Restatement of Problem

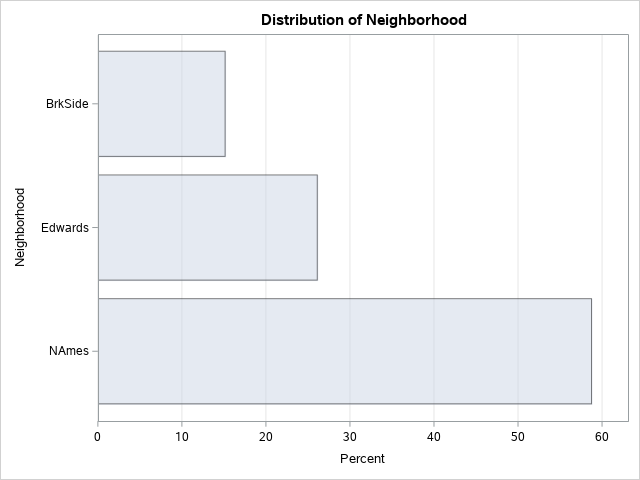
We would like to help the real estate company, Century 21 Ames , to derive meaningful insights from their data, which in turn will drive informed decision making. Particularly, we are interested in modeling the relationship between total square footage of the living area (GrLivArea) of houses in Ames Iowa USA with their Sale prices (SalePrice), specifically in three neighborhoods (Neighborhood) namely; Edwards NAmes and BrkSide.

Build and Fit the Model

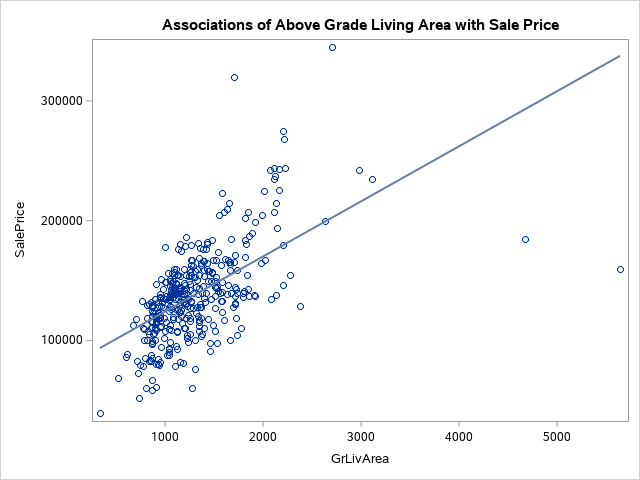
Before model fitting, we explore our data to understand the relationship between Sale Prices and other factors which will help us understand which variables have a strong association with SalePrice, further enabling us to build a model that will determine the effect f the selected variables on our outcome variable. We narrow this analysis down to the three neighborhoods specified above.

Describing the data

The Real Estate company sold 383 houses in BrkSide NAmes and Edwards. Most of which were from the NAmes Neighborhood. The plot below summarizes the distribution of houses sold across the three neighborhoods. The average general living area size is about 1301. Also the average sale price of a home in Ames Iowa USA is around $138062. Whereas the house with the largest general living area had a living area of 5642 ,the most expnsive house was worth $345,000.

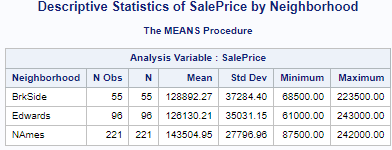


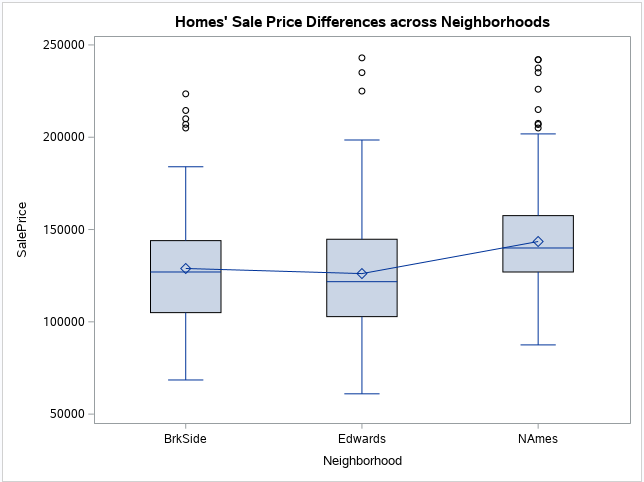
Here we can see an upward trend implying a positive linear relationship.between Sale price and square footage of the general living area of a house in ames iowa USA.we can deduce that an increase in size of general living area implies an increase in the house’s sale price.

