1 Data Prep, EDA, and Theory development

1.1 Varaible Selection:

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".

** Elaborate here on the theory **

Based on the theory, four categories of data were identified 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

WHICH VARIABLE IS THE MAIN REGRESSOR? WHICH MAIN REGRESSORS'????

To this end, this reasearch assignemnt draws data from an apprisal project conducted by... in the Ames district, Iowa (USA) (CITE THE INITAL DATA SET ADN TH CREATOR* LINK 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.

The outcome or dependent variable is SalePrice; (MoSold) Month Sold and (YrSold) Year Sold complement this information.

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