HOUSING INVESTMENT PROJECT

Submitted by:
Aneesha B Soman

ACKNOWLEDGMENT

Acknowledgement The success and final outcome of this project required a lot of guidance and assistance from Sajid Choudhary Sir and I am Extremely fortunate to have got this all along the completion of my project work Whatever I have done is only due to such guidance and assistance and I would not forget to thank him.

I respect and thank Sajid Choudhary Sir, for giving me an opportunity to do the project work in Data Modelling and Analytics and providing us all support and guidance which made me complete the project on time . I am extremely grateful to him for providing such a nice support and guidance though he had busy schedule managing the company affairs.

I have also referred to various articles in Towards Data Science and Kaggle to obtain codes on various visualisation methods.

INTRODUCTION

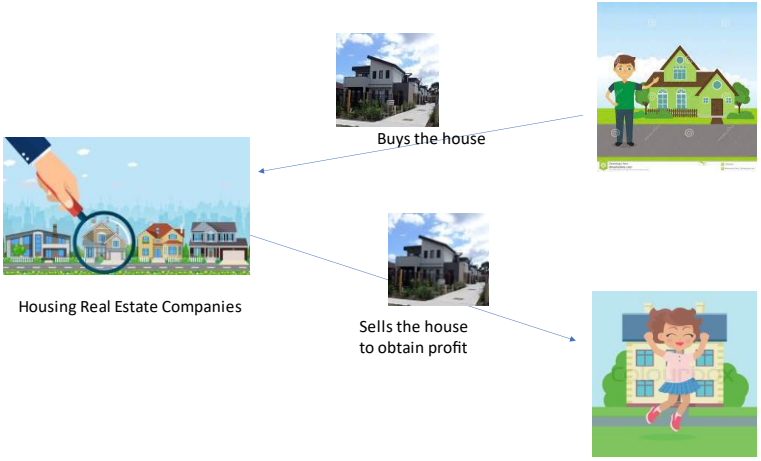
Business Problem Framing

The project deals with building a model for predicting the sales price for houses In Australia. It also deals in understanding the factors that are most significant for the Sales Price and Understanding of How the variables describe the price of the house

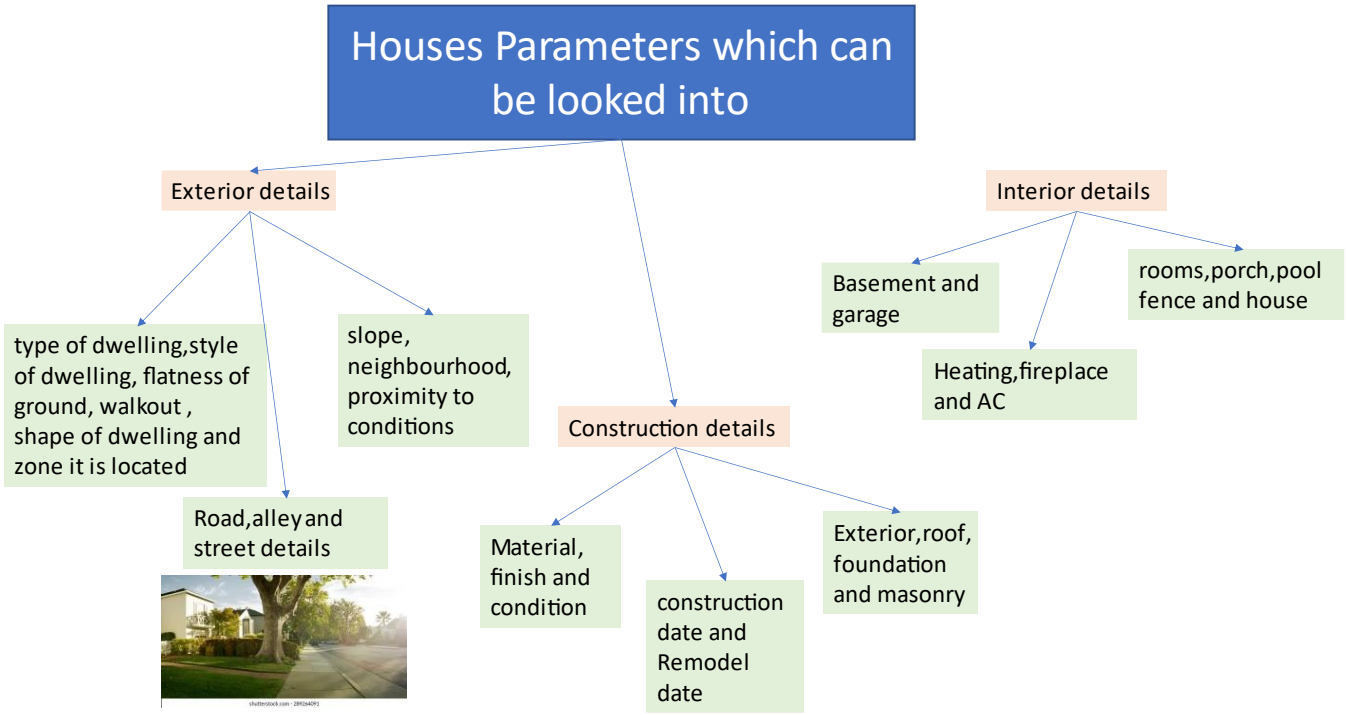
The Model will help the company to predict the value of the prospective properties and Accordingly manipulate the strategy of the firm and concentrate the investments on areas that will yield high returns.It will also the company to understand the pricing dynamics of a new market

Conceptual Background of the Domain Problem

Real estate sector comprises four sub sectors - housing, retail, hospitality, and commercial. The sector the project deals with is the housing sector.



Review of Literature



The Selling price depends on these parameters, with the highest dependence on Overall Quality, Above grade (ground) living area square feet, Full bathrooms above grade, Size of garage in car capacity and Year Built

Motivation for the Problem Undertaken

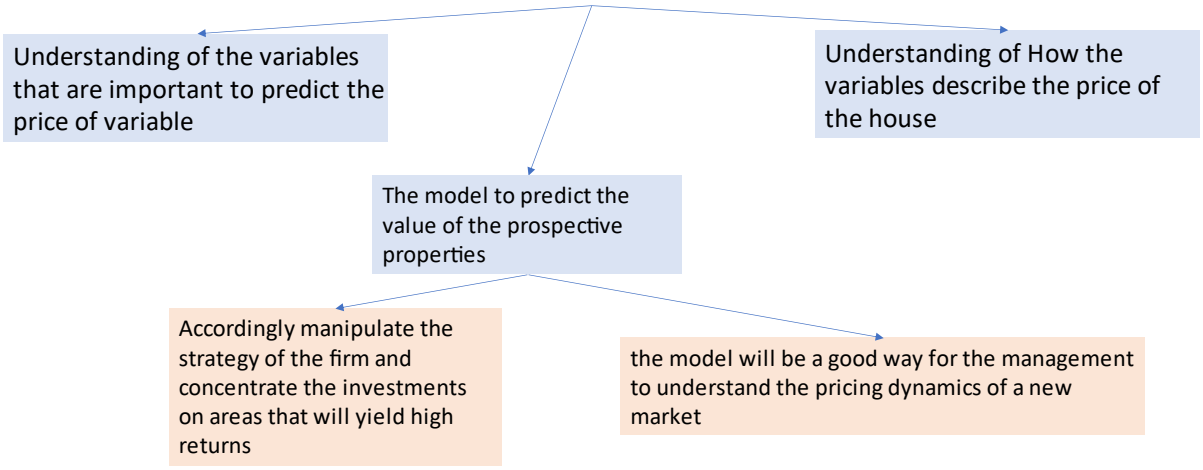
The growth of Real Estate sector is well complemented by the growth of the corporate environment and the demand for office space as well as urban and semi-urban accommodations.

As per the **outlook for Australian real estate in 2020** Positive outlook for 2020 with prices rising across Australia A year on from the clouds of uncertainty which welcomed 2019, the start of 2020 presents a much more positive outlook for Australia's residential markets. Strong price growth has returned to Sydney and Melbourne and is expected to spread to more affordable markets, Brisbane in particular.

Apartment supply cycles in the major capitals are past their peak and vacancy levels are well controlled. With cost of debt low and lending volumes starting to turn, investors should gradually return. This should encourage well placed developers to begin marketing larger projects again so they are at the forefront of the next development cycle post 2021.

Buisnessgoal

A US-based housing company named **Surprise Housing** to enter the Australian market, needs the



Monetary Benfits from the model

1. manipulate the strategy of the firm and concentrate the investments on areas that will yield high returns
2. Invest of properties showing marginal probability of loss

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem

Mathematical model	Data is analysed statistically Analysed through variance inflation factor Analysed through correlation and multicollinearity
Analytical graphs	Graphical modelling done through seaborn and matplotlib

Data Sources and their formats

1.Data orgin:

Data is obtained from US-based housing company named Surprise Housing

2.Description of data:

a.Data obtained was in csv format

b.Data had 80 columns and the column names were :

```
'Id', 'MSSubClass', 'MSZoning', 'LotFrontage', 'LotArea', 'Street',
      'Alley', 'LotShape', 'LandContour', 'Utilities', 'LotConfig',
      'LandSlope', 'Neighborhood', 'Condition1', 'Condition2',
'BldgType',
      'HouseStyle', 'OverallQual', 'OverallCond', 'YearBuilt',
'YearRemodAdd',
      'RoofStyle', 'RoofMatl', 'Exterior1st', 'Exterior2nd',
'MasVnrType',
      'MasVnrArea', 'ExterQual', 'ExterCond', 'Foundation',
'BsmtQual',
      'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinSF1',
      'BsmtFinType2', 'BsmtFinSF2', 'BsmtUnfSF', 'TotalBsmtSF',
'Heating',
      'HeatingQC', 'CentralAir', 'Electrical', '1stFlrSF', '2ndFlrSF',
      'LowQualFinSF', 'GrLivArea', 'BsmtFullBath', 'BsmtHalfBath',
'FullBath',
```

'HalfBath', 'BedroomAbvGr', 'KitchenAbvGr', 'KitchenQual',
'TotRmsAbvGrd', 'Functional', 'Fireplaces', 'FireplaceQu',
'GarageType',
'GarageYrBlt', 'GarageFinish', 'GarageCars', 'GarageArea',
'GarageQual',
'GarageCond', 'PavedDrive', 'WoodDeckSF', 'OpenPorchSF',
'EnclosedPorch', '3SsnPorch', 'ScreenPorch', 'PoolArea',
'PoolQC',
'Fence', 'MiscFeature', 'MiscVal', 'MoSold', 'YrSold', 'SaleType',
'SaleCondition'

c.Snapshot of data

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1
0	337	20	RL	86.0	14157	Pave	NaN	IR1	HLS	AllPub	Corner	Gtl	StoneBr	Norm
1	1018	120	RL	NaN	5814	Pave	NaN	IR1	Lvl	AllPub	CulDSac	Gtl	StoneBr	Norm
2	929	20	RL	NaN	11838	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm
3	1148	70	RL	75.0	12000	Pave	NaN	Reg	Bnk	AllPub	Inside	Gtl	Crawfor	Norm
4	1227	60	RL	86.0	14598	Pave	NaN	IR1	Lvl	AllPub	CulDSac	Gtl	Somerst	Feedr
5	650	180	RM	21.0	1936	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm

Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemodAdd	RoofStyle	RoofMatl	Exterior1st	Exterior2nd	MasVnrType
Norm	1Fam	1Story	9	5	2005	2006	Hip	CompShg	VinylSd	VinylSd	Stone
Norm	TwnhsE	1Story	8	5	1984	1984	Gable	CompShg	HdBoard	HdBoard	None
Norm	1Fam	1Story	8	5	2001	2001	Hip	CompShg	VinylSd	VinylSd	None
Norm	1Fam	2Story	7	7	1941	1950	Gable	CompShg	MetalSd	MetalSd	None
Norm	1Fam	2Story	6	5	2007	2007	Gable	CompShg	VinylSd	VinylSd	Stone
Norm	Twnhs	SFoyer	4	6	1970	1970	Gable	CompShg	CemntBd	CmentBd	None

MasVnrArea	ExterQual	ExterCond	Foundation	BsmtQual	BsmtCond	BsmtExposure	BsmtFinType1	BsmtFinSF1	BsmtFinType2	BsmtFinSF2
200.0	Gd	TA	PConc	Ex	TA	Gd	GLQ	1249	Unf	0
0.0	Gd	TA	CBlock	Gd	TA	Av	GLQ	1036	Unf	0
0.0	Gd	TA	PConc	Gd	TA	Av	Unf	0	Unf	0
0.0	TA	TA	CBlock	TA	TA	No	Rec	275	Unf	0
74.0	Gd	TA	PConc	Gd	TA	Mn	Unf	0	Unf	0

BsmtUnfSF	TotalBsmtSF	Heating	HeatingQC	CentralAir	Electrical	1stFlrSF	2ndFlrSF	LowQualFinSF	GrLivArea	BsmtFullBath	BsmtHalfBath
673	1922	GasA	Ex	Y	SBrkr	1922	0	0	1922	1	0
184	1220	GasA	Gd	Y	SBrkr	1360	0	0	1360	1	0
1753	1753	GasA	Ex	Y	SBrkr	1788	0	0	1788	0	0
429	704	GasA	Ex	Y	SBrkr	860	704	0	1564	0	0
894	894	GasA	Ex	Y	SBrkr	894	1039	0	1933	0	0

