Problem Set #1- Exploratory Data Analysis

Abhishek Pandit

11 October 2019

EXPLORATION AND COMPUTATION

This data set comes from the academic performance of students in Grades 5-8 at the government schools of an economically backward district in Bihar, India. A non-profit(in my last job) had started a collaboration with the district-level administration 3 years ago to reduce the dropout rate.

It has conducted a test to gauge the levels of reading and writing ability in Hindi- the state language. This analysis will provide a 'before' picture so that it can plan its interventions for improved education in the government school system for the next 3 years.

```
library(ggplot2)
```

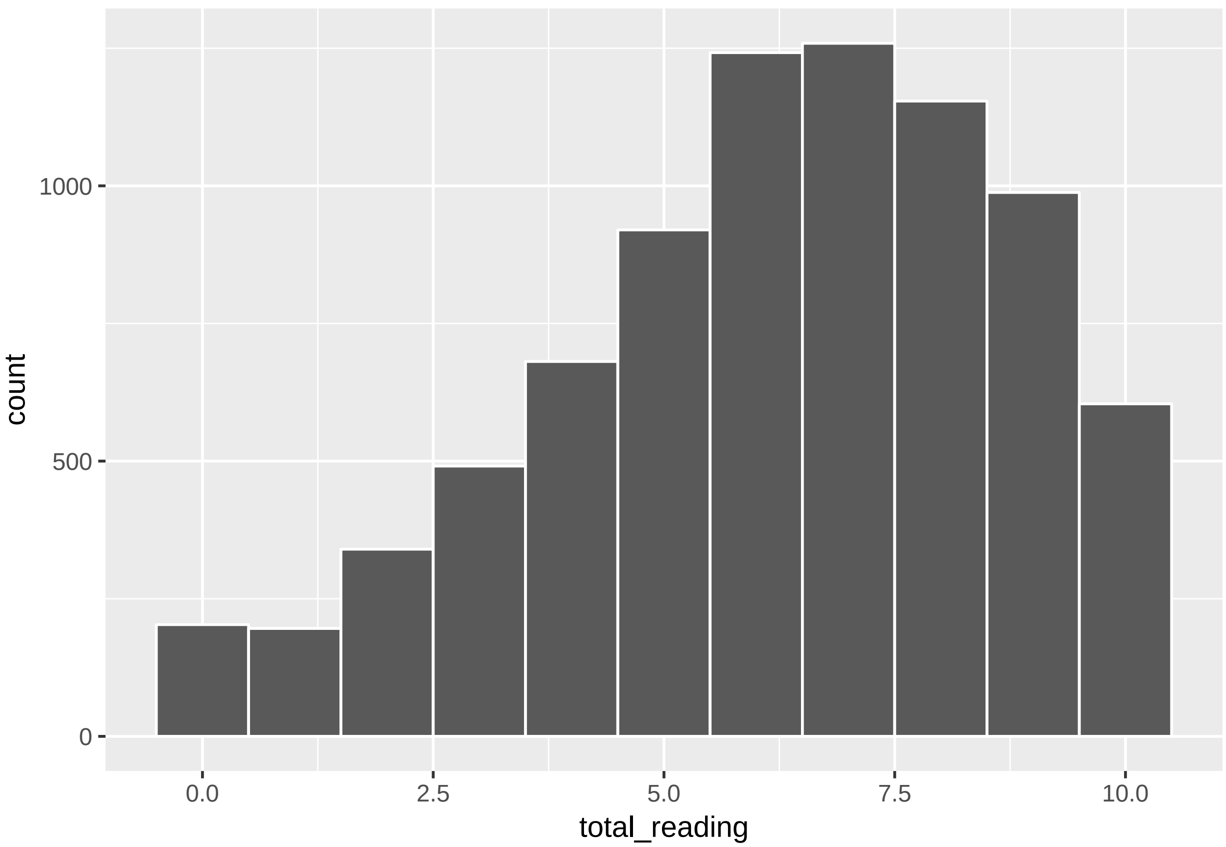
```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
school<-read.csv('student_baseline_clean.csv')
```

