FLIP ROBO

HOUSING: PRICE PREDICTION

Submitted by:

John Tojo

Data Science Intern

ACKNOWLEDGMENT

It gives me immense pleasure to deliver this report. Working on this project was a great learning experience that helped me attain in-depth knowledge on data analysis process.

Flip Robo Technologies (Bangalore) provided all of the necessary information and datasets, required for the completion of the project.

I express my gratitude to my SME, Gulshana Chaudhary, for providing the dataset and directions for carrying out the project report procedure.

My heartfelt gratitude to DataTrained institute and FlipRobo company for providing me this internship opportunity. Last but not least to my sincere thanks to my family and all those who helped me directly or indirectly in completion this project.

INTRODUCTION

- ## Business Problem Framing

  a. Housing and real estate market is one of the market which contributes to economy. Large no. of companies work in this domain and has become highly competitive to provide their customers the house that satisfies them within their budget without wasting the time of customers and the company

  b. By analysing the data, it is possible to predict the customers requirements, the price they can afford the house, by doing so helps the company to develop marketing strategies based on the trend on the current environment

  c. This helps to increase the sales, generate more revenue and provide a competitive edge over other companies.

- ## Conceptual Background of the Domain Problem

  a. 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.

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

  c. Build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know:

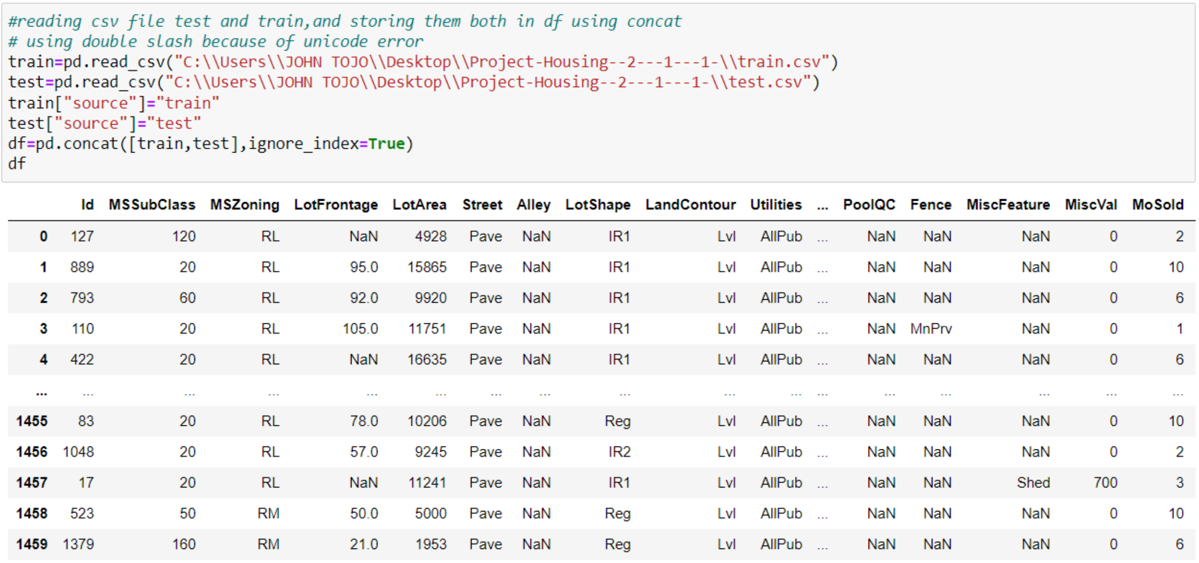
  - Which variables are important to predict the price of variable?
  - How do these variables describe the price of the house?

- ## Motivation for the Problem Undertaken

  Large amount of money and time has to be spend by the customers to select house that suits them, even then the customers are still concerned if the deal was beneficial for them or not. Customers are not aware of the base price of the house and if they are getting charged more than base price. Customers are not aware of the factors that contribute towards deciding the price of the house. By building a model it becomes easier to predict the price and map houses that suit our taste hence reducing the time to look for house and the sense of doubt is reduced.

