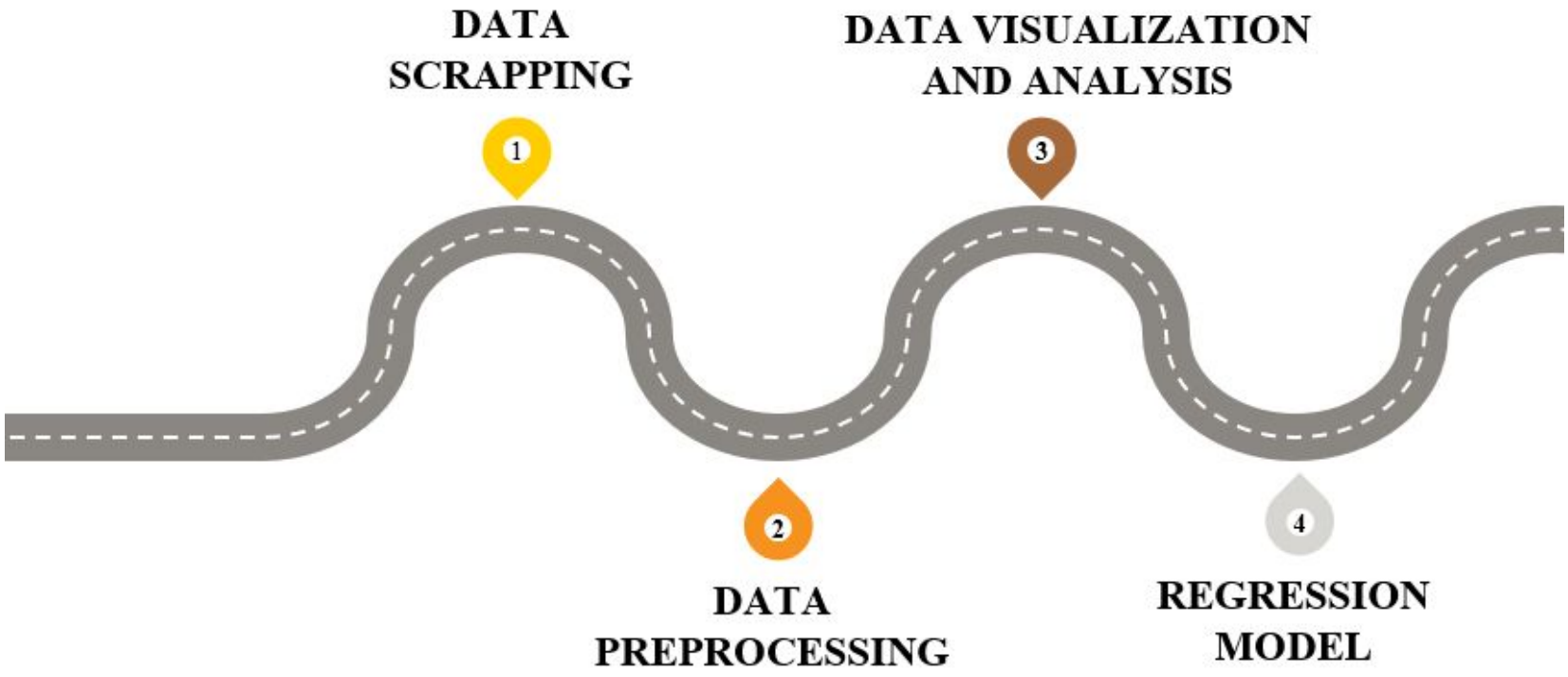
Real Estate Market Analysis in the Bay Area

BAN 612 Project

Instructor: Dr. Lan Wang

Agonda


DATA SCRAPPING
1
DATA VISUALIZATION AND ANALYSIS
3
2
DATA PREPROCESSING
4
REGRESSION MODEL

Objectives

Every individual need to deal with the real estate or housing market at certain point of time in life. Having a good overview on the market will help in buying or selling the house in the market at right price.

In our project we will be doing exploratory data analysis and prediction on housing prices at various cities in the Bay Area such as Fremont, South San Francisco, San Mateo, Burlingame, Palo Alto, Santa Clara, Cupertino, Milpitas etc. that are listed on Zillow.

DATA SCRAPING

WEB SCRAPING PROCESS – DATA COLLECTION

Zillow | AreaVibes | Modules - Urllib, Beautiful Soup, requests

Getting house listings and links

Getting the houses listed for each area (first 2 pages)

Scrapping Individual House data from listings

Scrapping details like facts & features, schools, price history using beautiful soup

Creating Data Frames

Creating data frames
Zillow – Basic Info of the house
Facts – Facts & Features
Price History – Price history of each house

Ratings

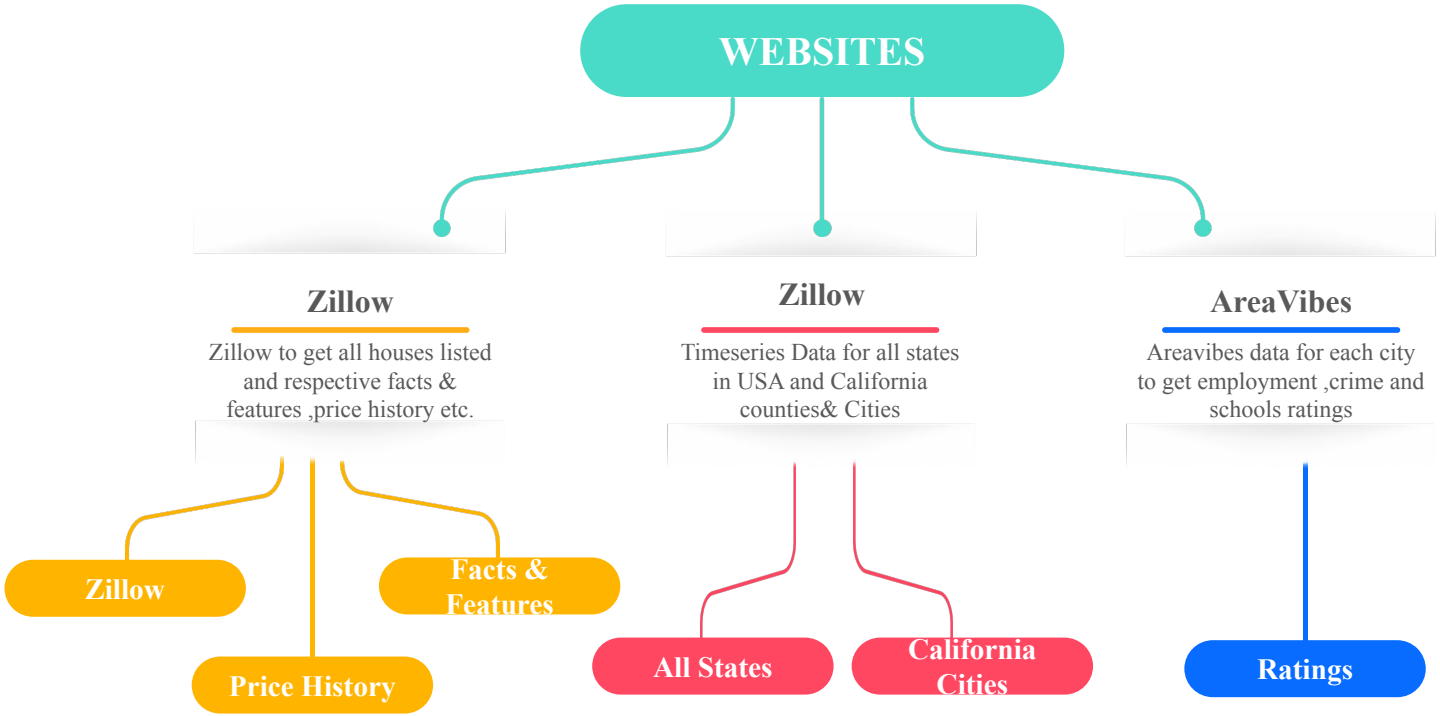
Getting ratings for schools, crimes, employment and livability from Area vibes

Storing Data in Data Frames

All scrapped data is converted to data frames and stores as CSV files

DATA COLLECTION

Data Is collected from various websites by scrapping & downloads



Data Preprocessing

DATA PREPROCESSING

1 Dropping unwanted Columns

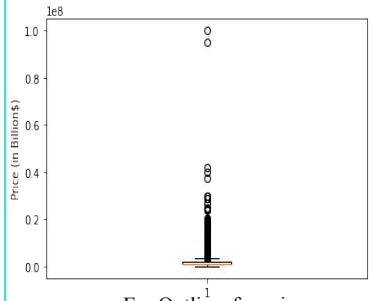
Purpose:
- Dropping columns that are not required for analysis

Applied on:
- Zillow
- Facts
- Price History

2 Removing Outliers

Purpose:
- Outlier will give biased results during analysis and model building.

Applied on:
- Zillow
- Facts
- Price History



Ex: Outliers for price

3 Removing House Types

Purpose:
- In our case we do not need LOT type of houses for our analysis. So we have dropped rows with house type as LOT.

Applied on:
- Zillow

```
SINGLE_FAMILY    1355
CONDO             603
TOWNHOUSE         275
LOT               165
MULTI_FAMILY      134
MANUFACTURED       61
Name: homeType, dtype: int64
```

Removed house with type as lot

4 Data Format and Conversions

Purpose:
- Making sure all values in single column has same datatype and correct format. Ex: Lot Area is acres in acres and sqft.

