



NAME OF THE PROJECT : Housing Price Prediction

Submitted by: Shirsath Vinayak M.

Internship No. 33

ACKNOWLEDGMENT

Used Kaggle for understanding & more analysis of project. Also used some you tube videos for better analysis.

INTRODUCTION

- **Business Problem Framing**

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.

- **Conceptual Background of the Domain Problem**

US-based housing company named **Surprise Housing** has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia

- **Review of Literature**

The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know:

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

- **Motivation for the Problem Undertaken**

By this we are going to predict the prices of houses considering the different types of features. Also we will find which feature is most affecting the label for model prediction.

Analytical Problem Framing

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

First we have to predict the prices of house. We have find out the shape of the data & found 1168 no. of rows & 81 columns. In that we have separated the columns in numerical & categorical columns & found 43 categorical columns & 38 numerical columns. There are so many null values are present in dataset. We have done the analysis features vs Selling Price of houses. Along with we have to find which features is affecting most for prediction of price.

- **Data Sources and their formats**

Data source is in the form of csv file. Two separate data files train & test csv files are available. The train data is having 81 columns. We need to predict the selling price of the house.

- **Data Preprocessing Done(Steps for data cleaning & assumptions):**

First of all we have separated the numerical & categorical columns. Numerical columns are 38 numbers & 43 columns are categorical columns. After that we have find out the null values present in the dataset & found huge null values in Alley, PoolQC, MiscFeatures, Fence hence I decided to remove those columns due to high number of missing values. Afterthat , I describe the numerical dataset to find the min, max, mean, outliers present in dataset & Plotted histplot, Boxplot for univariate analysis. Before treating I observed that so many outliers are present in data hence I decided to remove the outliers from data & then after filling the null values because if we fill numerical values with mean directly then due to outlier it will show some higher side mean this will impact on model training & prediction. Hence it is better to remove outliers first using IQR(Inter Quantile range). After removal of outlier I have filled numerical null values with mean method & categorical columns with mode method.

- **Data Inputs- Logic- Output Relationships**

Data is in the csv format. It is 81 columns one is label i.e selling price of house. Finding the correlation between each feature with bivariate analysis to validate which feature is contributing more. Each & every feature is contributing to finding the output (Selling Price). We need to find which feature is most important or prediction of selling price of the house. We have to do various combinations of data visualisation, histplot, boxplot, boxplot.

- **State the set of assumptions (if any) related to the problem under consideration**

Finding the correlation along with features & label.

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

Libraries used Seaborn, Matplotlib, Boxplot, Heatmap.

Model/s Development and Evaluation

- **Identification of possible problem-solving approaches (methods)**

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

First, I found the shape of the data. After that find the info of data in which I know that there are 81 nos. of columns & various rows also it is observed that there are null values are also present in some columns. Hence I have to deal with the outliers first before filling the null values because if I directly filled the null values that will impact on the mean of the data due to the outliers present in the dataset. While describing the data I found skewness are present in the data. I took various combinations of features & labels. Plotted histplot, boxplot in order to find outliers. After analysis of the data I removed the outliers using IQR

(Inter Quantile range). After removal of outliers label encoding is done for categorical columns & plotted heatmap in order to find the correlation ship between features. Then after train test split & model training. Model training done on linear regression, random forest, Knn regressor, Ada boost regressor. All the models are having good accuracy score hence I have chosen random forest as the good score & save this model in pickle.

- **Testing of Identified Approaches (Algorithms)**

Listing down all the algorithms used for the training and testing.

- a. Linear Regression is used for testing which has accuracy score of 86%, R2 score: 86%.
- b. Random Forest: - Which has accuracy score : 85%, R2 score: 85%.
- c. Ada Boost regressor : Which accuracy score is 89% & R2 score :82%.
- d. Knn Regressor: Accuracy score: 80% & R2 score : 80%

- **Run and Evaluate selected models**

Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

A. Linear regression Model Snaps:-

```
In [45]: x_train, x_test, y_train, y_test = train_test_split(x_scaled, y, test_size=0.25, random_state=1)

Initialising model for prediction Linear Regression

In [46]: lr = LinearRegression()
lr.fit(x_train, y_train)

Out[46]: LinearRegression()

In [47]: #Checking training score
lr.score(x_train, y_train)

Out[47]: 0.8994935825997183

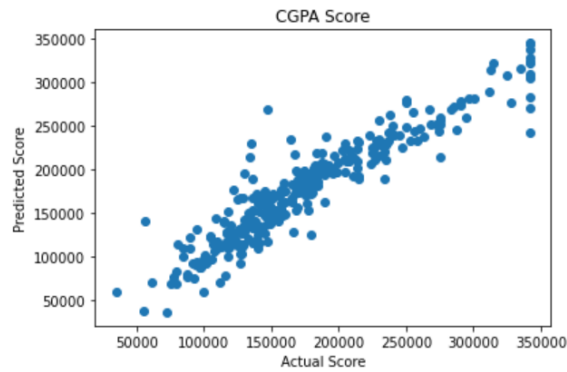
In [48]: #Checking testing score
lr.score(x_test, y_test)

Out[48]: 0.8683239264841851

In [49]: #plot to visualize data
y_pred = lr.predict(x_test)
y_pred
```

```
In [50]: #Plotting the graph of Actual score Vs Predicted score
```

```
plt.scatter(y_test, y_pred)
plt.xlabel("Actual Score")
plt.ylabel("Predicted Score")
plt.title("CGPA Score")
plt.show()
```



```
In [51]: from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
```

```
In [52]: y_pred= lr.predict(x_test)
```

```
In [53]: #Mean square error
mean_squared_error(y_test, y_pred)
```

Out[53]: 16596.183446446517

```
In [54]: #MSE
mean_squared_error(y_test, y_pred)
```

Out[54]: 16596.183446446517

```
In [55]: #RMSE
np.sqrt(mean_squared_error(y_test, y_pred))
```

Out[55]: 23858.070435352936

```
In [56]: r2_score(y_test, y_pred)
```

Out[56]: 0.8683239264841851

```
In [57]: conclusion=pd.DataFrame([lr.predict(x_test)[:],y_test[:]],index=["Predicted", "Original"])
conclusion
```

Out[57]:

	0	1	2	3	4	5	6	7	8	9	...	282
Predicted	218928.6847	153775.098256	184147.053893	69722.33645	233060.145145	190215.405614	116398.829332	110475.138854	92908.866156	77156.557886	...	218844.674717
Original	184000.0000	147500.000000	159000.000000	112000.00000	258000.000000	183200.000000	141000.000000	128000.000000	92000.000000	88000.000000	...	187500.000000

2 rows x 292 columns

Random Forest Codes:

Random Forest

```
In [58]: from sklearn.ensemble import RandomForestRegressor

regressor = RandomForestRegressor(n_estimators=20, random_state=0)
regressor.fit(x_train, y_train)
y_pred = regressor.predict(x_test)
```

```
In [59]: y_pred
```

```
Out[59]: array([[198275.25, 149882. , 158470. , 97509.1 , 288772. ,
180210. , 137798.4 , 135039.15, 89155. , 105660. ,
204556.05, 199262.5 , 158800. , 268756.75, 140400. ,
203342.95, 78010. , 103782.5 , 165431.7 , 260766.3 ,
78481.1 , 148771.85, 179100. , 216692. , 282214.9 ,
176709.25, 211207.5 , 137100. , 223315.75, 106892.5 ,
146055. , 169965. , 177255. , 244949.1 , 228062.5 ,
163265. , 247245. , 321658.4 , 199704.05, 212476.2 ,
116684. , 139310. , 163929.5 , 314500.275, 158593.35 ,
135952.75, 305400.85, 120502.5 , 188280. , 202305. ,
132565. , 115185. , 166437.5 , 129950. , 120514.35 ,
111969.15, 312508.225, 141607.65, 213170. , 188293. ,
157250. , 150920. , 159898.85, 279845.625, 151700. ,
130872.5 , 146627.5 , 159765. , 92120. , 83665. ,
274991.8 , 155307.5 , 158185. , 317591.875, 149342.5 ,
138950. , 153725. , 159016.25, 127375.3 , 218381.25 ,
100925.8 , 136327.5 , 120510. , 133812.5 , 201753.75 ,
88755. , 104062.5 , 234416.4 , 228853.25, 185439.75 ,
128381.6 , 253237.8 , 86684.15, 132135. , 177845. ,
218393.7 , 164275. , 264566.1 , 122525.4 , 234888.15 ,
150091.95, 111325. , 120415. , 159480. , 148322.65 ,
229908.2 , 119735. , 103220. , 169600. , 228711.25 ,
```

```
In [60]: #Training accuracy score

regressor.score(x_train, y_train)
```

```
Out[60]: 0.9806934027258574
```

```
In [61]: #Testing accuracy score

regressor.score(x_test, y_test)
```

```
Out[61]: 0.857217773364923
```

```
In [62]: #Mean square error

mean_absolute_error(y_test, y_pred)
```

```
Out[62]: 17387.08219178082
```

```
In [63]: #MSE

mean_absolute_error(y_test, y_pred)
```

```
Out[63]: 17387.08219178082
```


```
In [64]: #RMSE

np.sqrt(mean_squared_error(y_test, y_pred))
```

```
Out[64]: 24843.853045239128
```

```
In [65]: #R2 Score

r2_score(y_test, y_pred)
```

 Type here to search

Life is On | Schneider




```
In [65]: #R2 Score
r2_score(y_test, y_pred)

Out[65]: 0.8572177773364923

In [66]: conclusion=pd.DataFrame([regressor.predict(x_test[:]),y_test[:]],index=["Predicted", "Original"])
conclusion
```

	0	1	2	3	4	5	6	7	8	9	...	282	283	284	285	286	287
Predicted	198275.25	149882.0	158470.0	97509.1	288772.0	180210.0	137798.4	135039.15	89155.0	105660.0	...	206635.0	168375.0	321243.125	184479.5	117553.8	238486.15
Original	184000.00	147500.0	159000.0	112000.0	258000.0	183200.0	141000.0	128000.00	92000.0	88000.0	...	187500.0	175000.0	341937.500	196000.0	117000.0	341937.50

2 rows x 292 columns

Ada Boost Regressor:

Ada Boost Regressor

```
In [67]: from sklearn.ensemble import AdaBoostRegressor

In [68]: ada=AdaBoostRegressor()
ada.fit(x_train, y_train)

Out[68]: AdaBoostRegressor()

In [69]: #Model Prediction on train data
y_pred=ada.predict(x_train)

In [70]: accuracy=metrics.r2_score(y_train, y_pred)
print("R Square Score" , accuracy)

R Square Score 0.8946631003824975

In [71]: #Model Prediction on test data
y_pred=ada.predict(x_test)

In [72]: accuracy=metrics.r2_score(y_test, y_pred)
print("R Square Score" , accuracy)

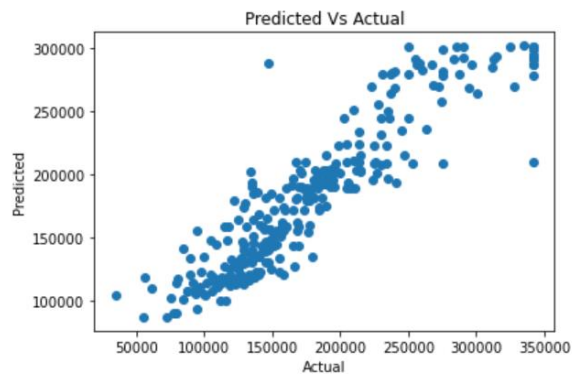
R Square Score 0.8278510445669562

In [73]: #Plotting the graph of Actual score Vs Predicted score
```

```

plt.ylabel('Predicted')
plt.title("Predicted Vs Actual")
plt.show()

```



```

In [74]: #Mean square error
mean_absolute_error(y_test, y_pred)

```

Out[74]: 19709.297880347665

```

In [75]: #MSE
mean_absolute_error(y_test, y_pred)

```

Out[75]: 19709.297880347665

```

In [75]: #MSE
mean_absolute_error(y_test, y_pred)

```

Out[75]: 19709.297880347665

```

In [76]: #RMSE
np.sqrt(mean_squared_error(y_test, y_pred))

```

Out[76]: 27279.35380170681

```

In [77]: r2_score(y_test, y_pred)

```

Out[77]: 0.8278510445669562

```

In [78]: #cross validation is to check whether the model is overfitting
from sklearn.model_selection import KFold, cross_val_score

```

```

In [80]: cv_score=cross_val_score(ada, x_scaled, y, cv=4)
cv_score

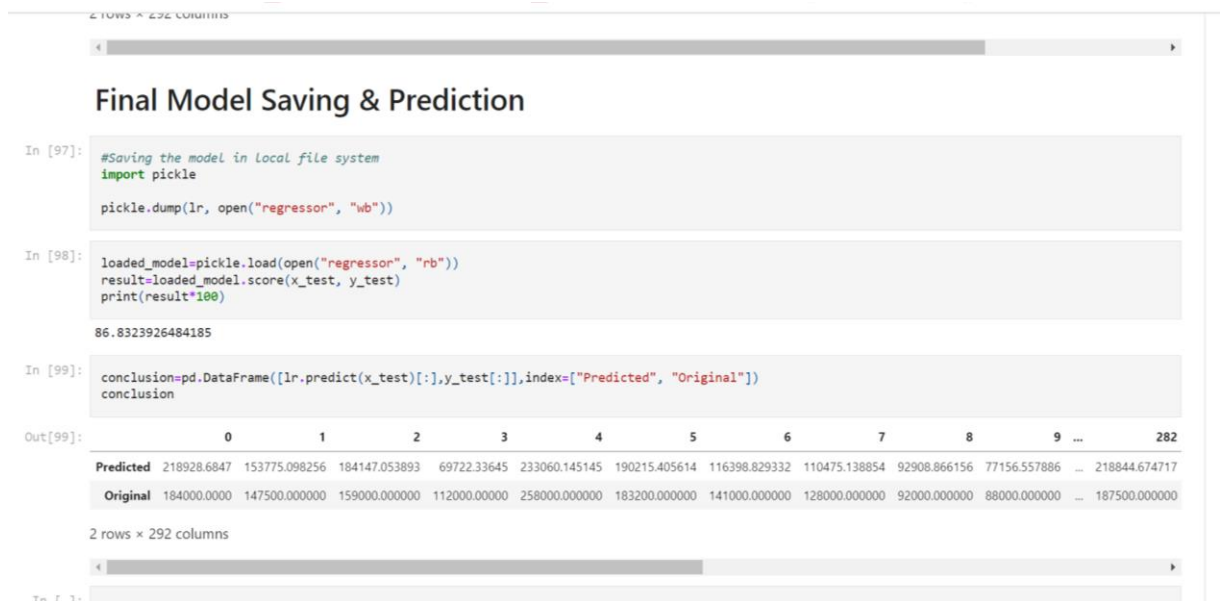
```

Out[80]: array([0.84587576, 0.84535161, 0.83351295, 0.82371406])

```

In [81]: cv_mean=cv_score.mean()
cv_mean

```



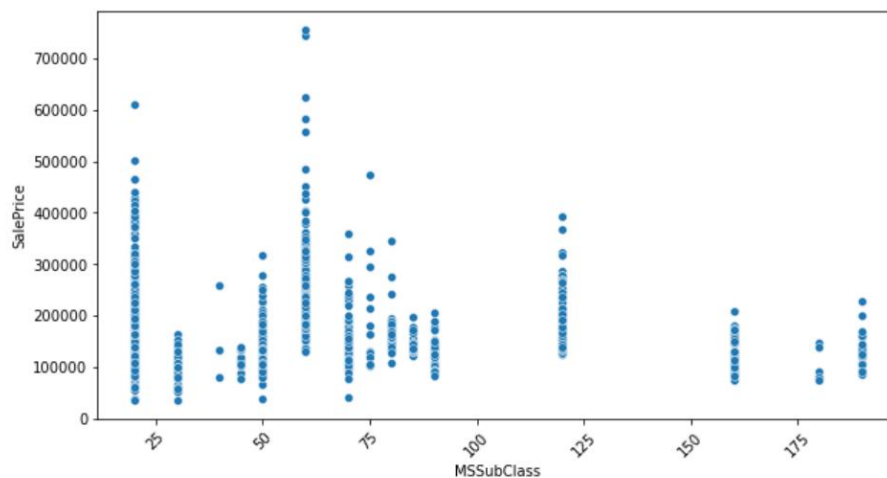
- Key Metrics for success in solving problem under consideration

