THE DETERMINING FACTORS OF RESIDENTIAL HOMES SALES PRICE IN AMES, STATE OF IOWA

Prepared By;

Jonathan Rocha and Samson Akomolafe

Being Final project for Statistical Foundation For Data Science in fulfillment of Master of Science in Data Science.

INTRODUCTION OF THE PROJECT

This Project explores the Ames Housing Market dataset, which contains information on residential homes sales in Ames, State of Iowa.

The dataset contains 79 explanatory variables describing various aspects of a homes, our goal is to predict the final sale price of each property in the test data using the provided trained data. The analysis is divided into two main parts:

Analysis 1: To examine how house sale prices is related to the ground living area of a house which are located in three specific neighborhoods - NAmes, Edwards and BrkSide for Century 21 Ames.

Analysis 2: To Build a predictive models for house prices across all neighborhoods in Ames, Iowa State using the trained data.

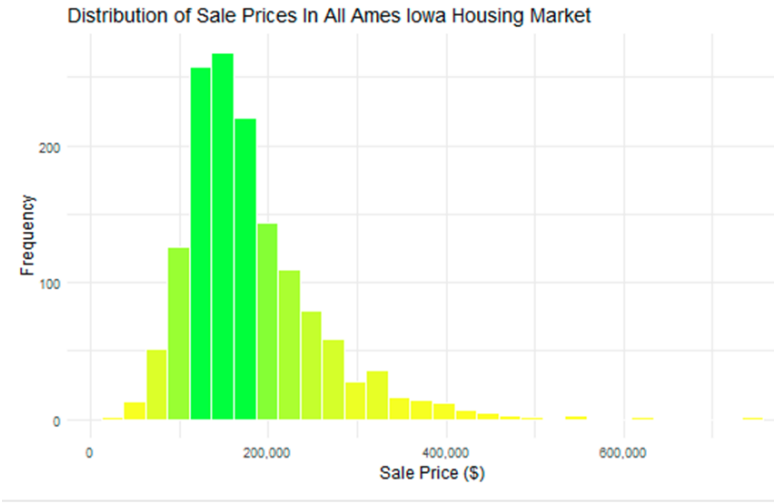
Data Description

The dataset contains information on residential homes in Ames, Iowa State, with 79 explanatory variables describing almost every aspect of a house up for sale and is suitable for both exploratory analysis and predictive modeling.

There are 2 datasets provided, a training dataset and testing dataset

Training set contains: 1,460 observations
Testing set contains: 1,459 observations

The Target variable: SalePrice (the property's sale price in US dollars) . The histogram of Sales price inside the entire dataset is shown in the next below.



From the histogram on the left hand side, we are able to deduce the following;

1. 70.21% of the homes in Ames Iowa State have prices below '$200,000'.
2. 27.33% of the homes in Ames Iowa State have prices between '$200,000 and $400,000'.
3. 1.918% of the home have price above '$400,000'.

Analysis Question 1

The analysis 1 is geared to analyzed the relationship between Sale Price and Ground Living Area in Selected Neighborhoods

Restatement of Problem

Century 21 Ames seeks to understand how the SalePrice of homes in the Selected neighborhood - NAmes, Edwards, and BrkSide is related to the square footage of the ground living area (GrLivArea). Specifically, the company want to quantify this relationship and see if it varies by the selected neighborhood.

Build and Fit the Model

Data Preparation

We filtered the dataset to include only observations from these 3 neighborhoods NAmes, Edwards, and BrkSide and transformed GrLivArea into increments of 100 sq. ft. by dividing its values by 100.

Model Specification

We built 4 models for analysis 1, in order to test the Sales Prices of homes by Ground living area and within the 3 neighborhoods using Edward neighborhood as the reference.

