House Price In Seattle

Anqian Li
Yaqiong Liu
Chen Chen

·Background Information

The dataset we analyze contains house sale prices for King County. King County is a county located in the U.S. State of Washington and it is the most populous county in Washington, and the 13th most popular County in the United States. The county seat in Seattle. which is the state's largest city.

·Analysis Goal

We know that house prices are affected by many different factors, as the size of the house is not the only factor that has a significant influence. For example, a house with more bedrooms and bathrooms might have a higher price compared to the other houses in the same community. A house has a view of waterfront might cost much than the houses with similar size but no views of waterfront. We want to deeply investigate, which factors affect the house price in King County and especially how much influence they have. We would like to focus on the house price changes in Seattle.

·Potential factors related to house prices

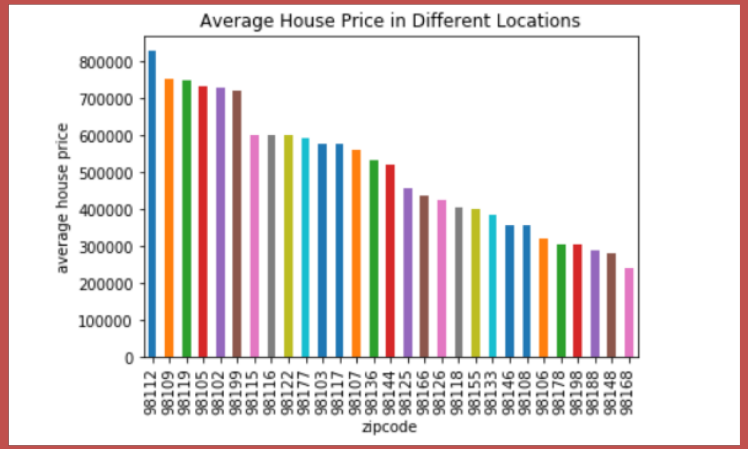
House Quality

Location

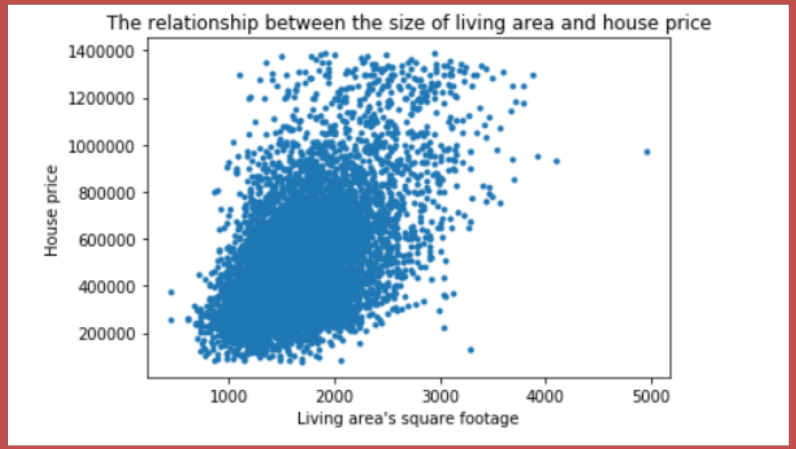
Neighborhood Environment

Living Area

Exploring the Potential Factors



Based on the plot we can see that area of 98112 has the highest housing price, and compare to 98168, which has the lowest price.



As the living area increases, the house price tends to increase.

Building the Model

Linear Regression Model

- Feature Selection: we want to analyze whether different locations, environments, quality and living areas will influence house prices, so we include zipcode, living area, condition, waterfront, number of bedrooms and bathrooms and the year built in our model.
- Preliminary Check: we first draw the pairplot of variables and notice that there are no large correlation between different independent variables, so we don’t need to worry about the collinearity problem.
- Model Budiling: Fit a linear model and check the summary of model, which shows us the coefficient of different variables and their statistical significance.

Conclusion Based on Analysis

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.278e+06	1.13e+05	11.292	0.000	1.06e+06	1.5e+06
zipcode[T.98198]	-3.868e+05	1.6e+04	-24.241	0.000	-4.18e+05	-3.55e+05
zipcode[T.98199]	-2.585e+04	1.57e+04	-1.644	0.100	-5.67e+04	4966.834
waterfront[T.1]	3.986e+05	2e+04	19.959	0.000	3.59e+05	4.38e+05
sqft_living15	179.8853	3.725	48.290	0.000	172.583	187.187
condition	2.757e+04	2228.809	12.371	0.000	2.32e+04	3.19e+04
bedrooms	1.751e+04	1658.770	10.557	0.000	1.43e+04	2.08e+04
bathrooms	8.175e+04	2659.614	30.736	0.000	7.65e+04	8.7e+04
yr_built	-614.3709	56.476	-10.878	0.000	-725.077	-503.665

(Note: this is only part of summary table. We leave out some levels of zipcode here for the display purpose)

A cheaper house or a nicer house?

As we can see from the summary of this linear model, all variables except some levels of zipcodes are statistically significance, so we may conclude that both house quality (condition, yr_built), locations(zipcode), environment (waterfront) and living area(sqft_living15, bedrooms, bathrooms) are related to house prices. However, a house with larger area and higher quality will generally associated with higher house price, so when we are planning to buy a house, we might need make some compromise.