FullBath	HalfBath	BedroomAbvGr	KitchenAbvGr	KitchenQual	TotRmsAbvGrd	Functional	Fireplaces	FireplaceQu	GarageType	GarageYrBlt
2	0	3	1	Gd	8	Typ	1	Gd	Attchd	2005.0
1	0	1	1	Gd	4	Typ	1	Ex	Attchd	1984.0
2	0	3	1	Ex	7	Typ	1	TA	Attchd	2001.0
1	1	3	1	Fa	7	Typ	1	Gd	Attchd	1941.0
2	1	4	1	Gd	9	Typ	1	Gd	BuiltIn	2007.0

GarageFinish	GarageCars	GarageArea	GarageQual	GarageCond	PavedDrive	WoodDeckSF	OpenPorchSF	EnclosedPorch	3SsnPorch	ScreenPorch
Fin	3	676	TA	TA	Y	178	51	0	0	0
RFn	2	565	TA	TA	Y	63	0	0	0	0
RFn	2	522	TA	TA	Y	202	151	0	0	0
Unf	1	234	TA	TA	Y	0	0	0	0	0
Fin	3	668	TA	TA	Y	100	18	0	0	0

PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition
0	NaN	NaN	NaN	0	7	2007	WD	Normal
0	NaN	NaN	NaN	0	8	2009	COD	Abnorml
0	NaN	NaN	NaN	0	6	2009	WD	Normal
0	NaN	NaN	NaN	0	7	2009	WD	Normal
0	NaN	NaN	NaN	0	1	2008	WD	Normal

d.Meaning of columns

MSSubClass	Identifies the type of dwelling involved in the sale
MSZoning	Identifies the general zoning classification of the sale
LotFrontage	Linear feet of street connected to property
LotArea	Lot size in square feet
Street	Type of road access to property
Alley	Type of alley access to property
LotShape	General shape of property
LandContour	Flatness of the property
Utilities	Type of utilities available
LotConfig	Lot configuration
LandSlope	Slope of property

Neighborhood	Physical locations within Ames city limits
Condition1	Proximity to various conditions
Condition2	Proximity to various conditions (if more than one is present)
BldgType	Type of dwelling
HouseStyle	Style of dwelling
	Rates the overall material and finish of the house
OverallCond	Rates the overall condition of the house
YearBuilt	Original construction date
OverallQual YearRemodAdd	Remodel date (same as construction date if no remodeling or additions)
RoofStyle	Type of roof
RoofMatl	Roof material
Exterior1st	Exterior covering on house
Exterior2nd	Exterior covering on house (if more than one material)
MasVnrType	Masonry veneer type
MasVnrArea	Masonry veneer area in square feet
ExterQual	Evaluates the quality of the material on the exterior
ExterCond	Evaluates the present condition of the material on the exterior
Foundation	Type of foundation
BsmtQual	Evaluates the height of the basement
BsmtCond	Evaluates the general condition of the basement
BsmtExposure	Refers to walkout or garden level walls
BsmtFinType1	Rating of basement finished area
BsmtFinSF1	Type 1 finished square feet
BsmtFinType2	Rating of basement finished area (if multiple types)
BsmtFinSF2	Type 2 finished square feet
BsmtUnfSF	Unfinished square feet of basement area
TotalBsmtSF	Total square feet of basement area
Heating	Type of heating
HeatingQC	Heating quality and condition
CentralAir	Central air conditioning
Electrical	Electrical system

1stFlrSF	First Floor square feet
2ndFlrSF	Second floor square feet
LowQualFinSF	Low quality finished square feet (all floors)
GrLivArea	Above grade (ground) living area square feet
BsmtFullBath	Basement full bathrooms
BsmtHalfBath	Basement half bathrooms
FullBath	Full bathrooms above grade
HalfBath	Half baths above grade
Bedroom	Bedrooms above grade (does NOT include basement bedrooms)
Kitchen	Kitchens above grade
KitchenQual	Kitchen quality
TotRmsAbvGrd	Total rooms above grade (does not include bathrooms)
Functional	Home functionality (Assume typical unless deductions are warranted)
Fireplaces	Number of fireplaces
FireplaceQu	Fireplace quality
GarageType	Garage location
GarageYrBlt	Year garage was built
GarageFinish	Interior finish of the garage
GarageCars	Size of garage in car capacity
GarageArea	Size of garage in square feet
GarageQual	Garage quality
GarageCond	Garage condition
WoodDeckSF	Wood deck area in square feet
OpenPorchSF	Open porch area in square feet
EnclosedPorch	Enclosed porch area in square feet
3SsnPorch	Three season porch area in square feet
ScreenPorch	Screen porch area in square feet
PoolArea	Pool area in square feet
PoolQC	Pool quality
Fence	Fence quality
MiscFeature	Miscellaneous feature not covered in other categories
MiscVal	$Value of miscellaneous feature

MoSold	Month Sold (MM)
YrSold	Year Sold (YYYY)
SaleType	Type of sale
SaleCondition	Condition of sale

e.There are both numerical and categorical columns.

f.label is Price of the house

3. Data engineering

a. renaming of data columns done:

	Id	type of dwelling	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	Slope of property	Physical locations within Ames city limits	Proximity to various conditions
0	127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	NPkVill	Norm
1	889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	Inside	Mod	NAmes	Norm
2	793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	CulDSac	Gtl	NoRidge	Norm
3	110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	NWAmes	Norm
4	422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	FR2	Gtl	NWAmes	Norm

Proximity to various conditions (if more than one is present)	Type of dwelling	Style of dwelling	OverallQual	OverallCond	YearBuilt	YearRemodAdd	Type of roof	Roof material	Exterior covering on house	Exterior covering on house (if more than one material)	Physical locations within Ames city limits	Masonry veneer area in square feet	quality of the material on the exterior
Norm	TwnhsE	1Story	6	5	1976	1976	Gable	CompShg	Plywood	Plywood	None	0.0	TA
Norm	1Fam	1Story	8	6	1970	1970	Flat	Tar&Grv	Wd Sdng	Wd Sdng	None	0.0	Gd
Norm	1Fam	2Story	7	5	1996	1997	Gable	CompShg	MetalSd	MetalSd	None	0.0	Gd
Norm	1Fam	1Story	6	6	1977	1977	Hip	CompShg	Plywood	Plywood	BrkFace	480.0	TA
Norm	1Fam	1Story	6	7	1977	2000	Gable	CompShg	CemntBd	CmentBd	Stone	126.0	Gd

present condition of the material on the exterior	Type of foundation	Evaluates the height of the basement	general condition of the basement	walkout or garden level walls	Rating of basement finished area	Type 1 finished square feet	Rating of basement finished area (if multiple types)	Type 2 finished square feet	Unfinished square feet of basement area	Total square feet of basement area	Heating quality and condition	HeatingQC	Central air conditioning
TA	CBlock	Gd	TA	No	ALQ	120	Unf	0	958	1078	GasA	TA	Y
Gd	PConc	TA	Gd	Gd	ALQ	351	Rec	823	1043	2217	GasA	Ex	Y
TA	PConc	Gd	TA	Av	GLQ	862	Unf	0	255	1117	GasA	Ex	Y
TA	CBlock	Gd	TA	No	BLQ	705	Unf	0	1139	1844	GasA	Ex	Y
TA	CBlock	Gd	TA	No	ALQ	1246	Unf	0	356	1602	GasA	Gd	Y

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	BedroomAbvGr	KitchenAbvGr	Kitchen quality	Total rooms above grade (does not include bathrooms
SBrkr	958	0	0	958	0	0	2	0	2	1	TA	5
SBrkr	2217	0	0	2217	1	0	2	0	4	1	Gd	8
SBrkr	1127	886	0	2013	1	0	2	1	3	1	TA	8
SBrkr	1844	0	0	1844	0	0	2	0	3	1	TA	7
SBrkr	1602	0	0	1602	0	1	2	0	3	1	Gd	8

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	PavedDrive	Wood deck area	OpenPorchSF	Enclosed porch area	Three season porch area
Typ	1	TA	Attchd	1977.0	RFn	2	440	TA	TA	Y	0	205	0	0
Typ	1	TA	Attchd	1970.0	Unf	2	621	TA	TA	Y	81	207	0	0
Typ	1	TA	Attchd	1997.0	Unf	2	455	TA	TA	Y	180	130	0	0
Typ	1	TA	Attchd	1977.0	RFn	2	546	TA	TA	Y	0	122	0	0
Typ	1	TA	Attchd	1977.0	Fin	2	529	TA	TA	Y	240	0	0	0

Screen porch area in square feet	Pool area in square feet	Pool quality	Fence	MiscFeature	Value of miscellaneous feature	Month Sold	Year Sold	Type of sale	Condition of sale	SalePrice
0	0	NaN	NaN	NaN	0	2	2007	WD	Normal	128000
224	0	NaN	NaN	NaN	0	10	2007	WD	Normal	268000
0	0	NaN	NaN	NaN	0	6	2007	WD	Normal	269790
0	0	NaN	MnPrv	NaN	0	1	2010	COD	Normal	190000
0	0	NaN	NaN	NaN	0	6	2009	WD	Normal	215000

b.Dropped the unique value column Utilities

c. dropped unnecessary columns-Id

d.Checked all the columns if contains any other symbols other than Null values, found that many columns contain 0, which would need to be replaced with mean/mode depending on the datatype

e.1244 null values found in the dataframe and 23 columns having 0 in their Columns. However some of the columns have 0 as genuine values, which will not be removed

f.Columns which have more than 70% ‘0’ values are dropped and on the others the 0 is replaced with mean/mode

g. The data is assumed to be linear, Homogeneity of variances, Normality and Independence. Eda is done to remove the outliers to make data normal and linear.
Columns which have more than 70% null values are dropped and on the others the null value is replaced with mean/mode

h.Outliers are found on the numerical values and it is removed through zscore

i.One hot encoding done and then PCA done. Obtained 129 columns to be optimum for 95% variance

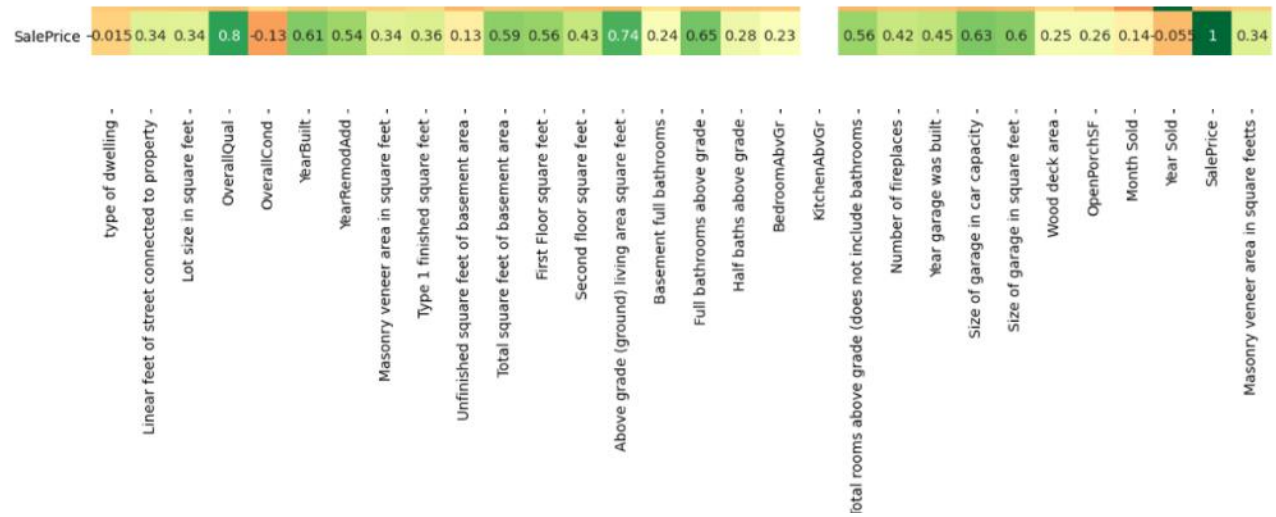
4.Data Inputs- Logic- Output Relationships

a.Input is numerical and categorical format and output is numerical format

b.Input and ouputs relationship is:

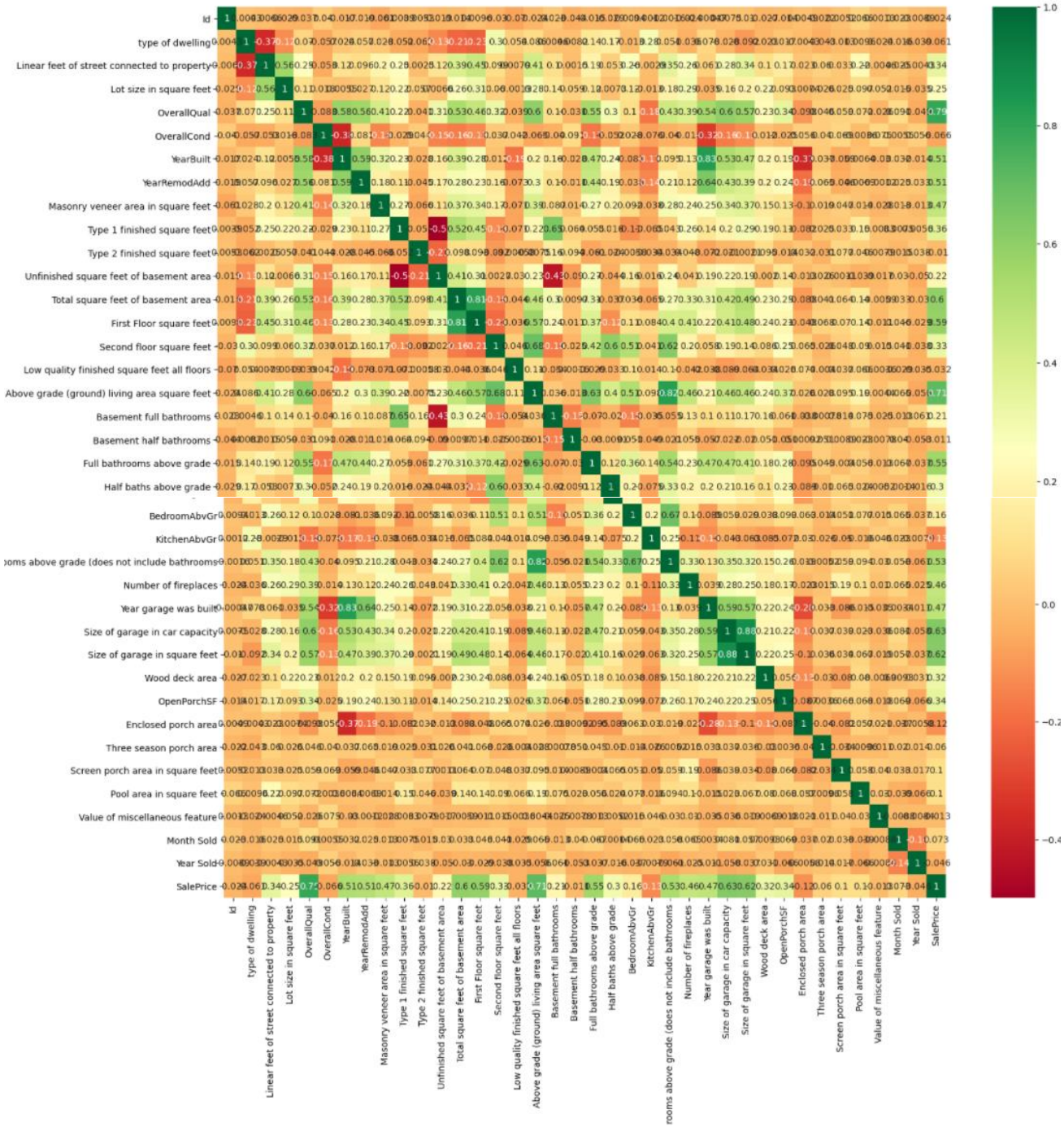
SalePrice	1.000000
OverallQual	0.798518
Above grade (ground) living area square feet	0.738217
Full bathrooms above grade	0.646966
Size of garage in car capacity	0.630109
YearBuilt	0.612338
Size of garage in square feet	0.599493
Total square feet of basement area	0.586724
First Floor square feet	0.562128
Total rooms above grade (does not include bathrooms	0.556870
YearRemodAdd	0.544277
Year garage was built	0.446236
Second floor square feet	0.425333
Number of fireplaces	0.422366
Type 1 finished square feet	0.359599
Masonry veneer area in square feetts	0.344676
Masonry veneer area in square feet	0.342551
Linear feet of street connected to property	0.339838
Lot size in square feet	0.336878
Half baths above grade	0.279214
OpenPorchSF	0.257764
Wood deck area	0.247307
Basement full bathrooms	0.238874
BedroomAbvGr	0.232722
Month Sold	0.142187
Unfinished square feet of basement area	0.131310
type of dwelling	0.015363
Year Sold	-0.055411
OverallCond	-0.133975
KitchenAbvGr	NaN

Name: SalePrice, dtype: float64



High relation can be seen between Overall Quality and above ground living area square feet

c.Relationship between inputs are:



High correlation seen between Total Square feet area and First floor square feet are Size of Garage in car capacity and size of garage in square feet has high correlation

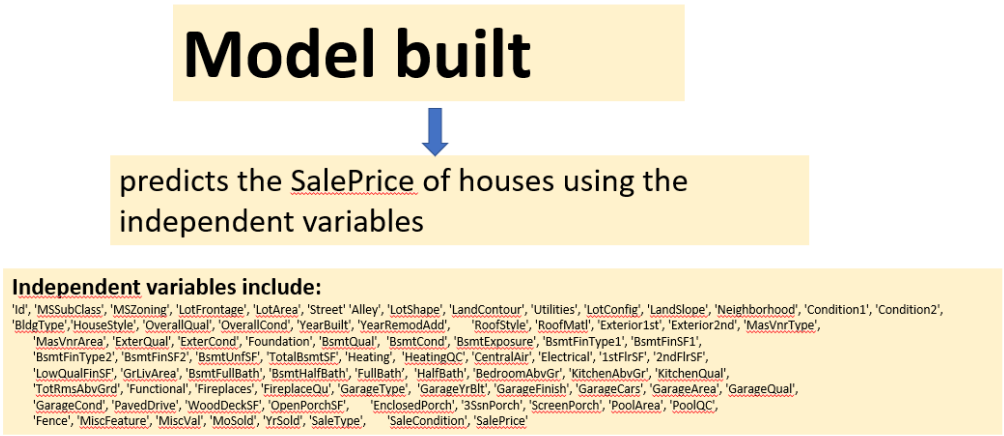
6.Hardware and Software Requirements and Tools Used

Library	Used in the project
Pandas library	1.Read the csv file, describe it,count values,converting date into usable format,dropping duplicates
Numpy library	Using zscore
Seaborn and matplotlib	For visualization
sklearn	Model building
GridSearchCV	hyperparameter tuning
pickle	saving data
Ridge,Lasso	Regularisation

Hardware: Windows 10

Softwares: Jupyter notebook

Model/s Development and Evaluation



1.Identification of possible problem-solving approaches

The data set was was analysed both statistically and graphically.
The statistical analysis showed that

- data to have outliers
- data to have null values
- data to have zero values
- datas independent variable had numerical datas and categorical datas

The **null values and zero values** were replaced with mean or mode, depending on the situation

Hence the datas outliers were removed(8%) and made **more normalised**

Features were engineered dropping columns which were not good for modelling

Performed one hot encoding on the categorical data and PCA was done to obtain optimum number of columns

The label was numerical data hence the approaches that can be applied are: LinearRegression, DecisionTreeRegressor and RandomForestRegressor

The data is split and the best random state is found. Then the data is split again with the best random state

2. Testing of Identified Approaches (Algorithms)

The label was categorical hence classification algorithms were used, which were LinearRegression, DecisionTreeRegressor and RandomForestRegressor

3.Run and Evaluate selected models

1.Models used:

```
#Linear Regression
from sklearn.linear_model import LinearRegression

LR=LinearRegression()
LR.fit(x_train,y_train)
predlr=LR.predict(x_test)
print("r2 score of LinearRegression model is",r2_score(y_test,predlr))
```

```
#DecisionTreeClassifier

dt=DecisionTreeRegressor()
dt.fit(x_train,y_train)
predlr=dt.predict(x_test)
print("r2 score of DecisionTreeRegressor model is",r2_score(y_test,predlr))
```

```
#Random forest regressor
rf=RandomForestRegressor()
rf.fit(x_train,y_train)
predlr=rf.predict(x_test)
print("r2 score of RandomForestRegressor model is",r2_score(y_test,predlr))
```

2.r2 score and cross val score was obtained for each of it:

LinearRegression	r2 score of LinearRegression model is 0.6688467396520761 mean_absolute_error is 18958.99830696684 mean squared error is 36587.22396862232
DecisionTreeClassifier	r2 score of DecisionTreeRegressor model is 0.6263201788492543 mean_absolute_error is 27982.723849372385 mean squared error is 38865.54522036389
Random forest regressor	r2 score of RandomForestRegressor model is 0.8367636601696511 mean_absolute_error is 17719.399665271965 mean squared error is 25687.590272304282

```
Cross validation score of Linear Regression model : 0.7502745352286656
Cross validation score of Decision Tree model : 0.5773745004771467
Cross validation score of Random Forest model : 0.8143902761815849
```

3.Key Metrics for success in solving problem under consideration

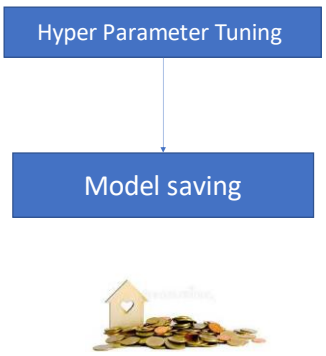
A.The **r2 score** was used.
R-squared is a metric of correlation. Correlation is measured by “r” and it **tells us how strongly two variables can be related**.

A correlation closer to **+1 means a strong relationship in the positive direction, while -1 means a stronger relationship in the opposite direction**.

A value closer to 0 means that there is not much of a relationship between the variables. R-squared is closely related to correlation.

B.Since r2 score of Randomforest Regressor is highest, we would be selecting it for hyperparamter tuning

C.Hyperparameter tuning



```
#RandomForestRegressor
paramters={'n_estimators':[2,4,5,6,8],
           'min_samples_split':[2,3,4],
           'min_samples_leaf':[2,3,4],
           'max_leaf_nodes':[2,3,4],
           'max_features':['auto','sqrt','log2'],
}
```

```
GCV=GridSearchCV(RandomForestRegressor(),paramters,cv=6)
```

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

```
GridSearchCV(cv=6, estimator=RandomForestRegressor(),
             param_grid={'max_features': ['auto', 'sqrt', 'log2'],
                         'max_leaf_nodes': [2, 3, 4],
                         'min_samples_leaf': [2, 3, 4],
                         'min_samples_split': [2, 3, 4],
                         'n_estimators': [2, 4, 5, 6, 8]})
```

```
GCV.best_params_
```

```
{'max_features': 'auto',
 'max_leaf_nodes': 4,
 'min_samples_leaf': 2,
 'min_samples_split': 3,
 'n_estimators': 8}
```

```
mod=RandomForestRegressor(n_estimators=8,max_leaf_nodes=4,min_samples_leaf=4,min_samples_split=4,max_features='auto')
```

```
mod.fit(x_train,y_train)
pred=mod.predict(x_test)
print(r2_score(y_test,pred)*100)
```

```
71.49402432757095
```

```
randomforest=RandomForestRegressor()
randomforest.fit(x_train,y_train)
```

```
7]: RandomForestRegressor()
```

Saving of model

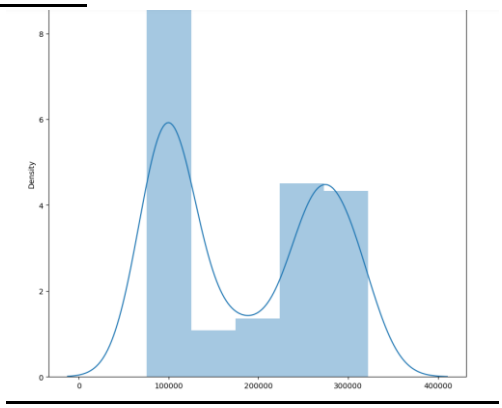
```
#saving RandomForestRegressor model
RandomForestRegressor_model=RandomForestRegressor()
RandomForestRegressor_model.fit(x_train,y_train)

filename='finalized_model.pickle'
pickle.dump(RandomForestRegressor_model,open(filename,'wb'))
```

```
#Adjusted R2
RandomForestRegressor_model.score(x_train,y_train)
```

```
9]: 0.9743467147146376
```

