

1 Data Prep, EDA, and Theory development

1.1 Variable Selection & Explanation

For the purpose of analyzing the determinants of house (sales) prices in the US a theory was developed based on combining two common valuation strategies in real-estate; "vergleichswert verfahren, sachwert verfahren".

Based on the theory, four categories of regressors were identified in the data, promising to best represent the population regression equation.

- Size related quantities (House size, lot size); garage
- District and Neighborhood dependent variables; Zoning
- type of housing (one family home, apartment, etc)
- Quality and condition of the house; including time since remodeling

To this end, this research assignment draws data from an appraisal project conducted by... in the Ames district, Iowa (USA) (CITE DATA SOURCE HERE!!!). Corresponding to the aforementioned data categories, the following variables were selected to be used in varying degrees in the model. It may be noted, that due to the data being limited to a city in the midwest of the USA, the generalization resulting from this research may only extend to similar cities. However, due to the data originating from one district alone results in the comparability of the sale instances recorded in the data; meaning that stark contrasts in sales prices may be less due to the simple fact that one sale may have been made in Iowa and the other in New York, which naturally yields higher prices.

To start, a total of 1,460 house sales were recorded between 2006 and 2010 for the district of Ames, Iowa (USA). The dependent variable was identified to be SalePrice. As can be observed in Table 1, the mean sale price of a house was \$180,921.20 (SD = 79,442.50). Combined with the range [34,900, 755,000] a positive skewness was to be expected (skew = 1.881), considering that the outcome variable is of financial nature.

Following, the first category of data pertains to size related dimensions of the property sold. More specifically, the total living area (tot_living_area)¹ displays a mean of 2,572.89 square feet (SD = 823.598) in addition to a large range of values [334, 11,752]; suggesting that the sales were conducted in neighborhoods (Neighborhood) included range from urban to (partially) rural. To this end, the second data category encompasses the zoning classification (MSZoning) which identifies neighborhoods and the corresponding sales as rural or not. Neighborhood consists of 25 distinctions and zoning of eight categories², which will be adjusted to three categories to decrease the complexity of the data analysis (see Appendix for a contingency table). Additionally, the number of bedrooms

¹defined as summing above- and below- ground or base living.

²only 5 categories actually contain data.

above ground level (mean = 2.866, SD = 0.816) (BedroomAbvGr) and the number of bathrooms (mean = 1.990, SD = 0.732) are included (tot_bathrooms)³.

Moreover, the third class of data was selected to balance size and neighborhood related associations by considering building type (BldgType), which consists of five categories. Interestingly, the majority of sold homes were one-family homes (n = 1220); this variable was adjusted to reduce the complexity of the data analysis and remove confusion about the definition of building type.

The fourth category contains quality and condition related variables. Both variables are on a discrete scale from one to ten. Quality displays values for each quality rating (mean = 6.099, SD = 1.383), while the condition ranges from one to nine (mean = 5.575, SD = 1.113).

TIME SINCE REMODELING AT YEAR OF SALE; Time since building is also negative but this one is clearer!; ALSO THIS RELATIONSHIP HOLDS FOR NEIGHBORHOODS!!!!

Table 1: Descriptive Statistics

| Statistic | N | Mean | St. Dev. | Min | Max |
|------------------------|-------|-------------|------------|--------|---------|
| SalePrice | 1,460 | 180,921.200 | 79,442.500 | 34,900 | 755,000 |
| YearBuilt | 1,460 | 1,971.268 | 30.203 | 1,872 | 2,010 |
| YearRemodAdd | 1,460 | 1,984.866 | 20.645 | 1,950 | 2,010 |
| LotArea | 1,460 | 10,516.830 | 9,981.265 | 1,300 | 215,245 |
| GrLivArea | 1,460 | 1,515.464 | 525.480 | 334 | 5,642 |
| TotalBsmtSF | 1,460 | 1,057.429 | 438.705 | 0 | 6,110 |
| BedroomAbvGr | 1,460 | 2.866 | 0.816 | 0 | 8 |
| BsmtFullBath | 1,460 | 0.425 | 0.519 | 0 | 3 |
| FullBath | 1,460 | 1.565 | 0.551 | 0 | 3 |
| PoolArea | 1,460 | 0.005 | 0.069 | 0 | 1 |
| GarageCars | 1,460 | 1.767 | 0.747 | 0 | 4 |
| OverallQual | 1,460 | 6.099 | 1.383 | 1 | 10 |
| OverallCond | 1,460 | 5.575 | 1.113 | 1 | 9 |
| tot_living_area | 1,460 | 2,572.893 | 823.598 | 334 | 11,752 |
| tot_bathrooms | 1,460 | 1.990 | 0.732 | 0 | 6 |
| Adjacent_features_bool | 1,460 | 0.137 | 0.344 | 0 | 1 |