Including Plots

We begin with an exploration of the key outcome variables- the two academic scores.

```
ggplot(data=school, aes(x=total_reading))+
  geom_histogram(binwidth=1,color="white")
```



summary

```
summary(school$total_reading)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   5.000   6.000   6.202   8.000  10.000
```

These numeric data help us to understand the distribution, central tendency and variability of the data

Reading just these summary statistics, we can see that students earned between 0-10 on the test. The mean figure of 6.2 is not too far from the median (6), so we can assume that the distribution is not heavily skewed. Since the gap between the median and the first quartile is smaller than between the median and the third quartile, it would be fair to assume that there is a higher concentration of scores right before the median (between 5 & 6), than right after it (between 6 & 7).

The histogram supports these ideas. However, it provides added dimension to our understanding by stating the exact counts for each possible reading score.

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).

Overall, numerical exploratory analysis provides conciseness and specificity. We are able to get a rough idea of our data and its characteristics with a few statistics.

Visual exploratory analysis helps understand patterns such as clustering, distributions of variables and identify anomalous data points. Here are a few examples:

Univariate Analysis- Numeric- Measures of central tendency(mean, median, mode), summaries of variation Visual- Distribution (e.g. histogram, bar plot)

Multivariate Analysis Numeric- Correlations Visual- Shape of distribution, outliers

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

For the next two sections, I will be referring to the ALbert Cairo's (2016), 'The Truthful Art' where he refers to 5 parameters to evaluate the quality of visualizations

`Truthful`

`Functional`

`Beautiful`

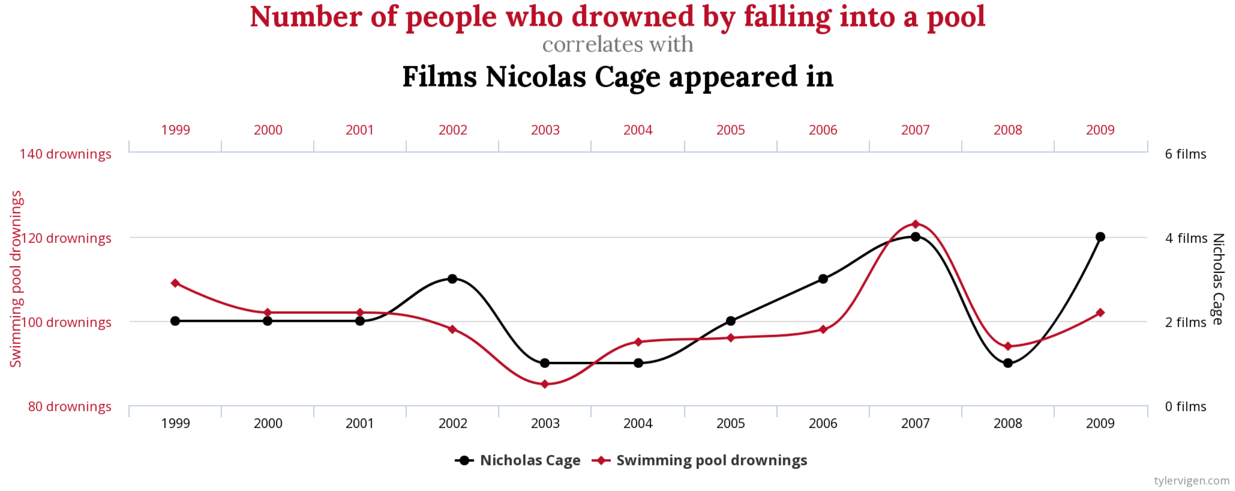
`Insightful`

`Enlightening`

BAD VISUALS

Bad Visual 1

The first bad visualization is entitled ' Number of People Who Died By Falling Into a Pool Correlates with Films Nicolas Cage Appeared in', available at this link: https://www.tylervigen.com/spurious-correlations



It fails by violating the principles of being 'Truthful', 'Insightful' or 'Enlightening'. By juxtaposing these entirely disparate phenomena, the graph's implicit claim appears to be that they are causally related. For example, that the release of Nicolas Cage Films somehow causes people to fall into pools. Furthermore, the scales used on either side of the graph are both not 'Truthful' (they attempt to make two vastly different scales appear comparable) and not Functional (since they are on either side of the core graph and are hard to follow). In its defence, the numbers displayed for each year by hovering over them do simplify understanding to some extent.

