FLIP ROBO

Housing Price Prediction Project

Submitted by:

Sahil Kumar

Acknowledgment

I'd like to acknowledge the invaluable assistance of the Data Trained Academy and FlipRobo Technologies teams in providing guidance to work on real-time data projects, which also assisted me in doing a lot of research and obtaining insights.

FlipRobo Technologies have provided all the necessary information as well as the dataset.

Introduction

Business Problem Framing

Houses are one of the necessary needs of 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 requirement is to build a model using Machine Learning to predict the actual value of the prospective properties and decide whether to invest in them or not. The primary questions that needed answering were:

- Which variables are important to predict the price of variable?
- How do these variables describe the price of the house?

The task was to model house prices using the available independent variables. The management will then use this model to understand how the prices vary with the variables. As an outcome, they can manipulate the firm's strategy and focus on areas that will yield high returns. Furthermore, the model will assist management in comprehending the pricing dynamics of a new market.

Conceptual Background of the Domain Problem

Despite the size and complexity of the real estate market, many people believe it is only comprised of brokers and salespeople. However, millions of people make a living in the real estate industry, not only in sales but also in appraisals, property management, financing, construction, development, counselling, education, and a variety of other fields.

Housing market is a critical driver of economic growth in most countries across the world.

Investors and analysts keep a close eye on housing because the numbers can provide a general sense of economic direction. Moreover, the type of new housing can give clues about how the economy is developing. (Chen 2021, Investopedia)

Based on the dataset provided, domain knowledge of the following areas was obtained:

- type of dwelling involved in the sale
- general zoning classification of the sale
- linear feet of street connected to property
- lot size in square feet
- type of road access to property
- type of alley access to property
- general shape of property
- flatness of the property
- type of utilities available
- lot configuration
- slope of property
- physical locations within Ames city limits
- proximity to various conditions
- type of dwelling

- style of dwelling
- rating of the overall material and finish of the house
- rating of the overall condition of the house
- original construction date
- remodel date
- type of roof
- roof material
- exterior covering on house
- masonry veneer type
- masonry veneer area in square feet
- quality of the material on the exterior
- present condition of the material on the exterior
- type of foundation
- height of the basement
- general condition of the basement
- walkout or garden level walls
- rating of basement finished area
- type 1 finished square feet
- type 2 finished square feet
- unfinished square feet of basement area
- total square feet of basement area
- type of heating
- heating quality and condition
- whether there is central air conditioning
- type of electrical system
- first floor square feet
- second floor square feet
- low quality finished square feet (all floors)
- above grade (ground) living area square feet
- basement full bathrooms
- basement half bathrooms
- full bathrooms above grade
- half baths above grade
- bedrooms above grade
- kitchens above grade
- kitchen quality
- total rooms above grade
- home functionality
- number of fireplaces
- fireplace quality
- garage location
- year garage was built
- interior finish of the garage
- size of garage in car capacity
- size of garage in square feet
- garage quality
- garage condition
- whether there is paved driveway
- wood deck area in square feet
- open porch area in square feet
- enclosed porch area in square feet
- three season porch area in square feet
- screen porch area in square feet
- pool area in square feet
- pool quality
- fence quality
- features not covered in other categories

- value of miscellaneous feature
- type of sale
- condition of sale

The problem statement is related to the 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.

Review of Literature

The relationship between house prices and the economy is a major motivator for forecasting house prices. In real estate transactions, the value of a property is critical. Housing price trends are not only a source of concern for buyers and sellers, but they also provide insight into the current economic situation. As a result, it is critical to predict housing prices without bias to assist both buyers and sellers in making decisions. Prospective homeowners, developers, investors, appraisers, tax assessors, and other real estate market participants, such as mortgage lenders and insurers, rely on accurate house price predictions. The advancement of civilization is the foundation for the increasing demand for housing on a regular basis. (Konwar et al. 2021)

Many studies have been conducted to train models to predict house prices in a specific region. There are studies in which the authors predict using various machine learning algorithms.

Bruno Klaus de Aquino Afonso, Luckeciano Carvalho Melo, Willian Dihanster Gomes de Oliveira, Samuel Bruno da Silva Sousa, and Lilian Berto conducted research. They predict using a dataset of 12,223,582 housing advertisements collected from Brazilian websites between 2015 and 2018. Each instance contains twenty-four different data types: integer, date, string, float, and image. To predict property prices, they combine two different Machine Learning architectures based on Random Forest (RF) and Recurrent Neural Networks (RNN). Their research on the application of machine learning algorithms shows that enriching the dataset and combining different ML approaches can be a better alternative for predicting housing prices in Brazil.

Motivation for the Problem Undertaken

The relationship between house prices and the economy is a major motivator for forecasting house prices. As a result, it is critical to predict housing prices without bias to assist both buyers and sellers in making decisions. This project is proposed to better predict house prices and obtain more accurate results. Various algorithms are used in this project to determine which algorithm produces the most accurate and precise results.

This would be extremely beneficial to the people because house pricing is a topic that many citizens, rich and poor, are concerned about, as one cannot judge or estimate the pricing of a house based on its location or the amenities available. Python programming language was used within the Jupyter Notebook to build the prediction model using machine learning.

This project provided insights on how to deal with data exploration using in-depth analysis skills to predict house prices using machine learning models. The model should prove to be a good start for the organization to understand the pricing dynamic of the new market they are planning to enter in Australia.

Analytical Problem Framing

Mathematical / Analytical Modelling of the Problem

The house pricing model relies on a housing demand function and a conventional utility life-cycle model for a typical household. This is a frequent method used in academic studies of housing prices. The goal of the research is to forecast the house's sale price by examining which features are essential and how they contribute to the prediction.

The target variable was the sale price of the house. We were provided with two datasets – a training dataset and a test dataset. The training dataset was used to obtain an understanding on the information provided, conduct exploratory data analysis, understanding what attributes were irrelevant, pre-processing the data, conduct data visualizations, encoding of categorical data, establishing correlations with sale prices of the houses, carrying out principal component analysis and finally, building the prediction model.

Columns containing records with mostly zeros (more than 85%) and null values were dropped, while other columns having nulls values were filled using appropriate techniques.

Also, the data containing years was converted to age for obtaining clearer insights and establishing effective correlations with the target variable.

The categorical and numerical features were analysed using appropriate bar and pie plots for conducting a thorough analysis of the features, and eventually understanding the importance of each attribute in relation to the target, as well as on its own. This helped to draw some unexpected insights, especially from attributes that provided a rating for certain factors affecting sale prices of houses.

Moreover, the statistical summary was obtained through the description function to gain some meaningful insights about the features in the dataset, including the presence of skewness, distribution of data, and outliers. The heatmap and bar plot help establish an understanding of the correlation of various features with each other, as well as with the target variable.

The training dataset was used to train the machine learning model for prediction of sale prices of houses, and the accuracy and cross validation scores were verified for four different algorithms – linear regression, random forest, support vector regression and decision tree. Based on the results, hyperparameter tuning was carried out for improving the accuracy of the prediction model, which was also plotted. The best model was used to predict the sale prices in the test dataset, post bringing the test dataset at the same level as the training dataset.

Data Sources and their formats

A US based housing company – Surprise Housing collected the data from the sale of houses in Australia, as they were expanding to this market. The training and test datasets, both, were provided by FlipRobo Technologies for the purpose of building the machine learning based prediction model, in the hopes of helping out Surprise Housing with a basis for purchasing properties in this new market at appropriate prices, primarily in order the mitigate the risk arising out of internationalization of operations, new market entry and lack of knowledge on key factors playing a definitive role in deciding the prices of houses on the real estate market. The datasets were provided in CSV formats. The training dataset contained 1168 records across 80 features, plus 1 target variable – the sale price of the house. The test dataset however, contained 292 records of houses on sale, or with the possibility of entering the on-sale market with the same 80 features, wherein the sale price was required to be predicted because of our prediction model. A snapshot of the training dataset is provided below, as viewed in MS Excel.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Id	MSSubCla	MSZoning	LotFrontag	LotArea	Street	Alley	LotShape	LandConto	Utilities	LotConfig	LandSlope	Neighborh	Condition1	Condition2	BldgType	HouseStyl	OverallQua	OverallCon	YearBuilt
2	127	120	RL		4928	Pave		IR1	Lvl	AllPub	Inside	Gtl	NPkVill	Norm	Norm	TwnhsE	1Story	6	5	1976
3	889	20	RL	95	15865	Pave		IR1	Lvl	AllPub	Inside	Mod	NAmes	Norm	Norm	1Fam	1Story	8	6	1970
4	793	60	RL	92	9920	Pave		IR1	Lvl	AllPub	CulDSac	Gtl	NoRidge	Norm	Norm	1Fam	2Story	7	5	1996
5	110	20	RL	105	11751	Pave		IR1	Lvl	AllPub	Inside	Gtl	NWAmes	Norm	Norm	1Fam	1Story	6	6	1977
6	422	20	RL		16635	Pave		IR1	Lvl	AllPub	FR2	Gtl	NWAmes	Norm	Norm	1Fam	1Story	6	7	1977
7	1197	60	RL	58	14054	Pave		IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story	7	5	2006
8	561	20	RL		11341	Pave		IR1	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	6	1957
9	1041	20	RL	88	13125	Pave		Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	4	1957
10	503	20	RL	70	9170	Pave		Reg	Lvl	AllPub	Corner	Gtl	Edwards	Feedr	Norm	1Fam	1Story	5	7	1965
11	576	50	RL	80	8480	Pave		Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1.5Fin	5	5	1947
12	449	50	RM	50	8600	Pave		Reg	Bnk	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin	6	6	1937
13	833	60	RL	44	9548	Pave		IR1	Lvl	AllPub	CulDSac	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	6	2003
14	277	20	RL	129	9196	Pave		IR1	Lvl	AllPub	Inside	Gtl	Mitchel	Norm	Norm	1Fam	1Story	7	5	2003
15	84	20	RL	80	8892	Pave		IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	5	1960
16	888	50	RL	59	16466	Pave		IR1	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1.5Fin	5	7	1955
17	1013	70	RL	55	10592	Pave		Reg	Lvl	AllPub	Inside	Gtl	Crawfor	Norm	Norm	1Fam	2Story	6	7	1923
18	1154	30	RM		5890	Pave		Reg	Lvl	AllPub	Corner	Gtl	IDOTRR	Norm	Norm	1Fam	1Story	6	8	1930
19	728	20	RL	64	7314	Pave		Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	7	5	2007
20	270	20	RL		7917	Pave		IR1	Lvl	AllPub	Corner	Gtl	Edwards	Norm	Norm	1Fam	1Story	6	7	1976
21	1105	160	RM	24	2016	Pave		Reg	Lvl	AllPub	Inside	Gtl	BrDale	Norm	Norm	TwnhsE	2Story	5	5	1970
22	259	60	RL	80	12435	Pave		Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	2001
23	1407	85	RL	70	8445	Pave		Reg	Lvl	AllPub	Corner	Gtl	CollgCr	Norm	Norm	1Fam	SFoyer	5	7	1972
24	1459	20	RL	68	9717	Pave		Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	6	1950
25	997	20	RL		10659	Pave		IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	6	1961

The datasets contained a mix of categorical (nominal and ordinal) and numerical (continuous and discrete) variables.

Data Preprocessing Done

1. Imported necessary libraries and loaded the training dataset for pre-processing
2. Conducted exploratory data analysis including analysing the shape of the dataframe, summarising the list of columns, their data types, and the sum of null records in each, and the analysis of unique values in each column
3. Dropped “Id” and “Utilities” as they contained all unique values and only one unique value respectively.
4. Value count of each column was analysed, followed by an analysis of the quantum of zero values in each column, which led to the dropping of columns containing zero values in more than 85% of the records.
5. Based on the year in which the house was sold, age of construction of house and garage, and any remodelling done were computed, following which the columns containing the respective values in terms of Year were dropped.
6. The quantum of null values across each independent feature was analysed, and columns containing mostly null values (over 75%) were dropped as no independent and collective method for filling these values could be ascertained. The null values in the remaining columns were filled appropriately, i.e., using mean for numerical values and mode for categorical values.
7. Post data visualization and the extraction of relevant insights therein, outliers were checked and removed to a great extent using the percentile method, since both z-score and IQR methods were resulting in massive data losses.
8. The skewness in the features was removed by using the power transformer with the Yeo-Johnson method.
9. The column KitchenAbvGr was dropped due to all zero values being present in the column post the steps carried out above, leaving us with 66 columns of data.
10. The data in categorical columns was encoded using the Ordinal Encoder, and post establishing the relevant correlations, Standard Scaler was applied to scale the data.
11. A principal component analysis was carried out and 50 independent variables were selected covering around 95% of the variance %.
12. Similar pre-processing was carried out on the test dataset for the purpose of predicting the sale price of the houses, after the best prediction model had been built, i.e., using the Random Forest Regression algorithm.

Data Inputs- Logic- Output Relationships

The relationship between the features and the target was analysed through exploratory data analysis and through various data visualizations (univariate, bivariate and multivariate). The univariate data visualizations involved deriving insights on the distribution and count of various unique records across distinct columns/features in the dataset. Visualizations for bivariate analysis included understanding the relationships between sale price of a house and the various features provided in the dataset. Using a pair plot, multivariate analysis was conducted to understand the relationships of various features with each other.

Visualizing the correlation between sale price of a house and various features revealed columns with highly positive, as well as highly negative correlations.

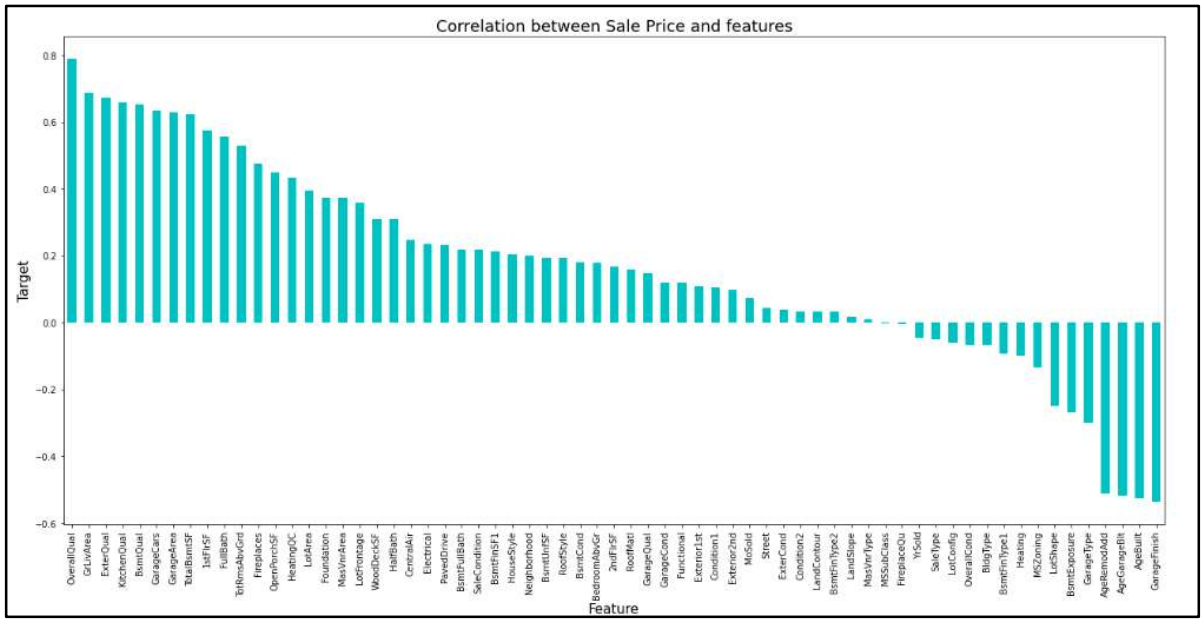
Features with highly positive correlations were as follows:

OverallQual	0.789185
GrLivArea	0.688210
ExterQual	0.672665
KitchenQual	0.659228
BsmtQual	0.653265
GarageCars	0.634573
GarageArea	0.627504
TotalBsmtSF	0.624311
1stFlrSF	0.575665
FullBath	0.554988
TotRmsAbvGrd	0.528564

Features with highly negative correlations with the sale price of a house were as follows:

LotShape	-0.248171
BsmtExposure	-0.268559
GarageType	-0.299470
AgeRemodAdd	-0.510784
AgeGarageBlt	-0.516445
AgeBuilt	-0.526644
GarageFinish	-0.537121

This has also been visualized using a bar plot as follows:



Assumptions related to the problem under consideration

1. Columns dropped because of the presence of zeros or null values did not adversely impact the prediction accuracy from the machine learning model built using the dataset.
2. The imputing techniques used to handle missing or null values in the dataset did not adversely impact the correlations derived post imputing the record values.

