Capstone Presentation

Washington House Price Prediction

By Anoop Raut

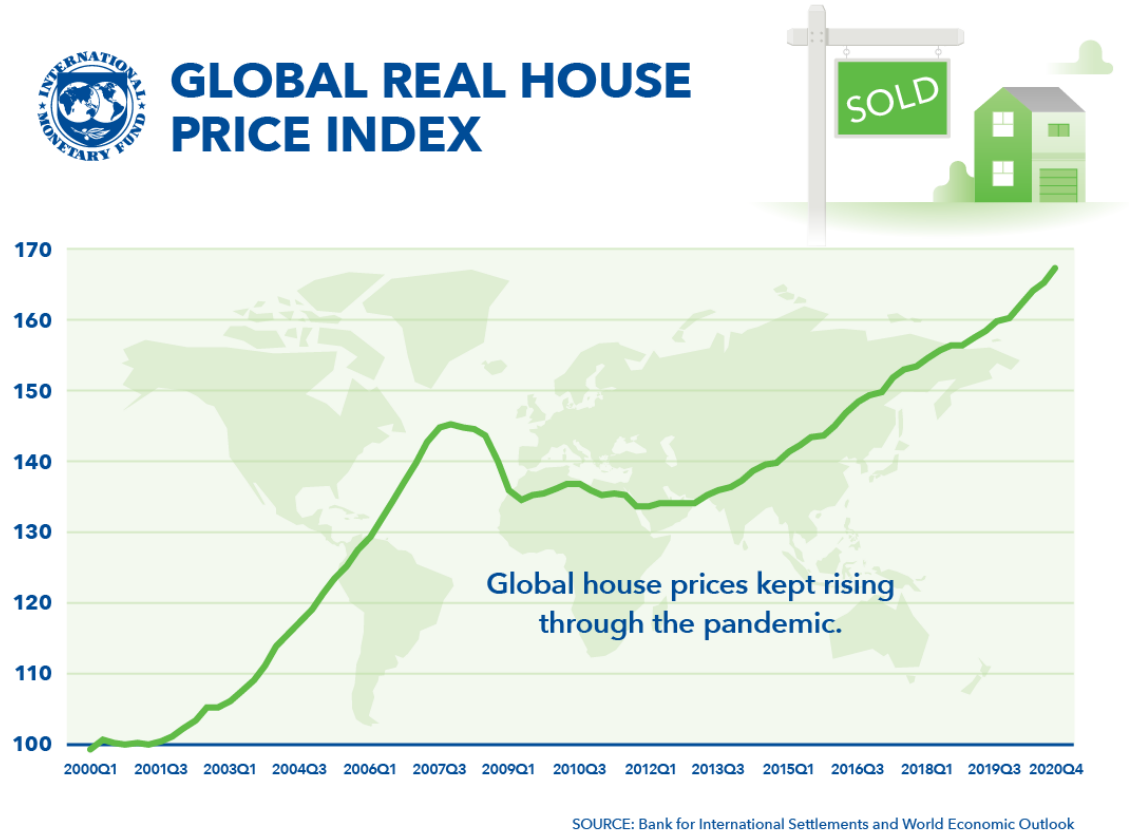
Agenda

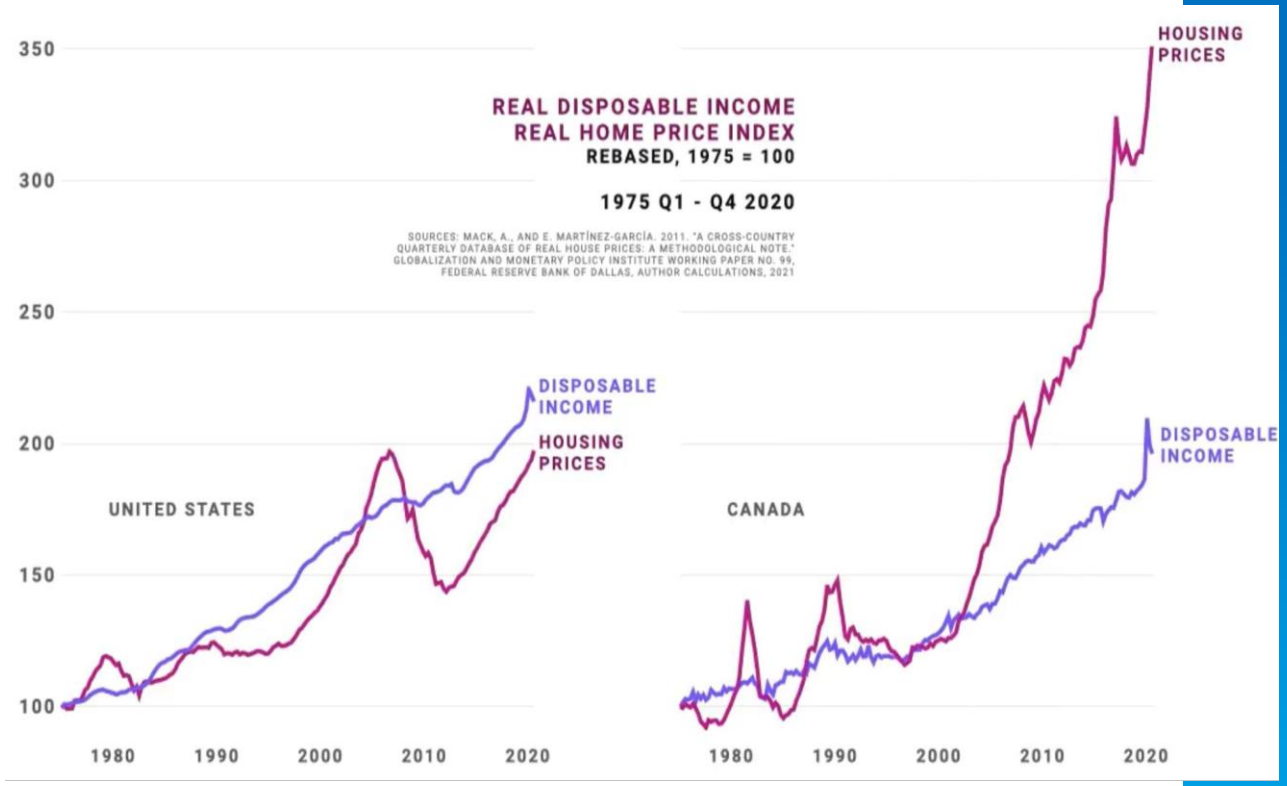
- Business Problem Understanding
- Data Modelling
- Insights & Recommendation

Business Problem Understanding

Skyrocketing Housing Prices

Housing Prices Inflating Globally. | Possible Reason: Increase of Disposable Income.

