Inspecting the KC Data set

Khalid Gharib
Jamie

Background

We are a Data scientist duo who has been approached by a real estate agent asking us to find out how best for them to use their resources most effectively.

Cleaning the Dataset

missing info in columns 'waterfront', 'yr_renovated' and a few in 'view’

‘?’ in the sqft_basement

Changed dtype of Date column

ways I have cleaned was by replacing Null values with the mode or median of those columns and in some cases changing dtypes

Further inspection of the Data

Looked at the correlation between different variables.

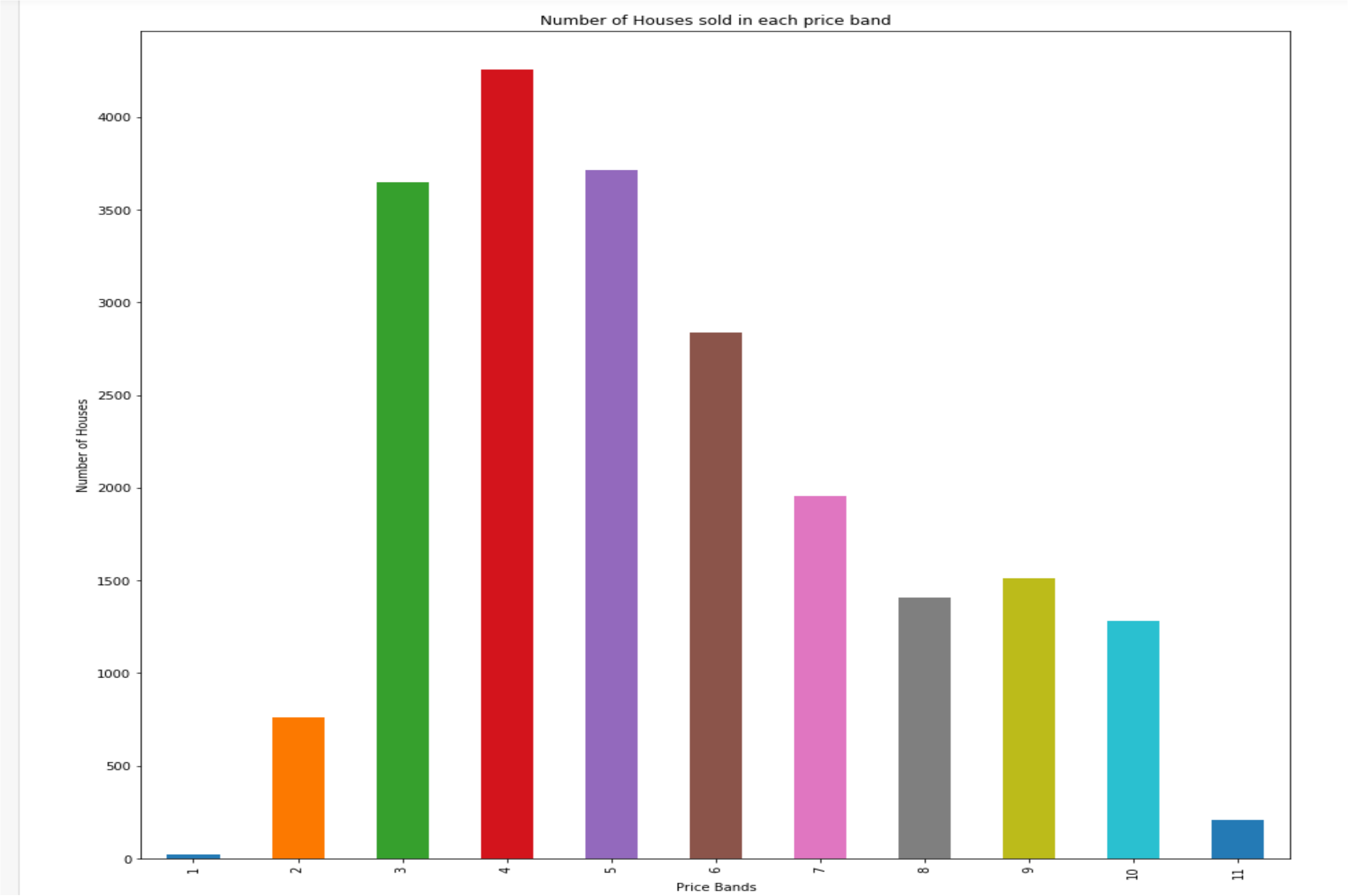
Also inspected the descriptive statistics of the price column.

