Project statement:

While searching for the dream house, the buyer looks at various factors, not just the basement ceiling height or proximity to an east-west railroad. Using the dataset, find the factors influencing price negotiations while buying the house.

There are 79 explanatory variables describing every aspect of the residential homes in Ames, Lowa.

To understand the dataset, we find the shape of the dataset, i.e., (1460, 81).
Missing numerical feature columns with missing values are;

LotFrontage: 259
MasVnrArea: 8
GarageYrBlt : 81

When we calculate the mean missing value for the numerical features, it is less than 30%. So, it can be neglected.

Missing Categorical feature columns with missing values are;

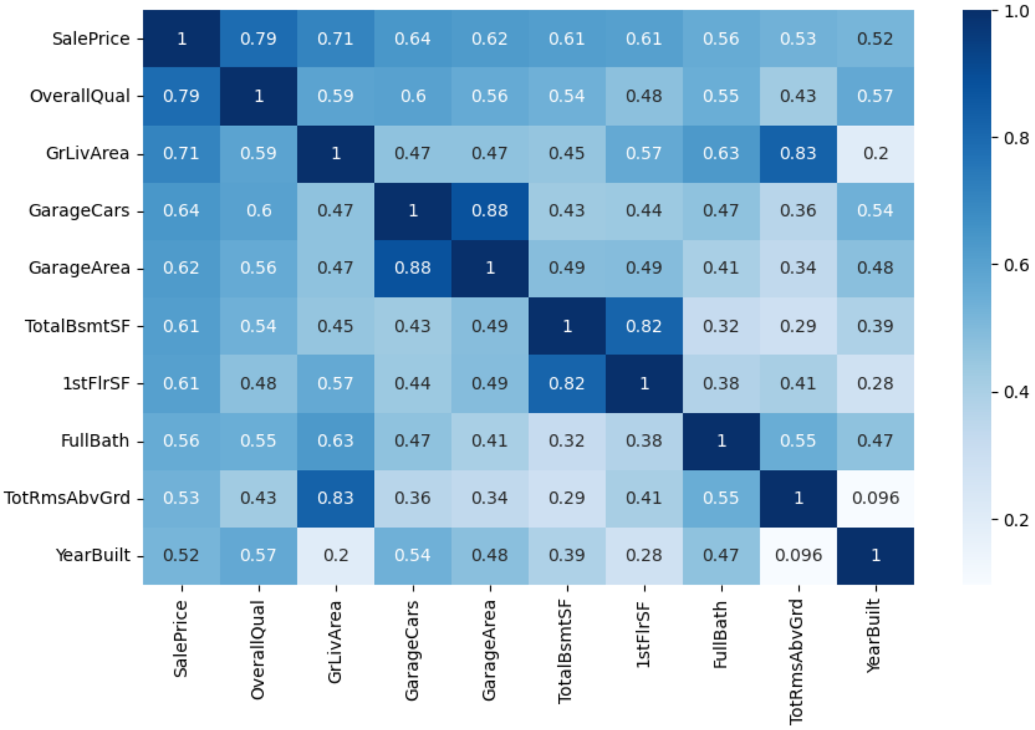
BsmtCond: 37
FireplaceQu : 690
BsmtFinType1 : 37
MasVnrType : 8
GarageQual : 81
PoolQC: 1453
BsmtExposure : 38
GarageCond : 81
GarageType : 81
Alley : 1369
GarageFinish : 81
Fence : 1179
Electrical : 1
BsmtFinType2 : 38
BsmtQual: 37
MiscFeature : 1406

Here, we treat the missing categorical columns by filling them with the missing ones.
The Unique value of the columns is calculated.
Column: ID, with Total unique values, is 1460.
Column: MSSubClass, Total unique values are 15.
And the unique values for the other columns can be seen in the attached file.

For the skewness and the determination, It is observed that there was a decline in the graph representing the building class (MSSubClass).

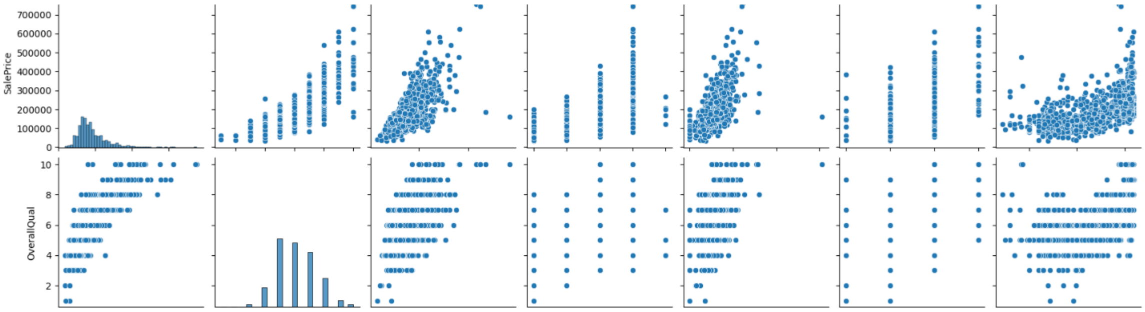
Linear feet of street connected to the property (lot frontage), exhibit a normal distribution between 50-150, Lot size in square feet (lot area) rises above 0 to 50,000 ,
The overall material and finish quality (overall quality), is lower at the start, then medium, and again it decreases, as can be seen in the graph, and all the other variables distribution can also be understood from the visualization.

