MSDS 601 | Fall 2023 | Final Project

Rithvik Donnipadu, Ronel Solomon, Belinda Ong

Contents

1.0 Description of your dataset: resource, dimension, variable description, etc.

Resource: Kaggle's House Prices - Advanced Regression Techniques (Link). We used the training dataset.

Dimension: 1460 rows x 81 columns

Variable description:

1. [Target Variable] SalePrice - the property's sale price in dollars
2. MSSubClass: The type of dwelling

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

3. MSZoning: The general zoning classification

A	Agriculture
C	Commercial
FV	Floating Village Residential
I	Industrial
RH	Residential High Density
RL	Residential Low Density
RP	Residential Low Density Park
RM	Residential Medium Density

4. LotFrontage: Linear feet of street connected to property
5. LotArea: Lot size in square feet
6. Street: Type of road access

Grvl	Gravel
Pave	Paved

7. Alley: Type of alley access

Grvl	Gravel
Pave	Paved
NA	No alley access

8. LotShape: General shape of property

Reg	Regular
IR1	Slightly irregular
IR2	Moderately Irregular
IR3	Irregular

9. LandContour: Flatness of the property

Lvl	Near Flat/Level
Bnk	Banked - Quick and significant rise from street grade to building
HLS	Hillside - Significant slope from side to side
Low	Depression

10. Utilities: Type of utilities available

AllPub	All public Utilities (E,G,W,& S)
NoSewr	Electricity, Gas, and Water (Septic Tank)
NoSeWa	Electricity and Gas Only
ELO	Electricity only

11. LotConfig: Lot configuration

Inside	Inside lot
Corner	Corner lot
CulDSac	Cul-de-sac
FR2	Frontage on 2 sides of property
FR3	Frontage on 3 sides of property

12. LandSlope: Slope of property

Gtl	Gentle slope
Mod	Moderate Slope
Sev	Severe Slope

13. Neighborhood: Physical locations within Ames city limits

Blmngtn	Bloomington Heights
Blueste	Bluestem
BrDale	Briardale
BrkSide	Brookside
ClearCr	Clear Creek
CollgCr	College Creek
Crawfor	Crawford
Edwards	Edwards
Gilbert	Gilbert
IDOTRR	Iowa DOT and Rail Road
MeadowV	Meadow Village
Mitchel	Mitchell
Names	North Ames
NoRidge	Northridge
NPkVill	Northpark Villa
NridgHt	Northridge Heights
NWAmes	Northwest Ames
OldTown	Old Town
SWISU	South & West of Iowa State University
Sawyer	Sawyer
SawyerW	Sawyer West
Somerst	Somerset
StoneBr	Stone Brook
Timber	Timberland
Veenker	Veenker

14. Condition1: Proximity to main road or railroad

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

15. Condition2: Proximity to main road or railroad (if a second is present)

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

RRAn	Adjacent to North-South Railroad
PosN	Near positive off-site feature--park, greenbelt, etc.
PosA	Adjacent to postive off-site feature
RRNe	Within 200' of East-West Railroad
RRAe	Adjacent to East-West Railroad

16. BldgType: Type of dwelling

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

17. HouseStyle: Style of dwelling

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

18. OverallQual: Overall material and finish quality

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

19. OverallCond: Overall condition rating

10	Very Excellent
9	Excellent
8	Very Good
7	Good

6 Above Average
5 Average
4 Below Average
3 Fair
2 Poor
1 Very Poor

20. YearBuilt: Original construction date
21. YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)
22. RoofStyle: Type of roof

Flat	Flat
Gable	Gable
Gambrel	Gabrel (Barn)
Hip	Hip
Mansard	Mansard
Shed	Shed

23. RoofMatl: Roof material

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

24. Exterior1st: Exterior covering on house

AsbShng	Asbestos Shingles
AsphShn	Asphalt Shingles
BrkComm	Brick Common
BrkFace	Brick Face
CBlock	Cinder Block
CemntBd	Cement Board
HdBoard	Hard Board
ImStucc	Imitation Stucco
MetalSd	Metal Siding
Other	Other
Plywood	Plywood
PreCast	PreCast
Stone	Stone

Stucco	Stucco
VinylSd	Vinyl Siding
Wd Sdng	Wood Siding
WdShing	Wood Shingles

25. Exterior2nd: Exterior covering on house (if more than one material)

AsbShng	Asbestos Shingles
AsphShn	Asphalt Shingles
BrkComm	Brick Common
BrkFace	Brick Face
CBlock	Cinder Block
CemntBd	Cement Board
HdBoard	Hard Board
ImStucc	Imitation Stucco
MetalSd	Metal Siding
Other	Other
Plywood	Plywood
PreCast	PreCast
Stone	Stone
Stucco	Stucco
VinylSd	Vinyl Siding
Wd Sdng	Wood Siding
WdShing	Wood Shingles

26. MasVnrType: Masonry veneer type

BrkCmn	Brick Common
BrkFace	Brick Face
CBlock	Cinder Block
None	None
Stone	Stone

27. MasVnrArea: Masonry veneer area in square feet
28. ExterQual: Exterior material quality

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

29. ExterCond: Present condition of the material on the exterior

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

30. Foundation: Type of foundation

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

31. BsmtQual: Height of the basement

Ex	Excellent (100+ inches)
Gd	Good (90-99 inches)
TA	Typical (80-89 inches)
Fa	Fair (70-79 inches)
Po	Poor (<70 inches
NA	No Basement

32. BsmtCond: General condition of the basement

Ex	Excellent
Gd	Good
TA	Typical - slight dampness allowed
Fa	Fair - dampness or some cracking or settling
Po	Poor - Severe cracking, settling, or wetness
NA	No Basement

33. BsmtExposure: Walkout or garden level basement walls

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

34. BsmtFinType1: Quality of basement finished area

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

35. BsmtFinSF1: Type 1 finished square feet
36. BsmtFinType2: Quality of second finished area (if present)

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

37. BsmtFinSF2: Type 2 finished square feet
38. BsmtUnfSF: Unfinished square feet of basement area
39. TotalBsmtSF: Total square feet of basement area
40. Heating: Type of heating

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

41. HeatingQC: Heating quality and condition

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

42. CentralAir: Central air conditioning

N	No
Y	Yes

43. Electrical: Electrical system

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

44. 1stFlrSF: First Floor square feet
45. 2ndFlrSF: Second floor square feet
46. LowQualFinSF: Low quality finished square feet (all floors)
47. GrLivArea: Above grade (ground) living area square feet
48. BsmtFullBath: Basement full bathrooms
49. BsmtHalfBath: Basement half bathrooms
50. FullBath: Full bathrooms above grade
51. HalfBath: Half baths above grade
52. Bedroom: Number of bedrooms above basement level
53. Kitchen: Number of kitchens
54. KitchenQual: Kitchen quality

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

55. TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
56. Functional: Home functionality rating

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

57. Fireplaces: Number of fireplaces
58. FireplaceQu: Fireplace quality

Ex	Excellent - Exceptional Masonry Fireplace
Gd	Good - Masonry Fireplace in main level

TA Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement
Fa Fair - Prefabricated Fireplace in basement
Po Poor - Ben Franklin Stove
NA No Fireplace

59. GarageType: Garage location

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

60. GarageYrBlt: Year garage was built
61. GarageFinish: Interior finish of the garage

Fin	Finished
RFn	Rough Finished
Unf	Unfinished
NA	No Garage

62. GarageCars: Size of garage in car capacity
63. GarageArea: Size of garage in square feet
64. GarageQual: Garage quality

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

65. GarageCond: Garage condition

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

66. PavedDrive: Paved driveway

Y	Paved
P	Partial Pavement
N	Dirt/Gravel

67. WoodDeckSF: Wood deck area in square feet
68. OpenPorchSF: Open porch area in square feet
69. EnclosedPorch: Enclosed porch area in square feet
70. 3SsnPorch: Three season porch area in square feet
71. ScreenPorch: Screen porch area in square feet
72. PoolArea: Pool area in square feet
73. PoolQC: Pool quality

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
NA	No Pool

74. Fence: Fence quality

GdPrv	Good Privacy
MnPrv	Minimum Privacy
GdWo	Good Wood
MnWw	Minimum Wood/Wire
NA	No Fence

75. MiscFeature: Miscellaneous feature not covered in other categories

Elev	Elevator
Gar2	2nd Garage (if not described in garage section)
Othr	Other
Shed	Shed (over 100 SF)
TenC	Tennis Court
NA	None

76. MiscVal: $Value of miscellaneous feature
77. MoSold: Month Sold
78. YrSold: Year Sold
79. SaleType: Type of sale

WD	Warranty Deed - Conventional
CWD	Warranty Deed - Cash
VWD	Warranty Deed - VA Loan
New	Home just constructed and sold

COD	Court Officer Deed/Estate
Con	Contract 15% Down payment regular terms
ConLw	Contract Low Down payment and low interest
ConLI	Contract Low Interest
ConLD	Contract Low Down
Oth	Other

80. SaleCondition: Condition of sale

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

2.0 Statement of the research problems, and a summary of methods

Problem statement: To predict the sale price of real estate based on size, location, condition, and transaction dimensions.

Summary of method:

Step 1: Determine the shape of the data as well as the names of the columns available - refer to *1.0 Description of your dataset: resource, dimension, variable description, etc.*

Step 2: Formulate an initial hypothesis on the most important variables because 79 is too overwhelming to perform a regression on ***SalePrice***. We achieved this by classifying the 79 variables into

1. Four dimensions:
    a. Size - related to how much space the property offers
    b. Location - related to where the property physically located
    c. Condition/Amenity - related to how well-maintained the property is and how valuable are its offerings (e.g., fence, air conditioning, etc.)
    d. Transaction - related to how/when the property was sold
2. Three prioritization categories:
    a. Must Have - directly impacts all of the liveable area or significantly impacts the quality of life of the inhabitants
    b. Good to Have - impacts a significant portion of the liveable area or impacts the quality of life of the inhabitants to a limited extent
    c. Nice to Have - Impacts a small portion of the liveable area or has very limited impact on the quality of life of the inhabitants

Our initial model (below) is based on all of the variables in the Must Have category and also contains variables from all four dimensions - refer to *Variable Classification Table*

Initial Model

SalePrice ~ LotArea + BsmtFinSF1 + TotalBsmtSF + GrLivArea + GarageCars + MSZoning + Condition1 + Neighborhood + C(OverallQual) + OverallCond + YearBuilt + Fence + YrSold + SaleCondition

Variable Classification Table

	Prioritization Categories		
Dimensions	**Must have**	**Good to have**	**Nice to have**
Size	5. LotArea 35. BsmtFinSF1 39. TotalBsmtSF 47. GrLivArea 62. GarageCars	2. MSSubClass [C] 4. LotFrontage 8. LotShape [C] 9. LandContour [C] 11. LotConfig [C] 12. LandSlope [C] 16. BldgType [C] 17. HouseStyle [C] 50. FullBath 51. HalfBath 52. Bedroom 53. Kitchen 59. GarageType [C] 63. GarageArea	37. BsmtFinSF2 38. BsmtUnfSF 44. 1stFlrSF 45. 2ndFlrSF 46. LowQualFinSF 48. BsmtFullBath 49. BsmtHalfBath 55. TotRmsAbvGrd 57. Fireplaces 67. WoodDeckSF 68. OpenPorchSF 69. EnclosedPorch 70. 3SsnPorch 71. ScreenPorch 72. PoolArea
Location	3. MSZoning [C] 13. Neighborhood [C] 14. Condition1 [C]	15. Condition2 [C]	6. Street [C] 7. Alley [C]
Condition / Amenity	18. OverallQual [O] 19. OverallCond [O] 20. YearBuilt 74. Fence [C]	54. KitchenQual [C] 56. Functional [C] 21. YearRemodAdd 10. Utilities [C] 33. BsmtExposure [C] 34. BsmtFinType1 [C] 40. Heating [C] 42. CentralAir [C] 43. Electrical [C] 61. GarageFinish [C] 66. PavedDrive [C]	28. ExterQual [C] 29. ExterCond [C] 31. BsmtCond [C] 41. HeatingQC [C] 58. FireplaceQu [C] 60. GarageYrBlt 64. GarageQual [C] 65. GarageCond [C] 73. PoolQual [C] 22. RoofStyle [C] 23. RoofMatl [C] 24. Exterior1st [C] 25. Exterior2nd [C] 26. MasVnrType [C] 27. MasVnrArea 30. Foundation [C] 31. BsmtQual [C] 36. BsmtFinType2 [C] 75. MiscFeature [C] 76. MiscVal
Transaction	78. YrSold 80. SaleCondition [C]	77. MoSold 79. SaleType [C]	

Legend

[C]: Categorical variables

[O]: Ordinal variables

Step 3: Conduct Exploratory Data Analysis to better understand the data. This included calculating a correlation matrix, identifying variables with a large proportion of NULL values, and also plotting out histogram and scatterplots of the hypothesized predictors to use.

Step 4: Conduct checks on the Data Structure to ensure that it does not violate key statistical inference and/or modeling assumptions. This included evaluating multicollinearity through VIF and an autocorrelation plot; and identified influential points through calculating Externalized Studentized Residuals and Cook's Distance. We removed predictors that exhibited strong multicollinearity and removed outliers from the dataset.

Step 5: Conduct checks on the model assumptions. This included evaluating heteroskedasticity through the Breush-Pagan test; evaluating normality through the Kolmogorov–Smirnov test, Jarque-Bera test, plotting out the residual versus fitted values, and creating a QQ plot; and evaluating the presence of a linear relationship between each of the predictors chosen and the response variable. Additionally, we also attempted a log normal transformation to address any deviations from the model assumptions.

Step 6: Finalize model selection. We ran the ANOVA typ=1 test to confirm the order of the predictors and identify the most important ones. We also calculated Mallow's Cp, Adjusted R^2, AIC, and BIC, for the pool of all possible models given our selected predictors.

From step 4 onwards, between each step, we ran the regression model and compared the adjusted R^2 between models, and also removed variables that had t-statistic p-values less than our chosen level of significance: 5%.