Finalised model

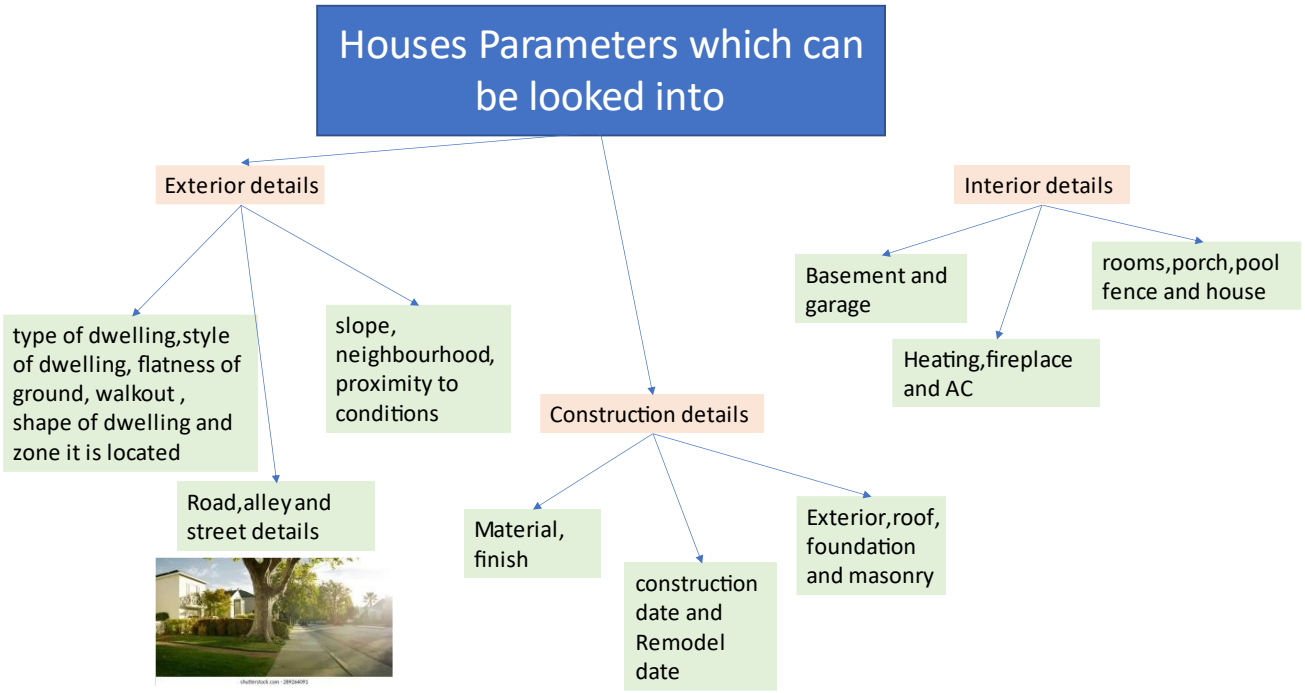


D.Model scores

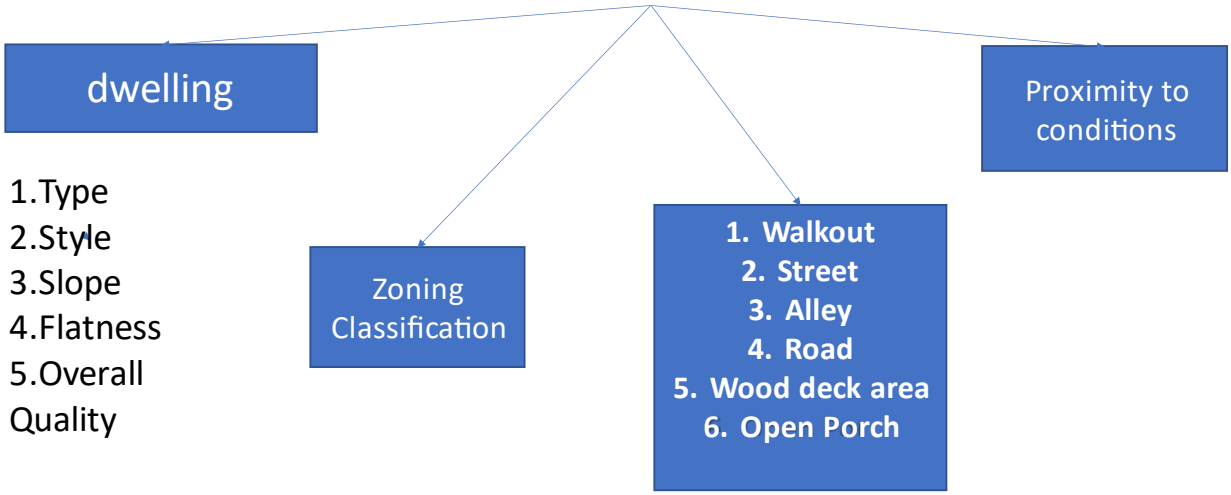
```
print("r2 score of RandomForestRegressor model is",r2_score(y_test,pred))
print("mean_absolute_error is",mean_absolute_error(y_test,pred))
print("mean squared error is",np.sqrt(mean_squared_error(y_test,pred)))
```

```
r2 score of RandomForestRegressor model is 0.7149402432757095
mean_absolute_error is 24212.989964939974
mean squared error is 33945.55365135884
```

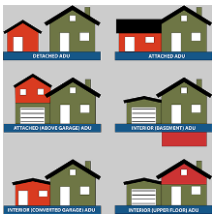
5.Visualizations



EXTERIOR DETAILS

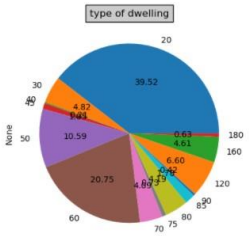


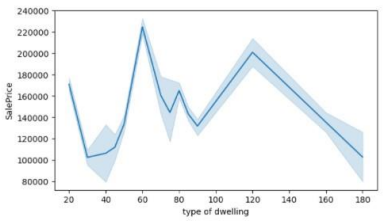
Dwelling

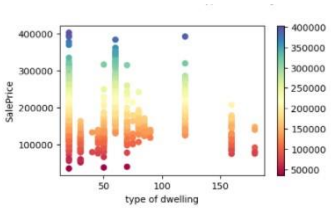


1.Type
2.Style
3.Slope
4.Flatness
5.Shape
6.Configuration

1.Type of dwelling





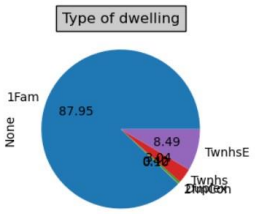


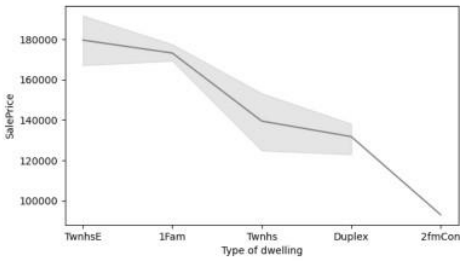

```
20    1-STORY 1946 & NEWER ALL STYLES
30    1-STORY 1945 & OLDER
40    1-STORY W/FINISHED ATTIC ALL AGES
45    1-1/2 STORY - UNFINISHED ALL AGES
50    1-1/2 STORY FINISHED ALL AGES
60    2-STORY 1946 & NEWER
70    2-STORY 1945 & OLDER
75    2-1/2 STORY ALL AGES
80    SPLIT OR MULTI-LEVEL
85    SPLIT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

1.The 1-STORY 1946 & NEWER ALL STYLES is highest among the types of dwellings

2.The Sale price is highest in 2-STORY 1946 & NEWER

Type of dwelling



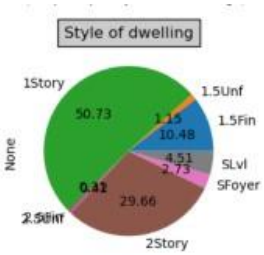


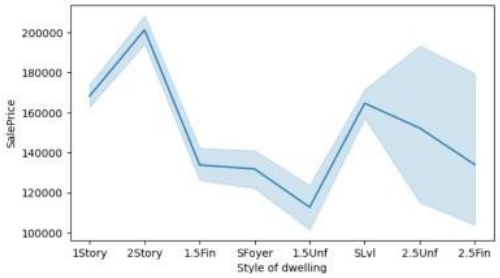
1.Maximum houses are single family detached

2.Maximum sale price is for Townhouse End Unit

```
1Fam     Single-family Detached
2FmCon   Two-family Conversion; originally built as one-family dwelling
Duplx    Duplex
TwnhsE   Townhouse End Unit
TwnhsI   Townhouse Inside Unit
```

Style of dwelling



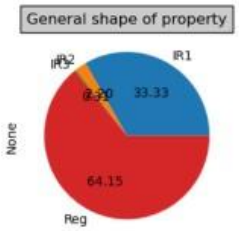


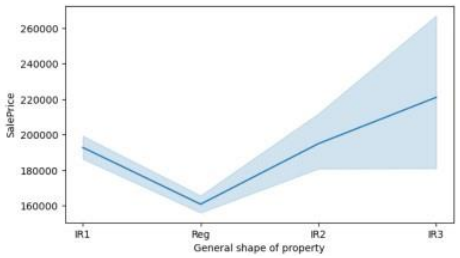
1Story	One story
1.5Fin	One and one-half story: 2nd level finished
1.5Unf	One and one-half story: 2nd level unfinished
2Story	Two story
2.5Fin	Two and one-half story: 2nd level finished
2.5Unf	Two and one-half story: 2nd level unfinished
SFoyer	Split Foyer
SLvl	Split Level

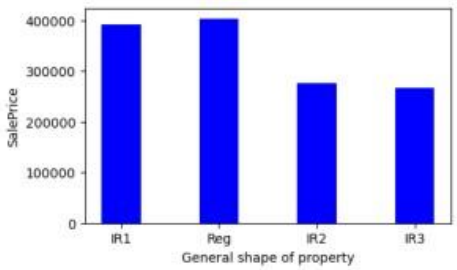
Maximum houses are 1 story but sale price of 2 storey is highest

General shape of property

Reg	Regular
IR1	Slightly irregular
IR2	Moderately Irregular
IR3	Irregular



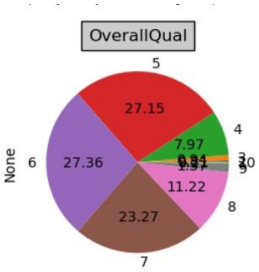


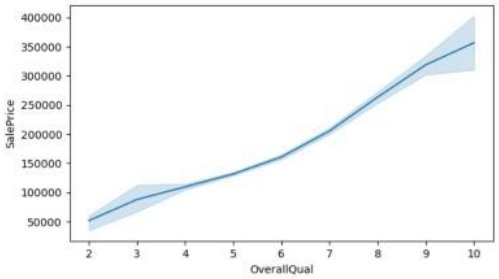


1.Maximum houses have regular shape

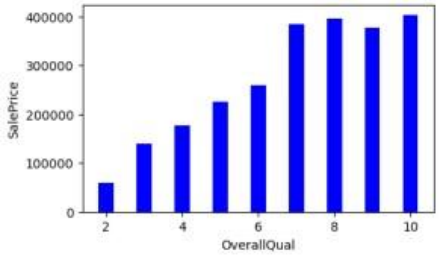
2.But the houses with slightly irregular shape shows the highest sales price

Overall Quality





10	Very Excellent
9	Excellent
8	Very Good
7	Good
6	Above Average
5	Average
4	Below Average
3	Fair
2	Poor
1	Very Poor

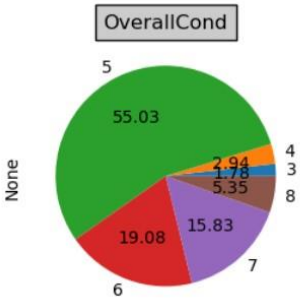


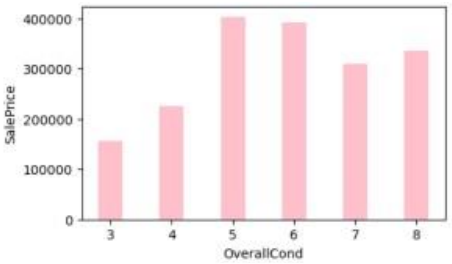
Maximum houses are above average

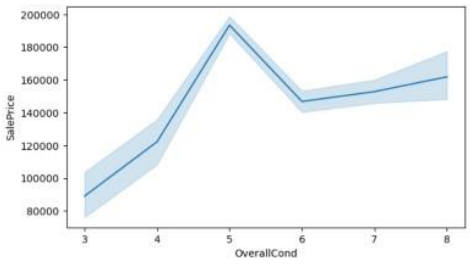
With increase in quality the price of the house increases

Overall Condition

10	Very Excellent
9	Excellent
8	Very Good
7	Good
6	Above Average
5	Average
4	Below Average
3	Fair
2	Poor
1	Very Poor







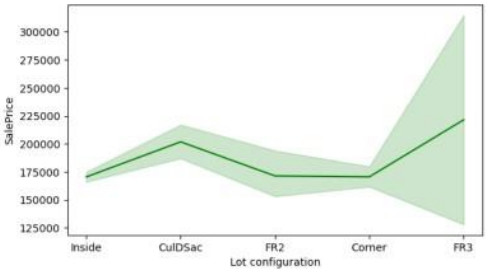
Most of the houses are average conditions.

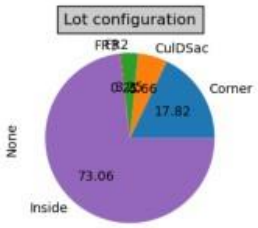
The prices increase upto average, then the price starts dropping towards very excellent

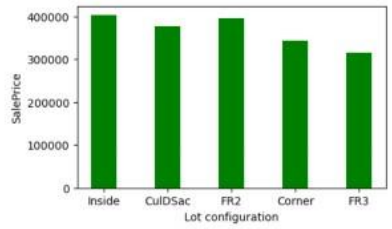
Property Details

- ❑Plot configuration
- ❑Flatness of the property
- ❑Slope of property

PLot configuration







Inside-Inside lot

Corner-Corner lot

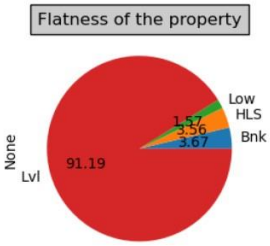
CulDSac-Cul-de-sac

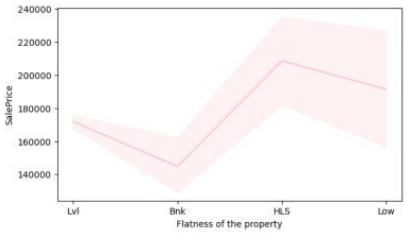
FR2 Frontage on 2 sides of property

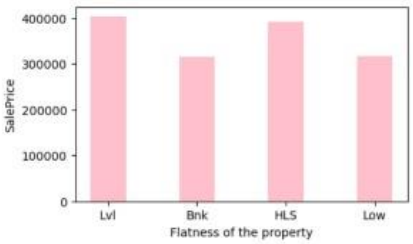
FR3 Frontage on 3 sides of property

1.Highest is Inside lot
2.Sale price is highest for Cul de sac

Flatness of the property







1.Most of the property was levelled

2.Property which were hillside with significant slope from side to side had highest sale price

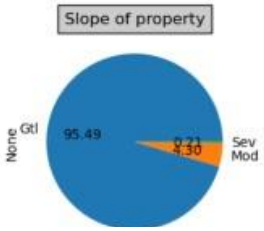
Lvl -Near Flat/Level

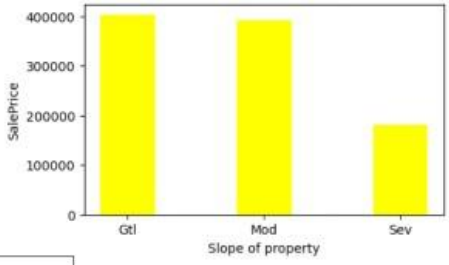
Bnk Banked - Quick and significant rise from street grade to building

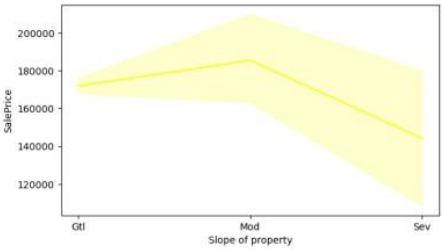
HLS Hillside - Significant slope from side to side

Low Depression

Slope of property







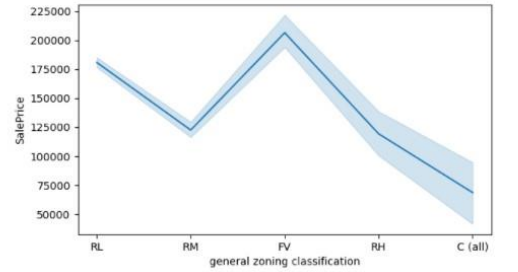
Gtl Gentle slope

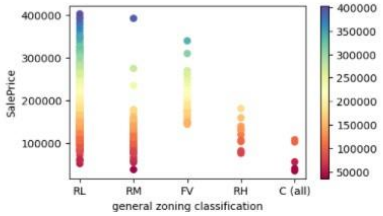
Mod Moderate Slope

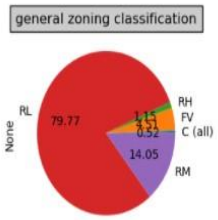
Sev Severe Slope

The highest number of properties have gentler slope and show higher sales price

general zoning classification







A Agriculture

C Commercial

FV Floating Village Residential

I Industrial

RH Residential High Density

RL Residential Low Density

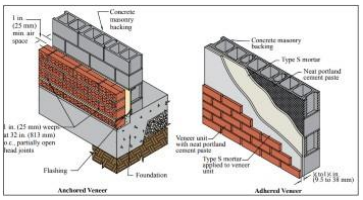
RP Residential Low Density Park

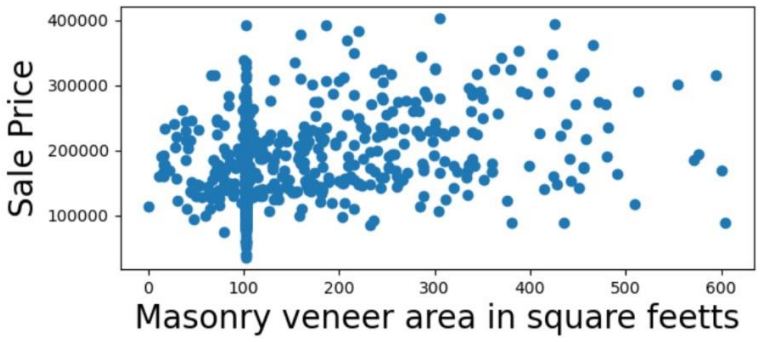
RM Residential Medium Density

1.Residential Low density is highest among the general zone classification

2.Residential Low Density and Residential Medium Density gives highest sale price

Masonry veneer area in square fee



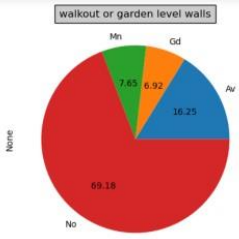


It does not impact the sale price much

OUTDOORS

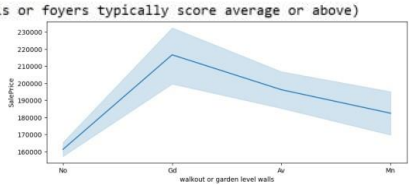
1.Walkout
2.Street
3.Alley
4.Road
5.Wood deck area
6.Open Porch

Walkout or garden level walls

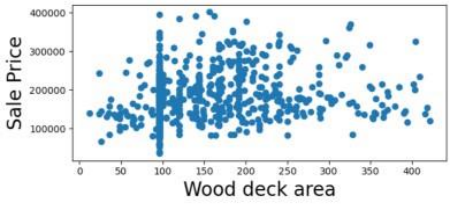


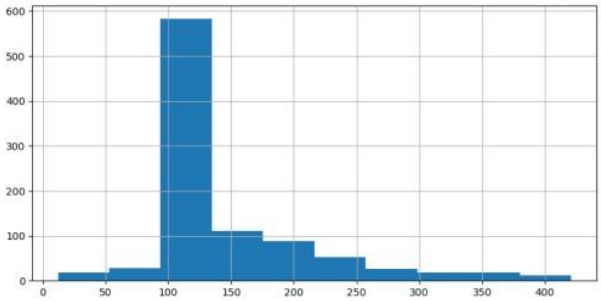
Maximum houses have no exposure to walkouts or gardens, but maximum sale price is for good exposure to it

