Kim_problemset1

Problem Set 1: Exploratory Data Analysis

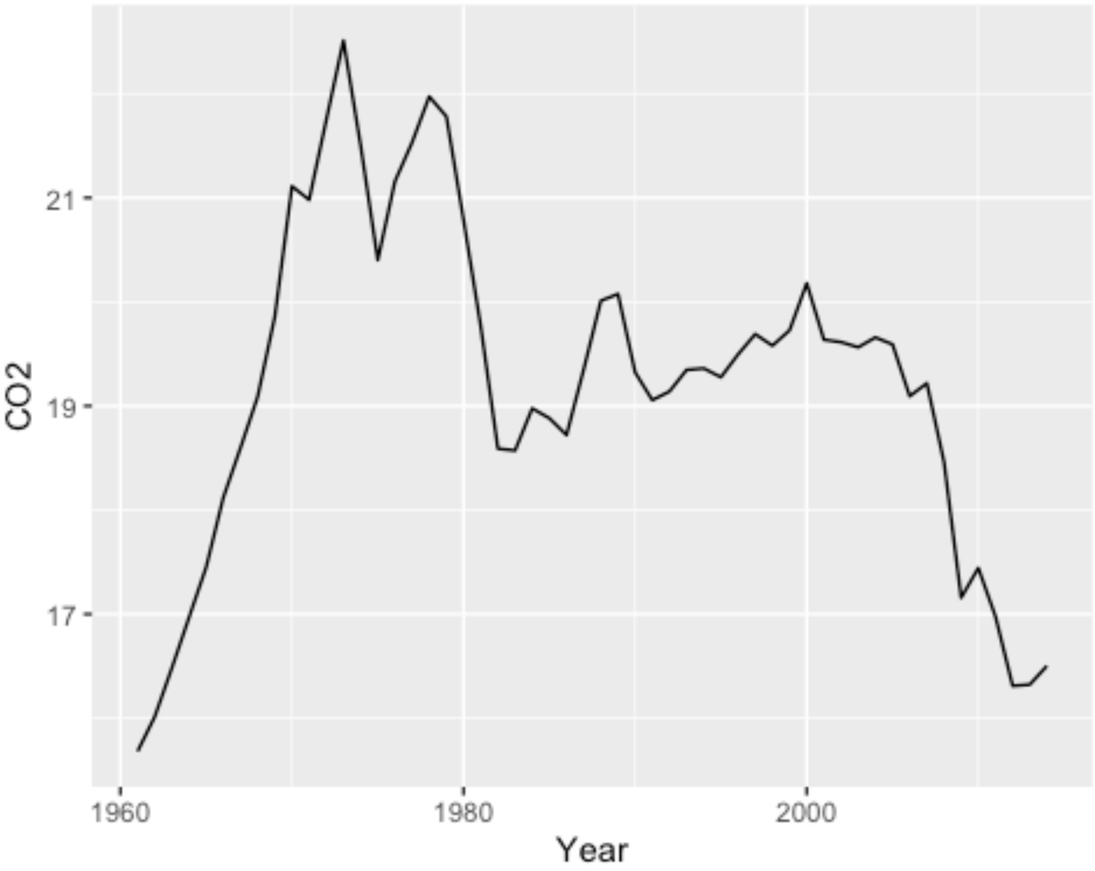
Exploration & Computation

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

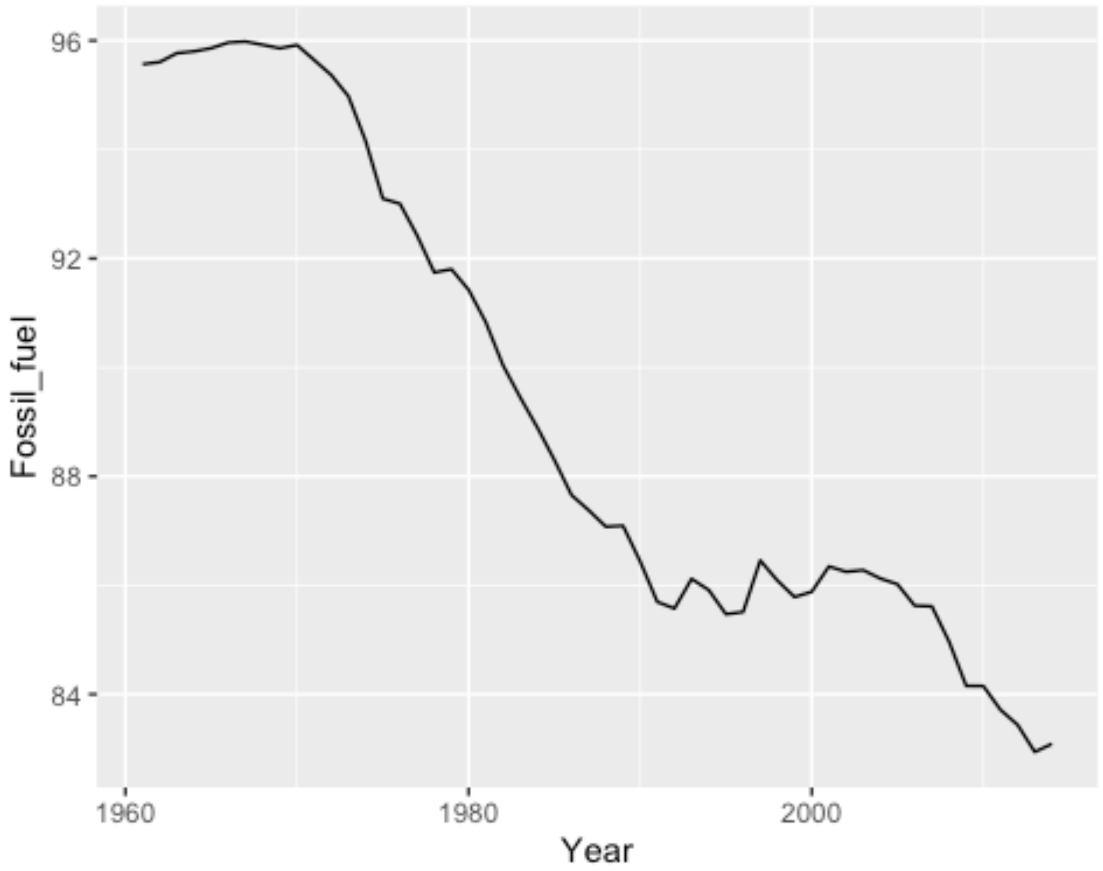
```
library(tidyverse)

## ── Attaching packages ─────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.2.1     ✔ purrr   0.3.2
## ✔ tibble  2.1.3     ✔ dplyr   0.8.3
## ✔ tidyr   1.0.0     ✔ stringr 1.4.0
## ✔ readr   1.3.1     ✔ forcats 0.4.0

## ── Conflicts ────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()

library(skimr)

##
## Attaching package: 'skimr'

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

WDI_dta<-read.csv("~/Desktop/Problem-Set-1/WDIData.csv")

skim(WDI_dta)

## Skim summary statistics
##  n obs: 54
##  n variables: 4
##
## ── Variable type:integer ───────────────────────────
##  variable missing complete  n   mean    sd   p0     p25    p50     p75
##      Year       0       54 54 1987.5 15.73 1961 1974.25 1987.5 2000.75
##  p100     hist
##  2014 ▇▇▇▇▇▇▇▇
##
## ── Variable type:numeric ───────────────────────────
##     variable missing complete  n  mean   sd    p0   p25   p50   p75  p100
##          CO2       0       54 54 19.19 1.62 15.68 18.58 19.35 19.97 22.51
##   Fossil_fuel       0        54 54 89.3   4.46 82.94 85.81 87.24 93.89 95.98
##    GDP_growth       0        54 54   3.09 2.13 -2.54   1.98   3.38   4.47   7.24
##       hist
##   ▂▁▁▄▆▂▂▂
##   ▂▅▃▂▁▂▂▅
##   ▁▂▁▂▅▅▃▂
```

