Project Name: Housing Price Prediction

Submitted by:

Dhrubajyoti Mandal

Acknowledgement

The success and final outcome of the machine learning requires a lot of guidance and assistance from some people and I am extremely privileged to have got this all among the completion of my course and few of the projects. All that I have done is only due to such supervision and assistance and I would not forget to thank them.

I respect and thank FLIP ROBO Technologies, for providing me this opportunity to do the carcerand project work and giving me all support and guidance, which made me complete the course.

I would like to thanks my mentor, Sapna Verma who guided me at every point of the project.

Introduction to Problem

AIM and IMPORTANCE

Aim These are the Parameters on which we will evaluate ourselves-

- Create an effective price prediction model
- Validate the model's prediction accuracy
- Identify the important home price attributes which feed the model's predictive power.

Business Problem Framing

Houses are one of the necessary needs 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. Our problem is related to one such housing company.

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. 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 file below.

The company is looking at prospective properties to buy houses to enter the market. You 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:

You 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.

Technical Requirements:

* Data contains 1460 entries each having 81 variables.

* Data contains Null values. You need to treat them using the domain knowledge and your own understanding.

* Extensive EDA has to be performed to gain relationships of important variable and price.

* Data contains numerical as well as categorical variable. You need to handle them accordingly.

* You have to build Machine Learning models, apply regularization and determine the optimal values of Hyper Parameters.

* You need to find important features which affect the price positively or negatively.

* Two datasets are being provided to you (test.csv, train.csv). You will train on train.csv dataset and predict on test.csv file.

General Description on features

MSSubClass: Identifies the type of dwelling involved in the sale.

20	1-STORY 1946 & NEWER ALL STYLES
30	1-STORY 1945 & OLDER
40	1-STORY W/FINISHED ATTIC ALL AGES
45	1-1/2 STORY - UNFINISHED ALL AGES
50	1-1/2 STORY FINISHED ALL AGES
60	2-STORY 1946 & NEWER
70	2-STORY 1945 & OLDER
75	2-1/2 STORY ALL AGES
80	SPLIT OR MULTI-LEVEL
85	SPLIT FOYER
90	DUPLEX - ALL STYLES AND AGES
120	1-STORY PUD (Planned Unit Development) - 1946 & NEWER
150	1-1/2 STORY PUD - ALL AGES
160	2-STORY PUD - 1946 & NEWER
180	PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
190	2 FAMILY CONVERSION - ALL STYLES AND AGES

MSZoning: Identifies the general zoning classification of the sale.

A	Agriculture
C	Commercial
FV	Floating Village Residential
I	Industrial

RH Residential High Density

RL Residential Low Density

RP Residential Low Density Park

RM Residential Medium Density

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access to property

Grvl Gravel

Pave Paved

Alley: Type of alley access to property

Grvl Gravel

Pave Paved

NA No alley access

LotShape: General shape of property

Reg Regular

IR1 Slightly irregular

IR2 Moderately Irregular

IR3 Irregular

LandContour: Flatness of the property

Lvl Near Flat/Level

Bnk Banked - Quick and significant rise from street grade to building

HLS Hillside - Significant slope from side to side

Low Depression

Utilities: Type of utilities available

AllPub All public Utilities (E,G,W,& S)

NoSewr Electricity, Gas, and Water (Septic Tank)

NoSeWa Electricity and Gas Only

ELO Electricity only

LotConfig: Lot configuration

Inside Inside lot

Corner Corner lot

CulDSac Cul-de-sac

FR2 Frontage on 2 sides of property

FR3 Frontage on 3 sides of property

LandSlope: Slope of property

Gtl Gentle slope

Mod Moderate Slope

Sev Severe Slope

Neighborhood: Physical locations within Ames city limits

Blmngtn Bloomington Heights

Blueste Bluestem

BrDale Briardale

BrkSide Brookside

ClearCr Clear Creek

CollgCr College Creek

Crawfor Crawford

Edwards Edwards

Gilbert Gilbert

IDOTRR Iowa DOT and Rail Road

MeadowV Meadow Village

Mitchel Mitchell

Names North Ames

NoRidge Northridge

NPkVillNorthpark Villa

NridgHt Northridge Heights

NWAmes Northwest Ames

OldTown Old Town

SWISU South & West of Iowa State University

Sawyer Sawyer

SawyerW Sawyer West

Somerst Somerset

StoneBr Stone Brook

Timber Timberland

Veenker Veenker

Condition1: Proximity to various conditions

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe	Adjacent to East-West Railroad

Condition2: Proximity to various conditions (if more than one is present)

Artery	Adjacent to arterial street

Feedr	Adjacent to feeder street

Norm	Normal

RRNn	Within 200' of North-South Railroad

RRAn	Adjacent to North-South Railroad

PosN	Near positive off-site feature--park, greenbelt, etc.

PosA	Adjacent to postive off-site feature

RRNe	Within 200' of East-West Railroad

RRAe	Adjacent to East-West Railroad

BldgType: Type of dwelling

1Fam	Single-family Detached

2FmCon	Two-family Conversion; originally built as one-family dwelling

Duplx	Duplex

TwnhsE	Townhouse End Unit

TwnhsI	Townhouse Inside Unit

HouseStyle: Style of dwelling

1Story	One story

1.5Fin	One and one-half story: 2nd level finished

1.5Unf	One and one-half story: 2nd level unfinished

2Story	Two story

2.5Fin	Two and one-half story: 2nd level finished

2.5Unf	Two and one-half story: 2nd level unfinished

SFoyer Split Foyer

SLvl Split Level

OverallQual: Rates the overall material and finish of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

OverallCond: Rates the overall condition of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

RoofStyle: Type of roof

Flat Flat

Gable Gable

Gambrel Gabrel (Barn)

Hip Hip

Mansard Mansard

Shed Shed

RoofMatl: Roof material

ClyTile Clay or Tile

CompShg Standard (Composite) Shingle

Membran Membrane

Metal Metal

Roll Roll

Tar&Grv Gravel & Tar

WdShake Wood Shakes

WdShngl Wood Shingles

Exterior1st: Exterior covering on house

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

Exterior2nd: Exterior covering on house (if more than one material)

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone	Stone

Stucco	Stucco

VinylSd	Vinyl Siding

Wd Sdng	Wood Siding

WdShing	Wood Shingles

MasVnrType: Masonry veneer type

BrkCmn	Brick Common

BrkFace	Brick Face

CBlock	Cinder Block

None	None

Stone	Stone

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior

Ex	Excellent

Gd	Good

TA	Average/Typical

Fa	Fair

Po	Poor

ExterCond: Evaluates the present condition of the material on the exterior

Ex	Excellent

Gd	Good

TA	Average/Typical

Fa	Fair

Po	Poor

Foundation: Type of foundation

BrkTil Brick & Tile

CBlock Cinder Block

PConc Poured Contrete

Slab Slab

Stone Stone

Wood Wood

BsmtQual: Evaluates the height of the basement

Ex Excellent (100+ inches)

Gd Good (90-99 inches)

TA Typical (80-89 inches)

Fa Fair (70-79 inches)

Po Poor (<70 inches

NA No Basement

BsmtCond: Evaluates the general condition of the basement

Ex Excellent

Gd Good

TA Typical - slight dampness allowed

Fa Fair - dampness or some cracking or settling

Po Poor - Severe cracking, settling, or wetness

NA No Basement

BsmtExposure: Refers to walkout or garden level walls

Gd Good Exposure

Av Average Exposure (split levels or foyers typically score average or above)

Mn Mimimum Exposure

No No Exposure

NA No Basement

BsmtFinType1: Rating of basement finished area

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

NA No Basement

BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Rating of basement finished area (if multiple types)

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

NA No Basement

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

Heating: Type of heating

Floor Floor Furnace

GasA Gas forced warm air furnace

GasW Gas hot water or steam heat

Grav Gravity furnace

OthW Hot water or steam heat other than gas

Wall Wall furnace

HeatingQC: Heating quality and condition

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

CentralAir: Central air conditioning

N No

Y Yes

Electrical: Electrical system

SBrkr Standard Circuit Breakers & Romex

FuseA Fuse Box over 60 AMP and all Romex wiring (Average)

FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)

FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)

Mix Mixed

1stFlrSF: First Floor square feet

2ndFlrSF: Second floor square feet

LowQualFinSF: Low quality finished square feet (all floors)

GrLivArea: Above grade (ground) living area square feet

BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms

FullBath: Full bathrooms above grade

HalfBath: Half baths above grade

Bedroom: Bedrooms above grade (does NOT include basement bedrooms)

Kitchen: Kitchens above grade

KitchenQual: Kitchen quality

Ex	Excellent
Gd	Good
TA	Typical/Average
Fa	Fair
Po	Poor

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

Functional: Home functionality (Assume typical unless deductions are warranted)

Typ	Typical Functionality
Min1	Minor Deductions 1
Min2	Minor Deductions 2
Mod	Moderate Deductions
Maj1	Major Deductions 1
Maj2	Major Deductions 2
Sev	Severely Damaged
Sal	Salvage only

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

Ex	Excellent - Exceptional Masonry Fireplace

Gd	Good - Masonry Fireplace in main level

TA	Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement

Fa	Fair - Prefabricated Fireplace in basement

Po	Poor - Ben Franklin Stove

NA	No Fireplace

GarageType: Garage location

2Types	More than one type of garage

Attchd	Atached to home

Basment	Basement Garage

BuiltIn	Built-In (Garage part of house - typically has room above garage)

