**Step 1: Collect data**

Before collecting data, figure out the following questions:

(Brainstorm) What attributes affect the dependent variable?

Are those attributes observable and measurable?

*Example ☺* we know the factors affecting a property’s price include location, how big the property is, number of bedrooms, number of bathrooms, property type. All theses factors are either quantitative variables or categorical variables. We can observe what these factors are for each property listing on Zillow.com.

*Example which is risky as a course project* people are interested in studying coronavirus. However, this topic can be too complicated for a course project, given limited tools and time we have before the end of semester. The risks are so many unknowns on this virus---how it transmits exactly, how long it survives in different environments---these factors thus are not observable. Chances are the result of statistical analysis may not be significant. Therefore, I do NOT recommend a course project like this.

*Two methods to collect data*

1. fill in a spreadsheet of a reasonable amount of observations (this is the case we show in class---the property market data in Austin)

2. download a dataset online directly.

**Step 2: Clean data**

- missing data: simple delete

- outliers: identify and delete

- visualize and check if association exist

**Step 3: Pre-process**

- categorical variables

- transform for linear association

**Step 4: Analysis**

- run regression model

- interpret results

- evaluate the model (critical thinking)