2. Choose a visual technique to illustrate your data (e.g., barplot, histogram, scatterplot).

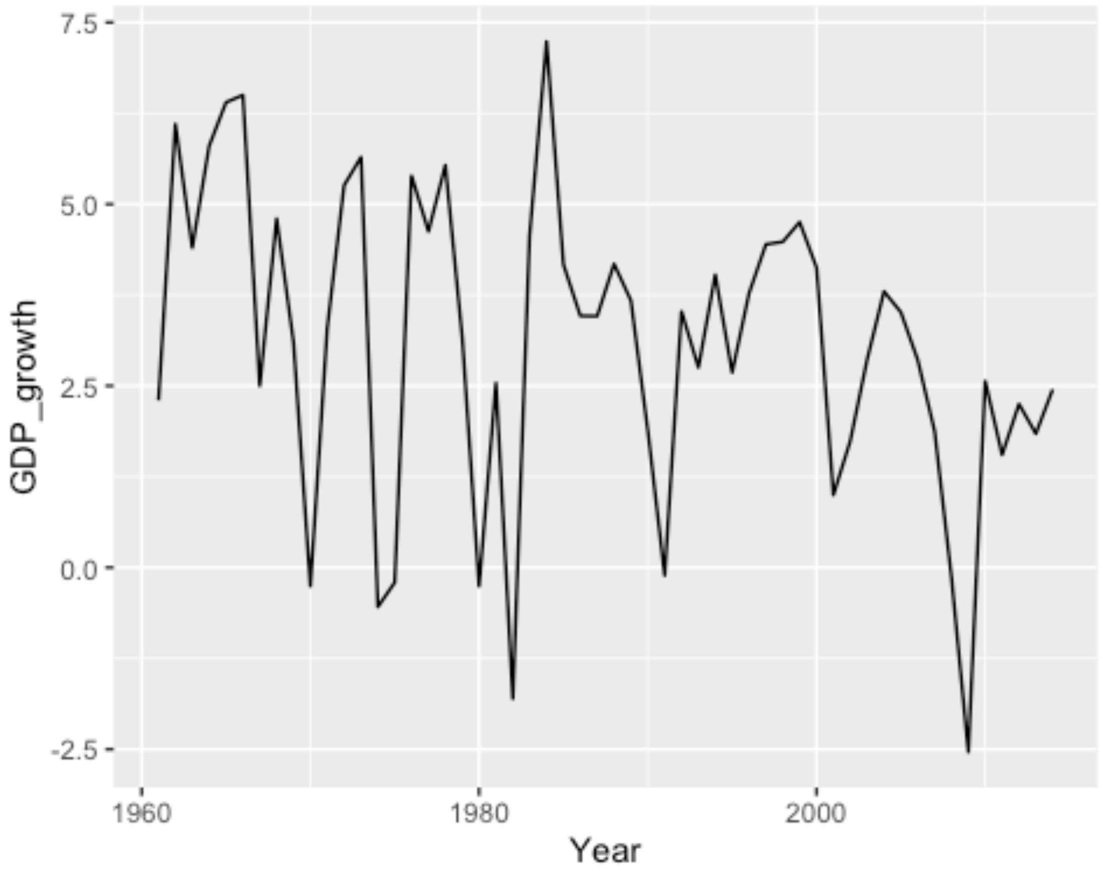
```
WDI_dta %>%
ggplot(aes(x=Year, y=CO2)) +
  geom_line()
```



```
WDI_dta %>%
ggplot(aes(x=Year, y=Fossil_fuel)) +
  geom_line()
```



```
WDI_dta %>%
ggplot(aes(x=Year, y=GDP_growth)) +
  geom_line()
```



3. Now generate and present the visualization and describe what you see.

The level of CO2 emissions has a dramatic increase reaching a peak around the 1970s. However, it experiences two steep decreases in the 1980s and 2010s. From these dramatic decreases, one might suggest that there were government regulations or campaigns to restrict CO2 emissions around the time of the 1980s and 2010s.

Overall, fossil fuel energy consumption seems to consistently decrease. However, it stays quite steady in 1960-1970 and 1990-2010. After this steady stage, it drops approximately in 1970-2010 and after 2000. This decreasing trend is occurring over a similar timeline with the CO2 emissions level. Comparing the CO2 level and fossil fuel energy consumption, one might suggest that fossil fuel is one of the significant contributors to CO2 emissions.