Model 1: GrLivArea
Model 2: GrLivArea + Neighborhood
Model 3: GrLivArea * Neighborhood,
Model 4: GrLivArea + Neighborhood

Fit the linear regression model

We found that the model 3 statistics is better among all the 4 models we used for analysis 1 and so we decided to adopt the coefficients of model 3. It is shown as below;

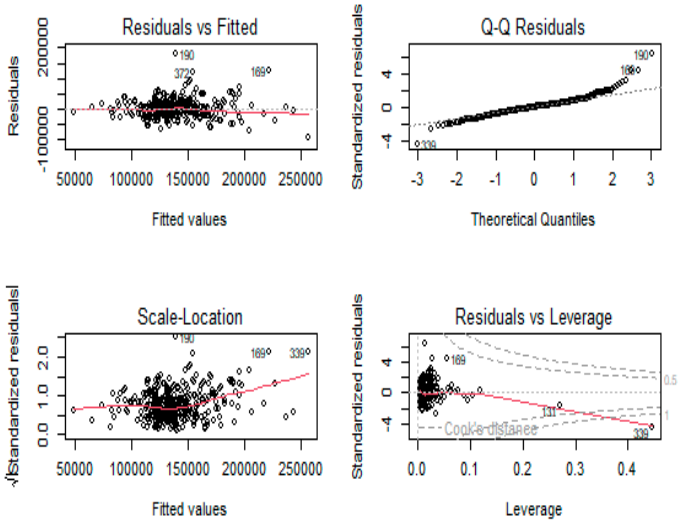
Interpretation

$\beta 0$ = Intercept = 88353.1
$\beta 1$ = GrLivArea_100 coefficient = 2975.0
$\beta 2$ = NeighborhoodBrkSide coefficient = -68381.6
$\beta 3$ = NeighborhoodNAmes coefficient = -13676.7
$\beta 4$ = GrLivArea_100:NeighborhoodBrkSide coefficient = 5741.2
$\beta 5$ = GrLivArea_100:NeighborhoodNAmes coefficient = 2456.6

The is our regression equation

Ames | SP = 88353.1 + 2975.0GrLivArea_100 - 68381.6BrkSide -13676.7NAmes 5741.2GrLivArea_100:BrkSide+2456.6GrLivArea_100:NAmes

Checking Assumptions

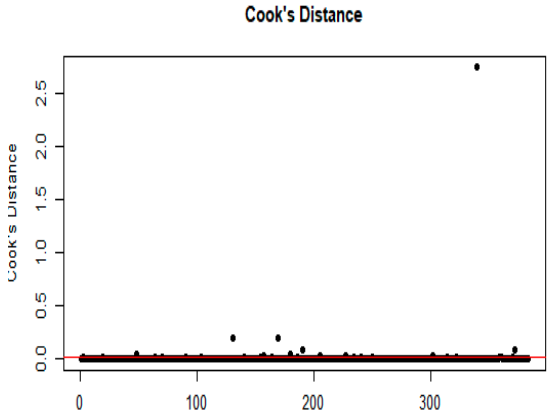


The Residual appears to pass through zero and the points concentrated around price of $130,000 and $185,000.

The QQplot appears to be normal with the points along the diagonal line.

Influential Point Analysis

	Id	Neighborhood	GrLivArea	SalePrice
	<int>	<fctr>	<int>	<int>
19	71	NAmes	2223	244000
48	176	Edwards	2158	243000
64	251	BrkSide	1306	76500
70	276	BrkSide	1548	205000
90	363	Edwards	1922	198500
131	524	Edwards	4676	184750
140	547	BrkSide	1635	210000
157	608	Edwards	2008	225000
164	629	NAmes	2080	135000
169	643	NAmes	2704	345000
180	667	NAmes	2380	129000
190	725	Edwards	1698	320000
205	808	BrkSide	1576	223500
227	889	NAmes	2217	268000
234	911	NAmes	2274	154300
240	935	NAmes	2069	242000
250	975	BrkSide	2019	167500
302	1169	Edwards	2108	235000
314	1224	NAmes	2136	137900
322	1251	NAmes	2110	244000
339	1299	Edwards	5642	160000
360	1382	NAmes	2117	237500
372	1424	Edwards	2201	274970



Model	Adj_R2	CV_PRESS	AIC	BIC
Model 1: GrLivArea	0.3406	393079209100	9012.303	9024.147
Model 2: GrLivArea + Neighborhood	0.3917	363653838311	8983.369	9003.109
Model 3: GrLivArea * Neighborhood	0.4400	340967460744	8953.631	8981.267
Model 4: GrLivArea + Neighborhood	0.3917	363653838311	8983.369	9003.109

From above, metrics, we can see that model3 (GrLivArea*Neighborhood) has the best metrics)

These points that have high residuals (far from zero) but they are not necessarily high leverage and so may indicate outliers. There are 23 points in the influential analysis

While they may not significantly affect the regression line, they can still affect the overall fit but we will proceed anyway.

