Problem Set 1

Adam Shelton

10/11/2019

Contents

1 Exploration and Computation

1.1 Loading Data

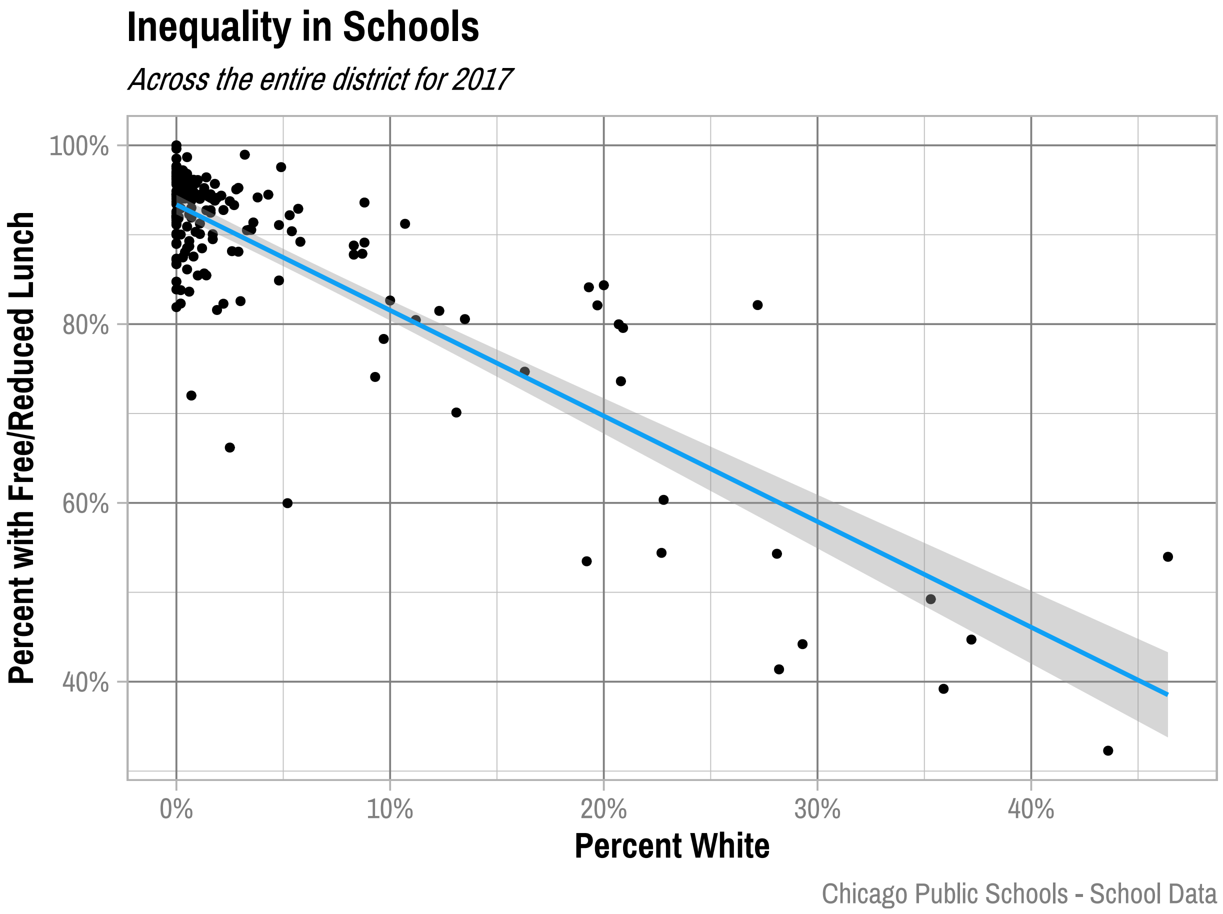
```r
cps_data = read_csv(here("Data", "cps_school_final.csv"))
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   network = col_character(),
##   surv_resp_rate = col_character(),
##   grades_offered = col_character(),
##   school_nm = col_character(),
##   school_typ = col_character(),
##   sch_str_prefix = col_character(),
##   sch_str = col_character()
## )

