Predicting Real Estate King County Selling Prices

Regression Models Performance

André
Jordi
Sandra
Francisco for moral support

Our Goals

1. Build a machine learning model to predict the selling prices of houses in King County, Washington State;

2. Compare performance of different Regression Machine Learning Models;

3. Choose the best Model after improvements;

4. Identify key points to improve Real Estate Business with Tableau Dashboard;

Before starting to build our Models

- Add and standardize column names
- Data types (datetime, floats…)
- Imperial vs Metric System
- Duplicates
- Nulls

- Drop Irrelevant Columns
- EDA on continuous numerical and discrete numerical
- Pre-processing (scaling, transformations, outliers, feature extraction)

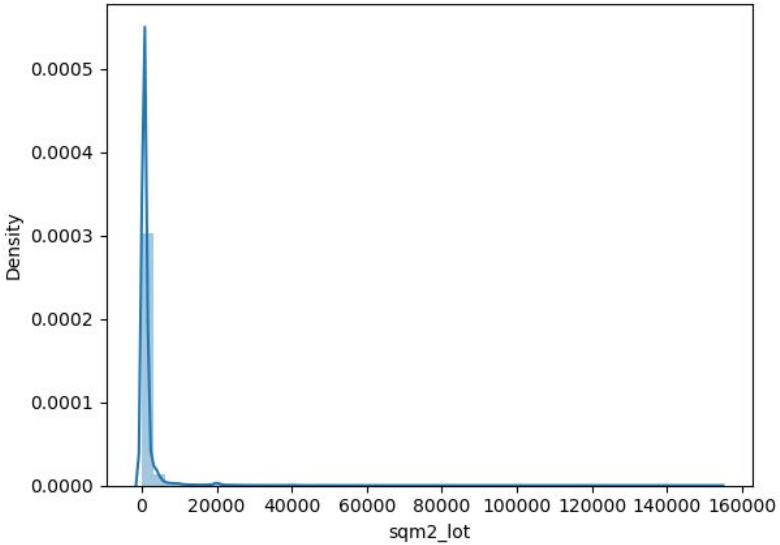
	7129300520	10/13/14	3	1	1180	5650	1.1	0	0.1	3.1
0	6414100192	12/9/14	3	2.25	2570	7242	2.0	0	0	3
1	5631500400	2/25/15	2	1.00	770	10000	1.0	0	0	3
2	2487200875	12/9/14	4	3.00	1960	5000	1.0	0	0	5
3	1954400510	2/18/15	3	2.00	1680	8080	1.0	0	0	3
4	7237550310	5/12/14	4	4.50	5420	101930	1.0	0	0	3

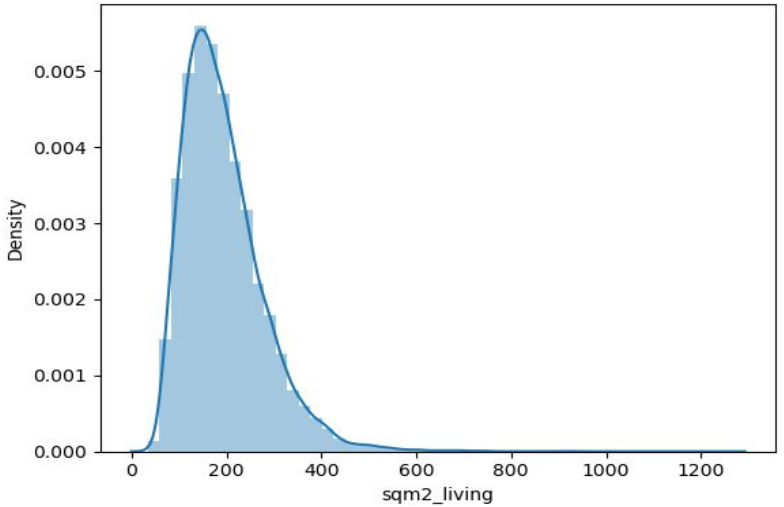
	id	date	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront
0	7129300520	10/13/14	3	1.00	1180	5650	1.0	0
1	6414100192	12/9/14	3	2.25	2570	7242	2.0	0
2	5631500400	2/25/15	2	1.00	770	10000	1.0	0
3	2487200875	12/9/14	4	3.00	1960	5000	1.0	0
4	1954400510	2/18/15	3	2.00	1680	8080	1.0	0

