Case Study: Regression

Dooin Kim, Guilherme, Sabina

TABLE OF CONTENTS

INTRODUCTION

22000

Dataset containing information on 22,000 properties

SOLD 2014-2015

sold between May 2014 and May 2015

2.OBJECTIVES OF BUSINESS CASE

Understand and perform the necessary EDA steps

build a ML Model that can accurately predict the selling prices

To identify the factors that influence the selling price

3. DATASET OVERVIEW AND METHODOLOGY

ID

DATE

WATERFRONT

CONDITION

GRADE

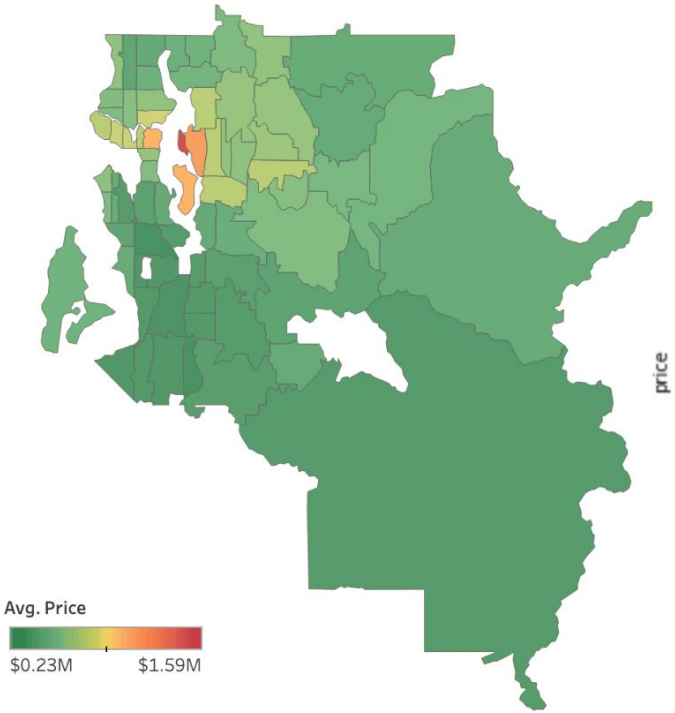
'SQFT_ABOVE'

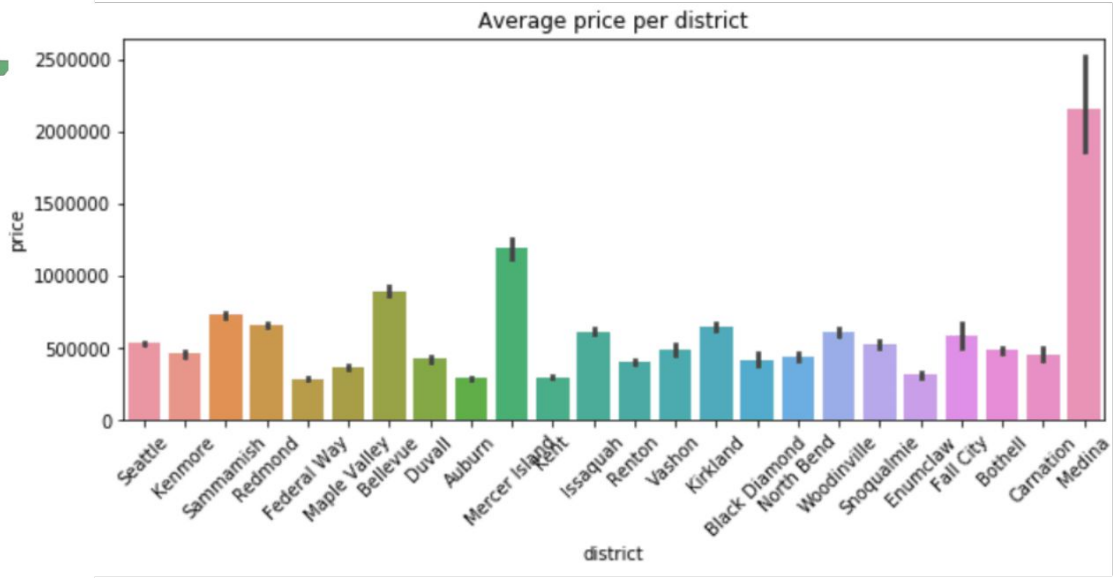
'SQFT_LIVING 15'

