King county house prediction

Business understanding

A Real Estate Agency is project management body tasked with giving advices to homeowners on when they can buy or sell homes. They also cite the relevancy to why it should be so.

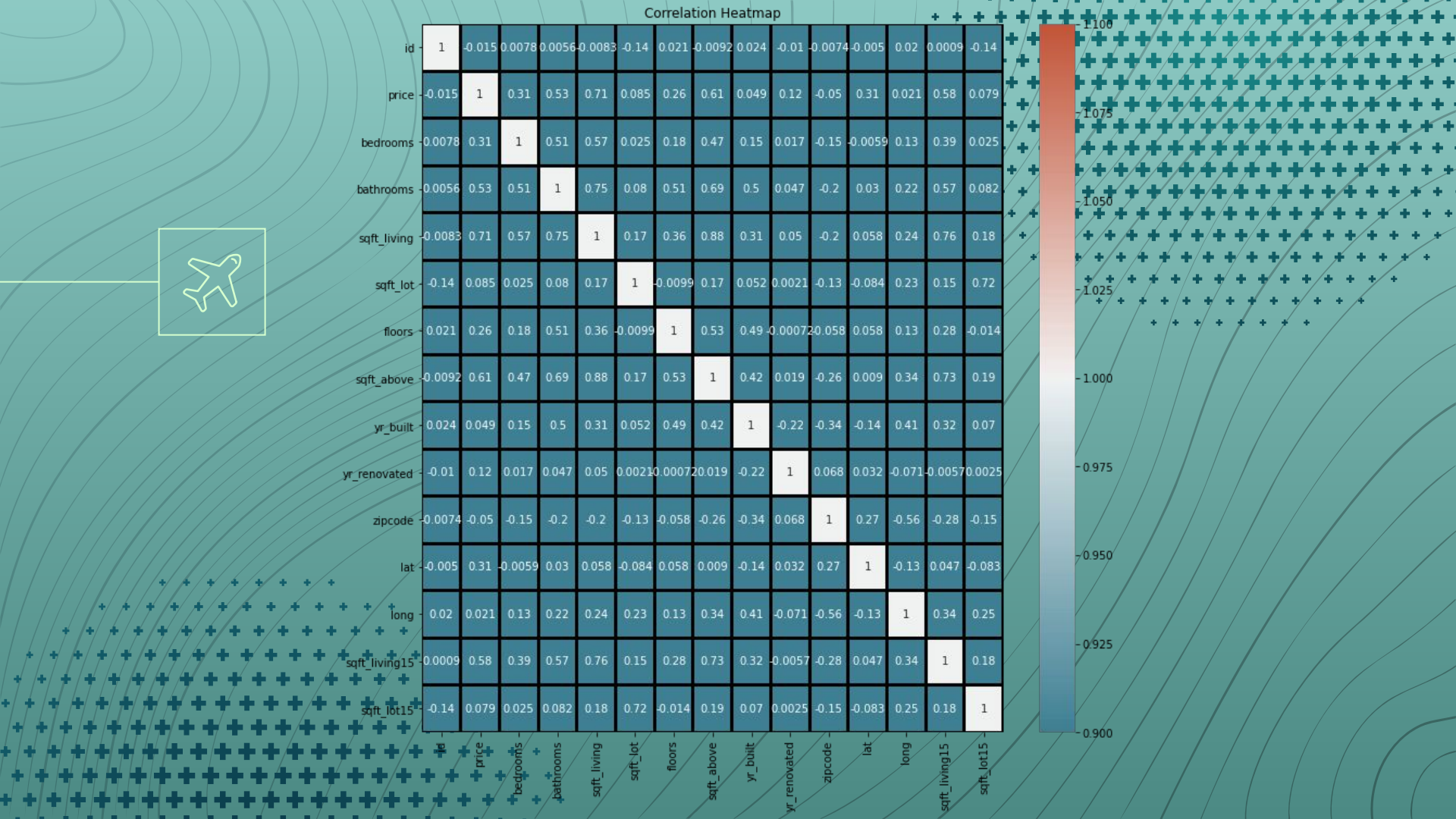
OBJECTIVES

We are therefore intrested in knowing how house renovation would help homeowners to be able to predict the current and future prices of their houses so that they be aware of what best time they can buy or sell the houses.