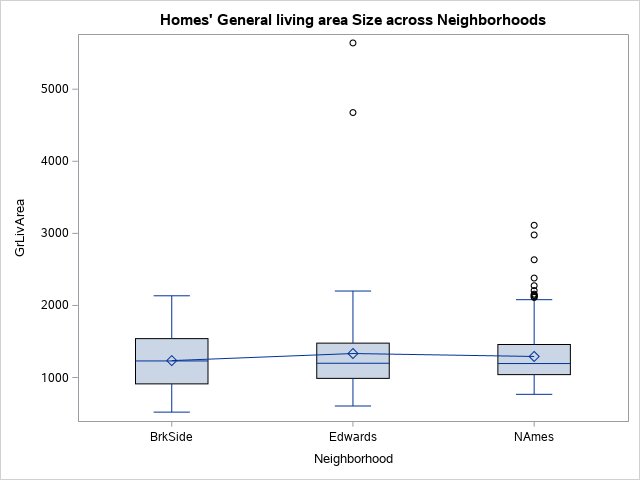


We can also see the average sale prices across the neighborhoods in the following table.

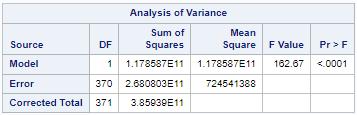
NAmes neighborhood had the house with the highest average sale price with the cheapest being $87500 and the most expensive being $242000.