Hardware and Software Requirements and Tools Used

Hardware:

Machine - Lenovo ThinkPad
Processor - Intel(R) Core (TM) i5-8265U CPU @ 1.60GHz 1.80 GHz
RAM – 8 GB
System Type - 64-bit operating system, x64-based processor

Software:

Windows 10 Pro
MS Office 365
Jupyter Notebook for coding in Python and building the machine learning model

Libraries:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
from scipy.stats import zscore
from sklearn.preprocessing import PowerTransformer
from sklearn.preprocessing import OrdinalEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
from sklearn.model_selection import train_test_split as TTS
from sklearn.decomposition import PCA
from sklearn.ensemble import RandomForestRegressor
from sklearn.linear_model import LinearRegression
from sklearn.svm import SVR
from sklearn.tree import DecisionTreeRegressor
from sklearn.metrics import classification_report
from sklearn.model_selection import cross_val_score
from sklearn import metrics
from sklearn.model_selection import GridSearchCV
import joblib

import warnings
warnings.filterwarnings('ignore')
```

- **NumPy** is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays
- **Pandas** is a software library written for the Python programming language for data manipulation and analysis. It offers data structures and operations for manipulating numerical tables and time series.
- **Seaborn** is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
- **Matplotlib** is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits.
- **Scipy.stats** contains several probability distributions, summary and frequency statistics, correlation functions and statistical tests, masked statistics, kernel density estimation, quasi-Monte Carlo functionality, and more
- **Scikit-learn (also known as sklearn)** is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. Scikit-learn is a NumFOCUS fiscally sponsored project.
- **Joblib** is a set of tools to provide lightweight pipelining in Python. Transparent disk-caching of functions and lazy re-evaluation (memorize pattern) and easy simple parallel computing. Joblib is optimized to be fast and robust on large data and has specific optimizations for NumPy arrays.

NumPy and Pandas libraries were used to process the dataset, seaborn and matplotlib were used in the visualization of data, Scipy.stats was used for attempting to remove outliers using z-score, skearn was used in pre-processing the dataset (power transformer, encoding categorical data and standard scaling), decomposition (in principal component analysis), creating a train test split, computing the metrics like MAE, MSE and R^2 score, importing relevant regression algorithms for building the model, computing the cross validation scores, and hyperparameter tuning. Joblib was used exclusively for saving the model built on the training dataset and reloading the model for predicting sale prices of houses on the test dataset.

Model Development and Evaluation

Identification of possible problem-solving approaches (methods)

- Use of appropriate and relevant imputation techniques for handling missing or null values
- Checking the best method for removal of outliers using z-score, IQR and percentile method, and using the percentile method to remove the outliers
- Removal of skewness using power transformation with the Yeo-Johnson method
- Encoding of categorical data using the ordinal encoder
- Understanding of relationships between variables using data visualization techniques
- Understanding of correlation between the Target and the Variables using correlation coefficient
- Data scaling to remove bias using the Standard Scaler
- Principal Component Analysis to choose the effective variables for building the prediction model
- Use of four machine learning models to build the best prediction model for sale price of a given house with the tools and libraries utilised.

Testing of Identified Approaches (Algorithms)

1. Linear Regression
2. Random Forest Regression
3. Support Vector Regression (SVR)
4. Decision Tree Regression

Run and evaluate selected models

Linear Regression: It is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

Random Forest Regression: It is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

Support Vector Regression (SVR): It is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.

Decision Tree Regression: Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

Finding best random state

```
maxAccu=0
maxRS=0
for i in range(1,200):
    x_train,x_test,y_train,y_test = train_test_split(train_df_x,y,test_size=.30, random_state=i)
    mod = RandomForestRegressor()
    mod.fit(x_train, y_train)
    pred = mod.predict(x_test)
    acc=r2_score(y_test, pred)
    if acc>maxAccu:
        maxAccu=acc
        maxRS=i
print("Maximum r2 score is ",maxAccu," on Random_state ",maxRS)
```

```
Maximum r2 score is  0.8929439261244873  on Random_state  181
```

```
x_train,x_test,y_train,y_test=train_test_split(train_df_x,y,test_size=.30,random_state=maxRS)
```

```
lr = LinearRegression()
rf = RandomForestRegressor()
svr = SVR()
dt = DecisionTreeRegressor()
```

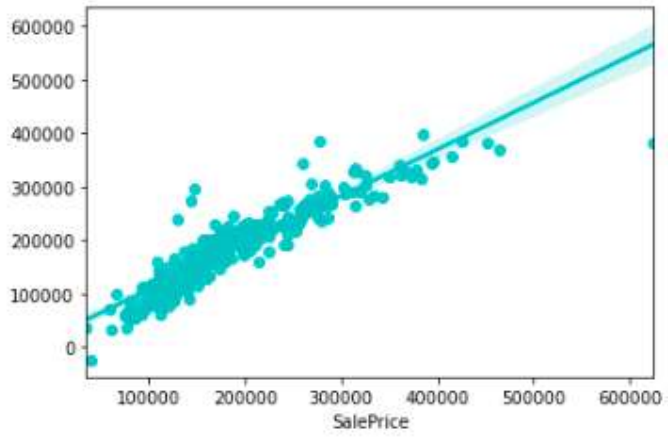
```
lr.fit(x_train,y_train)
rf.fit(x_train,y_train)
svr.fit(x_train,y_train)
dt.fit(x_train,y_train)
```

```
DecisionTreeRegressor()
```

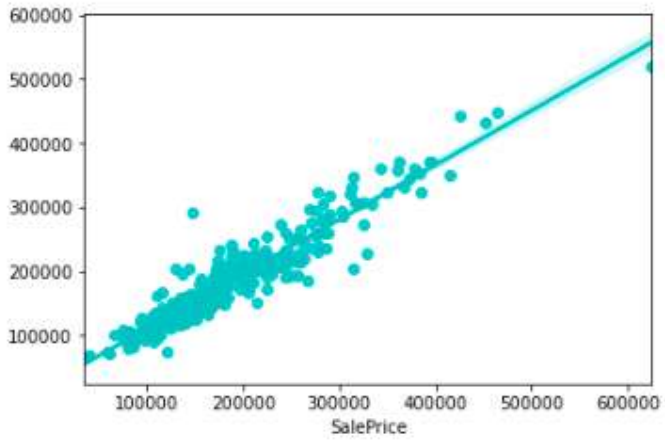
```
print("-"*50)
print("Linear Regression Model")
print("-"*50)
lr_pred = lr.predict(x_test)
print("R2 Score: ", r2_score(y_test,lr_pred), "\n")
print("Mean Squared Error: ", mean_squared_error(y_test,lr_pred), "\n"*2)

print("-"*50)
print("Random Forest Model")
print("-"*50)
rf_pred = rf.predict(x_test)
print("R2 Score: ", r2_score(y_test,rf_pred), "\n")
print("Mean Squared Error: ", mean_squared_error(y_test,rf_pred), "\n"*2)

print("-"*50)
print("Support Vector Regression Model")
print("-"*50)
svr_pred = svr.predict(x_test)
print("R2 Score: ", r2_score(y_test,svr_pred), "\n")
print("Mean Squared Error: ", mean_squared_error(y_test,svr_pred), "\n"*2)

print("-"*50)
print("Decision Tree Model")
print("-"*50)
dt_pred = dt.predict(x_test)
print("R2 Score: ", r2_score(y_test,dt_pred), "\n")
print("Mean Squared Error: ", mean_squared_error(y_test,dt_pred), "\n"*2)
```

```
--------------------------------------------------
Linear Regression Model
--------------------------------------------------
R2 Score:  0.841111736679627

Mean Squared Error:  955107746.7852867


--------------------------------------------------
Random Forest Model
--------------------------------------------------
R2 Score:  0.8936454816430158

Mean Squared Error:  639317355.8927604


--------------------------------------------------
Support Vector Regression Model
--------------------------------------------------
R2 Score:  -0.07869205387757017

Mean Squared Error:  6484224294.004828


--------------------------------------------------
Decision Tree Model
--------------------------------------------------
R2 Score:  0.6050881181639713

Mean Squared Error:  2373890869.954416
```

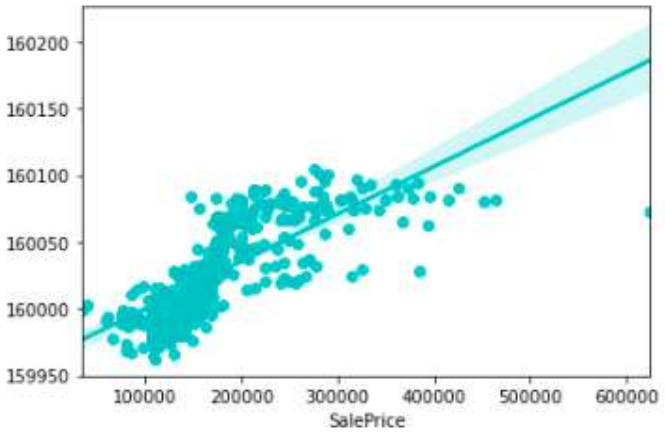
```
sns.regplot(y_test,lr_pred,color='c')
plt.show()
```



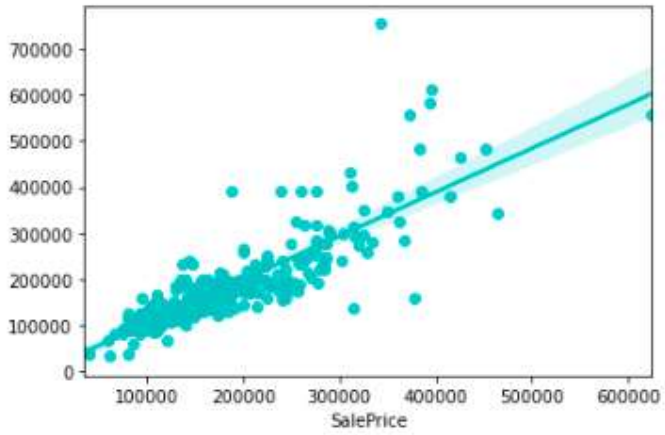
```
sns.regplot(y_test,rf_pred,color='c')
plt.show()
```




```
sns.regplot(y_test,svr_pred,color='c')
plt.show()
```




```
sns.regplot(y_test,dt_pred,color='c')
plt.show()
```



```
print("Cross Validation Score for Linear Regression Model: ", cross_val_score(lr,train_df_x,y,cv=5).mean(), "\n"*2)
print("Cross Validation Score for Random Forest Model: ", cross_val_score(rf,train_df_x,y,cv=5).mean(), "\n"*2)
print("Cross Validation Score for Support Vector Regression Model: ", cross_val_score(svr,train_df_x,y,cv=5).mean(), "\n"*2)
print("Cross Validation Score for Decision Tree Model: ", cross_val_score(dt,train_df_x,y,cv=5).mean())
```

```
Cross Validation Score for Linear Regression Model:  0.8010670703560805


Cross Validation Score for Random Forest Model:  0.8283693628416016


Cross Validation Score for Support Vector Regression Model:  -0.06166040986855133


Cross Validation Score for Decision Tree Model:  0.638444191216726
```

Random Forest appeared to be the best model for our prediction of sale prices of houses.

Hyperparameter Tuning

```
parameters = {'n_estimators':[10,50,100],
              'max_depth':[10,20,30],
              'max_features':['auto','sqrt'],
              'min_samples_leaf': [2, 3, 5],
              'min_samples_split': [2, 5, 8]
             }
```

```
GCV = GridSearchCV(RandomForestRegressor(),parameters,cv=5)
```

```
GCV.fit(x_train,y_train)
```

```
GridSearchCV(cv=5, estimator=RandomForestRegressor(),
             param_grid={'max_depth': [10, 20, 30],
                         'max_features': ['auto', 'sqrt'],
                         'min_samples_leaf': [2, 3, 5],
                         'min_samples_split': [2, 5, 8],
                         'n_estimators': [10, 50, 100]})
```

```
GCV.best_params_
```

```
{'max_depth': 10,
 'max_features': 'auto',
 'min_samples_leaf': 2,
 'min_samples_split': 8,
 'n_estimators': 50}
```

```
Final_model = RandomForestRegressor(max_depth=10, max_features='auto', min_samples_leaf=2, min_samples_split=8, n_estimators=50)
Final_model.fit(x_train,y_train)
pred = Final_model.predict(x_test)
print('R2_Score:',r2_score(y_test,pred)*100)
print("RMSE value:",np.sqrt(metrics.mean_squared_error(y_test, pred)))
print('MAE:',metrics.mean_absolute_error(y_test, pred))
print('MSE:',metrics.mean_squared_error(y_test, pred))
```

```
R2_Score: 89.35081751912671
RMSE value: 25301.04209793543
MAE: 17860.601375220773
MSE: 640142731.2415009
```

Saving and reloading the model for prediction

```
joblib.dump(Final_model,'Housing_Price_Prediction.pkl')
```

```
['Housing_Price_Prediction.pkl']
```

```
Model = joblib.load("Housing_Price_Prediction.pkl")

a = np.array(y_test)
predicted = np.array(Model.predict(x_test))
df_train_pred = pd.DataFrame({"Original":a,"Predicted":predicted},index= range(len(a)))
df_train_pred
```

	Original	Predicted
0	100000	109015.597245
1	85400	98976.106122
2	187500	211748.803968
3	145000	163036.925662
4	168500	173266.785154
5	237500	215243.305337
6	176500	190072.967187
7	157000	151898.789048
8	113000	116895.375859
9	155000	140852.264231
10	75000	100367.343230

```
plt.figure(figsize=(8,6))
plt.scatter(y_test, predicted)
plt.yscale('log')
plt.xscale('log')
p1 = max(max(predicted), max(y_test))
p2 = min(min(predicted), min(y_test))
plt.plot([p1, p2], [p1, p2], 'b-')
plt.xlabel('Actual Price', fontsize=15)
plt.ylabel('Predicted Price', fontsize=15)
plt.axis('equal')
plt.show()
```

Prediction of sale prices of houses on the test dataset

```
Prediction = pd.DataFrame()
Prediction['SalePrice'] = Predicted_SalePrice
Prediction
```

	SalePrice
0	88381.171924
1	160058.432855
2	120212.322375
3	224234.245770
4	116523.373577
5	291520.403893
6	151281.561178
7	97341.899887
8	122512.999688
9	152371.124143
10	451607.080624

Key Metrics for success in solving problem under consideration

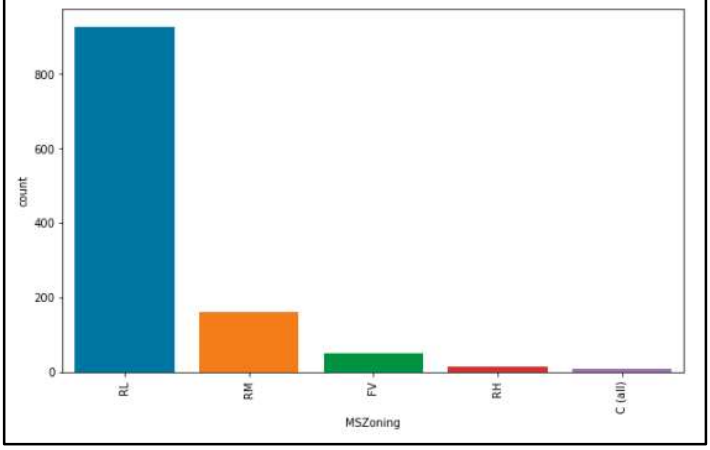
The crucial steps in any machine learning prediction model are to compute the accuracy and document the metrics on the error rates of the model. The following metrics were used:

1. **Mean Absolute Error (MAE)**: It is a popular error metric for regression problems which gives magnitude of absolute difference between actual and predicted values.
2. **Mean Squared Error (MSE)**: It is a most used and very simple metric with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference between actual and predicted value.
3. **Root Mean Squared Error (RMSE)**: RMSE is an extension of the mean squared error. The square root of the error is calculated, which means that the units of the RMSE are the same as the original units of the target value that is being predicted.
4. **R^2 Score**: It is the proportion of the variation in the dependent variable that is predictable from the independent variable.
5. **Cross Validation Score**: Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

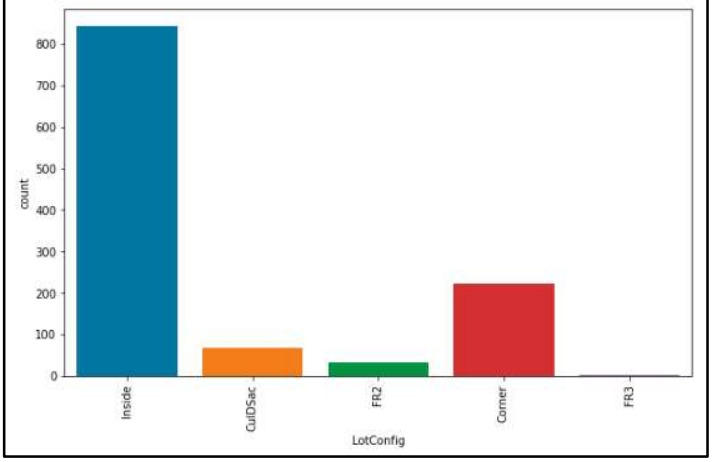
Visualizations

Univariate Analysis

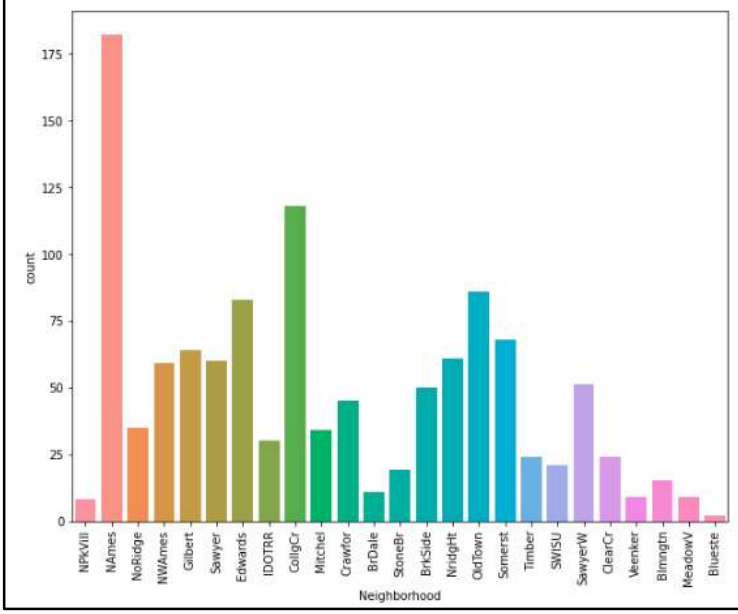
```
RL         928
RM         163
FV          52
RH          16
C (all)      9
Name: MSZoning, dtype: int64
```



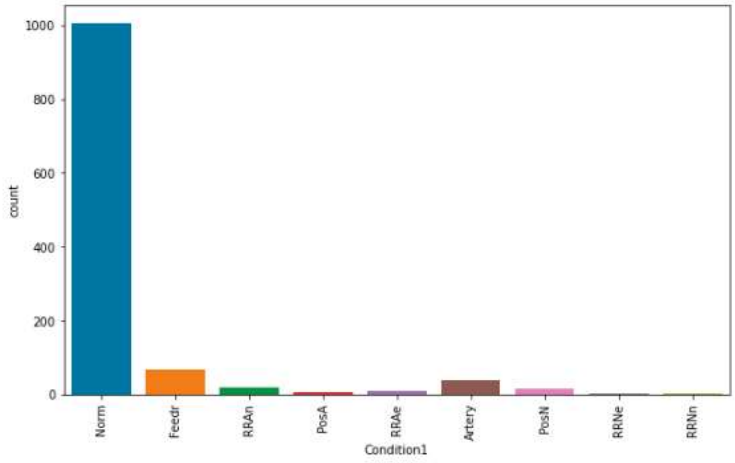

