HOUSING: PRICE PREDICTION

Submitted By:

Ganesh Kumbhar

Acknowledgement

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my project. 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 ***Ms. Khushboo Garg,*** *for providing me an opportunity to do the project work in FlipRobo Technologies and giving us all support and guidance which made me complete the project duly. I am extremely thankful to her for providing such a nice support and guidance.*

I owe my deep gratitude to everyone who guided me.

- ➢ *FlipRobo Technologies*
- ➢ *DataTrained Academy*
- ➢ *Krush Naik*
- ➢ *Code With Harry*
- ➢ *Dr. Abhinanda Sarkar*

Some mentors helped me with their research work mentioned as follow -

https://towardsdatascience.com/style-pandas-dataframe-like-a-master-6b02bf6468b0

https://towardsdatascience.com/feature-selection-correlation-and-p-value-da8921bfb3cf

https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

I am thankful to and fortunate enough to get constant encouragement, support and guidance, which helped me in successfully completing this project work.

INTRODUCTION

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.

Conceptual Background of the Domain Problem

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 a variable?
- How do these variables describe the price of the house?

Review of Literature

<u>Feature Selection</u> - To avoid the curse of dimensionality, and also to avoid overfitting and under filling we should select features which are very important to the data. All of the features we find in the dataset might not be useful in building a machine learning model to make the necessary prediction. Using some of the features might even make the predictions worse. So, feature selection plays a huge role in building a machine learning model.

I learned various methods to select the appropriate features.

- Variance
- P-Value
- Correlation
- Chi-Square Test
- Anova Test
- Co-Independence
- Visualization

<u>Conclusion and Need for Additional Research</u> - Removal of outliers plays very important role in as it manipulate a fine percentage of data and currently the known methods are zero, mean, median , mode , Z-score but i need to do more research in order to get the data which is outside the standard deviation.

Motivation for the Problem Undertaken

House Price Index is commonly used to estimate the changes in housing price. Since housing price is strongly correlated to other factors such as location, area, population, it requires other information apart from HPI to predict individual housing price. There has been a considerably large number of papers adopting traditional machine learning approaches to predict housing prices accurately, but they rarely concern themselves with the performance of individual models and neglect the less popular yet complex models. As a result, to explore various impacts of features on prediction methods, this paper will apply both traditional and advanced machine learning approaches to investigate the difference among several advanced models. This paper will also comprehensively validate multiple techniques in model implementation on regression and provide an optimistic result for housing price prediction.

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem

Outliers - When you get the description of a data set we see that data that has extreme outliers and these outliers are not 1 or 2 points above 3rd quartile meanwhile they are the distribution of data which is disturbing the mean,variance and standard deviation of data.

We cannot remove data through Z-score because in that way we remove the entire row and after passing all the features with outliers and removing the rows, we lose the entire data. To remove the outliers we chose a top 12 feature with highest correlation with the target variable and selected the boundary condition after looking at the scatter plot and removed the row, but we only did this to train the data set not test.

NaN Values - There are many features which have more then 50% data missing so we removed thode columns but the columns which have NaN value less then 50% we imputed the value with mean.

Grouping - The target data is given to us in continuous format and it will be very hard for us to classify continuous variables, so we divided the target variable into 6 bins.

Dropping Unnecessary columns

1. *By Uniqueness :-* First we separated the mobile number by I , to see the relationship between first and second number if any, but we see no relationship and as the number are unique even though some numbers are repeating but still their data is different so we can consider them as unique and they provide no information in data analysis and as well as machine learning whatsoever, so we dropped that column.

   We drop the column of serial number at it has all unique value

2. *By Zero Variance:-* We also dropped P-circle and Year column that is extracted from date column because they have zero variance.*By*

Data Sources and their formats

Data contains 1460 entries each having 81 variables..

1. ***MSSubClass**: Identifies the type of dwelling involved in the sale.*
2. ***MSZoning**: Identifies the general zoning classification of the sale.*
3. ***LotFrontage**: Linear feet of street connected to property*
4. ***LotArea**: Lot size in square feet*
5. ***Street**: Type of road access to property*
6. ***Alley**: Type of alley access to property*
7. ***LotShape**: General shape of property*
8. ***LandContour**: Flatness of the property*
9. ***Utilities**: Type of utilities available*
10. ***LotConfig**: Lot configuration*
11. ***LandSlope**: Slope of property*
12. ***Neighborhood**: Physical locations within Ames city limits*
13. ***Condition1**: Proximity to various conditions*
14. ***Condition2**: Proximity to various conditions (if more than one is present)*
15. ***BldgType**: Type of dwelling*
16. ***HouseStyle**: Style of dwelling*
17. ***OverallQual**: Rates the overall material and finish of the house*
18. ***OverallCond**: Rates the overall condition of the house*
19. ***YearBuilt**: Original construction date*
20. ***YearRemodAdd**: Remodel date (same as construction date if no remodeling or additions)*
21. ***RoofStyle**: Type of roof*
22. ***RoofMatl**: Roof material*
23. ***Exterior1st**: Exterior covering on house*
24. ***Exterior2nd**: Exterior covering on house (if more than one material)*
25. ***MasVnrType**: Masonry veneer type*
26. ***MasVnrArea**: Masonry veneer area in square feet*
27. ***ExterQual**: Evaluates the quality of the material on the exterior*
28. ***ExterCond**: Evaluates the present condition of the material on the exterior*
29. ***Foundation**: Type of foundation*
30. ***BsmtQual**: Evaluates the height of the basement*
31. ***BsmtCond**: Evaluates the general condition of the basement*
32. ***BsmtExposure**: Refers to walkout or garden level walls*

33. ***BsmtFinType1**: Rating of basement finished area*
34. ***BsmtFinSF1**: Type 1 finished square feet*
35. ***BsmtFinType2**: Rating of basement finished area (if multiple types)*
36. ***BsmtFinSF2**: Type 2 finished square feet*
37. ***BsmtUnfSF**: Unfinished square feet of basement area*
38. ***TotalBsmtSF**: Total square feet of basement area*
39. ***Heating**: Type of heating*
40. ***HeatingQC**: Heating quality and condition*
41. ***CentralAir**: Central air conditioning*
42. ***Electrical**: Electrical system*
43. ***1stFlrSF**: First Floor square feet*
44. ***2ndFlrSF**: Second floor square feet*
45. ***LowQualFinSF**: Low quality finished square feet (all floors)*
46. ***GrLivArea**: Above grade (ground) living area square feet*
47. ***BsmtFullBath**: Basement full bathrooms*
48. ***BsmtHalfBath**: Basement half bathrooms*
49. ***FullBath**: Full bathrooms above grade*
50. ***HalfBath**: Half baths above grade*
51. ***Bedroom**: Bedrooms above grade (does NOT include basement bedrooms)*
52. ***Kitchen**: Kitchens above grade*
53. ***KitchenQual**: Kitchen quality*
54. ***TotRmsAbvGrd**: Total rooms above grade (does not include bathrooms)*
55. ***Functional**: Home functionality (Assume typical unless deductions are warranted)*
56. ***Fireplaces**: Number of fireplaces*
57. ***FireplaceQu**: Fireplace quality*
58. ***GarageType**: Garage location*
59. ***GarageYrBlt**: Year garage was built*
60. ***GarageFinish**: Interior finish of the garage*
61. ***GarageCars**: Size of garage in car capacity*
62. ***GarageArea**: Size of garage in square feet*
63. ***GarageQual**: Garage quality*
64. ***GarageCond**: Garage condition*
65. ***PavedDrive**: Paved driveway*
66. ***WoodDeckSF**: Wood deck area in square feet*
67. ***OpenPorchSF**: Open porch area in square feet*
68. ***EnclosedPorch**: Enclosed porch area in square feet*
69. ***3SsnPorch**: Three season porch area in square feet*
70. ***ScreenPorch**: Screen porch area in square feet*
71. ***PoolArea**: Pool area in square feet*
72. ***PoolQC**: Pool quality*
73. ***Fence**: Fence quality*

