FLIP ROBO

HOUSING: PRICE PREDICTION

Submitted by:

ANCHAL AWASTHI

ACKNOWLEDGMENT

Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

INTRODUCTION

- **Business Problem Framing.**

It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. The company is looking at prospective properties to buy houses to enter the market. We have to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

- **Conceptual background of domain problem.**

The conceptual Background of Domain consists of Data science and Many predictive modelling, recommendation systems and lot of techniques used in data science.

- **Review of Literature.**

This is a review of literature which we undergone through a little research into it from plenty of websites as Wikipedia for Housing prices in Australia, Kaggle and GitHub for data analyzing. And, Research done on lots of estimated prices and cost of living in different cities to understand the data.

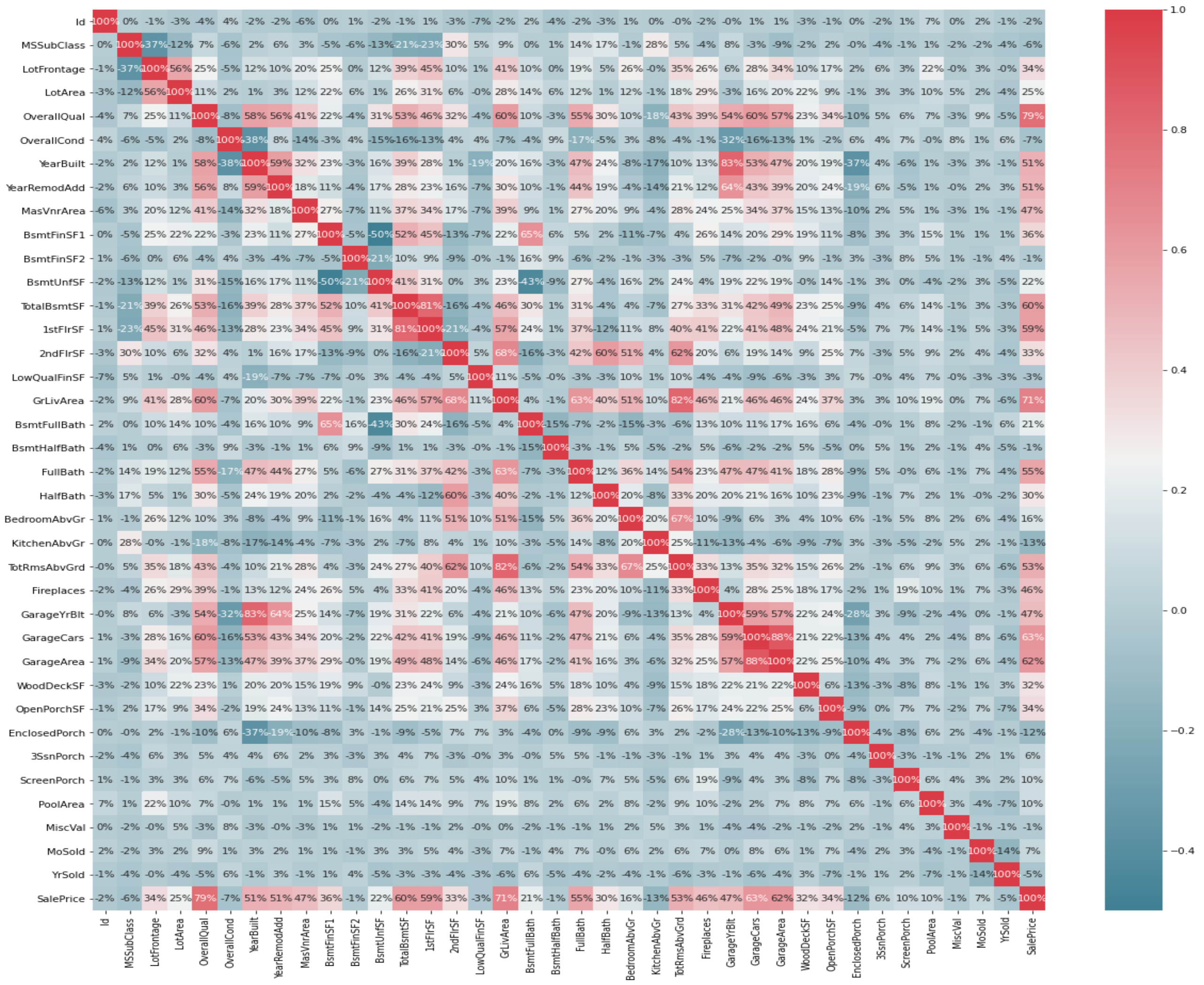
- **Motivation for the problem undertaken.**

The Problem undertaken is to get predictions of best predicted prices and this makes me to make this project because a lot of were suffering for small homes or medium range house with affordable pricing so they can get an idea about all things undergone through this. So, we are here with our technology as Data Science and this technology motivates me to work into this domain.

Analytical Problem Framing

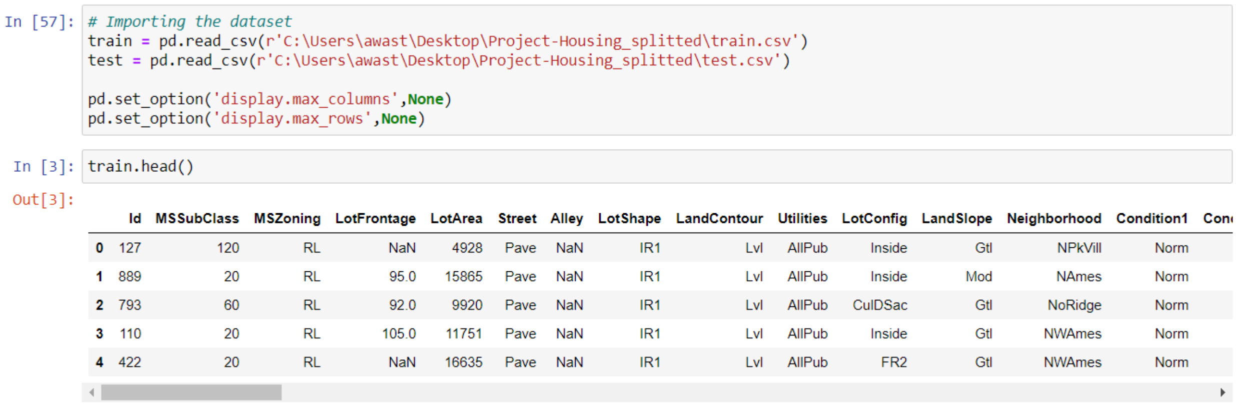
- **Mathematical/ Analytical Modelling of the Problem**

In this project we have performed various mathematical and statistical analysis such as we checked description or statistical summary of the data using describe, checked correlation using corr and also visualized it using heatmap. Then we have used zscore to plot outliers and remove them.



- **Data Sources and their formats**

The sample data is provided to us from our client database. It is provided in csv format and hence we import it using pandas. Then we further checked more about data using info, checked data types using dtypes, shapes using. shape, columns using. columns, null values using isnull.sum, and further visualize it through heatmap as follows:



In [57]:

```
# Importing the dataset
train = pd.read_csv(r'C:\Users\awast\Desktop\Project-Housing_splitted\train.csv')
test = pd.read_csv(r'C:\Users\awast\Desktop\Project-Housing_splitted\test.csv')

pd.set_option('display.max_columns',None)
pd.set_option('display.max_rows',None)
```

In [3]:

```
train.head()
```

Out[3]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Con
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	

- **Data Pre-processing Done**

First, we will determine whether there are any null values and since there were null values as well as NaN vales present in the dataset, we proceeded further by imputing them using Simple Imputer with mean and most frequent as strategies respectively. Next, we did Label encoding using label encoder. Then we performed some data visualization in which we observed certain attributes were having skewness and outliers that were plotted using distplot and boxplot. Outliers were removed with the help of Zscore in which 685 rows were removed.

- **Hardware and Software Requirements and Tools Used**

In this project we have used HP Pavilion PC with 64-bit operating system and have Windows 10 pro. We have used python to develop this project in which we have used various libraries.

Model/s Development and Evaluation

- **Identification of possible problem-solving approaches**

We have performed various mathematical and statistical analysis such as we checked description or statistical summary of the data using describe, checked correlation using corr and also visualized it using heatmap. Then we have used zscore to plot outliers and remove them. We have used distplot to find the distribution of all attributes.

