Data Mining HW2

Kaushal Vinay Nerkar

Link to Colab : https://colab.research.google.com/drive/1KV5fMNItfaMojmK6f5wYnN5xvXO-zCUg?authuser=1#scrollTo=LBncHmJ8urPE&uniqifier=1

Question 1 :

Attribute	Type
Id	numerical
MSSubClass	numerical
MSZoning	ordinal
LotFrontage	numerical
LotArea	numerical
Street	binary
Alley	ordinal
LotShape	ordinal
LandContour	ordinal
Utilities	binary
LotConfig	ordinal
LandSlope	ordinal
Neighborhood	ordinal
Condition1	ordinal
Condition2	ordinal
BldgType	ordinal
HouseStyle	ordinal
OverallQual	numerical
OverallCond	numerical
YearBuilt	numerical
YearRemodAdd	numerical
RoofStyle	ordinal
RoofMatl	ordinal
Exterior1st	ordinal
Exterior2nd	ordinal
MasVnrType	ordinal
MasVnrArea	numerical
ExterQual	ordinal
ExterCond	ordinal
Foundation	ordinal
BsmtQual	ordinal
BsmtCond	ordinal
BsmtExposure	ordinal
BsmtFinType1	ordinal
BsmtFinSF1	numerical
BsmtFinType2	ordinal

Here attributes are categorized into numerical , binary and ordinal data.

Above is the sample of data from output.csv (sheet2) ,which contains Attribute vs type.

