HOUSING : PRICE PREDICTION

By, UTKARSH VARDHAN

Background and Problem Statement:

- Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world's economy.
- It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies.
- A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price.

Background and Problem Statement:

- For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV format.
- The company is looking at prospective properties to buy houses to enter the market.
- We are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not.
- For this company wants to know:
    - Which variables are important to predict the price of variable?
    - How do these variables describe the price of the house?

Business Goal

- We are required to model the price of houses with the available independent variables.
- This model will then be used by the management to understand how exactly the prices vary with the variables.
- They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns.
- Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

Data Shared

- Data contains 1460 entries each having 81 variables.
- Data contains Null values.
- Data contains numerical as well as categorical variable.
- Data File is in csv format

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	La
0	1	60	RL	65.0	8450	Pave	NaN	Reg	
1	2	20	RL	80.0	9600	Pave	NaN	Reg	
2	3	60	RL	68.0	11250	Pave	NaN	IR1	
3	4	70	RL	60.0	9550	Pave	NaN	IR1	
4	5	60	RL	84.0	14260	Pave	NaN	IR1	
...	...	...	...	...	...	...	...	...	
1455	1456	60	RL	62.0	7917	Pave	NaN	Reg	
1456	1457	20	RL	85.0	13175	Pave	NaN	Reg	
1457	1458	70	RL	66.0	9042	Pave	NaN	Reg	
1458	1459	20	RL	68.0	9717	Pave	NaN	Reg	
1459	1460	20	RL	75.0	9937	Pave	NaN	Reg	