74. ***MiscFeature**: Miscellaneous feature not covered in other categories*
75. ***MiscVal**: $Value of miscellaneous feature*
76. ***MoSold**: Month Sold (MM)*
77. ***YrSold**: Year Sold (YYYY)*
78. ***SaleType**: Type of sale*
79. ***SaleCondition**: Condition of sale*
80. ***Id:** Id of House*
81. ***SalePrice** : Price of House*

The statical summary of the data are follow

In [47]:
```
df.describe().transpose()
```

Out[47]:

	count	mean	std	min	25%	50%	75%	max
Id	1168.0	724.136130	416.159877	1.0	360.50	714.5	1079.5	1460.0
MSSubClass	1168.0	56.767979	41.940650	20.0	20.00	50.0	70.0	190.0
LotFrontage	954.0	70.988470	24.828750	21.0	60.00	70.0	80.0	313.0
LotArea	1168.0	10484.749144	8957.442311	1300.0	7621.50	9522.5	11515.5	164660.0
OverallQual	1168.0	6.104452	1.390153	1.0	5.00	6.0	7.0	10.0
OverallCond	1168.0	5.595890	1.124343	1.0	5.00	5.0	6.0	9.0
YearBuilt	1168.0	1970.930651	30.145255	1875.0	1954.00	1972.0	2000.0	2010.0
YearRemodAdd	1168.0	1984.758562	20.785185	1950.0	1966.00	1993.0	2004.0	2010.0
MasVnrArea	1161.0	102.310078	182.595606	0.0	0.00	0.0	160.0	1600.0
BsmtFinSF1	1168.0	444.726027	462.664785	0.0	0.00	385.5	714.5	5644.0
BsmtFinSF2	1168.0	46.647260	163.520016	0.0	0.00	0.0	0.0	1474.0
BsmtUnfSF	1168.0	569.721747	449.375525	0.0	216.00	474.0	816.0	2336.0
TotalBsmtSF	1168.0	1061.095034	442.272249	0.0	799.00	1005.5	1291.5	6110.0
1stFlrSF	1168.0	1169.860445	391.161983	334.0	892.00	1096.5	1392.0	4692.0
2ndFlrSF	1168.0	348.826199	439.696370	0.0	0.00	0.0	729.0	2065.0
LowQualFinSF	1168.0	6.380137	50.892844	0.0	0.00	0.0	0.0	572.0
GrLivArea	1168.0	1525.066781	528.042957	334.0	1143.25	1468.5	1795.0	5642.0
BsmtFullBath	1168.0	0.425514	0.521615	0.0	0.00	0.0	1.0	3.0
BsmtHalfBath	1168.0	0.055651	0.236699	0.0	0.00	0.0	0.0	2.0
FullBath	1168.0	1.562500	0.551882	0.0	1.00	2.0	2.0	3.0
HalfBath	1168.0	0.388699	0.504929	0.0	0.00	0.0	1.0	2.0
BedroomAbvGr	1168.0	2.884418	0.817229	0.0	2.00	3.0	3.0	8.0
KitchenAbvGr	1168.0	1.045377	0.216292	0.0	1.00	1.0	1.0	3.0
TotRmsAbvGrd	1168.0	6.542808	1.598484	2.0	5.00	6.0	7.0	14.0
Fireplaces	1168.0	0.617295	0.650575	0.0	0.00	1.0	1.0	3.0
GarageYrBlt	1104.0	1978.193841	24.890704	1900.0	1961.00	1980.0	2002.0	2010.0
GarageCars	1168.0	1.776541	0.745554	0.0	1.00	2.0	2.0	4.0
GarageArea	1168.0	476.860445	214.466769	0.0	338.00	480.0	576.0	1418.0
WoodDeckSF	1168.0	96.206336	126.158988	0.0	0.00	0.0	171.0	857.0
OpenPorchSF	1168.0	46.559932	66.381023	0.0	0.00	24.0	70.0	547.0
EnclosedPorch	1168.0	23.015411	63.191089	0.0	0.00	0.0	0.0	552.0
3SsnPorch	1168.0	3.639555	29.088867	0.0	0.00	0.0	0.0	508.0
ScreenPorch	1168.0	15.051370	55.080816	0.0	0.00	0.0	0.0	480.0
PoolArea	1168.0	3.448630	44.896939	0.0	0.00	0.0	0.0	738.0
MiscVal	1168.0	47.315068	543.264432	0.0	0.00	0.0	0.0	15500.0
MoSold	1168.0	6.344178	2.686352	1.0	5.00	6.0	8.0	12.0
YrSold	1168.0	2007.804795	1.329738	2006.0	2007.00	2008.0	2009.0	2010.0
SalePrice	1168.0	181477.005993	79105.586863	34900.0	130375.00	163995.0	215000.0	755000.0

Data Preprocessing

1. We Checked the statistical summary of data.
2. We remove outliers of 12 highly correlated features by setting the boundary condition.
3. We merged various columns as made new feature
4. The data set has very high skewness and we removed it by using a square root method.
5. To make data in standard scale we used the Robust Scaling method.
6. We removed all the data which has 0 variance as they provide no value to machine learning.
7. We made new feature by combining new feature
8. We did encoding of categorical data by one hot encoding
9. We used train_test_split to split data for machine learning.

Data Inputs- Logic- Output Relationships

1. Most of the continuous features are related to Area and they have a direct positive relation with sale Price.
2. Some Features can be deleted.
3. Many features have high Skewness.

The set of assumptions related to the problem under consideration

Feature Selection -

1. TotalSF = TotalBsmtSF + 1stFlrSF+ 2ndFlrSF
2. TotalAreaExt = GrLivArea + GarageArea
3. TotalAreaInt = GrLivArea + TotalBsmtSF

Outliers -

We assumed the boundary conditions of some feature to remove outliers

Hardware and Software Requirements and Tools Used

Software

- ➔ Jupyter Notebook (Python 3.8)
- ➔ Microsoft Excel
- ➔ Microsoft Word

Hardware

- ➔ Processor - Intel i5 9th Gen
- ➔ RAM - 8 GB
- ➔ Graphic Memory - 4Gb , Nvidia 1060

Libraries

- ➔ Pandas
- ➔ Numpy
- ➔ Matplotlib
- ➔ Seaborn
- ➔ Scipy
- ➔ Sklearn

Model/s Development and Evaluation

Problem-solving approaches

Feature Selection - We used two approaches to select the feature first selection of feature by 0 variance and second selection of feature by the internal correlation.
We splitted the data by using train_test_split method and analysed the features on x_train and removed them from x_train and then simply removed the feature from x_test to reduce the chance of overfitting.

Scaling - The scale of data has high variance and to put data in in one scale which will increase the efficiency of our model for that we use a robust scaler library.

Skewness - The data is very much skewed and to remove the skewness we used square root method this method is capable of dealing with high skewness as well as features with 0 value but still skewness is not completely gone but we cannot remove the data any further as data is precious to us.

Testing of Identified Approaches.