Histograms for house prices and lot sizes

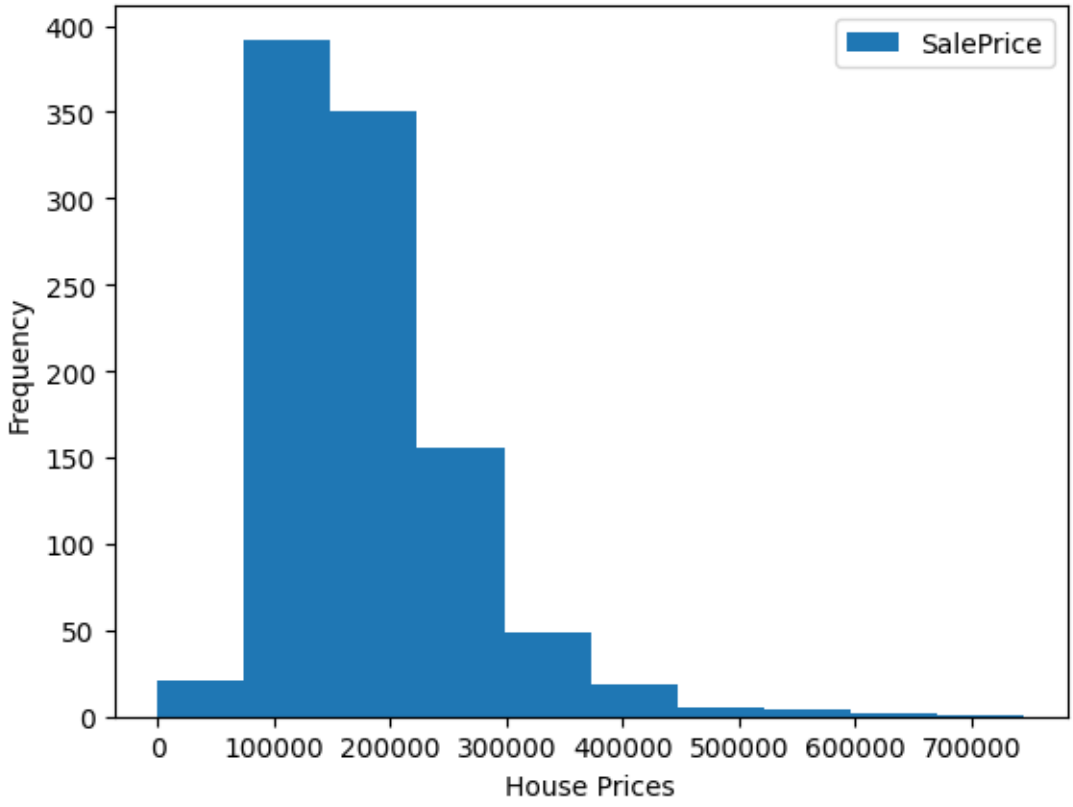


Figure 1: histogram of house prices

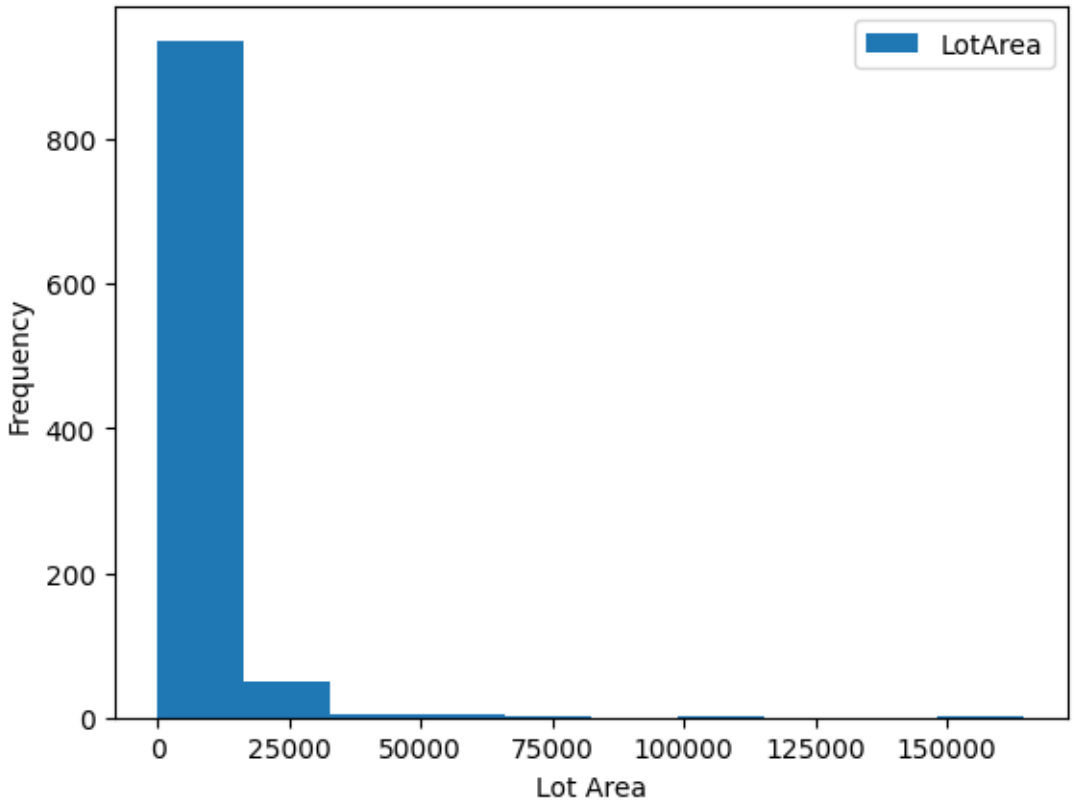


Figure 2: Histogram of lot sizes

Box plots

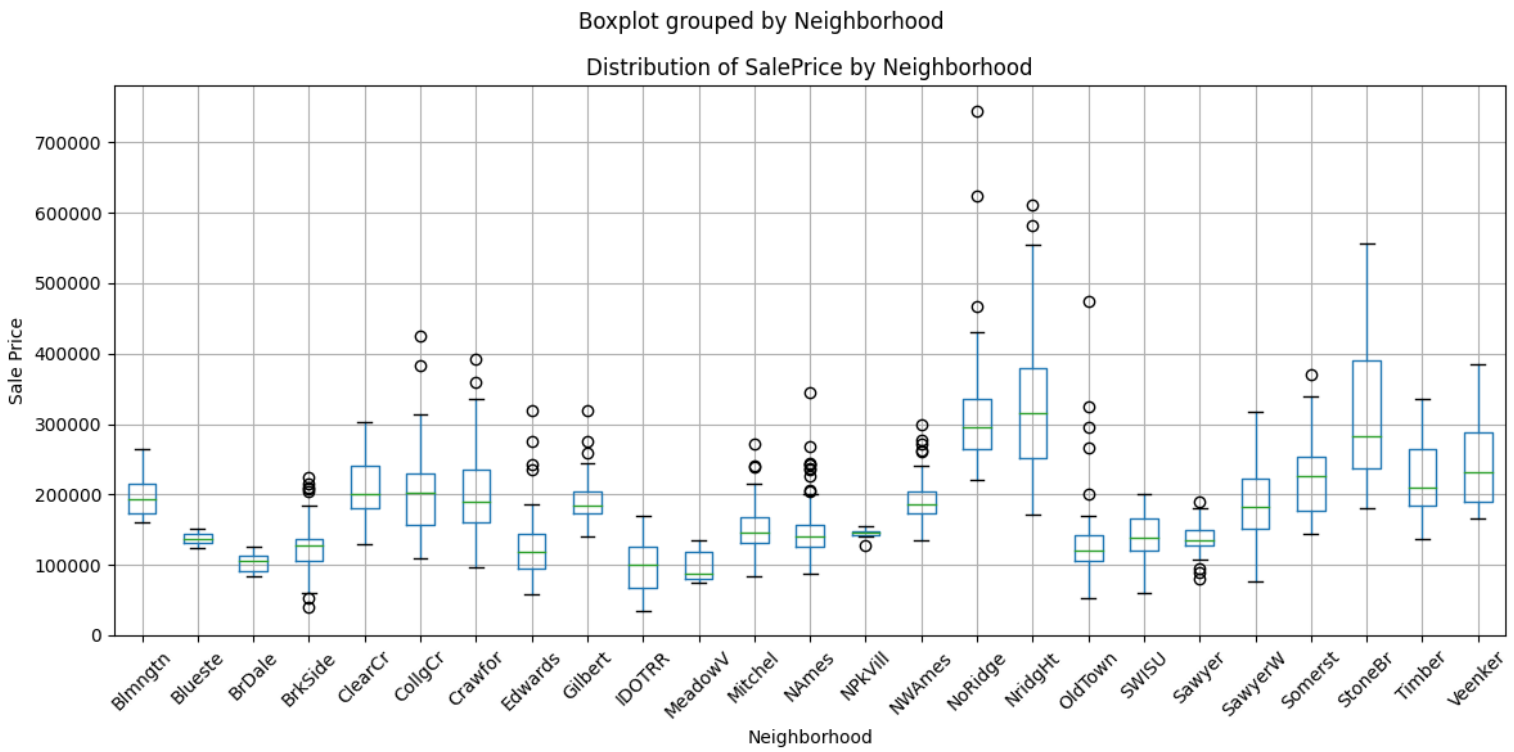


Figure 3: boxplot of house price for every category in neighborhood

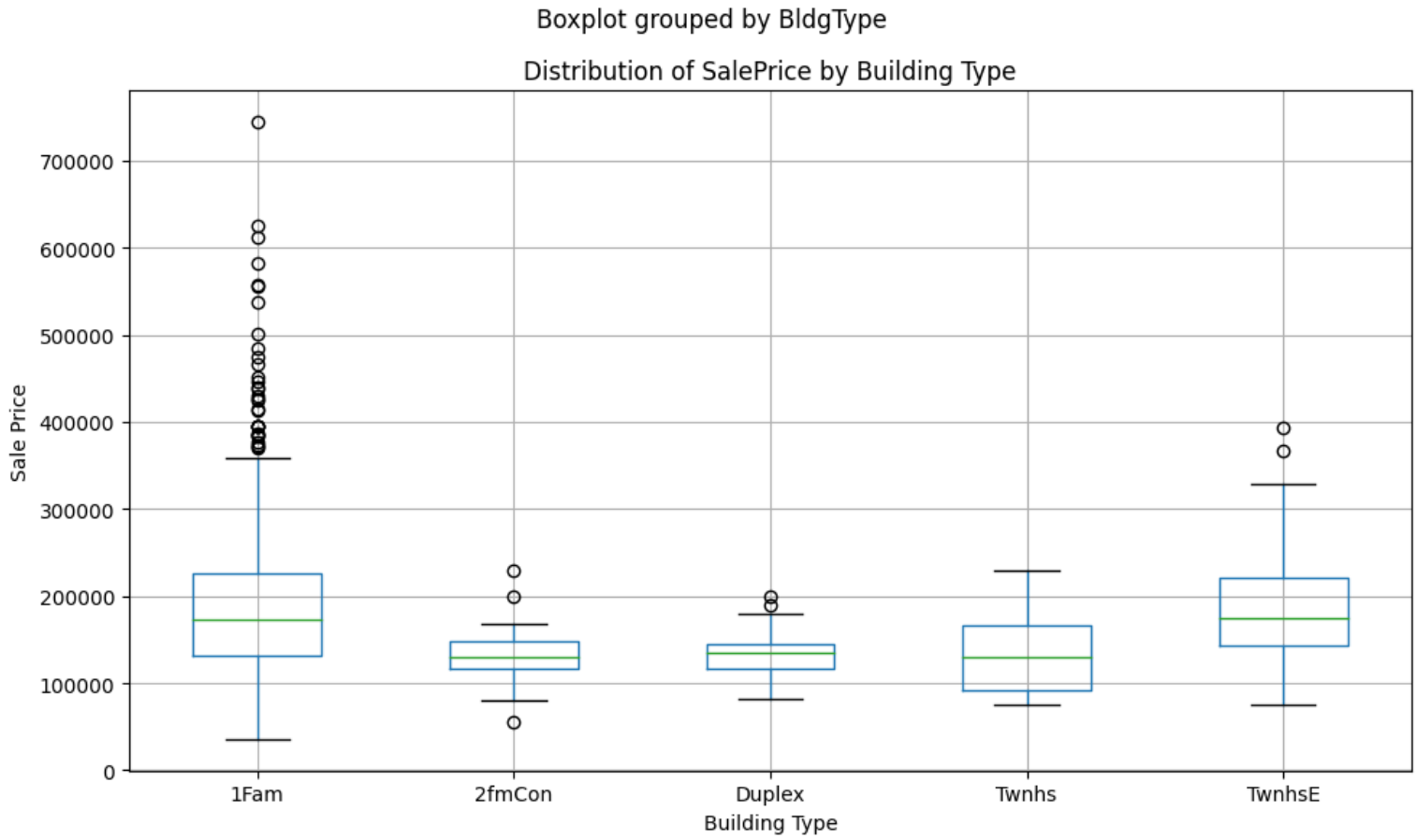


Figure 4: Boxplot for house price for every category in Building Type

Question 2 :

Correlation between house price and other ordinal attributes.

	Column	Spearman_corr	Spearman_pval
0	MSZoning	-0.328412	1.411646e-26