'SQFT_LOT15'

LIBRARIES	PANDAS	MATPLOTLIB	SEABORN	SCIKIT.LEARN
DATA CLEANING	DEALING WITH NULL VALUES	DROP COLUMNS	HANDLING OUTLIERS	
EDA	MY SQL	PYTHON	TABLEAU	
DATA MODELLING	PREDICTION MODELS	MODEL VALIDATION	MODEL IMPROOVEMENT	

4.1 DISTRICT

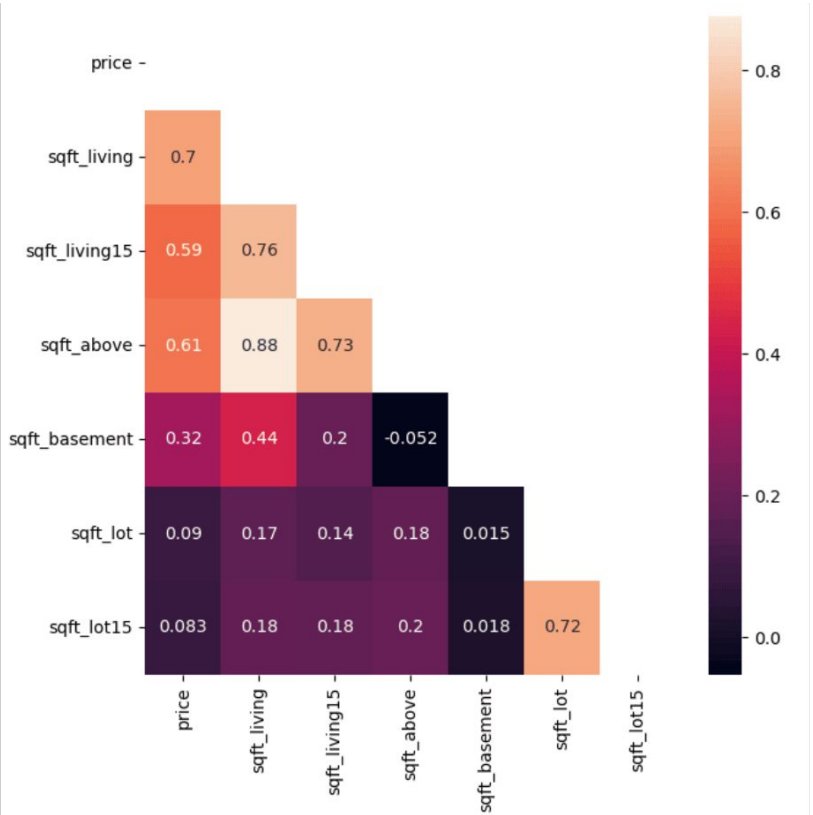

Avg. Price
$0.23M
$1.59M