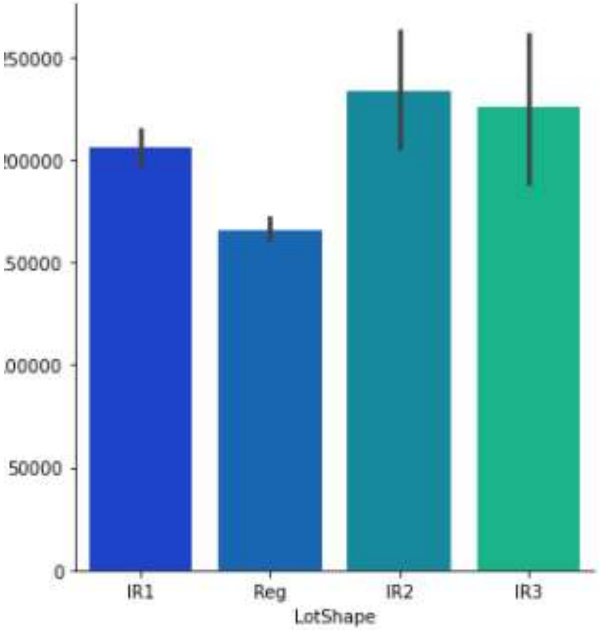
- **Run and evaluate selected models**

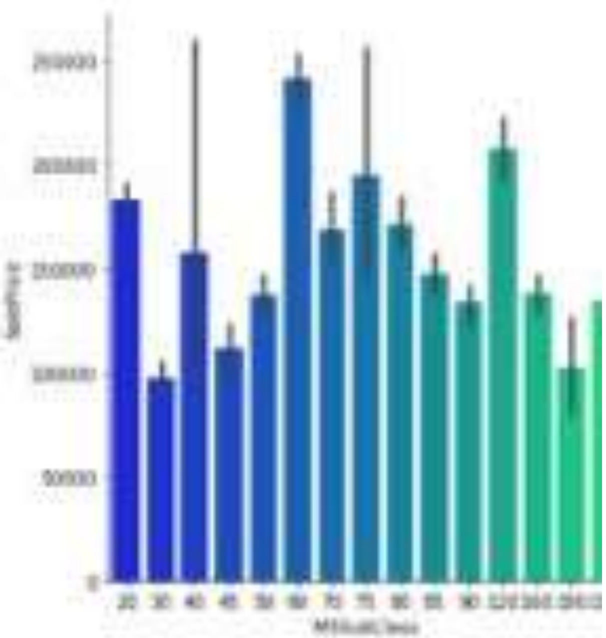
I chose GradientBoostingRegressor as our best model since it's giving us best score and it's performing well. It's r2_score is also satisfactory and it shows that our model is neither underfitting/overfitting. Then we performed hyperparamter tuning using GridSearchCV on GradientBoostingRegressor from which got 'learning_rate': 0.1, 'n_estimators': 350 as best parameters. We got score : `0.9963126920769413` after performing hyperparameter tuning and earlier it was `0.9846658425719441`. Its r2_score is also satisfactory.
Hence we saved GradientBoostingRegressor as our final model using joblib.

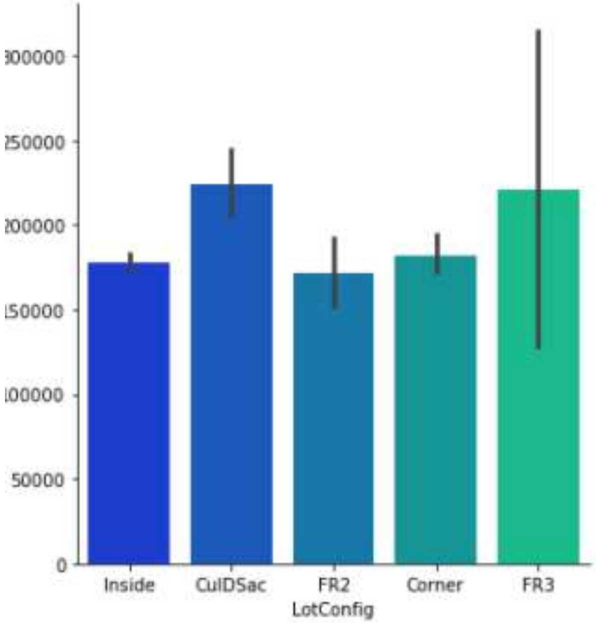
- **Key Metrics for success in solving problem under consideration**

Key metrics used for finalising the model was Score and r2_score. Since in case of GradientBoostingRegressor it's giving us good score among all other models and it's performing well. It's r2_score is also satisfactory and it shows that our model is neither underfitting/overfitting .

- **Visualizations**



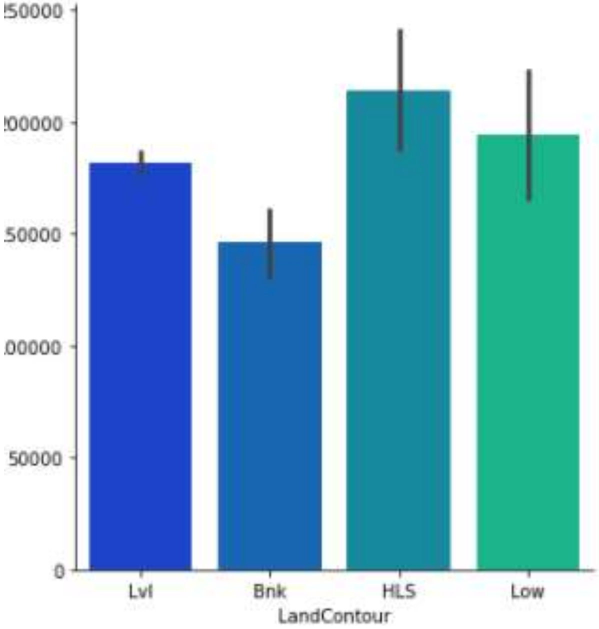


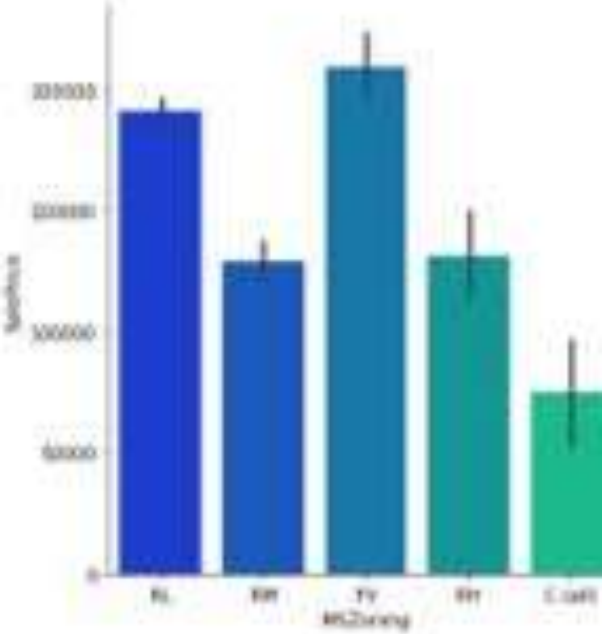


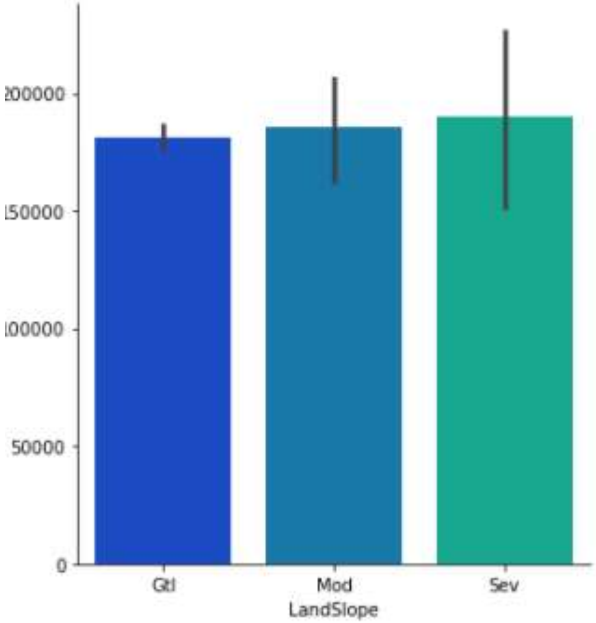
```
gure size 720x432 with 0 Axes>
```

```
<Figure size 720x432 with 0 Axes>
```