Looked at other variables and what correlation they have such as do bigger houses have waterfronts or does houses with a higher grade/condition tend to be houses built in more recent years or refurbished?

Question 1: relation between house sales and Zip codes

- Firstly I looked at the price distribution plot to see where the prices of houses are more focused.
- Created evenly distributed price bands based on the dist plot
- Put each house in a price band based on the price it was sold at.
- The final outcome was that most houses were being sold between price band 3-6(200k-599k dollars
- we can see: around 66.91% of houses sold fall in that price band(14450 houses)

Count of houses sold in each price band



Question 2: which location/zip code is best and most suitable to focus on.

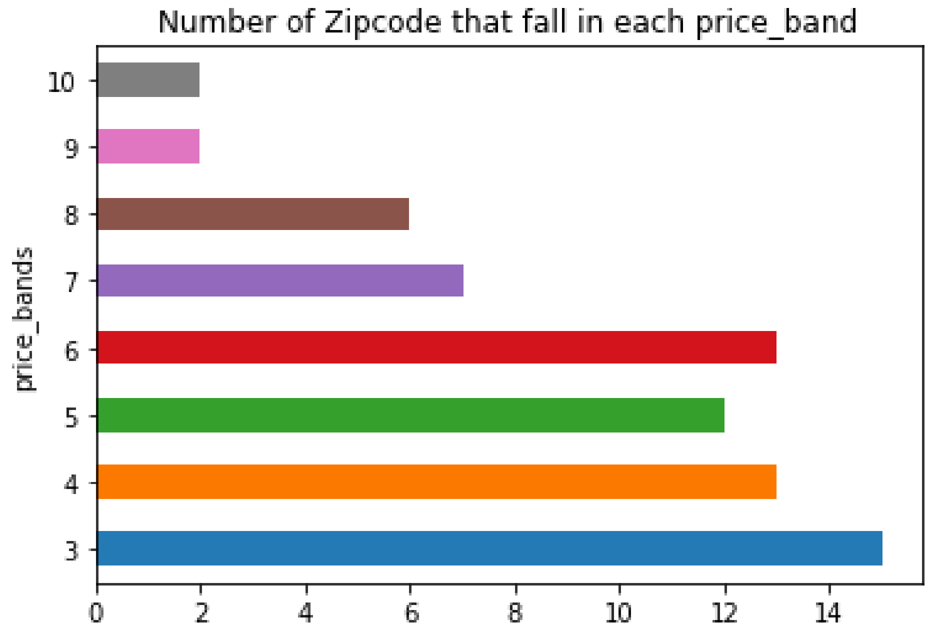
Investigated how many houses are sold in each zip code

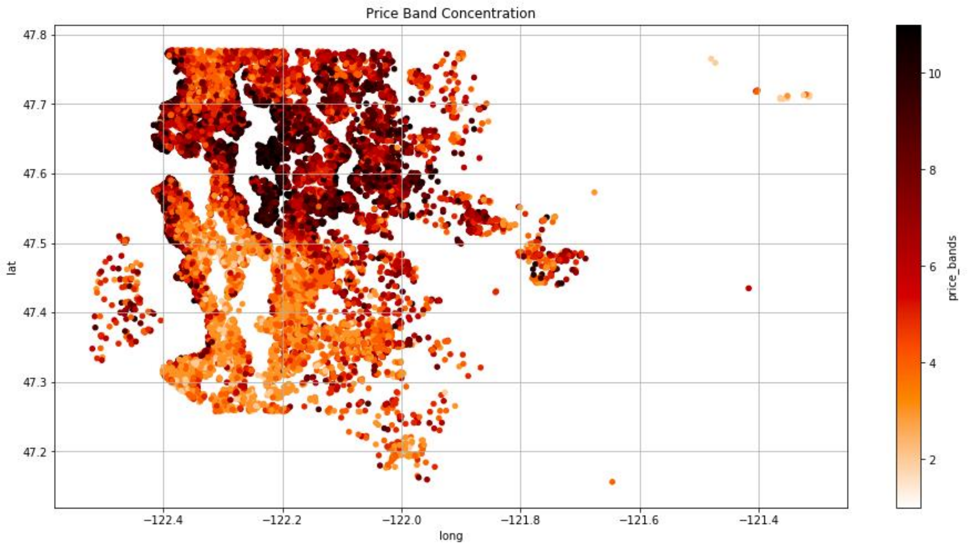
we will see the distribution of the price and price_band of houses based on location(zip code)