```
Gd      Good Exposure
Av      Average Exposure (split levels or foyers typically score average or above)
Mn      Mimimum Exposure
No      No Exposure
NA      No Basement
```



Wood deck area

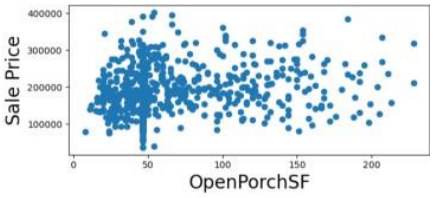


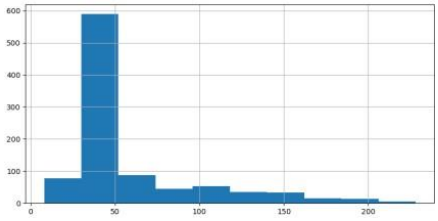


The maximum number of houses have
Their wood deck area has between 90 and 140 sq feet

The sale price is maximum for houses with wood deck area between 100 to 150 sq feet

Open Porch

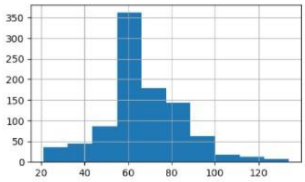


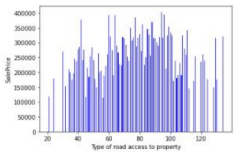


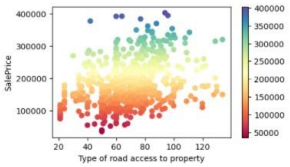
1.Maximum number of houses have open porch in the area between 30 and 50 sq feet

2.Maximum sales price is for houses with Open porch with area between 40 to 60. But otherwise it has little impact on sales price

Linear feet of street connected to property



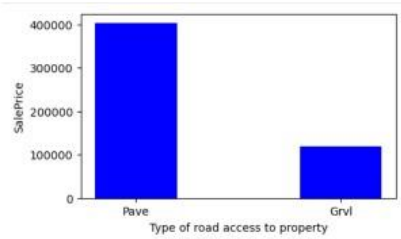


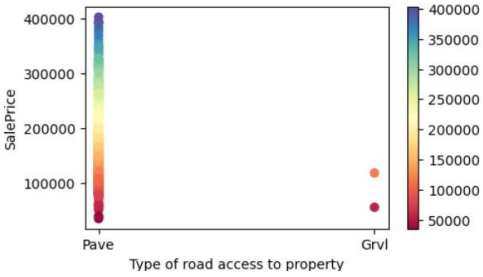


1.When the street is between 40 to 110, it gives sales price of upto 3lakh

2.But when between 60 to 90, it gives above 30lakh and upto 40lakh sales price

Type of road access to property





There is paved and gravel road access

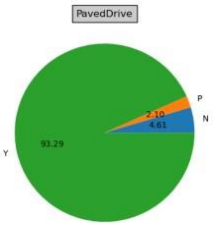
Paved road access is highest among the type of road acess to property and it shows highest sale price

No actual road access

Actual road access

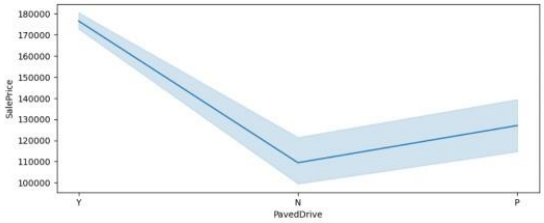
Paved Drive

Y	Paved
P	Partial Pavement
N	Dirt/Gravel



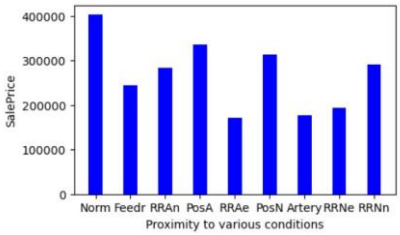
1.Maximum houses had paved drive

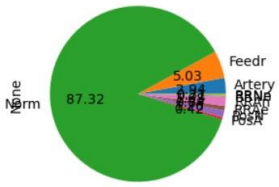
2.Houses with paved drive had maximum Sale price

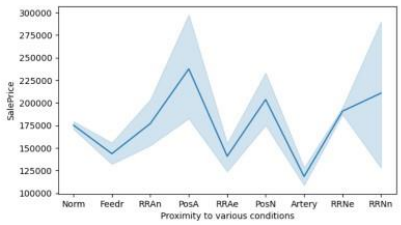


Proximity to various conditions

```
Artery   Adjacent to arterial street
Feedr    Adjacent to feeder street
Norm     Normal
RRNn     Within 200' of North-South Railroad
RRAn     Adjacent to North-South Railroad
PosN     Near positive off-site feature--park, greenbelt,
PosA     Adjacent to postive off-site feature
RRNe     Within 200' of East-West Railroad
RRAe     Adjacent to East-West Railroad
```

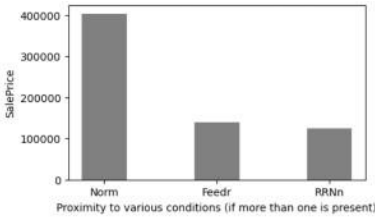


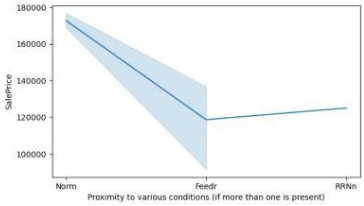




Maximum number of houses are normal and they show highest sale price

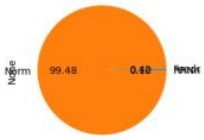
Proximity to various conditions (if more than one is present)





Maximum number of houses are normal and they show highest sale price

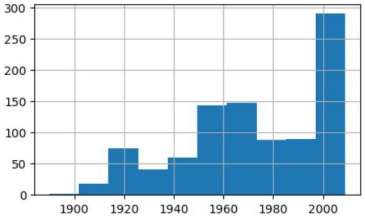
```
Feedr    Adjacent to feeder street
Norm     Normal
RRNn     Within 200' of North-South Railroad
```

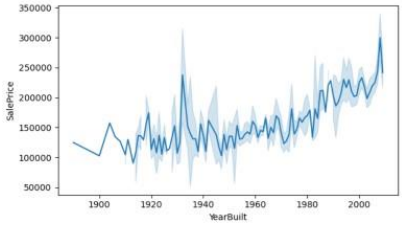


Construction details

- Material, finish
- construction date and Remodel date
- Exterior,roof, foundation and masonry

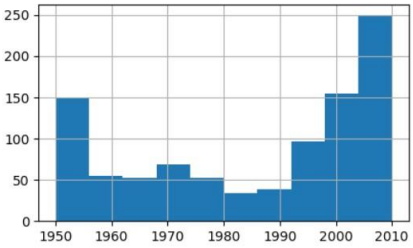
YearBuilt

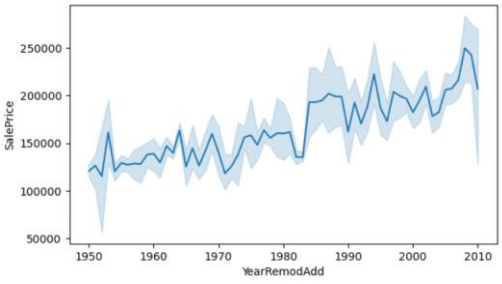




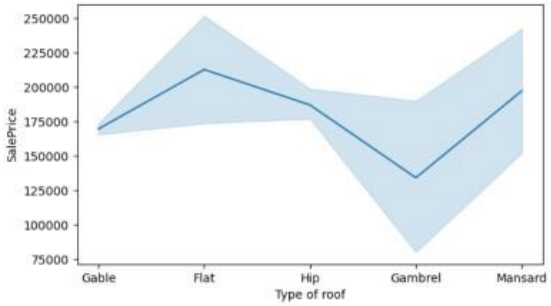
Sale price increases when the house is more New and ,most of the house were built in 2000

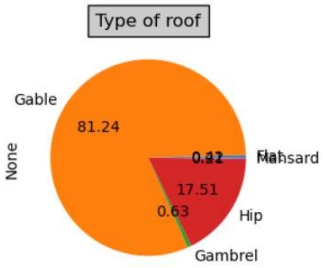
YearRemodAdd





The sooner the remodelling was done, the higher the prices

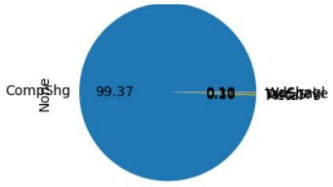


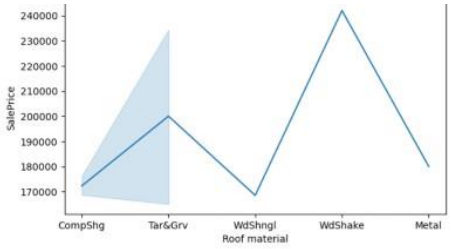


Maximum roofs are gable and hip, and they have given the highest sale price

Roof material