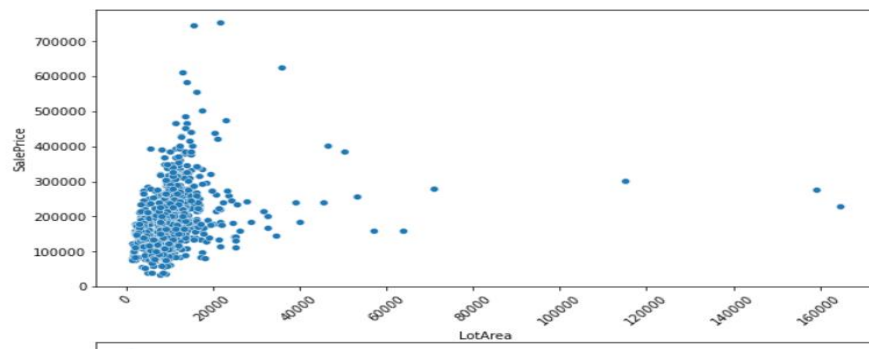
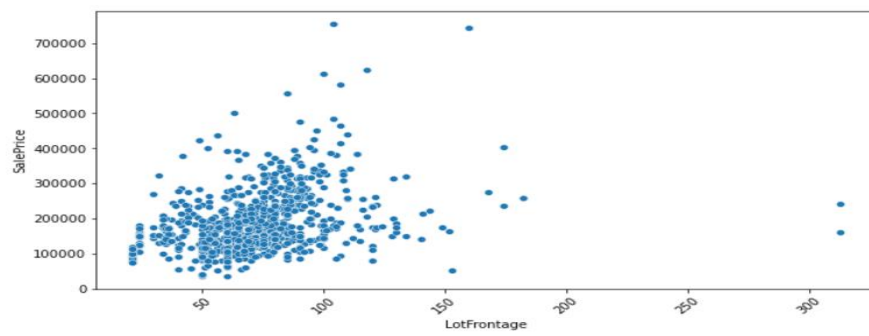
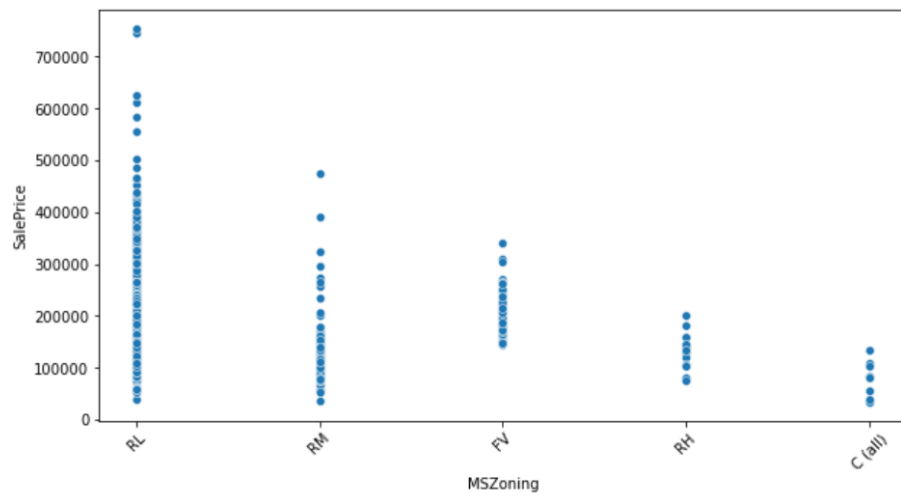
What were the key metrics used along with justification for using it?
You may also include statistical metrics used if any.

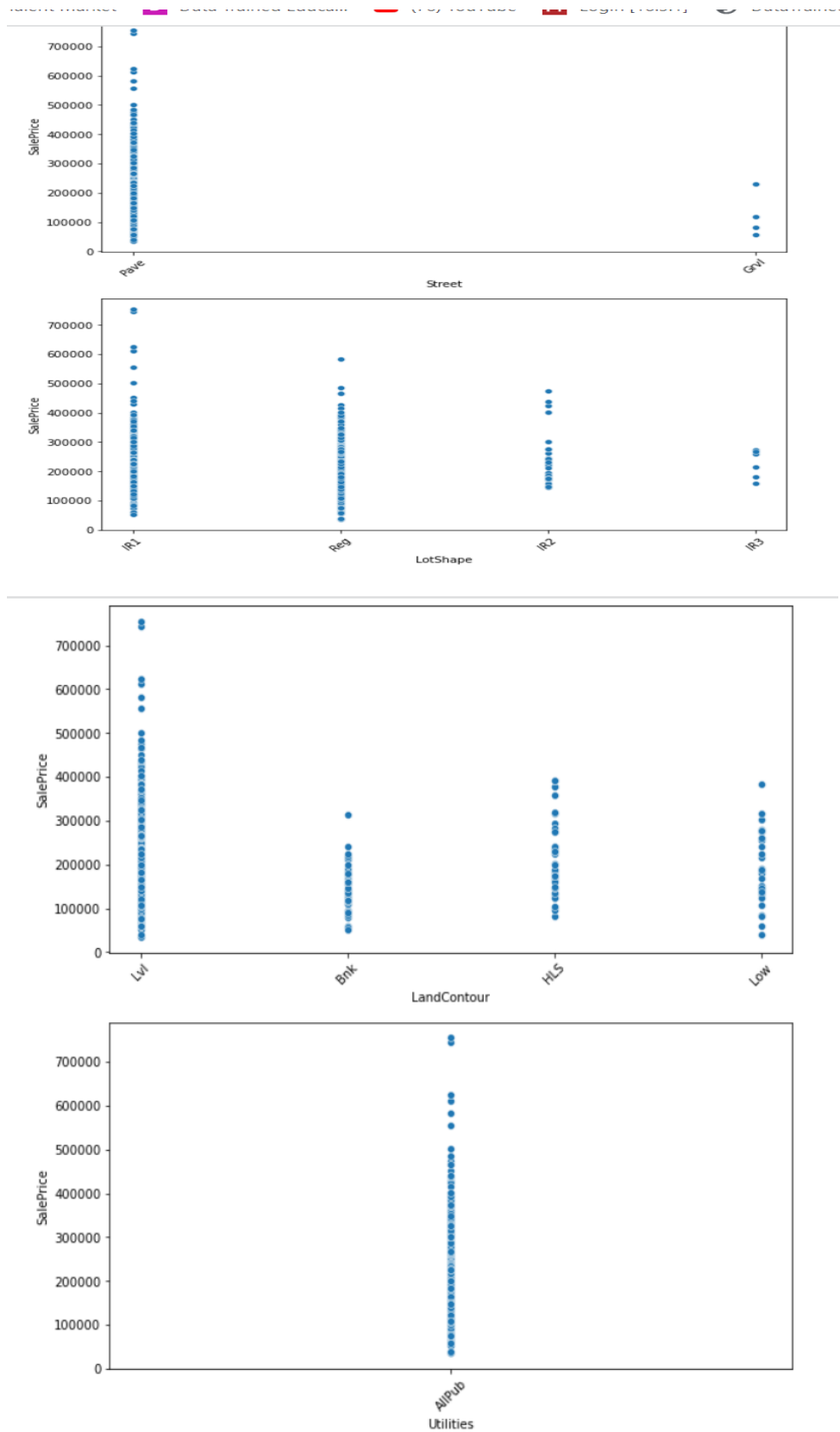
- Visualizations

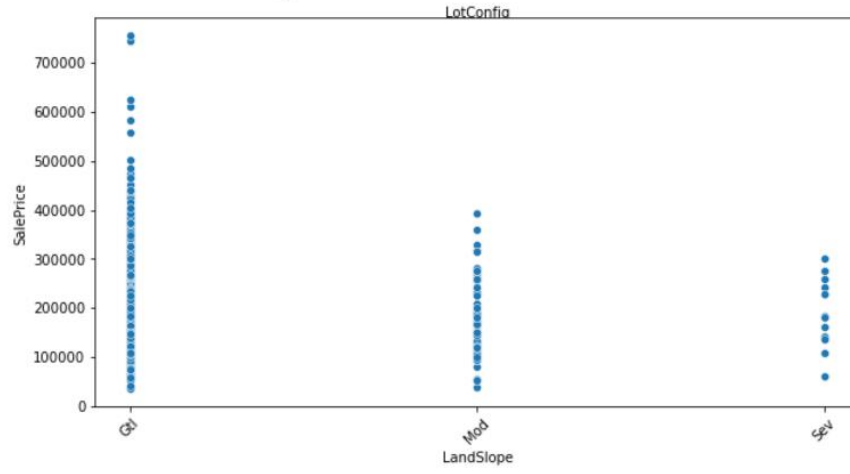
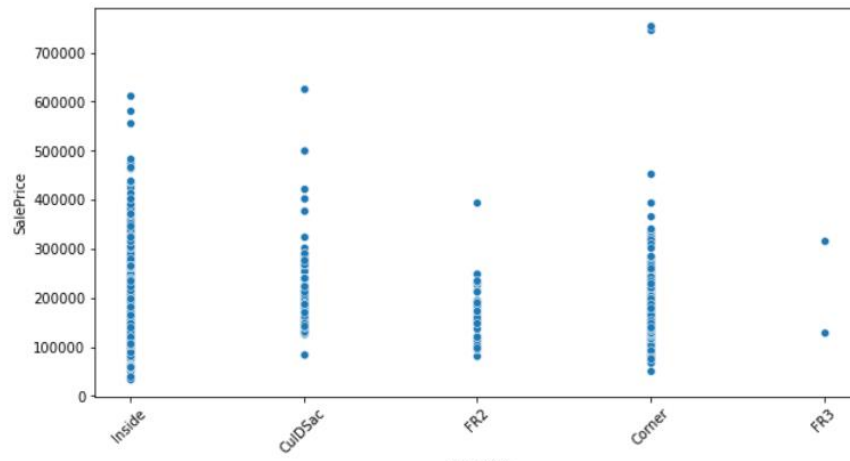
Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

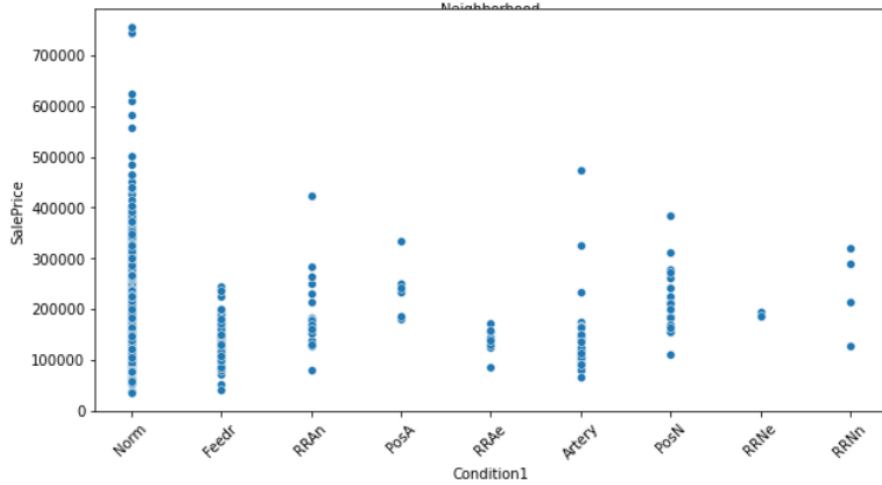
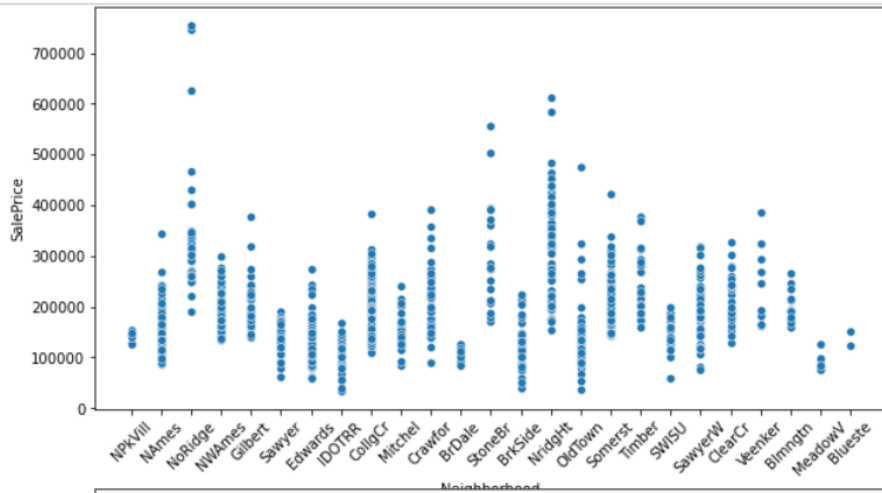
If different platforms were used, mention that as well.