## See spec(...) for full column specifications.
```

1.2 A Visualization

```r
cps_data %>% filter(year == 2017) %>% ggplot(aes(x = pct_white,
    y = pct_frl), group = 1) + geom_point() + geom_smooth(method = "lm",
    color = color_pal(1, "cool")) + scale_x_continuous(labels = percent_format(accuracy = 1)) +
    scale_y_continuous(labels = percent_format(accuracy = 1)) +
    labs(title = "Inequality in Schools", subtitle = "Across the entire district for 2017",
        x = "Percent White", y = "Percent with Free/Reduced Lunch",
        caption = "Chicago Public Schools - School Data") + theme_master()
```



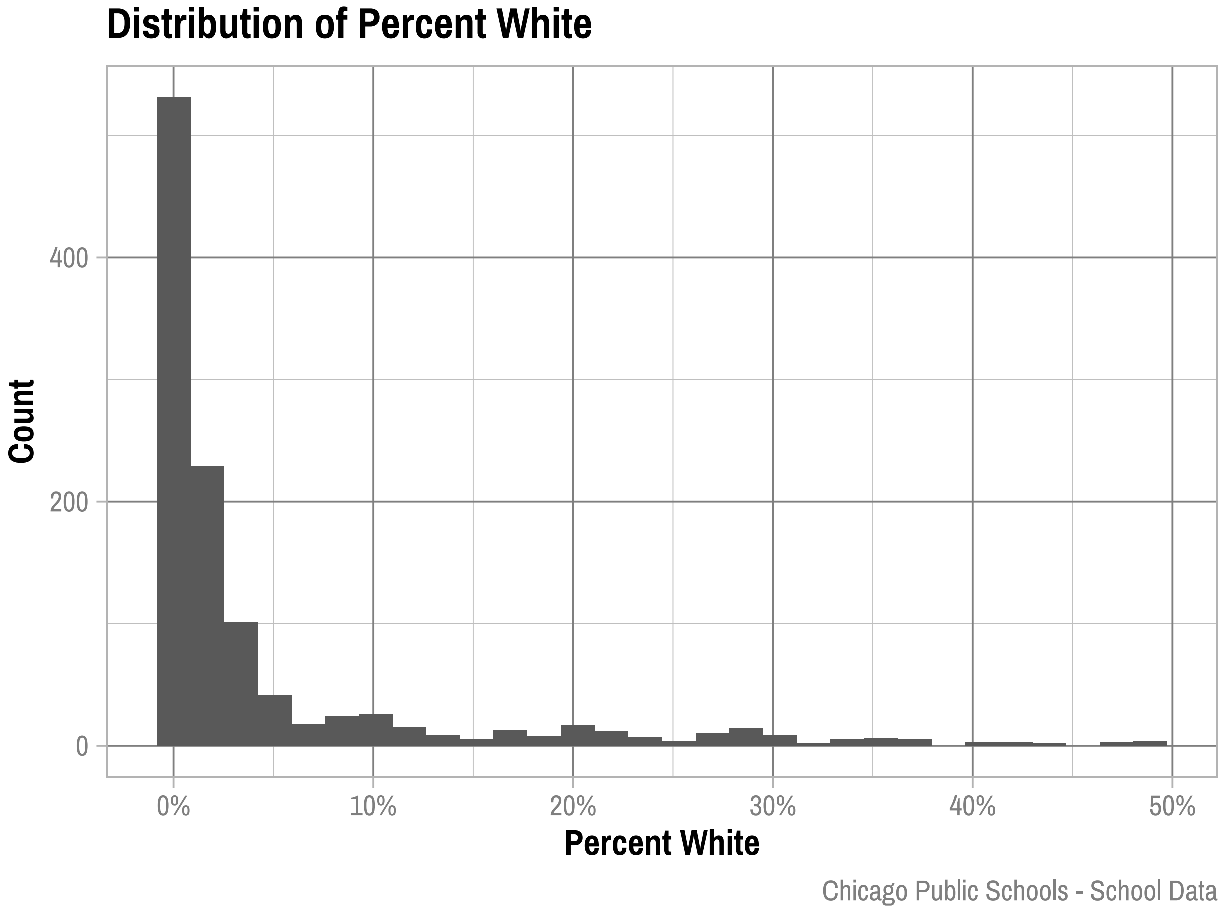
1.3 About the Variables

```r
cps_data %>% select(pct_white, pct_frl) %>% skim() %>% select(-level,
    -type, -value) %>% pivot_wider(names_from = stat, values_from = formatted) %>%
    select(-hist) %>% kable()
```

variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100
pct_white	0	1126	1126	0.048	0.09	0	0.001	0.01	0.038	0.49
pct_frl	0	1126	1126	0.88	0.13	0.31	0.86	0.93	0.96	1

```r
cps_data %>% ggplot(aes(x = pct_white)) + geom_histogram() +
    labs(title = "Distribution of Percent White", x = "Percent White",
      y = "Count", caption = "Chicago Public Schools - School Data") +