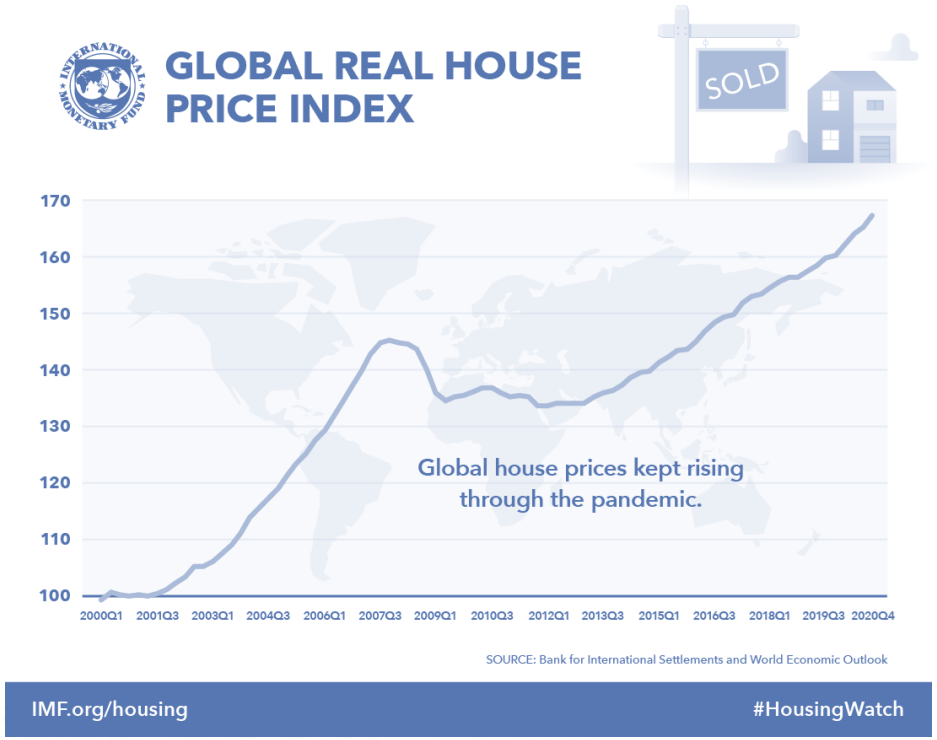


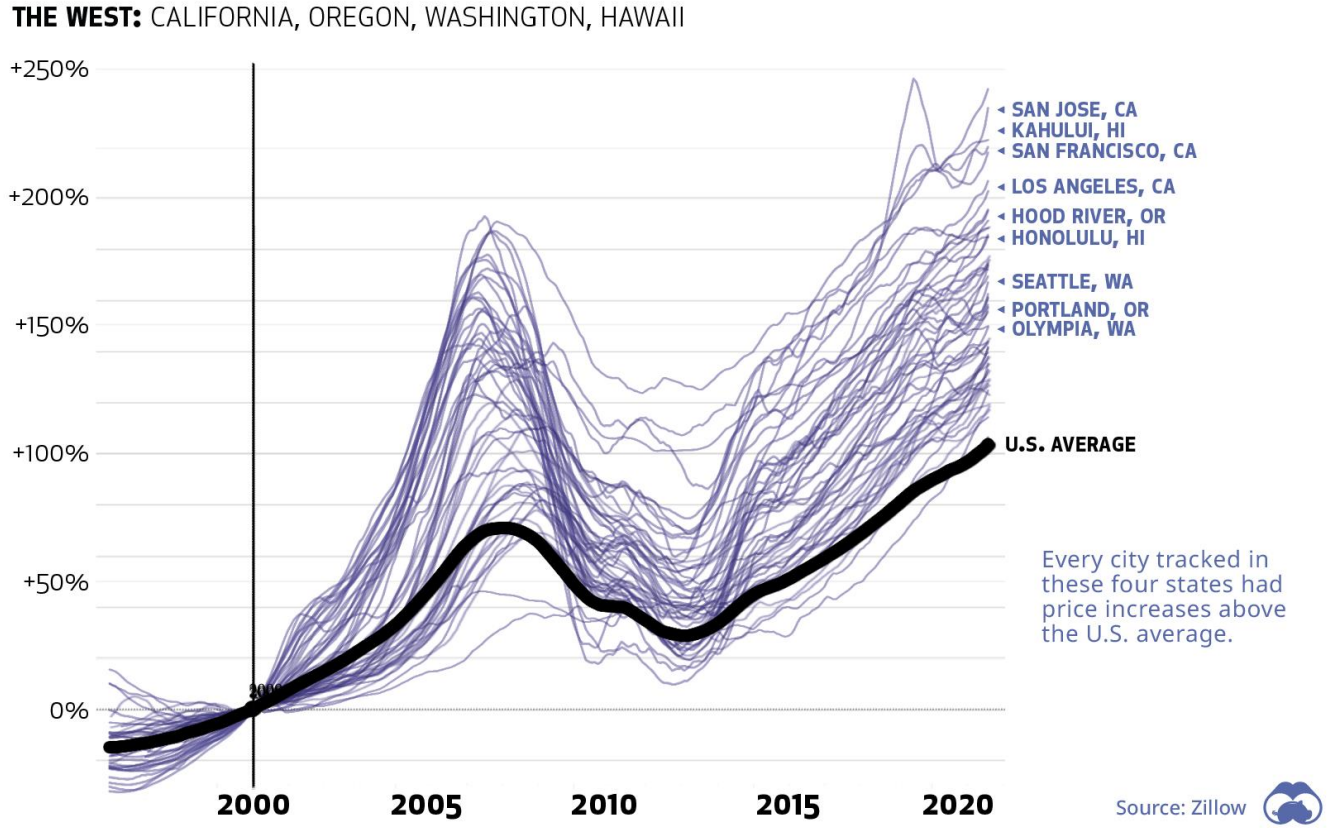


Ref: https://www.imf.org/external/research/housing/index.htm Dated: Nov/2021

Housing Prices in Western USA vs Global Trend

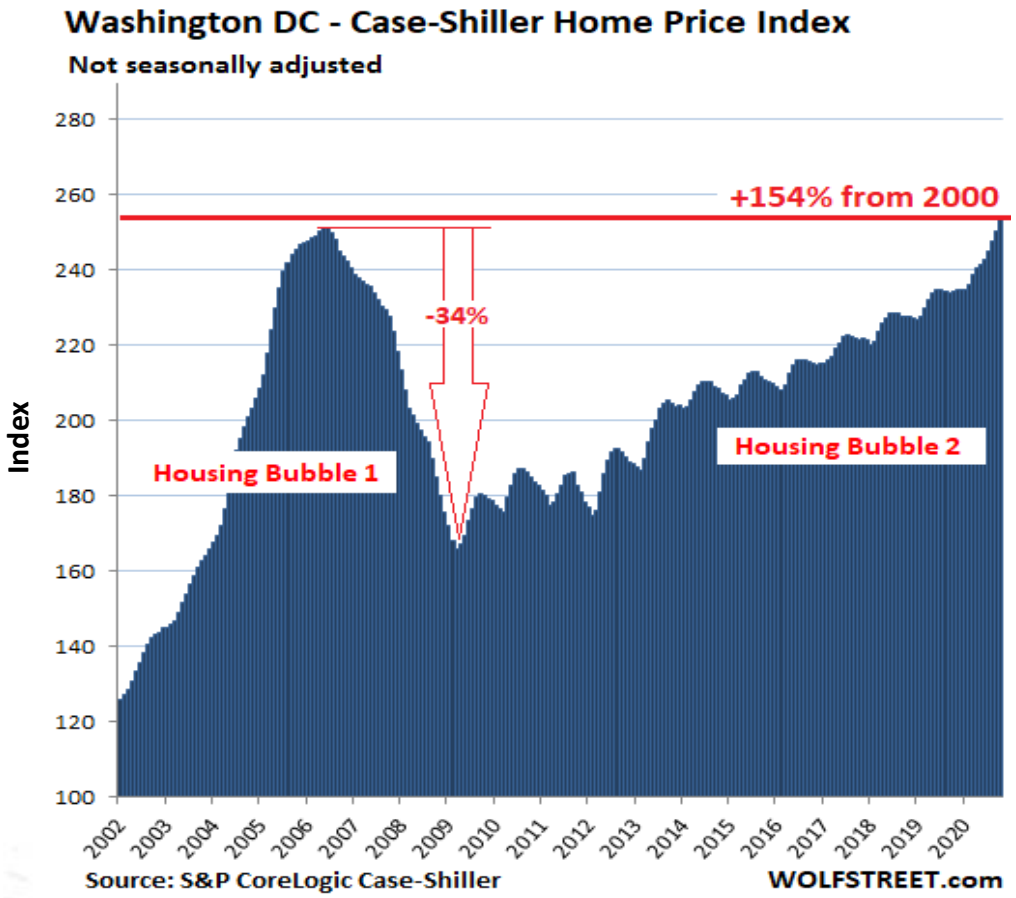
Housing Prices in Western USA inflating parallelly to Global Inflation.





Ref: https://www.visualcapitalist.com/20-years-of-home-price-changes-in-every-u-s-city/
dated: Oct/2020

Housing Market magnified to Washington DC



Ref: https://wolfstreet.com/2020/12/29/the-most-splendid-housing-bubbles-in-america-december-update-on-house-price-inflation/
Dated: Dec 2020

FYI: The Case-Shiller Home price Indices are calculated from data on repeat sales of single-family homes.

Increasing trend of House Price observed over years

Ref: https://www.thekproperties.com/vamd-news/december-2019-housing-market-update-washington-dc-and-baltimore-metro-median-sales-prices-at-record-levels-closed-sales-surge-in-both-areas-inventory-levels-at-decade-low Dated: Jan 2020

What's the Solution? Turn Chaos into Clarity

Problem

Inflating prices and Unpredictability of prices based on different affecting variables.

Creates a Chaos

Solution

Use Data Science and Machine Learning to generate predictability. Let numbers speak to You.

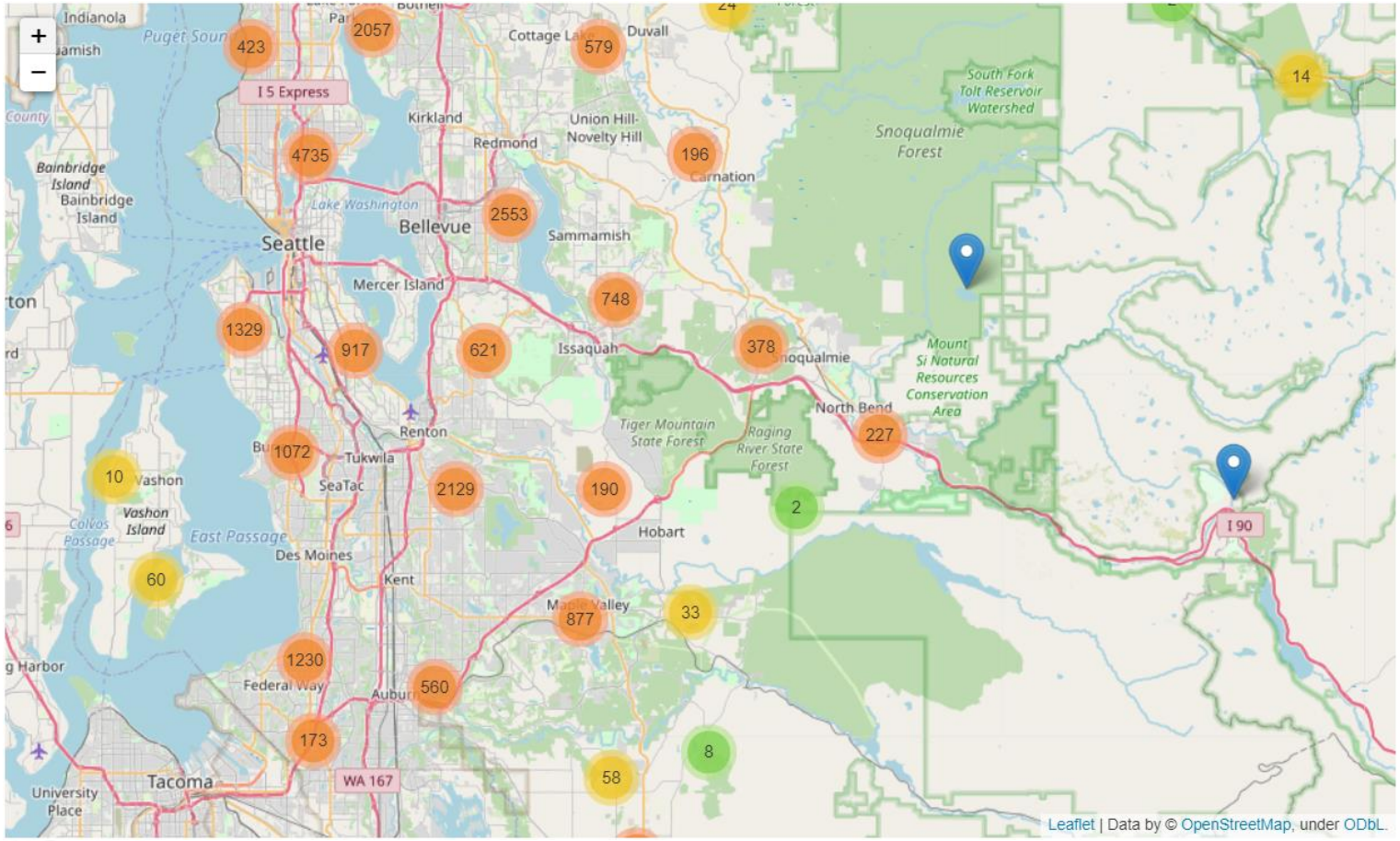
Gives the Clarity

What we are dealing with? Washington's Housing Market

***Washington, is a state in the Pacific Northwest region of the Western United States.**

Houses spread over cities of Washington
(as present in dataset)

Kirkland	Carnation
Seattle	Woodinville
Auburn	North Bend
Kent	Issaquah
Bellevue	Mercer Island
Snoqualmie	Vashon
Renton	Kenmore
Redmond	Duvall
Maple Valley	Fall City
Federal Way	Black Diamond
Sammamish	Medina
Bothell	Enumclaw



Our Resource: The Dataset

The dataset and different variables: Dataset contains 21613 entries and 23 columns.