Applied on:
- Zillow ,Facts

	lotAreaValue	lotAreaUnit
0	2.32461	acres
1	7500.00000	sqft
2	10400.00000	sqft

⬇

	lotAreaValue	lotAreaUnit
0	101260.000000	sqft
1	7500.000000	sqft
2	10400.000000	sqft

DATA PREPROCESSING

5 Data Imputation and Removal of Nulls

Purpose:

- Data Does not necessarily comes cleaner, so imputing null values with mean and mode help in avoiding nulls.
- **Nulls in Zestimate** are replaced by original price.

Applied on:

- Zillow,Facts

```
5   latitude          2406 non-null   float64
6   longitude         2406 non-null   float64
7   price             2406 non-null   float64
8   bathrooms         2329 non-null   float64
9   bedrooms          2333 non-null   float64
10  livingArea        2387 non-null   float64
11  homeType          2406 non-null   object
12  daysOnZillow      2406 non-null   int64
13  zestimate         1923 non-null   float64
```

```
4   state             2406 non-null   object
5   latitude          2406 non-null   float64
6   longitude         2406 non-null   float64
7   price             2406 non-null   float64
8   bathrooms         2406 non-null   float64
9   bedrooms          2406 non-null   float64
10  livingArea        2406 non-null   float64
11  homeType          2406 non-null   object
12  daysOnZillow      2406 non-null   int64
13  zestimate         2406 non-null   float64
```

6 Data Manipulation and Dropping Duplicates

Purpose:

- Manipulating data so that the data can be used for analysis and building regression models.
- Converting the datatypes of few columns from float to int ,like bedrooms, bathrooms, year built etc.
- Checking for duplicates and dropping duplicates

Applied on:

- Facts

HeatingHeating	HasHeating	SnowflakeCooling	HasCooling	ParkingParking	GarageSpaces	F
Wall Furnace	Yes	None	No	1 Covered Parking space	1	
Other	Yes	No Air Conditioning	No	2 Parking spaces	2	

Price per sqft

0	$543	543
1	$418	418

7 Merging Data Frames

Purpose:

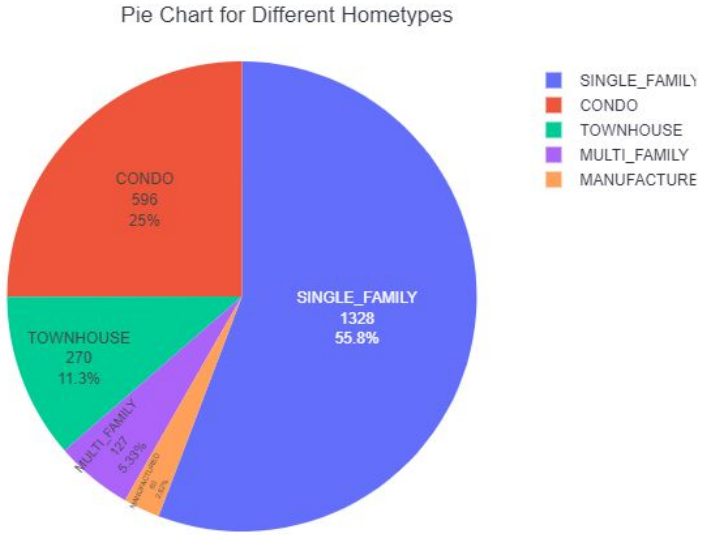
- Merging data frames after data preprocessing to create a big data frame with all the data.
- Inner join on Zillow and making sure

Applied on:

- Zillow and Facts
- Price History and Zillow

Data Visualization and Analysis

Most Popular Home Type




```
The total home in the dataset is 2381

The frequence of each home type
SINGLE_FAMILY    1328
CONDO             596
TOWNHOUSE         270
MULTI_FAMILY      127
MANUFACTURED       60
```

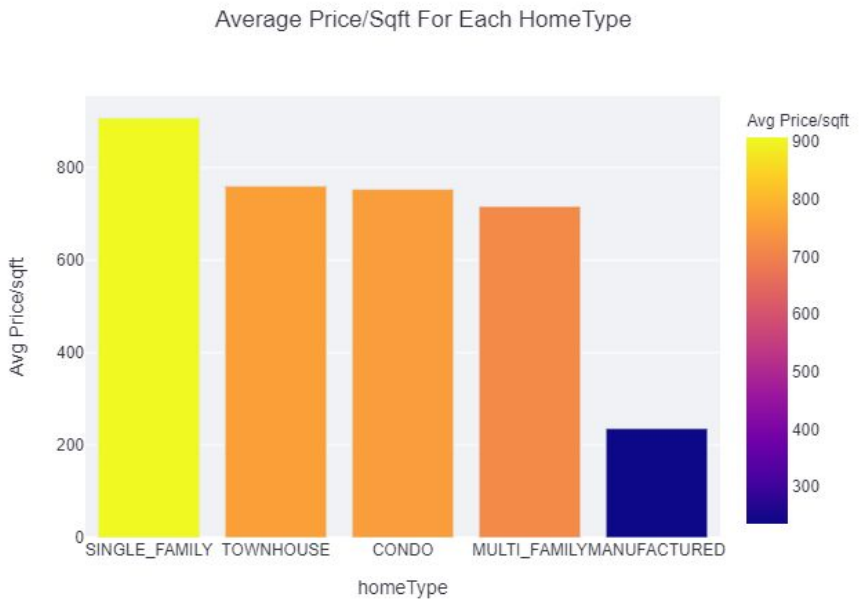
Work 🔨:

- Install plotly.express for better visualization
- Value count on home type column and plot pie chart

Analysis 💡 :

- With 55.8%, single family is the most popular home type in the Bay Area on Zillow.
- Following are condo and townhouse which contribute 25% and 11.3% of home type, respectively
- Multi family is pretty minor with 5.3%
- Manufactured is the less popular home type with 2.5%. This is predictable because the Bay Area one of the most expensive regions of the state

Average Price/ Sqft For Each Home Type



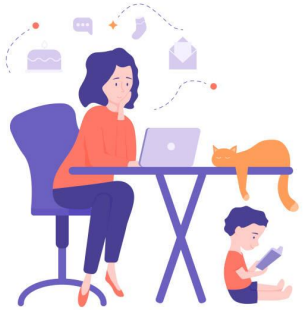
Work 🔨:

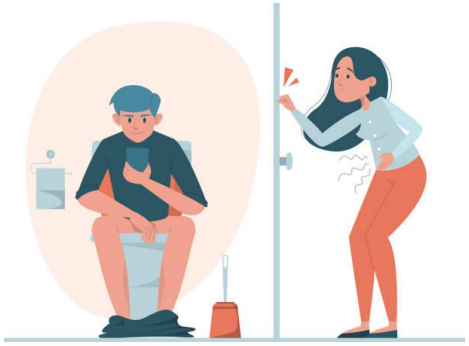
- Home type and price/sqft details were used; Grouped them by home type.
- Calculate average price/sqft for each home type

Analysis 💡 :

- Single Family_ the most popular home type has the highest average price/sqft being $907/sqft
- Following is townhouse and condo for approximately $755/sqft
- Multi family is just around $30/sqft less than townhouse and condo
- Manufactured is the cheapest home type in the bay area for $236/sqft

3 BEDS AND 2 BATHS





The Average Price For 3 Bedrooms And 2 Bathrooms In The Bay Area

$ 1,426,921

- As of Oct 2, 2021, there are 473* houses on Zillow with 3 bedrooms and 2 bathrooms in the Bay Area
- The average price is $ 1,426,921

(* used only limited information from zillow)

Average Price For 3 Bedrooms And 2 Bathrooms For Different Home Type

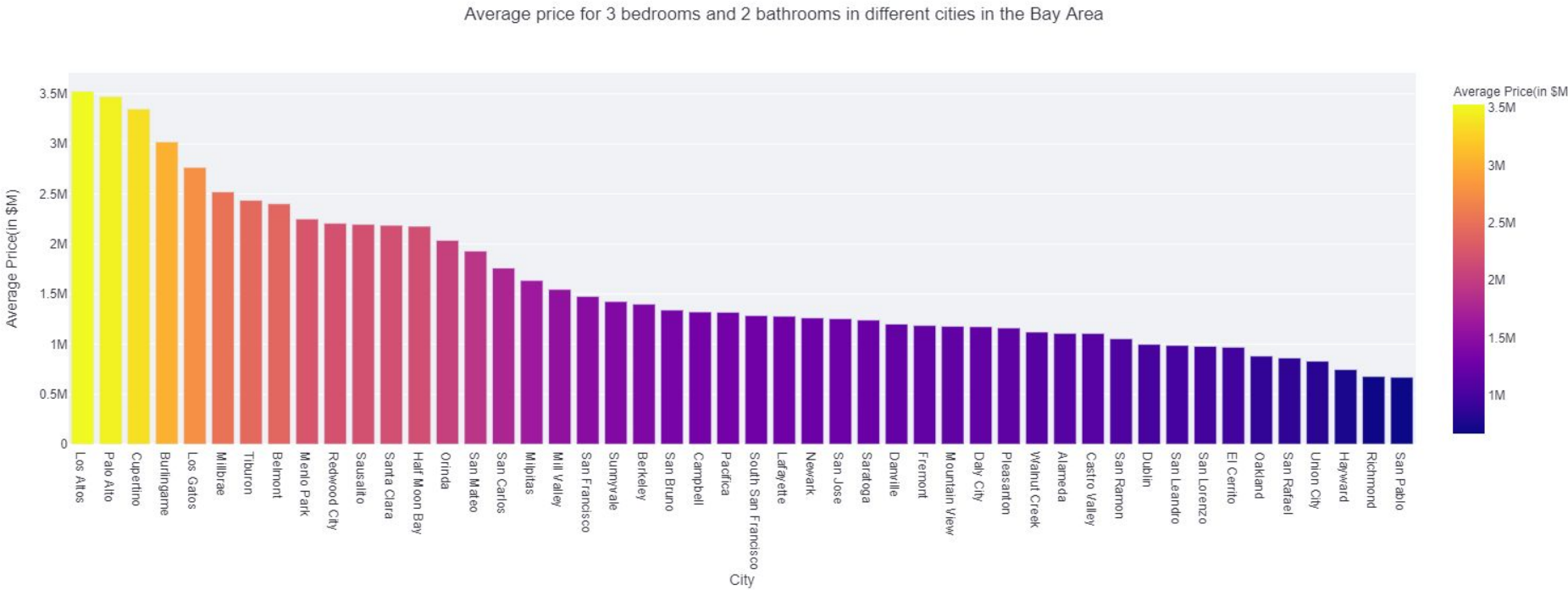
Work 🔨:

