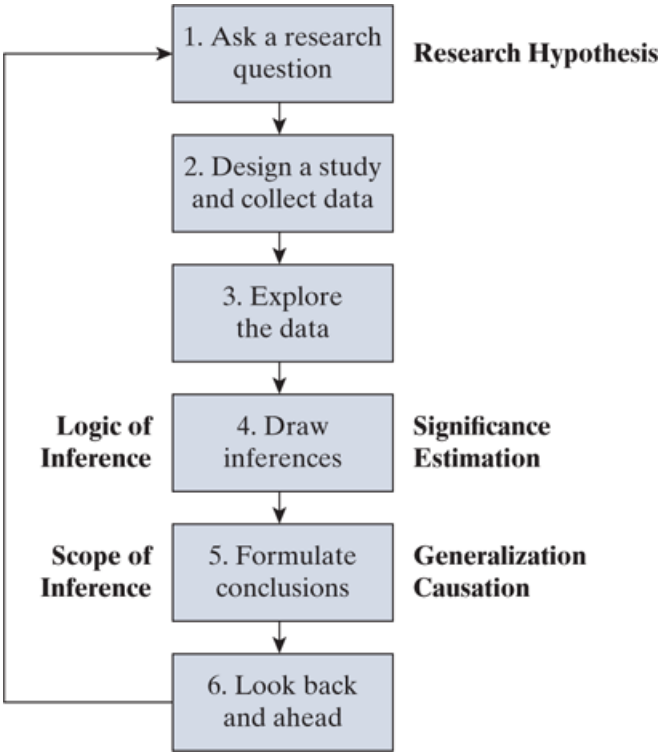
MA206, Lesson 2 - Preliminaries

Understand the 6-Step Process



Understand how to visually interpret the data

Shape

We want to look if the distribution is symmetric, where it is centered, and how many peaks there are.

Center

Where is the distribution centered on? What is a typical value?

Variability

How spread out or concentrated is the data? We often report this variation through **standard deviation**, which can be crudely explained as the average distance of our data from its mean

Unusual Observations

Are there **outliers** which are very different from most observations? Can this difference be explained? Could it be entered in error?

Looking forward, we want to look at a housing market to buy our house to live in. This could be your first duty station near a military base, or your retirement home in the perhaps not so distant future. You want to buy, but are unsure what you should spend. Think of a location and explore using data.

1: Ask a Research Question

What is your research question?

2: Design a Study and Collect Data

We will utilize redfin.com to collect the data from our analysis. Download the .csv file for your area and save it into your working directory to load it into R.

What are the **observational units**?

Give some examples of notable **variables of interest** and categorize them as either quantitative or categorical.

3: Explore the Data

Describe the **distribution** of the prices of the houses in your area of interest

Shape

Center

Variability

Unusual Observations

What is the mean and standard deviation of the housing prices in your area?

Conduct further data exploration and annotate any interesting findings or charts.

4: Draw Inferences

We will get into **Draw Inferences** throughout the course to build tools to indicate if there is an effect between variables. For example, we might provide evidence that as a home's square footage increases, the sales price increases. What else might we wish to infer?

5: Formulate Conclusions

To what underlying process or larger group can these conclusions be generalized? Again, we will get further into **Formulate Conclusions** throughout the course, but here we would not apply the conclusions from our analysis outside the houses in the local area because that is all our **sample** is from. What generalizations or causations can we infer from our data?

6: Look Back and Ahead

We could identify limitations of our study, such as the limited region of houses looked at. For example, there could be interactions between the variables, another tool we will explore later in this course. For future research, there are probably other factors not captured in Redfin housing data that impacts housing costs - for example, as Clermont is in Lake County, a "water front property" status could impact prices in a way that is not captured in our analysis.