```
Inside     842
Corner     222
CulDSac     69
FR2         33
FR3          2
Name: LotConfig, dtype: int64
```



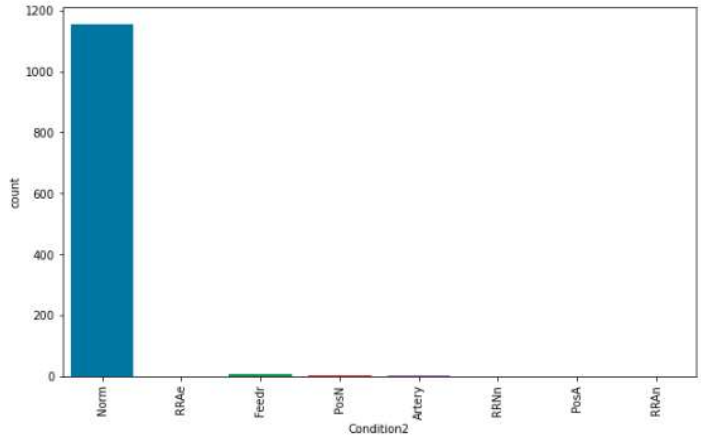
```
Name: Neighborhood, dtype: int64
```



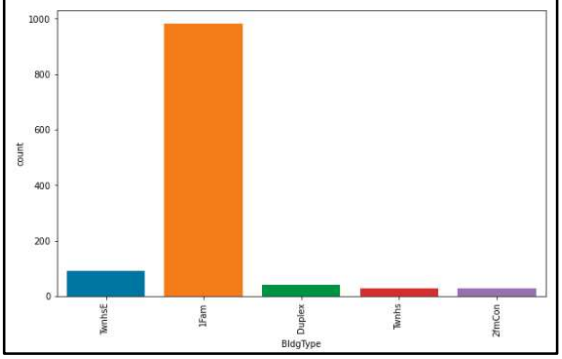

```
Norm       1005
Feedr        67
Artery       38
RRAn         20
PosN         17
RRAe          9
PosA          6
RRNn          4
RRNe          2
Name: Condition1, dtype: int64
```



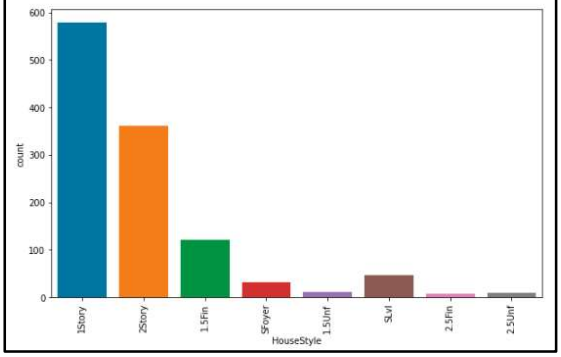
```
Norm      1154
Feedr        6
PosN         2
Artery       2
RRAn         1
RRNn         1
RRAe         1
PosA         1
Name: Condition2, dtype: int64
```



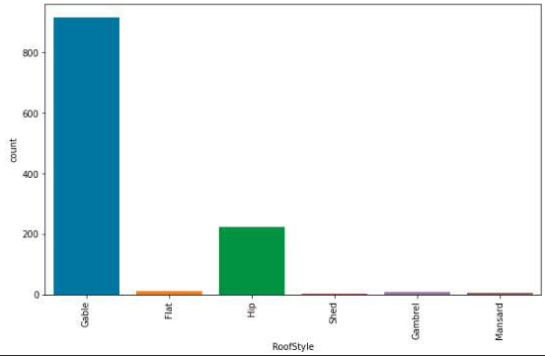

```
1Fam      981
TwnhsE     90
Duplex     41
Twnhs      29
2fmCon     27
Name: BldgType, dtype: int64
```



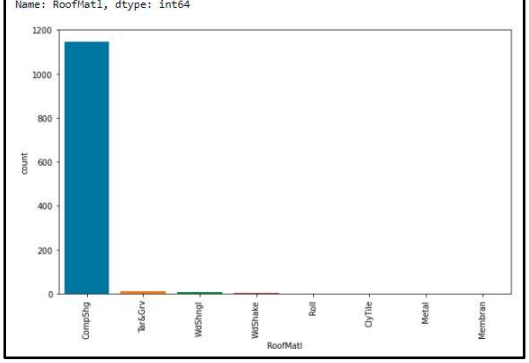

```
1Story    578
2Story    361
1.5Fin    121
SLvl       47
SFoyer     32
1.5Unf     12
2.5Unf     10
2.5Fin      7
Name: HouseStyle, dtype: int64
```



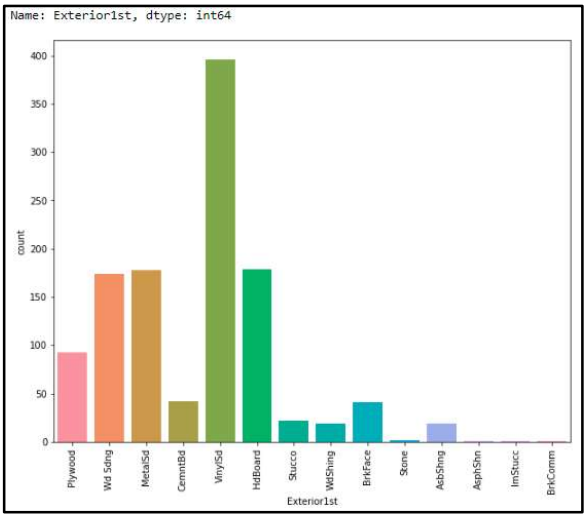

```
Gable      915
Hip        225
Flat        12
Gambrel      9
Mansard      5
Shed         2
Name: RoofStyle, dtype: int64
```



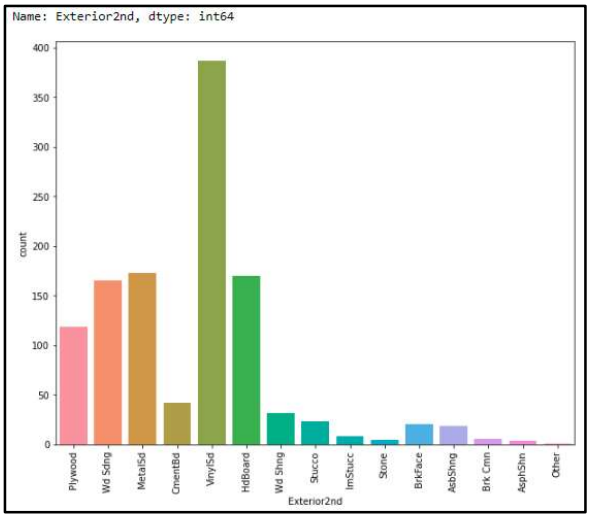

```
CompShg    1144
Tar&Grv      10
WdShngl       6
WdShake       4
Roll          1
Membran       1
Metal         1
ClyTile       1
Name: RoofMatl, dtype: int64
```



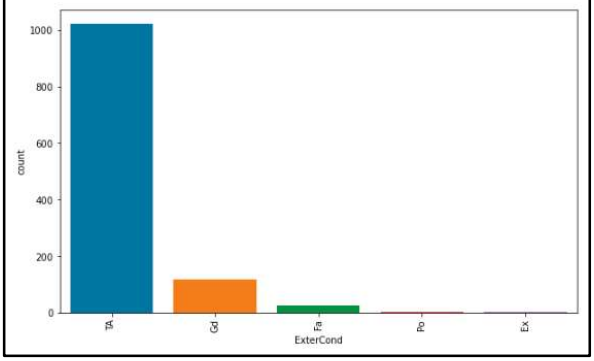
```
Name: Exterior1st, dtype: int64
```



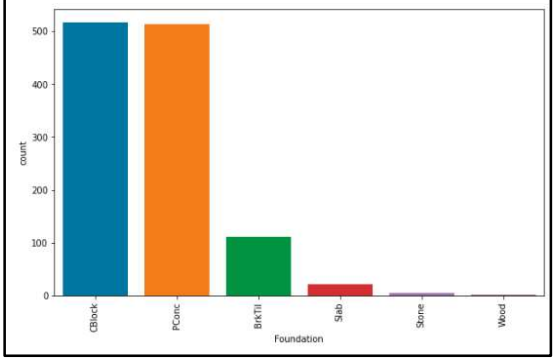

```
Name: Exterior2nd, dtype: int64
```



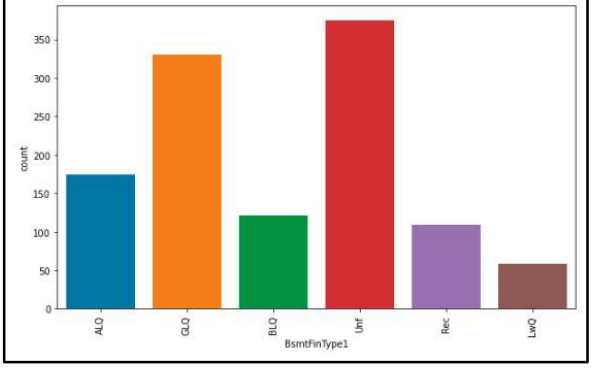

```
TA    1022
Gd     117
Fa      26
Ex       2
Po       1
Name: ExterCond, dtype: int64
```



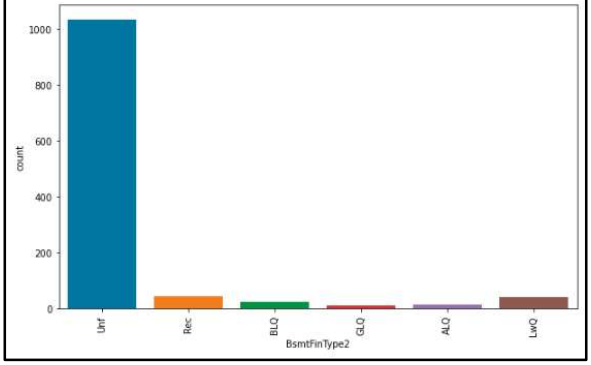

```
CBlock    516
PConc     513
BrkTil    112
Slab       21
Stone       5
Wood        1
Name: Foundation, dtype: int64
```



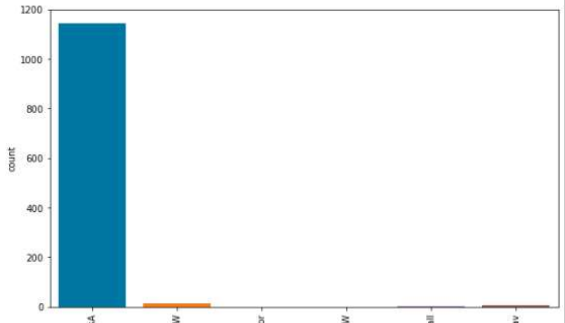

```
Unf    375
GLQ    330
ALQ    174
BLQ    121
Rec    109
LwQ     59
Name: BsmtFinType1, dtype: int64
```



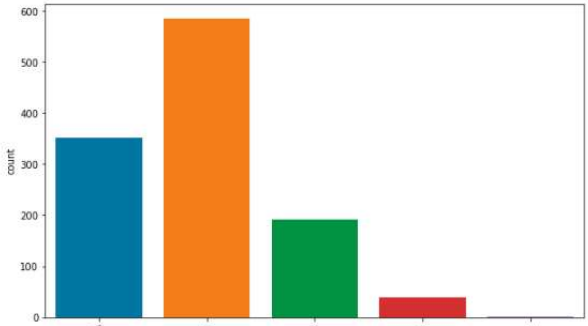

```
Unf    1033
Rec      43
LwQ      40
BLQ      24
ALQ      16
GLQ      12
Name: BsmtFinType2, dtype: int64
```



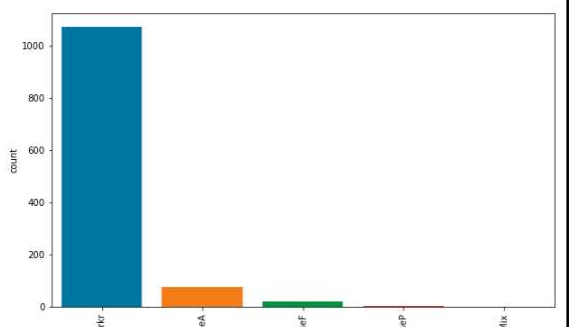
```
GasA    1143
GasW      14
Grav       5
Wall       4
OthW       1
Floor      1
Name: Heating, dtype: int64
```



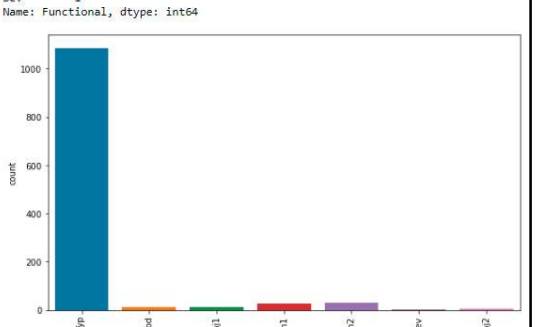

```
Ex    585
TA    352
Gd    192
Fa     38
Po      1
Name: HeatingQC, dtype: int64
```



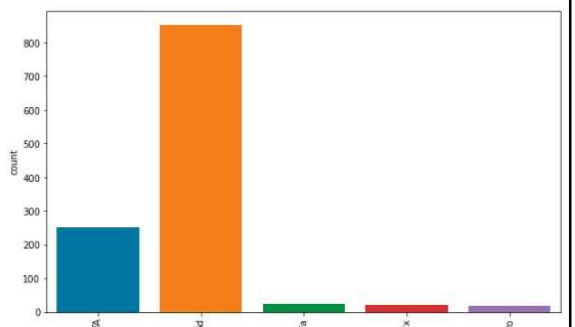

```
SBrkr    1070
FuseA      74
FuseF      21
FuseP       2
Mix         1
Name: Electrical, dtype: int64
```



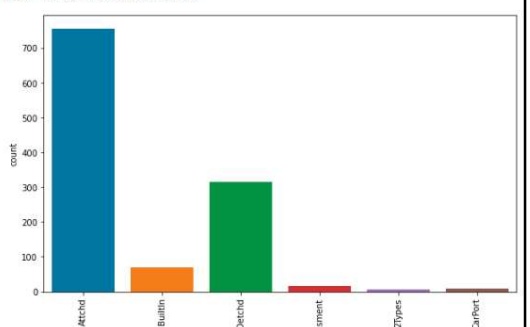

```
Typ     1085
Min2      30
Min1      25
Mod       12
Maj1      11
Maj2       4
Sev        1
Name: Functional, dtype: int64
```

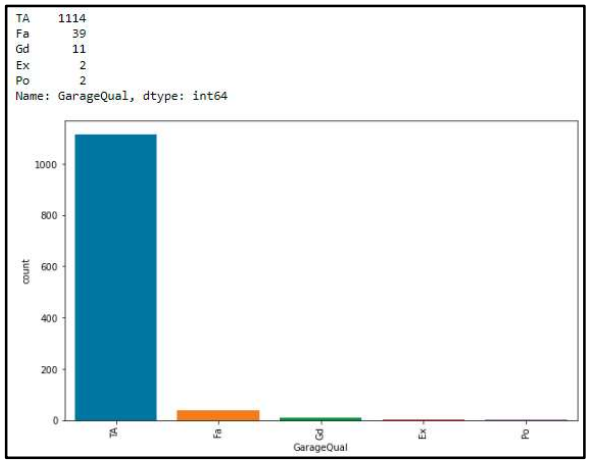


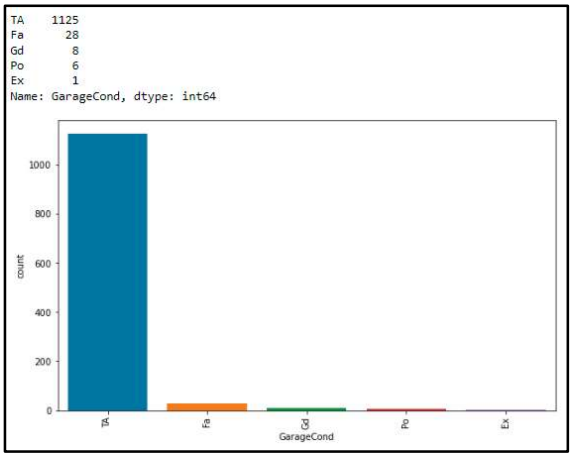

