

# Problem Set #1

Pedro Alberto Arroyo

10/8/2019

## Exploration & Computation

### 1. Obtain a dataset (preferably of substantive interest/domain expertise).

I used the following dataset on hate crimes, which I got Five Thirty Eight's GitHub page:

<https://github.com/fivethirtyeight/data/commit/fbc884a5c8d45a0636e1d6b000021632a0861986>

### 2. Choose a visual technique to illustrate your data (e.g., barplot, histogram, scatterplot). 3. Now generate and present the visualization and describe what you see.

```
#Load Libraries
```

```
library(tidyverse)
```

```
## -- Attaching packages -----  
## v ggplot2 3.2.1      v purrr  0.3.2  
## v tibble  2.1.3      v dplyr  0.8.3  
## v tidyr   1.0.0      v stringr 1.4.0  
## v readr   1.3.1      v forcats 0.4.0
```

```
## -- Conflicts -----  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag()     masks stats::lag()
```

```
library(skimr)
```

```
##  
## Attaching package: 'skimr'  
  
## The following object is masked from 'package:stats':  
##  
##   filter
```

```
# Load the data
```

```
HCdata <- read_csv("/Users/pedroLAPTOP/Documents/rstudiodef/Problem\ Set/Problem\ Set\ 1/rm/hate_crimes
```

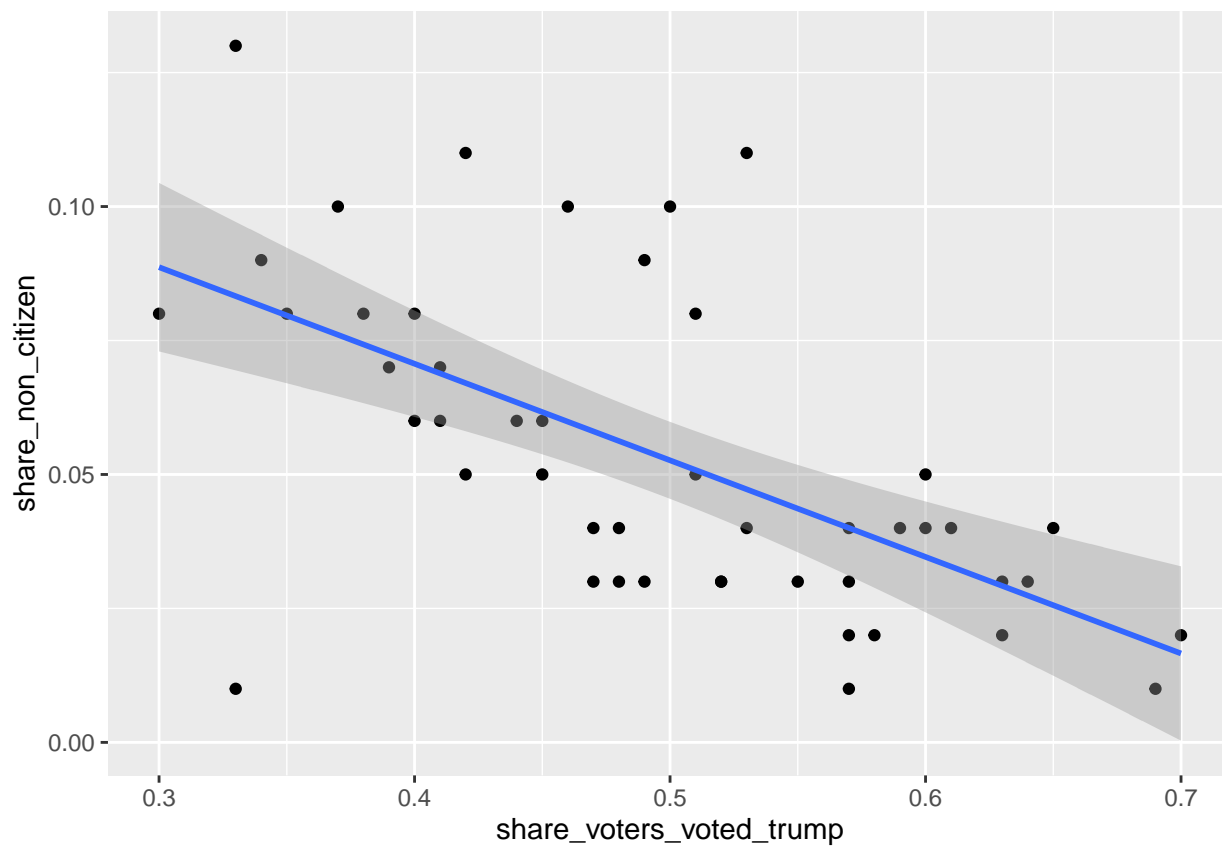
```
## Parsed with column specification:  
## cols(  
##   state = col_character(),  
##   median_household_income = col_double(),  
##   share_unemployed_seasonal = col_double(),  
##   share_population_in_metro_areas = col_double(),  
##   share_population_with_high_school_degree = col_double(),  
##   share_non_citizen = col_double(),  
##   share_white_poverty = col_double(),  
##   gini_index = col_double(),  
##   share_non_white = col_double(),  
##   share_voters_voted_trump = col_double(),  
##   hate_crimes_per_100k_splc = col_double(),  
##   avg_hatecrimes_per_100k_fbi = col_double()  
## )
```

```
#remove DC for being an egregious outlier
HCdata_sub <- HCdata[-c(9), ]
```

```
ggplot(HCdata_sub, (aes(x=share_voters_voted_trump, y=share_non_citizen))) +
  geom_point() +
  geom_smooth(method='lm')
```