```
ClyTile  Clay or Tile
CompShg  Standard (Composite) Shingle
Membran  Membrane
Metal    Metal
Roll     Roll
Tar&Grv  Gravel & Tar
WdShake  Wood Shakes
WdShngl  Wood Shingles
```

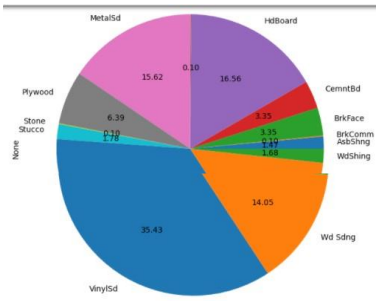


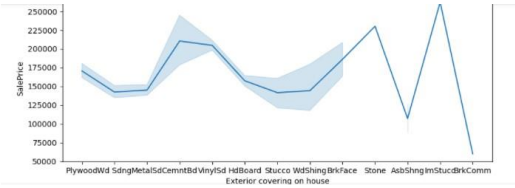


Maximum roofs are made of Standard Composite Shingle

Both standard composite shingle and Wood shakes show high price

Exterior covering on house

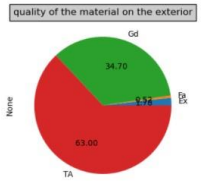


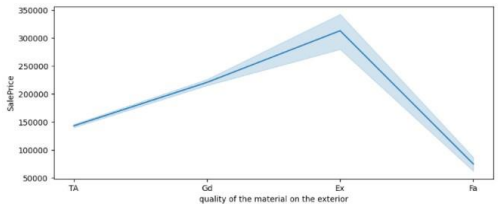


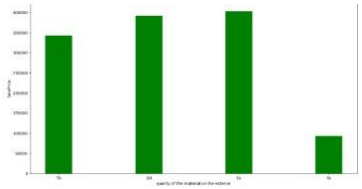
Vinyl siding has highest sales price

quality of the material on the exterior

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



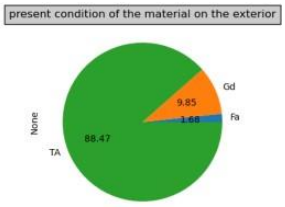


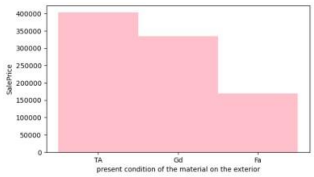


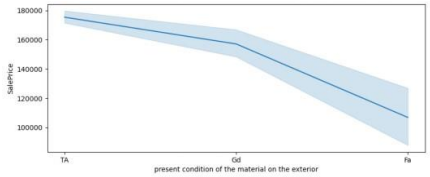
Maximum number of houses are average, but maximum sale price is for Excellent houses

present condition of the material on the exterior

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



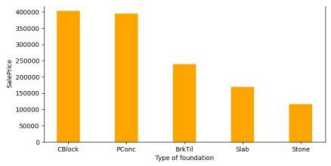




Maximum houses are having average rating for the condition of the exterior materials, but maximum price is obtained for good ratings of the exterior materials condition

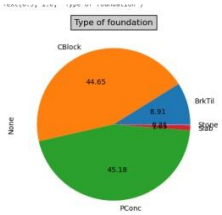
Type of foundation

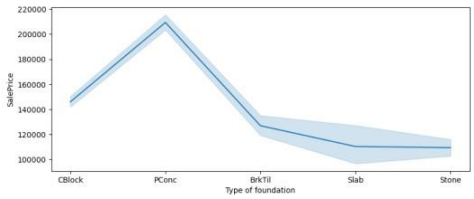
BrkTil	Brick & Tile
CBlock	Cinder Block
PConc	Poured Contrete
Slab	Slab
Stone	Stone
Wood	Wood



1.Maximum foundations are of cinder block and poured concrete

2.Maximum sale price is for poured concrete



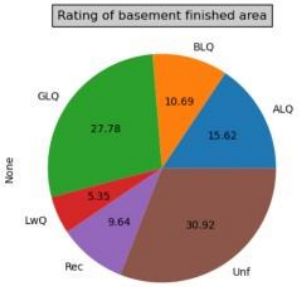


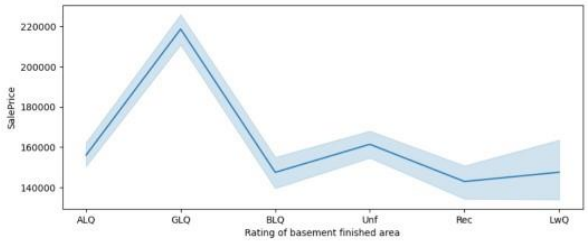
Basement

Rating of basement finished area

Rating of basement finished area (if multiple types)

Rating of basement finished area

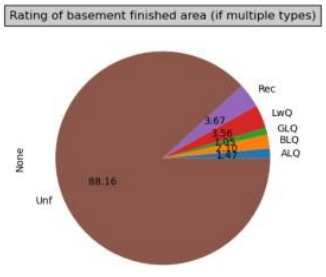


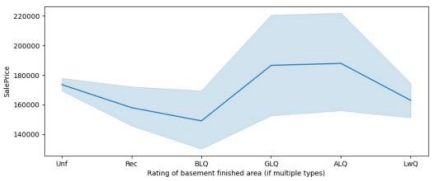


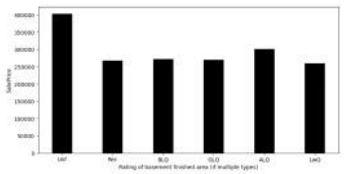
GLQ	Good Living Quarters
ALQ	Average Living Quarters
BLQ	Below Average Living Quarters
Rec	Average Rec Room
LwQ	Low Quality
Unf	Unfinshed
NA	No Basement

Maximum number of houses have basements unfinished but the maximum sale price is for the houses which are good living quaters

Rating of basement finished area (if multiple types)





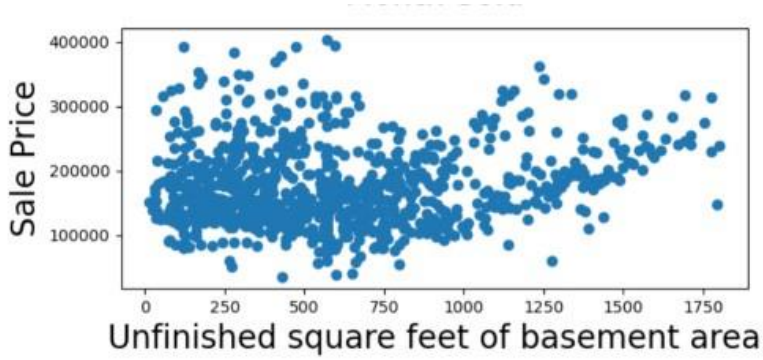


GLQ	Good Living Quarters
ALQ	Average Living Quarters
BLQ	Below Average Living Quarters
Rec	Average Rec Room
LwQ	Low Quality
Unf	Unfinshed
NA	No Basement

1.Maximum houses have their basements unfinished

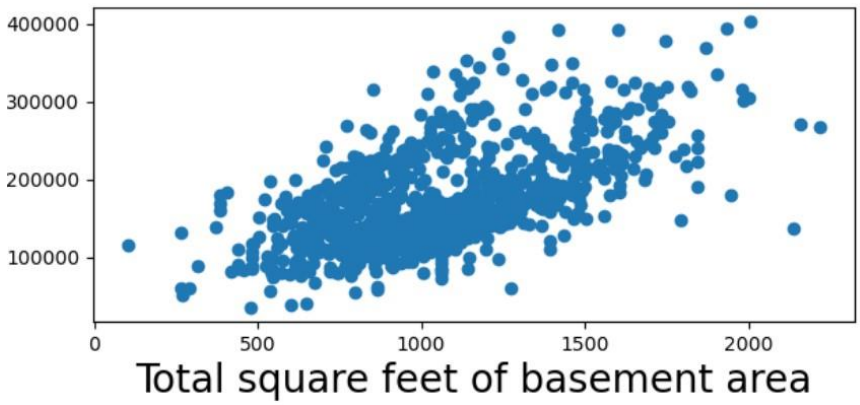
2.Sale price is maximum for good living quaters and average living quaters

Unfinished square feet of basement area



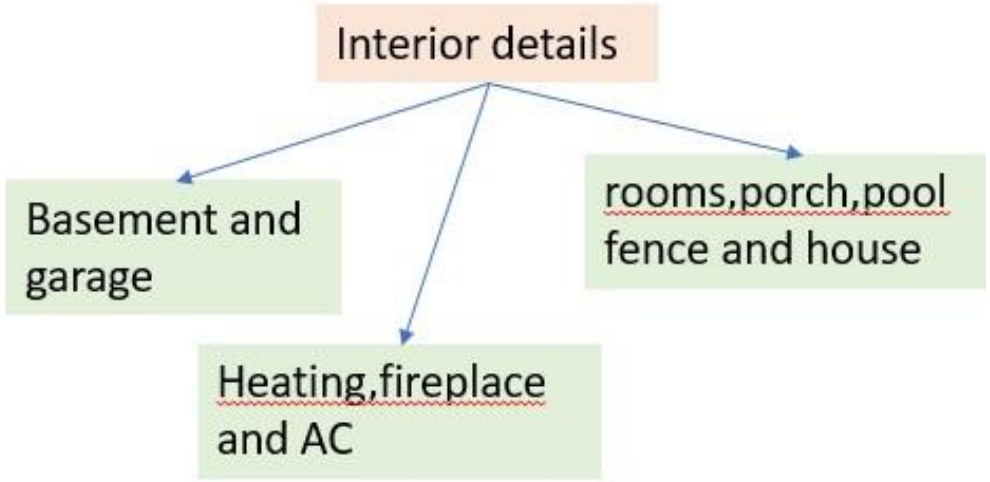
Sale price increases when unfinished basement is between 150 and 200 sq feet

Total square feet of basement area



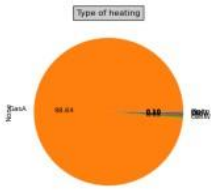
Sale price is highest when the total square feet of basement is higher

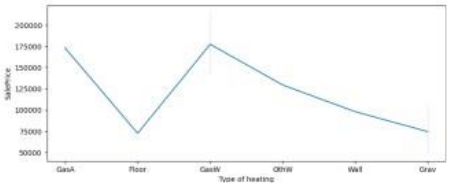
INTERIOR DETAILS

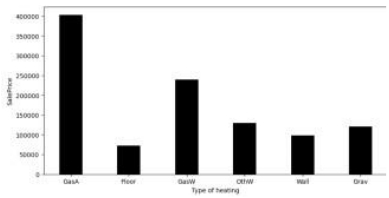


Type of heating

Floor	Floor Furnace
GasA	Gas forced warm air furnace
GasW	Gas hot water or steam heat
Grav	Gravity furnace
OthW	Hot water or steam heat other than gas
Wall	Wall furnace





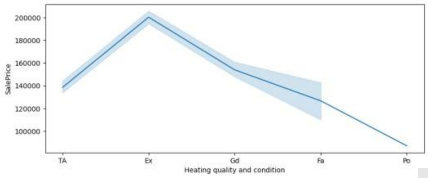


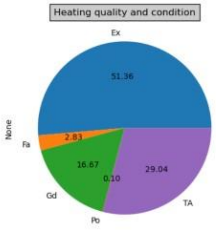
Maximum houses have has forced warm air furnance

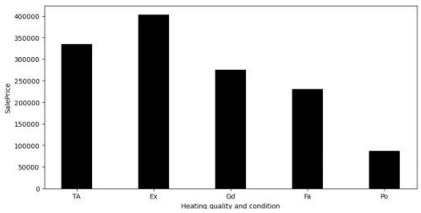
Sale price is highest for houses with forced warm air furnance and Floor furnance

Heating quality and condition

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

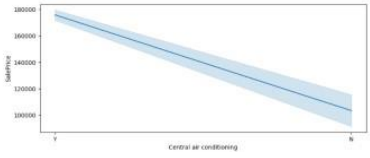


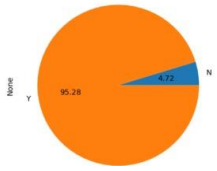


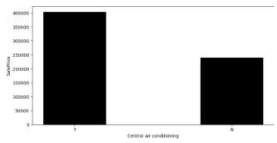


Most of the houses have excellent heating quality condition and they are seen to have maximum sale price

Central air conditioning



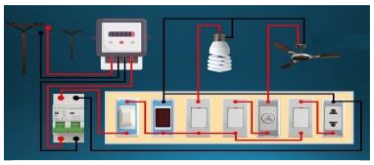




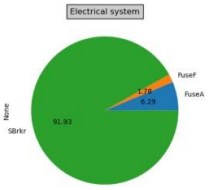
Most of the houses have central AC

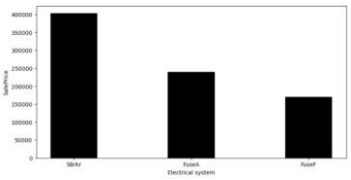
The houses with AC have higher sale price

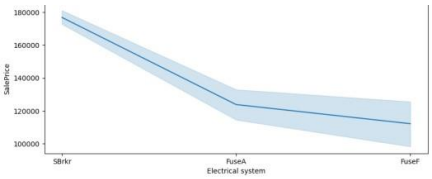
Electrical system



SBrkr	Standard Circuit Breakers & Romex
FuseA	Fuse Box over 60 AMP and all Romex wiring (Average)
FuseF	60 AMP Fuse Box and mostly Romex wiring (Fair)
FuseP	60 AMP Fuse Box and mostly knob & tube wiring (poor)
Mix	Mixed

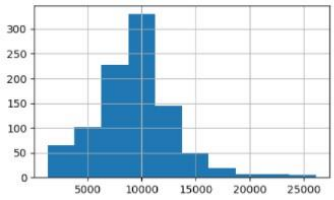


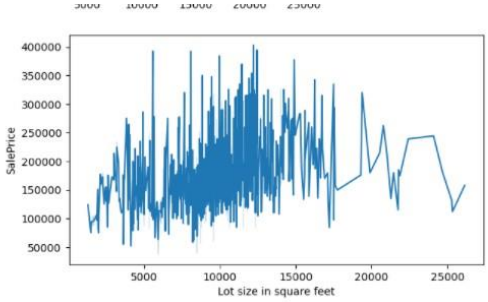


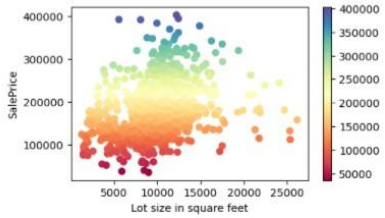


Maximum houses has Standard Circuit Breakers and Romex and houses with it shows maximum Sale price

Plot size in square feet



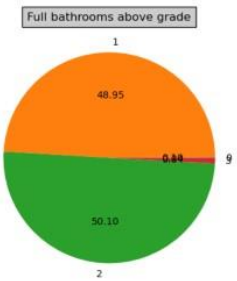


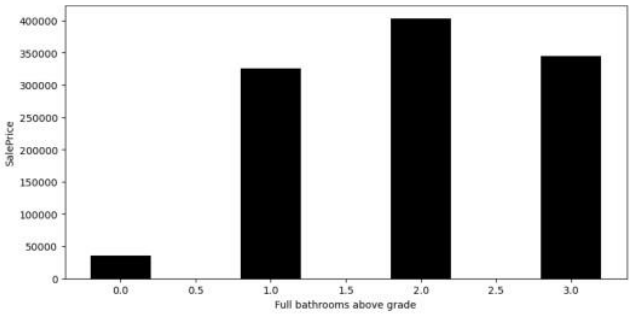


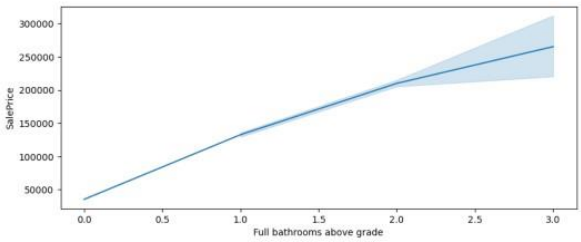
1.More than 300 customers have bought houses with land area of around 10,000 square feet

2.Sale prices are highest between 10,000sq feet and 13,000 sq feet and starts to decrease after it

Full bathrooms above grade

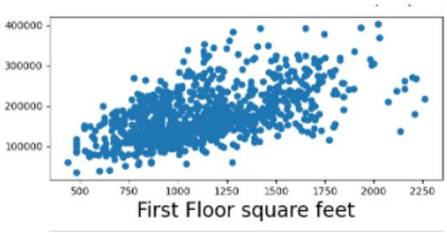




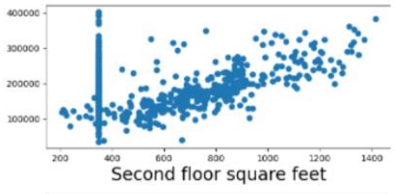


Maximum sale price is for houses with full bathroom above grade as 2

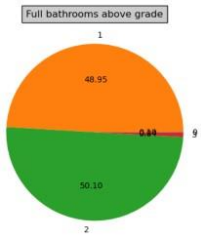
First Floor square feet

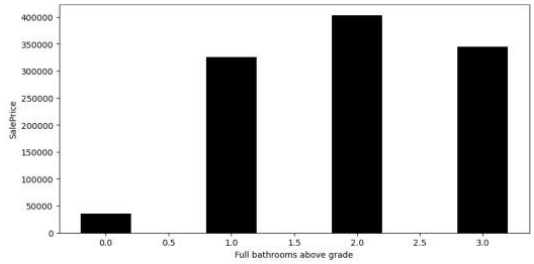


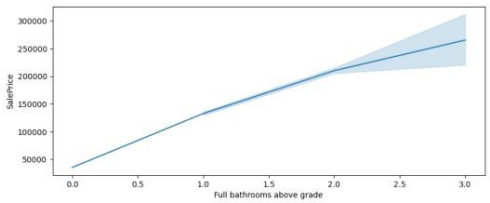
Second floor square feet



Full bathrooms above grade

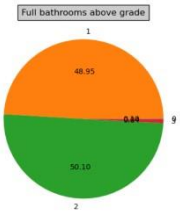


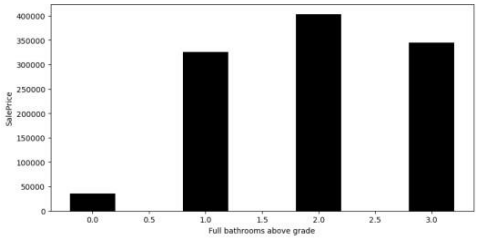


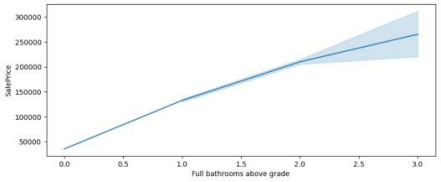


