1 Data Prep, EDA, and Theory development

1.1 Varaible Selection & Explanation

For the purpoase of analyzing the determinants of prices of home sales in the US, the followign variables were included in the analysis (Table 1):

- Sale Price (DV)
- Total Living Space (IV) & Years Since Remodeling (at point of Sale) (IV)
- Confounders & Controls: Quality, Lot Area, Condition

As can be seen in Table 1, a total of 1,460 house sales were recorded between 2006 and 2010 for the district of Ames, Iowa (USA). As can be observed in Table 1, the mean sale price of a house was (in 1000s) \$180.921 (SD = 79.443). Combined with the range [34,900, 755,000] a positive skewness was to be expected (skew = 1.881), considering that the outcome variable is a of financial nature. The Total Living Area displays a mean of 2,572.89 square feet (SD = 823.598) in addition to a large range of values[334, 11,752]. Years Since remodeling (at time of sale) shows that the average property did not undergo renovations for 22.95 (23) years (SD = 20.950). Contrary to expectation, this varaible distributes reasonably equally across the range, stopping out at a maximum of 60 years (See Figure 1B - quantiles). Furhermore, the variable Quality (and Condition) represents a rating from 1 to 10, similar to a Likert Scale. Quality has to be considered a categorical variable in this case i.a. because the distances between each rating level are not constant and the distribution is skewed **(SEE SUPPLEMENTARY APPENDIX REGRESSION AND PICTURE)**. However, while strictrly speakign there we cannot consider Quality a numerical variable, for the purpose of certain examples at a later point, this variable will be considered as both a categorical and numerical variable (no inference will be made is used as numerical). Finally, Lot Area, will be used as a confounder in the regressions to control for the association larger lot sizes creating larger houses (**AS AN INTERACTION**).

Table 1: Descriptive Statistics of Numeric Varaibles

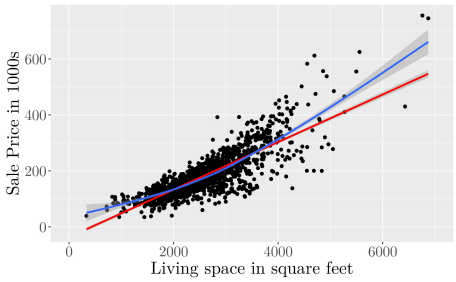
Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
SalePrice	180.921	79.443	34.900	129.975	163.000	214.000	755.000
Lot Area	10,516.830	9,981.265	1,300	7,553.5	9,478.5	11,601.5	215,245
Quality	6.099	1.383	1	5	6	7	10
Condition	5.575	1.113	1	5	5	6	9
Total Living Space	2,572.893	823.598	334	2,014	2,479	3,008.5	11,752
Years Since Remodeling	22.950	20.641	0	4	14	41	60

Notes: N = 1460. OLS estimates, robust standard errors in parentheses.*** p<0.01, ** p<0.05, * p<0.1

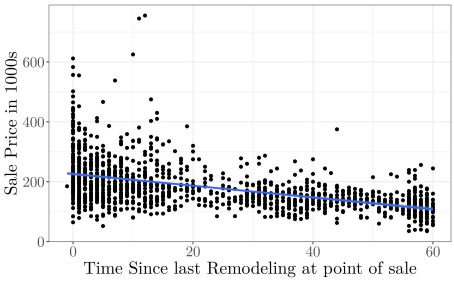
Beyond this table, there are multiple categorical variables **(see Appendix)**, such as Zoning and Year of Sale. The original (MS)Zoning variable contains seven categories, of which five contain data; these zones correspond to the administrative classification of the ground on which the properties are constructed (Commercial, Floating Village, Low-Density, Moderate-Density, High-Density contain data; Residential Low Density Park, Agriculutural, Industrial do not contain records). For the purpose of this analysis, this number was reduced to four categories based on the similar behaviour of Moderate and High Density properties **(SEE SUPLEMENTARY APPENDIX PLOT)** in the data as Ames, Iowa, represents the steretypical picture of a mid-western town in the US, thereby displaying fewer densly populated areas. Thus, the main question of this analysis section focuses on the difference between Low and higher density properties.[1] In addition, year of sale will be used to control for the effect of the 2008/2009 housing cricis.