- Price and home type of all the houses with 3 bed and 2 bath were used
- Grouped them by home type and average housing price was calculated

Analysis 💡:

- Multi family is the most expensive. It costs on average ~ 2.5 million dollars
- Following are single family and condo for approximately 1.2 million dollars
- Townhouse and Manufactured are the cheapest

Average Price For 3 Bedrooms And 2 Bathrooms In Different Cities



Average Price For 3 Bedrooms And 2 Bathrooms In Different Cities

Less than 800,000:

- Hayward
- Richmond
- San Pablo

Around 1 million dollar

- Mountain View
- Daly City
- Pleasanton
- Walnut Creek
- Alameda

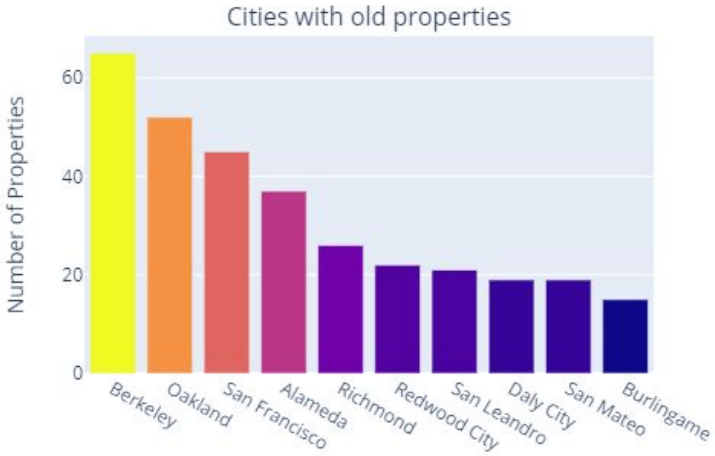
Around 2 millions:

- Redwood City
- Sausalito
- Santa Clara
- Half Moon Bay
- Orinda

Over 3 millions:

- Los Altos
- Palo Alto
- Cupertino
- Burlingame

Cities with oldest and newest properties



Work 🔨:

- 'Calendar Year built' and city information was used
- Sorted the data frame based on 'Calendar Year built' in a descending order
- Considered top and bottom 500 rows; Grouped them on city separately.

Analysis 💡 :

- Berkeley has the most number of old properties
- Mountain view has the most number of new properties

Appreciation rate with respect to city

Work 🔨:

- First and last events happened on all properties were identified
- Time gap(in years) and price change between these two events was calculated
- Price change rate per year was calculated
- Grouped the results by city and calculated the mean
- Top 10 cities were considered

Analysis 💡:

- Danville has the highest appreciation rate of 35% per year
- Palo Alto has the lowest appreciation rate of almost 20% per year (among top 10 cities)

Appreciation rate with respect to home type

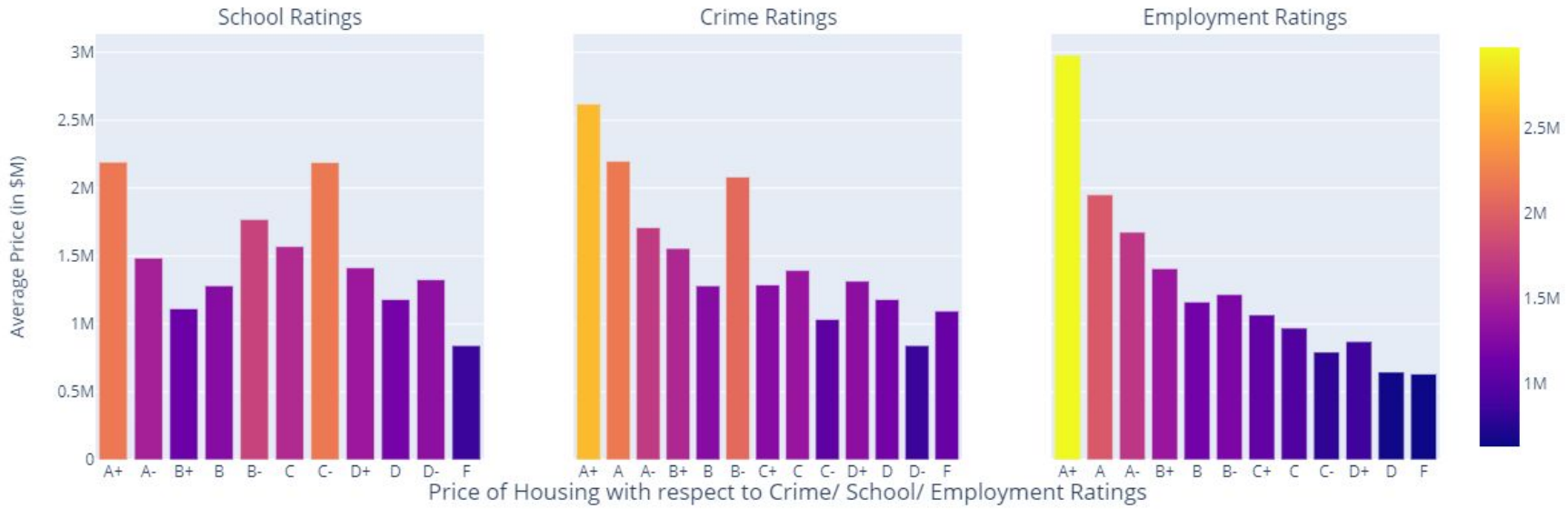
Work 🔨:

- First and last events happened on all properties were identified
- Time gap(in years) and price change between these two events was calculated
- Price change rate per year was calculated
- Grouped the results by home type and calculated the mean

Analysis 💡 :

- Prices of 'Multi family' type homes are increasing the most, which is 22% per year
- 'Single family' and 'Town houses' are next in line

Impact of Crime Rates, School ratings and Employment opportunities on housing prices in Bay Area



Impact of Crime Rates, School ratings and Employment opportunities on housing prices in Bay Area

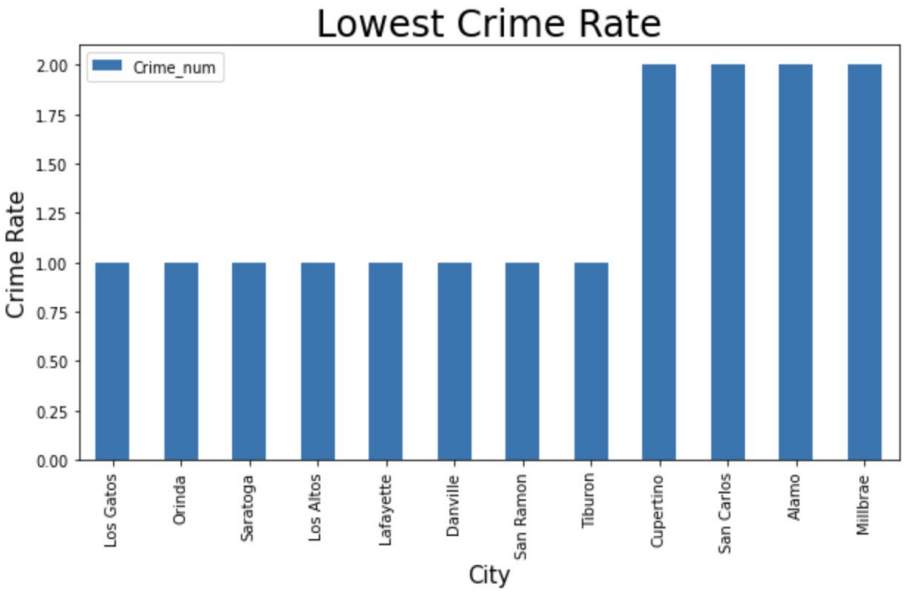
Work 🔨:

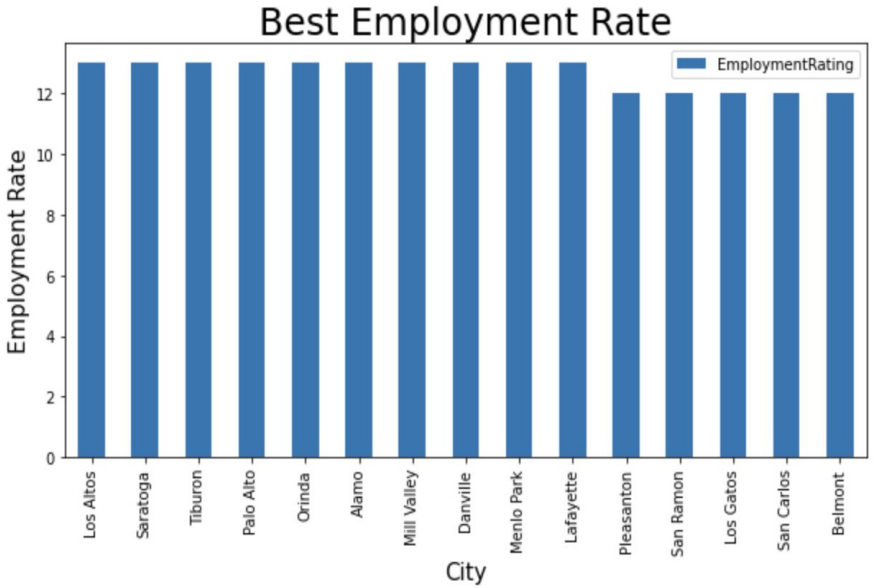
- School, crime and employment ratings for all the houses were considered along with the price details.
- Then calculated the average house prices by grouping the school, crime and employment details separately.
- Rating Vs average price bar charts were plotted for each category

Analysis 💡:

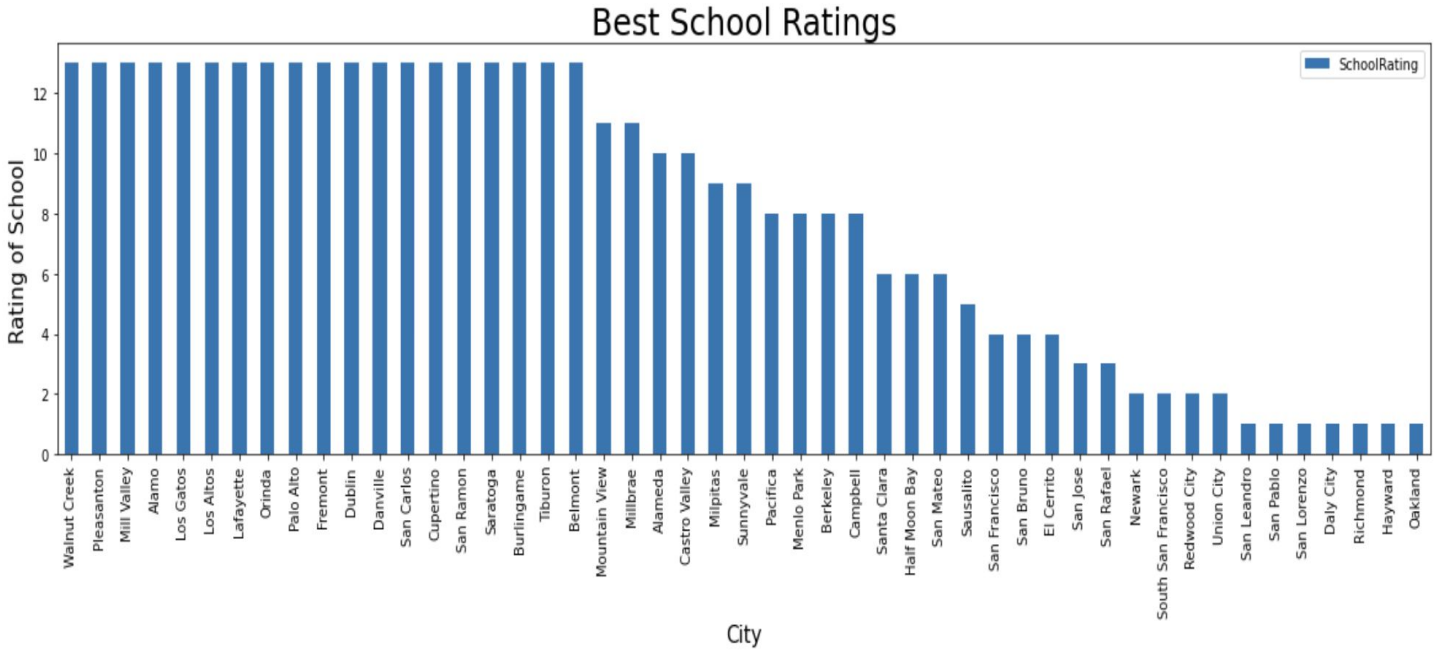
- Cities that have ratings in 'A', have a higher price range of houses.
- Areas with better Employment ratings matter the most and have highest range of house prices.
- Crime rating can be considered as a second priority while house hunting.

City/Cities with lowest Crime Rate and Best Employment Rates





Cities with highest to lowest school ratings



Cities with highest to lowest school ratings

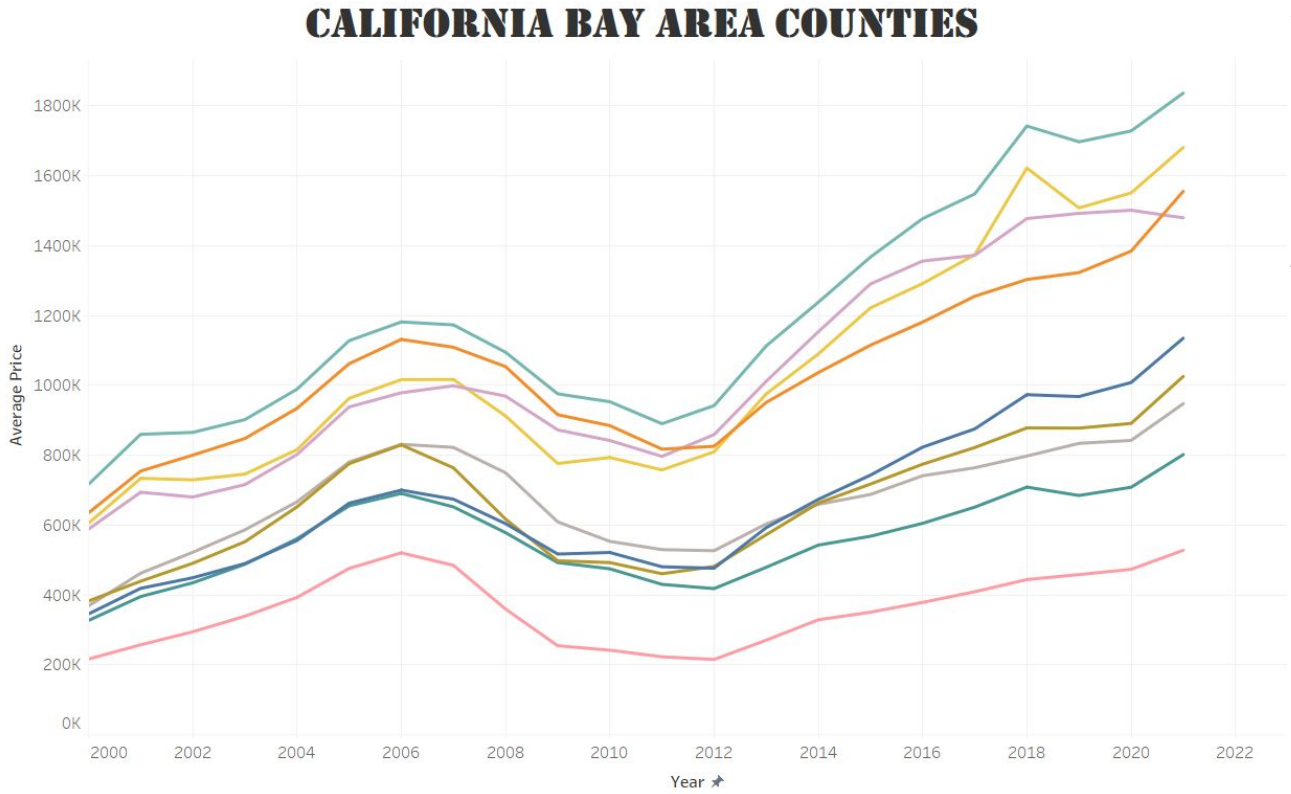
Work 🔨:

- Grouped by cities and calculated average crime, school and employment ratings for each city
- Plotted bar charts for crime, school and employment ratings with respect to city

Analysis 💡 :

- Los Gatos, Orinda, Saratoga, Los Altos, Lafayette, Danville, San Ramon, Tiburon are the safest cities of Bay Area.
- Los Altos, Saratoga, Tiburon, Palo Alto, Orinda, Alamo, Mill Valley, Danville, Menlo Park, Lafayette are the cities with highest employment ratings.
- Walnut Creek, Pleasanton, Mill Valley, Alamo, Los Gatos,Los Altos, Lafayette, Orinda, Palo Alto, Fremont, Dublin, Danville, San Carlos, Cupertino, San Ramon, Saratoga, Burlingame, Tiburon, Belmont are the cities with highest school ratings.