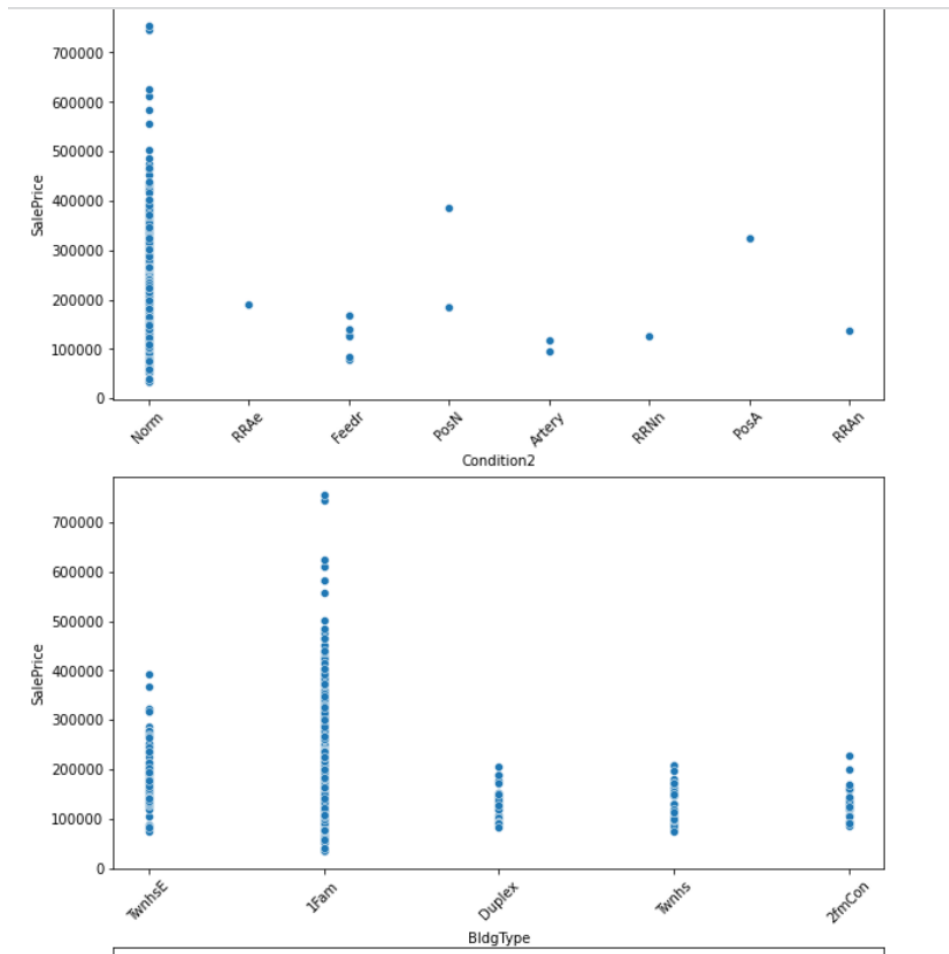


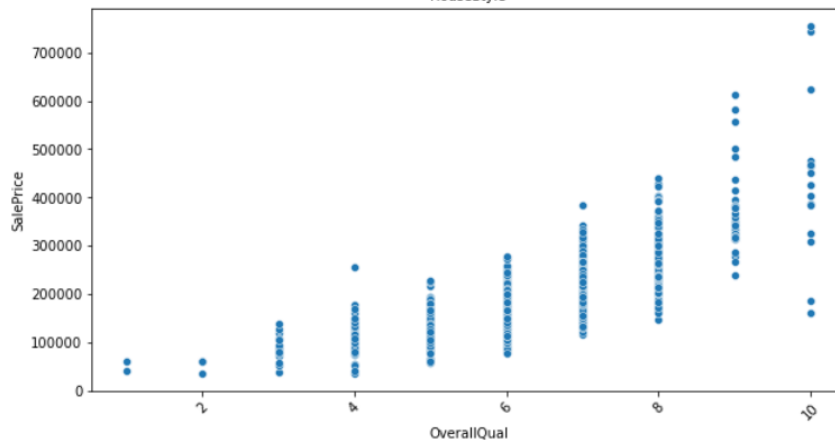
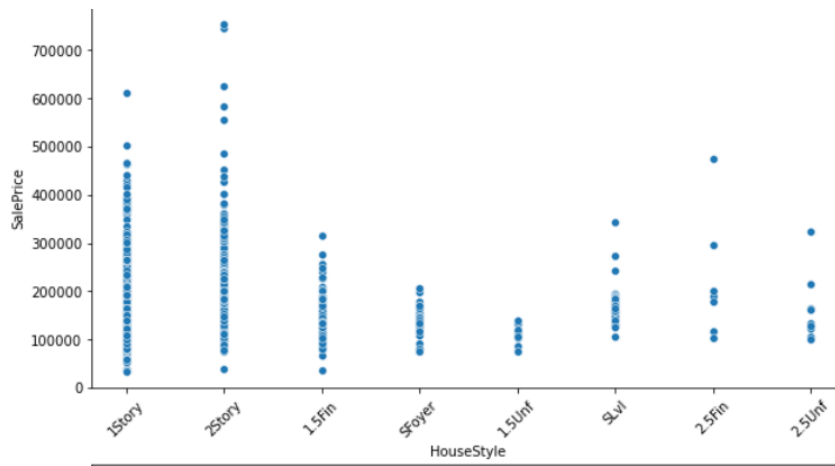


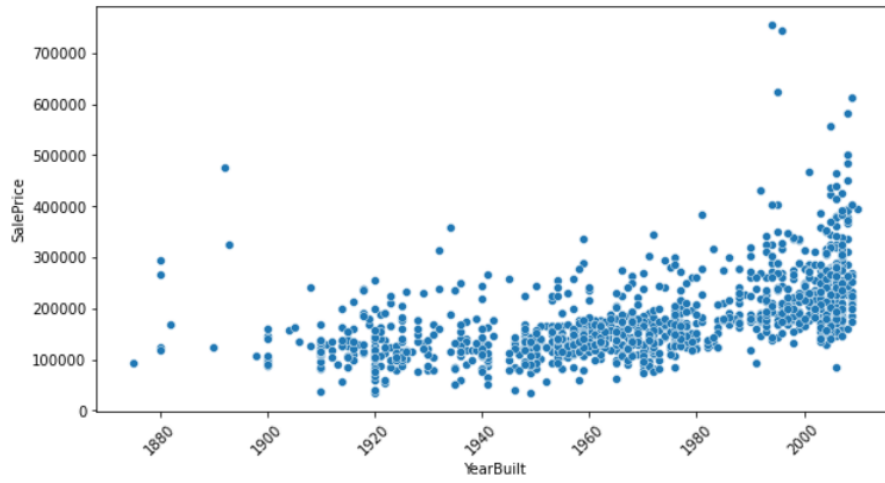
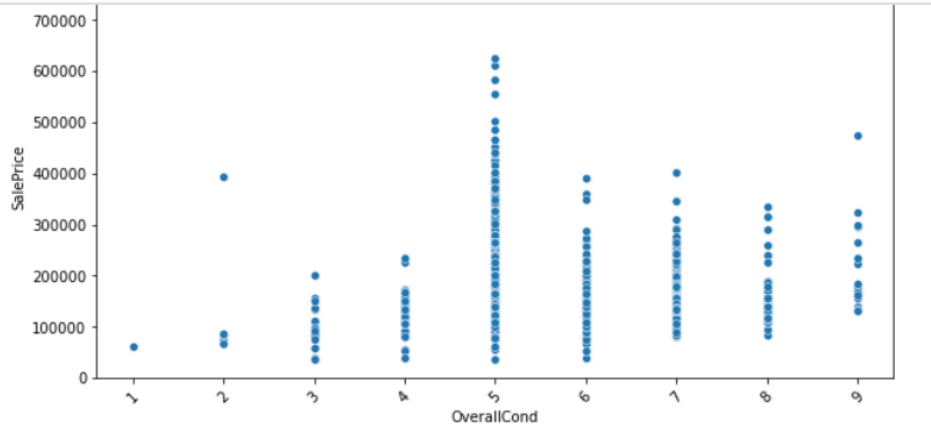


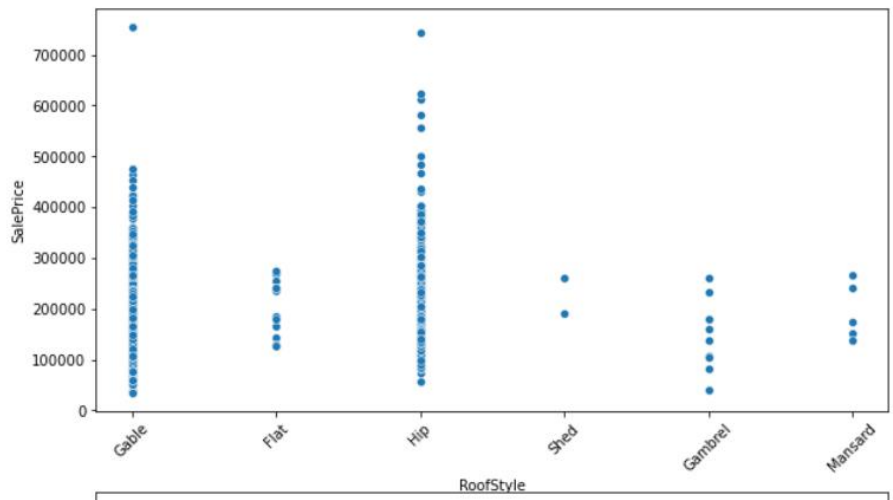
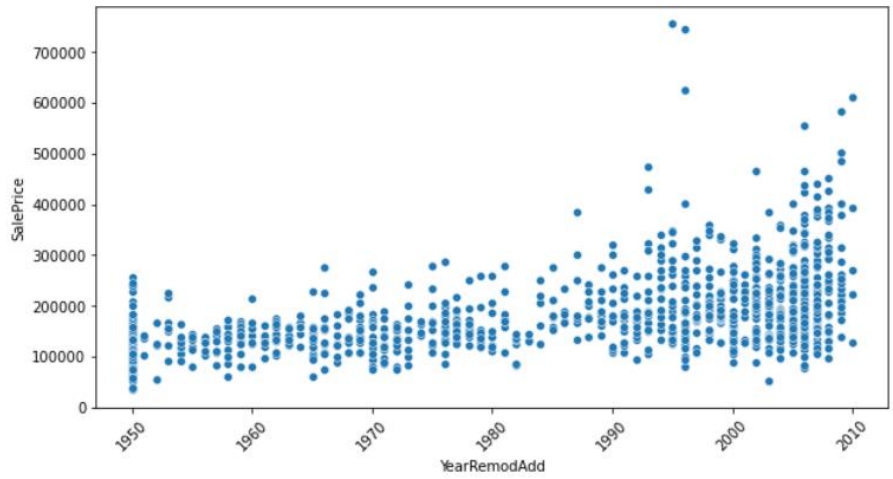


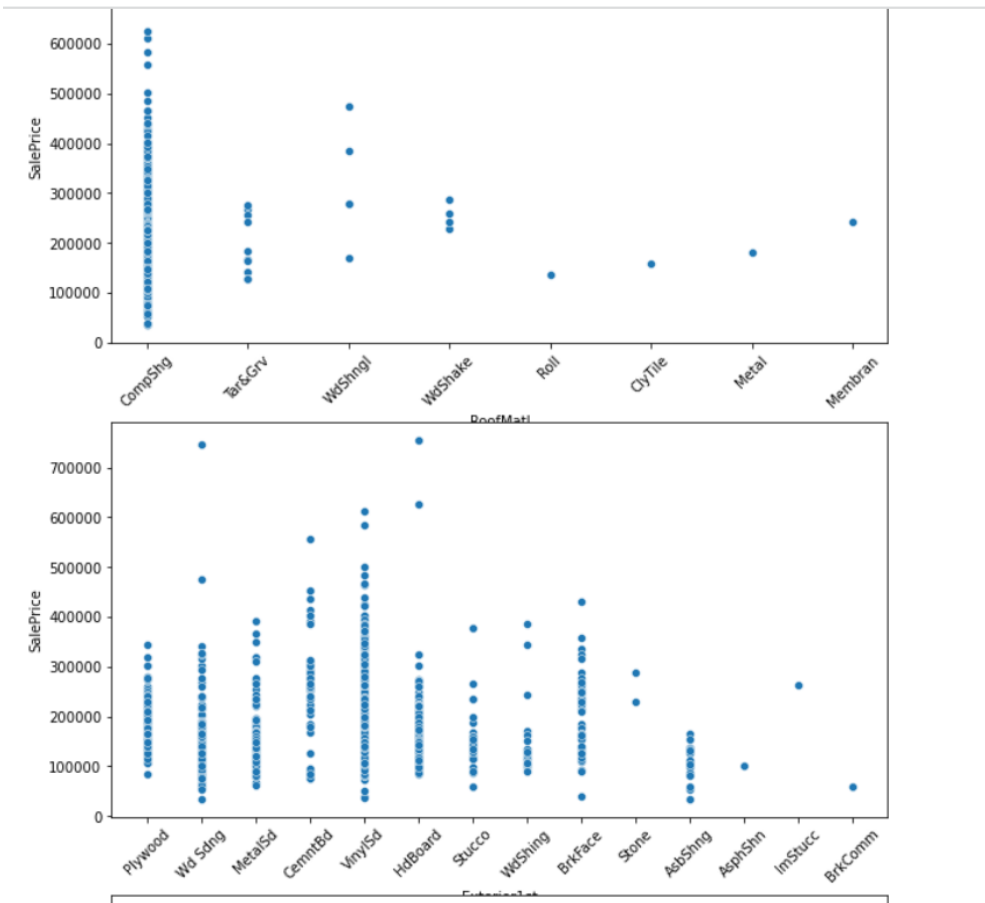


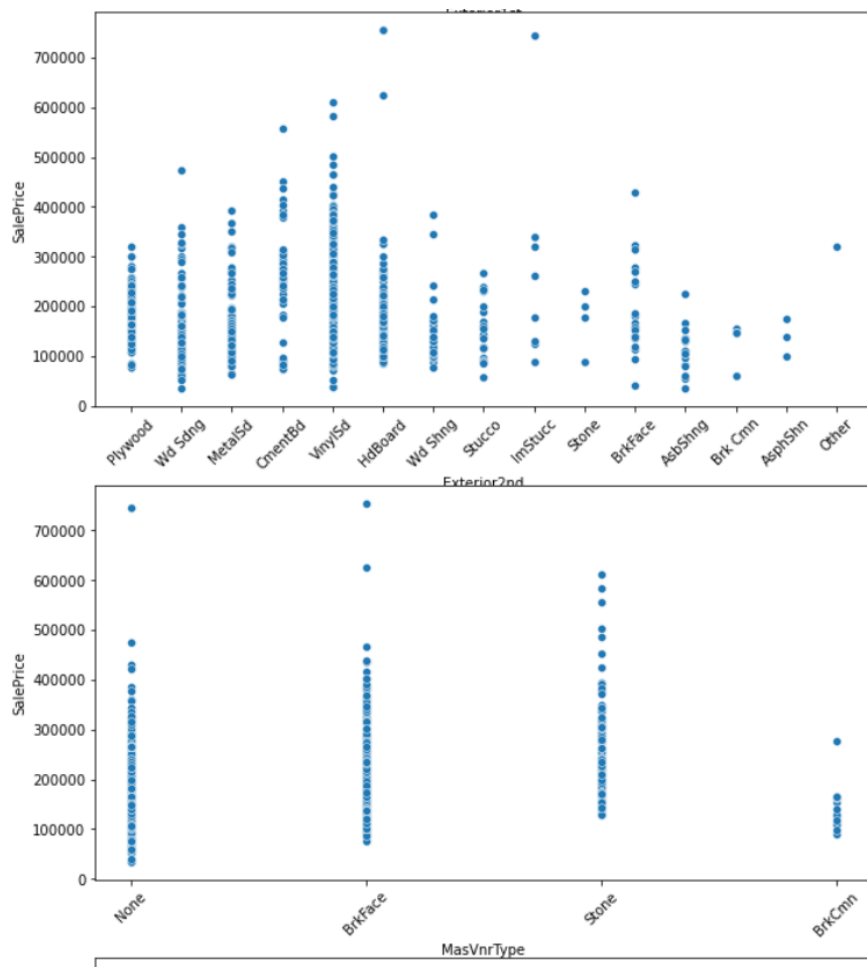


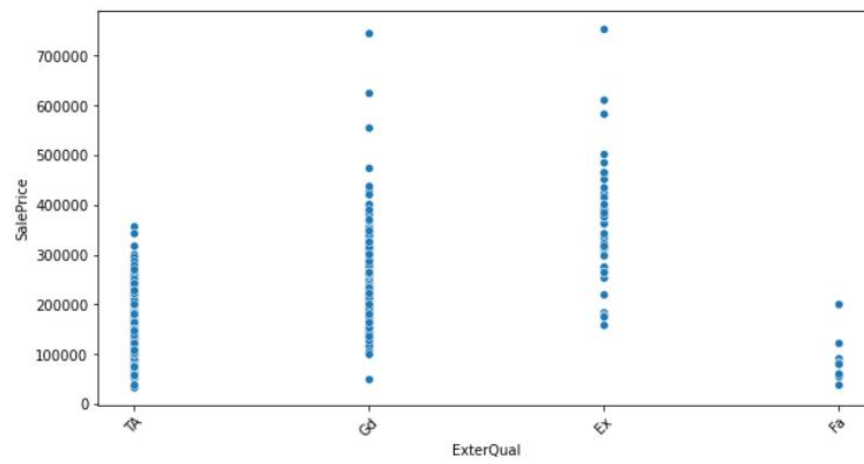
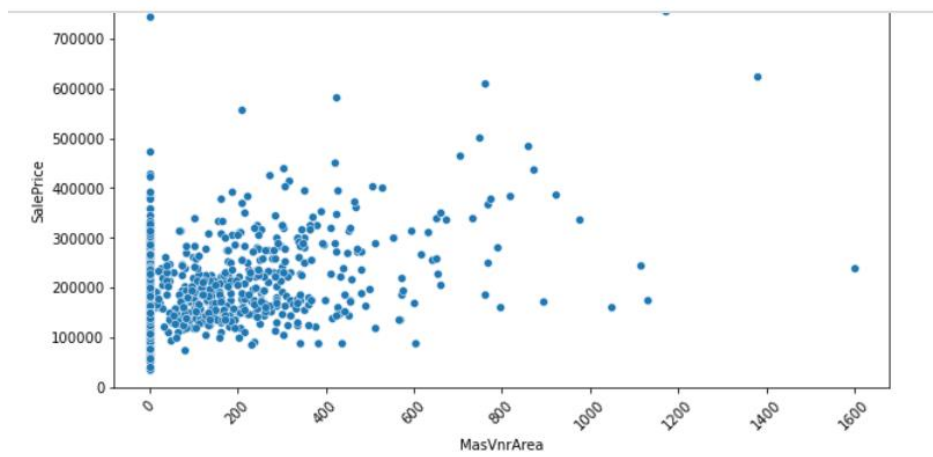