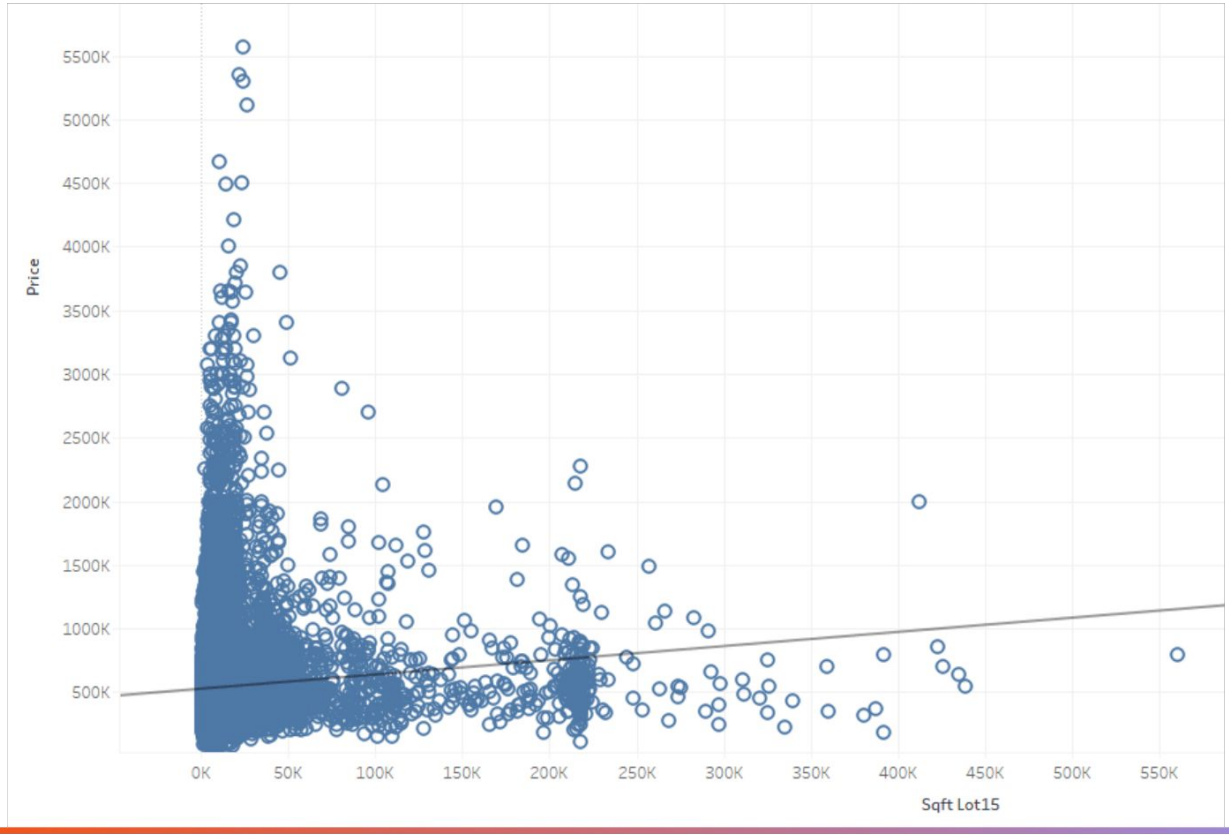

Average price per district
price
2500000
2000000
1500000
1000000
500000
0
Seattle
Kenmore
Sammamish
Redmond
Federal Way
Maple Valley
Bellevue
Duvall
Auburn
Mercer Island
Kent
Issaquah
Renton
Vashon
Kirkland
Black Diamond
North Bend
Woodinville
Snoqualmie
Enumclaw
Fall City
Bothell
Carnation
Medina
district