```
Gd    852
TA    252
Fa     25
Ex     21
Po     18
Name: FireplaceQu, dtype: int64
```

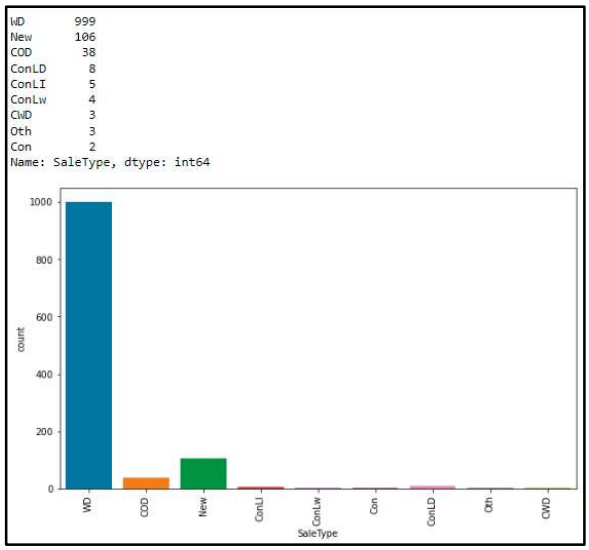


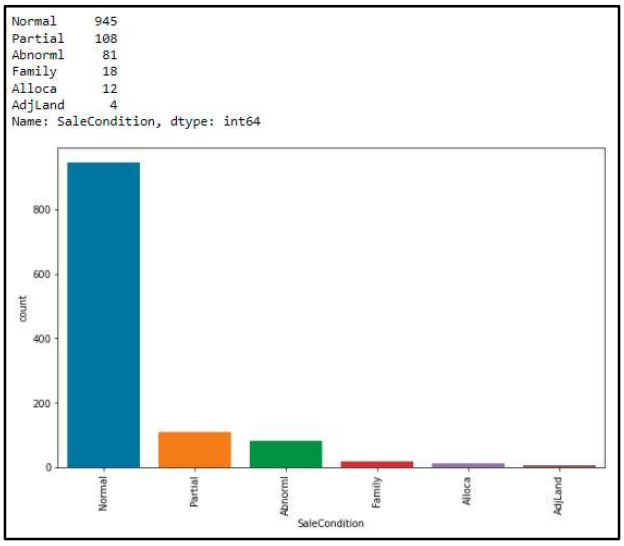

```
Attchd     755
Detchd     314
BuiltIn     70
Basment     16
CarPort      8
2Types       5
Name: GarageType, dtype: int64
```









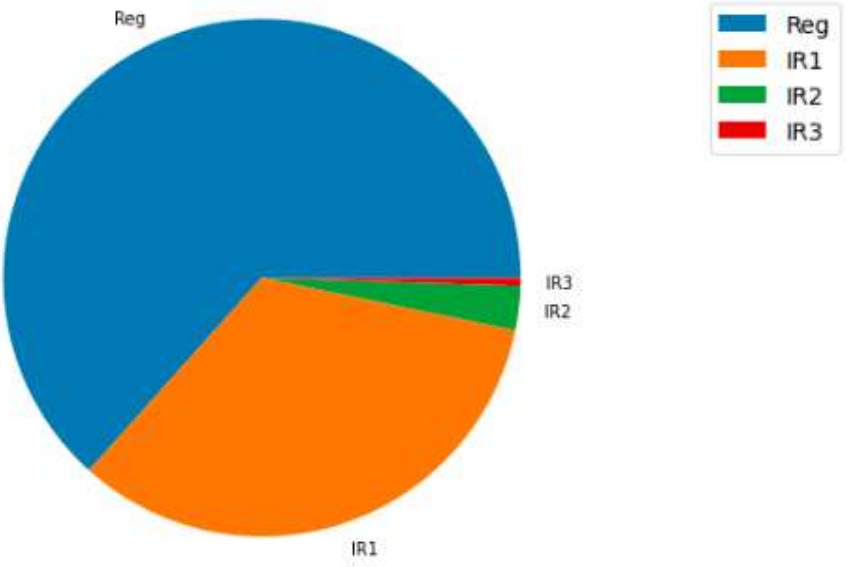


Observations:

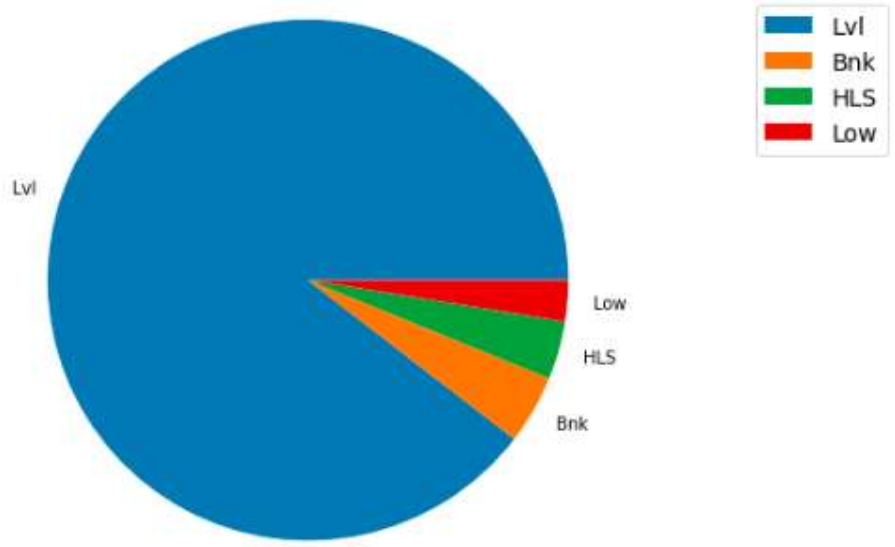
- **MSZoning**: In the general zoning classification of the sale, Residential Low Density has the highest numbers, followed by Residential Medium Density, with Commercial classification having the lowest numbers of houses sold.
- **LotConfig**: Inside lot configuration has the highest numbers, followed by corner lots and Cul-de-sacs. Houses with Frontage on 3 sides of property have the lowest numbers.
- **Neighborhood**: North Ames has the highest numbers, followed by College Creek, Old Town, and Edwards. Northpark Villa, Briardale, Veenker, Meadow Village and Bluestem have the lowest numbers in houses sold.
- **Condition1**: In terms of proximity to various conditions, Normal has the highest numbers, with all other conditions having very low numbers.
- **Condition2**: Same as in Condition1
- **BldgType**: Single-family Detached type of dwelling has the highest numbers in houses sold, with all other types having comparatively lower numbers.
- **HouseStyle**: One story styled dwelling has the highest numbers, followed by Two story, One and one-half story: 2nd level finished, and Split Level. All other styles of dwelling have lower numbers.
- **RoofStyle**: Gable type roof have the highest numbers, followed by comparatively low numbers of Hip style roofs, with very numbers of other styled roofs.
- **RoofMatl**: Roof of most houses have Standard (Composite) Shingle material
- **Exterior1st**: In terms of Exterior covering on house, Vinyl Siding is the highest in number, followed by Hard Board, Metal Siding and Wood Siding.
- **ExterCond**: The present condition of the material on the exterior is Typical in most cases, good and fair in the remaining, with some exceptions of Excellent and Poor.

- **Foundation**: Cinder Block and Poured Concrete have almost equal numbers, followed by Brick & Tile type of foundation.
- **BsmtFinType1**: The basement finished area of most houses are rated Unfinished or Good Living Quarters, followed by Average Living Quarters, Below Average Living Quarters, Average Rec Room, Low Quality respectively.
- **Heating**: Most houses have Gas forced warm air furnace
- **HeatingQC**: In terms of Heating quality and condition, most houses have a rating of Excellent, followed by either typical/average, or Good.
- **Electrical**: Most homes have Standard Circuit Breakers & Romex in terms of electrical systems installed
- **Functional**: In terms of home functionality, most homes have a Typical Functionality, with few numbers having minor to moderate deductions, 15 homes with major deductions, and 1 with severe damage.
- **FireplaceQu**: Most homes have Good or Average Fireplace Quality
- **GarageType**: Garages in most homes are attached, followed by houses having detached garages
- **SaleType**: The sale of most houses was through a conventional warranty deed, followed by homes that were recently constructed and sold
- **SaleCondition**: Most homes had normal conditions of sale, followed by comparatively low numbers of homes which were not completed when last assessed, and abnormal sales through trade, foreclosure, or short sale.

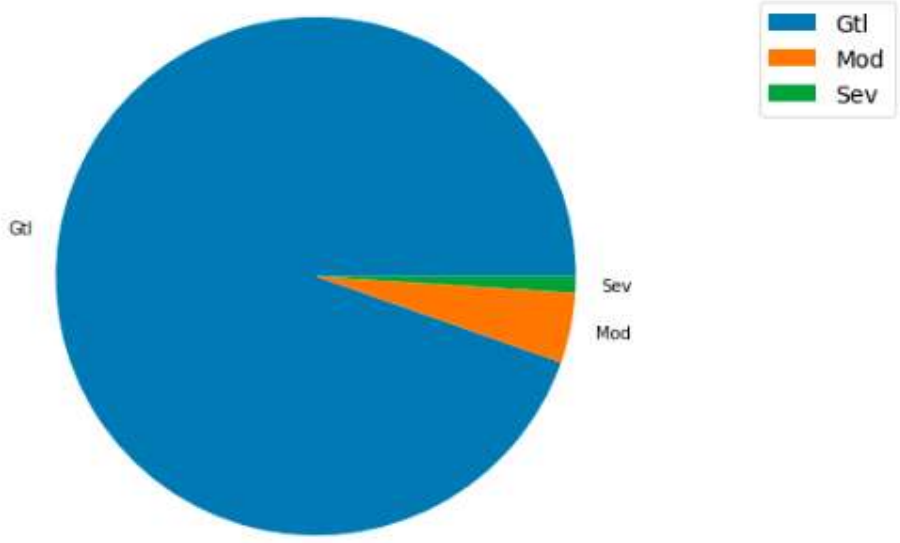
```
Pie plot for the column: LotShape
Reg    740
IR1    390
IR2     32
IR3      6
Name: LotShape, dtype: int64
```



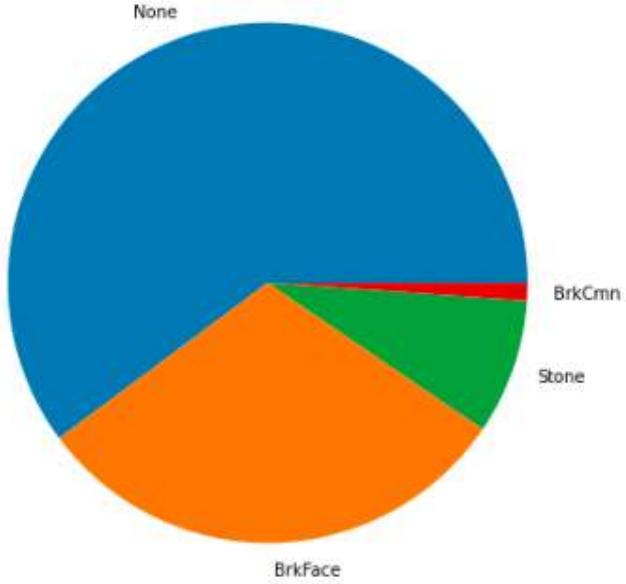
```
Pie plot for the column: LandContour
Lvl    1046
Bnk      50
HLS      42
Low      30
Name: LandContour, dtype: int64
```



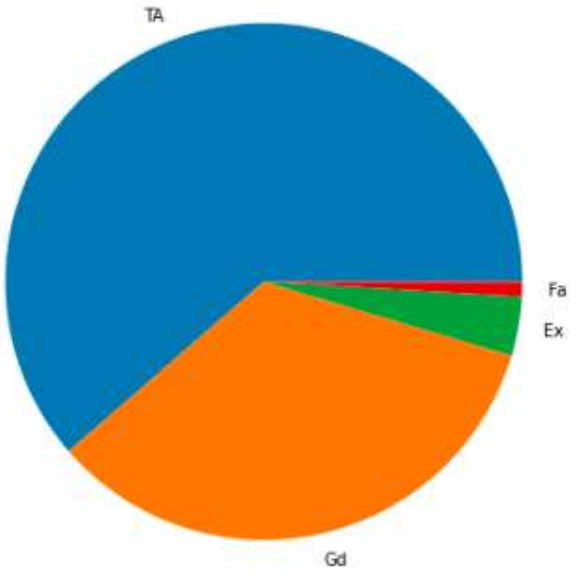

```
Pie plot for the column: LandSlope
Gtl    1105
Mod      51
Sev      12
Name: LandSlope, dtype: int64
```



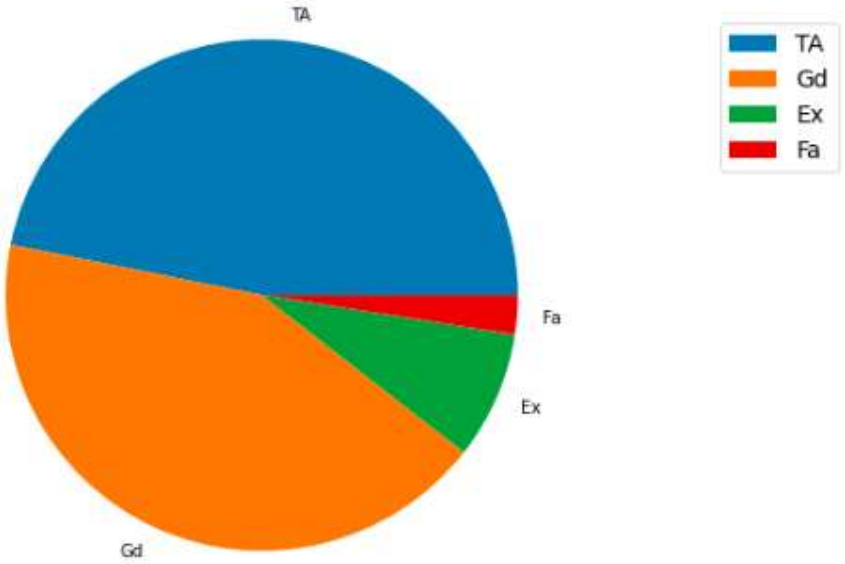
```
Pie plot for the column: MasVnrType
None       703
BrkFace    354
Stone       98
BrkCmn      13
Name: MasVnrType, dtype: int64
```



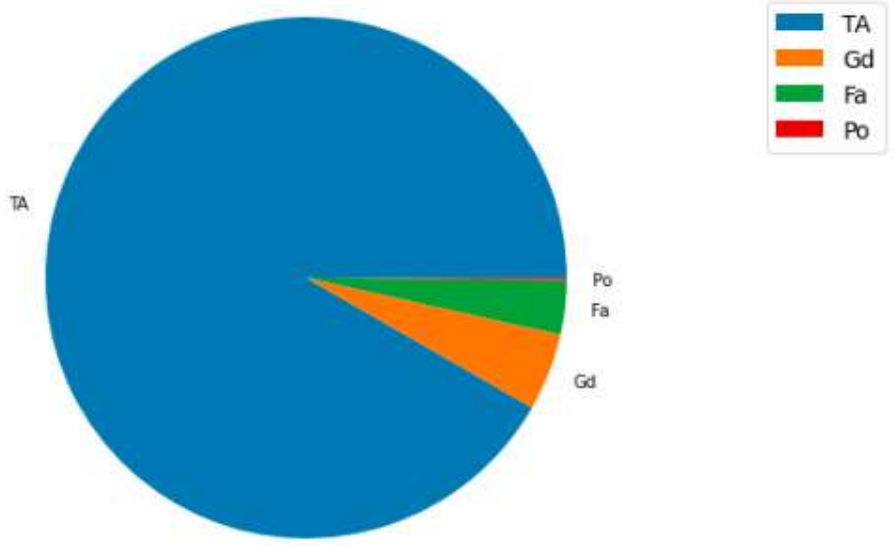