CarPort	Car Port

Detchd	Detached from home

NA	No Garage

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

Fin	Finished

RFn	Rough Finished

Unf	Unfinished

NA	No Garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

GarageCond: Garage condition

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

PavedDrive: Paved driveway

Y Paved

P Partial Pavement

NDirt/Gravel

WoodDeckSF: Wood deck area in square feet

OpenPorchSF: Open porch area in square feet

EnclosedPorch: Enclosed porch area in square feet

3SsnPorch: Three season porch area in square feet

ScreenPorch: Screen porch area in square feet

PoolArea: Pool area in square feet

PoolQC: Pool quality

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
NA	No Pool

Fence: Fence quality

GdPrv	Good Privacy
MnPrv	Minimum Privacy
GdWo	Good Wood
MnWw	Minimum Wood/Wire
NA	No Fence

MiscFeature: Miscellaneous feature not covered in other categories

Elev	Elevator
Gar2	2nd Garage (if not described in garage section)
Othr	Other
Shed	Shed (over 100 SF)
TenC	Tennis Court
NA	None

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

SaleType: Type of sale

WD	Warranty Deed - Conventional

CWD Warranty Deed - Cash

VWD Warranty Deed - VA Loan

New Home just constructed and sold

COD Court Officer Deed/Estate

Con Contract 15% Down payment regular terms

ConLw Contract Low Down payment and low interest

ConLI Contract Low Interest

ConLD Contract Low Down

Oth Other

SaleCondition: Condition of sale

Normal Normal Sale

Abnorml Abnormal Sale - trade, foreclosure, short sale

AdjLand Adjoining Land Purchase

Alloca Allocation - two linked properties with separate deeds, typically condo with a garage unit

Family Sale between family members

Partial Home was not completed when last assessed (associated with New Homes)

Need and Motivation

Having lived in India for so many years if there is one thing that I had been taking for granted, it's those housing and rental prices continue to rise. Since the housing crisis, housing prices have recovered remarkably well, especially in major housing markets. However, in the 4th quarter of 2016, I was surprised to read that US housing prices had fallen the most in the last 4 years. In fact, median resale prices for condos and coops fell 6.3%, marking the first time there was a decline since Q1 of 2017. The decline has been partly attributed to political uncertainty domestically and abroad and the 2014 election. So, to maintain the transparency among customers and also the comparison can be made easy through this model. If customer finds the price of house at some given website higher than the price predicted by the model, so he can reject that house.

Observation

Data exploration

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, Python. Before it can conduct analysis on data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working. We divided the data 9:1 for Training and Testing purpose respectively.

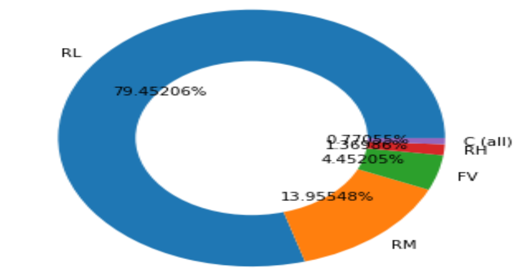
Some general observation when while doing Exploratory Data Analysis

- Dataset have shape for train dataset - ((1168, 81), test dataset - (292, 80))
- Columns having dtypes – int, float, bool, objects
- There are many null values is the dataset.

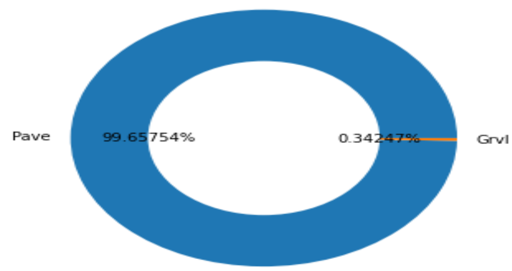
Observations after visualizations

- FV is highest in price followed by RL and RH.
- Streets having Pave and Alley having Grvl is having high Price.
- LotShape of IR2 is high in Price.
- LandContour with HLS ,LotConfig with FR3,LandSlope woth Sev are having higher prices than the other subcategories.
- Condition 1 withRRNn nad PosA have high price.
- Condition 2 with PonA follewd by PosN are having prices.
- BldgType of Twnhse,HouseStyle of 2.5Unf,RoofStyle of Shed, RoofMatl of Wdshngl, Exterior1st of stone and Imstucc are high prices whereas Exterior2nd with other and Imstucc have high price.

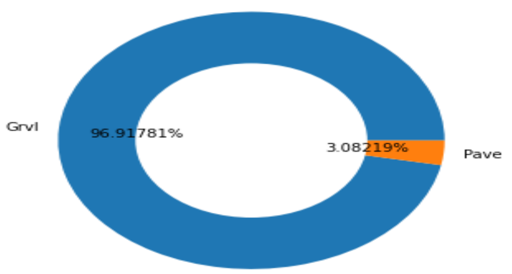
- MasVnrType with stone, ExterQual with Ex,ExterCond withEx,Foundation with Pconc, BsmtQual with ex BmstCond with Gd,BmstExposer with Gd,BsmtFinType1 with GQL,BsmtFinType2 with GQL and AQL are high in Price.
- Heating with GasA ,HeatingQc with Ex,CentralAir with Yes Electrical with SBrkr, KitchenQual with Ex ,Funtional with Typ,FireplaceQu with Ex, GarageType with BuiltIn, GarageFinish with Fin has high Price.
- GarageQual with Ex, GarageCond with Gd, PavesDrive with Y, PoolQc with Ex, Fence with MnPrv nad GdPrv are high in Price.
- SaleType of con and new ,SaleCondition with Partial are having highest SalePrice.
- Some features such as Id, YearRemodAdd, BsmFullBath, FullBath, HalfBathFirePlace, MoSold, YrSold are not having outliers.
- Rest of the features are more or less having outliers .
- Many features are skewed
- Some of the features are following standard deviation curve



