* Dealing with NAs - ~300
* Influential points

**Missing values:**

* Veneer type NA – replace with dummy variable ‘no’
* Veneer area NA – use mice for imputation – same for lot frontage
* Garage year built – remove NAs
  + Substituting with a value makes no sense – so, no imputation or dummy value replacement
* Utilities: predictor is removed because it had a single level
* ID: column is removed
* Step 1: Imputation required for:
  + Veneer area and lot frontage
* Step 2:
  + Standardize the betas
  + Lasso for variable selection
* Step 3: Reduce design matrix and fit OLS model
* Step 4: Calculate VIF (r function to check collinearity)
  + If collinear, either drop one/multiple variables OR combine the variables (Pg 115)
    - Need to figure out how we find collinear variables
  + If not collinear, stick with OLS from Step 3.

Notes from Oct 2 lecture

* **Tests for Normality:** This is important because otherwise you cannot do statistical tests for inference on betas.
  + People are running into many issues with this
  + Testing for outliers requires the normality assumption – this means we should *test for normality first*.
  + Be consistent and use the same normality test OR use multiple tests and discuss rigorously why their results differ.
* If you split a categorical variable and post-Lasso some of its levels go down to zero, it’s okay to use the original variable retaining only the levels that weren’t zeroed.
* Generalized linear model – allows non-constant variance. JW said it is acceptable to test for variance and use a glm for the case study.

---KK notes

1. Take cleaned data and fit OLS model. OLS model is trained on ALL data.
2. Test for normality
3. Test for outliers

* Start with test for normality
* If normal, check test for outliers

--Alvira notes

<https://stats.stackexchange.com/questions/117339/how-to-ensure-that-the-most-appropriate-value-for-lambda-is-chosen-in-lasso>