Based on multiple linear regression, the final model is

Final Model
SalePrice ~ LotArea + BsmtFinSF1 + SaleCondition + TotalBsmtSF + GarageCars + GrLivArea + Neighborhood + C(OverallQual) + OverallCond + YearBuilt

2.1 Extending multiple linear regression analysis

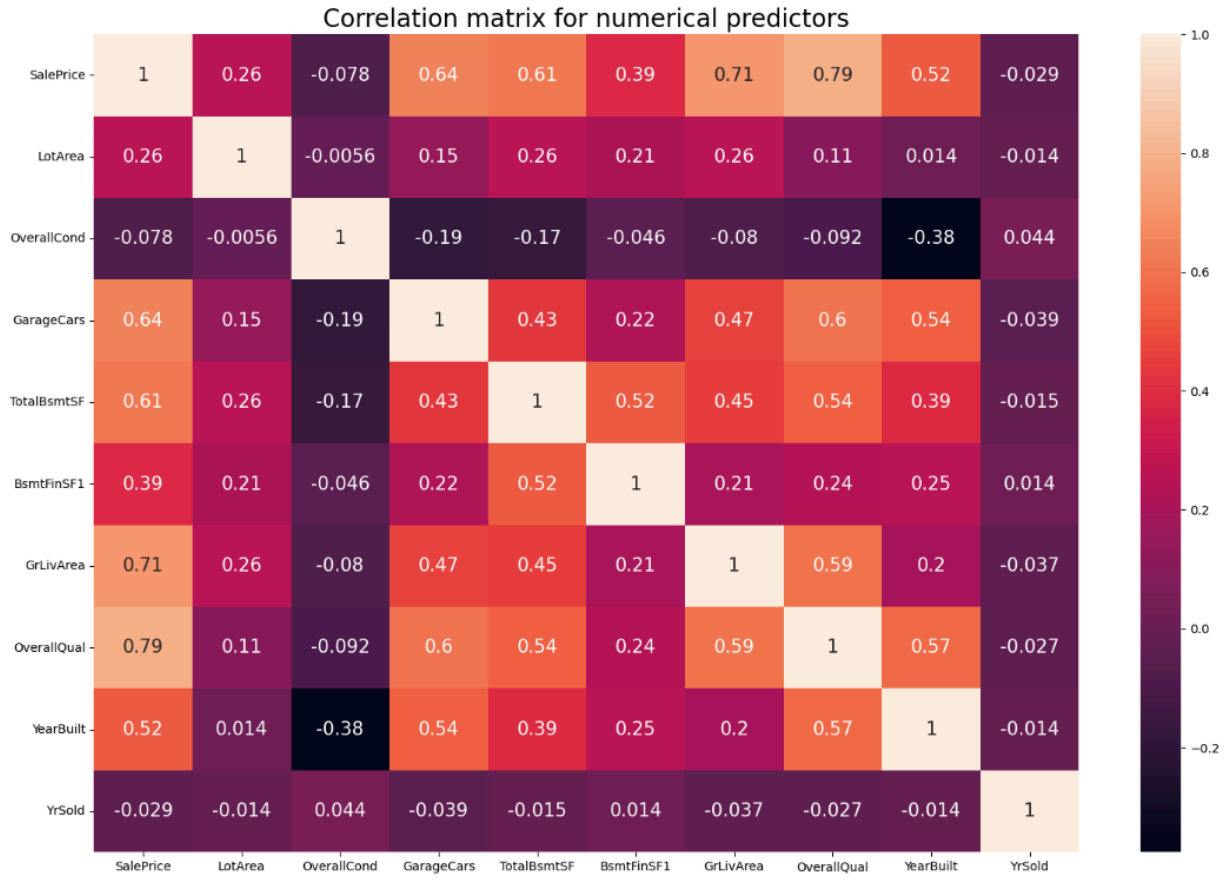
To extend our multiple linear regression analysis, we first calculated predicted values using the testing data set. Due to the high MSE calculated, we decided to try logistic regression. To that, we created two new binary variables

1. If the predicted sale price is above or equal to the median sale price
2. If the predicted sale price is above or equal to the mean sale price

The logistic regression yielded a 90%+ accuracy rate. Although high, we declined to continue the analysis as it was deemed that the findings did not directly address the original problem statement of predicting the sale price of real estate based on size, location, condition, and transaction dimensions.

3.0 Exploratory Data Analysis

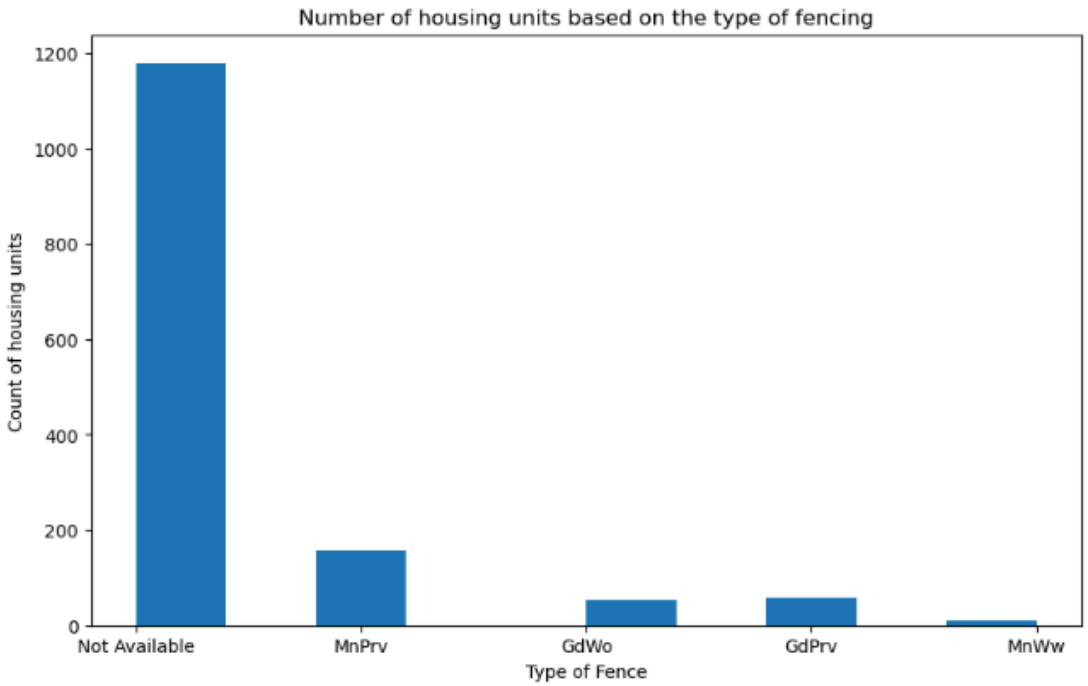
1. Based on the correlation matrix of our numerical predictors, OverallCond, YrSold, LotArea, and BsmtFinSF1 have low correlation and are good to have in the model



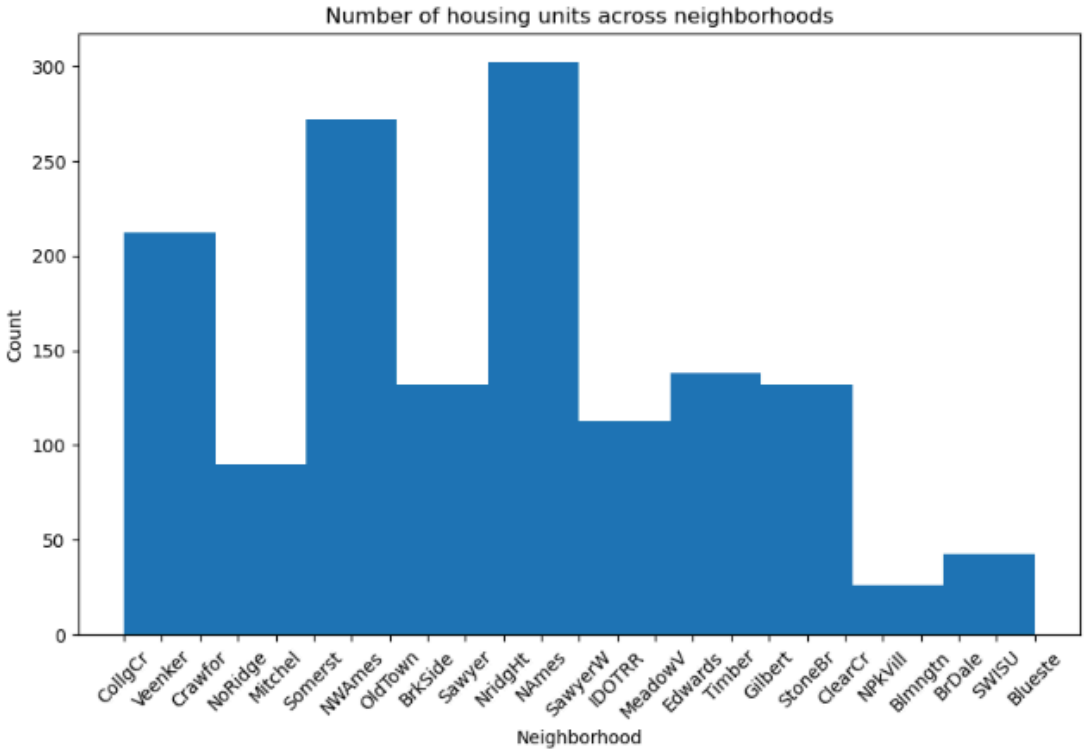
2. Fence was noted to have 80% NULL values so we filled those in as NA. The resulting distribution of values showcased strong skewness

```
df['Fence']=df['Fence'].fillna('NA')
df['Fence'].unique()
df['Fence'].value_counts()
```

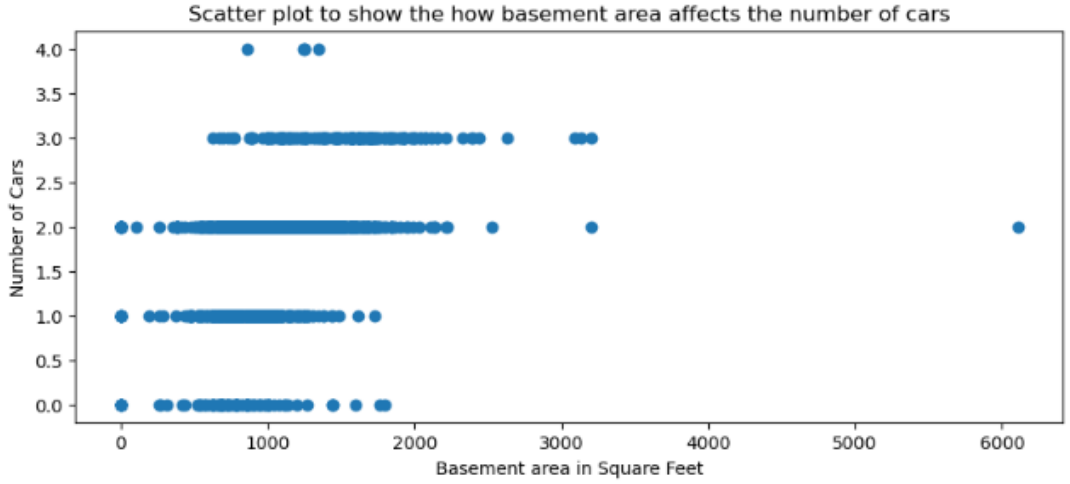
```
NA       1179
MnPrv     157
GdPrv      59
GdWo       54
MnWw       11
Name: Fence, dtype: int64
```

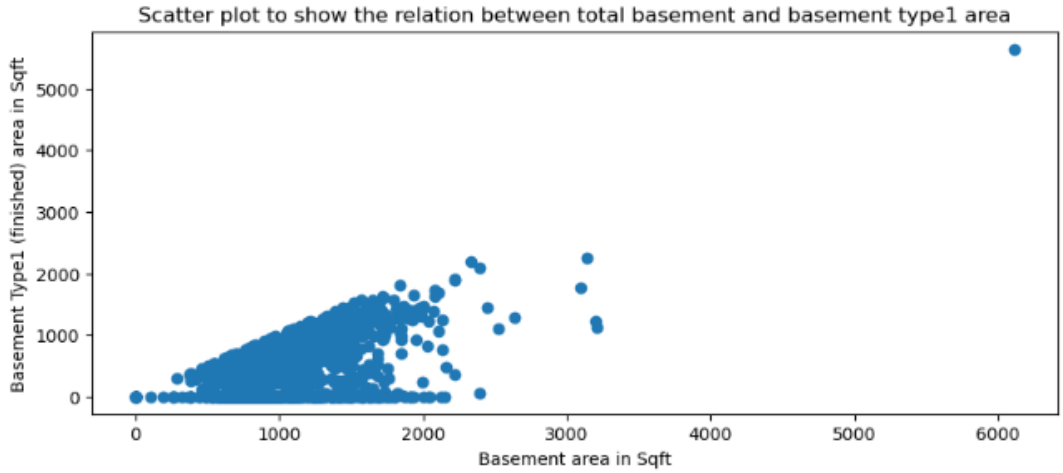


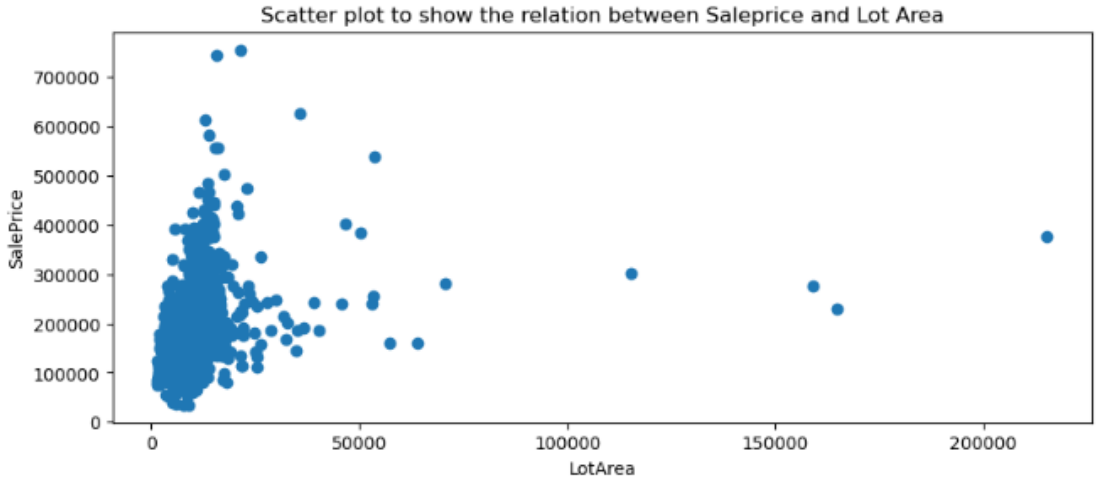
3. We plotted out a histogram of Neighborhoods and believe that the distribution is fairly evenly spread



4. We created scatterplots of GarageCars, BsmtFinSF1, and SalePrice and noticed clear outliers that we plant to address when correcting for issues in the data structure







With the above findings, we ran an initial regression analysis on the initial hypothesized model.

OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.937
Model:	OLS	**Adj. R-squared:**	0.923
Method:	Least Squares	**F-statistic:**	71.33
Date:	Sun, 08 Oct 2023	**Prob (F-statistic):**	3.04e-114
Time:	18:12:08	**Log-Likelihood:**	-3112.1
No. Observations:	281	**AIC:**	6322.
Df Residuals:	232	**BIC:**	6501.
Df Model:	48		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-9.101e+05	1.44e+06	-0.630	0.529	-3.75e+06	1.93e+06
MSZoning[T.FV]	6243.3285	3.93e+04	0.159	0.874	-7.12e+04	8.36e+04
MSZoning[T.RH]	2.6e+04	1.97e+04	1.323	0.187	-1.27e+04	6.47e+04
MSZoning[T.RL]	3.862e+04	1.38e+04	2.795	0.006	1.14e+04	6.58e+04
MSZoning[T.RM]	2.396e+04	1.24e+04	1.927	0.055	-531.542	4.84e+04
Condition1[T.Feedr]	4464.6895	6010.731	0.743	0.458	-7377.904	1.63e+04
Condition1[T.Norm]	7063.5235	5021.699	1.407	0.161	-2830.439	1.7e+04
Condition1[T.PosA]	-7.18e-08	1.14e-07	-0.630	0.530	-2.96e-07	1.53e-07
Condition1[T.PosN]	1.35e+04	1.44e+04	0.935	0.351	-1.5e+04	4.2e+04
Condition1[T.RRAe]	5502.8284	1.85e+04	0.297	0.767	-3.1e+04	4.2e+04
Condition1[T.RRAn]	3534.2799	8822.940	0.401	0.689	-1.38e+04	2.09e+04
Condition1[T.RRNe]	1.127e-08	1.79e-08	0.631	0.529	-2.39e-08	4.65e-08
Condition1[T.RRNn]	-1.011e+04	2.56e+04	-0.395	0.693	-6.05e+04	4.03e+04
Neighborhood[T.Blueste]	-5.09e+04	7.58e+04	-0.672	0.502	-2e+05	9.84e+04
Neighborhood[T.BrDale]	-4.696e-09	7.46e-09	-0.629	0.530	-1.94e-08	1e-08
Neighborhood[T.BrkSide]	-3.842e+04	7.37e+04	-0.522	0.602	-1.84e+05	1.07e+05
Neighborhood[T.ClearCr]	-6.169e+04	7.56e+04	-0.816	0.415	-2.11e+05	8.73e+04
Neighborhood[T.CollgCr]	-6.059e+04	7.48e+04	-0.811	0.418	-2.08e+05	8.67e+04
Neighborhood[T.Crawfor]	-2.624e+04	7.42e+04	-0.354	0.724	-1.72e+05	1.2e+05
Neighborhood[T.Edwards]	-5.769e+04	7.46e+04	-0.773	0.440	-2.05e+05	8.93e+04
Neighborhood[T.Gilbert]	-5.213e+04	7.72e+04	-0.675	0.500	-2.04e+05	1e+05
Neighborhood[T.IDOTRR]	-3.215e+04	7.38e+04	-0.436	0.663	-1.78e+05	1.13e+05
Neighborhood[T.MeadowV]	-5.84e+04	7.39e+04	-0.790	0.430	-2.04e+05	8.73e+04
Neighborhood[T.Mitchel]	-6.677e+04	7.44e+04	-0.898	0.370	-2.13e+05	7.98e+04
Neighborhood[T.NAmes]	-5.833e+04	7.41e+04	-0.787	0.432	-2.04e+05	8.77e+04
Neighborhood[T.NPkVill]	-1.23e-09	1.89e-09	-0.650	0.517	-4.96e-09	2.5e-09
Neighborhood[T.NWAmes]	-6.61e+04	7.43e+04	-0.890	0.374	-2.12e+05	8.02e+04
Neighborhood[T.NoRidge]	-7282.6623	7.46e+04	-0.098	0.922	-1.54e+05	1.4e+05
Neighborhood[T.NridgHt]	-3.965e-10	6.35e-10	-0.624	0.533	-1.65e-09	8.56e-10
Neighborhood[T.OldTown]	-4.82e+04	7.42e+04	-0.649	0.517	-1.94e+05	9.81e+04
Neighborhood[T.SWISU]	-5.107e+04	7.5e+04	-0.681	0.496	-1.99e+05	9.66e+04

Neighborhood[T.Sawyer]	-5.907e+04	7.44e+04	-0.794	0.428	-2.06e+05	8.76e+04
Neighborhood[T.SawyerW]	-4.956e+04	7.51e+04	-0.660	0.510	-1.98e+05	9.84e+04
Neighborhood[T.Somerst]	6243.3285	3.93e+04	0.159	0.874	-7.12e+04	8.36e+04
Neighborhood[T.StoneBr]	-4.158e-11	9.26e-12	-4.492	0.000	-5.98e-11	-2.33e-11
Neighborhood[T.Timber]	-6.501e+04	7.44e+04	-0.874	0.383	-2.12e+05	8.15e+04
Neighborhood[T.Veenker]	-6712.1937	7.54e+04	-0.089	0.929	-1.55e+05	1.42e+05
C(OverallQual)[T.2]	-2.785e-11	2.5e-11	-1.113	0.267	-7.72e-11	2.15e-11
C(OverallQual)[T.3]	-1.706e+05	2.07e+05	-0.824	0.411	-5.78e+05	2.37e+05
C(OverallQual)[T.4]	-1.831e+05	2.06e+05	-0.887	0.376	-5.9e+05	2.23e+05
C(OverallQual)[T.5]	-1.769e+05	2.06e+05	-0.859	0.391	-5.83e+05	2.29e+05
C(OverallQual)[T.6]	-1.729e+05	2.06e+05	-0.838	0.403	-5.79e+05	2.33e+05
C(OverallQual)[T.7]	-1.573e+05	2.06e+05	-0.762	0.447	-5.64e+05	2.49e+05
C(OverallQual)[T.8]	-1.362e+05	2.06e+05	-0.660	0.510	-5.43e+05	2.71e+05
C(OverallQual)[T.9]	-1.202e-11	3.41e-12	-3.522	0.001	-1.87e-11	-5.29e-12
C(OverallQual)[T.10]	8.694e+04	2.06e+05	0.421	0.674	-3.2e+05	4.94e+05
Fence[T.GdWo]	3045.9185	3846.181	0.792	0.429	-4531.989	1.06e+04
Fence[T.MnPrv]	4922.1765	3092.927	1.591	0.113	-1171.639	1.1e+04
Fence[T.MnWw]	2867.8372	6301.613	0.455	0.649	-9547.865	1.53e+04
SaleCondition[T.AdjLand]	0	0	nan	nan	0	0
SaleCondition[T.Alloca]	2.312e+04	1.34e+04	1.731	0.085	-3200.218	4.94e+04
SaleCondition[T.Family]	-6093.2085	7654.132	-0.796	0.427	-2.12e+04	8987.284
SaleCondition[T.Normal]	3494.8451	3558.091	0.982	0.327	-3515.456	1.05e+04
SaleCondition[T.Partial]	0	0	nan	nan	0	0
LotArea	1.6694	0.474	3.520	0.001	0.735	2.604
BsmtFinSF1	16.7162	4.113	4.064	0.000	8.612	24.820
TotalBsmtSF	12.7497	4.525	2.818	0.005	3.835	21.664
GrLivArea	51.1266	3.209	15.932	0.000	44.804	57.449
GarageCars	7246.3816	2019.352	3.588	0.000	3267.769	1.12e+04
OverallCond	8845.3628	1006.557	8.788	0.000	6862.201	1.08e+04
YearBuilt	694.6356	95.211	7.296	0.000	507.047	882.224
YrSold	-144.9681	856.538	-0.169	0.866	-1832.555	1542.619

Omnibus:	5.630	**Durbin-Watson:**	1.900
Prob(Omnibus):	0.060	**Jarque-Bera (JB):**	7.902
Skew:	0.074	**Prob(JB):**	0.0192
Kurtosis:	3.808	**Cond. No.**	1.05e+16

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 2.87e-22. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

The adjusted R^2 of 0.923 is good. However, we noticed that Condition1, Fence, and YearSold had p-values greater than our chosen significance level of 5% and removed them.

Model 2

SalePrice ~ LotArea + BsmtFinSF1 + TotalBsmtSF + GrLivArea + GarageCars + MSZoning + Neighborhood + C(OverallQual) + OverallCond + YearBuilt + SaleCondition

The new model then had adjusted R^2 of 0.853.

OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.857
Model:	OLS	**Adj. R-squared:**	0.853
Method:	Least Squares	**F-statistic:**	173.1
Date:	Sun, 08 Oct 2023	**Prob (F-statistic):**	0.00
Time:	18:12:25	**Log-Likelihood:**	-17122.
No. Observations:	1460	**AIC:**	3.434e+04
Df Residuals:	1410	**BIC:**	3.461e+04
Df Model:	49		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-7.444e+05	1.3e+05	-5.714	0.000	-1e+06	-4.89e+05
MSZoning[T.FV]	2.657e+04	1.44e+04	1.847	0.065	-1650.724	5.48e+04
MSZoning[T.RH]	2.484e+04	1.45e+04	1.719	0.086	-3505.391	5.32e+04
MSZoning[T.RL]	3.372e+04	1.2e+04	2.811	0.005	1.02e+04	5.72e+04
MSZoning[T.RM]	2.414e+04	1.13e+04	2.131	0.033	1921.679	4.64e+04
Neighborhood[T.Blueste]	-1.005e+04	2.34e+04	-0.430	0.667	-5.59e+04	3.58e+04
Neighborhood[T.BrDale]	-1.051e+04	1.19e+04	-0.882	0.378	-3.39e+04	1.29e+04
Neighborhood[T.BrkSide]	6472.6020	9941.268	0.651	0.515	-1.3e+04	2.6e+04
Neighborhood[T.ClearCr]	1.918e+04	1.01e+04	1.899	0.058	-629.870	3.9e+04
Neighborhood[T.CollgCr]	8383.5183	7949.656	1.055	0.292	-7210.907	2.4e+04
Neighborhood[T.Crawfor]	2.888e+04	9520.433	3.034	0.002	1.02e+04	4.76e+04
Neighborhood[T.Edwards]	-1.333e+04	8836.847	-1.508	0.132	-3.07e+04	4009.166
Neighborhood[T.Gilbert]	4182.6901	8419.337	0.497	0.619	-1.23e+04	2.07e+04
Neighborhood[T.IDOTRR]	5993.0420	1.14e+04	0.523	0.601	-1.65e+04	2.85e+04
Neighborhood[T.MeadowV]	-1.101e+04	1.2e+04	-0.921	0.357	-3.45e+04	1.25e+04
Neighborhood[T.Mitchel]	-2540.0366	9017.250	-0.282	0.778	-2.02e+04	1.51e+04
Neighborhood[T.NAmes]	-1375.0483	8384.315	-0.164	0.870	-1.78e+04	1.51e+04
Neighborhood[T.NPkVill]	-9277.1676	1.3e+04	-0.716	0.474	-3.47e+04	1.61e+04
Neighborhood[T.NWAmes]	-1368.6368	8642.594	-0.158	0.874	-1.83e+04	1.56e+04
Neighborhood[T.NoRidge]	5.373e+04	9313.900	5.768	0.000	3.55e+04	7.2e+04
Neighborhood[T.NridgHt]	3.41e+04	8547.759	3.989	0.000	1.73e+04	5.09e+04
Neighborhood[T.OldTown]	-5431.5998	1.02e+04	-0.532	0.595	-2.54e+04	1.46e+04
Neighborhood[T.SWISU]	-3594.6787	1.1e+04	-0.328	0.743	-2.51e+04	1.79e+04

Neighborhood[T.Sawyer]	-3911.5367	8898.972	-0.440	0.660	-2.14e+04	1.35e+04
Neighborhood[T.SawyerW]	5634.5717	8731.723	0.645	0.519	-1.15e+04	2.28e+04
Neighborhood[T.Somerst]	1.638e+04	1.02e+04	1.613	0.107	-3539.387	3.63e+04
Neighborhood[T.StoneBr]	4.53e+04	9949.908	4.552	0.000	2.58e+04	6.48e+04
Neighborhood[T.Timber]	1.09e+04	9279.585	1.175	0.240	-7299.449	2.91e+04
Neighborhood[T.Veenker]	2.686e+04	1.21e+04	2.213	0.027	3051.187	5.07e+04
C(OverallQual)[T.2]	-7048.0938	2.83e+04	-0.249	0.804	-6.26e+04	4.85e+04
C(OverallQual)[T.3]	-1.432e+04	2.3e+04	-0.624	0.533	-5.93e+04	3.07e+04
C(OverallQual)[T.4]	-1.213e+04	2.22e+04	-0.546	0.585	-5.57e+04	3.15e+04
C(OverallQual)[T.5]	-1.166e+04	2.23e+04	-0.524	0.600	-5.53e+04	3.2e+04
C(OverallQual)[T.6]	-5110.1462	2.24e+04	-0.228	0.819	-4.9e+04	3.88e+04
C(OverallQual)[T.7]	9565.1822	2.25e+04	0.425	0.671	-3.46e+04	5.38e+04
C(OverallQual)[T.8]	3.784e+04	2.28e+04	1.660	0.097	-6881.023	8.26e+04
C(OverallQual)[T.9]	1.039e+05	2.34e+04	4.442	0.000	5.8e+04	1.5e+05
C(OverallQual)[T.10]	1.177e+05	2.44e+04	4.820	0.000	6.98e+04	1.66e+05
SaleCondition[T.AdjLand]	2.331e+04	1.59e+04	1.461	0.144	-7980.736	5.46e+04
SaleCondition[T.Alloca]	2459.6204	9578.143	0.257	0.797	-1.63e+04	2.12e+04
SaleCondition[T.Family]	-3225.5044	7575.650	-0.426	0.670	-1.81e+04	1.16e+04
SaleCondition[T.Normal]	6833.7478	3262.910	2.094	0.036	433.068	1.32e+04
SaleCondition[T.Partial]	2.434e+04	4567.826	5.328	0.000	1.54e+04	3.33e+04
LotArea	0.4851	0.094	5.186	0.000	0.302	0.669
BsmtFinSF1	14.1427	2.208	6.406	0.000	9.812	18.474
TotalBsmtSF	9.0856	2.773	3.277	0.001	3.647	14.525
GrLivArea	46.6110	2.270	20.535	0.000	42.158	51.064
GarageCars	1.1e+04	1537.068	7.159	0.000	7988.385	1.4e+04
OverallCond	8484.0910	854.349	9.930	0.000	6808.160	1.02e+04
YearBuilt	363.2018	64.241	5.654	0.000	237.184	489.220

Omnibus:	1120.108	**Durbin-Watson:**	1.907
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	236368.519
Skew:	-2.610	**Prob(JB):**	0.00
Kurtosis:	65.115	**Cond. No.**	2.41e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.41e+06. This might indicate that there are strong multicollinearity or other numerical problems.