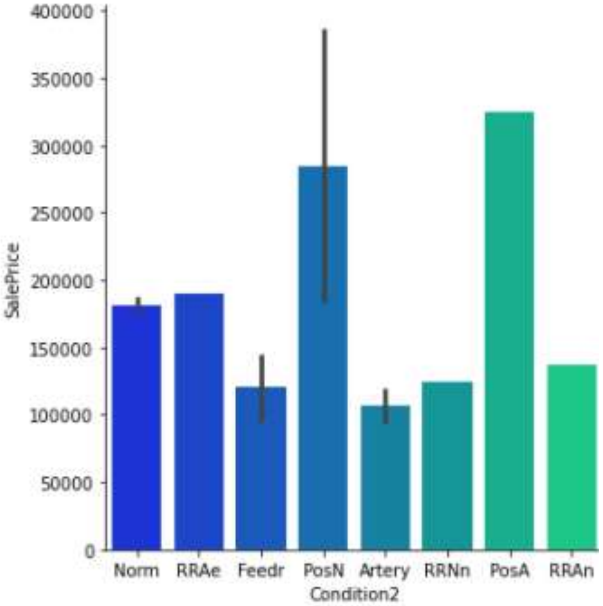
```
gure size 720x432 with 0 Axes>
```



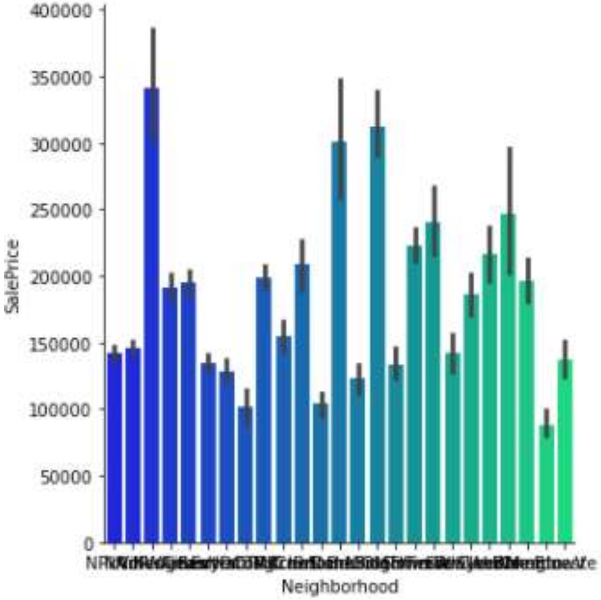




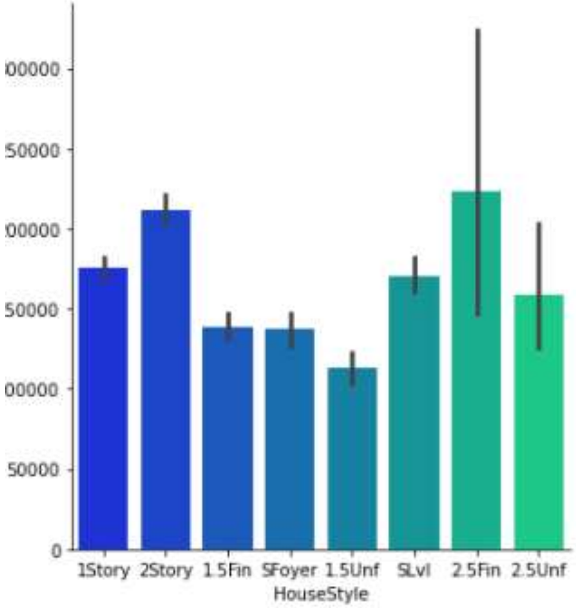
```
:Figure size 720x432 with 0 Axes>
```



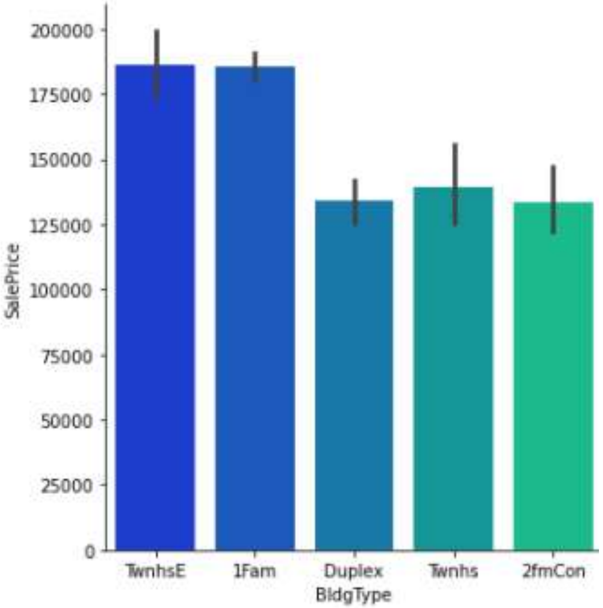
```
Figure size 720x432 with 0 Axes>
```



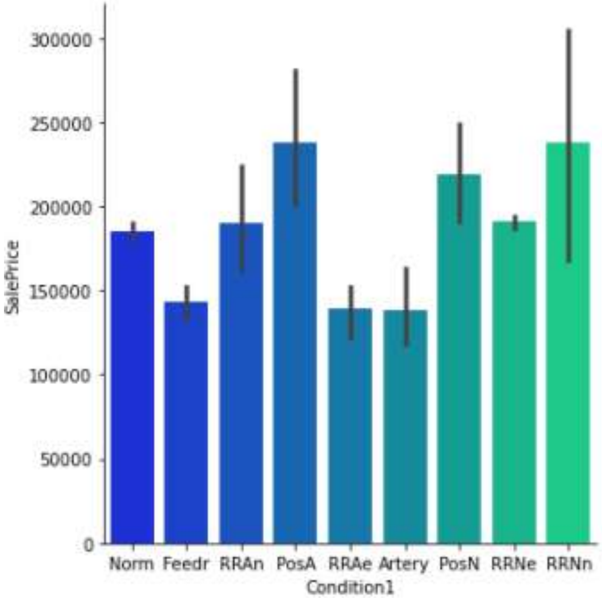
```
gure size 720x432 with 0 Axes>
```



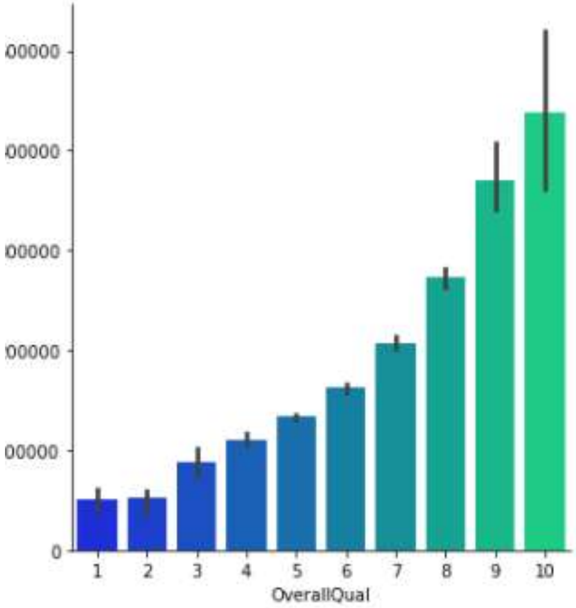
```
:Figure size 720x432 with 0 Axes>
```



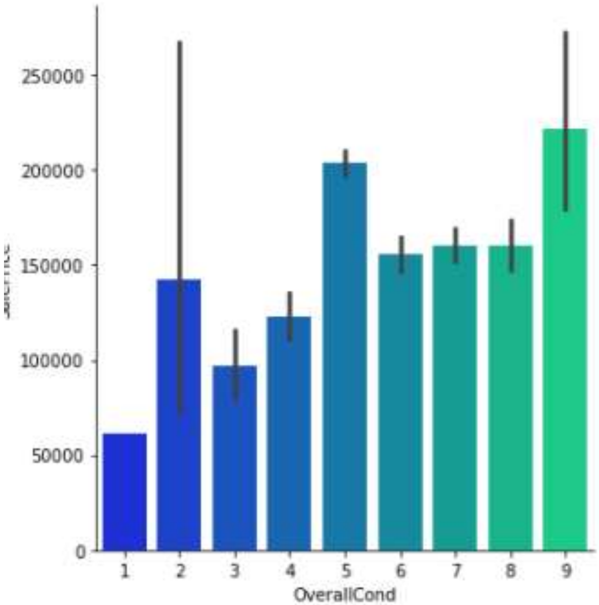
```
Figure size 720x432 with 0 Axes>
```