Include descriptive statistics of each zip code

Place each zip code into a price band based on it median

suggest which specific zip code based on that price band are most suitable


Number of Zipcode that fall in each price_band
price_bands
10
9
8
7
6
5
4
3
0
2
4
6
8
10
12
14

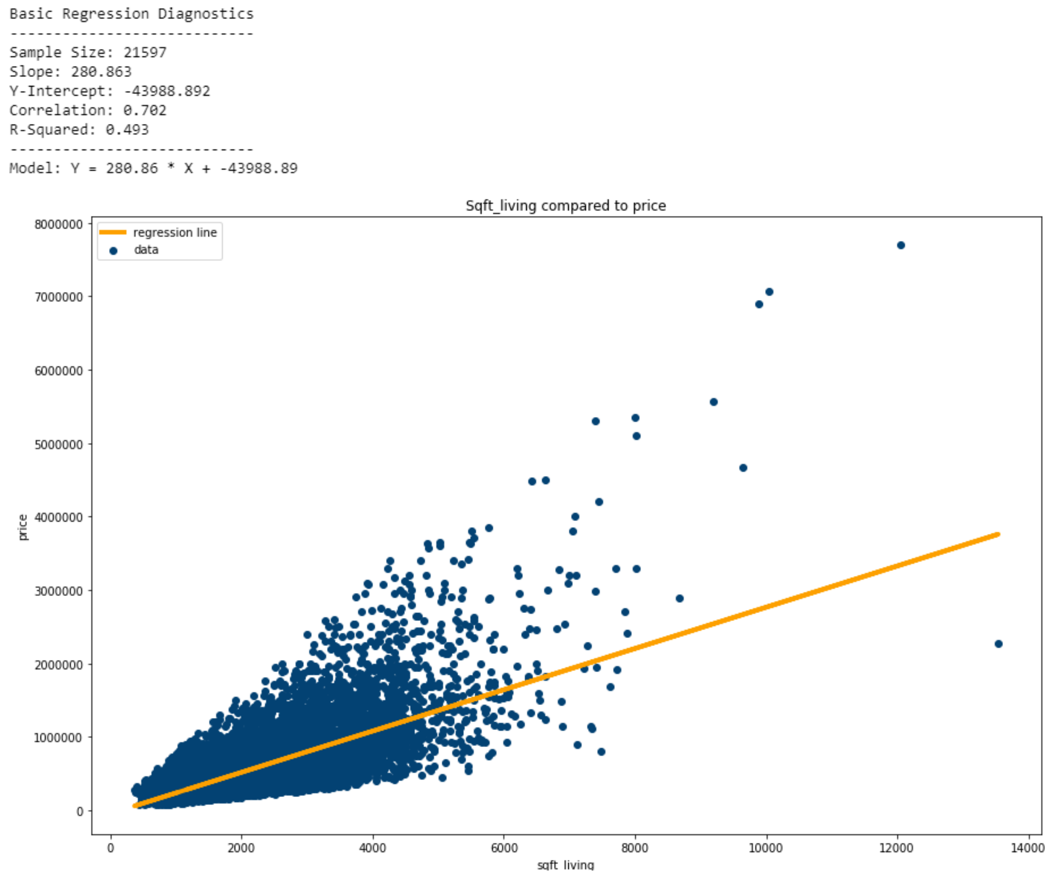

Price Band Concentration
lat
47.8
47.7
47.6
47.5
47.4
47.3
47.2
long
−122.4
−122.2
−122.0
−121.8
−121.6
−121.4
price_bands
10
8
6
4
2

Question 3: What is the relation between certain variables such as sq footage of the house and the price?

