Final Report:

King Country Housing Analysis

Problem Statement:

House prices have been changing rapidly over the past ten years and an accurate model to predict or even understand the changes of price of a home based on location, bedroom quantity, and square footage of the home and the differences between the effects of the world wide pandemic and its effect on income of individuals and families will most inevitably lead to a full blown public housing crisis. But is it possible to better understand the trends of the housing in a single location to give a more clear model to understand the features of other locations in the country or even the world? Is it possible to understand these features of King County to minimize the crisis in other locations?

By using the King Country Housing data, I created a few models to look at the different features that have traditionally affected the housing market and been instrumental in predicting the future prices for homes.

Data Wrangling:

The raw dataset from House Sales in King County, USA contained 21,613 rows with 21 columns. It was a very tidy dataset and although it did require some size reduction it was only done so with removing a few columns that would not be needed for model production.

Had there been more parameters to consider, reducing the dimensionality would have been a more critical step if I had desired to see more clear insights from the dataset, but this dataset did not contain many null values and only a couple columns contained non-significant information.

The final shape of my dataset was 21,613 rows with 14 columns.

Exploratory Data Analysis:

The dataset columns are as follows:

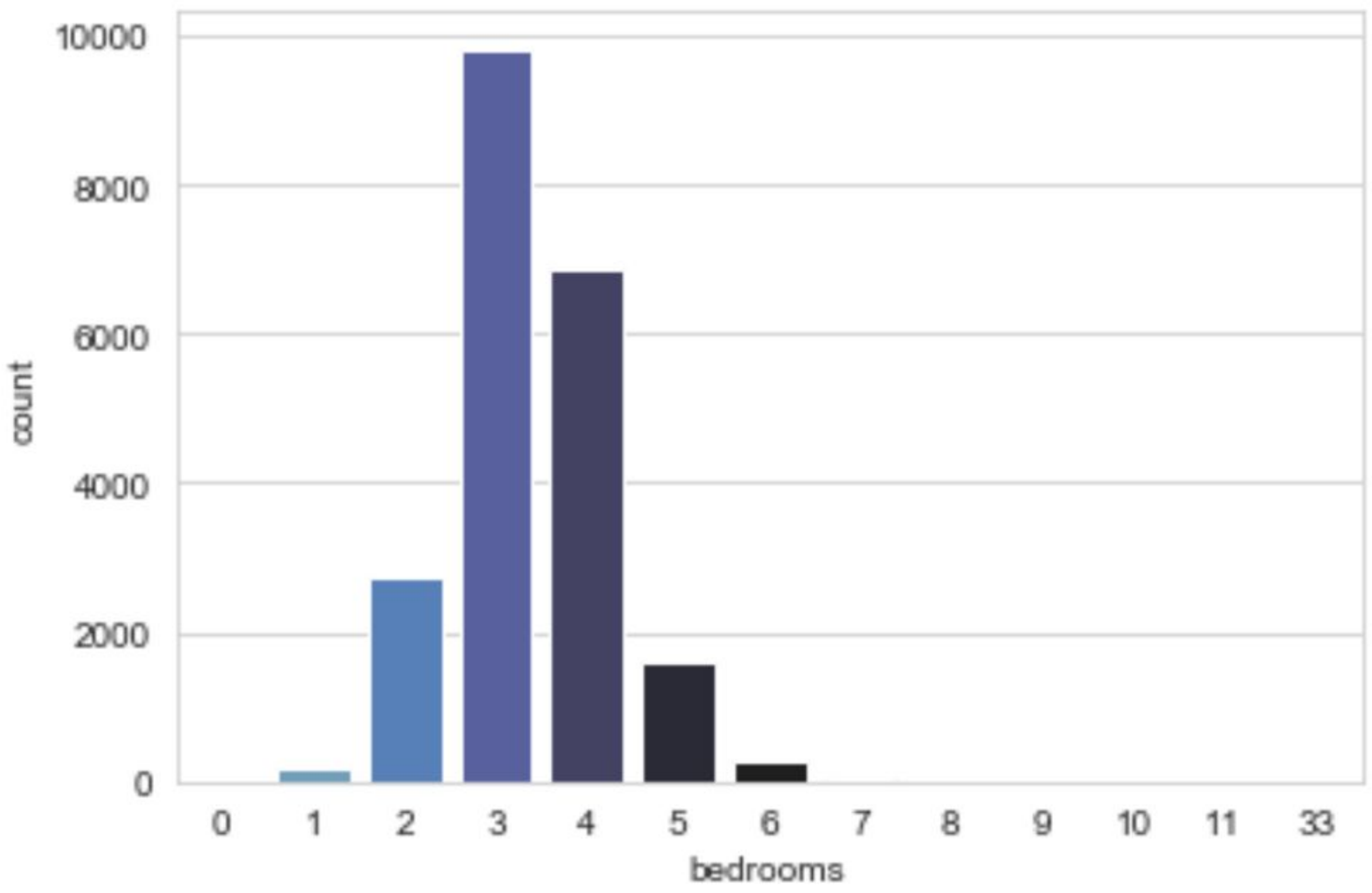
- id :a notation for a house
- date: Date house was sold
- price: Price is prediction target
- bedrooms: Number of Bedrooms/House
- bathrooms: Number of bathrooms/bedrooms
- sqft_living: square footage of the home
- sqft_lot: square footage of the lot
- floors :Total floors (levels) in house
- waterfront :House which has a view to a waterfront
- view: Has been viewed
- condition :How good the condition is Overall
- grade: overall grade given to the housing unit, based on King County grading system
- sqft_above :square footage of house apart from basement
- sqft_basement: square footage of the basement
- yr_built :Built Year
- yr_renovated :Year when house was renovated
- zipcode:zip code
- lat: Latitude coordinate
- long: Longitude coordinate
- sqft_living15 :Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area
- sqft_lot15 :lotSize area in 2015(implies-- some renovations)