4.0 Data Structure

4.1 Multicollinearity

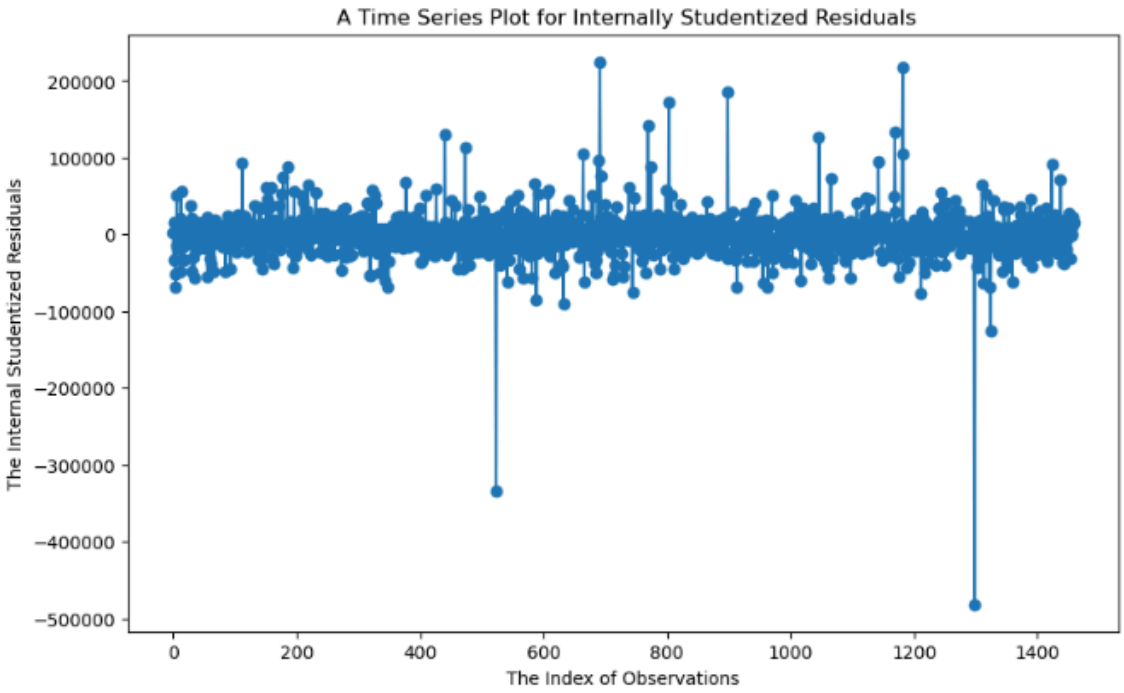
We calculated the VIF and noted that the following predictors exhibited high VIF > 10: MSZoning, Neighborhood, and OverallQual. We concluded that Neighborhood is ok because only two out of the 25 exhibited high VIF. We also recalculated the VIF of OverallQual by treating it as a numerical predictor and its VIF fell to 3. Therefore, the multicollinearity is likely due to the presence of multiple levels. Therefore, we decided to only remove MSZoning.

	VIF Factor	features
0	26627.878459	Intercept
1	13.814843	MSZoning[T.FV]
2	3.550552	MSZoning[T.RH]
3	37.652506	MSZoning[T.RL]
4	25.555713	MSZoning[T.RM]
5	1.171628	Neighborhood[T.Blueste]
6	2.415197	Neighborhood[T.BrDale]
7	5.914122	Neighborhood[T.BrkSide]
8	3.007831	Neighborhood[T.ClearCr]
9	9.138815	Neighborhood[T.CollgCr]
10	4.793196	Neighborhood[T.Crawfor]
11	7.815634	Neighborhood[T.Edwards]
12	5.691244	Neighborhood[T.Gilbert]
13	5.079332	Neighborhood[T.IDOTRR]
14	2.582931	Neighborhood[T.MeadowV]
15	4.137157	Neighborhood[T.Mitchel]
16	14.375246	Neighborhood[T.NAmes]
17	1.613234	Neighborhood[T.NPkVill]
18	5.565679	Neighborhood[T.NWAmes]
19	3.714155	Neighborhood[T.NoRidge]
20	5.725957	Neighborhood[T.NridgHt]
21	11.663692	Neighborhood[T.OldTown]
22	3.171007	Neighborhood[T.SWISU]
23	5.977304	Neighborhood[T.Sawyer]
24	4.637892	Neighborhood[T.SawyerW]
25	8.966904	Neighborhood[T.Somerst]
26	2.613730	Neighborhood[T.StoneBr]
27	3.424294	Neighborhood[T.Timber]
28	1.727315	Neighborhood[T.Veenker]
29	2.582932	C(OverallQual)[T.2]
30	11.164871	C(OverallQual)[T.3]
31	56.697084	C(OverallQual)[T.4]
32	153.827537	C(OverallQual)[T.5]
33	149.636307	C(OverallQual)[T.6]
34	135.930150	C(OverallQual)[T.7]
35	83.034631	C(OverallQual)[T.8]
36	24.519475	C(OverallQual)[T.9]
37	11.399223	C(OverallQual)[T.10]
38	1.090256	SaleCondition[T.AdjLand]
39	1.173122	SaleCondition[T.Alloca]
40	1.216362	SaleCondition[T.Family]
41	2.459228	SaleCondition[T.Normal]
42	2.562365	SaleCondition[T.Partial]
43	1.366836	LotArea
44	1.589486	BsmtFinSF1
45	2.319462	TotalBsmtSF
46	2.230128	GrLivArea
47	2.068390	GarageCars
48	1.416911	OverallCond
49	5.901477	YearBuilt

We then reran the VIF calculations to confirm that there were no further violations.

	VIF Factor	features
0	25831.318181	Intercept
1	1.138173	Neighborhood[T.Blueste]
2	2.123752	Neighborhood[T.BrDale]
3	5.492797	Neighborhood[T.BrkSide]
4	3.004099	Neighborhood[T.ClearCr]
5	9.133786	Neighborhood[T.CollgCr]
6	4.774964	Neighborhood[T.Crawfor]
7	7.762053	Neighborhood[T.Edwards]
8	5.681165	Neighborhood[T.Gilbert]
9	4.009672	Neighborhood[T.IDOTRR]
10	2.235409	Neighborhood[T.MeadowV]
11	4.126163	Neighborhood[T.Mitchel]
12	14.351189	Neighborhood[T.NAmes]
13	1.611821	Neighborhood[T.NPkVill]
14	5.560946	Neighborhood[T.NWAmes]
15	3.713373	Neighborhood[T.NoRidge]
16	5.721294	Neighborhood[T.NridgHt]
17	9.777602	Neighborhood[T.OldTown]
18	3.122536	Neighborhood[T.SWISU]
19	5.962311	Neighborhood[T.Sawyer]
20	4.611397	Neighborhood[T.SawyerW]
21	5.867467	Neighborhood[T.Somerst]
22	2.612111	Neighborhood[T.StoneBr]
23	3.423521	Neighborhood[T.Timber]
24	1.727222	Neighborhood[T.Veenker]
25	2.546595	C(OverallQual)[T.2]
26	11.142503	C(OverallQual)[T.3]
27	56.533680	C(OverallQual)[T.4]
28	153.251159	C(OverallQual)[T.5]
29	148.886284	C(OverallQual)[T.6]
30	135.366290	C(OverallQual)[T.7]
31	82.706853	C(OverallQual)[T.8]
32	24.434432	C(OverallQual)[T.9]
33	11.346727	C(OverallQual)[T.10]
34	1.088194	SaleCondition[T.AdjLand]
35	1.166083	SaleCondition[T.Alloca]
36	1.209735	SaleCondition[T.Family]
37	2.404594	SaleCondition[T.Normal]
38	2.532483	SaleCondition[T.Partial]
39	1.360648	LotArea
40	1.577756	BsmtFinSF1
41	2.258180	TotalBsmtSF
42	2.216769	GrLivArea
43	2.041008	GarageCars
44	1.412784	OverallCond
45	5.789996	YearBuilt

Additionally, we also plotted out the autocorrelation and concluded that there is no clearly discernible pattern and hence no further multicollinearity.



The resulting model without MSZoning had an adjusted R^2 of 0.852.

Model 3

SalePrice ~ LotArea + BsmtFinSF1 + TotalBsmtSF + GrLivArea + GarageCars + Neighborhood + C(OverallQual) + OverallCond + YearBuilt + SaleCondition

OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.856
Model:	OLS	**Adj. R-squared:**	0.852
Method:	Least Squares	**F-statistic:**	187.2
Date:	Sun, 08 Oct 2023	**Prob (F-statistic):**	0.00
Time:	18:13:12	**Log-Likelihood:**	-17128.
No. Observations:	1460	**AIC:**	3.435e+04
Df Residuals:	1414	**BIC:**	3.459e+04
Df Model:	45		
Covariance Type:	nonrobust		

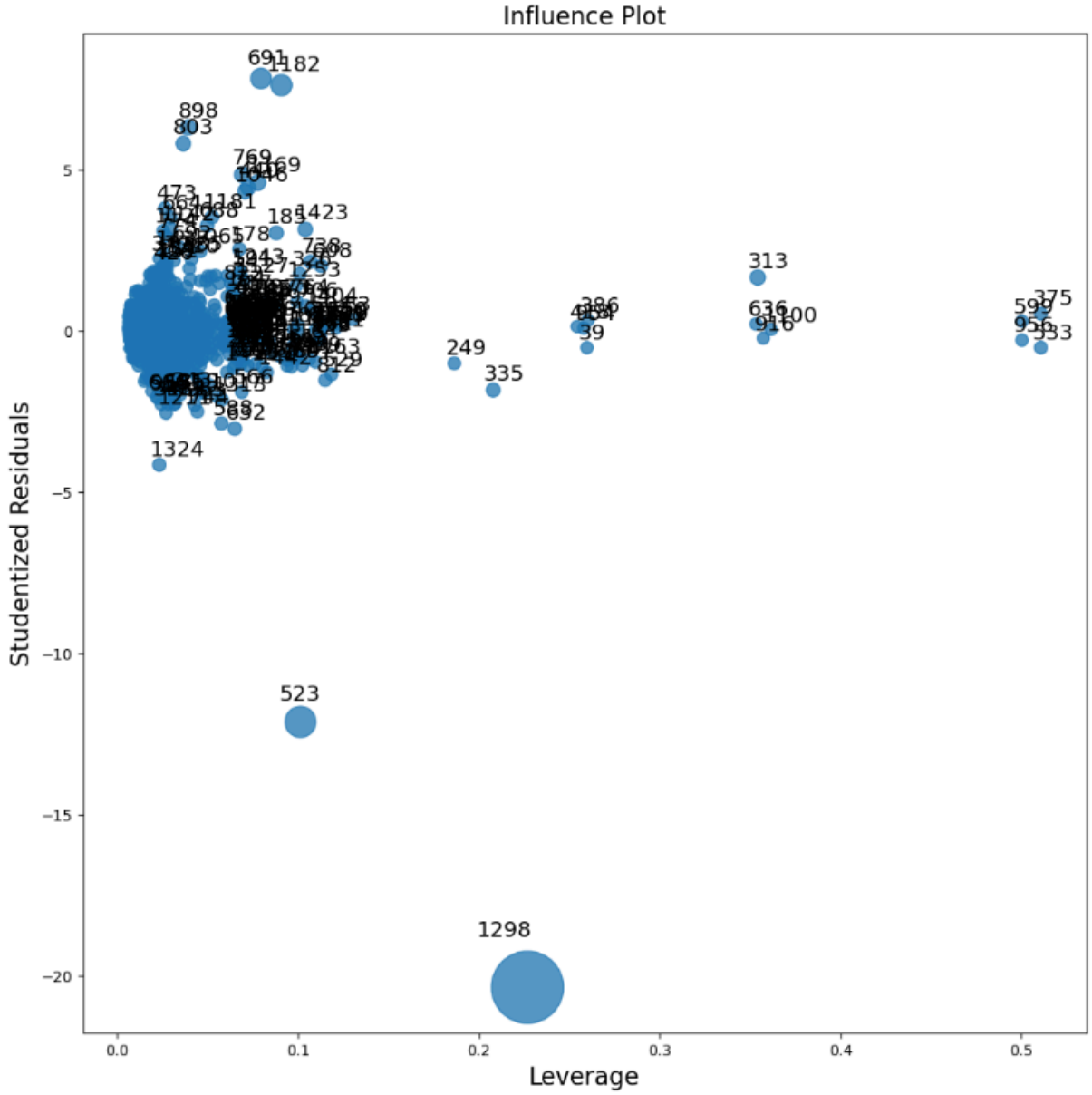
Omnibus:	1112.984	**Durbin-Watson:**	1.899
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	229275.528
Skew:	-2.588	**Prob(JB):**	0.00
Kurtosis:	64.173	**Cond. No.**	2.37e+06