- LinearRegression,
- DecisionTree,
- KNeighbors,
- ElasticNet
- Lasso_model,
- Ridge_model,
- SVR_model,
- RandomForest,

- Adaboost,
- GradientBoosting,
- BaggingRegressor,
- ExtraTreesRegressor

Run and Evaluate selected models

First we used Different models using pipeline to avoid any leakage of data

```
In [49]: #ML Models
from sklearn.linear_model import LinearRegression, Lasso, Ridge, ElasticNet
from sklearn.svm import SVR
from sklearn.tree import DecisionTreeRegressor
from sklearn.neighbors import KNeighborsRegressor

# Ensemble ML Models
from sklearn.ensemble import RandomForestRegressor, AdaBoostRegressor ,GradientBoostingRegressor, BaggingRegressor, ExtraTreesReg

#model selection
from sklearn.model_selection import train_test_split, GridSearchCV, cross_val_score

from sklearn.pipeline import make_pipeline


#metrics
from sklearn.metrics import r2_score,mean_absolute_error,mean_squared_error
```

Then we made object to test and print the result of all the models

```
In [36]: def evaluate(model, model_name):

    x_train,x_test,y_train,y_test=train_test_split(x, y, random_state=42, test_size=0.20)

    model.fit(x_train,y_train)
    pred=model.predict(x_test)
    print(f'Model :{model_name}')
    print(model.score(x_train,y_train))
    print('R2 Score')
    print(r2_score(y_test,pred))
    print('MAE')
    print(mean_absolute_error(y_test,pred))
    print('MSE')
    print(mean_squared_error(y_test,pred))
    print()
    print('cross_val_score')
    print(cross_val_score(model,x,y,cv=5).mean())
    print('........................................................................................................')
```

In result we got all the values of matrices then with our choice of matrix, we selected the model

```
Model :Linear
0.8940300139444367
R2 Score
0.8321697665421353
MAE
0.05100853502324266
MSE
0.006670749085572983

cross_val_score
0.8421224723410303
-----------------------------------------
Model :Decision
1.0
R2 Score
0.6940557357996802
MAE
0.08038583408656538
MSE
0.012160368120818728

cross_val_score
0.7208317178881897
-----------------------------------------
Model :KNeighbors
0.7995522286726331
R2 Score
0.6676990416297861
MAE
0.07821816807743276
MSE
0.013207967769047119

cross_val_score
0.6732650020678881
-----------------------------------------
Model :Elastic
0.0
R2 Score
-0.0004032639146507311
MAE
0.1575843280194127
MSE
0.03976303327754299

cross_val_score
-0.01508151483561444
-----------------------------------------
Model :Lasso_model
0.0
R2 Score
-0.0004032639146507311
MAE
0.1575843280194127
MSE
0.03976303327754299

cross_val_score
-0.01508151483561444
-----------------------------------------
Model :Ridge_model
0.8977440006687031
R2 Score
0.8342409861906588
MAE
0.049283400754752245
MSE
0.006588424308375569

cross_val_score
0.8427012950222948
-----------------------------------------
Model :SVR_model
0.9084628781875335
R2 Score
0.8018346116402677
MAE
0.05682970810537466
MSE
0.007876480631392182

cross_val_score
0.8284666193127947
-----------------------------------------
```

```
-----------------------------------------
Model :RamdomForest
0.9817063467884272
R2 Score
0.8315177029458485
MAE
0.05514461760844905
MSE
0.006696666660428515

cross_val_score
0.8587352726140637
-----------------------------------------
Model :Adaboost
0.8864008694894009
R2 Score
0.8137324863686425
MAE
0.06248488392055183
MSE
0.0074035757481102655

cross_val_score
0.8175738212614636
-----------------------------------------
Model :GradientBoosting
0.9681515709270445
R2 Score
0.8644695173723922
MAE
0.04835139612549563
MSE
0.005386930736066451б
```

we see that Gradaintbooster has R2 score = 86%, Lower MSE Score and average cross validation score = 88% but this can further be increased to we will first see all ensemble technique

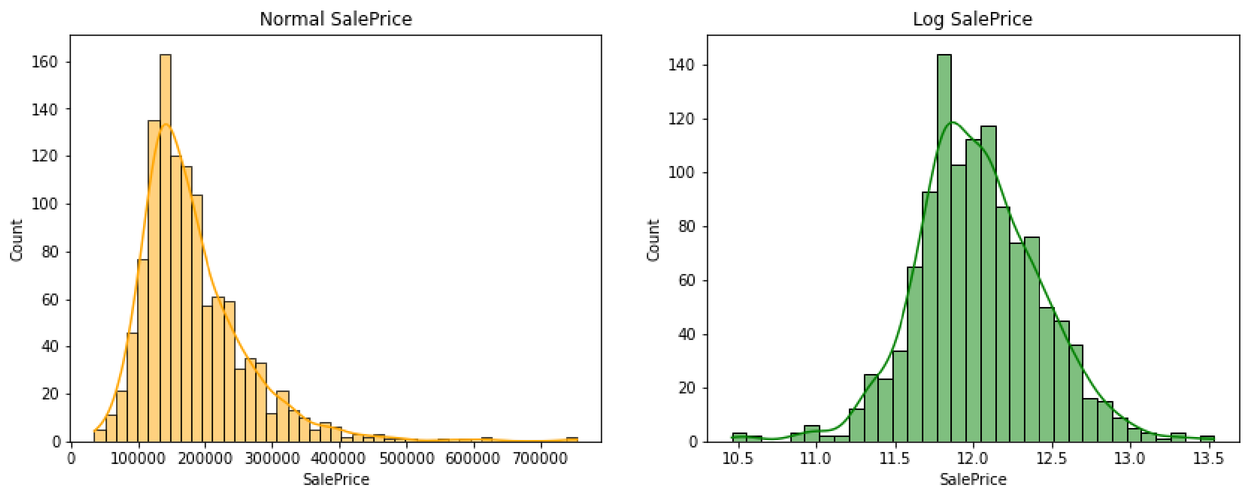
so we will process with **Gradient Boosting Regressor**

Key Metrics for success in solving problem under consideration

1. **R2 Score** - is a statistical measure that represents the goodness of fit of a regression model. The ideal value for r-square is 1. The closer the value of r-square to 1, the better is the model fitted.
2. **Mean Squared Error (MSE)** of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors — that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss.
3. **Cross Validation Score** - to check if our model is overfitting or not we use cross validation score, higher the cross validation score higher the cross validation score means the model is not overfitting.

Visualizations

<u>Label</u> -

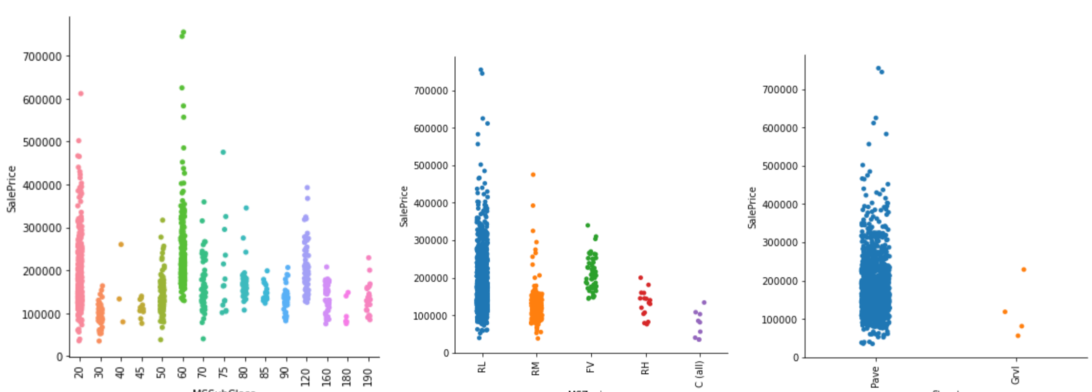


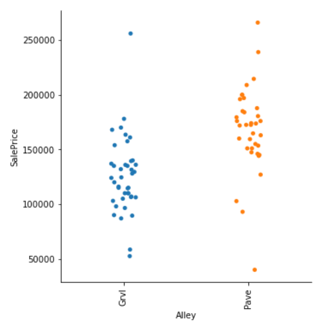
Using Logarithms helps us to have a normal distribution which could help us in a number of different ways such as outlier detection.