Bucketing

- Views - Viewed or not
- Property Age - Age group
- Renovated - renovated or not
- Quarter - months into quarters
- Fortnight - days into fortnights
- Bigger or not - if the lot or the living area became bigger after renovation or not

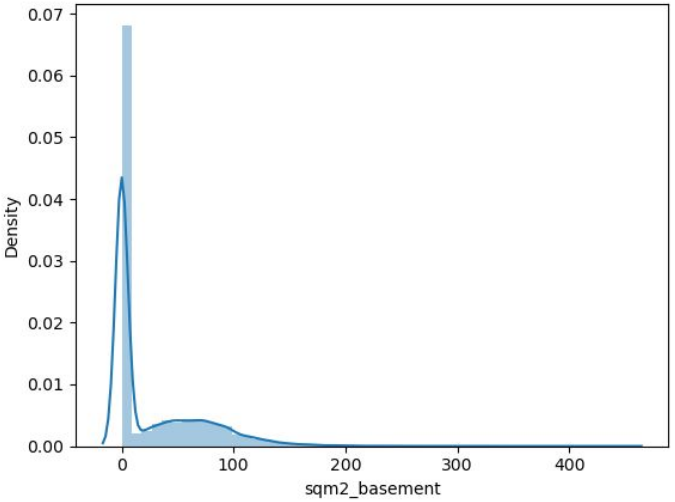
Exploratory Data Analysis

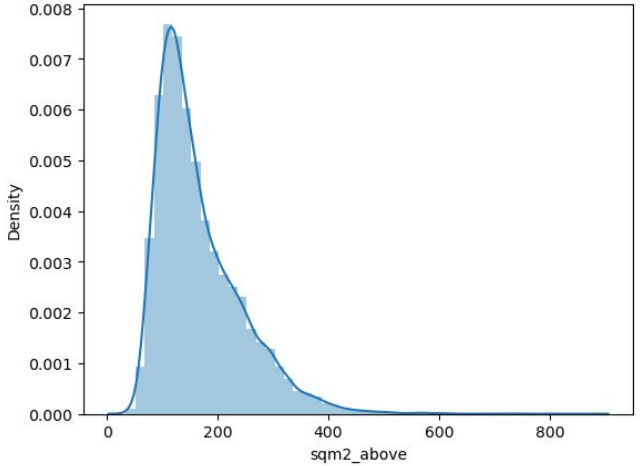




Continuous Numerical

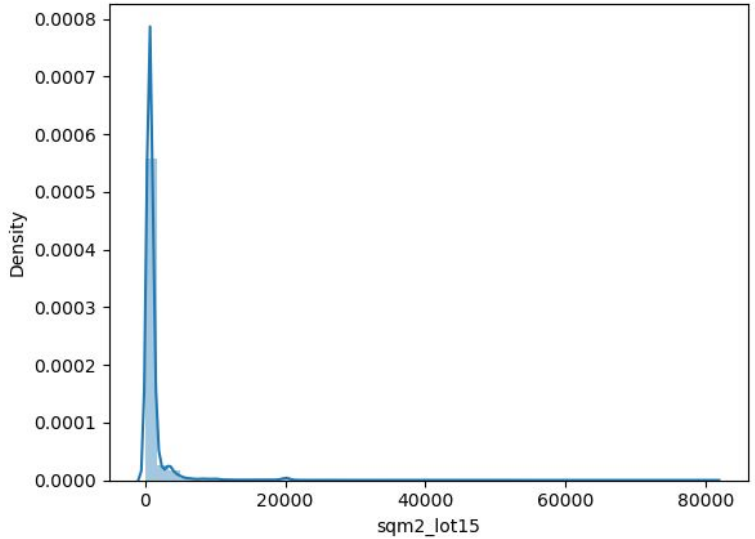