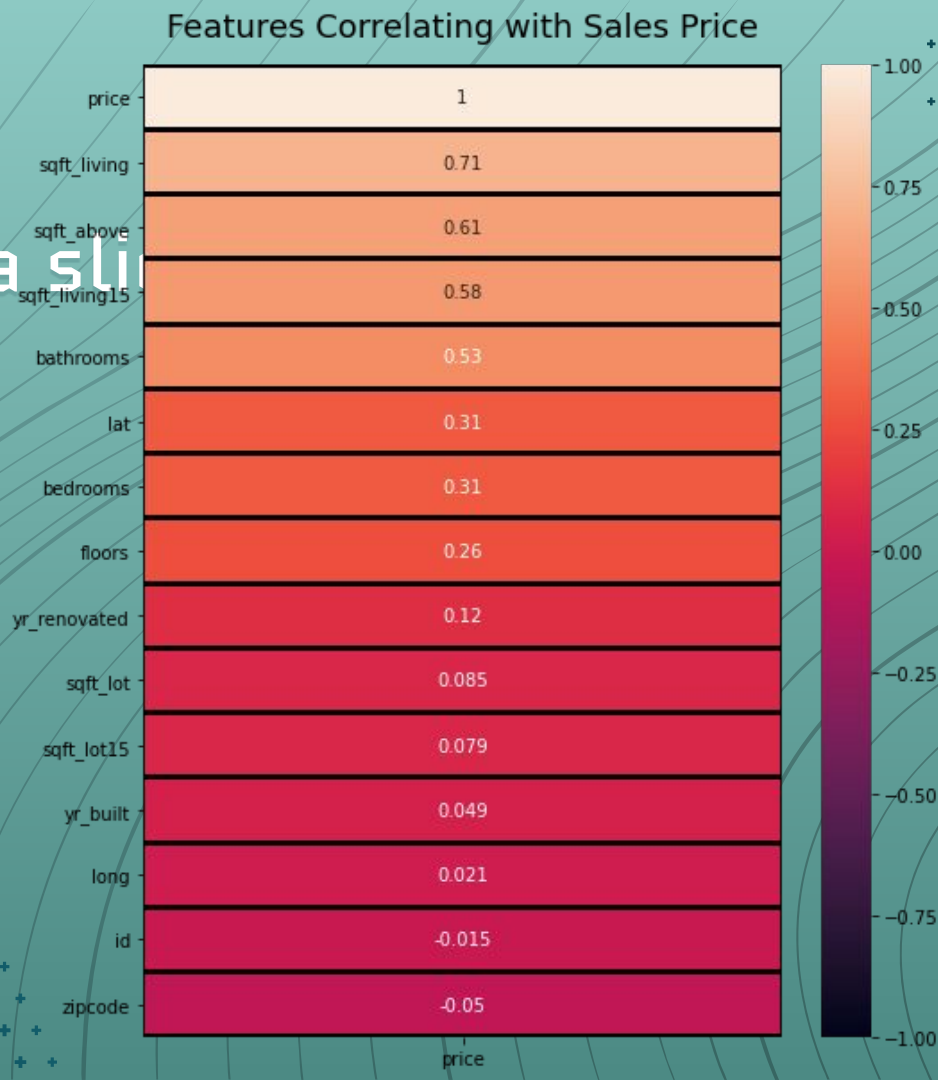
In addition, we want to be able to offer a constructive advice to the stakeholders about the current and future prices of the houses by building or creating a model that will predict the price of these houses.This models will help in coming up with a plan on when it is appropriate to buy or sell.

