AMES HOUSING DATA

ANALYSIS REPORT

Prepared by: SOURAV SYAL

Introduction

This report presents the findings of the data science project focused on analyzing housing sale price data. Through a combination of exploratory data analysis, feature engineering, and predictive modeling techniques, this project aims to understand the key factors influencing housing prices and develop a model that can accurately predict future sale prices. The analysis incorporates a variety of visualization techniques to effectively communicate insights and patterns within the data, providing a comprehensive understanding of the housing market dynamics.

Data set

Here's a glimpse of the data set describing the sale of individual residential property in Ames, Iowa from 2006 to 2010.

PID	MS SubClass	MS Zoning	Lot Frontage	Lot Area	Street	Alley	Lot Shape	Land Contour	Utilities	...	Pool Area	Pool QC	Fence	Misc Feature	Misc Val	Mo Sold	Yr Sold	Sale Type	Sale Condition	SalePrice
526301100	20	RL	141.0	31770	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	5	2010	WD	Normal	215000
526350040	20	RH	80.0	11622	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0	6	2010	WD	Normal	105000
526351010	20	RL	81.0	14267	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	Gar2	12500	6	2010	WD	Normal	172000
526353030	20	RL	93.0	11160	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	4	2010	WD	Normal	244000
527105010	60	RL	74.0	13830	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0	3	2010	WD	Normal	189900