Exploratory Data Analysis

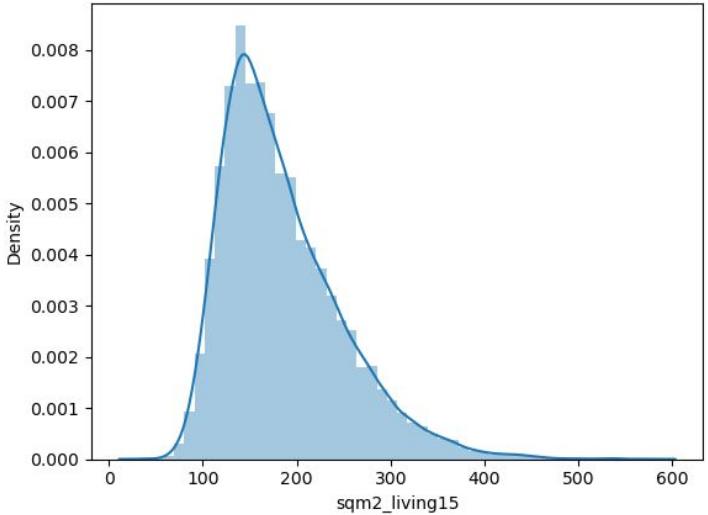




Continuous Numerical

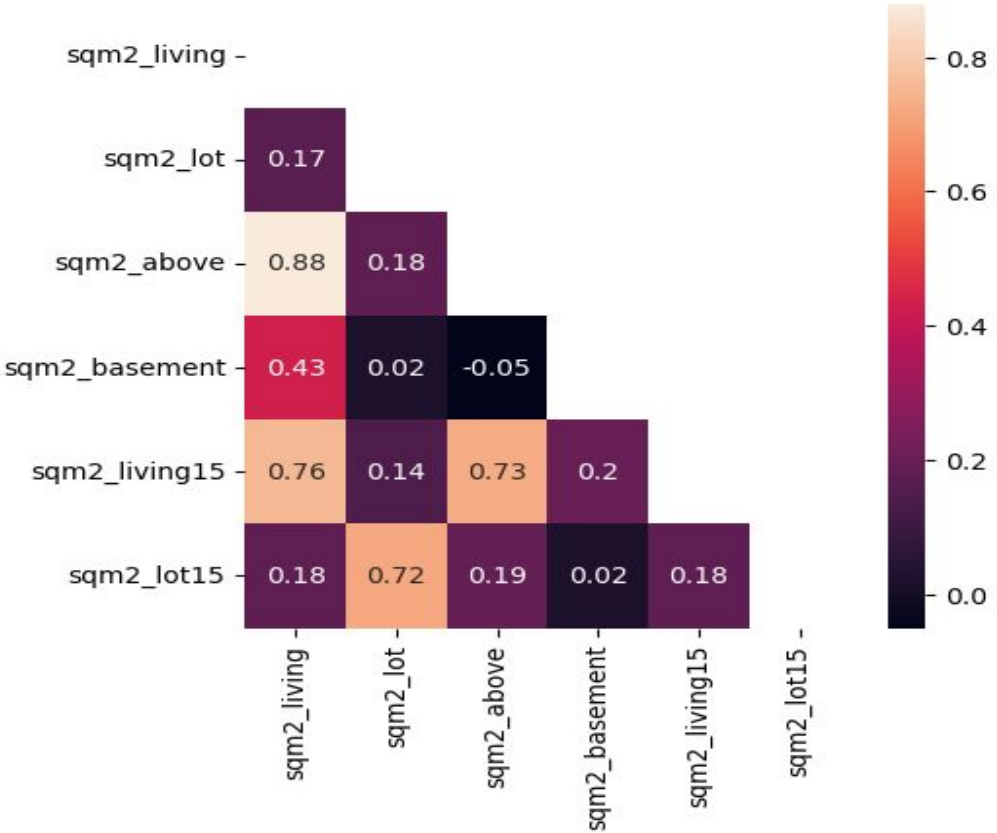
Exploratory Data Analysis





Continuous Numerical

Correlation Matrix



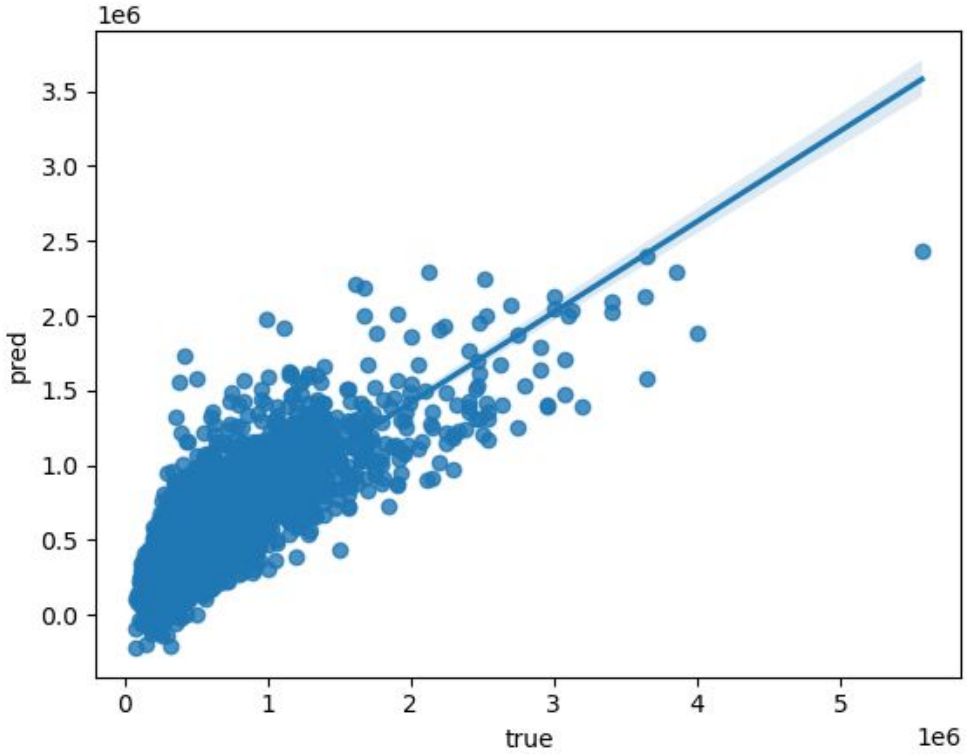
Regression Metrics
R2 Score

Linear Regression	KNN Regressor	Random Forest Regressor
0,37 - without outliers 0,64 0,30 - BoxCox	0,51 0,28 - without outliers 0,27 - Box Cox	0,73