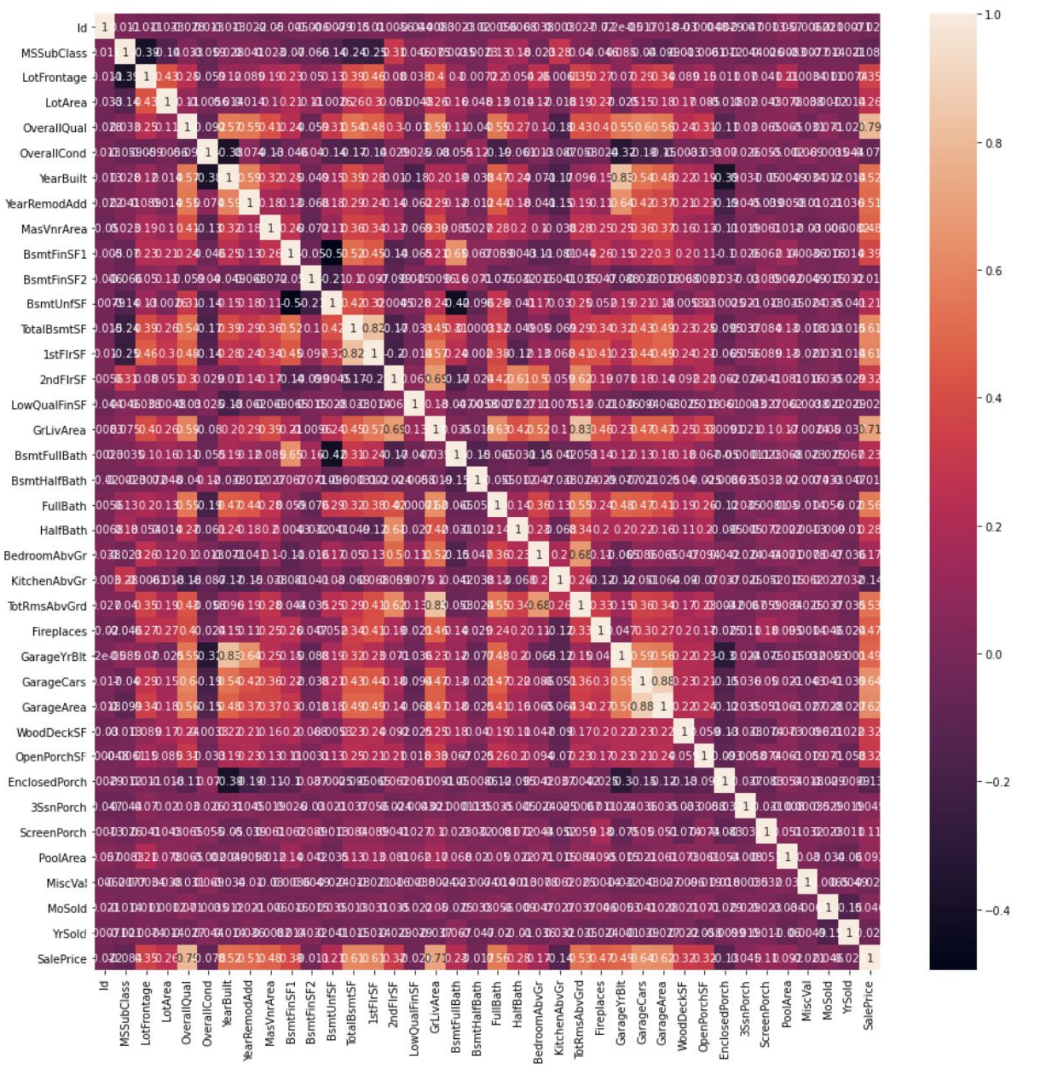
1460 rows × 81 columns

Data Cleaning

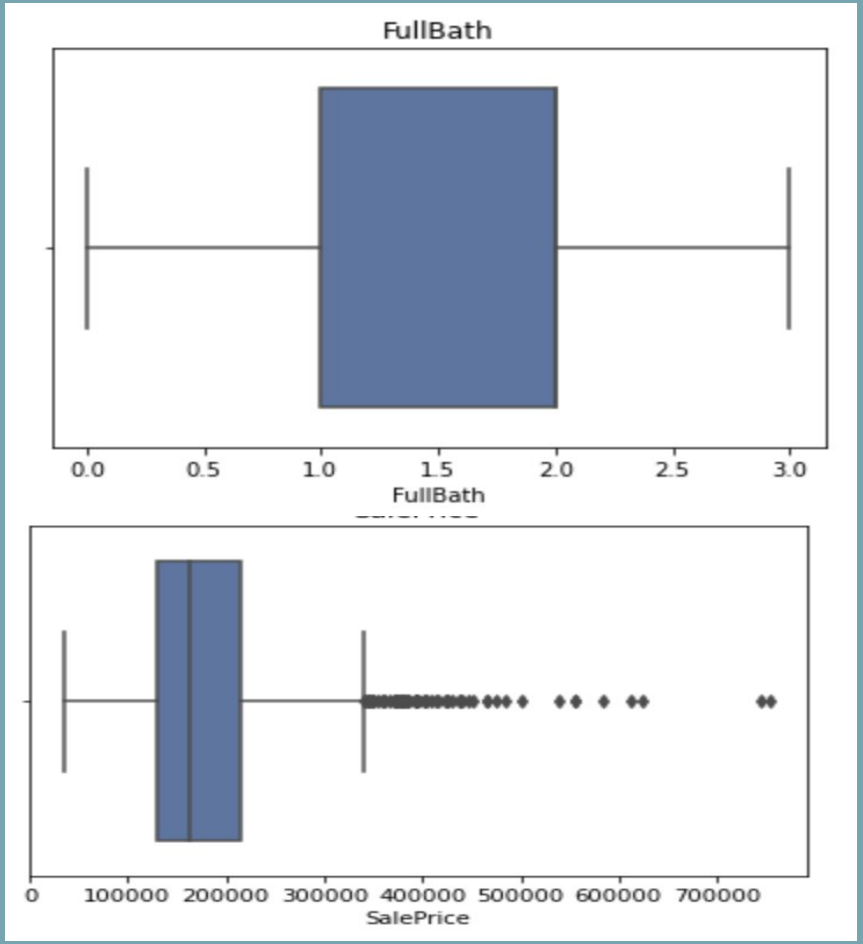
- Checking Shapes and Data Types
- Checking Uniqueness
- Dropping Columns
- Checking Missing Values
- Checking for correlation Features
- Removing Outliers
- PCA