1	Alley	NaN	NaN
2	LotShape	-0.301505	1.839253e-22
3	LandContour	0.015046	6.346236e-01
4	LotConfig	-0.067799	3.205084e-02
5	LandSlope	0.028327	3.708760e-01
6	Neighborhood	0.183473	5.085623e-09
7	Condition1	0.191524	1.024453e-09
8	Condition2	0.050832	1.081718e-01
9	BldgType	-0.129760	3.860655e-05
10	HouseStyle	0.246586	2.564927e-15
11	RoofStyle	0.153103	1.149211e-06
12	RoofMatl	0.088537	5.081890e-03
13	Exterior1st	0.066351	3.591516e-02
14	Exterior2nd	0.071779	2.321035e-02
15	MasVnrType	NaN	NaN
16	ExterQual	-0.643588	4.987498e-118
17	ExterCond	0.130620	3.425344e-05
18	Foundation	0.507092	1.896375e-66
19	BsmtQual	NaN	NaN
20	BsmtCond	NaN	NaN
21	BsmtExposure	NaN	NaN
22	BsmtFinType1	NaN	NaN
23	BsmtFinType2	NaN	NaN
24	Heating	-0.125196	7.195420e-05
25	HeatingQC	-0.488719	3.685762e-61
26	Electrical	NaN	NaN
27	KitchenQual	-0.565525	1.345082e-85
28	Functional	0.139497	9.536131e-06
29	FireplaceQu	NaN	NaN
30	GarageType	NaN	NaN
31	GarageFinish	NaN	NaN
32	GarageQual	NaN	NaN
33	GarageCond	NaN	NaN
34	PavedDrive	0.270180	3.457368e-18
35	PoolQC	NaN	NaN
36	Fence	NaN	NaN
37	MiscFeature	NaN	NaN
38	SaleType	-0.175334	2.392690e-08
39	SaleCondition	0.319923	3.117217e-25