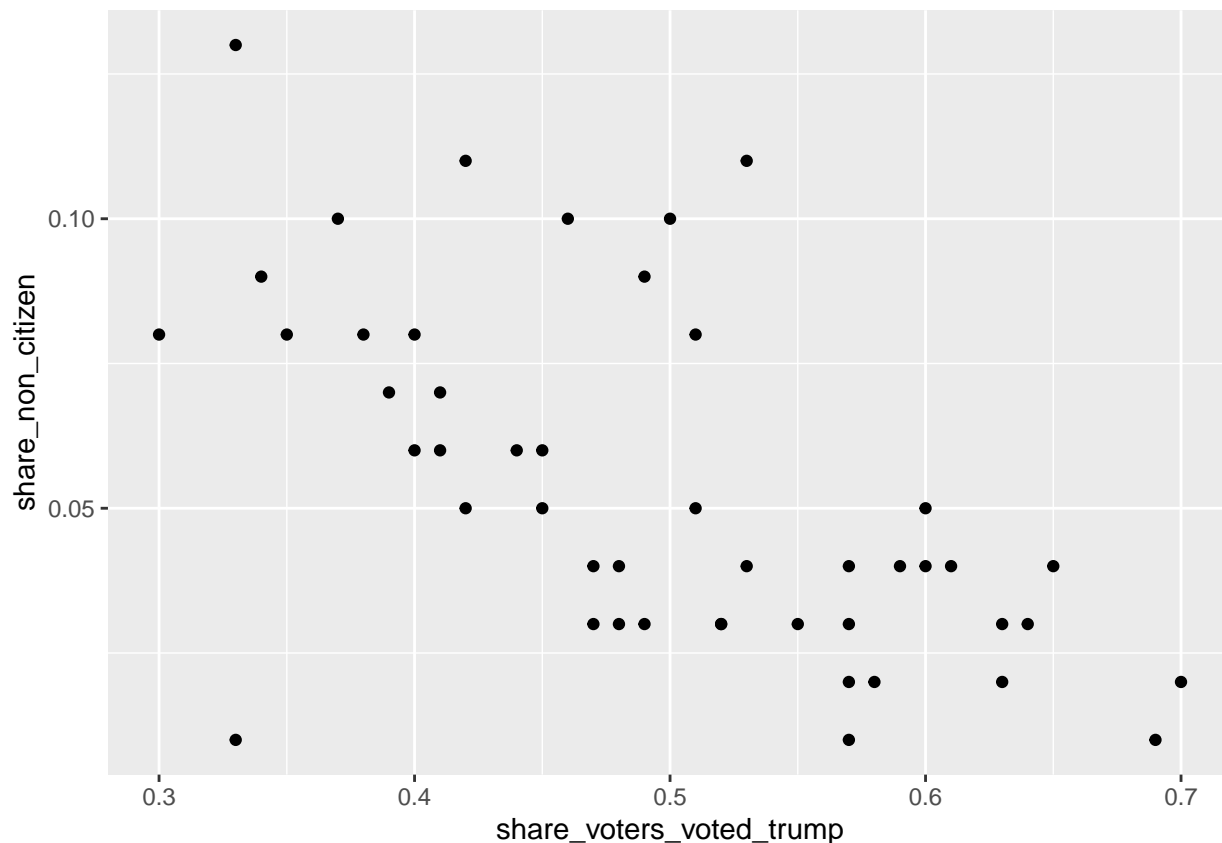
```
## Warning: Removed 3 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```