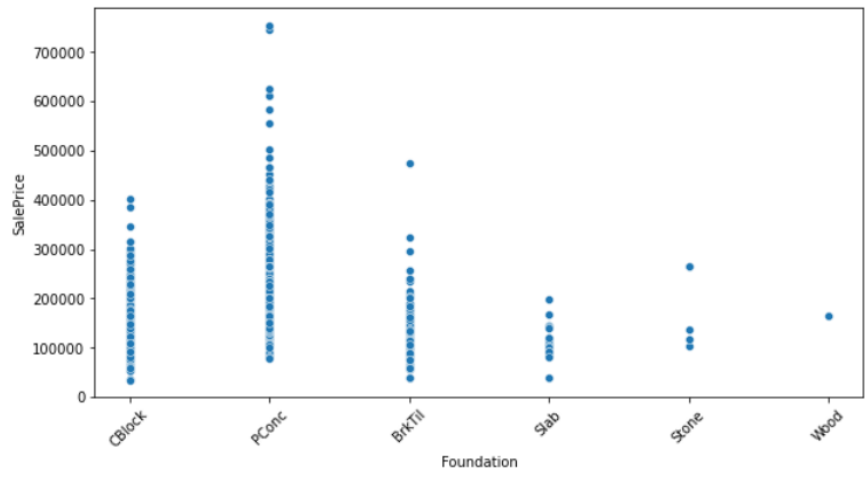
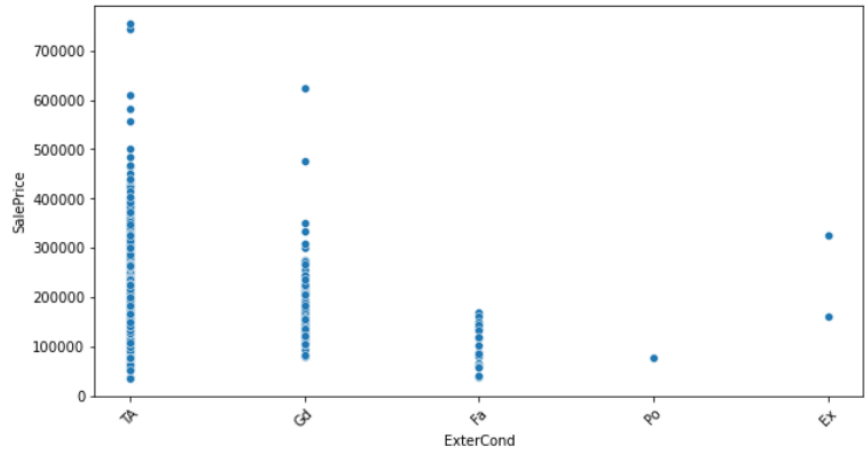


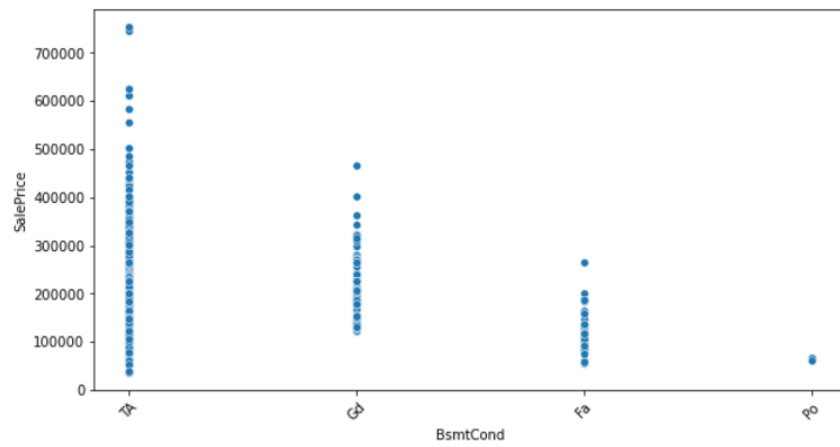
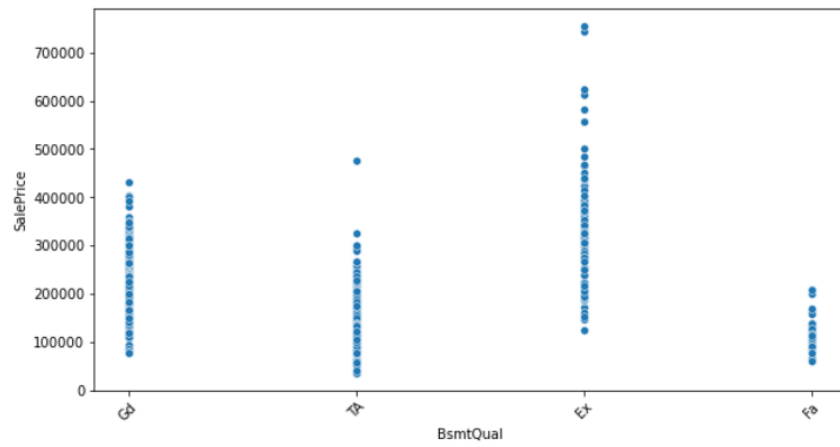


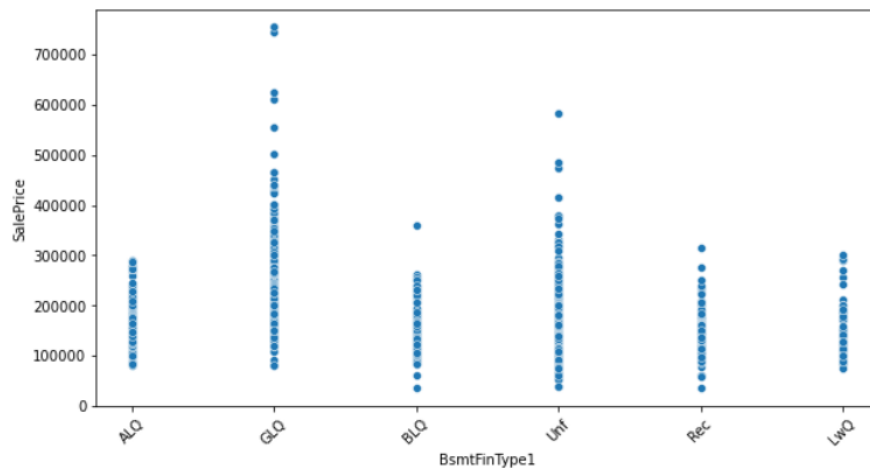
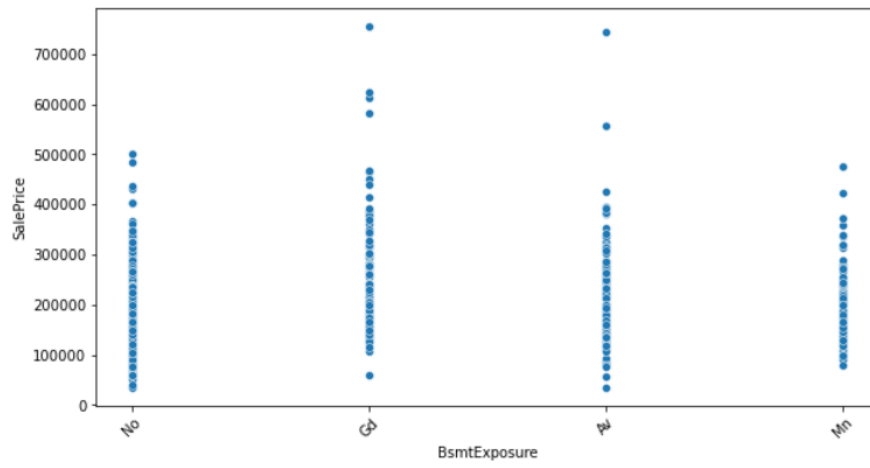


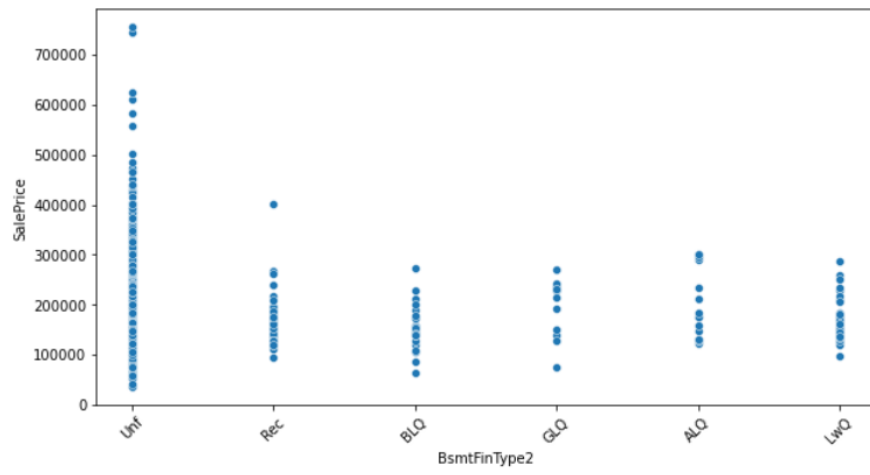
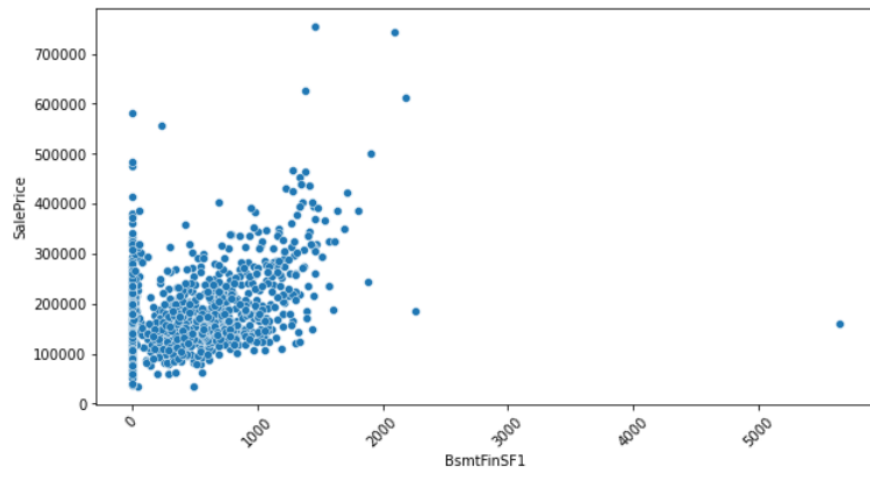


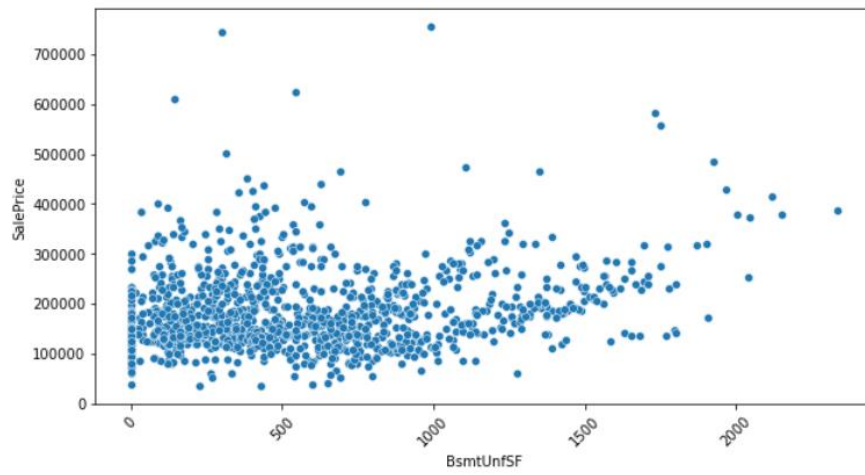
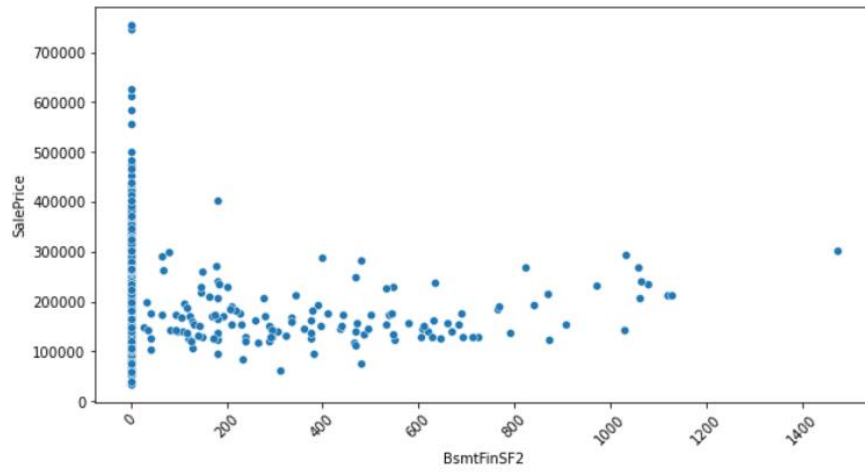


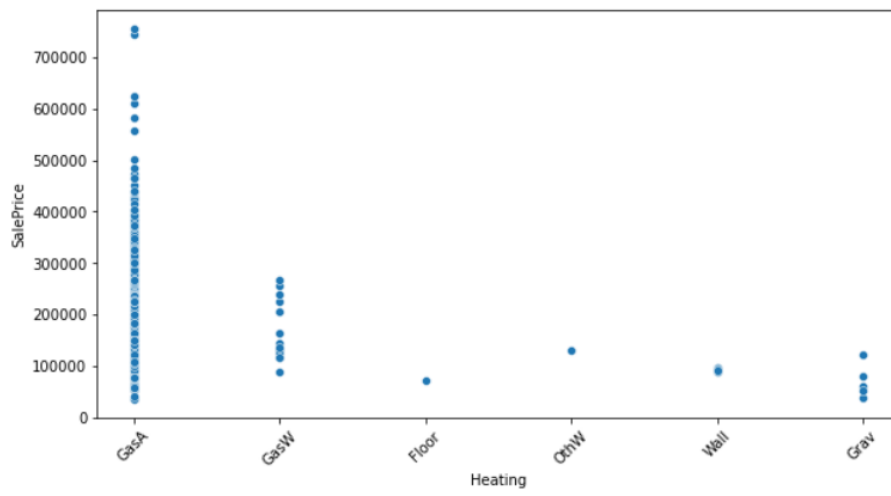
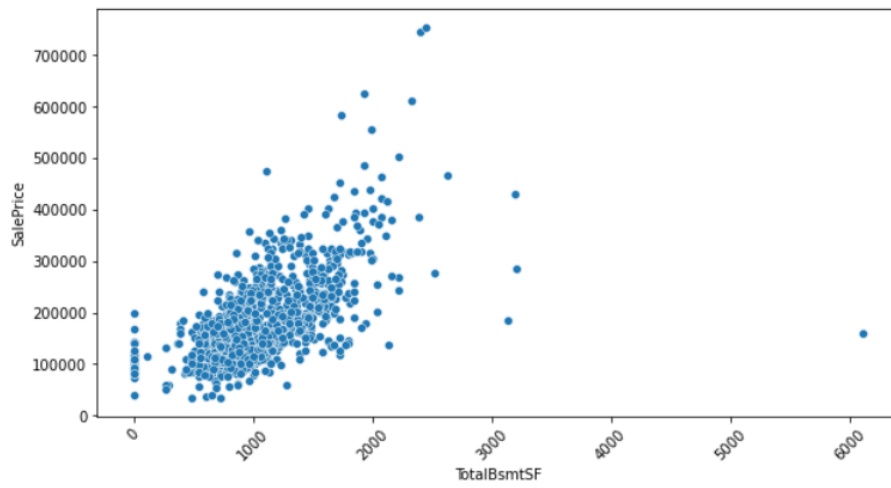