Real estate in the Bay Area



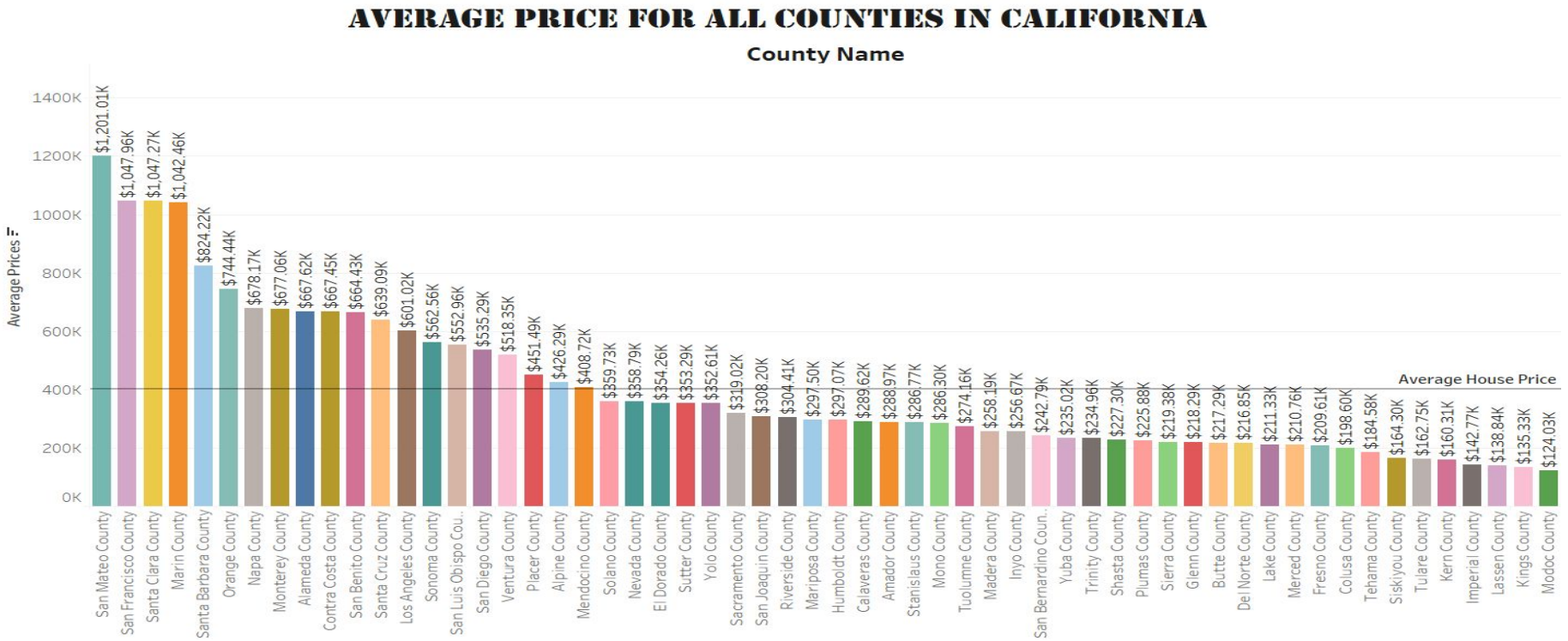
Work 🔨:

- Year and Average prices with respect to each year and county information was used

Analysis 💡:

- All the counties had an increase in the average prices of properties from 2000 to 2006.
- Later, all the counties have seen a downtrend in the average house prices from the end of 2006 to 2012 due to economic recession.
- After that, the average prices have again started to increase

Average housing prices for all counties in California



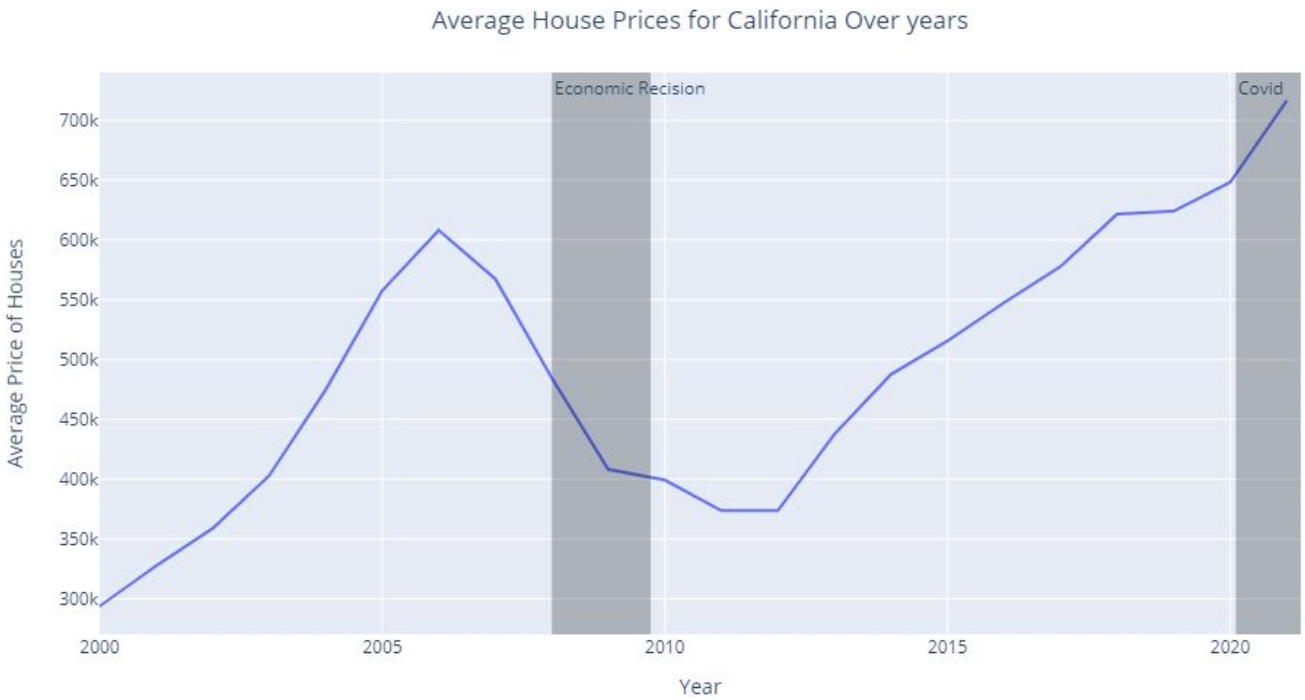
Work 🔨:

- All the counties in CA and Average Prices of each county information was used
- Arranged in the descending order of Average price

Analysis 💡:

- San Mateo County has the highest Average pricing
- Modoc County has the lowest average pricing

Trend in the housing prices over years in California



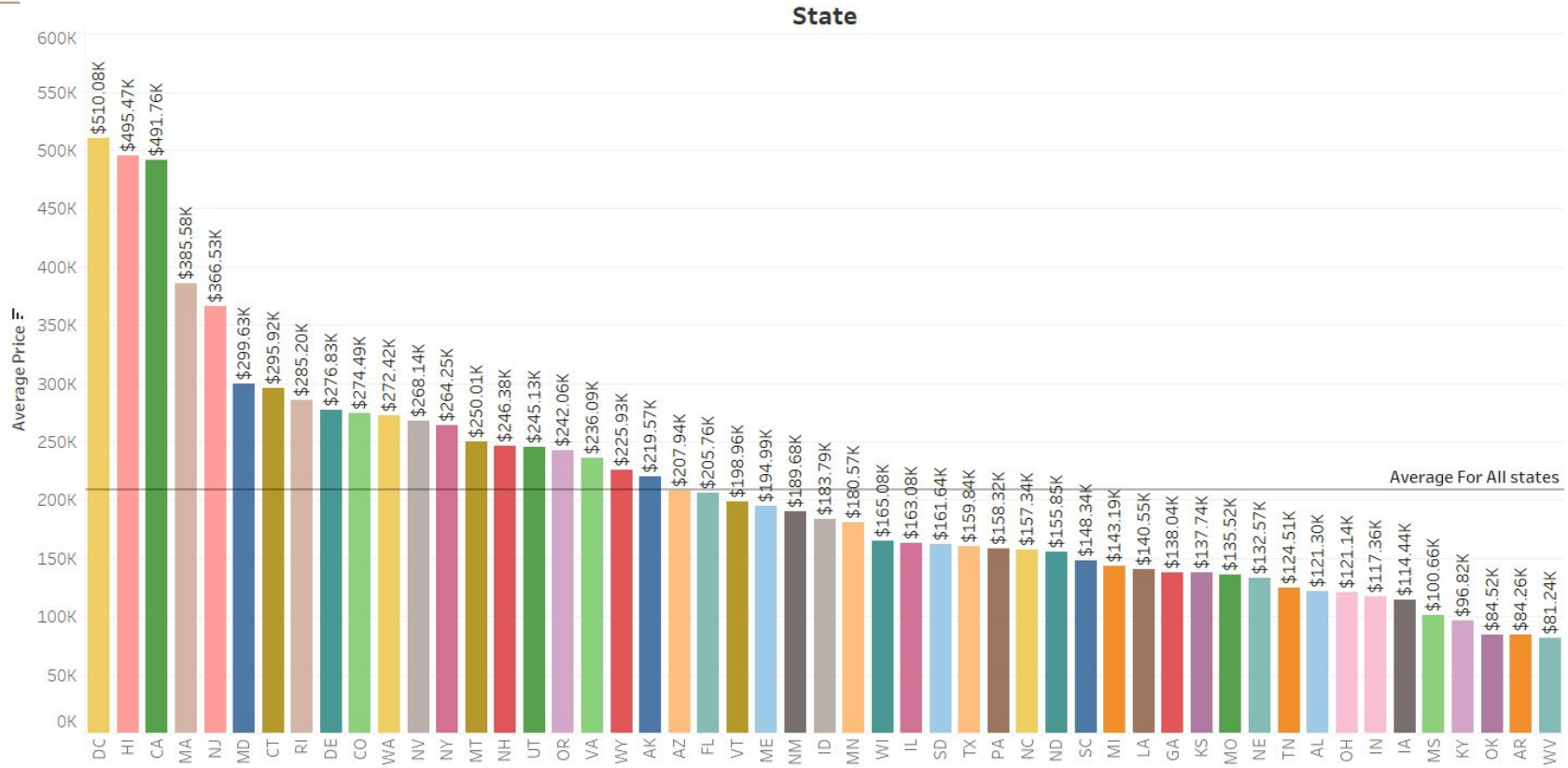
Work 🔨:

- Grouped by year and calculated the average prices of properties

Analysis 💡 :

- An increasing trend has been observed from 2000 till 2006.
- Downfall from 2006 to 2011.
- Since 2012 till date a tremendous increase in the prices can be observed clearly.

