Nikolas Goodrich

MSDS 410

Assignment 1

Section 1

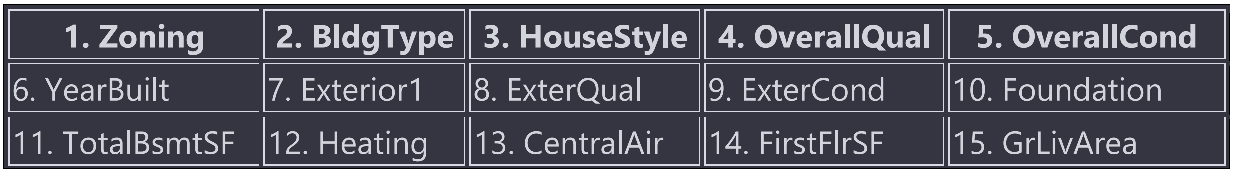
To ensure a focused and consistent analysis, the scope of the study has been narrowed to single-story houses, emphasizing characteristics intrinsic to the domicile itself. This delimitation excludes elements such as external land features, pools, and their geographical location, under the assumption that the primary determinant of property value within this scope is the square footage.

The selection process encompasses various criteria. Within the 'MS SubClass' category, only instances corresponding to one-story dwellings were retained. Zoning classifications have been filtered to encompass only residential properties. Similarly, every house's utility status was noted, with no exclusions. The dataset preserves details about the house's condition, year of construction, type of roofing material, and exterior. Foundations have been restricted to include basement quality, condition, and exposure. All relevant square footage metrics are considered, excluding garage spaces, wood decks, other external accessories, and pools. Sale type and condition are documented, along with the final sale price.

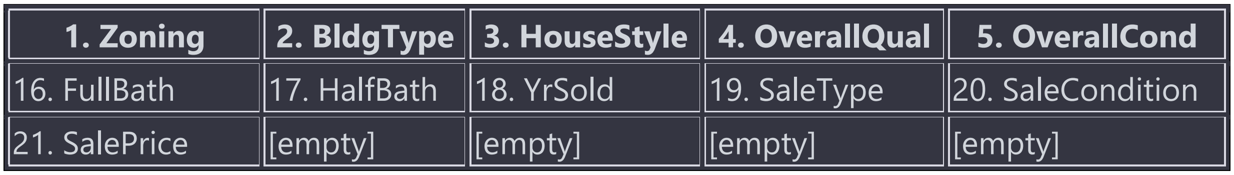
It is pertinent to recognize additional variables that may indirectly influence price, such as the timing of the sale. Initially, the dataset consisted of 2930 entries across 82 categories. The refinement to focus distinctly on single-family homes reduced the sample size to 1481 entries, each with 20 variables, thus streamlining the preliminary analysis.

Section 2

Here is a table of columns I kept.



1. Zoning	2. BldgType	3. HouseStyle	4. OverallQual	5. OverallCond
6. YearBuilt	7. Exterior1	8. ExterQual	9. ExterCond	10. Foundation
11. TotalBsmtSF	12. Heating	13. CentralAir	14. FirstFlrSF	15. GrLivArea