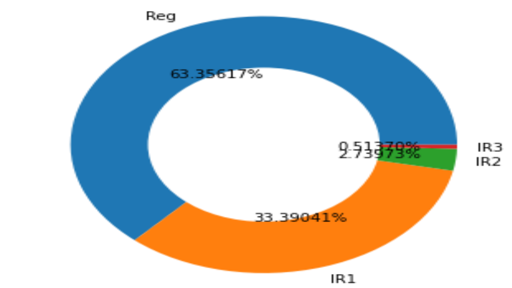
MSZoning



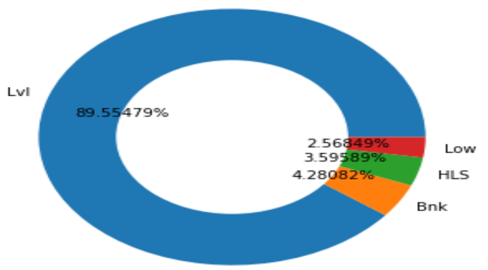
Street



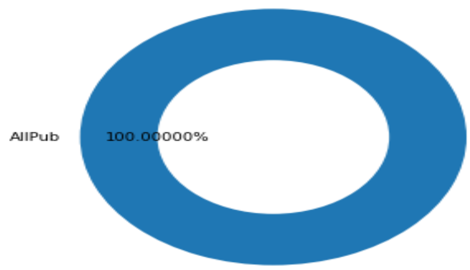
Alley



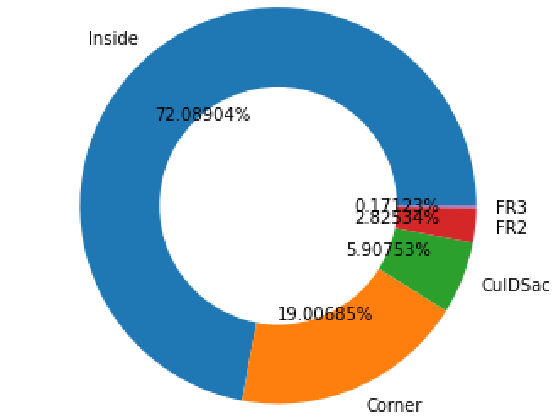
LotShape



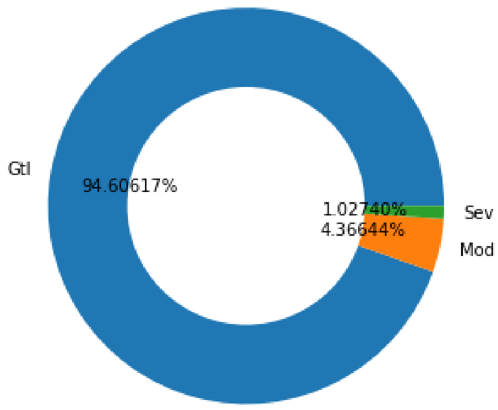
LandContour



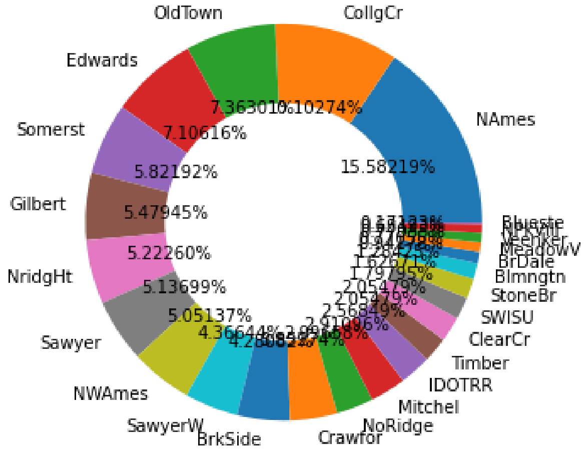
Utilities



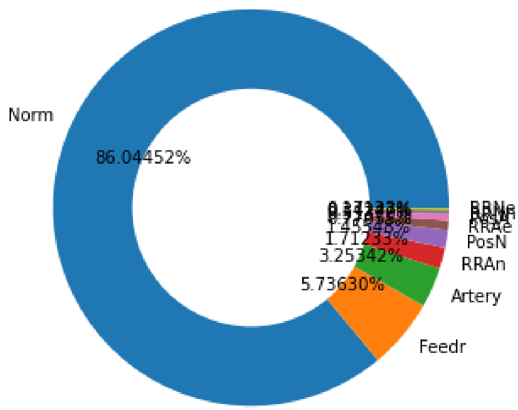
LotConfig



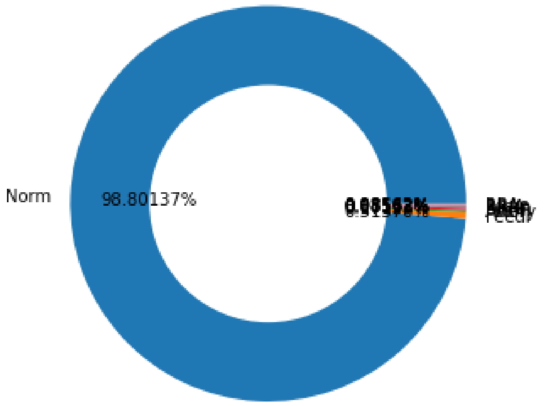
LandSlope



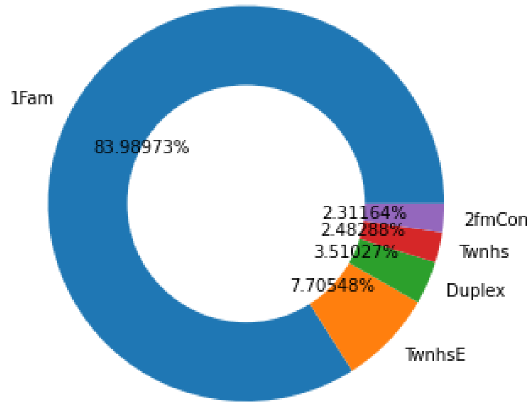
Neighborhood



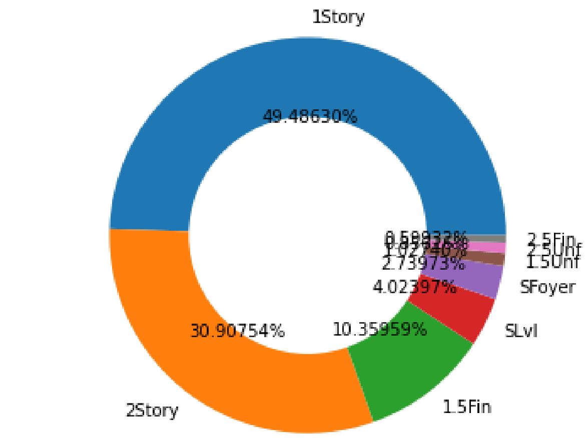
Condition1



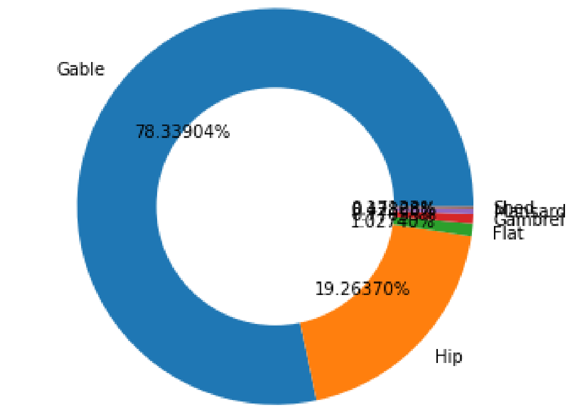
Condition2



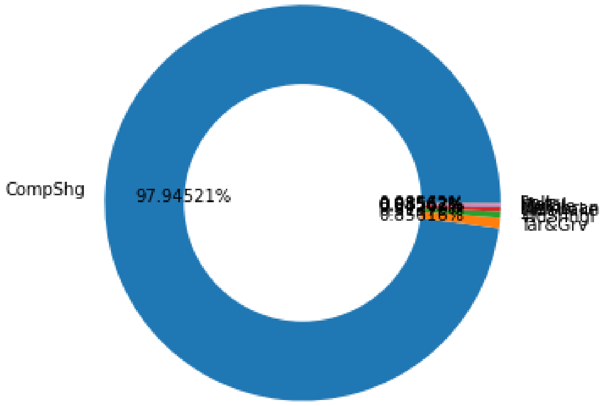
BldgType



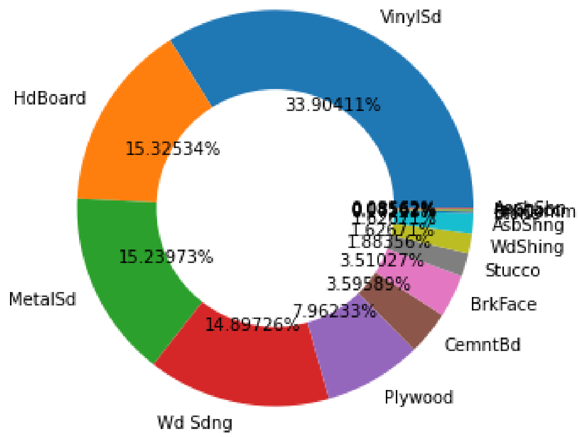
HouseStyle



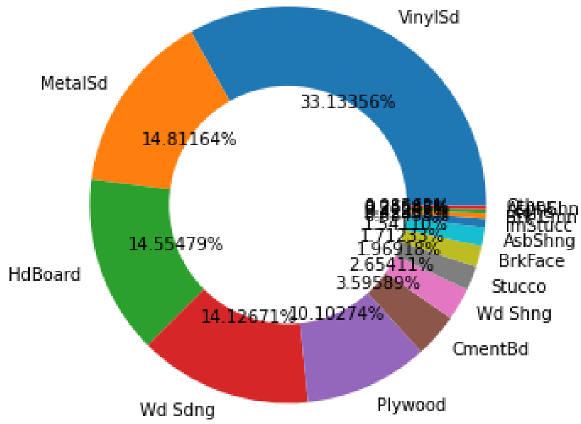
RoofStyle



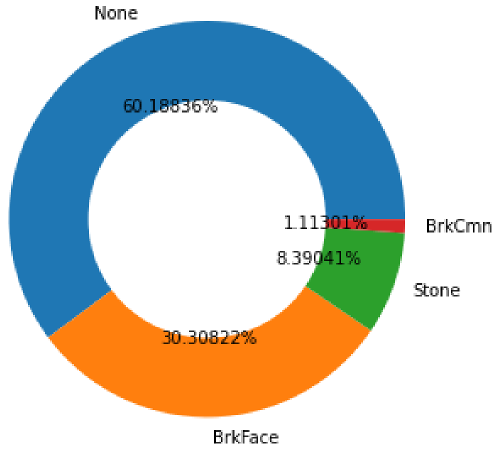
RoofMatl



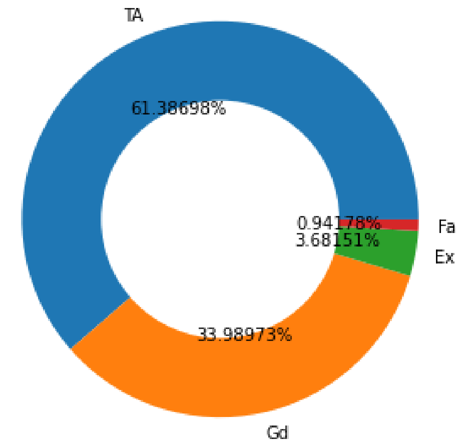
Exterior1st



