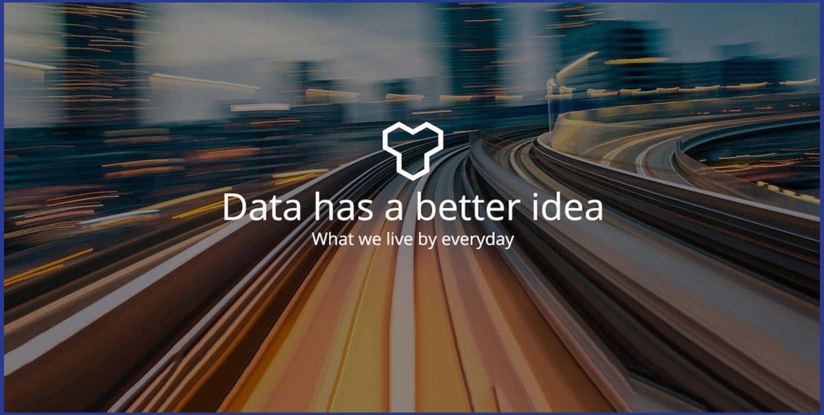


Comprehensive Data Analysis

Data Science Career Track
Egor Petrov

Agenda

Welcome Message

About

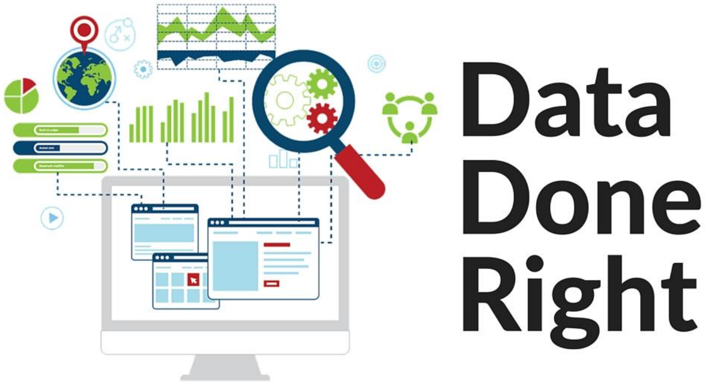
Team

Introduction to a project

Final Thoughts

Defining a Project

Main Course Project for Developing Data Science Skills



The problem

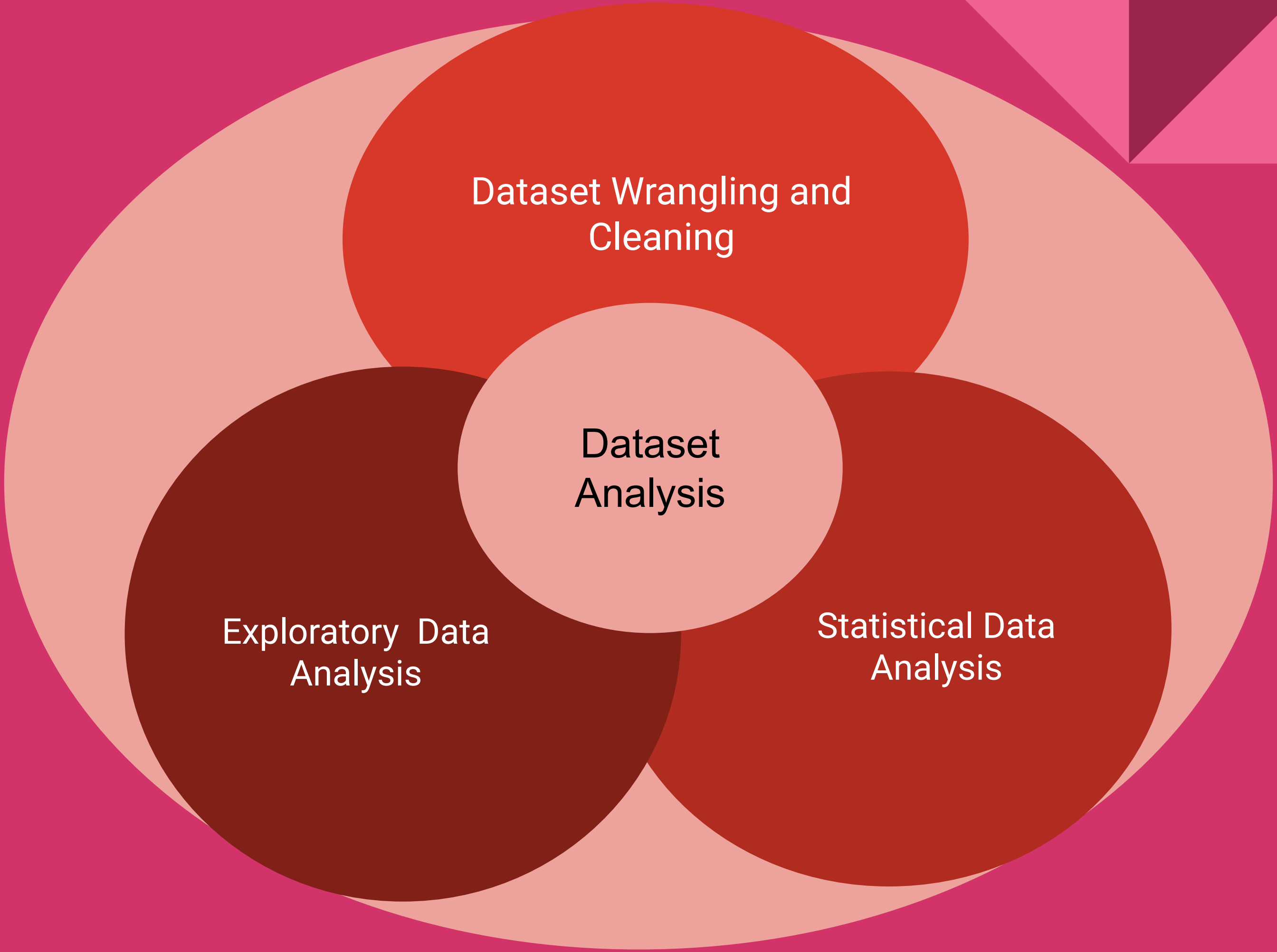
Predicting a Sale Price	Problem statement
My project is focused on challenges to predict the final price of each home.	This is too hard or impossible to calculate the final price. Each house is different and built from so many different structures. Each house has a different zone classification and many more differences. In my project I will show how this can be solved with a Data Scientist expertise.