Analytical Problem Framing

- ## Data Sources and their formats



```
#reading csv file test and train,and storing them both in df using concat
# using double slash because of unicode error
train=pd.read_csv("C:\\Users\\JOHN TOJO\\Desktop\\Project-Housing--2---1---1-\\train.csv")
test=pd.read_csv("C:\\Users\\JOHN TOJO\\Desktop\\Project-Housing--2---1---1-\\test.csv")
train["source"]="train"
test["source"]="test"
df=pd.concat([train,test],ignore_index=True)
df
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolQC	Fence	MiscFeature	MiscVal	MoSold
0	127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	...	NaN	NaN	NaN	0	2
1	889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	...	NaN	NaN	NaN	0	10
2	793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	...	NaN	NaN	NaN	0	6
3	110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	...	NaN	MnPrv	NaN	0	1
4	422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	...	NaN	NaN	NaN	0	6
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1455	83	20	RL	78.0	10206	Pave	NaN	Reg	Lvl	AllPub	...	NaN	NaN	NaN	0	10
1456	1048	20	RL	57.0	9245	Pave	NaN	IR2	Lvl	AllPub	...	NaN	NaN	NaN	0	2
1457	17	20	RL	NaN	11241	Pave	NaN	IR1	Lvl	AllPub	...	NaN	NaN	Shed	700	3
1458	523	50	RM	50.0	5000	Pave	NaN	Reg	Lvl	AllPub	...	NaN	NaN	NaN	0	10
1459	1379	160	RM	21.0	1953	Pave	NaN	Reg	Lvl	AllPub	...	NaN	NaN	NaN	0	6

There were two dataset one for training the model and other for predicting the sales price.

- ## Data Pre-processing Done

1. The two dataset were combined into a single dataframe df
2. The .info() used to get info
3. Check for nulls, drop column with more than 80% null datas
4. .nunique() fuction was used to check for unique data present in each feature, drop column for those whose value are all unique or has only 1 unique value
5. Check duplicates
6. Divide data into three categories
    a. Numerical continuous data
    b. Numerical discrete data
    c. Categorical data
7. Imputation for numerical and categorical data
8. Visualize and statistical interpretation
9. Treat skewness
10. Check correlation and check for multicollinearity
11. Encoding
12. Outliers check for train dataset

- Data Inputs- Logic- Output Relationships

  Scatterplot was used to identify the relation with numerical continuous data and target

  Box plot was used for categorical and numerical discrete to identify the relation with target

- Hardware and Software Requirements and Tools Used

  Software Technology being Used:-

  - Programming language: Python
  - Distribution: Anaconda Navigator
  - Browser based language shell: Jupyter Notebook

  Libraries/Packages Used:-

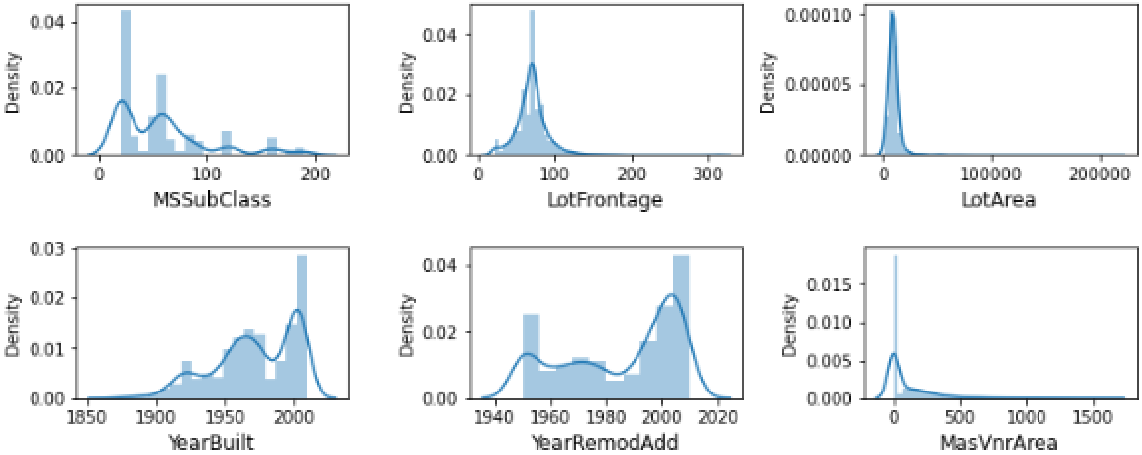
  - Pandas, NumPy, matplotlib, seaborn and scikit-learn

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)
    1. Data Pre-processing Done