	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade
count	2.161300e+04	21613.000000	21613.000000	21613.000000	2.161300e+04	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000
mean	5.400881e+05	3.370842	2.114757	2079.899736	1.510697e+04	1.494309	0.007542	0.234303	3.409430	7.656873
std	3.671272e+05	0.930062	0.770163	918.440897	4.142051e+04	0.539989	0.086517	0.766318	0.650743	1.175459
min	7.500000e+04	0.000000	0.000000	290.000000	5.200000e+02	1.000000	0.000000	0.000000	1.000000	1.000000
25%	3.219500e+05	3.000000	1.750000	1427.000000	5.040000e+03	1.000000	0.000000	0.000000	3.000000	7.000000
50%	4.500000e+05	3.000000	2.250000	1910.000000	7.618000e+03	1.500000	0.000000	0.000000	3.000000	7.000000
75%	6.450000e+05	4.000000	2.500000	2550.000000	1.068800e+04	2.000000	0.000000	0.000000	4.000000	8.000000
max	7.700000e+06	33.000000	8.000000	13540.000000	1.651359e+06	3.500000	1.000000	4.000000	5.000000	13.000000

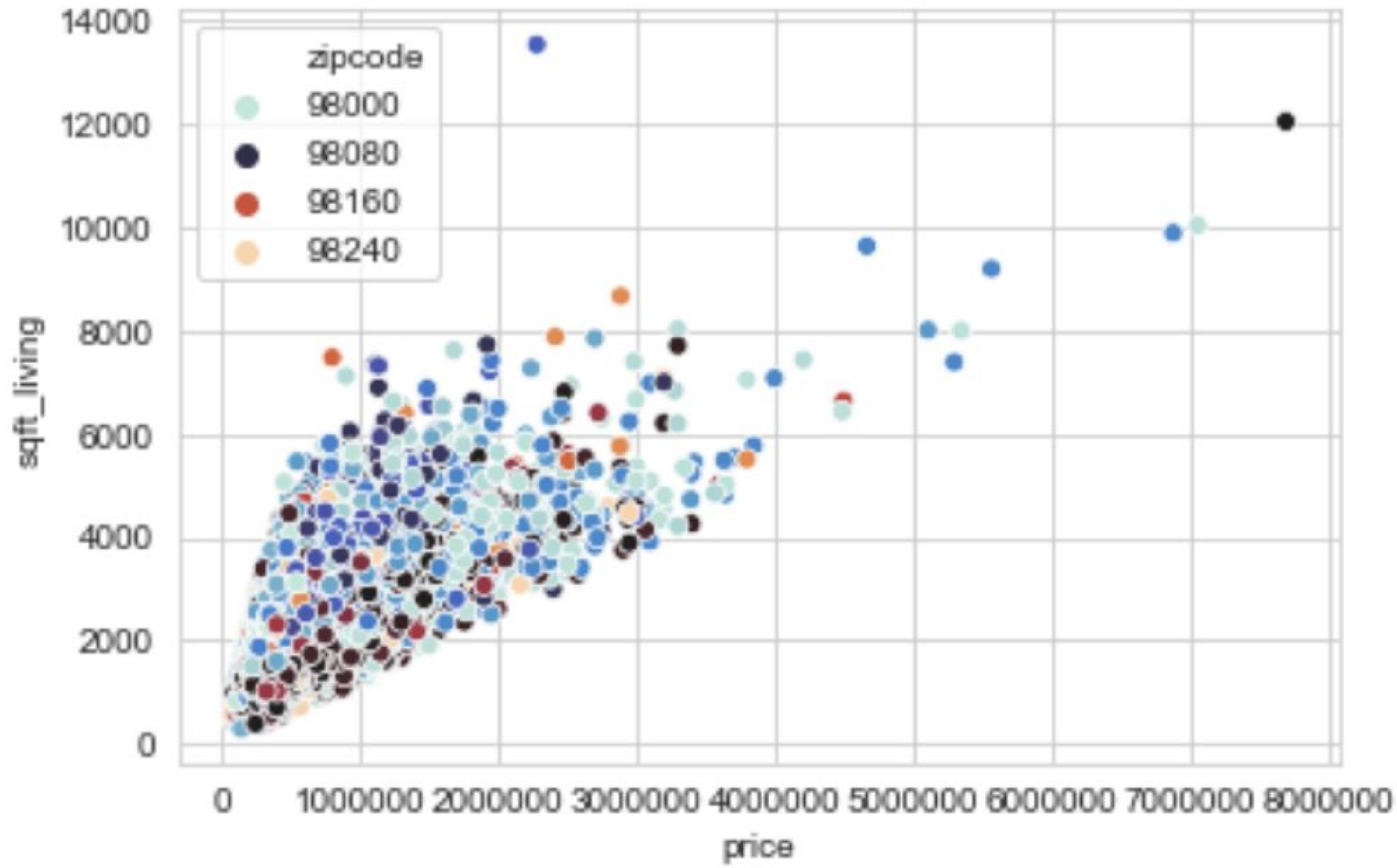
grade	sqft_above	sqft_basement	sqft_living15	sqft_lot15
21613.000000	21613.000000	21613.000000	21613.000000	21613.000000
7.656873	1788.390691	291.509045	1986.552492	12768.455652
1.175459	828.090978	442.575043	685.391304	27304.179631
1.000000	290.000000	0.000000	399.000000	651.000000
7.000000	1190.000000	0.000000	1490.000000	5100.000000
7.000000	1560.000000	0.000000	1840.000000	7620.000000
8.000000	2210.000000	560.000000	2360.000000	10083.000000
13.000000	9410.000000	4820.000000	6210.000000	871200.000000

