How to Formulate Good Research Questions for Meaningful Data Analysis

Learn how to bring good insights from the data

[[](https://towardsdatascience.com/@sucky00?source=post_page-----7bbb88bd546a--------------------------------)](https://towardsdatascience.com/@sucky00?source=post_page-----7bbb88bd546a--------------------------------)

[Rashida Nasrin Sucky](https://towardsdatascience.com/@sucky00?source=post_page-----7bbb88bd546a--------------------------------)

Follow

[May 28](https://towardsdatascience.com/how-to-formulate-good-research-question-for-data-analysis-7bbb88bd546a?source=post_page-----7bbb88bd546a--------------------------------) · 3 min read

Image for post



Source: Unsplash by [TAHA AJMI](https://unsplash.com/photos/dCAKIpxs3eE)

Nowadays, data are everywhere. There are so many different sizes and shapes of data. These data will not make any sense if nobody is analyzing them. So, Data Scientists and Data Analysts have a lot to do. But the analysis will not be effective if we do not know what information to infer from the data. No matter how authentic the data source is, how much money and time is spent to sample or collect data, that will be useless or even harmful if the research questions are not well formulated. I said harmful because if we make the wrong interpretation of a very sensitive dataset, that can lead to a wrong decision in an important issue. So, it is crucial for data scientists to learn, how to ask well-formulated research questions.

In this article, I want to explain some key elements that need to be considered to craft the research questions.

**The target population of interest**

A clear and concise description of the target population of interest has to be in the research question. A good example is a Hispanic population with a family income of more than 100,000 USD or more in the Chicago area. Or, African-American female with an education level of a college degree or higher.

**Descriptive or Analytic Question**

This is important to know what kind of calculations to perform. Here, descriptive means if there are specific parameters to calculate like the average age, gross income, or maximum height. On the other hand, analytic means a little more abstract parameters. It can be the relationship between income level and education. So, no straightforward calculation of a statistical parameter. Instead, the analysis of different parameters and finding conclusions from that.

**Is the question original**

Is the question original? Does it add a new area of knowledge or add more to any existing field or provide a different perspective about an old study? Many studies aim to build on some existing studies or add to the current knowledge. But that needs to be clear in the research statement or research question.

**Are the Data Already Available**

We need to make sure that the variables are either readily available or resources and tools are available to collect appropriate data. A careful study of the data has to be done so the data captures the core idea of the research.

These are the four essential components of a good research question. Without having a well thought out set of questions, data analysis can be meaningless or bring poor and misleading insights.

**A bad question**

“What is the relationship between income and the quality of health?”

Let’s analyze this question on our four key components. First, does it describe the target population precisely? It does not. So the area of research becomes too broad that may lead to an obscure analytic result. Second, Does it mention if the analysis is descriptive or analytic? Yes, it does. It is an analytic question. This question is asking to find the relationship between income and the quality of health. Third, is the question original? I guess not. But it depends on which part of the population. Forth, are the data already available for analysis? That depends on which part, area, and age group of population.

**A Well Crafted Question**

“What is the difference between the percentage of the White American population and the African American population who have diabetes in New York City, considering the age 50+.”

Now, let’s analyze if it covers all the key components. First, It describes the target population of the study clearly. Second, the question is descriptive. We need to calculate the percentage of White Americans and African Americans who have diabetes. Third, the question may not be original. But it will add new knowledge to existing knowledge. It may give scope to compare last year’s data to this year’s data. Forth, the data may be available already but there is scope to update the information. Most importantly, this question has a clear objective.

As a data scientist, you may know all the tools in the world. But learning data science is not complete without knowing how to infer the insights from the data properly.