Time Duration (May/2014 - April/2015)

	cid	dayhours	price
0	3876100940	20150427T000000	600000
1	3145600250	20150317T000000	190000
2	7129303070	20140820T000000	735000
3	7338220280	20141010T000000	257000
4	7950300670	20150218T000000	450000

Identity, Timestamp, Target (price) |

yr_built	yr_renovated
1966	0
1948	0
1966	0
2009	0
1924	0

House Age |

room_bed	room_bath	ceil	coast	sight	condition	quality	furnished
4.0	1.75	1	0	0.0	3	8.0	0.0
2.0	1.00	1	0	0.0	4	6.0	0.0
4.0	2.75	2	1	4.0	3	8.0	0.0
3.0	2.50	2	0	0.0	3	8.0	0.0
2.0	1.00	1	0	0.0	3	7.0	0.0

Variables defining Living Standard

living_measure	lot_measure	ceil_measure	basement	total_area	living_measure15	lot_measure15
3050.0	9440.0	1800.0	1250.0	12490	2020.0	8660.0
670.0	3101.0	670.0	0.0	3771	1660.0	4100.0
3040.0	2415.0	3040.0	0.0	5455	2620.0	2433.0
1740.0	3721.0	1740.0	0.0	5461	2030.0	3794.0
1120.0	4590.0	1120.0	0.0	5710	1120.0	5100.0

House Area measures

zipcode	lat	long
98034	47.7228	-122.183
98118	47.5546	-122.274
98118	47.5188	-122.256
98002	47.3363	-122.213
98118	47.5663	-122.285

| Location

Our Resource: The Dataset (Renamed)

The dataset and different variables: Dataset contains 21613 entries and 23 columns.

Time Duration (May/2014 - April/2015)

	house_id	date	price	yr_built	yr_renovated	bedroom	ratio_bathroom	total_floors	seaface	sight_viewed	condition	quality_grade	furnished
0	3876100940	2015-04-27	600000	1966	0	4.0	1.75	1.0	0.0	0.0	3	8.0	0.0
1	3145600250	2015-03-17	190000	1948	0	2.0	1.00	1.0	0.0	0.0	4	6.0	0.0
2	7129303070	2014-08-20	735000	1966	0	4.0	2.75	2.0	1.0	4.0	3	8.0	0.0

Identity, Timestamp, Target | House Age | Variables defining Living Standard

living_area	lot_area	floor_area	basement_area	basement_area	total_area	living_area_2015	lot_area_2015	zipcode	latitude	longitude
3050.0	9440.0	1800.0	1250.0	1250.0	12490.0	2020.0	8660.0	98034	47.7228	-122.183
670.0	3101.0	670.0	0.0	0.0	3771.0	1660.0	4100.0	98118	47.5546	-122.274
3040.0	2415.0	3040.0	0.0	0.0	5455.0	2620.0	2433.0	98118	47.5188	-122.256

House Area measures | Location

The Hybrid Dataset

Used present variables to introduce new synthetic variables to turn into a Hybrid dataset.

Introduced 10 meaningful synthetic variables.

house_id	prev_sold
3876100940	0
3145600250	0
7129303070	0

date	sold_month
2014-05-02	5
2014-05-02	5
2014-05-02	5

yr_built	house_age
1909	107
1948	68
1979	37

bedroom	ratio_bathroom	bathroom
4.0	1.75	10.00
2.0	1.00	5.25
4.0	2.75	10.00

Identity -> sold before?

timestamp -> sale month

built year -> house age

bathroom ratio -> no. of bathrooms

basement_area	basement_orNot
1010.0	1
1000.0	1
1330.0	1

yr_renovated	renovation_yrs	renovated_orNot
0	0	0
0	0	0
0	0	0

zipcode	city	population	population_density
98105	Seattle	50434.0	4717.7
98136	Seattle	16997.0	2854.7
98092	Auburn	47824.0	415.2

basement area -> basement present?

year renovated -> renovation age & renovated Boolean

Zipcode -> city, population & density

Data Modelling

Part.1. Basic Regression Modelling using StatsModel

Part.2 Advanced Regression & ML Modelling:

Phase.1 without any treatment to outliers and multi-collinearity

Phase.2 with all treatment to outliers and multi-collinearity

Basic Regression Modeling: Approach & Reasoning

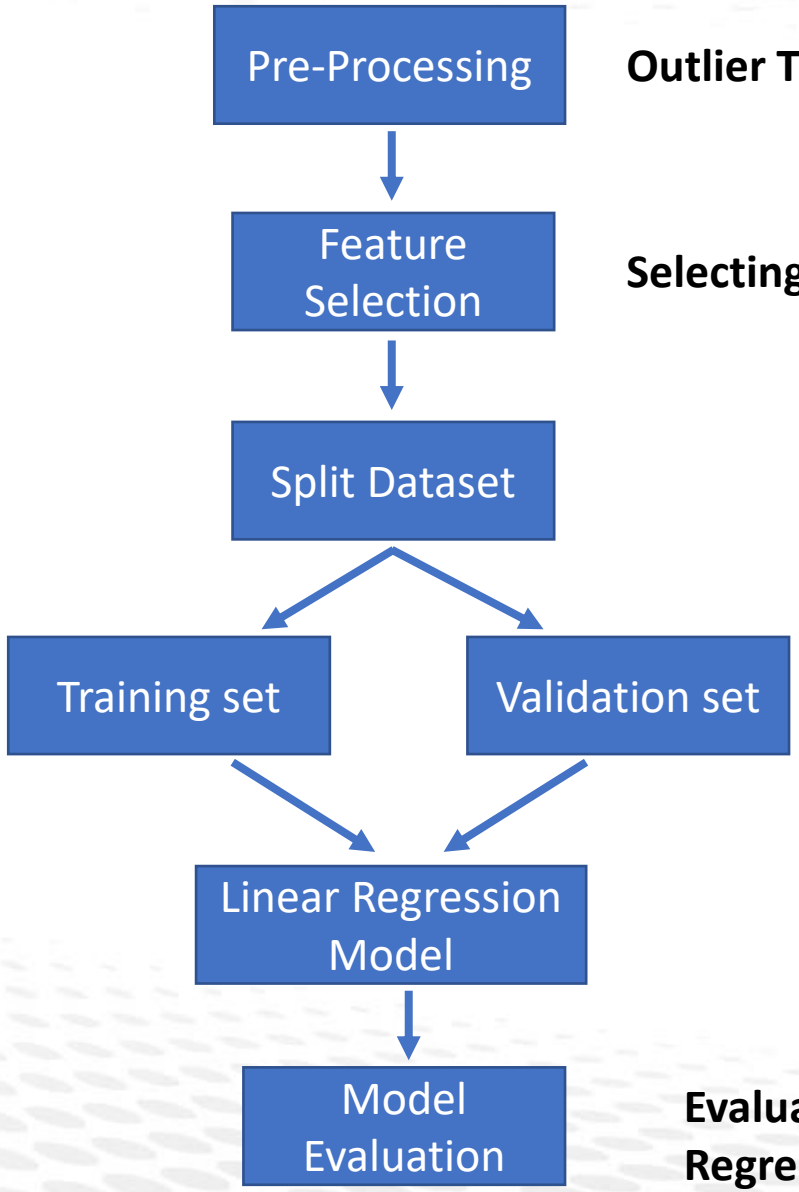
The Simple linear regression statistical method allows us to summarize and study the relationship between two continuous quantitative variables.

- **Variable, denoted as x, are regarded as the predictor, explanatory, or independent variable.**
- **The target variable, denoted as y, is regarded as the response, outcome, or dependent variable.**

In our case ‘Price’ is the target variable and it is assumed that other variables are independently affecting the Price.

This will give us a basic linear regression equation.

Basic Linear Regression using Statsmodel: Work Flow



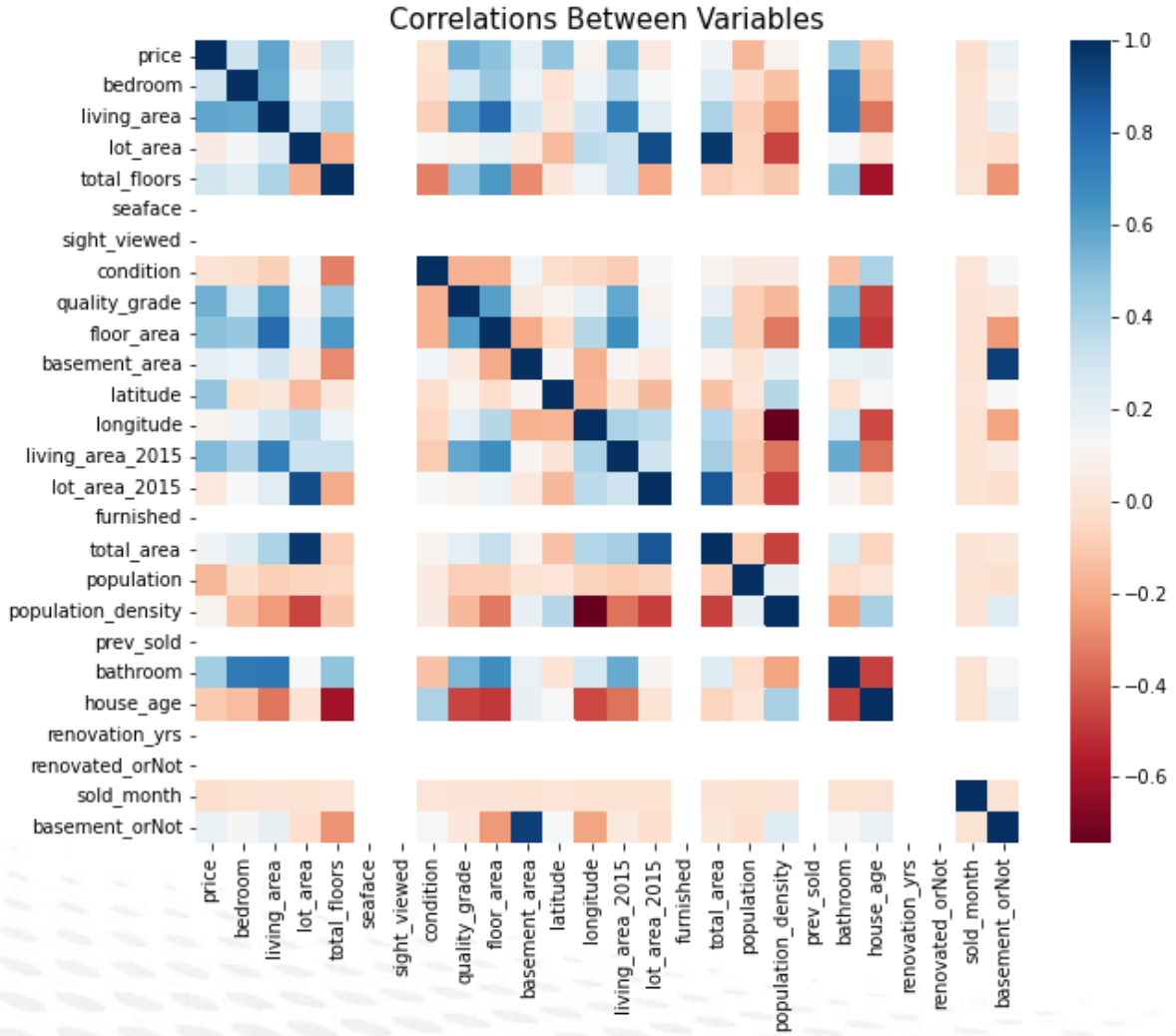
Outlier Treatment , missing value imputation

Selecting best features correlated with target variable

Training the model on Train dataset and testing on validation set.