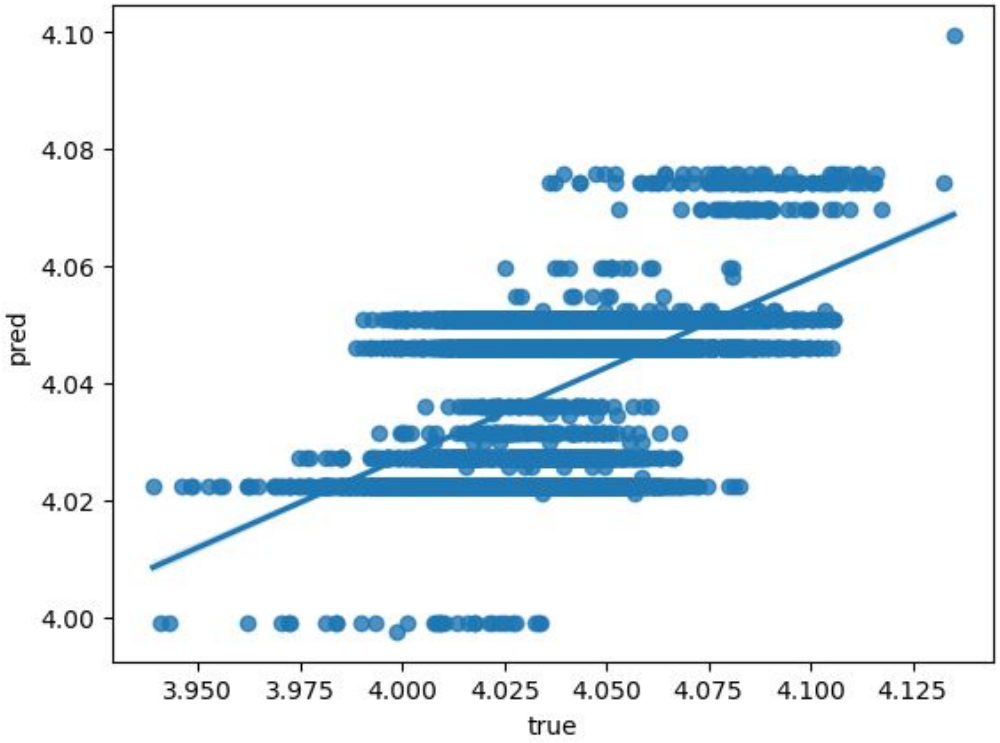
Visualizing Models

Linear Regression

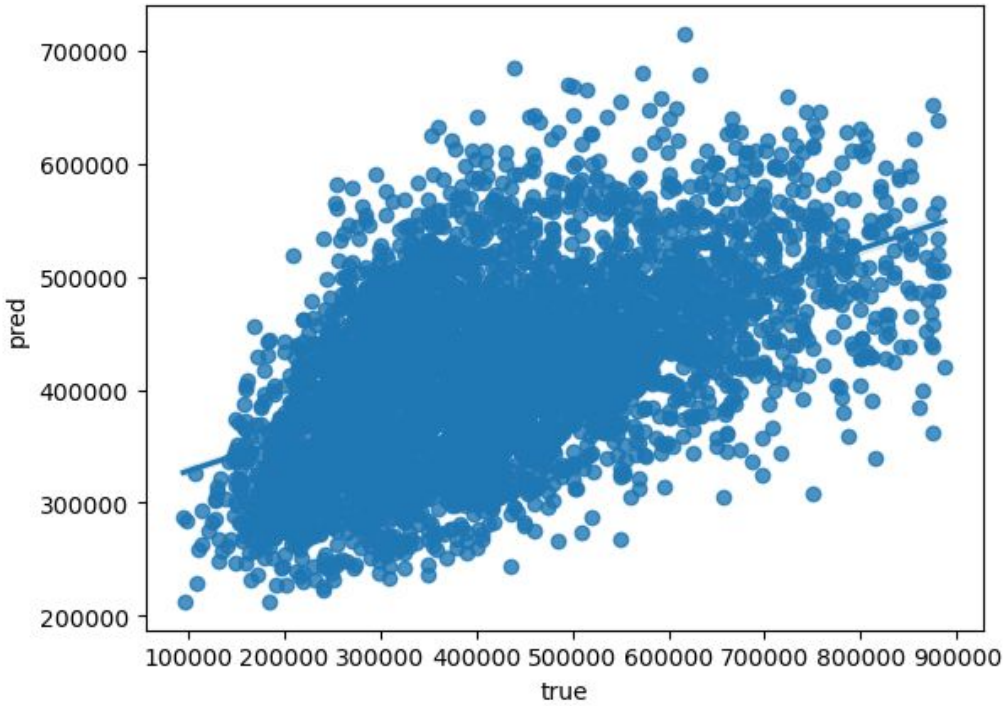