```
ggplot(HCdata_sub, (aes(x=share_voters_voted_trump, y=share_non_citizen))) +
  geom_point()
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```



I plotted the share of a state's non-citizen population against the share of the vote that President Trump received in the 2016 election. I also ran it again with a linear regression line.

The general trend seems to be that the lower the rate of non-citizens in the state, the better Trump performed in that state. This raises a host of interesting questions. Most obviously: given the emphasis that candidate Trump placed on immigration, it is noteworthy that, to a first approximation from this limited data, that message seems to have been more effective in states with fewer immigrants. However, though non-citizens cannot vote, they might be a proxy for demographically similar citizens who might have reacted negatively to the messaging of the Trump campaign. To further investigate this, I would love to see the breakdown of voting behavior by race to see whether the relationship holds when looking just at white voters.

Three states (Maine, Mississippi, South Dakota) had to be excluded because the dataset did not report non-citizen population data for those states. Washington, D.C. was also excluded because it was an extreme outlier across the board. Of the remaining states, two states where Trump performed poorly really diverge from the trendline - California and Vermont.

There's also a cluster of states above the trendline that catch my eye.

It's hard to draw any conclusions from the visualization, but it really does generate a lot of ideas. My gut feeling is that I would want to disaggregate the data: do these trends look different if we split out blue, red, and swing states? Obviously, something is going on here.

#### 4. Calculate the common measures of central tendency and variation, and then display your results.

```
#get just the columns I'm interested in
HCdata_sub_s <- HCdata_sub %>%
  dplyr::select(state, share_non_citizen, share_voters_voted_trump) %>%
  drop_na()
```

```
#get summary stats
summary(HCdata_sub_s)
```

```
##      state      share_non_citizen share_voters_voted_trump
## Length:47      Min.      :0.0100      Min.      :0.3000
## Class :character 1st Qu.:0.0300      1st Qu.:0.4150
## Mode  :character Median :0.0400      Median :0.4900
##                      Mean  :0.0534      Mean  :0.4957
##                      3rd Qu.:0.0800      3rd Qu.:0.5700
##                      Max.   :0.1300      Max.   :0.7000
```

### 5. Describe the numeric output in substantive terms, e.g.,

- a. What do these numeric descriptions of data reveal? - b. Why is this important? - c. What might you infer about the distribution or spread of the data? Why? d. Etc.