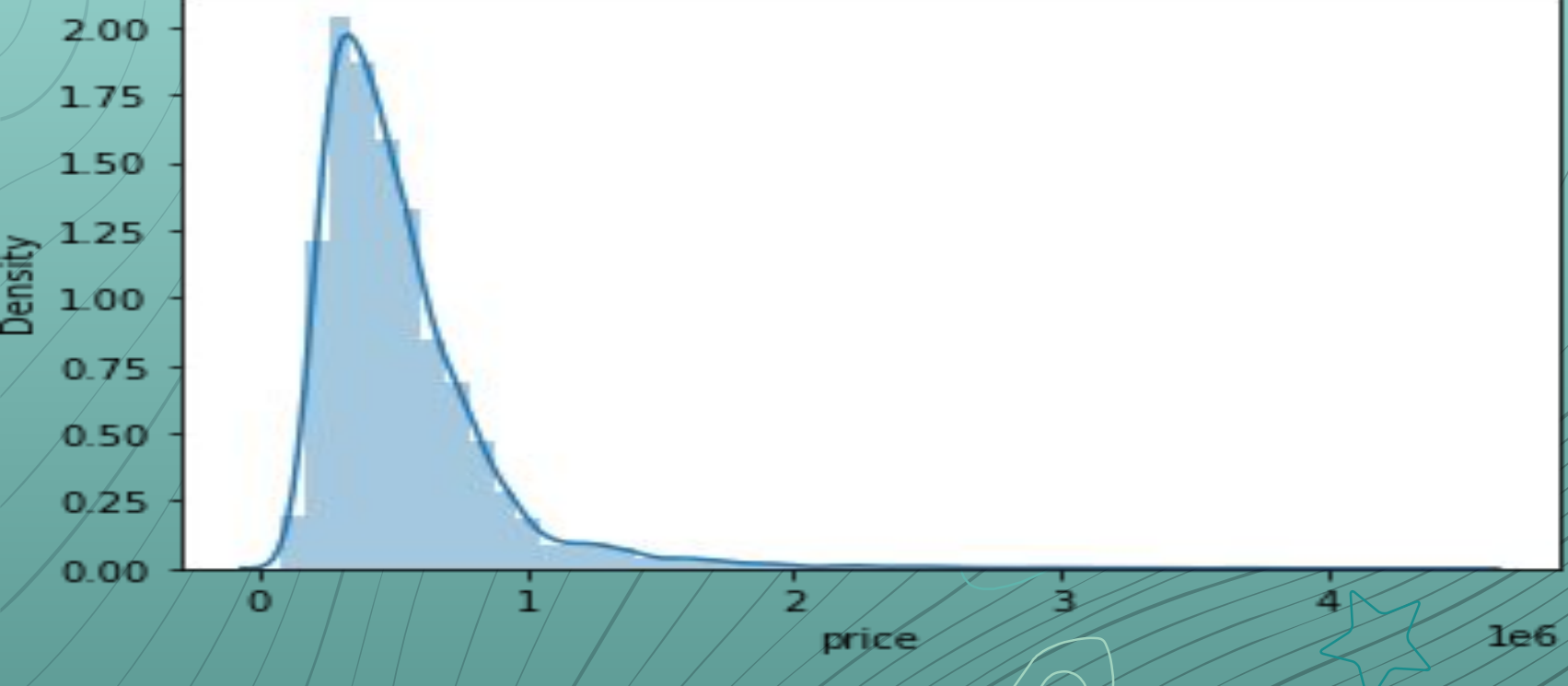
We also want to come up with an analysis so that we know how factors like location, bathrooms and the condition of the houses can be of what influence to the estimated value of the houses and by what amount. This will assist in knowing where to put more strenght so as to achieve the rationale of this project.