Attributes significantly correlated with house prices:

`MSZoning` (-0.328),

`LotShape` (-0.301)

`Neighborhood` (0.183),

Condition1` (0.191), `

HouseStyle` (0.246),

`Foundation` (0.507),

`ExterQual` (-0.644),

`HeatingQC` (-0.489),

`KitchenQual` (-0.566),
`Functional` (0.139),

`SaleType` (-0.175),

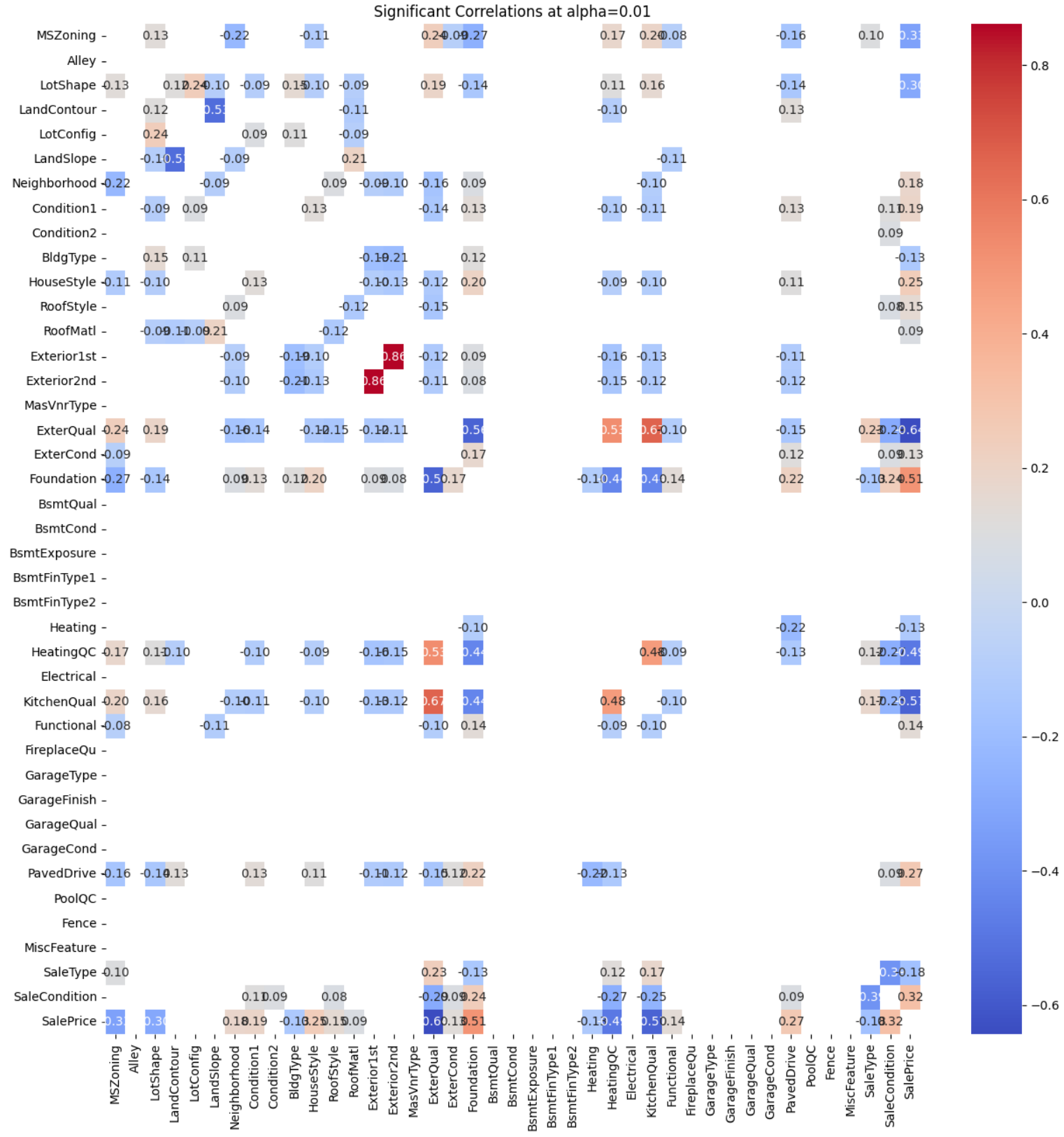
`SaleCondition` (0.320),

`PavedDrive` (0.270)

Statistically significant correlations (p-values ≤ 0.05):

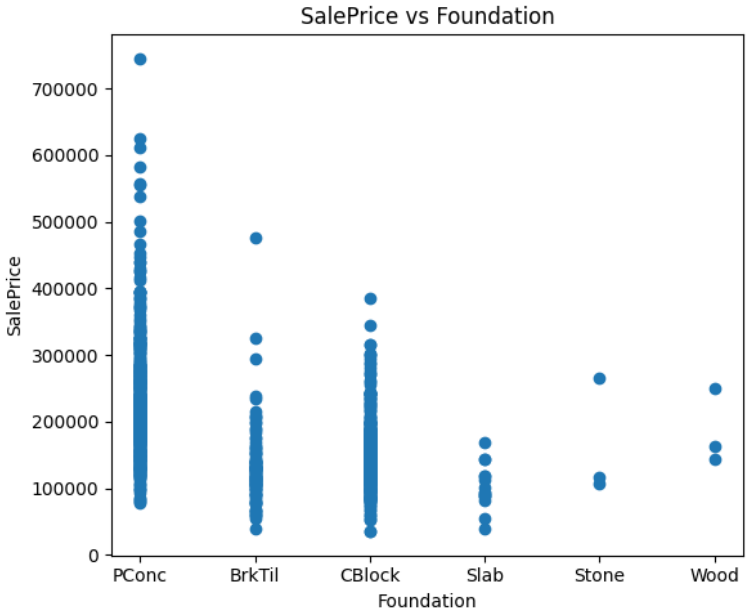
`MSZoning`, `LotShape`, `Neighborhood`, `Condition1`, `HouseStyle`, `Foundation`, `ExterQual`, `HeatingQC`, `KitchenQual`, `Functional`, `SaleType`, `SaleCondition`, `PavedDrive`.