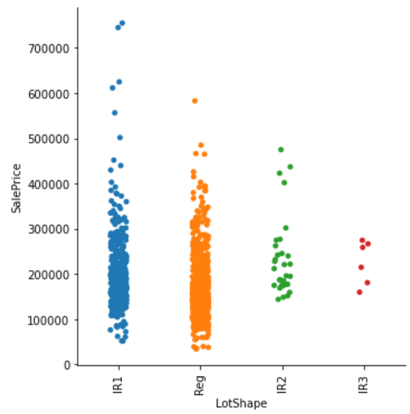
In this data We have a right skewed distribution in which most Sales are between 0 and 340K.

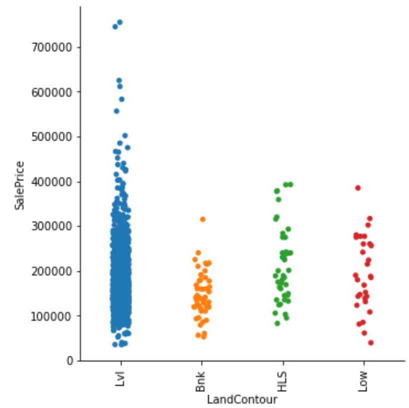
<u>Discrete Feature Analysis -</u>

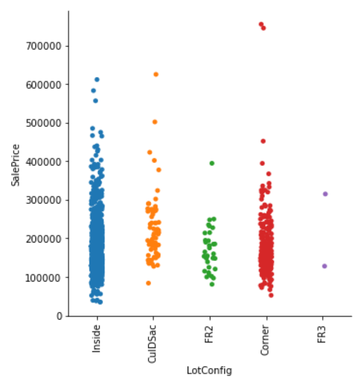
We plot various graphs of continuous Discrete data and its relationship with Sale Price, Below we will see strips in various graphs, and more the range of the srtip and denser the srtip more sales price depend on it.

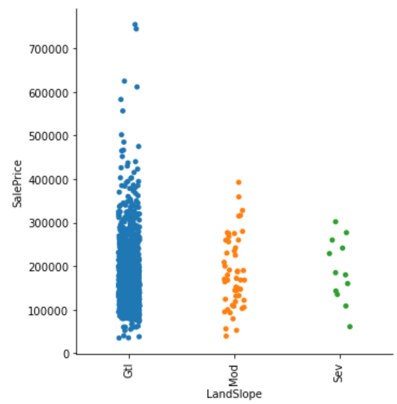


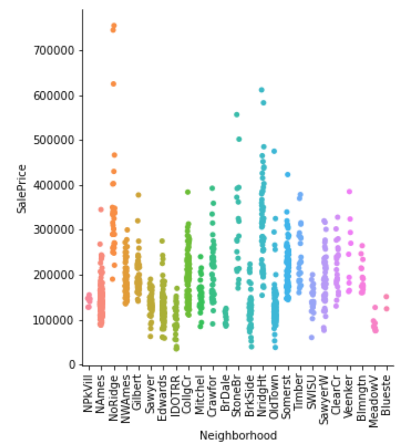

SalePrice
250000
200000
150000
100000
50000
Grvl
Pave
Alley

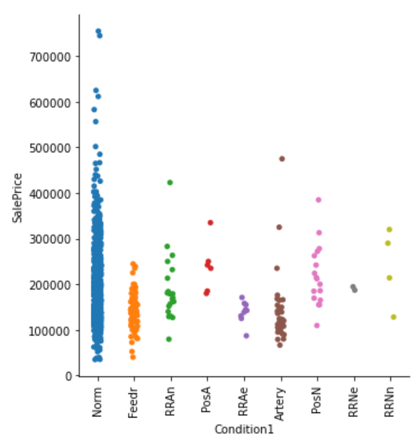

SalePrice
700000
600000
500000
400000
300000
200000
100000
0
IR1
Reg
IR2
IR3
LotShape

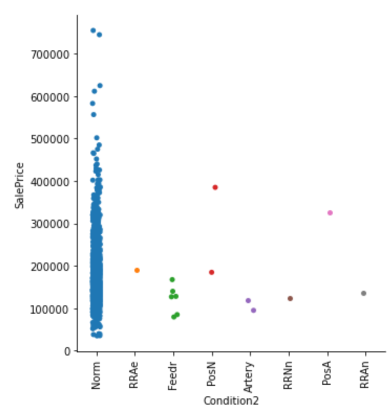

SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Lvl
Bnk
HLS
Low
LandContour

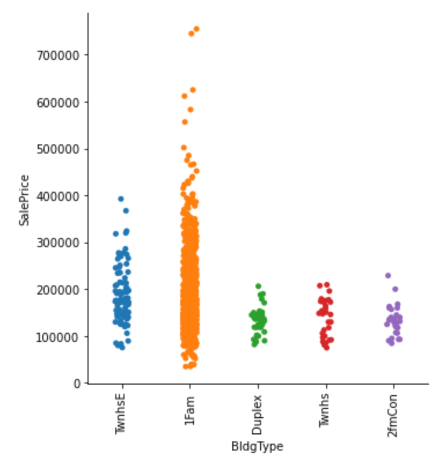

SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Inside
CulDSac
FR2
Corner
FR3
LotConfig