This visualization thus fails on just about every parameter.

Bad Visual 2

The second bad visual is entitled 'If Bush Tax Cuts Expire', available at the link below: https://flowingdata.com/2012/08/06/fox-news-continues-charting-excellence/

As mentioned in the accompanying post, the graph violates the three principles of 'Truthful', 'Insightful' or 'Enlightening'. The bar plot is 'Functional' as it does convey its intended message- that the change in the top tax rate (if the Bush tax cuts expire) would be huge and foreboding. However, this effect was created through incorrectly zooming in on the y axis (tax rates) with the cutoff at '34%' instead of the expected '0%'. This distorts a mere 4.6% to seem much larger than it actually is.

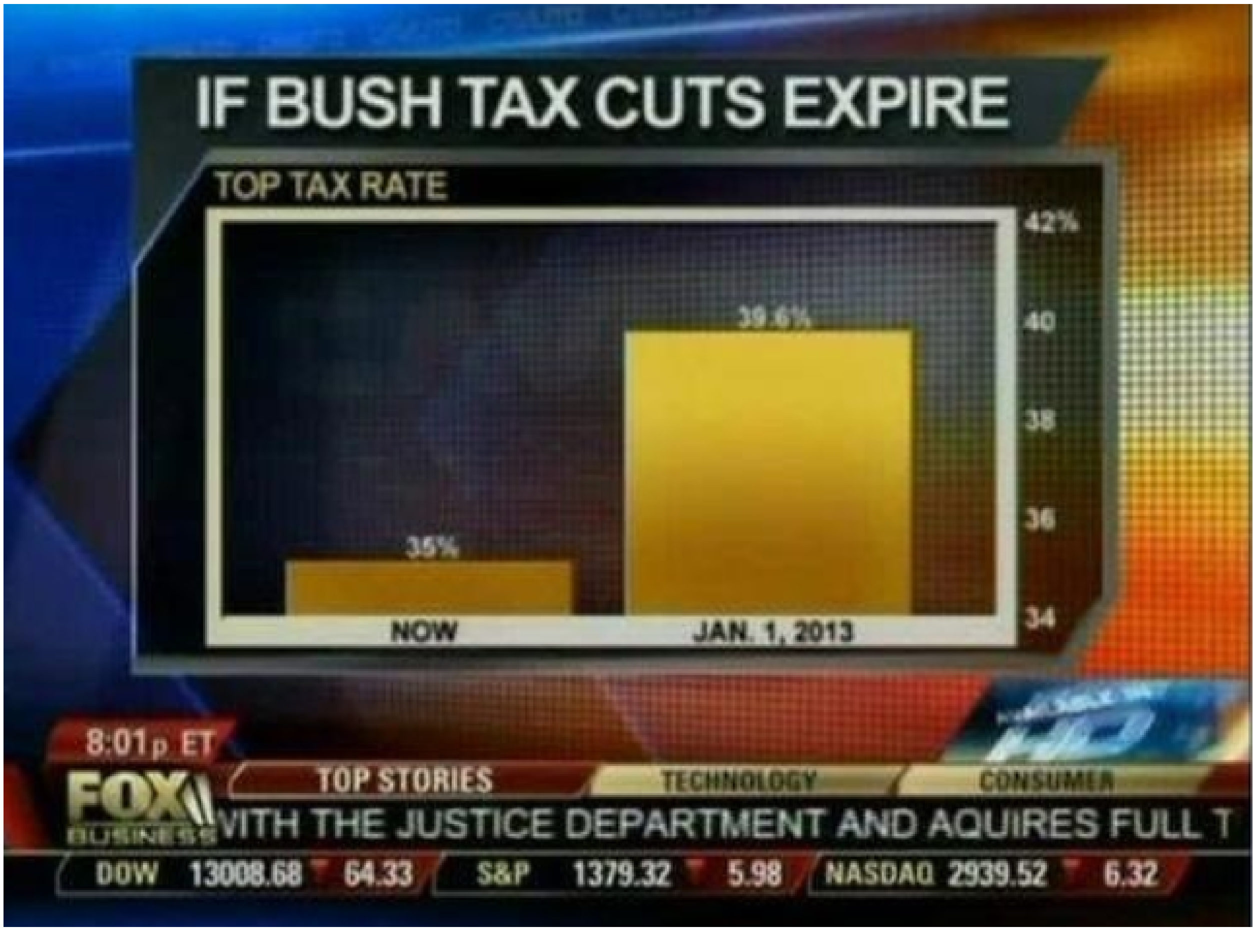


Figure 1: Bush Tax Cuts

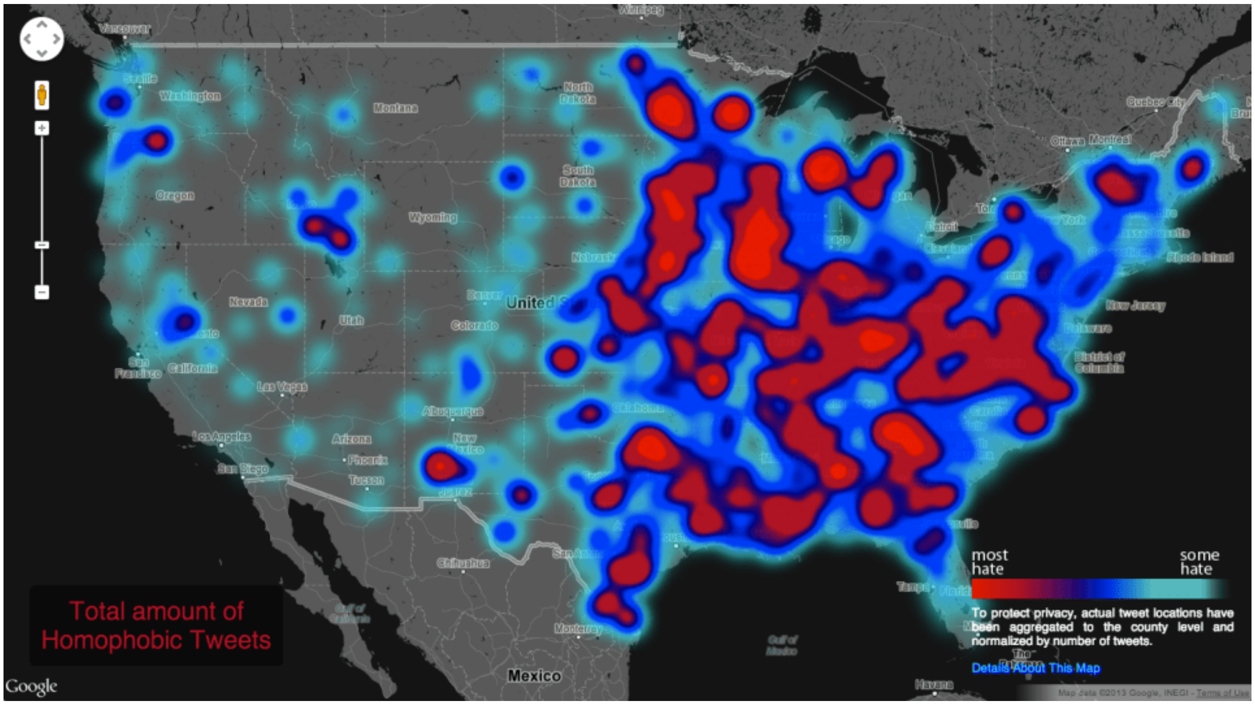


Figure 2: Homophobic Tweets

It is also not ‘Beautiful’ due to inclusion of an unnecessary lattice pattern in the background that only confuses the viewer further.