        a. The two dataset were combined into a single dataframe df
        b. The .info() function was used to get information regarding the data type of different features used
        c. Nulls were checked and 'Alley', 'PoolQC', 'Fence', 'MiscFeature' columns were deleted as it had more than 80% null datas and imputation were required for those columns which had null datas
        d. .nunique() fuction was used to check for unique data present in each feature, "Id" column was dropped as all its value were unique
        e. Duplicates were checked and no duplicates were found
        f. The data was divided into three categories
            i. Numerical continuous data
            ii. Numerical discrete data
            iii. Categorical data
        g. Imputation for numerical data was done by using mean value and for categorical data imputation was done using mode

b) Visualizing

i. Univariate analysis numerical continuous data



- MSSubClass
  - not uniformly distributed
  - skewed positively
  - presence of outliers

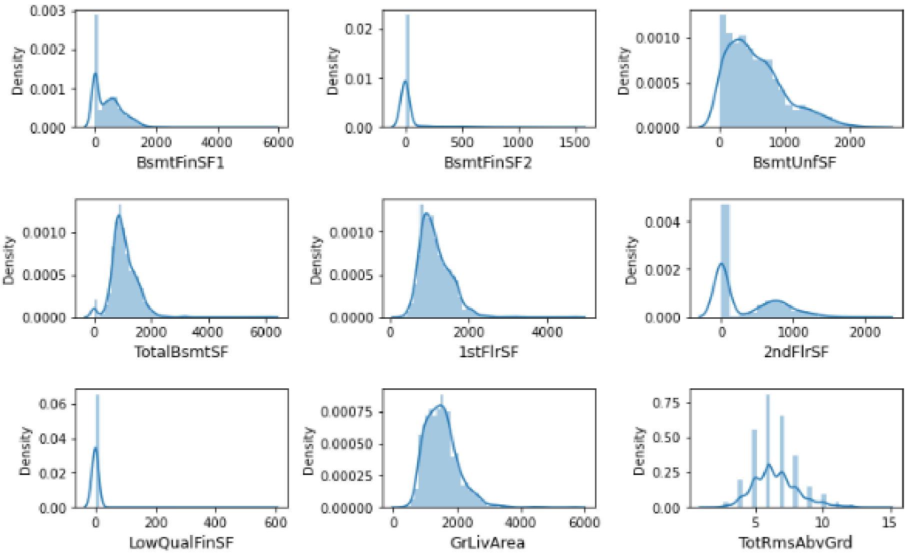
- LotFrontage
  - appears to be uniformly distributed
  - skewed positively

- LotArea
  - appears to be uniformly distributed

- YearBuilt
  - not uniformly distributed
  - negatively skewed

- YearRemodAdd
  - not uniformly distributed
  - negatively skewed

- MasVnrArea
  - not uniformly distributed
  - has high concentration of data at 0
  - skewed positively



- BsmtFinSF1
  - not uniformly distributed
  - has high concentration of data at 0
  - skewed positively

- BsmtFinSF2
  - not uniformly distributed
  - has high concentration of data at 0

- BsmtUnfSF
  - not uniformly distributed
  - skewed positively
  - presence of outliers

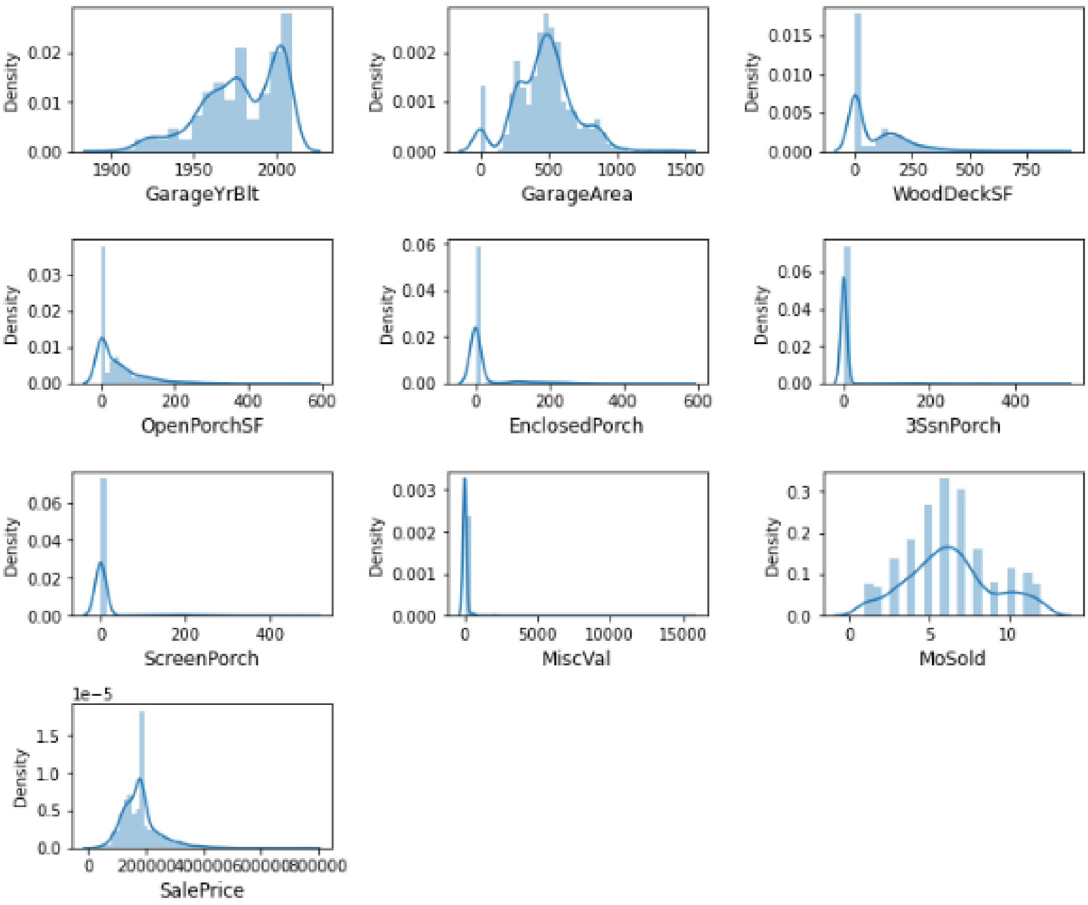
- TotalBsmtSF
  - not uniformly distributed
  - skewed positively
  - presence of outliers
  - more than one mode

- 1stFlrSF
  - not uniformly distributed
  - skewed positively
  - presence of outliers

- 2ndFlrSF
  - not uniformly distributed
  - presence of outliers
  - more than one mode

- GrLivArea
  - not uniformly distributed
  - skewed positively
  - presence of outliers

- TotRmsAbvGrd
  - not uniformly distributed



- GarageYrBlt
  - not uniformly distributed
  - skewed negitively
  - presence of outliers

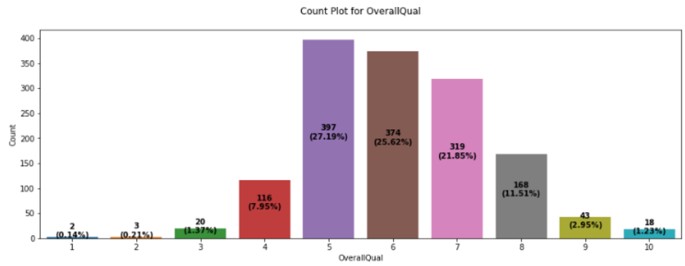
- GarageArea
  - not uniformly distributed
  - presence of outliers
  - more than one mode

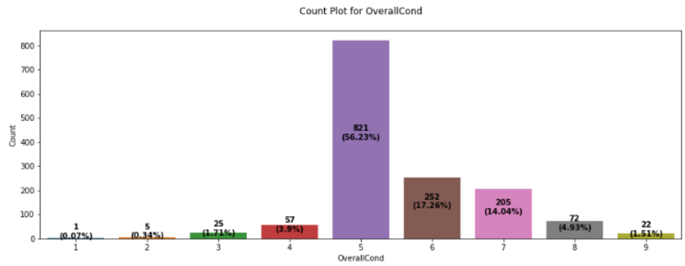
- WoodDeckSF
  - not uniformly distributed
  - skewed positively
  - presence of outliers

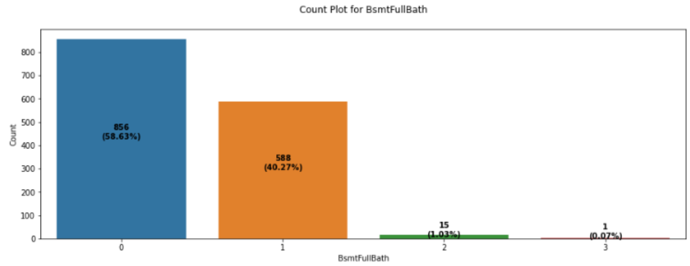
- OpenPorchSF
  - not uniformly distributed
  - skewed positively
  - presence of outliers

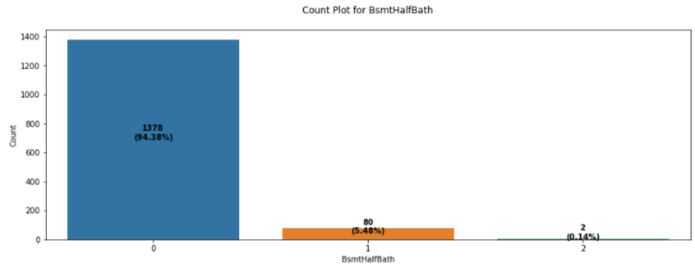
- MoSold
  - not uniformly distributed
  - skewed negatively
  - presence of outliers