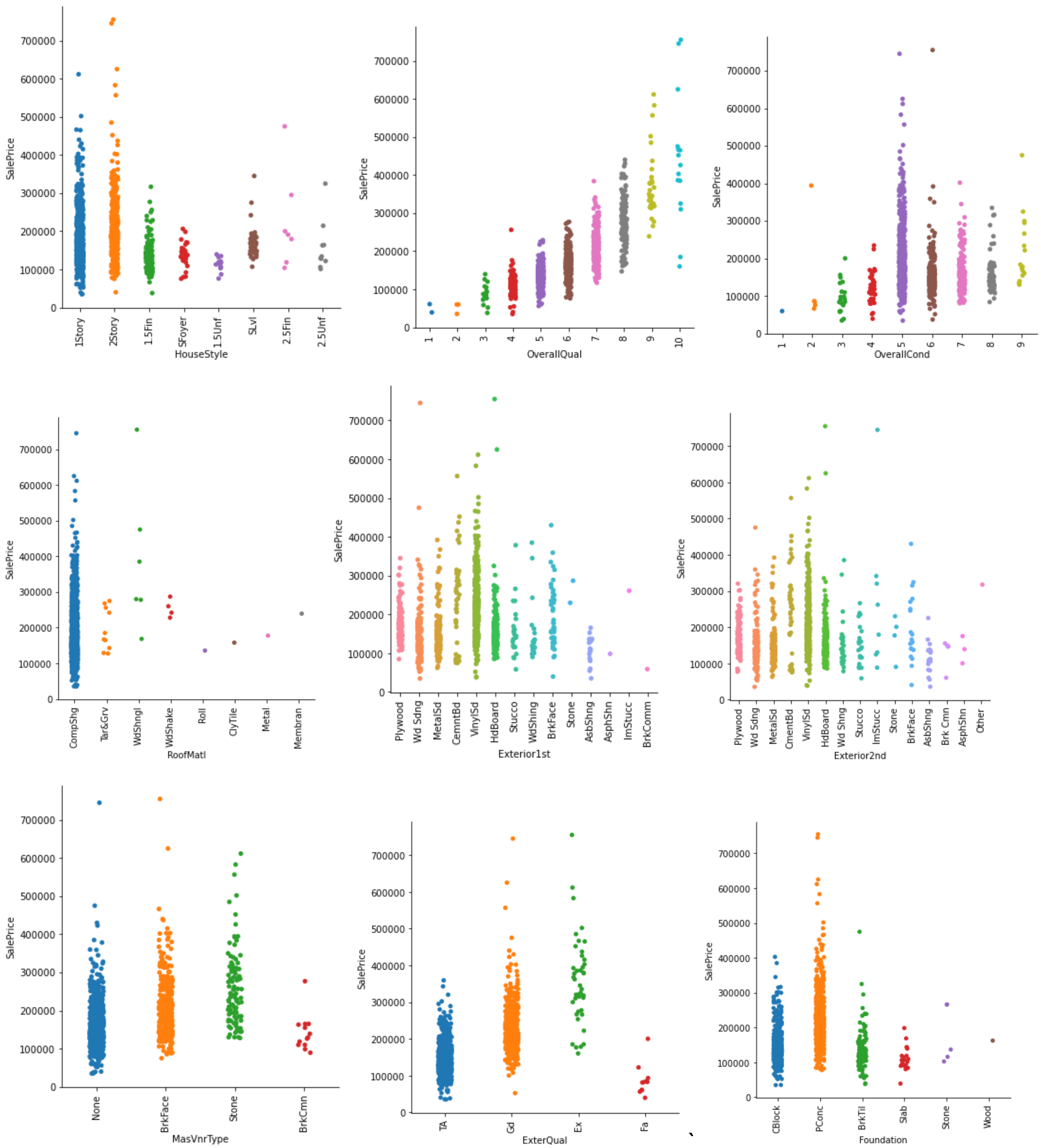

SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Gtl
Mod
Sev
LandSlope

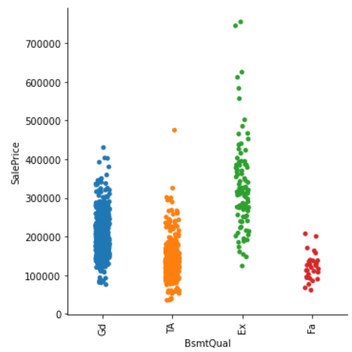

SalePrice
700000
600000
500000
400000
300000
200000
100000
0
NPkVill
NAmes
NoRidge
NWAmes
Gilbert
Sawyer
Edwards
IDOTRR
CollgCr
Mitchel
Crawfor
BrDale
StoneBr
BrkSide
NridgHt
OldTown
Somerst
Timber
SWISU
SawyerW
ClearCr
Veenker
Blmngtn
MeadowV
Blueste
Neighborhood

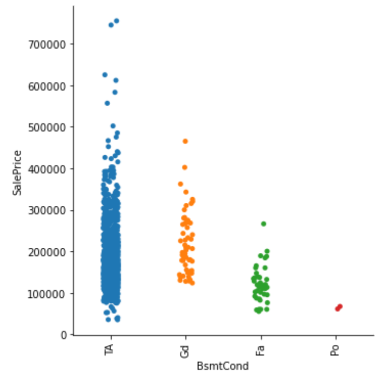

SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Norm
Feedr
RRAn
PosA
RRAe
Artery
PosN
RRNe
RRNn
Condition1

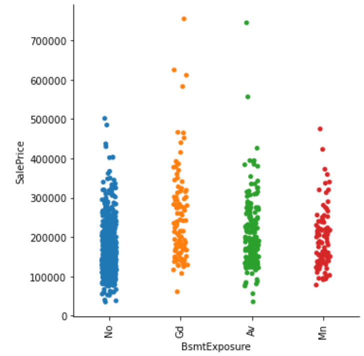

SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Norm
RRAe
Feedr
PosN
Artery
RRNn
PosA
RRAn
Condition2

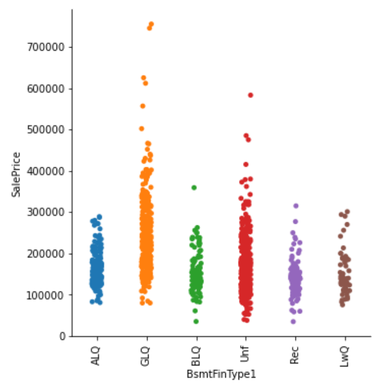

SalePrice
700000
600000
500000
400000
300000
200000
100000
0
TwnhsE
1Fam
Duplex
Twnhs
2fmCon
BldgType

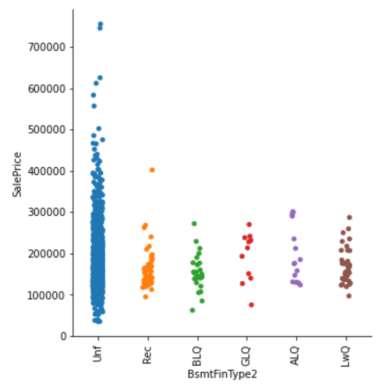

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
1Story
2Story
1.5Fin
SFoyer
1.5Unf
SLvl
2.5Fin
2.5Unf
HouseStyle
1
2
3
4
5
6
7
8
9
10
OverallQual
1
2
3
4
5
6
7
8
9
OverallCond
CompShg
Tar&Grv
WdShngl
WdShake
Roll
ClyTile
Metal
Membran
RoofMatl
Plywood
Wd Sdng
MetalSd
CemntBd
VinylSd
HdBoard
Stucco
WdShing
BrkFace
Stone
AsbShng
AsphShn
ImStucc
BrkComm
Exterior1st
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
CBlock
PConc
BrkTil
Slab
Stone
Wood
Foundation