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-6.897e+05	1.29e+05	-5.359	0.000	-9.42e+05	-4.37e+05
Neighborhood[T.Blueste]	-1.858e+04	2.31e+04	-0.804	0.421	-6.39e+04	2.67e+04
Neighborhood[T.BrDale]	-1.914e+04	1.12e+04	-1.708	0.088	-4.11e+04	2841.477
Neighborhood[T.BrkSide]	1506.8965	9607.593	0.157	0.875	-1.73e+04	2.04e+04
Neighborhood[T.ClearCr]	1.922e+04	1.01e+04	1.899	0.058	-631.047	3.91e+04
Neighborhood[T.CollgCr]	8297.7343	7969.846	1.041	0.298	-7336.260	2.39e+04
Neighborhood[T.Crawfor]	2.795e+04	9529.064	2.933	0.003	9256.857	4.66e+04
Neighborhood[T.Edwards]	-1.408e+04	8831.300	-1.594	0.111	-3.14e+04	3246.080
Neighborhood[T.Gilbert]	4997.6842	8435.564	0.592	0.554	-1.15e+04	2.15e+04
Neighborhood[T.IDOTRR]	-9182.2098	1.02e+04	-0.900	0.368	-2.92e+04	1.08e+04
Neighborhood[T.MeadowV]	-2.008e+04	1.12e+04	-1.799	0.072	-4.2e+04	1812.725
Neighborhood[T.Mitchel]	-2916.3690	9030.617	-0.323	0.747	-2.06e+04	1.48e+04
Neighborhood[T.NAmes]	-1188.8146	8400.884	-0.142	0.887	-1.77e+04	1.53e+04
Neighborhood[T.NPkVill]	-8205.3712	1.3e+04	-0.632	0.528	-3.37e+04	1.73e+04
Neighborhood[T.NWAmes]	-856.3263	8663.243	-0.099	0.921	-1.79e+04	1.61e+04
Neighborhood[T.NoRidge]	5.417e+04	9339.141	5.800	0.000	3.58e+04	7.25e+04
Neighborhood[T.NridgHt]	3.426e+04	8568.336	3.999	0.000	1.75e+04	5.11e+04
Neighborhood[T.OldTown]	-1.395e+04	9369.123	-1.489	0.137	-3.23e+04	4424.443
Neighborhood[T.SWISU]	-5201.6889	1.09e+04	-0.477	0.633	-2.66e+04	1.62e+04
Neighborhood[T.Sawyer]	-3963.8669	8912.830	-0.445	0.657	-2.14e+04	1.35e+04
Neighborhood[T.SawyerW]	5765.3322	8731.261	0.660	0.509	-1.14e+04	2.29e+04
Neighborhood[T.Somerst]	1.16e+04	8237.369	1.408	0.159	-4561.634	2.78e+04
Neighborhood[T.StoneBr]	4.553e+04	9974.833	4.564	0.000	2.6e+04	6.51e+04
Neighborhood[T.Timber]	1.115e+04	9304.665	1.199	0.231	-7099.047	2.94e+04
Neighborhood[T.Veenker]	2.704e+04	1.22e+04	2.222	0.026	3166.235	5.09e+04
C(OverallQual)[T.2]	-1.689e+04	2.82e+04	-0.599	0.549	-7.22e+04	3.85e+04
C(OverallQual)[T.3]	-1.78e+04	2.3e+04	-0.774	0.439	-6.29e+04	2.73e+04
C(OverallQual)[T.4]	-1.549e+04	2.23e+04	-0.696	0.487	-5.92e+04	2.82e+04
C(OverallQual)[T.5]	-1.479e+04	2.23e+04	-0.664	0.507	-5.85e+04	2.89e+04
C(OverallQual)[T.6]	-8801.5777	2.24e+04	-0.393	0.694	-5.27e+04	3.51e+04
C(OverallQual)[T.7]	6289.4960	2.25e+04	0.279	0.780	-3.79e+04	5.05e+04
C(OverallQual)[T.8]	3.496e+04	2.28e+04	1.532	0.126	-9802.957	7.97e+04
C(OverallQual)[T.9]	1.011e+05	2.34e+04	4.320	0.000	5.52e+04	1.47e+05
C(OverallQual)[T.10]	1.143e+05	2.44e+04	4.677	0.000	6.64e+04	1.62e+05
SaleCondition[T.AdjLand]	2.517e+04	1.6e+04	1.575	0.115	-6177.713	5.65e+04
SaleCondition[T.Alloca]	1761.2715	9576.254	0.184	0.854	-1.7e+04	2.05e+04
SaleCondition[T.Family]	-2599.1822	7576.257	-0.343	0.732	-1.75e+04	1.23e+04
SaleCondition[T.Normal]	7829.1740	3235.547	2.420	0.016	1482.186	1.42e+04
SaleCondition[T.Partial]	2.554e+04	4553.900	5.609	0.000	1.66e+04	3.45e+04
LotArea	0.5016	0.094	5.359	0.000	0.318	0.685
BsmtFinSF1	13.8256	2.206	6.268	0.000	9.499	18.153
TotalBsmtSF	9.6674	2.744	3.524	0.000	4.286	15.049
GrLivArea	46.5887	2.269	20.529	0.000	42.137	51.040
GarageCars	1.079e+04	1531.159	7.050	0.000	7791.007	1.38e+04
OverallCond	8638.2870	855.506	10.097	0.000	6960.090	1.03e+04
YearBuilt	352.9809	63.811	5.532	0.000	227.807	478.154

4.2 Influential Points

Based on calculating the externalized studentized residuals and Cook's Distance, we defined outliers as points identified by both methods and removed the following indexes: 769, 898, 4, 774, 523, 1423, 913, 1169, 1298, 1046, 664, 541, 1181, 1182, 1310, 1313, 1437, 803, 1065, 1322, 1324, 688, 178, 691, 692, 440, 185, 1211, 963, 585, 588, 1359, 343, 473, 218, 348, 608, 738, 744, 112, 1142, 632, 1017.

Going forward we will be using this newly filtered data set.



The Adjusted R^2 of the model with the existing predictors but without the outliers is 0.927.

OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.929
Model:	OLS	**Adj. R-squared:**	0.927
Method:	Least Squares	**F-statistic:**	400.2
Date:	Sun, 08 Oct 2023	**Prob (F-statistic):**	0.00
Time:	18:19:38	**Log-Likelihood:**	-15921.
No. Observations:	1417	**AIC:**	3.193e+04
Df Residuals:	1371	**BIC:**	3.218e+04
Df Model:	45		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-8.764e+05	7.98e+04	-10.975	0.000	-1.03e+06	-7.2e+05
Neighborhood[T.Blueste]	-1.032e+04	1.41e+04	-0.733	0.464	-3.79e+04	1.73e+04
Neighborhood[T.BrDale]	-1.179e+04	6845.623	-1.722	0.085	-2.52e+04	1641.839
Neighborhood[T.BrkSide]	9563.2683	5883.157	1.626	0.104	-1977.695	2.11e+04
Neighborhood[T.ClearCr]	1.691e+04	6280.718	2.693	0.007	4592.448	2.92e+04
Neighborhood[T.CollgCr]	6794.3546	4865.265	1.397	0.163	-2749.815	1.63e+04
Neighborhood[T.Crawfor]	2.739e+04	5909.548	4.634	0.000	1.58e+04	3.9e+04
Neighborhood[T.Edwards]	-2372.9115	5417.517	-0.438	0.661	-1.3e+04	8254.609
Neighborhood[T.Gilbert]	8793.5877	5151.314	1.707	0.088	-1311.723	1.89e+04
Neighborhood[T.IDOTRR]	-163.8084	6246.392	-0.026	0.979	-1.24e+04	1.21e+04
Neighborhood[T.MeadowV]	-1.56e+04	6809.452	-2.290	0.022	-2.9e+04	-2237.334
Neighborhood[T.Mitchel]	-3003.7523	5510.866	-0.545	0.586	-1.38e+04	7806.891
Neighborhood[T.NAmes]	1039.7920	5132.183	0.203	0.839	-9027.990	1.11e+04
Neighborhood[T.NPkVill]	-3420.4377	7921.426	-0.432	0.666	-1.9e+04	1.21e+04
Neighborhood[T.NWAmes]	575.3448	5296.788	0.109	0.914	-9815.343	1.1e+04
Neighborhood[T.NoRidge]	3.881e+04	5858.494	6.624	0.000	2.73e+04	5.03e+04
Neighborhood[T.NridgHt]	2.281e+04	5316.322	4.290	0.000	1.24e+04	3.32e+04
Neighborhood[T.OldTown]	-5916.4976	5738.693	-1.031	0.303	-1.72e+04	5341.073
Neighborhood[T.SWISU]	984.4784	6669.395	0.148	0.883	-1.21e+04	1.41e+04
Neighborhood[T.Sawyer]	-1582.9088	5442.608	-0.291	0.771	-1.23e+04	9093.832
Neighborhood[T.SawyerW]	7529.4353	5332.698	1.412	0.158	-2931.696	1.8e+04
Neighborhood[T.Somerst]	1.311e+04	5031.463	2.605	0.009	3237.748	2.3e+04
Neighborhood[T.StoneBr]	2.615e+04	6419.810	4.073	0.000	1.36e+04	3.87e+04
Neighborhood[T.Timber]	3947.9308	5731.488	0.689	0.491	-7295.505	1.52e+04
Neighborhood[T.Veenker]	2.452e+04	7424.033	3.303	0.001	9956.148	3.91e+04

C(OverallQual)[T.2]	-1.184e+04	1.72e+04	-0.689	0.491	-4.56e+04	2.19e+04
C(OverallQual)[T.3]	-1.532e+04	1.4e+04	-1.093	0.275	-4.28e+04	1.22e+04
C(OverallQual)[T.4]	-1.5e+04	1.36e+04	-1.106	0.269	-4.16e+04	1.16e+04
C(OverallQual)[T.5]	-1.437e+04	1.36e+04	-1.058	0.290	-4.1e+04	1.23e+04
C(OverallQual)[T.6]	-9662.2092	1.37e+04	-0.708	0.479	-3.64e+04	1.71e+04
C(OverallQual)[T.7]	5853.0826	1.38e+04	0.426	0.671	-2.11e+04	3.28e+04
C(OverallQual)[T.8]	3.364e+04	1.39e+04	2.414	0.016	6298.027	6.1e+04
C(OverallQual)[T.9]	9.228e+04	1.44e+04	6.426	0.000	6.41e+04	1.2e+05
C(OverallQual)[T.10]	1.166e+05	1.52e+04	7.670	0.000	8.68e+04	1.46e+05
SaleCondition[T.AdjLand]	2.131e+04	9746.181	2.186	0.029	2187.000	4.04e+04
SaleCondition[T.Alloca]	-1.821e+04	6674.683	-2.728	0.006	-3.13e+04	-5111.898
SaleCondition[T.Family]	2139.7636	4720.171	0.453	0.650	-7119.777	1.14e+04
SaleCondition[T.Normal]	1.006e+04	1991.866	5.052	0.000	6154.944	1.4e+04
SaleCondition[T.Partial]	2.601e+04	2882.933	9.020	0.000	2.03e+04	3.17e+04
LotArea	0.6016	0.058	10.385	0.000	0.488	0.715
BsmtFinSF1	18.5399	1.429	12.976	0.000	15.737	21.343
TotalBsmtSF	19.4673	1.733	11.233	0.000	16.068	22.867
GrLivArea	48.8629	1.466	33.322	0.000	45.986	51.740
GarageCars	8290.6296	951.581	8.712	0.000	6423.917	1.02e+04
OverallCond	8477.5550	527.385	16.075	0.000	7442.987	9512.123
YearBuilt	440.1416	39.577	11.121	0.000	362.504	517.779

Omnibus:	34.610	**Durbin-Watson:**	1.966
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	74.029
Skew:	0.066	**Prob(JB):**	8.41e-17
Kurtosis:	4.112	**Cond. No.**	2.34e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.34e+06. This might indicate that there are strong multicollinearity or other numerical problems.

5.0 Model Assumptions

5.1 Heteroskedasticity

We calculated the Breusch-Pagan test statistic:

LM Statistic: 270.392
LM Test p-value: 6.285e-34

Because the p-value of 6.285e-34 < 0.05, we reject the null hypothesis and conclude that there is heteroscedasticity.

5.2 Normality

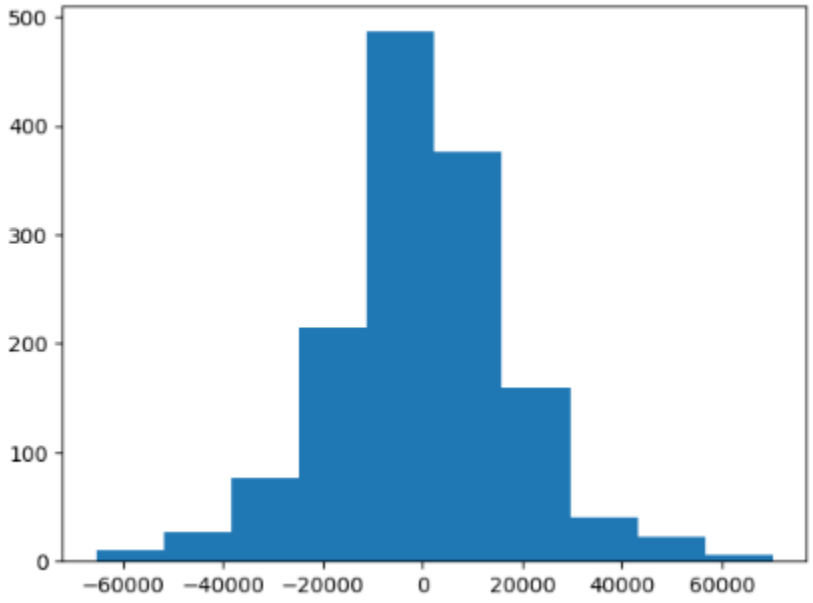
KS Statistic: 0.507
KS Test p-value: 0.000
KS Statistic Location: -25.967

Because the p-value of 0.000 < 0.05, we reject the null hypothesis and conclude that there is non-normality.