3. Find (and include) two examples of “good” visualizations and tell me precisely why they’re good.

GOOD VISUALS

Good Visual 1

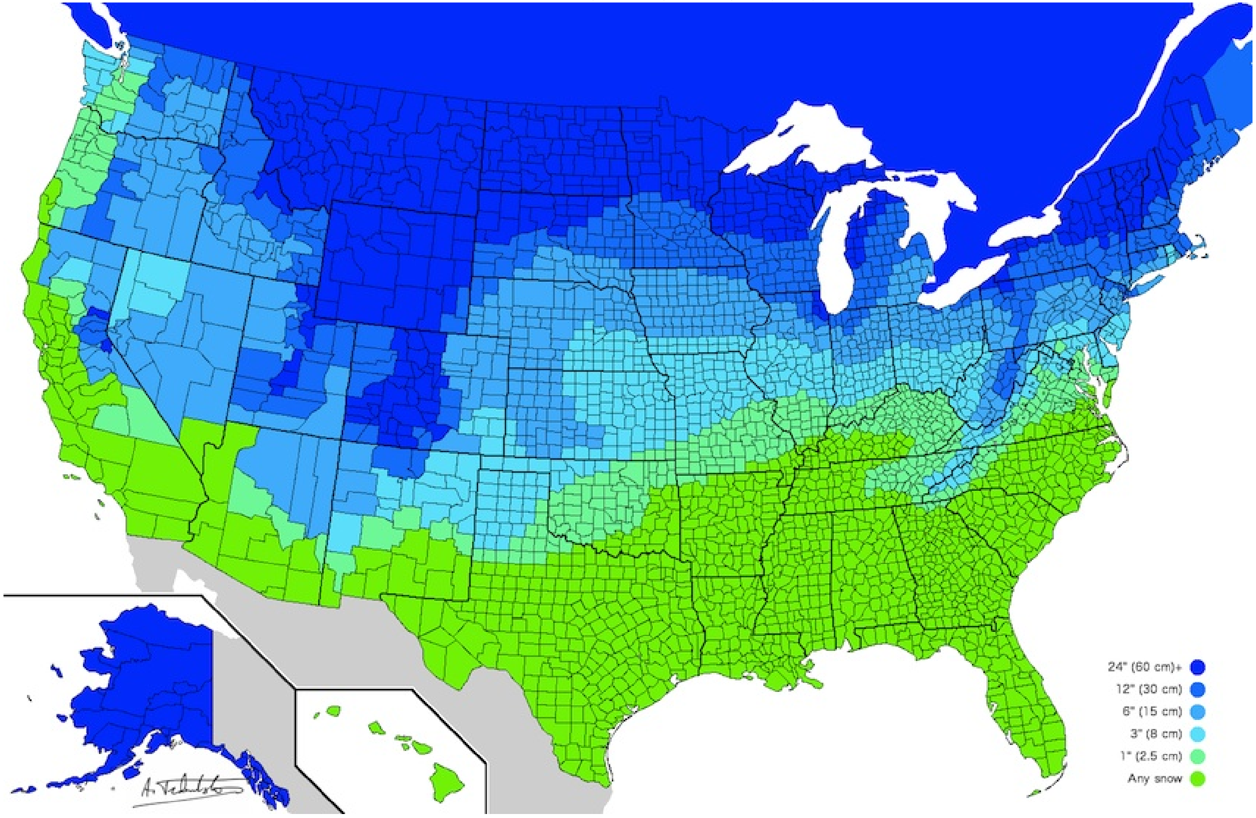
The first visualization is from Flowing Data, and is entitled ‘Total Amount of Homophobic Tweets’, at the top of this page: https://flowingdata.com/2013/05/13/geography-of-hate-against-gays-races-and-the-disabled/