    theme_master() + scale_x_continuous(labels = percent_format(accuracy = 1))
```

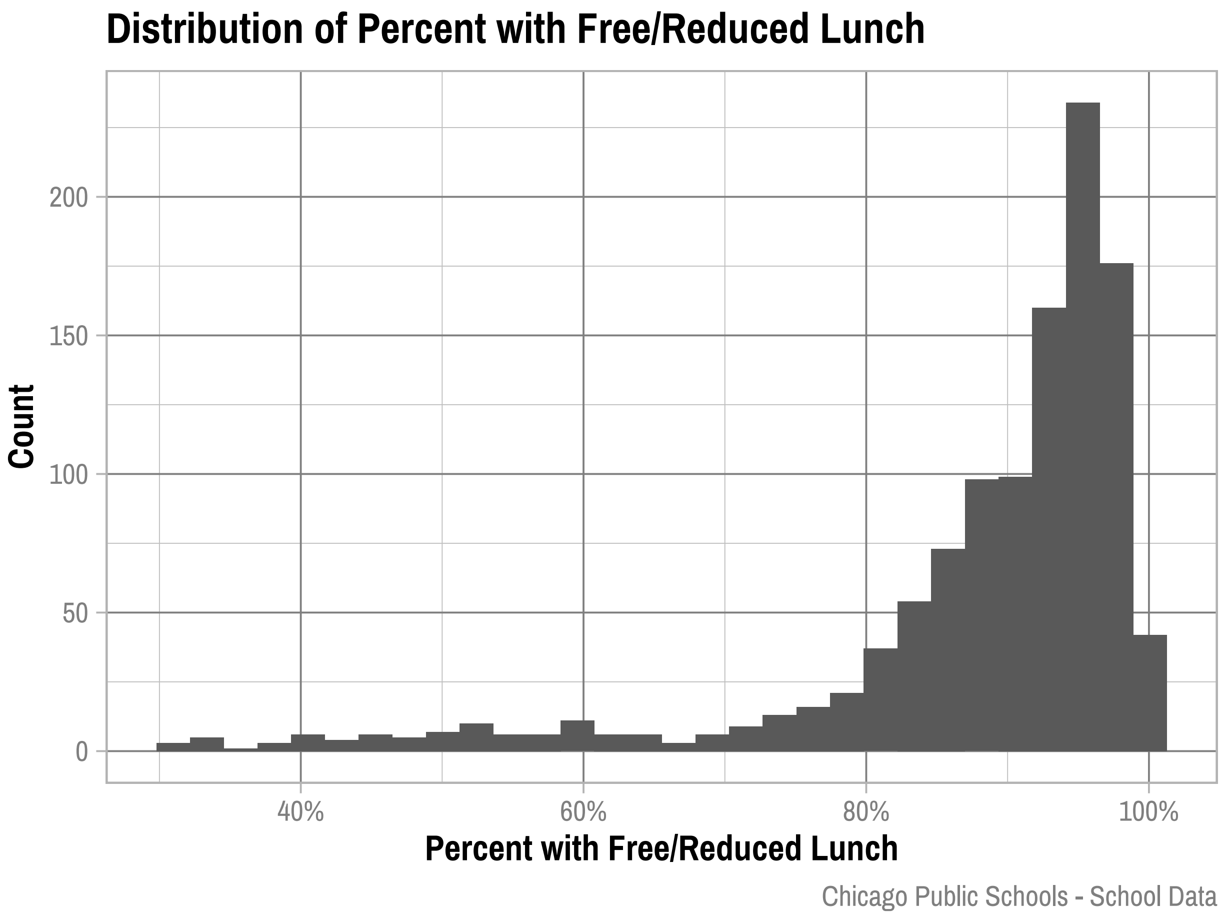
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Distribution of Percent White

Chicago Public Schools - School Data

```
cps_data %>% ggplot(aes(x = pct_frl)) + geom_histogram() + labs(title = "Distribution of Percent with F
    x = "Percent with Free/Reduced Lunch", y = "Count", caption = "Chicago Public Schools - School Data
    theme_master() + scale_x_continuous(labels = percent_format(accuracy = 1))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



1.4 Conclusion

Neither variable is normally distributed, and in fact, they are both heavily skewed, as the histograms show. It would appear that there are very few schools with large proportions of white students. This is important to understanding not only the data-set, but also the representativeness of CPS. While Chicago is roughly one-third white, it appears that most white children are not attending public schools. This creates unique challenges for battling inequality in K-12 education in Chicago. Free or Reduced Lunch Status is used to measure socioeconomic status, and this data shows, as we might expect, that socioeconomic status has a strong relationship with race.