4.2 SIZE

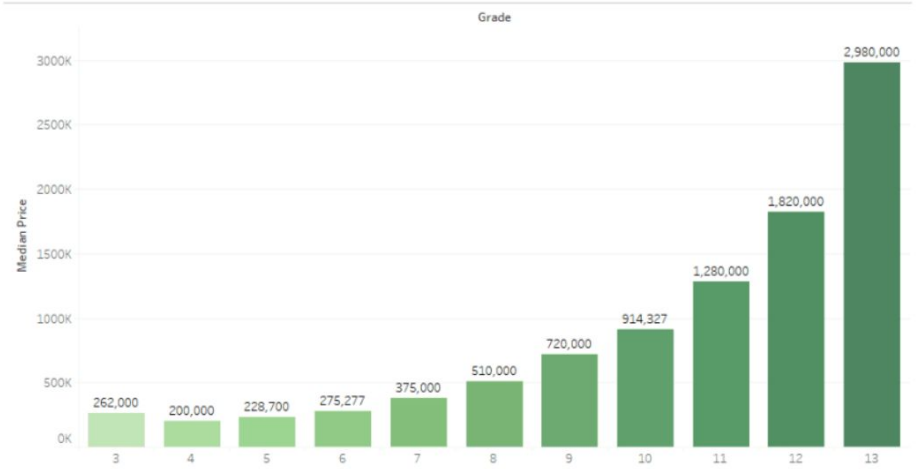
```
corr_matrix["price"].sort_values(ascending=False)
```

```
price            1.000000
sqft_living      0.701917
grade            0.667951
sqft_above       0.605368
sqft_living15    0.585241
bathrooms        0.525906
view             0.397370
sqft_basement    0.323799
bedrooms         0.308787
lat              0.306692
waterfront       0.266398
floors           0.256804
yr_renovated     0.126424
sqft_lot         0.089876
sqft_lot15       0.082845
yr_built         0.053953
condition        0.036056
long             0.022036
zipcode         -0.053402
Name: price, dtype: float64
```



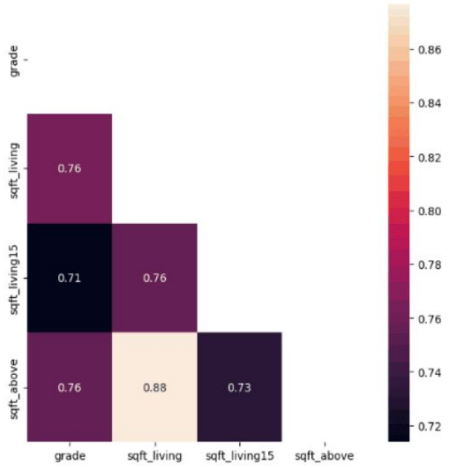

Price
5500K
5000K
4500K
4000K
3500K
3000K
2500K