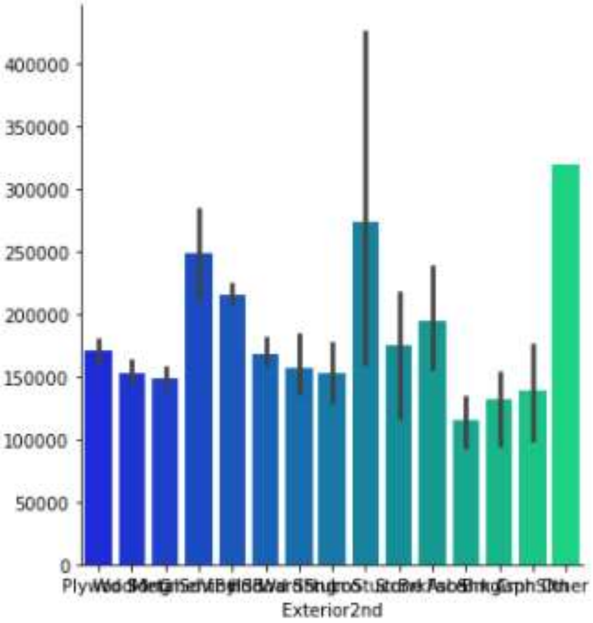


```
gure size 720x432 with 0 Axes>
```

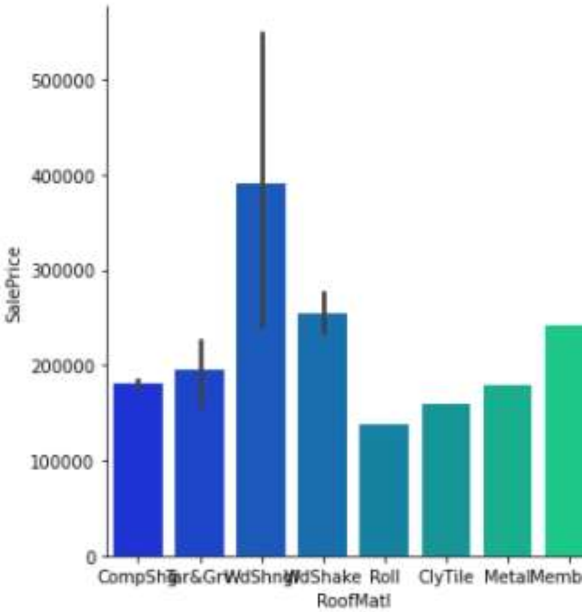


```
igure size 720x432 with 0 Axes>
```

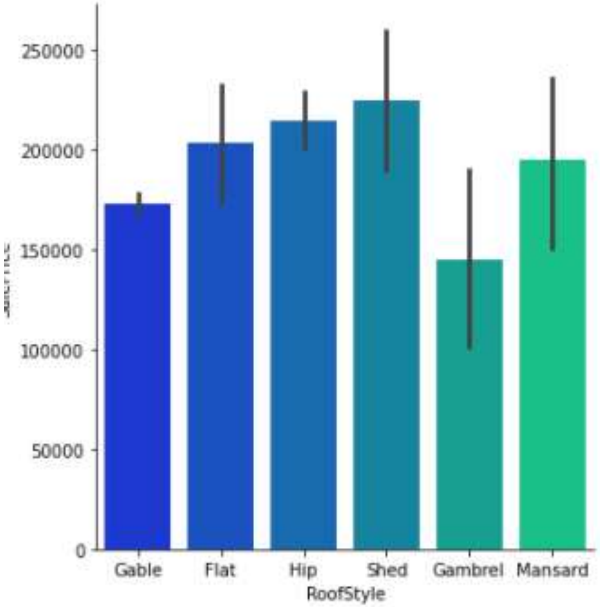




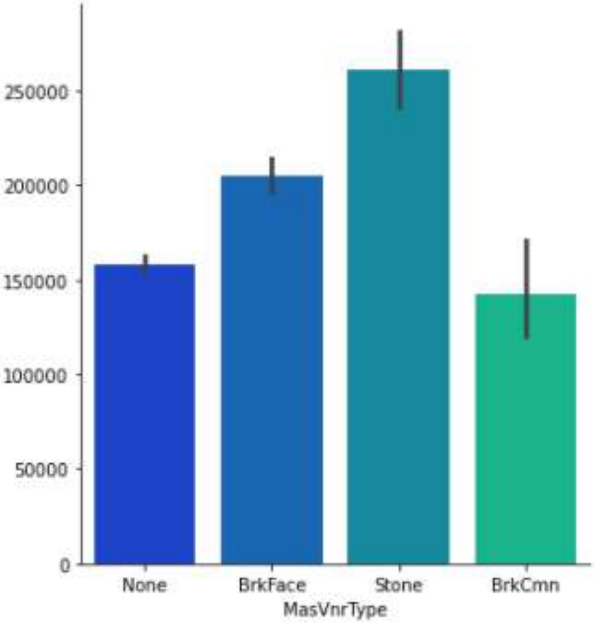
```
<Figure size 720x432 with 0 Axes>
```



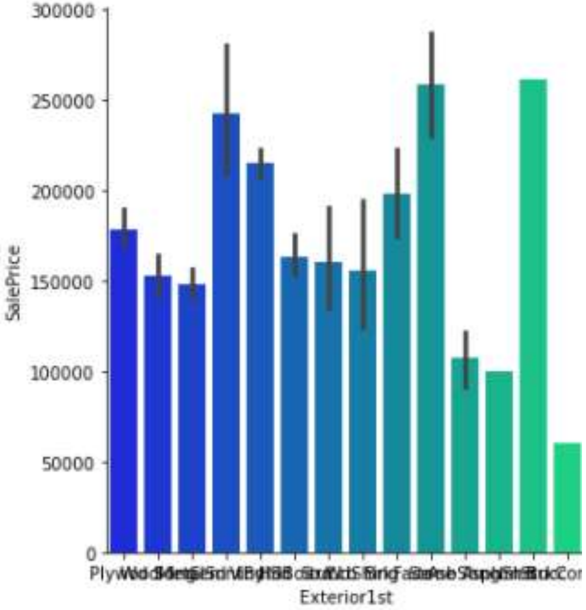
```
igure size 720x432 with 0 Axes>
```



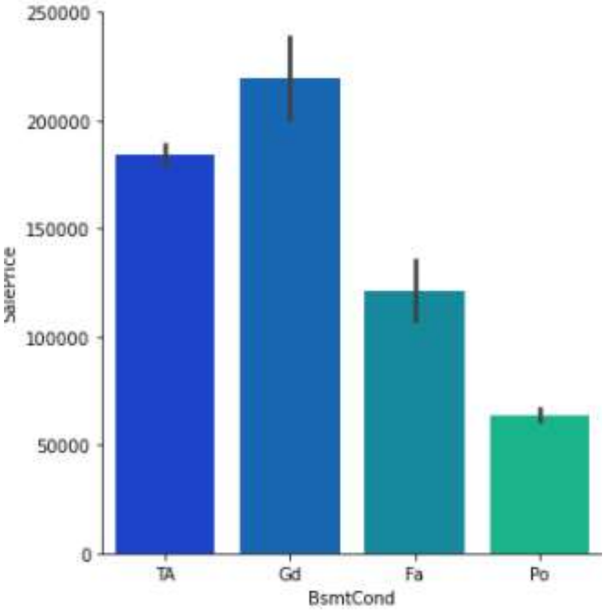
```
igure size 720x432 with 0 Axes>
```



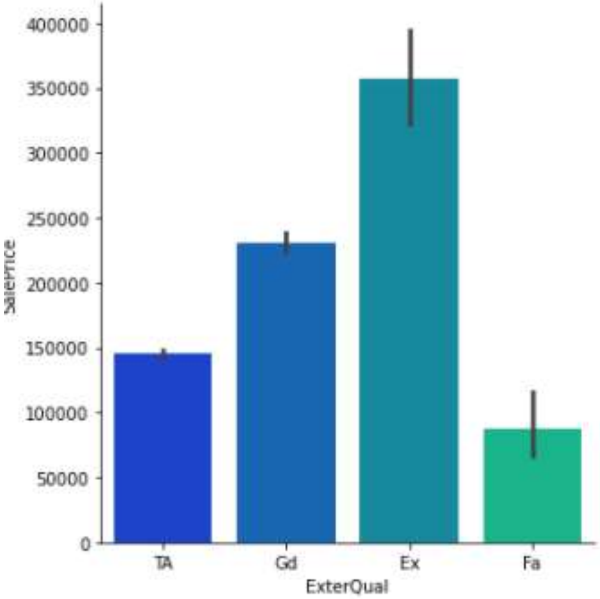
```
<Figure size 720x432 with 0 Axes>
```