The numeric descriptions indicate a couple of things. For the share of votes received by Trump, the median/mean are right around the 50/50 mark with states seemingly equally distributed throughout the range. My initial reaction was that this makes sense given a tight election - i.e. the states are more or less equally distributed above and below the 'endorse/reject' line, but upon further reflection I don't think it is such an obvious distribution. Given the partisan environment, you might expect a very tight cluster around the 50/50 mark (meaning a lot of states are tightly contested) and/or you'd expect a lot of states towards the outside of the distribution (meaning a lot of states are highly partisan). But, it doesn't seem like that's actually the case...which is (somehow) odder when you consider that the states represent vastly different population sizes.

Put differently, the way we talk about elections is in terms of flippable states... which puts a lot of states safely in the 'red' or 'blue' category based on secure voting outcomes. In electoral politics, a reliable 6% is absolutely huge; but in terms of actual people on the ground, a 53/47 is still a fairly moderate split.

The common wisdom on the partisan divide still holds as a predictive heuristic, which is almost entirely how it's used. But taking a look at the actual distribution of partisan preference raises questions about what structural features might become relevant in a slightly different political environment. In the language of social movements, the opportunity landscape might be different than people think it is.

## Critical Thinking

**1. Describe the different information contained in/revealed by visual versus numeric exploratory data analysis. (Hint: Think of different examples of each and then what we might be looking for when leveraging a given technique).**

Visual EDA is particularly useful for revealing patterns or underlying structures in the data that are not easily captured by descriptive statistics; for example, visual EDA can help detect outliers or unusual distributions. Generally, visual inspection of data is useful because humans are particularly adept at visual pattern recognition. Visual EDA can thus help researchers identify that there exists *a pattern* as well as gain insight as to what that pattern *may be* capturing. Having generated that insight, the researcher can then move on to other techniques in order to test for explanations that may establish the underlying causes of the observed data structure.

Numeric EDA is useful for interrogating the internal relationships between observations in a data set. For example, the variance of the distribution and skew of the distribution can be features of interest when conducting numeric EDA. This sort of approach is a form of data reduction; which, by definition, means that the researcher decreases the level of resolution at which the data is examined. One benefit of this reduction is that it allows for comparisons between distributions - e.g. a researcher can use numeric EDA to explore how voter clustering differs between demographic groups.

**2. Find (and include) two examples of "bad" visualizations and tell me precisely why they're bad.**

1. **Figure 1.** I'm presenting this as an example of 'bad' visualization, but it's very likely an example of purposefully deceptive visualization. The main problem is the misleading scale/range of the y-axis. By only presenting the top quartile of the range, the visualization presents the difference between Android and iOS as approximately 50/50; when it is, in truth, closer to 15/85.
2. **Figure 2.** This is an example of bad data visualization. The most obvious problem concerns Andrew Yang's fundraising number, where \$2.8M is depicted with a longer bar than \$10M, which I have to assume is plainly an error. But, the scale of the bars is otherwise inconsistent. For example, Michael Bennet's \$2.1M (Q3) is depicted with a bar that is roughly a fifth the length of Bernie Sanders' \$25.3M (Q3), despite the fact that the actual difference in scale is closer to 1:12.

3. **Find (and include) two examples of "good" visualizations and tell me precisely why they're good.**

**Figure 3.** This is an example of a good visualization because it represents data parsimoniously and clearly. The y-axis is clearly marked and the use of branded imagery for representing the actual batteries addresses a user-issue that might otherwise be easy to overlook: consumers are likelier to process batteries by their appearance than their brand and model number. This is particularly evident when we compare the two Panasonic batteries, which perform differently and are visually easily distinguishable even as consumers are unlikely to know or later recall their model numbers.

**Figure 4.** *Human Metabolic Rate* does a good job of presenting the difference demands of body functions on metabolism between two idealized people. The use of color, the close placement of the bars, and the location of the key are particularly helpful for making these comparisons quickly.