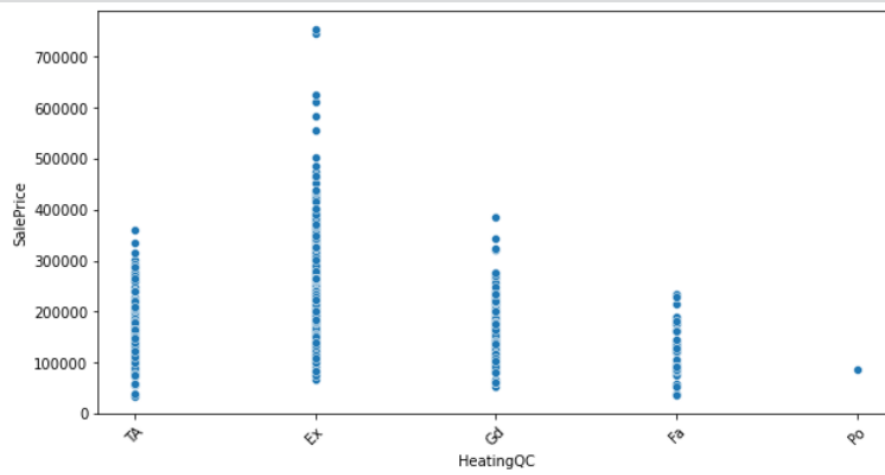


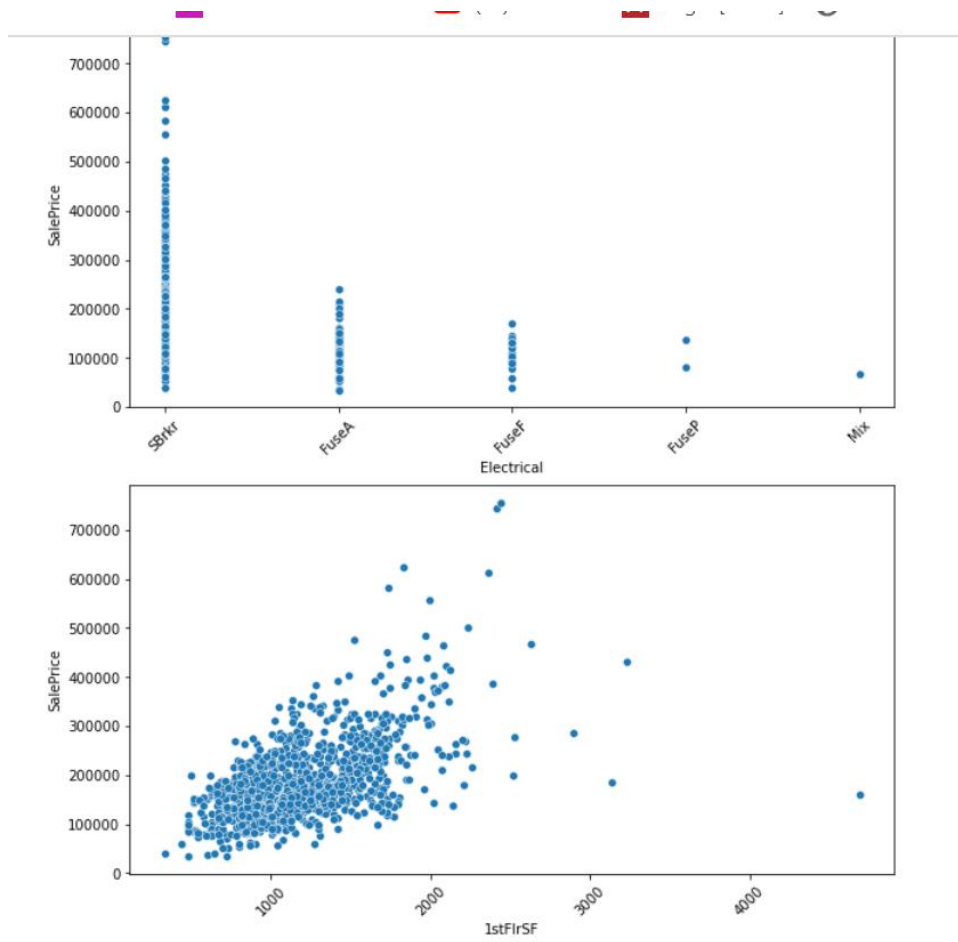


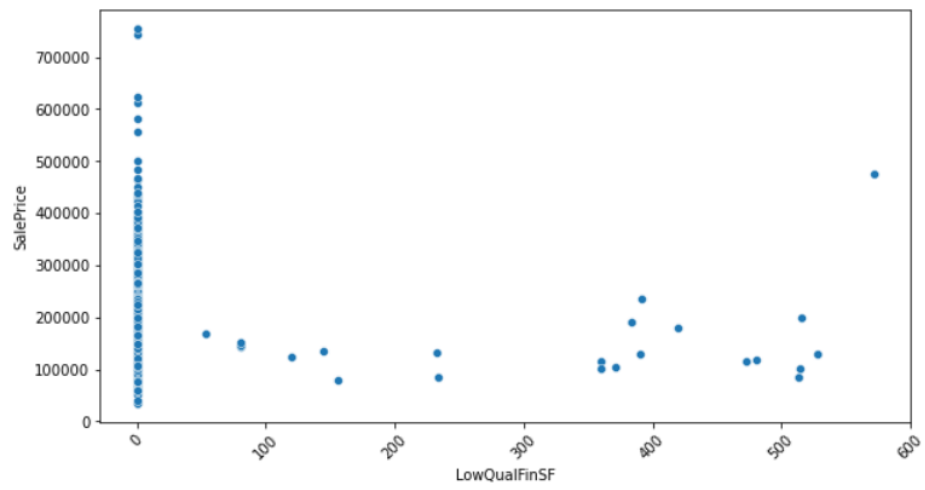
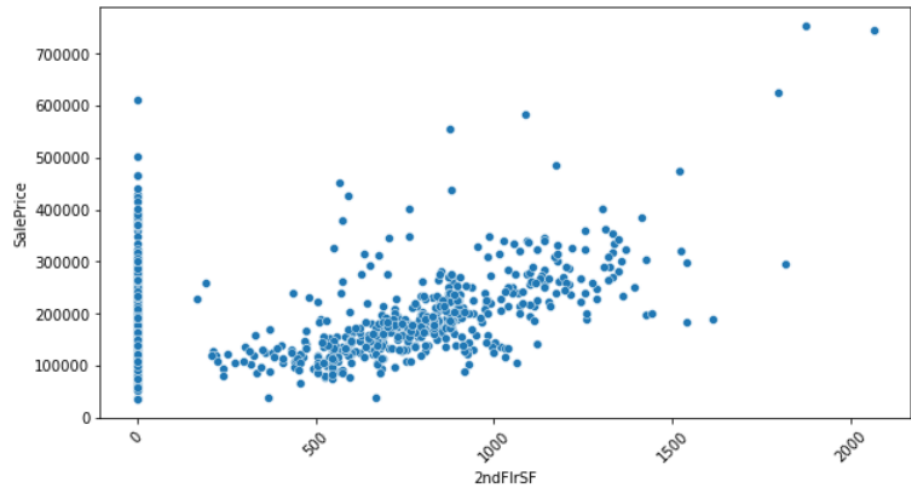


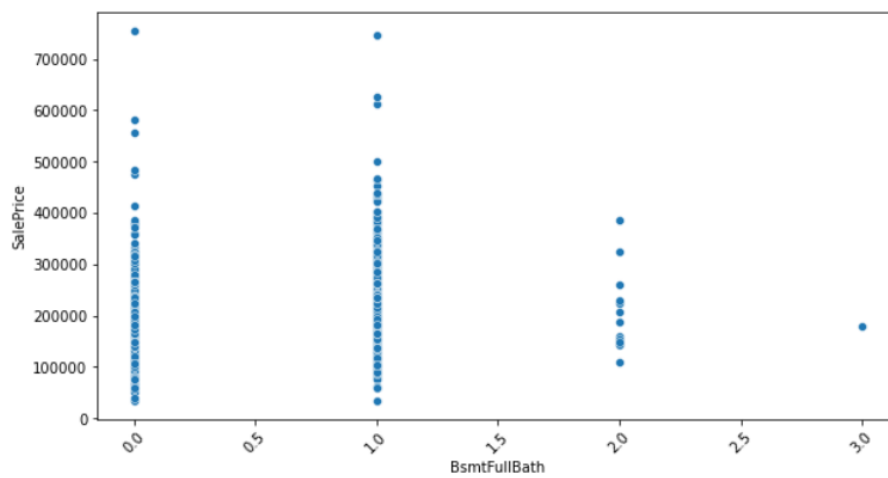
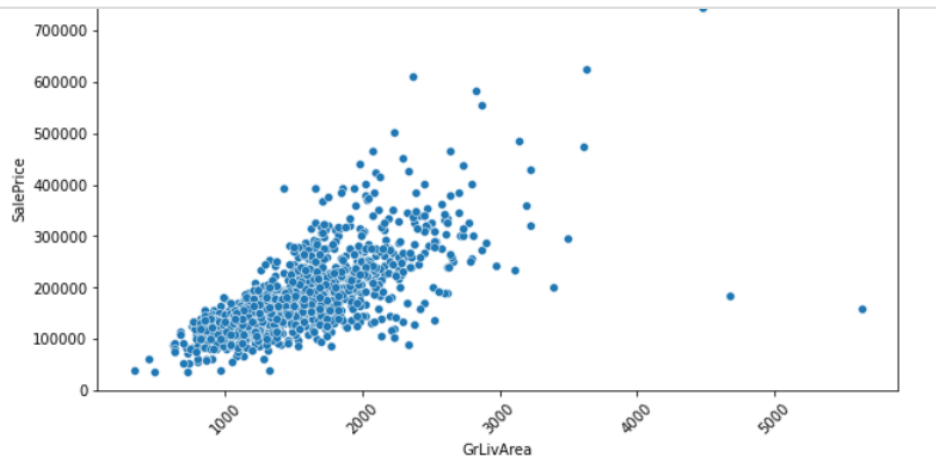


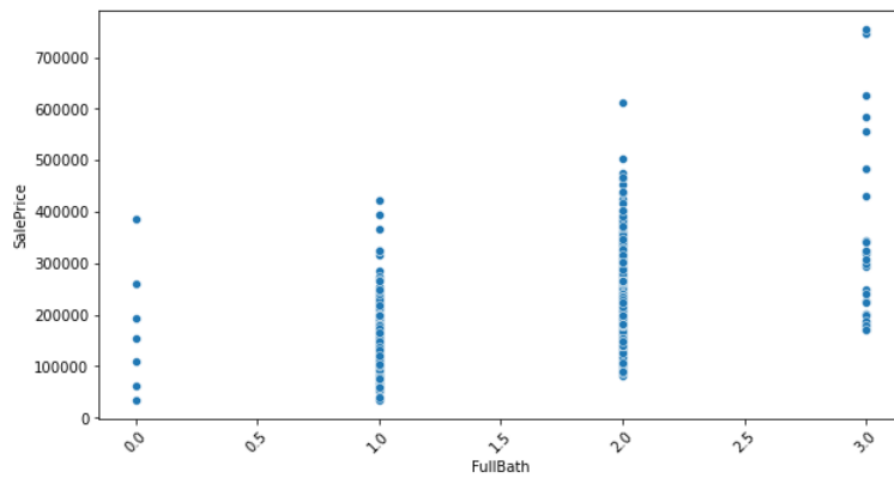
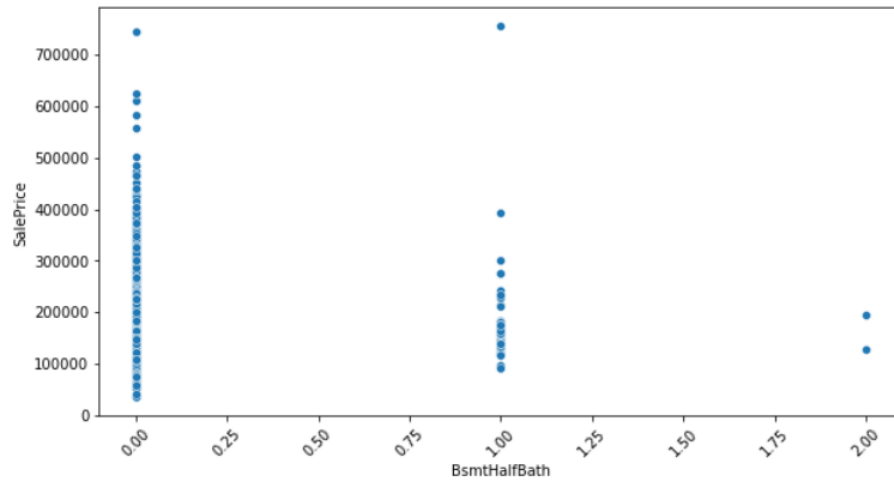


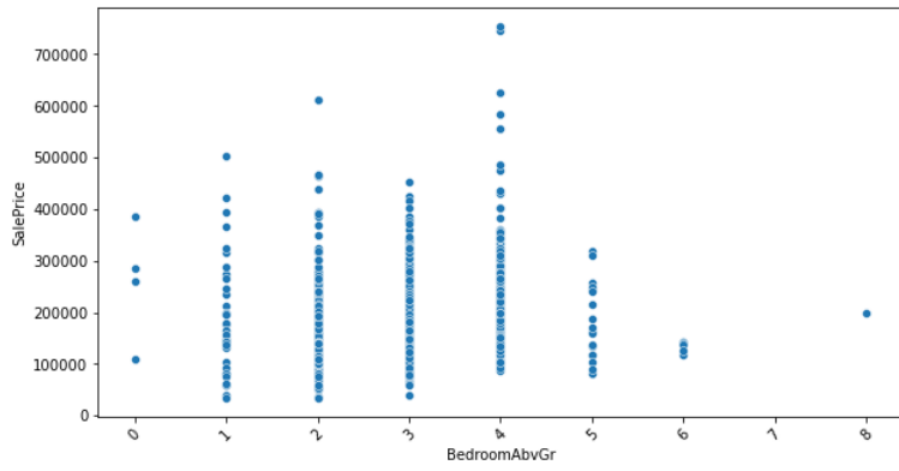
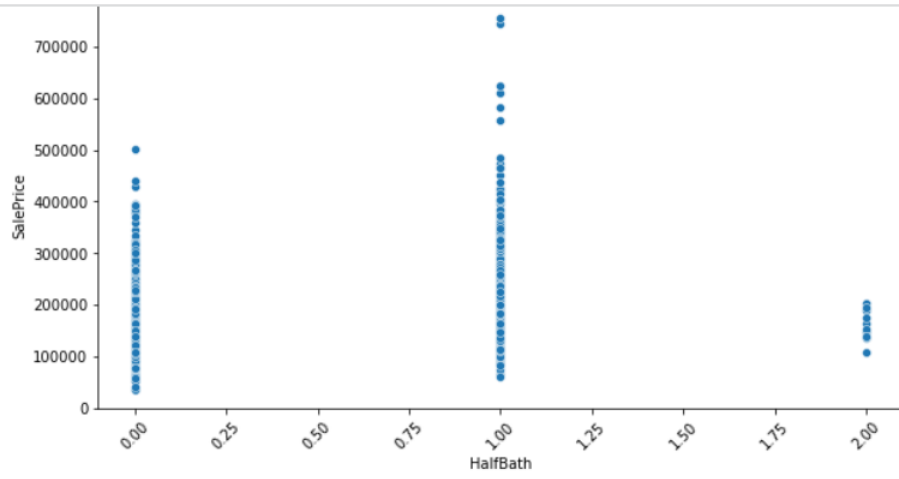