Solution

Train your ML model!

Use of new technologies.
New Python + AI
Very new approach!
Let the computer do the service for you!

Implementation



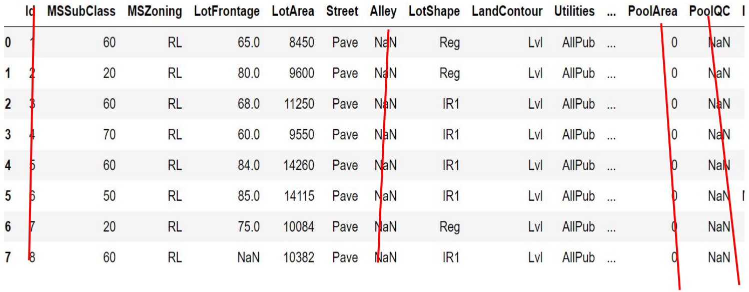
Dataset Wrangling and Cleaning

Dataset Analysis

Exploratory Data Analysis

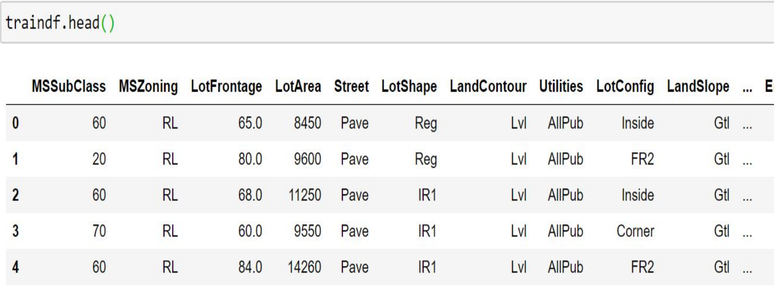
Statistical Data Analysis

Data Wrangling Results



	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	I
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	
5	6	50	RL	85.0	14115	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	I
6	7	20	RL	75.0	10084	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	
7	8	60	RL	NaN	10382	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	

Initially Data comes as a mess. It can have a lot of missing values in it.

