Project Proposal on San Francisco Bay Area Housing Data by Zillow Housing Group 1

Introduction to the Dataset:

This dataset (approx 20k records) gives us a detailed look at housing sales data in the San Francisco Bay Area from 2003 to 2006, sourced from Zillow, a big housing information source. In the dataset, we have details about these houses such as county and city, Zip codes, addresses, pricing data, number of bedrooms, size, basement details, construction details, the date they were sold, and exact locations on a map.

With all this information, we want to study the housing market in the San Francisco Bay Area. We'll look for patterns and trends in the data. For example, we'll figure out what makes some houses more expensive than others and see how the location and features of a house affect its price.

This analysis will help people who want to buy a house, real estate experts, investors, and even people who make rules about housing in the area. It'll give them a better idea of how the housing market works in one of the busiest and most exciting parts of the country.

Summary Statistics and Visualization:

Before diving into the proposed analyses, it's important to explore the dataset's characteristics. We will start by calculating some summary statistics and creating visualizations to gain insights into the data. This will include:

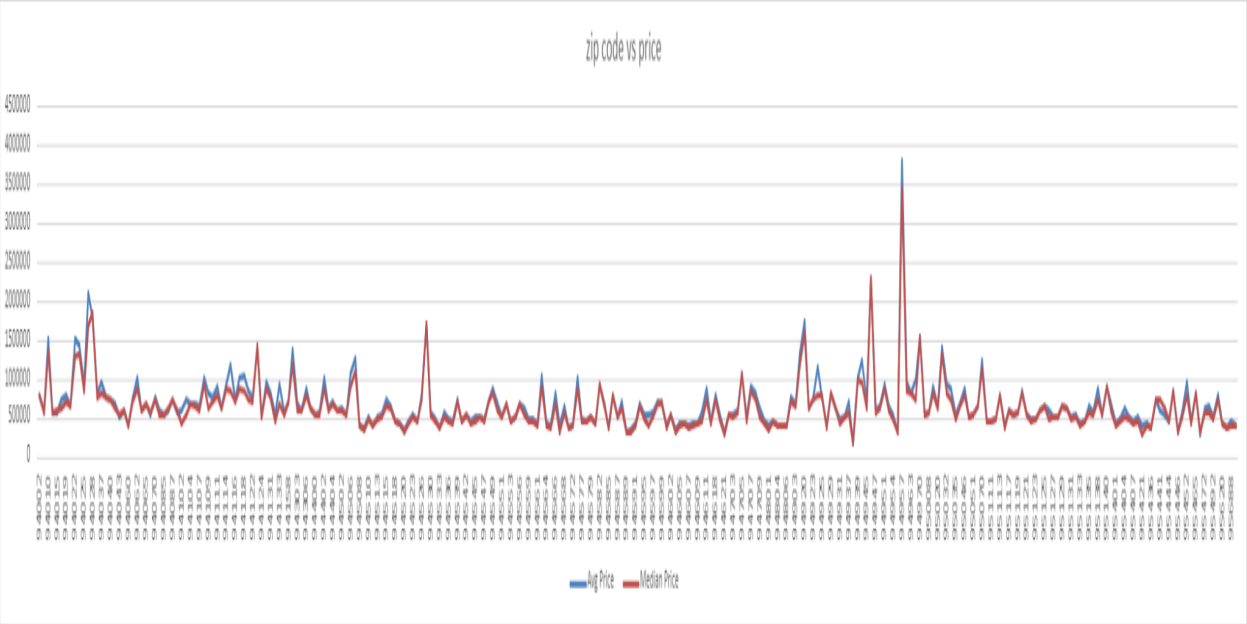
1. Basic Descriptive Statistics: We will compute measures such as Mean, Median, Standard Deviation, Quartiles for key numerical columns like ‘price’, ‘lsqft’, ‘bsqft’, ‘bedrooms’ and ‘year’. This will help us understand the central tendencies and variability of these variables.

	zip	price	br	lsqft	bsqft
count	19997.000000	2.000000e+04	16187.000000	1.657400e+04	17079.000000
mean	94691.279142	6.126222e+05	3.027182	5.348851e+04	1601.256338
std	394.443261	3.550735e+05	1.005055	2.630954e+06	736.156427
min	94002.000000	0.000000e+00	1.000000	2.500000e+01	370.000000
25%	94520.000000	4.020000e+05	2.000000	3.760000e+03	1119.000000
50%	94582.000000	5.350000e+05	3.000000	5.663000e+03	1432.000000
75%	95035.000000	7.150000e+05	4.000000	7.807000e+03	1899.000000
max	95694.000000	7.000000e+06	28.000000	3.136320e+08	22266.000000