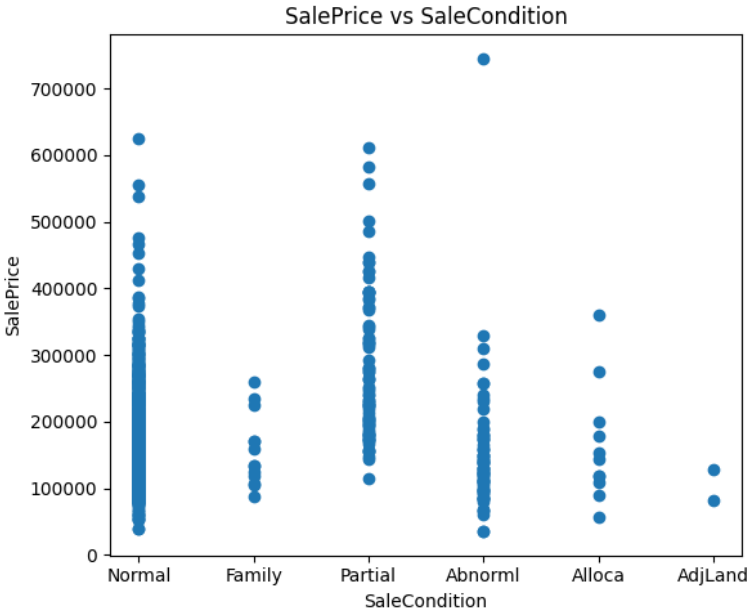
Heatmap


Significant Correlations at alpha=0.01
MSZoning
Alley
LotShape
LandContour
LotConfig
LandSlope
Neighborhood
Condition1
Condition2
BldgType
HouseStyle
RoofStyle
RoofMatl
Exterior1st
Exterior2nd
MasVnrType
ExterQual
ExterCond
Foundation
BsmtQual
BsmtCond
BsmtExposure
BsmtFinType1
BsmtFinType2
Heating
HeatingQC
Electrical
KitchenQual
Functional
FireplaceQu
GarageType
GarageFinish
GarageQual
GarageCond
PavedDrive
PoolQC
Fence
MiscFeature
SaleType
SaleCondition
SalePrice
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