heatmap showing correlation between different column pairs


Correlation Heatmap
id 1 -0.015 0.0078 0.0056 -0.0083 -0.14 0.021 -0.0092 0.024 -0.01 -0.0074 -0.005 0.02 0.0009 -0.14
price -0.015 1 0.31 0.53 0.71 0.085 0.26 0.61 0.049 0.12 -0.05 0.31 0.021 0.58 0.079
bedrooms 0.0078 0.31 1 0.51 0.57 0.025 0.18 0.47 0.15 0.017 -0.15 -0.0059 0.13 0.39 0.025
bathrooms 0.0056 0.53 0.51 1 0.75 0.08 0.51 0.69 0.5 0.047 -0.2 0.03 0.22 0.57 0.082
sqft_living -0.0083 0.71 0.57 0.75 1 0.17 0.36 0.88 0.31 0.05 -0.2 0.058 0.24 0.76 0.18
sqft_lot -0.14 0.085 0.025 0.08 0.17 1 -0.0099 0.17 0.052 0.0021 -0.13 -0.084 0.23 0.15 0.72
floors 0.021 0.26 0.18 0.51 0.36 -0.0099 1 0.53 0.49 -0.00072 0.058 0.058 0.13 0.28 -0.014
sqft_above -0.0092 0.61 0.47 0.69 0.88 0.17 0.53 1 0.42 0.019 -0.26 0.009 0.34 0.73 0.19
yr_built 0.024 0.049 0.15 0.5 0.31 0.052 0.49 0.42 1 -0.22 -0.34 -0.14 0.41 0.32 0.07
yr_renovated -0.01 0.12 0.017 0.047 0.05 0.0021 -0.00072 0.019 -0.22 1 0.068 0.032 -0.071 -0.0057 0.0025
zipcode -0.0074 -0.05 -0.15 -0.2 -0.2 -0.13 -0.058 -0.26 -0.34 0.068 1 0.27 -0.56 -0.28 -0.15
lat -0.005 0.31 -0.0059 0.03 0.058 -0.084 0.058 0.009 -0.14 0.032 0.27 1 -0.13 0.047 -0.083
long 0.02 0.021 0.13 0.22 0.24 0.23 0.13 0.34 0.41 -0.071 -0.56 -0.13 1 0.34 0.25
sqft_living15 0.0009 0.58 0.39 0.57 0.76 0.15 0.28 0.73 0.32 -0.0057 -0.28 0.047 0.34 1 0.18
sqft_lot15 -0.14 0.079 0.025 0.082 0.18 0.72 -0.014 0.19 0.07 0.0025 -0.15 -0.083 0.25 0.18 1
id price bedrooms bathrooms sqft_living sqft_lot floors sqft_above yr_built yr_renovated zipcode lat long sqft_living15 sqft_lot15
1.100 1.075 1.050 1.025 1.000 0.975 0.950 0.925 0.900

This is a slu


Features Correlating with Sales Price
price
sqft_living
sqft_above
sqft_living15
bathrooms
lat
bedrooms
floors
yr_renovated
sqft_lot
sqft_lot15
yr_built
long
id
zipcode
1
0.71
0.61
0.58
0.53
0.31
0.31
0.26
0.12
0.085
0.079
0.049
0.021
-0.015
-0.05
price
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00



BIG CONCEPT

Observation from the distribution for price column

1. The plot shows the price column is positively skewed
2. from our plotting is is evident that our data is not normally distributed
3. It also displays a leptokurtic kurtosis

Linear Regression Model Results

- The third model where all the predictor variables were were used explains about 85% of variance in price unlike the previous models that explains about 24%, 24% of variance respectively.
- All the models are statistically significant having a p_value of less than 0.005
- The value of our r_squared constantly increase as we add more predictors.
- The last model would be an ideal in predicting the current and future prices of the house.
- bathrooms , bedrooms, and condition affect the prices of the houses in that an increase in the number of bedrooms , bathrooms, and good condition of the house would lead to an increase in price of that house.

Limitations of the model

- The model perform poorly with the variables that are not linearl because linear regression assumes a linear relationship between the input and output variables

Recommendation

- I would recommend that the Agency adopt the last model since it has a higher r_squared value of 0.85 which translate to 85% accuracy of the model.
- Features such as condition , bedrooms, bathrooms, and sqft_living of the house should be put into consideration because they affect prices positively .
- By using this model the Agency is able to offer reasonable advices to the homeowners which will inturn help them in knowing how to maximize profits and minimize losses.
- I would also recommend the Agency that they advice the homeowners to improve features such as waterfront, view as they have a greater impact to the house prices as well