4. **When might we use EDA and why/how does it help the research process?**

EDA is useful for getting a *feel* for a data set before performing further analysis. It can be used to generate hypotheses for testing, which can be the result of observing patterns in the data that require further explication or which suggest interesting relationships. EDA can also be useful for understanding the data structure - is it clean? are there outliers? are there clustering effects that could skew descriptive statistic approaches?

One way to think about the use of EDA is that it is a tool that a researcher can deploy to gain insights into a dataset that precede formal analysis of that dataset. This is useful on two fronts: either because the dataset is new and therefore no previous insights exist to guide inquiry or (and I think this might be of at least equal importance) because the researcher wishes to look at a 'well-understood' dataset in a new way.

There are a lot of statistical artifacts in circulation that are familiar to researchers and which generate relevant *stylized facts*. For example, the unemployment rate or the poverty rate or the rate of inflation are common econometric measures. Their familiarity elides their constructed nature. To wit: the headline unemployment rate is one of multiple rates calculated by the Department of Labor and these rates diverge and converge over time in ways that reflect changes in the structure of the economy, the poverty rate fails to account for changes in spending priorities over the past 70 years, and the rate of inflation is highly sensitive to a host of choices regarding value and how it changes over time.

In each of these cases, the headline numbers are operationalized versions of the underlying phenomenon that is of actual interest. The quality of that operationalization changes over time as the structure of the world in which it is located changes. EDA offers researchers the opportunity to explore the continuing value of their initial assumptions.

5. **What did John Tukey mean by "confirmatory" versus "exploratory"? Give me an example for each.**

Tukey makes a distinction between confirmatory analysis (which tests specified hypotheses) and exploratory analysis (which is more concerned with getting a feel for the data landscape).

What Tukey calls confirmatory analysis might be more accurately called disconfirmatory analysis since the statistical techniques used in hypothesis testing, strictly speaking, are better at disproving null hypotheses than proving proposed hypotheses. A whole rigmarole of epistemological and methodological questions

