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. **INCLUDE FREQQUENCY TABLES FOR THE CATEGORICAL VAIRABLES**

Preanalysis: is quality as an interval? IMPORTANT: ARGUE WHY WE THING THAT QUALITY IS AN INTERVAL HERE LIKE LIKERT SCALE?!?!!?! 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. **IMPORTANT! TELL THAT WE HAVE LARGE RANGES ETC SO THIS MEAN WE NEED STANDARDIZED COEF**

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

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

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 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).

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

[2]only 5 categories actually contain data.

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

KEEP STAT PART IN STARGAZER USE QUANTILES IN THE SUMMARY STATISTIC; ANd remove the N; ALSO ADD DESCRIPTIVES

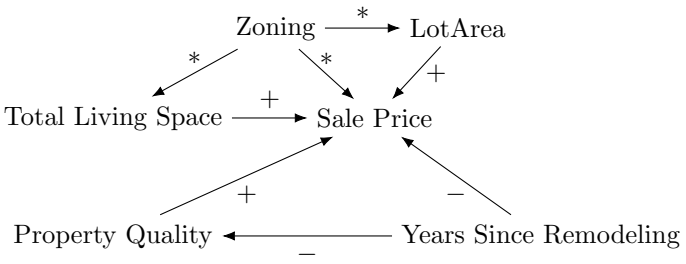
It is notable that upon selecting the aforementioned variables, no preprossessing in the form of imputation or data deletion had to be applied. However, in order to decrease the complecity of the interaction term, the variable MSZoning was binned into (rural, mixed rural, and urban) based on the corresponding 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 property. Thus, the plots represented, inter alia, drove the development of the hypothesis down the line in combination with the aforementioned valuation strategies.

2 Theoretical model and OLS assumptions

Plot 1: Causal relationship Scheme



$$SalePrice = \alpha + \beta_1 TotalLivingSpace - \beta_2 PropertyQuality - \beta_3 YearsSinceRemodeling + \beta_4 LotArea + \beta_5 Zonin$$

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.

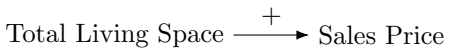
Plot XXX displays a potential direct positive association between Total Living Space (IV) and Sale Price (DV). Thus, one expects that larger houses have a higher sale price. Consequently,we assume that:

[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

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

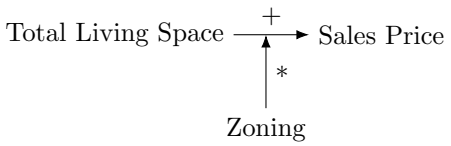
Plot 2: Hypothesis 1



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"[6] 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.*

Plot 3: Hypothesis 2



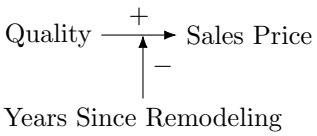
Finally, the sales price is not only dependent on the size and the location of the construction, but also its Quality (IV). Correspondingly, one might assume a positive relationship between the quality of a given property and the Years Since Remodeling (M-IV) or Construction. Thus, the final hypothesis states that[7]

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)*

[6] 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.

[7] Note that in order for this hypothesis to work, an initial association of Years Since Remodeling would have to be included in previous Hypotheses

Plot 4: 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

Assumtions for Linear Regression Considering the population regression model above, the first assumptions requires a linear relationship between the independent and the dependent variables. As could be observed during the EDA in part 1, this assumtion is likely to hold, with a weak nonliear relationship of Total Living Area and Sale Price; this can be rectivfied 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 assumtion is practically impossible if the functional for of the population regression is defined correctly. WHeterh the relationship between the indent and the dependent is of linear nature. 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. Additionally, the explanatory variables have to be linearly associated with the dependent. This assumption is easily violated. As can be seen in **plot xxx and the corresponding appendix with the best fitted line**, Total Living Space displayes a slight quadratic relationship with the 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 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: desribes that for a linear system to be solvable, no independent variable can be a linear function of other independent variables; e.g. this way 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. This assumption is commonly to some extend violated, as some multicollinearity is always present between the explanatory variables (actually making regression interesting in the first place). As such, a good 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.