Maximum sale price is for houses with full bathroom above grade as 2

Full bathrooms above grade

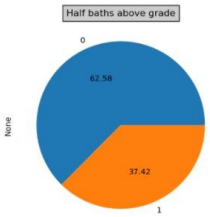


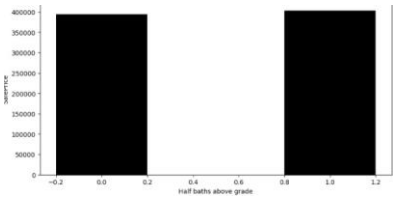


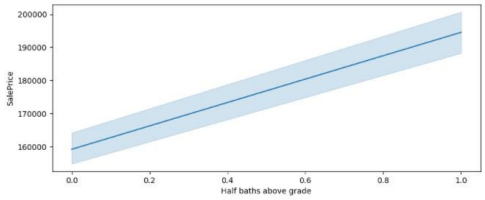


Maximum sale price is for houses with full bathroom above grade as 2

Half baths above grade



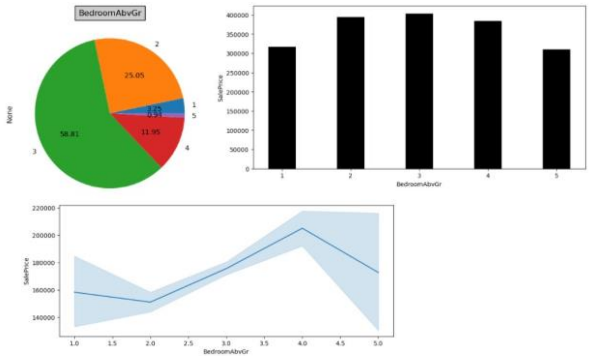




Maximum houses have zero half baths above grade

The maximum sale prices are for houses which have upto 2 houses as above grade

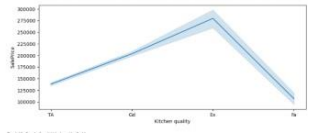
Bedroom Above Ground

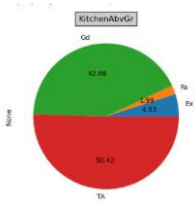


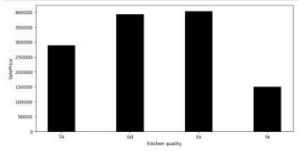
The maximum number of houses have 3 bedrooms above ground

The maximum sale price is for the houses which have 4 bedrooms above ground

Kitchen quality





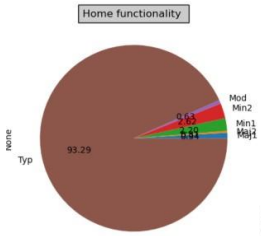


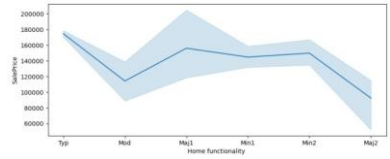
Ex	Excellent
Gd	Good
TA	Typical/Average
Fa	Fair
Po	Poor

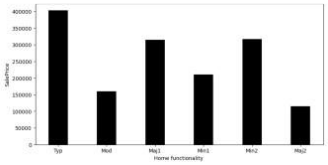
The kitchen quality of most of the houses are average

The highest sale price are for those houses whose kitchen quality is Excellent

Home functionality





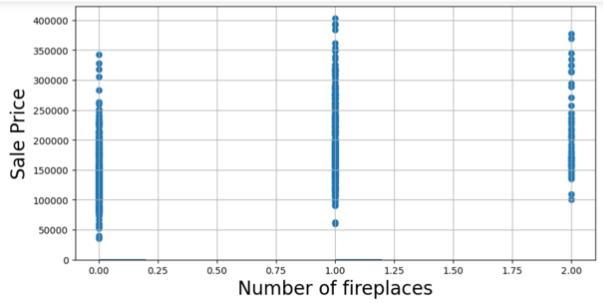


Typ	Typical Functionality
Min1	Minor Deductions 1
Min2	Minor Deductions 2
Mod	Moderate Deductions
Maj1	Major Deductions 1
Maj2	Major Deductions 2
Sev	Severely Damaged
Sal	Salvage only

Fireplaces

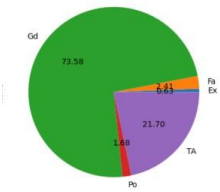
1.Number of fireplace
2.Quality of fireplaces

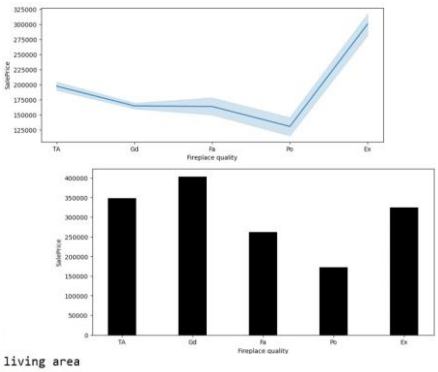
Number of fireplace



More number of houses have 1 fireplace, but houses with both 1 and 2 fireplace show high sale price

Fireplace quality





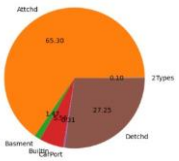
Ex	Excellent - Exceptional Masonry Fireplace
Gd	Good - Masonry Fireplace in main level
TA	Average - Prefabricated Fireplace in main living area
Fa	Fair - Prefabricated Fireplace in basement
Po	Poor - Ben Franklin Stove
NA	No Fireplace

Maximum houses have good fireplace quality, but houses with maximum sale price have excellent quality fireplace

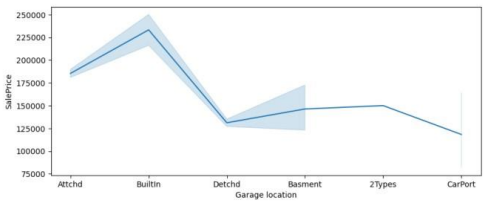
Garage

- 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

Garage location



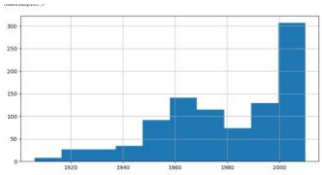
2Types	More than one type of garage
Attchd	Attached to home
Basment	Basement Garage
BuiltIn	Built-In (Garage part of house - typically has room above garage)
CarPort	Car Port
Detchd	Detached from home
NA	No Garage

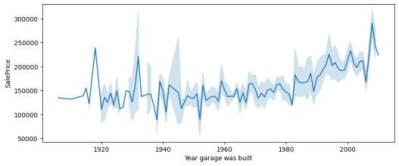


The maximum houses have the attached to home garage

The sale price is highest for Built in garage

Year garage was built

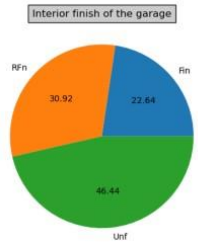




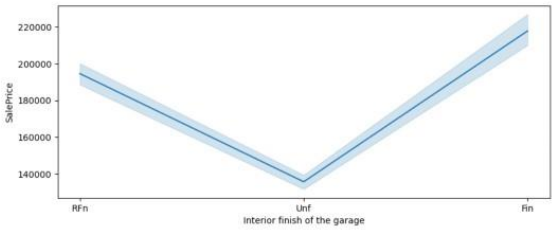
Maximum number of houses are built
After 2000

However the Sale price with increase in number of year after 2010

Interior finish of the garage



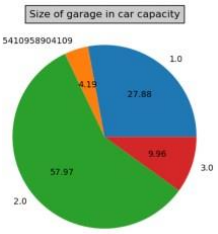
Fin Finished
RFn Rough Finished
Unf Unfinished
NA No Garage

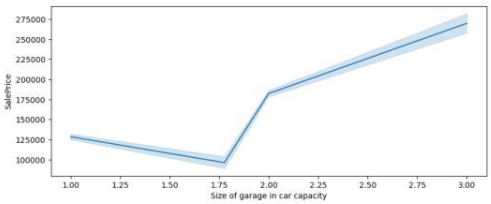


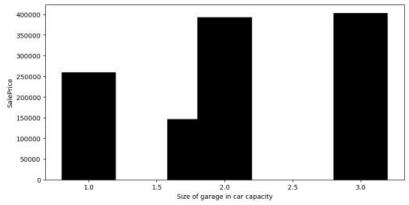
Maximum houses have Unfinished garade

Sale price is highest for houses which have interior of garage finished

Size of garage in car capacity



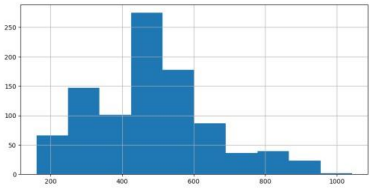




Sales price is highest for houses with size of garage in car capacity as 3

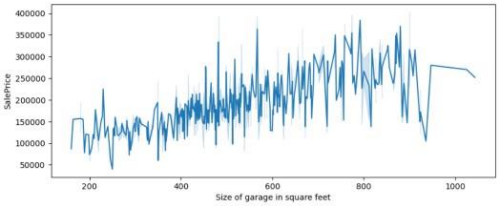
Maximum number of houses have 2 as size of garage in car capaicity

Size of garage in square feet

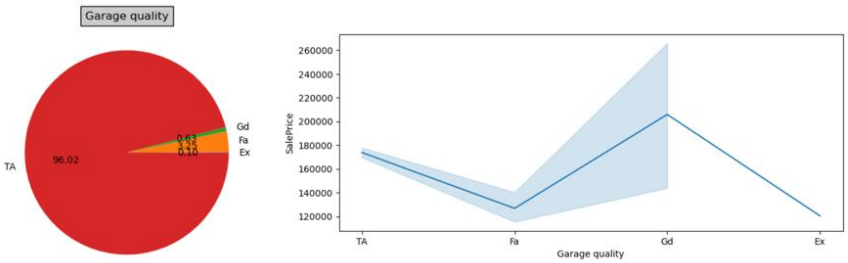


Maximum houses have garage size between 420 to 500square feet

Sales price is maximum for houses which have Size of garage In square feet till 800 square feet, then There is more drop in prices



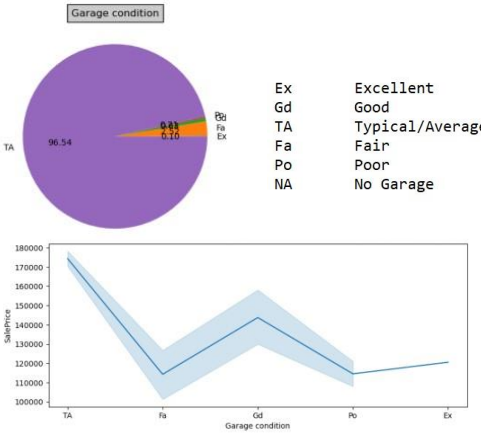
Garage quality



Ex	Excellent
Gd	Good
TA	Typical/Average
Fa	Fair
Po	Poor
NA	No Garage

1.

Garage condition



Ex	Excellent
Gd	Good
TA	Typical/Average
Fa	Fair
Po	Poor
NA	No Garage

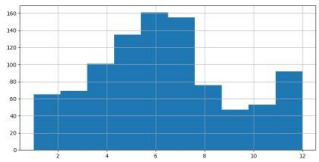
Maximum houses have Average garage
Condition

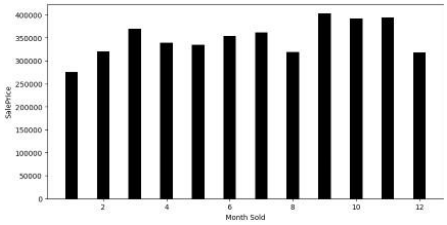
The sale price was highest for average
Condigtion garage, hence we can see the garage condition did not play much role in the sale price

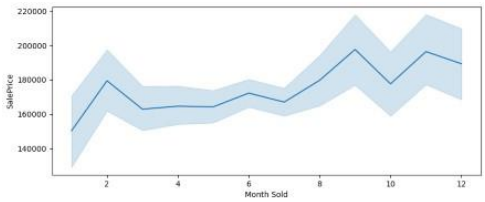
Sale details

- Month Sold
- Year Sold
- Type of sale
- Condition of sale

Month Sold



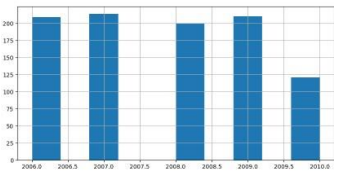


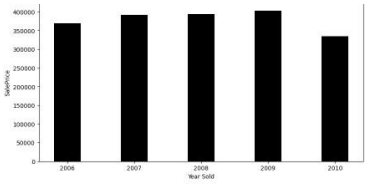


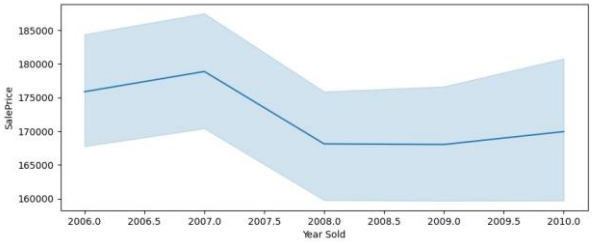
Maximum houses have been sold during May,June and July

Sales price is maximum during September and November

Year Sold



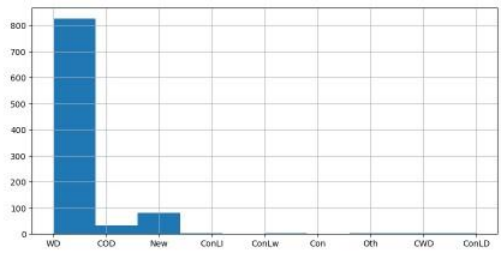


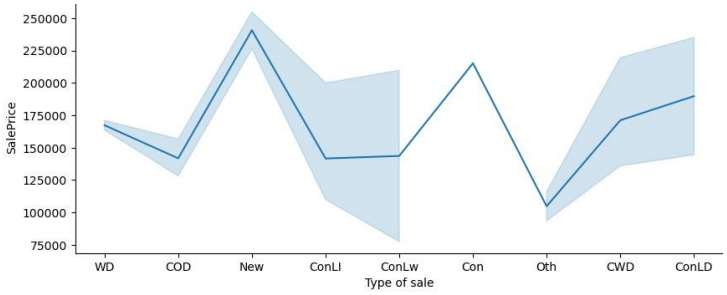


Houses were not sold in all years,
But the least was in 2010

The maximum sales price is obtained for houses Sold in 2007

Type of sale



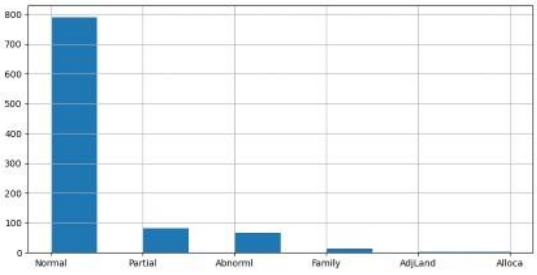


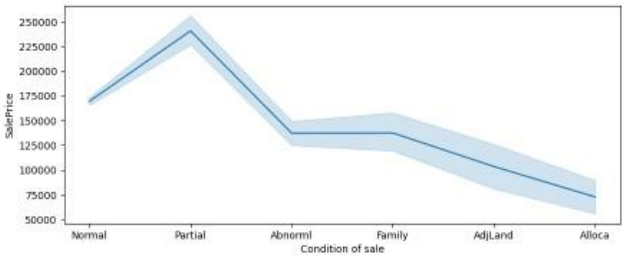
WD	Warranty Deed - Conventional
CWD	Warranty Deed - Cash
VWD	Warranty Deed - VA Loan
New	Home just constructed and sold
COD	Court Officer Deed/Estate
Con	Contract 15% Down payment regular terms
ConLw	Contract Low Down payment and low interest
ConLI	Contract Low Interest
ConLD	Contract Low Down
Oth	Other

1.Maximum houses were sold as warranty Deed-conventional

2.Maximum sales price is seen for home just constructed and sold

Condition of sale





Normal	Normal Sale
Abnorml	Abnormal Sale - trade, foreclosure, short sale
AdjLand	Adjoining Land Purchase