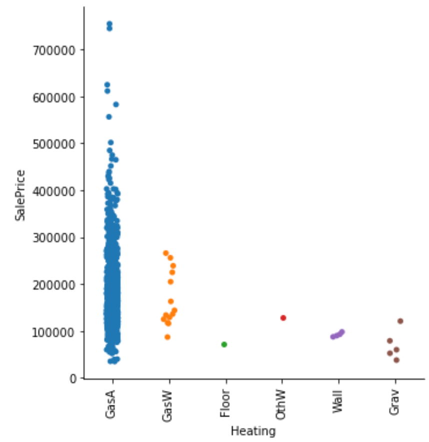

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
Gd
TA
Ex
Fa
BsmtQual

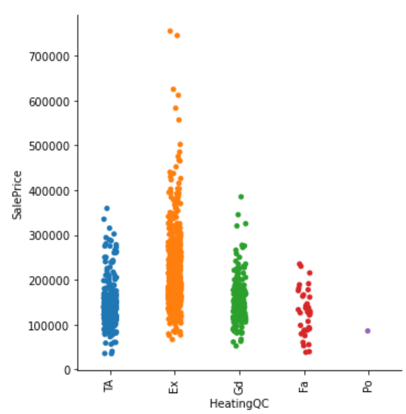

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
TA
Gd
Fa
Po
BsmtCond

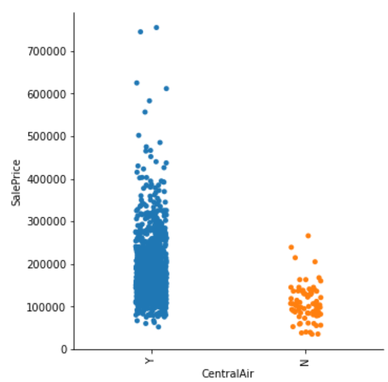

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
No
Gd
Av
Mn
BsmtExposure

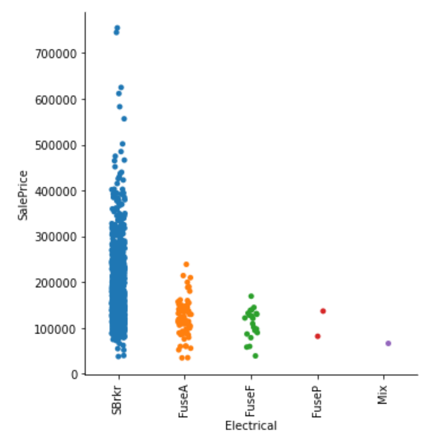

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
ALQ
GLQ
BLQ
Unf
Rec
LwQ
BsmtFinType1

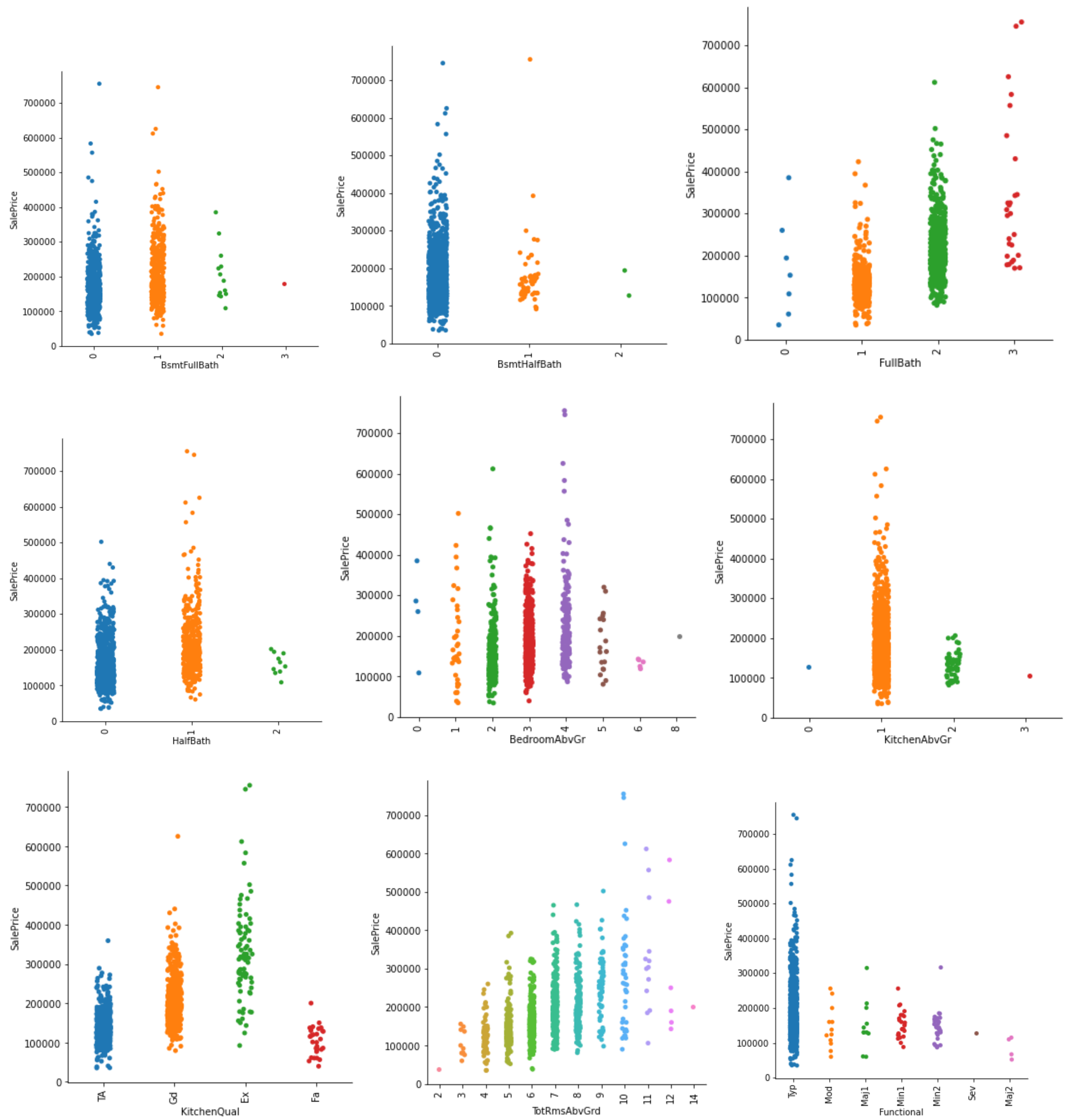

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
Unf
Rec
BLQ
GLQ
ALQ
LwQ
BsmtFinType2

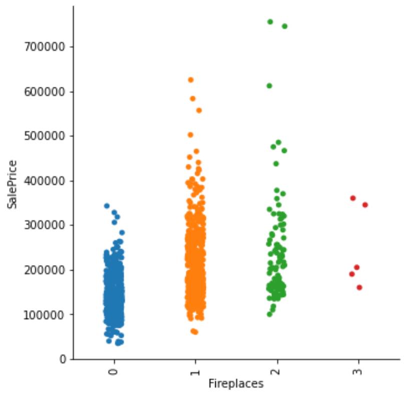

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
GasA
GasW
Floor
OthW
Wall
Grav
Heating

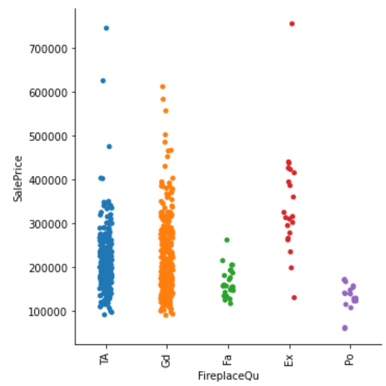

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
TA
Ex
Gd
Fa
Po
HeatingQC

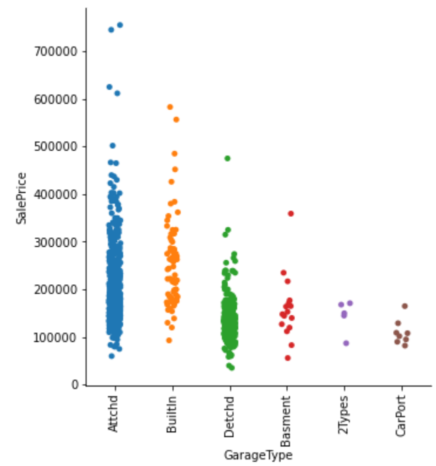

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
Y
N
CentralAir

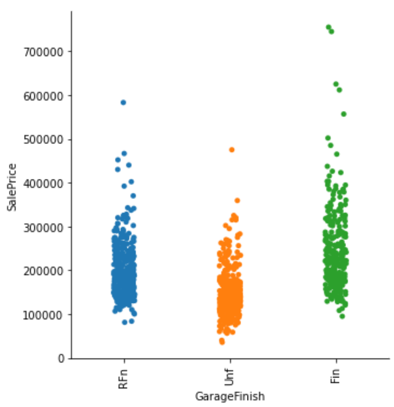

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
SBrkr
FuseA
FuseF
FuseP
Mix
Electrical