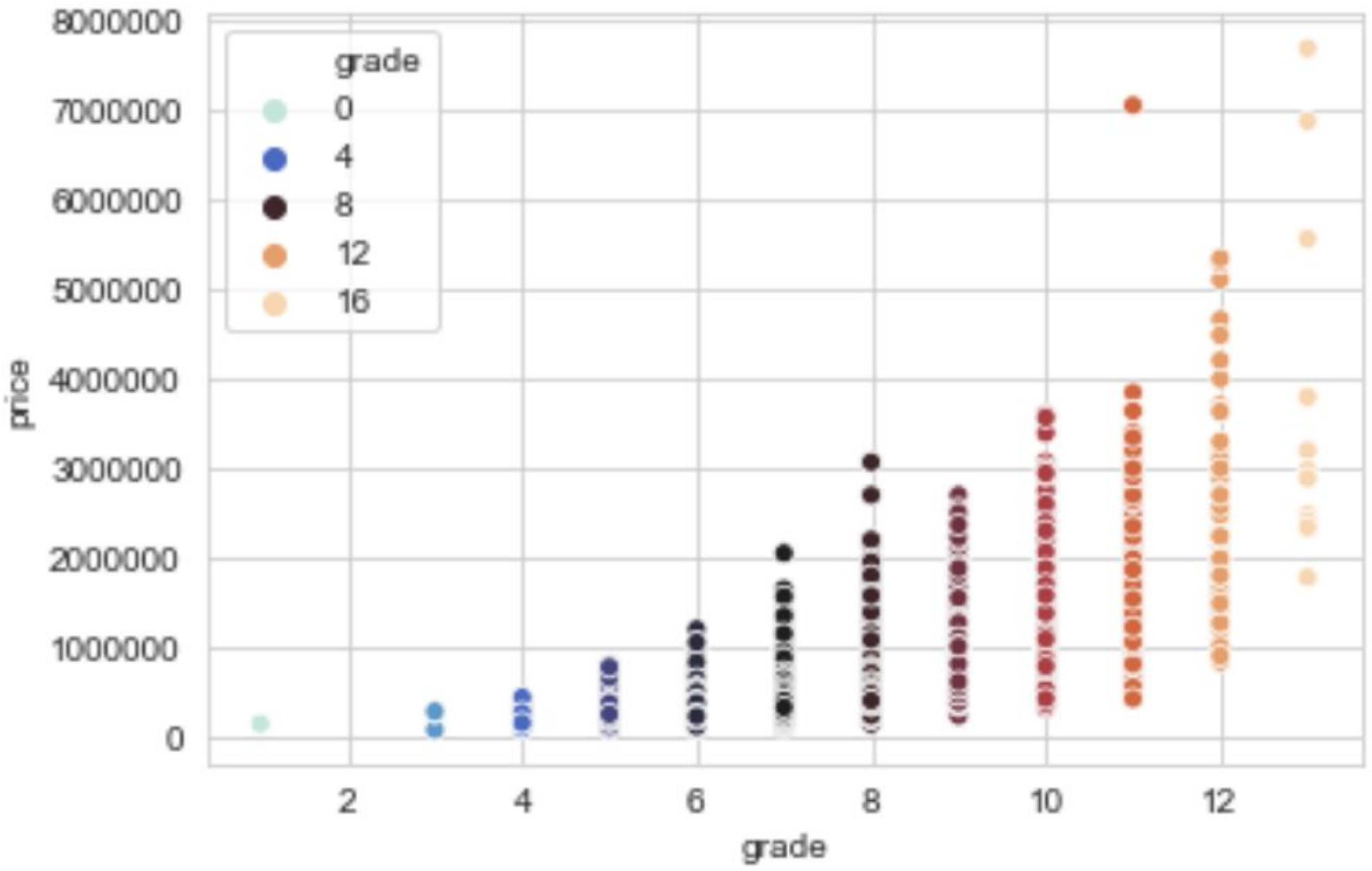
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21613 entries, 0 to 21612
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   id             21613 non-null  int64  
 1   date           21613 non-null  object 
 2   price          21613 non-null  float64
 3   bedrooms       21613 non-null  int64  
 4   bathrooms      21613 non-null  float64
 5   sqft_living    21613 non-null  int64  
 6   sqft_lot       21613 non-null  int64  
 7   floors         21613 non-null  float64
 8   waterfront     21613 non-null  int64  
 9   view           21613 non-null  int64  
 10  condition      21613 non-null  int64  
 11  grade          21613 non-null  int64  
 12  sqft_above     21613 non-null  int64  
 13  sqft_basement  21613 non-null  int64  
 14  yr_built       21613 non-null  int64  
 15  yr_renovated   21613 non-null  int64  
 16  zipcode        21613 non-null  int64  
 17  lat            21613 non-null  float64
 18  long           21613 non-null  float64
 19  sqft_living15  21613 non-null  int64  
 20  sqft_lot15     21613 non-null  int64  
dtypes: float64(5), int64(15), object(1)
memory usage: 3.5+ MB
```



10000
8000
6000
4000
2000
0
count
0
1
2
3
4
5
6
7
8
9
10
11
33
bedrooms


zipcode
98000
98080
