Planning a Data Analysis Project for Business

Spring 2019

The aim of business research is to create insights that are both interesting and actionable. Accomplishing this aim requires careful planning, creativity, and critical thought. Fancy software and complex statistics are not a substitute for this planning process. Like any other creative process, the planning of a data analysis project for business purposes will require time, patience, and lots of self-critique and reflection. This document provides an outline to guide planning. The side notes elaborate on key concepts. The links take you to worksheets, diagrams, and examples to help you carry out these steps for your project. The steps are flexible enough to encompass many types of business research. Although an outline may convey a linear process, this is definitely iterative and should be referred to and updated throughout your project.

1. **Frame the problem** in a clear and concise problem statement.[1]
   (a) What business needs are unmet or could be addressed more effectively?
   (b) What potential threats or opportunities does the organization face?
   (c) Are there findings from prior research or anecdotal evidence that point to a need for more research?
   (d) Who is the "customer" or stakeholder that cares about the need, opportunity, threat, or prior findings you've identified?

[1] A problem statement should identify a gap in existing information that will impact the business and the key stakeholder(s) who care about it.

2. Articulate an **interesting**[2] and **actionable**[3] research question.
   (a) What business problem does the question address? Does it match with the symptoms you identified?
   (b) What problems does it leave unaddressed or out of scope?
   (c) What business decision does the question inform?
   (d) What is already known about the likely answer to this question?[4]
   (e) What is unknown that your analysis will address?
   (f) What data are available for answering this question?
   (g) What assumptions must be made to answer this question?

[2] By interesting, we mean of interest to the business audience your work is intended for. There should be a clear business purpose for needing an answer to the question

[3] By actionable, we mean a question that can be answered in a satisfactory manner with available resources. Note that the answer may still require some assumptions or be subject to some limitations

[4] This may require web research or a literature review

3. Formulate **research objectives**.[5]

[5] These are statements of what your work will accomplish. Each should begin with a verb. In consulting, these are often the statements that sell the work.

4. Articulate one or more **testable hypotheses**[6] and the key variables required to test it.

5. Select methods suited to testing identified hypotheses.

   (a) What sort of analytical story will be told to evaluate this hypothesis?

      i. Descriptive
      ii. Predictive
      iii. Diagnostic or Causal
      iv. Prescriptive

   (b) What research design is appropriate?

      i. Observational
      ii. Quasi-Experimental
      iii. Experimental
      iv. Case Study

   (c) What sources of data are available for carrying out this design and how can they be accessed? [7]

      i. Publicly available secondary data.
      ii. Administrative data owned by the organization.
      iii. Survey data collection is required.
      iv. Proprietary data that must be purchased.

   (d) What is the unit of analysis or grouping of the data?

      i. Cross-sectional
      ii. Time-series
      iii. Pooled cross-section
      iv. Panel

   (e) How are the key variables in this hypothesis measured?[8]

      i. Is there a reason to believe available measures will be biased?
      ii. If collecting primary data, how will key measures avoid potential biases?

   (f) What analytic tools will be used?

      i. Descriptive statistics
      ii. Visualizations
      iii. Uni-variate and bi-variate inferential statistics
      iv. Multivariate statistical models
      v. Mathematical models

6. Identify the study population and consider potential risks and biases.

[6] A hypothesis is not testable unless it can be wrong. The goal of your project is not to "prove" or "disprove" a particular hypothesis. Your goal is to evaluate the evidence for and against hypothesis that together inform the answer to your research question.

[7] It's important to consider costs of accessing and preparing data for analysis at this step as well. Collecting survey data, purchasing proprietary data, or preparing unstructured data for analysis can be very costly. Weigh these costs against the anticipated benefits of your research.

[8] This includes identifying the level of measurement (nominal, ordinal, interval or ratio) for each variable, and the answers to this question will dictate which specific analytic tools are appropriate.

   (a) Are human subjects involved in the research?

   (b) What risks might this project create?

   (c) How do the study subjects available for research differ from the ideal sample, and will these differences likely bias your results?

7. Plan for curating and communicating results and documentation of your analysis.

   (a) Who is the audience for this work?[9]

   (b) What format(s) is(are) requested for the final deliverable(s)? [10]

   (c) What set of possible figures and tables will convey an honest yet efficient summary of key findings appropriate for the audience?

   (d) What assumptions or limitations of the research design are most likely to impact the key findings and how can these impacts be clearly communicated to the audience?

   (e) How should additional findings and work required to arrive at key findings be documented? [11]

8. Reflect on assumptions and potential biases or limitations.

[9] It is wise to determine how much of the technical information about your analysis the audience will desire to see. Most business presentations will be to decision-makers who wish to see convincing high level findings but do not want the details behind the analysis process. However, you should always be prepared to provide the details when requested.

[10] You may choose to present the same information in a different way on a slide than in a written report to comply with space and time constraints.