2000K
1500K
1000K
500K
0K
50K
100K
150K
200K
250K
300K
350K
400K
450K
500K
550K
Sqft Lot15

4.3 GRADE

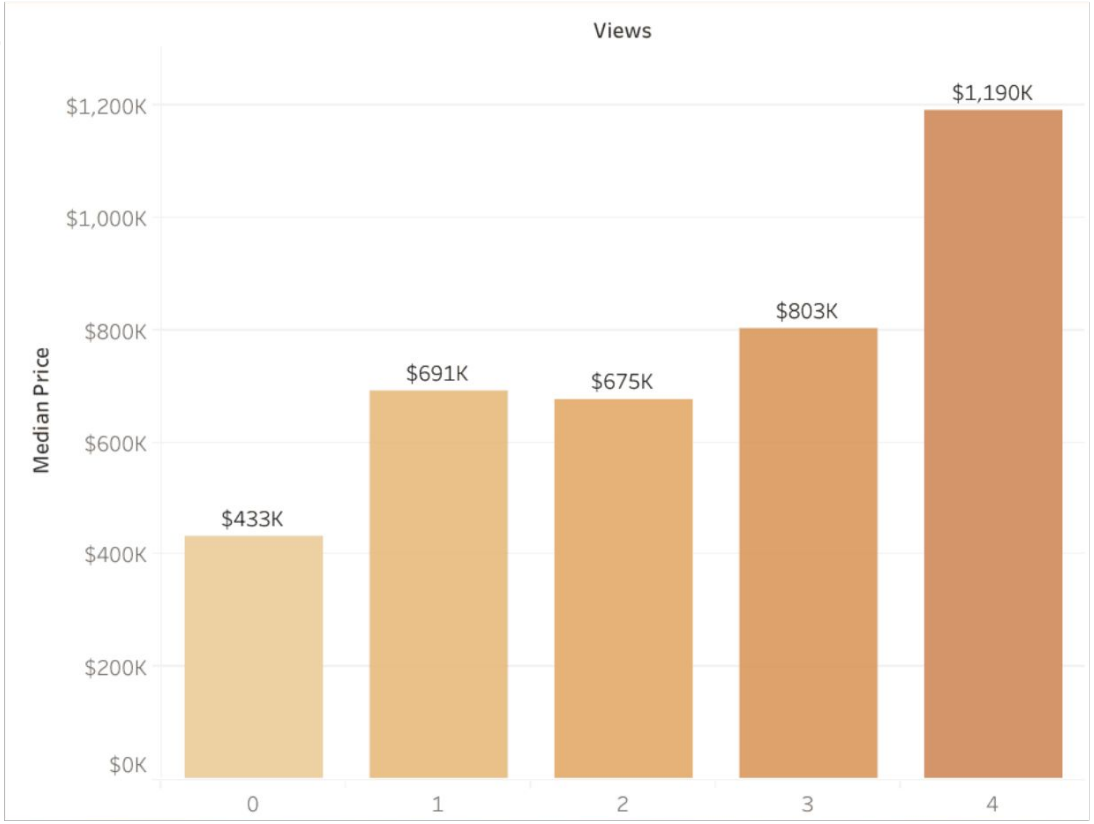



```
corr_matrix["price"].sort_values(ascending=False)
```

```
price            1.000000
sqft_living      0.701917
grade            0.667951
sqft_above       0.605368
sqft_living15    0.585241
bathrooms        0.525906
view             0.397370
sqft_basement    0.323799
bedrooms         0.308787
lat              0.306692
waterfront       0.266398
floors           0.256804
yr_renovated     0.126424
sqft_lot         0.089876
sqft_lot15       0.082845
yr_built         0.053953
condition        0.036056
long             0.022036
zipcode         -0.053402
Name: price, dtype: float64
```



