Kingston County House Data Analysis

Objective

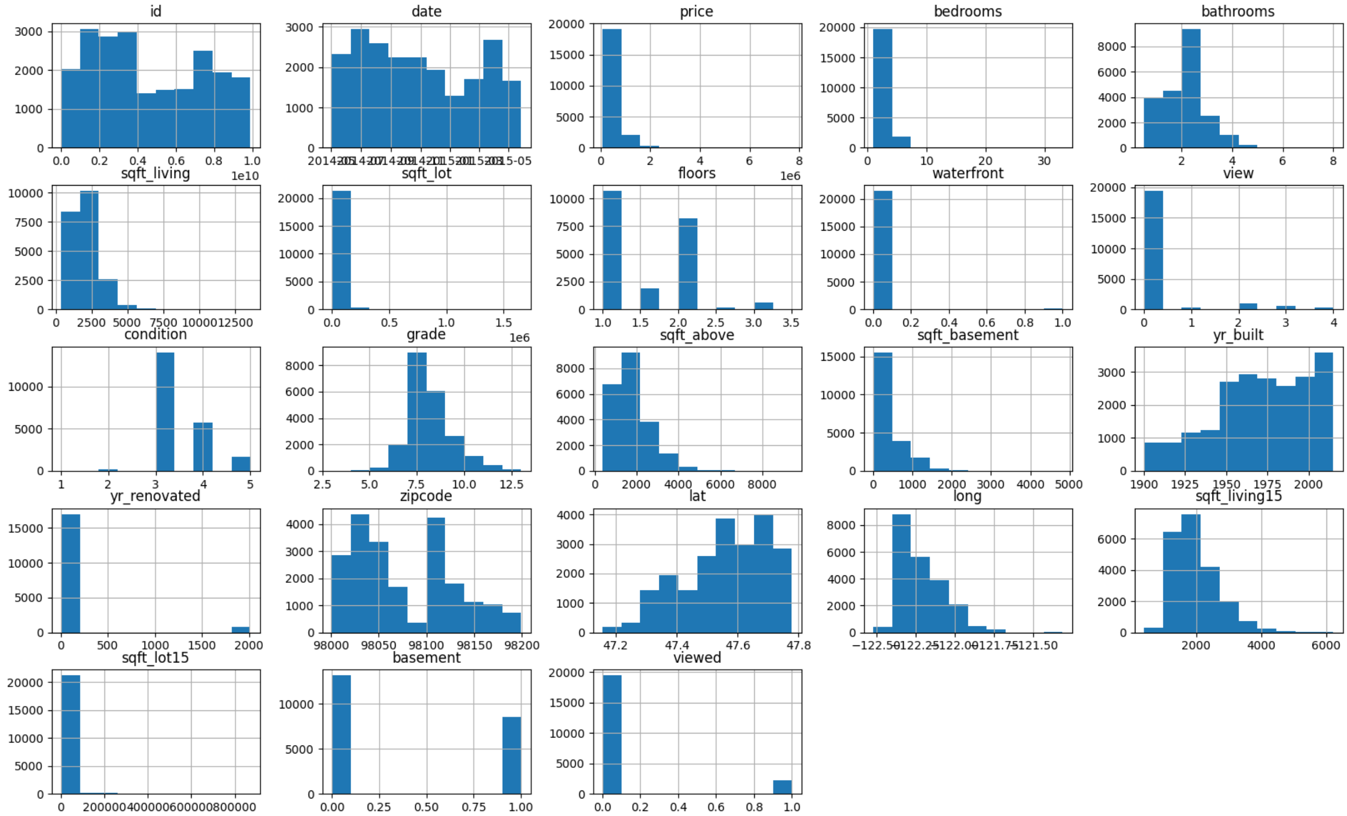
In this project, we examined the different factors that impact the pricing of houses in King's County. According to our best predicting multiple regression model, squarefoot living and the grade were the best predictors of housing prices in Kingston county