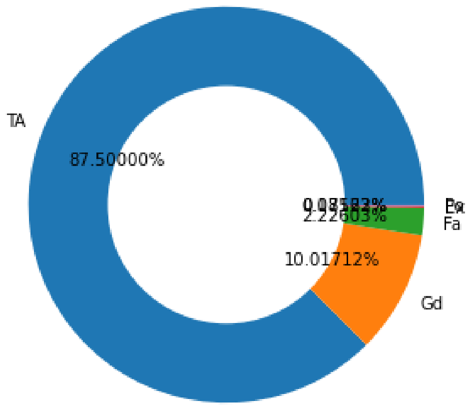
Exterior2nd



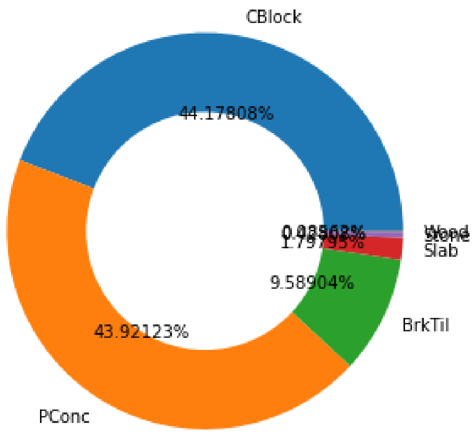
MasVnrType



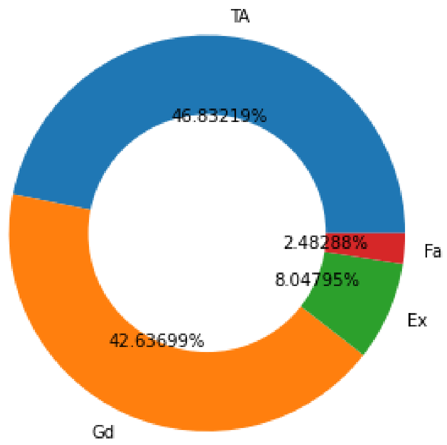
ExterQual



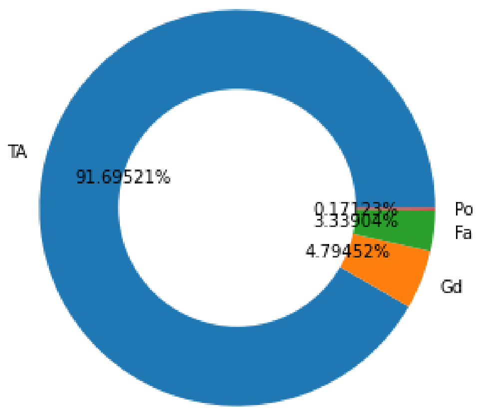
ExterCond



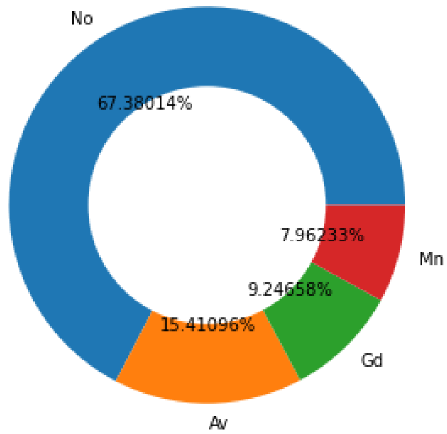
Foundation



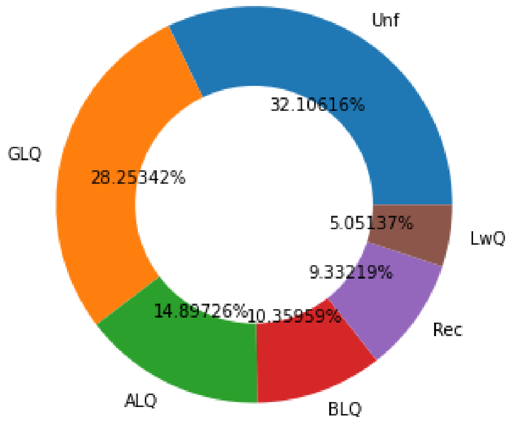
BsmtQual



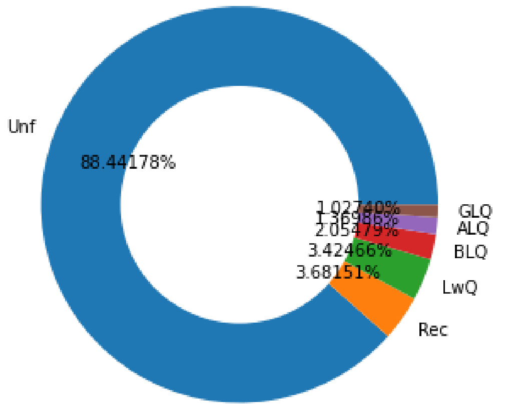
BsmtCond



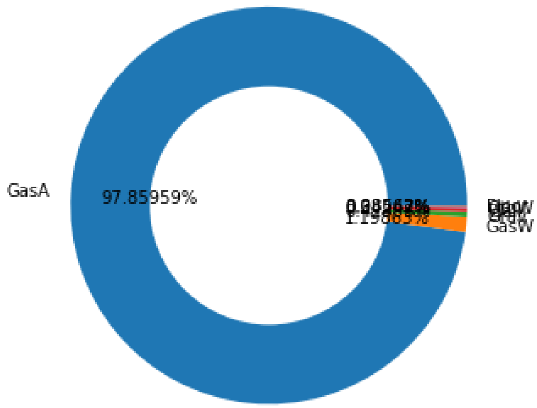
BsmtExposure



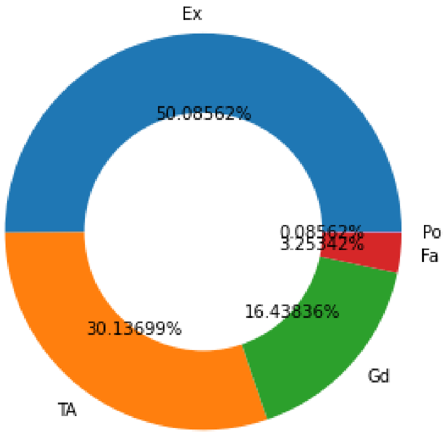
BsmtFinType1



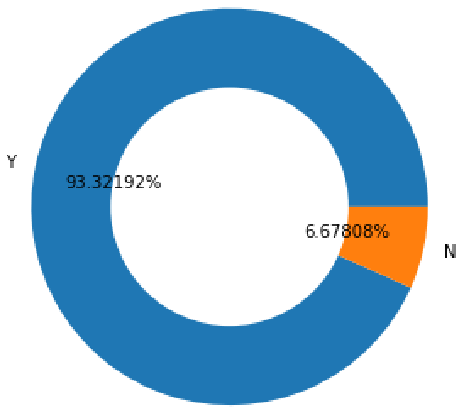
BsmtFinType2



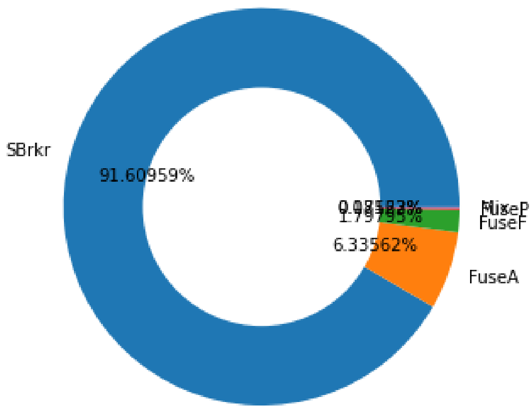
Heating



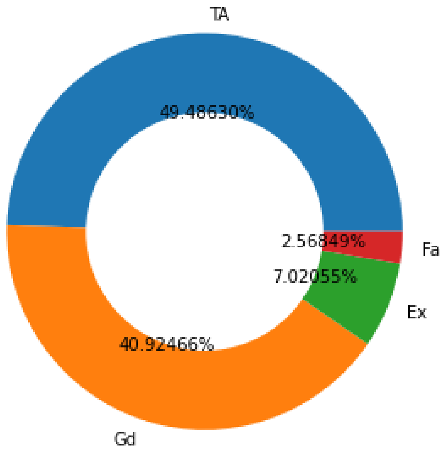
HeatingQC



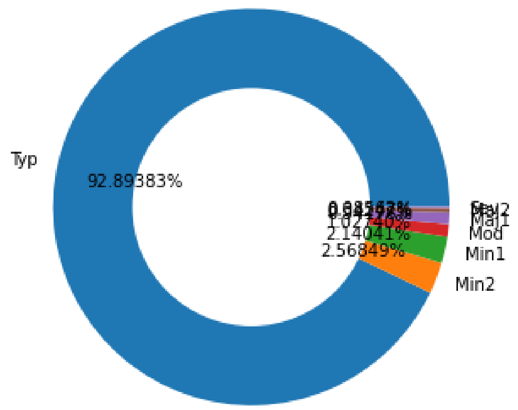
CentralAir



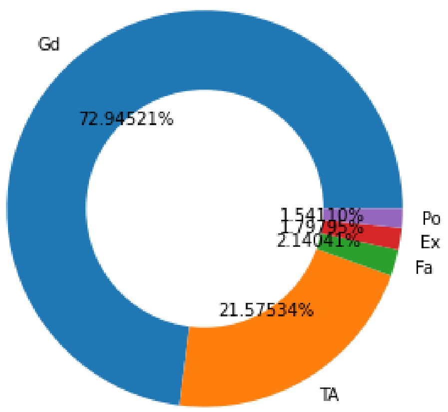
Electrical



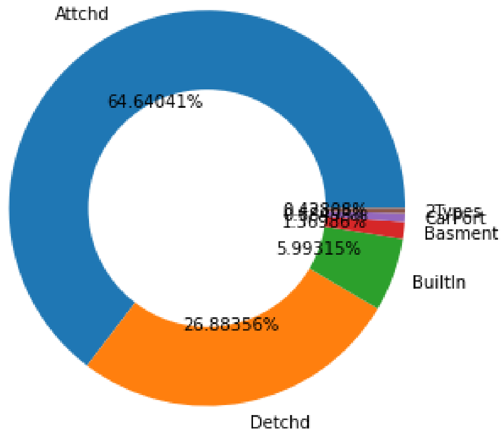
KitchenQual



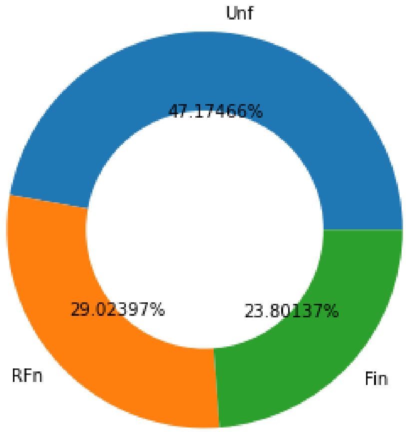
Functional