2 Critical Thinking

2.1 Visual vs. Numeric

Both visual and numeric exploratory analyses are summaries of data. Numeric EDA can provide a much higher level of precision than a visual EDA typically can, but it also has a much higher probability of obfuscating an aspect of the data. For example, a measure of central tendency is simple to understand and quick to interpret, but an outlier or other sub-optimal distribution, results in that measure having less accuracy.

Furthermore, a visual analysis, allows for much more data to be interpreted at once, giving more opportunity for a relationship in the data to be found. However, this often comes at the cost of less precision. It might be clear that there is a negative quadratic relationship between two variables, but visual EDA would not be able to provide much more information than that.

This is why both numeric and visual methods should be used throughout an EDA. Both methods supplement the other, allowing the researcher to verify the claims either method shows. Using both numeric and visual EDA methods allows for a much more comprehensive understanding of the data, before a more rigorous analysis is performed.

2.2 Bad Visualizations

Figure 1 shows murders committed by guns in Florida, and at first glance appears to show a drop in gun deaths after 'Stand Your Ground' polices were passed in the state. However, the y-scale for this visualization is actually *inversed*, meaning that zero is at the top of the y axis instead of the bottom, where the origin is typically found on a line plot. Therefore it is quite easy to misinterpret this visualization, which appears to be the author's intent. While murders by guns actually *rose* drastically after 'Stand Your Ground' policies were passed in Florida, supporters of these policies often tout that they will reduce gun violence. In this case the data appears to contradict this claim, possibly motivating supporters to create a poor and misleading visualization to make the data appear to reinforce their claim. In addition to the reversed y-axis this visualization also has an red 'area' under the curve, although this visualization is in no way showing any sort of proportions.

Figure 2 shows whether voters for each county in the US 2016 presidential election, had a majority of votes for the Republican or Democratic candidate. This visualization is not inherently misleading, but it does allow a lot of room for misinterpretation. This misinterpretation is best exemplified by President Trump's tweeting of a similar map with the caption "Try to impeach this.", insinuating that the Democratic party's efforts to impeach the president would be fruitless, as he garnered a majority of votes in the 2016 election, and therefore has a strong majority of American's support. However, this interpretation of this visualization is incorrect, as President Trump actually lost the popular vote in the 2016, meaning the majority of Americans voted for his Democratic rival, Secretary Clinton.

Democratic candidates typically perform better in urban counties, which while much smaller in area, typically also have much larger populations, and therefore more voters. For example, one of the largest counties that voted for Trump in 2016, Sweetwater, Wyoming, had an area of 10,491 square miles and a population of 43,806, according to the 2010 US Census. While Kings County, New York, which a majority of votes went to Secretary Clinton in 2016, only had a land area of 97 square miles, but contained a population of 2.5 million people in 2010. So while this choropleth makes it appear that President Trump has much greater support than Democrats (or really the Democratic presidential candidate in 2016), this is inherently untrue due to the distribution of voters across land area in the United States.

Figure 3 shows unemployment percentages from January 2011 to November 2011 during the Obama Administration. Disregarding that some data points on this graph are plotted completely incorrectly (see how 8.8% in March is lower then 8.6% in November), there is no context for what should be taken away from this visualization. Is the jump in unemployment rates in the summer months significant and/or unusual, and does this jump correspond to some policy change by the Obama administration? Are these good or bad unemployment rates for the current state of the economy? Is unemployment during the Obama administration doing better or worse than other administrations during similar economic conditions? Even if the data on this visualization was presented accurately, it would still be a very ineffective visualization, as more information is required to make it insightful in any manner.