Python Library Tools

- Pandas
- Numpy
- Seaborn
- Matplotlib
- Scikit Learn
- Statsmodels

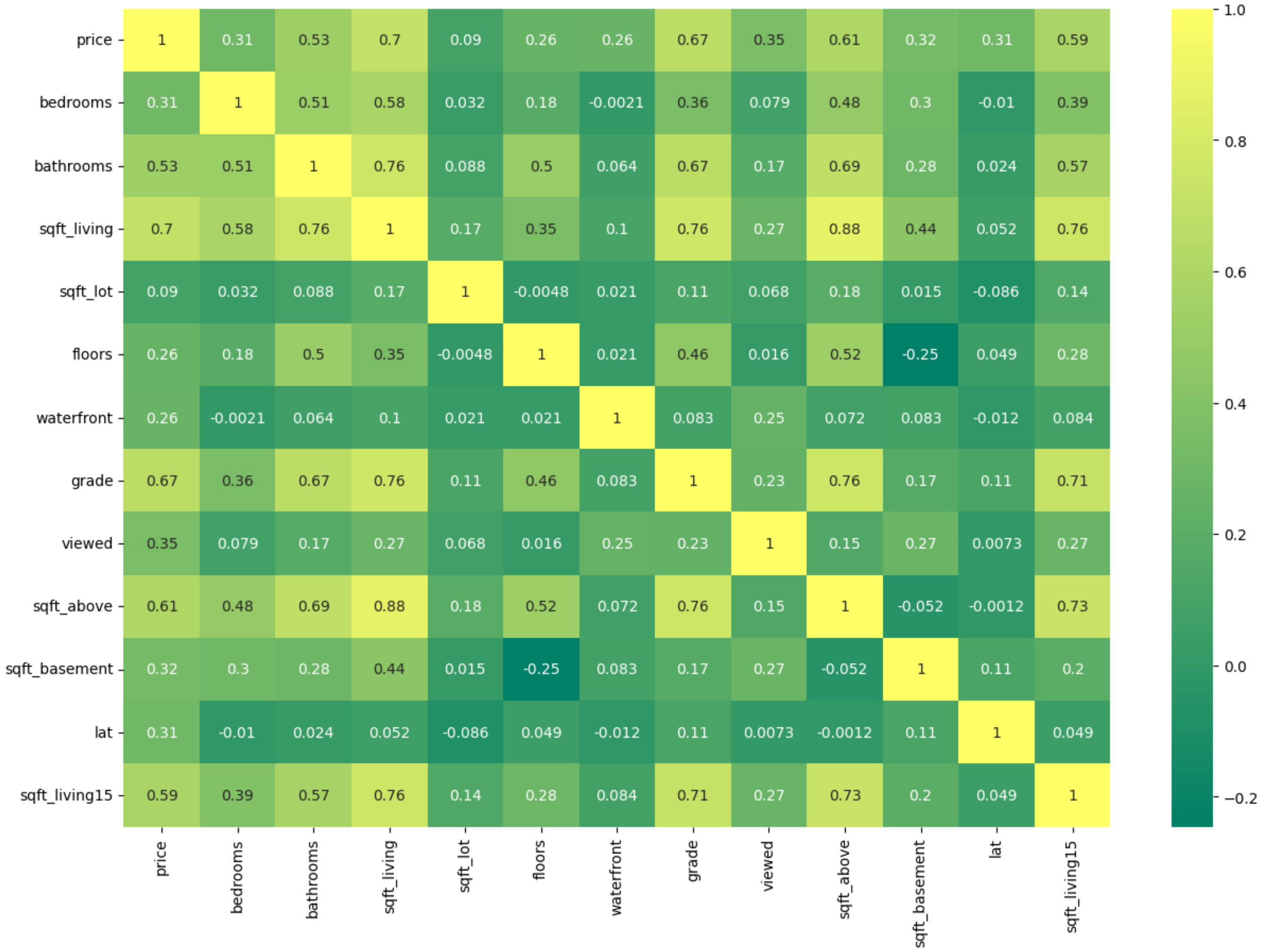
Data

The dataset utilized in this project was obtained in CSV format from Kaggle, consisting of 20,000 data points. The dataset includes the following features: id, date, price, bedrooms, bathrooms, sqft_living, sqft_loft, floors, waterfront, view, condition, grade, sqft_above, sqft_basement, yr_built, yr_renovated, zipcode, latitude, and longitude.