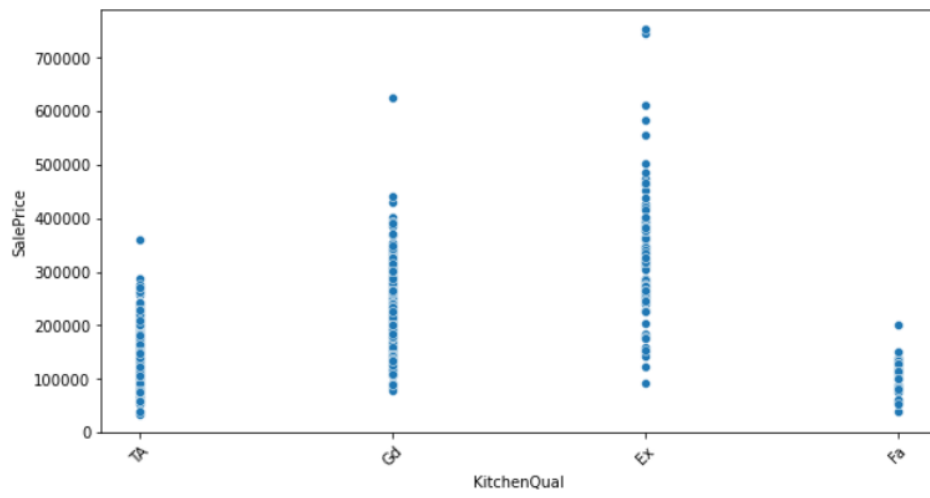
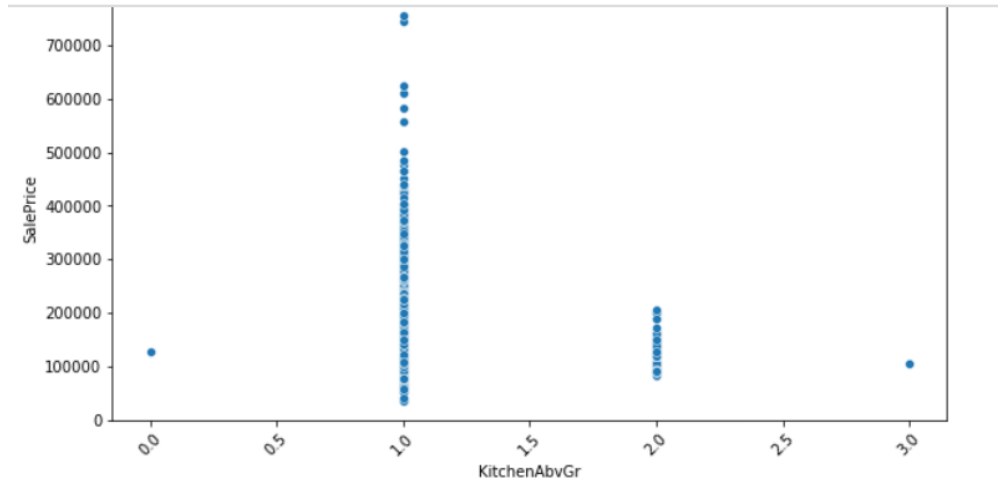


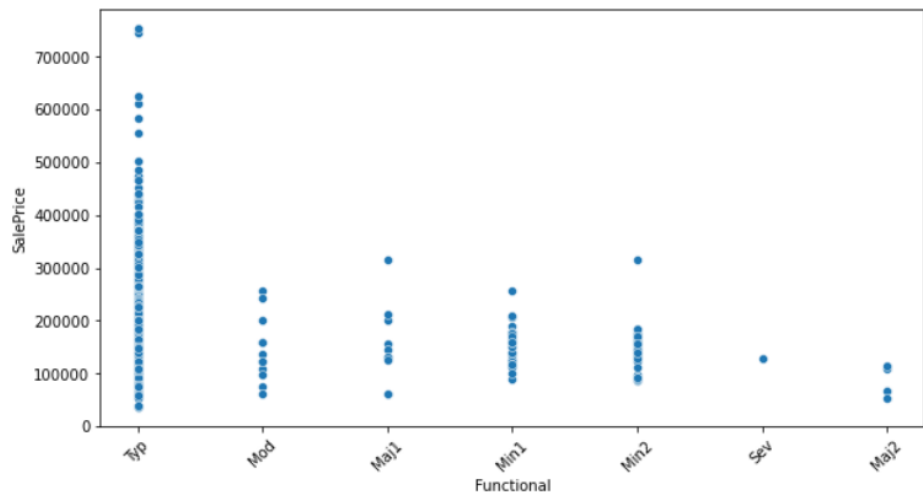
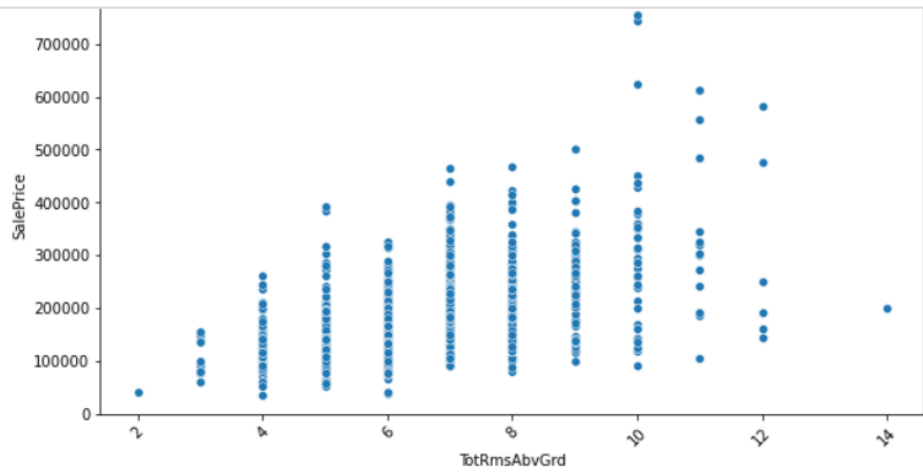


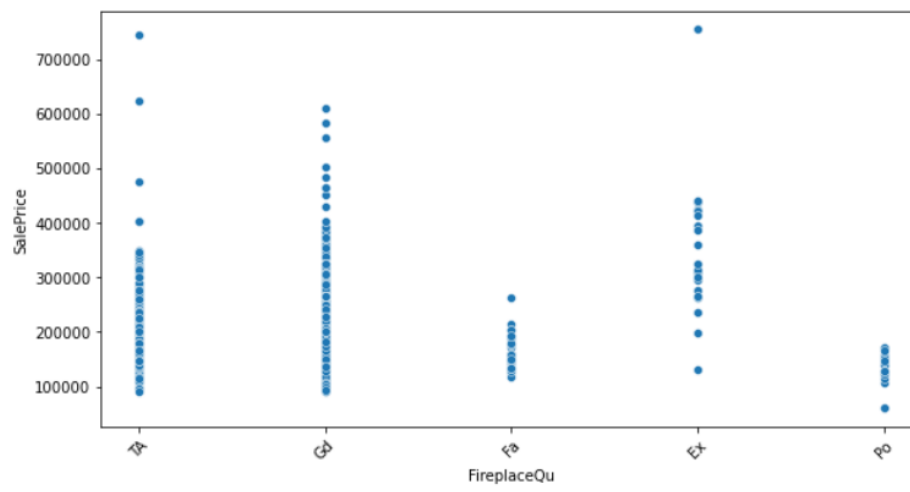
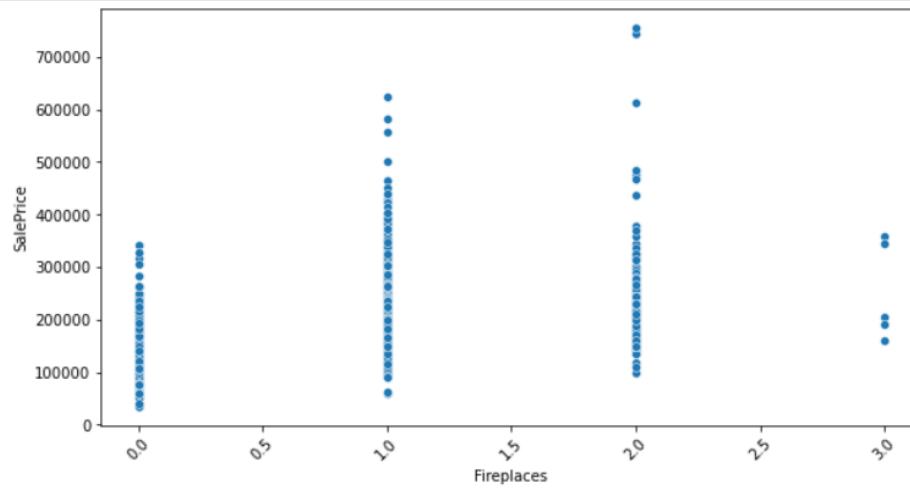


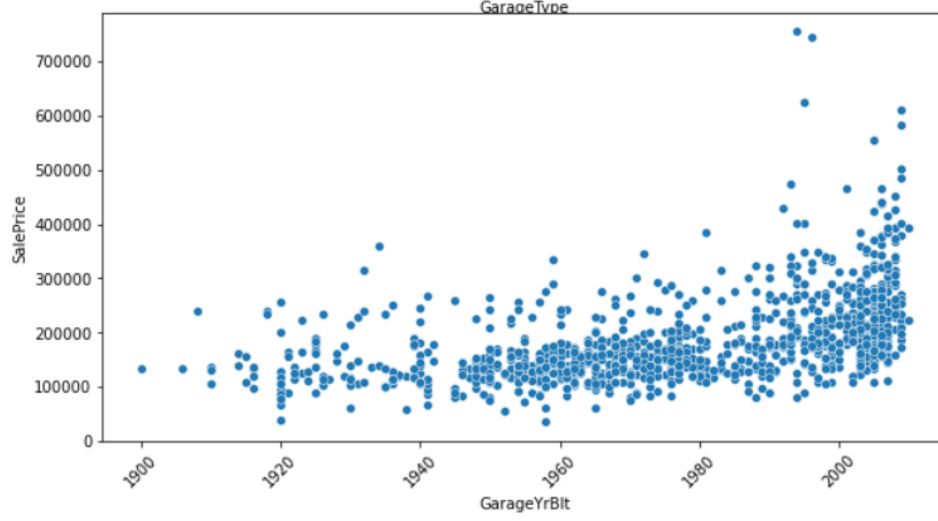
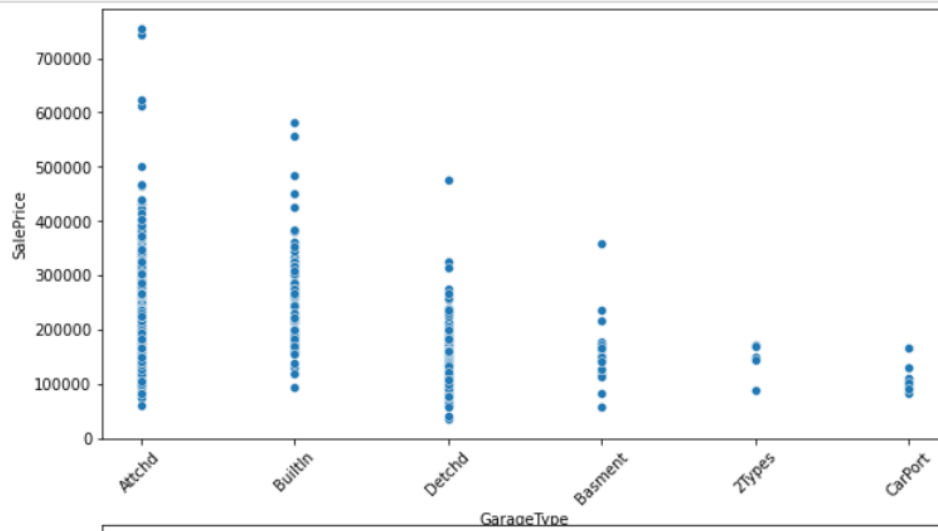


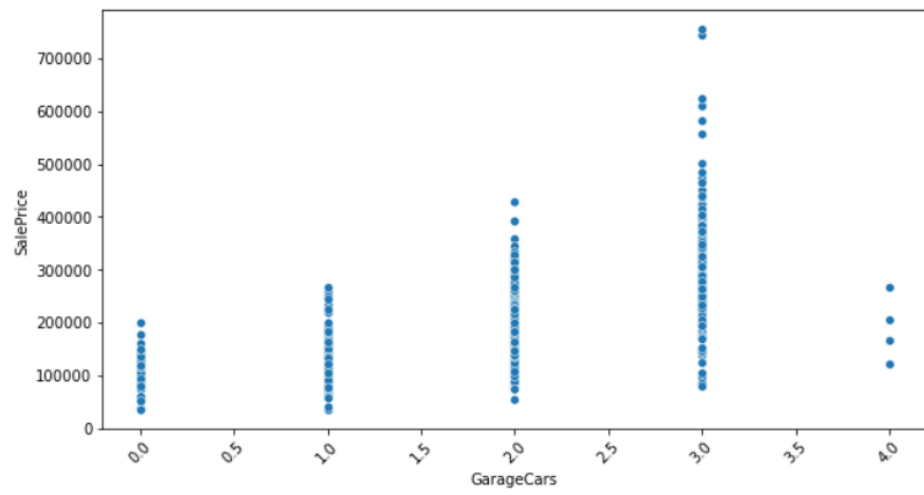
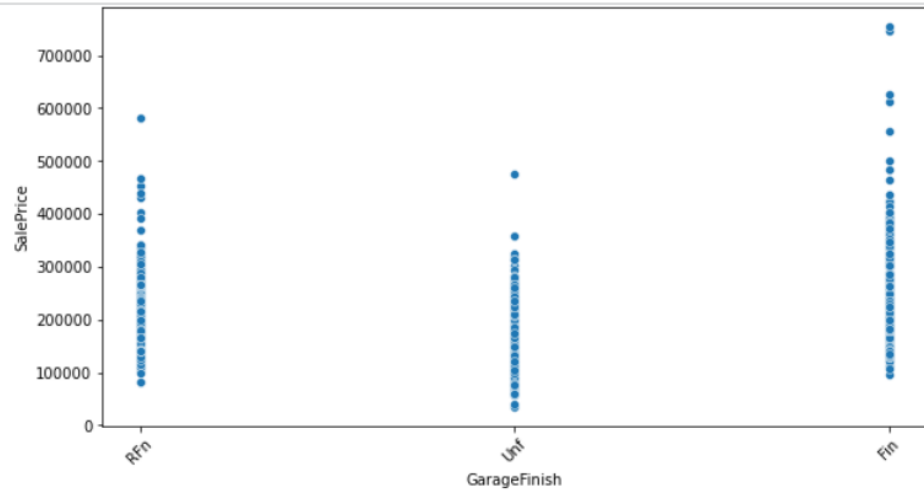