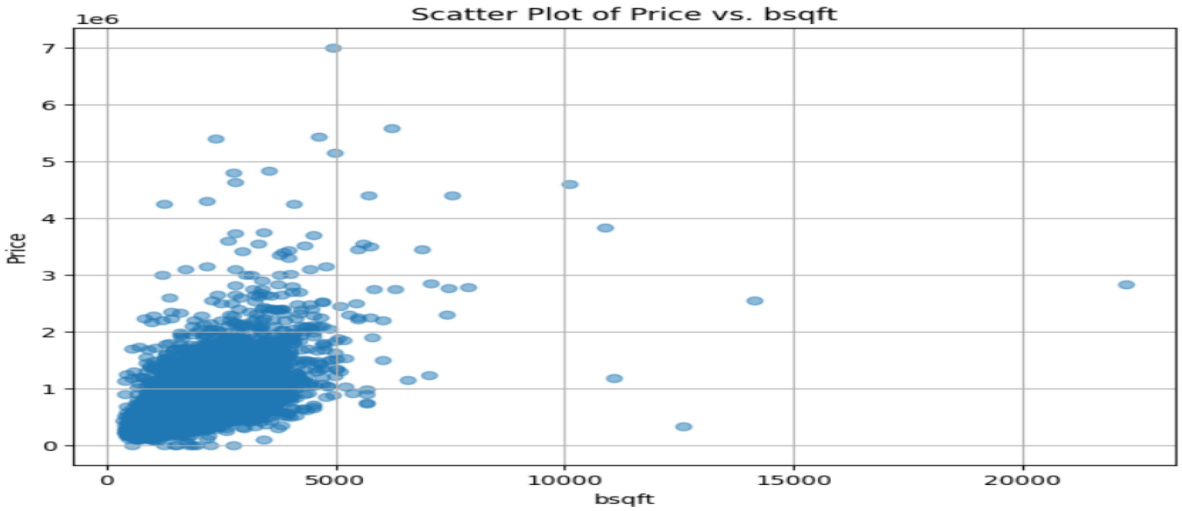
	year	long	lat
count	16493.000000	18104.000000	18104.000000
mean	1969.247317	-122.121888	37.787459
std	287.693629	0.952728	0.452125
min	0.000000	-123.557620	0.000000
25%	1953.000000	-122.304861	37.531792
50%	1970.000000	-122.072324	37.771202
75%	1985.000000	-121.921936	37.999453
max	20005.000000	0.000000	38.825318

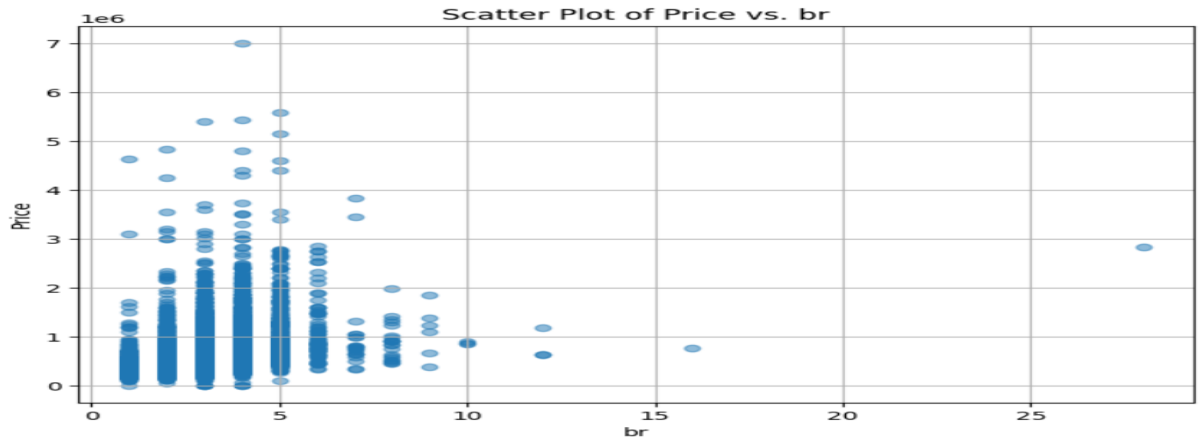
2. Data Distribution Visualization: We will create histograms, box plots, maps, and scatter plots to visualize the distribution of numerical variables and identify potential outliers.

Zipcode vs Price



City vs Price





3. Geospatial Visualization: In the future, using the latitude and longitude columns, we can create a geographical visualization, such as a scatter plot on a map, to see the spatial distribution of housing sales in the San Francisco Bay Area.