ii. univariate analysis numerical discrete data







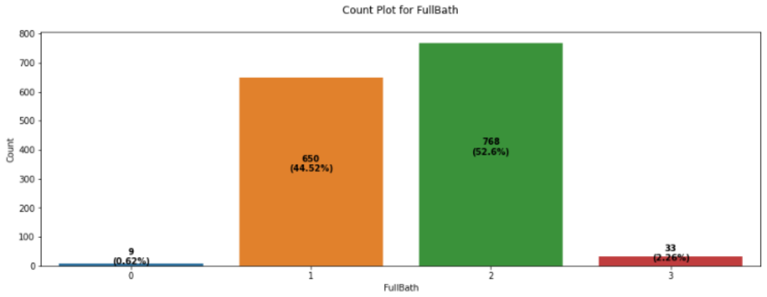


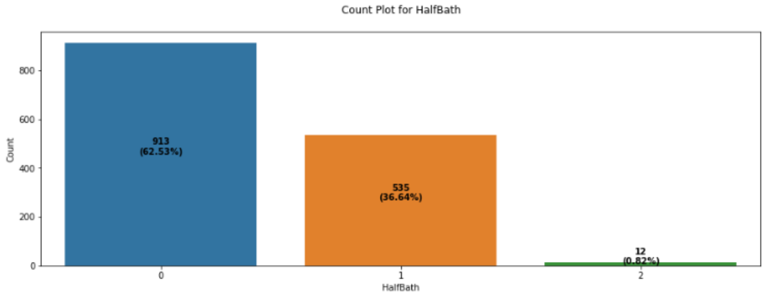
- OverallQual
  - the avg quality of the houses is around 6
  - most of the houses have rating 5,6 and 7
  - some houses have been given rating 1 and few rating 10
  - the houses quality needs to be improved

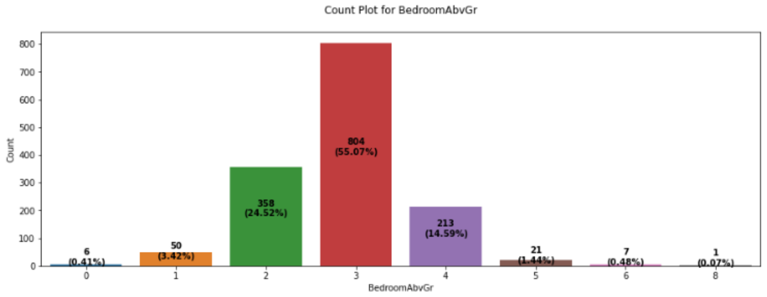
- OverallCond
  - the avg condition of house is 5.5
  - max ratng given is 9 and minimum is 1
  - the condition of the house needs to be improved

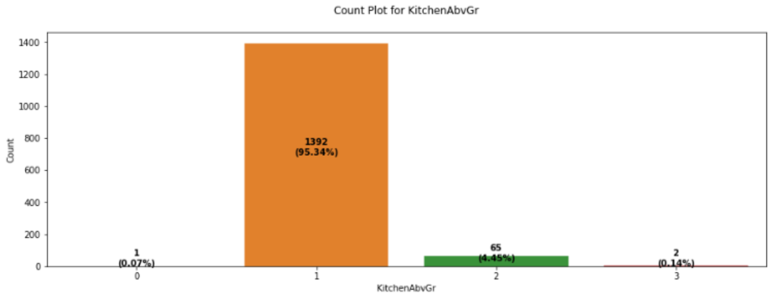
- BsmtFullBath
  - most of the houses doesnt have basement full bath
  - very few houses have 1 basement full bath

- BsmtHalfBath
  - most of the houses doesnt have basement half bath







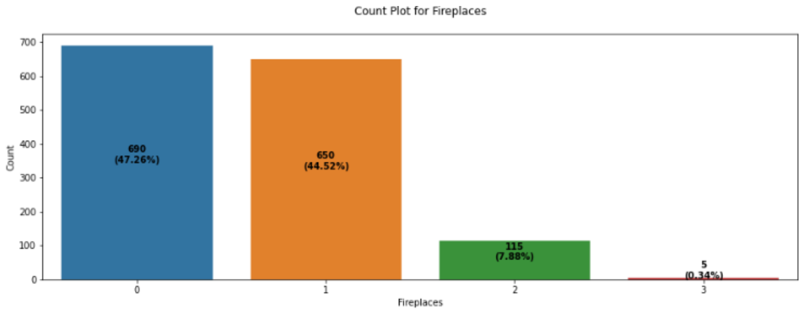


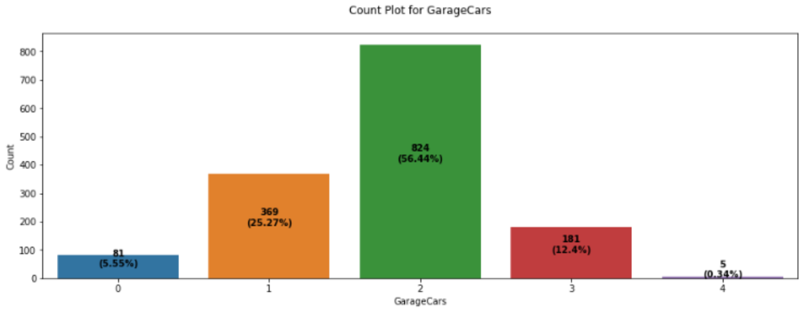
- FullBath
  - most of the houses have 2 Full bathrooms above grade, followed by 1

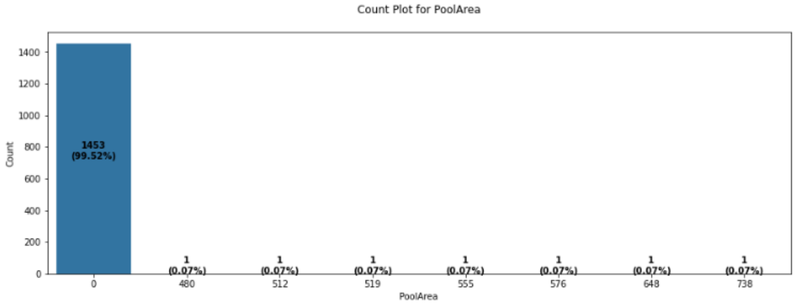
- HalfBath
  - most of the houses have 0 Full bathrooms above grade

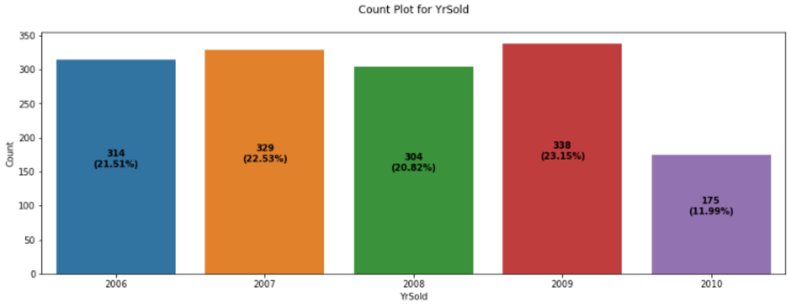
- BedroomAbvGr
  - most of the houses atleast 3 bedroom above grade, followed by 2 and 4 bedrooms above grade

- KitchenAbvGr
  - most of the houses atleast 1 kitchen above grade









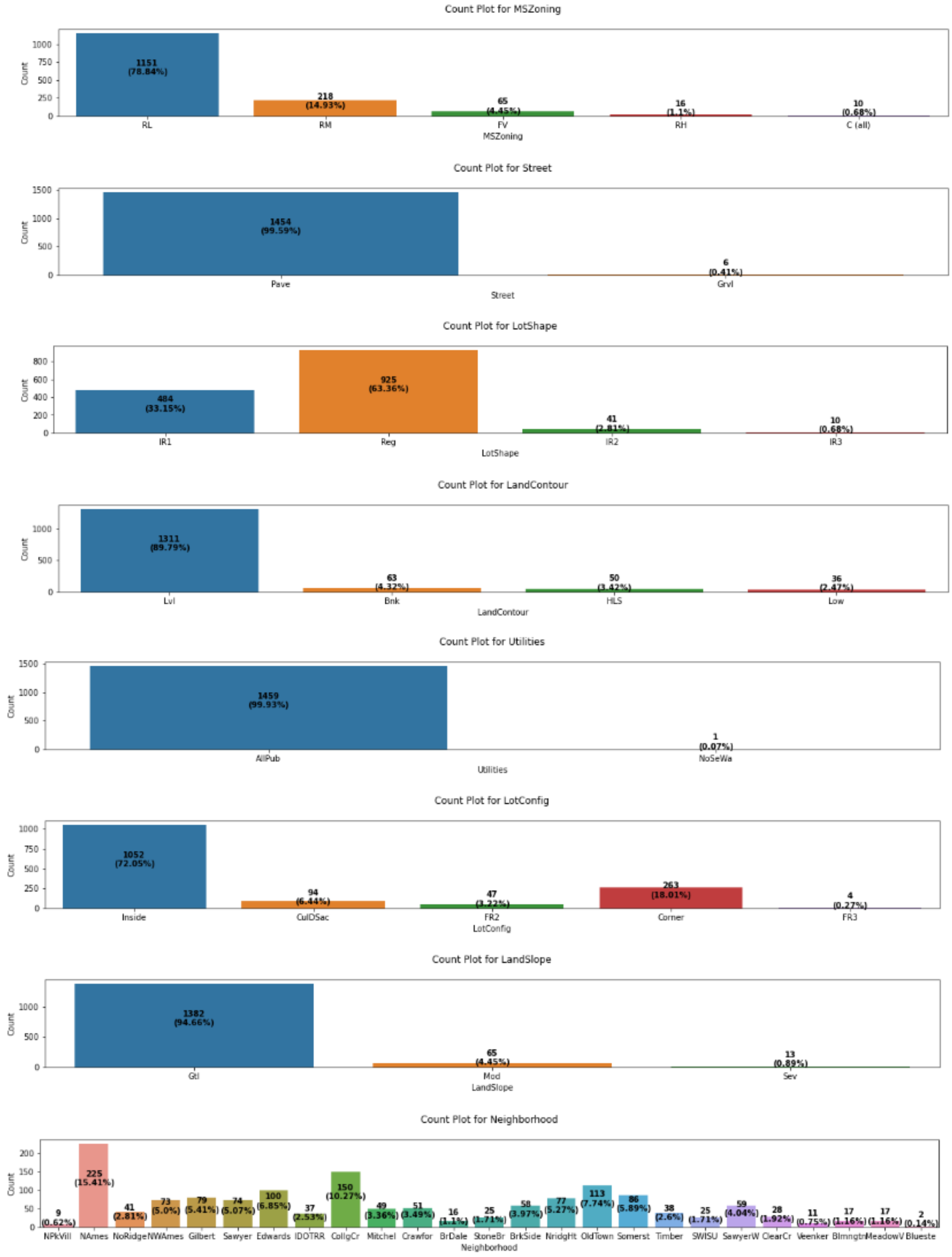
- Fireplaces
  - most of houses doesn thave fireplace
  - few houses have 1 fireplace

- GarageCars
  - most of the houses have space for 2 cars

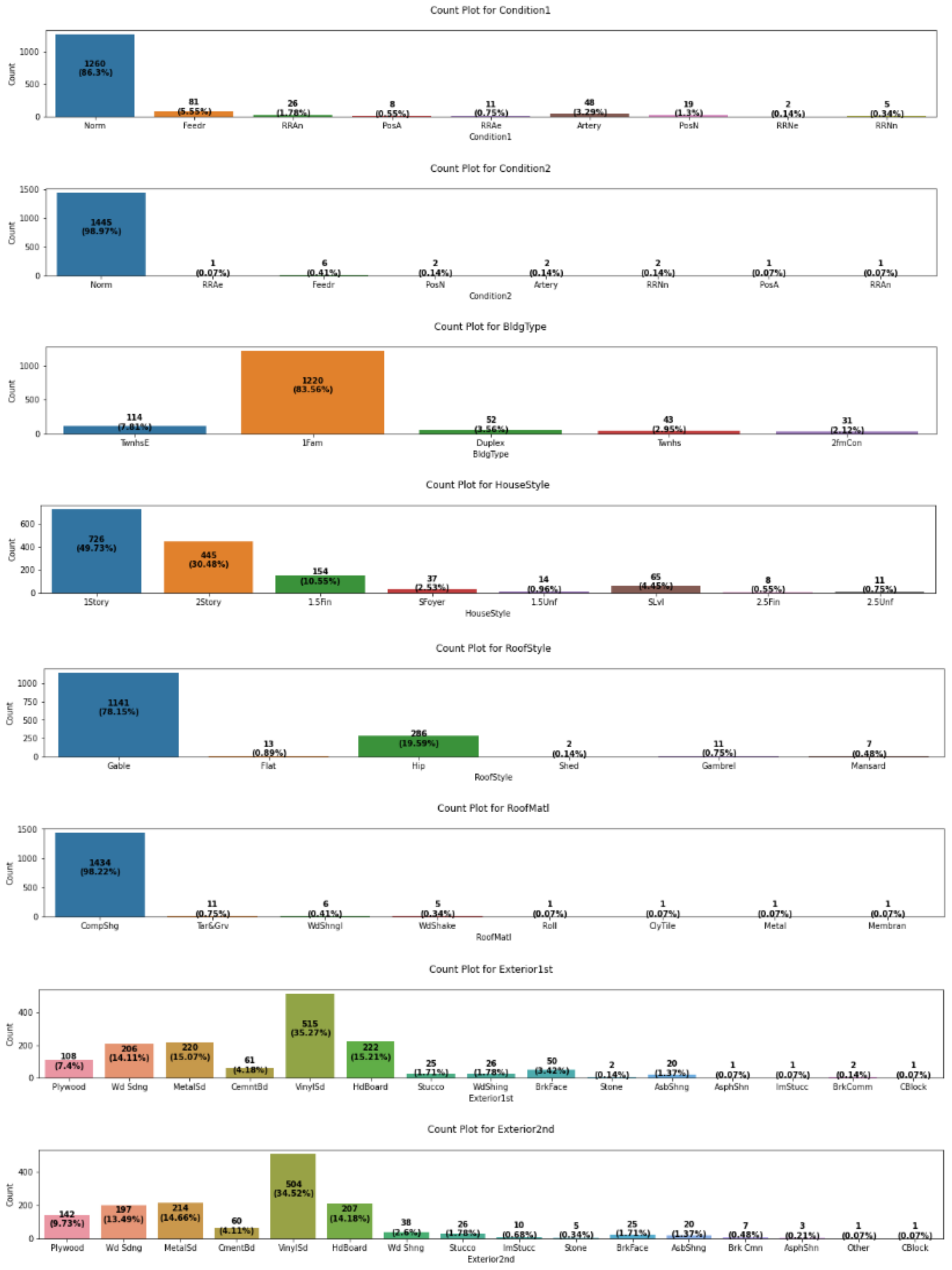
- PoolArea
  - most of the houses doesnt have a pool areas

- YrSold
  - the houses considered were sold from 2007 to 2010

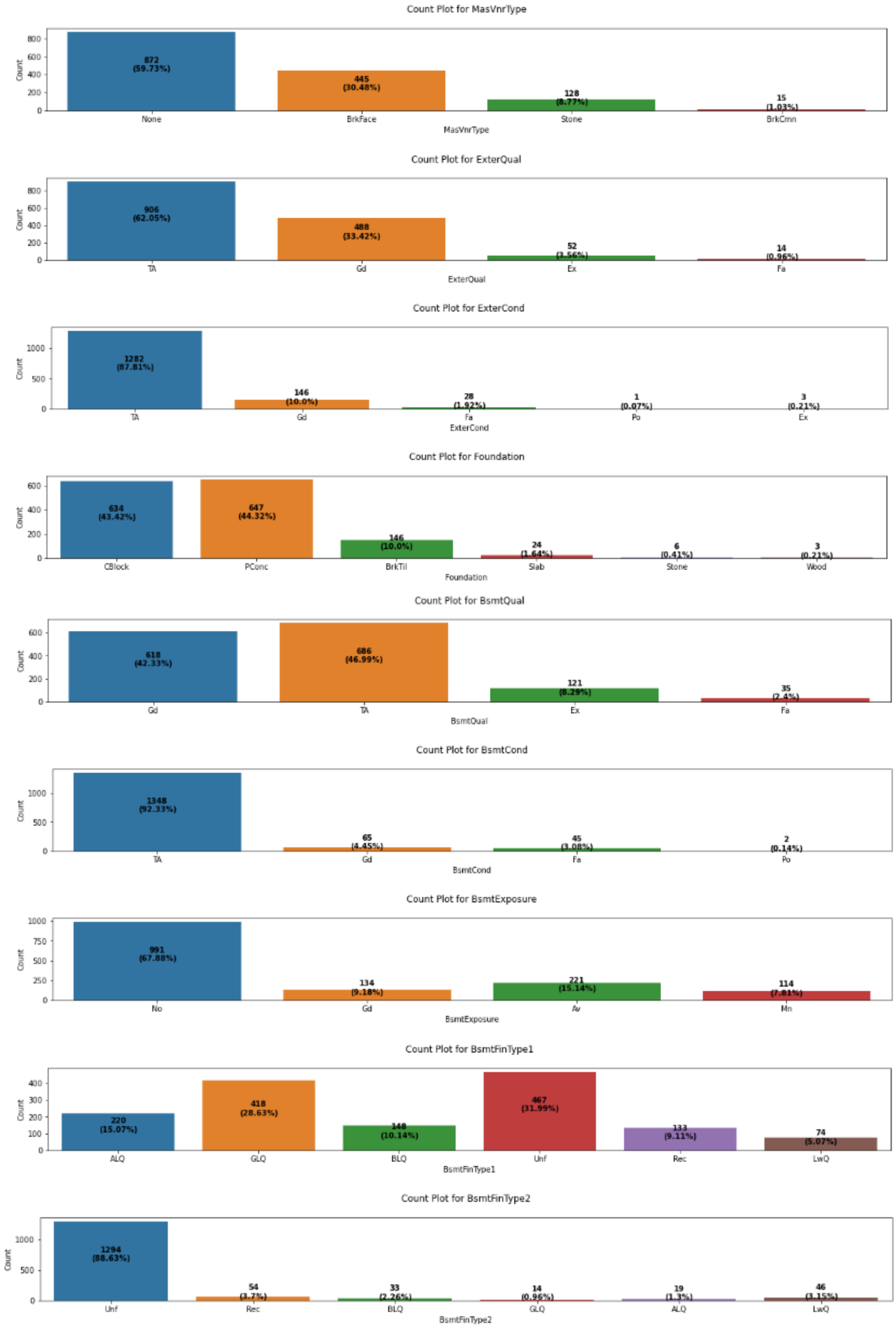
iii. univariate analysis categorical data



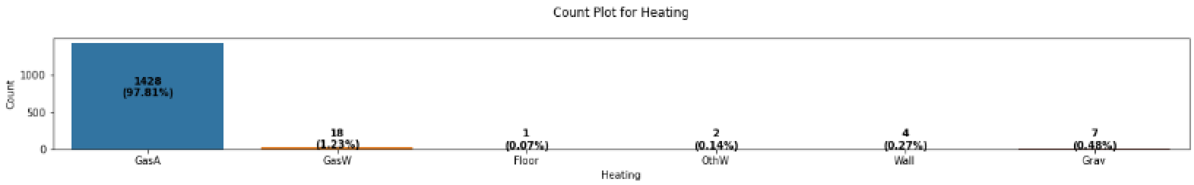
- RL is the most prefered, followed by RM as MSZoning
- most of the people like the street to be Paved
- Regular is the most prefered lot shape
- most of the people prefered level or near flat as their land contour
- most of the residents require all public utilities available
- residents prefer gentle slope for their land
- most of the residents who took part in the survey were from NAmes folowed by CollgCr

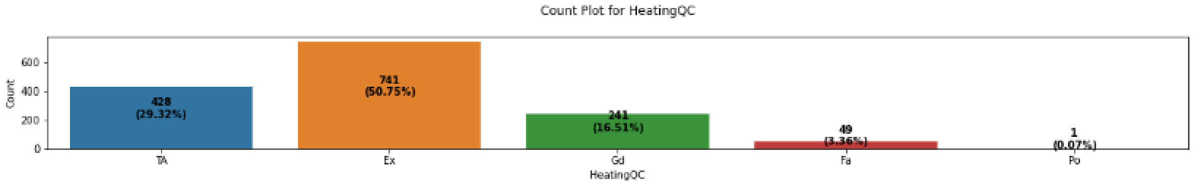


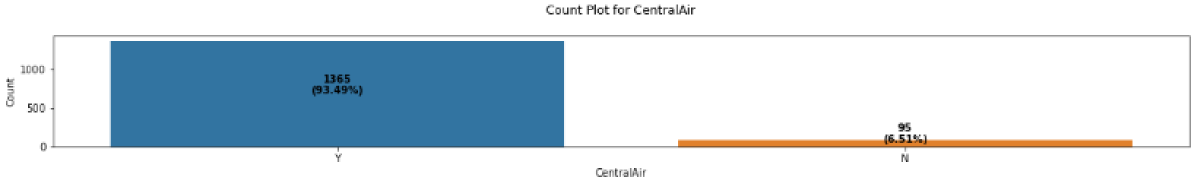
- most of the residents prefer condition 1 and condition 2 as normal
- most of the resdents prefered 1Fam as their building type
- most of the residents prefer single story building
- most of the residents prefer Gable as their roof style and prefer to have the roof material as Standard (Composite) Shingle
- residents prefer to have the exterior 1 and 2 with Vinyl Siding and have rated the condtion and quality of the exterior as average

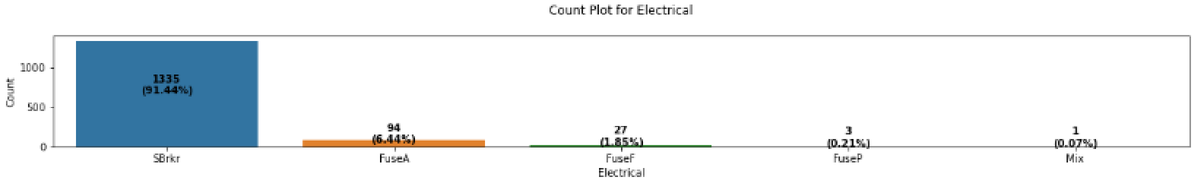


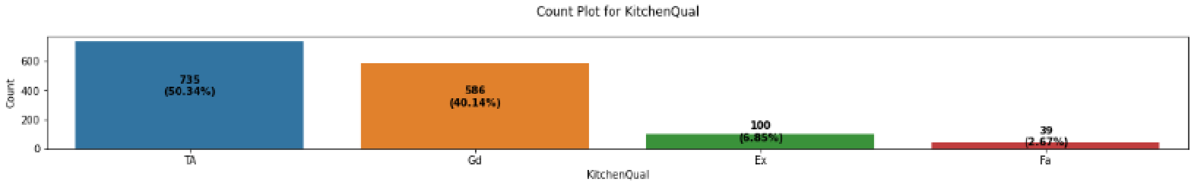
- residents dont prefer to have any masonry vaneer
