1 Data Prep, EDA, and Theory development

1.1 Varaible Selection & Explanation

For the purpose of analyzing the determinats 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 reasearch assignemnt draws data from an apprisal 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 varibale 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 a 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 reange 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 correpsonding sales as rural or not. Neighborhood consisists of 25 distinctions and zoning of eight categories[2], which will be adjusted to three categories to decrease the complexity of the data analysis (see Appendix for a contignecy table). Additionally, the number of bedrooms

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

[2] only 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 consideringh 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).

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
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

It is notable that upon selecting the aforementioned variables, no preprossessing in the form of imputation or data deletion had to be applied. However, in order to decrease the complecity of the interaction term, the variable MSZoning was binned into (rural, mixed rural, and urban) based on the corresponding zoning categories[4].

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

[3]The correlation betwen house size and number of bedrooms and bathrooms will be addressed later

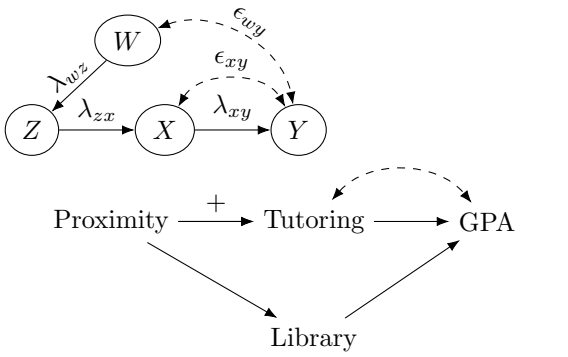
[4]$https://www.kaggle.com/competitions/home-data-for-ml-course/data$

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

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



Total Living Space Property Quality

Neighborhood Sale Price Areal Zoning

Property Quality Building Type Adjacent Features

The first category of data pertains to the size dimension of the house in question. Here the size of the house (1stFlrSF + 2ndFlrSF), combined with further information regarding the of bathrooms (BsmtFullBath + FullBath), Bedrooms (Bedroom), and Kitchen explain the overall living space of a house or appartment. Additionally, the lot size (LotArea) of the property and the garage (GarageCars/GarageArea) specification complement this category.

The Second category describes neighborhood and location related characteristics of the property. School districts (indirectly included as unobservable!!!) and afluent neighborhoods naturally have a large impact on the saleprice. Additionally, neighborhoods may, thus, function as cluster correction for similarities in the error term when correcting for heteroscedasticity; assuming that sales in the same neighborhood share similar underlying variation. Finally, the neighborhood may control for the size of the house; we would generally assume that big houses are more expensive. However, if we consider New York downtown, to large houses in the country side of Iowa, small flats (in new york) might induce that small properties cost more than large properties.

The third category of data distinguishes in the type of housing that was recorded as a sale in the data (MSSubClass, BldgType).

The fourth category focuses on the quality and condition aspect of the property as a function of years since remodeling (YearRemodAdd), building finalize date (YearBuilt), and the overall rating of quality (OverallQual) and condition (OverallCond) of the property.

Note, variables such as central heating have been discarded as those usually are contained in the type of house (appartments tend to have central administered heating from the overall building).

MAKE A CAUSAL SCHEME!!!! HErE IN LATEX