- The main 5 variables I analysed were:
  - sqft_living vs price
  - sqft_lot vs price
  - grade vs price
  - sqft_living vs grade
  - yr_built vs grade

Sqft_living vs price

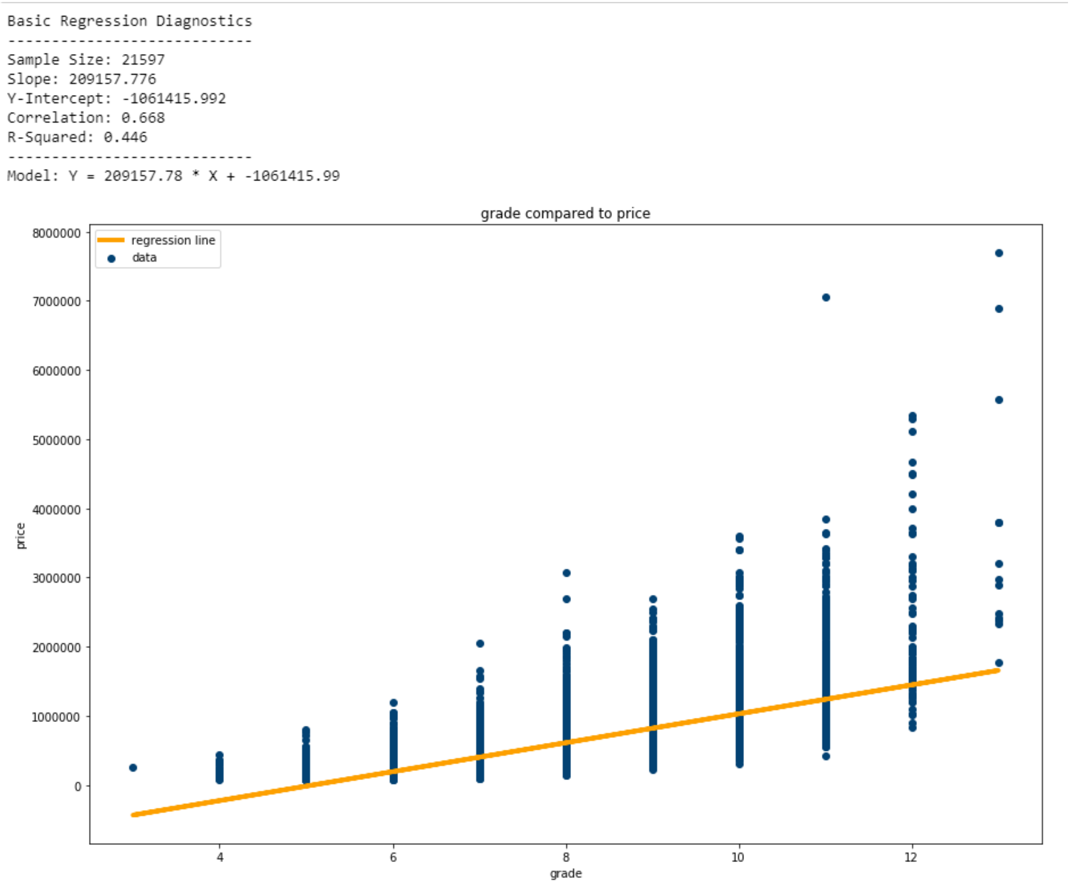
- This was an important regression model to show how price is effected by sqft_living
- In this case the regression model will help them to know what price range a certain sqft_living of a house will make it fall into and thus be able to firstly put accurate prices to future houses they look to buy or sell.



```
Basic Regression Diagnostics
----------------------------
Sample Size: 21597
Slope: 280.863
Y-Intercept: -43988.892
Correlation: 0.702
R-Squared: 0.493
----------------------------
Model: Y = 280.86 * X + -43988.89
```

Sqft_living compared to price

Grade vs price



```
Basic Regression Diagnostics
----------------------------
Sample Size: 21597
Slope: 209157.776
Y-Intercept: -1061415.992
Correlation: 0.668
R-Squared: 0.446
----------------------------
Model: Y = 209157.78 * X + -1061415.99
```

Findings & comments

i would respond to the real-estate agency that they should try to focus on the price range market between 200,000 to 600,000 Dollars as this is where around 70% of the houses in the data frame were sold.

we analyzed which zip codes actually fall in that price range based on its median so that we can tell them to help them specifically target certain zip codes.

there is also a very clear correlation between grade and price. Showing us that the higher grades/ or houses that have done custom works and gotten a grade 13 do sell for higher

we can see that this price band we suggest tend to be more towards the south of Seattle

finally we found out in our findings in the regression models that we did, is that there is a correlation between prices and sqft_living,