Mean Housing prices in all states of U.S



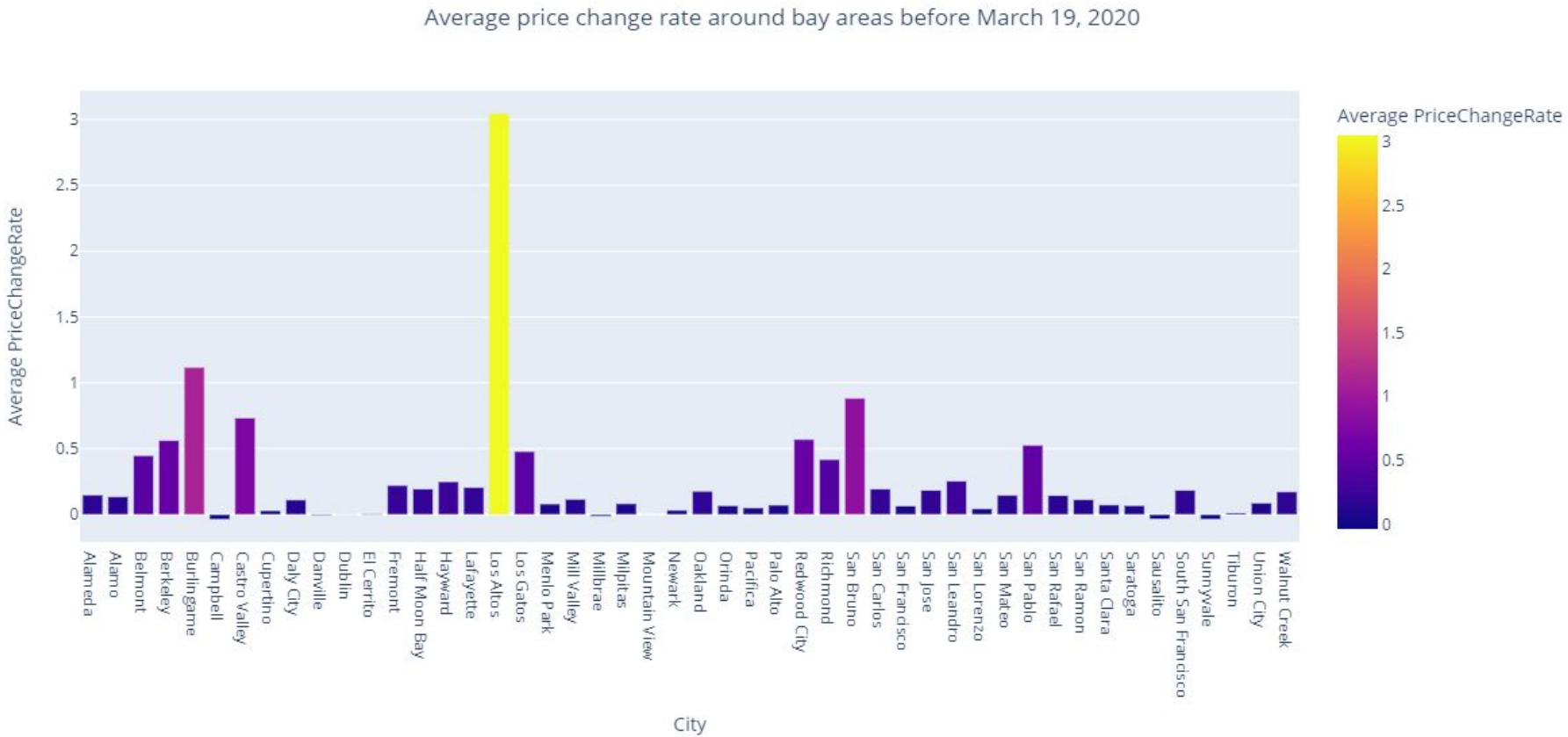
- Among all the states in the U.S, DC has the highest average housing price
- West Virginia has the least average housing price

How did housing prices change since the lockdown?

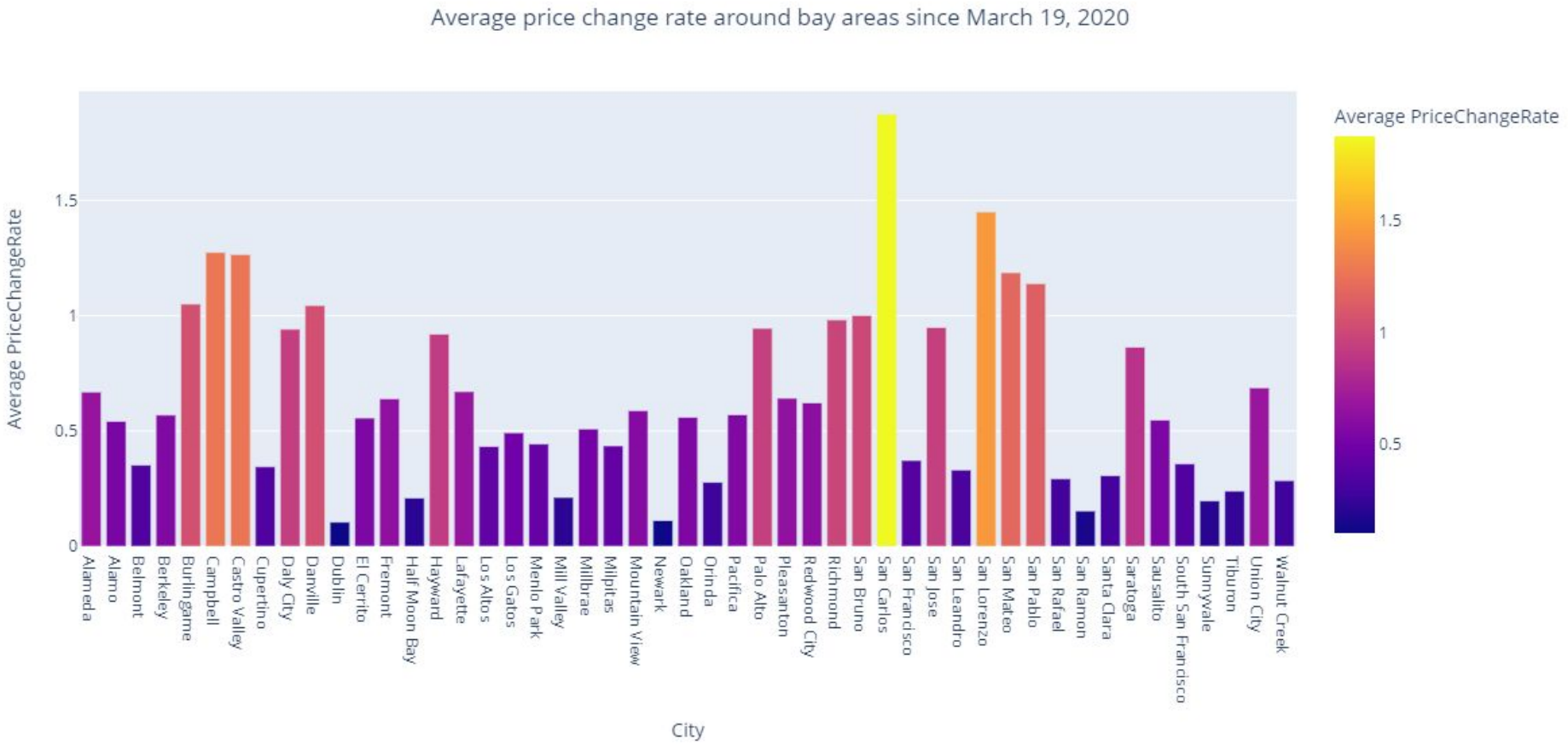
Work🔨:

- Mean price change rate for each city were captured for 2 years from Mar 19,2018 to March 18,2020 (before lockdown)and for 1.5 years from Mar 19,2020 to until now (after lock down).
- For each city in bay area, calculated the average mean price change rate before and after March 19,2020.