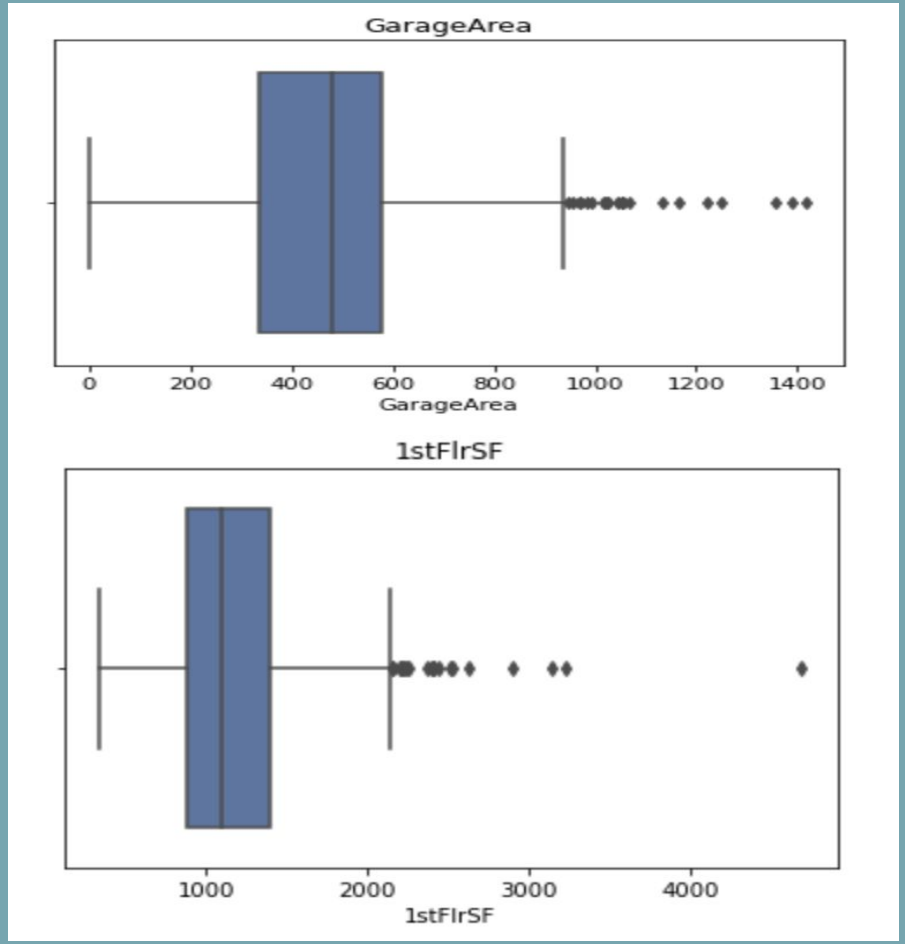
Data Visualization:

SalePrice	1.000000
OverallQual	0.790982
GrLivArea	0.708624
GarageCars	0.640409
GarageArea	0.623431
TotalBsmtSF	0.613581
1stFlrSF	0.605852
FullBath	0.560664
TotRmsAbvGrd	0.533723
YearBuilt	0.522897
YearRemodAdd	0.507101
GarageYrBlt	0.486362
MasVnrArea	0.477493
Fireplaces	0.466929
BsmtFinSF1	0.386420
LotFrontage	0.351799
WoodDeckSF	0.324413
2ndFlrSF	0.319334
OpenPorchSF	0.315856
HalfBath	0.284108
LotArea	0.263843
BsmtFullBath	0.227122
BsmtUnfSF	0.214479
BedroomAbvGr	0.168213
ScreenPorch	0.111447
PoolArea	0.092404
MoSold	0.046432
3SsnPorch	0.044584
BsmtFinSF2	-0.011378
BsmtHalfBath	-0.016844
MiscVal	-0.021190
Id	-0.021917
LowQualFinSF	-0.025606
YrSold	-0.028923
OverallCond	-0.077856
MSSubClass	-0.084284
EnclosedPorch	-0.128578
KitchenAbvGr	-0.135907



Data Visualization:





MODEL TESTING:

MODELS	ACCURACY	CROSS VAL SCORE
DecisionTreeRegressor	76.51%	59.65%
RandomForestRegressor	97.08%	73.38%
Lasso	92.24%	88.6%
KNeighborsRegressor	75.48%	66.36%
LinearRegression	92.24%	88.61%

THANK YOU