- residents prefer to have the Foundation using CBlock
- residents prefer their BsmtQual as TA ie within 80-89 inches, and wants their basement condition as typical with some dampness allowed, with no allowance for their basement
- most of the houses considered had their BsmtFinType 1 and 2 as unfinished

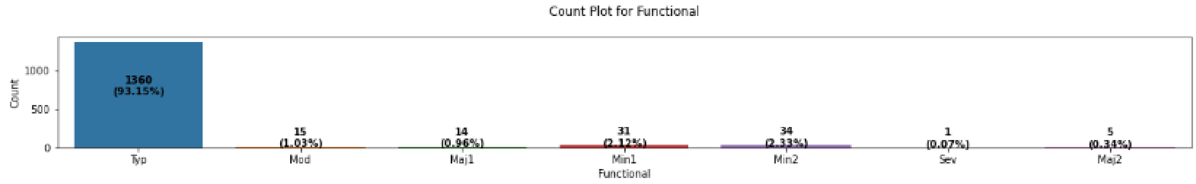


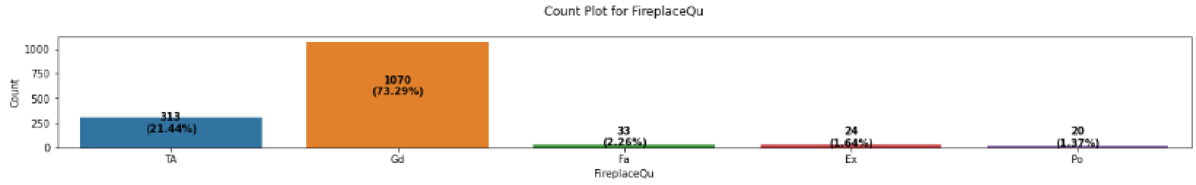




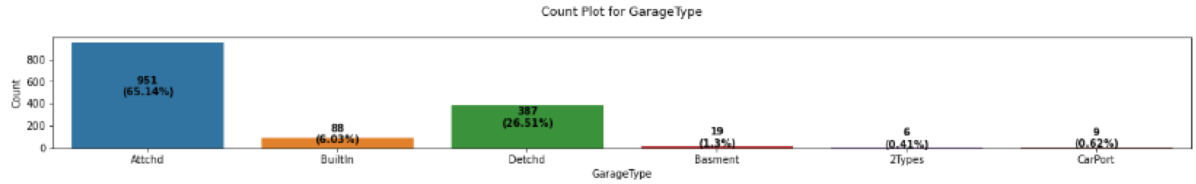


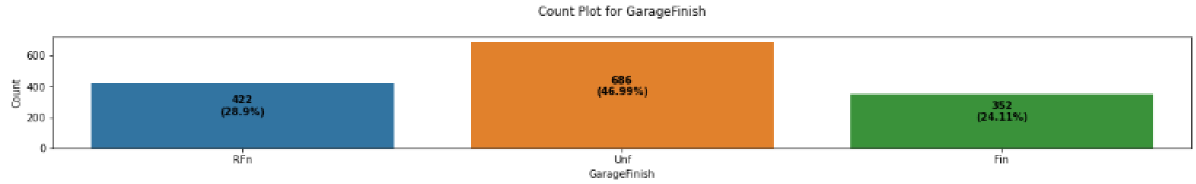


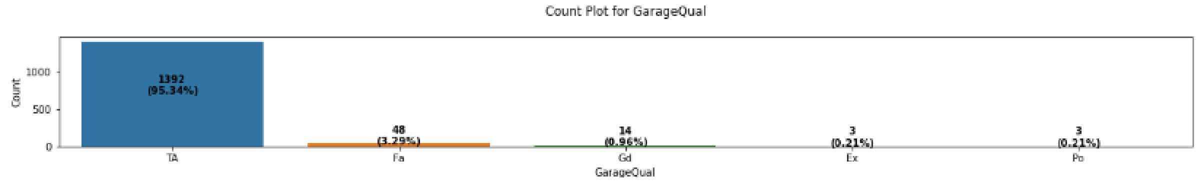


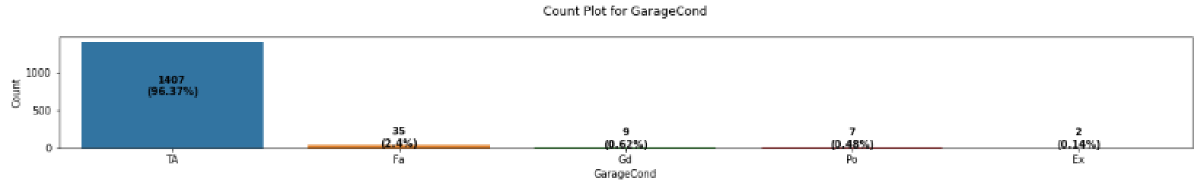


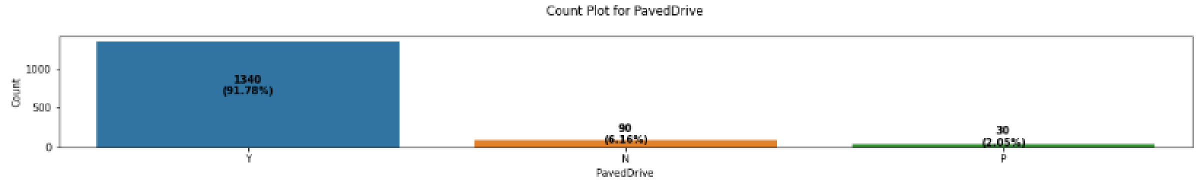
- the houses used GasA as their heating sources with central air conditioning, and have given excellent rating
- most of the houses had Standard Circuit Breakers & Romex as their electric systems
- KitchenQual was given as average
- residents prefer their houses to have typical functionality
- residents have given rating as good for quality of fireplace provided

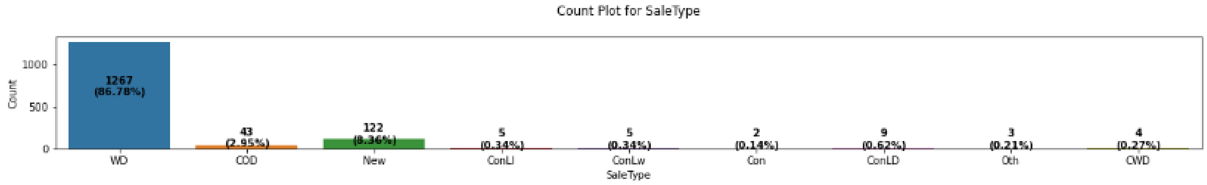


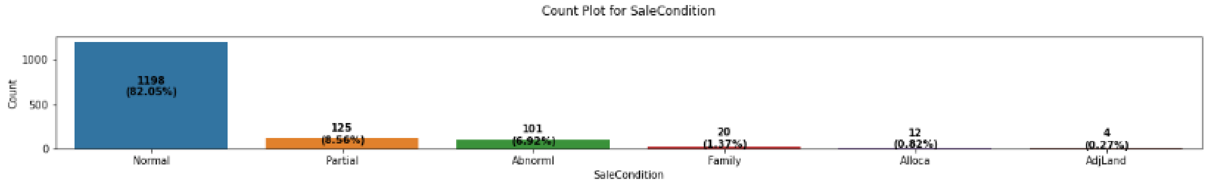






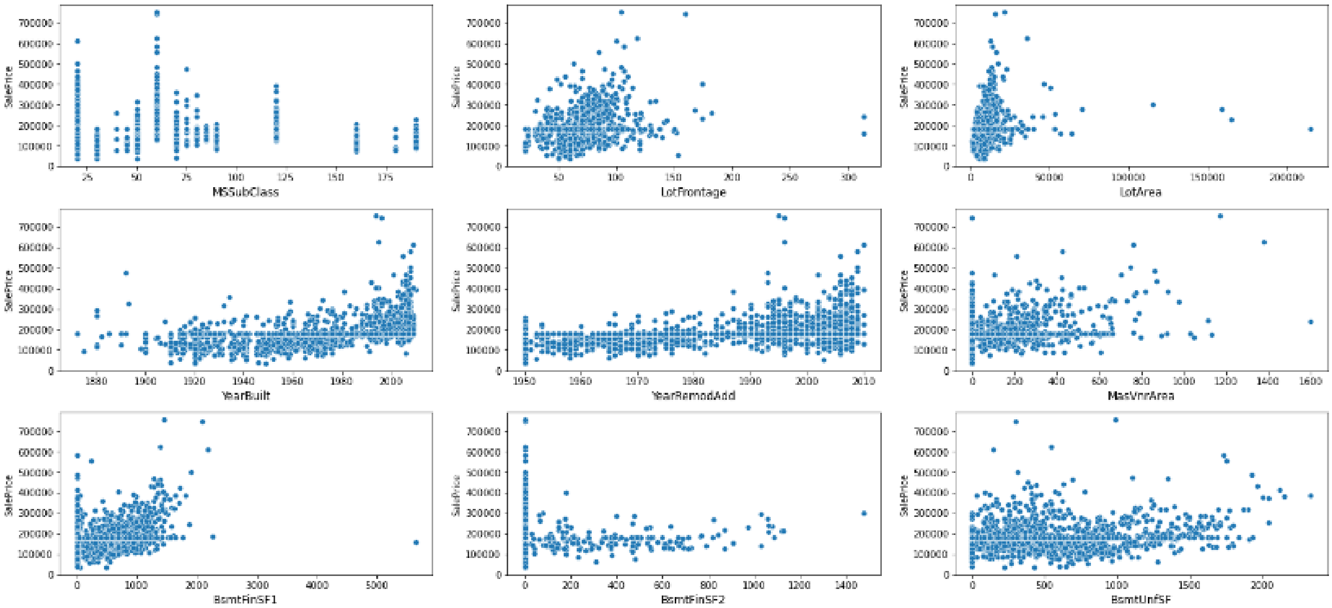






- the garage type prefered is attached with interior finish as unfinished, the quality of the garage and the condition of garage is average
- residents prefer their driveway as paved
- residents sale type was Warranty Deed - Conventional and sales condition was normal

iv. Bivariate analysis numerical continuous data



- MSSubClass
  - 2-STORY 1946 & NEWER (20) and 1-STORY 1946 & NEWER ALL STYLES (60) are having higher sale Price

- LotFrontage & LotArea
  - as the area increases the sale price increases

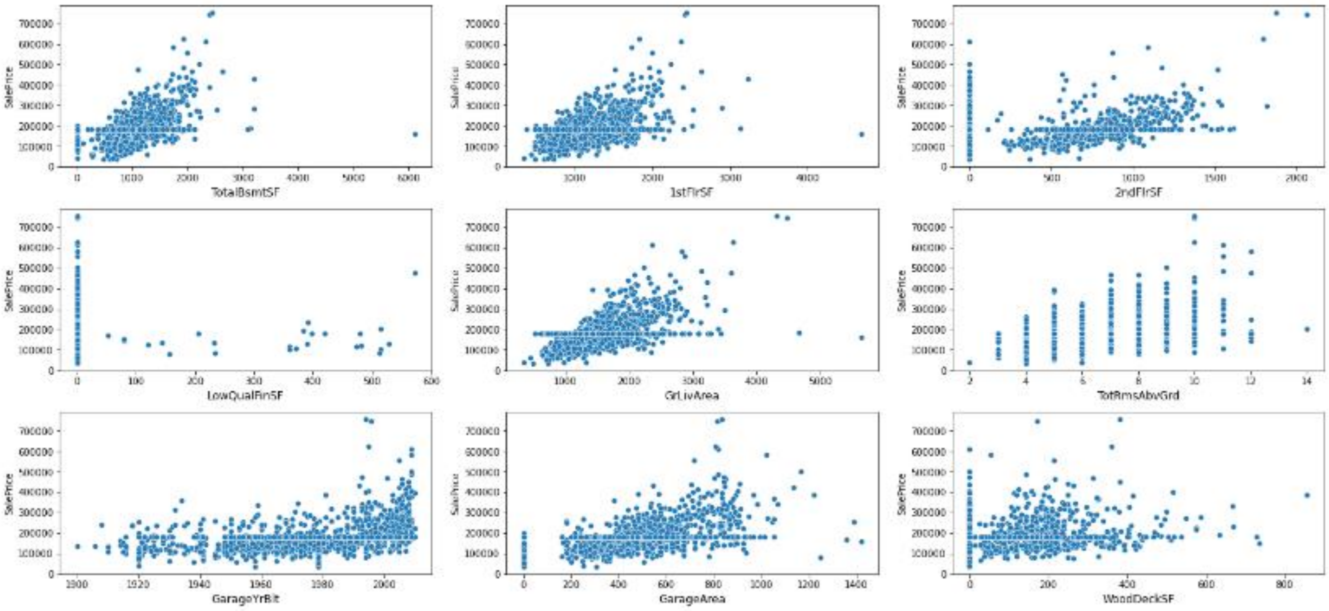
- YearBuilt
  - those houses bulit from 1900 to 1990 have the same sale price
  - sale price is high for houses built from 1991-2010

- YearRemodAdd
  - houses remodelled from 1950-1990 have the same price and linear price increase can be seen from 1991-2010

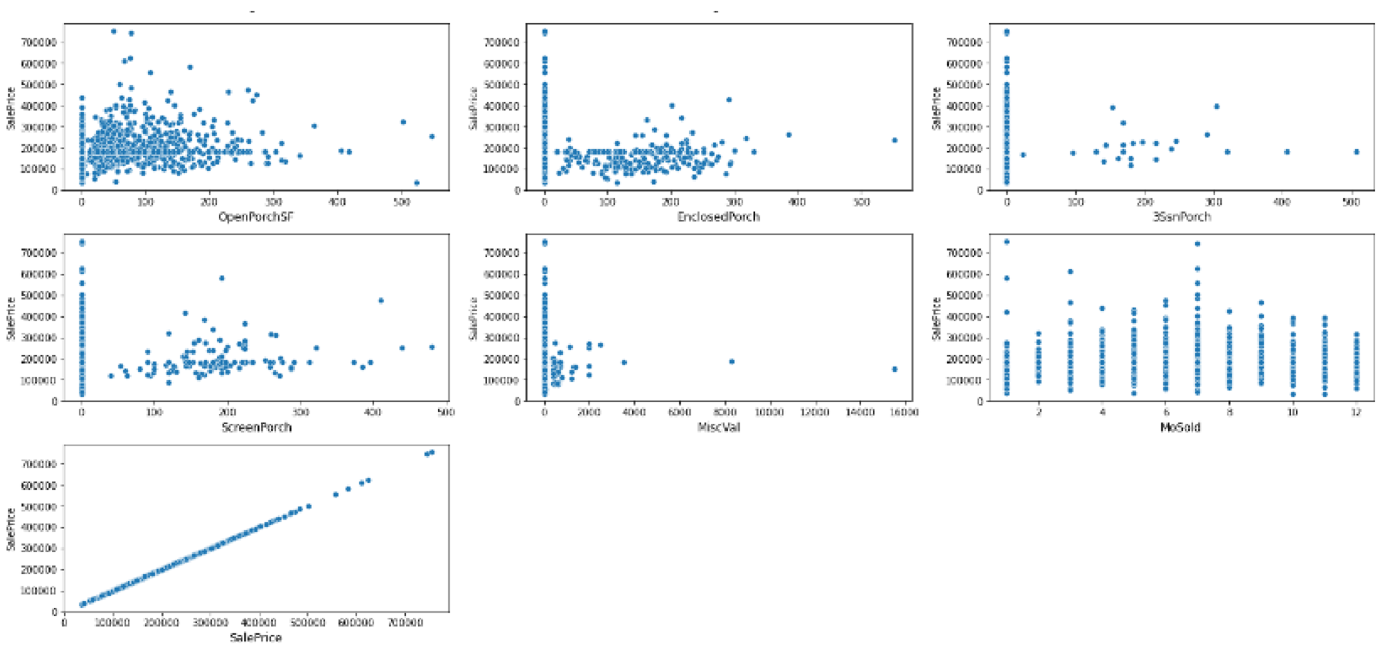
- MasVnrArea
  - as the area increase sale prices increases

- BsmtFinSF1
  - as the area increases price increases

- BsmtFinSF2 and BsmtuNFsf
  - the price remains almost a constant



- TotalBsmtSF
    - as the area increases price increases
- 1stFlrSF and 2ndFlrSF
    - as the area increases price increase
- LowQualFinSf
    - most of the houses had 0 and it sales price was around 500000
- GrLivArea
    - as area increases price increases
- TotRmsAbvGrd
    - most of the houses had 7-10 rooms which was above grade
- GarageYrBlt
    - garage were built from 1920 to 2010,
    - houses with garage built from 1990 to 2010 showed an increase in price with year and from 1940-1990 showed stable sale price with year
- GarageArea
    - as the area increase price increase
- WoodDeckSF
    - the price remains the for increase in area

