Why Do EDA

MCC

2022-04-05

I was discussing Data Science with a Journalist the other day, and a question came up about why are Boxplots important. Boxplots and Histograms are the first tools taught with statistical software, but why? Many courses gloss over the "WHY."

Since this discussion started with a journalist co-worker, I was reminded of the 5 W's of reporting:

1. Who
2. What
3. When
4. Where
5. Why

- How might a journalist think of the five W's when analyzing a dataset.
- Can these five questions be useful when carrying out a Data Analysis?

In short, **Yes**.

And, I believe the five W's should be asked by Data Scientists too.

- **Why** did your research come to find this result?
  - Why is this data being analyzed at all?
  - Are we hypothesis testing?
  - Are we looking for a correlation or a regression line?

  - Is the data normally distributed?
- **When** and **Where** was the data collected?
  - If a political poll is asked in 2 cities, is it reasonable to assume that you can apply 'your poll' to the entire country?
  - Could the time, day, or season the data was collected change over time?
- Is it important, **Who** gathered this data?
  - What prejudices did that person or group have while collecting the data?
- And finally, **What** can be learned from the dataset?

What do these examples of asking the five W's have in common? Bias.

Commonly, one of the first topics discussed in statistics classes is the concept of bias.

> Bias - Inclined to one side; swelled on one side; an inclination; to give a particular direction to; to influence; to prejudice; to prepossess.
>
> Webster, 1913

What is wrong with your data?

Even for a scientist, the five W's are a good means to focus your skepticism. In the world of scientific research, many people think about the simplest example to study. This is the reductionist approach. But what happens when political polls are asked of a number of people in one city or two cities. Can that poll be generalized to include everyone in that county, state and country?

- What is wrong with the data?
- Is there anything wrong with your study?
- Is it possible to fix the study or data?
- Can the mistake even be fixed?

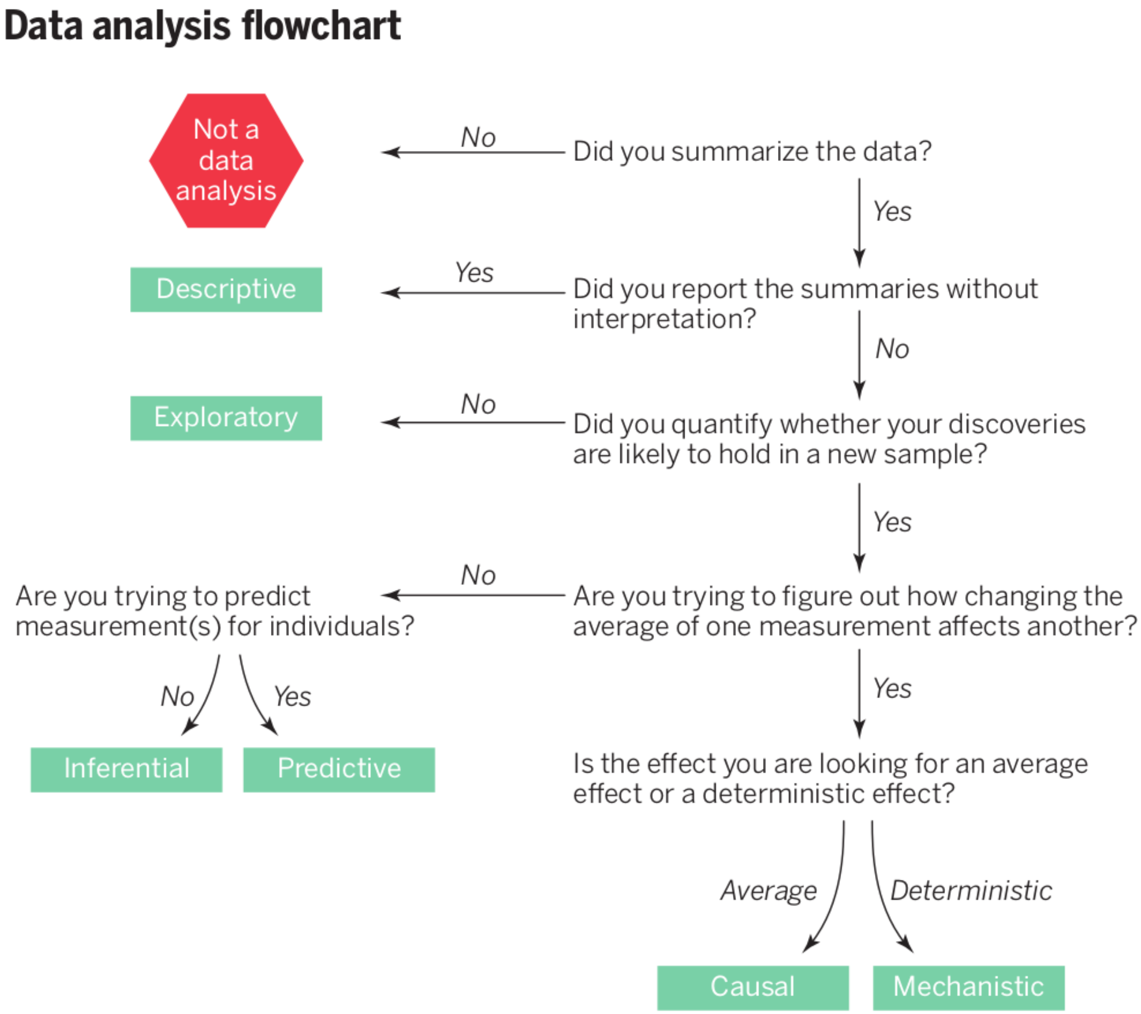
Exploratory Data Analysis is the first review of your data and information.