4.4 VIEWS


Views
Median Price
$1,200K
$1,000K
$800K
$600K
$400K
$200K
$0K
$433K
$691K
$675K
$803K
$1,190K
0
1
2
3
4

5. HOUSE PRICE PREDICTION MODEL

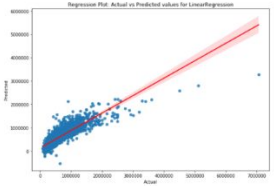
- Employed Prediction Models: Linear Regressor, KNN Regressor & Random Forest Regressor.

- Model Validation: R2 Score, MAE, RMSE

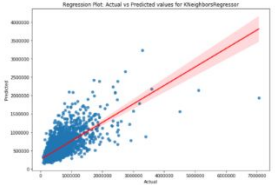
- Model Improvement: Scaling (Log Transform - to reduce outliers), Feature Selection (Avoid Multicollinearity using correlation matrix)

Cross- Examination of Different Algorithms

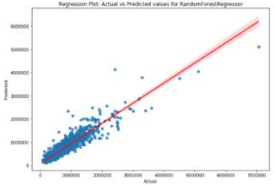
Before: Baseline Model



Linear Regressor
- **R2:** 0.74
- **RMSE:** $183,274
- **MAE:**$112,183

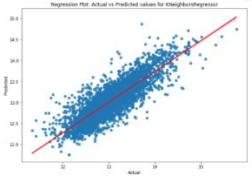


KNeighbor Regressor
- **R2:** 0.49
- **RMSE**:$255,644
- **MAE:** $156,184

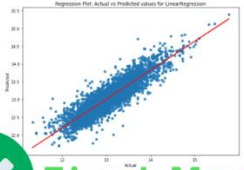


Random Forest Regressor
- **R2:**0.88
- **RMSE:**$123,522
- **MAE:** $67,809

After: Improved Model

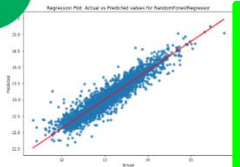


Linear Regressor
- **R2:**0.80
- **RMSE:**$159,856
- **MAE:** $94,731



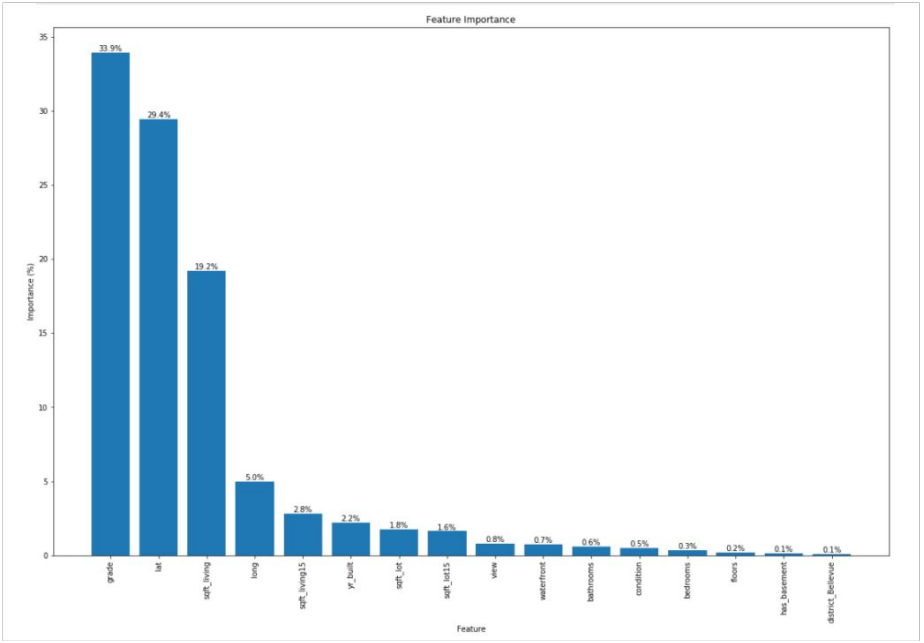
KNeighbor Regressor
- **R2:**0.73
- **RMSE:**$205,497
- **MAE:** $108,032

Final Model Selection



Random Forest Regressor
- **R2:**0.88
- **RMSE:**$133,140
- **MAE:** $68,544

Feature Importance

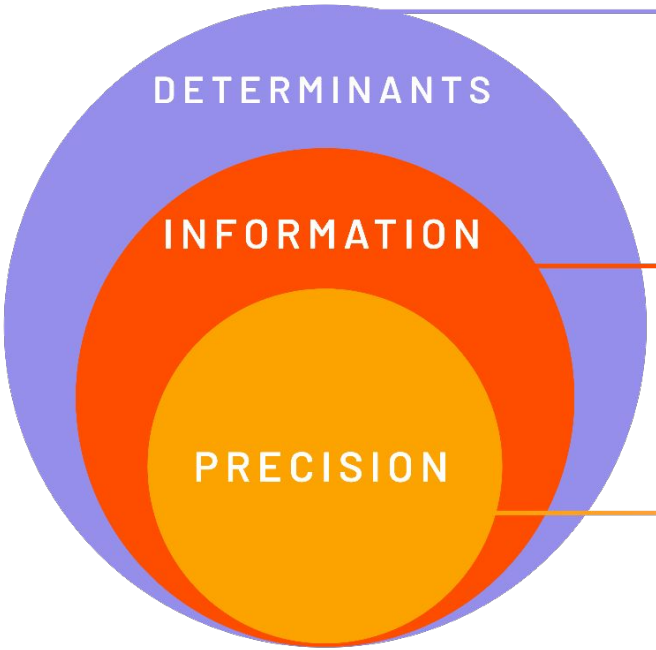


Important house price factors:

- Grade
- Location (lat & long)
- Size (sqft_living & sqft_living15, sqft_lot)
- View
- Waterfront

Bathrooms, Bedrooms, Condition, Floors, Basement

6. CONCLUSIONS



Our Forest Random Regression analysis indicates that property grade, location, and size are key price determinants.

Yet, we lack specific information about the grading system and our dataset omits influential external factors such as proximity to amenities and crime rates.

To increase the precision of our housing price estimation model, we recommend incorporating these social-economic indicators into further analyses.