It scores well on Cairo’s parameters in the following way: *Truthful- The visual uses scraped and geocoded Twitter data, and then controls for absolute number of Tweets by normalizing at the county level. The only major flaw is that they do not provide a caveat about certain areas with less use or penetration of Twitter. Homophobia on the ground may not be reflected online by means of Twitter.* Functional- The graph serves to display geographical patterns of online hate speech in the US. With the exception of territories like Alaska, Hawaii and Puerto Rico, the patterns of hate speech are clearly discernible across all of the United States. The use of red and blue relies on the human perception of red as an indicator of danger and violence. This allows the most hateful areas to stand out. * Beautiful- The choice of colours in the chloropleth also proves to be visually appealing. * Insightful- Pockets of hateful speech are clearly spread across large parts of the US. Interestingly, the central regions of the country do not show any clear activity, possibly due to the lower use of Twitter in these areas (this is just a conjecture- it certainly merits further investigration) * Enlightening- In many media depictions, the East Coast is often represented as an area of higher proportions of liberal attitudes and graduation rates from high school and college. However, hateful speech is clearly widely prevalent even in these states.

While not perfect, this first visual gets a number of elements right.

Good Visual 2

The second visualization is from the Boston Magazine's web portal, and is available at the top of this page: 'https://www.bostonmagazine.com/news/2014/02/03/much-snow-america-cancels-school/'



It scores well on Cairo's parameters in the following way:

- Truthful- The data are derived from user responses and interpolations of meteorological data from the NOAA. The author also provides 5 caveats for a more nuanced understanding of the visuals at this link: https://www.reddit.com/r/MapPorn/comments/1wiacl/how_much_snow_it_typically_takes_to_cancel_school/
- Functional- The map serves to visualize the variation in the snowfall required for the announcement of cancellation of school.
- Beautiful- The use of blues and greens gives the chloropleth aesthetic appeal
- Insightful- We find that the pattern isn't as clear-cut is simply a colder climate implying higher snow requirements for school closure. For example, Utah lies further south than Washington, but has higher 'snow thresholds'
- Enlightening- The visual leads us to ask deeper questions on school policy, resulting in the previously enumerated caveats.

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

We would ideally use EDA very early in our analysis, or when encountering an anomaly in the core confirmatory analysis that we had decided to undertake.

It helps identify unexpected relationships or distributional patterns. For example, the average value of support for President Trump in ANES 2016 Data may not accurately represent the distinct patterns of support under different party affiliations.

It can also assist in the development of theories and models or refine the understanding of the underlying logical assumptions. For example, the application of a certain method may require a variable to be normally distributed. However, a simple histogram reveals this to not be the case. EDA can also be used post-hoc to explain any unexpected results of our confirmatory analyses.

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

Tukey described Exploratory Data Analysis as an attitude, and a flexibility. It involves "a willingness to look for what can be seen, whether or not anticipated". Furthermore, rather than a single technique, it is the "recognition that the picture examining-eye is the best finder we have of the wholly unanticipated." For example, the discovery of high snowfall cutoffs for school closure would surprise a newcomer to the country like me, who assumed northern states bore most of the brunt of the cold climate.

Tukey refers to Confirmatory Analysis as 'an incomplete paradigm' involving a linear flow between "question, design, collection, analysis and answer". More broadly, it refers to statistical methods undertaken to prove or disprove an already formed hypothesis, often relying on tools of inference, confidence and significance. For example, in Bad Visual 1, a researcher already predisposed to dislike Nicolas Cage could well undertake regression analysis to link the number of his movies with released in a year the number of pool deaths in the same year.

REFERENCES Alberto Cairo. 2016. The Truthful Art: Data, Charts, and Maps for Communication (1st ed.). New Riders Publishing, Thousand Oaks, CA, USA.