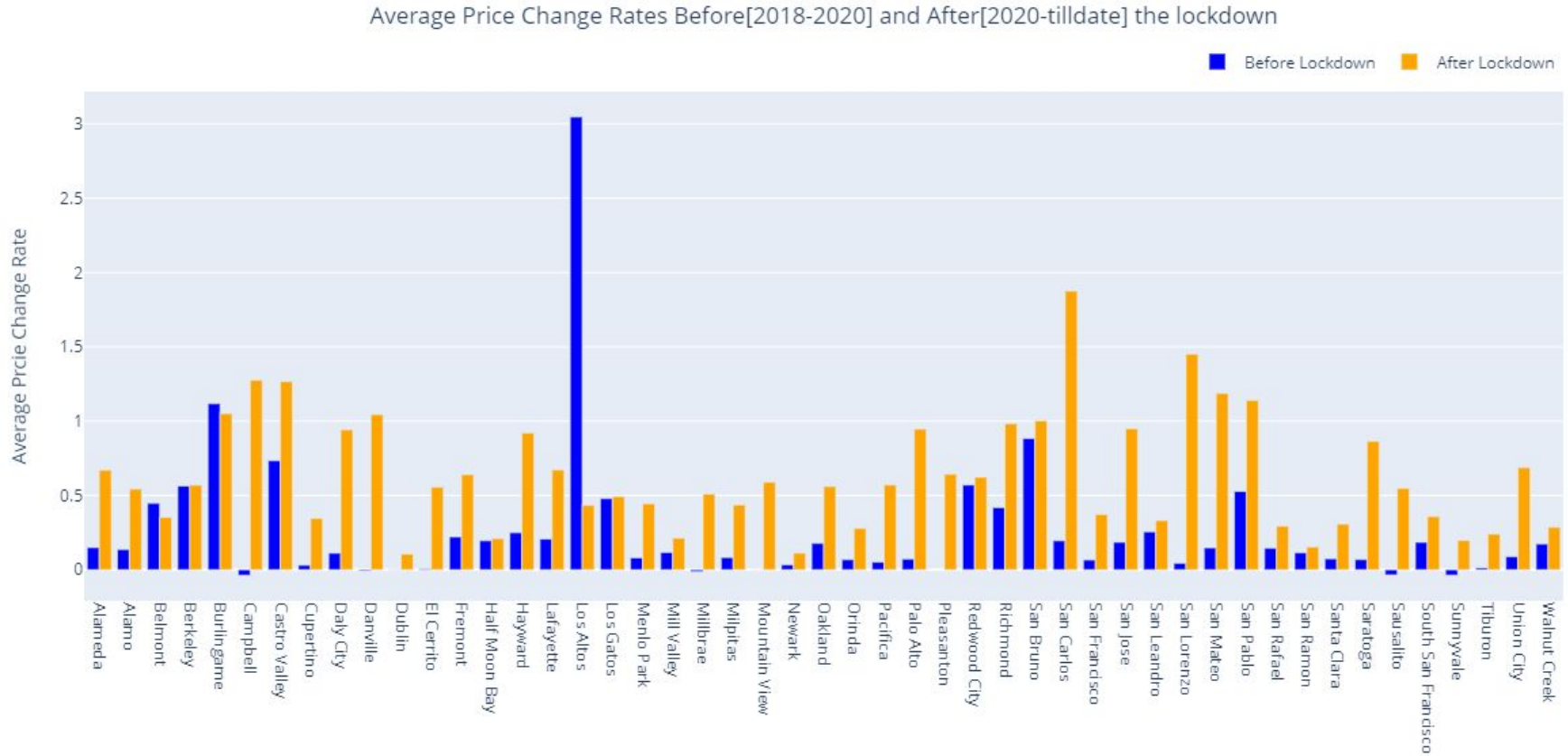
Pre - Lockdown (March 19, 2018 to March 18, 2020)



During Lockdown (March 19, 2020 onwards)



Comparing data before and After Lockdown



Comparing data before and After Lockdown

Analysis 💡 :

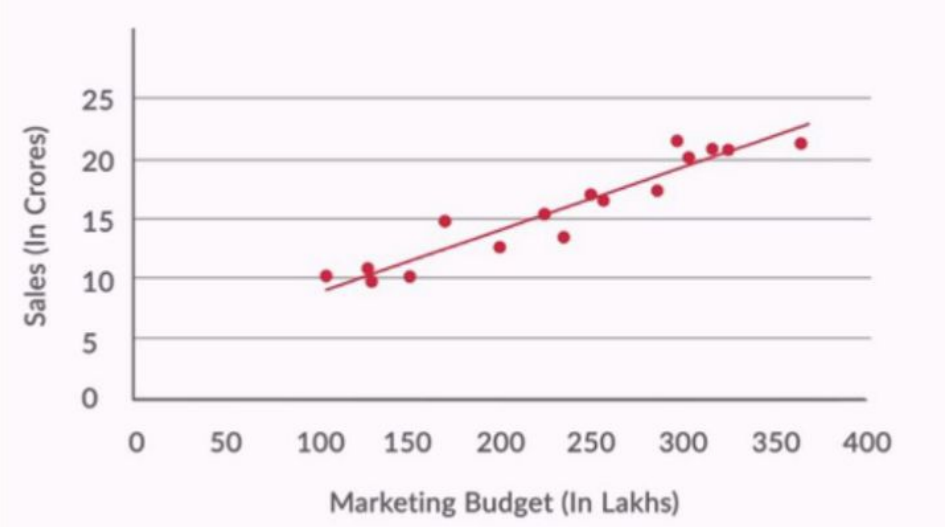
- Comparing average price change rate for data before and over the period of stay-at-home (i.e, March 19, 2020) , it can be observed that almost all the cities changed their housing prices during lockdown, except Belmont, Burlingame, Los Altos where they decreased the same.

REGRESSION

Regression is finding the best fit straight line between the dependent variable (Y) and independent variable. The output variable to be predicted is a continuous variable.

$$Y = \beta 0 + \beta 1 * X \square \text{ Single Independent Variable}$$

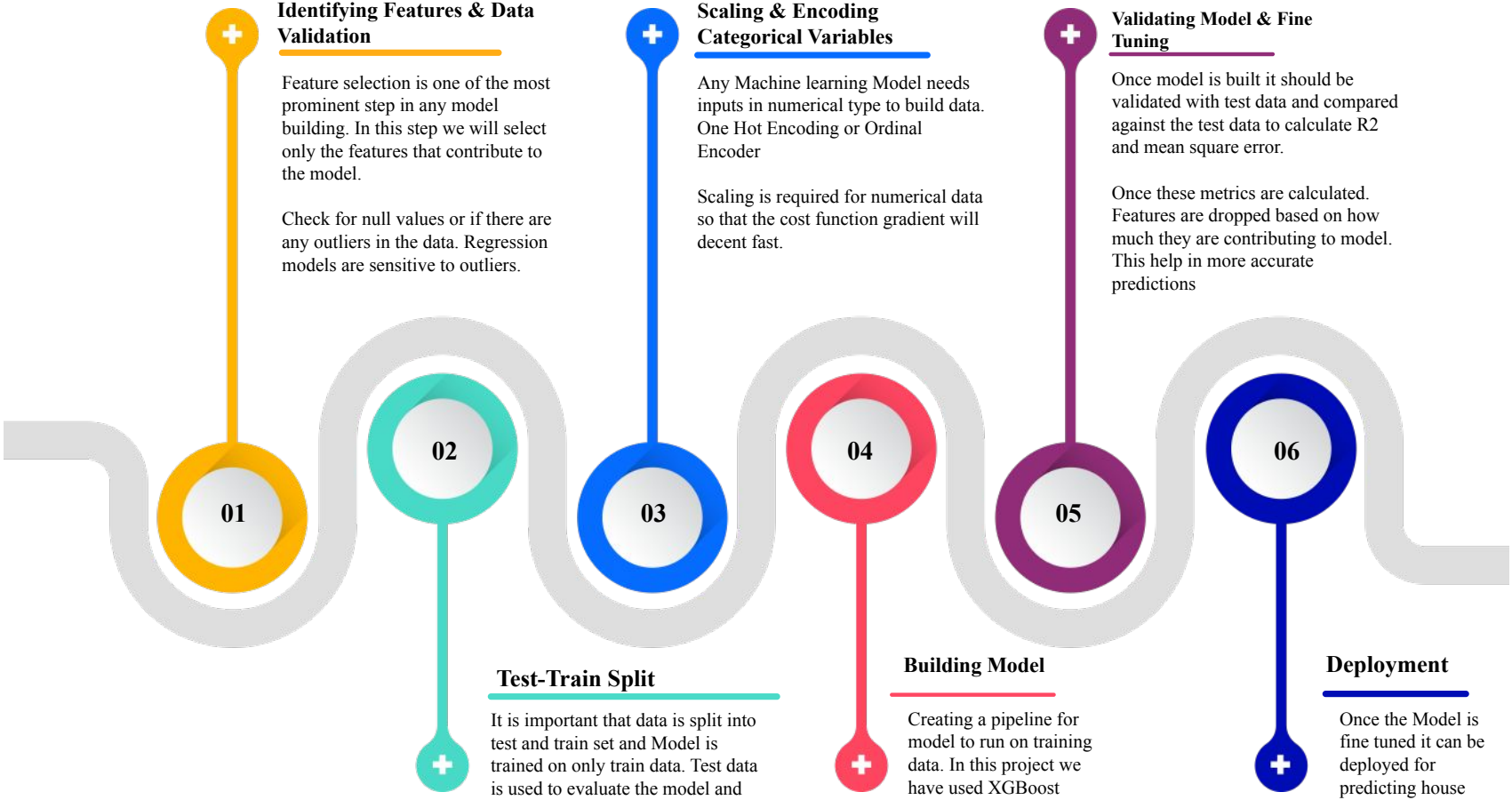
$$Y = \beta 0 + \beta 1 * X1 + \beta 2 * X2 + \ldots\ldots\ \beta n * Xn \square \text{ n Independent Variables}$$