2.3 Good Visualizations

Figure 4 contains a visualization comprised of stacked bar charts of Likert scores from survey responses. This visualization does many things right. The stacked bar charts show percentages so each bar chart is easily comparable to the others. Each bar chart also has a bracket below, combining the four raw categories

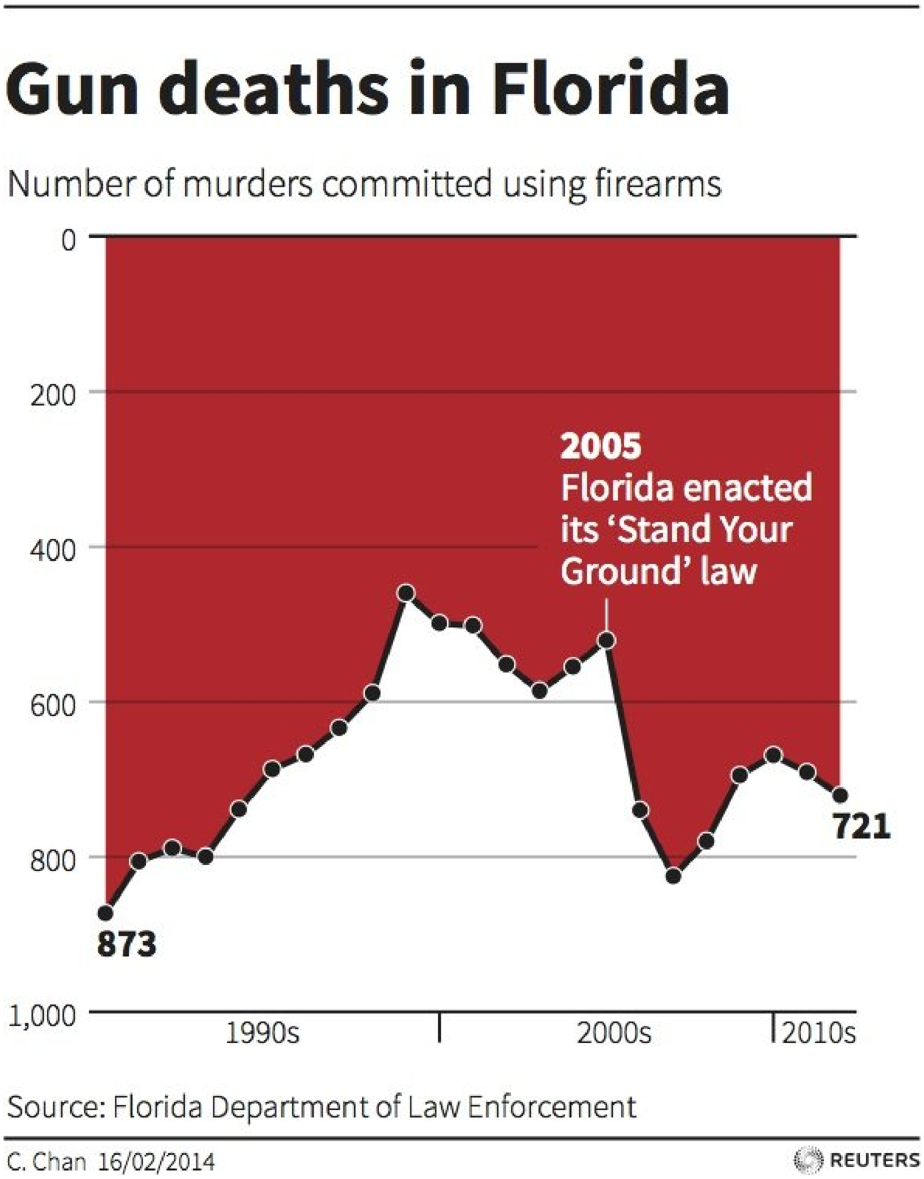


Figure 1: Florida Stand Your Ground Visualization - Source: https://www.livescience.com/45083-misleading-gun-death-chart.html