## Worldwide Smartphone Shipment OS Market Share Forecast

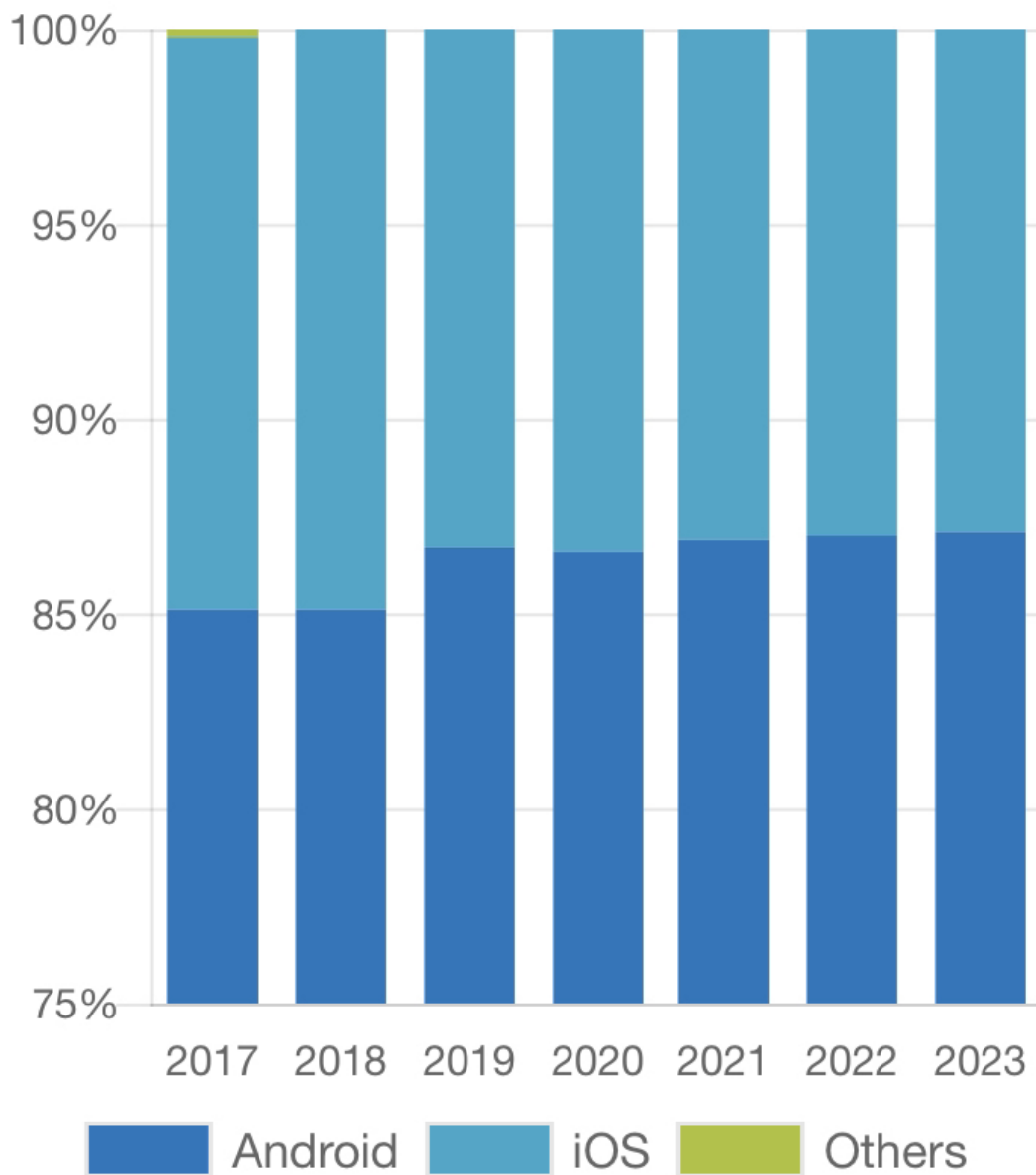





























Figure 1: Source: <https://www.idc.com/promo/smartphone-market-share/os>

# THE 2019 PRESIDENTIAL MONEY RACE

Who's Up and Who's Down?

APR - SEPT 2019

CANDIDATE'S  
CAMPAIGN  
(not including PACs)

		AMOUNT RAISED IN Q2	AMOUNT RAISED IN Q3
	BERNIE SANDERS	\$18M 	\$25.3M 
	ELIZABETH WARREN	\$19.2M 	\$24.6M 
	PETE BUTTIGIEG	\$24.9M 	\$19.1M 
	JOE BIDEN	\$21.5M 	\$15.2M 
	KAMALA HARRIS	\$11.8M 	\$11.6M 
	ANDREW YANG	\$2.8M 	\$10M 
	CORY BOOKER	\$4.5M 	\$6M 
	MARIANNE WILLIAMSON	\$1.5M 	\$3M 
	MICHAEL BENNET	\$2.8M 	\$2.1M 

abc NEWS

Figure 2: Source: ABC News

# WHICH BATTERIES LAST LONGEST?

11 different brands of AA batteries, tested in identical flashlights.