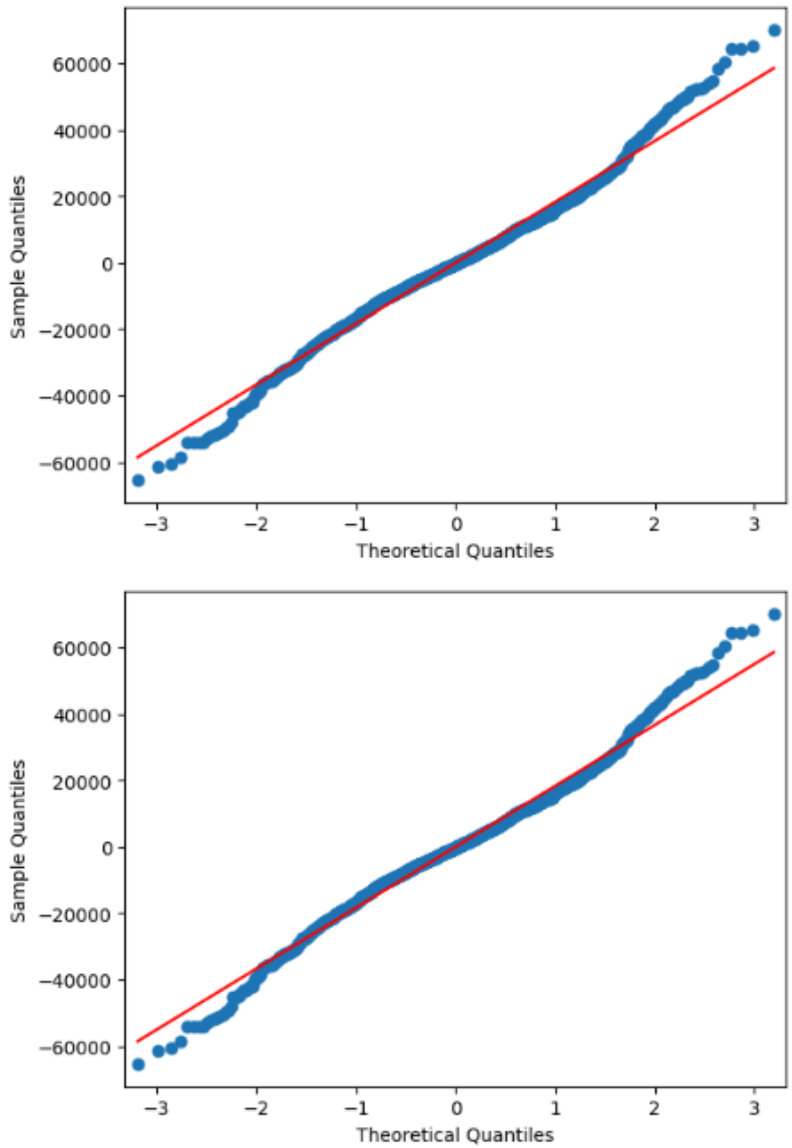
Jarque-Bera Statistic: 74.029
Jarque-Bera Test p-value: 8.409e-17
Skewness of Residuals: 0.066
Kurtosis of Residuals: 4.112

Because the p-value of 8.409e-17 < 0.05, we reject the null hypothesis and conclude that there is non-normality.

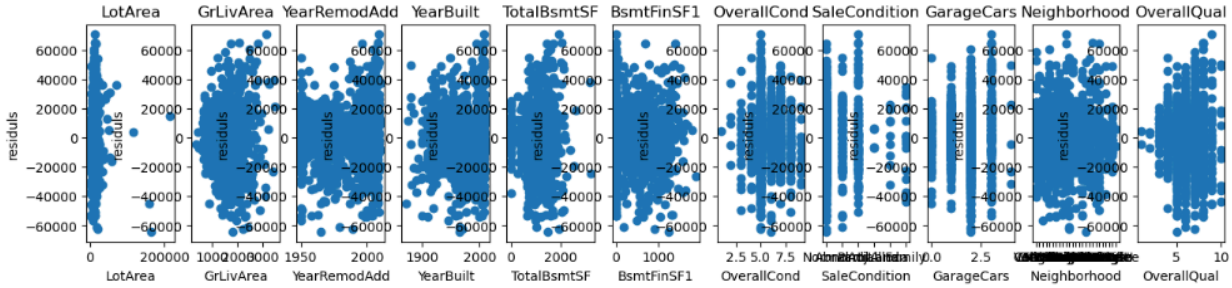
Based on a histogram of residuals, we noted that the distribution is approximately normal. Therefore, there is no violation of normality



Based on the QQ Plot, the distribution is approximately linear. Therefore, there is no violation of normality.

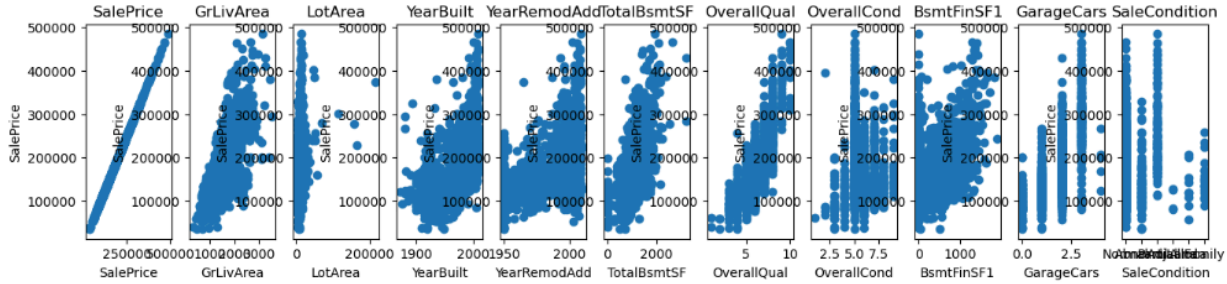


Based on the residual versus fitted values plot, there does not seem to be a violation of normality.



5.3 Linear relationship between predictor and response variables

Based on plotting out the response variable against each predictor, we confirmed that there is presence of a linear relationship.



Based on the findings regarding heteroskedasticity and conflicting findings around normality, we decided to try a log normal transformation on SalePrice. However, the adjusted R^2 fell to 0.913 against Model 3's 0.927. So we will stick to our original model.

Model 4

ln(SalePrice) ~ LotArea + BsmtFinSF1 + TotalBsmtSF + GrLivArea + GarageCars + Neighborhood + C(OverallQual) + OverallCond + YearBuilt + SaleCondition

Additionally, to remove heteroscedasticity we tried to use robust standard errors; and to deal with the non-normality we did a log transformation of the response variable SalePrice. But that did not meaningfully impact the model as the global F-statistic increased marginally from 325.6 to 422.3, while adjusted R^2 remained at 0.913.

OLS Regression Results

Dep. Variable:	log_price	**R-squared:**	0.916
Model:	OLS	**Adj. R-squared:**	0.913
Method:	Least Squares	**F-statistic:**	417.9
Date:	Mon, 09 Oct 2023	**Prob (F-statistic):**	0.00
Time:	14:48:10	**Log-Likelihood:**	1117.7
No. Observations:	1417	**AIC:**	-2141.
Df Residuals:	1370	**BIC:**	-1894.
Df Model:	46		
Covariance Type:	HC0		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.7875	0.638	5.939	0.000	2.536	5.038
SaleCondition[T.AdjLand]	0.1011	0.044	2.293	0.022	0.015	0.188
SaleCondition[T.Alloca]	-0.1151	0.058	-1.979	0.048	-0.229	-0.001
SaleCondition[T.Family]	0.0473	0.038	1.252	0.211	-0.027	0.121
SaleCondition[T.Normal]	0.0829	0.019	4.471	0.000	0.047	0.119
SaleCondition[T.Partial]	0.1295	0.021	6.218	0.000	0.089	0.170
Neighborhood[T.Blueste]	-0.0890	0.048	-1.859	0.063	-0.183	0.005
Neighborhood[T.BrDale]	-0.1603	0.034	-4.708	0.000	-0.227	-0.094
Neighborhood[T.BrkSide]	0.0237	0.037	0.642	0.521	-0.049	0.096
Neighborhood[T.ClearCr]	0.1142	0.038	3.037	0.002	0.040	0.188
Neighborhood[T.CollgCr]	0.0330	0.027	1.228	0.220	-0.020	0.086
Neighborhood[T.Crawfor]	0.1528	0.035	4.380	0.000	0.084	0.221
Neighborhood[T.Edwards]	-0.0338	0.034	-0.982	0.326	-0.101	0.034
Neighborhood[T.Gilbert]	0.0522	0.028	1.870	0.062	-0.003	0.107
Neighborhood[T.IDOTRR]	-0.1009	0.050	-2.033	0.042	-0.198	-0.004
Neighborhood[T.MeadowV]	-0.1821	0.036	-4.999	0.000	-0.254	-0.111
Neighborhood[T.Mitchel]	-0.0183	0.032	-0.576	0.565	-0.081	0.044
Neighborhood[T.NAmes]	0.0150	0.030	0.499	0.618	-0.044	0.074
Neighborhood[T.NPkVill]	-0.0308	0.031	-0.989	0.323	-0.092	0.030
Neighborhood[T.NWAmes]	0.0099	0.030	0.333	0.739	-0.048	0.068
Neighborhood[T.NoRidge]	0.0805	0.032	2.553	0.011	0.019	0.142
Neighborhood[T.NridgHt]	0.0749	0.028	2.645	0.008	0.019	0.130

Neighborhood[T.OldTown]	-0.0664	0.034	-1.948	0.052	-0.133	0.000
Neighborhood[T.SWISU]	0.0295	0.039	0.749	0.454	-0.048	0.107
Neighborhood[T.Sawyer]	-0.0082	0.033	-0.249	0.804	-0.073	0.056
Neighborhood[T.SawyerW]	0.0241	0.030	0.816	0.415	-0.034	0.082
Neighborhood[T.Somerst]	0.0727	0.027	2.666	0.008	0.019	0.126
Neighborhood[T.StoneBr]	0.1058	0.032	3.263	0.001	0.042	0.169
Neighborhood[T.Timber]	0.0307	0.030	1.015	0.310	-0.029	0.090
Neighborhood[T.Veenker]	0.1106	0.042	2.615	0.009	0.028	0.194
C(OverallQual)[T.2]	-0.0459	0.165	-0.278	0.781	-0.369	0.278
C(OverallQual)[T.3]	0.2178	0.144	1.515	0.130	-0.064	0.500
C(OverallQual)[T.4]	0.3099	0.138	2.249	0.025	0.040	0.580
C(OverallQual)[T.5]	0.3663	0.137	2.668	0.008	0.097	0.636
C(OverallQual)[T.6]	0.4094	0.137	2.978	0.003	0.140	0.679
C(OverallQual)[T.7]	0.4848	0.138	3.518	0.000	0.214	0.755
C(OverallQual)[T.8]	0.5679	0.138	4.107	0.000	0.297	0.839
C(OverallQual)[T.9]	0.7065	0.139	5.072	0.000	0.433	0.980
C(OverallQual)[T.10]	0.7284	0.141	5.168	0.000	0.452	1.005
LotArea	2.398e-06	4.51e-07	5.321	0.000	1.51e-06	3.28e-06
GrLivArea	0.0003	9.15e-06	29.771	0.000	0.000	0.000
YearRemodAdd	0.0007	0.000	3.079	0.002	0.000	0.001
YearBuilt	0.0027	0.000	9.236	0.000	0.002	0.003
TotalBsmtSF	0.0001	1.03e-05	11.164	0.000	9.49e-05	0.000
BsmtFinSF1	9.961e-05	8.36e-06	11.910	0.000	8.32e-05	0.000
OverallCond	0.0515	0.004	13.374	0.000	0.044	0.059
GarageCars	0.0602	0.007	8.970	0.000	0.047	0.073

Omnibus:	289.661	**Durbin-Watson:**	1.931
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1356.263
Skew:	-0.885	**Prob(JB):**	3.10e-295
Kurtosis:	7.454	**Cond. No.**	2.71e+06

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
[2] The condition number is large, 2.71e+06. This might indicate that there are strong multicollinearity or other numerical problems.

However, the adjusted R^2 fell to 0.913 against Model 3's 0.927. So we will stick to our original model, which is:

SalePrice ~ LotArea + BsmtFinSF1 + TotalBsmtSF + GrLivArea + GarageCars + Neighborhood + C(OverallQual) + OverallCond + YearBuilt + SaleCondition

Re-running the regression analysis for the final model:

OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.929
Model:	OLS	**Adj. R-squared:**	0.927
Method:	Least Squares	**F-statistic:**	400.2
Date:	Sun, 08 Oct 2023	**Prob (F-statistic):**	0.00
Time:	18:27:24	**Log-Likelihood:**	-15921.
No. Observations:	1417	**AIC:**	3.193e+04
Df Residuals:	1371	**BIC:**	3.218e+04
Df Model:	45		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-8.764e+05	7.98e+04	-10.975	0.000	-1.03e+06	-7.2e+05
Neighborhood[T.Blueste]	-1.032e+04	1.41e+04	-0.733	0.464	-3.79e+04	1.73e+04
Neighborhood[T.BrDale]	-1.179e+04	6845.623	-1.722	0.085	-2.52e+04	1641.839
Neighborhood[T.BrkSide]	9563.2683	5883.157	1.626	0.104	-1977.695	2.11e+04
Neighborhood[T.ClearCr]	1.691e+04	6280.718	2.693	0.007	4592.448	2.92e+04
Neighborhood[T.CollgCr]	6794.3546	4865.265	1.397	0.163	-2749.815	1.63e+04
Neighborhood[T.Crawfor]	2.739e+04	5909.548	4.634	0.000	1.58e+04	3.9e+04
Neighborhood[T.Edwards]	-2372.9115	5417.517	-0.438	0.661	-1.3e+04	8254.609
Neighborhood[T.Gilbert]	8793.5877	5151.314	1.707	0.088	-1311.723	1.89e+04
Neighborhood[T.IDOTRR]	-163.8084	6246.392	-0.026	0.979	-1.24e+04	1.21e+04
Neighborhood[T.MeadowV]	-1.56e+04	6809.452	-2.290	0.022	-2.9e+04	-2237.334
Neighborhood[T.Mitchel]	-3003.7523	5510.866	-0.545	0.586	-1.38e+04	7806.891
Neighborhood[T.NAmes]	1039.7920	5132.183	0.203	0.839	-9027.990	1.11e+04
Neighborhood[T.NPkVill]	-3420.4377	7921.426	-0.432	0.666	-1.9e+04	1.21e+04
Neighborhood[T.NWAmes]	575.3448	5296.788	0.109	0.914	-9815.343	1.1e+04
Neighborhood[T.NoRidge]	3.881e+04	5858.494	6.624	0.000	2.73e+04	5.03e+04
Neighborhood[T.NridgHt]	2.281e+04	5316.322	4.290	0.000	1.24e+04	3.32e+04
Neighborhood[T.OldTown]	-5916.4976	5738.693	-1.031	0.303	-1.72e+04	5341.073
Neighborhood[T.SWISU]	984.4784	6669.395	0.148	0.883	-1.21e+04	1.41e+04
Neighborhood[T.Sawyer]	-1582.9088	5442.608	-0.291	0.771	-1.23e+04	9093.832
Neighborhood[T.SawyerW]	7529.4353	5332.698	1.412	0.158	-2931.696	1.8e+04
Neighborhood[T.Somerst]	1.311e+04	5031.463	2.605	0.009	3237.748	2.3e+04
Neighborhood[T.StoneBr]	2.615e+04	6419.810	4.073	0.000	1.36e+04	3.87e+04
Neighborhood[T.Timber]	3947.9308	5731.488	0.689	0.491	-7295.505	1.52e+04
Neighborhood[T.Veenker]	2.452e+04	7424.033	3.303	0.001	9956.148	3.91e+04