The GDP growth is not steady but fluctuates almost every year, which makes it difficult to tell if there is a decreasing trend over the years. Therefore, unlike the other graphs, no clear increasing or decreasing tendency can be found.

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

```
mean(WDI_dta$CO2)

## [1] 19.193
```

```
var(WDI_dta$CO2)

## [1] 2.628577

mean(WDI_dta$Fossil_fuel)

## [1] 89.30125

var(WDI_dta$Fossil_fuel)

## [1] 19.87487

mean(WDI_dta$GDP_growth)

## [1] 3.093133

var(WDI_dta$GDP_growth)

## [1] 4.519087
```

The results show that the fossil fuel energy consumption has the largest variation of the three variables. This may suggest that the energy consumption of fossil fuel has changed more over the past fifty years. However, since the variables (CO2, GDP, fossil fuel) do not use the same scale, it cannot be judged whether the relatively large variance of the fossil fuel is meaningful.

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 graph of the CO2 emissions level seems to suggest some dramatic changes over the last 50 years. However, this is not clear when we look at the variance of 2.63. This might be because I calculated the variance over the entire years rather than some range of it. If a certain period was chosen to calculate the variance, it might have been larger. Thus, only looking at the numeric values can result in missing out the changes of CO2 emissions over the years.

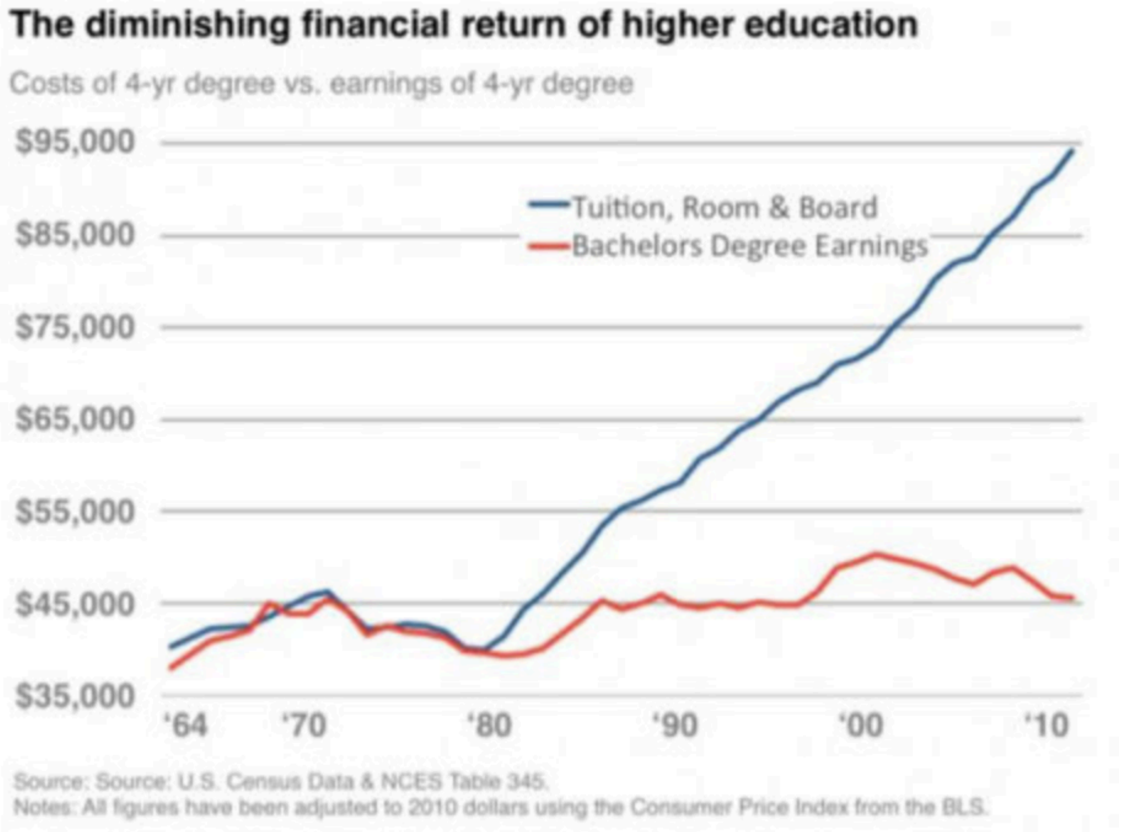
Looking at the numeric value for the GDP growth, there seems not to have much change, with a possible small variation (we cannot tell if this variance is truly small or not). However, the graph shows that it fluctuates over the years. As seen in this example, numeric descriptions of the data do not always give a full explanation of what the data look like. Although numeric values help to check for the statistical significance, it needs to be combined with visual analysis.