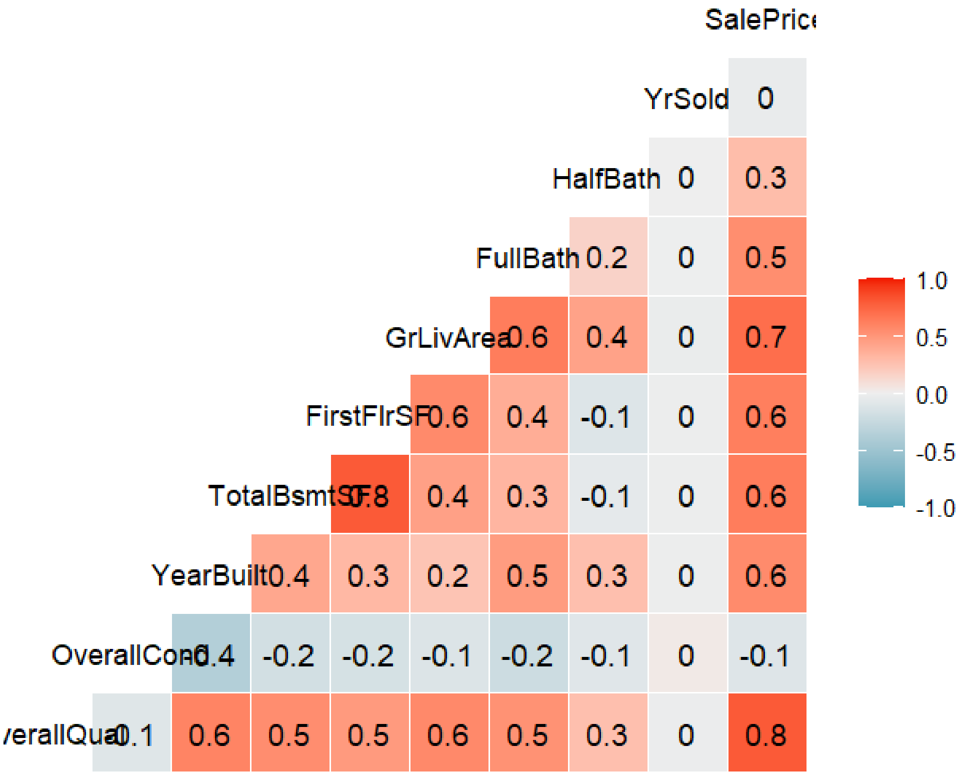


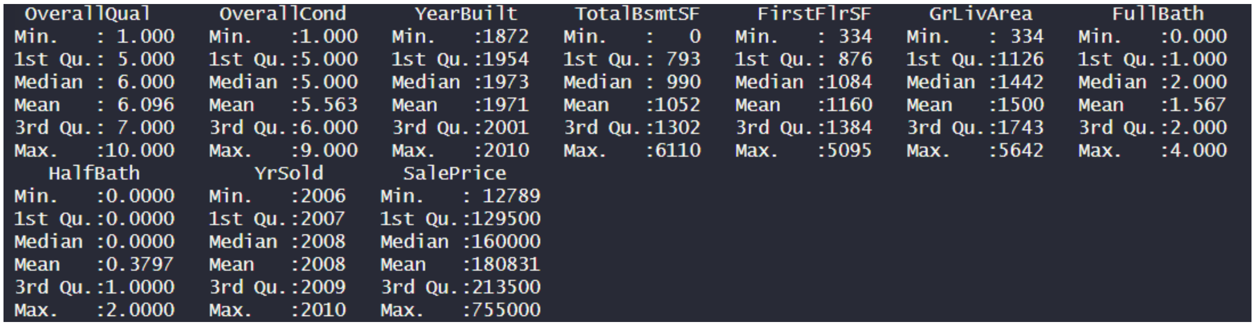
1. Zoning	2. BldgType	3. HouseStyle	4. OverallQual	5. OverallCond
16. FullBath	17. HalfBath	18. YrSold	19. SaleType	20. SaleCondition
21. SalePrice	[empty]	[empty]	[empty]	[empty]

I manually examined unique entries for each variable and complemented this with an automated data integrity check to ensure consistency across the dataset. Before starting the assessment process, I excluded instances with missing values, removing one row. I then explored the correlations among the 20 remaining variables.

My analysis clarifies that a property's square footage significantly influences its price, with larger spaces generally commanding higher financial values. While overall quality is also a significant factor, its quantification is more challenging due to its ordinal nature. This raises questions about the metrics and authority used in assessing house quality, leading to concerns about reliability. Despite its limitations, such as not accounting for degradation in older buildings, square footage shows a more consistent correlation with price than other factors.