Scalers but without Transformers



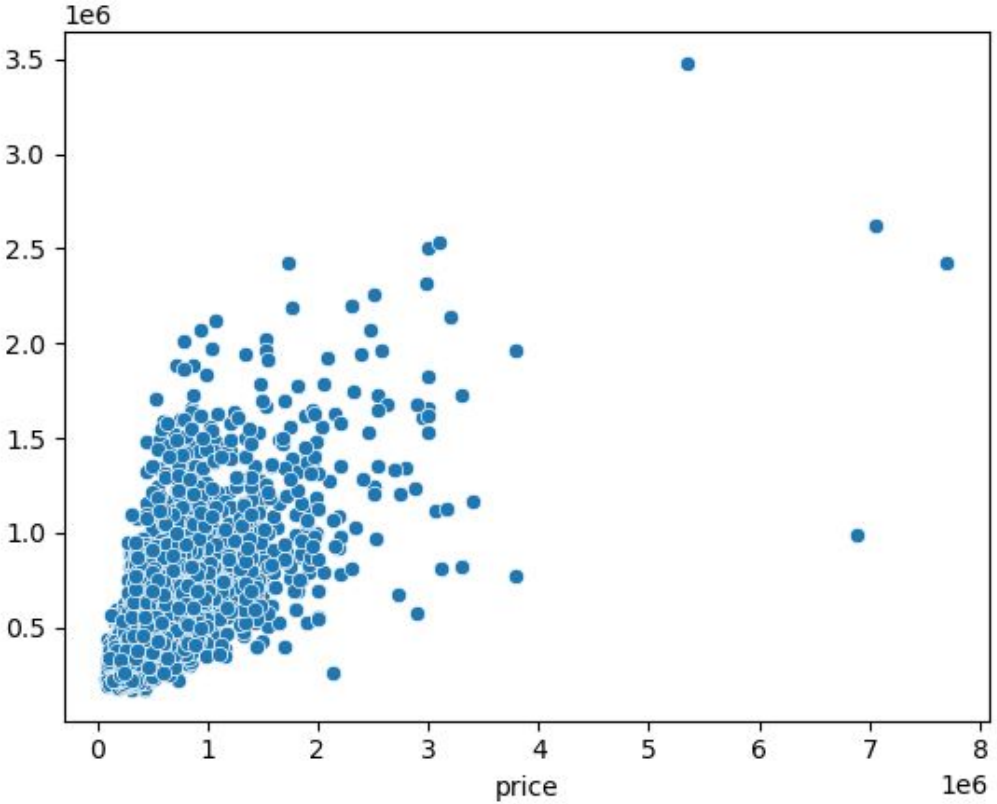
Linear Regression
Box-Cox



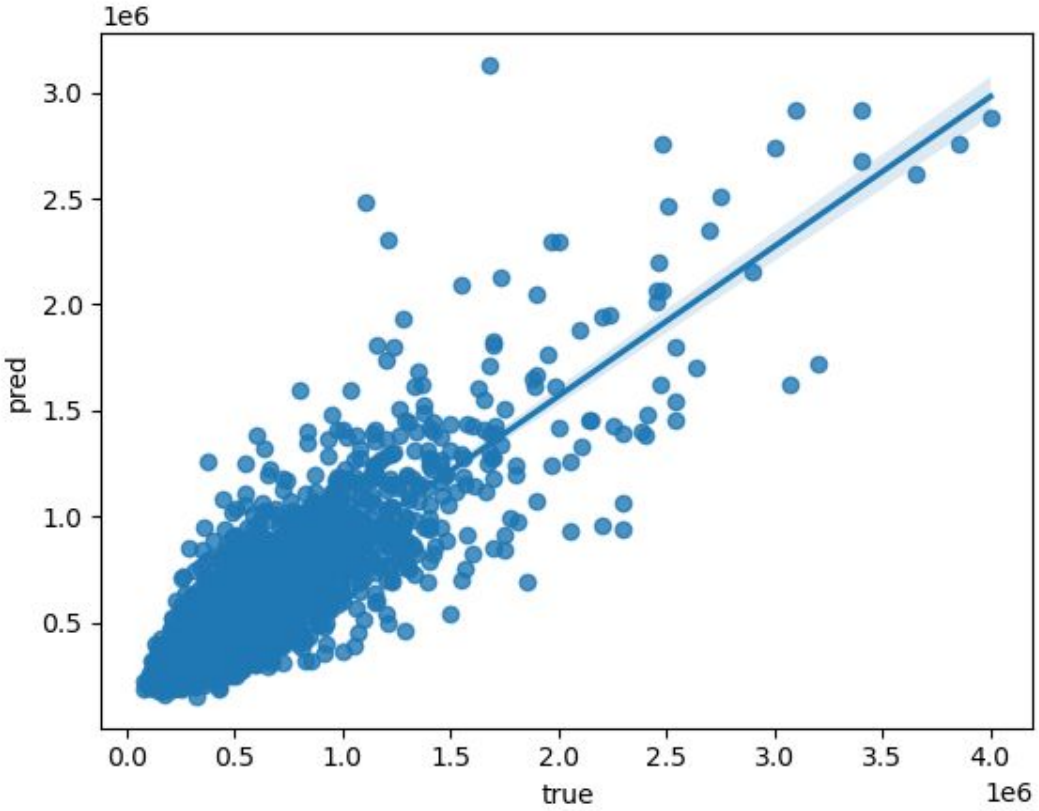
Linear Regression
Outliers Removed



KNN Regression



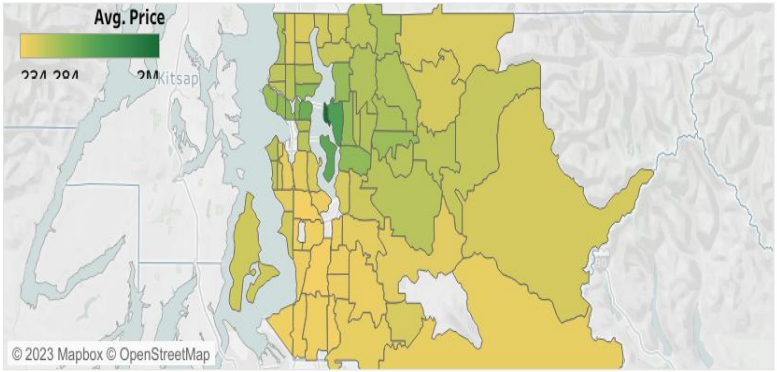
Random Forest Regression



Why RandomForest?