```
Pie plot for the column: ExterQual
TA    717
Gd    397
Ex     43
Fa     11
Name: ExterQual, dtype: int64
```



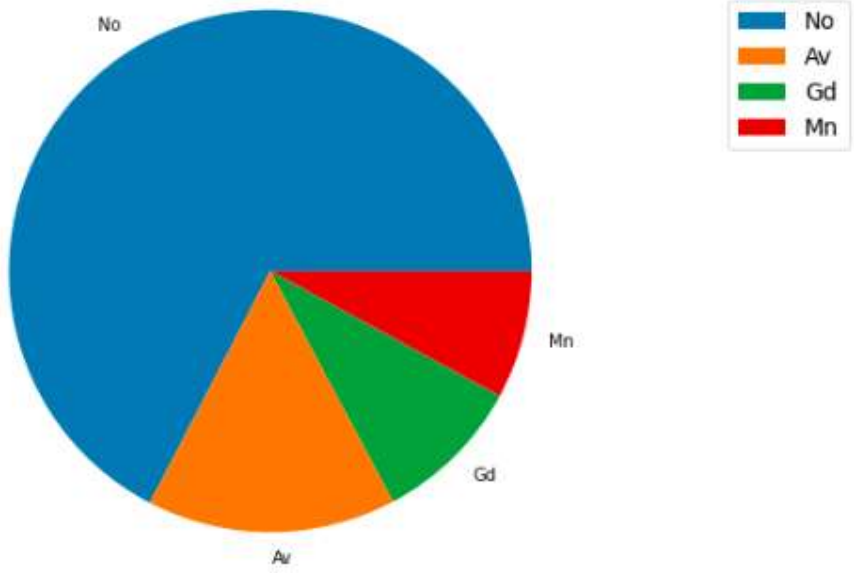
```
Pie plot for the column: BsmtQual
TA    547
Gd    498
Ex     94
Fa     29
Name: BsmtQual, dtype: int64
```



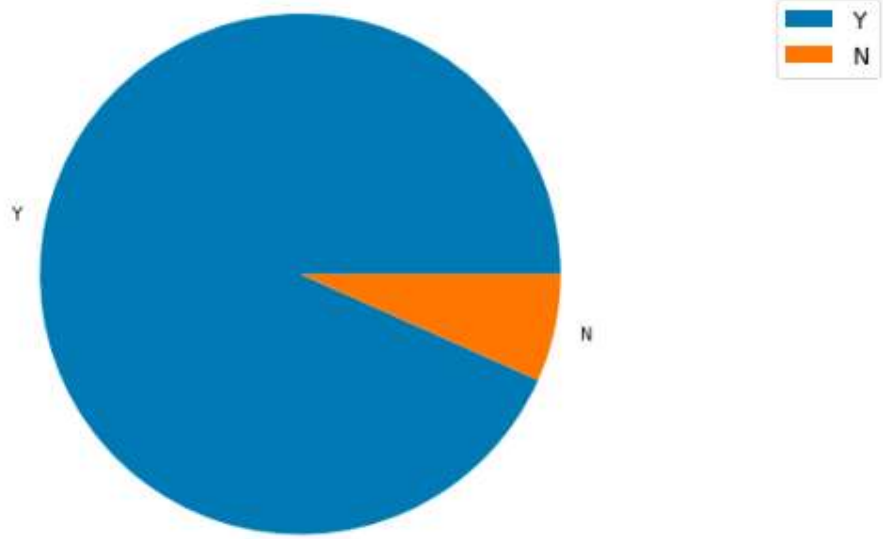

```
Pie plot for the column: BsmtCond
TA    1071
Gd      56
Fa      39
Po       2
Name: BsmtCond, dtype: int64
```



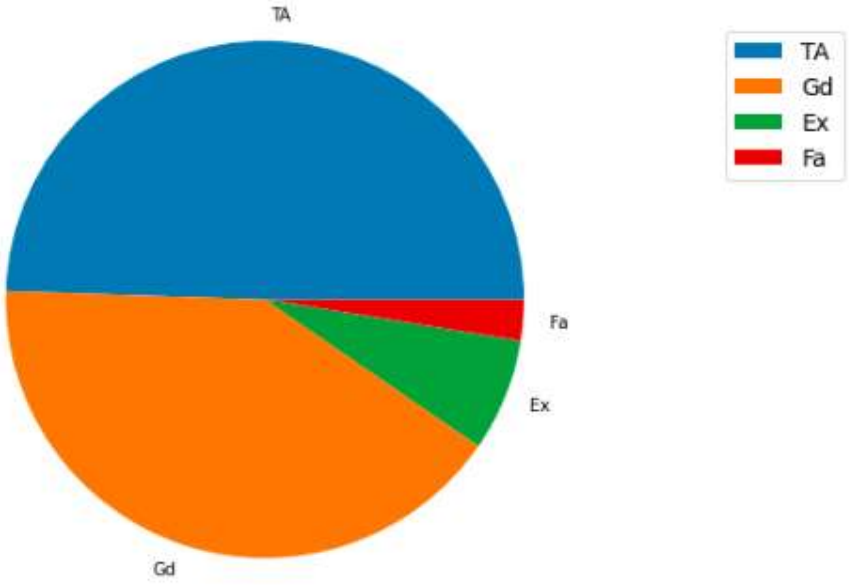
```
Pie plot for the column: BsmtExposure
No    787
Av    180
Gd    108
Mn     93
Name: BsmtExposure, dtype: int64
```



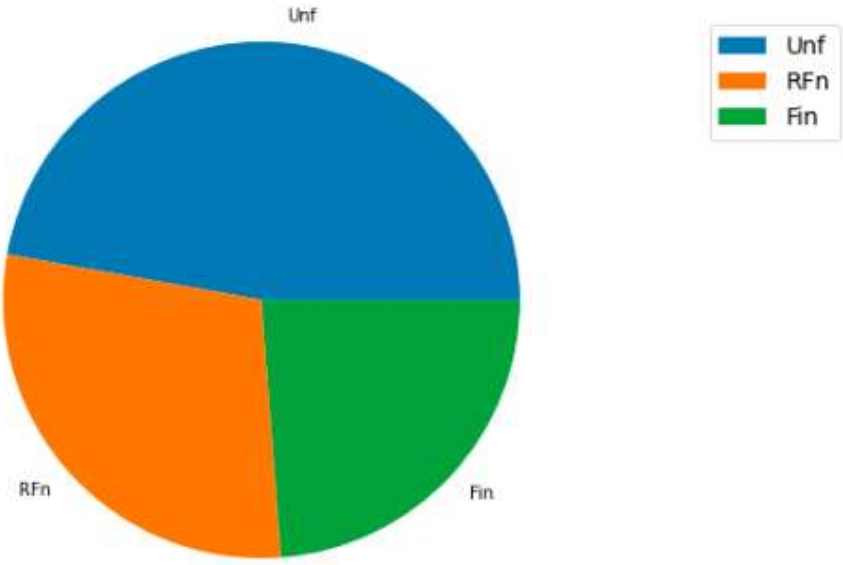

```
Pie plot for the column: CentralAir
Y    1090
N      78
Name: CentralAir, dtype: int64
```



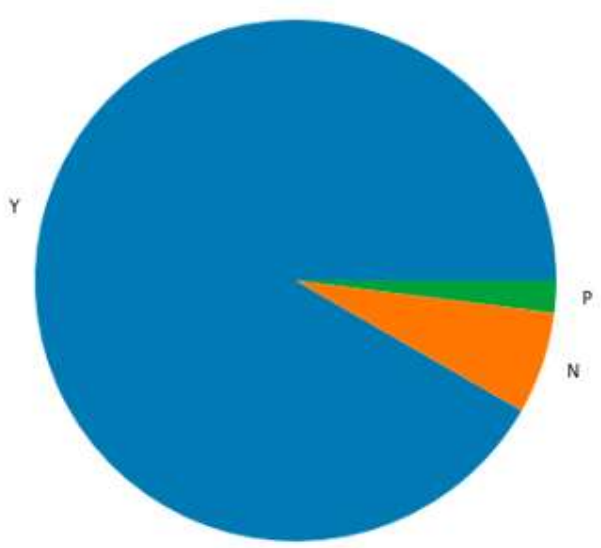
```
Pie plot for the column: KitchenQual
TA    578
Gd    478
Ex     82
Fa     30
Name: KitchenQual, dtype: int64
```