Evaluation on performance metrics & generating Linear Regression Equation

Basic Linear Regression using Statsmodel: Feature Selection



Feature selection based on correlation:

Corr wrt price>0.1 & Corr wrt price<-0.1

Selected features:

bedroom,
living_area,
total_floors,
quality_grade,
floor_area,
basement_area,
latitude,
living_area_2015,
total_area,
population,
population_density,
bathroom,
basement_orNot,
zipcode

Basic Linear Regression using Statsmodel: Results

Linear Regression Equation

Price =
(-13035181.5 * Intercept) + (3758.52 * bedroom) + (74.83 * living_area) + (-1484.73 * total_floors) + (51286.05 * quality_grade) + (44.3 * floor_area) + (3.95 * basement_area) + (592247.9 * latitude) + (61.13 * living_area_2015) + (0.75 * total_area) + (-2.66 * population) + (37.17 * population_density) + (-2866.85 * bathroom) + (19175.76 * basement_orNot) + (-156.7 * zipcode)

Performance Evaluation

OLS regression results

R-squared: 0.635
RMSE: 79389

Model Performance is only 63.5%

OLS Regression Results

Dep. Variable:	price	R-squared:	0.635
Model:	OLS	Adj. R-squared:	0.635
Method:	Least Squares	F-statistic:	2124.
Date:	Mon, 17 Jan 2022	Prob (F-statistic):	0.00
Time:	10:54:34	Log-Likelihood:	-2.1730e+05
No. Observations:	17109	AIC:	4.346e+05
Df Residuals:	17094	BIC:	4.348e+05
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.304e+07	1.41e+06	-9.272	0.000	-1.58e+07	-1.03e+07
bedroom	3758.5219	1960.847	1.917	0.055	-84.939	7601.983
living_area	74.8280	3.841	19.482	0.000	67.300	82.356
total_floors	-1484.7348	1945.650	-0.763	0.445	-5298.409	2328.939
quality_grade	5.129e+04	1701.727	30.138	0.000	4.8e+04	5.46e+04
floor_area	44.3043	4.039	10.968	0.000	36.387	52.222
basement_area	3.9452	9.041	0.436	0.663	-13.776	21.666
latitude	5.922e+05	7956.080	74.440	0.000	5.77e+05	6.08e+05
living_area_2015	61.1251	2.706	22.590	0.000	55.821	66.429
total_area	0.7467	0.334	2.233	0.026	0.091	1.402
population	-2.6624	0.088	-30.417	0.000	-2.834	-2.491
population_density	37.1702	1.259	29.517	0.000	34.702	39.639
bathroom	-2866.8453	575.095	-4.985	0.000	-3994.091	-1739.599
basement_orNot	1.918e+04	4136.395	4.636	0.000	1.11e+04	2.73e+04
zipcode	-156.6977	13.829	-11.331	0.000	-183.803	-129.592

Omnibus:	13.504	Durbin-Watson:	2.001
Prob(Omnibus):	0.001	Jarque-Bera (JB):	14.555
Skew:	0.035	Prob(JB):	0.000691
Kurtosis:	3.124	Cond. No.	2.42e+08

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.42e+08. This might indicate that there are strong multicollinearity or other numerical problems.

Part.1. Basic Regression Modelling using StatsModel

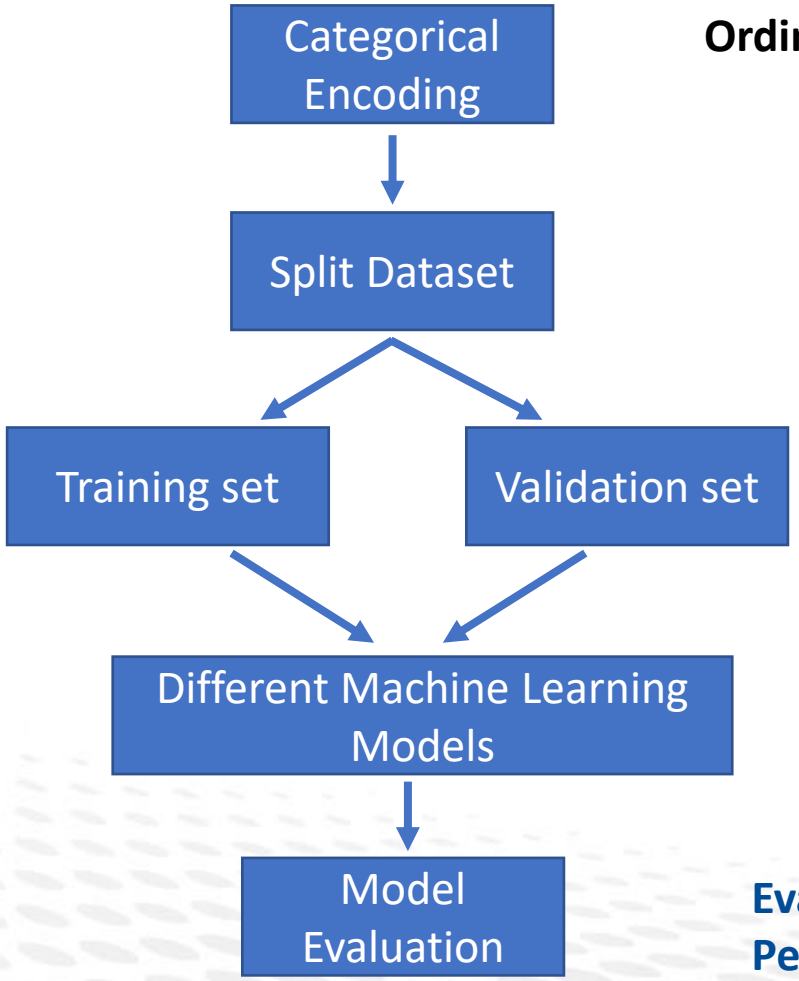
Part.2 Advanced Regression & ML Modelling:

Phase.1 without any treatment to outliers and multi-collinearity

Phase.2 with all treatment to outliers and multi-collinearity

Advanced Regression & ML modeling: Work Flow

Phase.1 without any treatment to outliers and multi-collinearity



Ordinal Encoding of City. Zipcode is already an encoded entity.

Training the model on Train dataset and testing on validation set.

Different ML models used: Linear Regression, Polynomial Regression , Ridge Regression, Lasso Regression, Elastic Regression, Support Vector regression, Random Forest Regressor, XG Boost Regressor.

Evaluation on performance metrics
Performance Metrics: MAE, MSE, RMSE, R2 Score, RMSE Cross-Validn

Advanced Regression & ML modeling: Approach & Reasoning

Advanced Regression techniques used:

1. Linear Regression: Because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and so the statistical properties of the resulting estimators are easier to determine.

2. Polynomial Regression: Polynomial provides the best approximation of the relationship between the dependent and independent variable. A Broad range of function can be fit under it. Polynomial basically fits a wide range of curvature.

3. Ridge Regression: Ridge is used for the analysis of multicollinearity in multiple regression data. It is most suitable when a data set contains a higher number of predictor variables than the number of observations. The second-best scenario is when multicollinearity is experienced in a set.

4. Lasso Regression: Lasso is used to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. The lasso does this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero.

5. Elastic Net Regression: A regularized regression model that combines l1 and l2 penalties, i.e., lasso and ridge regression. regularization helps in overfitting problems of the models. Elastic Net is a regression method that performs variable selection and regularization both simultaneously.

Advanced Regression & ML modeling: Approach & Reasoning

Advanced Regression techniques used:

7. **Support Vector Regressor:** SVR is a supervised learning algorithm that is used to predict discrete values. Unlike other Regression models that try to minimize the error between the real and predicted value, the SVR tries to fit the best line within a threshold value.
8. **Random Forest Regressor:** is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
9. **XG Boost Regressor:** The two main reasons to use XGBoost are execution speed and model performance. XGBoost dominates structured or tabular datasets on classification and regression predictive modeling problems.

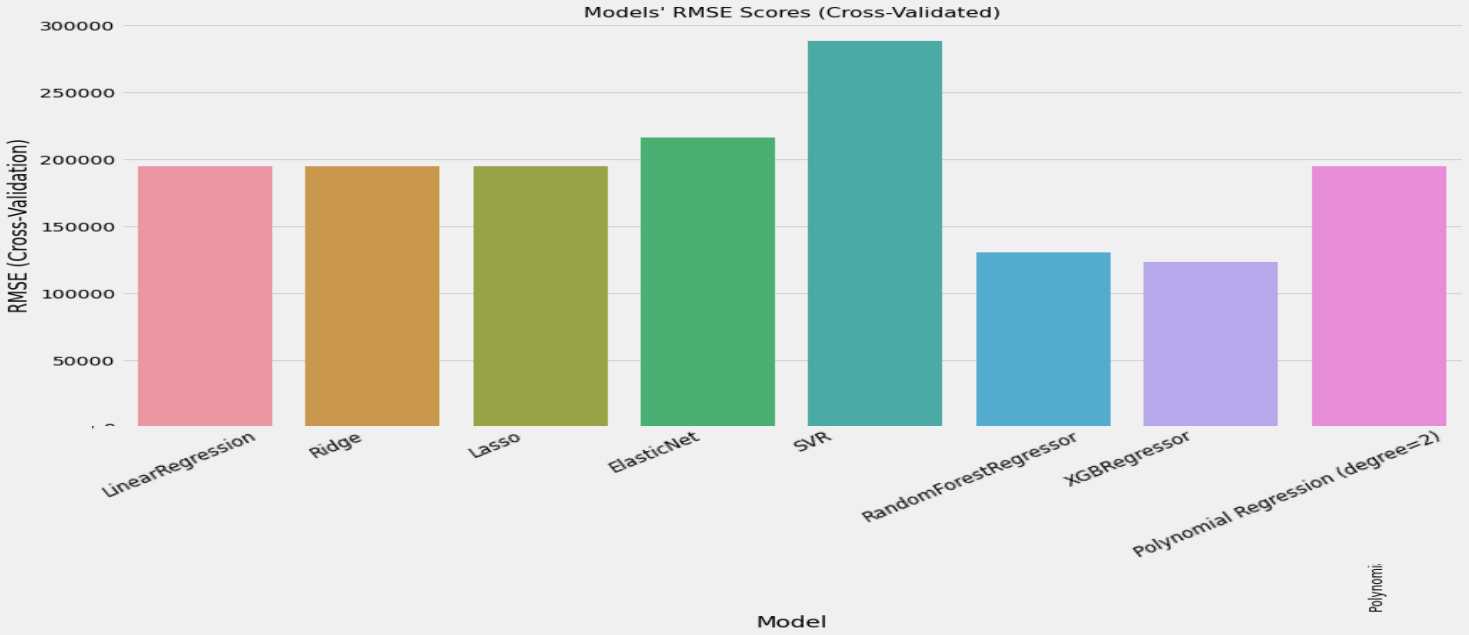
Performance Metrics:

1. **Mean Absolute Error (MAE):** shows the difference between predictions and actual values.
2. **Mean squared error (MSE):** tells you how close a regression line is to a set of points
3. **Root Mean Square Error (RMSE):** shows how accurately the model predicts the response.
4. **R squared (R2):** will be calculated to find the goodness of fit measure.
5. **RMSE cross validation:** The less the Root Mean Squared Error (RMSE), better the model is. (cv=5)

Advanced Regression & ML modeling: Results-1

Performance Evaluation & Comparison

	Model	MAE	MSE	RMSE	R2 Score	RMSE (Cross-Validation)
6	XGBRegressor	66778.233436	1.632371e+10	127764.282859	0.884653	123340.524755
5	RandomForestRegressor	69847.446951	1.805325e+10	134362.369443	0.872431	130698.273525
2	Lasso	120999.417531	3.851481e+10	196251.910285	0.727845	195105.910072
1	Ridge	121001.703622	3.852493e+10	196277.682617	0.727774	195129.586948
0	LinearRegression	121010.031719	3.851637e+10	196255.867875	0.727834	195130.606017
7	Polynomial Regression (degree=2)	95224.245472	2.444431e+10	156346.748958	0.827271	195130.606017
3	ElasticNet	134349.792172	4.810076e+10	219318.845637	0.660109	216292.839518
4	SVR	159320.250834	8.839376e+10	297310.881938	0.375389	288353.453162



LR, PR, Ridge, Lasso performing almost similar. SVR is worst performing.

Random Forest is performing better than rest of these.

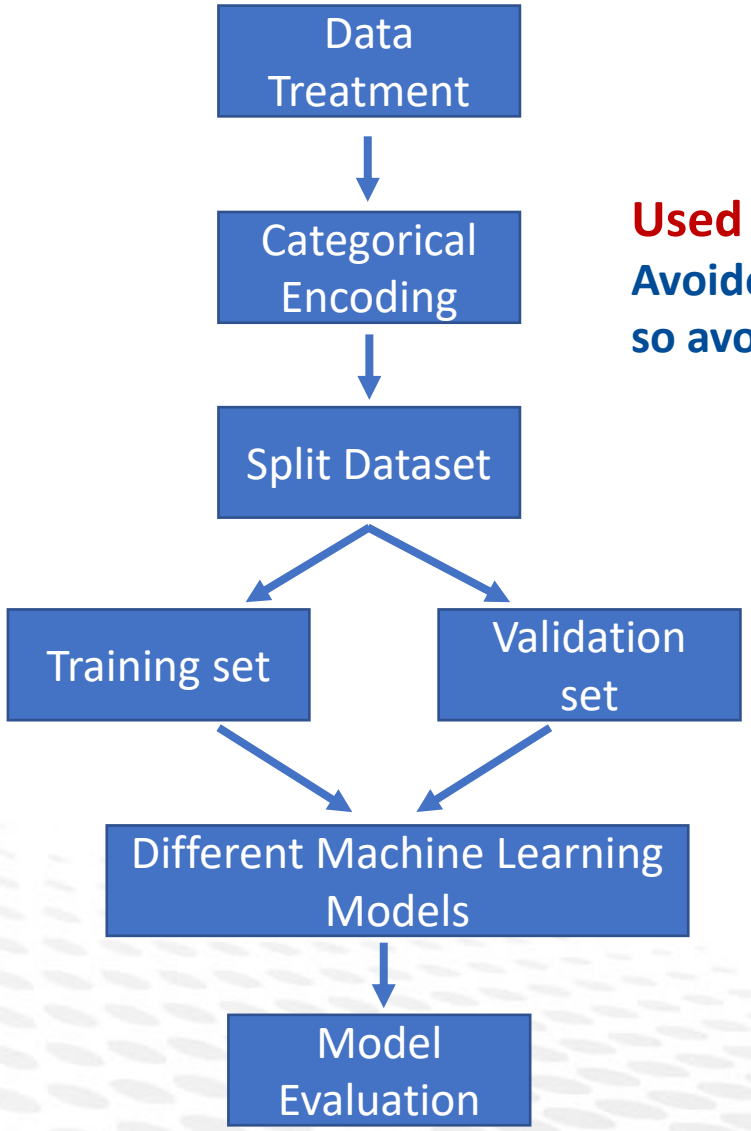
XGB is performing the best.

Best Performance is by XGB Regressor

88.5%

Advanced Regression & ML modeling: Work Flow

Phase.2 with treatment to outliers and multi-collinearity



Outliers were treated and VIF used to deal with multicollinearity.

Used One Hot Encoding. **Hot encoded zipcode, city, sold_month**
Avoided other variables because other variables were gradings (wither 1-n or 1/0) so avoided encoding them to avoid high cardinality.

Training the model on Train dataset and testing on validation set.

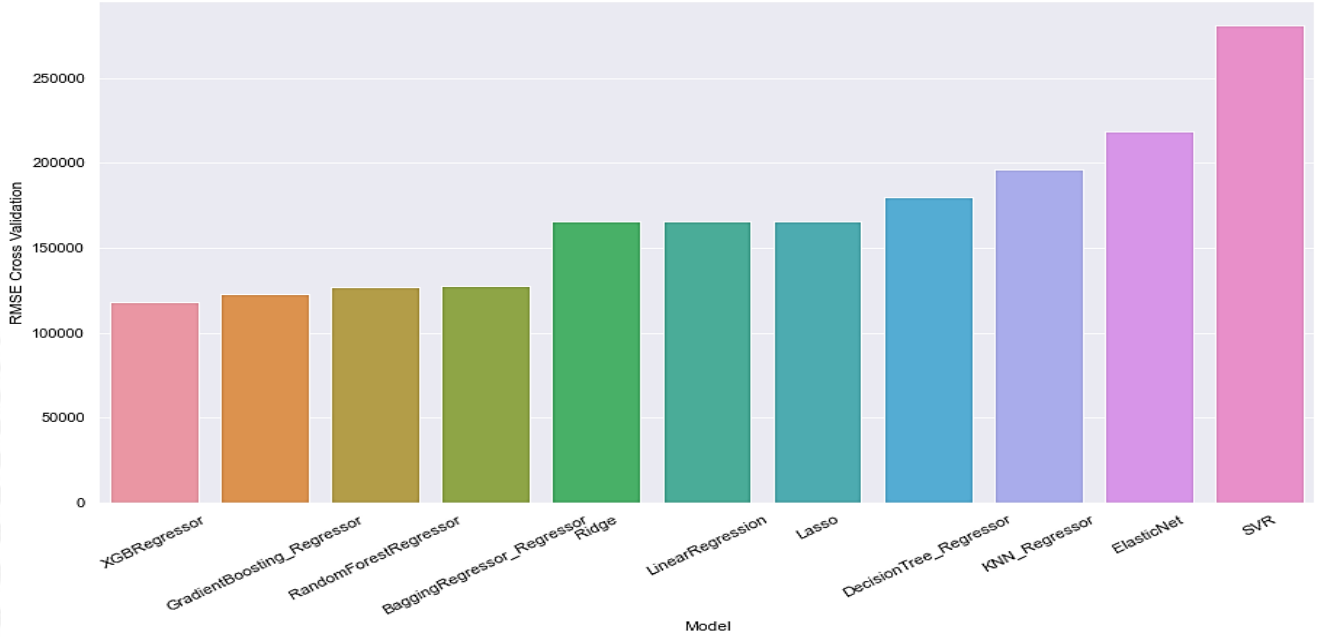
Different ML models used: Linear Regression, Polynomial Regression , Ridge Regression, Lasso Regression, Elastic Regression, Support Vector regression, Random Forest Regressor, XG Boost Regressor.

Evaluation on performance metrics
Performance Metrics: MAE, MSE, RMSE, R2 Score, RMSE Cross-Validn

Advanced Regression & ML modeling: Results-3

Performance Evaluation & Comparison

	Model	MAE	MSE	RMSE	R2 Score	RMSE (Cross-Validation)
5	XGBRegressor	69008.625809	1.417548e+10	119060.818140	0.887628	118112.883106
8	GradientBoosting_Regressor	73105.326965	1.516758e+10	123156.738557	0.879764	122860.539271
4	RandomForestRegressor	72025.008314	1.664007e+10	128996.391090	0.868091	126774.107475
9	BaggingRegressor_Regressor	73443.426515	1.746332e+10	132148.859318	0.861565	127781.028696
0	LinearRegression	101175.175211	2.863646e+10	169223.094598	0.772993	165432.176231
1	Lasso	101174.520648	2.863648e+10	169223.167826	0.772993	165468.910682
7	DecisionTree_Regressor	105557.265105	3.666178e+10	191472.665190	0.709375	180106.421719
6	KNN_Regressor	104967.421384	3.913409e+10	197823.382224	0.689777	196019.766844
2	ElasticNet	138158.011100	5.088374e+10	225574.248212	0.596635	218676.190549
3	SVR	154041.221249	8.370262e+10	289314.059432	0.336474	281061.736946