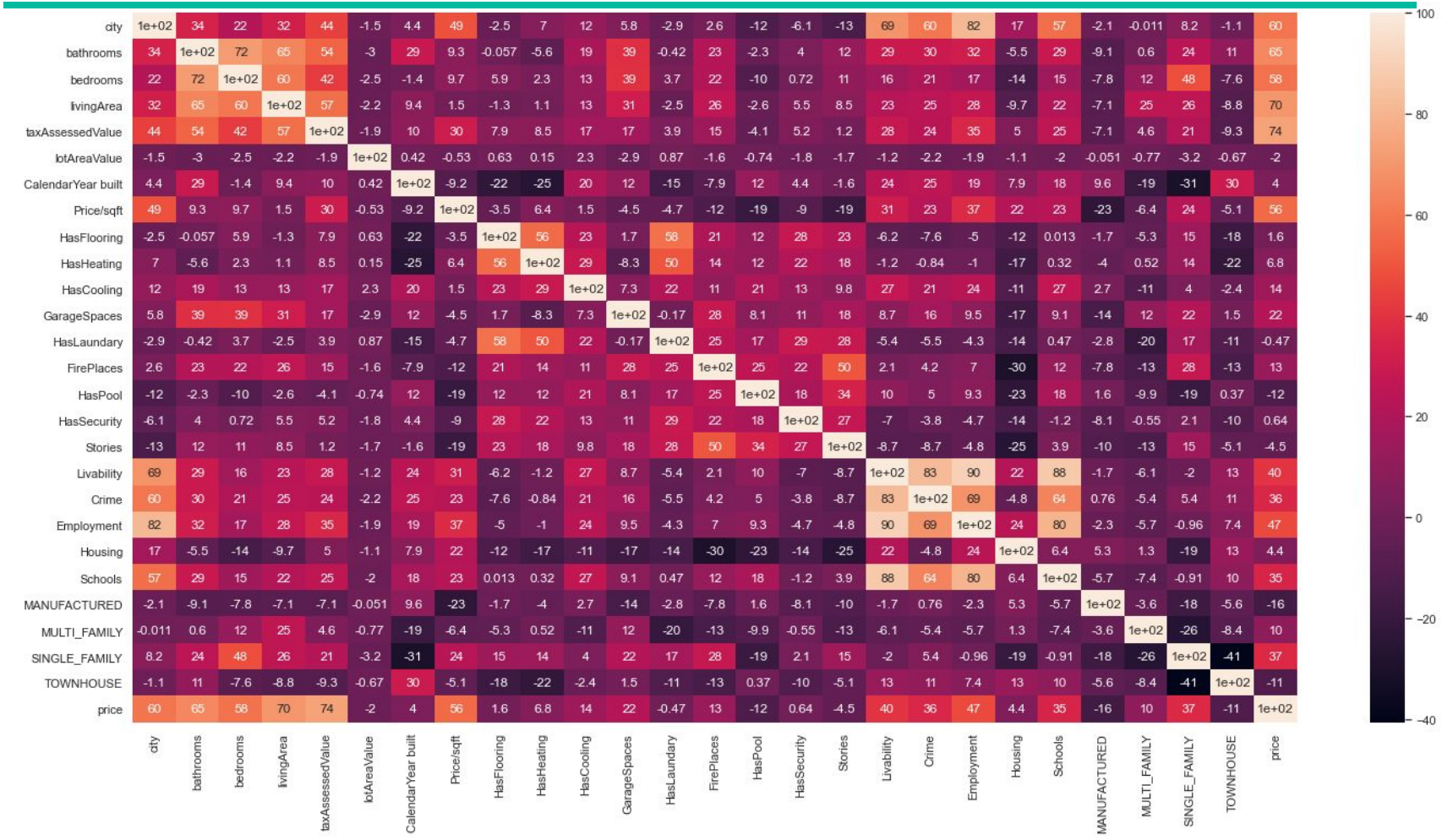
The best-fit line is found by minimizing the expression of **RSS (Residual Sum of Squares)** which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value and actual value.

R-Square is the amount variability in Y (dependent variable) is explained by independent variable(s) X.Usually higher R2 value means the model is good in capturing the variance of Y with respective to X.

REGRESSION PIPELINE



Correlation Heat Map



Metrics

Below are the metrics for the XGBoost model,r2 score close to 98 which indicates that model is able to capture Of the variance in y using the features. The mean square error for Zillow model is 156k and mean square error of XGB model is 167k.

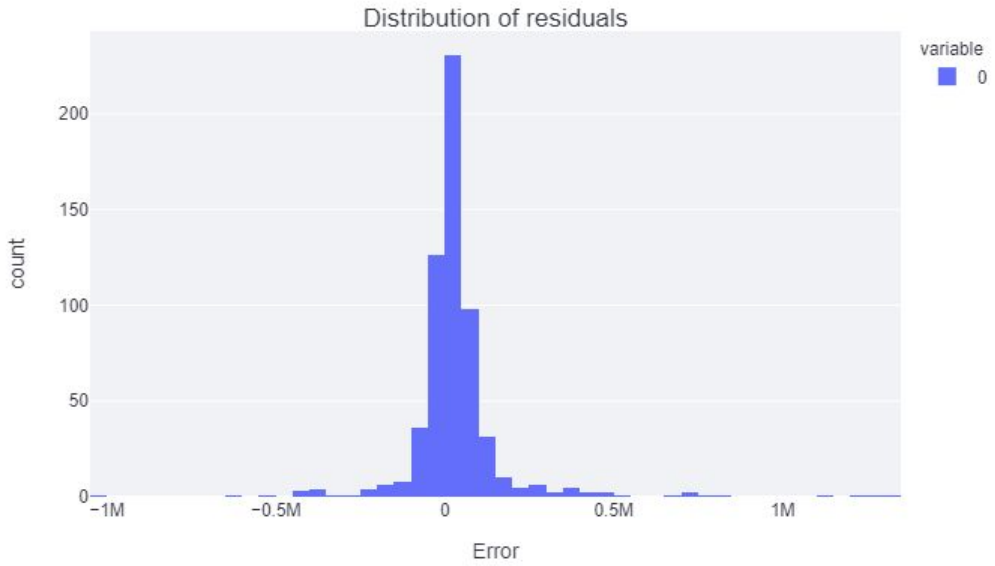
Test r2_score	MSE Zillow Model	MSE XGB Model
0.9800	156,059.5841	167,131.1152

Residuals

Price Predictions for few houses

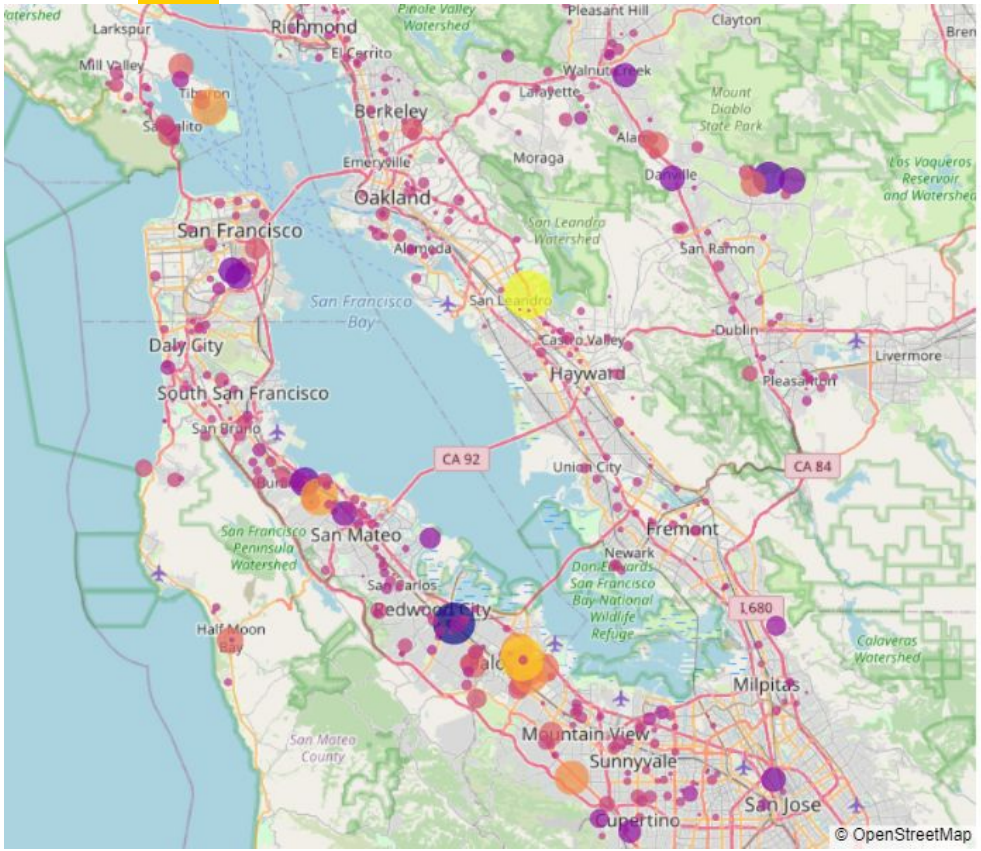
zpid	price	Predicted	zestimate
72547562	1849000.0	1751436.0	1745400.0
2068402452	395000.0	345977.0	395000.0
24857078	1198000.0	1216993.0	1198000.0
299074408	1195000.0	1173075.0	1195000.0
95427573	3999500.0	3845887.0	3999500.0
24902891	1248000.0	1233529.0	1373000.0
2112183577	1175000.0	1178008.0	1288300.0
25037443	1299900.0	1313353.0	1413900.0
24962903	699000.0	665630.0	833400.0
79844142	397000.0	439822.0	401178.0

Distribution of residuals



Residuals

Difference Between the predicted and listed



The size of the of the circle represents The magnitude of the error we made While predicting the house price.

If the circle color is **Dark Blue** that means model has over estimated the price.

If the circle color is Closer to orange or Yellow it indicates that we have under Estimated the price of the house.

THANK YOU


Q
A