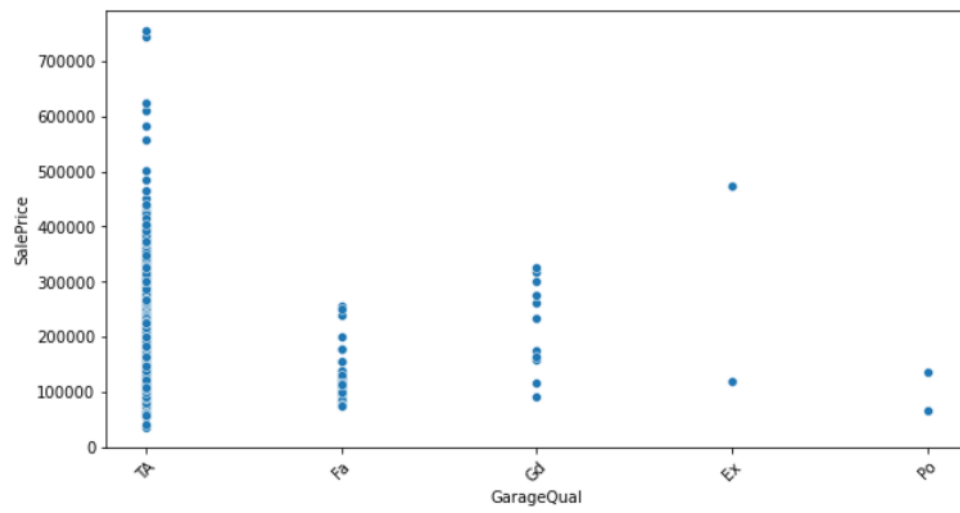
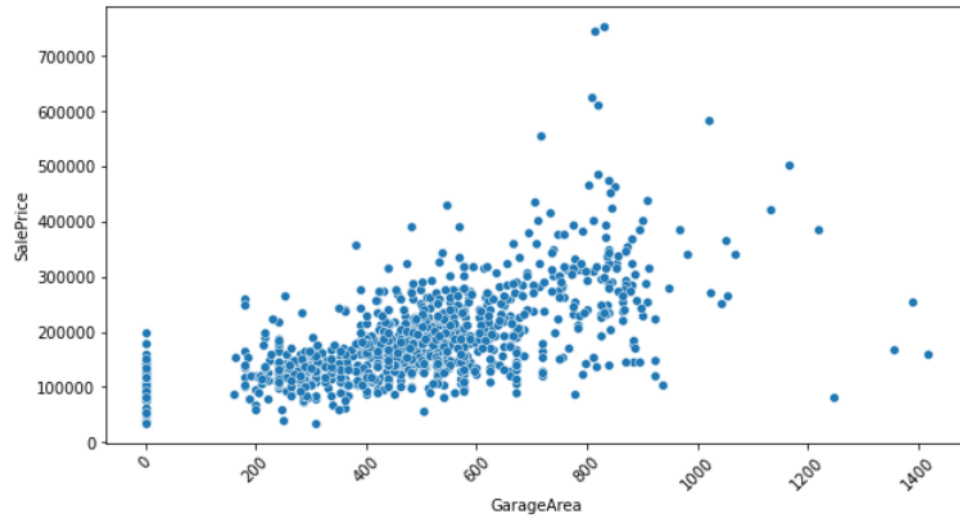


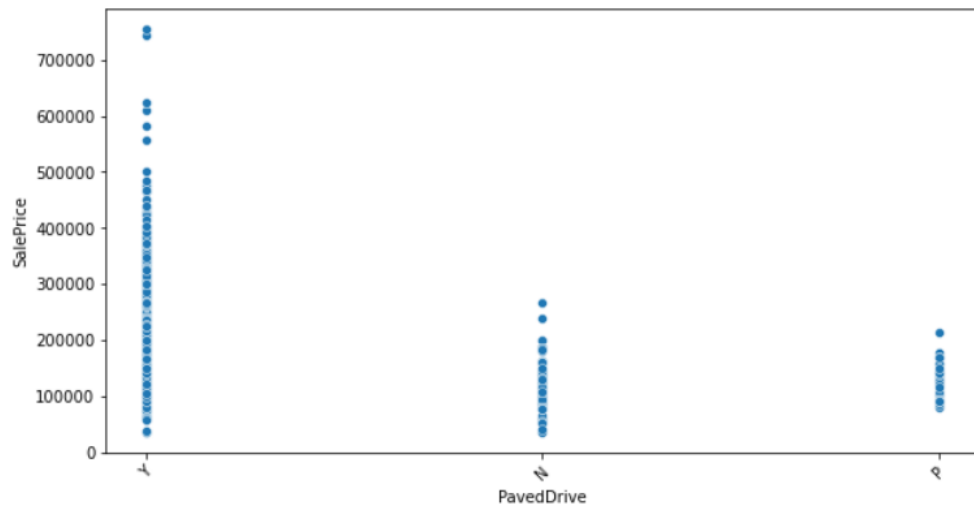
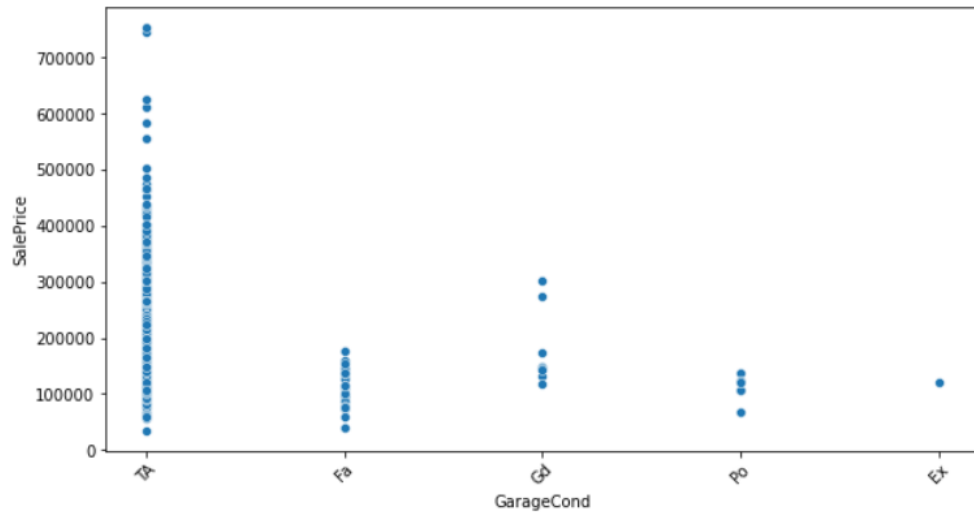


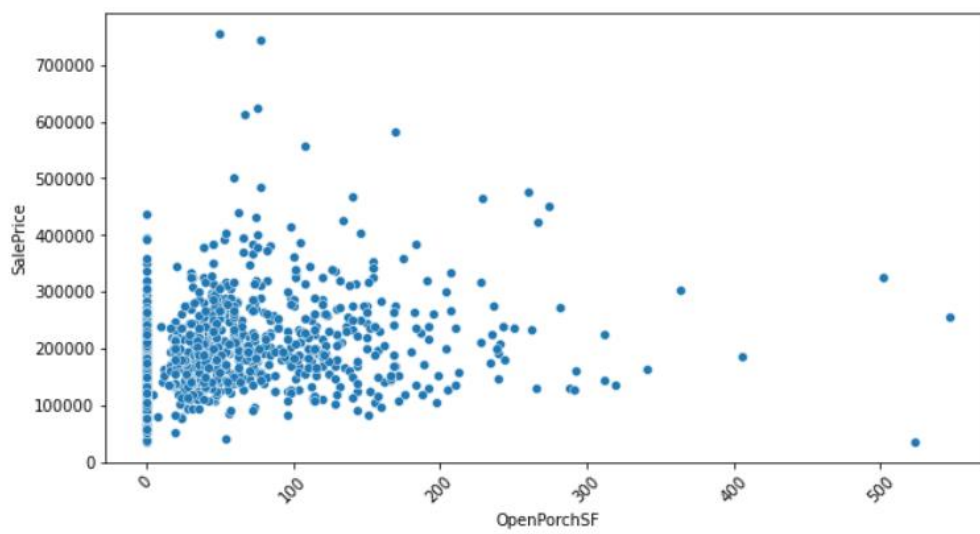
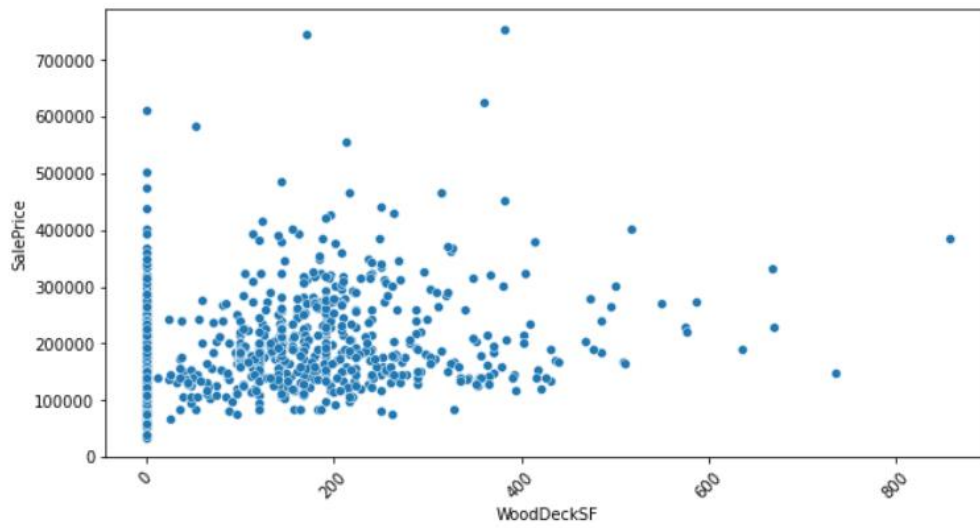


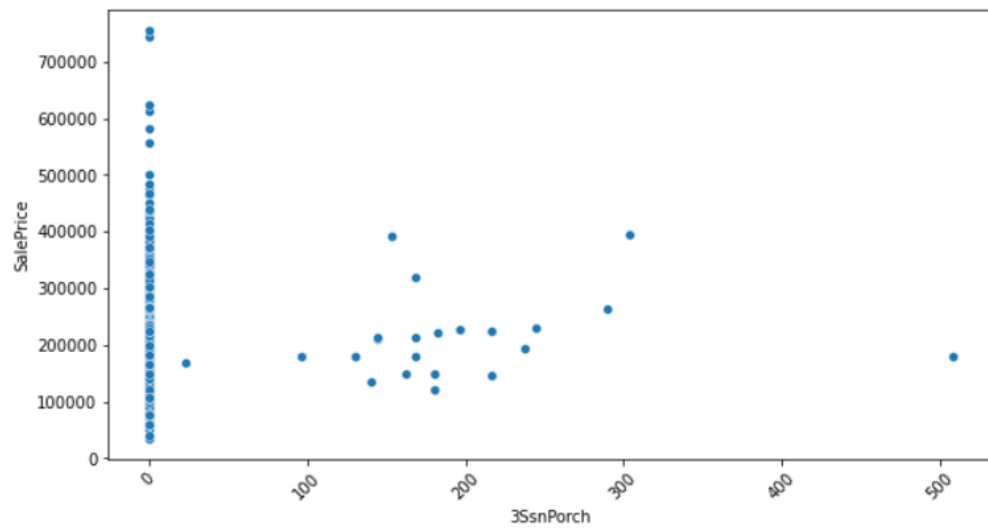
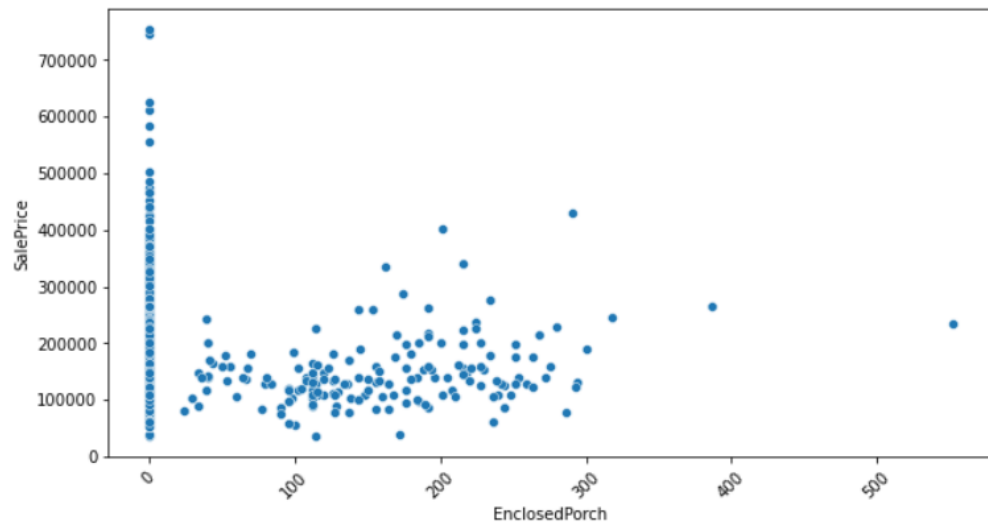


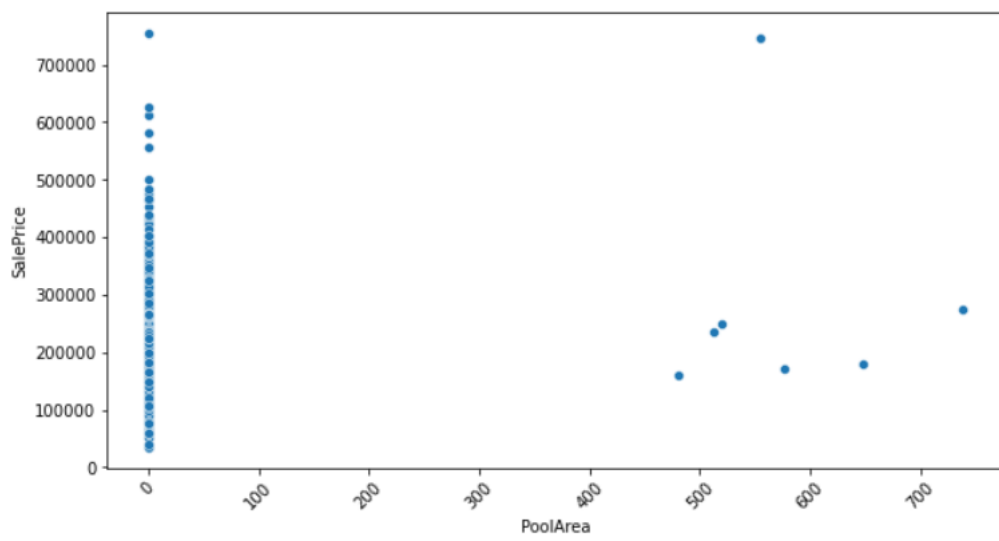
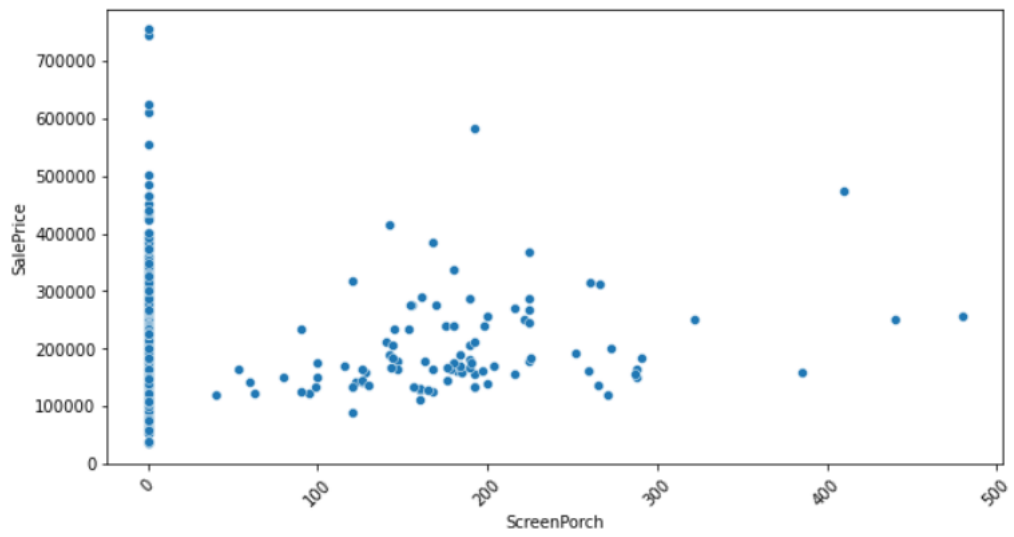