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 bais/ 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 neighborhood. 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.

A4: The error term has a constant variance for each observation expected! –¿ heteroscedasticity The assumption of homoscedasticity suggests 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.

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.

DISUSS WHAT CAN GO WRONG APPLIED TO MY DATA AND MY MODEL like last lecture in eci; Come up with a situation where this is violated; EG FULL RANK: NUMBER OF BEDROOMS AND TOTAL LIVING SPACE; HOUSE SIZE AND TOTAL LOT SIZE can pose multicollinearity problems

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.

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.

Normal distribution of disturbanc (The error term is normally distributed. generally, the means that the error term has a population mean of zeor and standard errror of 1. Generally, most distributions tend to approach normal distributions as sample sizes increase. For some distirbutions such as Sales Price (which is of a financial nature), this can also be remedied through tra´nsforming (here apply log transform). Additionally, multiple controls and test will be conducted down the line to investiate this issue.

3 OLS regression and model fit

The study of Sales Price of property in Ames, Iowa, has shown that the Total Living Area has a significant postive effect on the Sales Price ($\hat{\beta} = 0.040$, $p = 0.01$)

Effect Size regarding teh discussion which variable has the largest effect sized, the standardized coefficietns were used. This is primarily due to the large ranges in some varaibles such as total living area [334, 11,752] when compared to Quality [1,10]. This will cause the OLS to overstate the the cofificent with the large range in values, amking it appear to be greater than it is. Thus, to this end we use standardized coefficients

4 task 5

TIPP **run regression once with and once without the dummies to see the effect of the subsample**

Table 2:

	Dependent variable:	
	SalePrice	
	(1)	(2)
tot_living_area	0.040***	0.056***
	(0.002)	(0.004)
I(tot_living_areaˆ2)		−0.00000***
		(0.00000)
OverallQual	24.614***	31.638***
	(1.120)	(1.266)
low_density_zone	17.558***	−23.347***
	(2.896)	(9.011)
other_zone	11.981**	−63.223***
	(5.284)	(20.282)
time_since_remodeling	−0.460***	2.274***
	(0.060)	(0.219)
LotArea	0.001***	0.001***
	(0.0001)	(0.0001)
OverallQual:time_since_remodeling		−0.487***
		(0.038)
tot_living_area:low_density_zone		0.017***
		(0.004)
tot_living_area:other_zone		0.028***
		(0.008)
Constant	−81.443***	−137.445***
	(6.536)	(10.840)
Observations	1,460	1,460
R^2	0.763	0.799
Adjusted R^2	0.762	0.798
Residual Std. Error	38.725 (df = 1453)	35.744 (df = 1449)
F Statistic	781.201*** (df = 6; 1453)	575.803*** (df = 10; 1449)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 3:

	Dependent variable:
	ln_SalePrice
tot_liv_area	0.0004
	(0.00002)
I(tot_liv_areaˆ2)	-0.00000^{***}
	(0.000)
OverallQual	0.124
	(0.006)
low_density_zone	0.105
	(0.039)
other_zone	-0.344^{***}
	(0.089)
y_since_rem	-0.0004^{***}
	(0.001)
LotArea	0.00000
	(0.00000)
OverallQual:y_since_rem	-0.001^{***}
	(0.0002)
tot_liv_area:low_density_zone	0.00001
	(0.00002)
tot_liv_area:other_zone	0.0002
	(0.00004)
Constant	3.515
	(0.047)
Observations	1,460
R^2	0.848
Adjusted R^2	0.847
Residual Std. Error	0.156 (df = 1449)
F Statistic	807.324^{***} (df = 10; 1449)
Note:	*p<0.1; **p<0.05; ***p<0.01