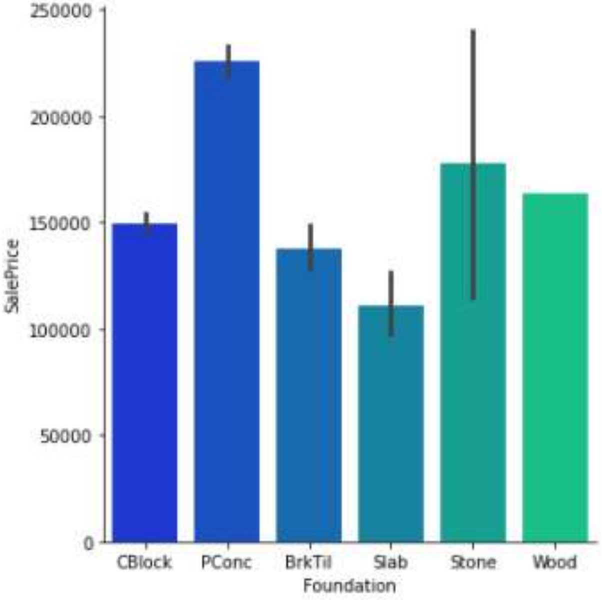
```
Figure size 720x432 with 0 Axes>
```



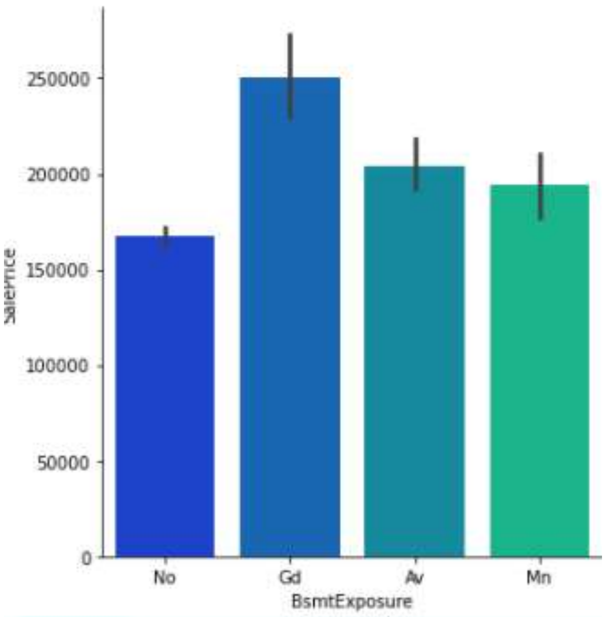
```
Figure size 720x432 with 0 Axes>
```



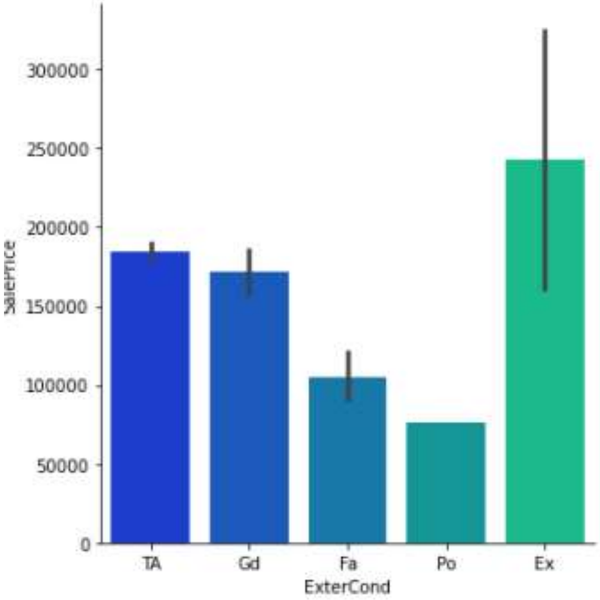
```
Figure size 720x432 with 0 Axes>
```



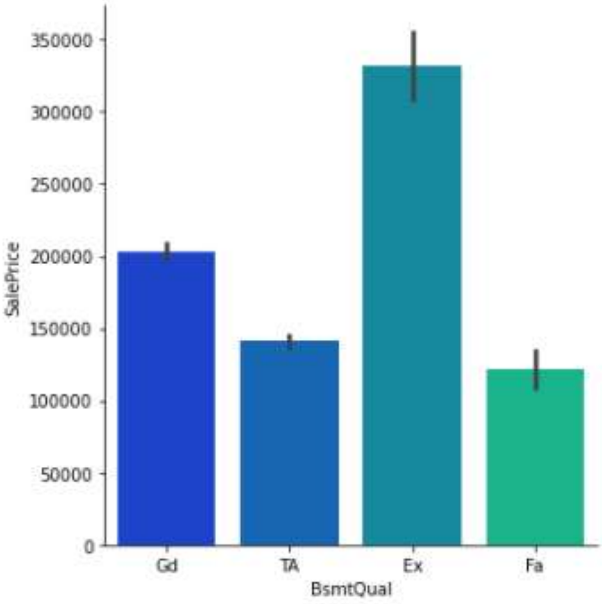
```
Figure size 720x432 with 0 Axes>
```

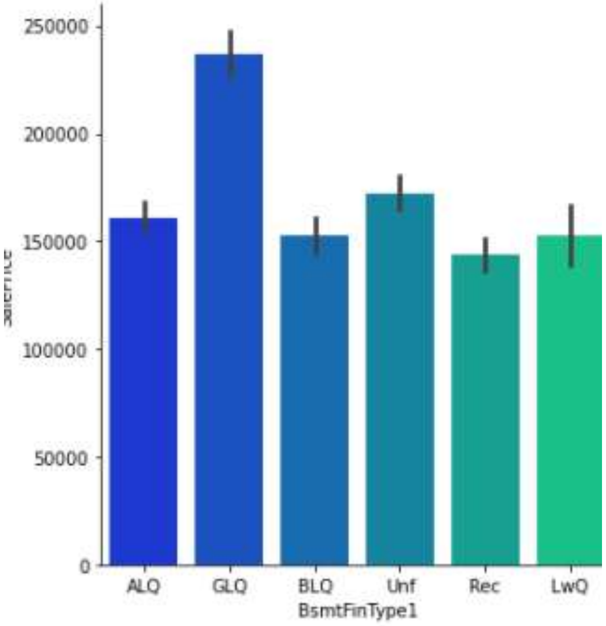


```
Figure size 720x432 with 0 Axes>
```

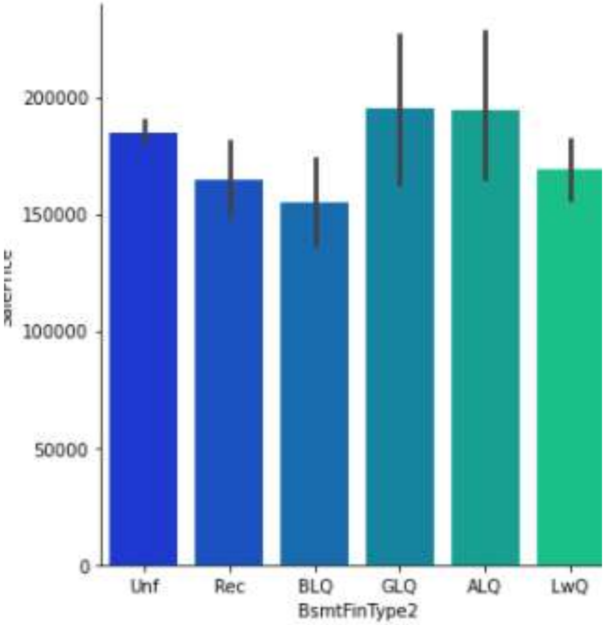


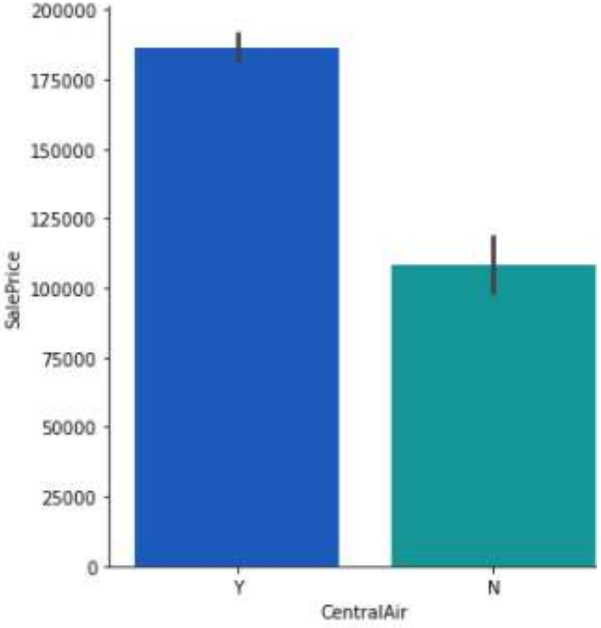
```
Figure size 720x432 with 0 Axes>
```