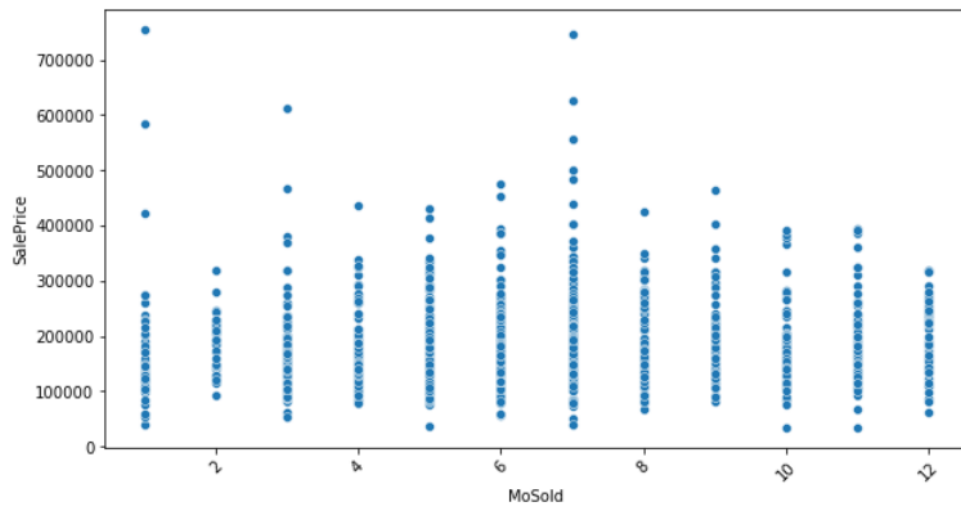
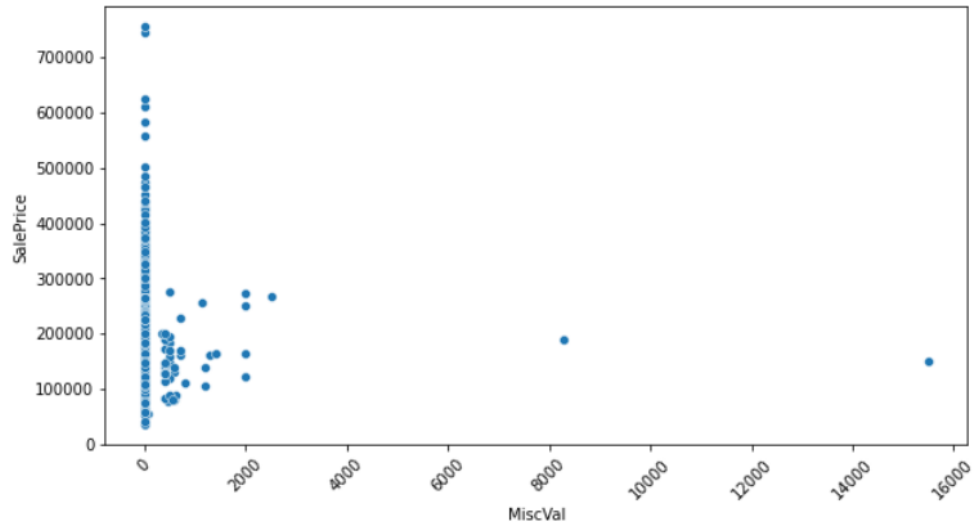


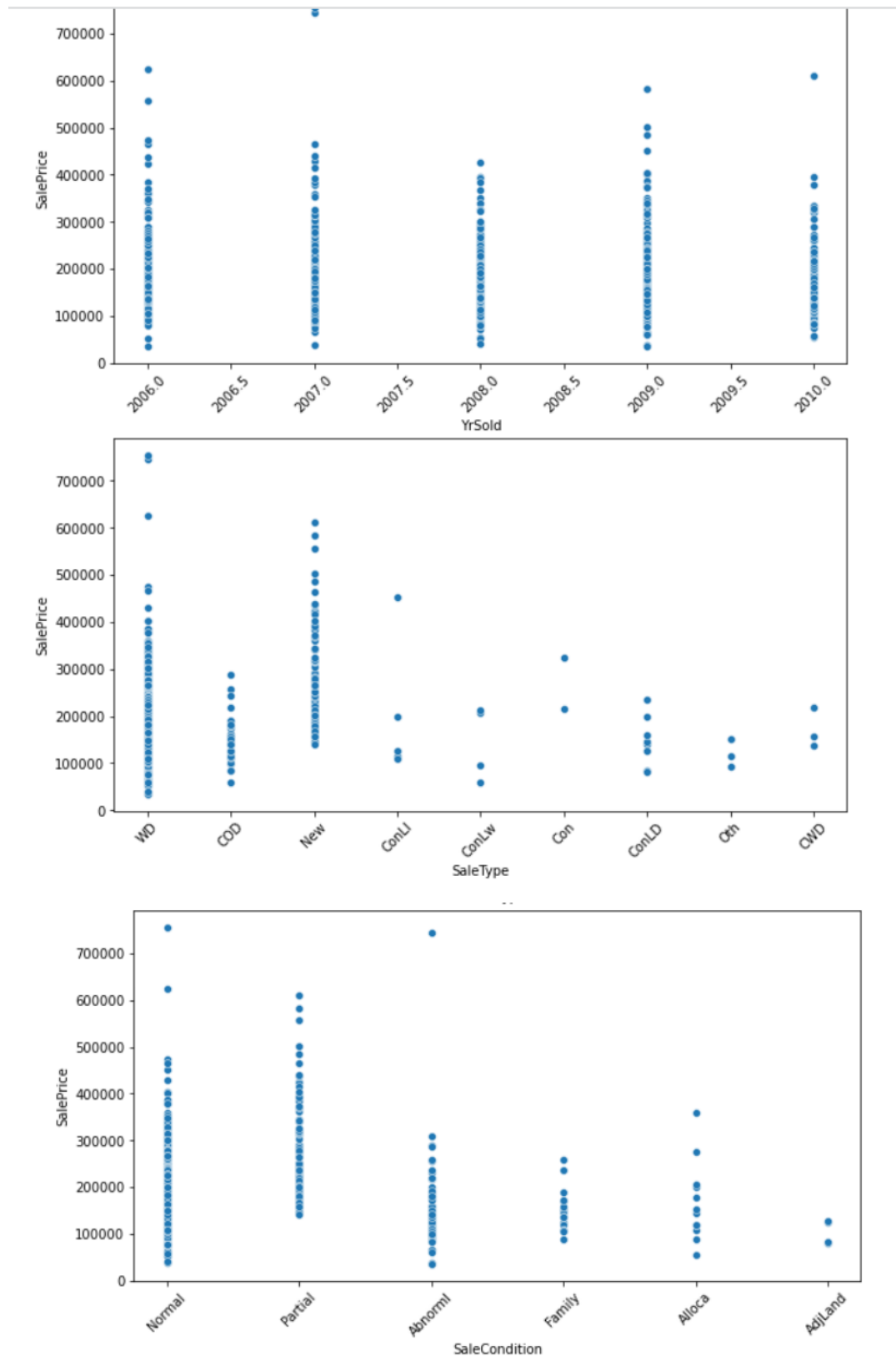












- Interpretation of the Results

Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

-MSS Class: Type 20(1-STORY 1946 & NEWER ALL STYLES) & type 60(2-STORY 1946 & NEWER) are generally getting more sale price than other. Type 180 (PUD - MULTILEVEL - INCL SPLIT LEV/FOYER) are not getting much selling price than other. Type 40(1-STORY W/FINISHED ATTIC ALL AGES) are very less nos. are available for sale.

-MSZoning: Identifies the general zoning classification of the sale.: RL (Residential Low Density) are getting more sale price than other zone. Followed by RM (Residential Medium Density) getting more sale price. Followed by FV (Floating village), RH: Residential High density & Commercial area are getting very less sale price.

-LotFrontage: Linear feet of street connected to property: Those houses are having less linear feet of street are generally getting more selling price. Some houses are very large linear feet more than 300.

-LotArea: Lot size in square feet. The lot area which is less than 20000 are getting very good selling price. Some houses lot area is more than 100000.

-Street: Type of road access to property. Pave street are getting more selling price as compared to Grvl.

-LotShape: General shape of property. IR1 means Slightly irregular are getting good selling price, followed by Reg: Regular shape, IR3: Irregular shape are getting very less sale price.

-LandContour: Flatness of the property. Lv1: Near Flat or level are getting more sale price, followed by HLS: Hillside - Significant slope from side to side, Low: Depression & Then after Bnk: Banked - Quick and significant rise from street grade to building: which is getting very less sale price than others.

-Utilities : All public Utilities (E, G, W, & S) are getting very good sale price.

-LotConfig: Lot configuration: Inside lot & Corner lot are generally getting good selling price than others. Frontage on 3 sides of property are less in nos. & generally not getting more selling price.

-LandSlope: EG: Gentle Slope are getting more sale price followed by moderate slope & severe slope.

-Neighborhood: Physical locations within Ames city limits. NoRidge (Northridge) and NridgHt (Northridge Heights) are generally getting good selling price.

-Condition1: Normal condition houses are getting good selling price. RRNe (Within 200' of East-West Railroad) are very less in nos. & also not getting much selling price.

-Condition2: Proximity to various conditions (if more than one is present), Normal condition is getting very good selling price. Other condition houses are very less in nos. & also not getting good selling prices.

-BldgType: Type of dwelling: Single-family Detached are getting good selling price. Followed by Townhouse End Unit. Whereas others are not getting much selling price.

-HouseStyle: 2Story are getting good selling price followed by one story houses getting good selling price. 1.5Unf (One and one-half story: 2nd level unfinished) are getting very less selling price than others.

-Overall Quality: The houses which overall quality rating is high they are getting very good selling price.

-Overall condition: Rated 5 houses are getting good selling price followed by rating 6 & above getting good selling price. The houses which are having overall condition less than 5 not getting good selling price.

-YearBuild: No doubt the houses which Year build is not older are getting good selling price & the houses which year build is older than that is not getting good selling price.

-YearRemodAdd: Remodel date (same as construction date if no remodeling or additions). Newly remodeling houses are getting good selling price in the year of 1995 to 2010.

-RoofStyle: Type of roof: Hip roof style are getting good price followed by Gable roof style. Gambrel & all other are not getting good selling price.

-RoofMatl: Roof material: generally people are preferring for: CompShg (Standard (Composite) Shingle) roof material are they are getting good selling price. Followed by other roof material. Other roof material are not commonly used.

-Exterior1st: Exterior covering on house: Generally people are preferring for VinylSd (Vinyl Siding) are getting good selling price. Followed by cement board. MetalSd (Metal Siding), BrkFace (Brick Face) & Plywood (Plywood) are also getting average selling price.

-Exterior2: are also getting prices as per exterior 1.

-Masonry veneer type: Stone type are getting good selling price followed by brick face.

-MasVnrArea: Masonry veneer area in square feet. Generally more the area more the selling price. But in some cases if area is not available they are also getting good selling price may be due to some other features.

-ExterQual: Evaluates the quality of the material on the exterior. Selling price is good as per quality of exterior.

-ExterCond: Evaluates the present condition of the material on the exterior. Average/Typical condition home are getting good selling price, followed by good condition getting good selling price.

-Foundation: Type of foundation. PConc(Poured Concrete) are getting good selling price followed by CBlock(Cinder Block) & BrkTil Brick & Tile.

-Basement Quality: Which house basement quality is good generally they are getting good selling price. Selling price is depend on the condition of basement.

-Basement Condition:- Typical/Average condition basement houses are getting good selling price followed by good condition.

-Basement Expouser:- Getting selling price as per the expouser of basement.

-BsmtFinType1: Rating of basement finished area: Generally good living type are getting good selling price followed by unfinished.

BsmtFinType1: Rating of basement finished area: Generally good living type are getting good selling price followed by unfinished.

-BsmtFinSF1: Type 1 finished square feet. Selling price increases with increase in square feet area

-BsmtFinType2: Rating of basement finished area (if multiple types): Unfinished basement area generally getting good selling price.

-BsmtFinSF2: Type 2 finished square feet. That generally not depend on sq. area. If no type 2 finished sq. ft area are also getting very good selling prices.

-BsmtUnfSF: Unfinished square feet of basement area. As the increase in area the selling price is also increases.

-TotalBsmtSF: Total square feet of basement area. Selling price is increase with increase in increase in basement sq. ft. area.

-Heating: Type of heating. GasA(Gas forced warm air furnace) are getting good selling price than other heating type.

-HeatingQC: Heating quality and condition. Getting good selling price as per heating quality.

-CentralAir: Central air conditioning. The houses are getting good selling price & peoples are preferring most are central air conditioning.

-Electrical: Electrical system.SBrkr(Standard Circuit Breakers & Romex) are getting good selling price followed by Fuse(Fuse Box over 60 AMP and all Romex wiring (Average))

-1stFlrSF: First Floor square feet. Selling price increaese with the increase in sq. ft. area of first floor.

-2ndFlrSF: Second floor square feet. Selling price increase with increase in second floor area.

-LowQualFinSF: Low quality finished square feet (all floors). No low quality all floors area getting good selling price.

-GrLivArea: Above grade (ground) living area square feet. Selling price increases with increase in ground living area in sq. ft.

-BsmtFullBath: Basement full bathrooms. Getting good selling price whose are having one full basement bathrooms follwed by no bathrooms.

-BsmtHalfBath: Basement half bathrooms.Getting good selling price whose does not having bathrroms followed by one bathrooms.

-FullBath: Full bathrooms above grade. Getting good selling price which having havigher bathrroms. 3 bathrooms houses are getting very good selling price followed by 2 & 1.

-HalfBath: Half baths above grade. The houses are having one half baths above grades are getting good selling price followed by no baths are getting good selling price.

-Bedrooms above ground: Price increases with incrase in bedrooms above ground. Generaly 4 ,bedrrrom are getting good selling price, followed by 3, 2 & 1.

-Kitchen above ground: Generally 1 kitchen are getting good selling price. 2 kitchen & above are very less in nos. & also not getting good selling price.

-Getting good selling price as per kitchen qaulity.

-TotRmsAbvGrd: Total rooms above grade (does not include bathrooms). Generally more the rooms are getting more selling price. 10 rooms are mainly getting more selling price followed by 11 & below 10 rooms.

Functional: Home functionality (Assume typical unless deductions are warranted). Typically functionality houses are getting good selling price.

-Fireplaces: 2 & 1 fireplaces are getting good price.

-Firequality: Good & typical quality are getting very good selling price.

-GarageType: Garage location. Attached to home are getting good selling price. followed by Built in garage are getting good selling price.

-Garage Year Build: Newly build garages are getting good selling price.

-Garage finished which has finished type are getting good price followed by roughfinished & unfinished.

-GarageCars: Size of garage in car capacity. Three cars capacity are getting good selling price followed by 3, 2.

-Garage Area: Price increase with the increase in area of garage.

-Garage Quality: Typical/Average quality garages are getting good selling prices.

-Same as garage condition.

-PavedDrive: Paved driveway. Paved drive are getting good selling price.

-WoodDeckSF: Wood deck area in square feet. Selling price increases with increase in Wood deck.

-OpenPorchSF: Selling price increase with increase in open porch.

-Enclosed orch: The houses which does not have enclosed porch are also having good selling porch.

-3SsnPorch: Three season porch area in square feet same as above.

-Screen Porch & Pool area, MiscVal: \$Value of miscellaneous feature which does not having area getting very good selling price.

-Month sold & Year Sold are of 2006 to 1010 year

-Sale Type:- New(Home just constructed and sold) are getting good selling price followed by WD (Warranty Deed - Conventional)

-Sale Condition:- Normal Normal Sale are getting very good selling price followed by Partial Home was not completed when last assessed (associated with New Homes).

CONCLUSION

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

Describe the key findings, inferences, observations from the whole problem.

Generally peoples are preferring for low residential area zone, Paved street, Slitely irregular shape of property, flat level of property, Inside lot, gentle slope of property, Normal condition, Single family detached dwelling, Newly build, Overall good condition of house, Hip roof top & CompShg(Standard (Composite) Shingle)roof material ,exterior covering VinylSd(Vinyl Siding), Poured concret foundation, overall good condition of centralised AC, heater, Basement area, Build up large area, Standard circuit breaker, no. of rooms , no. of bathrooms, quality of rooms & bathrooms, functionality of house,fireplaces, size of garages, Open porch size , newly constructed, All public Utilities (E,G,W,& S) are getting very good selling price.

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

Learned how to handle the outliers, cleaning of data set, How to fill null values, how to plot different graph using matplotlib & seaborn.

Challenges are to visualise the heatmap because of large nos. of features.

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

What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results.