[11] Planning ahead and completing technical appendices and annotations of workbooks and scripts as you work is far easier than returning to them after an analysis is finished.

Worksheet for Framing the Problem

This worksheet is adapted from Chapter 2 of **Keeping up with the Quants** *by Thomas H. Davenport.*

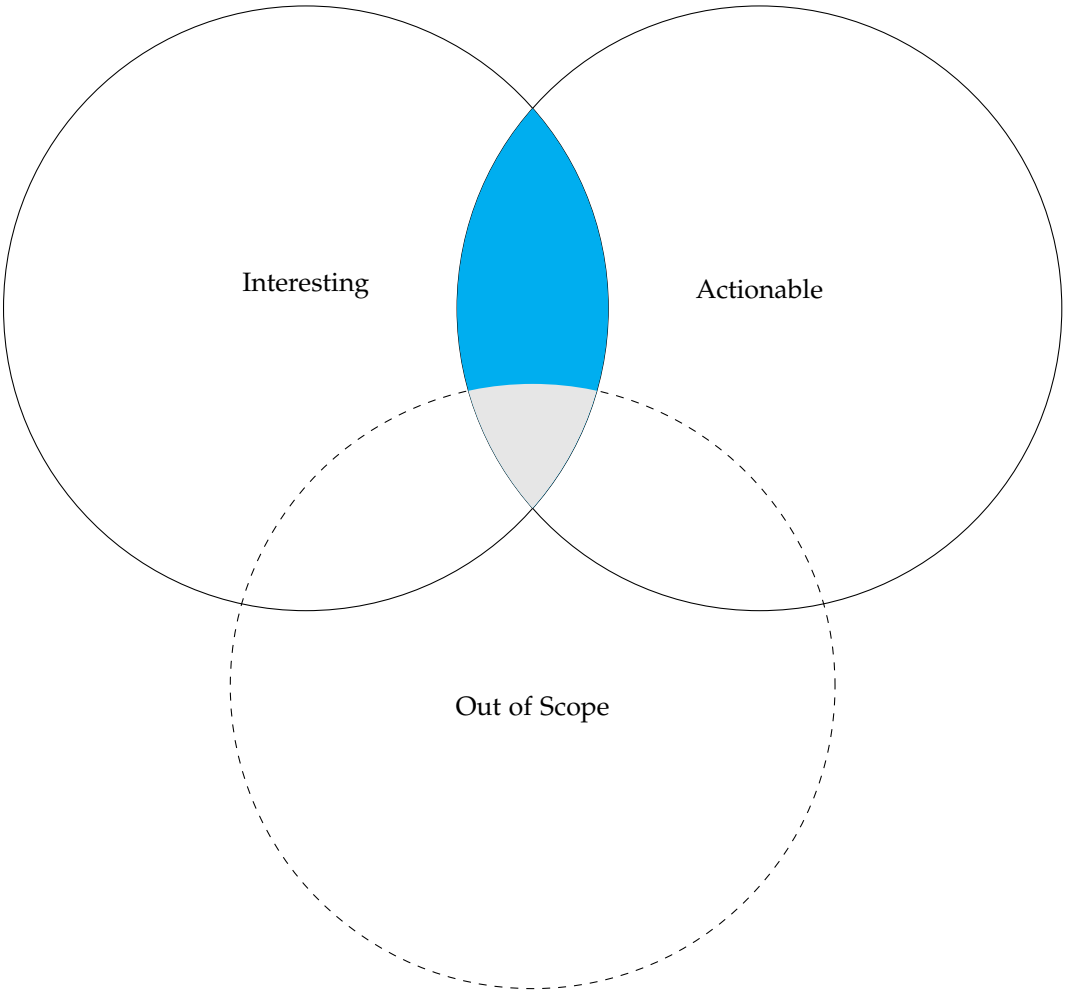
IT IS VERY TEMPTING TO MOVE TOO QUICKLY THROUGH THE PROBLEM FRAMING PROCESS, but doing so can lead to major problems later in the project. To avoid moving too quickly, see if you can answer all questions in the list below. If you cannot, you may need to further refine your problem statement.

1. What symptoms lead you to believe this problem is important to the organization?
2. Who is the problem you have defined important to and why? What decisions can they improve or mistakes can they avoid if you address this problem?
3. Do you see a feasible way to address this problem using available resources for research?[12]
4. Have you started with a broad definition of the problem, but then narrowed it down to a very specific problem with clear phrasing on the question to be addressed?
5. Are there previous findings that already address the problem in a satisfactory way? Is your research really necessary?

[12] This is similar to the idea of the research question being "actionable".

Venn Diagram for Research Questions

Identify one or more potential research questions in each region.



A QUESTION IS INTERESTING if answering it will serve a clear business purpose. What business decision will the answer to this question inform? What mistake will it avoid? [13]

A QUESTION IS ACTIONABLE if it can be answered with available resources and a feasible research design.