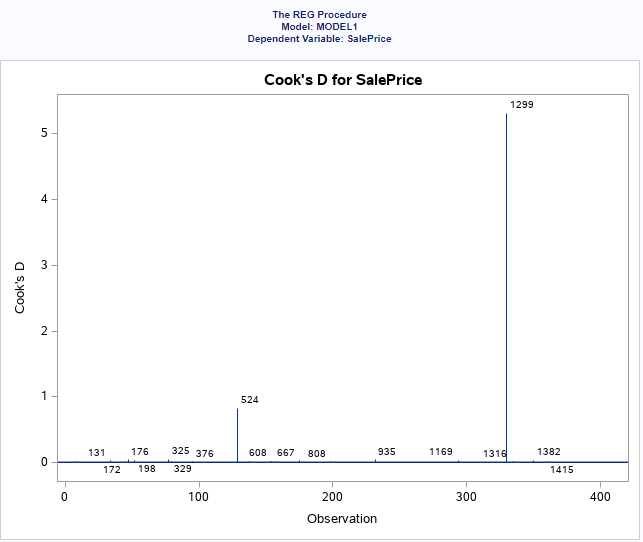


Modelling

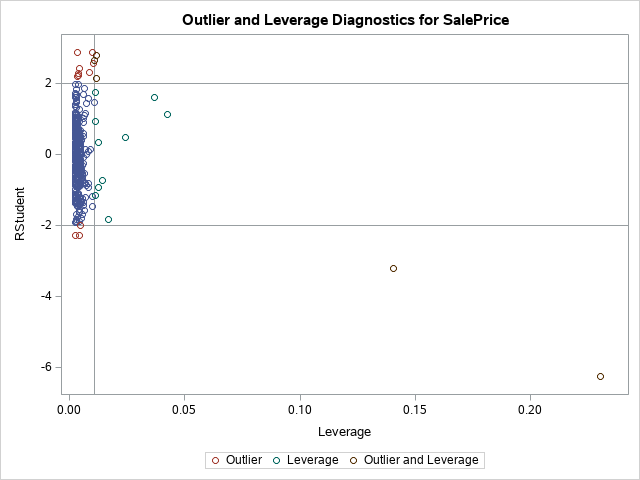


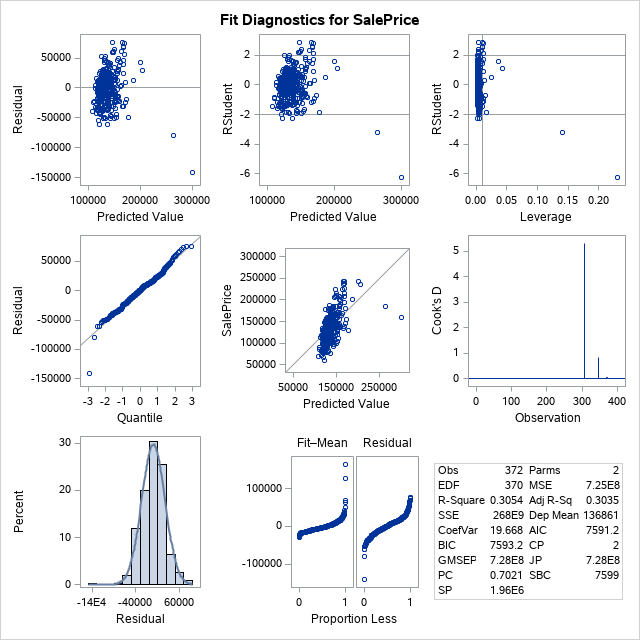
Checking Assumptions

Residual Plots and Influential point analysis (Cook’s D and Leverage)



we see that removing houses in indexes 1299 and 524 would lift the regression analysis up/or cause a sequential shift in the value o the predicted sales price





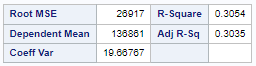
The Histogram of the residuals which is the plot in the bottom left suggests that the model meets the assumption that the error term follows a normal distribution as it assumes a bell-curve shape.

The residual plot shows that the dots are wrapped along the line tightly suggesting that the error term follows a normal distribution.

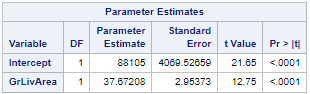
Homogeneity of variance is evidenced as shown by the evenly spread residuals on either side of the horizontal zero line, however we can spot outiers that may skew the distribution. The histogram at the bottom left shows the outlier indeed skews the distribution to the left.

Comparing Competing Models

Adj R2



Parameter estimates



Interpretation

An increase of the General living area of a house by a square foot causes an increase in its sales price by $38.

Holding all factors constant , The average sales price of a house is $88105.

Confidence Intervals

Conclusion

We have derived patterns within our data that are ready for presentation to out team. We have established that the size of the General living area of a house affects the sale rpice of a house in Ames Iowa USA.

Adjusting for the three neighborhoods added complexity in the model hence giving u even more information that could open way for further analysis. For example why did the Century 21 Ames company sell more houses in NAmes than the other neighborhoods?

**Analysis Question 2**

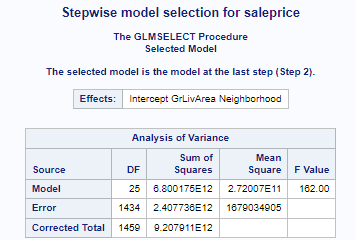
Restatement of Problem

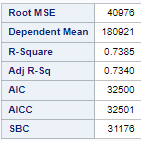
In this context the analysis problem is still the same (modeling the relationship between the sale price of a particular home in Ames Iowa USA with the size of it’s general living area in square feet) for our client company Century 21 Ames, variation being that we consider all Neighborhoods as a confounder, rather than only the specified three in the previous problem.

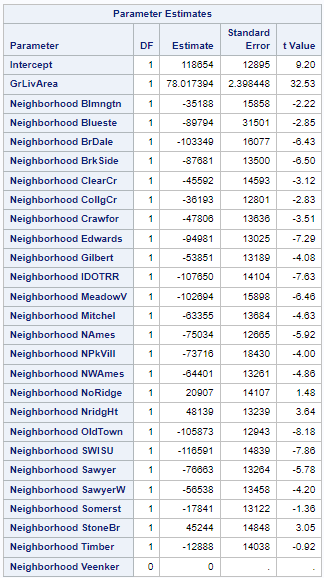
Model Selection

Type of Selection

Stepwise

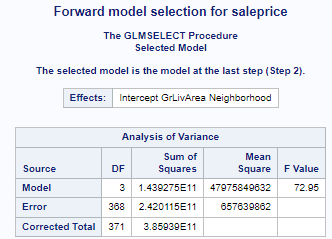


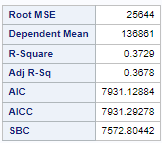


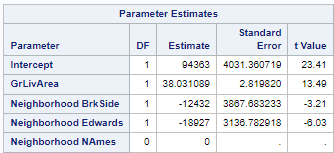


The stepwise model gives a detailed estimates for every Neighborhood and how the location of a house in these neighborhoods affect the saleprice, adjusting for the size of the general living area. The SWI SU neighborhood decreases the average sale priceof a house by $116,591

