Analysis and Prediction of House Sales in King County, USA

BY

SOPHIA MBATARU

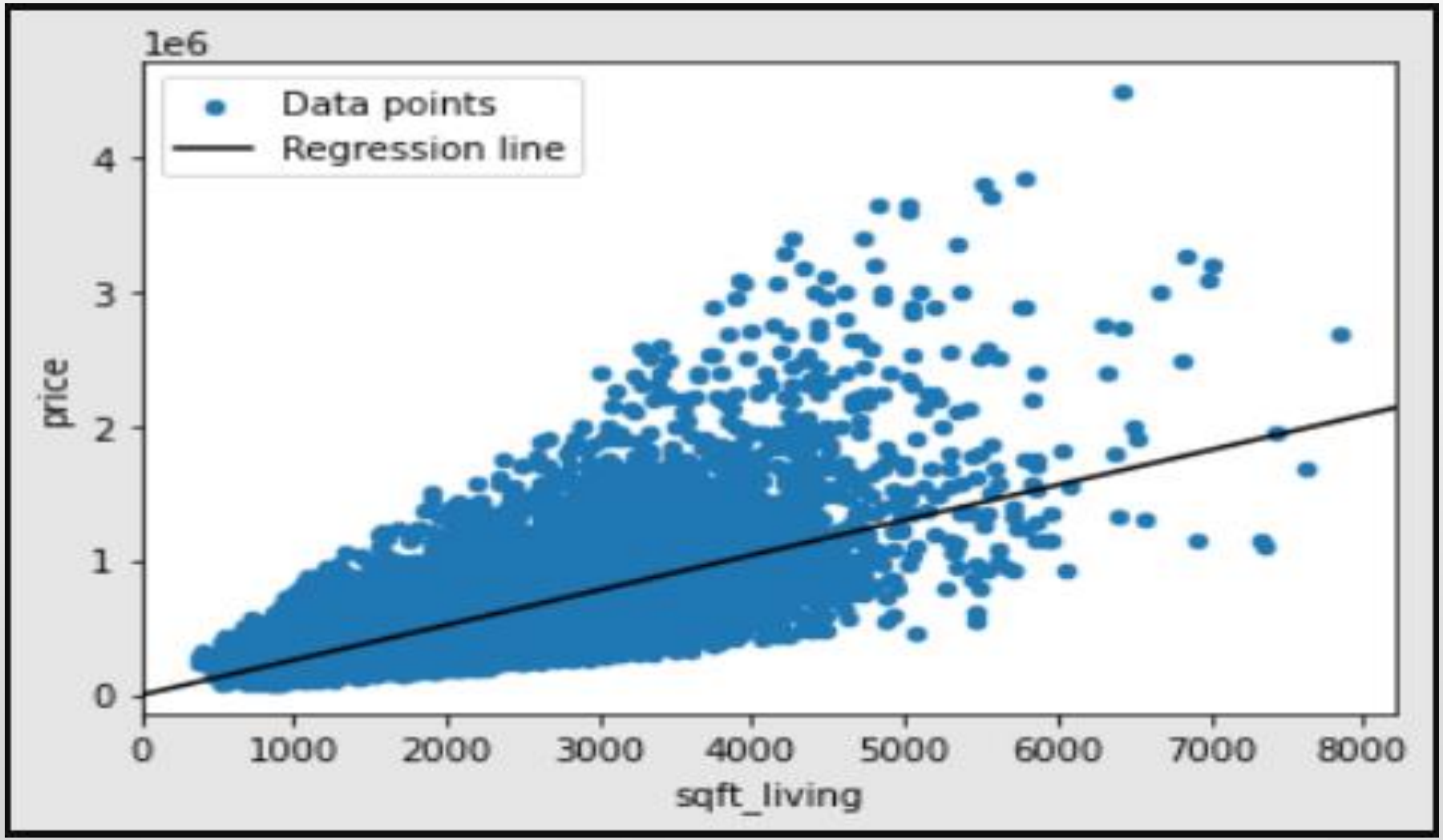
Business Understanding

- King County is a county located in the U.S. state of Washington. The house prices and its spatial distribution are important for stakeholders in the real estate business particularly in metropolitan areas. Stakeholders, such as, external customers looking to purchase or sell a house in King County, they would require to decide on the house to choose based on the variety of parameters associated with the house prices.
- The objective of the study is to use statistical analysis to find the dependence of these variables on the price of houses, and which parameters affect the housing prices and which variables have minimal affect on the price of houses and ultimately make recommendations to stakeholders. The statistical tools used are, Correlation and Regression. Insights between the variables are drawn from scatter and regression plots, and histogram.