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).

Numeric exploratory data analysis reveals the exact difference or significance of numbers in the data. For instance, a p-value can indicate exactly how significant the difference between the two distributions is. However, numeric analysis requires many assumptions. Visual exploratory data analysis can be used to confirm these assumptions, enabling researchers to see how the data behave in general. As an example, one can easily figure out a general trend of the data or outliers by looking at a scatter plot, which cannot be found in the numeric analysis.

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

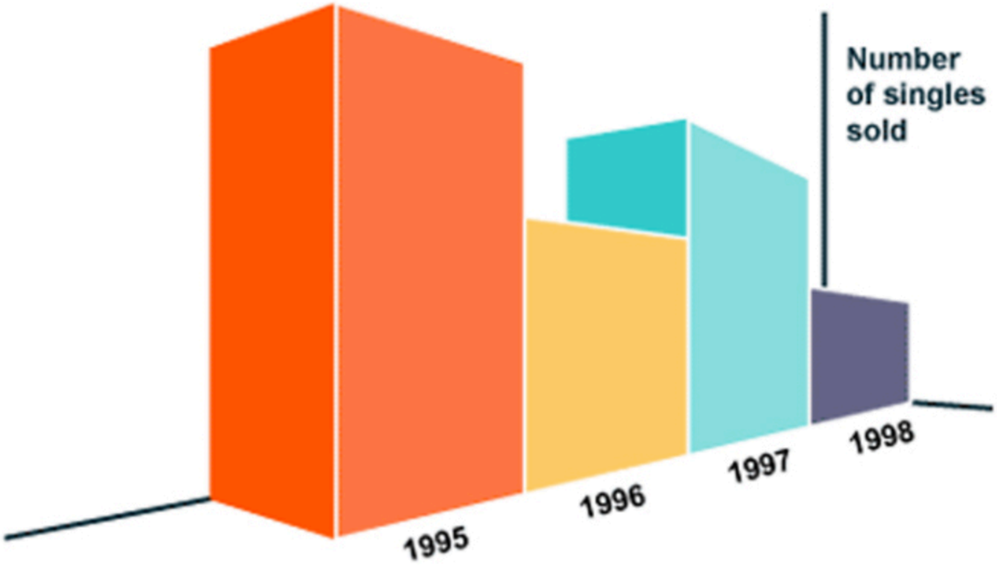

The diminishing financial return of higher education

Costs of 4-yr degree vs. earnings of 4-yr degree

Source: Source: U.S. Census Data & NCES Table 345.
Notes: All figures have been adjusted to 2010 dollars using the Consumer Price Index from the BLS.