Data Correlation

We took a better look at the data correlation



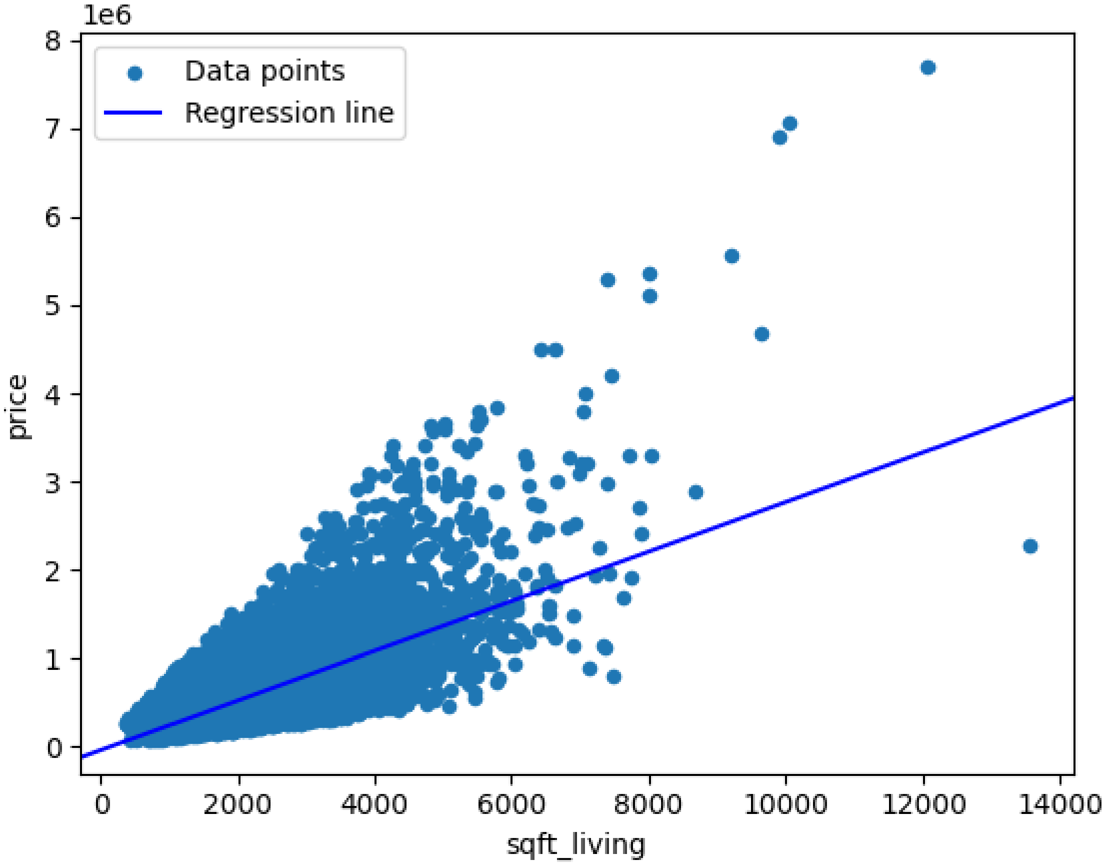
	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	grade	viewed	sqft_above	sqft_basement	lat	sqft_living15
price	1	0.31	0.53	0.7	0.09	0.26	0.26	0.67	0.35	0.61	0.32	0.31	0.59
bedrooms	0.31	1	0.51	0.58	0.032	0.18	-0.0021	0.36	0.079	0.48	0.3	-0.01	0.39
bathrooms	0.53	0.51	1	0.76	0.088	0.5	0.064	0.67	0.17	0.69	0.28	0.024	0.57
sqft_living	0.7	0.58	0.76	1	0.17	0.35	0.1	0.76	0.27	0.88	0.44	0.052	0.76
sqft_lot	0.09	0.032	0.088	0.17	1	-0.0048	0.021	0.11	0.068	0.18	0.015	-0.086	0.14
floors	0.26	0.18	0.5	0.35	-0.0048	1	0.021	0.46	0.016	0.52	-0.25	0.049	0.28
waterfront	0.26	-0.0021	0.064	0.1	0.021	0.021	1	0.083	0.25	0.072	0.083	-0.012	0.084
grade	0.67	0.36	0.67	0.76	0.11	0.46	0.083	1	0.23	0.76	0.17	0.11	0.71
viewed	0.35	0.079	0.17	0.27	0.068	0.016	0.25	0.23	1	0.15	0.27	0.0073	0.27
sqft_above	0.61	0.48	0.69	0.88	0.18	0.52	0.072	0.76	0.15	1	-0.052	-0.0012	0.73
sqft_basement	0.32	0.3	0.28	0.44	0.015	-0.25	0.083	0.17	0.27	-0.052	1	0.11	0.2
lat	0.31	-0.01	0.024	0.052	-0.086	0.049	-0.012	0.11	0.0073	-0.0012	0.11	1	0.049
sqft_living15	0.59	0.39	0.57	0.76	0.14	0.28	0.084	0.71	0.27	0.73	0.2	0.049	1