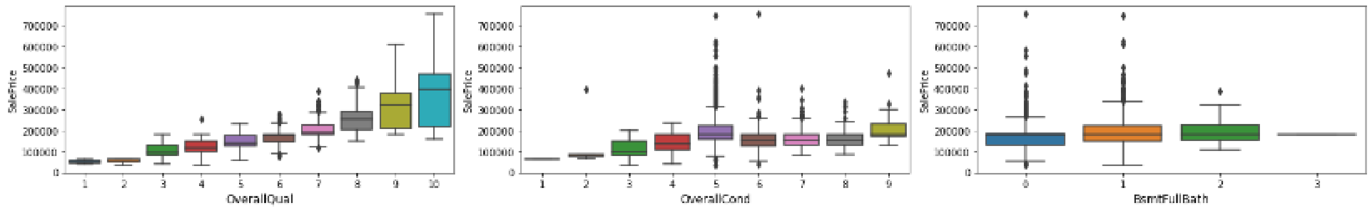


- OpenPorchSF, EnclosedPorch and ScreenPorch
  - as area increases price remains the same

- 3SsnPorch
  - most of the houses doesnt have three season porch

- MoSold
  - month 7 has the highest sales price and month 1 has the least sales price

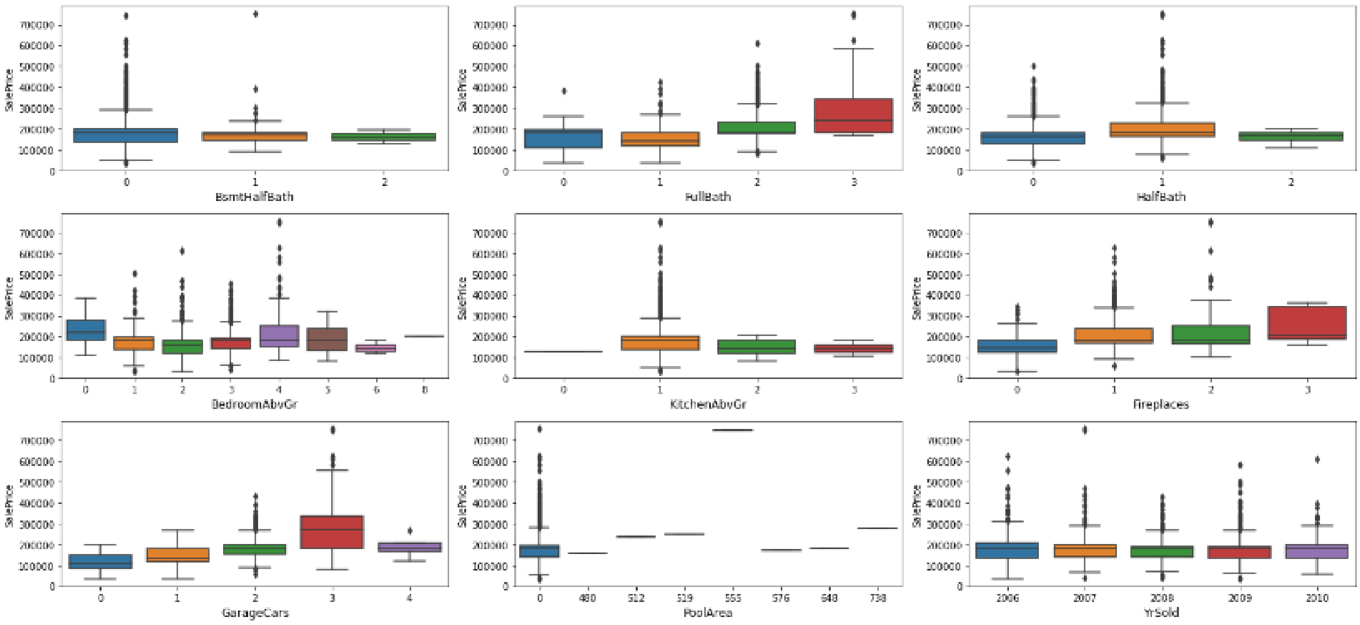
v. Bivariate analysis numerical discrete data



- OverallQual and OverallCond
  - as the quality increases the sales price increases

- BsmtFullBath
  - as the no. of full basement bathrooms increases price increases

- BsmtHalfBath
  - as the no, of half bathroom basment increases there is no significant change in price



- FullBath
  - as the no. of full bath above grade increases the price increases

- HalfBath
  - houses with 1 half bath has highest sales price

- BedroomAbvGr
  - houses with 0 and 4 bedroom above grade is expensive compared to other

- KitchenAbvGr
  - houses with 1 kitchen above grade is expensive compared to others

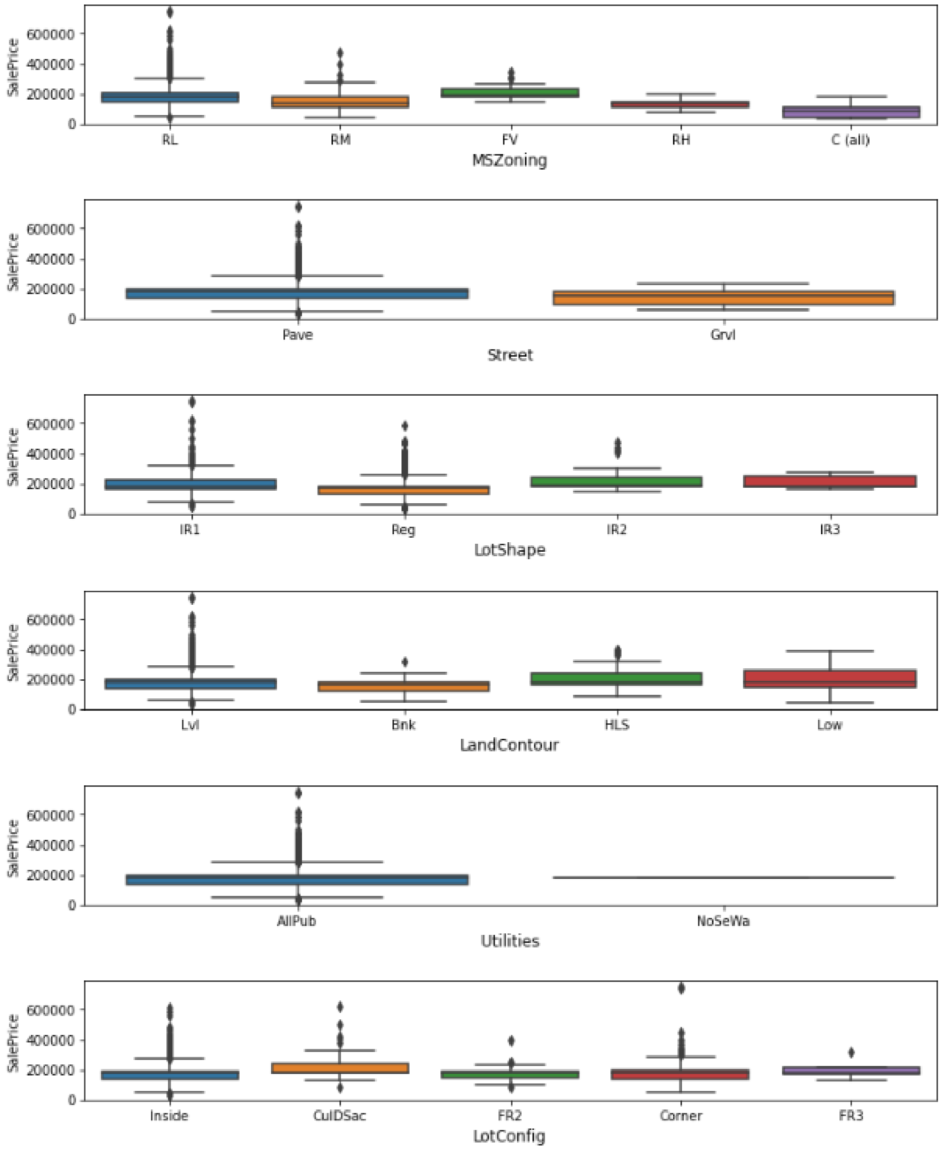
- Fireplaces
  - as the no of fireplaces increases the sales price also increases

- GarageCars
  - as the no of the cars which can be kept in garage increases the sales price increases with exception for 4 cars

- Pool Area
  - most of the houses dont have pools
  - as the pool area increases sales price increases

- YrSold
  - as the years change the sale price is almost constant

vi. Bivariate analysis categorical data



- MSZoning
    - RL has the highest sales price

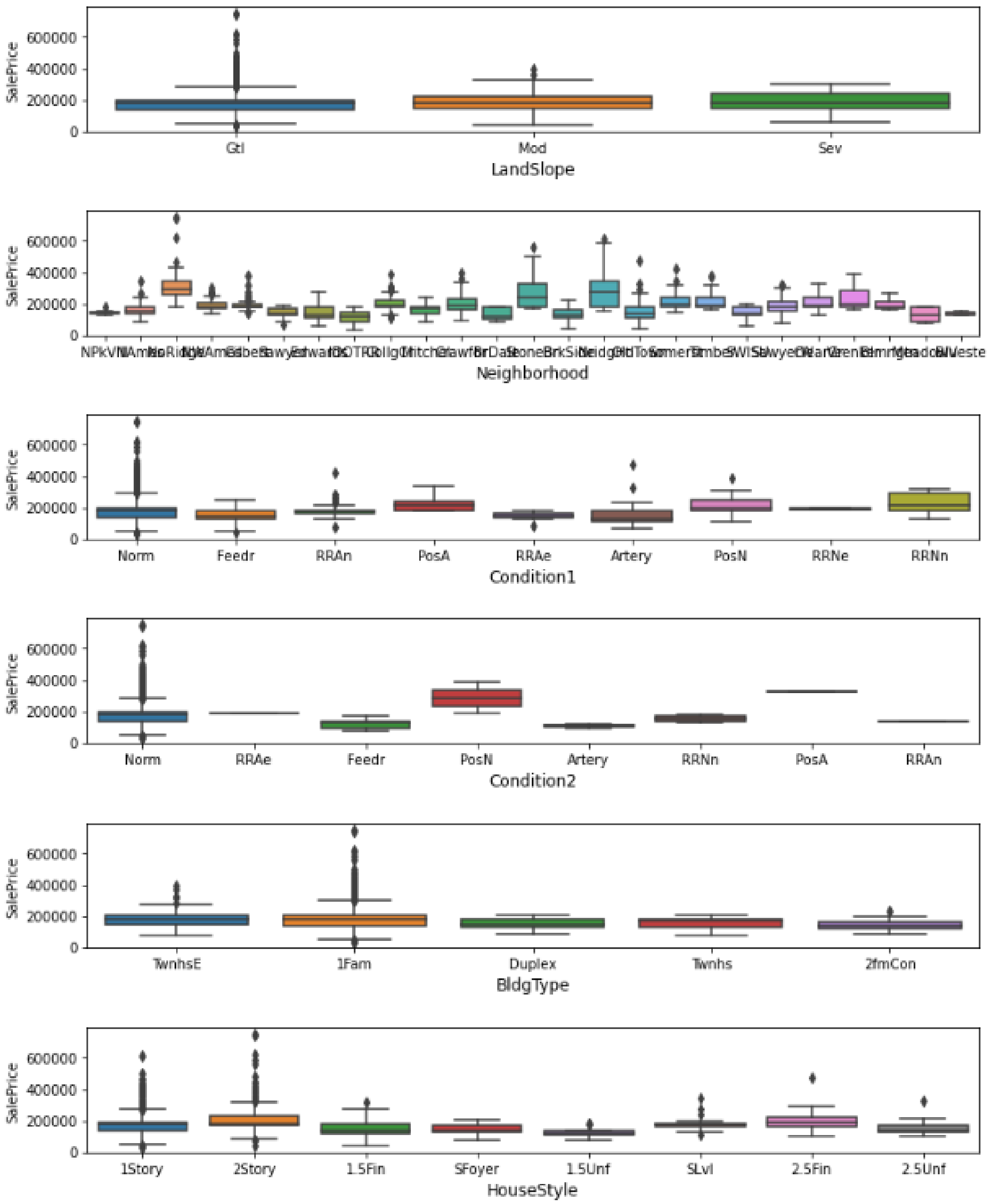
- Street
    - Price is higher for houses having ppaved roads

- LotShape
    - IR1 has the highest sales price, followed by IR2

- LandContour
    - Land Contour with low types has the highest sales price

- Utilities
    - houses with all public utilities are expensive

- LotConfig
    - houses having lots in Cul-de-sac configuration is expensive

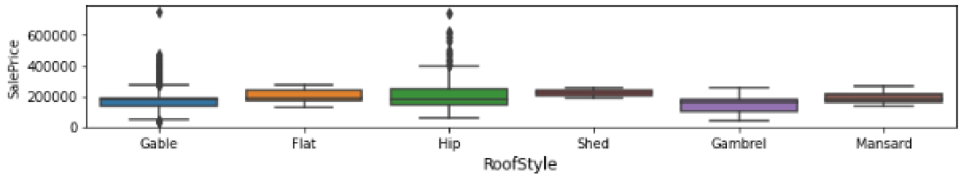


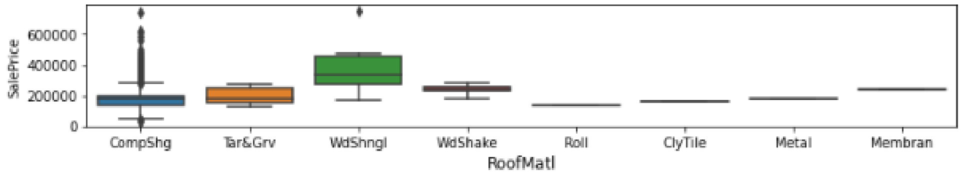
- Landslope
  - houses with Moderate slopes are expensive

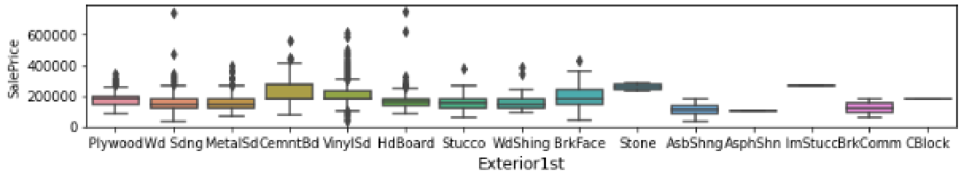
- Condition1 and condition2
  - Houses with PosA and PosN are expensive

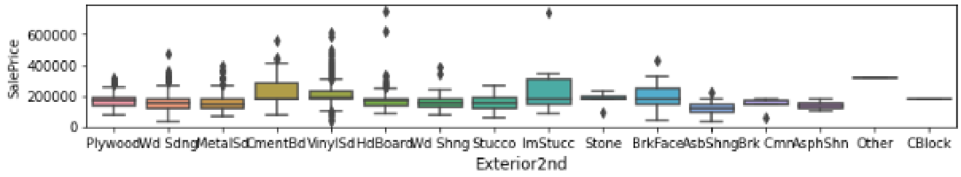
- BldgType
  - houses with Single-family Detached is expensive

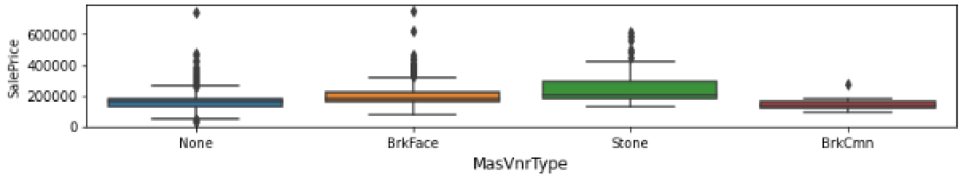
- HouseStyle
  - 2Story houses are expensive

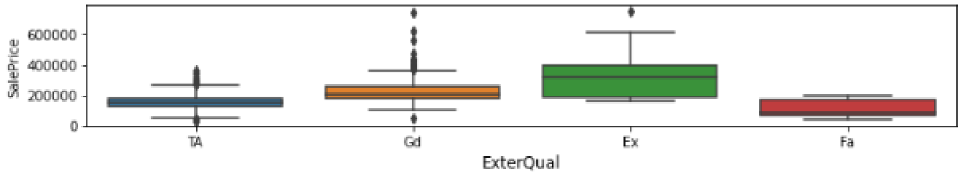


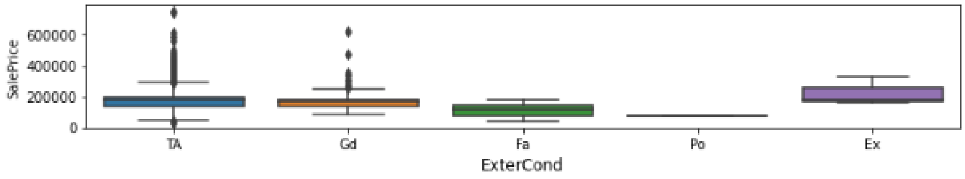










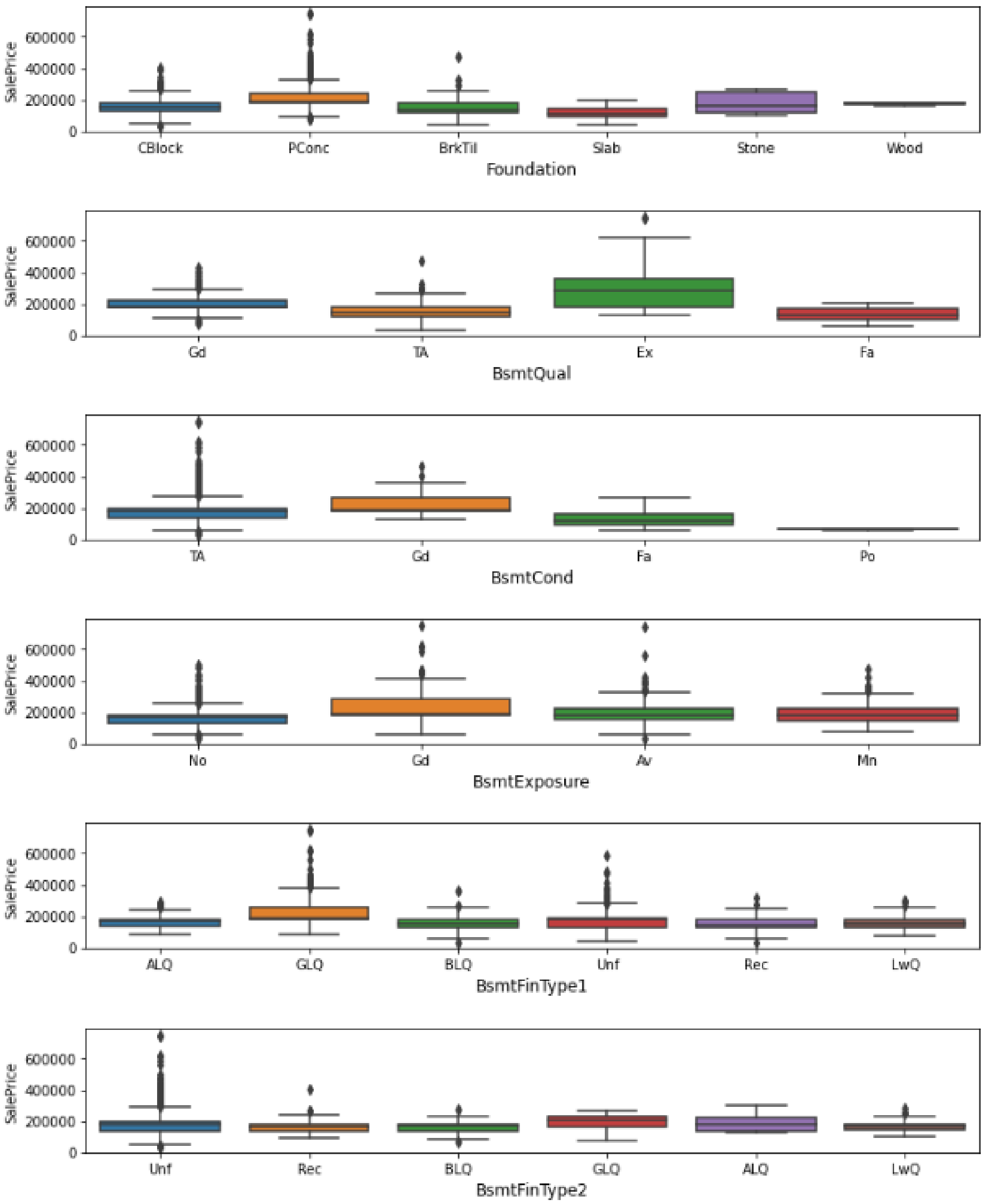


- RoofStyle and RoofMatl
  - houses with Hip as roofstyle and WdShngl as materail is expensive

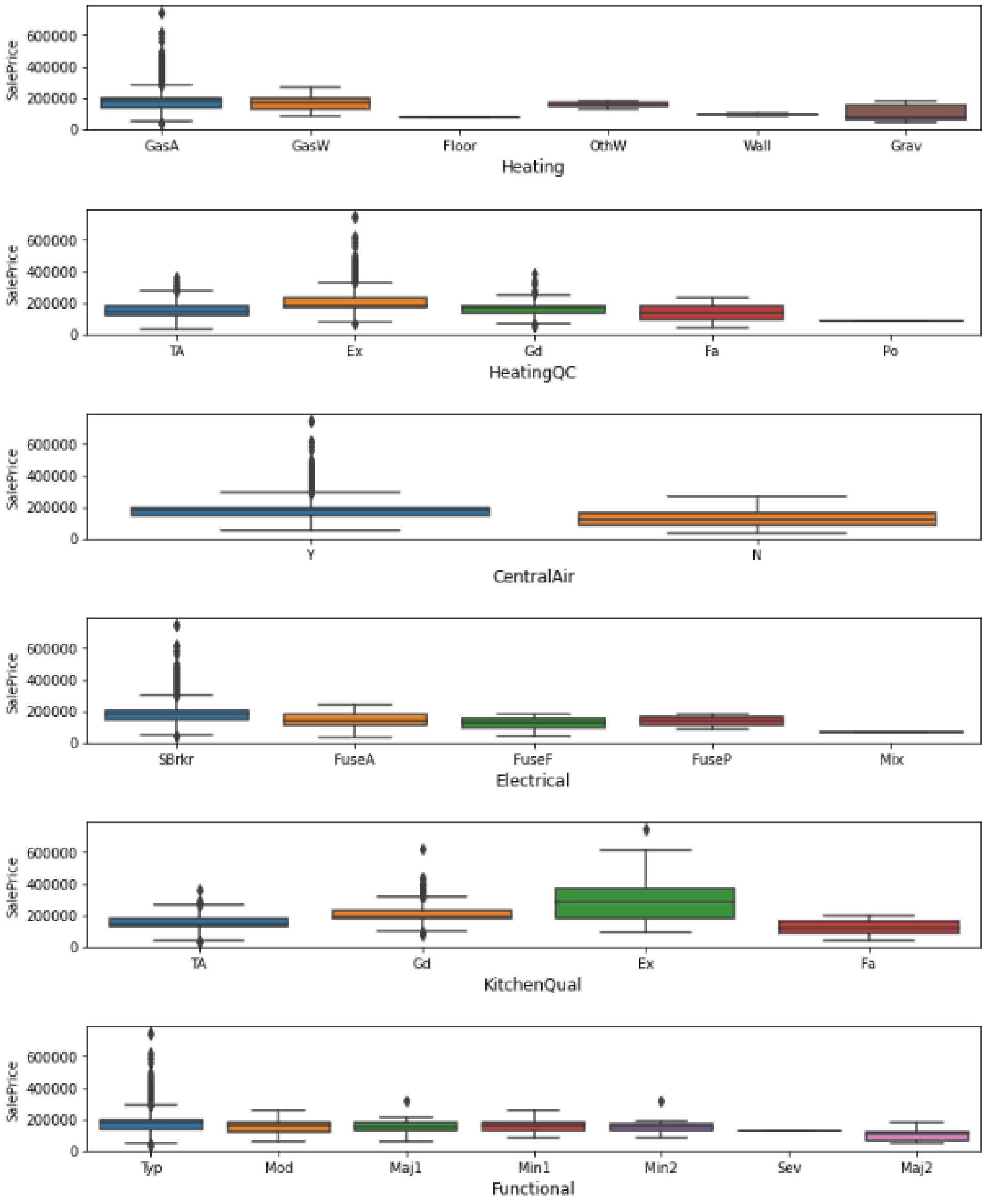
- Exterior1 and Exterior2
  - houses with Cement Board is expensive compare to others

- MasVnrType
  - Stone type masonary vaneer type is expensive

- ExterQal and ExterCond
  - houses with excellent rating is expensive