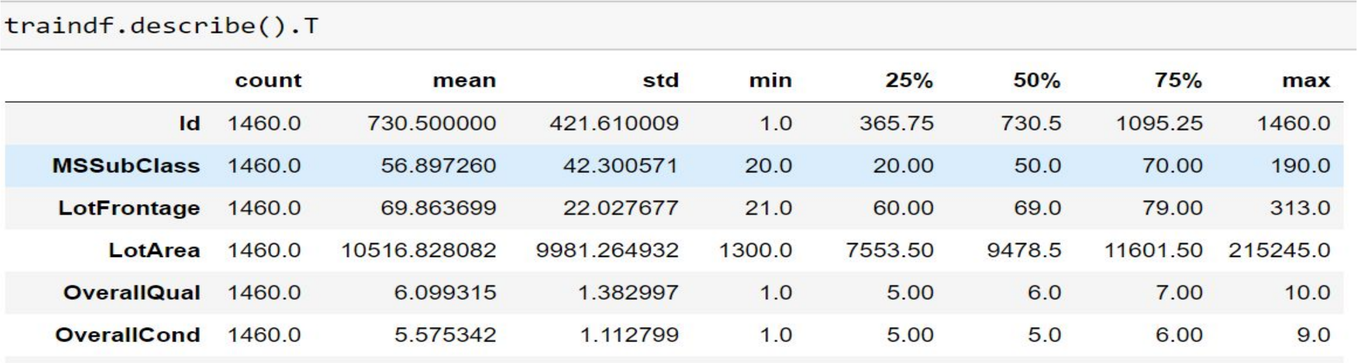


```
traindf.head()
```

	MSSubClass	MSZoning	LotFrontage	LotArea	Street	LotShape	LandContour	Utilities	LotConfig	LandSlope	...	E
0	60	RL	65.0	8450	Pave	Reg	Lvl	AllPub	Inside	Gtl	...	
1	20	RL	80.0	9600	Pave	Reg	Lvl	AllPub	FR2	Gtl	...	
2	60	RL	68.0	11250	Pave	IR1	Lvl	AllPub	Inside	Gtl	...	
3	70	RL	60.0	9550	Pave	IR1	Lvl	AllPub	Corner	Gtl	...	
4	60	RL	84.0	14260	Pave	IR1	Lvl	AllPub	FR2	Gtl	...	

Data Wrangling techniques used by a data scientist will update the same dataset as a complete DataFrame