In my mind, EDA is the first attempt to investigate

'What is wrong with your data?'. The second question might be 'Is it possible to understand where the errors inherent in your study came from? And lastly, Is it possible to fix these problems, with a focus on minimizing all the errors in a study.

Jeff Leek and Roger Peng have an interesting view of Data Science in the article, 'What is the Question?'[1]

[1] Science, 20 MARCH 2015, VOL 347, ISSUE 6228, P. 1314


Data analysis flowchart
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
Yes
Inferential
Predictive
Is the effect you are looking for an average effect or a deterministic effect?
Average
Deterministic
Causal
Mechanistic

In the flowchart above, Data Science and Statistics can be broken down into 6 categories. In this piece, I will be focusing on Exploratory Data Analysis (EDA).

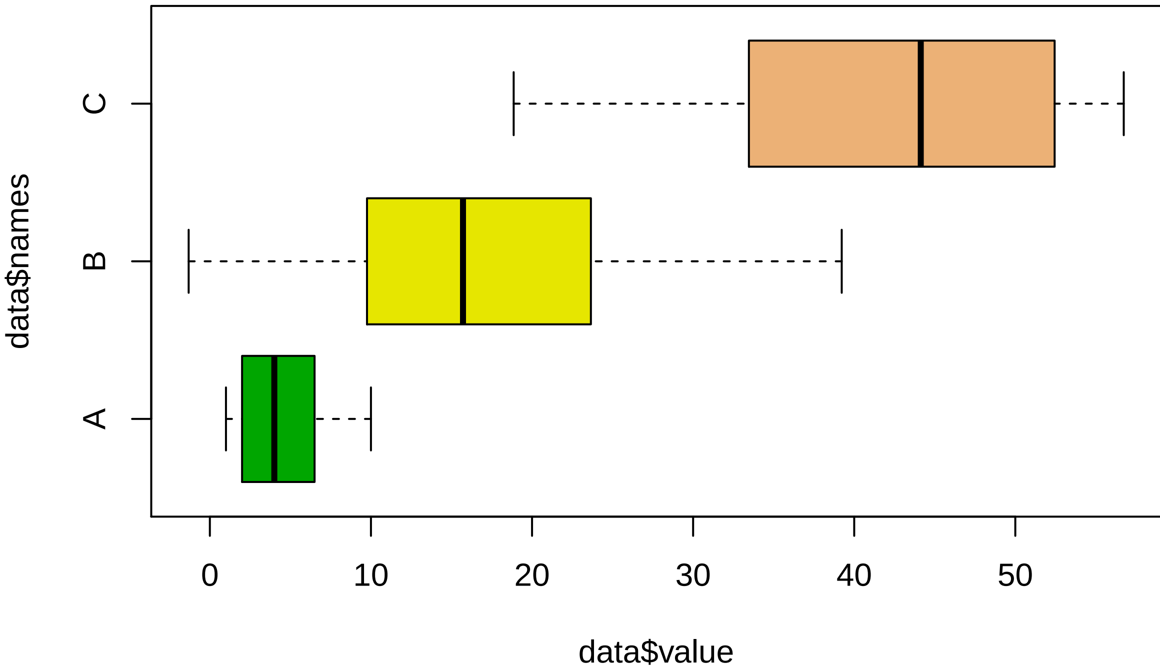
I plan to investigate two mathematical perturbations that can be interpreted using Boxplots and Histograms. Two common measures for distributions are skew and kurtosis.