Bad Visualziation 1

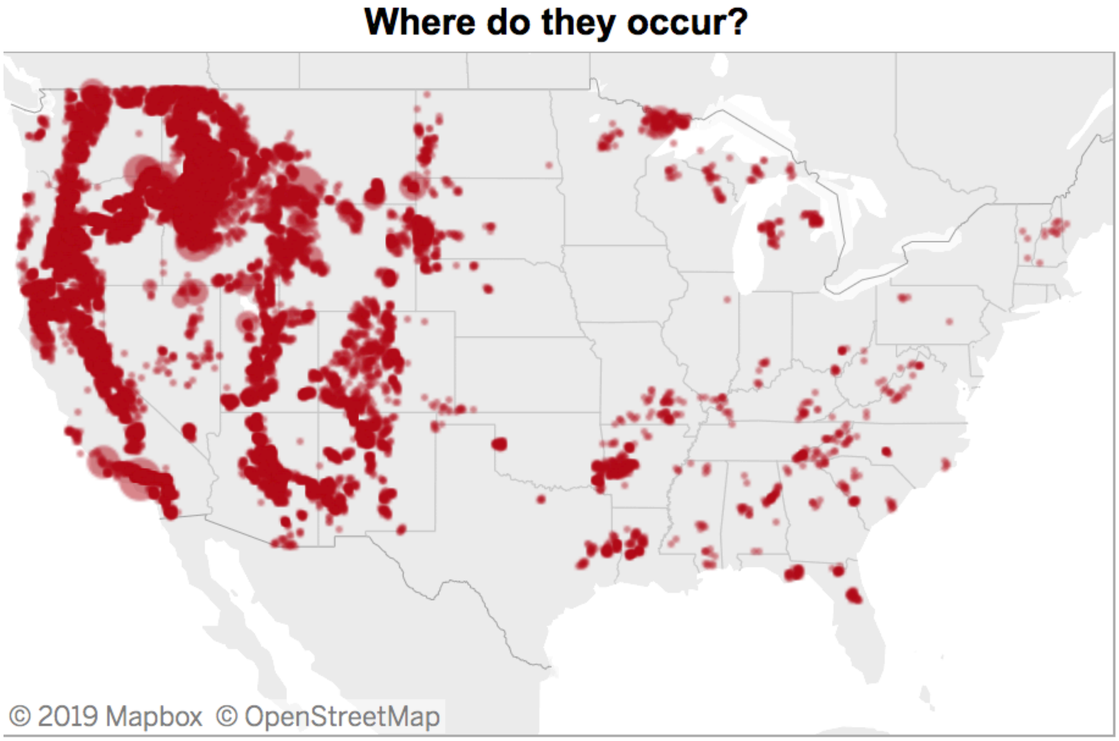
This graph indicates that the earnings of people with a bachelor's degree stay steady when the costs of tuition drastically increase over time. This implies that earning a bachelor's would not be financially beneficial. However, this graph does not show other groups to compare with. We cannot tell how much salary people without a bachelor's degree are earning. People without a degree might be doing even worse than those with a degree. URL:https://i.kinja-img.com/gawker-media/image/upload/c_fit,fl_progressive,q_80,w_470/jb5281qnqdobwumsgehx.jpg



Bad Visualziation 2

3-D graphs can be used to present the information more efficiently. However, they can be often misleading and confusing. In this graph, it is hard to compare the bars because the bar for 1995 is more exaggerated than the others. Even though the bar for 1995 is similar to the one for 1997, 1995 looks larger because of the 3-D effect. URL:https://visme.co/blog/wp-content/uploads/2016/02/Untitled-3.jpg