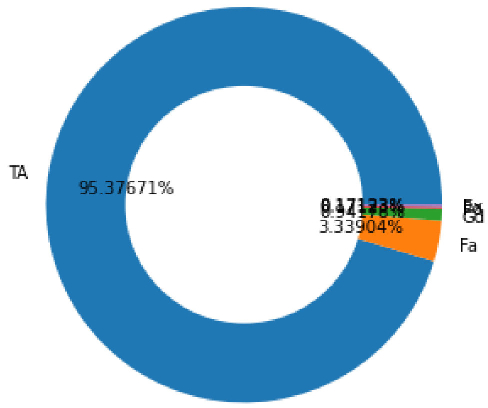
FireplaceQu



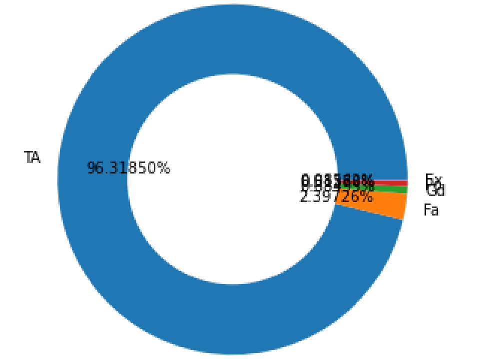
GarageType



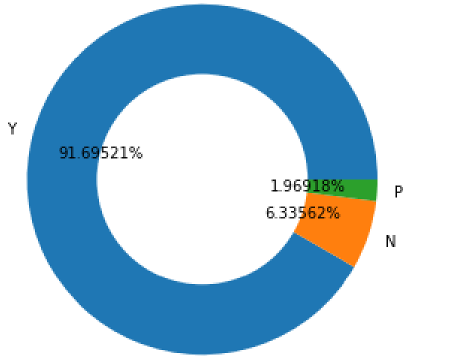
GarageFinish



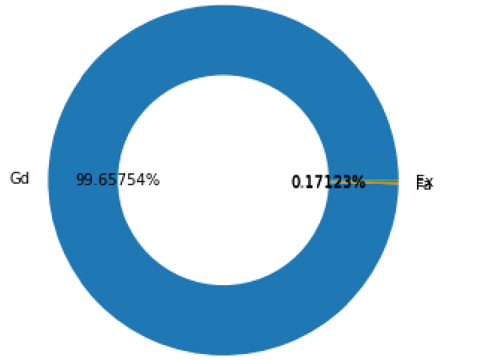
GarageQual



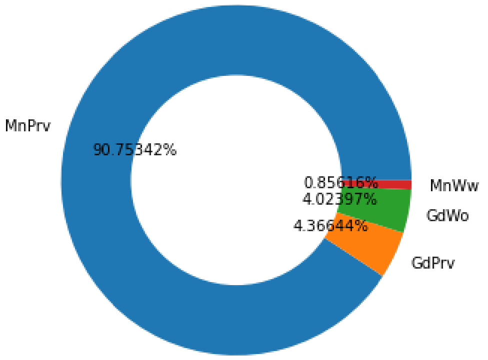
GarageCond



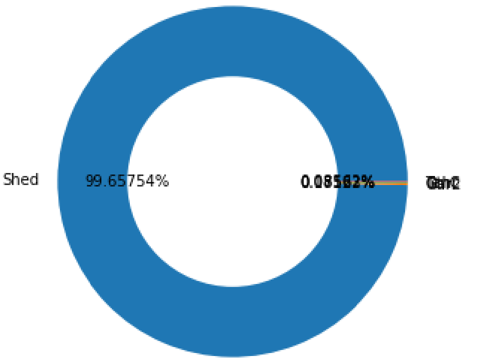
PavedDrive



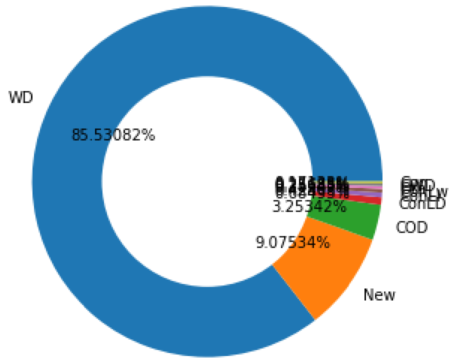
PoolQC



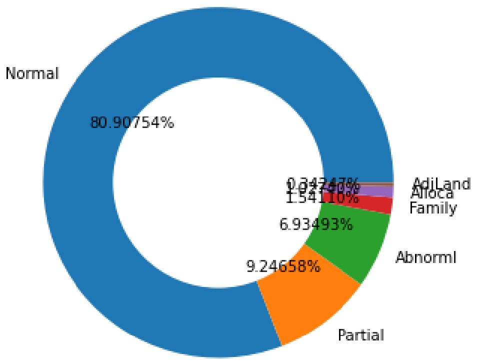
Fence



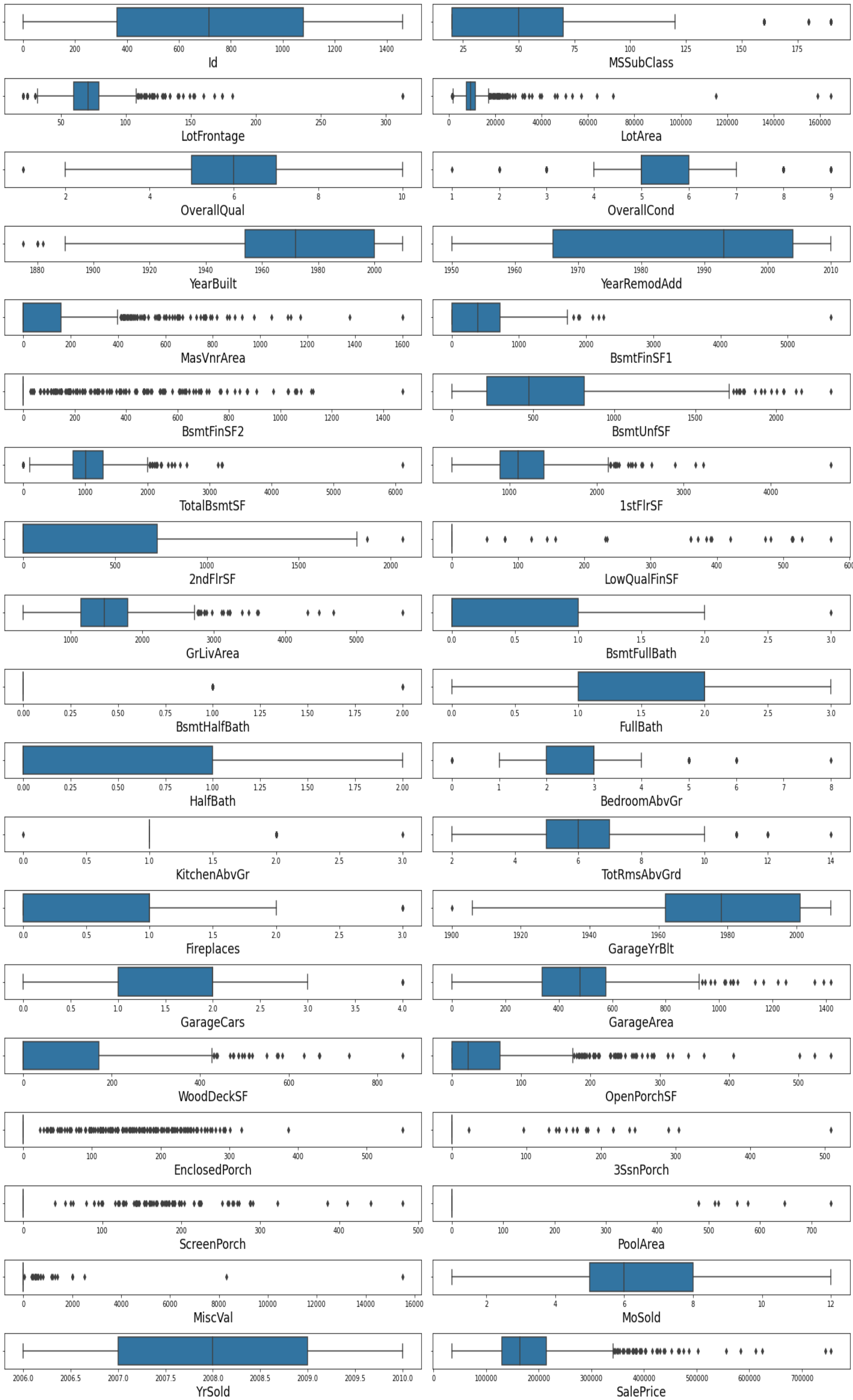
MiscFeature

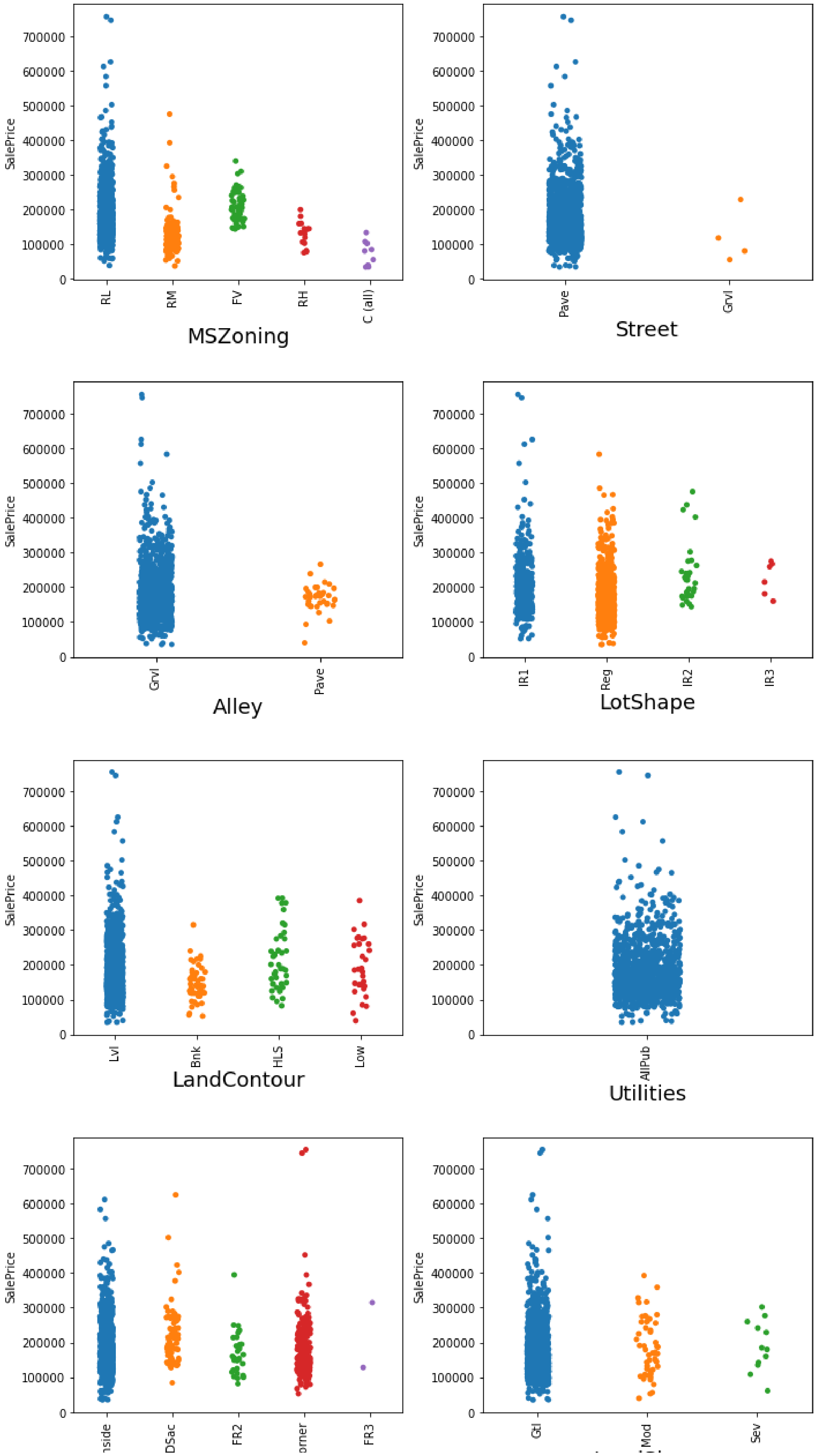


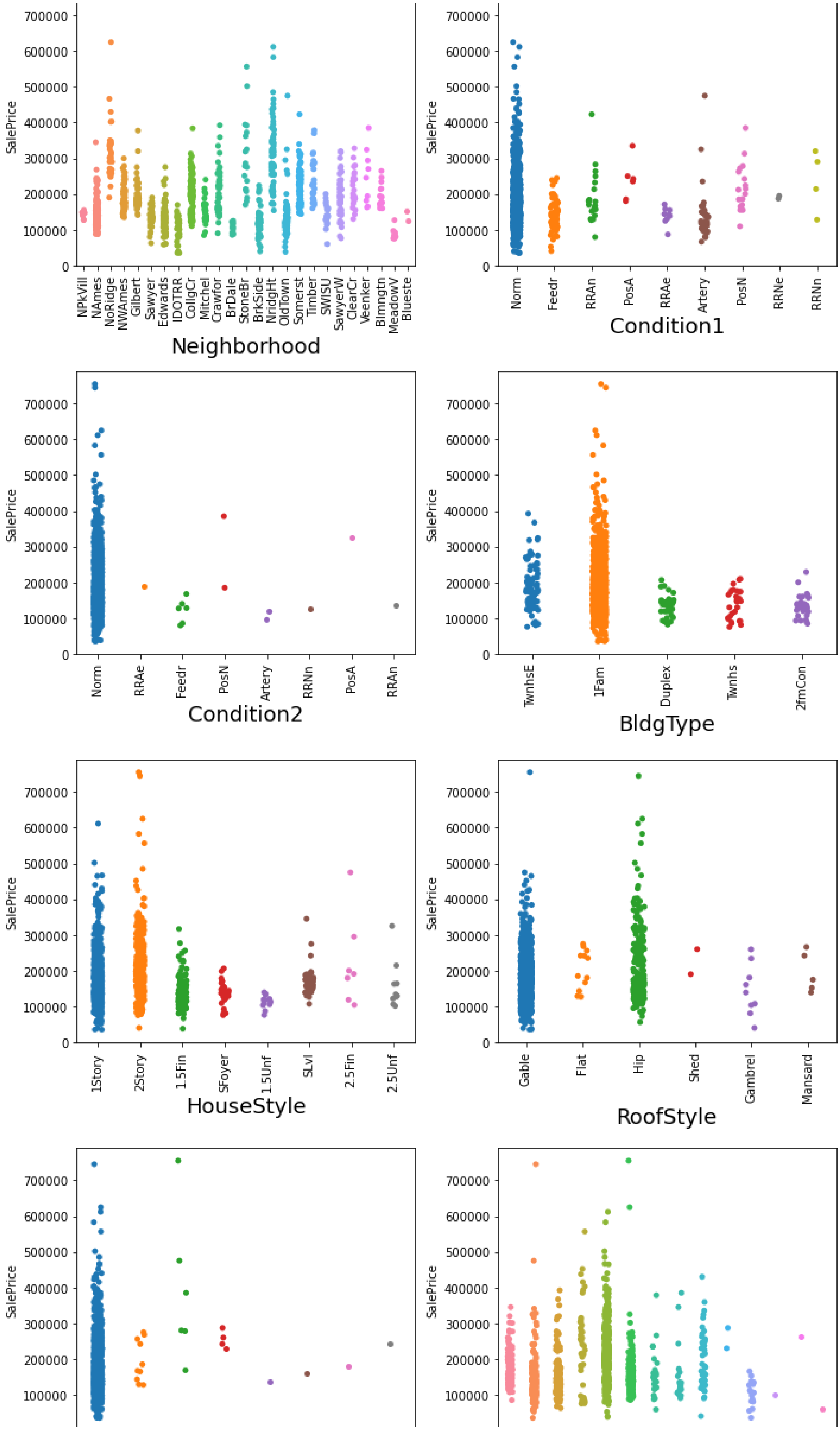
SaleType

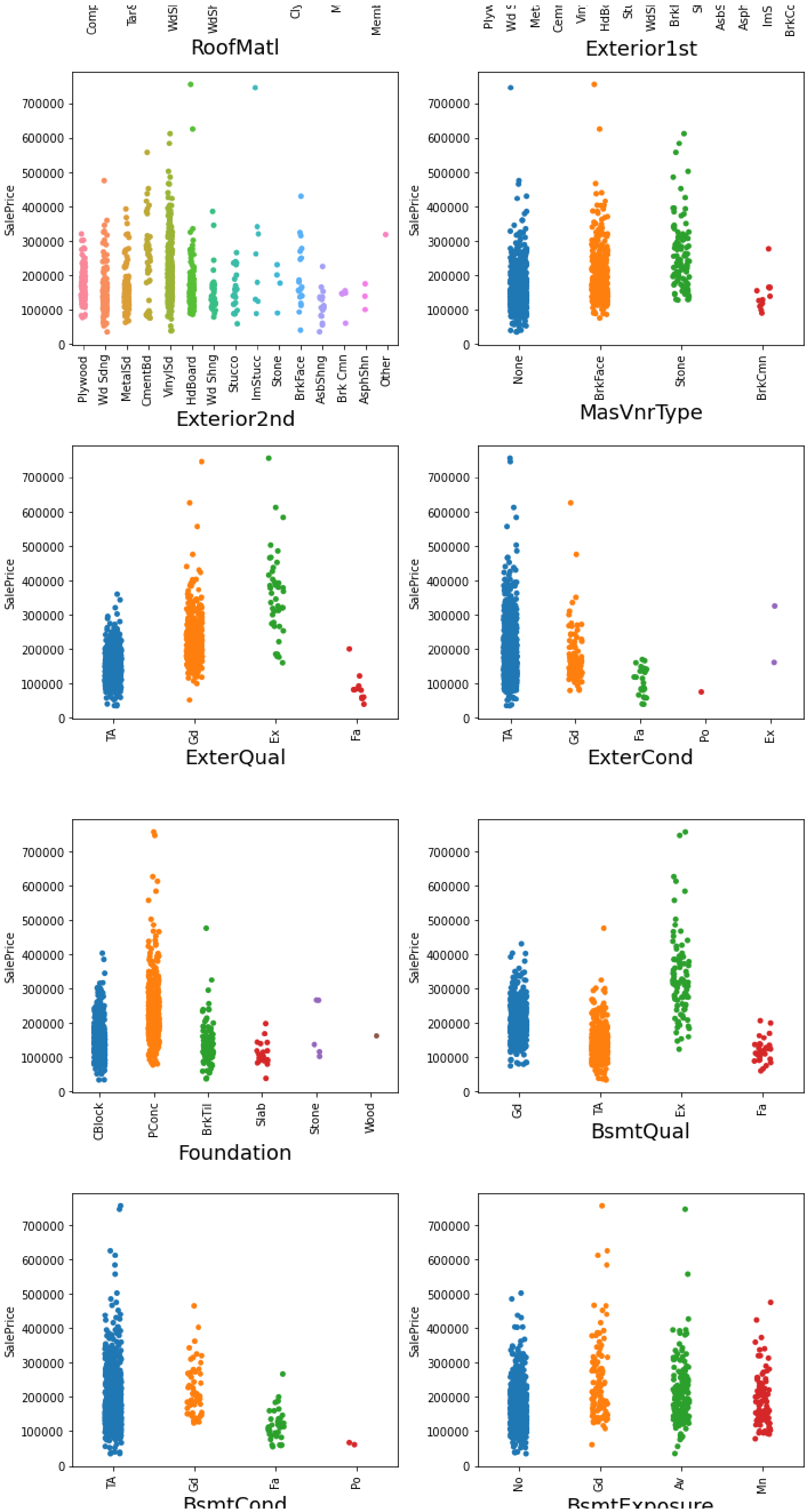


SaleCondition

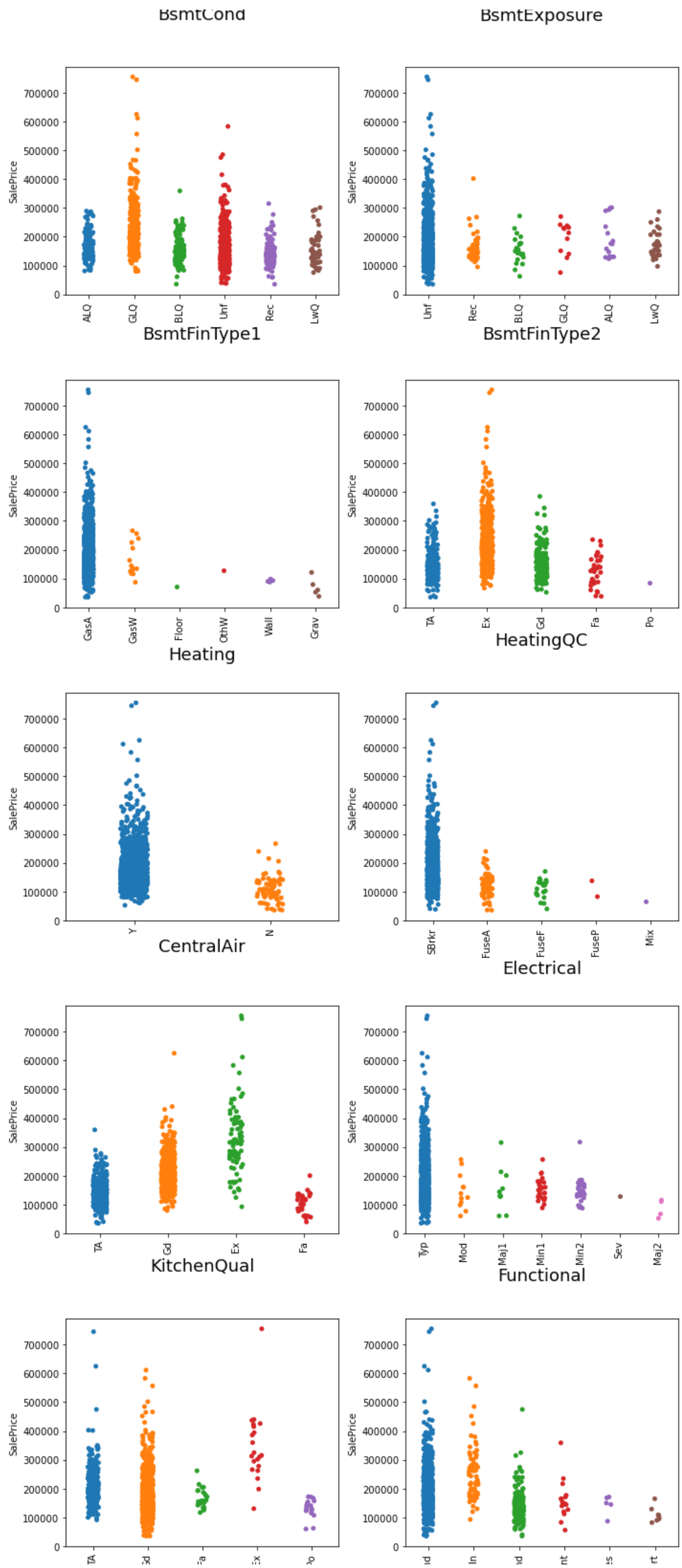

Id
MSSubClass
LotFrontage
LotArea
OverallQual
OverallCond
YearBuilt
YearRemodAdd
MasVnrArea
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF
TotalBsmtSF
1stFlrSF
2ndFlrSF
LowQualFinSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
TotRmsAbvGrd
Fireplaces
GarageYrBlt
GarageCars
GarageArea
WoodDeckSF
OpenPorchSF
EnclosedPorch
3SsnPorch
ScreenPorch
PoolArea
MiscVal
MoSold
YrSold
SalePrice


SalePrice
0
100000
200000
300000
400000
500000
600000
700000
RL
RM
FV
RH
C (all)
MSZoning
Pave
Grvl
Street
Grvl
Pave
Alley
IR1
Reg
IR2
IR3
LotShape
Lvl
Bnk
HLS
Low
LandContour
AllPub
Utilities
DSac
FR2
FR3
Gtl
Mod
Sev


SalePrice
0
100000
200000
300000
400000
500000
600000
700000
NPkVill NAmes NoRidge NWAmes Gilbert Sawyer Edwards IDOTRR CollgCr Mitchel Crawfor BrDale StoneBr BrkSide NridgHt OldTown Somerst Timber SWISU SawyerW ClearCr Veenker Blmngtn MeadowV Blueste
Neighborhood
Norm Feedr RRAn PosA RRAe Artery PosN RRNe RRNn
Condition1
Norm RRAe Feedr PosN Artery RRNn PosA RRAn
Condition2
TwnhsE 1Fam Duplex Twnhs 2fmCon
BldgType
1Story 2Story 1.5Fin SFoyer 1.5Unf SLvl 2.5Fin 2.5Unf