Financial Highlights

Quick Numbers

25%

Houses are sold below $129.5k

AVG

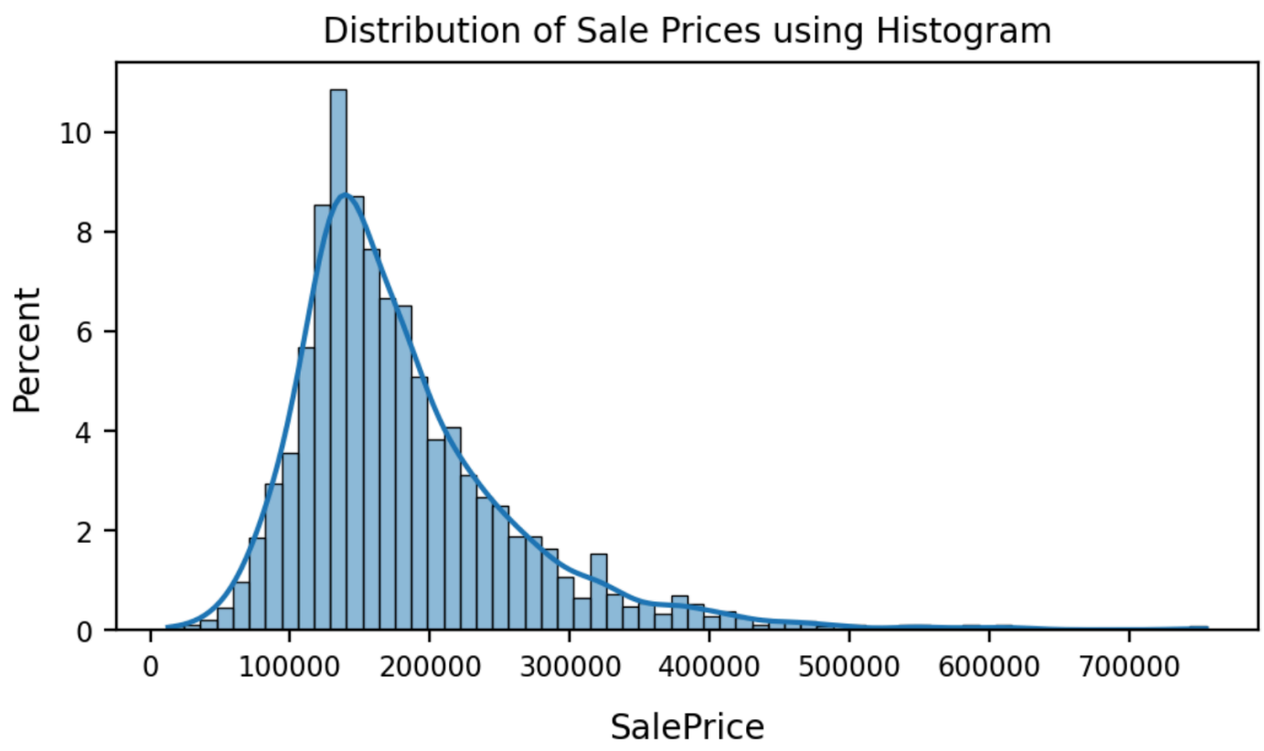
$180.8k Sold Price

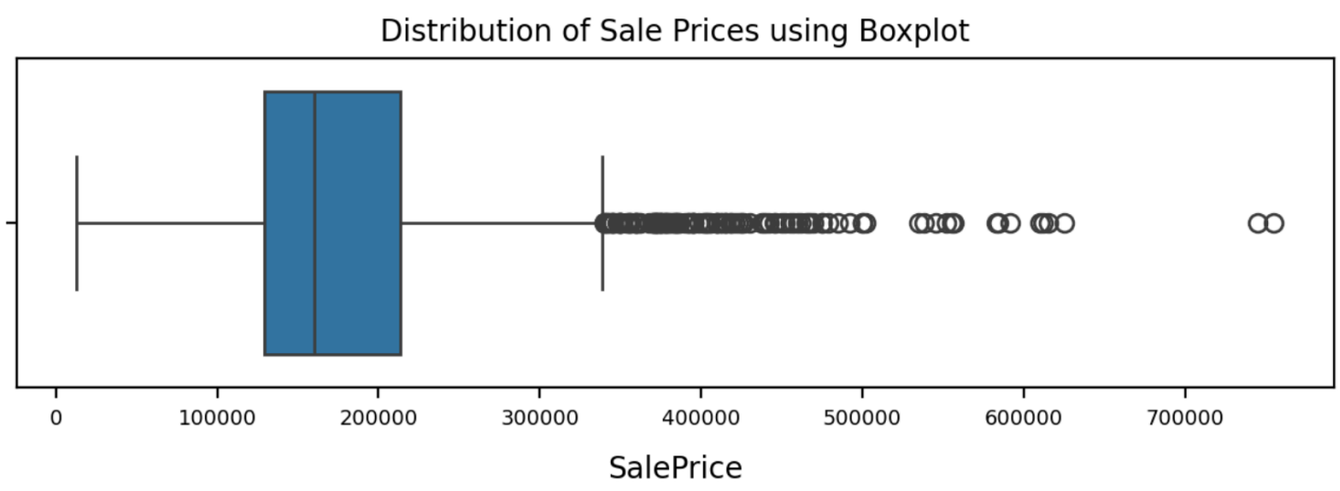
3K

Total Records

Sales Price Distribution

Right-skewed distribution showing the variation towards higher priced houses is lesser than the actual distribution in between 75k to 275k prices (most of the houses)





