**Hypothesis Building**

**Numeric Variables:**

There are 36 relevant numerical features. MSSubClass, which "identifies the type of dwelling involved in the sale", is encoded as numeric but is in reality a categorical variable.

There are 36 numerical features, of the following types:

* Square footage: Indicates the square footage of certain features, i.e. 1stFlrSF (First floor square footage) and GarageArea (Size of garage in square feet).
* Time: Time related variables like when the home was built or sold.
* Room and amenties: data that represent amenties like “How many bathrooms?”
* Condition and quality: Subjective variables rated from 1–10.

Most of the variables that deal with the actual physical space of the apartment are positively skewed — which makes sense, as people tend to live in smaller homes/apartments apart from the extremely wealthy.

Sale Price also has a similar positively skewed distribution — I hypothesize that the variables dealing with the actual dimensions of the apartment have a large impact on Sale Price.

Many features do not have a strong relationship with Sale Price, such as ‘Year Sold’. However, a few variables, like overall quality and lot square footage are highly correlated with Sale Price.

**Categorical Variables**

Similar to the numeric features, there is a range of categorical features. While many look like the sale price varies with category, there are many that don’t. Let’s identify a few features that affect value. Some include the presence or absence of central air, the neighborhood, the external quality, and the zoning.

There are also features that don’t vary in price a lot among different categories, including the roof style and land slope.

**Hypothesis Statement:**

 “ the variables dealing with the actual dimensions of the apartment have a large impact on Sale Price ”