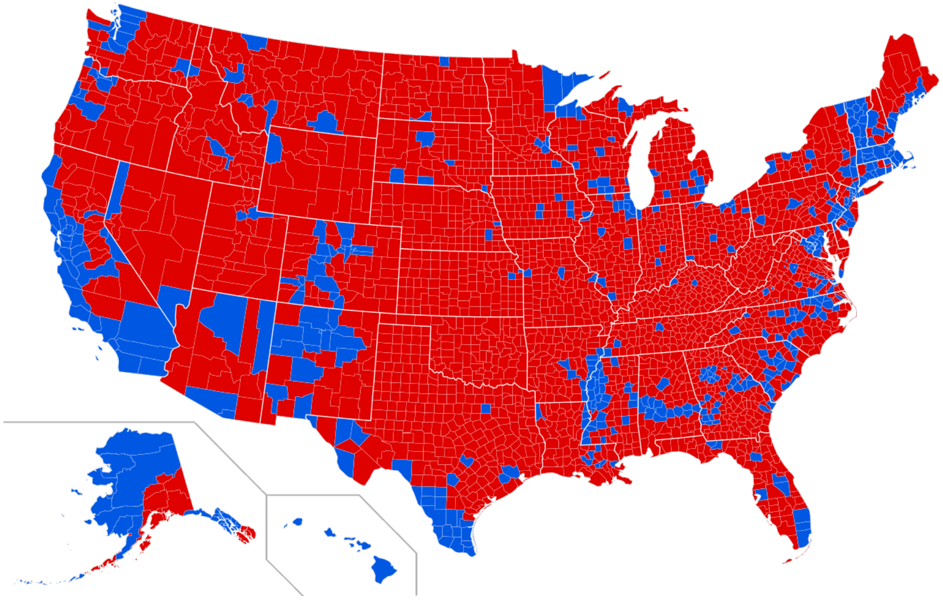


Figure 2: 2016 Presidential Election by County - Source: Wikipedia

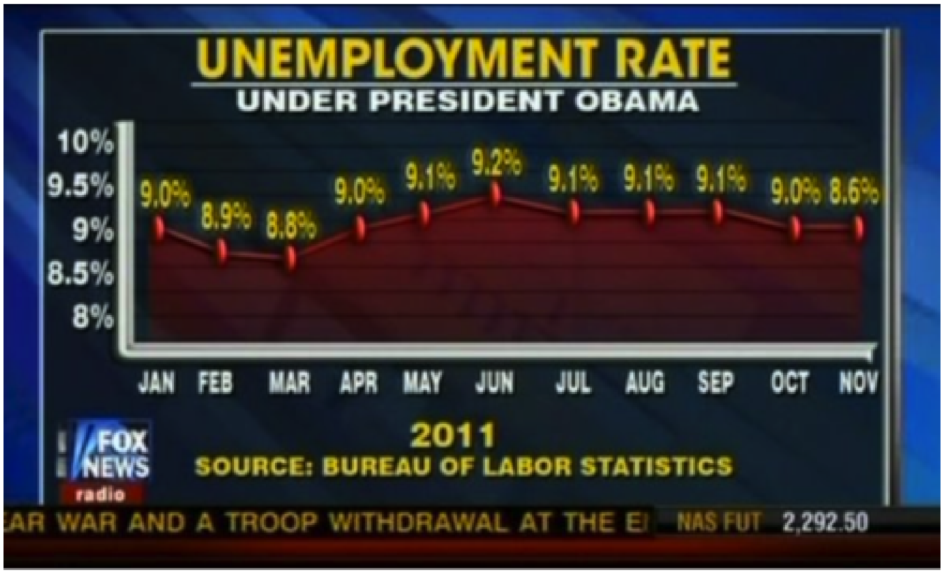
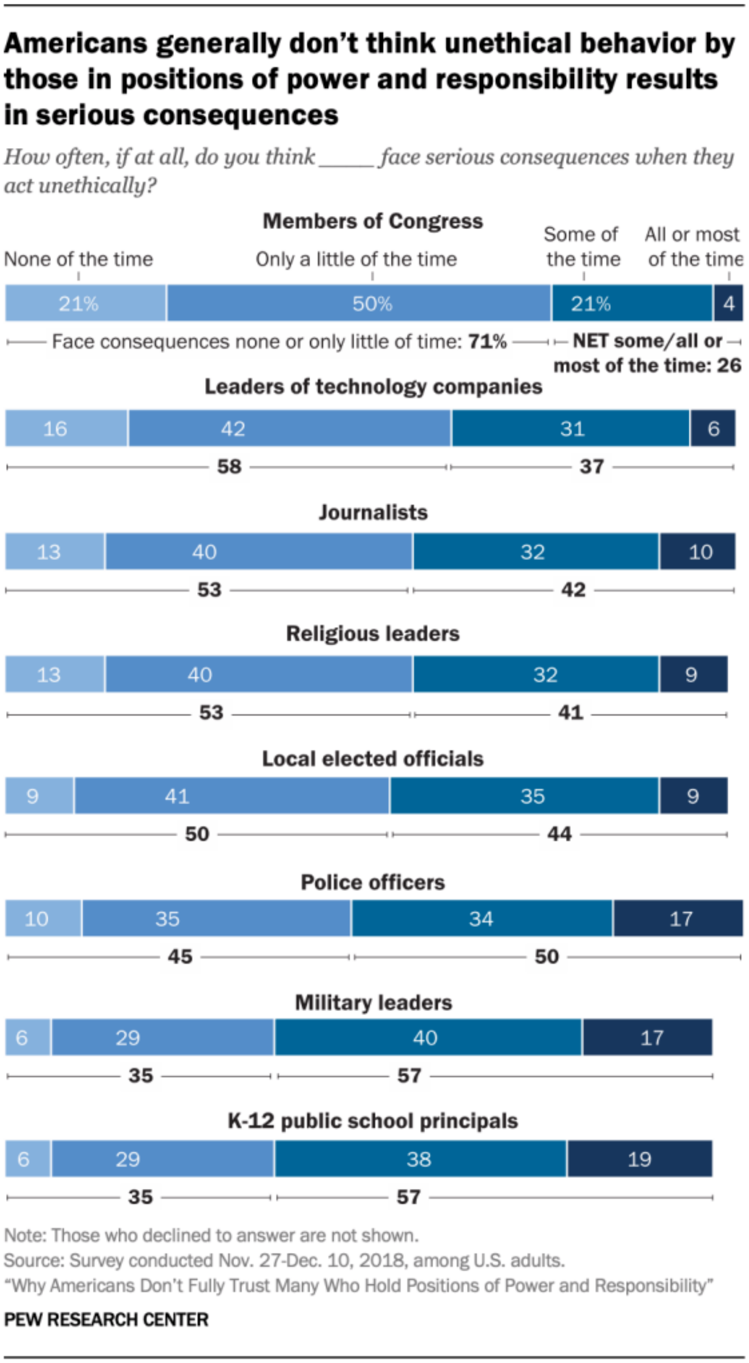


Figure 3: Fox News Obama Unemployment Graph - Source: https://rockthevizcomm.com/2012/04/24/ask-what-makes-a-good-data-visualization/



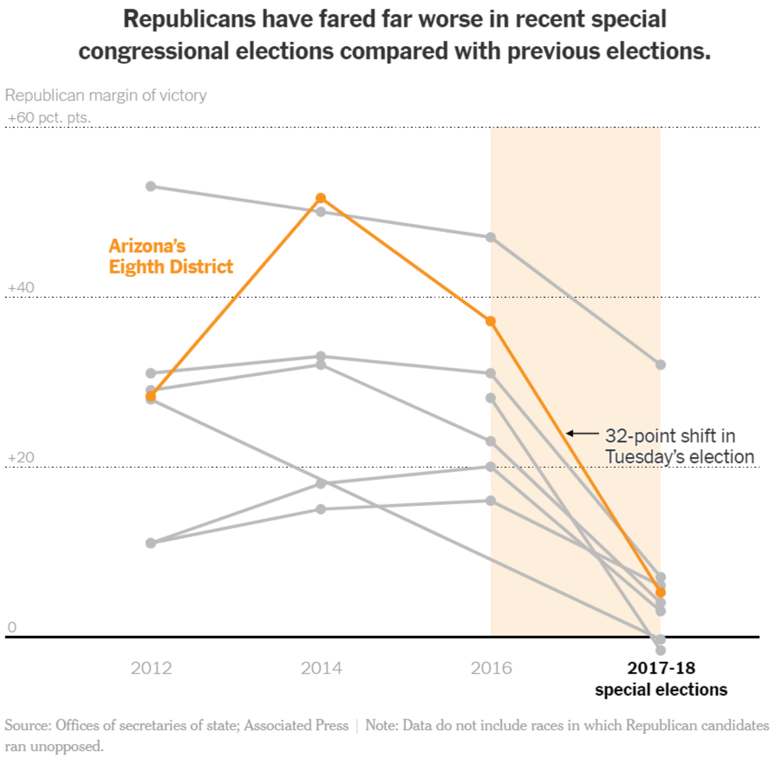
Americans generally don’t think unethical behavior by those in positions of power and responsibility results in serious consequences

How often, if at all, do you think _____ face serious consequences when they act unethically?

Note: Those who declined to answer are not shown.
Source: Survey conducted Nov. 27-Dec. 10, 2018, among U.S. adults.
“Why Americans Don’t Fully Trust Many Who Hold Positions of Power and Responsibility”

PEW RESEARCH CENTER

Figure 4: Pew Insitutional Trust Visualization - Source: Pew Research Center


Republicans have fared far worse in recent special congressional elections compared with previous elections.