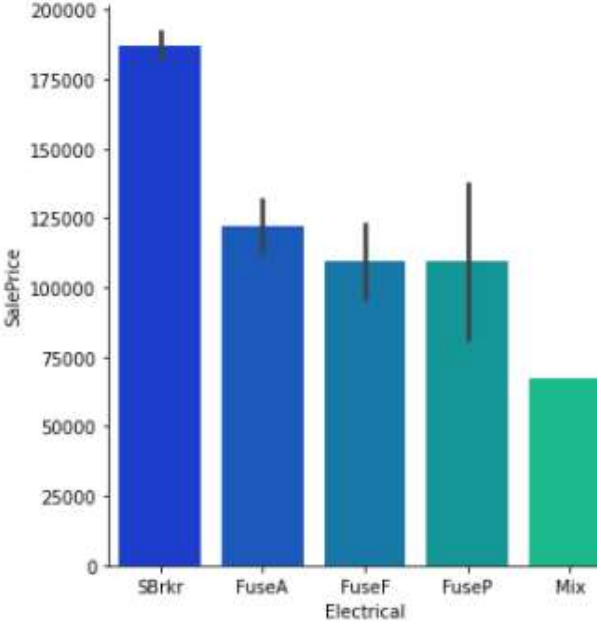


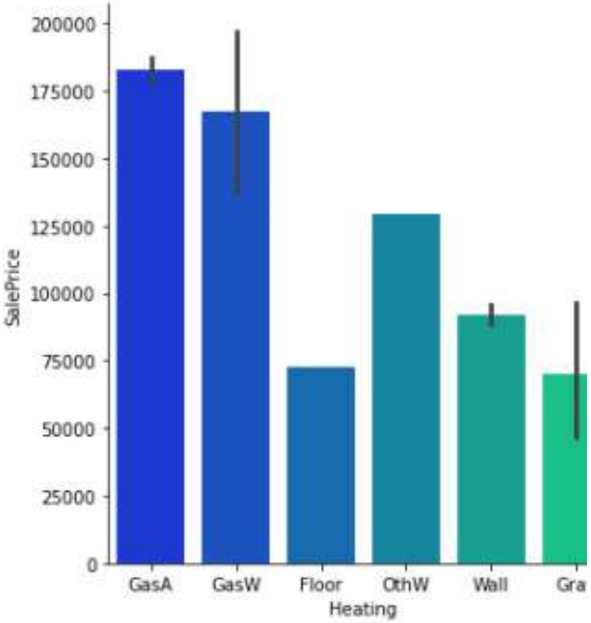
```
Figure size 720x432 with 0 Axes>
```



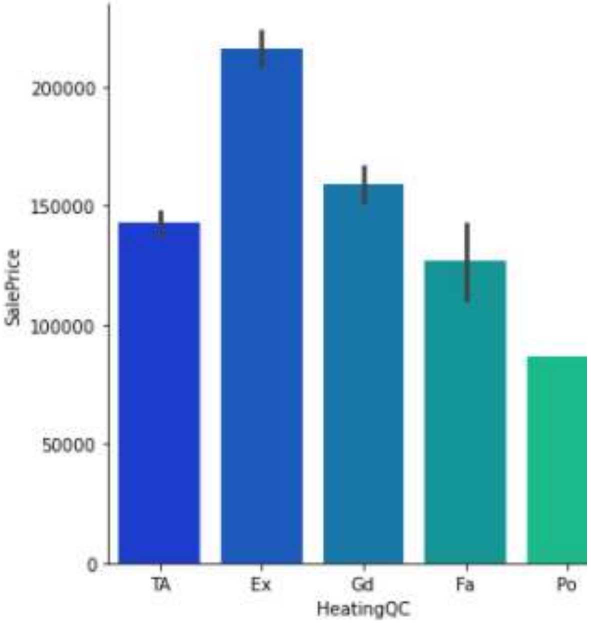


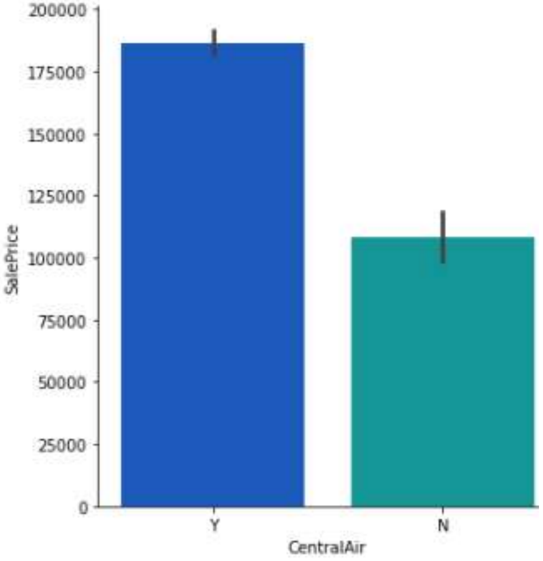
```
<Figure size 720x432 with 0 Axes>
```



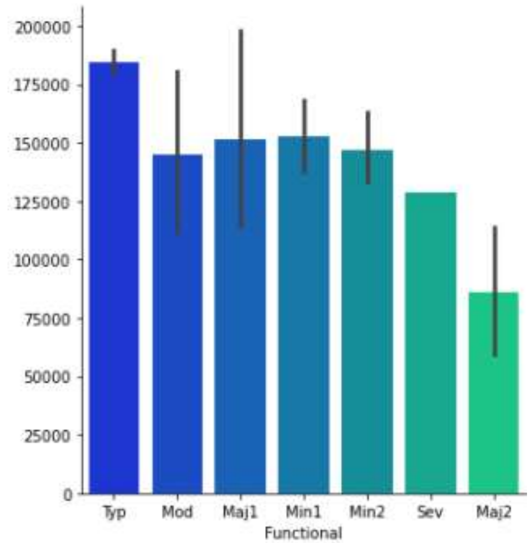


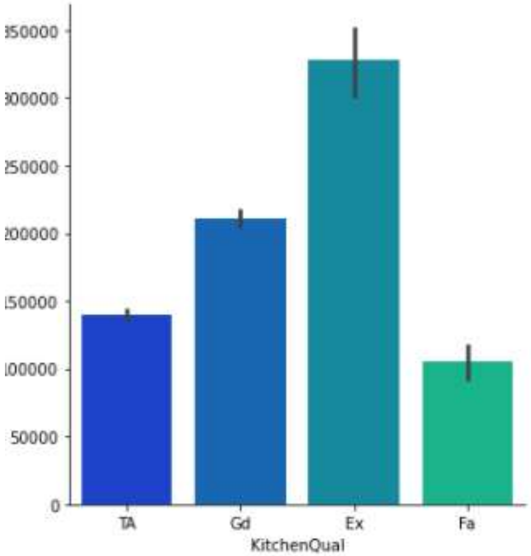
```
<Figure size 720x432 with 0 Axes>
```



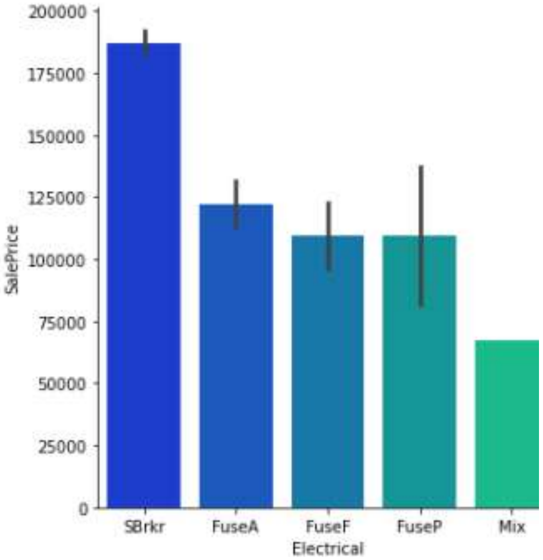


```
igure size 720x432 with 0 Axes>
```

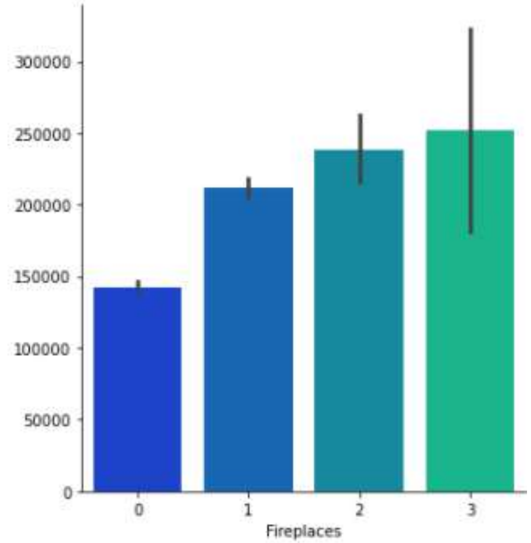




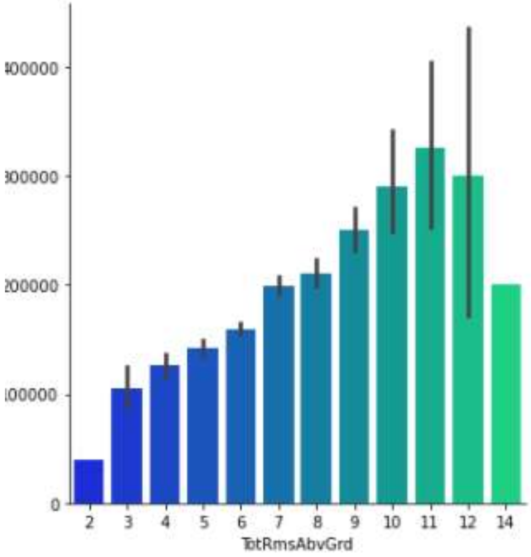
```
<Figure size 720x432 with 0 Axes>
```