Correlation Metrics

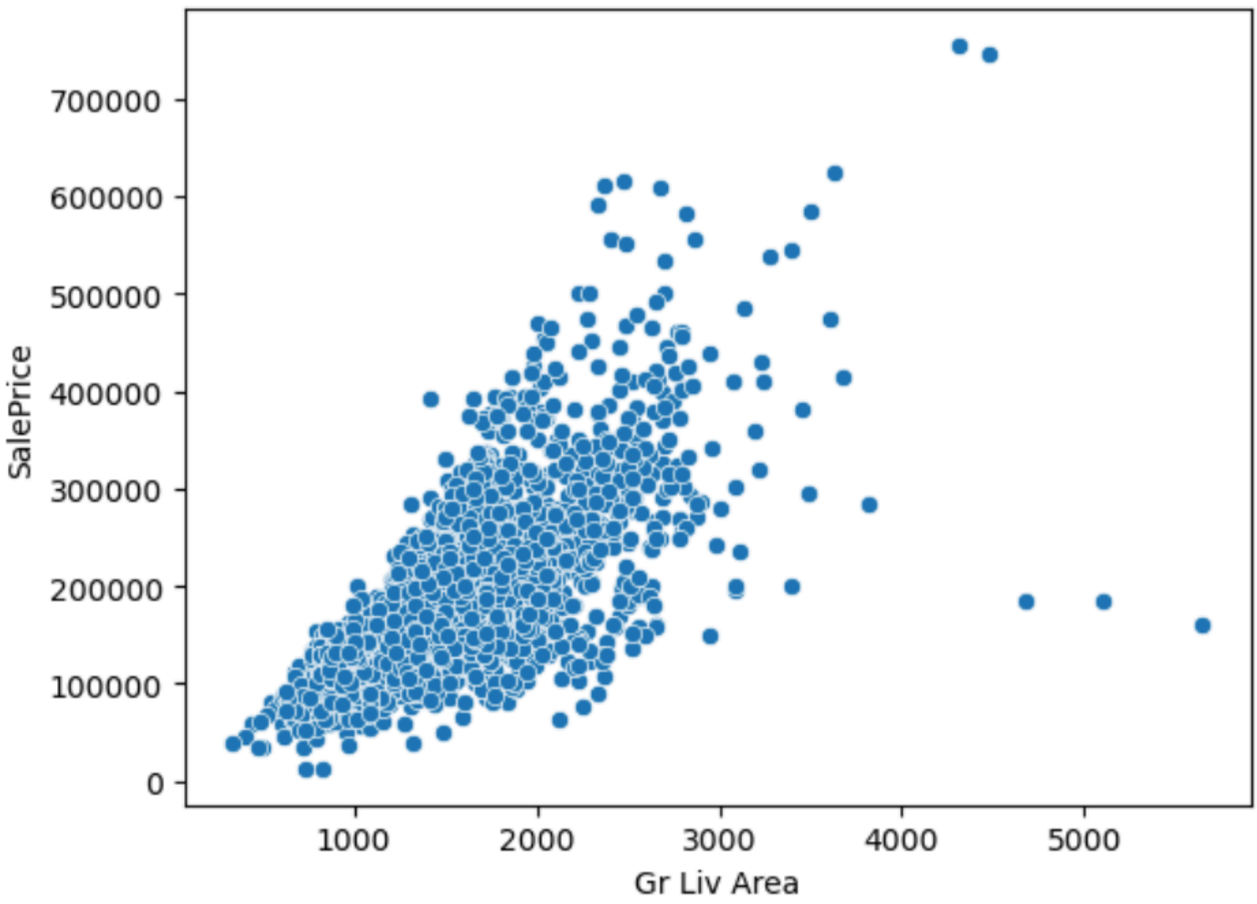


FIGURE: The relationship between the Ground Living Area(in sqft) and Sales Price(in USD)

Observation

- Trend: More Ground Living Area, the higher the house price
- The plot shows a good linear relationship towards this trend, although two houses above 700k outperform but are still under the trend as mentioned.
- Considering three points that are more than 4500 sqft area and don't appreciate much. Thus, it breaks the trend of being highly priced with more area. We can say that they are misleading the data points and will be crucial for removal.