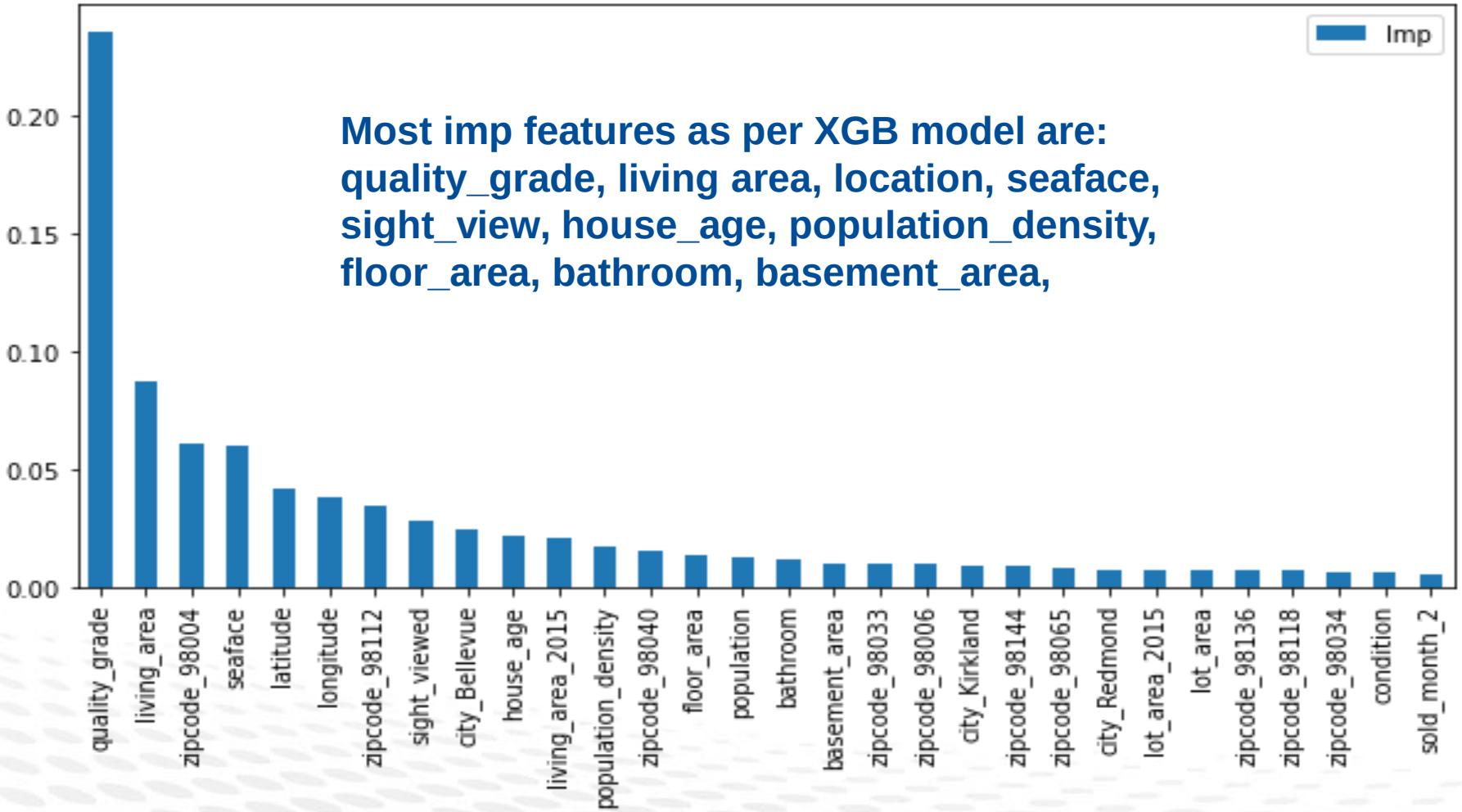


Best Performance is still from XGB Regressor

88.7%

Important Variables : Feature Importance

First 30 features are contributing 83% towards the target prediction.



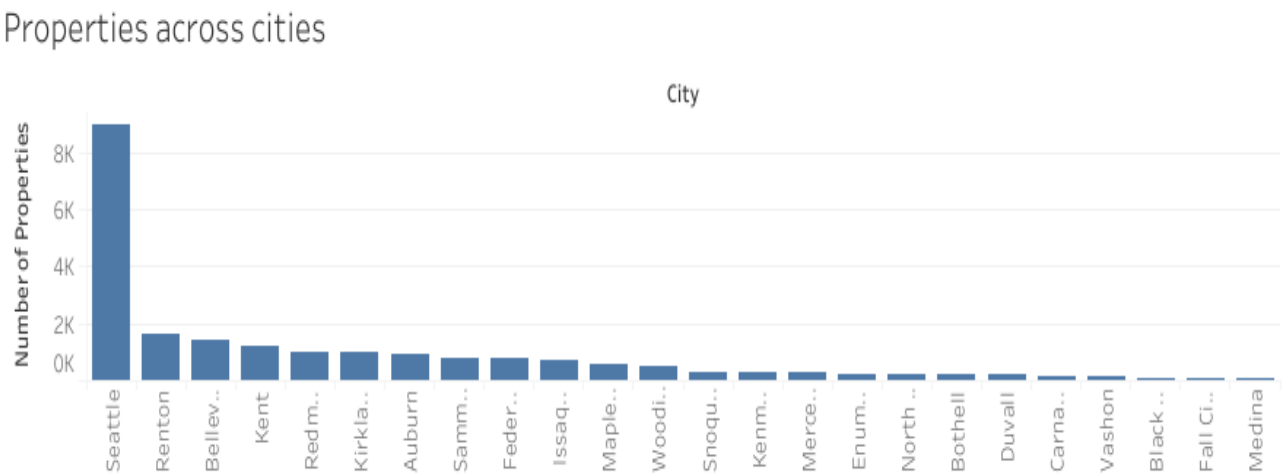
Most imp features as per XGB model are:
quality_grade, living area, location, seaface, sight_view, house_age, population_density, floor_area, bathroom, basement_area,

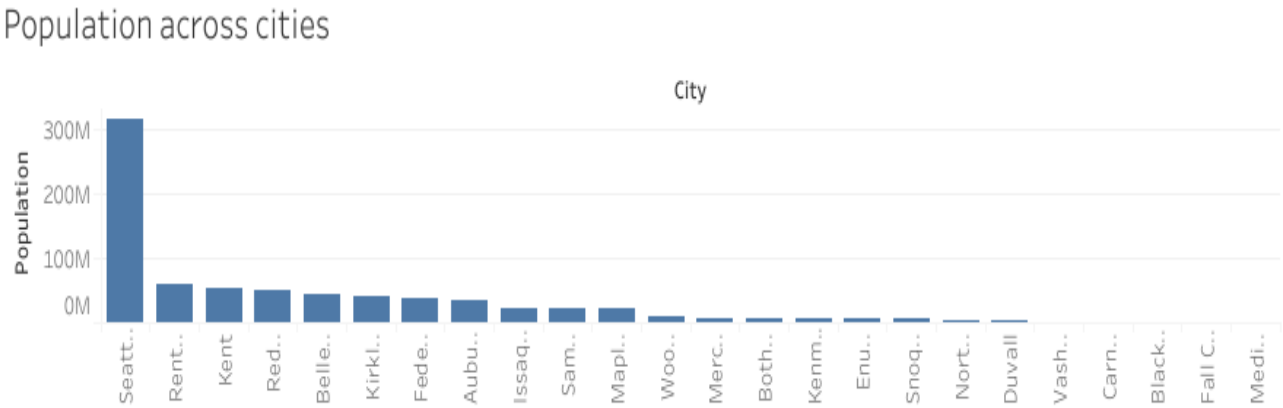
	Imp
quality_grade	0.23481
living_area	0.08709
zipcode_98004	0.06037
seaface	0.05984
latitude	0.04184
longitude	0.03856
zipcode_98112	0.03456
sight_viewed	0.02842
city_Bellevue	0.02434
house_age	0.02149
living_area_2015	0.02133
population_density	0.01705
zipcode_98040	0.01566
floor_area	0.01345
population	0.01270
bathroom	0.01194
basement_area	0.01025
zipcode_98033	0.00993
zipcode_98006	0.00985
city_Kirkland	0.00894
zipcode_98144	0.00878
zipcode_98065	0.00854
city_Redmond	0.00767
lot_area_2015	0.00725
lot_area	0.00716
zipcode_98136	0.00699
zipcode_98118	0.00689
zipcode_98034	0.00630
condition	0.00612
sold_month_2	0.00535

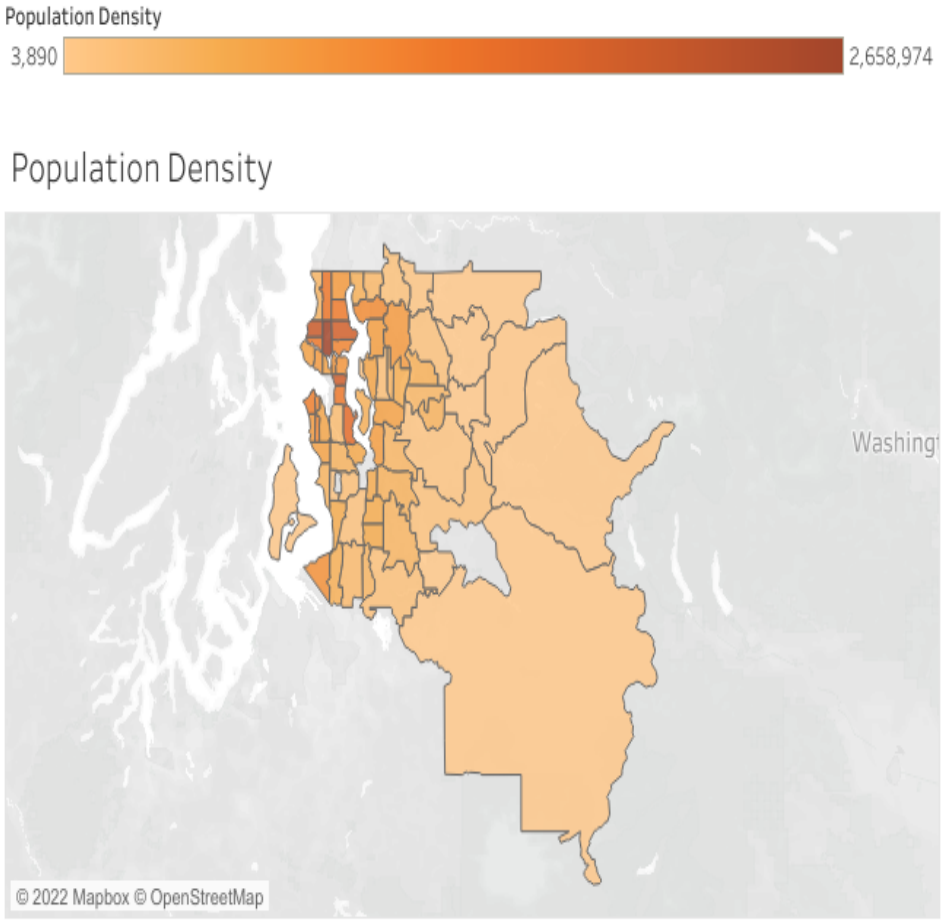
Insights from Analysis & Recommendations

Housing across Washington

Most number of properties are located in Seattle which is a very populated city.