A QUESTION MUST DEFINE THE BOUNDARIES OF THE PROJECT and you should be able to write down other questions that are both interesting and actionable but are not going to be tackled in your current project. [14]

[13] Remember that all research carries at least an opportunity cost if not tangible monetary costs. Answering questions that do not have a clear business purpose is an inefficient use of organizational resources.

[14] It's helpful to think about business priorities and compare the urgency or relative importance of questions to define which are out of scope.

Writing Research Objectives

RESEARCH OBJECTIVES ARE AN ACTION STATEMENT that explain what your research will accomplish. Objectives should be logically linked to your research question, [15] but they may be developed in tandem. Sometimes it is easier to brainstorm objectives that meet the needs of the business scenario and then articulate an overarching question that encompasses them. Other times, it is easier to start with the question and then move on to stating the objectives. Most times, this is an iterative process. Because objectives are action statements, they generally begin with a verb. Here is a helpful list of verbs to prompt ideas. Do not feel constrained to use only these.

[15] Writing down objectives is a key way to determine whether your question is actionable. Objectives specify actions that you must be able to complete with available resources and the planned study design.

Measure	Estimate	Describe	Identify
Depict	Categorize	Compare	Recommend
Assess	Rank	Explore	Forecast

Table 1: Here is a list of verbs you can use to begin a Research Objective statement. Do not feel constrained to just these words but avoid broad words like “analyze".

NOTE THAT THE WORD *analyze* DOES NOT APPEAR IN THE LIST. All of these words are forms of analysis, some more specific than others. Try to pick as specific a word as possible. It should provide your audience with an idea of what the final work produce will look like. If you say **depict** then a visual display of results is expected. If you say **forecast** then a statistical model that produces future values of a series is expected.

Here are some examples of these words in action.

- Estimate consumer willingness to pay for a reserved seat based on the number of tickets in their booking.
- Depict patterns in customer service call volumes by time of day and season.
- Rank possible benefit offerings by employee preference.
- Categorize unnecessary hospital emergency room by diagnoses and discharge status.
- Forecast future daily shipping costs based on last quarter.

Writing Testable Hypotheses

Simple hypotheses can be grouped into common types or "flavors" and more complex hypotheses are often just combinations of these simple forms. Referring to these forms can be helpful in the analysis planning process because they help you to formalize your ideas and ensure your research question is actionable.

Here are some hypothesis "flavors" with examples:

- Hypotheses that propose a relationship between two variables.
  - **Variable A** is positively associated with **Variable B.** [16].
  - **Variable A** is negatively associated with **Variable C**. [17].
  - When **Variable D** increases by *x units*, **Variable A** will increase by *z units*. [18].
  - The average of **Variable A** will be above the average of **Variable B**. [19].
  - The average of **Variable A** will differ across categories of **Variable E**. [20]
- Hypotheses that involve a single variable.
  - **Variable A's** mean is greater than #*c*. [21].
  - **Variable B's** median is below #*c.* [22]
  - The proportion of **Variable C observations that are Category 1** is less than *p*%.[23]

In each of the hypotheses above, you must figure out what your key variables are. Think of Variables A, B, C, and D in each example as columns in a data set. If you cannot think of a way to measure Variable A and store that information in a column, then you may not have a testable hypothesis.

Depending on the hypothesis you choose, you may have other "blanks" to fill in. For example, if you are saying there is a relationship of a particular size between two variables, you need to specify the sizes of the changes in each variable (like *x units* and *z units* above). For all of the hypotheses involving only one variable, you must specify threshold values (like #*c* and *p*% above) to compare to. All of these things are called "constants". Unlike the variables, they will not appear as columns in your data set. Instead, they are used when you interpret your results and when telling a statistical program how to run a hypothesis test for you. In business, these constants often comes from performance targets, break even points, aggregate competitor data, or other sources other than the data you

[16] **Quarterly advertising expenditures** are positively associated with **quarterly profits**

[17] **Employee hours of safety training** are negatively associated with **annual injury rates**

[18] For *every additional dollar* spent on **advertising**, **quarterly profits** increase by more than *$3*

[19] **Average annual patient expenditures on prescription drugs** will be above **average household spending on gym memberships and fitness equipment**

[20] **The average number of sick days taken per year** is greater **among non-managerial than among managerial employees.**

[21] **Average weekly hours worked among full-time employees** is below the conventional threshold of *40 hours per week*

[22] **Our median flight departure delay time** is below the industry median of *13 minutes*

[23] The proportion of **consumers who say they prefer to purchase cosmetics at a department store (rather than at a drug store, online, or in a big box retail store)** is greater than *25%*

are analyzing. The constant value should be chosen because of its importance for the business decision or insight you are creating. A performance target may be set at a given level because it is tied to year-end profit goals or to published guarantees about product safety and performance.