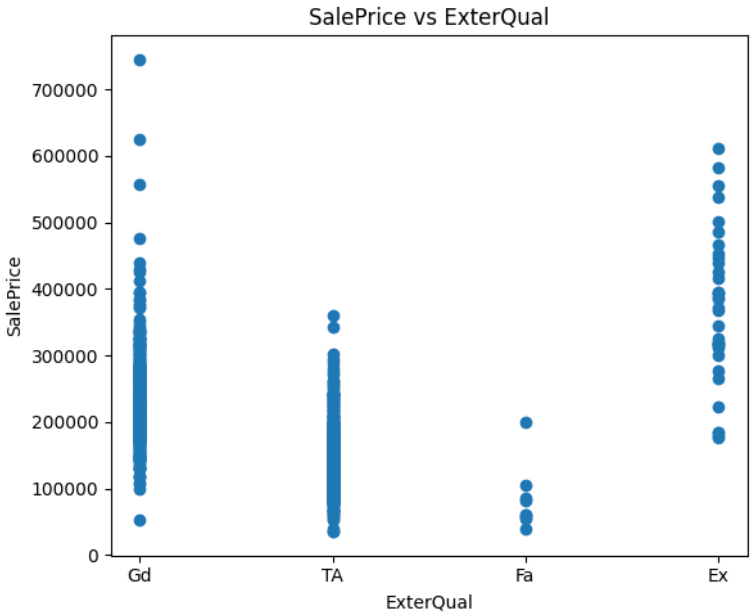
Scatter plots

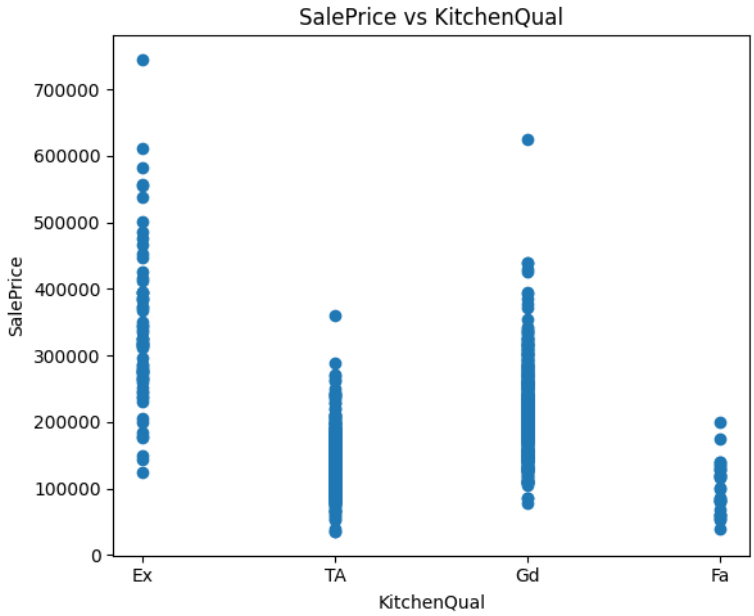
The 2 most positively correlated attributes with sale price are the following: Foundation and Sale Condition.





The 2 most negatively correlated attributes with sale price are the following: ExternalQual and KitchenQuality.