> Perturbation, in mathematics, a method for solving a problem by comparing it with a similar one for which the solution is known. Usually the solution found in this way is only approximate.
>
> https://www.britannica.com/science/perturbation-mathematics

But first, let us prepare some good and bad data.

```
# Create data
names <- c(rep("A", 15) , rep("B", 15) , rep("C", 15))
value <- c( rpois(15, 5) , rnorm(15, mean=10, sd=15) , rnorm(15, mean=40, sd=10) )
data <- data.frame(names,value)

boxplot(data$value ~ data$names , col=terrain.colors(4), horizontal = T )
```

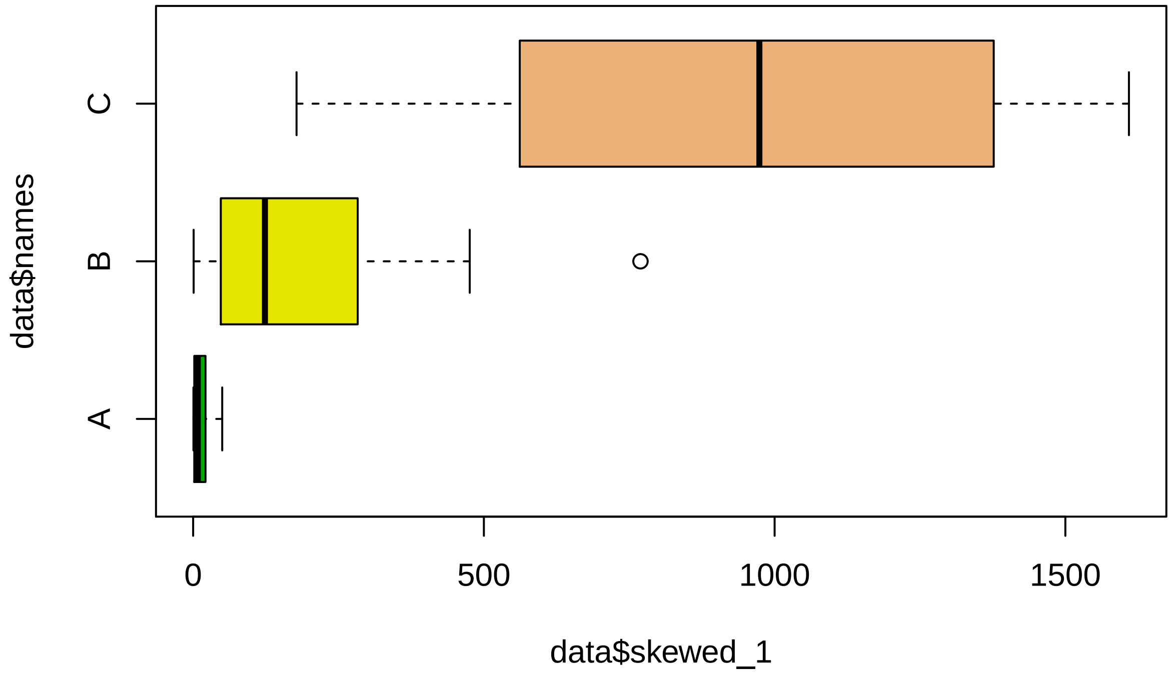


Generate a dataset

Square Values

```
skewed_1 = (1/2)*(data$value)^2
data <- cbind(data, skewed_1)

boxplot(data$skewed_1 ~ data$names , col=terrain.colors(4), horizontal = T )
```



https://statisticsglobe.com/jitter-r-function-example/ https://stackoverflow.com/questions/23675735/how-to-add-boxplots-to-scatterplot-with-jitter https://www.data-to-viz.com/caveat/boxplot.html