Finally, the plots are to be considered in the context of the hypothesis in the next section.

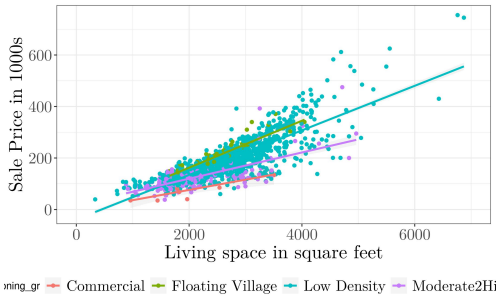
[1] Floating Village and Commercial behave too differently to be merged



(a) Positive association of total Living Space & Sale Price & optimal line



(b) Positive association of total Living Space & Sale Price



(c) Positive association of total Living Space & Sale Price; range [0, 7000]

Figure 1: Three Hypothesis Graphs displaying their repective association with the outcome variable

2 Theoretical model and OLS assumptions

2.1 Hypotheses

Based on the plots generated during the EDA, a mini theory was created to explain the variation in the sales price of properties (Figure 2).

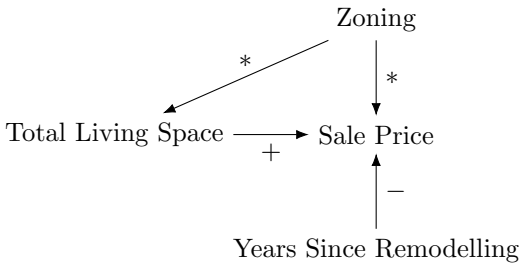


FIGURE 2: Causal relationship Scheme

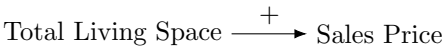
Based on this model, the following three hypothesis were created.

2.1.1 Hypotheses 1

Figure 1A displays a potential direct positive association between Total Living Space (IV) and Sale Price (DV), including an optimal fit, showing a small. Thus, one expects that larger houses have a higher sale price. Consequently,we assume that:

Hypothesis 1 (H1): *Total living space (IV) has a direct postive association with Sales Price (DV)*

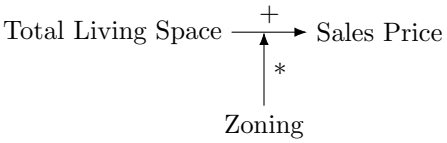
Figure 3: Hypothesis 1



Subsequently, when taking the Zoning (MSZoning or MSZoning_grouped - IV) into account to reflect the administrative borderrs of the Ames districts, larger houses in more densly populated areas of the city appear to have a lower price when compared to houses of same size in less densly populated areas as can be seen in **Plot XXXX2**. This suggests a separation between "downtown less affluent areas" and "suburban affluent areas"[2] and concludes in the hypothesis that

Hypothesis 2 (H2): *Zoning moderates (MIV) the direct postive association of Total Living Space (IV) and Sale Price (DV). The association of Space and Sale Price is proposed to be weaker for more densly populated areas than for more rural areas.*

Figure 4: Hypothesis 2



As can be observed on Figure 1c, older the houses 8not renovated (IV), are associated with lower sales prices (DV). Thus, the final hypothesis corresponds to:

Hypothesis 3 (H3): *Years Since Remodeling (M-IV) or Construction has an amplyfiying effect on the direct postive effect of Quality (IV) and Sale Price (DV)*

Figure 5: Hypothesis 3

IMPORTANT: DO YOU INCLUDE A QUADRATIC TERM IN THE POPULATION REGRESSION MODEL???? NO! BUt make more equations explaining each quadartic term and represent the interaction effects etc adn what not

(2-1)

$$SalePrice = \alpha + \beta_1 TotalLivingSpace - \beta_2 Zoning - \beta_3 YearsSinceRemodeling + \epsilon$$

(2-1)

2.2 Assumptions

A1: The linearity of model parameters and error term assumption suggests that a) the functional form of the underlying population regression model is linear and additive. Thus, this assumption is generally assumed to hold, i.a. not having introduced quadratic or polynominal (by parameter) terms into the population regression equation (2-1). However as Figure 1b shows, there might be a quadratic relationship present between Total Living Area and Sale Price. The reason why this might be the case may be that with increasion values for x ([X = x]), the effect of X on Y

[2]As an extention we will test whether the groups of Neighborhoods generally stay in the same zoning category; if Neighborhoods and Zoning are not related (so eg 50 % of one neighborhood is in rural zone while the other part is in moderately populated zone) then we have a problem that this woudl induce a bias. Otherwise, we can just proceed.

increases, which would be represented in later regression models by an additional quadratic term in the regression equation. However, as will be shown in later parts, this can be (partially) remedied.