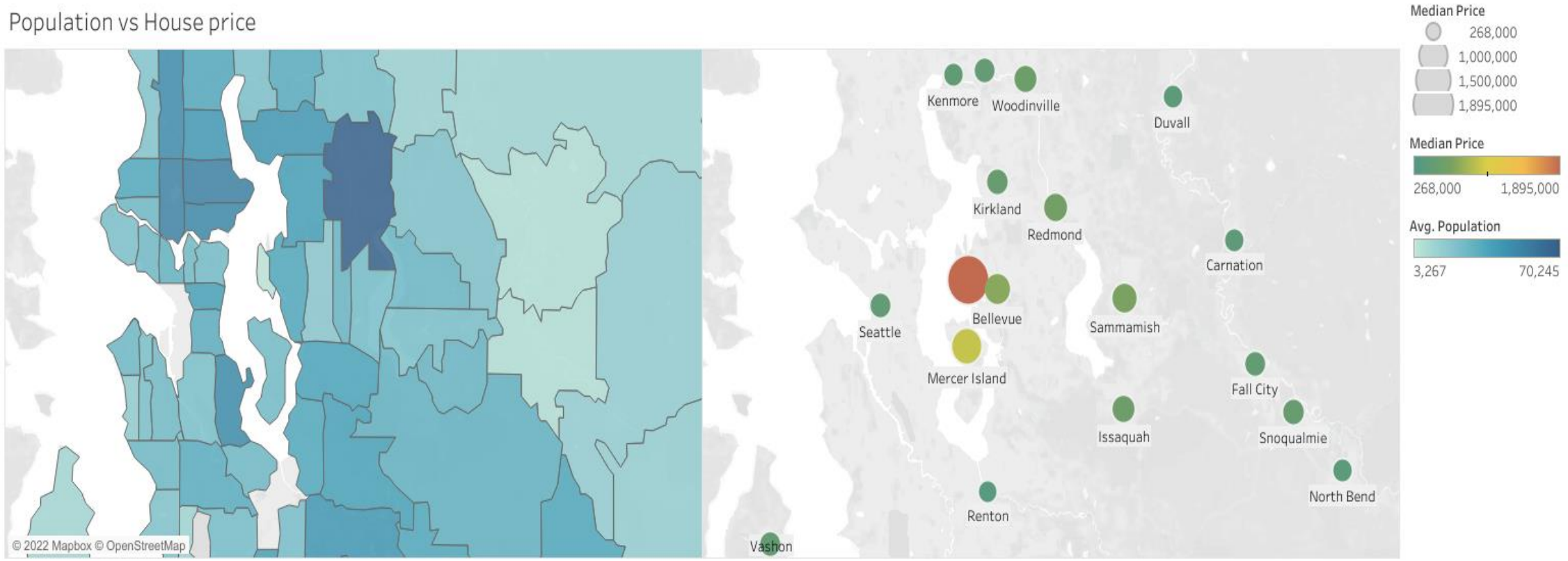




Recommendation: Stakeholders may be careful about the prices based on the population of area.

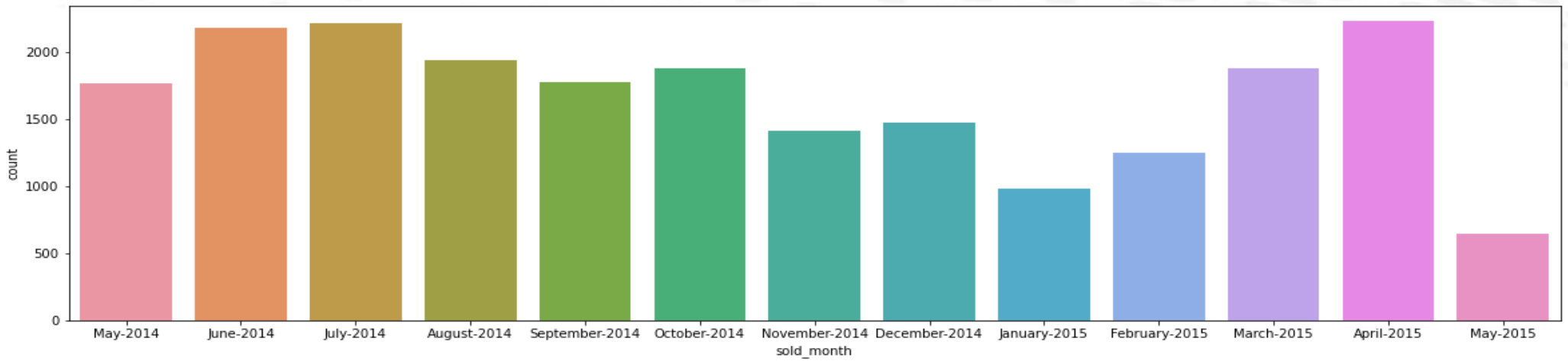
Population affecting the prices

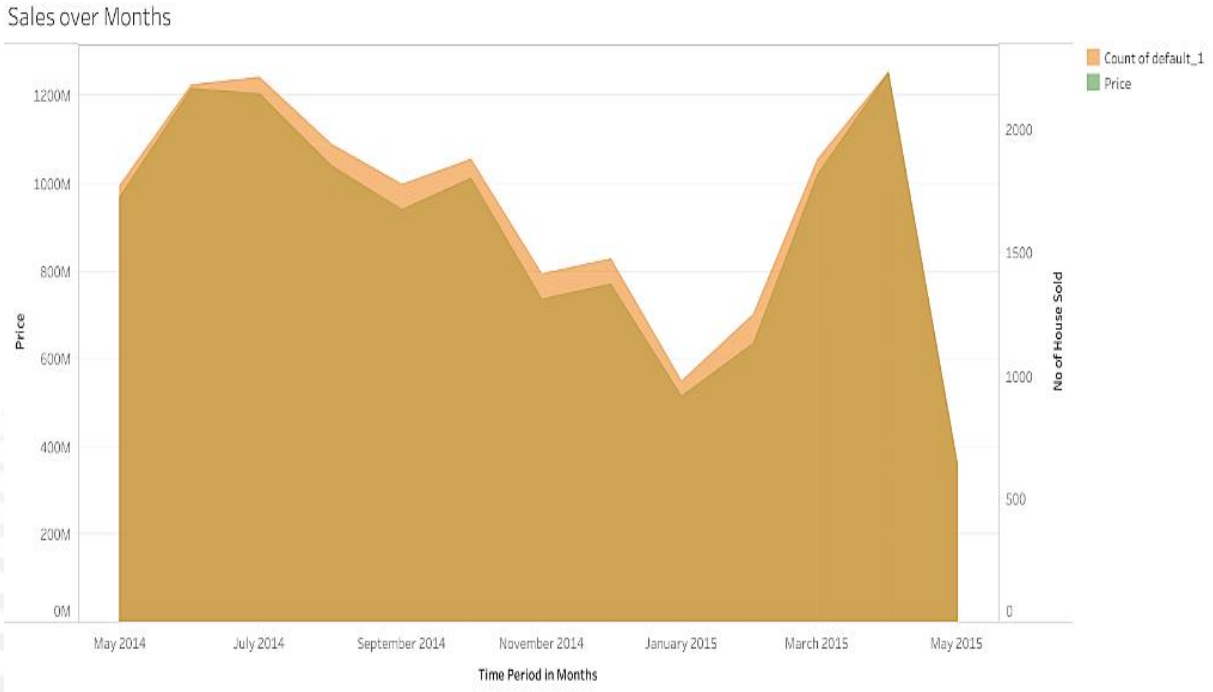
Highly dense cities are costly housing market.



Recommendation: Buyer should avoid densely populated area if possible for an affordable housing prices.

Housing Pricing trend over Months





Highest sales happened in the month of June, July of 2014 and April 2015.
During winters people prefer buying less may be because of discomfort to go out for sight-seeing

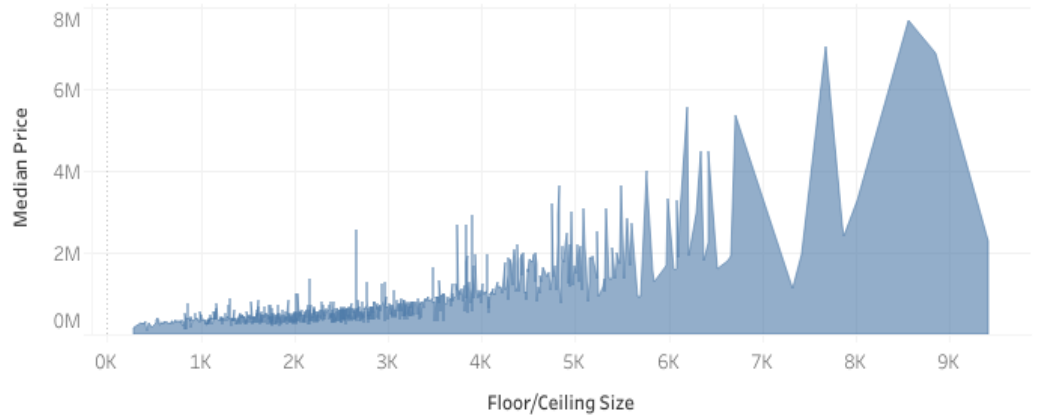
Recommendation 1: Stakeholders may bring more during offer summers to attract more buyers.

Recommendation 2: Buyers should look for prices during winters because prices are generally lowest in the year. They might get a best deal.