It is notable that upon selecting the aforementioned variables, no preprocessing in the form of imputation or data deletion had to be applied. However, in order to decrease the complexity of the interaction term, the variable MSZoning was binned into (rural, mixed rural, and urban) based on the corresponding zoning categories⁴.

³The correlation between house size and number of bedrooms and bathrooms will be addressed later

⁴<https://www.kaggle.com/competitions/home-data-for-ml-course/data>

1.2 Exploratory Data Analysis

⁵ A scatter plot matrix is provided in the appendix for the numeric variables. The EDA shows the four data categories with respect to the sale price of the property. Thus, the plots represented, inter alia, drove the development of the hypothesis down the line in combination with the aforementioned valuation strategies.

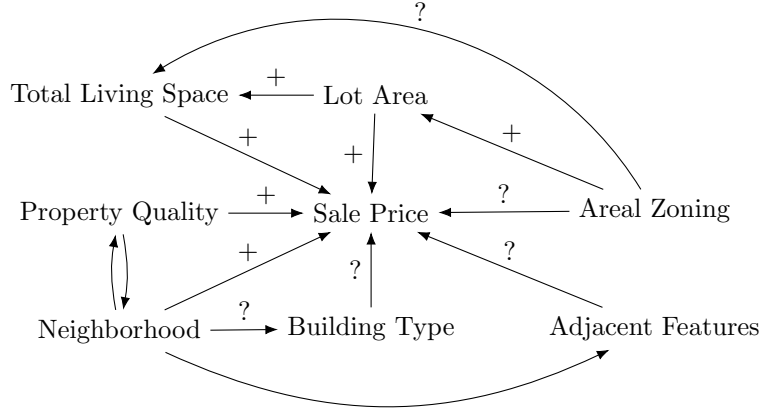
first analysis: Size of living area to sales price with respect to building type

Second analysis total rooms and sales price with respect to size zoning

Third analysis quality and condition to sales price

2 Theoretical model and OLS assumptions

Plot 1: Causal relationship Scheme

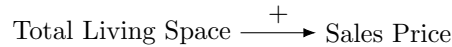


Based on the valuation strategies and EDA discussed above, a theory was set up to explain the variation in sales prices. The corresponding causal relationship scheme can be seen in Plot 1.

Plot XXX displays a potential direct positive association between Total Living Space (IV) and Sale Price (DV). 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 positive association with Sales Price (DV)*

Plot 2: Hypothesis 1

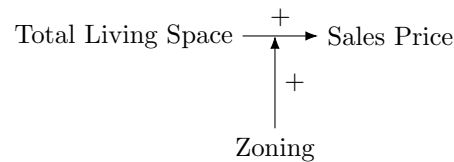


⁵The theory underlying the choice of the variables will be further elaborated upon in Section 2

Subsequently, when taking the Zoning (MSZoning or MSZoning_grouped - IV) into account to reflect the administrative borders of the Ames districts, larger houses in more densely populated areas of the city appear to have a lower price when compared to houses of same size in less densely populated areas as can be seen in **Plot XXXX2**. This suggests a separation between "downtown less affluent areas" and "suburban affluent areas"⁶ and concludes in the hypothesis that

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

Plot 3: Hypothesis 2



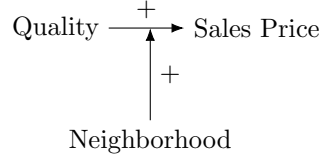
As the previous hypothesis suggested, location may be an important factor in explaining the variance in Sale Price. The assumption that affluent houses tend to cluster in affluent neighborhoods is, thus, close. We can thus assume that affluent neighborhood tend to pose the best qualitative properties. As was discovered during the EDA in part 1, Quality (IV) has a positive association with Sale Price (DV). Subsequently, this research investigates to what extent this clustering differs for the different Neighborhood(s) (IV). **However, due to Neighborhood not being a suitable variable for this analysis, Zoning again is used as a proxy.**

Hypothesis 3 (H3): *Neighborhood of a property moderates the relationship of Quality on Sales Price. Neighborhoods tend to cluster and display similar patterns observed.*⁷

⁶As an extension 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 would induce a bias. Otherwise, we can just proceed.