HouseStyle
Gable Flat Hip Shed Gambrel Mansard
RoofStyle



RoofMatl
Exterior1st
SalePrice
0
100000
200000
300000
400000
500000
600000
700000
Plywood
Wd Sdng
MetalSd
CmentBd
VinylSd
HdBoard
Wd Shng
Stucco
ImStucc
Stone
BrkFace
AsbShng
Brk Cmn
AsphShn
Other
Exterior2nd
None
BrkFace
Stone
BrkCmn
MasVnrType
TA
Gd
Ex
Fa
ExterQual
TA
Gd
Fa
Po
Ex
ExterCond
CBlock
PConc
BrkTil
Slab
Stone
Wood
Foundation
Gd
TA
Ex
Fa
BsmtQual
TA
Gd
Fa
Po
BsmtCond
No
Gd
Av
Mn
BsmtExposure

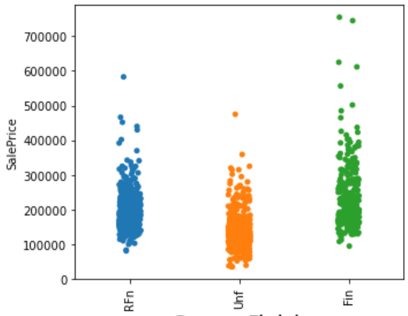

BsmtCond
BsmtExposure
SalePrice
0
100000
200000
300000
400000
500000
600000
700000
ALQ GLQ BLQ Unf Rec LwQ
BsmtFinType1
Unf Rec BLQ GLQ ALQ LwQ
BsmtFinType2
GasA GasW Floor OthW Wall Grav
Heating
TA Ex Gd Fa Po
HeatingQC
Y N
CentralAir
SBrkr FuseA FuseF FuseP Mix
Electrical
TA Gd Ex Fa
KitchenQual
Typ Mod Maj1 Min1 Min2 Sev Maj2
Functional