Based on these initial findings, I have decided to omit Overall Quality and Overall Condition from future models due to their ordinal nature and unreliable quantification. I will limit our selection to two or three variables to avoid redundancy from highly correlated square footage metrics. I will also emphasize features representing amenities and room types, such as the number of full bathrooms, which have shown a notable correlation (0.5) with the sale price.





```
  OverallQual      OverallCond      YearBuilt     TotalBsmtSF     FirstFlrSF      GrLivArea       FullBath    
 Min.   : 1.000   Min.   :1.000   Min.   :1872   Min.   :   0   Min.   : 334   Min.   : 334   Min.   :0.000  
 1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954   1st Qu.: 793   1st Qu.: 876   1st Qu.:1126   1st Qu.:1.000  
 Median : 6.000   Median :5.000   Median :1973   Median : 990   Median :1084   Median :1442   Median :2.000  
 Mean   : 6.096   Mean   :5.563   Mean   :1971   Mean   :1052   Mean   :1160   Mean   :1500   Mean   :1.567  
 3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2001   3rd Qu.:1302   3rd Qu.:1384   3rd Qu.:1743   3rd Qu.:2.000  
 Max.   :10.000   Max.   :9.000   Max.   :2010   Max.   :6110   Max.   :5095   Max.   :5642   Max.   :4.000  
    HalfBath          YrSold       SalePrice     
 Min.   :0.0000   Min.   :2006   Min.   : 12789  
 1st Qu.:0.0000   1st Qu.:2007   1st Qu.:129500  
 Median :0.0000   Median :2008   Median :160000  
 Mean   :0.3797   Mean   :2008   Mean   :180831  
 3rd Qu.:1.0000   3rd Qu.:2009   3rd Qu.:213500  
 Max.   :2.0000   Max.   :2010   Max.   :755000  
```