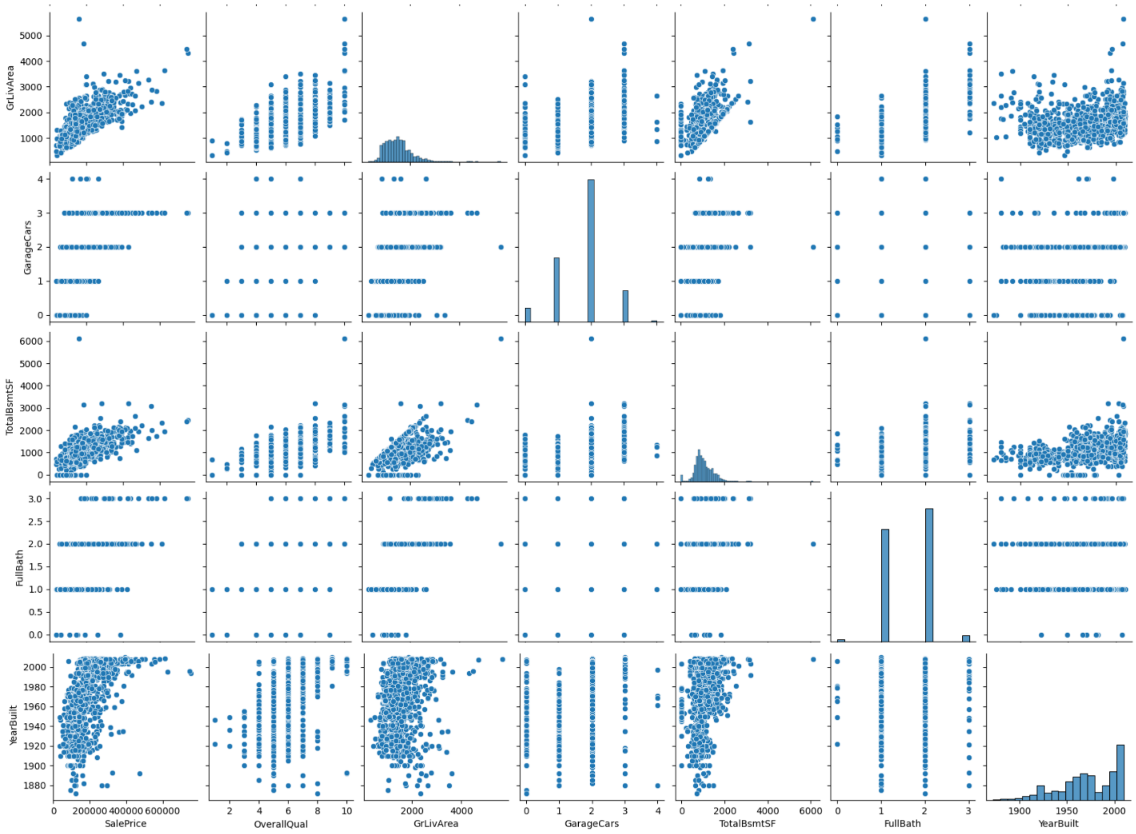
From the correlation matrix, we determine the significant variables by understanding their correlation. we can see the GrLivArea and TotRmsAbvGrd, 1STFLRSF and TotalBsmtSF ,TotRmsAbvGrd and GrLivArea ,are strongly correlated and TotRmsAbvGrd and yearbuilt and
GrLivArea and yearbuilt don’t have a good correlation.



	SalePrice	OverallQual	GrLivArea	GarageCars	GarageArea	TotalBsmtSF	1stFlrSF	FullBath	TotRmsAbvGrd	YearBuilt
SalePrice	1	0.79	0.71	0.64	0.62	0.61	0.61	0.56	0.53	0.52
OverallQual	0.79	1	0.59	0.6	0.56	0.54	0.48	0.55	0.43	0.57
GrLivArea	0.71	0.59	1	0.47	0.47	0.45	0.57	0.63	0.83	0.2
GarageCars	0.64	0.6	0.47	1	0.88	0.43	0.44	0.47	0.36	0.54
GarageArea	0.62	0.56	0.47	0.88	1	0.49	0.49	0.41	0.34	0.48
TotalBsmtSF	0.61	0.54	0.45	0.43	0.49	1	0.82	0.32	0.29	0.39
1stFlrSF	0.61	0.48	0.57	0.44	0.49	0.82	1	0.38	0.41	0.28
FullBath	0.56	0.55	0.63	0.47	0.41	0.32	0.38	1	0.55	0.47
TotRmsAbvGrd	0.53	0.43	0.83	0.36	0.34	0.29	0.41	0.55	1	0.096
YearBuilt	0.52	0.57	0.2	0.54	0.48	0.39	0.28	0.47	0.096	1

A pair plot is a data visualization that plots pair-wise relationships between all the variables in a dataset. From the pair-plot we can interpret ,
Year built and sale price, total bsmtsf and sale price, garage cars and sale price, and full bath and sale price have moderate correlations.
sale price and GrlivArea have a strong correlation.
Overall quality and sale prices have a strong correlation.
Overall quality and year built, have poor correlations.





Explanatory DATA ANALYSIS of the categorical variable is done using the count plot. Significant variables were identified using p-values and Chi-Square values, and the significant categorical variables were converted to numerical using one-hot encoding.And the box plot analysis of the categorical variable is done. We can see the outliers in the SaleType_con, Neighborhood_mes, BsmtQual_Gd, LotShape_IR2, LotConfig_FR3, KitchenQual_Ex, MasVnrType_Missing, and most of the variables .
Hence, the new data can be used for further training process.