T G F E P

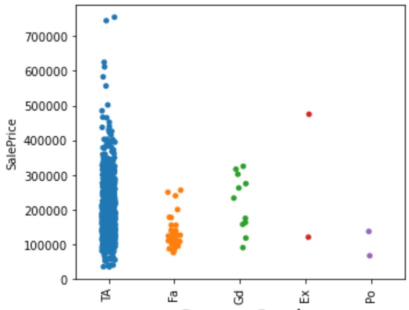
FireplaceQu

Attch BuiltI Detch Basmer 2Types CarPor

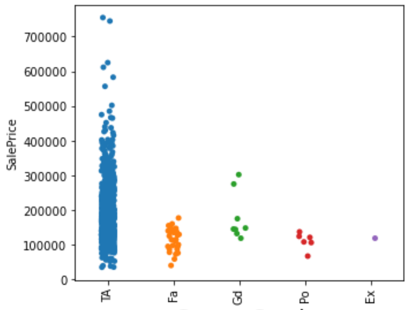
GarageType



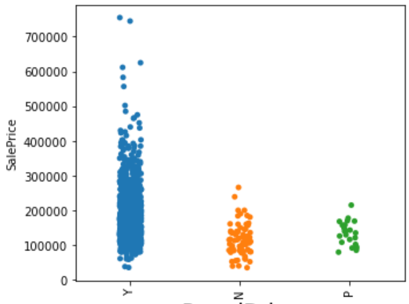
GarageFinish



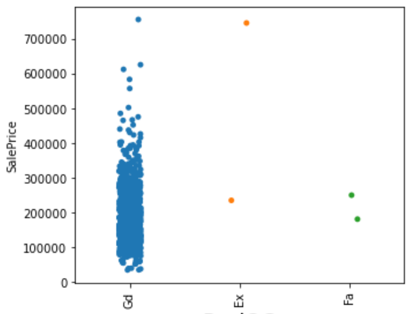
GarageQual



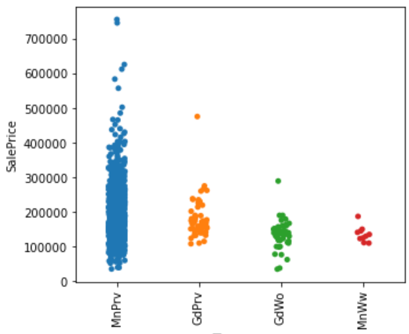
GarageCond



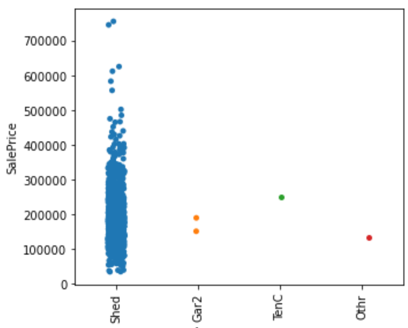
PavedDrive



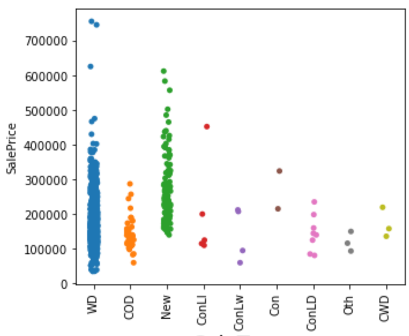
PoolQC



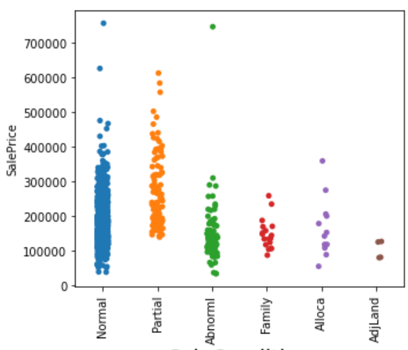
Fence



MiscFeature



SaleType



SaleCondition

Feature Engineering

- Used Log transformation for removing outliers and skewness

```
# There is a need to convert the skewed distribution into gaussian or normal distribution.
for col in train.columns:
    if train[col].dtype!='object':
        if (col=='Id' or col=='SalePrice'):
            continue;
        if train[col].skew() > 0.5:
            train[col] = train[col].apply(lambda x: np.log1p(x))
            test[col] = test[col].apply(lambda x: np.log1p(x))

train['SalePrice'] = train['SalePrice'].apply(lambda x: np.log1p(x))
```

- Used LabelEncoder for encoding every categorical features to encode with numeric codes.

```
from sklearn.preprocessing import LabelEncoder
label = LabelEncoder()

cat=['MSZoning', 'Street', 'Alley', 'LotShape', 'LandContour', 'Utilities',
       'LotConfig', 'LandSlope', 'Neighborhood', 'Condition1', 'Condition2',
       'BldgType', 'HouseStyle', 'RoofStyle', 'RoofMatl', 'Exterior1st',
       'Exterior2nd', 'MasVnrType', 'ExterQual', 'ExterCond', 'Foundation',
       'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinType2',
       'Heating', 'HeatingQC', 'CentralAir', 'Electrical', 'KitchenQual',
       'Functional', 'FireplaceQu', 'GarageType', 'GarageFinish', 'GarageQual',
       'GarageCond', 'PavedDrive', 'PoolQC', 'Fence', 'MiscFeature',
       'SaleType', 'SaleCondition']
train[cat] = label.fit_transform(cat)
test[cat] = label.fit_transform(cat)
```

- Split the dataset for training and testing
- Removed the mean and scales each feature/variable to unit variance.

```
# Now let's split our data into training and validation.
features = train.drop(['Id','SalePrice'],axis=1)
target = train['SalePrice']

test_set = test.drop(['Id'],axis=1)
```

Splited the dataset into features and targets.

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()

features = scaler.fit_transform(features)
test_set = scaler.transform(test_set)
```

Removed the mean and scales each feature/variable to unit variance.

Model/s Development and Evaluation

- **Testing of Identified Approaches (Algorithms)**
- **Techniques:**

- K-Neighbors Regressor
- Decision Tree Regressor
- Support Vector Machine
- Random Forest Regressor
- Gradient Boosting Regressor

Algorithms

```
# K-Neighbors Regressor
from sklearn.neighbors import KNeighborsRegressor
knr = KNeighborsRegressor()
beststate(knr)
```

```
Best Random State      :  73
Best R2_Score          :  0.8167488311602309
Cross Validation Score :  0.7814946878164013

Time taken by model for prediction 0.0890 seconds
```

```
# Decision Tree Regressor
from sklearn.tree import DecisionTreeRegressor
dt = DecisionTreeRegressor()
beststate(dt)
```

```
Best Random State      :  72
Best R2_Score          :  0.7288358196885846
Cross Validation Score :  0.6661584106762032

Time taken by model for prediction 0.2135 seconds
```

```
# Support Vector Machine
from sklearn.svm import SVR
svr = SVR()
beststate(svr)
```

```
Best Random State      :  74
Best R2_Score          :  0.8652776650382791
Cross Validation Score :  0.8229042578128875

Time taken by model for prediction 0.9164 seconds
```

```
# Random Forest Regressor
from sklearn.ensemble import RandomForestRegressor
rf = RandomForestRegressor()
beststate(rf)
```

```
Best Random State      :  72
Best R2_Score          :  0.8958856523693193
Cross Validation Score :  0.8521468505943737

Time taken by model for prediction 16.0362 seconds
```

```
# Gradient Boosting Regressor
from sklearn.ensemble import GradientBoostingRegressor
gbr = GradientBoostingRegressor()
beststate(gbr)
```

```
Best Random State      :  72
Best R2_Score          :  0.894155922461464
Cross Validation Score :  0.8725352557999217

Time taken by model for prediction 6.1228 seconds
```

We can clearly see that Gradient Boosting Regressor and Random Forest Regressor are giving almost the same and best scores but due to time factor, and cost factor, I think the Gradient Boosting Regressor is the best model.

Let’s Hyper parameter tune the model with GridSearchCV

Hyper Tuning the Model

```
# Hyper Parameter Tuning with Gradient Boosting Regressor

X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.20, random_state=72)

from sklearn.model_selection import GridSearchCV

param_grid = {"min_samples_leaf" : [1,2,3],
              "min_samples_split" : [2,3,4],
              "n_estimators" : [100,200],
              "learning_rate" : [0.1,0.2]}
grid_search = GridSearchCV(gbr, param_grid=param_grid)
grid_search.fit(X_train, y_train)
grid_search.best_params_
```

```
{'learning_rate': 0.1,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'n_estimators': 200}
```

Hyper Parameter tuning of the model having best r2 score is done to get the best parameters

```
# Final Model
best_model = GradientBoostingRegressor(learning_rate=0.1,min_samples_split=2,min_samples_leaf=1,n_estimators=200)
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.20, random_state=72)
best_model.fit(X_train, y_train)
y_pred = best_model.predict(X_test)
r2_score(y_test, y_pred)
```

```
0.8947724792730296
```

Using the best parameters to get the best hypertuned model

After hyper parameter tuning the r2 score is 89.5 % which is a good score.

Last let predict with test data and store it in a csv file

- Predicting the test dataset and view the first five prediction
- Saving a .csv file for storing the predicted Sale Price of the House

```
y_output = best_model.predict(test_set)

y_output = np.expm1(y_output)
pd.DataFrame({'Id':test.Id,'SalePrice':y_output}).to_csv('house price prediction - test dataset.csv', index=False)

out = pd.read_csv(r'house price prediction - test dataset.csv')
out.head()
```

	Id	SalePrice
0	337	353577.882240
1	1018	193341.979216
2	929	241558.706276
3	1148	172418.166948
4	1227	195998.090457

Conclusion:

- Learning Outcomes of the Study in respect of Data Science

  - Our customers' requirements are our highest priority so the project was built to satisfy their needs so the project works well and there is no customer churn

  - We should maintain the transparency among customers and also the comparison can be made easy through this model. If customer finds the price of house at some given website higher than the price predicted by the model, so he can reject that house.
  - So, we have to predict the pricing as per customers requirement and needs.

- Limitations of this work and Scope for Future Work

  - 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.
  - But still customers are always comparing the prices hence we should keep on updating our project to meet their necessity

Thank You