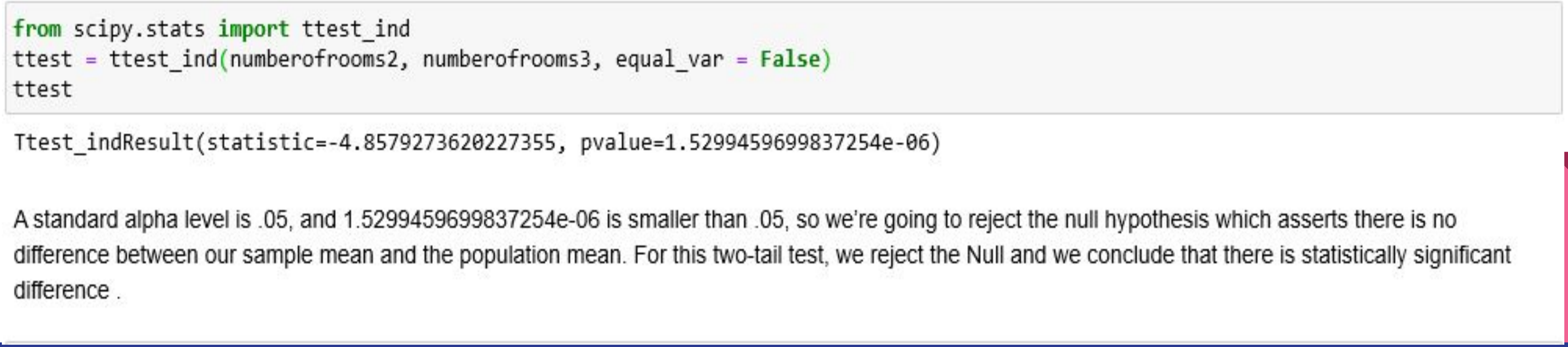
Statistical Analysis Results



```
traindf.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Id	1460.0	730.500000	421.610009	1.0	365.75	730.5	1095.25	1460.0
MSSubClass	1460.0	56.897260	42.300571	20.0	20.00	50.0	70.00	190.0
LotFrontage	1460.0	69.863699	22.027677	21.0	60.00	69.0	79.00	313.0
LotArea	1460.0	10516.828082	9981.264932	1300.0	7553.50	9478.5	11601.50	215245.0
OverallQual	1460.0	6.099315	1.382997	1.0	5.00	6.0	7.00	10.0
OverallCond	1460.0	5.575342	1.112799	1.0	5.00	5.0	6.00	9.0

Hypothesis testing

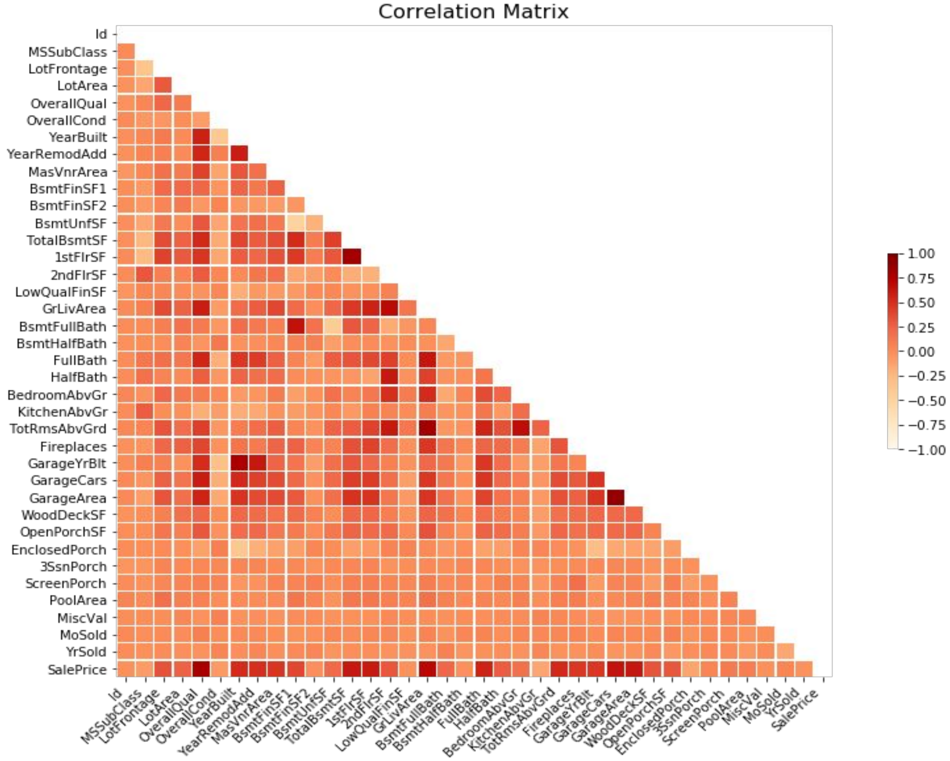