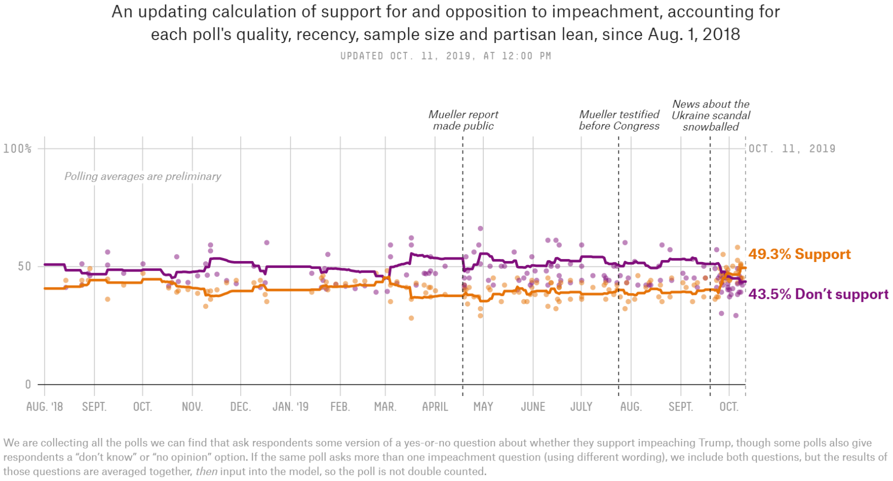
Source: Offices of secretaries of state; Associated Press | Note: Data do not include races in which Republican candidates ran unopposed.

Figure 5: New York Times Republican Performance in Special Elections Visualization - Source: The New York Times

from the survey into a binary rarely/often categorization, to make it easier to interpret the data. All the charts are sorted with those towards the top showing those with the most "cynical" responses (less likely to face consequences) toward the top. In general there is a clear aesthetic profile in layout, typography, and color palette. There is also a clearly stated source for the data on the visualization.

Figure 5 shows the margin of Republican victories for congressional districts with previous special elections. This visualization does a great job of not only showing overall trends but highlighting what is noteworthy. For example the time period under the Trump Administration has a clearly shaded background to highlight the drop in Republican margin of victories across all special elections. Arizona's Eighth District, which is notable in context of the article this visualization is from, is highlighted in a bright mustard yellow, contrasting against all the other data which is a dull gray. This district is specifically called out for its 32 point shift, which is clearly annotated on the visualization. Again, there is a clear aesthetic profile in layout, typography, and color palette. There is also a clearly stated source for the data on the visualization.

Figure 6 shows the predicted support from FiveThirtyEight's aggregate modeling of favorability polling of the impeachment of President Trump. As researchers are aggregating this data from third-party polls, FiveThirtyEight shows each individual poll as a point on the chart, and a line for their aggregation model's prediction of actual support. This is a welcome addition to increase the transparency of the data behind this visualization. The axes and scales are all sensible and well labeled, and notable events that are have likely to shifted public opinion are annotated on the visualization for reference. This visualization is interactive, allowing the viewer to mouse over each time point on the graph and see the exact modeled percentages in text for that time period. The contrasting colors for 'Support' and 'Don't support' make it clear that these are inverse measures. Again, there is a clear aesthetic profile in layout, typography, and color palette. There is also a clearly stated source for the data on the visualization.



An updating calculation of support for and opposition to impeachment, accounting for each poll's quality, recency, sample size and partisan lean, since Aug. 1, 2018

UPDATED OCT. 11, 2019, AT 12:00 PM

We are collecting all the polls we can find that ask respondents some version of a yes-or-no question about whether they support impeaching Trump, though some polls also give respondents a "don't know" or "no opinion" option. If the same poll asks more than one impeachment question (using different wording), we include both questions, but the results of those questions are averaged together, *then* input into the model, so the poll is not double counted.

Figure 6: FiveThirtyEight Support for Trump Impeachment Visualization - Source: FiveThirtyEight

2.4 Using EDA

EDA is an essential first step to most research projects. Using EDA gives us a stronger idea of what data we have, how it was gathered, if it is representative, and what relationships there are in it. Any modeling process is dependent on an understanding of the data, not only to make sure the data meets any assumptions for the model method used, but also to aid in the interpretation of any models built from the data. EDA can also help researchers to understand how they might need to wrangle the data to get it into the right form for a specific analysis, whether that is further EDA or a more advanced modeling process.

2.5 Confirmatory vs. Exploratory

John Tukey felt that traditionally researchers had focused too much on confirmatory data analysis methods, like hypothesis testing. Confirmatory methods take specific assumptions and impose a certain form/relationship on the data. While these assumptions and imposed restrictions may be valid, it can also distort or restrict the perception of the data. By looking at this narrow sliver of the data that is a confirmatory method, other relationships entirely may be missed.

Exploratory data analysis methods allow the data to do the talking. In this way, as much of the data is considered at once as possible, with very loose forms being imposed on the possible relationships in the data. Instead of asking "is this relationship/pattern valid" exploratory data analysis is like asking "what relationships/patterns are there".