- Foundation
  - houses with Pconc is expensive
- BsmtQal
  - houses with excellent as basment quality is expensive
- BsmtCond
  - houses with basement condition as Good is expensive
- BsmtFintype1
  - houses with GLQ is expensive
- BsmtFintype2
  - houses with Unf is expensive



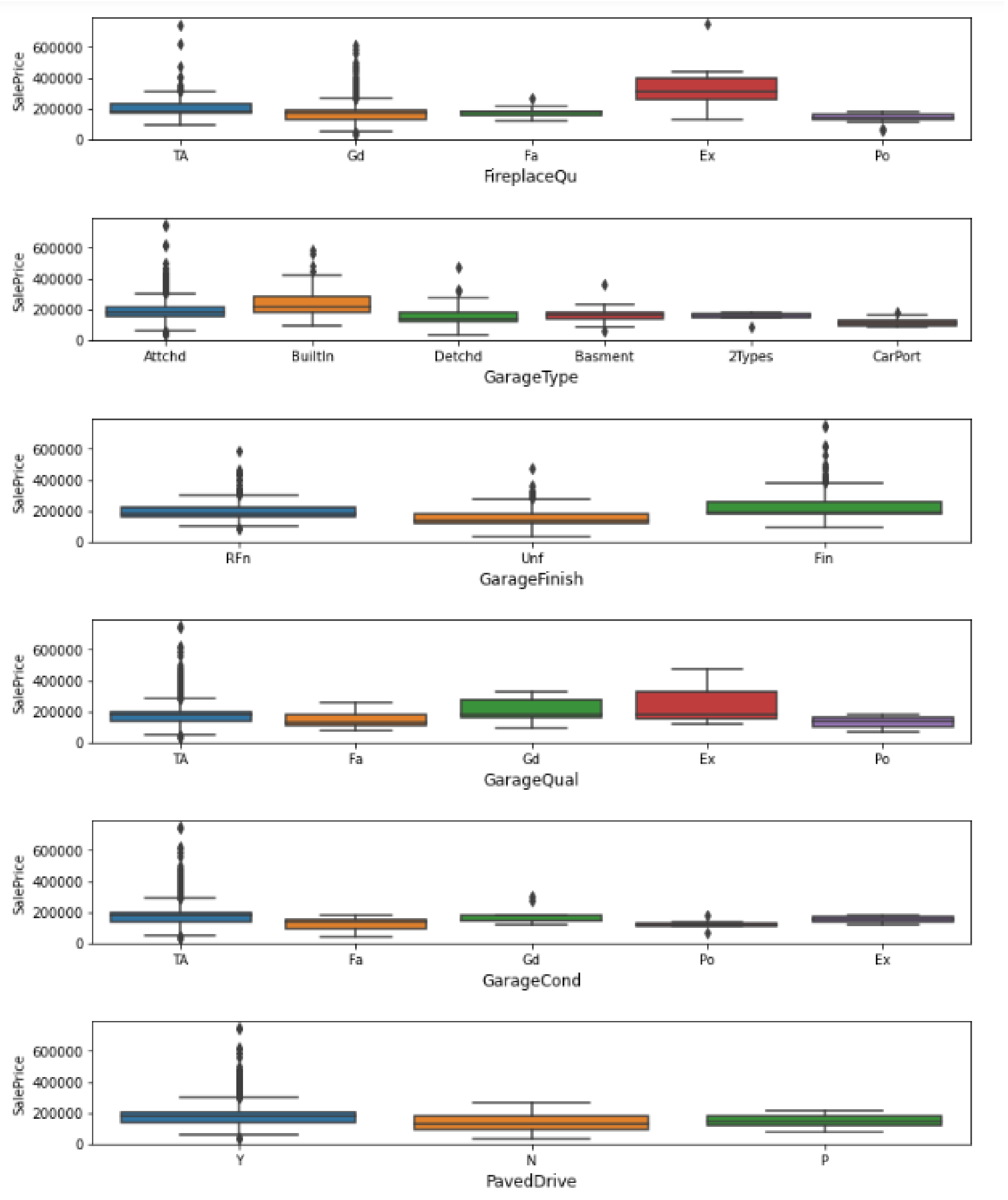
- Heating and HeatingQC
  - houses having heating type as GasA and GasW, with heating quality as excellent are expensive

- CentralAir
  - sales price remains the same irrespective of provision for central air conditioning

- Electrical
  - Houses with Sbrkr type of electric systems are expensive

- KitchenQual
  - houses with excellent Kitchen quality is expensive

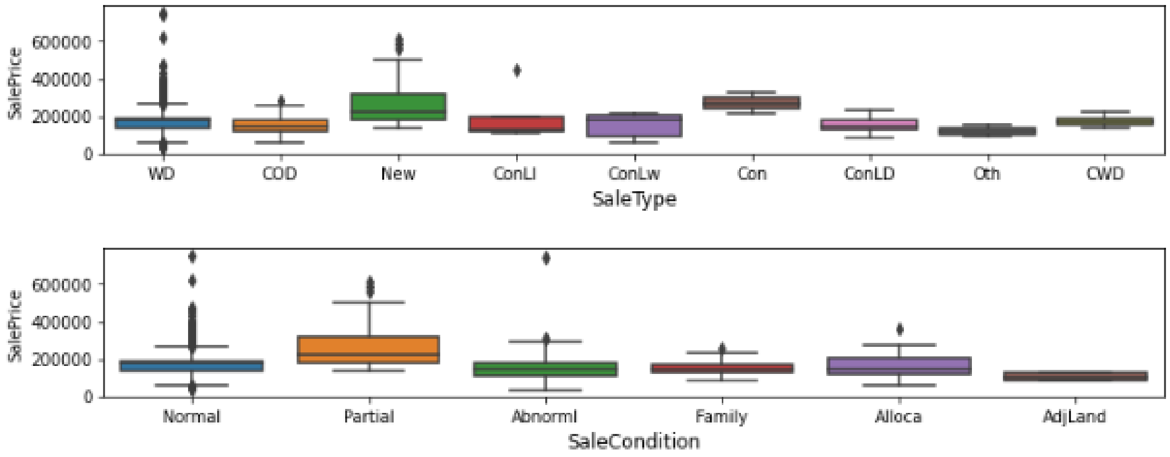
- Fucntional
  - houses with Typical functionality is expensive



- FireplaceQU
  - houses with excellent rating fireplace are expensive

- GarageType, GarageFinish, GarageQual and Garage Condition
  - houses with builtin garage having its interior finished, with excellent quality and average condition is expensive

- PavedDrive
  - houses with paved drives are expensive



- SaleType
  - Houses which are newly constructed is the most expensive

- SaleCondition
  - houses with sale condition partial is expensive, this is because new sale types houses are expensive

c) Data cleaning

a. Skewness

The skewness of the numerical continuous features is checked and all those which exceeds the skewness limit is treated using power transform some features skewness was not reducing even after trying other methods so dropped the columns 'EnclosedPorch', 'BsmtFinSF2', 'ScreenPorch', 'MiscVal', 'LowQualFinSF', '3SsnPorch' from the dataframe

b. Correlation

The correlation of target with numerical continuous features were checked to see how they were correlated and later feature to feature correlation was plotted on a heatmap. To avoid the problem of multicollinearity vif was used and column 'GrLivArea' was dropped as it exceeded vif limits

c. Encoding

- The feature which were having rating values were mapped as the following
  'Ex':5, 'Gd':4, 'TA':3, 'Fa':2, 'Po':1, 'None':0
- Rest of the categorical columns were encoded using ordinal encoding

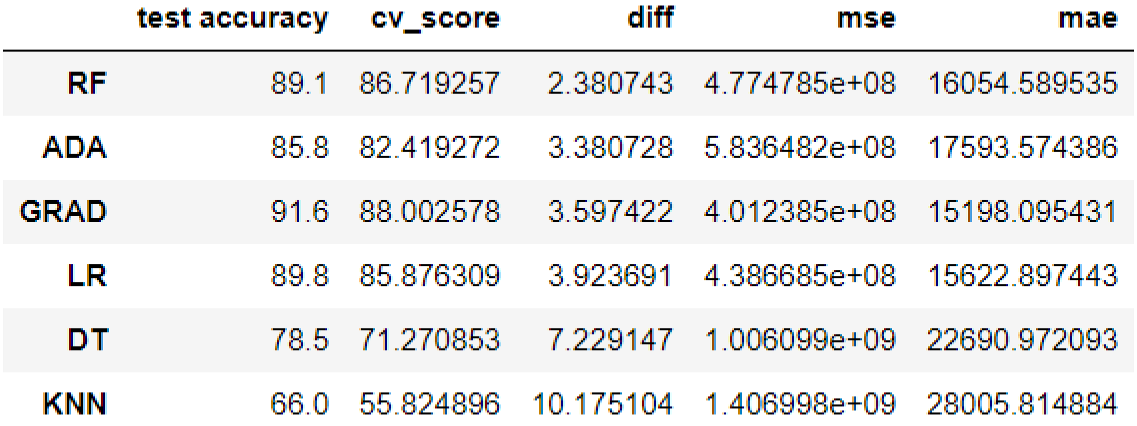
d. Outliers
Outliers were checked for train dataset after separating train and test . As outliers were present it was treated using ZScore method

e. Splitting dataset into two X and Y, where X had all feature and Y had target

Testing of Identified Approaches (Algorithms)

The target variable is continuous data hence it was regression problems. Among the available models available the models used for the analysis were

1. Random Forest
2. Adaboost regressor
3. Gradientboost regressor
4. Linear Regressor
5. Decision Tree
6. KNN



	test accuracy	cv_score	diff	mse	mae