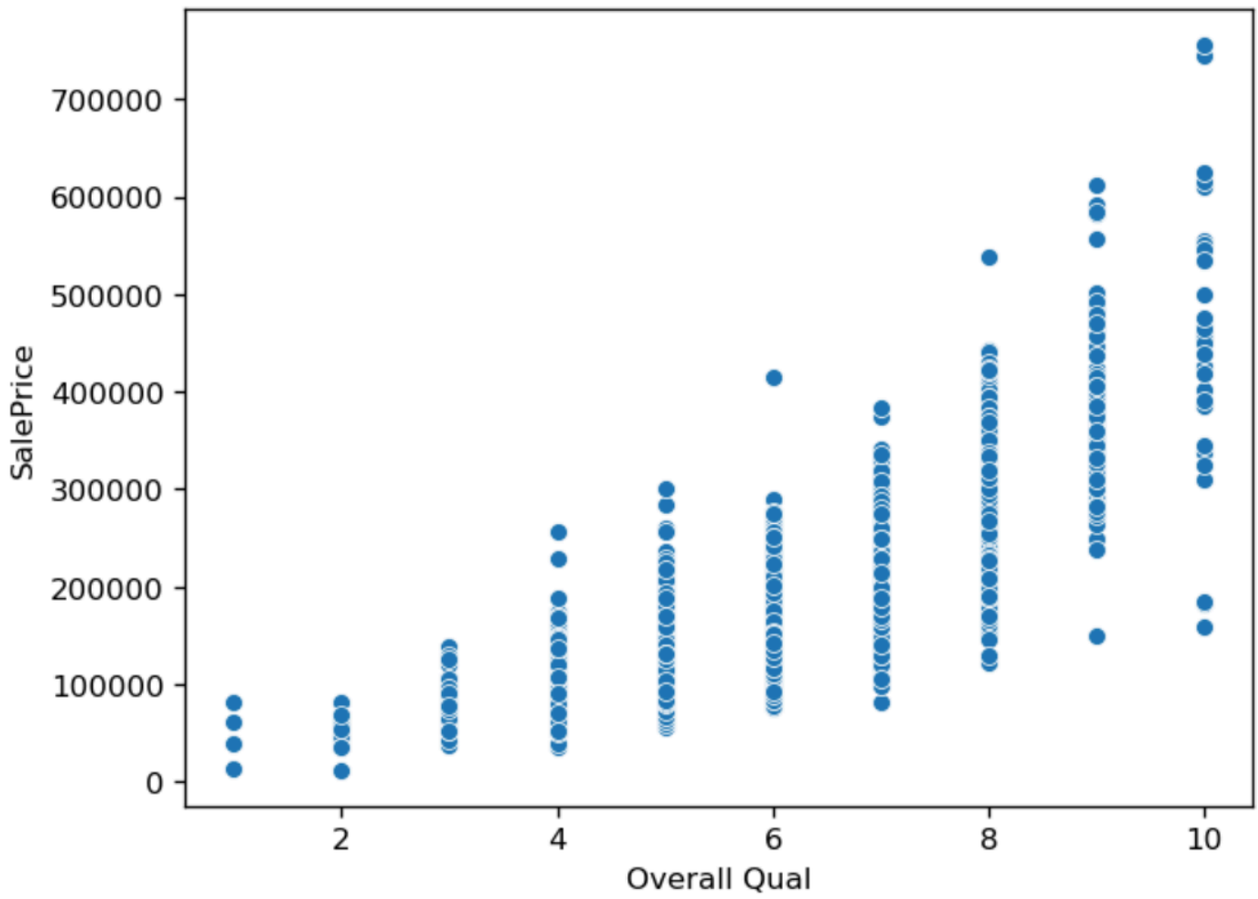


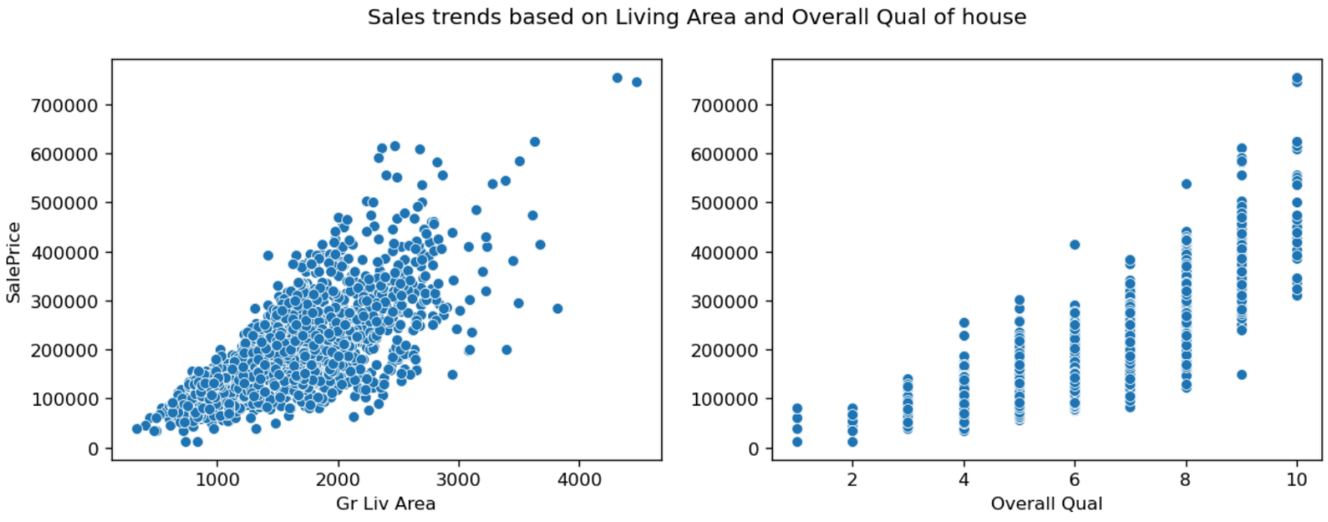
FIGURE: The relationship between the Overall Qualification (in rating) and Sales Price(in USD)

Observation

- It is clear that when the Overall Quality Rating (from 1 to 10) increases, the Sale Price of the house must be higher. So, we can say that points that are not following this trend must be an outliers and will give problems further in developing an optimized model.
- Consider three points here that are below $200k value but have an excellent overall rating of 9 or 10. It means that this misleads the overall trend of being a “highly qualified house having more price"