C(OverallQual)[T.2]	-1.184e+04	1.72e+04	-0.689	0.491	-4.56e+04	2.19e+04
C(OverallQual)[T.3]	-1.532e+04	1.4e+04	-1.093	0.275	-4.28e+04	1.22e+04
C(OverallQual)[T.4]	-1.5e+04	1.36e+04	-1.106	0.269	-4.16e+04	1.16e+04
C(OverallQual)[T.5]	-1.437e+04	1.36e+04	-1.058	0.290	-4.1e+04	1.23e+04
C(OverallQual)[T.6]	-9662.2092	1.37e+04	-0.708	0.479	-3.64e+04	1.71e+04
C(OverallQual)[T.7]	5853.0826	1.38e+04	0.426	0.671	-2.11e+04	3.28e+04
C(OverallQual)[T.8]	3.364e+04	1.39e+04	2.414	0.016	6298.027	6.1e+04
C(OverallQual)[T.9]	9.228e+04	1.44e+04	6.426	0.000	6.41e+04	1.2e+05
C(OverallQual)[T.10]	1.166e+05	1.52e+04	7.670	0.000	8.68e+04	1.46e+05
SaleCondition[T.AdjLand]	2.131e+04	9746.181	2.186	0.029	2187.000	4.04e+04
SaleCondition[T.Alloca]	-1.821e+04	6674.683	-2.728	0.006	-3.13e+04	-5111.898
SaleCondition[T.Family]	2139.7636	4720.171	0.453	0.650	-7119.777	1.14e+04
SaleCondition[T.Normal]	1.006e+04	1991.866	5.052	0.000	6154.944	1.4e+04
SaleCondition[T.Partial]	2.601e+04	2882.933	9.020	0.000	2.03e+04	3.17e+04
LotArea	0.6016	0.058	10.385	0.000	0.488	0.715
BsmtFinSF1	18.5399	1.429	12.976	0.000	15.737	21.343
TotalBsmtSF	19.4673	1.733	11.233	0.000	16.068	22.867
GrLivArea	48.8629	1.466	33.322	0.000	45.986	51.740
GarageCars	8290.6296	951.581	8.712	0.000	6423.917	1.02e+04
OverallCond	8477.5550	527.385	16.075	0.000	7442.987	9512.123
YearBuilt	440.1416	39.577	11.121	0.000	362.504	517.779

Omnibus:	34.610	**Durbin-Watson:**	1.966
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	74.029
Skew:	0.066	**Prob(JB):**	8.41e-17
Kurtosis:	4.112	**Cond. No.**	2.34e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.34e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

6.0 Model Selection

6.1 Order of predictors

We ran the ANOVA Typ=1 test on the existing order and reordered the predictors according to F-statistic value.

	df	sum_sq	mean_sq	F	PR(>F)
Neighborhood	24.0	3.948332e+12	1.645138e+11	473.426979	0.000000e+00
C(OverallQual)	9.0	1.401620e+12	1.557356e+11	448.165508	0.000000e+00
SaleCondition	5.0	4.847727e+10	9.695454e+09	27.900930	5.496072e-27
LotArea	1.0	1.237309e+11	1.237309e+11	356.064643	8.692749e-71
BsmtFinSF1	1.0	1.164268e+11	1.164268e+11	335.045321	3.936810e-67
TotalBsmtSF	1.0	8.467559e+10	8.467559e+10	243.673777	1.094633e-50
GrLivArea	1.0	3.854240e+11	3.854240e+11	1109.147404	1.077829e-178
GarageCars	1.0	3.493778e+10	3.493778e+10	100.541626	6.965227e-23
OverallCond	1.0	7.126671e+10	7.126671e+10	205.086574	1.871164e-43
YearBuilt	1.0	4.297845e+10	4.297845e+10	123.680503	1.436466e-27
Residual	1371.0	4.764166e+11	3.474957e+08	NaN	NaN

Trial and error yielded the following as the optimal order where SaleCondition is moved up.

	df	sum_sq	mean_sq	F	PR(>F)
SaleCondition	5.0	8.628379e+11	1.725676e+11	496.603447	1.383730e-304
Neighborhood	24.0	3.191238e+12	1.329683e+11	382.647191	0.000000e+00
C(OverallQual)	9.0	1.344354e+12	1.493727e+11	429.854654	0.000000e+00
LotArea	1.0	1.237309e+11	1.237309e+11	356.064643	8.692749e-71
BsmtFinSF1	1.0	1.164268e+11	1.164268e+11	335.045321	3.936810e-67
TotalBsmtSF	1.0	8.467559e+10	8.467559e+10	243.673777	1.094633e-50
GarageCars	1.0	9.011864e+10	9.011864e+10	259.337405	1.425102e-53
GrLivArea	1.0	3.302431e+11	3.302431e+11	950.351625	5.630969e-159
OverallCond	1.0	7.126671e+10	7.126671e+10	205.086574	1.871164e-43
YearBuilt	1.0	4.297845e+10	4.297845e+10	123.680503	1.436466e-27
Residual	1371.0	4.764166e+11	3.474957e+08	NaN	NaN

Model 5

SalePrice ~ LotArea + BsmtFinSF1 + SaleCondition + TotalBsmtSF + GarageCars + GrLivArea + Neighborhood + C(OverallQual) + OverallCond + YearBuilt

6.2 Optimal predictors from pool of all possible models

We then calculated Mallow's Cp, BIC, AIC, Adjusted R^2, and R^2 for all possible models based on the predictors identified. We then sorted the possible models based on the lowest Mallow's Cp,

BIC, AIC, and highest Adjusted R^2 to yield our final model where the adjusted R^2 is 0.927.

	model	Predictors	Cp	BIC	AIC	adj_R^2	R^2
1022	<statsmodels.regression.linear_model.Regressio...	LotArea + BsmtFinSF1 + SaleCondition + TotalBs...	46.000000	32175.392796	31933.603123	0.926933	0.929255
1017	<statsmodels.regression.linear_model.Regressio...	LotArea + BsmtFinSF1 + SaleCondition + TotalBs...	119.907242	32244.495712	32007.962336	0.922944	0.925338
1019	<statsmodels.regression.linear_model.Regressio...	LotArea + BsmtFinSF1 + TotalBsmtSF + GarageCar...	142.067348	32244.703599	32029.195412	0.921566	0.923782
1015	<statsmodels.regression.linear_model.Regressio...	LotArea + BsmtFinSF1 + SaleCondition + TotalBs...	286.765814	32272.081539	32156.443000	0.913065	0.914354
1021	<statsmodels.regression.linear_model.Regressio...	BsmtFinSF1 + SaleCondition + TotalBsmtSF + Gar...	151.845496	32275.433599	32038.900223	0.921243	0.923690
...	...	...	...	...	...	...	...
2	<statsmodels.regression.linear_model.Regressio...	SaleCondition	15491.464326	35444.027077	35412.489294	0.125037	0.128126
32	<statsmodels.regression.linear_model.Regressio...	SaleCondition + OverallCond	15453.287311	35447.909969	35411.115889	0.126498	0.130199
17	<statsmodels.regression.linear_model.Regressio...	LotArea + OverallCond	16571.924643	35510.563200	35494.794308	0.070751	0.072064
0	<statsmodels.regression.linear_model.Regressio...	LotArea	16696.327944	35513.232258	35502.719663	0.064881	0.065541
8	<statsmodels.regression.linear_model.Regressio...	OverallCond	17834.894673	35599.633161	35589.120566	0.006088	0.006790

Final Model (Same as Model 5)

SalePrice ~ LotArea + BsmtFinSF1 + SaleCondition + TotalBsmtSF + GarageCars + GrLivArea + Neighborhood + C(OverallQual) + OverallCond + YearBuilt

OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.929
Model:	OLS	**Adj. R-squared:**	0.927
Method:	Least Squares	**F-statistic:**	400.2
Date:	Sun, 08 Oct 2023	**Prob (F-statistic):**	0.00
Time:	18:59:10	**Log-Likelihood:**	-15921.
No. Observations:	1417	**AIC:**	3.193e+04
Df Residuals:	1371	**BIC:**	3.218e+04
Df Model:	45		
Covariance Type:	nonrobust		

Omnibus:	34.610	**Durbin-Watson:**	1.966
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	74.029
Skew:	0.066	**Prob(JB):**	8.41e-17
Kurtosis:	4.112	**Cond. No.**	2.34e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.34e+06. This might indicate that there are strong multicollinearity or other numerical problems.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-8.764e+05	7.98e+04	-10.975	0.000	-1.03e+06	-7.2e+05
SaleCondition[T.AdjLand]	2.131e+04	9746.181	2.186	0.029	2187.000	4.04e+04
SaleCondition[T.Alloca]	-1.821e+04	6674.683	-2.728	0.006	-3.13e+04	-5111.898
SaleCondition[T.Family]	2139.7636	4720.171	0.453	0.650	-7119.777	1.14e+04
SaleCondition[T.Normal]	1.006e+04	1991.866	5.052	0.000	6154.944	1.4e+04
SaleCondition[T.Partial]	2.601e+04	2882.933	9.020	0.000	2.03e+04	3.17e+04
Neighborhood[T.Blueste]	-1.032e+04	1.41e+04	-0.733	0.464	-3.79e+04	1.73e+04
Neighborhood[T.BrDale]	-1.179e+04	6845.623	-1.722	0.085	-2.52e+04	1641.839
Neighborhood[T.BrkSide]	9563.2683	5883.157	1.626	0.104	-1977.695	2.11e+04
Neighborhood[T.ClearCr]	1.691e+04	6280.718	2.693	0.007	4592.448	2.92e+04
Neighborhood[T.CollgCr]	6794.3546	4865.265	1.397	0.163	-2749.815	1.63e+04
Neighborhood[T.Crawfor]	2.739e+04	5909.548	4.634	0.000	1.58e+04	3.9e+04
Neighborhood[T.Edwards]	-2372.9115	5417.517	-0.438	0.661	-1.3e+04	8254.609
Neighborhood[T.Gilbert]	8793.5877	5151.314	1.707	0.088	-1311.723	1.89e+04
Neighborhood[T.IDOTRR]	-163.8084	6246.392	-0.026	0.979	-1.24e+04	1.21e+04
Neighborhood[T.MeadowV]	-1.56e+04	6809.452	-2.290	0.022	-2.9e+04	-2237.334
Neighborhood[T.Mitchel]	-3003.7523	5510.866	-0.545	0.586	-1.38e+04	7806.891
Neighborhood[T.NAmes]	1039.7920	5132.183	0.203	0.839	-9027.990	1.11e+04
Neighborhood[T.NPkVill]	-3420.4377	7921.426	-0.432	0.666	-1.9e+04	1.21e+04
Neighborhood[T.NWAmes]	575.3448	5296.788	0.109	0.914	-9815.343	1.1e+04
Neighborhood[T.NoRidge]	3.881e+04	5858.494	6.624	0.000	2.73e+04	5.03e+04
Neighborhood[T.NridgHt]	2.281e+04	5316.322	4.290	0.000	1.24e+04	3.32e+04
Neighborhood[T.OldTown]	-5916.4976	5738.693	-1.031	0.303	-1.72e+04	5341.073
Neighborhood[T.SWISU]	984.4784	6669.395	0.148	0.883	-1.21e+04	1.41e+04
Neighborhood[T.Sawyer]	-1582.9088	5442.608	-0.291	0.771	-1.23e+04	9093.832
Neighborhood[T.SawyerW]	7529.4353	5332.698	1.412	0.158	-2931.696	1.8e+04
Neighborhood[T.Somerst]	1.311e+04	5031.463	2.605	0.009	3237.748	2.3e+04
Neighborhood[T.StoneBr]	2.615e+04	6419.810	4.073	0.000	1.36e+04	3.87e+04
Neighborhood[T.Timber]	3947.9308	5731.488	0.689	0.491	-7295.505	1.52e+04
Neighborhood[T.Veenker]	2.452e+04	7424.033	3.303	0.001	9956.148	3.91e+04
C(OverallQual)[T.2]	-1.184e+04	1.72e+04	-0.689	0.491	-4.56e+04	2.19e+04
C(OverallQual)[T.3]	-1.532e+04	1.4e+04	-1.093	0.275	-4.28e+04	1.22e+04
C(OverallQual)[T.4]	-1.5e+04	1.36e+04	-1.106	0.269	-4.16e+04	1.16e+04
C(OverallQual)[T.5]	-1.437e+04	1.36e+04	-1.058	0.290	-4.1e+04	1.23e+04
C(OverallQual)[T.6]	-9662.2092	1.37e+04	-0.708	0.479	-3.64e+04	1.71e+04
C(OverallQual)[T.7]	5853.0826	1.38e+04	0.426	0.671	-2.11e+04	3.28e+04
C(OverallQual)[T.8]	3.364e+04	1.39e+04	2.414	0.016	6298.027	6.1e+04
C(OverallQual)[T.9]	9.228e+04	1.44e+04	6.426	0.000	6.41e+04	1.2e+05
C(OverallQual)[T.10]	1.166e+05	1.52e+04	7.670	0.000	8.68e+04	1.46e+05
LotArea	0.6016	0.058	10.385	0.000	0.488	0.715
BsmtFinSF1	18.5399	1.429	12.976	0.000	15.737	21.343
TotalBsmtSF	19.4673	1.733	11.233	0.000	16.068	22.867
GarageCars	8290.6296	951.581	8.712	0.000	6423.917	1.02e+04
GrLivArea	48.8629	1.466	33.322	0.000	45.986	51.740
OverallCond	8477.5550	527.385	16.075	0.000	7442.987	9512.123
YearBuilt	440.1416	39.577	11.121	0.000	362.504	517.779

7.0 Summary of results

Our final model is:

SalePrice ~ LotArea + BsmtFinSF1 + SaleCondition + TotalBsmtSF + GarageCars + GrLivArea + Neighborhood + C(OverallQual) + OverallCond + YearBuilt

This is a combination of predictors that comprehensively cover key dimensions across size, location, condition/amenity, and transaction to predict the sale price of real estate.