We focused on these as the top correlation features

price
1
0.53
0.7
0.67
0.61
0.59
1.0

Modelling

We used a simple Linear Regression model between Sqrft Living and price



We also used two models to conduct a multiple linear regression analysis between the price, waterfront, sqrft_living and bathrooms

Results

R-squared (uncentered): The R-squared value of 0.899 indicates that the model explains approximately 89.9% of the variance in the dependent variable (price). This suggests that the independent variables included in the model (bathrooms, sqft_living, waterfront) collectively have a strong association with the price.

Adjusted R-squared (uncentered): The adjusted R-squared value is also 0.899, which means that the inclusion of the three independent variables in the model is not significantly impacting the overall goodness of fit. The adjusted R-squared value is useful for comparing models with different numbers of predictors.

F-statistic: The F-statistic has a very large value of 6.378e+04, and the associated probability (Prob (F-statistic)) is 0.00. This indicates that the overall model is statistically significant, suggesting that at least one of the independent variables has a significant impact on the price.

Coefficients: The coefficients for the independent variables indicate the magnitude and direction of their relationship with the dependent variable (price).

Bathrooms: The coefficient for the "bathrooms" variable is 4.4689, indicating a positive relationship with the price. A one-unit increase in the number of bathrooms is associated with an increase in the price by approximately 4.4689 units.

Sqft_living: The coefficient for the "sqft_living" variable is 0.0011, indicating a positive relationship with the price. A one-unit increase in the square footage of living area is associated with an increase in the price by approximately 0.0011 units.

Waterfront: The coefficient for the "waterfront" variable is -1.3921, indicating a negative relationship with the price. A property with a waterfront location is associated with a decrease in the price by approximately 1.3921 units.

All three variables have p-values close to zero, indicating that they are highly statistically significant in relation to the price.

Notes from the model

- R^2 is computed without centering (uncentered) since the model does not contain a constant.
- Standard Errors assume that the covariance matrix of the errors is correctly specified.
- The condition number is large, 2.79e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Recommendations

The stakeholders should therefore consider large living area in square feet for their house sales, more bathrooms and no waterfront **bold text

---

In [ ]:

```

```