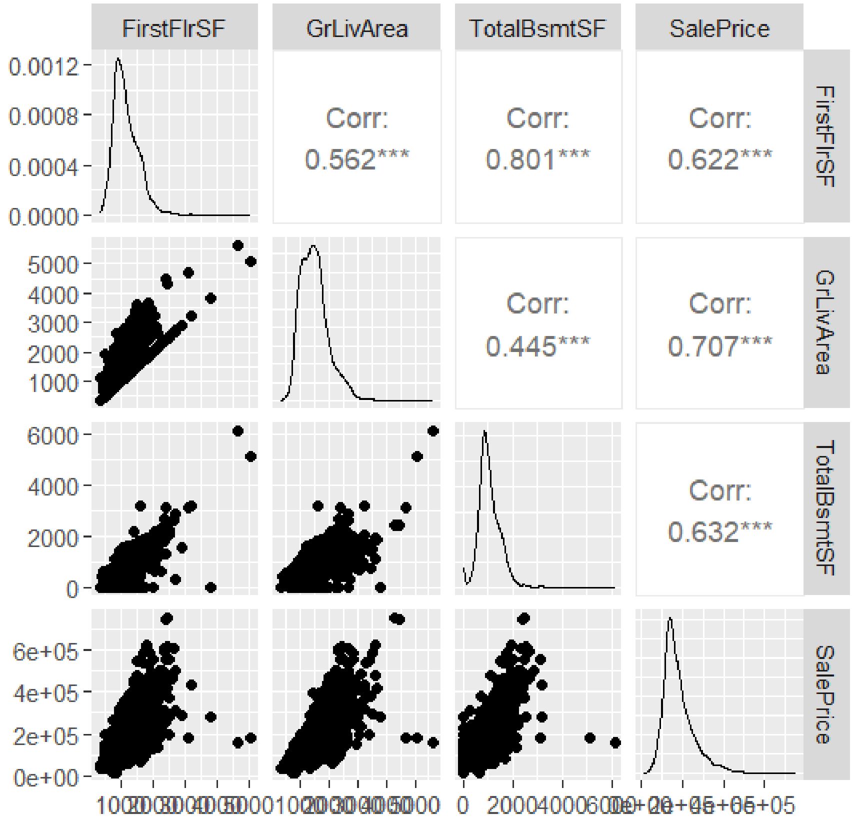
Section 3

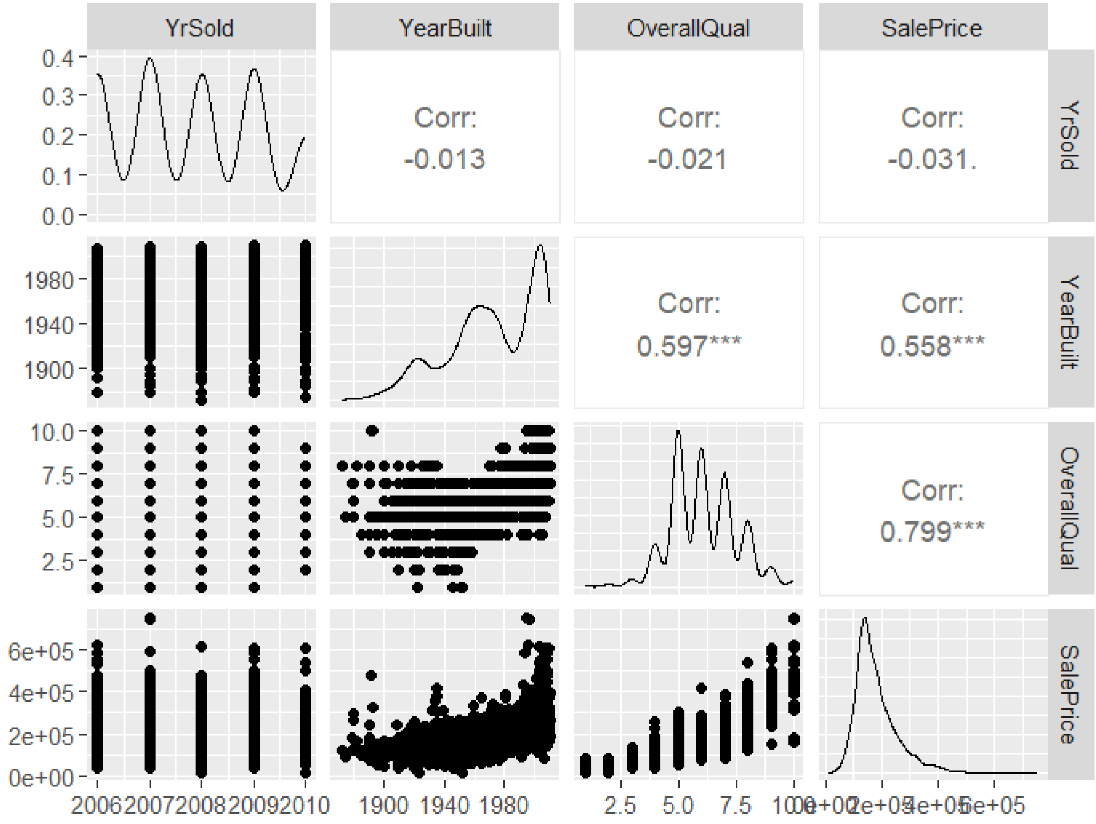
I focused my analysis on integer-based columns, narrowing down my list of variables to ten significant predictors.