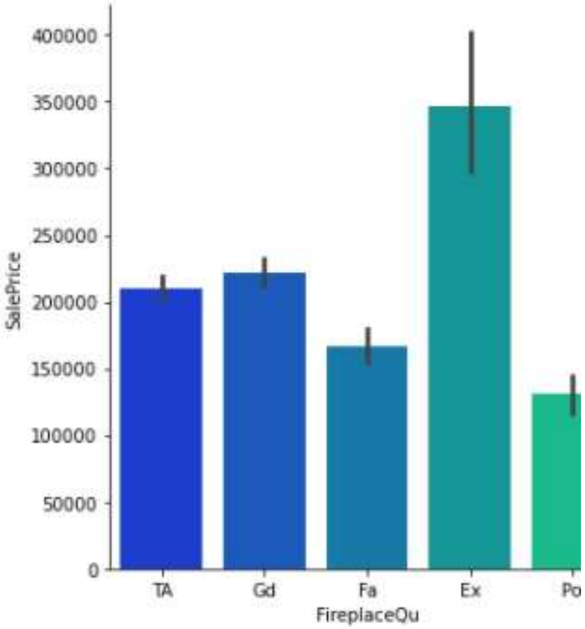


```
igure size 720x432 with 0 Axes>
```

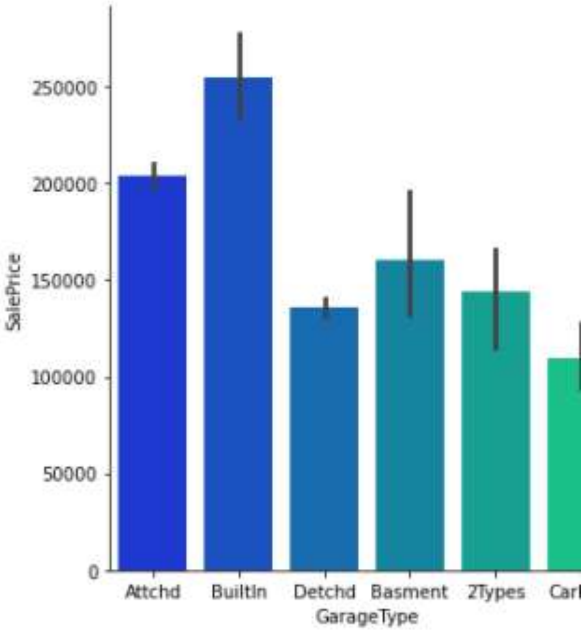


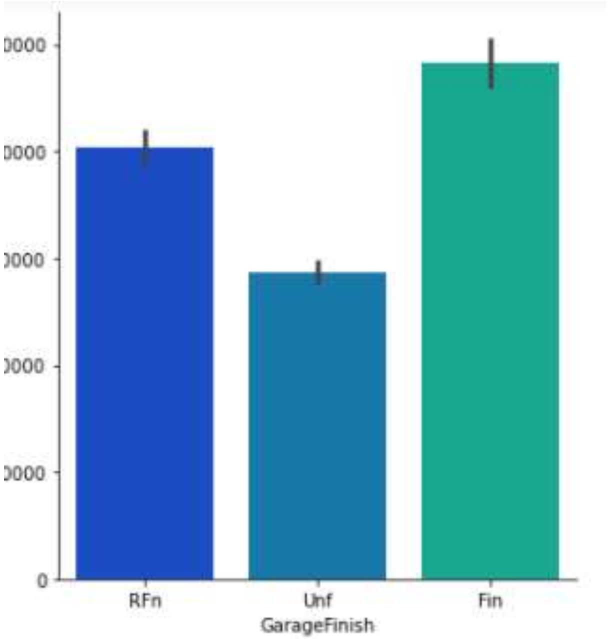
```
gure size 720x432 with 0 Axes>
```



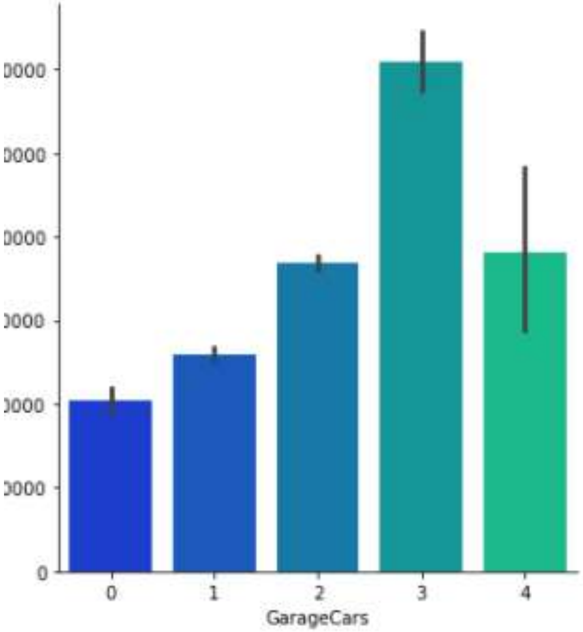


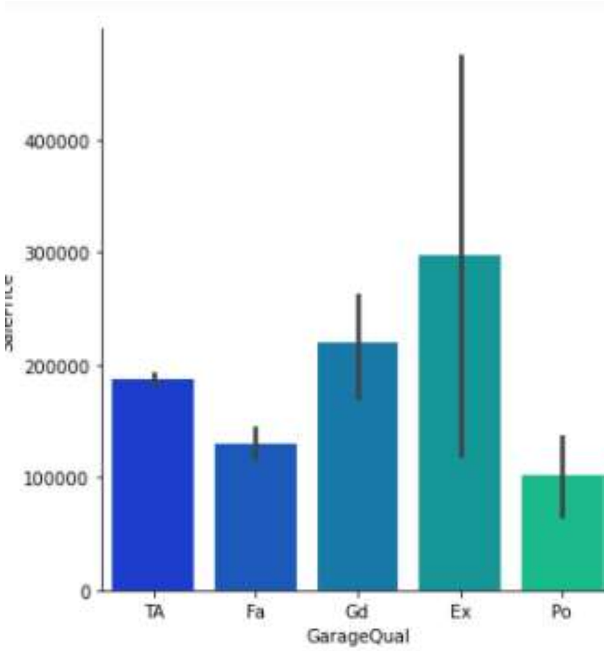
```
<Figure size 720x432 with 0 Axes>
```