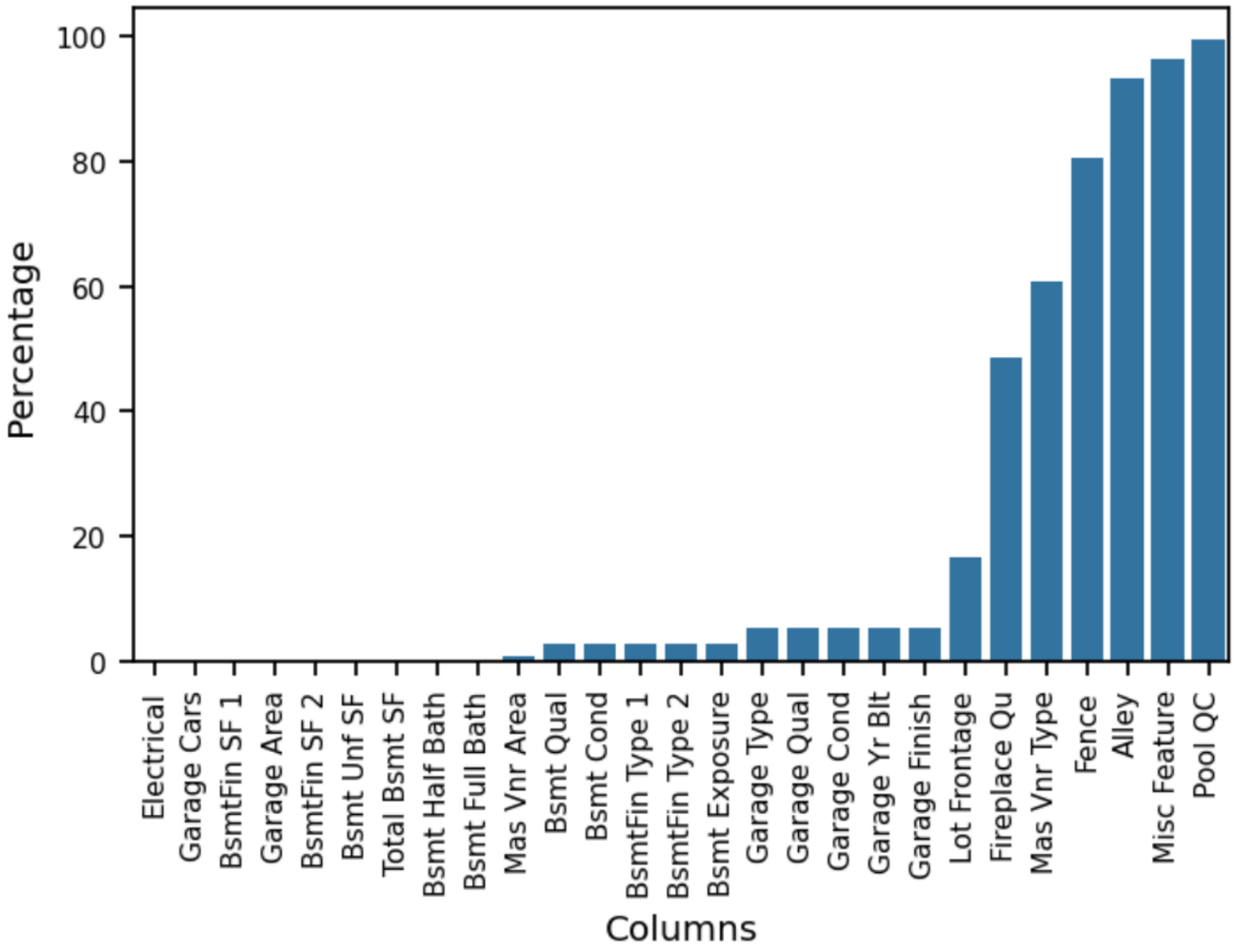
Outlier Removal

Removal of outliers with several techniques brings us to the following results:



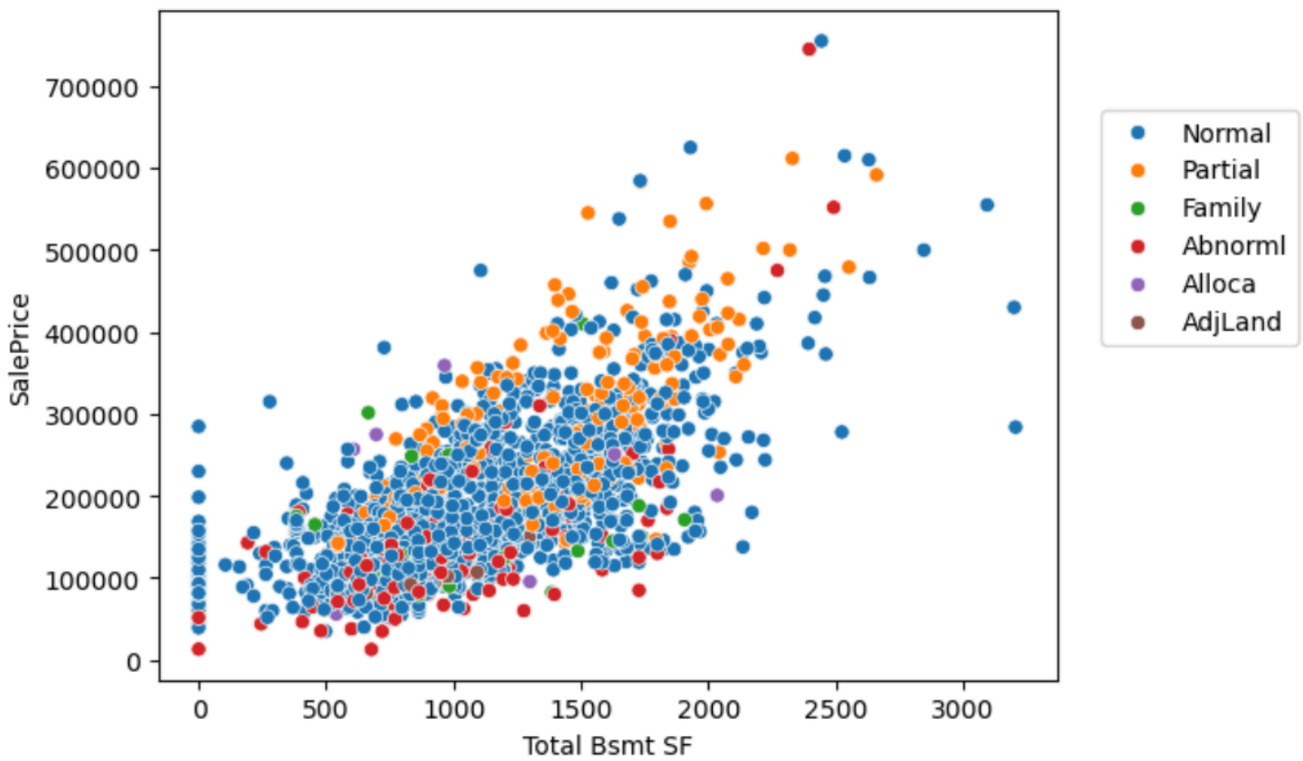
Missing Data

Bar Graph to showcase the missing values in different columns:



Total Basement Area v/s Sale Price

The below scatterplot defines the relationship between the Basement Area(in sqft) with the Sale Price(in USD) w.r.t the different Sale Conditions:



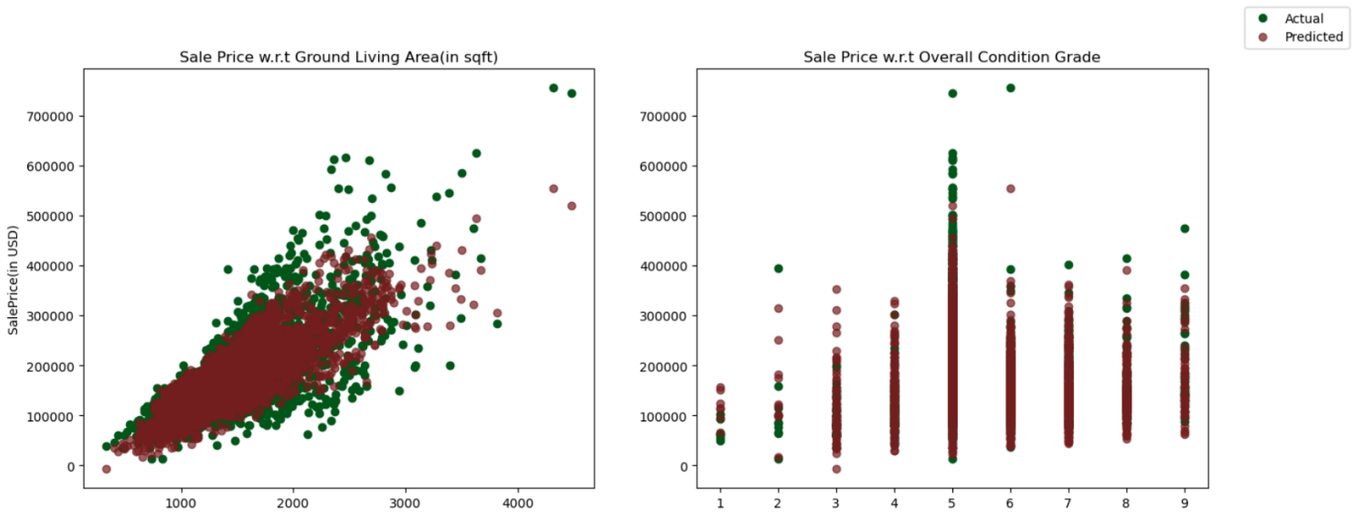


FIGURE 1(Left Graph): Shows the scatterplot of the correlation between Sales Price and Living Area with two distinctive points – Brown shows the predictive values by ML Model

FIGURE 2(Right Graph): Shows the scatterplot of the correlation between overall quality and Living Area with two distinctive points – Brown shows the predictive values by ML Model

Conclusion

- The best-fit model for the dataset is the LassoCV Regression Model.
- The analysis is off by $14414 (the average predicted price would be off by this margin)
- We have to take a marginal 8-9% gap while predicting future prices of Houses