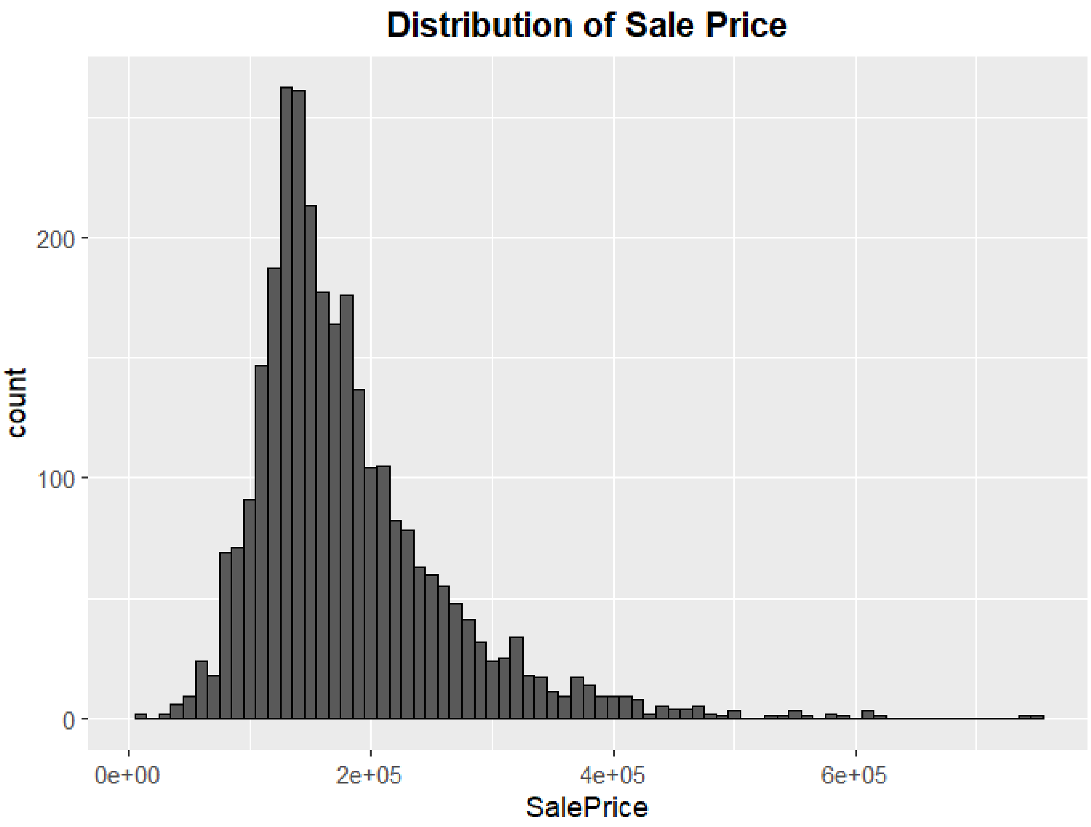
When juxtaposing these variables with Sale Price, I noted a strong (>0.5) correlation with square footage metrics, in line with conventional wisdom. Their simplicity makes them practical for financial institutions processing mortgage loans and predicting sale prices effectively. Although I acknowledge lot sizes as contributing to the cost, I temporarily excluded them to refine my analytical scope.

Interestingly, my correlation analysis showed that the Year Sold and Year Built have low relevance to sale prices. The Year Sold variable showed slight seasonal price variations, suggesting a tactical advantage in timing sales for profit maximization. Also, a newer construction date positively correlates with higher sale prices, likely due to lower anticipated maintenance costs such as repairs.

The Sales Price distribution I observed was right-skewed, with a substantial clustering around the $200k mark. This suggests that the majority of house sales occur within this price range.


FirstFlrSF
GrLivArea
TotalBsmtSF
SalePrice
0.0012
0.0008
0.0004
0.0000
Corr:
0.562***
Corr:
0.801***
Corr:
0.622***
5000
4000
3000
2000
1000
Corr:
0.445***
Corr:
0.707***
6000
4000
2000
0
Corr:
0.632***
6e+05
4e+05
2e+05
0e+00
10002000300040005000
10002000300040005000
0 2000400060000e+002e+054e+056e+05
FirstFlrSF
GrLivArea
TotalBsmtSF
SalePrice


YrSold
YearBuilt
OverallQual
SalePrice
0.4
0.3
0.2
0.1
0.0
Corr:
-0.013
Corr:
-0.021
Corr:
-0.031.
1980
1940
1900
Corr:
0.597***
Corr:
0.558***
10.0
7.5
5.0
2.5
Corr:
0.799***
6e+05
4e+05
2e+05
0e+00
20062007200820092010
1900 1940 1980
2.5 5.0 7.5 10.0
0e+002e+054e+056e+05
YrSold
YearBuilt
OverallQual
SalePrice