Alloca	Allocation - two linked properties with separate deeds, typically condo with a garage unit
Family	Sale between family members
Partial	Home was not completed when last assessed (associated with New Homes)

Maximum houses condition of sale was normal

Houses sold in maximum price had condition of sale as partial, that is home was not completed when lastassesed

CONCLUSION

1.Key Findings and Conclusions of the Study

This project has built a model that can detect Sale Prices of House. In doing so, the model can reduces loses for companies in Investment. The challenge behind Sale Price finding in machine learning is the number of data in dataset.Also some other issues are the excess number of zeroes and null values in the data.

Five different regressor Linear Regression,Decision Tree regressor and Random forest Regressor. The

Inferences from the Problem are:

Type of dwelling	1.The 1-STORY 1946 & NEWER ALL STYLES is highest among the types of dwellings 2.The Sale price is highest in 2-STORY 1946 & NEWER
general zoning classification	1.Residential Low density is highest among the general zone classification 2.Residential Low Density and Residential Medium Density gives highest sale price
Linear feet of street connected to property	1.When the street is between 40 to 110, it gives sales price of upto 3lakh 2.But when between 60 to 90, it gives above 30lakh and upto 40lakh sales price
Lot size in square feet	1.More than 300 customers have bought houses with land area of around 10,000 square feet 2.Sale prices are highest between 10,000sq feet and 13,000 sq feet and starts to decrease after it
Type of road access to property	Pave is highest among the type of road acess to property and it shows highest sale price

General shape of property	1.Maximum houses have regular shape 2.But the houses with slightly irregular shape shows the highest sales price
Flatness of the property	1.Most of the property was levelled 2.Property which were hillside with significant slope from side to side had highest sale price
Lot configuration	1.Highest is Inside lot 2.Sale price is highest for Cul de sac
Slope of property	The highets number of properties have gentler slope and show higher sales price
Proximity to various conditions	Maximum number of houses are normal and they show highest sale price
Proximity to various conditions (if more than one is present)	Maximum number of houses are normal and they show highest sale price
Type of dwelling	1.Maximum houses are single family detached 2.Maximum sale price is for Townhouse End Unit
Style of dwelling	Maximum houses are 1 story but sale price of 2 storey is highest
Overall Quality	Maximum houses are above average With increase in quality the price of the house increases
Overall Condition	Most of the houses are average conditions. The prices increase upto average, then the price starts dropping towards very excellent
Year house was built	Sale price increases when the house is more New
Year house was remodelled	The sooner the remodelling was done, the higher the prices
Type of roof	Maximum roofs are gable and hip, and they have given the highest sale price

Roof material	Maximum roofs are made of Standard Composite Shinglea Both standard composite shingle and Wood shakes show high price
Exterior covering on house	Vinyl siding has highest sales price
Exterior covering on house (if more than one material)	Largest number of houses have Vinyl but the highest price is for Cement board
quality of the material on the exterior	Maximum number of houses are average, but maximum sale price is for Excellent houses
present condition of the material on the exterior	Maximum houses are having average rating for the condition of the exterior materials, but maximum price is obtained for good ratings of the exterior materials condition
Type of foundation	1.Maximum foundations are of cinder block and poured concrete 2.Maximum sale price is for poured concrete
walkout or garden level walls	Maximum houses have no exposure to walkouts or gardens, but maximum sale price is for good exposure to it
Masonry veneer area in square feet	maximum sale price is between 100 to 400 square feet
Rating of basement finished area	Maximum number of houses have basements unfinished, but the maximum sale price is for the houses which are good living quaters
Rating of basement	1.Maximum houses have their basements unfinished

finished area (if multiple types)	2.Sale price is maximum for good living quaters and average living quaters
Type 1 finished square feet	Type 1 finished square feet increases, the sale price increases
Unfinished square feet of basement area	the sale prices decreases as it increases
Total square feet of basement area	the sale price increases as it increases
First Floor square feet	As it increases, the sale price increases
Second floor square feet	the sale price increases as it increases
Above grade (ground) living area square feet	the sale price increases as it increases
Size of garage in square feet	the sale price increases as it increases
Wood deck area	doesnt have much impact on sale price and maximum houses have 100 to 250 wood deck area
Open porch area in square feet	doesnt have much impact on sale price

2.Limitations of the problem

- The number of datas of house is less
- the excess number of zeroes and null values in the data.
- there can also be seen high collinearity problem.