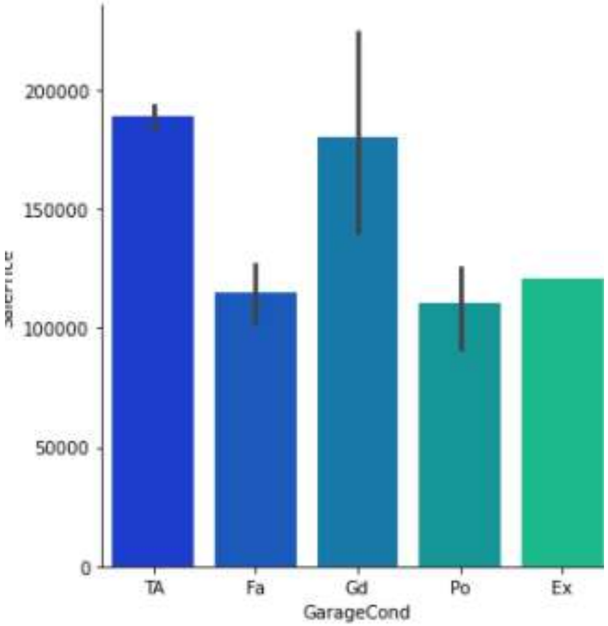


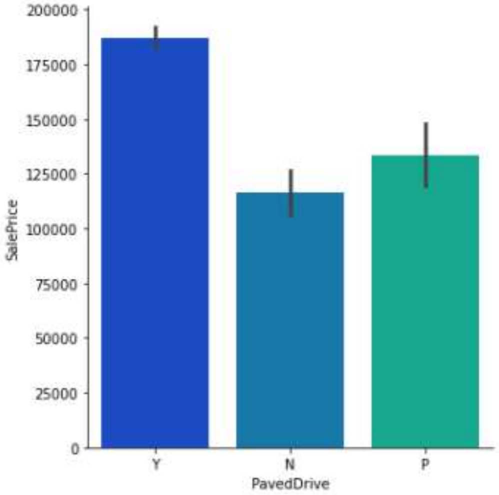
```
ure size 720x432 with 0 Axes>
```



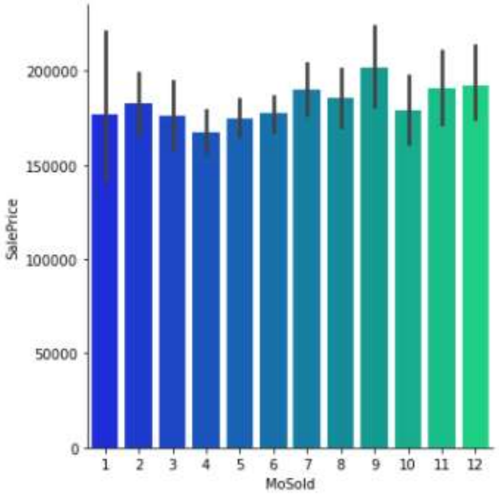


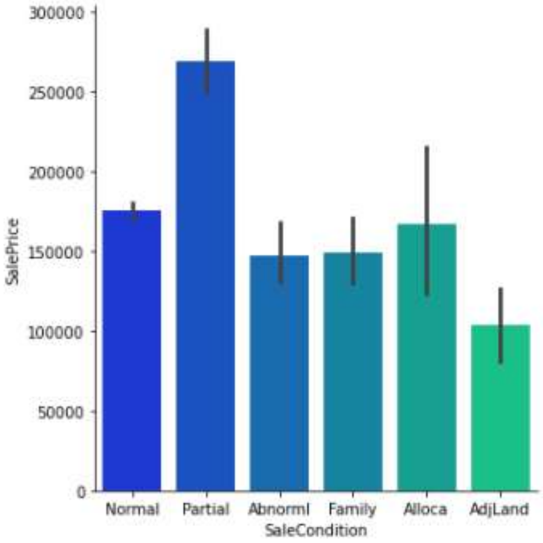
```
Figure size 720x432 with 0 Axes>
```



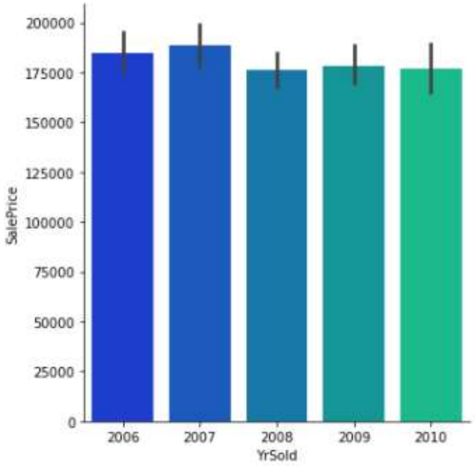

200000
175000
150000
125000
100000
75000
50000
25000
0
SalePrice
Y
N
P
PavedDrive

```
<Figure size 720x432 with 0 Axes>
```

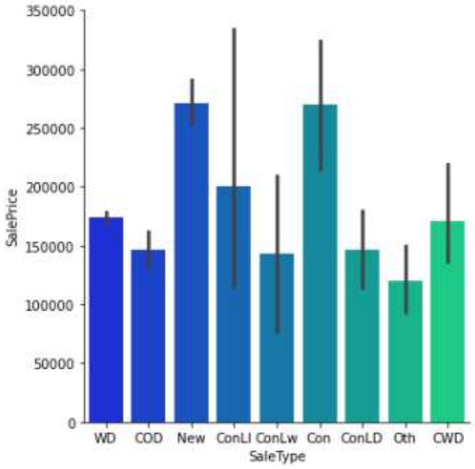

200000
150000
100000
50000
0
SalePrice
1
2
3
4
5
6
7
8
9
10
11
12
MoSold


300000
250000
200000
150000
100000
50000
0
SalePrice
Normal
Partial
Abnorml
Family
Alloca
AdjLand
SaleCondition

```
<Figure size 720x432 with 0 Axes>
```


200000
175000
150000
125000
100000
75000
50000
25000
0
SalePrice
2006
2007
2008
2009
2010
YrSold

```
<Figure size 720x432 with 0 Axes>
```


350000
300000
250000
200000
150000
100000
50000
0
SalePrice
WD
COD
New
ConLI
ConLw
Con
ConLD
Oth
CWD
SaleType

- **Interpretation of the Results**

- Least SalePrice is for 30:1-STORY 1945 & OLDER and maximum for 60:2-STORY 1946 & NEWER
- In MSZoing maximum is for category 1 i.e, Floating Village Residential
- Lotshape 1 and 2 have almost similar price and 3 has least.
- Landconotur corresponding to 1 i.e, HLS Hillside - Significant slope from side to side has maximum price.
- Lotconfig corresponding to 1 and 3 have similar price.
- Neighborhoot with (15)NPkVill Northpark Villa has maximum sales price and (10)IDOTRR Iowa DOT and Rail Road has least.
- Normal condition houses have highest saleprice
- 1Fam Single-family Detached and TwnhsI Townhouse Inside Unit have maximum saleprice.
- In HouseStyle category 3: 2Story Two story has max sale price.
- In OverallQual: SalePrice increase as Ratings increase.
- Similary for OverallCond 5 and 9 have max sale price
- In RoofStyle 5:Shed has maximum.
- In Exterior1st 6:HardBoard and 9:Other have Saleprice
- In Exterior2nd 8:MetalSd Metal Siding
- In MasVnrType, 3:stone has max saleprice and 0:BrkCmn Brick Common has least
- In ExterQual 0:Excellent has maximum price. Similary for ExterCond
- In Foundation 2:PConc Poured Contrete has max price
- In BsmtQual 0: Ex Excellent (100+ inches), In BsmtCond 1: Gd Good, In BsmtExposure 1: Av Average Exposure (split levels or foyers typically score average or above) have max sale prices
- In BsmtFinType1: Rating of basement finished area - 2:GLQ Good Living Quarters has max price
- In HeatingQC: Heating quality and condition 0:Ex Excellent has max price.
- Houses with CentralAir has higher saleprice
- In FireplaceQu: Fireplace quality 0:Ex Excellent - Exceptional Masonry Fireplace has max saleprice
- GarageType 3:BuiltIn Built-In (Garage part of house - typically has room above garage) has max saleprice
- Finished Garage has more price
- Paved Driveway has more price
- In 2007 maximum houses are sold followed by 2006
- In saletype category 2 and 6 have max sale price
- Normal sale condition has max price.

Conclusion

- **Key Findings and Conclusions of the Study**

- Lotshape 1 and 2 have almost similar price and 3 has least.
- Landconotur corresponding to 1 i.e, HLS Hillside - Significant slope from side to side has maximum price.
- Neighborhoot with (15)NPkVill Northpark Villa has maximum sales price and (10) IDOTRR Iowa DOT and Rail Road has least.
- Normal condition houses have highest saleprice
- 1Fam Single-family Detached and TwnhsI Townhouse Inside Unit have maximum saleprice.
- In HouseStyle category 3: 2Story Two story has max sale price.
- In RoofStyle 5: Shed has maximum.
- In Exterior1st 6: HardBoard and 9: Other have Saleprice
- In MasVnrType, 3: stone has max saleprice and 0: BrkCmn Brick Common has least
- Houses with CentralAir has higher saleprice
- GarageType 3: BuiltIn Built-In (Garage part of house - typically has room above garage) has max saleprice
- In 2007 maximum houses are sold followed by 2006
- In LotArea, initially the price keeps on increasing as LotArea increases but after 70000 it becomes constant till 160000 and then drops.
- In MasVnrArea, at 1200 saleprice is maximum and then it drops drastically.
- For 1stFlrSF:first floor square feet till 2500 the price is increasing uniformly but after that it decreases and drops after 3000
- For 2ndFlrSF: Second floor square feet the price is increasing as the area increases.

- **Learning Outcomes of the Study in respect of Data Science**

With the help of visualization tools such as matplotlib and seaborn we have visualized the impact of each attribute on our target variable. For cleaning the data and plotting outliers we have used distplot and boxplot and for removing outliers we have used zscore which is a statistical tool. At last, we got GradientBoostingRegressor as our best model.

- **Limitations of this work and Scope for Future Work**

The model is working well and we have performed hyperparameter tuning and we have concluded our project by choosing GradientBoostingRegressor as our best model.