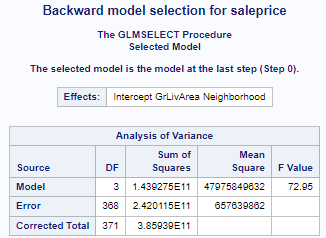
Forward

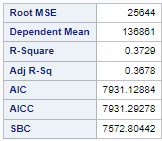


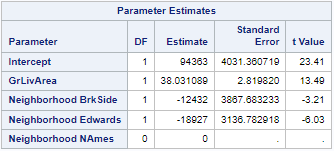




Backward



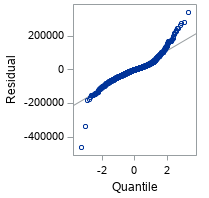


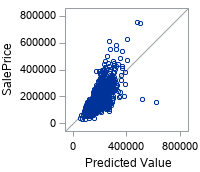


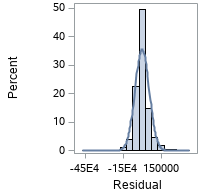
Custom

Checking Assumptions

Residual Plots



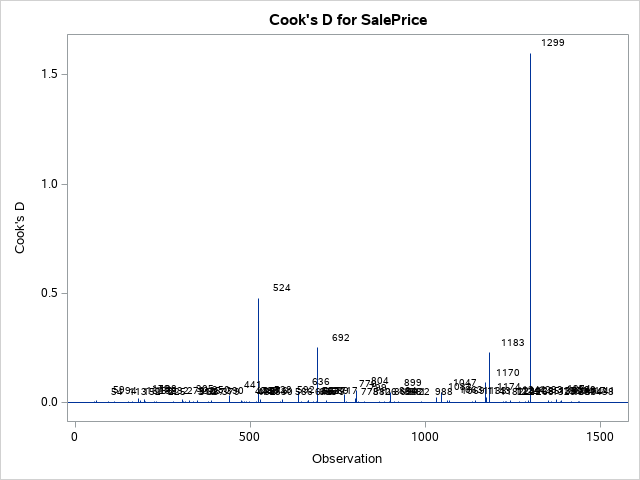




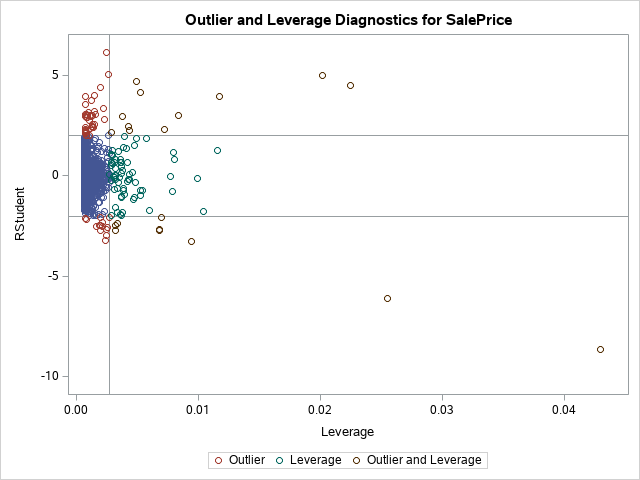
The Histogram of the residuals which is the plot in the bottom left suggests that the model meets the assumption that the error term follows a normal distribution as it assumes a bell-curve shape.

The residual plot shows that the dots are wrapped along the line tightly sugeesting that the error term follows a normal distribution.

Influential point analysis (Cook’s D and Leverage)



Comparing to analysis one, we see that alongside observation 1299 and 524 removing 692 and 1183 would lift the regression analysis up or cause a sequential shift in the value of the predicted sales price.



Comparing Competing Models

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictive Models** | **Adjusted R2** | **CV PRESS** | **Kaggle Score** |
| Forward | .3678 | 19.66767 |  |
| Backward | .3678 | 19.66767 |  |
| Stepwise | .7340 | 19.66767 |  |
| CUSTOM | .3678 | 19.66767 | . |

Conclusion:

The stepwise model explained 73% of the variation in the sales price as affected by the size of the genera living room area as well as the neighborhood the house was located in.

Factoring in all the neighborhood yields a more complex model of how the size of the general living area of a house sold in Ames Iowa USA affects its sale price but also where it is located matters.

Appendix

Detailed comments and code for Analysis 1 and 2