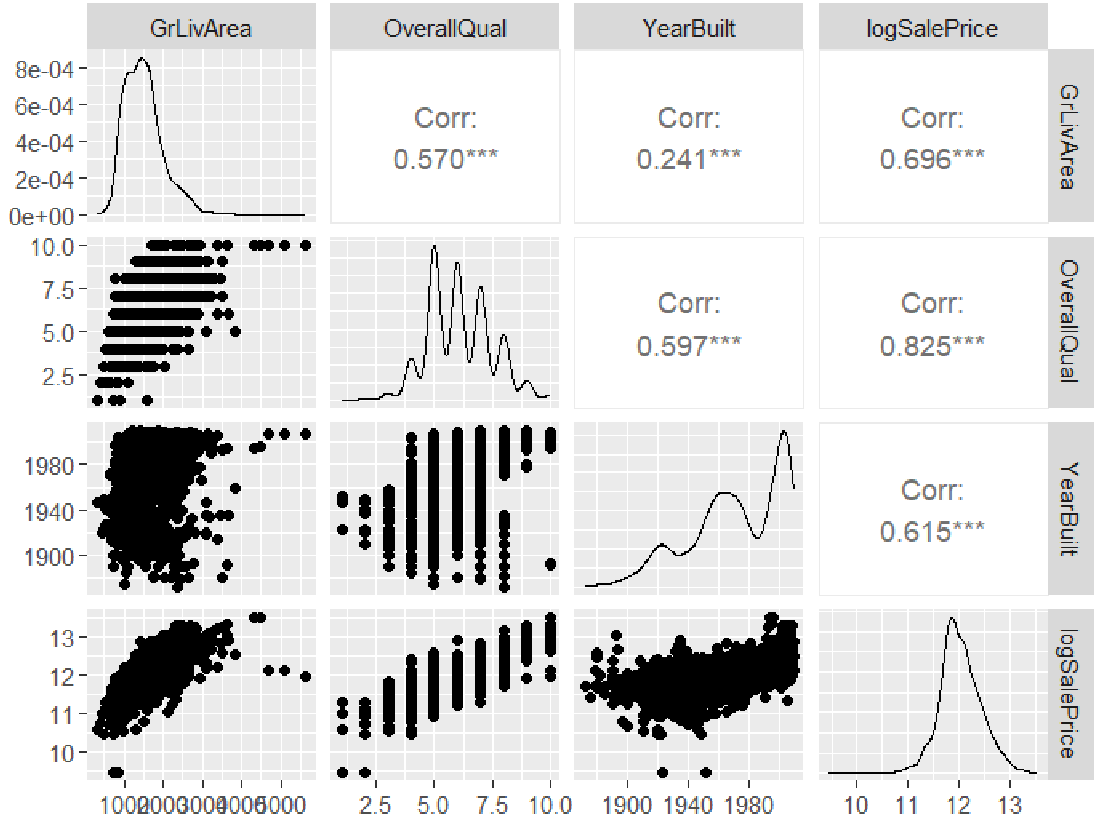


Section 4

To optimize my interpretation of the data, I applied a logarithmic transformation exclusively to the Sale Price variable. I intended this statistical adjustment to normalize the distribution and mitigate the influence of outliers.

My preliminary exploratory data analysis suggests that variables like Above Ground Living Area, Overall Quality, and Year Built will likely be the most robust predictors of Sale Price. I will reevaluate these variables as part of my preparatory work before progressing to model development. However, for the objectives of this project, they stand out as prime candidates for prediction.

Given the observed linear relationship between Living Area and Sale Price, constructing linear regression models seems promising. My dataset's narrowed focus on numerical variables has streamlined my analysis process. I may benefit from including categorical variables through dummy encoding in future model refinement.



Section 5

In my dive into the Ames housing dataset, I've picked up some key insights for my future models. I started by carefully choosing twenty main variables, mostly integers, which were easy to analyze. I focused on variables that matter for Sale Price, our main target.

Square footage stood out as a top predictor of price, showing how much property size affects market value. Even though Overall Quality seemed important, its subjective nature and the lack of a clear way to measure it made me leave it out. Instead, I looked at clear-cut features, like the number of full bathrooms, which lined up closely with Sale Price.

My look at correlations showed that while the year a house was sold or built didn't directly tie to its price, there were hints of seasonal price changes and a liking for newer houses, probably because they need fewer repairs.

I tackled the Sales Price distribution's skew to the right by using a log transformation. This made the data more normal and less affected by outliers. From the start, it seemed like Above Ground Living Area, Overall Quality, and Year Built were key for predicting Sale Price.

As I move to building predictive models, these early findings have set me up for a focused and smart approach. Down the line, I plan to bring in categorical variables using dummy encoding to make my models better and get a deeper sense of what drives house prices in Ames.