⁷This hypothesis is partially meant to show the problematic of clustering in the data; this will be further investigated in part 3 & 4 of this research assignment

Plot 4: Hypothesis 3



FURTHERMORE, a possible fourth hypothesis pertains to the capital that has to be invested upon buying the house; such as remodeling etc. This value is assumed to be larger for houses that have not been remodeled for along time before being sold. As such, the fourth hypothesis states that (ADDITIONALLY, we assume this relationship to hold for neighborhoods as well! so neighborhoods do not have an impact here)

Hypothesis 4 (H4): *Years Since Last Remodelling (IV) has a negative association with Sale Price (DV) IMPORTANT! ZONING MAY BE AN ALTERNATIVE TO NEIGHBORHOOD HERE WHICH WOULD MAKE THE ANALYSIS WAY EASIER IN GENERAL*

Plot 4: Hypothesis 3



$$\begin{aligned} SalePrice = & \alpha + \beta_1 TotLivingSpace - \beta_2 TotLivingSpace^2 + \beta_3 Quality + \\ & \beta_4 Zoning_{LowDensity} + \beta_5 Zoning_{Other} + \beta_6 YearsSinceRemodeling + \\ & \beta_7 LotArea + \beta_8 TotLivingSpace * Zoning_{LowDensity} + \beta_9 TotLivingSpace * \\ & Zoning_{Other} + \epsilon \end{aligned}$$

IMPORTANT: DO YOU INCLUDE A QUADRATIC TERM IN THE POPULATION REGRESSION MODEL????

$$\begin{aligned} SalePrice = & \alpha + \beta_1 TotLivingSpace - \beta_2 TotLivingSpace^2 + \beta_3 Quality + \\ & \beta_4 Zoning - \beta_5 YearsSinceRemodeling + \beta_6 LotArea + \beta_7 TotLivingSpace * \\ & Zoning + \beta_8 AdjacentFeatures + \beta_9 BuildingType + \epsilon \end{aligned}$$

Assumptions for Linear Regression Considering the population regression model above, the first assumption requires a linear relationship between the independent and the dependent variables. As could be observed during the EDA in part 1, this assumption is likely to hold, with a weak nonlinear relationship of Total Living Area and Sale Price; this can be rectified using a quadratic term.

<https://statisticsbyjim.com/regression/ols-linear-regression-assumptions/>

A1: Linearity of model parameters and error term; through the formulation of the (population) regression model, addressing the models' functional form.

Thus, a violation of this assumption is practically impossible if the functional form of the population regression is defined correctly.

A2: describes that for a linear system to be solvable, no independent variable can be a linear function of other independent variables. This implies that for dummy variables to have one "comparative" category to be dropped. Furthermore, the "almost" failure of this assumption, multicollinearity also brings problems with it. While some multicollinearity is to be expected between the coefficients; strong multicollinearity leads to a lower precision in the estimate. Further tests will be conducted to probe for this assumption violation.

A3: This assumption is referred to as mean independence (or exogeneity); assuming no correlation/systematic association between independent variables and residuals (or conditional mean error is dependent on independent variables). However, as could be seen during the EDA in part 1, the linearity assumption holds for most of the variables we have observed except for a weak quadratic relationship. Thus this assumption might induce a bias or reduce consistency of the estimator; this is why this assumption must be treated with care. But it is also difficult to be found as omitted variable bias is very likely to have occurred during this issue; one example might be crime which has an impact on whether a certain neighborhood is affluent or not.

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

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, heteroscedasticity 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: ϵ_i can be random or fixed

A6: Normal distribution of disturbance (The error term is normally distributed. generally, the means that the error term has a population mean of zero and standard error of 1. Generally, most distributions tend to approach normal distributions as sample sizes increase. For some distributions such as Sales Price (which is of a financial nature), this can also be remedied through transforming (here apply log transform). Additionally, multiple controls and tests will be conducted down the line to investigate this issue.

3 OLS regression and model fit