```
from scipy.stats import ttest_ind
ttest = ttest_ind(numberofrooms2, numberofrooms3, equal_var = False)
ttest
```

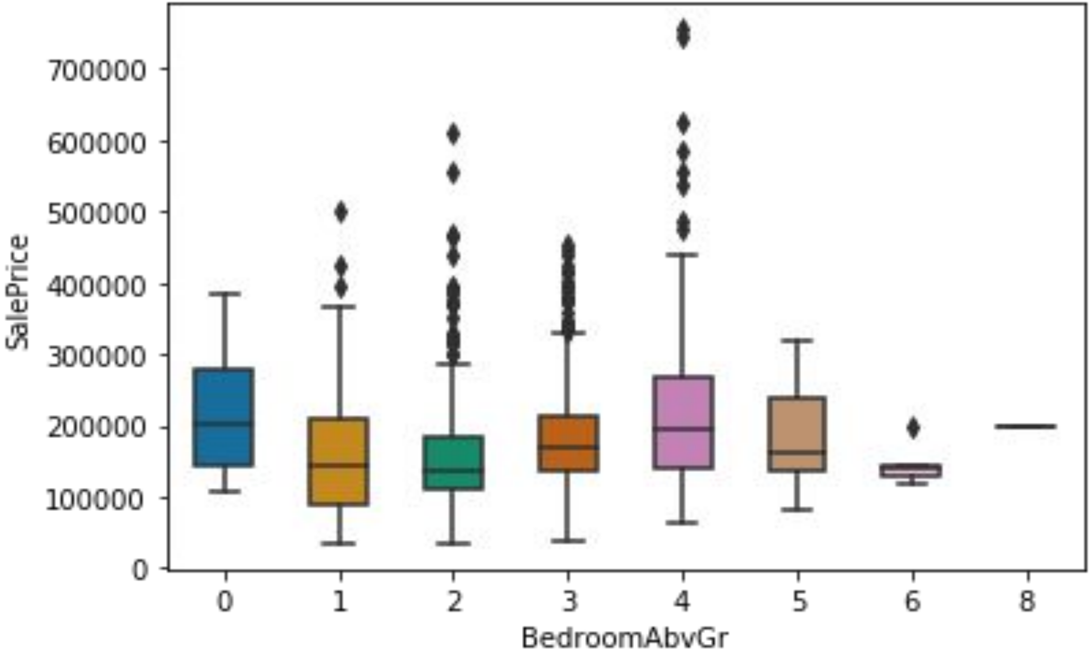
```
Ttest_indResult(statistic=-4.8579273620227355, pvalue=1.5299459699837254e-06)
```

A standard alpha level is .05, and 1.5299459699837254e-06 is smaller than .05, so we're going to reject the null hypothesis which asserts there is no difference between our sample mean and the population mean. For this two-tail test, we reject the Null and we conclude that there is statistically significant difference .

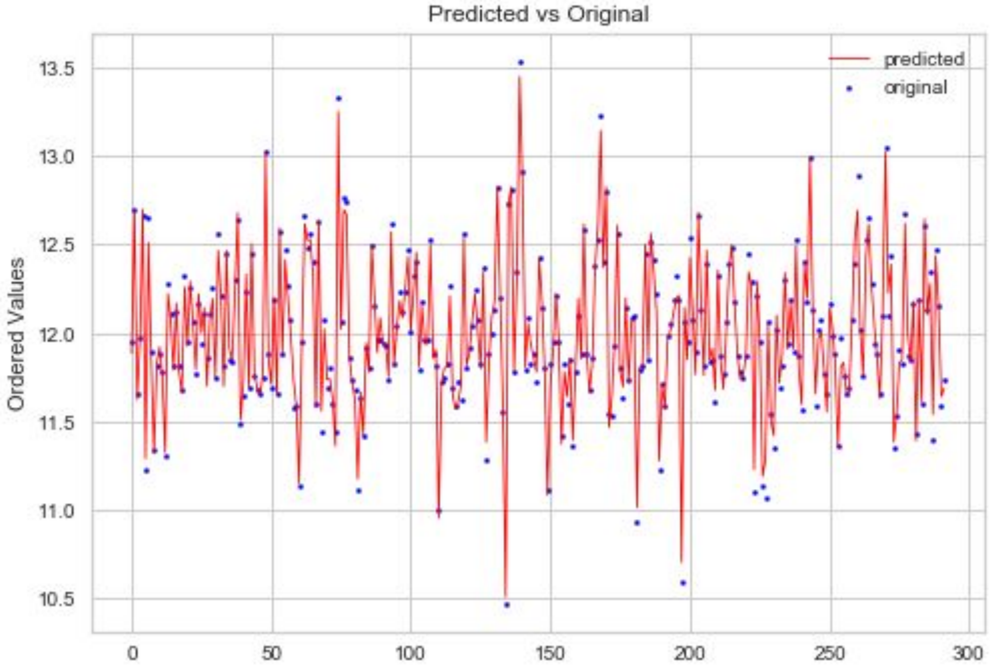
Visualizations



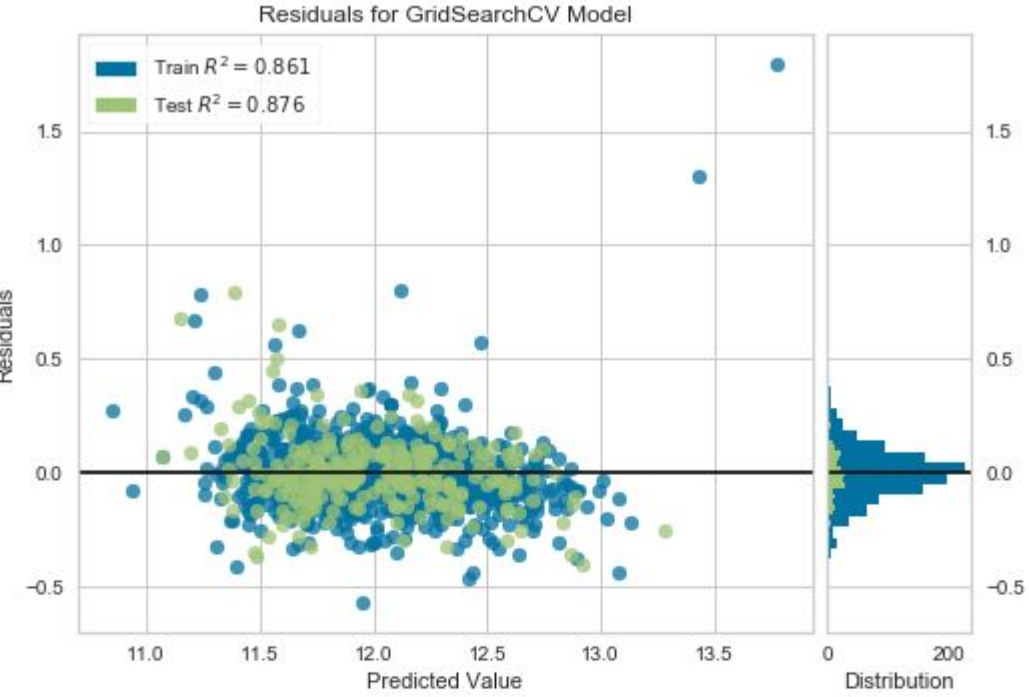
Visualizations



Training a model



Visualizations



Metrics

Ridge: RMSE scores mean: 0.143

Ridge after removing not correlated features: RMSE scores mean: 0.143

Pipeline Ridge: RMSE scores mean: 0.141

Random Forest: RMSE scores mean: 0.144

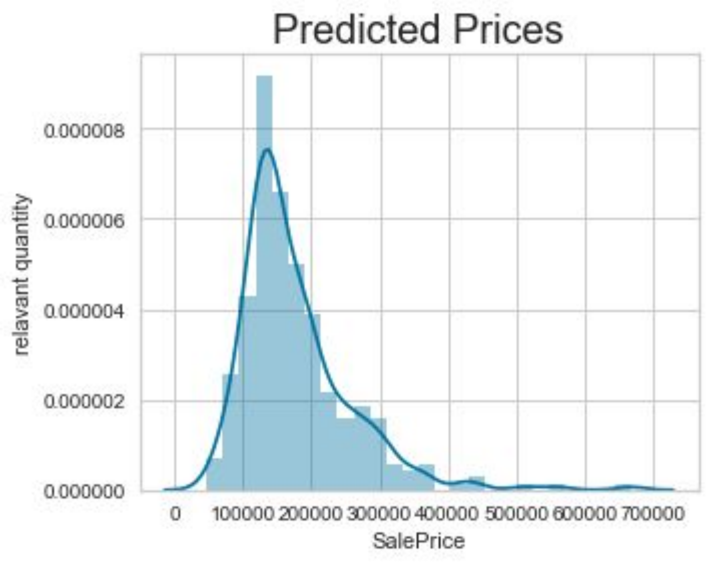
Random Forest after removing NOT correlated features: RMSE scores mean: 0.144

XGBoost : RMSE scores mean: 0.130

XGBoost after removing NOT correlated features: RMSE scores mean: 0.130

Target

Predicted Prices



Final Thoughts

XGBoost : RMSE scores mean: 0.130 is low error rate which is good indicator of a good model and the best among the list of models I tested. This makes me pick XGBoost in my case.

A ML model is fun, fast and accurate.
Realtors and Construction businesses are encouraged to take advantage of training a ML model and expertise of a Data Scientist. This is crucial to adapt new technologies and be ahead of competitors.

More Final Thoughts

End Use:

As a standalone application to calculate new predicted prices on each house and correlation of prices.

For education end use, teachers are advised to use this project as a textbook and a resource to tell students about house selling factors.

Possible Improvements:

Lasso and Elastic regressions. And a Robust regression in case if I deal with a lot of outliers in my dataset. I want to eliminate as many outliers as possible by applying weights to my variables or residuals.