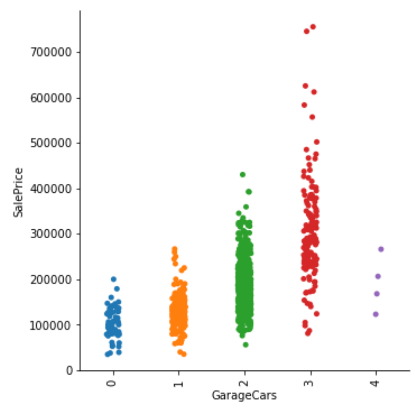

SalePrice
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
KitchenQual
TA
Gd
Ex
Fa
TotRmsAbvGrd
Functional
Typ
Mod
Maj1
Min1
Min2
Sev
Maj2

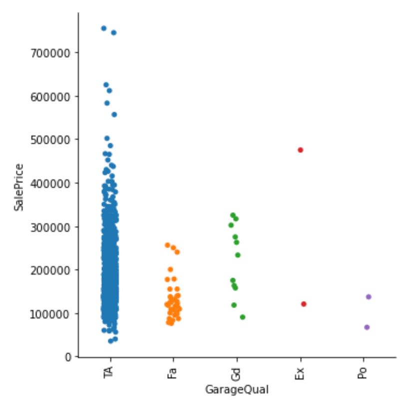

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
0
1
2
3
Fireplaces

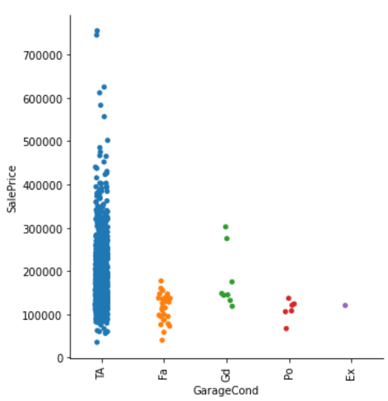

SalePrice
100000
200000
300000
400000
500000
600000
700000
TA
Gd
Fa
Ex
Po
FireplaceQu

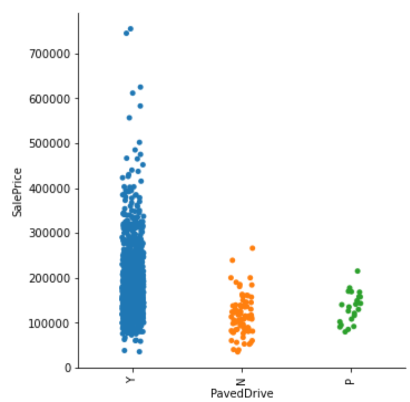

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
Attchd
BuiltIn
Detchd
Basment
2Types
CarPort
GarageType