```
Pie plot for the column: GarageFinish
Unf    551
RFn    339
Fin    278
Name: GarageFinish, dtype: int64
```



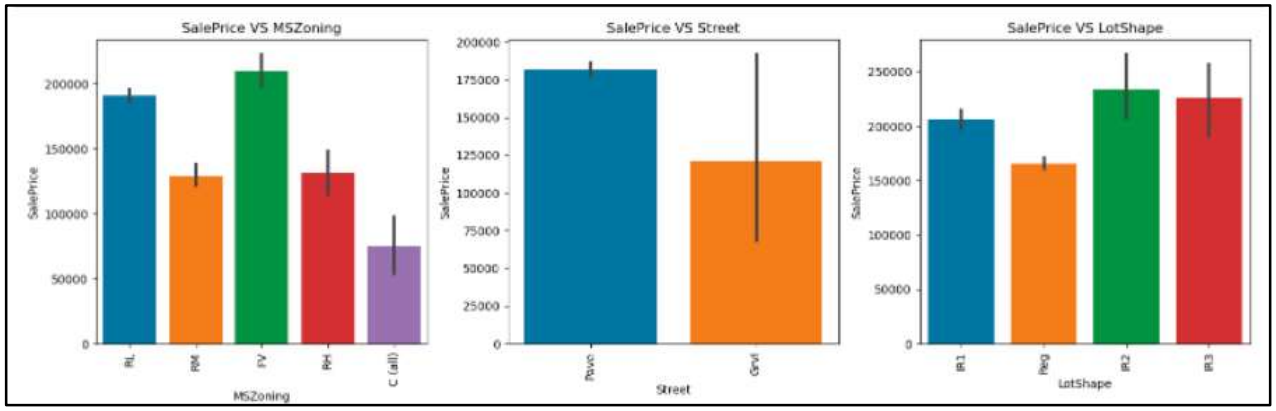
```
Pie plot for the column: PavedDrive
Y    1071
N      74
P      23
Name: PavedDrive, dtype: int64
```

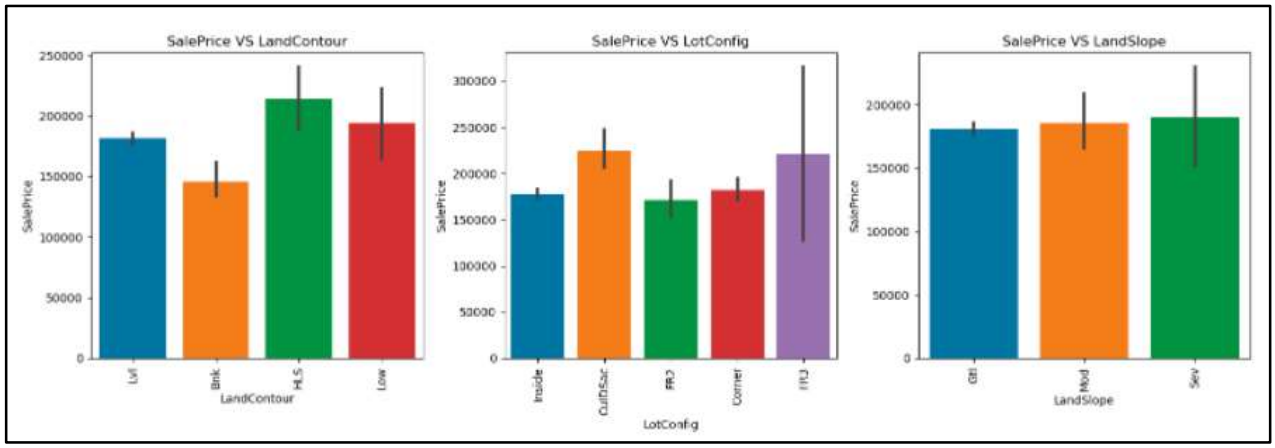


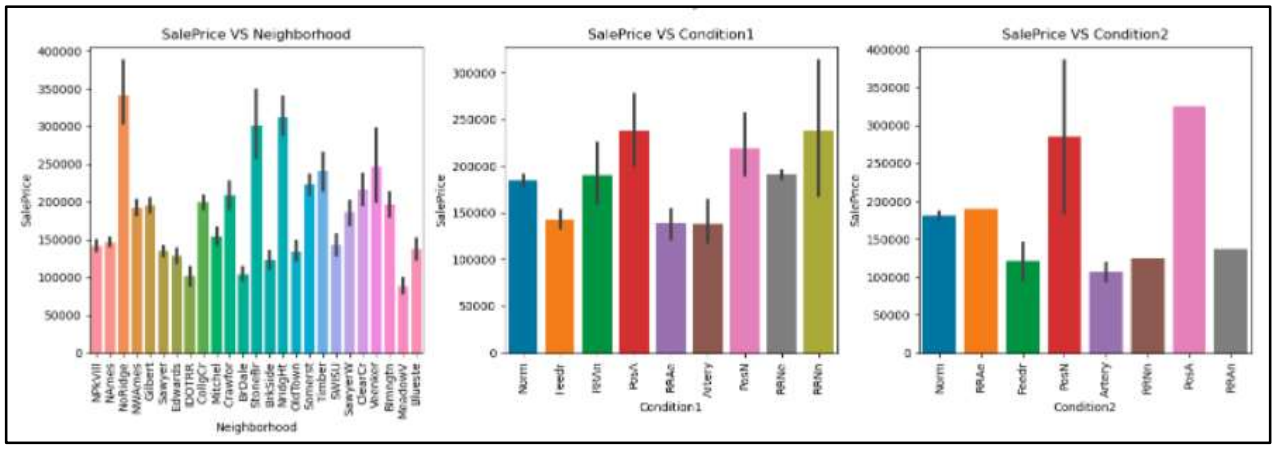
Observations:

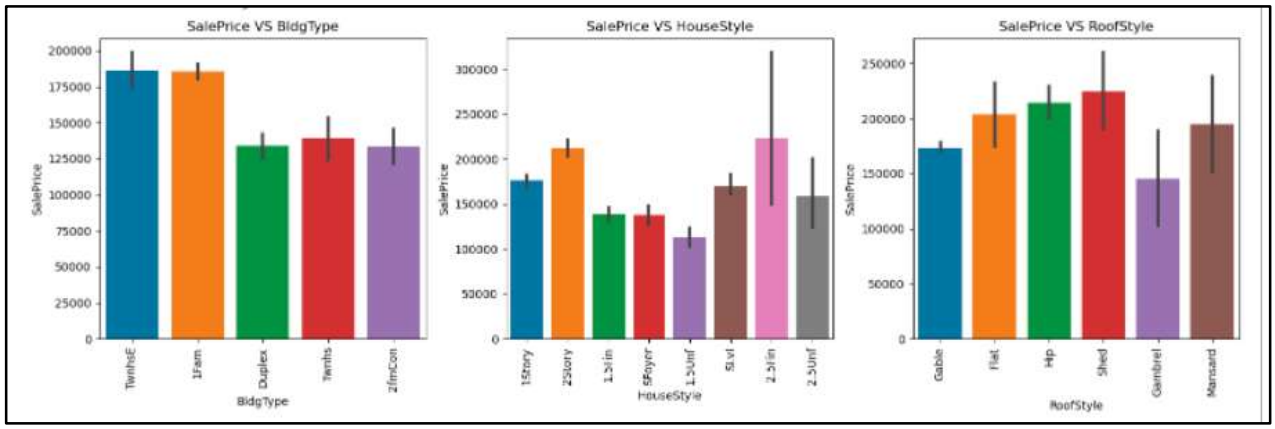
- **Street**: Most houses have a paved road access to the property
- **LotShape**: Most properties have a regular shape, followed by those that are slightly irregular. There are few properties that are moderately or completely irregular
- **LandContour**: Most properties have Near Flat or Levelled land contour
- **MasVnrType**: Although most homes have no masonry veneers, there are quite a few that have a brick face veneer, followed by stone or brick commons.
- **BsmtExposure**: Most homes have no exposure to walkout or garden level walls, followed by those with average exposure, good exposure, and minimum exposure.
- **CentralAir**: Almost all homes have central air conditioning
- **KitchenQual**: Most homes in the dataset have either an average/typical or good kitchen quality
- **GarageFinish**: The interior finish of garages in most homes is unfinished, followed by those with a rough finish, or are finished.

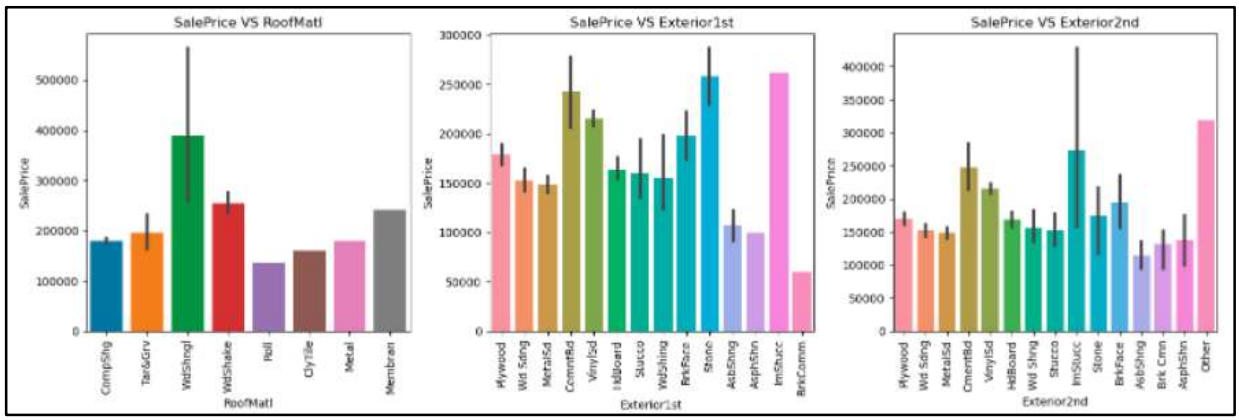
Bivariate Analysis




SalePrice VS LandContour
SalePrice
LandContour
Lvl
Bnk
HLS
Low
SalePrice VS LotConfig
SalePrice
LotConfig
Inside
CulDSac
FR2
Corner
FR3
SalePrice VS LandSlope
SalePrice
LandSlope
Gtl
Mod
Sev

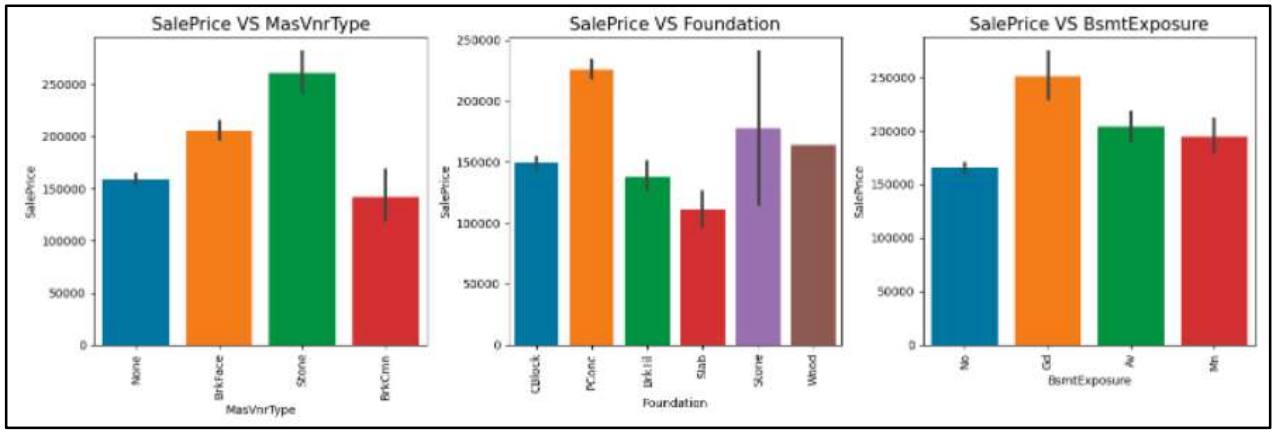

SalePrice VS Neighborhood
SalePrice
Neighborhood
NPkVill
Names
NoRidge
NAmes
Gilbert
Sawyer
Edwards
BrkSide
OldTown
Somerst
Timber
SWISU
SawyerW
ClearCr
Veenker
Blmngtn
Meadowv
Blueste
SalePrice VS Condition1
SalePrice
Condition1
Norm
Feedr
RRAn
PosA
RRAe
Artery
PosN
RRNe
RRNn
SalePrice VS Condition2
SalePrice
Condition2
Norm
RRAe
Feedr
PosN
Artery
RRAn
PosA
RRNn

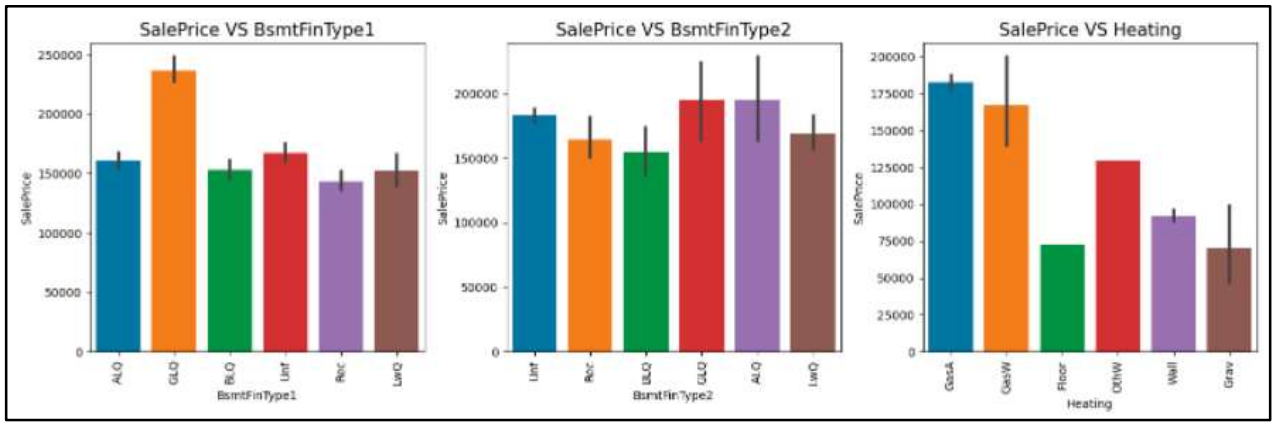

SalePrice VS BldgType
SalePrice
BldgType
TwnhsE
1Fam
Duplex
Twnhs
2fmCon
SalePrice VS HouseStyle
SalePrice
HouseStyle
1Story
2Story
1.5Fin
SFoyer
1.5Unf
SLvl
2.5Fin
2.5Unf
SalePrice VS RoofStyle
SalePrice
RoofStyle
Gable
Flat
Hip
Shed
Gambrel
Mansard

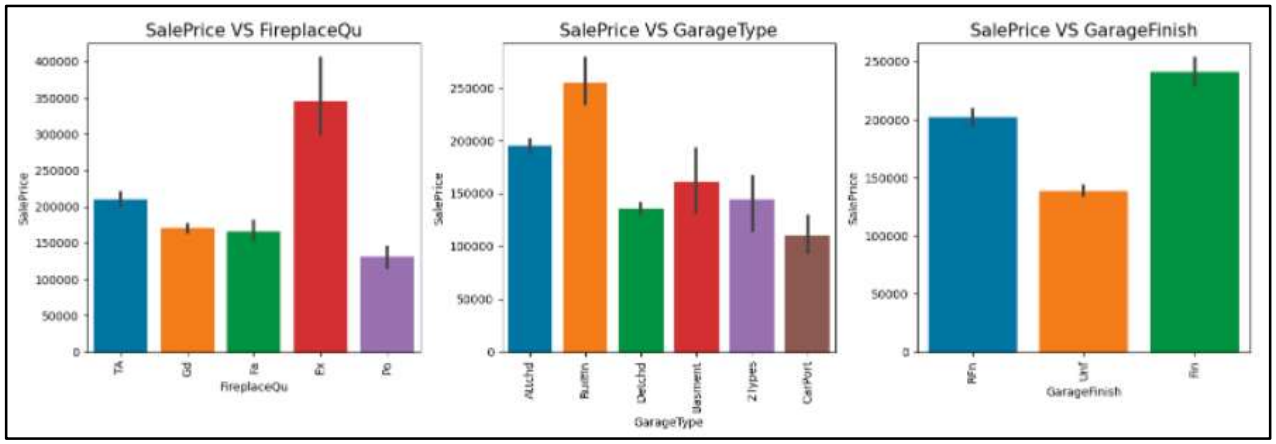

SalePrice VS RoofMatl
SalePrice
RoofMatl
CompShg
Tar&Grv
WdShngl
WdShake
Roll
ClyTile
Metal
Membran
SalePrice VS Exterior1st
SalePrice
Exterior1st
AsbShng
AsphShn
BrkComm
BrkFace
CemntBd
HdBoard
ImStucc
MetalSd
Plywood
Stone
Stucco
VinylSd
Wd Sdng
WdShing
SalePrice VS Exterior2nd
SalePrice
Exterior2nd
Brk Cmn
Other

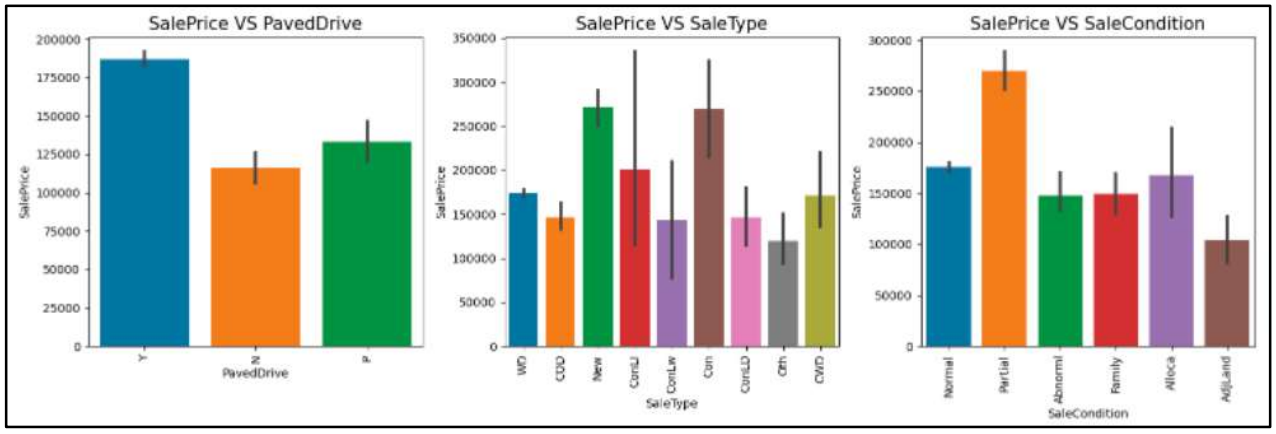
Observations:

- Floating Village Residential have the highest sale prices, followed by Residential Low Density. Residential Medium Density and Residential High Density are similar in terms of highest prices, finally followed by Commercial properties.
- Properties with Paved Road access to property have higher sale prices
- Sale prices for regularly shaped properties are lower as compared to irregularly shaped properties.
- Hillside properties are sold at higher prices, compared to banked, depressed or near flat properties
- Properties with frontage on 3 sides have highest prices, with Cul-de-sacs having higher sale prices in general.
- Properties have similar sale pricing irrespective of slope of the property
- Houses in the neighbourhoods of Northridge, Stone Brook and Northridge Heights have higher sale prices, while the neighbourhoods of Iowa DOT and Railroad, Briardale and Meadow Village have sale prices on the lower end of the price spectrum.
- Houses within 200' of North-South Railroad and those Adjacent to positive off-site feature have higher sale prices, combined with secondary conditions of Near positive off-site feature--park, greenbelt, etc., or Adjacent to positive off-site feature also have higher sale prices.
- Townhouse End Units and Single-family Detached dwellings have similar and high sale prices, followed by similar pricing for Townhouse Inside Units, Duplex and Two-family Conversion units.
- Two and one-half story: 2nd level finished has the highest sale pricing, followed by Two story houses.
- The houses having the roof style Flat, Hip and Shed have high sale price and the houses having gravel roof style have less sale price.
- Houses with Wood Shingles root materials have high sale prices.
- Houses having Imitation Stucco, Stone and Cement Board as 1st exterior cover have high sale price.
- Houses having Imitation Stucco and other as 2nd cover have high sale price.





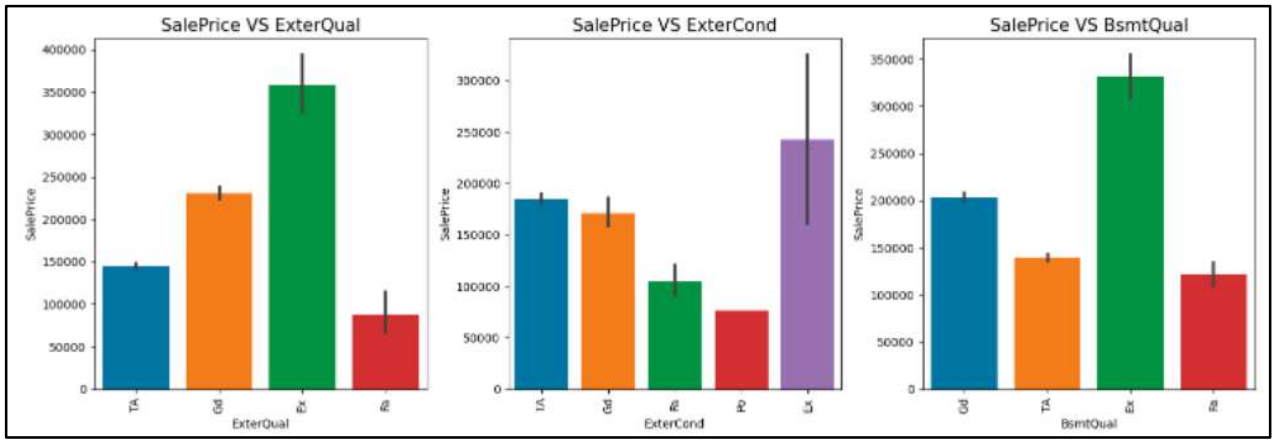


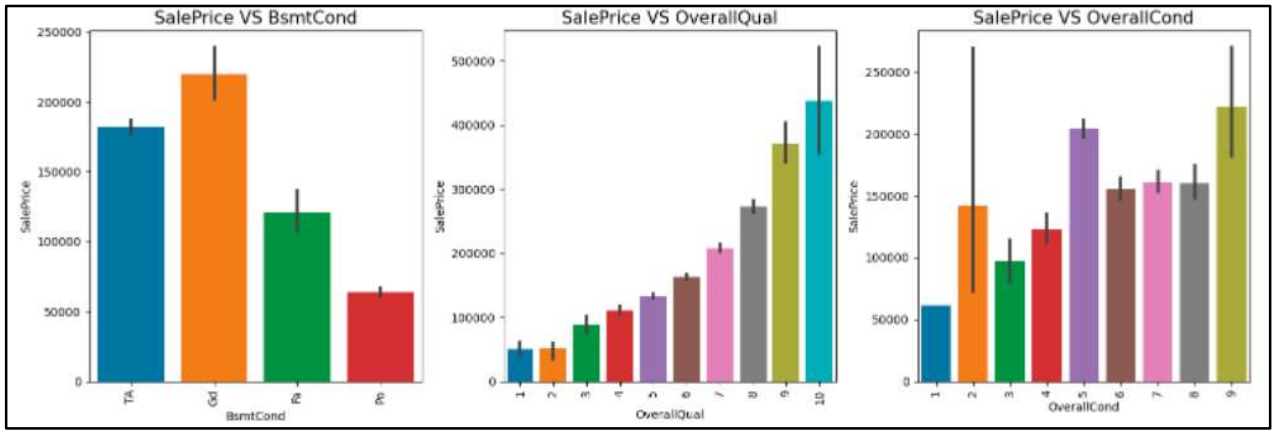


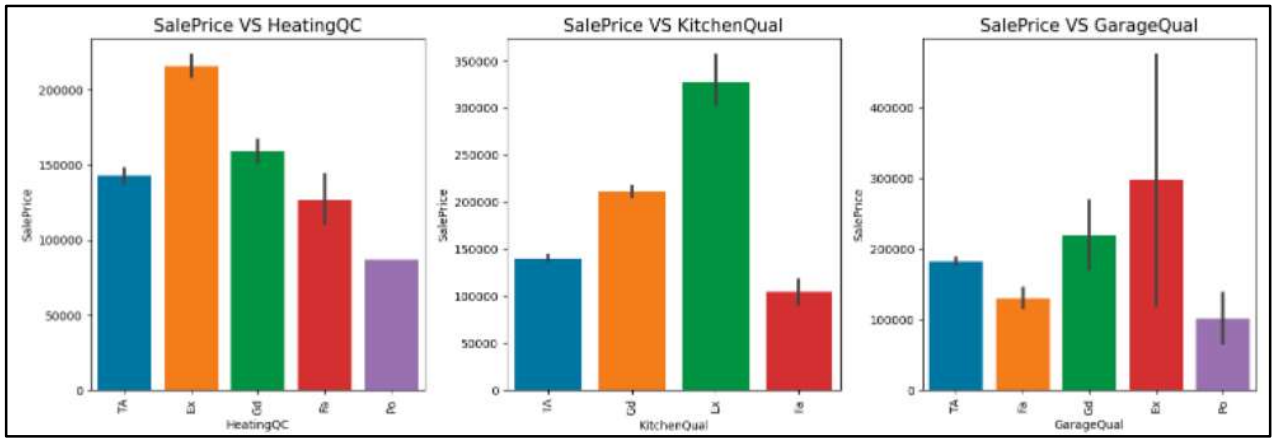
Observations:

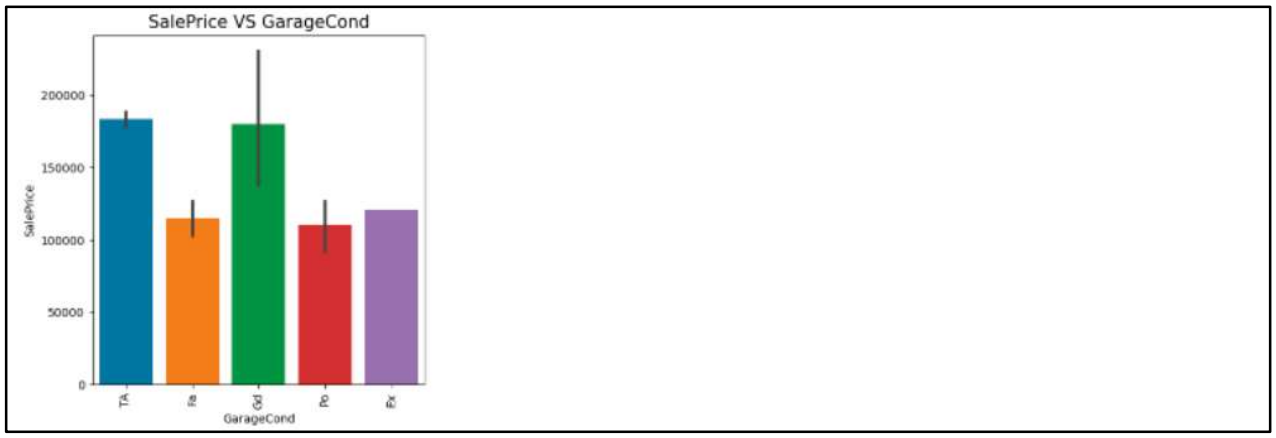
- Homes with Stone masonry veneers have highest prices, followed by those with Brick Face. Houses with no masonry veneers and those with brick commons have similar sale prices.
- Poured concrete foundation houses have higher sale prices, followed by houses having Stone foundations, which in some instances have been sold at prices almost as high as houses with poured concrete foundations. This is followed by houses with wood, cinder block, brick and tile, and slab foundations.
- Properties with good exposure to walkout or garden level walls have the highest sale prices.
- Homes with Good Living Quarters finishing rating for basements have the highest prices, followed by all other categories, which are at almost similar sale prices.
- Sale prices are the highest for homes with Gas forced warm air furnace and Gas hot water or steam heat heating systems.
- Centrally air-conditioned homes have higher sale prices as compared to those that do not.

- In terms of electrical systems, homes with Standard Circuit Breakers & Romex have the highest sale prices, while those with mixed electrical systems are priced the lowest in comparison.
- Sale prices are the highest for homes with typical functionality, followed by a variable pricing for other categories of home functionality.
- Sale prices are by far the highest for homes with excellent fireplace quality
- Houses with built-in garages are priced the highest, followed by those with attached garages, and basement garages.
- From the plot, it is clear as expected that homes with finished garages have the highest prices, followed by roughly finished and finally with unfinished garages.
- Homes with a paved driveway have the highest sale prices, as expected, followed by those with partial pavements.
- Properties sold with contracts with low interest have witnessed highest sale prices in on off instances, however, in general, homes sold with Contract 15% Down payment regular terms, and Homes just constructed and sold have the highest sale prices.