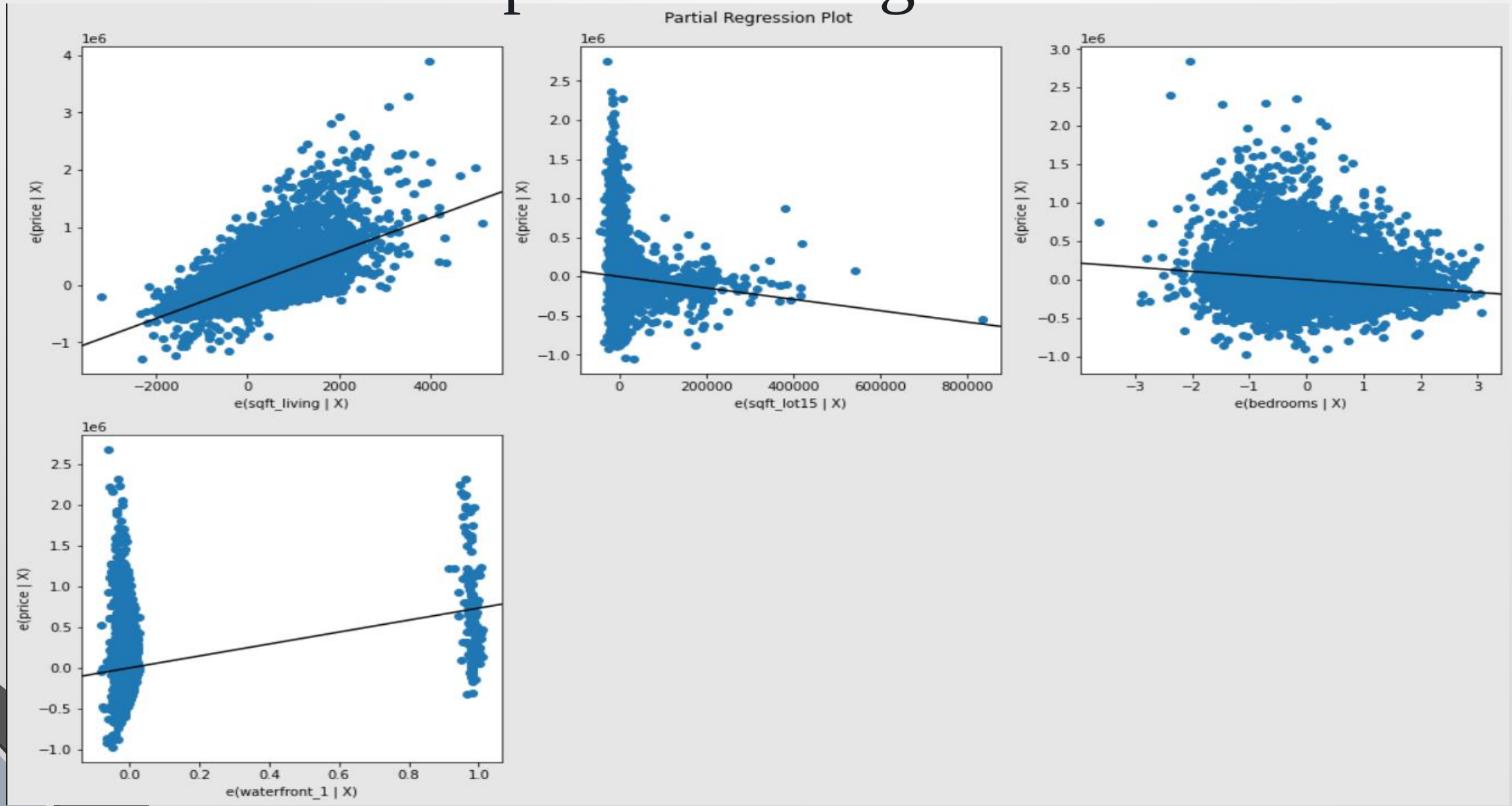
Data Understanding

- The dataset we have taken is House sales in King County, which can be found in kc_house_data.csv in the data folder. The data contains the prices of houses against a variety of parameters, for example, bedrooms, bathrooms/bedroom, square foot area of the house and lot, presence of a waterfront, views, condition of the house, grade assigned by the county, built year, renovated year and the location of the house.
- The data was cleaned and analyzed using python library, Pandas. Missing values were either dropped or replaced with appropriate values. Outliers were removed.

Modeling
Simple Linear Regression



Modeling
Multiple linear Regression



Regression Results

- The adjusted R-Squared for simple linear regression: 0.4610716981513801 6
- The adjusted R-Squared for multiple linear regression: 0.5127811244220679
- The Error-Based Metric (mean absolute error) for simple linear regression: 166350.62192683897
- The Error-Based Metric (mean absolute error) for multiple linear regression: 159875.09332006477

Conclusion

- The data understanding, data preparation and data cleaning allowed me analyze, model and evaluate the data on the King County dataset.
- The key takeaways are that sqft_living, waterfront, sqft_lot15 and bedrooms are the best predictors of a house's price in King County.

Recommendations

- Based on these findings the recommendations to the external customers looking to purchase/sell a house in King County are:

1. Homeowners interested in selling their homes at a higher price should focus on expanding square footage of the living and lot are thus improving the quality of construction.
2. When expanding square footage, homeowners should consider building additional bedrooms and waterfronts, as this analysis suggests that number of bedrooms and presence of waterfronts are positively related to price.

Next Steps

- The next steps I would pursue would be:

1. To explore the best predictors of the prices of homes outside of King County.
2. Given that outliers were removed, the model may also not accurately predict extreme values. I would also explore the predictors of the prices of homes with extreme price values.

THANK YOU!!!