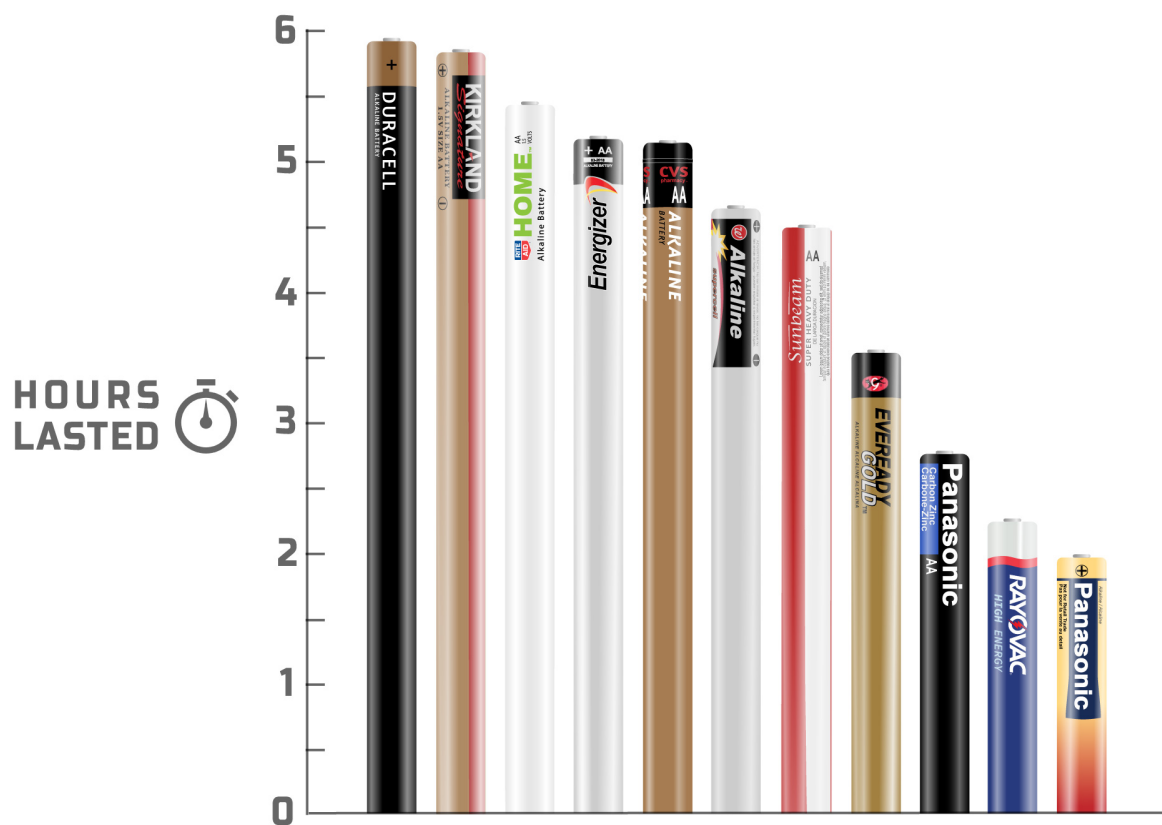


Figure 3: Source: r/dataisbeautiful; author: u/thecrispiestbacon



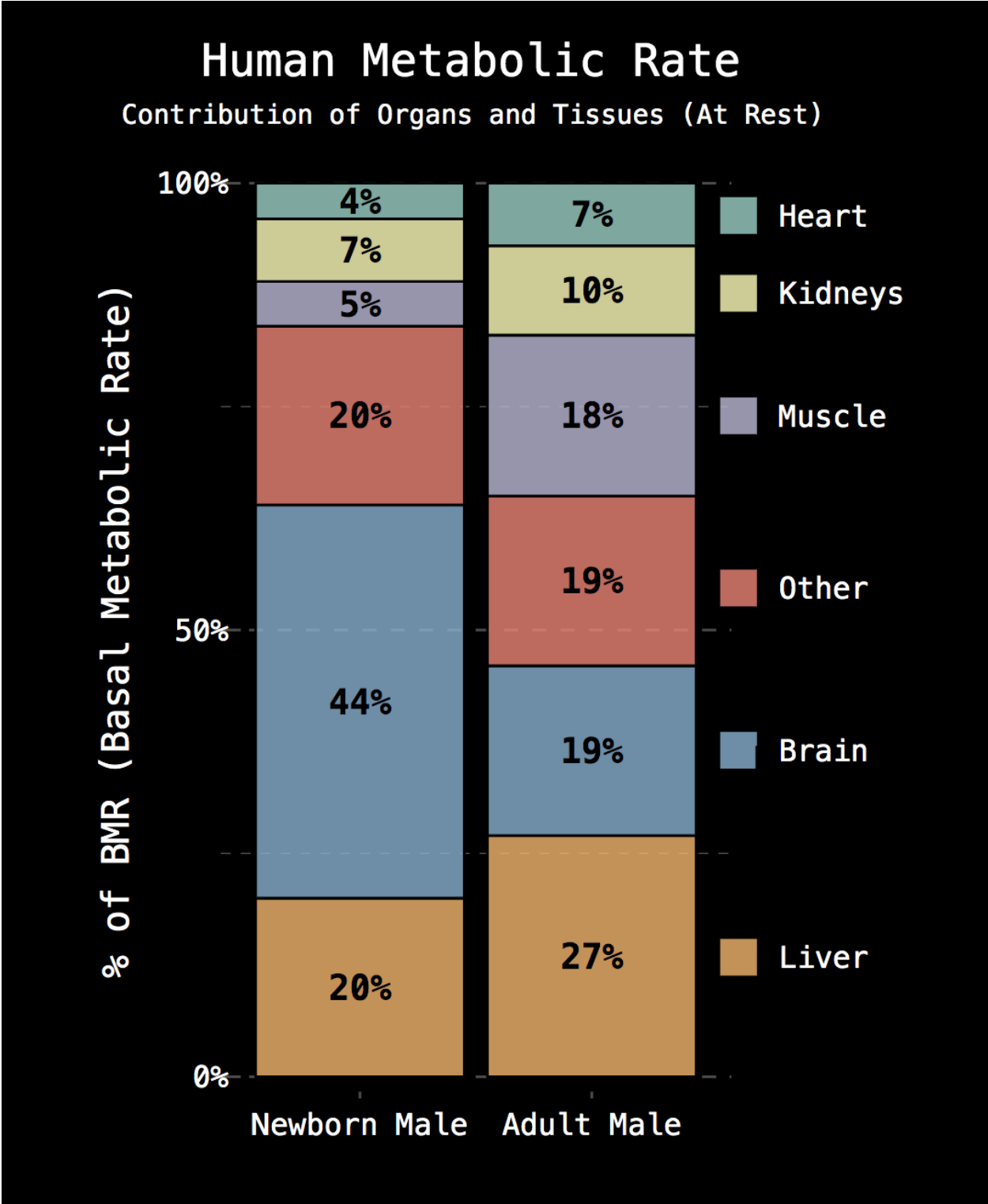


Figure 4: Source: [r/dataisbeautiful](#); author: [u/takeasecond](#)

plausibly arise at this point. But, leaving those aside, Tukey refers to statistical testing of specified hypotheses as confirmatory analysis. An example of this would be randomized drug trial.

(Though I do have to point out that Tukey specifically points to the pitfalls of confirmatory statistics when multiple questions are asked simultaneously - a concern as valid now as when he expressed it in 1980.)

Exploratory analysis is distinguished from confirmatory analysis in that it does not begin with a question, but rather with an interest in an area of inquiry. Tukey proposes some ways in which researchers might do this, with a strong emphasis on visualization. But I find the aphorism that EDA is more of an attitude than a set of tools to be the most useful way of thinking about it. I'm reminded of the observation that *the most important expression in science is not 'eureka', but 'that's funny'*. At heart, EDA is about nurturing an active curiosity coupled with a flexible imagination.

Strictly speaking, I am not sure if the example that I will provide is, in fact, an instance of EDA. But, I can't help but think of the construction of the periodic table at a time before the atomic structure of elements was understood. Working to create a visual representation of shared characteristics between elements, Mendeleev produced an instrument that accurately identified patterns that allowed scientists to make predictions about yet-undetected constituents of the universe *even as they could not yet explain what rules governed the emergence of the properties that led to those predictions*.