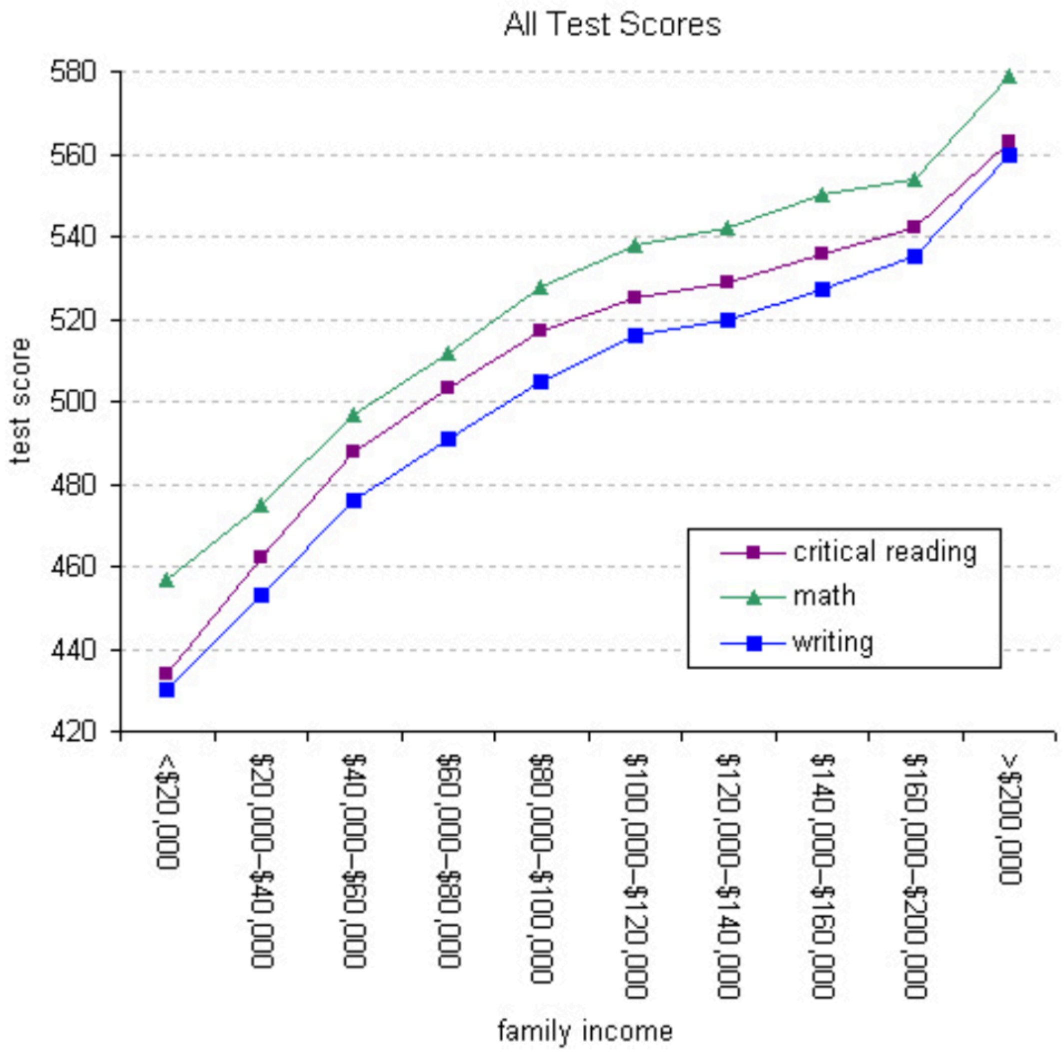
3. Find (and include) two examples of “good” visualizations and tell me precisely why they’re good.


Where do they occur?

© 2019 Mapbox © OpenStreetMap

Good Visualziation 1

This visualization captures where the fire occurred in 2002-2012. By spotting the fires on the map, it efficiently shows adjacent areas of the fire origin. This can help to figure out the usual path of fire and prevent reoccurrence of fire. URL:https://public.tableau.com/en-us/gallery/forest-fire-hot-spots



Source: College Board

Good Visualziation 2

This line graph shows the relationship between family income and SAT scores for the three sections, critical reading, math, and writing. This graph efficiently shows the increasing trend for all sections of the SAT with family income. Moreover, one can easily see how much the SAT score increases over family income by looking at the scale. URL:https://economix.blogs.nytimes.com/2009/08/27/sat-scores-and-family-income/?_php=true&_type=blogs&_php=true&_type=blogs&_r=1

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

It might be a good idea to use EDA as the first step to analysis without any assumptions about the data. With certain assumptions, you might look at a specific part of the data because you already know what to look for. However, exploring the data using EDA, you might stumble on something you did not realize there was.

Also, EDA is more helpful and better to understand the data. Humans are more equipped to find patterns in images or visuals than in numbers. Therefore, exploring graphs and many visualizations of the data will make the data clear and accessible to you.

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

Tukey describes confirmatory data analysis as a “routine” that is “easy to computerize.” Statistics can be easily calculated by computers. In fact, computers outperform humans in precision and speed.

Moreover, the confirmatory analysis starts with a specific question. Researchers design methods, collect data, and analyze them to answer the question. Therefore, they already know what to look for before having the data, which makes the analaysis narrow. For example, the confirmatory analysis is often used by psychologists to answer specific questions such as the age difference in memory. They specify how to measure memory capacity and target the data for different age groups.

On the other hand, Tukey treats the exploratory data analysis as an “attitude”, “flexibility”, and “some graph paper”. It is crucial to actively look for what is in the data, not merely taking what one needs from them. To better understand the data, one needs to explore even what they did not think important without holding any assumptions.