Confidence Intervals

	2.5 %	97.5 %
(Intercept)	-4314.212	44257.239
GrLivArea_100	6792.850	10639.657
NeighborhoodEdwards	40913.670	95849.512
NeighborhoodNAmes	27408.383	82001.393
GrLivArea_100:NeighborhoodEdwards	-7848.612	-3633.834
GrLivArea_100:NeighborhoodNAmes	-5411.269	-1158.065

Impact of Living Area - The coefficient for GrLivArea_100 is consistently positive and statistically significant across models, indicating that larger living areas are associated with higher sale prices. Specifically, for every additional hundred square feet of living area, the sale price increases by approximately $4,576 to $8,716, depending on the model used for prediction.

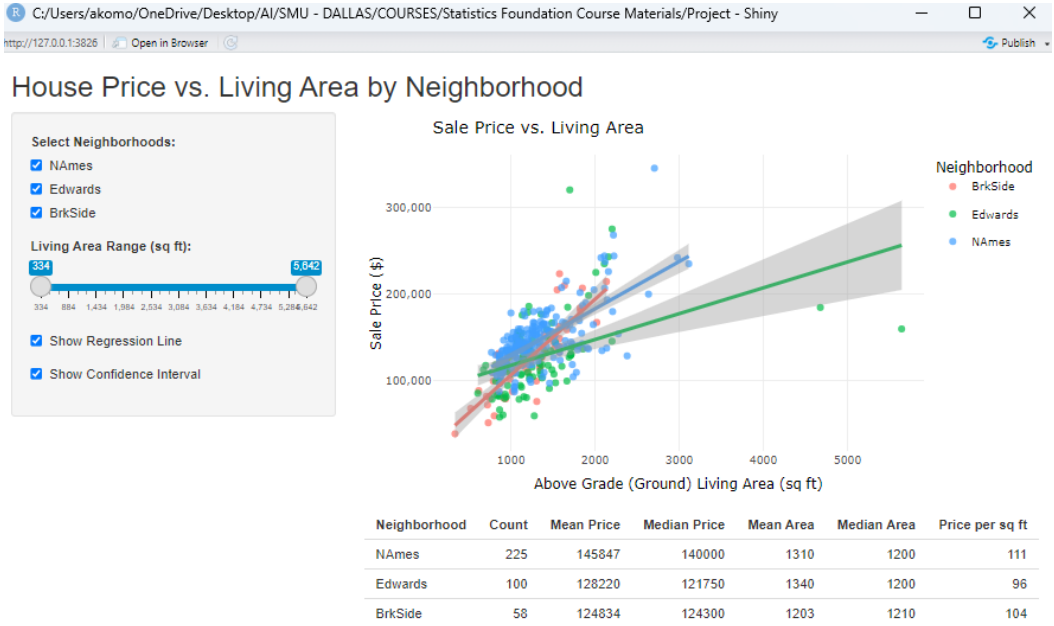
Neighborhood Effects - Different neighborhoods exhibit various impacts on sale prices. For instance, being in the Edwards neighborhood adds approximately $28,882 to $68,381 to the sale price compared to the reference neighborhood (BrkSide), while NAmes also contributes positively with a range of $27,408 to $54,705.

Interaction Effects - In Edwards, the impact of additional 100 square footage is reduced by about $5,741 compared to the baseline neighborhood, indicating localized market force at play.

Home Pricing Distribution - A substantial portion of homes (70.21%) sold for below $200,000 as earlier mentioned above, with only 1.92% exceeding $400,000. This suggests a market skewed towards lower-priced homes, which may reflect economic conditions or buyer demographics.