A2: Full rank assumes that no independent variable can be a linear function of other independent variables; e.g.a 2-dimensional sphere might be compressed to a line, figuratively speaking, meaning that there is no optimal (or none at all) solution for the parameters in question. This first part of the assumtion might easily be violated when not dropping a "comparative" category of a categorical varaible. However, violations of full rank might also come in the form of multicollinearity. Multicollinearity is the "almost" violation of the full rank violation as a given variable may, for isntance, be highly correlated with another given variable. This way the sphere is almost compressed to a single line, which leads to a multitute of problems regarding inference of the model: primairly, the standard errors of the model become unstable to the inclusion of other explanatory or control variables. Multicollinearity is almost always present with observational data (actually making regression interesting in the first place); however, the extend is more relevant. For example might be if we included the Total Linving Space and the TotalNumberofRooms, both of which will be strongly correlated. Further tests will be conducted to probe for this assumption violation.

in what direction is the estimator baised when this variable is left out

A3: This assumption is referred to as mean independence (or exogeniety); assuming no correlation/systematic association between independent variables and residuals (or conditional mean error is dependent on independent variables). Thus this assumption might induce a bias/ or reduce consistency of the estimator and is, thus, critical to the model itself. A common way, this variable is violated is via unobserved confounders or ommitted variables. If a given variable is left out, part of the error term can be explained by a given variable which suffers under the influence of the confounder. An example in this case will be the missign information regariding crime by the area a property is located. It is obvious that higher crime levels will reduce the value in a given area. Thus, given a certain Neighborhood in question, part of the variance that would have been explained by the Crime variable is now falsely attributed to Neighborhood, biasing the estimate. Omitted variables bias the estimate of the population coefficient either up or down, thereby reducing the consistency of the estimate.

A4: The error term has a constant variance for each observation expected! –¿ heteroscedasticity

The assumption of homoscedasticity assumes constance variance of each observation. However, if the variance of the estimate is varying we face heteroscedastic variances. While this is not necessarily problematic regarding the estimated parameter (coefficient), heteroscedastictiy impacts the standard errors of the estimate, leading to problems regarding inference; heteroscedastitiy can lead to both type I and II error. In this research, this violation might occur if the eg. sales prices vary stronger for large houses than for small houses. The resulting (averaged) standard error would neither be rebresenative for small and large house standard errors and, thus, inference itself.

NOTE AUTOCORRELATION AND YEARS SINCE BUILT

Generally, as the sample size increases, this assumption becomes less important for the estimate itself, considering that OLS is a consistent estimator; which is the case here (N is quite large). However, heteroscedastictiy may lead to problems regarding inference, as the standard error of the estimate may still be biased. As such, solutions will be later implemented to control for such violations.

A5: Data generation: xi can be random or fixed; the data was collected for predictive purposes so we might not be able to veryfy this. As the data was not generated via a random experiment,none of the variables at hand are **random variables** Due to the data originating from sales in a town in Iowa (Ames), we have to assume that the data at ahnd is "fixed". Thus, inference regardign a wider population cannot be made from this data, as its fixed nature only applies to small towns in the mid-western USA. However, we can assume that the fixed variables collected are measured without error; particualrly as the sample collected can be somewhat representative of the wider popualtion of small towns in the mid-western USA.

generally, due to the nature of the data, being a

in order to

A6: This assumption regards inference; if the residuals do not follow a standard normal distribution, this might result in incorrect decisions as the distribution might for instance be "fatter" in the tails, resulting in potentially Type I or Type II errors. This might be the case if for instance we do not have a lot of observations for e.g. subcategories. Subsequently, the resulting inference

might be biased. However, as the sample size increases, most distributions approach the normal form. However, in order to answer the question, this assumtion is violated