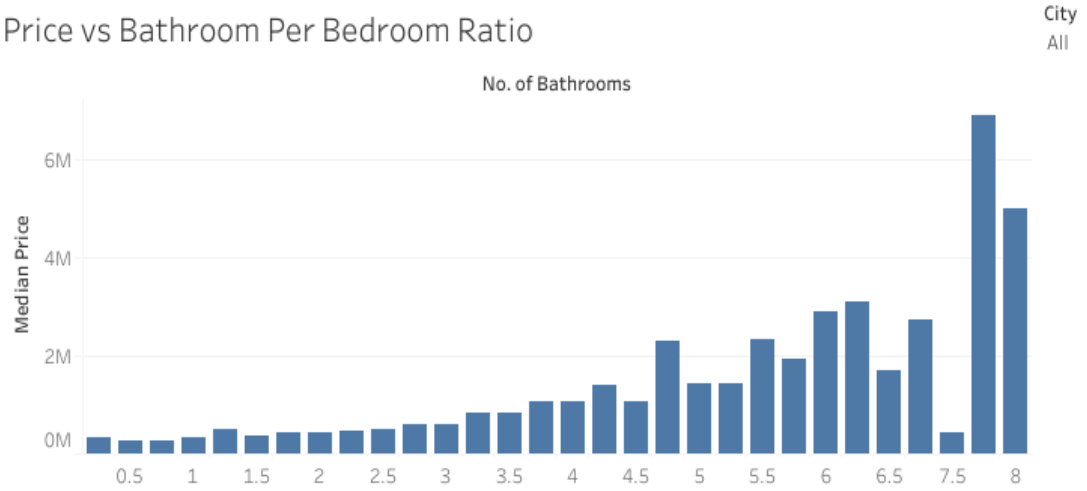
Housing Pricing vs Property

Increase in Median price wrt Floor & Basement Area. And wrt Bedroom & Bathroom ratio.

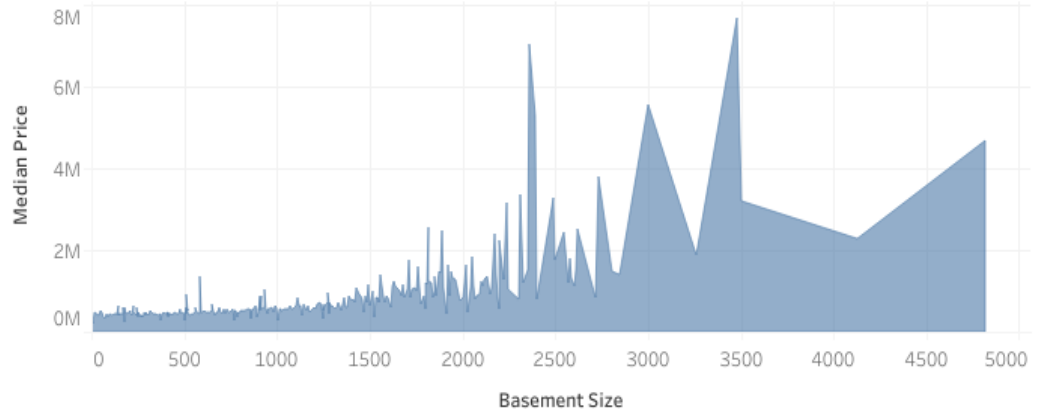
Price vs Floor Size



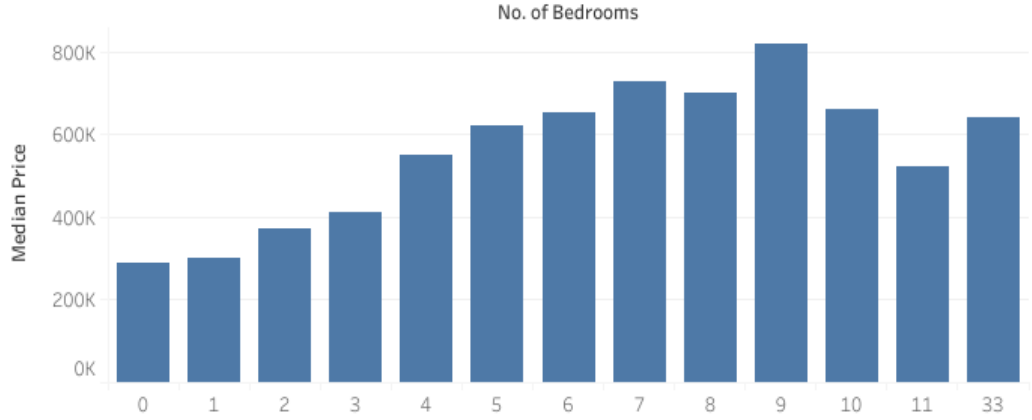
Price vs Bathroom Per Bedroom Ratio



Price vs Basement Size

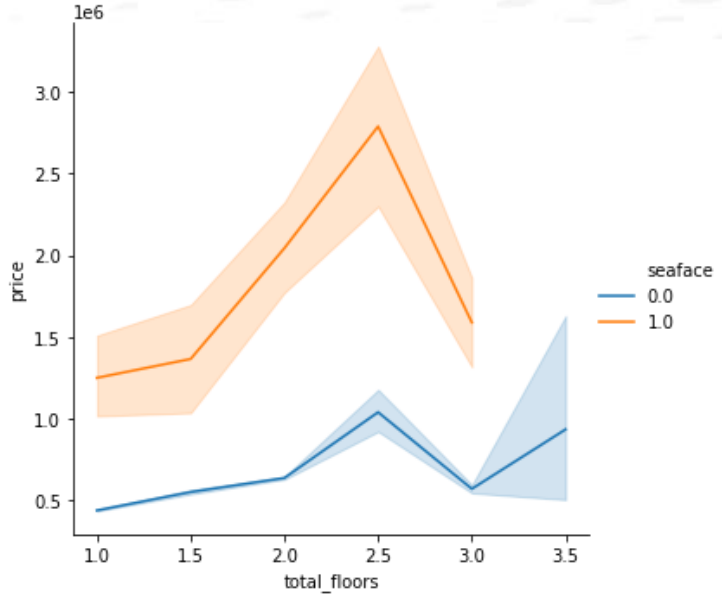


Price vs Bedrooms



Prices increase with number of bedrooms, bathrooms, floor area and basement area.

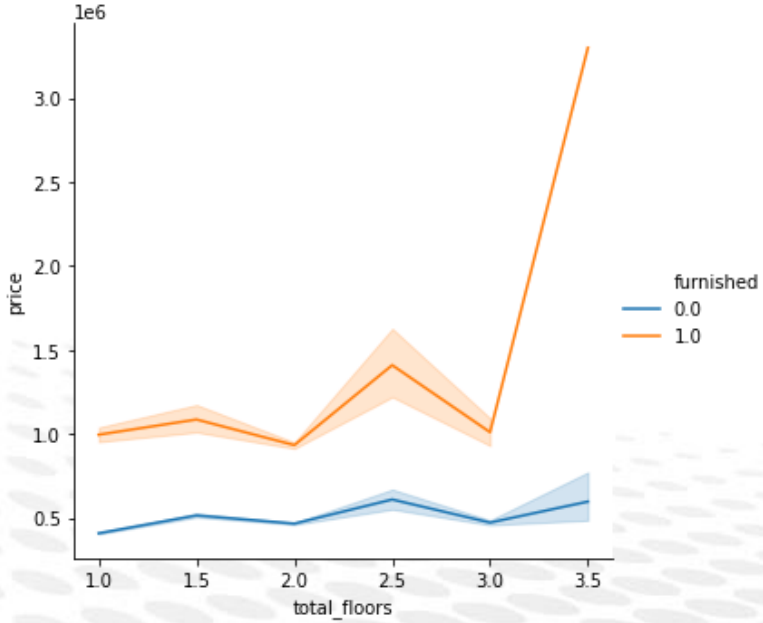
Housing Pricing vs Property

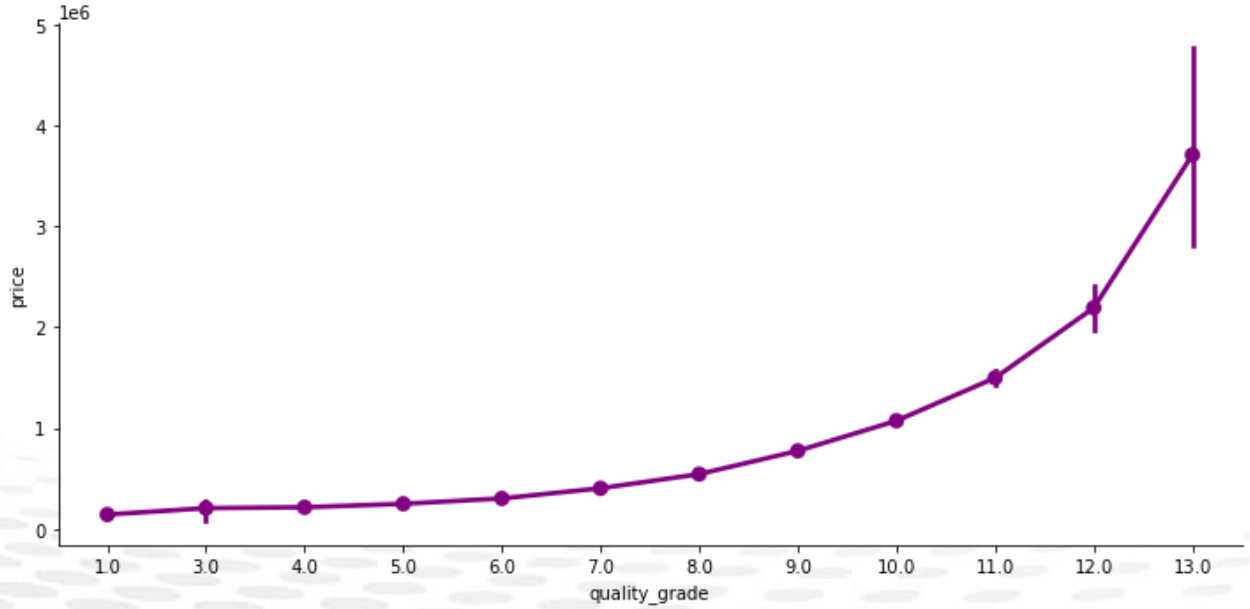


Most of seaface house has higher price even very spike in price is seen with increase in number of floors as compared to non-seafacing houses.

The furnished houses shows higher price with increase in number of floors.

With increase in quality grading price increases.





Final Insights & Recommendation

- Price of premium houses with better conditions and better-quality grading are costlier.
- The house at premium locations and seafront locating houses are costlier.
- Cities like Seattle are over populated because of high density of population and hence price and also high compared to other cities.
- The average price is high at medina but median price are high at Seattle.
- The costlier cities have lower floor area as compared to cheaper cities where big house cost the less for the same floor area in over populated cities.
- The house those are furnished cost more.
- House with a greater number of floors cost more.
- The lot area has not much impact on price yet the large the size of lot area mean higher would-be land and higher would be the house price.

Recommendation: most imp features that affect the price of a property are:
quality_grade, living area, location, seaface, sight_view, house_age, population_density, floor_area, bathroom, basement_area.

People should make their decision based on these factors.
The stakeholder/sellers should price the property while keeping these features in mind.

Thank You