Additionally, we are confident that this is the best model because Adjusted R^2 increased from 0.853 to 0.927. Additionally, we

1. Removed predictors with t-test p-values for all levels < 0.05
2. Conducted EDA to confirm that there is no high correlation between numerical predictors and removed predictors with a high proportion of NULL values
3. Conducted data structure validation to remove predictors with high multicollinearity (VIF, autocorrelation plot), and removed outlier data points (External studentized residuals, Cook's Distance)
4. Conducted model assumption checks around heteroskedasticity (Breush-Pagan Test) , normality (KS Test, JB Test, Histogram and scatter plot of residuals, QQ plot) , and linear relationship between chosen predictors and the response variable (scatter plot of predictors versus response variable)
5. Selected the model based on the optimal order of predictors (ANOVA Type=1) and also from the pool of all possible combinations of predictors (Adjusted R^2, Mallow's Cp, AIC, BIC)

8.0 Potential problems of the data and results

We note that there continues to be high multicollinearity (VIF) for some neighborhoods and transformations did not manage to address the heteroskedasticity identified. However, we chose to keep Neighborhood in the predictor set as the p-value of other levels were sufficiently low and the overall adjusted R^2 is good at 0.927. Also, we may consider changing 'OverQuality' to numeric, as OverallQual is not a reliable predictor when it is categorical.

9.0 Extending our Discussion

We determined the predicted values for the test data set from the same population and calculated the MSE as 18,336.166. This was deemed to be high. Therefore, we decided to try logistic regression to improve the accuracy of our predictions. To do so, we created binary response variables and ran the regression analysis with the same set of predictors.

9.1 Sale price greater than or equal to mean sale price

For the binary response variable: sale price greater than or equal to mean sale price, despite an accuracy rate of 94%, the Pseudo R^2 was low at 0.6624. Therefore we did not believe that this model is superior to the linear regression model.

Further study may include trying other models like PCA, Partial LSE, or other Regularization methods to reduce the Multicollinearity in Neighborhood. Additionally, we also tried to introduce Dummy Variables, and removed the Neighborhood levels that had high p-values. However, the effect was limited with an adjusted R2 of 0.921, a small decrease of 0.927. Based on the fact that the removal of specific Neighborhoods makes interpretation challenging due to the optics of cherry picking, we left all Neighborhoods in.

Generalized Linear Model Regression Results

Dep. Variable:	Sale_Price_binary_mean	**No. Observations:**	1417
Model:	GLM	**Df Residuals:**	1366
Model Family:	Binomial	**Df Model:**	50
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-191.26
Date:	Sun, 08 Oct 2023	**Deviance:**	382.52
Time:	18:50:01	**Pearson chi2:**	610.
No. Iterations:	24	**Pseudo R-squ. (CS):**	0.6624
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-281.4654	1.59e+05	-0.002	0.999	-3.12e+05	3.12e+05
MSZoning[T.FV]	5.6340	6.29e+04	8.95e-05	1.000	-1.23e+05	1.23e+05
MSZoning[T.RH]	24.7926	4.83e+04	0.001	1.000	-9.45e+04	9.46e+04
MSZoning[T.RL]	24.9513	4.83e+04	0.001	1.000	-9.45e+04	9.46e+04
MSZoning[T.RM]	22.4574	4.83e+04	0.000	1.000	-9.46e+04	9.46e+04
SaleCondition[T.AdjLand]	-15.7969	8.4e+04	-0.000	1.000	-1.65e+05	1.65e+05
SaleCondition[T.Alloca]	3.8852	1.916	2.028	0.043	0.131	7.640
SaleCondition[T.Family]	-0.3272	1.346	-0.243	0.808	-2.966	2.311
SaleCondition[T.Normal]	2.2762	0.689	3.305	0.001	0.926	3.626
SaleCondition[T.Partial]	1.5371	0.819	1.878	0.060	-0.067	3.141
Neighborhood[T.Blueste]	-19.2510	1.46e+05	-0.000	1.000	-2.87e+05	2.87e+05
Neighborhood[T.BrDale]	-15.7103	4.61e+04	-0.000	1.000	-9.04e+04	9.03e+04
Neighborhood[T.BrkSide]	4.4929	1.440	3.121	0.002	1.671	7.315
Neighborhood[T.ClearCr]	2.2694	1.282	1.770	0.077	-0.243	4.782
Neighborhood[T.CollgCr]	0.9817	0.859	1.143	0.253	-0.702	2.665
Neighborhood[T.Crawfor]	4.5907	1.317	3.487	0.000	2.010	7.171
Neighborhood[T.Edwards]	0.3986	0.997	0.400	0.689	-1.556	2.353
Neighborhood[T.Gilbert]	0.7759	0.839	0.925	0.355	-0.868	2.419

Neighborhood[T.IDOTRR]	-14.1343	2.83e+04	-0.000	1.000	-5.55e+04	5.55e+04
Neighborhood[T.MeadowV]	-22.0481	3.71e+04	-0.001	1.000	-7.28e+04	7.28e+04
Neighborhood[T.Mitchel]	-2.6431	1.187	-2.227	0.026	-4.969	-0.317
Neighborhood[T.NAmes]	-0.5339	0.988	-0.540	0.589	-2.470	1.403
Neighborhood[T.NPkVill]	-21.9283	6.7e+04	-0.000	1.000	-1.31e+05	1.31e+05
Neighborhood[T.NWAmes]	0.2749	0.931	0.295	0.768	-1.551	2.101
Neighborhood[T.NoRidge]	17.8789	2.85e+04	0.001	0.999	-5.58e+04	5.58e+04
Neighborhood[T.NridgHt]	0.0941	0.983	0.096	0.924	-1.832	2.021
Neighborhood[T.OldTown]	1.1723	1.825	0.642	0.521	-2.405	4.750
Neighborhood[T.SWISU]	1.2599	1.584	0.795	0.426	-1.844	4.364
Neighborhood[T.Sawyer]	-2.5848	1.472	-1.756	0.079	-5.469	0.300
Neighborhood[T.SawyerW]	0.4062	0.941	0.432	0.666	-1.437	2.250
Neighborhood[T.Somerst]	20.4734	4.04e+04	0.001	1.000	-7.92e+04	7.92e+04
Neighborhood[T.StoneBr]	1.4869	2.060	0.722	0.470	-2.551	5.525
Neighborhood[T.Timber]	-0.4008	1.162	-0.345	0.730	-2.678	1.876
Neighborhood[T.Veenker]	0.0462	1.306	0.035	0.972	-2.514	2.606
C(OverallQual)[T.2]	-0.3298	1.87e+05	-1.77e-06	1.000	-3.66e+05	3.66e+05
C(OverallQual)[T.3]	-9.9283	1.55e+05	-6.39e-05	1.000	-3.05e+05	3.05e+05
C(OverallQual)[T.4]	6.3981	1.52e+05	4.21e-05	1.000	-2.98e+05	2.98e+05
C(OverallQual)[T.5]	6.2177	1.52e+05	4.1e-05	1.000	-2.98e+05	2.98e+05
C(OverallQual)[T.6]	8.1766	1.52e+05	5.39e-05	1.000	-2.98e+05	2.98e+05
C(OverallQual)[T.7]	9.2428	1.52e+05	6.09e-05	1.000	-2.98e+05	2.98e+05
C(OverallQual)[T.8]	10.2726	1.52e+05	6.77e-05	1.000	-2.98e+05	2.98e+05
C(OverallQual)[T.9]	26.5767	1.55e+05	0.000	1.000	-3.03e+05	3.03e+05
C(OverallQual)[T.10]	26.8758	1.6e+05	0.000	1.000	-3.13e+05	3.13e+05
LotArea	0.0002	3.57e-05	4.756	0.000	0.0001	0.000
GrLivArea	0.0062	0.001	9.607	0.000	0.005	0.007
YearRemodAdd	0.0331	0.013	2.561	0.010	0.008	0.059
YearBuilt	0.0789	0.016	4.993	0.000	0.048	0.110
TotalBsmtSF	0.0035	0.001	6.354	0.000	0.002	0.005
BsmtFinSF1	0.0022	0.000	4.652	0.000	0.001	0.003
OverallCond	1.1559	0.221	5.236	0.000	0.723	1.589
GarageCars	0.7658	0.393	1.947	0.052	-0.005	1.537

9.2 Sale price greater than or equal to median sale price

For the binary response variable: sale price greater than or equal to median sale price, despite an accuracy rate of 93%, the Pseudo R^2 was low at 0.6607. Therefore we did not believe that this model is superior to the linear regression model.

Generalized Linear Model Regression Results

Dep. Variable:	Sale_Price_binary_median	**No. Observations:**	1417
Model:	GLM	**Df Residuals:**	1366
Model Family:	Binomial	**Df Model:**	50
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-216.37
Date:	Sun, 08 Oct 2023	**Deviance:**	432.75
Time:	18:50:08	**Pearson chi2:**	690.
No. Iterations:	25	**Pseudo R-squ. (CS):**	0.6607
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-192.6869	2.63e+05	-0.001	0.999	-5.16e+05	5.16e+05
MSZoning[T.FV]	8.6822	9.8e+04	8.86e-05	1.000	-1.92e+05	1.92e+05
MSZoning[T.RH]	26.1759	8.04e+04	0.000	1.000	-1.58e+05	1.58e+05
MSZoning[T.RL]	28.0970	8.04e+04	0.000	1.000	-1.58e+05	1.58e+05
MSZoning[T.RM]	25.7544	8.04e+04	0.000	1.000	-1.58e+05	1.58e+05
SaleCondition[T.AdjLand]	-17.6599	1.51e+05	-0.000	1.000	-2.95e+05	2.95e+05
SaleCondition[T.Alloca]	-0.2066	2.499	-0.083	0.934	-5.104	4.691
SaleCondition[T.Family]	-0.3820	1.108	-0.345	0.730	-2.553	1.789
SaleCondition[T.Normal]	1.9753	0.616	3.209	0.001	0.769	3.182
SaleCondition[T.Partial]	2.2226	0.907	2.451	0.014	0.445	4.000
Neighborhood[T.Blueste]	-24.2762	2.44e+05	-9.95e-05	1.000	-4.78e+05	4.78e+05
Neighborhood[T.BrDale]	-20.8453	7.74e+04	-0.000	1.000	-1.52e+05	1.52e+05
Neighborhood[T.BrkSide]	1.7797	1.540	1.155	0.248	-1.239	4.799
Neighborhood[T.ClearCr]	2.2320	1.732	1.289	0.197	-1.162	5.626
Neighborhood[T.CollgCr]	-0.1360	1.265	-0.108	0.914	-2.615	2.343
Neighborhood[T.Crawfor]	2.5711	1.484	1.733	0.083	-0.337	5.480
Neighborhood[T.Edwards]	-1.3190	1.299	-1.016	0.310	-3.864	1.226
Neighborhood[T.Gilbert]	1.4415	1.394	1.034	0.301	-1.291	4.174
Neighborhood[T.IDOTRR]	3.5728	1.878	1.902	0.057	-0.108	7.254
Neighborhood[T.MeadowV]	-26.2770	6.34e+04	-0.000	1.000	-1.24e+05	1.24e+05
Neighborhood[T.Mitchel]	-0.2104	1.315	-0.160	0.873	-2.787	2.366
Neighborhood[T.NAmes]	-0.8671	1.290	-0.672	0.501	-3.395	1.660
Neighborhood[T.NPkVill]	-26.5572	1.11e+05	-0.000	1.000	-2.18e+05	2.18e+05
Neighborhood[T.NWAmes]	0.0842	1.301	0.065	0.948	-2.465	2.634
Neighborhood[T.NoRidge]	18.0948	4.3e+04	0.000	1.000	-8.42e+04	8.42e+04
Neighborhood[T.NridgHt]	21.4211	3.21e+04	0.001	0.999	-6.28e+04	6.29e+04
Neighborhood[T.OldTown]	0.3118	1.560	0.200	0.842	-2.746	3.370
Neighborhood[T.SWISU]	-0.1379	1.596	-0.086	0.931	-3.266	2.990
Neighborhood[T.Sawyer]	-0.6607	1.351	-0.489	0.625	-3.308	1.986
Neighborhood[T.SawyerW]	-0.9826	1.316	-0.747	0.455	-3.561	1.596

Neighborhood[T.Somerst]	19.2747	5.6e+04	0.000	1.000	-1.1e+05	1.1e+05
Neighborhood[T.StoneBr]	19.8606	6.1e+04	0.000	1.000	-1.2e+05	1.2e+05
Neighborhood[T.Timber]	1.6469	1.695	0.972	0.331	-1.675	4.968
Neighborhood[T.Veenker]	23.5465	8.96e+04	0.000	1.000	-1.76e+05	1.76e+05
C(OverallQual)[T.2]	-0.4257	3.04e+05	-1.4e-06	1.000	-5.96e+05	5.96e+05
C(OverallQual)[T.3]	-7.4165	2.58e+05	-2.87e-05	1.000	-5.06e+05	5.06e+05
C(OverallQual)[T.4]	11.7107	2.51e+05	4.67e-05	1.000	-4.92e+05	4.92e+05
C(OverallQual)[T.5]	12.3063	2.51e+05	4.91e-05	1.000	-4.92e+05	4.92e+05
C(OverallQual)[T.6]	13.7262	2.51e+05	5.47e-05	1.000	-4.92e+05	4.92e+05
C(OverallQual)[T.7]	14.4823	2.51e+05	5.77e-05	1.000	-4.92e+05	4.92e+05
C(OverallQual)[T.8]	18.6546	2.51e+05	7.44e-05	1.000	-4.92e+05	4.92e+05
C(OverallQual)[T.9]	32.1227	2.54e+05	0.000	1.000	-4.98e+05	4.98e+05
C(OverallQual)[T.10]	32.7293	2.64e+05	0.000	1.000	-5.18e+05	5.18e+05
LotArea	3.749e-05	3.43e-05	1.092	0.275	-2.98e-05	0.000
GrLivArea	0.0052	0.001	9.878	0.000	0.004	0.006
YearRemodAdd	0.0184	0.009	1.939	0.053	-0.000	0.037
YearBuilt	0.0486	0.012	3.995	0.000	0.025	0.072
TotalBsmtSF	0.0022	0.001	4.304	0.000	0.001	0.003
BsmtFinSF1	0.0016	0.000	3.746	0.000	0.001	0.002
OverallCond	0.7196	0.165	4.371	0.000	0.397	1.042
GarageCars	1.3699	0.306	4.479	0.000	0.770	1.969

Overall, we declined to continue the analysis as it was deemed that the findings did not directly address the original problem statement of predicting the sale price of real estate based on size, location, condition, and transaction dimensions. Moreover, we observed that logistic regression does not meaningfully increase accuracy linear regression, while MSE remains high for linear regression. For future studies, we may consider making SalesPrice a categorical variable and bucket our predictions into ranges instead of trying to arrive at a fixed value; this can greatly improve the model prediction results and can be better suited for any future analyses