The Shiny App

Shiny app is a web application framework for R, is allows us to create web applications directly inside R environment. So we have built a web application with shiny app which allows any user to check the sales prove of a house using the ground living area in square footage. A screenshot of the page is included below



Neighborhood	Count	Mean Price	Median Price	Mean Area	Median Area	Price per sq ft
NAmes	225	145847	140000	1310	1200	111
Edwards	100	128220	121750	1340	1200	96
BrkSide	58	124834	124300	1203	1210	104

Analysis 2: Predictive Modeling for All Neighborhoods

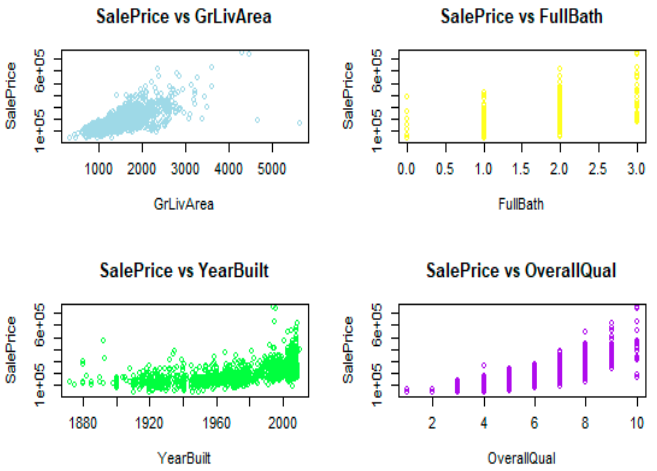
Restatement of Problem

For this analysis, we need to build the most predictive model for saleprices of homes in all of Ames, Iowa. We'll create at least three competing models:

1. A simple linear regression model2.
2. A multiple linear regression model with GrLivArea and FullBath as predictors
3. At least one additional multiple linear regression model with selected variables.

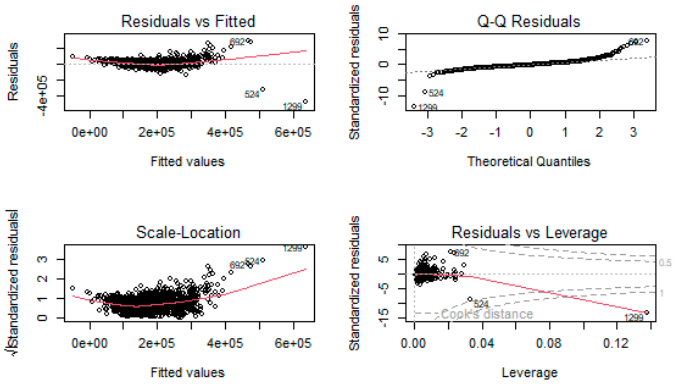
We'll compare these models using adjusted R^2, CV PRESS, AIC, and KaggleScore.

<u>Sale price relationship with the predictors we are using for analysis 2</u>

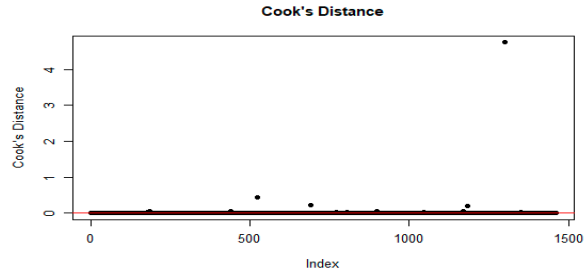


1. Ground living area sales is mostly around 1000 – 3000 sqf.
2. Houses for sale are around zero to 3 bath rooms
3. Most of the house were year 2000s, though other old houses are also in the market.
4. The ratings of 9 and 10 have the highest selling prices

Checking Assumptions
Residual Plots



Influential point analysis (Cook's D and Leverage)