- Homes that were not completed when last assessed (i.e., primarily new homes) saw the highest sale prices.



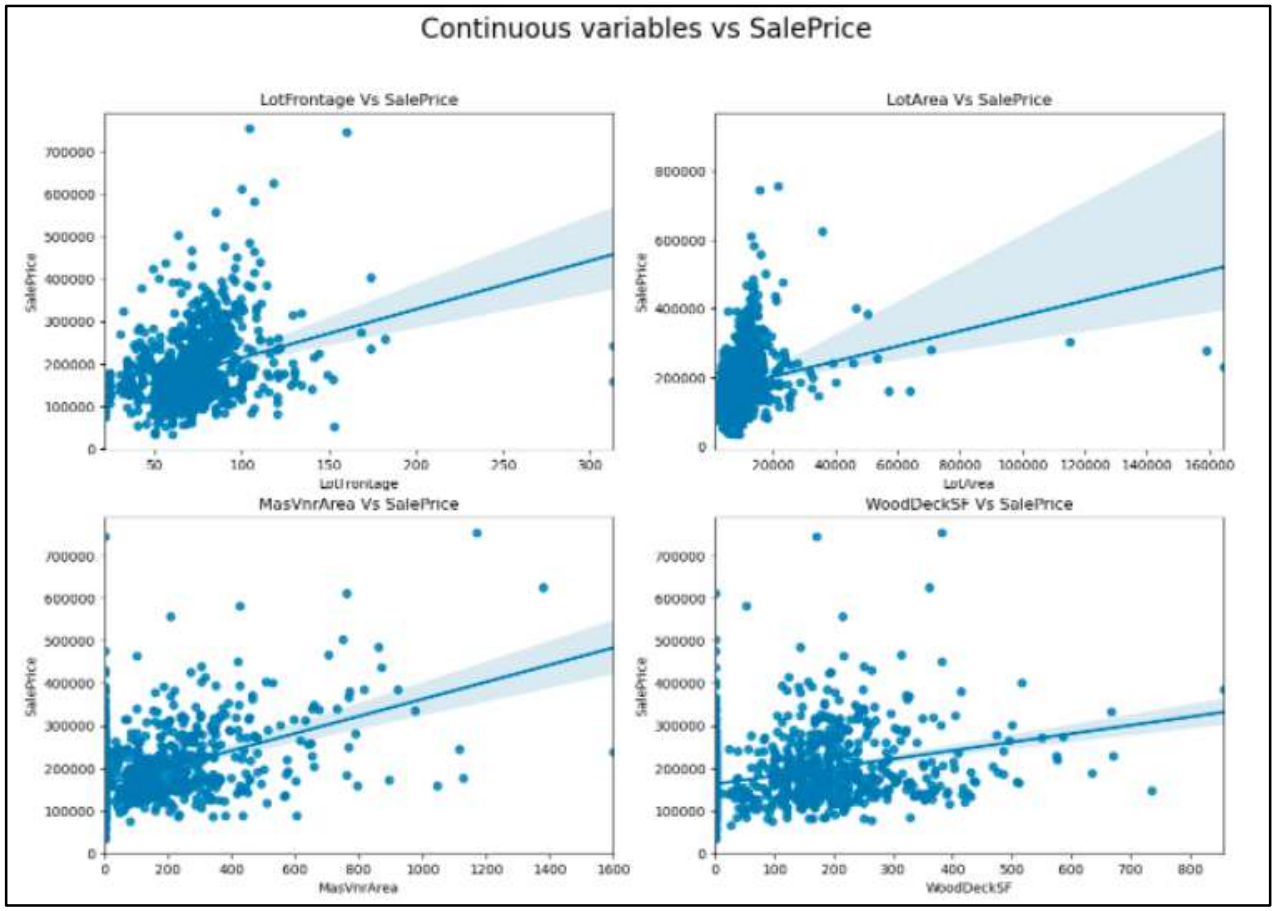




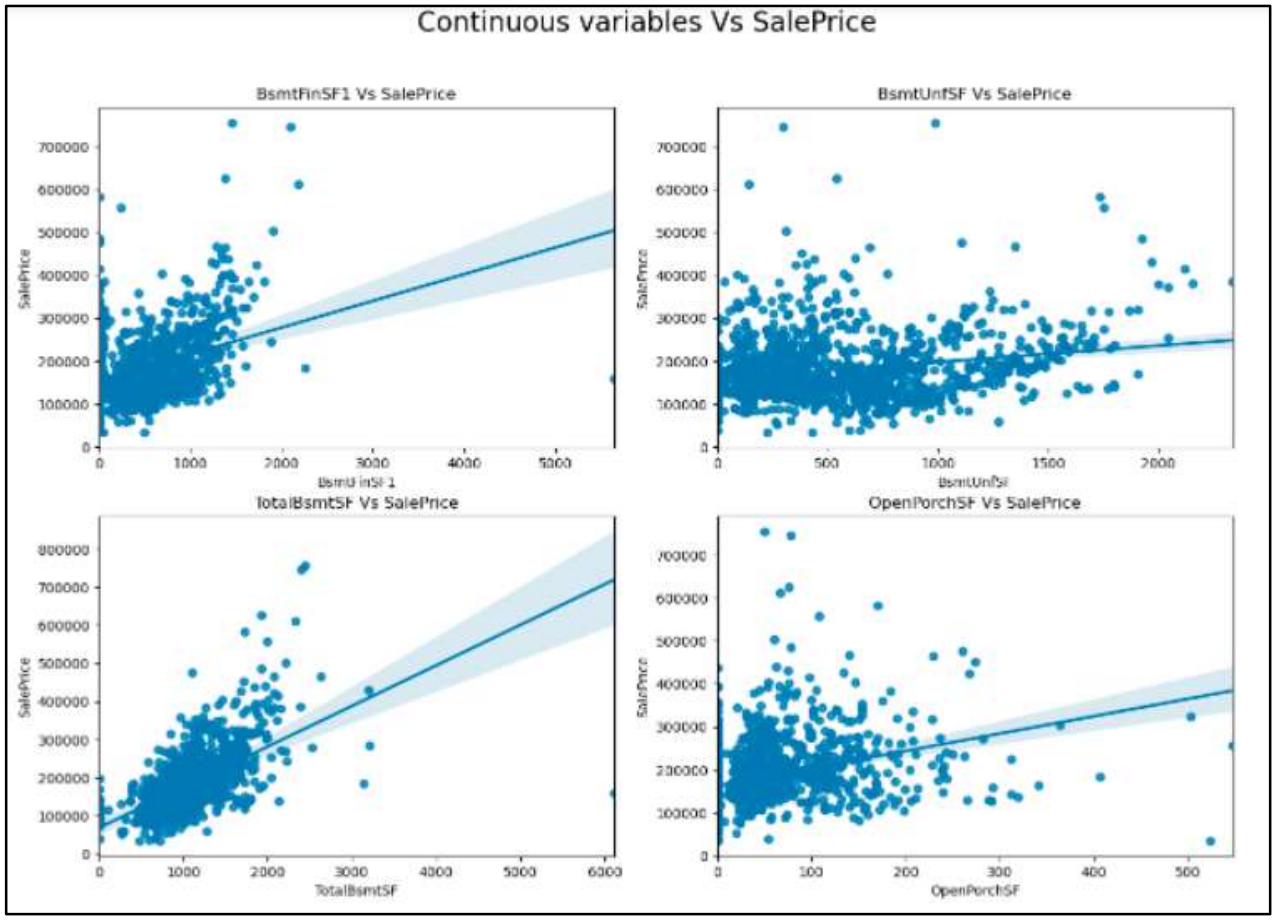


Observations:

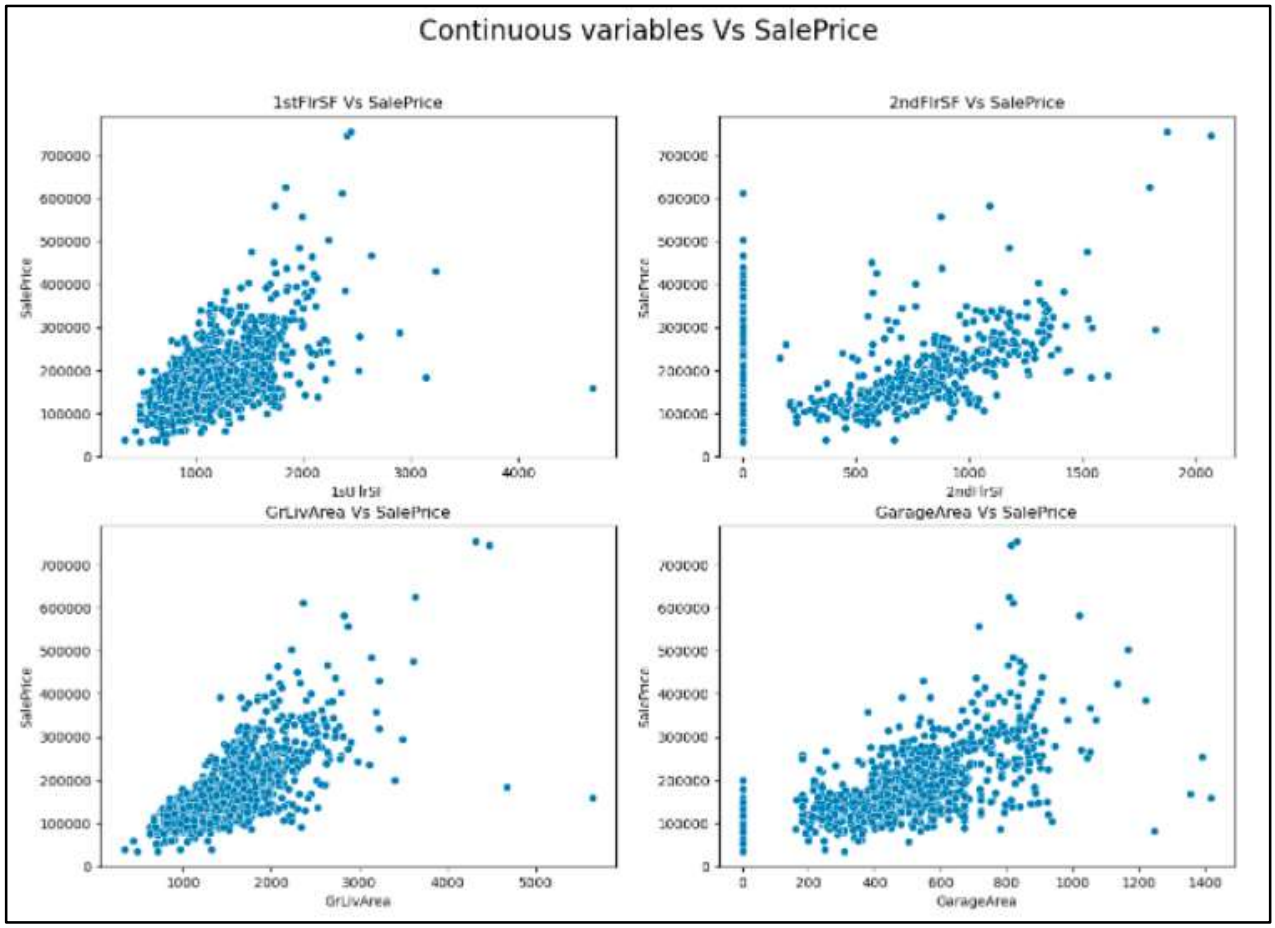
- Sale pricing of houses increases with an increase in the exterior quality rating of the house, with excellent rated houses fetching the highest prices, followed by those rated as Good, Average and Poor.
- Exterior condition of the house records a similar trend as in the case with exterior quality of the house.
- In terms of basement quality, i.e., the height of the basement, those rated excellent (100+ inches) fetch highest sale prices in the housing market.
- General condition of the basement is directly related to the sale price of the house, with houses having better condition basements fetching prices in the market.
- Overall quality of the house experiences a linear relationship with the sale price of the house, with price increasing with each higher rating of the house.
- Ratings on the overall condition of the house does not experience a linear relationship with the sale pricing of the house. The highest pricing going for those rated 9 (no data for houses rated 10), followed by those rated 5, and some instances of those rated 2, but 8,7, and 6 in general.
- An excellent heating quality and condition rating fetches highest sale prices, followed by Good, typical, fair, and poor rated houses respectively.
- Garage Condition rating of good sees highest sale prices in the market, followed by typical/average rated garages of houses, with Excellent, Fair and Poor rated ones experiencing similar pricing levels.



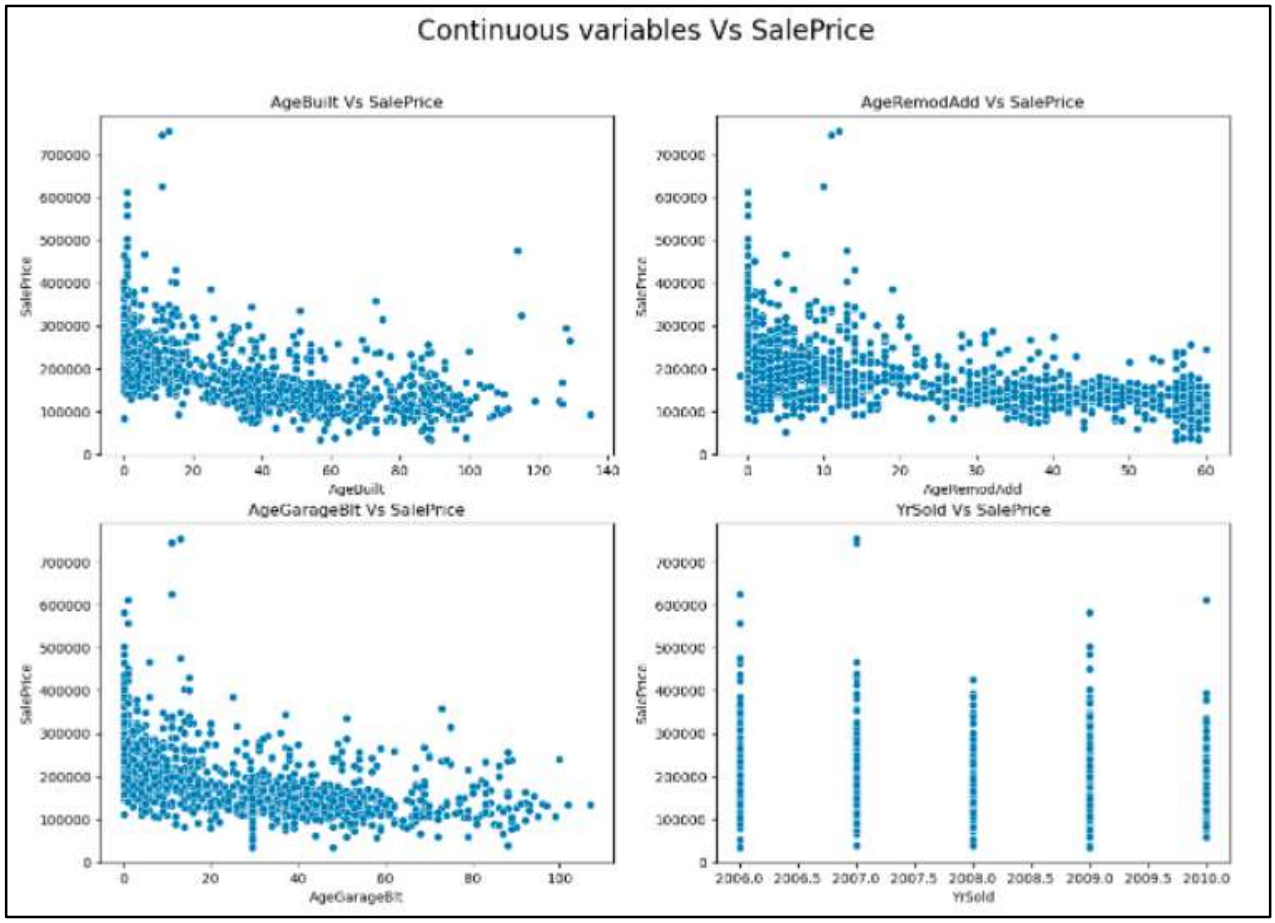
Observation: We can see a weak and positive relationship between sale price and LotArea, MasVnrArea and WoodDeckSF respectively. Sale prices increase invariably as the aforementioned variables increase in values.



Observation: There is a clear positive linear relationship between Sale Price and the variables. The sale price is high that is 100000-300000 when basement square feet lie up to 1500 square feet. When the unfinished basement area is below 1000 square feet, the sale price is high. As total basement area increases, sale price also increases. The sale price is high when Open porch area is below 200sf.



Observation: As the values of the variables are increasing, there appears to be an increase in sale price of the houses/properties, thereby implying a somewhat/clear positive linear relationship between them. The sale prices increase with the increase in the square footage of the 1st and 2nd floors, along with a similar trend with the above grade living area and the area of the garage in the house.



Observation: Price appears to be declining as the age variables increase. Year in which a building/property is sold does not seem to have a significant impact on the sale price of the building/property.

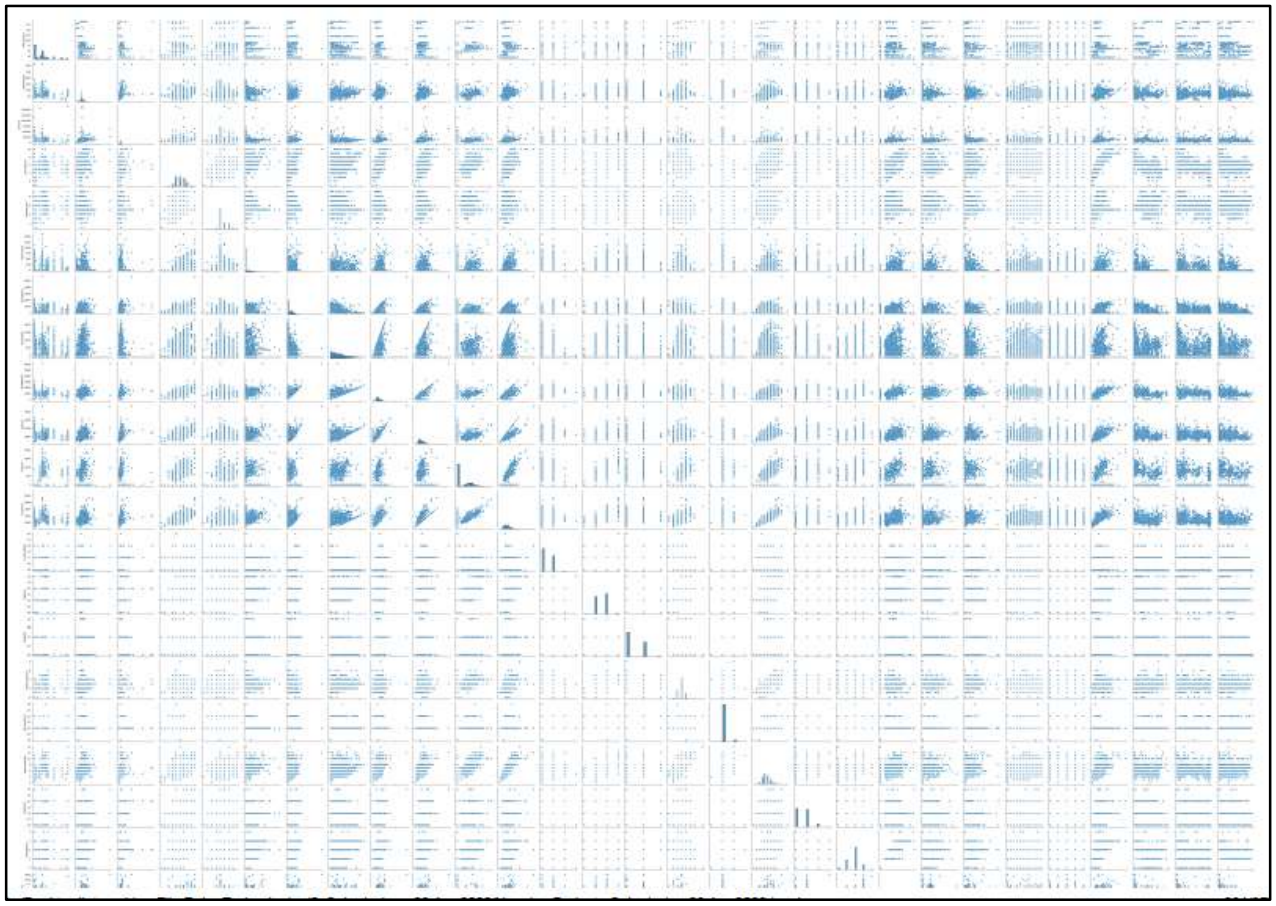
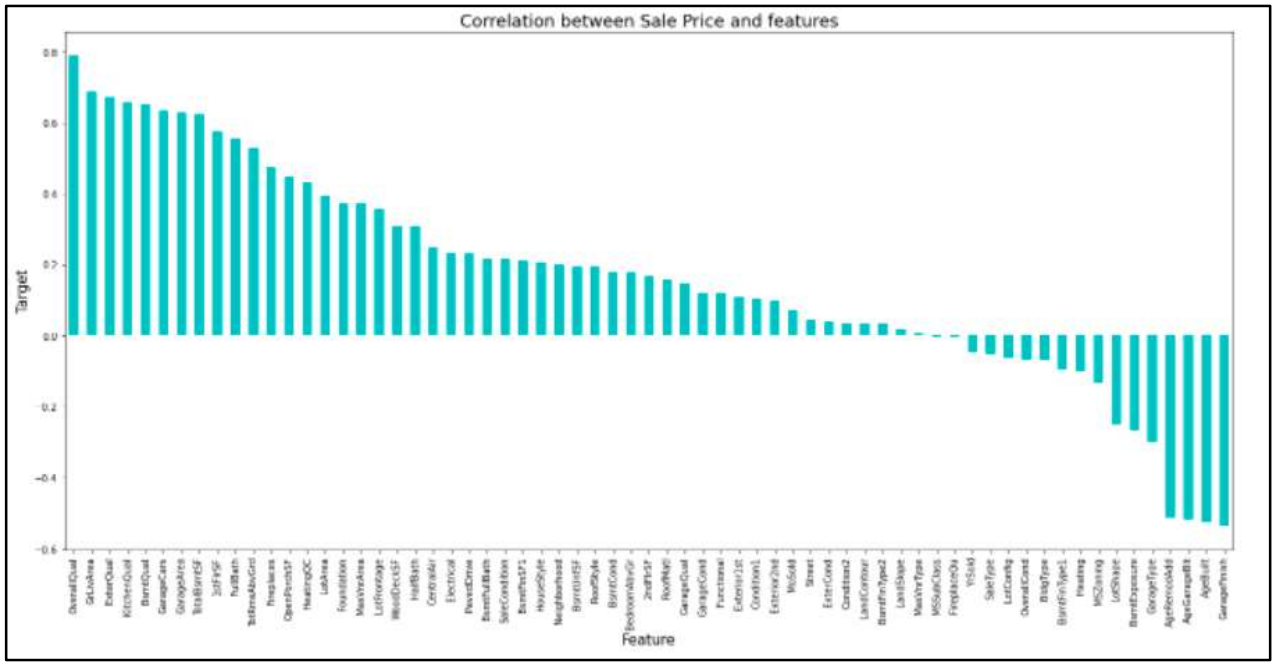


Fig: Partial snapshot of the pair plot

Observation: This pair plot depicts the pairwise relationship between the columns plotted based on the target variable "SalePrice." The relationship between the features and the label can be seen here.

Here, we can see the correlation between the features, and on the diagonal, we can see the distribution plot, which shows whether the columns are skew.

The linear relationship between the features is also visible.



Variable	Correlation Coefficient
SalePrice	1.000000
OverallQual	0.789185
GrLivArea	0.688210
ExterQual	0.672665
KitchenQual	0.659228
BsmtQual	0.653265
GarageCars	0.634573
GarageArea	0.627504
TotalBsmtSF	0.624311
1stFlrSF	0.575665
FullBath	0.554988
TotRmsAbvGrd	0.528564
Fireplaces	0.475531
OpenPorchSF	0.449961
HeatingQC	0.431516
LotArea	0.394343
Foundation	0.374169
MasVnrArea	0.373364
LotFrontage	0.358470
WoodDeckSF	0.310128
HalfBath	0.309808
CentralAir	0.246754

Variable	Correlation Coefficient
Electrical	0.234621
PavedDrive	0.231707
BsmtFullBath	0.218620
SaleCondition	0.217687
BsmtFinSF	0.212858
HouseStyle	0.205502
Neighborhood	0.198942
BsmtUnfSF	0.193140
RoofStyle	0.192654
BsmtCond	0.181625
BedroomAbvGr	0.177452
2ndFlrSF	0.167816
RoofMatl	0.159865
GarageQual	0.147272
GarageCond	0.119962
Functional	0.118673
Exterior1st	0.108451
Condition	0.105820
Exterior2nd	0.097541
MoSold	0.072764
Street	0.044753
ExterCond	0.038282

Variable	Correlation Coefficient
Condition	0.033956
LandContour	0.032836
BsmtFinType	0.032285
LandSlope	0.015485
MasVnrType	0.007732
MSSubClass	-0.001252
FireplaceQu	-0.004503
YrSold	-0.045508
SaleType	-0.050851
LotConfig	-0.060452
OverallCond	-0.065642
BldgType	-0.066028
BsmtFinType	-0.092109
Heating	-0.100021
MSZoning	-0.133221
LotShape	-0.248171
BsmtExposure	-0.268559
GarageType	-0.299470
AgeRemodAdd	-0.510784
AgeGarageBlt	-0.516445
AgeBuilt	-0.526644
GarageFinish	-0.537121

Interpretation of the Results

Visualizations:

- Through the univariate analysis conducted using the plotting of bar plots and pie plots, we were able to understand the distributions of individual variables across the data, while drawing some unique insights into the features of houses sold and preferences of housing customers.
- The bivariate analysis presented us with invaluable insights into the relationship between the features and the sale price of the house.
- Multivariate analysis through a pair plot helped us understand the distribution of data and relationships each feature held with each other
- The visualization of outliers in the data helped us decide the optimum course for removal of outliers, while the heatmap documented the correlation each feature had with each other and the target.

Pre-processing:

- Several data pre-processing steps, as detailed earlier in the report, were undertaken to make the dataset easier to interpret and use. Unwanted or inappropriate records / elements in the data were removed or fixed to avoid their negative effects on the prediction model's accuracy.

Modelling:

- Once the training dataset was ready for building the prediction model, a 70:30 train test split was created. The best random state was also computed for the purpose of training the model on Random Forest Regression Algorithm and document an initial R^2 accuracy score. Four different algorithms (LR, RF, SVR and DT) were trained on the dataset, following which, their R^2 scores, mean squared errors and cross validation scores were computed and documented. It was evident from the results that random forest regression would provide us with the best accuracy in predicting the sale prices of houses. Post the hyperparameter tuning, we had our best model for predicting the sale prices of houses. On the training dataset, the results on key metrics for the best model were as follows:

```
R2_Score: 89.35081751912671
RMSE value: 25301.04209793543
MAE: 17860.601375220773
MSE: 640142731.2415009
```

- The best prediction model from the steps undertaken was saved for predicting the housing prices on the training dataset.
- Once the test dataset was evaluated and pre-processed and ready for prediction, the prediction model saved was reloaded and the predictions were processed and documented on the test dataset. A snapshot of the dataframe created of the predicted sale prices is attached below:

	SalePrice
0	88381.171924
1	160058.432855
2	120212.322375
3	224234.245770
4	116523.373577
5	291520.403893
6	151281.561178
7	97341.899887
8	122512.999688
9	152371.124143
10	451607.080624

CONCLUSION

Key Findings and Conclusions of the Study

In this study, we have used multiple machine learning models to predict the house sale price. We have gone through the data analysis by performing feature engineering and finding the relation between features and labels. And got the important feature and predicted the price by building ML models.

The house's structural features the structural attributes of the house, such as lot size, lot shape, quality and condition of the house, garage capacity, rooms, Lot frontage, number of bedrooms, bathrooms, overall finishing of the house, and so on, all have a significant impact on the house price. The characteristics of the neighbourhood can be considered when determining the price of a home. Various plots aided in visualising the feature-label relationships, confirming the significance of structural and locational attributes in estimating Sale Prices. Because the Training dataset was so small, some of the outliers had to be kept for the models to be properly trained. Despite working on a small dataset, the Random Forest Regressor performed well due to its robustness to outliers and indifference to nonlinear features.

- **Which variables are important to predict the price of the house?**

  Overall quality is the most important and has the greatest positive impact. Furthermore, features such as GarageArea, LotArea, 1stFlrSF, TotalBsmtSF, and so on have a somewhat linear relationship with the price variable.

- **How do these variables describe the price of the house?**

  - ✓ Houses with very high overall quality, such as material and finish, have a high sale price. Also, we can see from the plot that as the overall quality of the house improves, so does the sale price. That is, there is a strong linear relationship between the SalePrice and the OverallQual. As a result, if the seller constructs the house with these qualities in mind, the house's sale price will rise.

  - ✓ The SalePrice and 1stFlrSF have a linear relationship. As we can see, as the size of the first-floor increases, so does the price. As a result, people prefer to live in houses with only 1-2 floors, and the cost of the house rises as a result.

  - ✓ We've also seen a positive linear relationship between the SalePrice and the GarageArea. As the size of the garage increases, so does the sale price.

  - ✓ There is positive linear relation between sale price and TotalBsmtSF. As total basement area increases, sale price also increases.

  - ✓ We have built many ML models using features that have some relationship with the target and have seen an increase in the accuracy of the best model.

Learning Outcomes of the Study in respect of Data Science

While working on this project, insights were gained about the housing market and how machine learning models can predict the price of a house, which helps sellers and buyers understand the future value of the house. The project piqued high interest because the dataset contained a variety of data types and variables. To visualise the relationship between target and features, a variety of plotting techniques were used. The graphical representation assisted in the understanding the importance of the features and how they affect the sale price of a house. One of the most important aspects of this project was data cleaning and pre-processing, which required the use of imputation methods to replace all null values and dealt with features with zero values and time variables.

Finally, the goal of predicting the house price based on the test data was achieved based on the random forest prediction model built. This may help sellers and buyers understand the housing market further. Machine learning models and data analytics techniques are surely critical in solving such problems. It also informs customers about the future value of their homes.

Limitations of this work and Scope for Future Work

Because of advancements in computing technology, it is now possible to examine social data that could not previously be captured, processed, and analysed. Machine learning analytical techniques can be used in housing research. This study is an exploratory attempt to estimate housing prices using machine learning algorithms.

While this model is beneficial to the city of Ames, it does have limitations. Certain predictors' importance may not translate well to other cities. In such cities, the size of the garage or the square footage of the second floor may not have the same numerical value. That is, the measurements of those values may be important in other cities, but they may have different numerical weights.

During the data pre-processing stages, concatenation of the train and test datasets was unsuitable due to potential of data leakage. Several records were filled using data imputation techniques because of the presence of null values and zeros.

Demographic data (age, income, buyer regional preferences, and reason for purchasing a home) is critical for understanding the housing market. Interest rates have an impact on the price and demand for homes. Economic cycles influence real estate prices, as do government policies, regulations, and legalizations, which are all important factors that may influence house sales. The availability of data on the aforementioned features would aid in the development of a predictive model that would more accurately understand the relationship between the features and the target variable, resulting in more accurate predictions.

One of the major specific prospects is to expand the estate database to include more cities, allowing users to explore more houses and make more informed decisions.

It is recommended that the organization use this model to get an idea of the actual price when buying a house in the area covered by the dataset. The model can also be used with datasets from different cities and areas if they contain the same features. It is advised to consider the features that were deemed most important in this study to better estimate the sale price of houses.

References

1. https://www.investopedia.com/terms/r/realestate.asp
2. Konwar, Robart & Kakati, Angshuman & Das, Bhagyashree & Shah, Borah & Muchahari, Monoj. (2021). House Price Prediction Using Machine Learning. The Journal of Philosophy Psychology and Scientific Methods. 9. 2455-6211.
3. Bruno Klausde Aquino Afonso, Luckeciano Carvalho Melo, Willian Dihanster Gomesde Oliveira, Samuel Brunoda Silva Sousa, Lilian Berton. Housing Prices Prediction with a Deep Learning and Random Forest Ensemble
4. https://en.wikipedia.org/wiki/NumPy
5. https://en.wikipedia.org/wiki/Pandas_(software)
6. https://seaborn.pydata.org/
7. https://en.wikipedia.org/wiki/Matplotlib
8. https://docs.scipy.org/doc/scipy/reference/stats.html
9. https://en.wikipedia.org/wiki/Scikit-learn
10. https://joblib.readthedocs.io/en/latest/
11. https://www.javatpoint.com/linear-regression-in-machine-learning
12. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84
13. https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0
14. https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/
15. https://www.wfmj.com/story/43657444/valley-realtors-relax-in-worrying-about-a-housing-market-crash