RF	89.1	86.719257	2.380743	4.774785e+08	16054.589535
ADA	85.8	82.419272	3.380728	5.836482e+08	17593.574386
GRAD	91.6	88.002578	3.597422	4.012385e+08	15198.095431
LR	89.8	85.876309	3.923691	4.386685e+08	15622.897443
DT	78.5	71.270853	7.229147	1.006099e+09	22690.972093
KNN	66.0	55.824896	10.175104	1.406998e+09	28005.814884

The best model was selected as Random Forest based on the following factors

1. least difference between test accuracy and cv_score
2. second highest test accuracy and cv_score
3. least error

Hyper Parameter Tuning

The following parameters were considered for tuning

- n_estimators
- max_features
- criterion
- max_depth
- min_samples_split

After tuning the cv_score and test accuracy increased

The model was saved in pickle format and later used to predict based on the different conditions and stored in a dataframe named House_Price_Predictions

CONCLUSION

- Key Findings and Conclusions of the Study

The customers find it difficult to buy houses to their liking when its not within their expected budget and makes them very frustrated as they have selected the house after considering various factors and sometimes think if the house they are going to buy is actually a good deal for them and wait for better opportunities.

A large amount of this problem is solved by using the machine learning and hence the customers are able to get an idea of how much the house could cost and save a lot of time

There were many factors which contributed towards the pricing of the house such as the condition of the house, when the house was built, when it was remodelled etc.

- Learning Outcomes of the Study in respect of Data Science

The dataset had two sets of data to predict and for training the model. Various features were taken into consideration for determining the price of house. There were nulls present which was imputed and graphical representation helped to gain an insight on how the data was spread and how the data needed to be cleaned and how these features could be related to target variable. Data cleaning and feature selections plays a vital role in the performance of the model. The skewness of column was treated to some extend.

- Limitations of this work and Scope for Future Work
    - There were lot of features that had zero values this tends to create a bias
    - It is not always preferred to remove outliers as it is possible for outliers to be special case
    - Imputation for numerical data was done by taking the mean of that column it would have been better to quantify value based on the various factors that contribute to it
    - Skewness couldn’t be treated to a large extend and some columns were dropped as skewness couldn’t be reduced by using various treatment methods
    - Multicollinearity was present and a feature had to be removed because of it