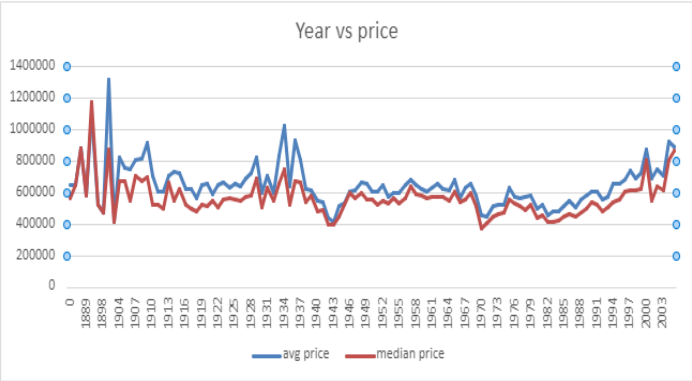
Proposed Data Generation Model and Hypotheses:

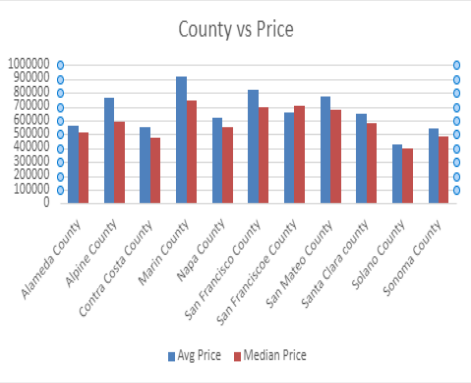
1. Data Generation Model: We can use the housing data to build a predictive model for housing prices in the San Francisco Bay Area. A suitable model might be a regression model, where the target variable is 'price,' and the predictors include features like 'lsqft', 'bsqft', 'year','bedrooms' and geographical location (latitude and longitude). This model can be used to generate price estimates for new properties.

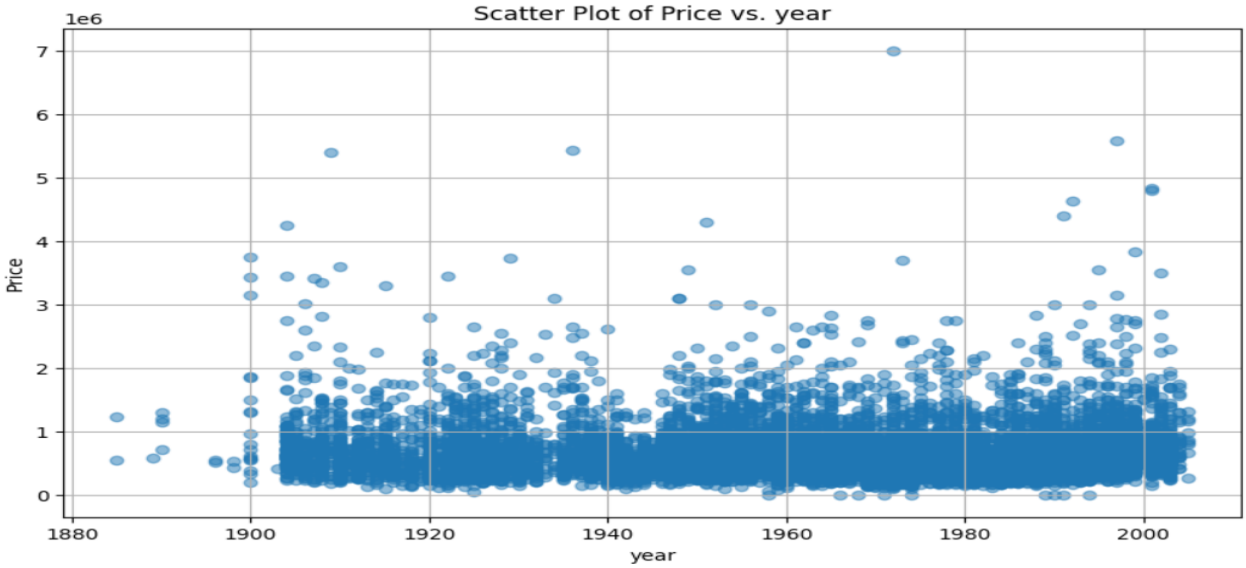
2. Hypotheses:

Hypothesis 1: Housing prices are positively correlated with living square footage ('lsqft'). We can test this by performing a correlation analysis and running a regression model.

Hypothesis 2: The year of construction ('year') has an impact on housing prices. Newer properties may be priced higher. We can test this hypothesis by analyzing the relationship between 'year' and 'price.' Below are few plots for Year & Price after initial evaluation of the dataset.







Conclusion:

This project includes data cleaning, exploratory data analysis (EDA), model development, and hypothesis testing. It aims to provide insights into the San Francisco Bay Area housing market and create a predictive model for housing prices. Additionally, we will visualize the geographical distribution of housing sales. The project also includes feature engineering, model evaluation, and if possible, development of a web application or dashboard for housing price estimation based on these results. This project proposal outlines the initial steps for analyzing the housing data in the San Francisco Bay Area, with the potential to expand into a comprehensive analysis and application development.