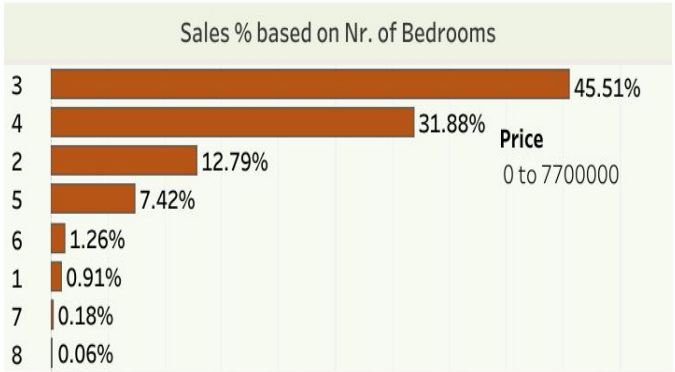
1. The Random Forest Model is a great tool to use to improve predictive models.
2. Adding more trees can improve the performance of the model, but decreases the returns, and too many trees can lead to overfitting or slow training/prediction times.
3. It uses Cross-Validation, a useful method that splits the data in a training and validation sets. It finds the number of trees that give the best performance, by trying out different values and evaluating performance on a validation set.
4. Preprocessing the data and considering feature selection/engineering techniques are also important for optimizing random forest performance.

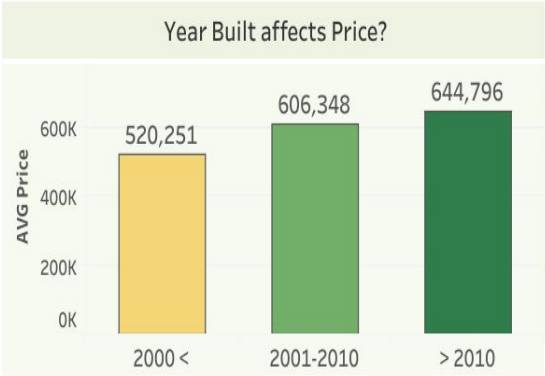
King County Housing Dashboard

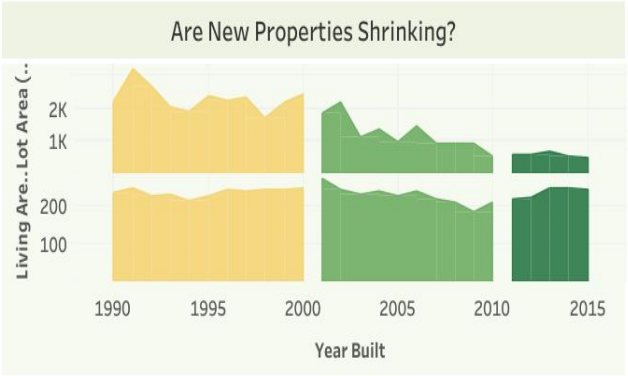


Sales by Condition	
Average	91.2%
Good	7.9%

Sales by Grade	
7-9	81.7%
10-13	7.6%







https://public.tableau.com/app/profile/sandra.cunha/viz/MIDPROJECT_V2/10_DASHBOARD?publish=yes

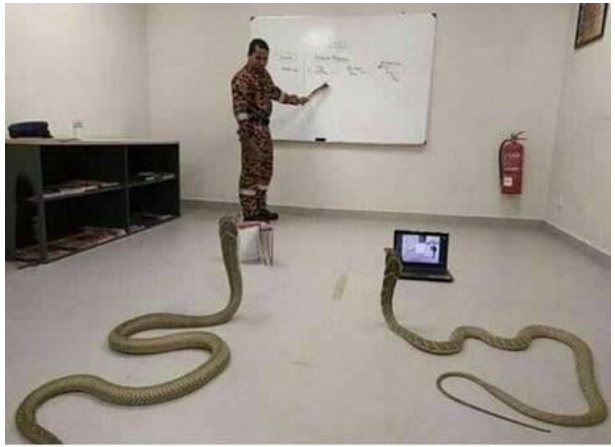
What to look for if Real Estate Company wants to improve sales in King County?

Add to Real Estate Property Portfolio:

- 3 - 4 Bedrooms Houses;
- Prioritize Living Area (Lot Area has been decreasing over time while Living Area remains the same);
- Grade (7-9 higher selling rates);
- New properties instead of Old ones;

Thank You for Listening!

Questions?



tiger-in-the-flightdeck

The lack of context here is thrilling

mark-watney-spacepirate

introductory python programming course