Question 3 :

Normalized data

	Id	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	...	WoodDeckSF	OpenPorchSF	EnclosedPorch	3SsnPorch	ScreenPorch
0	0.000062	0.061385	0.014590	0.030419	0.045134	0.031935	0.036535	0.036310	0.000000	0.025377	...	0.000000	0.013889	0.00000	0.0	0.135887
1	0.000123	0.023019	0.029181	0.028670	0.033850	0.019161	0.035308	0.035473	0.000000	0.000000	...	0.000000	0.000000	0.05146	0.0	0.000000
2	0.000247	0.010231	0.036476	0.038090	0.045134	0.031935	0.036737	0.036492	0.008637	0.001394	...	0.033820	0.048150	0.00000	0.0	0.000000
3	0.000370	0.010231	0.034044	0.028989	0.039492	0.031935	0.036627	0.036383	0.039763	0.081298	...	0.051435	0.016667	0.00000	0.0	0.000000
4	0.000432	0.010231	0.036476	0.031059	0.039492	0.031935	0.036627	0.036401	0.027866	0.000000	...	0.017380	0.108338	0.00000	0.0	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
761	0.061363	0.046039	0.029181	0.027218	0.016925	0.025548	0.035711	0.035473	0.000000	0.000000	...	0.000000	0.000000	0.00000	0.0	0.000000
762	0.061487	0.025577	0.024317	0.019114	0.022567	0.038322	0.035455	0.035473	0.000000	0.039023	...	0.000000	0.000000	0.00000	0.0	0.000000
763	0.061548	0.010231	0.029181	0.022172	0.022567	0.038322	0.036078	0.035837	0.000000	0.021776	...	0.022547	0.000000	0.00000	0.0	0.000000
764	0.061610	0.010231	0.037935	0.032512	0.045134	0.031935	0.036755	0.036510	0.076266	0.001916	...	0.033820	0.045835	0.00000	0.0	0.000000
765	0.061734	0.097193	0.036476	0.037032	0.028209	0.025548	0.035986	0.035746	0.000000	0.048837	...	0.000000	0.000000	0.00000	0.0	0.000000

766 rows × 38 columns

Figure 5: Snippet of Normalized training dataset

	Id	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	...	WoodDeckSF	OpenPorchSF	EnclosedPorch	3SsnPorch	ScreenPorch
0	0.000200	0.015492	0.056154	0.039590	0.050746	0.046762	0.052682	0.052327	0.097490	0.081059	...	0.000000	0.000000	0.000000	0.0	0.156461
1	0.000401	0.069713	0.054751	0.023488	0.059203	0.046762	0.053759	0.053423	0.050382	0.102025	...	0.000000	0.000000	0.000000	0.0	0.000000
2	0.001002	0.139426	0.024568	0.012226	0.050746	0.046762	0.053974	0.053610	0.020153	0.035775	...	0.000000	0.018212	0.000000	0.0	0.000000
3	0.001603	0.023238	0.039308	0.029809	0.042288	0.056114	0.051875	0.052139	0.000000	0.000000	...	0.000000	0.000000	0.102329	0.0	0.000000
4	0.001803	0.046475	0.067385	0.037621	0.076118	0.046762	0.054055	0.053690	0.038794	0.072953	...	0.000000	0.066342	0.000000	0.0	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
350	0.090966	0.054221	0.056154	0.055094	0.050746	0.074819	0.052009	0.052139	0.000000	0.039204	...	0.000000	0.000000	0.000000	0.0	0.000000
351	0.091166	0.092951	0.022462	0.014971	0.050746	0.046762	0.053786	0.053423	0.111597	0.093607	...	0.000000	0.129432	0.000000	0.0	0.000000
352	0.091567	0.038729	0.052645	0.151707	0.050746	0.074819	0.051363	0.053396	0.000000	0.000000	...	0.170558	0.026016	0.000000	0.0	0.142608
353	0.091967	0.054221	0.046327	0.030082	0.059203	0.084171	0.052252	0.053637	0.000000	0.021434	...	0.000000	0.039025	0.000000	0.0	0.000000
354	0.092168	0.061967	0.056154	0.031938	0.050746	0.056114	0.052628	0.053369	0.000000	0.064769	...	0.000000	0.018212	0.000000	0.0	0.145052

355 rows × 38 columns

Figure 6: Snippet of Normalized test dataset

Question 4 :

Regression Test Error:

Mean Square Error: ***Mean Squared Error on Test data (MSE): 4.557550034021317e-21***

Prediction results are attached in Regression_results.csv file.