	Id	GrLivArea	GrLivArea.1	OverallQual	YearBuilt	TotalBsmtSF	GarageCars	SalePrice
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
4	4	1717	1717	7	1915	756	3	140000
54	54	1842	1842	9	1981	1842	3	385000
59	59	2945	2945	10	2006	1410	3	438780
67	67	2207	2207	7	1970	1947	2	180000
94	94	2291	2291	6	1910	1214	2	133900
113	113	2696	2696	7	2007	1264	3	383970
138	138	1959	1959	7	1988	1907	3	171000
152	152	1710	1710	8	2007	1710	3	372402
179	179	2234	2234	9	2008	2216	3	501837
186	186	3608	3608	10	1892	1107	3	475000

These points that have high residuals (far from zero) but they are not necessarily high leverage and so may indicate outliers. There are 10 points in the influential analysis

While they may not significantly affect the regression line, they can still affect the overall fit but we will proceed anyway.

Model	Adj_R2	CV_PRESS	AIC	Kaggle_Score
Simple Linear Regression	0.5018	4.631820e+12	36075.76	NA
Multiple Linear Regression 1	0.5231	4.441743e+12	36012.90	NA
Multiple Linear Regression 2	0.7672	2.248292e+12	34969.20	NA

```
Call:
lm(formula = SalePrice ~ GrLivArea + OverallQual + YearBuilt +
    TotalBsmtSF + GarageCars, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-476135  -20008   -2706   16431  285876

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.720e+05  8.485e+04  -7.920 4.69e-15 ***
GrLivArea    5.083e+01  2.564e+00  19.825  < 2e-16 ***
OverallQual  2.039e+04  1.156e+03  17.633  < 2e-16 ***
YearBuilt    3.014e+02  4.459e+01   6.760 1.99e-11 ***
TotalBsmtSF  2.998e+01  2.821e+00  10.628  < 2e-16 ***
GarageCars   1.451e+04  1.824e+03   7.957 3.52e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38330 on 1454 degrees of freedom
Multiple R-squared:  0.768,     Adjusted R-squared:  0.7672
F-statistic: 962.4 on 5 and 1454 DF,  p-value: < 2.2e-16
```

Interpretation

$\beta 0$ = Intercept = -672,000

$\beta 1$ = GrLivArea coefficient = 50.53

$\beta 2$ = OverallQual coefficient = $20,390

$\beta 3$ = YearBuilt coefficient = $301.40

$\beta 4$ = TotalBsmtSF coefficient = 29.28

$\beta 5$ = GarageCars coefficient = 14,510

Ames | SP = $\beta 0$ + $\beta 1$GrLivArea + $\beta 2$OverallQua + $\beta 3$NAmes + $\beta 4$GrLivArea_100:BrkSide + $\beta 5$GrLivArea_100:NAmes.

The is our regression equation

Ames | SP = -672000 + 50.53GrLivArea + 20390OverallQual + 301.40YearBuilt + 29.28 TotalBsmtSF +14510GarageCar

Coefficients:

1. Intercept: -672,000 This indicates the estimated sale price when all predictors are zero, this scenario may not be realistic though.
2. GrLivArea - For each additional square foot of above-ground living area, the sale price increases by approximately $50.83.
3. OverallQual - For each one-unit increase in overall quality, the sale price increases by about $20,390.
4. YearBuilt - Each additional year in the age of the home is associated with an increase in sale price of approximately $301.40, this is if newer homes are more valuable than older build houses.
5. TotalBsmtSF - For each additional square foot of basement area, the sale price increases by about $29.98.
6. GarageCars - Each additional car capacity in the garage is associated with an increase in sale price of about $14,510.
7. All predictors are statistically significant (p-values < 0.001), indicated by the asterisks. This suggests that each variable significantly contributes to the model.

Model Fit

Residual Standard Error (38,330)This value indicates the average distance that the observed values fall from the regression line.Multiple

R-squared (0.768) About 76.8% of the variability in SalePrice can be explained by the model's predictors.

Adjusted R-square (0.7672) This adjusted value accounts for the number of predictors in the model, providing a more accurate measure of lack-of-fit.

F-statistic: 962.4 with a p-value < 2.2e-16 - This indicates that the model is statistically significant overall, suggesting that at least one of the predictors significantly predicts SalePrice.

Kindly check our appendix for the codes used to do this project.

Thank you for your time.