98160
98240
sqft_living
price
0
2000
4000
6000
8000
10000
12000
14000
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
8000000

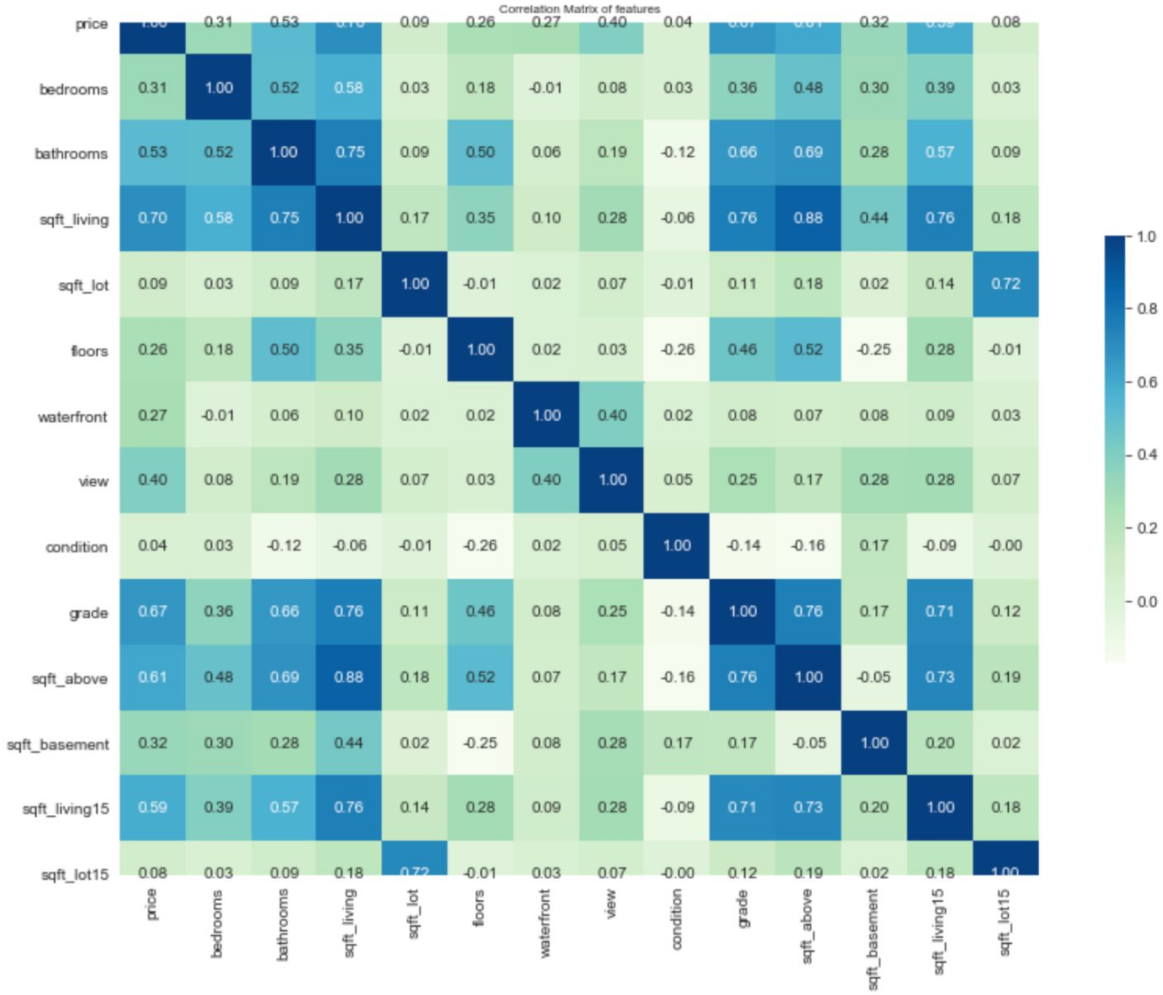


grade
0
4
8
12
16
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
price
2
4
6
8
10
12
grade

In-Depth Analysis

Now that a we have a better understanding of the data and how what the dataset consists of we need to determine some of the features to model around:


SqFt Living vs Price (Test set)
SqFt Living
Price
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
0
2000
4000
6000
8000
10000
12000
14000


Correlation Matrix of features
price
bedrooms
bathrooms
sqft_living
sqft_lot
floors
waterfront
view
condition
grade
sqft_above
sqft_basement
sqft_living15
sqft_lot15
1.0
0.8
0.6
0.4
0.2
0.0

Heatmap:

The Heatmap is showing the different strengths of overlap between all of the different features of the dataset. It is clear to see that we will not be using the different years of columns as they do not factor as viable features in the different models but will simply be used to help determine the overal usage in the future and help with more questions as we move forward with implementing the models into other locations at other times.

This heatmap has shown to highlight the value of price in relation to bedrooms, bathrooms, square foot of living space, and the grade of the home.

Model Selection:

I tested 3 different machine learning classification models: XGBoost, Linear Regression, and Random Forest.

XGBoost:

Varianace score (Best possible score is 1.0, lower values are worse.) : 0.999636475208102

rmse : 7379.707111785109

R2 (Best possible score is 1.0): 0.9996362814561216

Linear Regression: R2 (Best possible score is 1.0): 0.27166936680438114

Random Forest: R2 (Best possible score is 1.0): 0.5000063123558642

Takeaways:

Without doubt, the XGBoost is the best model according to the R2 score.

This project has given me a lot to consider as I move to improve the way I take datasets and model the features to better understand and explain what is going on with the area of study. I would like to be able to better understand the different ways I can take the individual features and model them against each other to see the different trends and create different models to learn how they relate.

There is a lot of different locations that similar study can be conducted and as we work with the effects of the world wide pandemic there will be new features that will affect the values of homes.