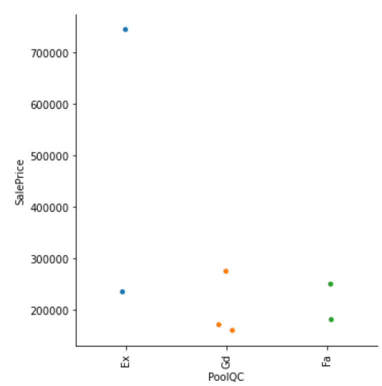

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
RFn
Unf
Fin
GarageFinish

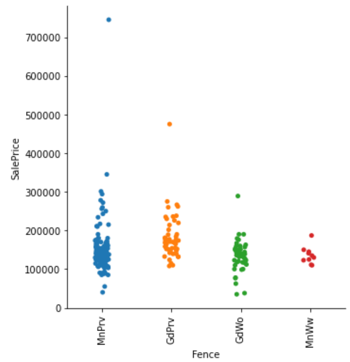

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
0
1
2
3
4
GarageCars

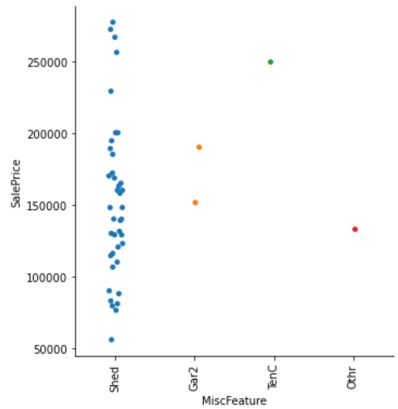

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
TA
Fa
Gd
Ex
Po
GarageQual

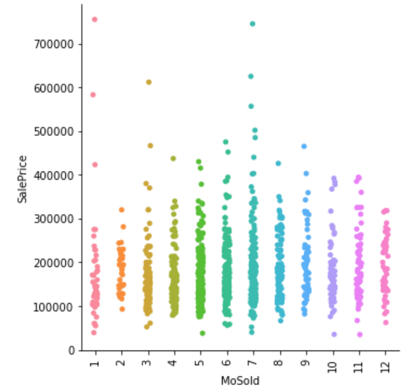

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
TA
Fa
Gd
Po
Ex
GarageCond

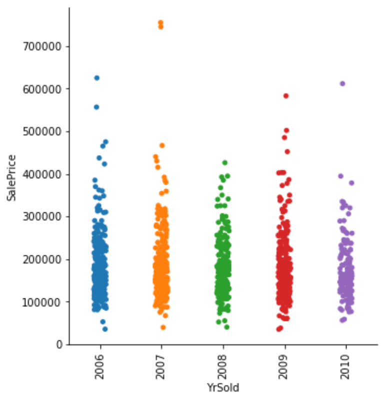

SalePrice
0
100000
200000
300000
400000
500000
600000
700000
Y
N
P
PavedDrive

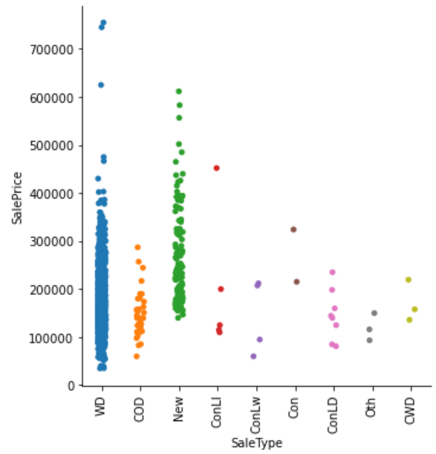

SalePrice
200000
300000
400000
500000
600000
700000
Ex
Gd
Fa
PoolQC

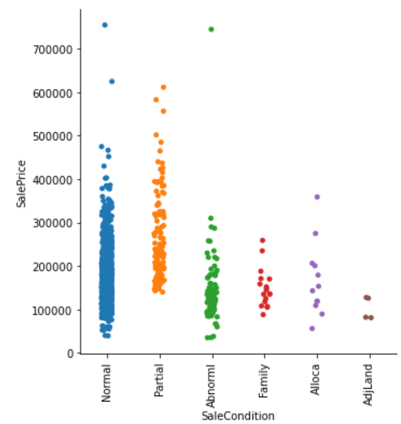








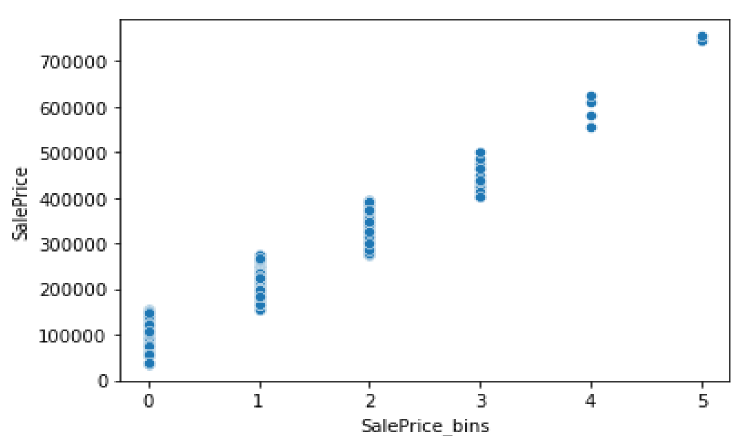




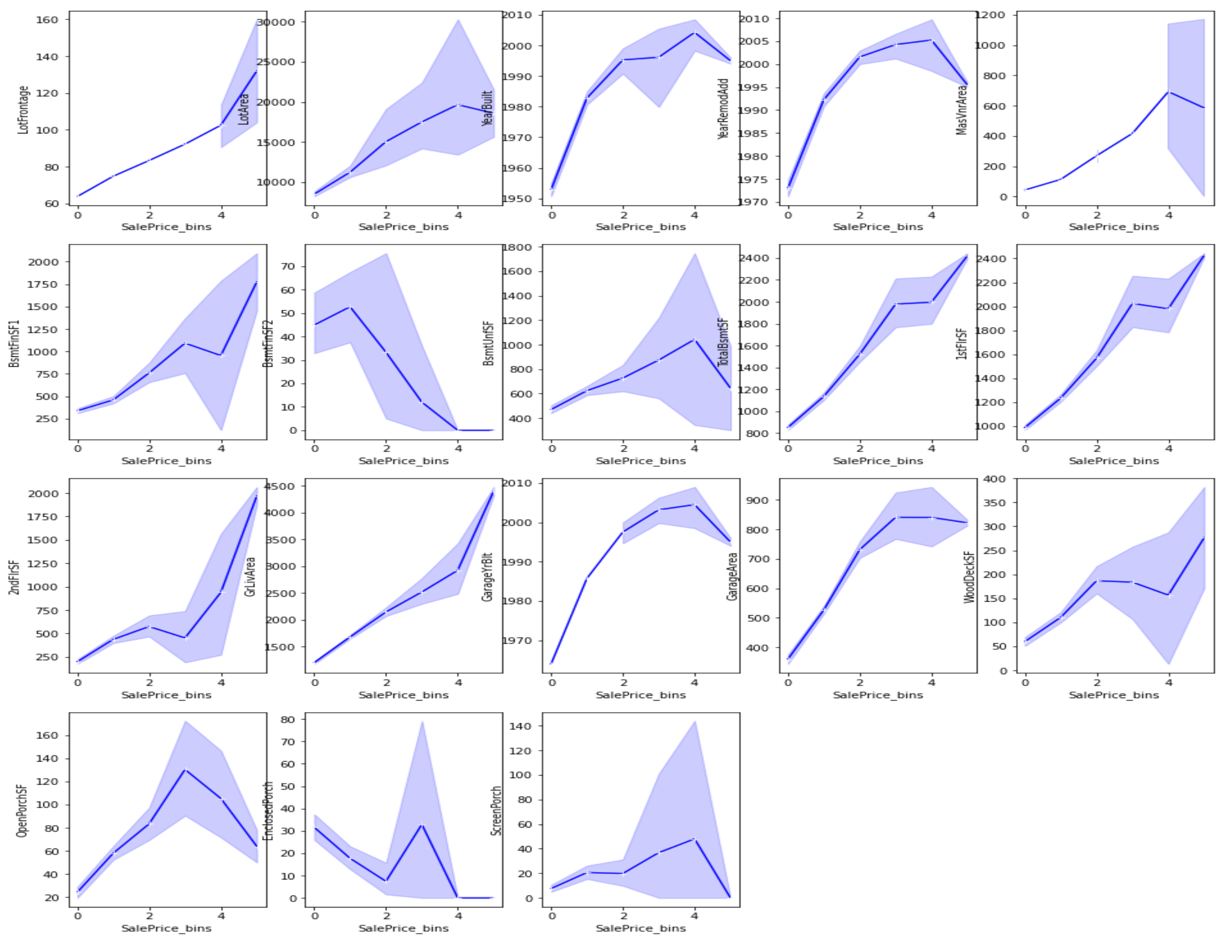
The almost uniform distribution shows us that there are more new people to

Continuous Data Analysis

We put divided the price into 6 bins to compare the continuous data,



When the slope is positive that means at that Element of feature will effect that range of price.



CONCLUSION

Key Findings and Conclusions of the Study

1. There are many elements of features which directly depend on the Sales, we can see in the table below.

FEATURES	*ELEMENT*
MSSubClass	*60*
MSZoning	*RL*
Street	*Pave*
Alley	*Pave*
Lotshape	*IR1*
Landcontour	*Lvl*
LotConfig	*Corner*
LandSlope	*Gtl*
Neighbourhood	*NoRidge*
Condition1	*norm*
Condition2	*norm*
BldgType	*1Fam*
HouseStyle	*2Story*
OverallQuall	*10*
OverallCond	*5*
RoofMatl	*CompShg*
Exterior 1st	*Wd Sdng*
Exterior 2nd	*HdBoard*
MasVnrType	*none*

ExterQuall	*Gd*
Foundation	*PConc*
BsmtQuall	*Ex*
BsmtCond	*TA*
BsmtExposure	*Av*
BsmtFin Type1	*GLQ*
BsmtFin Type2	*UNA*
Heating	*GasA*
HeatingQC	*Ex*
CentralAir	*Y*
Electrical	*SBrkr*
BsmtFullBath	*1*
BsmtHalfBath	*0*
FullBath	*3*
HalfBath	*1*
BedroomAbvGr	*4*
KitchenAbvGR	*1*
KitchenQual	*Ex*
TotRmsAbvGrd	*10*
Functional	*Typ*

FirePlaces	*2*
FirePlaceQU	*TA*
GarageTypeAttchd	*Attchd*
GarageFinish	*Fin*
GarageCars	*3*
GarageQual	*TA*
GarageCond	*TA*

PavedDrive	*Y*
PoolQc	*Ex*
Fence	*MnPrv*
MiscFeature	*Shed*
MoSold	*7*
YrSold	*2007*
SaleType	*WD*
SaleCondition	*Normal*

2. Most of the Continuous Feature have positive overall positive relation with sale price.
3. Removing High Variance features can decrease the effectiveness of the model.

Learning Outcomes of the Study in respect of Data Science

1. We see how to deal with outliers when all the rows has at least one value Z>3.
2. To do a visualisation when data has high standard deviation and no classification
3. Ways to select features and to do hyperparameter tuning efficiently
4. Ways of removing skewness and what are the best methods still not versatile when it comes to data with 0 value
5. How to make a model using a pipeline.

Limitations of this work and Scope for Future Work

The high skewness of data reduces the effectivity

Many features have NaN value more than 50%, and imputation of them can decrease the effectiveness. And dropping them had the loss of data.

we can increase the efficiency of a model by selecting a better method to remove outliers and skewness also how to make the search of perfect model in a way that if we want to change some parameters in model then we don't have to run all the model again