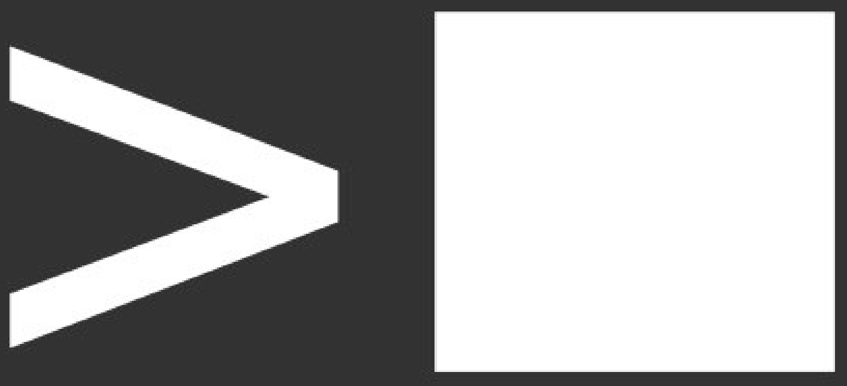
The Elements of Data Analytic Style



Jeff Leek

The Elements of Data Analytic Style

A guide for people who want to analyze data.

Jeff Leek

This book is for sale at http://leanpub.com/datastyle

This version was published on 2015-03-02

This is a Leanpub book. Leanpub empowers authors and publishers with the Lean Publishing process. Lean Publishing is the act of publishing an in-progress ebook using lightweight tools and many iterations to get reader feedback, pivot until you have the right book and build traction once you do.

An @simplystats publication.

Thank you to Karl Broman and Alyssa Frazee for constructive and really helpful feedback on the first draft of this manuscript. Thanks to Roger Peng, Brian Caffo, and Rafael Irizarry for helpful discussions about data analysis.

Contents

1. Introduction

The dramatic change in the price and accessibility of data demands a new focus on data analytic literacy. This book is intended for use by people who perform regular data analyses. It aims to give a brief summary of the key ideas, practices, and pitfalls of modern data analysis. One goal is to summarize in a succinct way the most common difficulties encountered by practicing data analysts. It may serve as a guide for peer reviewers who may refer to specific section numbers when evaluating manuscripts. As will become apparent, it is modeled loosely in format and aim on the Elements of Style by William Strunk.

The book includes a basic checklist that may be useful as a guide for beginning data analysts or as a rubric for evaluating data analyses. It has been used in the author's data analysis class to evaluate student projects. Both the checklist and this book cover a small fraction of the field of data analysis, but the experience of the author is that once these elements are mastered, data analysts benefit most from hands on experience in their own discipline of application, and that many principles may be non-transferable beyond the basics.

If you want a more complete introduction to the analysis of data one option is the free Johns Hopkins Data Science Specialization[1].

As with rhetoric, it is true that the best data analysts sometimes disregard the rules in their analyses. Experts usually do

[1] https://www.coursera.org/specialization/jhudatascience/1

this to reveal some characteristic of the data that would be obscured by a rigid application of data analytic principles. Unless an analyst is certain of the improvement, they will often be better served by following the rules. After mastering the basic principles, analysts may look to experts in their subject domain for more creative and advanced data analytic ideas.

2. The data analytic question

2.1 Define the data analytic question first

Data can be used to answer many questions, but not all of them. One of the most innovative data scientists of all time said it best.

> The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.
>
> John Tukey

Before performing a data analysis the key is to define the type of question being asked. Some questions are easier to answer with data and some are harder. This is a broad categorization of the types of data analysis questions, ranked by how easy it is to answer the question with data. You can also use the data analysis question type flow chart to help define the question type (Figure 2.1)

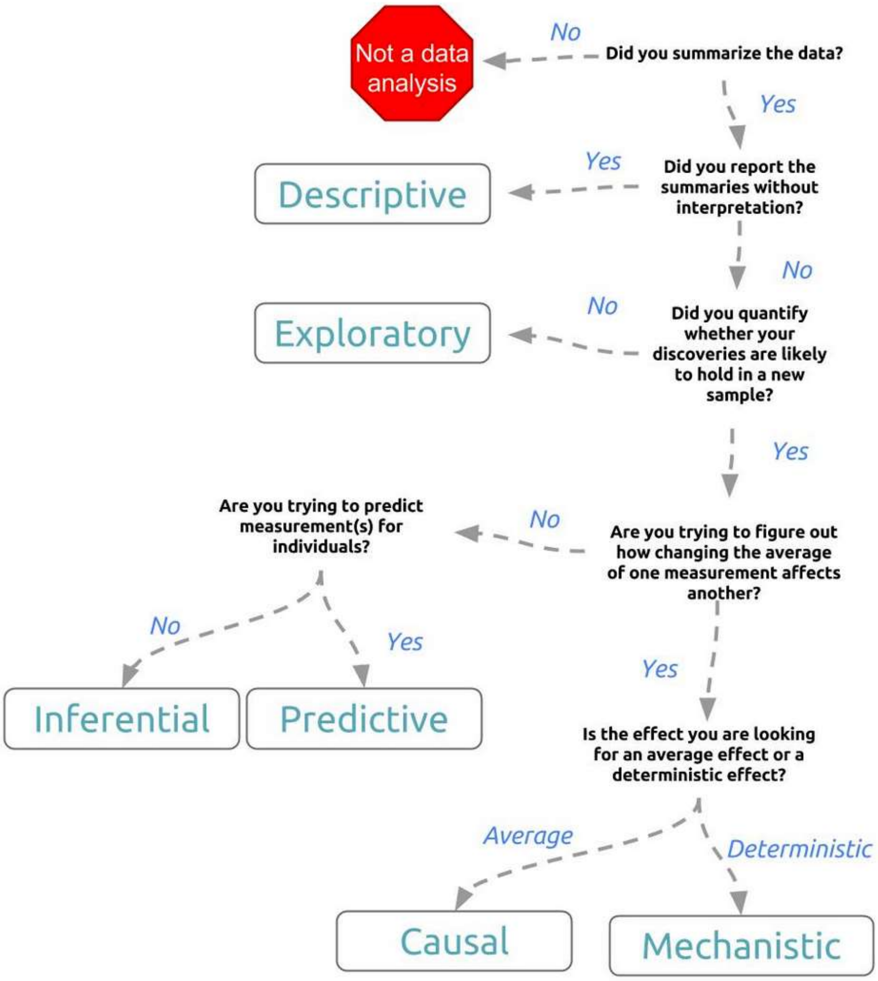

Not a data analysis
No
Did you summarize the data?
Yes
Descriptive
Yes
Did you report the summaries without interpretation?
No
Exploratory
No
Did you quantify whether your discoveries are likely to hold in a new sample?
Yes
Are you trying to predict measurement(s) for individuals?
No
Are you trying to figure out how changing the average of one measurement affects another?
No
Yes
Inferential
Predictive
Yes
Is the effect you are looking for an average effect or a deterministic effect?
Average
Deterministic
Causal
Mechanistic

Figure 2.1 The data analysis question type flow chart

2.2 Descriptive

A descriptive data analysis seeks to summarize the measurements in a single data set without further interpretation. An example is the United States Census. The Census collects data on the residence type, location, age, sex, and race of all people in the United States at a fixed time. The Census is descriptive because the goal is to summarize the measurements in this fixed data set into population counts and describe how many